Dr. Dave Tarrant
@davetaz – davetaz@theodi.org

Aim

Equip you with the knowledge and skills to work practically and theorectically with data.

Session 1

Statistics for data scientists

Session 2

Machine learning for data scientists

Statistics for Data Scientists

Outcomes

Explain the importance of good statistical methods
Describe a number of key statistical techniques
Apply a number of statistical techniques to data
Use R-Studio to explore data through applied statistics

A look back

A look back

19th century: large data sets, simple questions

20th century: small data sets, complex questions

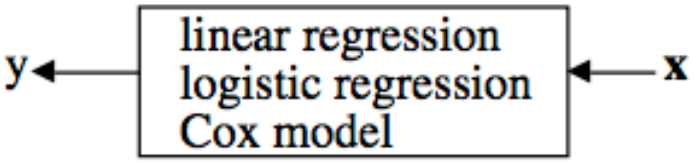
21th century: large data sets, complex questions

Statistical Modeling: The Two Cultures

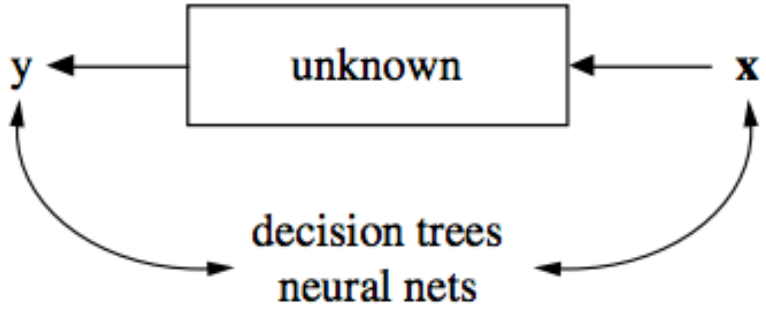
Leo Breiman:

“There are two cultures in the use of statistical modeling to reach conclusions from data.”

The *Two Cultures*



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

(Tukey spoke of algorithmic and algebraic models.)

Stereotypes

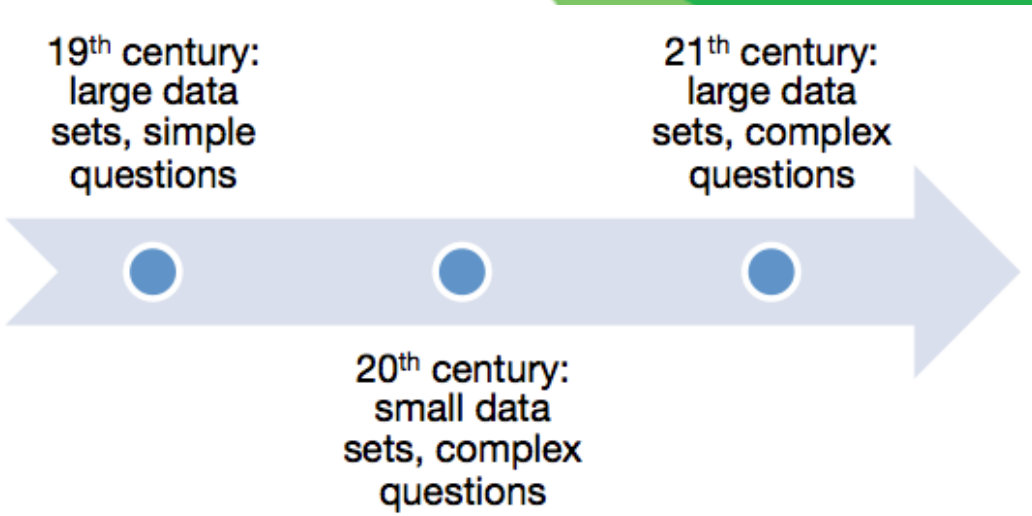
Statistics	Data Science
Causal	Predictive

Stereotypes

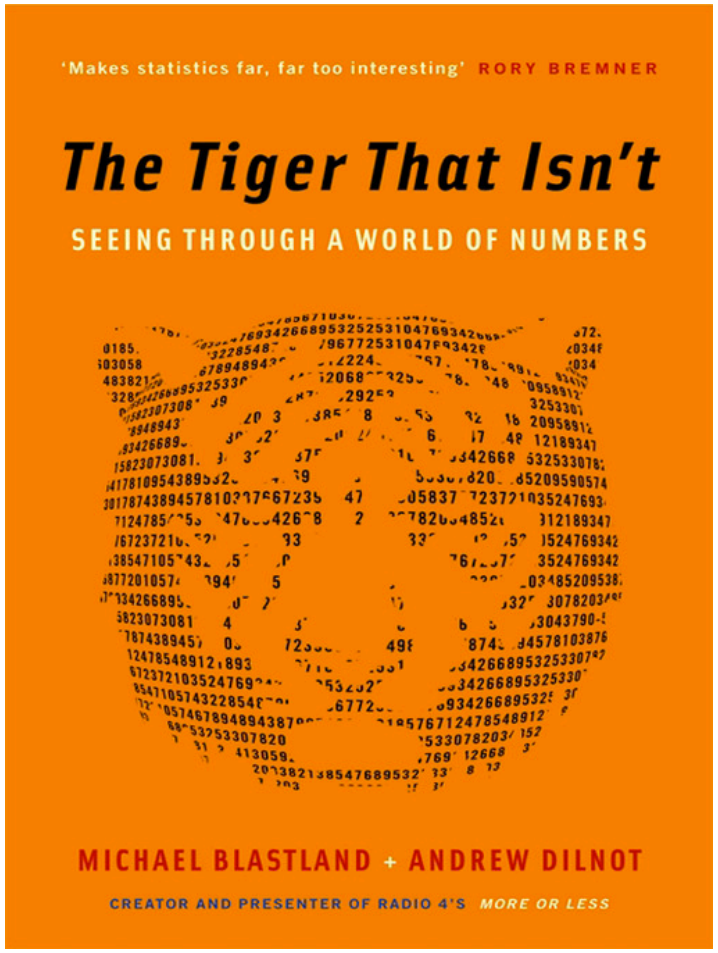
Statistics	Data Science
Sampling	Use *all* the data

