Week 1: DSA (Prof. Lei Chen)

Week 2: DSA

Week 3: DSA – TA presentation

Week 4: IoT

Week 5: IoT

Week 6: IoT – TA presentation

Week 7: AI

Week 8: AI

Week 9: AI – TA presentation

Week 10: CMA

Week 11: CMA

Week 12: CMA – TA presentation

Week 13: (Selected) project presentation

TA presentation

- 30 minutes
- Why he/she chose his/her thrust

INFH5000, Data Science and Analytics (DSA)
An Introduction to Data Science

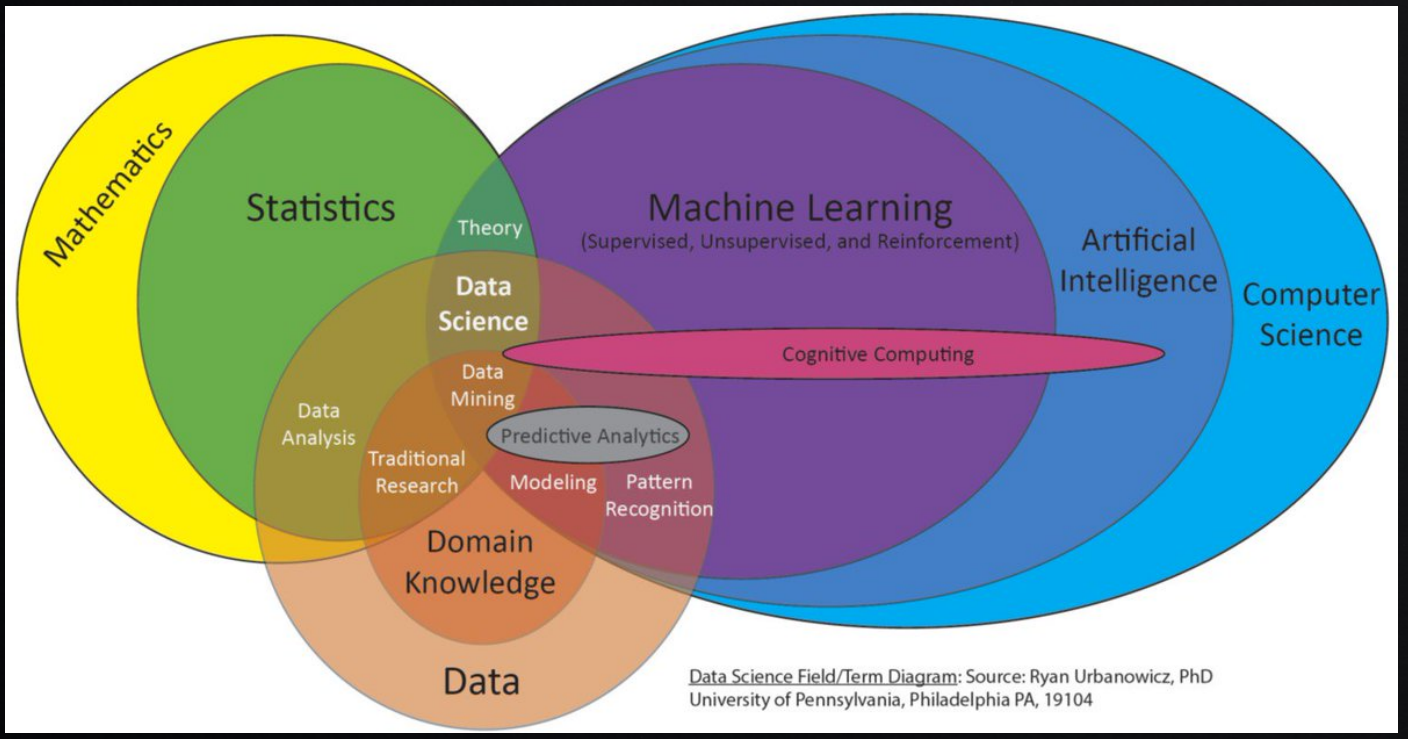
Lecture 1: An overview of data science
Lecture 2: Key Methods
Lecture 3: Data-centric AI

Nan Tang
Data Science and Analytics (DSA), Information Hub, HKUST (GZ)

