Business analytics and data science in business strategy

Case Nokia & Alteryx

Tomi Sarpola

Opinnäytetyö
Tietojenkäsittelyn koulutusohjelma
2020

Tiivistelmä

Author Tomi Sarpola	
Department Information Technology Program	
Title Business analytics and data science in business strategy Case Nokia & Alteryx	**Pages and appendixes** 93 + 6

This thesis work explains methods available for applying data science in business, identifying enablers, and evaluating data science development success factors in organizations.

The theoretical framework gives an overview of data science and related standards, summarizes examples of successful AI adoption in different business areas, and then explains data strategy and organization analytics maturity concepts. The maturity of AI development in the industry has come to the point where literature examples are easily available on success factors for setting up data science teams, technology platforms, and selecting business problems for implementing successful data science solutions.

Empirical work in the thesis consists of two parts. The first part was to study the analytics maturity of case organization in Nokia, and to propose guidelines for improving and focusing on analytics development processes. The second part focused on improving organization data science capabilities by producing a prototyping environment and architecture for scaling the prototypes to operation using Alteryx software. This approach was applied in a project work to produce a predictive algorithm in the product field failure management area.

Thesis work and project deliverables were taken into operative use by the customer organization and further activities planned to improve analytics capabilities and strategy.

Keywords
machine learning, data strategy, predictive analytics, Alteryx, Azure, Python

Tiivistelmä

Haaga-Helia
ammattikorkeakoulu Oy

Tekijä(t)
Tomi Sarpola

Koulutusohjelma
Tietojenkäsittely

Raportin/Opinnäytetyön nimi Business analytics and data science in business strategy Case Nokia & Alteryx	**Sivu- ja liitesivumäärä** 93 + 6

Lopputyö on kirjoitettu englanniksi asiakkaan vaatimuksesta hyödyntää raporttia kansainvälisen yrityksen henkilöstön analytiikkaosaamisen kehittämiseksi.

Tässä opinnäytetyössä käyn läpi menetelmiä datatieteen periaatteiden ja analytiikan hyödyntämiseksi liiketoiminnassa. Työssä tutkitaan onnistuneiden analytiikkahankkeiden arvioinnin menetelmiä, menestyksen tekijöitä ja menetelmiä organisaatioiden analytiikkavalmiuksien kehittämiseksi. Teoreettinen viitekehys antaa yleiskuvan datatieteestä ja siihen liittyvistä käytännöistä sekä standardeista. Teoria antaa tiivistettynä esityksenä esimerkkejä onnistuneista tekoälyhankkeista eri liiketoiminta-alueilla, käy läpi käsitteet datastrategian käytäntöihin liittyen, sekä ja organisaation analytiikkavalmiuksien kypsyyden määrittämiseksi. Tekoälyhankkeiden kehityksessä on tultu pisteeseen, jossa kirjallisuusesimerkkejä on helposti saatavilla. Näistä on tunnistettu menestystekijöitä ja ohjeita toimivien datatiedetiimien kehittämiseksi, analytiikan teknologia-alustojen valintaan, sekä liiketoiminnan analytiikkahankkeiden määrittelyyn.

Opinnäytetyön empiirinen työ koostuu kahdesta osasta. Ensimmäinen osa tutkii Nokia Oyj:n asiakasorganisaation analytiikan kypsyyttä. Tutkimuksen pohjalta ehdotetaan parannuksia analytiikan sovelluskehityksen hallintaan. Toinen osa keskittyy tietotieteiden osaamisen parantamiseen organisaatiossa käytännönprojektin avulla. Projekti tuotti arkkitehtuurin ja alustan koneoppimisen algoritmien prototyyppien ja tuotantomallien kehittämiseen Alteryx ohjelmistoa hyödyntäen. Projektissa toteutettiin myös ennustavan algoritmin tuotevikojen hallinnan alueelle operatiiviseen käyttöön.

Ohjeet ja projektituotokset otettiin operatiiviseen käyttöön, ja jatkotoimet suunniteltiin analyysikyvyn ja strategian parantamiseksi.

Asiasanat
koneoppiminen, tietostrategia, ennustava analytiikka, Alteryx, Azure, Python

Abbreviations & terminology

AI	Artificial intelligence
Alteryx	Analytics platform aiming at self-service analytics
APIs	Application programming interfaces
Azure	Cloud computing service developed by Microsoft
BI modal IT	Framework enabling two development modes for combining business and IT expertise in different scale
Cadence of analysis	Culture in DevOps and ML Ops to agree rhythm for executing set of practices e.g. analytics activities
CI/CD	Continuous integration and continuous development practices
CRISP-DM	Cross-industry standard process for data mining
Crowd-sourcing	Sourcing services or information from groups of participants
Data Ops	Practices of managing operative data pipelines
DevOps	IT management framework for combining software development and IT operations
Data analytics maturity	Framework for assessing organizations ability to produce analytic solutions and to utilize analytics in decision making
Data dissemination	Phase in analytical processes where datasets or data summaries are distributed to end users
ERP	Enterprise Resource Planning system
GIT	Version-version-control system
IoT	Internet of things
KDD	Knowledge Discovery and Data Mining
ML	Machine learning
ML Ops	Practices of managing ML algorithms and ML service
PLM	Product Lifecycle Management system
REST	Representational state transfer
Stage	Data repository for collecting uncleaned data
SQL	Structured Query Language

Table of contents

1. Introduction

Objectives, information on customer, thesis methods and concepts are explained in the introductory chapters.

1.1 Thesis introduction, objectives, concepts, and methods

The objective of this thesis work is to produce an overview of the latest development of industry standards, processes, and technologies in areas of business and data strategy development, business analytics, and automated predictive systems using machine learning. As an outcome, the objective is to create practical guidelines for a customer organization for applying data strategies for business development, in applying technological innovations for developing a machine learning-enabled IT architecture for iterative business analytics, and guide for building a team culture among business owners, stakeholders, business analysts and technical experts to evolve the organization towards a data science enabled data-driven organization.

For the past years, there has been a high demand for increased efficiency in applying business management principles and data-driven decision making, creating automation to reduce manual labour from data-intensive tasks, and taking the latest trends from the data science boom to operative use. Although there have been many use cases and presentations of successful machine learning and artificial intelligence projects, very few of such projects succeed to meet the planned business targets, timelines, and cost objectives. (Aiken and Harbour, 2017, p. 9-29)

Amount of available data increases at the moment at such velocity that organizations have difficulty keeping up with it. New data is introduced when setting up new services. However, the vast majority of organizations are dependent on the legacy data they currently operate on. And it is a real challenge to keep track of what data organizations have at which quality and who is using them. To create benefits of the collected data, the business needs a data strategy to ensure the data processes and analytics projects can be consistent, repeatable, and focused on delivering the business objectives. (Aiken and Harbour, 2017, p. 9-29)

To understand how to take machine learning-enabled systems into operative use in business, the thesis produces first a literature overview of what analytics means to business and strategy development, what success factors and challenges are generally identified in organization structures, processes, and teams developing analytics systems, and what technologies are involved in big data operation.

The driving thesis questions for the study are: How are business analytics and data science related to business strategy? How does data strategy help businesses? What makes a data science project successful, e.g., what are the requirements to deploy predictive models to operative use in business processes? What are the elements which make a business problem suitable for data science? And regarding our use case, how do the different analytics platforms (code-free, code-based) from evaluated vendors (Microsoft, Alteryx, etc.) enable efficient and iterative business data analytics processes?

The theoretical component in the thesis consists of a literature study on data strategy development, data quality principles, business analytics, and data science processes and methods. The thesis aims to study team development practices in analytics and data science areas, identify key roles in data science projects, standards used in data technologies handling big data, and best practices on data science productization. A technology review is focused on comparing analytics platforms of the following vendors: Microsoft, Alteryx, Rapid Miner. Additionally, technologies and architecture design principles are studied to produce big data capable data processing for an operationalized analytics system, which uses machine learning models to predict future process events and creates alerts and automated dashboard reports for end-users.

The empirical part comprises of two parts. Firstly, the thesis will evaluate the customer organization's current analytics capabilities and experiences in a maturity assessment. The assessment conclusions are used to develop guidelines and identify key success criteria and best practices for business analytics strategy development and ensure data science projects success. The project will focus on business strategy development and analytics maturity assessment evaluation on a high level and will utilize globally available standards and methods as far as possible. Analytics maturity assessment will be executed as a survey among stakeholders and summary discussion to conclude the current state in data science evolution.

The second part of the empirical part focuses on providing a roadmap proposal for analytics strategy and IT architecture development in Nokia. This architecture should enable iterative data strategy execution, development of business analytics projects, and evolution to data science development. The organization will produce key architectural components in the Nokia MN FiRe Program (Phases 1 and 2). Work packages in the program include team competence development, key concepts development, architecture development, building understanding on business need, development of analysis hypothesis, and data science model development. Key deliverables are a high-level architecture description, user scenarios description, technology demonstrations, data automation pipeline in cloud environment, and a proposal for of architecture development.

As an outcome of the thesis, objective is to produce recommendations for business development on building data strategies and applying automation support for decision making. Thesis will also create a roadmap proposal for a machine learning-enabled IT architecture for use in iterative business analytics development. Finally, thesis will collect practical guidelines for an organization for building a data-driven team culture, tools for evaluating business analytics opportunities, and steps to evolve the organization to utilize data science capabilities.

1.2 Thesis customer and research motivation

1.2.1 Nokia

Nokia is a Finnish telecommunications company operating worldwide, whose main businesses are network infrastructure, network management software, and services and technology development and licensing. In 2019, Nokia reported net sales of EUR 23.3 billion, of which profit was EUR 1.23 billion. In 2017, Nokia was Finland's largest company in terms of net sales (Kujansuu, 2018), And in 2018, the second largest after Neste Corporation (Suomen Asiakastieto Oy, 2020).

In recent history, Nokia has become best known as one of the world's largest mobile phone companies. In April 2014, Nokia transferred the mobile business to Microsoft. Nokia's current mission is to be a trusted partner for critical networks, we are committed to innovation and technology leadership across mobile, fixed and cloud networks. We create value with intellectual property and long-term research, led by the award-winning Nokia Bell Labs. (Nokia Oyj, 2020)

At the end of 2019, Nokia employed 98,322 people worldwide. Finland has just under 6,000 employees, of which about 3,000 are in Espoo and Helsinki, about 750 in Tampere, and about 2,250 in Oulu. At the turn of the millennium, there were about 25,000 employees in Finland. Nokia currently operates in more than 100 countries worldwide. With the merger of Alcatel-Lucent in 2016, there were 113,000 employees. The product development staff is 37,000. (Nokia Oyj, 2020)

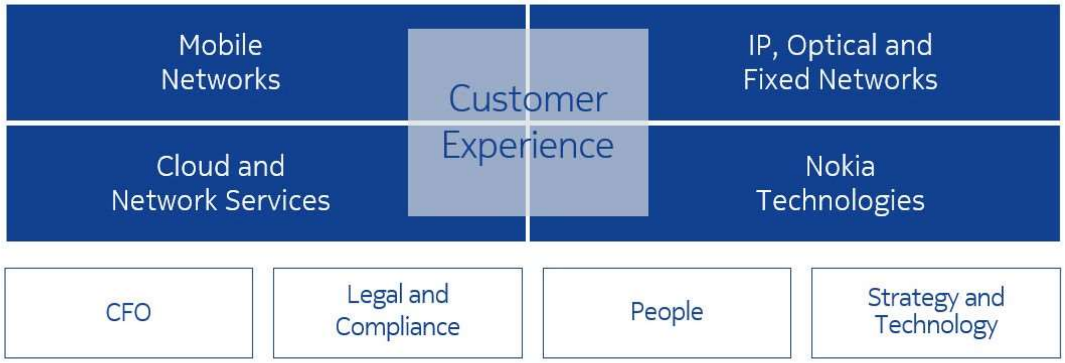


Figure 1: Nokia organization and business groups 2020

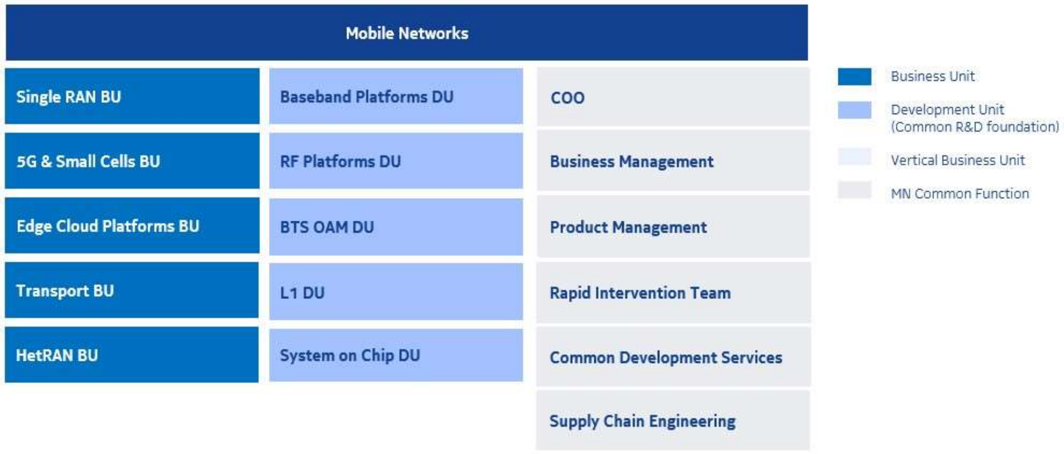
At the beginning of 2020, there were 7 business units in Nokia Corporation, including Mobile Networks (MN), which focuses on telecommunications network services, radio networks, and backbone data centre services, -Fi) Fixed Network (FN), IP Routing and IP / Optical Network (ION) for Optical Network Devices, Global Service (GS) for Network

Entity Solution Design and Maintenance, Nokia Software for Network Service Providers (NSW), Nokia Enterprise, which specializes in digitization and automation solutions for large enterprises, and Nokia Technologies (TECH), which provides technology development and licensing services. At the moment, Nokia has launched an update for the mode of operation to focus the organization structure to 4 business units: Mobile Networks; IP, Optical and Fixed Networks; Cloud and Network Services; and Nokia Technologies. These are supported with common functions of Customer Experience, CFO, Strategy and Technology, Legal and Compliance, and People. (Nokia Oyj, 2020)

MN Supply Chain Engineering

Our organization unit Supply Chain Engineering (SCE), is part of the Mobile Networks (MN) business group. The role of the Supply Chain Engineering unit is to drive the profitability of MN's business portfolio and performance of the supply chain for MN's products at all stages of the product life cycle. SCE accelerates Time-to-Profit and value creation towards our customers' expectations by providing technology development, supply chain planning, and production ramp-up, product information management, product testing, product HW maintenance support and engineering, customer maintenance support, and MN product portfolio and MN supply chain analytics services.

The team is distributed globally to all regions. It works closely with R&D teams, regional factory teams, operations supply chain management teams, finance and control teams, and customer-facing teams to ensure high-quality supply chain and product portfolio management.



Mobile Networks

Single RAN BU	Baseband Platforms DU	COO
5G & Small Cells BU	RF Platforms DU	Business Management
Edge Cloud Platforms BU	BTS OAM DU	Product Management
Transport BU	L1 DU	Rapid Intervention Team
HetRAN BU	System on Chip DU	Common Development Services
		Supply Chain Engineering

Business Unit
Development Unit (Common R&D foundation)
Vertical Business Unit
MN Common Function

Figure 2: MN organization and business units 2020

MN Supply Chain Analytics Team

Our team Supply Chain Analytics (SCA), is part of the Supply Chain Engineering (SCE) unit. We are responsible for the development of a common and agile platform for data analytics in SCE. The objective is to drive and promote Digital Portfolio Management Tools and Processes. We support digitizing, automating, and robotizing the SCE processes. We also operate as a centre of excellence for automation and reporting-related activities for our unit.

We develop and apply Advanced Analytics and AI/ML concepts, strategies, and tools for the whole SCE organization to enable descriptive, deductive, predictive, and preventive operations in E2E Supply Chain. Our goal is to collect and validate essential data from the product development departments and Supply Chain operations and make them available in a harmonized Data Pool as a basis for automation, analytics, and AI/ML development.

We develop and operate MN Supply Digital Twin to provide all SC relevant information on a self-service 24/7 basis. The objective is to enable Analytics Assisted Decision Making in SCE and drive to simulate and predict the MN E2E Supply Chain by applying Analytics and AI/ML.

In my current role in the Supply Chain Analytics (SCA) team, my responsibilities include reporting, analytics, and automation solutions development for various processes and tasks of the SCE unit. Solutions in my responsibility area include project quality management and reporting, continuous improvement (CI) monitoring and reporting, delivery and demand planning reporting and analytics, product quality analytics. Additionally my responsibilities include machine learning (ML) research and solutions development, maintenance operative ML services, automation development, and infrastructure development and maintenance on cloud platforms.

The largest work topic in 2020 for myself is to develop the AI / ML strategy for our unit. Strategy aims to influence supply chain performance, cost-effectiveness, and product reliability and quality. This requires getting more familiar with the broader enterprise information management models, data strategies, and how the capabilities of our unit meet these models and strategies. Objective is to create a data strategy that describes what data is critical to our functions, key processes, and business decisions, and which are intended to be automated and supported by machine learning. This thesis helps consolidate the different aspects and considerations of data-driven decision making into a more a compact overview, which can be discussed within the organization.

1.2.2 MN Field Reliability Analytics Initiative

Our organization in Nokia MN Supply Chain Engineering manages data analytics initiatives to develop processes, methods, and tools to enable predictive product field failure detection in the product maintenance business area. In particular, the project aims to produce analytics and automation solutions for the supply chain of 4G and next generation 5G products to find information that promotes the prediction and reduction of product field failures. The scope and focus of SCE in product maintenance is to impact HW design and product supply chain to improve reliability. We operate as a key stakeholder in all operational process improvements (manufacturing, supply/demand chain, customer repair, production testing) related to MN products.

Previous advanced analytics programs in field reliability area

The MN journey in product maintenance and field reliability analytics started with an external project in collaboration with IBM to analyse failure patterns for products to identify if the field is applicable for advanced analytics.

MN continued the study with Nokia Bell Labs and Nokia Global Services (GS) to boost the Nokia service business. This study had a somewhat different scope, as GS's focus is on developing a preventative maintenance system out of operational system data. Supply and return logistics analytics provide supplementary information for preventative maintenance.

The results received from earlier studies were promising. Still, results were either not fitting to current business processes at the time, or capabilities were not in place to continue building the maturity of the model prototypes to full solutions. Later on, MN decided to invest in a new program to build the internal organization capabilities of product maintenance and field reliability analytics to a degree where the organization can integrate predictive solutions into the business processes.

MN Field Reliability improvement by advanced machine learning (MN FiRe)

The program that our organization launched is called MN Field Reliability improvement by advanced machine learning, MN FiRe for short. The objective was to identify statistically significant variables and correlation patterns indicating Field Failure that can be detected in data early on in the product lifecycle and to create predictions of probable field failure risk patterns using advanced analytics and ML algorithms.

Predictive modelling and prevention of HW failures are expected to be one of the key enablers for successful business operation in MN, helping MN improve customer perceived quality and helping to maintain supply chain performance. Goals set for the program are expected to improve utilizing and combining existing data sources to find new ways of product reliability improvement, improve organization capabilities in machine learning analytics, and use-case development in MN to investigate machine learning and predictive analytics opportunities in product reliability improvements. Predictive modelling is identified as a large opportunity. However, the identified risk for the program was that many strategic projects and process improvements had been stopped before deployment to operative use due to the maturity of analytics deployment within the organization.

Use cases discussed in the planning of the first project iteration in the program included analysis and design of the predictive model to identify:

i) Impact of repair debug variables against all field returns,
ii) Impact of manufacturing test variables against field failure probability and evaluation of failure time frame (1yr, 2yr, +2yr)
iii) Predict field failure risk of the HW product before shipment.

Also, the first project objective was to produce several additional quality improvement use-cases based on the concept as input for the next project iterations.

The first project iteration was a research/innovation project applying a Design Thinking mindset and improving the organization's understanding of the business data processes, data quality, data science processes, machine learning methods, and technologies. First iterations were conducted to analyse the product manufacturing and testing processes data among the data of the product returns. The development team consists of technical product experts, business decision-makers and analysts, one data scientist, and many contributors from both non-IT and IT backgrounds who participate in developing the analytics solution.

The overall setup of the program and different work packages are explained in the attached figure. I will describe these packages in more details in chapter 5, which explains our use case "Application of machine learning enabled IT architecture for business analytics in Nokia MN".

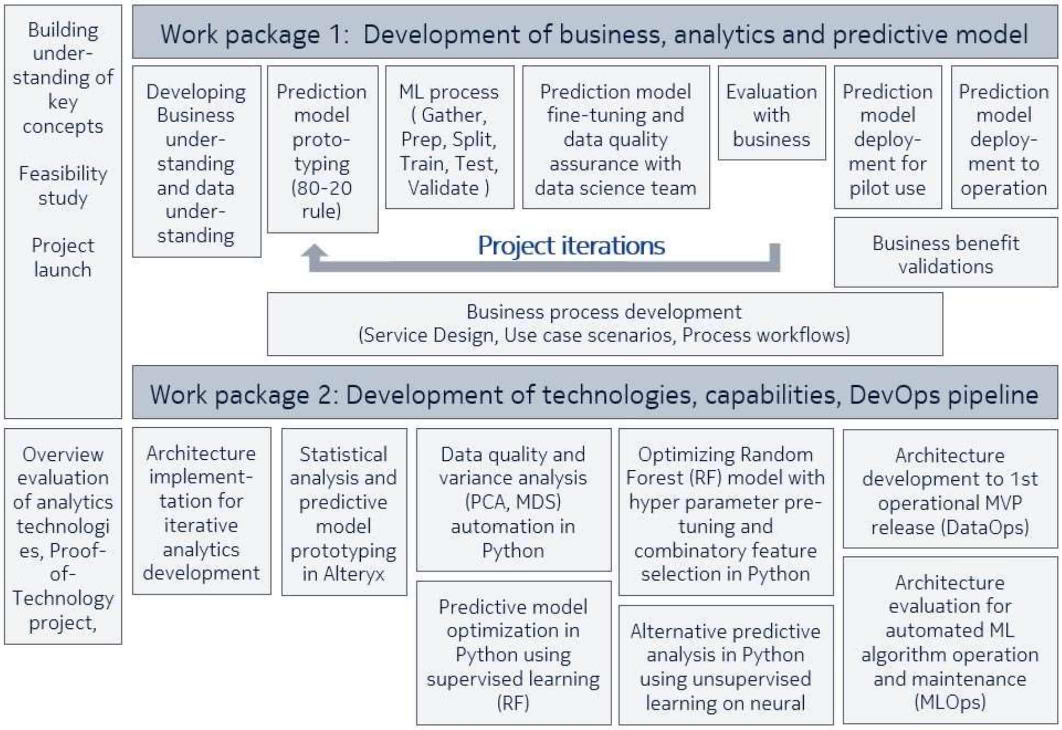

Building under-standing of key concepts
Feasibility study
Project launch
Work package 1: Development of business, analytics and predictive model
Developing Business under-standing and data under-standing
Prediction model proto-typing (80-20 rule)
ML process (Gather, Prep, Split, Train, Test, Validate)
Prediction model fine-tuning and data quality assurance with data science team
Evaluation with business
Prediction model deploy-ment for pilot use
Prediction model deploy-ment to operation
Project iterations
Business benefit validations
Business process development (Service Design, Use case scenarios, Process workflows)
Work package 2: Development of technologies, capabilities, DevOps pipeline
Overview evaluation of analytics technologi es, Proof-of-Technology project,
Architecture implement-tation for iterative analytics development
Statistical analysis and predictive model prototyping in Alteryx
Data quality and variance analysis (PCA, MDS) automation in Python
Predictive model optimization in Python using supervised learning (RF)
Optimizing Random Forest (RF) model with hyper parameter pre-tuning and combinatory feature selection in Python
Alternative predictive analysis in Python using unsupervised learning on neural
Architecture development to 1st operational MVP release (DataOps)
Architecture evaluation for automated ML algorithm operation and maintenance (MLOps)

Figure 3: Nokia MN Field Reliability Project development activities overview

In the first project, my responsibility was to develop an information technology platform to enable the analytics team to process the business data, produce competencies required for data-oriented development, and enable solution development with data scientist teams. I began research with a technology survey on the previous year to understand what tools are available to use ML / AI methods and what practices are technically required to process big data information sets.

The first project chose key technology platforms based on my initial feasibility study and comparison between Microsoft PowerBI, Azure cloud, Alteryx analytics platforms and Python development environments. We acquired and evaluated Alteryx machine learning and analytics software by producing a prototype predictive model in a 6-week development iteration in the next phase. The evaluation consisted of analytics capabilities overview, integration possibilities with Microsoft Azure data systems (Azure SQL, Data lake) and ML services (Databricks, Python), and study on how the organization can utilize ML components in our current production systems (Oracle, MS SQL, SAP BO BI, PowerBI, etc.) operated both in on-premises and cloud environments.

The second project was to analyse additional use cases, focusing on organization KPIs on product returns. KPI selected for the project was the monthly return rate (MRR), an industry-standard quality metric for following the performance of the product maintenance process and product reliability, including the long-term historical perspective of product delivery data and returns. This thesis work was executed as part of the activities for the second project to document the key information about the processes used, the architecture capabilities and to produce proposals for the organization towards advanced analytics strategy development.

We now improved the project organization with an additional focus on business process experts, product technical experts, analytics, and a new collaboration with Nokia internal data science team.

My focus area for the second project was overseeing data collection and data quality assessments between multiple sources, technical documentation for requirements collection, use cases description, service design and IT architecture design, and support in the preparation of business process improvement. As the data science team focused on practical data analysis and prediction algorithm development, I had the opportunity to focus on preparing the IT architecture for operative use, to enable sufficient automation to support both Data OPS and ML OPS, to review architecture guidelines with enterprise IT, cybersecurity, and cloud platform vendors.

The project extended the use of key technologies of Microsoft PowerBI, Azure cloud, Python development environment and Alteryx analytics platforms. During the project, an iterative approach was applied to extend the use of the architecture. Initially, we used manual data export analysis in Alteryx. This was developed to a semi-automated data pipeline and continued with evolution to integrated and fully automated DataOps and MLOps solution.

2. Business analytics and data management implications to business strategy

To understand how analytics relates to business, few fundamental concepts need to be explained. Before organizations can develop overall plans for data analytics, often consolidated to data strategies, it is important to understand the relation of business information and data strategy to a company's business strategy.

2.1 Business analytics and Data Strategy overview

In overview, strategy is the fundamental representation of the company's vision, and the company's high-level decisions to make to be successful. The term originates from military tactics but has been evolved as an integral part of modern business. There are several attempts to explain strategy as a term and as examples. Henry Mintzberg offers a traditional definition for the term as being "the highest-level guidance available to an organization, focusing activities on articulated goal, and provides direction and specific guidance when faced with a stream of decisions or uncertainties (Mintzberg, 1978, p. 1-15).

In business, the purpose of strategy is to turn vision into plans for the company's future. A strategy is good when people know what to do and how, they are motivated to do it, and can measure its progress with regular management reviews. (Stroh, 2014, p. 1-13). A simple and often cited example of strategy in action has been developed by Wayne Gretzky, who is considered the greatest hockey player. Gretzky's mission is to score a goal. And his strategy guidance is to "skate to where the puck will be" and "move away from the person who cuts you off when a pass is done and to intercept the puck at where the puck is going". This strategy is easy to understand and to share with others. (Farber and Illustrated, 2012, p. 1-13)

In modern companies and their strategic planning, data and information are considered valuable and function at the core of the business. E-business companies like Google, Amazon, Spotify, Netflix, among others, collect vast amounts of data from their customers. By analysing it, the company can predict customer behaviour and improve existing or create new services and digital products. A typical example is to analyse, for example, what is being bought, when, where and in which order. (Hämäläinen et al., 2016). These systems are vital to a company's operation and are called Business Information Systems (BIS). A more detailed overview of examples and use of these systems is found in the book 'Business Information Systems' by Bocij et al. (2015).

What is considered as business information?

Data quality is a key focus area when designing and developing business-critical systems and analytics solutions. Often, information management and information quality concepts are represented as a hierarchical structure established by Russell Ackoff called ’DIKW model’ or data pyramid (Bocij et al., 2015, p. 5-11).

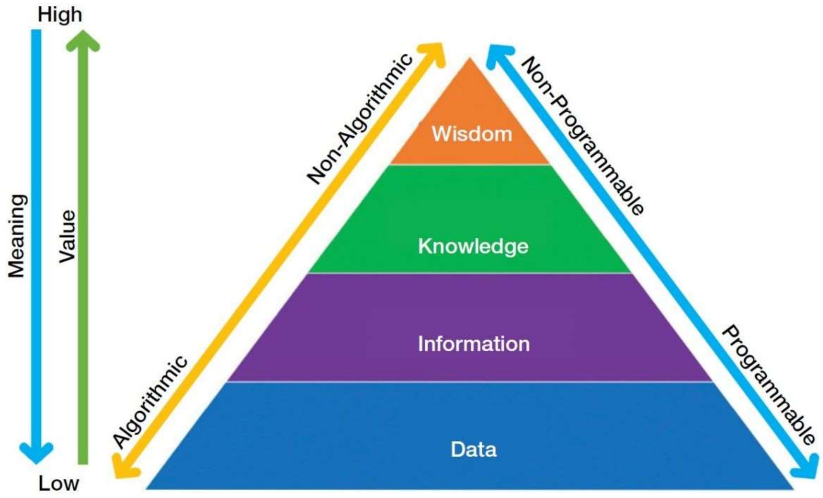


Figure 4: DIKW model for knowledge management by Russell Ackoff

The model consists of four layers:
i) Data, e.g. raw facts on the first level, are produced by business processes and collected company’s data systems (Bocij et al., 2015, p. 5-11).

ii) Information is achieved when data is processed for a purpose, e.g. information need, is applied through a transformation using a specified process, and is placed to an appropriate context to become meaningful and understood by the recipient. Information is required to reduce uncertainty and to influence managerial decisions.
We can therefore calculate the value of information in a simplified manner as:
total value =
the tangible value of information
+ intangible value of improvement in decision making
- cost of gathering the information.