Discussion

1. Where are you at the moment?
2. Where are the challenges for you With:
    - data
    - teams
    - technical skills



12 problems with numbers



5 number theory

1. Counting

Flickr: mattbrittaine

2. Big numbers

£300m

boost for childcare

1,000,000

new places

£1.15

per week

per child

3. Chance

Random events cluster

They do not evenly distribute

Introducing R

Download at

https://www.rstudio.com/

Run online at

https://www.rollapp.com/app/rstudio

Exercise sheet

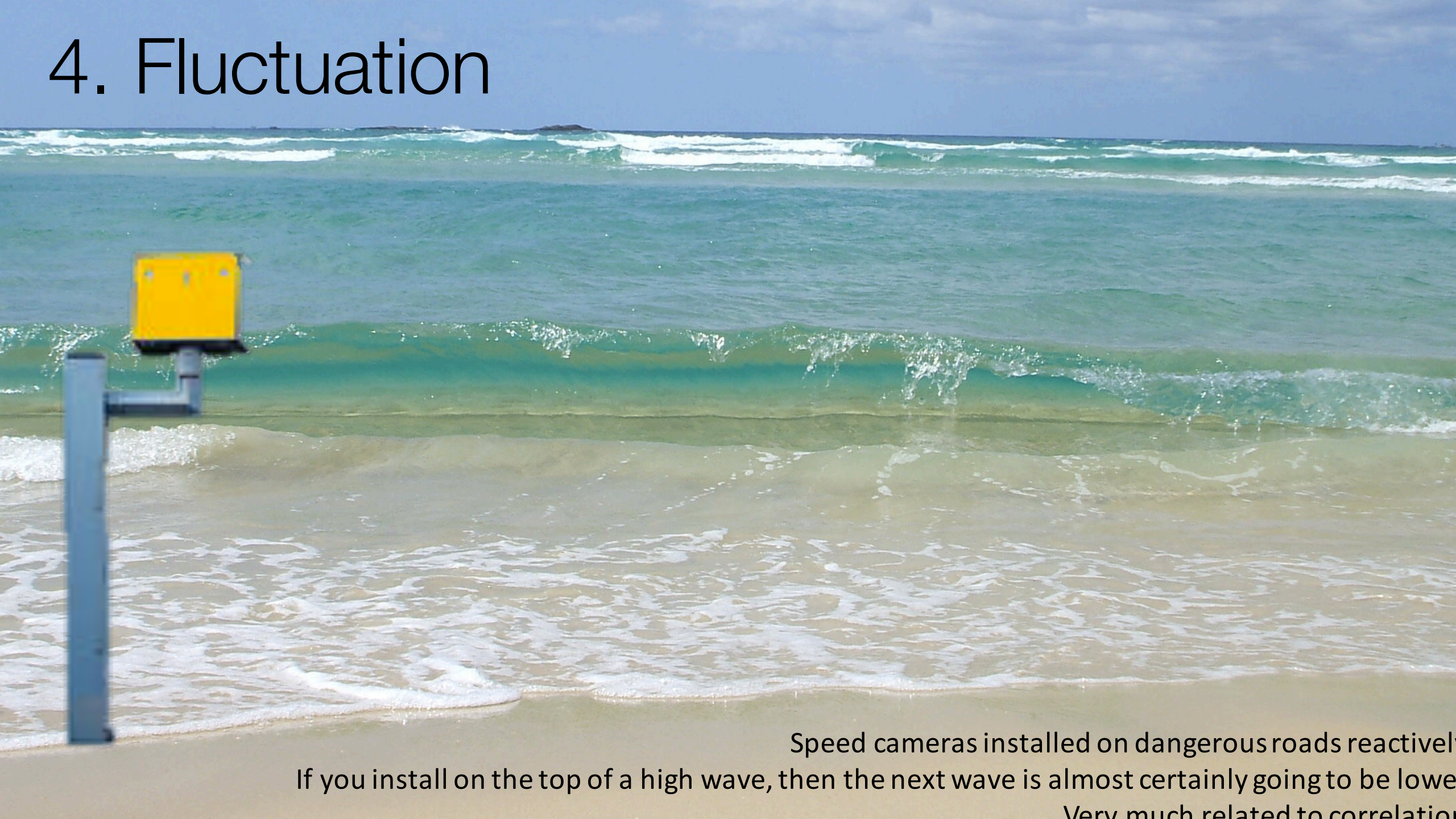
http://bit.ly/odiStats

Exercise

Discovering R (Part 1)

Averages

4. Fluctuation

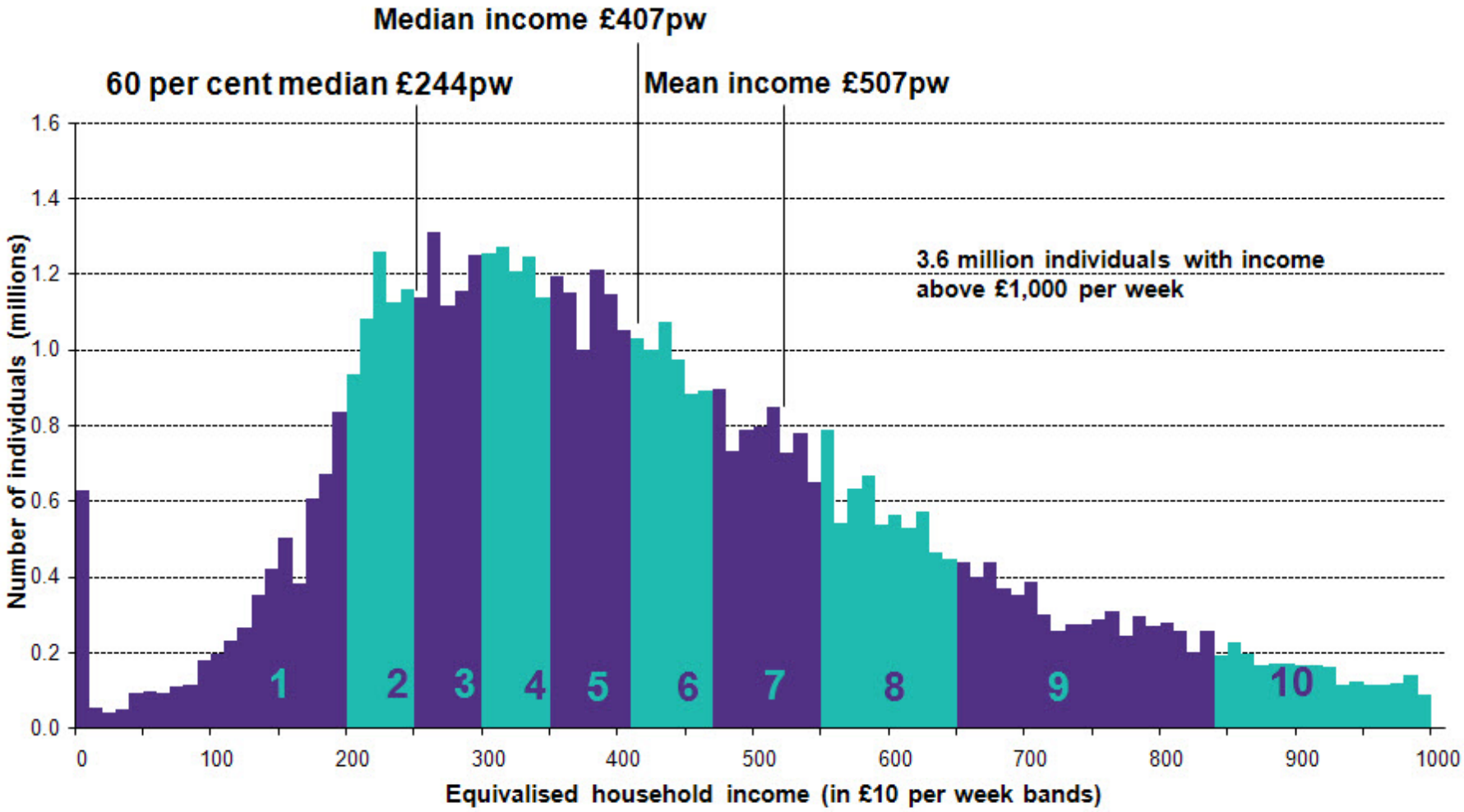


Speed cameras installed on dangerous roads reactivel

If you install on the top of a high wave, then the next wave is almost certainly going to be lowe

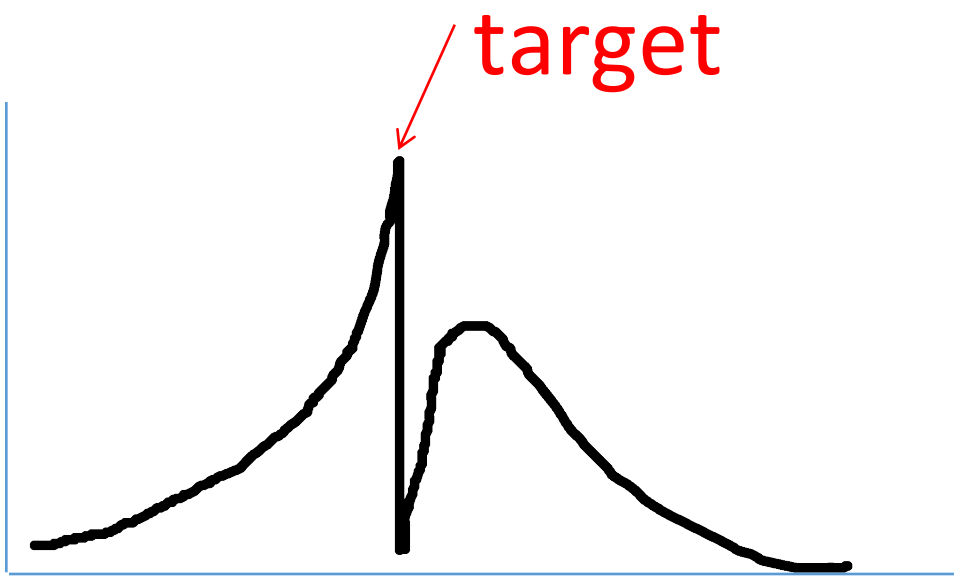
Very much related to correlatio

5. Averages


Median income £407pw
60 per cent median £244pw
Mean income £507pw
3.6 million individuals with income above £1,000 per week
Number of individuals (millions)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
0
100
200
300
400
500
600
700
800
900
1000
Equivalised household income (in £10 per week bands)

6. Targets

Look at one aspect to measure an entire service.



Encourage gamification

Exercise

Discovering R (Part 2)

Risk, samples and hypothesis testing

7. Risk


SECTIONS
OTTAWA CITIZEN
HOME
NEWS
NATIONAL
LOCAL NEWS
Combined va
seizure risk ir
ELIZABETH PAYNE
More from
Published on: June 9, 2014 | Last Upda
CANCER RESEARCH UK
Donate
HOME
MENU
SEARCH
Home › About us › Cancer News › News report › Global cancer incidence predicted to increase by 75 per cent by 2030
Global cancer incidence predicted to increase by 75 per cent by 2030
News report
31 May 2012
In collaboration with the Press Association
The number of worldwide cancer cases is set increase by 75 per cent in the next two decades, according to researchers in France.
The scientists predict cancer cases will increase from 12.7 million in 2008 to 22.2 million by 2030.
Recent news
Investing in cancer research boosts economy as well as

8. Sampling

Between 2,000 and 5 million cases of norovirus in winter 2007/08.

