



# THE ART OF ANALYZING BIG DATA- THE DATA SCIENTIST'S TOOLBOX - LECTURE 1

DR. MICHAEL FIRE



# The Big Data Revolution

More than 6  
years of  
experience?



## אושרה התכנית הלאומית לקידום מדעי הנתונים

בתקצוב ות"ת של כ-150 מיליוני שקלים אושרה התכנית הלאומית לקידום מדעי הנתונים  
עיקרי התכנית הלאומית לקידום מדעי הנתונים במערכת ההשכלה הגבוהה:

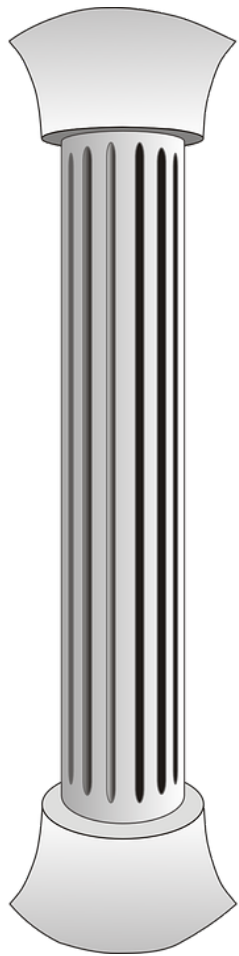


- תמיכת בהקמה/ביסוס של מרכזי-גג למחקר באוניברסיטאות
- הקמת מערך לסנכרון פעילות מרכזי המחקר המוסדיים עם התעשייה וגורמי ציבור
- תכנית למענקי מחקר במדעי הנתונים תוך התבססות על מאגרי הנתונים הנצברים בגופים ציבוריים אשר צפויה לתרום לחברה בישראל על ידי הבנת תהליכים הקורים בה, שיפור הקשר בין הממסד לאזרח וייעול השירות הניתן לו.
- מימון מלגות לדוקטורנטים ובת-דוקטורטים מצטיינים בתחום

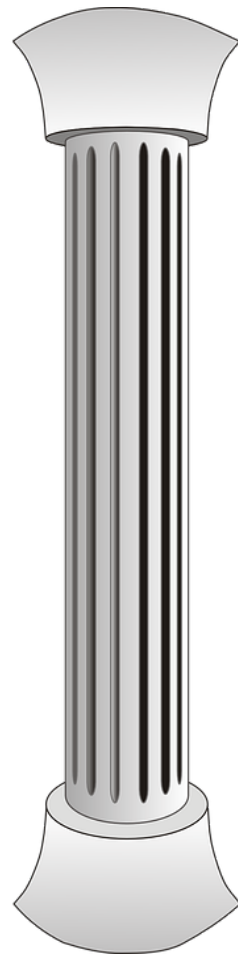


# Pillars of Science

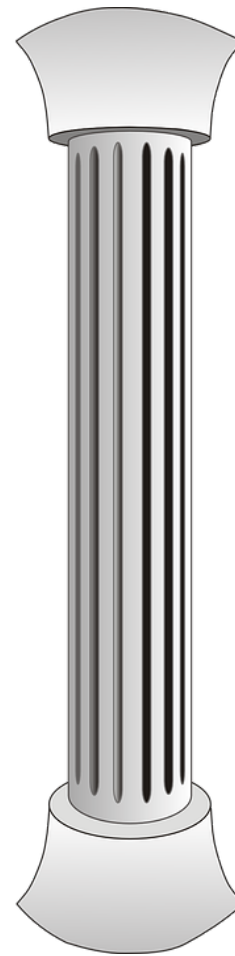
**Theory**



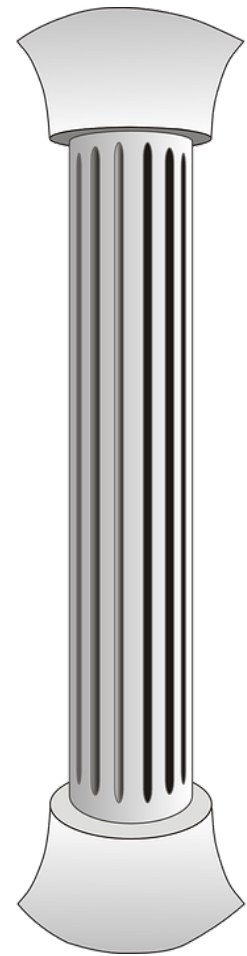
**Experimentation**



**Computational  
Science**



**Data-Intensive  
Science**



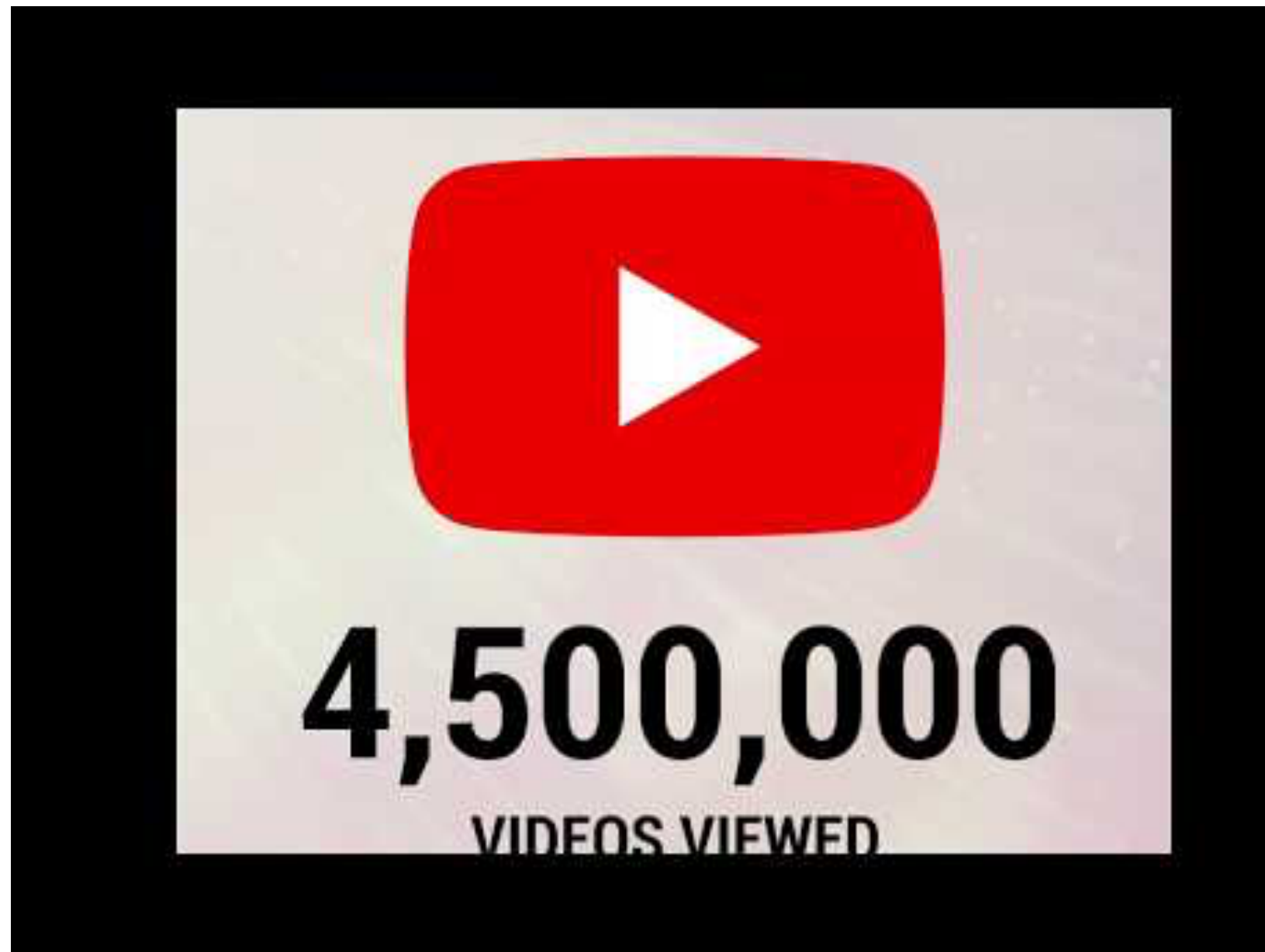


*“There was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing”*

Eric Schmidt, 2010



# The Data Tsunami



[A Day in Data Infographic](#)