However, decision making on a strategic level seldom is simple, and this makes calculating benefits for Business Information Systems (BIS) challenging and complex. (Bocij et al., 2015, p. 5-11)

iii) Knowledge is obtained by applying information or sets of information in practice along with people’s decision-making styles, competencies, specialized skills, intuitions, and motivations. Knowledge applied to current business problems accumulates learnings. Knowledge is often divided into explicit, e.g. "know-what" or tangible, formal, structured knowledge and tacit, e.g. "know-how" or intangible, internal, intuitive knowledge. (Bocij et al., 2015, p. 15-16)

iv) Wisdom is ultimately derived from knowledge by consolidating learnings through a longer period of time (Bocij et al., 2015, p. 15-16).

Systems operating on the different layers of the data pyramid are explained, for example, in the book ’Artificial Intelligence for Business’.

Data processing of routine business transactions are done in structured transaction processing system (TPS) environments by operational staff using routines of fact findings and predefined procedures. Automation of such structured systems and their predefined procedures is a common and easy activity as of now. (Akerkar, 2019, p. 1-8)

Information is then collected and analysed using concepts and rules to tactical systems like decision support system (DSS) or management information system (MIS) to use middle management on reports and decision making. MIS usually forms standard and exceptional reports from TPS data. DSS works on structured and semi-structured information to utilize models and databases to produce detailed analysis, solution alternatives or cost-benefit ratios for effective decision making. TPS, MIS, and DSS focus on business transaction processes; however, they lack proper knowledge and do not take decisions or justify decisions based on explanations and reasoning. (Akerkar, 2019, p. 1-8)

Knowledge is synthesized and collected from higher management decision making to knowledge base systems (KBS) using heuristics and models. Information judged with ethics and principles, which collect sufficient maturity, can be generalized, and produced into knowledge. (Akerkar, 2019, p. 1-8)

Wisdom, experience, morals, and principles are applied by strategy makers when creating policies and visions in strategy planning. (Akerkar, 2019, p. 1-8)

In summary, to acquire knowledge that can be acted upon, a convergence of data requires activities including acting, interaction, research, reflection, and engagement. The decision-making events on this level are rare, and due to requirements set for knowledge, the accumulation of it is slow. (Akerkar, 2019, p. 1-8)

Autonomous distributed systems and situational awareness

The contribution of AI development to computer science has been the ability to equip distributed computer systems with a degree of autonomy, creating independent agents in place of passive software components. These systems are supplied with capabilities like choosing an action from a set of possible actions, being able to sensor their environment, deciding of becoming active or staying inactive, being able to communicate and cooperate with other agents and even humans and have the ability to maintain goals.
Such computer agents control large scale installations, and personal assistants can manage our daily information in our wearable computer systems and even drive our cars. (Akerkar, 2019, p. 8-10)

Situation awareness is closely related to autonomy, as autonomous systems must act in variable situations based on their location, orientation, and environment. This involves solving two main problems: detecting the situation and the environment, and then choosing appropriate decision and reaction. In big data environments, for example, the Internet, the task is no longer simple as the highly unstructured and dynamic environment. Information creation based on learning, autonomy and situatedness are having a high focus on AI research and development still, and a large number of single methods need to be integrated into greater systems. (Akerkar, 2019, p. 8-10)

Big data and data acquisition for business use

As the development of agents and AI applications require enormous quantities of data, demand for high-quality data is increasing. As businesses start to realize the market value of their recorded business history, data is becoming a real business asset and valued commodity. First, the competition for the most advanced data reserves has started from search queries, web page clicks, online purchases. Next, the offline world started to be digitized with corporate surveillance strategies like Amazon’s grocery store monitoring, AI personal assistants and Internet of Things (IoT) sensory equipment.
However, one key criterion for businesses to finally benefit from the procured data and to develop their business strategy, the data must be of high quality. Crowd-sourcing and having the human in the loop, used at Facebook and Google when having users generate text posts and queries, is a common practice for data collection and producing classification for uncleaned data in the business of data providers and collaborators. This requires

introducing effective and high-quality data acquisition strategies for the businesses procuring datasets for AI and machine learning development. (Akerkar, 2019, p. 10-11)

Data quality has traditionally been managed with statistical techniques. While statistics can produce patterns from data, additional techniques have been developed in data science to produce automated pattern recognition.
Data mining is a discipline that aims to derive information from data by applying algorithms and techniques to map available large datasets and possibly unstructured data to digestible patterns. It is also part of a broader concept of knowledge discovery in databases (KDD). These techniques include pattern recognition, classification, partitioning, clustering, and production of statistical models.
Machine learning is closely related to data mining. The emphasis is on producing automation to enable machines to map and learn data structures and apply that information model and data to solve a problem. The field applies mathematical and statistical techniques in implementing automation algorithms.
Data science is the broad term covering all fields of finding useful patterns from data and valuable insights from data to aid in solving analytical problems. (Akerkar, 2019, p. 11-12)

Applying AI/ML and predictions to practical business problems

Several examples of efficient use of AI in day to day business have been produced for the past decade. We have a good amount of information to identify best practices and common solution models to business problems.

The book ’Artificial Intelligence for Business’ Akerkar (2019) lists typical application areas for utilizing AI in the business domain. In customer relations, regression analysis and clustering techniques help systems analyse customer demographics and transaction history to create customer profiles, map them to customer cohorts and segments, and target marketing efforts for better efficiency.
The financial sector makes frequent use of outlier detection and predictive analytics for anomaly and fraud detection. Previously this has been performed purely via statistical techniques, but with AI, outlier detection has been made a critical tool in other business areas. In the demand planning area, predictions are used for analysing time-series data to make general forecasts for the demand of products and services. With the help of AI, online retailers can use a large volume of customer behaviour data and external data sources to predict demand fluctuations.
Product operation efficiency benefits from predictions based on historical data and real-time sensory data to imply which machines and parts are anticipated to perform poorly

and require maintenance. Such predictions are useful for manufacturers, energy producers and other businesses who rely on complex and sensitive machinery. (Akerkar, 2019, p. 15-16)

Experts panel ’Artificial Intelligence and Machine Learning application in finance and technology’ reported by Antonov (2018) highlights also typical examples of application areas for AI & ML seen in the industry:

1. Process automation applications
    a. Logistics: DART (Dynamic Analysis and Replanning Tool)
    b. Filling of input forms and identification of missing data in forms
    c. Robotic automation, e.g. analysis and automation of click-through paths
    d. Recommendation system and use of activity databases
    e. Natural language supported translation system
2. Customer engagement
    a. Recommender applications, e.g. Similar products, Cross-sell, Up-sell, Novelty
    b. Propensity analysis and action history analysis
    c. Conversational agents: rule-based, NLP & AI-supported dialogue learning systems
3. Automated advanced analytics
    a. Anomaly and outlier identification in large databases
    b. Data summarization over large datasets
    c. Variable analysis, e.g. correlation analysis, significance
    d. Data quality analysis
    e. Dimension reduction at scale - Search engines

In a book called ’Artificial Intelligence in Practice’, Marr and Ward (2019) presents 50 successful companies that utilize machine learning and predictions to solve their business problems.

One example is a Chinese multinational e-commerce network and the world’s largest cloud computing provider Alibaba Group.
Their customers use artificial intelligence methods to find the best recommendations for products customers may need via individually customized catalogue pages. Visitor pages are created with a semi-supervised learning engine that includes vast amounts of historical customer behaviour data, customer profiling algorithms, and reinforcement learning to map current customers to the best matching profiles. Alibaba also develops an automated content generator, a natural language processing AI using deep learning neural networks to produce descriptions to sales item texts and produces alternatives to be tested on the customer behaviour models. The system determines which content is most likely to result in customer clicks and visit ending to a purchase. By using cloud platforms, Alibaba can serve millions of customers while efficiently collecting valuable data on how customers behave. (Marr and Ward, 2019, p. 13-19)

Another example is Apple Inc. who aim to move machine learning from the central cloud to the very far end of mobile devices, where each system will monitor sensors to train the model responsible for security, facial recognition, camera image processing, augmented reality and battery life management. This is faster than uploading data to the cloud, waiting for it to be processed, and downloading back to the device. Yet the algorithms are provided with limited data and are not benefiting from learning from cloud-based and crowdsourced datasets. (Marr and Ward, 2019, p. 37-39)

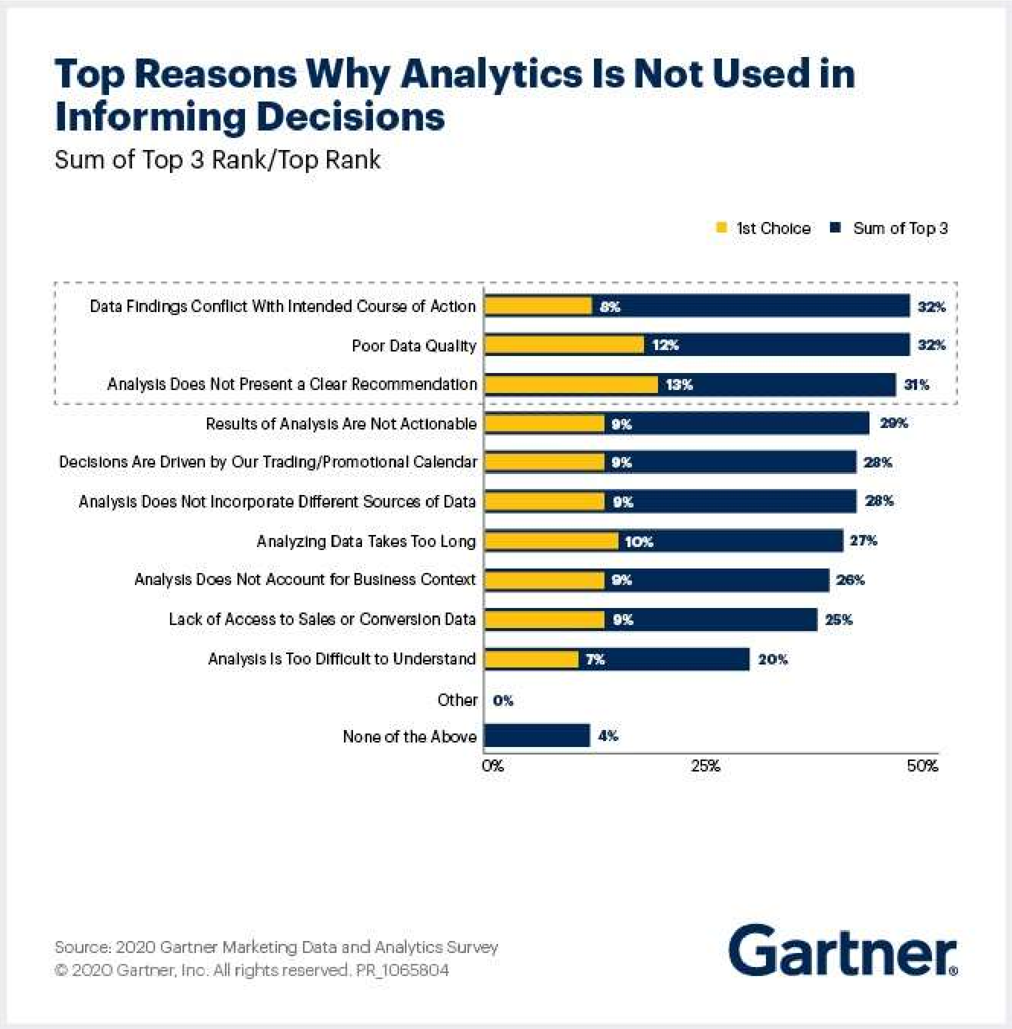
On the logistics and supply chain automation, JD.com, high tech online retail company from China, has built their flagship delivery fulfilment centre in Shanghai to process 200 000 daily orders using automation while maintained by a total of 4 persons. Factory robots and process automation enhanced with machine learning are responsible for receiving orders and collecting materials for packing robots, which dispatch them for delivery on next-day delivery service levels to any of China’s residents. Their Beijing store is designed as a human-free operation, where customers can collect their products and pay with their registered facial identification information. While business is booming, this has raised questions of ethics on expanding automation to reduce 50% of the company human staff. (Marr and Ward, 2019, p. 61-67)

Burberry, a fashion retail company in Britain, sells luxury goods online and in over 500 bricks-and-mortar stores spread to some 50 countries. These stores represent a strategic advantage in providing the luxury customer experience, assistance in evaluating fine craft and high-quality products - a service that is extremely hard to be replicated with robots and AI. To compete with the convenience and scale of online shopping, Burberry’s strategy is to use AI to improve store experience by using advanced data technologies to produce personalized loyalty programs from customer data and profiles. This expands on the sales assistant’s ability to recommend products to customer, not only from their previous purchase preferences but also from the customer profile. Stores also give real-life test for online recommendation algorithms where it’s possible to compare if products sell in different volumes online vs in-store and identify root causes, i.e. in case of outdated images or other detail, they can easily improve online. (Marr and Ward, 2019, p. 83-86)

Pit falls in applying AI and Data science

As it is with any new technology development and engineering effort, there are also challenges and pitfalls demonstrated with many projects. Those give light to typical problem areas to avoid. Currently, the general understanding of what AI is capable of providing us and what we think it can provide also contains several misconceptions.

Looking at web-based marketing, where transactions data are a vast resource, Garner survey on marketing analytics cite poor data quality, unactionable results as the top reasons for not relying on marketing analytics solutions in decision making (Omale, 2020).



Source: 2020 Gartner Marketing Data and Analytics Survey
© 2020 Gartner, Inc. All rights reserved. PR_1065804

Figure 5: Gartner’s Top reasons why analytics not used in decision making

Other findings include data findings conflicts, confirmation bias, deprioritizing skill development in data science, focus on traditional business areas for analytics, and suboptimization instead of exploring new business areas, challenges of deploying analytics to operative decision making. Many companies find it difficult to quantify the correlation of analytics insights and the company bottom line. Lack of clear calculation and regular evaluation of ROI gives a poor image of the importance of the company analytics team. (Omale, 2020)

Book called ’The 9 Pitfalls of Data Science’ Gary Smith (2019) point out that dangerous conclusions can be made when machines are left unattended to find patterns in data. Among the great tales of success and hype of data science, it is easy to be fooled into thinking that the solutions are easily at hand’s reach and all the found patterns are meaningful. The book lists the pitfalls as:

- Using Bad Data
- Putting Data Before Theory
- Worshiping Math
- Worshiping Computers
- Torturing Data
- Fooling Yourself
- Confusing Correlation with Causation
- Being Surprised by Regression Toward the Mean
- Doing Harm

Again, data quality and domain understanding are highlighted as key topics, along with an unbalanced focus on individual areas of AI system development. (Gary Smith, 2019)

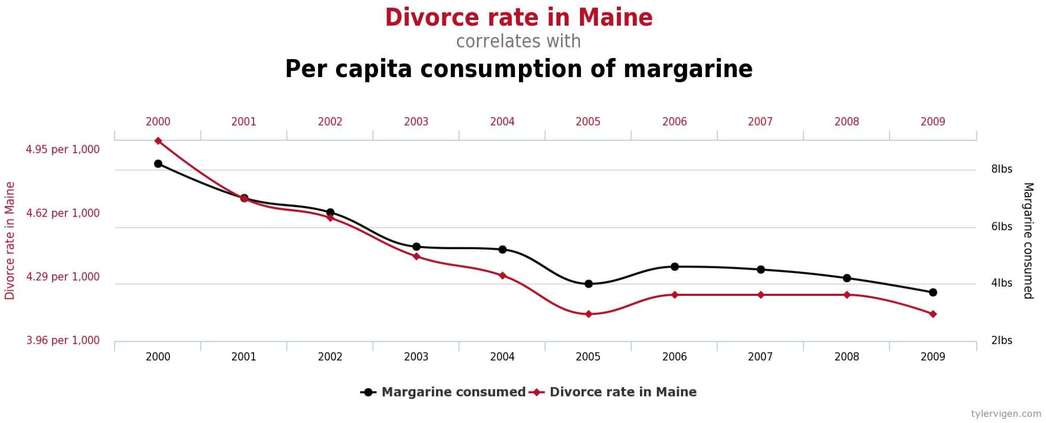


Figure 6: High correlation does not mean causality, tylervigen.com

At a recent conference, ’Data Science Salon Miami 2018’ held for 300 data scientists and data analysts, a key panel discussion ’Artificial Intelligence and Machine Learning application in finance and technology’ also highlighted few possible pitfalls in applying AI and Data Science. Here are some key points, to name a few.

Firstly, it was pointed that current general capabilities for AI applications, which should be named ’Weak AI’, aim to adequately perform one or a few specific cognitive tasks in an automated manner. However, Weak AI does not aim to replicate human approaches to problem-solving. (Antonov, 2018)

Furthermore, there seem to be areas where high performance is confused with competence or skill in solving a particular problem. Only when data indicates a phenomenon can the system identify and utilize it for problem-solving. And the flow of problem-solving may differ greatly from the human approach. Often the highest performance may be received using black-box systems where logic could not be inspected by humans (i.e. deep neural networks). In areas where the business requires high-level explainability and interpretability, such AI tools are not feasible. Efficient AI capability may not, therefore, transfer to other problem domains of similar-sounding settings. This has resulted from a lack of knowledge on how AI algorithms work and what data is available. (Antonov, 2018)

Additionally, the problem with high performance is the somehow assumed exponential growth expectation of AI's advances, similar to advances in general computation power. It was argued that in terms of AI advancement, we have already reached its top progress level for sound processing, we are nearing the top for image and video processing, and we're halfway up the curve for text processing advancements. Advancements in any field of science require several parallel researches to be executed to discover and confirm new knowledge. This process traditionally takes years, even decades. (Antonov, 2018)

Often the complexity of the problem area can be misunderstood by examples in similar fields. Panellists gave an example from the Health & Finance industry where the application of AI to finance is somewhat easier than in the health industry since financial data tends to be well-curated. In general, different areas of patient health data can be messy (and incomplete), and besides, the human body itself is inherently complex. In contrast, health insurance financial data is well-curated and fairly simple, as transactions and constraints are highly documented, and people need to get paid swiftly. (Antonov, 2018)

A similar pitfall example is explained in the book 'Data science for Business' by Provost and Fawcett (2013). Fraud detection in the Finance sector to identify credit card transactions is a classic supervised data mining task where it is safe to assume that nearly all fraud events are identifiable and reliably labelled in the datasets, minimum by having different persons acting in the transactions than normally expected. Fraud detection in the Medicare sector may seem a similar conventional problem. In contrast, the business problem does not have similar descriptions in the data. Persons who commit frauds, medical providers who submit false claims, are also normal actors in the legal transactions - there is no distinction of an illegal transaction by the person.

Furthermore, the billing data have no reliable target variable, which would enable supervised learning to use credit card fraud to work in this scenario. Different approaches need to be utilized, such as profiling, clustering, anomaly detection, and co-occurrence grouping. The key is to match the actual business data to the known data mining tasks fit for the problem. (Provost and Fawcett, 2013, p. 29)

Two key factors were pointed out for making Data Science more accessible for business: One is to improve the ability to communicate AI and ML concepts, ideas, and functionalities to business stakeholders in a way they understand.
Secondly, it needs to be acknowledged that many of the underlying algorithms require mastery of multiple scientific disciplines such as computer science, mathematics, linguistics, physics, psychology, etc. (Antonov, 2018)

This means the work done in the AI project needs data scientist able to explain the AI and ML functionalities in layman's terms to demonstrate they know their work. On the other hand, businesspersons may not be able to lead data science work unless they acquire appropriate knowledge of the scientific disciplines and AI development work. (Antonov, 2018)

Data strategy development

To become an effective data-driven organization company should have good data alignment and direction, i.e. a strategy focused on data.

Book called 'Data strategy and Enterprise data executive' (Aiken and Harbour, 2017) summarizes recent articles and academic studies, which show that data-driven organizations outperform the competition by being more profitable and retaining more customers. Further improved financial performance is achieved by having a top executive (CDO, Chief Data Officer or EDE, Enterprise Data Executive) responsible for data management and data strategy in the company. (Aiken and Harbour, 2017, p. 9-13)

Developing data strategies is the initial step toward organizations becoming data-centric and having data transformed into strategic assets. Leadership needs to be able to communicate: What the organisation wants from the data? Which of the data are strategical business assets and which are not? And how they want the assets to be managed, measured, and reported, to ensure the organization uses the assets to their full potential.

Aiken’s book lists three key requirements for improving the efficiency of operational data utilization and overcoming data-related issues in organizations:

’Data literacy’ meaning the ability to read, create and communicate data as information needs to be moved from individuals’ abilities to organization ability. This encourages everyone in the organization to gain awareness of the shared best practices and methods to operate the company data supply chain. (Aiken and Harbour, 2017, p. 15-18)

A clear ’Data Supply Chain’ is needed, with a uniform, documented, and repeatable set of enterprise backed processes, to provide reliable data towards decision making and to enable predictable business results. Projects tend to finish and have no mechanism to maintain data supply. However, organizations can implement pervasive data supply logistics and can standardize the data supply chain to ensure continuous data quality. (Aiken and Harbour, 2017, p. 15-18)

Well defined set of ’Standard Data Assets’ form the baseline for meaningful communication in an organization to extract value from data assets. For organizations to become data-centric, documenting the business-critical data assets using data dictionaries and similar techniques is a mandatory activity to perform. The company would not manage any other strategic assets, such as material inventories, without having meticulous financial bookkeeping. (Aiken and Harbour, 2017, p. 15-18)

As organizations evolve, they usually become more and more complex. The majority of the organization’s activities are done based on ’legacy data’, which means the data systems already existing in production. When organizations acquire or merge with other organizations, they must develop the legacy systems to support critical business functions. Oftentimes organizations fail to develop data management practices, resulting in problems with data quality. Any activities on data analytics development in such an environment regularly underperform since most resources are spent on IT instead of managing the data assets. The key activity to improve the clarity of development direction is to create a data strategy, which guides data management and usage of data assets, defines data goals, and describes the application of the data towards achieving business objectives. (Aiken and Harbour, 2017, p. 15-18)

Aiken underlines that the focus of the data strategy is not on technology, but it is a balance of people, process, policy, and technology (P3T). It should bring together discipline areas of organizational data governance, data quality management, metadata management, BI and analytics management, data architecture management, and data security management. (Aiken and Harbour, 2017, p. 29-38)

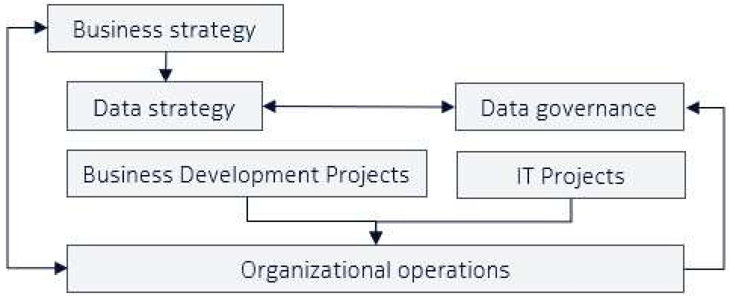


Figure 7: Data strategy and governance support to organization business strategy

Aiken lists three main requirements for data strategy (Aiken and Harbour, 2017, p. 29-38):

- Strategy should be concise and actionable, in contrast to organizational strategy.
- Supports current organizational business strategy.
- Easily understood by both business and IT

A well-formulated data strategy can help harden and reinforce business strategies of any quality (Aiken and Harbour, 2017, p. 29-38).

Data strategy should consolidate results of three core strategy development activities. ’Analysis’ meaning the actions taken to investigate company position in the market, analysis and inventory of existing organizational data assets, improvement prioritization of the data assets and individual data collections. (Aiken and Harbour, 2017, p. 29-38)

’Choice’ activity happens after analysis. Strategy must convey choice on directions by evaluating the trade-offs. Knowing the business strategy decisions organizations need to make, data strategy should determine the ability of the data assets and data collections to support in organization business decision making. (Aiken and Harbour, 2017, p. 29-38)

’Implementation plan’ activity in data strategy definition should document a concise, actionable, and easily understandable plan, so that all organization personnel understand their roles and responsibilities by which they contribute to the larger strategic intent. (Aiken and Harbour, 2017, p. 29-38)

Transforming the organizational habit requires a large fortitude and commitment, and continuous communication to shift the organization to think that data is an important corporate asset. Language of the data strategy must be engaging to both business and IT persons, and particularly to budget authorities. Language needs to be business-friendly and framed as much as possible to business conversations, where experts discuss data form and function and propose technical solutions to business problems. (Aiken and Harbour, 2017, p. 29-38)

Data science relation to data strategy

A good data strategy helps in the grand scheme of all things data related and in managing the complexity and quality of data. All organizations should have Data Management in some shape or form already in place. Based on the Data Management Book of Knowledge (DMBoK2) published by DAMA, reviewer Peter Vennel has produced the following pyramid diagram showing the different levels data management is touching and how data science is related to data management. (Vennel, 2019)

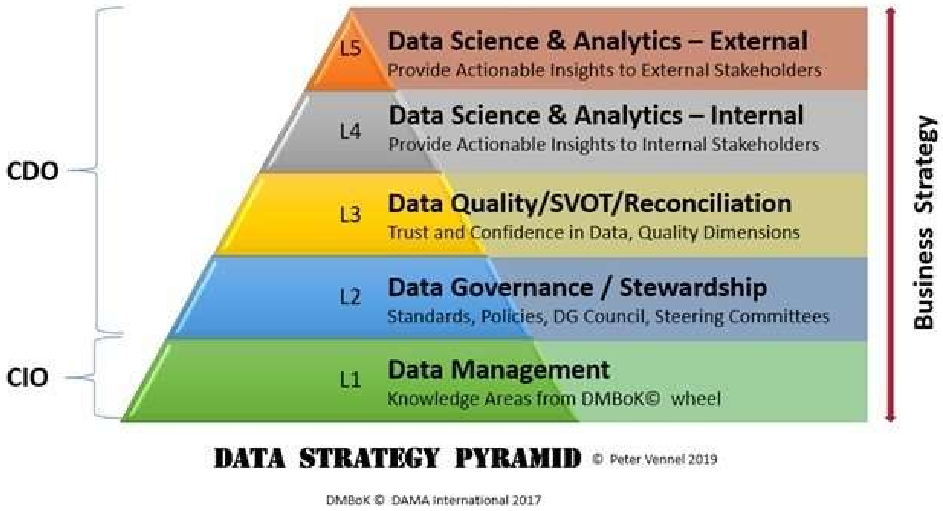

CDO
CIO
L5 Data Science & Analytics – External
Provide Actionable Insights to External Stakeholders
L4 Data Science & Analytics – Internal
Provide Actionable Insights to Internal Stakeholders
L3 Data Quality/SVOT/Reconciliation
Trust and Confidence in Data, Quality Dimensions
L2 Data Governance / Stewardship
Standards, Policies, DG Council, Steering Committees
L1 Data Management
Knowledge Areas from DMBoK© wheel
Business Strategy
DATA STRATEGY PYRAMID © Peter Vennel 2019
DMBoK © DAMA International 2017

Figure 8: Data Strategy Pyramid based on DMBoK2 by Peter Vennel

DMBoK2 defines the 10 knowledge areas for Data Management to consider as the baseline for data strategy efforts. The knowledge areas are (DAMA International, 2009):

- Data Quality
- Data Architecture
- Metadata
- Data Modelling & Design
- Data Warehousing & BI
- Reference and Master Data
- Document & Content Management
- Data Integration & Interoperability
- Data Storage & Operations
- Data Security

The area of data management is very IT-centric. It has various levels of implementation extent and quality depending on a different mode of development, e.g. testing, daily operations, tactical services, etc. Chief Information Officer (CIO) will be the executive driver for

Data Management Level, whereas Chief Data Officer (CDO) will be the executive driver for all other levels (DAMA International, 2009).

On top of the Data Management layer, efficient Data Governance and Data Stewardship formulate standards and policies to ensure data-centric culture and well-formed working practices across the organization. Typically, a matrix type of organization leads the Data Governance activities, working closely with the data stewards from business and IT. Successful data governance requires attention and support from C-level management in the organization (DAMA International, 2009).

