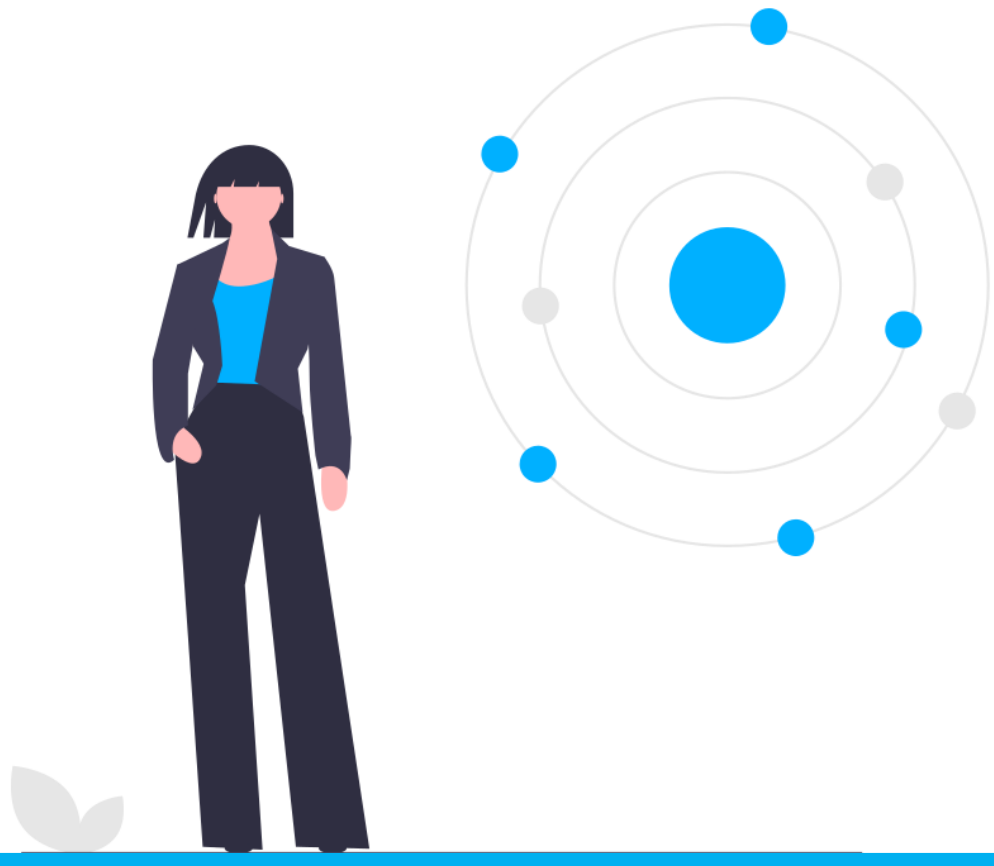
WHITE PAPER
SERIES

2021

Executive Guide To Data Science and AI



adsp.ai

APPLIED DATA SCIENCE

Introduction

This series of white papers lays out the essential steps you need to take to ensure your business stays at the forefront of an increasingly data-driven and automated world.

We cover everything you need to know about growing a data science team, the current trends in machine learning, data visualisation, essential data science tools and technologies, automation, deployment, how to prove value with data science and also give our thoughts on the next 'big ideas' that are set to take the world of AI in business by storm.

We reveal the best approach to implementing an AI strategy across each of these areas, providing expert advice and industry best practices throughout the guide. We hope you enjoy it!

David Foster
Partner
Applied Data Science Partners (ADSP)

APPLIED
DATA SCIENCE

Focus Areas

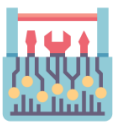
This whitepaper series is a practical guide to developing your data strategy into 2021 and beyond.

It is organised around 8 commonly occurring focus areas that we encounter when working with data science teams from 1 to 100 people strong.

1 Team

2 Machine Learning

3 Visualisation



4 Tools

5 Automation

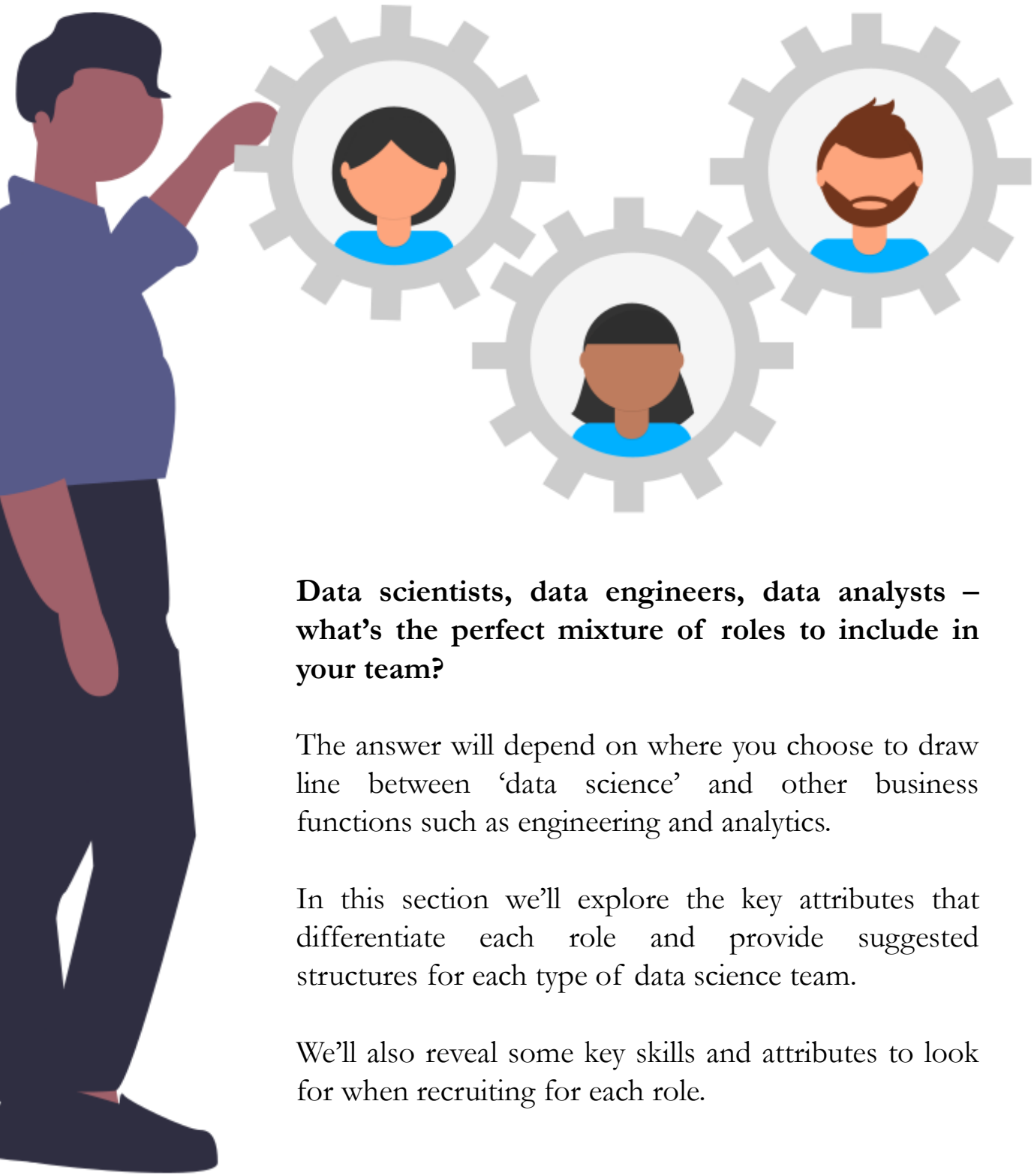
6 Deployment

7 Proving Value

8 Big Ideas

APPLIED
DATA SCIENCE

1 Team



Data scientists, data engineers, data analysts – what's the perfect mixture of roles to include in your team?

The answer will depend on where you choose to draw line between 'data science' and other business functions such as engineering and analytics.

In this section we'll explore the key attributes that differentiate each role and provide suggested structures for each type of data science team.

We'll also reveal some key skills and attributes to look for when recruiting for each role.

APPLIED
DATA SCIENCE

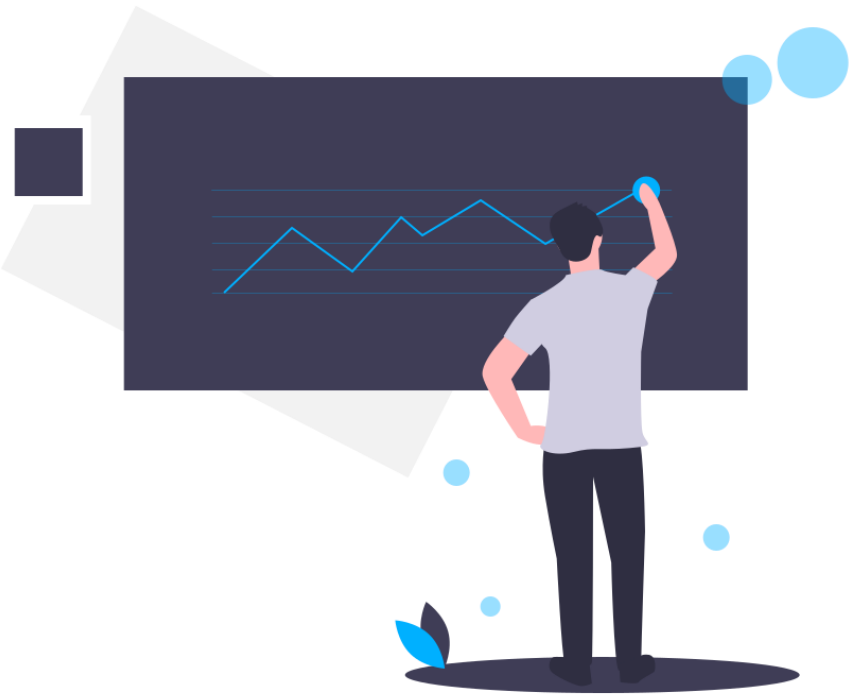
Data scientists

Data scientists have a wealth of practical expertise building AI systems for a range of applications. They bring deep expertise in **machine learning**, **clustering**, **natural language processing**, **time series modelling**, **optimisation**, **hypothesis testing** and **deep learning** to the team.

The most common data science languages are **Python** and **R** – **SQL** is also a must have skill for acquiring and manipulating data.

Data scientists are **inquisitive, curious** and pay **great attention to detail**. They use these skills to ensure trained models are accurate and free from unwanted bias.

> A typical day involves testing a range of ML models, extracting data using SQL and building a validation framework to ensure my models are accurate on unseen data. Text modelling is my specialism – the more complex the documents, the better!
>
> **Data Scientist - ADSP**



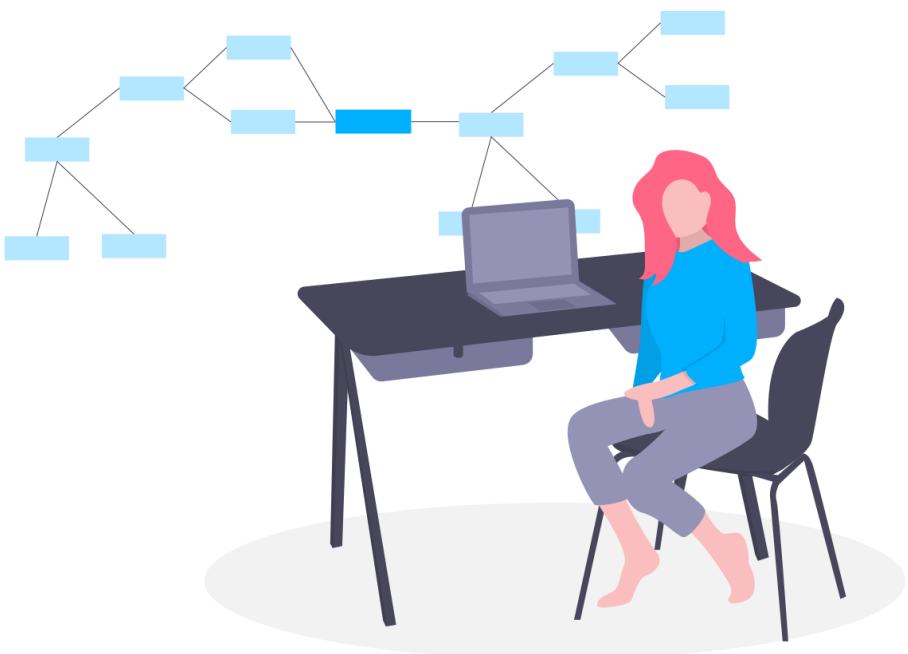
Data engineers

Data engineers understand how to build robust **data pipelines** from source to output. They build **production-ready systems** using best-practice **containerisation** technologies, **ETL tools** and **APIs**. They are skilled at deploying to any **cloud** or **on-premises** infrastructure.

Data engineers must be comfortable working with platforms such as **Azure**, **AWS** and **Google Cloud Platform. Docker** is also a hugely beneficial skill for seamless deployment.

Data engineers are **methodical** and **innovative**. They use these skills to ensure deployments are **robust** and stand the test of time.

> My job is to ensure that the models built by the data scientists are deployed seamlessly into our clients' infrastructure. I particularly enjoy the challenge of building pipelines from multiple disparate data sources and automating everything as much as possible.
>
> **Data Engineer - ADSP**



Data analysts

Data analysts build interactive **dashboards** and **reports** that stakeholders use to consume model outputs and insights. They are experts at **asking the right questions** and being **inquisitive** to ensure that the output from your data science team is **actionable**, **practical** and **functional**.

Data analysts have expertise using visualisation platforms such as **Tableau** or **PowerBI** – **SQL** is also crucial for the role.

Data analysts understand how to present data **beautifully** and are excellent **communicators**. They use these skills to ensure they can present data in an **engaging**, **insightful** manner.

