Data Science Graduate Programme Induction

Art of the Possible

Jake Marshall

Trainee Data Science Lecturer

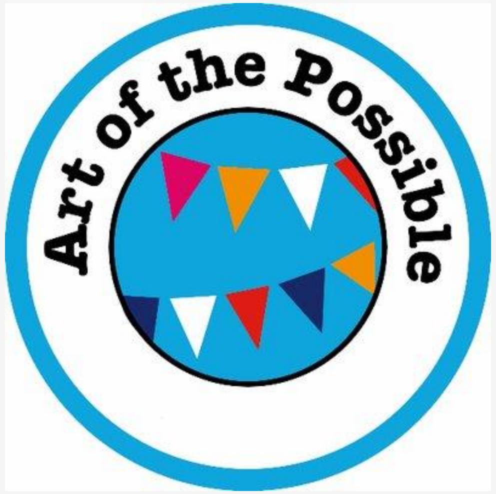
Data Science Campus Faculty Team

17 October 2023

Data Science Campus

Purpose of this session

- What is Data Science?
- Demystifying common Data Science terms
- Data science in government case studies
- Support and opportunities at the Data Science Campus



Interaction

- We will be utilising slido in this session.
- Polls and anonymous Q&A will allow us to chat!
- Go to slido.com and use the code #1957688

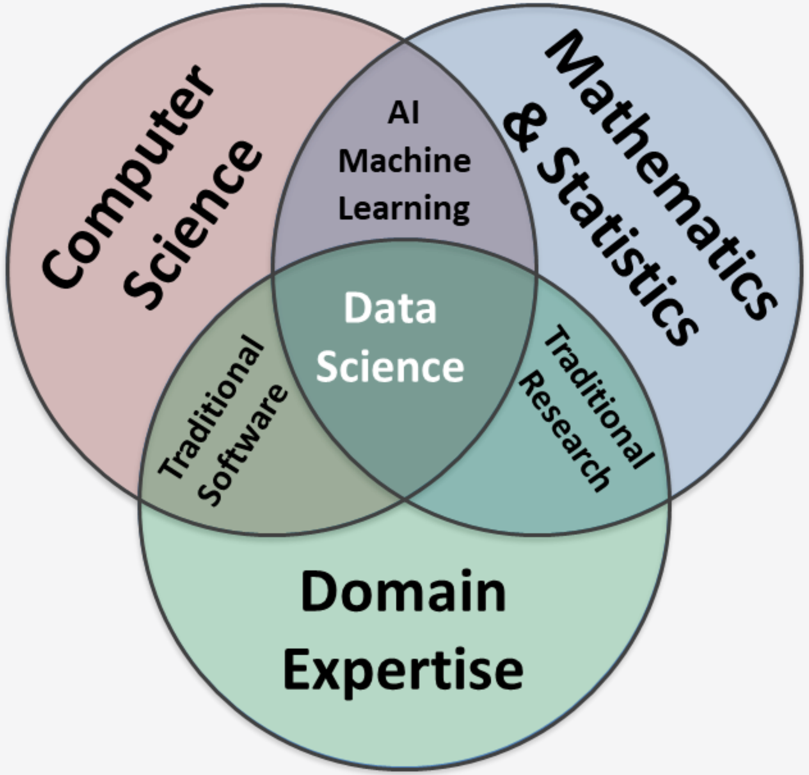
What is Data Science?

“ Work that takes more programming skills than most statisticians have, and more statistical skills than most programmers have.”

David Taylor

Attempting to define Data Science

Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

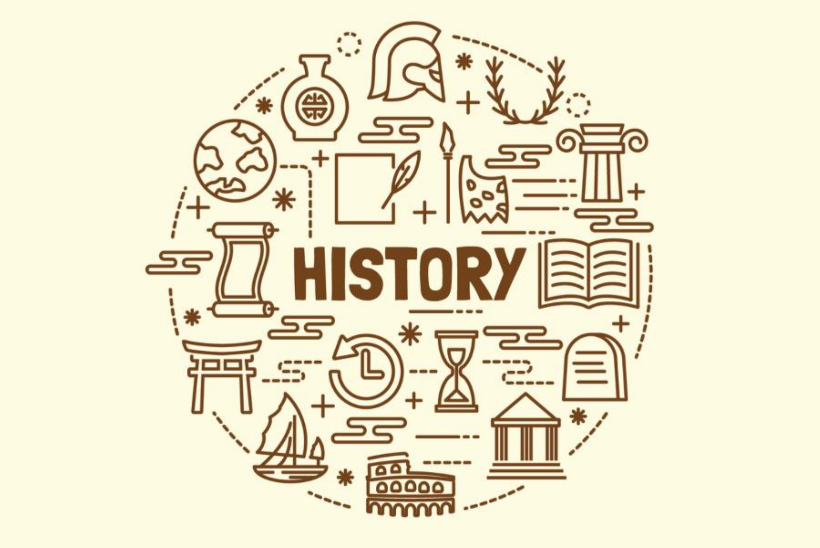


What do Data Scientists do?

“Data Scientists solve complex business problems using a combination of domain expertise, coding knowledge, machine learning and statistics skills on large and varied datasets.”

History of the name

- 1962 - John Tuckey described a field he called “data analysis”.
- 1974 - Peter Naur proposed “data science” as an alternative name to computer science.
- 1985 - Chien-Fu Jeff Wu used “data science” as an alternative name to statistics.
- 1998 - Hayashi Chikio argued for data science as a new, interdisciplinary concept, with three aspects: data design, collection, and analysis.



Origin of Data Science

Importance of Data Ethics

- We must always think about the ethics of collecting, handling and analysing data.
- In government, we are bound by the General Data Protection Regulation (GDPR).
- The UK's implementation is the Data Protection Act (2018) which controls how personal information is used by organisations, businesses and government.

