

CS 105

# DATA ANALYSIS METHODS

---

# Outline

- Data Science
  - Introduction
  - Why Data Science
- Data Scientists
  - What do they do?
- Course Logistics

# Introductions ...

## Course Instructor

- Prof. Mariam Salloum
- **Office:** Bourns A (Room 159B)
- **Email:** [msalloum@cs.ucr.edu](mailto:msalloum@cs.ucr.edu)
- **Office Hours:** MWF 1 - 2 & by appointment
- **Webpage:** [www.cs.ucr.edu/~msalloum](http://www.cs.ucr.edu/~msalloum)

## Lab Instructor / TA

Al Amin Hossain

**Email:** [ahoss005@ucr.edu](mailto:ahoss005@ucr.edu)

We would like to learn a little about you. Please complete the pre-class survey.

<http://bit.ly/cs105-pre-survey>

# What is Data Science?

- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

# What is Data Science?

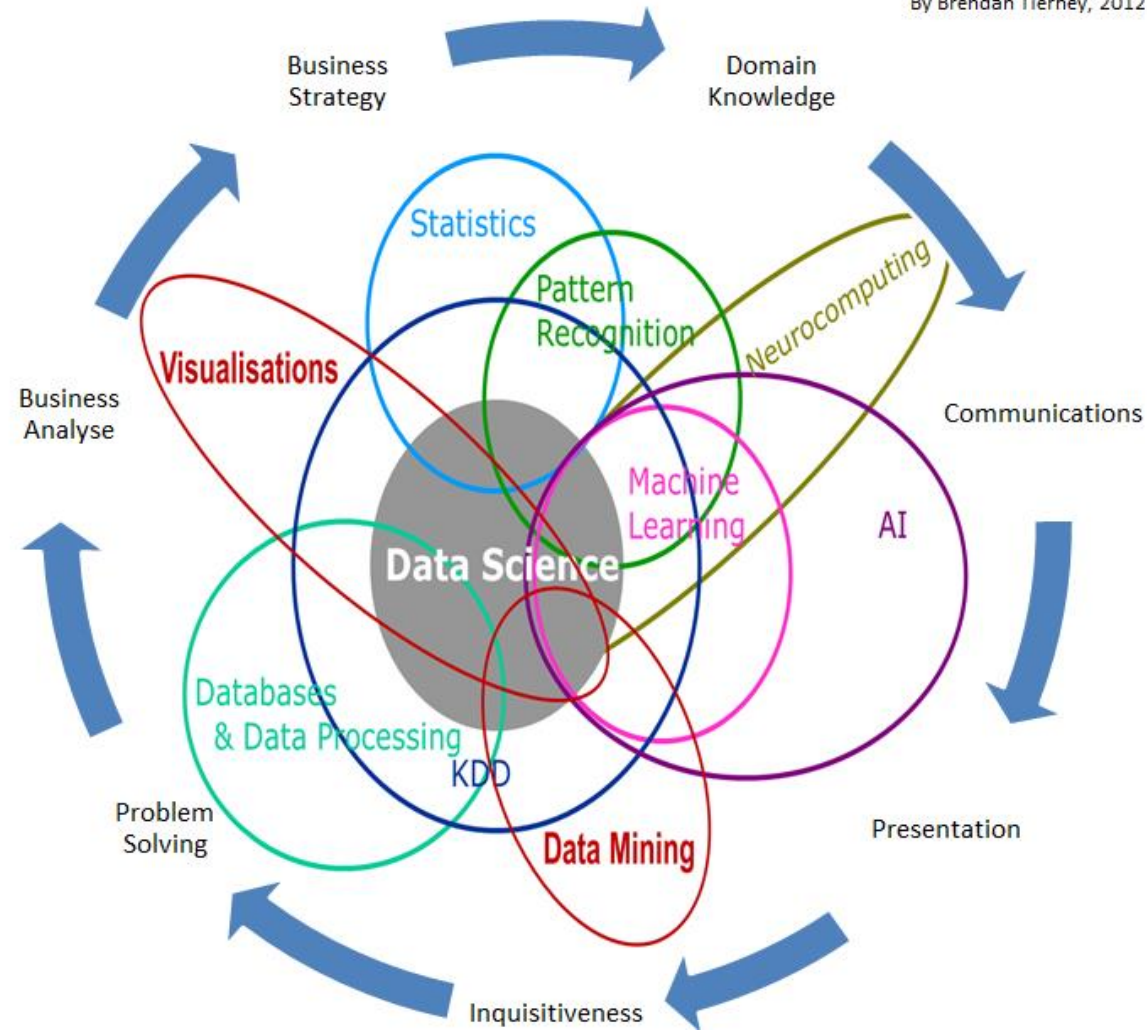
- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
- Computer Science
  - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
- Mathematics
  - Mathematical Modeling
- Statistics
  - Statistical and Stochastic modeling, Probability.

# Big Data and Data Science

- “... the sexy job in the next 10 years will be statisticians,” Hal Varian, Google Chief Economist
- The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute’s June 2011
- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...
- New degree programs, courses, boot-camps:
  - at most UCs, including UCR!!!
  - There are numerous MS programs as well
- There are a lot of internships and research opportunities:
  - Check this out: <http://www.dssgfellowship.org>

# Data Science

By Brendan Tierney, 2012



# What is Data Science?

...on any given day, a team member could author a **multistage processing pipeline in Python**, design a **hypothesis test**, perform a **regression analysis** over data samples with **R**, design and implement an **algorithm** for some data-intensive product or service in **Hadoop**, or **communicate** the results of our analyses to other members of the organization.

- *Jeff Hammerbacher describing the data science group he put together at Facebook.*

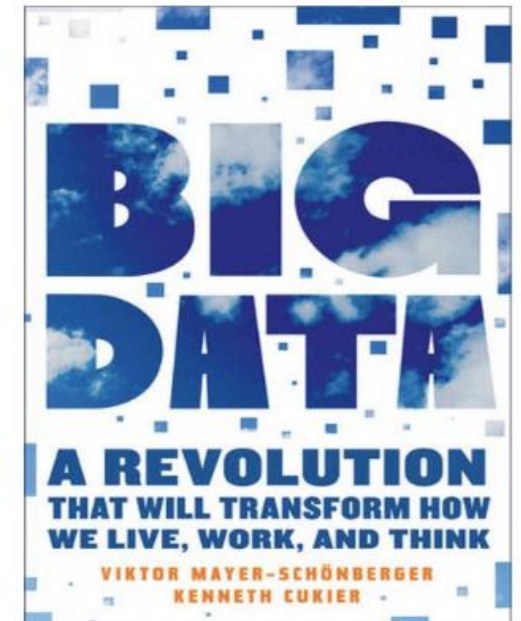
- *What is Data Science?*

*An O'Reilly Radar Report – Mike Loukides*



“The “data scientist,” which combines the skills of the **statistician, software programmer, infographics designer, and storyteller.**”

- *Big Data: A Revolution that Transform How We Live, Work, and Think, Viktor Mayer-Schönberger and Kenneth Cukier.*





# What is Data Science?

## Responsibilities



- Work with large, complex data sets. Solve difficult, non-routine analysis problems, applying advanced analytical methods as needed. Conduct analysis that includes data gathering and requirements specification, processing, analysis, ongoing deliverables, and presentations.
- Build and prototype analysis pipelines iteratively to provide insights at scale. Develop comprehensive knowledge of Google data structures and metrics, advocating for changes where needed for product development.
- Interact cross-functionally, making business recommendations (e.g., cost-benefit, forecasting, experiment analysis) with effective presentations of findings at multiple levels of stakeholders through visual displays of quantitative information.
- Research and develop analysis, forecasting, and optimization methods to improve the quality of Google's user facing products.



