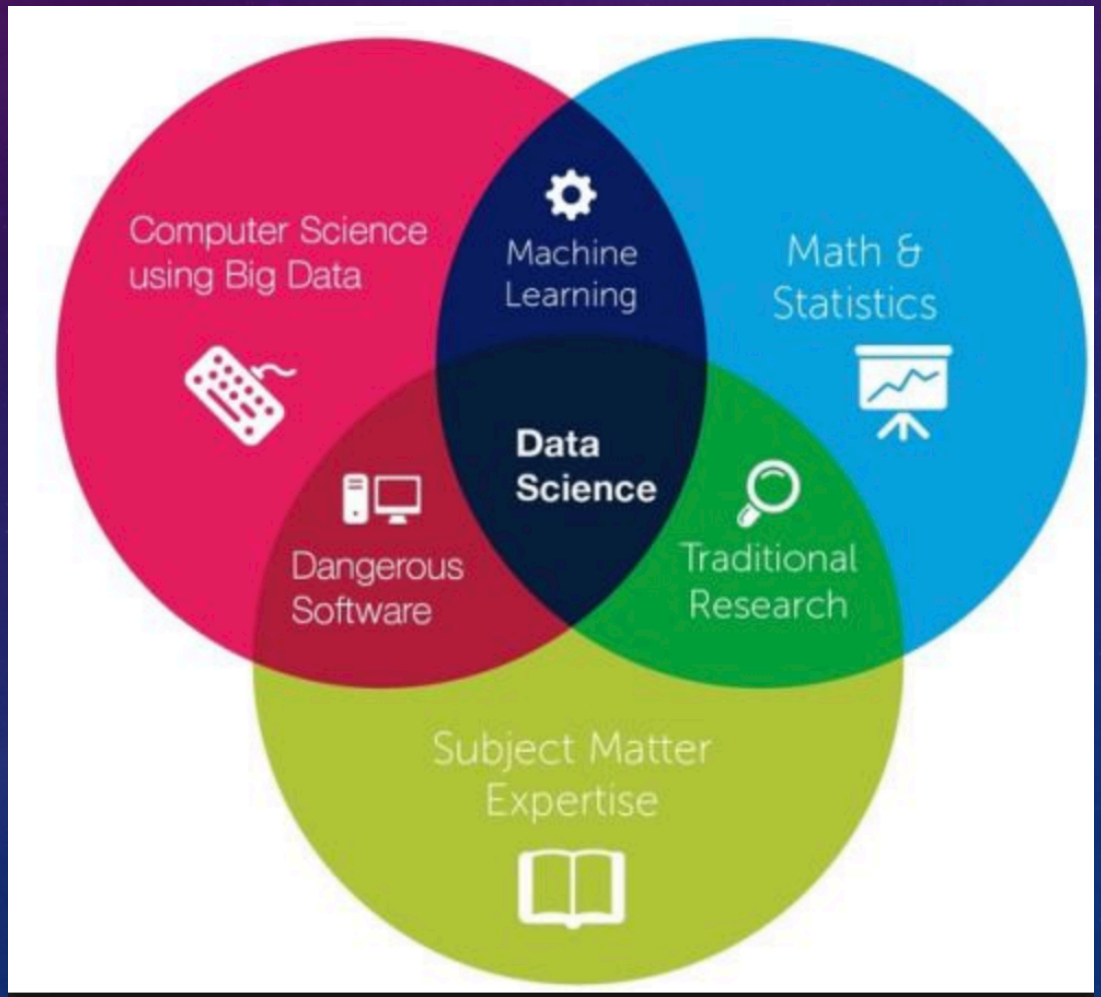
DATA SCIENCE

ADRIANO ZAMBOM

WHAT IS DATA SCIENCE?

- There are several ways to define it.
- Most people define data science as a collection of methods (statistical analysis), processes and algorithms (computer science) that are used to extract knowledge/ information about data to to help in decision making.
- The term Data Science has appeared in different areas long ago, but only in the past decade it has gained strength and become relevant. Why?

WHAT IS DATA SCIENCE?



WHAT IS DATA SCIENCE?

- A famous statistician, John Tukey, said:
- “The best thing about being a statistician is that you get to play in everyone's backyard.”
- This quote is actually very much fit for data scientists.

WHAT IS DATA SCIENCE?

