IDT403: Lecture Notes

Practical data analysis and machine learning in manufacturing

Unit objectives

- **Understand** what Industry 4.0 is, and the **implications** for the manufacturing sector.
- **Understand** what **role** data science and machine learning play in Industry 4.0.
- **Remember** the **differences and similarities** between data mining, data science and machine learning.
- **Remember** the **skills and roles** required to implement Industry 4.0 in the manufacturing sector.
- **Understand** what machine learning (ML) is and **identify** the main ML techniques.
- **Understand** the **limitations** of ML and **identify** the main types of data used by ML.
- **Identify** the **different phases** within the data science software development life cycle (SDLC).
- **Understand** the **value** of the **data maturity model** for organisations.

1) Introduction to data analytics in manufacturing

Welcome to the "Introduction to data analysis and machine learning course". This is an exciting and fast changing field. The course introduces the methods (algorithms), tools (systems) and skills involved in the process of gaining insights from data in the manufacturing industry. The course will cover topics within the field of data analytics in the manufacturing context.

Suggested reading

- Tibco: What is Manufacturing Analytics?
- Forbes: The Data Analytics Profession And Employment is Exploding
- Coursera: 7 In-Demand Data Analyst Skills to Get Hired in 2021

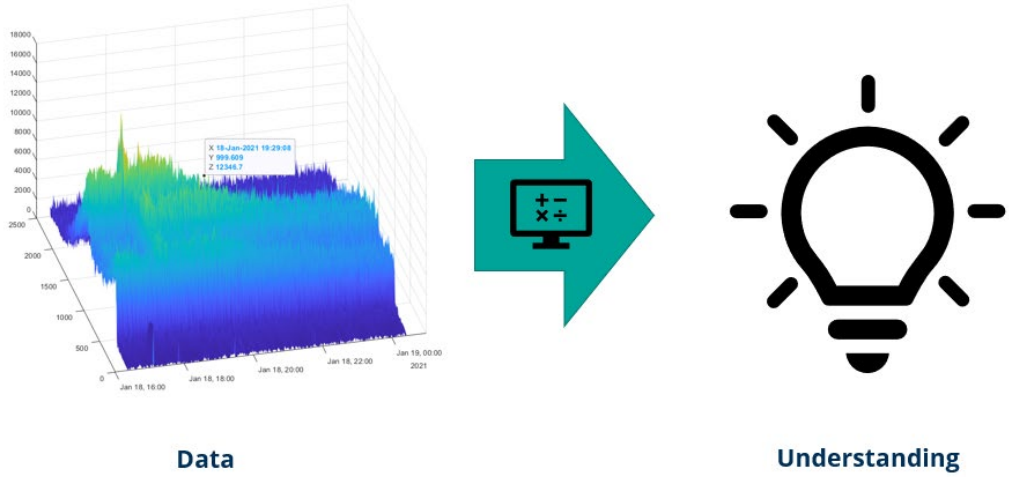
OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Published articles describing the use of analytics, data driven approaches, machine learning and other related topics often do not explain how people (skills), processes and tools (systems) are utilised to obtain the intended results. These articles sometimes reflect a particular perspective (often commercially motivated) which may or may not provide the insights of relevance to yourself. The course will provide you with information that will help you to navigate this field by debunking opinions expressed in online literature.

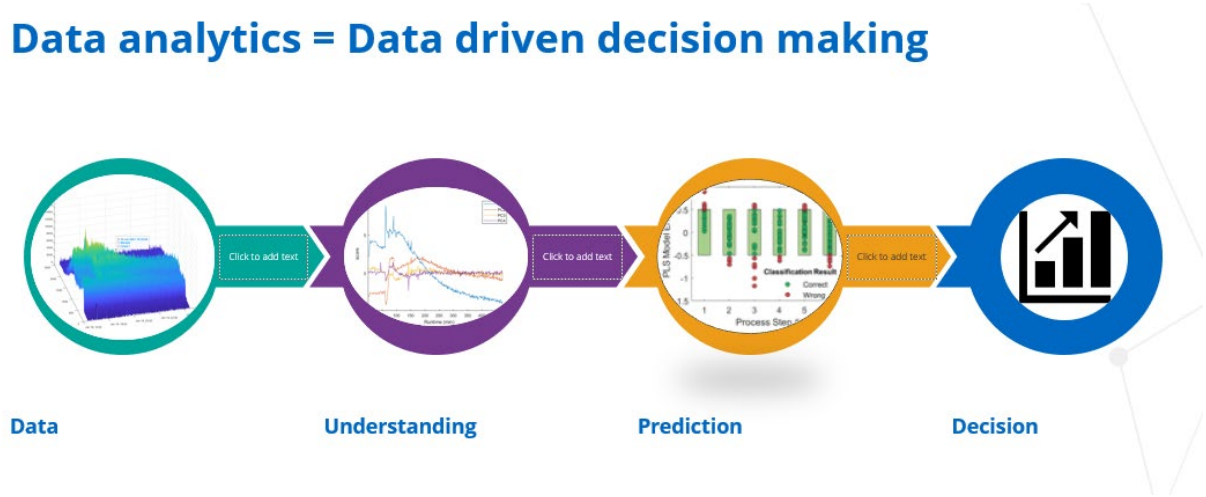
Technical terms are sometimes used out of context during technical discussions. Misunderstandings regarding terminology, processes and tools can lead to significant problems in executing projects. Data analytics projects require stakeholders with different roles and backgrounds to have a common understanding of the field. It is suggested that you discuss the terms that you are unsure about with your peers and tutor during the course. In your work or study context it is important to communicate clearly and to clarify uncertainties that may arise. Keep your communications as succinct and simple as possible.

Data analytics essentially describes the acquisition, storage, processing, and most importantly, the utilisation of data, to drive business processes and business decisions. The Gartner glossary defines "data and analytics" as: "data and analytics is the management of data for all uses (operational and analytical) and the analysis of data to drive business processes and improve business outcomes through more effective decision making and enhanced customer experiences."

What is the role of data analytics in Industry 4.0?



OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0



Manufacturing analytics is the use of operations and events data and technologies in the manufacturing industry to ensure quality, increase performance and yield, reduce costs, and optimize supply chains. Use cases include the following:

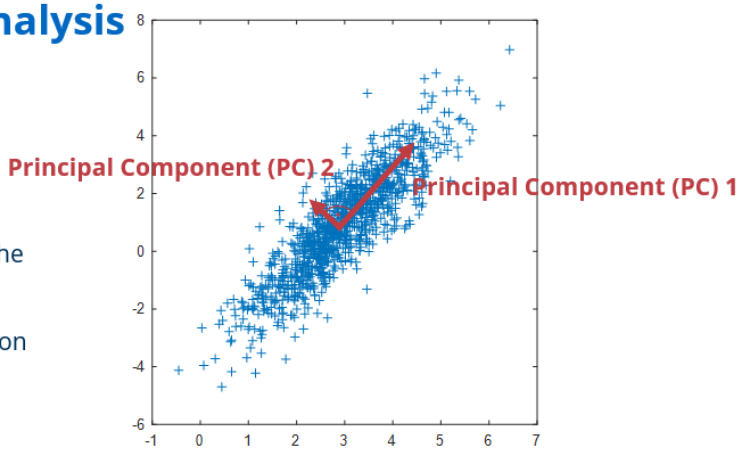
- Supply chain
    - Demand forecasting, Order management, Inventory optimization, Supplier performance, Transportation analytics, Early warning systems
- Product quality
    - Real time quality monitoring, Root cause, Reliability, Warranty
- Field service and support
    - Inventory management, Supplier performance, Transportation analytics
- Creating an efficient factory
    - Real time equipment and process monitoring, Process capability, Optimize maintenance, OEE and factory productivity

Different analytic methodologies are typically utilised within the industry 4.0 context. These are broadly speaking as follows:

Statistical Data Analysis	Mechanistic Modelling	Machine Learning / AI
Understand the math	Understand the system	Have lots of data
Apply statistical methods to find a more meaningful representation of the data	Build a mathematical representation of the actual system	Find structure and patterns in large amounts of data and apply that pattern to new data

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Statistical analysis example: Principal Component Analysis



Reduced representation of data capturing the maximum amount of variance

Purely algebraic, no modelling or optimization involved, simple, fast, robust

Mechanistic Modelling example: How much water is flowing down the river?

