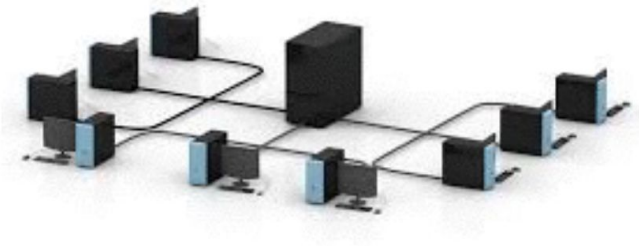
Data Science Intro

Agenda

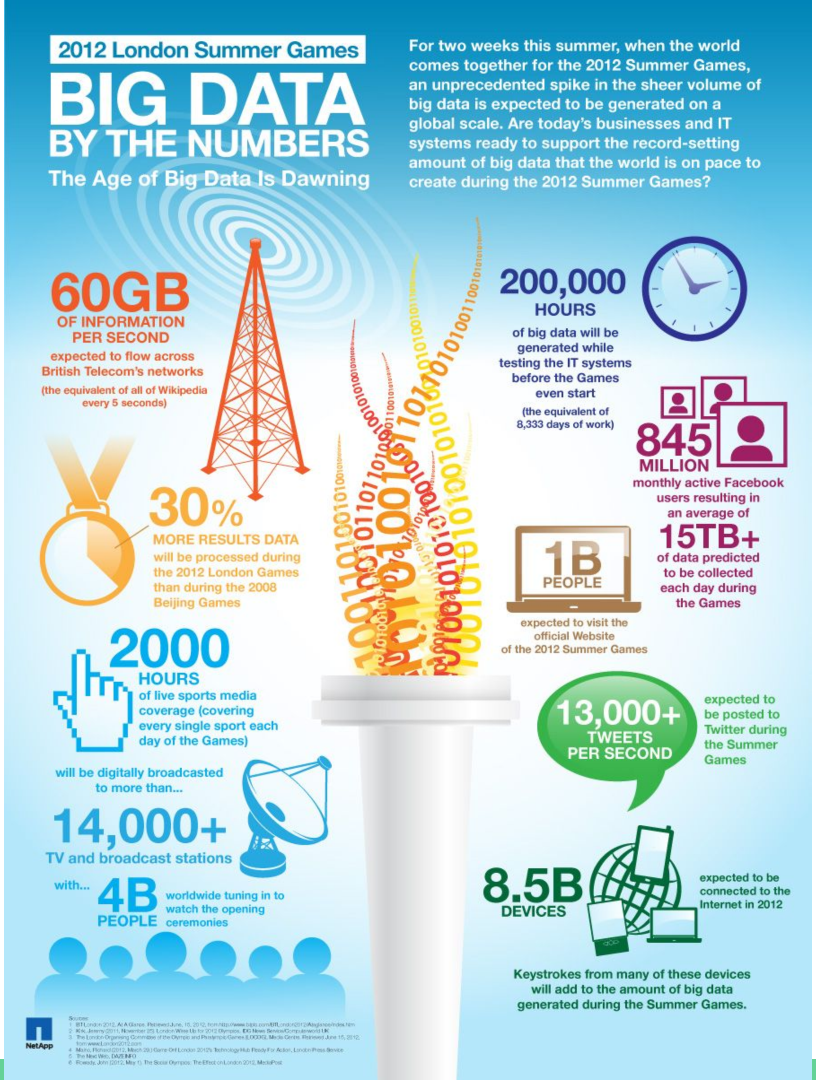
1. What is big data, really
2. Accessing the data
3. Analysing the data
4. Discussion

