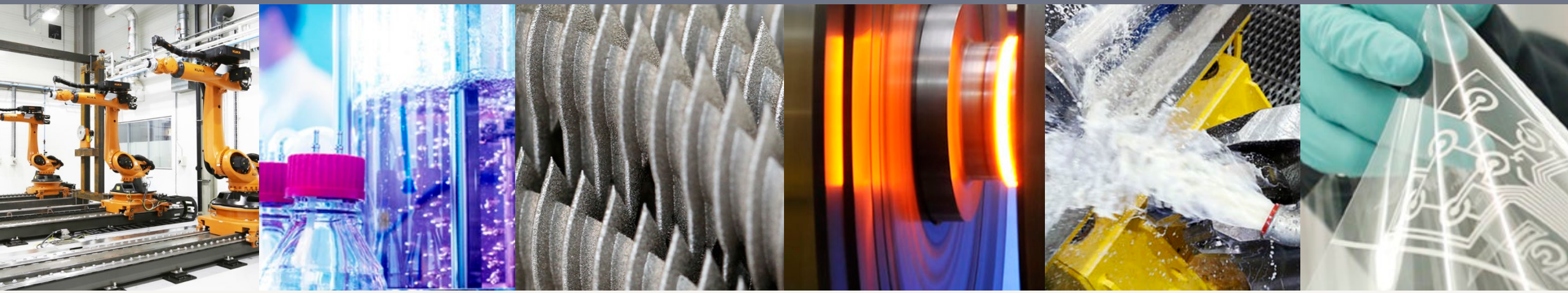
Emerging Skills Programme

Open Government Licence Guidelines

Emma Carroll-Walsh
Programme Exec
27/05/2021



What Is Open Government Licence (OGL)

Basic Guidelines

- User is free to copy, publish and transmit the information
- User is free to adapt the information
- User is free to exploit commercially and non-commercially
  - Combine it with other information
  - Include their own product or application
- User must acknowledge source of information by linking to any attribution statement specified by the provider
- Cannot use material in any way that suggests official status or that the provider endorses you or your use of the information
- Information is licensed 'as-is' so has no cover of warranties, obligations or liabilities
- Must follow UK Government Licensing Framework (UKGLF)

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

IDT403 - Introduction to data analysis and machine learning

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

Unit 1: Practical Data Analysis and Machine Learning in Manufacturing

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

Learning Objectives

1) Understanding practical data analysis: Processes and tools	
a. Understanding of data science, data modelling and data management tools as part of digital manufacturing;	☑
b. Understand what makes collectable data valuable and improve manufacturing processes for industry and industrial automation;	☑
c. Application of suitable data protection protocols including regulatory requirements, data encryption, and data protection.	
2) Understanding the role of data ethics, regulation and data protection	
a. Basic knowledge of ethics and data protection in manufacturing industry;	
b. Understanding principles, tools and techniques of ethical data use and data protection implications.	
3) Knowledge of machine learning	
a. Knowledge of taxonomy and principles of machine learning;	☑
b. Understanding principle of machine learning and the difference between supervised and unsupervised learning.	☑
4) Ability to use machine learning algorithm and tools for digital manufacturing process	
a. Understanding machine learning tools and programming;	☑
b. Be able to manipulate a dataset through RStudio, Python or others.	

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Unit Objectives

- ***Understand*** what Industry 4.0 is, and the ***implications*** for the manufacturing sector.
- ***Understand*** what ***roles*** 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.

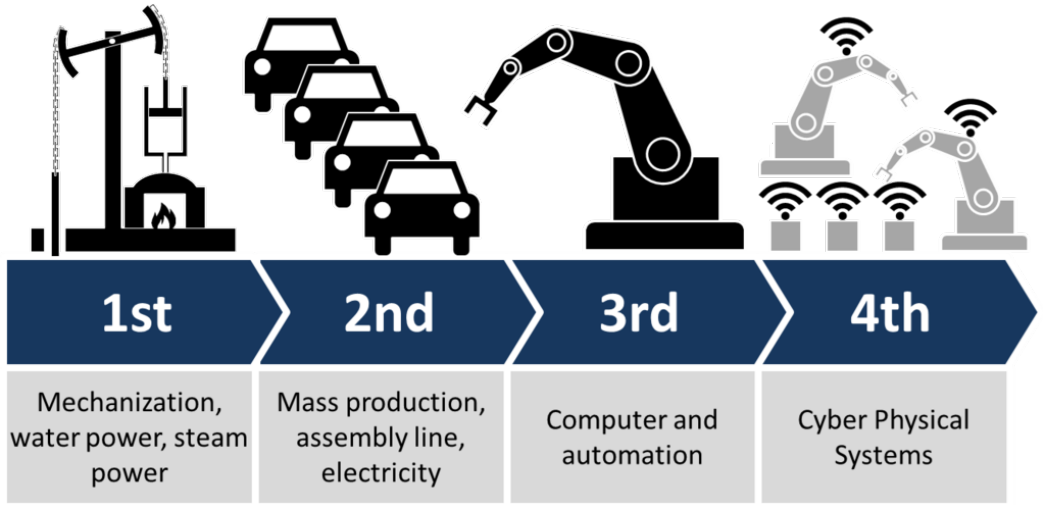
OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Introduction to Data Analytics in Manufacturing

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

What is Industry 4.0?

Industry 4.0 is smart work environments and spaces comprising of cyber-physical systems that communicate and cooperate with each other, software services and humans across the value chain real time or via the internet.



Digital manufacturing in the context of Industry 4.0 from R&D, through manufacture and to aftersales services

Smart application domain areas:
Healthcare, city spaces, transportation, smart homes...etc

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Applications in Industry 4.0

Digital Manufacturing

Automation & Smart Factories

Energy

Intelligent Use of Energy & Smart Grid

Healthcare

Tele-medicine & Smart Devices

Digital & Cloud Services

Digitised Services & Business Models

Digital Economy and Society

Supply Chain Integration

Logistics & Supply Chain Management

Smart Cities & Infrastructure

Telecoms. Building & Spaces

Living Spaces

Smart Consumer Products and Homes

Smart Integrated Transport

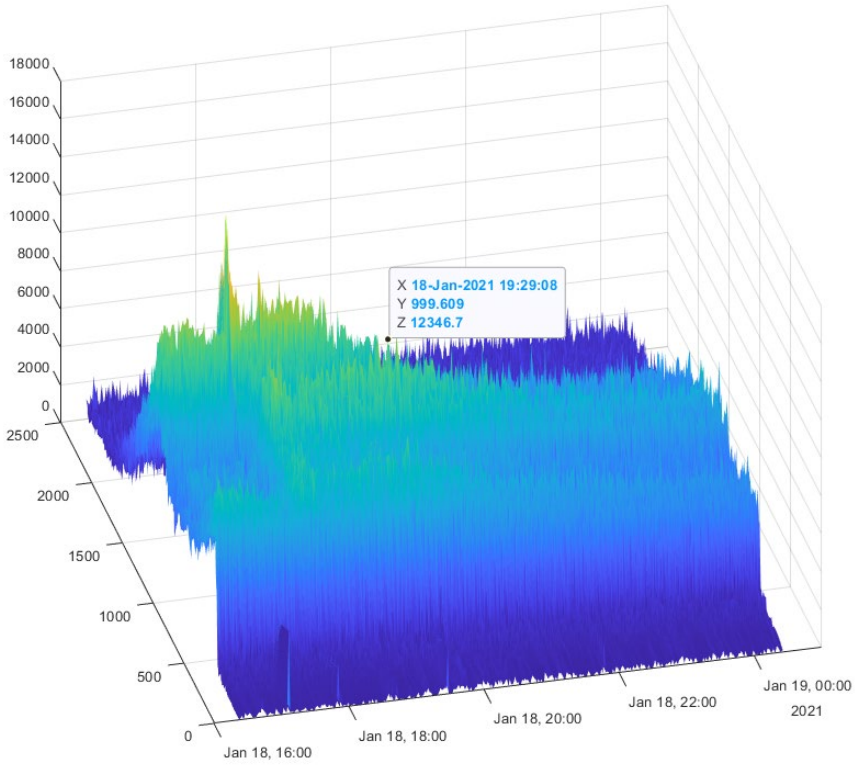
Enabling technologies such as cloud, sensors seamlessly connects products, equipment and structures to digital and big data services and user interfaces

It's a critical capability to the future of manufacturing and new product innovation

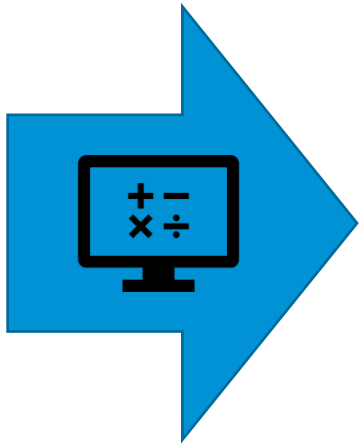
Critical to next generation UK manufacturing

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

What is the Role of Data Analytics in Industry 4.0?



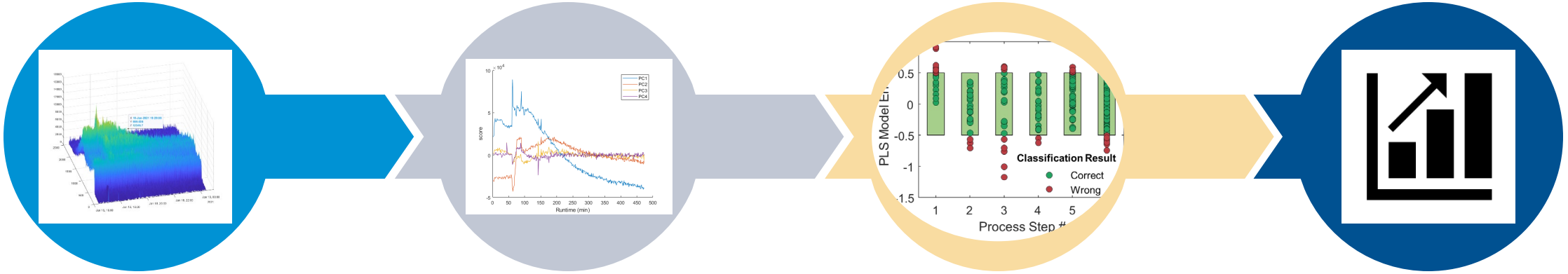
Data



Understanding

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Data Analytics = Data Driven Decision Making



Data

Understanding

Prediction

Decision

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 Analysis Methods Relevant to Industry 4.0

Statistical Data Analysis

- Understand the maths.
- Apply statistical methods to find a more meaningful representation of the data.

Mechanistic Modelling

- Understand the system.
- Build a mathematical representation of the actual system.