As discussed in previous chapters, data quality is essential when you venture out to do analytics. One of the key aspects of data quality is to manage the costs of data and to ensure savings are produced for the organization. Unifying data management and harmonizing data utilization on standard reports and analytics ensures trust and confidence in the organization data assets. Ensuring compliance to data policies and guidelines has a significant impact in avoiding fines when being audited for violations against statutory guidelines, policies about handling information about our customers, products, or persons, and related data protection regulations like GDPR. (Vennel, 2019)

Firstly, the base condition is to efficiently operate the foundations of data management, governance, and data quality activities. In later phases, it is possible to enable data monetization with the help of Data Science & Analytics. This helps businesses grow and improve operational efficiencies and profit margin. (Vennel, 2019)

A mature organization will have seasoned data scientists and analysts working with the data landscape to gather all necessary basic statistics of the data and gathering insights, hidden patterns and trends using techniques like data mining. (Vennel, 2019)

Evolution from the level of a mature data-centric organization towards organization developing predictive systems is likely to produce successful projects and high-quality results, helping the organization utilize digital services and predictions in their decision-making processes. (Vennel, 2019)

The Pinnacle of data strategy for analytics is working on data analytics projects with external parties. Starting with the supply chain and vendors to investigate how to improve your business data related to your products and supply is a great help for driving the business cost-effectively and efficiently. You will soon be able to provide actionable insights about your business to Customers, Vendors, suppliers, and other external stakeholders. (Vennel, 2019)

Analytics work can also be enhanced by working with external analytics vendors. Naturally, they produce external factors and constraints influencing your work. However, external vendors who have been in the business for years have efficient working practices in place for developing their data services. And in addition, vendors who operate with several different organizations have the ability to bring in a large amount of experience on what the industry common practices are and how to revolutionize the business. (Vennel, 2019)

2.2 Data science and ML/AI methods overview and application in discreet and practical business operations

Basic concepts of Machine Learning (ML) and Artificial Intelligence (AI) are described in the widely used university textbook ’Artificial Intelligence: A Modern Approach’ written by Russell (2016). What we currently understand as artificial intelligence (AI) on a practical level is a technological solution capable of combining multiple sources of data, being able to sense and cognize environment, being able to perform activities with the ability to learn from the outcomes and to adapt performance over time. (Russell, 2016, p. 1-32)

Concepts of AI have been developed since the age of computers began and the development of intelligent systems. One official start for AI is considered to be the "Dartmouth conference" of 1956. However, key concepts such as the Turing test to identify "intelligent machine" predates that event. We have also received ideas in science fiction of what AI could be. While some AI projects are still pursuing the original goal of achieving machine intelligence matching to living beings and robotics with humanoid appearances, most of the AI projects today focus on automating and solving complex but discreet practical problems. (Russell, 2016, p. 1-32)

Solutions to practical problems include AI systems being able to sense the environment using computer vision, media processing and sensor signal processing, being able to cognize information using knowledge representation and natural language processing (NLP), to perform actions using machine learning (ML) methods and knowledge base systems (KBS). The concrete applications for these systems are already covering almost every industry. They promise to transform and create new business models on the market and produce new competitive advantage for companies adapting AI methods. A showcase example in Russell's book is AI-optimized fraud-detection system in the financial services industry, which improves process accuracy and speed. It is estimated to have a market size of $3 billion in 2020. (Akerkar, 2019, p. 3-6)

Predictive analytics as such is the process of using historical data to make predictions. In the last decade, the application of predictive analytics and the development of AI has become increasingly faster, smarter, and more actionable than before. (Akerkar, 2019, p. 14-15)

At the core of the predictive analytics is the model, which can be generally classified into two types of categories:
i) Regression models, which are used to map the correlation between specific variables and outcomes. Correlation coefficients give you a mathematically quantified measure of the relationships and probability of how likely a certain outcome is based on a selection of variables.
ii) Classification models, which use a regression model to assign a probability to an outcome or an event and determine based on variables to which category or cluster the event belongs to. These methods, their difference, and way of utilizing the data are explained in the next graphic. (Akerkar, 2019, p. 14-15)

Examples of how these different types of models produce outputs are described in the image below

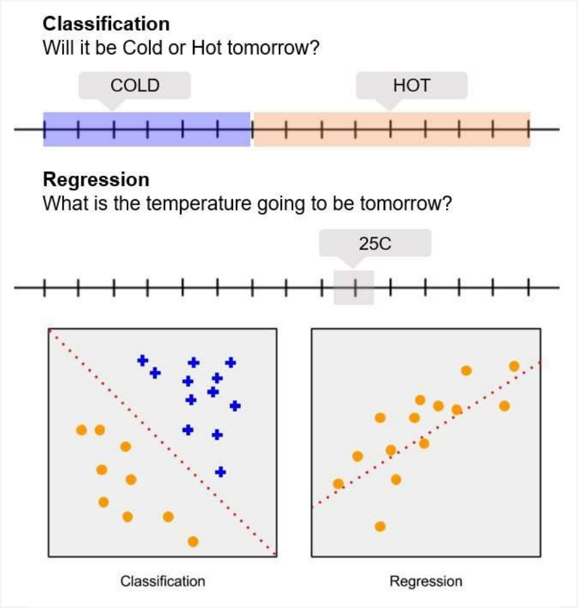

Classification
Will it be Cold or Hot tomorrow?
COLD
HOT
Regression
What is the temperature going to be tomorrow?
25C
Classification
Regression

Figure 9: Regression vs classification models

With the help of advancements in the field of AI, we are now able to create prediction models based on large volumes of real-time information and a multitude of variables. This makes the outcomes and predictions much more reliable and precise. A further benefit of AI is the cognitive ability to take actions based on predictions and available data. (Akerkar, 2019, p. 14-15)

Data science in decision making processes

To utilize a data-driven approach for operational decision making in the organization, companies need to adjust their processes to bring AI into the mix. And in some occasions to remove the human participation entirely to produce AI-driven decision making. Examples of models for implementing AI in business decision making are presented in a Harvard Business Review article by Colson (2019).

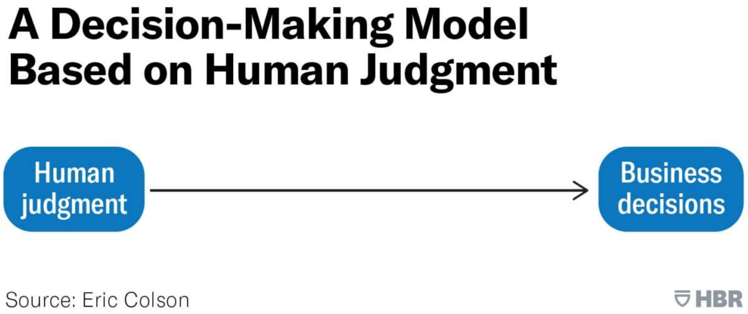
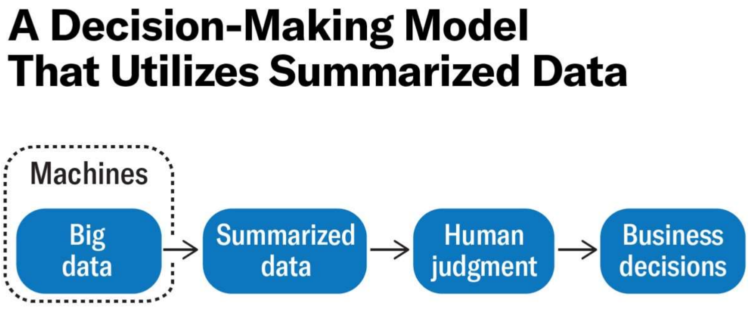


Figure 10: Decision making based on human involvement

The first figure represents the most utilized form of decision-making throughout history, where humans applied their judgment based on learnings, collected knowledge, intuition, and even gut feeling. Moving from simple scenarios to more complex scenarios, our human capacity quickly becomes less effective and can be easily impacted by several cognitive biases that impair our judgement. (Colson, 2019)

Cognitive biases imply that our reasoning is based on multiple simple heuristics or rules of-thumb to enable quick processing of information, even in life-threatening situations. In complex scenarios, these shortcuts don’t often result in optimal or accurate outcomes. (Colson, 2019)

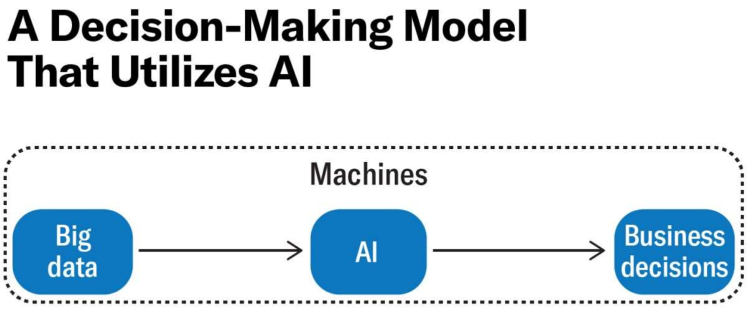


Source: Eric Colson HBR

Figure 11: Decision making based on human and machine involvement

Second phase is a data-driven approach to assist humans in decision making.
In the modern world, challenges of processing a large amount of information by the human brain can be diverted to machines. They produce repeatable and accurate outputs to inform better decisions. IT services are designed to reduce the data volume to digestible summaries for human consumption in information analytics environments like spreadsheets, dashboards, and analytics applications. The final output in such a data-driven approach is a relatively small set of data presented for human judgment and decision making. (Colson, 2019)

All the same, the "data-driven" approach has limitations as humans act in the central processor role. Cognitive bias is still a factor, and the effect may vary by day. Using summarized data removes a large set of information describing the data like distribution and outliers, displaying relationships between data elements. There is a concern we may miss important aspects of data from decision-making. Summarized data may also produce conflicting contradicting results compared to reality as contributing factors are not shown. Related information can be studied, for example, from the theory of Simpson's paradox and A/B testing with randomized control trials in analytics. (Colson, 2019)

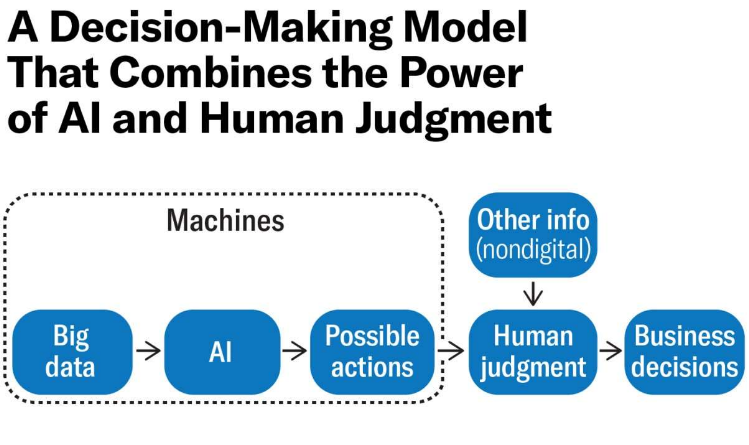


Source: Eric Colson HBR

Figure 12: Decision making based on machine involvement

The next evolution stage requires the introduction of ’AI-based decision making’. For routine decisions done on structured data where decision-making processes are clearly defined and documented, it is possible to introduce AI as an actor to the decision-making flow. AI system needs to be trained to identify operative and decision-making patterns that best explain the decision outcomes. Once trained and fully automated, the AI system can repeat the process accurately in the same way. And since the AI system does not get tired, it’s relatively easy to expand the processing to thousands or millions of transactions. (Colson, 2019)

While covering a large portion of daily operations with the ’AI-based decision making’ approach, there are still many business decisions, which does not rely solely on structured data. More complex managerial decisions on tactical and strategic levels often utilise information from several sources, knowledge from explicit sources, and tacit knowledge from experience and intuition. (Colson, 2019)



Source: Eric Colson HBR

Figure 13: Decision making based on machine and human involvement

On the final level, the process of ’Prescriptive AI-supported decision making’ takes leverage of both AI and Human actors. This enables human to make objective and rational decisions based on actions proposed by AI while being fully aware of the tacit information, wisdom, strategies and policies, which is inaccessible to AI and extremely relevant to the business decision. (Colson, 2019)

Data driven decision-making approach in our FiRe program

In this thesis, we focus on machine learning (ML) methods, which aim to develop computational approaches to making sense of big volumes of data and automating the use of such data. Instead of predefined rules, the fundamental goal is to let the machine identify important patterns in the data to create the logic and rules. The organization will then use logic to produce predictions in the subject area of product reliability and forecasting of product returns.

2.3 Developing predictive systems - Overview of Bayesian theorem

Descriptive analytics, i.e. the use of data to describe what has happened, is the starting point for most businesses in business analytics. (Akerkar, 2019, p. 65)

The amount of information is increasing by 2.5 quintillion bytes per day. However, most of the information is not of high quality or particularly useful. This presents a problem that the noise level in the data is increasing faster, the number of hypotheses to test is increasing. Still, there are a rather small amount of objective truths that have been tested to be true, as described by Nathan Silver in his book ’The signal and the noise’ (Silver, 2012).

Prediction solutions connect subjective and objective reality, meaning that we can test scientific hypotheses in the real world. Nate Silver presents in his book ’The signal and the noise’ an overview of the theorem of Karl Popper. Theorem stated that a hypothesis is not scientific unless it can be falsifiable and proven wrong in the real world. For example, the hypothesis that "all swans are white" can be falsified by setting an observation for a black swan, which we know to occur in nature. This gives us the paradox that not all theories are possible to be tested, and in addition, we have many theories we yet haven’t had the opportunity to test. An alternative approach is provided in Bayes’ theorem, which implies that we must think differently about our ideas and about testing them.
Silver summarizes that Bayes’ theorem is a mathematical approach where we rely on probability and understanding of the uncertainty, assumptions and beliefs embedded in the formula. This raises the emphasis on data quality and evaluating the usefulness of the collected data and information. (Silver, 2012, p. 14-15)

Critical thinking and fox-like attitude in prediction expert evaluation

One of the key disciplines in producing high-quality predictive solutions is to keep the right attitude in thinking. Silver makes use of the prediction expert classification spectrum from Philip Tetlock. The spectrum describes the far ends with names “hedgehogs” and “foxes”. The critical thinking style by the ’foxes’ is applied by using a multidisciplinary empirical approach, learning and adaptation to new information, self-criticism and correction of mistakes, tolerance for complexity, and caution towards probabilities and quality of assumptions behind statements.

At the opposite end are the ’hedgehogs’, who are specialized and confront new opinions sceptically. They prefer to expand their existing approach as "all-in-one", confidence to own creation is unquestionable, reason and causes for mistakes are collected from outside circumstances, and they seek to find simple control of the order. They tend only to worsen in performance when presented with new information. They resort to permuting and manipulating the data to validate their existing standing and confirm their own biases. However, Silver states that they do make excellent entertainment and television guests. (Silver, 2012, p. 52-57)

Pareto Principle of Prediction

There is an inherent learning curve on prediction development and learning, meaning that progress is not linear. This is often presented as the ’Pareto Principle of Prediction’, which demonstrates the ’80-20 rule’. Learning the rules of the game and producing a prediction system in real life takes effort. Silver points out an example that happened in the boom years of online poker. Few players with experience of several years earned a large portion of the money as new players lacking experience and skill entered the games. Because of the excitement of the new game, they forgot to follow the basic rules: fold the bad hands, bet on good hands, and estimate the probability if your opponent has a better hand.
This 20% of knowledge executed 80% of the time gives a good portion of the profits, even against the best poker players. (Silver, 2012, p. 308-313)

Pareto principle presents an opportunity for business analytics and data science application. Due to the high competition of our economy, there is a constant "water level" from the competition. To make a profit, companies need to produce the tip of the iceberg, a competitive advantage to float above the surface.
Having organizations adopt data analytic thinking and basic information of mathematical, statistical and data science principles gives the ability to create necessary strategic assets within the organizations:
Knowing only 20% of data science executed properly for 80% of the time would give you the power of understanding your business data and the opportunity to produce prototypes of prediction models to test. (Silver, 2012, p. 313-315)

The first 20% should be about producing the right quality of data, using the right technology, and having the right objective. In prototyping, the ideal is not on the most pleasing or most convenient prediction but to have accurate and objective outcomes. This applied with few heuristic rules, a systematic approach, and a review of experienced experts in the field will set the organization in speed with the competition. Challenges arise when predictions of the competition are equally performing well, or in the worst case, even better. Organizations should direct the focus of effort in the areas where profit to investment can be expected, where competition lacks good objectives, expresses bad habits, adherence to tradition, or are lagging in technology and data quality - to those areas where forecasters are not doing even the first 20% correctly. (Silver, 2012, p. 313-315)

Failures and pitfalls for predictive solutions

There are several examples of failures and pitfalls for predictive solutions in Nathan's book, where data problems have been disregarded.
These include:
- Data being out of sample, e.g. the collected data do not present evidence the event attempted to be predicted. Silver provides a simple example: you may assess your driving abilities at a Christmas party involving drinking based on your yearly safe car trips vs 2 minor accidents. When none of the trips would contain evidence of driving while drunk, it would be a misinterpretation of the information to say that you'll perform well. Nothing in the past data describes your odds under these conditions would be.
- Overfitting, e.g. providing an overly specific answer to a general question. Silver's example is an assignment to produce a high probability of picking a lock successfully, where 3 true answers were provided as solutions for experimented solutions. Still, they were not applicable if any of the conditions change. In contrast, prediction can also be underfitting, where prediction is based on a broadly generic model and does not represent the actual signal. This is likely when understanding of the fundamental relationships is weak and received data is small in quantity and noisy. This has happened in several earthquake prediction systems, i.e. Keilis-Borok's model, David Bowman's model, Fukushima nuclear plant earthquake prediction model, as earthquakes are complex processes involving simple objects behaving in unexpected ways, among others during sudden and catastrophic failure events. (Silver, 2012, p. 163-172)

Bayesian thinking in the development of predictions

As the volume of recorded information is increasing currently exponentially, the signal to noise ratio may not improve in the same manner, and events do not necessarily become more predictable. Strategy for shortening the gap between what we actually know and what we assume we know requires shifting the mindset to Bayesian thinking of predictions and probabilities of occurrences of real-world events. Secondly, we need to repeatedly revise and improve the predictions when we recognize something to improve. This strategy has been utilized successfully in military and medical fields in multiple use cases with reasonably good results. (Silver, 2012, p. 446-451)

The starting point is a statement of your beliefs and what is your foundation. This is required to establish your subjective view of realities. Problems arise when we blindly believe the approximation to be the reality. Simple universal statements are good for communicating initial highlights to strangers of a subject. But approximations leave out all the grit and details which make predictions possible. And information becomes knowledge only after being applied in a context. (Silver, 2012, p. 446-451)

The first requirement in Bayes' theorem is to provide an explicit measure of likelihood if the analysed event occurs prior to evaluating the outcomes. Here common sense, 50–50 chance, or first approximation can serve as the first prior check to reduce the biases affect to your prediction. The only unacceptable starting point is to state no first approximation. This indicates you have only unknown biases, and many affect your prediction. Your outcomes may become swamped with noise and false positives. (Silver, 2012, p. 446-451)

Data utilized for machine learning is affected by several decisions and phenomenon occurring in the analytical environment. Different types of measurement bias were summarized in an article by Lionbridge AI (2020). Measurement bias occurs when either faulty measurement causes data distortion or collected data do not represent the real world correctly. Sample bias occurs when a dataset does not reflect the realities of the environment in which a model will run. Exclusion bias is commonly occurring at the data pre-processing stage when evaluating which data points are important. Recall bias arises when you label similar types of data inconsistently. Confirmation bias or observer bias affects seeing what you expect to see or want to see in data. Racial bias occurs when data skews in favour of particular demographics. Google's Inclusive Images competition included good examples of how this can occur as technology failed to recognize people of colour. Association bias occurs when the data for a machine learning model reinforces and/or multiplies a cultural bias. Association bias is best known for creating gender bias. (Lionbridge AI, 2020)

The second important rule of the Bayesian theorem is to make multiple forecasts, score them, and choose the best one - to trial and error in volumes. The forecasts should also update when we receive new information, and we should have a good discipline to do so to apply proper evaluation for the new data points. Big companies, like Google, run thousands of experiments with their data and test them with customers to see which apply to the real world and can succeed as general features. Big innovations are created from small, incremental improvements. The sooner we can start to correct our predictions, we start to move away from the biases and pitfalls and towards the real-life events. (Silver, 2012, p. 451-454)

Developing business cases and measuring the performance of predictions

Practical tools for evaluating the business problem and the value of the solution are discussed in the book Data Science for Business by Provost and Fawcett (2013). The first step should be to formulate the discreet business problem where clear data science methods like clustering, regression etc., can be selected to produce an outcome. In the fundamental concepts, Accuracy is defined as a method for describing how a model performs. However, accuracy as such is too high-level variable to evaluate if the algorithm performs well and produces the output expected to solve the business problem. In a typical situation for classification problems, one class is usually rare. In large datasets, most of the population is normal values, and outstanding values are very few in portion. It also makes no distinction between false positive and false negative errors. (Provost and Fawcett, 2013, p. 187-194)

In summary, there is no single metric to evaluate the usefulness and performance of an algorithm in solving a business problem. Analytical engineering is required first to investigate how much we care about errors in predictions and how harmful they are. Secondly, to devise a set of measures to confirm the correctness of the outputs. As a starting point, evaluation of the confusion matrix should be done to identify how many false positives and false negatives are produced. (Provost and Fawcett, 2013, p. 187-194)

A more advanced analytical tool called expected value evaluation framework is introduced in the book to help in defining (i) structure of the problem, (ii) information that we can acquire from data, (iii) information that needs to be acquired from other sources .e.g. from business subject matter experts. The expected value is calculated as the weighted average of the possible outcomes from different situations, where the weight is the probability of occurrence. (Provost and Fawcett, 2013, p. 187-199)

In a classifier example from targeted marketing presented by Provost and Fawcett, we may have classes for consumers being “likely buyers” vs “not likely buyers”. Using a common coin toss probability of 50% would result in most customers being classified as ’not likely buyers’. In a scenario where a consumer can only buy via the advertisement, we get values for response and for no response. Our history data can show the likelihood of a customer buying, and our benefit for buying is product revenue - marketing cost. Assuming product revenue is $100, and marketing cost is $1, giving the benefit of $99. And if the customer does not respond to buy, the expected benefit for no response is zero, and deducting incurred marketing costs brings benefit total to -$1. Expected value for customer x in a scenario where we evaluate if we are making profit or no would be therefore be calculated as: p(x)*$99 - [1 - p(x)]*$1 > 0, and this results to p(x) > 0.01. Therefore, this operation brings benefit according to the expected value as long as the probability of responding with a buy decision is greater than 1%. This example shows how to use the prediction model's output and helps formulate the problem and analysis. Finally, a cost and benefits matrix is produced with the same dimensions as confusion matrix dimensions of true and false positives and negatives. This gives an overview of where the benefits and costs are produced. Business value or cost is not always simple to find, and often the average value is used. Additional or alternative data sources can also have different associated costs involved. (Provost and Fawcett, 2013, p. 187-199)

Additional basic comparative baseline performance analysis methods are used when predictions have been calculated to evaluate the performance. When evaluating the performance of weather forecast models, two baseline measurements are produced as a metric: the first one is to assume that the previous value will also repeat in future, the second baseline assumes that historical statistical value from the same time period from history will also repeat in future time periods. When the dataset is being modified, a simple metric to produce a quick evaluation of performance utilises a majority class identification with a naive classifier. This assumes the most probable occurrence in an imbalanced dataset will dominate in classification producing the highest accuracy available for the data. Any complex model needs to be able to perform better than the simple majority classifier. Another similar method uses a decision stump classifier, meaning a tree with only one node where the algorithm identifies the most significant feature. When increasing the datasets available or adding to the volume of data, evaluation of reduced data models can be useful to understand if the combined data sets perform any better than the previous or individual dataset. (Provost and Fawcett, 2013, p. 204-206)

Finally, a consideration of the dataset sampling method's impact on the selection bias and the outcome of the target variable should be evaluated. Understanding the shape of the data used in modelling is important for the successful evaluation of the expected value framework. Selection bias is tackled with the additional branch of methods and tools outside the scope of this study. (Provost and Fawcett, 2013, p. 279-290)

2.4 Overview of Data Analytics Maturity Models

It is rare today for an organization to develop software that is critical to its business without a defined software development methodology being used; In contrast, it is relatively common for an organization to build analytic models that are critical to its business without using any analytic methodology. (Grossman, 2018, p. 3)

Over a hundred different maturity models have been developed since the 1970s in the information systems field until today to identify the strengths and weaknesses of information management and assist in finding remedial action. A traditional approach to the assessment of analytics capabilities includes self-assessment, qualitative interviews, and quantitative studies, which are good for checking if certain tools and technologies are in use. However, these approaches do not answer if a particular organization uses analytics methods fully to make business decisions. Analytics maturity models provide an alternative to these methods, as a schematic and generalized representation of essential capabilities, to study the implemented "depth and width" of these capabilities.
(Grossman, 2018, p. 3)

For the thesis, I studied a set of 11 most utilized analytics maturity models based on scientific literature and reports, and from analytic sector publications to a paper named 'Analytics Maturity Models: An Overview' written by Król and Zdonek (2020).
The list of models includes Analytic Processes Maturity Model (APMM), Analytics Maturity Quotient Framework, Blast Analytics Maturity Assessment (AMA) Framework, DAMM: Data Analytics Maturity Model for Associations, DELTA Plus Model, Gartner's Maturity Model for Data and Analytics, Logi Analytics Maturity Model, Online Analytics Maturity Model (OAMM), SAS Analytic Maturity Scorecard, TDWI Analytics Maturity Model, Web Analytics Maturity Model (WAMM).

Most of the analysed models comprised of five analytics maturity levels. Models described the levels in a detailed manner with a set of capabilities required to meet each level. This allowed an independent assessment of an organization's analytics maturity using the criteria from the models.

Analytics maturity can be described as the evolution of an organization to integrate, manage, and leverage all relevant internal and external data sources into key decision points. It means creating an ecosystem that enables insight and action. In other words, analytics maturity is not simply about having some technology in place; it involves technologies, data management, analytics, governance, and organizational components. It can take years to create and instil an analytics culture in an organization. (Król and Zdonek, 2020)

Analytic Processes Maturity Model (APMM)

One generally available framework described in the overview for evaluating the analytic maturity of an organization is called the Analytic Processes Maturity Model (APMM) developed by Robert Grossman. The framework describes the analytics processes and capabilities within organizations aiming to operate highly efficient analytical strategies.

The APMM 6 key process areas related to analytics (Grossman, 2018):

i) building analytic models.
ii) deploying analytic models.
iii) managing and operating analytic infrastructure.
iv) protecting analytic assets through appropriate policies and procedures.
v) operating an analytic governance structure.
vi) identifying analytic opportunities, making decisions, and allocating resources based upon an analytic strategy.

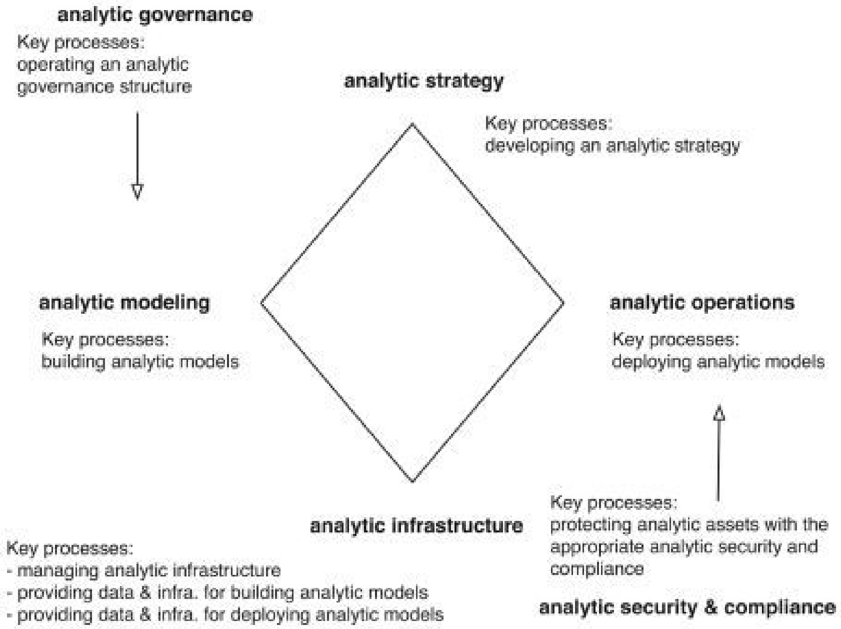

analytic governance
Key processes:
operating an analytic governance structure
analytic strategy
Key processes:
developing an analytic strategy
analytic modeling
Key processes:
building analytic models
analytic operations
Key processes:
deploying analytic models
Key processes:
protecting analytic assets with the appropriate analytic security and compliance
analytic security & compliance
analytic infrastructure
Key processes:
- managing analytic infrastructure
- providing data & infra. for building analytic models
- providing data & infra. for deploying analytic models

Figure 14: Key Processes in Analytic Processes Maturity Model by Grossman

The outcome of the framework analysis is Analytic Maturity Level, or AML score from 1 to 5, which indicates the probability that the organization's processes for building and deploying analytic models will result in analytic models that are statistically valid and are completed according to schedule. The process should ensure the organization can deploy analytical models into an organization's products, services, or operations. The process should also ensure that development efforts meet the organization's goals for the analytical model. (Grossman, 2018, p. 1)

Approach chosen in thesis for evaluating analytics maturity

This thesis utilized APPM for producing an overview understanding of development goals. Analytics Maturity Assessment Framework by Grossman defines key processes and development goals within each area for improving an organization’s analytics maturity.
In addition, I utilized a survey tool called Blast Analytics Maturity Assessment (AMA) for survey questions and scoring. Out of the evaluated maturity assessment frameworks, Blast AMA provided a wide enough range of survey questions for general use and consistent outcomes from the scoring tool.

