Slide01 Introduction to Data Science

Prof. Ryan Leong Hou U

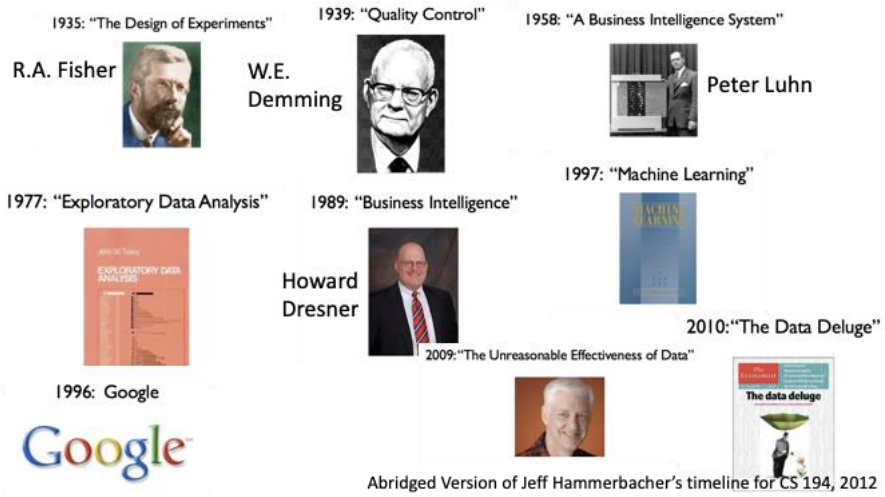
Interim head of Centre for Data Science

(Part of the materials from COMPSCI 194 -UC Berkeley)

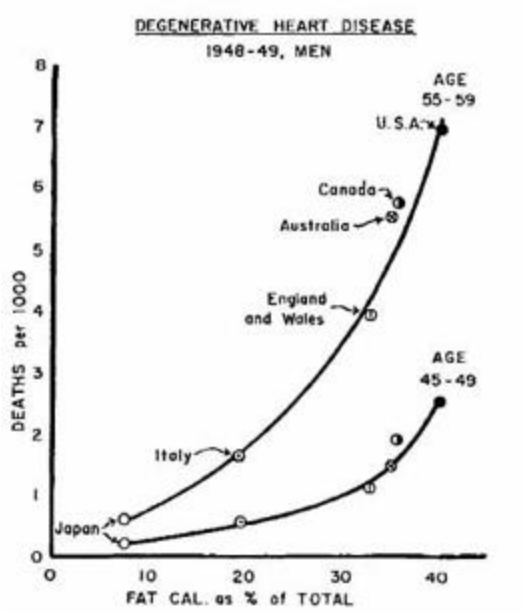
Outline

- Data Science – Why all the excitement?
- Where does data come from
- So what is Data Science
- Doing Data Science About the course
  - What we'll cover
  - Data Science first, Big data later
  - Requirements, workload etc.

Data Analysis Has Been Around for a While



*1 - **Data Analysis Has Been Around for a While***



*2 - **Data makes everything clearer***

Seven Countries Study (Ancel Keys, UCB 1925,28) 13,000 subjects total, 5-40 years follow-up.

Data Science: Why all the Excitement?

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.

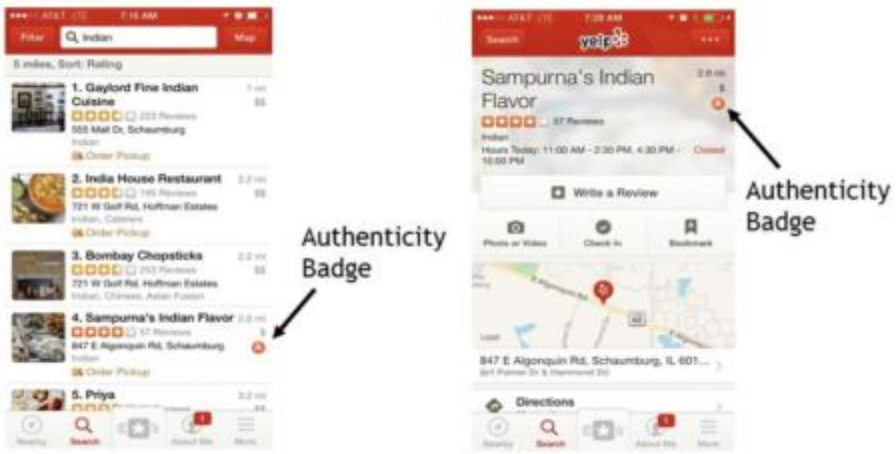
*3 - **Data and Election***

"Eat, Rate, Love" — An Exploration of R, Yelp, and the Search for Good Indian Food (*Beginner*)

- *The number of restaurant reviews by a single person of a particular cuisine (in this case, Indian food). He was able to justify this parameter by looking at reviewers of other cuisines, such as Chinese food.*

- *The apparent ethnicity of the reviewer in question. If the reviewer had an Indian name, he could infer that they might be of Indian ethnicity, and therefore more familiar with what constituted good Indian food.*

Yelp's data has become popular among newcomers to data science. You can access it here. Find out more about Robert's project here.

