NPA Data Science

1: Data in Society and the Principles of Data Science

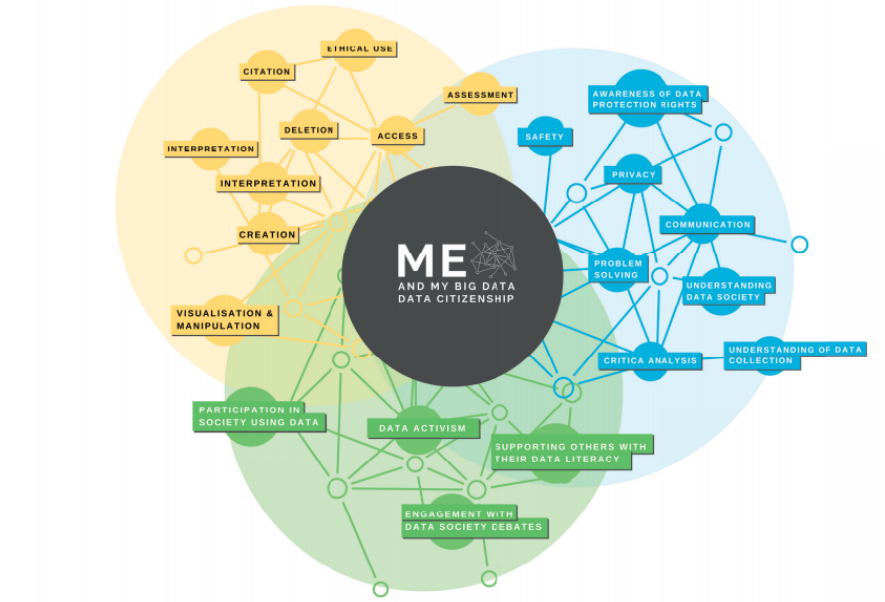
Topic Outcomes

- Explain the use of data in society
- Explain the principles of data science

Data in Society and Data Citizenship

Outcome 1 of the data citizenship component of the NPA is to explain the use of data in society. You should have a detailed understanding of data and how data impacts our lives every day. The concept of data or digital citizenship is relatively recent. There have been a number of attempts to define this term. Yates et al. (2020) have developed a framework of data citizenship shown in the graphic below:

The Data Citizenship Model



The framework covers three main areas: **data thinking, data doing and data participation.**

Data Thinking	Data Doing	Data Participation
Data Thinking incorporates critical skills as they view and analyse the world through data. The process of data decoding [11] requires critical data literacy abilities such as understanding the online ecosystem, solving problems with data, communicate using data, development and evaluation data-based explanations.	Data Doing incorporates practical skills involving data handling and data management. Data Doing advocates that, for example, social media users should be provided with the abilities to identify and highlight the source of the information they share with others.	Data Participation examines the collective and interconnected nature of data society. Through Data Participation, citizens seek opportunities to exercise their rights as well as to contribute to and shape their collective data experiences. Examples of Data Participation might include a person who actively contributes to online forums, uses open data for the benefits of their community, helps others to set up a secure password, engages in privacy or misinformation debates or takes steps to protect their personal information.

