http://training.theodi.org/ods

David Tarrant · @davetaz

Course bought to you in collaboration with the
European Data Science Academy Project

Course aim

Equip you with the knowledge and tools to help you upskill as modern data scientists.

Session 1

Discovering Open Data Science

Session 2

Machine learning and classification

Session 3

Visualisation and communication

Introductions

Your name & role?

Why is data science important to you?

What do you want to do differently after this course?

Session 1

Discovering Open Data Science

Outcomes

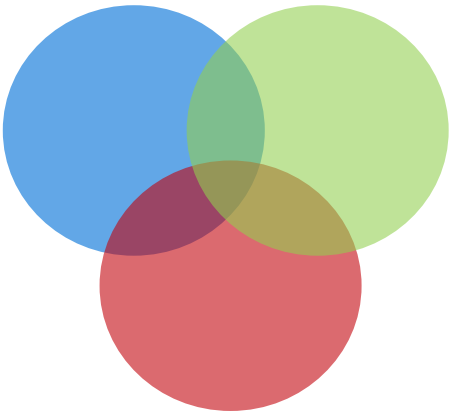
Describe the key skills of an open data scientist

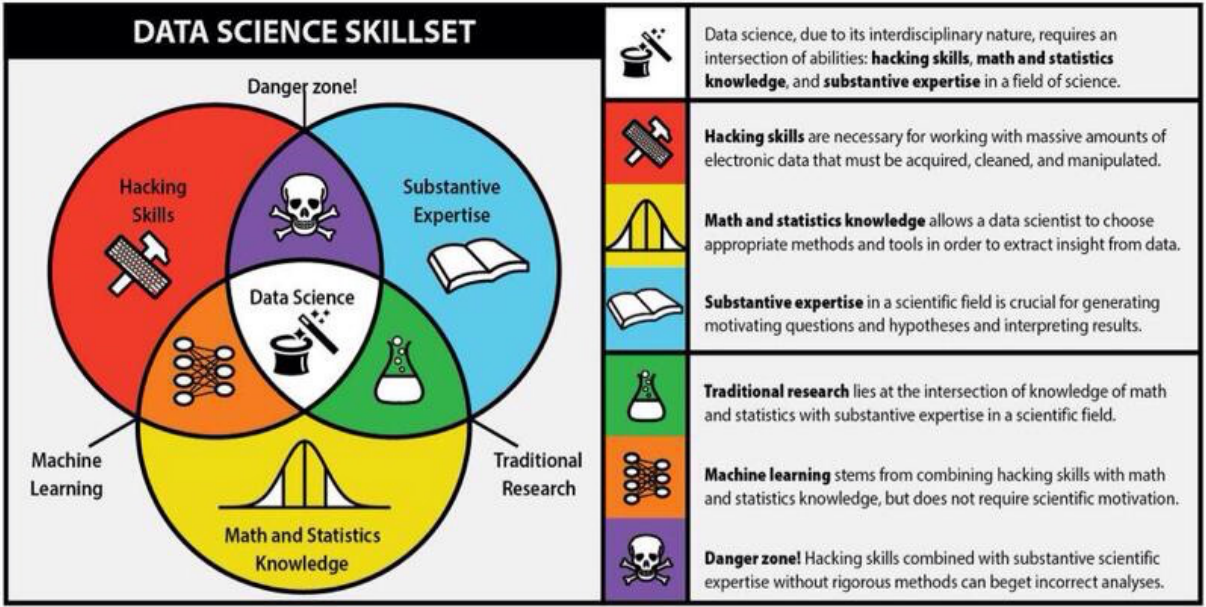
Apply pivot tables on big data in the cloud

Explain and use ”data on the web” and the ”web of data”

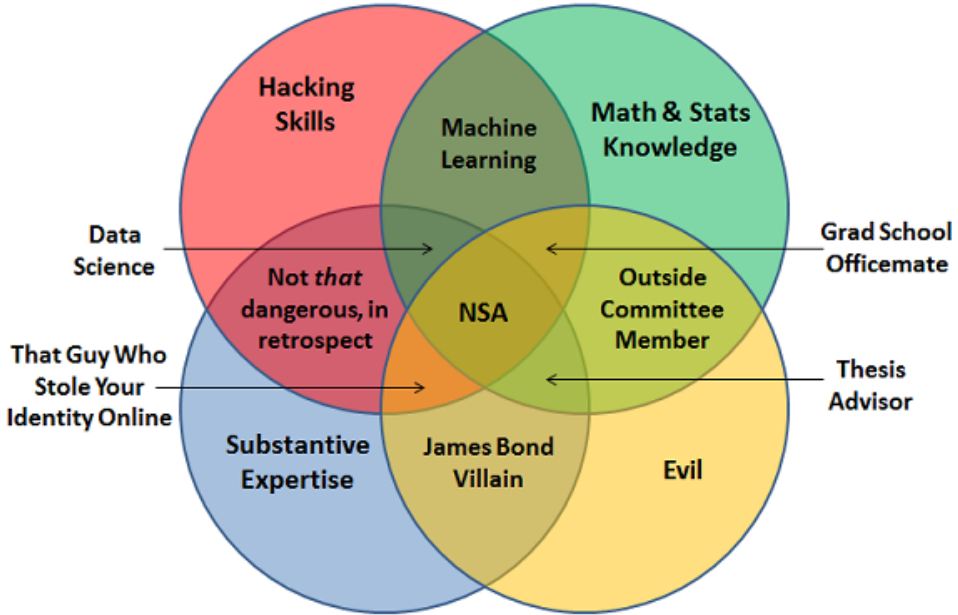
Before we begin

What key skills and competencies does a modern open data scientist require?





Courtesy of Natalia Bilenko, modified from Drew Conway (the Data Science Venn Diagram)


Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Grad School Officemate
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil

Data Science

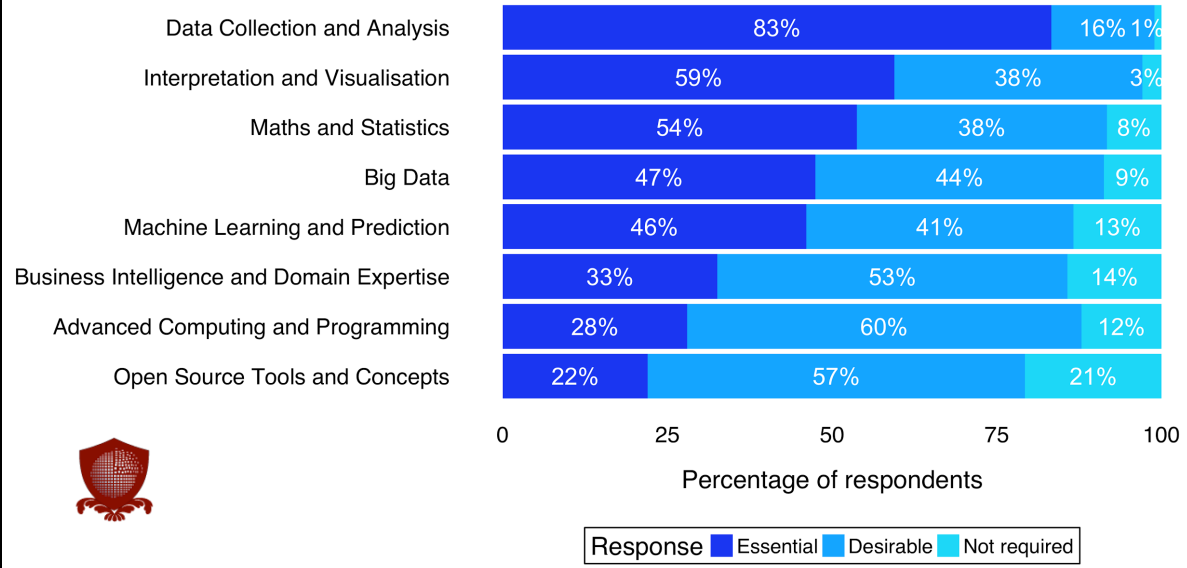
“part analyst, part artist”