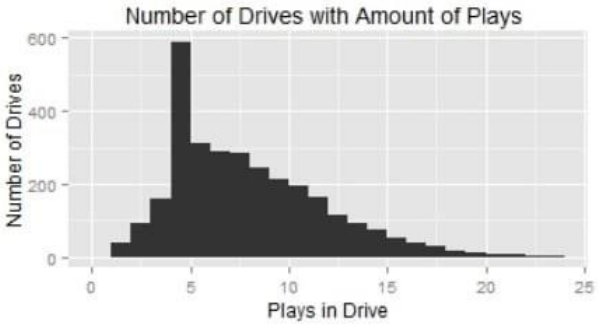


***Third and Goal* (Intermediate)**

The intersection of sports and data is full of opportunities for aspiring data scientists. A lover of both, Divya Parmar decided to focus on the NFL for his capstone project during Springboard's Introduction to Data Science course.

Divya's goal: to determine the efficiency of various offensive plays in different tactical situations. Here's a sample from Divya's project write-up.

"To investigate 3rd down behavior, I obtained play-by-play data from Armchair Analysis; the dataset was every play from the first eight weeks of this NFL season. ..."

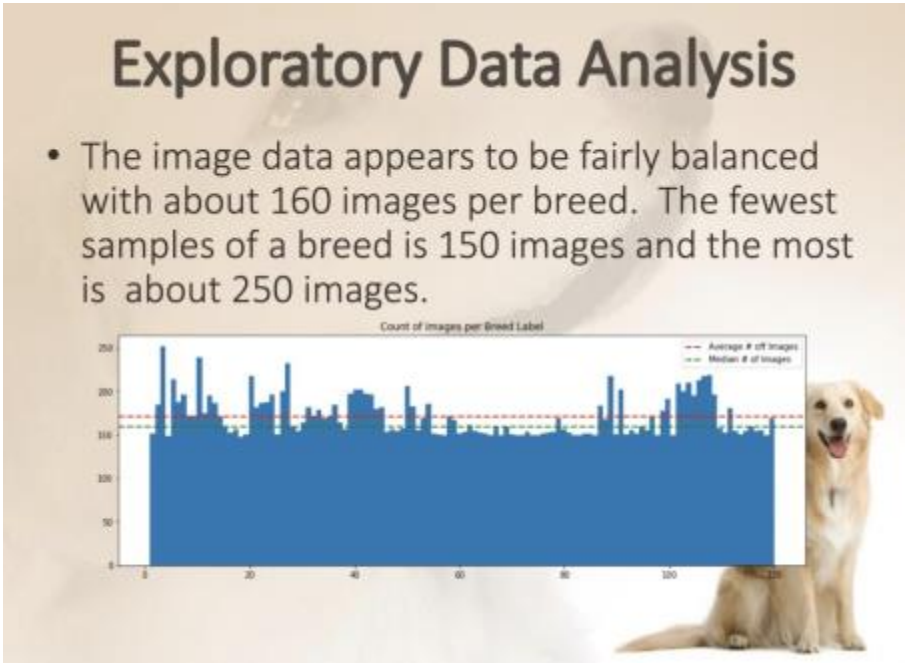


***Who's a Good Dog?* (Intermediate)**

Garrick Chu, another Springboard alum, chose to work on an image classification project, identifying dog breeds using neural networks.

- *Working with large data sets*

- *Effective processing of images (rather than traditional data structures)*
- *Network design and tuning*
- *Avoiding over-fitting*
- *Transfer learning (combining neural nets trained on different data sets)*
- *Performing exploratory data analysis to understand model outputs that people can't directly interpret*



***Amazon vs. eBay* (Advanced)**

Ever pulled the trigger on a purchase only to discover shortly afterward that the item was significantly cheaper at another outlet?

In support of a Chrome extension he was building, Chase Roberts decided to compare the prices of 3,500 products on eBay and Amazon. With his biases acknowledged, Chase walks readers of this blog post through his project, starting with how he gathered the data and documenting the challenges he faced during this process.

	A	B	C	D	E
1	UPC	Amazon Price	Ebay Price	Category	Price Diff
2	746591610623	$27.99	$21.85	BISS Basic	-21.94%
3	746591610586	$47.99	$31.99	BISS Basic	-33.34%
4	770094957060	$17.40	$10.98	Personal Compu	-36.90%
5	753070352042	$10.86	$11.94	BISS Basic	9.05%
6	793518717156	$71.20	$30.00	Personal Compu	-57.87%
7	746591610364	$53.99	$57.54	BISS Basic	6.17%
8	719377200400	$14.99	$19.05	BISS Basic	21.31%
9	747100058301	$27.88	$31.31	PC Accessory	10.95%
10	689145744835	$23.91	$27.50	Book	13.05%
11	746591610470	$17.65	$10.17	BISS Basic	-42.38%
12	845156603565	$21.77	$19.99	BISS	-8.18%
13	758399181809	$22.95	$21.95	Personal Compu	-4.36%

Fake News! (Advanced)

These days, it's hard enough for the average social media user to determine when an article is made up with an intention to deceive. So is it possible to build a model that can discern whether a news piece is credible? That's the question a four-person team from the University of California at Berkeley attempted to answer with this project.

To develop a classifier that would be able to detect clickbait and propaganda articles, the foursome scraped data from news sources listed on OpenSources, preprocess articles for content-based classification using natural language processing, trained different machine learning models to classify the news articles, and created a web application to serve as the front end for their classifier.

Find out more and try it out here.

Audio Snowflake (Advanced)

The purpose of this Hackbright Academy project was to create a stunning visual representation of music as it played, capturing a number of components, such as tempo, duration, key, and mood. The web application Wendy created uses an embedded Spotify web player, an API to scrape detailed song

data, and trigonometry to move a series of colorful shapes around the screen. Audio Snowflake maps both quantitative and qualitative characteristics of songs to visual traits such as color, saturation, rotation speed, and the shapes of figures it generates.