## Data Scientist, Engineering

Google Mountain View, CA, USA San Bruno, CA, USA  
Seattle, WA, USA + 1 more location

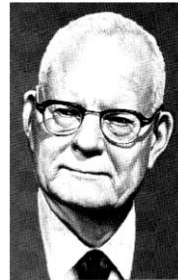
# Data Analysis Has Been Around for a While...

1939: "Quality Control"

1958: "A Business Intelligence System"

R.A. Fisher

W.E.  
Deming

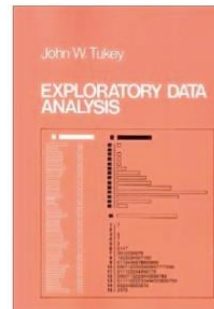


Peter Luhn

1997: "Machine Learning"

1977: "Exploratory Data Analysis"

1989: "Business Intelligence"



Howard  
Dresner

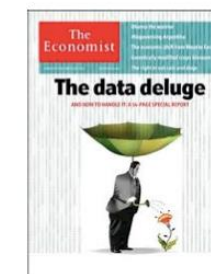
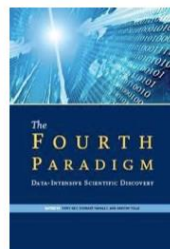


2010: "The Data Deluge"

2007: "The Fourth Paradigm"

2009: "The Unreasonable Effectiveness of Data"

1996: Google



# Data Science: Why all the Excitement?



Exciting new effective applications of data analytics

e.g.,  
Google Flu Trends:

Detecting outbreaks  
two weeks ahead  
of CDC data

New models are estimating  
which cities are most at risk  
for spread of the Ebola virus.

Prediction model is built on  
Various data sources,  
types and analysis.

# Why the all the Excitement?

## elections2012

Live results President Senate House Governor Choose your

### Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

**Luke Harding**

[guardian.co.uk](http://guardian.co.uk), Wednesday 7 November 2012 10.45 EST



## Data and Election 2012 (cont.)

- ...that was just one of several ways that Mr. Obama's campaign operations, some unnoticed by Mr. Romney's aides in Boston, **helped save the president's candidacy**. In Chicago, the campaign recruited a team of behavioral scientists to build an **extraordinarily sophisticated database**
- ...that allowed the Obama campaign not only to alter the very nature of the electorate, making it younger and less white, but also to create a portrait of shifting voter allegiances. **The power of this operation stunned Mr. Romney's aides on election night**, as they saw voters they never even knew existed turn out in places like Osceola County, Fla.

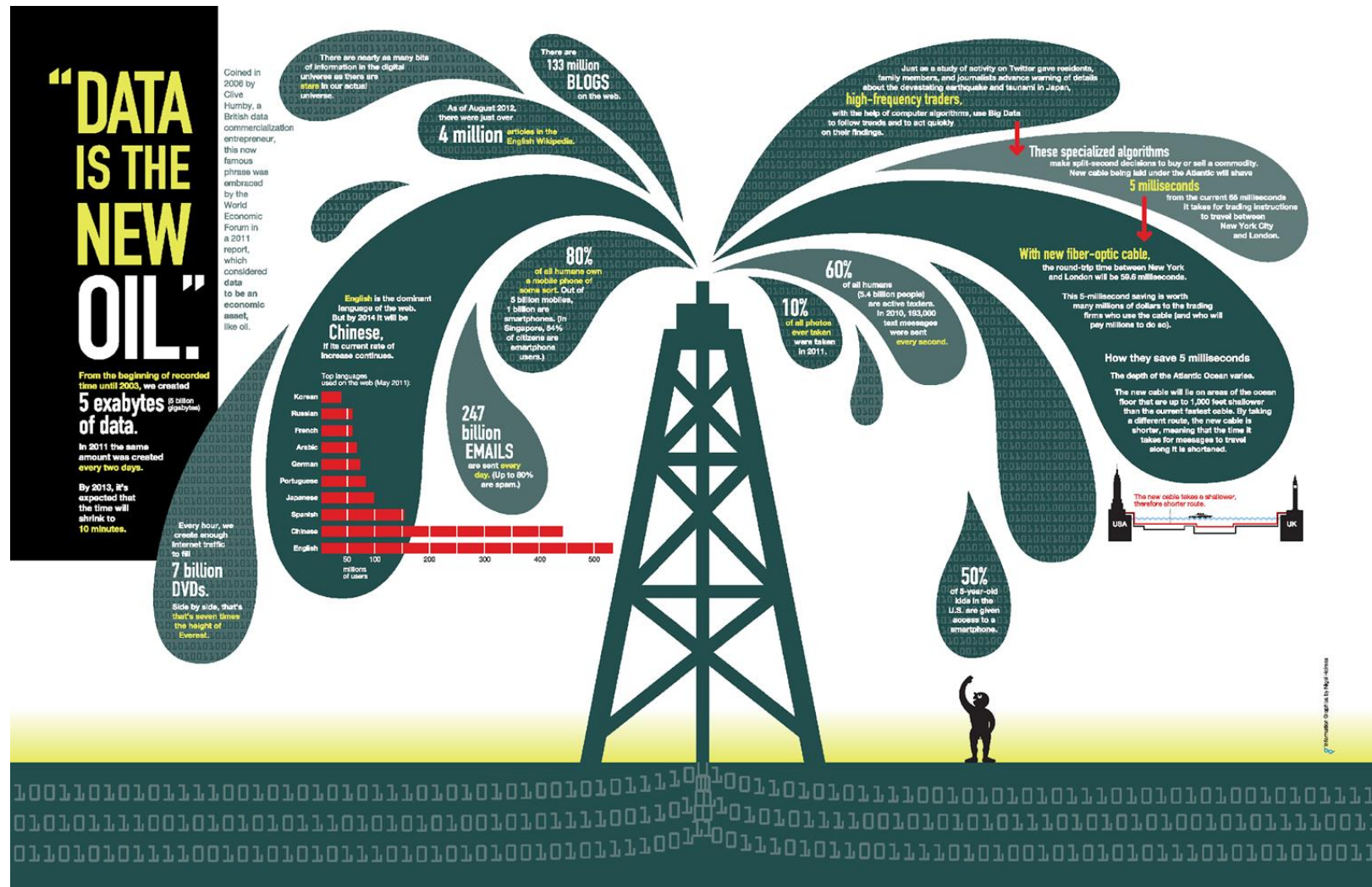
-- New York Times, Wed Nov 7, 2012

- The White House Names Dr. DJ Patil as the First U.S. Chief Data Scientist, Feb. 18<sup>th</sup> 2015



