



# Open Data Science

<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

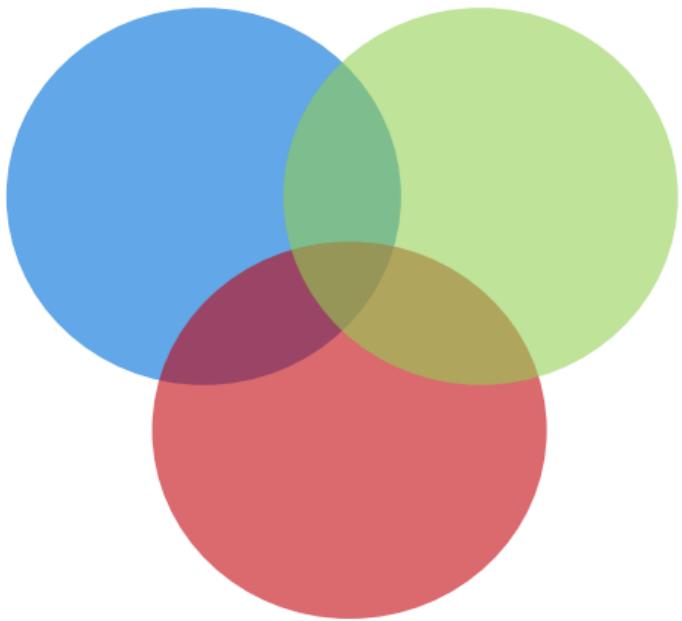
Apply pivot tables on big data in the cloud

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

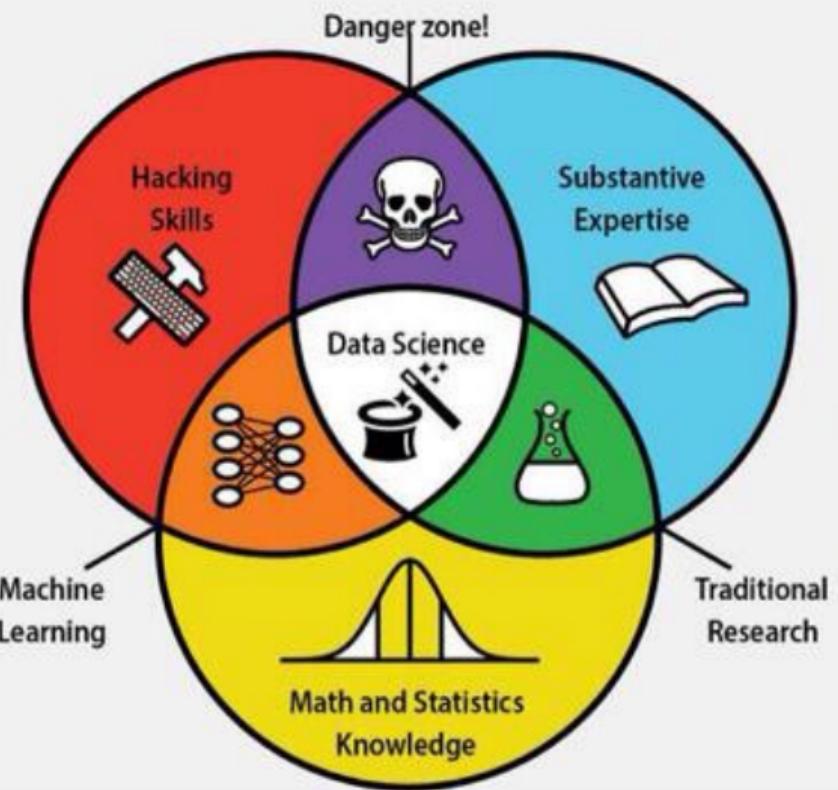


# Before we begin

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



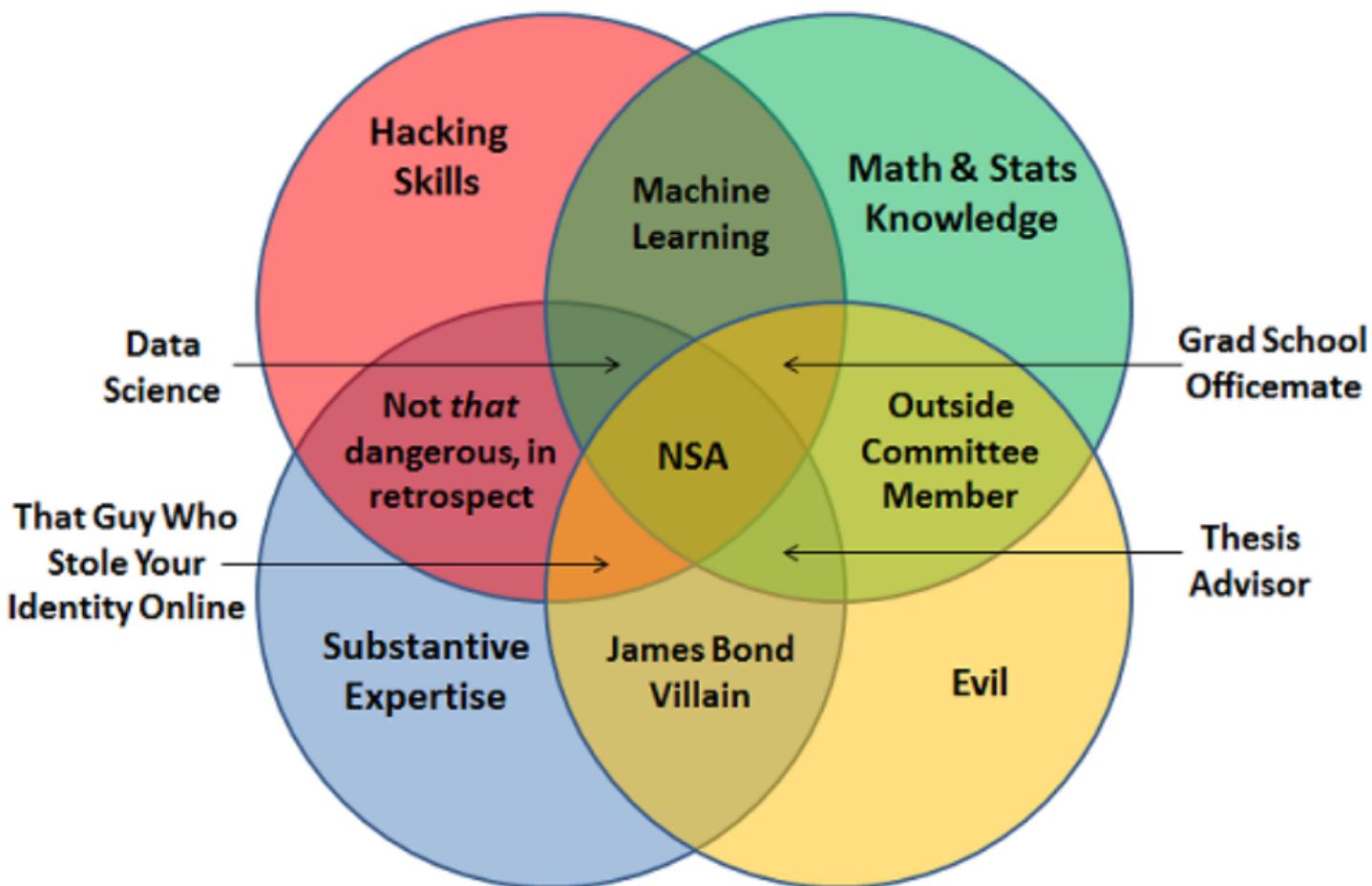
# DATA SCIENCE SKILLSET



	Data science, due to its interdisciplinary nature, requires an intersection of abilities: <b>hacking skills, math and statistics knowledge</b> , and <b>substantive expertise</b> in a field of science.
	<b>Hacking skills</b> are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.
	<b>Math and statistics knowledge</b> allows a data scientist to choose appropriate methods and tools in order to extract insight from data.
	<b>Substantive expertise</b> in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.
	<b>Traditional research</b> lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.
	<b>Machine learning</b> stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.
	<b>Danger zone!</b> Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.



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



# Data Science

“part analyst, part artist”