Find out more here.

Bonus Data Sets for Data Science Projects

Here are a few more data sets to consider as you ponder data science project ideas:

- VoxCeleb: an audio-visual data set consisting of short clips of human speech, extracted from interviews uploaded to YouTube.
- Titanic: a classic data set appropriate for data science projects for beginners.
- Boston Housing Data: a fairly small data set based on U.S. Census Bureau data that's focused on a regression problem.
- Big Mart Sales: a retail industry data set that can be used to predict store sales.
- FiveThirtyEight: Nate Silver's publication shares the data and code behind some of its articles and graphics so admirers can create stories and visualizations of their own.

A history of the (Business) Internet: 1997

BackRub Search: university

university Search

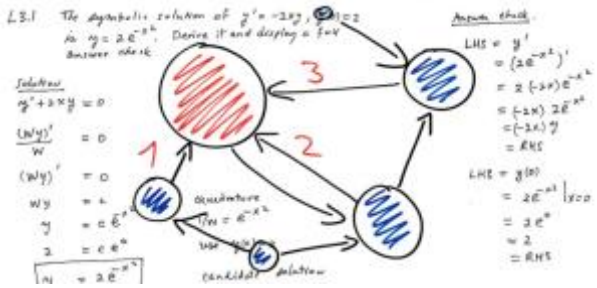
BackRub Query Results

BackRub's Highest Ranked Sites

University of Illinois at Urbana-Champaign
http://www.uiuc.edu/
694.687 8460 backlinks 12k - 10/25/96 - 11/1/96

Stanford University Homepage
http://www.stanford.edu/
609.303 8857 backlinks 4k - none - 11/1/96

Stanford University: Portfolio Collection
http://www.stanford.edu/home/administration/portfolio.html
167.919 34 backlinks



*4 - **PageRank: The web as a behavioral dataset***

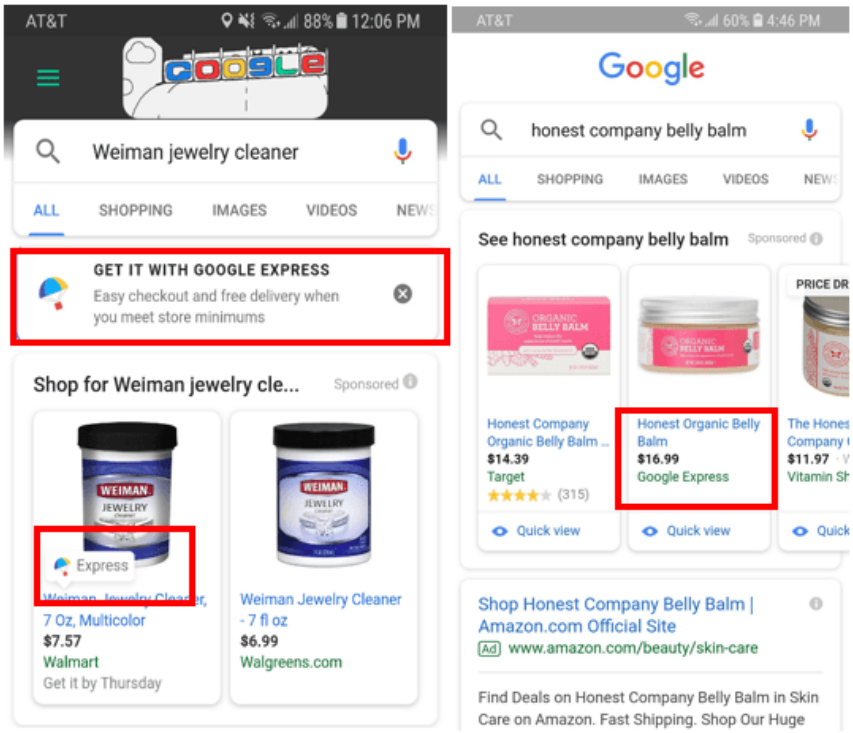
*5 - **Google server farms 2 million machines***