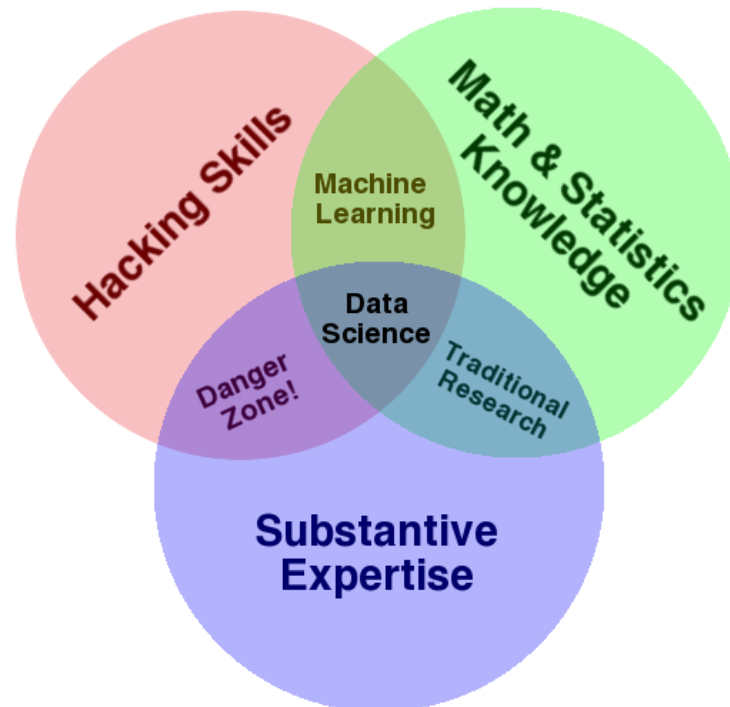
# “Data is the New Oil”

## – World Economic Forum 2011



# Data Science – A Definition

**Data Science** is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with **data** to **create data products**.



# Data All Around

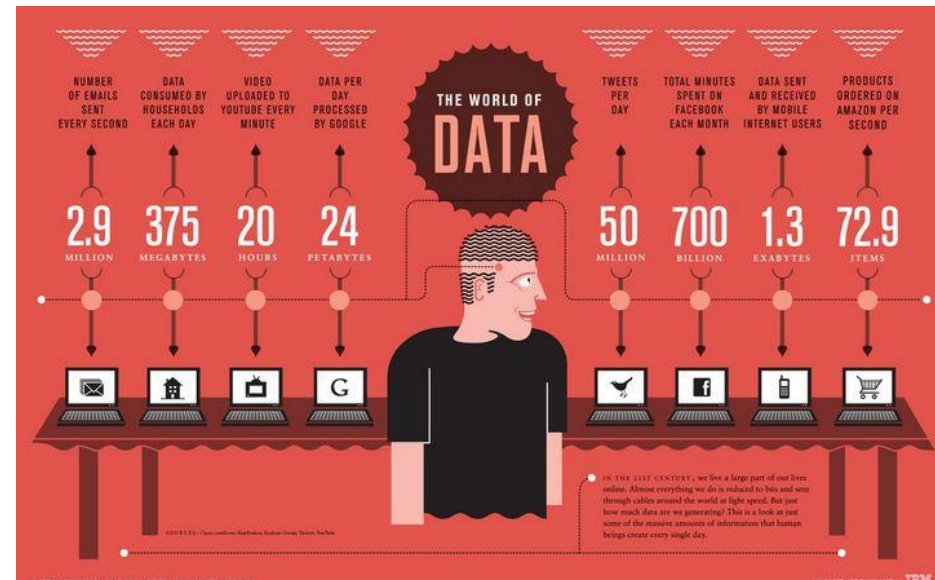
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network





# How Much Data Do We have?

- Google processes 20 PB a day (2008)
  - Facebook has 60 TB of daily logs
  - eBay has 6.5 PB of user data + 50 TB/day (5/2009)
  - 1000 genomes project: 200 TB
- 
- Cost of 1 TB of disk: \$35
  - Time to read 1 TB disk: 3 hrs  
(100 MB/s)



# Big Data

- Big Data is any data that is expensive to manage and hard to extract value from
- **Volume**
  - The size of the data
- **Velocity**
  - The latency of data processing relative to the growing demand for interactivity
- **Variety and Complexity**
  - the diversity of sources, formats, quality, structures.

# Types of Data We Have

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), ...
- Streaming Data
- You can afford to scan the data once

# Logistics

- Class Attendance is expected !
- Lectures
  - Class will meet MWF 11:00 AM - 11:50 AM
  - Lectures will include PPT slides and examples in Jupyter notebooks
- Labs
  - Tue 11:00 AM - 01:50 PM
  - Thur 02:00 PM - 04:50 PM
- Lab attendance is required and points will be deducted for unexcused absences

# Course Logistics (1/2)

- Lets go over the syllabus ...
- We will use the following tools to manage the course:
- **iLearn** : Will be used to turn-in programming assignments and post grades
- **CampusWire**: Will be used for class discussions. We will enroll everyone to the course discussion page on the 2nd week of classes, or you can enroll yourself via code.
- **Google Drive**: Will use a shared drive to post lecture slides, labs, reading material, solutions, etc. Be sure to use your UCR email to gain access to the Google drive.
- You can find the links on our course page [www.cs.ucr.edu/~msalloum/Teaching](http://www.cs.ucr.edu/~msalloum/Teaching)

Make sure you can  
login to iLearn,  
and access this  
course.



# Course Logistics (2/2)

- Lets go over the syllabus ...
- Class is composed of
  - Labs x 8
  - Midterms x 2
  - Final Project

## Grading

Item	Percentage
Labs	40%
Midterms	35%
Final Project	25%
<b>Total</b>	<b>100%</b>

- Labs
  - We will have a lab assignment each week
  - Labs will be turned in online and demoed to the TA
  - You have upto 1 – week to complete the lab assignment. Incomplete labs must be demoed to the TA at the beginning of the next lab session.
- Midterms
  - We will have 2 in-class written midterms
  - Midterms are closed book/ notes / electronics
- Final Project
  - We will have a team-based final project that is composed of several phases.
    - Phase 1 – project proposal
    - Phase 2 – data gathering and exploration
    - Phase 3 – data analysis
    - Phase 4 – Final report / presentation

# Final Project

- The final project is team-based because
    - Such projects are often team based and its important to learn the skills to communicate and work with others
  - The final project will be completed in several phases:
  - Phase 1 – Project Proposal
    - Find a team member that is excited about the same project idea
    - Find a team member that is in the same lab section and can meet outside of lab
  - Phase 2 – Data Gathering / Exploration
    - Gather data using a crawler or API from 1+ data sources
    - Based on the complexity, you may need to use multiple data sources.
    - Clean, integrate, and summarize / visualize the dataset
  - Phase 3 – Preliminary Data Analysis
    - Explore the data further and start preliminary analysis using ML techniques
  - Phase 4 – Final Data Analysis / Jupyter Notebook / Presentation
    - Final results, visualizations, etc.
- \*\* Results will be presented as a Jupyter notebook \*\***

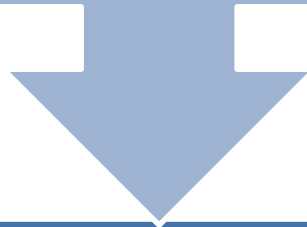
# Suggested Texts

- Due to the rapidly evolving nature of the material, there is no single textbook that covers the course in its entirety.
- The following textbook covers fundamental concepts for 'dealing with data'
  - Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, 1st Edition by Foster Provost, Tom Fawcett
  - Data Mining and Analysis: Fundamental Concepts and Algorithms Mohammed J. Zaki and Wagner Meira Jr
  - The Wall Street Journal Guide to Information Graphics: The Dos and Dont's of Presenting Data, Facts, and Figures by Dona M. Wong
  - Doing Data Science: Straight Talk from the Frontline 1st Edition, Kindle Edition by Cathy O'Neil , Rachel Schutt
- The class schedule will reference the appropriate reading material for each topic.
- Please note, these books are available online via the UCR Library.
  - Note – you must be on campus or connected via VPN



# How the course works

This course provides an overview of Data Science, covering a broad selection of key challenges in and methodologies for working with big data.