Data Scientist

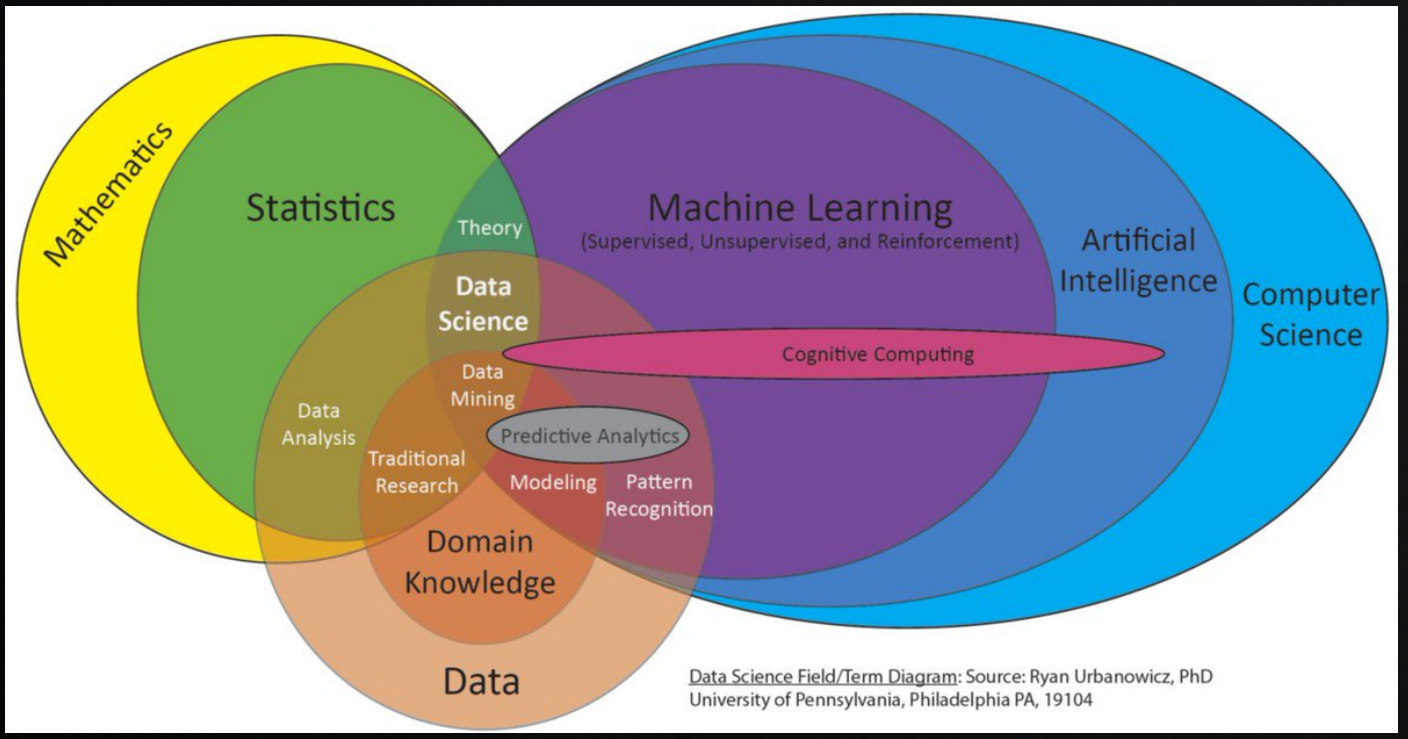
Data Scientist

A person whose