> The most important part of my role is ensuring that the dashboards I build are easy to use and refreshed daily. I work really closely with our clients to understand what information they want to see and in what format - I particularly like this communicative aspect of my role.
>
> **Data Analyst - ADSP**

Building a modern data science team

The modern data science team should contain all three roles – a common mistake is to hire only data scientists and no data engineers or data analysts. This has the effect of making the team less efficient when it comes to model deployment and handling ad-hoc requests for analysis.

The balance of the roles should ultimately be determined by the level of responsibility the data science team takes for these crucial supporting tasks that are not part of the standard remit of a data scientist, but are more related to engineering or traditional business analytics.

Here we present four team structures that we've seen work well in practice – exact team sizes will vary with size of company.

The Data Science Research Team

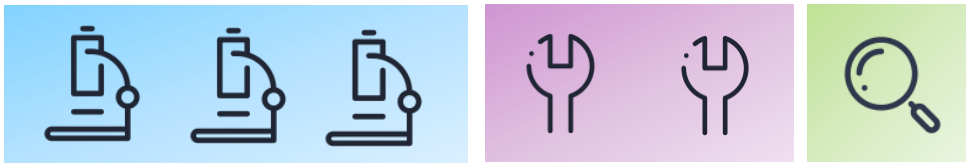
This structure consists mostly of data scientists and is best suited to larger companies who want to experiment with a wide range of data science and AI applications. The data engineer within the team liaises with an existing engineering team to translate work from the data scientists into live applications. A existing analytics team handles ad-hoc requests from the business, allowing the data analyst within the team to focus on presenting results from deployed models to key stakeholders.

APPLIED
DATA SCIENCE

The Analytics Team, Transformed

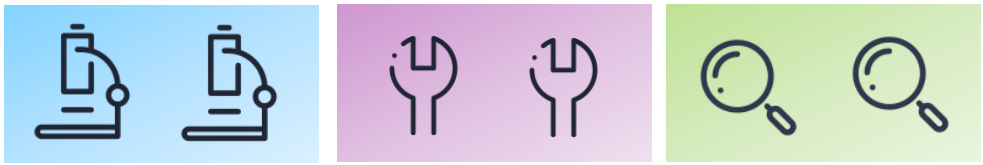
Many data science teams start out life as an existing analytics team. In this structure, analysts are upskilled with specific data science capabilities and data engineering skillsets are brought into the team. It offers great progression opportunities for existing staff and also allows the company to gradually experiment with new technologies without major investment or restructuring. There is then the option to split into a dedicated data science function, as the team grows in capability and confidence. Providing training opportunities for staff is a key component of this structure.

The Data Science and Engineering Team



This structure is a self-contained team that is able to deploy models autonomously, through enhanced in-team data engineering capabilities. Bringing the deployment step inside the team reduces bottlenecks in the process from experimentation to go-live. For example, a model to predict customer churn could be efficiently deployed inside a Docker container on cloud architecture managed and maintained by the team's data engineers, and visualised by the data analyst within Tableau. Crucially, in this structure, ad-hoc requests for analysis are still handled by a separate analytics team, allowing the data science team to focus purely on model build and deployment.

The Data Centre of Excellence



Finally, the Data Centre of Excellence structure aims to bring all data related activities across the business under one roof, for complete strategic alignment. This avoids any duplication of activities or confusion between the respective roles of the data science and analytics teams. In this structure, best-practice methodologies are standardised across the team and common ad-hoc requests can be quickly automated through close alignment between the data engineers and data analysts. This structure requires oversight and direction at the board level – for example through a Chief Data Officer or CIO.

In summary…

Whilst there are no concrete rules to building a data science team, there are certainly best-practices that should be followed.

In this section we have presented the three key data roles that should form part of any data science team and four team structures that allow you to quickly realise the potential that data science and AI can bring to your company.

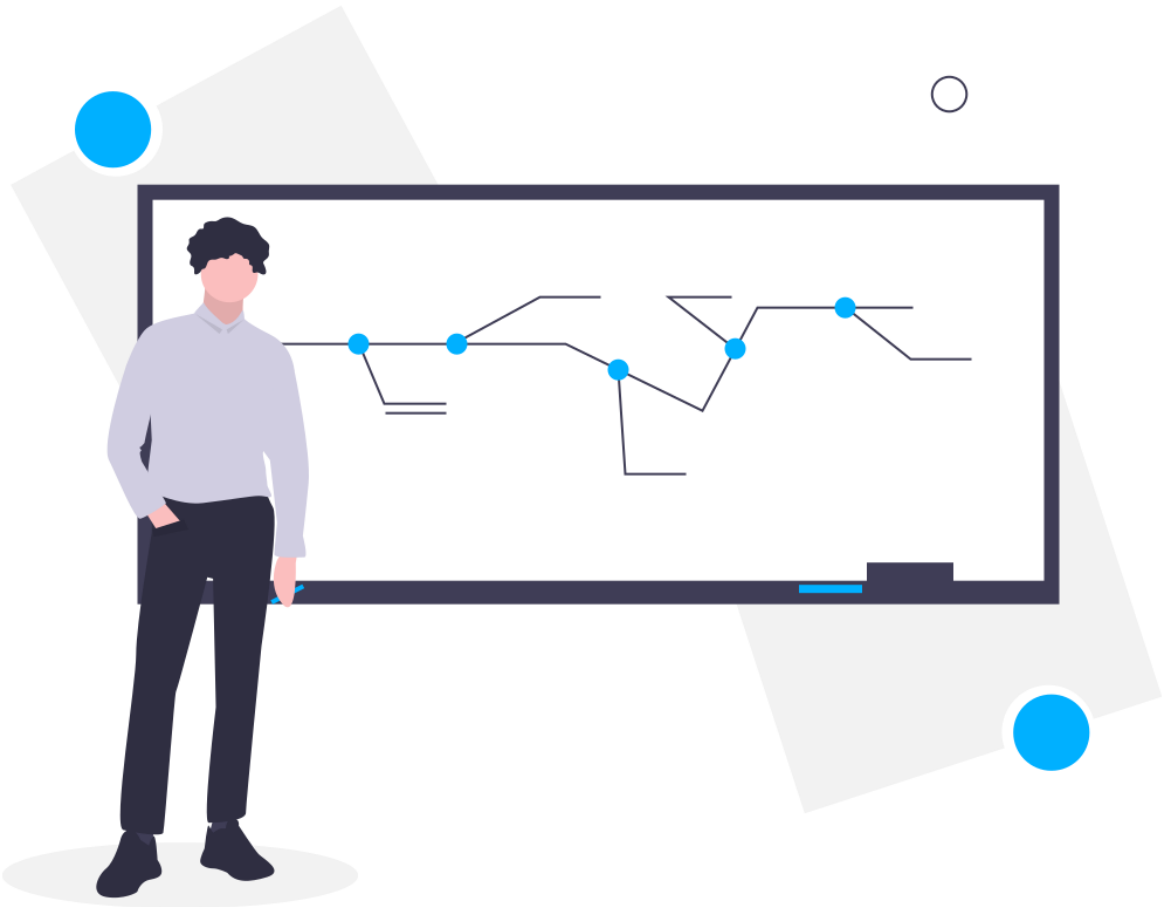
ADSP is a trusted partner for companies wanting to implement cutting-edge data science and AI solutions. Get in touch with us at hello@adsp.ai to hear more.

2 Machine Learning

Machine learning (ML) is the driving force behind most modern AI applications. Instead of hard-coding rules, ML is the process through which a machine can discover the rules for itself, with the overall aim of minimising error.

Much has been written about the different types of machine learning – for example, **supervised learning** for prediction of a value or label and **unsupervised** learning for segmentation.

In this section, we look beyond 'standard' ML practices and explore six ML trends that will set you apart from the pack in 2021.



APPLIED DATA SCIENCE

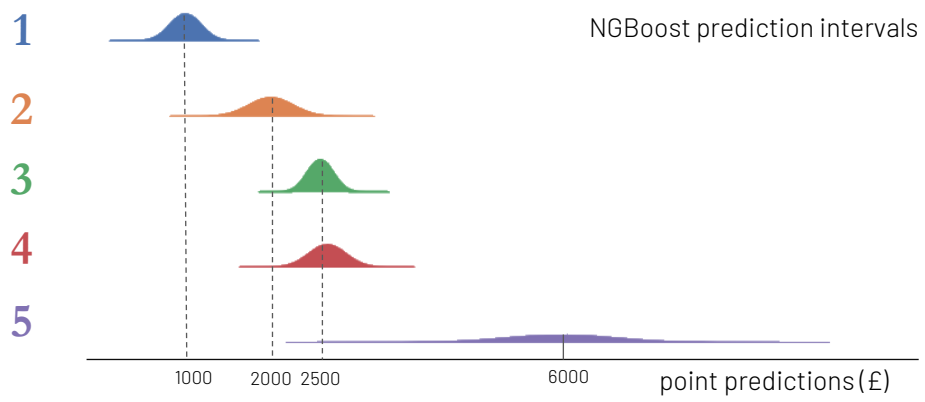
From XGBoost to NGBoost

XGBoost is a gradient boosting technique that enables you to build supervised machine learning models on structured data. We find that it is a more powerful alternative to random forests and more easily scalable to high-dimensional data than techniques such as support vector machines.

We also love using a new library called **NGBoost** – Natural Gradient Boosting for Probabilistic Prediction. Instead of providing only a point estimate for each observation, NGBoost outputs a **prediction interval**, allowing for a much richer interpretation of the predictions. It achieves this whilst still taking advantage of the power of the original XGBoost algorithm.

CASE STUDY – Car price prediction

Below are NGBoost price prediction intervals for 5 used cars.



NGBoost allows us to see that cars 3 and 4 have similar point price predictions, but the model is more confident in its prediction for car 3, as it has a narrower prediction interval – i.e. we can say:

> "There is 90% chance the price for car 3 falls between £2400 and £2600", whereas for car 4, there is a 90% chance the price falls between £2200 and £2800.

Self-supervised learning

A major hurdle of supervised machine learning is acquiring enough labelled data on which to train a ML model. **Self-supervised learning** is a technique through which an algorithm creates labels from the data itself, without human supervision.

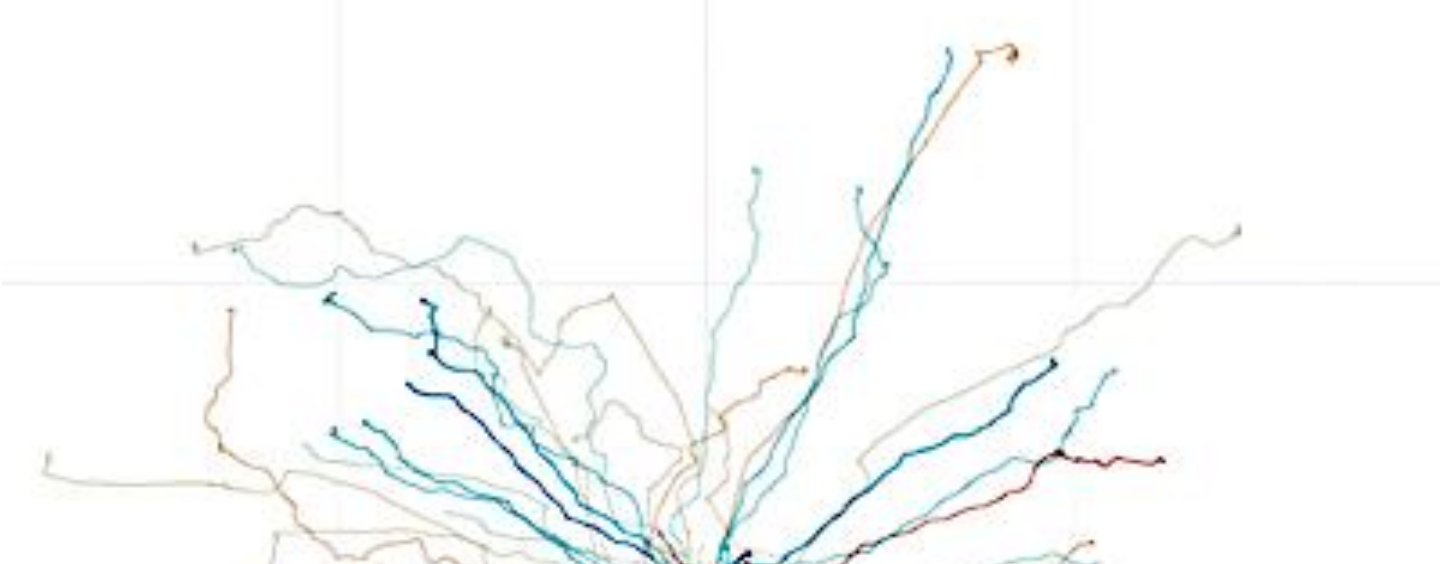
Self-supervised learning can be used for **image and text generation tasks** – for example, by removing elements of a sentence or picture and training the AI to fill in the blanks.

CASE STUDY – Detecting Driver Fraud

A insurance company wanted to develop a machine learning algorithm to identify if a given journey was taken by a given driver. Data relating to known examples of driver fraud is limited but self-supervised learning can solve the problem.

To create the labelled dataset, a random number of 'fraudulent' trips that were not driven by the driver of interest were planted in amongst real trips. The fraudulent trips were trips driven by other drivers in the dataset. A machine learning model is then trained to predict which trips were planted by the algorithm.

Read more: https://www.kaggle.com/c/axa-driver-telematics-analysis/



APPLIED DATA SCIENCE

Human-in-the-loop AI

We would expect students to ask questions to clarify understanding during training sessions – training an AI system is no different.

Human-in-the-loop AI systems siphon off portions of validation data for human review, especially where prediction confidence is low or prediction error is high. During development, the AI system can receive targeted feedback (additional labelled data) on which to continue training and in a live environment can defer marginal predictions to a human for manual consideration.