Topics to be covered include:

Data  
gathering

Data storage

Business  
intelligence

Basic  
statistics

Visualization

Data  
Analysis

# High level course goals and learning objectives

- Students will acquire a working knowledge of data science through hands-on projects and case studies in a variety of business,, engineering, social sciences, or life sciences domains.
- Professional skills, such as
  - Communication
  - Presentation
  - Teamwork,
  - Storytelling with data, will be fostered

# Computing background for the course

- **Prerequisite:** CS 14
- You will need to be able to get your hands dirty playing with, processing, and plotting data using Python.
  - The course will not teach you Python. You should be able to pick-up the syntax through outside readings and practice.
- Why Python? It is more commonly used in industry
  - Kdnuggets Python VS R [LINK](#)

# What Python skills do we need?

- Data exploration/processing
- The Python scientific stack
  - numpy, scipy, pandas, matplotlib
- Formatting and presenting results:
  - JuPyter notebooks

# What statistics skills do we need?

- Basics:
  - Estimation
  - Hypothesis testing
  - Inference
- Applied statistics:
  - Classification

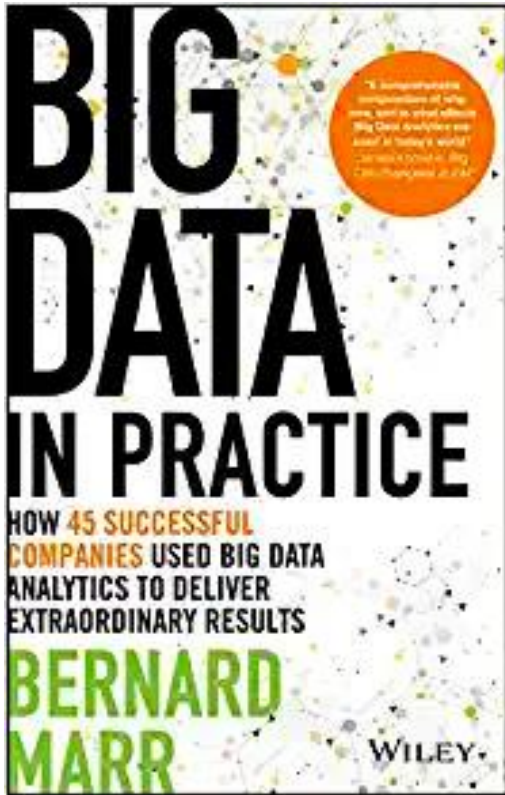
# What other skills do we need?

- Combining statistics and programming:
  - Machine learning
- Communicating your results
  - Use Jupyter Lab and notebooks
  - Data Visualization
- Adapting to different types of data
  - Evaluating data sets
  - Cleaning messy data

# Python

- If you don't know Python, then this is a good place to start:
  - Learning Python by Mark Lutz O'Reilly Media, September 2013
  - Available online via the UCR Library (must be on-campus to access)
- We will also be making use of iPython/Jupyter notebooks.
  - Makes developing Python code easier and sharing / visualizing results
  - Each student will be given access to a Jupyter notebook

Lets take a look at a few use-cases of data



Bernard Marr

[LINK:](#)

<https://ebookcentral.proquest.com/lib/ucr/reader.action?docID=4455265>

I suggest you browse the book to look at a few case studies

Select ones that interest you



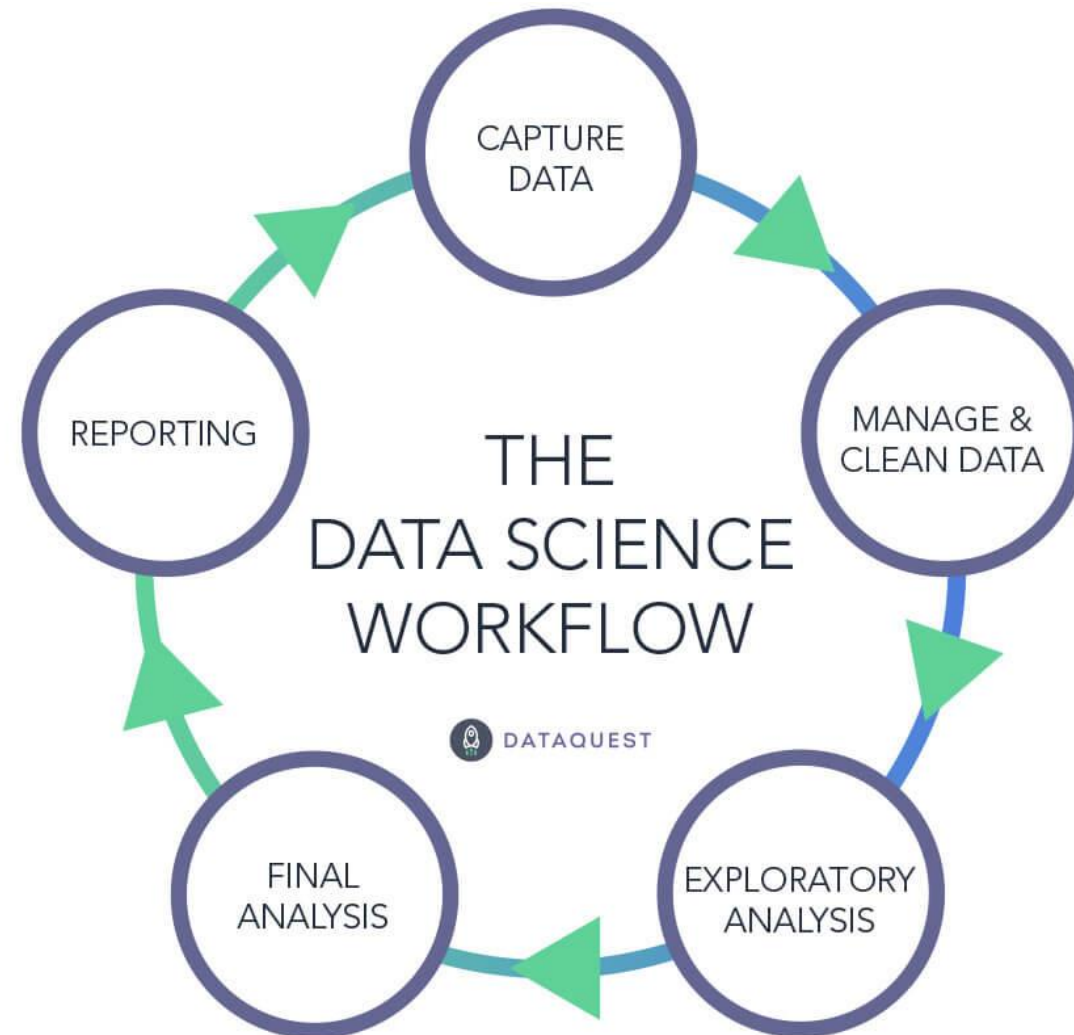
# Netflix

- How does Netflix use data?
- Obviously, Netflix relies on customer data (what customers watch, how long they watch a show, what they search for, etc.) to understand the user's preferences.
- Data is used to build a profile on each customer / user and make recommendations.
- In 2007 the Netflix Prize, offering \$1M to the group that could come up with the best algorithm predicting how customers would rate a movie based on prior ratings.

# Types of Data

- Data Science Lifecycle
- Types of data :
  - Tables
  - Graphs & networks
  - Time-series
  - textual data
- And go over
  - simple statistics that describe the data, and
  - data analysis techniques of each type of data.