- A data scientist has to adapt to the area he works in. For example:

- Engineer may ask these questions about their data:
  - What can we do to optimize this process based on the current results?
  - How can we use past data to minimize the probability of error?

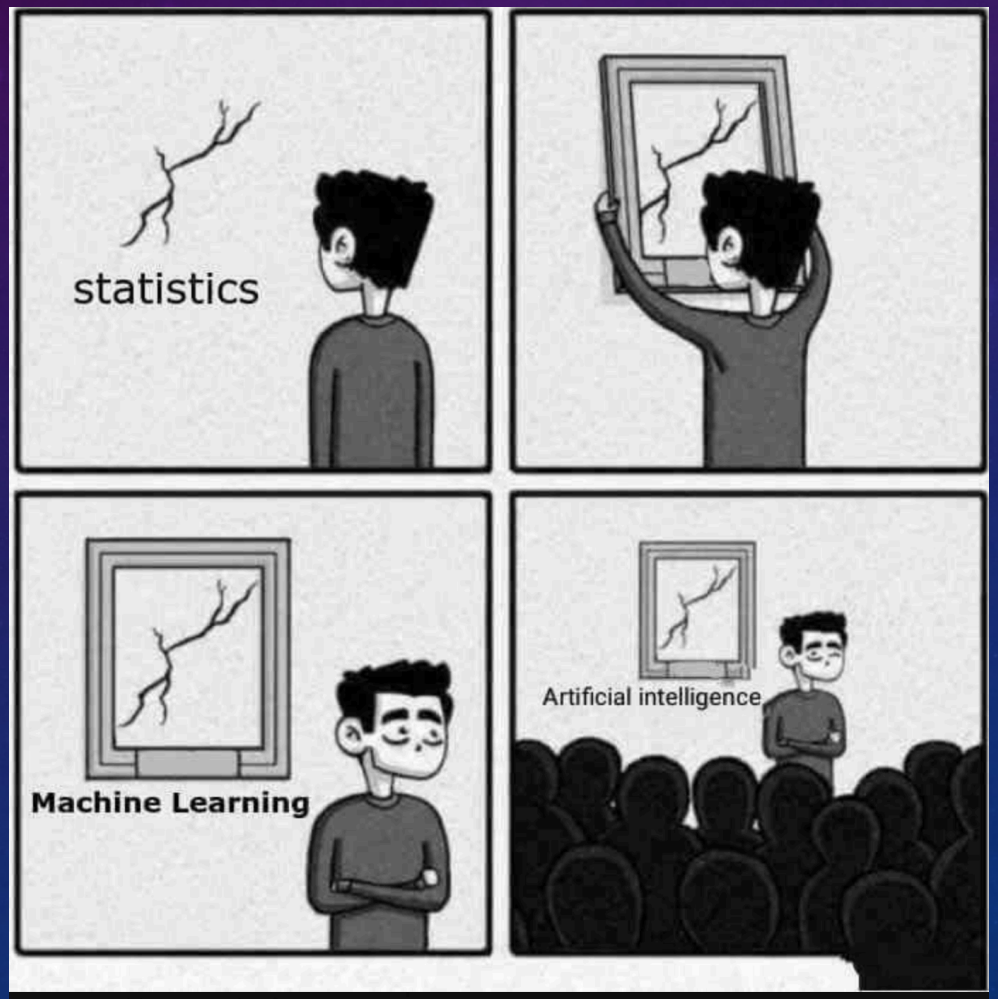
- Biologist may want to know
  - What genes should I look at when trying to predict a disease?
  - Which type of fertilizer should I use to obtain the maximum growth of a specific type of plant?

- HR questions about their data:
  - What should we do to meet or exceed the organization's hiring and retention goals for next year?
  - What measures should we look for when hiring?

WHAT IS DATA SCIENCE?

- General guidelines for a data science program:
- https://www.stat.berkeley.edu/~nolan/Papers/Data.Science.Guidelines.16.9.25.pdf

WHAT IS DATA SCIENCE?