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

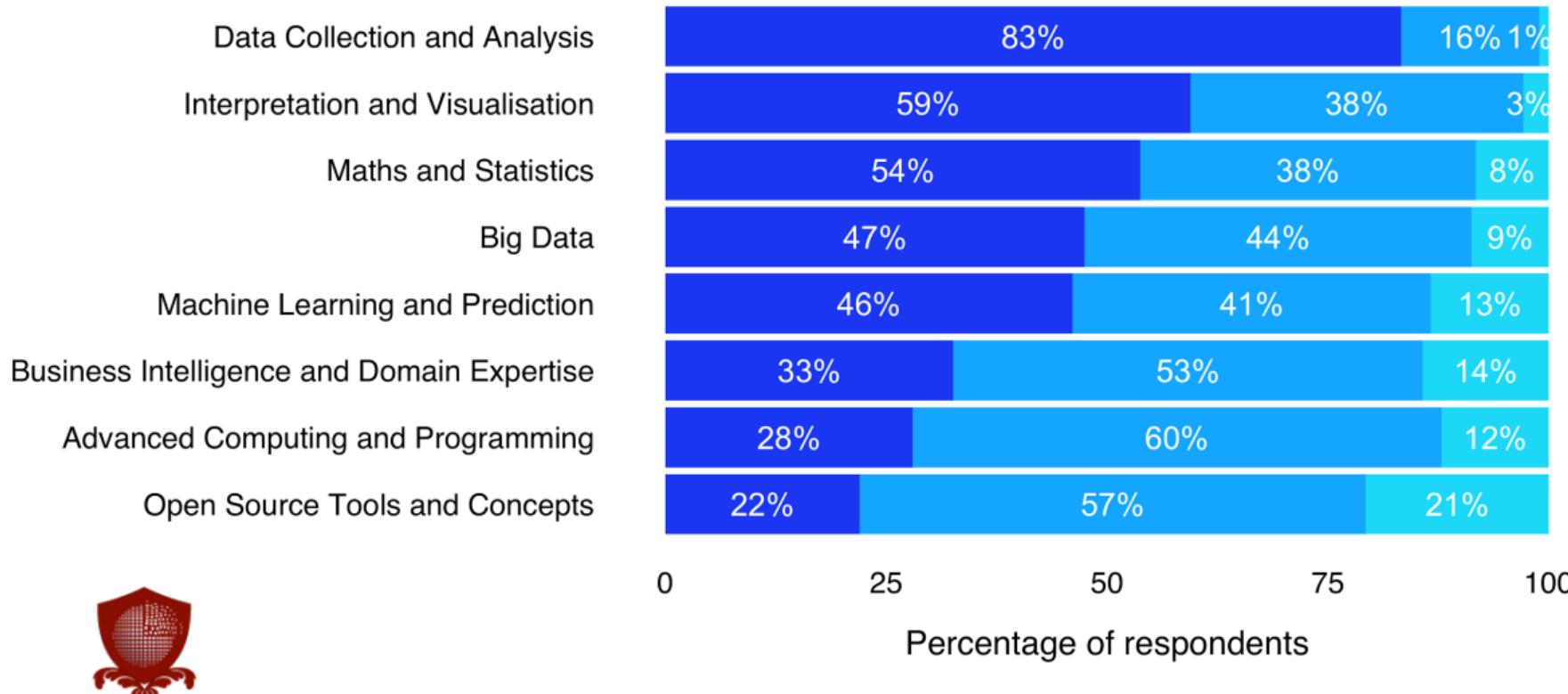


# Key areas

- 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

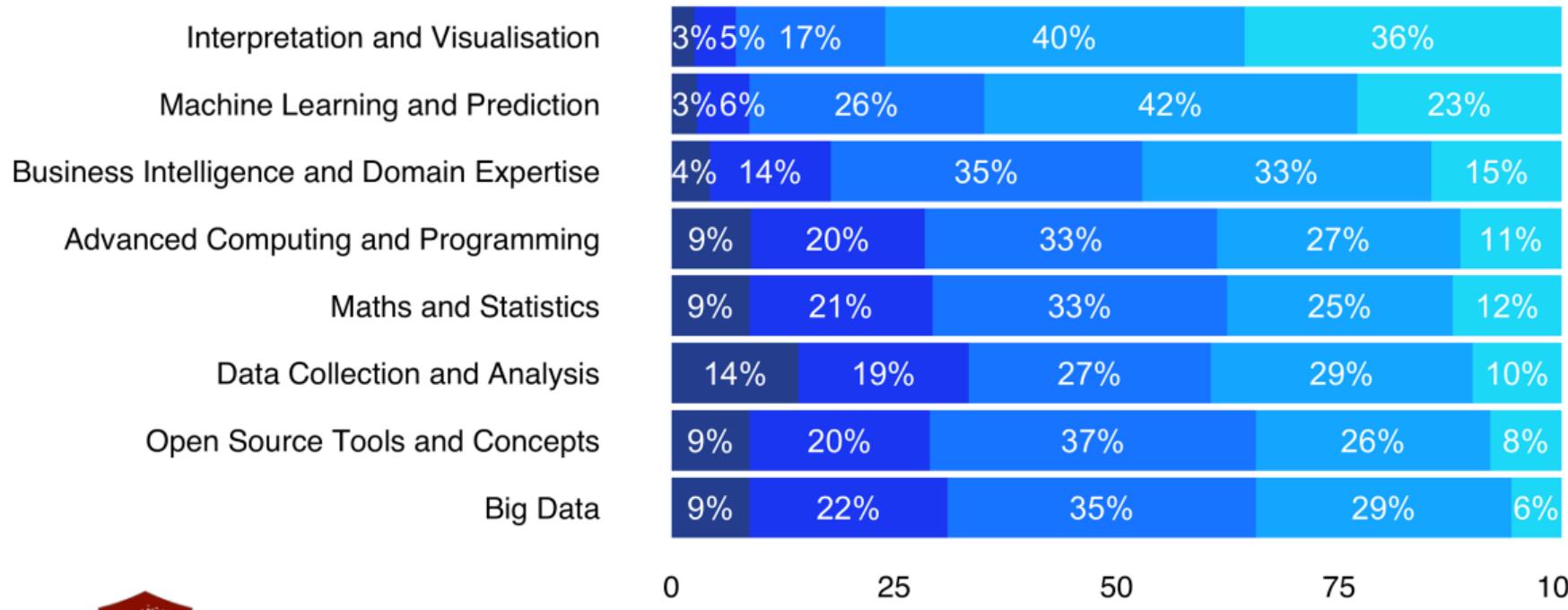


## *Skills that a data scientist should have (N = 633)*



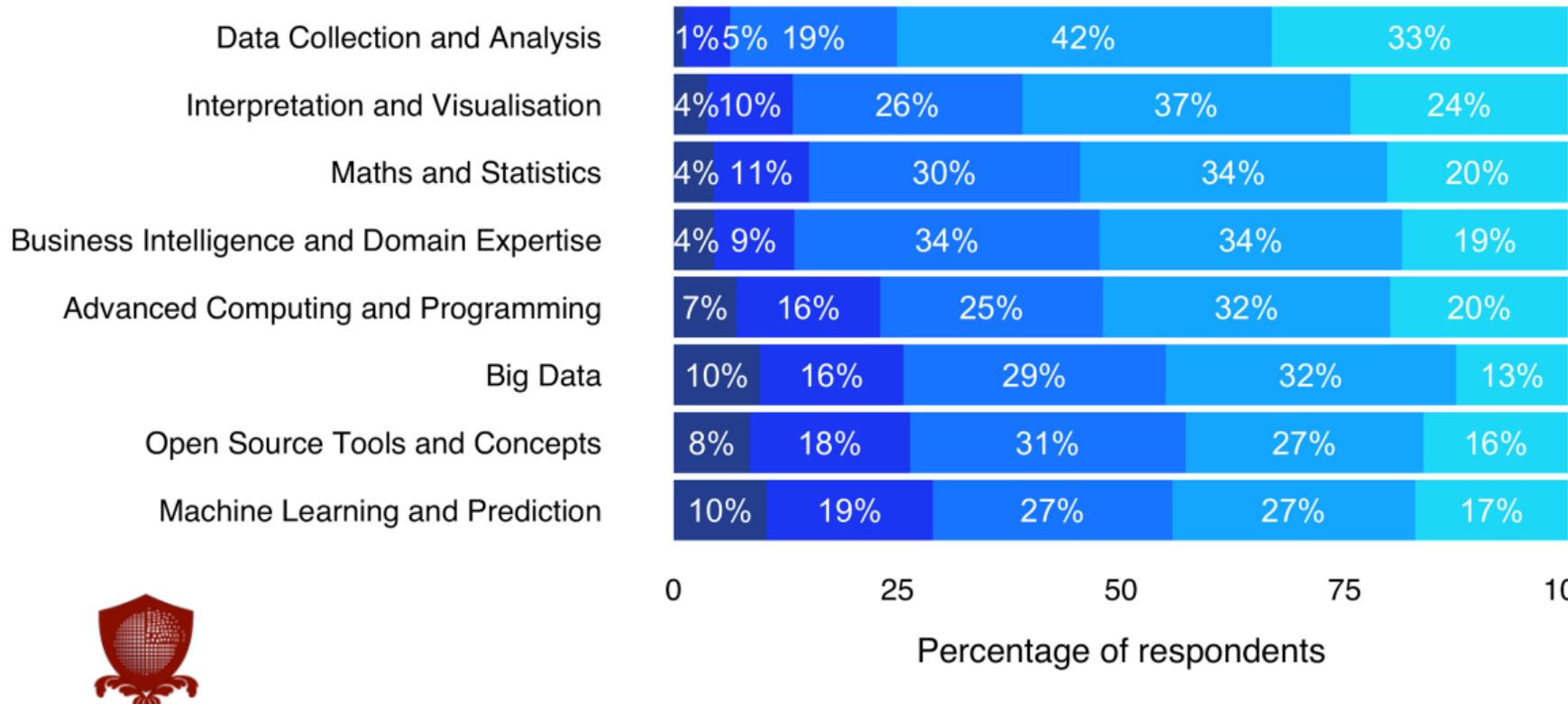
Response ■ Essential ■ Desirable ■ Not required