# Data Science Lifecycle



# Exploratory Data Analysis (EDA)

- EDA is applied to explore a data set, which means:
  - Use methods that help you understand the data:
    - to help you understand the events that generated the data
    - detect outliers (e.g. assess data quality)
    - test assumptions (e.g. normal distributions or skewed?)
    - identify useful raw data & transforms (e.g.  $\log(x)$ )
- Before diving into EDA, we need to understand different types of data.

# Types of Data I

- Categorical (Qualitative)

- Nominal scales – number is just a symbol that identifies a quality
  - 0=male, 1=female
  - 1=green, 2=blue, 3=red, 4=white
- Ordinal – rank order

- Quantitative (continuous and discrete)

- Interval – units are of identical size (i.e. Years)
- Ratio – distance from an absolute zero (i.e. Age, reaction time)

# Quantitative data is a measurement

- Every measurement has 2 parts:
  - The True Score (the actual state of things in the world)
  - ERROR! (mistakes, bad measurement, report bias, context effects, etc.)

$$X = T + e$$

- EDA can help identify outliers, etc.

# Variable Summaries

- Indices of central tendency:
  - Mean – the average value
  - Median – the middle value
  - Mode – the most frequent value
- Indices of Variability:
  - Variance – the spread around the mean
  - Standard deviation
  - Standard error of the mean (estimate)

Variable summaries helps us understand the data quickly and detect outliers.

Don't need to report all of these: Bottom line...do these numbers make sense???

# The Mean

Mean = sum of all scores divided by number of scores

$$\frac{X_1 + X_2 + X_3 + \dots X_n}{n}$$

Is the mean the best summarization of a column?

What potential issues can arise?

Subject	before	during	after
1	3	2	7
2	3	8	4
3	3	7	3
4	3	2	6
5	3	8	4
6	3	1	6
7	3	9	3
8	3	3	6
9	3	9	4
10	3	1	7

Sum =	30	50	50
/n	10	10	10
Mean =	3	5	5



# The Mean

Subject	before	during	after
1	3	2	1000
2	3	8	4
3	3	7	3
4	3	2	6
5	3	8	4
6	3	1	6
7	3	9	3
8	3	3	6
9	3	9	4
10	3	1	7

Sum =	30	50	1043
/n	10	10	10
Mean =	3	5	104.3

One value ( possibly an outlier) can influence the mean.

In this case, the median could be a better summary.

# The Variance: Sum of the squared deviations divided by number of scores

Subject	before	during	after
1	3	2	7
2	3	8	4
3	3	7	3
4	3	2	6
5	3	8	4
6	3	1	6
7	3	9	3
8	3	3	6
9	3	9	4
10	3	1	7

Sum = 30 50 50

/n 10 10 10

Mean = 3 5 5

Before - mean	Before - mean <sup>2</sup>	During - mean	During - mean <sup>2</sup>	After - mean	After - mean <sup>2</sup>
0	0	-3	9	2	4
0	0	3	9	-1	1
0	0	2	4	-2	4
0	0	-3	9	1	1
0	0	3	9	-1	1
0	0	-4	16	1	1
0	0	4	16	-2	4
0	0	-2	4	1	1
0	0	4	16	-1	1
0	0	-4	16	2	4

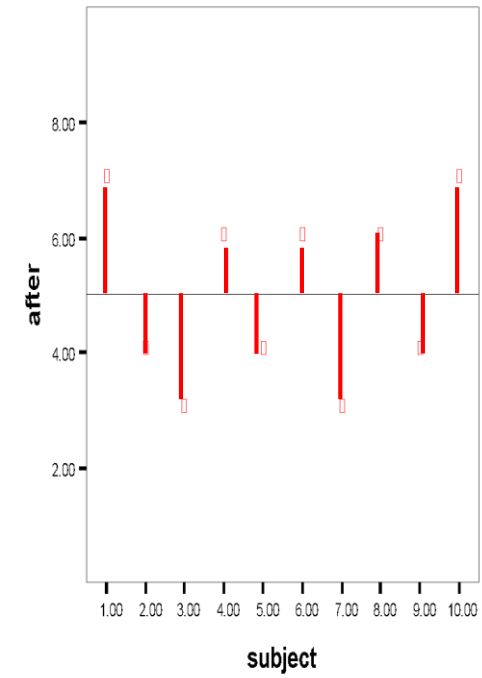
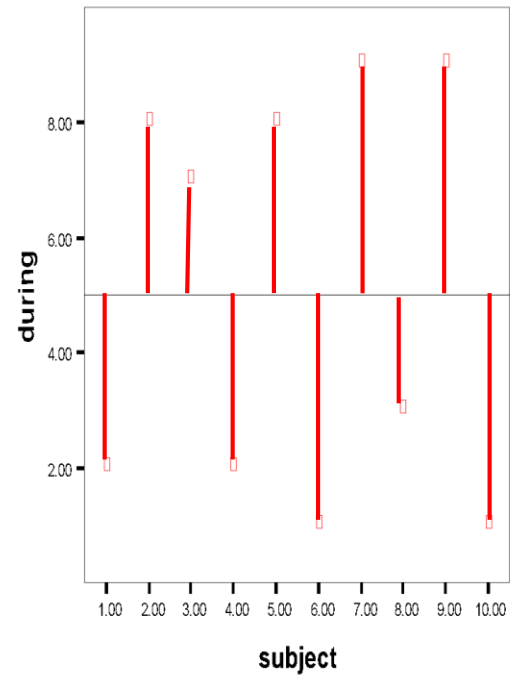
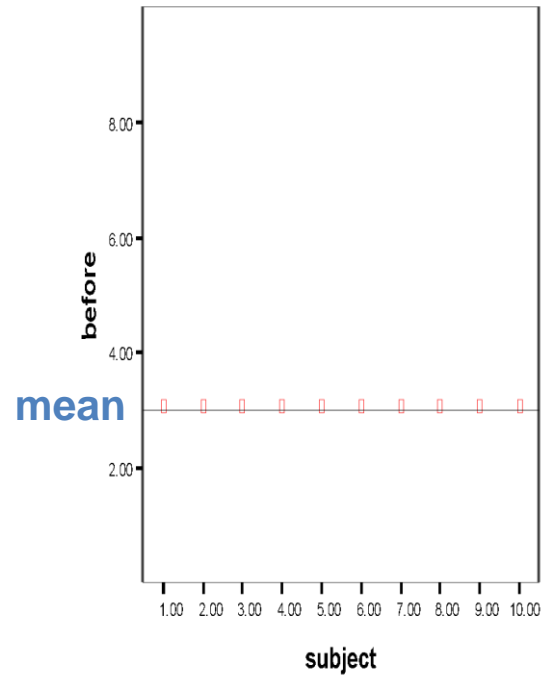
0 0 0 108 0 22

10\* 10 10

VAR = 0 10.8 2.2

\*actually you divide by n-1 because it is a sample and not a population, but you get the idea...