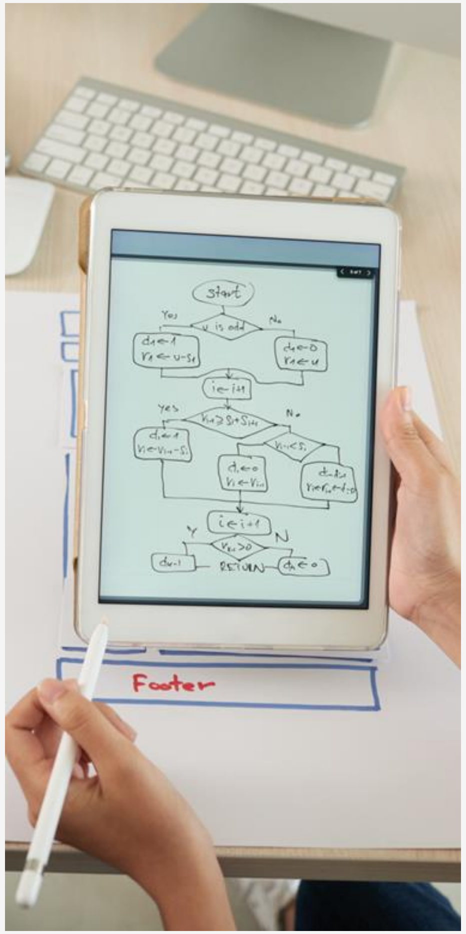
Data Ethics Canvas – The ODI

Jargon Busting time!

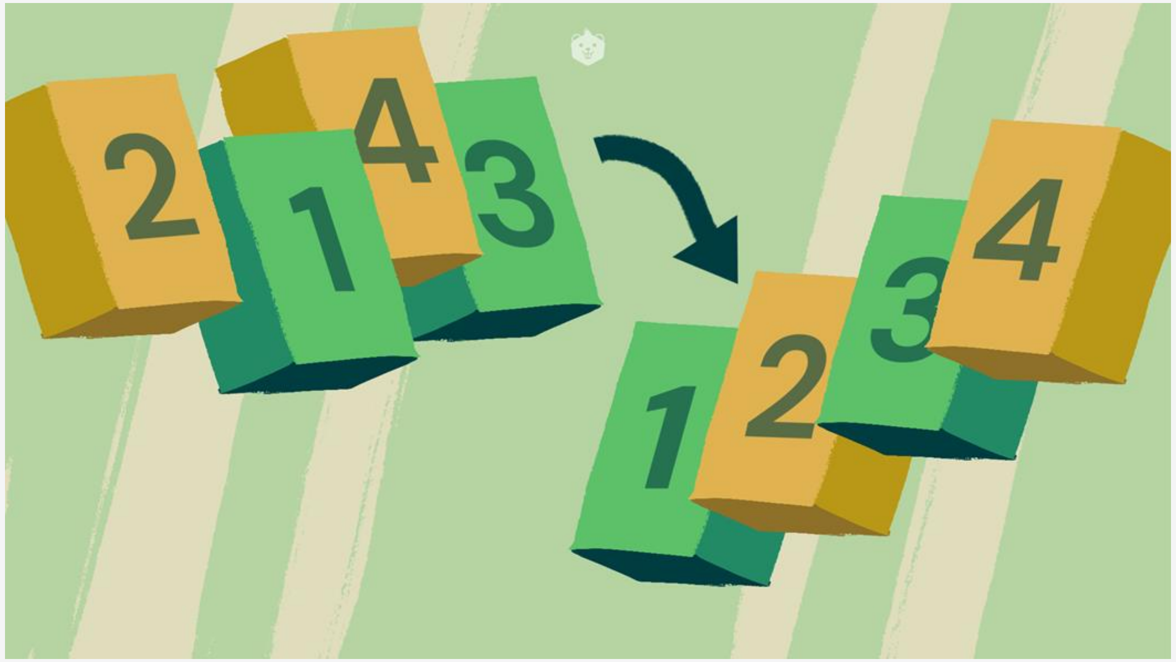
- Let’s go through some of the most common keywords associated with the practice of Data Science.
- It is very important that we as Data Scientists can describe and define our terminology to those unfamiliar or less experienced.
- You have likely heard many of these terms, but are unsure of what they really mean!

What is an Algorithm?

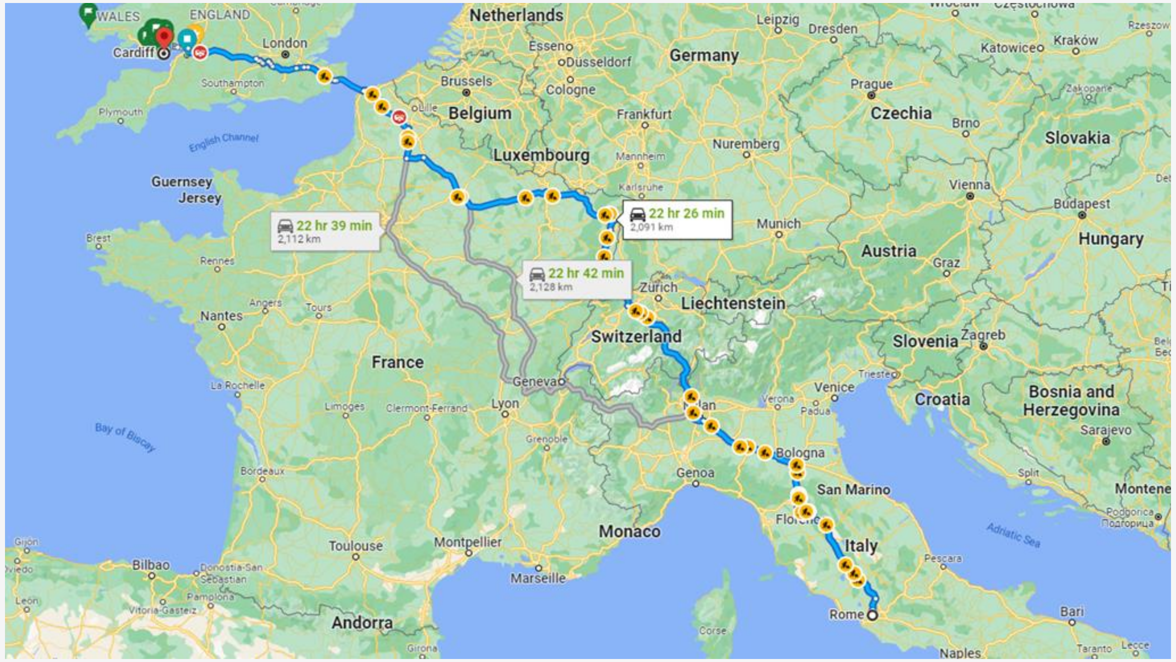
- A precise finite set of instructions, given in a specific order, that carry out a specified task
- Commonly implemented in a computer using a programming language.
- Algorithms try to minimize the number of operations needed to achieve the result.