Data in society

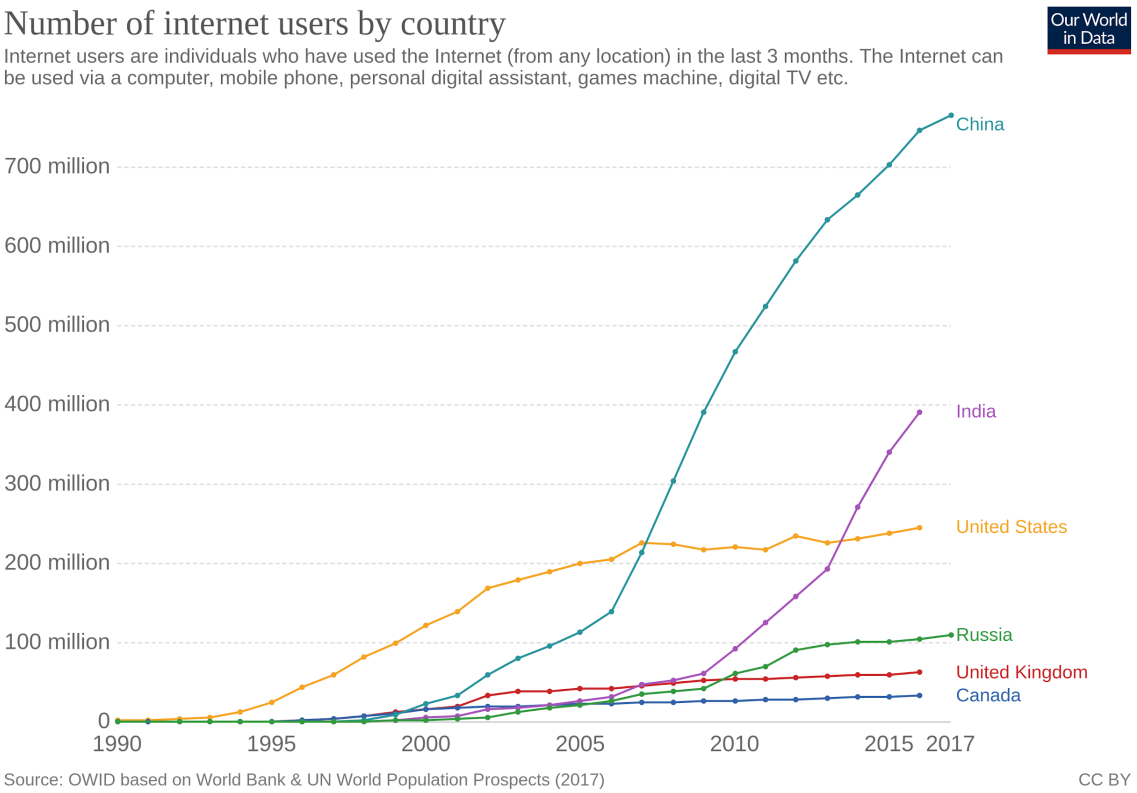
In the 1990s and 2000s internet usage gradually increased, however in the 2020s it would be hard to imagine life without the internet, every day millions of emails, social media posts, instant messages, digital financial transactions and many other digital messages are sent across the internet.

Each time we interact with a communication device connected to the internet, data is being generated and this data creates a footprint, or an image of us as individuals in society. This data is slowly becoming a more and more important part of our digital identity, our data citizenship. Within a modern digital society, how do we define data citizenship?

It is important to distinguish data citizenship from digital or computer literacy, if an individual or group of individuals have different levels of ability in usage of computer systems this does not make them any less a digital citizen than someone who is highly engaged with the latest technology.

Data cross-nationally

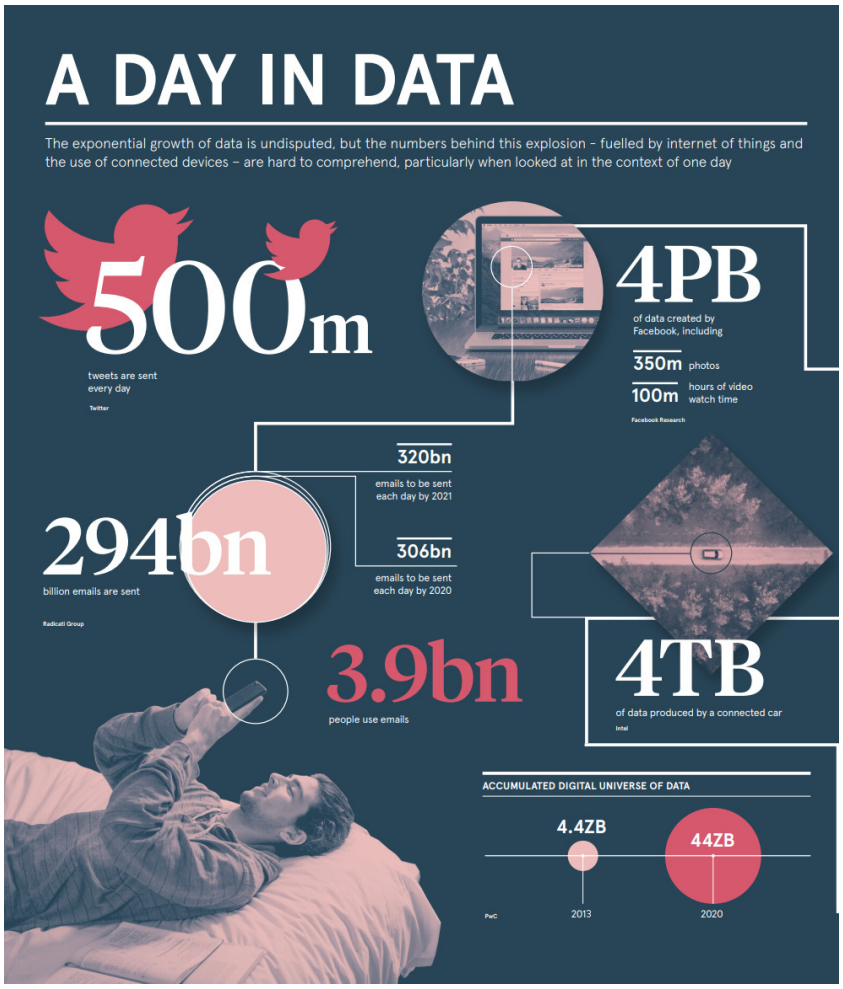
In Scotland and in the wider United Kingdom our impact as internet users is very small compared to other countries. Consider the following line chart which shows the number of internet users by country. China has by far the most data citizens in the world, followed by India and then the USA. If we consider Scotland as around 8% of the UK in terms of population, then you can see how small we are in comparison to some of the larger countries in the world.