MODERN WORLD GENERATES TONS OF DATA

DATA IS EVERYWHERE

- Internet
- Hospitals
- Industry
- Pharmaceutical
- Autonomous Vehicles
- Stock Market
- Advertising
- Biology/ Microarray
- Much more

EX: AUTONOMOUS VEHICLES

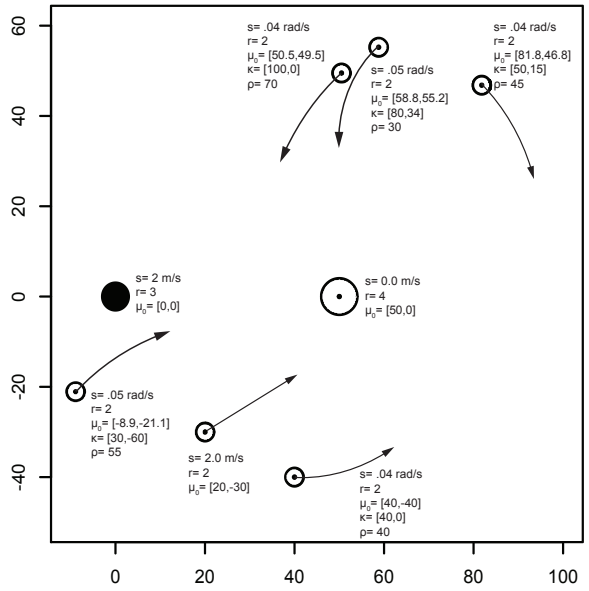
- Darpa Grand Challenge – Department of Defense – 2004 – In the desert
- A video
- But The Darpa Grand Challenge moved to the city – Urban Challenge

SIMPLIFIED EXAMPLE OF AUTONOMOUS VEHICLE DATA

Robot Path Planning

Robot Path Planning

- Plan a trajectory for an autonomous vehicle avoiding obstacles and minimizing the travel time



SIMPLIFIED EXAMPLE OF AUTONOMOUS VEHICLE DATA

- Data may look like this:

time	Obstacle Car1	Obstacle Car2	Obstacle Car3	Obstacle Car4
1	(10.2, 28.7)	(-3.5, 7.2)	(39.4, 21.8)	(5.3, 0)
2	(10.2, 28.8)	(-3.8, 7.2)	(39.4, 21.8)	(5.1, 0)
3	(10.2, 29.9)	(-3.9, 7.2)	(39.4, 21.8)	(4.7, 0)
4	(10.2, 30.0)	(-4.5, 7.2)	(39.4, 21.8)	(4.5, 0)
5	(10.2, 30.1)	(-5.5, 7.2)	(39.4, 21.8)	(4.2, 0)
6	(10.2, 30.2)	(-6.8, 7.2)	(39.4, 21.8)	(3.9, 0)
7	(10.2, 30.3)	(-7.9, 7.2)	(39.4, 21.8)	(3.3, 0)
8	(10.2, 30.4)	(-9.1, 7.2)	(39.4, 21.8)	(2.9, 0)
9	(10.2, 30.5)	(-11.3, 7.2)	(39.4, 21.8)	(2.2, 0)
…	…	…	…	…

SIMPLIFIED EXAMPLE OF AUTONOMOUS VEHICLE DATA

- How to predict the next position of the obstacles?