Sorting a list of numbers



Finding the shortest route



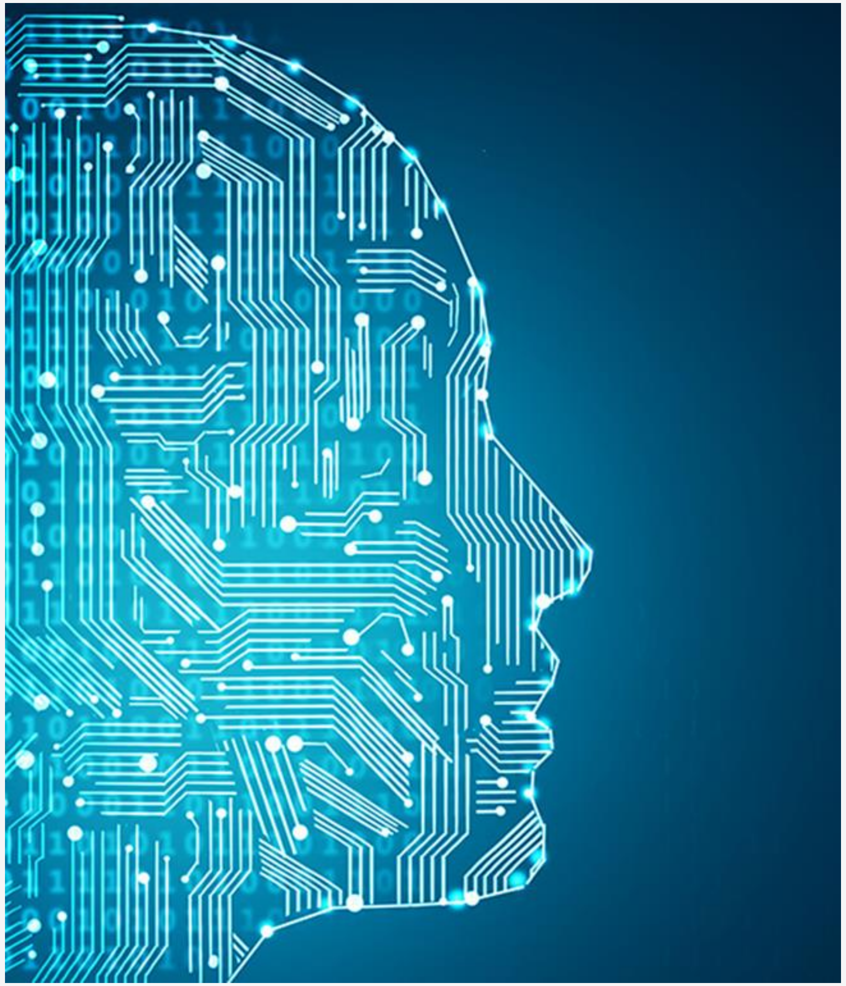
What is Machine Learning?

Field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

Tom Mitchell, 1997

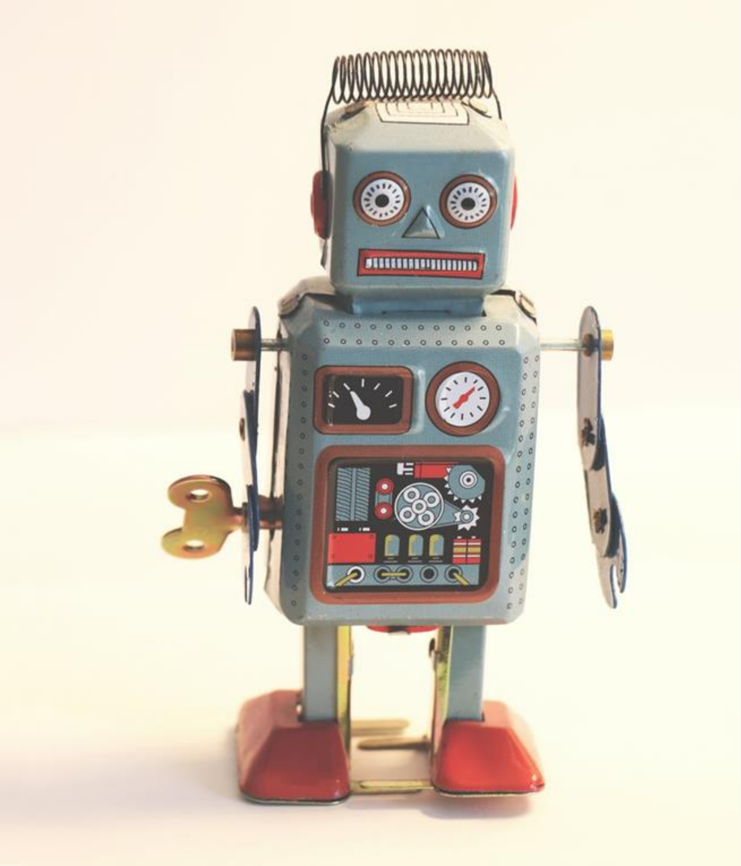


Breaking down Machine Learning

- Machine Learning algorithms search data for patterns and structures.