How much data do we create every day?

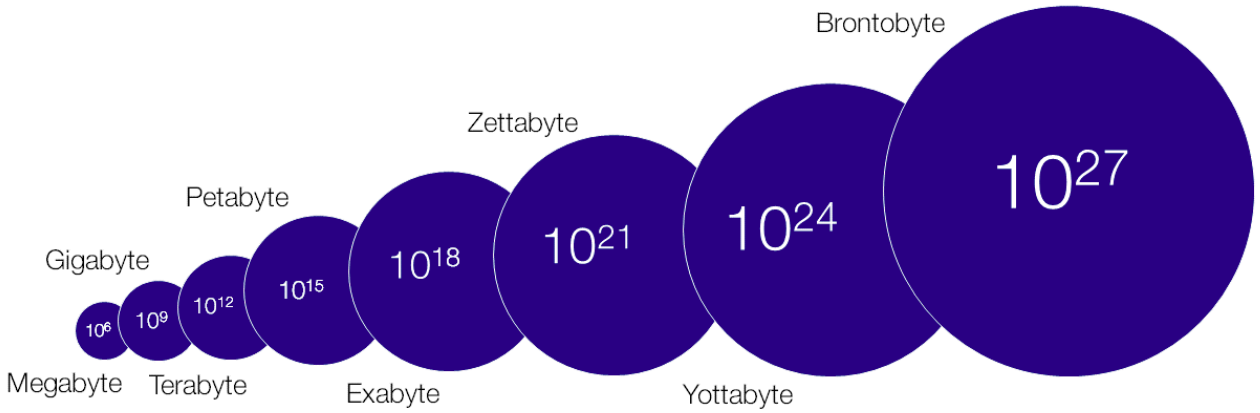
Each day we walk past CCTV cameras. Most of us use a smart phone of some kind. Many smartphones track our GPS location, even when we do not want them to. We may send messages to friends and these are recorded in massive databases by large American corporations. These corporations perform analysis on the trends in our activities and use these to target advertising. Have you ever searched for something online and later noticed an advert for something very similar?

A shopping centre is an example of a place where lots of data is collected. CCTV monitors customer activity and wireless networks register movement throughout the site. If you go into a coffee shop and buy a coffee your presence is noted when you pay with a bank card, unless you use cash - either way you are on CCTV. The coffee shop analyses the time and type of drink that you bought - this allows them to build trends on what is popular and when. Perhaps you may go to see a film at the cinema or make a purchase from a clothes shop, the same data gathering and analysis takes place.



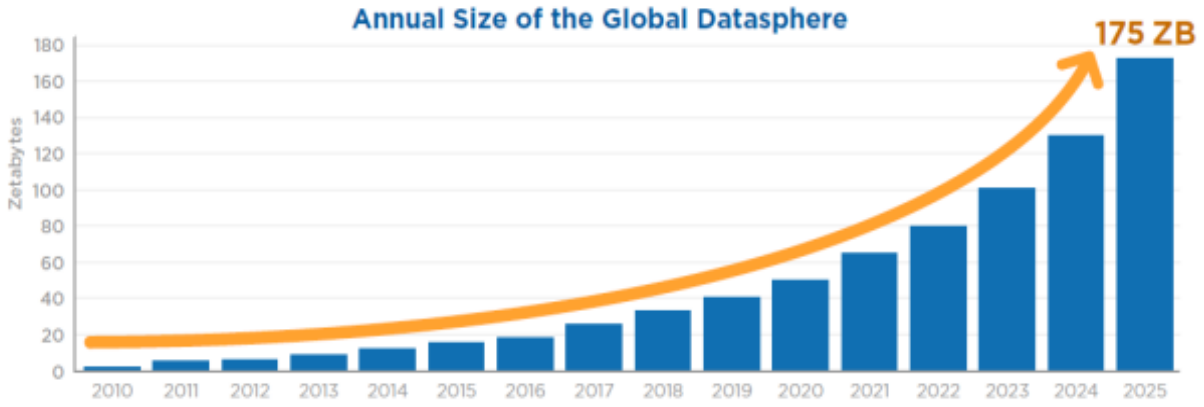
Consider the [Day in Data Infographic,]() each day this data is generated and stored. Much of this data is transparent to the everyday users however, the vast number of smart devices connected to the internet generating data is known as the internet of things, each day these smart devices are generating millions of bytes of data.