expertise spans across a significant number of areas

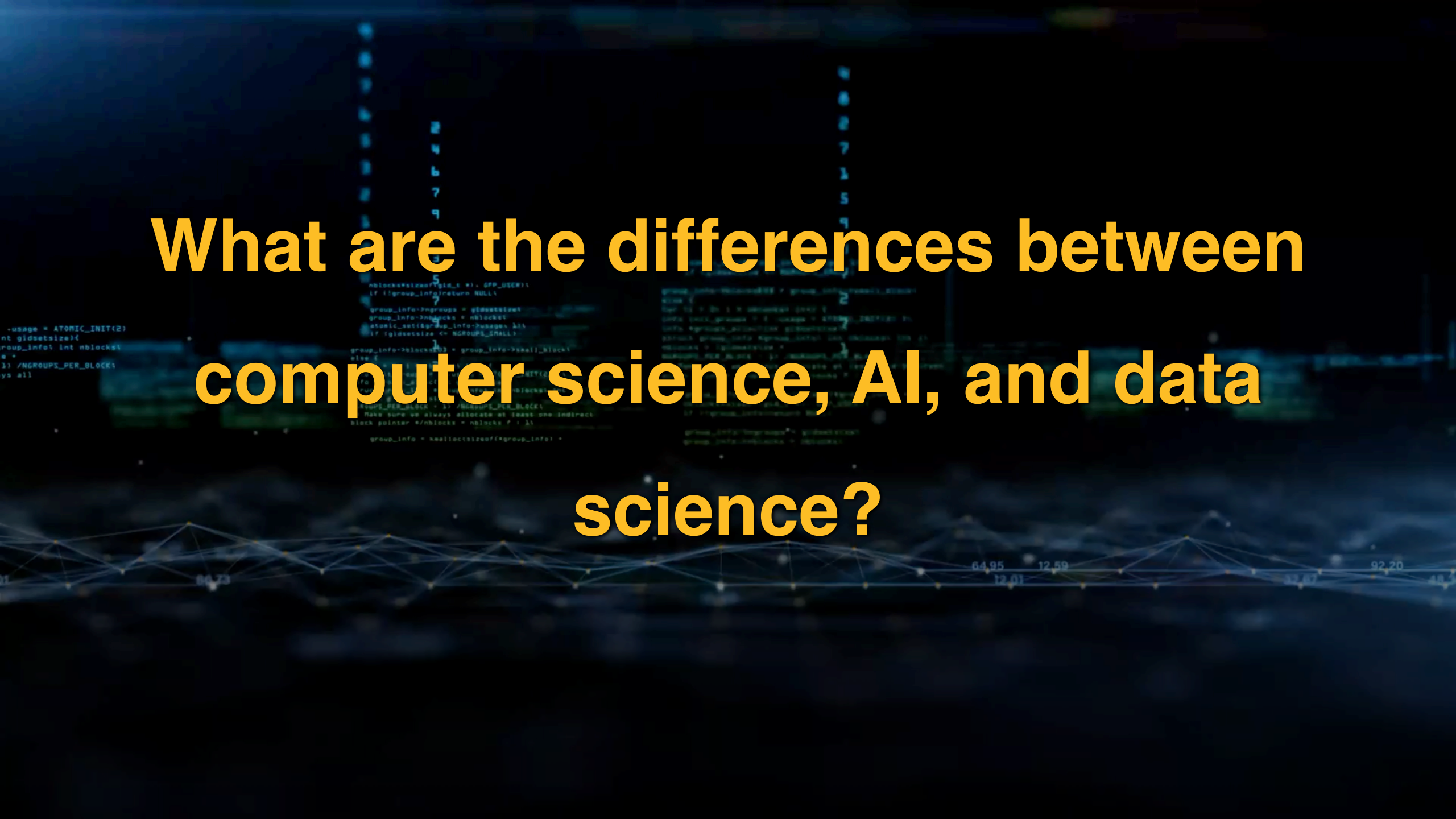