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



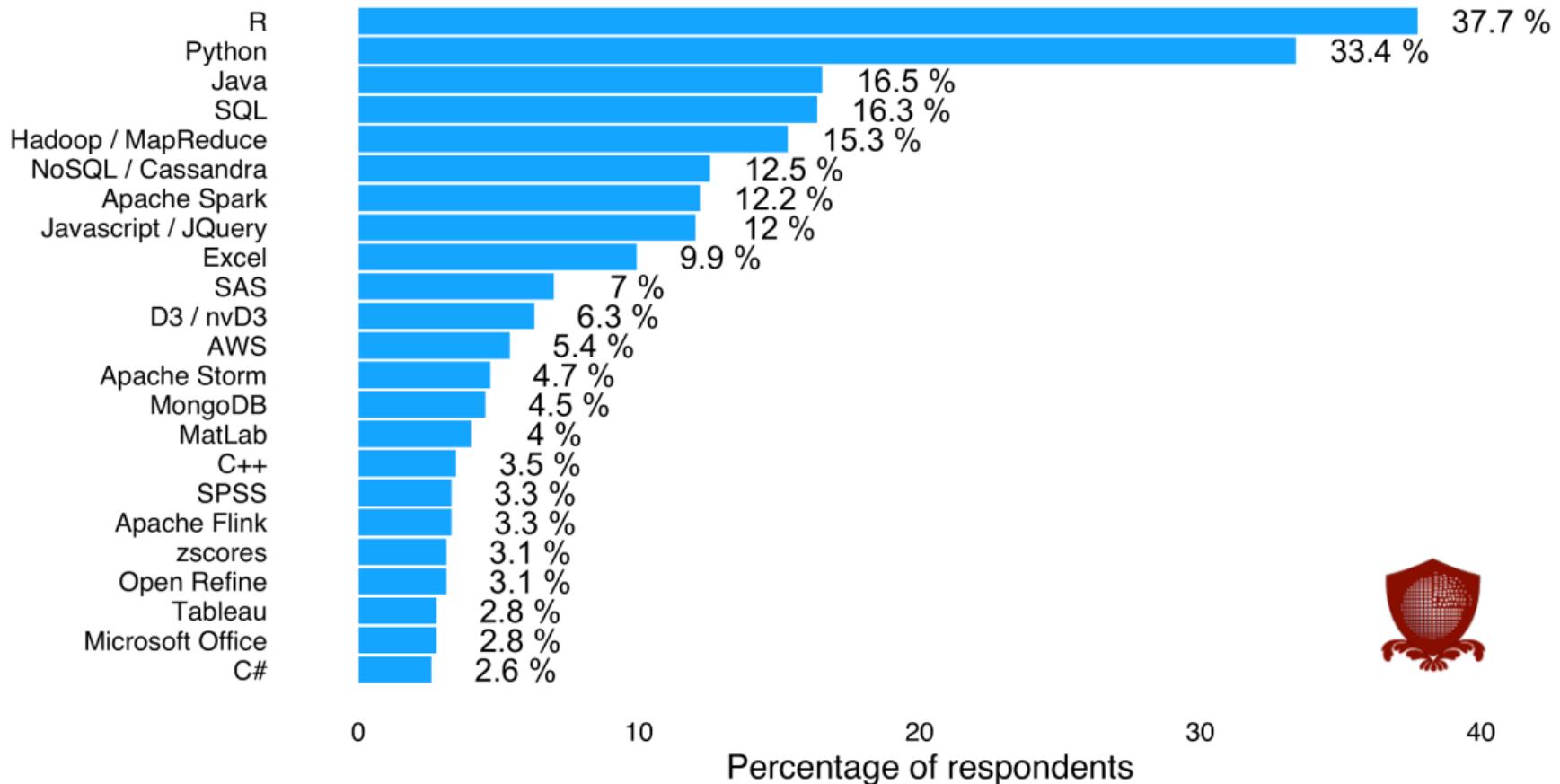
Score ■ 1 ■ 2 ■ 3 ■ 4 ■ 5

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



Score ■ 1 ■ 2 ■ 3 ■ 4 ■ 5

## *Key skills and tools*



# Key areas

## 1) Big Data

2) Machine Learning and Prediction

3) Data Engineering and ETL

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

6) Application Development and Programming

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

The screenshot shows the DATA.GOV.UK homepage with a search bar and navigation links for Home, Data, Apps, Interact, Datasets, Map Search, Data Requests, Publishers, Public Roles & Salaries, Spend Reports, Site Analytics, and Reports. A sidebar shows a map of the United States and Mexico with a marker for Phoenix. The main search results page displays 19266 Results for "Live traffic information from the Highways Agency". It includes a snippet about live traffic information data from the Highways Agency and a link to "Published datasets (5186)". Below the search results is a section for "OpenData Burkina Faso" with themes like Agriculture, Education, Diplomacy, Infrastructures, Environment, TIC, Health, Territories, Security, Tourism, and Culture.

The screenshot shows the DATOS.GOB.MX BETA website with a header for "OPEN TO PARTICIPATE". It features a map of Mexico and Central America with a marker for Phoenix. The main content area shows a search bar with the placeholder "Buscar conjuntos de datos...", a list of "127 datasets found" (sorted by Relevance), and a section titled "RATING SOCIAL PROGRAMS" describing a database of social program evaluations. The footer includes links for "Data", "Stories", "Advances", "Datasets", "Organizations", and "Portal Rasmi Open Data Kerajaan Malaysia" (The Government of Malaysia's Open Data Official Portal).

The screenshot shows the data.gov.my website with a header for "HOME", "ABOUT US", "CATALOGUE", "INFOGRAPHICS", "MOBILE APPS", and "CONTACT". It features sections for "DATASETS INFO" (117 Datasets, 11 Ministries, 10 Sectors), "RESOURCES" (Malaysia Directory, Malaysian Open Source Centre, Open Government Data), "MOBILE APPS" (myHealth | myJalan | KPONIK SPAD | SELAWAT | myMAHTAS, 6 Android Mobile Apps, 4 iOS Mobile Apps), and "QUICK LINKS" (Malaysia Informative Data Centre, Malaysian Population Quick Info, Tourism Malaysia - Facts & Figures). The "Latest News and Events" section indicates there are no latest events.

# Google advanced

Google site:gov filetype:xls

Web Images Maps Shopping More Search tools

About 4,150,000 results (0.22 seconds)

[XLS] [Code List or Concept \(Acronym\)](#)

[www.acquisition.gov/short\\_codelistsTS.xls](http://www.acquisition.gov/short_codelistsTS.xls) Share  
File Format: Microsoft Excel - [View as HTML](#)  
A, B, C, D, F, G, H, 1, Code List or Concept (Acronym), Definition, Authoritative or Maintenance Agency, Current Version, Update Frequency and/or ...

[XLS] [Approps - Foreign Assistance.gov](#)

[www.foreignassistance.gov/Full\\_ForeignAssistanceData.xls](http://www.foreignassistance.gov/Full_ForeignAssistanceData.xls)  
File Format: Microsoft Excel  
A, B, C, D, E, F, G, H, 1, Planning Data, 2, 3, 4, Fiscal Year, Fiscal Year Type, Account Name, Agency Name, Operating Unit, Category, Sector, Amount ...

[XLS] [TSB Monthly Cash Flow Projection](#)

[www.dia.iowa.gov/tsbselocashflow.xls](http://www.dia.iowa.gov/tsbselocashflow.xls)

**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

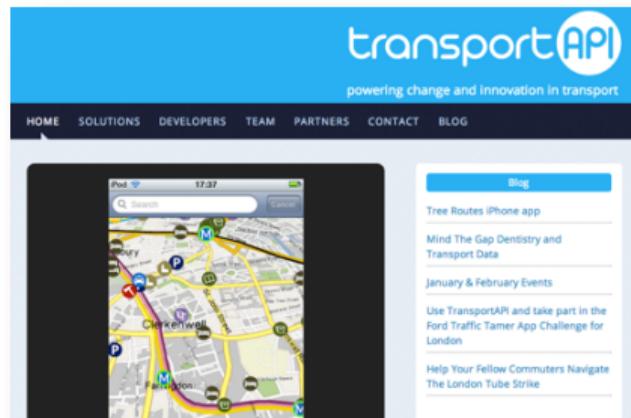


# Aggregators and portals

Collect together data from across the web into one place.



enigma.io



transportAPI



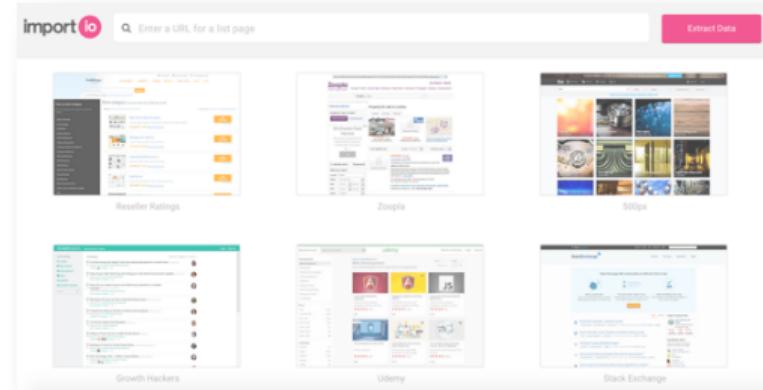
# Scraping

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



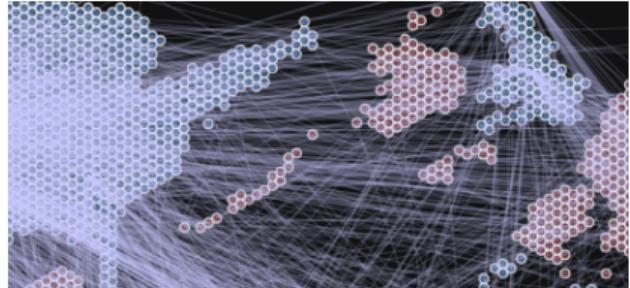
“excellent, so excited beyond description”  
*George Ofosu, Doctoral Student, UCLA*

[pdftables.com](http://pdftables.com)



[magic.import.io](http://magic.import.io)

# Hierarchical data

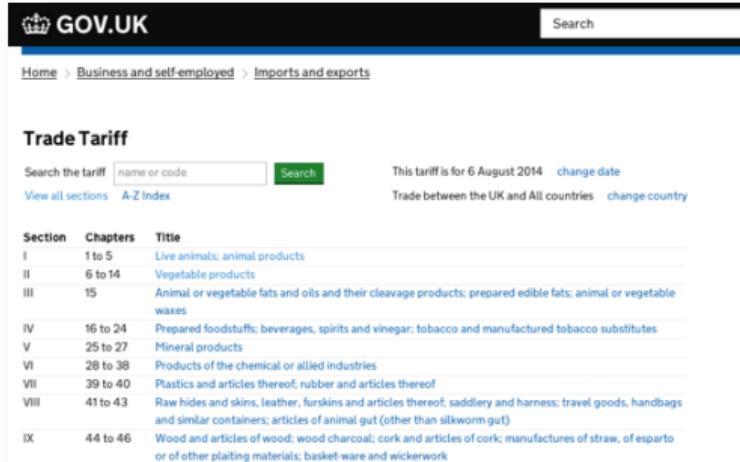


A screenshot of the OpenCorporates website. At the top, the logo 'opencorporates' is displayed in a stylized font with five dots below the 'ates'. Below the logo, the text 'The largest open database of companies in the world' is shown. A search bar contains the placeholder 'Search 108,723,509 companies'. To the right of the search bar is a dropdown menu set to 'All jurisdictions' with a downward arrow. A teal-colored search button with a white magnifying glass icon is positioned to the right of the dropdown. At the bottom of the search area, there are two links: 'Browse all jurisdictions' and 'Search officers'.