More than 10^{25} molecules of water flow down the river, impossible to measure all of them individually

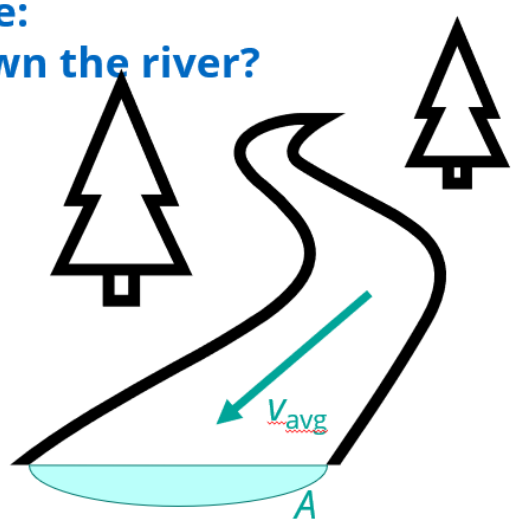
Build a mathematical representation of the real system

1st step: simplest model $\frac{dV}{dt} = v_{avg} \cdot A$

2nd step: refine $\frac{dV}{dt} = \int v(x) dA$

3nd step: refine more $\frac{dV}{dt} = \int (v(x) - \tau(x)) dA$

....



Modelling both gives us a way to handle the data and improves our understanding of the system

"All models are wrong, but some are useful"

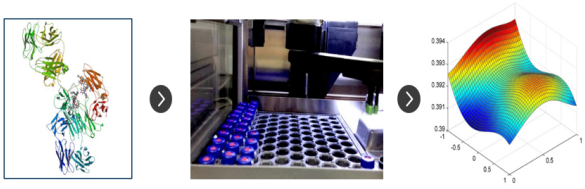
Some case studies of machine learning (ML) in in manufacturing based on these use cases are as follows:

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Machine learning in manufacturing

CASE STUDY

Machine learning to lower risk of development failure and increase efficiency of the supply chain



Using advanced analytical techniques (ML), CPI was able to model data in order to establish a rapid prototyping capability for mAb based on important properties such as colloidal and conformational stability as well as propensity to aggregate during process development.

Outcomes and impact

- *Applying this approach increases the ability to identify failing biopharmaceuticals earlier in development allowing for suitable drug candidate selection.*
- *This enables biopharmaceutical drug developers to test the suitability of their drug molecules earlier and reduce risk of development failure. Efficiency for the drug commercial supply chain is guaranteed.*

cpi

For funding
AMSCI

Collaborative project

6 months

£11.3m

Machine learning in manufacturing

CASE STUDY

Machine learning to monitor the freeze-drying process of substances

Freeze drying enables long term preservation and storage of the substance. Using the advanced machine learning techniques CPI was able to analyse spectra data from novel instrumentation to enable the freeze-drying of biological medicines.

Outcomes and impact

- *This approach enables the identification of process points which helps monitor complex freeze-drying processes.*
- *Allows for prompt interventions during process anomalies to avoid batch loss.*
- *Leads to more efficient processes and demonstrates that this approach can be used to manage downstream processes.*

AtlasBio
Analytical Technologies for the Stabilization of Biopharmaceuticals

Innovate UK

Collaborative project

36 months

£557K

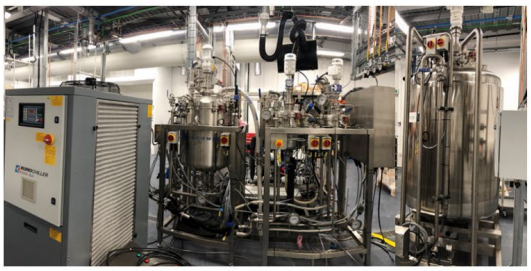
OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

CATAPULT High Value Manufacturing

Machine learning in manufacturing

CASE STUDY

Innovative tools that allow Advanced Process Control models to be rapidly built, maintained and reapplied.



Test the latest digital manufacturing solutions in a broad range of applications, without investment. Quick-fire projects in industry-like environment with expert support, and reduce the risk before they are deployed in your own facilities.

Using advanced modelling techniques to allow rapid development and validation of inferential soft sensors, chemometric calibration models and model predictive control.

cpi | INDUSTRIAL STRATEGY | UK Research and Innovation | MADE SMARTER SMART FACTORY INNOVATION HUBS

Collaborative project

6 months

£200k

Outcomes and impact

- *Reducing Waste and Increasing Productivity in Development and Manufacture.*
- *Demonstration of innovative tools that allow APC models to be rapidly built, maintained and reapplied across the manufacturing.*

The amount of work and types of skills required to successfully execute a data analytics project are often underestimated. There is a significant technology cost to projects due to the complexity of unforeseen factors such as: acquisition of data, setting up data pipelines, changes to internal business processes and human change management requirements. Apart from these often-unconsidered factors, analytic tools (systems) utilised could also include several different approaches, which means decisions need to be made regarding procurement. Different toolsets are required for different types of analytic services such as: ad-hoc analyses, scheduled reports, or real-time alerts.

Different organisations have different levels of “data maturity” which impacts on analytic projects. Maturity levels range from no formal information systems and processes in place to complex information systems such as data warehouses and bespoke data marts being in place to provide insights to the respective business units. Availability of existing systems increases the productivity of new and ad-hoc resources. Building basic information management systems before proceeding with extracting business insights should be factored into data analytic projects. Trade-offs between different analytic solutions should be considered in terms of time to create, cost to create, reproducibility, and accuracy. Depending on your needs some approaches may be better than others.

These are some of the themes we will be expanding on during this unit and will allow you to execute manufacturing analytics projects of your own successfully.

Suggested reading

- Confluent: Databases, data lakes and data warehouses compared

Data analytics can mean different things in different contexts, depending on the industry, data maturity, infrastructure, and resources available within an organisation. The data analytics field is ever-changing with new methodologies and tools being introduced to the field at a high rate. The principle of the underlying science however remains largely the same. The following excerpt from a presentation at the “Tukey Centennial workshop” in 2015 expresses this sentiment:

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

"More than 50 years ago, John Tukey called for a reformation of academic statistics. In `The Future of Data Analysis', he pointed to the existence of an as-yet unrecognized science, whose subject of interest was learning from data, or 'data analysis'. Ten to twenty years ago, John Chambers, Bill Cleveland and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland even suggested the catchy name Data Science" (Donoho, 2015)