Based upon 2,000 confirmed cases extrapolated from BMJ report based on sample size of …. 1

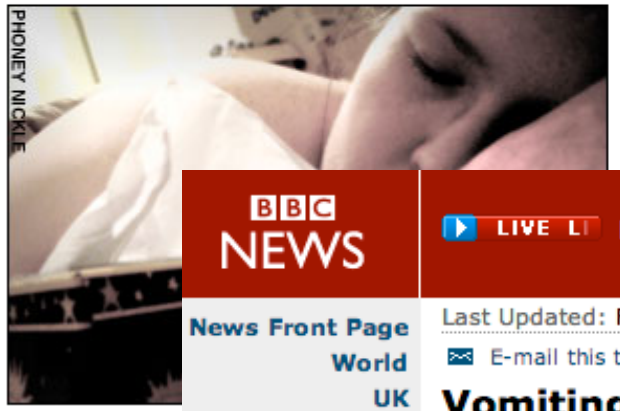
The Telegraph

Home | News | World | Sport | World Cup | Finance | Comment | Cultu

Politics | Investigations | Obits | Education | Earth | Science | Defence

HOME » NEWS » UK NEWS

GPs urge millions hit by bug to stay at home



The NHS advises symptoms

BBC NEWS

LIVE L BBC NEWS CHANNEL

News Front Page
World
UK
England
Northern Ireland
Scotland
Wales
Business
Politics
Health
Medical notes
Education
Science & Environment
Technology

Last Updated: Friday, 11 January 2008, 12:02 GMT

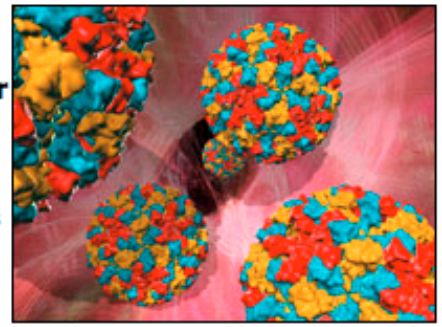
E-mail this to a friend Printable version

Vomiting bug 'hits three million'

Almost three million people have been affected by the norovirus stomach bug so far this winter, figures suggest.

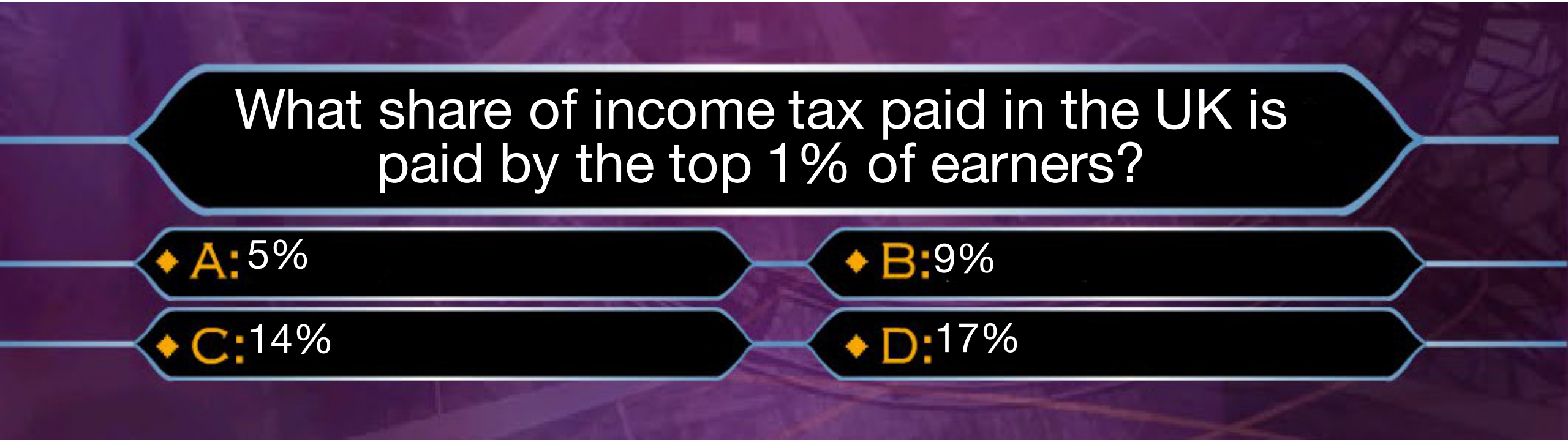
Surveillance from the Health Protection Agency shows cases in England and Wales are double those seen last year.

Doctors advise people to stay at home for 48 hours after

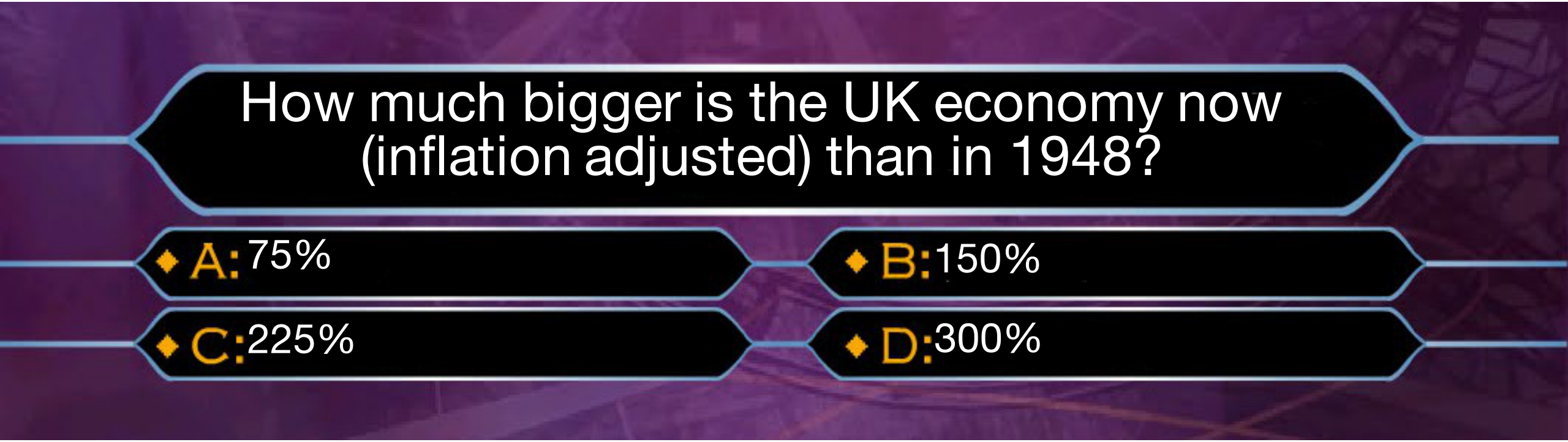


Norovirus causes sudden vomiting and diarrhoea

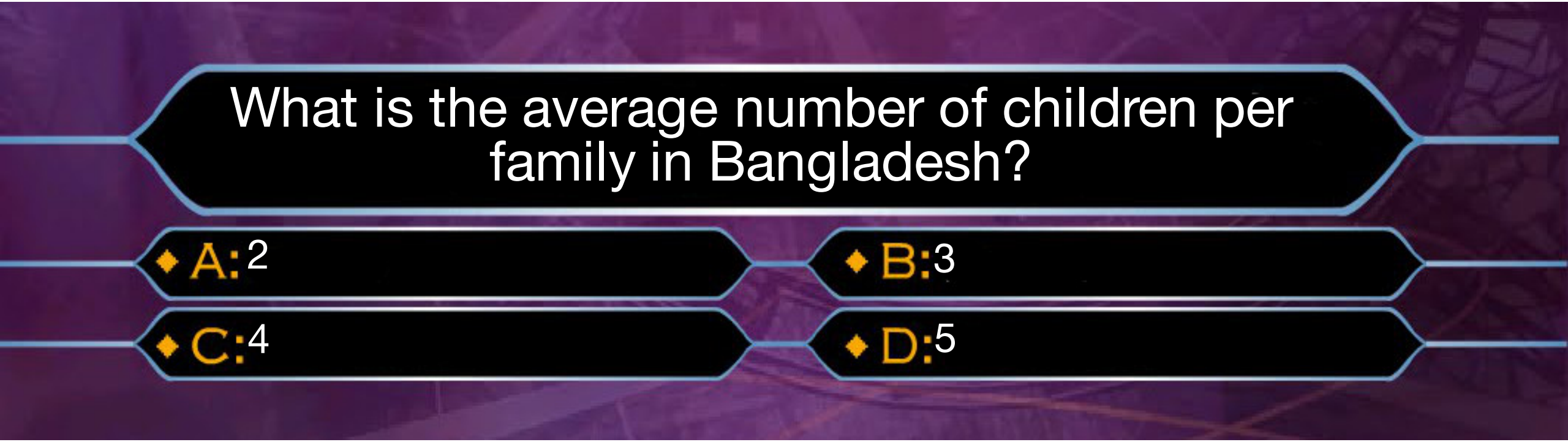
9. Data (known unknowns)