What is big data, really

Scale

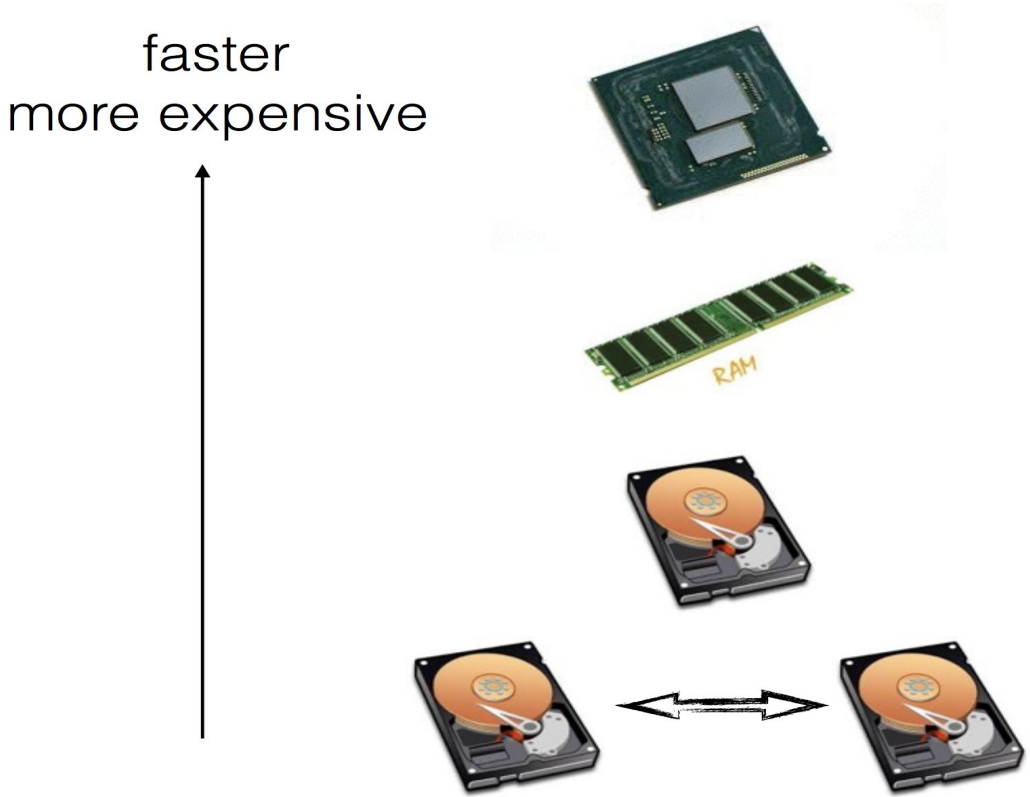


? megabytes



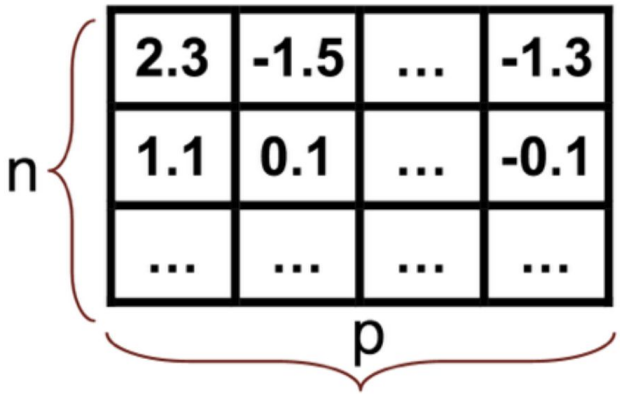
2012 London Summer Games
BIG DATA BY THE NUMBERS
The Age of Big Data Is Dawning
For two weeks this summer, when the world comes together for the 2012 Summer Games, an unprecedented spike in the sheer volume of big data is expected to be generated on a global scale. Are today's businesses and IT systems ready to support the record-setting amount of big data that the world is on pace to create during the 2012 Summer Games?
60GB OF INFORMATION PER SECOND
expected to flow across British Telecom's networks
(the equivalent of all of Wikipedia every 5 seconds)
30% MORE RESULTS DATA
will be processed during the 2012 London Games than during the 2008 Beijing Games
2000 HOURS
of live sports media coverage (covering every single sport each day of the Games)
will be digitally broadcasted to more than...
14,000+
TV and broadcast stations
with... 4B PEOPLE
worldwide tuning in to watch the opening ceremonies
200,000 HOURS
of big data will be generated while testing the IT systems before the Games even start
(the equivalent of 8,333 days of work)
845 MILLION
monthly active Facebook users resulting in an average of
15TB+
of data predicted to be collected each day during the Games
1B PEOPLE
expected to visit the official Website of the 2012 Summer Games
13,000+ TWEETS PER SECOND
expected to be posted to Twitter during the Summer Games
8.5B DEVICES
expected to be connected to the Internet in 2012
Keystrokes from many of these devices will add to the amount of big data generated during the Summer Games.
Sources:
NetApp

Scale - Memory Hierarchy



Format

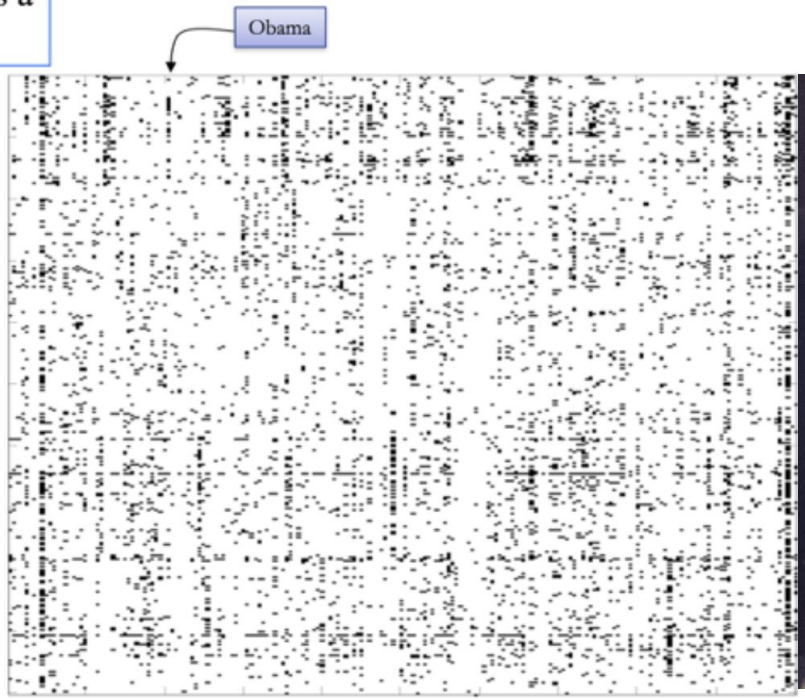
Types of Data: Flat File Data

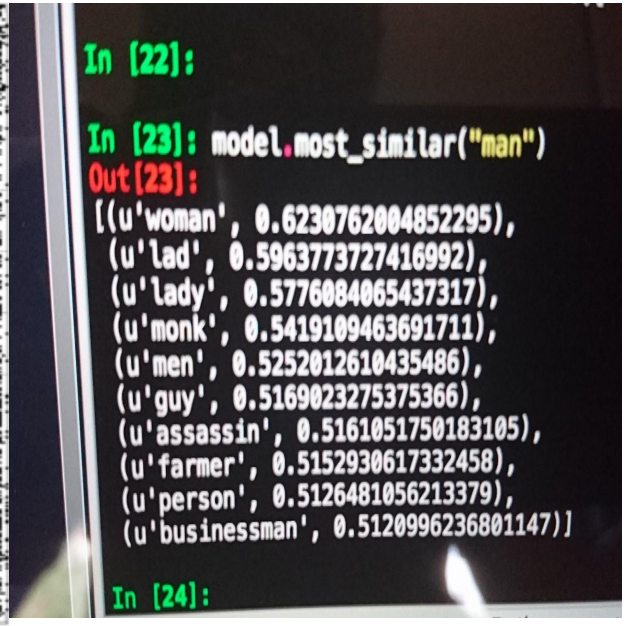


- Rows = objects
- Columns = measurements on objects
- Both n and p can be very large in data mining (also p>>n)
- Matrix can be quite sparse

Types of Data: Text Data

Can be represented as a sparse matrix





Types of Data: Transactional Data

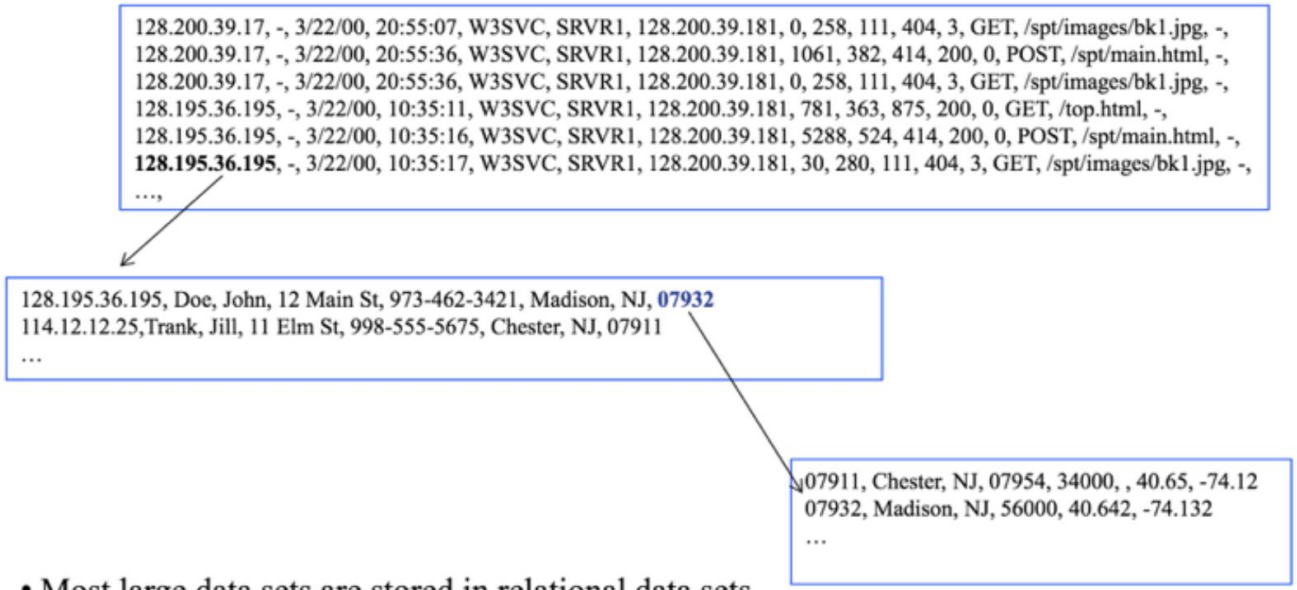
Date stamped events (logs, phone calls):

