Decoded: Machine learning

Dr David Tarrant
@davetaz
theODI.org

Violeta Mezeklieva

theODI.org

Our founders

Dr Jeni Tennison
CEO

Sir Nigel Shadbolt
Chairman

Sir Tim Berners-Lee
President

Founded in 2012, the Open Data Institute (ODI) is an international, independent and not-for-profit organisation based in London, UK.

Me

Dr David Tarrant
Learning & Technology

12+ years experience in Open Data

Established first degree level module in Open Data at the University of Southampton

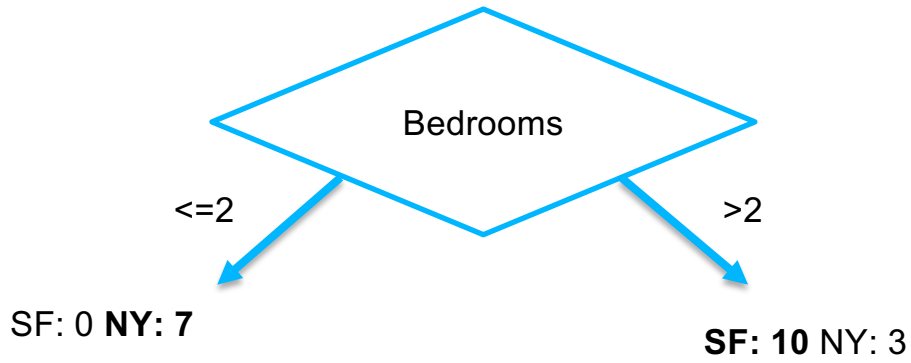
Part of the team that established the ODI

Have since helped transform governments and unlock over $15m for startups

Aim

Equip you with the skills and knowledge to both develop and critique applications of machine learning and AI.

Remember me?



NY

Correct: 17
Wrong: 3

SF

85%
Confidence

Each table has a set of "Top Trump" cards relating to properties in two cities.

Build a ONE LEVEL decision tree to sort them into "New York" and "San Francisco".

You cannot use the name of the city to sort them.

Review

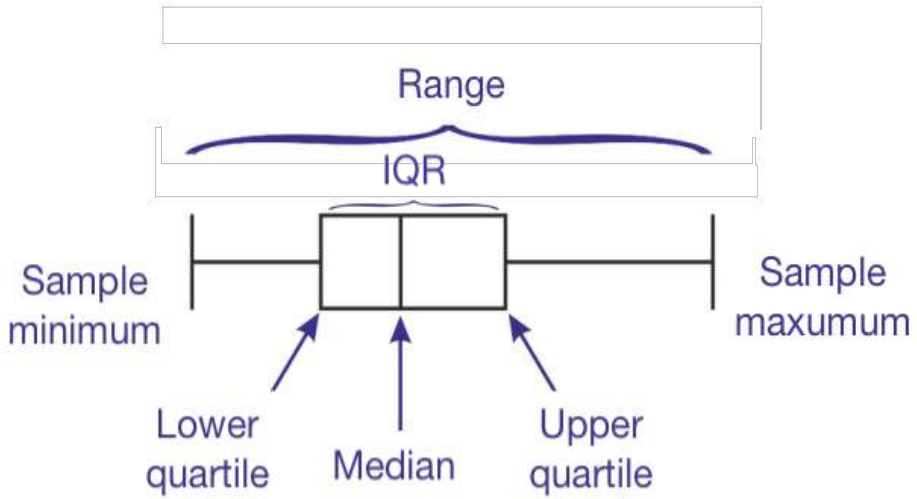
What approach should we take?

What do we need to avoid?

How do we pick the right variable to use?

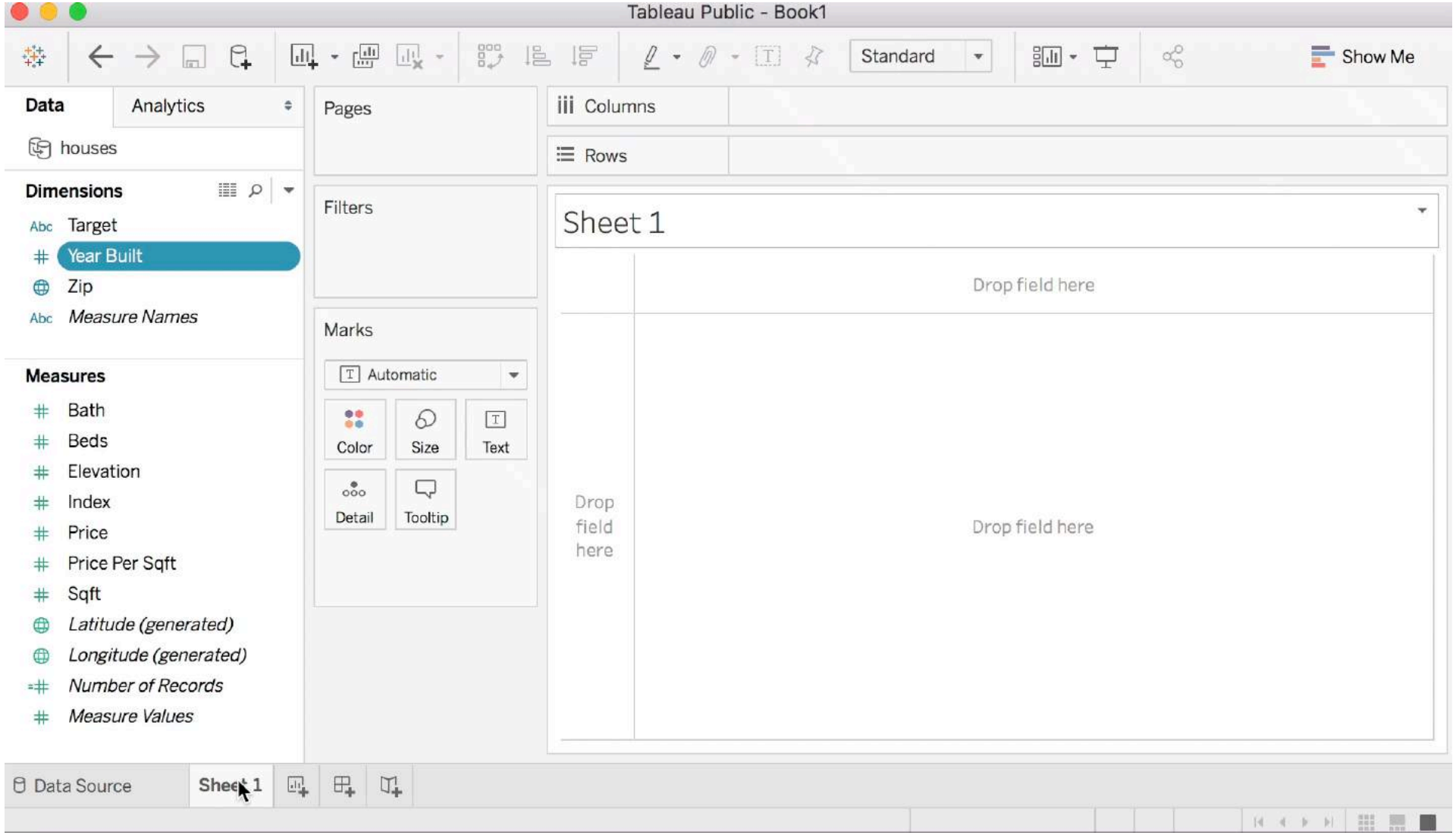
What is the right threshold?

Which one is better statistically?

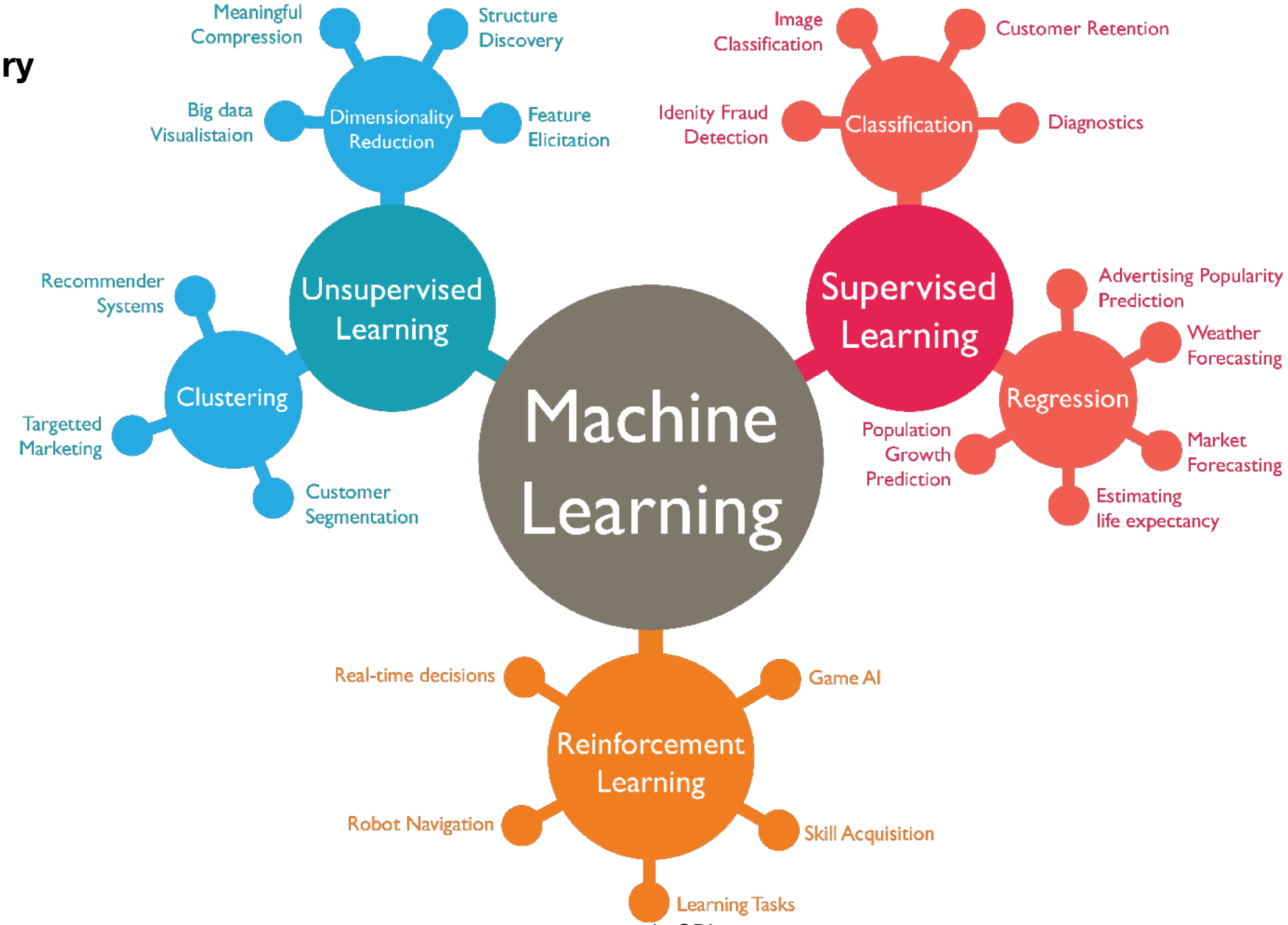


open data institute

Data analysis for the houses dataset in Tableau



Summary


Machine Learning
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Agenda:

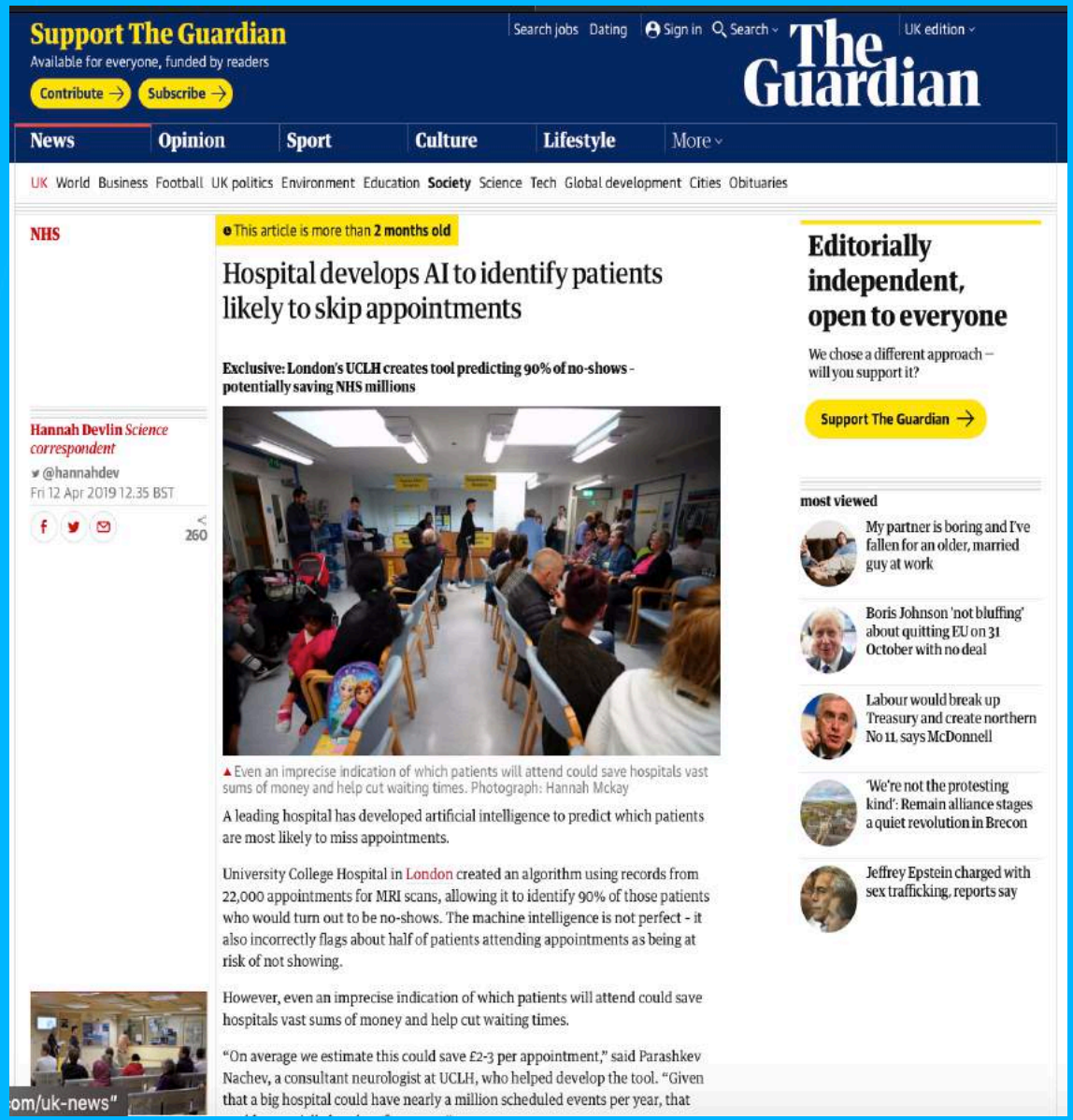
Session 1	Decoding the terminology
Workshop 1	Can machine learning save the NHS?
Session 2	Holding algorithms to account
Workshop 2	Injustice the justice system
Session 3	Evaluating black box systems
Workshop 3	Deep fake detection

Can machine learning save the NHS?

