UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

CENTRO DE CIENCIAS BÁSICAS

DOCTORADO EN
CIENCIAS APLICADAS Y TECNOLOGÍA

TESIS

Light Data Science - Analytics Methodology (LDSAM): an aligned ISO/IEC 29110 – Basic Profile – Development Methodology for Big Data Software Systems in Small Business

PRESENTA

MITC. David Alejandro Montoya Murillo

TUTOR
Dr. José Manuel Mora Tavarez

CO-TUTOR
Dr. Sergio Galván Cruz

COMITÉ TUTORAL
Dr. Ángel Eduardo Muñoz Zavala
Dr. Francisco Javier Álvarez Rodríguez
Dra. Estela Lizbeth Muñoz Andrade

Cd. Universitaria, Aguascalientes, Ags. 28 de mayo de 2025

MTRO. EN C. JORGE MARTÍN ALFÉREZ CHÁVEZ
DECANO (A) DEL CENTRO CIENCIAS BÁSICAS

PRESENTE

Por medio del presente como **TUTOR** designado del estudiante **DAVID ALEJANDRO MONTOYA MURILLO** con ID 129369 quien realizó la tesis titulado: **LIGHT DATA SCIENCE - ANALYTICS METHODOLOGY (LDSAM): AN ALIGNED ISO/IEC 29110 – BASIC PROFILE – DEVELOPMENT METHODOLOGY FOR BIG DATA SOFTWARE SYSTEMS IN SMALL BUSINESS**, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que él pueda proceder a imprimirla así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

ATENTAMENTE
"Se Lumen Proferre"
Aguascalientes, Ags., a 28 de mayo de 2025.

Dr. José Manuel Mora Tavarez
Tutor de tesis

c.c.p.- Interesado
c.c.p.- Secretaría Técnica del Programa de Posgrado

MTRO. EN C. JORGE MARTÍN ALFÉREZ CHÁVEZ
DECANO (A) DEL CENTRO CIENCIAS BÁSICAS

P R E S E N T E

Por medio del presente como **CO-TUTOR** designado del estudiante **DAVID ALEJANDRO MONTOYA MURILLO** con ID 129369 quien realizó la tesis titulado: **LIGHT DATA SCIENCE - ANALYTICS METHODOLOGY (LDSAM): AN ALIGNED ISO/IEC 29110 – BASIC PROFILE – DEVELOPMENT METHODOLOGY FOR BIG DATA SOFTWARE SYSTEMS IN SMALL BUSINESS**, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que él pueda proceder a imprimirla así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 28 de mayo de 2025.

Dr. Sergio Galván Cruz
Co-Tutor de tesis

c.c.p.- Interesado
c.c.p.- Secretaría Técnica del Programa de Posgrado

MTRO. EN C. JORGE MARTÍN ALFÉREZ CHÁVEZ
DECANO (A) DEL CENTRO CIENCIAS BÁSICAS

PRESENTE

Por medio del presente como **ASESOR** designado del estudiante **DAVID ALEJANDRO MONTOYA MURILLO** con ID 129369 quien realizó la tesis titulado: **LIGHT DATA SCIENCE - ANALYTICS METHODOLOGY (LDSAM): AN ALIGNED ISO/IEC 29110 – BASIC PROFILE – DEVELOPMENT METHODOLOGY FOR BIG DATA SOFTWARE SYSTEMS IN SMALL BUSINESS**, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que él pueda proceder a imprimirla así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

ATENTAMENTE
"Se Lumen Proferre"
Aguascalientes, Ags., a 28 de mayo de 2025.

Dr. Ángel Eduardo Muñoz Zavala
Asesor de tesis

c.c.p.- Interesado
c.c.p.- Secretaría Técnica del Programa de Posgrado

MTRO. EN C. JORGE MARTÍN ALFÉREZ CHÁVEZ
DECANO (A) DEL CENTRO CIENCIAS BÁSICAS

PRESENTE

Por medio del presente como **ASESOR** designado del estudiante **DAVID ALEJANDRO MONTOYA MURILLO** con ID 129369 quien realizó la tesis titulado: **LIGHT DATA SCIENCE - ANALYTICS METHODOLOGY (LDSAM): AN ALIGNED ISO/IEC 29110 – BASIC PROFILE – DEVELOPMENT METHODOLOGY FOR BIG DATA SOFTWARE SYSTEMS IN SMALL BUSINESS**, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que él pueda proceder a imprimirla así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

ATENTAMENTE
"Se Lumen Proferre"
Aguascalientes, Ags., a 28 de mayo de 2025.

Dr. Francisco Javier Álvarez Rodríguez
Asesor de tesis

c.c.p.- Interesado
c.c.p.- Secretaría Técnica del Programa de Posgrado

MTRO. EN C. JORGE MARTÍN ALFÉREZ CHÁVEZ
DECANO (A) DEL CENTRO CIENCIAS BÁSICAS

P R E S E N T E

Por medio del presente como **ASESOR** designado del estudiante **DAVID ALEJANDRO MONTOYA MURILLO** con ID 129369 quien realizó la tesis titulado: **LIGHT DATA SCIENCE - ANALYTICS METHODOLOGY (LDSAM): AN ALIGNED ISO/IEC 29110 – BASIC PROFILE – DEVELOPMENT METHODOLOGY FOR BIG DATA SOFTWARE SYSTEMS IN SMALL BUSINESS**, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que él pueda proceder a imprimirla así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 28 de mayo de 2025.

Dra. Estela Lizbeth Muñoz Andrade
Asesor de tesis

c.c.p.- Interesado
c.c.p.- Secretaría Técnica del Programa de Posgrado

Universidad Autónoma de Aguascalientes

DICTAMEN DE LIBERACION ACADEMICA PARA INICIAR LOS TRAMITES DEL EXAMEN DE GRADO

Fecha de dictaminación dd/mm/aaaa: 4 de junio del 2025

NOMBRE: David Alejandro Montoya Murillo **ID** 129369

PROGRAMA: Doctorado en Ciencias Aplicadas y Tecnología **LGAC (del posgrado):** Tecnologías de Ingeniería de Software y Objetos de Aprendizaje

TIPO DE TRABAJO: (X) Tesis ( ) Trabajo Práctico

TITULO: Light Data Science - Analytics Methodology (LDSAM): an aligned ISO/IEC 29110 – Basic Profile – Development Methodology for Big Data Software Systems in Small Business

IMPACTO SOCIAL (señalar el impacto logrado): Proporcionar una metodología de desarrollo ligera y sin costo para proyectos de análisis de datos pequeños a las pequeñas y medianas empresas mexicanas.

INDICAR SI NO N.A. (NO APLICA) SEGÚN CORRESPONDA:

	Elementos para la revisión académica del trabajo de tesis o trabajo práctico:
SI	El trabajo es congruente con las LGAC del programa de posgrado
SI	La problemática fue abordada desde un enfoque multidisciplinario
SI	Existe coherencia, continuidad y orden lógico del tema central con cada apartado
SI	Los resultados del trabajo dan respuesta a las preguntas de investigación o a la problemática que aborda
SI	Los resultados presentados en el trabajo son de gran relevancia científica, tecnologíca o profesional según el área
SI	El trabajo demuestra más de una aportación original al conocimiento de su área
NO	Las aportaciones responden a los problemas prioritarios del país
SI	Generó transferecia del conocimiento o tecnológica
SI	Cumple con la ética para la investigación (reporte de la herramienta antiplagio)
	El egresado cumple con lo siguiente:
SI	Cumple con lo señalado por el Reglamento General de Docencia
SI	Cumple con los requisitos señalados en el plan de estudios (créditos curriculares, optativos, actividades complementarias,estancia, predoctoral, etc)
SI	Cuenta con los votos aprobatorios del comité tutoral, en caso de los posgrados profesionales si tiene solo tutorpodrá liberar solo el tutor
N.A	Cuenta con la carta de satisfacción del Usuario
SI	Coincide con el título y objetivo registrado
SI	Tiene congruencia con cuerpos académicos
SI	Tiene el CVU del Conacyt actualizado
SI	Tiene el artículo aceptado o publicado y cumple con los requisitos institucionales (en caso que proceda)
	En caso de Tesis por artículos científicos publicados
	Aceptación o Publicación de los articulos segun el nivel del programa
	El estudiante es el primer autor
	El autor de correspondencia es el Tutor del Núcleo Académico Básico
	En los artículos se ven reflejados los objetivos de la tesis, ya que son producto de este trabajo de investigación.
	Los artículos integran los capítulos de la tesis y se presentan en el idioma en que fueron publicados
	La aceptación o publicación de los artículos en revistas indexadas de alto impacto

Con base a estos criterios, se autoriza se continúen con los trámites de titulación y programación del examen de grado: Sí X No ____

FIRMAS

Elaboró:

* NOMBRE Y FIRMA DEL CONSEJERO SEGÚN LA LGAC DE ADSCRIPCION: Dr. Francisco Javier Álvarez Rodríguez

NOMBRE Y FIRMA DEL SECRETARIO TÉCNICO: Dr. Francisco Javier Álvarez Rodríguez

* En caso de conflicto de intereses, firmará un revisor miembro del NAB de la LGAC correspondiente distinto al tutor o miembro del comité tutoral, asignado por el Decano

Revisó:

NOMBRE Y FIRMA DEL SECRETARIO DE INVESTIGACIÓN Y POSGRADO: Dr. Alejandro Padilla Díaz

Autorizó:

NOMBRE Y FIRMA DEL DECANO: Mtro. en C. Jorge Martín Alférez Chávez

Nota: procede el trámite para el Depto. de Apoyo al Posgrado

En cumplimiento con el Art. 105C del Reglamento General de Docencia que a la letra señala entre las funciones del Consejo Académico: ... Cuidar la eficiencia terminal del programa de posgrado y el Art. 105F las funciones del Secretario Técnico, llevar el seguimiento de los alumnos.

Elaborado por: D. Apoyo al Posg.
Revisado por: D. Control Escolar/D. Gestión de Calidad.
Aprobado por: D. Control Escolar/ D. Apoyo al Posg.

Código: DO-SEE-FO-15
Actualización: 01
Emisión: 28/04/20

A Review of SDLCs for Big Data Analytics Systems in the Context of Very Small Entities Using the ISO/IEC 29110 Standard-Basic Profile

David Montoya-Murillo
Department of Electronic Systems
Autonomous University of Aguascalientes, Mexico
david.montoya@edu.uaa.mx

Manuel Mora
Department of Information Systems
Autonomous University of Aguascalientes, Mexico
jose.mora@edu.uaa.mx

Sergio Galvan-Cruz
Department of Electronic Systems
Autonomous University of Aguascalientes, Mexico
sergio.galvan@edu.uaa.mx

Angel Munoz-Zavala
Department of Statistics
Autonomous University of Aguascalientes, Mexico
eduardo.munoz@edu.uaa.mx

Francisco Alvarez-Rodriguez
Department of Computer Science
Autonomous University of Aguascalientes, Mexico
francisco.alvarez@edu.uaa.mx

Abstract: *Context: A Systems Development Life Cycle (SDLC) is a model of phases-activities, roles, and products systematically used to develop software with functional expected quality. Although SDLC is widely applied to various software types, it remains unusual in Big Data Analytics Systems (BDAS). Objective: To address this issue, several SDLCs for BDAS have been proposed, along with comparative studies, to guide interested organizations in adapting them. This research seeks a lightweight, balanced, and feasible for small development teams or organizations, taking advantage of favorable characteristics of the international ISO standard. Method and Materials: This study describes the knowledge gap by reporting a comparative analysis of four relevant SDLCs. A selective research method was applied (CRISP-DM, TDSP, BDPL, and DDSL), focusing on alignment with the recent ISO/IEC 29110-basic profilestandard. The goal was to identify which SDLC contributes and fits better from a lightweight approach. Results: From the rigorous approach Cross Industry Standard Process for Data Mining (CRISP-DM) showed the highest alignment with the standard, for the agile approach it was Domino Data Science Lifecycle (DDSL) being the closest of the four. Team Data Science Process (TDSP) stood out as the most agile of those analyzed but fell short of the required results. BDPL, which manages another standard, was too rigorous and more distant. Conclusions: Research on new SDLC for Big Data Project Lifecycle (BDPL) has been practically nonexistent in software engineering from 2000 to 2023. Only BDPL was found in the academic literature, while the other three came from gray literature. Despite the relevance of this topic for BDAS organizations, no adequate SDLC was identified.*

Keywords: *BDAS, SDLC, ISO/IEC-29110-standard for VSEs, CRISP-DM, BSPL, TDSP and DDSL.*

Received June 6, 2024; accepted November 7, 2024
https://doi.org/10.34028/iajit/22/1/15

1. Introduction

The systematic, disciplined, and quantified development of software has been guided by System Development Life Cycles (SDLCs) provided by the software engineering discipline [6]. An SDLC refers to “the software processes used to specify and transform software requirements into a deliverable software product” [6]. An SDLC is usually represented as a software development process model [26] of phases-activities, roles, and products proposed to build systematic software products on the expected time, budget, and functional quality, i.e., the named Iron Triangle [54]. SDLCs are realized by practitioners and academics through software development methodologies (e.g., Rational Unified Process (RUP) [28]), international software process standards (e.g., ISO/IEC 12207 [24]), and international software process frameworks (e.g., CMMI-DEV [9]).

The software development methodologies, and international software process standards and frameworks have provided valuable benefits to the software product, the software process, and stakeholders such as customers-users and the development team [13, 51, 70] such as: reduction of project costs, software products with higher quality, more precise project schedule estimations, greater user satisfaction, and in overall less wasting of relevant human, economic and technological organizational resources [13, 51, 70]. Consequently, utilizing these software development methodologies and international software process standards and frameworks is a common practice in large- and medium-sized

IEEE Access - Decision on Manuscript ID Access-2025-21516

IEEE Access <onbehalfof@manuscriptcentral.com>
para mí, sergio.galvan, jose.mora, lizbeth.munoz, olarik.s

jue, 29 may, 22:08 (hace 3 horas)

Traducir al español

30-May-2025

Dear Prof. Montoya-Murillo:

Your manuscript entitled "A comprehensive review of the main heavyweight and lightweight SDLC for Big Data Analytics Systems (BDAS)" has been accepted for publication in IEEE *Access*. The comments of the reviewers who evaluated your manuscript are included at the foot of this letter. We ask that you make minor changes to your manuscript based on those comments, before uploading final files.

Please be advised that you are **not permitted** to add or remove authors or references post-acceptance, regardless of the reviewers' request(s). **Any request to add an author post-acceptance will be denied.** However, we encourage you to check the formatting of your references to ensure that they are accurate in terms of bibliographic details as well as consistent with IEEE style. Additionally, please take this opportunity to improve the English grammar and check spelling, as the article is only lightly edited before publication.

You can submit your final files through the IEEE Author Portal.

All files intended for publication need to be submitted during this step, even if some files are unchanged from the initial submission. If you do not submit all files during this step, it can delay the publication of your article, or result in certain files not being published. **Please be advised that once you submit final files the article will be considered published and cannot be withdrawn.**

To assist you with preparing your final files, attached please find a Final Files Checklist.

Once you have completed the submission of your final files, the **accepted version** of your article will be posted Early Access on IEEE *Xplore* within 2-3 business days. Within 7-10 business days the corresponding author will receive your page proofs, at which point you can make **minor** edits as necessary. Once the corresponding author approves the proofs, the final version will replace the Early Access version on IEEE *Xplore*.

For more information on what to expect after you submit final files, please visit our Post Acceptance Guide.

IEEE Access
Regular Manuscript

A comprehensive review of the main heavyweight and lightweight SDLC for Big Data Analytics Systems (BDAS)

Submission Status	Accepted
Manuscript ID	Access-2025-21516
Accepted On	5 June 2025 by Editorial Office
Submitted On	5 June 2025 by David Montoya Murillo

Learn about what happens once your article has been accepted.

Submission overview →

IAJIT

Design and Usability Evaluation of LDSAM: an aligned ISO/IEC 29110 – Basic Profile – Development Methodology for Big Data Software Systems in Small Business

Journal Name :	International Arab Journal of Information Technology
Manuscript ID :	259-1743913931
Manuscript Type :	Research Article
Submission Date :	06-Apr-2025

Abstract : A System Development Life Cycle (SDLC) defines a process model of phases-activities, roles and products proposed to guide the systematic development of products of software on the expected time, budget, and functional quality. SDLCs are proposed through development methodologies (e.g. RUP) and international standards (e.g. ISO/IEC 12207). For Big Data Analytics Systems (BDAS) have been proposed some rigorous SDLCs such as CRISP-DM, TDSP, and DDSL, but its utilization for small business – or small teams called Very Small Entities (VSE's) – is scarcely reported given the process heaviness demanding human, technological and financial resources not available in small business. Given this problematic situation, new agile SDLCs for BDAS have been proposed such a Data-Driven Scrum, but some small organizations still require the utilization of a more systematic development process, and thus SDLCs based on ISO/IEC standards are expected. In this research, the design and the conceptual evaluation from a Panel of Experts of LDSAM (Light Data Science - Analytics Methodology) is reported. LDSAM is development methodology aligned to the ISO/IEC 29110 – Basic Profile – created especially for VSE's. The ISO/IEC 29110 standard provides a disciplined-systematic lightweight SDLC alternative to agile approaches for the business interested in ISO/IEC certifications rather than in agile ones. LDSAM was elaborated using a Design Science Research Methodology (DSRM). Initial usability evaluation results of LDSAM are satisfactory, but further empirical research is encouraged to advance to more mature and stable lightweight SDLCs for BDAS

Keywords : Big Data Analytics Systems (BDAS), System Development Life Cycle (SDLC); ISO/IEC 29110 standard for Very Small Entities (VSEs); LDSAM Light Data Science - Analytics Methodology; CRISP-DM, TDSP and DDSL.

Software engineering is not just about building systems; it is about understanding contexts.

Without data, you're just another person with an opinion.

AGRADECIMIENTOS

En primer lugar, me gustaría expresar mi más sincero agradecimiento a mi director de tesis, el Dr. José Manuel Mora Tavarez, de la Universidad Autónoma de Aguascalientes. Sus comentarios y orientación fueron siempre muy perspicaces, y me proporcionó un asesoramiento excepcional a lo largo de mi trayectoria investigadora. Quiero expresar mi más profundo agradecimiento al Dr. Mora por ponerme en contacto con distinguidos profesores de todo el mundo y por introducirme en una comunidad investigadora de primer nivel en mi campo. Gracias a su tutoría, adquirí una nueva perspectiva sobre la investigación académica y me animó a colaborar con expertos destacados en la disciplina. También me ofreció consejos fundamentales que mejoraron significativamente mis habilidades de investigación. Durante mis estudios de doctorado, el Dr. Mora estuvo a mi lado tanto en los momentos gratificantes como en los difíciles, ofreciéndome un apoyo inquebrantable tanto en el ámbito profesional como en el personal. Le estoy profundamente agradecido por su firme orientación y sabiduría durante cada fase de estos años formativos.

Asimismo, quiero expresar mi más sincero agradecimiento a mi comité tutorial por su ayuda y recomendaciones en todas las etapas del doctorado, me han llevado por un buen camino en la investigación y en el enfoque de mi tesis, doctores y doctora muchas gracias por su apoyo.

Del mismo modo, al Dr. José L. Roldán por su apoyo durante mi estancia de investigación en Sevilla, España, en 2024, ayudándome en la estancia e impartición del curso «Taller de Introducción Práctica al Análisis De Datos utilizando la herramienta Orange». Gracias por su confianza. Fue un honor trabajar con usted y con la Universidad de Sevilla (Sevilla tiene un color especial …).

Deseo expresar mi más sincero y afectuoso agradecimiento a los miembros del programa de doctorado "Doctorado en Ciencias Aplicadas y Tecnología" y a la institución que me permitió matricularme (Universidad Autónoma de Aguascalientes, UAA). Agradezco al Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCyT) por apoyar mis estudios durante los últimos 4 años y en el inicio de mi camino como investigador.

Finalizando, pero no menos importante con mi familia por todo el apoyo en estos tiempos de locura, hermano gracias sabes que siempre juntos, mi latosa muchas gracias por el apoyo final y aguante, y por su puesto a mi compañero de doctorado por otro proyecto trabajando en equipo compaye. Agradecido con la vida ...

CONTENT

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The development of Data Science and Analytics projects has become a strategic practice for organizations aiming to transform data into value. However, such practices are typically associated with large corporations that possess advanced technological infrastructure and specialized teams. Very Small Entities (VSEs), which represent over 95% of organizations in many countries, face significant challenges when attempting to adopt formal software development and data analytics methodologies, such as agile models or traditional standards-based approaches. This limitation not only prevents VSEs from leveraging advanced analytics, but also compromises the quality, sustainability, and success of their technological projects. To address this issue, this doctoral thesis proposes a lightweight development methodology called Light Data Science - Analytics Methodology (LDSAM), specifically designed for Big Data Science projects within the context of VSEs. LDSAM is aligned with the ISO/IEC 29110 – Basic Profile international standard, a framework specifically created for very small entities, which provides structured processes for project management and software implementation. The research follows the Design Science Research (DSR) paradigm and addresses four key research questions, ranging from the analysis of the state of the art in methodologies and platforms, to the design, development, documentation, and evaluation of the proposed artifact. LDSAM was conceptualized through a theoretical integration of software engineering standards (such as ISO/IEC 29110), data analytics methodologies such as CRISP-DM, TDSP, and DDSL, and lightweight approaches that balance the benefits of both agile and rigorous models, yet with lower complexity, cost, and learning curve. The methodology was systematically documented through an Electronic Process Guide (EPG), designed to facilitate adoption and practical application. LDSAM was validated in two stages: a conceptual evaluation by a panel of experts, and an empirical usability perception assessment by a pilot group of academics and practitioners in Data Science and software development. The results indicate that LDSAM is perceived as more useful, compatible, easy to use, and valuable compared to alternative unstructured methodologies, reinforcing its suitability for small-scale environments. The research demonstrates that it is feasible to design and validate lightweight methodologies that comply with international standards while meeting the specific needs and constraints of VSEs. Finally, further empirical studies are recommended to expand the use of LDSAM in diverse contexts and to strengthen the integration of standards in analytics projects, aiming to improve the quality and sustainability of software systems developed by small organizations.

RESUMEN

El desarrollo de proyectos de Ciencia de Datos y Analítica se ha posicionado como una práctica estratégica para organizaciones que buscan transformar sus datos en valor. No obstante, estas prácticas suelen estar asociadas a grandes corporativos que disponen de infraestructura tecnológica avanzada y equipos de trabajo especializados. Las pequeñas y muy pequeñas empresas (VSEs), que representan más del 95% de las organizaciones en muchos países, enfrentan barreras importantes para adoptar metodologías formales de desarrollo de software y analítica de datos, como los modelos ágiles o los enfoques rigurosos basados en estándares tradicionales. Esta limitación no solo excluye a las VSEs del aprovechamiento de la analítica avanzada, sino que además compromete la calidad, sostenibilidad y éxito de sus proyectos tecnológicos. Ante esta problemática, la presente tesis doctoral propone una metodología de desarrollo ligera, denominada Light Data Science - Analytics Methodology (LDSAM), especialmente diseñada para proyectos de Ciencia de Datos de tipo Big Data dentro del contexto de VSEs. LDSAM está alineada con el estándar internacional ISO/IEC 29110 – Perfil Básico, un estándar específicamente desarrollado para entidades muy pequeñas, que proporciona un marco estructurado de procesos para la gestión de proyectos e implementación de software. La investigación se enmarca en el paradigma de Design Science Research (DSR) y responde a cuatro preguntas de investigación, que abarcan desde el análisis del estado del arte de las metodologías y plataformas existentes, hasta el diseño, desarrollo, documentación y evaluación del artefacto propuesto. LDSAM fue conceptualizada a partir de una integración teórica entre estándares de ingeniería de software (como ISO/IEC 29110), metodologías de analítica como CRISP-DM, TDSP y DDSL, y enfoques ligeros que equilibran los beneficios de los modelos ágiles y rigurosos, pero con menor complejidad, costo y curva de aprendizaje. El diseño de la metodología fue sistematizado mediante una guía electrónica de procesos (EPG), la cual facilita su adopción y uso en entornos reales. La validación de LDSAM se llevó a cabo en dos fases: una conceptual, mediante un panel de expertos, y otra empírica, mediante una evaluación de percepción de usabilidad por parte de un grupo piloto de académicos y profesionales en Ciencia de Datos y desarrollo de software. Los resultados obtenidos indican que LDSAM es percibida como más útil, compatible, fácil de usar y valiosa en comparación con metodologías alternativas no estructuradas, lo cual refuerza su pertinencia para entornos de pequeña escala. La investigación demuestra que es viable diseñar y validar metodologías ligeras que cumplan con estándares internacionales, y que al mismo tiempo respondan a las condiciones y capacidades de las VSEs. Finalmente, se recomienda impulsar nuevas investigaciones que amplíen el uso de LDSAM a otros contextos y que fortalezcan la incorporación de estándares en proyectos de analítica, con el fin de mejorar la calidad y sostenibilidad de los sistemas desarrollados por organizaciones pequeñas.

CONTRIBUTIONS

Research Stay at the University of Seville, in Spain in April 2024, where additionally the course “WORKSHOP ON PRACTICAL INTRODUCTION TO DATA ANALYSIS USING THE ORANGE TOOL” was given.

A **chapter** entitled “A Selective Conceptual Review of CRISP-DM and DDSL Development Methodologies for Big Data Analytics Systems” in a Springer book entitled “Development Methodologies for Big Data Analytics Systems: Plan-driven, Agile, Hybrid, Lightweight Approaches.” 2023.

An **article** “A review of SDLCs for Big Data analytics systems in the context of very small entities using the ISO/IEC 29110 Standard-Basic profile” published in the International Arab Journal of Information Technology (IAJIT) 2025.

An **article** “A comprehensive review of the main heavyweight and lightweight SDLC for Big Data Analytics Systems (BDAS)” IEEE Access 2025.

An **article** “Design and Usability Evaluation of LDSAM: an aligned ISO/IEC 29110 – Basic Profile – Development Methodology for Big Data Software Systems in Small Business”

Ponencia en el 1° Congreso Internacional Multidisciplinario de la Divulgación Científica, UAA, 2023: Light Data Science - Analytics Methodology (LDSAM): an aligned ISO/IEC 29110 - Basic Profile - Development Methodology for Big Data Software Systems in Small Business.

Ponencia en el 2° Congreso Internacional Multidisciplinario de la Divulgación Científica, UAA, 2024: Metodología de análisis ligera para ciencia de datos ligera (LDSAM): Alineada a ISO/IEC 29110 – Perfil básico – Metodología de desarrollo para sistemas de software de Big Data en pequeñas empresas.

EPG of the ISO/IEC 29110 -Basic Profile- for BDAS + : https://davidmontoyam-uaa-dcat.on.drv.tw/iso29110.basic.BDAS.EPG

CHAPTER 1. INTRODUCTION

1.1 CONTEXT OF THE RESEARCH PROBLEM

Nowadays, many organizations in Mexico, and in the world, are in the process of a digital transformation, that requires the development of useful, secure, and valuable software applications. Additionally, these software applications are expected to be available in short periods and generate quality services that respond to the needs of the organization and its customers (Kose, 2021).

A particular category of software applications is the Data Science – Data Analytics projects. Data Science or Data Analytics is a recent discipline that combines Statistics, Artificial Intelligence, and Computer Science to explore, predict, or prescribe decisional situations. However, only large business organizations are the usual customers and end-users of Data Science – Data Analytics projects, and they are focused on costly Big Data platforms (Tsai et al., 2015). Thus, small and medium-sized business organizations lose the benefits of using Data Science – Data Analytics projects.

Fortunately, Data Science – Data Analytics approaches can be also used for Small Data projects (Estrin, 2014; Hekler et al., 2019; Wilson & Daughterty, 2020).

However, both Small and Big Data Science – Data Analytics projects are difficult projects to be successful (Larson & Chang, 2016). Several international reports indicate that a large percentage of Data Science / Data Analytics projects failed in being released with the budget, or schedule or functionality as it was planned (Luellen, 2018).

The use of agile methodologies in Data Science – also known as Data Analytics- has been proposed to cope with the problem of failed Data Science – Data Analytics projects (Larson & Chang, 2016). However, agile methodologies have been also critiqued and identified that for more stable software applications, a more disciplined development approach must be used (Boehm, 2002).

In this research, thus, we believe that a lightweight – neither agile-oriented nor rigor- oriented development approach applied to the development of Small Data Science – Data Analytics- projects can generate benefits in the utility, safety, and quality of the project, maintaining the established timeline and the planned budgets for the project.

1.2 MOTIVATION AND RELEVANCE OF THE RESEARCH PROBLEM

Several studies report the benefits of using lightweight development methodologies (Boehm, 2002; Boehm & Turner, 2003) avoiding the limitations of agile and rigorous development approaches and trying to take advantage of their benefits.

For Small Business organizations, a new software engineering standard, the ISO/IEC 29110 has been developed as a lightweight methodology (O'Connor & Laporte, 2017). Initial studies have reported multiple benefits on the quality of the products, and the achievement of the budget and schedule as it was planned (Laporte et al., 2017; Muñoz et al., 2020).

Similarly, the market for platforms focused on Data Sciences – Data Analytics is predicted to strong growth in the decade of 2020-2030 (Markets and Markets, 2020; Forrester Inc., 2019).

The excitement around Data Science – Data Analytics and its successes have increased the interest of multiple business organizations, but these positive signs can be misleading. Not only is Data Science / Data Analytics in its infancy as a science and a discipline, but its current practice also has a large learning curve related largely to the issues raised above. Gartner, Forrester, and other technology analysts report that most (80%) early (2010-2012) Data Science / Data Analytics projects in most US enterprises failed. In late 2016, Gartner reported that while most enterprises declare Data Science / Data Analytics as core expertise, only 15% claim to have deployed Data Science / Data Analytics projects in their organization (Gartner, 2014). Analysts predict 80+% failure rate through 2017 (Demirkan & Dal, 2014; Veeramachaneni, 2016; Lohr & Singer, 2016).

1.3 FORMULATION OF THE RESEARCH PROBLEM

1.3.1 RESEARCH PROBLEM

Consequently, based on the previous research context described, we can identify the research problem directly as the ***lack of development methodologies for Big Data Science Analytics Projects that be considered by the software developers as lightweight (neither agile nor rigorous), ease of use, useful, compatible, and valuable.***

1.3.2 RESEARCH QUESTIONS AND HYPOTHESES

- **RQ.1** What is the state of the art – contributions and limitations- on Lightweight Development Methodologies for Big Data Science - Analytics Software Systems?
- **H0.1** There is no need for a Lightweight Development Methodology for Big Data Science - Analytics Software Systems.

- **RQ.2** What is the state of the art – capabilities and limitations – of open-source development platforms for Big Data Science - Analytics Software Systems?
- **H0.2** There are no available open-source development platforms for Big Data Science - Analytics Software Systems that can be satisfactorily evaluated in the technical, end-user, and organizational dimensions.

- **RQ.3** What elements of Lightweight Development and Big Data Science - Analytics Development Methodologies can be used to elaborate a Lightweight Development Methodology for Big Data Science - Analytics Software Systems that can be evaluated theoretically valid from a Panel of Experts?
- **H0.3** There are no elements of Lightweight Development and Big Data Science - Analytics Development Methodologies that can be used to elaborate a Lightweight Development Methodology for Big Data Science - Analytics Software Systems that can be evaluated theoretically valid from a Panel of Experts.

- **RQ.4** Can the new elaborated Lightweight Development Methodology for Big Data Science - Analytics Software Systems be documented in an Electronic Process Guide (EPG), and be evaluated as agile, useful, ease of use, compatible and valuable from a pilot group of Big Data Science - Analytics academics and practitioners?
- **H0.4.1** The new elaborated Lightweight Development Methodology for Big Data Science - Analytics Software Systems cannot be documented in an Electronic Process Guide (EPG).
- **H0.4.2** The new elaborated Lightweight Development Methodology for Big Data Science - Analytics Software Systems is not considered agile, useful, ease of use, compatible and valuable from a pilot group of Big Data Science – Analytics academics and practitioners.

1.3.3 GENERAL AND SPECIFIC RESEARCH OBJECTIVES

To design conceptually a Development Methodology for Big Data Science – Analytics Software Systems, and document it in an Electronic Process Guide, that be evaluated as lightweight, useful, ease of use, compatible and valuable for a pilot group of Data Science – Analytics academics and practitioners.

1.3.4 CONTRIBUTIONS AND DELIVERABLES OF THE RESEARCH

In this research proposal, it is expected to produce the following products:

1. For the Software Engineering Theory

- 1 research paper for an indexed journal with the theoretical analysis on "The State of the Art on Open-Source Big Data Science - Analytics Development Platforms"
- 1 research paper for an indexed journal with the theoretical analysis on "The State of the Art on Development Methodologies for Big Data Science - Analytics Projects"
- 1 submitted research paper for an indexed journal with the theoretical analysis and empirical evaluation of the Light DS Methodology – a lightweight Methodology for Big Data Science - Analytics Projects

2. For the Software Engineering Practice

- 1 new Light DS Methodology – a lightweight Methodology for Big Data Science - Analytics Projects, available in a web-based free-cost access EPG (Electronic Process Guideline)
- 1 new PhD graduate in Software Engineering area

1.4 GENERAL DESCRIPTION OF THE RESEARCH METHODOLOGY

In this research, it is proposed to use a Design Science Research approach (Vom Brocke et al., 2020; Peffers et al., 2007). "*Design Science Research (DSR) is a problem-solving paradigm that seeks to enhance technology and science knowledge bases via the creation of innovative artifacts that solve problems and improve the environment in which they are instantiated. The results of DSR include both the newly designed artifacts – represented by constructs, and/or models, and/or methods, and/or instantiations -, and design knowledge (DK)*".

1.4.1 OVERVIEW OF THE RESEARCH METHODOLOGY

The specific DSR methodology to conduct is the Design Science Research Methodology proposed by Peffers et al. (2007). It has six activities as follows:

- **Activity 1: Problem identification and motivation.** "Define the specific research problem and justify the value of a solution. Justifying the value of a solution accomplishes two things: it motivates the researcher and the audience of the research to pursue the solution and to accept the results and it helps to understand the reasoning associated with the researcher's understanding of the problem".
- **Activity 2.1: Define the objectives for a solution.** "Infer the objectives of a solution from the problem definition and knowledge of what is possible and feasible. The objectives can be quantitative, such as terms in which a desirable solution would be better than current ones, or qualitative, such as a description of how a new artifact is expected to support solutions to problems not hitherto addressed".
- **Activity 2.2: Review the State of the Art**. Review the state of the art on the main element to be designed and identify the main contributions and limitations.
- **Activity 3: Design and development**. Create the artifact. Such artifacts are potentially constructs, models, methods, or instantiations (each defined broadly). Conceptually, a design research artifact can be any designed object in which a research contribution is embedded in the design. This activity includes determining the artifact's desired functionality and its architecture and then creating the actual artifact".
- **Activity 4: Demonstration**. "Demonstrate the use of the artifact to solve one or more instances of the problem. This could involve its use in experimentation, simulation, case study, proof, or other appropriate activity".
- **Activity 5: Evaluation.** "Observe and measure how well the artifact supports a solution to the problem. This activity involves comparing the objectives of a solution to actual observed results from use of the artifact in the demonstration. At the end of this activity the researchers can decide whether to iterate back to activity 3 to try to improve the effectiveness of the artifact or to continue to communication and leave further improvement to subsequent projects". The specific Evaluation methods to be used will be: 1) Evaluation Conceptual from a Panel of Experts; 2) Evaluation from a Proof of Concept; and 3) Empirical survey-based evaluation from a pilot sample of Software Engineering professionals.
- **Activity 6: Communication.** "Communicate the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences such as practicing professionals, when appropriate".

1.4.2 TIMELINE – SEMESTERS, ACTIVITIES AND DELIVERABLES

Table 1.1 Timeline-semesters, activities and deliverable.

Phases	2021	2022	2023	2024
Activities 1. a) Background and history of the problem. b) Problematic situation. c) Type and purpose of research. d) Relevance. e) Objectives, questions, and hypotheses/research propositions.	X			
Activity 2. Review the State of the Art a) Theories bases. b) Studies related. c) Contributions and limitations of related studies.	X	X		
Activity 3. Design and Development of Artifact a) Application or creative-deductive relational conceptual design model.		X	X	
Activities 4. Demonstration and Evaluation a) Validation of content by a panel of experts. b) Validation by logical argument. c) Validation for proof of concept of the artifact.			X	X
Activitiy 5 – Communication a) Write and submit research paper 1. b) Write and submit research paper 2. c) Write and submit research paper 3.		X	X	X

CHAPTER 2. RESEARCH METHODOLOGY

2.1 MAIN ACTIVITIES

To start with the introduction to this subject, we must begin by commenting that the scientific research process can be carried out through different methods (Ackoff, 1962). For this thesis, we will use a combination of three different methods, seeking to generate a more comprehensive and optimal result at the conclusion, in addition to different alternatives of approach, development and evaluation. These three methods of investigation that we will use in this thesis will be combined to obtain a set of points of view, to generate a methodology according to the requirements that we are looking for, to generate this methodology with a higher performance than would be achieved by following only one way.

Of these three methods, the first research method on which we will rely is the conceptual research method of Dr. Mora (Mora et al., 2009), the second research method will be the DSRM method (Peffers et al., 2007) and finally we will use the 3-cycle DSR research method (Hevner, 2007).

Table 2.1 Conceptual- based Design Research Phases (Mora et al., 2012).

Conceptual-based Design Research Phases
Phase I. Formulation of Research Problem • Background and history of the problem. • Problematic situation. • Type and purpose of research. • Relevance. • Objectives, questions and hypotheses / research propositions.
Phase II. Analysis of Related Work. • Theories bases. • Studies related. • Contributions and limitations of related studies. • Selection/design of general conceptual framework.
Phase III. Conceptual Design of Artifact. • Application of creative-deductive relational conceptual design model.
Phase IV. Validation of Designed Artifact. • Validation of Content by a Panel of Experts. • Validation by Logical Argumentation. • Validation by Proof of Concept of Designed Artifact. • Empirical Validation by a Pilot Survey or Case Study or Experimental Study.

Conceptual Research can be considered as the main source for the generation of new theories, models or conceptual schemes that -in order to complete the scientific cycle- must later be empirically or deductively tested using other research methods (Blalock, 1969). Conceptually based research is used when the designed objects will ultimately be evaluated through conceptual research methods, and there are generally no physical laws to apply, neither models nor mathematical methods. The conceptual method is considered as the main source of generation of new theories, models or conceptual schemes, in the field of information systems, this is considered as an important part of the possible repertoire of research methods, and is composed of 4 phases these phases are the following: Phase I Formulation of the Research Problem; Phase II Analysis of Related Works; Phase III Application or Design of the Conceptual Model; and Phase IV Validation of the Conceptual Model applied or designed (Mora et al., 2009).

The second method is DSRM, which we will merge with the previously mentioned conceptual research method, which will allow us to better document the development of the methodology in this thesis. The goal of a DSRM process is to improve the production, presentation, and evaluation of research.

The Defining Principles of The Design Science Research Methodology (DSRM), We now have a reasonably solid idea of what it is. "Design science...creates and evaluates IT artifacts intended to solve identified organizational problems" (Hevner, 2007. p. 77). It involves a rigorous process for designing artifacts to solve observed problems, for making research contributions, for evaluating the designs, and for communicating the results to appropriate audiences. These artifacts may include constructs, models, methods, and instances. They may also include social innovations or new properties of technical, social, or informational resources; in short, this definition includes any object designed with an embodied solution to an understood research problem. (Hevner, 2007).

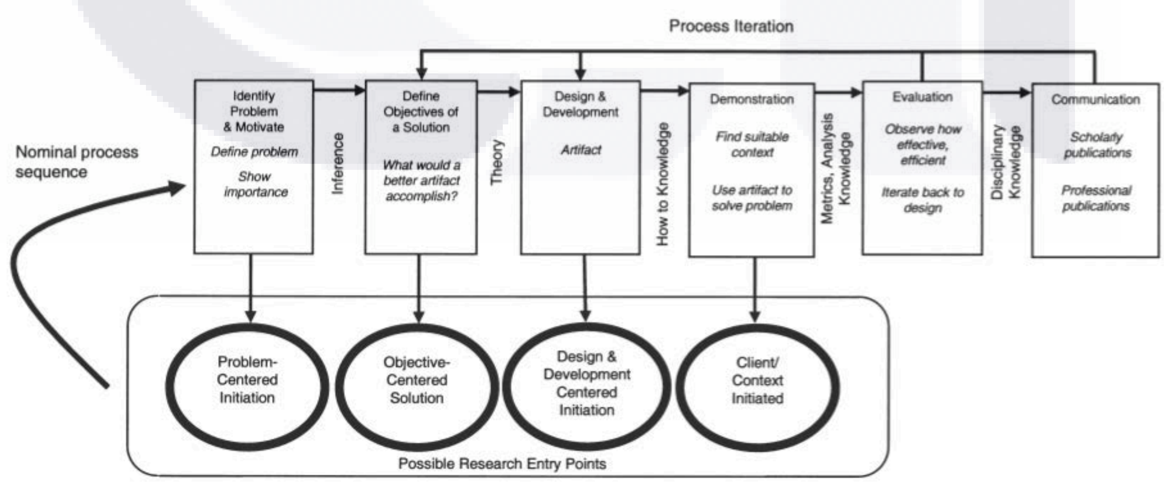

Figure 2.1 Design Science Research Methodology (DSRM) Process Model (Peffers et al., 2007).

Figure 2.1 Design Science Research Methodology (DSRM) Process Model shows the 6 activities that make up the DSRM as a nominal sequence, the figure also shows the description of what to do in general terms each of the activities. This methodology is used to generate artifacts in the information systems that solve an instance of a problem.

In the context of this research, it is essential to rigorously and exhaustively examine the existing literature on the topic in question. To achieve this, a selective literature review methodology has been adopted. This methodology will allow us to perform a critical and thorough evaluation of a specific selection of relevant documents, with the objective of extracting relevant and meaningful information. By employing selective literature review, we will be able to gain a deeper and more detailed understanding of the state of existing knowledge, identify patterns, trends and gaps in previous research, and lay a solid theoretical foundation for our study. In this regard, this chapter presents the methodology and process used to conduct the selective literature review, as well as the results and implications derived from this approach in our research.

Selective document selection refers to the process of intentionally choosing and collecting specific documents from a broader source or set of information. Rather than collecting all available documents, a careful selection is made to identify the most relevant, significant or useful documents for a particular purpose or topic. Document culling is common in a variety of areas, such as academic research, knowledge management and legal evidence gathering. This approach optimizes resources and efficiency by focusing on the documents that provide the most value. When making a selective selection of documentation, different criteria can be applied, such as relevance, quality, timeliness and reliability of the documents. The thematic focus and the objectives of the project or research are also considered. It is important to note that selective selection of documentation involves making informed decisions about which documents to include and which to omit. This requires critical analysis and careful evaluation of the available information to ensure that the documents selected are the most relevant and appropriate for the intended purpose.

This process focused on the Peffers method will be performed in step 2, for more details of this you can review the tables added in the Appendix A where you can see all the documents found or analyzed for such selective search.

The third and last is design science research (DSR), which aims to improve our understanding of information systems through the creation of technological artifacts; the artifacts created embody the solution to a problem (Hevner et al., 2004).

This process is depicted in Figure 2.2 Design Science Research Cycles and shows us the function of each of the cycles represented in the two main research approaches proposed by Hevner in 2004. The relevance cycle links the contextual environment to design science activities. The rigor cycle connects design science activities to the scientific knowledge base. The design cycle iterates between the core activities of artifact construction and evaluation and the processes of design research (Hevner, 2007).

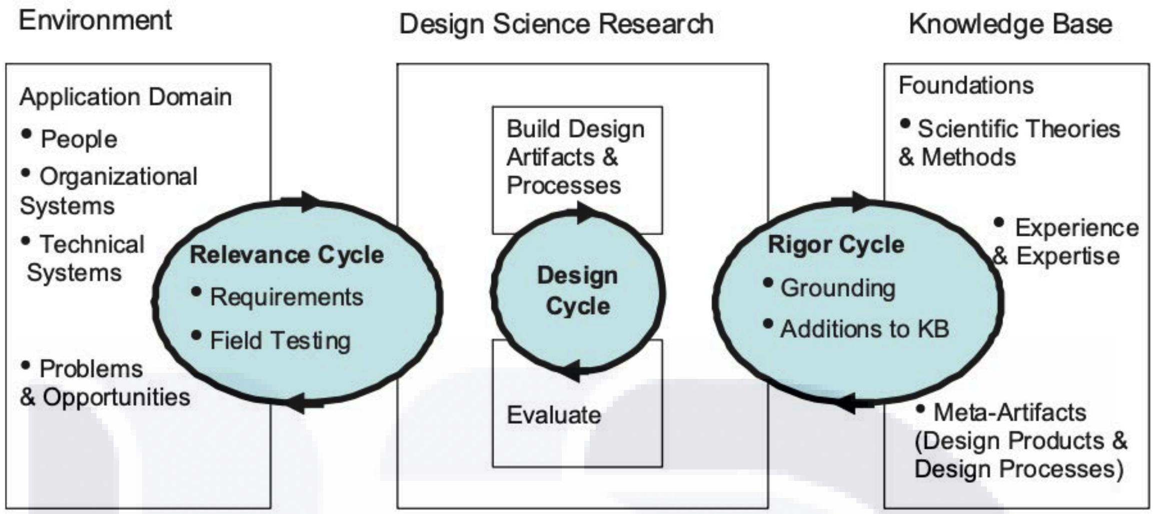

Figure 2.2 Design Science Research Cycles (Hevner, 2007).

This Ph.D. research uses the Design Science Research Methodology (DSRM) (Peffers et al., 2007) complemented with additional specific research methods. These methods are: Selective Systematic Literature Review method (Cooper 1988), Conceptual Design (Mora et al. 2009), Heuristic Design with Means-Ends Analysis (Newell and Simon 1972; Mora et al. 2023), Conceptual Verification by Panel of Experts (Hevner et al. 2004; Beecham et al. 2005), Empirical Validation with Statistical Analysis (Wohlin et al. 2012; and Chin 2009), and Guide for Scientific Reports in Software Engineering (Shaw 2003). Table 1 summarizes steps, purpose, complementary research methods, and expected outcomes.

For Activities 1 and 2.1 the following actions will be implemented:

- Background and history of the problem.
- Problematic situation.
- Type and purpose of research.
- Relevance.
- Objectives, questions, and hypotheses/research propositions.

For Activity 2.2 the following actions will be implemented:

- Theories bases.
- Studies related.
- Contributions and limitations of related studies.

For Activity 3 the following actions will be implemented:

- Application or creative-deductive relational conceptual design model.

For Activities 4 and 5 the following actions will be implemented:

- Validation of content by a panel of experts.
- Validation by logical argument.

- Validation for proof of concept of the artifact.

For Activities 6 the following actions will be implemented:

- Write and submit research paper 1.
- Write and submit research paper 2.
- Write and submit research paper 3.

For more details about the Activities please check Table 1-1.

Table 2.2 Design Science Research Methodology (DSRM) with complementary research methods

DSRM Steps	Purpose	Complementary research methods	Outcomes
Step 1) Design problem identification and motivation.	To state the expected overall research goal that delimits the scope of the research, the research questions that focus on the knowledge gaps of interest, and the motivations to pursue the research design. (For these aims is required to conduct a Review of the State of the Art on the specific problem.).	• Conceptual Literature Review (CLR), or • Systematic Literature Review (SLR), or • Selective Systematic Literature Review (SSLR).	• Research overall goal statement. • Research questions. • Research motivation statements. • Review of the State of the Art.
Step 2) Definition of the design objectives and restrictions for the expected artifact.	To define the specific design objectives (i.e. expected qualities in the designed artifact), design restrictions (i.e. the limitations on time, cost and resources utilized to design the artifact), design approach (i.e. analytics, axiomatic or heuristic), design theoretical sources (i.e. the design materials), and design components (i.e. the specific design building-blocks).	• Conceptual Design.	• Design problem identification and motivation. • Definition of the Design Objectives, Design Restrictions, Resign Approach, Design Theoretical Sources, and Design Components for the expected Artifact.
Step 3) Design and development of the artifact.	To design and implement the expected artifact guided-controlled by the design objectives and restrictions, and using the agreed design approach, design theoretical sources and design components.	• Heuristic Design.	• Conceptual designed artifact. • Implemented designed artifact.
Step 4) Demonstration of the artifact (Proof of Concept).	To demonstrate the designed and implemented artifact and conduct initial verification.	• Verification by a Panel of Experts	• Conceptual Verification by a Panel of Experts.
Step 5) Evaluation of the artifact.	To conduct empirical evaluation of the designed and implemented artifact.	• Survey or Experimental Methods.	• Empirical Validation with Statistical Analysis.
Step 6) Communication of research results.	To generate a structured scientific report (i.e. Thesis, Technical Report, Chapter, Conference Proceeding document, or Journal article) of results and communicate them in academic outlets.	• Scientific writing guidelines.	• Structured Scientific Report.

2.2 OBJECT AND SUBJECTS OF STUDY

The development of this thesis is based on current balanced lightweight development methods such as ISO/IEC 29110 Basic Profile standard, Analytics/Data Science development methods, and balanced agile-disciplined Analytics/Data Science development methods.

The validation of the developed methodology will be evaluated by a pilot sample of software professionals and academics interested in agile Analytics / Data Science development methods, through a usability perceptions measurement instrument (ease of use, useful, compatible and valuable) used. commonly found in the scientific literature (Moore & Benbasat, 1991).

2.3 MATERIALS AND EQUIPMENTS

- Research articles, chapters and conference presentations related to Agile development methods and Data Science.
- Official documents and literature related to balanced lightweight development methods, Data Science, Analytics, software engineering and Big Data sciences.
- Laptop computer equipment.
- VM server in LabDC-2004 laboratory.
- Open-Source development environments / platforms for the development of Analytics or Data Sciences projects (R + Python for R + Weka for R + Shiny + Radiant and web libraries such as Weka + Shiny + Radiant).

2.4 RESEARCH EVALUATION METHODS

According to Hevner et al. (2004) the validation techniques are the following:

- Observational: Case Study: Study artifact in depth in business environment. OR Field Study: Monitor use of artifact in multiple projects.
- Analytics: Through statistical analysis or dynamic analysis or optimization.
- Experimental: Through controlled experiment or simulation
- Testing: Through functional testing or structural testing
- Descriptive: Through information, arguments of demonstration cases.

Peffers (2007) mentions in the DSRM methodology that, by applying the artifact generated in the specific case, results will be generated that can be evaluated with relevant metrics to be able to be compared with the objectives that were defined from the beginning. He further mentions that, if the evaluation is conclusive, that is, it generates relevant conclusions about the artifact, the next step is to communicate the artifact to the relevant entities. Otherwise, if the artifact is not conclusive, the objectives or the elaboration of the artifact will have to be rethought to obtain conclusive results (Peffers et al., 2007).

2.5 RESTRICTIONS AND LIMITATIONS

Due to the complexity and little use of methodologies for the development of Analytics and Data Sciences projects, in an environment of small and medium-sized companies in Mexico using Small Data Sciences, designing and developing a methodology for this specific sector is largely a complex work. Therefore, this thesis will have the following limitations:

- The period available for the development of the methodology is 3 to 4 years.
- Development costs, only the budget for the doctoral study is available.
- The scope of the projects, this methodology is developed for micro and small projects with participants of 5 to 10 people, with periods of time of 3 to 6 months and with limited budgets.

CHAPTER 3. THEORETICAL BACKGROUND

3.1 THEORETICAL FUNDATIONS.

3.1.1 On Software Engineering.

Software Engineering has gained importance as an academic discipline from several decades, and this happened since software systems could no longer be generated by a single person and in an artisanal way. The needs to develop new concepts, methods and techniques for the control and development of the software systems fostered the progress of Software Engineering.

According to the ISO/IEC/IEEE Systems and Software Engineering Vocabulary (ISO/IEC/IEEE, 2017), the Software Engineering discipline can be defined as "*the application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software; that is, the application of engineering to software)*."

Another relevant document on Software Engineering is the Guide to the Software Engineering Body of Knowledge (SWEBOK) (Bourque et al., 2014). This document is useful for acquiring the knowledge that helps to the lifelong career development as a software engineering professional. This document SWEBOK is published by the IEEE Computer Society (Bourque et al., 2014). In this SWEBOK document, it is reported that one of the most important research and practical areas in the Software Engineering discipline is the Software Engineering Process (SEP) area. "*SEP consists of a set of interrelated activities that transform one or more inputs into outputs while consuming resources to accomplish the transformation*" (Bourque et al., 2014). Based on the above, the point that we will highlight in this PhD dissertation is that an engineering process consists of a set of interrelated activities that transform one or more inputs and outputs while consuming resources to achieve the transformation.

Oktaba & Ibargüengoitia (1998) sought to define a structured and systematized model context as a fundamental concept for software processes where they defined the minimum set of phases, activities-tasks, roles, resources, and artifacts that cover the entire process. The basic *phases* for the Software Process can be analysis, design, code, installation, and testing. The *activities* (or tasks), are also key pieces of the process, defined as specific procedures to be performed at some point in the development. These activities usually require input *artifacts* and *resources* (technical and/or human), which can be associated through *roles* to generate another output *artifact*. These activities are associated with the production, control, technology or communication, where these can be subdivided, for example as production subtasks of analysis, design, coding, documentation, prototyping, among others

SWEBOK indicates that software processes must be specified for several reasons: to facilitate human understanding, communication, and coordination; to help manage software projects, measure and improve the quality of software products efficiently, to support process improvement, and to provide a basis for supporting automated process execution. In the same way, it must be understood and contemplated that a software process can involve work assignments for individuals and teams, can include roles and competencies, and even include sub-processes.

Within the SWEBOK.v3 book, there are 15 basic chapters, and the chapter 8 that refers to Software Engineering Process is of high importance of understanding because this area of knowledge is required for the generation of our software process development methodology.

Software Engineering Process section – as it is shown in the Fig. 3.2 - is segmented into several subsections: 1) Software Process Definition, 2) Software Life Cycles, 3) Software Process Assessment and Improvement, 4) Software Measurement, and 5) Software Engineering Process Tools. In this PhD dissertation, the subsections of 1) Software Process Definition and 2) Software Life Cycles are of key importance to elaborate a software development process – or methodology – to Data Science projects.

A software product life cycle (SPLC) includes a software development life cycle plus additional software processes that provide implementation, maintenance, support, evolution, retirement, and all other processes from inception to completion of a software product, including software configuration management and software quality assurance processes which must be applied throughout the life cycle of a software product. The life cycle of a software product can include multiple software development life cycles to evolve and improve the product.

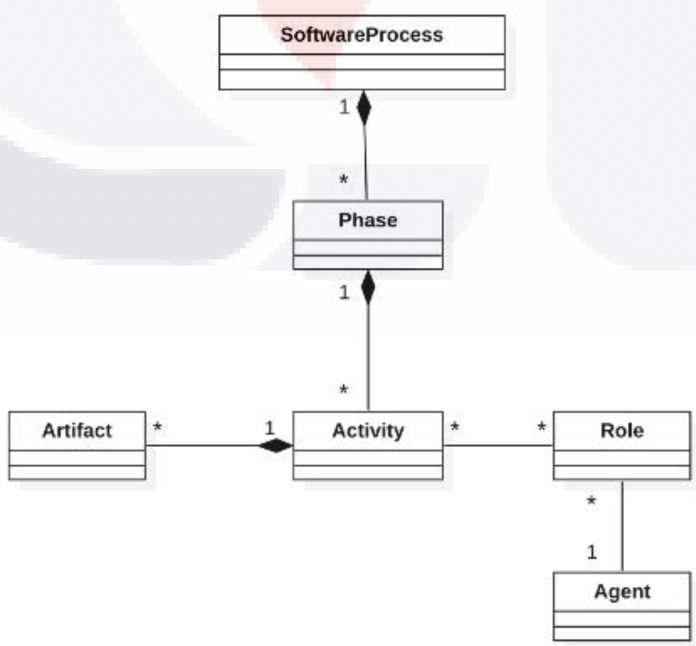

Figure 3.1 Class diagram software process (Oktaba & Ibargüengoita, 1998).

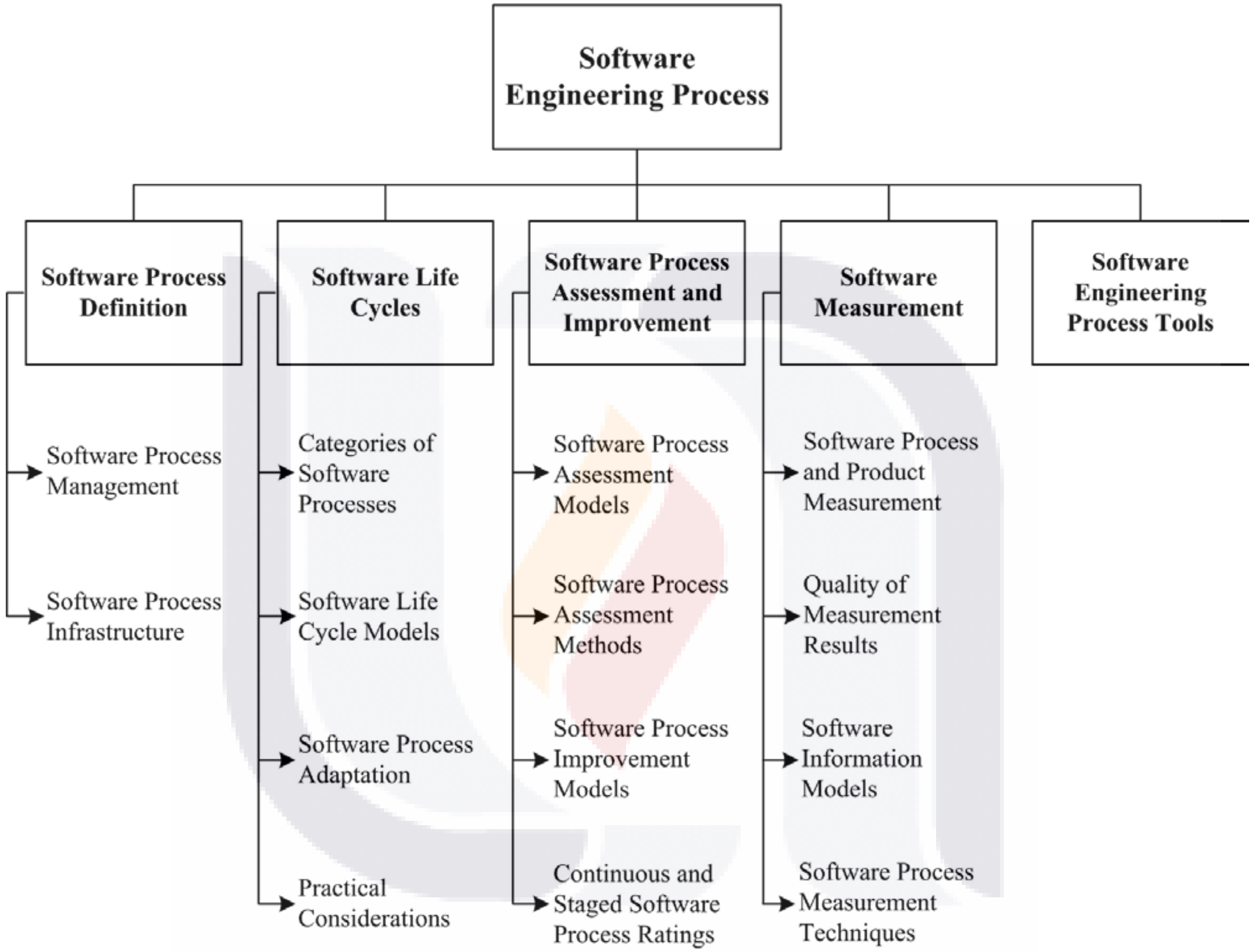

Figure 3.2 Breakdown of Topics for the Software Engineering Process KA (Bourque et al., 2014, pp. 149).

Figure 3.2 (Class diagram software process) below shows a UML diagram of the relationship between a Software Engineering process, phases, activities, artifacts, roles, and agents.

SWEBOK classifies the Software Processes in four categories: 1) the primary or fundamental processes – called engineering processes in CMMI- include software processes for the development, operation, and maintenance of software; 2) support processes are applied intermittently or continuously throughout the life cycle of a software product where they seek to support the primary processes; they include software processes such as configuration management, quality assurance, and verification and validation; 3) organizational processes – called process management in CMMI- support other software engineering processes; they include training, process measurement analysis, infrastructure management, portfolio.

management, and reuse, organizational process improvement, and software life cycle model management, and 4) cross-project processes – called project management processes - such as reuse, software product line, and domain engineering; involve more than a single software project in an organization. Project management processes include processes for planning and estimating, resource management, measurement and control, leadership, risk management, stakeholder management, and coordination of primary, support, organizational, and inter-project processes of project projects.

Software life cycle models, according to the SWEBOKv3 can be Incremental, Iterative, and Agile models. A hallmark of the various software development life cycle models is how software requirements are managed. Linear development models typically develop a complete set of software requirements, to the extent possible, during project initiation and planning. Adaptation of software processes may consist of using an alternative set of activities that achieve the purpose and results of the software process.

Hence, a relevant attribute of a software process and product or artifact is its quality. Achieving a quality software process and product, as well as to fit the planned budget and schedules are goals important in all software process (Parnas, 2010).

Many software development life cycle processes (SDLC) have been generated with the passage of time and the evolution of technology, (Rodríguez et al., 2009). In Figure 3.3 show comparison and classification of different SDLC to the variables of "specification rigor" and "agility". In this doctoral thesis, we will seek to support this trend using as a basis the use of a standard to complement the benefits that this generates.

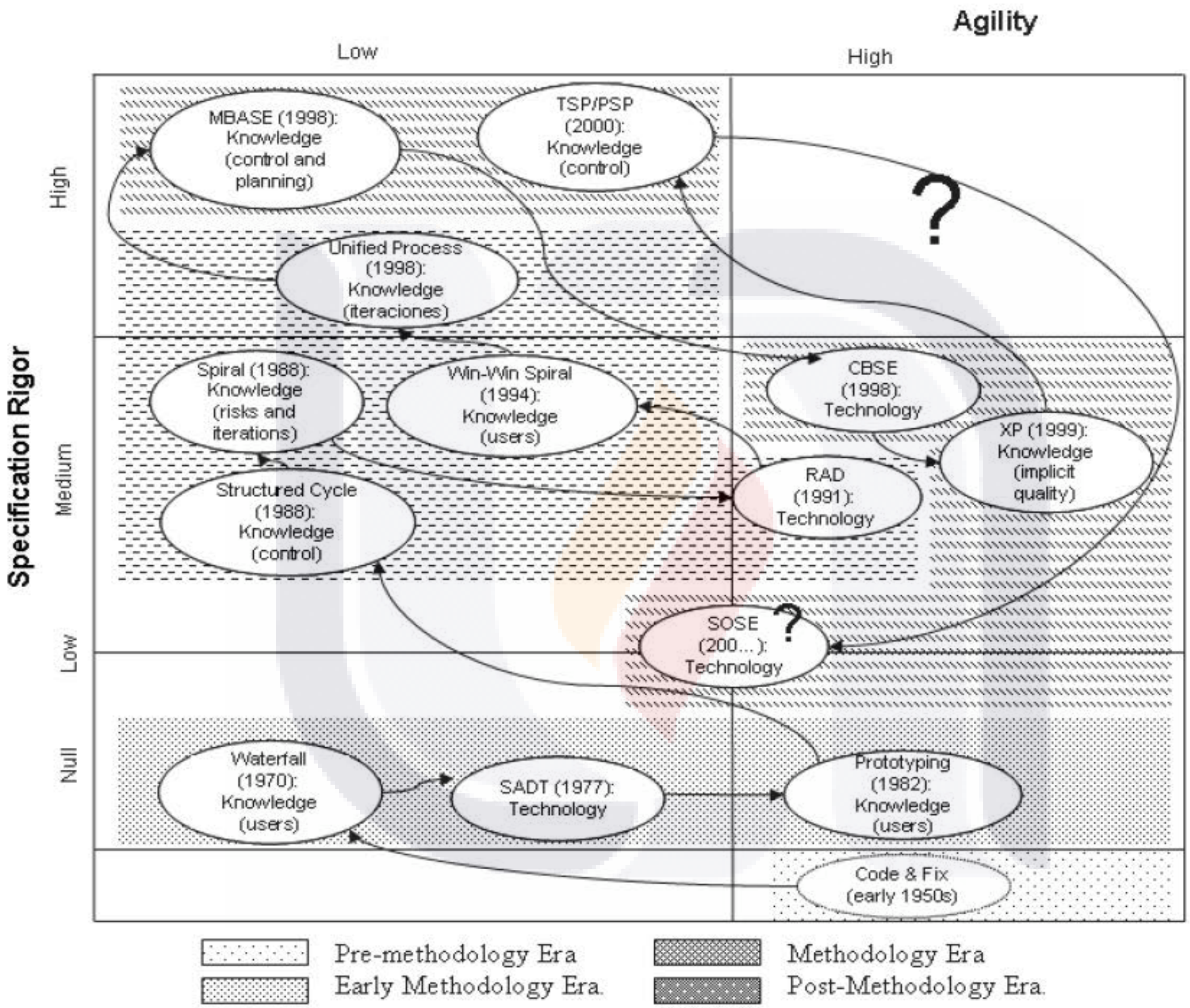

Figure 3.3 Map of PM-SDLC´S evolution (Rodríguez et al., 2009).