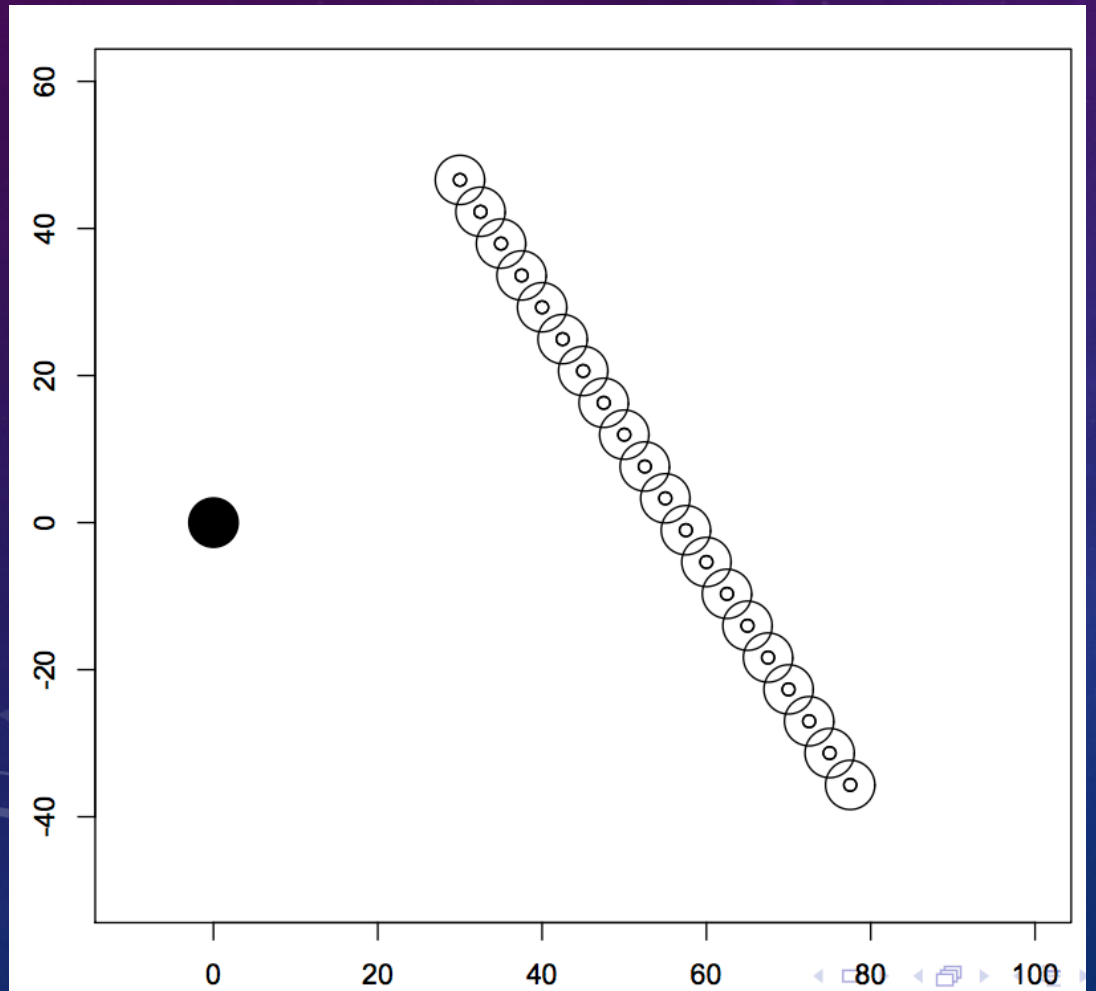
time	Obstacle Car1	Obstacle Car2	Obstacle Car3	Obstacle Car4
1	(10.2, 28.7)	(-3.5, 7.2)	(39.4, 21.8)	(5.3, 0)
2	(10.2, 28.8)	(-3.8, 7.2)	(39.4, 21.8)	(5.1, 0)
3	(10.2, 29.9)	(-3.9, 7.2)	(39.4, 21.8)	(4.7, 0)
4	(10.2, 30.0)	(-4.5, 7.2)	(39.4, 21.8)	(4.5, 0)
5	(10.2, 30.1)	(-5.5, 7.2)	(39.4, 21.8)	(4.2, 0)
6	(10.2, 30.2)	(-6.8, 7.2)	(39.4, 21.8)	(3.9, 0)
7	(10.2, 30.3)	(-7.9, 7.2)	(39.4, 21.8)	(3.3, 0)
8	(10.2, 30.4)	(-9.1, 7.2)	(39.4, 21.8)	(2.9, 0)
9	(10.2, 30.5)	(-11.3, 7.2)	(39.4, 21.8)	(2.2, 0)
…	…	…	…	…