Blast Analytics Maturity Assessment Framework is described in ’Analytics Maturity Models: An Overview’ to be based on the Online Analytics Maturity Model developed by S. Hamel. The Blast AMA evaluates six key process areas and key success factors for strategy, governance, data management, insights, evolution, and resources. Evaluation is proposed to be carried out quarterly to produce a benchmark to assess different circumstances and progress. Analysis of key success factors in the six process areas recommends analytics development strategy and action plan for implementing the strategy. (Król and Zdonek, 2020, p. 1)

Improving Analytic governance and strategy

Grossman states in his ’Framework for evaluating the analytic maturity of an organization’ that goals of Analytic Governance should include ensuring long-term decisions about analytics in data strategy are reached and that investments in analytics generate business value. Governance should be operated so that data, derived data, and analytic products are well protected and managed in a secure and compliant fashion. Governance should ensure accountability, transparency, and traceability for those funding analytic resources, those developing and supporting analytic resources, and those using analytic resources. The organization structure and governance should ensure that the necessary analytic resources are available; that data is available to those building analytic models; that analytic models can be deployed; and that the impact of analytic models is quantified and tracked in a continuous development fashion. (Grossman, 2018, p. 2)

3. Building data analytics platforms and skilled analytics teams

Overview of technologies available for data analytics platforms, industry standards on data science processes, methods for organizations and team's development towards data driven organizations.

3.1 Technology overview: Alteryx, Microsoft Azure and PowerBI, Python development platforms, Rapid Miner

In the FiRe analytics program in Nokia, the selection of business analytics platform vendors was chosen for the analytics IT architecture based on current partnership and overview of industry market analysis, and additional analytics tool comparison analysis reports. Evaluation included analytics solutions from Alteryx, Microsoft, Rapid Miner, and these were compared to Python development platforms.

Both Microsoft and Alteryx have a history of providing business analytics tools and self-service functional capabilities to end-users to democratize the use of data, and they continue to be at the forefront of self-service and advanced analytic innovation. An additional commercial platform chosen for business analytics vendor comparison was Rapid Miner, a similar offering as Alteryx. Python development provides open-source libraries and tools for analytics development by programmers and data scientists, lacking in more advanced visual editor tools and requiring a code-only way of working. Programmers can choose Python development platforms from desktop environments to big data capable server clusters.

Gartner has produced Magic Quadrant series of market research reports that analyse and demonstrate market trends in the IT area. A Magic Quadrant provides a graphical positioning of four types of competitive technology providers and gives a good first understanding of technology providers organizations might consider investing in. Magic Quadrant for Business Intelligence and Analytics Platforms 2017 highlighted that market is shifting from IT-led reporting to modern business-led analytics as mainstream. Modern business intelligence (BI) and analytics platforms are characterized by easy-to-use tools and visual-based data discovery, a defining feature of modern platforms. Business analytics platforms have support to a full range of analytic workflow capabilities. They do not require significant IT involvement to predefine data models upfront as a prerequisite to analysis work. The objective of these platforms is to enable business teams to focus on analytics instead of tool issues (Gartner, 2017). Magic Quadrant for Data Science and Machine Learning Platforms 2019 concluded how expert data scientists, citizen data scientists and application developers require professional capabilities for building, deploying, and managing analytical models. Data science platforms provide a cohesive software application that offers a mixture of basic building blocks essential for creating all kinds of data science solutions and incorporating those solutions into business processes, surrounding infrastructure, and products. (Gartner, 2019)

Alteryx overview

Alteryx offers a workflow-based end-to-end platform for data preparation and building parameterized analytic applications exploiting open-source Python and R-based packages. Alteryx remained in the Leader segment for 6 consecutive years but was now positioned to Challengers. Emphasis is on making data science accessible to citizen data scientists and others across the end-to-end analytic pipeline is resonating in the market. Its approach provides a natural extension for a client base focused on data preparation but ready to take the next step into data science. With the Alteryx Server, data modellers can create interactive, parameterized dashboards published on-premises or in the cloud via the Analytics Gallery. (Gartner, 2017)

Alteryx strengths

Collaborative enablement of a broad user base: Alteryx's no-code approach is attractive to a broad spectrum of users, from business and data analysts to citizen data scientists. A focus on the ease of use and cohesiveness of its platform enables collaboration between users. Marketing execution: Alteryx's focus on addressing the end-to-end analytic process easily and clearly positions it as a vendor of a comprehensive platform. Alteryx's value proposition is clear and resonates with business needs. Business benefits (2017): Alteryx is in the top quartile for achieving business benefits, for both qualitative and the hard benefits of data monetization. User enablement and skilled resources (2017): User enablement is important for self-service analytics, where business users become the data stewards preparing data and application authors. (Gartner, 2017)

Cautions on Alteryx

Data preparation legacy reputation and market perception. Market understanding. Subscription costs. (Gartner, 2017)

Microsoft overview

Microsoft offers a broad range of BI and analytics capabilities with its Power BI suite, delivered via the Azure cloud. Power BI offers data preparation, data discovery and interactive dashboards via a single design tool. Microsoft Reporting Services and Analysis Services are for traditional enterprise reporting platforms, as on-premises offerings. Excel is frequently used for data analysis, and while it is not considered here as a BI and analytics tool per se. (Gartner, 2017)

Microsoft strengths

Cost: Microsoft is placing downward pricing pressure on the BI and analytics market with a free desktop product and a low subscription price per user per month. However, potential customers should be aware that additional data scale-out options incur additional costs when leveraging Microsoft SQL Azure or HDInsight in the cloud. Microsoft's Azure ML Service development environment is designed for all skill levels by providing visual development tools to code-only development. The environment also consolidates ML model management services to a single place in Azure. Ease of use plus complex analysis: Microsoft's customer reference scores place it in the top quartile for ease of use and complexity of analysis. Ease of use for content consumers was also the most-cited reason for customers choosing Microsoft Power BI. Vision: Microsoft is furthest to the right on the Completeness of Vision axis and has also continued to execute its roadmap with frequent (monthly) product releases. Active community: Microsoft has a strong community of partners, resellers, and individual users. (Gartner, 2017)

Cautions on Microsoft

Product immaturity and cloud-only. The breadth of use. Support. Not the only standard. (Gartner, 2017)

Python environments overview

Python development provides open-source libraries and tools for analytics development by programmers and data scientists, lacking in more advanced visual editor tools and requiring a code-only way of working. Python development platforms can be chosen from desktop environments to big data capable server clusters. Python is quickly becoming the go-to language of machine learning and is used to create models for Bayesian networks, decision trees, and much more.

Azure Databricks (Gartner top vendor in the data science sector) enables programmers with easy transition from a single machine to scalable ML cluster using Python libraries of Pandas, Koalas, and Spark.
Pandas is a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool built on top of the Python programming language. It provides high performance, easy-to-use data structures and data analysis tools. Apache Spark is an open-source parallel processing framework that supports in-memory processing to boost applications that analyse big data. Koalas implements the pandas DataFrame API on top of Apache Spark for enabling big data on parallel computing architecture.
Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms, including support vector machines, random forests, gradient boosting, k-means and DBSCAN. It is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library and is also used for machine learning applications such as neural networks. It is used for both research and production at Google.
PySpark offers a versatile interface for using powerful Spark clusters. Still, it requires a completely different way of thinking and being aware of the differences of local and distributed execution models. (Gartner, 2017)

Python environments strengths

Python has a vast software ecosystem and community able to extend to any data science task. Python is often praised for being a general-purpose language with an easy-to-understand syntax. Most deep learning research is done in Python, so tools such as Keras and PyTorch have "Python-first" development. Python has an edge in deploying models to other pieces of software. Python is a general-purpose programming language, so if you write an application in Python, the process of including your Python-based model is seamless. (Chiu, 2019)

Cautions on Python environments

Python has a vast software ecosystem that may be difficult to absorb. Python was originally developed as a programming language for software development (the data science tools were added later) - people with a computer science or software development background is required for the development team. (Chiu, 2019)

Rapid Miner overview

Rapid Miner is a Leader in Data Science and Machine Learning Platforms, with no BI and Analytics platforms position. Striking a good balance between ease of use and data science sophistication. Its platform's approachability is praised by citizen data scientists, while the richness of its core data science functionality, including its openness to open-source code and functionality, make it appealing to experienced data scientists, too. (Gartner, 2017)

Rapid Miner strengths

Sophisticated simplicity: Features such as Auto Model, augmented analytics capabilities such as Turbo Prep, and an above-average UI make RapidMiner Studio a favourite of citizen data scientists. Advanced features: Ease of use does not preclude the presence of power. Beyond deep learning and GPU support, RapidMiner's platform now includes data augmentation functionality and enhanced time-series features. Coherent end-to-end platform: Reference customers made many complimentary comments about the coherence of RapidMiner's user experience — from its scalable repository management to its real-time scoring. (Gartner, 2017)

Cautions on Rapid Miner

Data preparation and visualization not matching other analytics components of the platform. License and pricing models have complicated pricing schemes and difficult-to-navigate pricing conditions. Model operationalization challenges. (Gartner, 2017)

3.2 Data science process overview, team roles, key competencies

A summary overview of standard processes, team composition and roles, and key competencies required in data science, machine learning and AI solutions development. Having a consistent and repeatable process is the foundation of any scientific discipline.

There are many existing industry-wide best practices and processes available also for the data science area so that teams are not required to reinvent the wheel. Irrespective of the approach taken, the main process for producing machine learning applications remains the same and iteratively repeats once additional data is available, learning is achieved, or the model needs higher accuracy. (Akerkar, 2019, p. 21)

The steps in typical process for producing machine learning applications:

1. Gather data
2. Prepare data
3. Split data
4. Train a model
5. Test and validate the model
6. Utilize model in operation
7. Iterate based on learning
   (Akerkar, 2019, p. 21)

CRISP-DM

Cross-Industry Standard Process for Data Mining gives a useful description of common data mining and analysis process. Emphasis on the process is on the iterative approach of firstly, exploring the data and building a well-informed data science project.

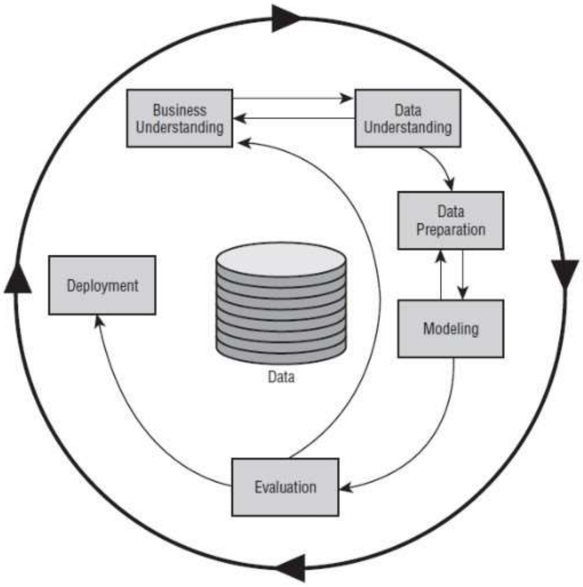


Figure 15: Cross Industry Standard Process for Data Mining (CRISP-DM) model

The process usually starts with 1. developing business understanding and business problem definition, 2. collecting data to support solving the business, 3. preparing the data by cleaning-up, conversion and removal of leaks, 4. modelling the regularities and patterns in the data, 5. results evaluation in laboratory settings to identify accuracy, causalities, and business benefit, and finally 6. deployment of the model and automation of data mining techniques in the target information system. The book ’Data science for business’ details each step of the CRISP-DM process and gives examples of practical use (Provost and Fawcett, 2013, p. 27-33).

Data-analytic thinking motto often emphasized is “You get what you measure – if you do not measure it, you have no evidence to show that it exists”. This idea promotes building data literate organizations, looking at transactions in the processes and actively evaluating the data collected of the processes. A critical skill in organizing data science work based on business and data problems is splitting the problems into smaller pieces that can be matched to known analytical tools. This set of analytical tools are addressing only a small number of fundamentally different problem-solving tasks. (Aiken and Harbour, 2017, p. 62-64)

These tasks are: 1. classification, 2. regression, 3. similarity matching, 4. clustering, 5. association grouping, 6. profiling, 7. linked item prediction, 8. dimension reduction, and 9. causal modelling (Provost and Fawcett, 2013, p. 27-33)

In the early stage designing a solution should involve structuring sub-problems which involve, i.e. performing classification, regression, dimension reduction, and so on. The design team responsibility is to consider the problem to be solved carefully, the user scenarios, and the value of the solution to the business. (Provost and Fawcett, 2013, p. 27-33)

KDD

Knowledge Discovery and Data Mining (KDD) is a field of analytics closely linked to Machine Learning. Both targets to find useful or informative patterns in data, and they share many common techniques and algorithms. For understanding the perspective and terminology used, it is worth noting some differences between the two. Machine learning contains subfields of robotics and computer vision, which are not in the scope of KDD. Also, KDD scope is limited to data mining activities and information discovery and do not concern decision making and cognitive behaviours. KDD also has a much larger emphasis on the process of data analytics: data preparation, model learning, evaluation, etc. (Provost and Fawcett, 2013, p. 40-41)

TDSP

Microsoft Team Data Science Process (TDSP) provides a structured methodology to deliver predictive analytics solution iteratively. It closely resembles CRISP-DM and KDD, and those much in common. TDSP describes development lifecycles as a key concept. They are ’Business Understanding’, ’Data Acquisition and Understanding’, ’Modelling’, and ’Deployment’, and they often repeat in the lifetime of a solution or a service. Methodology details also goals, tasks, and documentation artefacts for each stage of the lifecycle. The attached diagram provides a grid view of the tasks and related artefacts involved in each task. The tools and utilities provided by Microsoft in their DevOps environment support in defining and tracking tasks for project execution. (Microsoft.com, 2020)

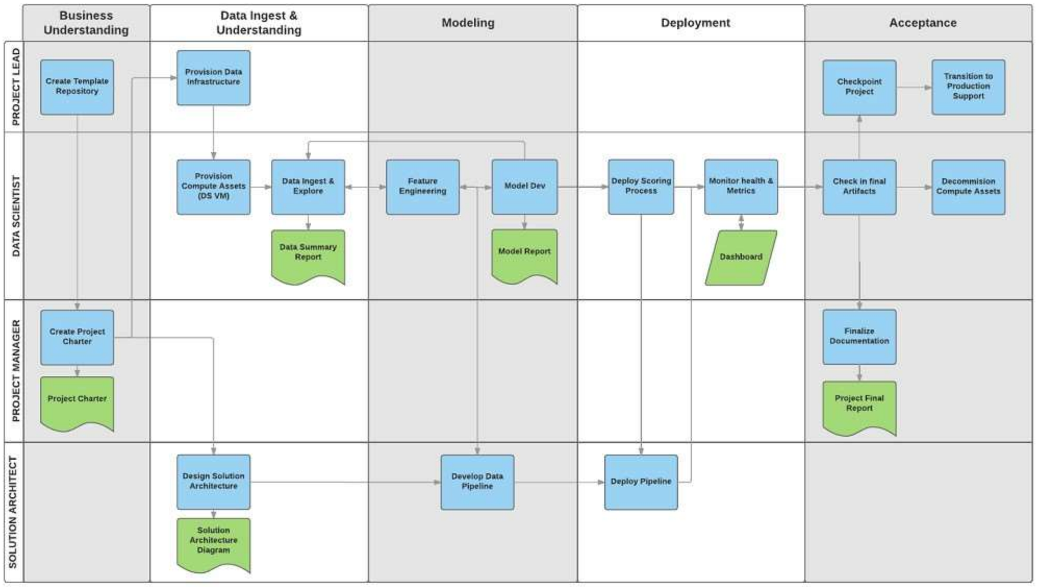


Figure 16: TDSP data science methodology lifecycle phases, tasks, and artifacts

Roles and staffing of analytical projects

CRISP-DM itself does not define a team or specific roles for executing the process. Other sources define core roles that can operate in a data science project as: Product Owner, Subject Matter Expert, Data scientist, Solution Developer. Additional roles that bring key competencies help to productize the data service: Solution Architect, Data Engineer, Process Master, Business Analyst.

Microsoft Team Data Science Process (TDSP) provides a structured set of key roles and tasks for a standardized data science process. Project leadership involves a 1. Group Manager, who manages the entire data science unit in an enterprise. 2. Team Lead, who is responsible for managing a team of several data scientists in the data science unit. 3. Project Leader is in charge of the daily activities of individual data scientists on a specific data science project. 4. Project Individual Contributors: Data Scientists, Business Analysts, Data Engineers, Solution Architects, Application Developers, and others execute the data science project. (Microsoft.com, 2020).

In most cases, you can outsource portions of the analytical process. These are the steps to consider outsourcing data acquisition, data loading, data profiling, data integration, and data visualization. The steps that carry more risk if outsourced are model selection, modelling, model tuning, hypothesis definition/testing, pilot/prototype/production, and model management, which should be tasks left for the core data science team. (Douglas B. Laney, 2020, p. 3-6)

3.3 Development of data-driven use cases in organizations

Data-driven decision making (DDD) refers to organization practices of utilizing data analysis to complement gathered experience and intuition to remove uncertainty in decision making. The benefits of DDD have been studied and demonstrated conclusively. Companies utilizing DDD are more productive (up to 6%), have a higher return on assets and return on investments, have higher assets utilization, and market value. (Provost and Fawcett, 2013, p.4-17).

Typical areas where data science helps in producing solutions fall into two categories:
i) Discovering insights and patterns in data to remove uncertainty (predicting purchase behaviour of customers, customer retention discovery, profiling, anomaly detection, etc.)
ii) Decision making is done on a large scale, where even small accuracy increase based on complementary data analysis gives large benefits in total (direct marketing, online advertising, credit scoring, financial trading, customer service management, fraud detection, search ranking, product recommendation, etc.).
Managers who understand these aspects and possess great analytical skills play a key role in leveraging data science solutions. (Provost and Fawcett, 2013, p.4-17)

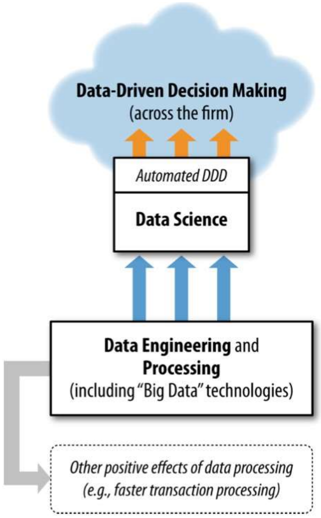

Data-Driven Decision Making
(across the firm)
Automated DDD
Data Science
Data Engineering and Processing
(including "Big Data" technologies)
Other positive effects of data processing
(e.g., faster transaction processing)

Figure 17: Data-driven decision making (DDD) supported with data science and automation

The attached diagram shows how data science support decision making, and as the level of discovery and automation increases, business decision is supported with automated decisions made by IT systems. Data engineering and processing and related literacy and governance are the fundamental backbone to support data science. Data and the capability in the organization, the expert talent acquired to extract useful insights from data should be regarded as a strategic asset in organizations. (Provost and Fawcett, 2013, p.4-17)

Approaches for learning about ML

The book ’Artificial Intelligence for Business’ states two approaches for learning about and how to apply machine learning in practice. The first approach is to start using machine learning libraries and functions as black-box services for providing different but expected functionalities. Several existing packages and visual development environments do just that and can get you easily started. With a small amount of learning, you can get your job done. The second, more robust approach is to learn the mathematical concepts, write algorithms based on those, find coefficients in data, fit the model to the business problem, find optimization points and further improve the predictive function according to business requirements. The main process for producing machine learning applications remains the same and iteratively repeats once additional data is available, learning is achieved, or the model needs higher accuracy. (Akerkar, 2019, p. 21)

Bi-modal IT operations enabling business teams with digital transformation

Gartner introduced in 2014 a term called bi-modal IT, which is described as the practice of managing two separate, well-defined modes of information systems delivery:
Mode 1 is focused on stability, is traditional and sequential, emphasising safety, accuracy, and profit on investments. These are the standard IT operations and processes already existing in organizations. Mode 2 is exploratory and nonlinear, emphasising agility and speed. Here the business involvement is significantly higher. In past years, this may have been called ’hidden IT’. Business teams need capabilities to analyse the data they own to produce digital services for automated data analysis. They utilize IT platforms and unified development patterns to explore data science problems, applications prototypes, and small-scale deployments. Mode 2 also recognizes that in other areas of your enterprise, requirements are unclear, changing, and less understood at the start.
Iterative development focused on continuous learning is required to fully utilize your Processes, People, Policies and Technology investments in digital transformation. (Gartner: Mingday, S., Scott. D., 2017)

4. Assessment of business analytics and data science maturity status in business team and development of improvement strategy

The theoretical framework and learnings of analytics maturity assessment, data strategy development, and data science team development principles were applied to analyse the organisation's current state capabilities and produce guidelines for developing analytics capabilities in the organization.

4.1 Business problems, the current state in analytics evolution, data maturity, and organization experiences

Current state assessment of analytics maturity in our customer organization MN Supply Chain Engineering (MN SCE). Later on team is referred to more generically as "our team" for purposes of readability.

Organization maturity Interviews and survey

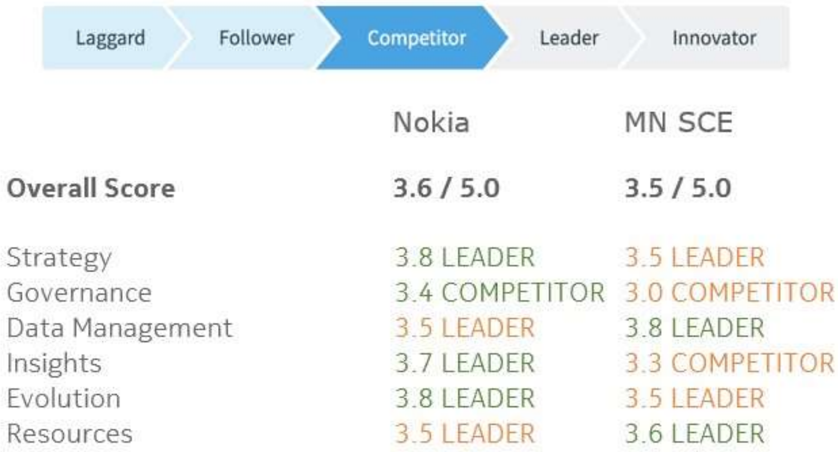
For this thesis, I conducted a survey with our organization key stakeholders to discover the current perception of analytics maturity and capabilities in our current organization. I collected results based on the arranged online survey. The survey questions and scoring were provided by Blast Analytics Maturity Assessment tool and were used as input in the thesis to evaluate the current state. Survey questions had Likert-scale evaluation for participants to indicate their level of agreement to key statements. Also, open questions about their experiences on analytics use cases benefit and improvement ideas.

Responses were collected and analysed to identify overall company analytics maturity status, overall organization unit status, strong points, weak points, inconsistencies, and differences in overall results. Response overview was summarized using a median of responses to identify the typical organization response. Additional characteristics were produced to understand the quality of responses: average of responses, interquartile range (IQR) representing dispersion, first quotient value, third quotient value. These were then used to further understanding the confidence level of responses. I then used a typical response in the Blast Analytics Maturity Tool to collect an analysis of responses with a summary of maturity level and recommended improvement actions.

Overall result of Analytics Maturity Assessment in Nokia overall and in our team

The overall Score analysed for Nokia was 3.6 out of 5.0, indicating a solid maturity level of '3 - Competitor' for the company overall. For our team, the analysed score was 3.5 out of 5.0, resulting in the same category.

This brought together the overall impression and received feedback that most of the analytics work in the organization is proceeding in a good direction, and value is being constantly delivered to the business teams from the analytics development. Break-down of the key areas evaluated in the analytics maturity assessment are described below.



Laggard → Follower → Competitor → Leader → Innovator

	Nokia	MN SCE
Overall Score	**3.6 / 5.0**	**3.5 / 5.0**
Strategy	3.8 LEADER	3.5 LEADER
Governance	3.4 COMPETITOR	3.0 COMPETITOR
Data Management	3.5 LEADER	3.8 LEADER
Insights	3.7 LEADER	3.3 COMPETITOR
Evolution	3.8 LEADER	3.5 LEADER
Resources	3.5 LEADER	3.6 LEADER

Figure 18: Analytics Maturity Assessment score in organization

The majority of the responses to the analytics maturity assessment statements were towards agreement and neutral responses, with still a large portion towards disagreement. Extreme values of full agreement and full disagreement were not used as often. This also implies that statements in the survey also included conditions that are not easy to meet in full and could be achieved only by the industry innovator organizations. There was in addition few concerns raised on how applicable some of the extensive questions were on our organization.

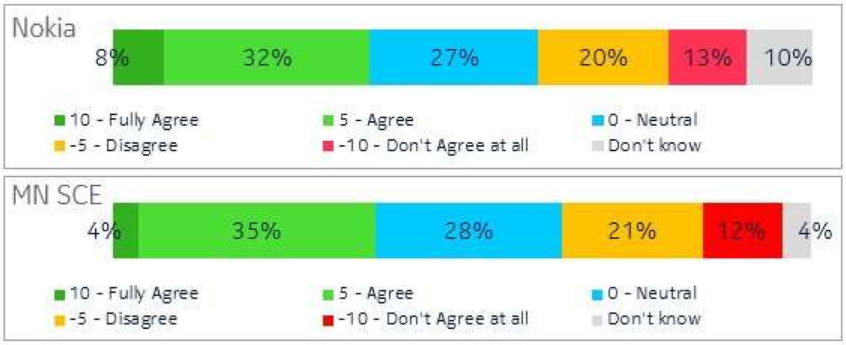


Figure 19: Analytics Maturity Assessment survey responses from organization

On Strategy, Nokia overall analytics maturity score was 3.8 LEADER, while the analytics maturity score for our team was 3.5 LEADER. Analysis summary stated that organization focus and effort is beginning to pay off. The organization has dedicated resources to help the company evolve and a vision for analytics that business leaders support.

Assessment summary proposed key steps for organization to evolve in maturity to be implemented. Organization should develop and establish analytics strategy to be viewed as a key part of business strategy. Organization should create additional business cases for advanced analytics investment to foster greater executive buy-in.

On Governance, Nokia overall analytics maturity score was 3.4 COMPETITOR, while the analytics maturity score for our team was 3.0 COMPETITOR. Analysis summary concluded that along with senior-level sponsorship and adequate funding for organization analytics efforts, the organization had dedicated resources to help the company to evolve and a vision for analytics that business leaders support. Organization has widespread adoption and focus on business analytics. Organization produces analysis that is now focused on what will happen instead of looking back.

Assessment summary proposed following key steps for organization to evolve. Organization should setup a steering committee with cross functional representation. Organization should have documented policies and processes to drive governance. Organization should engage with an executive-level sponsor to ensure analytics adoption across all necessary departments, and to form business working groups. Organization should establish security and privacy governance-related processes.

On Data Management, Nokia overall analytics maturity score was 3.5 LEADER, while the analytics maturity score for our team was 3.8 LEADER. Analysis summary stated that the organization is on its way to developing a cohesive data ecosystem that accurately provides the answers to the business. Organization has dedicated resources to help the company evolve in analytics. Organization has formulated a vision for analytics that business leaders support. Organization has established defined architectural patterns to increase levels of data control and ownership, and data warehousing in place and is being used regularly.

Assessment summary proposed following key steps for organization to evolve. Organization should continue collecting increasingly complex data (e.g. event, time series, social, etc.). Data collection quality monitoring should be in place. Organization should focus on increasing compliance and establish standards for managing the 4 Vs (Volume, Velocity, Variety, Veracity) of Big Data to enable full data control and ownership. Organization should also implement self-service data preparation processes.

On Insights, Nokia overall analytics maturity score was 3.7 LEADER, while the analytics maturity score for our team was 3.3 COMPETITOR. Analysis summary concluded that organization is starting to see how insights can clearly guide what business should, or shouldn’t, do. Organization has proactively delivered reporting with context and recommendations for action. Organization has started to understand customer journeys through analytics.

Assessment summary proposed following key steps for organization to evolve. Organization should develop customer personas and segments. Organization should establish a regular cadence of analysis and insights. Organization should initiate a testing and personalization program to take advantage of new analytics data and insights

On Evolution, Nokia overall analytics maturity score was 3.8 LEADER, while the analytics maturity score for our team was 3.5 LEADER.
Analysis summary stated that organization investments in evolution, future and sustainable success are clearly visible. The organization shows the following characteristics are in place: dedicated resources to help the company evolve, a vision for analytics that business leaders support, and alignment around data and analytics.

Assessment summary proposed following key steps for organization to evolve. Organization should setup analytics related KPIs and performance measured to manage maturity. Individual contributors in organization should be encouraged to act on available data to improve utilization of analytics investments.

On Resources Nokia overall analytics maturity score was 3.5 LEADER, while the analytics maturity score for our team was 3.6 LEADER. Analysis summary stated organization analytics strategy had gained considerable momentum with increased amount of people on board. Organization includes business analysts and a dedicated program manager who are participating in intermediate training led by external resources.

Assessment summary proposed following key steps for organization to evolve. Organization analytics team should include next also a single statistician and/or data scientists. Organization should appoint an internal analytics leader and develop a small supporting team as centre of excellence. Organization should begin to administer advanced analytics training to continue building internal expertise.