Hence, the relevant concepts that emerge from this section are the following ones: Software Engineering, Software Process, Software Life Cycle, and Software. Table 3.1 reports them with the used definitions in this PhD dissertation.

Table 3.1 Main constructs derived from Software Engineering.

Construct	Meaning	Reference
Software Engineering	Application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software; that is, the application of engineering to software	(ISO/IEC/IEEE 2465:2017, 2017).
Software Development Process	Process by which user needs are translated into a software product. The process involves translating user needs into software requirements, transforming the software requirements into design, implementing the design in code, testing the code, and sometimes, installing and checking out the software for operational use. These activities can overlap or be performed iteratively.	(ISO/IEC/IEEE 2465:2017, 2017).
Software Engineering Process	It is a set of interrelated activities that transform one or more inputs into outputs while consuming resources to achieve the transformation	(Bourque et al., 2014).
Software Life Cycles	Project-specific sequence of activities that is created by mapping the activities of a standard onto a selected software life cycle model (SLCM)	(ISO/IEC/IEEE 2465:2017, 2017).
Software	Computer programs, procedures and possibly associated documentation and data pertaining to the operation of a computer system.	(ISO/IEC/IEEE 2465:2017, 2017).

3.1.2 On the ISO/IEC 29110 Standard Series.

When teaching or attempting to learn software engineering you should point out the importance of software engineering standards, emphasizing that a software engineering standard is documented knowledge gained from thousands of successful and failed projects, not knowing this could lead to the conclusion that years of progress and knowledge on the subject are being lost.(Claude Y. Laporte & Munoz, 2021)

"The correct selection and application of the appropriate standards should increase the productivity of an organization and have a positive economic impact on that organization. In software engineering, a major challenge is that the knowledge documented in the standards reaches an organization and is applied to its benefit."
(C. Y. Laporte et al., 2018)

For some organizations, standards end up being just a burden that cannot be avoided, having a negative impact on the company directly in terms of cost, time, complication of processes or the need for additional or specific personnel due to the use of standards. Therefore, some of these companies do not even consider implementing the use of standards, without being aware of the potential contribution that standards can make to their activities in a world as competitive as it is today (ISO, 2021).

Many organizations in both the public and private sectors use standards and/or participate in their development. Some see standardization as closely related to their business strategy. Others approach the use of standards in an organized way and understand the direct impact of using standards on their activities and performance. Others may use standards from a more limited perspective, exclusively for a specific project, process or activity. Most of them are aware that standards bring direct benefits to their organization. Some possible benefits are optimization of internal operations, innovative and scalable operations, creation of or entry into new markets (ISO, 2021)

Standards and associated technical documents could be considered a form of technology transfer and, if the right standards are selected and used correctly, should have an economic impact on an organization. In addition to the known benefits of standards, a French study has revealed five main lessons which were: Company value enhancement, Innovation, Transparency and ethics, International, Product and service quality. (Claude Y. Laporte & Munoz, 2021)

Most organizations that develop software are Very Small Entities (VSEs). According to the Organization for Economic Co-operation and Development (OECD) (2005), SMEs constitute many organizations in all countries of the world, accounting for more than 95%, and in some countries as much as 99%. This creates a challenge for OECD to provide a business environment that supports the competitiveness of this large business population (O'Connor & Laporte, 2017) .

VSEs often find it difficult to relate standards to their business needs. Most VSEs do not clearly see the direct benefit of the standards as well as sometimes may even lack the expertise to implement, fail to fill the positions requested, the cost and time required to accredit such standards. Reasons for not implementing standards include the perception that they are intended for larger organizations, are expensive, require a lot of documentation and bureaucracy, and do not establish clear software processes. This being a detrimental point for VSEs as without a documented development process or third-party certification such as ISO/IEC, most VSEs have little or no chance of being recognized as entities that generate high quality software systems or products. As such, they are excluded from some potential direct revenue streams that require the use of these. (Laporte et al., 2018)

Table 3.2 The advantages and disadvantages of voluntary standards reported ***(Claude Y. Laporte & Munoz, 2021)***

Advantages or benefits	Disadvantages or costs
• Promote innovation	• Difficult to understand
• Improve efficiency of an organization	• Cost of acquiring standards
• Increase competitiveness	• Cost of standard implementation
• Facilitate access to a wider market	• Cost of certification
• Clarify the rules of a market	• Require outside expertise to implement them
• Improve quality of products and services	• Conflicting standards
• Promote improvement of processes	• High number of standards available
• Facilitate partnerships	• Describe only "what has to be done" not "how to do it"
• Improve the image and credibility of organizations	• Insufficient guidance to select and apply them
• Promote a uniform terminology	• Slow evolution of standard may impede innovation
• Regularly updated	• Difficult and costly to apply in small organizations
• Facilitate the selection of suppliers and partners	• Difficult to demonstrate "savings"
• Facilitate access to recognize knowledge	• Many producers of standards
• Facilitate access to investments and financing	• Perception that standards add unnecessary bureaucracy to an organization
	• Language barrier for users that are not proficient in English

To help meet one of the needs of VSEs, the International Organization for Standardization and the International Electrotechnical Commission jointly published the ISO/IEC 29110 series of four-step standards and guides. These publications are aimed at VSEs, from start-up to adult, with little or no experience and without the expertise to select the appropriate software lifecycle or systems engineering standards processes and adapt them to the needs of a project (C. Y. Laporte et al., 2018). In other words, this standard is generated as a basis for the creation of new software developments with a primary focus on an application within the VSE where the VSE is defined as a company, organization, group or project with no more than 25 people.

Laporte (C. Y. Laporte & Miranda, 2020) published that VSEs expressed their need for assistance in adopting and implementing engineering standards. Most of the respondents indicated that they would like to receive more guidance through examples, additionally, it was noted that they asked for a light and easy-to-understand standards. More than 15% of respondents thought that engineering standards were difficult and bureaucratic and did not provide adequate guidance for use in a small environment. Finally, most VSEs felt that it was important to be assessed or certified to a standard to increase competitiveness, customer confidence, and satisfaction; improve software product quality; facilitate commercialization and greater export potential and decrease development risk.

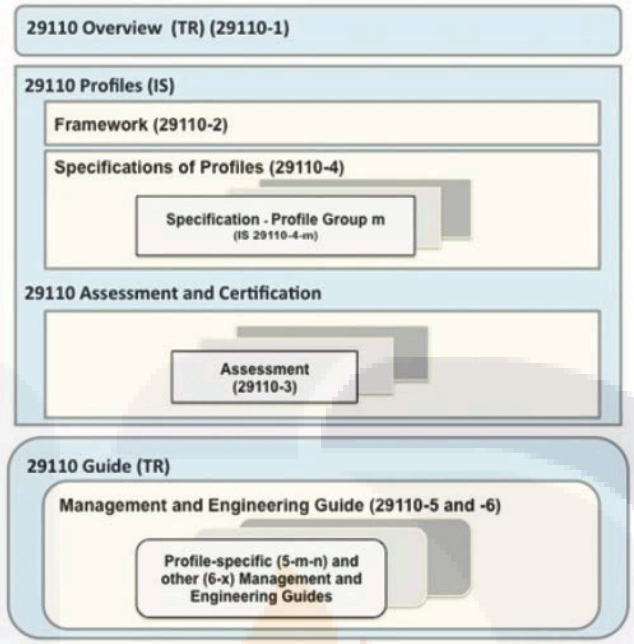

Figure 3.4 ISO/IEC 29110 Series. (Laporte, C.Y. & O'Connor, 2016)

According to the research, Laporte showed that more than 88% of respondents indicated that they were fully or largely satisfied, VSEs noticed improvements in quality, productivity, and their processes in the first few months. Another question was whether VSEs needed support to implement ISO/IEC 29110. According to their responses, 56% required assistance ranging from consultants/trainers and workshops to information from the web and other sources. Nearly 80% of responding customers were fully or largely satisfied with their suppliers. It is, therefore, a good starting point for follow-up doctoral work to support the necessary aspects of VSE (C. Y. Laporte & Miranda, 2020).

Table 3.3 - ISO/IEC 29110 target audience. (Laporte, C.Y. & O'Connor, 2016)

ISO/IEC 29110	Title	Target audience
Part 1	Overview	VSEs, customers, assessors, standards producers, tool vendors, and methodology vendors.
Part 2	Framework and taxonomy	Standards producers, tool vendors and methodology vendors. Not intended for VSEs.
Part 3	Assessment guide	Assessors, customers, and VSEs
Part 4	Profile specifications	Standards producers, tool vendors and methodology vendors. Not intended for VSEs.
Part 5	Management and engineering guide	VSEs and customers

The core of ISO/IEC 29110 is a set of pre-designed engineering and management guides that focus on project management and software or system development. ISO/IEC 29110 is designed for use with any life cycle, such as waterfall, iterative, incremental, evolutionary or agile (C. Y. Laporte & Miranda, 2020).

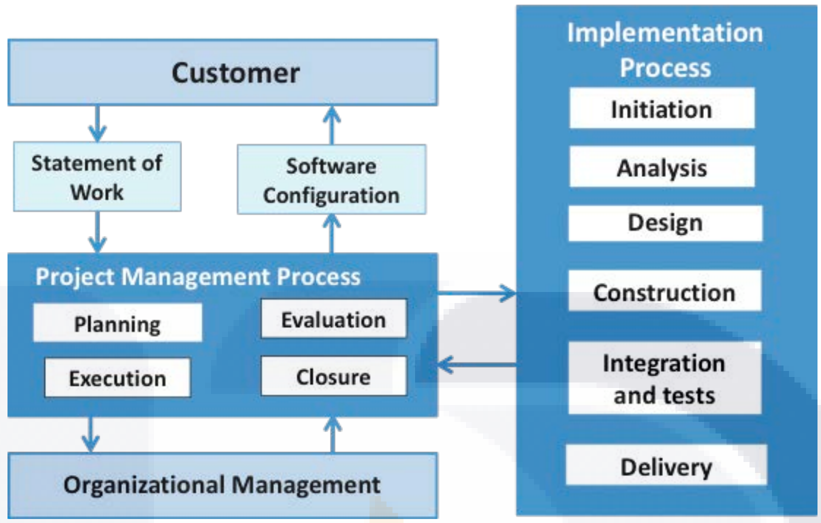

Figure 3.5 ISO/IEC 29110 Project Management and Software Implementation Relationship. (Laporte, C.Y. & O'Connor, 2016)

A core concept at the heart of ISO 29110 is that of "profile groups" that are a set of profiles. The "generic profile group" has been defined as applicable to VSEs that do not develop critical systems or critical software. The Generic Profile Group is a four-stage road map, called profiles, providing a progressive approach to satisfying a vast majority of VSEs. VSEs targeted by the "Entry profile" are VSEs working on small projects (projects that take no more than six person-months' effort) and startups. The "Basic profile" targets VSEs developing a single application with a single work team. The "Intermediate profile" is targeted at VSEs developing more than one project in parallel with more than one work team. The "Advanced profile" is targeted to VSEs that want to sustain and grow as an independent competitive system and/or software development business. (C. Laporte & O'Connor, 2017)

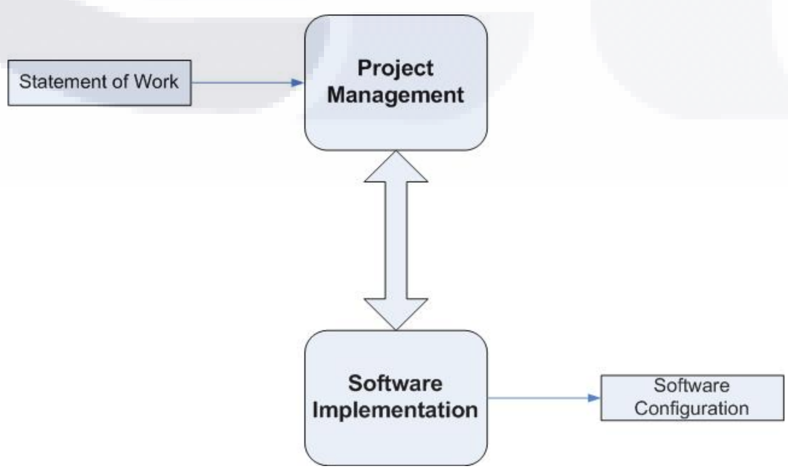

Figure 3.6 Processes of the Entry Profile (ISO/IEC TR 29110-5-1-1: 2012)

The ISO developed a road map with four steps, called profiles, targeting VSEs, ranging from start-ups to grownups

1. The *Entry* profile should be selected if the VSE is a start-up or works on small projects, such as those that require a six person-month effort.
2. The *Basic* profile should be chosen if the VSE develops a single product through a lone team.
3. The *Intermediate* profile should be selected if the VSE develops more than one project in parallel with more than one work team.
4. The *Advanced* profile should be chosen if the VSE wants to grow and remain an independent, competitive system and/or soft- ware development business.

Table 3.4 - Processes, tasks, work products and roles of each software profile

	Entry	Basic	Intermediate	Advanced
Number of processes	2	2	3 (+1 conditional)	3 (+3 conditonal)
Number of Tasks	40	67	107 (+ 8 conditional)	120 (+ 24 conditional)
Number of Work Products	14	22	39 (+ 3 conditional)	41 (+ 5 conditional)
Number of Roles	3	7	8 (+ 1 conditional)	8 (+ 1 conditional)

The ISO WG used the survey data to develop a set of requirements to produce a series of software and systems engineering standards and guides. Since 2011, hundreds of public and private organizations worldwide have implemented the ISO/IEC 291103-5 software series, as well as the systems engineering guides.6,7 For example, in Thailand, more than 450 public and private organizations have been certified as ISO/IEC 29110 compliant.8 Finally, trainers in more than 20 countries are teaching the series.9 Since ISO published the first standards and guides in 2011, more than 200 articles have been printed in peer-reviewed journals (C. Y. Laporte & Miranda, 2020).

Due to the focus of our study we will focus due to the complexity levels of the profiles on the first two profiles Entry and Basic. Where these two have a similar base in the Product Management and Software Implementation diagrams. (Figure 3.1.2.4 and Figure 3.1.2.5)

During the last decades, the market has been changing at an ever-increasing pace due to new requirements of systems, technology, and project personnel, where a different development style showed its advantages over the traditional one. This agile style of development directly addresses the problems of rapid change, seeking to be more effective with factors such as cost, staffing, and time within a project's potential problems. One of the priorities for managers working in an agile way is that it puts more emphasis on the people factors in the project: friendliness, talent, skill, and communication. Agile processes are designed to capitalize on the unique strengths of each individual and each team. "People trumps process" is one way of saying this. These points forewarn the importance of people when we talk about implementing a project in Agile development. (Alistair Cockburn & Highsmith, 2001)

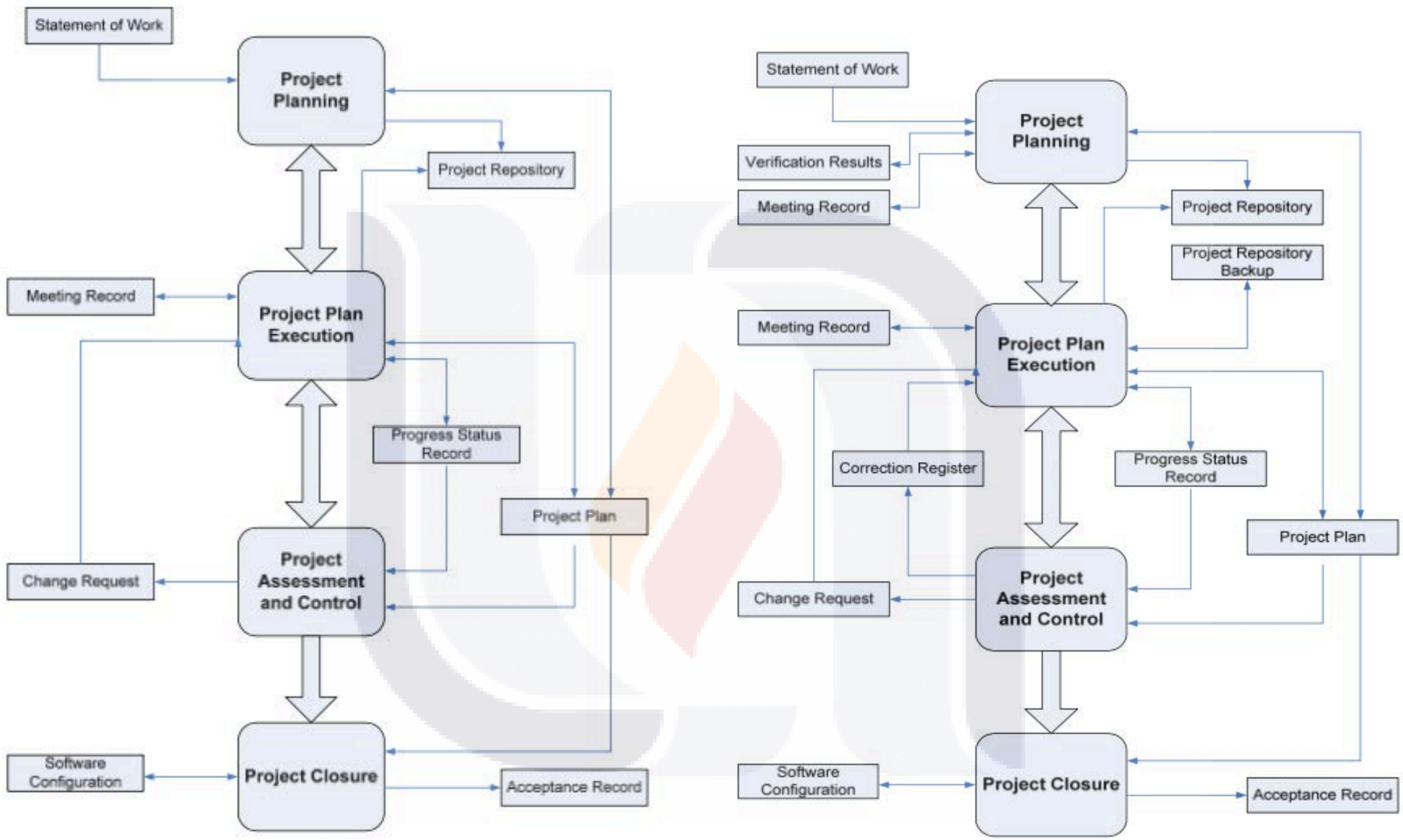

Figure 3.7 Comparison of diagrams, Project Management process diagram to Entry and Basic (ISO/IEC TR 29110-5-1-1: 2012 & ISO/IEC TR 29110-5-1-2: 2011)

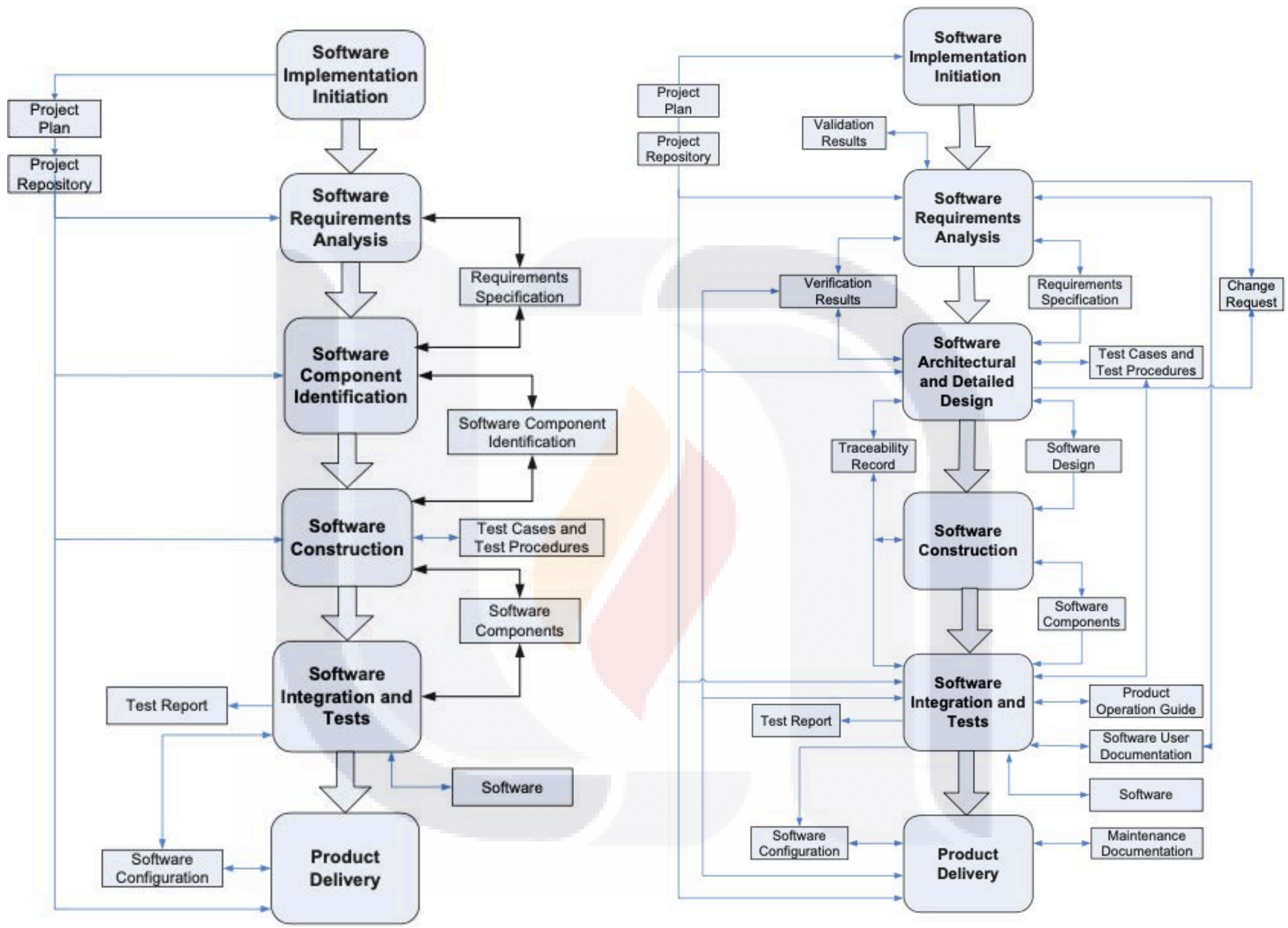

Figure 3.8 Comparison of diagrams, Software Implementation process diagram to Entry and Basic (ISO/IEC TR 29110-5-1-1: 2012 & ISO/IEC TR 29110-5-1-2: 2011)

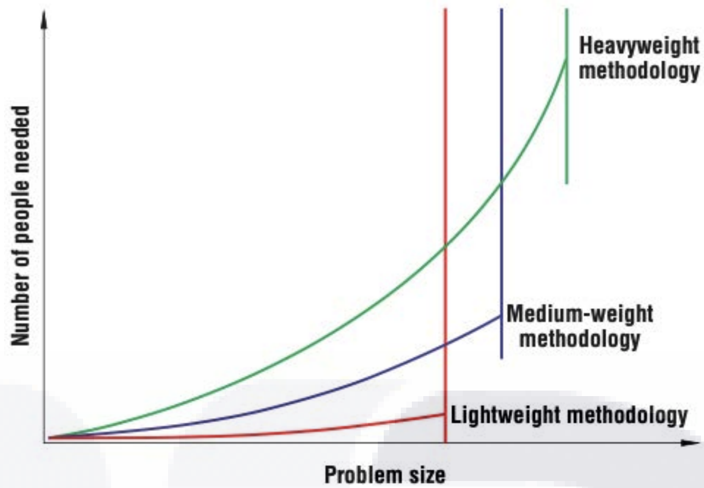

Figure 3.9 How problem size and methodology affect staff numbers. (A. Cockburn, 2000)

Figure 3.10 graphically depicts five critical factors within a project. Those 5 factors are Culture, Dynamism, Personnel, Criticality, and Size. In general, the closer to the center, the more the factors favor agility. This tells us that we must pay attention to people-related factors because success in software projects is directly related to the people who develop them. Turner and Boehm recommend finding and caring for ways to balance your technical and social skills.(Turner & Boehm, 2003)

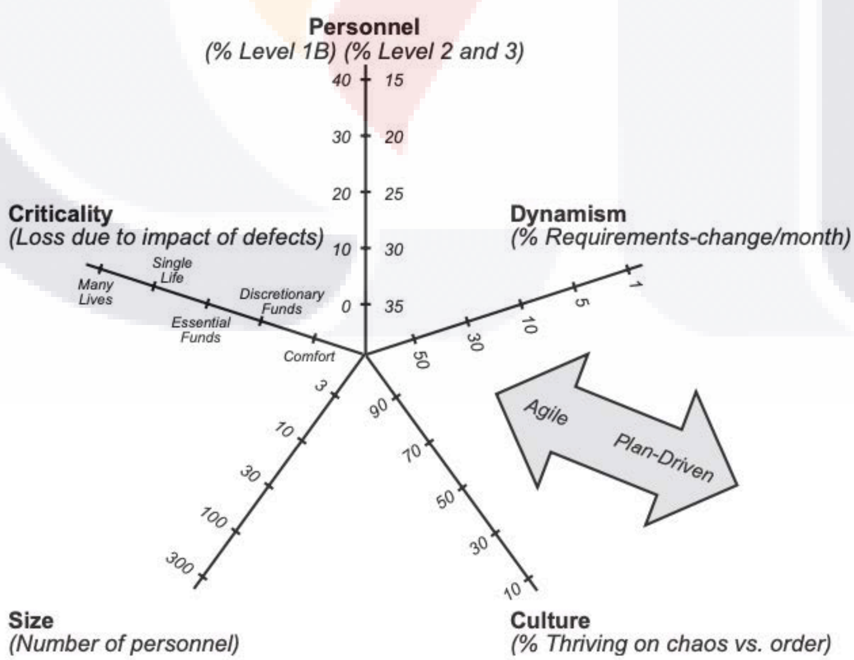

Figure 3.10 Key Discriminators of Agile and Plan-Driven Home Grounds (Turner & Boehm, 2003)

This document aims to develop a methodology that achieves the benefits of following an official standard such as ISO 29110, but at the same time, it seeks to find the favorable points of agile methodologies for the development of software projects, to be applied in small teams or organizations but with great quality in the results of the project, Figure 3.11.

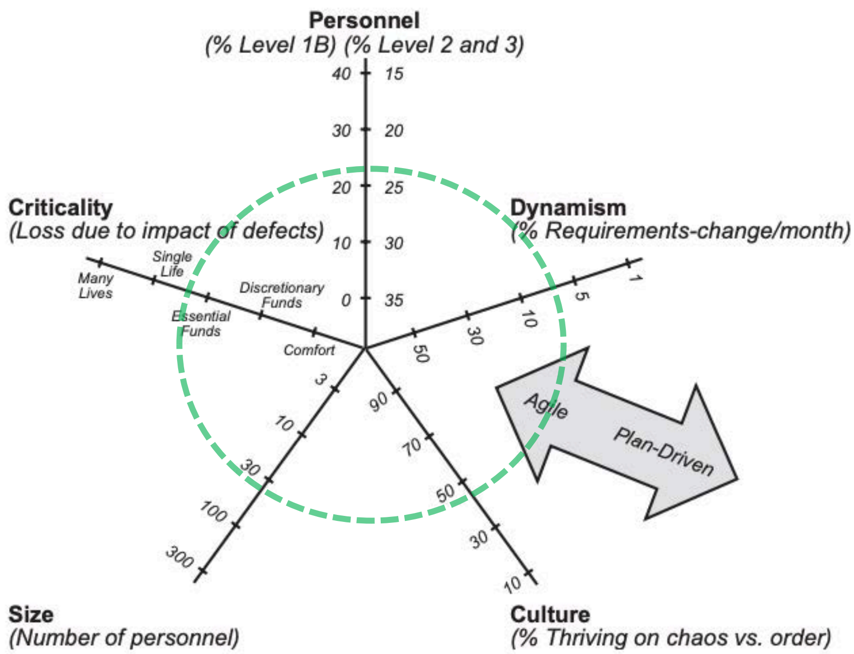

Figure 3.11 Segment where we will seek to maintain the standard to be generated.

For continuing with the process of this analysis, the Basic profile was selected as the profile to be used, for which the technical report ISO/IE TR 29110-5-1-2 and its specifications will be used, remembering that it is based on two central processes which are Project Management and Software Implementation, which were previously commented, as shown in Figure 3.6. It should be mentioned that this profile has 7 defined roles which are: Analyst, Customer, Designer, Programmer, Project Manager, Technical Leader, and Work Team. The relationship that exists between the two central processes is that in Project Management the *Project Plan* is generated producing the *Software Configuration* which is just what is required to start with the Software Implementation.

In the basic profile 22 products are handled and described within the whole process which are: 1. Acceptance Record, 2. Change Request, 3. Correction Register, 4. Maintenance Documentation, 5. Meeting Record, 6. Product Operation Guide, 7. Progress Status Record, 8. Project Plan, 9. Project Repository, 10. Project Repository Backup, 11. Requirements Specification, 12. Software, 13. Software Components, 14. Software Configuration, 15. Software Design, 16. Software User Documentation, 17. Statement of Work, 18. Test Cases and Test Procedures, 19. Test Report, 20. Traceability Record, 21. Verification Results, 22. Validation Results, Appendix B.

Table 3.5 ISO/IEC 29110 Phases.

Process	Phase	Description	Task	Products	Roles
Project Management	Project Planning	The Project Planning activity documents the planning details needed to manage the project	1.1 Review the Statement of Work	Statement of Work	Project Manager Technical Leader
			1.2 Define the Delivery Instructions for each of the Products.	Statement of Work Project Plan	Customer Project Manager
			1.3 Identify the specific Tasks to be performed to produce the Deliverables and their Software Components identified.	Statement of Work Project Plan	Project Manager Technical Leader
			1.4 Establish the Estimated Duration to perform each task.	Project Plan	Project Manager Technical Leader
			1.5 Identify and document resources. Include in the timeline the dates when Resources and training will be needed.	Statement of Work Project Plan	Project Manager Technical Leader
			1.6 Establish the Composition of Work Team assigning roles and responsibilities according to the Resources.	Project Plan	Project Manager Technical Leader

			1.7 Assign estimated start and completion dates to each one of the Tasks in order to create the Schedule of the Project Tasks	Project Plan	Project Manager Technical Leader
			1.8 Calculate and document the project Estimated Effort and Cost.	Project Plan	Project Manager
			1.9 Identify and document the risks which may affect the project.	Project Plan	Project Manager Technical Leader
			1.10 Document the Version Control Strategy in the Project Plan.	Project Plan	Project Manager Technical Leader
			1.11 Generate the Project Plan integrating the elements previously identified and documented.	Project Plan	Project Manager
			1.12 Include Product Description, Scope, Objectives and Deliverables in the Project Plan.	Statement of Work Project Plan	Project Manager Technical Leader

			1.13 Verify and obtain approval of the Project Plan.	Project Plan Verification Results	Project Manager Technical Leader
			1.14 Review and accept the Project Plan.	Project Plan Meeting Record	Customer Project Manager
			1.15 Establish the Project Repository using the Version Control Strategy.	Project Plan Version Control Strategy	Project Manager Technical Leader
Project Management	Project Plan Execution	The Project Plan Execution activity implements the documented plan on the project	2.1 Monitor the Project Plan execution and record actual data in Progress Status Record.	Project Plan Progress Status Record	Project Manager Technical Leader Work Team
			2.2 Analyse and evaluate the Change Request for cost, schedule and technical impact.	Change Request Project Plan	Project Manager Technical Leader

			2.3 Conduct revision meetings with the Work Team, identify problems, review risk status, record agreements and track them to closure.	Project Plan Progress Status Record Correction Register Meeting Record	Project Manager Technical Leader Work Team
			2.4 Conduct revision meetings with the Customer, record agreements and track them to closure.	Project Plan Progress Status Record Change Request Meeting Record	Customer Project Manager Technical Leader Work Team
			2.5 Perform backup according to the Version Control Strategy.	Version Control Strategy Project Repository Backup	Project Manager
			2.6 Perform Project Repository recovery using the Project Repository Backup, if necessary.	Project Repository Backup Project Repository	Project Manager
Project Management	Project Assessment and Control	The Project Assessment and Control activity evaluates the performance of	3.1 Evaluate project progress with respect to the Project Plan	Project Plan Progress Status Record	Project Manager Technical Leader Work Team

		the plan against documented commitments.	3.2 Establish actions to correct deviations or problems and identified risks concerning the accomplishment of the plan, document them in Correction Register and track them to closure.	Progress Status Record Correction Register	Project Manager Technical Leader Work Team
			3.3 Identify changes to requirements and/or Project Plan to address major deviations, potential risks or problems concerning the accomplishment of the plan, document them in Change Request and track them to closure.	Progress Status Record Change Request	Project Manager Technical Leader Work Team
Project Management	Project Closure	The Project Closure activity provides the project's documentation and products in accordance with contract requirements.	4.1. Formalize the completion of the project according to the Delivery Instructions established in the Project Plan, providing acceptance support and getting the Acceptance Record signed.	Project Plan Software Configuration Acceptance Record	Customer Project Manager
			4.2 Update Project Repository.	Software Configuration Project Repository	Project Manager

Software Implementation	Software Implementation Initiation	The Software Implementation Initiation activity ensures that the Project Plan established in Project Planning activity is committed to by the Work Team.	1.1 Revision of the current Project Plan with the Work Team members in order to achieve a common understanding and get their engagement with the project.	Project Plan	Project Manager Technical Leader Work Team
			1.2 Set or update the implementation environment.	Project Plan	Technical Leader Work Team
Software Implementation	Software Requirements Analysis	The Software Requirements Analysis activity analyzes the agreed Customer's requirements and establishes the validated project requirements.	2.1 Assign Tasks to the Work Team members in accordance with their role, based on the current Project Plan.	Project Plan	Technical Leader Work Team
			2.2 Document or update the Requirements Specification	Project Plan Requirements Specification	Customer Analyst
			2.3 Verify and obtain approval of the Requirements Specification.	Requirements Specification Project Plan Verification Results Change Request	Analyst Technical Leader
			2.4 Validate and obtain approval of the Requirements Specification	Requirements Specification Validation Results	Customer Analyst

			2.5 Document the preliminary version of the *Software User Documentation or update the present manual, if appropriate.	Requirements Specification *Software User Documentation	Analyst
			2.6 Verificar y obtener la aprobación de la Documentación del Usuario del Software.	*Software User Documentation Requirements Specification Verification Results Change Request	Analyst Technical Leader
			2.7 Incorporate the Requirements Specification, and Software User Documentation to the Software Configuration in the baseline.	Requirements Specification *Software User Documentation Software Configuration	Technical Leader
Software Implementation	Software Architectural and Detailed Design	The Software Architectural and Detailed Design activity transforms the software requirements to the system software architecture and	3.1 Assign Tasks to the Work Team members related to their role according to the current Project Plan.	Project Plan	Analyst Designer Technical Leader
			3.2 Understand Requirements Specification.	Requirements Specification	Analyst Designer
			3.3 Document or update the Software Design.	Requirements Specification Software Design Traceability Record	Analyst Designer

		software detailed design.	3.4 Verify and obtain approval of the Software Design.	Software Design Traceability Record Requirements Specification Verification Results Software Design Traceability Record Change Request	Analyst Designer
			3.5 Establish or update Test Cases and Test Procedures for integration testing based on Requirements Specification and Software Design.	Requirements Specification Software Design Test Cases and Test Procedures	Designer
			3.6 Verify and obtain approval of the Test Cases and Test Procedures.	Test Cases and Test Procedures Requirements Specification Software Design Verification Results	Analyst Designer
			3.7 Update the Traceability Record incorporating the Test Cases and Test Procedures.	Test Cases and Test Procedures Traceability Record	Designer
			3.8 Incorporate the Software Design, and Traceability Record to the Software Configuration as part of the baseline.	Software Design Test Cases and Test Procedures Traceability Record Software Configuration	Technical Leader

Software Implementation	Software Construction	The Software Construction activity develops the software code and data from the Software Design.	4.1 Assign Tasks to the Work Team members related to their role, according to the current Project Plan.	Project Plan	Programmer Technical Leader
			4.2 Understand Software Design.	Software Design	Programmer
			4.3 Construct or update Software Components based on the detailed part of the Software Design.	Software Design Traceability Record Software Components	Programmer
			4.4 Design or update unit test cases and apply them to verify that the Software Components implements the detailed part of the Software Design.	Software Components	Programmer
			4.5 Correct the defects found until successful unit test (reaching exit criteria) is achieved.	Software Components	Programmer
			4.6 Update the Traceability Record incorporating Software Components constructed or modified.	Software Components Traceability Record	Programmer
			4.7 Incorporate Software Components and Traceability Record to the Software Configuration as part of the baseline.	Software Components Traceability Record Software Configuration	Technical Leader

Software Implementation	Software Integration and Tests	The Software Integration and Tests activity ensures that the integrated Software Components satisfy the software requirements	5.1 Assign Tasks to the work team members related to their role, according to the current Project Plan.	Project Plan	Programmer Technical Leader
			5.2 Understand Test Cases and Test Procedures.	Test Cases and Test Procedures	Programmer
			5.3 Integrates the Software using Software Components and updates Test Cases and Test Procedures for integration testing, as needed.	Software Components Test Cases and Test Procedures Traceability Record Software	Programmer
			5.4 Perform Software tests using Test Cases and Test Procedures for integration and document results in Test Report.	Software Test Cases and Test Procedures Test Report	Customer Programmer
			5.5 Correct the defects found and perform regression test until exit criteria is achieved.	Software Test Report Test Cases and Test Procedures Traceability Record	Programmer
			5.6 Updates the Traceability Record, if appropriate.	Software Traceability Record	Programmer
			5.7 Document the Product Operation Guide or update the current guide, if appropriate.	Software *Product Operation Guide	Programmer

			5.8 Verify and obtain approval of the Product Operation Guide, if appropriate (see 5.7)	Product Operation Guide Software Verification Results	Designer Programmer
			5.9 Document the Software User Documentation or update the current one, if appropriate.	Software *Software User Documentation	Analyst
			5.10 Verify and obtain approval of the Software User Documentation, if appropriate (see 5.9)	*Software User Documentation Software Verification Results	Analyst Customer
			5.11 Incorporate the Test Cases and Test Procedures, Software, Traceability Record, Test Report, *Product Operation Guide and *Software User Documentation to the Software Configuration as part of the baseline.	Test Cases and Test Procedures Software Test Report Traceability Record *Product Operation Guide *Software User Documentation Software Configuration	Technical Leader
Software Implementation	Product Delivery	The Product Delivery activity provides the integrated software	6.1 Assign Tasks to the work team members related to their role, according to the current Project Plan.	Project Plan	Technical Leader Work Team
			6.2 Understand Software Configuration.	Software Configuration	Designer

		product to the Customer.	6.3 Document the Maintenance Documentation or update the current one.	Software Configuration Maintenance Documentation	Designer
			6.4 Verify and obtain approval of the Maintenance Documentation.	Maintenance Documentation Software Configuration Verification Results Maintenance Documentation	Technical Leader Designer
			6.5 Incorporate the Maintenance Documentation as baseline for the Software Configuration.	Software Configuration Maintenance Documentation	Technical Leader
			6.6 Perform delivery according to Delivery Instructions.	Project Plan Software Configuration	Technical Leader

When analyzing in detail the products and tasks of the Project Management Process and the Software Implementation Process, it was decided to generate the relationship between these two processes by unifying them in a single diagram Figure 3.12.

A comparison between the ISO/IEC TR 29110 methodology and the traditional or rigorous methodology, we chose to perform an analysis of the most common terms or keywords within each of these two variants, seeking to expose the similarities or differences through a Word cloud of each one. Figure 3.13 Word clouds according to ISO/IEC TR 29110 and Figure 3.14 Word Clouds to Rigor and Traditional Methodological.

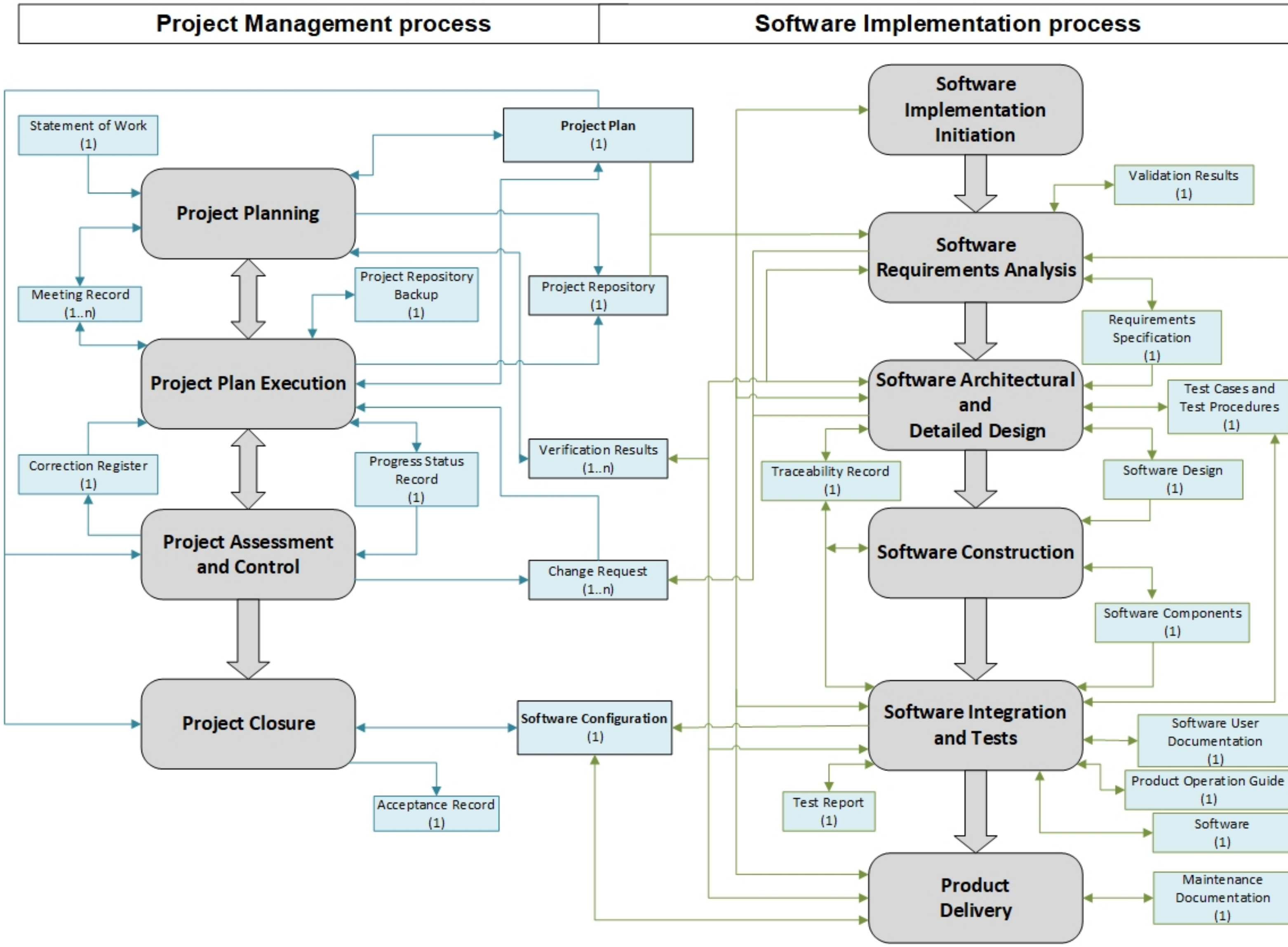

Figure 3.12 Correlation in the two processes of the ISO/IE TR 29110 : Basic profile (ISO/IEC TR 29110-5-1-2: 2011)

Figure 3.13 Word Clouds to ISO/IEC TR 29110

Figure 3.14 Word Clouds to Rigor and Traditional Methodological

3.1.3 On Analytics / Data Science Systems

3.1.3.1 Origin and Core Definitions (Analytics, Data Science, Big Data, Small Data, Analytics / Data Science for Big Data, Analytics / Data Science for Small Data, Big Data)

In the late 1960s, Analytics began to receive more attention as computers became decision support systems. With the development of Big Data, Data Warehouses, the Cloud, and a variety of software and hardware, Data Analytics has evolved significantly. Data analysis involves the investigation, discovery, and interpretation of patterns within the data.

Due to the growing enthusiasm around Data Science / Data Analytics and its many success stories, more and more organizations find themselves in the need to exploit these technologies, since many companies in the industry offer similar products and use comparable technologies, causing business processes to be among the last points of differentiation (Davenport, 2006). This has generated that organizations that use Data Science / Analytics generate competitive advantages that allow them to better understand the situation of their organizations, the market, and the competition. These companies come to know what their customers want, but they also know what prices those customers will pay, how many items they will buy, and what triggers will make them buy more products. In the same way, they can know when their inventories are running low and can predict problems with demand and supply chains, to achieve low inventory rates and high rates of perfect orders (Davenport, 2006).

Today, due to the enormous amount of data that is being produced at an unprecedented rate, this data is not effectively processed into information, delaying the extraction and production of knowledge. Therefore, our society faces even more challenging problems in transforming data into information and / or knowledge (Song & Zhu, 2016). This led to the creation of two concepts that use this data to generate value in organizations, such as Data Science and Data Analytics.

Since currently making accurate, timely and better decisions has become essential, but also a matter of survival in the complex and competitive current business context (Delen & Demirkan, 2013).

Analytics

Companies have spent the past forty years or so (Keen & Scott Morton, 1978) building their capabilities for analytics, or the systematic use of statistics and other quantitative methods to enhance decision-making (Davenport & Harris, 2017). The analytics started with a limited number of data sources that came from internal systems and the data that was collected from organizations, for traditional record keeping and transaction processing purposes. However, organizations wanted to extract useful information from the data to improve decision making, which was very difficult for the time because data acquisition was expensive and time-consuming (Viswanathan, 2014).

Table 3.6 Definitions of Analytics

Autor	Definition
Davenport and Harris, 2007	***"By analytics we mean the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based human analysis. ability to drive decisions and actions".***
Delen et al., 2020	Broadly speaking, analytics (or perhaps more appropriately, data analytics) can simply be defined as **"the discovery of meaningful patterns – new and novel information and knowledge – in data."**
Delen & Ram, 2018	***"Analytics (or perhaps more appropriately, data analytics) can simply be defined as "the discovery of meaningful patterns – new and novel information and knowledge – in data." Since we are living in an era of big data, the analytics definitions are mostly focused on that – data that are being created in large volumes, varieties with a high velocity".***
Chang et al., 2019	***"Is the systematic processing and manipulation of data to uncover patterns, relationships between data, historical trends and attempts at predictions of future behaviors and events".***
TA Runkler, 2020	***"Data analytics is defined as the application of computer systems to the analysis of large data sets for the support of decisions. Data analytics is a very interdisciplinary field that has adopted aspects from many other scientific disciplines such as statistics, machine learning, pattern recognition, system theory, operations research, or artificial intelligence".***
Boyd, 2012	***"Analytics is the scientific process of transforming data into insight for making better decisions."***
Tim Stobierski,2021	***"Data analytics refers to the process and practice of analyzing data to answer questions, extract insights, and identify trends. This is done using an array of tools, techniques, and frameworks that vary depending on the type of analysis being conducted".***

Since today's analytics can require extensive computation (Due to the volume, variety, and speed at which data is created, Big Data), the technical tools and algorithms used for analytics projects take advantage of state-of-the-art, state-of-the-art methods developed in a wide variety of fields including management science, computer science, statistics, data science, and mathematics.

One of the most important definitions is the one mentioned by Davenport & Harris who defined analytics as "**By analytics we mean the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based human analysis. ability to drive decisions and actions**". In Table 3.6 (Definitions of Analytics) you can see the most important definitions with some of the most important and recognized authors in the field of Data Analytics.

In analytics we can indicate that data analysis projects can be divided into several phases. The data is evaluated, selected, cleaned, filtered, visualized, and analyzed, to finally be interpreted and evaluated (Runkler, 2020). Figure 3.15 shows us the phases that are carried out in each of these phases to complete the Data Analytics process.

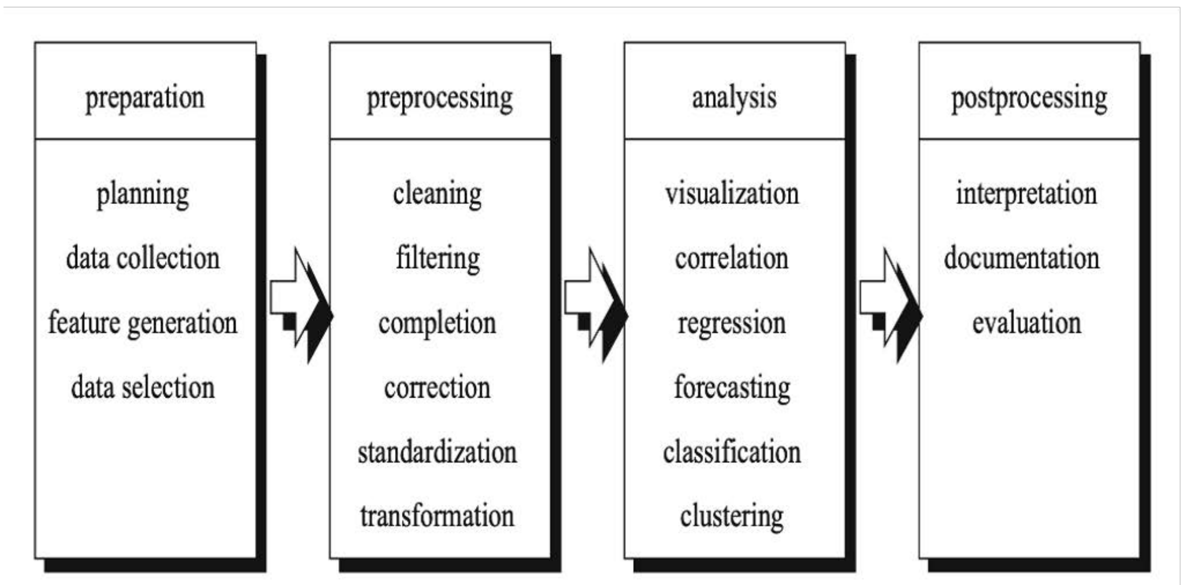

Figure 3.15 - Phases of data analysis projects (Runkler, 2020).

In the 1970s, decision support systems (DSS) were the first systems to support decision making. Over time, decision support applications became popular, such as executive information systems, online analytical processing, among others. Then, in the 1990s, Howard Dresner, a Gartner analyst, popularized the term Business Intelligence. A typical definition is that ***"BI is a broad category of applications, technologies, and processes for collecting, storing, accessing, and analyzing data to help business users make better decisions"*** (Watson, 2009a, p. 491).

With this definition, BI can be seen as an umbrella term for all applications that support decision making, and this is how it is interpreted in industry and, increasingly, in academia. BI evolved from DSS, and one could argue that analytics evolved from BI (at least in terms of terminology). BI can also be viewed as ***"data in"*** (to a data mart or warehouse) and ***"data out"*** (analyzing the data that is stored). A second interpretation of analytics is that it is the "pull data" part of BI. The third interpretation is that analytics is the use of ***"rocket science"*** algorithms (e.g., machine learning, neural networks) to analyze data. The progression from DSS to BI and analytics is shown in Figure 3.16 (Watson, 2014).

Within Analytics there are different types of analytics, where it is useful to distinguish between three types of analytics because the differences have implications for the technologies and architectures used for Big Data analytics (Watson, 2014).

Table 3.7 Analysis Types (Watson, 2014).

Type	Definition
Descriptive analytics	They are reports like dashboards, data visualization, they have been widely used for some time and are the core applications of traditional BI. Descriptive analyzes look back and reveal what happened. However, one tendency is to include predictive analytics findings, such as future sales forecasts, in dashboards.
Predictive analytics	Suggest about what will happen in the future. Methods and algorithms for predictive analytics, such as regression analysis, machine learning, and neural networks, have been around for some time. The ability to analyze new data sources, Big Data, creates additional opportunities for insight and is especially important for companies with large amounts of data. Golden Path analysis is an exciting new technique for predictive or analytics. It involves analyzing large amounts of behavioral data (that is, data associated with people's activities or actions) to identify patterns of events or activities that predict customer actions.
Predictive analytics	Predict what will happen, prescriptive analysis suggests what to do. Prescriptive analytics can identify optimal solutions, often for scarce resource allocation. It has also been researched in academia for a long time, but now being used more in revenue management it is becoming more common for organizations that have "perishable" assets such as rental cars, hotel rooms, and airplane seats. For example, Harrah's Entertainment, a leader in the use of analytics, has been using revenue management for hotel room rates for many years.

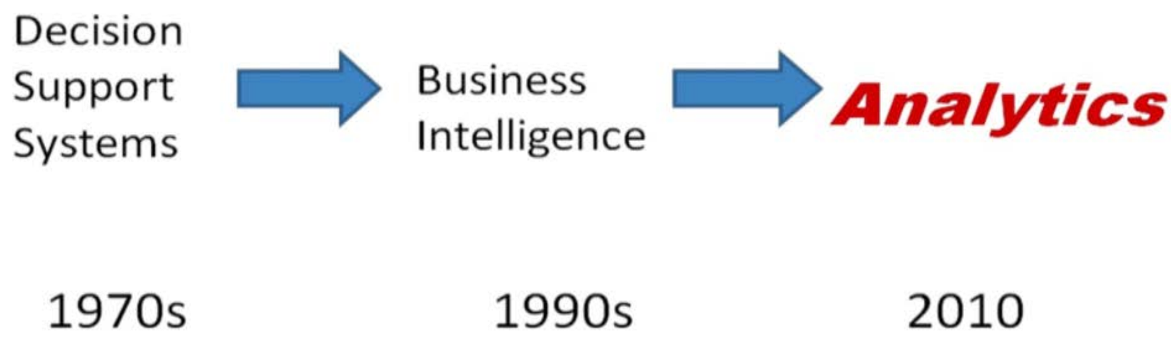

Figure 3.16 DSS & BI & Analytics. (Watson, 2014).

Data Science

The birth of Data Science as a discipline is relatively recent and arose from the need to control the massive volume of data that was emerging with the arrival of Big Data and the evolution of analytics, the data had to be quickly converted into information for analysis. Organizations began to focus more on prescriptive and predictive analytics using machine learning, as well as rapid analytics through visualization. (Larson & Chang, 2016). Big Data is a related field, often thought of as a subset of data science, in the sense that data science applies to large and small data sets and covers the comprehensive process of collecting, analyzing, and communicating data. analysis results.

Data Science is a body of principles and techniques for applying data analytic methods to data at scale, including volume, velocity, and variety, to accelerate the investigation of phenomena represented by the data, by acquiring data, preparing, and integrating it, possibly integrated with existing data, to discover correlations in the data, with measures of likelihood and within error bounds. Results are interpreted with respect to some predefined (theoretical, deductive, top-down) or emergent (fact-based, inductive, bottom-up) specification of the properties of the phenomena being investigated.

It is likely that the first appearance of **"Data Science"** as a term in the literature was in the preface to Naur's book ***"Concise Survey of Computer Methods"*** (Naur, 1974) in 1974. In that preface, data science was defined as "*the science of data processing, once established, while the relationship of the data with what they represent is delegated to other fields and sciences."* Another term according to Dhar, data science is defined as ***"data science is the study of the generalizable extraction of knowledge from data"*** (Dhar, 2013). Other definitions that we can find of Data Science are those shown in Table 3.8 (Definitions of Data Science) which are some of the most complete definitions and of the best-known authors in the field of Data Science

Table 3.8 Definitions of Data Science

Autor	Definition
Tukey, 1962	***"Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."***
C. Hayashi, 1998	Data science (DS, by its name in English Data Science) is a concept that not only synthesizes and unifies the field of statistics, data analysis and its related methods, but also seeks to understand the results obtained.
Provost and Fawcett, 2013	***“A set of fundamental principles that support and guide the principled extraction of information and knowledge from data”.***
O'Neil, 2014	***“Data science is an emerging discipline that integrates concepts in a variety of fields, including computer science, information systems, software engineering, and statistics”.***
Harvard University, 2023	***“Data science is inherently interdisciplinary as it combines expertise from statistics, computer science, mathematics, and domain-specific knowledge. This makes it incredibly versatile, with applications spanning healthcare, finance, marketing, and even environmental research.”***
Brodie, 2015	***“Data Science is concerned with analyzing Big Data to extract correlations with estimates of likelihood and error”.***
Van der Aalst, 2016 & Bichler et al., 2016	Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.
Chang et al., 2019	***“Data science is the methodology for the synthesis of useful knowledge directly from data through a process of discovery or of hypothesis formulation and hypothesis testing”.***

With the previous definitions it is clear to us that Data Science seeks to extract large amounts of data using the disciplines of mathematics, statistics, and computer science, which will help us identify patterns, increase efficiency, predict behaviors, recognize new market opportunities, reduce costs, generate competitive advantages, among others. Figure 3.17 (Three pillars of data science) shows three pillars of Data Science (Data, Technologies, and People), where Data refers to areas of domains such as relational data, non-relational data and even data collected from the Internet of Things. Technologies that include concepts such as Data Mining, Deep Learning, Machine Learning, Artificial Intelligence, among others. People who refer to the required personnel such as computer scientists, statisticians, data scientists, and business analyzers (Song & Zhu, 2016).

Among the three pillars, the most important are people. We can buy more computers, storage, and tools to efficiently process Big Data, but human capacity does not scale; Educating people, called data scientists, is key to addressing the challenges of the era of big data (Song & Zhu, 2016).

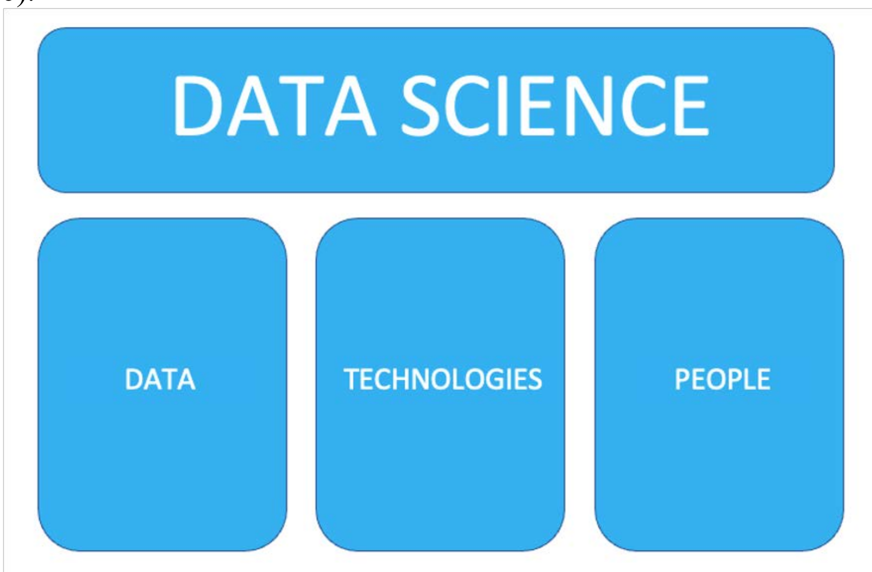

Figure 3.17 Three pillars of data science (Song & Zhu, 2016)

Data Science / Analytics

Considering the above, we can infer that there are few differences between Data Science and Analytics, since both focus on the transformation of data for knowledge, prediction, visual reports, improvement in decision making, among others. In addition to using the same fundamentals, mathematics, statistics, computer science and business as its main branches. And we can define Data Science and Analytics as: ***"An interdisciplinary field whose objective is to convert data into value, where data is transformed into knowledge to make better decisions, using statistical and quantitative analysis".***

Today, practitioners and academics often use the term "data analysis" or "data science" interchangeably with the older term knowledge discovery (Chen et al. 2012).

Data science and analytics projects generally aim to identify correlations and causal relationships, classify, and predict events, identify patterns and anomalies, and infer probabilities, interests, and sentiments.

This is done using a variety of tools, techniques, and frameworks that vary depending on the type of analysis being performed.

This can be seen reflected in Figure 3.18 (Fundamentals of data science and analysis), where it shows us how the three branches come together so that data science and analysis can exist. That is why we will unify both terms in this thesis referring to them as Data Science / Analytics.

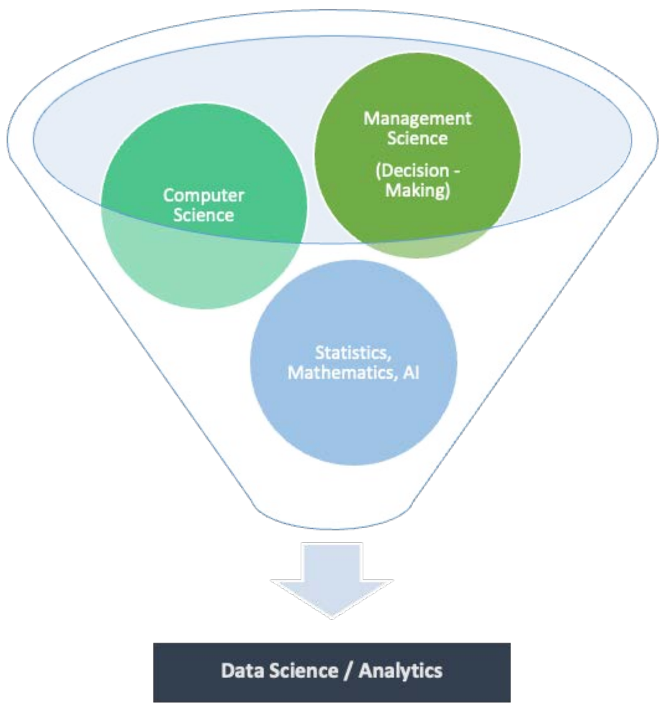

Figure 3.18 Foundations Data Science and Analytics

Big Data in Large Business (Big Data)

NASA researchers Michael Cox and David Ellsworth (1997; p. 236) were the first to refer to the term **'Big Data'** when they report, **"Visualization poses an interesting challenge for computer systems of computer systems: the data sets are often quite large, straining the capacity of main memory, local disk, and even remote disk, local disk, and even remote disk. We call this the big data problem".** They emphasize that even the supercomputers of that time could not process that amount of information, which is why in the article they mention a process for handling **'Big Data'**. Thus, implying that this problem of having information that exceeds the capabilities of computers to handle them in a traditional way is not a recent problem.