Individually read the guardian article.

Share your thoughts on the article with the rest of the group.

Do you think this is significant and important?



Support The Guardian
Available for everyone, funded by readers
Contribute → Subscribe →
Search jobs Dating Sign in Search The Guardian UK edition
News Opinion Sport Culture Lifestyle More
UK World Business Football UK politics Environment Education Society Science Tech Global development Cities Obituaries

NHS

This article is more than 2 months old

Hospital develops AI to identify patients likely to skip appointments

Exclusive: London's UCLH creates tool predicting 90% of no-shows - potentially saving NHS millions

Hannah Devlin *Science correspondent*
@hannahdev
Fri 12 Apr 2019 12.35 BST
260

▲ Even an imprecise indication of which patients will attend could save hospitals vast sums of money and help cut waiting times. Photograph: Hannah Mckay

A leading hospital has developed artificial intelligence to predict which patients are most likely to miss appointments.

University College Hospital in London created an algorithm using records from 22,000 appointments for MRI scans, allowing it to identify 90% of those patients who would turn out to be no-shows. The machine intelligence is not perfect - it also incorrectly flags about half of patients attending appointments as being at risk of not showing.

However, even an imprecise indication of which patients will attend could save hospitals vast sums of money and help cut waiting times.

"On average we estimate this could save £2-3 per appointment," said Parashkev Nachev, a consultant neurologist at UCLH, who helped develop the tool. "Given that a big hospital could have nearly a million scheduled events per year, that

om/uk-news"

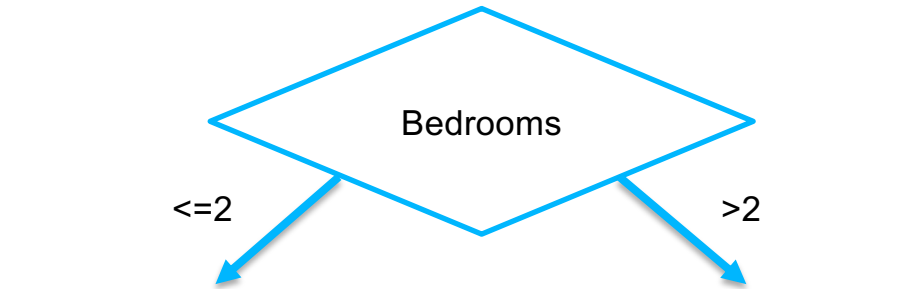
Editorially independent, open to everyone
We chose a different approach – will you support it?
Support The Guardian →

most viewed
My partner is boring and I've fallen for an older, married guy at work
Boris Johnson 'not bluffing' about quitting EU on 31 October with no deal
Labour would break up Treasury and create northern No 11, says McDonnell
'We're not the protesting kind': Remain alliance stages a quiet revolution in Brecon
Jeffrey Epstein charged with sex trafficking, reports say

University College Hospital in London created an algorithm using records from 22,000 appointments for MRI scans, allowing it to identify 90% of those patients who would turn out to be no-shows. The machine intelligence is not perfect – it also incorrectly flags about half of patients attending appointments as being at risk of not showing.

Guardian UK

Building up



NY

Correct: 17
Wrong: 3

SF

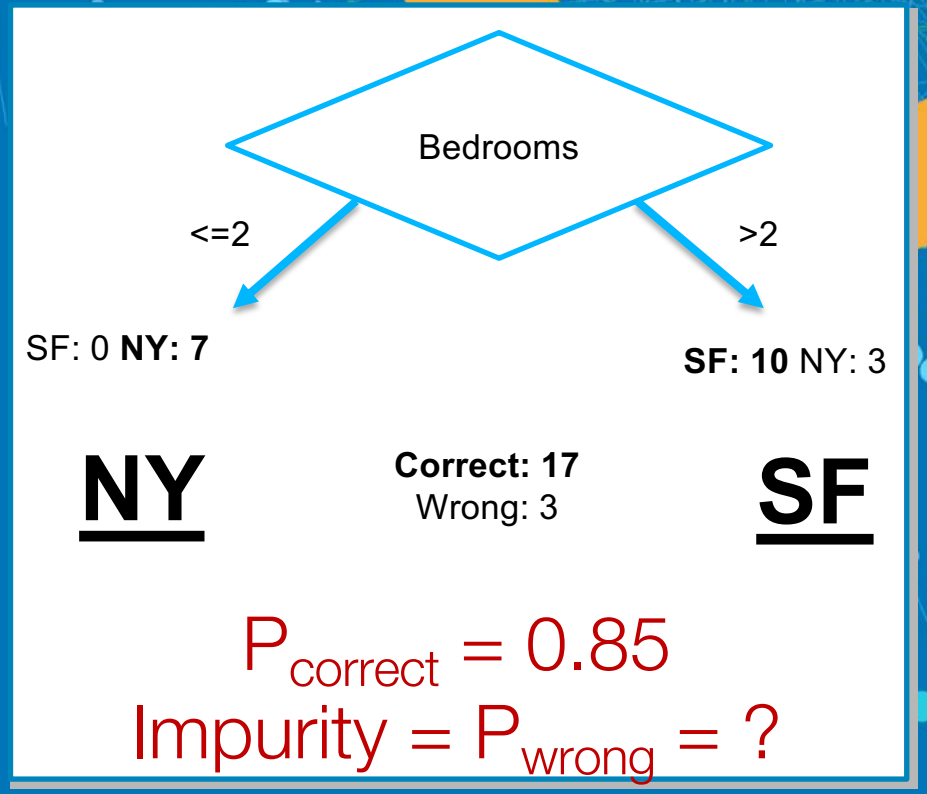
85%
Confidence

Terminology overload

Warning: This session is an overload of terminology.

We will give you access to slides, but it is best if you also note down the terminology as a group on a piece of A3.

Gini Impurity



Gini Impurity

A measurement of the likelihood of an incorrect classification of a new instance of a random variable.

Entropy

As it relates to machine learning, is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.

Flipping coins

Take an unbiased coin.

Work out the entropy:

P_{head} = ?

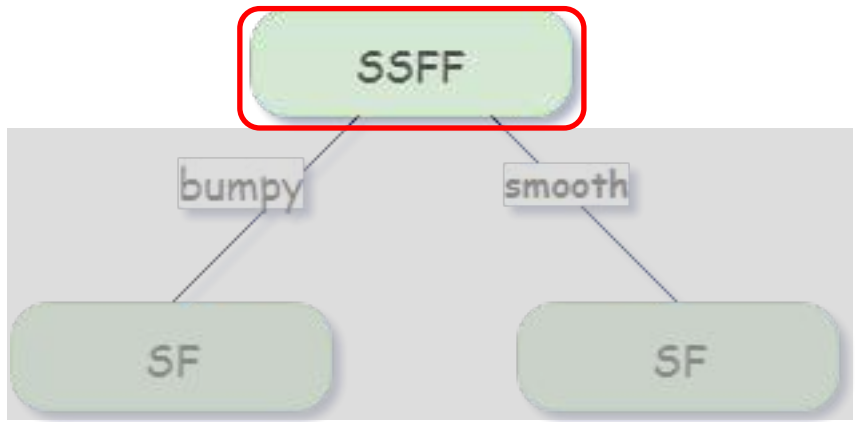
P_{tail} = ?

$$entropy = - P_{head} \log_2(P_{head}) + P_{tail} \log_2(P_{tail})$$

In excel = -1 * (P_{head} * log(P_{head},2) + P_{tail} * log(P_{tail},2))

Speed of cars is influenced by what?

Road condition



S = Slow
F = Fast

Parent	Left Child	Right Child
$P_{fast} = ?$ $P_{slow} = ?$	$P_{fast} = ?$ $P_{slow} = ?$	$P_{fast} = ?$ $P_{slow} = ?$
$Entropy_{parent} = ?$	$Entropy_{leftchild} = ?$	$Entropy_{rightchild} = ?$

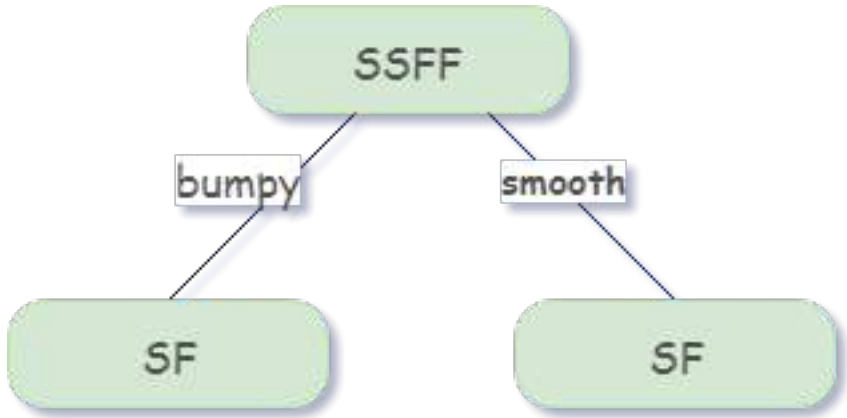
Weighted average

$Entropy_{parent} = 1$

$Entropy_{leftchild} = 1$

$Entropy_{rightchild} = 1$

Average Child Entropy = ?

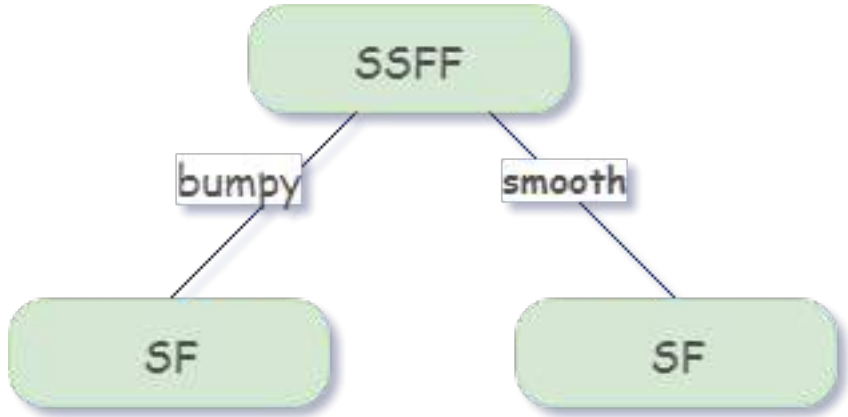


Information gain = $Entropy_{parent}$ - Average Child Entropy

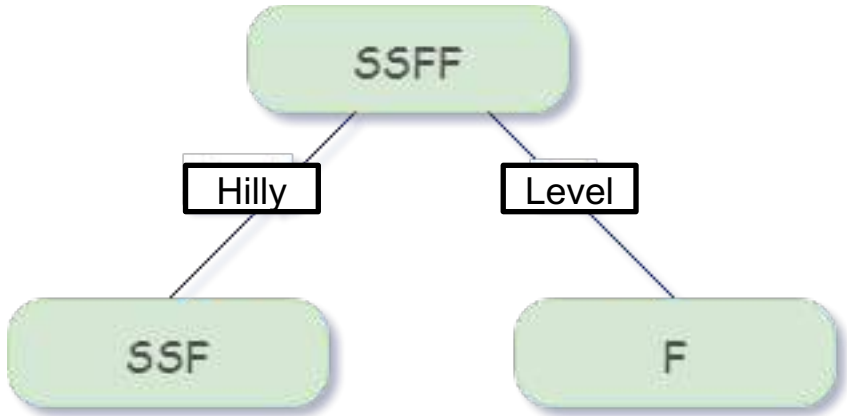
Speed of cars is influenced by what?

S = Slow
F = Fast

Road condition
Information gain = 0



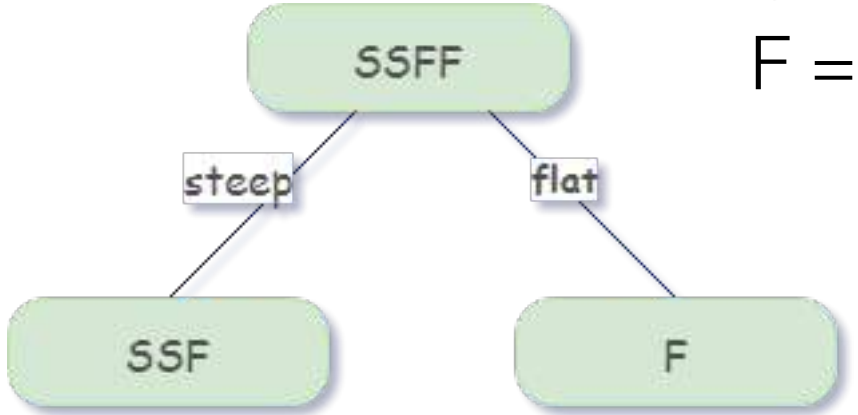
Gradient



Speed of cars is influenced by what?