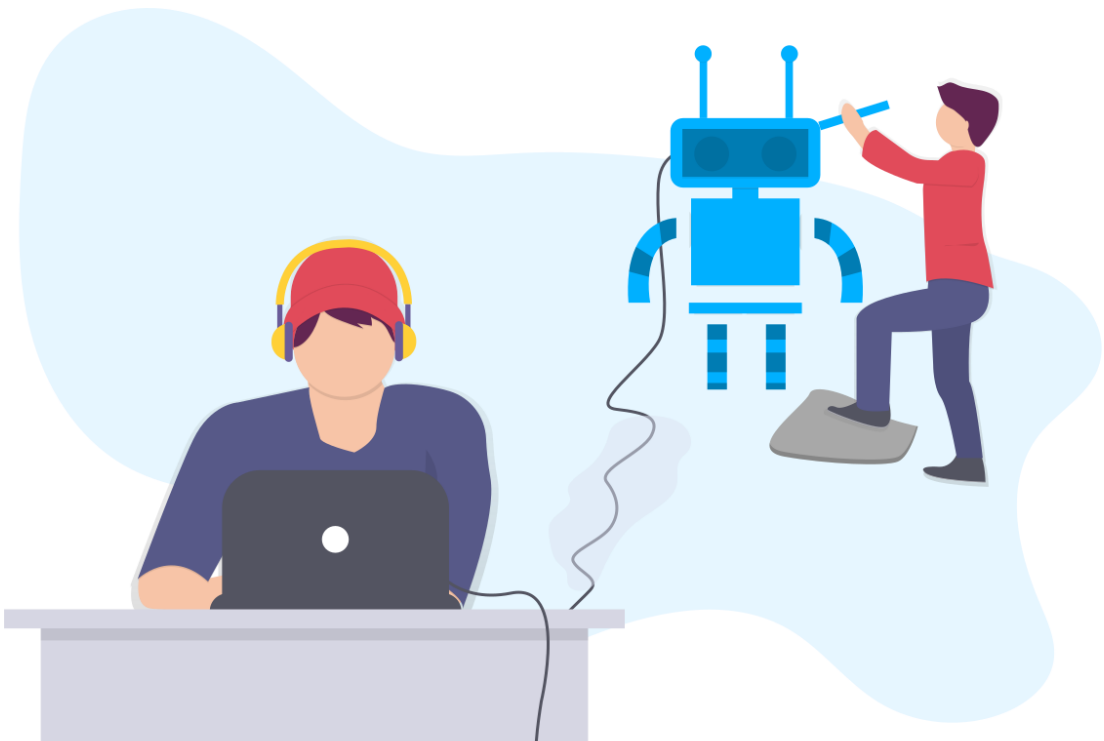
By designing your AI system to tell you where it needs extra help, you ensure it remains **flexible to deal with uncertainty,** has the ability to **adapt quickly to new scenarios** and **finds blind-spots** in prediction capability.

EXAMPLE

AWS Augmented AI

Whilst it is possible to set up your own human-AI loop, there are some platforms that offer this service out of the box, such as **AWS Augmented AI (A2I).**

Read more: https://aws.amazon.com/augmented-ai/

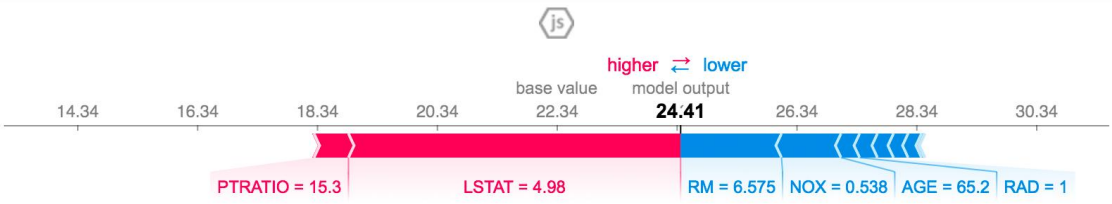


APPLIED DATA SCIENCE

Explainable ML

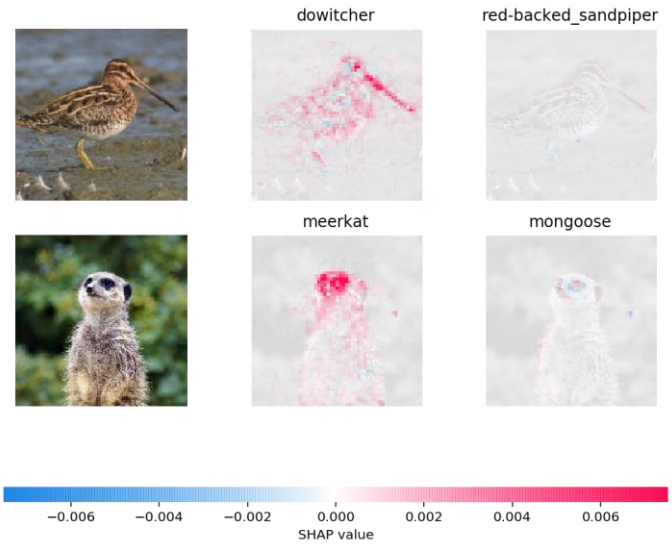
In order understand the rationale behind a ML model's predictions, we recommend using **SHAP** as a wrapper around the model output.

SHAP values directly measure the impact of a given feature on each prediction - some features pull up the predicted score and some push it down. The overall prediction can therefore be broken down as the sum of the individual SHAP values.



source: https://github.com/slundberg/shap

You can use SHAP values to build **human-interpretable explanations** of every prediction. This is invaluable for providing users of your ML model not only with the 'what' but also the '**why**' for every prediction made by the model.



source: https://github.com/slundberg/shap

SHAP works for images and text too – highlighting exactly which pixels or words are pushing the overall prediction towards a particular category. It's very easy to use and a game-changer for explainable ML.

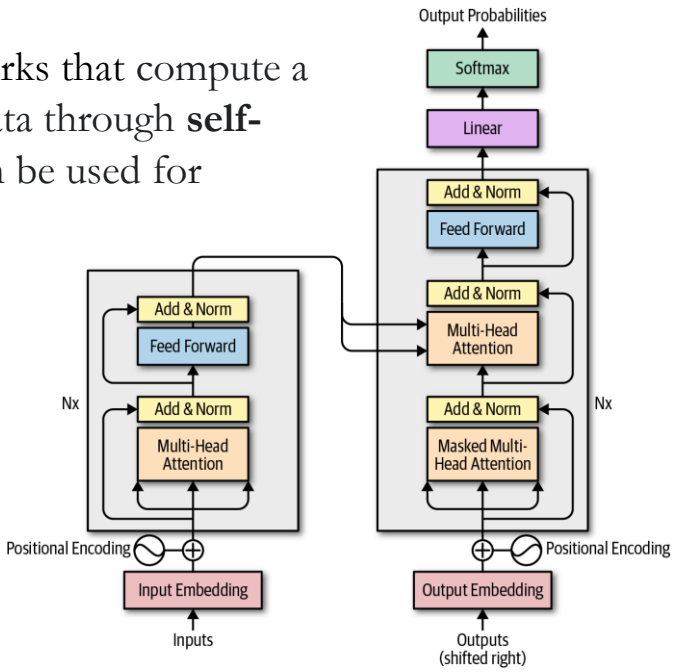
APPLIED DATA SCIENCE

Transformers for NLP

Natural Language Processing (NLP) has been turbo-charged by the recent development of a new kind of machine learning model called a **Transformer**.

Transformers are neural networks that compute a representation of sequential data through **self-attention** mechanisms and can be used for supervised or generative tasks.

They surpass the predictive power of more traditional NLP techniques such as recurrent neural networks as they are easily able to capture **long-range dependencies** in sequential data and are **highly parallelisable**.



The Transformer architecture
source: https://arxiv.org/pdf/1706.03762.pdf

The Transformer architecture is used by powerful NLP models such as Google's **BERT** and OpenAI's **GPT-3** and can also be trained from scratch using open-source libraries such as **Keras** and **PyTorch**.

KEY RESOURCES

You can learn more about the uses and implementation of Transformers through the **Keras** and **PyTorch** websites**:**

https://keras.io/

https://pytorch.org/

APPLIED DATA SCIENCE

Eliminating AI bias

It is essential that companies develop a framework to ensure deployed algorithms do not unfairly discriminate against certain characteristics. Below we outline **three simple steps** you can take to mitigate against unwanted bias in AI models.

Build an AI bias test suite
For every model, build a test suite that checks your model output against certain controlled scenarios. Every time you retrain the model, run the test suite to ensure unwanted bias hasn't crept into the training process.

Know your data
Most causes of AI bias stem from the training data itself. Ensure your data contains sufficient quantity across all groups and that variables aren't acting as a proxy for other protected characteristics.

Transparency is key
To eliminate unwanted bias, you must first understand **why** the model is biased (see Explainable ML section). You can also apply monotonic constraints to ensure pairwise relationships between each feature and the response can be explained.

ADSP is a trusted partner for companies wanting to implement cutting-edge data science and AI solutions. Get in touch with us at hello@adsp.ai to hear more.

3 Visualisation

Data visualisation isn't only about building dashboards with lots of charts and filters.

Sometimes, a single number, presented in the right place, at the right time, in the right format to the right person can be far more compelling.

We define data visualisation as simply the process through which raw analysis and statistics are converted into insight and actions via a human interpreter.

In this section we'll cover the essential steps to ensure data visualisation at your company delivers this simple but powerful function.

APPLIED DATA SCIENCE

Choosing a platform

Most modern businesses use a central platform to host and group visualisations together, so that users can easily access the information they need.

Two of the most popular are **Tableau** and **PowerBI.** Both platforms are mature and an excellent choice for a centralised data visualisation platform.

The best way to choose which platform is right for your team is to run a trial project for a single use-case and grade each platform against the following six attributes:

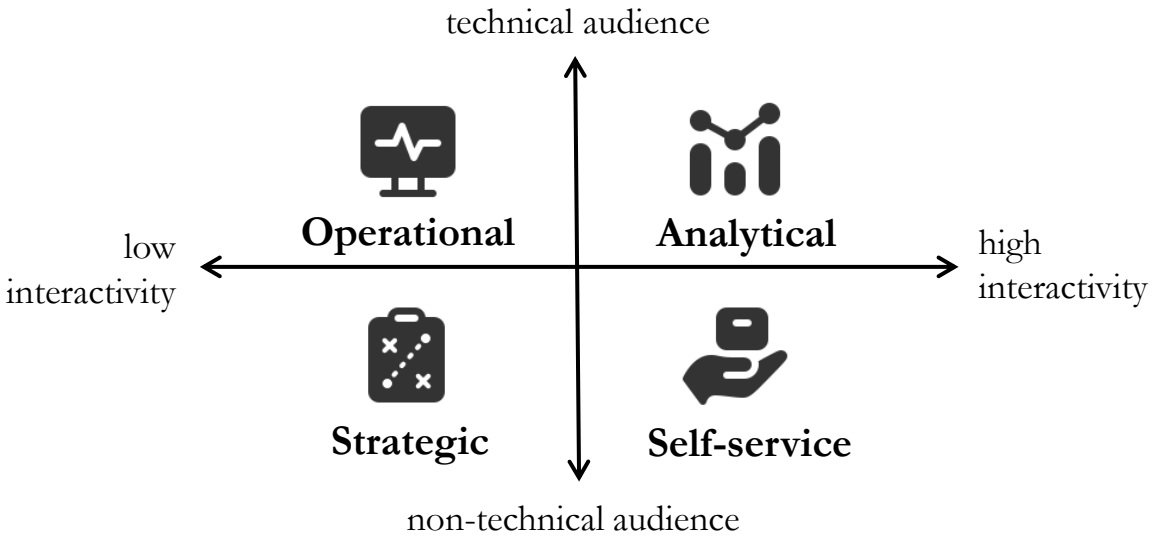
Ease of use	How easily are analysts able to quickly design and deploy useful dashboards and visualisations?
Functionality	Does the platform allow you to build the types of visualisations and reports users require?
Integration	Can you connect the platform directly to your data and does the platform fit within your existing architecture?
Accessibility	How would you grant users access to only the dashboards they require? (both internal and external)
Scalability	Test the platform with large datasets and multiple filters – do you foresee any scalability issues?
Pricing	How is the platform priced? If per user, are there minimum requirements? Are there running costs?

It is also important to consider existing skills within the business and the role each team will play in managing the deployment (e.g. content creation, technical administration, training for users etc.).

APPLIED DATA SCIENCE

Types of Dashboard

There are four kinds of dashboard – **operational, analytical strategic and self-service.** These can be defined along two primary dimensions – **target audience** and **interactivity**.



Operational dashboards display real-time metrics to monitor business processes and alert technical teams to anomalies. They are often displayed on communal screens and are less interactive.

Analytical dashboards are used by data science teams to interactively mine datasets for insight. They are often shared within the team but not deployed to the wider business.

Strategic dashboards present key performance metrics to senior management and aim to tell the 'big-picture' story behind the data. Important areas of concern or success are highlighted.

Self-service dashboards allow users from across the business to quickly find data they need through highly interactivity and easy-to-use dashboards. They encourage exploration and discovery.

Your data visualisation platform should be a mixture of these four types of dashboard. Remember to establish which type of dashboard you require before starting the build!

Operational dashboards

An operational dashboard should provide quick access to information in near real-time. A car dashboard is a good example of an operational dashboard – it provides the driver with live information on speed, fuel consumption and warnings during the journey, so that appropriate action can be taken.

Typically, there is less interactivity on an operational dashboard because users do not want to have to 'search' for the answer. It needs to be **fast**, **simple** and **immediately actionable**.

If your users would like more interactivity, then consider building an **analytical** dashboard instead. If they need a higher level view, over a longer time period, a **strategic** dashboard is more appropriate. If you can imagine the dashboard on a big communal screen in the workplace, an operational design is the right choice!