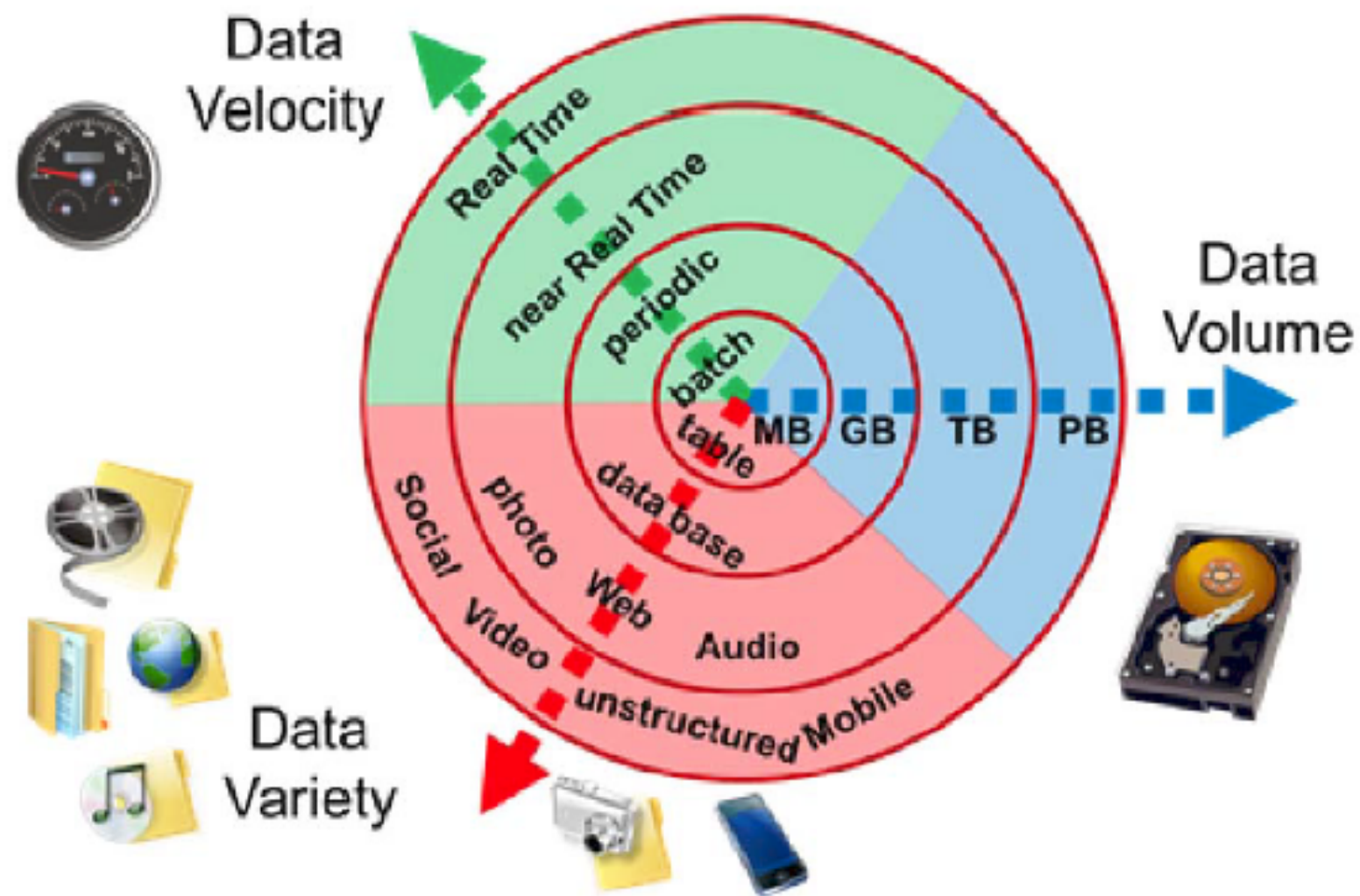
# What is Big Data?

- “Big data is a term used to refer to data sets that are too large or complex for traditional data-processing application software to adequately deal with. Data with many cases (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data **challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source**. Big data was originally associated with three key concepts: **volume**, **variety**, and **velocity**. Other concepts later attributed with big data are **veracity** (i.e., how much noise is in the data) and value.” (Wikipedia)
- “Big data is **high-volume, high-velocity and/or high-variety information** assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.” (Gartner)





# Data 3Vs (or 4vs)





# Example: Big Data at Netflix

## **Big Data at Netflix:**

- **167 million users**
- **160+ millions hours of video watched each day**
- **4000 different devices**
- **700+ billion events a day**
- **60 peta bytes of data**

## **Some of Netflix data related challenges:**

- **Building Big Data Infrastructure**
- **Personal recommendation of movies**
- **Creating Data Visualization Tools**
- **Improving Marketing Effectiveness**
- **Creating video previews**
- **Minimize the playback startup time**

More can be found on the [Netflix Technology Blog](#)





# Exciting Times

We are living in exciting times with a lot of new things to discover using new **datasets**, data analysis **tools**, new data **infrastructures**

*“The next Kinsey, I strongly suspect, will be a data scientist. The next Foucault will be a data scientist. The next Freud will be a data scientist. The next Marx will be a data scientist. The next Salk might very well be a data scientist”*

Seth Stephens Davidowitz, 2017



**OPEN  
DATASETS**

The logo consists of a white circle containing the text "Open Data" in a blue, sans-serif font. The circle is set against a rectangular background of blue binary code (0s and 1s).

Open  
Data



# How many stars will there be in the second Gaia data release?





# Diverse Datasets

The image is a collage of four overlapping screenshots from the Kaggle website, illustrating a variety of datasets. The top-most screenshot shows the 'UFO Sightings' dataset, which includes reports of unidentified flying objects from the last century, sourced from the National UFO Reporting Center (NUFORC). Below it, the 'European Soccer' dataset is visible, featuring 25k+ matches. To the left, the 'Poker' dataset is shown, with 721 poker games. The bottom-most screenshot displays the 'The ultimate' dataset, which includes a list of what you get: +25,000 matches, +10,000 players, 11 European countries, Seasons 2008-2017, and Players and Teams. The screenshots are arranged in a way that they overlap, creating a sense of depth and showcasing the diversity of data available on the platform.

**UFO Sightings**  
Reports of unidentified flying object reports in the last century  
National UFO Reporting Center (NUFORC) · updated 15 days ago (Version 2)  
Download (28 MB) New Notebook

**European Soccer**  
25k+ matches  
Hugo Math

**Poker**  
721 Pok  
Alb

**The ultimate**  
What you get:  
• +25,000 match  
• +10,000 players  
• 11 European Co  
• Seasons 2008  
• Players and Tea  
• Team line-up w  
• #: ID for each pokemon



# Notable Open Datasets

- [Kaggle](#) - over 28,000+ datasets
- [Microsoft Academic Graph](#) - over 231 million papers
- [data.gov](#) - U.S. Government's open data
- [pushshift.io](#) - full Reddit dataset
- [Common Crawl](#) - 8 years of web pages data
- [YouTube-8M Dataset](#) - a large-scale labeled video dataset that consists of millions of YouTube video
- [Data4Good.io](#) - over 1TB of compressed networks data

:-)