-- Anjul Bhambhri (VP for Big Data, IBM)

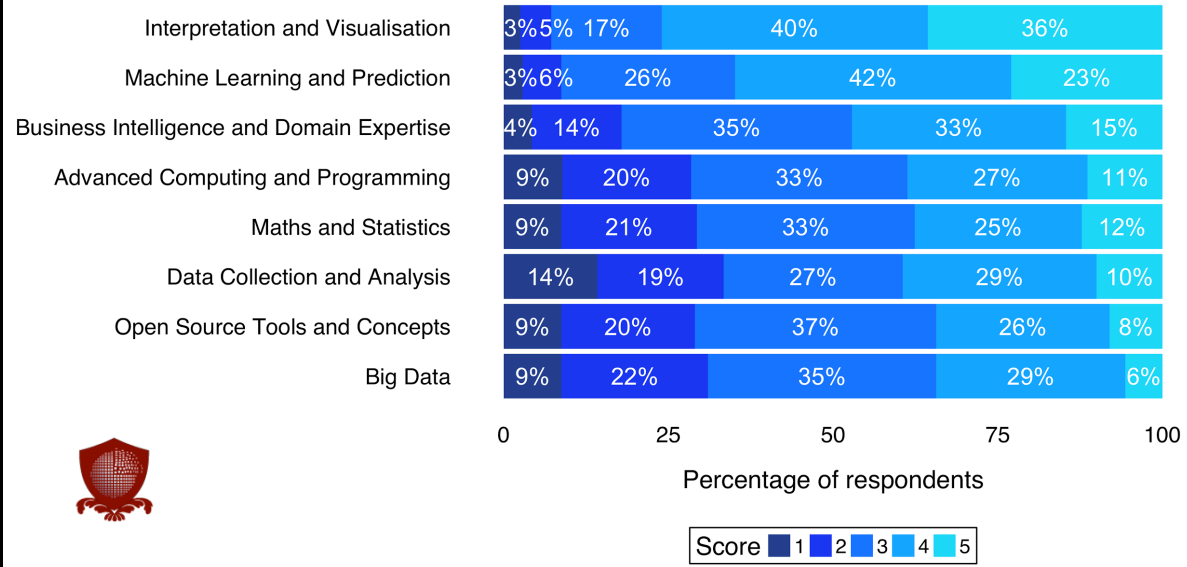
Key areas

1) Big Data
2) Data Collection and Analysis
3) Machine Learning and Prediction
4) Maths and Statistics
5) Interpretation and Visualisation
6) Advanced Computing and Programming
7) Business Intelligence and Domain Expertise
8) Open Source Tools and Concepts

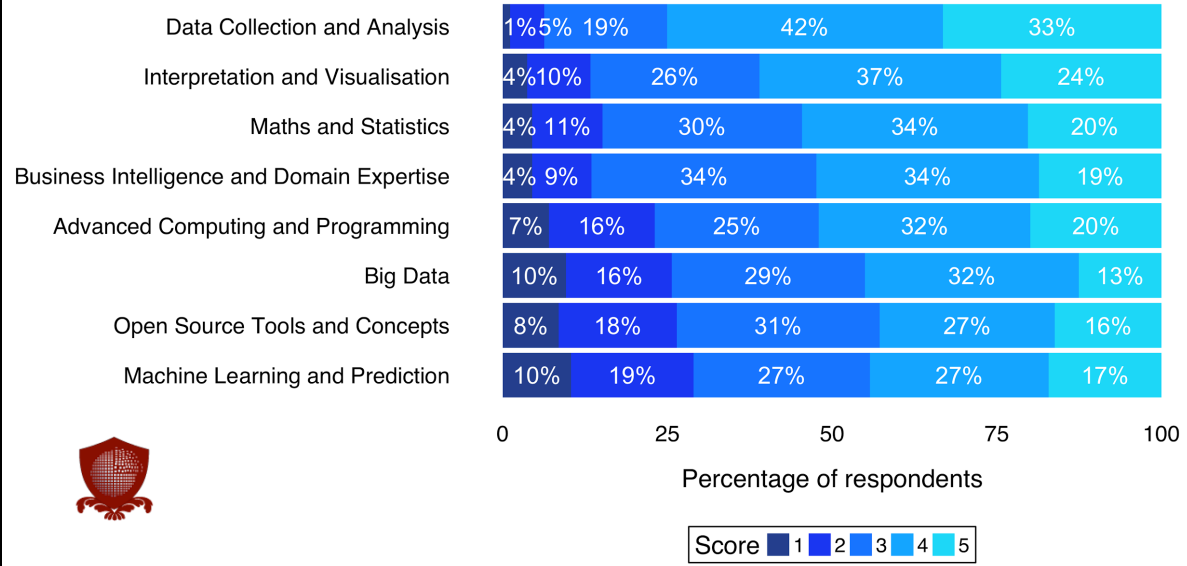
Skills that a data scientist should have (N = 633)