Survey responses summary

Survey responses for our team were collected in total from 18 stakeholders. 64% of respondents were decision-makers and analytics users, 36% of respondents, were analytics developers. Response ratings were overall positive and in agreement with Strategy, Governance, Data management, Insights, and Evolution statements. One exception being Resourcing area, where responses to statements were only neutral. The survey contained 100 questions or statements about analytics capabilities. In total, there were 1237 replies evaluated, and 59 replies with either ’Don’t know’ or empty were excluded. I calculated the median response to give the organization a typical response score. There were no statements in the survey where the median response was in either extreme of ’Fully agreed’ or ’1 - Don’t Agree at all’. 73 of statements had a median response of ’Agreed’. 17 statements had a median response as ’Neutral’. And 1 statement had a median response of ’Disagree’. Review of analysis outcomes and recommendations to evolve confirmed the majority opinion in the organization and overall positive impression of the analytics maturity progress.

Analytics maturity considerations for MN SCE

This survey outcome can be used as an initial guideline for data and analytics strategy development. However, for more concrete data and analytics strategy implementation, it is advised to further extend the analysis to concrete competence and technology areas with the help of an external consultant company. With the overall maturity score and summary provided, it is advised to evaluate the proposed key steps to evolve with analytics maturity in more detail and to plan concrete actions in analytics teams.

Regarding the individual statements, strong points, weak points, stress points and differentiators, it is advised to discuss within the teams what measures can be taken to improve on weak points, to align the organization by training to avoid conflicting opinions, and to evaluate what is the impact of differentiators identified to overall Nokia responses.

Organization experiences in AI/ML project development and implementation

In the interview and review feedback, the impression of AI/ML development projects was that they take longer to implement and may be subject to project cancellation if the business benefit is not evident in the short term. Within the area of MN product maintenance, and especially the sub-processes that define product failure rate at a customer, processes are typically complex and not well interlocked to produce high-quality datasets. This produces challenges in developing AI/ML solutions in the area. A large amount of time is spent in data maintenance, investigating the processes, and solution needs to cover a wide area of subjects to produce business benefits. It is important to define development programs, and solution platforms to be continuous and split the problem area into smaller pieces to be iteratively developed.

The importance of having high-quality data in the area of structured data collection could not be emphasised more, as that is one of the highest cost forms of data and is also a key prerequisite for developing AI/ML solutions. Uniform and standardized data pipelines for consolidating raw datasets for use in reporting and analytics have been in Nokia digital transformation objectives on the enterprise IT level since 2015. Nokia Enterprise Data Platform now produces over 100 key information sets from 9 data domain areas as API based datasets. Areas include Product master data, Customer master data, Common reference data (calendar, geo locations), Supplier Master data, CRM data, Procurement data, Finance and control data, Customer transaction data, Supply Chain data. Solution projects can use those datasets after access approval, data stewards of each domain area maintain data, and sets have a high level of data security measures to ensure credential lineage up to reporting layer.

4.2 Identification of recommended approach for successful business analytics organization and development projects implementation

How are business analytics and data science related to business strategy? What makes a data science project successful, e.g. what are the requirements to deploy predictive models to operative use in business processes? What are the elements which make a business problem suitable for data science?

Analytics governance and organization

Data analytics governance’s key objective is to drive business value and reduce risks in investments and analytics quality. According to Blast Analytics Maturity Assessment, analytics data governance is an ongoing initiative and oversight over the goals, communication, policies, processes, metrics, and data management to operationalize analytics and bring up competitive advantage from analytics insights. (Blast, 2020)

In more detail, the objective is to (Grossman, 2018, p. 2):

- Develop and disseminate data strategies, goals, policies, standards, processes, and KPIs.
- Establish roles, set expectations, and enforce accountability.
- Manage and increase analytics maturity.
- Plan, sponsor and oversee analytics projects.
- Manage data quality and resolve data-related issues with data dictionary, data layer, workflows for your tag management, automated QA monitoring, disaster recovery plan
- Promote the value of data and vision for analytics.
- Manage data mining models and tools.
- Provide training and best practices for decision-making.
- Share insights and results across the organization.
- Support self-service analytics capabilities for business users.

Improving data analytics maturity often takes years and must always be maintained through organizational changes. The steering committee, which needs to oversee analytics strategy and governance oversight, requires stakeholders with strong influence and decision-making power from the various functional areas and business units since analytics data governance involves large cultural changes.

Speeding up business problem decomposition to analytical solution

The literature review proposed that organizational mindset is to be shifted towards data analytic thinking: “you get what you measure – if you do not measure it, you have no evidence to show that it exists”. Demonstrating evidence is also critical for communicating change and progress. (Aiken and Harbour, 2017, p. 62-64)

This statement gives the emphasis and motivation towards building data literate organizations. The organization should define and understand what information they will need to accumulate so that analytics is possible. People in business teams should be looking at transactions in the processes and actively evaluate the processes’ data. One inherent issue of typical organization structures results from having siloed departments of business and support service teams. Data resides in IT departments, and businesspeople rarely have access to the actual data. Well managed data governance ensures data analysts working in business departments have a clear and straight forward way of accessing the business data and producing analytical solutions.

A critical skill in organizing data science work based on business and data problems is dividing the business problems into smaller pieces that match known analytical tools. This set of analytical tools are addressing only a small number of fundamentally different but discreet problem-solving tasks.

These tasks are worth recognizing and defining clearly. A list of the common tasks was provided in the book ’Data science for Business’ by (Provost and Fawcett, 2013, p. 19-23): i) classification, ii) regression, iii) similarity matching, iv) clustering, v) association grouping, vi) profiling, vii) linked item prediction, viii) dimension reduction and ix) causal modelling. In this early stage designing a solution should involve structuring subproblems which involve, i.e. performing classification, regression, dimension reduction, and so on. The design team responsibility is to carefully consider the problem to be solved and the user scenarios.

A framework to be used as an organizational tool and in analytical engineering called Expected Value Evaluation was introduced in ’Data science for Business’ by Provost and Fawcett (2013). Such methods assist in decomposing the business problem into smaller parts. Expected Value evaluation also demonstrates how the analytical outputs can be utilized and show what can be calculated based on the available data.

Project selection for AI/ML solution development

There has been several techniques and tools developed for selecting projects for AI/ML solution development. One such practical tool for describing data problems and requirements is the ’Analytics investigation pyramid’ introduced by Dr. Roy from PA Consulting.

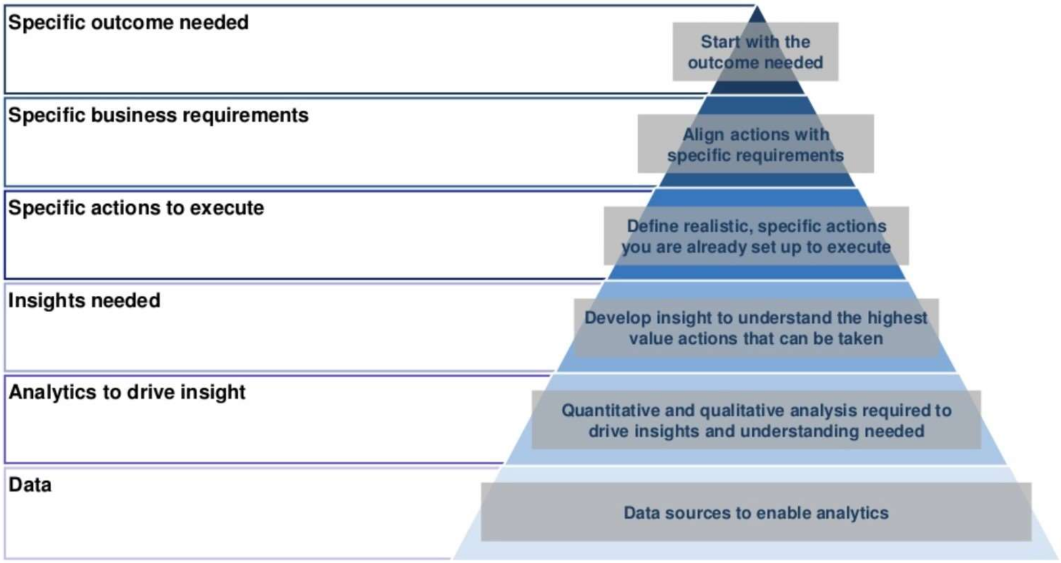
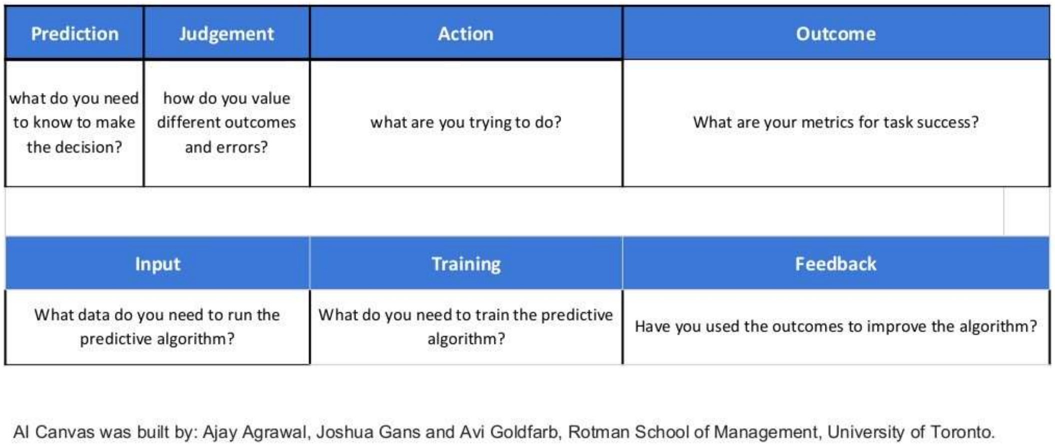

Specific outcome needed
Start with the outcome needed
Specific business requirements
Align actions with specific requirements
Specific actions to execute
Define realistic, specific actions you are already set up to execute
Insights needed
Develop insight to understand the highest value actions that can be taken
Analytics to drive insight
Quantitative and qualitative analysis required to drive insights and understanding needed
Data
Data sources to enable analytics

Figure 20: Analytics investigation pyramid template (adjusted from Dr. Roy, PA Consulting s.a.)

The descriptions of a use case for analytics in ’Analytics investigation pyramid’ includes:

- Outcome: ultimate business problem or goal where an analytical solution is required
- Business requirements: Description of detailed use case and actions for aligning business intelligence objective to corporate and organization strategy
- Actions to execute: Decision actions in the current way of working which need to be taken by named teams/individuals, considering existing constraints
- Insights: What is specifically needed to know to decide on action towards a specific outcome
- Analytics: What analysis activities are required to be performed
- Data: Which data sources (structured, unstructured, internal, external) need to be evaluated to perform the analysis, and what constraints are known for the data (quality, legal, cost).

A similar tool was introduced by the University of Toronto’s Rotman School of Management: the AI Canvas.



Prediction	Judgement	Action	Outcome
what do you need to know to make the decision?	how do you value different outcomes and errors?	what are you trying to do?	What are your metrics for task success?
Input		Training	Feedback
What data do you need to run the predictive algorithm?		What do you need to train the predictive algorithm?	Have you used the outcomes to improve the algorithm?

AI Canvas was built by: Ajay Agrawal, Joshua Gans and Avi Goldfarb, Rotman School of Management, University of Toronto.

Figure 21: AI Canvas (adjusted from Rotman School of Management s.a.)

These tools were evaluated by the team and agreed to be piloted in the next projects to define a structure how to collect and document AI/ML ideas, project proposals and use cases.

User experience (UX) impact on data analytics success

A clear step in deploying analytical solutions to business teams is the usability and user experience of the solution. To create a solid user base and conversion of investments to business benefit, the users need to be able and enjoy the process of navigating the analytics landscape. This is defined by user experience (UX), which balances four important elements: design, technology, user needs and business goals. Aligning and integrating these elements leads to optimum investment conversion. UX is also a data-driven process utilizing testing methods (e.g. PSSUQ, Post-Study System Usability Questionnaire), digital analytics, persuasion architecture, and user feedback to better understand users’ needs.

Confirmed also with our end user surveys, user needs to be provided with modern analytical visualization tools and a more personalized online experience to increase business benefit from analytics investments.

Logically structured content strategy, information architecture, functionalities, user interfaces and accessibility are relatively easy to implement and can be developed in small increments.

The current state of the organization shows good progress and opportunity for success. The organization has experience in utilizing several tools and techniques for usability evaluation during solution development. The organization has collected use case scenarios, process flows, user stories and information on constraints and requirements to help guide the solution design. At deployment, User Acceptance Testing (UAT), open interviews and mid-term piloting are often used to ensure efficient system and process operation. After solution deployment, usually open discussions, service backlog tracking and performance measurement produce a starting point for development. Surveys for user experience evaluation and interviews based on open discussion have been utilized in some of the projects in the analytics area to ensure efficient information discovery and user acceptance. More advanced and detailed user experience measurements could produce standardized data for usability and more tangible results of how an organization uses analytical solutions and their front end. It is recommended for the organization to strengthen the analytical and data-driven mindset and culture with additional knowledge sharing and time invested in developing teams to be fluent in discussing business data, analytical solutions, and usability heuristics.

Training for managers and importance of Analytics Centre of Excellence

Managing AI/ML applications development requires that the basic concepts in mathematics and statistics are understood well. Organization development is not a short-term activity but often requires long term effort of even years to change the capabilities and behaviours in the organization.

One of the fundamental principles of agile development is to improve the organizational knowledge base continuously. Training and documentation are the essential tools for growing an organization-wide knowledge base. It is advised to set up this knowledge base as a well-formulated Analytics Centre of Excellence (CoE) for the organization. This department or function is in the lead position to produce alignment, to build the internal analytics community and to ensure education on all organization levels. Training should cover data understanding, reporting and dashboards development, data storytelling, user experience, and data governance. Having the Analytics Centre of Excellence (CoE) ensure accessibility of documented policies, best practices, training, processes, tools, and other resources will ensure a solid organisational knowledge base is in place.

In one of the interviews and review feedback, the key challenge recognized for the organization was in having business teams accumulate sufficient competence in the analytics field and having concrete experience from AI/ML projects. This knowledge development approach was also seen as the only reliable way to mitigate the typical pitfalls in advanced analytics solutions development and integration of solutions to organizations processes. Similarly, this phenomenon was visible when Nokia implemented 6 Sigma processes and basic statistics skills in the organization. After the initial hype, the actual transformation was taking several years. To mitigate this issue, Nokia has been running an initiative for analytics training continually since 2016.

Continuous learning and development using the continuous improvement principles give a method for implementing the organization vision of capabilities. There are few key success factors in applying continuous improvement, as described in the book ’How to succeed with continuous improvement’ by Ahlstrom (2014).

Successful organizations, based on the book, apply these five principles for solution development:

Focus:
An organization should have a clear vision for solution development and the purpose the solution serves, and clear targets to achieve those targets. Applying techniques to divide, break down, prioritize, simplify, and discard, the solution will be developed according to critical features, which end users really need. Thinking inside the box was found to be the most productive and innovative method - tasks and obstacles are limited and clarified. (Ahlstrom, 2014, p. 46-54)

Positive visualization:
Solution vision and purpose should be forged to the organization mindset to understand the positive benefits of the solution and atmosphere during the transition. Successful organizations can treasure their problems as learning opportunities. Measuring your opportunities and current state enables you to visualize your organization targets and standards to check your progress constantly. Using tools like improvement boards, improvement

comms meetings, and improvement newsletters are an effective way of focusing on development priorities and the organization informed about the progress. Peer support and coaching from other experts and leaders in other projects and teams can improve the development atmosphere, drive, and confidences. Tools, for example, include improvement voting, rewards and prizes, and demonstration visits of best improvements done outside of the company. (Ahlstrom, 2014, p. 46-54)

Simplicity:

Solutions developed should be simple so that everyone in the organization can and wants to use them. Simple techniques like fishbone diagrams, 5 Why's, among others, can be used to find good solutions to difficult problems and enable everyone in the team to contribute. To avoid the development culture being framed to solving only problems and focusing on bad behaviours, it is most useful to measure and analyse the root causes of success. Questions to start should involve examining, e.g. 'Why was the best team able to perform their activity in a shorter time than others?', 'What contributed to the customer or end-user satisfaction of the best group?'. This enables identifying key behaviours and patterns on the scale instead of applying only short-term fixes in specific areas. With a focus on success, it is possible to develop solutions that are easy to do things right the first time. People with high ambitions clarify their expectations and produce insurances that it's difficult for this to go wrong. Occasionally, it's not possible to make the right behaviour easy. In such conditions, the book states that 'fun doing right' is an excellent substitute for the 'easy to do right'. (Ahlstrom, 2014, p. 46-54)

Ownership:

Solution end-users should have a clear role and transition to develop their own working practices, rather than having outside factors trying to change them. Often traditional organizations learn the pattern of having leaders tell their employees or colleagues what to do, counting that leaders have the most information on the subject and are willingly sharing their top knowledge. Successful organizations adopt trust and responsibility-based approach. Leaders ask for methods for achieving expected results, fact-based reports on the current state, identifying reasons behind high priority problem areas, and updates on improvement activities. This enables the development team to think about overcoming the project challenges and the organization to think on how to leverage the capabilities available. (Ahlstrom, 2014, p. 46-54)

Systematic approach:
An organization should have clear rules on how to participate in development so that everyone can be expected and have a fair chance to contribute to the joint targets. This echoes through the organization by ensuring not only the individuals or teams are creating success, but the entire organization is also participating in joint success. (Ahlstrom, 2014, p. 46-54)

Nokia has designed a specific Data Science learning path in the company online training portal for all Nokia employees who want to improve their Data Science skills. This learning collection is structured into specific areas of Data Science competence, including Math and Statistics, Computer Programming and Domain Knowledge, Data storytelling and communication. Some of the courses are internally developed expert training, and some are done in collaboration with external online training platform vendors. Offering now consists of 33 individual courses and 3 programs that aim for certifications on foundation and specialization levels.

5. Application of machine learning enabled IT architecture for business analytics in Nokia MN

The theoretical framework learnings of data science and machine learning principles were applied in the solution development program to build team capabilities and an IT platform that supports iterative data analytics and machine learning solution development.
The focus of the thesis is on analytics platform development, development methodologies and key tools used in development activities, instead of in-depth documentation of data science principles and software implementation. A key motivation behind this scope is business administration and business analytics specialization instead of the data science field.

5.1 Prediction system development in Nokia MN Field Reliability Projects

The 'MN Field Reliability improvement by advanced machine learning (FiRe)' program aims to produce analytics and automation solutions for the supply chain of 4G and next generation 5G radio products to find information that promotes the prediction and reduction of product field failures. The deliverable objective is to identify statistically significant variables and correlation patterns on Field Failure data that can be detected early on in the product lifecycle and create probability predictions of field failure using advanced analytics and ML algorithms.

MN FiRe program and solution development is currently comprised of two larger project iterations. An overview of the key activities from the solution development is summarized in the attached figure. In the first project started in May of 2019, we developed an architecture and initial prototypes on the iterative analytics platform to investigate manufacturing test correlation to product failures and returns. The second project started in February 2020, focused on analysing patterns in returns data and similarity of product characteristics to predict if a high probability of failure is expected for a product in the near future.

The first project team consisted of a Project manager, Product technical specialists and operative business team members, and an ML solution development team of two persons. My role in the solution development team was to lead IT platform development and architecture implementation, and to evaluate the ML algorithm prototypes with the support of a summer trainee assisting in IT & ML development, and with the company Data scientist team who were able to provide consultation more detailed insights on different phases of the data science process.

The second project involved a larger team with an enhanced focus on advanced machine learning concepts and business process development. A data scientist team of 3 persons from AI LAB were included in the project to focus on machine learning algorithm development. Additional resources were also introduced to data engineering to cover the analysis of different data sources, data quality, and the operative data flow for the solution. Business process development involved now 4 persons provide an in-depth understanding of the different processes affecting the prediction environment.

5.2 Development of business process, analysis objectives, development of analytics hypothesis and predictive model

Overview of the activities from the MN FiRe program and iteration projects is explained in the next chapters in more detail. Overview will cover the development of business area knowledge, operative business process, data analysis objectives, development of analytics hypothesis and predictive model, and integration towards the operative business process.

5.2.1 Developing Business understanding and data understanding

The initial project started by framing the business problem, studying the key concepts for predictive solution development in the chosen field, and executing a feasibility study. Key areas in the feasibility study were investigating what would be the scope of the business problem and benefit of the business solution, what would be the available development methods in developing a technical solution to the business problem, what are the organizational and IT capabilities at the current to support the work, who are the key stakeholders for the solution, and who would be organization resources available for project execution.

5.2.2 Prediction model prototyping in business with 80-20 data science rule

As discussed in the literature study, the application of data science requires expertise. Consequently, to develop strategies and development initiatives involving data science, it is mandatory to acquire a proper knowledge level on those concepts. The data science field being vast, it needs to be accepted in business development conditions that it is sometimes more useful to start applying the available knowledge and to initiate a process of moving towards the end goal, instead of waiting until expert knowledge on the entire field of science can be obtained and route to end goal is fully evident and clear. In the early stages of the process, organizations can collect practical experience and further learnings in testing the limits of their capabilities.

One approach we agreed to utilize in the first project, and which has proved to give excellent experience and results in many complex scenarios and can be successfully applied to the field of data science as well is the 80-20 Pareto rule. As explained by Nate Silver (2012), understanding 20% of data science and executing that properly for 80% of the time would give organizations the benefit of understanding your business data and the opportunity to produce experiences in prediction model prototyping. This was demonstrated in our program by first having experts with mathematics, statistics, and data engineering background investigate the fundamentals of data science, by executing own learning projects, and then consulting the algorithm principles with the available data science teams.

This first 20% of the data science pareto principle should begin with producing the right quality of data, using the right technology, and having the right objective. In prototyping, the ideal is not on the most pleasing or most convenient prediction but accurate and objective outcomes. This applied with few heuristic rules, a systematic approach, and a review of experienced experts in the field will set the organization in speed with the competition. Productizing concepts that have been identified as useful in the business would be a good basis for setting up fast and cost-efficient AI/ML implementation project. (Silver, 2012, p. 313-315)

The main process for producing machine learning applications iterates through key steps to produce a model and repeats once additional data is available, more learning is achieved, or it is decided that the model needs higher accuracy.
The high-level process explained in background material (Akerkar, 2019, p. 21) was applied in our program activities in following way.

In the ’Gather data’ step, we collected data samples from the source systems as local files and examined the data for consistency and quality, input variables, and variation in the data.

In the ’Prepare data’ step, we performed basic clean-up operations to remove records with bad information. Null values in datasets were converted to data type values of zeros or empty strings and produced feature representation pivoting the data of input variables from rows to columns.

’Split data’ step involved choosing partition for a dataset for training, algorithm testing, and one set reserved to validate predictions. Random selection was used with a 60-20-20 split of a dataset to examine what type of initial accuracy the data would hold.

’Model training’ step consisted of configuring the algorithm with chosen target variable and prediction features. This produced a report where algorithm logic was possible to be inspected.

In the ’Test and validate model’ step, we produced several iterations of tests to evaluate how executing the model would perform with a different selection of data.

Next step of ’Utilize model in operation’ meant prototyping that we produced a recommendation towards the feasibility study showing the algorithm capabilities and initial results. This demonstrated both the team and organization capability for advancing in the solution development and the technological capabilities available for implementing the solution.

The final step, ’Iterate based on learning’, was to define the next iteration objectives and to expand the model.

A more technical description of these activities is explained in the next chapters.

Technology overview and development phases in FiRe program

The technology we chose for data analysis and prototyping is the Alteryx analytics platform. This enables code-free tooling, fast workflow creation with good automated visual documentation and out-of-the-box quality assurance tools. The initial concept for our predictive analytics use case was created within 6 weeks to validate data availability, project feasibility and to produce an overview of the application of our use case for predictive ML analysis.

In the first phase, we focused solely on the Alteryx tool to evaluate the available dataset samples.

Second phase introduced a more robust data pipeline to connect to source data systems, data storage to Azure cloud SQL database, separate workflows for data handling, statistics and machine learning, and data visualizations using PowerBI.

In the third phase, we developed the architecture towards operative machine learning solution, where data pipeline would connect to source data systems, data processing would be done with Alteryx, storage of data would be secured in Azure cloud data platform, and operative machine learning processing would be done using Python in Databricks computation cluster. Visualization of prediction reports and data quality reports would be again done using PowerBI. This architecture also enabled prediction score serving towards business applications via containerized and secure web APIs.

Overview of the different phases, their timeline, and the technical components applied at each phase to scale the solution are demonstrated in the attached diagram

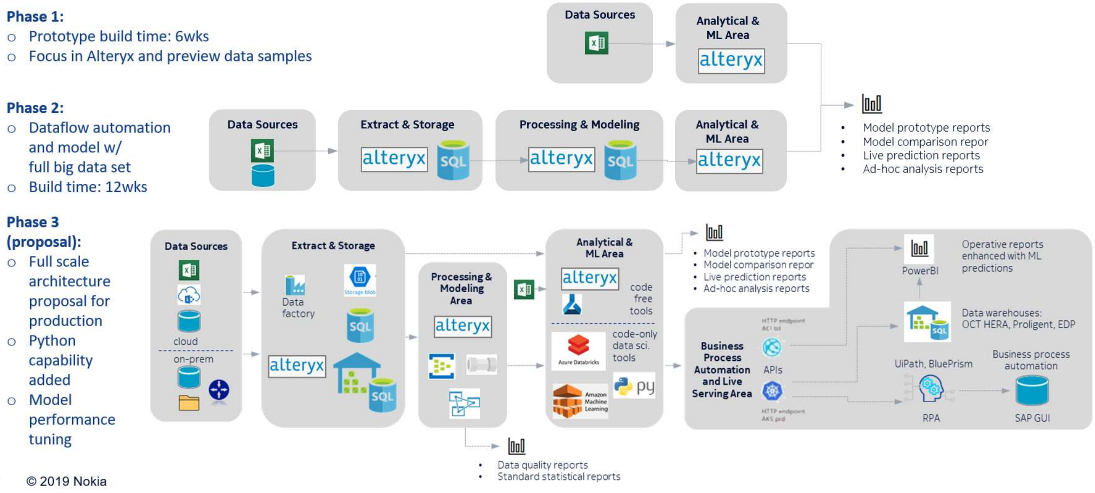


Figure 22: Alteryx utilization in iterative analytics architecture

5.2.3 Prediction model fine-tuning and data quality assurance with data science team

Our organization has a beneficial collaboration with the company supply chain analytics community, where data science expertise and support are available. We approached the community with our analytics idea and initial results during our second phase of the project. We could secure a data scientist with high-level academic education and long-term experience in multi-variate analysis to join our project iteration work.

He was then tasked to define the key concepts for ensuring data quality and fine-tuning the selected machine learning model to the appropriate level of accuracy. Data quality concepts were defined to ensure datasets used for machine learning were of high quality, have a sufficient and identified amount of variation and dissimilarity between data points, and are stored and selected in a manner that the algorithms produce systematic results. Evaluation of possible issues affecting results was also evaluated to avoid measurement bias, sample bias, exclusion bias, recall bias, confirmation bias.

Regarding model fine-tuning, we evaluated methods for ensuring expected values are produced and metrics for evaluating the model's performance. Concepts were then communicated and trained for our project team so that our summer trainee and I could define the concepts in our software workflow and produce the results for evaluation within our project team.

The main concepts for evaluating classifier models are described in detail in a book called 'Data science for business (Provost and Fawcett, 2013). Book describes fundamental concepts for data similarity and clustering analysis. A common measure for the similarity of data points is the variance, which can be visualized as a vector presentation of the distance between two data points. This measure of dissimilarity also means variance and information gain in data. The model performance evaluation methods in the book are covered from plain accuracy, use of confusion matrix, defining expected value, ranking of models with profit curves, ROC graphs, AUC statistic (Area under ROC curve), and using Bayesian rules for evaluating model prediction accuracy with probabilistic reasoning and tests for conditional independence. (Provost and Fawcett, 2013, p. 141-249)

Data evaluation using variance, Pearson correlation, and application of PCA (Primary Component Analysis) and iterative MDS (Multidimensional Scaling) methods were demonstrated in the project iteration. Evaluation of the model performance was produced by using confusion matrices, variable importance calculations, and model rankings using both numerical recall/precision rankings and also graphical views of ROC and AUC. More details on those methods and techniques are in the next chapter discussing technical aspects of the system.

5.2.4 Business process development for enabling use of ML system in operation (Service Design, Use case scenarios, Process workflows)

Preparation of the business process to utilize new data services is a critical activity to ensure development success. The program evaluated the current state of the decision-making process where prediction information was planned to be used. To produce a new version of the process, the development team utilized design techniques and tools to clarify the future business process and solution features and their operation.

Business process requirements were collected from key stakeholders using use case scenarios, service blueprint, problem statements including 5 why and fishbone tools, process flow documentation and value stream mapping, and finally, user story descriptions and service blueprint designs. These process documents were then discussed with the business organization to understand how the decision-making work of the end-users is done and how prediction information would need to be delivered when information is needed, and the required changes in roles and decision process content to utilize the predictions. My responsibilities included designing the solution IT architecture and deriving the IT requirements from available business requirements, process flows and feedback documentation.

For the first FiRe project, the front end designed for prediction end-users was designed to inspect analytics outcomes. The solution produced identification of risk pattern in the product, which could potentially cause a failure within a short timeframe of product operation in the customer environment. The business process was utilizing multiple sources of information, and it was proposed that prediction would be additional information available for the end-users. I produced a visual dashboard to inform the product maintenance owners about possible product quality issues. Dashboard summarized the overall supply chain status, including key figures of product units delivered vs units returned, prediction score and probability for the product failure, and a large set of statistics from the production test measurements.

