510 DATA SCIENCE

Lecture 01

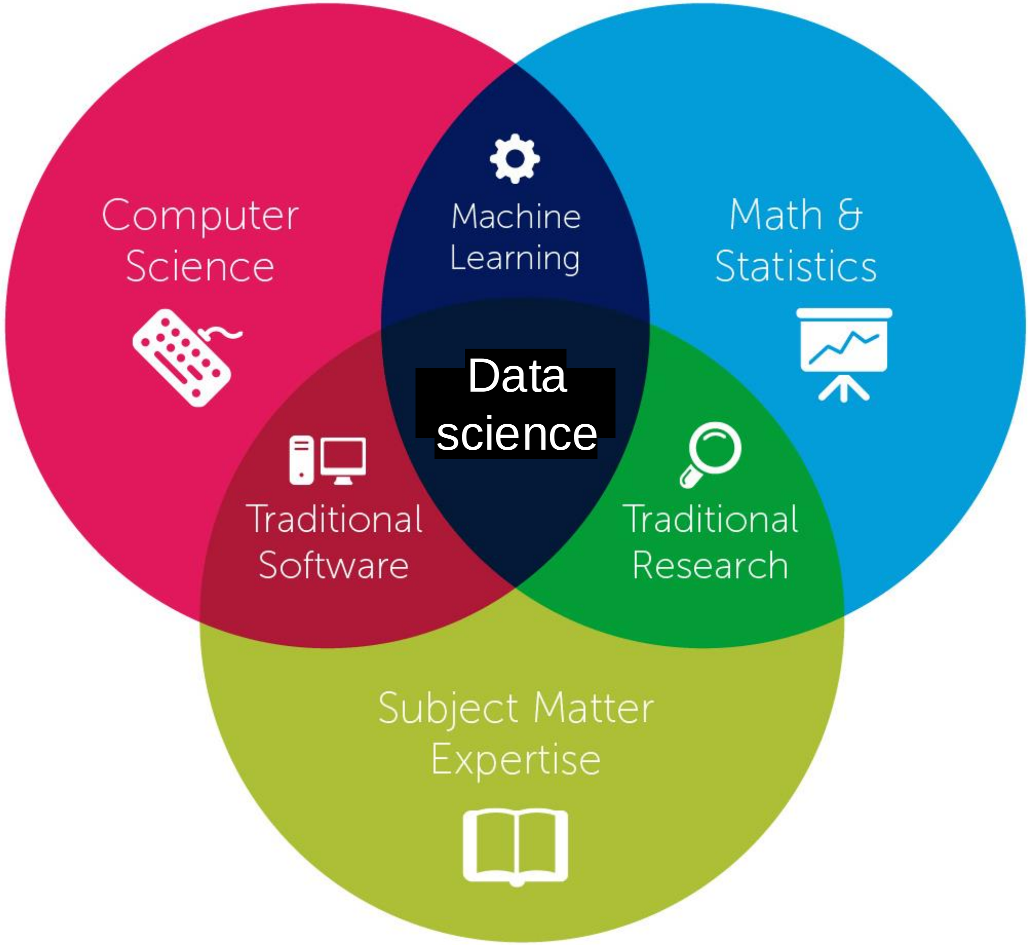
Fall 2024

Instructor: Assoc. Prof. Şener Özönder

Email: sener.ozonder@bogazici.edu.tr

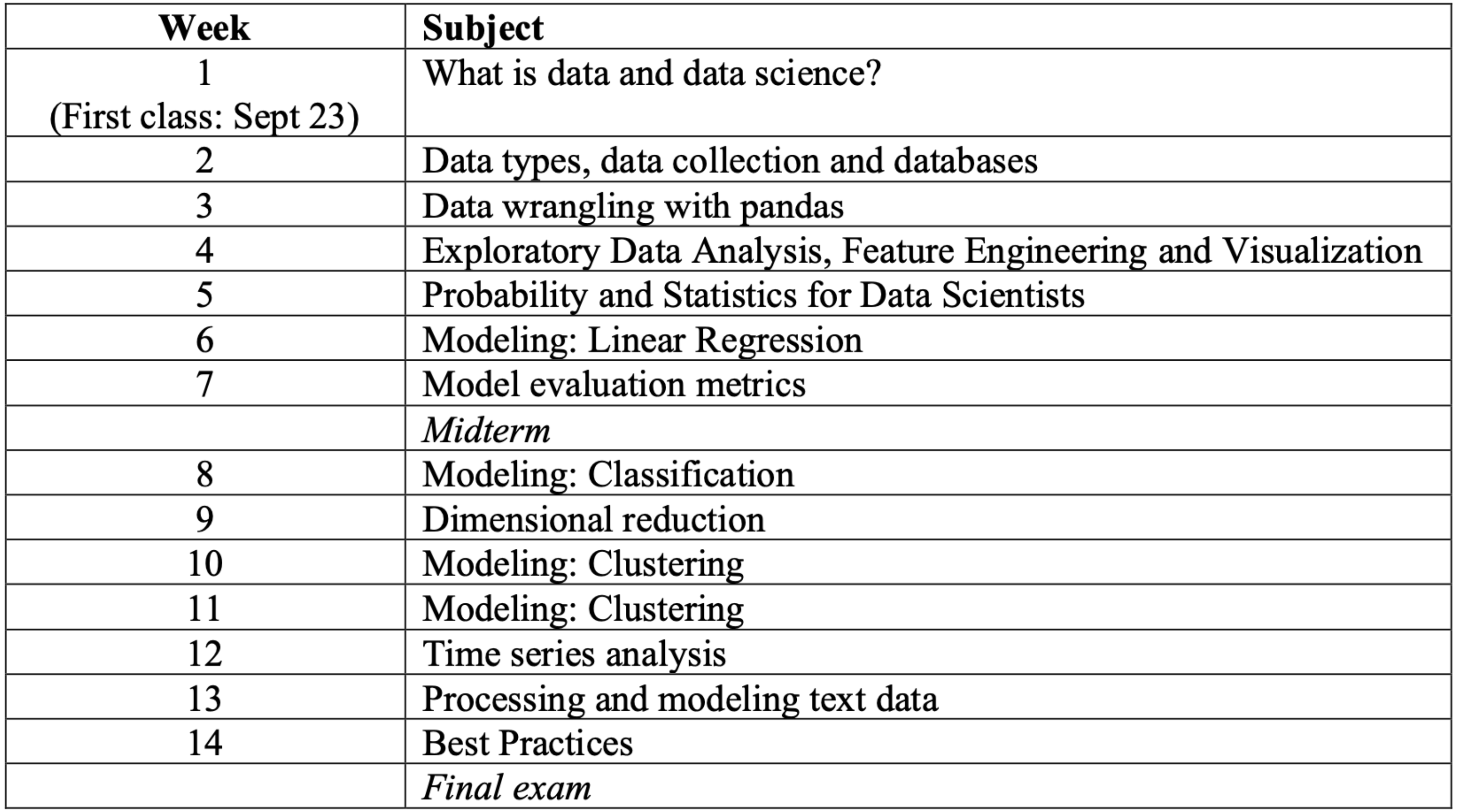
Institute for Data Science & Artificial Intelligence

Boğaziçi University



What’s in this class?

- In this course, we will focus on core data science concepts such as data cleaning, descriptive statistics, transforming data, modelling (regression, classification, clustering) and prediction/inference. (Check syllabus & lecture slides at http://moodle.bogazici.edu.tr)



Week	Subject
1 (First class: Sept 23)	What is data and data science?
2	Data types, data collection and databases
3	Data wrangling with pandas
4	Exploratory Data Analysis, Feature Engineering and Visualization
5	Probability and Statistics for Data Scientists
6	Modeling: Linear Regression
7	Model evaluation metrics
	Midterm
8	Modeling: Classification
9	Dimensional reduction
10	Modeling: Clustering
11	Modeling: Clustering
12	Time series analysis
13	Processing and modeling text data
14	Best Practices
	Final exam

What's *not* in this class?

- We'll do some of the things below a little bit but these subjects will be treated in depth in their respective courses.

- **No data engineering**

520 Big Data Systems
523 Cloud Computing and Distributed Systems
524 Software Design for Data Science

- **No dashboarding but some viz**

522 Business Intelligence and Analytics
521 Data Visualization for Data Scientists

- **Very little image processing**

543 Image Processing with Machine Learning
544 Computer Vision with Machine Learning

- **A bit text data processing**

545 Natural Language Processing

- **No web scraping**

526 Web Mining
531 Social Media Analytics

- **Some basic machine learning, but not deep learning**

512 Machine Learning
541 Deep Learning (2nd semester)

- **Some statistics, but not in depth.**

514 Statistical Inference

Python *vs* R

- We will use Python because with Python you can not only do data science and machine learning, but also things like web and application development, dashboarding etc.

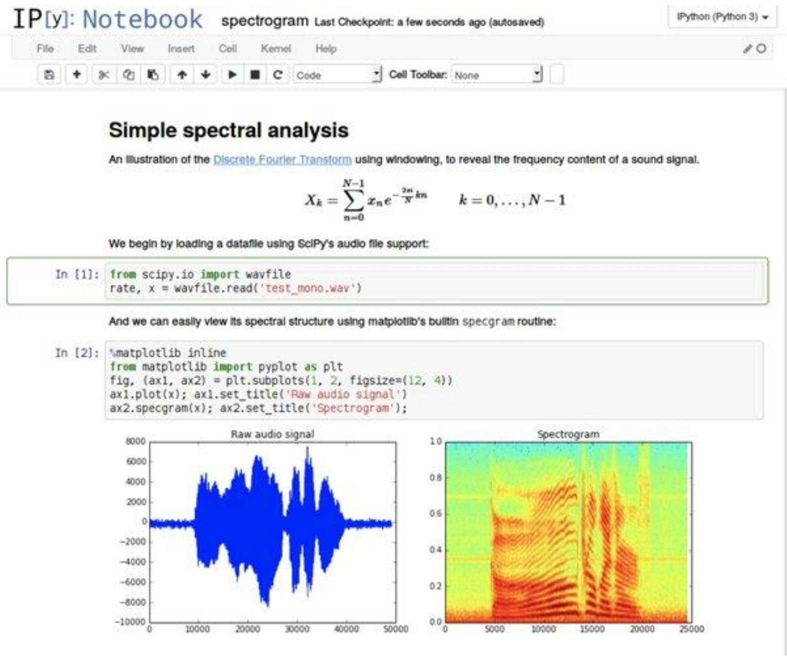
IDE (Integrated Development Environment) is where you write code. For Python, what's mostly used is Jupyter Notebook (Anaconda). You can also use Google Colab to code in Python even in your cellphone; it's an online IDE and it uses Google's computational resources. Another option is Microsoft's VS Code, where you can code in almost in all languages. Start with Jupyter and Google Colab.

- R is originally developed for statistical analysis. Mostly statisticians use it. Transforming data and statistical tests sometimes easier in R than Python, but it will *not* allow you to build an application to deploy except some web dashboards. Still, if you're collaborating with statisticians or bio-statistics people and if they're already using R, then you can easily learn R code in a week.

To use R, install RStudio IDE: https://posit.co/products/open-source/rstudio/

Python *vs* R

Jupyter Notebook



Simple spectral analysis

An illustration of the Discrete Fourier Transform using windowing, to reveal the frequency content of a sound signal.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

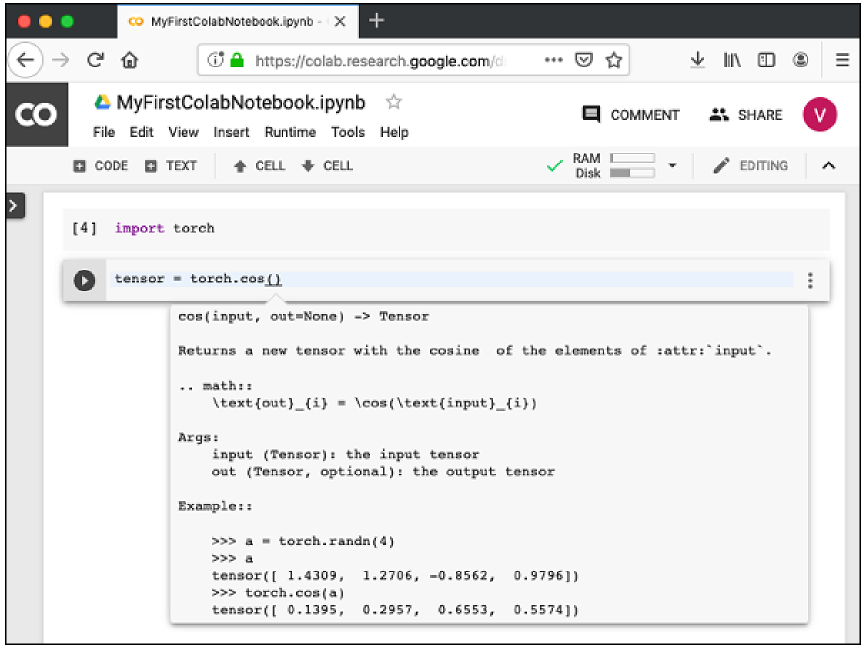
We begin by loading a datafile using SciPy's audio file support:

```
In [1]: from scipy.io import wavfile
        rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin specgram routine:

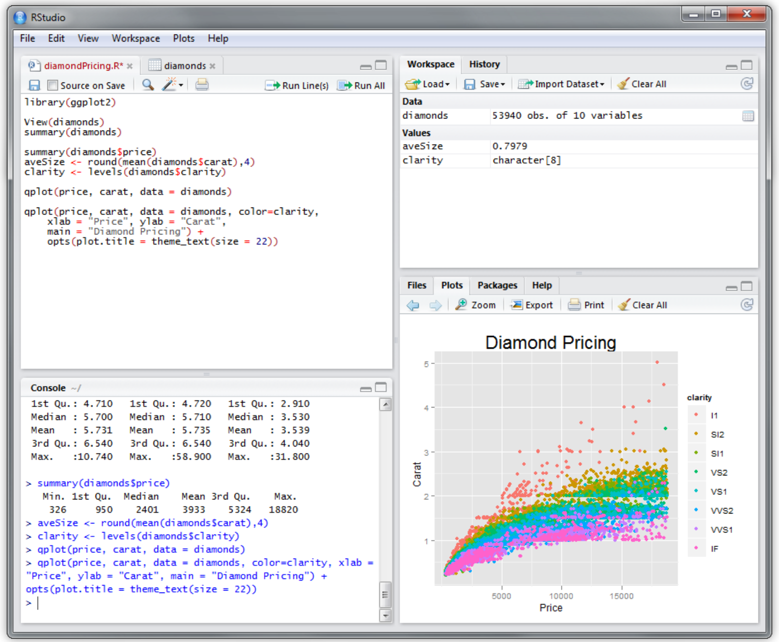
```
In [2]: %matplotlib inline
        from matplotlib import pyplot as plt
        fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
        ax1.plot(x); ax1.set_title('Raw audio signal')
        ax2.specgram(x); ax2.set_title('Spectrogram');
```

Google Colab


MyFirstColabNotebook.ipynb -
https://colab.research.google.com/d
CO
MyFirstColabNotebook.ipynb
COMMENT
SHARE
V
File
Edit
View
Insert
Runtime
Tools
Help
CODE
TEXT
CELL
CELL
RAM
Disk
EDITING

```
[4]  import torch

tensor = torch.cos()

cos(input, out=None) -> Tensor

Returns a new tensor with the cosine  of the elements of :attr:`input`.