DATA FROM LIDAR

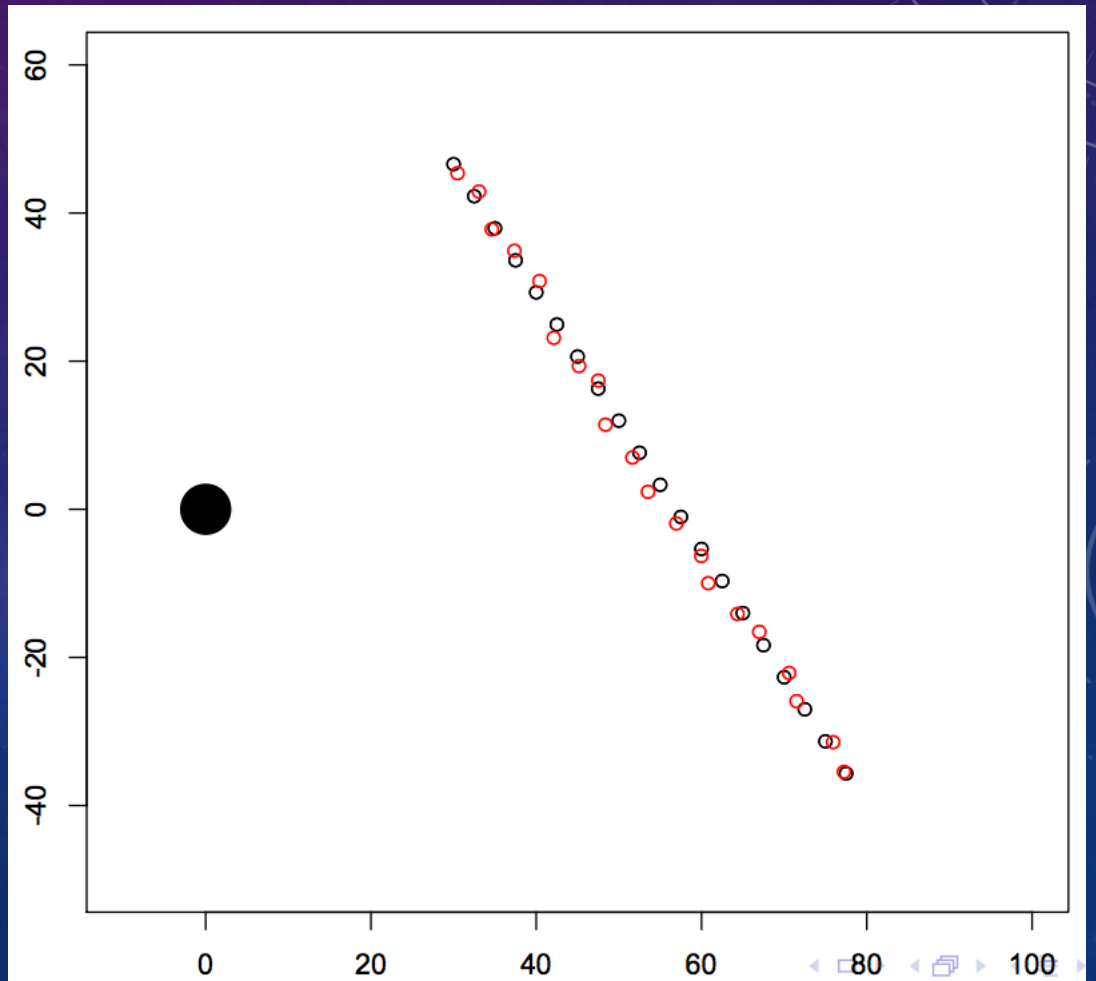
- Lidar is a detection system that works on the principle of radar, but uses light from a laser.
- In autonomous vehicles it is used like this.

WHAT IS THE DATA HAS SENSOR ERROR?

Suppose this is the true locations



But we observe this: with error



WHAT IS THE DATA HAS SENSOR ERROR?

