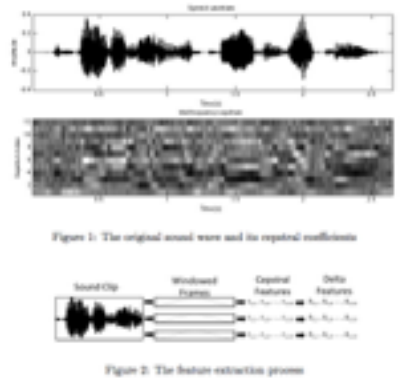
Data Science Intro

About me



NEURO SPLASH

VISA

nugit

tripadvisor

GENERAL ASSEMBLY

General Skills

Programming
Golang
shell
javascript
C
Python
…

Data Stack
hadoop
hive
spark
custom
…

Web Stack
node.js
gorilla
meteor, angular
d3.js, bootstrap
SQL, noSQL
…

Other
vanilla machine learning
auction dynamics
selling tech
…

“Techniques & tools to summarise & analyse large data sets”

Agenda

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

0. Tell me about you!

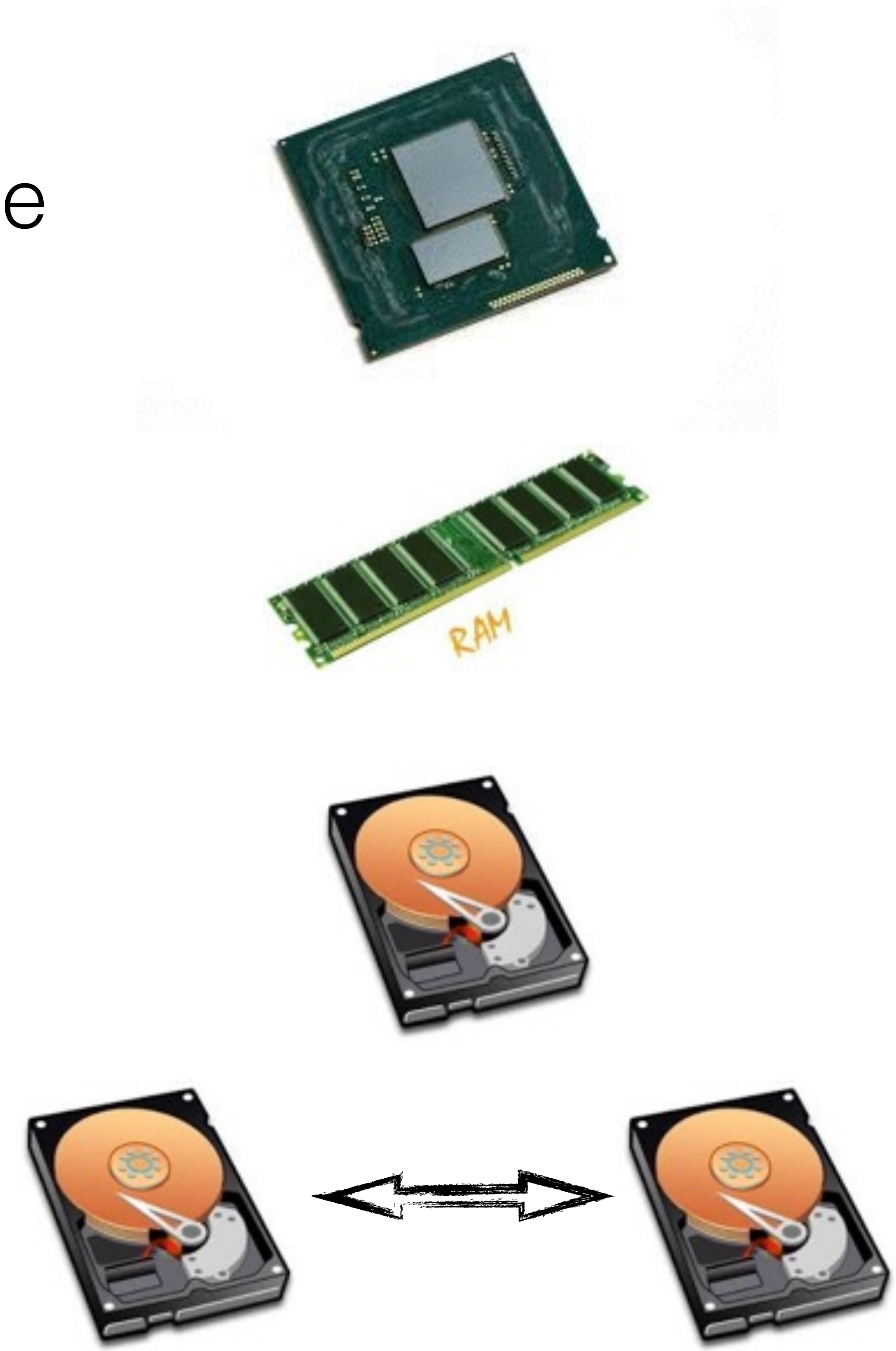
1. What is big data, really

Scale

? megabytes

Scale - Memory Hierarchy

faster
more expensive



Format

Types of Data: Flat File Data

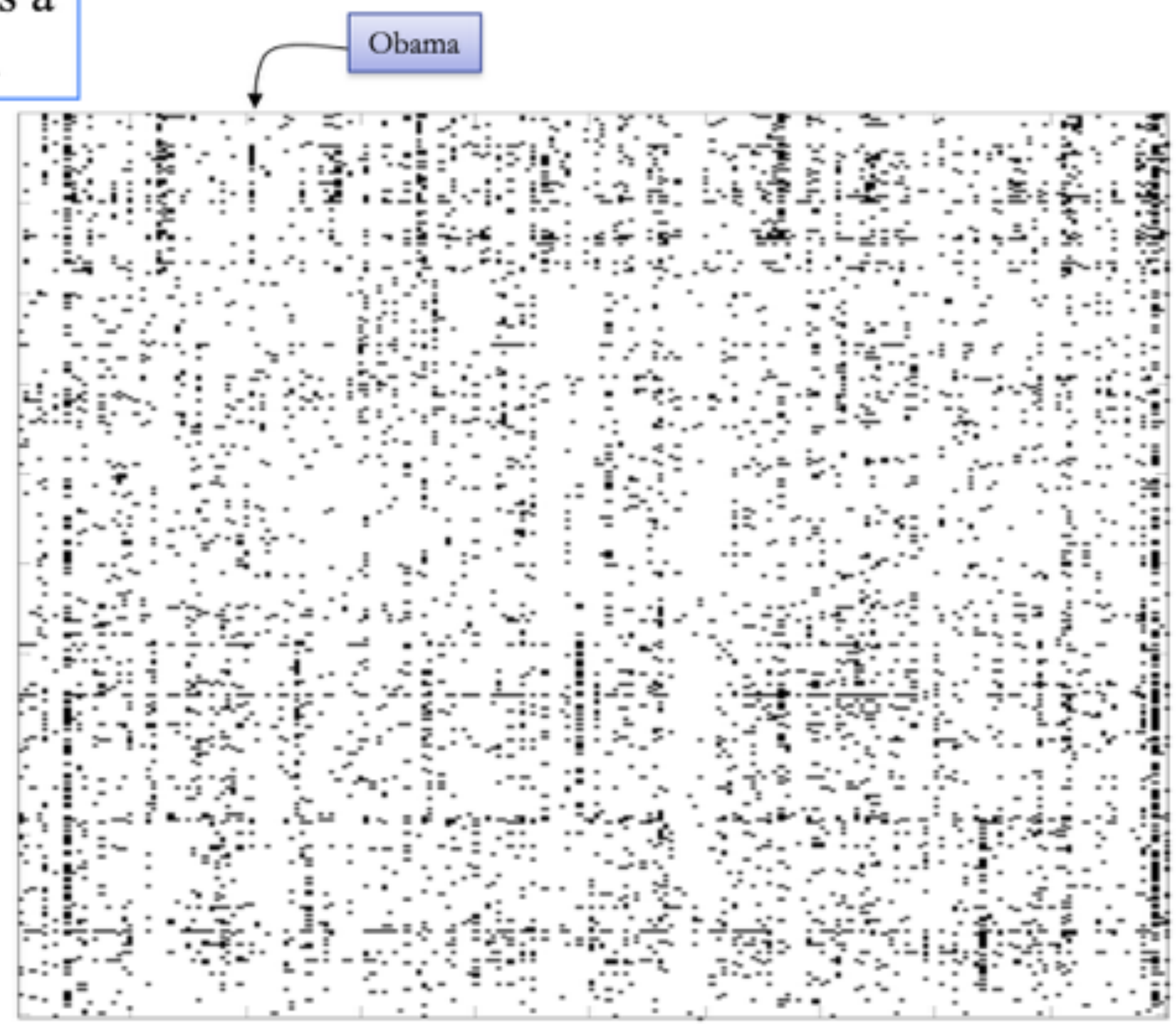
$$n\left\{\underbrace{\begin{array}{|c|c|c|c|}\hline 2.3 & -1.5 & \ldots & -1.3 \\ \hline 1.1 & 0.1 & \ldots & -0.1 \\ \hline \ldots & \ldots & \ldots & \ldots \\ \hline \end{array}}_{p}\right.$$

- 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

Text Documents



Word ID

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

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, -,
128.195.36.195, -, 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, -,
…,

128.195.36.195, Doe, John, 12 Main St, 973-462-3421, Madison, NJ, **07932**
114.12.12.25,Trank, Jill, 11 Elm St, 998-555-5675, Chester, NJ, 07911
…

07911, Chester, NJ, 07954, 34000, , 40.65, -74.12
07932, Madison, NJ, 56000, 40.642, -74.132
…