.. math::
    \text{out}_{i} = \cos(\text{input}_{i})

Args:
    input (Tensor): the input tensor
    out (Tensor, optional): the output tensor

Example::

    >>> a = torch.randn(4)
    >>> a
    tensor([ 1.4309,  1.2706, -0.8562,  0.9796])
    >>> torch.cos(a)
    tensor([ 0.1395,  0.2957,  0.6553,  0.5574])
```

R Studio



```
library(ggplot2)

View(diamonds)
summary(diamonds)

summary(diamonds$price)
aveSize <- round(mean(diamonds$carat),4)
clarity <- levels(diamonds$clarity)

qplot(price, carat, data = diamonds)

qplot(price, carat, data = diamonds, color=clarity,
    xlab = "Price", ylab = "Carat",
    main = "Diamond Pricing") +
    opts(plot.title = theme_text(size = 22))
```

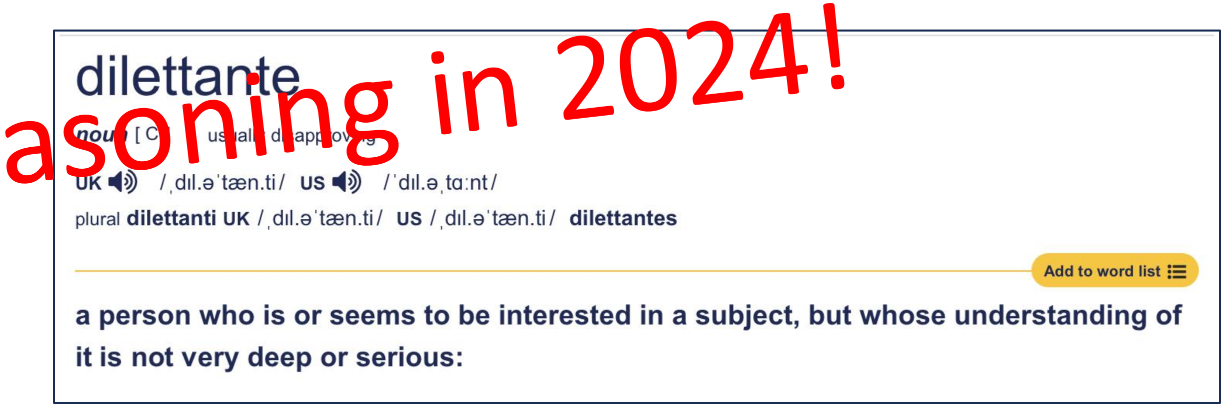
Data	
diamonds	53940 obs. of 10 variables
Values	
aveSize	0.7979
clarity	character[8]