- Supervised ML uses these patterns to make predictions based on what we know about an outcome of interest.
  - ➢ Predict what a customer will purchase

- Unsupervised ML uses these patterns to group data into meaningful clusters, or reduce the ‘dimensionality’ (the number of variables) without affecting what the data is able to tell us.
  - ➢ Group similar customers



Everyday uses of Machine Learning

- Virtual assistants – Siri, Alexa, Google
- Online customer support – chatbots
- Recommender systems – Amazon, Netflix, Spotify
- Fraud detection – card payments
- Video Games

- Image recognition – Facebook
- Text parsing – Google Translate
- Flights (93% AI controlled)
- Driving/walking – Sat Nav

What is Data Mining?

- Extract information from a dataset and transform the information into a comprehensible structure for further use.
- Comprises six common classes of task – anomaly detection, association rule learning, clustering, classification, regression and summarization.

Example – Retail Chains

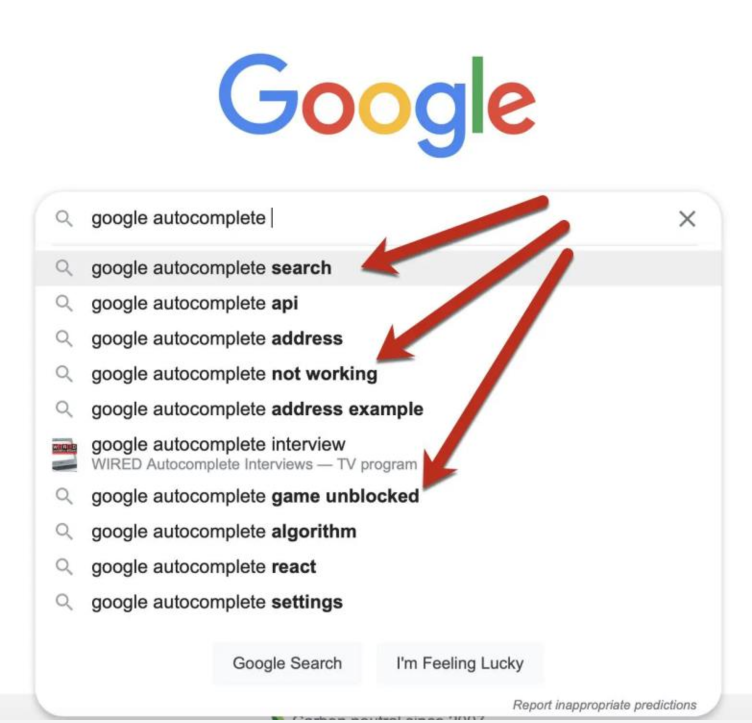
- Supermarket chains will create association “rules” between products that are frequently bought together.

- These rules are used to inform marketing strategies, recommender systems, chatbot responses and even physical locations of products in brick and mortar stores.

What is Natural Language Processing (NLP)?

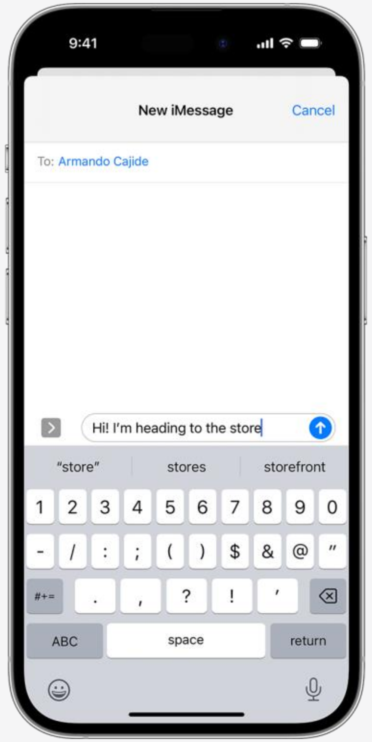
- The application of Data Science approaches to the analysis and synthesis of human language (text) and speech.
- The UK government collects immense amounts of unstructured, messy text data, via customer feedback, social media and more.
- Incorporate approaches such as cleaning and processing text, exploring and visualising relationships, predicting topics, sentiment and more!

NLP Examples



google autocomplete |

google autocomplete **search**
google autocomplete **api**
google autocomplete **address**
google autocomplete **not working**
google autocomplete **address example**
google autocomplete interview
WIRED Autocomplete Interviews — TV program
google autocomplete **game unblocked**
google autocomplete **algorithm**
google autocomplete **react**
google autocomplete **settings**



New iMessage

To: Armando Cajide

Hi! I'm heading to the store

"store" stores storefront

What is Big Data?

- Large data sets, usually too big to easily process on a single computer.