APPLIED DATA SCIENCE

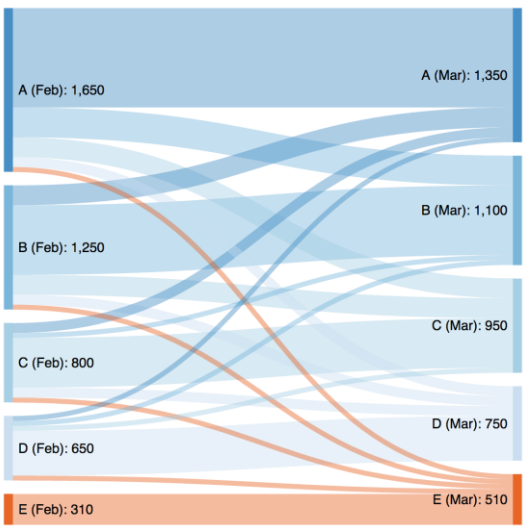
Analytical dashboards

Analytical dashboards allow users to dig into the 'why' behind the data. They are used by analysts and data scientists to explore new datasets quickly, in order to uncover patterns and signal. For example, an analytical dashboard could be used to understand if a dataset of predictions is biased or inaccurate for certain groups.

A common mistake is to build an analytical dashboard instead of a **self-service** dashboard. Whilst both allow the users to filter, the difference is in how the information is presented. Analytical dashboards make use of a variety of different chart types and interactive filters to **subjectively** reveal potential statistical insights to the user. Self-service dashboards on the other hand provide users with **objective** key metrics that are unambiguously defined.

Sankey charts

Sankey charts are a great way to visualise **movement between groups** – e.g. to show how customers have moved between segments month-on-month. This can help analysts recommend actions for each customer segment to marketing.



APPLIED DATA SCIENCE

Strategic dashboards

Strategic dashboards are used by senior management to assess the overall health of the business. They should be **unambiguous, immediately actionable** and at a higher level than **operational** dashboards.

Strategic dashboards are often presented as part of board reports and are therefore less interactive – the key to building a good strategic dashboard is to tell a story, rather than expect the user to find the story on their own. This often means including a **commentary** alongside charts, which ideally is auto-generated but also can be added manually.

Showing change

As strategic dashboards use slower moving data than operational dashboards, showing the change between time periods is just as important than the metrics themselves.

Groups with the biggest change should 'pop-out' from the dashboard, through consistent and intuitive use of colour. Beware that percentage increases can often be misleading when numbers are small – or undefined if originally zero!

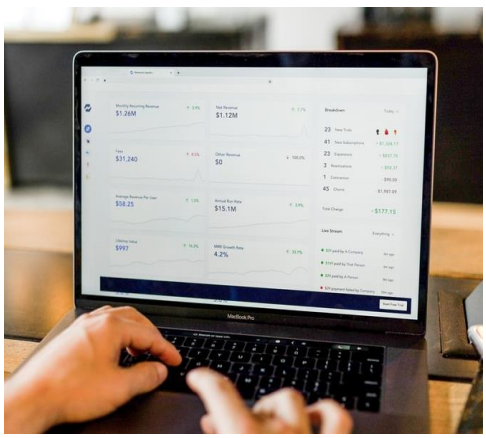
APPLIED DATA SCIENCE

Self-service dashboards

A self-service dashboard has all the interactivity of an analytical dashboard, but is specifically designed for users to quickly access information, rather than for in-depth analysis.

For example, a sales executive might use a self-serve dashboard to find out how much a client spent with the company last year and on what products. The information should be easy to download and not locked within charts – inclusion of a button to 'download' the relevant data into Excel is an often requested and useful feature of self-service dashboards!

Self-service dashboards should take ad-hoc reporting workload away from the analytics team. This means running training sessions for staff how to use the dashboards and showing why they are a quick and easy way to access commonly requested information.



In summary…

In this section we have presented the four types of dashboard that should form part of your overall data visualisation platform – whether in Tableau, Power BI or another vendor.

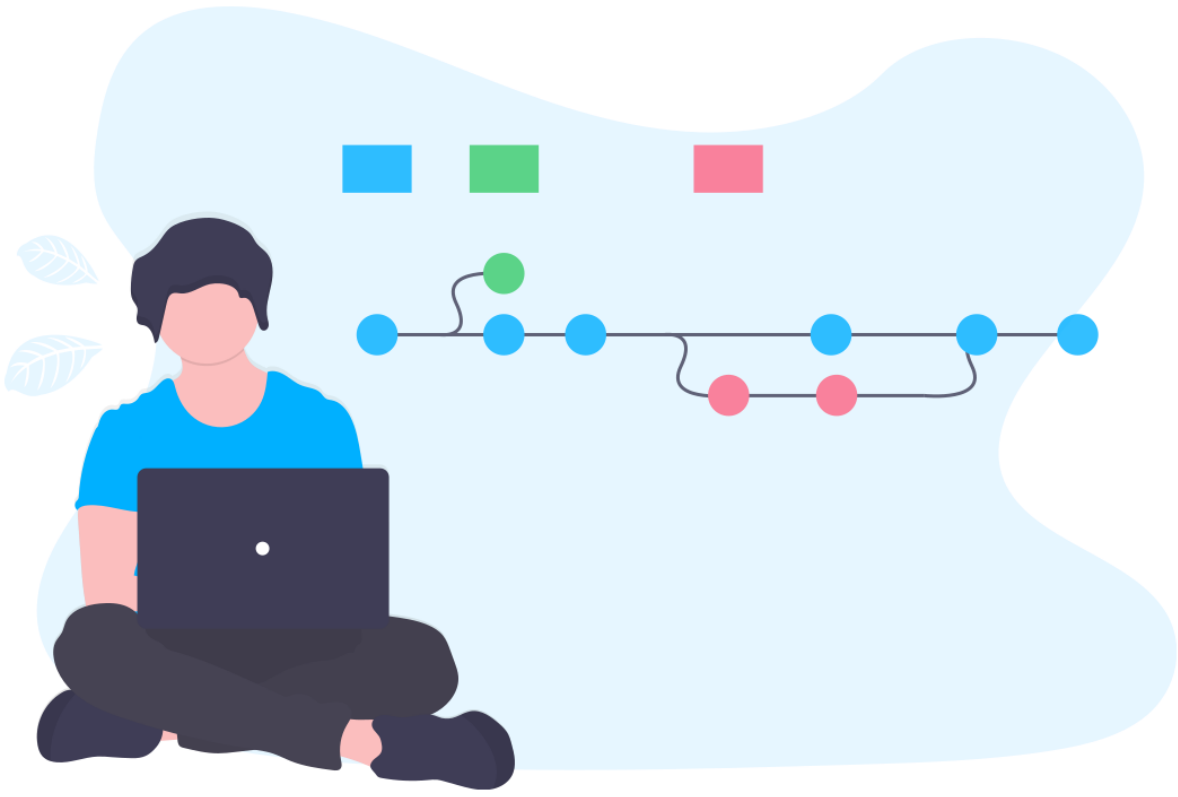
ADSP can help your company build a world-class data visualisation platform from scratch or improve your existing setup. Get in touch with us at hello@adsp.ai if you'd like to hear more.

4 Tools

With a vast array of data platforms, products, frameworks and libraries available, it is sometimes difficult to know how to choose the right tools for your data science team.

In this section we'll present the most commonly used tools across three sections – Coding and Visualisation, Machine Learning and Cloud & Storage, as reported in the 2020 Kaggle Machine Learning & Data Science Survey.

We'll also explore two additional tools not mentioned in the survey that we feel are increasingly important for any modern data team to adopt – Git and Docker.



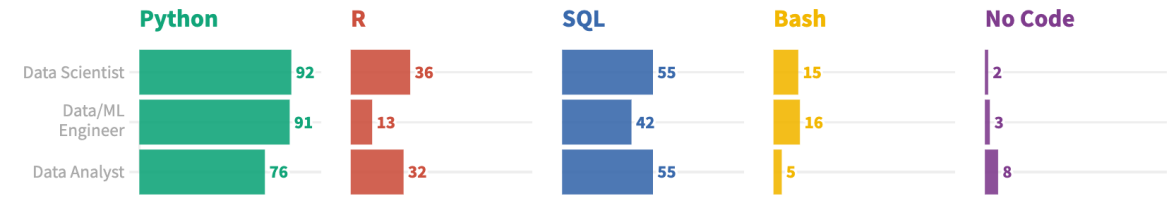
APPLIED DATA SCIENCE

Coding and visualisation

Python and R are the dominant languages for data professionals, supported by SQL and Bash. Whilst Jupyter notebooks are common and useful IDEs for testing ideas, we would recommend more sophisticated tools such as VSCode for writing production scripts. Power BI and Tableau are the most utilised dashboarding tools, with open-source visualisation libraries such as Matplotlib and Seaborn commonly used for ad-hoc analysis and charting.

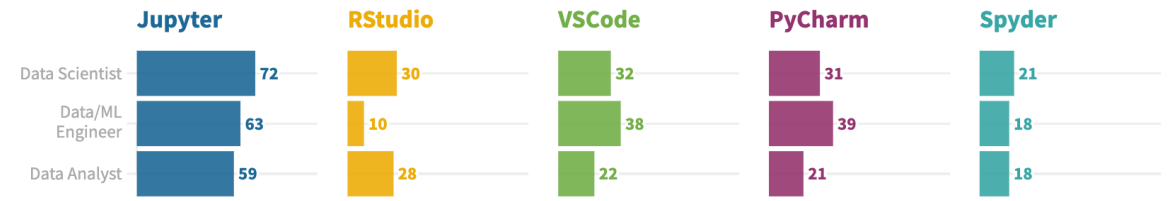
Languages

What percentage of respondents use each language on a regular basis?



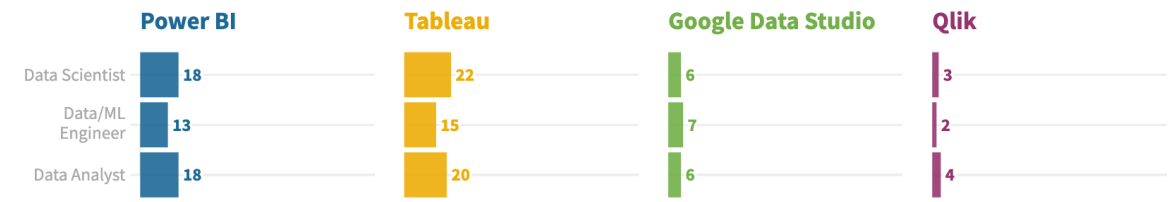
Integrated Development Environments (IDEs)

What percentage of respondents use each tool on a regular basis?



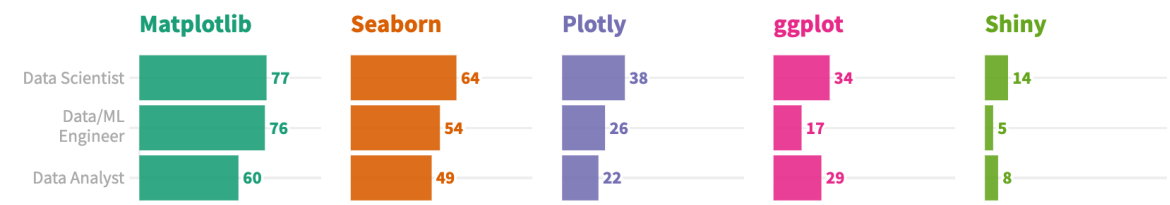
Visualisation tools

What percentage of respondents use each tool on a regular basis?



Visualisation libraries

What percentage of respondents use each library on a regular basis?



Data source: https://www.kaggle.com/c/kaggle-survey-2020

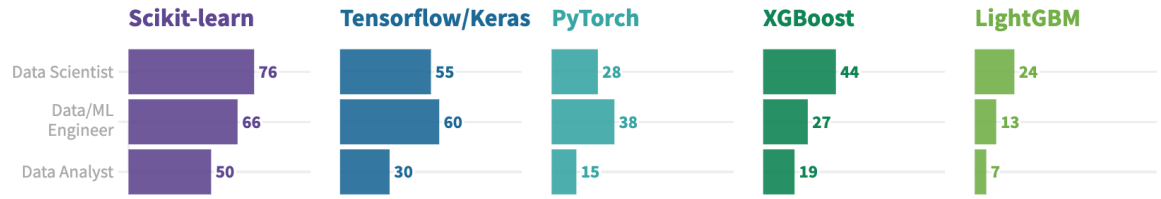
APPLIED DATA SCIENCE

Machine learning

Scikit-learn is the most common general ML library, with Tensorflow and PyTorch focused particularly on neural networks. Gradient boosting is now regarded one of the most powerful modern algorithms for structured data, though regressive and simple tree-based techniques remain popular. Data Engineers are most likely to use computer vision and NLP methods, as both typically involve the deployment of large neural network architectures.

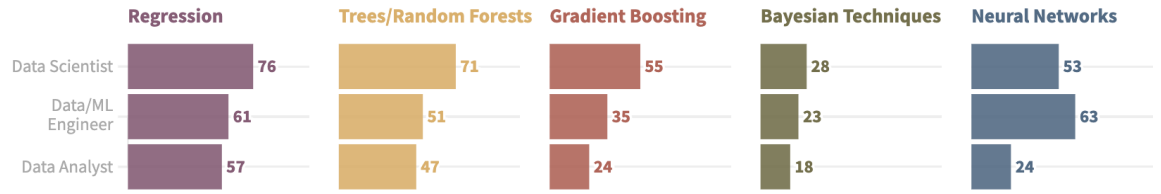
ML Frameworks

What percentage of respondents use each framework on a regular basis?



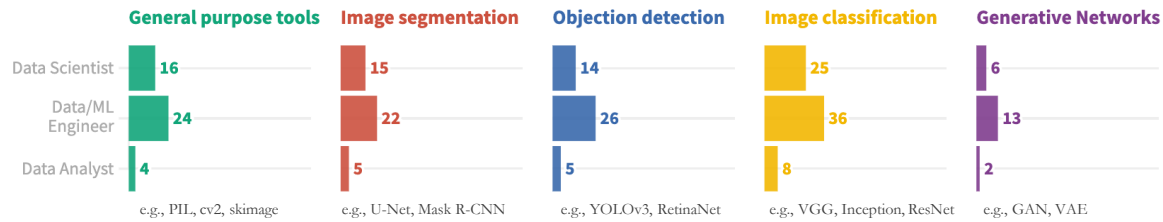
ML Algorithms

What percentage of respondents use each algorithms on a regular basis?