# Variance continued

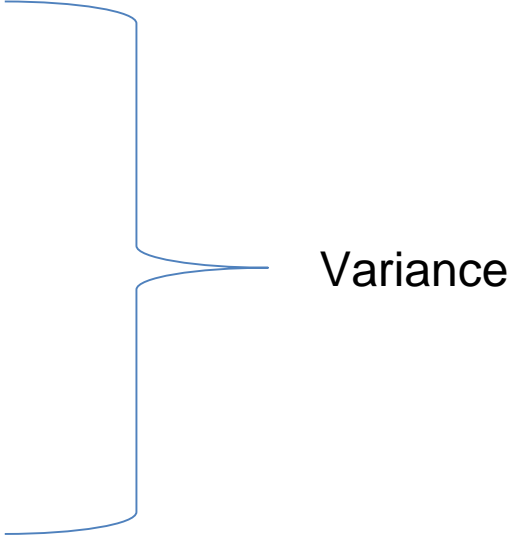


# Standard deviation

- Variance, as calculated earlier, is arbitrary.
- What does it mean to have a variance of 10.8? Or 2.2? Or 1459.092? Or 0.000001?
- Nothing. But if you could “standardize” that value, you could talk about any variance (i.e. deviation) in equivalent terms.
- Standard Deviations are simply the square root of the variance

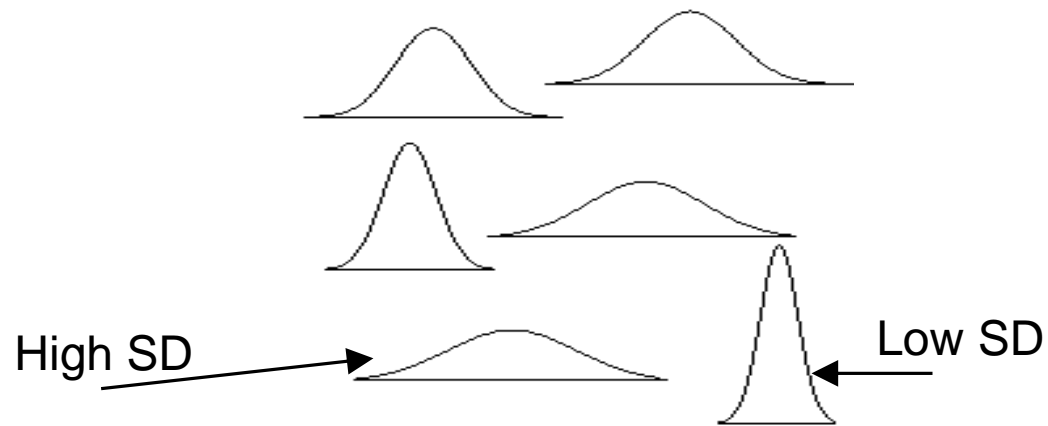
# Standard deviation

The process of standardizing deviations goes like this:

1. Score (in the units that are meaningful)
  2. Mean
  3. Each score's deviation from the mean
  4. Square that deviation
  5. Sum all the squared deviations (Sum of Squares)
  6. Divide by  $n$  (if population) or  $n-1$  (if sample)
  7. Square root – now the value is in the units we started with!!!
- 
- Variance

# Interpreting standard deviation (SD)

- First, the SD will let you know about the distribution of scores around the mean.
- High SDs (relative to the mean) indicate the scores are spread out
- Low SDs tell you that most scores are very near the mean.



# Interpreting standard deviation (SD)

- Second, you can then interpret any individual score in terms of the SD.
- For example: mean = 50, SD = 10      versus      mean = 50, SD = 1
- A score of 55 is:
  - 0.5 Standard deviation units from the mean (not much)      OR
  - 5 standard deviation units from mean (a lot!)

# Standardized scores (Z)

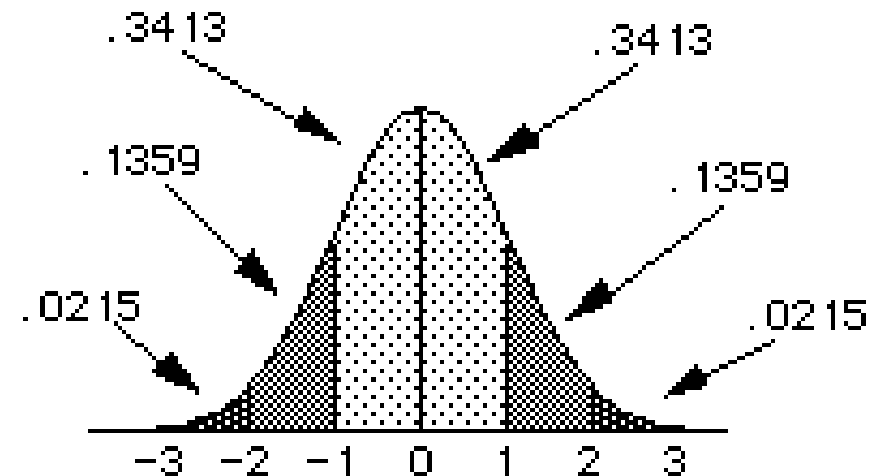
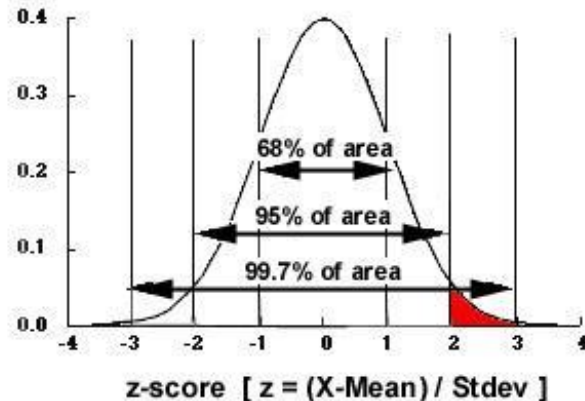
- Third, you can use SDs to create *standardized* scores – that is, force the scores onto a normal distribution by putting each score into units of SD.
- Subtract the mean from each score and divide by SD

$$Z = (X - \text{mean})/\text{SD}$$



# Standardized normal distribution

- ALL Z-scores have a mean of 0 and SD of 1. Nice and simple.
- From this we can get the proportion of scores anywhere in the distribution.

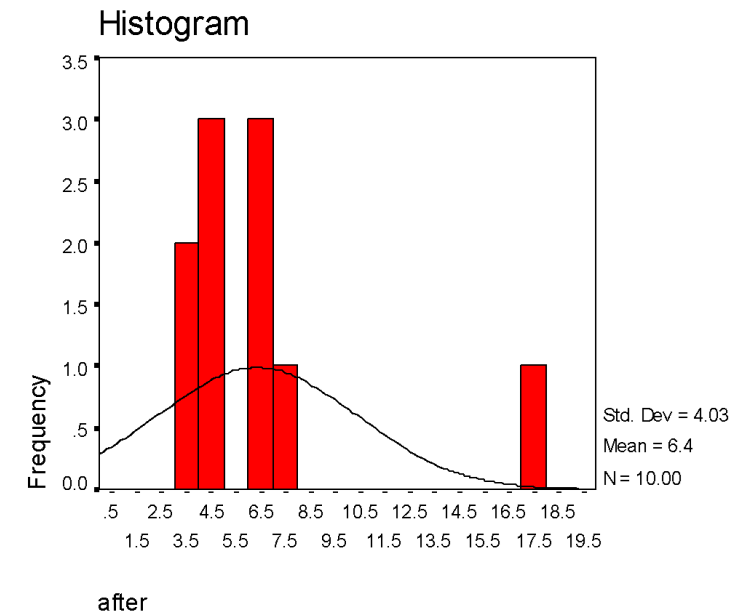
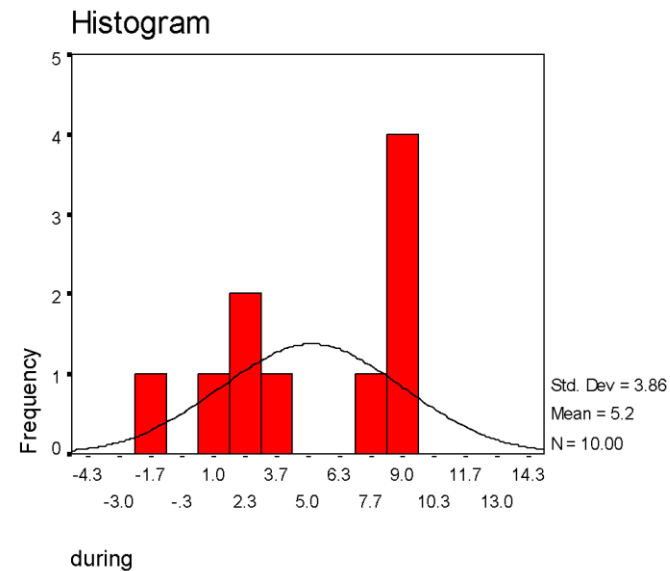
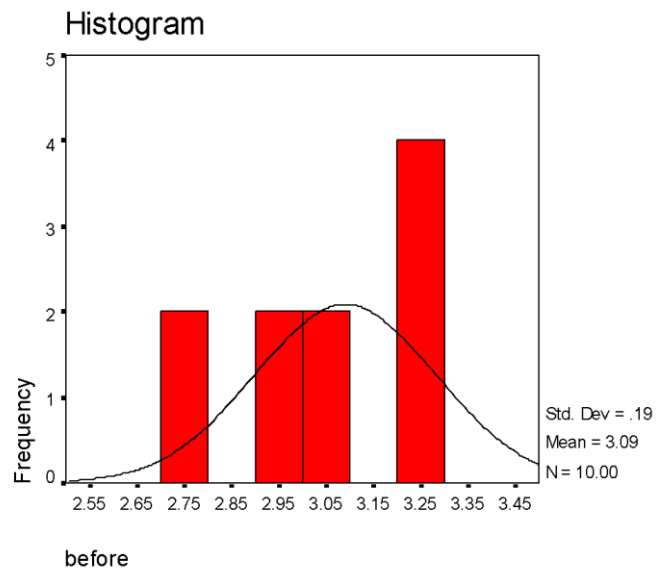