The digital and connected nature of modern day life has resulted in vast amounts of data being generated by people and organisations alike. The amount of global data is currently measured in Zettabytes, if you consider the storage capacity of a USB storage device of a smartphone could be 32-128Gb compare that to a quantity of Peta, Zetta or Brontobytes.



The quantity of data generated every day is stored in large databases and is growing every day. Consider the UK 100 years ago, how was data stored? Financial accounts were hand written onto ledgers and information about history was written down and stored in archives or libraries. Information has never been so easily accessible.

This phenomena of an unprecedented growth of information and our ability to collect, process, protect, and exploit it has been described with the catchall term of Big Data. Big Data is a field that treats ways to analyse, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.This trend of growing amounts of data is predicted to continue into the future, as shown below:



There are some important technological, economic and societal reasons for the growth of data.

Technological reasons behind the growth include advances in mobile devices, digital sensors, communications, computing and storage which provide means for collecting data. For example, there has been an exponential increase of computer memory - exponentially increasing storage capacity and decreasing storage costs.

IoT data are generated by GPS devices, smart cars, PDSs, alarms, lighting and heating fixtures.

One economic reason for the growth of data, is that the value in data has increased. More data means more possible uses. Data can be considered the fuel of the digital economy, with 22% of the world economy is attributed to digital skills, capital and goods and services (Knickrehm et al. 2016).

One societal reason is the use of internet-based communication technologies such as social media sites Facebook, Twitter, LinkedIn, Youtube, Instagram etc. and emailing. For example, an email is sent every 3.5x 10-7 s; 100 terabytes of data are uploaded to Facebook daily. The generation of multimedia data occurs through text, images, audio, video, and these are not just generated for personal use but through the digitisation of public services and commercial products on a global scale.

A variety of real examples can be used to illustrate why data is used, but also the limitations and impact it can have on decision making in business and society.

How data is used and misused by individuals, organisations and society.

Individuals

Self-quantification data is generated by individuals through wristbands for monitoring movement/exercise etc.

Individuals may be responsible for misusing data by creating ‘fake news’ through graphics and sharing these on social media. There are also issues related to stalking for example, where a person can use the data of others maliciously, including accessing location tracking data which is a common feature of many applications.

Some individuals may want to exploit the data of others in criminal ways. For example, identity theft which could be (medical/financial/child etc.) or credit card cloning.

Hence there are issues of personal data privacy to consider, and individuals should understand how to protect themselves online.

Organisations

Two-thirds of companies worldwide have completed at least one big data implementation so far, typically starting with focused initiatives to improve personalization (for instance, data-driven insights feed location-based services such as special offers to customers), or to optimise operations (for instance, data mined from smart devices such as pipelines or planes allowing for predictive maintenance and asset optimization).

Data Analytics for decision-making

With the availability of advanced big data analysing technologies (e.g. NoSQL Databases, BigQuery, MapReduce, Hadoop, WibiData and Skytree), insights can be better attained to

enable in improving business strategies and the decision-making process in critical sectors such as healthcare, economic productivity, energy futures, and predicting natural catastrophe, to name but a few (Yi et al., 2014).

Data can be misused by organisations, for example in the Cambridge Analytica Scandal, Facebook data was used in order to influence voting in America. Another example might be phone hacking, in one case there was the illegal acquisition and use of personal data by the News of the World to report on the Milly Dowler case.

Society

Data can be used to improve public services and is collected across society in areas of healthcare, transport, criminal justice, etc. and at national, local and regional scales.

Data can be misused by governments in order to control their populations. For example, data collected about citizens that can be used for surveillance and oppression. Data can also be used to monitor citizens behaviour; for example research conducted by the Scottish Centre for Crime and Justice Research and based on a survey of all Scottish local authorities, identified that there were over 2,200 public space CCTV cameras in Scotland.