Machine Learning / AI

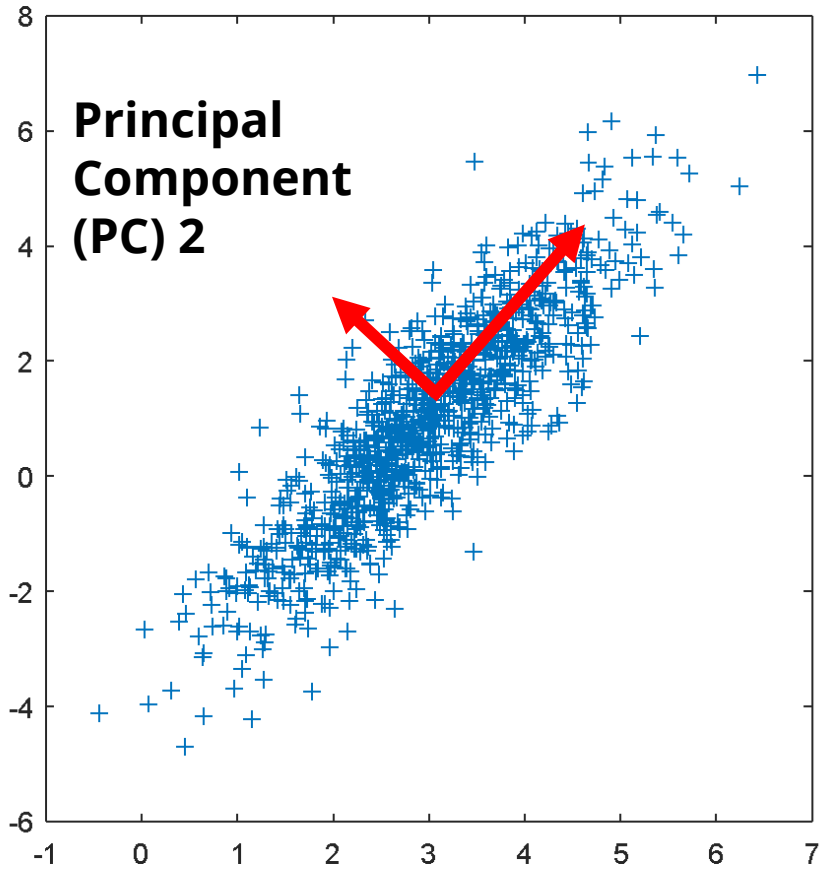
- Have lots of data.
- Find structure and patterns in large amounts of data and apply that pattern to new data.

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.



Principal Component (PC) 1

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

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 $\frac{dV}{dt} = v_{avg} \cdot A$

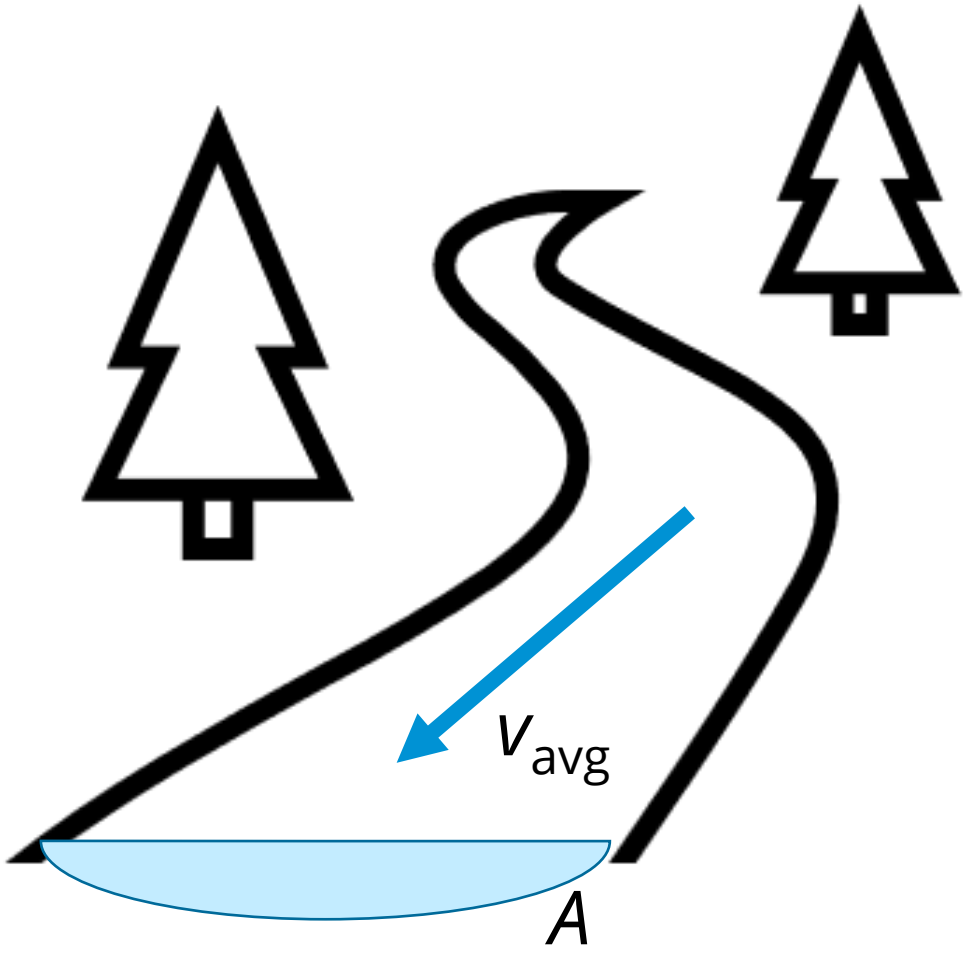
2nd step: refine $\frac{dV}{dt} = \int v(x) dA$

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.

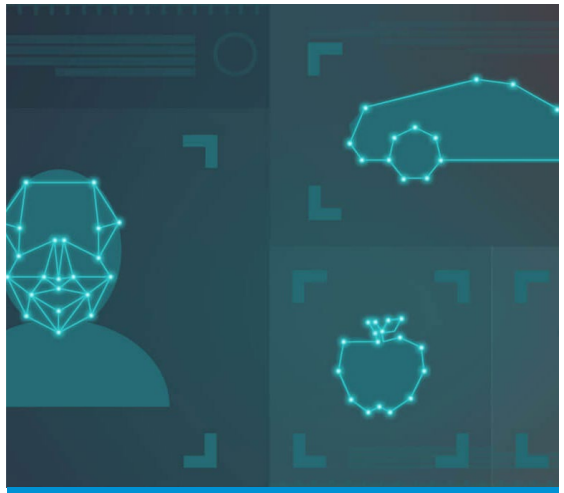
"All models are wrong, but some are useful".



This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

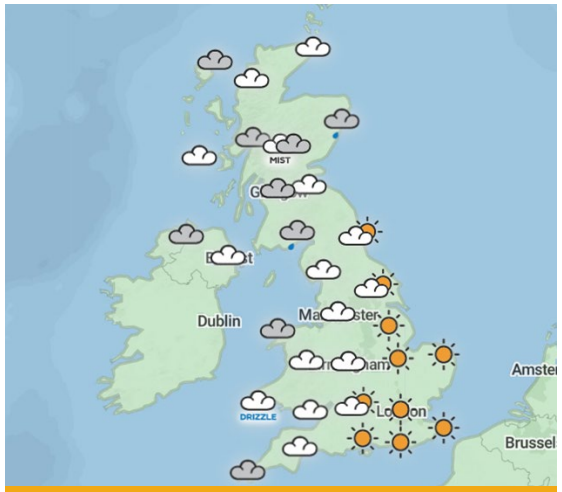
Machine Learning in Everyday Life

Online Shopping
E-commerce relies on algorithm to learn shopping patterns & tastes to develop custom recommendations.



Social Media
Image filters or tagging photos – image recognition techniques used to recognise the place, person or object.

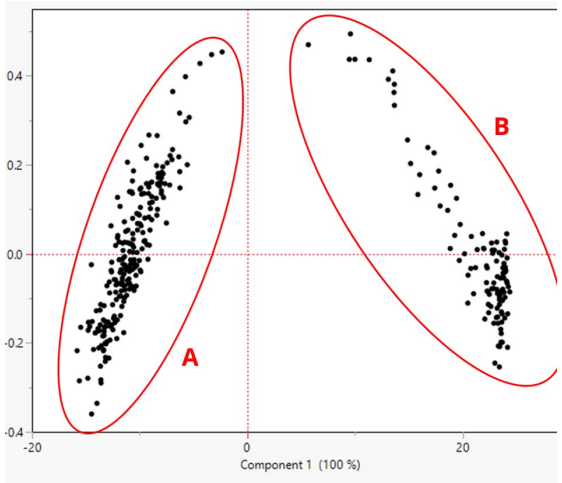
Internet Search
Search engines use data science algorithms to come up with the best results within a fraction of a second.



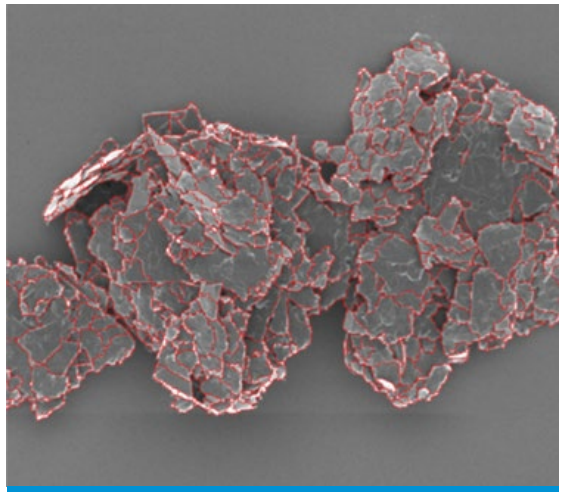
Weather Forecasting
Massive amount of data is collected across the globe to monitor and predict environmental and weather conditions.

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Machine Learning in Industry 4.0

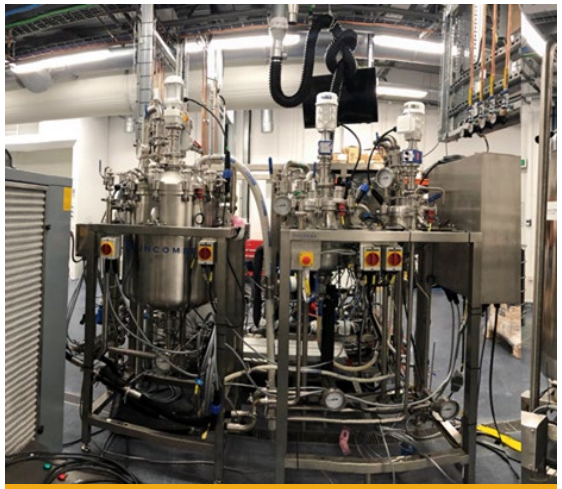


Automatically filtering unwanted/bad quality data in real-time.



Optimise experiments and determine best results for rapid product development.

Optimise experiments and determine best results for rapid product development.



Predict product characteristics from process conditions and monitoring sensors.

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 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

cpi

Innovate UK

Collaborative project

36 months

£557K

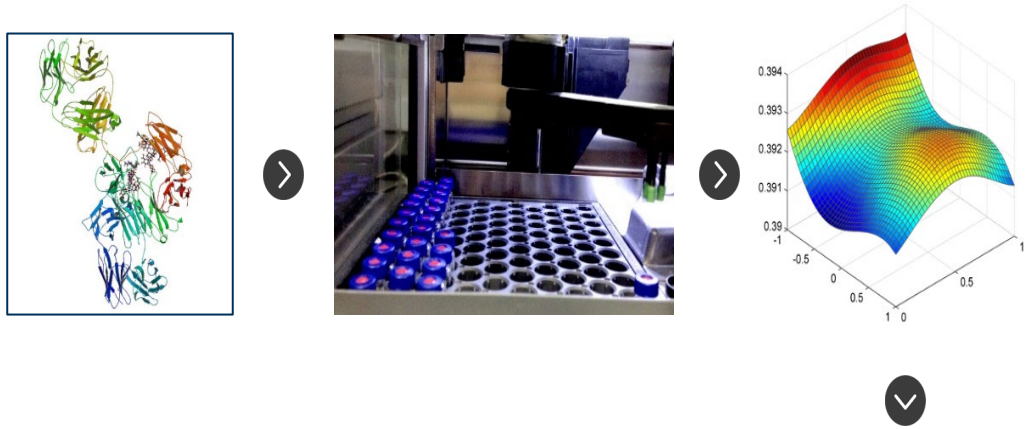
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Machine Learning in Manufacturing (cont.)