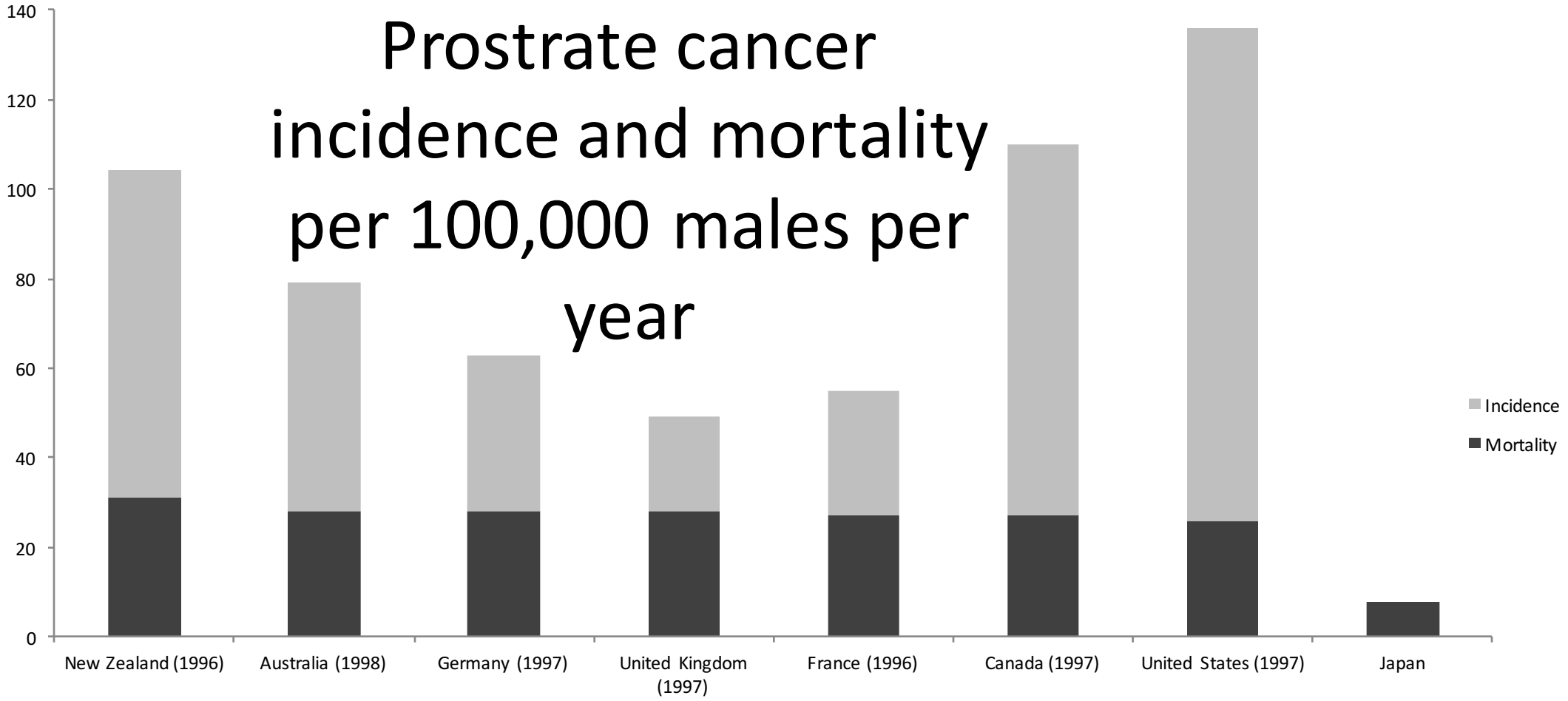
9. Data (known unknowns)



9. Data (known unknowns)



10: Comparison

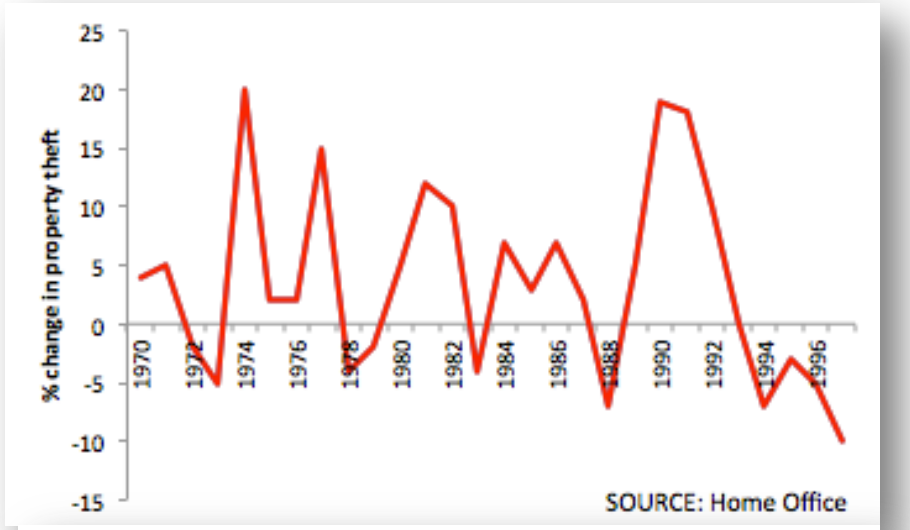


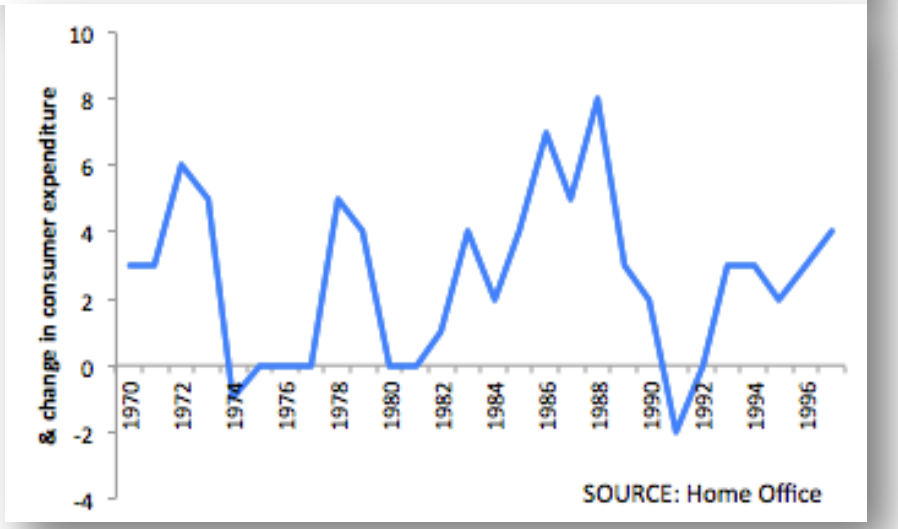
Exercise

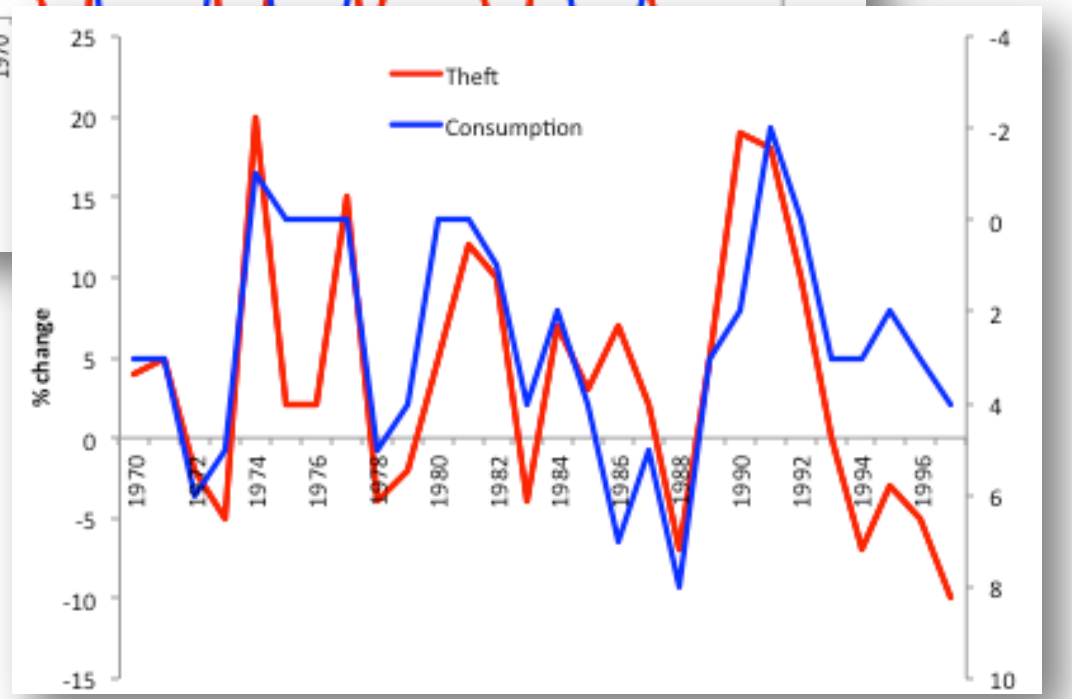
Discovering R (Part 3)

Correlation and percentages

11. Correlation







12. Percentages

Know the difference between a **percentage** and a **percentage point**.