Another example of the misuse of data relates to recent uses of social media information, as Carmi and Yates (2020) explain:

> *In the UK context, the Facebook/Cambridge Analytica scandal in 2017 revealed that people received disinformation content and advertisements based on their social media profiles and activity, designed to influence their decisions on the 2016 UK Referendum to leave the European Union, and the 2016 US presidential election. These two cases made clear the extent to which citizens are unaware of the uses and abuses to which our data can be put. This lack of data literacy opens citizens up to risks and harms – personal, social, physical and financial – but also limits their ability to be proactive citizens in an increasingly datafied society.*

Data is also used within global society for example WHO, and World Bank information about country metrics can tell us information about a range of indicators.

From the above, we can consider the fact that there are forms of data which are public, and forms which are private. Public data would include census data, and other information published by the office for national statistics, whereas banking information would be private personal data.

Summary points:

- The amount of data in society is increasingly exponentially
- The number of internet users and data citizens varies around the world
- There are a huge variety of uses, and misuses of data in society

Reflective Questions:

- Which country has the most data citizens in the world?
- What are some examples of how data is generated from your actions in society?
- How many bytes of data are being generated by smart devices each day?
- What are the three main components of the data citizenship framework?

The Principles and Concepts of Data Science

Although data science isn't a new profession, it has evolved considerably over the last 50 years. A trip into the history of data science reveals a long and winding path that began as early as 1962 when mathematician John W. Tukey predicted the effect of modern-day electronic computing on data analysis as an empirical science.

Co-founder of LinkedIn Allen Blue said recently: "It's hard to believe that just 20 years ago, the whole notion of data science was brand new. I don't think anybody had even used the term yet." The increasing popularity of the internet was the catalyst, he said. Every action an internet user took was visible and trackable, leading to a huge and continuous volume of captured information. This, in turn, enabled novel types of research. Computer scientists began doing real-time analytics of how people were using particular products.

Today, data science is defined as a multidisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from data. It emerged thanks to the convergence of a wide range of factors: New ideas among academic statisticians, the spread of computer science across various fields, and a favorable economic context.

As the falling cost of hard drives allowed companies and governments to store more and more data, the need to find new ways to value it arose. This boosted the development of new systems, algorithms, and computing paradigms. Since data science was particularly appropriate for those wanting to learn from big data, and thanks to the emergence of cloud computing, it spread quickly across various fields. With faster processing speeds than ever before, technology took a giant leap into the new decade, blazing a trail for individuals ready and willing to conquer the mountain of big data that had only begun to grow.

Companies had also begun to view data as a commodity upon which they could capitalize. Thomas H. Davenport, Don Cohen, and Al Jacobson wrote in a 2005 Babson College Working Knowledge Research Center report, "Instead of competing on traditional factors, companies are beginning to employ statistical and quantitative analysis and predictive modeling as primary elements of competition."

Machine learning, AI and Data Science

The fields of Machine Learning, Artificial Intelligence and Data Science are huge. Due to this, there are a lot of confusions between the three. Even though the terms all fall in the same domain and are connected to each other, they have their specific applications and

meaning. There may be overlaps in these domains every now and then, but essentially, each of these three terms has unique uses of their own.

Artificial Intelligence

Artificial intelligence is a field of computer science in which computer systems are developed such that they can perform tasks with human-like intelligence. There are a variety of tasks such as speech recognition, image recognition, decision making systems etc. The term artificial intelligence itself is self-explanatory. This intelligence in computer systems is built by humans using some techniques and algorithms like Natural Language Processing (NLP) or computer vision etc. The machines are developed with the intent of making it intelligent enough to work and react like humans. Data plays an important part of AI. It is with the help of the large amount of data known as big data that these systems perform well. The field of artificial intelligence is vast and it comprises subfields like machine learning and deep learning.

Machine learning:

Machine learning is a subfield of artificial intelligence. Machine learning is used to make future predictions for a particular problem based on the historical data. The most common example of application of machine learning that you can relate to, is detecting whether an email is spam or not. You may have noticed that your email provider helps to detect the spam emails based on the previous email that you marked as spam. This prediction is done using various algorithms like regression or classification.

Machine Learning (ML) is considered a subset of AI. You can even say that ML is an implementation of AI. So whenever you think of AI, you can think of applying ML there. As the name makes it pretty clear, ML is used in situations where we want the machine to learn from the huge amounts of data we give it, and then apply that knowledge on new pieces of data that streams into the system. There are different ways of making a machine learn. Different methods of machine learning are supervised learning, non-supervised learning, semi-supervised learning, and reinforced machine learning.