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

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

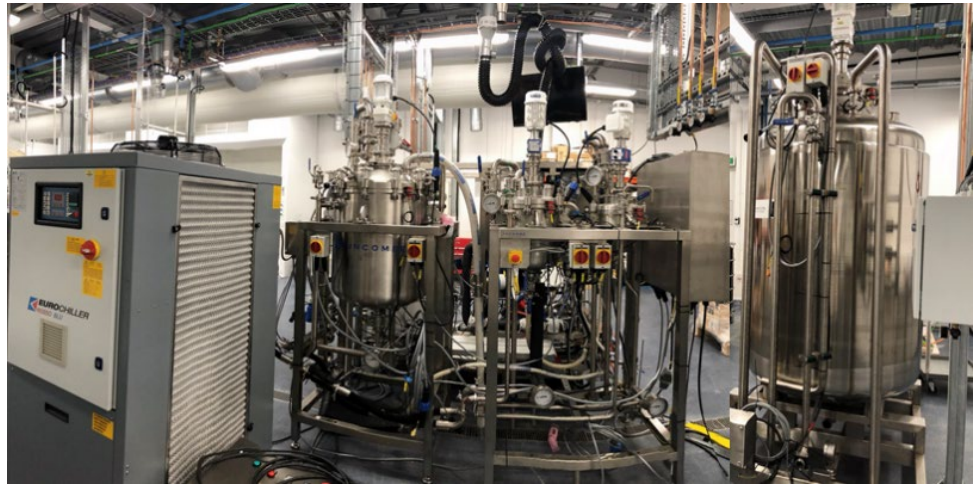
Machine Learning in Manufacturing (cont.)

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.



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.

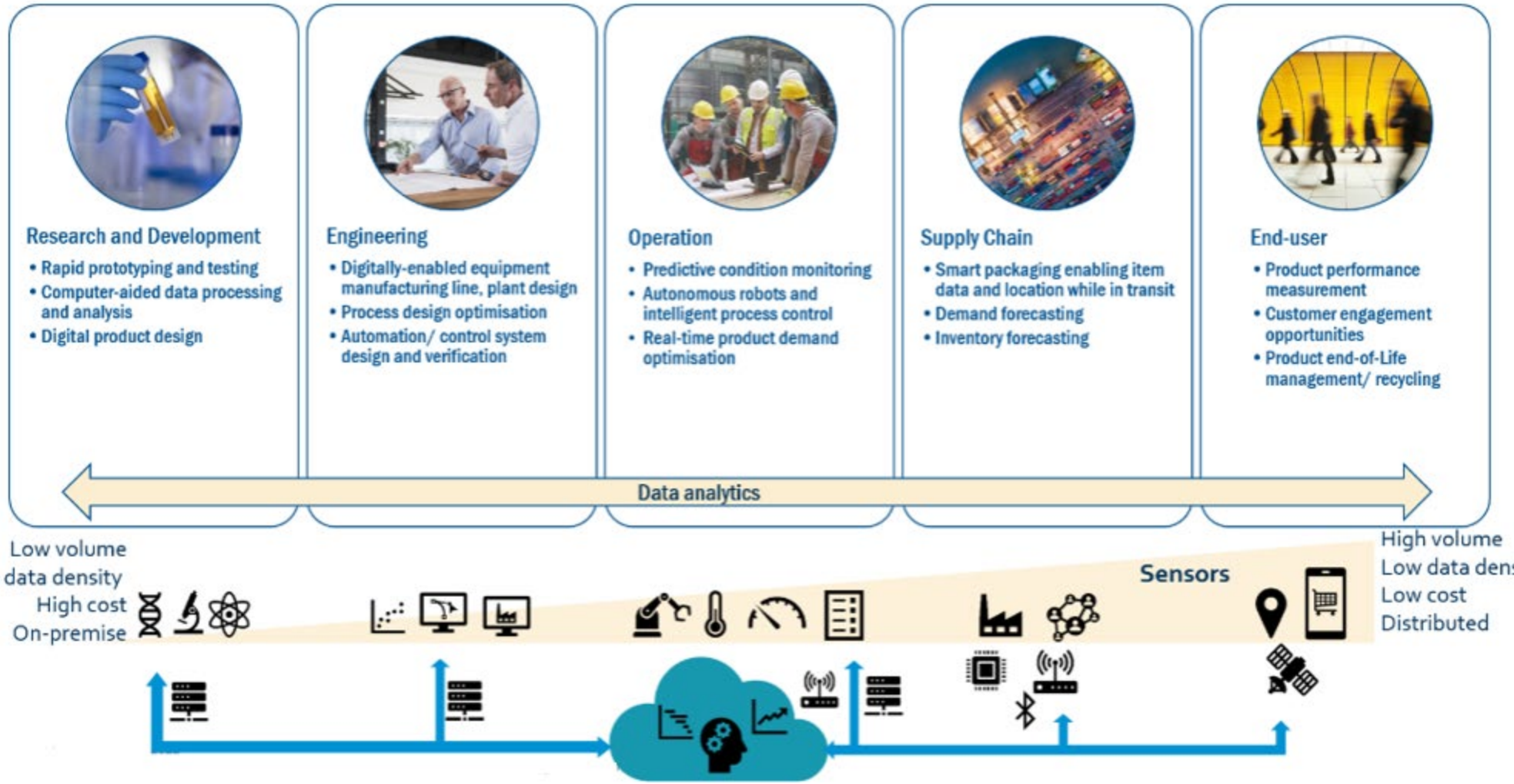
Collaborative project

6 months

£200k

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

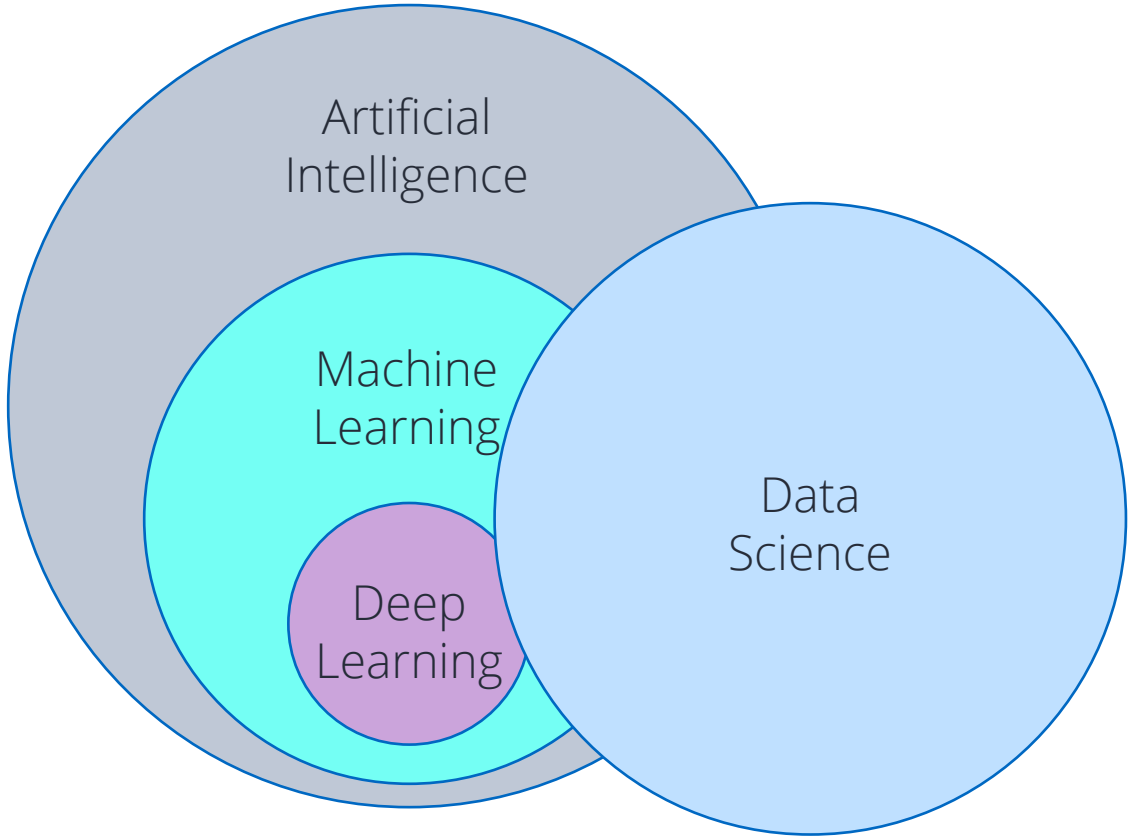
Data Analysis Enables End to End Manufacturing Optimisation



This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

What Skills Are Required For Industry 4.0?

- **Data Science** – understanding finding insights in data, sometimes using computers to do so.
- **Artificial Intelligence** – techniques which enable computers to mimic humans.
- **Machine Learning** – subset of AI which uses statistics to enable machines to improve predictions with experience.
- **Deep Learning** – advanced machine learning using complex neural networks.



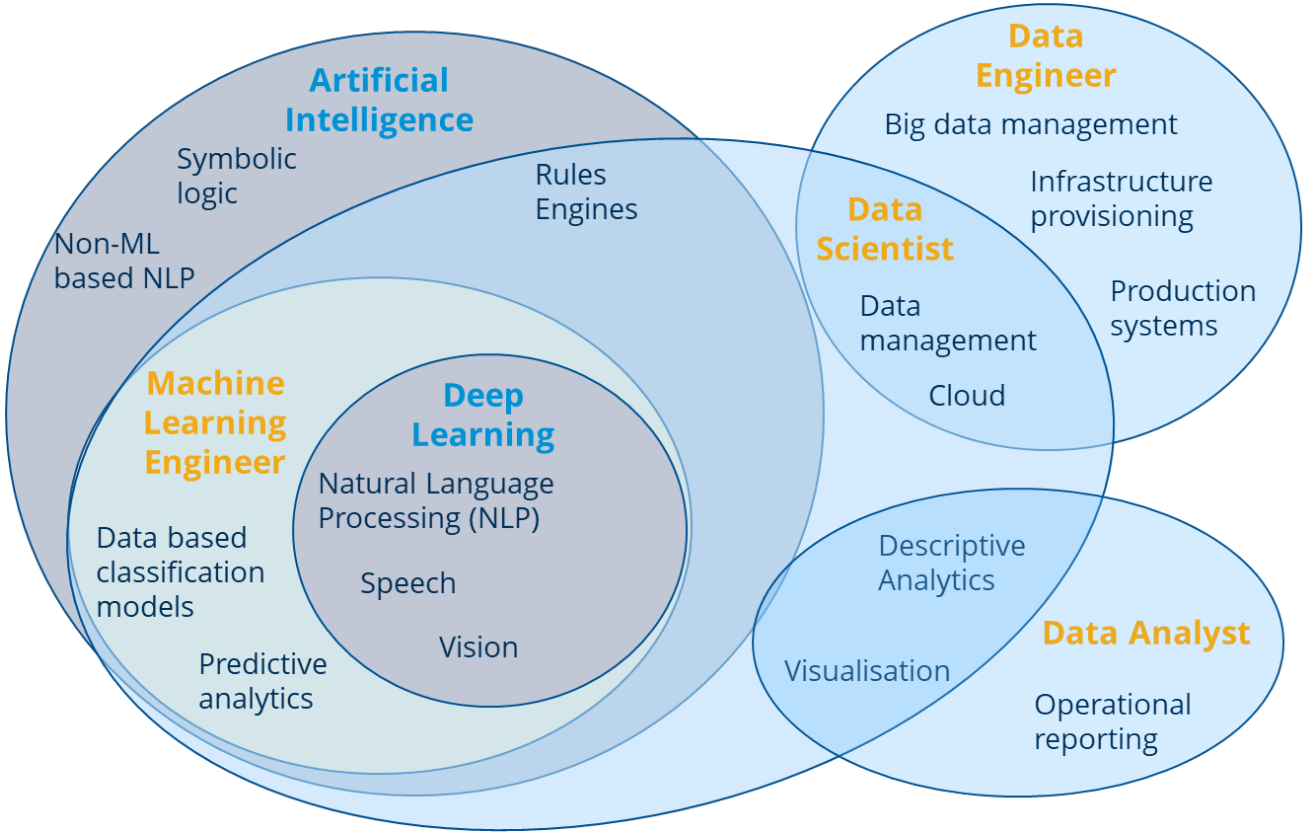
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

What Roles Are Required For Industry 4.0?