VAT increased from 17.5% to 20% on January 2011.

This is a rise of 2.5 percentage points **not** a rise of 2.5%.

How much would a rise in 2.5% actually be?

From: Stories and Statistics by Frank Swain

Summary

1) Counting
2) Big numbers
3) Chance
4) Fluctuation
5) Averages
6) Targets
7) Risk
8) Sampling
9) Known unknowns
10)Comparison
11)Correlation
12)Percentages

Review

Given the new houses dataset.

How might you analyse the relationship between houses in New York and those in San Francisco?

Outcomes

Explain the importance of good statistical methods
Describe a number of key statistical techniques
Apply a number of statistical techniques to data
Use R-Studio to explore data through applied statistics

Machine learning and classification

Outcomes

Explain the core aspects of Machine Learning

Apply classification to a set of data to create a decision tree

Write a machine learning algorithm

Machine Learning

"construction and study of systems that can learn from data"

Can be seen as building blocks to make computers learn to behave more intelligently

There are various *techniques* with various *implementations*.

Slide ideas thanks to @rahuldausa

Question?

What instances of machine learning have you come across?

What types are they?

Use-Cases

- Spam Email Detection
- Machine Translation (Language Translation)
- Image Search (Similarity)
- Clustering (KMeans) : Amazon Recommendations
- Classification : Google News

Use-Cases (contd.)

- Text Summarization - Google News
- Rating a Review/Comment: Yelp
- Fraud detection : Credit card Providers
- Decision Making : e.g. Bank/Insurance sector
- Sentiment Analysis
- Speech Understanding – iPhone with Siri
- Face Detection – Facebook’s Photo tagging

Not Spam

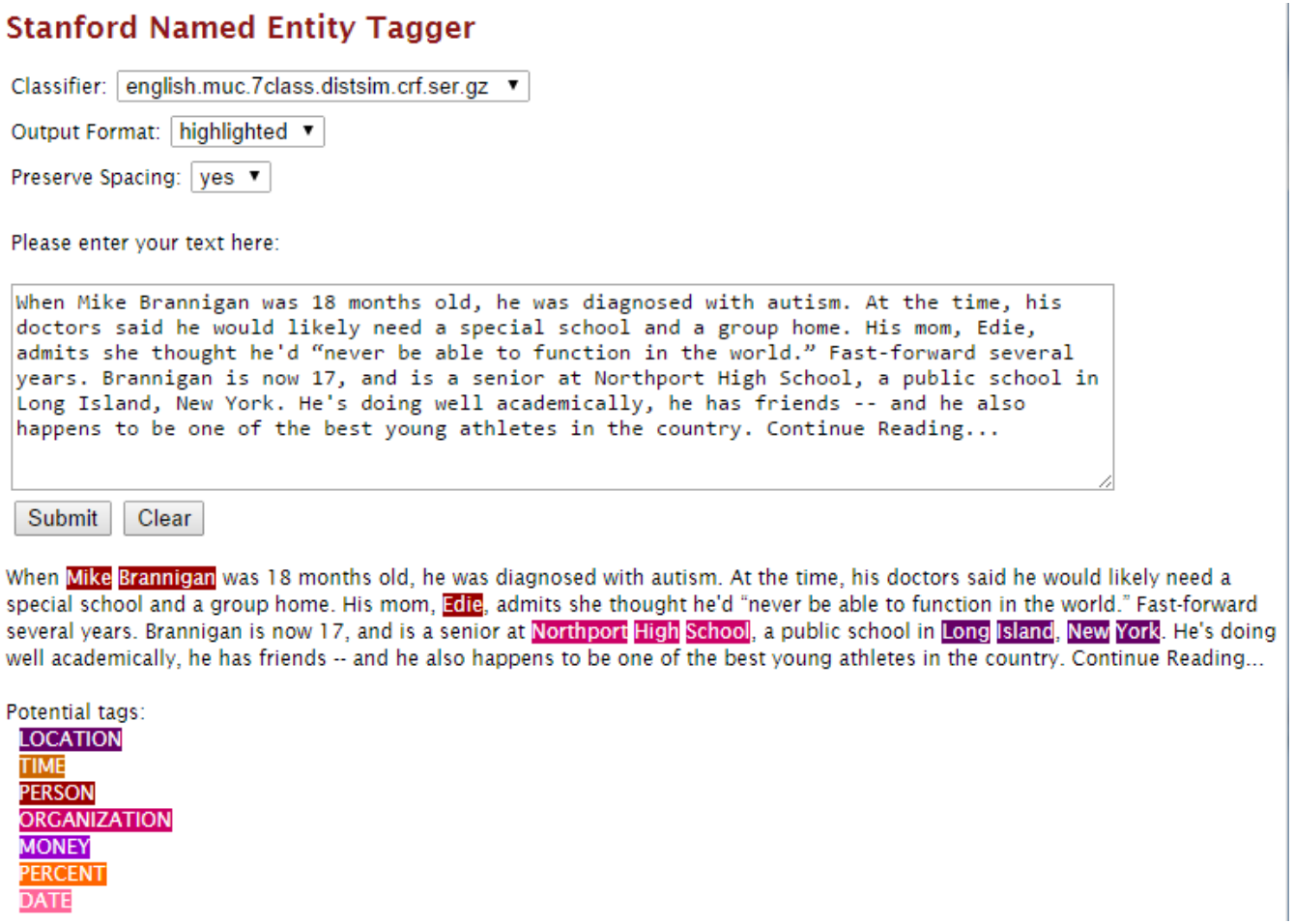
Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

HDFC Bank	**LOAN upto Rs 25 lac - Disbursal in 2 days** - open in fresh tab -- If you do not want to receive any more newsletters, please click here	9:40 pm
iEntry	**Welcome iEntry Member** - Ultimate Guide To Assessing	9:23 pm
New-Zealand-Jobs.067L	**Come to New Zealand to find a great job and settle here (Search for all Jobs from diffe...** - Search for all Jobs from different kinds of industries Find a Job in Enchanting	8:18 pm
CarSizzler	**Assured Free Luxurious Ride worth Rs.300 with Uber Cabs** - Home Home Buy New Car Buy New Car Sell Car Sell Car Tech Tics Tip & Tale Facebook 41727 others	6:05 pm
Supermarket Promotion	**Enjoy Rs.1700 voucher valid at any supermarket!** - If you are unable to view this mailer Click here HOW TO CONTACT US? BY EMAIL: support@savethedeals.in	4:51 pm
Entireweb Newsletter	**Hire an SEO the Right Way – 6 Tips You Must Remember for Life** - Unsubscribe me View web version Become a fan on Facebook Follow us on Twitter September 5th, 2	1:24 pm
Max Bupa	**A policy that understands your family's medical need** - open in fresh tab -- If you do not want to receive any more newsletters, pleas	11:08 am
Scoop.it	**Your Scoop.it Daily Summary - How to Maximize Your LinkedIn Publishing Exposure | SME a...** - Scoop.it Facebook Twitter G+ Hi re the top	9:30 am
standard charterer Bank	**Instant approval on your Credit Card`**	7:27 am
CAR TRADE	**Sell your car at no cost at all** - If you are having trouble viewing this email,view web version | View this message in your mobile |	Sep 4
Uday	**VPS Web Hosting Services Provider** - Dear Sir, I am Uday Sharma, Business development executive. We are providing quality VPS hosting for	Sep 4
Mark Regan, SPN	**How to Find Your Most Valuable Keywords [Free Guide]** - This is a SiteProNews/ExactSeek Webmaster Exclusive Mailing! To drop your subscription, use the link	Sep 4
HDFC Bank	**LOAN upto Rs 25 lac - Disbursal in 2 days** - open in fresh tab You have received this mailer from Shop@Best on behalf of HDFC Bank because you	Sep 4
CAR TRADE	**Sell your car at no cost at all** - open in fresh tab -- If you do not want to receive any further newsletters, please click here	Sep 4
ICICI Bank	**Home Loan Interest Rate starting from 10.15%*. Get Instant Approval!** - open in fresh tab -- If you do not want to receive any more newsletters, please Click Here	Sep 4
calculateyourwealth	**It's good when your bank helps you manage your wealth and fulfill your ambitions** - Calculate Now Dreams you wish to realize in your lifetime require enough wealth. C	Sep 4
Angel Broking	**Get Low Brokerage - Free Demat & Trading Account** - open in fresh tab -- If you do not want to receive any further newsletters, please click here	Sep 4
Bankbazaar	**7 Minute Instant Online Approval for your PESONAL LOAN** - Now get instant online Personal Loan approval in 7 minutes by BankBazaar.com from leading Banks in	Sep 4
Jayde	**Welcome To The Jayde Newsletter!** - Welcome To WebProNews Welcome To The Jayde Newsletter! Before we begin, make sure to add	Sep 4
ineedhits noreply	**[ineedhits] Your ineedhits Account and Password** - ACCOUNT CREATION Account ID : A1588368 Dear Rah, Welcome to ineedhits. Yo r	Sep 4
Rekha	**Mobility Apps for Your Business** - While we look at the span of last 20 years, we could broadly look at two distinct eras, - Life in	Sep 4
SlideShare Newsletter	**Top Tips From the World Champions of PowerPoint** - View online version Remember to display images Meet the PowerPoint World Champs Top Tips From the	Sep 4
Dilshad Pathan	**Feeling Hesitate to Discuss personal Health Queries** - My Life Care Follow Us on facebook twitter linkedin Google+ Feeling Hesitate to Discuss personal	Sep 4
Vaishu	**TAKE YOUR PICK. Register in SimplyMarry** - TAKE YOUR PICK. Register in SimplyMarry -- Regards Vaishu	Sep 4