```
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800

> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat),4)
> clarity <- levels(diamonds$clarity)
> qplot(price, carat, data = diamonds)
> qplot(price, carat, data = diamonds, color=clarity, xlab =
"Price", ylab = "Carat", main = "Diamond Pricing") +
opts(plot.title = theme_text(size = 22))
> |
```

ChatGPT policy

I think it started reasoning in 2024!

- ChatGPT is a 30 year old dillattente or encyclopedist; can't reason much.



dilettante

noun [C] usually disapproving

UK /ˌdɪl.əˈtæn.ti/ US /ˈdɪl.əˌtɑːnt/

plural dilettanti UK /ˌdɪl.əˈtæn.ti/ US /ˌdɪl.əˈtæn.ti/ dilettantes

Add to word list

a person who is or seems to be interested in a subject, but whose understanding of it is not very deep or serious:

- Can you use ChatGPT/Claude/Gemine in this class? Not "you can", "you should." Why look up Google rather than simply asking ChatGPT to find the piece of code you need?
- In the pre-ChatGPT era, people relied on websites like Stack Overflow for codes. Now, ChatGPT can provide the code you need through a simple chat interface.
- But here's the caviat: ChatGPT is great as a conversational assistant or an encyclopedist to lead you to the right direction, but don't rely on it as it may mislead you. So use it with caution!

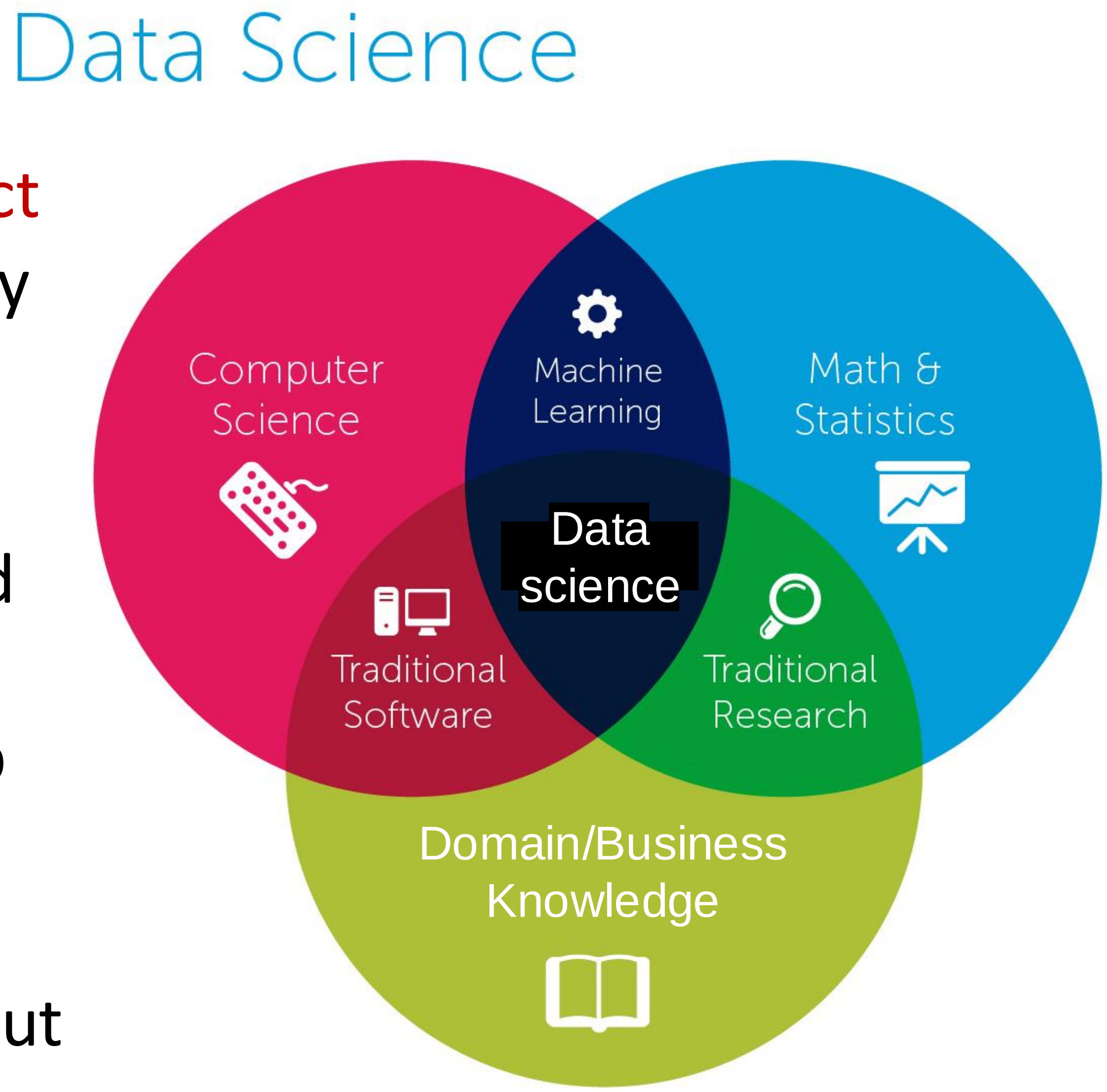
ChatGPT policy

From the Syllabus:

Collaboration and ChatGPT policy for homeworks: Collaborating with fellow students on assignments is both permitted and encouraged. You're also welcome to use ChatGPT for conceptual clarity and guidance on Python coding. However, all submitted work must fundamentally be your own. Verbatim copying of answers from peers or ChatGPT is prohibited. While you can use code directly from ChatGPT, ensure you understand its structure and logic thoroughly, as the instructor will ask you to explain any code you submit in the midterm and final exams.

What is data science?

- An umbrella term for creating impact for business or research problems by using both statistics/math and computer science together.
- Sometimes “data science” is defined very broadly to include collection, cleaning of data including setting up the hardware and software infrastructure as well as creating dashboards to show the analytics, but these roles have different names.



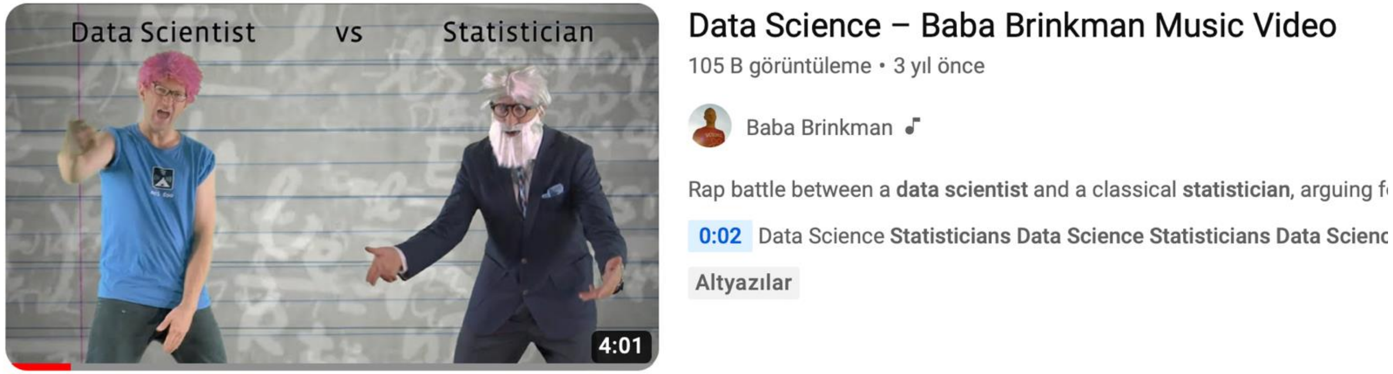