Gradient

S = Slow
F = Fast



$$entropy = - P_{fast} log_2(P_{fast}) + P_{slow} log_2(P_{slow})$$

Parent	Left Child	Right Child
$P_{fast} = 0.5$	$P_{fast} = ?$	$P_{fast} = ?$
$P_{slow} = 0.5$	$P_{slow} = ?$	$P_{slow} = ?$
$Entropy_{parent} = 1$	$Entropy_{leftchild} = ?$	$Entropy_{rightchild} = ?$

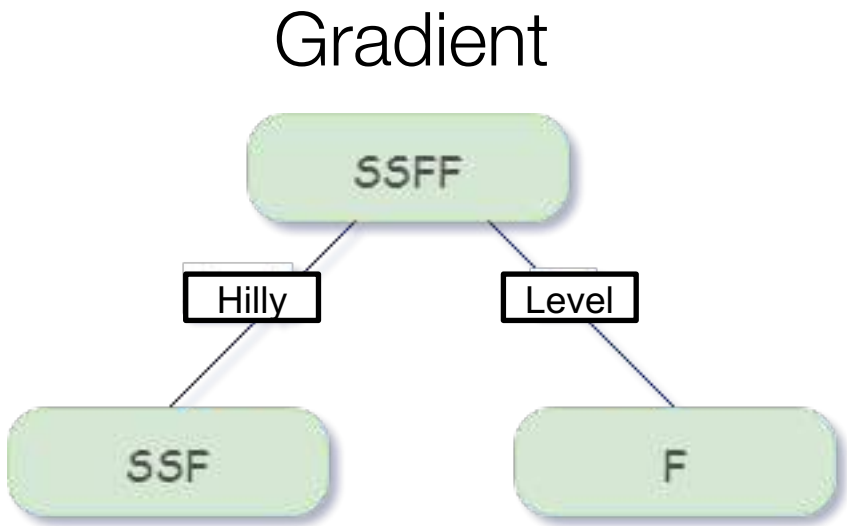
Weighted average

$Entropy_{parent} = 1$

$Entropy_{leftchild} = 0.92$

$Entropy_{rightchild} = 0$

Average Child Entropy = ?



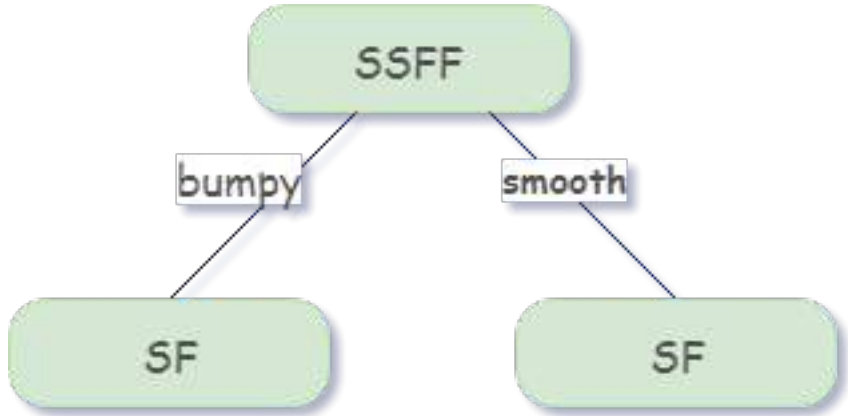
$[weighted_{avg}](children) = ¾\ Entropy_{leftchild} + ¼\ Entropy_{rightchild}$

Information gain = Entropy(parent) - $[weighted_{avg}](children)$

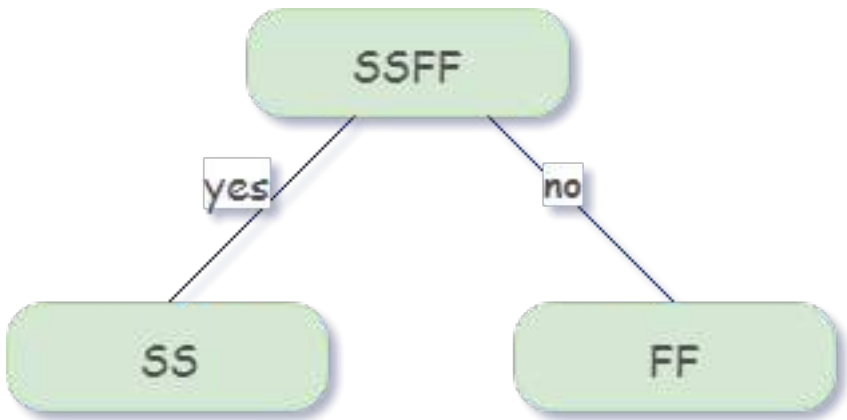
Speed of cars is influenced by what?

S = Slow
F = Fast

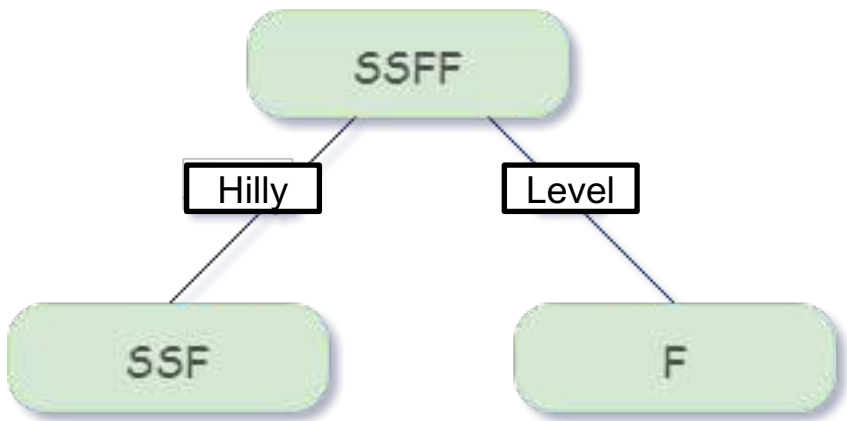
Road condition
Information gain = 0



Speed limit?
Information Gain = ?



Gradient
Information gain = 0.31



Information gain

A measure of decrease of “uncertainty” of the result.

Specifically of each feature in a decision tree.

Which is better?

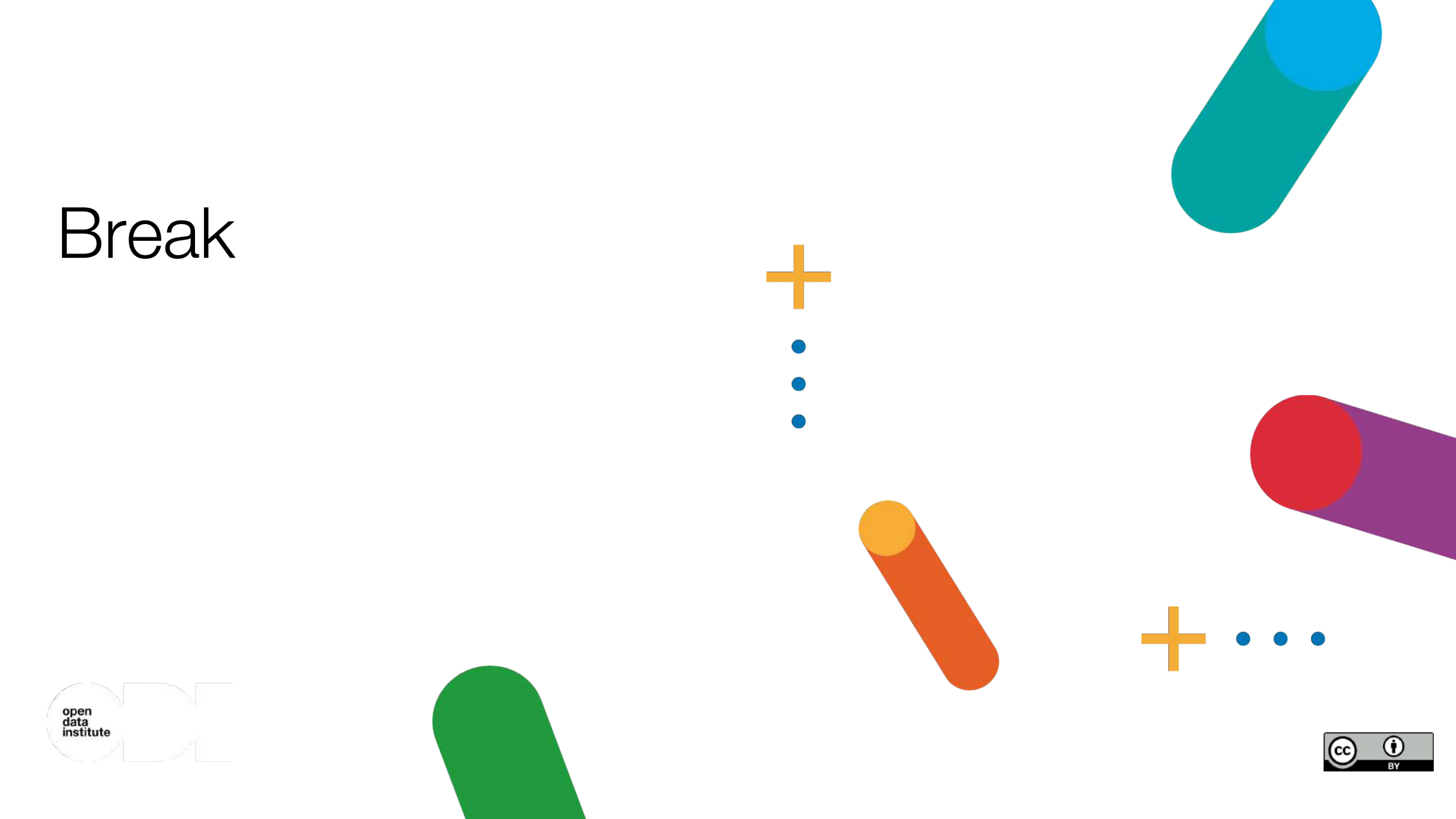
Elevation Threshold = 27

Price per sqft Threshold = 1000

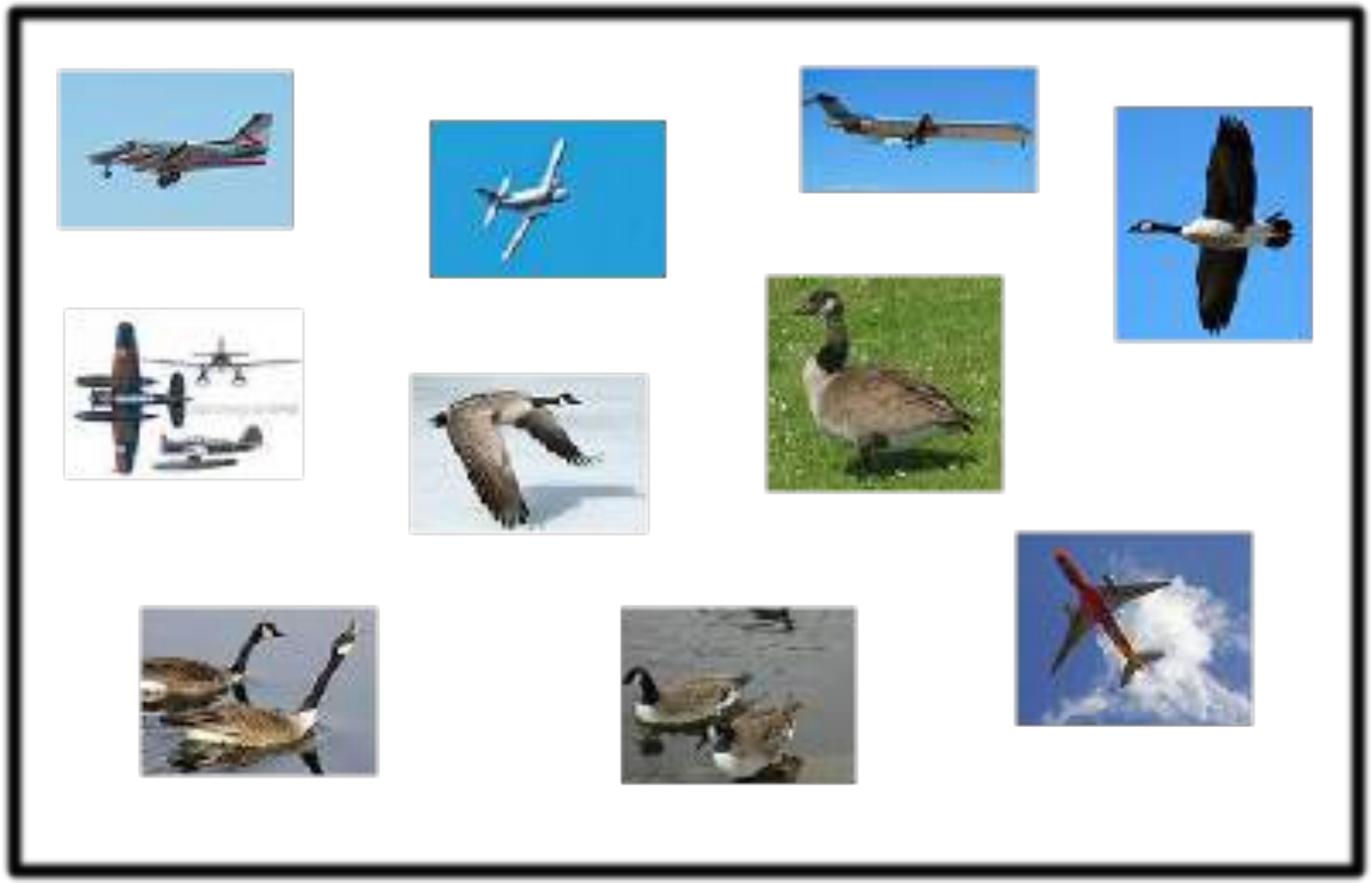
$$entropy = - P_{SF} \log_2(P_{SF}) + P_{NY} \log_2(P_{NY})$$