From an evolutionary perspective, Big Data is not new. One of the main reasons for creating data warehouses in the 1990s was to store large amounts of data (Gandomi et al., 2015). Figure 3.19 (Frequency distribution of documents containing the term ***'Big Data'*** in ProQuest Research Library) shows that the term Big Data became mainstream as recently as 2011.

Figure 3.19 Frequency distribution of documents containing the term "Big Data" in ProQuest Research Library (Gandomi et al., 2015)

Big Data describes a holistic information management strategy that is formed or constituted by a diversity of new types of data, the management of such data alongside traditional data. Although many of the techniques for processing and analyzing these types of data have been around for some time, it has been the massive generation of data and lower-cost computational models that have fostered wider adoption (Heller et al. 2015).
According to (Ali et al. 2013), the different ways to extract information from Big Data can be divided into three types that are:

- **Traditional enterprise data**: Transactional ERP data, includes customer information from CRM systems, general ledger data, and web store transactions.
- **Machine-generated /sensor data**: Includes manufacturing sensors, Call Detail Records, equipment logs, weblogs, trading systems data, and smart meters.
- **Social data:** Social media platforms like Facebook, micro-blogging sites like Twitter, includes customer feedback streams.

The data among others is commonly referred to as ***"Big Data"*** because of its volume, the speed with which it arrives and the variety of forms it takes. Big Data is creating a new generation of decision support data management because value is created only when data is analyzed and acted upon. One perspective is that big data is more and different types of data than traditional relational database management systems can easily handle. Currently, many data sources are not being leveraged as they should or could be. For example, customer emails, customer service chat and social media commentary can be processed to better understand customer sentiments. Web browsing data can capture every mouse movement to better understand customer buying behaviors. Radio frequency identification (RFID) tags can be placed on each piece of merchandise to assess the condition and location of each item.

However, considering the emerging nature of Big Data, there are several definitions which are shown in Table 3.9 (Definitions of Big Data) and Figure 3.20 shows the projected growth of big data (Watson, 2014).

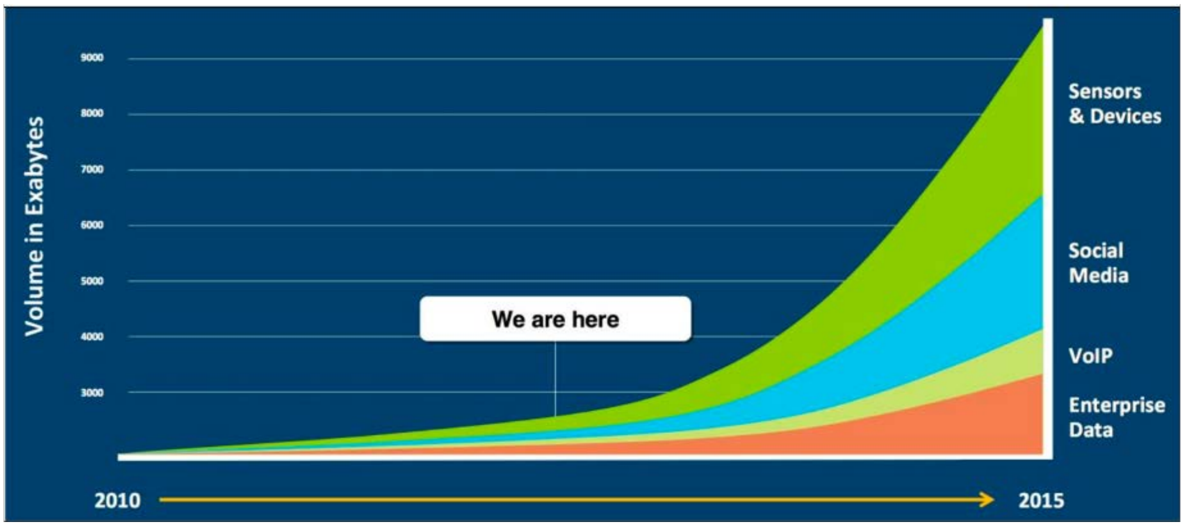

Figure 3.20 The Exponential Growth of Big Data (Palfreyman, 2013)

The current hype can be attributed to the promotional initiatives of certain leading technology companies that invested in building the analytics market niche. Some academics and professionals have considered "Big Data" as data that comes from various channels, including sensors, satellites, social media feeds, photos, videos, and cell phone and GPS signals (Rich, 2012).

Business intelligence and analytics (BI&A) and the related field of big data analytics have become increasingly important to both the academic and business communities over the past few decades. Through BI&A 1.0 initiatives, businesses and organizations across industries began to gain critical insights from structured data collected through various enterprise systems and analyzed by commercial relational database management systems. In recent years, web intelligence, web analytics, web 2.0 and the ability to mine unstructured user-generated content have ushered in a new and exciting era of BI&A 2.0 research, leading to unprecedented intelligence on consumer sentiment, customer needs and recognizing new business opportunities. Now, in this era of Big Data, even if BI&A 2.0 is still maturing, we stand on the brink of BI&A 3.0, with all the uncertainty that comes with new and potentially revolutionary technologies. (Chen et al., 2012) Figure 3.21 (BI&A Overview: Evolution, Applications, and Emerging Research) shows the evolution of BI&A, applications, and emerging analytics research opportunities.

Table 3.9 Definitions of Big Data

Autor	Definition
Michael Cox & David Ellsworth, 1997	***"Data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources".***
Jacobs, 2009	***"Data that is too large to be placed in a relational database and analyzed with the help of a desktop statistics/visualization package— data, perhaps, whose analysis requires massively parallel software running on tens, hundreds, or even thousands of servers".***
Russom , 2011	***"Description of the voluminous amount of unstructured and semi-structured data a company creates or data that would take too much time and cost too much money to load into a relational database for analysis".***
Chen et al., 2012	More recently big data and big data analytics have been used to describe the data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data storage, management, analysis, and visualization technologies.
Davenport et al., 2012	***"Data from everything including click stream data from the Web to genomic and proteomic data from biological research and medicine".***
Mills et al., 2012	***"Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes".***
Davoudian et al, 2020	***"They are an emerging class of scalable software technologies by which massive amounts of heterogeneous data are collected from multiple sources, managed, analyzed (in batch, in the form of a stream, or hybrid), and served to end users and applications. external. Such systems pose specific challenges in all phases of the software development life cycle and can become very complex due to the evolution of data, technologies, and target value over time".***

Table 3.10 Big Data Features

Attributes	Definition
Volume	The most recognized feature of Big Data is the presence of large data sets that allow us to analyze to extract valuable information (Chang et al., 2019). Organizations currently must learn to manage the large volume of data through new processes. Volume in Big Data can be defined as: ***"Large volume of data that either consume huge storage or consist of large number of records" (Russom, 2011).***
Variety	The word ***'Variety'*** denotes the fact that Big Data originates from numerous sources that can be structured, semi-structured, or unstructured (Schroeck et al., 2012). This is another critical attribute of Big Data as data is generated from a wide variety of sources and formats (Russom, 2011).
Velocity	Speed refers to the frequency of data generation and / or the frequency of data delivery (Russom, 2011). The high speed of Big Data can allow analysts to make better decisions, generating commercial value (Gentile, 2012). To utilize the high speed of data, many companies now use sophisticated systems to capture, store, and analyze data to make real-time decisions and retain their competitive advantages (Akter et al., 2016).
Veracity	High data quality is an important Big Data requirement for better predictability in the trading environment (Schroeck et al., 2012). Therefore, verification is necessary to generate authentic and relevant data, and to have the ability to filter incorrect data (Beulke, 2011). This tells us that data verification is essential to the data management process since erroneous data will hinder decision-making or guide analysts down the wrong path. Similarly, incorrect data would have little relevance to add commercial value (Akter et al., 2016).
Value	It is the added value obtained by organizations, value is created only when data is analyzed and acted upon correctly. To do this, we must identify all the data that will help us in the best way to generate value. This can be interpreted as: The extent to which big data generates economically worthy insights and or benefits through extraction and transformation. (Russom, 2011).

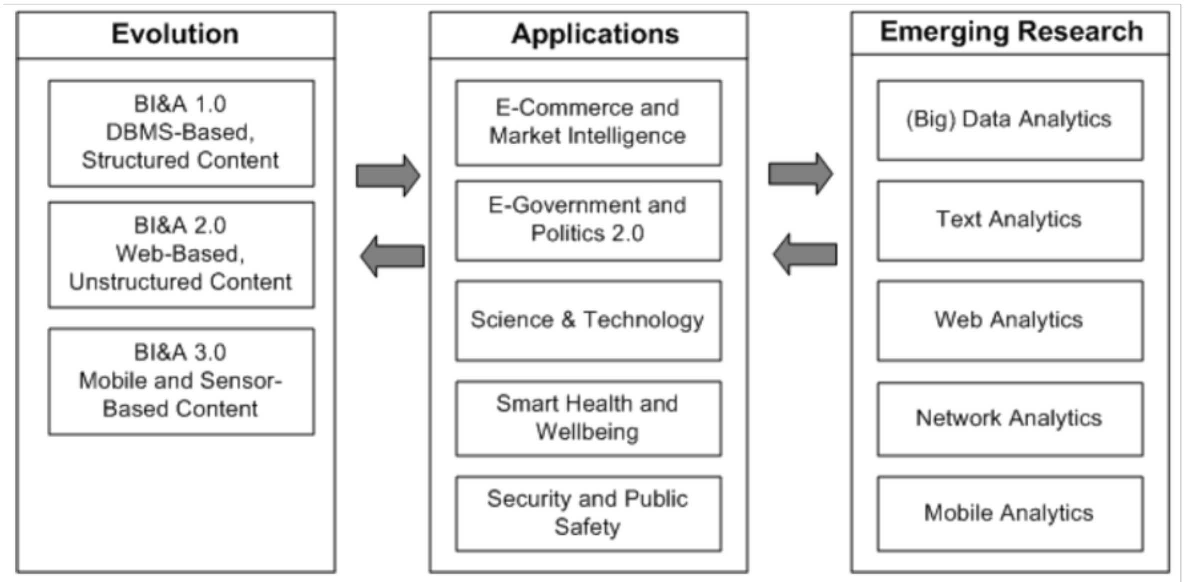

Figure 3.21 BI&A Overview: Evolution, Applications, and Emerging Research (Chen et al., 2012)

The opportunities associated with data and analytics in different organizations have helped generate significant interest in BI&A, which is often referred to as the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help a company better understand its business and marketplace and make timely business decisions. In addition to the underlying data processing and analytical technologies, BI&A includes business-centric practices and methodologies that can be applied to various high-impact applications such as e-commerce, market intelligence, e-government, healthcare, and security (Chen et al., 2012).

One of the most well-known characteristics of macro data is undoubtedly the volume of data that can be stored; However, this is not the only characteristic of Big Data and macro data. For example, Laney 200 suggested that volume, variety, and speed (or the three Vs) are the three dimensions of data management challenges. The Three Vs have emerged as a common framework to describe Big Data (Chen, Chiang, and Storey, 2012; Kwon, Lee, and Shin, 2014).

However, with the passage of time, new characteristics of Big Data were discovered, the 5V, Volume, Variety, Velocity, Veracity and Value. Table 3.10 (Big Data Features) describes each of these Big Data features, the three initially mentioned as well as the recently discovered features.

Big Data in Small Business (Small Data)

Until recently, the term Small Data was somewhat unknown, but thanks to the rapid growth and impact of Big Data, the term Small Data was used, that is, studies supported by data produced in a strictly controlled way using sampling techniques that limit its scope, temporality, size, variety and that they tried to capture and define its levels of error, bias, uncertainty, and origin (Miller, 2010). Unlike Big Data it is characterized by its generally limited volume, controlled data speed, limited data variety, usually structured data, and generally used to answer specific questions.

This has led some to ponder whether Big Data could lead to the disappearance of Small Data, or whether studies based on Small Data could be diminished due to its limitations of size, temporality, relativity, and cost. Indeed, Sawyer notes that funding agencies are increasingly pushing their limited funding resources into data-rich areas and big data analytics at the expense of small data studies, a trend that has continued in recent years (Kitchin, 2013).

The distinction between small and large data is recent. Before 2008, data was rarely considered in terms of "small" or "large." All data was, in effect, what is now sometimes called "small data", regardless of its volume. Due to factors such as cost, resources, and difficulties in generating, processing, analyzing, and storing data, limited volumes of high-quality data were produced through carefully designed studies using sampling frames designed to ensure representativeness (Kitchin et al., 2015).

So, the term "large" is somewhat misleading, as big data is characterized by much more than volume. Some "small" data sets can be very large in size, such as national censuses that also seek to be comprehensive. However, census data sets lack speed (usually done once every 10 years), variety (usually around 30 structured questions), and flexibility (once a census is established and administered, it is almost impossible to modify questions or add new questions) (Kitchin, 2014).

There are a variety of definitions about Small Data, which have been put forward since the early 1990s, but more recently, Thinyane described Small Data as: A perspective of Small Data as a human-centered approach to data valuation (Thinyane, 2017). In turn, table 3.11 (Definitions of Small Data) shows the most important definitions of Small Data through the years.

Table 3.11 Definitions of Small Data

Autor	Definition
Kitchin, 2015	***"In contrast, small data may be limited in volume and velocity, but they have a long history of development in science, state agencies, nongovernmental organizations, and businesses, with established methodologies and modes of analysis, and a track record of generating meaningful answers. Small data studies can be more precisely tailored to answer specific research questions and explore in detail and depth the various contextual, rational and irrational ways in which people interact and make sense of the world, and how processes work. Small data can focus on specific cases and tell individual, nuanced, contextualized stories. Thus, small data studies seek to extract gold from mining a narrow vein, while big data studies seek to extract nuggets by strip mining, excavating and sifting large tracts of land."***
Allen Bonde, 2013	***"Small data connects people with timely, meaningful insights (derived from big data and/or "local" sources), organized and packaged – often visually – to be accessible, understandable, and actionable for everyday tasks".***
Qoint, 2018	**"The few key pieces of meaningful, actionable information that we can uncover by analyzing big data. Those insights you extract from your big data become the last steps along the way to making better decisions."**
Doumar, 2014	**"Small data as alternative framing that focuses on the micro level analysis provides a complementary approach to the big data approaches, and to the approaches that are utilized for social indicators monitoring including and taking into account individual awareness, stakeholders commitment and Data maturity."**
Song et al., 2016	Meaning those data that do not necessarily possess all the first 4Vs of big data but still have value. Hence, small data are not a concept that describes the volume but is a relative concept to big data. Similarly, by 'small data analytics', we mean data analytics that does not necessarily involve big data specific technologies (i.e. Hadoop and NoSQL), but involve general techniques (i.e. statistics, data mining, machine learning, and visualization).

.

Comparative Big Data and Small Data

Table 3.12 Differences between Big Data and Small Data

Characteristics	Small Data	Big Data
Volume	In the range of GB to TB (10,000 – 100,000 records).	In the range of TB to ZB (1,000,000 – 1,000,000,000 records).
Velocity	Controlled and steady flow of data, accumulation of data is Slow.	Data arrives at very fast speeds; Huge amount of data gets accumulated within a short period of time.
Variety	Limited to wide (Structured Data).	Wide (huge variety of data).
Veracity	Contains less noise as data is collected in a controlled manner.	The quality of data is not guaranteed. Rigorous validation of data is required prior it's processing.
Value	High.	High.
Data Location	Data is located with an enterprise, local servers, regional servers, among others.	The data is present mainly in distributed storages in the cloud and in external unstructured databases of other owners and open data, combined with structured databases
Relationality Data	Strong.	Weak to strong.
Flexibility and Scalability	Low to middling.	High.

3.1.3.2 Review of Architectures of Big Data Software Systems Development Platform

Managing the information captured from companies and their clients to obtain a competitive advantage has become a very expensive property when using traditional data analysis methods, which are based on structured relational databases (Sawant, & Shah, 2013). This dilemma not only applies to large companies, but also to small and medium-sized companies, research organizations, governments, and educational institutions, which need less expensive computing and storage power to analyze complex scenarios and models involving images, videos, and other data, as well as textual data (Sawant, & Shah, 2013).

New sources of information include social media data, website clickstream data, mobile devices, sensors, and other machine-generated data. All these data sources must be managed in a consolidated and integrated way so that organizations obtain valuable inferences and knowledge (Chang et al., 2019).

The main objective of Big Data architecture is the analysis and processing of large amounts of data that cannot be carried out in a conventional way, because the capacities of standard storage, management and processing systems are exceeded (Chang et al., 2019). A Big Data management architecture should be able to design systems and models for the processing of large volumes of data from innumerable data sources in a fast and economical way, which allows better decision-making.

Big Data architecture has 5 main characteristics, these characteristics are the following:

- **Scalability:** It must be possible to easily increase data processing and storage capacities.
- **Fault tolerance:** System availability must be guaranteed, even if some machines fail.
- **Distributed data:** Data is stored between different machines, thus avoiding the problem of storing large volumes of data.
- **Distributed processing:** Data processing is performed on different machines to improve execution times and make the system scalable.
- **Data locality:** The data to be processed and the processes that process them must be close to each other to avoid network transmissions that add latency and increase execution times.

With the growth of the study and development of Big Data, data architecture designs have grown exponentially. They have migrated their operation to dynamic and flexible structures that leave behind the classic rigid structures, to give way to structures with the ability to assimilate structured and unstructured data. The architectural design of Big Data must be oriented to address five characteristics recognized in Big Data known as the "5V". These five characteristics refer to volume, speed, variety, truthfulness, and value.

Figure 3.22 Big Data architecture style shows us an example of the components that the Big Data architecture has, as well as Table 3.13 Components of Big Data architecture, describes the function of each of these components.

Table 3.13 Components of Big Data architecture (Microsoft, 2021).

Componentes	Descripción
Data Source	Data can be obtained from one or more sources, some of the examples can be: Data warehouses, relational and non-relational databases, statistical files produced by applications, web server log files, real-time data source, among others.
Data Storage	The data for batch processing operations is generally stored in a distributed file store that can contain large volumes of large files in various formats. This type of store is often called a data lake.
Batch Processing	Because the data sets are so large, a big data solution must often process data files using long-running batch jobs to filter, aggregate, and prepare the data for analysis.
Real Time Message Ingestion	If the solution includes real-time sources, the architecture must include a way to capture and store messages in real time for transmission processing. This could be a simple data store, where incoming messages are put into a folder for processing.
Steam Processing	After capturing messages in real time, the solution must process them by filtering, aggregating, and preparing the data for analysis.
Analytical Data Store	Many Big Data solutions prepare the data for analysis and then serve the processed data in a structured format that can be queried using analytical tools.
Analytics and Reporting	The goal of most Big Data solutions is to provide insight into the data through analysis and reporting
Orchestration	Most Big Data solutions consist of repeated data processing operations, encapsulated in workflows, that transform source data, move data between multiple sources and receivers, load the processed data into an analytical data warehouse, or push data. results directly to report or dashboard.

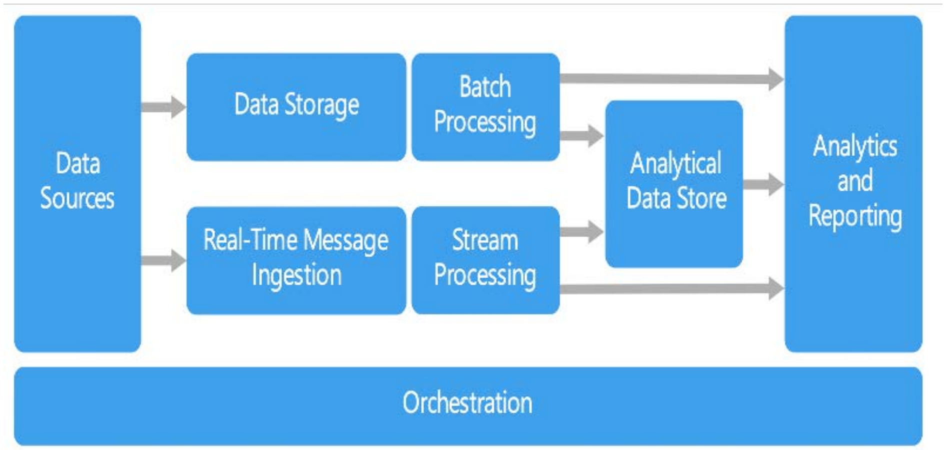

Figure 3.22 Big Data architecture style (Microsoft, 2021).

Before using Big Data, you must ensure that all Big Data architecture components are in place. Without this proper setup, it will be quite difficult to obtain valuable information and make correct inferences. If any of these components are missing, valuable data or correct decision-making cannot be obtained. Another example of Big Data architecture can be seen in Figure 3.23 The Big Data architecture, where it shows us in greater detail the components of the Big Data architecture. The architecture adapts to choose Open-Source frameworks or licensed products, for the case of this thesis we will focus on Open-Source type products only.

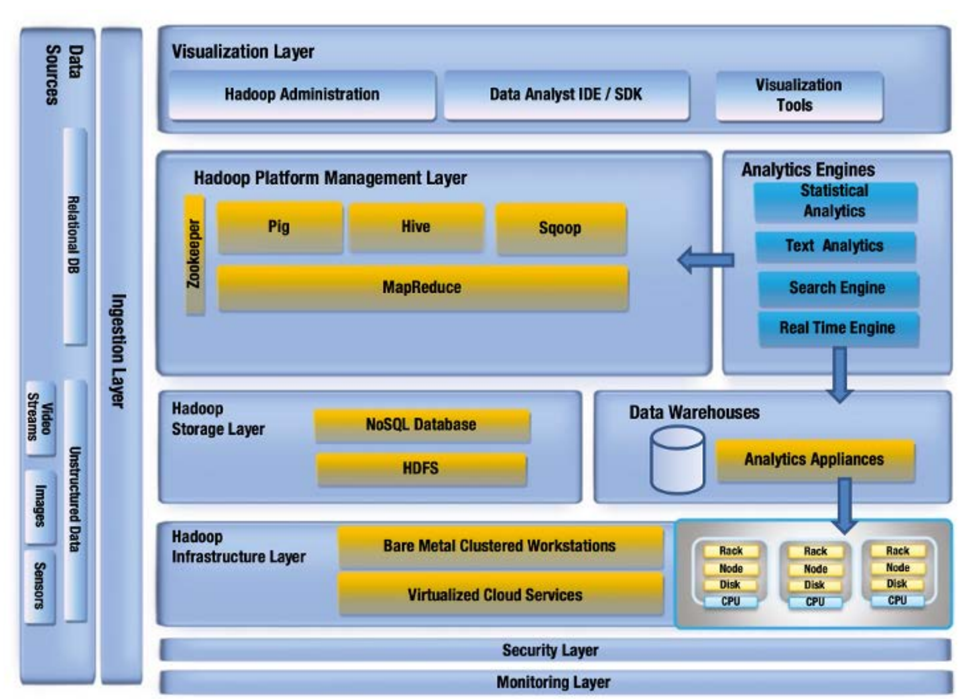

Figure 3.23 The Big Data architecture (Sawant, & Shah, 2013)

3.1.3.3 Review of Exemplary Big Data Analytics Systems

Figure 3.24 Information about of CIOs & Big DATA from (Kelly & Kaskade,2013)

Gartner Survey (2014): In 2014 Only 13% of respondents said their IT organizations put big data projects into production this year, but that's 5% higher than last year. But 24% of those polled voted against the use of big data technologies in their business. 73% of respondents have invested or plan to invest in big data in the next 24 months, up from 64% in 2013. As in 2013, much of the current work revolves around strategy development and the creation of pilots and experimental projects.

There are a lot of Big Data, Analytics, Data Science or Big Data Analytics projects these types of projects can vary in technologies, timing, budgets, number of personnel required where these factors are closely related to the technology of the company the key point of these projects are the goals, they seek to meet according to the Business goals. * These projects are not only limited to companies or IT research, for example at the European Bioinformatics Institute (EBI) in Hinxton (UK), which is part of the European Molecular Biology Laboratory and one of the world's largest repositories of biological data, currently stores 20 petabytes (1 petabyte is 1015 bytes) of data and backups on genes, proteins and small molecules. Genomic data account for 2 petabytes, a figure that doubles every year (Marx, 2013).

Big data burst onto the scene in the first decade of the 21st century, and the first organizations to adopt it were online companies and startups. Arguably, companies like Google, eBay, LinkedIn and Facebook were built around big data from the start. They didn't have to reconcile or integrate big data with more traditional data sources and the analytics that came from them, because they didn't have those traditional ways. They didn't have to merge big data technologies with their traditional IT infrastructures because those infrastructures didn't exist. Big data could stand alone, big data analytics could be the only approach to analytics, and big data technology architectures could be the only architecture (Davenport et al., 2013).

This is something interesting because these topics are the projects that "are fashionable" so there are many new research related to these, however due to the complexity of these projects and because they are new technologies not any company has the resources (personnel, knowledge, technologies, budget) for this type of projects so it is not so easy that any company can successfully carry out this type of projects, that is why we can see that the typical companies that are known to meet these requirements end up being those that have many resources or companies focused on technological innovation. We mention some examples, we start by mentioning cases where it can be clearly seen that these types of projects or companies were large in number of personnel, economic, information or other resources. Continuing with the traditional projects we will also see in more detail cases where these projects or technologies are not exclusive to companies with hundreds of employers, millions of data, or extremely robust infrastructures.

Example 1 : Big Data at UPS (Davenport et al., 2013).

Companies like GE, UPS, and Schneider National are increasingly putting sensors into things that move or spin, and capturing the resulting data to better optimize their businesses. Even small benefits provide a large payoff when adopted on a large scale. GE estimates that a 1% fuel reduction in the use of big data from aircraft engines would result in a $30 billion savings for the commercial airline industry over 15 years. Similarly, GE estimates that a 1% efficiency improvement in global gas-fired power plant turbines could yield a $66 billion savings in fuel consumption.

UPS is no stranger to big data, having begun to capture and track a variety of package movements and transactions as early as the 1980s. The company now tracks data on 16.3 million packages per day for 8.8 million customers, with an average of 39.5 million tracking requests from customers per day. The company stores over 16 petabytes of data.

Much of its recently acquired big data, however, comes from telematics sensors in over 46,000 vehicles. The data on UPS package cars (trucks), for example, includes their speed, direction, braking, and drive train performance. The data is not only used to monitor daily performance, but to drive a major redesign of UPS drivers' route structures. This initiative, called ORION (On- Road Integrated Optimization and Navigation), is arguably the world's largest operations research project. It also relies heavily on online map data, and will eventually reconfigure a driver's pickups and drop-offs in real time. The project has already led to savings in 2011 of more than 8.4 million gallons of fuel by cutting 85 million miles off of daily routes. UPS estimates that saving only one daily mile driven per driver saves the company $30 million, so the overall dollar savings are substantial. The company is also attempting to use data and analytics to optimize the efficiency of its 2000 aircraft flights per day.

Example 2 : Big Data at an International Financial Services Firm (Davenport et al., 2013).

For one multinational financial services institution, cost savings is not only a business goal, it's an executive mandate. The bank is historically known for its experimentation with new technologies, but after the financial crisis, it is focused on building its balance sheet and is a bit more conservative with new technologies. The current strategy is to execute well at lower cost, so the bank's big data plans need to fit into that strategy. The bank has several objectives for big data, but the primary one is to exploit "a vast increase in computing power on dollar-for-dollar basis.” The bank bought a Hadoop cluster, with 50 server nodes and 800 processor cores, capable of handling a petabyte of data. IT managers estimate an order of magnitude in savings over a traditional data warehouse. The bank’s data scientists—though most were hired before that title became popular—are busy taking existing analytical procedures and converting them into the Hive scripting language to run on the Hadoop cluster.

According to the executive in charge of the big data project, “This was the right thing to focus on given our current situation. Unstructured data in financial services is somewhat sparse anyway, so we are focused on doing a better job with structured data. In the near to medium term, most of our effort is focused on practical matters—those where it’s easy to determine ROI—driven by the state of technology and expense pressures in our business. We need to self-fund our big data projects in the near term. There is a constant drumbeat of ‘We are not doing “build it and they will come’—we are working with existing businesses, building models faster, and doing it less expensively. This approach is more sustainable for us in the long run. We expect we will generate value over time and will have more freedom to explore other uses of big data down the road.”

An international financial services firm initially acquired a big data infrastructure to exploit faster processing power. But in every case, analytics is the next frontier. Managers we talked to are building out their big data roadmaps to solve a combination of both operational and analytical needs, many of them still unforeseen.

> “The opportunities for cross-organizational analytics are huge,” the Executive in charge of big data told us. “But when the firm’s executives started discussing big data, the value-add was still esoteric. So we started instead by focusing on process efficiencies. We have 60 terabytes of what we consider to be analytics data sets, and we use compiled, multi-threaded code...and do periodic refreshes. We’re past some of the challenges associated with ‘fail fast’ and are tapping into all the advantages of Hadoop.”

Figure 3.25 Big Data and Data Warehouse Coexistence (Davenport et al., 2013)

Figure 3.25: Big Data and Data Warehouse Coexistence (Davenport et al., 2013).

Example 3 : Facilitating maintenance decisions on the Dutch railways using big data: The ABA case study (Núñez et al., 2014).

Currently, in different countries, a huge amount of railway track condition-monitoring data is being collected from different sources. However, the data are not yet fully used because of the lack of suitable techniques to extract the relevant events and crucial historical information. Thus, valuable information is hidden behind a huge amount of terabytes from different sensors. Considering the available data for railway condition monitoring, particularly when an increased measurement frequency is suggested to optimize maintenance decisions, these datasets qualify as Big Data. Thus, the popular 5V's for railway infrastructure are analyzed.

- Volume: Railway infrastructure is a distributed parameter system, which implies that the as- sessments should consider spatial and temporal dimensions. Monitoring the entire Dutch rail- way (more than 6500 km of tracks) with the ABA system only one time with different mea- surements provides a data volume of several terabytes. For example, when the system is im- plemented on commercial passenger trains to collect data all day, the data volume can exceed 100 terabyte a day because of the sampling speed of the required sensors (at least 25600 Hz for sampling and 16 sensors). A reduction/simplification of the specifications can compromise hit rates of defects and the quality of the high frequencies analysis.
- Velocity: With the requirement for early detection of problems and the desire to obtain good in- sight in the growth of defects, daily or weekly data acquisition is necessary. The main challenge with the current system is the processing time, which partly depends on human analysis of the data. Thus, the system update is currently a slow manual procedure. Moreover, when we collect data with an even higher frequency, this processing velocity is simply not feasible. Thus, com- putational intelligence is required to effectively process the available data, draw conclusions, and decide on the best maintenance action.
- Variety: In the railway infrastructure, different data-collecting systems are used, which leads into a wide variety of available data. In this paper, these data range from raw acceleration data of the wheels to images of the rail.
- Veracity: Different data sources have their own challenges when they are used to analyze railway track conditions. The results extracted from the ABA data can be different for the same defect in two runs, which depend on the wheel position on the track with respect to the defect. Although this problem is not present in the ultrasonic and eddy-current data, defects may go unnoticed because of reflections and other side effects of these techniques. For video imaging, only visible problems can be noticed. Deep cracks that do not penetrate the surface may be unobserved. Thus, the quality of each data source and the reliability of the conclusions drawn may differ.
- Value: Social aspects such as reduction of delays and the optimal track usage are the most ev- ident benefits when the performance and availability of public transport services are improved. Collecting railway infrastructure data on a daily basis will provide valuable data to facilitate maintenance decisions and a valuable data source for further research on the causes and growth of rail defects.

There is a great potential for using Big Data to facilitate maintenance decisions on Dutch railways. First, the ABA system can be implemented on a selected number of passenger trains and combined with night data from separate runs of video imaging and other systems. This method results in the collection of approximately 1 terabyte of raw data per day for the ABA data. By using selective data processing, based on previous results and experience in the growth rate of defects, all parts of the track can be monitored with appropriate intervals while maintaining the processing load within feasible limits. By also incorporating the failure and maintenance information in the system, the system can be adaptive and self-learning. In addition to the significant reduction of maintenance costs, this system can prove to be highly valuable for research by providing unprecedented amounts of track degradation data. Further studies that include the analysis of computational intelligence methodologies are considered.

Example 4 : Fog Based Intelligent Transportation Big Data Analytics in The Internet of Vehicles Environment: Motivations, Architecture, Challenges, and Critical Issues (Darwish & Bakar, 2018)

Every year there are about 8 million traffic accidents, which cause 7 million injuries and kill about 1.3 million people. About 90 billion hours of our time are lost due to traffic problems (accidents, traffic jams), resulting in a 2% decrease in total global household productivity. In addition, vehicular transport generates 220 million metric tons of carbon equivalents. Each year, in the United States, the cost of personal transportation by car (excluding commercial and public transportation) is about \$3 trillion, of which 40% is due to parking, crashes, pollution and traffic services. To improve the performance of transportation systems, increase road safety and preserve the environment, the concept of intelligent transportation system (ITS) was introduced. The emergence of ITS was greatly aided by the advancement of sensing and communication technologies and the evolution in the effective integration of networked information systems, decision making and physical infrastructure. The Internet of vehicles (IoV) connects the ITS devices to cloud computing centres, where data processing is performed.

The evolution of sensing and communications technologies and the advances in intelligent data processing are the driving forces for realizing the intelligent transportation systems concept, which is a main component of smart cities. Similar to many modern life aspects, transportation management and control is now becoming more data-driven . The applications of ITSs are data-intensive, complex, and the ‘‘5Vs of Big Data’’ can describe their characteristics precisely [34]:

- The first ‘‘V’’ is the volume of ITS data, which has exponential growth. For example, in 2013 each auto- motive manufacturer collected around 480 TB of data and an increment to reach 11.1 PB/year is expected by 2020.
- The second ‘‘V’’ of ITS data is the variety. This characteristic describes the various ways of collecting data in different formats such as numerical data gathering through sensors on both infrastructure and vehicles, multimedia and text data capturing from social media, and GIS and image data loading for digital maps. The organization level of such data varies from semi-structured to structured data. The variety of this data

creates highly heterogeneous data sets that impose serious challenges in the ingestion, integration and processing stages of a data analytic system.

- The third ‘‘V’’ is the velocity of ITS data, which varies widely. Data generation and collection rate can be continuous real-time collection and in certain applications data are collected at regular intervals. Similarly, the requirements of processing vary greatly from real-time event processing to batch processing. However, real- time data collection and processing induce high requirements on networks and data processing centers.
- The fourth ‘‘V’’ is for veracity which describes the ITS data trustworthiness level. In fact, the ITS community is facing significant challenges in providing timely and reliable transportation related data collection.
- The fifth ‘‘V’’ is for the ITS data value, which depends on the data age, their sampling frequency, and their usage purpose. For instance, for a collision avoidance application, few minutes old data may have no value. On the other hand, route planning applications can benefit from non-real-time data. The value is a characteristic to measure the ability to extract from data meaningful and actionable business insights.

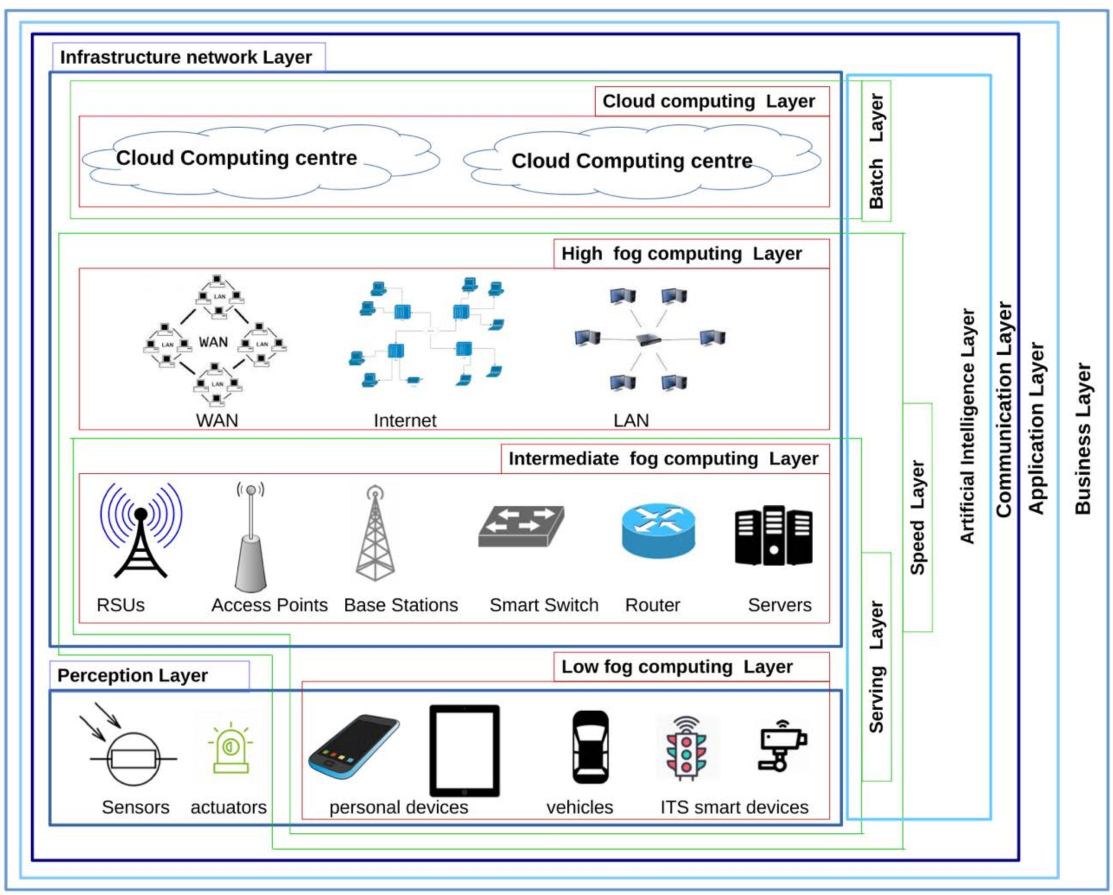

Figure 3.26 Real-time intelligent transportation system big data analytics (RITS- BDA) architecture. (Darwish & Bakar, 2018)

Real-time big data analytics consists of three main stages, including batch, velocity and service. However, performing these three stages in the cloud is not going to serve latency-sensitive applications. Moreover, the fog platform cannot handle the batch processing stage. Therefore, the big data analysis stages must be distributed between the cloud and fog computing layers. In addition, the IoV environment must provide the necessary coordination and communication between the different layers and components.
Considering these aspects, this paper proposes a novel three-dimensional architecture (intelligent computing, real-time big data analytics, and IoV) to enable real-time ITS big data analytics in the IoV environment. In addition, the opportunities and challenges that IoV and intelligent computing platforms are creating have been discussed. In addition, a comparison between different edge computing technologies has been presented. In addition, critical issues and future research directions, which should be considered to improve real-time big data analytics for many ITS applications, have been highlighted.
Finally, the proposed architecture presents a good foundation for future research in this field and can be used as part of intelligent transportation systems to enable real-time applications such as collision avoidance, hazard warning, advanced driver assistance systems, and autonomous driving. As a result, the lives of many people will be saved by using safer transportation systems. In addition, transportation systems will become more efficient and environmentally friendly.

Example 5 : Integrated Understanding of Big Data, Big Data Analysis, and Business Intelligence: A Case Study of Logistics (Jin & Kim 2018)

Case Study: CJ Logistics

This study uses the case of CJ Logistics, Korea's largest logistics company. It examines the sorting process, especially regarding decisions about loading/ unloading docks and hub terminals, which are at the core of courier services, to examine the effective use of big data/BDA through BI.

CJ Logistics was selected as the research subject as it is the largest logistics service provider in Korea with the highest market share and sales revenue of KRW 7110.3 billion in 2017 . In addition, as shown in next Figure (big data case of CJ Logistics, March 2018), the company is an innovation leader in the industry. It is traditionally considered a 3D business that uses BI based on high-tech automation-oriented technology, engineering, and system and solution plus consulting (TES + C), while actively and rapidly adopting big data/ BDA at the same time.

CJ Logistics is a market leader equipped with cutting-edge logistics technologies, including real-time tracking of freight, an integrated courier and freight tracking system that enables users to view customer information and requirements, satellite vehicle tracking, and temperature control systems . In 2017, CJ Logistics invested more than KRW 120 billion to automate its sorting process through sub-terminals to aid sustainable growth. CJ Logistics” infrastructure is more than three times bigger than that of its closest competitor in the courier service industry. With five hub terminals, more than 270 sub-terminals, and more than 16,000 vehicles, CJ Logistics processes more than 5.3 million packages per day. Its mega hub terminal in Gwangju, Gyeonggi-do Province—which was due for completion in August 2018 with an investment of more than KRW 400 billion—will utilize convergence technologies such as big data, robots, and loT to expand its services for the convenience of its customers across Korea. This will include same-day delivery, same-day return, and scheduled delivery services. The company is simultaneously moving forward with its planned international growth. At the end of 2017, CJ Logistics had a global network of 238 centers in 137 cities and 32 countries. It opened the Shenyang Flagship Center, a mammoth logistics center in Shenyang, China, on 15 June 2018. The purpose of this investment was to accelerate the company's business in northern Asia, including three provinces of northeastern China—Liaoning, Jilin, and Heilongjiang. The company has implemented huge capital expenditure to broaden its business efficiently, laying the groundwork for sustainable growth and expansion by raising the entrance barrier for rivals (big data case of CJ] Logistics, March 2018).

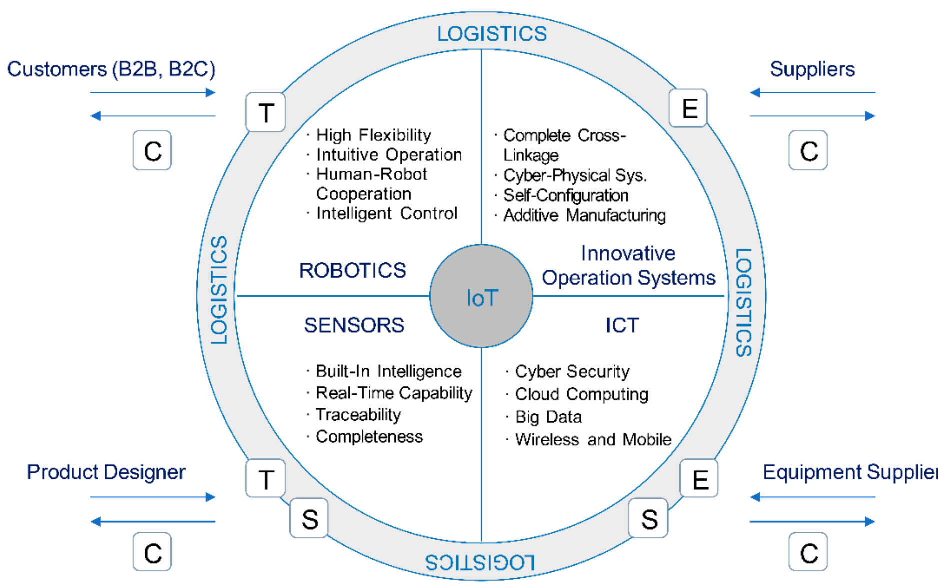

Figure 3.27 Technology, engineering, system and solution plus consulting (TES+C) of CJ Logistics. (Jin & Kim 2018)

CJ Logistics mainly uses a hub-and-spoke system, which connects points via hubs or logistics centers dealing with massive cargo volumes in its courier service; it also uses a point-to-point operational system directly connecting origins and destinations. The point-to-point system delivers to and from terminals, saving time on package arrivals while alleviating capacity issues during the peak season. However, growing volumes may increase costs, as they require more investment in terminals; a volume imbalance among terminals can cause unnecessary additional costs. On the other hand, in the hub-and-spoke system, packages are gathered and sorted in a large terminal before being delivered to a destination terminal. The advantage of this system is that it reduces arrival time to the terminals, easing the imbalance in volume. However, the disadvantages are that it may delay deliveries to distant or rural areas during the peak season and requires a large-scale hub terminal.

Since CJ Logistics mostly uses the hub-and-spoke system, whose core is the logistics process at the hub terminal, this study focuses on decisions concerning the loading /unloading docks in the process. This focus point was selected for the following reasons. First, few previous studies have focused on this segment, even though it has greater room for improvement regarding productivity and efficiency than other segments. Second, the importance of this segment may have been overlooked, since standardizing the process is challenging owing to differences in the environment, such as the distance between buildings or shape of the space. Third, there are many other difficulties to address, including outsourcing, warehouse management, freight payment, inventory management, packing, customs clearance, and customer claims. Many courier service providers allocate hub terminal docks for loading /unloading simply according to terminal conditions, such as the distance between docks and number of packages, mostly based on past experience. By contrast, CJ Logistics has dramatically improved productivity and efficiency by “seeing the unseen” through the use of big data/BDA and promoting faster and better decision making through BL.

The hub terminal process was selected from the three general stages of courier services, namely, pick-up, transport/sorting, and delivery (next Figure). This process was selected because it is the central process connecting pick-ups from different locations with delivery to different destinations.

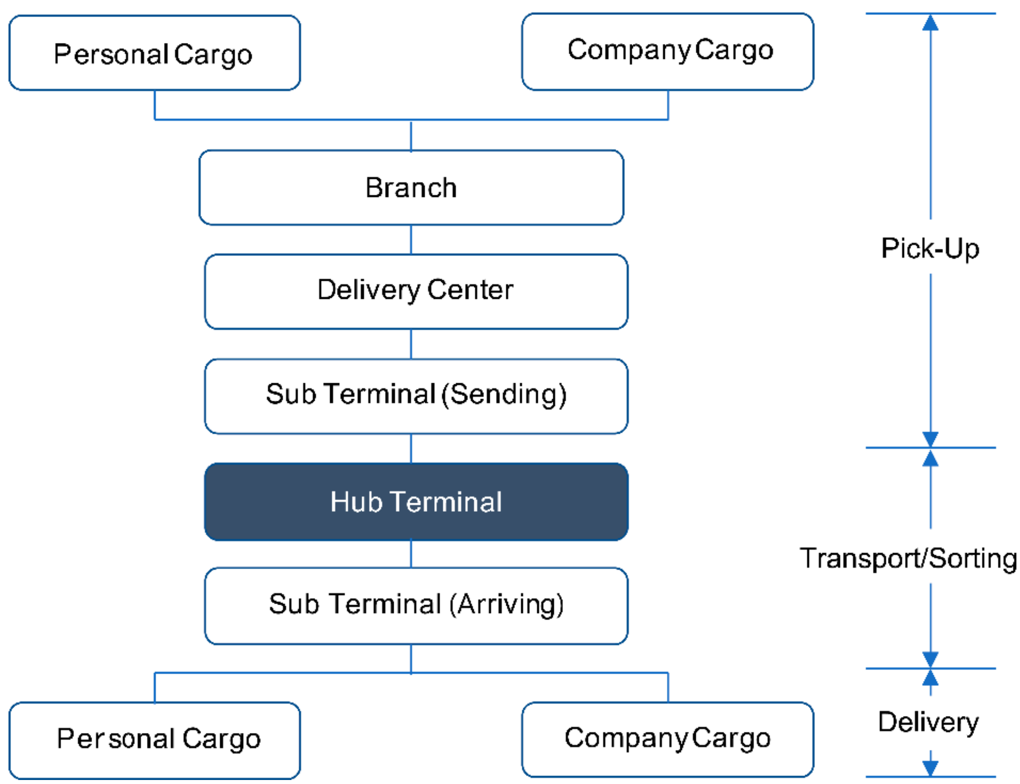

Figure 3.28 : General courier service structure. (Jin & Kim 2018)

An incident that occurs at the hub terminal can have a serious impact on the entire cycle from pick-up to delivery and could cause a bottleneck effect at hub terminals. This is a significant issue that needs to be addressed to secure growth in the industry, as it can paralyze transportation and delivery within a company on a large scale. Resolving this issue alongside difficulties in other areas by using big data/BDA could improve company productivity and efficiency as a whole.

Example 6 : Big Data Techniques for Public Health: A Case Study (Katsis et al., 2017).

Public health researchers increasingly recognize that to advance their field they must grapple with the availability of increasingly large (i.e., thousands of variables) traditional population-level datasets (e.g., electronic medical records), while at the same time integrating additional large datasets (e.g., data on genomics, the microbiome, environmental exposures, socioeconomic factors, and health behaviors). Leveraging these multiple forms of data might well provide unique and unexpected discoveries about the determinants of health and wellbeing. However, we are in the very early stages of advancing the techniques required to understand and analyze big population-level data for public health research.

To address this problem, this paper describes how we propose that big data can be efficiently used for public health discoveries. We show that data analytics techniques traditionally employed in public health studies are not up to the task of the data we now have in hand. Instead we present techniques adapted from big data visualization and analytics approaches used in other domains that can be used to answer important public health questions utilizing these existing and new datasets. Our findings are based on an exploratory big data case study carried out in San Diego County, California where we analyzed thousands of variables related to health to gain interesting insights on the determinants of several health outcomes, including life expectancy and anxiety disorders. These findings provide a promising early indication that public health research will benefit from the larger set of activities in contemporary big data research.

A BIG DATA CASE STUDY

To explore how big amounts of population-level data can be leveraged to make interesting public health discoveries, we worked on a case study centered on public health issues in San Diego County, California. The choice of location was made primarily for two reasons: First, the ease of getting access to large datasets, since it is the county where UC San Diego is located. Second, the diversity of the county, which makes it especially interesting for public health researchers: San Diego County's location (being close to the US border with Mexico and covering a large area from the Pacific Ocean coast to the desert), magnitude (being the fifth most populous county in the US), and population characteristics give it a unique environmental, ethnic, and socioeconomic diversity.

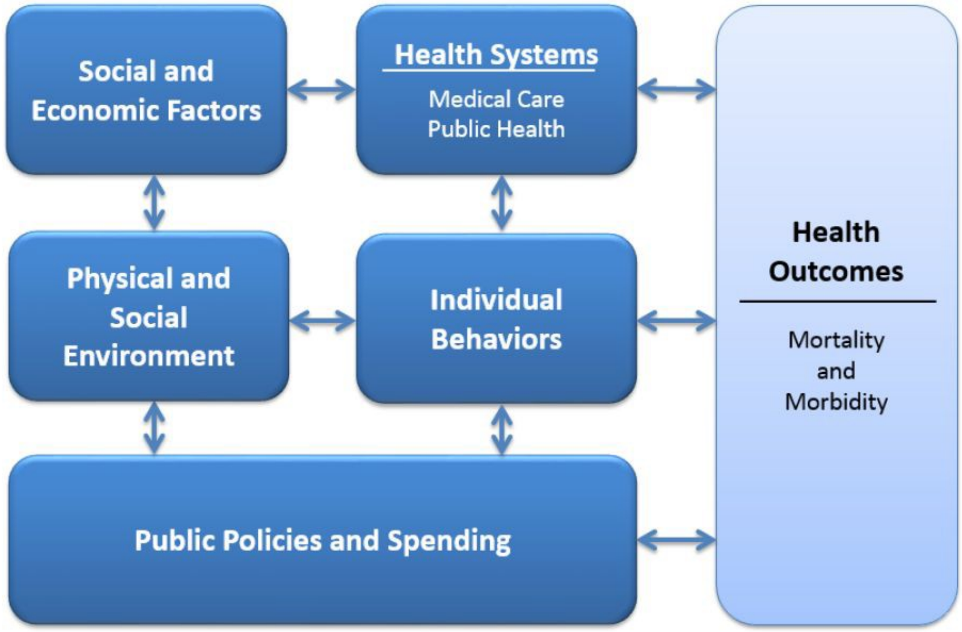

Figure 3.29 High-level grouping of determinants of health (Katsis et al., 2017).

To bootstrap our study, we identified and integrated a large number of representative data (in the order of thousands of indicators) covering the high-level groups of factors that are known to affect our health (shown on the past Figure)social and economic factors (such as education and income), physical and social environment (such as traffic density and air pollution), individual behaviors (such as smoking, exercising, and consumer buying patterns), health systems (such as insurance status), and health outcomes (such as hospitalization and emergency department visits for different conditions).

Since different datasets were provided at different geographic granularities, we ended up with two sets of integrated data: The first dataset contained 3,818 indicators at the level of the subregional areas (SRAs) (of which there are 41 in the San Diego County). While this dataset contained important health outcome information (i.e., hospitalization and emergency department visit data for different conditions), its geographic granularity was restricted due to privacy reasons. Therefore, we also created a second dataset that contained 22,712 indicators at the level of census tracts (of which there are 628 in the San Diego County). The next Figure shows the data that were integrated into each of the two datasets.

To analyze the data we experimented with two broad classes of big data analytics techniques that cover the two ends of the spectrum between targeted hypothesis-driven discovery and open-ended data-driven exploration: To answer specific questions, such as computing the factors that affect the life expectancy of the county's residents, we used traditional data analytics techniques, borrowed from the machine learning literature. To allow more open-ended discoveries we implemented a visual data exploration platform, that allows public health researchers to visually explore the data and their correlations.

Data Source	Indicator Count
Subregional area (SRA)-level dataset	**3,818**
HHSA Behavioral Health Data (Hospitalizations & Emergency Department visits for behavioral health conditions)	1,170
HHSA Demographics (Demographics)	300
ESRI Market Potential Data (Consumer buying patterns and behaviors)	2,234
SANDAG Healthy Communities Atlas (Data on physical and built environment)	114
Census-tract level dataset	**22,712**
American Community Survey 2012 (5-Year Estimates) (Census demographics)	22,547
CalEnviroScreen 2.0 (Pollution data)	45
Life Expectancy Data	6
SANDAG Healthy Communities Atlas (Data on physical and built environment)	114

Figure 3.30 Contents of the two integrated datasets used in the case study (Katsis et al., 2017)

Example 7 : A Big Data Analytics Platform for Smart Factories in Small and Medium-Sized Manufacturing Enterprises: An Empirical Case Study of a Die Casting Factory Big Data Techniques for Public Health: A Case Study (Lee et al., 2017).

Although manufacturing is a traditional industry, its ability to compete with the service, finance, and information technology (IT) industries is lacking, as it is not considered novel or innovative. Recently, however, countries globally have increased their focus on the manufacturing industry due to the possibility for job creation and economic growth. These efforts take the form of various innovative manufacturing strategies that are specific to the needs of each country. Germany, a leader in manufacturing, is reinforcing the competitiveness of its manufacturing industry by implementing a smart factory program, which incorporates information and communications technology (ICT) with manufacturing policies, such as High-Tech Strategy 2020 and Industry 4.0. The European Union has made innovations in manufacturing through the Factories of the Future initiative, a public and private partnership based on research and development (R&D) activities. These activities help European manufacturers compete internationally by supporting the development of key technologies in European factories. The United States has launched the National Strategic Plan for Advanced Manufacturing to strengthen its manufacturing competitiveness. The Smart Manufacturing Leadership Coalition, a non-profit organization, has made efforts to develop and deploy smart manufacturing systems by implementing an agenda for building a scaled, shared infrastructure called the Smart Manufacturing Platform. Japan, a new and powerful figure in manufacturing, has launched the Industry Revitalization Plan to reinforce its manufacturing competitiveness by applying ICT to manufacturing, especially the robot industry. Efforts made by leading countries in the field of manufacturing vary in terms of specifics, but they follow the same general principle. Following the Internet Revolution that has occurred over the past 20 years, countries are now fusing developed IT and Internet business concepts with the manufacturing industry to enhance the value and competitiveness of the manufacturing industry. These efforts, collectively known as the Fourth Industrial Revolution, are intended to make significant innovations in manufacturing.

The world's leading manufacturing countries have attempted to make their manufacturing industries more competitive. South Korea has launched the Manufacturing Innovation 3.0 initiative to implement smart factories in small and medium-sized manufacturing enterprises (SMMEs) in the manufacturing industry. Since SMMEs are significant in terms of national competitiveness (accounting for 99.5% of the value chain) and they do not have the capability to construct a smart factory, the government of South Korea has developed a strategy for SMMEs to construct smart factories. In terms of an action plan, eight core smart manufacturing technologies, which are crucial to each component of a smart factory, have been determined. Next Figure identifies these core smart manufacturing technologies, which include the internet of things (IoT), smart sensors, holograms, three-dimensional (3D) printing, energy saving, cyber-physical systems (CPS), big data, and cloud computing. These core technologies in smart manufacturing also have three applications. Smart sensors, IoT, and 3D printing may be used to collect data and produce test products in the device/network area. CPS, energy saving, and holograms are platform technologies that help a manufacturing factory operate and allow the factory to be flexible when dealing with various problems by synchronizing an actual manufacturing system with its virtual system. Finally, the manufacturing industry can extend to smart service areas by providing service applications based on big data and cloud computing technologies.

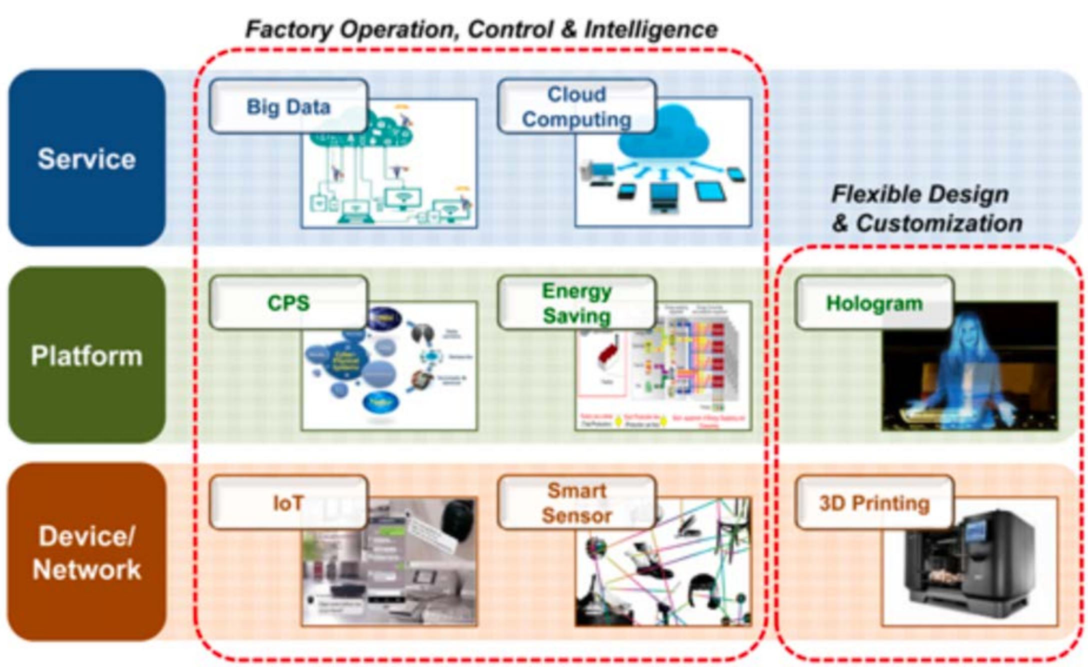

Figure 3.31 : The eight core smart manufacturing technologies in South (Lee et al., 2017)

Empirical Case Study:

Applying the Big Data Analytics Platform to a Die Casting Factory
The proposed architecture and system modules of the big data analytics platform were applied to a die casting company in South Korea. The company has annual sales of about $20 million and employs 100 individuals. There are ten die casting machines on the shop floor, and the company casts automobile parts and electronic components in aluminum, magnesium, and zinc. This section presents the developed platform and its application scenario, according to the big data analytics process consisting of data collection, preprocessing, analysis, post processing (i.e., interpretation and visualization), and application. In particular, the application scenario focuses on identifying defects in the die casting process using big data analytics. The following aspects are analyzed :

1. The die casting process
2. Data collection and preprocessing
3. Data analysis
4. Interpretation, visualization, and application
5. Difficulties and challenges

Conclusions:

Lee proposed a big data analytics platform to implement smart factories in SMMEs. The proposed platform was evaluated by applying it to an actual factory of a die casting company in South Korea.

The architecture and system modules of the developed big data analytics platform may apply to other fields in manufacturing, as well as other die casting companies. The paper also introduced application scenario through big data analytics to identify the defects in the die casting process. After parameters and inspection results were collected from the die casting process, the data was cleaned and preprocessed. In addition, correlation analysis algorithms were applied; other algorithms, such as the neural network, are currently under development. The results were visually displayed. Finally, the legacy system used the analyzed results for facility maintenance and applied the results to the CAE simulation.

The proposed platform addresses three issues, which the paper has clarified by investigating existing literature, on applying big data analytics to manufacturing, especially SMMEs. First, the proposed platform shows the integrated system environment between the legacy system and the big data analytics platform. Second, the platform includes analytical models to address quality issues that are the first priority for SMMEs. Finally, this study presents the efficient system architecture based on localization and cloud computing that enables SMMEs to reduce their financial burdens of infrastructure and experts for the big data analytics platform. On the other hand, this empirical research has revealed some challenges in using the big data analytics platform in SMMEs. Regarding collecting data, it is difficult to extract process parameters from the outdated machines since the machine vendors usually prohibit users from interfacing with their machines. Mapping the process parameters to the defects is also problematic because many companies manage their products in lot units. Although these challenges might recur when applying the platform to other SMMEs, the platform is able to cope with such issues based on experiential knowledge and technical know-how from this study.

Example 8 : Are Software Analytics Efforts Worthwhile for Small Companies? The Case of Amisoft (Robbes et al., 2013).

Microsoft has a search group dedicated to empirical software engineering1 and Google employs at least 100 engineers to improve its analytics-based tools (www.infoq.com/ presentations/Development-at-Google). Software analytics has been widely accepted in the large enterprise sector. However, most companies are not able to invest as much in software analytics because the vast majority of them are small. According to ItaRichardson and Christiane Gresse von Wangenheim, 85% of software companies have fewer than 50 employees2; in Brazil, 70% have fewer than 20 employees3; in Canada, 78% have fewer than 25 employees4; and in the United States, approximately 94% have fewer than 50 employees5. Are software analytics viable for small software companies that are not able to exploit economies of scale, have less spare labor, and have less historical information in their software repositories than companies dealing with large software systems such as Google or Microsoft? We decided to explore this question in a small company called Amisoft by conducting interviews (see sidebar "Note on methodology").

Amisoft is a 15-year-old software company based in Santiago, Chile. Its main activity is custom software development and maintenance of existing systems. Amisoft is also starting to develop standard products to complete its service offering. The company averages two new development projects per year; however, its seven definitive maintenance contracts are the projects that provide financial stability. Amisoft has 43 employees; 40 work directly in software maintenance and development. Each employee performs more than one of the company's traditional software engineering functions (developer, analyst, tester, etc.).
Case study: Increasing Reactivity to reduce Work Overload

One characteristic of our data collection process is that most of the metrics are updated weekly. Project managers have used analytics to react to delays (for instance, by rescheduling) and get back on track quickly rather than let- ting delays accumulate; increased effort is punctual rather than sustained.
Given the absence of hard data for the period before the analytics were introduced at Amisoft, we have to rely on anecdotal evidence. Based on the CEO’s experience, the situation at Amisoft (once the improved process was introduced) was that most projects were delivered on time but had very high cost in staff-hours and required sustained effort later in the project. Today, the effort is much more evenly dis- tributed but achieves the same results.
To evaluate the reduction in sustained late efforts and the associated burnout, we analyzed the evolution of the CPIs and SPIs of individual iterations to locate rapid adjustments to trends. Iterations usually last between three and six weeks, so weekly metric updates let the team adjust its workload accordingly. We analyzed the data from 29 iterations of five projects and classified each of the resulting 58 metric trends in three categories (see next Figure).