128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.101, -, 3/22/00, 16:18:50, W3SVC, SRVR1, 128.200.39.181, 60, 425, 72, 304, 0, GET, /top.html, -,
128.195.36.101, -, 3/22/00, 16:18:58, W3SVC, SRVR1, 128.200.39.181, 8322, 527, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.101, -, 3/22/00, 16:18:59, W3SVC, SRVR1, 128.200.39.181, 0, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:54:37, W3SVC, SRVR1, 128.200.39.181, 140, 199, 875, 200, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 17766, 365, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:39, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:03, W3SVC, SRVR1, 128.200.39.181, 1081, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:56:04, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:33, W3SVC, SRVR1, 128.200.39.181, 0, 262, 72, 304, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:56:52, W3SVC, SRVR1, 128.200.39.181, 19598, 382, 414, 200, 0, POST, /spt/main.html, -,

Can be represented as a time series:

User 1	2	3	2	2	3	3	3	1	1	1	3	1	3	3	3	3
User 2	3	3	3	1	1	1										
User 3	7	7	7	7	7	7	7	7								
User 4	1	5	1	1	1	5	1	5	1	1	1	1	1	1		
User 5	5	1	1	5												
…		…														

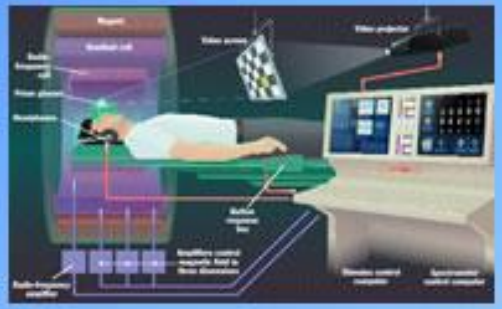
Types of Data: Relational Data



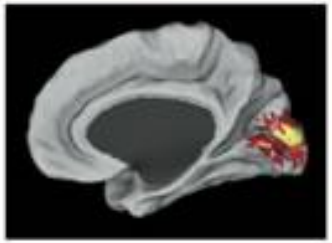
- Most large data sets are stored in relational data sets
- Special data query language: SQL