*"Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world"*

Atul Butte, 2012



# DATA SCIENCE TOOLS







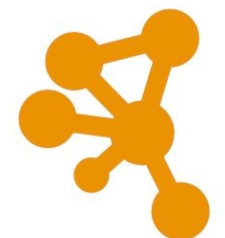
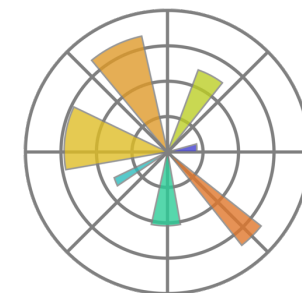
IP[y]: IPython  
Interactive Computing



Gephi



elasticsearch



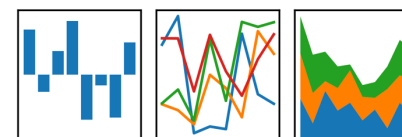
+ a b l e a u

Spark



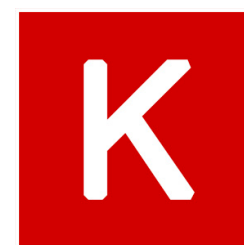
neo4j

pandas  
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

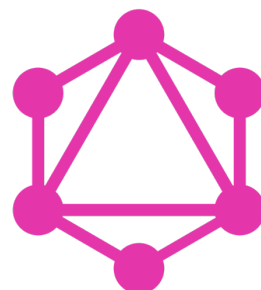


PyTorch

mongoDB



SQLite



WEKA  
The University  
of Waikato



# Wide Variety of Easy to Use Tools

image\_path = "/1280px-White Lion.jpg"  
Image(  
In [88]: def  
Out [88]:

predictions, probabilities  
predictions, probabilities  
(['lion', 'ice\_bear', 'ram',  
[53.90625596046448,

cls = tc.sound\_classifier.create(train\_sf, target='label', feature='audio')  
cls

/anaconda3/envs/venv/lib/python3.6/site-packages/coremltools/\_deps/\_init\_.py:118: DeprecationWarning: The 'warn' function is deprecated, use 'warning' instead  
% (tensorflow.\_\_version\_\_, TF\_MAX\_VERSION))  
WARNING:root:TensorFlow version 1.13.1 detected. Last version known to be fully compatible is 1.12.0 .

Creating a validation set from 5 percent of training data. This may take a while.  
You can set ``validation\_set=None`` to disable validation tracking.

Preprocessing audio data -  
Preprocessed 282 of 398 examples

Extracting deep features -  
Extracted 383 of 2464  
Extracted 771 of 2464  
Extracted 1162 of 2464  
Extracted 1554 of 2464  
Extracted 1944 of 2464  
Extracted 2334 of 2464  
Preparing validation set

cls.evaluate(test\_sf)

{'accuracy': 0.8125,  
'auc': 0.79908267733471,  
'precision': 0.8069767441860465,  
'recall': 0.5145502645502645,  
'f1\_score': 0.5175808720112518,  
'log\_loss': 0.531309188922887,  
'confusion\_matrix': Columns:  
target\_label str  
predicted\_label str  
count int

Rows: 5

Data:

target_label	predicted_label	count
murmur	murmur	4
extrastole	normal	5
murmur	normal	3
normal	normal	35
normal	murmur	1



# Using Data Science Tools

My Personal belief:

Using data science tools is similar to using electricity - we can start using most of the tools without knowing the details behind the underline algorithms





# CLOUD INFRASTRUCTURE





# Cloud Computing





# Increasing affordable Computational Power

x1.16xlarge

64

174.5

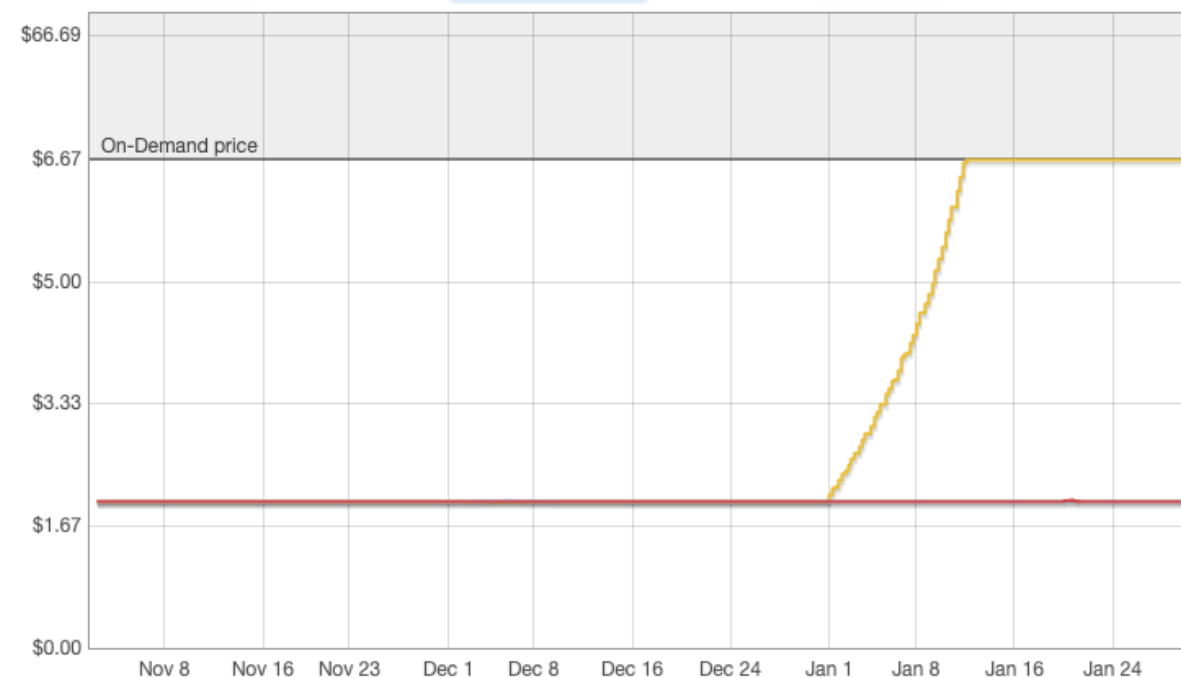
976 GiB

1 x 1920 SSD

\$6.669 per Hour

## Spot Instance Pricing History

Product: Linux/UNIX Instance type: x1.16xlarge Date range: 3 months



Date

11/4/2018

6:56:08 AM UTC+0200

On-Demand price

\$6.6690

Availability Zone

us-west-2a	\$2.0007
us-west-2b	\$2.0007
us-west-2c	\$2.0007

Calculations per second per constant dollar

1E+09  
1E+07  
1E+05  
1,000  
10  
0.1  
0.001  
1E-05  
1E-07  
1E-09

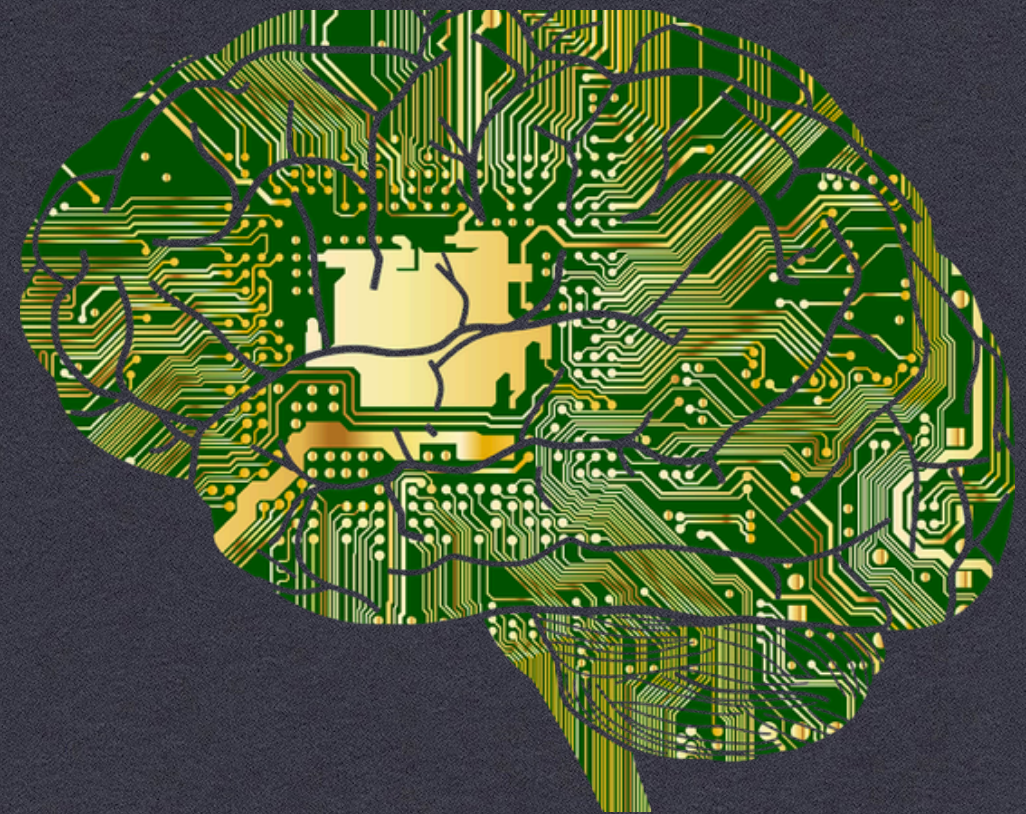
1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020