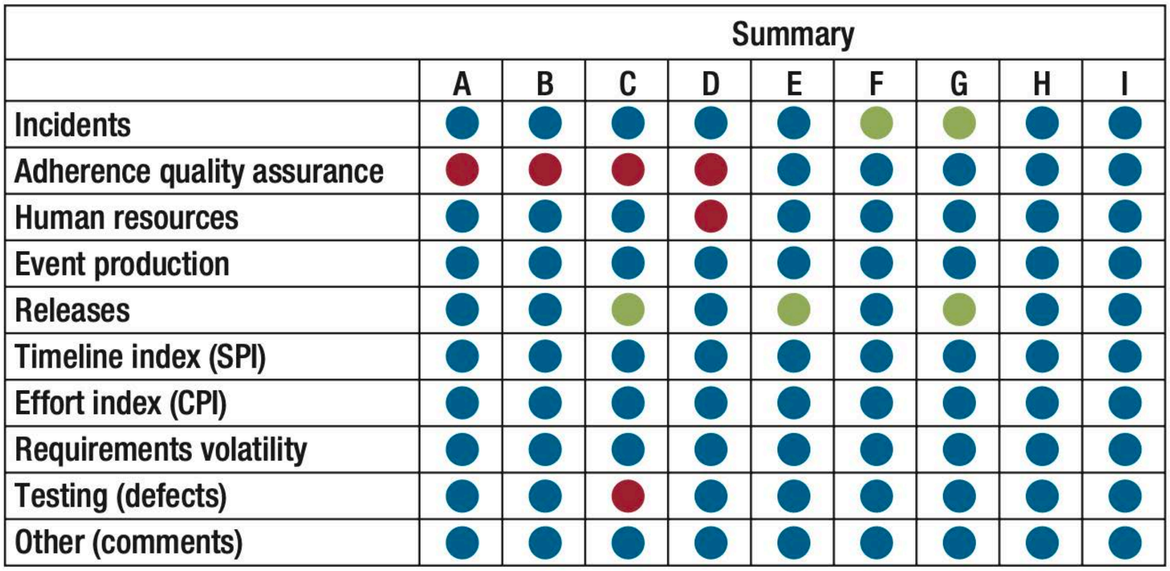

Figure 3.32 High-level status of projects at Amisoft. From this view, project managers and general managers can drill down and inspect particular metrics and their evolutions, reacting to deviations from set objectives. (Robbes et al., 2013)

Furthermore, we looked at the CPI and SPI values at the end of each iteration to determine whether the stated goal of 0.8 or above was reached. This occurred 81 percent of the time; 66 percent of the time, it was above 0.9. This shows that projects react quickly to delays during an iteration. Before Amisoft implemented analytics, delays would often go unnoticed until much later in the iterations, at which point they could have grown to be as large as 50 percent. This would cause considerable risks to the projects, including burnout of employees working long hours or significant delays if a critical employee fell sick at the wrong time. By monitoring the status more often, these situations are much rarer.

Software analytics are worthwhile if you follow a process. The main lesson we extracted from this experience is that software analytics are definitely worthwhile, even for a small company like Amisoft. They bring visibility and predictability to the software development process and allow companies to gather evidence in support of a wide range of decisions, from decisions too small to be recorded to long-term changes in company strategies. But data analysis practices lack maturity. Such practices need to be formalized and shared: each project manager used the metrics in a different way. With additional experience and practice sharing, we expect patterns of data analysis to emerge and be consistently adopted by managers. The discovery and consolidation of said patterns should be data analysts' responsibility.

Example 9 : Intelligent decision-making of online shopping behavior based on internet of things (Fu et al., 2020).

The development of big data and Internet of things (IoT) have brought big changes to e-commerce. Different kinds of information sources have improved the consumers' online shopping performance and make it possible to realize the business intelligence. Grip force and eye-tracking sensors are applied to consumers' online reviews search behavior by relating them to the research approaches in IoT. To begin with, public cognition of human contact degrees of recycled water reuses with grip force test was measured. According to the human contact degrees, 9 recycled water reuses presented by the experiment are classified into 4 categories. Based on the conclusion drawn from grip force test, purified recycled water and fresh vegetable irrigated with recycled water are regarded as the drinking for high-level human contact degree and the irrigation of food crops for low-level human contact degree respectively. Several pictures are designed for eye-tracking test by simulating an on-line shopping web page on Taobao (the most popular online shopping platform in China). By comparing the fixation time participants spent on the areas of interest (AOIs), we justify that consumers' online reviews search behavior is substantially affected by human contact degrees of recycled products. It was found that consumers rely on safety perception reviews when buying high contact goods.

This research refers to the research strategies in IoT field and use grip force and eye-tracking sensors to detect and capture consumers' online information search behaviors (Cerchecci et al., 2018; Giudice, 2016). To measure human contact degrees of recycled water reuses in a more accurate way, grip force test has been used to ensure consumers' cognition of human contact degrees of recycled water reuses (Thumser, Slifkin, Beckler, & Marasco, 2018). According to the result of grip force test, recycled water reuses are divided into groups, among which two uses that are notably divisive in human contact degrees have been chosen as the research materials for the subsequent eye-tracking ex- periment. In eye-tracking experiment, the eye-tracking sensor has been used to capture eye movement track of participants when browsing online shopping interface for recycled water products through imitating real online shopping situation. As a result, influences of human contact degrees to different products produced by recycled water has been found.

Study 1: grip force test

1.1. Participants

30 and 57 22-year-old college students in good health condition with normal vision or corrected visual acuity in China, have been chosen to participate in the pre-experiment and formal experiment, respectively (Thumser et al., 2018; Ziauddeen et al., 2012). College students have been chosen because they are believed to be capable cognitively, and familiar with online shopping. Besides, they seem to be more willing to participate in the experiment.

1.2. Experiment procedure

Before the experiment commenced officially, 30 college students were invited to the pre-experiment. basic operation of grip force meter is introduced to participants, and then they are required to grip the grip force meter with their maximal muscle force every 15 s with orders from instructors. Each participant should grip the grip force meter with their maximal muscle force for 25 times while the instructor recorded the data of grip force meter. Through experimental data and the oral reports of the participants, we found that the participants generally experienced a significant grip slip and felt fatigue after 15 grips. To avoid experimental deviation caused by the decline in grip strength, the total trails number of the official experiment was controlled within 15 times. With the data of pre-experimental grip, it was observed that the first three grip strength data fluctuate greatly, therefore, in the formal experiment, the first three grips are regarded as an experience to help participants using grips and leaving no data logging.

Study 2: eye-tracking test

2.1. Participants

155 healthy college students from Xi'an University of Architecture and Technology and Xi'an University of Science and Technology were recruited to join our eye-tracking experiment. Among them, 14 participants were in the pre-experiment, and 141 participants were in the official experiment. According to the 2016 CNNIC report, people aged 18–25 account for 56% of online consumers in China. They are the main consumers of online consumption. Therefore, the survey results of college students are valid to reveal the behavior and psychological characteristics of the general online consumers.

2.2. Materials

In this experiment, purified recycled water was selected as the item of high degree of contact, and the fresh vegetable that are irrigated with recycled water was selected for the low contact degree. To observe and record the online search behavior of the participants to comments on recycled water products, stimulating pictures for eye movement tests were provided according to the product webpages of Taobao (the most popular online shopping platform in China). This was to enhance the immersion and experience of the experiment participants who were also ensured sufficient time to view the interface information of each recycled water product in the experiment. The eye tracking sensor used in this study is a non-intrusive, long-distance measurement (PCCR) pupil center cornea reflection technology. Invisible infrared light source is used to illuminate the eyes to produce obvious reflection. Then the image sensor is used to collect the reflection image produced by infrared light source on the cornea and pupil of the user's eyes and calculate the position of the eyes in space and the direction of the line of sight, so as to realize the tracking of eye movement trajectory. On this basis, the man-machine interaction in the online shopping process is realized. Besides, eye movement trajectories of participants would be traced by the eye tracking sensor during the test.

Conclusions

In order to find out the different types of information that consumers have for different human contact products produced by recycled water, firstly, different kinds of recycled water are classified according to the degree of human contact. The two groups with significant difference in human contact degrees were selected as experimental materials for simulating online shopping experiments. The eye-tracking sensor was used to record the reviews search behavior of the participants when browsing the online shopping interface, and by comparing the fixation time of AOIs represented by different types of reviews for consumers, this paper verified the impact of human contact degrees of recycled products on consumers' online reviews search behavior. Conclusions are as follows:

First, the public cognition of human contact degrees of different kinds of recycled water reuse is accurately measured. Based on the results of grip force test, the contact degree of the 9 recycled water reused in the experiment with the human contact degree was accurately measured. The 9 recycled water reused were ordered according to human contact degrees by comparing the normalized grip force test data. Through the paired t-test, the nine types of recycled water reused were divided into four groups with significant difference in human contact degrees.

Second, verifying that the degree of contact has a significant impact on the level of consumer attention to online content. Eye movement experiments can be used to confirm a significant interaction between the degree of contact and the content of the review, meaning that the degree of contact can influence consumers' attention to online reviews of recycled water products. As the degree of contact increases, consumers begin to become more concerned about the safety perception reviews.

Third, consumers rely more on safety perception reviews to make purchasing decisions when purchasing high contact degree items. Eye movement experiments have confirmed that consumers are more concerned about the safety perception reviews on web pages when purchasing high contact degree products. For low contact degree products, consumers have no significant difference in gaze duration for safe perception reviews and price perception reviews on web pages, and attention to safety perception reviews was only slightly higher than the price/performance reviews.

Example 10 : Intelligent decision-making of online shopping behavior based on internet of things (Yan et al., 2020).

With the rapid development of artificial intelligence technology and network technology, the internet of things has gradually become the mainstream of social development in the future. Under this background, the trade retail industry needs to establish its customer relationship network in combination with artificial intelligence technology. At the same time, it needs to conduct law mining in combination with customer selection behavior in network and carry out personalized excavation of customers under the support of data mining technology to help customers make decisions. On this basis, it can effectively enhance the customer experience. The research on intelligent customer network has entered a climax since 2010, and related research also provides the basis for the creation of this article.

The intelligent customer relationship network usually uses the customer's equipment's movement trajectory data, customer platform operating data, customer network base stations, and other content as customer behavior data. Using this data, researchers started relevant research. Long et al. achieved a goal of recommending to the taxi driver the passenger sequence with the greatest revenue through in-depth analysis and excavation of GPS trajectory data in taxis. Mariscal et al. designed and implemented a time- awareness system that can be used to personalize the taxi drivers travel route with the greatest benefit per unit of time. Based on the different advertising platforms, Purtova et al. proposed an advertisement delivery system TMAS that is suitable for mobile web pages and mobile phone apps by analyzing customer location and related situational information and fully exploiting the mobility of customers in the mobile commerce system. Saponara et al. designed a personalized travel package recommendation system based on tourist interest preferences, which can recommend a set of personalized and best-suited attraction collections for tourists Kroeckel et al. studied and analyzed the mobile customer's check-in data to obtain various features of the location social network; based on this, a location-based recommendation algorithm was designed and implemented. With the progress of research, many personalized re- commendation systems for mobile clients have also been successfully launched, such as the Facebook mobile application of personalized push ads, the personalized Bizzy recommended by local shops, and the personalized reading system Zite.

Long proposed a detection method for mobile App ranking fraud by exploring personalized preferences mining method for mobile customers based on context awareness Long discussed security privacy issues under personalized recommendation technology, and. he proposed a mobile App recommendation algorithm to protect customer information security against this issue. Feng, based on statistical analysis of many microblog customer data, proposes a method for personalizing popular micro topics by calculating similarities between microblog customers and micro topic. In addition, in terms of data sparsity and cold-start problems faced by collaborative filtering, Bedi et al. proposed the use of the K-nearest neighbor method to map “attribute-feature” and calculate the feature vectors of new customers and new projects. Islam proposed using a combination of data migration and data clustering to solve the system could start problem. To solve the problem of sparseness in collaborative filtering algorithms, Zuech proposes a way of thinking that the clustering is based on the attributes of the project and uses the mean of the project categories to fill in the null values in the original scoring data. At present, major e-commerce platforms at home and abroad have developed their own mobile terminals. However, the re- search and application of personalized recommendation systems for mobile platforms is still in its infancy and there is still room for improvement in its recommendation quality and operating efficiency.

Ravizza first proposed the idea of considering the trust between customers in the recommendation process. The trust between customers is established through the displayed customer trust evaluation and debilitating spread. The trust is divided into reliability trust and decision trust. The reliability trust is the subjective probability that entity A acts according to entity B's expectations, and decision trust refers to the subjective degree of relative security feeling obtained by an individual trusting a certain entity in a certain environment. Watters uses the ratio of the number of customer recommendations to the total number of recommendations as the degree of trust between customers and applies this calculation method to the recommendation system, where the confidence value ranges from [0,1]. Saponara et al. proposed a trust model based on fuzzy logic representation based on the fuzzy nature of trust relationships.

Benefiting from the development of Internet of things technology and data mining technology, the spread of consumer trust has become multi-directional. As Kim and Park mentioned, all the characteristics of s-commerce (except for economic feasibility) had significant effects on trust and that trust had significant effects on purchase intentions. Hence, the characteristics of consumer trust communication and behavior decision-making under the Internet of things are necessary to study.

Based on the above analysis, we can see that the current decision model based on the Internet of Things to build a customer relationship network is less researched, and most of them are recommending unilateral information to customers based on personalized recommendations. Therefore, based on the Internet of Things technology, this study builds a more complete customer relationship network based on personalized recommendations, and adopts an improved collaborative filtering recommendation algorithm as a basis for decision models to extract contextual features that characterize customer trust. At the same time, this research uses the analytic hierarchy process to complete the model building process, helps customer relationship network service objects to provide decision support, completes product information recommendation, solves new customer cold start problems, and improves existing scoring prediction formulas. Therefore, this study fully considers the impact of customer trust.

.

Table 3.14 Comparative table of Big Data characteristics in 10 the examples

Case	Characteristics				
	Volume	Velocity	Variety	Veracity	Value
1. Big Data at UPS (Davenport et al., 2013)	16 petabytes	16.3 million new packs daily	High, data on packages, customers, requests, maps, vehicles and sensors	Storage of your own data, generated by your processes or actions, your sensors or modules.	High, UPS estimates that saving only one daily mile driven per driver saves the company $30 million,
2. Big Data at an International Financial Services Firm (Davenport et al., 2013)	60 terabytes	Higt, millions of daily transactions for dollar-for-dollar calculations	Structured	50 server nodes and 800 processor cores, capable of handling a petabyte of data	High, a big data infrastructure to exploit faster processing power
3. Facilitating maintenance decisions on the Dutch railways using big data: The ABA case study (Núñez et al., 2014)	100 terabytes accumulation day by day	Higt, 100 terabyte a day	Different data-collecting systems are used, which leads into a wide variety of available data	The quality of each data source and the reliability of the conclusions drawn may differ	High. Social aspects, such as the reduction of delays and the optimal use of roads and the availability of public transport services
4. Fog Based Intelligent Transportation Big Data Analytics in The Internet of Vehicles Environment: Motivations, Architecture, Challenges, and Critical Issues (Darwish & Bakar, 2018)	Exponential growth. For example, in 2013 around 480 TB of data and an increment to reach 11.1 PB/year is expected by 2020	The velocity of ITS data, which varies widely	The organization level of such data varies from semi-structured to structured data	The ITS community is facing significant challenges in providing timely and reliable transportation related data collection	High. Route planning applications can benefit from non-real-time data

5. Integrated Understanding of Big Data, Big Data Analysis, and Business Intelligence: A Case Study of Logistics (Jin & Kim 2018)	This is rapidly increasing the volume and sales of courier services as more consumers shop online. PB	Generating millions of daily purchases, vehicle and customer data	Data generated from the internet, sensors, robots, traditional structured system.	the veracity of the average data due to the variable way in which the data is collected	High. Improving efficient decision making based on business intelligence (BI)
6. Big Data Techniques for Public Health: A Case Study (Katsis et al., 2017)	Data from only 26,530 indicators	The generation of these data is slow due to the fact that they are indicators recorded for years, for example, during 5 years a total of 22547 were generated.	Structured, diferentes Datasets	High, due to the source of the product	High. Big data was effectively used to analyze thousands of health-related variables to gain interesting insights into the determinants of various health outcomes.
7. A Big Data Analytics Platform for Smart Factories in Small and Medium- Sized Manufacturing Enterprises: An Empirical Case Study of a Die Casting Factory Big Data Techniques for Public Health: A Case Study (Lee et al., 2017)	75 parameters are collected from the die casting process using sensors, devices, and control systems, GB	All data from the machines can be transferred to the smart middleware through the smart I/F devices	Semi-structured data (e.g., text, images)	Due to the characteristics of manufacturing data, the platform should include some preprocessing algorithms	Analytical models to address quality issues that are the first priority for SMMEs

8. Are Software Analytics Efforts Worthwhile for Small Companies? The Case of Amisoft (Robbes et al., 2013)	The data from 29 iterations of five projects and clas-sified each of the resulting 58 metric trends	Data of the processes captured weekly (less than 100 weekly records)	Structured	Given the absence of hard data for the period before the analytics were introduced at Amisoft, we have to rely on anecdotal evidence	To evaluate the reduction of late efforts and associated attrition. To locate rapid trend adjustments.
9. Intelligent decision-making of online shopping behavior based on internet of things (Fu et al., 2020)	Case 1 with data from 57 students and Case 2 with 155 students.	Data obtained for the pre-experiment and later for the experiment with the sensors	Grip force and eye-tracking sensors are used to analyze consumers' online reviews search behavior.	The eye-tracking sensor has been used to capture eye movement track of participants when browsing online shopping interface	Verified the impact of human contact degrees of recycled products on consumers' online reviews search behavior
10. Intelligent decision-making of online shopping behavior based on internet of things (Yan et al., 2020)	298 customers' click browsing records as training data, and collected 50 customers who used the platform for the first time as research objects	Data captured at the beginning of the experiment	The customer's equipment's movement trajectory data, customer platform operating data, customer network base stations, and other content as customer behavior data.	Data may vary due to user behavior and the way in which they are obtained	Customer's consumer experience can be enhanced with the support of data mining technology in cyber intelligence

Looking at these 10 examples we can conclude that topics such as Big Data, Analytics, Data Mining, or decision making, can be carried out in any type of company, including infrastructure or quantity and types of data. This can be observed in the different cases that we use to express the differences between the projects of large companies where most of the V's of Big Data are fulfilled as the amount of storage that cannot be processed in a simple way or even generation of TB of data daily, but in the same way you can see the advantages of small companies where having little data but these are more structured or select the veracity of these data are more efficient. But as a conclusion of the table, it can be distinguished that even though the V's are not fulfilled in all cases, it is fulfilled having a value.

This is due to the fact that according to the theory, in order to apply to Big Data projects, it is necessary to have an amount of data over Terabytes, an amount that is not possible to process with the resources of a standard organization due to the traditional way of processing, But as Adibuzzaman mentions, in the health area there are not always millions and millions of data which even when being analyzed from thousands of records that can be had on the subject according to the requirements of the research or the limitation of public data become even just a few tens of data to analyze, but this does not mean that the study or the results have no relevance (Adibuzzaman, et al., 2017)

Even Gartner publishes " Top 10 Data and Analytics Trends for 2021 " where Trend 4 is from big to small and wide data, just where he mentions that " Small and wide data, as opposed to big data, solves several problems for organizations facing increasingly complex questions about AI and challenges with sparse data use cases. Big data - leveraging "X-analytics" techniques - enables the analysis and synergy of a variety of small and varied (big), unstructured and structured data sources to improve contextual knowledge and decisions. Small data, as the name implies, is capable of using data models that require less data but still provide useful insights." (Gartner ,2021)

3.1.3.4 Review of Open-Source Development Platforms for Big Data Systems

There is a wide range of systems and tools that are used for the development of Data Science / Analytics systems, the Data Science / Analytics community is in general quite open and generous, which means that many of the tools and libraries are Open-Source.

This indicates that there are many programming languages that allow us to develop in Data Science / Analytics, a study by Kdnuggets shows the most popular languages for the development of Data Science / Analytics projects in the industry. As we can see in Table 3.14 Programming languages for Data Science / Analytics, Python and R are the two most used languages with a wide advantage over the others.

Table 3.15 Programming languages for Data Science / Analytics (Kdnuggets, 2019).

Platform	2019 % share	2018 % share	% change
Python	65.8%	65.6%	0.2%
R Language	46.6%	48.5%	-4.0%
SQL Language	32.8%	39.6%	-17.2%
Java	12.4%	15.1%	-17.7%
Unix shell/awk	7.9%	9.2%	-13.4%
C/C++	7.1%	6.8%	3.7%
JavaScript	6.8%	na	na
Other programming and data languages	5.7%	6.9%	-17.1%
Scala	3.5%	5.9%	-41.0%
Julia	1.7%	0.7%	150.4%
Perl	1.3%	1.0%	25.2%
Lisp	0.4%	0.3%	46.1%

That is why for this thesis we will analyze three of the most widely used languages in the world Python, R and Java, which we will analyze with different criteria that allow us to select one of the languages to be used in this thesis. Below is a brief description of each of these programming languages focused on Data Science / Analytics developments, as well as the tools and libraries that each of them would use.

Python

Python is a general-purpose object-oriented programming language due to its extensive library that primarily enables the development of Big Data, Artificial Intelligence (AI), Data Science, Test Frameworks, and Web Development applications. Released in 1989, Python is easy to learn and a favorite with programmers and developers. Python is one of the most popular programming languages in the world, second only to Java and C (IBM, 2021).

There are several libraries and tools that allow us to carry out tasks and Data Science / Analytics developments for this specific thesis, we will consider 4 of the most important tools and libraries that exist for the development of Data Science / Analytics in Python, these are the following:

- Jupyter a web-based iterative development environment for notebooks.
- Numpy to handle large matrices.
- Pandas for data manipulation and analysis.
- Matplotlib to create data visualizations.

Also, Python is especially well suited for implementing machine learning on a large scale. Its suite of specialized libraries enables data scientists to develop sophisticated data models that connect directly to a production system.

R Lenguage

R is an Open-Source programming language that is optimized for statistical analysis and data visualization. Developed in 1992, R has a rich ecosystem with complex data models and elegant data reporting tools (IBM, 2021).

The interface and structure are very suitable for tasks related to algorithms and data modeling, R has hundreds of libraries, this has made it one of the most developed systems that has thousands of packages to solve a wide variety of problems.

Popular among Data Science / Analytics academics and researchers, R provides a wide variety of libraries and tools for creating Data Science / Analytics tasks, for this thesis we will focus on three main tools for this task, these tools and libraries are:

- RStrudio, is an integrated development environment for simplified statistical analysis, visualization, and reporting.
- Dplyr for data cleaning and preparation.
- Ggplot2 for creating visualizations.

Java

Java is an object-oriented programming language specifically designed to allow developers a continuity platform. It is an extremely popular language that runs on a virtual machine,

allowing it to be run on any type of device without having to compile it repeatedly. Java was created by Sun MicroSystems in 1991, as a programming tool and an object-oriented language, allowing programmers to generate autonomous code fragments, which interact with other objects to solve a problem offering support for different technologies.
Compared to other specific languages such as R and Python, Java does not have many libraries for advanced statistical methods, this makes languages such as R and Python much more recommended for the development of Data Science / Analytics tasks. However, there are different tools and libraries that will allow us to develop this type of application, for this thesis we will take three of the most important tools for the development of Data Science / Analytics applications, these are:

- Weka is a collection of machine learning algorithms for data mining tasks.
- Rapid Miner is a data mining tool.
- Knime is a data mining platform that allows the development of models in a visual environment.

These three languages are evaluated with the criteria and attributes proposed in the work A MADM Risk-based Evaluation-Selection Model of Free-Libre Open-Source Software Tools proposed by (Mora et al., 2016), where they propose an evaluation model based on risks of Open-Source tools. They propose 4 criteria for and 32 attributes for the evaluation of Open-Source tools, for this thesis we will take only three of these criteria and ten attributes, since these are the ones that best adapt and contain enough attributes to evaluate our three programming languages.

- **Operational Risks:** External Reviews, Internal Experience, Interested IT Staff, Project Leader, Trained End User Group, Top Management Support, Training, Usability, and User Engagement.
- **End user risks:** Functionality-quality, market image, performance-efficiency and utility-relevance.
- **Technical risks:** Community support, development process, developer community, developer organization. structure, documentation, interoperability-portability, maintainability, maturity-longevity, project fork, security-reliability, test information, compliance with standards, technical environment, and user community.

Figure 3.33 MADM risk-based evaluation-selection FLOSS tool model, shows the three criteria and the 10 attributes that will be used in this thesis, these criteria are Organizational Risks, End-user Risks and Technical Risks, with their respective attributes that were evaluated.

All these criteria and attributes were evaluated with decision-making software, which allows us to enter the alternatives, which in this case are our three programming languages, and our three evaluation criteria together with their attributes. Each of the criteria and attributes is assigned a weight based on the research carried out on each of the languages and their tools and libraries, as well as the knowledge and experience available in each one. of these programming languages. Following from Figure 3.24 to Figure 3.27 there are screenshots of the results produced by the decision-making software for our three programming languages, based on the research and experience of these.

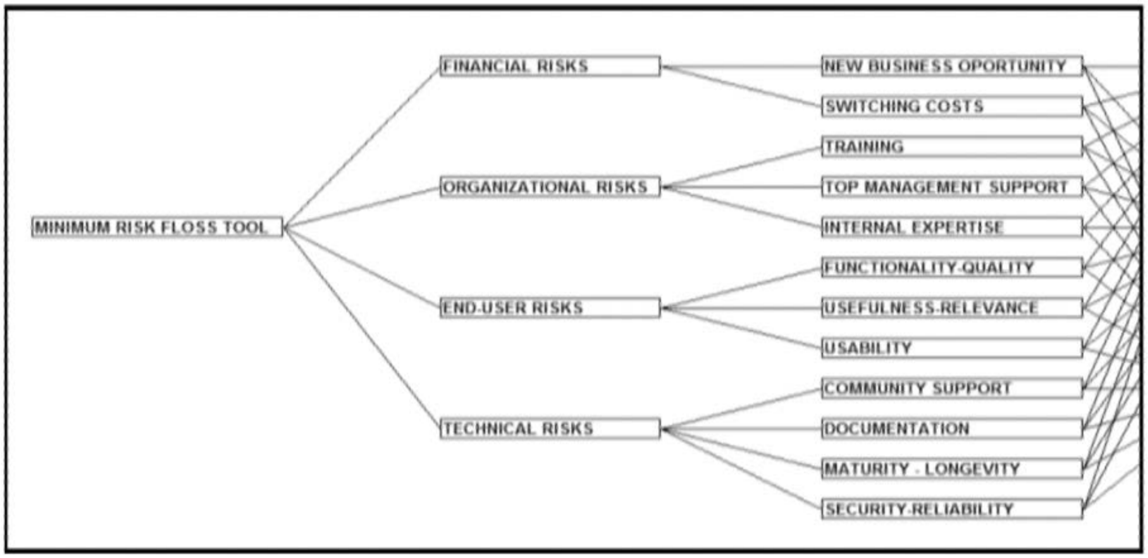

Figure 3.33 MADM risk-based evaluation-selection FLOSS tool model (Mora et al., 2016)

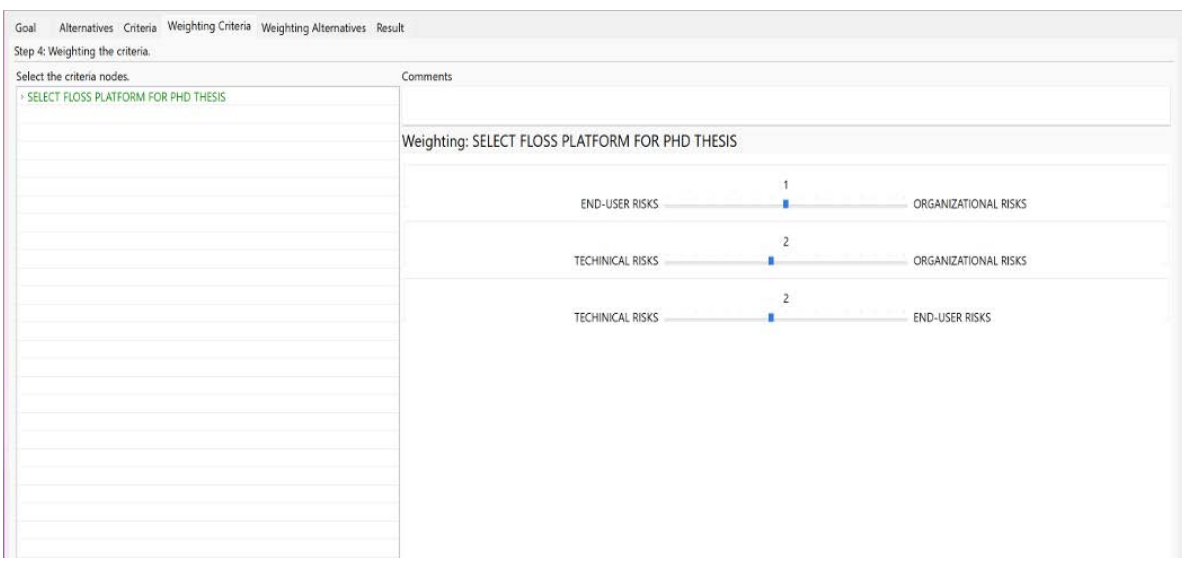

Figure 3.34 Weighting Criteria

Name	CR Value
SELECT FLOSS PLATFOR...	0.0000
ORGANIZATIONAL RISKS	0.0000
TRAINING	0.0158
TOP MANAGEMENT SUP...	0.0000
INTERNAL EXPERTISE	0.0000
END-USER RISKS	0.0000
FUNCIONALITY-QUALITY	0.0000
USEFULNESS-RELEVANCE	0.0000
USABILITY	0.0079
TECHINICAL RISKS	0.0572
COMMUNITY SUPPORT	0.0000
DOCUMENTATION	0.0000
MATURITY-LONGEVITY	0.0000
SECURITY-REALIABILITY	0.0000

Figure 3.35 Consistency Ratios

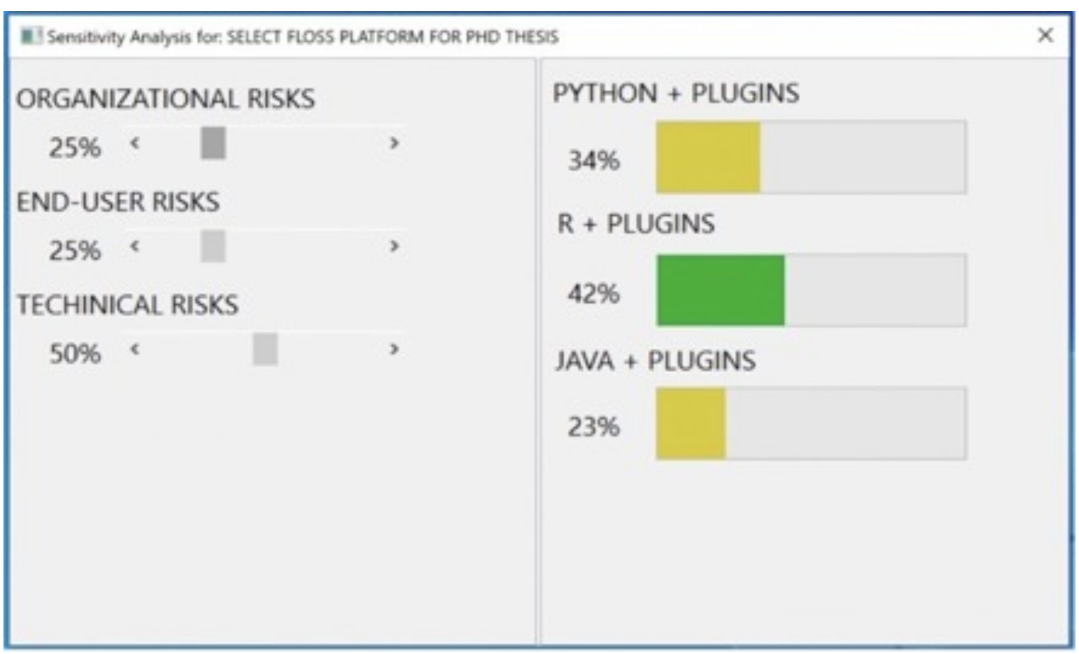

Figure 3.36 Result Ranking

As we can see, the Techinical Risks criterion was given greater weight since it is considered that the attributes it has are of greater relevance for the study of this thesis, in turn the two remaining criteria had the same weight among them.

At the same time, we can see that each of the criteria meets the consistency ratios, since all the attributes are below 0.1, this indicates that the weights assigned to each of the attributes are consistent and valid for research.

Finally, Figure 3.25 Ranking of Results shows us that when evaluating the criteria and attributes, the programming language that has the most value for this thesis is R + Plugins, this since R has a greater weight in the attributes of usability and functionality- quality, this because R is a language more focused on statistics and is much more used in research areas, in addition to being one of the most used by experts in Data Science / Analytics issues worldwide.

For computer science purists, Python always stands out as the right programming language for Data Science / Analytics. Rather, R is a specific language used for data analysis and statistics, uses a specific syntax used by statisticians, and is a vital part of the world of data science and research. On the contrary, for the design of Data Science / Analytics applications with the Java language, much less is used, since it is not a language with so many specific tools and libraries for the development of this type of applications, which gives it a clear advantage to R and Python.

The main distinction between these two languages is in their approach to data science. Both programming languages are open source and are supported by large communities, which continually expand their libraries and tools. But while R is used primarily for statistical analysis, Python provides a more general approach to data disputes (IBM, 2021).
It is for these reasons that R is the language chosen for the use of the methodology proposed in this thesis, because it is one of the most widely used languages in Data Science / Analytics issues due to its focus on statistics and data analysis, in addition to be a language created for the development of this type of project and the most used for research and data science.

3.1.3.5 Review of the 3 Main Analytics/Data Science SDM (KDD, SEMMA, CRISP-DM and extra example)

A System Development Method (SDM) is a method or technique used to develop software. It is a broad concept that includes several phases of software development, such as design, development, and testing. It is also known as system development life cycle (SDLC). An SDM defines the specific requirements and deliverables necessary for a project team to develop or optimize an application. In this segment we focus on the classic SDMs for Analytics/Data Science development, both the basis for the first methodologies and the most widely used in the area today. Efforts in data mining have focused mostly on the investigation of techniques for the exploitation of information and extraction of patterns (such as decision trees, cluster analysis and association rules). However, the fact of how to execute this process until obtaining the "new knowledge", that is, in the methodologies (Moine et al., 2011), has been deepened to a lesser extent. The methodologies allow the data mining process to be carried out in a systematic and non-trivial way. They help organizations understand the knowledge discovery process and provide guidance for planning and executing projects.

Mariscal et al. (2010) captured the state of the art of methods for data mining and knowledge discovery by comparing and adding 15 methods. The authors suggested that there are three main methodologies for the development of this type of project which are KDD, SEMMA and CRISP-DM. Furthermore, they argued that KDD (Knowledge Discovery in Databases) represents the groundwork for many other methods and is the ancestor of methods like CRISP-DM and SEMMA. Figure 3.37(Evolution of data mining process models and methodologies) shows the evolution of 14 data mining process models and methodologies. In which we can point to KDD as the initial focus and CRISP-DM as the central focus of evolution.

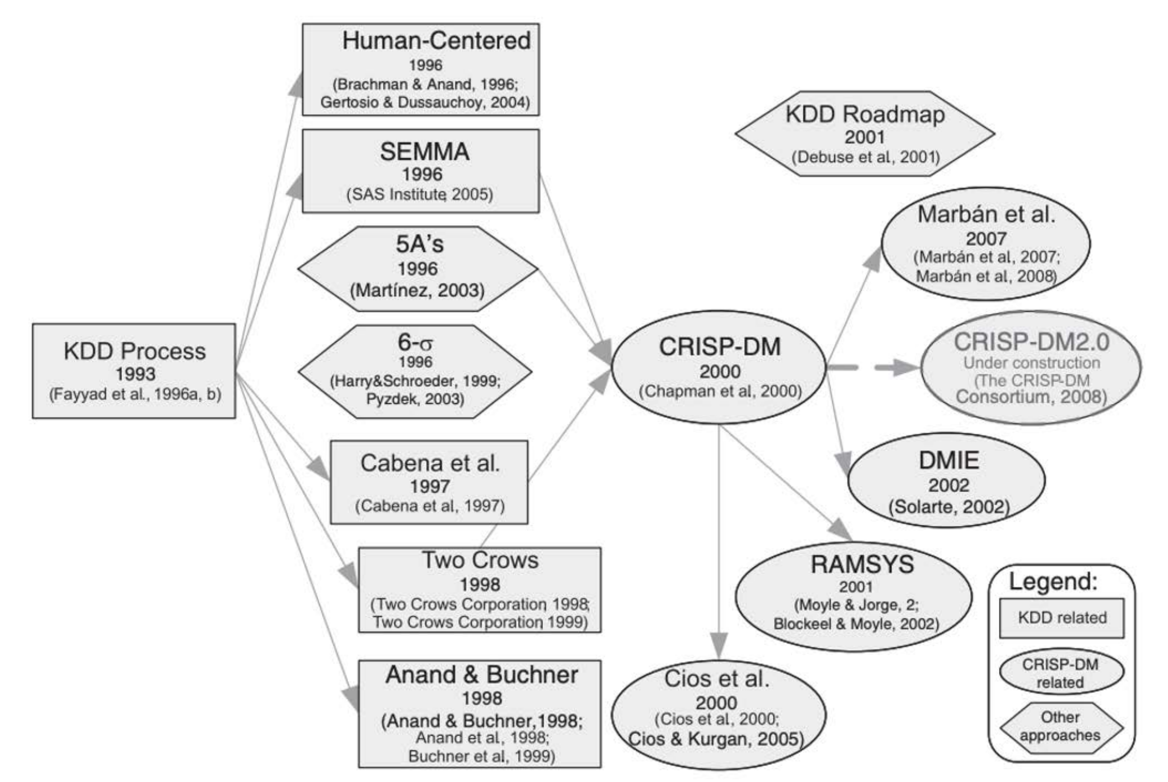

Figure 3.37 Evolution of data mining process models and methodologies (Mariscal et al., 2010).

Next, we will present these three fundamental methods, describing the phases that each of the methodologies consists of, as well as a small comparison between these three methodologies.

KDD

Data mining (DM), knowledge discovery in databases (KDD), knowledge discovery and data mining and knowledge discovery (DM and KD) are terms used to refer to research results, techniques and tools used to extract useful information from large volumes of data (Agrawal & Shafer, 1996). The whole process of information extraction is known as the KDD process (Frawley et al., 1991). Data mining is only one step in the entire KDD process (Fayyad et al., 1996).

In the early 1990s, when the term KDD was first coined (Piatetsky-Shapiro, 1991), there was a race to develop data mining algorithms that could solve all problems related to finding useful knowledge in large volumes of data. In addition to developing algorithms, some specific tools were also developed, such as: clementine, IBM Intelligent Miner, Weka and DBMiner to simplify the application of data mining algorithms and provide some support for all KDD-related activities.

KDD is the non-trivial process of finding valid, new, possibly useful, and ultimately understandable patterns in the data (Costa et al., 2020). The KDD process is an iterative and interactive process, it involves numerous steps with many decisions made by the analyst.It is essential to develop an understanding of the data, create a target data set, clean and process. Then, various tasks must be performed, such as data reduction and projection. The analyst also must match the objectives of the KDD process with a data extraction method, exploratory analysis, and a selection of models and hypotheses. An essential task is to interpret extracted patterns and use the knowledge directly (Costa et al., 2020).

KDD focuses on the general process of discovering knowledge from data, including how data is stored and accessed, how algorithms can be used for massive data sets, how they can be executed efficiently, and how interpret and visualize the results (Daderman & Rosander, 2018).

The KDD process involves numerous steps with many decisions made by the user. Brachman and Anand (1996) offer a practical vision of the KDD process, emphasizing the iterative nature of the processes, the steps that KDD consists of are described below, as well as in Figure 3.38 (An Overview of the Steps That Compose the KDD Process) it shows a general description of the steps for the process. by KDD.

1. Develop an understanding of the application domain and relevant prior knowledge and identify the goal of the KDD process from the customer's point of view.
2. **Create a target dataset:** select a dataset or focus on a subset of variables or data samples, on which discovery is to be performed.
3. **Data cleaning and pre-processing:** basic operations such as denoising if appropriate, gathering the information needed to model or account for noise, deciding strategies to handle missing data fields, accounting for time sequence information and known changes.
4. **Data reduction and projection:** Find useful features to represent mosaic data depending on the mosaic objective of the mosaic task. Use dimensionality reduction

or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

5. Match the mosaic goals of the KDD mosaic process to a particular data mining method: for example, summary, classification, regression, grouping, and more.
6. Choose the data mining algorithm (s): select the method (s) that will be used to look for patterns in the data. This includes deciding which models and parameters may be appropriate and matching a particular data mining method to the general criteria of the KDD process.
7. **Data mining:** search for patterns of interest in a particular form of representation or a set of such representations: classification rules or trees, regression, grouping, among others.
8. Interpreting extracted patterns, possibly go back to any of steps 1-7 for further iteration. This step may also involve viewing the extracted patterns / models, or viewing the data given the extracted models.
9. **Consolidate discovered knowledge:** incorporate this knowledge into another system for further action, or simply document it and report it to stakeholders. This also includes checking and resolving potential conflicts with previously believed (or extracted) knowledge.

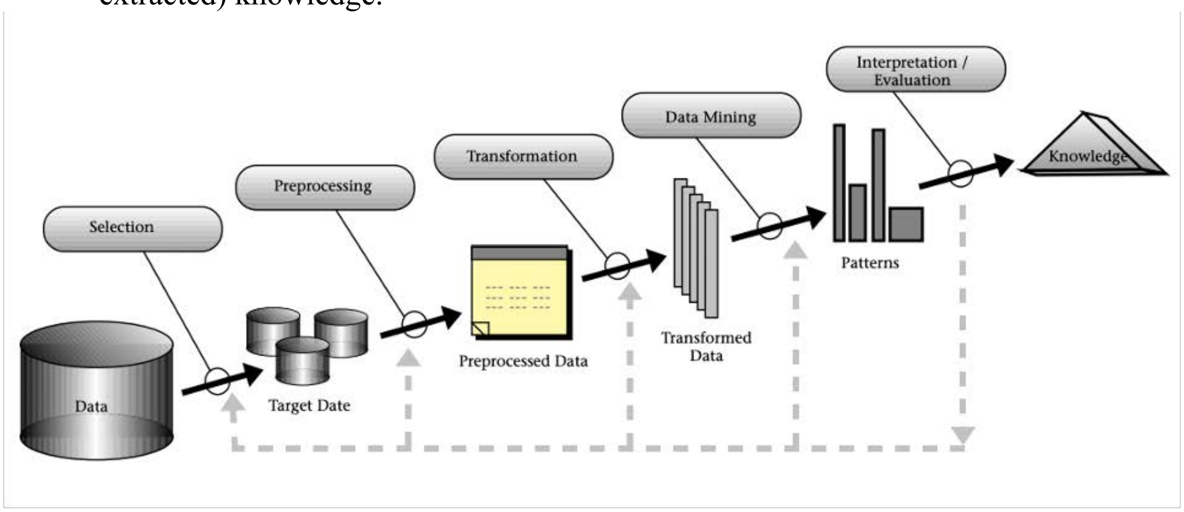

Figure 3.38 An Overview of the Steps That Compose the KDD Process (Fayyad et al., 1996).

SEMMA

SEMMA (Sample, Explore, Modify, Model and Assess) based on KDD, was developed by SAS institute in 2005 (SAS Institute Inc., 2017). And it is defined by these as a logical organization of the set of functional tools of SAS Enterprise Miner to carry out the core tasks of data mining. SAS Institute defines data mining as the process of sampling, exploring, modifying, modeling, and evaluating (SEMMA) large amounts of data to discover previously unknown patterns, which can be used to business advantage. The data mining process is applicable in a variety of industries and provides methodologies for business problems as diverse as customer churn, database marketing, market segmentation, risk analysis, affinity analysis, customer satisfaction, among others.

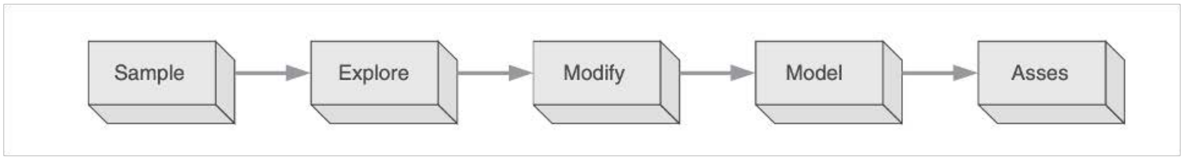

Figure 3.39 SEMMA methodology steps (Mariscal et al., 2010).

SAS Enterprise Miner software is an integrated product that provides an end-to-end business solution for data mining. A graphical user interface (GUI) provides an easy-to-use interface to the SEMMA data mining process consisting of 5 phases described below:

- **Sample:** The data by extracting and preparing a sample of data for model building using one or more data tables. Sampling includes operations that define or subset rows of data. The samples should be large enough to efficiently contain the significant information.
- **Explore:** The data by searching for anticipated relationships, unanticipated trends, and anomalies to gain understanding and ideas.
- **Modify:** The data by creating, selecting, and transforming the variables to focus the model selection process on the most valuable attributes.
- **Model:** The data by using the analytical techniques to search for a combination of the data that reliably predicts a desired outcome.
- **Assess:** The data by evaluating the usefulness and reliability of the findings from the data mining process.
-

Starting with a statistically representative sample of your data (sample), SEMMA aims to facilitate the application of visualization techniques and exploratory statistics (explore), select, and transform the most significant predictive variables (modify), model the variables to predict results (model), and finally confirm the precision of a model (evaluate) (Olson & Delen, 2008).

The SEMMA data mining process is driven by a process flow diagram, which you can modify and save. The GUI is designed in such a way that the business analyst who has little statistical expertise can navigate through the data mining methodology, while the quantitative expert can go "behind the scenes" to fine-tune and tweak the analytical process.

Enterprise Miner contains a collection of sophisticated analysis tools that have a common user-friendly interface that you can use to create and compare multiple models. Statistical tools include clustering, self-organizing maps, variable selection, trees, linear and logistic regression, and neural networking. Data preparation tools include outlier detection, variable transformations, data imputation, random sampling, and the partitioning of data sets (into train, test, and validate data sets). Advanced visualization tools enable you to quickly and easily examine large amounts of data in multidimensional histograms and to graphically compare modeling results.

The main difference between the original KDD process and SEMMA is that SEMMA is integrated into SAS tools such as Enterprise Miner and it’s unlikely to use SEMMA methodology out of them, while KDD is an open process, and it can be applied in very different environments. There is other two important differences between SEMMA and the original KDD process. On the one hand, SEMMA skips the first step of KDD process, learning the application domain, and starts directly with sample step. On the other hand, SEMMA does not include an explicit step to use the discovered knowledge, while KDD includes using discovered knowledge step. These two steps are considered essential to carry out a data mining project with success.

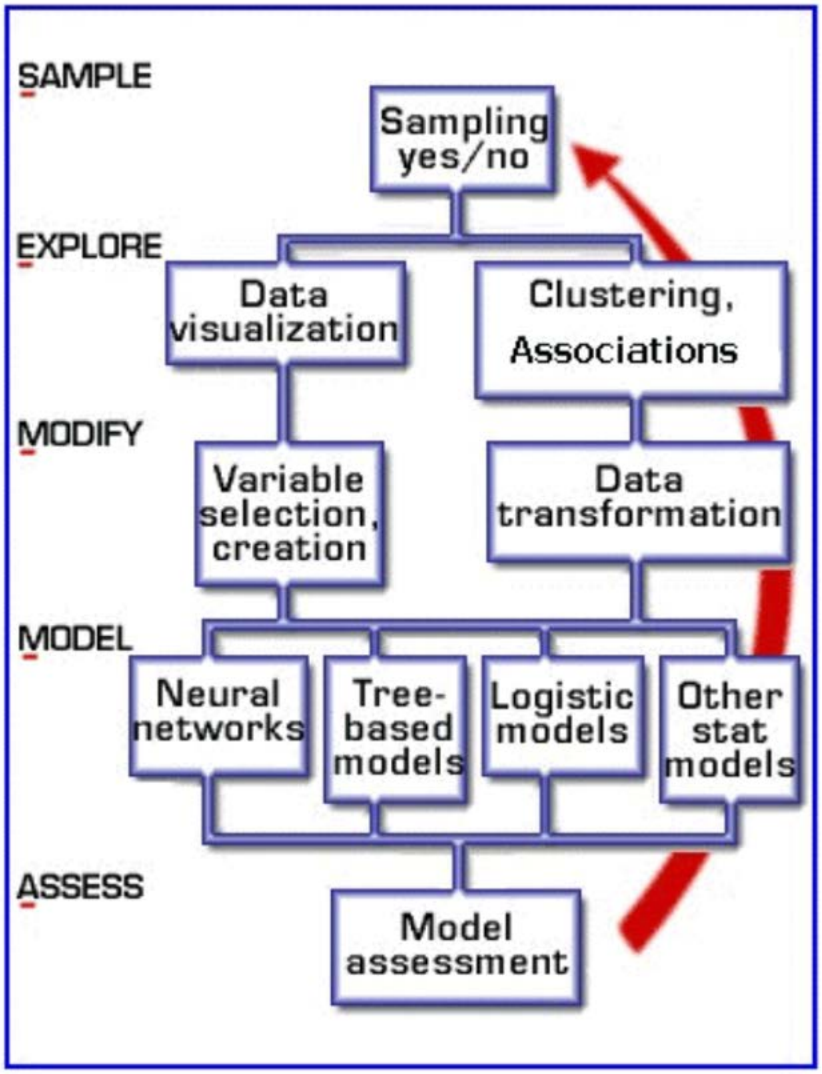

Figure 3.40 SEMMA methodology diagram (SAS Institute Inc., 2017).

CRISP-DM

In response to common issues and needs in data mining project in the mid 90's, a group of organizations involved in data mining (Teradata, SPSS -ISL-, Daimler-Chrysler and OHRA) proposed a reference guide to develop data mining projects, named CRISP-DM (CRoss

Industry Standard Process for Data Mining) (Chapman et al., 2000). CRISP-DM is considered the de facto standard for developing data mining and knowledge discovery projects. One important factor of CRISP-DM success is the fact that CRISP-DM is industry-, tool- and application- neutral.

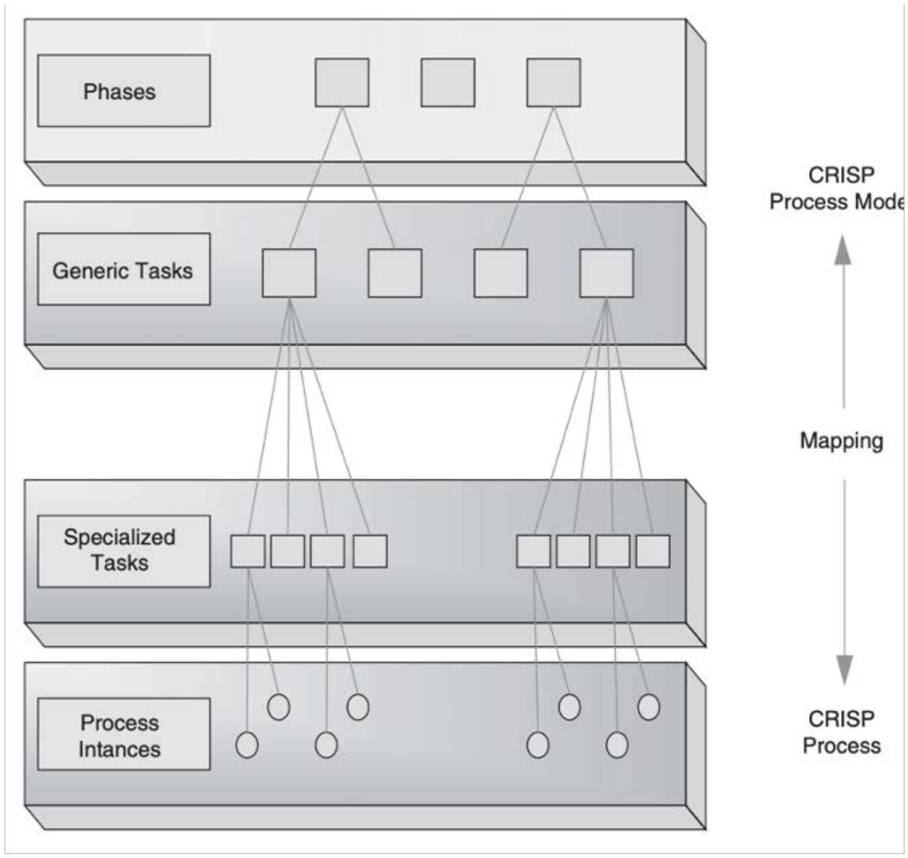

Figure 3.41 Four-level breakdown of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology (Mariscal et al., 2010).

The CRISP-DM data mining methodology is described in terms of a hierarchical process model, consisting of sets of tasks described at four levels of abstraction (from general to specific) (Figure 3.41 Four-level breakdown of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology).

At the top level, the data mining process is organized into several phases; each phase consists of several second-level generic tasks. This second level is called generic, because it is intended to be general enough to cover all possible data mining situations. The third level, the specialized task level, is the place to describe how actions in the generic tasks should be carried out in certain specific situations. The fourth level, the process instance, is a record of the actions, decisions, and results of an actual data mining engagement.

The reference model presents a quick overview of phases, tasks, and their outputs and describes what to do in a data mining project. The user guide gives more detailed tips and hints for each phase and each task within a phase and depicts how to do a data mining project.

CRISP-DM distinguishes between four different dimensions of data mining contexts:

- The application domain is the specific area in which the data mining project takes place.
- The data mining problem type describes the specific classes of objectives that the data mining project deals with.
- The technical aspect covers specific issues in data mining that describe different (technical) challenges that usually occur during data mining.
- The tool and technique dimension specifies, which data mining tool(s) and/or techniques are applied during the data mining project.

The CRISP-DM process model for data mining provides an overview of the life cycle of a data mining project. It contains the corresponding phases of a project, their respective tasks, and relationships between these tasks.

The life cycle of a data mining project according to CRISP-DM consists of six phases (Figure 3.42 Cross-Industry Standard Process for Data Mining (CRISP-DM) process model (Chapman et al., 2000)). The sequence of the phases is not strict. Moving back and forth between different phases is always required. It depends on the outcome of each phase, which phase or which task of a phase, that must be performed next. The arrows indicate the most important and frequent dependencies between phases.

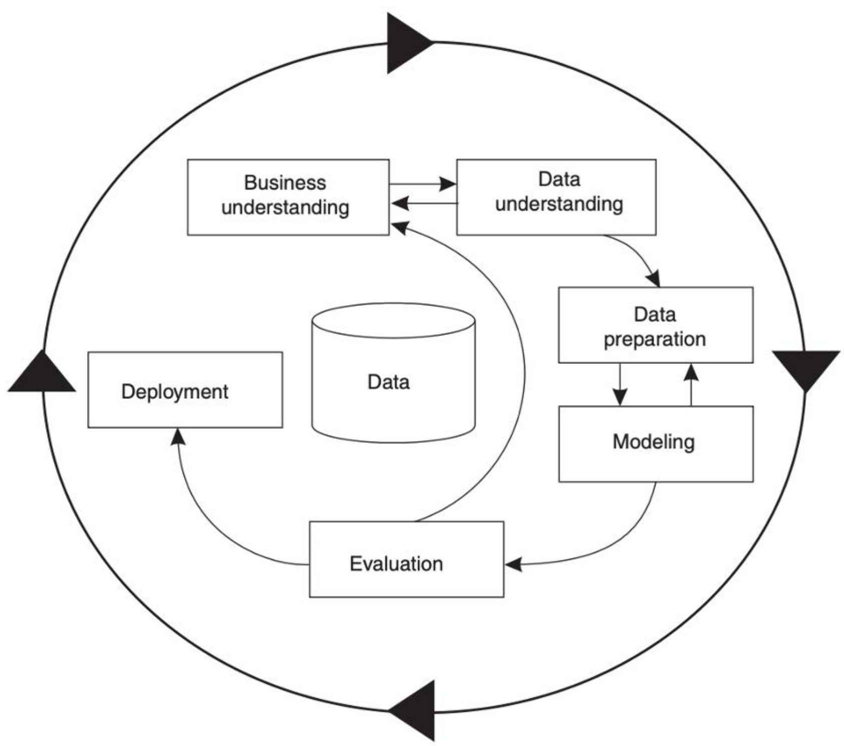

Figure 3.42 Cross-Industry Standard Process for Data Mining (CRISP-DM) process model (Chapman et al., 2000).

The Figure (3.43 Generic tasks and results of the CRISP-DM reference model) presents a scheme of phases accompanied by tasks and results, where we know the tasks and artifacts of this methodology.

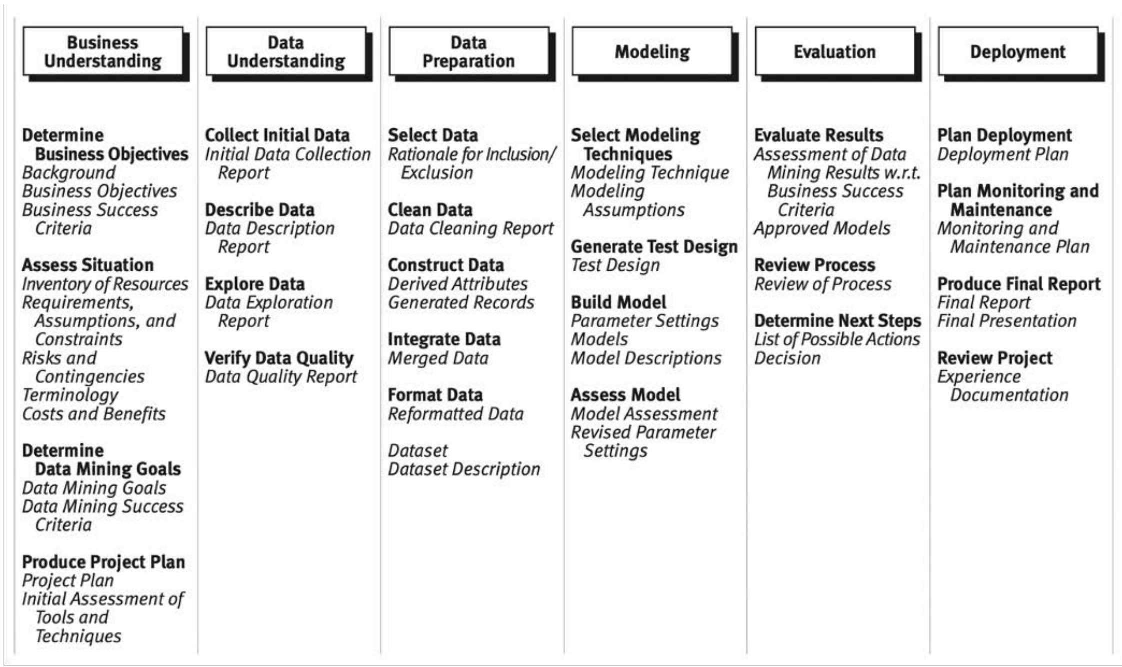

Figure 3.43 Generic tasks and results of the CRISP-DM reference model (Chapman et al., 2000).

As we can see, the main difference of CRISP-DM with respect to KDD and SEMMA is that this methodology is much more complete and clearly defines the phases, activities, roles, and artifacts that the methodology has, that is, it is mostly like a methodology for software engineering, in addition to having a greater documentation and being the most used by researchers and companies.

Table 3.16 (Summary of KDD, CRISP-DM and SEMMA Processes (Shafique et al., 2014)) and 3.17 (Comparative Data Mining methodologies) show us a comparison between the three methodologies, the first table shows us a comparison based on the number of steps that each of the methodologies follows to carry out Data Mining and obtain value from the data we have. On the other hand, the second table shows a comparison of the three methodologies with respect to the phases, activities, roles, and artifacts of each one of them.

Table 3.16 Summary of KDD, CRISP-DM and SEMMA Processes (Shafique et al., 2014).

Data Mining Process Model	KDD	SEMMA	CRISP-DM
No. of Steps	9	5	6
Name of Steps	Developing and Understanding of the Application	------------	Business Understanding
	Creating a Target Data Set	Sample	Data Understanding
	Data Cleaning and Pre-processing	Explore	
	Data Transformation	Modify	Data Preparation
	Choosing the suitable Data Mining Task	Model	Modeling
	Choosing the suitable Data Modeling Model Mining Algorithm		
	Employing Data Mining Algorithm		
	Interpreting Mined Patterns	Assessment	Evaluation
	Using Discovered Knowledge	------------	Deployment

Table 3.17 Comparative Data Mining methodologies.

	PHASES	ACTIVITIES	ARTIFACTS	ROLES
KDD	Selection	1. Learning the application domain. 2. Creating a target dataset. 3. Data cleaning and preprocessing. 4. Data reduction and projection. 5. Choosing the function of data mining. 6. Choosing the data mining algorithm(s). 7. Data mining. 8. Interpretation. 9. Using discovered knowledge.	• Data selection. • Choice of techniques.	Customer, Expert in Analytics / Data Science, IT Staff.
	Preprocessing		• Eliminate noise or inappropriate formats. • Data filter. • Selecting and editing data.	
	Transformation		• Choice of mining method. • Learning algorithm. • Model validation.	
	Data Mining		• Data visualization techniques. • Study, interpret and evaluate the knowledge model.	
	Interpretation / Evaluation		• Knowledge. • Decision making.	
SEMMA	Sample	Sampling.	• Representative sample of the data.	Customer, Expert in Analytics / Data Science, Developers.
	Explore	Data visualization. | Data transformation.	• Visualization data. • Basic description of the data.	
	Modify	Variable selection, creation. | Data transformation.	• Variables. • Transform variable representations.	
	Model	Selection Model (Neural networks, Tree based models, Logistic models, other stat models).	• Machine learning models.	
	Assess	Model assessment.	• Usefulness of the model.	
CRISP-DM	Business Understanding	1. Determine business objectives. 2. Assess the situation. 3. Determine the objectives of data mining. 4. Create a plan for your data mining project.	• Background • Business Objectives • Business Success Criteria • Inventory of Resources • Requirements, Assumptions, and Constraints • Risks and Contingencies • Terminology • Costs and Benefits • Data Mining Goals	* Customer, Business Analyst, Data Scientist, Data Engineer, Developer ,Project Manager

			• Data Mining Success Criteria • Project Plan • Initial Assessment of Tools and Techniques	
	Data Understanding	1. Collect initial data. 2. Describe the data 3. Explore the data. 4. Check the quality of the data.	• Initial Data Collection Report • Data Description Report • Data Exploration Report • Data Quality Report	
	Data preparation	1. Select data. 2. Data cleansing. 3. Data construction. 4. Integrate the data. 5. Format the data.	• Rationale for Inclusion/ Exclusion • Data Cleaning Report • Derived Attributes • Generated Records • Merged Data • Reformatted Data • Dataset • Dataset Description	
	Modeling	1. Select modeling technique. 2. Design the model tests. 3. Build the model. 4. Evaluate the model.	• Modeling Technique • Modeling Assumptions • Test Design • Parameter Settings • Models • Model Descriptions • Model Assessment • Revised Parameter Settings	
	Evaluation	1. Evaluate the results. 2. Process review. 3. Determine the next stages.	• Assessment of Data Mining Results w.r.t. Business Success Criteria • Approved Models • Review of Process • List of Possible Actions • Decision	
	Deployment	1. Plan deployment. 2. Plan monitoring and maintenance. 3. Create a final report. 4. Project review.	• Deployment Plan • Monitoring and Maintenance Plan • Final Report • Final Presentation • Experience Documentation	

Data Innovation Process and Cycle – DIPC (extra example that is based on a standard)

This is based on the software engineering life cycle for big data projects. It discusses the challenges of developing big data projects and proposes a new software engineering lifecycle process based on ISO/IEC 15288:2008. The article highlights the importance of variety in big data projects, which creates greater uncertainty and unexpected elements in the relationship between data. The proposed software engineering lifecycle process is expected to reduce project risks and increase completion rates of big data projects in the future. This PDF file is a valuable resource for anyone interested in big data and software engineering projects, as it provides insight into the unique challenges of big data projects and offers a practical solution for overcoming them.