A middle ground  
for data



# Other sources of JSON/XML



The screenshot shows the GOV.UK website for the Trade Tariff. At the top, there's a search bar and a navigation menu with links to Home, Business and self-employed, and Imports and exports. Below this, the title "Trade Tariff" is displayed. A search bar allows users to search by name or code, with a "Search" button. To the right, it says "This tariff is for 6 August 2014" and "change date". Below the search bar, there are links to "View all sections" and "A-Z Index". The main content area is titled "Section Chapters Title" and lists categories from I to IX. Category I includes "Live animals; animal products" and "Vegetable products". Category II includes "Animal or vegetable fats and oils and their cleavage products; prepared edible fats; animal or vegetable waxes". Category III includes "Prepared foodstuffs; beverages, spirits and vinegar; tobacco and manufactured tobacco substitutes". Category IV includes "Mineral products". Category V includes "Products of the chemical or allied industries". Category VI includes "Plastics and articles thereof; rubber and articles thereof". Category VII includes "Raw hides and skins; leather, furskins and articles thereof; saddlery and harness; travel goods, handbags and similar containers; articles of animal gut (other than silkworm gut)". Category VIII includes "Wood and articles of wood; wood charcoal; cork and articles of cork; manufactures of straw, of esparto or of other plaiting materials; basket-ware and wickerwork". Category IX includes "Food, drink and their愁制 products; live animals; animal products".

UK Trade Tariff



The screenshot shows the BBC Music and Programmes website for the TV show Doctor Who. The header features the BBC One logo and the show's title "DOCTOR WHO" with the TARDIS icon. Below the header, there are navigation links for Home, Episodes, Clips, Galleries, Latest News, Characters, Monsters, Fun and Games, and More. The main content area has two sections: "On iPlayer" which shows a thumbnail of the Doctor and a woman, and "On TV" which shows a thumbnail of three Doctors. Below these, there's a promotional message for "The Day of the Doctor" with broadcast details: SATURDAY 19:00 BBC THREE, All upcoming (0 NEW AND 1 REPEAT).

BBC Music and Programmes

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



# Linked data

Linked data is pure,  
but not always  
useful...

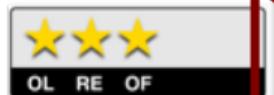


# 5-Stars



<http://5stardata.info/>

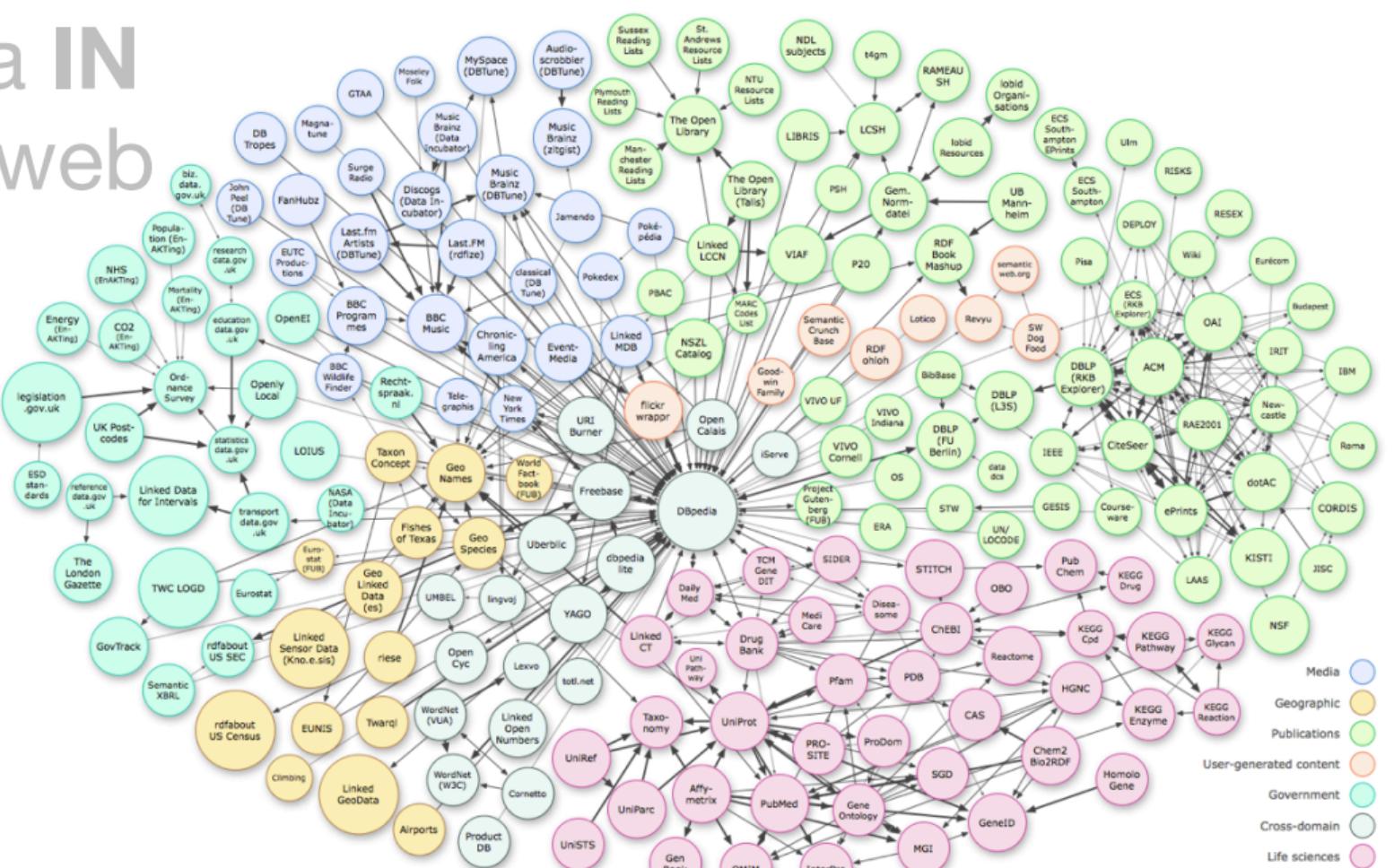
## ON THE WEB



## IN THE WEB



# Data IN the web



# Wikidata

## Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

- 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	<a href="#">human</a> <a href="#">edit</a> <a href="#">+ 2 references</a>
	<a href="#">+ add</a>

birth name	<a href="#">Douglas Noël Adams (English)</a> <a href="#">edit</a> <a href="#">+ 1 reference</a>
	<a href="#">+ add</a>

given name	<a href="#">Douglas</a> <a href="#">edit</a> <a href="#">+ 1 reference</a>
	<a href="#">Noel</a> <a href="#">edit</a> <a href="#">+ 1 reference</a>

[Wikipedia \(69 entries\)](#) [edit](#)

<a href="#">ar</a> <a href="#">دونالد آدمز</a>
<a href="#">arz</a> <a href="#">دونالد آدمز</a>
<a href="#">ast</a> <a href="#">Douglas Adams</a>
<a href="#">az</a> <a href="#">Duqas Adams</a>
<a href="#">bar</a> <a href="#">Douglas Adams</a>
<a href="#">be_x-old</a> <a href="#">Дуглас Адамз</a>
<a href="#">be</a> <a href="#">Дуглас Адамс</a>
<a href="#">bg</a> <a href="#">Дъглас Адамс</a>
<a href="#">bn</a> <a href="#">ডাউগলাস আডামস</a>
<a href="#">bs</a> <a href="#">Douglas Adams</a>
<a href="#">ca</a> <a href="#">Douglas Adams</a>
<a href="#">cs</a> <a href="#">Douglas Adams</a>
<a href="#">cy</a> <a href="#">Douglas Adams</a>
<a href="#">da</a> <a href="#">Douglas Adams</a>
<a href="#">de</a> <a href="#">Douglas Adams</a>
<a href="#">el</a> <a href="#">Ντόγκλας Άντομς</a>
<a href="#">en</a> <a href="#">Douglas Adams</a>
<a href="#">eo</a> <a href="#">Douglas Adams</a>
<a href="#">es</a> <a href="#">Douglas Adams</a>
<a href="#">et</a> <a href="#">Douglas Adams</a>
<a href="#">eu</a> <a href="#">Douglas Adams</a>
<a href="#">fa</a> <a href="#">دانکل آدمز</a>
<a href="#">fi</a> <a href="#">Douglas Adams</a>
<a href="#">fr</a> <a href="#">Douglas Adams</a>
<a href="#">ga</a> <a href="#">Douglas Adams</a>
<a href="#">gl</a> <a href="#">Douglas Adams</a>
<a href="#">he</a> <a href="#">דואגלס אדמס</a>
<a href="#">hr</a> <a href="#">Douglas Adams</a>
<a href="#">hu</a> <a href="#">Douglas Adams</a>
<a href="#">hy</a> <a href="#">Դուգլաս Ադամս</a>
<a href="#">id</a> <a href="#">Douglas Adams</a>
<a href="#">io</a> <a href="#">Douglas Adams</a>



# 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 Statistics
  - 5) Interpretation and Communication
  - 6) Advanced Analytics
  - 7) Business Intelligence
  - 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