Year

Source: Ray Kurzweil, Steve Jurvetson



# DEEP LEARNING





# Deep Learning

- Deep learning is part of a broader family of machine learning methods based on artificial neural networks
- Deep learning architectures have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs
- They have produced results comparable to and in some cases surpassing human expert performance



# OUR ACADEMIC COURSE





# Course Goals

During this course, you will learn to:

- How to **collect** data
- How to **create** data from various sources
- How to **manipulate** data
- How to **handle with massive datasets**
- How to **identify patterns** in the data



# Course Goals

During this course, you will also learn to:

- Learn how to work with various data analytics tools
- Learn how to work with graphs
- Learn some practical text analytics
- Learn to visualize data

We will learn how to transfer **data to knowledge**



# Course Assignments

- Weekly **relatively small code** tasks to check you understand the material of each lesson (you get one in the end of today lesson)
- **Course Project** (in **pairs only**) - doing something cool with a real dataset
- **Test**



# WORKING WITH DATA





# *“Data Scientist: The Sexiest Job of the 21st Century”*

Thomas H. Davenport and D.J. Patil, 2012



# Some Things to Remember

*“If you torture the data long enough, it will confess”*

*Ronald H. Coase*



# The Bonferroni Principle

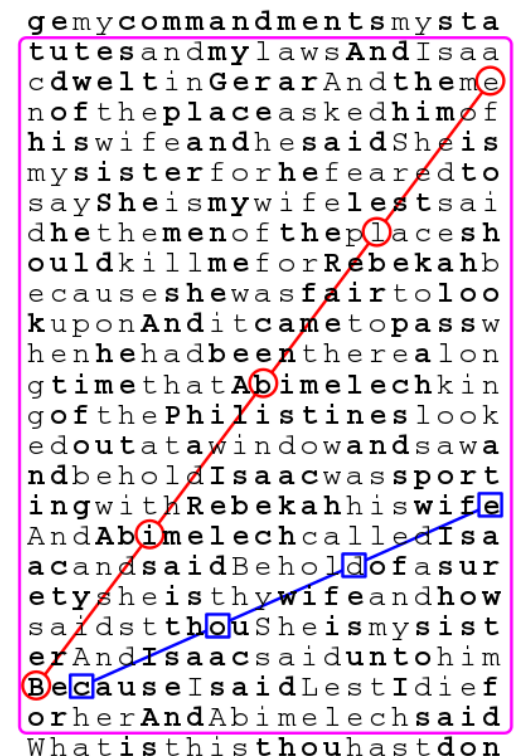
- In a completely random dataset still there are interesting events that may occur
- If you look hard enough you will find them
- In big datasets there are many “interesting” patterns that occur by chance.

For example, in a large geolocation dataset, if we want to identify people that are friends according to repeating joint locations over time. We will probably match pairs of people that were in the same places by chance.



# The Look-Elsewhere Effect

- An apparently statistically significant observation may have actually a space to be searched
- “The Bible Code” - with enough options something significant will be discovered

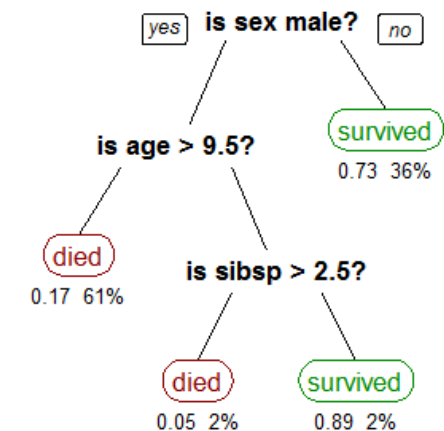
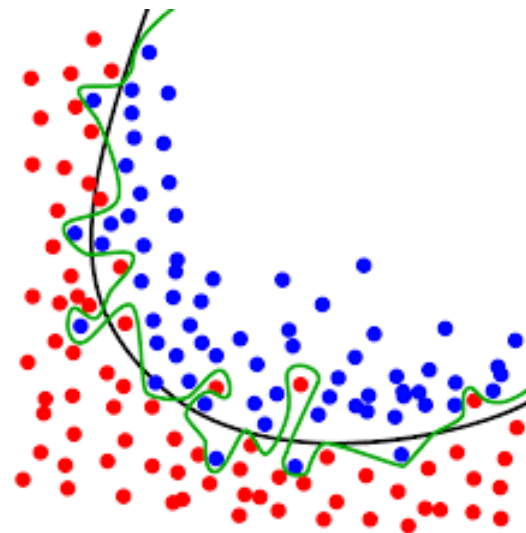
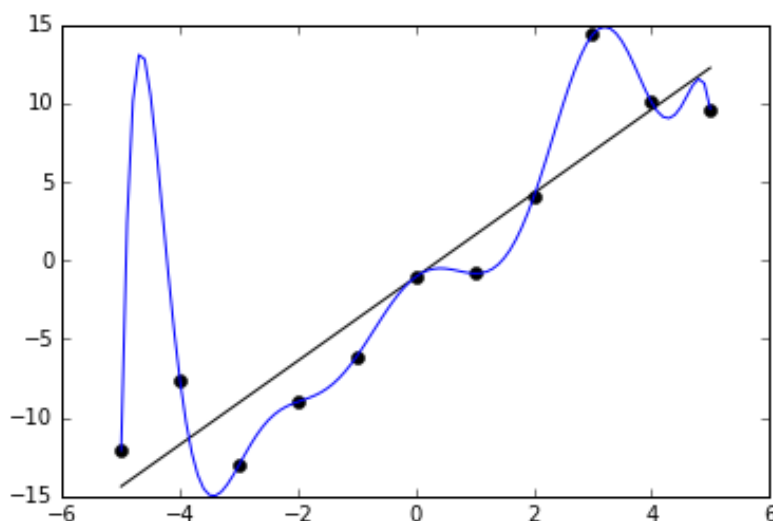


gemycommandmentsmysta  
tutesandmylawsAndIsaa  
cdweltinGerarAndthene  
noftheplaceaskedhimof  
hiswifeandhesaidSheis  
mysisterforhefearedto  
saySheismywifelestsai  
dhethemenofthelacesh  
ouldkillmeforRebekahb  
ecauseshewasfairtooo  
kuponAnditcametopassw  
henhehadbeentherealon  
gtime thatAbimelechkin  
gofthePhilistineslook  
edoutatawindowandsawa  
ndbeholdIsaacwassport  
ingwithRebekahhiswife  
AndAbimelechcalledisa  
acandsaidBeholdofasur  
ety sheisthywifeandhow  
saidstthouSheismysist  
erAndIsaacsaiduntohim  
BecauseIsaidLestIdief  
orherAndAbimelechsaid  
Whatisthisthouhastdon