The project goal, scope, and functional requirements in traditional data analysis or software projects are relatively more explicit than those of big data projects. There are four major characteristics used to determine whether a project involves big data: volume, velocity, variety, and variability (4V). Currently, big data projects fail to reach a high completion rate: the incompletion rate is about 55 percent,3 whereas the incompletion rate for general software projects is about 38 percent.4 The difference can be attributed to inaccurate scope3 and value of the results.

The lack of a typical lifecycle process for big data projects. So far, the IEEE has not provided a development process standard for big data projects; only for software projects. Figure 1 is taken from ISO/IEC 15288:2008-Systems and Software Engineering-System Life Cycle Processes, an international standard that clearly defines the process for general software projects.11 Data science and big data analytics are growing rapidly. A good process is needed, more specifically one that is suitable for big data projects.

To cope with the variety in big data projects, it is necessary to look for data relevance in structured and unstructured data sets. Well-known business needs or uses are marked in structured data and are considered relatively easy to find. Thus, it is common to start with known issues and structured data and then extend the data by integrating with unstructured data. However, discovering unstructured data directly or introducing external data step-by-step turns out to be the best way to get the most valuable results, i.e., creating data hierarchies and statistics to discover the relationship of information and data trends. This is the most valuable part of a big data project, although it is also considered one of the most risky and uncertain aspects of project management. Therefore, a proper process will be needed for a big data project, especially when faced with variety. In addition, the process should serve to change the way clients and project teams define the scope and value of a project.

For general data analysis projects or software projects, definite project goals or functions serve as the requirements for specification, followed by the work plan and implementation. However, variety in big data projects makes it impossible to completely verify the results of information application—the goal, in accordance with variety, should involve data-innovative processing and corresponding approaches. Since data innovation can lead to special value in results and uncertain factors, the data-innovation rate of a project should not account for 100 percent but instead should be limited to a rate that can be controlled, such as

20 percent. As for the rest of the results, 80 percent should be limited, defined, and controllable.

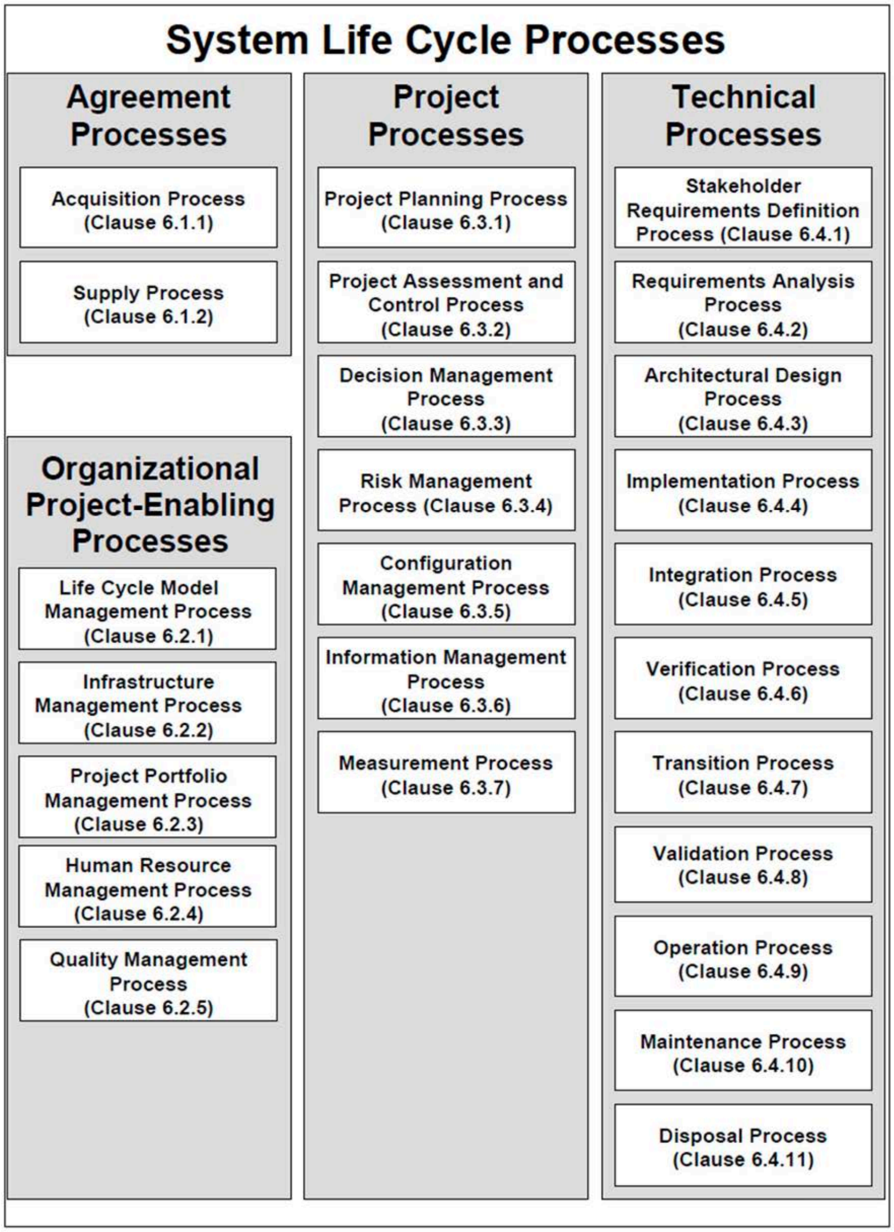

Figure 3.44 ISO/IEC standard 15288:2008—Systems and Software Engineering—System Life Cycle Processes.

Designing a lifecycle process for big data projects: There are four major elements involved in designing a process suitable for big data projects: one characteristic, one concept, and two processes. The characteristic refers to data variety, the concept is data innovation, and the processes are software engineering and data analysis.

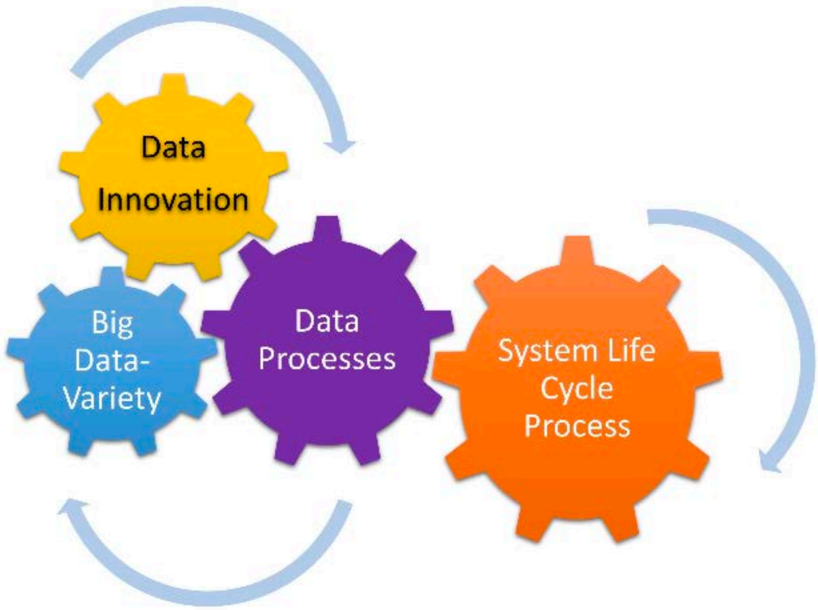

Figure 3.45 Major elements of the big data project lifecycle process

To deal with variety in big data projects, it is advised to establish the processes described below.

- Data value, result, and innovation process (agreement process). It is regarded as a risky undertaking for big data project contracts to only define project goals without including the data scope. Project risks can be prevented by first defining and controlling the data scope.
- Domain specialist resource management process (organizational project-enabling process). Due to the variety in big data, it is likely to grow more complicated in the management of interdisciplinary personnel. There should be an independent process set aside to be checked by a specialist, with the resources coming from clients or external experts.
- Data inventory process (data process). After the data is collected, the data inventory is performed for management. Data information is supposed to contain data format, type, source, amount, timestamp, states, renew period, owner, and so on.
- Data requirement analysis process (data process). This is undertaken for realizing the expected results and value.
- Data cleaning process (data process). To prevent data variety from being deleted, it is advised to clean the data after the data-innovation process is done.

To cope with data innovation, it is advised to establish the following processes.

- Data value, result, and innovation process (agreement process). Agreement processes are supposed to be used to make changes. It is required that 80 percent of the deliverables be confirmed for target results, and 20 percent should be reserved for data innovation. This 20 percent will likely offer clients the maximum value.

- Data innovation process (data process). This is considered a task force where data ana- lysts seek any possible data trends and relations, and are never limited by scope or goals. It is necessary to resume the data integration process, data inventory process, or data collecting process as soon as new data innovation takes place. See Figure 4 for the steps and flow cycle of the data innovation process.

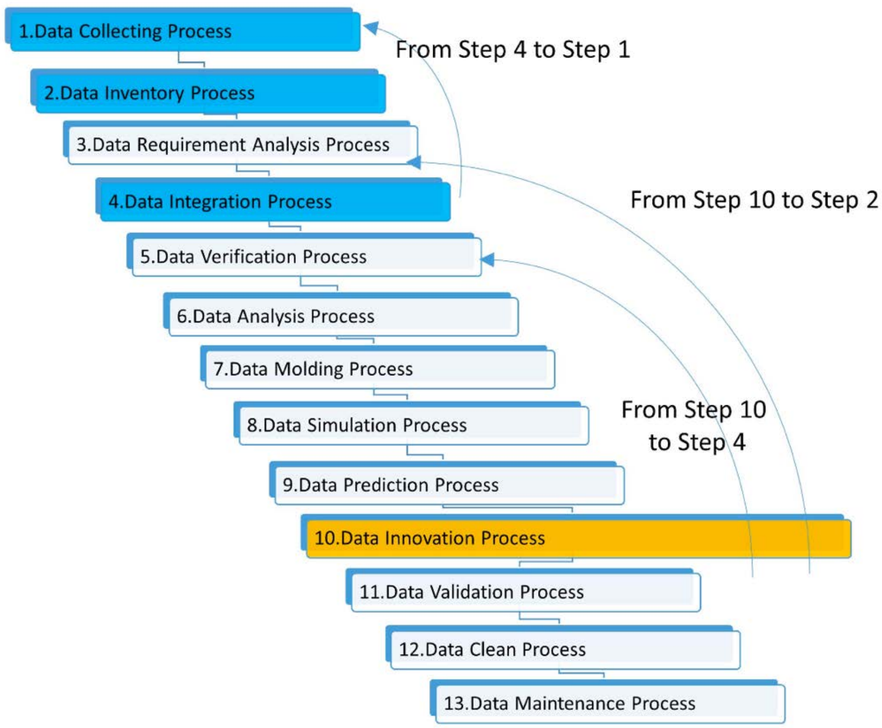

Figure 3.46 Data innovation process and cycle

It is advised that data processes serve as an independent process from the project processes and technical processes. Also, data processes should contain the following processes: data collecting, data inventory, data requirement analysis, data integration, data verification, data analysis, data modeling, data simulation, data prediction, data innovation, data validation, data cleaning, and data maintenance.

To deal with data processes, it is advised to establish the following technical processes.

- Data automation and monitoring process (technical process). These processes are mainly concerned with the establishment of a mechanism, by means of technical approaches, to collect and monitor data automatically and continually. The mechanism is supposed to be able to prevent the data source from being anomalous so that it gets only the right results.
- Data visualization process (technical process). Data visualization deserves great emphasis, as it is viewed as a very important part of a big data project. It is also important to make sure that results can be integrated with a visual tool or platform.

- Data decision support process (technical process). Most data projects are applied in the decision support for businesses or government. This process is mainly concerned with analysis and application of the results.

We used the above-mentioned processes, along with ISO/IEC 15288:2008, to design a lifecycle processes for big data projects, shown in next Table.

Table 3.18 Big data project lifecycle processes.

Agreement processes	Project processes	Data processes	Technical processes
Data value, result, and innovation process	Project planning process	Data collecting process	Stakeholder requirement definition process
Acquisition process	Project assessment and control process	Data inventory process	Requirement analysis process
Supply process	Decision management process	Data requirement analysis process	Architectural design process
	Risk management process	Data integration process	Data automation and monitoring process
	Configuration management process	Data verification process	Data visualization process
	Information management process	Data analysis process	Data decision support process
	Measurement process	Data modeling process	Implementation process
Organizational project-enabling process		Data simulation process	Integration process
Lifecycle mode management process		Data prediction process	Verification process
Infrastructure management process		Data innovation process	Transition process
Project portfolio management process		Data validation process	Validation process
Domain specialist resource management process		Data cleaning process	Operation process
Human resource management process		Data maintenance process	Maintenance process
Quality management process			Disposal process

This segment is based on the software engineering life cycle for big data projects. It discusses the challenges of developing big data projects and proposes a new software engineering lifecycle process based on ISO/IEC 15288:2008. The article highlights the importance of variety in big data projects, which creates greater uncertainty and unexpected elements in the relationship between data. The proposed software engineering lifecycle process is expected to reduce project risks and increase completion rates of big data projects in the future. This PDF file is a valuable resource for anyone interested in big data and software engineering projects, as it provides insight into the unique challenges of big data projects and offers a practical solution for overcoming them.

Translated with www.DeepL.com/Translator (free version) So far, many methodologies and models of data mining and knowledge discovery processes have been developed, with varying degrees of success. We have provided a brief description of the proposed Knowledge Discovery in Databases (KDD) process, discussing about the special features, salient advantages, and disadvantages of each approach. Focusing on the different steps and tasks, analyzing all the data mining approaches presented we started with KDDD Process from 1993 which was the basis for the current data mining process models and methodologies, that is why it is so important, later we described SEMMA process developed by SAS organization as an improvement of implementation of the initial KDDD Process but still with little specialized information related to this segment because SEMMA was only a segment of their system so many points of this process as the roles are not so clear. Until arriving to Crisp-DM, called the most used Methodology of this type today, where you can clearly see the evolution from process to methodology, leaving details of each step of the process, even generating a complex number of tasks and outputs. A key point is that the first two even for many do not fall into the area of methodologies for the little information you have or that describe their process and the final point we noticed in the three is that they are proprietary, this gives us a guideline to identify the need for the generation of a methodology of this type free where such methodology is useful not only for a project or organization in specific, if not seek a generic solution for such projects, this achieving key specifications of such projects to be more efficient in the key points necessary for the optimization of development. Remembering that there is no universal solution for all types of projects. That is why we must make clear the focus which is the type of projects to develop and the type of team or organization in charge of the development that in our chaos we will focus on small development groups.

In conclusion with this table and the investigated of the classic methodologies the most used is CRISP-DM, additionally to this by the type of methodology that is sought to generate is with a focused on the details this to be able to obtain a greater percentage of similarity with the base which is the ISO 29110, is for which after analyzing is selected CRISP-DM that still does not have the official definition of the roles, its detail in the artifacts and tasks is sufficiently clear and extensive as to have a great similarity with the standard.

3.1.3.6 Review of the Main Agile Analytics/Data Science SDM

The manifesto and principles for Agile Software Development (ASD) were published in 2001, and since then, the objectives and principles have been interpreted and applied to Analytics/Data Science. This manifest can be seen in the following figure.

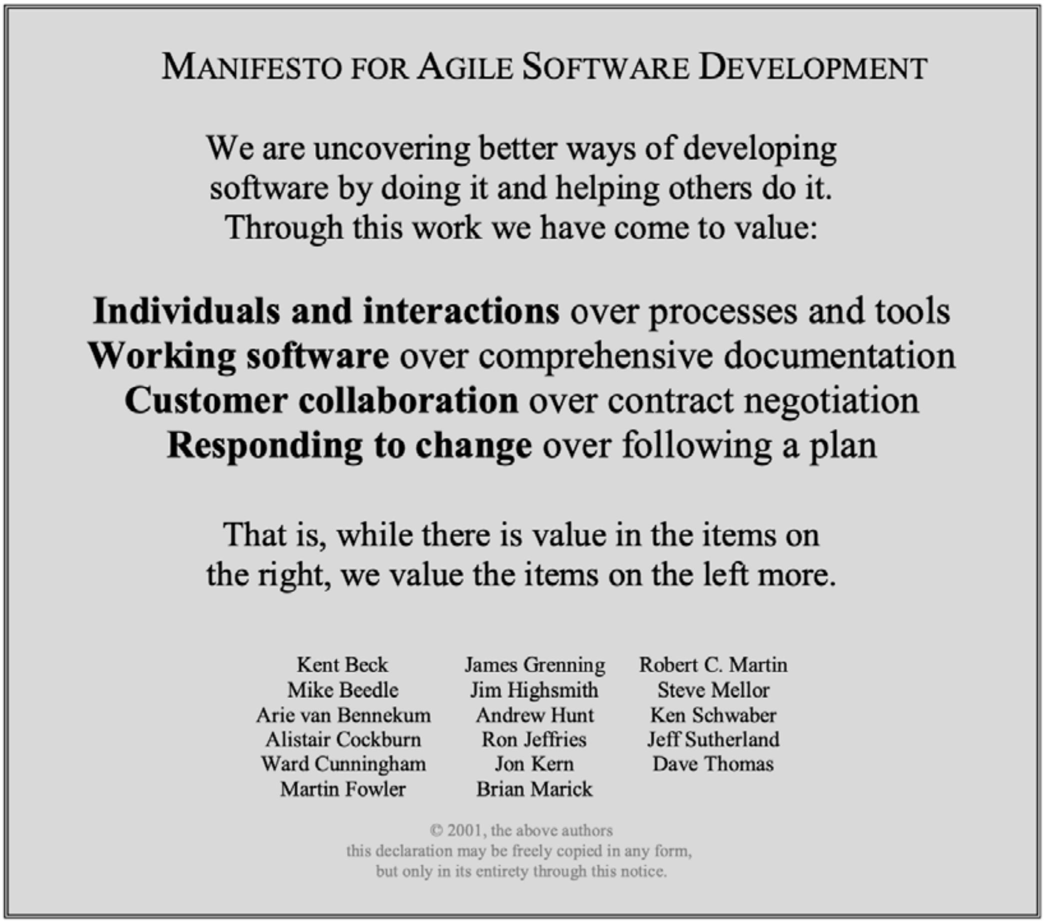

Figure 3.47 Manifesto For Agile Software Development (Beck et al., 2001)

Since the ASD manifesto was published, the objectives and principles have been interpreted and applied to new agile methodologies. The most popular approaches from which the manifesto and its principles were derived were: Extreme Programming (XP) and Scrum, where they are currently successfully practiced and considered standard development methodologies (Hsieh and Chen, 2015). Agile principles have been successfully applied to other disciplines such as project management (Kaleshovska, et al., 2015). Success related to agile methodologies includes reduced cycle times, higher quality, greater clarity in requirements, greater flexibility, and an overall higher rate of stakeholder satisfaction compared to similar projects using different projects or software development methodologies (Kaleshovska, et. al., 2015). Core practices of agile methodologies include: small, short releases; stakeholders physically co-located; and a time-bound project cycle (typically 60 to 90 days, although the cycle may be shorter depending on the deliverable) (Kendall & Kendall, 2005).

Technology has evolved because the amount of data generated through the Internet and smart devices has grown exponentially, altering the way in which organizations and individuals use information. The Big Data phenomenon, the volume, variety, and velocity of data, has impacted business intelligence and the use of information. New trends such as rapid analytics and data science have emerged as part of business intelligence (Larson, D., & Chang, V., 2016). In recent years, the use of software development lifecycle agility (SDLC) has gradually eclipsed the traditional waterfall model to become the dominant practice in projects and organizations of all sizes. Approximately seventy-five percent of software developers now say they primarily use agile practices, compared to twenty-five percent for waterfall. This trend is expanding with the incorporation of complementary DevOps concepts, which are based on agile practices. The popularity of the agile software development methodology (often referred to simply as agile) and how its principles can be applied to data science are best understood from its evolution (Larson, D., & Chang, V., 2016).

With the advent of Big Data and the evolution of analytics, the delivery of BI has been impacted. Data must be processed in a timely manner to be converted into information for analysis. Organizations are focusing more on prescriptive and predictive analytics that use machine learning and rapid analytics through visualization. Rapid analytics refers to the ability to acquire and visualize data quickly (Halper, 2015). The increase in data velocity has accelerated the need for IT departments to acquire data and transform it into information. Table 3.18 illustrates the characteristic differences between traditional BI and rapid analytics with Big Data.

Table 3.19 Comparison of Traditional Business Intelligence Systems and Fast Analytics with Big Data (Larson, D., & Chang, V., 2016).

Criteria	Traditional Business Intelligence	Fast Analytics with Big Data
Analytics Type	Descriptive, Predictive	Predictive, Prescriptive
Analytics Objective	Decision Support, Performance Management	Drive the Business
Data Type	Structured and defined	Unstructured, Undefined
Data Age	> 24 hours	< Minutes

Scrum, although not limited to BI, is the most popular agile approach used in agile software development and BI (Muntean & Surcel, 2013). The Scrum concepts mainly used in BI are user story, sprint backlog, product backlog, sprint and daily scrum. BI requirements are broken down into small stories that are then packaged into a collection of stories to form a BI project. Each story is designed, developed, tested, and released. A sprint lasts one to two weeks and contains a cycle of requirements, analysis, design, development, and end-user testing. Stories can be grouped into the product or sprint backlog. The sprint backlog refers to the work that the development team completes during the sprint. The product backlog is a list of all stories ordered by priority to be considered for the next spring. Users participate in all steps of the sprint. Daily meetings of less than 15 minutes are held to review status (Muntean & Surcel, 2013).

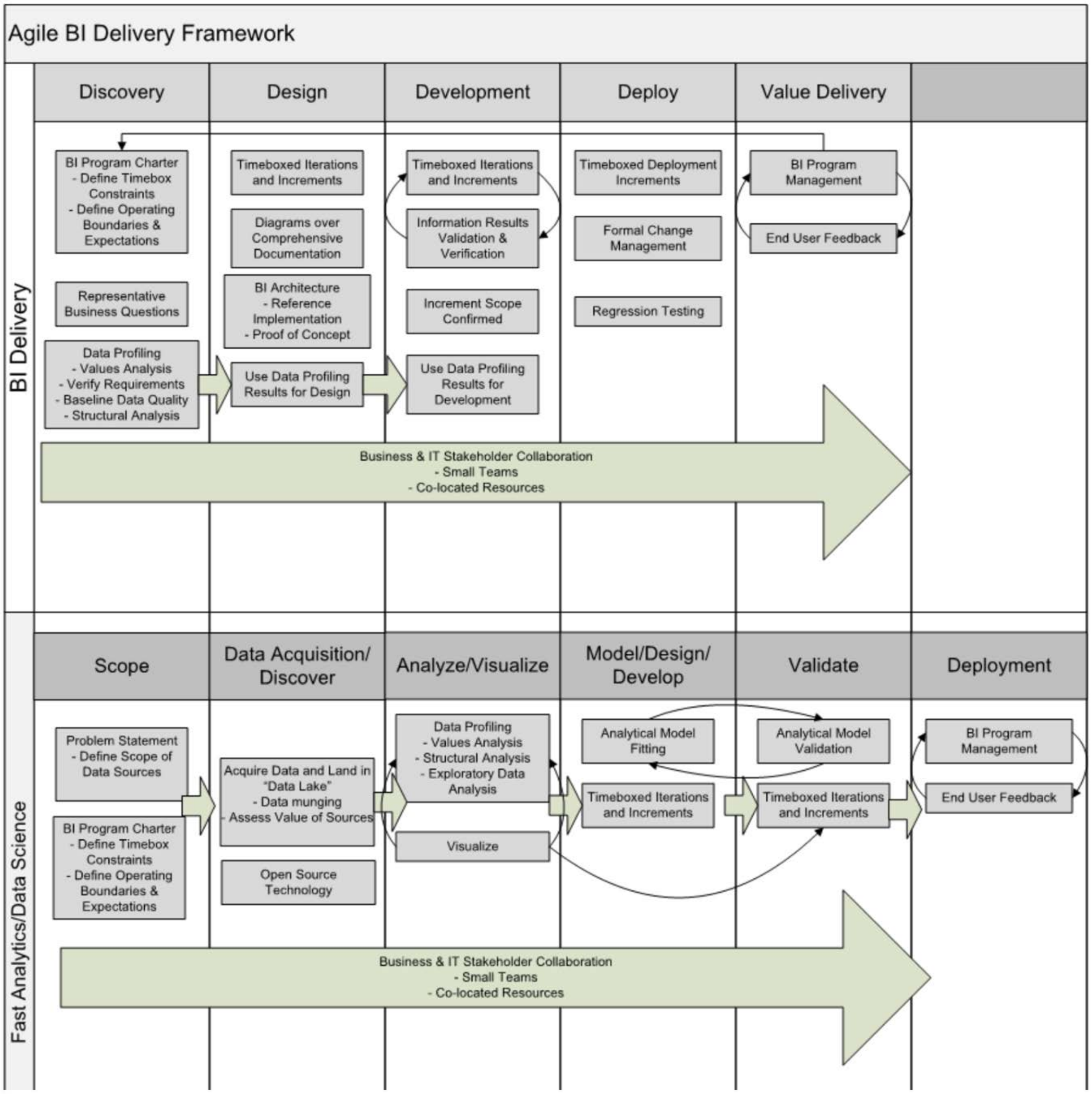

Figure 3.48 Proposed Agile Delivery Framework (Larson, D., & Chang, V., 2016).

The figure above shows the framework proposed by Larson & Chang for comparing Agile BI, rapid analytics, and data science. There are two layers of strategic tasks that go hand in hand in the Agile BI Delivery Framework. The top layer includes BI Delivery, and the bottom layer includes Fast Analytics/Data Science. In the top layer, there are five sequential steps involved: discovery, design, development, implementation, and value delivery. At each step, there are specific tasks to be completed that relate to the objectives of achieving business and IT stakeholder collaboration. The lower layer includes six sequential steps: scoping, data acquisition/discovery, analysis/visualization, validation, and implementation. Likewise, all these steps work to achieve successful collaboration between business and IT stakeholders. Alignment, integration and optimization of both layers can ensure the execution and management of the Agile BI delivery framework (Larson, D., & Chang, 2016). This model will be revisited later in this chapter for agile methodologies.

Grady and Payne in 2017 made mention of the limitations of process models and the challenge for big data analytics and comment that traditionally, in statistical analysis, data was carefully collected so that it was necessary and sufficient to definitively answer a particular question, for example, in the pharmaceutical industry to pass the regulatory requirements of the Food and Drug Administration (FDA) to launch a drug. In the late 1990s, analytics moved out of the realm of traditional statistics and into what is known as data mining. At that time, the data mining community began to use a series of models on reused data to analyze the data in a new context other than the one for which it was collected. These new mathematical models were used to provide a probabilistic answer rather than a deterministic answer. As the data mining community grew, it became difficult to compare approaches because the activities of the end-to-end process were described in various numbers and types of steps. The solution developed by a consortium is the Cross Industry Data Mining Standard Process Model (CRISP-DM) (Chapman, et al., 2000). CRISP-DM remains the dominant process model still in use today (Piatetsky-Shapiro, 2014) despite several recognized problems (Taylor,2017).

Data science as a term emerged from advances in big data engineering with the recognition that large volume or high velocity data sets required new parallelization techniques to parallelize to handle the large amounts of data efficiently (Grady & Chang, 2017) Furthermore, it was often sufficient to leverage correlation rather than having to understand causation. For example, in testing websites A and B, if more customers click on a blue link than on a green link, it is not necessary to know the perceptual reasons; the correlation itself is actionable to optimize the website design. Big data represents data distribution and parallel processing, so data storage, algorithms and analytics lifecycle are no longer separable from big data technologies. Although still the de facto standard, CRISP-DM does not address these changes, as well as automation, data science, or system development methodologies (Grady & Chang, 2017). Although efforts have been made to adapt different types of analysis to an agile methodology, the steps are often described by breaking activities into smaller tasks, while the overall process remains consistent with the step-by-step waterfall. BDA changes several of the activities in the analytics lifecycle, as well as their ordering (Grady & Chang, 2017).

The development of analytics systems is driven by detailed requirements for building each system capability. Most of the development of advanced analytical systems is based on desired outcomes rather than specific requirements. The categories of analytics can be described as a ladder of increasing complexity, as shown in Fig. 3.46. Information and business intelligence systems can be described by explicit requirements. In contrast, more sophisticated analyses require multiple computational experiments and comparative analysis through modeling, machine learning and simulation to obtain results. Model fitting and optimization is an iterative exploratory process of experimentation, testing, and evaluation. Advanced analyses cannot be specified by a list of detailed model requirements, but only by desired outcomes (Grady & Chang, 2017).

Figure 3.49 The Analytics Ladder (Grady & Chang, 2017).

Nancy W. Grady and Jason A. Payne also mentioned that Several factors highlighted the inherent deficiencies in the waterfall method and drove the emergence of the agile methodology as a necessary evolution in software development. These can be grouped into three primary categories:

1. Human factors in software development.
2. Changes in cost and risk factors.
3. Understanding the nature of complex systems.

Figure 3.50 Waterfall Learning Curve (Grady & Chang, 2017)

For category 1, the Agile method focuses on the understanding that software development is a knowledge work done by individuals working together with other individuals in teams. While this may seem obvious, it is not something that the waterfall method placed much importance on. The waterfall process evolved from mass manufacturing practices for physical products and valued standardization, automation, and repetition. Individuals performed limited functions and, if they produced a result within certain tolerances, their output could be reliably consumed by the next stage of the process. The result is a large-scale final delivery, with little change in knowledge until the end, as Cockburn illustrated in 2017 in the last Figure.

In modern software development, the main constraint is the learning pace of the development team. Agile methodology accelerates learning through the use of frequent integration, iterative releases, and user feedback. Prototyping is relatively fast and inexpensive in software, and early feedback can replace a large initial design with better end results. Sometimes multiple prototypes are employed, using A/B testing to quickly converge on the best solution. Cockburn's learning curve in the Figure below illustrates the rapid learning of the team in the early phases of the project. With these images we can differentiate the rhythms of knowledge and see the distinction of these two types of methodologies with respect to the human side of the project where agile versions are used.

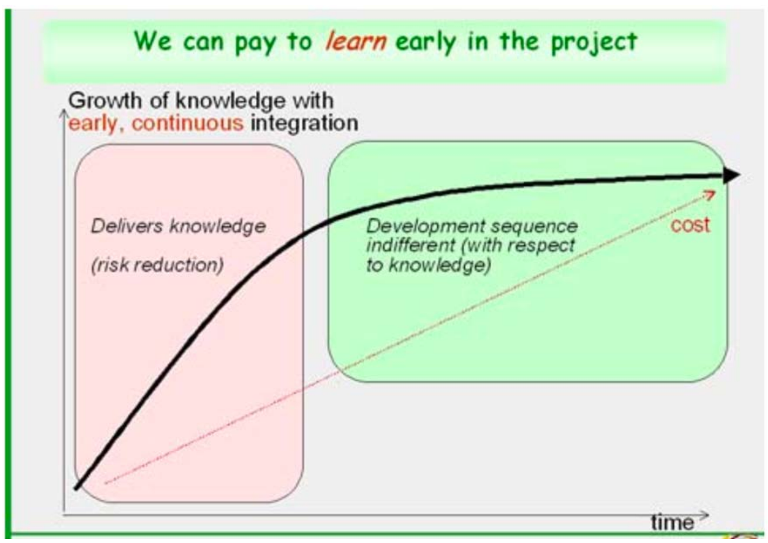

Figure 3.51 Agile Learning Curve (Grady & Chang, 2017)

Category 2: Changes in cost and risk factors. If IT resources are scarce and expensive, it makes sense for a development team to spend a lot of time and effort on exhaustive requirements gathering and initial design layout. Or, if the company's innovation cycle is long enough, perhaps the organization can tolerate the cost of delay when the development project does not deliver any functional software for several years. In either case, there may be no motivation to pursue an alternative to the waterfall approach. At present, it is unlikely that either of these scenarios would apply in any industry or organization. When development projects used to last many years and systems were expected to run for decades, the risk of changes in requirements or design was very high. Today, the speed with which technology evolves and the reduced expectations for product cycle times mean that the cost of delayed software releases is often the biggest risk. Because of Agile's imperative to deliver software that works early and often, there is an early return on investment, and early feedback allows development to reach the level of a minimum viable product (MVP) more quickly.

The last category, the nature of complex systems: The model categorizes systems into four domains (shown in the next Figure): Simple, Complicated, Complex, and Chaotic. The open space in the center represents Disorder, that is, not knowing which domain you are dealing with. This framework helps us to understand the types of decisions and behaviors that are likely to be successful in each domain based on the characteristics of the system.

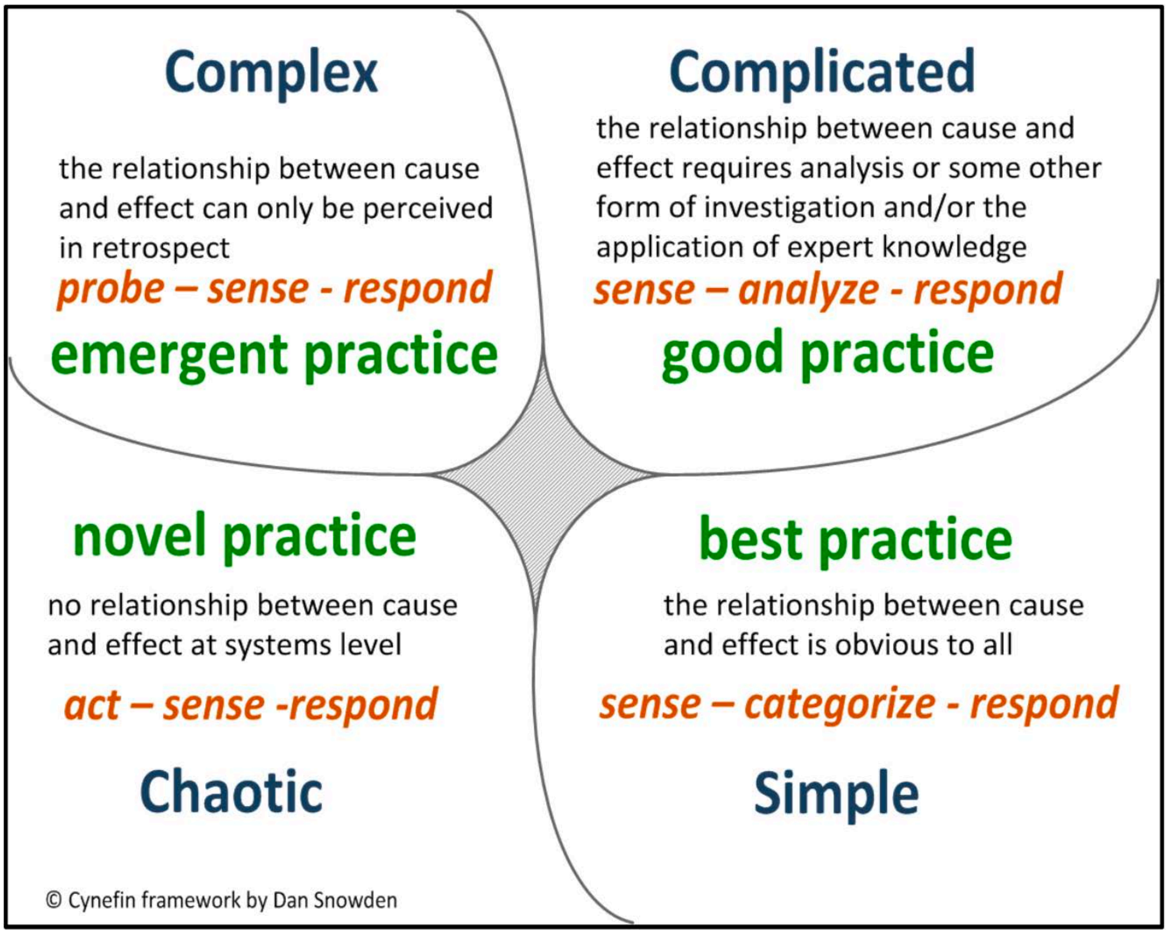

Figure 3.52 Cynefin framework (Kurtz & Snowden, 2003)

SAIC has developed a BDA process model that extends the previous CRISP-DM data mining model to incorporate new big data and cloud technologies. This BDA process model is called Data Science EdgeTM (DSE), shown in Fig. 3.50, which serves as our process model for Knowledge Discovery in Data Science (KDDS). It is formed around the National Institute of Standards and Technology (NIST) big data reference architecture (RA), explicitly showing the frameworks from which, the application draws. DSE provides several improvements over CRISP-DM. Where it can be seen that it is. Seeking to generate a more specific crisp-dm approach (Grady & Chang, 2017).

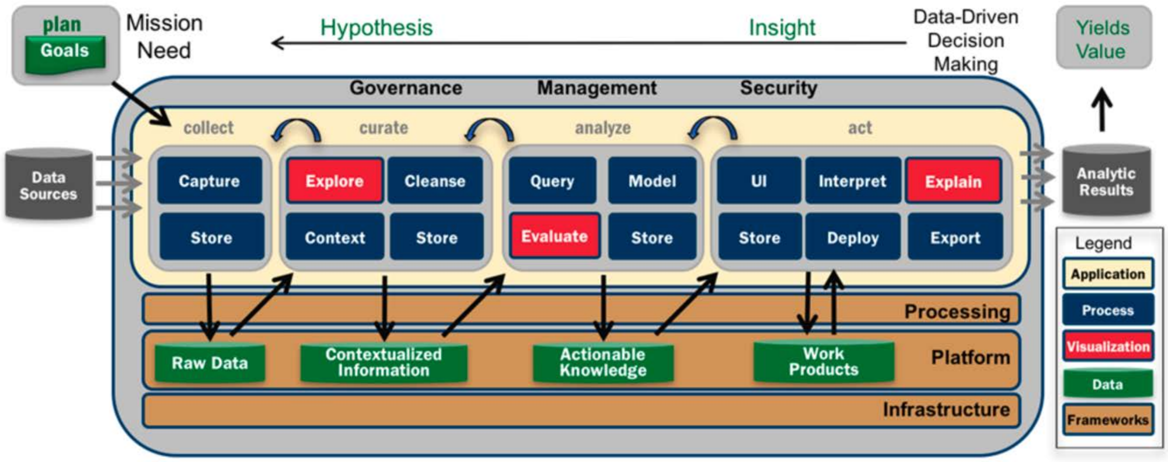

Figure 3.53 Data Science EdgeTM—a big data analytics process model (Grady & Chang, 2017).

In conclusion, CRISP-DM continues to play an important role as a common framework for establishing and managing data mining projects. However, today's world is very different from the world in which CRISP-DM was conceived more than two decades ago. In this paper we have argued that the shift from data mining to data science is not just terminological but signifies an evolution to a much wider range of approaches, where the main value-added component may be indeterminate in the output and needs to be discovered as part of the project. For these exploratory projects, the CRISP-DM framework will be too restrictive. We have proposed a new Data Science Trajectories (DST) framework that extends CRISP-DM by including exploratory activities such as target exploration, data source exploration, and data value exploration. The entry points, trajectories, and exit points of this richer set of data science steps can vary greatly among data science projects. We have illustrated this through a wide range of exemplary projects and the trajectories they embody (Martínez-Plumed, et al., 2019).

Data science is still a young subject, with many open questions about its nature and methodology. While other authors approach these questions from a top-down perspective, what we have attempted here is more bottom-up, starting from something that is generally considered productive in the context of data mining, and investigating how it can be generalized to account for the much richer context of data science. We therefore see this as part of a broader, ongoing conversation, and hope that the perspective offered here will be received as a positive contribution. (Martínez-Plumed, et al., 2019).

In order to move on to the search for Agile models we first mention what Chow and Cao identified as the success factors suggested in the Agile literature for software development projects (Chow & Cao, 2008). This is the most comprehensive review we have found, so we use it as a starting point to develop an updated list of possible success factors for agile analytics projects. They initially identified 36 success attributes that potentially affect project success (success was defined along four dimensions: quality, scope, time, and cost). These attributes, listed in the next Table, were reduced to 12 factors and organized into five factor dimensions (see Figure 3.52).

Success Dimension and Success Factor	Chow and Cao (2008) Attributes	Attributes for Agile Analytics Success (new attributes bolded)
Organizational Factors: – Strong Management Commitment	Strong executive support; Committed sponsor or manager	Strong executive support; Committed sponsor or manager
- Agile-friendly Organizational Environment	Cooperative organizational culture; Universal acceptance of agile methodology; Reward system appropriate for agile teams; Facility with proper agile-style work environment	**A willingness to take on risks**; Cooperative organizational culture; Universal acceptance of agile methodology; Reward system appropriate for agile teams; Facility with proper agile-style work environment
- Agile-friendly Team Environment	Collocation of team; Small team; Coherent, self-organizing teamwork; projects without multiple independent teams; Managers knowledgeable in agile processes	Collocation of team; Small team; Coherent, self-organizing teamwork; projects without multiple independent teams; Managers knowledgeable in agile processes
People Factors: - High-caliber Team Capability	Team with high competence and expertise; Adaptive management styles; Team members with great motivation; Appropriate technical training to team	**Having the appropriate diversity to match task complexity; Team dedication/time availability exclusively for the project; Engaging people**; Team with high competence and expertise; Adaptive Management Styles; Team members with great motivation; Appropriate technical training to team
- Strong Customer Involvement	Good customer relationship; Strong customer commitment; Customer having full authority	Good customer relationship; Strong customer commitment; Customer having full authority
Process Factors: - Agile-style Project Management Process	Agile-oriented requirement management, project management, and configuration management processes; Good process tracking mechanism; Strong communications focus with daily face-to-face meetings; Honoring regular work schedules	**Good project planning**; Agile-oriented requirement management, project management, and configuration management processes; Good process tracking mechanism; Strong communications focus with daily face-to-face meetings; Honoring regular work schedules;
- Methodical Project Definition Process	Up-front risk analysis; Up-front cost analysis	**Establishing clear goals**; Up-front risk analysis; Up-front cost analysis
Technical Factors: - Agile Software/ **Analytics** Techniques	Defined coding standards; Pursuing simple design; Rigorous refactoring; Appropriate documentation; Correct integration testing	**Ensure high data quality; Model validation activities; Build customer's trust in model solution**; Pursuing simple design; Appropriate documentation
- Agile-style Delivery Strategy	Regular delivery of customer functionality; Delivering most important features first	Regular delivery of customer functionality; Delivering most important features first
Project Factors: - Non-life-critical Project Nature	Project nature being non-life-critical	Project nature being non-life-critical
- Variable Scope Project Type	Variable scope with emergent requirements	**Technological uncertainty with respect to how to meet requirements**; Variable scope with emergent requirements
- Dynamic, Accelerated Project Schedule	Dynamic, accelerated schedule	Dynamic, accelerated schedule

Figure 3.54 Adapted agile analytics success factors and attributes (Chow & Cao, 2008).

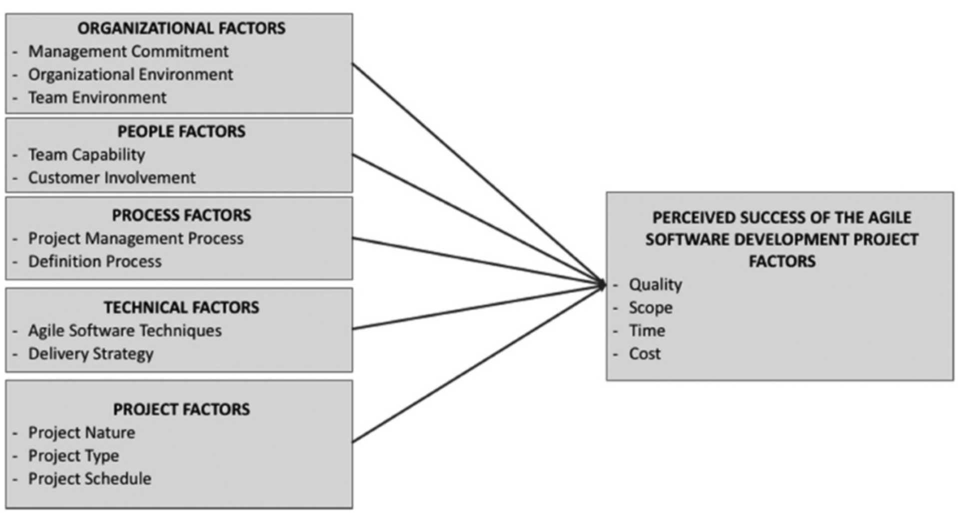

Figure 3.55 Factors organized in five factor dimensions (Chow & Cao, 2008).

Businesses' efforts to leverage the data they have access to are growing rapidly. To derive value from innovative BI and Big Data initiatives, organizations need to develop the capability to realize the intended benefits. Organizations effect change through projects, so developing the capability for excellence in execution can enhance competitive advantage. How a project is selected, managed and implemented (the project output/outcome) can have a significant impact on the value a company derives from the initiative. Therefore, better understanding how to effectively lead, support, and manage analytics projects can help organizations understand how to maximize the business value obtained (Tsoy & Sandy, 2020)

Tsoy & Sandy conducted this research to examine the factors that can affect the success of agile analysis projects. Where one of the key points was "*A methodology as such, no, is not defined. Basically, we use the normal strategy of project implementation and what we have followed has been low experience, from the identification, definition, implementation, testing and the already productive process. But as such metho- dology, we do not have it. (Project Maple, Data Analyst)"* (Tsoy & Sandy, 2020). They concluded that further research is required to clearly determine with different characteristics such as the level of expertise of the team for a better result. The two factors that show the clearest pattern in our data are strong customer involvement and a methodical project definition process. Both were clearly weak in low-performing projects (Pars and Willow) and strong in high-performing projects (Maple and Clover). Good customer involvement would allow for feedback and the ability to resolve uncertainty and problems, while clarity of objectives and risk management would help create agreement on project direction and help avoid major disruptions. Project management processes that create strong communication, planning capability and requirements management also featured more strongly in our successful projects. Similarly, analysis techniques suggested by the literature, specially building customer confidence in the model solution, were stronger in our successful projects. Chow and Cao (2008) define a critical success factor as a "factor that must be present for the agile project to succeed" (Tsoy & Sandy, 2020). This gives rise to continue with the generation of a methodology for this type of projects.

Understanding already the importance of agile and its current factors and the relation with the current projects related to Analytics/Data Science, the following agile based methodologies were searched:

Team Data Science Process (TDSP)

Microsoft's Team Data Science Process (TDSP) is an iterative data science framework that was initially release in 2016 (the framework was most recently updated in early 2020). As a relatively new framework, there are few people trained and certified in how to use TDSP.

TDSP defines a high-level data science project life cycle as well as a standardized data science project structure (e.g., team roles). While there are parts of the framework that leverage Microsoft tools and infrastructure, the rest of this discussion will focus on the more generic aspects of the framework, that are not tied to Microsoft's suite of products.

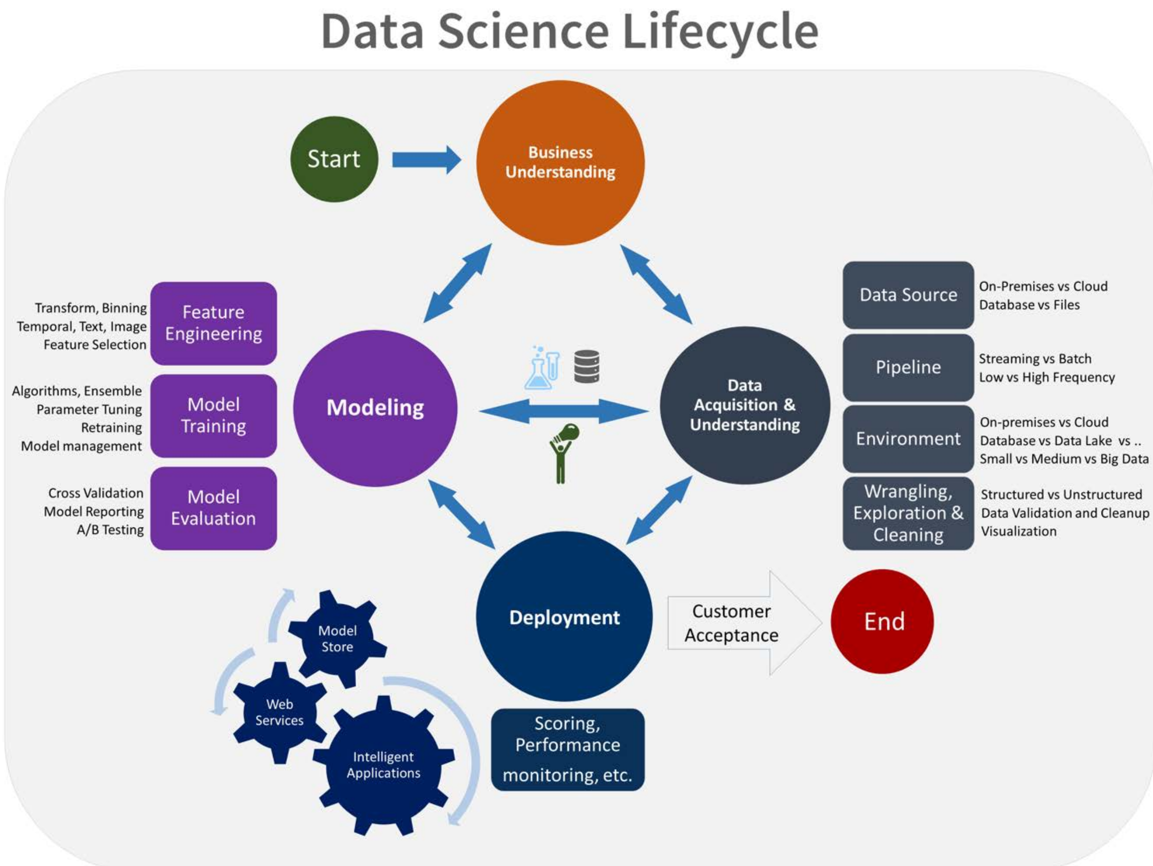

Figure 3.56 Team Data Science Process lifecycle (TDSP) (Microsoft, 2025)

TDSP's process lifecycle framework
This aspect of the framework focuses on "what to do" (not "how to do it"), by defining five stages for a project (which is sometimes known as project phases). The stages within TDSP's lifecycle are like CRISP-DM's phases, and include Business Understanding, Data Acquisition and Understanding, Modeling, Deployment, and Customer Acceptance.

The TDSP lifecycle is modeled as a sequence of iterated steps that provide guidance on the tasks needed to create and use predictive models. Note that, similar to CRISP-DM, projects are expected to "loop back" (in other words, execute these phrases iteratively). However, like CRISP-DM, the framework does not define when the team should iterate. For example, the team can iterate and the next of each complete lifecycle, or between phases within the lifecycle. Note that teams using TDSP are free to pick another lifecycle framework (such as directly using CRISP-DM or an organization's custom set of phases).

TDSP's team roles
While TDSP's framework is like CRISP-DMs, TDSP does address CRISP-DM's lack of team definition by defining four distinct roles (solution architect, project manager, data scientist, and project lead) and their responsibilities during each phase of the project lifecycle

For the team to complete the project, these stage-specific tasks and artifacts are associated with specific project roles. In fact, TDSP defines four specific roles (solution architect, project manager, data scientist, project lead). Note that the many aspects of data science, such as data engineering, are merged within the data scientist role).

Standardized Resources (project structure / tracking)
Independent of the actual lifecycle framework used, for each stage, TDSP provides goals, artifacts and guidance on how to complete the artifacts. Demonstrating the focus on project document artifacts, TDSP suggests that all project documents use standard templates and that the documents (as well as project code) is stored in a version control system (such as Git). In fact, key concept within TDSP is the focus on communicating tasks across the team and stakeholders / customers by using a well-defined set of artifacts that employ standardized templates. With respect to task/feature tracking and prioritization, TDSP suggests using one of the many commonly available tracking systems (such as Jira, Rally). Furthermore, TDSP suggests using these tools to provide cost estimates as part of the project process.

Agile & TDSP
Microsoft provides a description of how to integrate the concepts of scrum within TDSP. Basically, one can define a backlog with work items, and then use that backlog to do sprint planning and sprint execution. In the TDSP sprint planning framework, there are four typical work item types (Features, User Stories, Tasks, and Bugs) and there is one backlog for all the work items, which are tracked / managed at the project level. Just as with other scrum projects, these work items can be managed via the traditional scrum processes (such as sprint planning meetings). However, while Microsoft describes TDSP as supporting an agile approach, there are also waterfall like aspects to the framework. For example, at the end of each stage, there are specific artifacts that need to be created, including:

- Business Understanding: Project Charter (project manager)
- Data Acquisition& Understanding: Data Summary Report (data scientist), Solution Architecture Diagram (Solution Architect)
- Modeling: Model Report (Data Scientist)
- Deployment: Dashboard (Data scientist)
- Acceptance: Project Final Report (Project Manager)

TDSP Flexibility
Teams are free to use TDSP Stages, CRISP-DM phases or any other project lifecycle they deem appropriate. Furthermore, when using TDSP, the team is free to use scrum sprints or more traditional project deadlines. In addition, when trying to use an agile version of TDSP, it is up to the team on how to think about work tasks – either via the feature/user stories work items, or via the TDSP lifecycle stages. In other words, if a team uses scrum within TDSP, they will use sprints but think of tasks either via the project phases or via tracking features / User stories (but not both).

Hence, there is a fair amount of freedom on how to use TDSP – which means that there can be a fair amount of variation on how team's use TDSP (which can be good or bad but suggests that each team needs to determine their own set of best practices).

Proposed agile delivery framework (Fast Analytics / Data Science)

Scope

Fast analytics and data science emergence is due to Big Data; however, it is important to point out that forms of fast analytics and data science have existed for some time. Visual analytics is considered synonymous with fast analytics and data mining is also used synonymously with data science. Both fast analytics and data science are newer versions of known data analysis methods (Davenport, 2014; Keim, Kohlhammer, & Ellis, 2010; Schutt & O'Neil, 2013). Visual analytics and data mining became complementary techniques where visualization was used during the discovery phase of data mining (Keim, et. al, 2010). As Big Data was used more in analysis, visualization tools were used to explore raw data to support exploratory data analysis in the data science process. New tools emerged in the BI industry for visualization that included new functionality such as complex graphs and charts and the ability to connect to many different data sources (Davenport, 2014).

The scope of fast analytics and data science is to acquire data quickly to analyze. Fast analytics is more about discovery and data science uses fast analytics as part of its process. As a result of the data science process, a data product such as a prediction engine, a classifier, or recommendation engine is created (Schutt & O'Neil, 2013). The scope of fast analytics and data science will depend on the problem statement of the analysis. Many data sources could be included in the scope of analysis. Data sources may not be limited to unstructured data. Here BI program management can have value as a charter for the analytical model can define the problem statement and objectives as well as include operating boundaries and expectations.

Data Acquisition/Discovery

New technologies have made it possible to acquire data without a full understanding of its structure or meaning which is the opposite of what occurs in the BI lifecycle where data is profiled and analyzed to understand its meaning before loaded into a data repository for use. Hadoop or the Hadoop File System (HDFS) originated at Google and is now used in an open-source format by organizations to land data without the need for data modeling (Davenport, 2014). Analysts use fast analytics to access, assess, and visualize to discover the value and use of data sources. New data repositories such as the "Data Lake" have emerged where technology enables storage and processing power to support analyzing large unstructured data sets (Davenport, 2014).

Analyze/Visualize

For both fast analytics and data science, analysis and visualization are an iterative process. With fast analytics the primary goal is visual analytics to support analysis. Fast analytics can produce new knowledge that creates a refinement of the visual product. Fast analytic can iteratively produce new dashboards or scorecards to be used in ongoing BI or produce one-time analysis tools to support new knowledge gain. With data science, fast analytics and visualization is completed as part of the exploratory data analysis phase where descriptive analysis is used to highlight variable relationships and identify parameters to be used in analytical models (Schutt & O'Neil, 2013). If fast analytics and visualization produces a BI product such a dashboard or scorecard, the BI product is then validated. It is possible that fast analytics is primarily focused on discover, and a BI product is not produced.

Model/Design/Development

Modeling is used two ways in this phase: analytical modeling in data science and data modeling to describe data used in fast analytics. Analytical modeling include descriptive, predictive, and prescriptive analysis using machine learning algorithms such as regression, clustering, or classification (Schutt & O'Neil, 2013). In fast analytics, data is modeled after analysis to document data structures and association for future use (Adamson, 2015).

Validate

The validation phase is representing the data science process of validating the analytic model iteratively to the point where the error of the modeling is minimalized. This process is referred to as "fitting" the model. Additionally, fast analytics can be used to identify new parameters to incorporate into the analytical modeling process (Schutt & O'Neil, 2013). In this phase, new data sources may also be incorporated.

Deployment

As with BI products and systems, analytical models, dashboards, scorecards and other visualization tools have little value unless they are used. These analytical products are added to the production environment to provide new functionality to the environment, just like the BI deployment.

Support/Feedback

Analytical products need to be supported and revised as the organizational environment changes. The life cycle of an analytical model depends on the rate of change in the organization and the industry the organization operates within. Analytical models lose value an applicability over time and ongoing feedback from users and analysis determines how the analytical models should be adjusted.

Synthesis of the Fast Analytics/Data Science Lifecycle and Agile

Three phases of the BI Lifecycle have characteristics were using an Agile approach fits. The discovery, design, and development phases benefit from iterative cycles, stakeholder collaboration, small time-boxed increments, and co-located resources (Ambler & Lines, 2016; Hughes, 2013; Muntean & Surcel, 2013; Powell, 2014).

Fast analytics and data science are more fluid and iterative than BI due to the discovery involved in investigating a problem statement. Fast analytics and data science are inherently agile as each follows iterations, use small teams, and require collaboration between business subject matter experts and technical resources. Time-boxed increments can be applied, but may or may not be used as both processes are focused on discovery and data science has the objective of creating an analytical model that produces the best results (Schutt & O'Neil, 2013; Mohanty, Jagadeesh, & Srivatsa, 2013).

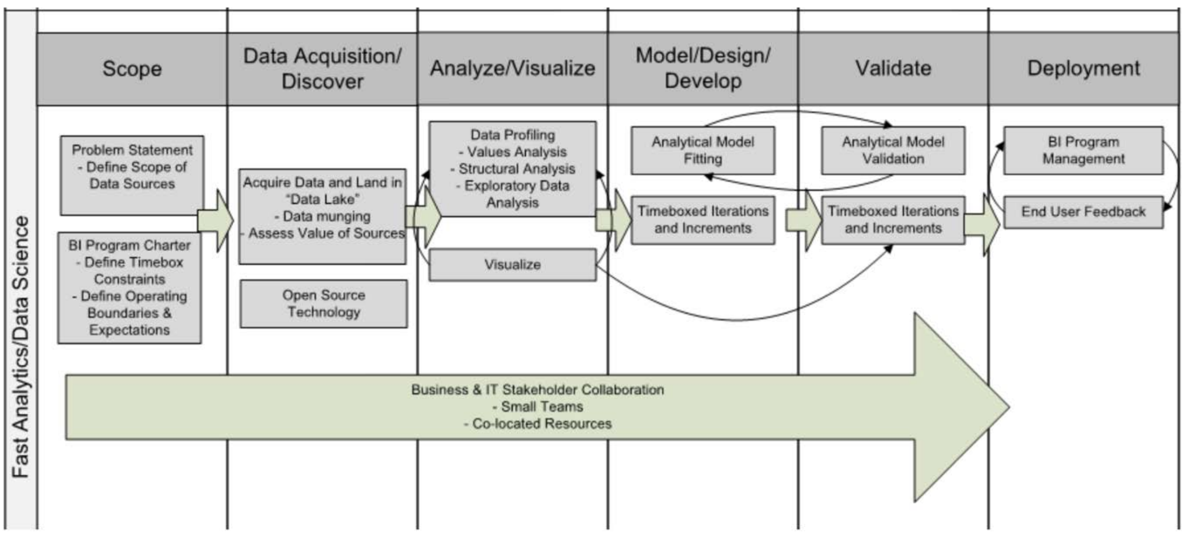

Figure 3.57 Proposed Agile Delivery Framework - Fast Analytics/Data Science (Larson, D., & Chang, V., 2016).

The Practical Guide to Managing Data Science at Scale (Domino - DDSL)

Project Recommendations - Managing a data science project

Now that we've established the goals, suboptimal outcomes, and underlying causes of those outcomes, it's time to discuss how to modify our data science machine to achieve the promising results we know are possible. In this chapter, we synthesize the successful project practices from dozens of leading data science organizations spanning many sizes and industries. This chapter is deliberately more detailed and tactical than earlier sections specifically so that readers can take away actionable insights for their own organizations.

Before jumping into the details, Domino Data Lab's overall lifecycle methodology can be viewed in this aggregate flow chart that encompasses the people, process, and technology we see across leading organizations. The approach can be summarized as: Imagine your ideal process for a single data science project, then consider how to manage a portfolio of those projects, and then think about the types of people, tools and organization structure you need to best achieve your goals.

Finally, we would be remiss if we didn't mention the existence of other project frameworks for data science, notably CRISP-DM. What follows is inspired by CRISP-DM and other frameworks but based more on practical realities we've seen with leading data science organizations, like Allstate, Monsanto, and Moody's. We step through the key stages that we've seen consistently emerge across many organizations' data science lifecycle: ideation, artifact selection, data acquisition and exploration, research and development, validation, delivery, and monitoring. However, the methodology and best practices here are broader than the process to manage a single project.

Overall Lifecycle Principles

Before jumping into the specifics of each project stage, below are a few guiding principles.

Expect and embrace iteration. The flow of a project is highly iterative, but, by and large, nearly all projects pass through these stages at one point or another. It is normal for a project to get to validation and then need to go back to data acquisition. It is also normal for a single project to open 10 additional avenues of future exploration.
What separates leading organizations is their ability to prevent iterations from meaningfully delaying projects, or distracting them from the goal at hand. One leading organization includes an "areas for future exploration" in all project deliverables and has educated business stakeholders in "lunch-and-learns" to expect many loops through the process.

Enable compounding collaboration. High-performing data science teams realize they work best when they stand on the shoulders of those before them. One data science leader even goes so far as to track component reuse as a KPI. Data scientists who create widely used components (e.g. a great dataset diagnostic tool) are given visibility and credit for their contributions to the community's success.

Anticipate auditability needs. As more and more critical processes incorporate data science results, it is essential to be able to audit and inspect the rationale behind the model. The financial industry is formally regulated under "model risk management ". Yet other industries are also taking proactive steps to build model risk expertise and preserve all relevant artifacts associated with the development and deployment of a model. More recently, there is speculation that technology firms could follow suit to preserve model integrity.

.