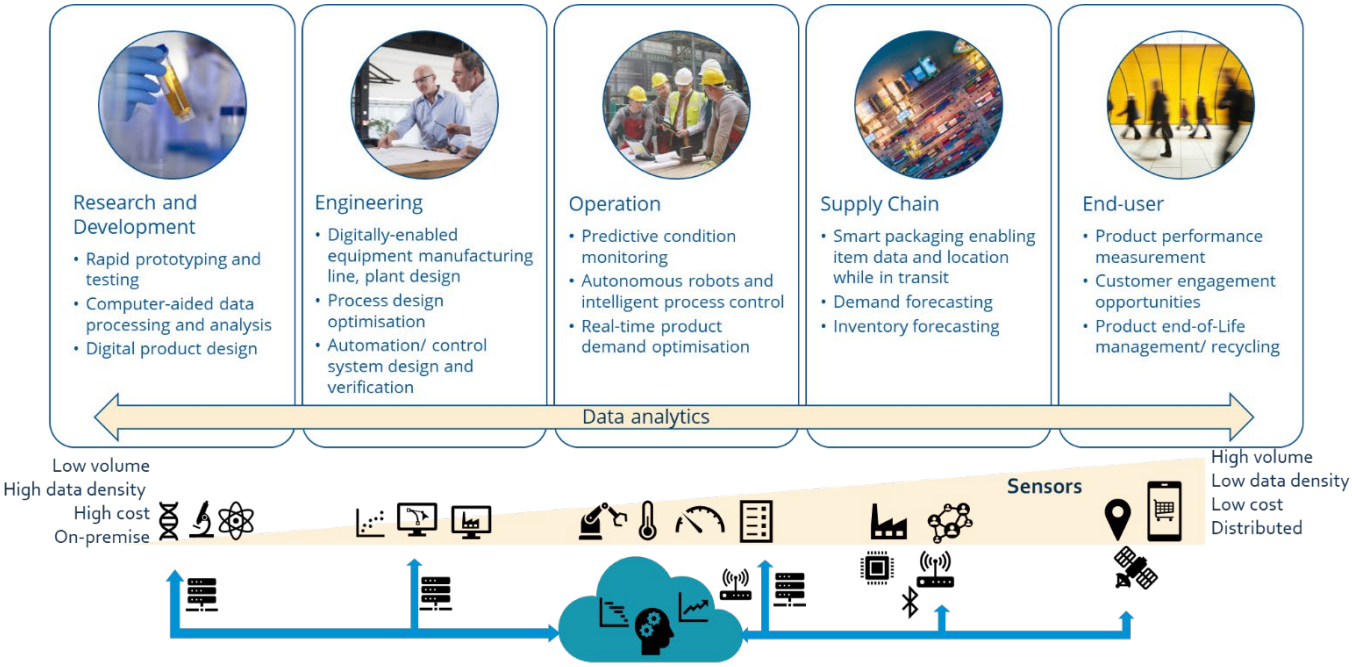
As you can see from the excerpt, some of the *new* advances in the field during the past 2 decades, such as the birth of the field called "data science" itself, has in fact been around for much longer than expected.

One theme that remains constant is that data can be used to make better decisions. To make better decisions one typically needs to be better informed and capable of acting on information. Raw data needs to be transformed into information. Better access to information and automated decision capabilities can assist with efficiency in making decisions. Manufacturing analytics is a broad term that encompasses various methods capable of transforming data into insights that then can drive desirable business outcomes.

Suggested reading

- CIO: What is data analytics? Analysing and managing data for decisions.

Manufacturing analytics has the exciting prospect of supporting end to end manufacturing optimisation. The following idealised diagram shows how data is generated and utilised at the various points in the manufacturing process from research and development through to end-user usage of the product.



The diagram depicts a state where data is efficiently captured, transformed, analysed, and disseminated at all the different stages of manufacturing. This is currently not a reality in most manufacturing processes but is what Industry 4.0 sets out to achieve.

A scenario where data is managed as depicted has several advantages, in that it allows you to:

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

- Learn faster, focus on what matters
- Leverage learning in research and development for process design, development, and operations
- Ensure manufacturability early on
- Improve economics, reproducibility, flexibility of manufacturing assets
- Efficient production, enhancing quality and reducing capital outlays and overall cost of goods
- Improve availability of products
- Perform effective ongoing process optimisation
- Implement modularized, multi-purpose manufacturing processes
- Implement machine learning assisted performance condition monitoring
- Integrate supply chains and distribution with overall process

To achieve such a system, one must understand the different analytic elements required to build a supportive analytic system. One can start by understanding the basic four types of analytics described in literature, each with a different focus:

Descriptive analytics: What has happened? This area is historically known as *business intelligence (BI).*

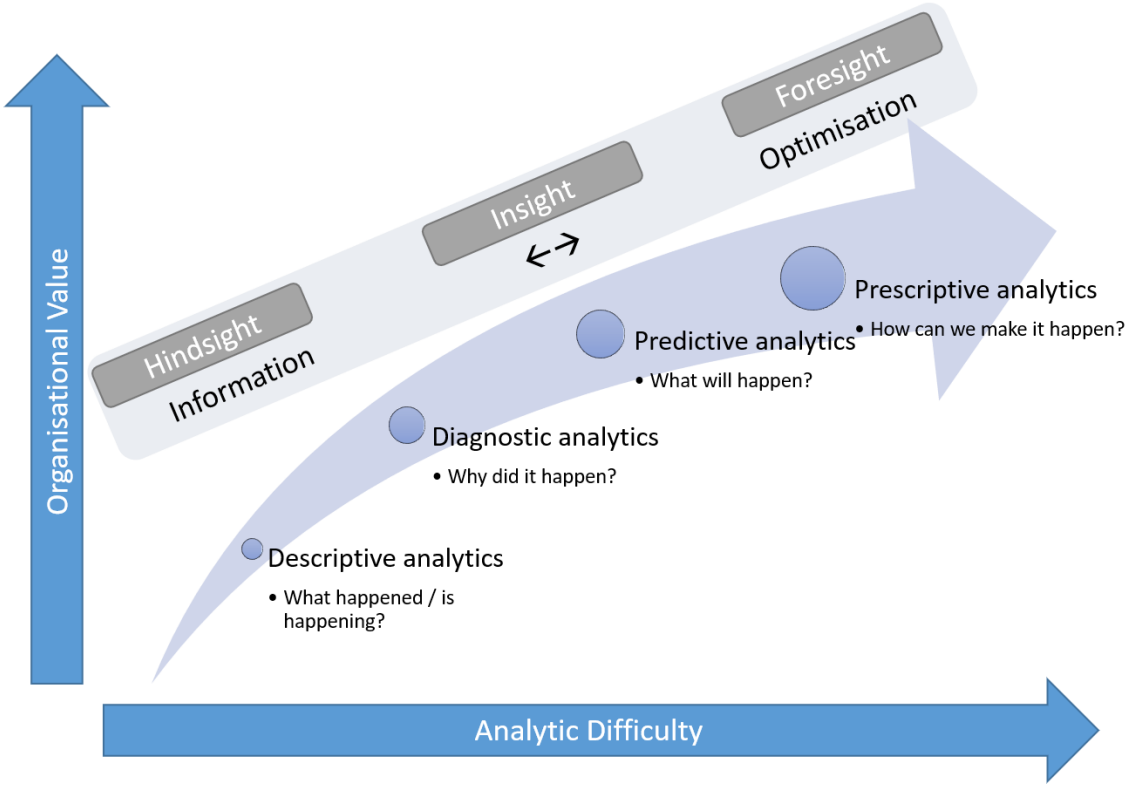
Diagnostic analytics: Why did it happen? This area focuses on explaining the reasons for observations made using *descriptive analytics*. It is also known as interpretation on insights derived from data.

Predictive analytics: What will happen in future? This area is often referred to as *advanced analytics* and typically depends on advanced statistical techniques combined with machine learning and deep learning.

***Prescriptive analytics*:** What needs to be done? This area recommends solutions that will deliver desired outcomes. Similar techniques to predictive analytics are used, but with a different objective. Examples are recommender systems or rule based *decisioning systems*.

All these types of analytics are required by a manufacturing process to achieve Industry 4.0. The different types of analytics are often used within a framework or roadmap to better understand the journey to insight and foresight (Industry 4.0 in the case of manufacturing) for an organisation. Higher levels of data maturity enable decision making utilising the best possible information available. We will expand on this concept in Section 4.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0



Different categories of *Analytics* practised in industry

Data driven decision making is made possible by advances in technology, increase in amount of data generated and stored digitally, and advances in ML/ AI systems. Technology and processes supporting the analysis of structured data have been in use for decades. Nonetheless effective management of data and elicitation of information remains challenging due to the complexity of the field. Industry 4.0 requires effective processing of *unstructured data* which increases data management challenges. Unit 2 focuses on different types of data and how to best manage these resources.

The past decade has seen unprecedented advances in the field of data analytics. Much of this advancement is within the ML/ AI field and was driven by scientific research. The rate of adoption within industry has however been slow as organisations struggle to transition from experience-based decision-making processes to data driven approaches developed within the academic area. This is particularly true for manufacturing organisations on the road to Industry 4.0.

Organisations typically cannot justify the time and resources required to develop new decision-making frameworks. In business the focus is typically on profit, while the academic focus is on knowledge. Existing business systems and processes are typically very advanced and costly to build and maintain with respect to extracting information from data in its original format, which drives decisions. To transition these systems (change management) to rely more on data-driven decision making is a costly and time-consuming process.