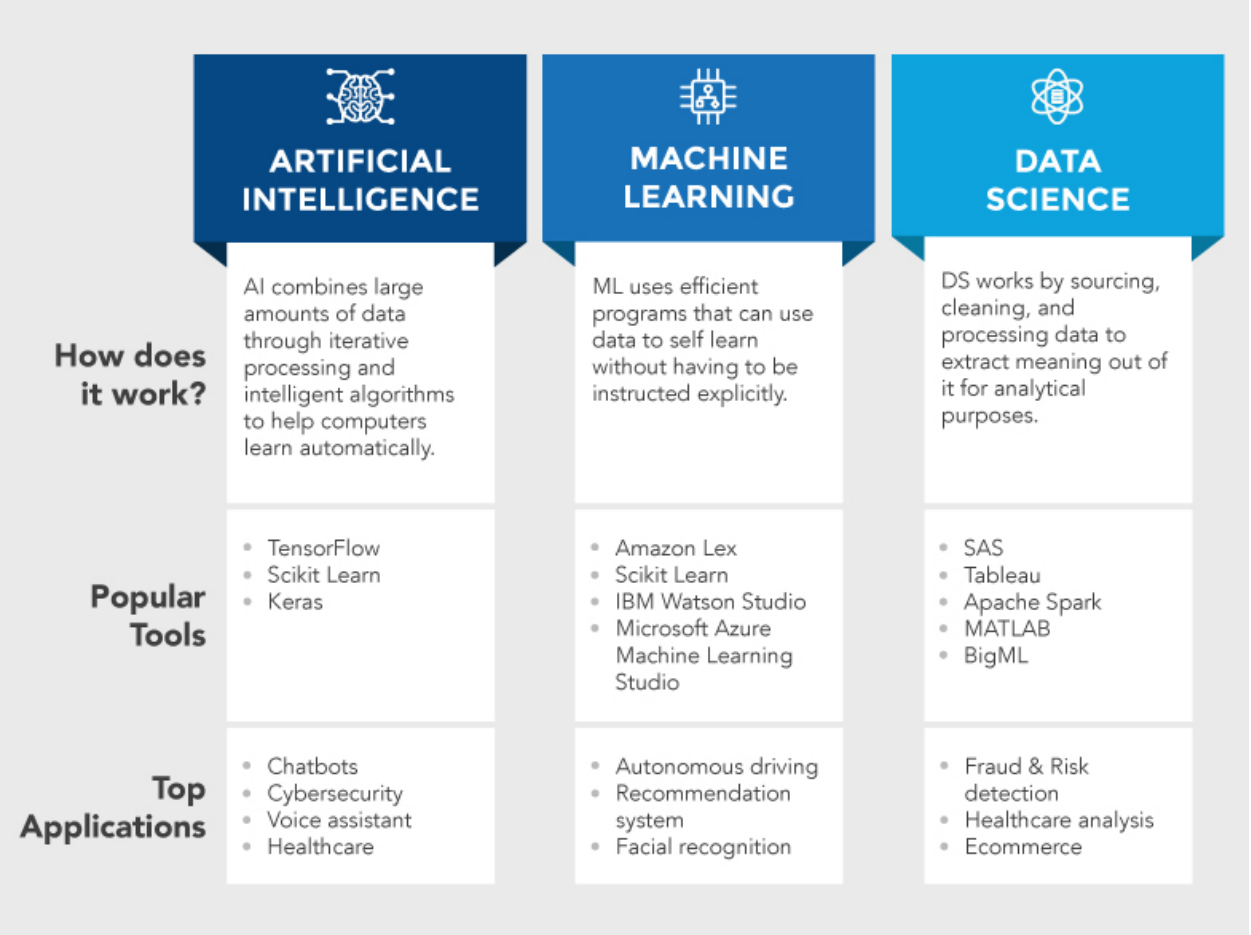
Data Science:

Data science is a field where data scientists derive valuable insights from large volumes of data. The insights derived by data scientists helps companies grow their business. Data science involves analysis of data, modelling of data etc. This field involves part of machine learning for making predictions and for modelling. The other sections of data science include data extraction, data exploration, data preparation, data visualisation etc. This field is growing rapidly as the amount of data being generated day by day is growing exponentially. This generated data needs to be processed and used to grow business. The data scientists are people who have good expertise in programming, machine learning, domain knowledge, mathematics and statistics. A Data Scientist is also known as a person

who is better at statistics than any software engineer and better at software engineering than any statistician.