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

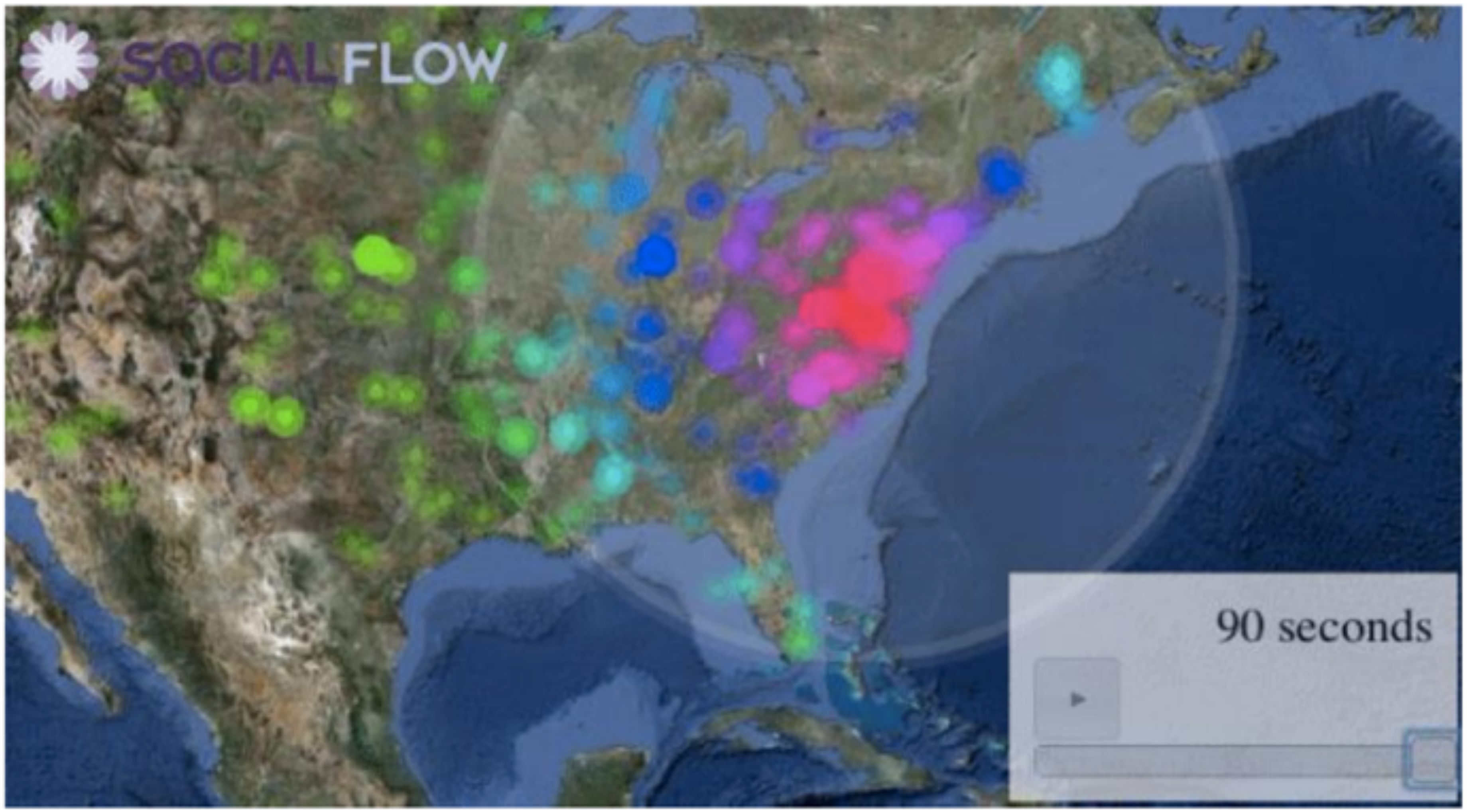
Types of Data: Time Series Data

Types of Data: Image Data

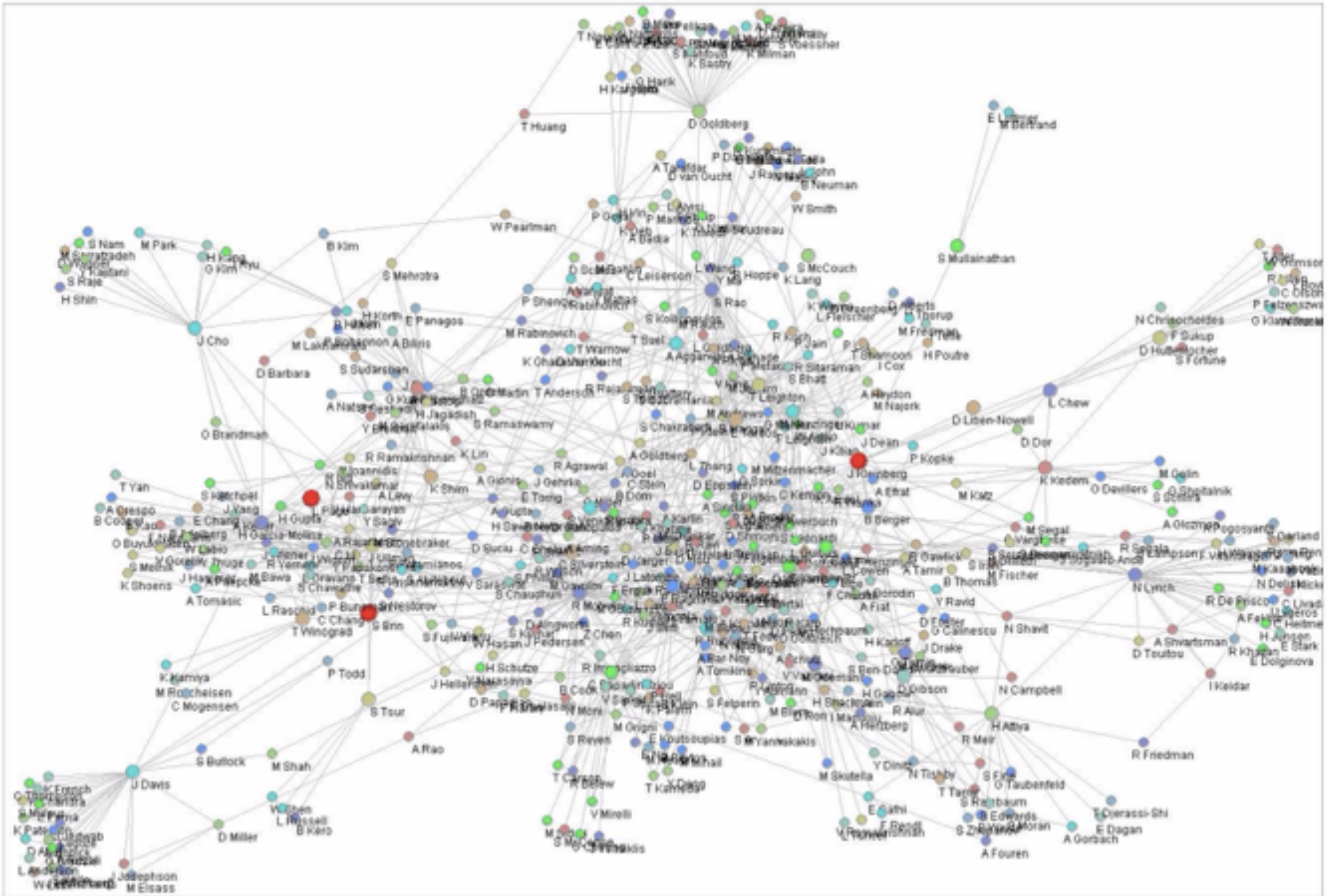
Types of Data:
Spatio-Temporal Data

@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.