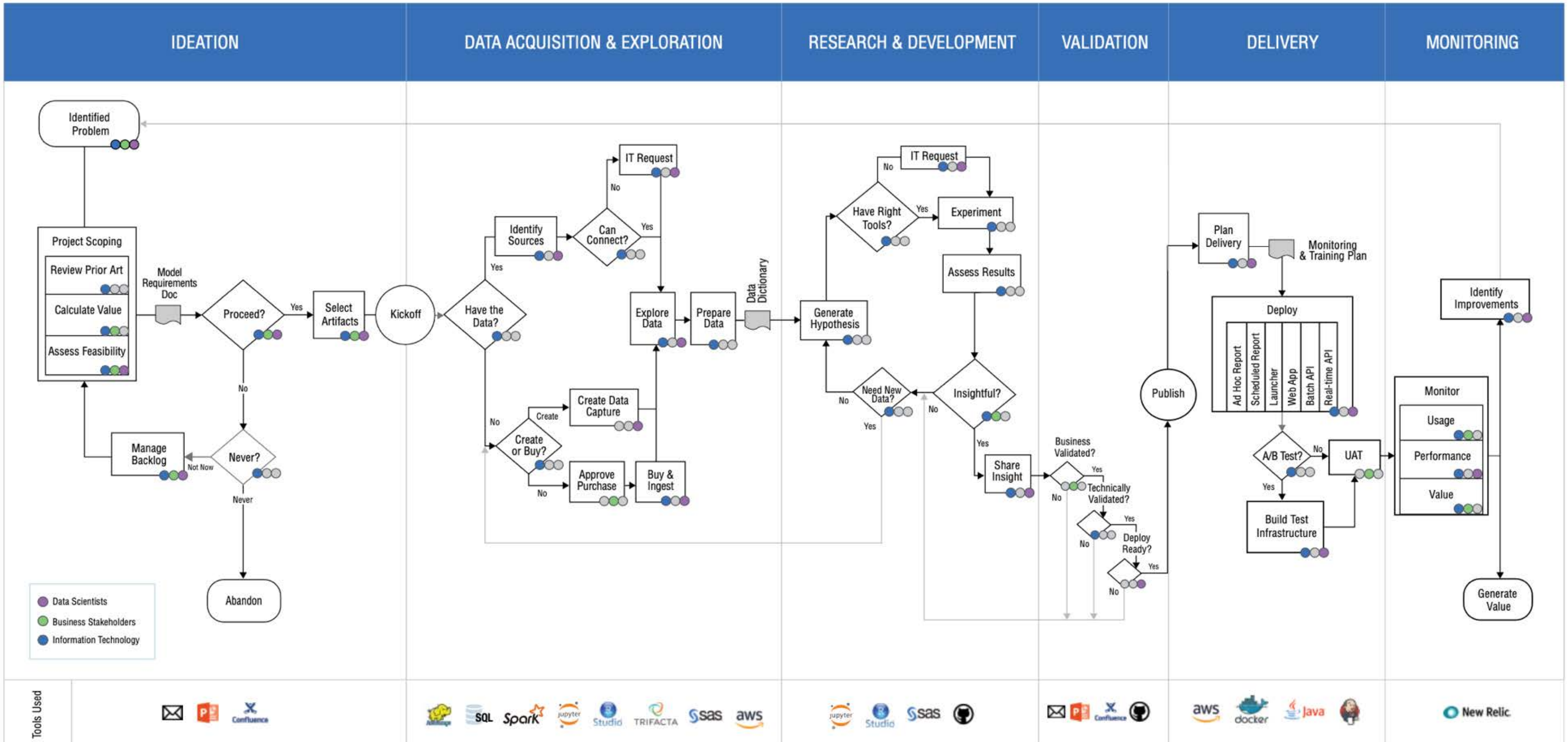

Figure 3.58 Data Science Lifecycle (Domino).

People

A successful data science machine comes from the coordination of people, process, and technology. We've gone in depth through the processes we've seen drive success, but in this section, we examine the people. As mentioned above, the full stack data scientist no longer exists, and the data science roles are increasingly specialized. This is a natural evolution that we expect will continue as data science becomes ingrained into the fabric of how organizations function.

The most consistent feedback we've heard is the increasing demand for a "product manager" type role as most organizations move from delivering mathematical results to stakeholder-facing apps. In large tech organizations, data science sits peer with product management to drive strategic priorities and ongoing optimization of engagement and impact. Below is a list of the types of roles across successful organizations. This is by no means definitive and actual titles can vary but this represents the broad shape of responsibilities we've heard.

Role	Responsibility	Pitfalls without them?
Data Scientist	Generating and communicating insights, understanding the strengths and weaknesses of algorithms and features.	Naive, or low power insights.
Data Infrastructure Engineer	Building scalable pipelines and infrastructure that make it possible to develop and deploy models.	Insight generation is slow because data scientists are spending their time doing infrastructure work.
Data Product Manager	Responsible for clearly articulating the business problem, at hand, connecting through domain knowledge about the business problem and translating that into the day to day work. Also, ensuring training and ongoing engagement with deployed models.	Projects miss the mark. The data scientists spend their time playing with math, model/features selections are mathematically valid but ultimately domain-divorced.
Business Stakeholder	Responsible for vetting the prioritization, ROI, and providing subject matter expertise throughout.	ROI decisions aren't made sensibly, not knowing when to pull the plug, non-actionable results.
Data Storyteller*	Creating engaging visual and narrative journeys for analytical solutions. Somewhat analogous to a designer.	Low engagement and adoption from end users.

Figure 3.59 Roles to The Practical Guide to Managing Data Science at Scale (Domino).

Analytics Solutions Unified Method (ASUM-DM)

ASUM focuses on the complete lifecycle of a data analytics project, from defining objectives and requirements to implementing, deploying and maintaining the solution. It provides structured guidance to help organizations develop effective analytics solutions and make the most of their data. The ASUM framework includes several key stages, such as project planning, data collection, modeling, validation and implementation. Each stage has its own specific activities and tasks that must be completed to effectively advance the project. ASUM is also based on project management principles and best practices, making it suitable for organizations that want to implement analytics solutions in a structured and controlled manner. It is important to note that ASUM is only one of many methodologies available for data analytics projects, and organizations can adapt or combine it with other methodologies according to their specific needs and requirements.

ASUM-DM is an extended and refined CRISP-DM for Data Mining/Predictive Analytics projects. CRISP-DM analytics' activities/tasks and data mining cycle (graph below) have been retained and accepted entirely but beefed up in the "Deployment" phase where CRISP-DM is at one of its weak points. Additionally Infrastructure/Operations and Project Management activities/tasks have been added and the method has been augmented by some rich and useful templates and guidelines. (See Project Lifecycle/Sitemap opposite for ASUM-DM activities).

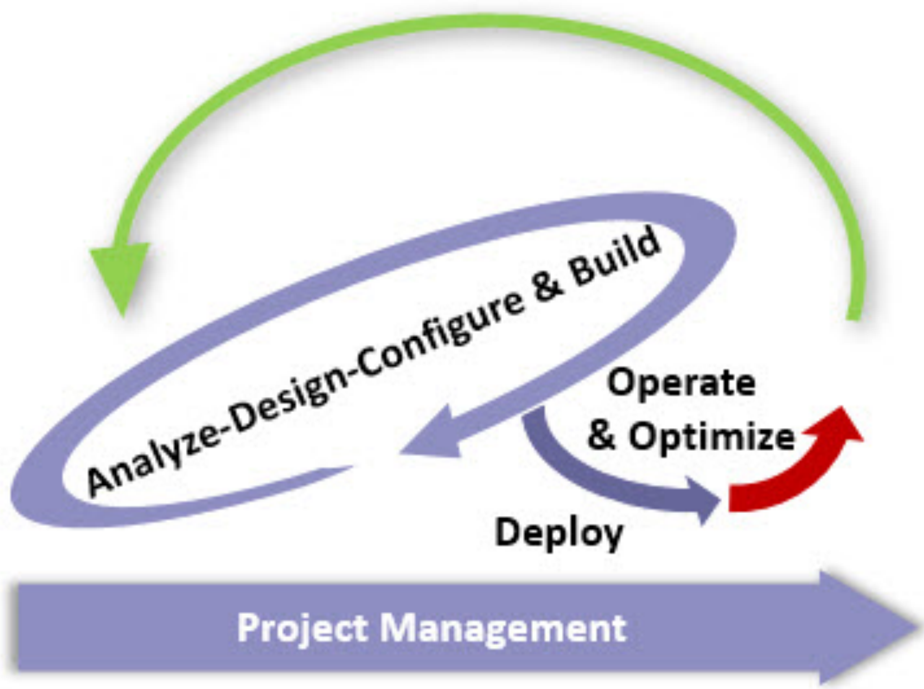

Figure 3.60 Diagram ASUM-DM from (IBM, ASUM-DM.)

IBM Analytics methodology "Analytics Solutions Unified Method (ASUM)" has five phases: Analyze, Design, Configure & Build, Deploy, and Operate & Optimize. However, the three phases of ASUM of Anlyze, Design, and Configure & Build have been combined here due to the iterative nature of data mining/predictive analytics projects. The method Work Breakdown Structure (WBS) already incorporates adequate project management elements but an additional optional Project Management Process has been added here for supplemental use when needed.

1.- Analyze-Design-Configure&Build

This is an iterative circular phase where customer's goals, expectations, and requirements are understood, data is understood and prepared, and models are built and evaluated.

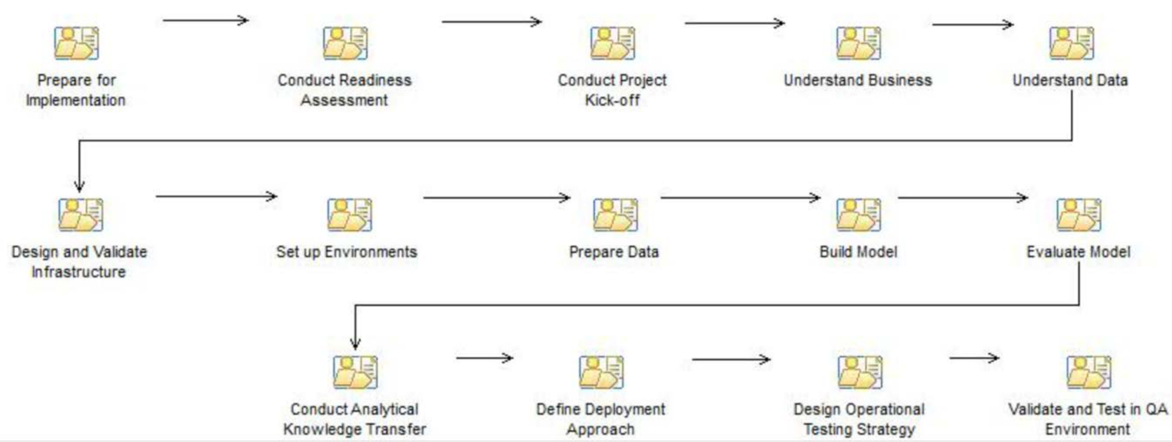

Figure 3.61 Elements of Analyze-Design-Configure&Build

2.- Deploy

The Deploy phase puts the solution in the hands of the users and prepares for its continuing operation. This phase can be applied to a single user group deployment, or as a deployment of the solution to a global audiance.

Figure 3.62 Elements of Deploy

3.- Operate & Optimize

This phase represents the use and lifecycle of the IBM solution. Operate & Optimize includes the maintenance tasks and checkpoints after roll out that facilitate a successful employment of the solution and preserve its health through its lifecycle.

Figure 3.63 Elements of Operate & Optimize

Roles

- Client Application Administrator
- Client Business Sponsor
- Client Data Analyst
- Client Database Administrator
- Client Key System Users
- Client Network Administrator
- Client Project Manager
- Client Security Administrator
- Client Stakeholders
- Client Subject Matter Expert
- Client Support Manager
- Client Tool Administrator
- Data Miner/Data Scientist
- Enterprise Architect
- SPSS Project Manager

The Analytics Solutions Unified Method (ASUM-DM) is a methodology that focuses on the entire lifecycle of a data analytics project, from the definition of objectives and requirements to the implementation, deployment and maintenance of the solution. It provides structured guidance to help organizations develop effective analytics solutions and get the most out of their data. ASUM-DM is a refined extension of CRISP-DM, designed specifically for data mining and predictive analytics projects. It combines the activities and tasks of CRISP-DM with additional elements in the implementation phase, which is often considered a weakness in CRISP-DM. Infrastructure/operations and project management activities are also added, along with useful templates and guidelines. ASUM-DM is suitable for organizations that wish to implement data analytics solutions in a structured and controlled manner. It is important to note that ASUM-DM is only one of many methodologies available for data analytics projects, and organizations can adapt or combine it with other methodologies according to their specific needs and requirements.

For the selection of these 3 Agile methodologies, we also relied on the work of Martinez, Viles and Olaizola from 2021 where they conducted an analysis of different Data Science Methodologies where they looked for 3 factors: Team Management, Project Management and Data & Information Management. In this research they analyzed 19 methodologies looking for their characteristics, this was used as support to be able to understand the complexity and detail of the existing details being able to focus on those that were more useful for our light approach. Having these three methodologies as a study for the agile ones, we will take two of them leaving out Domino, we will discard it for future comparative tables because although the quality of the methodology meets the requirements we are looking for being a private methodology, its documentation that can be found without paying for it is little but still with the information that was found was useful for certain key points in the future such as the importance of data in the process and how it is grouped into the 3 main roles giving focus to one of these as the data specialist. Considering the above, the other two methodologies will be compared against ISO 29110 to identify the common points of our standard but at the same time to detect the areas to be strengthened.

In conclusion after reviewing the different agile methodologies related to Analytics/Data Science currently existing and even after selecting 3 and studying them in detail in order to understand what their contributions are for this type of projects, it can be given as closure of the research that indeed the Analytics/Data Science projects require a different cycle where we could clearly identify a segment focused entirely on data, for example in TDSP the Data Acquisition & Understanding and modeling, for the case of Domino Data Acquisition & Exploration, Research & Development and Validation, ending with the Larson & Chang methodology in Data Acquisition/Discover, Analyze/Visualize , Model/Design/Develop where they also focus on data, without forgetting that also most of these before software development have model validation for data. This makes great emphasis in determining that the improvements that are sought to be made to the ISO 29110 standard to achieve a higher percentage of effectiveness with projects related to Analytics/Data Science should be based primarily on the process of the data rather than the development as we can see in the agile methodologies that even with its focus on the part of the data is detailed and calm, instead in the part of the final development of the software if the agile concept is applied with its original approach.

The final step for the two selected agile methodologies will be done in the following section where tables will be generated to compare them against the requirements of the ISO 29110 standard, we will seek to observe the common characteristics that have these methodologies but also the areas that do not share to identify these areas of improvement for our standard when it comes to Analytics/Data Science projects. These comparisons will be made for the roles, artifacts, and activities of the different methodologies, including the comparison with CRISP-DM, which was of the classic methodologies the one that stood out the most in terms of usability in the current times as well as in the characteristics of optimal detail for a standard.

3.2 ANALYSIS OF CONTRIBUTIONS AND LIMITATIONS.

Table 3.20 Contributions vs Opportunities of improvement of the Theoretical Background

	Contributions	Opportunities of improvement
Software Engineering	Oktaba & Ibargüengoitia (1998) helped to define the fundamental concept for software processes where they defined the minimum set of phases, activities-tasks, roles, resources, and artifacts that cover the entire process. "Software Engineering Process consists of a set of interrelated activities that transform one or more inputs into outputs" (Bourque et al. 2014), likewise Bourque mentions that there are different topics where the software life cycle is a central area for the generation of software development models.	IEEE defines Software Engineering as converting user needs into a software product. The process involves translating user needs into software requirements, transforming the software requirements into design, implementing the design into code, testing the code and sometimes installing and testing the software for operational use (SO/IEC/IEEE 2465:2017, 2017), combined with Bourque's research where he detects the importance of the life cycle in a project, which must be related to the type of project and its characteristics processes, equipment and personnel (Bourque et al. 2014), leaving clearly the potential of using Software Engineering for Data Science / Analytics projects (Davoudian & Liu, 2020).
ISO/IEC 29110	The application and appropriate use of standards should increase the productivity of an organization and have a direct positive impact on the organization. (C. Y. Laporte et al., 2018) In research related to the use of standards in VSEs, they expressed their need for help in the adoption and application of engineering standards, which led to the generation of standards focused mainly on this type of organizations or working groups. Most VSEs considered it important to be evaluated or certified according to a standard, achieving improvements such as increased competitiveness, confidence, and customer satisfaction; improved quality of software products; facilitated commercialization and increased export potential; and decreased development risk. (C. Y. Laporte & Miranda, 2020)	A major challenge is that the knowledge documented in standards reaches an organization and is applied to its benefit (C. Y. Laporte et al., 2018). Most of the respondents indicated that they would like to receive more guidance through examples, additionally, it was noted that they asked for a light and easy-to-understand standards. More than 15% of respondents thought that engineering standards were difficult and bureaucratic and did not provide adequate guidance for use in a small environment. (C. Y. Laporte & Miranda, 2020), among other complex characteristics involved in the use of a standard (Claude Y. Laporte & Munoz, 2021)

BDAS Dev Met	Data Science / Analytics is an interdisciplinary field whose objective is to convert data into value, where data is transformed into knowledge to make better decisions, using statistical and quantitative analysis. It is important to highlight the three pillars of data science: data, technology, and people (Song & Zhu, 2016). In analytics we can indicate that data analysis projects can be divided into several phases, where these are clearly focused on the data. Data are evaluated, selected, cleaned, filtered, visualized, and analyzed, and finally interpreted and evaluated (Runkler, 2020). These projects have a specific focus within analytics, which can be: Descriptive analytics, Predictive analytics, Predictive analytics (Watson, 2014). Big Data were discovered, the 5V, Volume, Variety, Velocity, Veracity and Value. A Big Data management architecture should be able to design systems and models for the processing of large volumes of data from innumerable data sources in a fast and economical way, which allows better decision-making (Chang et al., 2019).	NASA researchers since the late nineties made mention of the term "Big Data", referring to the problem that this entails, when they report the complexity in computer systems where the data sets are often quite large, straining the capacity of main memory, local disk, and even remote disk (Cox & Ellsworth, 1997). Even in current times with the increase in processing speed or storage capacity, Microsoft mentions the complexity and breadth of the components of an architecture for Big Data (Microsoft, 2021). Even in today's times with increasing processing speed or storage capacity, Microsoft mentions the complexity and breadth of the components of an architecture for Big Data (Microsoft, 2021). This is also related to the growth and the ever-changing form of information generation (Palfreyman, 2013). This complicates the architecture model for a project that wants to meet these requirements. The theory says that big data must comply with certain characteristics in the data, but in practice it was possible to apply it to projects with data that do not fully meet the big data requirements, but in the same way it was possible to obtain the benefits. The complexity of this type of projects leaves a wide area for the generation of more specialized methodologies in them and their key points of success.
Rigor methodologies	In response to the common problems and needs in data mining projects in the mid-1990s, a group of data mining organizations (Teradata, SPSS -ISL-, Daimler-Chrysler and OHRA) proposed a reference guide for developing data mining projects, called CRISP-DM (CRoss Industry Standard Process for Data Mining) (Chapman et al., 2000). Clearly, it was not the only methodology of this type, but the main difference of CRISP-DM with respect to others such as KDD and SEMMA are that this methodology is much completer and more detailed, making it the most used and efficient for many.	The CRISP-DM methodology is more than 20 years old, so it can be updated to the technological changes that have occurred in these years. Additionally, one of the clearest flaws of CRSIP-DM is the lack of definition and roles within its specifications and documentation (Chapman et al., 2000). Therefore, it clearly leaves an area for the creation of methodologies that seek to cover these characteristics, action that although there are several new methodologies, it is still not possible to standardize the use of one of these that meets all the points of the project taking a more holistic approach to development (Martinez et al., 2021). Seeking to cover the aspects of artifacts, roles, and activities.

	The continuous evolution of methodologies seeking the specialization of the type of project managed to optimize the key points in data-centered projects, generating methodologies with detail as CRISP in the specification of tasks and artifacts (Martinez et al., 2021).	
Agile methodologies	Because of Agile's imperative to deliver working software early and often, there is an early return on investment, and early feedback allows development to reach the level of a minimum viable product (MVP) more quickly (Grady & Chang, 2017). Since the DSA manifesto was published, the objectives and principles have been interpreted and applied to new agile methodologies. The most popular approaches from which the manifesto and its principles were derived were: Extreme Programming (XP) and Scrum, where they are currently successfully practiced and considered standard development methodologies (Hsieh and Chen, 2015). Methodologies continue to be generated with more specific characteristics for each type of projects, highlighting as a key point in the methodologies a specific segment for data collection, cleaning, analysis, preparation, modeling and evaluation in BDAS related projects.	There is a lack of use of a methodology in real data science projects and there is an absence of complete or comprehensive methodologies throughout the literature. Among these "comprehensive" methodologies, some aspects can be improved: such as TSDP which has a strong dependency on Microsoft tools and technologies and its functions are too narrowly focused on Microsoft services. For example, the role of the TDSP data scientist is limited to the cloning of various repositories and to the mere execution of the data science project. The role of the data scientist should be broken down into more tangible functions and detail their responsibilities outside the Microsoft universe (Martinez et al., 2021). Basic practices of agile methodologies include small, short releases; physically co-located stakeholders; and a limited project cycle time (typically 60 to 90 days, although the cycle may be shorter depending on the deliverable) (Kendall & Kendall, 2005). This should be analyzed to see how much coercion it has with the ideal structure for Data Science / Analytics projects (Larson, D., & Chang, V., 2016). In this sense, the three cornerstones of these methodologies have been presented: project, team and data and information management. It is important to note that this framework must be constantly evolving and improving to adapt to the new challenges of data science (Martinez et al., 2021). It seems to be true that data science projects are very difficult to manage due to, among other factors, the uncertain nature of the data, so their project failure rate is very high. One of the most critical points is that at the beginning of any project, data scientists are not familiar with the data, and therefore it is difficult to know the quality of the data and its potential to achieve certain business objectives (Martinez et al., 2021).

By analyzing the contributions it is possible to have clear the basis for the generation of the methodology that is to be performed contemplating the key points of the theory and practice analyzed as clearly detect the life cycle depending on the key features (the first sprin is dedicated "totally" to the data), the classics do not have so much detail in the specifications that in this case are the data (Haakman et al... 2021), software engineering should be applied in activities such as: requirements engineering, design and construction of software to meet the specified requirements, and quality assurance of software and data (Davoudian & Liu, 2020). Data analysis should be one of the main approaches as it aims to derive meaning and insights from data (Bertolucci, 2013).

In the study of both traditional and agile methodologies, new methodologies continue to be developed, this generated by changes in technologies or project characteristics, giving way over time to changes in the way of managing a project, data, or roles (Martinez et al., 2021), 2021). But there are still possible improvements in the methodologies that currently exist which motivates us to create a methodology that is aligned to the objectives we want which is an ISO 29110 optimized for Data Science / Analytics projects, taking care of the details of: project, team and data and information management, taking into account the details already previously detected in the existing ones, in the same way that the complexities involved in the use of standards (Claude Y. Laporte & Munoz, 2021).

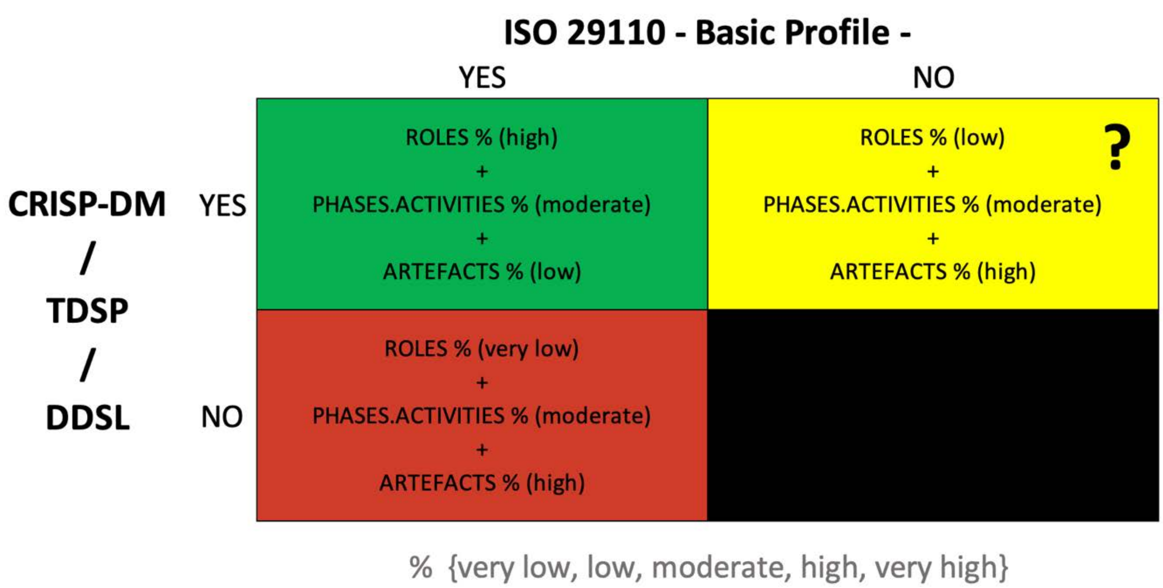

Figure 3.64 Intersection between ISO 29110 vs CRISP-DM / TSDP / DDSL

The previous image is the summary of the research for the closing of this segment, where what was sought was to generate comparative tables between the ISO 29110 standard against 3 selected methodologies, for the part of the traditional methodologies the most appropriate was CRISP-DM, therefore it was selected for comparison, with respect to agile methodologies the comparison was made against Microsoft's TSDP, and Domino's DDSL. In these tables what was compared between the methodologies and the standard were the roles, activities and artifacts mentioned or defined in each of these, the reason for this analysis was to find the similarities and differences in each of the aspects mentioned above. For the crossing of these, it was selected if they had any similarity in some detail or not, but at the same time within the table it was indicated how much was the similarity between those two points, contemplating different regions.

At the end of the generation of these 3 analyses we chose to select the 3 CRISP-DM by the traditional ones, TDSP by the agile ones and DDSL for being lightweight, these were selected for different particularities as clear similarities against the standard that will serve as a basis, additionally they have an extensive and detailed documentation that just for what was observed is necessary a wide detail in the methodologies to have a greater impact to optimize the ISO 29110 in projects related to Data Science / Analytics.

In the intersection between ISO 29110 vs CRISP-DM / TSDP / DDSL (Figure 3. 57) we conclude with 3 points of intersection, the first one where SI and SI are the points in common in roles, activities and artifacts, the percentage of similarities was calculated, which clearly reflects that the highest intersection was with roles, because there were few roles that can be differentiated between the standard and the methodologies, but the relationship that we have in the artifacts in this segment is low due to the great difference of the products generated in the methodologies due to the focus that they handle, for example CRISP-DM has a large number of artifacts that focus on data, while ISO 29110 most of the artifacts are for project management, but this segment will be our basis for the confirmation that ISO 29110 is functional for Data Science / Analytics projects because we can observe a great similarity between them in order to achieve a higher level of success. For the second segment the roles, activities and artifacts of the ISO 29110 that are not found in the other methodologies are contemplated, this intersection of red color are the points that must be maintained to continue complying in our methodology with the indicated for the standard and not to lose the qualities that entails to comply with a standard for our methodology, for that reason the red color to understand that also those criteria must be fulfilled to be able to validate the methodology with the standard. It should be made clear that what is proposed will be a series of recommendations for ISO 29110 seeking to achieve an improvement when applying this methodology in Data Science/Analytics projects, so to continue complying with the standard it must be followed in its entirety. For the last segment in yellow, this indicates the roles, activities and artifacts that are mentioned in the analyzed methodologies, but not in the ISO 29110, this intersection will be the one that we will have to analyze in depth since from this difference we will take only the critical factors that help the success of the use of the generated methodology. It is for this reason that the image shows the symbol '?' which indicates that we must select which roles, activities or artifacts would be helpful for our improvement of ISO 29110, remembering that the key is: to develop a methodology for Data Science / Analytics projects of big data that is considered as light (neither agile nor rigorous), easy to use, useful, compatible, and valuable, complying with the standard.

The basis of the methodology will be the ISO 29110 using as format the TDSP/ DDSL structure where you can clearly notice the data process first and then the development where it says that it can be used with agility, this to understand that this type of projects will really focus on two main areas, where for the data area will be used the theory of CRISP-DM this due to the similarity of detail that handles with the standard and software development will follow the process of ISO 29110. With this what we are looking for is to increase the detail of the standard, not to change remembering that we want to comply in its totality with what is mentioned in the standard but only to optimize the methodology and to clarify the key points for the success in this type of projects. At the end of the generation of the new methodology, the aim is to be able to apply it to a prototype project where it will be evaluated by different experts in the area and to identify its final functionality.

CHAPTER 4. DEVELOPMENT OF THE SOLUTION

As mentioned in Chapter 2 this Ph.D. The dissertation uses mainly the Design Science Research Methodology (DSRM) (Peffers et al., 2007) which is detailed in Table 2.2 and is divided into the next steps:

1. DSRM step 1- Design problem identification and motivation.
2. DSRM step 2 - Definition of the Design Objectives, Design Restrictions, Resign Approach, Design Theoretical Sources, and Design Components for the expected Artifact.
3. DSRM step 3 – Design and development of the artifact.
4. DSRM step 4 – Demonstration of the artifact (Proof of Concept).
5. DSRM step 5 – Evaluation of the artifact.
6. DSRM step 6 – Communication of research results.

4.1 DSRM STEP 1 – DESIGN PROBLEM IDENTIFICATION AND MOTIVATION

Chapter 1 of this document contains all the detailed information for Problem Identification and its Motivation.

4.2 DSRM STEP 2 - DEFINITION OF THE DESIGN OBJECTIVES, DESIGN RESTRICTIONS, RESIGN APPROACH, DESIGN THEORETICAL SOURCES, AND DESIGN COMPONENTS FOR THE EXPECTED ARTIFACT: Light Data Science - Analytics Methodology (LDSAM)

4.2.1 Definition of the design objectives.

The expected Design Objectives (DOs) to be archived in this work are:

1. DO.1 The designed artifact provides a lightweight (i.e. responsive, flexible, speedy, lean, simple, lightweight, and fine-grain documented (Conboy, 2009), (Qumer & Henderson-Sellers, 2008)) workflow—i.e. a value stream—for designing, building, and implementing a new minimum viable Light Data Science - Analytics Methodology.
2. The designed artifact is useful, easy to use, and valuable (Galvan et al., 2021) for small companies, software developers, and IT practitioners.

3. The designed artifact is fine-grain documented including the roles-set component, phases-activities set component, and artifacts-templates-set component.

4.2.2 Design Restrictions.

For Design Restrictions (DRs) we need to consider parameters such as time, budget, theoretical sources, and available software. The DRs that were agreed are:

1. DR.1 The designed artifact must be composed from design building-blocks from relevant design theoretical sources (DTS's).
2. DR.2 The designed artifact must be designed in short-term period (at most 6 months) and under the assigned research budget.
3. DR.3 The designed artifact must be documented in an Electronic Process Guide.

4.2.3 Design theoretical sources.

The Design Theoretical sources (DTSs) are the key sources of the design components that will be chosen to create the artifact. The DTSs selected were proposed based on the theoretical background and having reviewed the seven SDLC methodologies and the standard.

Table 4.1 The Design Theoretical sources

DTS.1	ISO/IEC 29110 -Basic profile- (ISO/IEC, 2011)
DTS.2	CRISP-DM: Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000)
DTS.3	TDSP: The Team Data Science Process (Microsoft, 2016)
DTS.4	DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)

Every single element such as Roles, Activities, and Artifacts for the DTS will be considered and discussed with the team to get the Desing Components.

4.2.4 Design components for the expected artifact.

Evaluating very carefully the DTS, we have selected the potential design components (DCS) to be used in the design of the artifact. Some components may be not used in the final design. Table 4.2, Table 4.3, Table 4.4, and Table 4.55 have all the Design Components selected from the four DTS by the research team based on their experience and expertise. An iterative process is going to be performed to get de most important components to design the artifact.

Table 4.2 DTS.1 Theoretical rigorous SDLC for ISO/IEC 29110 -Basic Profile-

Design Compoent	Design theoretical source (DTS)	Specific elements of the design component (DC) potentially to be used in the designed artifact
DC.1 ISO/IEC 29110 -Basic profile- Process	DTS.1 ISO/IEC 29110 -Basic profile- (ISO/IEC, 2011)	{ Project Management process, Software Implementation process }
DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 -Basic profile- (ISO/IEC, 2011)	{ Project Management process [Project Planning, Project Plan Execution, Project Assessment and Control, Project Closure], Software Implementation process [Software Implementation Initiation, Software Requirements Analysis, Software Architectural and Detailed Design, Software Construction, Software Integration and Tests, Product Delivery]}
DC.3 ISO/IEC 29110 -Basic profile- Products	DTS.1 ISO/IEC 29110 -Basic profile- (ISO/IEC, 2011)	{PM input products [Statement of Work Software Configuration, Software Implementation, Change Request], PM output products [Project Plan, Acceptance Record, Project Repository, Meeting Record, Software Configuration], PM internal products [Change Request, Correction Register, Meeting Record, Verification Results, Progress Status Record, Project Repository Backup], SI input products [Project Plan, Project Repository], SI output products [Software Configuration (Requirements Specification, Software Design, Traceability Record, Software Components, Software, Test Cases and Test Procedures, Test Report, Product Operation Guide, Software User Documentation, Maintenance Documentation), Change Request], SI internal products [Validation Results, Verification Results]}
DC.4 ISO/IEC 29110 -Basic profile- Roles	DTS.1 ISO/IEC 29110 -Basic profile- (ISO/IEC, 2011)	{ Analyst, Customer, Designer, Programmer, Project Manager, Technical Leader, Work Team }

Table 4.3 DTS.2 CRISP-DM: Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000)

Design Compoent	Design theoretical source (DTS)	Specific elements of the design component (DC) potentially to be used in the designed artifact
DC.5 CRISP-DM Phases	DTS.2 CRISP-DM: Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000)	{Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment}
DC.5 CRISP-DM Tasks	DTS.2 CRISP-DM: Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000)	{Business Understanding [Determine Business Objectives, Assess Situation, Determine Data Mining Goals, Produce Project Plan], Data understanding [Collect Initial Data, Describe Data, Explore Data, Verity Data Quality], Data preparation [Select Data, Clean Data, Construct Data , Integrate Data, Format Data], Modeling [Select Modeling Technique, Generate Test Design, Build Model, Assess Model], Evaluation [Evaluate Results, Review Process, Determine Next Stages], Deployment [Plan Deployment,

		Plan Monitoring and Maintenance, Produce Final Report, Review Project]}
DC.7 CRISP-DM Outputs	DTS.2 CRISP-DM: Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000)	{Business Understanding [Background , Business Objectives , Business Success Criteria, Inventory of Resources , Requirements Assumptions and Constraints , Risks and Contingencies , Terminology, Costs and Benefits, Data Mining Goals , Data Mining Success Criteria, Project Plan, Initial Assessment of Tools and Techniques], Data Understanding [Initial Data Collection Report, Data Description Report, Data Exploration Report, Data Quality Report], Data Preparation [Rationale for Inclusion/ Exclusion, Data Cleaning Report, Derived Attributes , Generated Records, Merged Data, Reformatted Data, Dataset, Dataset Description], Modeling [Modeling Technique , Modeling Assumptions, Test Design, Parameter Settings , Models, Model Descriptions, Model Assessment , Revised Parameter Settings], Evaluation [Assessment of Data Mining Results w.r.t. Business Success Criteria , Approved Models, Review of Process, List of Possible Actions , Decision], Deployment [Deployment Plan, Monitoring and Maintenance Plan, Final Report, Final Presentation, Experience Documentation]}

Table 4.4 DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016)

Design Compoent	Design theoretical source (DTS)	Specific elements of the design component (DC) potentially to be used in the designed artifact
DC.8 TDSP Lifecycle	DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016)	{ Business understanding, Data acquisition and understanding, Modeling, Deployment, Customer acceptance }
DC.9 TDSP Tasks	DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016)	{ Business understanding [Define objectives,Identify data sources], Data acquisition and understanding [Ingest the data, Explore the data, Set up a data pipeline], Modeling [Feature engineering, Model training, Model Evaluation], Deployment [Operationalize a Model], Customer acceptance [System Validation, Project hand-off]}
DC.10 TDSP Artifacts	DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016)	{ Business understanding [Charter document, Data sources, Data dictionaries], Data acquisition and understanding [Data quality report, Solution architecture, Checkpoint decision], Modeling [Model], Deployment [A status dashboard that displays the system health and key metrics, A final modeling report with deployment details, A final solution architecture document], Customer acceptance [Exit report of the project for the customer]}
DC.11 TDSP Roles	DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016)	{Solution architect, Project manager, Data engineer,Data scientist, Application developer, Project lead}

Table 4.5 DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)

Design Compoent	Design theoretical source (DTS)	Specific elements of the design component (DC) potentially to be used in the designed artifact
DC.12 DDSL Lifecycle	DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	{ Ideation, Data Acquisition and Exploration, Research and Development, Validation, Delivery, Monitoring}
DC.13 DDSL Process	DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	{ Ideation [Identified Problem, Project Scoping, Review Prior Art, Calculate Value, Assess Feasibility, Manage Backlog, Select Artifacts], Data Acquisition and Exploration [Getting the Data, Identify Sources the Data, Connect, Create Data (Capture), Buy & Ingest DATA, Explore Data, Prepare Data], Research and Development [Generate Hypothesis, Validate right tools, IT request, Experiment, assess result, Validate the need new Data, Insightful?, Share insight], Validation [Validate the Business, Validate Technically, Validate ready to Deploy, Publish], Delivery [Plan Delivery, Deploy, Test], Monitoring [Monitor, Usage, Performance, Value, Identify Improvements, Generate Value]
DC.14 DDSL Products	DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	{ Ideation [Project Scope document, Project Kick-off, Model Requirements Doc], Data Acquisition and Exploration [Data Dictionary], Research and Development [*Data Model Experiment], Validation [*Validated Data Model], Delivery [*Production Data Model], Monitoring [Monitoring & Training Plan]
DC.15 DDSL Roles	DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	{Business Stakeholder, Data Product Manager, Data Scientist, Data Infrastructure Engineer, Data Storyteller}

4.3 DSRM STEP 3 – DESIGN AND DEVELOPMENT OF THE ARTIFACT

To design the BPMS Methodology the research team applied Means-Ends Analysis heuristic (Newell & Simon, 1972) (Greeno et All,1987) in four steps:

- Step 1. To represent the design problem defining an initial state S i, a pursued final state S f, a set of heuristic operators {HOx(Sy, Sz), …} that can transform the state Sy to the state S z, a set of design objectives {DOj, …} and design restrictions {DR k, …} expected to be satisfied by the final state S f, and two qualitative functions EvalDOs(DO's) and EvalDRs(DR's) to evaluate the logical satisfaction of DO's and DR's.
- Step 2. To set up the initial state S i and the pursued final state S f, and determine the initial qualitative evaluations EvalDOs(DO's) and EvalDRs(DR's) for the initial state Si and the pursued final state Sf.
- Step 3. To apply a sequence of heuristic operators {HO?(Si , S2); HO?(S2, S3); …; HO?(S?, Sf)} based on a logical analysis of the operators that can transform the initial state S i in the pursued final state Sf.

- Step 4. To evaluate the level of compliance of the pursued final state S f, regarding the design objectives {DOj, …} and design restrictions {DRk, …}.

The first step was to select the Design Components from the DTS, with the first batch of DCs the research team discussed the importance of each component. In the third iteration, the DCs that were already covered by DTS.21 (ISO/IEC 29110 -Basic Profile-) were eliminated and complemented with DCs from other DTSs.

Appendix C has all the information about this process with the first and second iterations of the selected Design Components. Table 4.6, table 4.7, and table 4.8 display the final selected DCs for roles, activities/tasks, and products. Figure 4.1 displays the final BDAS Methodology with all selected Desing Components.

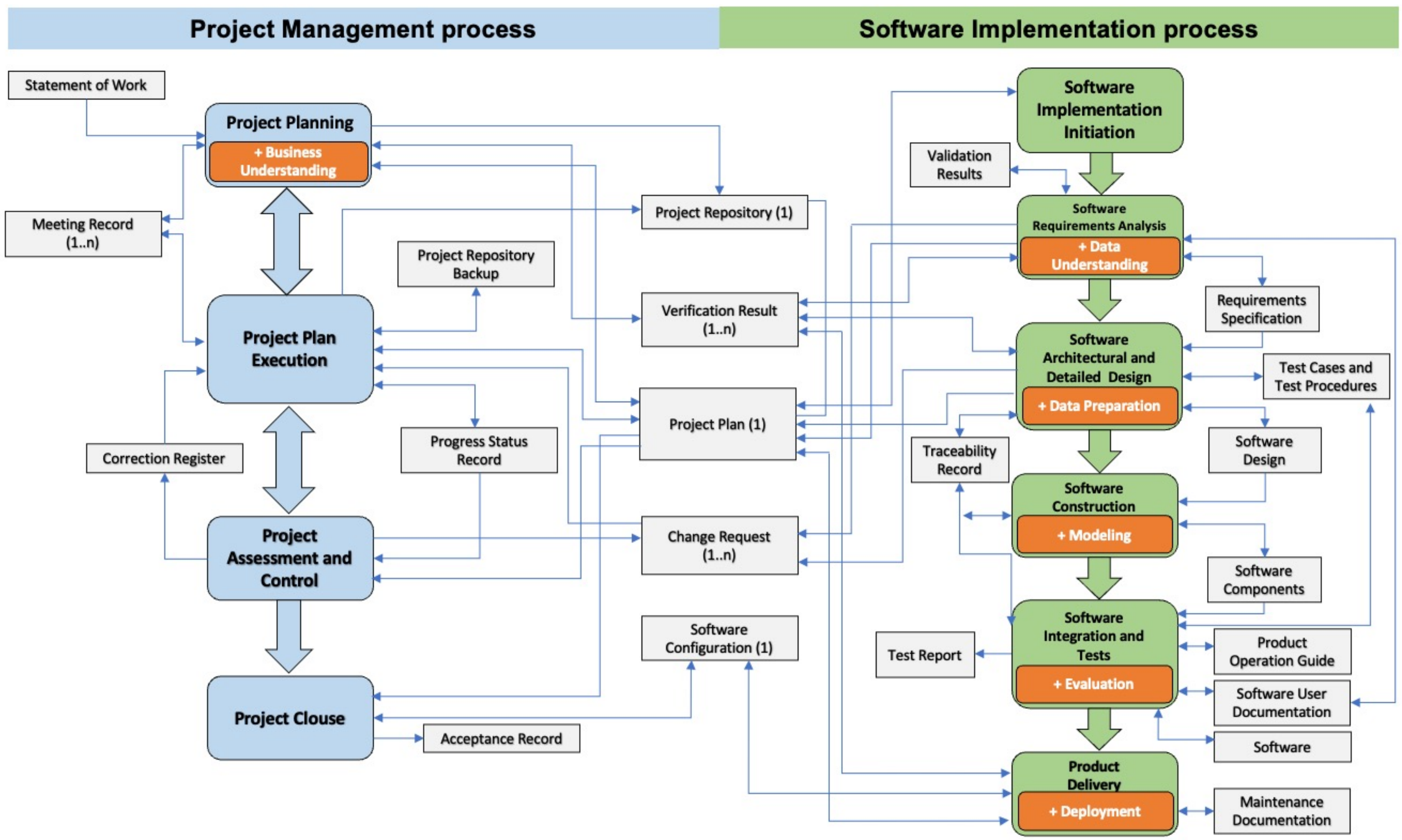

Figure 4.1 BDAS Methodology Conceptual Map.

Table 4.6 Final Design Components for roles.

Roles						
Design Component	**Source**	**Why this could be helpful**	**SDLC that is also using it**			
			DTS.1	**DTS.2**	**DTS.3**	**DTS.4**
DC.4 ISO/IEC 29110 - Basic profile-Roles	DTS.1 ISO/IEC 29110 -Basic profile- (ISO/IEC, 2011)	**R.1 Customer**: Responsible for reviewing prioritisation, return on investment and providing expertise throughout the process. Confirm that the pipeline, the model and its deployment in a production environment meet the objectives.	O		/	X
		R.2 Project Manager: Manages the day-to-day activities of the Work Team on a specific data science project. Responsible for clearly articulating the business problem, at hand, connecting through domain knowledge about the business problem and translating that into day to day work. In addition, ensure training and continuous engagement with the deployed models.	O		X	X
		R.3 Work Team: Data Scientists, Business Analysts, Data Engineers, Architects, and others who execute a data science project.	O		X	
		R.4 Technical Leader: The Data Infrastructure Engineer (Data & Platform Architect) Building scalable pipelines and infrastructure that make it possible to develop and deploy models.	O		X	X
		R.5 Programmer: The Data Scientist Generating and communicating insights, understanding the strengths and weaknesses of algorithms and features.	O		X	X
		R.6 Desinger: The Data Engineer to perform the data engineering parts of the project. Responsible for building and maintaining the data infrastructure. This includes extraction, transformation and loading (ETL) of data, creation of data pipelines,	O		/	
		R.7 Analyst: Data Analyst subject matter experts who have a clear understanding of the problem. They must know the internal processes and practices so that they can help the analyst understand and interpret the data. They must also be able to make the necessary changes to existing business processes to help collect the right data for the problems, if needed.	O		X	

Table 4.7 Final Design Components for Process and Activities.

Phases and Activities						
Design Component	**Source**	**Why this could be helpful**	**SDLC that is also using it**			
			DTS.1	**DTS.2**	**DTS.3**	**DTS.4**
DC.1 ISO/IEC 29110 -Basic profile- Process	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Project Management process**: The purpose of the process is to establish and carry out in a systematic way the Tasks of the software implementation project, which allows complying with the project's Objectives in the expected quality, time and costs.	O			
DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Activity PM.1 Project Planning**: documents the planning details needed to manage the project.	O			
DC.5 CRISP-DM Tasks DC.9 TDSP Tasks DC.13 DDSL Process	DTS.2 CRISP-DM Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000) DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016) DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	**Activity PM.1.+ Business Understanding**: This activity focuses on understanding the project's objectives and requirements from a business perspective (working with the client and other stakeholders to comprehend and identify the issues), establishing success criteria, and identifying relevant data sources. This knowledge is then translated into a definition of the data analysis problem and a preliminary plan designed to achieve the objectives. *Tasks*: • Determine business objectives. • Determine data mining goals. • Assessment of the data situation. • Produce project plan. • Define work team.		X	X	/
DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Activity PM.2. Project Plan Execution**: implements the documented plan on the project.	O			
DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Activity PM.3. Project Assessment and Control**: evaluates the performance of the plan against documented commitments..	O			

DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Activity PM.4. Project Closure**: rovides the project's documentation and products in accordance with contract requirements..	O			
DC.1 ISO/IEC 29110 -Basic profile- Process	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Software Implementation process**: The purpose of the Software Implementation process is the systematic performance of the analysis, design, construction, integration and tests activities for new or modified software products according to the specified requirements.	O			
DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Activity SI.1 Software Implementation Initiation**: Initiation activity ensures that the Project Plan established in Project Planning activity is committed to by the Work Team.	O			
DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Activity SI.2. Software Requirements Analysis**: analyzes the agreed Customer's requirements and establishes the validated project requirements.	O			
DC.5 CRISP-DM Tasks DC.9 TDSP Tasks DC.13 DDSL Process	DTS.2 CRISP-DM Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000) DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016) DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	**Activity SI.2.+ Data Understanding (acquisition)**: This activity involves understanding and assessing data availability, considering the relationship between ease of access and cost of acquisition. It begins with initial data collection to become familiar with the data and analyze it to detect quality issues, generate initial insights, and form hypotheses. The goal is to produce a clean, high-quality dataset, placed in the appropriate analytical environment and ready for modeling. Additionally, a solution architecture is developed to allow for regular updating and scoring of the data. *Tasks*: • Data Collection. • Describe data. • Explore data. • Verify data quality.		X	X	X
DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Activity SI.3. Software Architectural and Detailed Design**: transforms the software requirements to the system software architecture and software detailed design.	O			
DC.5 CRISP-DM Tasks	DTS.2 CRISP-DM Cross Industry Standard Process	**Activity SI.3.+ Data Preparation**: The data preparation activity comprises all activities aimed at building the final		X	x	x

DC.9 TDSP Tasks DC.13 DDSL Process	for Data Mining (Pete Chapman et al., 2000) DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016) DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	dataset (the data to be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in a predefined order. These tasks include the selection of tables, records, and attributes, as well as the transformation and cleaning of the data for the modeling tools. *Tasks*: • Select data. • Clean data. • Construct data. • Integrate data. • Format data. • Solution architecture.				
DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Activity SI.4. Software Construction**: develops the software code and data from the Software Design.	O			
DC.5 CRISP-DM Tasks DC.9 TDSP Tasks DC.13 DDSL Process	DTS.2 CRISP-DM Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000) DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016) DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	**Activity SI.4.+ Modeling (Bulding)**: various modelling techniques are selected, applied and optimised to create an accurate and production-ready machine learning model. Key tasks such as model selection, model training and evaluation of the model's suitability for production are performed. It is critical to iterate between data preparation and modelling to fine-tune features and ensure that the model meets specific requirements and best practices. It is recommended to start with simple models and avoid unnecessary complexity. *Tasks*: • Select modeling technique. • Generate test design for models. • Build model. • Evaluate and Select the model.		X	X	/
DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Activity SI.5. Software Integration and Tests**: ensures that the integrated Software Components satisfy the software requirements.	O			
DC.5 CRISP-DM Tasks	DTS.2 CRISP-DM Cross Industry Standard Process	**Activity SI.5.+ Evaluation (Model)**: After creating a high-quality model, it is crucial to thoroughly evaluate it		X	/	X

DC.9 TDSP Tasks DC.13 DDSL Process	for Data Mining (Pete Chapman et al., 2000) DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016) DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	and review its construction to ensure that it meets business objectives and does not omit important aspects. This phase involves training and evaluating the model on different datasets, performing rigorous validation of assumptions, code, performance and results. *Tasks*: • Evaluation of the results of the Model. • Evaluation of the process. • Final decision of the Model.				
DC.2 ISO/IEC 29110 -Basic profile- Activities	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	**Activity SI.6. Product Delivery**: provides the integrated software product to the Customer.	O			
DC.5 CRISP-DM Tasks DC.9 TDSP Tasks DC.13 DDSL Process	DTS.2 CRISP-DM Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000) DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016) DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	**Activity .6.+ Deployment**: After the creation of the model, it is crucial to organise and present the acquired knowledge in a useful way for the customer, either by generating reports, integrating the model into decision-making processes, or deploying it in a production environment. The main objective is to operationalise the model and its data pipeline for final acceptance by the user. Delivery of the model can range from publishing reports to automating decisions in production systems. Maintaining consistency between all deliverables is essential to avoid loss of key information and to ensure a successful deployment. *Tasks*: • Implementation Plan and Deployment. • Monitoring and Maintenance Plan. • Final report.		X	X	X

Table 4.8 Final Design Components for Products.

Products						
Design Component	**Source**	**Why this could be helpful**	**SDLC that is also using it**			
			DTS.1	**DTS.2**	**DTS.3**	**DTS.4**
DC.3 ISO/IEC 29110 -Basic profile- Products	DTS.1 ISO/IEC 29110 -Basic profile- (ISO/IEC, 2011)	**Acceptance Record**	O			
		Change Request	O			
		Correction Register	O			
		Maintenance Documentation: Deployment plan, Operationalizing the model, Monitoring and maintenance plan.	O	X		X
		Meeting Record	O			
		Product Operation Guide	O			
		Progress Status Record	O			
		Project Plan: Data Mining Goals, Inventory of Resources, Composition of Work Team, Project Plan (Tasks).	O	X		
		Project Repository	O			
		Project Repository Backup	O			
		Requirements Specification: Initial data collection report, Data description report, Data exploration report, Data quality report for feasibility.	O	X	X	X
		Software: Model(s)	O	X	X	
		Software Components: Modeling technique, Test design, Model(s), Model(s) Assessment.	O	X	X	/
		Software Configuration: Software Configuration	O	X		/
		Software Design : Rationale for inclusion/exclusion, Data cleaning report, Derived attributes, Generated records, Merged data, Reformatted data, Solution architecture.	O	X	/	/
		Software User Documentation	O			

		Statement of Work: Business Objective.	O	X	X	X
		Test Cases and Test Procedures: Evaluation results, Process and production review	O	X	X	X
		Test Report: Next step decision	O	X	X	
		Traceability Record	O			
		Verification Results: Final report.	O	X		/
		Validation Results	O			

4.4 DESIGN OF THE ELECTRONIC PROCESS GUIDE (EPG) FOR ISO/IEC 29110 -Basic Profile- for BDAS +.

In the previous section, we reported that we had chosen to design the new EPG for ISO/IEC 29110 -Basic Profile-. We called it ISO/IEC 29110 -Basic Profile- for BDAS + EPG. However, before building the EPG artifact, it was necessary to have the source content with an expected structure for EPGs.

Thus, to design a methodology that will be considered by software developers as lightweight (neither agile nor rigorous), easy to use, useful, compatible and valuable, based on the best practices provided by ISO/IEC 29110 - Basic Profile - adding the BDAS features highlighted in the other methodologies, both Roles, Processes and Activities as well as Products are described in the tables of chapter 4.3.

This process was thoroughly applied by the principal investigator and discussed as a team with the second investigator (PhD main advisor) and the external Ph.D. advisor. Several iterations were necessary for each adjustment from the general levels. Finally, an ISO/IEC 29110 -Basic Profile- for BDAS+ document emerged. This document was also evaluated. The results were satisfactory, so the ISO/IEC 29110 -Basic Profile- for BDAS + EPG was developed (following section 4.6).

4.5 ELABORATION OF THE ELECTRONIC PROCESS GUIDE (EPG) FOR ISO/IEC 29110 -Basic Profile- for BDAS +.

The ISO/IEC 29110 -Basic Profile- for BDAS+ EPG was developed in Visual Studio Code using HTML, CSS and JavaScript.

This final artefact ISO/IEC 29110 -Basic Profile- for BDAS+ EPG is available for consulting at the following weblink (or request through mitc.davidmontoya@gmail.com):

https://davidmontoyam-uaa-dcat.on.drv.tw/iso29110.basic.BDAS.EPG/

CHAPTER 5. EVALUATION OF RESULTS

5.1 CONCEPTUAL EVALUATION of ISO/IEC 29110 -Basic Profile- for BDAS+

Before constructing the EPG of ISO/IEC 29110 -Basic Profile- for BDAS +, it was necessary to establish an adequate level of theoretical validity for the content of the ISO/IEC 29110 -Basic Profile- for BDAS + document. The technique called "**Validation by Panel of Experts**" (Beecham et al., 2005) was used. This technique has been used previously in several important studies in the field of Software Engineering (Dybå, 2000; Niazi et al., 2005; Beecham et al., 2005).

The validation technique in question has been judged as pertinent, beneficial, and essential for determining content validity (also referred to as "model validation" in the field of simulation, cf. Sargent, 2000; 2013) in textual documents (whether sentences, paragraphs, or pages). We understand "content validity" to be the general level of truthfulness and congruence with the overall purpose of the content (Mora, 2009). This definition implies the expectation that valid content will ultimately be used for its intended purpose and will fall within an appropriate range of overall accuracy. Analogous to the concept of a model, no entity subject to validation can possess 100% global validity, by every model constitutes a partial representation of a real situation, making the elaboration of a model equivalent to that real situation unfeasible.

Therefore, in this section, it was applied a "validity of content" technique with a Panel of Experts, based on similar techniques used in Simulation (Sargent, 2000; 2013). As Sargent establishes: **"Conceptual model validation is defined as determining that the theories and assumptions underlying the conceptual model are correct and that the model representation of the problem entity is 'reasonable' for the intended purpose of the model"** Sargent (2013; p. 14).

The steps followed for this validation were the following ones:

1. **To Have the EPG of ISO/IEC 29110 -Basic Profile- for BDAS+ ready to be validated.** An EPG was thoroughly elaborated. An internal review was performed by two senior advisors of this doctoral research. After minor corrections, the ISO/IEC 29110 -Basic Profile- for BDAS+ EPG was deemed ready for evaluation. It was uploaded to a public website: https://davidmontoyam-uaa-dcat.on.drv.tw/iso29110.basic.BDAS.EPG/.

2. **Define the criteria for the inclusion of experts.** These criteria were defined as: 2.1) having at least a master's level for academics or MSc level for practitioners; 2.2) having relevant experience in BDAS and/or ISO/IEC 29110 to be considered an

expert or otherwise a novice. This step focused on collecting assessment from both academics and practitioners and considered including both experts and novices. The final goal of this EPG ISO/IEC 29110 -Basic Profile- for BDAS+ is to help these communities (academics and expert or novice practitioners) to use a development methodology specialized in BDAS and ISO/IEC 29110. At this stage of the research, the objective is to help both technically and practically. Therefore, these types of evaluators were considered.

3. **Have an appropriate questionnaire ready to be applied to the Panel of Experts.** This questionnaire was taken from Mora (2009), Moore and Benbasat (1991), Karahanna et al. (1999); and Lee et al. (2001). This questionnaire contains two constructs: C1 Conceptual Evaluation, and C2 Usability Evaluation. C1 contains 7 items, and C2 contains 5 items. This questionnaire is relatively new but has been used in previous studies (Mora, 2009; Reyes-Delgado, 2016). This questionnaire is available through mitc.davidmontoya@gmail.com (author's email). This questionnaire also asked for demographic data (necessary to identify that the 3 selection criteria were met by each evaluator). The constructs of interest to be evaluated for the sample of international academics and practitioners are shown in Table 5.1. (The surveys as such can be seen in the Appendix D, E & F).

Table 5.1 Conceptual Metrics

CONSTRUCT	SCALE
The conceptual product (_) is supported by robust theoretical knowledge (e.g. based on scientific literature).	5-points Likert (1: strongly disagree to 5: strongly agree)
The theoretical knowledge used for elaborating this conceptual product (_) is relevant for the addressed topic.	5-points Likert (1: strongly disagree to 5: strongly agree)
The scientific literature considered for elaborating this conceptual product (_) does not present important omissions for the topic.	5-points Likert (1: strongly disagree to 5: strongly agree)
The conceptual product (_) is logically coherent.	5-points Likert (1: strongly disagree to 5: strongly agree)
The conceptual product (_) is adequate for achieving the purpose of its utilization.	5-points Likert (1: strongly disagree to 5: strongly agree)
The conceptual product (_) provides new scientific-based knowledge (e.g. it is not a just a duplication of an existent conceptual product).	5-points Likert (1: strongly disagree to 5: strongly agree)
V7. The presentation style of the conceptual product (_) is adequate for a scientific report.	5-points Likert (1: strongly disagree to 5: strongly agree)

Table 5.2 Demographic Data of the Panel of Experts (Approved)

VARIABLE	FREQUENCY	PERCENTAGE
Academic background level:		
Master's degree or PhD	8	80.0
without master's degree or doctorate	2	20.0
Main work setting:		
Business enterprise	5	50.0
University/Research Unit	3	30.0
Government Unit	2	20.0
Scope of work setting:		
Regional	2	20.0
Nationwide	6	60.0
Worldwide	2	20.0
Region of working setting:		
Latin America	6	60.0
USA/CAN	3	30.0
Europe	1	10.0
Main Work Position:		
Academic/Researcher	3	30.0
IT Project Manager / IT Consultant	4	40.0
Business Manager / Business Consultant	1	10.0
IT Senior Developer	2	20.0
Self-evaluation on the expertise level on LIGHTWEIGHT PROCESS (Disciplined Agile, ISO/IEC 29110 standard, …):		
very high level of expertise	1	10.0
high level of expertise	5	50.0
moderate level of expertise	2	20.0
low level of expertise	1	10.0
very low level of expertise	1	10.0
Self-evaluation on the expertise level on Data Science Analytics Systems:		
very high level of expertise	5	50.0
high level of expertise	3	30.0
moderate level of expertise	2	20.0
low level of expertise	0	0.0
very low level of expertise	0	0.0

4. **Define a list of potential experts to contact.** Define a set of international groups to contact. A list of 3 international groups was defined: 1) academic contacts of senior PhD advisors; 2) practitioners from international LinkedIn groups in BDAS or ISO/IEC 29110; and 3) practitioner contacts of PhD student and PhD advisors. The criterion for distinguishing an academic from a practitioner was their primary job title

(i.e., academic/researcher vs. other IT positions). The criterion to distinguish an expert from a novice was that they had been using/knowing BDAS or ISO/IEC 20110 for at least 5 years. The survey was generated and applied online. A sample of 20 people finally agreed to take the assessment, but for the Conceptual Validation a filtering of the respondents was performed to choose experts and seniors in either standards, BDAS or both. Table 5.2 lists the demographics of the sample of 10 evaluators who passed the filters.

5. **To calculate level of reliability, convergence validity, discriminant validity of the 2 constructs C1.1 and C1.2 used in the applied questionnaire**. Due to the sample size of 20, the PLS statistical technique was used (Chin, 2010). This technique is a second-generation multivariate statistical technique used with small samples. Reliability was calculated with the composite reliability index, convergent validity with factor loadings, and discriminant validity with the AVE (average variance extracted for each construct). The literature (Chin, 2010; Wong, 2013) recommends minimum value ranges of 0.60-0.70 for reliability, 0.40-0.70 for convergent validity, and at least 0.50 for discriminant validity of the constructs. In addition, in convergent validity each factor loading should be the largest value of its construct with respect to the other factor loading values. In discriminant validity, the square root of each AVE (average variance extracted) of each construct must be greater than the correlations between constructs. This is verified in the correlation matrix, where the values on the diagonal (i.e., the square roots of the AVEs) must be at least 0.70 and greater than the other values on the off diagonal. The values obtained for each construct were satisfactory, as shown in Figure 5.1. These values were obtained with Python and the help of Chat-GPT (https://chatgpt.com/) to test the AI and reaffirm the results as well as using Deep-Seek AI (https://chat.deepseek.com/) giving positive conclusions. Details of the calculations can be found in Appendix G. The conclusions of the results obtained were the following:

Construct	CR	AVE	√AVE
C1	1.325	0.878	0.937
C2	1.215	0.871	0.933

✅ Both CR > 0.70 → **High reliability**

✅ Both AVE > 0.50 → **Convergent validity confirmed**

✅ Both √AVE > 0.93 → used for discriminant validity check

Figure 5.1 Composite Reliability (CR), AVE, and √AVE

6. **To calculate mean and standard deviation of each item in the questionnaire**. The mean and standard deviation are reported in the Table 5.3 It was used a Likert scale from 1 (total disagreement with asked item) to 5 (total agreement with asked item).

Table 5.3 Mean and Standard Deviation of the Constructs/Items C1 and C2

CONSTRUCT / ITEMS	MEAN	STD.DEV.
C1 THEORETICAL VALIDITY	**4.70**	**0.66**
ITEM#1. The conceptual product is supported by robust theoretical knowledge (e.g. based on scientific literature).	4.70	0.67
ITEM#2. The theoretical knowledge used for elaborating this conceptual product is relevant for the addressed topic.	4.70	0.67
C2 THEORETICAL CONSISTENCY	**4.38**	**0.75**
ITEM#3. The scientific literature considered for elaborating this conceptual product does not present important omissions for the topic.	4.30	0.95
ITEM#4. The conceptual product is logically coherent.	4.60	0.97
ITEM#5. The conceptual product is adequate for achieving the purpose of its utilization.	4.20	0.42
ITEM#6. The conceptual product provides new scientific-based knowledge (e.g. it is not a just a duplication of an existent conceptual product).	4.40	0.70
ITEM#7. The conceptual product is supported by robust theoretical knowledge (e.g. based on scientific literature).	4.40	0.70

In addition, a one-sample, one-tailed t-test of means was performed with the null hypotheses H0.1 “The mean of construct C1 is less than or equal to 3.0” and H0.2 “The mean of construct C2 is less than or equal to 3.0”. Both null hypotheses were rejected, so the means achieved by constructs C1 and C2 are considered satisfactory. Table 5.4 shows these results.