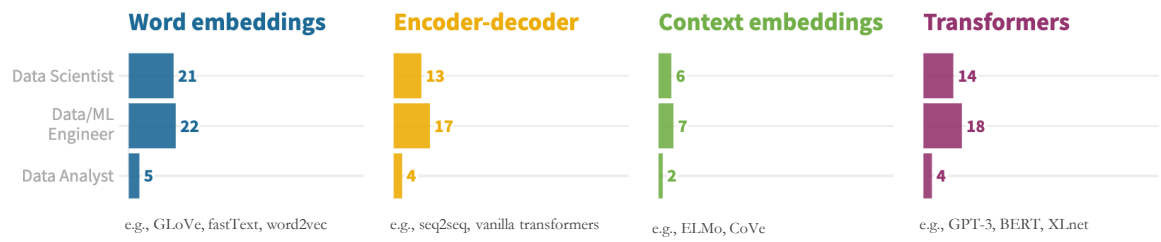
Computer vision methods

What percentage of respondents use each method on a regular basis?



Natural Language Processing (NLP) methods

What percentage of respondents use each method on a regular basis?



Data source: https://www.kaggle.com/c/kaggle-survey-2020

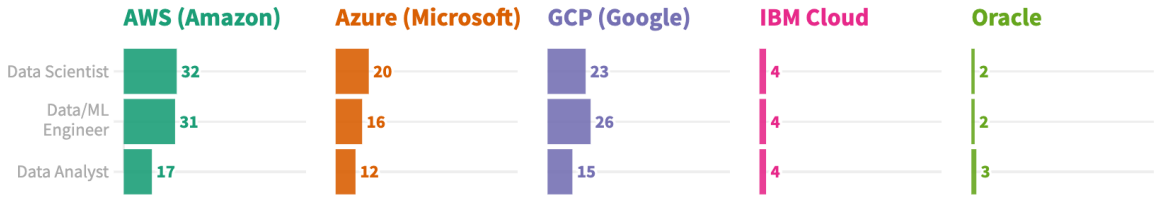
APPLIED DATA SCIENCE

Cloud & Storage

Whilst less than one-third of data professionals report using one of the 'big three' cloud providers regularly, they are a crucial part of any modern, scalable tech stack, often managed by DBA roles. Cloud ML and AutoML are not often utilised, with most data professionals instead opting for self-maintained applications deployed onto virtual machines. MySQL and PostgreSQL are commonly used open-source alternatives to SQL Server.

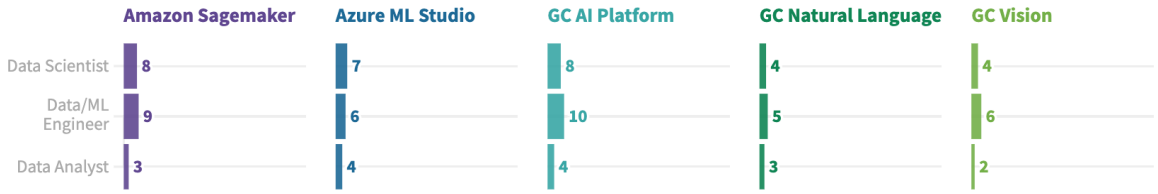
Cloud computing platforms

What percentage of respondents use each platform on a regular basis?



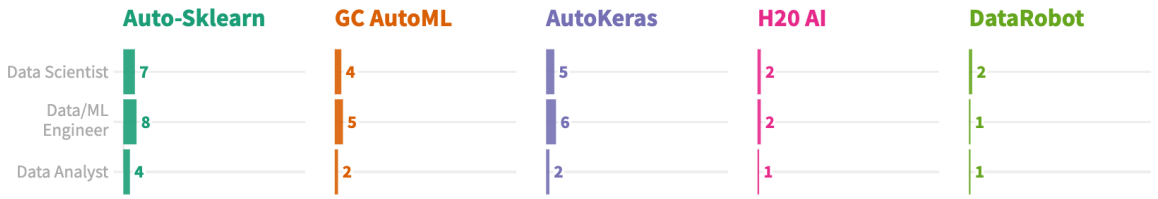
Cloud ML products

What percentage of respondents use each product on a regular basis?



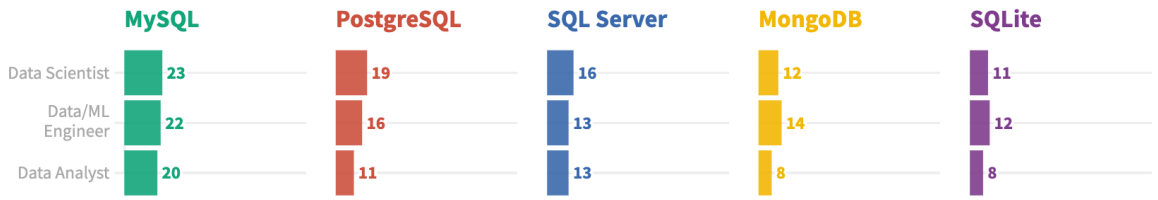
AutoML tools

What percentage of respondents use each tool on a regular basis?



Databases

What percentage of respondents use each database on a regular basis?

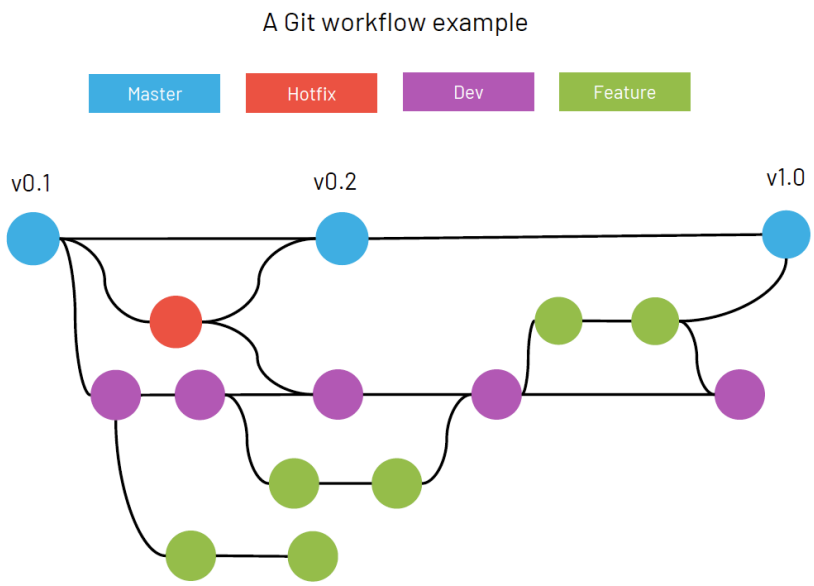


Data source: https://www.kaggle.com/c/kaggle-survey-2020

Git

Coding is messy – it is rarely a linear process from idea to production ready system. To keep track of changes to source code over time, most modern data science teams use **Git**, the most commonly used version control system in the world.

Git allows your team to introduce new features in a controlled manner, roll back to previous versions and make changes to colleagues' code without interrupting their workflow. This is achieved through an elegant system of 'branches' and 'merges', that creates an audit trail of the entire codebase over time.

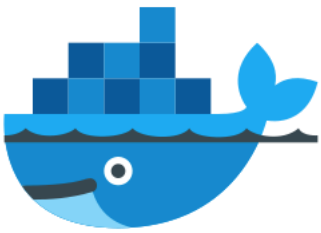


There are also websites such as **GitHub** or **GitLab** where your team can privately upload codebases using Git. Having this remote repository means that other can **pull** a copy of the codebase and start working on it. When they are ready, they can **push** their version back to the remote, for others to review.

GitHub GitLab Bitbucket Google Cloud Source Repositories AWS CodeCommit Azure DevOps

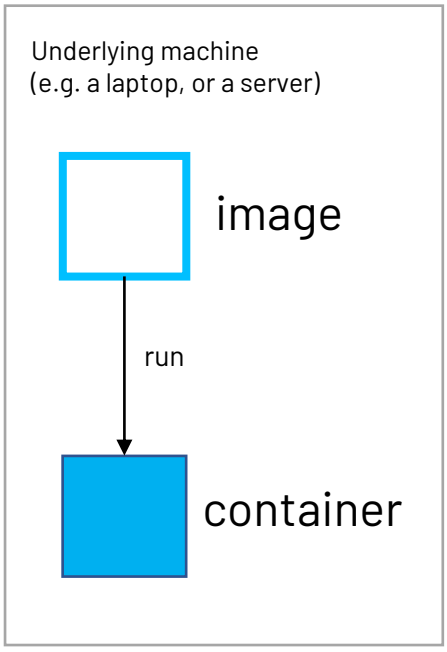
Docker

A frequent frustration when developing data science solutions is ensuring portability. Team members and servers will rarely have the exact same computer setup, meaning the solution may not work when deployed to production or when sharing code between machines.



Docker solves this problem. A Docker **container** is an entire virtual operating system. A running container is therefore a 'machine within a machine' – for example, you can run an Linux container on a Windows laptop. Or a Windows container, on a Linux server.

A Docker **image** is a 'cookie-cutter' that includes everything needed to run an container: code, operating system, system tools, system libraries and settings. Therefore if you data science team packages up codebases as Docker images, the containers they describe will run on any machine, making deployment a much smoother process.



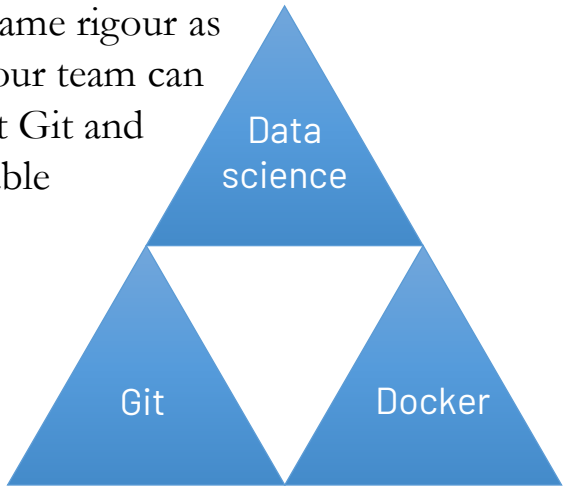
Containers in the cloud

All major cloud providers allow you to run containers directly in the cloud, without worrying about infrastructure:

Amazon AWS
https://aws.amazon.com/containers/
Microsoft Azure
https://azure.microsoft.com/en-gb/product-categories/containers/
Google Cloud Platform
https://cloud.google.com/compute/docs/containers

Modern Data Science Delivery

Data science is now subject to the same rigour as other branches of programming. Your team can deliver data science projects without Git and Docker, but to truly launch sustainable applications into production, it is crucial to adopt these technologies.



Most public AI codebases are now distributed through GitHub and give the option to use Docker to install the codebase.

Check out our Git and Docker training courses specifically tailored for data scientists - **https://adsp.ai/downloads/training**

In summary…

Whilst there are hundreds of data science tools and platforms available, some have become established as essential for any modern data scientist, engineer or analyst.

In this section we have presented the core technologies that are commonly used by data professionals and have explored why Git and Docker are a core part of any modern tech stack.

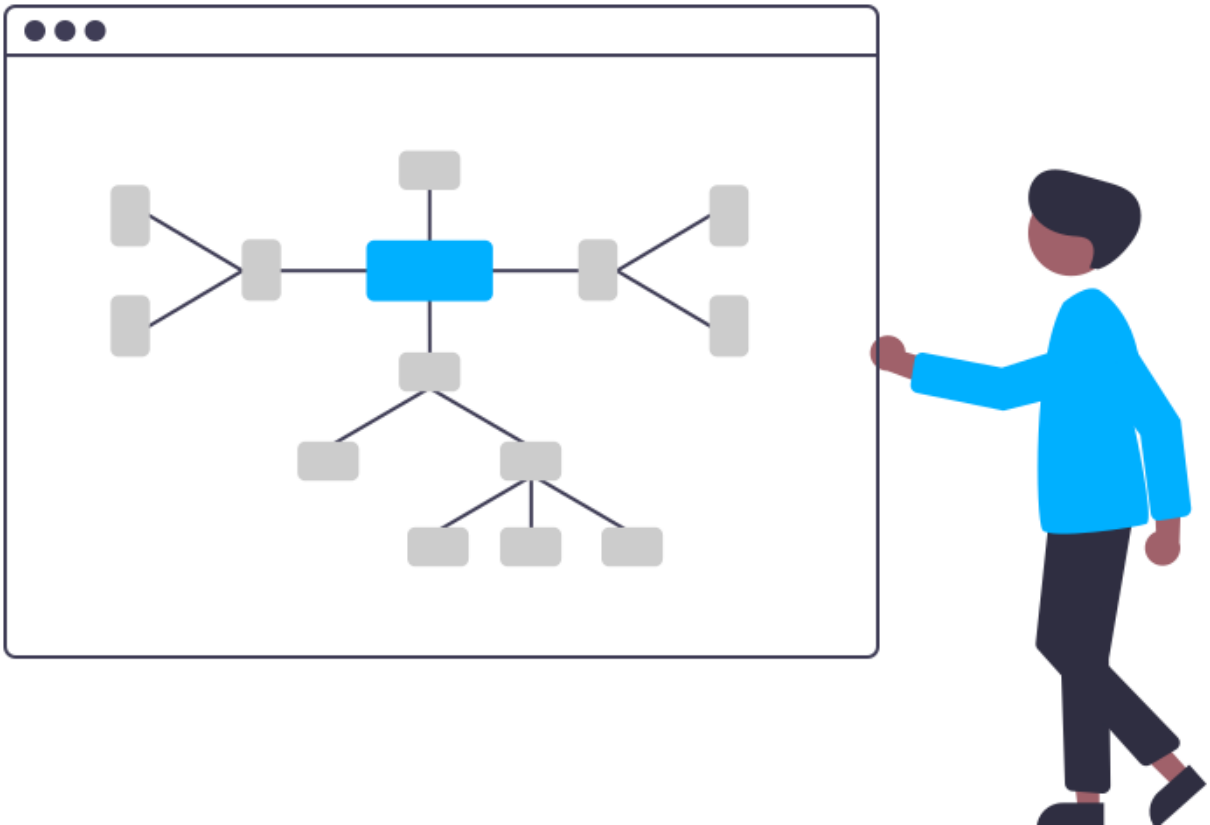
ADSP can help you find the right set of tools for your team and provides training courses on a wide range of technologies. Get in touch with us at hello@adsp.ai if you'd like to hear more.