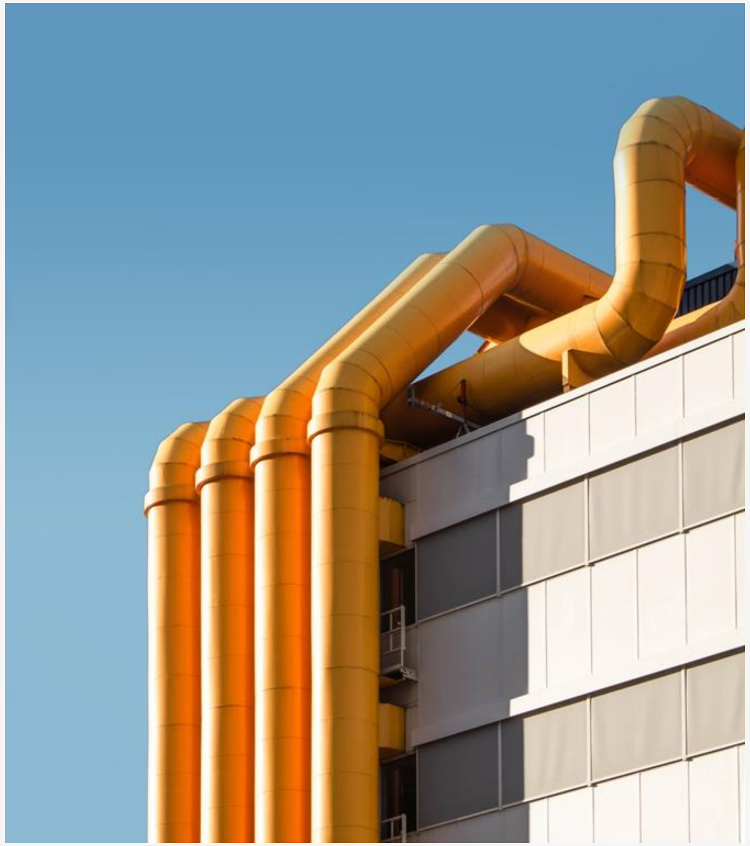
- Often requiring specialist infrastructure – distributed computing.

- A popular definition is the Three V's:
  1. Volume – amount of data.
  2. Velocity – speed of generation.
  3. Variety – different types of data.

Big Data platforms

What is a Reproducible Analytical Pipeline?

- An automated, reproducible and auditable end to end Data Science production workflow.
- Analysis has a clear audit trail that explains how, when where and why it was carried out.
- Move away from manual processing and data collection to an automated system.



Elements of a RAP

- Open-source coding languages (R/Python)
- Version Control (Git/GitHub/Gitlab)
- Modular Programming and Unit Testing
- Continuous Integration (Jenkins/Travis)

Reproducible Analytical Pipelines

Data Science in government

"Radical transformation of how the government understands and unlock the value of its own data to improve a range of public services and inform decisions at scale"

Rt Hon. Oliver Dowden CBE MP, Ministerial Foreword to the National Data Strategy

“It is imperative that we learn the hugely valuable lessons that lie buried in our data”.

“Government must also ask itself if its people have the skills necessary for the challenges that I have set out”

Rt. Hon Michael Gove, Ditchley Annual Lecture, June 2020

UK Government Data Strategy

Challenges

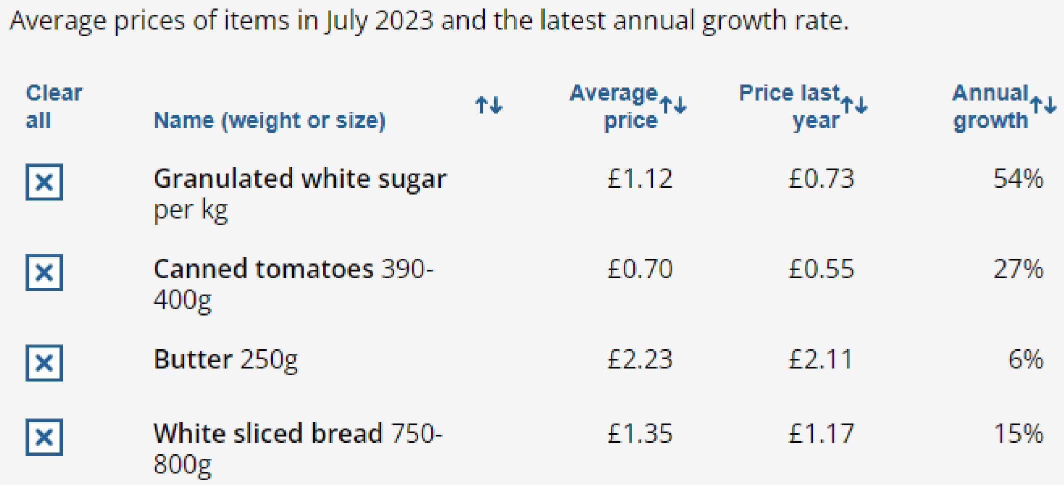
- Size
- Quality
- Rate of acquisition
- Ethics

Opportunities

- Scope of acquisition
- Rate of acquisition
- Insight from linkage
- Building capability

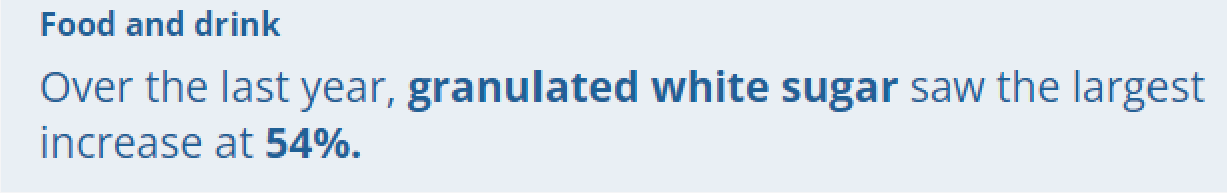
Case Study: Shopping prices comparison tool

- This interactive shows how the price of 450 different items have changed since 2018.
- Uses published indices for each item as well as monthly price information.
- You can add items to your basket and generate an immediate analysis of the change in price.



Average prices of items in July 2023 and the latest annual growth rate.