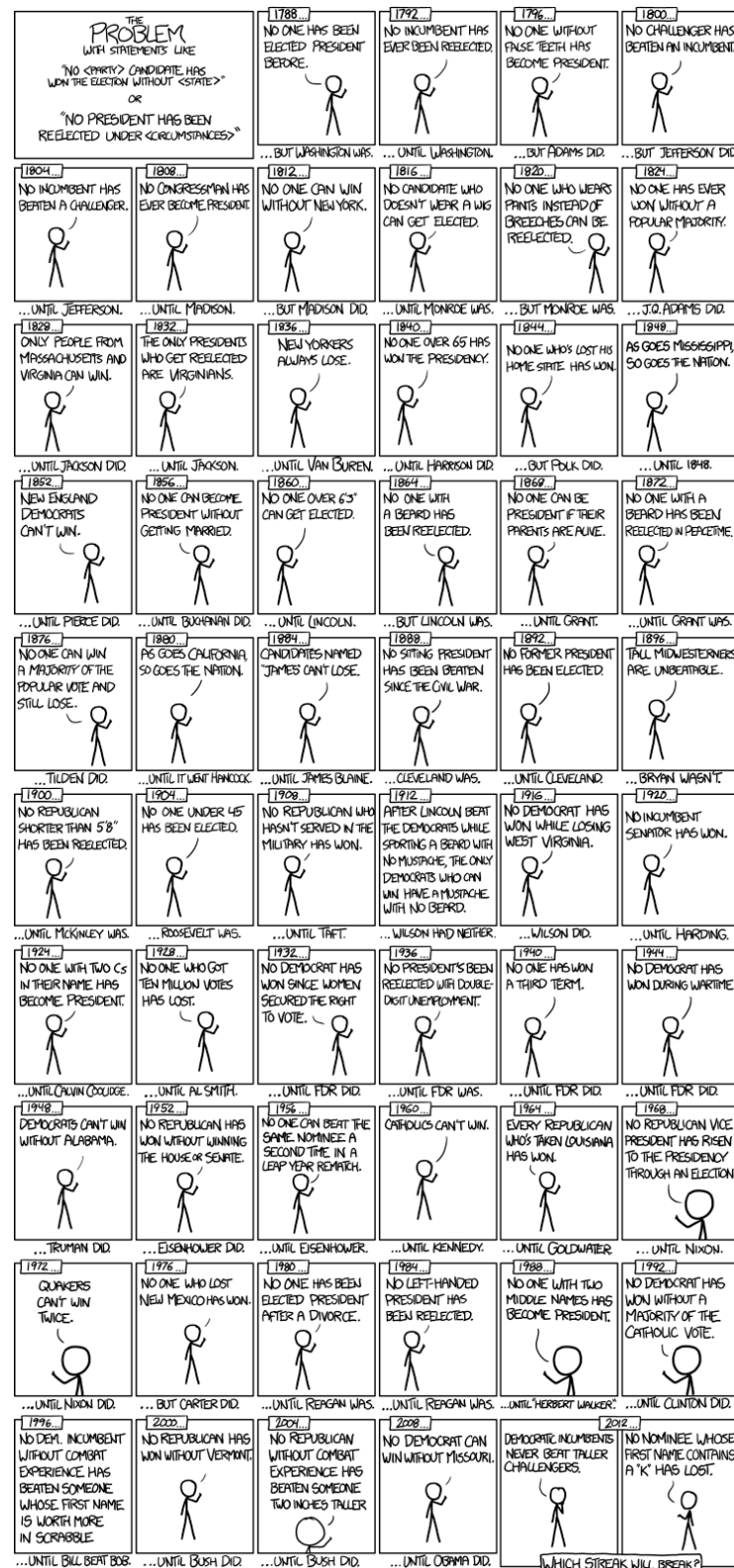
# Underfitting & Overfitting

- We use data and machine learning algorithms to create prediction models
- The goal of a good machine learning model is to generalize well from the training data
- **Underfitting** is when the model is **too simple**
- **Overfitting** is when the model is **too complex**
- **A rule of thumb** - if at first your model's performances is too good to be true on the first runs - you are probably overfitting