What is a data scientist?

WHAT IS A DATA SCIENTIST?

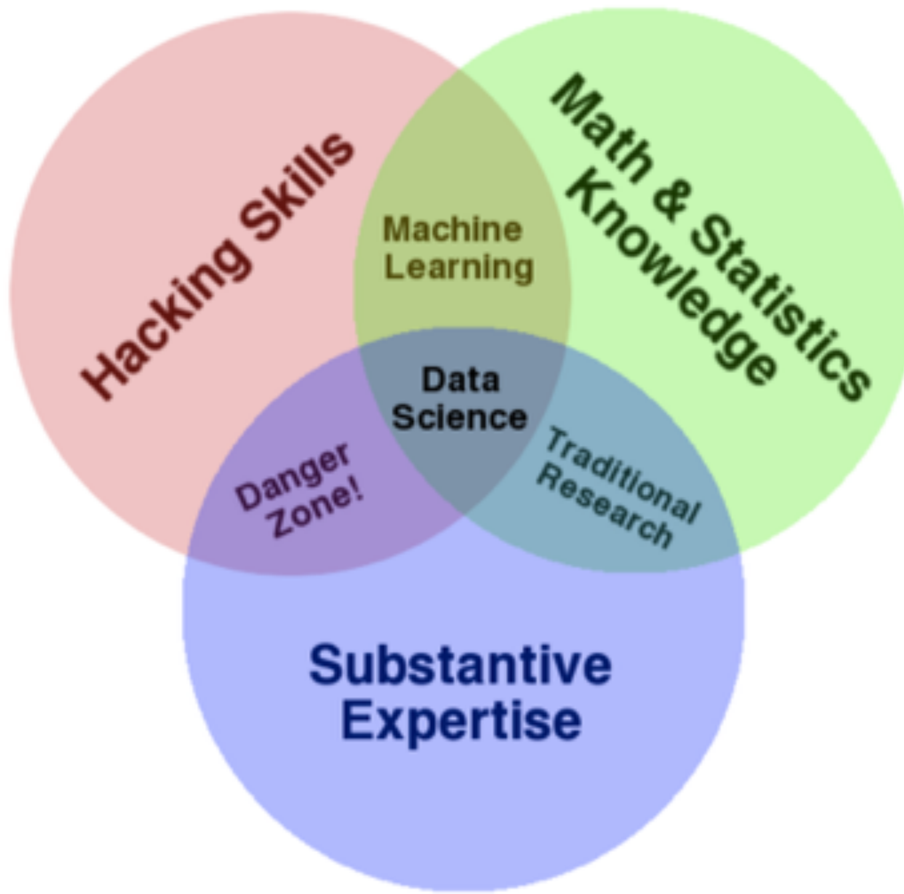
Javier Nogales
@fjnogales

Follow

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

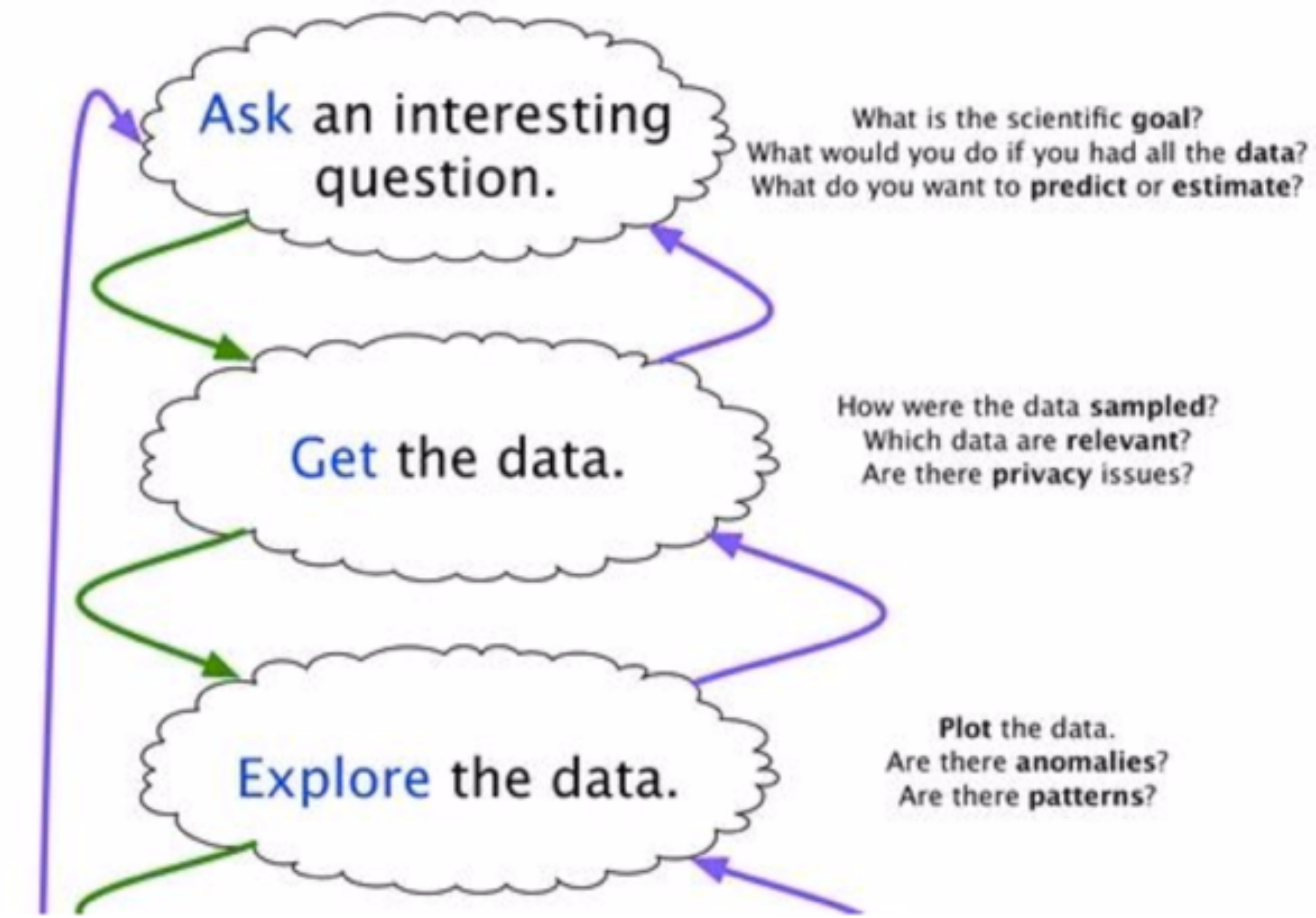
RETWEET 1 FAVORITES 5