Values created by a data scientist

Some examples:

- **Industry**: Process Optimization, Quality Control, Predictive Maintenance, Energy Efficiency
- **Business**: Customer Insights, Marketing Optimization, Sales Forecasting, Risk Management, Inventory Management, Fraud Detection, Churn Prediction, Pricing Strategy, Operational Efficiency and anything else that will increase revenue.
- **Academia**: Gather insights from data, test hypotheses, discover unknown structures and relations within the data, simulate and model complex systems,

What happened to good old statistics?

- Data science and machine learning is also called “statistical learning”. Data science sometimes has more emphasis on black box modeling such as deep learning of big data for the purpose of prediction/inference where explainability is often sacrificed. Statisticians don’t like this.
- When you need to not just model but understand the data (“data mining”) and extract knowledge and causal relations from data (“knowledge discovery”), good old statistics may become more important than data science.
- Having said that, the intersection of these two approaches grows everyday.



Data Science – Baba Brinkman Music Video

105 B görüntüleme • 3 yıl önce

Baba Brinkman

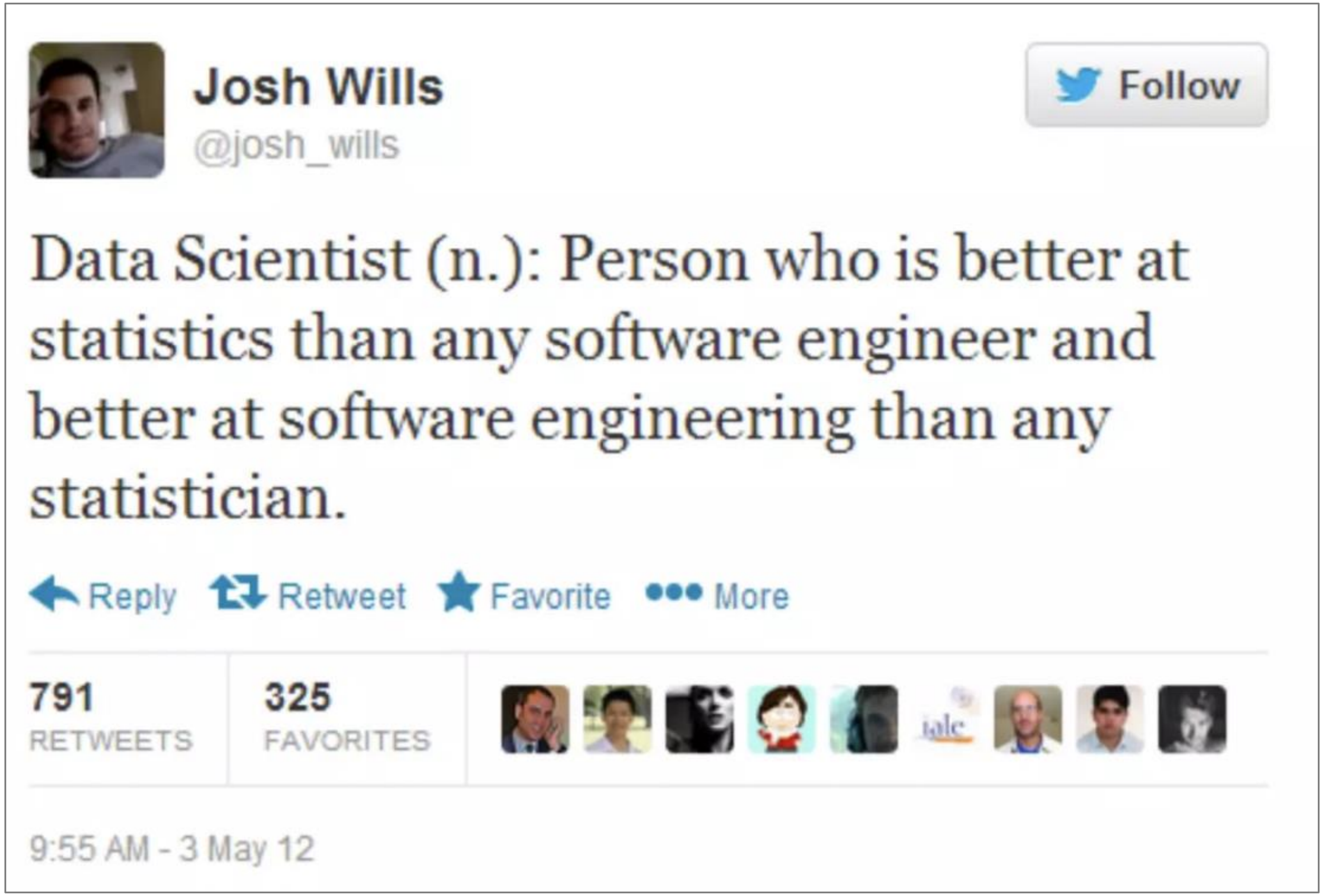
Rap battle between a **data scientist** and a classical **statistician**, arguing f

0:02 Data Science **Statisticians Data Science Statisticians Data Scienc**

Altyazılar

http://youtu.be/uHGlCi9jOWY

Who is a data scientist?



Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

791 RETWEETS

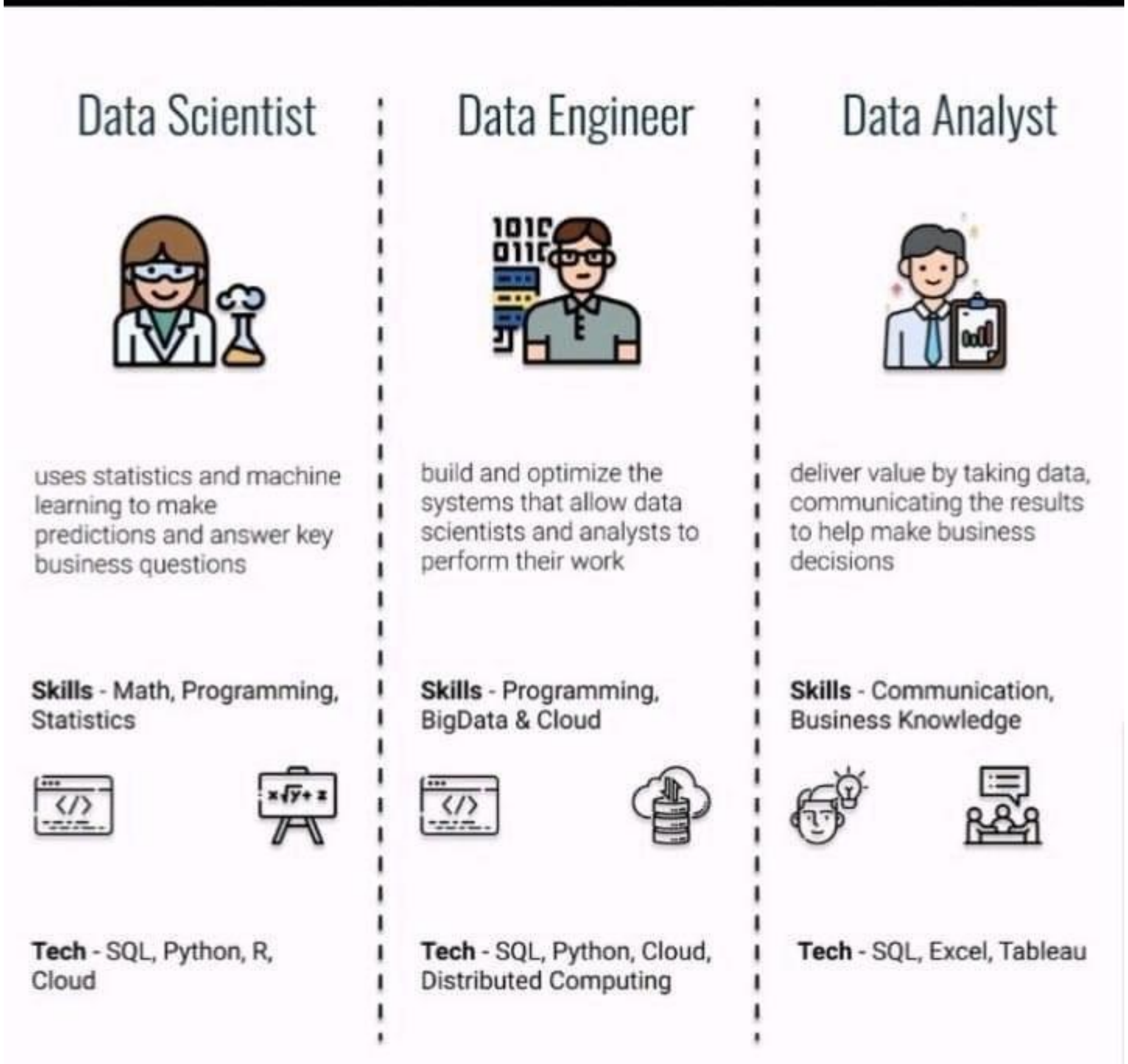
325 FAVORITES

9:55 AM - 3 May 12

Data science and ML is industry driven!

- Data science and machine learning are industry driven because big data is mostly produced by industry not academia. Also big data is not publicly shared for academicians to play with. (Industry: Hospitals, tech companies, financial sector etc.)
- Machine learning (ML) is not like theoretical physics where you can discover new particles with only math. Most of the time, ML models are designed and optimized specifically for the dataset at hand, so without real data you can’t develop good models on fake/synthetic data. *ML models learns from the data!*