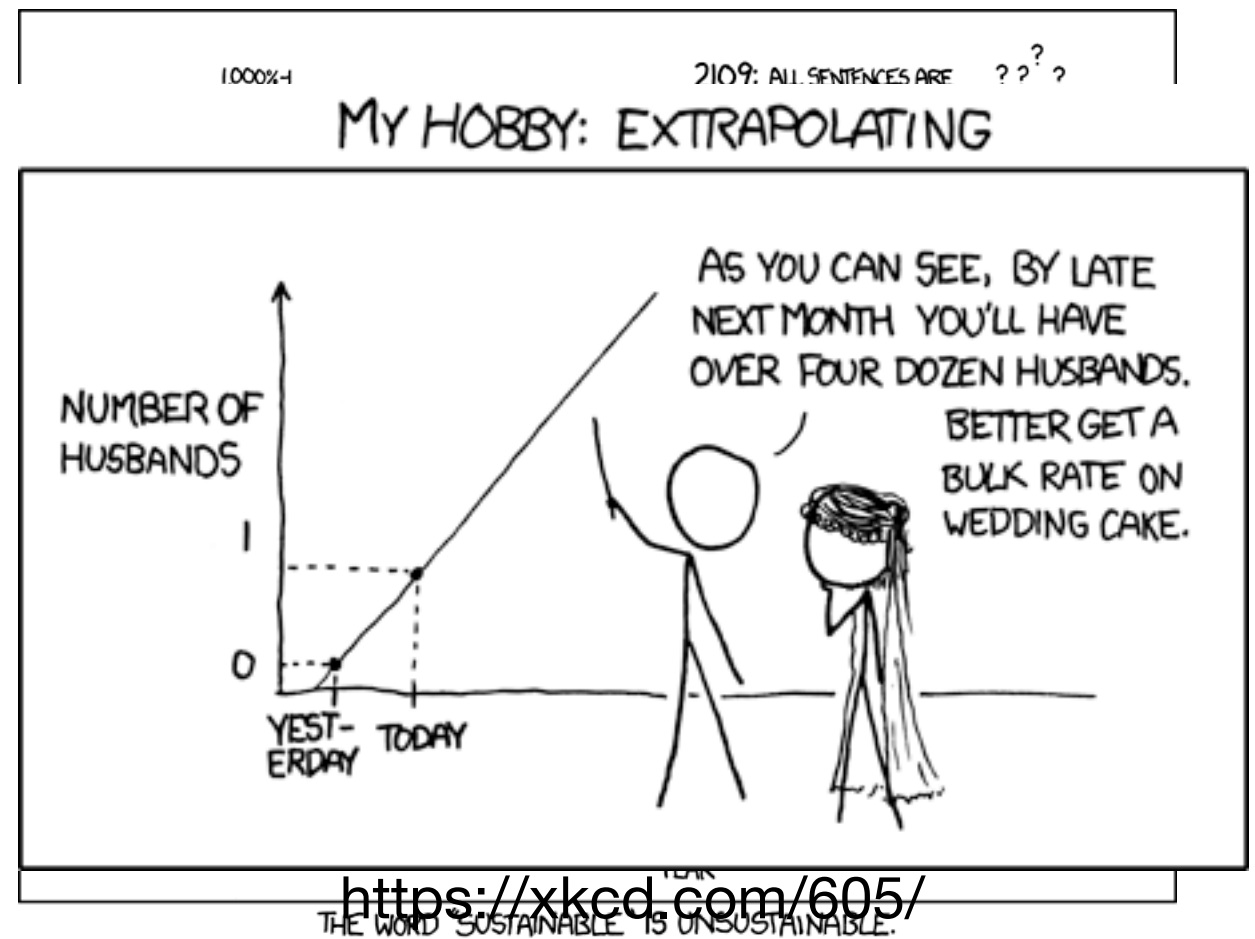




# Overfitting according to XKCD

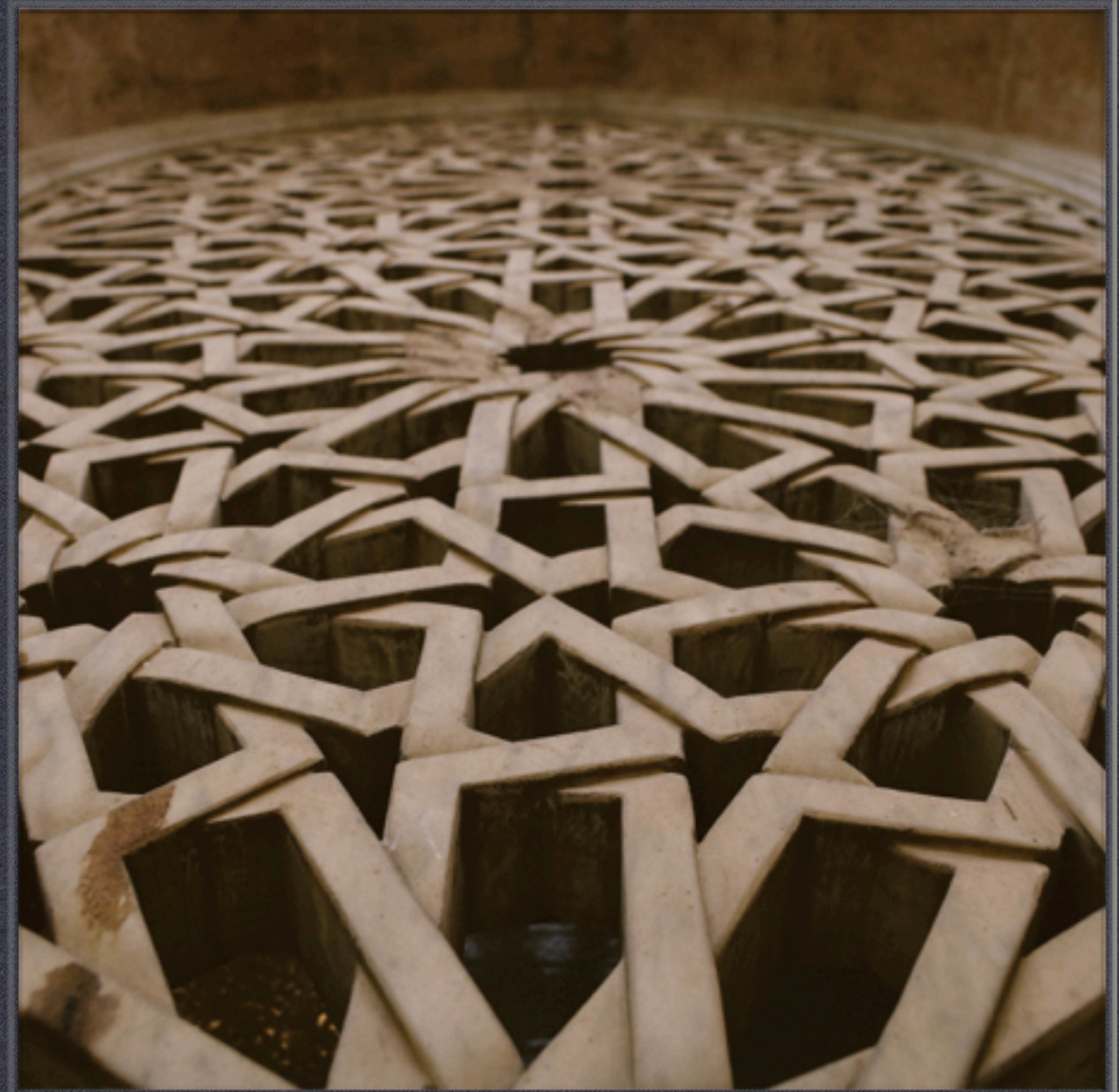


# Underfitting according to XKCD





# WORKING WITH STRUCTURED DATA





# Working with DBMS

Craig Kerstiens

About

Travel & Wine

My Recommendations

Top Content

Archive

## SQL: One of the Most Valuable Skills

Feb 12th, 2019

PostgreSQL, Postgres,

Like 258

Tweet

G+

I've learned a lot of skills over the course of my career, but no technical skill more useful than SQL. SQL stands out to me as the most valuable skill for a few reasons:

1. It is valuable across different roles and disciplines
2. Learning it once doesn't really require re-learning
3. You seem like a superhero. *You seem extra powerful when you know it because of the amount of people that aren't fluent*

Let me drill into each of these a bit further.

## SQL a tool you can use everywhere





# Data Science and Databases

## From my personal experience:

### When to use databases:

- Working with structured/tabular data
- Working with relatively small datasets (up to several million rows)
- Doing relatively simple analytics
- Needing to work with many subsets of the datasets

### When not to use databases:

- Working with unstructured data
- Working with data that contains dictionary/lists structures
- Working with relatively large datasets (several hundreds of millions of rows)
- Doing complex analytics



# SQL - A Very Quick Review

**Select <Col\_1>,<Col\_2>,....,<Col\_N>  
From <Table1>, <Table2>, .....,<Table\_N>  
Where <RowCondntion>  
Order by <Col\_i>**

**SELECT FirstName, LastName**

**FROM Users**

**WHERE firstName='John' and LastName like 'Sm%'**

**ORDER BY Age**



# Data Definition Language (DDL)

Used to **Create/Drop/Alter/Truncate** tables

```
CREATE TABLE "flavors_of_cacao" (  
  "Company" TEXT,  
  "SpecificBeanOriginorBarName" TEXT,  
  "REF" INTEGER,  
  "Review Date" INTEGER,  
  "Cocoa_Percent" TEXT,  
  "Company_Location" TEXT,  
  "Rating" REAL,  
  "Bean_Type" TEXT,  
  "BroadBean_Origin" TEXT  
);
```

```
UPDATE User  
SET Country = 'USA'  
WHERE Country = 'United States';
```

```
TRUNCATE Table Users;
```

```
ALTER Table User  
ADD LastPost varchar(255);
```

```
ALTER Table User  
Drop LastPost varchar(255);
```



# Data Manipulation Language (DML)

Used to manipulate data using **Select/Insert/Update/Delete**

**Select** u1.firstname, u2.firstname  
From Links l, Users u1, Users u2  
Where l.user1 = u1.userid, l.user2 = u2.userid

**Select** GroupNumber, **AVG**(JoinYear), **Max**(JoinYear)  
From Users  
**Group by** GroupNumber  
**Order by** **AVG**(JoinYear)