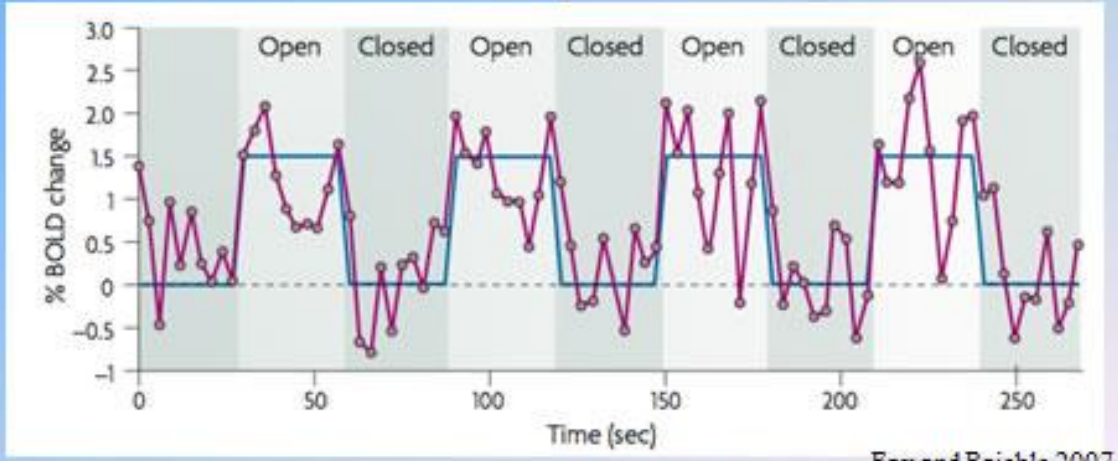
Types of Data: Time Series Data

Functional MRI



Open – Closed =





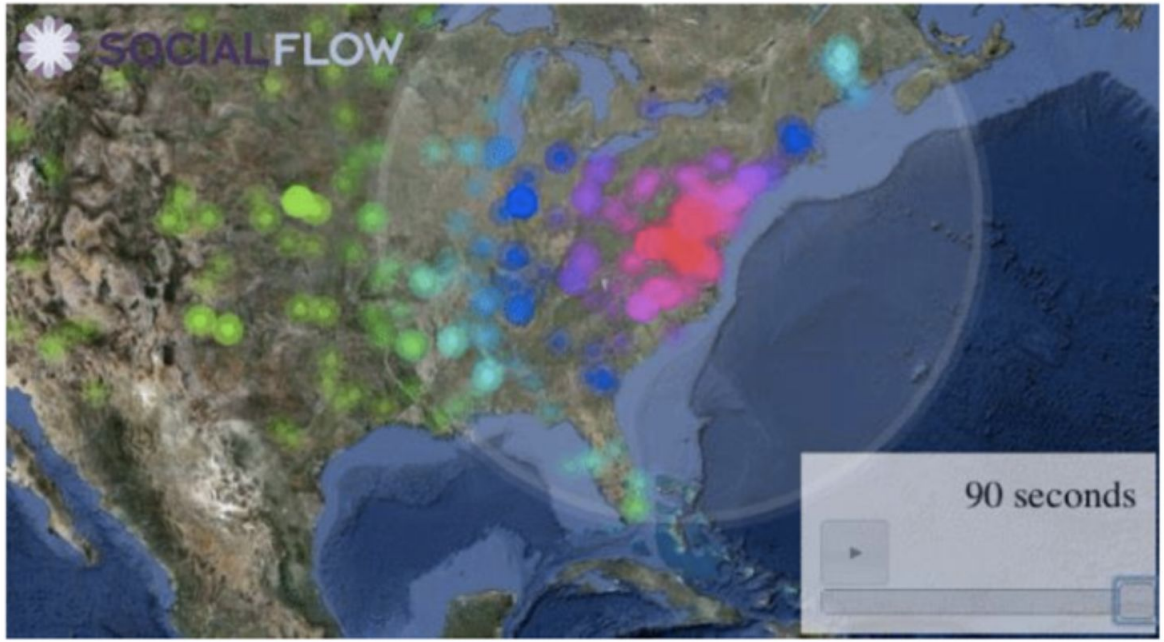
Fox and Raichle 2007

Types of Data: Image Data

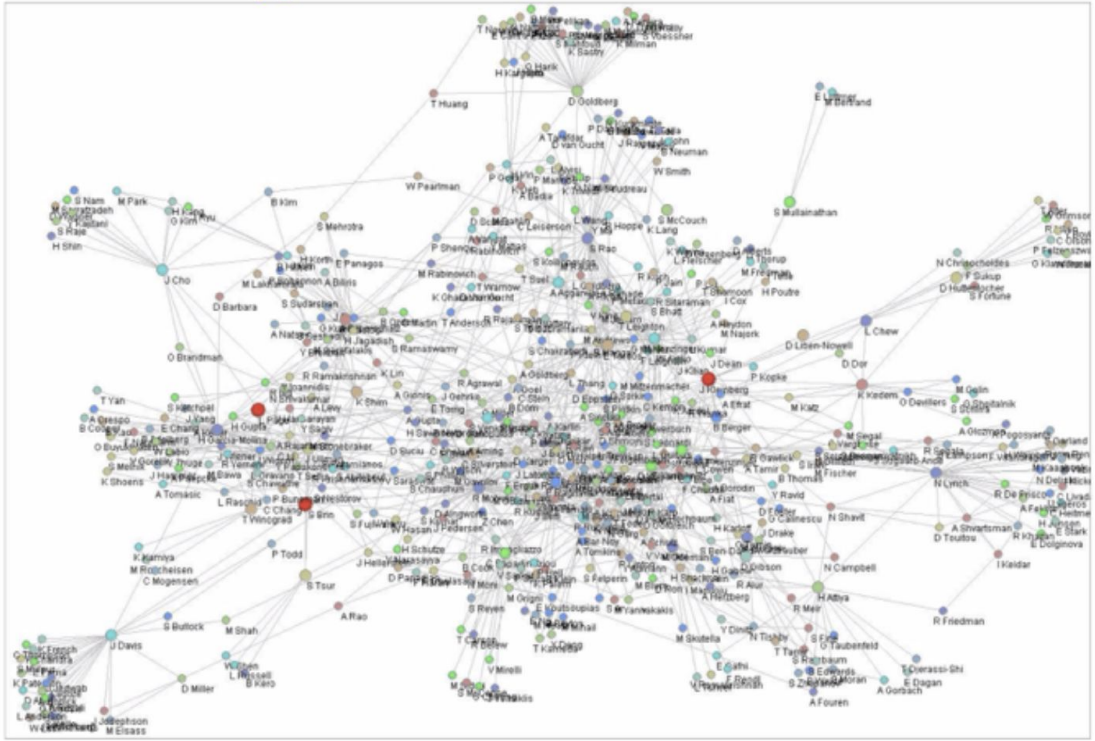
Types of Data: Spatio-Temporal Data

@b_mc817
Glendaaaaa

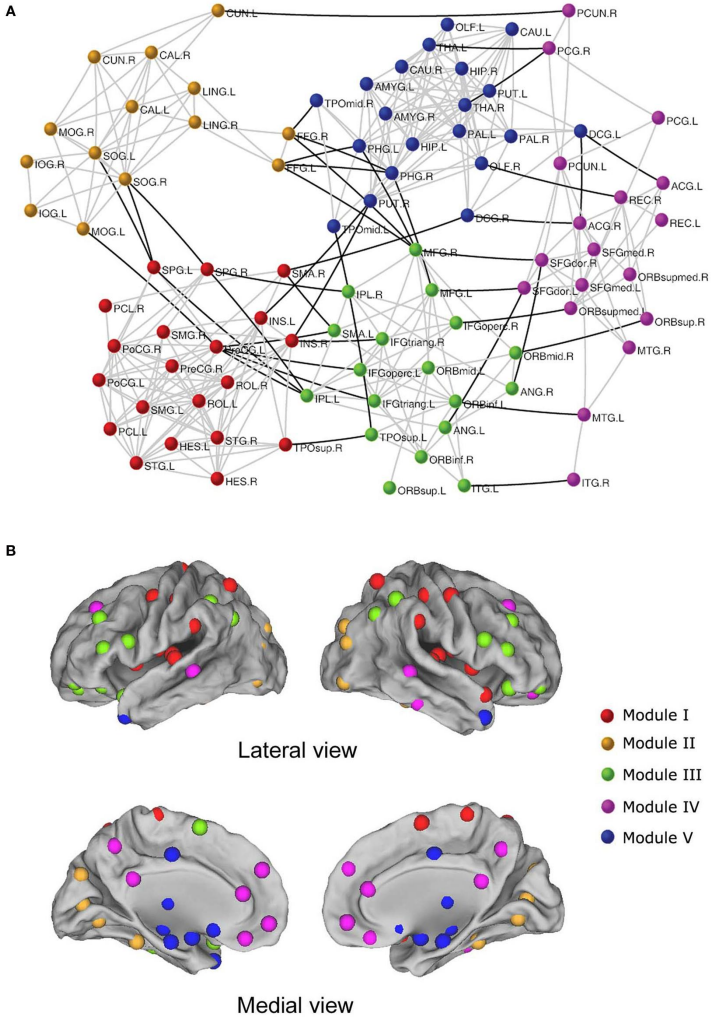
Omg earthquake!!!



Types of Data: Network Data



Algorithms for estimating relative importance in networks
S. White and P. Smyth, *ACM SIGKDD*, 2003.


A
CUN.L
PCUN.R
OLF.L
CAU.L
THA.L
PCG.R
CUN.R
CAL.R
CAU.R
HIP.R
AMYG.L
PUT.L
LING.L
TPOmid.R
THA.R
CAL.L
AMYG.R
PCG.L
LING.R
MOG.R
FEG.R
PAL.L
PAL.R
DCG.L
PHG.L
HIP.L
SOG.L
IOG.R
EFG.L
OLF.R
PCUN.L
PHG.R
SOG.R
ACG.L
REC.R
PUT.R
IOG.L
DCG.R
REC.L
TPOmid.L
ACG.R
MOG.L
SFGmed.R
MFG.R
SPG.L
SPG.R
SFGdor.R
SMA.R
ORBsupmed.R
IPL.R
MFG.L
SFGdor.L
SFGmed.L
PCL.R
ORBsupmed.L
INS.L
IFGoperc.R
ORBsup.R
SMG.R
SMA.L
IFGtriang.R
INS.R
MTG.R
PoCG.R
PreCG.L
ORBmid.R
PreCG.R
IFGoperc.L
ORBmid.L
PoCG.L
ROL.R
ANG.R
IPL.L
IFGtriang.L
ROL.L
ORBinf.L
SMG.L
MTG.L
PCL.L
ANG.L
HES.L
STG.R
TPOsup.L
TPOsup.R
ORBinf.R
STG.L
HES.R
ORBsup.L
ITG.L
ITG.R
B
Lateral view
Medial view
Module I
Module II
Module III
Module IV
Module V

What is a data scientist?

Data Scientists

What my mom thinks I do

What my friends think I do

What society thinks I do

What investors think I do

What I think I do

What I really do

SiSense
The Big Data Analytics Company
www.SiSense.com

Zvi
@nivertech

Follow

"Data Scientist" is a Data Analyst who lives in California.