- Check out this application in my webpage

MICROARRAY DATA

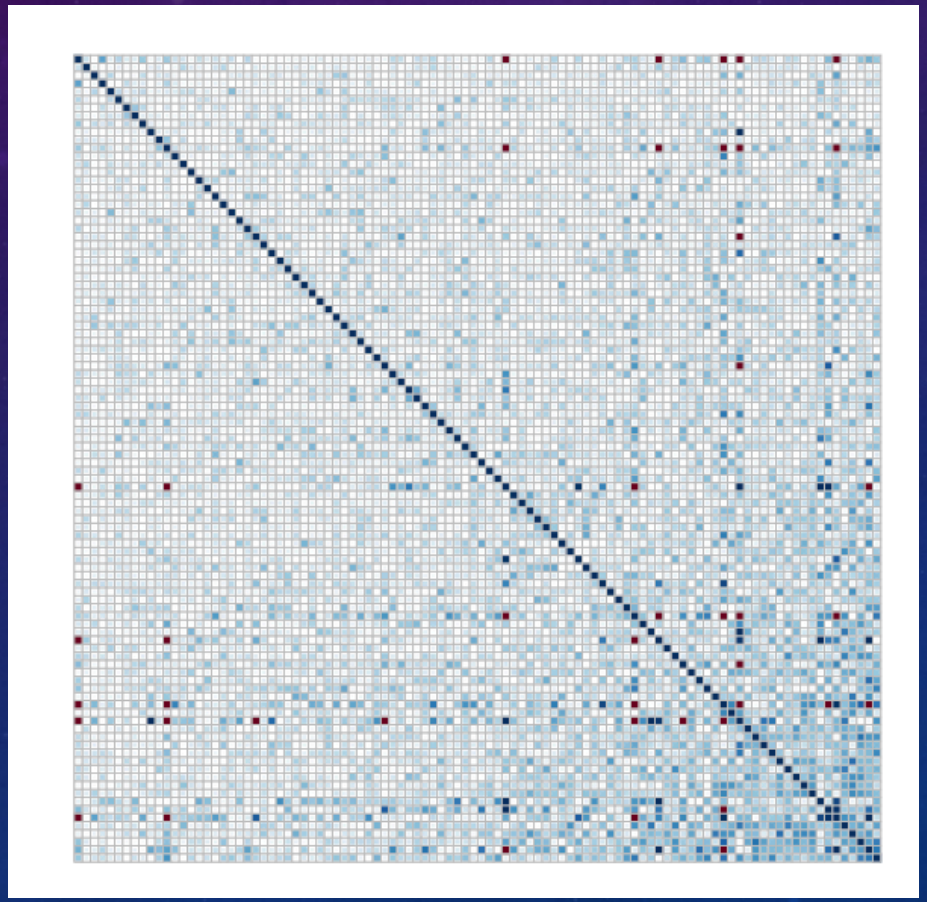
- Microarrays are very important in molecular biology
- They allow monitoring the expression levels of tens of thousands of genes simultaneously
- Applications include:
  - gene expression,
  - genome mapping,
  - SNP (single nucleotide polymorphisms) discrimination,
  - transcription factor activity,
  - toxicity,
  - pathogen identification and many others

MICROARRAY DATA EXAMPLE – MISSING DATA

- Gene expression study on prostate cancer - Tomlins et al. (2007)
- 104 observations of cDNA micro-array data across 20,000 gene locations
- Of these 20,000 genes, 18,106 contain at least 1 missing value and 2,570 variables are missing more than half of their observations
- Genes Previously associated with Prostate Cancer: BRCA1, BRCA2 (Pritchard et al. (2016); Castro and Eeles (2012)), HOXB13 (Ewing et al. (2012); Mills (2014); Beebe-Dimmer et al. (2015); Pilie et al. (2016)), and STAT3 (Abdulghani et al. (2008); Pencik et al. (2015))

MICROARRAY DATA EXAMPLE – MISSING DATA

Correlation plot for some of the Genes



MICROARRAY DATA EXAMPLE – MISSING DATA

Data may look like this:

Person	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	…
1	23.4	356	924	-	36	45.6	345.7	…
2	45.5	56	834	-	64	25.8	234.1	…
3	12.2	-	845	456	66	82.8	833.4	…
4	63.1	325	683	477	64	-	-	…
5	19.9	54	-	966	23	28.34	126.7	…
…	…	…	…	…	…	…	…	…

MICROARRAY DATA

- What is the role of a data scientist here?

- Hypothesis testing:
  - Compare arrays
  - Are differences significant?
  - What are your criteria for statistical significance?

- Clustering: Can we find patterns in the data? Groups of genes?

- Classification: Do the expression of genes predict well a disease?

MICROARRAY DATA

- Two main repositories:

- Gene expression omnibus (GEO) at NCBI

- ArrayExpress at the European Bioinformatics Institute (EBI)

FACEBOOK ADS

- **How does Facebook decide which ads to show?**

- Information you share on Facebook (example: posts or comments you make) and your activity on Facebook (such as liking a Page or a post, clicking on ads you see).
- Other information about you from your Facebook account (example: your age, your gender, your location, the devices you use to access Facebook).
- Information advertisers and our marketing partners share with us that they already have, like your email address.
- Your activity on websites and apps off of Facebook.
- Extracted from: https://www.facebook.com/help/562973647153813/

FACEBOOK ADS

- A simples dataset of activity may look like:

Date	Time	Action	Target	Object Type	Object Info
6/10/2018	8:37pm	View Pic	Peter	Pic	Nature
6/11/2018	8:10pm	Watch Vid	Maria	Video	Music
6/11/2018	8:24pm	Liked	Anthony	Comment	Key words: Democrat, vote
6/11/2018	8:35pm	Liked	Peter	Comment	World Cup
6/12/2018	8:01pm	View Pic	John	Pic	Nature
6/13/2018	8:56pm	Loved	Peter	Pic	-
6/13/2018	8:59pm	Comment	Barbara	Post	Key words: Democrat, campaign
6/13/2018	9:17pm	Post	All	Pic	Family (face recognition)
…	…	…	…	…	…