Not a Spam

Not a Spam

NER (Named Entity Recognition)



http://nlp.stanford.edu:8080/ner/process

Similar/Duplicate Images

About 81 results (0.70 seconds)

Image size:
250 × 321

No other sizes of this image found.

Best guess for this image: ***taj mahal***

Visually similar images Report images

Credit: https://www.google.co.in/

Remember

Features ?

(Feature Extraction)

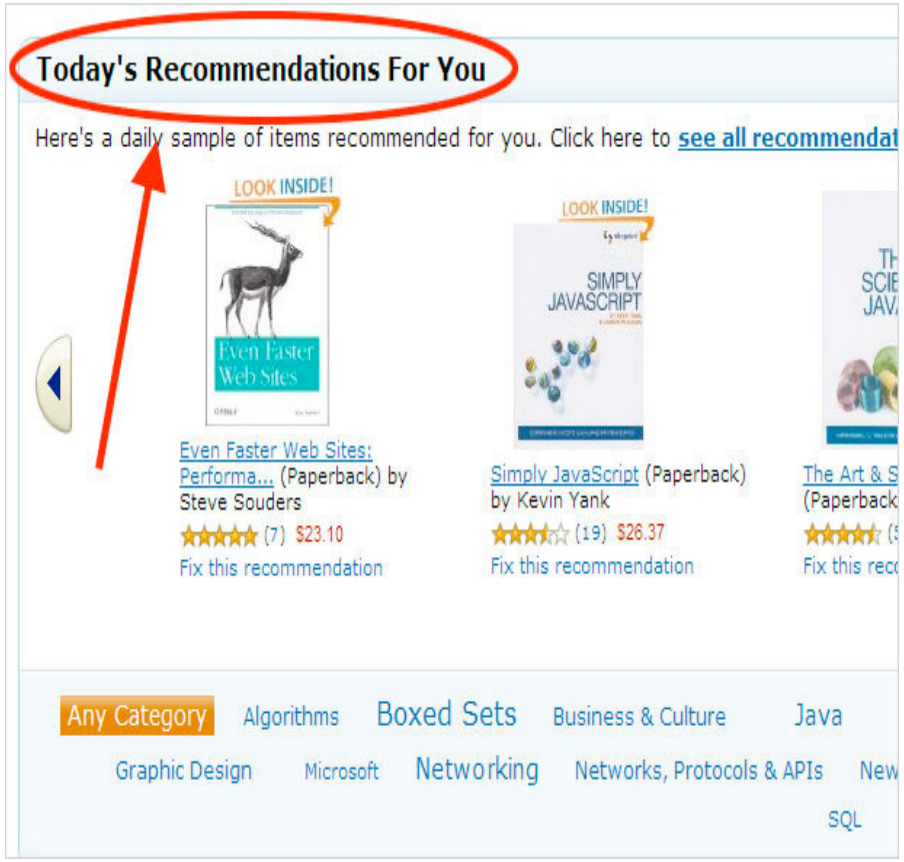
Can be :

- Width
- Height
- Contrast
- Brightness
- Position
- Hue
- Colors

Check this :

LIRE (Lucene Image REtrieval) library - https://code.google.com/p/lire/

Recommendations



http://www.webdesignerdepot.com/2009/10/an-analysis-of-the-amazon-shopping-experience/

Terminology

Features

The number of features or distinct traits that can be used to describe each item in a quantitative manner.

Samples

A sample is an item to process (e.g. classify). It can be a document, a picture, a sound, a video, a row in database or CSV file, or whatever you can describe with a fixed set of quantitative traits.

Feature vector

is an n-dimensional vector of numerical features that represent some object.

Feature extraction

Preparation of feature vector

Transforms the data in the high-dimensional space to a space of fewer dimensions.

Training/Evolution set

Set of data to discover potentially predictive relationships.

Slide ideas thanks to @rahuldausa

Learning (Training)

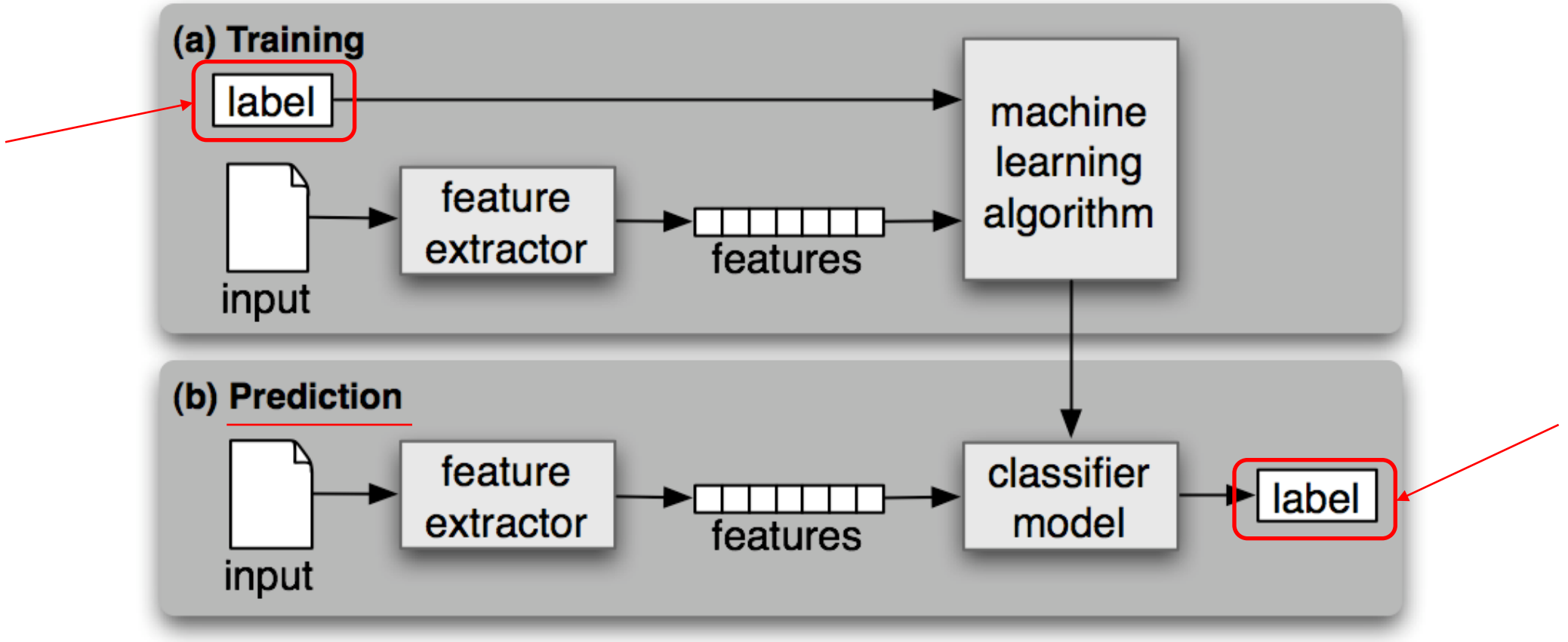
Features:
1. Color: **Radish/Red**
2. Type : **Fruit**
3. Shape
etc…

Features:
1. Sky Blue
2. **Logo**
3. Shape
etc…

Features:
1. **Yellow**
2. **Fruit**
3. Shape
etc…

Slide ideas thanks to @rahuldausa

Workflow



Slide ideas thanks to @rahuldausa

Property locations?

