How does the amount of data affect a sorting activity?

Objective

In this lesson students will develop their own rules/ algorithms for sorting objects and also look at the effect that increasing the size of the data set has.

Introducing the experiment

Watch the Oxford Sparks *Machine Learning* video https://youtu.be/f_uwKZIAeM0

Based on the video, ask students (in pairs) to come up with three different things that machine learning is already used for or might be used for in the future.

Things mentioned include:

- Facial, text and speech recognition (including photo tagging/predictive text).
- Spam filters.
- Online viewing or shopping recommendations.
- Credit card fraud recognition.
- Medical diagnosis.
- Social media.

During the experiment

Introductory activity: Thingies and Whatsits

- Two new types of creature have been discovered, Thingies and Whatsits.
- Hand out or display diagram 1 from the Thingies and Whatsits diagram sheet.
- Ask students in pairs or small groups to decide which group the next three creatures belong to and why.
- Reveal the answers: Whatsit, Thingy, Whatsit
- Several groups are likely to have got the wrong answers. Emphasize that this is very likely at this stage and getting the right answer is probably just luck. In fact, in machine learning, getting the answer wrong helps a machine create better rules.
- Show them diagram 2. How might they change their rule now?
- Ask them to choose either a ‘thingy’ or ‘whatsit’ and draw a version that hasn’t been seen in any of the pictures.
- Ask a few students to show their creations.
- Reveal the rule – Whatsits have 4 straight sides and Thingies don’t (nothing else matters – colour, number of eyes etc). Do their new creations fit the rule?
- Optional – Students play a game with each other where they decide the rules for a new creature (a whojamacallit) and other students have to try to guess the rules either by asking questions or sketching a creature and finding out whether it is a whojamacallit or not.
- Machine learning is very dependent on the data that we feed it. In order to produce accurate results we need to feed it a large amount of data but also data that is as varied as possible. For instance if we were creating a voice recognition program you would probably want to use a large number of words but also different accents or voice types. If we only fed it words spoken by men, for instance, it might struggle to understand women.

Main activity:

- Give students the student worksheet.
- Give the students 4 – 6 different liquorish allsorts, dolly mixtures, Haribo Starmix or similarly varied sweets (you could use photographs if you prefer).
- Ask them to create an algorithm to sort the sweets by asking a series of yes/no questions. eg Is it square? Is it yellow? Try to encourage students to focus on more general properties of the sweets rather than just asking 'is it a cola bottle?' or similar.
- The simplest way of representing the algorithm is to use a flow chart, decision tree or similar graphical representation.
- Get students to swap their algorithm and sweets with other groups and check that they can use them to get the same answers.

Discussion points after the activity

- Discuss with students what types of questions are best for creating algorithms.
- Ask them to practice what they have learnt by creating an algorithm to identify species of dogs. They could use the internet to source images of dogs (or cats).
- Ask students to decide whether they agree/disagree with some of the questions and answers they have come up with.
- Ask the students how such a machine learning system can improve the accuracy of the results. They should realise that having more data (species of dogs) will usually improve the results.

You could also ask them to try this game at home:

- 20 Questions is an app/website and stand alone toy that uses machine learning to guess the object that you are thinking of by asking a series of questions. The computer algorithm uses the answers that it gets from thousands of previous users to improve how quickly and accurately it can guess the object you are thinking of.
- The program reports back on how other users have answered the same questions.
- Although this algorithm is actually very good at playing 20 Questions it does rely on the users answering correctly. If users answer incorrectly either by accident or on purpose it alters the accuracy of the result.

For more school experiments and to access the accompanying videos, visit **royalsociety.org/schoolexperiments**

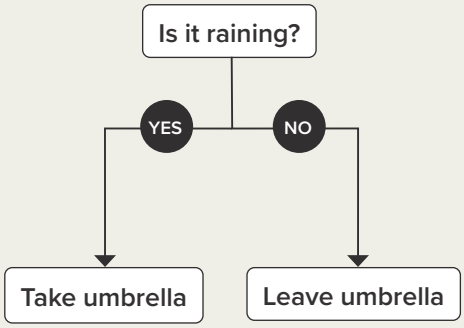
THANKS

This activity is based on an original activity, called *All sorted*, designed by Oxford Sparks (from the University of Oxford; www.oxfordsparks.ox.ac.uk) who have kindly agreed to its use in the Royal Society Brian Cox School Experiments.

How does the amount of data affect a sorting activity?

Computer algorithms frequently use decision trees. A decision tree is often a set of questions with yes/no answers that determine which commands the computer program runs. Each branch is the answer to a different question.