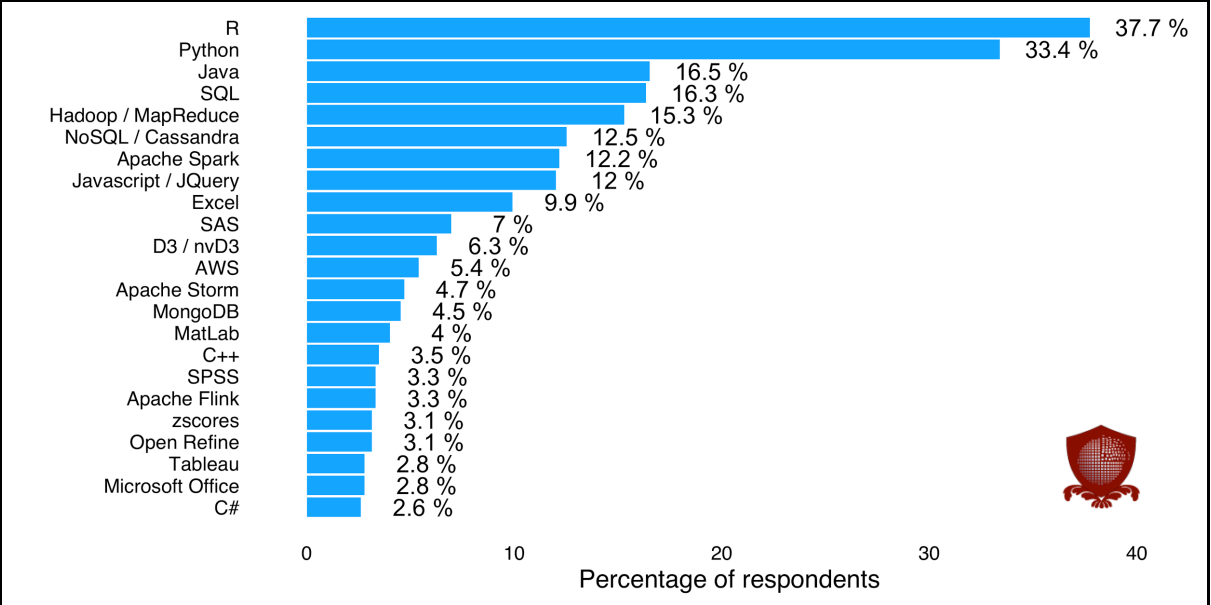
Self-assessment of own skills by data scientists (N = 355)



Assessment of team's skills by manager (N = 278)



Key skills and tools



Key areas

1) Big Data

Filtering and processing
6 million rows of data in
5 minutes.

2) Machine Learning and Prediction

3) D

4) M

5) In

6) A mming

7) Business Intelligence and Domain Expertise

8) Open Source Tools and Concepts

Conclusion

Commodity scalable cloud computing is ready to be used. Build and test small and then scale on demand.

Amazon EC2, Heroku, Cloudflare are all great examples. Socrata and tableau are others.

Key areas

1) Big Data

2) Data Collection and Analysis

3) Machine Learning and Prediction

4) Maths and Statistics

5) Interpretation and Visualisation

6) Advanced Computing and Programming

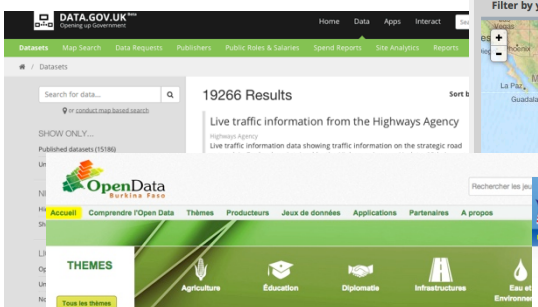
7) Business Intelligence and Domain Expertise

8) Open Source Tools and Concepts

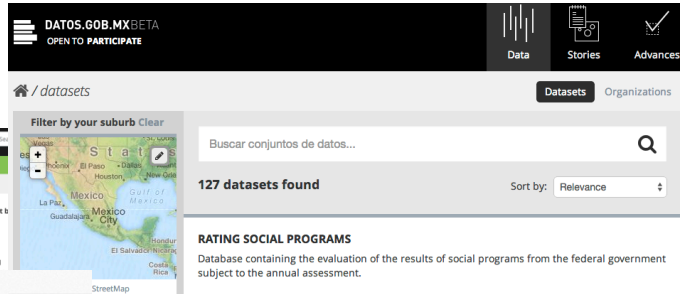
Structures

Not all data that looks tabular, is…

data.gov.XX

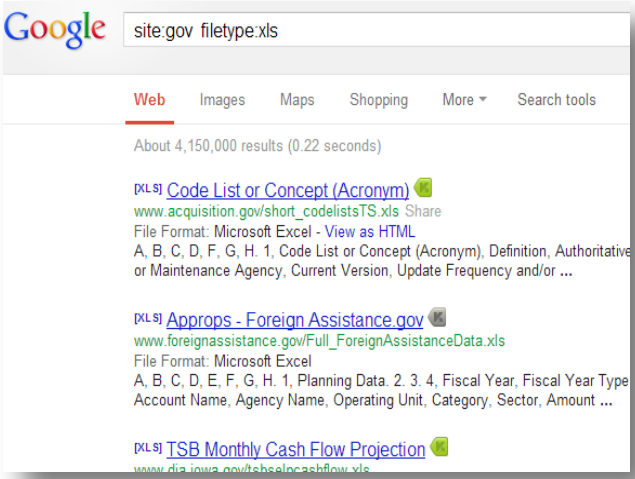

DATA.GOV.UK
Opening up Government
Home
Data
Apps
Interact
Datasets
Map Search
Data Requests
Publishers
Public Roles & Salaries
Spend Reports
Site Analytics
Reports
Datasets
Search for data...
or conduct map based search
19266 Results
Live traffic information from the Highways Agency
Highways Agency
Live traffic information data showing traffic information on the strategic road
SHOW ONLY...
Published datasets (15186)

OpenData
Burkina Faso
Accueil
Comprendre l'Open Data
Thèmes
Producteurs
Jeux de données
Applications
Partenaires
A propos
THEMES
Tous les thèmes
Agriculture
Éducation
Diplomatie
Infrastructures
TIC
Santé
Collectivités territoriales
Sécurité Publique
Open Data Burkina Faso statistiques
78 jeux de données
28 organisations
11 groupes


DATOS.GOB.MXBETA
OPEN TO PARTICIPATE
Data
Stories
Advances
/ datasets
Datasets
Organizations
Filter by your suburb Clear
Houston
Mexico City
Gulf of Mexico
StreetMap
Buscar conjuntos de datos...
127 datasets found
Sort by: Relevance
RATING SOCIAL PROGRAMS
Database containing the evaluation of the results of social programs from the federal government subject to the annual assessment.


data.gov.my
Portal Rasmi Open Data Kerajaan Malaysia
The Government of Malaysia's Open Data Official Portal
HOME
ABOUT US
CATALOGUE
INFOGRAPHICS
MOBILE APPS
CONTACT
DATASETS INFO
117 Datasets
11 Ministries
10 Sectors
RESOURCES
Malaysia Directory
Malaysian Open Source Centre
Open Government Data
MOBILE APPS
myHealth I myJakim I KPDNKK
SPAD I SELAWAT I myMAHTAS
6 Android Mobile Apps
4 iOS Mobile Apps
QUICK LINKS
Malaysia Informative Data Centre
Malaysian Population Quick Info
Tourism Malaysia - Facts & Figures
Latest Datasets
Top Publishers
Feedback
Aplikasi Pemetaan Belia Malaysia
Info belia berdasarkan domain
by : Ministry of Youth and Sports
Last Updated
2014-06-26 11:30:19
Aplikasi Pemetaan Belia Malaysia
Statistik Sikap Belia Malaysia mengikut
by : Ministry of Youth and Sports
Last Updated
2014-06-26 11:27:30
Aplikasi Pemetaan Belia Malaysia
Statistik Pemetaan Belia Malaysia dalam bentuk aplikasi
by : Ministry of Youth and Sports
Last Updated
2014-06-26 11:24:52
Latest News and Events
Currently, there are no latest event.