- Also developments in computer hardware and software, emergence of cheat IoT sensors, collection of big data by tech companies (Google, Facebook etc.) created the field of data science besides statistics. Data science started to become famous around 2010s.

Data roles



Data Scientist

uses statistics and machine learning to make predictions and answer key business questions

Skills - Math, Programming, Statistics

Tech - SQL, Python, R, Cloud

Data Engineer

build and optimize the systems that allow data scientists and analysts to perform their work

Skills - Programming, BigData & Cloud

Tech - SQL, Python, Cloud, Distributed Computing

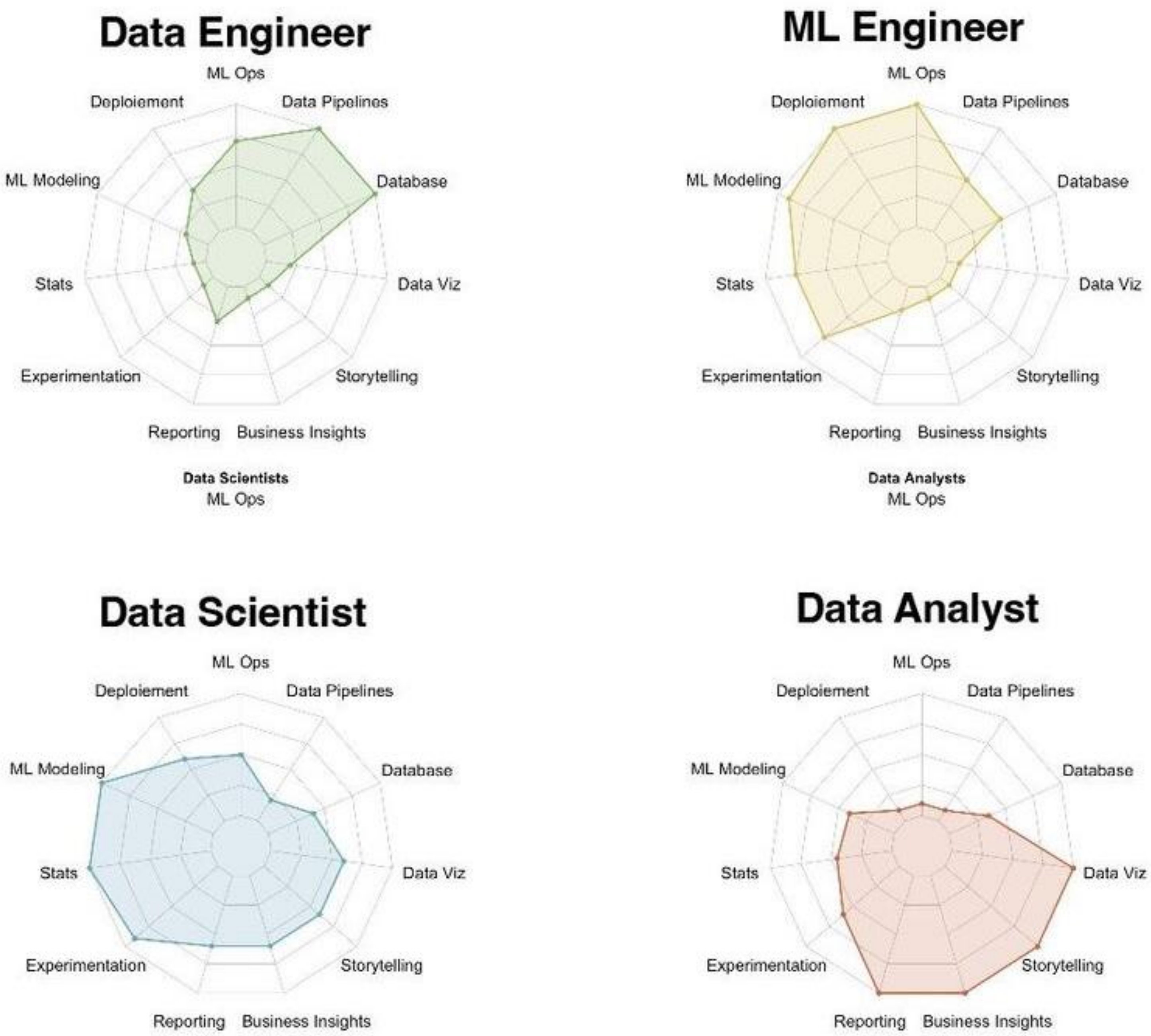
Data Analyst

deliver value by taking data, communicating the results to help make business decisions

Skills - Communication, Business Knowledge

Tech - SQL, Excel, Tableau

Data roles


Data Engineer
ML Ops
Data Pipelines
Database
Data Viz
Storytelling
Business Insights
Reporting
Experimentation
Stats
ML Modeling
Deploiement
ML Engineer
ML Ops
Data Pipelines
Database
Data Viz
Storytelling
Business Insights
Reporting
Experimentation
Stats
ML Modeling
Deploiement
Data Scientists
ML Ops
Data Analysts
ML Ops
Data Scientist
ML Ops
Data Pipelines
Database
Data Viz
Storytelling
Business Insights
Reporting
Experimentation
Stats
ML Modeling
Deploiement
Data Analyst
ML Ops
Data Pipelines
Database
Data Viz
Storytelling
Business Insights
Reporting
Experimentation
Stats
ML Modeling
Deploiement

Data roles

- Even though there are overlaps between them, some roles are directly related to some DSAI courses.

(See course descriptions at https://dsai.boun.edu.tr/course-descriptions)

- You can choose to specialize in one role, but it's always good to have basic level skills in all.
- The most important skill is to understand the business/research problem at hand and being able to formulate the solution and determine the necessary tools for it. This is why you're here!

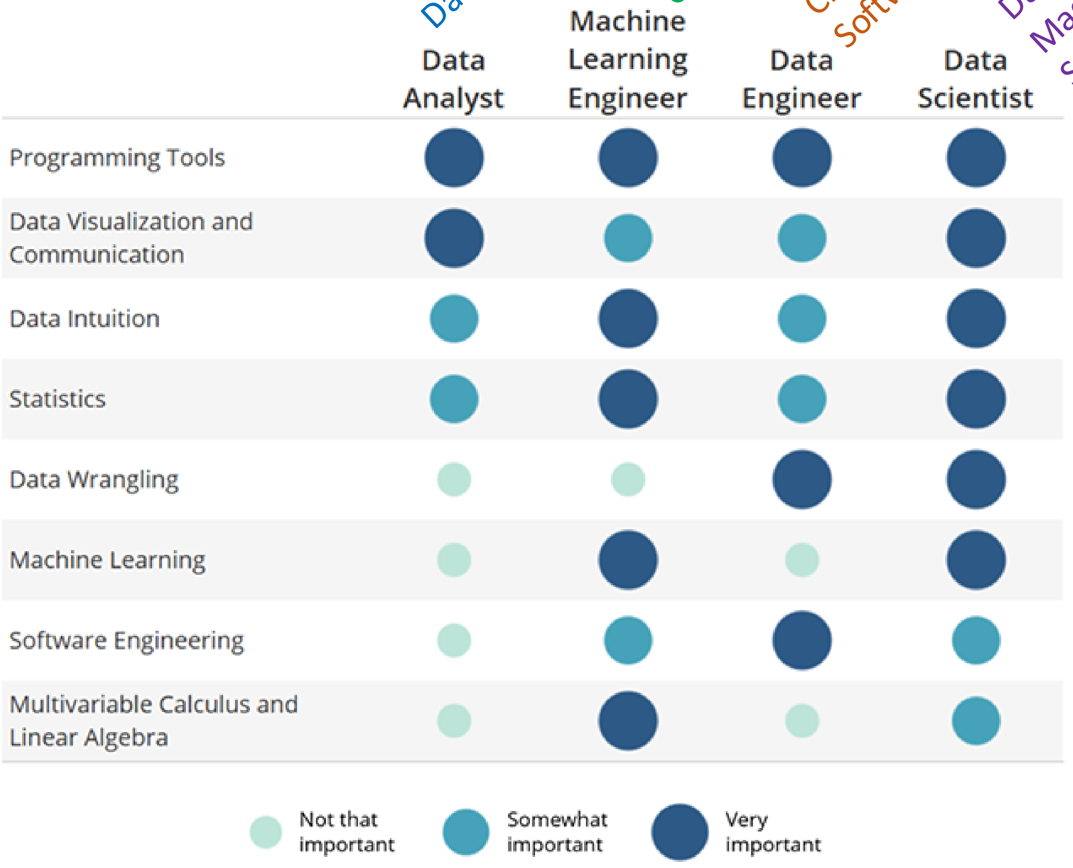
DSAI courses

Data Science
Business Intelligence and Analytics
Data Visualization for Data Scientists

Data Science
Machine Learning
Cloud Computing and Distributed Systems

Big Data Systems
Cloud Computing and Distributed Systems
Software Design for Data Science

Data Science
Machine Learning
Statistical Inference

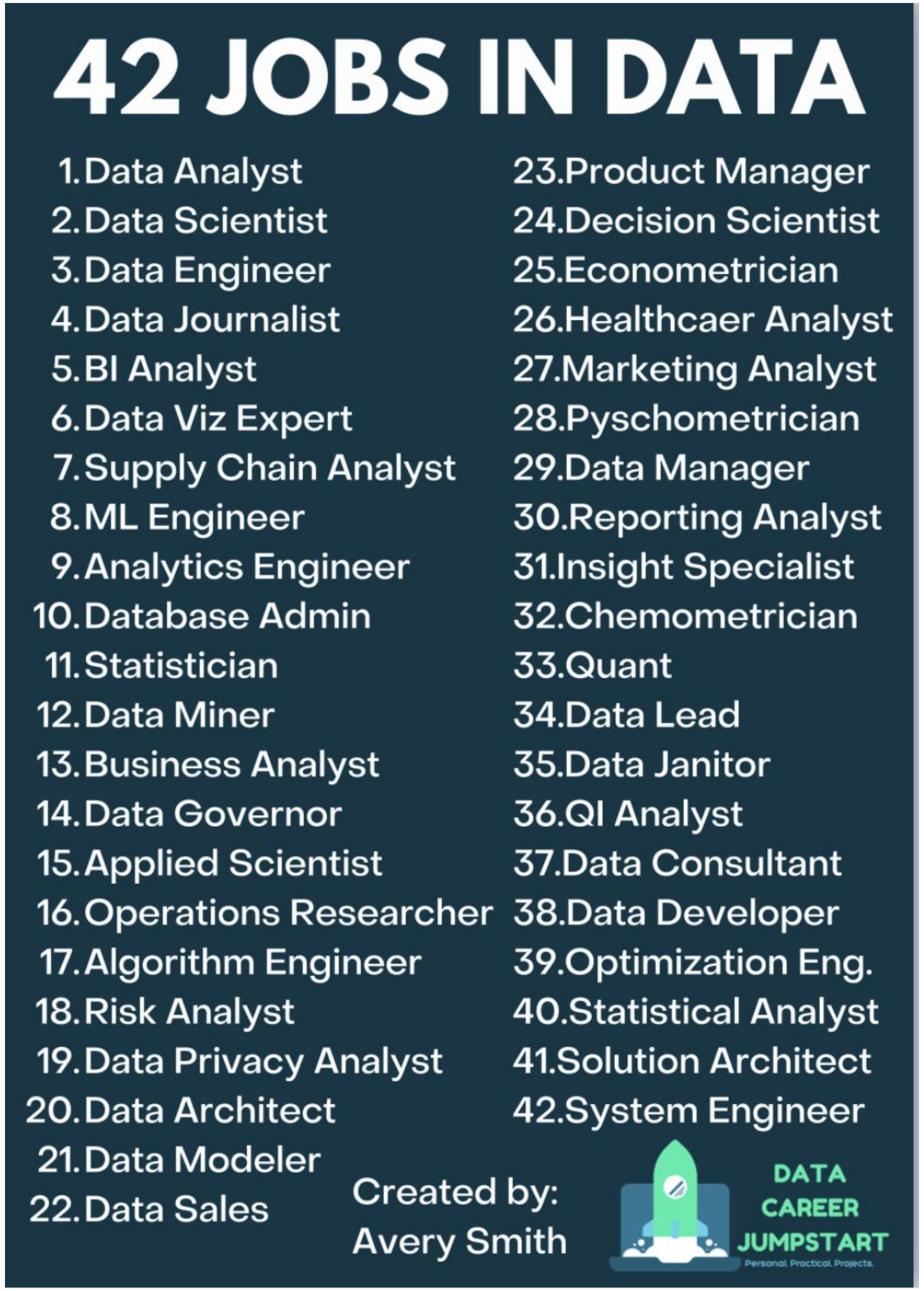


	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	●	●	●	●
Data Visualization and Communication	●	◉	◉	●
Data Intuition	◉	●	◉	●