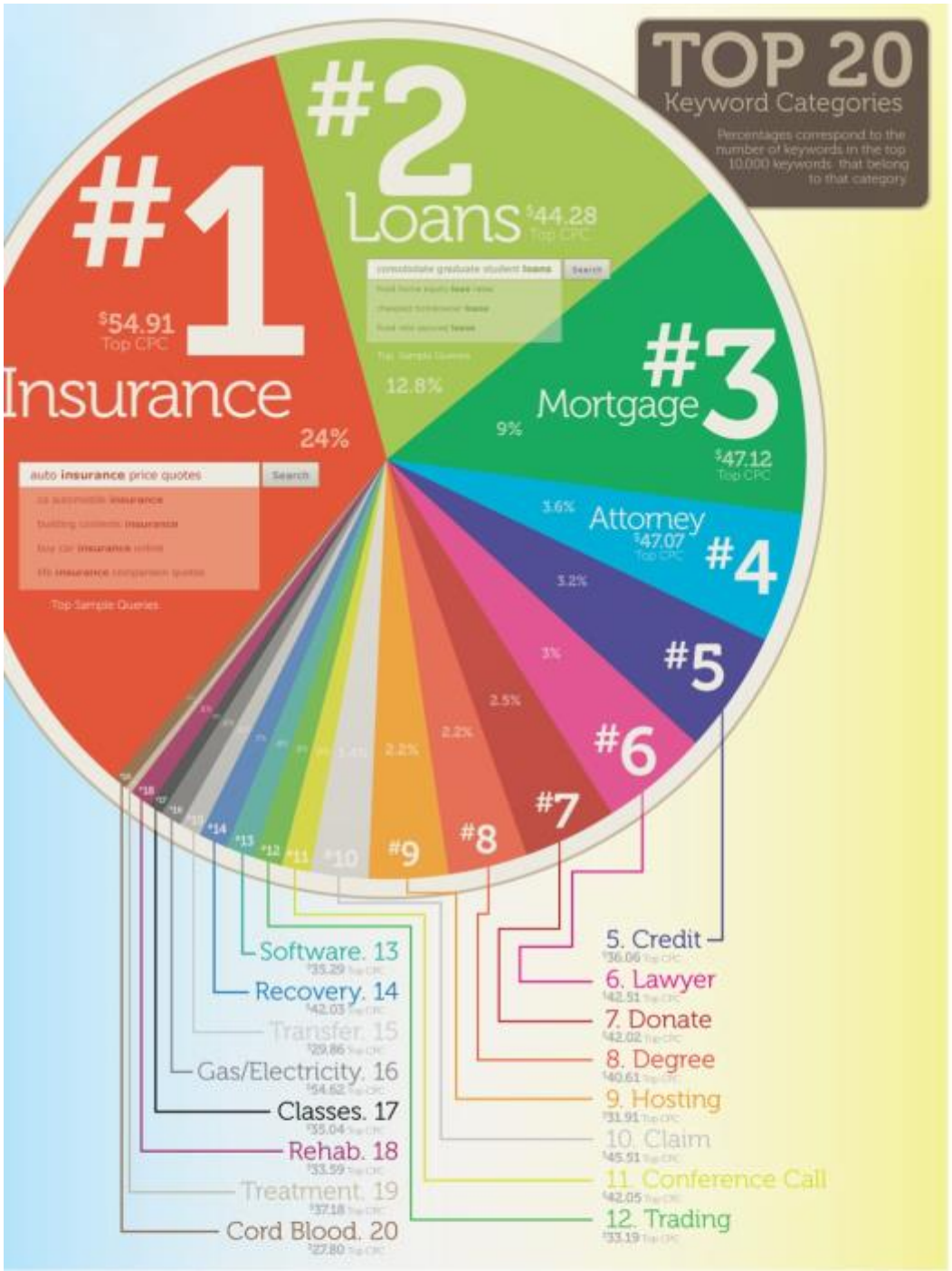
(DB size = 50 billion sites)

1998 – sponsored search

- Google revenue around $50 bn/year from marketing, 97% of the companies revenue.
- Sponsored search uses an auction – a pure competition for marketers trying to win access to consumers.

- In other words, a competition for **models** of consumers – their likelihood of responding to the ad – and of determining the right bid for the item.
- There are around 30 billion search requests a month. Perhaps a **trillion events** of history between search providers.





presented by WordStream

Data Makes Everything Clearer?

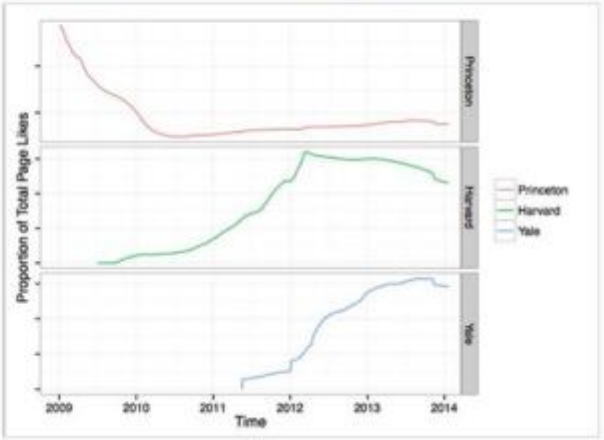
Epidemiological modeling of online social network dynamics

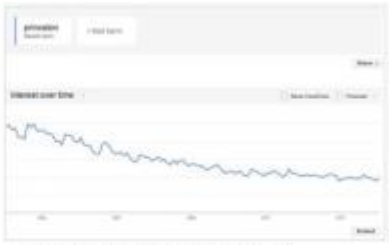
John Cannarella[1], Joshua A. Spechler[1,*]
1 Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA
* E-mail: Corresponding spechler@princeton.edu

Abstract

The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for "MySpace" as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for "Facebook," which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.

In keeping with the scientific principle "correlation equals causation," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely. Looking at page likes on Facebook, we find the following alarming trend:

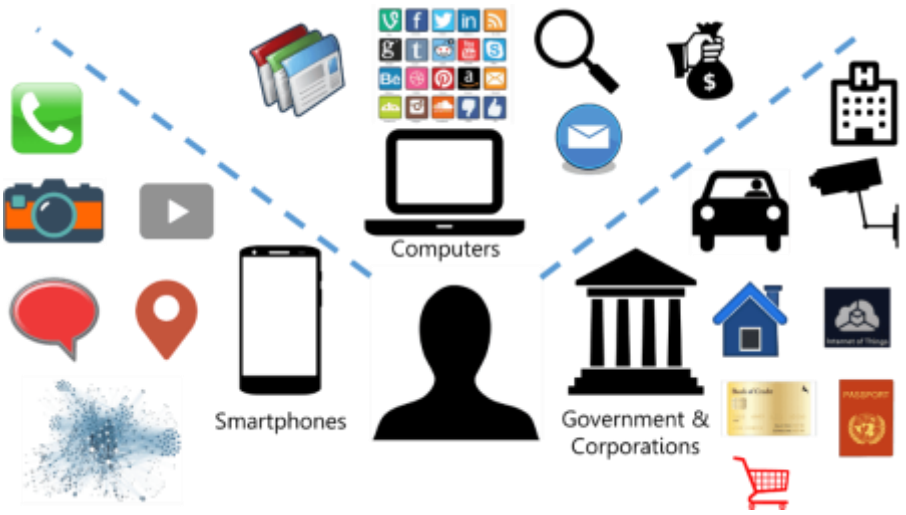




and based on Princeton search trends:

"This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,...