FACEBOOK ADS

- But another dataset with your personal info exists:
  - Location
  - Age
  - Gender
  - Ethnicity
  - Nationalty
  - …

FACEBOOK ADS

- Match person’s data to advertisement requirements/requests/target
- Classification: Which personality types are most likely to buy specific types of products
- Optimize revenue
- History of number of clicks − with info on links to the clicked objects

DATING WEBSITES AND APPS

- Sites such as
  - Match.com
  - eHarmony
  - OkCupid
  - Tinder
  - Note how you can register using your Facebook profile
- Use data science to match profiles based on their data
- *Compatibility matching models* – identify potential matches based on a client's core compatibility: common personality/psychology characteristics - user profile. Example: age, distance, religion, ethnicity, income, or education, etc.
- *Affinity matching models* – predict the probability of interaction and good communication between two people based on their profile.

DATING WEBSITES AND APPS

- Examples of anonymous personality data can be found here
- Can we match users?

PREDICTING BODYFAT EXAMPLE

- From website: http://lib.stat.cmu.edu/datasets/bodyfat

- Percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men
- Use Multiple regression
- Accurate measurement of body fat is inconvenient/costly and it is desirable to have easy methods of estimating body fat that are not inconvenient/costly.

PREDICTING BODYFAT EXAMPLE

- D = Body Density (gm/cm^3)
- A = proportion of lean body tissue
- B = proportion of fat tissue (A+B=1)
- a = density of lean body tissue (gm/cm^3)
- b = density of fat tissue (gm/cm^3)

- D = 1/[(A/a) + (B/b)]

- solving for B we find

- B = (1/D)*[ab/(a-b)] - [b/(a-b)]

Siri's equation: Percentage of Body Fat = 495/D - 450.

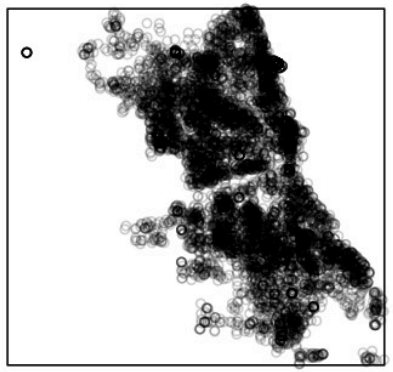
PREDICTING BODYFAT EXAMPLE

- But our data is composed of:

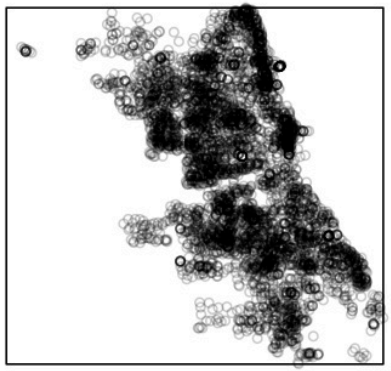
- Density determined from underwater weighing
- Percent body fat from Siri's (1956) equation
- Age (years)
- Weight (lbs)
- Height (inches)
- Neck circumference (cm) Chest circumference (cm) Abdomen circumference (cm) Hip circumference (cm) Thigh circumference (cm) Knee circumference (cm) Ankle circumference (cm) Biceps (extended) circumference (cm) Forearm circumference (cm) Wrist circumference (cm)