Each table has a set of “Top Trump” training-set cards.

Build a decision tree to sort them into “New York” and “San Francisco”.

You cannot use the target to sort them.

Property locations? (part 2)

Build the machine learning algorithm in R

Run online at

https://www.rollapp.com/app/rstudio

Exercise sheet

http://bit.ly/odimach

Categories

Supervised Learning

Unsupervised Learning

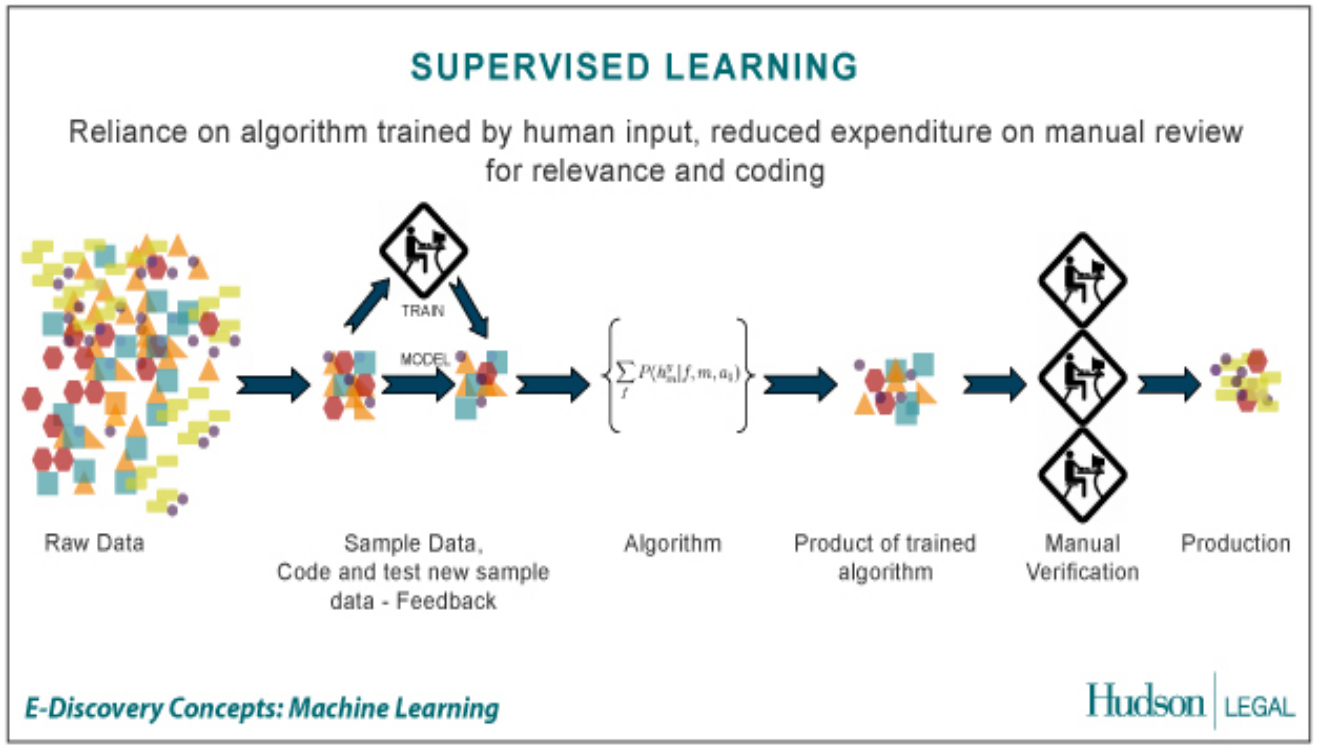
Semi-Supervised Learning

Reinforcement Learning

Slide ideas thanks to @rahuldausa

Supervised Learning

The correct labels of the training data are known

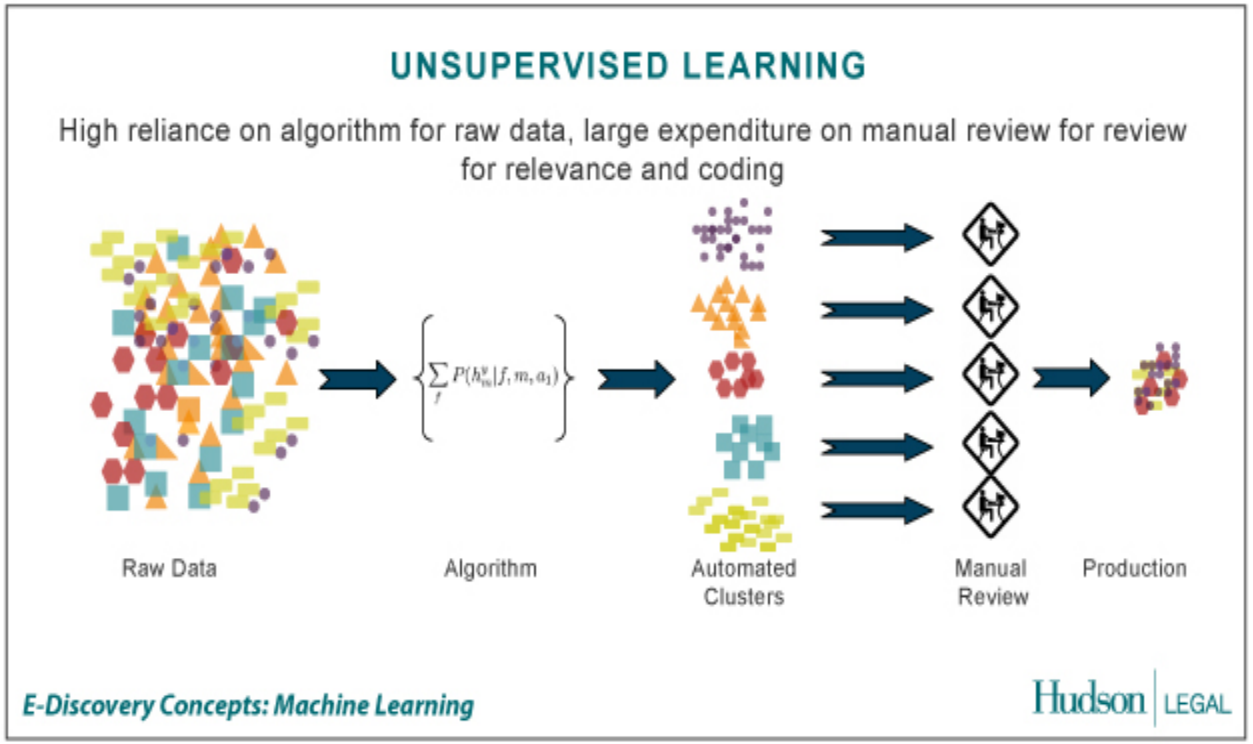


Credit: http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth

Slide ideas thanks to @rahuldausa

Unsupervised Learning

The correct labels of the training data are not known.
Cluster discovery and pattern generation.



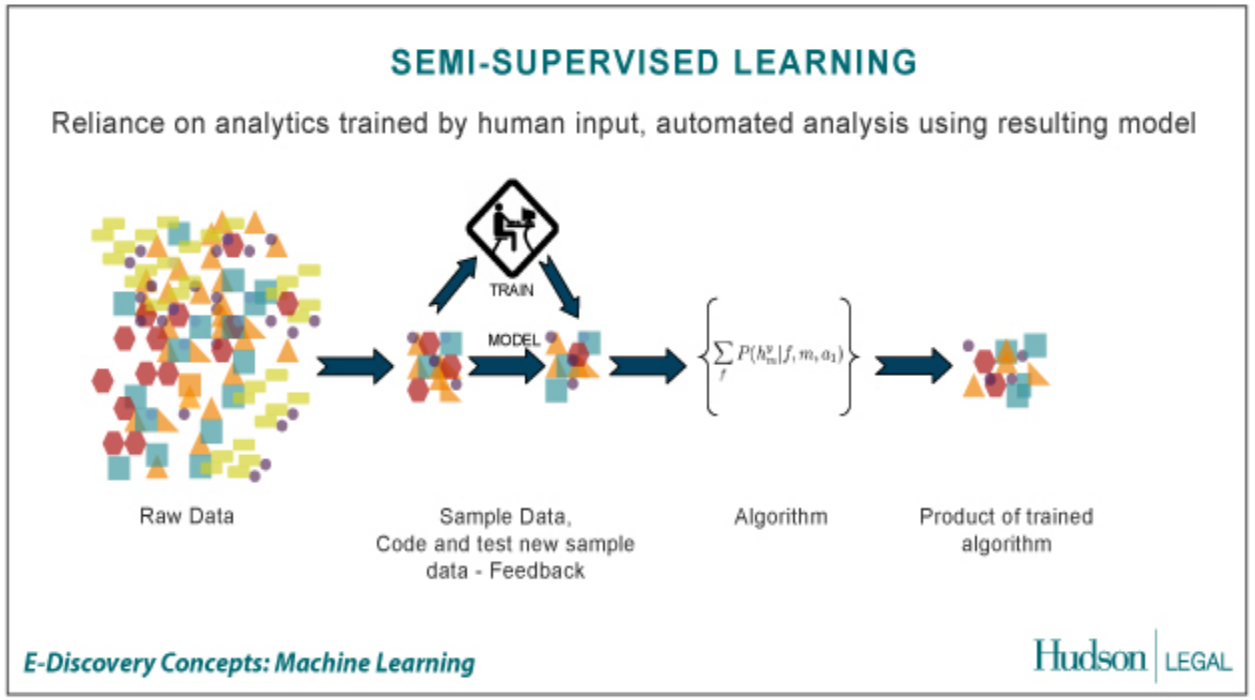
Credit: http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth

Slide ideas thanks to @rahuldausa

Semi-Supervised Learning

A Mix of Supervised and Unsupervised learning.

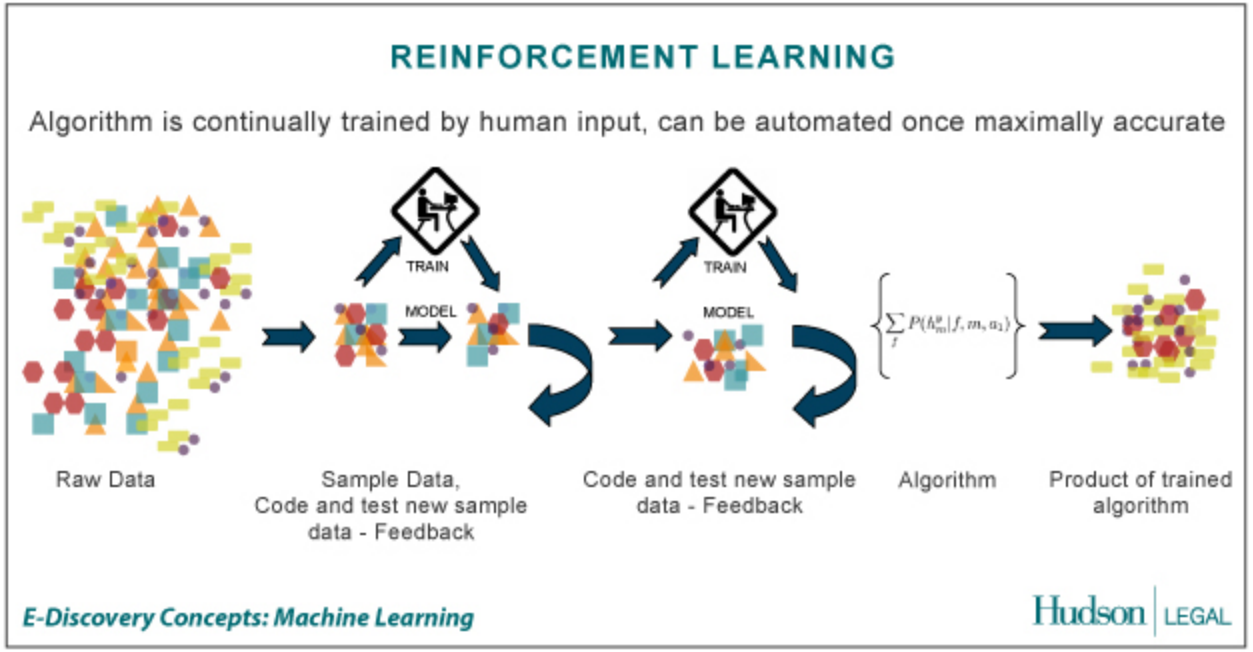
Some labels are known, but majority are not.



Credit: http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth

Reinforcement Learning

Software agent learns based on feedback and/or reward.



Credit: http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth

Techniques

classification

predict class from observations

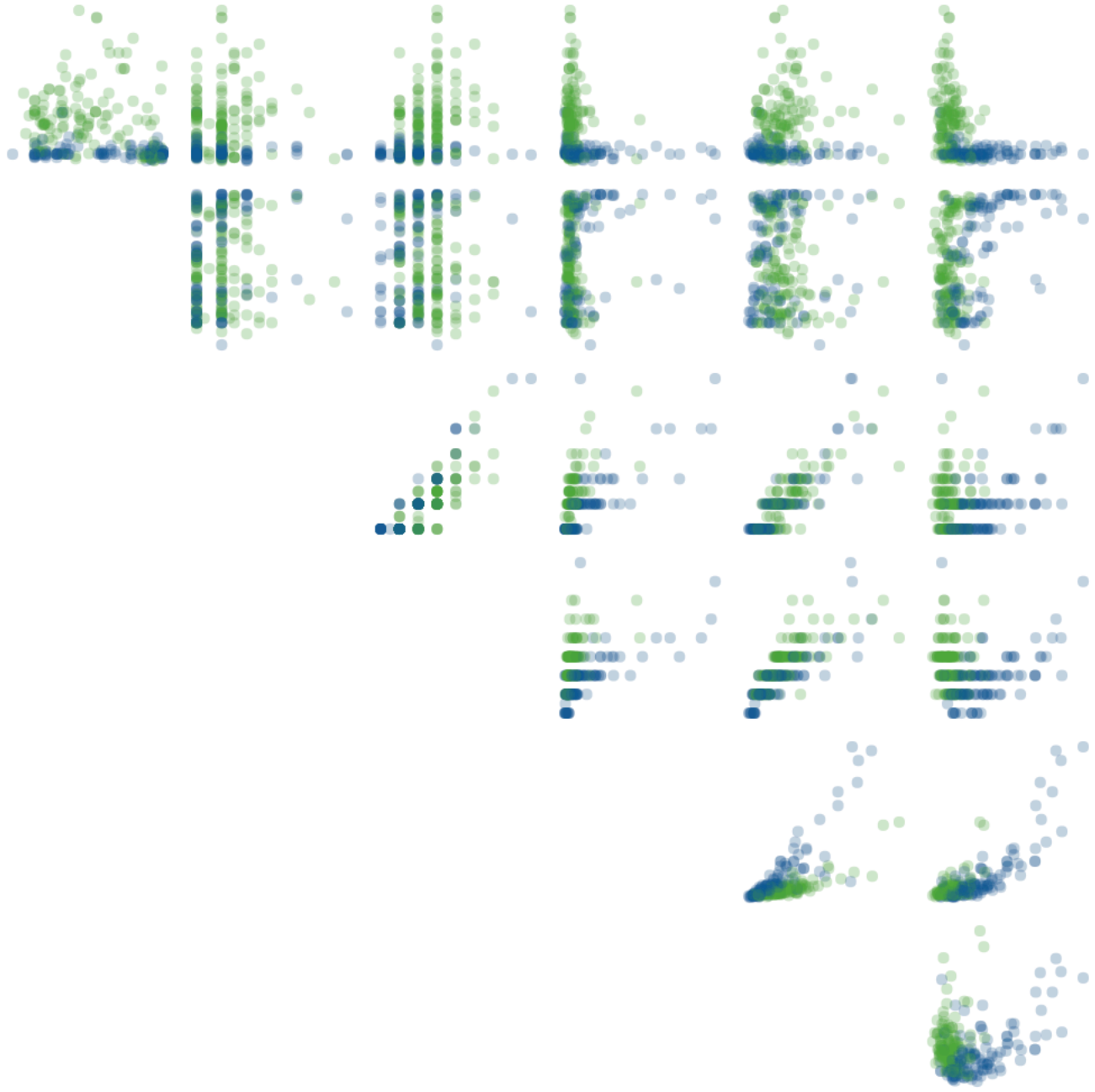
clustering

group observations into “meaningful” groups

regression (prediction):

predict value from observations

r2d3.us



Frameworks/Tools

R
Weka
Carrot2
Gate
OpenNLP
LingPipe
Stanford NLP
Mallet – Topic Modelling
Gensim – Topic Modelling (Python)
Apache Mahout
MLib – Apache Spark
scikit-learn - Python
LIBSVM : Support Vector Machines
and many more…

Outcomes

Explain the core aspects of Machine Learning

Apply classification to a set of data to create a decision tree

Write a machine learning algorithm