## 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

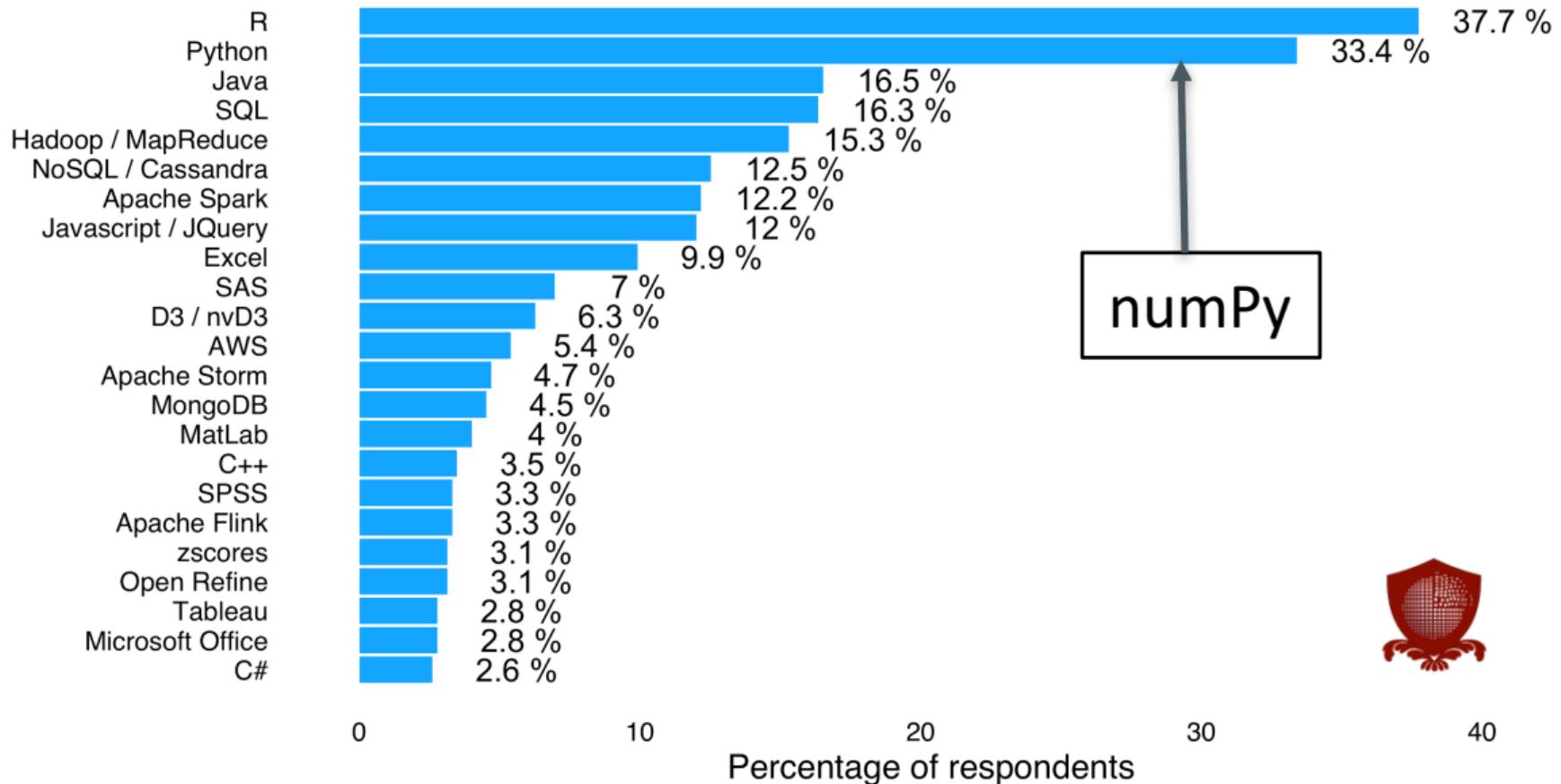


# Key areas

- 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.

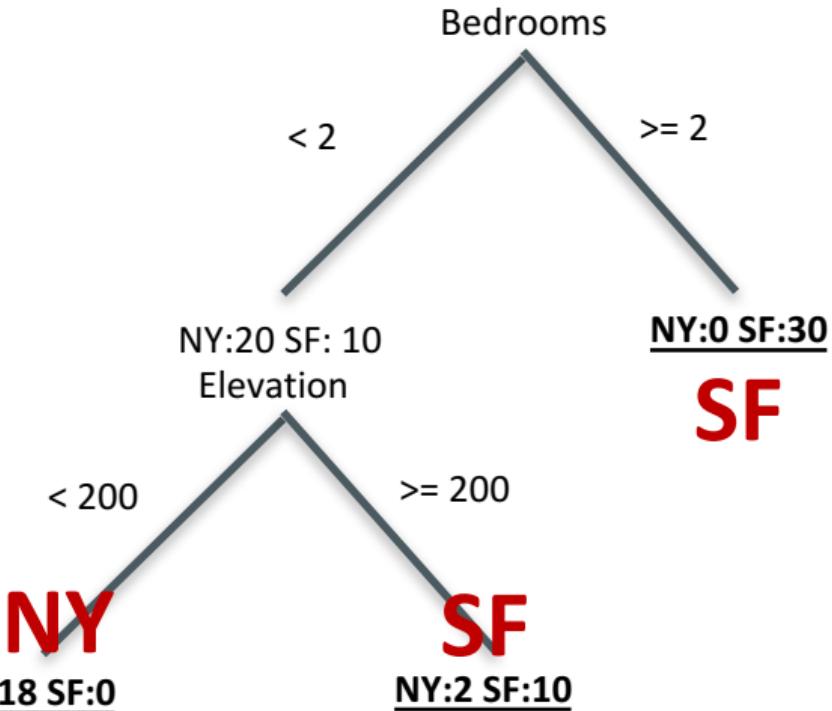
Build a decision tree to sort them into “New York” and “San Francisco”.

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*.



# 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.



# Learning (Training)



Features:

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



Features:

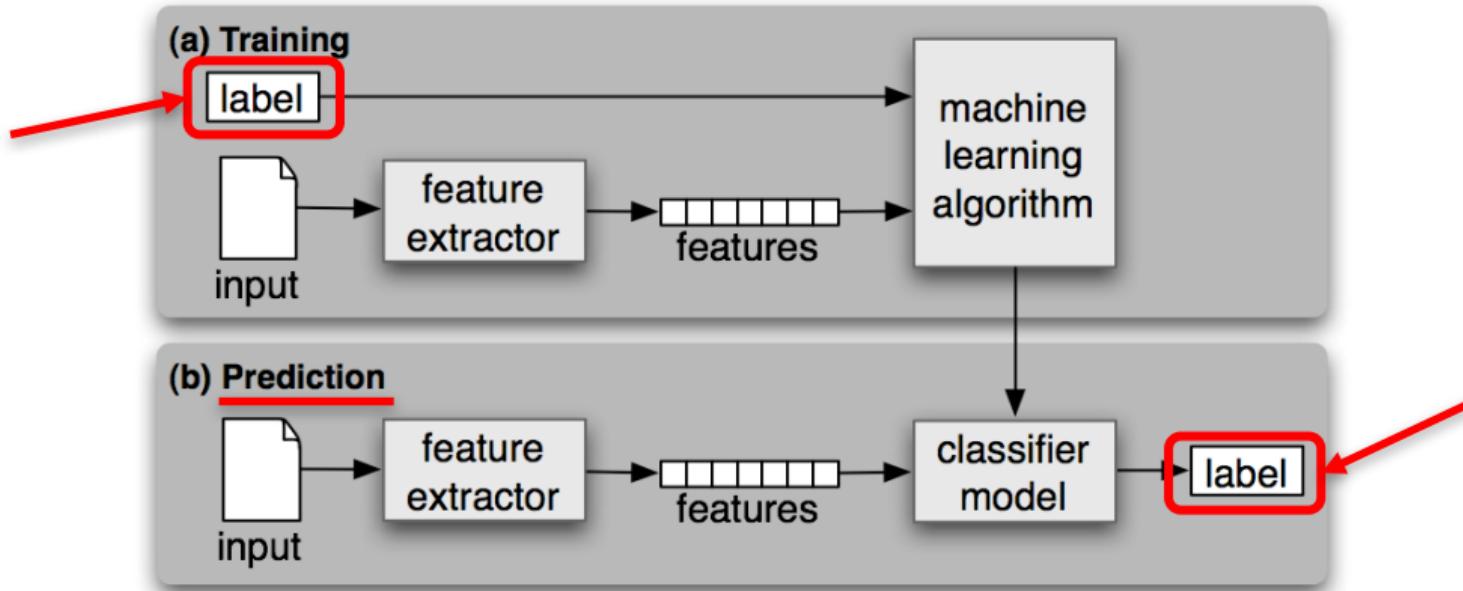
1. Sky Blue
2. **Logo**
3. Shape
- etc...



Features:

1. **Yellow**
2. **Fruit**
3. Shape
- etc...