Clear all	Name (weight or size) ↑↓	Average price ↑↓	Price last year ↑↓	Annual growth ↑↓
☒	**Granulated white sugar** per kg	£1.12	£0.73	54%
☒	**Canned tomatoes** 390-400g	£0.70	£0.55	27%
☒	**Butter** 250g	£2.23	£2.11	6%
☒	**White sliced bread** 750-800g	£1.35	£1.17	15%



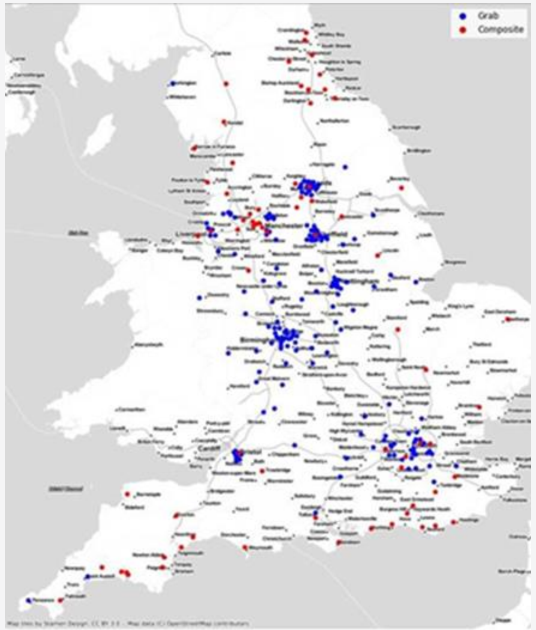
Food and drink

Over the last year, **granulated white sugar** saw the largest increase at **54%.**

Case Study: Wastewater testing and COVID-19

UK Health Security Agency

- Wastewater testing involves taking samples from wastewater and testing it for COVID-19.
- This was detected earlier than when symptoms emerge in people.
- There were 500 sites across England at it’s peak, testing at a community level to cover 40 million people at once.

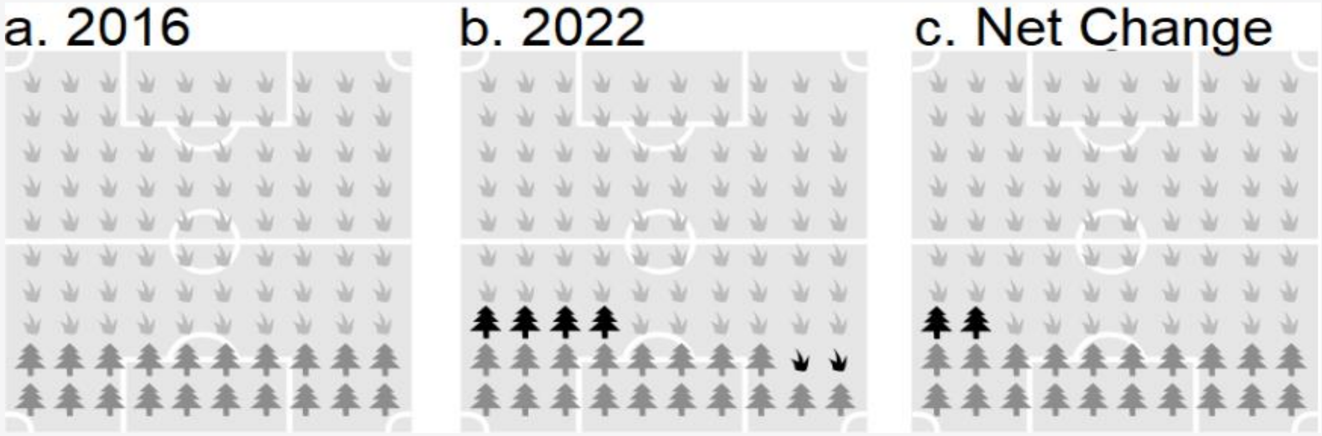


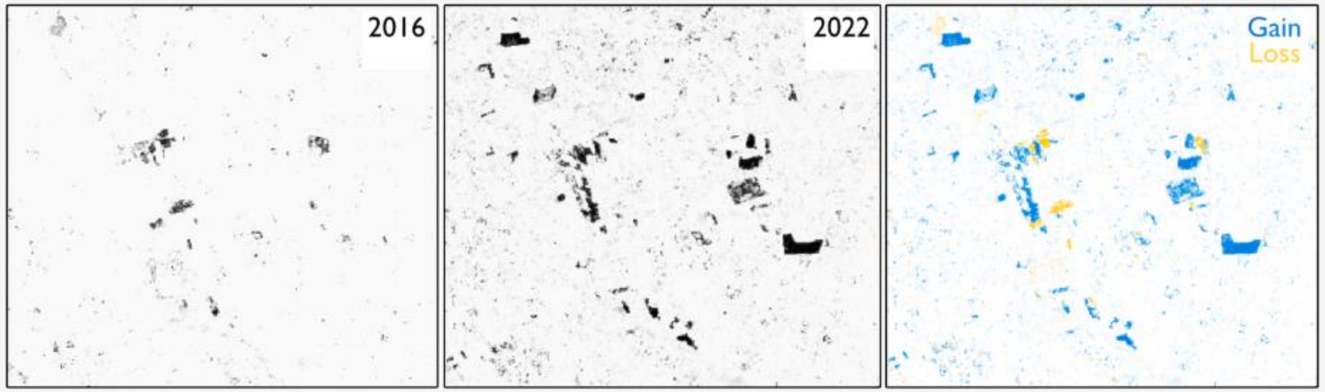
Case Study: Tree coverage in Eastern Uganda

- Used European Space Agency open access Copernicus Sentinel 2 satellite imagery to identify changes in tree cover from 2016-2022.

- Found a net positive change in tree cover, particularly near the Mbale Tree Programme’s nursery sites.

- Relationship is a correlation, more comparison to a baseline is needed to establish whether the programme is the **cause** of the net gain.