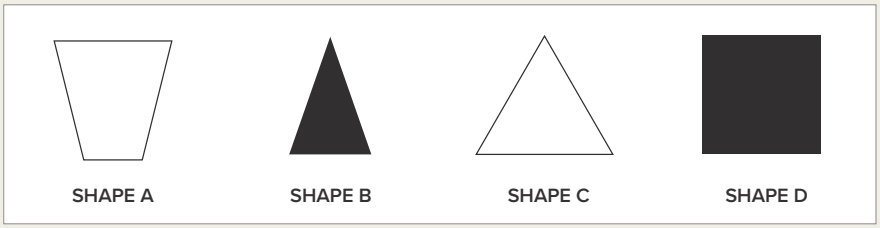
A simple decision tree might be:

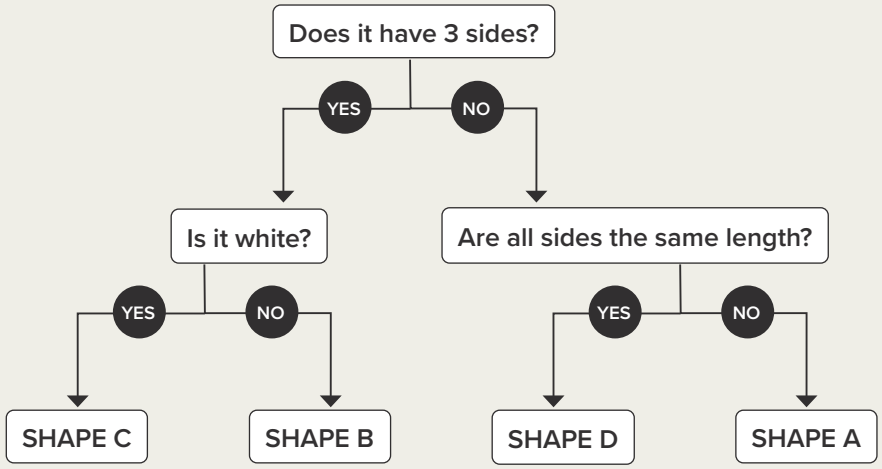


There are often several different ways of creating decision trees.

Can you think of a different way of sorting these shapes?

Decision trees can be used to sort or classify objects.
If we wanted to write a decision tree to sort the following four objects, we could use the following decision tree:





Sorting sweets

Put each different sweet into a box below. If you have fewer than 6 sweets you will have some spare boxes.

(A)	(B)	(C)	(D)	(E)	(F)

Create a decision tree below that will sort the sweets by asking a series of yes/no questions about the sweets until the answer is one of the sweets above. When you are finished swap your sweets and decision tree with another group and check that you get the same answers as they do. You might want to ask questions about shape/colour/number of different layers/parts etc.

Your decision tree

THANKS

This activity is based on an original activity, called *All sorted*, designed by Oxford Sparks (from the University of Oxford; www.oxfordsparks.ox.ac.uk) who have kindly agreed to its use in the Royal Society Brian Cox School Experiments.

How does the amount of data affect a sorting activity?

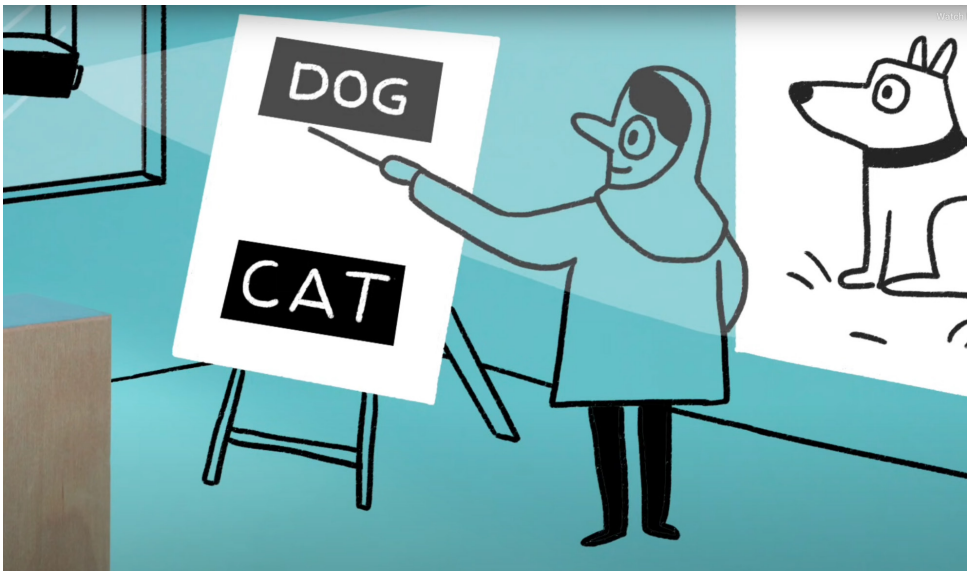
Background

Machine learning is a system where rather than a computer programmer deciding the best way to sort, organise, classify or use information, the computer program develops its own set of instructions (algorithm) based on information that users feed it. Scientists are working on ways to improve the speed and accuracy of these algorithms.