It is the opinion of many that the next decade is likely to focus on “decision science” rather than “data science” i.e., the ability to transition an organisation to data-driven decision making. Large organisations such as Google are already starting to employ “decision scientists” in addition to “data scientists”.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

It is important to note that data science is rooted in the "scientific method", combined with existing fields such as data mining and data analytics. The following reading on the scientific method is recommended as additional material:

Suggested reading

- Purdue: The Scientific Method
- OpenLearn: Science, the scientific method and scientific laws and Origins of the scientific method

The fields of data mining, data science and machine learning contain a large degree of overlap. One such overlap is the similarity of project phases within each field. These include the acquisition of data, the preparation of data, analysing the data and finally presenting and acting on insights gained. We will view data science as a central topic in this course. The project phases of data science are presented in a following section, but a high-level overview of the disciplines is presented here.

Suggested reading

- *Data analytics*

Data mining is a field that has been around for as long as databases have been around i.e., at least 3 decades. The objective of data mining is to extract value from data in ways that was not originally intended when gathering the data and prepare the data for future use. Tools used are typically statistical, machine learning and data management e.g., data warehousing. Data utilised are always structured and tabular. The work is typically performed by data scientists and data engineers. Many of the functions relate to data processing and statistical knowledge. Classical examples of data mining are fraud detection and credit scoring from financial or insurance data, marketing campaign targeting, customer attrition and churn analysis from sales and demographic data. During the early years of data mining there was a heavy reliance on Multiple Linear and Logistic Regression techniques on large datasets. During later years this was replaced by newer approaches such as Neural Networks. Data mining typically utilised structured data. Data Mining typically incorporates the following fields as applied to mainly structured data found in databases:

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Machine learning typically refers to the use of algorithms to learn or model patterns found in data. Algorithms are used instead of direct programming to achieve a particular goal. Data preparation and transformation of data still require programming, but an ML application's main logic is implemented by a model derived from data. Hence the term "data driven". ML has seen huge development during the past decade after breakthroughs in the ability to process large amounts of unstructured data such as images, videos, and text. Although ML has very successfully been implemented on unstructured data, it has proven to be as effective on structured data and is often used in this setting if the use case is appropriate.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Data science is in many ways a modern version of the field of "data mining". The field however includes unstructured data and the use of various new tools developed during the past decade. It can be viewed as a set of tools and approaches that can be applied to solve a wide variety of problems. It combines skills from programming, mathematics and statistics, visualisation, logic, business and domain knowledge, and finally soft skills. The diagram by Drew Conway is a good representation of the various domains in involved in data science.

The definition of data science is however ever evolving and changes as the area matures. Although data scientists are expected to function across many domains, in practice working as a team rather than a single contributor (or 'unicorn') across all the domains is typically a more pragmatic approach. Many organisations are currently retraining or reskilling existing resources to function in exactly this manner i.e., as a member of a diverse team with different skills contributed to by different individuals.

For such teams to function effectively, business processes and culture needs to change. Education and the creation of an awareness of the opportunities available is an ongoing process. The culture should also be conducive to experimentation and self-learning. An aspect that is often overlooked is the complexity in starting on the journey of hiring, managing, and retraining talent in this area. Organisations often struggle to define their needs, especially when starting out their respective journeys. Working with partners with experience to assist with this journey is often a good idea.

Suggested reading

- Data Science Central: Data Science is Multidisciplinary
- Simplilearn: Data Science vs data analytics vs machine learning
- Educba: Data science vs data mining

Advanced data analysis tools and techniques have matured considerably during the past 2 decades. When introducing new processes or tools to an organisation one should consider the impact on people and existing technology. Where possible, one should build and extend on existing capabilities rather than replacing them. Data science, data mining and machine learning should be viewed as ways to extend capabilities. Each capability has its own place, advantages, disadvantages, cost, and maintenance implications.

2) Introduction to machine learning

One possible definition of machine learning is to: "Give computers the ability to learn from data rather than being programmed explicitly". Many definitions can however be found in literature:

- "Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world." – Nvidia
- "Machine learning is the science of getting computers to act without being explicitly programmed." – Stanford
- "Machine learning is based on algorithms that can learn from data without relying on rules-based programming."- McKinsey & Co.
- "Machine learning algorithms can figure out how to perform important tasks by generalizing from examples." – University of Washington

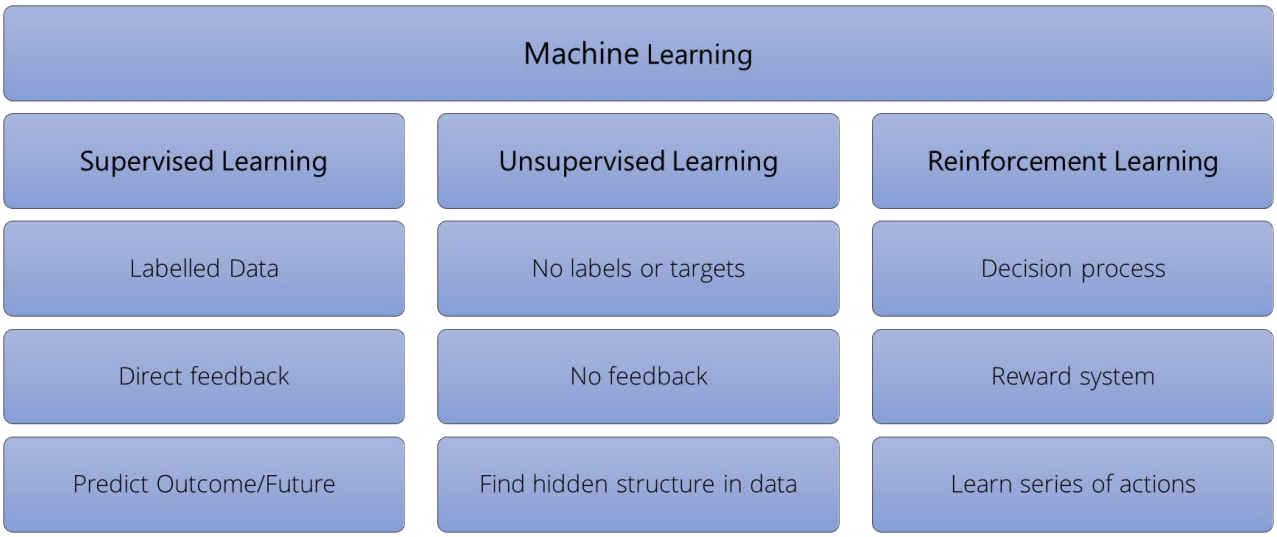
OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

- "The field of Machine Learning seeks to answer the question "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" – Carnegie Mellon University

A common element in all the previously discussed definitions is the term "data". A key characteristic of ML is that it is "data driven". Typically, a data scientist/ programmer parameterises an existing learning algorithm. The computer then receives both the parameterised model as well as a training set of data and examples, and starts learning on its own, changing the algorithm as it learns more about the information it is processing. The implication here is that it is the computer that creates the final system. It also means that the ultimate shape and nature of the system depends on data. If the quality of the data is poor, or if the data is biased, then so is the system.

The utility of this approach can be illustrated by a system built to determine whether an image contains an apple or an orange. Using the traditional software engineering approach, such a system must iterate over all the pixels in an image using a series of selection statements or rules programmed individually in advance. The drawback of this approach is that edge cases that do not fit into the rules specified will always occur, requiring extensive rework to specify the system as completely as possible e.g., some of these edge cases might be the existence of black and white images, or an image of a cross-section of an orange instead of the whole orange. ML is a good fit for such a use case. An ML model can be trained on thousands of images of apples and oranges including all the different angles and colours possible which will allow the model to generalise to images it has not seen before (provided sufficient example images are included).

Machine learning algorithms are typically grouped into three categories: supervised, unsupervised, and reinforcement learning. Many more categorisations do exist, but these cover these three cover the main areas of ML pragmatically.