CRIME IN CHICAGO

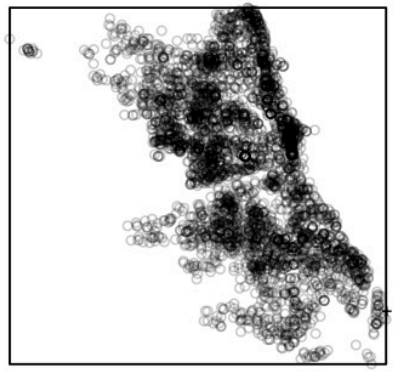
Sex Offenses 2003-2007



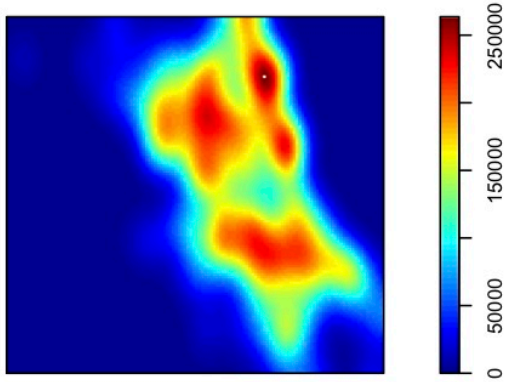
Sex Offenses 2008-2012



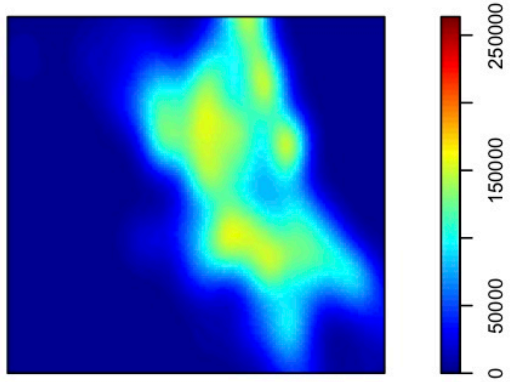
Sex Offenses 2013-2017



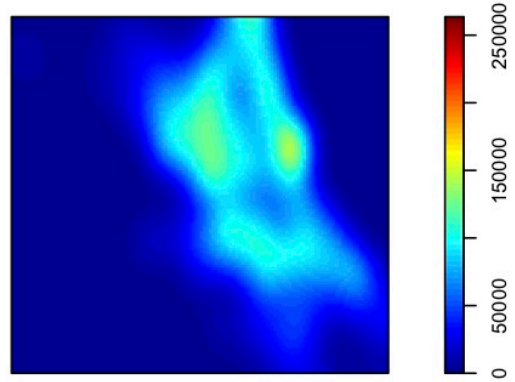
Intensity of Sex Offenses 2003-2007



Intensity of Sex Offenses 2008-2012



Intensity of Sex Offenses 2013-2017



CRIME IN CHICAGO

- Given the past years of observed data, can you predict how many crimes will occur in Chicago?
- What data would you need?
- Where can we find this kind of data? Here for Chicago


BIG
DATA
ALSO
TOOLS
DISK
TARGET
EXAMPLES
APPLIED
SENSOR
SHARED
DEFINITION
CURRENT
MAY
MOVING
RECONSIDER
WITHIN
ZETTABYTES
PRACTITIONERS
CAPTURE
THOUGHT
INTERNET
BUSINESS
DISTRIBUTED
SETS
GENOMICS
DESCRIBING
RADIO-FREQUENCY
COMPLEXITY
MANAGEMENT
TERABYTES
CASE
ELAPSED
PETABYTES
INCLUDE
ABILITY
SYSTEMS
TOLERABLE
SIZE
STORAGE
MPP
SAN
QUALITIES
SINCE
PARALLEL
MASSIVELY
GROW
SEARCH
RECORDS
NETWORKS
DATABASES
HUNDREDS
DESKTOP
CAPACITY
BURIED
WORLD'S
NOW
BIOLOGICAL
PROCESSING
UBIQUITOUS
TYPES
SOLID
CURRENTLY
GARTNER
WORKING
OPPORTUNITIES
AMOUNT
FORMS
DIFFICULTY
ANALYTICS
TENS
COMPLEX
INDEXING
CITATION
LARGER
CONTINUES
SOCIAL
USE
SET
CONNECTOMICS
ORGANIZATIONS
RELATIONAL
EVERY

WHAT IS BIG DATA?

- This term is usually used to describe datasets whose sizes are beyond the ability of current software/ hardware to process in a reasonable amount of time.

Common Data Storage Measurements

UNIT	VALUE
bit	1 bit
byte	8 bits
kilobyte	1,024 bytes
megabyte	1,024 kilobytes
gigabyte	1,024 megabytes
terabyte	1,024 gigabytes
petabyte	1,024 terabytes

WHAT IS BIG DATA?

- Sometimes big data is used to refer to predictive analytics, or user behavior analytics.
- In fact, the challenge is to use computer science tools to appropriately store, query, sample, transfer, and then use statistics to analyze the data.
- Correctly applying these concepts, one can obtain useful information to make decisions on his business
- Careful: Misuse of computer science and statistics may lead to biased conclusions!
  - Ex: Spurious correlations

STATISTICS AS THE KEY

- Data Science and Statistics