Data analyst – descriptive statistics, visualise data, and communicate results and conclusions (skills include – subject matter expertise, mathematics/ statistics, visualisation).

Data engineer – prepares data. develop, construct, test and maintain complete data architecture (skills include – data modelling, data architecture and pipelining, data warehousing, ETL, data lake, ELT).

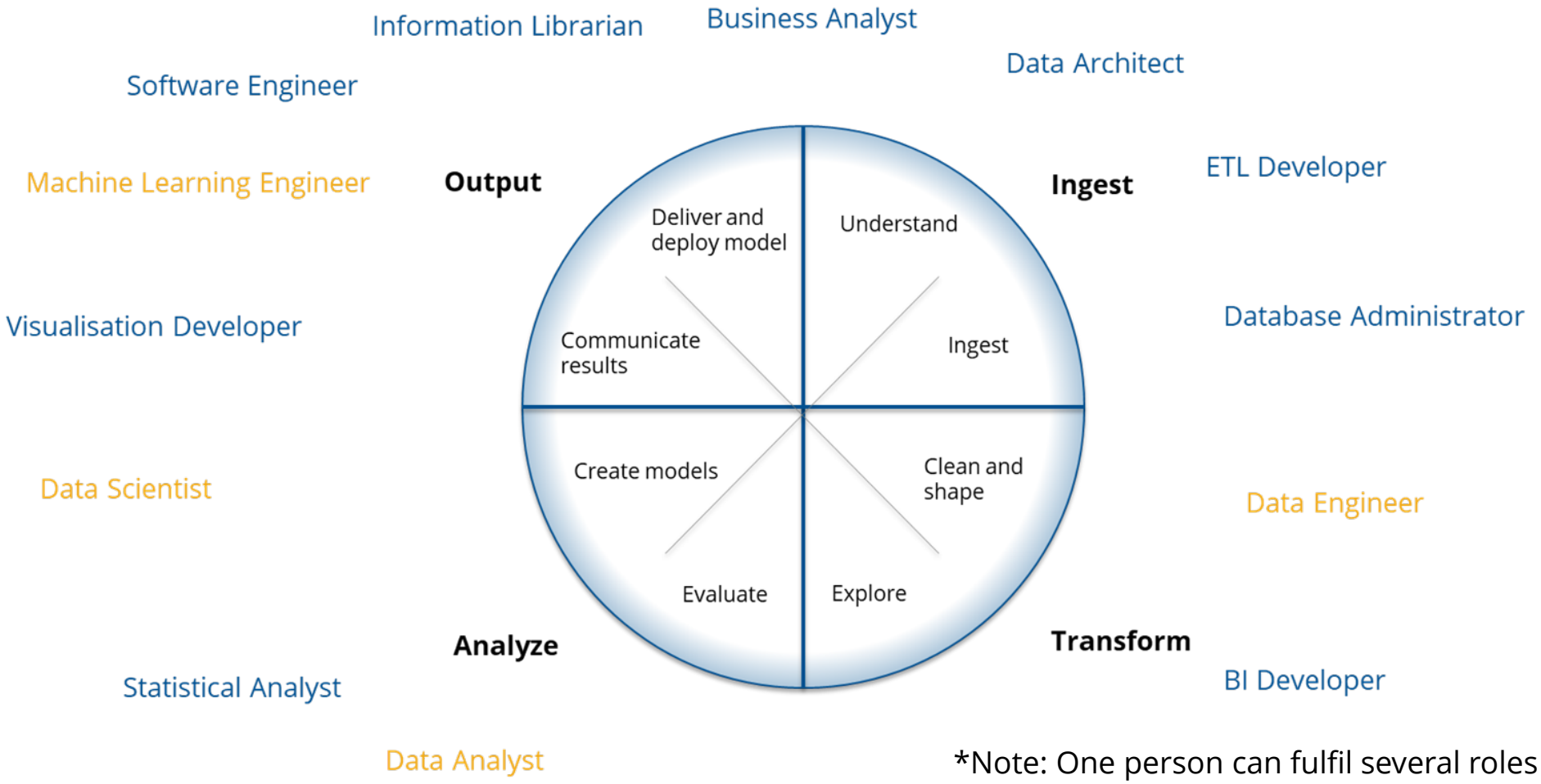
Machine learning engineer – deploy models developed by data scientists and scale them out to production-level, develop data pipelines and CI/ CD (skills include – software engineering, data architecture and pipelining, QA, testing).



***Note:** One person can fulfil several roles

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Operational Deployments Require Additional Roles



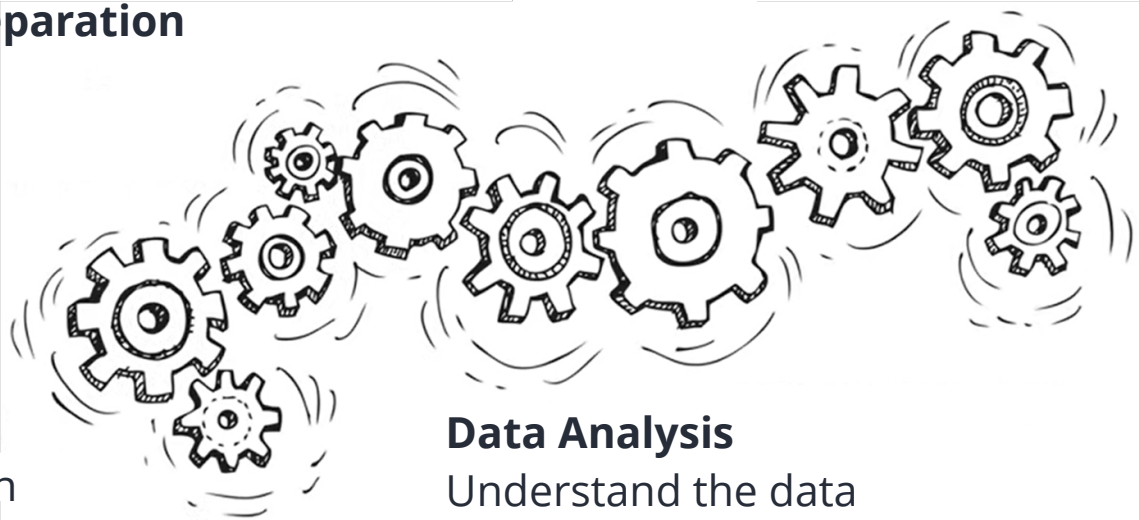
*Note: One person can fulfil several roles

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

What Does a Typical Analytic Workflow Look Like?

Cleaning
Formatting
Transforming
Data Preparation

Iterative application of algorithms to identify best model fitting the problem
Validate predictions
Predictive Modelling



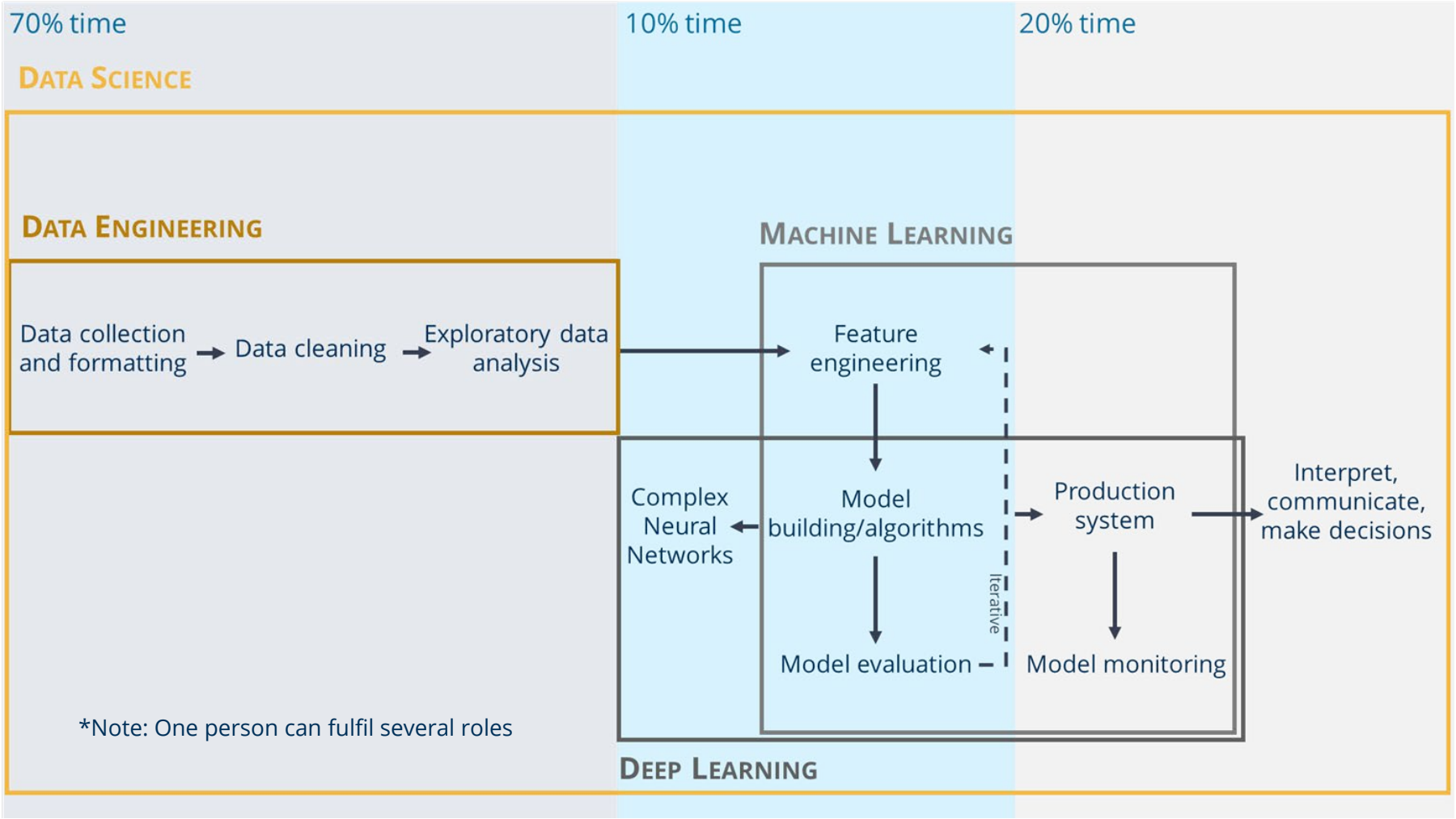
Data collection
Find relevant data and organise – often via automation / programming

Data Analysis
Understand the data
Variation, bias, outliers

Visualisation
Simple, effective manner to communicate the results

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Which Roles Are Involved in The Workflow?