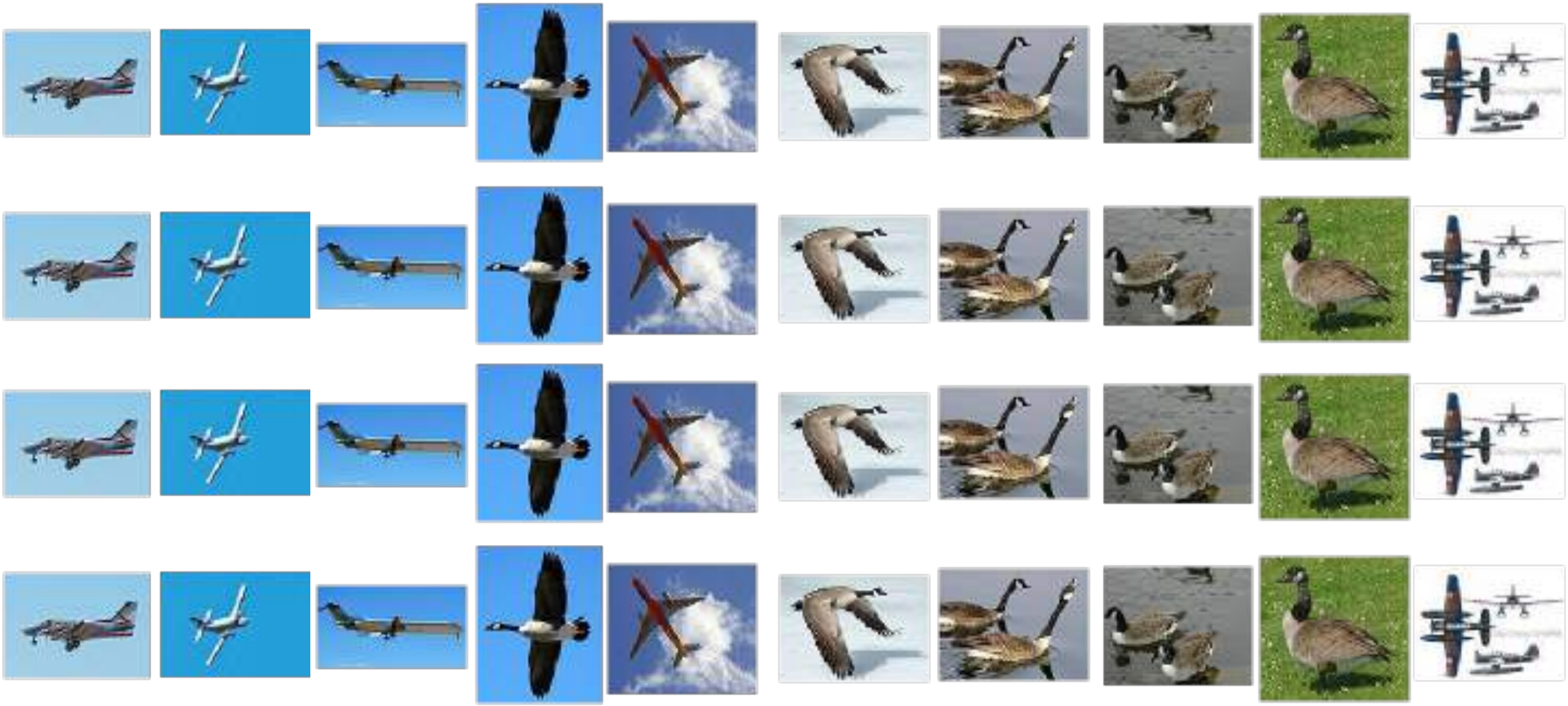
$$\text{Information gain} = \text{Entropy(parent)} - [\text{weighted}_{avg}](\text{children})$$

Break

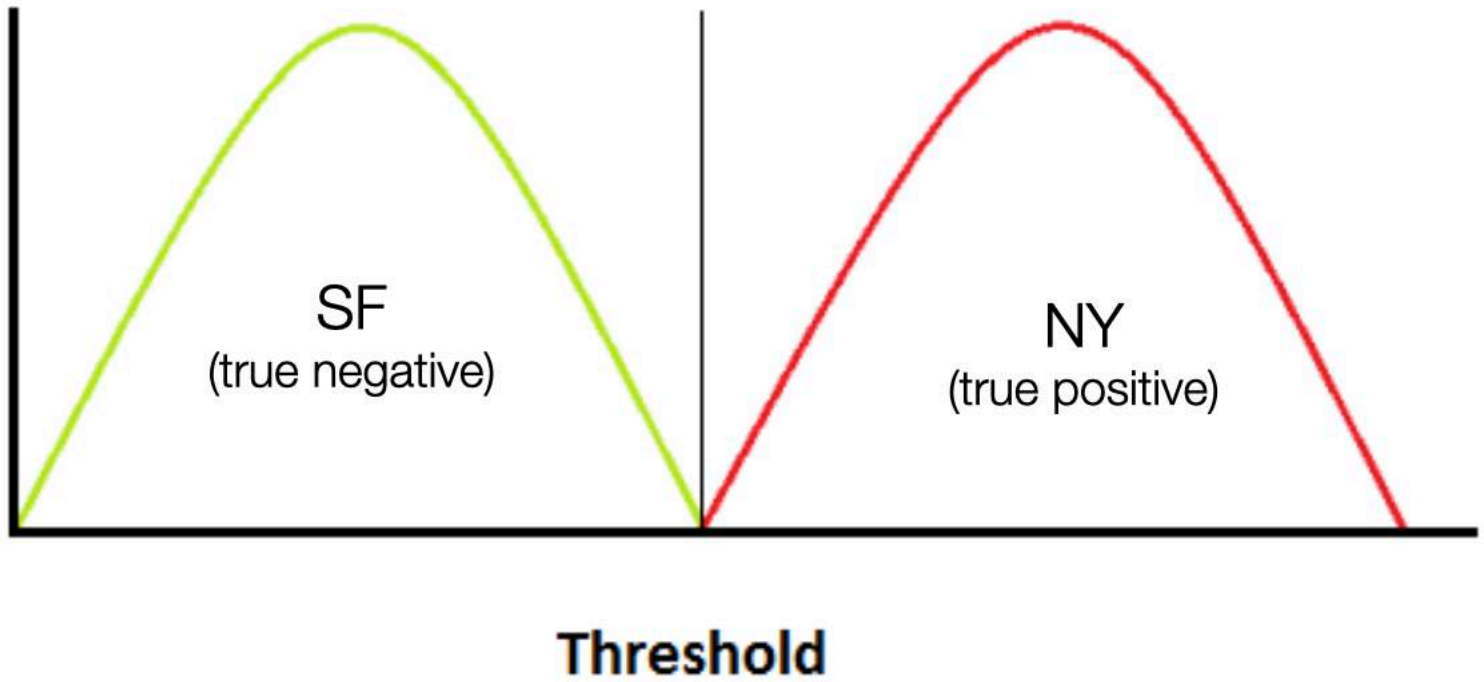


Put these in order

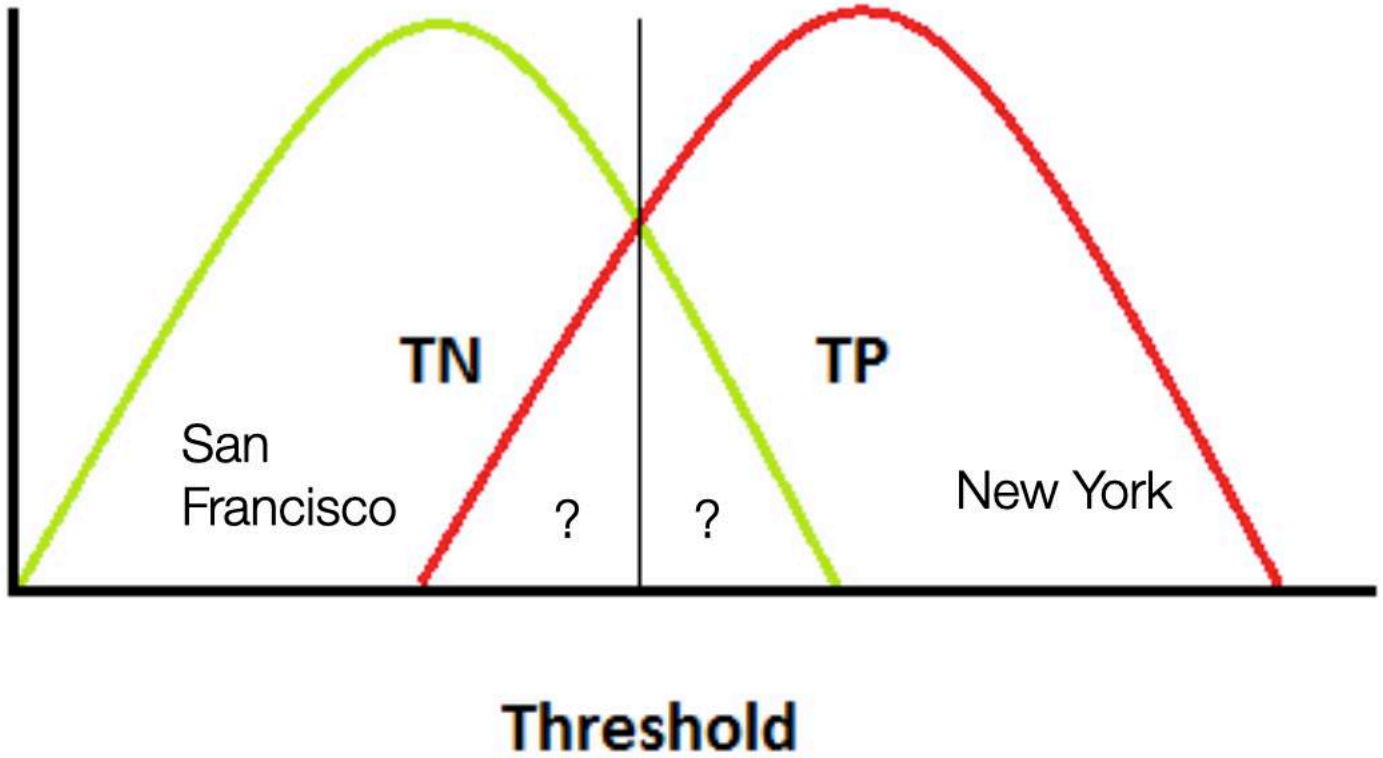




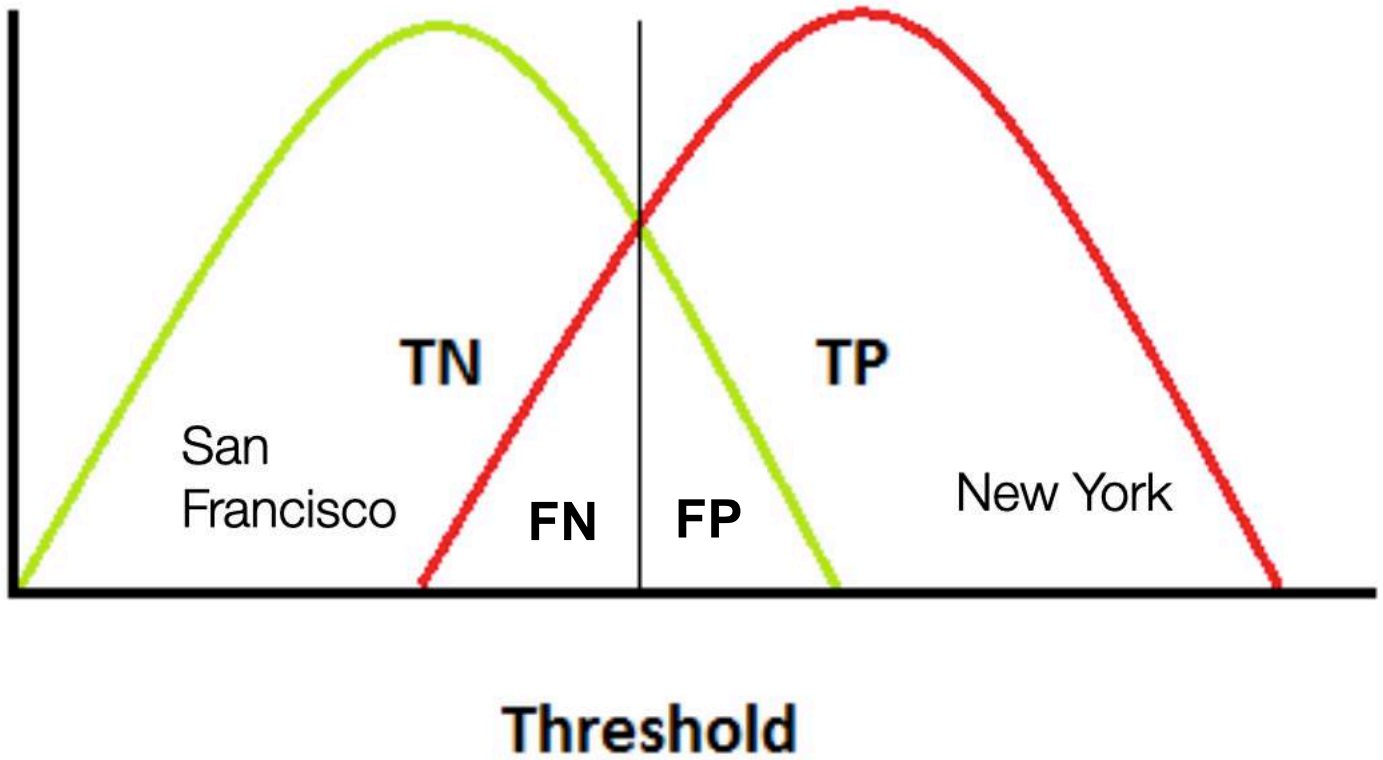
True Positives and True Negatives



True Positives and True Negatives



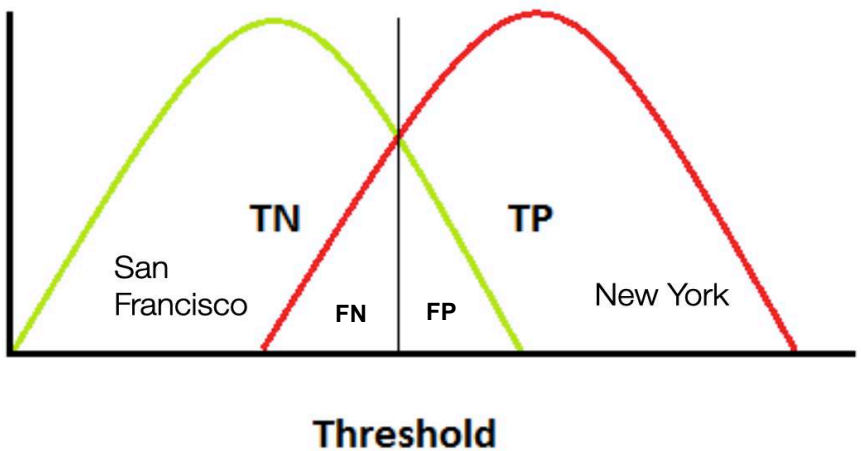
True Positives and True Negatives



Precision and recall

$$recall = \frac{true\ positives}{true\ positives\ +\ false\ negatives}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$