**INSERT INTO** Links (User1, User2)  
**VALUES** (5,4);

**UPDATE** Users  
**SET** GroupNumber = 3  
**WHERE** GroupNumber = 1;

**DELETE FROM** Users **WHERE** UserId=4;

Links

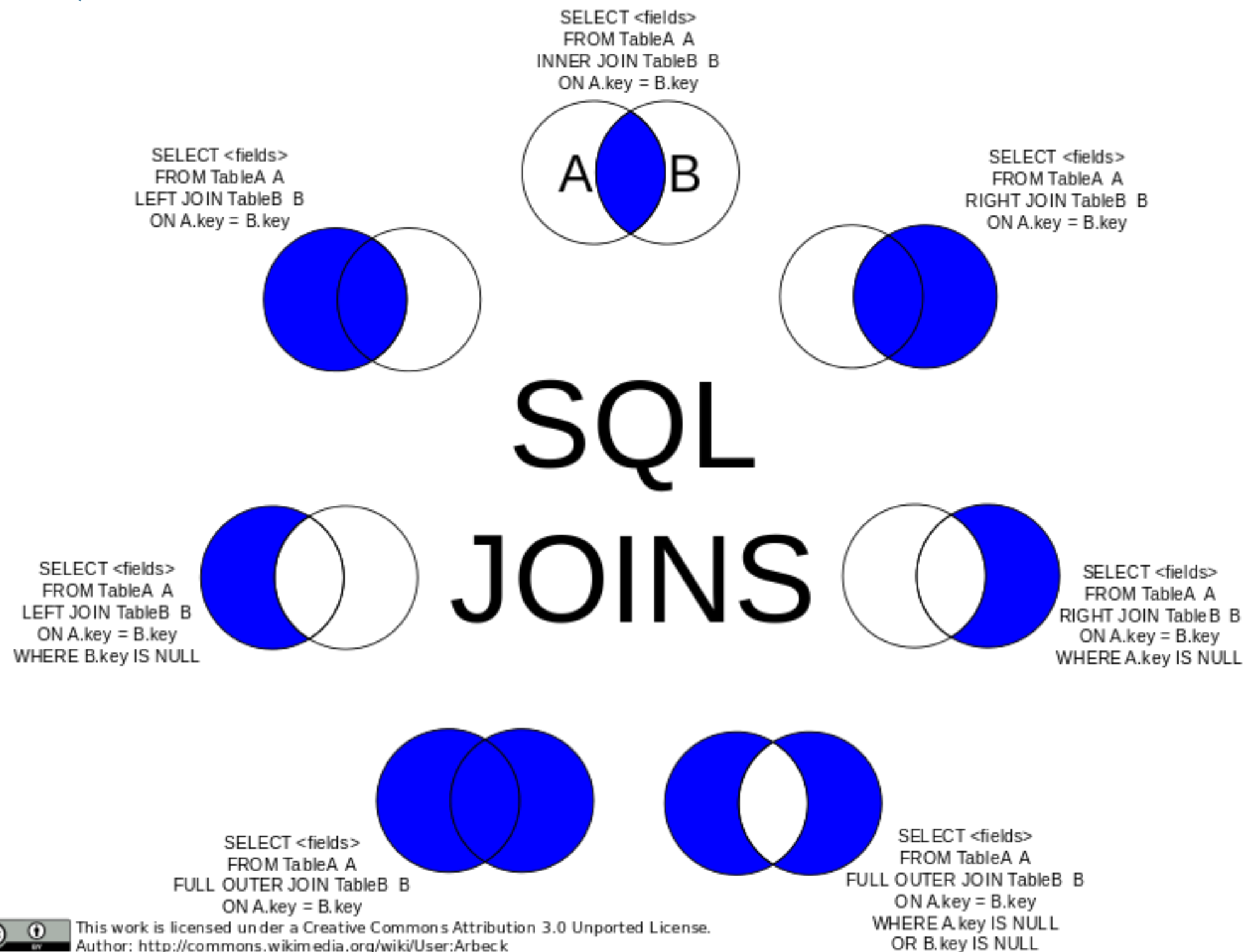
User1	User2
1	2
2	3
2	1
3	1
4	1

Users

UserId	FirstName	LastName	JoinYear	GroupNumber
1	Jhon	Smith	2018	1
2	Marry	Perry	2019	1
3	William	Brown	2018	2
4	Daniel	Miler	2017	2



# SQL Joins



This work is licensed under a Creative Commons Attribution 3.0 Unported License.  
Author: <http://commons.wikimedia.org/wiki/User:Arbeck>



# SQLite

In this course, we will be working with SQLite

Useful Links:

- [SQLite.org](https://sqlite.org)
- [DB Browser for SQLite](#)
- [sqlite3 module](#)