Google advanced



site: Get results only from certain sites or domains

link: Find pages that link to a certain page

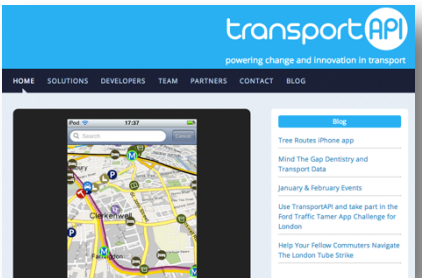
related: Find sites similar to one you already know

filetype: Find certain file types only

Aggregators and portals

Collect together data from across the web into one place.

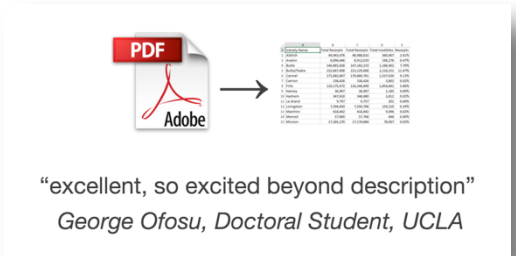
enigma.io



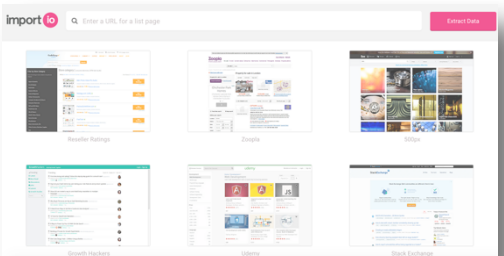
transportAPI

Scraping

If you can't obtain usable data (csv, xls) then you may have to resort to scraping.

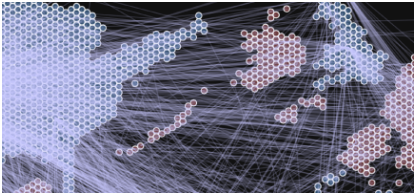


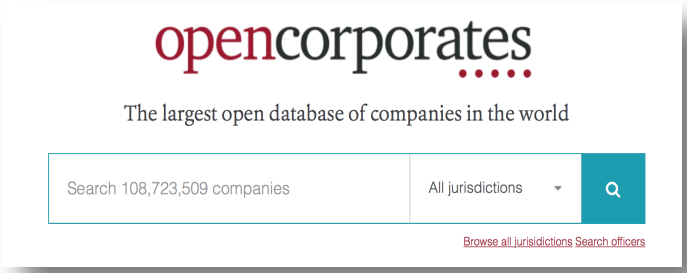
pdftables.com



magic.import.io

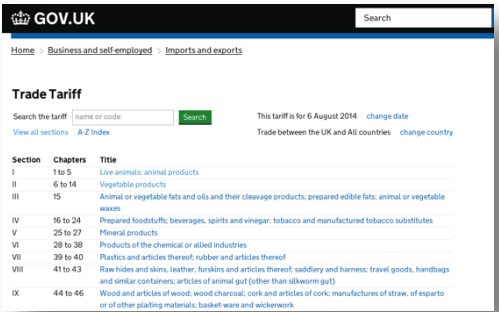
Hierarchical data





A middle ground for data

Other sources of JSON/XML



UK Trade Tariff

BBC Music and Programmes

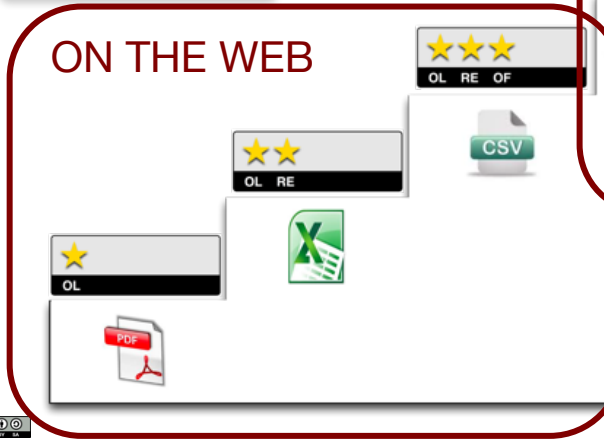
Try using the following: .csv .json .xml .rss .rdf

Linked data

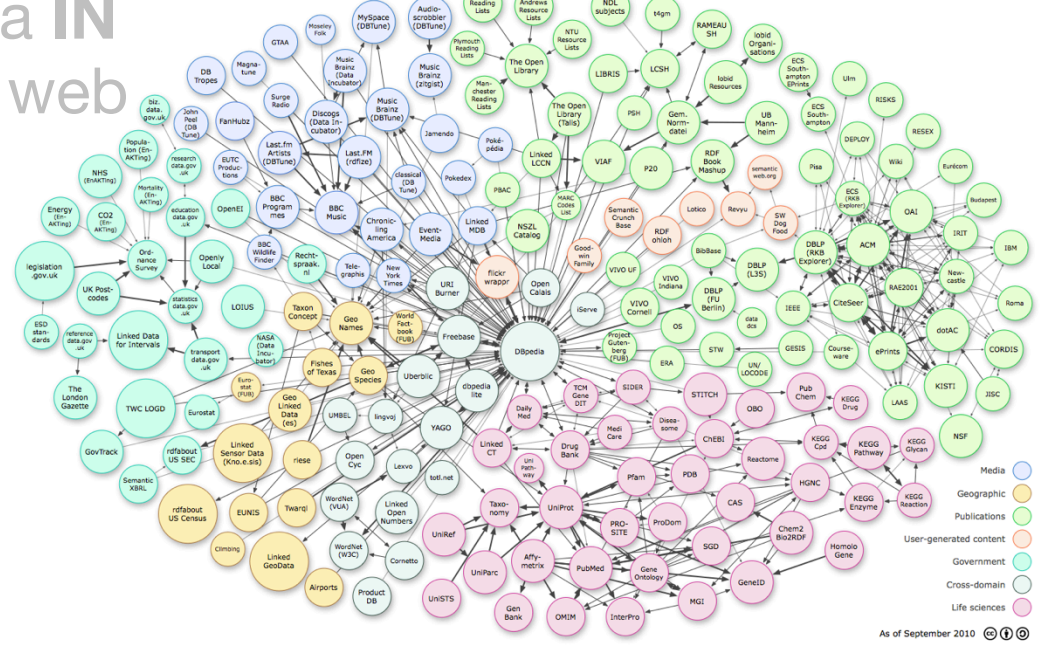
Linked data is pure, but not always useful…