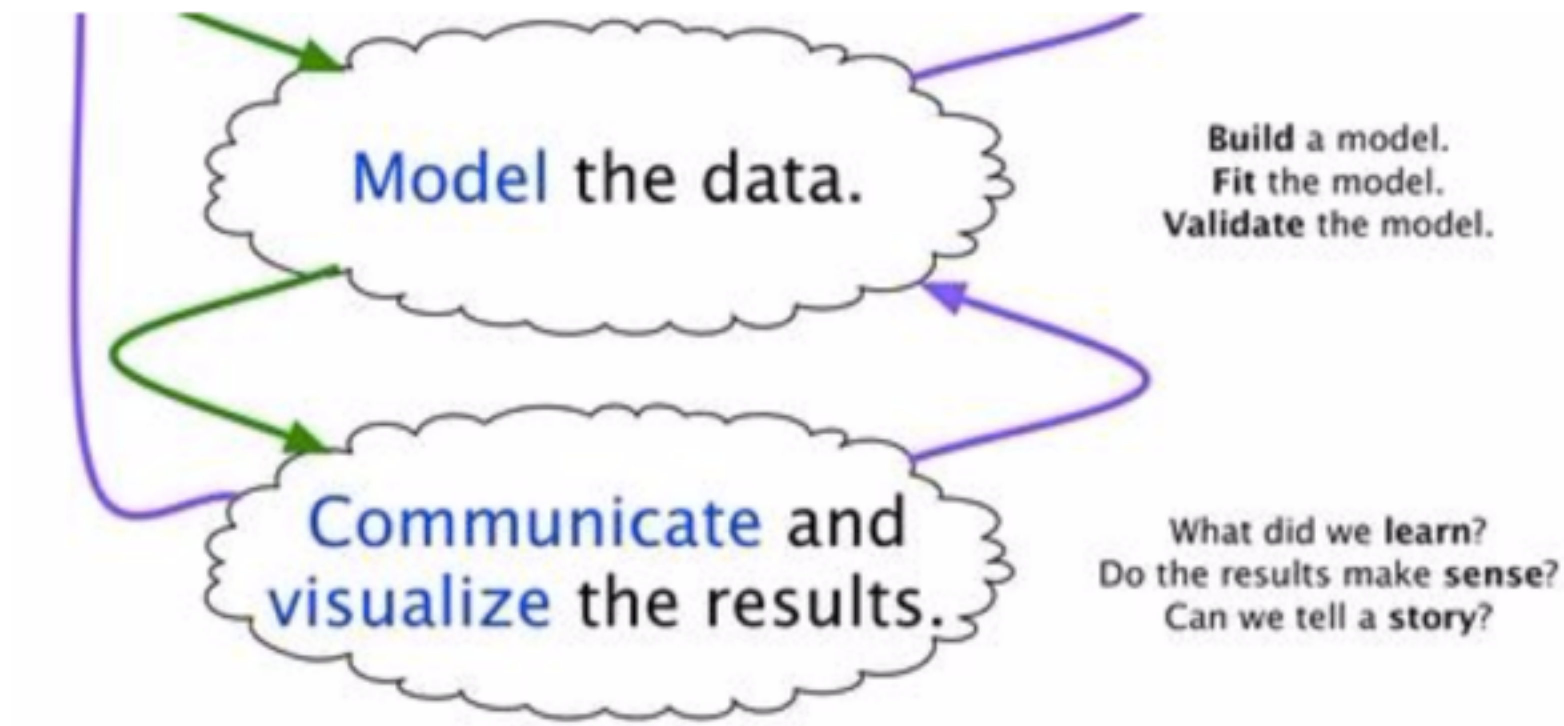
9:08 AM - 27 Jan 2014



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

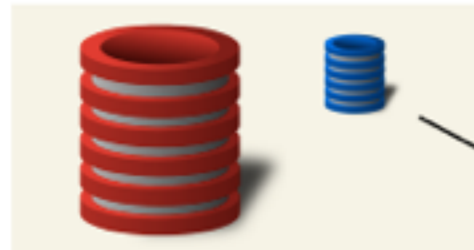
Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram


Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?