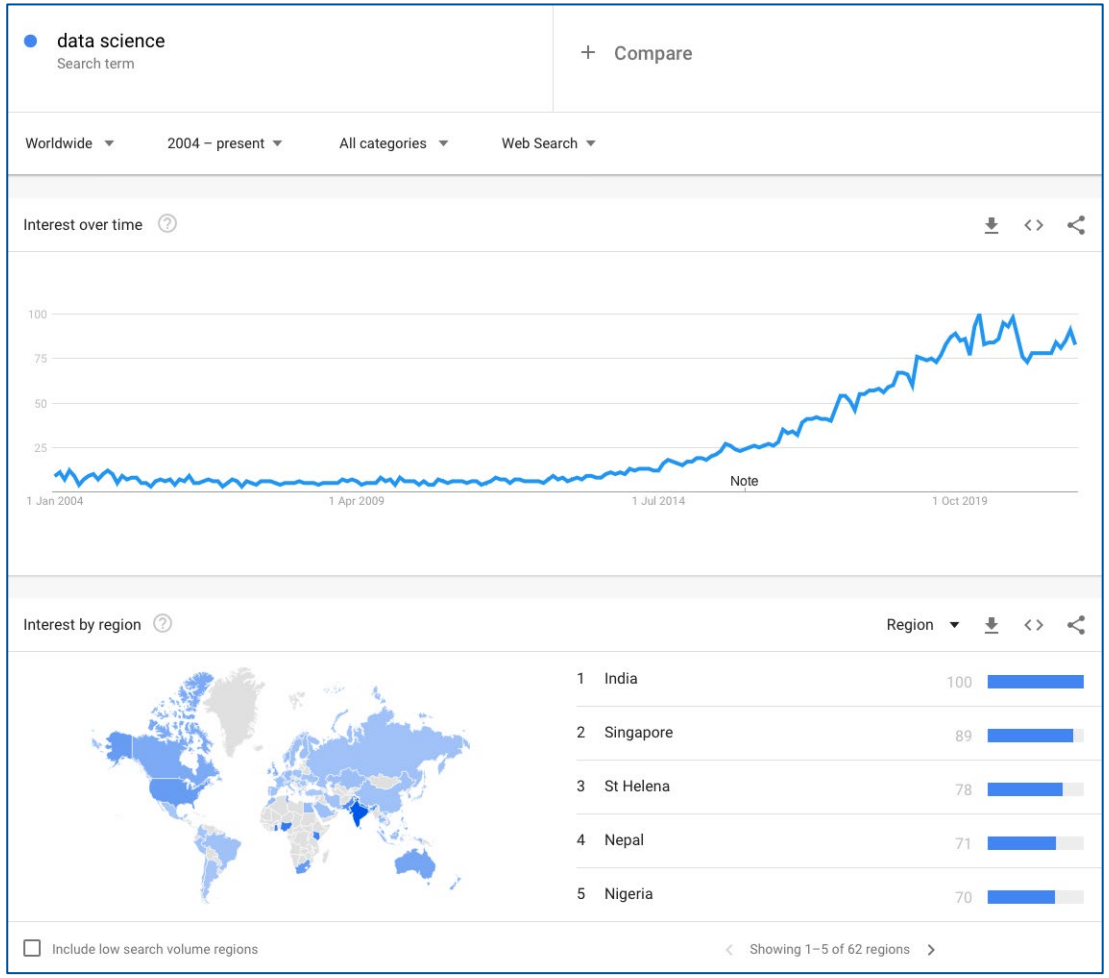


This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Trends – Interest in Term Data Science!

History of term

- 1962 – John Tukey (statistician) defined concept of *data science.*
- 1985 – Jeff Whu (statistician) used term *data science* in lectures.
- Many uses 1985 – 2008.
- 2008 – DJ Patil and Jeff Hammerbacher define *data scientist* as their professional title.



Google Trends: Data Science

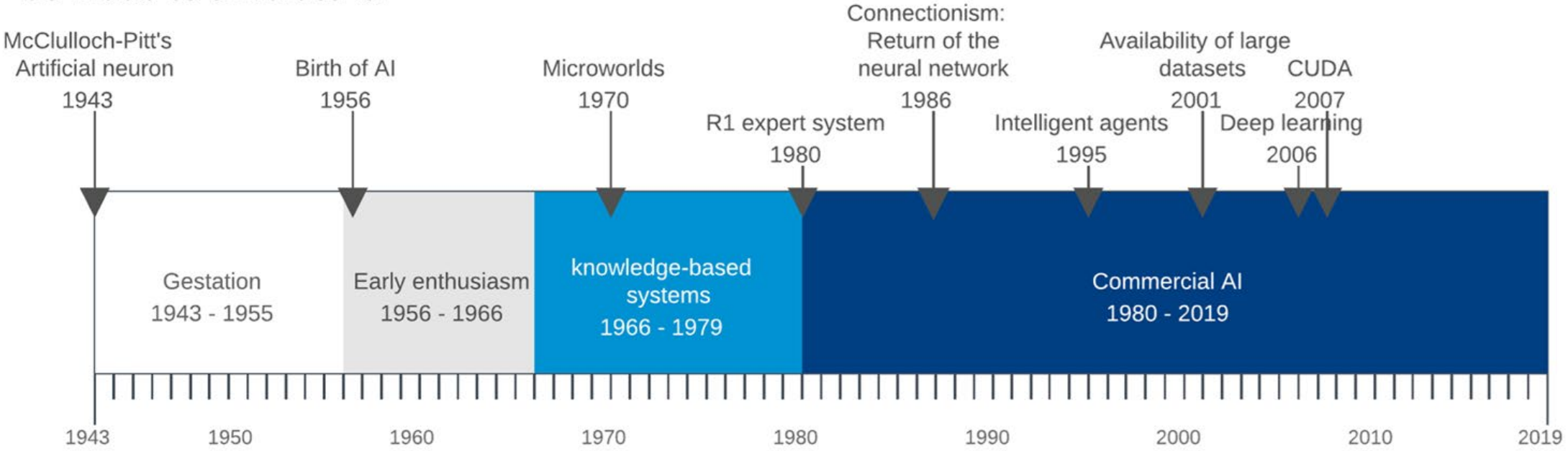
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Foundations of AI

AI phrase coined at Dartmouth College summer school

1956 – John McCarthy

”The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.”

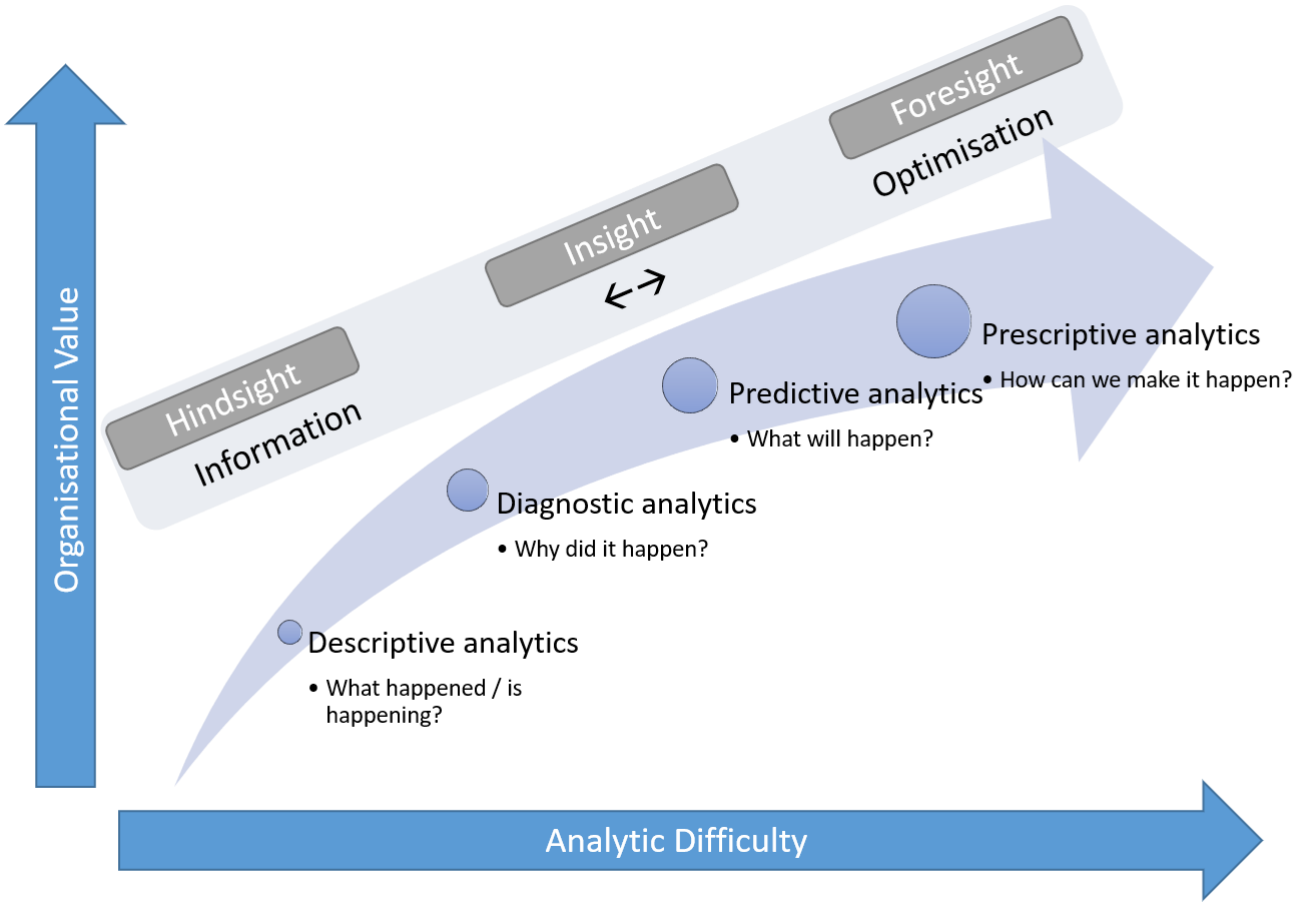


This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Different Types of Data Analytics

- **Descriptive** – also known as business intelligence (BI).
- **Diagnostic** – interpretation of descriptive analytics.
- **Predictive** - depends on techniques such as machine learning and deep learning.
- **Prescriptive** - recommends solutions that will deliver desired outcomes .

- All four types of analytics required for Industry 4.0.



This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

What is Data Mining?

The term was first coined in 1990s and it has been used in practice since the 1980s.

It is described as knowledge discovery in databases, is a predecessor to data science, and will probably be replaced by data science in future. A great deal of literature and tools exist, so important to know about the field.

Best known tools (proprietary and open source):

- SAS Enterprise Miner
- IBM SPSS Modeller
- RapidMiner (open source)
- KNIME (open source)

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Reflection and Discussion

Do you have challenges that could be solved by Data Science in your workplace?