5-Stars

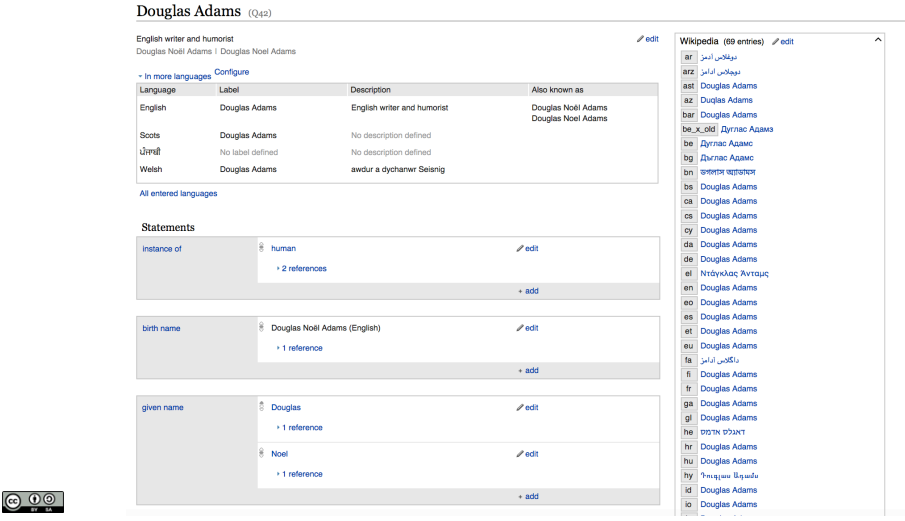
http://5stardata.info/



Data **IN** the web



Wikidata



Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

edit

In more languages Configure

Language	Label	Description	Also known as
English	Douglas Adams	English writer and humorist	Douglas Noël Adams Douglas Noel Adams
Scots	Douglas Adams	No description defined	
ਪੰਜਾਬੀ	No label defined	No description defined	
Welsh	Douglas Adams	awdur a dychanwr Seisnig	

All entered languages

Statements

instance of	human 2 references	edit
		+ add
birth name	Douglas Noël Adams (English) 1 reference	edit
		+ add
given name	Douglas 1 reference	edit
	Noel 1 reference	edit
		+ add

Wikipedia (69 entries) edit

ar	دوغلاس آدمز
arz	دوجلاس ادامز
ast	Douglas Adams
az	Duqlas Adams
bar	Douglas Adams
be_x_old	Дуглас Адамз
be	Дуглас Адамс
bg	Дъглас Адамс
bn	ডগলাস অ্যাডামস
bs	Douglas Adams
ca	Douglas Adams
cs	Douglas Adams
cy	Douglas Adams
da	Douglas Adams
de	Douglas Adams
el	Ντάγκλας Άνταμς
en	Douglas Adams
eo	Douglas Adams
es	Douglas Adams
et	Douglas Adams
eu	Douglas Adams
fa	داگلاس ادامز
fi	Douglas Adams
fr	Douglas Adams
ga	Douglas Adams
gl	Douglas Adams
he	דאגלס אדמס
hr	Douglas Adams
hu	Douglas Adams
hy	Դուգլաս Ադամս
id	Douglas Adams
io	Douglas Adams

Data cleaning

All data is dirty, so you will need to clean it first.

Open refine is a free tool that can help you do this.

Enriching data

Combining data with other data will also be essential in your data science work.

Refine can help if you have the right services available.

Key areas

1) Big Data
2) Data Collection, Preparation and Analysis
3) Machine Learning and Prediction
4) Maths and
5) Interpretat
6) Advanced
7) Business I
8) Open Source Tools and Concepts

As a modern data scientist, 60-80% of your time will be spent preparing data

Session 1

Discovering Open Data Science

Session 2

Machine learning and classification

Session 3

Visualisation and communication

Session 2

Machine learning and classification

Outcomes

Explain the core aspects of Machine Learning

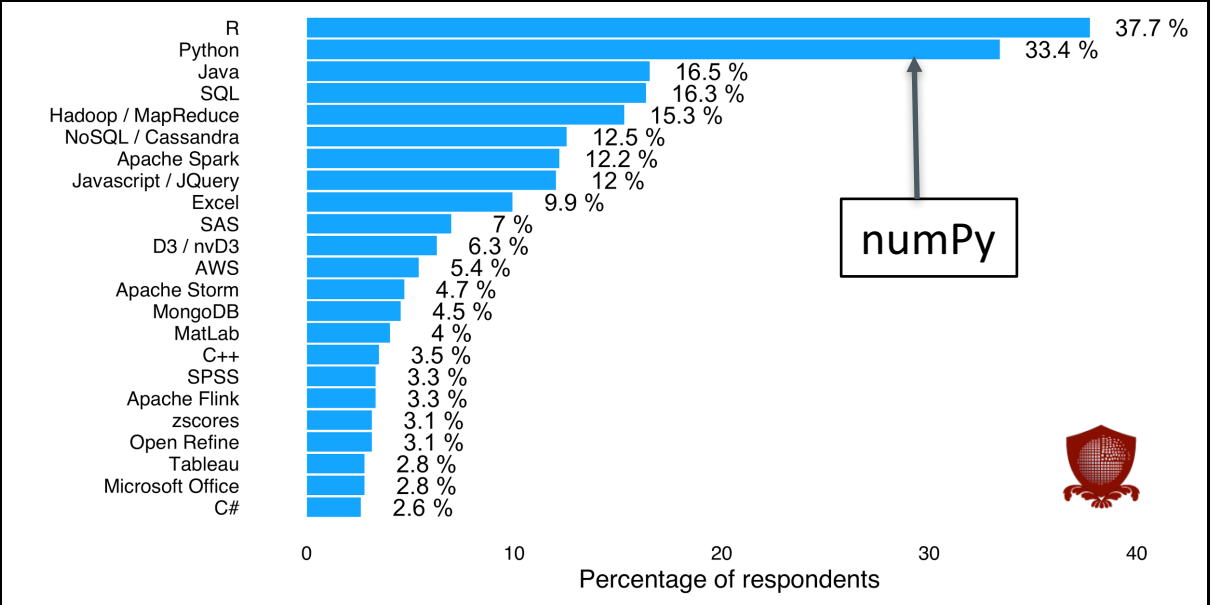
Apply classification to a set of data to create a decision tree

Write a machine learning algorithm

Key areas

1) Big Data
2) Data Collection and Analysis
3) Machine Learning and Prediction
4) Maths and Statistics
5) Interpretation and Visualisation
6) Advanced Computing and Programming
7) Business Intelligence and Domain Expertise
8) Open Source Tools and Concepts

Key skills and tools



Property locations?