Statistics	◉	●	◉	●
Data Wrangling	○	○	●	●
Machine Learning	○	●	○	●
Software Engineering	○	◉	●	◉
Multivariable Calculus and Linear Algebra	○	●	○	◉

○ Not that important ◉ Somewhat important ● Very important

Data roles

- Caution: The definition of these roles are not set in stone yet somewhat company and country dependent and still evolving.
- Always define well what you mean by these words while communicating.



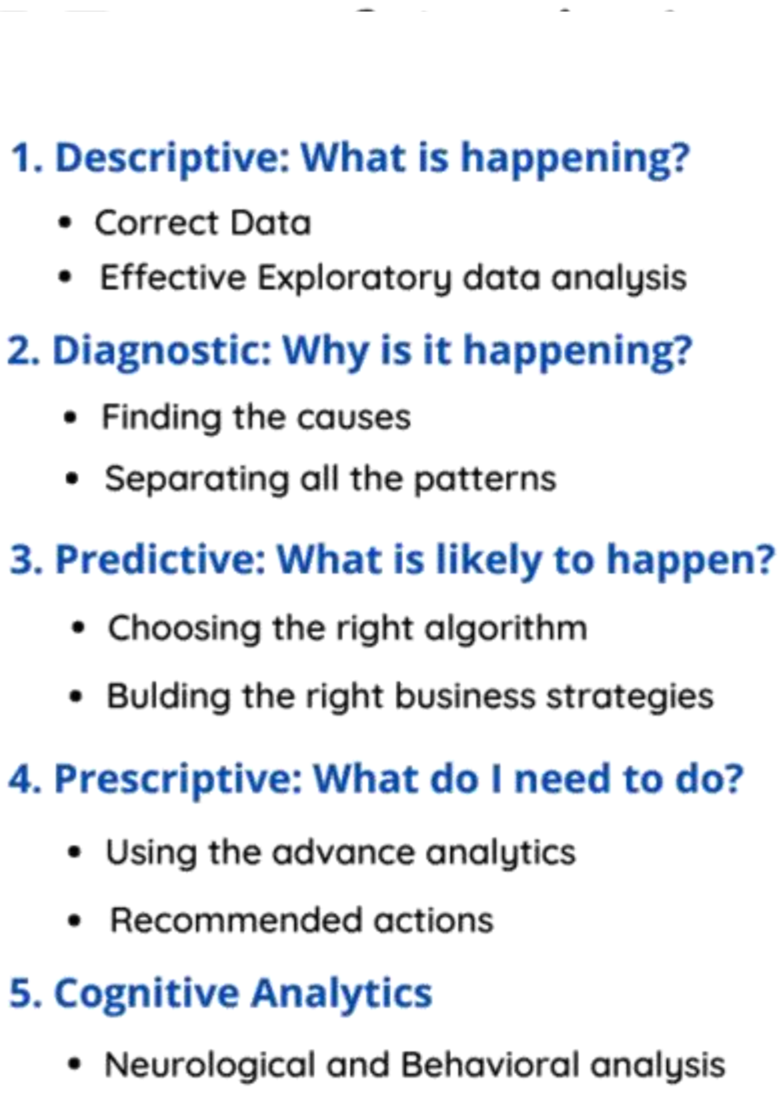
42 JOBS IN DATA

1. Data Analyst
2. Data Scientist
3. Data Engineer
4. Data Journalist
5. BI Analyst
6. Data Viz Expert
7. Supply Chain Analyst
8. ML Engineer
9. Analytics Engineer
10. Database Admin
11. Statistician
12. Data Miner
13. Business Analyst
14. Data Governor
15. Applied Scientist
16. Operations Researcher
17. Algorithm Engineer
18. Risk Analyst
19. Data Privacy Analyst
20. Data Architect
21. Data Modeler
22. Data Sales
23. Product Manager
24. Decision Scientist
25. Econometrician
26. Healthcaer Analyst
27. Marketing Analyst
28. Pyschometrician
29. Data Manager
30. Reporting Analyst
31. Insight Specialist
32. Chemometrician
33. Quant
34. Data Lead
35. Data Janitor
36. QI Analyst
37. Data Consultant
38. Data Developer
39. Optimization Eng.
40. Statistical Analyst
41. Solution Architect
42. System Engineer

Created by:
Avery Smith

DATA
CAREER
JUMPSTART
Personal. Practical. Projects.

5 Types of Analytics



1. Descriptive: What is happening?

- Correct Data
- Effective Exploratory data analysis

2. Diagnostic: Why is it happening?

- Finding the causes
- Separating all the patterns

3. Predictive: What is likely to happen?

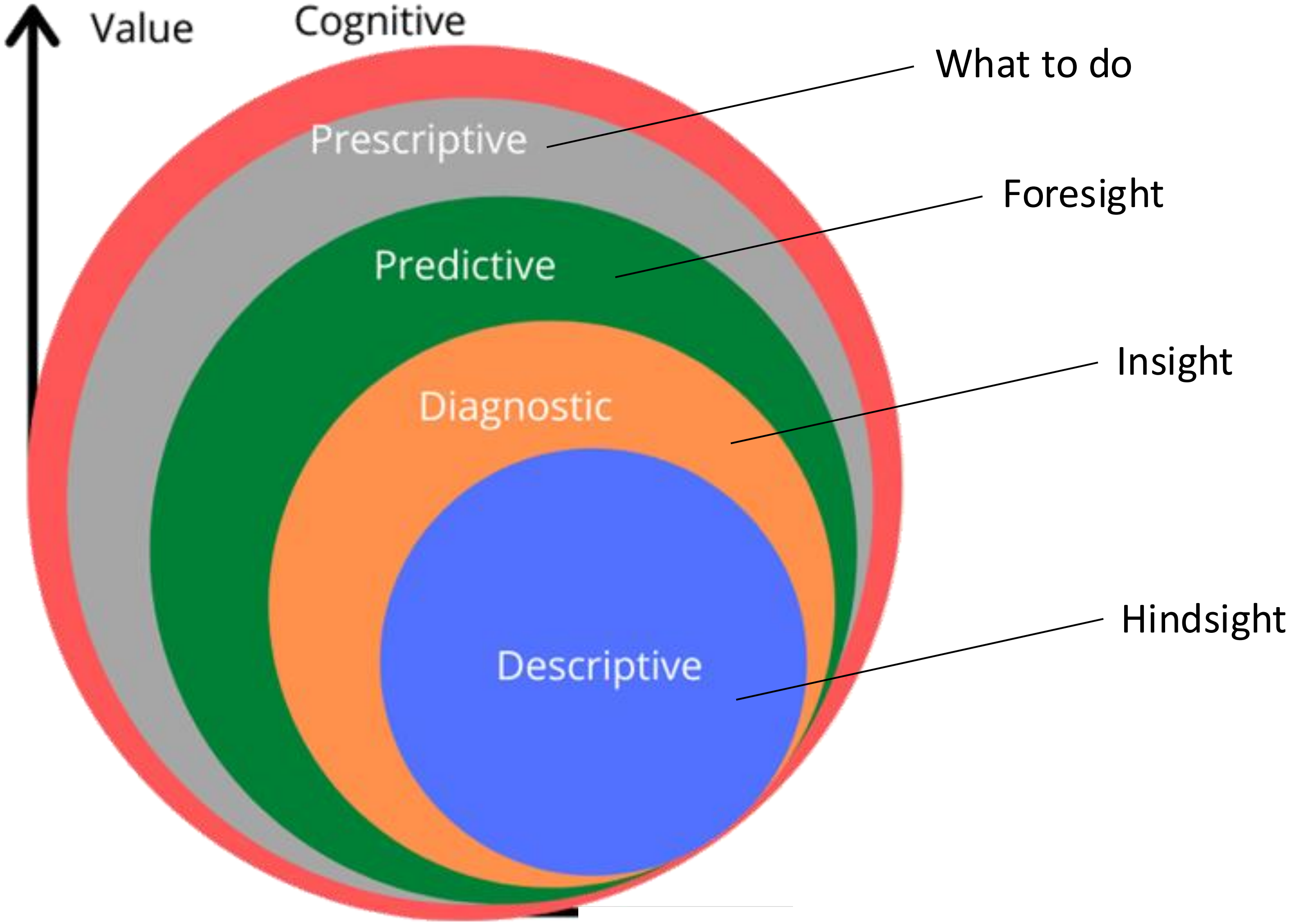
- Choosing the right algorithm
- Bulding the right business strategies

4. Prescriptive: What do I need to do?

- Using the advance analytics
- Recommended actions

5. Cognitive Analytics

- Neurological and Behavioral analysis



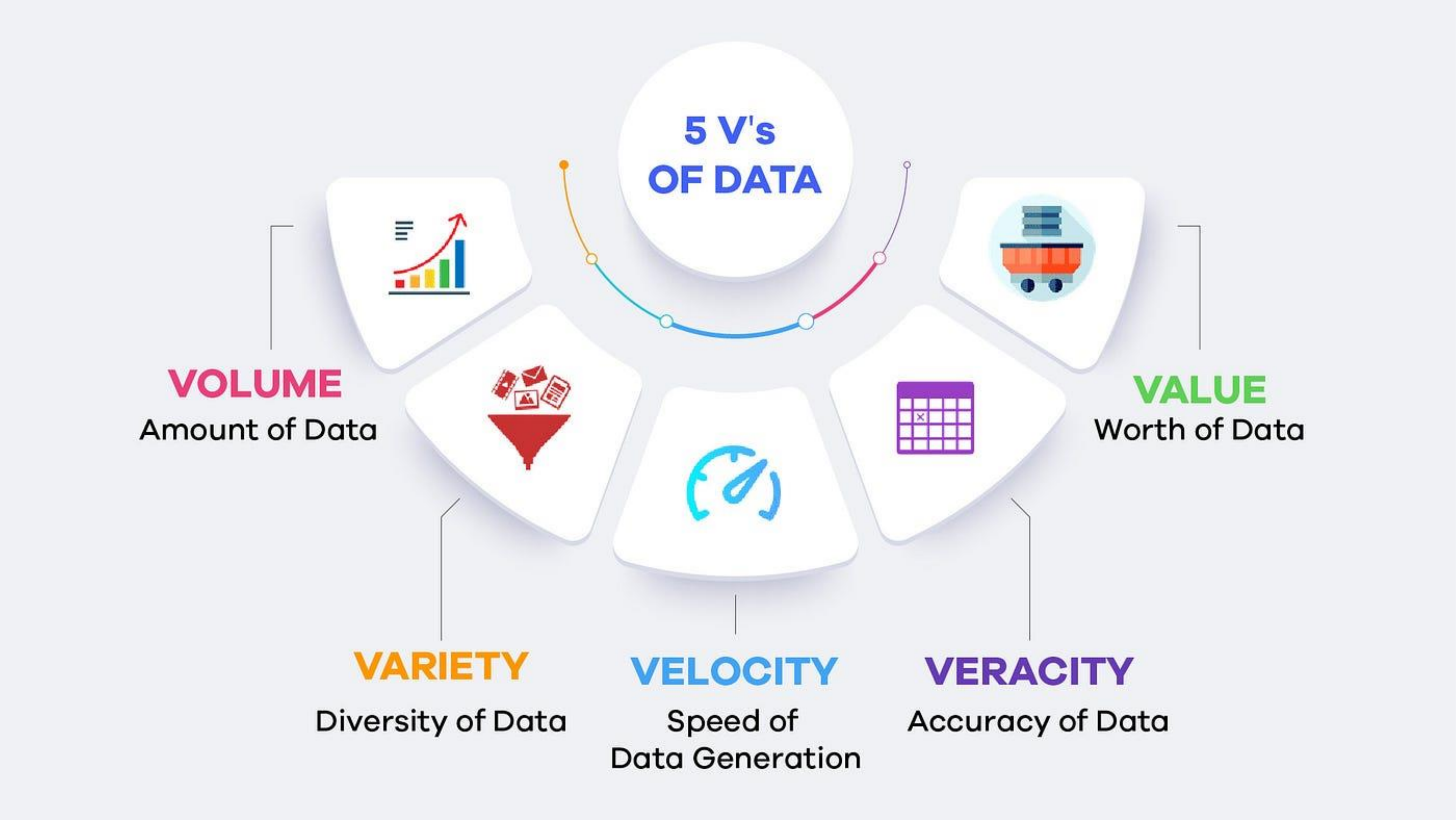
What to do

Foresight

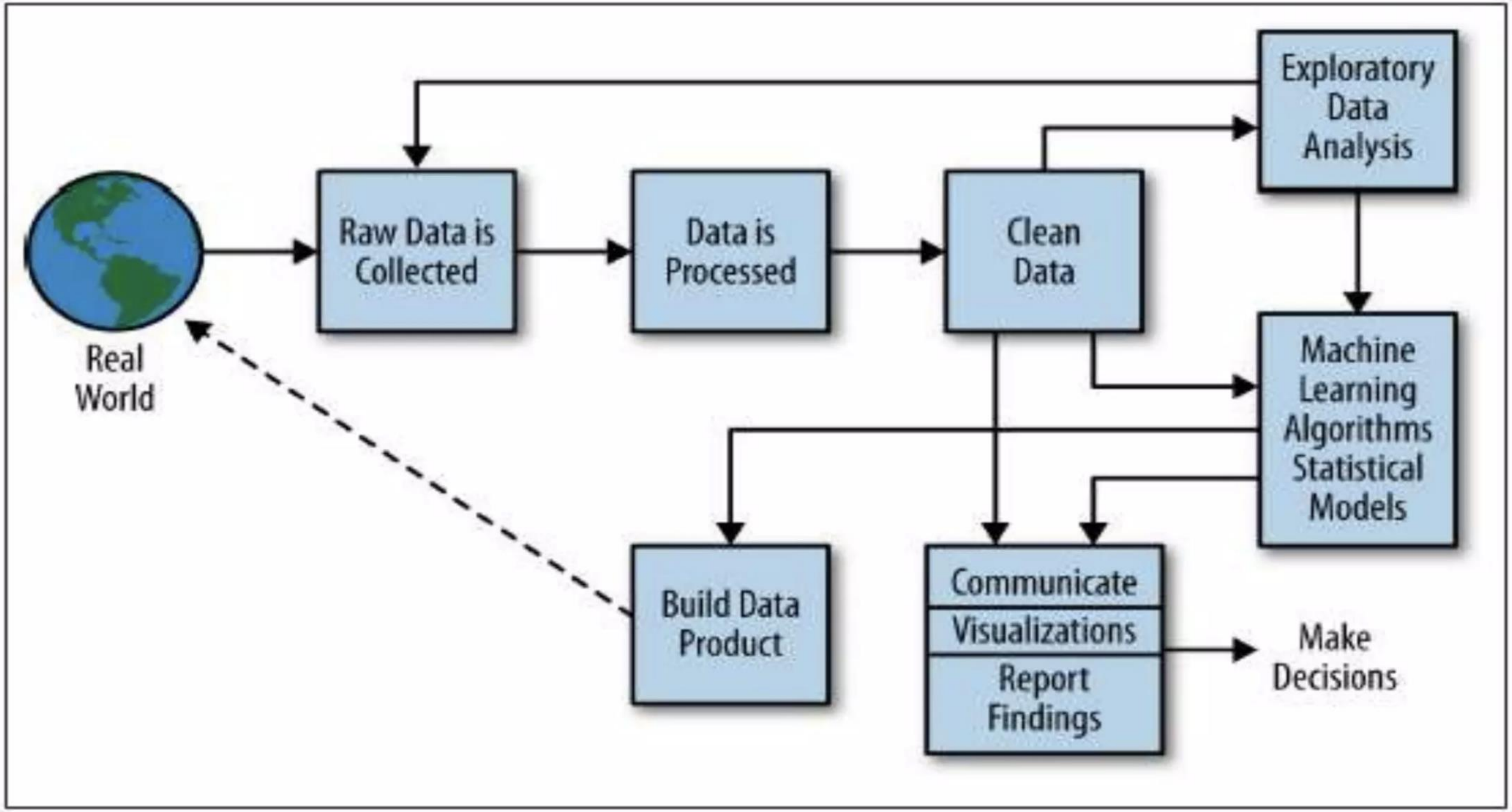
Insight

Hindsight

5 V’s of Big data

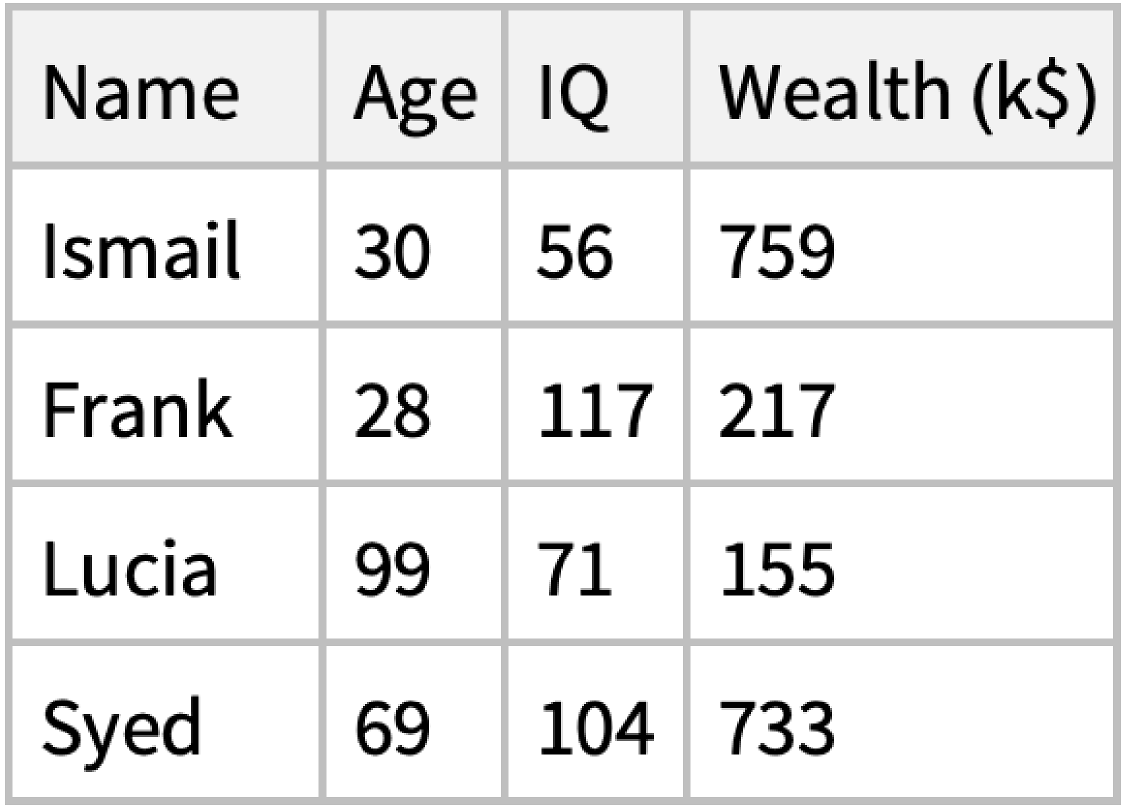


Data Science Lifecycle



Tabular Data

Column, feature, attribute, variable.

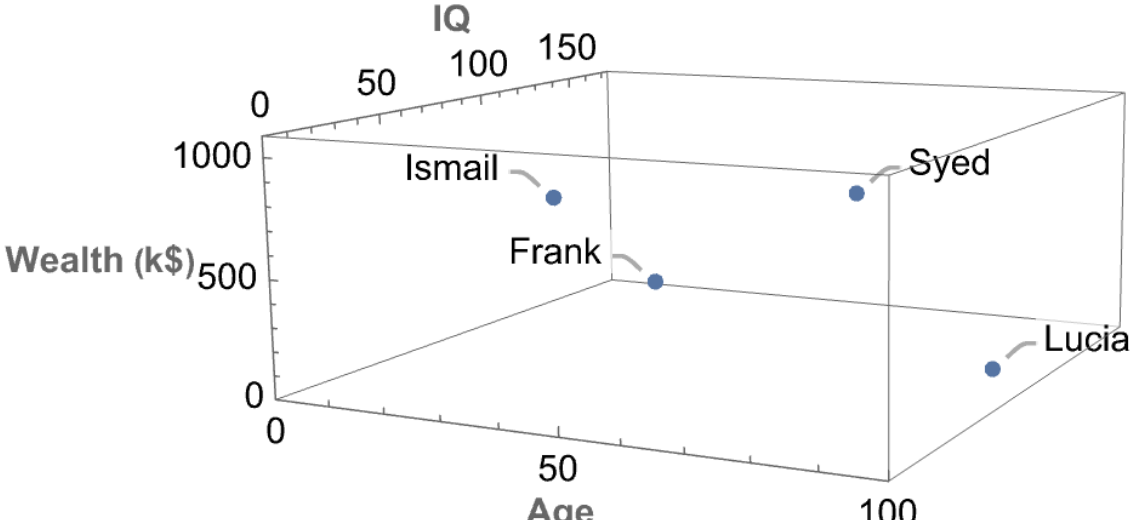


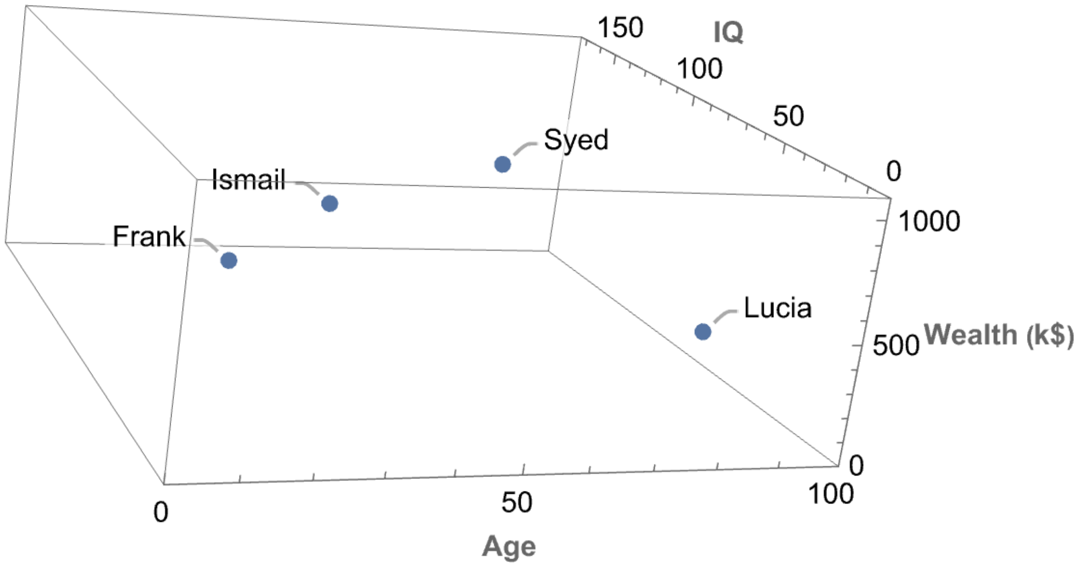
Name	Age	IQ	Wealth (k$)
Ismail	30	56	759
Frank	28	117	217
Lucia	99	71	155
Syed	69	104	733

Row, instance, point, vector, record, entry.