# Workflow



# 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

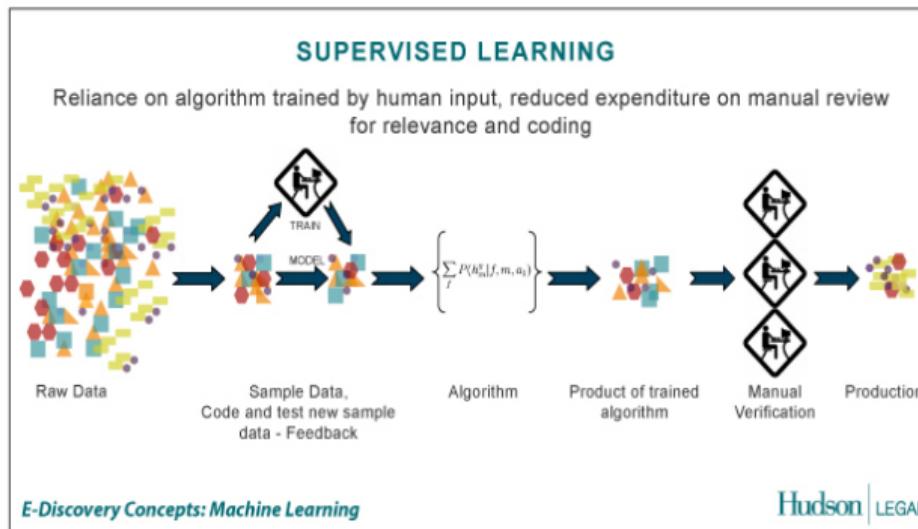
Semi-Supervised Learning

Reinforcement Learning



# Supervised Learning

The correct labels of the training data are known



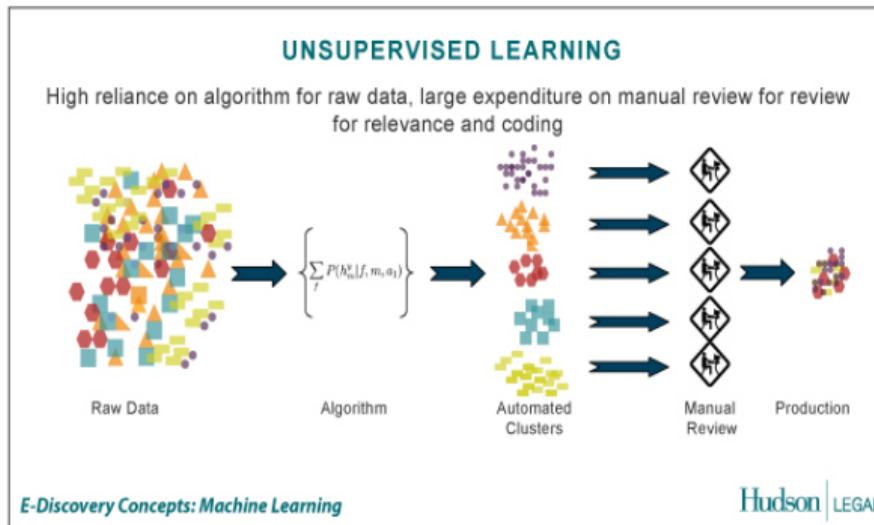
Credit: <http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth>



Slide ideas thanks to @rahuldausa

# Unsupervised Learning

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



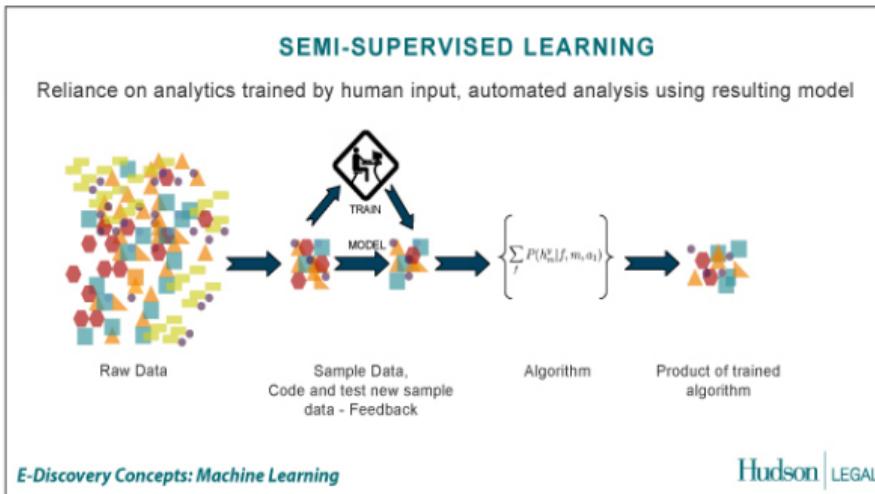
Credit: <http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth>



Slide ideas thanks to @rahuldausa

# Semi-Supervised Learning

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

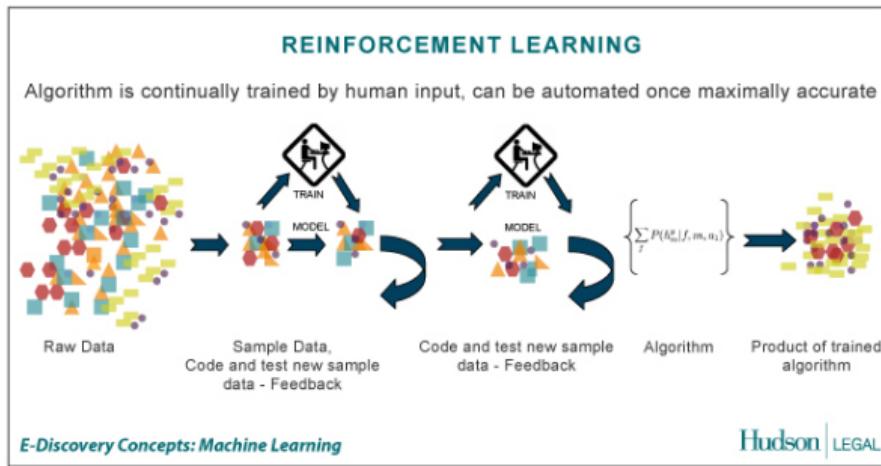


Credit: <http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth>



# Reinforcement Learning

Software agent learns based on feedback and/or reward.



# 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)

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	HD <b>F</b> C 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
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	iEntry	Welcome iEntry Member - Ultimate Guide To Assessing	9:23 pm
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	New-Zealand-Jobs.067L	Come to New Zealand to find a great job and settle here ( <a href="#">Search for all Jobs from differ...</a> - Search for all Jobs from different kinds of industries Find a Job in Enchanting	8:18 pm
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Supermarket Promotion	Enjoy Rs.1700 voucher valid at any supermarket! - If you are unable to view this mailer Click here <a href="#">HOW TO CONTACT US? BY EMAIL:</a> support@savethedeals.in	4:51 pm
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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, 21	1:24 pm
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Max Bupa	A policy that understande your family's medical need - open in fresh tab – If you do not want to receive any more newsletters, please click here	11:08 am
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Scoop.it	Your Scoop.it Daily Summary - How to Maximize Your LinkedIn Publishing Exposure   SME a... - Scoop.it Facebook Twitter G+ H	9:30 am
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	standard charterer Bank	Instant approval on your Credit Card	7:27 am
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Uday	VPS Web Hosting Services Provider - Dear Sir, I am Uday Sharma, Business development executive. We are providing quality VPS hosting for	Sep 4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Mark Regan, SPN	How to Find Your Most Valuable Keywords [ <a href="#">Free Guide</a> ] - This is a SiteProNews/ExactSeek Webmaster Exclusive Mailing! To drop your subscription, use the link	Sep 4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	HD <b>F</b> C 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 HD <b>F</b> C Bank because you	Sep 4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Jayde	Welcome To The Jayde Newsletter! Welcome To The Jayde Newsletter! Before we begin, make sure to add	Sep 4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	ineedhits noreply	[ineedhits] Your ineedhits Account and Password - ACCOUNT CREATION Account ID : A1588368 Dear Rah, Welcome to ineedhits. Y	Sep 4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Rekha	Moving Apps for Your Business - Will you be able to move to the top of the list? You could easily look at the distinct areas - Click in	Sep 4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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)

## Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz ▾

Output Format: highlighted ▾

Preserve Spacing: yes ▾

Please enter your text here:

When Mike Brannigan was 18 months old, he was diagnosed with autism. At the time, his doctors said he would likely need a special school and a group home. His mom, Edie, admits she thought he'd "never be able to function in the world." Fast-forward several years. Brannigan is now 17, and is a senior at Northport High School, a public school in Long Island, New York. He's doing well academically, he has friends -- and he also happens to be one of the best young athletes in the country. Continue Reading...

When **Mike Brannigan** was 18 months old, he was diagnosed with autism. At the time, his doctors said he would likely need a special school and a group home. His mom, **Edie**, admits she thought he'd "never be able to function in the world." Fast-forward several years. Brannigan is now 17, and is a senior at **Northport High School**, a public school in **Long Island, New York**. He's doing well academically, he has friends -- and he also happens to be one of the best young athletes in the country. Continue Reading...

Potential tags:

LOCATION

TIME

PERSON

ORGANIZATION

MONEY

PERCENT

DATE



# 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](#)

## 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/>

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



# Recommendations

The screenshot displays two main sections of an Amazon product page:

**More Items to Consider:**

- You looked at:** JavaScript: The Good Parts Paperback by Douglas Crockford (\$29.99 \$19.79)
- You might also consider:** JavaScript: The Definitive Guide Paperback by David Flanagan (\$49.99 \$31.49)

**Related to Items You've Viewed:**

- You looked at:** Forms that Work: Designing Web Forms... Paperback
- You might also consider:** Don't Make Me Think: A Common Sense Approach to Web Usability Paperback by Steve Krug

**Today's Recommendations For You:**

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#)

- Even Faster Web Sites: Performance Optimization for Web Developers Paperback by Steve Souders
- Simply JavaScript (Paperback) by Kevin Yank
- The Art & Science of Java (Paperback)

Category filters at the bottom include: Any Category, Algorithms, Boxed Sets, Business & Culture, Java, Graphic Design, Microsoft, Networking, Networks, Protocols & APIs, New SQL.