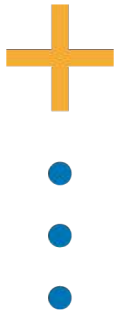
Precision

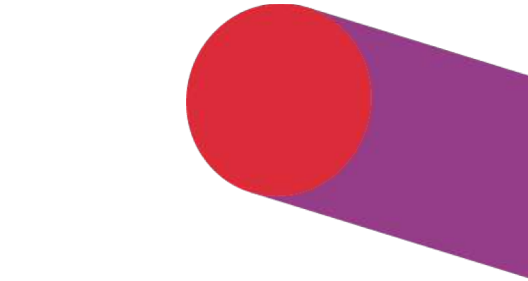
aka. Specificity

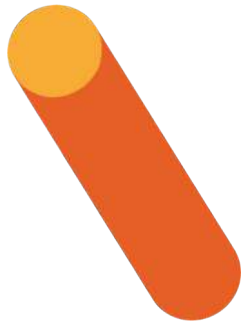
Recall

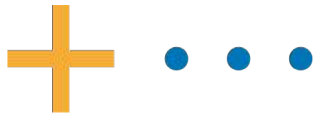
aka. True positive rate

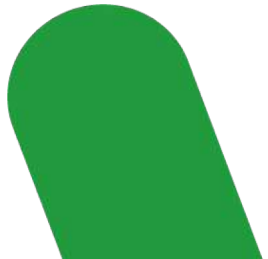
aka. Sensitive









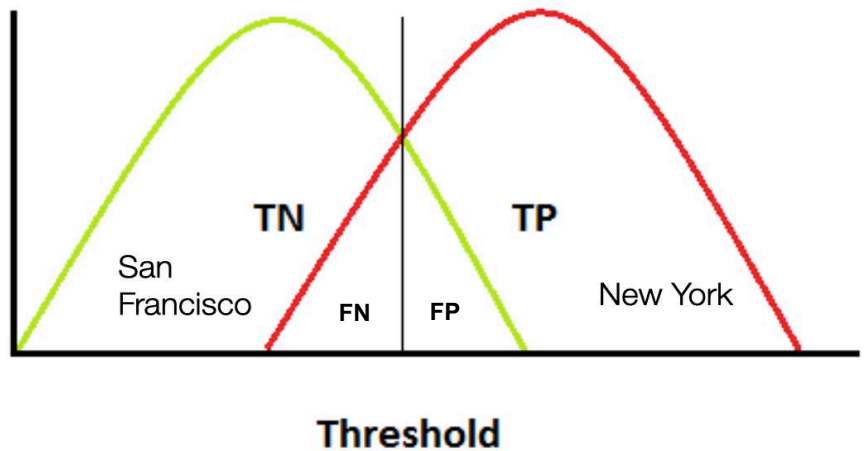


What is the FP?

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$fpr = ?$$

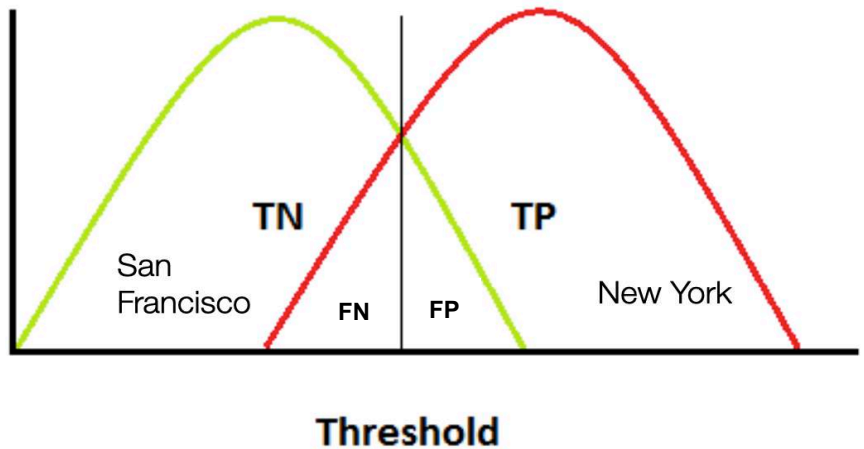


What is the FPR

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

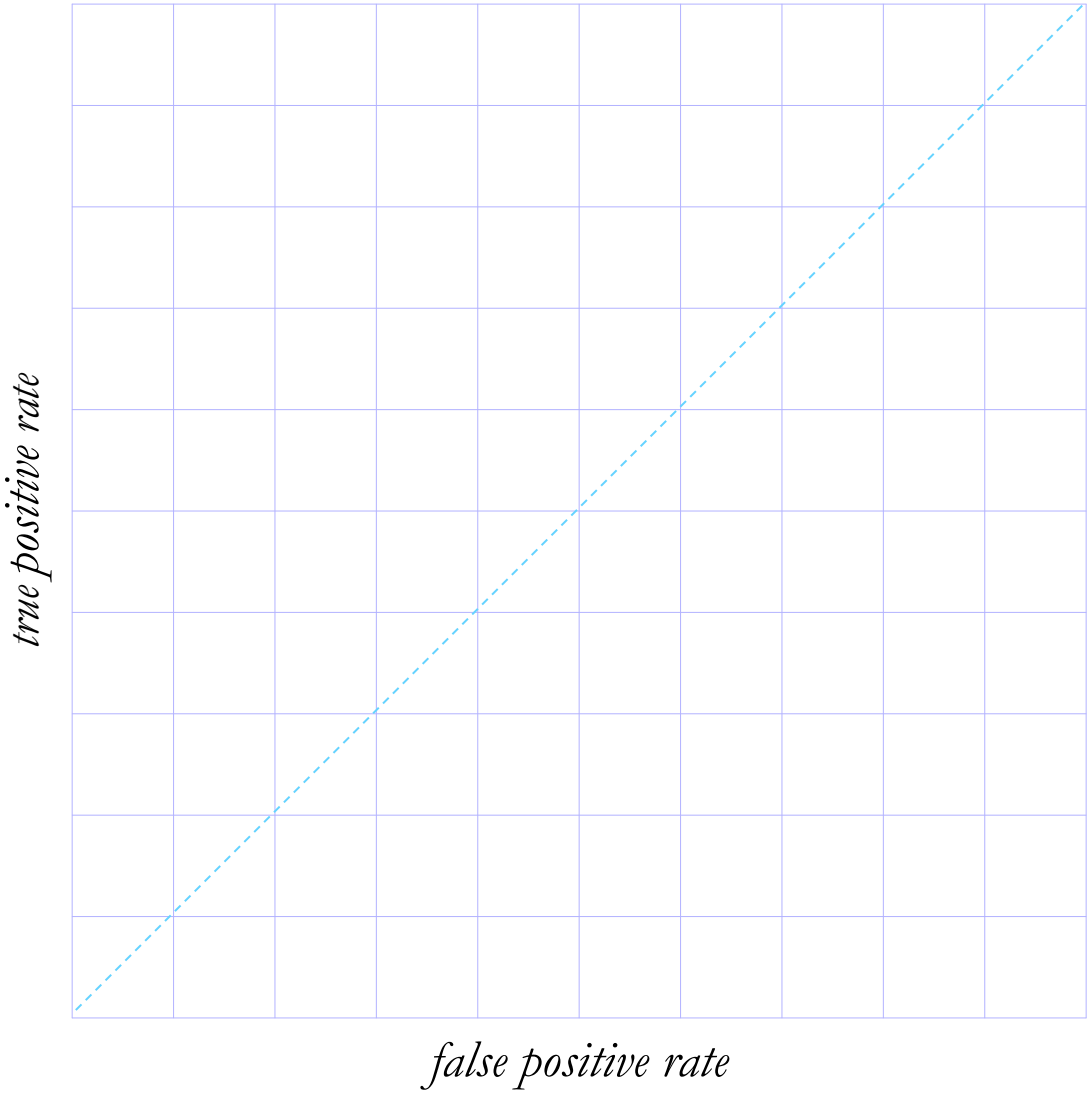
$$fpr = \frac{false\ positives}{false\ positives + true\ negatives}$$

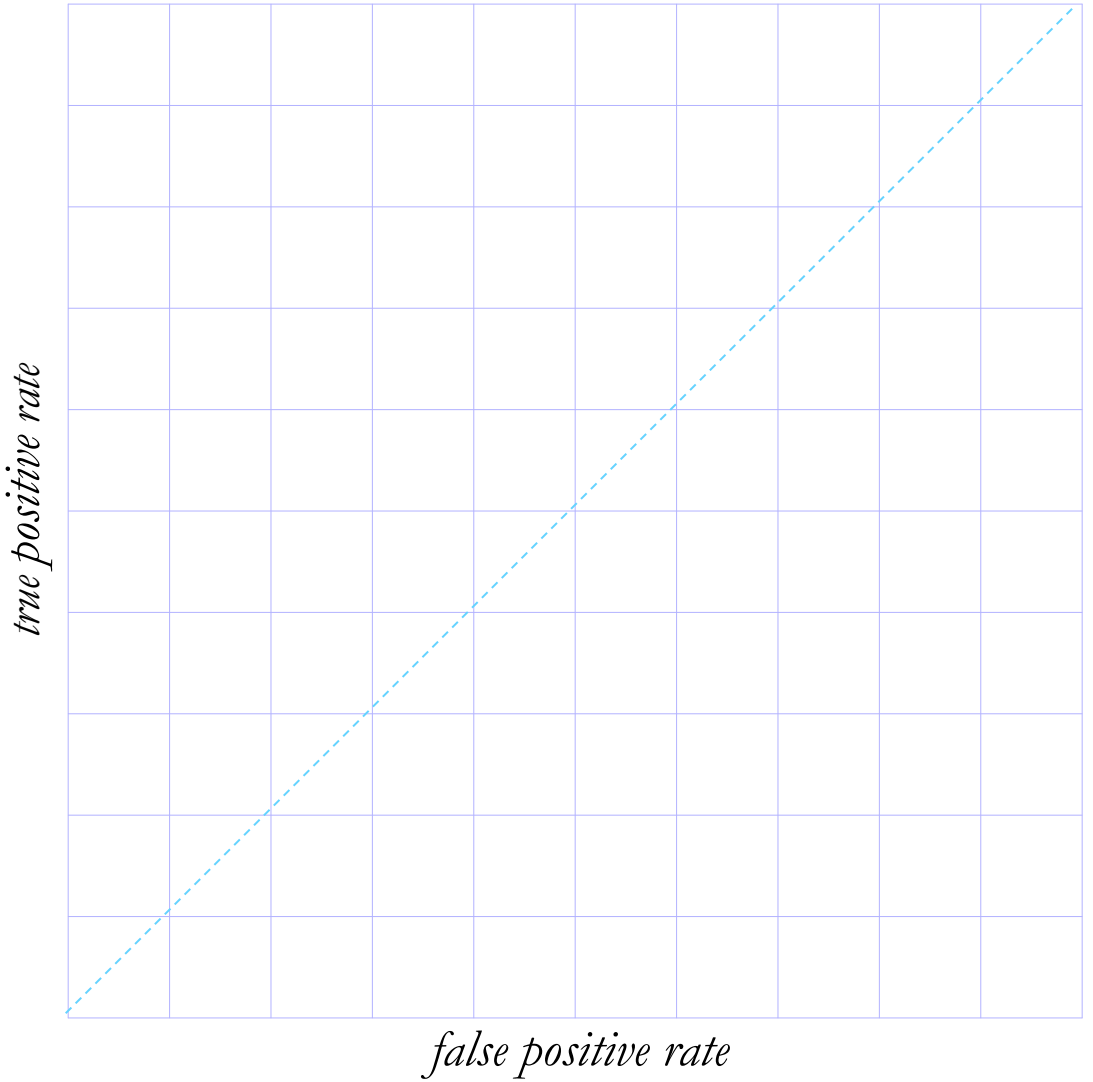


Precision and Recall table

	Below Threshold		Above Threshold						
	Planes	Birds	Planes	Birds					
Threshold					Precision	Recall	TPR	FPR	F_1 Score
First image									
First 2									
First 3									
First 4									
First 5									
First 6									
First 7									
First 8									
First 9									
First 10									

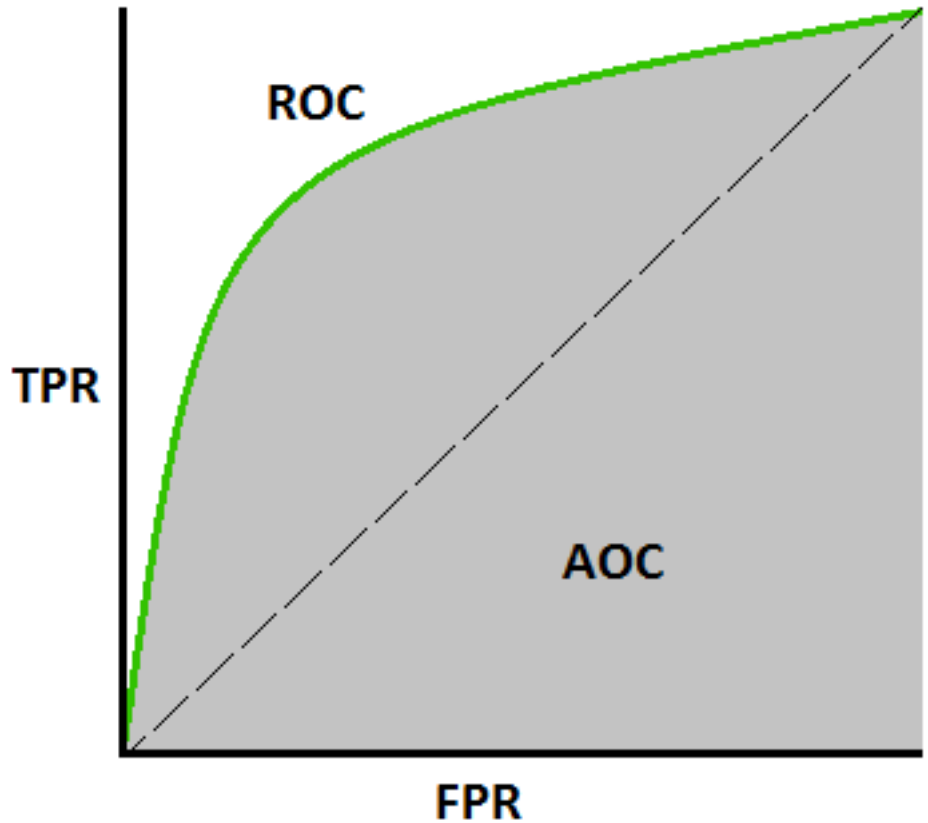
Roc Curves


true positive rate
false positive rate


true positive rate
false positive rate

Plot the ROC? (Can’t plot the ROC?)