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

2. Accessing the data - database technologies

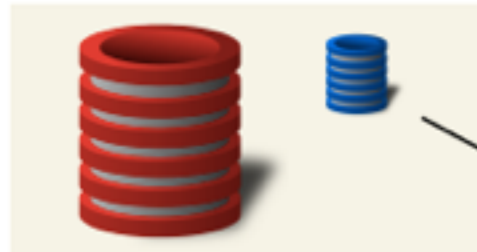


Relational

- 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

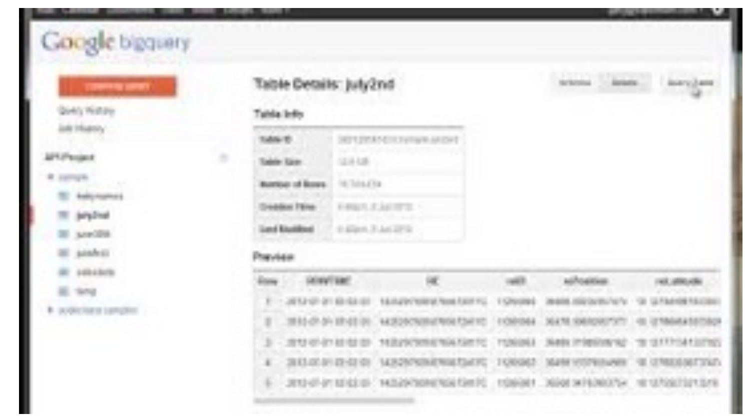
Examples

- MongoDB
- CouchDB
- Redis
- Casssandra

3. Analysing 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

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)

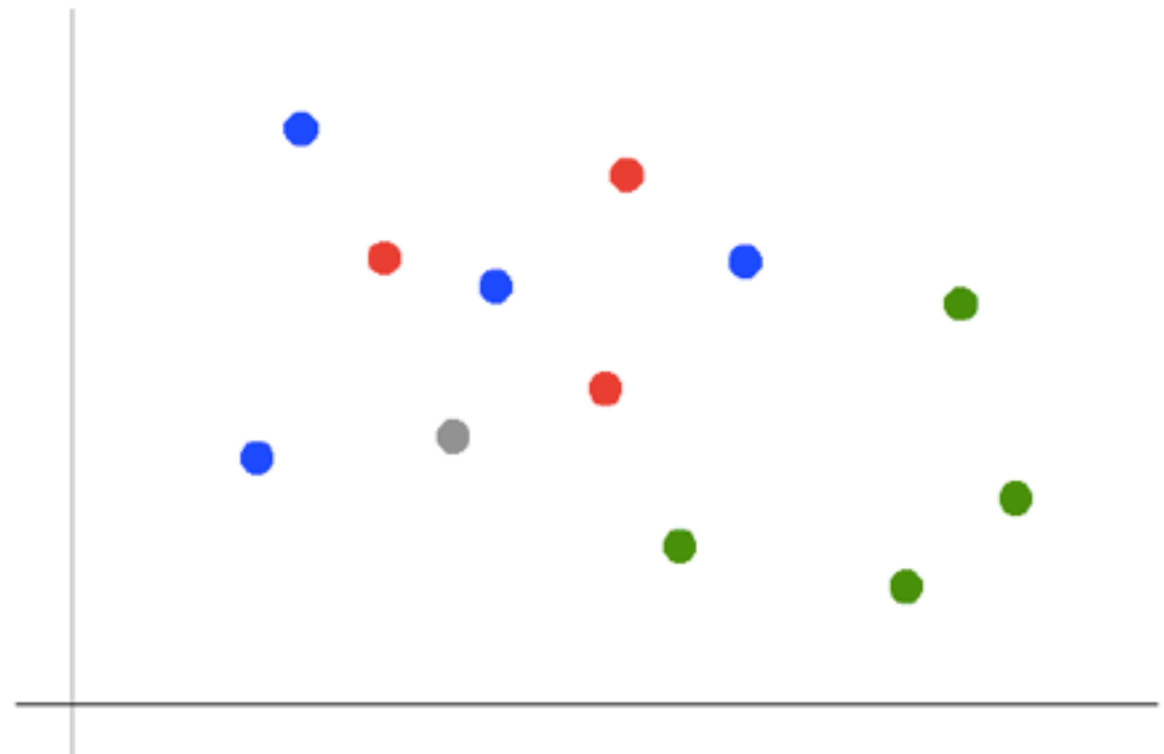
response

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.

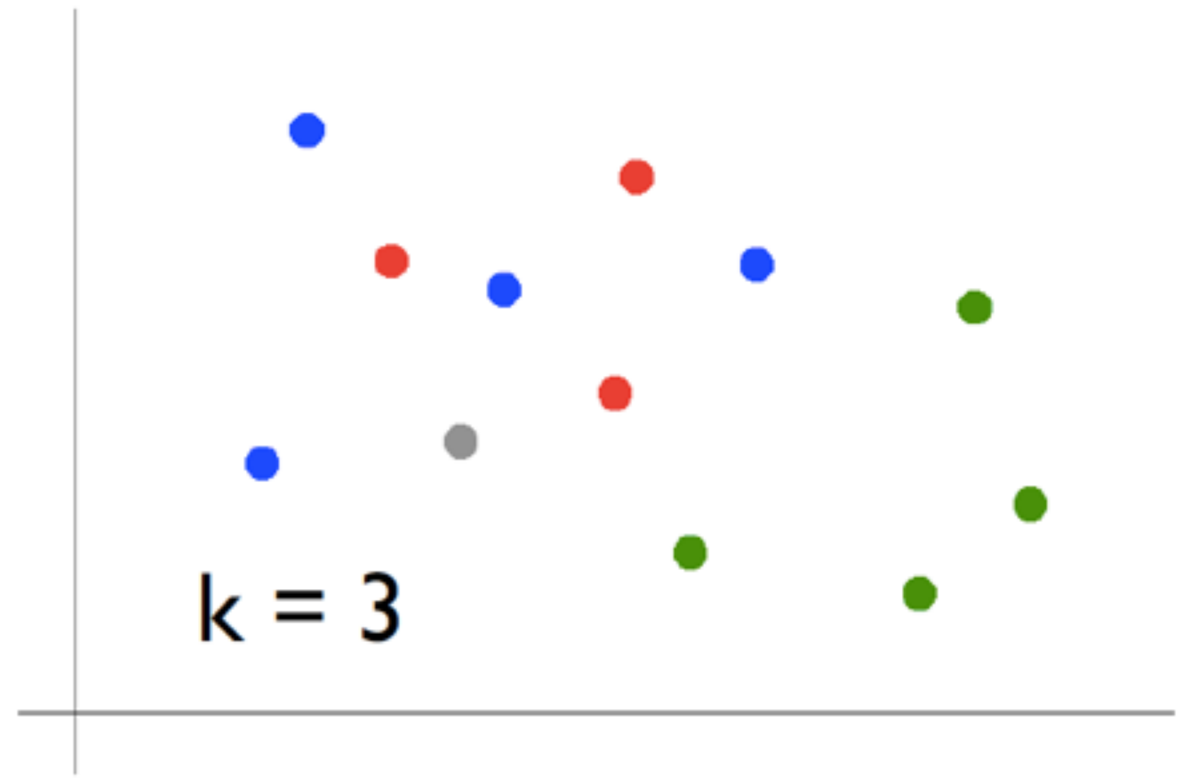
QUESTION:

What are the predictors?
What is the response?



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

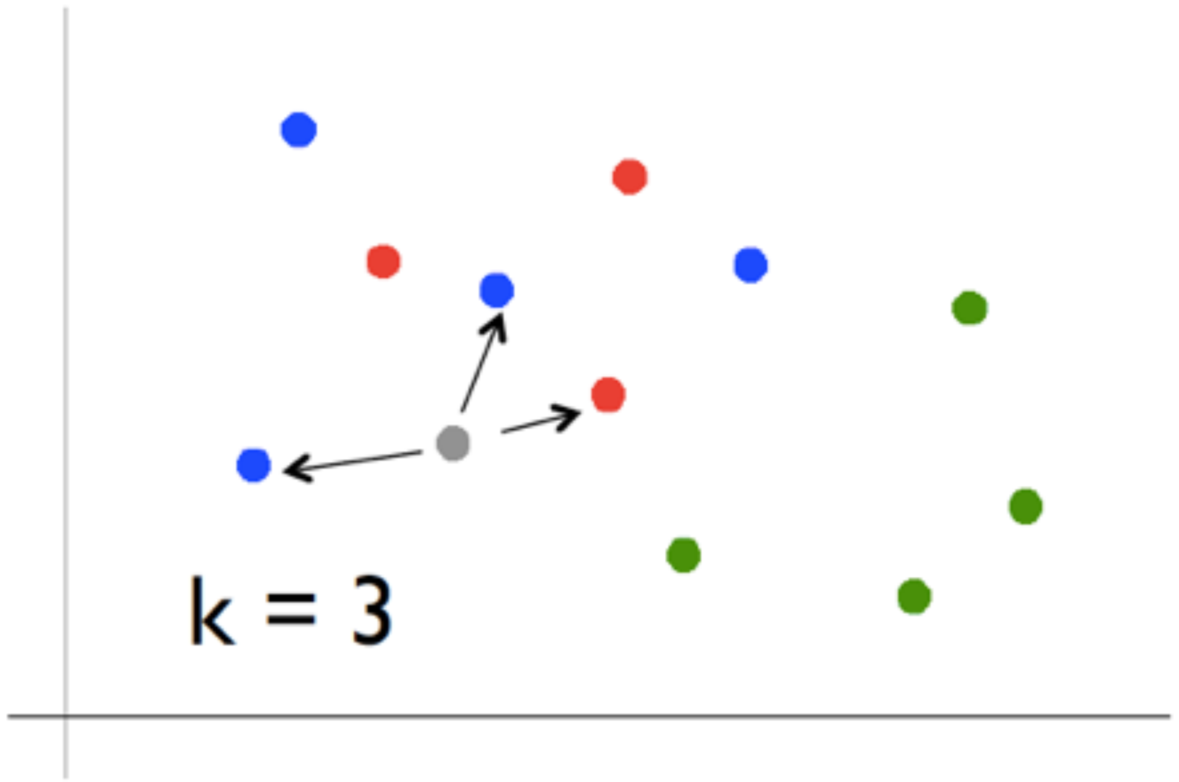
1) Pick a value for k.



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

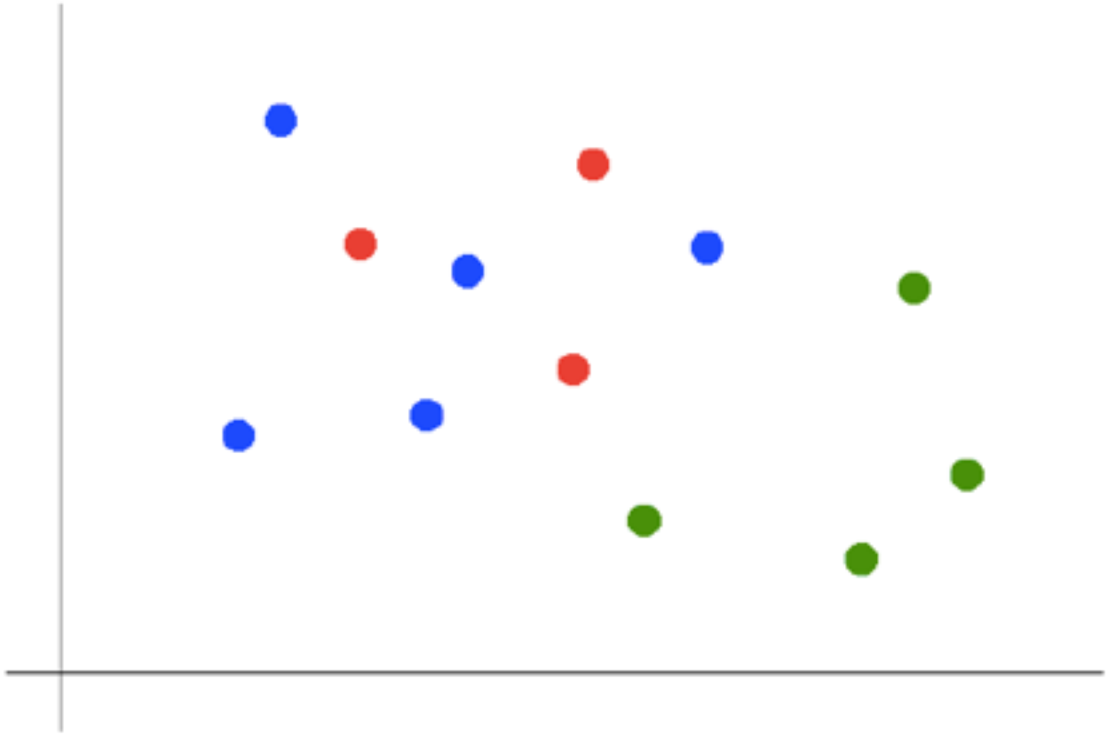
1) Pick a value for k.