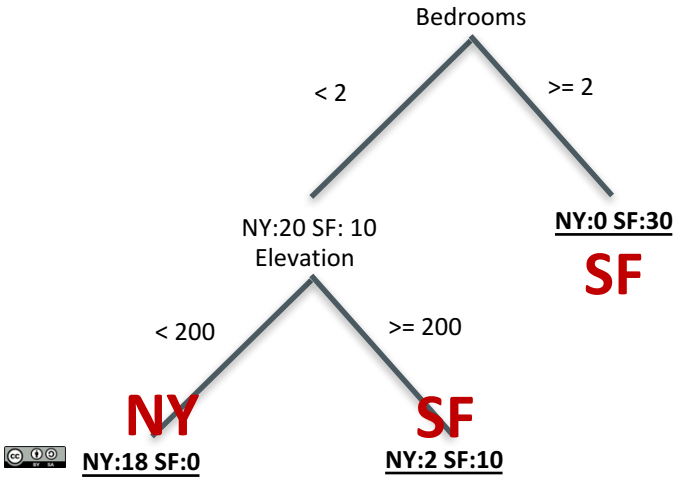
Each table has a set of “Top Trump” training-set cards.

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

You cannot use the target to sort them.

Instructions

Build a decision tree to sort your cards



Confidence

Correct: 18 + 10 + 30

Wrong: 0 + 2 + 0

= 58 / 60

97%

No compound decisions

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

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)

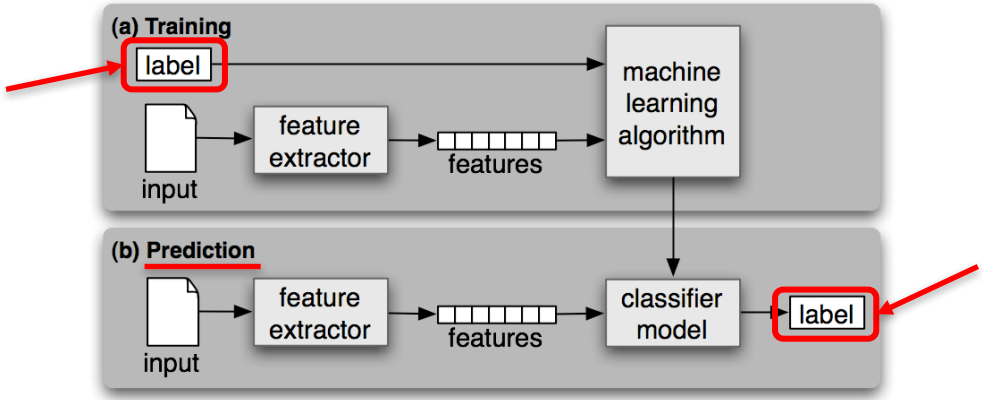
Features:
1. Color: **Radish/Red**
2. Type : **Fruit**
3. Shape
etc…

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.

Categories

Supervised Learning

Unsupervised Learning

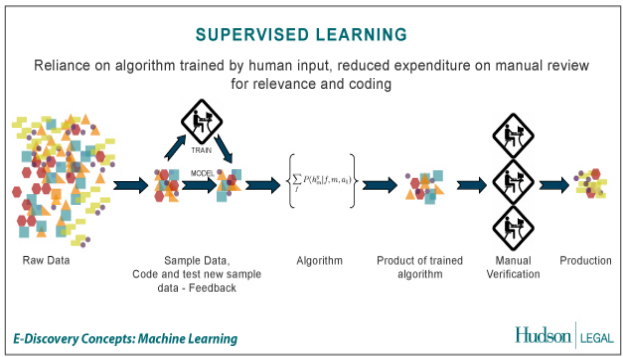
Semi-Supervised Learning

Reinforcement Learning

Slide ideas thanks to @rahuldausa

Supervised Learning

The correct labels of the training data are known

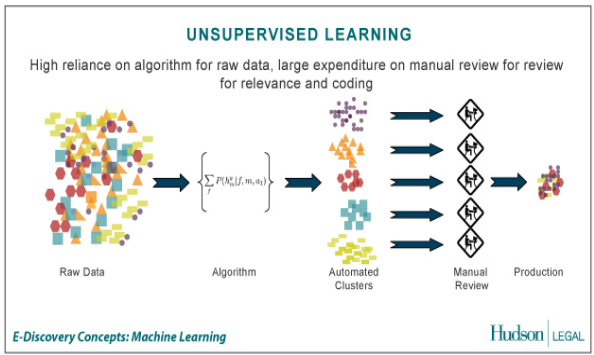


Credit: http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth

Slide ideas thanks to @rahuldausa

Unsupervised Learning

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

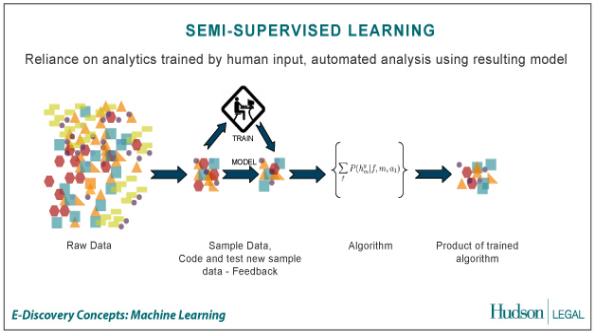


Credit: http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth

Slide ideas thanks to @rahuldausa

Semi-Supervised Learning

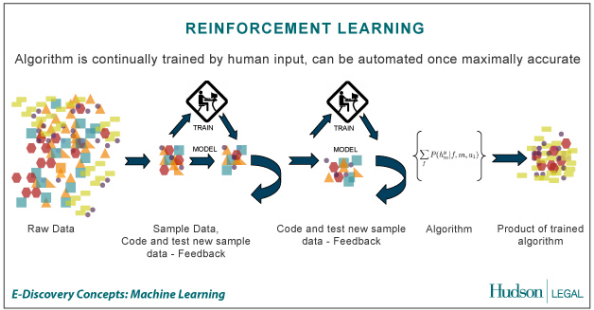
A Mix of Supervised and Unsupervised learning.
Some labels are known, but majority are not.



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

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.)