Reply Retweet Favorite More

RETWEETS	FAVORITES
140	40

9:55 PM - 14 Mar 2012

Javier Nogales
@fjnogales

Follow

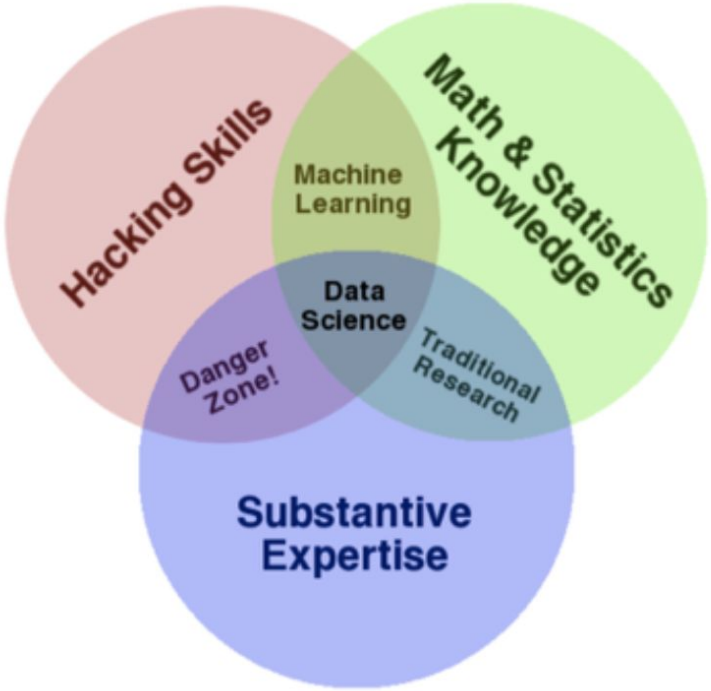
Data Scientist (2/2): person who is worse at statistics than any statistician and worse at software engineering than any software engineer

RETWEET 1

FAVORITES 5

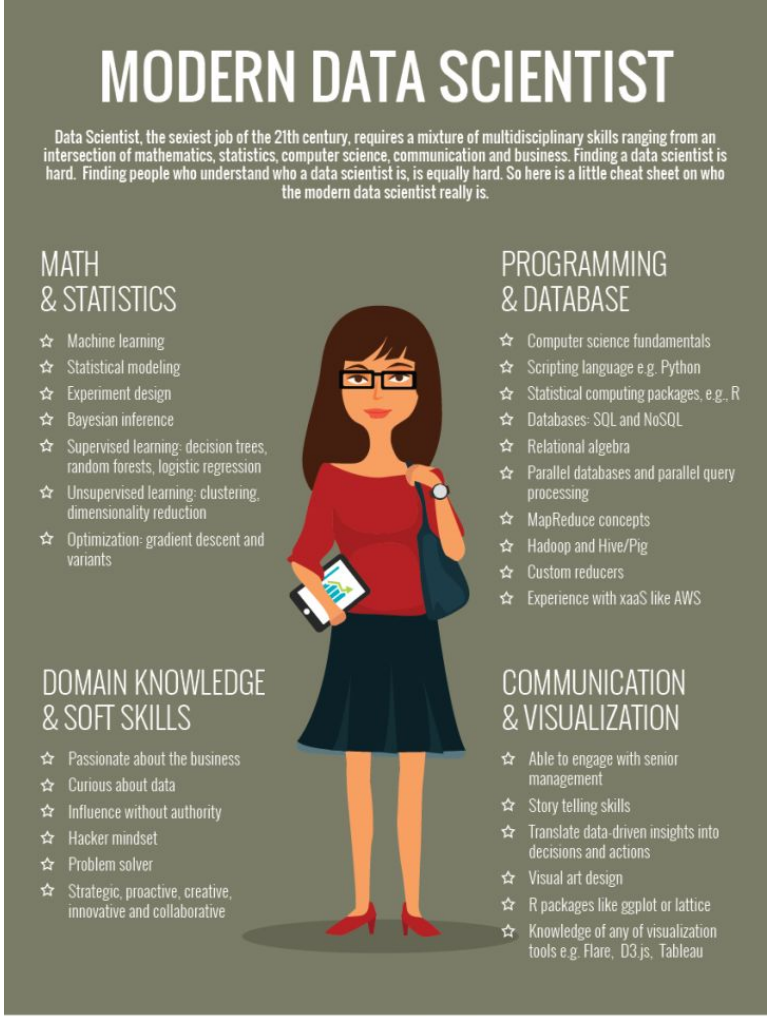
9:08 AM - 27 Jan 2014

What is a data scientist?



Wide variance in terms of skillsets: many job descriptions are more appropriate for a **team of data scientists!**

Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



Use a set of tools and techniques to extract useful information from data

Interdisciplinary, problem-solving oriented

Apply scientific techniques to practical problems

Who uses Data Science

Netflix - movie recommendations

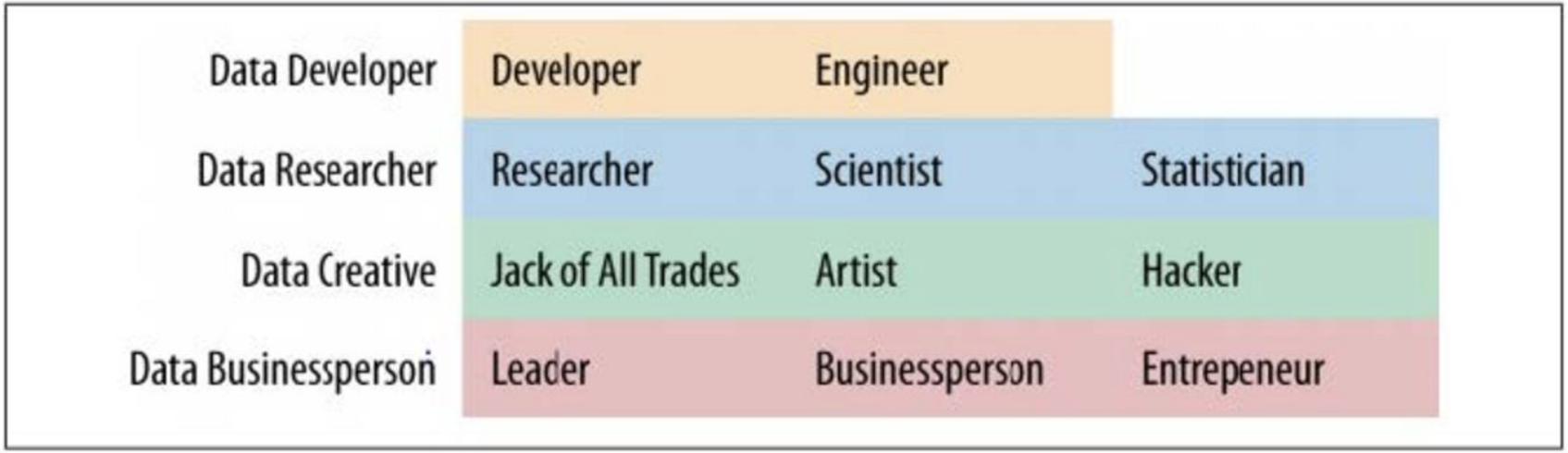
Amazon's algorithm - "you might also like x"