Make a note of the following point:

- Idea(s).
- Advantages of implementing your ideas (e.g., increased yields, increase resource efficiency,...).
- Concepts/ideas from the previous section that would help you.
- Potential barriers to implementing the ideas.

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

Introduction to Machine Learning

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

What is Machine Learning?

Machine learning is using statistics to find patterns in data.

Computers learning without being explicitly programmed – training machine how to learn from data.

Making predictions on future data from historical data.

Learning from data without explicit rule-based programming.

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

What is Machine Learning? (cont.)

Machine learning is using statistics to find patterns in **data.**

Computers learning without being explicitly programmed – training machine how to learn from **data.**

Making predictions on future data from historical **data.**

Learning from **data** without explicit rule-based programming.

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Machine Learning (ML)

Provides systems the ability to automatically learn and improve from experience without explicit instructions (Deepmind - AlphaGo?).

Supervised learning

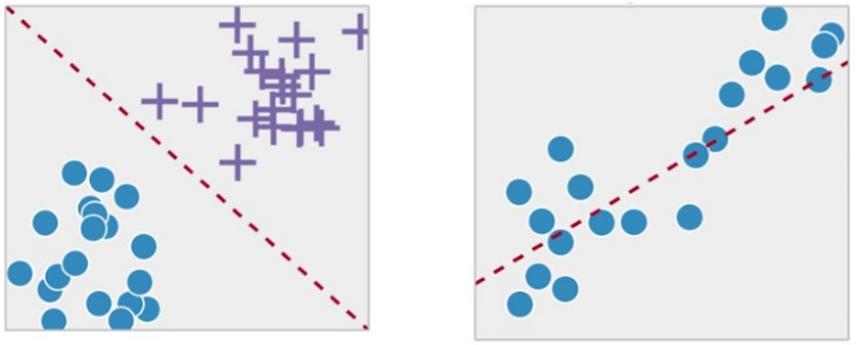
All data is labelled, and the algorithms learn to predict the output from the input data.

Unsupervised learning

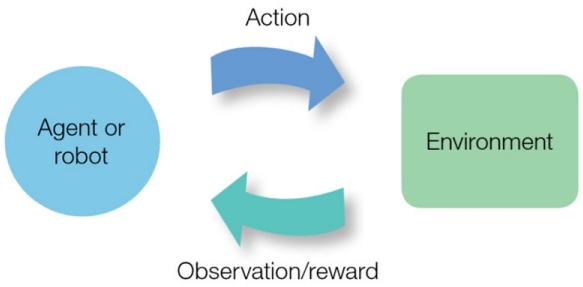
Unlabelled data and the algorithms learn to inherent structure from the input data.

Reinforcement learning (RL)

Agent learns how to behave in an environment by performing actions and receiving rewards. (How we learn as children).



Classification Regression

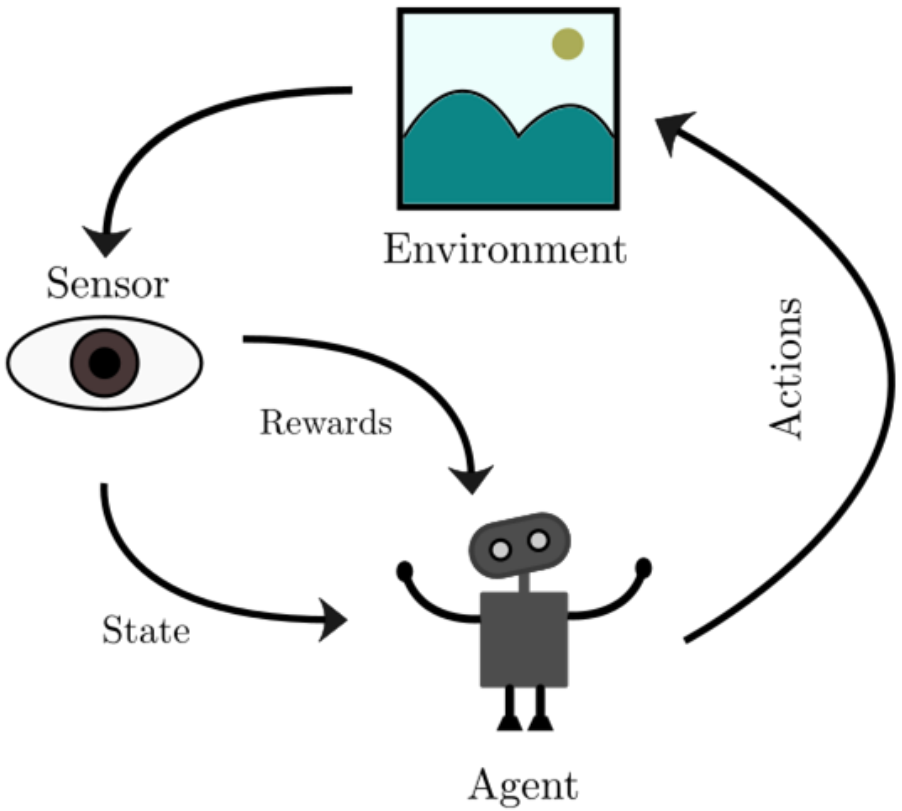


AI, Data Science and Machine Learning: a Dstl biscuit book

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Reinforcement Learning

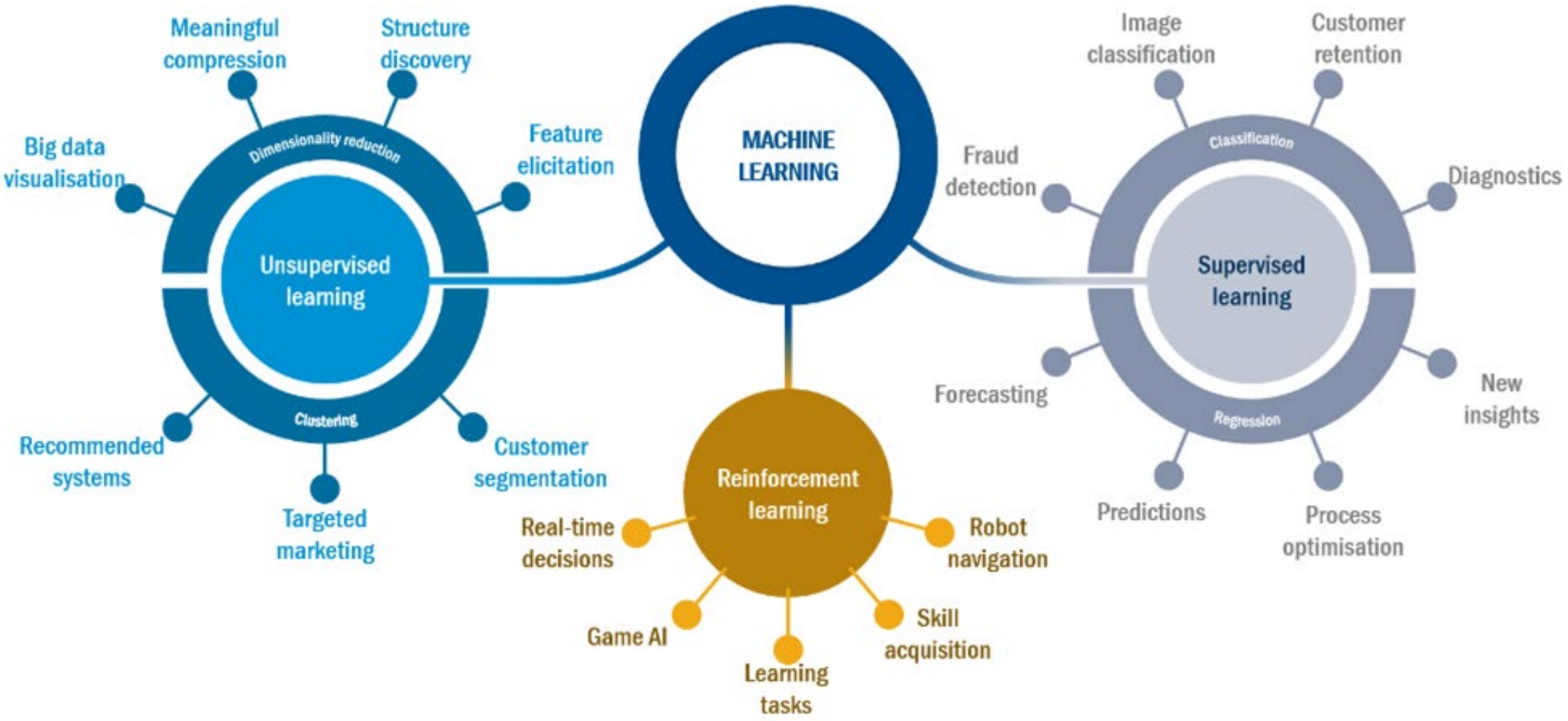
- Train an agent to make decision.
- Receives reward for each correct decision.
- Goal is to maximise agent reward.
- Based on Markov Decision Process (MDP):
  - The state of the agent encloses all the necessary information for it to take the right action.
  - We only worry about the present state.
- Many applications, main success in gaming:
  - AlphaGo (Google DeepMind).
  - Poker (Deep Q-Learning).



The typical framing of a Reinforcement Learning (RL) problem

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Types and Examples of Machine Learning



This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Why/ When to Use ML

When not to use ML:

- Only looking at historical trends.
- Regurgitating memorised examples.
- Can programme instruction or mathematics to carry out the task.
- Well understood physical phenomena (mass balance).
- Noisy or low quantity of high quality data (features which are relevant to prediction).
- If the past is not like the future.

When to use ML:

- Complex data (audio, images).
- Rules depend on too many factors, overlap or need to be tuned very finely.
- Cannot scale up solution – good at handling large scale problems.
- Data with good variation (examples of many possibilities).
- Theoretical framework solution/evaluation is too costly to justify; falls apart under certain conditions; or requires certain assumptions.
- Forecasting.

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Limitations/ Problems

- **Overfitting** - model exhibits bias towards the training data and does not generalize to new data, and/or variance i.e., learns random things when trained on new data.
- **Dimensionality** - algorithms with more features work in higher/multiple dimensions, making understanding the data more difficult.
- **Data** quality - access to a large enough data set of good enough quality.
  - Most of the time when algorithms don't perform well, it's due a to a problem with the training data (i.e., insufficient amounts/skewed data/bias; noisy data; or insufficient features describing the data for making decisions).
  - Garbage in, garbage out.
- **Interpretability** - may be able to predict for future data but sometimes difficult to determine how it came to those conclusions (black box between input and predictions).
- **Correlations** - If there is no pattern that connects the inputs to the outputs – although nothing short of magic can help here!
- **Generalise** - i.e., still apply to the new predictions, new data is fundamentally different or has different rules.
  - E.g., pandemic changes all the rules.

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Data Processing Requirements For ML

- Important to distinguish between different types of data as they are *processed differently:*
  - Structured.
  - Unstructured.

- Depending on 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).
  - 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.

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

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 Tidy Data]

"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.

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Unstructured Data

- **Text**
  - Unstructured free-form text e.g., survey results, call centre logs, product reviews, social media feeds, blogs etc.
  - By using pre-processing, 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]

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Reflection and Discussion

Do you have challenges that could be solved by Data Science in your workplace?