# Summary

- Examine every variable for:
  - Out of range values
  - Normality
  - Outliers
- Visual display of data is useful as well
  - Histograms
  - Stem and Leaf plots
  - Boxplots
  - QQ Plots

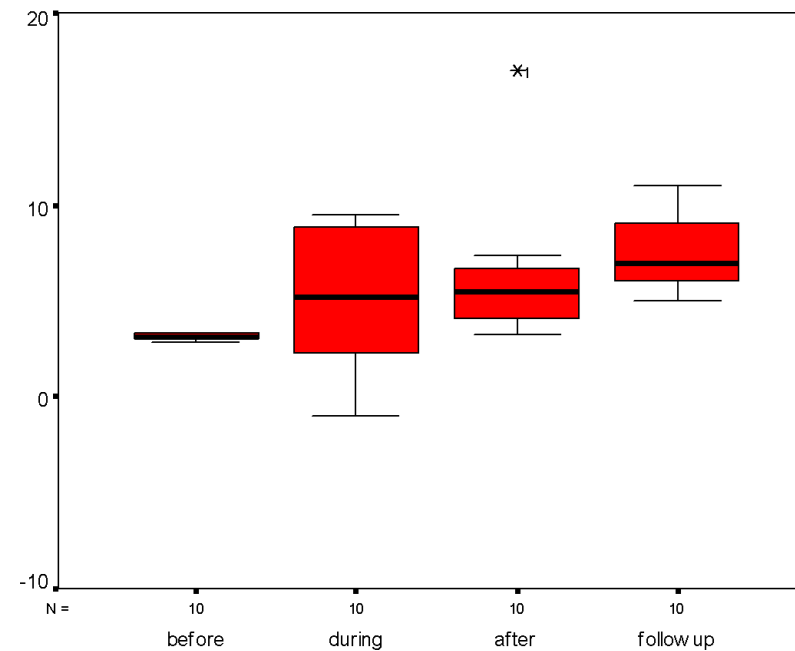
# Histograms

- # of bins is very important:



# Boxplots

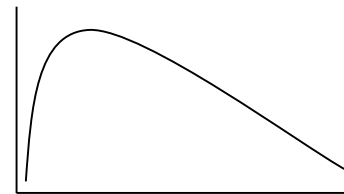
- Upper and lower bounds of boxes are the 25<sup>th</sup> and 75<sup>th</sup> percentile (interquartile range)
- Whiskers are min and max value unless there is an outlier
- An outlier is beyond 1.5 times the interquartile range (box length)



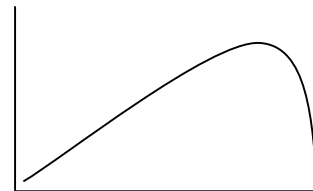
## So, now what?

- If you find a mistake, fix it.
- If you find an outlier, trim it or delete it.
  - Otherwise the outlier will affect your analysis.
- If your distributions are askew, transform the data

- Positive skew is reduced by using the square root or log



- Negative skew is reduced by squaring the data values



# Graph Data

- **What is a Graph?**

- *A graph is a collection of interconnected nodes*

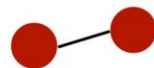
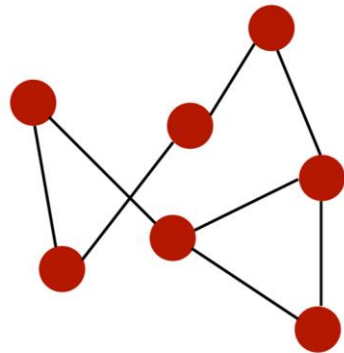
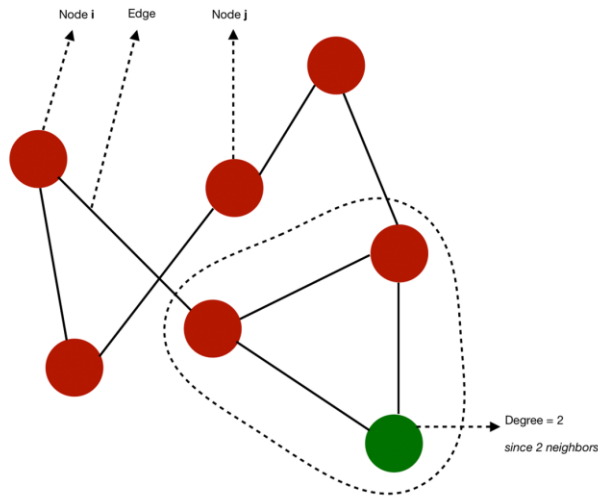
- An “**edge**” is a connection or tie between two n

- A **neighborhood** N for a vertex or node is the set of its immediately connected nodes.

- **Degree:** The degree  $k_i$  of a vertex or node is the number of other nodes in its neighborhood.



## Graph Data (Cont.)



- A **degree** of a node is its number of neighbors
- A graph is **complete** if all nodes have  $n-1$  neighbors. This would mean that all nodes are connected in every possible way.
- A **path** from  $i$  to  $j$  is a sequence of edges that goes from  $i$  to  $j$ . This path has a **length** equal to the number of edges it goes through.
- If all the nodes can be reached from each other by a given path, they form a **connected component**. A graph is **connected** if it has a single connected component

# Types of graphs

- Graphs can be used to represent :
  - social networks
  - web pages
  - biological networks
- Social networks are essentially graphs
  - a social structure made up of individuals (or organizations) called "nodes", which are tied (connected) by one or more specific types of interdependency, such as friendship, common interest
- **Social Network analysis** - views social relationships in terms of network theory consisting of *nodes* and *ties* (also called *edges*, *links* or *connections*).