Five Thirty Eight - election and sports coverage

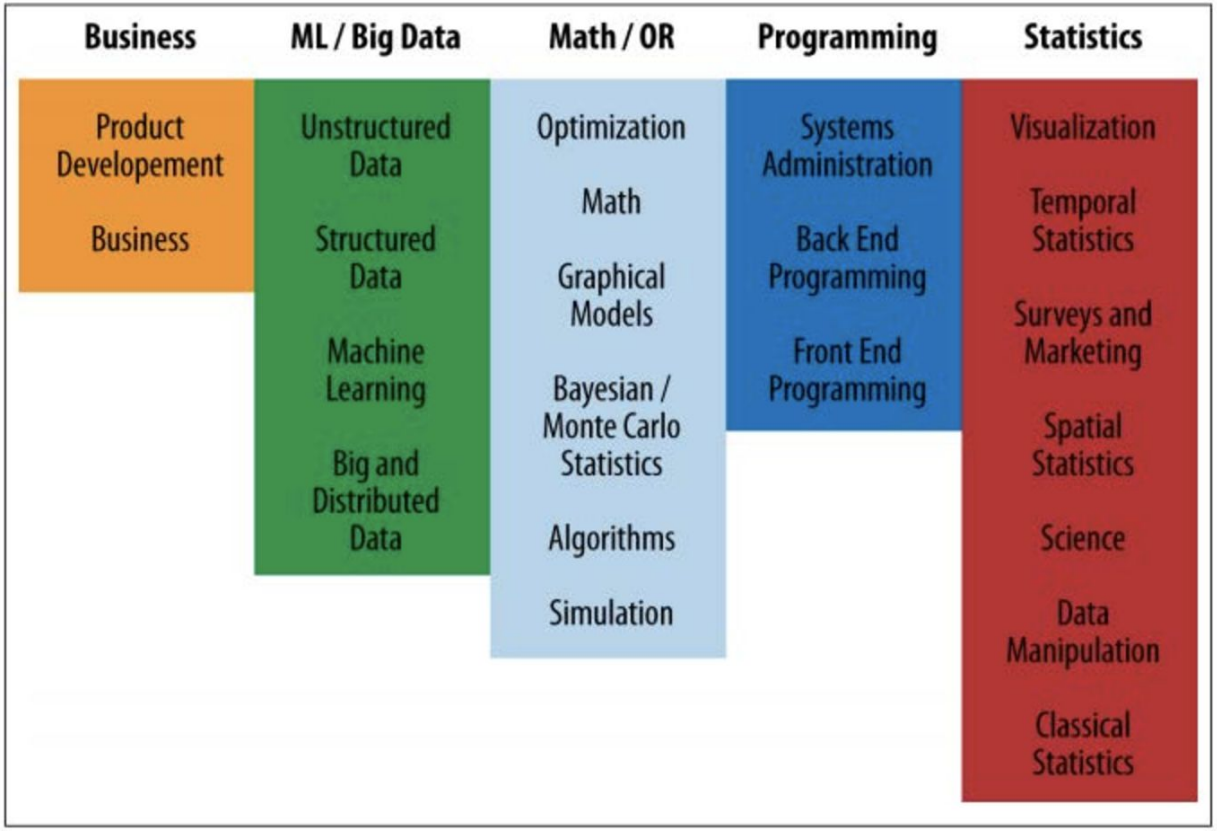
Draft Kings - using data science to predict daily bets

Google - auto-translate and search results

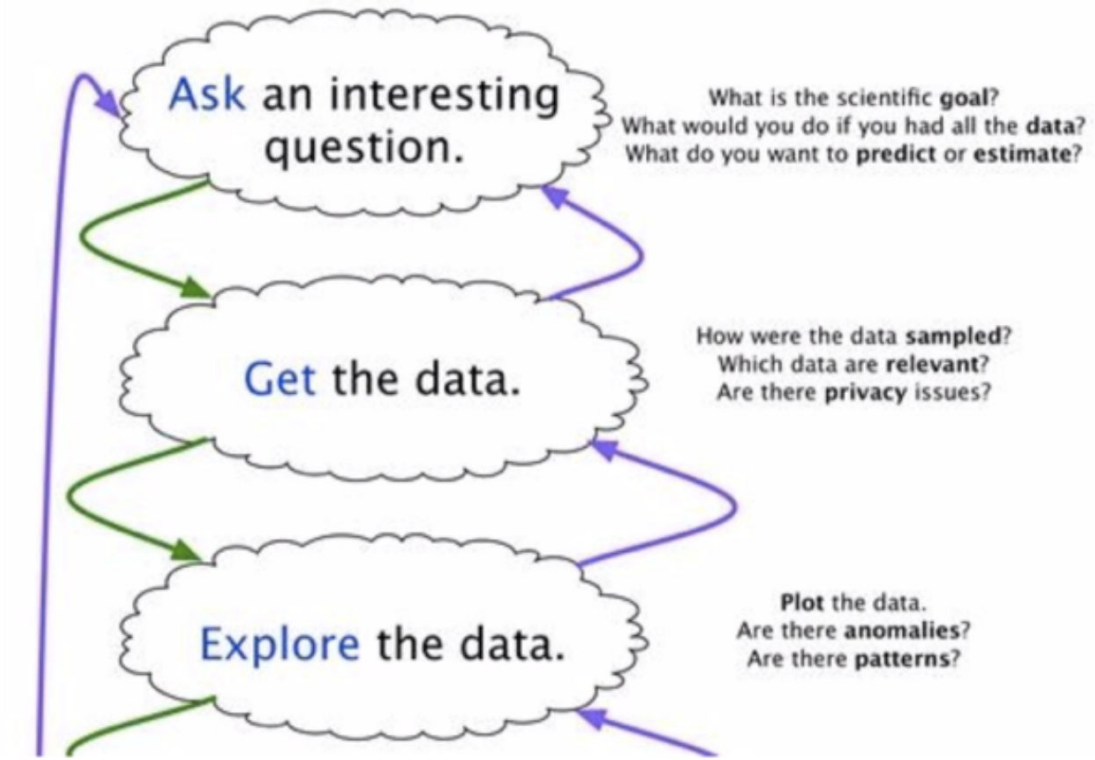
What are the roles?