In the second FiRe project in 2020, the user operation was designed to allow additional control parameters to be provided for the prediction algorithm execution. A JSON file containing the parameters can be stored in the data environment by a key user before launching the algorithm. Similarly, a visual dashboard in PowerBI was produced to inform the end-users about possible product quality issues. The dashboard summarizes the risk alarms based on prediction scores, shows the key drivers behind the prediction, and summarizes the overall supply chain status. An extra feature to enhance the end user experience of utilizing the analytics results was also introduced. The tabular view developed in PowerBI about prediction and product information helps the end-user easily extract the details and query additional information from other systems.

Each project iteration was regularly evaluated with the end-user group to confirm that development is going in the correct direction regarding the expected features and user experience.

5.2.5 Prediction model deployment for pilot use and business validation

The theoretical framework introduced an overview of evaluating how predictions apply in the real world. Further practical tools for evaluating the business problem are discussed in more detail in the book ’Data science for Business’. These include expected value evaluation framework, comparative baseline performance analysis, analytical engineering, selection bias evaluation, and implications for investments consideration (Provost and Fawcett, 2013, p. 187-208; p. 279-290).

During the evaluation of the first FiRe project, it was decided that the prediction model would not be taken into operative use since signals produced by the model were seen still relatively weak by end-users, and the provided key drivers were coming from a relatively complex model. The second FiRe project iteration was then planned based on this learning to have a different approach to calculating a prediction or risk pattern and having a key objective in producing supervised algorithms, which can be easily communicated and understood.

The pilot phase for deploying the outcomes of the second FiRe project was agreed upon after successful proof-of-concept evaluation with the business stakeholders and implementation of the scalable cloud version of the solution. The key objective in the pilot is to verify the operation of the planned business process and validate that impact to the operative KPIs can be measured after the proactive identification and investigation of quality problems before customers escalations. Metrics and practical aspects are under planning at the time of writing this report.

5.3 Development of technologies, key capabilities, competencies and MLOps architecture

This section describes the activities from the MN FiRe projects from a more technical perspective, addressing the development of technology platforms, organizational competencies, solution capabilities, iterative machine learning architecture development, and DevOps pipeline produced for data and machine learning service operation and maintenance.

5.3.1 Building understanding of key concepts, & evaluation of technologies, Proof-of-Technology, Project launch

Key objectives in the early stages of the project were in two areas. The first area was evaluating and selecting ML platforms and tools, enabling iterative development, the building of experience, fast development cycles, and good integration to existing data architecture at Nokia. For comparison of ML platforms, I evaluated code-free tools: Weka, Azure ML Studio (Classic), Rapid Miner, Alteryx, and code-only tools enabling Python development: Azure ML Studio Notebook (Classic), Jupyter Notebook.

The key benefits and weak points of each are explained in the theoretical framework as a summary from Gartner research. A team of developers were involved in evaluating each tool also in practice, and we discussed findings in the analytics development community to conclude the best options.

Alteryx was our chosen tool for the program, since it provided us with a solution to key criteria seen as critical to data science competence development and efficient execution. One of top priorities was to have single user interface and environment to enable business with accelerated advanced analytics development instead of large-scale service-based environment with multiple different types of tools and different user interfaces. Secondly there was a priority to have a tool enabling shared data workflows to improve joint development activities. Also, the chosen tool should have a low threshold to ML / AI model development, clear and easy to use components, and optimally interface developed with business team usage in mind instead of code-based IT team users. Tool should also enable fast concepts and data workflow creation and enhance communication within the team with visual and automated workflow documentation.

Benefits identified by evaluation team in the organization:
i) Team had quick adaption of the tool, user interface is intuitive and guides the work sufficiently, there is low requirement on technical knowledge, and tool enables fast speed to market with the code-free development option.
ii) Having tool with out-of-the-box features for basic data sourcing, quality assurance, cleansing and transformation is a quick win in rapid iterative development, and in quality assurance to verify workflow operation and correct results.
iii) Choosing a tool with ML engine based on the R programming language was seen as beneficial option since it is widely utilized in the data science field and documentation is widely available.
iv) program overall cost was reduced since there was no need for external programmer resources in our ML project and tool provides also self-service BI & ML capabilities for team.
v) Tool demonstrated good re-use of outputs, it integrates and interfaces our other platform services like Azure SQL, Power BI, and Azure Databricks.
vi) Tool provided good transparency to features, clear and visual documentation of analytic workflow steps. This enabled fast code-review and knowledge transfer in the team and enabled quick adoption of workflows and models to different platforms.

After selecting the Alteryx tool, my task was to produce a concept for proof-of-technology with the chosen platform. After procurement of licenses and training received of the tool features, it was possible to create proof-of-technology quickly from the available dataset.

Within only 6 weeks, we were able to procure and set up the development environment and validate our prediction model feasibility and business use case for ML-based analytics development. This is a considerably short time in a large enterprise to introduce new platforms for practical use.

5.3.2 Architecture implementation for iterative analytics development and prototype development with Alteryx

During the next Iteration 2 executed in a 3-month timeframe, I developed with the support of our summer trainee the prototype using full-scale datasets for 2 products to create a predictive ML model. My main role was to design the architecture, create the concept for the first product, document the development steps and enable summer trainee to replicate the development for the second product.

We created an automated data workflow to collect lifetime production volume data (up to 300k product units) of product system testing measurements (up to 28 million records), customer returns data (up to 6 000 cases) with identification of product failures, transformed the datasets to features (up to 431) to be utilized in ML analysis, and created a data model to blend the datasets.

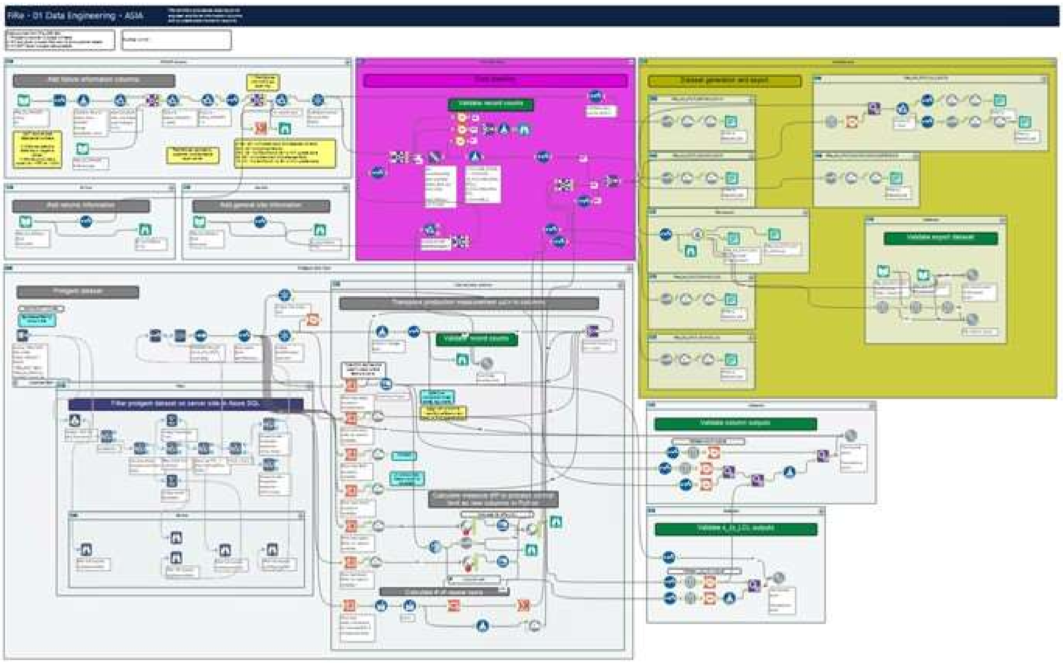


Figure 23: View of Alteryx workflow to clean-up data, to blend datasets, and to transform data to ML model features

Impact of Alteryx towards development cycle speed

Alteryx impact on the development speed and resources spent during the development of the prototype analytical model was significant. 3 months enabled the team to evaluate the data in detail, produce data processing workflows to improve data quality significantly, evaluate both statistical and predictive capability with the data at hand, and compare multiple different predictive models to evaluate their performance to produce accurate prediction outcomes.

We collected several learnings and identified success factors in the program. Fast design and implementation of data workflows were achieved using out-of-the-box workflow elements inside one desktop development environment. Complex data engineering tasks were possible to be taken under work by the project team immediately in the code-free environment, without the need for configuring and patching together code snippets or going through to read extensive documentation libraries about code parameters and module interoperability issues. In later releases Alteryx also provided automated ML workflow tools to further speed up the development of prototypes and concepts for team evaluation.

Excellent outcomes were achieved finally by using hybrid development with code-free workflow elements and code-based programming of specific functions. Having experienced programming skills inside the development team gave confidence that a large team of external 3rd party programmers will not be required to create the data and business understanding, evaluation of the prediction model system, and prototype of the solution.

Rapid response capability on questions about workflow details or taking requests to change workflow setup according to new requirements was constantly supported with the visual documentation capability of the tool. Anyone in the project team was able to follow each chain and elements of the complex data workflows on their screen instead of reading pages and pages of cryptic code, which the non-programmer members would not necessarily be able to comprehend.

Tasks for managing database structure was very fast using the tool. For example, table design was a 1-minute task instead of hours or days required for specifying and defining traditional SQL databases. This was possible by designing the target SQL database directly from data - using the Alteryx data processing workflow, all key attributes for the target SQL database were possible to be configured in an automated fashion. It was a rather simple task to review the schemas and table definition from T-SQL scripts available from the environment.

5.3.3 Statistical analysis and predictive model prototyping in Alteryx

Once the data collection and clean-up processing were completed on the workflow demonstrated in the attached image, we performed the statistical analysis to identify feasible variables for prediction modelling. Statistical reports in the first project were developed in Alteryx to understand the form and quality of the data. Statistical reports were exported to PDF format, and results stored in SQL database to share between the development team and product owner organization.

Following analysis was performed for the collected data:

- Dataset descriptive reports (quantities, min/max/std/mean summary, missing value info)
- Filtering of production system testing measurement variables to numerical records with sufficient variance (variables reduced from up to 431 to 271).
- Correlation analysis (PCC) between target variable (product failure) and product system testing measurement variables (up to 271), As Pearson Correlation (PCC), requires normalized data and we used standardization to have all variables comparable to each other within scale between 0 to 1.
- Filtering of the variable set to highest PCC scores (highest scored for product 1: 0.15, product 2: 0.01)

Overview of the workflows and timelines of activities spent in processing and analysing the data compared to previous manual operation are demonstrated in the diagrams below.

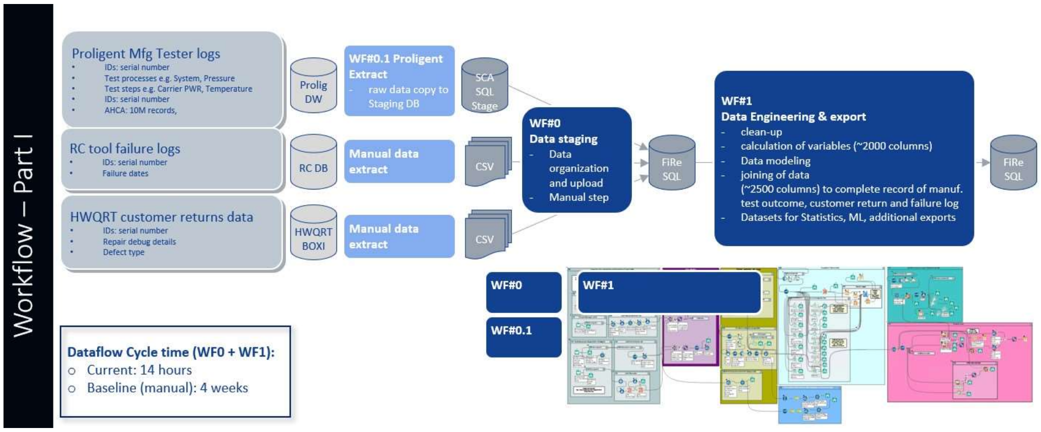
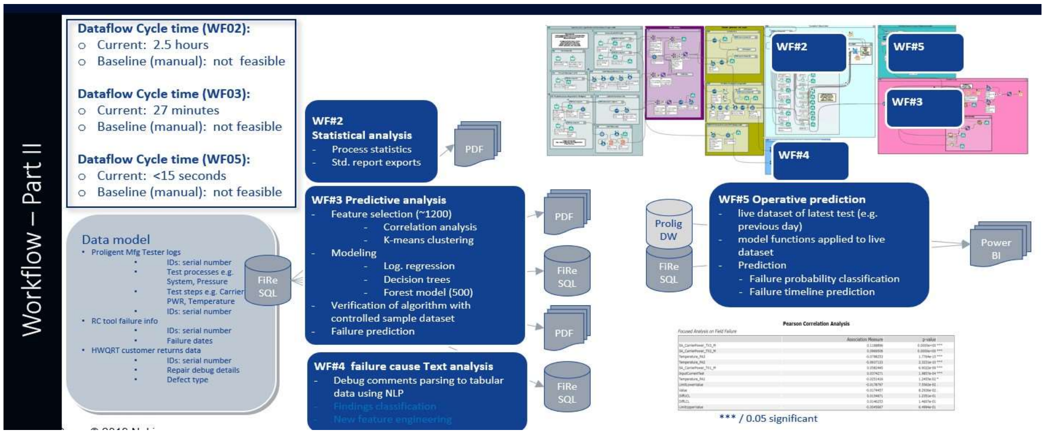

Workflow – Part I
Proligent Mfg Tester logs
IDs: serial number
Test processes e.g. System, Pressure
Test steps e.g. Carrier PWR, Temperature
IDs: serial number
AHCA: 10M records,
RC tool failure logs
IDs: serial number
Failure dates
HWQRT customer returns data
IDs: serial number
Repair debug details
Defect type
Prolig DW
RC DB
HWQRT BOXI
WF#0.1 Proligent Extract
- raw data copy to Staging DB
Manual data extract
Manual data extract
SCA SQL Stage
CSV
CSV
WF#0
Data staging
- Data organization and upload
- Manual step
FiRe SQL
WF#1
Data Engineering & export
- clean-up
- calculation of variables (~2000 columns)
- Data modeling
- joining of data (~2500 columns) to complete record of manuf. test outcome, customer return and failure log
- Datasets for Statistics, ML, additional exports
FiRe SQL
WF#0
WF#0.1
WF#1
Dataflow Cycle time (WF0 + WF1):
○ Current: 14 hours
○ Baseline (manual): 4 weeks

Figure 24: Alteryx workflows part 1: data cleaning, blending and storage



Workflow – Part II

Dataflow Cycle time (WF02):
- Current: 2.5 hours
- Baseline (manual): not feasible

Dataflow Cycle time (WF03):
- Current: 27 minutes
- Baseline (manual): not feasible

Dataflow Cycle time (WF05):
- Current: <15 seconds
- Baseline (manual): not feasible

Data model
- Proligent Mfg Tester logs
  - IDs: serial number
  - Test processes e.g. System, Pressure
  - Test steps e.g. Carrier PWR, Temperature
  - IDs: serial number
- RC tool failure info
  - IDs: serial number
  - Failure dates
- HWQRT customer returns data
  - IDs: serial number
  - Repair debug details
  - Defect type

WF#2
Statistical analysis
- Process statistics
- Std. report exports

WF#3 Predictive analysis
- Feature selection (~1200)
  - Correlation analysis
  - K-means clustering
- Modeling
  - Log. regression
  - Decision trees
  - Forest model (500)
- Verification of algorithm with controlled sample dataset
- Failure prediction

WF#4 failure cause Text analysis
- Debug comments parsing to tabular data using NLP
- Findings classification
- New feature engineering

WF#5 Operative prediction
- live dataset of latest test (e.g. previous day)
- model functions applied to live dataset
- Prediction
  - Failure probability classification
  - Failure timeline prediction

Figure 25: Alteryx workflows part 2: statistical analysis, predictive modelling, text analysis, operative prediction scoring

Predictive analysis prototyping overview

After completing the statistical analysis, the next task in the project iteration was to develop predictive model prototypes in Alteryx to evaluate the performance and quality of the data. Dataset oversampling was done to produce a balanced dataset with product failure scores divided as 50/50. The training and validation dataset sampling were done with a ratio of 70%-20%-10% of the previously oversampled data.

Prototyping of usable models was then done by comparing the performance of different models using the top 20 PCC scores. Models evaluated included simple logarithmic regression, decision tree (unlimited), random forest (500 iterations with all variables and random sampling), support vector machine (using feature reduction and PCA vectors), and basic neural network (1-layer unsupervised learning). Alteryx had a good set of tools for evaluation of the model performance. Iterative model scoring was produced by taking a random sample from the training dataset to the Cross-Validation tool and validation dataset to Lift Chart Tool and Model Comparison tool. Alteryx produces reports for each model, including overall accuracy and confusion matrix details for performance evaluation. Reports were exported to PDF format for sharing between the development team and product owner organization. Overview of the predictive workflows is shown in diagram.

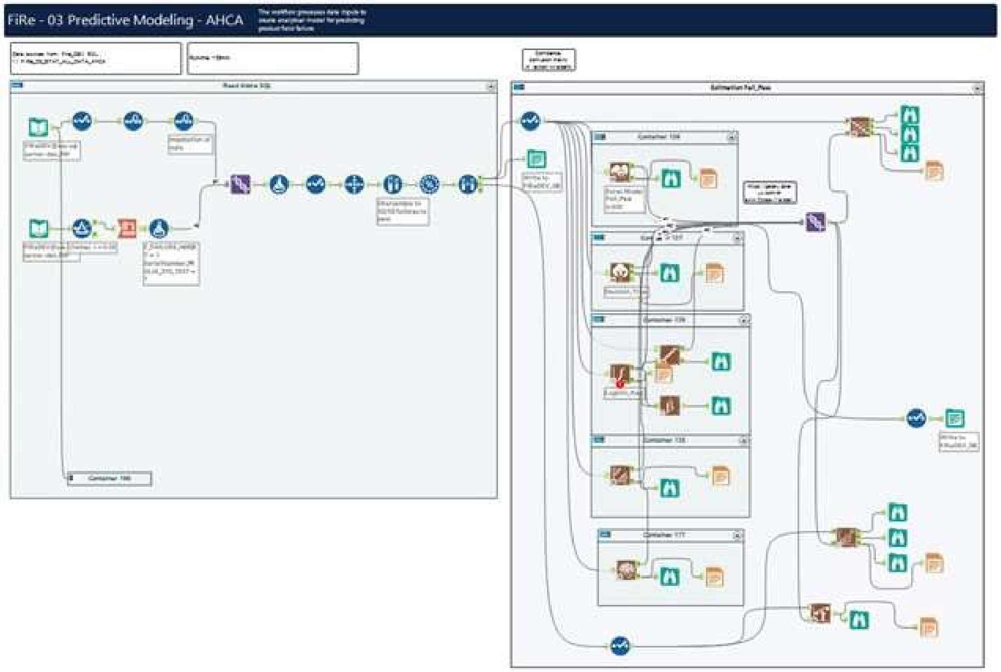


Figure 29: Alteryx workflow for predictive model comparison

Initial results with full datasets (without model optimization) showed the highest accuracy towards correct failure prediction on classifier models at up to 85 per cent.

There was a small number of failure records available for product1, but the decision was to continue research to the next steps of model improvement with a focus on supervised learning and random forest model optimization in Phase 3.

5.3.4 Data statistics automation and data variance/dissimilarity analysis (PCA, MDS) in Python

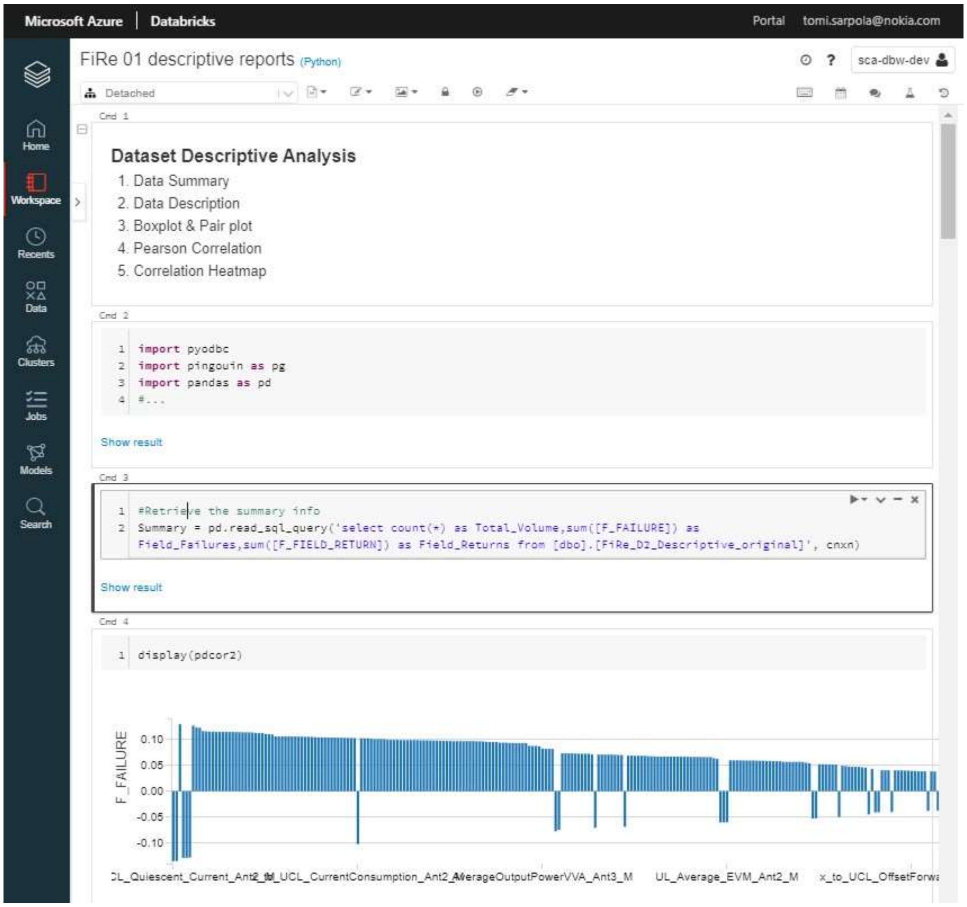
This activity had dual objectives as we wanted to validate data sample statistics used in Alteryx and improve data understanding. Additionally, due to identified limitations in the desktop-based setup used in previous phases, the intention was to apply the recent learnings of cloud-based computing capabilities and develop the solution to handle big data.

We decided to introduce Azure cloud data centre services to improve the computation performance and utilise Azure Databricks environment for our Python code development.

Python development is largely different to Alteryx development as the second one is focusing on code-free development of complex workflows. Python development environment, in contrast, is focusing on fully code-based development of specific functions and procedures, which will be expanded iteratively to full workflows. Code development started with a desktop-based Anaconda Jupyter Notebook. We further evaluated cloud-based team development environments and the code compatibility in Alteryx Jupyter Notebook, Azure ML Studio Notebook, and Azure Databricks.

Python development in Azure Databricks is based on notebook files executed on Spark enabled server nodes configured to a computation cluster. The environment is web based, enabling all developers to access the same workspace resource for better collaboration. Internal version management ensures that code is maintained. Results of the execution are either stored in the notebook file, to export files or databases. When operating on high performance (e.g. high cost) cluster configuration, storing the results to external data storage is highly advised.

Attached graphic shows an overview of a typical Databricks notebook we used.



Dataset Descriptive Analysis

1. Data Summary
2. Data Description
3. Boxplot & Pair plot
4. Pearson Correlation
5. Correlation Heatmap

```
import pyodbc
import pingouin as pg
import pandas as pd
#...
```

```
#Retrieve the summary info
Summary = pd.read_sql_query('select count(*) as Total_Volume,sum([F_FAILURE]) as
Field_Failures,sum([F_FIELD_RETURN]) as Field_Returns from [dbo].[FiRe_D2_Descriptive_original]', cnxn)
```

```
display(pdcor2)
```

Figure 26: Databricks development environment

Data descriptions were produced in Databricks notebooks using

- Dataframe & pandas basic functions to show data summary (min, max, etc.),
- Seaborn for data distribution Boxplots
- Seaborn for Timeseries of manufacturing weeks by failure codes
- pingouin for Pearson Correlation statistics (Pingouin pairwise correlation plots shown if compatible, otherwise Pandas.df.corr used, e.g. in ML Studio)

Data distribution boxplots produced a promising view of clear centre points for each variable and large groups of data points outside of the centre area, which may be useful for the prediction algorithm. Additionally, it was evident that some of the features contained very exceptional outliers, which would need to be evaluated for possible bias effect.

Variance in data was also analysed with dimensionality reduction and variance or dissimilarity validation using PCA and MDS methods available from Python Scikit Learn libraries. With the PCS visualization, we were able to inspect if there was clear variance and visual patterns evident in the data to identify useful features for the prediction algorithm training.

PCS provides an approximation of a high-dimensional data set that we had to a low-dimensional linear subspace. MDS has a similar function to extract underlying dimensions from data, but it targets to preserve the distance between data points. Using both techniques, we were able to identify clear clusters visible in the overall data. The attached image shows the two main clusters.

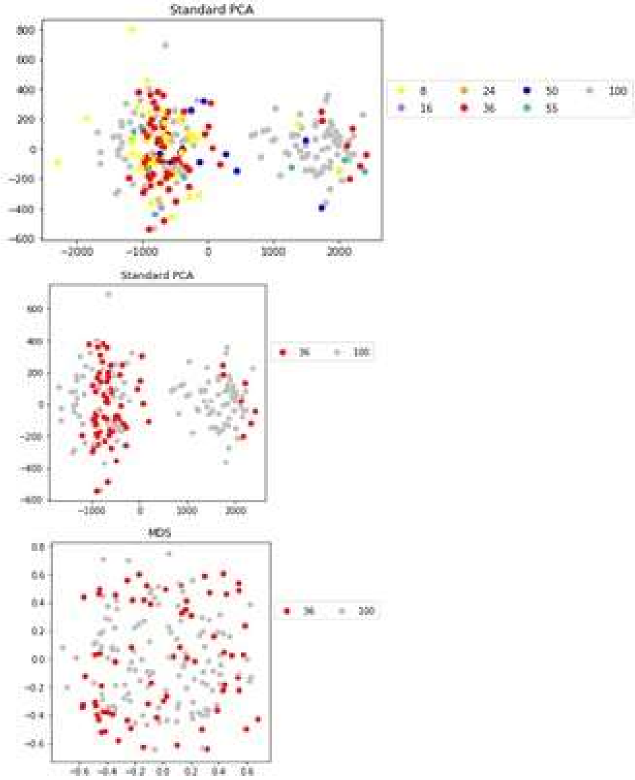


Figure 28: Data clustering example using PCA & MDS

Consequently, when inspecting the visualization of clusters and target category profiles, it was evident in the PCA plot that the category had data points occurring in both main clusters. In the MDS plot, the distance between data points from centre was relatively constant and provided a limited amount of information gain. Conclusion from this was that correlation of product failures in the target category and the system test variables did not have outstanding signals to indicate strong correlation, but data was more homogenous with only weak signals inside.

5.3.5 Supervised predictive model (RF) optimization using Python

In the model optimization phase (Sept - Nov), we had two objectives in ML analysis, first to improve the random model performance by enhancing understanding of data and improving parameter and variable selections, secondly to study available non-linear correlation by having a deep neural network (2-layers) system to determine which variables to use. As this part of optimization work required the introduction of Python to the toolset, and the decision was to continue this phase of project iteration focusing to development in Python. Model accuracy evaluation was developed to include details on precision and recall. Product failure accuracy evaluation was decided to be done based on recall as the highest cost was evaluated from false-negative results, which has the impact of products with high fault risk being delivered to the customer. As an outcome, we produced the evaluation of predictions of products with risk of failure and the most significant system test variables were produced for business team for evaluation.

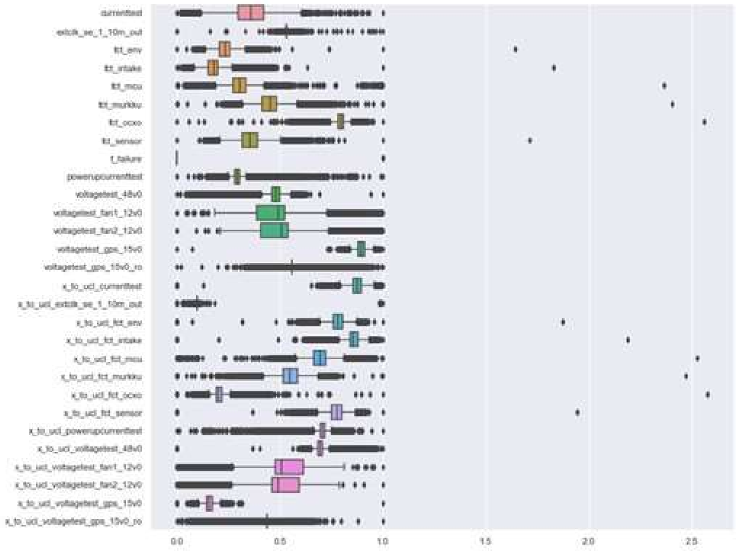


Figure 27: Data variance boxplot example

Let's reiterate the difference between model performance metrics of precision and recall. Precision of a model is calculated by dividing true positive results by both false and true predictions, TP/(TP+FP). Precision answers the following: How many of those who we labelled as faulty is actually faulty? Such products should be re-evaluated. Non-faulty products may have the same characteristics as faulty products but have not failed yet.
Recall is calculated by dividing true positive results by the total of true positive and false negative predictions, TP/(TP+FN). Recall answers the question: Of all the faulty products, how many of those we correctly predicted? Such products should be evaluated if there is a business risk in delivering them to customer.