# 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

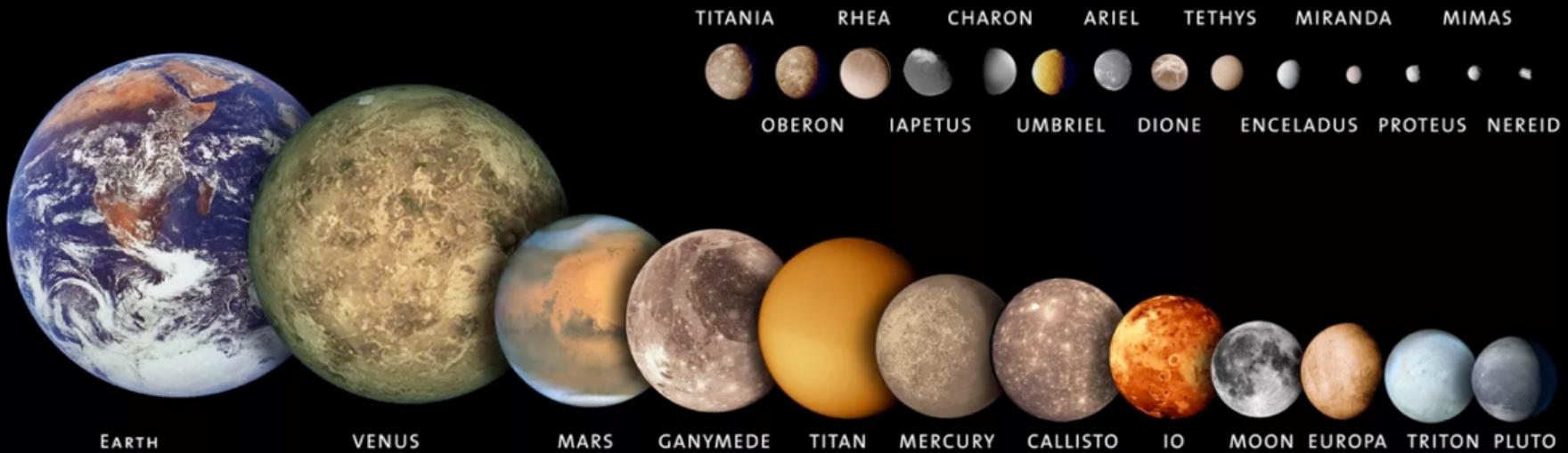
## Visualisation and communication

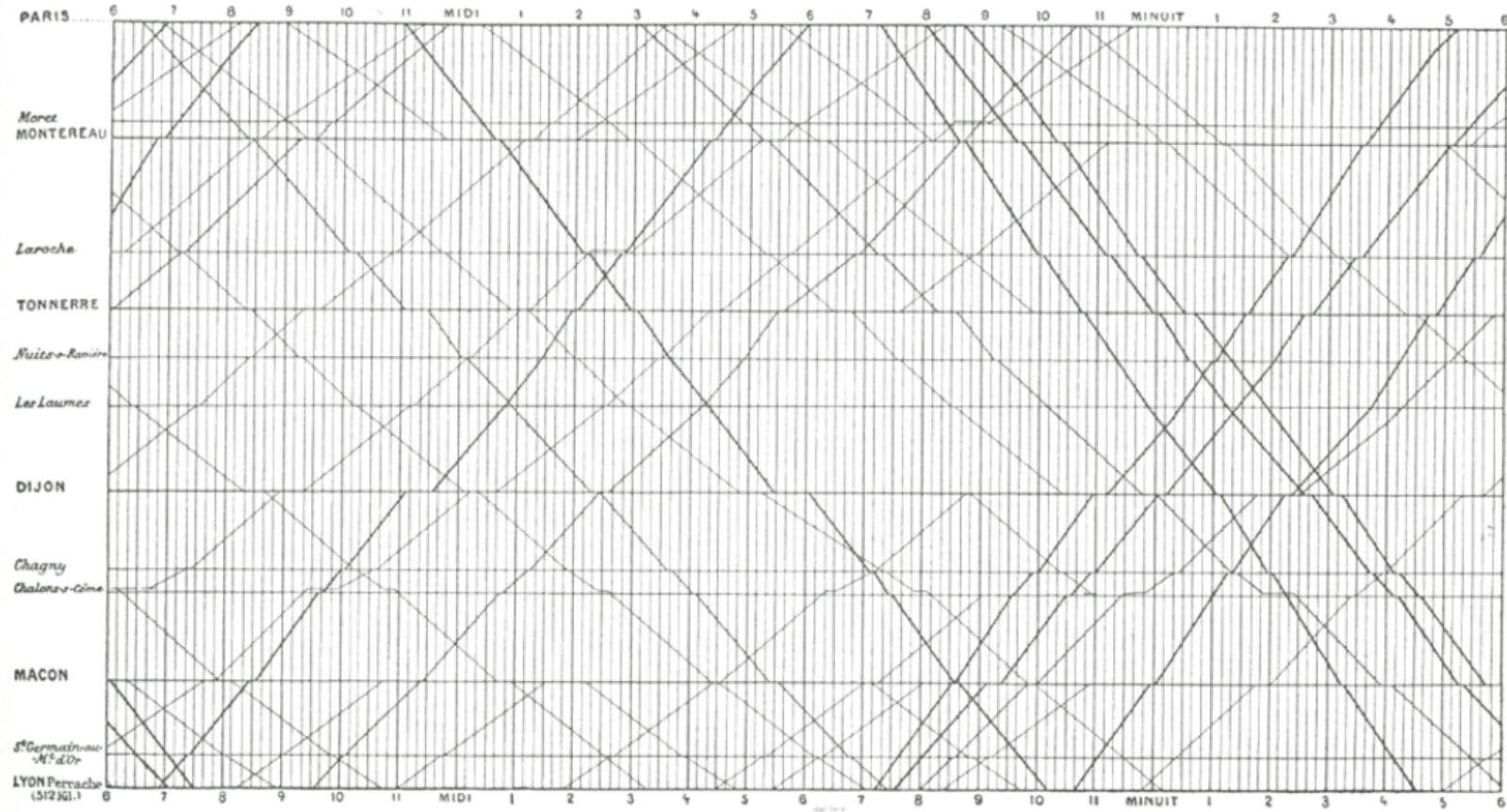


# Outcomes

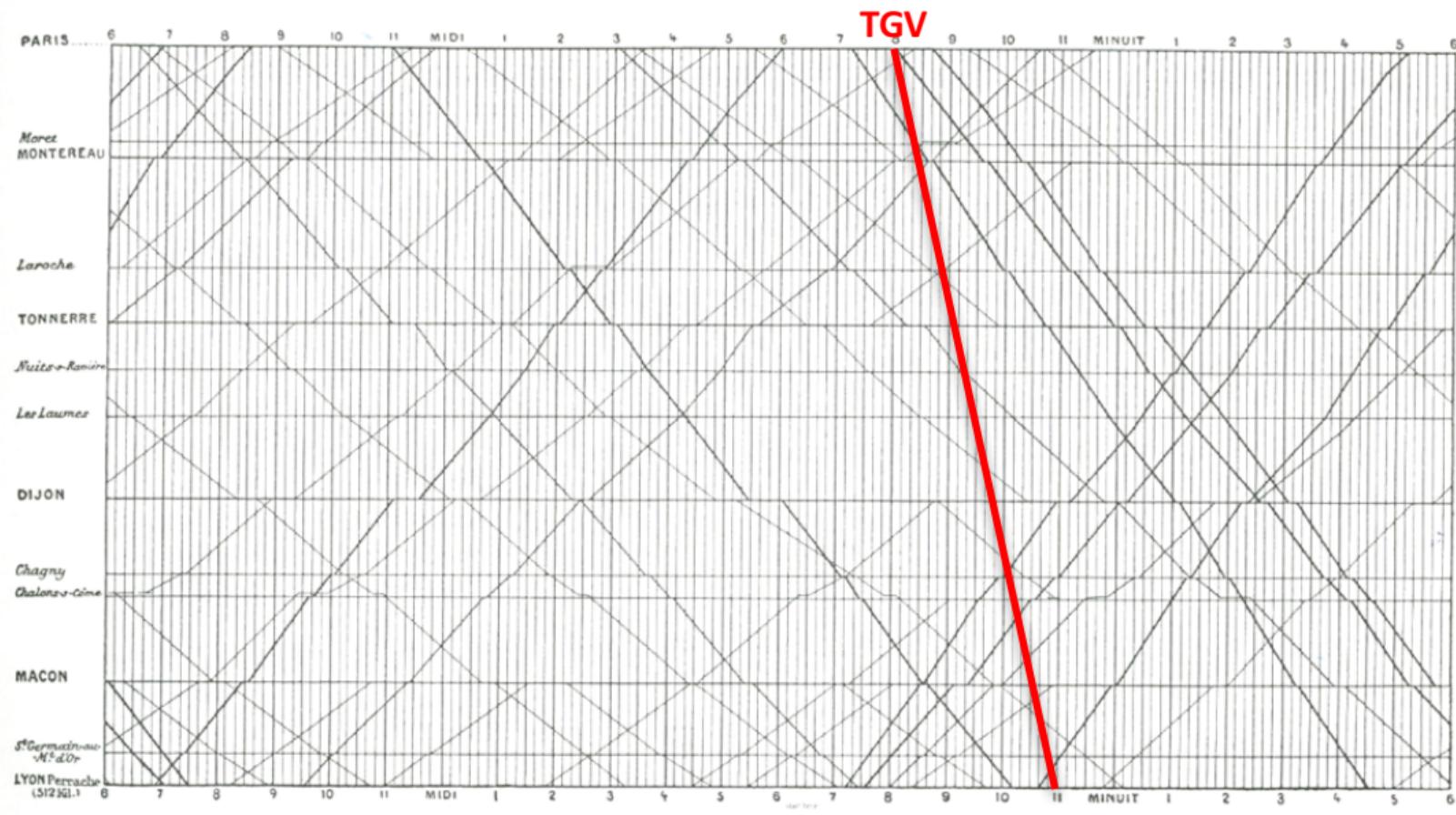
Differentiate between visualising for communication and understanding.  
Explain the challenges of visualising complex data.  
Create a visualisation to aid understanding of complex data.