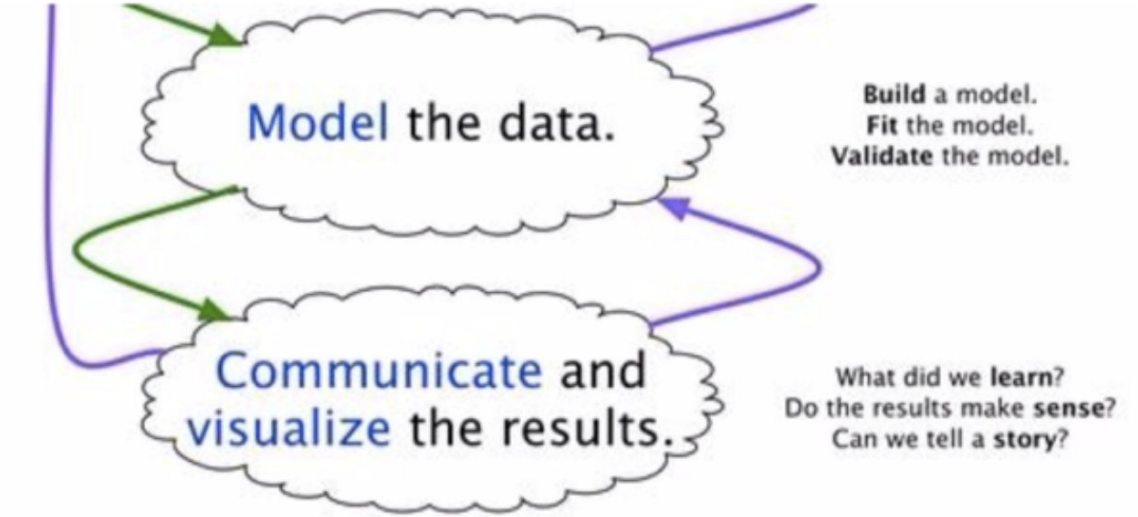
Skills


Business
ML / Big Data
Math / OR
Programming
Statistics
Product Developement
Business
Unstructured Data
Structured Data
Machine Learning
Big and Distributed Data
Optimization
Math
Graphical Models
Bayesian / Monte Carlo Statistics
Algorithms
Simulation
Systems Administration
Back End Programming
Front End Programming
Visualization
Temporal Statistics
Surveys and Marketing
Spatial Statistics
Science
Data Manipulation
Classical Statistics

Data Science WorkFlow

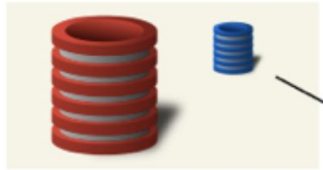


Data Science WorkFlow



Source: https://www.quora.com/What-is-the-work-flow-or-process-of-a-data-scientist-analyst-and-what-tools-do-you-use-for-this/answer/Ryan-Fox-Squire

Accessing the data
Database technology

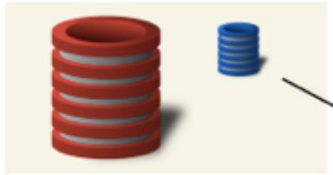


Relational

- Traditional rows and columns data
- **Strict** structure / Primary Keys
- Entire column for each feature
- Industry standard

NoSql

- No well defined data structure
- Works better for unstructured data
- Cheaper hardware
- Popular among Startups



Relational

Examples

- MySQL
- Oracle
- Postgres
- SQLite

NoSql

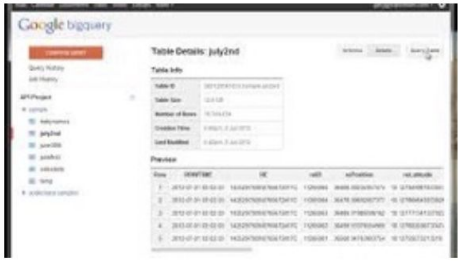
Examples

- MongoDB
- CouchDB
- Redis
- Casssandra

Analyzing the data

What is Machine Learning?

Exploring the data - from Excel to BigQuery



- summarising: min, max, mean, variance
- cleaning: outliers, junk data
- initial visualisation: pie, histogram, line
- analytical transformations: machine learning

Google BigQuery

https://www.youtube.com/watch?v=D-YrpJkuGqE

Visualisation: beyond pie charts https://d3js.org/

Beyond the basics: Machine Learning

Mario

https://www.youtube.com/watch?v=qv6UVOQ0F44

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

150 observations *(n = 150)*

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

4 features (p = 4)

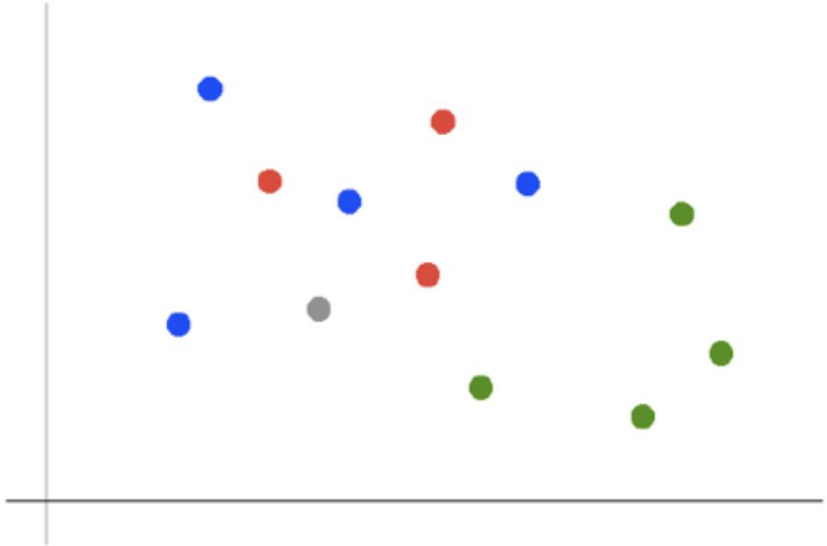
response

Clustering, or **cluster analysis**, is the task of grouping observations such that members of the same group, or **cluster**, are more similar to each other by some metric than they are to the members of the other clusters

Suppose we want to predict the color of the gray dot.

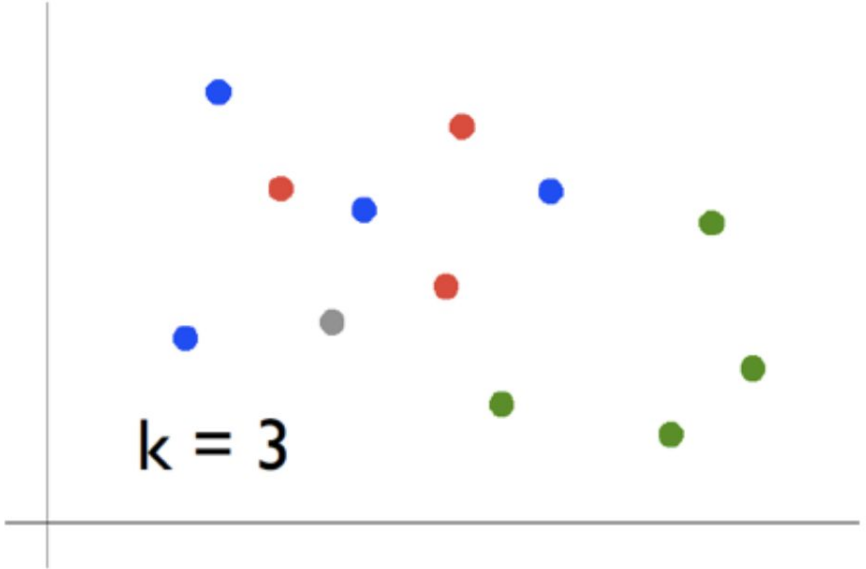
QUESTION:

What are the predictors?
What is the response?