Predictive analysis was further extended to improve model performance. We made evaluation of impact change to model performance by extending the set of variables from the Top 20 PCA variables to all variables to produce a baseline comparison. Analysis used Skitit-Learn and evaluated performance between Logistic regression, SVM, decision tree, and random forest. Best performance was found on random forest model and we decided to improve on that model.

5.3.6 Alternative predictive analysis in Python using unsupervised learning on neural networks

Additional research was conducted to investigate non-linear correlations in data and use a deep neural network (2-layer) to let the machine choose the best parameters and features for the predictive model. Our summer trainee conducted most of the implementation, and results were then presented and reviewed with the team.

Deep Neural network model (DNN) hyperparameter pre-tuning was done using Scikit-Learn functions RandomSearchCV (finds parameters for epochs, batch size. etc. to model to find optimal parameters as a random set) and GridSearchCV (takes input from RandomSearchCV and validates parameter tuning with grid search). The output of the best parameter set given as the mean average for the neural net. Python Keras and TensorFlow were used to create the DNN model, and we compared performance to the supervised models.

Results: overall performance of DNN was similar as with optimized random forest (RF); however, DNN optimization would have required more time to understand all considerations, and as RF did expose the decision logic in such a way that it would be easy to communicate to end customer, the decision was to discontinue development activities on DNN and to focus on finalizing optimized RF model for production.

The theoretical framework proposed that focusing on most simple algorithms in the prototyping phase would be most practical as the gained accuracy is not always justified by the cost of additional development and, in the case of neural networks, the loss of being able to clearly describe what decisions the model makes to produce the outputs.

5.3.7 Optimizing Random Forest (RF) model with hyperparameter pretuning and combinatory feature selection in Python

The Random Forest (RF) model was then further improved by optimizing forest configuration using hyperparameter pre-tuning and combinatory feature selection analysis to determine the most important feature combinations. The assumption was that combinatory analysis with N-2 selections from N=271 features would show clear accuracy variation when features are changed, we can identify the most important variables, and we can filter out less important variables. Amount of combinations for N-2 features is C(271,269) =

36585 possible subsets. Larger variable filtering, e.g. N-3 with 3 280 455 combinations were not considered feasible to be evaluated due to extensive processing time and high server cost.
Random forest hyperparameter pre-tuning was done using Scikit-learn functions RandomSearchCV (finds parameters for min depth, split, etc., to model to find optimal parameters as a random set) and GridSearchCV (takes input from RandomSearchCV and validates parameter tuning with grid search). The output of the best parameter set given as the mean average for the random forest.

We then executed random forest combinatory analysis with 36585 iterations. Forest was trained on each iteration with 200 random samples from the training dataset and used N-2 random variables from the possible 271. The analysis produced a comparison of model improvement from simple PCC filtered variable set, measurement of performance to Random Forrest, which uses all variables, and finally the Random Forrest, which uses optimized parameters and variables.

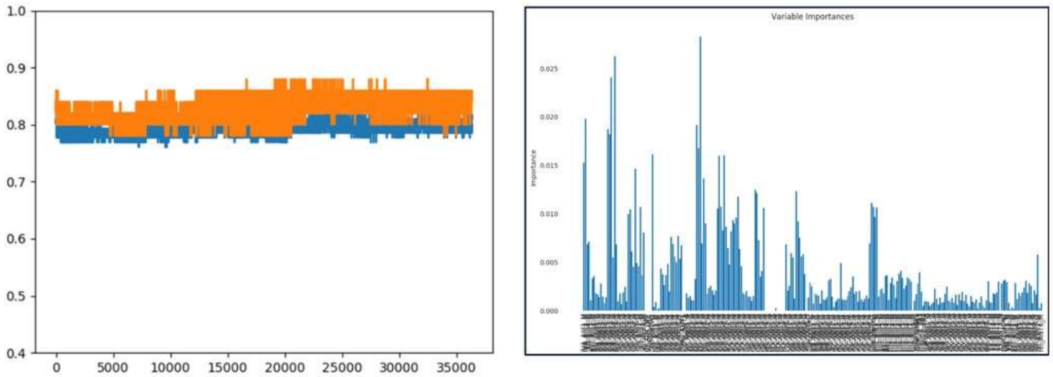
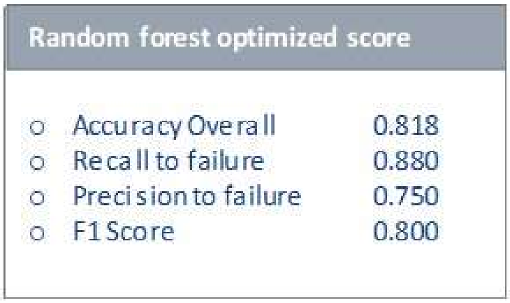


Figure 30: Example RF combinatory analysis accuracy variation and variable importance



Random forest optimized score

○	Accuracy Overall	0.818
○	Recall to failure	0.880
○	Precision to failure	0.750
○	F1 Score	0.800

Figure 31: RF model results

The final optimized model produced adequate overall accuracy of 0.818 and a recall of 0.880. After this large amount of evaluation of model performance, we were confident in providing the business team with the outcomes of the model development.

A similar evaluation of the prediction concepts and model performance was also produced in the second FiRe project iteration, which analysed the correlation of logistics drivers, e.g. manufacturing location, customer locations, and product characteristics, towards increases of product returns over time.

5.3.8 Architecture development to 1st operational MVP release (Data Ops)

Azure cloud environment was initially developed in the first project to study the capabilities and system integration. Work continued in the second project. The objective was to produce capabilities for the developed prediction solution so that machine learning could be operated in an automated fashion for full big datasets of the prediction environment. Data capability requirements were evaluated based on PoC of 10 products and 2,2 million records, measuring finally to some 300 million records of data for 250 products in the current portfolio. The operation of the models was also trialled using Azure Databricks and prediction score serving in API based architecture using Azure Container Image (ACI) or Azure Kubernetes (AKS) service.

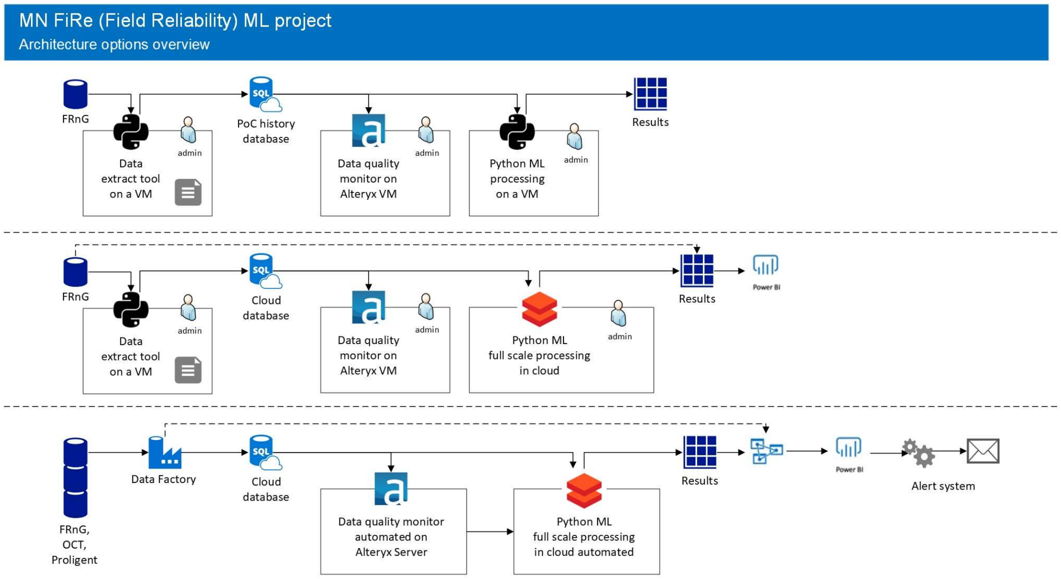

MN FiRe (Field Reliability) ML project
Architecture options overview
FRnG
SQL
PoC history database
Results
Data extract tool on a VM
admin
Data quality monitor on Alteryx VM
admin
Python ML processing on a VM
admin
FRnG
SQL
Cloud database
Results
Power BI
Data extract tool on a VM
admin
Data quality monitor on Alteryx VM
admin
Python ML full scale processing in cloud
admin
Data Factory
SQL
Cloud database
Results
Power BI
Alert system
FRnG, OCT, Proligent
Data quality monitor automated on Alteryx Server
Python ML full scale processing in cloud automated

Figure 32: ML architecture iterative development stages

The initial stage of the second project was to collect the data samples again via manual reports. However, after a short development cycle, the project produced automation using Selenium robotics process automation to collect data from existing reporting interfaces towards project use. The architecture setup was then enhanced with data quality monitoring produced with the Alteryx tool to quickly compare different data sources to understand the completeness of data. The data science team was tasked with developing the predictive

algorithm using the field reliability database - work was started on a local desktop environment to evaluate the data and examine the proof of concept.
I then developed, with the team, an architecture roadmap to iterate the Proof of Concept system towards cloud architecture in the second step and finally to a full cloud automation system.

Development to cloud version as MVP (Minimum Viable Product) was decided to be produced as deliverable for end of 2020. MVP architecture is illustrated in attached image and larger view is provided in appendixes.

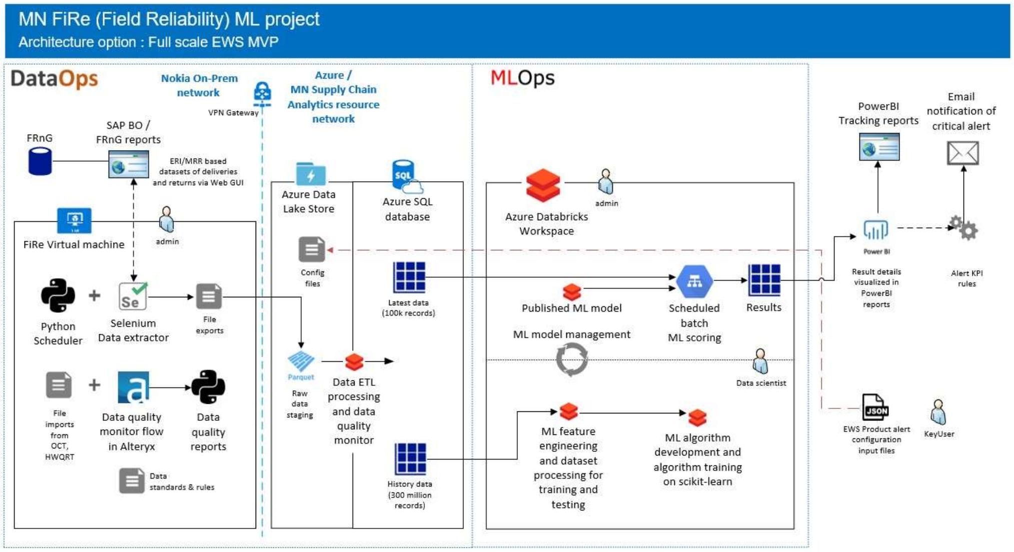

MN FiRe (Field Reliability) ML project
Architecture option : Full scale EWS MVP
DataOps
Nokia On-Prem network
VPN Gateway
Azure / MN Supply Chain Analytics resource network
FRnG
SAP BO / FRnG reports
ERI/MRR based datasets of deliveries and returns via Web GUI
FiRe Virtual machine
admin
Python Scheduler
+
Selenium Data extractor
File exports
File imports from OCT, HWQRT
+
Data quality monitor flow in Alteryx
Data quality reports
Data standards & rules
Azure Data Lake Store
Azure SQL database
Config files
Latest data (100k records)
Parquet
Raw data staging
Data ETL processing and data quality monitor
History data (300 million records)
MLOps
Azure Databricks Workspace
admin
Published ML model
ML model management
Scheduled batch ML scoring
Results
Data scientist
ML feature engineering and dataset processing for training and testing
ML algorithm development and algorithm training on scikit-learn
PowerBI Tracking reports
Email notification of critical alert
Power BI
Result details visualized in PowerBI reports
Alert KPI rules
EWS Product alert configuration input files
KeyUser

Figure 33: Architecture description of MVP cloud version

MVP data sourcing had interfaces to collect data from the on-premises system using a virtual machine to process the raw datasets towards Azure cloud environment.
Azure cloud setup contained a data lake feature for storing the data. Azure Databricks was utilized to produce a data pipeline for cleaning-up, blending and transforming of data in cloud.

The machine learning model production and ML operations were setup also in Azure Databricks, which enabled high capacity data centre cluster for the processing of big datasets efficiently. Outcomes of the predictive system would be produced as visual dashboards and email alerts towards the end-users.

5.3.9 Evaluation of full-scale automated Data and ML algorithm operation and maintenance (ML Ops)

The program included an evaluation of the machine learning model serving architecture options on the Azure cloud platform. In the production scale architecture, the prediction score can be produced as API based functions in Azure Container Image (ACI) or Azure Kubernetes (AKS) service.

Microsoft and Databricks presented an overview in one of their seminars of what bringing up ML / AI system to an operational state often requires. The below picture was produced by Google Inc. while evaluating their projects. (Gary Holt, 2015)

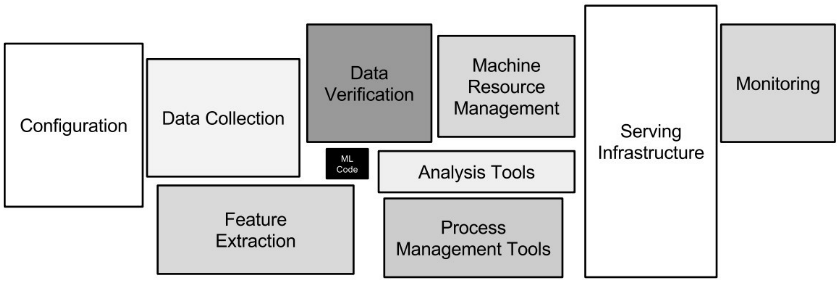


Figure 36: Fraction of real-world ML systems is composed of the ML code compared to surrounding infrastructure

AI systems are typically made up of a series of workflow pipelines. Also, data itself influences ML system behaviour. Mitigation actions for the common problems of implementing technical ML system setups were studied from the background material. One such concept is reducing technical dept of development team.
Technical debt is the ongoing cost of fast decisions made when implementing code. It is all the shortcuts or workarounds in technical decisions that give short-term benefit in earlier software releases and faster time-to-market (Chia, 2019).

We investigated the three common issues in machine learning systems summarized by Derek Chia, and how we could overcome those in our area.

Hidden Feedback Loops, i.e. the indirect influence of AI/ML system on the source data and performance metrics (Chia, 2019). Hidden feedback loops were studied with the collaboration of business and technical expert teams to avoid tunnel vision in the algorithm development, and to also enable exploration of potential improvements.

Pipeline Confusion, i.e. growing complexity of workflow pipelines, number of requirements, messy code, and undeclared consumers of data streams (Chia, 2019).
Data pipeline was managed with large focus in the projects, with high priority on co-development, documentation and minimizing features in operative flow to only the necessary parts. Target with this was to avoid larger problems from confusing code which tend to snowball to bigger problems and delays in projects.

Data Dependency i.e. similar to software system module and library dependencies, ML/AI system depend on datasets of a certain quality (Chia, 2019).
Data dependency issues e.g. when changing interfaces and dataset versions was managed with data quality monitoring, frozen copies of datasets and their data mapping. New features and datasets were also introduced to the system in a controlled manner.
Technical debt can be kept to minimal by asking a few useful questions during development. We should ask ourselves: Do we know precisely how a new model or update of a model would affect the entire system? Where and how can we measure the impact of this change? What does our metric of success look like? Does it align with the business objective? Are we keeping track of all the dependencies, producers, and consumers of the entire system? Do we have a systematic way of identifying faults when we need to? (Chia, 2019)

ML operations in Databricks

We utilized code integration from personal environments to a consolidated test environment during the development process and finally to a production environment. Having four persons work on the production environment setup and working on different system areas

We had the opportunity to discuss with the Databricks team on CI/CD best practices and development patterns. Important topics to note was the available Databricks Connect SDK, CI/CD process overview in Microsoft Azure documentation, and available templates for testing sample projects, integration automation, and ML pipeline integration. The attached Databricks reference architecture proposes the information system elements to be utilized in a modern application to produce automated machine learning and analytics solutions.

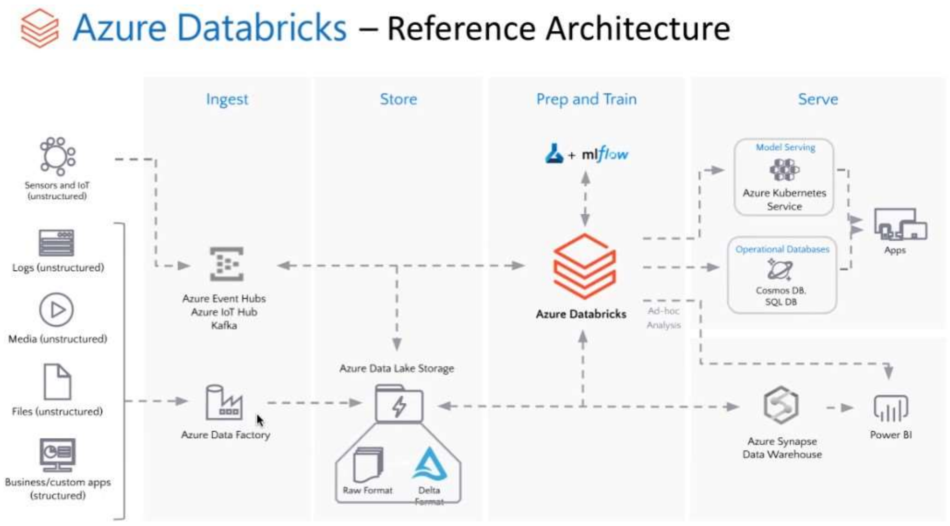


Figure 37: Databricks reference architecture

Model evaluation metrics

Automating machine learning model integration pipeline requires additional focus on model measurement and evaluation. Silver points out in his book that simply looking at improvements in overall Accuracy performance is not sufficient measurement for evaluating models. Additional measurement techniques for measuring models were described in the book ’Data Science for Business’ 2013.

When evaluating the performance of the forecast, we need to outperform baseline measures of i) value will remain the same, ii) historical value from the same timeframe from history will repeat. Additionally, we need to test that all forecasts do not fall to the majority class or most significant feature of an imbalanced dataset.

Challenges and limitations identified with Alteryx platform

Lessons learned during development program about graphical workflow development tool Alteryx also contained challenges and limitations. As an alternative we decided to propose code-only based platform of Python on Databricks as production environment.
Here are few learnings we identified in Alteryx platform suitability for our use case.

From system integration perspective, level of integration and iterative development is a limitation since Alteryx does not produce program code outputs of the workflows, which could be easily adapted and integrated to planned Python production environment. This needs to be considered in the system design and Alteryx environment included on necessary scale. Additional consideration is the requirements for desktop and server hosting.

So far, there was no cloud service readily available. Server based approach fits easily to fully on-premises architectures. Hybrid and full Cloud architectures need additional focus in design.

Alteryx workflows are possible to be automated. Server license is required for automation. So far, the limitation has been with the CPU based licensing model. It is challenging for the organization to predict how much CPU capacity or how many CPUs in total long-term development project will require, and to calculate the ROI in advance for all projects to justify the high first-time setup costs.

Regarding team management and collaboration, Alteryx would benefit from additional integrated version control and team collaboration tools. Modern organizations apply continuous integration and continuous delivery models in their development work. Typical code and artifact management is done with version control tools like Gitlab, MS DevOps or similar, where also development backlogs, software deployment and integration pipelines can be managed.

FiRe MMP architecture proposal

Second FiRe project iteration developed the architecture proposal for full-scale MMP phase based on the previously introduced considerations. MMP features would support the collection of datasets from different sources, data quality monitoring, data pipeline to collect on-premises data to processing in the cloud, and ML operations utilizing automated ML training, metrics evaluation, and deployment to parallel computation clusters.

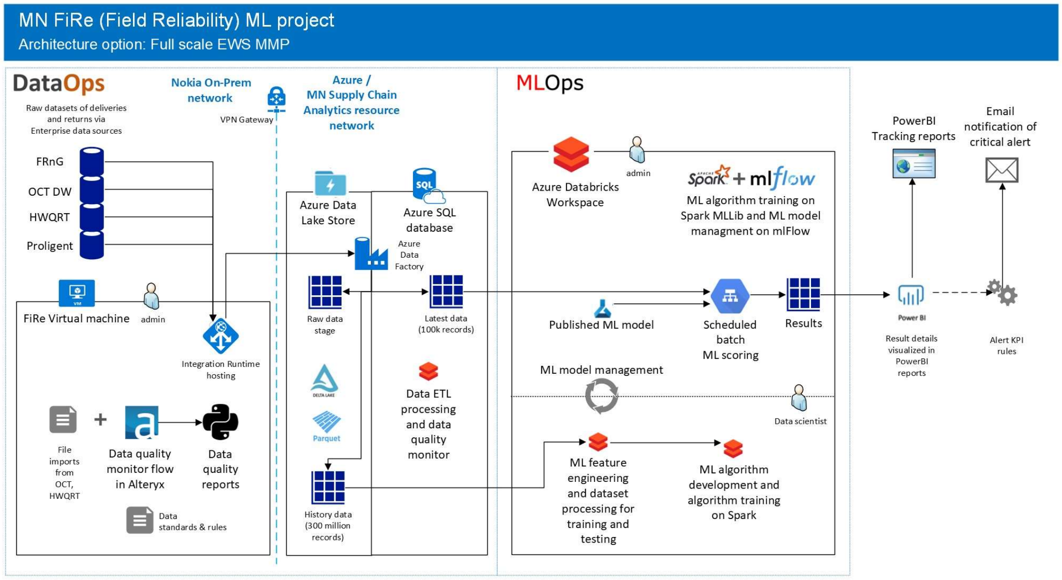

MN FiRe (Field Reliability) ML project
Architecture option: Full scale EWS MMP
DataOps
Raw datasets of deliveries and returns via Enterprise data sources
Nokia On-Prem network
VPN Gateway
Azure / MN Supply Chain Analytics resource network
FRnG
OCT DW
HWQRT
Proligent
VM
FiRe Virtual machine
admin
Integration Runtime hosting
File imports from OCT, HWQRT
+
Data quality monitor flow in Alteryx
Data quality reports
Data standards & rules
Azure Data Lake Store
SQL
Azure SQL database
Azure Data Factory
Raw data stage
Latest data (100k records)
DELTA LAKE
Parquet
Data ETL processing and data quality monitor
History data (300 million records)
MLOps
Azure Databricks Workspace
admin
APACHE Spark + mlflow
ML algorithm training on Spark MLLib and ML model managment on mlFlow
Published ML model
Scheduled batch ML scoring
Results
ML model management
Data scientist
ML feature engineering and dataset processing for training and testing
ML algorithm development and algorithm training on Spark
PowerBI Tracking reports
Email notification of critical alert
Power BI
Result details visualized in PowerBI reports
Alert KPI rules

Figure 38: Architecture description of MVP cloud version

6. Conclusions

6.1 Evaluation of the research activities, organization maturity assessment, and business analytics implementation activities

The thesis objective was to study how are business analytics and data science related to business strategy, what are the current team development practices in analytics and data science area, and what are the key roles in data science projects, standards used in handling big data, and best practices on data science productization.

The outcome of the thesis met the objective. The organization gained an understanding of their analytics maturity and produced actions to improve business strategy and relation of data management and governance towards business strategy. The analytics maturity assessment survey in the thesis was limited due to timeframe constraints but demonstrated the principle and provided sufficient detail for the organization to evaluate and plan activities for areas identified to require improvement. Recommendations for further improving the organization analytics maturity are summarized in the next chapter.

The team was also able to develop the capabilities in the organization through implementation projects to improve development platforms and team competencies according to the collected learnings of development practices, data science principles, big data handling standards, and data science algorithm deployment to operative services. In FiRe program the team spent a sufficient amount of time before each iteration to collect key learnings to reduce technical debt. However, certain aspects and details of technological problems were only discovered during development and in-depth training and research was required during development. Support from the analytics platform vendors to provide clear documentation and additional education of their products was instrumental for the success of the program.

The empirical research produced a platform and a roadmap for IT services, which enabled the team to develop rapid prototypes using data exploration, functional predictive solutions as cloud applications, and full-scale AI service to integrate into the company enterprise data systems. Project iteration objectives were met on time, and we successfully delivered the first functional MVP release for operative pilot use according to the planned schedule and extent.

6.2 Recommendations for Data Strategy, Analytics processes, and organization development

Based on interviews and organization analytics maturity evaluation, there is clearly good consensus on benefits in the current state and positive progress of the capabilities. Large investments are already made to keep the organization up to speed in analytics development. Nonetheless, current state analysis indicated that while many projects are being developed in the area of analytics, data strategy and analytics strategy is in a current form not well perceived or clear in all organization levels. Data management principles are applied in most areas, and a large amount of development has been done in the development of data-related IT platforms.

However, data is not perceived as a strategic asset for the organization, and typical problems of underutilized data exist in the organization information systems. To evolve in analytics maturity and maintain a competitive edge on the markets from analytics and automation, the journey to the next level may require more significant long-term investment and push to become successful.

Key findings and recommendations for organization analytics capabilities development:

i) Organization would benefit from a clearly structured data strategy formulated on a business group level. Digital transformation of organizations is a long-term activity that requires repeated push and focus from top management to be successful. Currently, the focus of digitalization is in the high-cost migration of legacy ERP and PLM systems inherited from several company mergers towards modern ERP and PLM tooling. While being an imperative foundation to have in the company to enable high-quality data analytics, too high focus on the core systems enables the surrounding environment to erode.

In the development and implementation of a machine learning solution for business process the team identified several typical problems preventing efficient analytics solution development. Key issues identified were high complexity of the data landscape and restricted access to data, varying capability of the business processes to produce accurate data, poor interlock of the processes to provide full lineage of data, lack of high-quality data documentation, and finally, the existence of competing data pipelines that produced inconsistencies when examining data quality and completeness of datasets. Data quality challenges are also expected to continue as many business teams are still utilizing mostly spreadsheets, presentations, and emails in daily information tasks - as no better alternatives seem to be easily available or perform wildly different from the familiar local tools.

From collected examples, challenges are found in areas like tracking of business development tasks in a systematic manner for development benefits evaluation and business case formulation; product data, especially in the product development phase, is maintained largely in spreadsheet tools, and this information is not easy to utilize in analytics and business process automation as the quality of the spreadsheet content and availability to different environments is not guaranteed. Clearly defining a role for the business in data governance to regularly utilize and evaluate the condition of the collected data is a prerequisite for ensuring we can evolve data to strategic assets.

ii) Building a well-formed Analytics Centre of Excellence as a community for business analytics and data scientists would benefit the organization to transform. Several data-oriented teams are working on transforming the organization reporting and analytical practices to online dashboards and machine-based analytics in the current state. Many projects start to build their own datasets and collections, while enterprise data pipelines remain seldomly utilized and without active development focus.

From the survey and collected measurement of analytics utilization, it is seen that in those areas where development is done to produce automated dashboard reports and analytics, often the utilization rate is on a good level only inside the requestor team, and wide-scale adoption of the developed tools may not occur. Considering many simple prototypes and tests done by the team, there is an alarming quantity of the tens of thousands PowerBI reports deployed to workspaces that have only a few active users.

The organization’s analytics Centre of Excellence community would provide the organization methods for driving analytics utilization within the different teams, methods for sharing knowledge on data, available models, analytics techniques and tooling best practices, and finally, methods for transforming the development towards standardized and efficient analytics processes. Regular measurement and investigation of analytics maturity of organization would also be one key channel of information on progress, and clear activity to be executed by the analytics Centre of Excellence.

iii) Organization is recommended to continue building centralized data pipelines, to produce systematic outputs for the whole organization to use. In the current state key programs exist on the enterprise IT level for producing the data pipelines. However, they are not utilized to their full potential as long as the development is being driven by other organization groups, who may have different objectives towards using the data and different perspective of how critical the data is for the organization. Conflicting interests produce, for example, data bias by removing important elements of data when doing data warehouse modelling. The effect is that provided datasets fit only for use in specific processes or business areas, and the dataset is not utilized any longer organization wide. Continuing data governance and building common business data objects would benefit the organization in uniform baseline data and understanding for analytics development projects. It is recommended that data strategy is formulated to define the focus of data quality and data pipeline improvement activities are linked to critical processes related to business decision making and digital assets supporting decision making are clearly mapped and documented.

The development programs managing the data pipelines should be centrally managed and empowered to execute iterative development. Logically structured content strategy and information architecture are required to enable functionalities and solutions to be inefficient and consistent.

iv) Organization would benefit from hiring a data statistician or data scientist to support AI supported and data-driven decision making. As described in the theoretical framework, personnel experienced in managing large sets of information and solving mathematical problems in novel ways are pivotal drivers in organization development. One aspect is the ability to apply their experiences and knowledge of their own domain area and adapt it to produce results for the business.

The second aspect is related to driving machine enabled decision making and culture, as data-oriented persons are typically fluent in educating management level personnel on how to adapt their decision-making strategies to enable machine-based decision support.