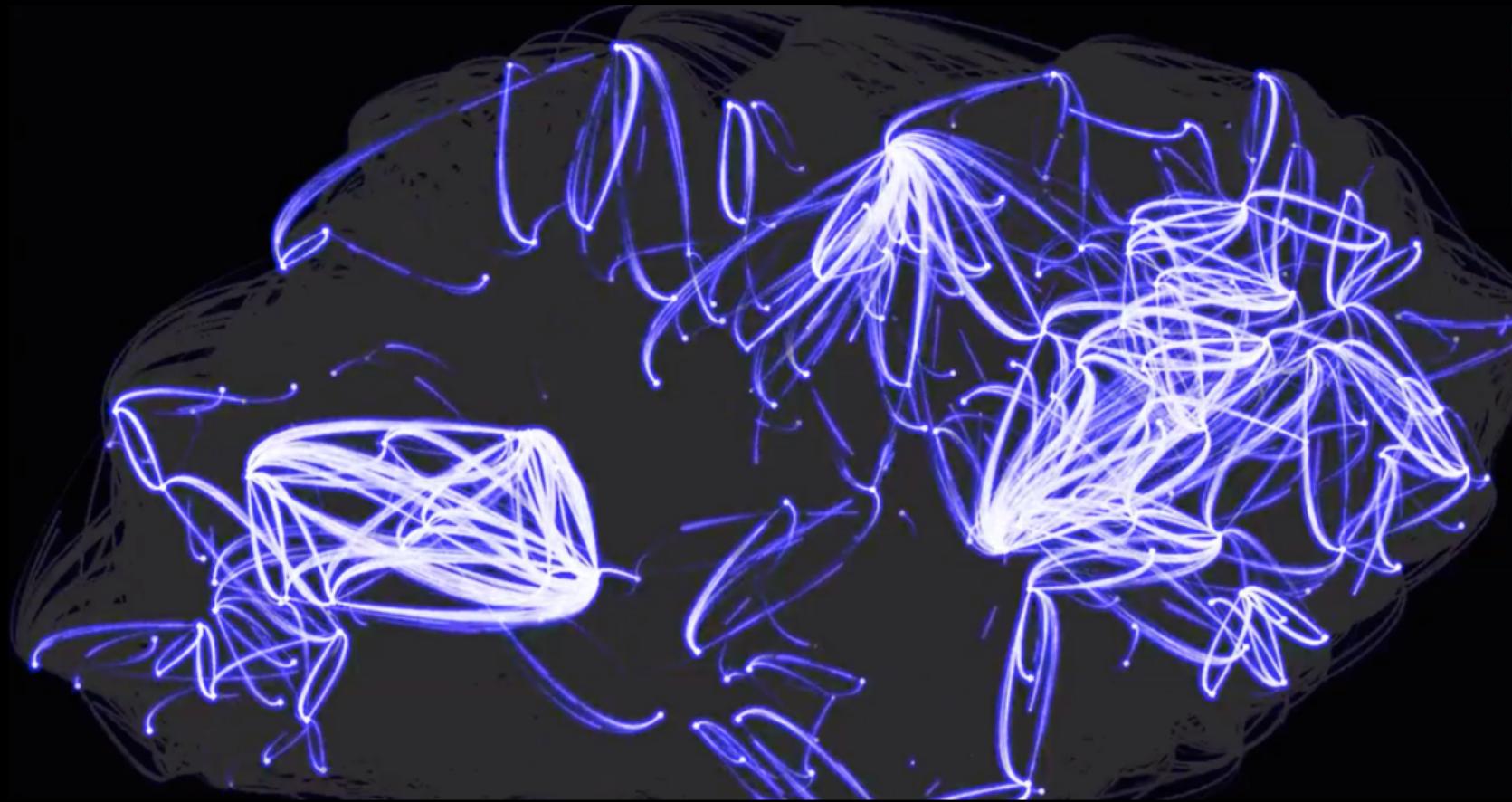




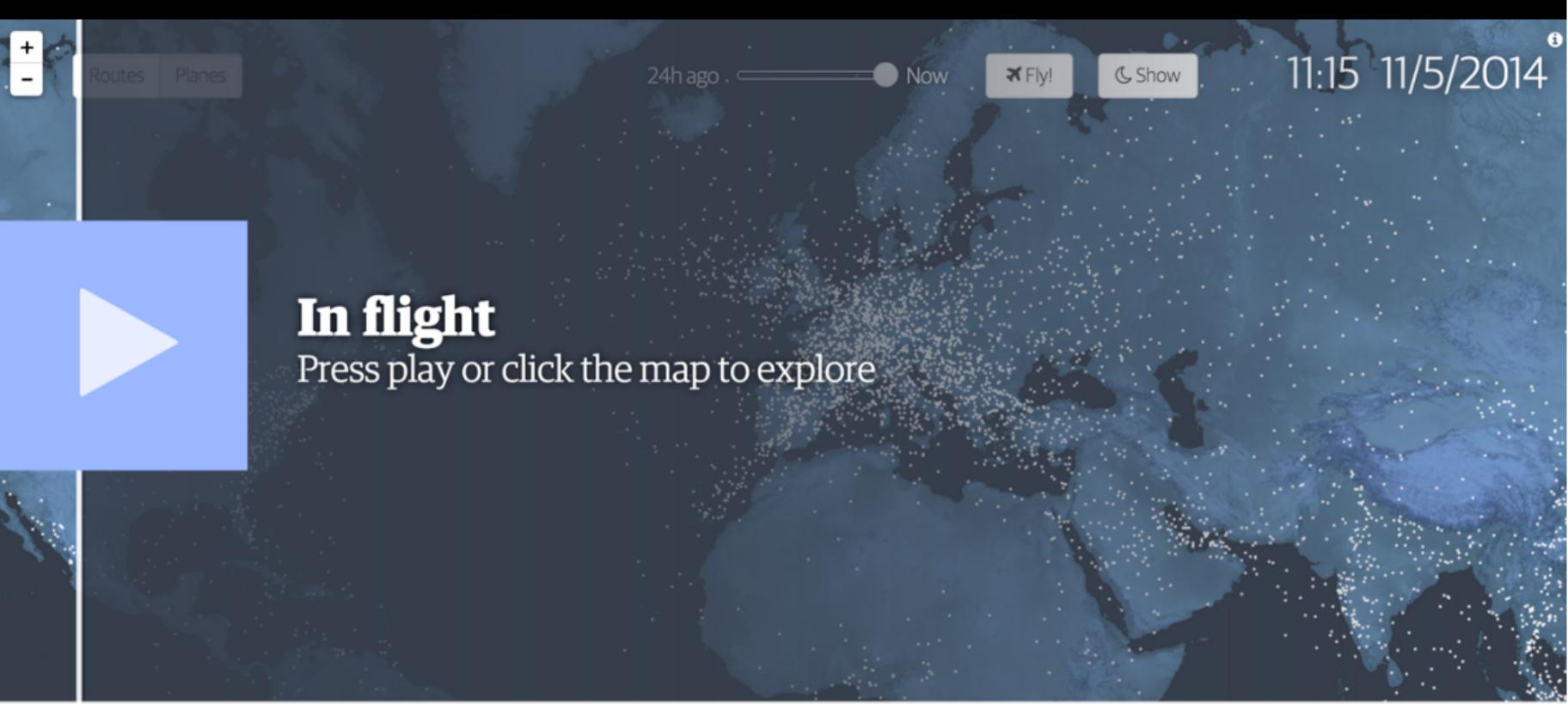
E. J. Marey, *La méthode graphique* (Paris, 1885), 20. The method is attributed to the French engineer, Ibry.



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



<https://vimeo.com/33712288>



Routes Planes

24h ago Now

Fly!

Show

11:15 11/5/2014

## In flight

Press play or click the map to explore

g

1 Mapping the skies

2 Birth of an industry

3 A century of growth

4 Hitting the limits?



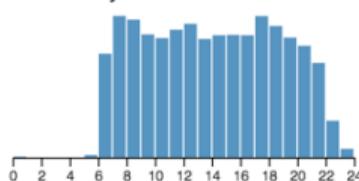
Share

Tweet

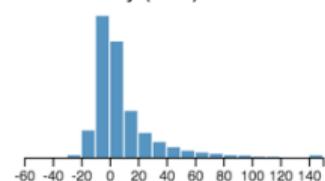


<http://www.theguardian.com/world/ng-interactive/2014/aviation-100-years>

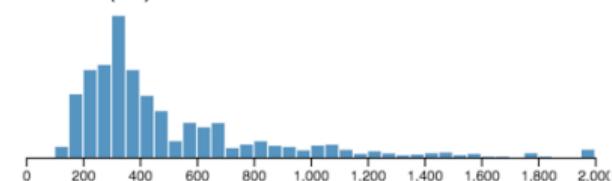
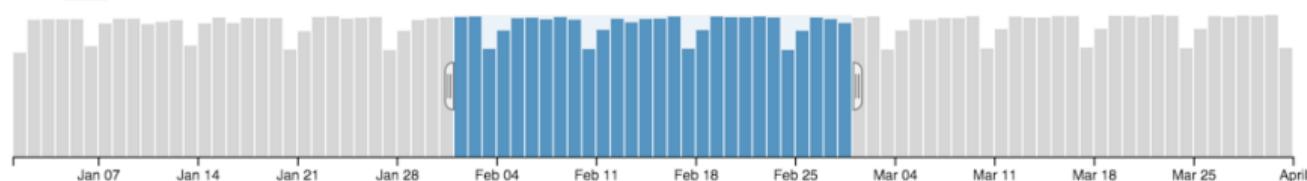
Time of Day



Arrival Delay (min.)



Distance (mi.)

Date [reset](#)

February 28, 2001

71,818 of 231,083 flights selected.

11:59 PM	LAS	LAX	236 mi.	+139 min.
11:58 PM	PHX	SAN	304 mi.	+83 min.
11:49 PM	SJC	PDX	569 mi.	+172 min.
11:42 PM	PHX	OAK	646 mi.	+97 min.
11:41 PM	PHX	LAX	370 mi.	+73 min.
11:40 PM	PHX	ONT	325 mi.	+92 min.
11:35 PM	PHX	ONT	325 mi.	+16 min.
11:25 PM	ONT	OAK	361 mi.	+75 min.



# Exercise

Building a crossfilter for our houses



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





Thank-you