5 Automation

Much of the promise of data science and AI centres on the ability of machines to automate processes. It is therefore crucial that the process of monitoring and maintaining AI solutions is also automated as much as possible.

For example, machine learning is well known for being able to tackle **complex tasks at scale**, such as identifying broken or defective items on a production line conveyor belt from a live video stream. For such a solution to thrive, processes surrounding the AI system should also be automated – for example, data pipeline validation, monitoring model performance and establishing schedules to retrain the model.

In this section, we'll explore some of the key tools and paradigms that can that can help ensure your AI models are truly automatic.



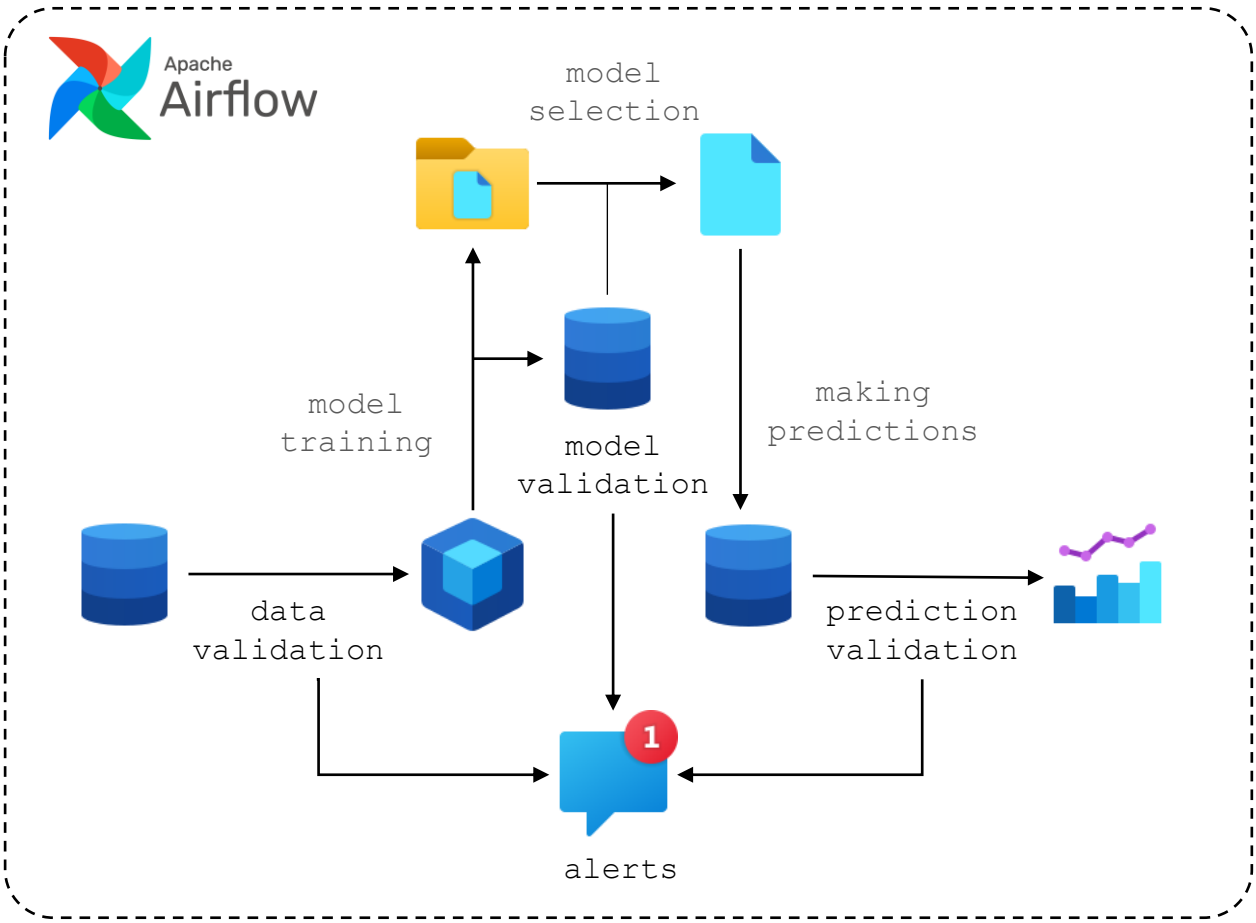
AI model workflows

AI models do not exist in isolation. They should be supported by an automated framework that operates and validates each of the key workflows in the live environment.

There are broadly six key workflows that can be considered for automation – three operational and three based on validation.

Operational	Model training	Model selection	Making predictions
Validation	Data validation	Model validation	Prediction validation

Apache Airflow is an open-source tool that can be used to programmatically author, schedule and monitor these workflows and create alerts if any part of the process fails or deviates from expected norms. We'll first explore Airflow and then zoom-in on each workflow.

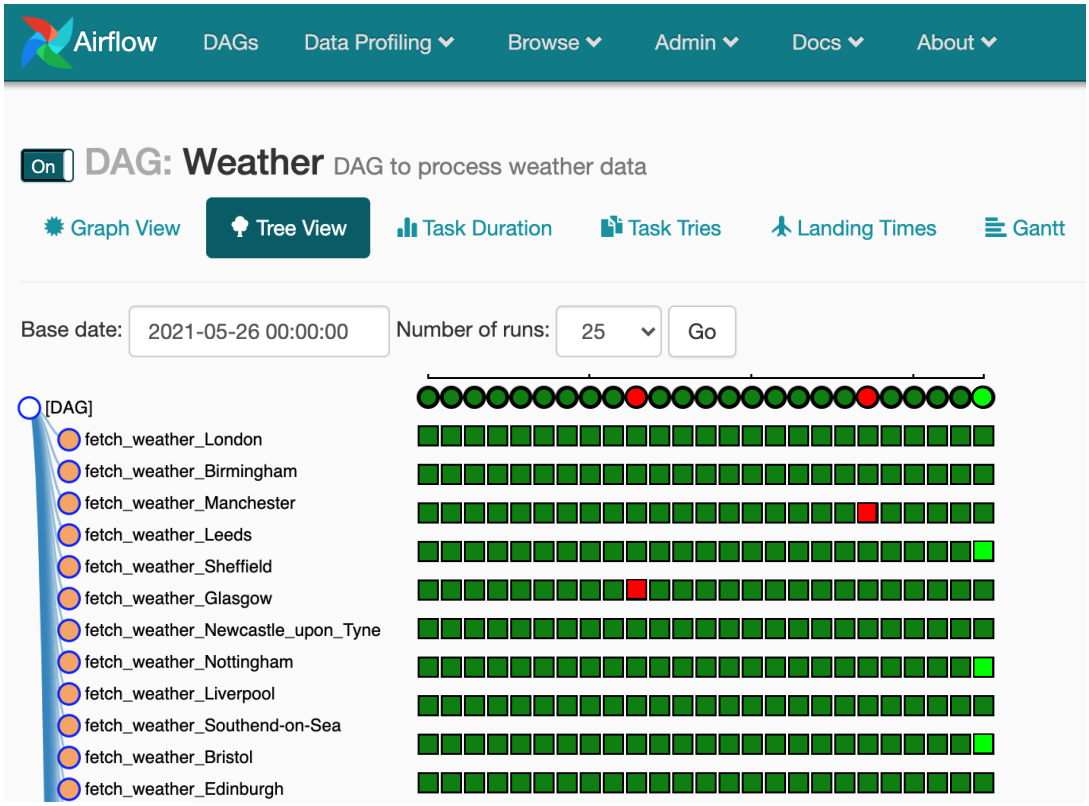


Airflow

Created by Airbnb in 2014, **Apache Airflow** is an open-source workflow management platform. It allows users to write and schedule workflows in Python and monitor processes in an intuitive browser-based user interface.

For example, suppose your AI team wanted to scrape weather data daily from an API, across multiple locations. Airflow can help orchestrate the workflow, by allowing users to specify how many processes should run in parallel, any downstream tasks that should be run after each location is captured and the time or trigger that kicks off the workflow every day.

The Airflow interface shows which tasks have completed successfully (dark green), failed (red) and are in progress (green). Your team can inspect log files and rerun failed tasks directly from the interface.



APPLIED DATA SCIENCE

Data validation

The first step of any AI retraining schedule is to check that the data being used to train the model passes a set of validation criteria. There are three broad validation themes:

Missingness

Check that the row count of the dataset is in line with expectations and that all missing field values are explainable. Missing field values or rows can indicate a problem in the data creation pipeline, such as incorrect joins or out-of-date lookup tables.

Distribution

Check the distribution of each feature – for continuous variables, are there outliers that require special treatment; for discrete variables, can sparse categories be grouped? Does the dataset sufficiently represent the full spectrum of response values?

Drift

Check if distributions and missingness metrics are consistent over time – if the training data contains historical observations that are no longer representative of the ground truth, the model may perform poorly in the live environment.

Model training

Most machine learning models are typically not continually retrained in the live environment, but instead are deployed as part of a process triggered by a data scientist or engineer. There are two key reasons for this:

- Machine learning models should be version controlled in the same way as other software, to ensure stability and so that the performance of each version can be accurately monitored.
- Usually, the signal in the data does not change rapidly enough to justify constant retraining.

However, the pipeline for retraining models can be automated, using Airflow to manage the workflow (e.g. creating the train / validation data split, saving out trained models and accuracy metrics etc.)

Model validation and selection

To establish whether a model is better than a previous version, it is tested against a set of data that wasn't used during training - this is called **model validation**.

EXAMPLE

Suppose we withhold a validation dataset of 1000 bank transactions, 50 of which are fraudulent. Which model below would you say is 'better' at detecting fraud in the validation dataset?

Model A		Predicted	
		Not Fraud	Fraud
Actual	Not Fraud	930	20
	Fraud	30	20

Model B		Predicted	
		Not Fraud	Fraud
Actual	Not Fraud	850	100
	Fraud	10	40

The answer is that it depends on whether you want to prioritise **precision** or **recall**. Model A is more **precise** – 50% (20/40) of its positive fraud predictions are correct, whereas only 29% (40/140) are correct for Model B. However, Model B can **recall** more examples of fraud – 40/50 (80%), in comparison to only 20/50 (40%) for Model B. Therefore Model A might be a better choice if false positives are costly for your business, whereas Model B might be preferred if you want to catch most fraud, at the expense of some mislabelled mistakes.

Precision and recall can also be changed by adjusting the **threshold** at which the model predicts positive values (i.e. lowering the threshold means more positive predictions are made).

The process of model validation and selection revolves around balancing **model performance metrics** with **model complexity** – a simpler, more interpretable model is sometimes better, even at the expense of accuracy. Whilst lots of the model validation steps can be automated, a data scientist is often involved in the final model selection decision.

Making predictions

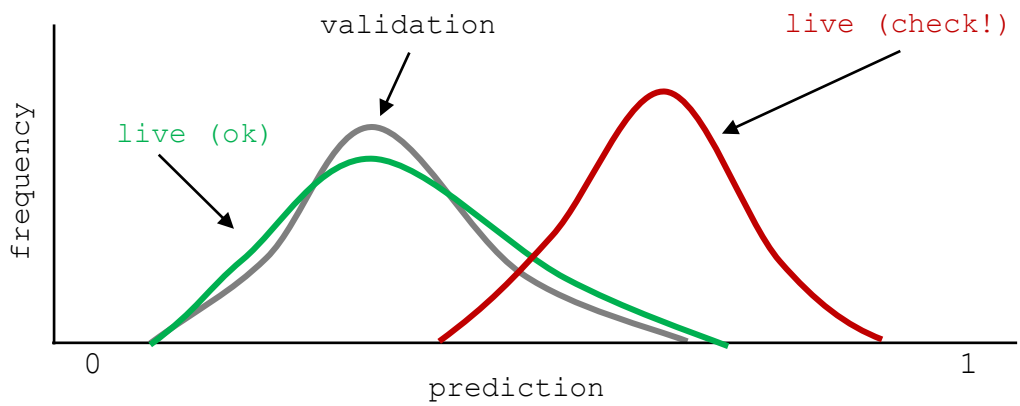
The process of using the trained machine learning model to predict new observations can either be conducted in **batch** or in **real-time.**

Batch processing is more computationally efficient and involves sending many examples to the model at once for prediction. For example, customer segments could be reallocated overnight based on usage statistics from the previous 30 days. This can be managed using Airflow, with predictions being stored back in the database for use downstream as part of other processes, tools and visualisations.

Real-time processing involves placing the model behind an API, so that it can be called with a request containing data on the observation to be predicted. For example, interactions from a web browsing session could be sent to the API to predict the offer most likely to convince the customer to purchase in real-time.

Prediction validation

The distribution of predictions in the live test set should approximately match the distribution of predictions in the validation set, as shown below. If this isn't the case, then the data used to train the model isn't representative of the new data being predicted in the live environment and the predictions may therefore not be accurate.



Alerts

Scripts deployed through Airflow can be used to alert data scientists and engineers to workflows that need attention. For example, if the latest predictions do not match the distribution of past predictions, then rather than overwriting the latest values, the process can terminate early and send a process report to the engineer in charge of the workflow for analysis.

This kind of process control is common across all forms of software development – machine learning and AI are not exceptions. Operational dashboards are an excellent way to monitor live machine learning models and we strongly recommend building a simple dashboard at the start of any ML project to be used as a window into the running of the model in the live environment. This can be used as part of the alert system, or simply to check on key metrics.

In summary…

Continuous improvement of existing machine learning models and development of new models is only possible if the right frameworks are put in place to facilitate the scheduling, maintenance and orchestration of AI workflows.

In this section we have seen how Airflow can deliver this aim and have presented the six key areas where automation can take place, supported by oversight from data experts.

ADSP are experts at designing, building and deploying automated machine learning solutions. Get in touch with us at hello@adsp.ai if you'd like to hear more.