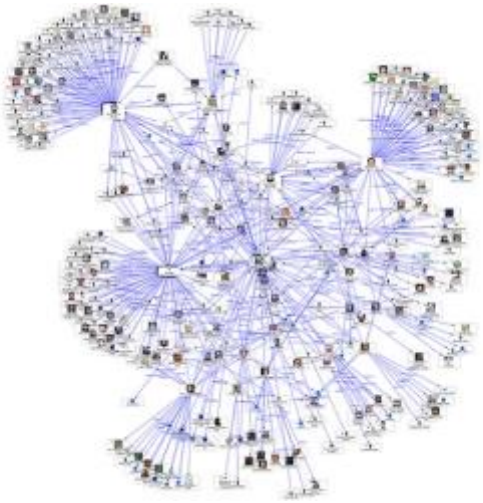
"Big Data" Sources

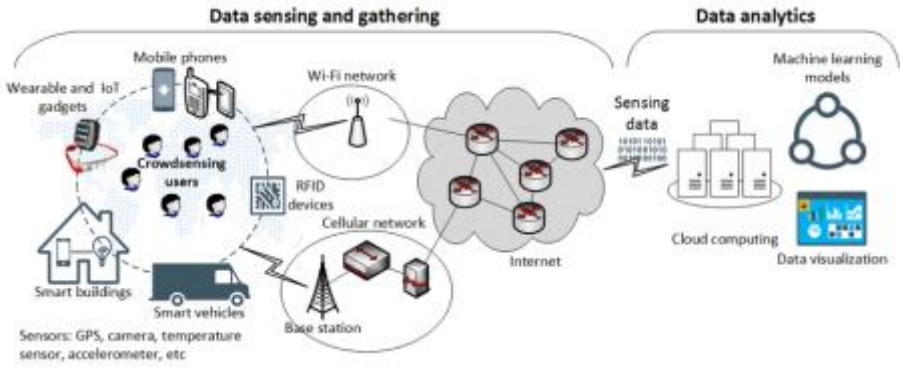


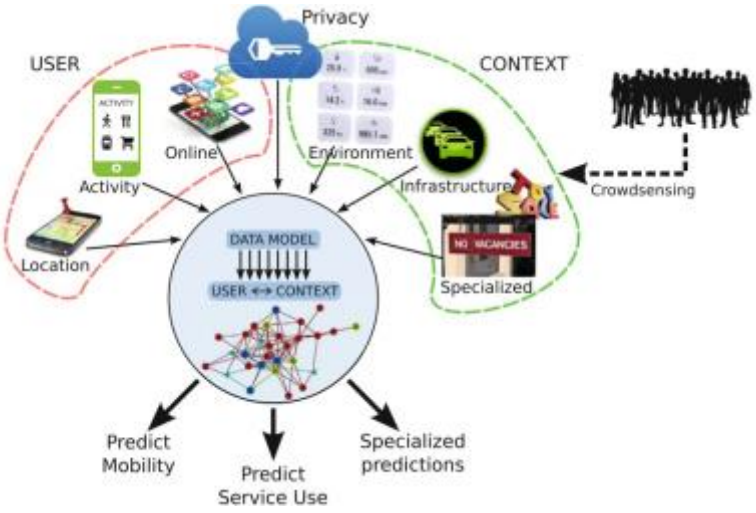
Graph Data

Lots of interesting data has a graph structure:

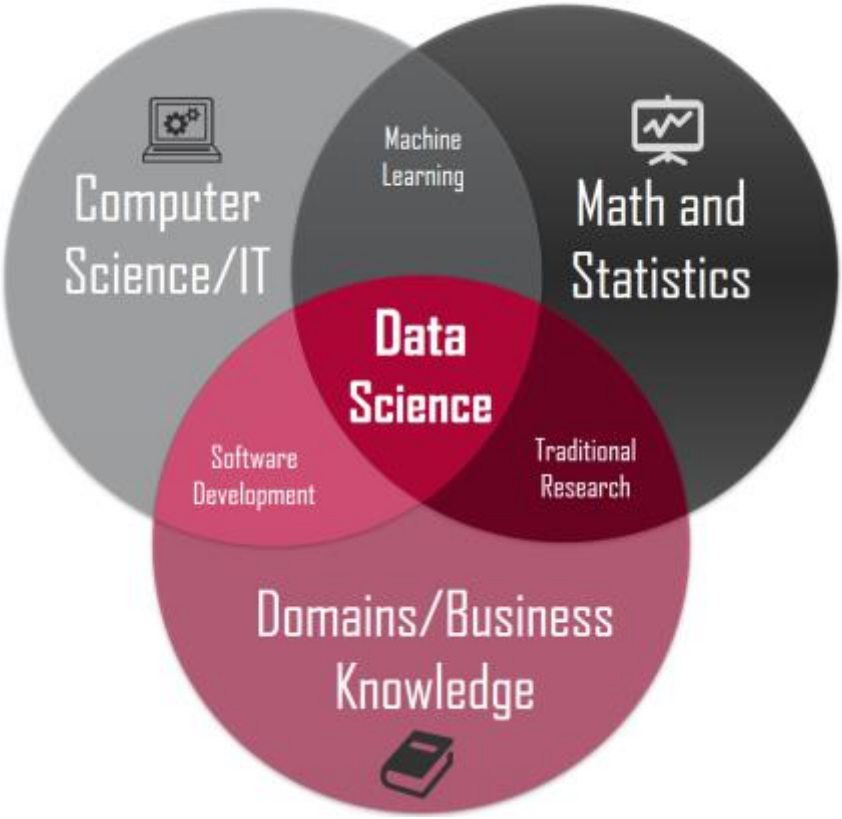
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- Some of these graphs can get quite large (e.g., Facebook* user graph)