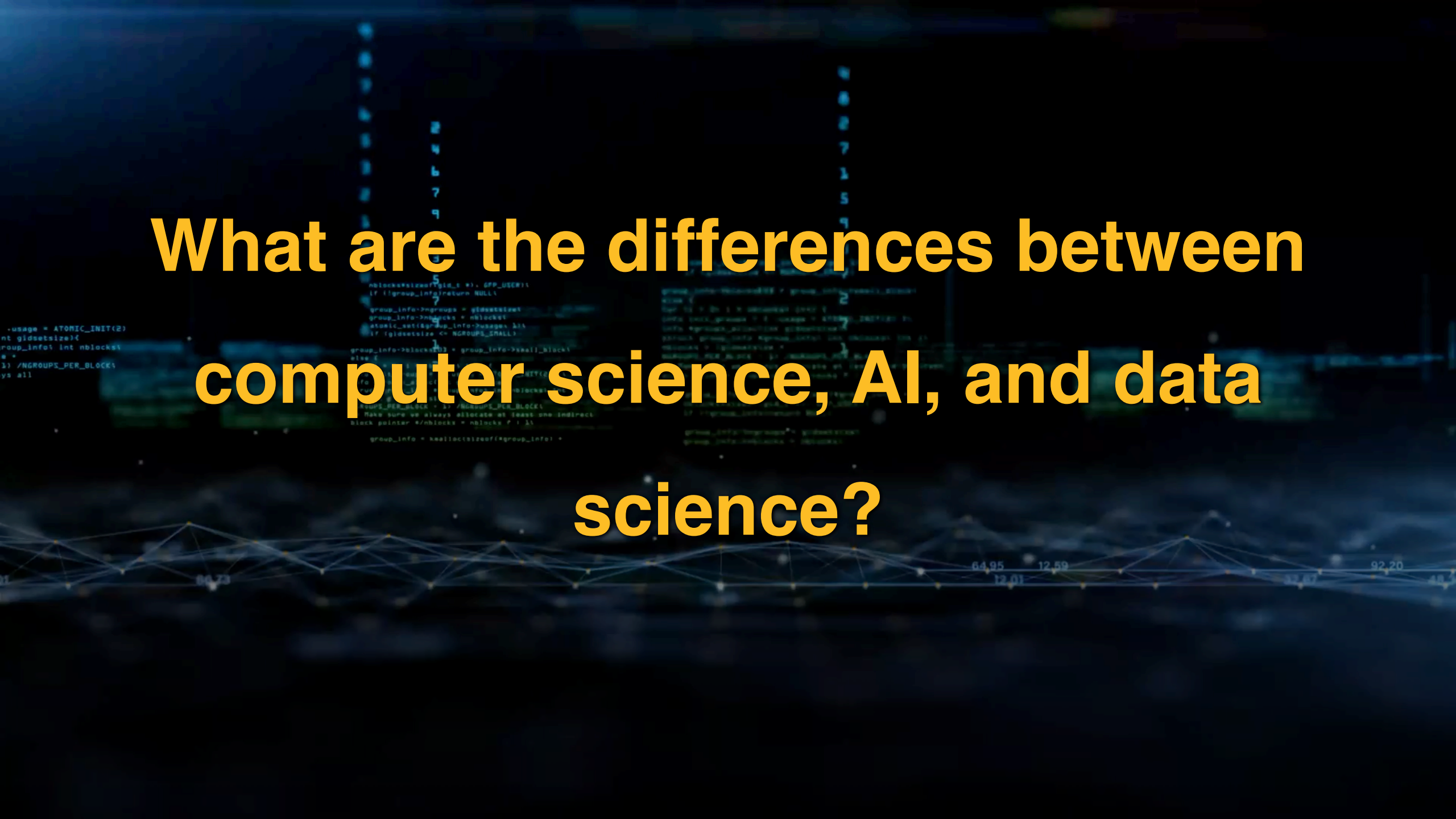
Data Scientist

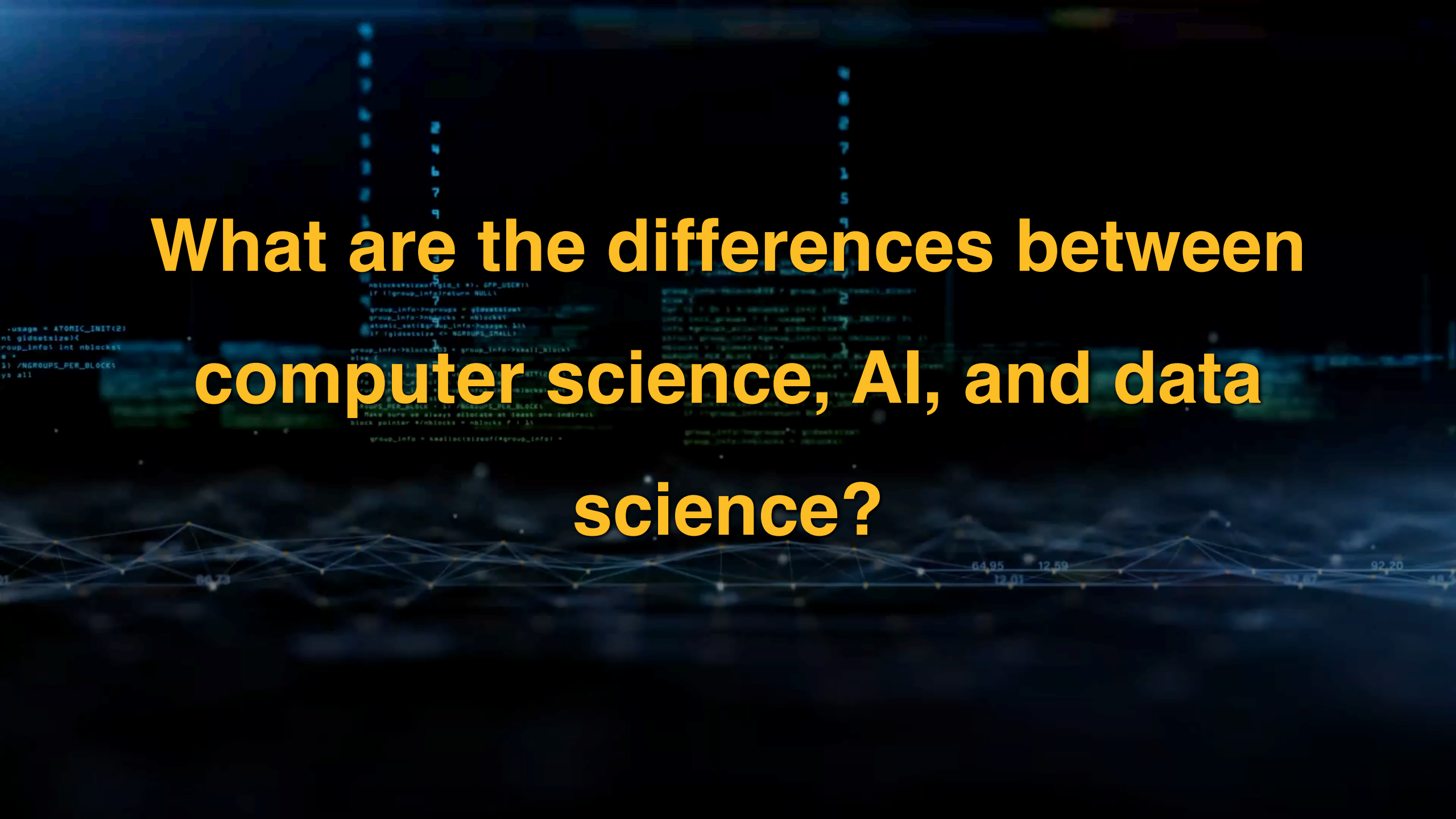
A person whose expertise spans across a significant number of areas