Plot FPR against TPR



Calculating F1 score

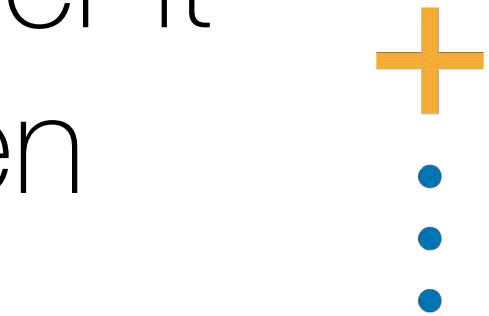
$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

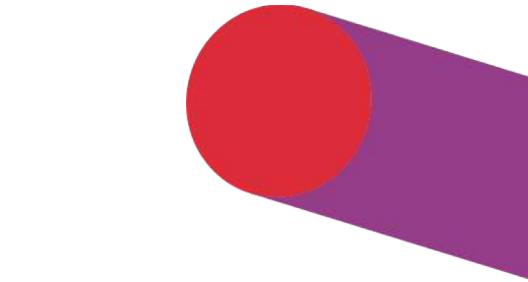
$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

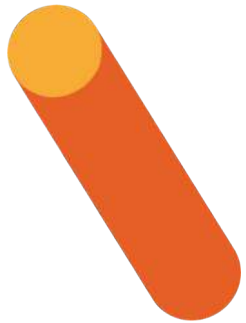
$$fpr = \frac{false\ positives}{false\ positives + true\ negatives}$$

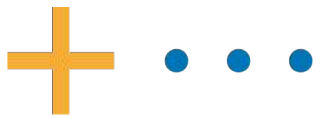
$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

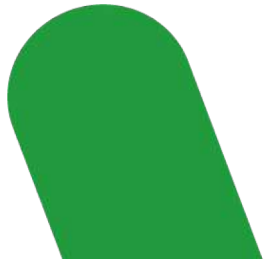
Take a pack of blue cards and order it by your chosen feature.
Make a ROC curve.

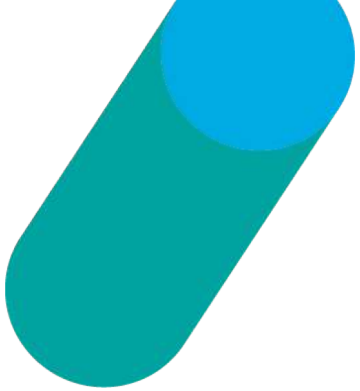








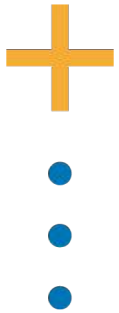


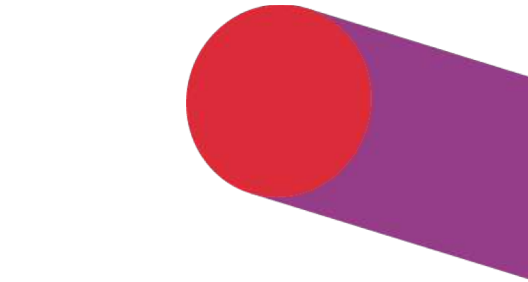


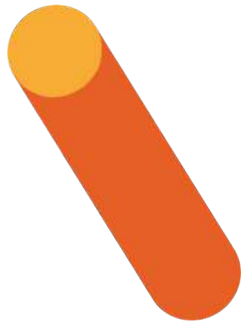
	Below Threshold		Above Threshold						
	NY	SF	NY	SG					
Threshold					Precision	Recall	TPR	FPR	F_1 Score
First card									
First 2									
First 3									
First 4									
First 5									
First 6									
First 7									
First 8									
First 9									
First 10									
First 11									
First 12									
First 13									
First 14									
First 15									
First 16									
First 17									
First 18									
First 19									
All 20									

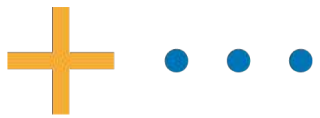
search: EasyROC
download: bit.ly/NYSFBlue

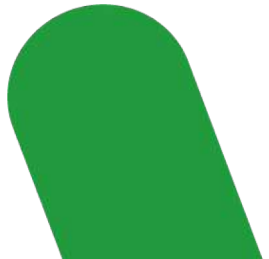
Using EasyROC,
gather the AUC and
information gain for
each feature.
Plot each of these on a
column chart.





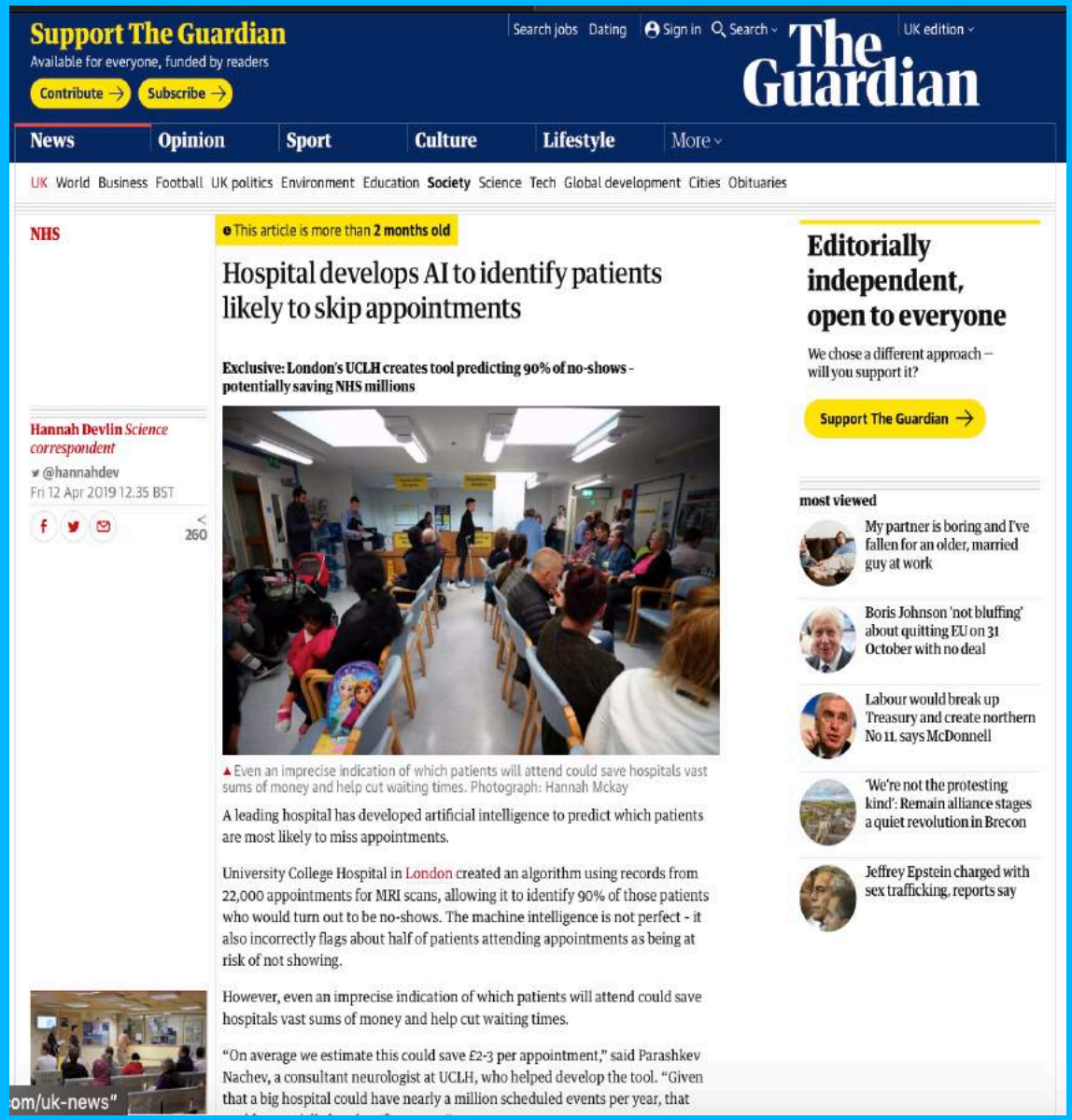






Can machine learning save the NHS?

What to the 90% and 50% relate to?



Support The Guardian
Available for everyone, funded by readers
Contribute → Subscribe →

Search jobs Dating Sign in Search The Guardian UK edition

News Opinion Sport Culture Lifestyle More

UK World Business Football UK politics Environment Education Society Science Tech Global development Cities Obituaries

NHS

This article is more than 2 months old

Hospital develops AI to identify patients likely to skip appointments

Exclusive: London's UCLH creates tool predicting 90% of no-shows - potentially saving NHS millions

Hannah Devlin Science correspondent
@hannahdev
Fri 12 Apr 2019 12.35 BST

260

▲ Even an imprecise indication of which patients will attend could save hospitals vast sums of money and help cut waiting times. Photograph: Hannah Mckay

A leading hospital has developed artificial intelligence to predict which patients are most likely to miss appointments.

University College Hospital in London created an algorithm using records from 22,000 appointments for MRI scans, allowing it to identify 90% of those patients who would turn out to be no-shows. The machine intelligence is not perfect - it also incorrectly flags about half of patients attending appointments as being at risk of not showing.

However, even an imprecise indication of which patients will attend could save hospitals vast sums of money and help cut waiting times.

"On average we estimate this could save £2-3 per appointment," said Parashkev Nachev, a consultant neurologist at UCLH, who helped develop the tool. "Given that a big hospital could have nearly a million scheduled events per year, that

om/uk-news"

Editorially independent, open to everyone

We chose a different approach – will you support it?

Support The Guardian →

most viewed

My partner is boring and I've fallen for an older, married guy at work

Boris Johnson 'not bluffing' about quitting EU on 31 October with no deal

Labour would break up Treasury and create northern No 11, says McDonnell

'We're not the protesting kind': Remain alliance stages a quiet revolution in Brecon

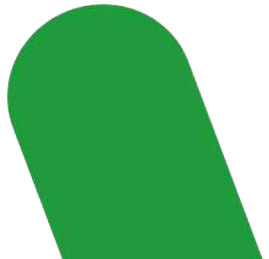
Jeffrey Epstein charged with sex trafficking, reports say

Case study

Is it:

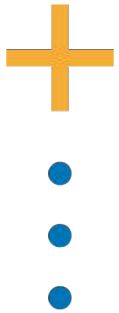
Sound?
Breakthrough?
Significant?
Worth it?

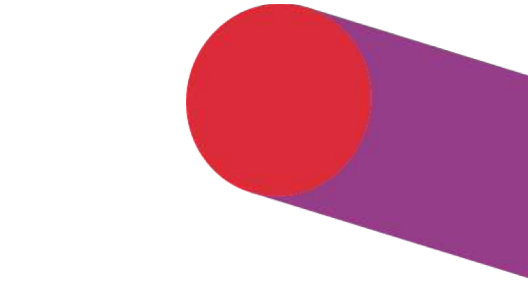
How would you report it?

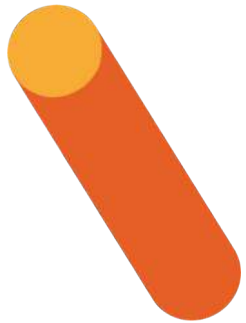


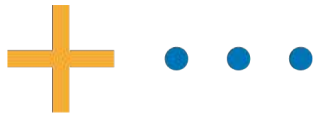
Workshop

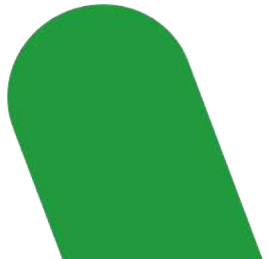
Create a checklist to help guide analysis and reporting of articles involving machine learning











Any questions?