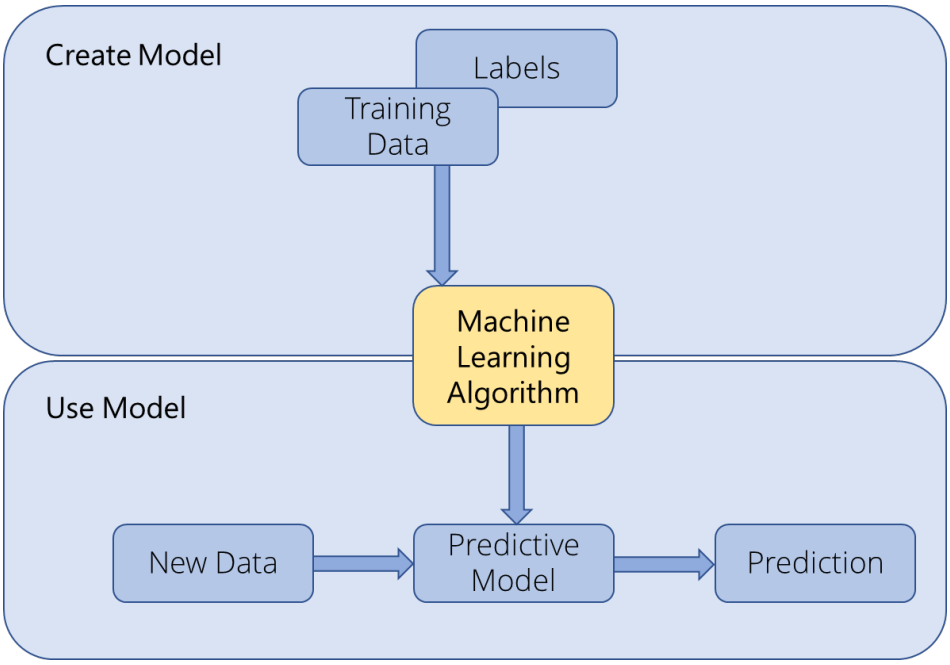


Three different types of machine learning

There is an important yet subtle distinction between supervised and unsupervised learning (more specifically, the implications of the distinction are subtle). For supervised learning the right answer is known before training the model, but for unsupervised learning the data are unlabelled or of unknown structure. Using unsupervised learning techniques, the structure of the data can be explored to extract meaningful information without the guidance of a known outcome variable. This distinction requires unsupervised learning to typically involve a large degree of interpretation by a data scientist to make sense and interpret the model results after analysis.

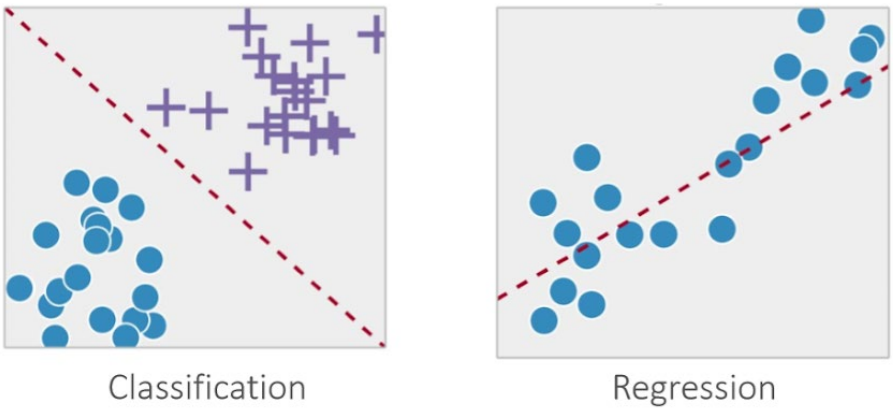
OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

In a typical *supervised* learning scenario, an outcome measurement is available, usually quantitative (such as a component lifetime) or categorical (such as component failure/non-failure). The objective of the algorithm is to predict an outcome based on a set of features such as measurements taken from sensors on a production line. A model can be trained using a supervised machine learning algorithm on such an existing dataset of labelled sensor measurements, each measurement marked as either signifying failure or non-failure. The *supervised* machine learning process therefore typically looks as follows:



Supervised machine learning

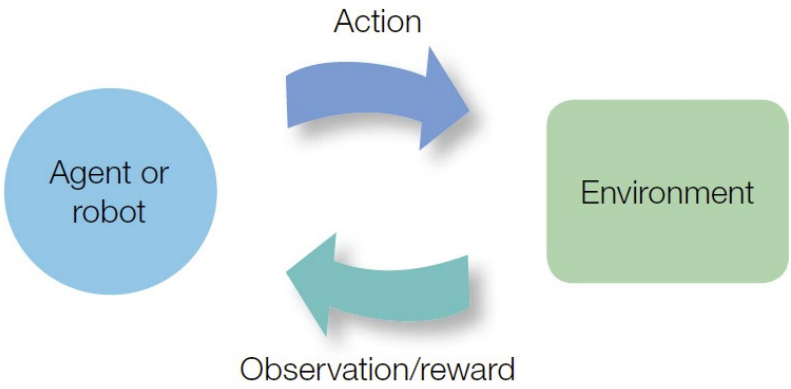
Supervised ML algorithms are typically further categorised as regression type problems, where continuous numeric outcomes are predicted, or classification problems where we categorise or classify items into two or more classes (typically also converted to discrete numeric representation).



For regression, a problem could be to predict an optimal value, a price point, or time to failure of a specific piece of equipment. The objective here would be to use existing data to predict the outcome for a specific set of circumstances. Classification problems are similar, but the predicted outcome is a category drawn from a set of pre-identified classes. While predicting multiple classes is possible, in many cases you will work with binary solutions where the solution can be one of two classes, or the question can be stated as true or false. Think about sorting solutions where you are trying to identify and separate two types of items as an example of a classification tasks.

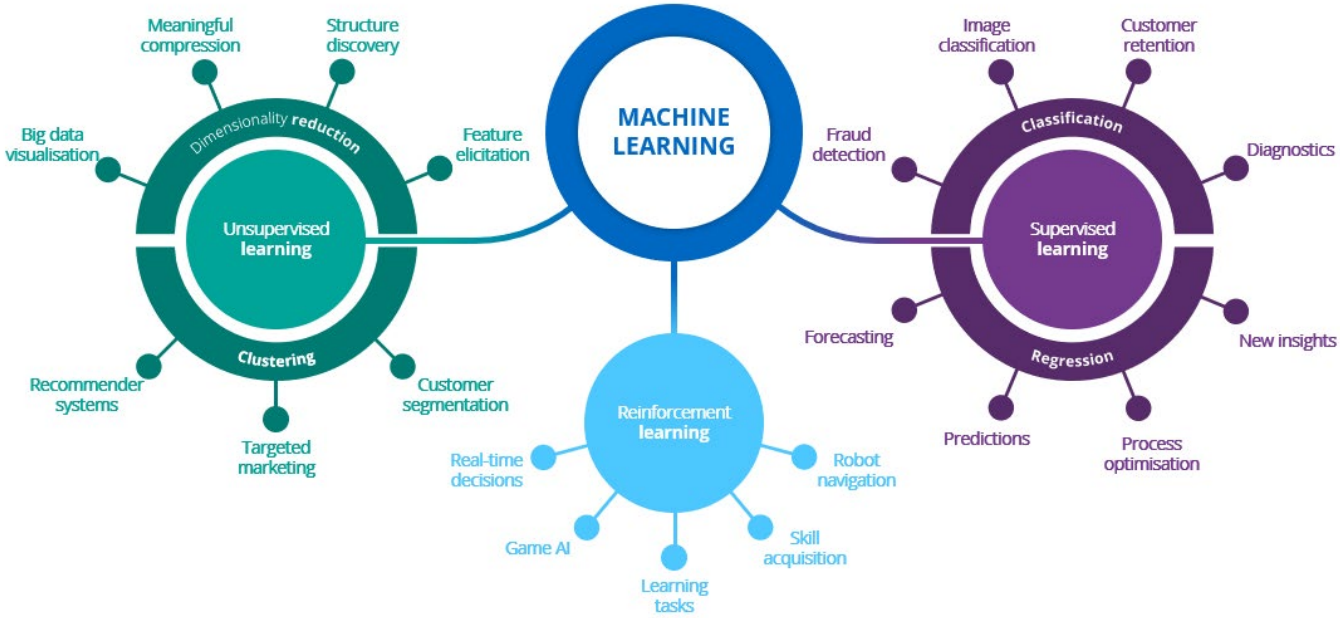
OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

For reinforcement learning, the goal is to develop a system that improves its performance based on interactions with the environment. The system typically takes the form of an agent. Since the information about the current state of the environment typically also includes a so-called reward signal, we can think of reinforcement learning as a field related to supervised learning. However, in reinforcement learning this feedback is not the correct ground truth label or value, but a measure of how well the action was measured by a reward function. Through its interaction with the environment, an agent can then use reinforcement learning to learn a series of actions that maximizes this reward via an exploratory trial-and-error approach or deliberative planning (Raschka & Mirjalili, 2017).