Data lives mathematical spaces

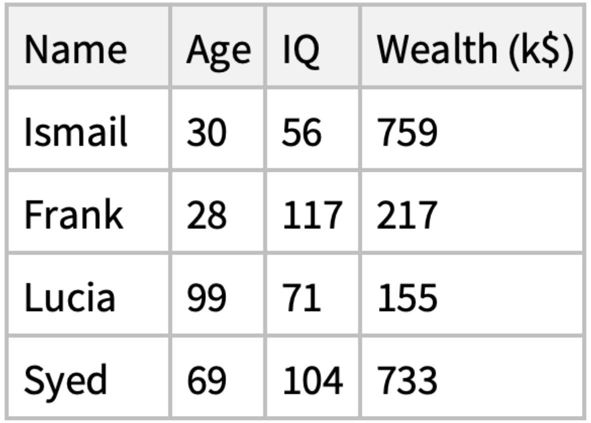
Same box in two different angles:



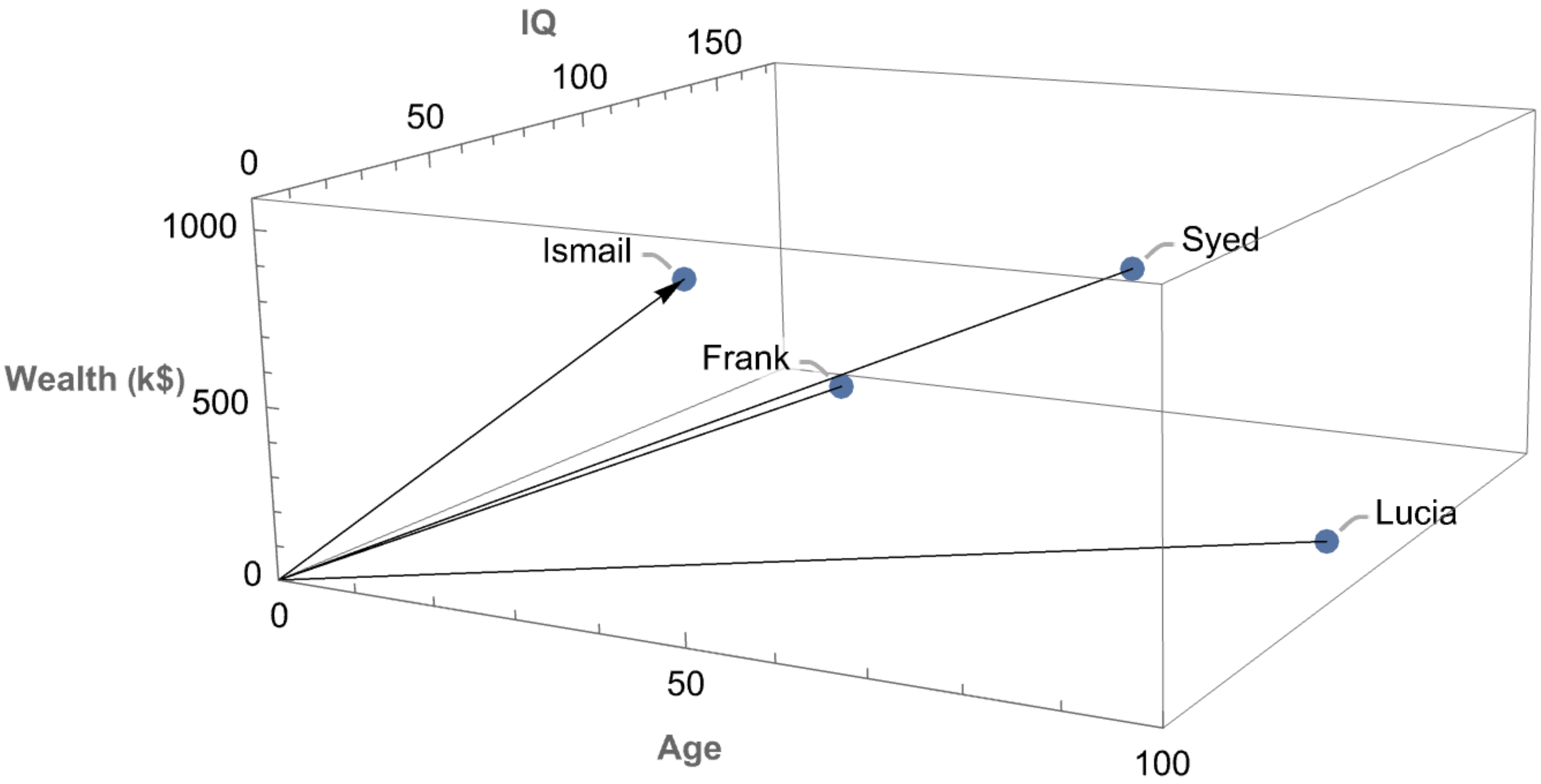


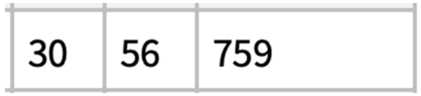
Name	Age	IQ	Wealth (k$)
Ismail	30	56	759
Frank	28	117	217
Lucia	99	71	155
Syed	69	104	733

Tabular Data



Name	Age	IQ	Wealth (k$)
Ismail	30	56	759
Frank	28	117	217
Lucia	99	71	155
Syed	69	104	733

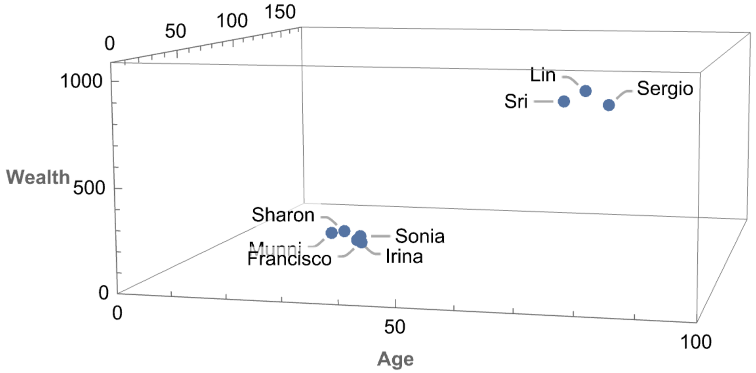


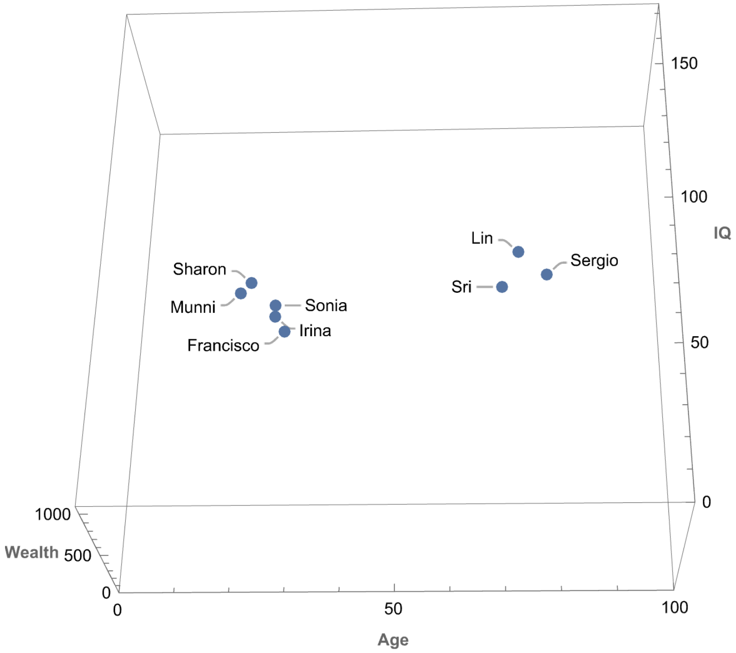


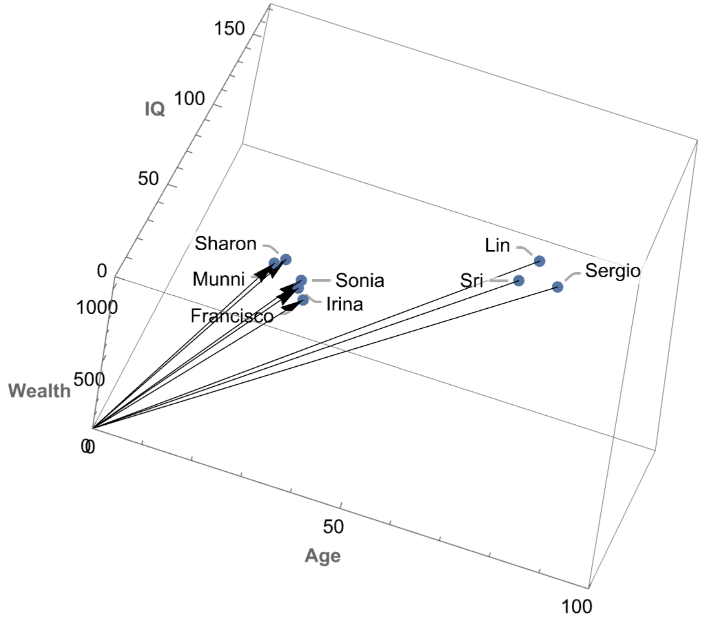
30	56	759

is one data “point” or “vector” in the 3D space where all data points live. In computer, this data point/vector is shown as [30, 56, 759]. You can consider these three numbers as the coordinates in the 3D space of “Age”, “IQ” and “Wealth (k$)”, or you can show this data as a vector from origin [0, 0, 0] to [30, 56, 759].

Tabular Data







- *What story is this plot conveying?*

There are clearly two groups (technically called “clusters”).

Group 1: Old, rich people. Group 2: Young, middle-class people.

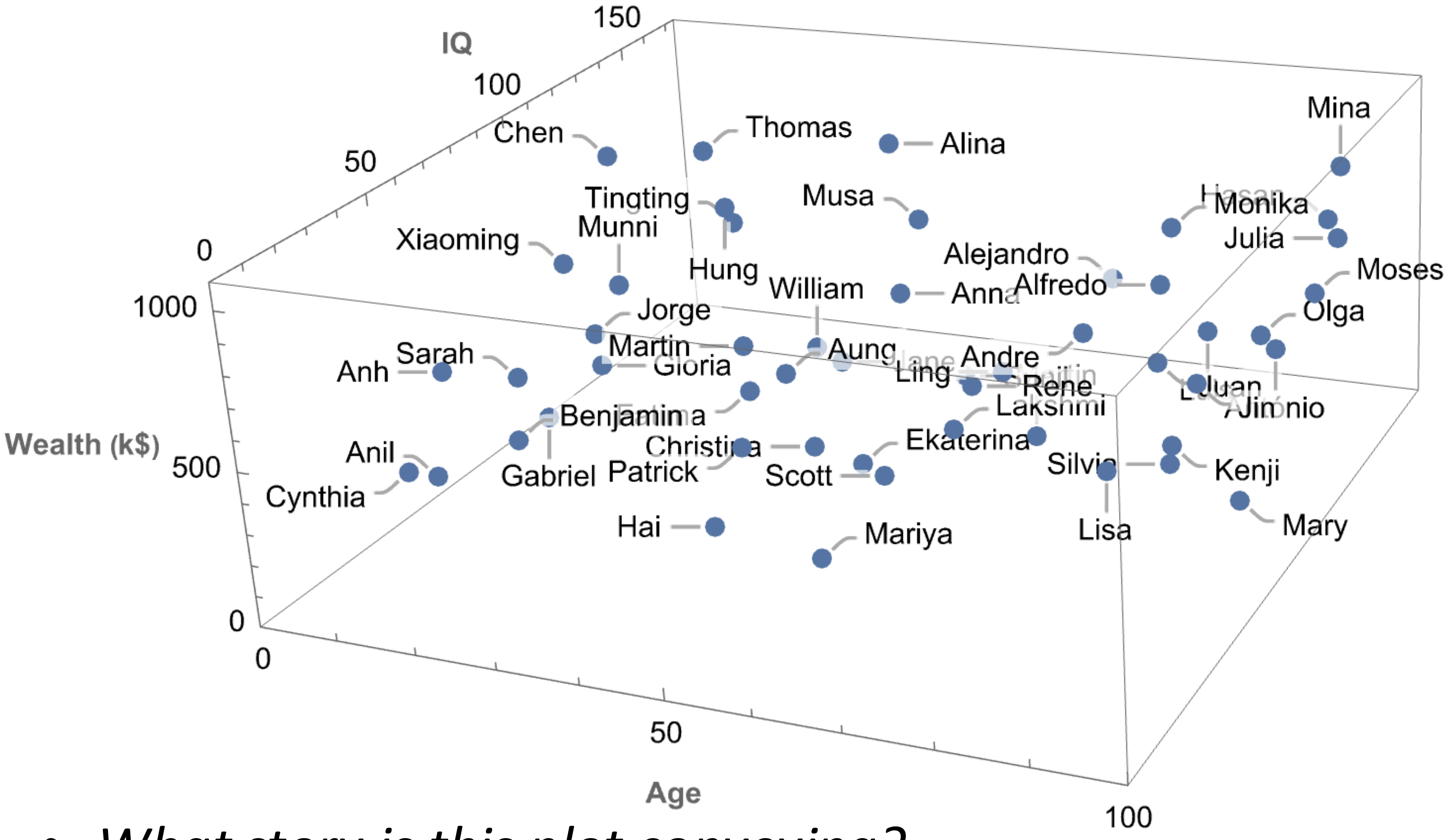
Both group have similar IQs.

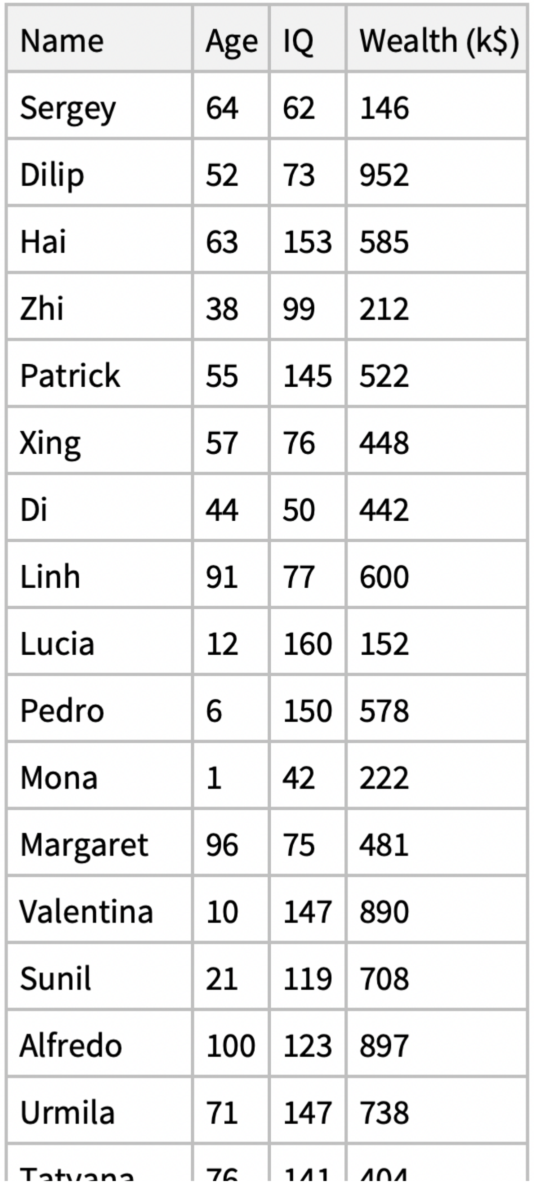
- *Can you see the clusters in the data table?*

Maybe yes for this data, but in general you won’t be able to. That’s why plotting/visualization is an essential part of data science!

Name	Age	IQ	Wealth (k$)
Francisco	28	92	16
Irina	26	98	10
Sergio	78	84	800
Munni	19	107	17
Lin	73	90	860
Sharon	21	111	18
Sonia	26	102	20
Sri	70	79	822

Tabular Data





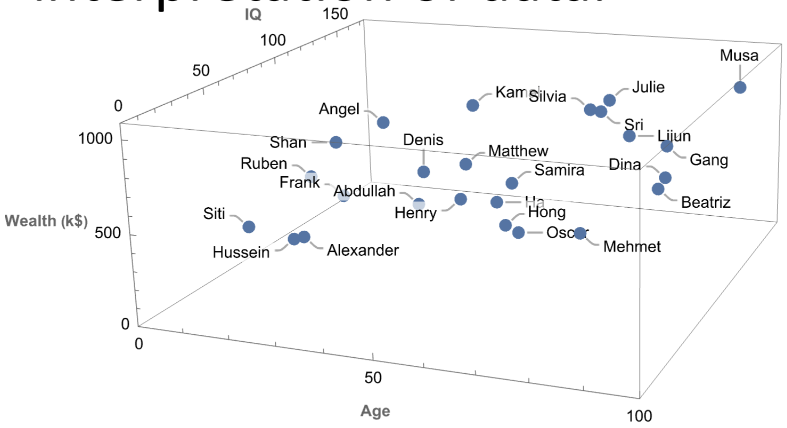
Name	Age	IQ	Wealth (k$)
Sergey	64	62	146
Dilip	52	73	952
Hai	63	153	585
Zhi	38	99	212
Patrick	55	145	522
Xing	57	76	448
Di	44	50	442
Linh	91	77	600
Lucia	12	160	152
Pedro	6	150	578
Mona	1	42	222
Margaret	96	75	481
Valentina	10	147	890
Sunil	21	119	708
Alfredo	100	123	897
Urmila	71	147	738
Tatyana	76	141	404

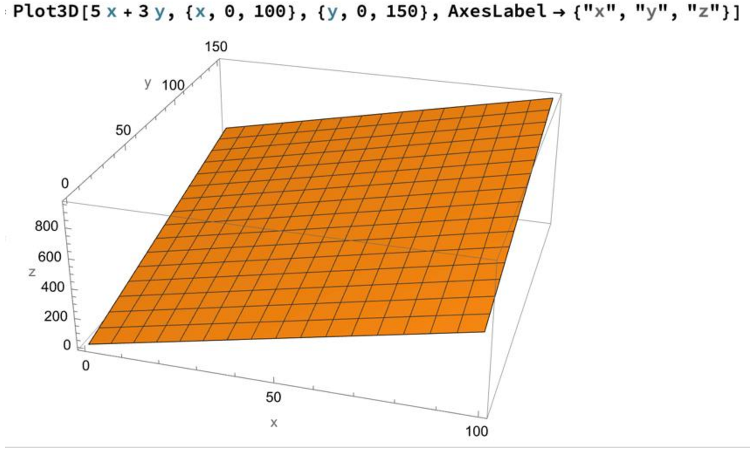
- *What story is this plot conveying?*

Well, this dataset looks like randomly distributed in all directions without clear clusters or correlations or trends. It looks random to eye; maybe there are structures we can’t see by eye.

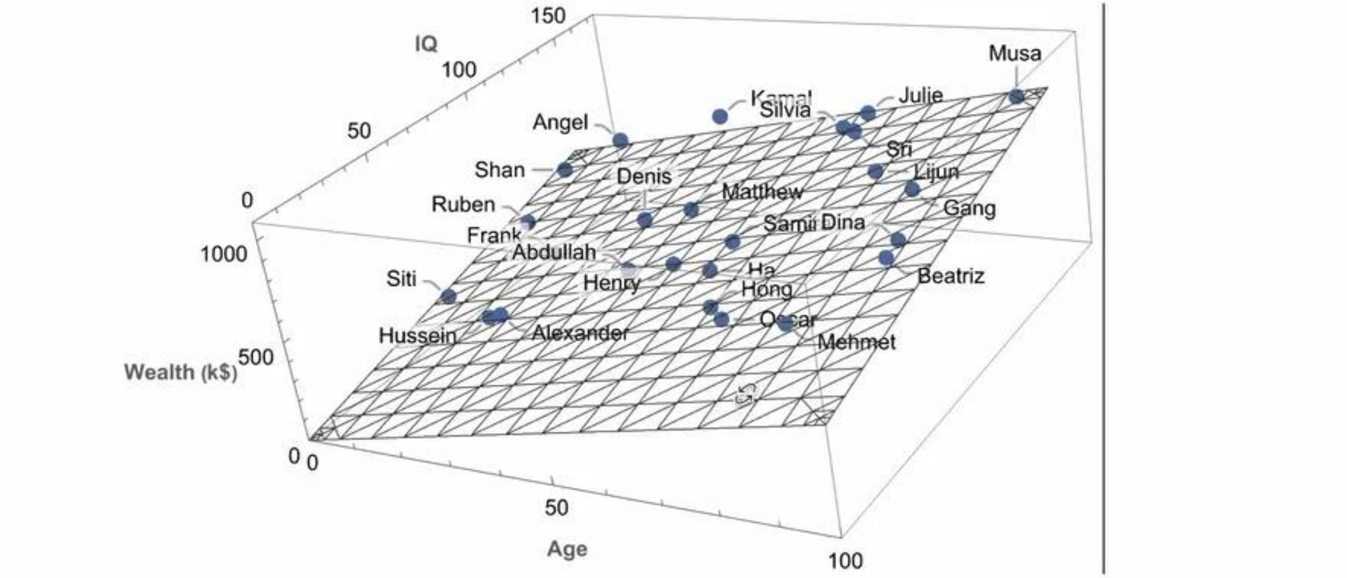
Tabular Data

- Data itself may be 3D (=has 3 columns) but it maybe living on a 2D (plane), 1D (line) or 0D (point) subspace . The data below is actually living on a two dimensional plane. The coordinates of this plane is a mixture of Age, IQ and Wealth. In the following weeks we’ll do “dimensional reduction” to get rid of extra dimensions for easier analysis and interpretation of data.