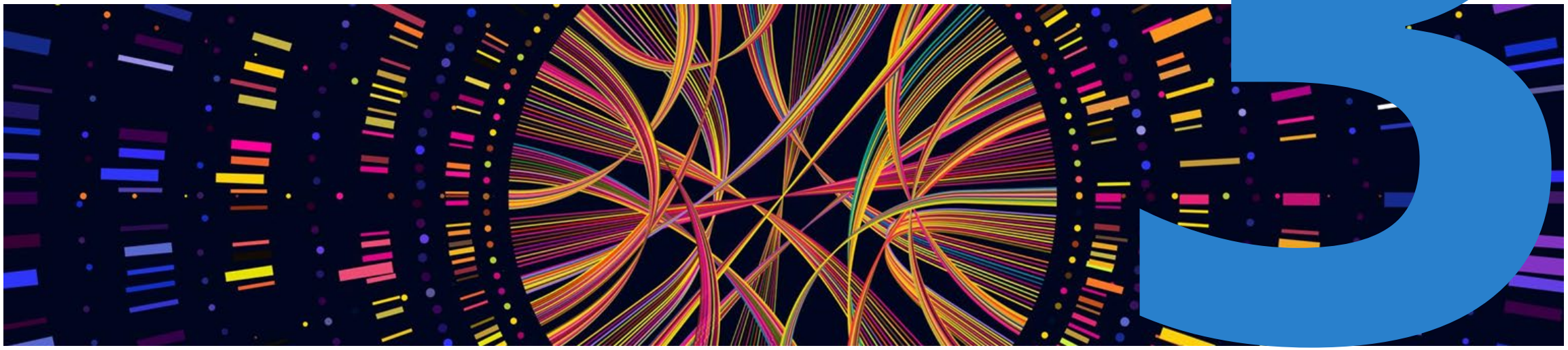
Make a note of the following point:

- Idea(s).
- Advantages of implementing your ideas (e.g., increased yields, increase resource efficiency,...).
- Concepts/ideas from the previous section that would help you.
- Potential barriers to implementing the ideas.

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 Software Development Lifecycle

3



This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Why Use The Software Development Lifecycle (SDLC)?

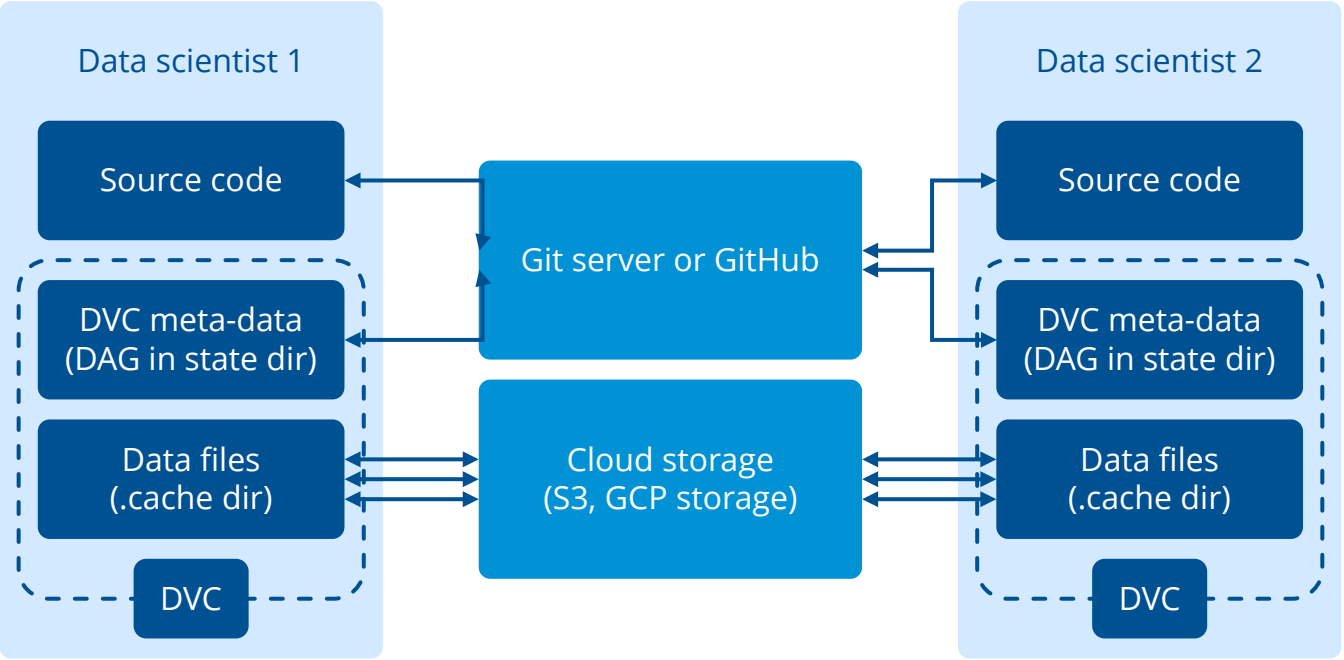
The ML development process is:

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

Governance policies are required to manage complex processes.

As is the case for software development, these policies should include:

- Documentation.
- Data privacy and access.
- Data quality / lineage.
- Model QA / test.
- Model versioning.

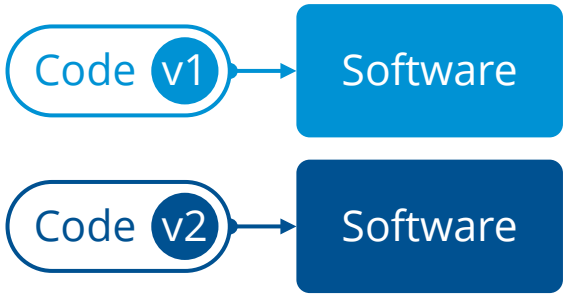


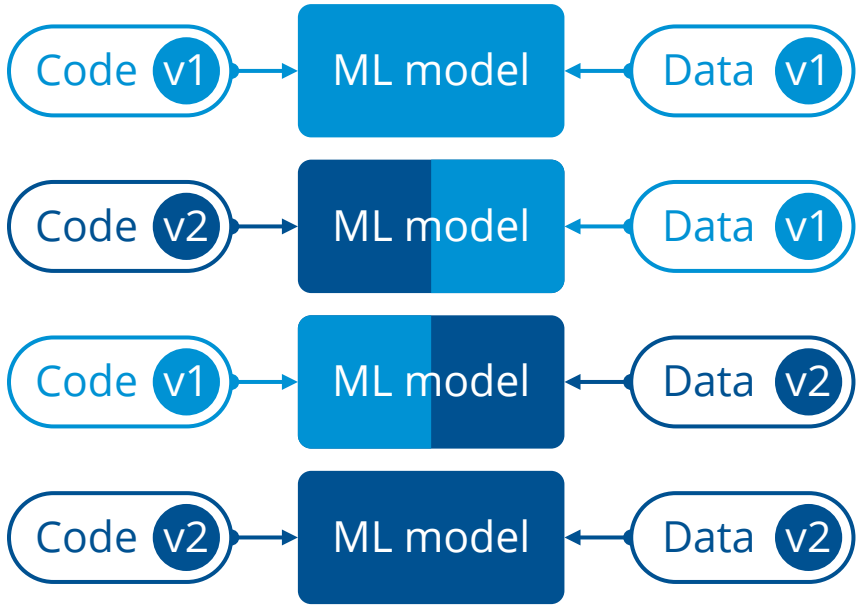
One possible approach to version control of ML models

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

How is ML Different From Software Engineering?

- Data is a new variable in the development process.
- Increases complexity of process.
- Increases complexity of process automation.
- Longer development, testing, QA, and release times.
- Versioning becomes challenging.

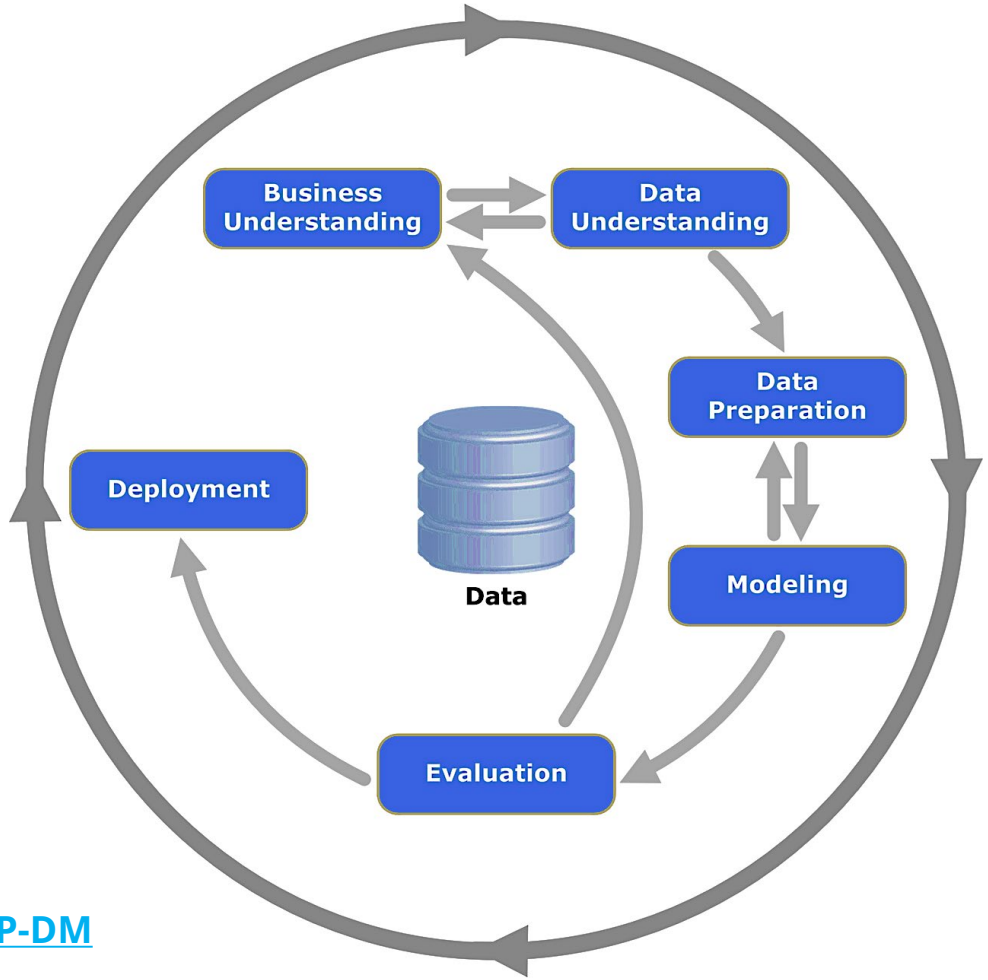




This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Data Science Software Development Lifecycle (CRISP-DM)

- CRISP-DM is the CRoss-Industry Standard Process for Data Mining, the industry standard data mining methodology.
- Published in 1999.
- Built into IBM SPSS Modeller.
- Closest thing to a standard data science SDLC to date.
- Promotes:
  - Business involvement at all steps.
  - Ongoing evaluation.
  - Measure business value at all steps.



CRISP-DM

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Tom Khabaza (CRISP-DM) – 9 Rules of Data Mining

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.