Case Study: ONS R&D Survey

The Challenge

- Producing official statistics for publications is a key problem: as it is a time consuming meticulous process
- It is time consuming as the analysis has to pass through multiple systems and multiple individuals
- The systems are diverse and do not always conform to good software engineering practice

Solution

- Use of software engineering tools and techniques such as version control.
- Automated generation of tables/charts and statistical verification
- Process from data storage to report generation

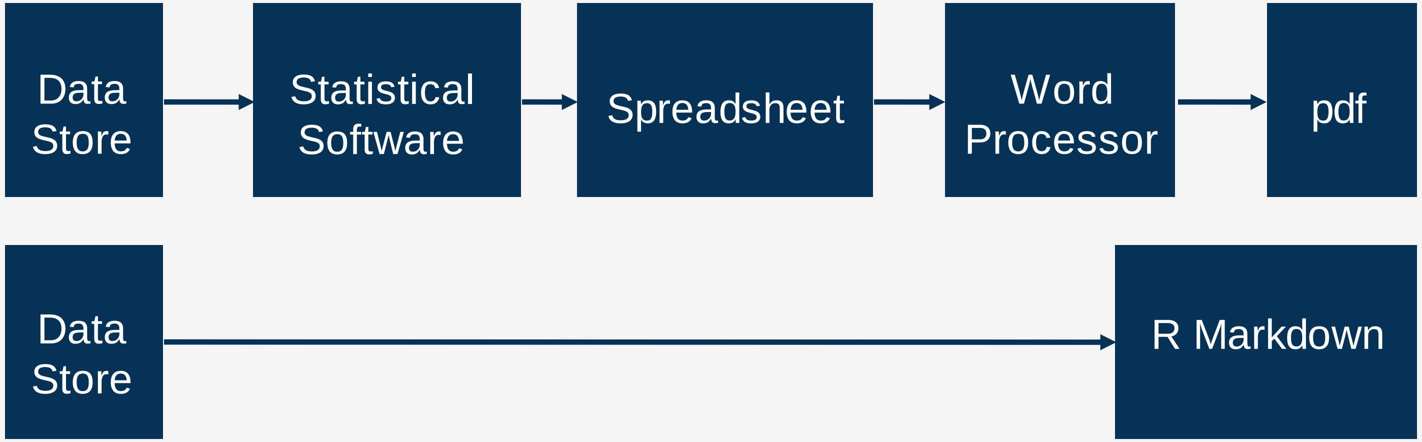
£118m

Estimated annual efficiency savings across government stats publications

£8.8k

Estimated average annual saving per publication

R&D Survey RAP



Emerging Technologies – Large Language Models (LLMs)

- Artificial Intelligence (AI) algorithm that uses deep learning techniques and massive datasets to understand, summarize, generate and predict new content.

- The campus is designing a LLM that runs as a search engine on the ONS and gov.uk websites, producing much finer grained outputs and categorizations, allowing for better search results among the 88,000+ articles.

ChatGPT

Definition – Large Language Model

Let’s hear from you!

Data Science Capability Level	Associated Roles	Programmes Available
Awareness	Any with a curiosity for data and upskilling in this field.	• Art of the Possible • Data Masterclasses • Introduction to Python/R
Working (Undergraduate)	Trainee/Junior Data Scientist	• Public Sector Data Science Graduate Programme • Apprenticeships • DS Accelerators • International Accelerator • Learning Hub Courses
Practitioner (Graduate)	Data Scientist	• Masters in Data Analytics for Government (MDataGov) • CPD Modules • MSc Projects
Expert (Post-Graduate)	Senior Data Scientist	• PhD Placements • Expert capability building sessions • Community of interest

The Data Science Competency Framework

- Describes the role of Data Scientists and the skills required.

- The job levels in this role are detailed, starting at trainee data scientist, up to head of data science.

- Each of these has specific skills as well as the capability level of that skill detailed (we can’t be masters of everything!)

The Data Scientist Role Profile

Core Competencies

- Applied maths, statistics and scientific practices
- Data engineering and manipulation
- Data science innovation
- Delivering business impact

- Developing data science capability
- Ethics and privacy
- Programming and build
- Understanding product delivery

Accreditation

- The Alliance for Data Science Professionals (AfDSP) has come together to create industry wide professional standards and certifications for Data Science.

- Aims to give Data Scientists an identity of their own in the professional landscape, like Chartered Data Analysts and Statisticians have.

- Currently two levels, Data Science Professional and Advanced Data Science Professional.

AfDSP

The Alliance for Data Science Professionals

The Data Science Campus (DSC)

- DSC was established within the Office for National Statistics (ONS) in 2017.

- Our mission is to work at the frontier of Data Science and AI, building skills and applying tools, methods and practices to create new understanding and improve decision making for the public good.

- Breaks down to two goals: Investigate new data sources for the public good and build Data Science capability for the benefit of the UK.

Data Science Campus

[Data Science Campus Website]

Capability Building at the Campus

Data Science Campus Faculty

Data Science Graduate programme

Data Science Apprenticeships

Masters in Data Analytics for Government

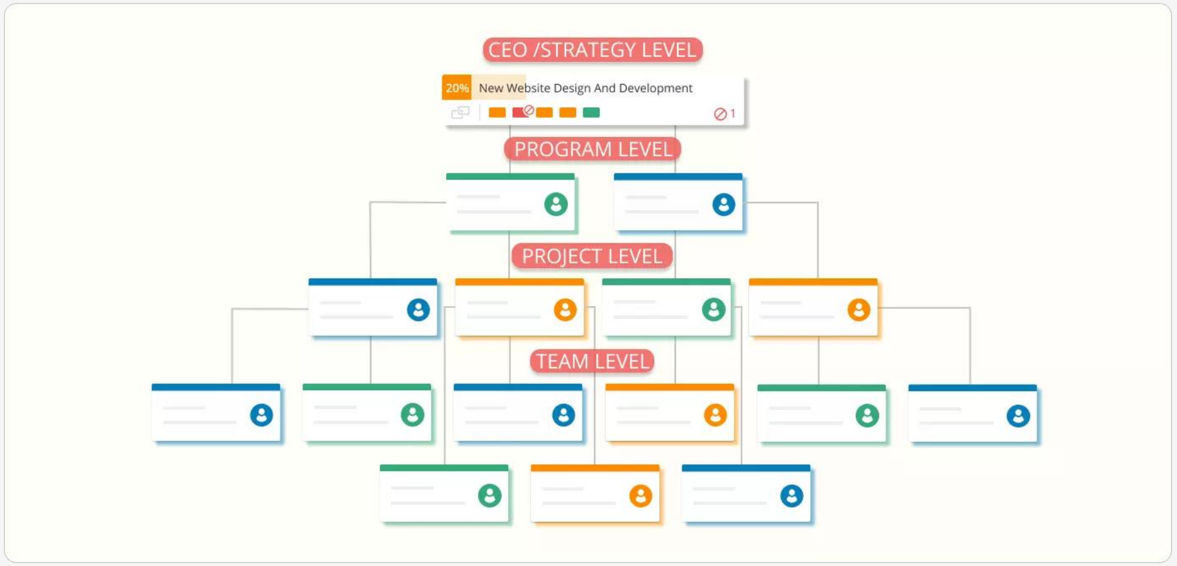
Data Science and Data Viz Accelerator programmes