Reinforcement learning rewards an agent for making smaller decisions (Torres et al., 2019)

Use cases for the different types of ML algorithms are as follows:



Machine learning algorithms are often implemented as libraries of code which enables us to use computers to perform tasks without explicitly programming the tasks, as previously discussed. You still need to prepare the data, select the appropriate algorithm, and evaluate the output. If we were to consider the classification task earlier, the classification algorithm could automatically group images of two different types of components such as nuts and bolts as being separate items without having an understanding what these elements represent. Depending on the data type, data must be transformed before use by ML algorithms when:

- Algorithms require input data to be numbers (raw data often not in numeric form)

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

- Algorithms based on assumptions about the data e.g., normally distributed residuals
- Statistical noise and errors in the data need to be corrected
- Complex nonlinear relationships need to be teased out of the data

It is Important to distinguish between different types of data used for ML as they are processed differently. The two main categories of data used by ML are:

- Structured data
- Unstructured data

Structured data

- Tabular form:
  - Each variable has its own column
  - Each observation has its own row
  - Each value has its own cell
- Common format: database tables, comma separated value files (.csv), Excel files (.xlsx), XML files (.xml)
- Even though this is the easiest of all data types to work with, quality is often very poor, and most "structured data" sources do not comply with above properties but are still *treated* as structured.

Day	Outlook	Temp.	Humidity	Wind	Play tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

AI, Data Science and Machine Learning: a Dstl biscuit book

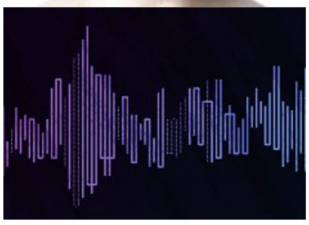
"Happy families are all alike; every unhappy family is unhappy in its own way." – Leo Tolstoy

"Tidy datasets are all alike, but every messy dataset is messy in its own way." – Hadley Wickham

Tidy Data

Unstructured data

- Text
  - Unstructured freeform text e.g., survey results, call centre logs, product reviews, social media feeds, blogs etc.
  - By using preprocessing, it is possible to analyse these along with other existing structured data
- Image
  - Includes video and image data
  - Pictures are represented by a grid of pixels
  - The pixels by themselves are difficult to interpret by non-ML systems and are classified as unstructured for data processing and storage purposes
- Sound
  - Digital signal is measured as amplitude over time which is represented digitally
  - Again, stored and managed as unstructured digital data



AI, Data Science and Machine Learning: a Dstl biscuit book

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

3) Data analytics project life cycle

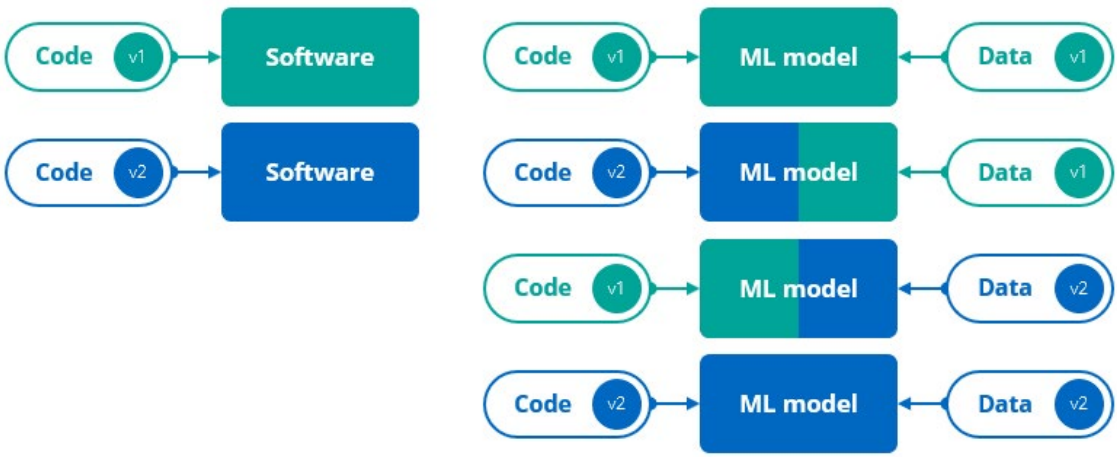
Why should data scientists use the software development lifecycle (SDLC) as data science is not software engineering? This is a good question. Although data scientists 20 years ago would have gotten away with this argument, the same does not remain true anymore. Due to the astronomic advances in the field data science is growing ever closer to software engineering in terms of systems and processes utilised. In particular, the ML development process has become:

- Complex
- Iterative
- Collaborative
- Dependent on data which changes over time

Governance policies are required to manage these complex processes, and these processes already largely exist within the software engineering field. These policies and processes include:

- Documentation
- Data privacy and access
- Data quality/ lineage
- Model QA/ test
- Model versioning/ source control
- Change control

The ML process is however different from the software engineering process in several ways. The most obvious one is the inclusion on data in the development process. The complexity introduced can best be illustrated as follows:



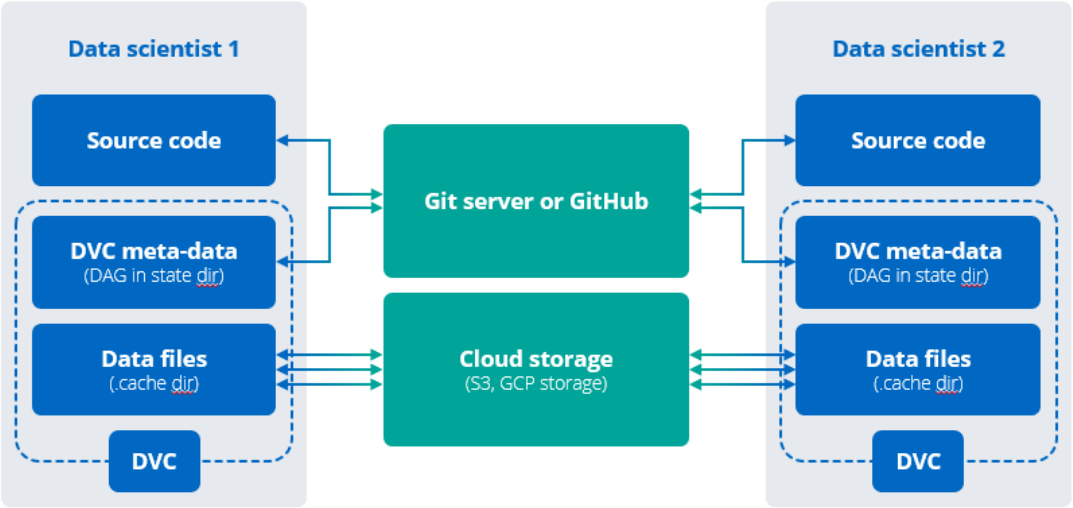
Practical MLOps: How to Get Ready for Production Models

As can be seen data is a new variable in the development process, this has the following implications:

- Complexity of process increases
- Complexity of process automation increases
- Development, testing, QA, and release times are increased
- Versioning becomes challenging

One possible version control system to manage this complexity looks as follows:

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0



The closest thing to an industry standard for a data science software development lifecycle is called the CRoss-Industry Standard Process for Data Mining (CRISP-DM). The process was published in 1999 by Tom Khabaza and was built into IBM SPSS Modeller system which is widely used as a data science workbench. The CRISP-DM process promotes:

- Business involvement at all steps of development
- Ongoing evaluation
- Measure business value at all steps

You should recognise the phases from previous articles. The execution of the life cycle would depend on the specific project. The importance, duration and resources required for each stage will depend largely on the use case and the environment. The importance lies in understanding the framework and applying it in such a manner that you add value to your business. The following are key aspects that need to be kept in mind throughout all the phases of the process (also known as the 9 rules of data mining by Tom Khabaza):

Know the problem

Business objectives are the origin of every data mining solution.