In this lesson students will develop their own rules/algorithms for sorting objects and also look at the effect that increasing the size of the data set has.

Preparation notes

The access to the internet is for sourcing images of dog species (plenary activity) or the 20Q website (although preferable to set this as a homework activity).



EQUIPMENT LIST

Materials for each student

- Paper and pen.
- Printout of *Thingies and whatsits* diagram sheet.
- Student worksheet.
- Liquorish allsorts, dolly mixtures, Haribo Starmix or similarly varied selection of sweets.
- Access to the internet (optional – see preparation notes).

SAFETY PRECAUTIONS

- When choosing the sweets, make sure to check for allergens that may be an issue.

FURTHER LINKS

- *What is Machine Learning?*
- Machine learning resources

THANKS

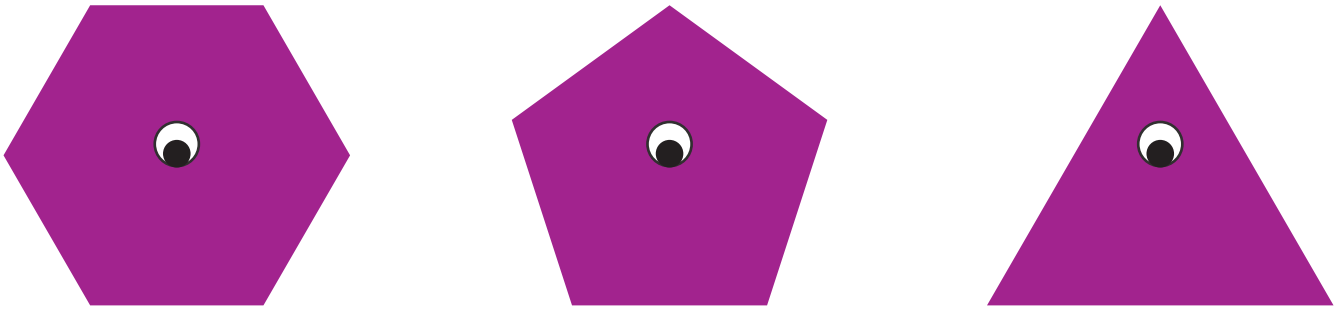
This activity is based on an original activity, called *All sorted*, designed by Oxford Sparks (from the University of Oxford; www.oxfordsparks.ox.ac.uk) who have kindly agreed to its use in the Royal Society Brian Cox School Experiments.

How does the amount of data affect a sorting activity?

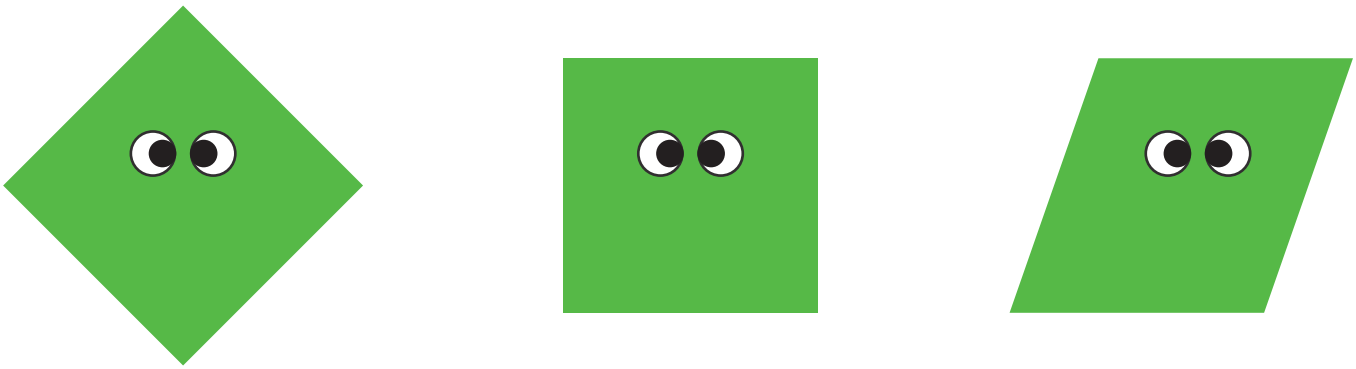
Introductory activity: Thingies and Whatsits

Diagram 1

Thingies



Whatsits



What are these?