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.



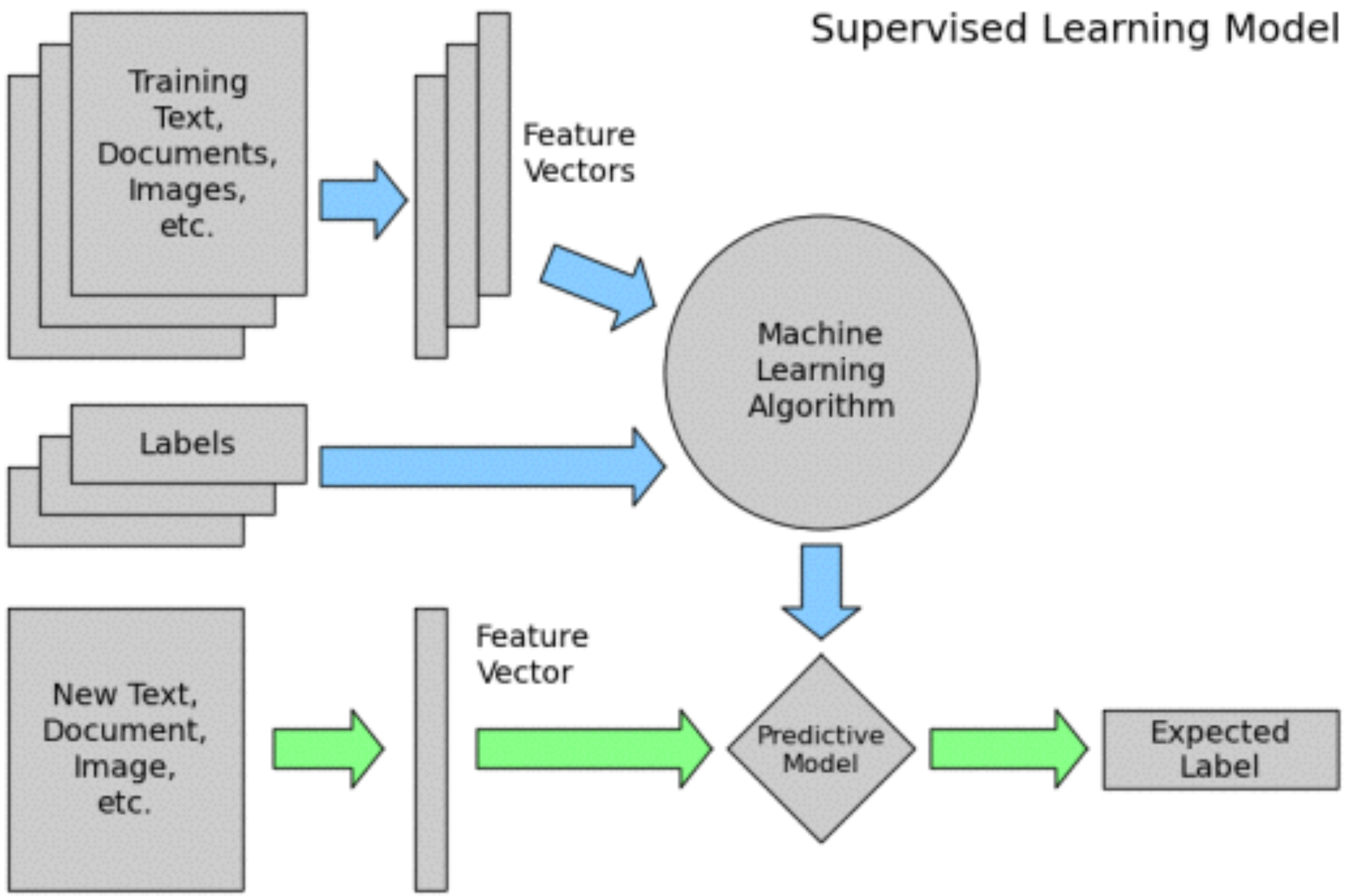
Fisher's *Iris* Data

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*

150 observations *(n = 150)*

4 features (p = 4)

response


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

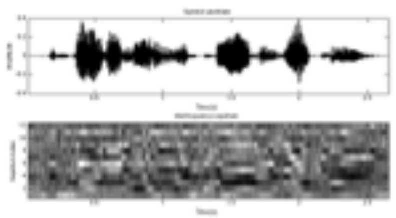


Figure 1: The original sound wave and its cepstral coefficients

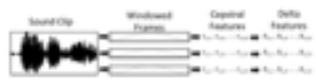


Figure 2: The feature extraction process

4. Discussion