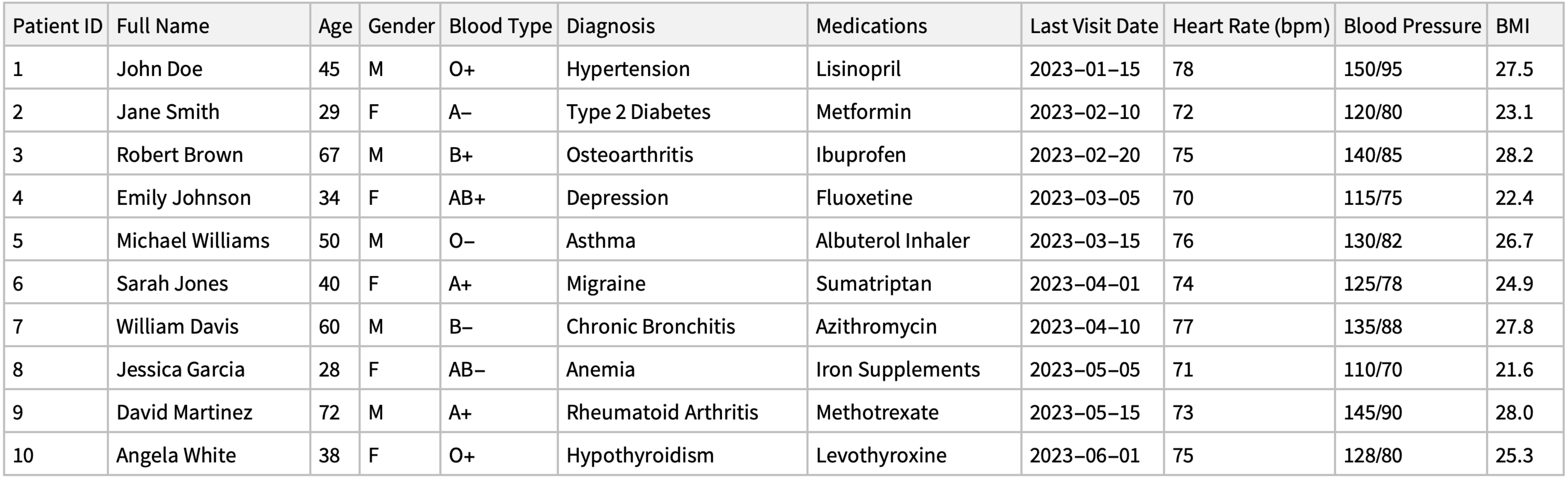
```
Plot3D[5 x + 3 y, {x, 0, 100}, {y, 0, 150}, AxesLabel → {"x", "y", "z"}]
```



Tabular Data

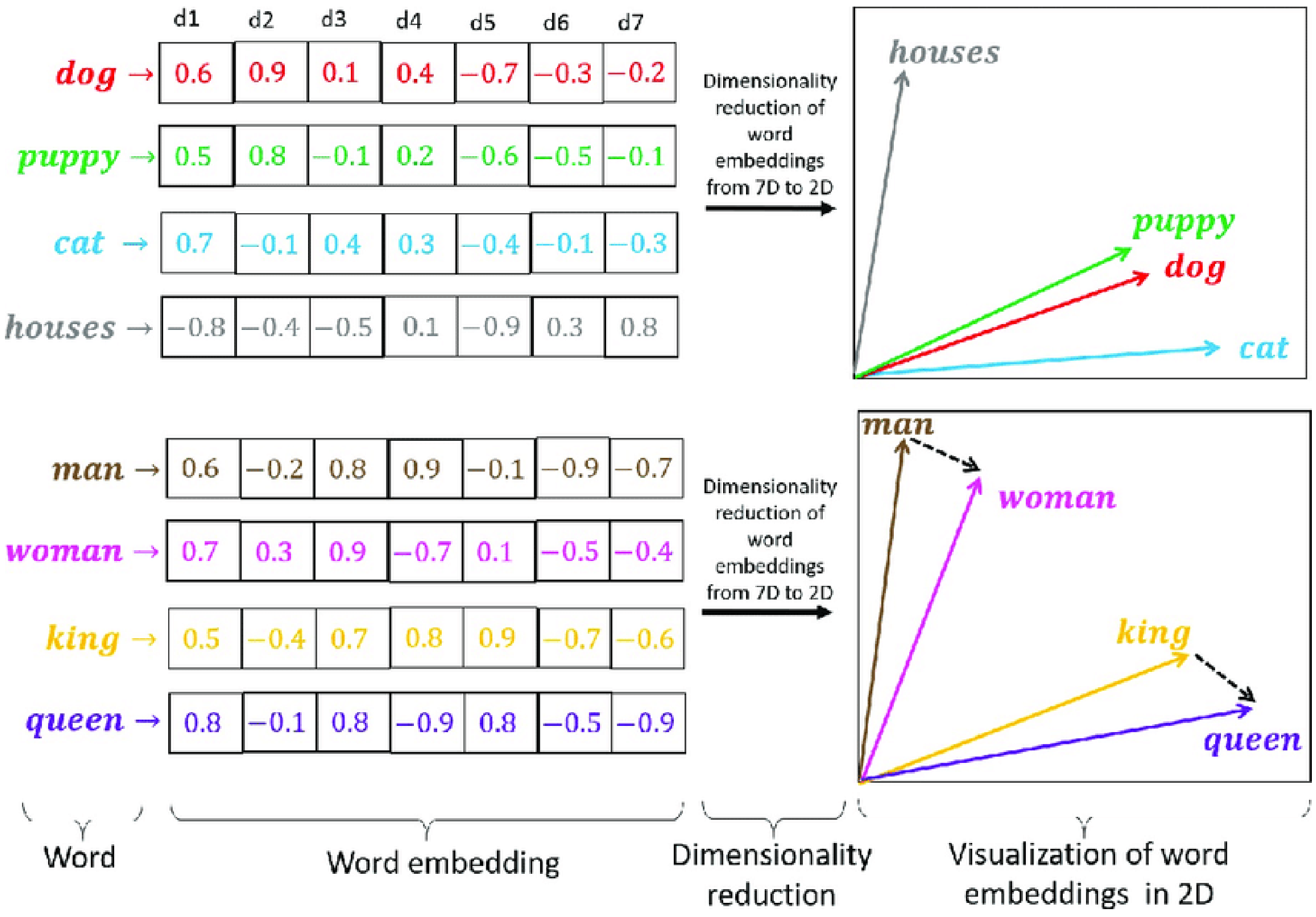
- Real world data has more than 3 dimensions. Human brain cannot imagine spaces with more than 3 dimensions.

- Sample high dimensional EHR (electronic health records) dataset:



Patient ID	Full Name	Age	Gender	Blood Type	Diagnosis	Medications	Last Visit Date	Heart Rate (bpm)	Blood Pressure	BMI
1	John Doe	45	M	O+	Hypertension	Lisinopril	2023–01–15	78	150/95	27.5
2	Jane Smith	29	F	A–	Type 2 Diabetes	Metformin	2023–02–10	72	120/80	23.1
3	Robert Brown	67	M	B+	Osteoarthritis	Ibuprofen	2023–02–20	75	140/85	28.2
4	Emily Johnson	34	F	AB+	Depression	Fluoxetine	2023–03–05	70	115/75	22.4
5	Michael Williams	50	M	O–	Asthma	Albuterol Inhaler	2023–03–15	76	130/82	26.7
6	Sarah Jones	40	F	A+	Migraine	Sumatriptan	2023–04–01	74	125/78	24.9
7	William Davis	60	M	B–	Chronic Bronchitis	Azithromycin	2023–04–10	77	135/88	27.8
8	Jessica Garcia	28	F	AB–	Anemia	Iron Supplements	2023–05–05	71	110/70	21.6
9	David Martinez	72	M	A+	Rheumatoid Arthritis	Methotrexate	2023–05–15	73	145/90	28.0
10	Angela White	38	F	O+	Hypothyroidism	Levothyroxine	2023–06–01	75	128/80	25.3

Text data



	d1	d2	d3	d4	d5	d6	d7
dog →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
puppy →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
cat →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8

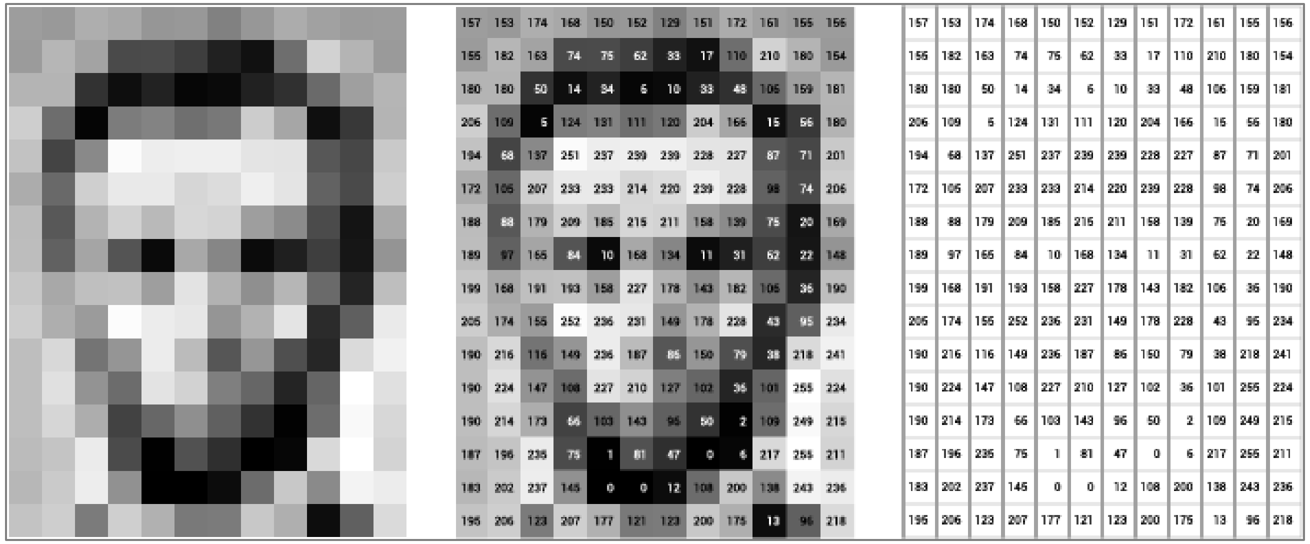
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Dimensionality reduction of word embeddings from 7D to 2D
houses
puppy
dog
cat
Dimensionality reduction of word embeddings from 7D to 2D
man
woman
king
queen
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

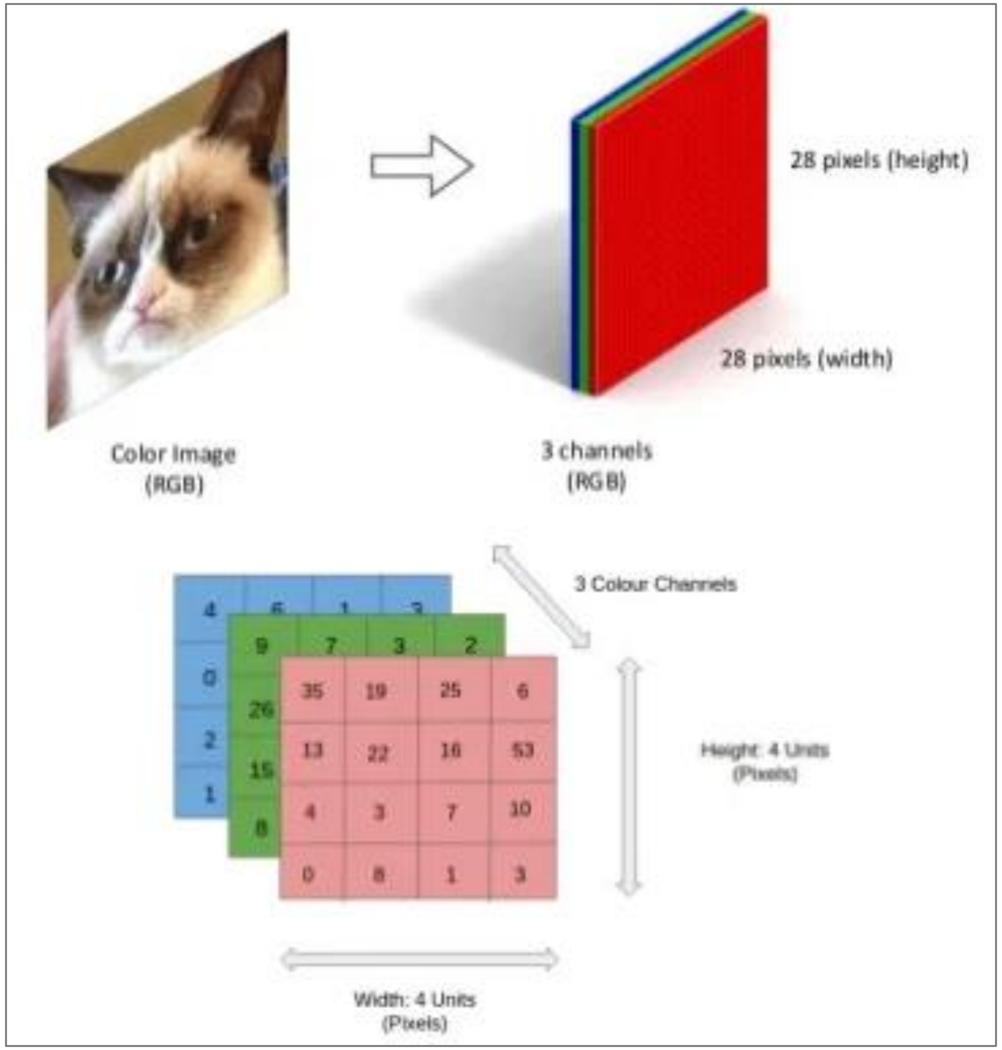
Image data

RGB colored images are rank-3 tensors (3D).

Grayscale images are matrices (2D).



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

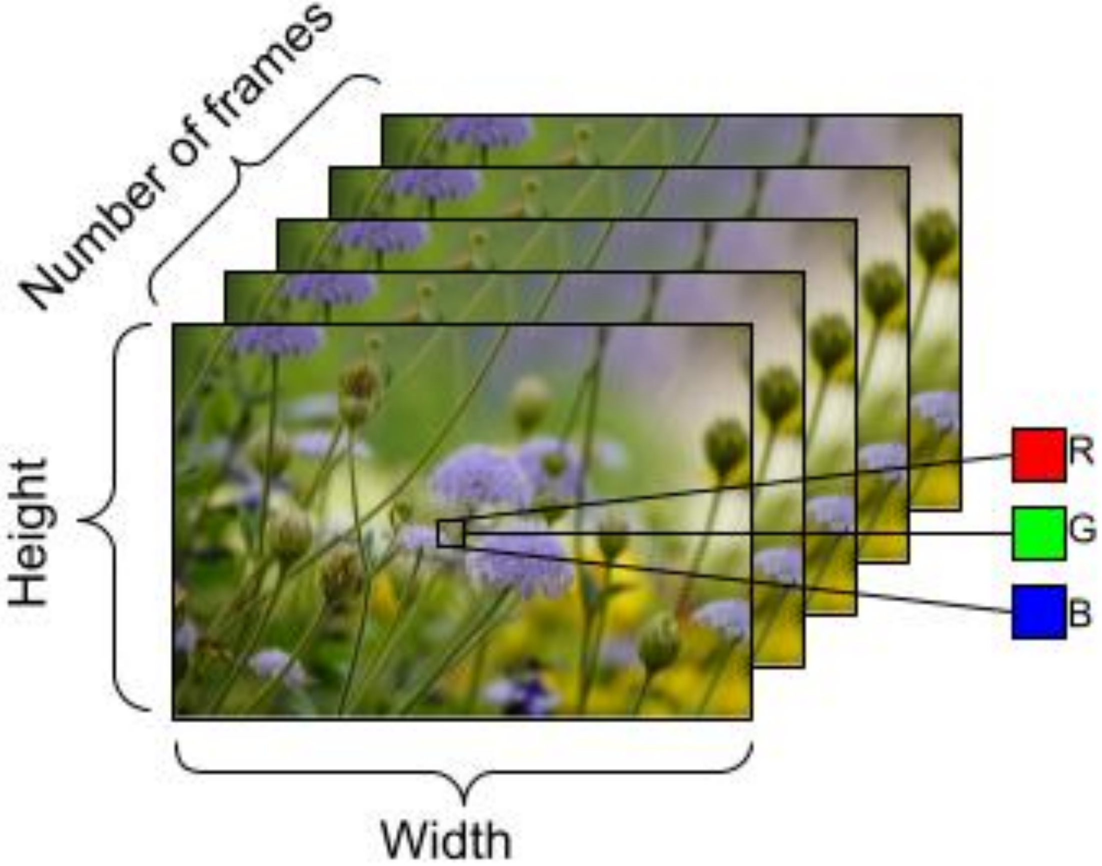


35	19	25	6
13	22	16	53
4	3	7	10
0	8	1	3

Video data

- Video data is a rank-4 tensor (4D) with dimensions:

number of frames (time) x height x width x color channels.


Number of frames
Height
Width
R
G
B