International Data Science Accelerator Programme

How do we decide on our projects?

- The Data Science Campus Portfolio Board is the decision-making body that controls which projects the campus undertakes.

- Decisions underpinned by Strategic Objectives, Key Performance Indicators (KPIs) and Accountability Framework Objectives (AFOs).



The Cross Government Public Sector Data Science Community

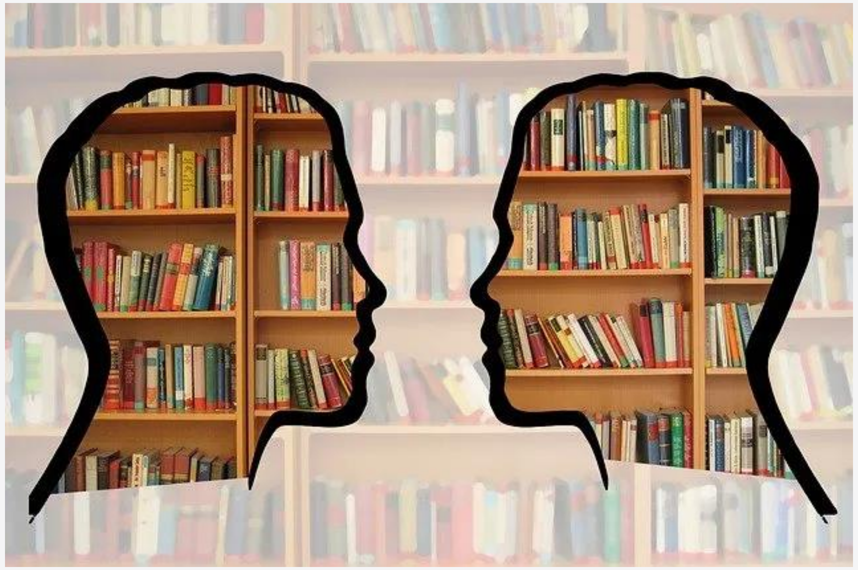
- Brings together data scientists and analysts to build capability across the public sector.

- Enable knowledge sharing and networking, promoting collaboration and learning at all levels.

- Subcommunities provide members with specialism to network and promote best practice in their area of expertise.

Get involved with the community

- Cross-government data science Slack
- Knowledge Hub group
- Monthly meetups
- Mailing list
- Living Library



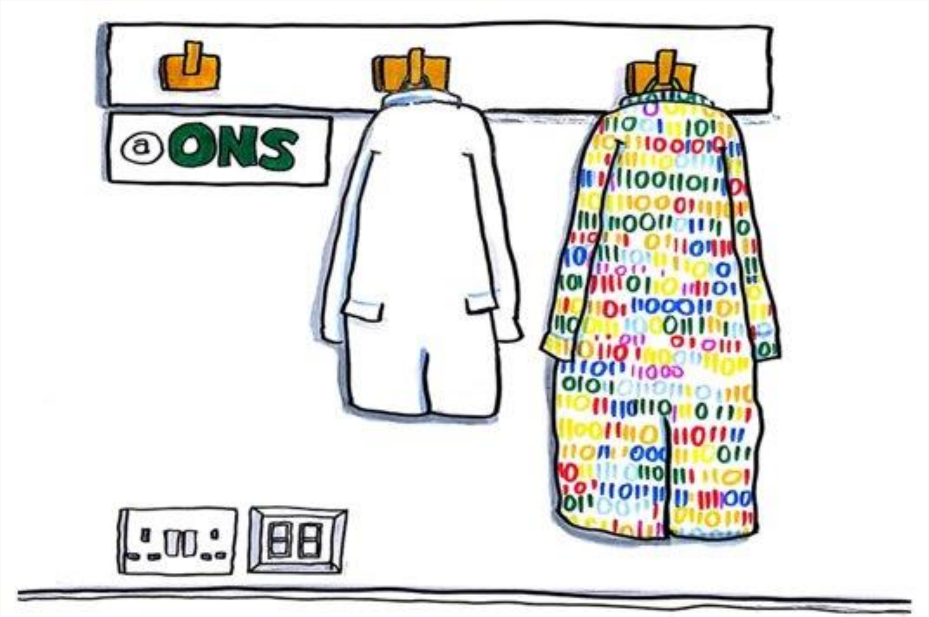
[Cross-Gov and Public Sector Data Science Community]

Contact us!

Campus: datasciencecampus@ons.gov.uk
Faculty: Data.Science.Campus.Faculty@ons.gov.uk

Accelerators: Data.Science.Accelerator@ons.gov.uk
International.Data.Science.Accelerator.Programme@ons.gov.uk
MDataGov: acp@ons.gov.uk

Apprenticeships: GSS.Careers@ons.gov.uk
Community of Interest:
Government.Data.Science.Community@ons.gov.uk



Resources

Shopping prices comparison tool

EMPH Wastewater monitoring of COVID-19

Measuring change in tree coverage in Uganda

Research indices using web scraped price data

ONS Research and Development Survey

Questions?

17 October 2023

Data Science Campus