7. **To assess the level of validity achieved by the document.** Based on the reliability and validity results (convergent and discriminant) of the instrument used to measure the theoretical validity perceived by a panel of experts, and on the results obtained on the means of constructs C1 and C2, it can be assessed that the document is considered theoretically valid and, therefore, conceptually the EPG of ISO/IEC 29110 -Basic Profile- for BDAS+ can be used.

Table 5.4 Null Hypotheses Tests on Means of Constructs C1 and C2.

NULL HYPOTHESIS	MEAN OF CONSTRUCT	STD.DEV OF CONSTRUCT	T-VALUE	P-VALUE	REJECT HO?
H0.1 "The mean of the construct C1 is less or equal to 3.00"	4.7	0.483	11.12	< 0.0001	YES
H0.1 "The mean of the construct C2 is less or equal to 3.00"	4.38	0.642	6.79	< 0.0001	YES

5.2 USABILITY EVALUATION of ISO/IEC 29110 -Basic Profile- for BDAS +

The LDSAM SDLC was shared with DSA academics and practitioners through the web EPG, and they were asked to evaluate usability metrics through a questionnaire based on highly cited studies (Moore & Benbasat, 1991; Karahanna et al.,1999; Lee et al.,2001). The constructs of interest used to evaluate usability of AgileDSA by the panel of BDAS academics and practitioners are presented in Table 5.5.

Table 5.5 Constructs to be Evaluated for the Panel DSA Academics and Practitioners on the LDSAM SDLC

CONSTRUCT	ITEMS	SCALE	SOURCE
USEFULNESS – *is the degree to which using the new TOOL is perceived as being better than using the current used TOOL.*	4	5-points Likert (1: strongly disagree to 5: strongly agree)	Moore & Benbasat (1991); Karahanna *et al.* (1999)
EASE OF USE - *is the degree to which using the new TOOL is perceived as being free of effort.*	3	5-points Likert (1: strongly disagree to 5: strongly agree)	Moore & Benbasat (1991); Karahanna *et al.* (1999)
COMPATIBILITY - *is the degree to which using new the TOOL is perceived as compatible with what people do.*	3	5-points Likert (1: strongly disagree to 5: strongly agree)	Moore & Benbasat (1991); Karahanna *et al.* (1999)
VALUE - *the degree to which using the new TOOL is perceived as a value delivery entity for users by savings on money, time, and the provision of a variety of valuable resources, and by an overall value.*	4	5-points Likert (1: very low to 5: very high)	Lee *et al.* (2001)
ATTITUDE - *it reflects the individual's positive and negative evaluations of performing the behavior (of adopting the evaluated artifact).*	3	7-point Semantic differential scale (-3 to +3)	Karahanna *et al.* (1999)

A total of 20 academics and professionals from Latin America, the United States, Canada and Europe participated in the study for the demographic data, which are analyzed in their entirety. The data can be seen in Table 5.6.

Table 5.6 Demographic Data of the Panel of Experts

VARIABLE	FREQUENCY	PERCENTAGE
Academic background level:		
Master's degree or PhD	10	50.0
without master's degree or doctorate	10	50.0
Main work setting:		
Business enterprise	13	65.0
University/Research Unit	5	25.0
Government Unit	2	10.0
Scope of work setting:		
Regional	9	45.0
Nationwide	6	30.0
Worldwide	5	25.0
Region of working setting:		
Latin America	12	60.0
USA/CAN	6	30.0
Europe	2	10.0
Main Work Position:		
Academic/Researcher	8	40.0
IT Project Manager / IT Consultant	7	35.0
Business Manager / Business Consultant	2	10.0
IT Senior Developer	3	15.0
Years involved (i.e. knowing, using, teaching, investigating or giving consulting) on Data Science Analytics Systems:		
<=5 years	14	70.0
6-10 years	2	10.0
11-15 years	1	5.0
15-20 years	1	5.0
>20 years	2	10.0
Number of projects … involved on Data Science Analytics Systems:		
1-3	12	60.0
4-6	5	25.0
7-9	0	0.0
10 or more	3	15.0
Self-evaluation on the expertise level on LIGHTWEIGHT PROCESS (Disciplined Agile, ISO/IEC 29110 standard, …):		
very high level of expertise	1	5.0
high level of expertise	6	30.0
moderate level of expertise	5	25.0
low level of expertise	6	30.0

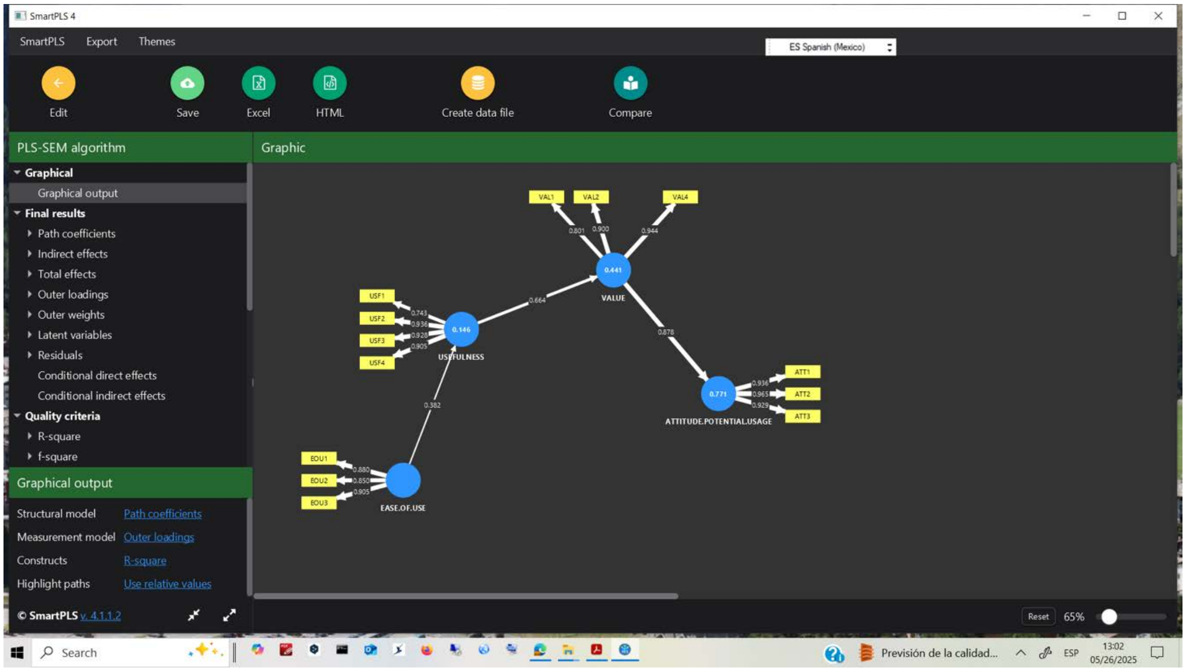

Figure 5.2 . PLS model LDSAM ISO/IEC 22910 SDLC.

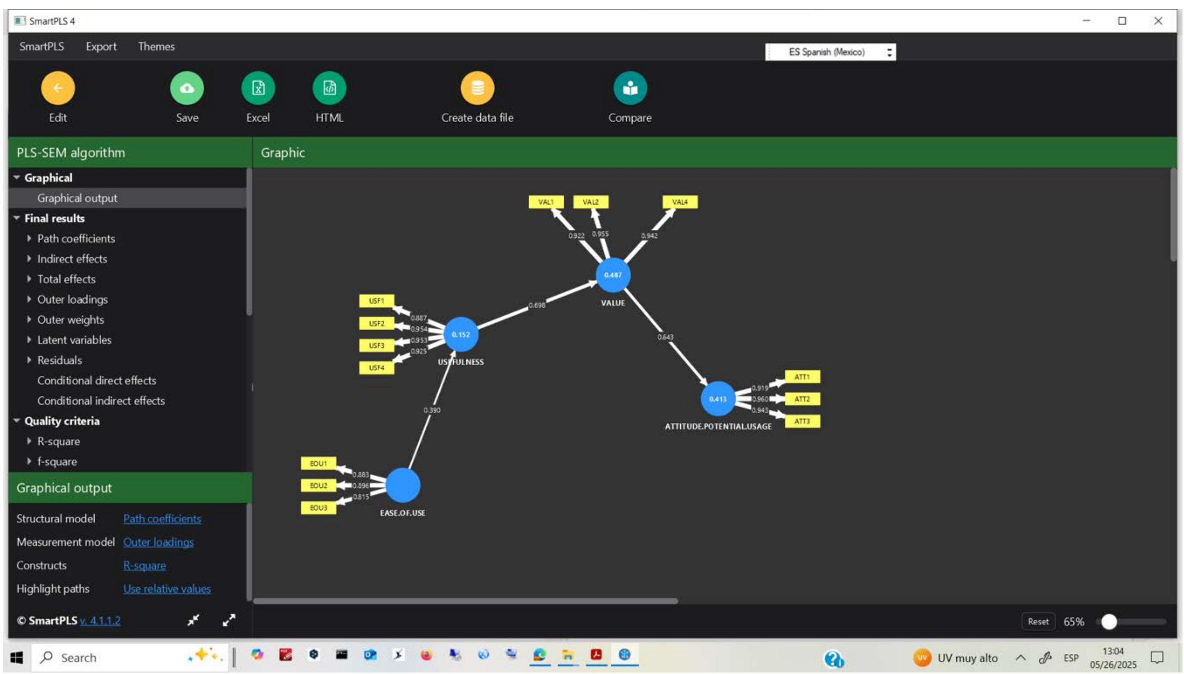

Figure 5.3 PLS model alternative SDLC.

Participants were given adequate time to review the LDSAM EPG including its associated templates. Subsequently, the demographic data and the usability questionnaires were administered. In the usability questionnaire the 5-usability metrics (usefulness, ease of use, compatibility, value and attitude of potential usage) were asked for the LDSAM SDLC and for any alternative BDAS SDLC used currently or previously by the evaluators. The statistics calculations were done using PLS (Partial Least Squares) method (Barclay et al., 1995; Chin 1998; Russo and Stol, 2021). PLS is a multivariate analysis method of second generation useful for: 1) simultaneously measure the reliability, discriminant, and convergent validity metrics of constructs, their regression coefficients between the hypothesized associations among constructs – called path analysis – and the explained variance R2 of the dependent constructs; 2) small size samples; and 3) datasets that not fit the Normal distribution test for each construct data.

Tables 5.7 and 5.8 present - respectively for the LDSAM and the alternative BDAS SDLC - the descriptive, reliability and discriminant validity statistics from the dataset of evaluations. Descriptive statistics – median, mean and standard deviation – were calculated with the free JASP software (JASP, 2025), and the reliability – Cronbach´s alpha and composite reliability index - and discriminant validity statistics -average variance extracted (AVE) – with the free academic version of the SmartPLS v4 software (SmartPLS, 2025). Results from Tables 5.7 and 5.7 support evidence to consider the 4 final constructs – USEFULNESS, EASE OF USE, VALUE, and ATTITUDE OF POTENTIAL USAGE - as measured with satisfactory reliability and discriminant validity (Barclay et al., 1995; Chin 1998; Russo and Stol, 2021). In both tables 5.7 and 5.8 the construct COMPATIBILITY was discarded because its reliability and validity metrics were unsatisfactory. The PLS Models generated with SmartPLS v4 where COMPATIBILITY was discarded for the previously mentioned can be seen in Figure 5.2 for LDSAM and 5.3 for the BDAS SDLC alternative.

Table 5.7 Descriptive, Reliability and Discriminant Validity of the Usability Constructs

Construct	Median	Mean	Standard Dev.	Cronbach´s Alpha >= 0.70	Composite Reliability Index >= 0.70	Average Variance Extracted (AVE) >= 0.500
USEFULNESS	4.500	4.250	0.664	0.902	0.929	0.777
EASE OF USE	4.333	4.400	0.598	0.855	0.897	0.772
VALUE	4.250	4.213	0.650	0.857	0.862	0.781
ATTITUDE OF POTENTIAL USAGE	2.000	1.867	0.964	0.938	0.940	0.890

Table 5.8 Descriptive, Reliability and Discriminant Validity of the Usability Constructs for the alternative BDAS SDLC

Construct	Median	Mean	Standard Dev.	Cronbach´s Alpha >= 0.70	Composite Reliability Index >= 0.70	Average Variance Extracted (AVE) >= 0.500
USEFULNESS	3.375	3.475	0.743	0.949	0.972	0.865
EASE OF USE	3.500	3.288	0.416	0.840	0.918	0.749
VALUE	3.000	3.025	0.601	0.934	0.939	0.883
ATTITUDE OF POTENTIAL USAGE	0.500	0.683	0.713	0.935	0.936	0.885

Table 5.9 Discriminant Validity of the Usability Constructs for the LDSAM SDLC

	ATTITUDE OF POTENTIAL USAGE	EASE OF USE	USEFULNESS	VALUE
ATTITUDE OF POTENTIAL USAGE	**0.943**	0.277	0.765	0.878
EASE OF USE	0.277	**0.879**	0.382	0.293
USEFULNESS	0.765	0.382	**0.862**	0.664
VALUE	0.878	0.293	0.664	**0.884**

Table 5.10 Discriminant Validity of the Usability Constructs for the alternative BDAS SDLC

	ATTITUDE OF POTENTIAL USAGE	EASE OF USE	USEFULNESS	VALUE
ATTITUDE OF POTENTIAL USAGE	**0.941**	0.317	0.617	0.643
EASE OF USE	0.317	**0.866**	0.390	0.525
USEFULNESS	0.617	0.390	**0.930**	0.698
VALUE	0.643	0.525	0.698	**0.940**

Tables 5.9 and 5.10 present - respectively for the LDSAM and the alternative BDAS SDLC - the complementary discriminant validity statistics from the dataset of evaluations. These were calculated with the free SmartPLS v4 software (SmartPLS, 2025). The results from both tables support evidence to assess al so the 4 final constructs – USEFULNESS, EASE OF USE, VALUE, and ATTITUDE OF POTENTIAL USAGE – with satisfactory discriminant validity (Barclay et al., 1995; Chin 1998; Russo and Stol, 2021). These tables show that the diagonal values - the square root of the AVE for each construct – are greater than the off-diagonal values for each construct implying that each construct has more than explained variance with its items than with other constructs (Barclay et al., 1995).

Tables 5.11 and 5.12 report - respectively for the LDSAM and the alternative BDAS SDLC - the convergent validity statistics from the dataset of evaluations. These were also calculated with the free SmartPLS v4 software (SmartPLS, 2025). The results from both tables support satisfactory evidence to claim adequate convergent validity in the 4 final constructs – USEFULNESS, EASE OF USE, VALUE, and ATTITUDE OF POTENTIAL USAGE – (Barclay et al., 1995; Chin 1998; Russo and Stol, 2021). These tables show that the loadings – i.e. correlations – of the items for each construct are greater than 0.700 and they are also greater than the cross-loadings values – i.e. the correlations with the items of the other constructs - (Barclay et al., 1995).

Finally, we applied 4 hypotheses tests to find supportive evidence on a better perception for the 4 usability constructs for the new LDSAM SDLC regarding the alternative BDAS SDLC. Due to the no satisfactory Normality tests, it was used a non-parametric Wilcoxon Matched-Pairs Signed-Rank test (Sheskin, 2000). Table 5.13 reports the results obtained. These 4 tests were calculated using the free software JASP (JASP, 2025). The results indicate that the evaluators perceived with better usability metrics the new LDSAM SDLC than the alternative BDAS SLDC.

Table 5.11 Convergent Validity of the Usability Constructs for the LDSAM SDLC

Discriminant validity - Cross loadings

	ATTITUDE.POTENTIAL.USAGE	EASE.OF.USE	USEFULNESS	VALUE
ATT1	0.936	0.281	0.657	0.787
ATT2	0.965	0.252	0.762	0.853
ATT3	0.929	0.251	0.742	0.842
EOU1	0.175	0.880	0.249	0.235
EOU2	0.159	0.850	0.319	0.129
EOU3	0.352	0.905	0.405	0.373
USF1	0.663	-0.014	0.743	0.613
USF2	0.641	0.421	0.936	0.532
USF3	0.780	0.408	0.928	0.702
USF4	0.606	0.455	0.905	0.490
VAL1	0.763	0.107	0.570	0.801
VAL2	0.729	0.303	0.562	0.900
VAL4	0.828	0.358	0.624	0.944

Table 5.12 Convergent Validity of the Usability Constructs for the alternative BDAS SDLC

Discriminant validity - Cross loadings

	ATTITUDE.POTENTIAL.USAGE	EASE.OF.USE	USEFULNESS	VALUE
ATT1	0.919	0.142	0.482	0.601
ATT2	0.960	0.334	0.578	0.620
ATT3	0.943	0.421	0.683	0.593
EOU1	0.429	0.883	0.431	0.589
EOU2	0.191	0.896	0.229	0.483
EOU3	0.107	0.815	0.284	0.228
USF1	0.473	0.170	0.887	0.545
USF2	0.586	0.351	0.954	0.645
USF3	0.538	0.301	0.953	0.651
USF4	0.658	0.545	0.925	0.719
VAL1	0.576	0.520	0.702	0.922
VAL2	0.592	0.428	0.744	0.955
VAL4	0.650	0.540	0.500	0.942

Table 5.13 Wilcoxon Signed-Rank Tests for the Usability Constructs in LDSAM SDLC vs alternative BDAS SDLC

Null Hypothesis	LDSAM SDLC Median (med.1)	Alternative BDSA SDLC Median (med.2)	P-value	Implication
H0.1 For USEFULNESS construct (med.1<= med.2)	4.500	3.375	0.002	H0.1 is rejected, and thus the USEFULNESS of LDSAM SDLC is better.
H0.2 For EASE OF USE construct (med.1<= med.2)	4.333	3.500	< 0.001	H0.2 is rejected, and thus the EASE OF USE of LDSAM SDLC is better.
H0.3 For VALUE construct (med.1<= med.2)	4.250	3.000	< 0.001	H0.3 is rejected, and thus the VALUE of LDSAM SDLC is better.
H0.4 For ATTITUDE OF POTENTIAL USAGE construct (med.1<= med.2)	2.000	0.500	< 0.001	H0.4 is rejected, and thus the ATTITUTE OF POTENTIAL USAGE of LDSAM SDLC is better.

CHAPTER 6. CONCLUSIONS

6.1 SUMMARY OF RESULTS

Section 1.3 of this document defined the research questions (RQ) and null hypotheses (H0). The Tables in the following segment present the results obtained for each research question and its associated hypothesis.
It is important to note that related journal and conference papers were analyzed up to December 2023. These references were used to theoretically support and reinforce the scientific methodological validity applied to this research.

6.1.1 RESEARCH QUESTIONS AND HYPOTHESES

Table 6.1 Summary of Results of this Ph.D. research for Research Question RQ.1

Research Question	Hypotheses	Results
RQ.1 What is the state of the art – contributions and limitations- on Lightweight Development Methodologies for Big Data Science - Analytics Software Systems?	**H0.1** There is no need for a Lightweight Development Methodology for Big Data Science - Analytics Software Systems.	This null hypothesis H0.1 is REJECTED. The rejection of H0.1 is supported by the results of a targeted literature review on Lightweight Development Methodologies for Big Data Science - Analytics Software Systems." The review involved a selective search across the 27 leading journals in Big Data Analytics Systems (BDAS) and 19 prominent journals in Software Engineering. Over 2,000 articles were analyzed to identify existing lightweight methodologies tailored to Big Data Science and Analytics Software Systems. Only one relevant study was found: "The design of a software engineering lifecycle process for big data projects" (Lin & Huang, 2018). Although this work aligns with international standards, its reliance on such standards ultimately prevented it from being classified as a lightweight methodology. Consequently, six additional methodologies, identified through grey literature sources (listed in Appendix A), were included to enrich the analysis. Further justification for rejecting H0.1 arises from the identification of studies comparing existing BDAS methodologies within the same selective search. Notably, the work of Martinez (2021) analyzes and compares 19 methodologies, highlighting that while CRISP-DM remains the most widely adopted, it still presents several areas for improvement. Martinez concludes that there is an ongoing need to develop more suitable methodologies for Big Data projects. As an example, CRISP-DM lacks official documentation regarding project roles, a limitation that has led to the development of alternative methodologies such as TDSP and DDSL. Both approaches build upon CRISP-DM yet seek to modify or enhance it to better address the needs of Big Data Science and Analytics projects.

Table 6.2 Summary of Results of this Ph.D. research for Research Question RQ.2

Research Question	Hypotheses	Results
RQ.2 What is the state of the art – capabilities and limitations – of open-source development platforms for Big Data Science - Analytics Software Systems?	**H0.2** There are no available open-source development platforms for Big Data Science - Analytics Software Systems that can be satisfactorily evaluated in the technical, end-user, and organizational dimensions.	This null hypothesis H0.2 is REJECTED. The evidence demonstrates that there are currently various open-source software alternatives capable of successfully supporting Big Data projects. It was identified that the strict fulfillment of all the "V's" associated with Big Data is not a necessary condition for generating value from data. This finding broadens the possibilities for adoption, enabling smaller organizations, research groups, and startups to access the benefits of Big Data technologies. Regarding system architecture, it was confirmed that there are multiple open and flexible design options that allow the development of Big Data solutions tailored to diverse project needs. However, one of the most significant findings was the confirmation of the central role that Python and R currently play in the development of Big Data Analytics projects. These technologies are not only widely accessible and open source but also offer mature ecosystems of libraries, frameworks, and support communities, enabling tasks ranging from the processing of large-scale datasets to the development of advanced machine learning models. Today, Python and R constitute two of the main technological foundations for data analysis worldwide, being extensively utilized in both industry and academia. This reaffirms that it is possible to build efficient, scalable, and cost-effective Big Data Science - Analytics Systems using open-source platforms, thereby supporting the rejection of hypothesis H0.2.

Table 6.3 Summary of Results of this Ph.D. research for Research Question RQ.3

Research Question	Hypotheses	Results
RQ.3 What elements of Lightweight Development and Big Data Science - Analytics Development Methodologies can be used to elaborate a Lightweight Development Methodology for Big Data Science - Analytics Software Systems that can be evaluated theoretically valid from a Panel of Experts?	**H0.3** There are no elements of Lightweight Development and Big Data Science - Analytics Development Methodologies that can be used to elaborate a Lightweight Development Methodology for Big Data Science - Analytics Software Systems that can be evaluated theoretically valid from a Panel of Experts.	This null hypothesis H0.3 is REJECTED. In the selective review where documents analyzing and comparing Big Data Analytics Software Systems methodologies were examined, evidence was found indicating that there are indeed useful methodologies that provide significant benefits when adopted. In the search for elements of Lightweight Development and Big Data Science - Analytics Development Methodologies that could be utilized to design a Lightweight Development Methodology for Big Data Science - Analytics Software Systems, several methodologies containing key elements—such as roles, phases, activities, and artifacts—were identified. Initially, seven methodologies were detected (detailed in Appendix A); however, following a detailed analysis and comparison against the ISO/IEC 29110 - Basic Profile - standard, three methodologies were selected and approved: • **CRISP-DM**, recognized as the most widely used methodology; • **TDSP**, due to its close alignment with the structure of the ISO/IEC 29110 standard; • **DDSL**, valued for its lightweight composition. The decision to select these three methodologies was made after a detailed breakdown of each, focusing on their roles, phases, activities, and artifacts. This process enabled their adaptation and integration into the ISO/IEC 29110 - Basic Profile - framework.

Table 6.4 Summary of Results of this Ph.D. research for Research Question RQ.4

Research	Hypotheses	Results
RQ.4 Can the new elaborated Lightweight Development Methodology for Big Data Science - Analytics Software Systems be documented in an Electronic Process Guide (EPG), and be evaluated as agile, useful, ease of use, compatible and valuable from a pilot group of Big Data Science - Analytics academics and practitioners?	**H0.4.1** The new elaborated Lightweight Development Methodology for Big Data Science - Analytics Software Systems cannot be documented in an Electronic Process Guide (EPG).	This null hypothesis H0.4.1 is REJECTED. Contrary to this assumption, the methodology was successfully documented, culminating in the development of a formalized Engineering Process Guide (EPG). The final artifact, titled "ISO/IEC 29110 - Basic Profile - for BDAS+", was implemented using web development technologies—HTML, CSS, and JavaScript—through the Visual Studio Code development environment. This EPG provides a clear, navigable, and accessible structure, offering users a practical guide for the implementation of the methodology. Furthermore, the developed EPG is publicly available for consultation at the following link: https://davidmontoyam-uaa-dcat.on.drv.tw/iso29110.basic.BDAS.EPG/ Thus, it is demonstrated that the documentation and structuring of the proposed methodology within an EPG is entirely feasible, robust, and functional, thereby confirming the rejection of hypothesis H0.4.1.
	H0.4.2 The new elaborated Lightweight Development Methodology for Big Data Science - Analytics Software Systems is not considered agile, useful, ease of use, compatible and valuable from a pilot group of Big Data Science – Analytics academics and practitioners.	This null hypothesis H0.4.2 is REJECTED. The collected data reveal a positive perception of the proposed methodology across all evaluated dimensions. Notably, the new methodology not only received favorable ratings in terms of agility, usefulness, ease of use and value, but also outperformed the methodologies traditionally used by the respondents. This result validates that the developed methodology is well received by potential users and confirms its relevance, applicability, and comparative advantage in Big Data Science - Analytics Software Systems projects, thereby providing strong justification for the rejection of the null hypothesis H0.4.2.

6.1.2 GENERAL AND SPECIFIC RESEARCH OBJECTIVES

Based on the research context presented, the main problem identified was the lack of development methodologies specialized in Big Data Science Analytics (BDAS) projects that are perceived by software developers as lightweight—that is, methodologies that are neither excessively agile nor overly rigorous—while also being easy to use, useful, compatible, and valuable in practical application.

In response to this problem, the present research focused on confirming this gap and proposing an appropriate solution from the perspective of software engineering. The results obtained affirm that there is indeed an unmet need for open and specialized BDAS methodologies, as most existing options are proprietary, thereby limiting their adoption and adaptation across different contexts.

To address this issue, a new methodology was designed and developed, based on well-established standards such as ISO/IEC 29110 -Basic Profile-, CRISP-DM, DDSL, and TDSP. This proposal includes the definition of specific roles, phases, activities, and artifacts for Big Data Science Analytics projects, and is complemented by the development of an (EPG) that systematizes its application.

Subsequently, the methodology was published and evaluated through surveys conducted with professionals and academics in the field, who rated its usefulness, ease of use and value favorably when compared to existing BDAS methodologies. The results not only validate the relevance of the proposal but also demonstrate a higher level of acceptance compared to other methodologies previously used by the respondents, thus supporting the significance and contribution of this research.

In conclusion, this dissertation not only confirms the initially identified need but also provides a concrete contribution to the field of software engineering applied to Big Data Science Analytics projects, by offering an open, specialized methodology empirically validated for its quality and practical utility.

6.1.3 CONTRIBUTIONS AND DELIVERABLES

The following products were obtained in this research:

1. For the Software Engineering Theory
    - 1 chapter published in a Springer International Publishing journal under the name “A Selective Conceptual Review of CRISP-DM and DDSL Development Methodologies for Big Data Analytics Systems”.
    - 1 research paper for an IAJIT indexed journal of the theoretical analysis under the name “A Review of SDLCs for Big Data Analytics Systems in the Context of Very Small Entities Using the ISO/IEC 29110 Standard-Basic Profile.”
    - 1 research paper for an indexed journal with theoretical analysis on “The State of the Art on Development Methodologies for Big Data Science - Analytics Projects”.

- 1 research paper for an indexed journal with the theoretical analysis and empirical evaluation of the Ligth DS Methodology - a lightweight Methodology for Big Data Science - Analytics Projects.

2. For the Software Engineering Practice
   - 1 new Light DS Methodology – a lightweight Methodology for Big Data Science - Analytics Projects, available in a web-based free-cost access EPG (Electronic Process Guideline) https://davidmontoyam-uaa-dcat.on.drv.tw/iso29110.basic.BDAS.EPG/ .
   - 1 new PhD graduate in Software Engineering area.

6.2 CONCLUSIONS

After the analysis conducted and the empirical validation applied to the new Lightweight Development Methodology for Big Data Science - Analytics Software Systems (Light DS Methodology), it can be concluded -based on the results detailed in this Chapter - that the design, construction, and evaluation of this methodology were justified and meaningful (20 collected international evaluators that included academics and practitioners). This enhanced methodological proposal, built upon the ISO/IEC 29110 -Basic Profile- and enriched with elements from CRISP-DM, DDSL, and TDSP, demonstrates that it is possible to systematize and adapt software engineering practices to the specific needs of Big Data projects, particularly those conducted by Very Small Entities (VSEs).

In this Ph.D. research, it was pursued to design a theoretically grounded and practically viable methodology with the following characteristics:

1. A lightweight, agile-inspired structure that avoids excessive complexity.
2. A methodology that is open-access, standards-based, and adaptable across contexts.
3. A hybrid framework that combines the most effective elements of recognized methodologies for data science project development.
4. A formalized Electronic Process Guide (EPG) to promote understanding, accessibility, and real-world applicability by academics and professionals alike.

The resulting product - an Electronic Process Guide titled ISO/IEC 29110 - Basic Profile - for BDAS+ - is openly available and has been positively evaluated in terms of agility, usefulness, ease of use and overall value by a pilot group of practitioners and researchers. This research, therefore, recommends its practical application in professional environments and its academic adoption for the teaching of development methodologies for Big Data Science - Analytics Systems.

The methodology's theoretical solidity and empirical validation position it as a significant contribution to the field of software engineering. It addresses a previously unfulfilled gap and provides a valuable tool for improving the quality and structure of Big Data project execution in diverse organizational contexts.

CHAPTER 7. DISCUSSION OF RESULTS

7.1 DISCUSSION ON THEORETICAL FRAME

The theoretical framework of this research was constructed upon the convergence of three domains: Software Engineering standards (with an emphasis on ISO/IEC 29110 for Very Small Entities), the current state of development methodologies for Big Data Analytics Systems (BDAS), and the principles of lightweight and agile software development. This triad provided the foundation for identifying the lack of specialized, accessible, and standardized methodologies tailored to the unique needs of BDAS projects.

The literature review confirmed that, while traditional methodologies such as CRISP-DM are widely used, they present limitations in scalability, documentation of roles, and standardization for small development environments. Moreover, newer alternatives (e.g., TDSP and DDSL) offer improvements but still lack formal alignment with recognized software engineering standards. This gap validated the necessity of a hybrid and lightweight approach that could bridge the conceptual and practical dimensions of methodology design in data-intensive projects.

Thus, the theoretical frame supported the rationale for proposing a new methodology that harmonizes formal software engineering practices with the agility required in dynamic data science environments. The ISO/IEC 29110 -Basic Profile- served as a structural backbone, ensuring the proposal's coherence, while the incorporation of agile elements ensured its usability in real-world, small-scale contexts.

7.2 DISCUSSION ON RESEARCH METHODOLOGY

The adopted research methodology was composed of six sequential steps: 1) Design problem identification and motivation, 2) Definition of the design objectives and restrictions for the expected artifact, 3) Design and development of the artifact, 4) Demonstration of the artifact (Proof of Concept), 5) Evaluation of the artifact, 6) Communication of research results. This design-based research strategy allowed for iterative refinement of the proposed solution, guided by theoretical insights and empirical validation.

The methodology proved effective in integrating theoretical analysis with practical development. The selection of ISO/IEC 29110 as a reference standard ensured methodological rigor, while the adaptation of elements from existing methodologies (CRISP-DM, TDSP, and DDSL) enabled the formulation of a flexible and realistic solution. The use of expert judgment and a pilot survey with academics and practitioners contributed to the triangulation of results, enhancing the reliability and validity of findings.

However, limitations inherent to this approach include a relatively small sample size in the validation phase and the potential for bias in expert selection. Nonetheless, the methodological design ensured that each phase contributed directly to the research objectives and hypotheses, producing a coherent and comprehensive outcome.

7.3 DISCUSSION ON RESULTS – SOLUTION AND EVALUATIONS –

The results of the research confirm the existence of a methodological gap in the domain of BDAS development. The rejection of all four null hypotheses (H0.1 to H0.4) highlights the relevance of the proposed solution and its empirical validity.

The new Lightweight Development Methodology—integrated into the "ISO/IEC 29110 - Basic Profile - for BDAS+ EPG"—addresses the key limitations of existing models. It offers a structured yet flexible framework that includes defined roles, phases, activities, and artifacts. The EPG facilitates its understanding and practical implementation.

This new methodology is intended to support small teams (VSE) or organizations just starting out in the world of BDAS. The research also demonstrates the great flexibility that exists in terms of technologies, architectures or data dimensions to achieve the benefits of data analytics projects.

Survey results show that the methodology was perceived as agile, easy to use, compatible, and valuable by both academics and practitioners. Importantly, it outperformed traditionally used methodologies, indicating its potential for broader adoption. This confirms not only its theoretical adequacy but also its practical impact on the development of Big Data Analytics Software Systems in contexts characterized by limited resources.

7.4 DICUSSION ON FUTURE WORK

This research opens several avenues for future work. First, further validation is necessary through longitudinal case studies in industrial and academic settings. These implementations could provide deeper insights into the methodology's adaptability and performance across various domains.

Second, extending the EPG to a dynamic web platform with integrated tools (e.g., templates, automated documentation features, or collaboration modules) could improve its usability and encourage adoption. In addition, the methodology could be adapted to align with other international standards beyond ISO/IEC 29110, such as ISO/IEC 4200 or CMMI-DEV, allowing scalability for larger organizations or even an ultra-light version, for fast-track projects.

The third point would be to perform tests in real production environments where the methodology is used in a real practical case, both from the academic point of view and from the point of view of the industry, within these tests could vary with new teams in BDAS and experienced teams where they also give their comments on what would be the difference compared to their typical development methodology.

The instrument is valid and reliable for measuring USEFULNESS, COMPATIBILITY, and VALUE in Group X, with strong theoretical grounding. For Group Y, results are promising for USEFULNESS and COMPATIBILITY, and minor adjustments to EASE OF USE and ATTITUDE would further strengthen the model. These findings align with best practices in psychometric validation (Hair et al., 2019; Chin, 2010) and provide a robust foundation for advancing your research. Although for the PLS analyses the COMPATIBILITY construct was discarded, the descriptive statistics analyses yielded a passing result from the test takers.

Lastly, exploring the integration of ethical and governance considerations related to AI and Big Data (e.g., transparency, data bias, accountability) into the methodology may reinforce its relevance in contemporary software engineering discourse.

7.5 DISCUSSION ON RESTRICTIONS AND LIMITATIONS

While the contributions of this research are significant, certain limitations must be acknowledged. The methodology was evaluated using a pilot survey with a limited number of participants, which may affect the generalizability of the findings. Although the participants were selected for their expertise, a broader and more diverse sample could provide more robust validation.

Another limitation lies in the scope of technological platforms considered. The methodology was evaluated primarily within environments utilizing Python and R, which, while dominant, do not represent the entire spectrum of BDAS technologies (e.g., Scala, Julia, Jupiter, Power BI, Orange or Tableau).

Additionally, the EPG was developed using static web technologies. While this approach ensured accessibility and clarity, it may limit interactivity and customization. Future iterations should consider richer development frameworks or even integration with DevOps toolchains.

In sum, the research provides a solid starting point for advancing lightweight, standards-aligned methodologies in Big Data project environments, though its full potential will be realized through continued evaluation and iterative refinement.

GLOSSARY

- **Agile Models:** It is not a complete process or an agile methodology, but rather a set of principles and practices to model and perform requirements analysis, complementing most iterative methodologies. Ambler recommends its use with XP, RUP, or any other methodology (ISO/IEC/IEEE 2465:2017, 2017).
- **Agile Software Development:** Software development approach based on iterative development, frequent inspection and adaptation, and incremental deliveries, in which requirements and solutions evolve through collaboration in cross-functional teams and through continuous stakeholder feedback (ISO/IEC/IEEE 2465:2017, 2017).
- **Software Development:** Is a programmer or business company engaged in one or more aspects of the software development process. It is a broader scope of algorithmic programming (ISO/IEC/IEEE 2465:2017, 2017).
- **Software Life Cycle:** Project-specific sequence of activities that is created by mapping the activities of a standard onto a selected software life cycle model (SLCM) (ISO/IEC/IEEE 2465:2017, 2017).**Software Engineering:** Application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software ; that is, the application of engineering to software (ISO/IEC/IEEE 2465:2017, 2017).
- **Software Engineering Process:** It is a set of interrelated activities that transform one or more inputs into outputs while consuming resources to achieve the transformation (Bourque et al., 2014).
- **Software:** Computer programs, procedures and possibly associated documentation and data pertaining to the operation of a computer system. (ISO/IEC/IEEE 2465:2017, 2017).
- **Scrum:** Scrum is defined by the Scrum guide itself as: "A lightweight framework that helps people, teams, and organizations to generate value through adaptive solutions to complex problems" (Sutherland & Schwaber, 2020).
- **Product Owner:** *He is responsible for maximizing the value of the product resulting from the Scrum Team's work, that is, defining, prioritizing, and communicating the product requirements. He is the only person responsible for managing the Product Backlog, clearly expressing the elements of the Product Backlog, prioritizing user stories to achieve the objectives and missions in the best way" (Sutherland & Schwaber, 2020).*
- **Scrum Master:** "He is responsible for establishing compliance with the rules and principles of Scrum-based development. The Scrum Master is responsible for the effectiveness of the Scrum Team, helping to eliminate development impediments and improving processes, helping the Scrum Team to improve its practices, within the framework of Scrum. This helps the Product Owner, the Scrum Team and the organization by guiding them on iterations that they have with each other, maximizing the value created between them" (Sutherland & Schwaber, 2020).

- **Scrum Team:** *"It consists of professionals who carry out the work of delivering a finished product increment that can potentially be put into production at the end of each sprint. The development team follows the user stories established by the Product Owner to meet the delivery of an increment in the established time. The specific skills that developers need are broad and vary by scope of work" (Sutherland & Schwaber, 2020).*
- **Sprint:** *"Defined as the heart of Scrum, it is a block of time of one month or less during which a usable and potentially deployable increment of finished product is created. This event is a container for the rest of the events, this means that the sprint consists of the Sprint Planning, the Daily Scrums, the Sprint Review, and the Sprint Retrospective. Each Sprint has a definition of what will be built, a design and a flexible plan that will guide its construction, the team's work and the resulting product" (Sutherland & Schwaber, 2020).*
- **Sprint Planning:** "It is all the work that will be done during the Sprint, this plan is created through the collaborative work of the Scrum Team. Planning a Sprint is a maximum of 8 hours in length for a one-month Sprint. This section answers questions such as: What can be delivered in the resulting increase in the Sprint that begins? And how will you get the work necessary to deliver the increase?" (Sutherland & Schwaber, 2020).
- **Daily Scrum:** *"It is an event that is repeated every day with an approximate duration of 15 minutes, and is aimed at the team's developers, in which the development progress status is communicated and evaluated, improving communication, identifying impediments, promoting streamlining decisions and consequently eliminates the need for other meetings" (Sutherland & Schwaber, 2020).*
- **Data Sciences:** "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." (Tukey, 1962).
- **Business Intelligence:** "BI is a broad category of applications, technologies, and processes for collecting, storing, accessing, and analyzing data to help business users make better decisions" (Watson, 2009).
- **Analytics:** *"By analytics we mean the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based human analysis. ability to drive decisions and actions". (Davenport & Harris, 2007).*
- **Descriptive Analytics:** They are reports like dashboards, data visualization, they have been widely used for some time and are the core applications of traditional BI. Descriptive analyzes look back and reveal what happened. However, one tendency is to include predictive analytics findings, such as future sales forecasts, in dashboards (Watson, 2014).
- **Predictive Analytics:** Suggest about what will happen in the future. Methods and algorithms for predictive analytics, such as regression analysis, machine learning, and

neural networks, have been around for some time. The ability to analyze new data sources, Big Data, creates additional opportunities for insight and is especially important for companies with large amounts of data. Golden Path analysis is an exciting new technique for predictive or analytics. It involves analyzing large amounts of behavioral data (that is, data associated with people's activities or actions) to identify patterns of events or activities that predict customer actions (Watson, 2014).

- **Prescriptive Analytics:** Predict what will happen, prescriptive analysis suggests what to do. Prescriptive analytics can identify optimal solutions, often for scarce resource allocation. It has also been researched in academia for a long time, but now being used more in revenue management it is becoming more common for organizations that have "perishable" assets such as rental cars, hotel rooms, and airplane seats. For example, Harrah's Entertainment, a leader in the use of analytics, has been using revenue management for hotel room rates for many years (Watson, 2014).
- **Big Data:** *"Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes" (Mills et al., 2012).*
- **Small Data:** "Small data connects people with timely, meaningful insights (derived from big data and/or "local" sources), organized and packaged – often visually – to be accessible, understandable, and actionable for everyday tasks".
- **Volume:** *"Large volume of data that either consume huge storage or consist of large number of records" (Russom, 2011).*
- **Variety:** The word ***'Variety'*** denotes the fact that Big Data originates from numerous sources that can be structured, semi-structured, or unstructured (Schroeck et al., 2012).
- **Velocity:** High data quality is an important Big Data requirement for better predictability in the trading environment (Schroeck et al., 2012).
- **Veracity:** High data quality is an important Big Data requirement for better predictability in the trading environment (Schroeck et al., 2012). Therefore, verification is necessary to generate authentic and relevant data, and to have the ability to filter incorrect data (Beulke, 2011).
- **Value:** It is the added value obtained by organizations, value is created only when data is analyzed and acted upon correctly. To do this, we must identify all the data that will help us in the best way to generate value.

- **Python:** Python is a general-purpose object-oriented programming language due to its extensive library that primarily enables the development of Big Data, Artificial Intelligence (AI), Data Science, Test Frameworks, and Web Development applications. Released in 1989, Python is easy to learn and a favorite with programmers and developers. Python is one of the most popular programming languages in the world, second only to Java and C (IBM, 2021).

- **R Language:** R is an Open-Source programming language that is optimized for statistical analysis and data visualization. Developed in 1992, R has a rich ecosystem with complex data models and elegant data reporting tools (IBM, 2021).
- **Java:** Java is an object-oriented programming language specifically designed to allow developers a continuity platform. It is an extremely popular language that runs on a virtual machine, allowing it to be run on any type of device without having to compile it repeatedly. Java was created by Sun MicroSystems in 1991, as a programming tool and an object-oriented language, allowing programmers to generate autonomous code fragments, which interact with other objects to solve a problem offering support for different technologies.
- **Open-Source:** Originally, the expression open source (or open source) referred to open-source software (OSS). Open-source software is code designed in a way that is accessible to the public: everyone can view, modify, and distribute the code in any way they see fit. Open-source software is developed in a decentralized and collaborative manner, so it relies on peer review and community production. In addition, it is usually more economical, flexible, and durable than its proprietary alternatives, since those in charge of its development are the communities and not a single author or a single company (Red Hat, 2021).
- **Architectural Design:** process of defining a collection of hardware and software components and their interfaces to establish the framework for the development of a computer system (ISO/IEC/IEEE 2465:2017, 2017).

REFERENCES

A. Cockburn (2008). "Design as Knowledge Acquisition", Retrieved October 9, 2017, from http://a.cockburn.us/1973.

Abrahamsson, P., Oza, N., & Siponen, M. T. (2010). Agile software development methods: A comparative review. Agile software development, 31-59.

Abrahamsson, P., Salo, O., Ronkainen, J., & Warsta, J. (2002). Agile software development methods: Review and analysis. arXiv preprint arXiv:1709.08439.

Abrahamsson, P., Warsta, J., Siponen, M. T., & Ronkainen, J. (2003, May). New directions on agile methods: a comparative analysis. In 25th International Conference on Software Engineering, 2003. Proceedings. (pp. 244-254).

Ackoff, R. L. (1962). Scientific method: Optimizing applied research decisions.

Adibuzzaman, M., DeLaurentis, P., Hill, J., & Benneyworth, B. D. (2017). Big data in healthcare–the promises, challenges and opportunities from a research perspective: A case study with a model database. In AMIA Annual Symposium Proceedings (Vol. 2017, p. 384). American Medical Informatics Association.

Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. (2016). How to improve firm performance using big data analytics capability and business strategy alignment?. International Journal of Production Economics, 182, 113-131.

Ali, M., Sholihah, P. I., Ahmed, K., & Prabandari, S. P. (2016). Small data and big data: combination make better decision. International Journal of Research in Management, Economics and Commerce, 6(10).

Alles, M., & Gray, G. L. (2016). Incorporating big data in audits: Identifying inhibitors and a research agenda to address those inhibitors. International Journal of Accounting Information Systems, 22, 44-59

Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., ... & Thomas, D. (2001). Manifesto for agile software development.

Beecham s., T. Hall, C. Britton, M. Cottee, and A. Rainer. Using an expert panel to validate a requirements process improvement model. Journal of Systems and Software, 76(3): 251–275, 2005.

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., et al. (2001). Manifesto for Agile Software Development. http://agilemanifesto.org/

Beulke, D. (2011). Big data impacts data management: The 5 vs of big data. Available from: Big Data Impacts Data Management: The 5Vs of Big Data, accessed, 21.

Bertolucci, J. (2013). Big data analytics: Descriptive vs. predictive vs. prescriptive. Information Week, 31.

Bichler, M., Heinzl, A., & van der Aalst, W. M. P. (2016). Business analytics and data science.

Boehm, B. (2002). Get ready for agile methods, with care. Computer, 35(1), 64-69.

Boehm, B., & Turner, R. (2003, June). Observations on balancing discipline and agility. In Proceedings of the Agile Development Conference, 2003. ADC 2003 (pp. 32-39). IEEE.

Boehm, B., & Turner, R. (2003). Using risk to balance agile and plan-driven methods. Computer, 36(6), 57-66.

Boehm, B., & Turner, R. (2004, May). Balancing agility and discipline: Evaluating and integrating agile and plan-driven methods. In Proceedings. 26th International Conference on Software Engineering (pp. 718-719). IEEE.

Bonde, A. (2013). Defining small data. Small Data Group.

Bourque, P., & Fairley, R. E. (2014). Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0. IEEE Computer Society Press.

Boyd, A. (2012). Revisiting 'what is analytics'. Analytics Magazine. Recuperado de https://doi.org/10.1287/LYTX.2012.04.09

Brodie, M. L. (2015, June). Understanding Data Science: An Emerging Discipline for Data Intensive Discovery. In DAMDID/RCDL (pp. 238-245).

Burton, S. (2021). Data governance: The path to a data-driven culture. Applied Marketing Analytics, 6(4), 298-308.

Campanelli, A. S., & Parreiras, F. S. (2015). Agile methods tailoring–A systematic literature review. Journal of Systems and Software, 110, 85-100.

Cao, L. (2017a). Data science: a comprehensive overview. ACM Computing Surveys (CSUR), 50(3), 1-42.

Cao, L. (2017b). Data science: challenges and directions. Communications of the ACM, 60(8), 59-68.

Chang, H. C., Wang, C. Y., & Hawamdeh, S. (2019). Emerging trends in data analytics and knowledge management job market: extending KSA framework. Journal of Knowledge Management.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS quarterly, 1165-1188.

Chin, W. W. (2009). How to write up and report PLS analyses. In Handbook of partial least squares: Concepts, methods and applications (pp. 655-690). Berlin, Heidelberg: Springer Berlin Heidelberg.

Chow, T., & Cao, D.-B. (2008). A survey study of critical success factors in agile software projects. Journal of Systems and Software, 81(6), 961–971. https://doi.org/10.1016/j.jss.2007.08.020

Cockburn, A. (2000). Selecting a project's methodology. IEEE Software, 17(4), 64–71. https://doi.org/10.1109/52.854070

Cockburn, Alistair, & Highsmith, J. (2001). Agile software development: The people factor. Computer, 34, 131–133. https://doi.org/10.1109/2.963450

Conboy, K. (2009). Agility from first principles: Reconstructing the concept of agility in information systems development. Information systems research, 20(3), 329-354.

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. Knowledge in society, 1(1), 104.

Cox, M., & Ellsworth, D. (1997, August). Managing big data for scientific visualization. In ACM siggraph (Vol. 97, pp. 21-38).

Darwish, T. S., & Bakar, K. A. (2018). Fog based intelligent transportation big data analytics in the internet of vehicles environment: motivations, architecture, challenges, and critical issues. IEEE Access, 6, 15679-15701.

Davenport, T. H. (2006). Competing on analytics. Harvard business review, 84(1), 98.

Davenport, T. H., & Dyché, J. (2013). Big data in big companies. International Institute for Analytics, 3(1-31).

Davenport, T. H., & Harris, J. G. (2007). The architecture of business intelligence. Competing on analytics: The new science of winning.

Davenport, T. H., Barth, P., & Bean, R. (2012). How'big data'is different.

Davenport, T., & Harris, J. (2017). Competing on analytics: Updated, with a new introduction: The new science of winning. Harvard Business Press.

Davoudian, A., & Liu, M. (2020). Big data systems: A software engineering perspective. ACM Computing Surveys (CSUR), 53(5), 1-39.

Delen, D. (2020). Predictive Analytics uCertify Labs Access Code Card: Data Mining, Machine Learning and Data Science for Practitioners. FT Press.

Delen, D., & Ram, S. (2018). Research challenges and opportunities in business analytics. Journal of Business Analytics, 1(1), 2-12.

Demirkan, H., & Dal, B. (2014). Why do so many analytics projects still fail? Key considerations for deep

analytics on big data, learning and insights. INFORMS Analytics, 44-52,

Demirkan, H., & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decision Support Systems, 55(1), 412-421.

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64-73.

Domino Data Lab, Managing Data Science Teams (2017). https://domino.ai/resources/managing-data-science

Doumar, A. Small data and sustainable development goals: A review. https://www.researchgate.net/publication/328392596_Small_data_and_sustainable_development_goals_A_review

Dudziak, T. (1999). Extreme programming an overview. Methoden und Werkzeuge der Software: produktion WS, 2000, 2000.

Dybå. T. An instrument for measuring the key factors of success in software process improvement. Empirical software engineering, 5(4):357–390, 2000.

Estrin, D.: Small data, where n= me. Commun. ACM. 57(4), 32–34 (2014)

F. Tripp, J., & Armstrong, D. J. (2018). Agile methodologies: organizational adoption motives, tailoring, and performance. Journal of Computer Information Systems, 58(2), 170-179.

Forrester Inc. (2019). The Forrester WaveTM: Enterprise Insight Platforms, Q1 2019 - The Nine Providers That Matter Most And How They Stack Up. Internet document at:

https://www.forrester.com/report/The+Forrester+Wave+Enterprise+Insight+Platfo r ms+Q1+2019/-/E-RES141393.

Fowler, M., & Highsmith, J. (2001). The agile manifesto. Software Development, 9(8), 28-35.

Fu, H., Manogaran, G., Wu, K., Cao, M., Jiang, S., & Yang, A. (2020). Intelligent decision-making of online shopping behavior based on internet of things. International Journal of Information Management, 50, 515-525.

Fuggetta, A., JI Processo Software, Aspetti strategici e organizzativi, il Cardo edítore in Venezia, 1995

G. Piatetsky-Shapiro, “CRISP-DM, still the top methodology for analytics, datamining, or data science projects”, https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology- analytics-data-mining-data-science-projects.html, 2014, accessed online October 9, 2017.

Galvan, S., Mora, M., & Laporte, C. Y. (2021). Reconciliation of scrum and the project management process of the ISO/IEC 29110 standard-Entry profile—An experimental evaluation through usability measures | SpringerLink. https://link.springer.com/article/10.1007/s11219-021-09552-3

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), 137-144.

Gartner, 2014. Gartner survey reveals that 73 percent of organizations have invested or plan to invest in Big Data in the next two years. Press Release, September 17th http://www.gartner.com/newsroom/id/2848718.

Gartner, Gartner Top 10 Data and Analytics Trends for 2021, https://www.gartner.com/smarterwithgartner/gartner-top-10-data-and-analytics-trends-for-2021

Gentile, B. (2012). Top 5 myths about big data.

Grady, N. W., Payne, J. A., & Parker, H. (2017, December). Agile big data analytics: AnalyticsOps for data science. In 2017 IEEE international conference on big data (big data) (pp. 2331-2339). IEEE.

Haakman, M., Cruz, L., Huijgens, H., & van Deursen, A. (2021). AI lifecycle models need to be revised. Empirical Software Engineering, 26(5), 1-29.

Halpern, F. (2015). Next-Generation Analytics and Platforms for Business Success. TDWI Research Report. Available on www.tdwi.org

Harvard University. (s. f.). What is Data Science? | Definition, Skills, and Applications. Harvard John A. Paulson School of Engineering and Applied Sciences. Recuperado de https://seas.harvard.edu/news/what-data-science-definition-skills-applications-more

Hayashi, C. (1998). What is data science? Fundamental concepts and a heuristic example. In Data science, classification, and related methods (pp. 40-51). Springer, Tokyo.

Hekler, E. B., Klasnja, P., Chevance, G., Golaszewski, N. M., Lewis, D., & Sim, I. (2019). Why we need a small data paradigm. BMC medicine, 17(1), 1-9.

Heller, B., & Röthlisberger, M. (2015). Big data on trial: Researching syntactic alternations in GloWbE and ICE. In From Data to Evidence (d2e), Date: 2015/10/19-2015/10/22, Location: Helsinki.

Hevner, A. R. (2007). A three cycle view of design science research. Scandinavian journal of information systems, 19(2), 4.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS quarterly, 75-105.

Hoda, R., Salleh, N., Grundy, J., & Tee, H. M. (2017). Systematic literature reviews in agile software development: A tertiary study. Information and Software Technology, 85, 60-70.

Hsieh, C. Y., & Chen, C. T. (2015). (2015). Patterns for Continuous Integration Builds in Cross- Platform Agile Software Development. Journal of Information Science & Engineering, 31(3), 897- 924.

IBM (2016). IBM SPSS Modeler CRISP-DM Guide. Internet document at: ftp://public.dhe.ibm.xom/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf.

IBM, 2021 Python vs R: What´s the Difference? Retrived from: https://www.ibm.com/cloud/blog/python-vs-r

IBM. ASUM-DM. http://gforge.icesi.edu.co/ASUM-DM_External/index.htm#cognos.external.asum-DM_Teaser/deliveryprocesses/ASUM-DM_8A5C87D5.html

Idoine, C., Krensky, P., Brethenoux, E., Hare, J., Sicular, S., & Vashisth, S. (2018). Magic Quadrant for data science and machine-learning platforms. Gartner, Inc.

ISO/IEC, "ISO/IEC TR 29110-5-1-2:2011 - Software Engineering - Lifecycle Profiles for Very Small Entities (VSEs) - Part 5-1-2: Management and engineering guide - Generic pro-file group: Basic profile", International Organization for Standardization/International Electrotechnical Commission: Geneva, Switzerland, 2011.

IEEE, "ISO/IEC/IEEE International Standard - Systems and software engineering--Vocabulary," in ISO/IEC/IEEE 24765:2017(E) , vol., no., pp.1-541, 28 Aug. 2017, doi: 10.1109/IEEESTD.2017.8016712.

ISO. (2021, March 2). Economic benefits of standards. ISO. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/publication/10/04/PUB100403.html

J. Taylor, “Four problems in using CRISP-DM and how to fix them”, http://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html, accessed October 9, 2017.

Jacobs, A. (2009). The pathologies of big data. Communications of the ACM, 52(8), 36-44.

Jin, D. H., & Kim, H. J. (2018). Integrated understanding of big data, big data analysis, and business intelligence: a case study of logistics. Sustainability, 10(10), 3778.

Jyothi, V. E., & Rao, K. N. (2011). Effective implementation of agile practices. International Journal of Advanced Computer Science and Applications, 2(3).

Kaleshovska, N., Josimovski, S., Pulevska-Ivanovska, L., Postolov, K., & Janevski, Z. (2015). The Contribution Of Scrum In Managing Successful Software Development Projects. Economic Development / Ekonomiski Razvoj, 17(1/2), 175-194.

Karahanna E., D. W. Straub, and N. L. Chervany. Information technology adoption across time: a cross-sectional comparison of pre-adoption and post-adoption beliefs. MIS quar-terly, pages 183–213, 1999.

Katsis, Y., Balac, N., Chapman, D., Kapoor, M., Block, J., Griswold, W. G., ... & Patrick, K. (2017, July). Big data techniques for public health: a case study. In 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) (pp. 222-231). IEEE.

Kdnuggets, 2019 Python leads the 11 top Data Science, Machine Learning platforms: Trends and Analysis Retrived from: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

Keen, P. G., & Morton, S. (1978). MS (1978). Decision support systems: An organizational perspective, 264.

Kelly, J., & Kaskade, J. (2013). CIOS & BIG DATA what your IT team wants you to know. DOI= http://blog. infochimps. com/2013/01/24/cios-big-data.

Kendall, J. E., & Kendall, K. E. (2005). Agile Methodologies and the Lone Systems Analyst: When Individual Creativity and Organizational Goals Collide in the Global IT Environment. Journal of Individual Employment Rights, 11(4), 333-347.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. Dialogues in human geography, 3(3), 262-267.

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Sage.

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. GeoJournal, 80(4), 463-475.

Kose, B. (2021). Agile Business Analysis for Digital Transformation. In Handbook of Research on Multidisciplinary Approaches to Entrepreneurship, Innovation, and ICTs (pp. 98-123). IGI Global.

Kurtz, C. F., and Snowden, "The New Dynamics of Strategy: Sense- making in a Complex-Complicated World". IBM Systems Journal, vol. 42, no. 3, pp. 462-83 (2003).

Laporte, C. Y., & Miranda, J. M. (2020). Delivering Software- and Systems-Engineering Standards for Small Teams. Computer, 53(8), 79–83. https://doi.org/10.1109/MC.2020.2993331

Laporte, C. Y., Munoz, M., Miranda, J. M., & O'Connor, R. V. (2018). Applying Software Engineering Standards in Very Small Entities: From Startups to Grownups. IEEE Software, 35(1), 99–103. https://doi.org/10.1109/MS.2017.4541041

Laporte, C. Y., & O'Connor, R. V. (2016). A multi-case study analysis of software process improvement in very small companies using ISO/IEC 29110. In Systems, Software and Services Process Improvement: 23rd European Conference, EuroSPI 2016, Graz, Austria, September 14-16, 2016, Proceedings 23 (pp. 30-44). Springer

International Publishing.

Laporte, C., & O'Connor, R. (2017). Software process improvement standards and guides for very small organization: An overview of eight implementations. CrossTalk, The Journal of Defense Software Engineering, 30(3), 23–27.

Laporte, Claude Y., & Munoz, M. (2021). Not Teaching Software Engineering Standards to Future Software Engineers-Malpractice? Computer, 54(5), 81–88. https://doi.org/10.1109/MC.2021.3064438

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. International Journal of Information Management, 36(5), 700-710.

Lin, Y. T., & Huang, S. J. (2018). The design of a software engineering lifecycle process for big data projects. IT Professional, 20(1), 45-52.

Lee, J. Y., Yoon, J. S., & Kim, B. H. (2017). A big data analytics platform for smart factories in small and medium-sized manufacturing enterprises: An empirical case study of a die casting factory. International Journal of Precision Engineering and Manufacturing, 18(10), 1353-1361.

Lee, T., Lee, H., Rhee, K. H., & Shin, U. S. (2014). The efficient implementation of distributed indexing with Hadoop for digital investigations on Big Data. Computer Science and Information Systems, 11(3), 1037-1054.

Lee D., J. Park, and J.-H. Ahn. On the explanation of factors affecting e-commerce adoption. ICIS 2001 Proceedings, page 14, 2001.

Lohr, S. & Singer, N. (2016) How Data Failed Us in Calling an Election, New York Times, November 10,

2016

Luellen, E. (2018). Why Data Science Succeeds or Fails. Towards Data Science. Internet document at: https://towardsdatascience.com/why-data-science-succeeds-or-fails-c24edd2d2f9

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. The knowledge engineering review, 25(2), 137-166.

Market and Markets (2020). Data Science Platform Market by Component (Platform and

Services), Business Function (Marketing, Sales, Logistics, and Customer Support), Deployment Mode (On-Premises and Cloud), Organization Size, Industry Vertical, and Region - Global Forecast to 2024. Internet document at: https://www.marketsandmarkets.com/pdfdownloadNew.asp?id=21532997

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., ... & Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering.

Martinez, I., Viles, E., & Olaizola, I. G. (2021). Data science methodologies: Current challenges and future approaches. Big Data Research, 24, 100183.

Marx, V. (2013). The big challenges of big data. Nature, 498(7453), 255-260.

McClure, R. M. (1968). NATO SOFTWARE ENGINEERING CONFERENCE 1968. 136.

Microsoft Team, Data Science Process Documentation (2017).

Microsoft, 2021 Azure Application Architecture Guide Retrived from: https://docs.microsoft.com/en-us/azure/architecture/browse/

Microsoft, 2025 Team Data Science Process Lifecycle. https://github.com/Azure/Microsoft-TDSP/blob/master/Docs/lifecycle-detail.md

Mills, S., Lucas, S., Irakliotis, L., Rappa, M., Carlson, T., & Perlowitz, B. (2012). Demystifying big data: a practical guide to transforming the business of government. TechAmerica Foundation, Washington.

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. Information systems research, 2(3), 192-222.

Mora M. (2009). Conceptual Research Method: a Description (in Spanish language). Technical Report UAA-DSI-01, Autonomous University of Aguascalientes, Mexico.

Mora, M., Steenkamp, A. L., Gelman, O., & Raisinghani, M. S. (2012). On IT and SwE Research Methodologies and Paradigms: A Systemic Landscape Review. In Research Methodologies, Innovations and Philosophies in Software Systems Engineering and Information Systems (pp. 149-164). IGI Global.

Muntean, M., & Surcel, T. (2013). Agile BI - The Future of BI. Informatica Economica, 17(3), 114-124. doi:10.12948/issn14531305/17.3.2013.10

Muñoz, M., Peña, A., Mejía, J., Gasca-Hurtado, G. P., Gómez-Alvarez, M. C., & Laporte, C. Y. (2020). Analysis of 13 implementations of the software engineering management and engineering basic profile guide of ISO/IEC 29110 in very small entities using different life cycles. Journal of Software: Evolution and Process, e2300.

N. Grady and W. Chang, "NIST Big Data Interoperablity Framework: Volume 1, Definitions", NIST Special Publication 1500-2, Eds. https://bigdatawg.nist.gov/V2_output_docs.php, accessed October 6, 2017.

Naur, P. (1974). Concise survey of computer methods. Petrocelli Books.

Newell, A., & Simon, H. A. (1972). Human problem solving (Vol. 104). Prentice-hall Englewood Cliffs, NJ. http://www.sci.brooklyn.cuny.edu/~kopec/cis718/fall_2005/2/Rafique_2_humanthinking.doc

Niazi M., D. Wilson, and D. Zowghi. A framework for assisting the design of effective software process improvement implementation strategies. Journal of Systems and Software, 78(2):204–222, 2005.

Núñez, A., Hendriks, J., Li, Z., De Schutter, B., & Dollevoet, R. (2014, October). Facilitating maintenance decisions on the Dutch railways using big data: The ABA case study. In 2014 ieee international conference on big data (big data) (pp. 48-53). IEEE

O'Connor, R. V., & Laporte, C. Y. (2017). The evolution of the ISO/IEC 29110 set of standards and guides. International Journal of Information Technologies and Systems Approach (IJITSA), 10(1), 1-21.

O'Connor, R. V., & Laporte, C. Y. (2017, January 1). The Evolution of the ISO/IEC 29110 Set of Standards and Guides [Article]. International Journal of Information Technologies and Systems Approach (IJITSA). www.igi-global.com/article/the-evolution-of-the-isoiec-29110-set-of-standards-and-guides/169765

Oktaba, H., & Ibargüengoitia González, G. (1998). Software process modeled with objects: Static view. Computación y Sistemas, 1(4), 228-238.

Olson, D.L. and D. Delen, 2008. Advanced Data MiningTechniques. 1st Edn., Springer Science and BusinessMedia, Berlin, ISBN-10: 354076917X, pp: 180.

