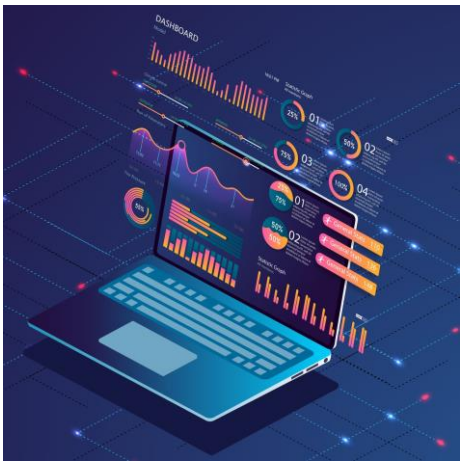
An introduction to data science and machine learning

This resource was developed by teachers within the Royal Society Schools Network



iStock image: credit NatalyaBurova

Curriculum key words

Data

Machine learning

Curriculum links

Computing:

- select, use and combine a variety of software (including internet services) on a range of digital devices to design and create a range of programs, systems and content that accomplish given goals, including collecting, analysing, evaluating and presenting data and information

Equipment needed

– sheet of paper to record data around a person or book character

Resources

– Example of Google map

KS2

Introduction

The aims of the National Curriculum for computing state that:

"A high-quality computing education equips students to use computational thinking and creativity to understand and change the world. Computing has deep links with mathematics, science, and design and technology, and provides insights into both natural and artificial systems. The core of computing is computer science, in which students are taught the principles of information and computation, how digital systems work, and how to put this knowledge to use through programming."

In today's world it is impossible to ignore the presence of data, data science and machine learning. As more and more data is collected about us and is used by data scientists, it is vital that we educate our students on how data is collected, how it is used and how it impacts our daily lives both now and in the future.

The purpose of this lesson is to introduce students to the concept of data, data science and machine learning. It could take place over several lessons and could link with data handling in maths.

Learning objective:

To understand how data science is used to solve real world problems.

Success criteria (SC):

- SC1: I can identify the types of data that can be collected about people
- SC2: I can explain how data sets can be used to understand and therefore solve important problems.

An introduction to data science and machine learning

Starter activity: How google collects data

(Approximately 15 – 20 mins) [SC1]

Ask the students to write down / tell an adult as many facts as they can either about a character from a book that the class is reading or about someone they all know well. Add these around a central image of the character.

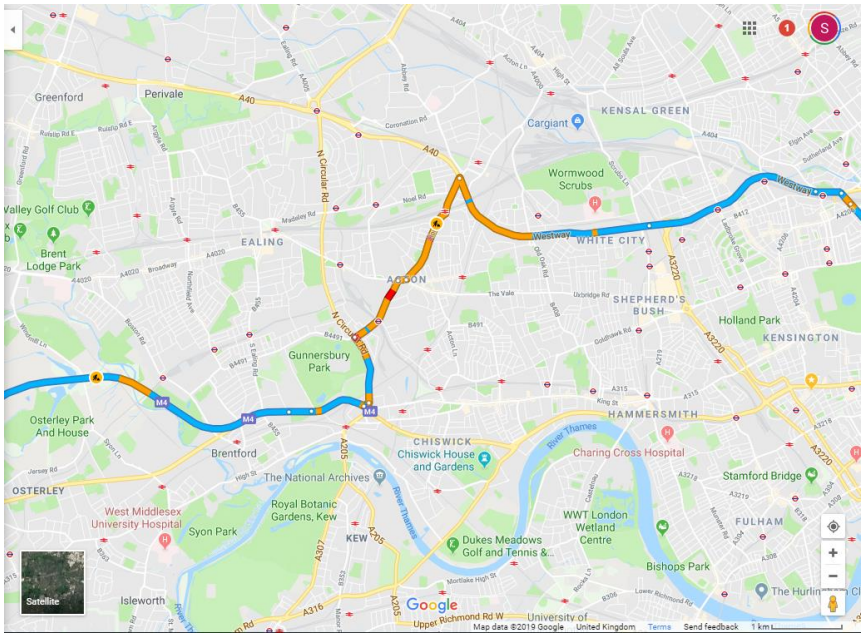
Explain that each of these facts is a piece of data about the person.

Explain that data is collected about all sorts of things in lots of different ways and all the time.

Give some examples about how data is collected e.g. browsing a website providing data about what websites we like to visit, use of a loyalty card collects data about what we buy, mobile phones and GPS data giving data about location – show them the location icon and ask if they recognise this.

Location icon

Open Google Maps, select a a local area to you where you know there is traffic and find a busy journey. Show the students how the route changes colour to show where there is really busy traffic.



Example Google map screenshot: Showing where busy traffic is located in a part of London.

Ask the students how Google Maps knows when a route is busy (Google uses anonymized data from mobile phones, calculates the speed users are traveling along a length of road and then generates traffic data accordingly).

This shows us how computer scientists use data to help us find out about something.

Activity A: Class discussion about data and data science

(Approximately 10 – 20 mins) [SC1]

Ask the students the following key questions:

- What is data?
- What is data science?

Gather their responses and discuss each, eventually coming up with class definitions along the lines of:

- Data is facts and figures that we can use to help us learn something.
- Data science is using scientific methods to investigate large volumes of data to learn about something.