What's Data Science?

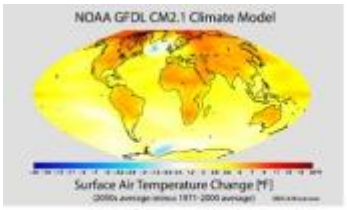


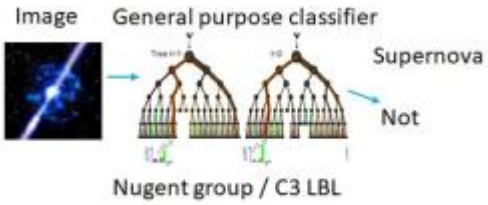
	Databases	Data Science
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,...

*6 - **Contrast: Databases***

Databases	Data Science
Querying the past	Querying the future

*7 - **Contrast: Databases***



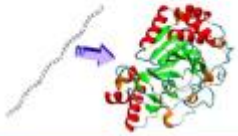


Scientific Modeling
Physics-based models
Problem-Structured
Mostly deterministic, precise
Run on Supercomputer or High-end Computing Cluster

Data-Driven Approach
General inference engine replaces model
Structure not related to problem
Statistical models handle true randomness, and **unmodeled complexity**.
Run on cheaper computer Clusters (EC2)

*8 - **Contrast: Scientific Computing***

CASP: A Worldwide, Biannual Protein Folding Contest



Quark	Raptor-X
Rich, Complex Energy Models	Data-intensive, general ML models
Faithful, Physical Simulation	Feature-based inference
	Conditional Neural Fields

Brain Mapping: Allen Institute, White House, Berkeley

Techniques (Massive ML)
Principal Component Analysis
Independent Component Analysis
Sparse Coding
Spatial (Image) Filtering

*9 - **Contrast: Computational Science***

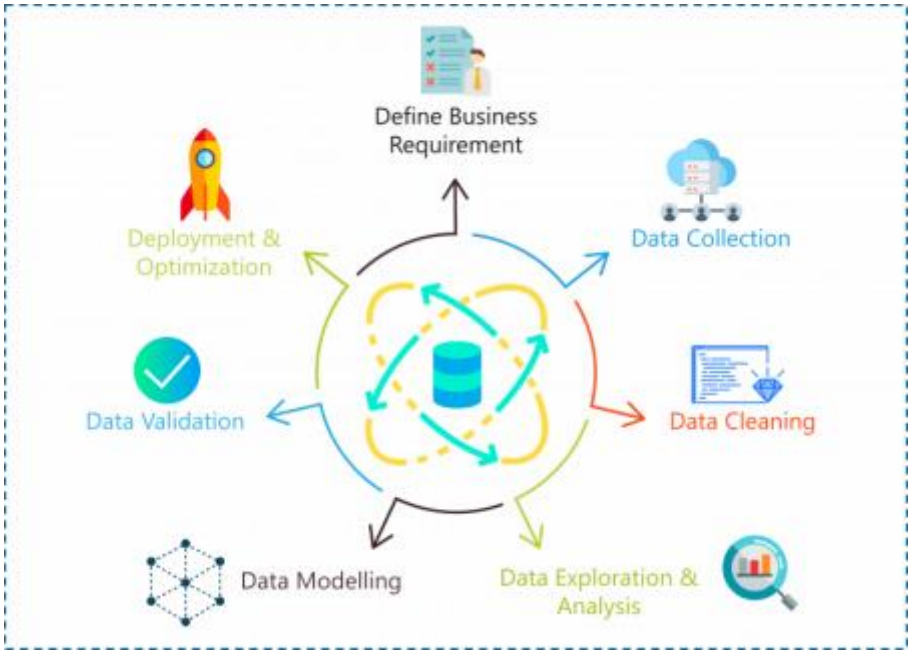
Machine Learning	Data Science
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/validate on a few, relatively clean, small datasets	Develop/use tools that can handle massive datasets
Publish a paper	Take action!

*10 - **Contrast: Machine Learning***

Data Science Project Life Cycle

A problem statement in Data Science can be solved by following the below steps:

1. Define Problem Statement/ Business Requirement
2. Data Collection
3. Data Cleaning
4. Data Exploration & Analysis
5. Data Modelling
6. Deployment & Optimization



Step 1: Define Problem Statement

Before you even begin a Data Science project, you must define the problem you're trying to solve. At this stage, you should be clear with the objectives of your project.