Know the business

Business knowledge is central to every step of the data mining process.

Know the data

Data preparation is more than half of every data mining process.

Decide on the tool by means of exploration

The right model for a given application can only be discovered by experiment or “There is No Free Lunch for the Data Miner”.

Find the pattern

There are always patterns.

Provide Insight

Data mining amplifies perception in the business domain.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Find value

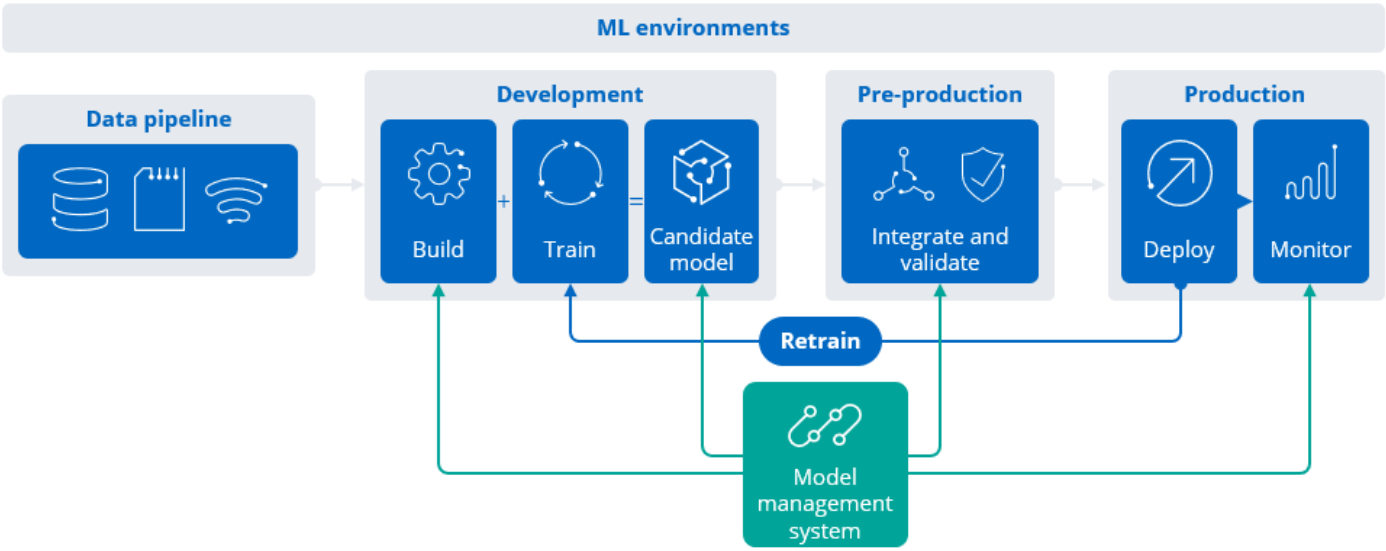
The value of data mining results is not determined by the accuracy or stability of predictive models.

Update models

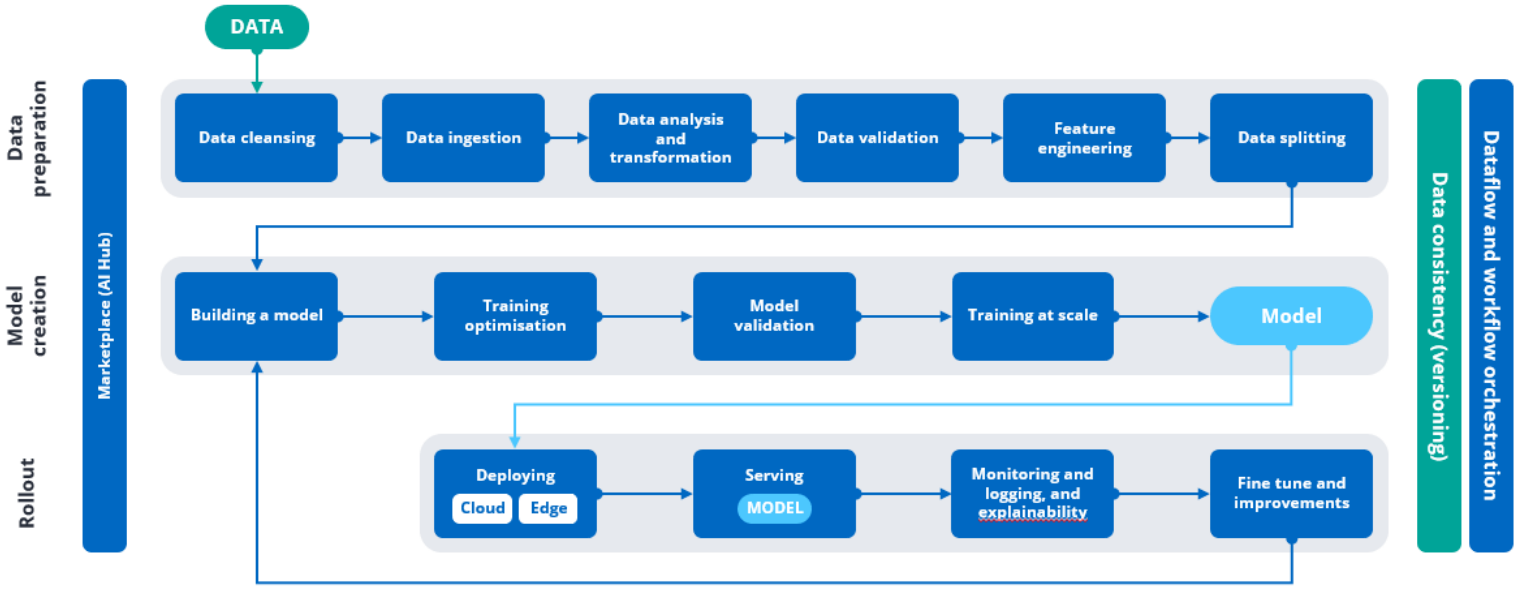
All patterns are subject to change.

ML Ops is closely related to the data science SDLC and specifies a set of practices that aims to deploy and maintain machine learning models in production reliably and efficiently.

Some examples of these are:



Nvidia ML Ops process



CI/CD Foundation's ML Ops process

Some characteristics of these frameworks are:

- Continuous Integration/ Continuous Delivery (CI/ CD)
- Model inside a container that runs on many machines

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

- Versioning data as well as code

Suggested reading

- Analytics Vidya: Introduction to Life Cycle of Data Science projects
- KDNuggets: Data Science Process
- Microsoft: What is the Team Data Science Process?

4) Organisational data maturity

A phased approach to becoming "data-driven" is generally suggested. The application of a data maturity model is useful to plan the process. A maturity model consists of different levels which can be used as goals to grow capabilities and resources whilst implementing ever more advanced use cases. Note that it is not only the tools and algorithms which need to evolve, but also the skillsets of resources, complexity of the applications and input requirements. Maturity models include classifications of types of analytics used discussed previously. The framework below is used to depict increasing organisational value against analytic complexity and difficulty.