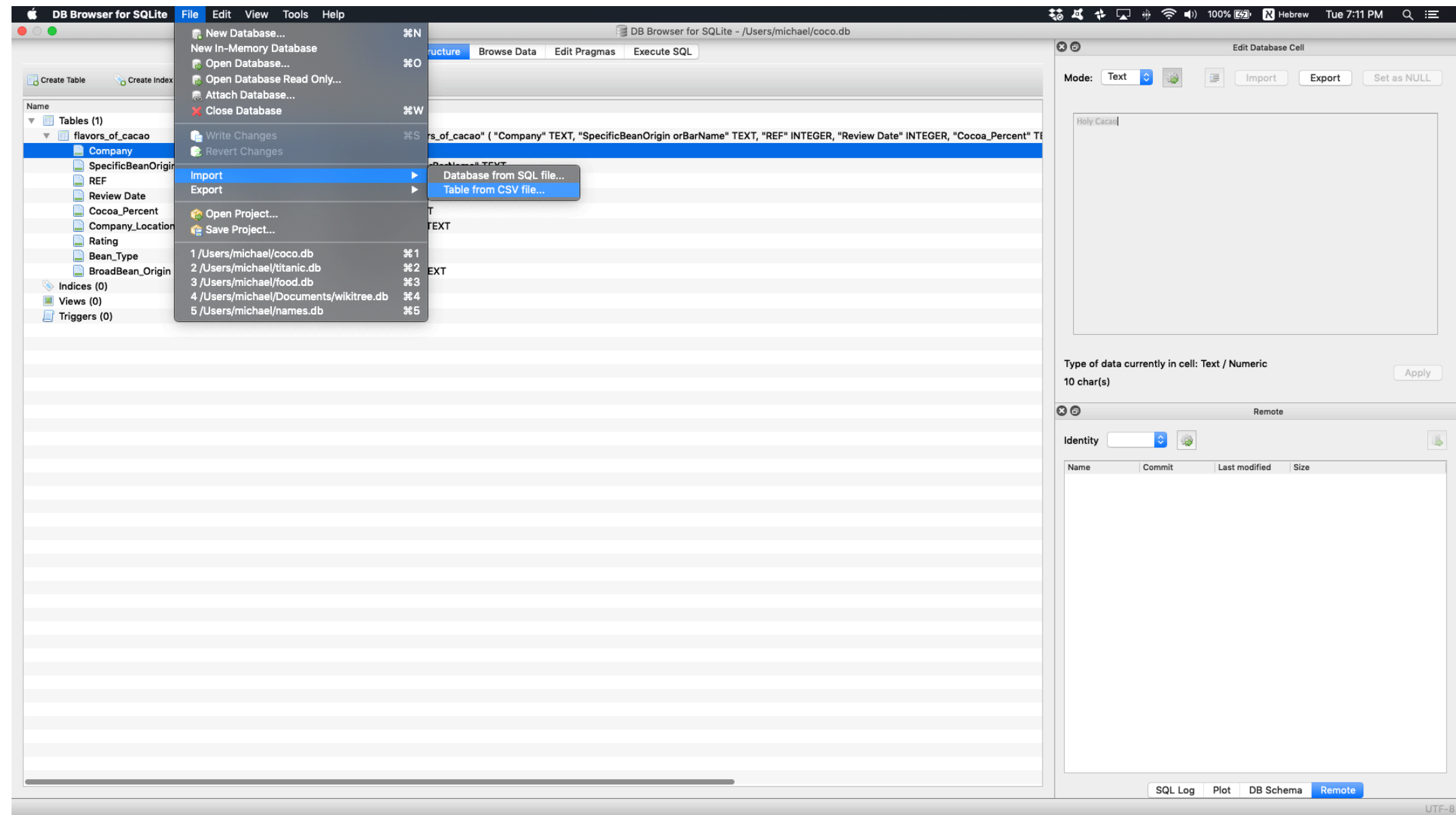


# WORKING WITH REAL WORLD DATA





# Importing Dataset from CSV

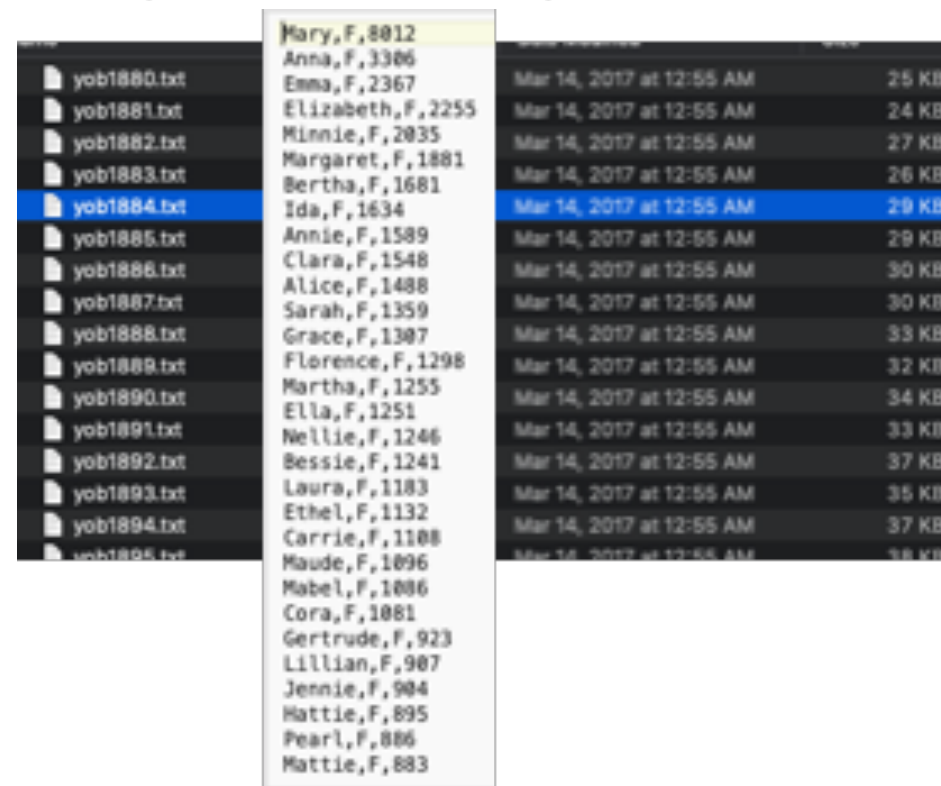




# Example 1: Baby Names

There is an [open datasets](#) containing the names of babies that was born each years.

Let's use this dataset to discover various trends



The image shows a file explorer interface. On the left, a list of files named yob1880.txt through yob1895.txt is displayed. The file yob1884.txt is selected. To the right, the contents of yob1884.txt are shown, listing baby names, their frequency, and the year. The first few entries are Mary, F, 8012; Anna, F, 3306; Emma, F, 2367; Elizabeth, F, 2255; Minnie, F, 2035; Margaret, F, 1881; Bertha, F, 1681; and Ida, F, 1634. The list continues with many other names and their frequencies.

yob1880.txt	Mary, F, 8012	Mar 14, 2017 at 12:55 AM	25 KB
yob1881.txt	Anna, F, 3306	Mar 14, 2017 at 12:55 AM	24 KB
yob1882.txt	Emma, F, 2367	Mar 14, 2017 at 12:55 AM	27 KB
yob1883.txt	Elizabeth, F, 2255	Mar 14, 2017 at 12:55 AM	26 KB
yob1884.txt	Minnie, F, 2035	Mar 14, 2017 at 12:55 AM	29 KB
yob1885.txt	Margaret, F, 1881	Mar 14, 2017 at 12:55 AM	29 KB
yob1886.txt	Bertha, F, 1681	Mar 14, 2017 at 12:55 AM	30 KB
yob1887.txt	Ida, F, 1634	Mar 14, 2017 at 12:55 AM	30 KB
yob1888.txt	Annie, F, 1589	Mar 14, 2017 at 12:55 AM	33 KB
yob1889.txt	Clara, F, 1548	Mar 14, 2017 at 12:55 AM	32 KB
yob1890.txt	Alice, F, 1488	Mar 14, 2017 at 12:55 AM	34 KB
yob1891.txt	Sarah, F, 1359	Mar 14, 2017 at 12:55 AM	33 KB
yob1892.txt	Grace, F, 1307	Mar 14, 2017 at 12:55 AM	37 KB
yob1893.txt	Florence, F, 1298	Mar 14, 2017 at 12:55 AM	36 KB
yob1894.txt	Martha, F, 1255	Mar 14, 2017 at 12:55 AM	37 KB
yob1895.txt	Ella, F, 1251	Mar 14, 2017 at 12:55 AM	38 KB
	Nellie, F, 1246		
	Bessie, F, 1241		
	Laura, F, 1183		
	Ethel, F, 1132		
	Carrie, F, 1108		
	Maude, F, 1096		
	Mabel, F, 1086		
	Cora, F, 1081		
	Gertrude, F, 923		
	Lillian, F, 907		
	Jennie, F, 904		
	Hattie, F, 895		
	Pearl, F, 886		
	Mattie, F, 883		

See also Kaggle [Dataset](#)



# Example 1 - Questions

- How many rows in the dataset?
- How many distinct names in the dataset?
- What is the most common name? (males/female)
- How many names starts with 'B'? How common is the name Beyonce?
- What is the rarest name for female babies, and starts with z, and were born at 2001?



# Example 2: 1000 Netflix Shows

title,rating,ratingLevel,ratingDescription,releaseyear,userratingscore,userratingsize  
White Chicks,PG-13,"crude and s... 80 2004 82 80  
Lucky Number Slevin,R,"strong viole... 100 2006 NA 82  
Grey's Anatomy,TV-14,Parents stro... 90 2016 98 80  
Prison Break,TV-14,Parents stro... 90 2008 98 80  
How I Met Your Mother,TV-PG,Parental gui... 70 2014 94 80  
Supernatural,TV-14,Parents stro... 90 2016 95 80  
Breaking Bad,TV-MA,For mature a... 110 2013 97 80  
The Vampire Diaries,TV-14,Parents stro... 90 2017 91 80  
The Walking Dead,TV-MA,For mature a... 110 2015 98 80  
Pretty Little Liars,TV-14,Parents stro... 90 2016 96 80  
Once Upon a Time,TV-PG,Parental gui... 70 2016 98 80  
Sherlock,TV-14,Parents stro... 90 2016 98 80  
Death Note,TV-14,Parents stro... 90 2006 77 80  
Naruto,TV-PG,Parental gui... 70  
The Hunter,R,language a

Table name

Column names in first line ☒

Field separator

Quote character

Encoding

Trim fields? ☐

Advanced

	title	rating	ratingLevel	ratingDescription	releaseyear	userratingscore	userratingsize
1	White Chicks	PG-13	crude and s...	80	2004	82	80
2	Lucky Numb...	R	strong viole...	100	2006	NA	82
3	Grey's Anato...	TV-14	Parents stro...	90	2016	98	80
4	Prison Break	TV-14	Parents stro...	90	2008	98	80
5	How I Met Y...	TV-PG	Parental gui...	70	2014	94	80
6	Supernatural	TV-14	Parents stro...	90	2016	95	80
7	Breaking Bad	TV-MA	For mature a...	110	2013	97	80
8	The Vampire...	TV-14	Parents stro...	90	2017	91	80
9	The Walking...	TV-MA	For mature a...	110	2015	98	80
10	Pretty Little ...	TV-14	Parents stro...	90	2016	96	80
11	Once Upon ...	TV-PG	Parental gui...	70	2016	98	80

,90,2016,98,80  
90,2008,98,80  
,2014,94,80  
90,2016,95,80  
,97,80  
under.,90,2017,91,80  
2015,98,80  
under.,90,2016,96,80  
,98,80  
016,95,80  
,2006,77,80



# Example 2: Questions

- How many movies?
- How many movies in each rating category?
- What is the highest rated movie in each category?



# Working with Jupyter Notebooks

In this course, we are going to work with Python

- I recommend to install [Anaconda](#) and [PyCharm](#)
- It is recommended to work with virtual environment

```
$ conda create -n venv python=3.7 anaconda  
$ source activate venv
```

- We will use Jupyter Notebooks

```
$ jupyter notebook
```

- I also recommend to get familiar with [ipython](#)



# Practice SQL Online

- [SQLZOO](#)
- [SQL Murder Mystery](#)



Let's move to reviewing the  
course first notebook