6 Deployment

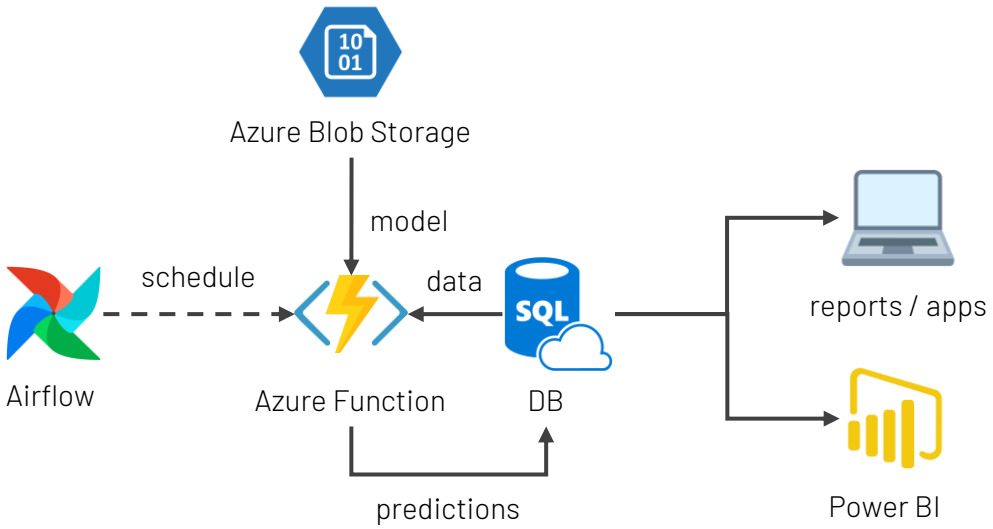
Data science and AI solutions start delivering real value once they are deployed to production.

Production solutions need to be **stable, scalable and sustainable.** In this section we'll reveal three common architecture designs for deploying data science solutions to a live environment that meet these crucial criteria.

For each architecture design, we'll provide practical steps for delivery within Microsoft Azure (the same patterns are of course equally applicable within AWS or GCP).

Batch processing

A **batch process** runs at regular intervals, (usually overnight) so that the output has been refreshed by the following morning.



The key components of this architecture are as follows:

Airflow

Airflow can be used to build and execute processes according to a given schedule. It can be configured using **Python** and has more functionality than simpler scheduling tools, such as cron on Linux or Task Scheduler on Windows.

Azure Blob Storage

Azure Blob Storage can be used to store a trained model as a file – this is then loaded at the start of the batch prediction process.

Azure Function

The code that runs the pre-processing and prediction logic can be run through a serverless **Azure Function** – alternatively, a server can be used, though would require infrastructure provisioning and maintenance.

Database

For batch processing, a database is required to store the output.

Power BI

Dashboards (e.g. Power BI), downstream applications and reports can then be connected directly to the database, to read up-to-date predictions and other output from the previous run.

Top tips

Here are our three top tips for ensuring batch processes run smoothly:

1. Rather than overwriting previous predictions, store new predictions as **new rows** with a timestamp, so you can perform meta-analysis on how predictions have changed over time.

2. Batch predictions into **small chunks**, to avoid long running processes and write detailed **logging** to help with debugging

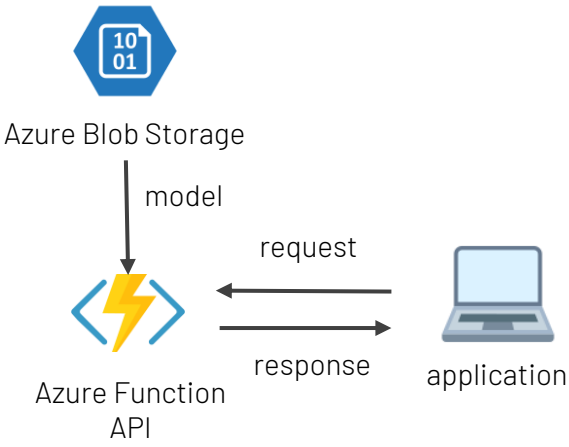
3. Make batch processes **idempotent** and pick up from where they left off in the case of failure – if only the last prediction failed, you don't want to have to predict the whole dataset again!

Example: Customer Segmentation

Customer segmentation is an excellent example of a business process well suited to batch processing. Overnight, new data from the previous day's activity by each user can be processed to reassign the segment of each customer. Users who have moved segment can be piped through to a dashboard or email marketing tools to be targeted with specific messaging relevant to the type of segmentation movement.

API

Putting a model behind an **API**, allows you to call it with specific data – this is useful if you want to build the model into processes that require a response based on live information.



The key components of this architecture are as follows:

Azure Blob Storage

Azure Blob Storage can be used to store the trained model as a file. This is loaded each time the API is called.

Azure Function (API)

The code that runs the pre-processing and prediction logic can be run through a serverless **Azure Function.** You can define the **trigger** for the function to be an HTTP request (i.e. an API call) and also define the routing of your API.

Application

The application sends a **request** containing data to the API (e.g. information relating to the customer session on a website) and receives back a **response** (e.g. the predicted best offer to present to the user).

Top tips

Here are our three top tips for deploying models behind APIs:

1. Ensure your API is **well documented** – you can use OpenAPI (formerly Swagger) to automatically generate documentation from the codebase

2. You can ensure **scalability** by defining the maximum number of instances of the Azure Function to run concurrently – this is one key advantage of a serverless implementation.

3. Make use of frameworks such as **Django**, **Flask** or **FastAPI** to handle more complicated API with routing and authentication

Example: Car price prediction

A good example of a machine learning model that is well suited to being deployed behind an API is a **used car price prediction model**.

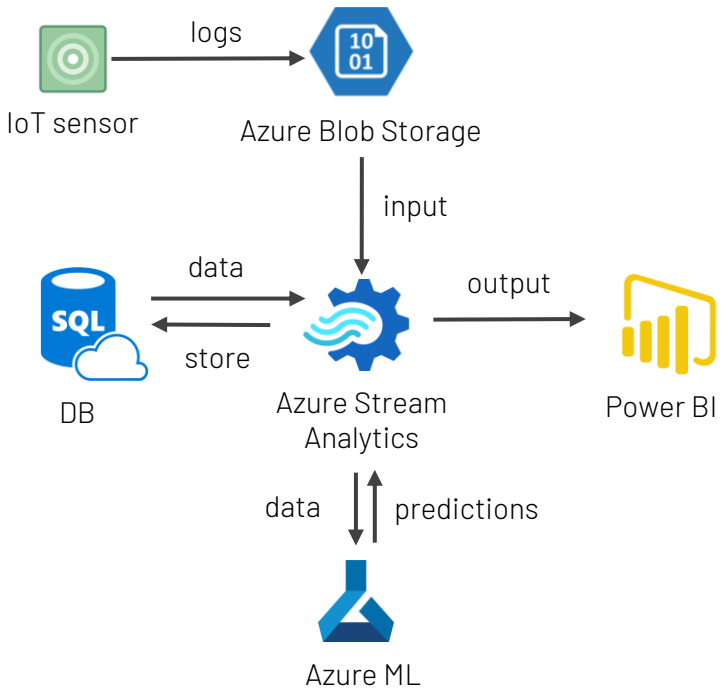
Suppose you run a business that trades used cars – you might want to place a tool on your website that allows users to enter details of the car they are looking to sell through an online form and receive back an instant quote.

To achieve this, your data scientists could train a machine learning to predict the price of a used car, given attributes of the car such as the model, engine size and colour. You would then deploy this behind an **API** (e.g. an Azure Function) which can then be called with data for a single vehicle to generate a prediction.

When a user enters details of their car on the website, the API is hit with this data to return a prediction back to the browser.

Streaming

Some devices constantly output data that requires visualisation, processing and prediction. There are many ways of handling this data, depending on the exact use case - below we show a suitable solution for applying a machine learning model to predict anomalies from data logged by an IoT sensor



The key components of this architecture are as follows:

Azure Stream Analytics

Azure Stream Analytics processes real time data from a queue or data store – it can be connected to **Azure ML** to generate live predictions against the data. **Power BI** can be directly connected to Stream Analytics, for real-time visualisation.

Azure ML

Machine learning models can be built and deployed through **Azure ML** which can then integrate directly with Stream Analytics.

APPLIED DATA SCIENCE

Example: Predicting failures in water pipes

Acoustic sensors attached to water pipes generate a stream of time series data that can be used to detect leaks before they irreparably damage the pipe.

A water company that wanted to take advantage of this technology could deploy an anomaly detection model to identify pipes that are imminently about to break, given historic acoustic sensor data and prior work order data from previous breakages.

The data and predictions could be surfaced through a Power BI dashboard, for the leak management team. Engineering teams could then be deployed to at-risk areas, to pre-emptively prevent major damage to infrastructure before it occurs.

In summary…

Machine learning solutions generate value if they are deployed in a way that is **stable , scalable**, **sustainable**.

In this section we have presented three ways in which machine learning solutions can be deployed – through **batch processing**, **APIs** or **streaming**. The correct method to choose will depend on your specific use case.

ADSP can help your company build and deploy machine learning and AI models, using best-practice architectures patterns and technologies. Get in touch with us at hello@adsp.ai if you'd like to hear more.

7 Proving Value

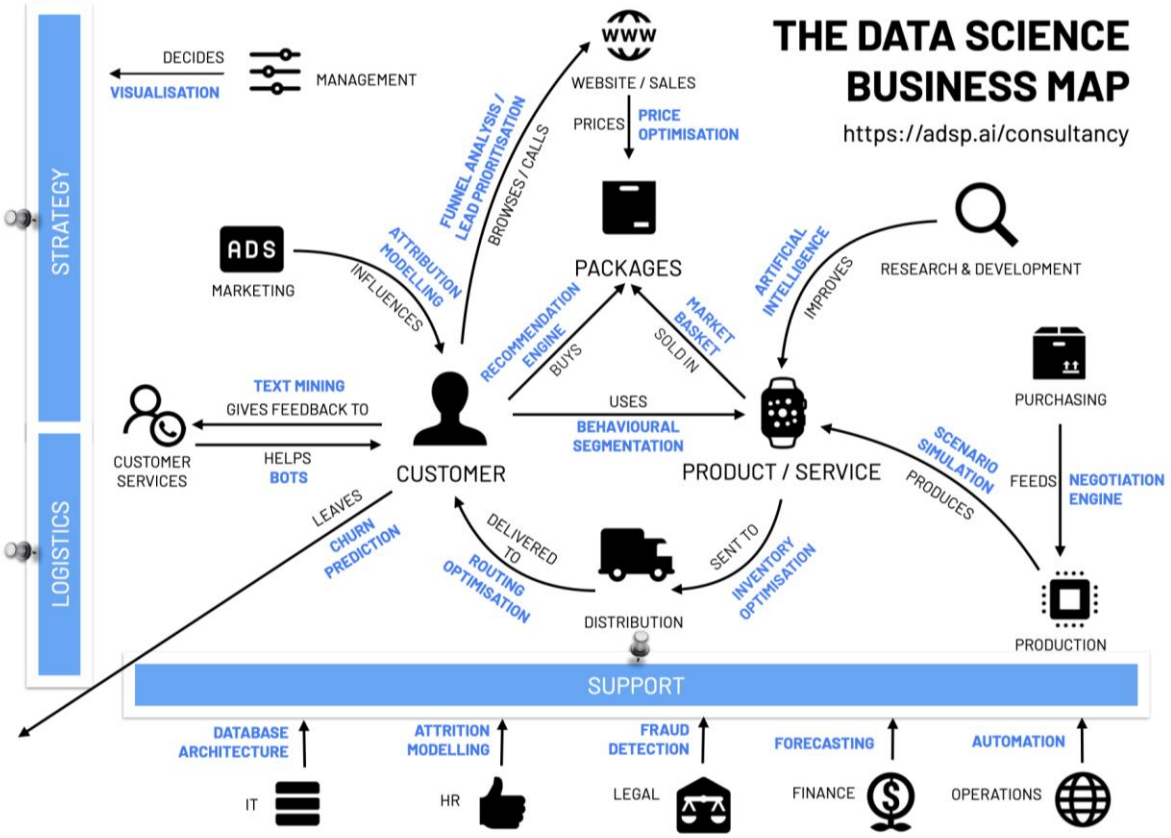
Any business undertaking should be justified by the value it brings to the business. Data science projects are no different.

Calculating return on investment is not an easy task and requires a structured approach to measuring value, that doesn't track 'vanity' metrics but assesses the true impact on business objectives.

In this section we'll explore the four different ways in which data science projects can create tangible value and how to measure each type of value generation.

Where can data science create value?

The diagram below depicts the connections between key business entities (the customer, the product, the distribution network etc.) and the typical data science projects that improve the efficiency or quality of each connection.



For example, **behavioural segmentation** projects uncovers the different ways in customers use your product or service. **Inventory / routing optimisation** projects seeks to improve the physical movement of product to the customer through the distribution network. Creating a business map such as this for your own organisation can help to visualise how individual data science projects fit into the wider business operation.

Crucially, the impact of each project should be measured according to the business objectives of the underlying connection, rather than with so-called 'vanity' metrics that do not accurately describe the true impact of the project. We'll explore what this means in detail in the following section.

How can data science create value?

There are four primary ways that a data science project can directly create tangible value for an organisation:

Customer Acquisition

Projects such as **funnel analysis**, **attribution modelling** and **lead prioritisation** can help you get more customers through the door, thereby increasing revenue.

Lifetime value

Projects such as **behavioural segmentation**, **churn prediction** and **recommendation engines** can help to increase the average amount a customer spends overall (for example, per transaction or per time period).

Cost saving

Projects such as **inventory optimisation**, **logistics optimisation** and **scenario simulations** aim to reduce costs – for example, by improving throughput from the warehouse to the distribution network.

Time saving

Projects such as **text mining and data visualisation** seek to reduce the time spent performing routine tasks, such as handling customer requests and collating key metrics and performance statistics from across the business.