Limited	Reactive	Stable	Pro-active	Exemplar
Level 1	*Level 2*	*Level 3*	*Level 4*	*Level 5*
• Data used rather than exploited • **Silos** and firefighting • **Low trust in data accuracy** • Ad-hoc analysis • **Spreadsheets and spreadmarts** • Transactional	• Attempts at information strategies • Data cultural difficulties • **Lack of high-level leadership** • **IT strategies rather than business strategies enabled by data**	• Siloed approach to sources and standards • Strategy and vision maturing • **Working closely with business users** • **Changes in internal processes**	• **Executive sponsors and champions** • **Business led, data-driven** • **Metrics redefined and informed by data**	• **Data central to business strategy** • Continuous improvement of processes, standards and capabilities • Boardroom representatives

The process is measured across different organisational functions, typically divided across:

- People (skills)
- Process (leadership & culture, data governance, quality, and standards)
- Technology (tools & architecture)

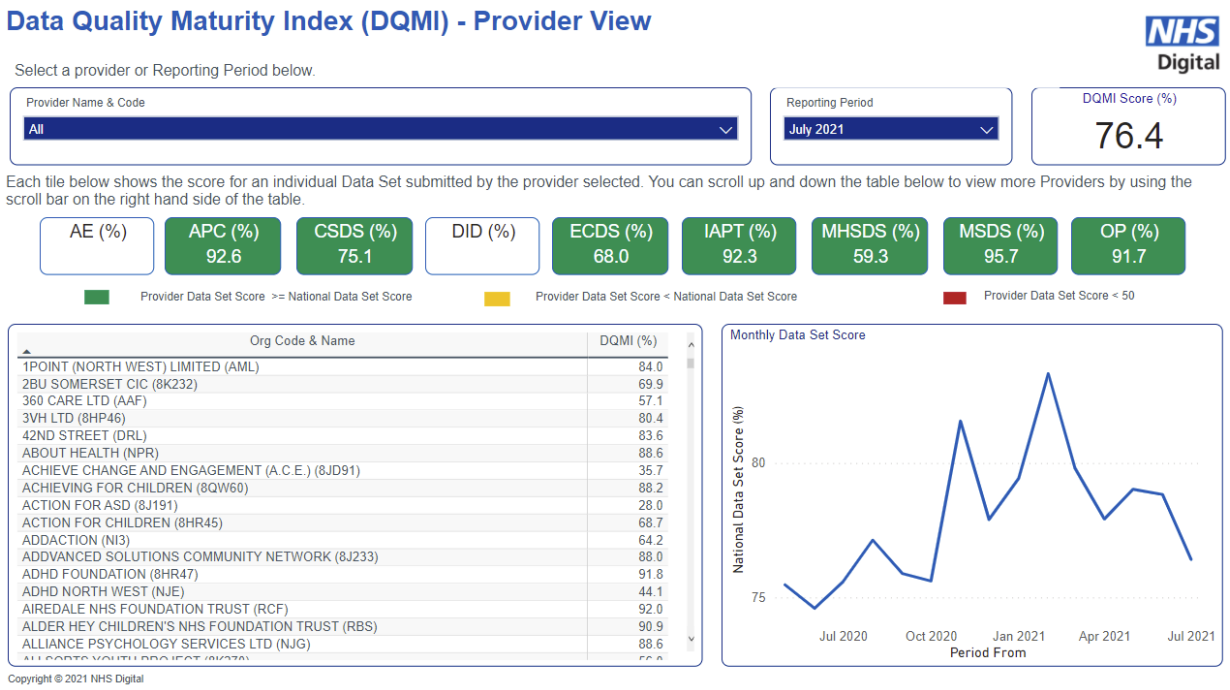
The NHS Digital Quality Maturity Index is shown in the next diagram.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

	Level 1 (e.g. limited)	Level 2 (e.g. reactive)	Level 3 (e.g. stable)	Level 4 (e.g. proactive)	Level 5 (e.g. exemplar)
Leadership and culture		Now	**Future**		
Skills	Now		**Future**		
Tools and architecture			Now	**Future**	
Data governance		Now		**Future**	
Quality and standards	Now				**Future**

The Government Data Quality Framework: guidance

The example shows that maturity should be tracked over time, with a future goal measured against the current state. Each area should identify meaningful Key Performance Indicators (KPIs), with achievable targets set for each KPI. The following image shows the NHS data quality KPI.



From the diagram it can be seen that:

- The KPI is made accessible for relevant stakeholders to address gaps in performance
- KPI is tracked in real-time
- Is transparent – all stakeholders can assess and continuously improve

This particular KPI would be one of many that make up the overall Data Quality metric for NHS, using the Balanced Scorecard methodology.

Suggested reading

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0
CATAPULT High Value Manufacturing

- Gartner [2012]
- The four stages of the data maturity model

Note that these frameworks are used as conceptual tools to understand current capabilities and future possibilities. Certain areas within a business might be more mature than others and depending on the organisation and industry, the criteria per phase may be different.

You will be introduced to additional concepts and material in subsequent units, but you should recognise that there is no single tool, role or resource that will deliver on all the needs of the organisation. As is the case with the scientific method, data analytics is an iterative process of balancing the costs and efforts of your actions against the potential business value based on the context and current state of the organisation. The iterative process stops when a point is reached where additional effort or resources can no longer justify the cost. The data maturity frameworks from Gartner and Dell indicate that more mature organisations may find it trivial to implement new use cases for predictive maintenance, whilst a less mature organisation might not be able to do so. There is more to growing your capabilities than introducing new algorithms to your organisation. The management of the approach, infrastructure and resources is also required.

Make sure to reflect on your own organisation.

- Where are you on your journey?
- Are all departments at the same level?
- What is required to move to the next level?
- Where are your personal interests/capabilities, and where does this fit into the needs/capabilities of the organisation?

6) Conclusion

In this unit you were introduced to basic concepts in the data analysis space. While many of concepts are applicable to all industries, we will start to introduce more industry specific use cases and content were applicable.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

7) Supplemental material

- *Example big data use cases in the manufacturing industry*

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

8) References

Donoho, D. (2015). 50 Years of Data Science. *Tukey Centennial Workshop*, 1–41.

Raschka, S., & Mirjalili, V. (2017). Python Machine Learning. In J. Huffman & S. Huai-En (Eds.), *Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow* (Second, Issue June). Packt Publishing.

Torres, M. T., Hart, G., & Emery, T. (2019). *The Dstl Biscuit Book Artificial Intelligence, Data Science and (mostly) Machine Learning*. 1–40.

Hastie, T. et. all. (2009). *Springer Series in Statistics The Elements of Statistical Learning*. *The Mathematical Intelligencer*, *27*(2), 83–85.

VanderPlas, J. (2017). *Python Data Science Handbook*. O'Reilly.

Jason Brownlee (2021), *Machine Learning Mastery*

Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly.

Valohai. (2020). *Practical MLOps: How to Get Ready for Production Models*. 52.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

9) List of further courses

Course	Topic	Level
Making Friends with Data Science – Free	Very accessible and entertaining free YouTube resource created by Google and designed to give everyone no matter what their background the tools for effective participation in machine learning for solving business problems and for being a good citizen in an increasingly AI-fuelled world.	Beginner
Programming for Data Science with Python - Udacity	Learn the programming fundamentals required for a career in data science. By the end of the program, you will be able to use Python, SQL, Command Line, and Git.	Beginner
Machine Learning – University of Stanford/ Coursera – Free	Introduction to machine learning. Teaches the theoretical underpinnings of machine learning and how to apply these techniques to new problems.	Beginner
Data Science with Python – University of Cape Town/ GetSmarter/ U2	Explore supervised learning using tree-based models and neural networks, as well as unsupervised learning using K-means and hierarchical clustering. You'll also learn about the process of revealing more robust patterns to ensure models are useful.	Beginner
Applied Data Science with Python Specialization – University of Michigan/ Coursera	The 5 courses in this University of Michigan specialization introduce learners to data science through the python programming language.	Beginner/ Intermediate
Become a Data Analyst NanoDegree – Udacity	You'll learn to manipulate and prepare data for analysis and create visualizations for data exploration. Finally, you'll learn to use your data skills to tell a story with data.	Intermediate

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

10) Copyright notices

Various copyrights utilised in these lecture notes:

Reference	Copyright
The DSTL Biscuit Book	http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3
NHS Digital Quality Maturity Index	https://digital.nhs.uk/about-nhs-digital/terms-and-conditions
The typical framing of a Reinforcement Learning (RL) problem	CC0 1.0 Universal Public Domain Dedication

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0