Find value

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

Update models

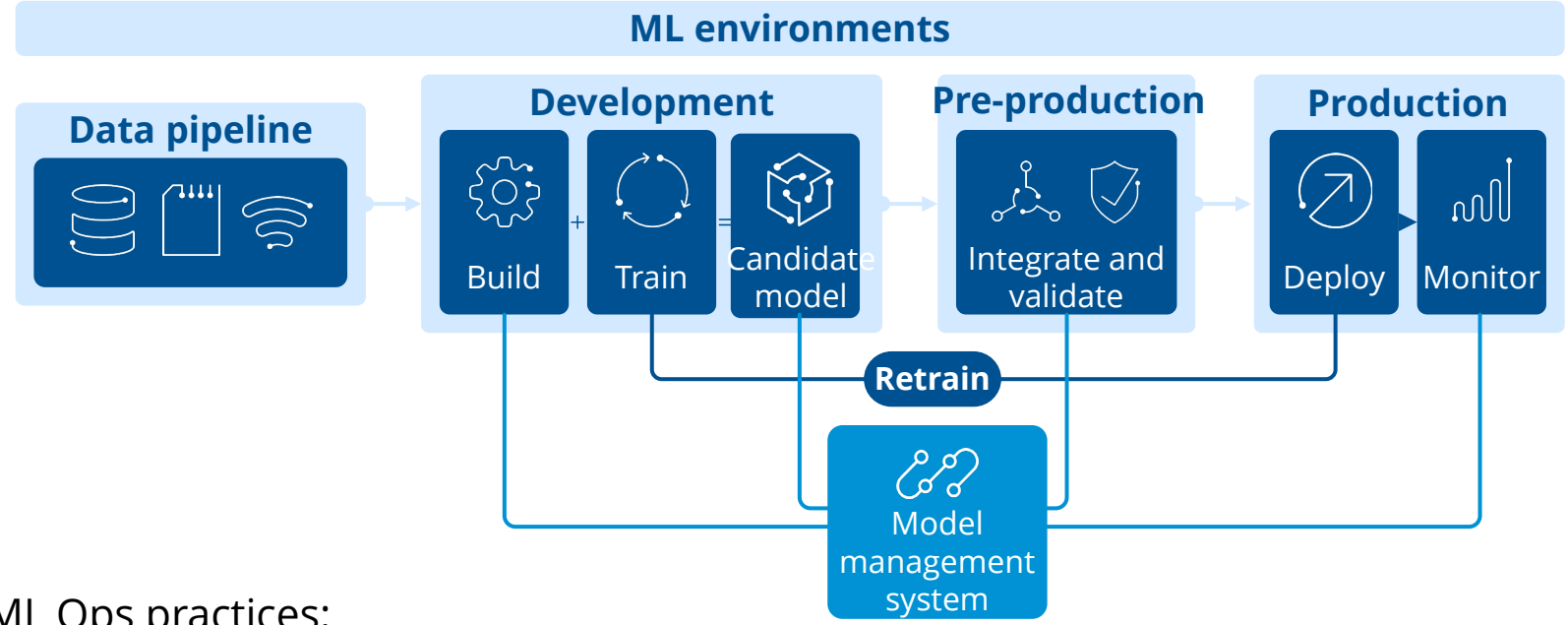
- All patterns are subject to change.

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

CRISP – DM implementation best practice

Use SDLC processes:

- Source / version control.
- Documentation.
- Change control.
- Data management (lineage).
- Information management.



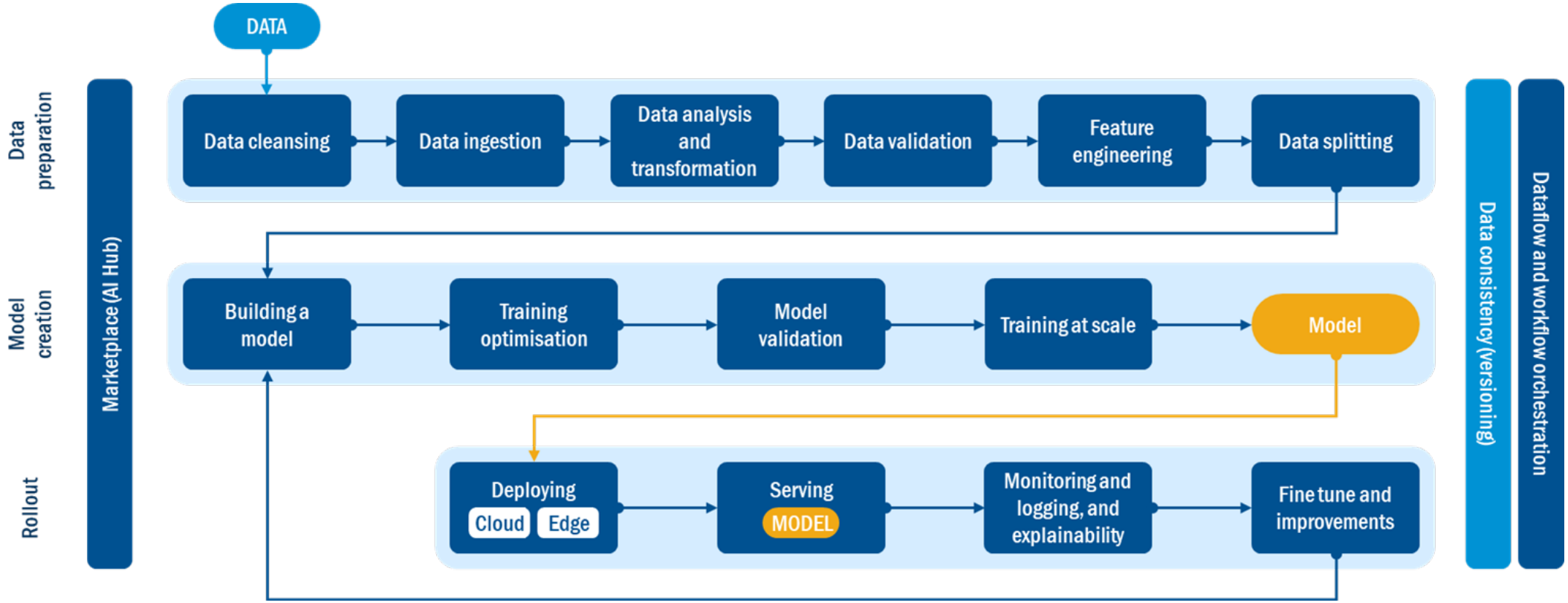
The approach is incorporated into ML Ops practices:

ML Ops is a set of practices that aims to deploy and maintain machine learning models in production reliably and efficiently.

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Model deployment – CI/CD framework

- Continuous integration and continuous delivery (CI/CD).
- Model inside a container that runs on many machines.
- Versioning data as well as code.



This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Reflection and Discussion

Do you have challenges that could be solved by Data Science in your workplace?

Make a note of the following point:

- Idea(s).
- Advantages of implementing your ideas (e.g., increased yields, increase resource efficiency,...).
- Concepts/ideas from the previous section that would help you.
- Potential barriers to implementing the ideas.

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Data Maturity Model

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Why Use an Organisational Data Maturity Model?

- Assess an organisation's level of data capability.
- Highlight areas where progress can be made.
- Set strategic direction for data capability.
- Benchmark progress over time.
- Promotes a culture of continuous improvement.
- Measured across different organisational functions, typically divided across:
  - People (*skills*).
  - Process (*leadership & culture, data governance, quality and standards*).
  - Technology (*tools & architecture*).

	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]

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Data Maturity Level Definitions

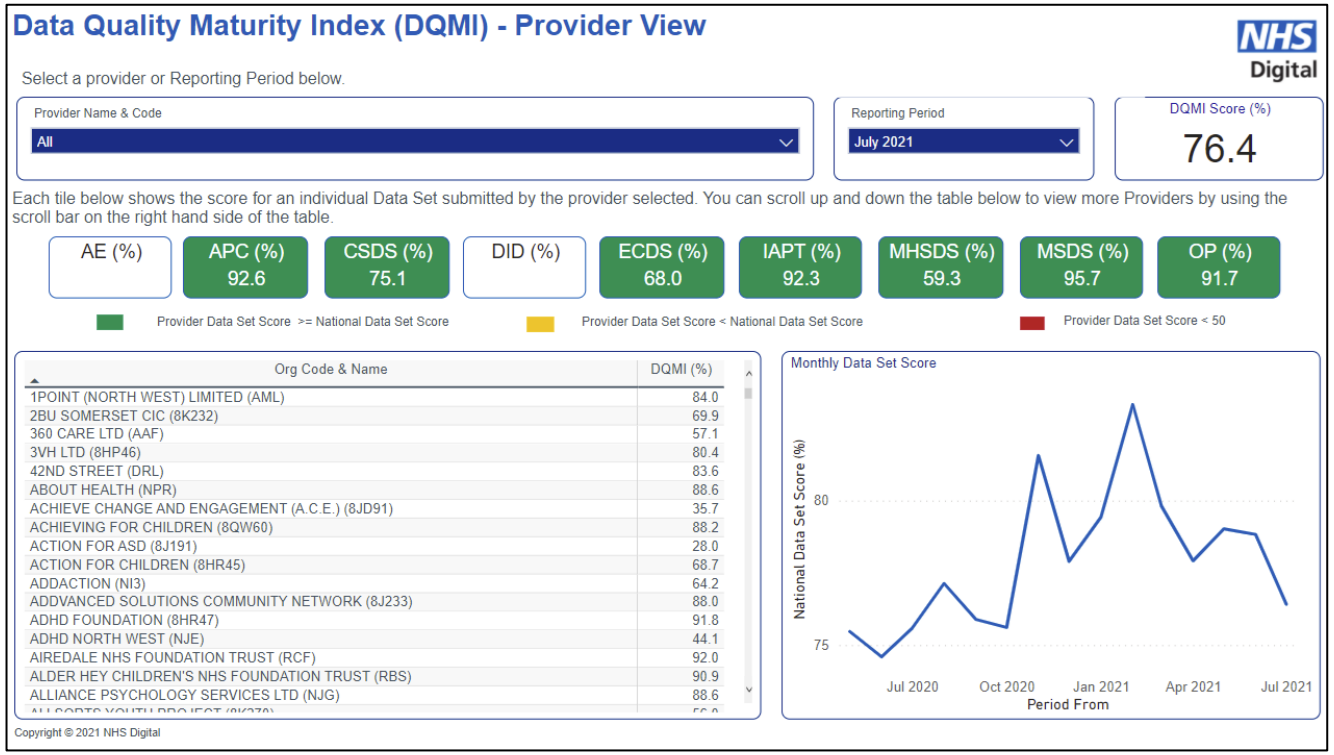
- Many possible frameworks to choose from depending on industry.
- Indicative maturity level descriptions are given below – key focus areas highlighted.

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.

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Tracking Organisational Data Maturity

- Maturity should be tracked over time.
- Each area should identify meaningful Key Performance Indicators (KPIs).
- Achievable targets should be set for KPIs.
- KPIs should be made accessible for relevant stakeholders to address gaps in performance.
- Example of a single *Quality* KPI for NHS given.
  - KPI is tracked in real-time.
  - Is transparent – all stakeholders can assess and *continuously improve.*
  - KPI is one of many that make up the overall *Data Quality* metric for NHS – **Balanced Scorecard** methodology applies.



Contains information from NHS Digital, licenced under the current version of the Open Government Licence – see Copyright slide

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Reflection and Discussion

Do you have challenges that could be solved by Data Science in your workplace?

Make a note of the following point:

- Idea(s).
- Advantages of implementing your ideas (e.g., increased yields, increase resource efficiency,...).
- Concepts/ideas from the previous section that would help you.
- Potential barriers to implementing the ideas.

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

End

Thank You

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

References

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

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

3. Jason Brownlee (2021), **Machine Learning Mastery**

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

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

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

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

What Next?

List of possible 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 *(Highly recommended)*
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

This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Copyright Utilised

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