Data science on the other hand is a field which is related to drawing insights and making predictions using data. But those predictions are built using machine learning algorithms. Data science involves statistics and machine learning. Machine learning is the link between data science and Artificial Intelligence. But for making predictions you need clean and well prepared data. The process of data extraction, data engineering, data pre-processing etc. is done in data science which is then used for predictive modelling in machine learning.

For example, in the case of self-driving cars, the data are collected through images and sensors which are then processed using machine learning techniques like deep learning.



All these three technologies are related because they all rely on data. Data is an important factor in these technologies. Usually you will see that the terms Data Science, Artificial Intelligence and machine learning are being used interchangeably in the industry. But it is important that in all these fields data is a component that connects these technologies.

Data science is intricately intertwined with other important concepts also of growing importance, such as big data and data-driven decision making.

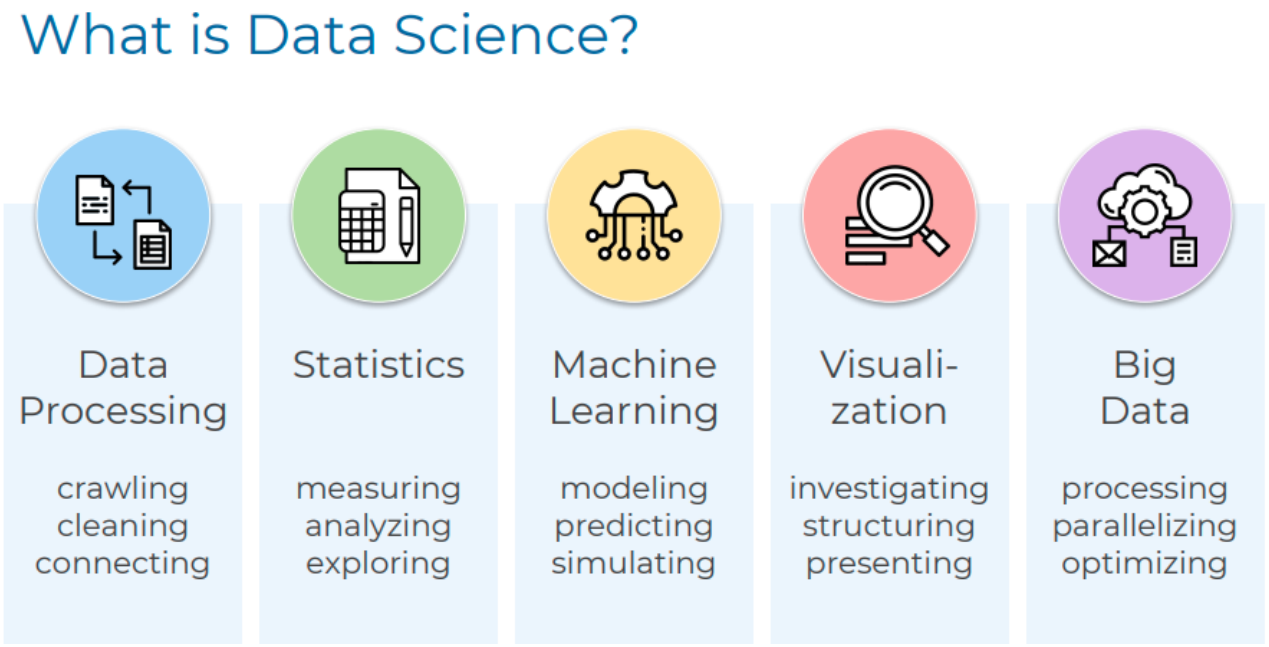
Data Science, artificial intelligence, and machine learning work in tandem to exploit data for a wide variety of business benefits.

Simply put, machine learning is the link that connects Data Science and AI.

That is because it's the process of learning from data over time. So, AI is the tool that helps data science get results and the solutions for specific problems. However, machine learning is what helps in achieving that goal. A real-life example of this is Google's Search Engine. Google's search engine is a product of data science. It uses predictive analysis, a system used by artificial intelligence, to deliver intelligent results to the users. For instance, if a person types "waterproof jacket" on Google's search engine, then the AI collects this information through machine learning. Now, as soon as the person writes these two words in the search tool "best place to buy," the AI kicks in, and with predictive analysis completes the sentence as "best place to buy waterproof jackets" which is the most probable suffix to the query that the user had in mind.