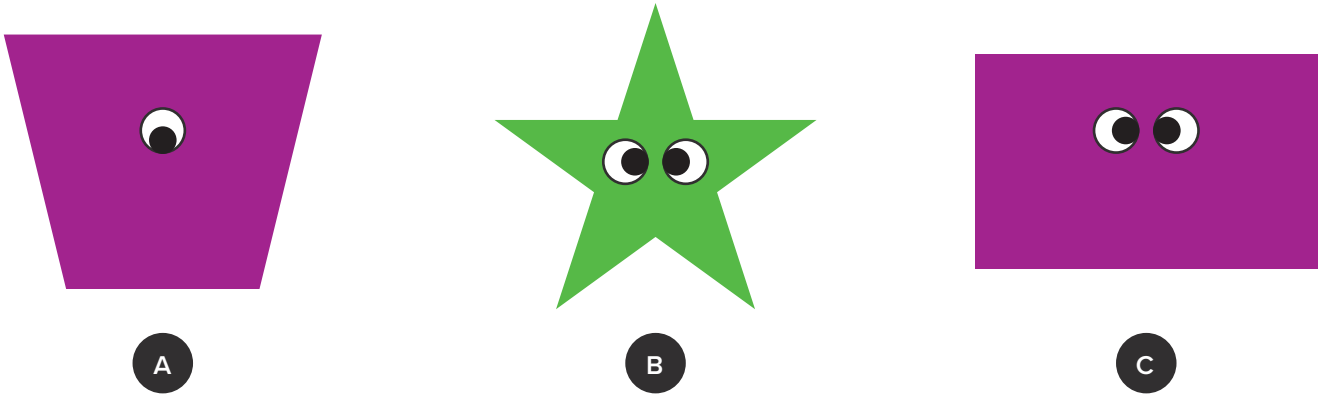
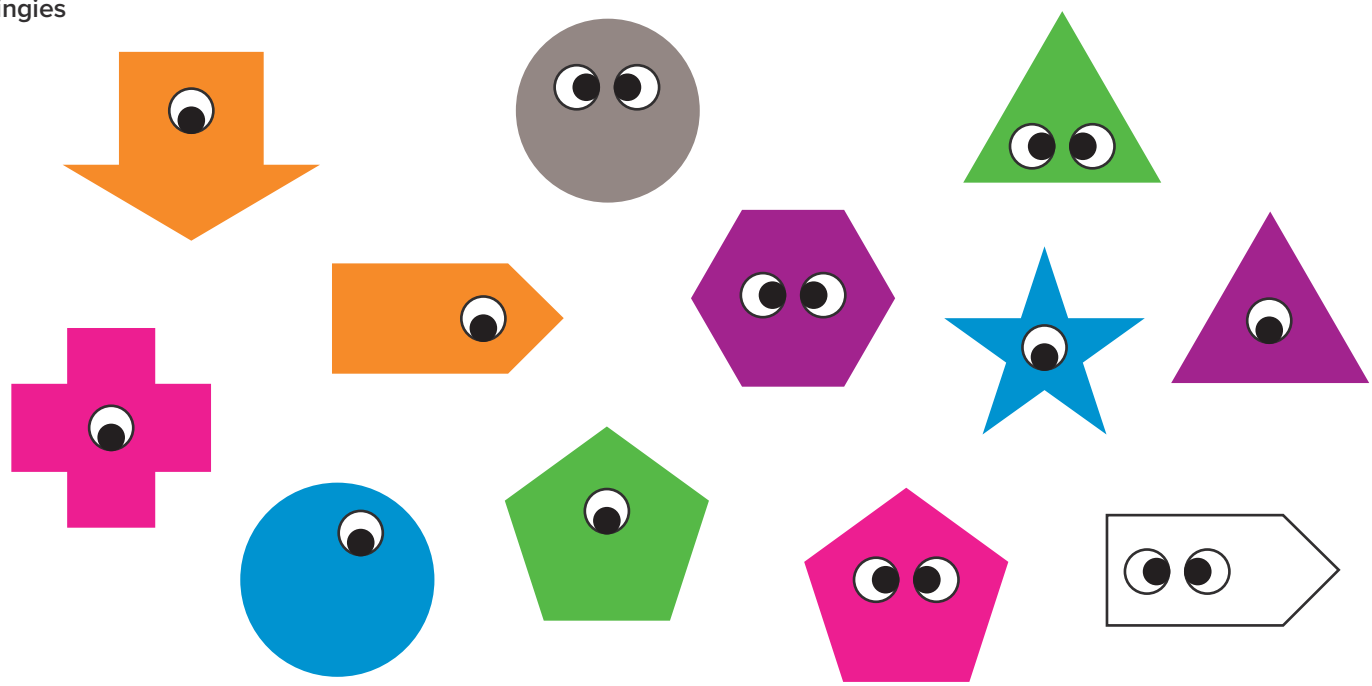
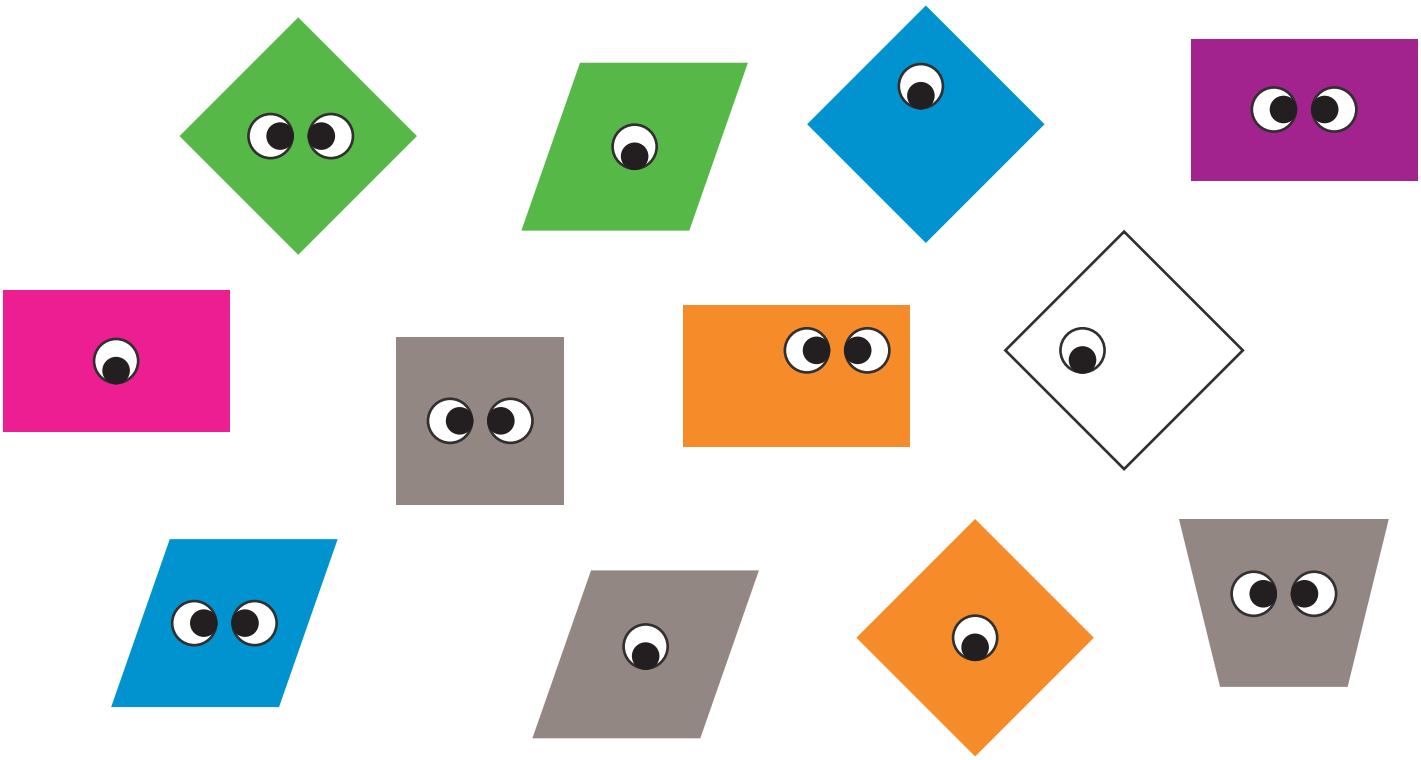


Diagram 2

Thingies



Whatsits



THANKS

This activity is based on an original activity, called *All sorted*, designed by Oxford Sparks (from the University of Oxford; www.oxfordsparks.ox.ac.uk) who have kindly agreed to its use in the Royal Society Brian Cox School Experiments.