Understanding where we can utilize ML/AI methods and which scenarios require entirely different approaches is critical when evaluating the development of high-complex and usually high-cost AI solutions. If the organization does not evolve and enable the development of data-driven solutions by documenting and digitizing decision processes, the investments in this arena will quickly stumble to the typical data science development pitfalls, and the organization reverts to using old and proven practices.

v) Thesis produced a small collection of tools and guidelines for use in the organization for selecting and defining data science projects. It is recommended that the organisation invest further effort in systematically applying and further developing those guidelines to fit the organization data strategy objectives.

Guidelines should drive towards producing identification and measuring success behind business analytics and AI/ML projects. Lack of evidence of success after solution implementation has two critical effects: not showing investment return and not having experience and data to support defining business cases for future investments.

Finally, the extent of implementing these recommendations depends heavily on the company business strategy and mode of operation, which defines if traditional, yet quite flexible human-based decision making is still the best fit for the company or if the next level of digital transformation and analytics maturity is the key objective to aim for.

6.3 Considerations for IT architecture roadmap development and evolution to data-driven decision making

In the FiRe program and project work, we demonstrated highly complex prediction solution development initiated from business teams and iteratively developed from prototype to first operative solution.

Following considerations is the collected learnings from projects, findings of success factors and identified challenges to consider in IT architecture roadmap development.

In the early phases, the impact of existing data understanding, and literature is emphasized as the team spent most of its time souring and investigating the data. Clear data governance and standardized data pipelines to support raw data delivery towards data science workspaces would speed up the data science development. Data literature produced and data stewardship actions reduce the amount of work clarifying data inconsistencies and possible gaps.

Enabling business teams to take ownership of the data their business processes are producing is a complimentary activity on top of data governance. Business analysts who already can describe in great detail the shape and internal relationships of the data are key drivers for speeding up the business understanding in the data science processes. As it was witnessed that many teams operate extensively with spreadsheets, they may not have sufficient experience operating with relational database queries and extensive software code. To overcome this bottleneck, code-free tooling should be established in the general IT capability portfolio. Organizations are no longer stuck in the usage of single tools like spreadsheets.

Clear governance and definition of roles will help the organization to ensure efficient analytics-related work. Introducing concepts like Bi-modal IT and ML Ops enables traditional

IT approaches to be scaled for business team use. This enables an agile and iterative way of evolving data solutions from tests and prototypes to mature and robust Enterprise-grade digital services. A large amount of competitive advantage can be achieved with the ’80-20 Pareto rule of Data Science’ applied in a wide area of organization’s teams. Simultaneously, having the centralized teams responsible for operative environments and PhD grade data scientists working in development laboratories and driving centre of excellence activities provides clear channels for finding the necessary skills and resources for productizing data service prototypes and ideas.

The first functional MVP release of an early warning system developed in the 2020 FiRe project iteration was deployed to operative use by the organization. The pilot is agreed upon for the first half of 2021. Objective of the pilot is to evaluate the predictive system’s business benefits. Piloting will also collect insights and ideas on improving the predictive model to better fit into the business process and produce advanced automation to support technical investigations and decision making.

An additional development project iteration is also under preparation to address the proposals for enhancements.

The second project iteration to produce MVP is now finished successfully and closed.

References

Ahlstrom, J. 2014, How to Succeed with Continuous Improvement: A Primer for Becoming the Best in the World, McGraw-Hill. Visited 10.10.2020.
URL: https://haaga-helia.finna.fi/Record/nelli21.3710000000372640

Aiken, P. and Harbour, T. 2017, Data Strategy and the Enterprise Data Executive,
1st edn, Technics Publications, LLC, Denville, NJ, USA.

Akerkar, R. 2019, Artificial intelligence for business, Springer, Cham.
URL: https://haaga-helia.finna.fi/Record/3amk.271567

Antonov, A. 2018, Applying Artificial Intelligence and Machine Learning to Finance and Technology, TowardsDataScience.com. Visited 20.10.2020.
URL: https://towardsdatascience.com/applying-artificial-intelligence-and-machinelearning-to-finance-and-technology-378cbd5e5c85

Bocij, P., Greasley, A. and Hickie, S. 2015, Business information systems : technology, development and management for the e-business, fifth edition edn, Pearson, Harlow, England ; New York.
URL: https://haaga-helia.finna.fi/Record/3amk.272969

Chia, D. 2019, ‘3 common technical debts in machine learning and how to avoid them’. Visited 23.11.2020.
URL: https://towardsdatascience.com/3-common-technical-debts-in-machinelearning-and-how-to-avoid-them-17f1d7e8a428

Chiu, J. 2019, Why Your Company Needs Python for Business Analytics, Datacamp.com. Visited 23.11.2020.
URL: https://www.datacamp.com/community/blog/why-your-company-needspython-for-business-analytics

Colson, E. 2019, ‘What ai-driven decision making looks like’.
URL: https://hbr.org/2019/07/what-ai-driven-decision-making-looks-like

DAMA International 2009, The DAMA Guide to the Data Management Body of Knowledge - DAMA-DMBOK, Technics Publications, LLC, Denville, NJ, USA.

de Graaf, R. 2019, Managing Your Data Science Projects, Apress.
URL: https://haaga-helia.finna.fi/Record/nelli21.4100000008403485

Douglas B. Laney, J. K. T. 2020, Building Analytics Teams, Packt Publishing.
URL: https://haaga-helia.finna.fi/Record/nelli21.4100000011345030

Farber, M. and Illustrated, S. 2012, The Great One: The Complete Wayne Gretzky Collection, Sports Illustrated, McClelland & Stewart. Visited 09.10.2020. URL: https://books.google.fi/books?id=eSGPtgAACAAJ

Gartner 2017, Gartner Magic Quadrant for Business Intelligence and Analytics Platforms 2017, Vol. G00301340, Gartner. Visited 4.10.2020.
URL: https://www.gartner.com/doc/3611117/magic-quadrant-business-intelligenceanalyt-ics

Gartner 2019, Magic Quadrant for Data Science and Machine Learning Platforms 2019, Vol. G00385005, Gartner. Visited 4.10.2020.
URL: https://www.gartner.com/doc/reprints?id=1-1YCR6NY7&ct=200213&st=sb

Gartner: Mingday, S., Scott. D. 2017, ‘Scaling bimodal — fusing it with the business: A gartner trend insight report’. Visited 25.9.2020.
URL: https://www.gartner.com/en/documents/3772092/scaling-bimodal-fusing-itwith-the-business-a-gartner-tr

Gary Holt, Daniel Golovin, E. D. T. P. 2015, ‘Hidden technical debt in machine learning systems’. Visited 23.11.2020.
URL: https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463ebaPaper.pdf

Gary Smith, J. C. 2019, The 9 Pitfalls of Data Science, Oxford University Press.
URL: https://books.google.fi/books?id=u8SbDwAAQBAJ

Grossman, R. L. 2018, ‘A framework for evaluating the analytic maturity of an organization’, International Journal of Information Management. Visited 20.09.2020.
URL: http://www.sciencedirect.com/science/article/pii/S0268401217300026

Hills, P. 2006, ‘International journal of information management’, International journal of information management 263.
URL: http://dx.doi.org/10.1016/j.ijinfomgt.2006.01.003

Hämäläinen, V., Maula, H. and Suominen, K. 2016, Digiajan strategia, Alma, Helsinki.
URL: https://haaga-helia.finna.fi/Record/3amk.210503

Król, K. and Zdonek, D. 2020, ‘Analytics maturity models: An overview’, Information 11, 142.

Kujansuu, V. 2018, ‘Suomen 50 suurinta yritystä’. Visited 25.5.2020.
URL: https://www.itewiki.fi/blog/2018/11/suomen-50-suurinta-yritysta/

Lionbridge AI 2020, 7 Types of Data Bias in Machine Learning, Medium.
URL: https://becominghuman.ai/7-types-of-data-bias-in-machine-learning2198cf1bccfd

Marr, B. and Ward, M. 2019, Artificial Intelligence in Practice: How 50 Successful Companies Used AI and Machine Learning to Solve Problems, Wiley. URL: https://books.google.fi/books?id=UbaIDwAAQBAJ

Microsoft 2018, ‘Sql server integration services - sql server integration services ssis’. Visited 10.10.2020.
URL: https://docs.microsoft.com/en-us/sql/integration-services/sql-serverintegration-services

Microsoft 2020, ‘Azure documentation’. Visited 02.10.2020.
URL: https://docs.microsoft.com/en-us/azure/

Microsoft 2020, ‘Power bi documentation - power bi’. Visited 08.08.2020.
URL: https://docs.microsoft.com/en-us/power-bi/

Microsoft.com 2020, Team Data Science Process Documentation, Microsoft.com. URL: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-scienceprocess/

Mintzberg, H. 1978, ‘Patterns in strategy formation’, Manage. Sci.
Visited 09.10.2020.
URL: https://doi.org/10.1287/mnsc.24.9.934

Nokia Oyj 2020a, ‘Nokia oyj:n vuoden 2019 viimeisen neljänneksen ja koko vuoden 2019 katsaus’. Visited 25.5.2020.

URL: https://www.nokia.com/fifi/about−us/news/releases/2020/02/06/nokia− oyjn − vuoden −2019− viimeisen − neljanneksen − ja − koko − vuoden − 2019− katsaus/

Nokia Oyj 2020b, ‘Tietoa nokiasta’. Visited 09.05.2020.
URL: https://www.nokia.com/fifi/tietoa − nokiasta

Omale, G. 2020, ‘More than half of marketing leaders are disappointed in their analytics results’. Visited 23.11.2020.
URL: https://www.gartner.com/en/newsroom/press-releases/2020-10-07-gartnerreveals-more-than-half-of-marketing-leaders-a

Patil, D. J. 2011, Building Data Science Teams, O’Reilly Media, Inc.
URL: https://haaga-helia.finna.fi/Record/nelli21.2550000001100841

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. 2011, ‘Scikit-learn: Machine learning in Python’, Journal of Machine Learning Research 12.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. 2012, ‘Scikit-learn: Machine learning in python’, CoRR abs/1201.0490.
URL: http://arxiv.org/abs/1201.0490

Provost, F. and Fawcett, T. 2013, Data Science for Business, O’Reilly Media, Inc.
URL: https://haaga-helia.finna.fi/Record/nelli21.3460000000129360

Russell, S. 2016, Artificial Intelligence: A Modern Approach, Global Edition, 3rd edition edn, Pearson.
URL: http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1419715&site=ehostlive

Sharda, R., Delen, D. and Turban, E. 2018, Business intelligence : a managerial approach, 4th edition, global edition edn, Pearson, Harlow, England. URL: https://haaga-helia.finna.fi/Record/3amk.269895

Silver, N. 2012, The signal and the noise : why so many predictions fail - but some don’t, Penguin Press, New York.
URL: https://haaga-helia.finna.fi/Record/3amk.89255

Stroh, P. J. 2014, Evolution of Strategic Planning and Today’s Role: Chief Strategy Officer, John Wiley & Sons, Ltd. Visited 09.10.2020.
URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118896006

Suomen Asiakastieto Oy 2020, ‘Top-listat’. Visited 25.5.2020.
URL: https://www.asiakastieto.fi/yritykset/top-listat

Vennel, P. 2019, ‘Formulating a pragmatic data strategy’.
URL: https://www.linkedin.com/pulse/formulating-pragmatic-data-strategy-petervennel

Vuorinen, T. 2013, Strategiakirja : 20 työkalua, Talentum, Helsinki. Saatavana Alma Talent Bisneskirjasto -palvelussa.
URL: https://haaga-helia.finna.fi/Record/3amk.211729

Widjaja, J. T. 2020, ‘How analytics maturity models are stunting data science teams’. Visited 02.09.2020.
URL: https://towardsdatascience.com/how-analytics-maturity-models-are-stuntingdata-science-teams-962e3c62d749

Appendix

Appendix 1. Maturity Assessment improvement goals

Analytics Maturity Assessment Framework by Grossman defines key processes and development goals for improving analytics maturity.

Goals for building analytic models according to Grossman:

- Models should be built from data ("empirically derived") and use generally accepted statistical procedures.
- The performance of models should be quantified with metrics and a process developed so that new models can be developed that outperform the current models with respect to these metrics.
- When rules are used that are not empirically derived using generally accepted statistical procedures, then the business, compliance, or other reason for the rule should be known and managed. The performance impact on the model of the rule should be quantified if possible.
- Models should be robust in the sense that small changes to the data result in substantially similar models.
- The processes used to clean and transform data to create the features of models should be separately managed, automated and documented, as should any pre- and post-processing require. (Grossman, 2018, p. 5)

Goals for deploying analytic models according to Grossman:

- The performance and business impact of the model in operations should be quantified and monitored on a regular basis.
- It should be possible to update the model without writing code that impacts operational products, services or systems.
- There should be a process for validating and verifying models before they are deployed broadly.
- There should be a mechanism for checking that models are being used as per compliance policies. In particular, it is important that the models and the resulting actions as deployed in the operational systems are consistent with the organization's security, privacy, and data use policies.
- Deployed models should be robust in the sense that missing or malformed data, delayed feeds, etc. do not disrupt the systems or operations associated with the model. (Grossman, 2018, p. 5)

Improving model building and deploying models:

Usually models do not generate value for an organization until they are deployed either to products, into services, or to improve operations. Typically, a modelling team develops an analytic model using statistical or other specialized applications, and then IT team deploys the model into the appropriate product, service or operational system. It is important that there would be an efficient mechanism for moving the models between the environments - for example using Python coded model and version control system able to integrate the code to deployment environment.

Consequently, there should be an architecture in place, which clearly separates environments in which models are produced, such as development environments, and environments, in which models are consumed, such as in products or in operational systems. (Grossman, 2018, p. 3)

Goals for managing and operating analytic infrastructure:

- The analytic infrastructure for managing the data required for analytics should be adequate given the volume, velocity and variety of the data and the analytic objectives and strategy of the organization.
- The analytic infrastructure available to the modelling group should be such that the data required for building analytic models is available in a timely fashion to those that build the models.
- The analytic infrastructure for deploying models should allow analytic models to be deployed efficiently and reliably into operational systems, products and services. - The analytic infrastructure should support the management of models over their entire life cycle.
- The analytic infrastructure should integrate the security and compliance needed to protect the data as required. (Grossman, 2018, p. 5)

Goals for operating an analytic governance structure:

- The analytic governance structure should include the groups responsible for building models, deploying models, and managing the analytic infrastructure, and include the appropriate stakeholders and business owners from these organizations.
- The analytic governance structure should include executive committees that involve the appropriate businesses owners and stakeholders for making the decisions required so that the analytic strategy developed can be developed and executed.
- The analytic governance structure should include technical committees for evaluating and making recommendations on analytic processes and technology that span more than one group or impact more than one stakeholder or business owner.
- The analytic governance structure should include the necessary stakeholders, decision makers, and executives so that the policies required for the security and compliance for analytic assets can be developed and implemented.
- The analytic governance structure should include a process for assessing the analytic competence of the organization and improving the analytic maturity of the organization. (Grossman, 2018, p. 5)

Goals for developing an analytic strategy and for selecting analytic opportunities:

- Analytics should used by the organization to help differentiate itself from competitors and to provide a competitive advantage.
- The analytic strategy should identify long-range analytic directions for the organization. - There should be a process for selecting analytic opportunities that optimizes the value to the organization as a whole, given the limited resources that most organizations have for building and deploying models.
- The value brought by the analytic opportunities selected should be quantified and tracked.
- The analytic strategy should manage data as corporate assets. (Grossman, 2018, p. 5)

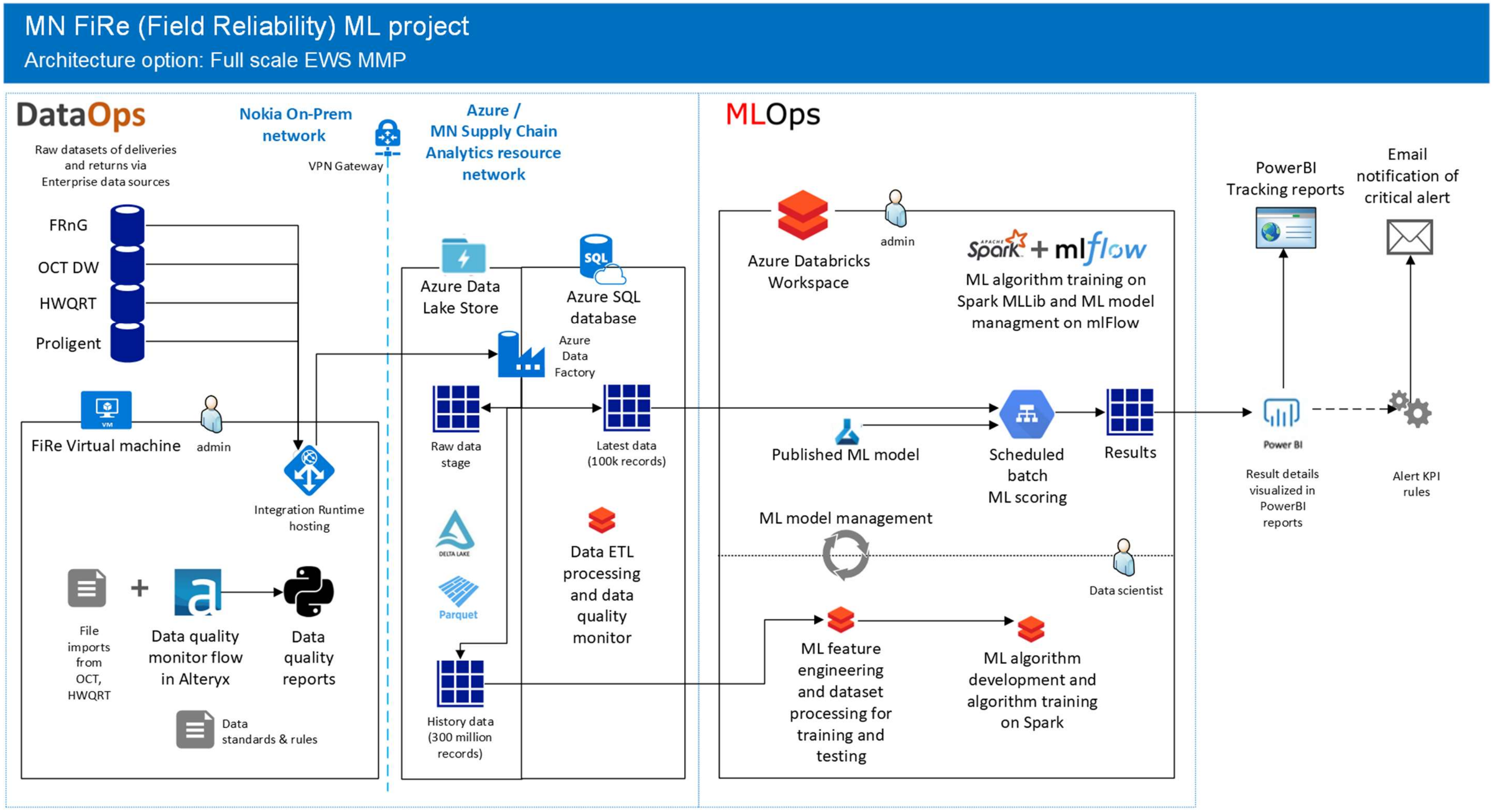
Goals for providing security and compliance for analytic assets

The assumption is that the company or organization has a Chief Information Security Officer (CISO) and perhaps a Chief Compliance Officer or Chief Risk Officer and that the goals below are supplementary to the organization's security and compliance policies and procedures.

Some specific goals for security and compliance related to analytics include:

- Protecting data assets used in, and produced by, analytics should be integrated into the organization's security plans, policies, procedures and controls. By protecting data, we mean protecting the confidentiality, integrity and availability (Information Technology Laboratory (National Institute of Standards and Technology), 2004) of data assets.
- The analytic group should work with the company's inside or outside counsel so that the collection of data assets, modelling practices, and the deployment of analytic models are compliant with all relevant local, state and national and international laws, regulations and policies.
- As the size of data grows, it is important to make greater and greater use of automation and continuous monitoring (Dempsey et al., 2011) to ensure that data is being properly protected and relevant policies, procedures and controls are being followed.
- Analytic security and compliance should cover not only analytics within the company, but also the compliance of data made available by the company to third parties through service contracts and other contractual relationships.
- When the company sells data to third parties, then it is important that there is a system for monitoring how third parties use the data and whether it is consistent with the terms of the sale. (Grossman, 2018, p. 5)

Appendix 2. Architecture description of MVP cloud version

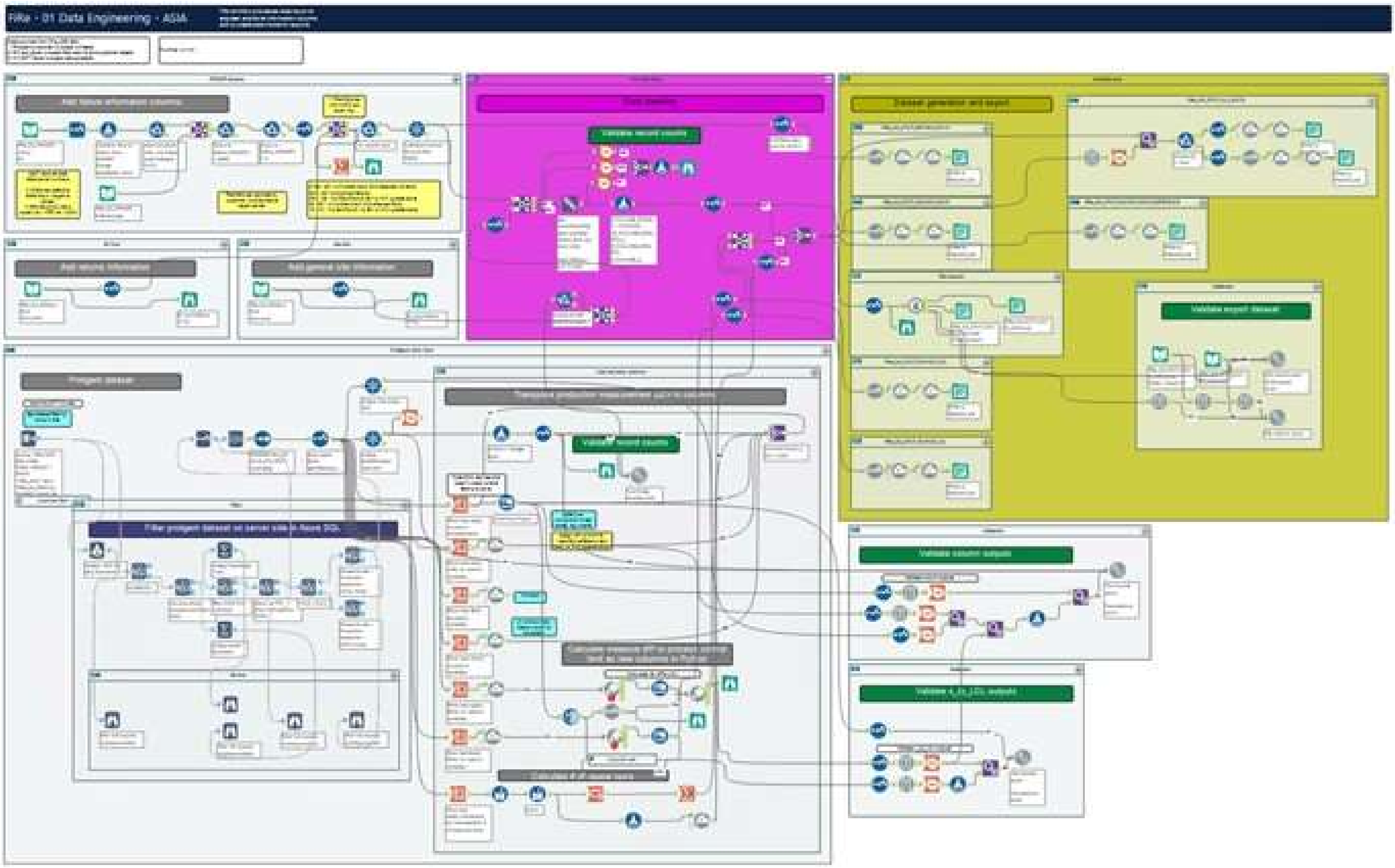


MN FiRe (Field Reliability) ML project

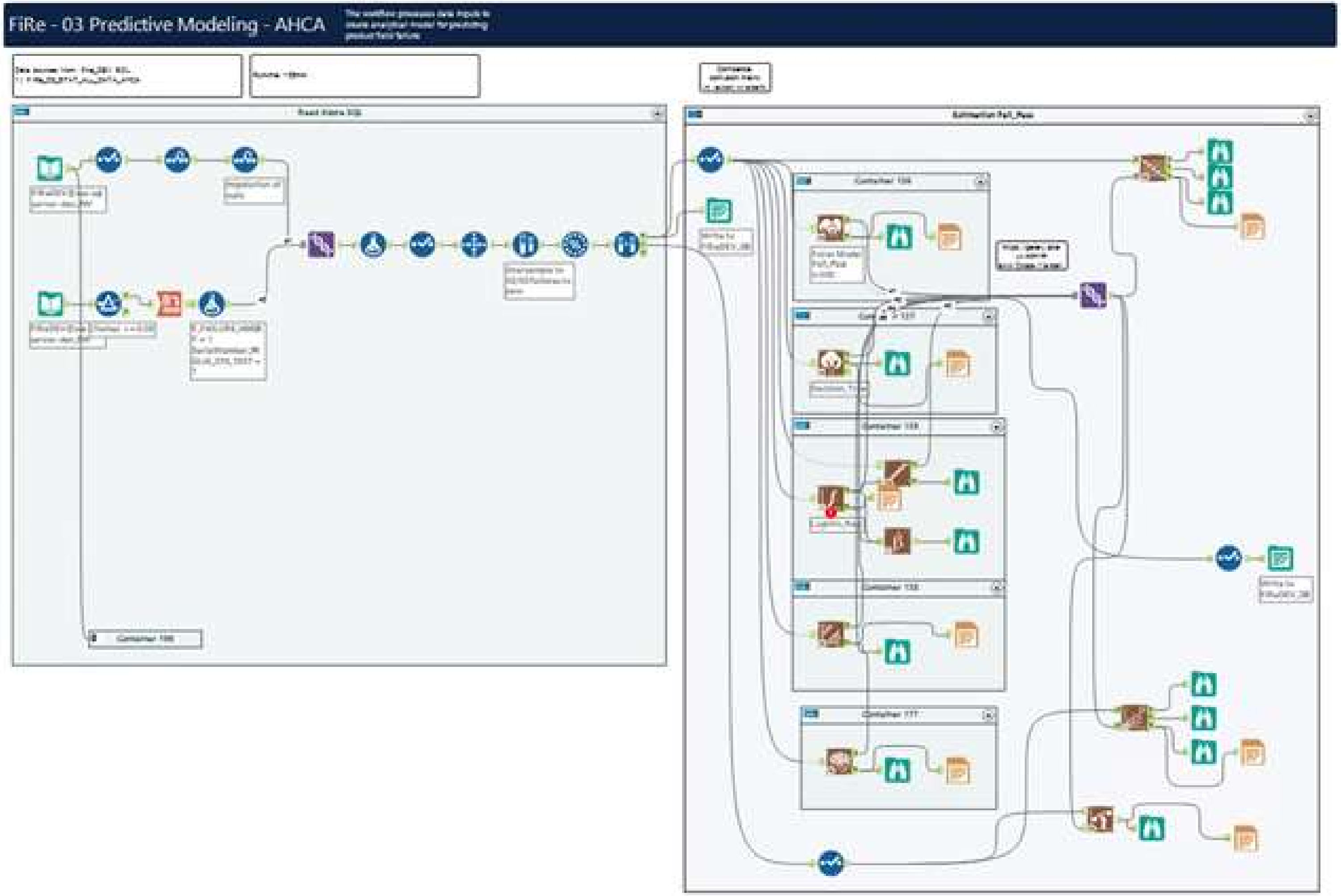
Architecture option: Full scale EWS MMP

DataOps
Raw datasets of deliveries and returns via Enterprise data sources
FRnG
OCT DW
HWQRT
Proligent
Nokia On-Prem network
VPN Gateway
Azure / MN Supply Chain Analytics resource network
FiRe Virtual machine
VM
admin
Integration Runtime hosting
File imports from OCT, HWQRT
+
Data quality monitor flow in Alteryx
Data quality reports
Data standards & rules
Azure Data Lake Store
Azure SQL database
SQL
Azure Data Factory
Raw data stage
Latest data (100k records)
DELTA LAKE
Parquet
Data ETL processing and data quality monitor
History data (300 million records)
MLOps
Azure Databricks Workspace
admin
Spark + mlflow
ML algorithm training on Spark MLLib and ML model managment on mlFlow
Published ML model
Scheduled batch ML scoring
Results
ML model management
Data scientist
ML feature engineering and dataset processing for training and testing
ML algorithm development and algorithm training on Spark
PowerBI Tracking reports
Email notification of critical alert
Power BI
Result details visualized in PowerBI reports
Alert KPI rules

Appendix 3. Alteryx workflow example of complex process: data clean-up, dataset blending, transformation, statistics and quality reports


FiRe - 01 Data Engineering - ASIA
Add failure information columns
Add record information
Add general site information
Validate record counts
Dataset generation and export
Validate export dataset
Filter progent dataset on server side in Azure SQL
Transpose production measurement pairs to columns
Validate record counts
Validate column outputs

Appendix 4. Alteryx workflow example of ML prediction model training, validation, and result comparisons


FiRe - 03 Predictive Modeling - AHCA