Suppose we want to predict the color of the gray dot.

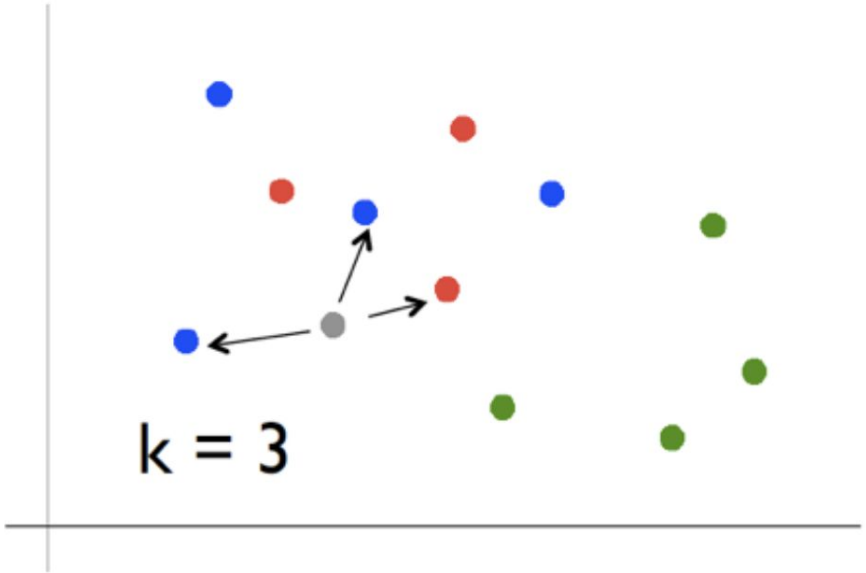
1) Pick a value for k.



Suppose we want to predict the color of the gray dot.

1) Pick a value for k.

2) Find colors of k nearest neighbors.

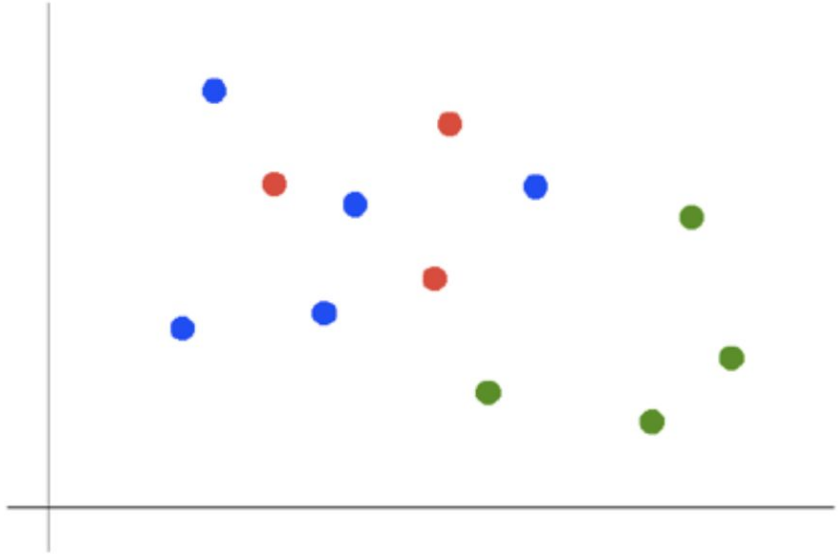


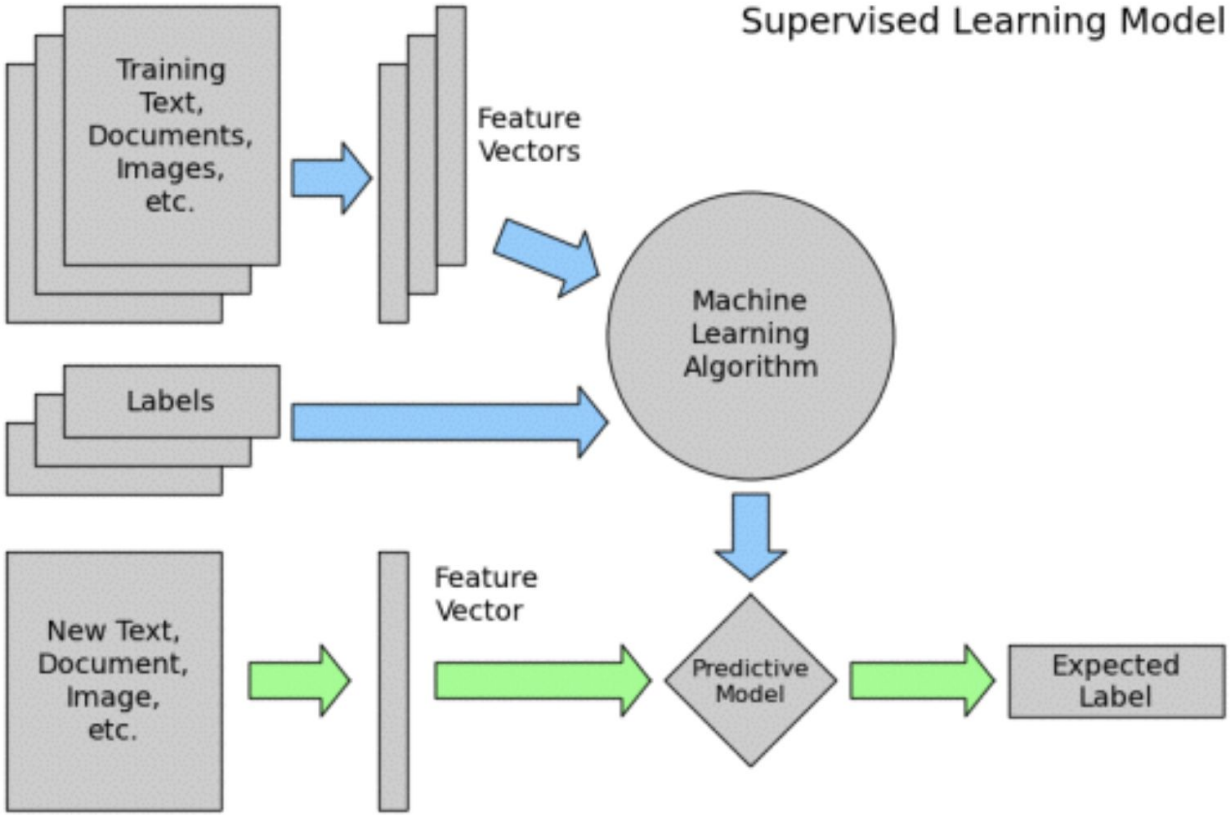
Suppose we want to predict the color of the gray dot.

1) Pick a value for k.

2) Find colors of k nearest neighbors.

3) Assign the most common color to the gray dot.




Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Discussion