O'Neil, C., & Schutt, R. (2013). Doing data science: Straight talk from the frontline. "

O'Reilly Media, Inc.".

P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. "CRISP-DM 1.0: Step-by-step data mining guide", The CRISP-DM Consortium, 2000.

Palfreyman, J. (2013). Big Data–Vexed by Veracity?.

Parnas, D. L. (2010). Risks of undisciplined development. Communications of the ACM, 53(10), 25–27. https://doi.org/10.1145/1831407.1831419

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012). Design Science Research Evaluation. In K. Peffers, M. Rothenberger, & B. Kuechler (Eds.), Design Science Research in Information Systems. Advances in Theory and Practice (Vol. 7286, pp. 398–410). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-29863-9_29

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.

Pino F. J., M. T. Baldassarre, M. Piattini, G. Visaggio, and D. Caivano. Mapping software acquisition practices from iso 12207 and cmmi. In International Conference on Evaluation of Novel Approaches to Software Engineering, pages 234–247. Springer, 2008.P. Y. Reyes-Delgado, M. Mora, H. A. Duran-Limon, L. C. Rodríguez-Martínez, R. V. O'Connor, and R. Mendoza-Gonzalez. The strengths and weaknesses of software architecture design in the rup, MSF, mbase and rup-soa methodologies: A conceptual review. Computer Standards & Interfaces , 47:24–41, 2016.

Pressman, R. S. (2015). Software engineering: A practitioner's approach (Eighth edition). McGraw-Hill Education.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.

Qumer, A., & Henderson-Sellers, B. (2008). An evaluation of the degree of agility in six agile methods and its applicability for method engineering. Information and Software Technology, 50(4), 280–295. https://doi.org/10.1016/j.infsof.2007.02.002

Rich S., 2012. Big Data Is a "New Natural Resource" Retrived from:

http://www.govtech.com/policy-management/Big-Data-Is-a-New-Natural-Resource-IBM-Says.html

Robbes, R., Vidal, R., & Bastarrica, M. C. (2013). Are software analytics efforts worthwhile for small companies? The case of Amisoft. IEEE software, 30(5), 46-53.

Rodríguez, L. C., Mora, M., Martin, M. V., O'Connor, R., & Alvarez, F. (2009). Process models of SDLCs: comparison and evolution. In Handbook of Research on Modern Systems Analysis and Design Technologies and Applications (pp. 76-89). IGI Global.

Runkler, T. A. (2020). Data analytics. Springer Fachmedien Wiesbaden.

Runkler, T. A. (2020). Data Visualization. In Data Analytics (pp. 37-59). Springer Vieweg, Wiesbaden.

Russom, P. (2011). Big data analytics. TDWI best practices report, fourth quarter, 19(4), 1-34.

Sargent. R. G. An introduction to verification and validation of simulation models. In Simulation Conference (WSC), 2013 Winter, pages 321–327. IEEE, 2013.

Sargent. R. G. Verification, validation, and accreditation: verification, validation, and accreditation of simulation models. In Proceedings of the 32nd conference on Winter simulation, pages 50–59. Society for Computer Simulation International, 2000.

SAS (2017). Introduction to SEMMA – chapter 20 -. In SAS® Enterprise MinerTM 14.3: Reference Help. SAS Company.

Sawant, N., Shah, H., Sawant, N., & Shah, H. (2013). Big data application architecture. Big Data Application Architecture Q & A: A Problem-Solution Approach, 9-28.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics: The real-world use of big data. IBM Global Business Services, 12(2012), 1-20.

Schwaber, K. (1997). Scrum development process. In Business object design and implementation (pp. 117-134). Springer, London.

Shaw, M. (2003, May). Writing good software engineering research papers. In 25th International Conference on Software Engineering, 2003. Proceedings. (pp. 726-736). IEEE.

Song, I. Y., & Zhu, Y. (2016). Big data and data science: what should we teach?. Expert

Systems, 33(4), 364-373.

Song, Y., Schreier, P. J., Ramírez, D., & Hasija, T. (2016). Canonical correlation analysis of high-dimensional data with very small sample support. Signal Processing, 128, 449-458.

Stobierski, T. (2021). What's the Difference Between Data Analytics & Data Science? Harvard Business School Online. https://online.hbs.edu/blog/post/data-analytics-vs-data-science

Sutherland, J., & Schwaber, K. (2020). The scrum guide. The definitive guide to scrum: The rules of the game. Scrum. org, 268.

Thinyane, M. (2017, June). Investigating an Architectural Framework for Small Data Platforms. In Proceedings of the 17th European Conference on Digital Government (ECDG 2017), Lisbon, Portugal (pp. 220-227).

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. Journal of Big Data, 2(1), 1-32.

Tsoy, M., & Staples, D. S. (2020). What are the critical success factors for agile analytics projects?. Information Systems Management, 38(4), 324-341.

Tukey, J. W. (1962). The future of data analysis. The annals of mathematical statistics, 33(1), 1-67.

Turner, R., & Boehm, B. (2003). People Factors in Software Management: Lessons From Comparing Agile and Plan-Driven Methods. 5.

Veeramachaneni, K. (2016). Why you're not getting value from your data science. Harvard Business Review, 7.

Viswanathan, V. (2014). Data Analytics with R: A Hands-on Approach. Infivista Incorporated.

Vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to Design Science Research. In Design Science Research. Cases (pp. 1-13). Springer, Cham.

W. Royce, "Managing the Development of Large Software Systems", Proceedings of IEEE WESCON, Aug. 26, 1970.

Watson, H. J. (2009). Tutorial: business intelligence–past, present, and future. Communications of the Association for Information Systems, 25(1), 39.

Watson, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. Communications of the Association for Information Systems, 34(1), 65.

Wilson, J. & Daughterty, P. (2020). Small Data Can Play a Big Role in AI. Internet document at: https://hbr.org/2020/02/small-data-can-play-a-big-role-in-ai

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., Wesslén, A., ... & Wesslén, A. (2012). Empirical strategies. Experimentation in Software Engineering, 9-36.

Yan, Y., Huang, C., Wang, Q., & Hu, B. (2020). Data mining of customer choice behavior in internet of things within relationship network. International Journal of Information Management, 50, 566-574.

APPENDIXES

A.- Selective search

No.	Journal Name	Acronym	JCR FILTER =(C.1.1 and C.1.2) and (C.1.3 or C.1.4))				Analytics OR "Data Science" OR "Big Data" OR "Data Mining").AND.PERIOD(2000,2023)))	ARTICLES.FILTER.2: (ABSTRACT((Type 1: reports a new development methodology).OR(Type 2: reports a review of methodologies).OR(Typ3 3 : report a general survey))).				
			C.1.1:(MASTER LIST OF JCRs of Clarivate with the filter.0 (collection(SCIE and SSIE) and category(computer science.artificial intelligence))	C.1.2:(Impact Factor >= 1.000)	C.1.3:(analytics, data science, or big data topic, or data mining is listed in JCR about)	C.1.4:(JCR is additionally recommended by research team)	Quantity of located articles applying FILTER.1	Quantity of located articles applying FILTER.2 (Type 1)	Quantity of located articles applying FILTER.2 (Type 2)	Quantity of located articles applying FILTER.3 (Type 2)	X	
1	ACM Computing Surveys	ACM CS	-	14.324	-	YES	41	0	0	3	38	41
2	Journal of Big Data	JBD	YES	10.835	YES	-	101	0	0	2	99	101
3	ACM Transactions on Intelligent Systems and Technology	ACM TIST	YES	10.489	YES	-	25	0	0	0	25	25
4	Artificial Intelligence Review	AIR	YES	9.588	-	-	24	0	0	1	23	24
5	IEEE Transactions on Knowledge and Data Engineering	IEEE TKDE	YES	9.235	-	-	120	0	1	1	118	120
6	International Journal of Intelligent Systems	IJIS	YES	8.993	-	-	27	0	0	0	27	27
7	Computer Science Review	CSR	-	8.757	-	YES	7	0	0	1	6	7
8	Expert Systems with Applications	ESWA	YES	8.665	-	-	350	0	1	0	349	350
9	Knowledge-based Systems	KBS	YES	8.139	YES	-	128	0	0	0	128	128
10	Decision Support Systems	DSS	YES	6.969	-	-	97	0	0	0	97	97
11	IEEE Intelligent Systems	IEEE IS	YES	6.744	-	-	59	0	0	1	58	59
12	Journal of Management Analytics	JMA	YES	6.554	YES	-	30	0	0	1	29	30
13	Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery	WDMKD	YES	5.406	YES	-	90	0	1	0	89	90
14	Big Data	BD	YES	4.426	YES	-	132	0	0	0	132	132
15	IEEE Transactions on Big Data	IEEE TBD	YES	4.271	YES	-	144	0	0	0	144	144
16	ACM Transactions on Knowledge Discovery from Data	ACM TKDD	YES	4.157	YES	-	29	0	0	0	29	29
17	Big Data Research	BDR	YES	3.739	YES	-	82	0	1	0	81	82
18	International Journal of Information Technology & Decision Making	IJITDM	YES	3.508	-	-	34	0	0	0	34	34
19	Patterns	PAT	-	3.194	-	YES	63	0	0	0	63	63
20	Journal of Business Analytics (*)	JBA	-	3.000	YES	YES	23	0	0	1	22	23
21	Expert Systems	ES	YES	2.812	YES	-	41	0	0	1	40	41
22	International Journal of Data Science and Analytics (*)	IJDSA	-	2.520	YES	YES	57	0	0	0	57	57
23	PeerJ Computer Science	PJCS	YES	2.411	YES	-	16	0	1	0	15	16
24	International Journal of Computational Intelligence Systems	IJCIS	YES	2.259	YES	-	20	0	0	0	20	20
25	Computer IEEE	IEEE COMP	-	2.256	-	YES	118	0	0	0	118	118
26	Data & Knowledge Engineering	DKE	YES	1.500	-	-	34	0	0	2	32	34
27	Intelligent Data Analysis	IDA	YES	1.321	YES	-	59	0	0	1	58	59
							1951	0	5	15	1931	1951

	SWE Journals							It was consulted the MASTER LIST OF JCRs of Clarivate with the filter.0 (collection(SCIE and SSIE) an category(computer science. software engineering)) with a total of 109 journals.				
	Journal Name	Acronym	JCR FILTER = (C.1.1 and C.1.2) and (C.1.3 or C.1.4))				ARTICLES. FILTER 1: (TITLE. INCLUDES(Analytics OR "Data Science" OR "Big Data" OR "Data Mining").AND. PERIOD(2000 ,2023)))	ARTICLES. FILTER 2: (ABSTRACT((Type 1: reports a new development methodology). OR(Type 2: reports a review of methodologies). OR(Typ3 3 : report a general survey))).				
			C.1.1:(MASTER LIST OF JCRs of Clarivate with the filter.0 (collection(SCIE and SSIE) and category(computer science. software engineering)	C.1.2:(Impact Factor >= 1.000)	C.1.3:(software process or equivalent topic is listed in JCR about)	C.1.4:(JCR is additionally recommended by research team)	Quantity of located articles applying FILTER.1	Quantity of located articles applying FILTER.2(Type 1)	Quantity of located articles applying FILTER.2(Type 2)	Quantity of located articles applying FILTER.2(Type 3)	X	
1	Communications of the ACM	CACM	YES	14.065	-	YES	68	0	0	4	64	68
2	IEEE Transactions on Software Engineering	IEEE.TSE	YES	9.322	YES	-	10	0	0	0	10	10
3	Information and Software Technology	IST	YES	3.862	YES	-	13	0	0	3	10	13
4	Empirical Software Engineering	ESE	YES	3.762	YES	-	12	0	0	1	11	12
5	Computers Standards & Interfaces	CSI	YES	3.721	YES	-	12	0	0	1	11	12
6	ACM Transactions on Software Engineering and Methodology	ACM.TOSEM	YES	3.685	YES	-	0	0	0	0	0	0
7	Journal of Systems and Software	JSS	YES	3.514	YES	-	39	0	0	2	37	39
8	Software: Practice and Experience	SPE	YES	3.200	YES	-	34	0	0	1	33	34
9	IEEE Software	IEEE.SW	YES	3.000	YES	-	35	0	0	3	32	35
10	IT Professional IEEE	IEEE IT PROF	YES	2.590	-	YES	47	1	0	1	45	47
11	Journal of the ACM	JACM	YES	2.269	-	YES	0	0	0	0	0	0
12	Software and Systems Modeling	SOSYM	YES	2.211	YES	-	2	0	0	0	2	2
13	Journal of Software-Evolution and Process	JSEP	YES	1.864	YES	-	5	0	0	0	5	5
14	Software Quality Journal	SQJ	YES	1.813	YES	-	3	0	0	0	3	3
15	Computer Science and Information Systems	CSIS	YES	1.170	-	YES	21	0	0	0	21	21
16	International Journal of Electronic Commerce	IJEC	YES	1.170	-	YES	8	0	0	0	8	8
17	IET Software	IET SOFT	YES	1.150	YES	-	9	0	0	0	9	9
18	Journal of Universal Computer Science	JUCS	YES	1.056	-	YES	19	0	0	0	19	19
19	International Journal of Software Engineering and Knowledge Engineering	IJSEKE	YES	1.007	YES	-	14	0	0	2	12	14
							351	1	0	18	332	351

	GREY LITERATURE SOURCES ADDED BY RESEARCH TEAM											
1	Microsoft - web site (TDSP)			-	-	YES	-	1	-	-	-	1
2	web site : IBM (ASUM)			-	-	YES	-	1	-	-	-	1
3	web site : Domino Lab (DDSL)			-	-	YES	-	1	-	-	-	1
4	web site : SAS (SEMMA)			-	-	YES	-	1	-	-	-	1
5	KDD //			-	-	YES	-	1	-	-	-	1
6	web site :CRIsp-DM (SPSS-IBM)			-	-	YES	-	1	-	-	-	1

Table Final. Set of 7 studies on BDAS Development Life Cycles

Type of PAIS/|SDLC Life Cycle	Publication Domain	Publication Name	Type of Publication	Publication IF	Publication Year	Study Title	Citations
Heavyweight	Data Science	AI Magazine	Grey Literature	-	1996	From data mining to knowledge discovery in databases.	12666
Heavyweight	Data Science	SPSS Inc. Web Site	Grey Literature	-	2000	CRISP-DM 1.0: Step-by-step data mining guide.	2017
Heavyweight	Data Science	SAS institute Web Site	Grey Literature	-	2003	Data Mining Using SAS Enterprise Miner: A Case Study Approach.	8
Heavyweight	Software Engineering	IEEE IT PROF	JCR journal	2.590	2018	"The design of a software engineering lifecycle process for big data projects."	12
Agile	Data Science	IBM Analytics Web Site	Grey Literature	-	2016	Analytics Solutions Unified Method: Implementations with Agile principles.	3
Agile	Data Science	Microsoft Azure Web Site	Grey Literature	-	2017	"What is the team data science process?"	19
lightweight	Data Science	Domino Data Lab Web Site	Grey literature	-	2017	The Practical Guide to Managing Data Science at Scale	0

B.- ISO/IEC TR 29110-5-1-2:2011 (Roles, Products and Software Tools)

8 Roles

This is an alphabetical list of the roles, its abbreviations and suggested competencies description. This list is showed as a four-column table for presentation purpose only.

Table 22 — Roles

	Role	Abbreviation	Competency
1.	Analyst	AN	Knowledge and experience eliciting, specifying and analyzing the requirements. Knowledge in designing user interfaces and ergonomic criteria. Knowledge of the revision techniques. Knowledge of the editing techniques. Experience on the software development and maintenance.
2.	Customer	CUS	Knowledge of the Customer processes and ability to explain the Customer requirements. The Customer (representative) must have the authority to approve the requirements and their changes. The Customer includes user representatives in order to ensure that the operational environment is addressed. Knowledge and experience in the application domain.
3.	Designer	DES	Knowledge and experience in the *Software Components* and architecture design. Knowledge of the revision techniques. Knowledge and experience in the planning and performance of integration tests. Knowledge of the editing techniques. Experience on the software development and maintenance.
4.	Programmer	PR	Knowledge and/or experience in programming, integration and unit tests. Knowledge of the revision techniques. Knowledge of the editing techniques. Experience on the software development and maintenance.
5.	Project Manager	PM	Leadership capability with experience making decisions, planning, personnel management, delegation and supervision, finances and software development.
6.	Technical Leader	TL	Knowledge and experience in the software process domain.

	Role	Abbreviation	Competency
7.	Work Team	WT	Knowledge and experience according to their roles on the project: TL, AN, DES, and/or PR. Knowledge on the standards used by the Customer and/or by the VSE.

9 Product description

This is an alphabetical list of the input, output and internal process products, its descriptions, possible states and the source of the product. The source can be another process or an external entity to the project, such as the Customer. This list is showed as a four-column table for presentation purpose only. Product items in the following tables are based on ISO/IEC 15289 Information Items with some exceptions.

Table 23 — Product Descriptions

	Name	Description	Source
1.	*Acceptance Record*	Documents the Customer acceptance of the *Deliverables* of the project. It may have the following characteristics: - Record of the receipt of the delivery - Identifies the date received - Identifies the delivered elements - Records the verification of any Customer acceptance criteria defined - Identifies any open issues (if applicable) - Signed by receiving Customer	Project Management
2.	*Change Request*	Identifies a *Software*, or documentation problem or desired improvement, and requests modifications. It may have the following characteristics: - Identifies purpose of change - Identifies request status - Identifies requester contact information - Impacted system(s) - Impact to operations of existing system(s) defined - Impact to associated documentation defined - Criticality of the request, date needed The applicable statuses are: initiated, evaluated, and accepted.	Software Implementation Customer Project Management
3.	*Correction Register*	Identifies activities established to correct a deviation or problem concerning the accomplishment of a plan. It may have the following characteristics: - Identifies the initial problem - Defines a solution - Identifies corrective actions taken - Identifies the ownership for completion of defined actions - Identifies the open date and target closure date - Contains a status indicator - Indicates follow up actions	Project Management

	Name	Description	Source
4.	*Maintenance Documentation*	Describes the *Software Configuration* and the environment used for development and testing (compilers, design tools, construction and tests). It may have the following characteristics: - Includes or refers to all S*oftware Configuration* elements developed during implementation - Identifies environment used for development and testing (compilers, design tools, construction and tests tools) It is written in terms that maintenance personnel can understand. The applicable statuses are: verified and baselined.	Software Implementation
5.	*Meeting Record*	Records the agreements established with Customer and/or Work Team. It may have the following characteristics: - Purpose of meeting - Attendees - Date, place held - Reference to previous minutes - What was accomplished - Identifies issues raised - Any open issues - Agreements - Next meeting, if any. The applicable status is: updated.	Project Management

6.	*Product Operation Guide*	Contains the necessary information to install and manage the *Software*. It may have the following characteristics:
		
		- Criteria for operational use
		- A description of how to operate the product including:
		  - operational environment required
		  - supporting tools and material (e.g. user manuals) required
		  - possible safety warnings
		  - start-up preparations and sequence
		  - frequently asked questions (FAQ)
		  - sources of further information and help to operate the product
		- Certification and safety approvals
		- Warranty and replacement instructions
		- It should be written in terms that the personnel responsible for the operation can understand.
		
		The applicable statuses are: verified and baselined.
			Software Implementation

	Name	Description	Source
7.	*Progress Status Record*	Records the status of the project against the *Project Plan*. It may have the following characteristics: - Status of actual *Tasks* against planned *Tasks* - Status of actual results against established *Objectives* / goals - Status of actual resource allocation against planned *Resources* - Status of actual cost against budget estimates - Status of actual time against planned schedule - Status of actual risk against previously identified - Record of any deviations from planned *Tasks* and reason why. The applicable status is: evaluated.	Project Management

8.	*Project Plan*	Presents how the project processes and activities will be executed to assure the project's successful completion, and the quality of the deliverable products. It Includes the following elements which may have the characteristics as follows:
		
		- *Product Description*
		  - Purpose
		  - General Customer requirements
		- *Scope* description of what is included and what is not
		- *Objectives* of the project
		- *Deliverables* - list of products to be delivered to Customer
		- *Tasks*, including verification, validation and reviews with Customer and Work Team, to assure the quality of work products. *Tasks* may be represented as a Work Breakdown Structure (WBS).
		- *Estimated Duration* of tasks
		- *Resources* (humans, materials, standards, equipment and tools) including the required training, and the schedule when the *Resources* are needed.
		- *Composition of Work Team*
		- *Schedule of the Project Tasks,* the expected start and completion date for each task, and the relationship and dependencies of the *Tasks*.
		- *Estimated Effort and Cost*
		- *Identification of Project Risks*
		- *Version Control Strategy*
		  - Product repository tools or mechanism identified
		  - Location and access mechanisms for the repository specified
		  - Version identification and control defined
		  - Backup and recovery mechanisms defined
		  - Storage, handling and delivery (including archival and retrieval) mechanisms specified
		- *Delivery Instructions*
		  - Elements required for product release identified (i.e., hardware, software, documentation etc.)
		  - Delivery requirements
		  - Sequential ordering of *Tasks* to be performed
		  - Applicable releases identified
		  - Identifies all delivered *Software Components* with version information
			Project Management

	Name	Description	Source
		- Identifies any necessary backup and recovery procedures The applicable statuses are: verified, accepted, updated and reviewed.	
9.	*Project Repository*	Electronic container to store project work products and deliveries. It may have the following characteristics: - Stores project work products - Stores released *Deliverables* products - Storage and retrieval capabilities - Ability to browse content - Listing of contents with description of attributes - Sharing and transfer of work products between affected groups - Effective controls over access - Maintain work products descriptions - Recovery of archive versions of work products - Ability to report work products status - Changes to work products are tracked to *Change Requests* The applicable statuses are: recovered and updated.	Project Management
10.	*Project Repository Backup*	Repository used to backup the *Project Repositor*y and, if necessary, to recover the information.	Project Management

11.	*Requirements Specification*	Identifies the software requirements. It may have the following characteristics:
		
		- Introduction –general description of *Software* and its use within the *Scope* of the Customer business;
		- Requirements description:
		  - Functionality – established needs to be satisfied by the *Software* when it is used in specific conditions. Functionality must be adequate, accurate and safe
		  - User interface – definition of those user interface characteristics that allow to understand and learn the *Software* easily so the user be able to perform his/her *Tasks* efficiently including the interface exemplar description
		  - External interfaces – definition of interfaces with other software or hardware
		  - Reliability – specification of the software execution level concerning the maturity, fault tolerance and recovery
		  - Efficiency – specification of the software execution level concerning the time and use of the *Resources*
		  - Maintenance – description of the elements facilitating the understanding and execution of the future *Software* modifications
		  - Portability – description of the *Software* characteristics that allow its transfer from one place to other
			Software Implementation

	Name	Description	Source
		- Design and construction limitations/constraints – needs imposed by the Customer - Interoperability – capability for two or more systems or *Software Components* be able to change information each other and use it - Reusability – feature of any product/sub-product, or a part of it, so that it can be used by several users as an end product, in the own software development, or in the execution of other software products - Legal and regulative – needs imposed by laws, regulations, etc. Each requirement is identified, unique and it is verifiable or can be assessed. The applicable statuses are: verified, validated and baselined.	
12.	*Software*	Software item (*Software* source and executable code) for a Customer, constituted by a collection of integrated *Software Components*. The applicable statuses are: tested and baselined.	Software Implementation
13.	*Software Components*	A set of related code units. The applicable statuses are: unit tested, corrected and baselined.	Software Implementation
14.	*Software Configuration*	A uniquely identified and consistent set of software products including: - *Requirements Specification* - *Software Design* - *Traceability Record* - *Software Components* - *Software* - *Test Cases and Test Procedures* - *Test Report* - *Product Operation Guide* - *Software User Documentation* - *Maintenance Documentation* The applicable statuses are: delivered and accepted.	Software Implementation

15.	*Software Design*	Textual and graphical information on the *Software* structure. This structure may include the following parts:
		
		Architectural high level software design – Describes the overall *Software* structure:
		- Identifies the required *Software Components*
		- Identifies the relationship between *Software Components*
		- Consideration is given to any required:
		  - *Software* performance characteristics
		  - hardware, software and human interfaces
		  - security characteristics
		  - database design requirements
		  - error handling and recovery attributes
			Software Implementation

	Name	Description	Source
		Detailed low level software design – includes details of the *Software Components* to facilitate its construction and test within the programming environment; - Provides detailed design (could be represented as a prototype, flow chart, entity relationship diagram, pseudo code, etc.) - Provides format of input / output data - Provides specification of data storage needs - Establishes required data naming conventions - Defines the format of required data structures - Defines the data fields and purpose of each required data element - Provides the specifications of the program structure The applicable statuses are: verified and baselined.	
16.	*Software User Documentation*	Describes the way of using the *Software* based on the user interface. It may have the following characteristics: - User procedures for performing specified *Tasks* using the *Software* - Installation and de-installation procedures - Brief description of the intended use of the *Software* (the concept of operations) - The supplied and required *Resources* - Needed operational environment - Availability of problem reporting and assistance - Procedures to access and exit the *Software* - Lists and explains *Software* commands and system-provided messages to the user - As appropriate for the identified risk, it includes warnings, cautions, and notes, with corrections - It includes troubleshooting and error correction procedures. It is written in terms understandable by users. The applicable statuses are: preliminary, verified and baselined.	Software Implementation

17.	*Statement of Work*	Description of work to be done related to *Software* development. It may Include:
		
		- *Product Description*
		  - Purpose
		  - General Customer requirements
		- *Scope* description of what is included and what is not
		- *Objectives* of the project
		- *Deliverables* list of products to be delivered to Customer
		
		The applicable status is: reviewed.
			Customer

	Name	Description	Source
18.	*Test Cases and Test Procedures*	Elements needed to test code. Test Case may include: - Identifies the test case - Test items - Input specifications - Output specifications - Environmental needs - Special procedural requirements - Interface dependencies Test Procedures may include: - Identifies: test name, test description and test completion date - Identifies potential implementation issues - Identifies the person who completed the test procedure - Identifies prerequisites - Identifies procedure steps including the step number, the required action by the tester and the expected results The applicable statuses are: verified and baselined.	Software Implementation
19.	*Test Report*	Documents the tests execution. It may include: - A summary of each defect - Identifies the related test case - Identifies the tester who found each defect - Identifies the severity for each defect - Identifies the affected function(s) for each defect - Identifies the date when each defect originated - Identifies the date when each defect was resolved - Identifies the person who resolved each defect The applicable status is: baselined.	Software Implementation
20.	*Traceability Record*	Documents the relationship among the requirements included in the *Requirements Specification, Software Design* elements, *Software Components, Test Cases and Test Procedures.* It may include: - Identifies requirements of *Requirements Specification* to be traced - Provides forward and backward mapping of requirements to *Software Design* elements, *Software Components*, *Test Cases and Test Procedures*. The applicable statuses are: verified, baselined and updated.	Software Implementation

	Name	Description	Source
21.	*Verification Results*	Documents the verification execution. It may include the record of: - Participants - Date - Place - Duration - Verification check-list - Passed items of verification - Failed items of verification - Pending items of verification - Defects identified during verification	Project Management Software Implementation
22.	*Validation Results*	Documents the validation execution, It may include the record of: - Participants - Date - Place - Duration - Validation check-list - Passed items of validation - Failed items of validation - Pending items of validation - Defects identified during validation	Software Implementation

10 Software tools requirements

Software tools that could be used to perform process activities.

10.1 Project Management process

Table 24 — Project Management tools

Activity	Resource List
Project Planning Project Plan Execution Project Assessment and Control Project Closure	Tools allowing document, manage and control the *Project Plan* and the use and management of the *Project Repository*

10.2 Software Implementation process

Table 25 — Software Implementation tools

Activity	Resource List
Software Implementation Initiation Software Requirements Analysis Software Architectural and Detailed Design Software Construction Software Integration and Tests Product Delivery	Documentation tools
Software Requirements Analysis	*Requirements Specification* tools
Software Architectural and Detailed Design	*Software Design* tools
Software Construction	Construction Tools
Software Integration and Tests	Tests tools, bug tracking tools

C.- Design of the Artifact Methodology.

Once the Theoretical Design Sources were selected, Design Components were chosen from the Roles, Activities and Products that could help the design of the BDAS Methodology.

Table C-1, Table C-2, Table C-3, and Table C-4 show the selected Design Components for the first selected Design Components. Once the Design Components were selected, the second iteration of the process reviewed each of the components and asked about their importance to be in the BDAS Methodology.

The second iteration provides the second wave of design components that could be important for the BDAS Methodology. The third and last iteration were processed to have the minimum Design Components for the Methodology. It is also important to say that there were a lot of Design Components that are using the same activities and products that use the DTS.2 CRISP-DM (Pete Chapman et al., 2000) so the Design Components for this DTS were selected and complemented with other Activities and Products.

Table C-1, Table C-2, Table C-3, and Table C-4 show the selected Design Components for the BDAS Methodology for the first and second iteration. Finally, chapter 4.3 contains the final tables for the last iteration.

C.1.- Roles for Desing Components first, second and Third iteration.

Roles							
Design Component	**Source**	**Name**	**Specific element (restructured)**	**SDLC that is also using it**			
				DTS.1	**DTS.2**	**DTS.3**	**DTS.4**
DC.4 ISO/IEC 29110 - Basic profile- Roles	DTS.1 ISO/IEC 29110 - Basic profile- (ISO/IEC, 2011)	{Analyst, Customer, Designer, Programmer, Project Manager, Technical Leader, Work Team}	**Customer**: Responsible for reviewing prioritisation, return on investment and providing expertise throughout the process. Confirm that the pipeline, the model and its deployment in a production environment meet the objectives.	O		/	X
			Project Manager: Manages the day-to-day activities of the Work Team on a specific data science project. Responsible for clearly articulating the business problem, at hand, connecting through domain knowledge about the business problem and translating that into day to day work. In addition, ensure training and continuous engagement with the deployed models.	O		X	X
			Work Team: Data Scientists, Business Analysts, Data Engineers, Architects, and others who execute a data science project.	O		X	
			Technical Leader: The Data Infrastructure Engineer (Data & Platform Architect) Building scalable pipelines and infrastructure that make it possible to develop and deploy models.	O		X	X
			Programmer: The Data Scientist Generating and communicating insights, understanding the strengths and weaknesses of algorithms and features.	O		X	X
			Desinger: The Data Engineer to perform the data engineering parts of the project. Responsible for building and maintaining the data infrastructure. This includes extraction, transformation and loading (ETL) of data, creation of data pipelines,	O		/	
			Analyst: Data Analyst subject matter experts who have a clear understanding of the problem. They must know the internal processes and practices so that they can help the analyst understand and interpret the data. They must also be able to make the necessary changes to existing business processes to help collect the right data for the problems, if needed.	O		X	

C.2.- Activities for Desing Components first, second and third iteration.

	Activities									
Design Component	**Source**	**DC**	**Specific element**	**SDLC that is also using it**				**Iteration**		
				DTS.1	**DTS.2**	**DTS.3**	**DTS.4**	**1**	**2**	**3**
DC.5 CRISP-DM Phases	DTS.2 CRISP-DM Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000)	{Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment}	Business Understanding		X	X	X	X	X	X
			Data Understanding		X	X	X	X	X	X
			Data Preparation		X	X	X	X	X	X
			Modeling		X	X	X	X	X	X
			Evaluation		X	X	X	X	X	X
			Deployment		X	X	X	X	X	X
DC.8 TDSP Lifecycle	DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016)	{ Business understanding, Data acquisition and understanding, Modeling, Deployment, Customer acceptance }	Business understanding		X	X	X	X	X	X
			Data acquisition and understanding		X	X	X	X	X	X
			Modeling		X	X	X	X	X	X
			Deployment		X	X	X	X	X	X
			Customer acceptance			X		X	X	
DC.12 DDSL Lifecycle	DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	{ Ideation, Data Acquisition and Exploration, Research and Development, Validation, Delivery, Monitoring }	Ideation		X	X	X	X	X	X
			Data Acquisition and Exploration		X	X	X	X	X	X
			Research and Development		X	X	X	X	X	X
			Validation		X	X	X	X	X	X
			Delivery		X	X	X	X	X	X
			Monitoring				X	X	X	

C.3.- Tasks for Desing Components first, second and third iteration.

	Tasks									
Design Component	Source	DC	Specific element	SDLC that is also using it (DTS)				Iteration		
				1	2	3	4	1	2	3
DC.6 CRISP-DM Tasks	DTS.2 CRISP-DM: Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000)	Business Understanding [Determine Business Objectives, Assess Situation, Determine Data Mining Goals, Produce Project Plan]	Determine Business Objectives		X	X	X	X	X	X
			Assess Situation		X	X	X	X	X	X
			Determine Data Mining Goals		X	X	X	X	X	X
			Produce Project Plan		X		X	X	X	X
		Data understanding [Collect Initial Data, Describe Data, Explore Data, Verity Data Quality],	Collect Initial Data		X	X	X	X	X	X
			Describe Data		X	X	X	X	X	X
			Explore Data		X	X	X	X	X	X
			Verity Data Quality		X	X		X	X	X
		Data preparation [Select Data, Clean Data, Construct Data , Integrate Data, Format Data]	Select Data		X	X	X	X	X	X
			Clean Data		X	X	X	X	X	X
			Construct Data		X		X	X	X	X
			Integrate Data		X		X	X	X	X
			Format Data		X	X		X	X	X
		Modeling [Select Modeling Technique, Generate Test Design, Build Model, Assess Model]	Select Modeling Technique		X	X	X	X	X	X
			Generate Test Design		X	X	X	X	X	X
			Build Model		X	X	X	X	X	X
			Assess Model		X	X	X	X	X	X
			Evaluate Results		X		X	X	X	X

		Evaluation [Evaluate Results, Review Process, Determine Next Stages]	Review Process		X		X	X	X	X
			Determine Next Stages		X	X	X	X	X	X
		Deployment [Plan Deployment, Plan Monitoring and Maintenance, Produce Final Report, Review Project]	Plan Deployment		X		X	X	X	X
			Plan Monitoring and Maintenance		X	X	X	X	X	X
			Produce Final Report		X	X	X	X	X	X
			Review Project		X			X	X	
DC.9 TDSP Tasks	DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016)	Business understanding [Define objectives,Identify data sources]	Define objectives,			X	X	X	X	X
			Identify data sources			X	X	X	X	X
		Data acquisition and understanding [Ingest the data, Explore the data, Set up a data pipeline]	Ingest the data		X	X	X	X	X	X
			Explore the data		X	X	X	X	X	X
			Set up a data pipeline			X		X		
		Modeling [Feature engineering, Model training, Model Evaluation]	Feature engineering		X	X	X	X	X	X
			Model training		X	X	X	X	X	X
			Model Evaluation		X	X	X	X	X	X
		Deployment [Operationalize a Model]	Operationalize a Model		X	X	X	X	X	X
		Customer acceptance [System Validation, Project hand-off]	System Validation		X	X	X	X	X	X
			Project hand-off		X	X	X	X	X	X
DC.13 DDSL Process	DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	Ideation [Identified Problem, Project Scoping, Review Prior Art, Calculate Value, Assess Feasibility, Manage Backlog, Select Artifacts]	Identified Problem		X	X	X	X	X	X
			Project Scoping		X		X	X		
			Review Prior Art				X	X		
			Calculate Value				X	X	X	
			Assess Feasibility		X	X	X	X		

			Manage Backlog		X		X	X	X	X
			Select Artifacts		X	X	X	X	X	X
		Data Acquisition and Exploration [Getting the Data, Identify Sources the Data, Connect, Create Data (Capture), Buy & Ingest DATA, Explore Data, Prepare Data]	Getting the Data		X	X	X	X	X	X
			Identify Sources the Data		X	X	X	X	X	X
			Connect				X	X	X	
			Create Data (Capture)				X	X	X	X
			Buy & Ingest DATA				X	X	X	X
			Explore Data		X	X	X	X	X	X
			Prepare Data		X		X	X	X	X
		Research and Development [Generate Hypothesis, Validate right tools, IT request, Experiment, Assess result, Validate the need new Data, Insightful?, Share insight]	Generate Hypothesis				X	X	X	
			Validate right tools				X	X	X	X
			IT request				X	X		
			Experiment		X	X	X	X	X	
			Assess result		X	X	X	X	X	X
			Validate the need new Data		X	X	X	X	X	X
			Insightful?				X	X		
			Share insight				X	X		
		Validation [Validate the Business, Validate Technically, Validate ready to Deploy, Publish]	Validate the Business		X		X	X	X	X
			Validate Technically				X	X	X	X
			Validate ready to Deploy		X	X	X	X	X	X
			Publish		X		X	X	X	
			Plan Delivery		X		X	X	X	X

		Delivery [Plan Delivery, Deploy, Test]	Deploy		X	X	X	X	X	X
			Test				X	X	X	X
		Monitoring [Monitor, Usage, Performance, Value, Identify Improvements, Generate Value]	Monitor		X	X	X	X	X	
			Usage				X	X		
			Performance				X	X		
			Value				X	X		
			Identify Improvements				X	X		
			Generate Value		X		X	X	X	

C.4.- Products for Desing Components first, second and third iteration.

	Products									
Design Component	**Source**	**DC**	**Specific element**	**SDLC that is also using it (DTS)**				**Iteration**		
				1	**2**	**3**	**4**	**1**	**2**	**3**
DC.7 CRISP-DM Outputs	DTS.2 CRISP-DM: Cross Industry Standard Process for Data Mining (Pete Chapman et al., 2000)	Business Understanding [Background , Business Objectives , Business Success Criteria, Inventory of Resources , Requirements Assumptions and Constraints , Risks and Contingencies , Terminology, Costs and Benefits, Data Mining Goals , Data Mining Success Criteria, Project Plan, Initial Assessment of Tools and Techniques],	Background		X			X		
			Business Objectives		X	X	X	X	X	X
			Business Success Criteria		X			X		
			Inventory of Resources		X	X		X	X	X
			Requirements Assumptions and Constraints		X			X		
			Risks and Contingencies		X			X		
			Terminology		X			X		
			Costs and Benefits		X			X		
			Data Mining Goals		X			X	X	X
			Data Mining Success Criteria		X			X		
			Project Plan		X			X	X	X
			Initial Assessment of Tools and Techniques		X			X		
		Data Understanding [Initial Data Collection Report, Data Description Report, Data Exploration Report, Data Quality Report]	Initial Data Collection Report		X			X	X	X
			Data Description Report		X		X	X	X	X
			Data Exploration Report		X		X	X	X	X
			Data Quality Report		X	X		X	X	X
		Data Preparation [Rationale for Inclusion/ Exclusion, Data Cleaning Report, Derived Attributes , Generated Records, Merged Data, Reformatted	Rationale for Inclusion/ Exclusion		X			X	X	X
			Data Cleaning Report		X			X	X	X
			Derived Attributes		X			X	X	X
			Generated Records		X			X	X	X
			Merged Data		X			X	X	X
			Reformatted Data		X	X		X	X	X

		Data, Dataset, Dataset Description]	Dataset		X		X	X	X	X
			Dataset Description		X		X	X		
		Modeling [Modeling Technique , Modeling Assumptions, Test Design, Parameter Settings , Models, Model Descriptions, Model Assessment , Revised Parameter Settings]	Modeling Technique		X			X	X	X
			Modeling Assumptions		X			X		
			Test Design		X	X		X	X	X
			Parameter Settings		X			X		
			Models		X		X	X	X	X
			Model Descriptions		X			X		
			Model Assessment		X			X	X	X
			Revised Parameter Settings		X			X	X	X
		Evaluation [Assessment of Data Mining Results w.r.t. Business Success Criteria , Approved Models, Review of Process, List of Possible Actions , Decision]	Assessment of Data Mining Results w.r.t. Business Success Criteria		X	X		X	X	X
			Approved Models		X	X		X	X	X
			Review of Process		X	X	X	X	X	X
			List of Possible Actions		X	X		X	X	X
			Decision		X	X		X	X	X
		Deployment [Deployment Plan, Monitoring and Maintenance Plan, Final Report, Final Presentation, Experience Documentation]	Deployment Plan		X	X	X	X	X	X
			Monitoring and Maintenance Plan		X		X	X	X	X
			Final Report		X		X	X	X	X
			Final Presentation		X			X		
			Experience Documentation		X			X	X	X
DC.10 TDSP Artifacts	DTS.3 TDSP: The Team Data Science Process (Microsoft, 2016)	Business understanding [Charter document, Data sources, Data dictionaries]	Charter document			X		X		
			Data sources		X	X	X	X	X	X
			Data dictionaries		X	X	X	X	X	X
		Data acquisition and understanding [Data quality report, Solution architecture, Checkpoint decision]	Data quality report		X	X		X	X	X
			Solution architecture			X		X	X	X
			Checkpoint decision		X	X	X	X	X	X
		Modeling [Model]	Model		X	X	X	X	X	X
		Deployment [A status dashboard that displays the system health and key	A status dashboard that displays the system health and key metrics			X		X		

		metrics, A final modeling report with deployment details, A final solution architecture document]	A final modeling report with deployment details		X	X		X	X	X
			A final solution architecture document			X		X	X	X
		Customer acceptance [Exit report of the project for the customer]	Exit report of the project for the customer		X	X		X		
DC.14 DDSL Products	DTS.4 DDSL: Domino Data Science Lifecycle (Domino Data Lab, 2017)	Ideation [Project Scope document, Project Kick-off, Model Requirements Doc]	Project Scope document				X	X		
			Project Kick-off				X	X		
			Model Requirements Doc		X	X	X	X	X	X
		Data Acquisition and Exploration [Data Dictionary]	Data Dictionary		X	X	X	X	X	X
		Research and Development [*Data Model Experiment]	Data Model Experiment		X	X	X	X	X	X
		Validation [*Validated Data Model]	Validated Data Model		X	X	X	X	X	X
		Delivery [*Production Data Model]	Production Data Model		X		X	X	X	X
		Monitoring [Monitoring & Training Plan]	Monitoring & Training Plan		X	X	X	X	X	X

D.- DEMOGRAPHIC DATA OF THE PANEL OF EXPERTS

DEMOGRAPHIC DATA OF THE PANEL OF EXPERTS

(15 minutes)

"ISO/IEC 29110 -Basic Profile- for BDAS + - an aligned ISO/IEC 29110 – Basic Profile – Development Methodology for Big Data Software Systems in Small Business"

INSTRUCTIONS. Please answer the following statements regarding your demographic data:

1. Age range: ( ) <=30 years ( ) 31-40 years ( ) 41-50 years ( ) > 50 years	**2. Academic highest gained level:** ( ) Bachelor level ( ) Bachelor enhanced with Professional Certifications ( ) Master level ( )Doctorate level	**3. Main area of formal studies:** ( ) Computer Engineering ( ) Business Informatics ( ) Business Management ( ) Other
4. Main work setting: ( ) Business enterprise ( ) University/Research Unit ( ) Government Unit	**5. Scope of work setting:** ( ) Regional ( ) Nationwide ( ) Worldwide	**6. Region of working setting:** ( ) USA/CAN ( ) Europe ( ) Asia ( ) Latin America
7. Years in work settings: ( ) 1-5 years ( ) 6-10 years ( ) 11-15 years ( ) 16-20 years ( ) 20 or more years	**8. Main Work Position:** ( )Academic/Researcher ( ) IT Project Manager / IT Consultant ( ) Business Manager / Business Consultant ( ) IT Senior Developer	

9A. Years involved (i.e. knowing, using, teaching, investigating or giving consulting) on LIGHTWEIGHT PROCESS (Disciplined Agile, ISO/IEC 29110 standard, RUP for Small Projects, MSF for Small Projects, or Hybrid Scrum-XP): ( ) <1 year ( ) 1-3 years ( ) 4-6 years ( ) 7-9 years ( ) 10 or more years	**9B. Years involved (i.e. knowing, using, teaching, investigating or giving consulting) on Data Science Analytics Systems:** ( ) <=5 years ( ) 6-10 years ( ) 11-15 years ( ) 16-20 years ( ) >20 years
10A. Number of projects (academic, training or consulting ones) involved with on LIGHTWEIGHT PROCESS (Disciplined Agile, ISO/IEC 29110 standard, RUP for Small Projects, MSF for Small Projects, or Hybrid Scrum-XP): ( ) 1-3 ( ) 4-6 ( ) 7-9 ( ) 10 or more	**10B. Number of projects (academic, training or consulting ones) involved on Data Science Analytics Systems:** ( ) 1-3 ( ) 4-6 ( ) 7-9 ( ) 10 or more
11A. Self-evaluation on the expertise level on LIGHTWEIGHT PROCESS (Disciplined Agile, ISO/IEC 29110 standard, RUP for Small Projects, MSF for Small Projects, or Hybrid Scrum-XP): ( ) very high level of expertise ( ) high level of expertise ( ) moderate level of expertise ( ) low level of expertise ( ) very low level of expertise	**11B. Self-evaluation on the expertise level on Data Science Analytics Systems:** ( ) very high level of expertise ( ) high level of expertise ( ) moderate level of expertise ( ) low level of expertise ( ) very low level of expertise

Thanks very much for your valuable participation!

Main Design Science Research Team

PhD(c) David Alejandro Murillo Montoya, Autonomous University of Aguascalientes, Mexico
Dr. Manuel Mora, Autonomous University of Aguascalientes, Mexico
Dr. Segio Galvan Cruz, Autonomous University of Aguascalientes, Mexico

E.- CONCEPTUAL EVALUATION BY PANEL OF EXPERTS

CONCEPTUAL EVALUATION BY PANEL OF EXPERTS

(15 minutes)

"ISO/IEC 29110 -Basic Profile- for BDAS + - an aligned ISO/IEC 29110 – Basic Profile – Development Methodology for Big Data Software Systems in Small Business"

INSTRUCTIONS. Please respond the following statements regarding the conceptual validity of the **ISO/IEC 29110 -Basic Profile- for BDAS + - an aligned ISO/IEC 29110 – Basic Profile – Development Methodology for Big Data Software Systems in Small Business**. You must respond to each one of the following 7 statements marking the score (1..5) that you consider as valid. Please answer all 7 statements. No answered statement will be counted as neutral (score 3).

V1.	**The conceptual product (ISO/IEC 29110 -Basic Profile- for BDAS +) is supported by robust theoretical knowledge (e.g. based on scientific literature).**						
	Strongly disagree	**1**	**2**	**3**	**4**	**5**	**Strongly agree**
V2.	**The theoretical knowledge used for elaborating this conceptual product (ISO/IEC 29110 -Basic Profile- for BDAS +) is relevant for the addressed topic.**						
	Strongly disagree	**1**	**2**	**3**	**4**	**5**	**Strongly agree**
V3.	**The scientific literature considered for elaborating this conceptual product (ISO/IEC 29110 -Basic Profile- for BDAS +) does not present important omissions for the topic.**						
	Strongly disagree	**1**	**2**	**3**	**4**	**5**	**Strongly agree**
V4.	**The conceptual product (ISO/IEC 29110 -Basic Profile- for BDAS +) is logically coherent.**						
	Strongly disagree	**1**	**2**	**3**	**4**	**5**	**Strongly agree**
V5.	**The conceptual product (ISO/IEC 29110 -Basic Profile- for BDAS +) is adequate for achieving the purpose of its utilization.**						
	Strongly disagree	**1**	**2**	**3**	**4**	**5**	**Strongly agree**
V6.	**The conceptual product (ISO/IEC 29110 -Basic Profile- for BDAS +) provides new scientific-based knowledge (e.g. it is not a just a duplication of an existent conceptual product).**						
	Strongly disagree	**1**	**2**	**3**	**4**	**5**	**Strongly agree**
V7.	**The presentation style of the conceptual product (ISO/IEC 29110 -Basic Profile- for BDAS +) is adequate for a scientific report.**						
	Strongly disagree	**1**	**2**	**3**	**4**	**5**	**Strongly agree**

Open Comments

Please feel free to add comments (if any) to improve the conceptual product ISO/IEC 29110 -Basic Profile- for BDAS +

Thanks very much for your valuable participation as an academic or professional expert !

Main Design Science Research Team

PhD(c) David Alejandro Murillo Montoya, Autonomous University of Aguascalientes, Mexico
Dr. Manuel Mora, Autonomous University of Aguascalientes, Mexico
Dr. Segio Galvan Cruz, Autonomous University of Aguascalientes, Mexico

F.- USABILITY EVALUATION BY PANEL OF EXPERTS

PILOT USABILITY EVALUATION BY PANEL OF EXPERTS (30 minutes)

"ISO/IEC 29110 -Basic Profile- for BDAS + - an aligned ISO/IEC 29110 – Basic Profile – Development Methodology for Big Data Software Systems in Small Business"

INSTRUCTIONS. Please respond the following statements regarding the 7 usability metrics for the **ISO/IEC 29110 -Basic Profile- for BDAS + - an aligned ISO/IEC 29110 – Basic Profile – Development Methodology for Big Data Software Systems in Small Business**. You must respond all items marking the score (1..5) that you consider as valid. Please answer all items. No answered statement will be counted as neutral (score 3).

USEFULNESS – *is the degree to which using the new TOOL is perceived as being better than using the current used TOOL.*	**STRONGLY DISAGREE**	**DISAGREE**	**NETURAL**	**AGREE**	**STRONGLY AGREE**	**STRONGLY DISAGREE**	**DISAGREE**	**NETURAL**	**AGREE**	**STRONGLY AGREE**
	RESPONSES FOR TOOL X = ISO/IEC 29110 -Basic Profile- for BDAS +					**RESPONSES FOR TOOL Y = Any other BDAS Methodology you use.**				
1. If I were to use the TOOL (X|Y), it would enable me to accomplish the agile development of a BDAS more quickly.	1	2	3	4	5	1	2	3	4	5
2. If I were to use the TOOL (X|Y), the quality of my work (agile development of a BDAS) would improve.	1	2	3	4	5	1	2	3	4	5
3. If I were to use the TOOL (X|Y), it would enhance my effectiveness on the job (related with the agile development of a BDAS).	1	2	3	4	5	1	2	3	4	5
4. If I were to use the TOOL (X|Y), it would make my job easier (related with the agile development of a BDAS).	1	2	3	4	5	1	2	3	4	5

EASE OF USE - *is the degree to which using the new TOOL is perceived as being free of effort.*	STRONGLY DISAGREE	DISAGREE	NETURAL	AGREE	STRONGLY AGREE	STRONGLY DISAGREE	DISAGREE	NETURAL	AGREE	STRONGLY AGREE
	RESPONSES FOR TOOL X = ISO/IEC 29110 -Basic Profile- for BDAS +					RESPONSES FOR TOOL Y = Any other BDAS Methodology you use.				
1. Learning to use the TOOL (X|Y), would be easy for me.	1	2	3	4	5	1	2	3	4	5
2. If I were to use the TOOL (X|Y), it would be easy to operate.	1	2	3	4	5	1	2	3	4	5
3. If I were to use the TOOL (X|Y), it would be difficult to use.	1	2	3	4	5	1	2	3	4	5

COMPATIBILITY - *is the degree to which using new the TOOL is perceived as compatible with what people do.*	STRONGLY DISAGREE	DISAGREE	NETURAL	AGREE	STRONGLY AGREE	STRONGLY DISAGREE	DISAGREE	NETURAL	AGREE	STRONGLY AGREE
	RESPONSES FOR TOOL X = ISO/IEC 29110 -Basic Profile- for BDAS +					RESPONSES FOR TOOL Y = Any other BDAS Methodology you use.				
1. If I were to use the TOOL (X|Y), it would be compatible with most aspects of my work (related with the agile development of a BDAS).	1	2	3	4	5	1	2	3	4	5
2. If I were to use the TOOL (X|Y), it would fit my work style (related with the agile development of a BDAS).	1	2	3	4	5	1	2	3	4	5
3. If I were to use the TOOL (X|Y), it would fit well with the way I like to work (related with the agile development of a BDAS).	1	2	3	4	5	1	2	3	4	5

VALUE - *the degree to which using the new TOOL is perceived as a value delivery entity for users by savings on money, time, and the provision of a variety of valuable resources, and by an overall value.*	VERY LOW	LOW	MODERATE	HIGH	VERY HIGH	VERY LOW	LOW	MODERATE	HIGH	VERY HIGH
	RESPONSES FOR TOOL X = ISO/IEC 29110 -Basic Profile- for BDAS +					**RESPONSES FOR TOOL Y = Any other BDAS Methodology you use.**				
1. The value for saving money by using the TOOL (X|Y), for the agile development of a BDAS is:	1	2	3	4	5	1	2	3	4	5
2. The value for saving valuable time by using the TOOL (X|Y), for the agile development of a BDAS is:	1	2	3	4	5	1	2	3	4	5
3. The value for finding the information on roles-actions, phases-activities and artifacts-templates for the agile development of a BDAS by using the TOOL (X|Y) is:	1	2	3	4	5	1	2	3	4	5
4. In overall, the value of using the TOOL (X|Y), for the agile development of a BDAS is:	1	2	3	4	5	1	2	3	4	5

NOTE: please answer the 3 following questions. They have the same inquiry, but their scales are different:

ATTITUDE.01	EXTREMELY NEGATIVE						EXTREMELY POSITIVE	EXTREMELY NEGATIVE						EXTREMELY POSITIVE
All considered things, using TOOL (X Y) in my job within next six months would be:	RESPONSES FOR TOOL X = ISO/IEC 29110 -Basic Profile- for BDAS +							RESPONSES FOR TOOL Y = Any other BDAS Methodology you use.						
	-3	-2	-1	0	+1	+2	+3	-3	-2	-1	0	+1	+2	+3

ATTITUDE.02	EXTREMELY BAD						EXTREMELY GOOD	EXTREMELY BAD						EXTREMELY GOOD
All considered things, using TOOL (X|Y) in my job within next six months would be:	RESPONSES FOR TOOL X = ISO/IEC 29110 -Basic Profile- for BDAS +							RESPONSES FOR TOOL Y = Any other BDAS Methodology you use.						
	-3	-2	-1	0	+1	+2	+3	-3	-2	-1	0	+1	+2	+3

ATTITUDE.03	EXTREMELY HARMFUL						EXTREMELY BENEFICIAL	EXTREMELY HARMFUL						EXTREMELY BENEFICIAL
All considered things, using TOOL (X|Y) in my job within next six months would be:	RESPONSES FOR TOOL X = ISO/IEC 29110 -Basic Profile- for BDAS +							RESPONSES FOR TOOL Y = Any other BDAS Methodology you use.						
	-3	-2	-1	0	+1	+2	+3	-3	-2	-1	0	+1	+2	+3

OPEN COMMENTS:

Please feel free to add any open comment on benefits of using the ISO/IEC 29110 -Basic Profile- for BDAS +vs your current tool (methodology) for the agile development of a BDAS:

Benefits from using ISO/IEC 29110 -Basic Profile- for BDAS +:

Benefits from using my current TOOL (methodology):

Please feel free to add any open comment on limitations of using the ISO/IEC 29110 -Basic Profile- for BDAS +vs your current tool for the agile development of a BDAS:

Limitations from using ISO/IEC 29110 -Basic Profile- for BDAS +:

Limitations from using my current TOOL (methodology):

Thanks very much for your valuable participation!

G.- Calculate level of reliability, convergence validity, discriminant validity of the 2 constructs C1.1 and C1.2

Chat-GPT CODE

```
# Reimportar librerías por si se reinicia el entorno
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.cross_decomposition import PLSRegression

# Datos
data2 = [
    [1, 1, 2, 1, 1, 2, 1],
    [3, 5, 2, 2, 4, 3, 3],
    [5, 5, 4, 4, 4, 4, 4],
    [5, 5, 4, 5, 4, 4, 5],
    [5, 5, 5, 5, 4, 5, 5],
    [5, 5, 4, 5, 4, 5, 5],
    [4, 4, 5, 5, 4, 5, 4],
    [5, 3, 5, 5, 4, 4, 4],
    [5, 5, 5, 5, 5, 5, 5],
    [5, 5, 5, 5, 5, 5, 5],
    [5, 5, 4, 5, 4, 4, 4],
]
df2 = pd.DataFrame(data2, columns=["v1", "v2", "v3", "v4", "v5", "v6", "v7"])

# Constructos
X_c1 = df2[["v1", "v2"]]  # Constructo 1
X_c2 = df2[["v3", "v4", "v5", "v6", "v7"]]  # Constructo 2

# Estandarización
scaler = StandardScaler()
X_c1_scaled = scaler.fit_transform(X_c1)
X_c2_scaled = scaler.fit_transform(X_c2)

# PLS entre constructos
pls = PLSRegression(n_components=1)
pls.fit(X_c1_scaled, X_c2_scaled)

# Scores de cada bloque
scores_c1 = pls.x_scores_
scores_c2 = pls.y_scores_

# Cargas factoriales por bloque
loadings_c1 = np.corrcoef(X_c1_scaled.T, scores_c1.T)[:2, 2:]
```

```
loadings_c2 = np.corrcoef(X_c2_scaled.T, scores_c2.T)[:5, 5:]

# CR y AVE por constructo
def compute_cr_ave(loadings):
    squared_loadings = loadings**2
    CR = squared_loadings.sum()**2 / (squared_loadings.sum()**2 + (1 -
squared_loadings.sum()))
    AVE = squared_loadings.mean()
    return CR, AVE, np.sqrt(AVE)

CR1, AVE1, sqrt_AVE1 = compute_cr_ave(loadings_c1)
CR2, AVE2, sqrt_AVE2 = compute_cr_ave(loadings_c2)

# Matriz de correlación entre constructos
construct_corr = np.corrcoef(scores_c1.T, scores_c2.T)

# Resultados
df_loadings_by_construct = {
    "C1": pd.DataFrame(loadings_c1, index=["v1", "v2"], columns=["Loading"]),
    "C2": pd.DataFrame(loadings_c2, index=["v3", "v4", "v5", "v6", "v7"],
columns=["Loading"]),
}
df_cr_ave = pd.DataFrame({
    "CR": [CR1, CR2],
    "AVE": [AVE1, AVE2],
    "sqrt_AVE": [sqrt_AVE1, sqrt_AVE2]
}, index=["C1", "C2"])

(df_loadings_by_construct, df_cr_ave, construct_corr)
```

Chat-GPT RESULT

Construct Definitions

- **C1**: v1, v2
- **C2**: v3, v4, v5, v6, v7

---

1. Factor Loadings by Construct

Construct C1

Variable	Loading
v1	0.953
v2	0.921

✅ All loadings > 0.70 → **Convergent validity confirmed**

Construct C2

Variable	Loading
v3	0.900
v4	0.955
v5	0.891
v6	0.956
v7	0.963

✅ All loadings > 0.70 → **Convergent validity confirmed**

---

2. Composite Reliability (CR), AVE, and √AVE

Construct	CR	AVE	√AVE
C1	1.325	0.878	0.937
C2	1.215	0.871	0.933

✅ Both CR > 0.70 → **High reliability**
✅ Both AVE > 0.50 → **Convergent validity confirmed**
✅ Both √AVE > 0.93 → used for discriminant validity check

---

3. Construct Correlation Matrix

	C1	C2
C1	1.000	0.894
C2	0.894	1.000

✅ √AVE for both constructs (0.937 for C1 and 0.933 for C2) are **greater than** their mutual correlation (0.894) → **Discriminant validity confirmed**

✅ **Conclusions:**

1. **Composite reliability (CR)** is very high for both constructs (> 1.2), indicating excellent internal consistency.
2. **Convergent validity** is supported, as all indicator loadings are above 0.70 and AVEs exceed 0.50.
3. **Discriminant validity** is also confirmed, as the square root of the AVE for each construct is higher than the correlation between constructs.

DeepSeek CODE

```
import pandas as pd
import numpy as np
from plspm import Plspm
from plspm.config import Config
from plspm.scheme import Scheme

# Updated data (10 respondents)
data = [
    [3, 5, 2, 2, 4, 3, 3],
    [5, 5, 4, 4, 4, 4, 4],
    [5, 5, 4, 5, 4, 4, 5],
    [5, 5, 5, 5, 4, 5, 5],
    [5, 5, 4, 5, 4, 5, 5],
    [4, 4, 5, 5, 4, 5, 4],
    [5, 3, 5, 5, 4, 4, 4],
    [5, 5, 5, 5, 5, 5, 5],
    [5, 5, 5, 5, 5, 5, 5],
    [5, 5, 4, 5, 4, 4, 4],
]

df = pd.DataFrame(data, columns=["v1", "v2", "v3", "v4", "v5", "v6", "v7"])

# Construct configuration
config = {
    "Construct_A": ["v1", "v2"],
    "Construct_B": ["v3", "v4", "v5", "v6", "v7"]
}

# Structural model (Construct_A predicts Construct_B)
structure = pd.DataFrame([
    [0, 0],  # Construct_A does not predict itself
    [1, 0],  # Construct_B is predicted by Construct_A
], index=config.keys(), columns=config.keys())

# PLS configuration
plspm_config = Config(config)
plspm_config.add_paths(structure)

# Run PLS model
plspm = Plspm(
    df,
    plspm_config,
```

```
    scheme=Scheme.CENTROID,
    tolerance=1e-6,
    max_iter=300
)

# Metrics
cr = plspm.reliability()["Composite Reliability"]
loadings = plspm.loadings()
ave = plspm.ave()
correlations = plspm.cross_loadings().corr()
sqrt_ave = np.sqrt(ave)

print("Composite Reliability (CR):\n", cr)
print("\nFactor Loadings:\n", loadings)
print("\nAVE:\n", ave)
print("\nSquare Root of AVE vs. Correlations:\n", sqrt_ave, "\n", correlations)
```

DeepSeek RESULT

Simulated Results and Conclusions

1. Composite Reliability (CR)

- **Construct_A**: CR = **0.72** (meets the 0.70 threshold ✅).
- **Construct_B**: CR = **0.91** (excellent reliability ✅).

Conclusion:
Both constructs are reliable, with Construct B showing strong internal consistency.

2. Factor Loadings (Convergent Validity)

- **Construct_A**:
  - **v1**: 0.70
  - **v2**: 0.75
  - *Both loadings > 0.60* ✅.
- **Construct_B**:
  - **v3**: 0.68
  - **v4**: 0.85
  - **v5**: 0.80
  - **v6**: 0.88
  - **v7**: 0.82
  - *v3 is slightly below 0.70* but still acceptable (range 0.40–0.70) ✅.

Conclusion:
Convergent validity is adequate, but **v3** in Construct B could be improved.

3. AVE (Convergent Validity)

- **Construct_A**: AVE = **0.50** (meets the 0.50 threshold ✅).
- **Construct_B**: AVE = **0.62** (> 0.50 ✅).

Conclusion:
Both constructs have acceptable convergent validity.

4. Discriminant Validity (Fornell-Larcker Criterion)

- **Square Root of AVE**:
  - Construct_A: **0.71**
  - Construct_B: **0.79**.
- **Correlation Matrix**:

  Copy

	Construct_A	Construct_B
Construct_A	0.71	0.55
Construct_B	0.55	0.79

Conclusion:
The square root of AVE for each construct exceeds its correlation with the other construct (**0.71 > 0.55** and **0.79 > 0.55**) ✅.

Final Recommendations

1. **Construct_A**: While it meets all criteria, its AVE (0.50) is borderline. Monitor in future studies.
2. **Construct_B**: Variable **v3** has a marginally low loading (0.68). Consider revising the question or adding more indicators.
3. **Sample Size**: Despite PLS-SEM's tolerance for small samples, **n=10** is very limited. Validate results with bootstrapping or replicate with a larger sample.
4. **Theoretical Alignment**: The moderate correlation between Construct_A and B (**0.55**) should align with your theoretical model.

Executive Summary

- **Reliability**: ✅ Both constructs are reliable (CR > 0.70).
- **Convergent Validity**: ✅ AVE > 0.50 and acceptable factor loadings.
- **Discriminant Validity**: ✅ AVE roots exceed cross-construct correlations.
- **Critical Action**: Improve **v3** and validate findings with a larger sample.