Step 2: Data Collection

Like the name suggests at this stage you must acquire all the data needed to solve the problem. Collecting data is not very easy because most of the time you won't find data sitting in a database, waiting for you. Instead, you'll have to go out, do some research and collect the data or scrape it from the internet.

Step 3: Data Cleaning

If you ask a Data Scientist what their least favorite process in Data Science is, they're most probably going to tell you that it is Data Cleaning. Data cleaning is the process of removing redundant, missing, duplicate and unnecessary data. This stage is considered to be one of the most time-consuming stages in Data Science. However, in order to prevent wrongful predictions, it is important to get rid of any inconsistencies in the data.

Step 4: Data Analysis and Exploration

Once you're done cleaning the data, it is time to get the inner Sherlock Holmes out. At this stage in a Data Science life-cycle, you must detect patterns and trends in the data. This is where you retrieve useful insights and study the behavior of the data. At the end of this stage, you must start to form hypotheses about your data and the problem you are tackling.

Step 5: Data Modelling

This stage is all about building a model that best solves your problem. A model can be a Machine Learning Algorithm that is trained and tested using the data. This stage always begins with a process called Data Splicing, where you split your entire data set into two proportions. One for training the model (training data set) and the other for testing the efficiency of the model (testing data set).

This is followed by building the model by using the training data set and finally evaluating the model by using the test data set.

Step 6: Optimization and Deployment:

This is the last stage of the Data Science life-cycle. At this stage, you must try to improve the efficiency of the data model, so that it can make more accurate predictions. The end goal is to deploy the model into production or production-like environment for final user acceptance. The users must validate the performance of the models and if there are any issues with the model then they must be fixed in this stage.

Now that you know how a problem can be solved using Data Science, let's get to the fun part. In the following section, I will be providing you with five high-level Data Science projects that can get you hired in the top IT firms.

Classification of 1994 Census Income Data

Problem Statement: To build a model that will predict if the income of any individual in the US is greater than or less than USD 50,000 based on the data available about that individual.

Data Set Description: This Census Income dataset was collected by Barry Becker in 1994 and given to the public site *http://archive.ics.uci.edu/ml/datasets/Census+Income*. This data set will help you understand how the income of a person varies depending on various factors such as the education background, occupation, marital status, geography, age, number of working hours/week, etc.

Here's a list of the independent or predictor variables used to predict whether an individual earns more than USD 50,000 or not:

- Age
- Work-class
- Final-weight
- Education
- Education-num (Number of years of education)
- Marital-status
- Occupation
- Relationship
- Race
- Sex
- Capital-gain
- Capital-loss

- Hours-per-week
- Native-country

The dependent variable is the "income-level" that represents the level of income. This is a categorical variable and thus it can only take two values:

1. <=50k
2. >=50k

https://www.edureka.co/blog/data-science-projects/

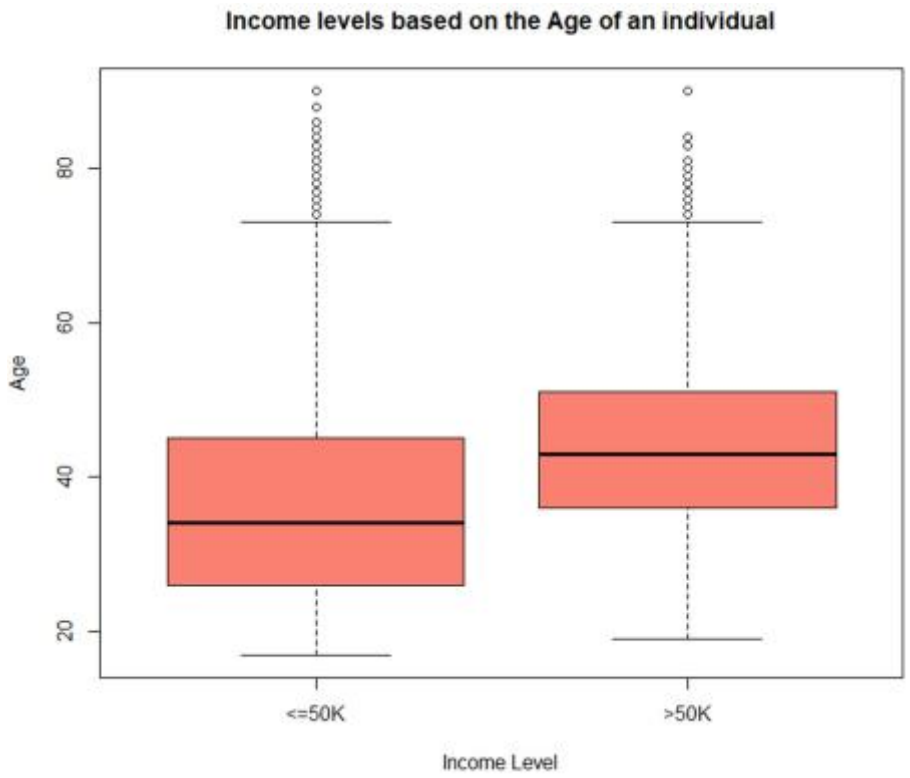
Step 1: Import the data

Step 2: Data Cleaning

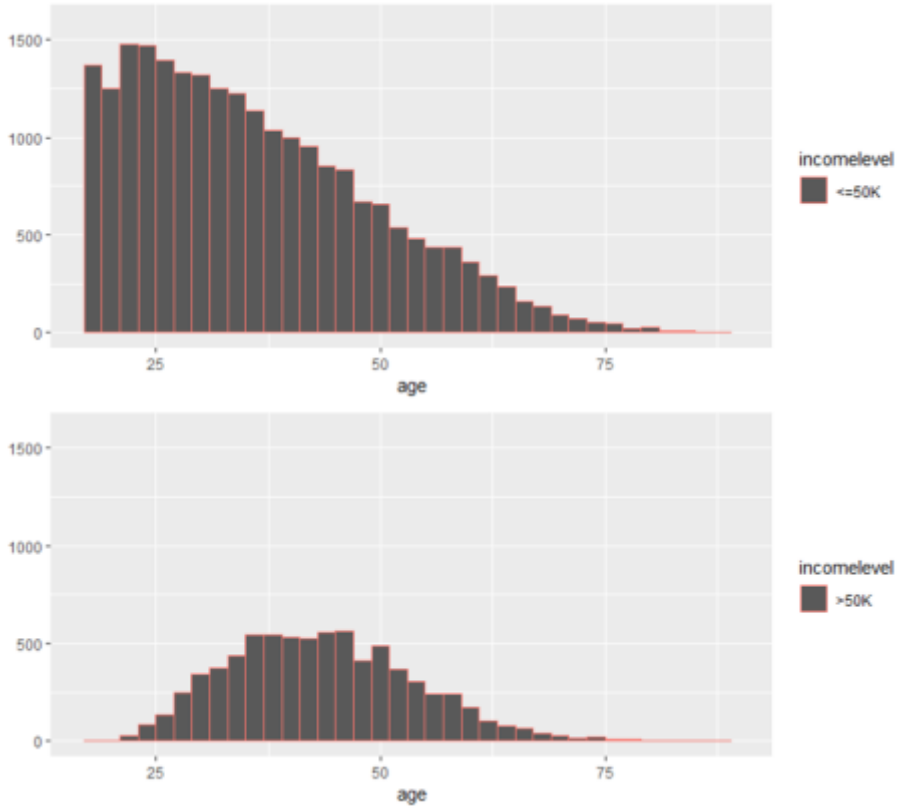
The data cleaning stage is considered to be one of the most time-consuming tasks in Data Science. This stage includes removing NA values, getting rid of redundant variables and any inconsistencies in the data.

Step 3: Data Exploration

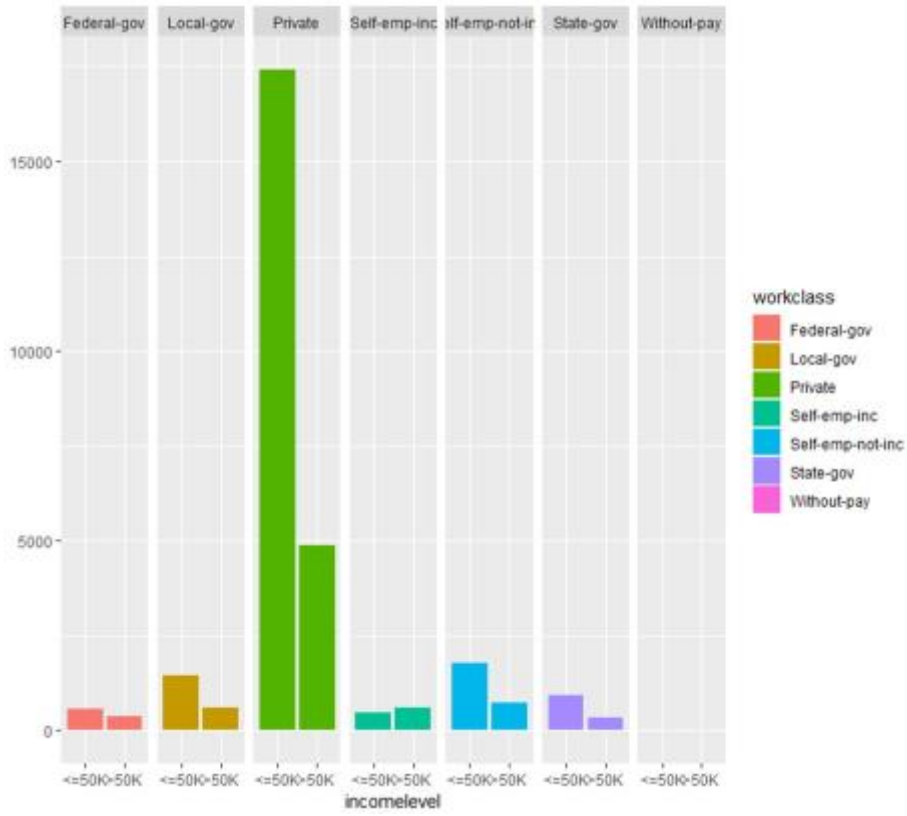
Data Exploration involves analyzing each feature variable to check if the variables are significant for building the model.



*11 - **Exploring the age variable***



*12 - **Histogram for age variable***



*13 - **Exploring work-class variable***

Step 4: Building A Model

So, after evaluating all our predictor variables, it is finally time to perform Predictive analytics. In this stage, we'll build a predictive model that will predict whether an individual earns above USD 50,000 or not based on the predictor variables that we evaluated in the previous section.

Step 5: Load and evaluate the test data set

Just like how we cleaned our training data set, our testing data must also be prepared in such a way that it does not have any null values or unnecessary predictor variables, only then can we use the test data to validate our model.

Step 6: Validate the model

The test data set is applied to the predictive model to validate the efficiency of the model.