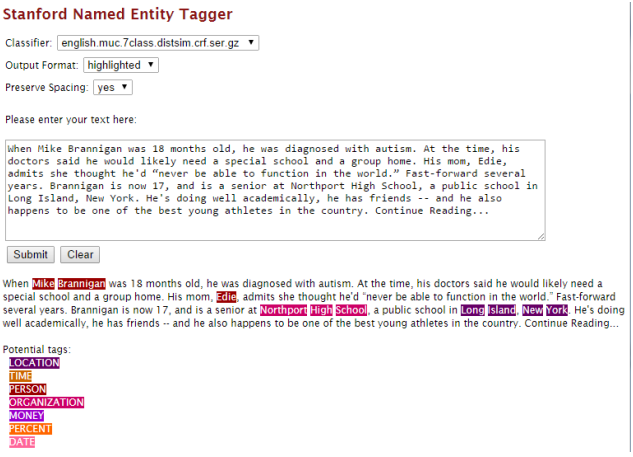
- Text Summarization - Google News
- Rating a Review/Comment: Yelp
- Fraud detection : Credit card Providers
- Decision Making : e.g. Bank/Insurance sector
- Sentiment Analysis
- Speech Understanding – iPhone with Siri
- Face Detection – Facebook’s Photo tagging

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, plea	11:08 am
Scoop.it	Your Scoop.it Daily Summary - How to Maximize Your LinkedIn Publishing Exposure | SME a... - Scoop.it Facebook Twitter G+ H **Not a Spam** e 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. Y **Not a Spam**	Sep 4
Rekha	Mobility Apps for Your Business - While we look at the span of last 30 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