Activity B: early data science - John Snow's cholera map

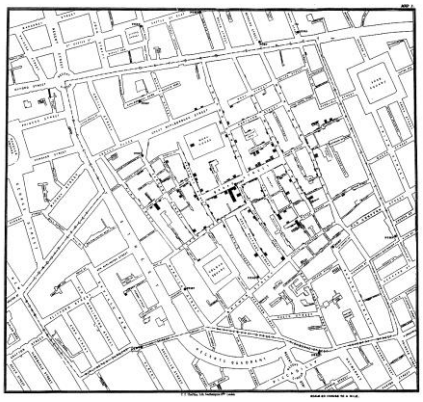
(Approximately 10 – 20 mins) [SC2]

This is an example of how data was used in the 19th Century to solve a very serious problem.

Background information to share with the students:

In 1854 it was believed that a horrible disease called cholera was spread through the air. There was one scientist, however, who believed that cholera entered the body through the mouth. His name was John Snow and he published his ideas in an essay called 'On the Mode of Communication of Cholera'. In August, just a few years later, he was able to use data to prove his theory. There was an outbreak of cholera in Soho in London. John Snow collected data, plotting each outbreak on a map.

Provide groups of students with the following map (enlarged - avaialbe in the public domain from Wikipedia https://en.wikipedia.org/wiki/John_Snow).



Instruct the students to look carefully at the map and ask the following three questions:

- Can you see the lines used to identify the outbreaks?
- What do you notice?
- Can you draw any conclusions from the map?

As a result of John Snow's analysis of the data he identified that many of the outbreaks were centred around a particular part of London.

- Can you find that on the map?
- What else can you see in that area?
- Can you see the water pump situated in Broad Street?
- Have a look at the number of outbreaks near it.

They removed the handle of the pump so that no one could use it and the incidents of cholera reduced dramatically.

- What does this suggest?

This just shows how important collecting and analysing data can be. It literally saved lives!

Teacher note: On Wikipedia there is an extract from a letter to the editor of the Medical Times and Gazette regarding the cholera map which might be interesting to share with the students.

Activity C: how is data collected today?

(Approximately 5 – 10 mins - class discussion or direct instruction)

Q) Why is data science so important today?

Data science is so important today because there is so much data available and computers are much better at processing it than they were in the past.

Q) What is big data?

Big data is the description used for huge volumes of data. The more data available, the more we can learn from it. Some of it is 'open' which means anyone can have access to it and use it.

Q) Why is there so much data available now?

Data is collected in many different ways. How many different ways can you think of?

- Loyalty cards – shopping choices
- Mobile phones - location

- Web browsing – which sites you visit
- Doctors – what treatment you have received
- Photographs – where and when photos were taken
- CCTV – images of who has been where

Q) Can you think of any other examples of where data might be collected?

Activity D: Machine Learning Infographic

(Approximately 10 – 25 minutes) [SC1, SC2]

Computers can now use large amounts of data to help them learn. Find out how using the Royal Society's What is machine learning infographic (https://royalsociety.org/topics-policy/projects/machine-learning/what-is-machine-learning-infographic/)

Teacher note: The infographic consists of explanations, illustrations and interactive elements.

It starts by giving a short explanation of what machine learning is and then leads students to do a short quiz to see if they can identify which situations use machine learning in the real world. The first example is a driverless car. When a student makes their selection they then get a short explanation associated with that example. There are several examples to click through.

The next part of the infographic explains that more data is being collected than ever before. A diagram of 'Big Data Universe' compares data stored by different online services and the human brain. The data is measured in petabytes (1,000,000 gigabytes).

Some questions you could discuss using the illustration:

- What data is stored by each service?
- Why is Spotify so much smaller than Google?
- The illustration is from 2016 – would the sizes have changed by 2019?
- Allow the students to ask their own questions relating to the illustration.

There is then an animated graph showing how much faster computers have become at processing data from 1970s to 2014 – there is a dramatic increase up to 2014.

Following this is a ‘how machine learning works’ photo challenge. Students can compete against Google to identify images of dogs and cats. Once the student has completed the challenge, get them to click on the play arrow to see an animation of Google carrying out the same task (takes less than a millisecond!).

An explanation for how Google trained its computers to carry out this task is in written form and in gallery form.

The next interactive element shows how a driverless car uses machine learning to help it adapt its driving to different objects that are found on roads. Prior to the students viewing this it would be good to have a discussion along the lines of:

- You may have seen a lot about driverless cars in the news. These are cars which can drive themselves by using lots of data to help them make decisions when driving along a road.
- Let’s think of the decisions that we might make when we are driving along the road.
- What would you do if you saw a person crossing the road in front of you?
- What would you do if a cat ran out in front of the car?
- What would you do if you were going to turn left but there was a sign saying that the road was closed?
- What would you do if the car in front of you slowed down?

It would then be possible to click through the different parts that the car learns to deal with and write pseudo code for the car:

E.g. Edge Detection

If car is > 20cm from left white line

Then turn 10 degrees right

Else if car is > 20cm from middle white line

Then turn 10 degrees left

E.g. Vehicle Detection

If car is > 0.5m from car in front

Then decelerate by 5 miles an hour

Repeat forever

Activity E: machine learning in the world around you

(Approximately 10 – 25 minutes) [SC1, SC2]

Allow students to independently explore the Royal Society Machine l earning in the world around you Infographic **(**https://royalsociety.org/topics-policy/projects/machine-learning/machine-learning-in-the-world-around-you-infographic/ **)**

In groups, select one of the parts of the town e.g. Shopping Centre and discuss the pros and cons for machine learning in that environment. Present these back to class.

It would also be possible to complete a P4C (philosophy for children) activity around the theme 'machine learning is good'.

Activity F: Next Steps (anything from 20 mins to have a quick go to a whole lesson)

Use the site: https://machinelearningforkids.co.uk/ to allow students to create their own machine learning programs using Scratch!

Plenary:

Students make a poster explaining what data is collected about people, how all this collected data can be used to help us but also why we should possibly be worried.