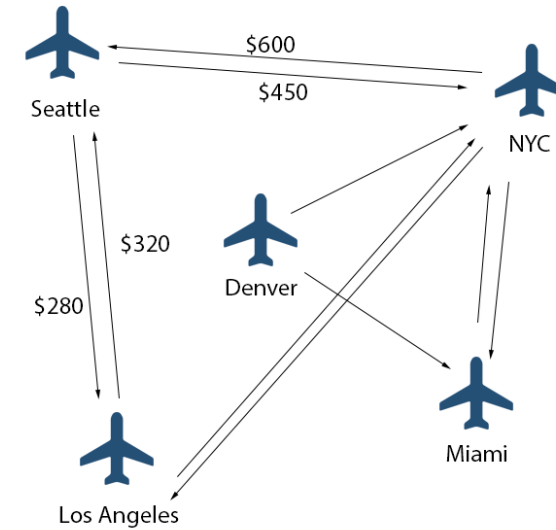


# Some concepts

- What kind of analysis can we perform on a graph?
  - study **topology and connectivity**
  - **community** detection
  - identification of **central nodes**
  - predict missing nodes
  - predict missing edges
  - ...

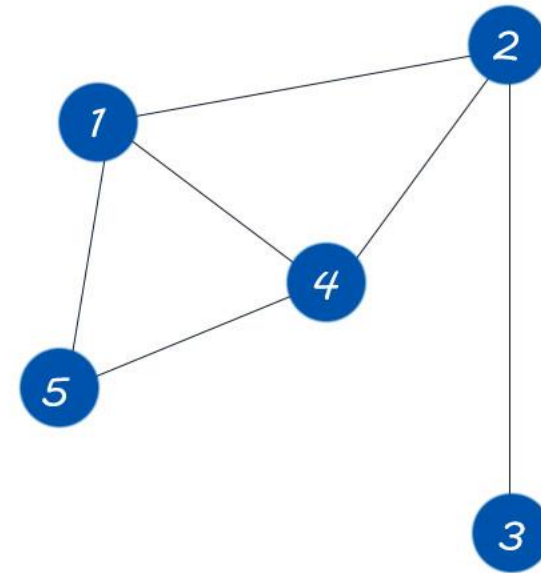
# Some concepts

- In an **undirected** graph or network, the edges are reciprocal—so if A is connected to B, B is by definition connected to A.
- In a **directed** graph or network, the edges are not necessarily reciprocal—A may be connected to B, but B may not be connected to A (think of a graph with arrows indicating direction of the edges.)



# How to represent a graph

- A graph can be represented using an adjacency matrix
- For each possible pair in the graph, set it to 1 if the 2 nodes are linked by an edge. A is symmetric if the graph is undirected.
- The adjacency matrix is sparse, hence we use a efficient representation
  - Example
    - 1 : [ 2, 3, 4]
    - 2 : [1, 3]
    - 3 : [2, 4]
    - ....

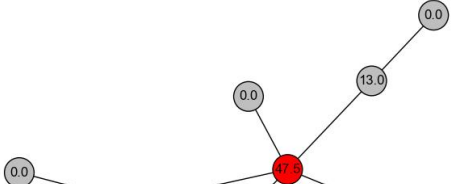


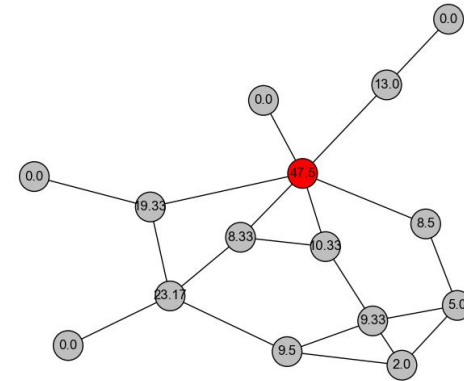
$$A = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix} \end{matrix}$$

# Graph algorithms

- There are 3 main categories of graph algorithms that are currently supported in most frameworks (networkx in Python, or in Neo4J for example) :
- **Pathfinding**: identify the *optimal path* depending on availability and quality for example. We'll also include **search algorithms** in this category. This can be used to identify the *quickest route or traffic routing* for example.
- **Centrality**: determine the importance of the nodes in the network. This can be used to *identify influencers* in social media for example or identify potential attack targets in a network.
- **Community detection**: evaluate how a *group is clustered*. This can be used to *segment customers* and *detect fraud* for example.

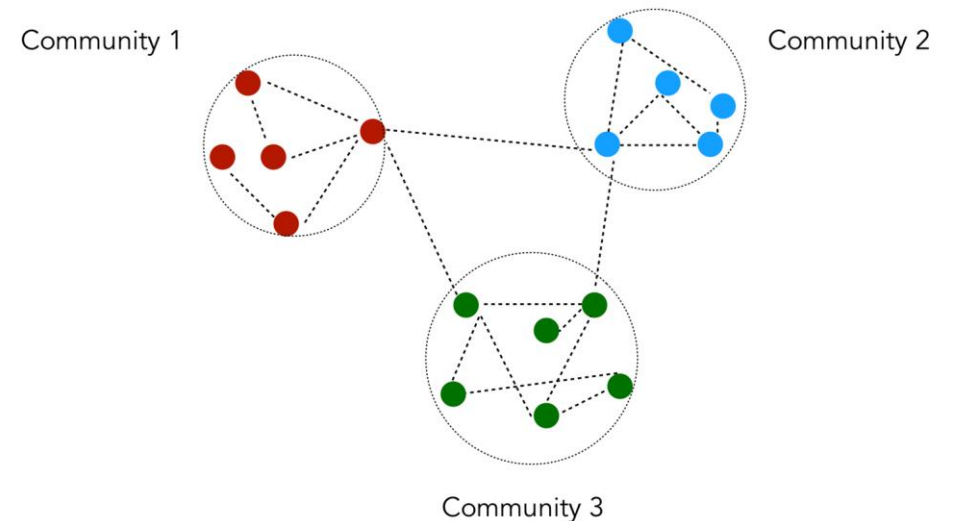
# Centrality algorithms

- Centrality measures **how important a node is**.
  - This is not a clear definition, but it's useful when we want to identify
    - important web pages,
    - Influential users in a social network
    - bottlenecks in transportation networks...
- 
- A small network diagram illustrating a central node. A red node is connected to four gray nodes. The gray nodes have numerical values: 0.0, 13.0, 0.0, and 0.0. The red node is the central hub, and the gray nodes are peripheral nodes.



# Community Detection

- Community detection partitions the nodes into a several **groups** according to a given **quality criterion**.
- It is typically used to identify social communities, customers behaviors or web pages topics.
- A *community* is a set of connected nodes.



# Time Series Data

- A **time series** is a series of data points indexed (or listed or graphed) in time order.
- Example of time series
  - Opening and closing prices of stocks
  - Measurement of temperature on a daily or hourly basis
  - Measurement of air quality throughout the year
- Objectives of time series analysis
  - Compact description of the data
  - interpretation
  - forecasting

# Textual Data

## Topics:

People  
Events  
Products  
Services, ...



## Sources:

Blogs  
Microblogs  
Forums  
Reviews ,...





# Textual Data Analysis

- Category or topic summarization / concept extraction - determining the topic(s) being discussed in a document
- Document clustering - grouping and categorizing documents
- Natural Language Processing (NLP)- low-level language processing and understanding (tagging part of speech)
- Sentiment analysis - understanding the sentiment and object in a body of text
- Information extraction - identification and extraction of relevant facts and relationships from unstructured text
- Sentence completion / suggestion - auto completion or suggestion of similar questions, replies, etc.