Doing Data Science

In practice doing data science involves a whole range of activities outlined in the graphic below, such as data processing, statistics, programming, visualization and working with big data. Data science techniques have been applied in many spheres of society, such as medical research and healthcare provision, sports, politics and commerce.



Summary points:

- Data science is a relatively new concept emerging only as early as the 1960s
- Data science, artificial intelligence and machine learning are interconnected but separate domains

- Data science involves different techniques for working with data to generate insights
- Data science is a rapidly growing field and data scientists are required across many different parts of society

Reflective Questions:

- Can you explain the relationship between Data Science, AI and Machine Learning?
- Who predicted the emergence of modern day computing as an empirical science?
- What is one example of a product of data science?

Self-assessed Topic Activity:

- Consider your footprint in data, how many photos you have on social media, how many on your phone or laptop, how much data you have stored in cloud services. How long has it taken to accumulate this data, where is it stored?
- Make a list of the organisations that hold data about you, who has your email address, who knows your address - can you think of 10 organisations?

References

Carmi, E. & Yates, S. J. & Lockley, E. & Pawluczuk, A. 2020. Data citizenship: rethinking data literacy in the age of disinformation, misinformation, and malinformation. Internet Policy Review, 9(2). Available at: https://policyreview.info/articles/analysis/data-citizenship-rethinking-data-literacy-age-disinformation-misinformation-and

Chatterjee, M. 2020. Data Science Vs Machine Learning And Artificial Intelligence. [online] Available at: <https://www.mygreatlearning.com/blog/difference-data-science-machine-learning-ai/ > [Accessed 26 January 2021].

Desjardins, J. 2021. How Much Data Is Generated Each Day?. [online] Visual Capitalist. Available at: <https://www.visualcapitalist.com/how-much-data-is-generated-each-day/ > [Accessed 26 January 2021].

Ghosh, P. 2019. Data Science Vs. Machine Learning Vs. Artificial Intelligence. Dataversity. [online] Available at: <https://www.dataversity.net/data-science-vs-machine-learning-vs-artificial-intelligence/> [Accessed 26 January 2021].

Knickrehm, M., Berthon, B, and Daugherty, P. 2016. Digital disruption:The growth multiplier Optimizing digital investments to realize higher productivity and growth. Accenture. Available at: <https://www.accenture.com/_acnmedia/pdf-14/accenture-strategy-digital-disruption-growth-multiplier-brazil.pdf> [Accessed 26 January 2021].

Knowledge@Wharton. 2019. What's Driving The Demand For Data Scientists? - Wharton University of Pennsylvania. [online] Available at:

<https://knowledge.wharton.upenn.edu/article/whats-driving-demand-data-scientist/ > [Accessed 26 January 2021].

Provost, F. and Fawcett, T. 2013. Data Science And Its Relationship To Big Data And Data-driven Decision Making [online] Available at: <https://www.liebertpub.com/doi/pdfplus/10.1089/big.2013.1508 > [Accessed 26 January 2021].

McGregor, C. and Emejulu, A. 2019 Towards a radical digital citizenship in digital education, Critical Studies in Education, 60(1): 131-147.

Roser, M., Ritchie, H. and Ortiz-Ospina, E., 2021. Internet. [online] Our World in Data. Available at: <https://ourworldindata.org/internet > [Accessed 26 January 2021].

Sivarajah, U. et al. 2017. Critical analysis of Big Data Challenges and Analytical methods. Journal of Business Research. 70: 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

Srinidhi, S. 2019. Data Science Vs. Artificial Intelligence Vs. Machine Learning Vs. Deep Learning. Towards Data Science. [online] Available at: <https://towardsdatascience.com/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-9fadd8bda583 > [Accessed 26 January 2021].

Shreyas, S.. 2019. Relationship Between Artificial Intelligence, Machine Learning And Data Science. [online] Available at: <https://medium.com/@shreyasb494/relationship-between-artificial-intelligence-machine-learning-and-data-science-15a87e2cc758 > [Accessed 26 January 2021].

Yates, S., Carmi, E., Pawluczuk, A., Wessels, B., Lockley, E., & Gangneux, J. 2020. Understanding citizens data literacy: thinking, doing & participating with our data (Me & My Big Data Report 2020). Me and My Big Data project, University of Liverpool: Available at: https://www.liverpool.ac.uk/humanities-and-social-sciences/research/research-themes/centre-for-digital-humanities/projects/big-data/publications/

Yi X., Liu F., Liu J., Jin H. 2014. Building a network highway for big data: architecture and challenges. IEEE Network, 28(4): pp. 5-13