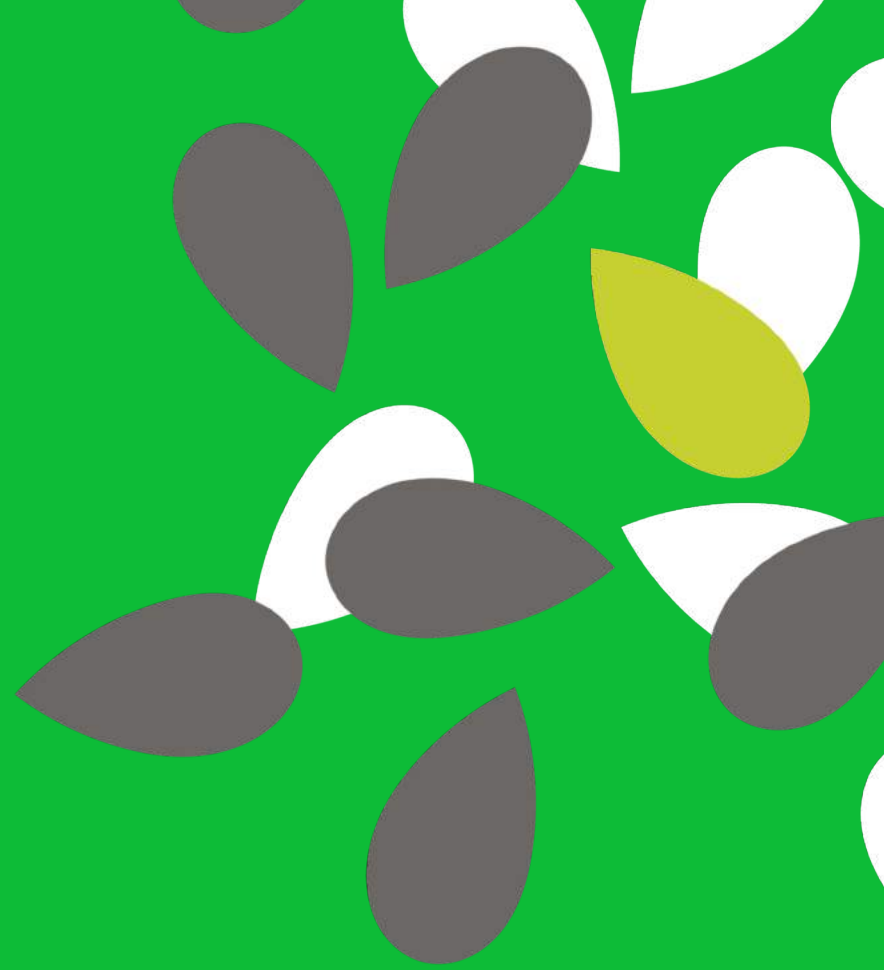
Get in touch

If you would like to talk to us about collaborating, partnering, supporting our work, or anything else, we'd love you to get in touch.

info@theodi.org

+44 (0)20 3598 9395

@ODIHQ



open data institute

Decoded: Machine learning Part 3

Dr David Tarrant
@davetaz
theODI.org

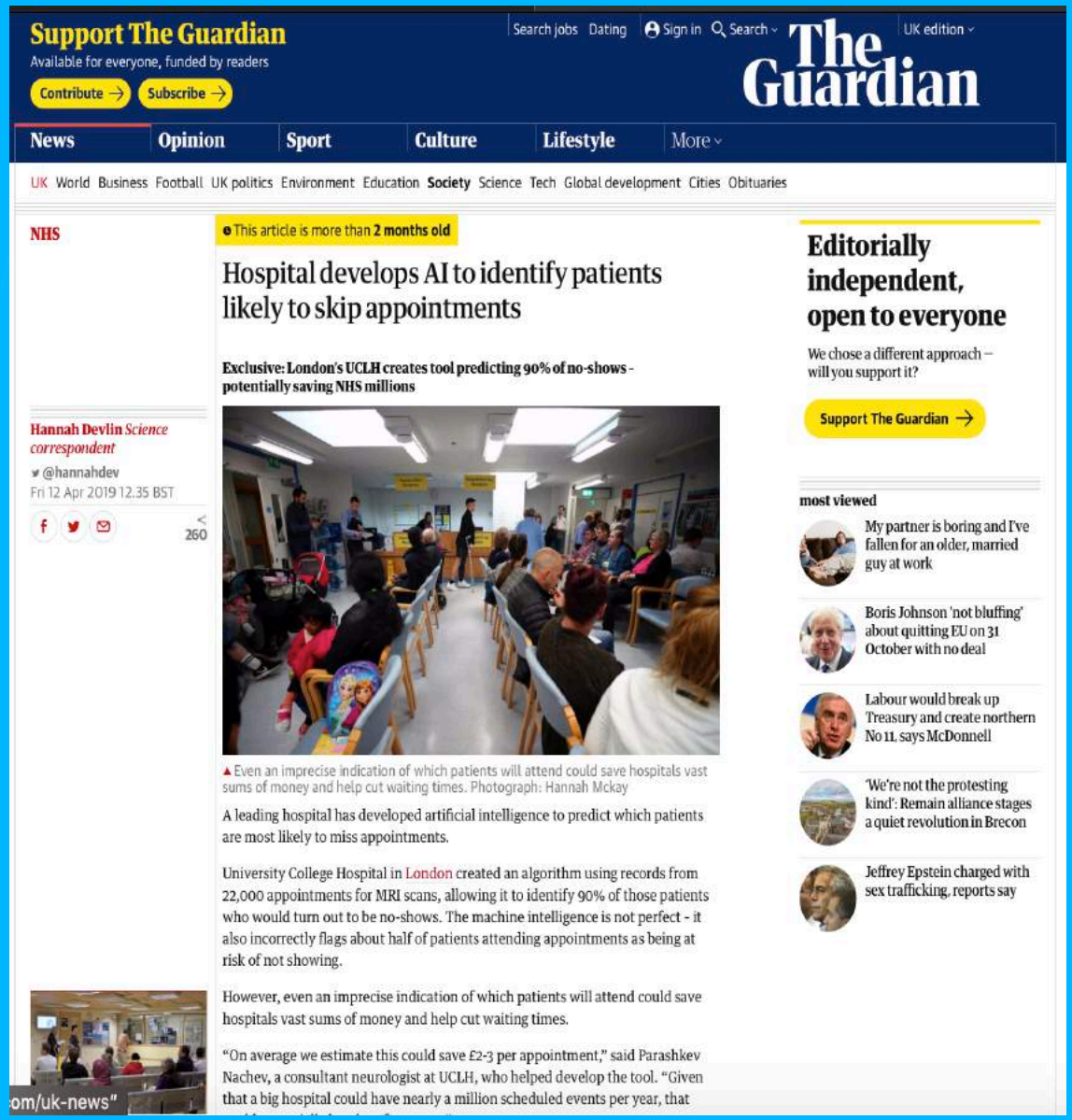
Violeta Mezeklieva

theODI.org

Can machine learning save the NHS?

Is the data sample correct?

How would you collect a sample to discover if the findings are significant outside of the London hospitals?



Support The Guardian
Available for everyone, funded by readers
Contribute → Subscribe →
Search jobs Dating Sign in Search UK edition
The Guardian
News Opinion Sport Culture Lifestyle More
UK World Business Football UK politics Environment Education Society Science Tech Global development Cities Obituaries

NHS

This article is more than 2 months old

Hospital develops AI to identify patients likely to skip appointments

Exclusive: London's UCLH creates tool predicting 90% of no-shows - potentially saving NHS millions

Hannah Devlin Science correspondent
@hannahdev
Fri 12 Apr 2019 12.35 BST
260

▲ Even an imprecise indication of which patients will attend could save hospitals vast sums of money and help cut waiting times. Photograph: Hannah Mckay

A leading hospital has developed artificial intelligence to predict which patients are most likely to miss appointments.

University College Hospital in London created an algorithm using records from 22,000 appointments for MRI scans, allowing it to identify 90% of those patients who would turn out to be no-shows. The machine intelligence is not perfect - it also incorrectly flags about half of patients attending appointments as being at risk of not showing.

However, even an imprecise indication of which patients will attend could save hospitals vast sums of money and help cut waiting times.

"On average we estimate this could save £2-3 per appointment," said Parashkev Nachev, a consultant neurologist at UCLH, who helped develop the tool. "Given that a big hospital could have nearly a million scheduled events per year, that

om/uk-news"

Editorially independent, open to everyone

We chose a different approach – will you support it?

Support The Guardian →

most viewed

My partner is boring and I've fallen for an older, married guy at work

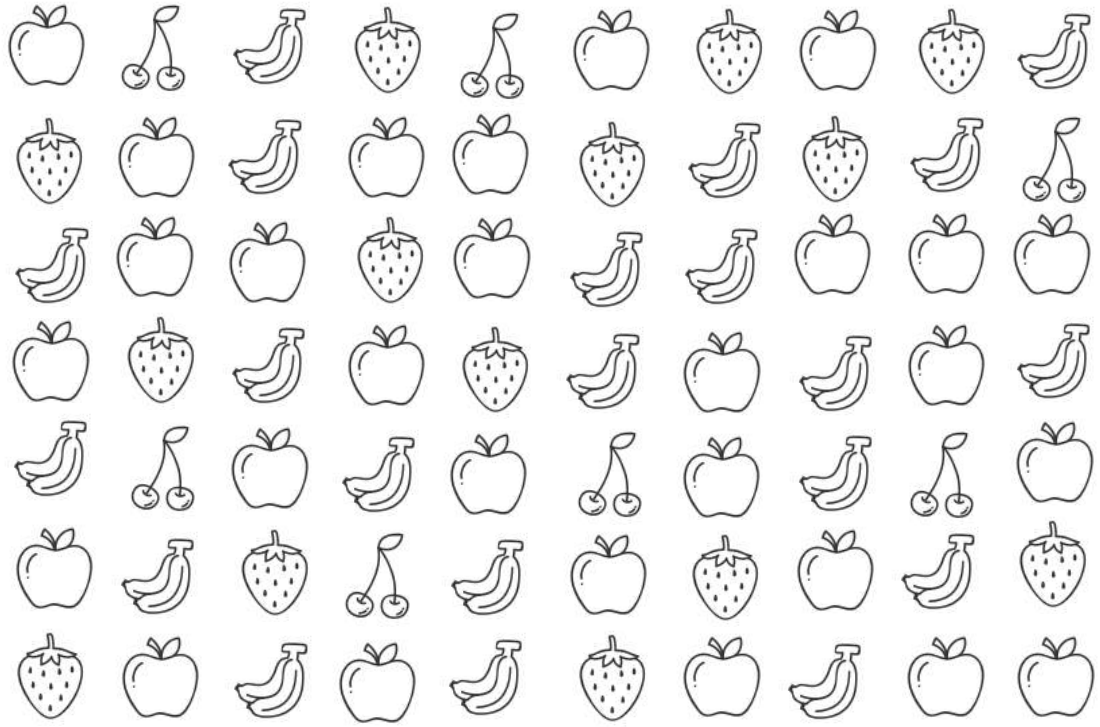
Boris Johnson 'not bluffing' about quitting EU on 31 October with no deal

Labour would break up Treasury and create northern No 11, says McDonnell

'We're not the protesting kind': Remain alliance stages a quiet revolution in Brecon

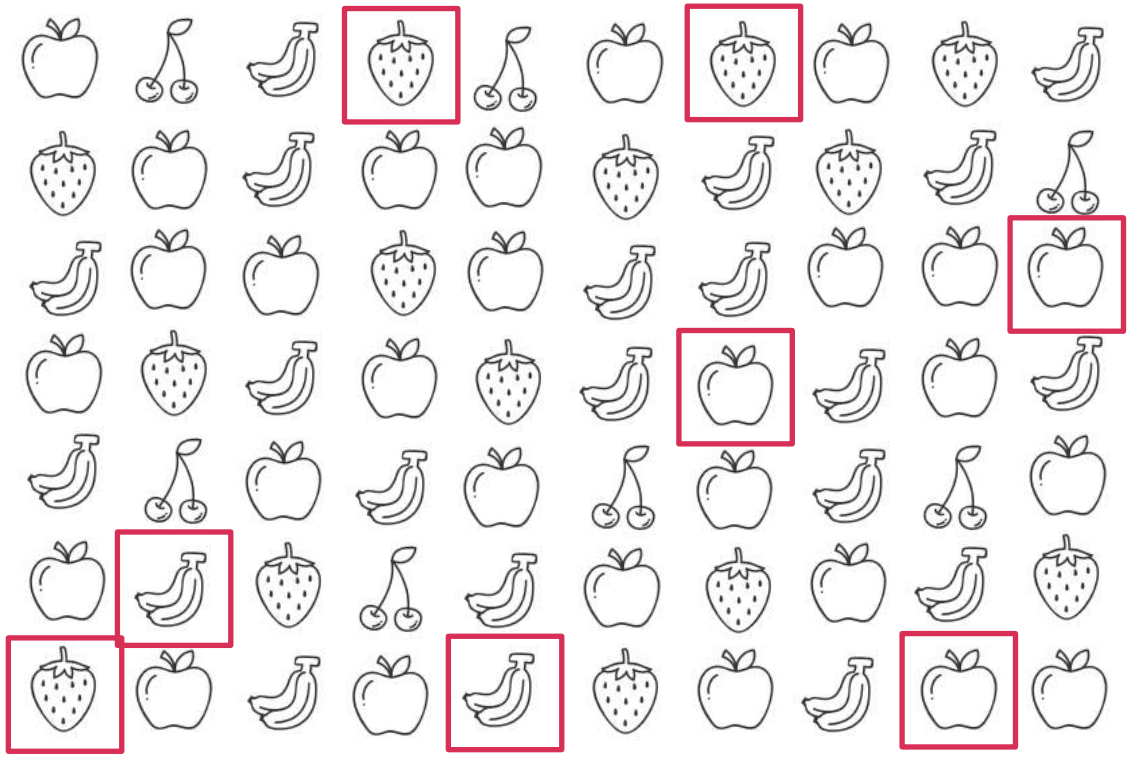
Jeffrey Epstein charged with sex trafficking, reports say