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/

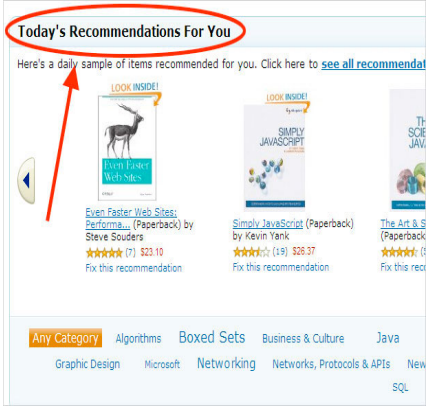
Remember

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/

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…

Exploring R

Improve your classification tree using R

Exploring in R

```
attach(houses)
str(houses)
summary(houses)
plot(houses)
Hist(elevation)
plot(x=price_per_sqft,y=elevation,col=target+1)
boxplot(formula = beds ~ target, data=houses)
Library(rpart)
fit <- rpart(formula = target ~ elevation+ price_per_sqft, data=houses,
method="class")
summary(fit)
plot(fit, uniform=TRUE, main="Classification Tree for Houses")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

Outcomes

Explain the core aspects of Machine Learning

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

Perform a number of statistical experiments on data

Session 1

Discovering Open Data Science

Session 2

Machine learning and classification

Session 3

Visualisation and communication

Session 3

Visualisation and communication

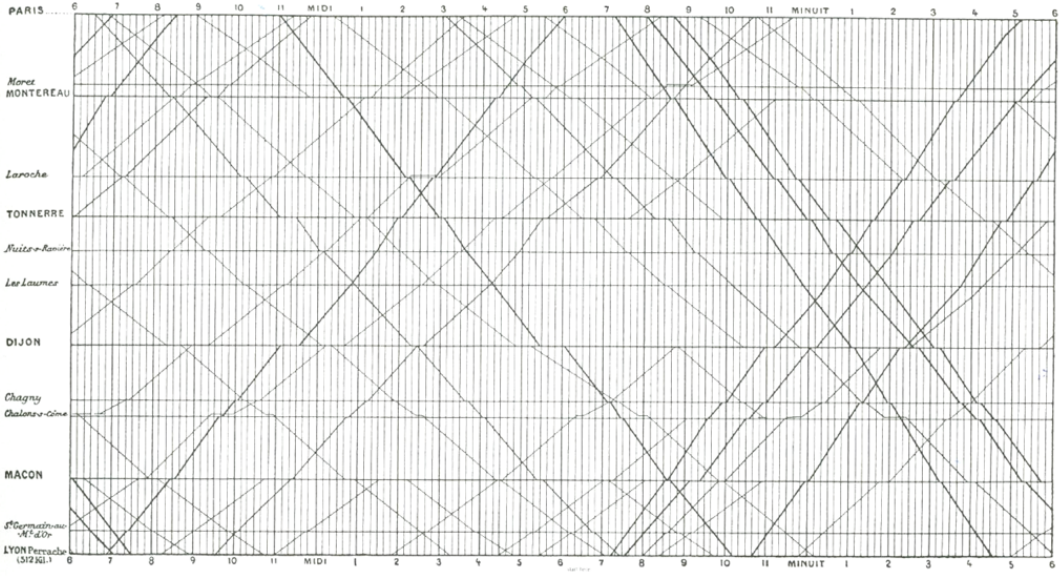
Outcomes

Differentiate between visualising for commication and understanding.

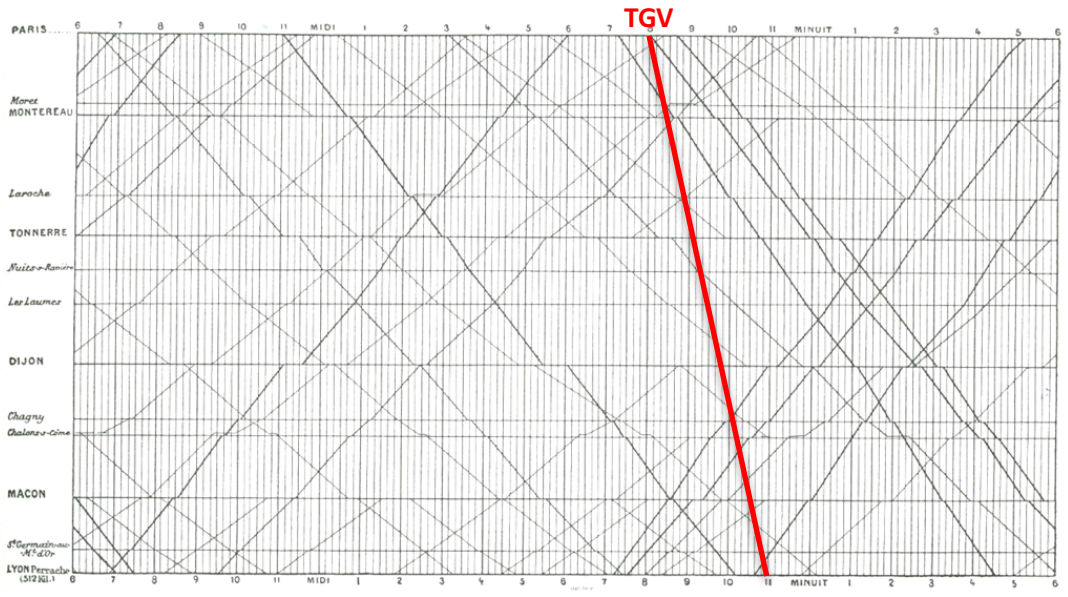
Explain the challenges of visualising complex data.

Create a visualisation to aid understanding of complex data.

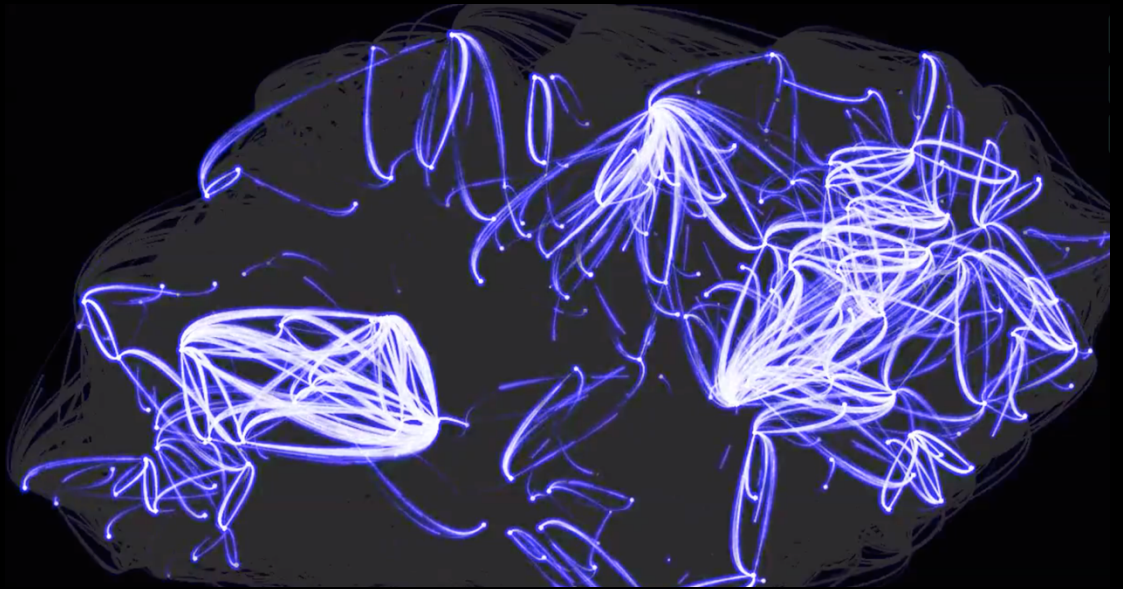

TITANIA
RHEA
CHARON
ARIEL
TETHYS
MIRANDA
MIMAS
OBERON
IAPETUS
UMBRIEL
DIONE
ENCELADUS
PROTEUS
NEREID
EARTH
VENUS
MARS
GANYMEDE
TITAN
MERCURY
CALLISTO
IO
MOON
EUROPA
TRITON
PLUTO



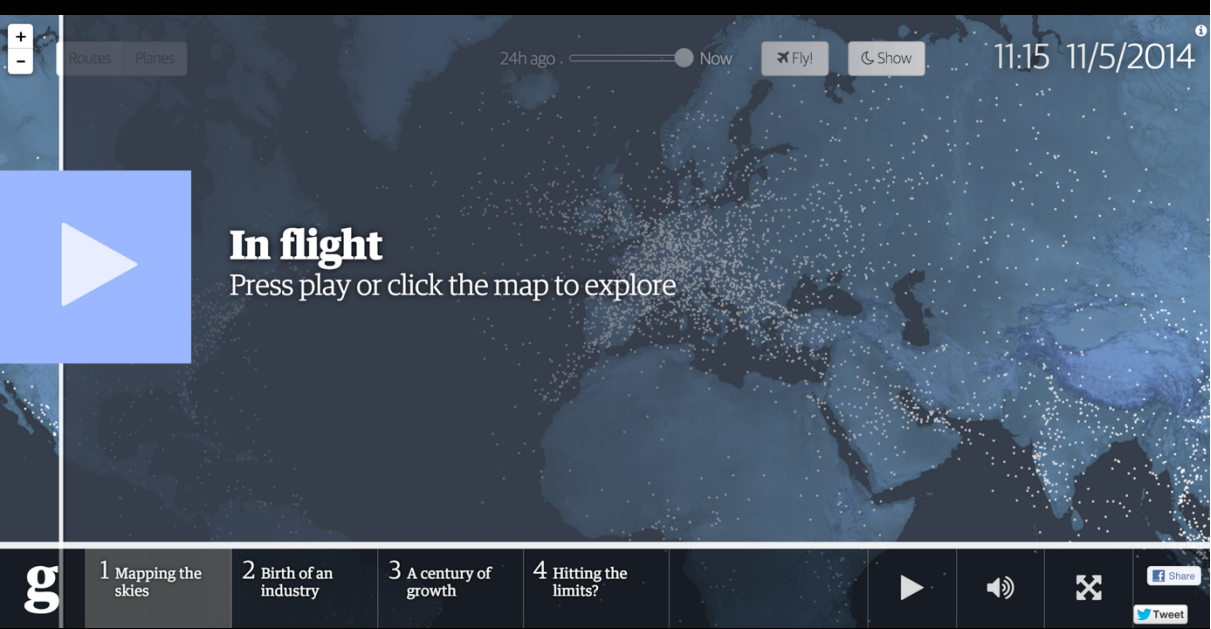
E. J. Marey, *La méthode graphique* (Paris, 1885), 20. The method is attributed to the French engineer, Ibry.



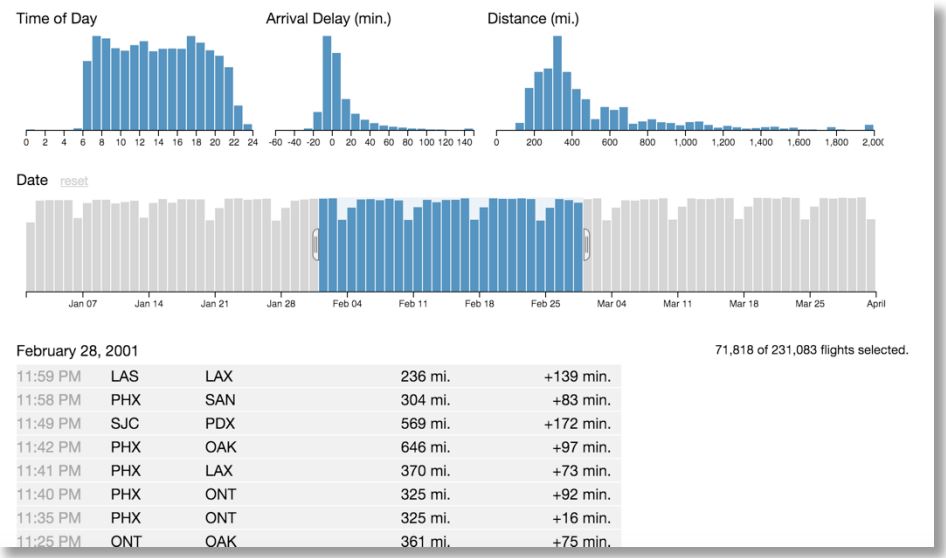
E. J. Marey, *La méthode graphique* (Paris, 1885), 20. The method is attributed to the French engineer, Ibry.



https://vimeo.com/33712288



http://www.theguardian.com/world/ng-interactive/2014/aviation-100-years



http://square.github.io/crossfilter/

Exercise

Building a crossfilter for our houses

TED

Gary Flake:

Is Pivot a turning point for web exploration?

TED2010 · 6:25 · **Filmed** Feb 2010

24 subtitle languages

View interactive transcript

http://www.ted.com/talks/gary_flake_is_pivot_a_turning_point_for_web_exploration?language=en


Thank-you