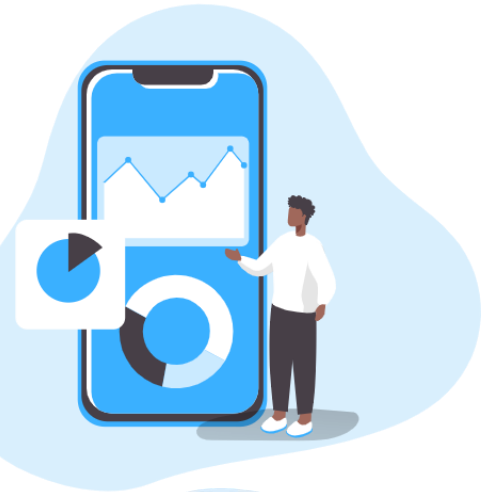
The precise meaning of each category will depend on the nature of the organisation. In general, **customer acquisition** and **lifetime value** focus on increasing revenue, whereas **cost saving** and **time saving** focus on reducing costs, either directly or indirectly through improved efficiency.

For each category of value creation, there are specific ways in which the value can be measured. In the following sections we explore each focus area in detail.

APPLIED DATA SCIENCE

Customer Acquisition

The following three examples of **customer acquisition** projects showcase how to set up experiments to directly measure their impact.



Funnel analysis

Learn more

Funnel analysis involves mining website clickstream data to identify areas of the site with high bounce rate or low engagement.

Experiment - A/B test a change to the site, based on findings from the funnel analysis.

KPIs - Conversion rate.

Attribution modelling

Learn more

Attribution modelling calculates the amount of credit to assign to each marketing touchpoint.

Experiment – Track marketing impact by touchpoint over time, using calculations from the multi-touch attribution modelling.

KPIs - Conversion rate and actionability of results.

Lead prioritisation

Learn more

Machine learning can be used to predict which leads are most likely to convert and produce a list of prioritised opportunities for the sales team.

Experiment – Action a mixture of prioritised and non-prioritised leads and log the results.

KPIs – Measure the different in conversion between prioritised and non-prioritised leads.

APPLIED DATA SCIENCE

Lifetime Value

The following three examples of **lifetime value** projects showcase how to set up experiments to directly measure their impact.

Behavioural segmentation

Learn more

Clustering customers by behaviour rather than broad demographics allows you to deliver a personalised experience that resonates.

Experiment – Personalise content based on segment vs control group with no personalisation.

KPIs – Uplift in engagement / purchases vs control.

Churn prediction

Learn more

Predict which customers are most likely to stop using your product or service using machine learning and intervene before it's too late.

Experiment – Target high potential churn customers with offers to encourage continuation.

KPIs – Lifetime value vs control group.

Recommendation engine

Learn more

Cross-sell products by mining your data for bundles and packages that work well together.

Experiment – Surface recommended products through the website, based on prior purchases.

KPIs – Uplift in engagement / purchases vs control group with random recommendations.

APPLIED DATA SCIENCE

Cost saving

The following three examples of **cost saving** projects showcase how to set up experiments to directly measure their impact.

Inventory optimisation

Learn more

Forecasting demand in order to maintain optimal stock levels reduces warehouse and insurance costs and improves picking efficiency.

Experiment – Test the new inventory strategy for a selected group of product lines.

KPIs – Reduced volatility/cost in stocking position.

Logistics optimisation

Learn more

Routing algorithms can determine the optimal placement of shipment hubs and allocate individual dispatches in the most efficient manner.

Experiment – Backtest the optimised strategy on historic data to prove the concept.

KPIs – Reduction in transportation costs.

Scenario simulation

Learn more

By building a digital twin of a business operation, you can accurately simulate potential future scenarios, to optimise efficiency and strategy.

Experiment – Backtest how closely the simulation matches reality in a variety of circumstances.

KPIs – Reduction in opportunity costs.

APPLIED DATA SCIENCE

Time saving

The following examples of a **time saving** project showcases how to set up an experiment to directly measure its impact.

Text mining

Learn more

Machine learning can automatically classify and respond to incoming emails, and track sentiment against each topic.

Experiment – assign a proportion of inbound email to be classified by the ML solution.

KPIs – Classification time / accuracy vs control

Dashboards

Learn more

An engaging, functionality suite of interactive dashboards can significantly reduce the time spent creating ad-hoc reports.

Experiment – For the most common requests, build a suite of dashboard to allow users to self-serve.

KPIs – Faster time from question to answer.

In summary…

In this section, we've summarised the four ways data science projects can create value and provided examples of specific experiments you can run for each type of project.

Get in touch with us at hello@adsp.ai if you'd like to hear more about how we deliver tangible value for our clients.

8 Big Ideas

In this final chapter, we look to the future to explore the next big ideas from data science and AI that will impact businesses and create competitive advantage in the years to come.

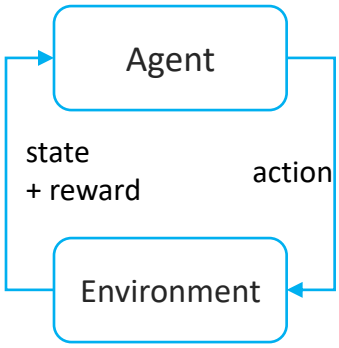
Given the rapid rate of progress in the fields of data science and AI, it is clear that businesses need to start preparing now for the next big idea, to get ahead of the curve before it's too late.

Over the following pages, we present five interesting concepts from cutting-edge AI research with potential business applications.

Reinforcement Learning

Reinforcement Learning (RL) is often described as the third pillar of AI, alongside supervised learning (e.g. prediction) and unsupervised learning (e.g. clustering). It is the technique used to teach agents how to act to maximise rewards in a given environment.

For example, DeepMind have successfully applied RL to learn near-optimal strategies for complex games such as chess and Go. The agent is able to learn 'tabula-rasa' (with no prior knowledge), by playing against itself millions of times inside a simulated environment, to gradually develop more intricate and sophisticated strategies.



The RL interaction loop

The application of RL to business remains novel, largely because an accurate simulation of the environment is required in order to provide the agent with a 'playground' in which to learn. However, with the rise of 'digital-twin' technologies, businesses now have the potential to train RL agents to act optimally inside the digital-twin simulation.

Example applications

Purchasing and selling
Given a digital twin of a marketplace with dynamic pricing (e.g. an ad platform or spot market for goods), train an agent using RL to predict the optimal times and quantities to buy and sell, in order to maximise profit.

Operational design
Managing the flow of patients through a hospital, by training an agent to adjust parameters relating to bed numbers and staffing to minimise waiting times.

Creative AI

Traditionally, the field of machine learning has focused on predicting a given label, using a training set of labelled data. For example, predict if a given photo of a face is smiling or frowning. **Creative AI** or 'generative modelling', is the idea that models can be trained to output completely new images that are indistinguishable from those in the training set. For example – is the image below created by a machine, or a real human? (answer on the next page!)

Real or fake?

The field of creative AI has been driven largely by the invention of generative adversarial networks (GANs) and other types of generative model. GANs consist of two competing models that are trained in parallel - the generator (that creates new images), and the discriminator (that tries to guess if the image is real or fake).

Example application

Product Design

A model called DALL-E, developed by OpenAI, can output a variety of designs for a new product, given nothing more that a text-based prompt (see example below).

This is opening new avenues for designers, working with AI solutions such as DALL-E to develop novel and interesting product concepts. Creative AI is also being explored by game designers and film studios to complement and enhance human-based design.

Prompt: an armchair in the shape of an avocado

source: https://openai.com/blog/dall-e/

APPLIED DATA SCIENCE

Language Modelling

Language modelling is a branch of creative AI, that is based on producing realistic text-based output. The model at the cutting-edge of language modelling is called GPT-3, developed by OpenAI.

GPT-3 is able to form a realistic continuation of any text based prompt. Crucially, the model does not need to be retrained for each application – you simply tell it what you want it to do!

For example, this example shows GPT-3 creating ad-copy, given a short description of the product. GPT-3 is now available as an API, meaning that businesses can easily built it into their own applications and apply it to specific use-cases, such as product name generation.

Prompt

Write a creative ad for the following product to run on Facebook:
"""
Airee is a line of skin-care products for young women with delicate skin. The ingredients are all-natural.
"""
This is the ad I wrote for Facebook aimed at teenage girls:
"""

Sample response

Do you have sensitive skin? If so, then Airee is for you! Airee products are made with natural ingredients to help you feel good about your skin. Try out Airee for a happy, healthy you!

source: https://beta.openai.com/examples/default-ad-product-description

Example application

Language models are also opening exciting new possibilities for text summarisation and knowledge extraction – for example, summarising lengthy legal or technical documents into a short abstract or summary, for human consumption, as demonstrated in the adjacent example.

Prompt

My second grader asked me what this passage means:
"""
Jupiter is the fifth planet from the Sun and the largest in the Solar System. It is a gas giant with a mass one-thousandth that of the Sun, but two-and-a-half times that of all the other planets in the Solar System combined. Jupiter is one of the brightest objects visible to the naked eye in the night sky, and has been known to ancient civilizations since before recorded history. It is named after the Roman god Jupiter.[19] When viewed from Earth, Jupiter can be bright enough for its reflected light to cast visible shadows,[20] and is on average the third-brightest natural object in the night sky after the Moon and Venus.
"""
I rephrased it for him, in plain language a second grader can understand:
"""

Sample response

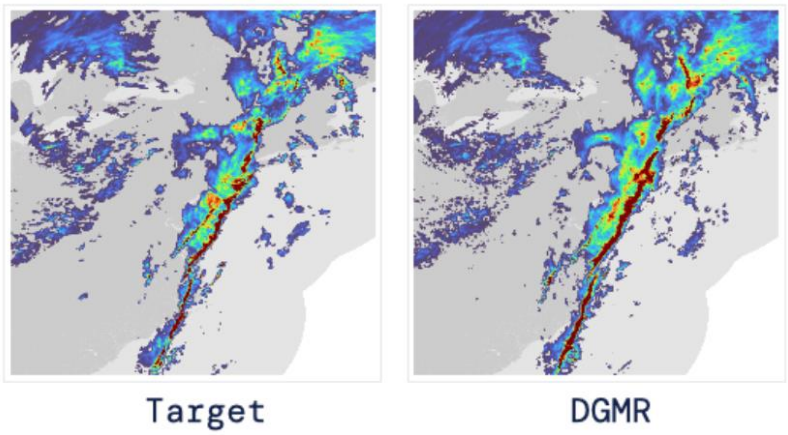
Jupiter is a huge planet made of gas. It is the fifth planet from the sun. It is very big and bright. It is the third brightest object in the night sky after the moon and Venus. Jupiter has many moons. One of the moons is bigger than the planet Mercury.

Answer to face question: It's fake – generated from https://thispersondoesnotexist.com/

Nowcasting

Most businesses use some form of forecasting to predict trends at a daily, monthly or yearly level to inform strategy. In contrast, **nowcasting** focuses on forecasting the very near future, as data arrives from multiple sources in real time.

The idea comes from weather forecasting and econometrics, which need to react immediately to short term changes in situation. DeepMind have recently released an example of nowcasting 1-2 hours of precipitation, DGMR, beating previous benchmarks.



DeepMind nowcasting model (DGMR) for precipitation
https://deepmind.com/blog/article/nowcasting

Example application

Social Media Nowcasting

Nowcasting can be used to take advantage of short term social media trends. For example, if nowcasting models can predict which hashtags will trend in the next hour, content writers can get ahead of the curve and promote their content ahead of competitors. A key challenge of nowcasting lies in data engineering – ensuring models are able to be constantly fed with a stream of data from a variety of unstructured sources, to predict quickly and accurately.

Human-AI interaction

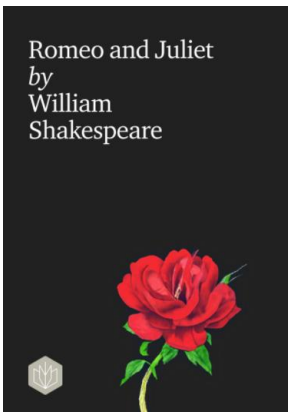
One of the most exciting ways that the field of AI is developing, is in the interface between AI and humans.

Traditionally, AI has been used by businesses to tackle specific 'narrow' problems, such as demand forecasting, recommending products to customers or detecting anomalies. The output from these problems is often a number, that can be validated against known data and surfaced in other applications or dashboards.

A big idea for the future is that AI can tackle more general problems and interface with humans in a more assistive and strategic way. DeepMind have recently published an example of how an AI model can be trained to discover novel patterns within a particular field of mathematics and present back suggestions for a human to validate. In a similar way, it is not unreasonable to imagine AI models that can in the future find abstract patterns in business related fields across diverse datasets and present these insights back as human interpretable conjectures for discussion.

Example application

Recursive Task Decomposition
OpenAI has recently shown how the complex task of summarising a whole book or play can be divided into smaller subtasks, assisted by humans. The concept of AI assisting businesses at a strategic level, by synthesising output from smaller, more tactical tasks is an interesting area to keep in mind for businesses wanting to stay at the cutting edge of AI research.



source: https://openai.com/blog/summarizing-books/

Conclusion

Thank you for taking the time to read this Executive Guide To Data Science and AI! We hope you have enjoyed it and have found it informative and useful.

In Chapter 1, we started by discussing the typical roles in a data science team. Chapter 2 focused on specific trends in machine learning and how it is being applied within businesses. Chapter 3 on the different types of data visualisation and how to establish the best platform for your company.

Chapter 4 provided a summary of the most important tools and technologies currently used by data science teams and in Chapter 5, we explored current best-practices regarding the automation of data processes.

In Chapter 6, we focused on deployment and how to practically move solutions to the cloud. Chapter 7 explored the different ways that data science projects can create value and Chapter 8 concluded the guide with a summary of the 'big ideas' from AI research that business can already start to explore.

We hope this guide has been a useful summary of the current state of AI and data science with a nod to the future – do let us know if you have any feedback or would like any further info!

ADSP is a team of data scientists and engineers who develop world class data-driven solutions. Get in touch with us at hello@adsp.ai to hear how we can help your business to prosper and take advantage of the most recent developments in data science and AI.

hello@adsp.ai

APPLIED DATA SCIENCE