Whole population



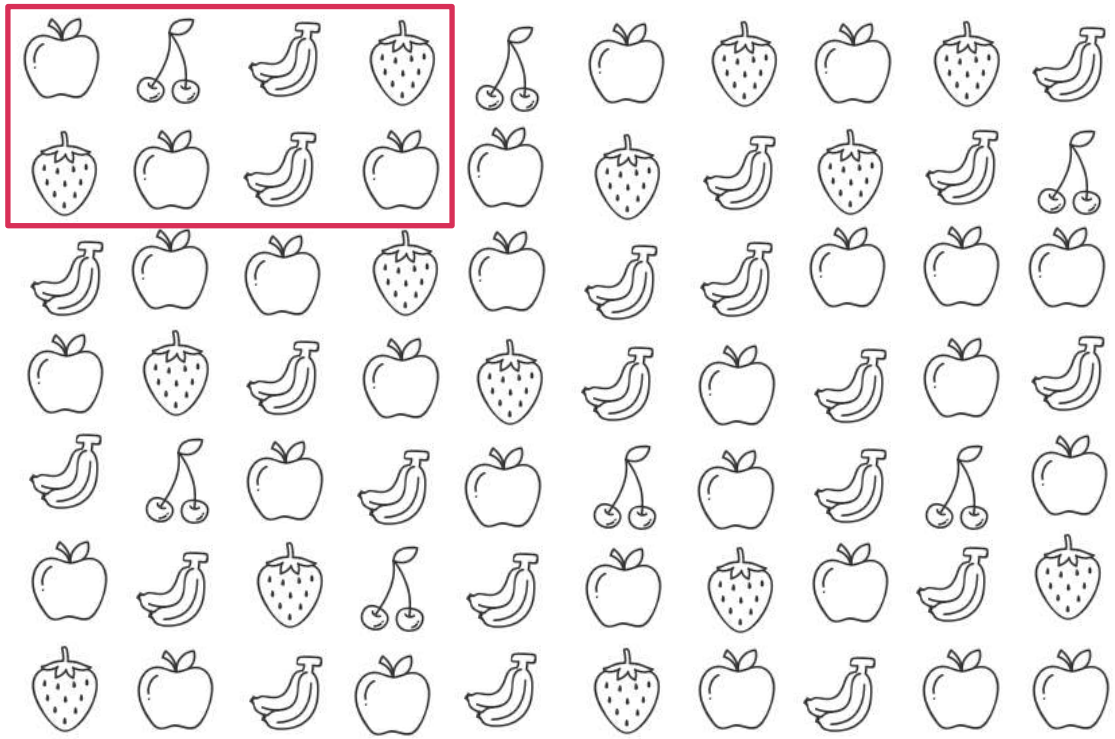
open
data
institute

Random sample



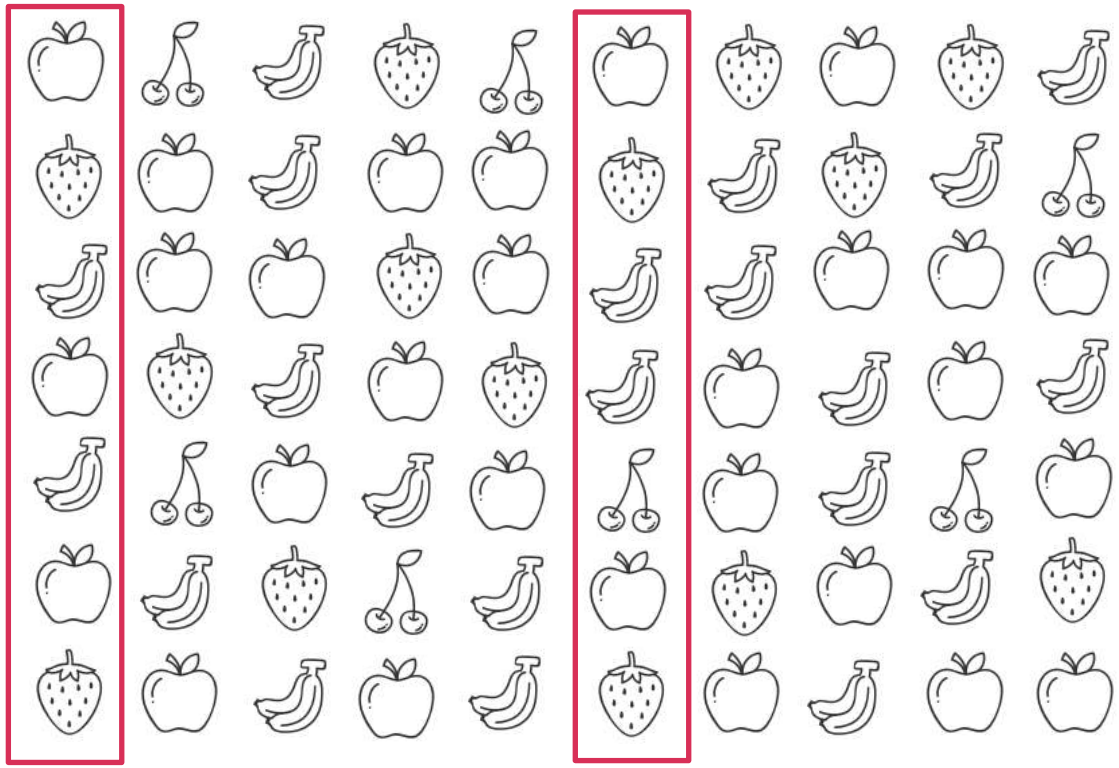
open
data
institute

Convenience



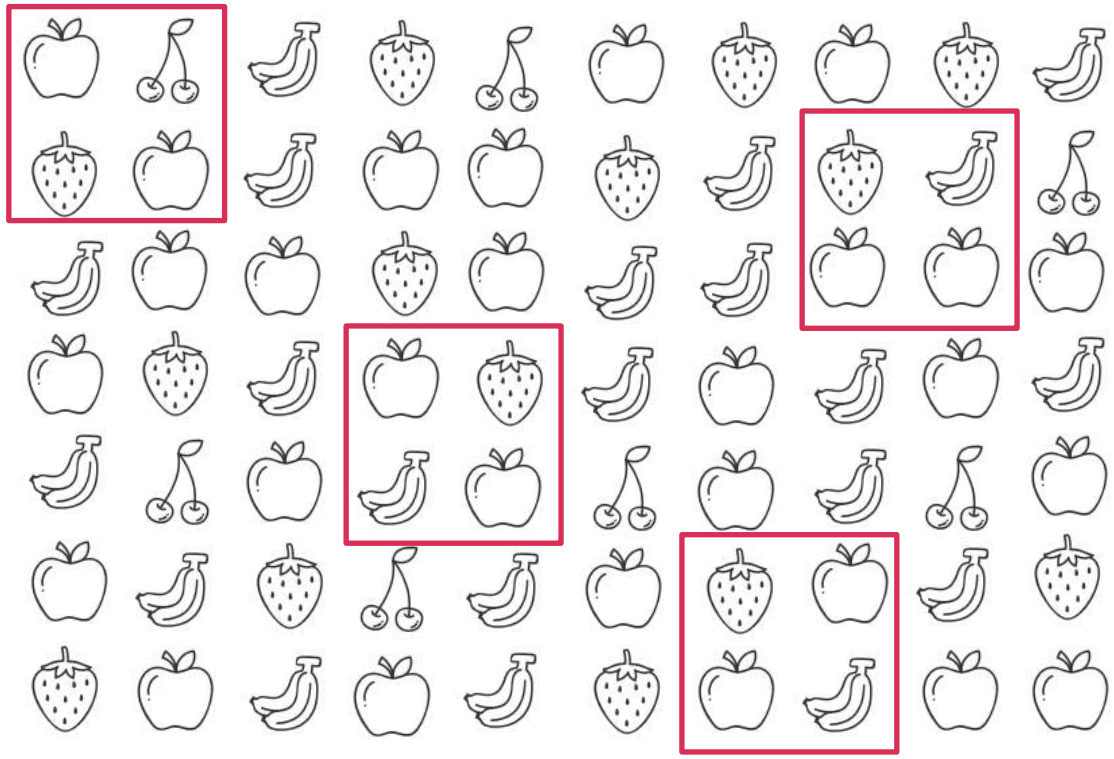
open
data
institute

Systematic



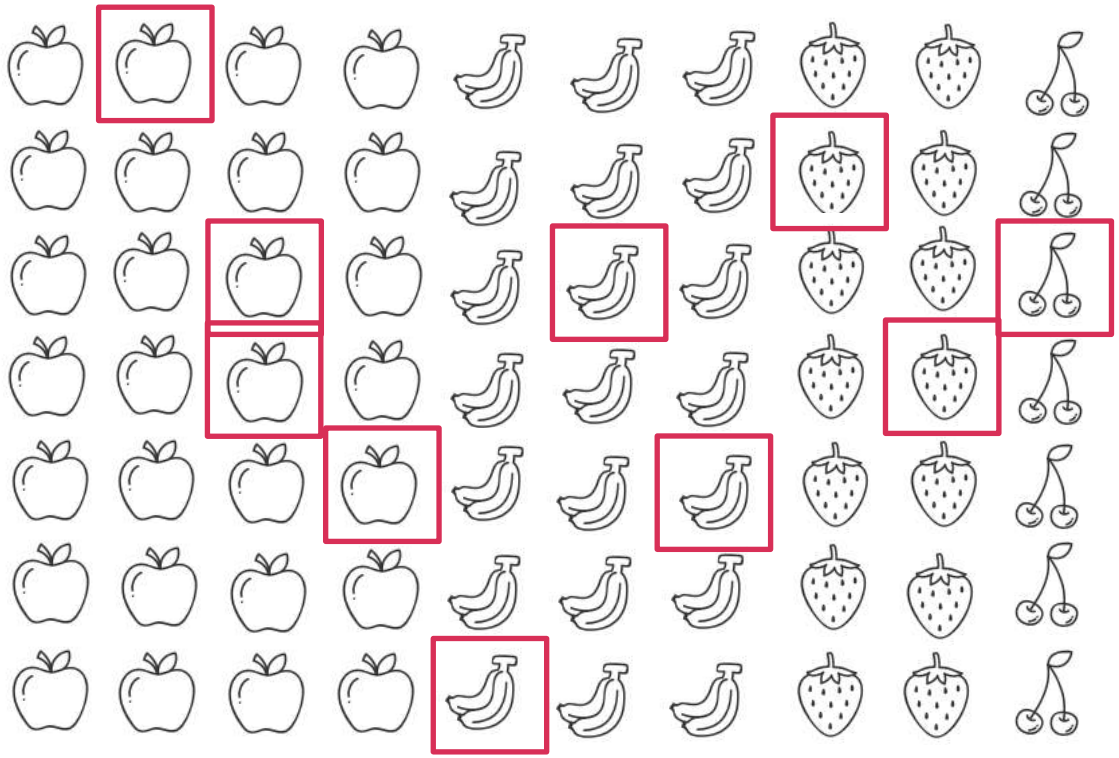
open
data
institute

Clustered



open data institute

Stratified (random)



open data institute

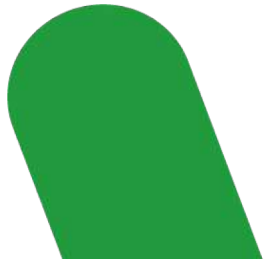
Spectrum of sampling techniques

Image classification

In Chrome

machinelearningforkids.co.uk

Username: **bbcstudent**[1..8]
Password: **fence.ants.tired**



Datasets

https://theodi.github.io/ML102/Car_Cup/index.html

https://theodi.github.io/ML102/faces/gender.html

https://theodi.github.io/ML102/faces/real-fake.html

Any questions?

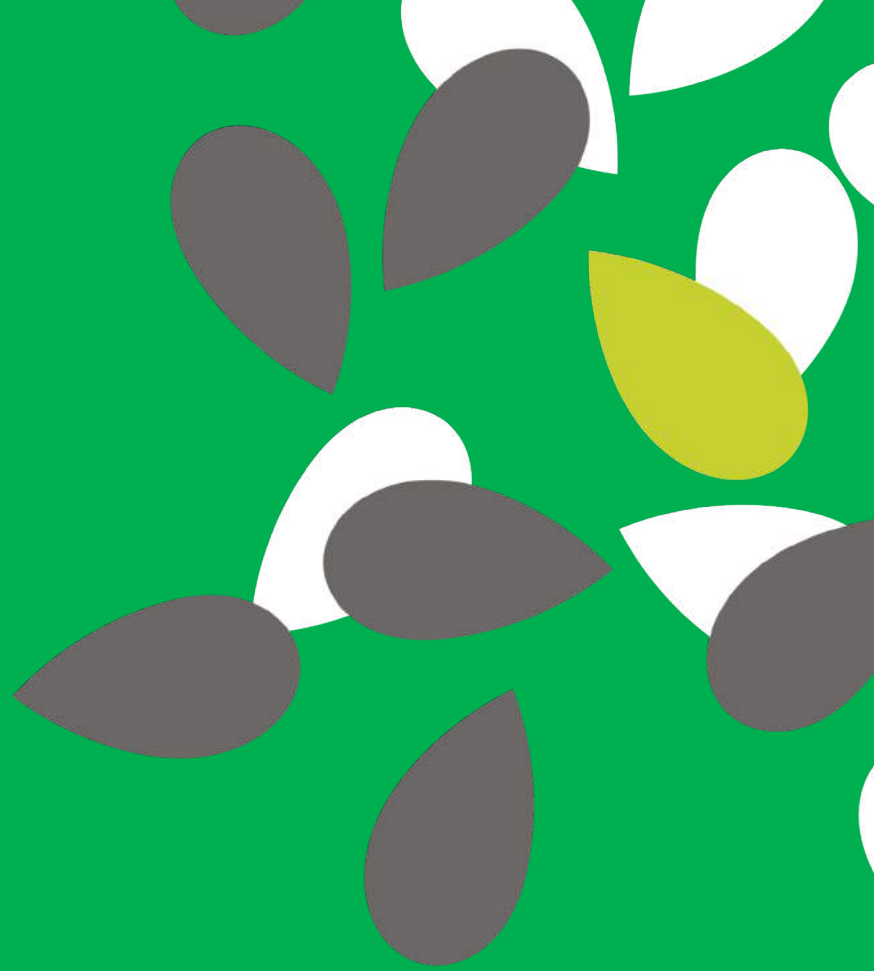
Get in touch

If you would like to talk to us about collaborating, partnering, supporting our work, or anything else, we'd love you to get in touch.

info@theodi.org

+44 (0)20 3598 9395

@ODIHQ



open data institute