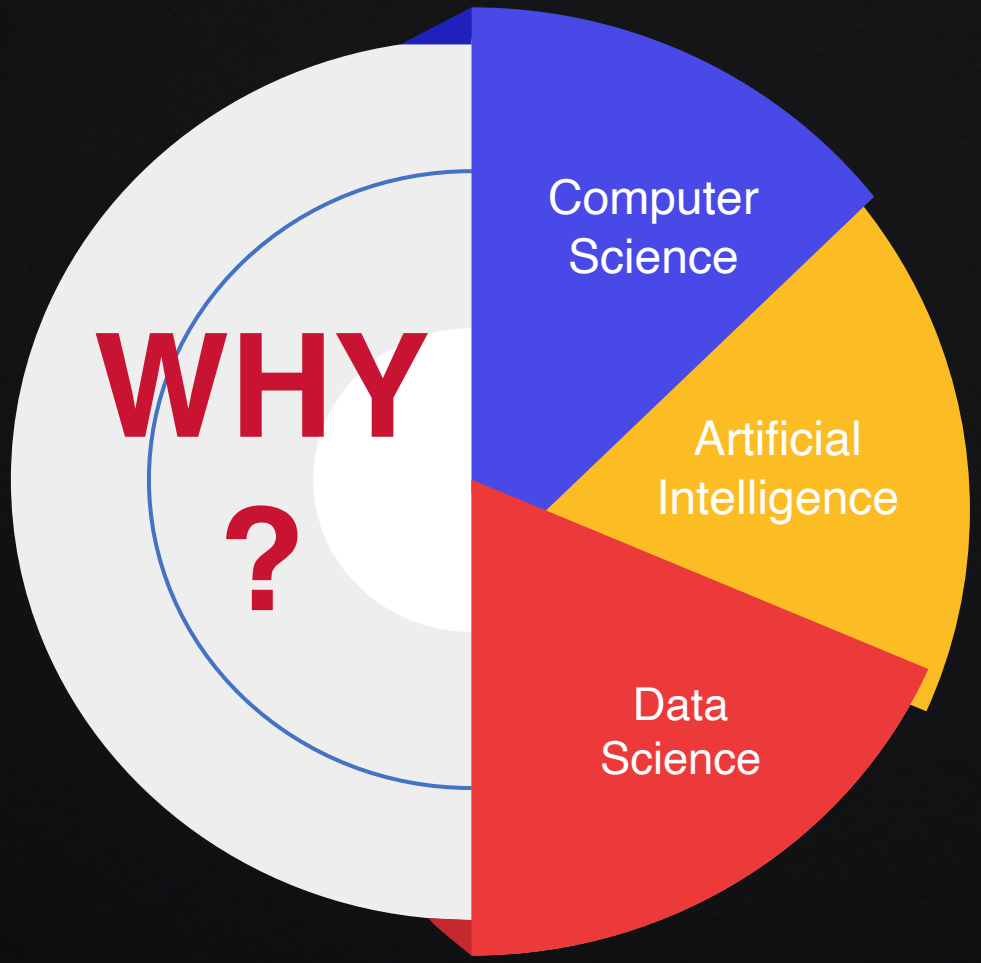


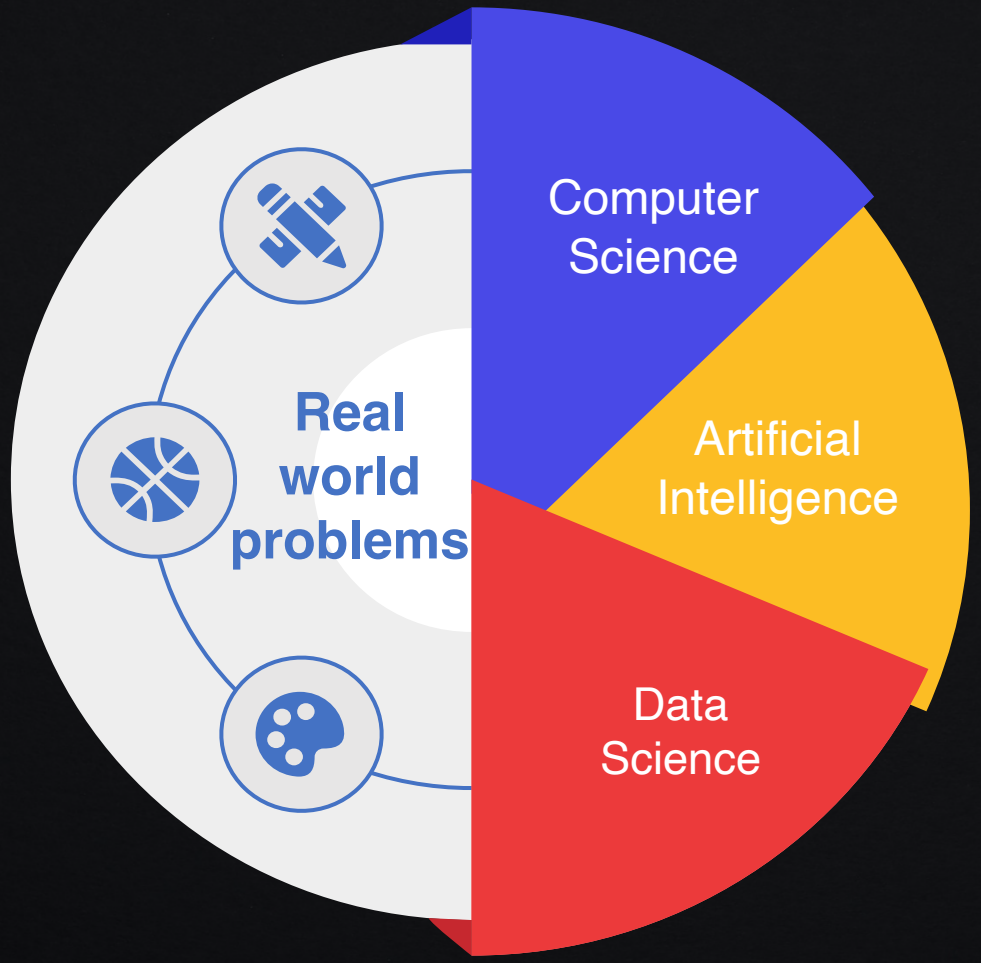
A person who has a curiosity and inquisitive mindset

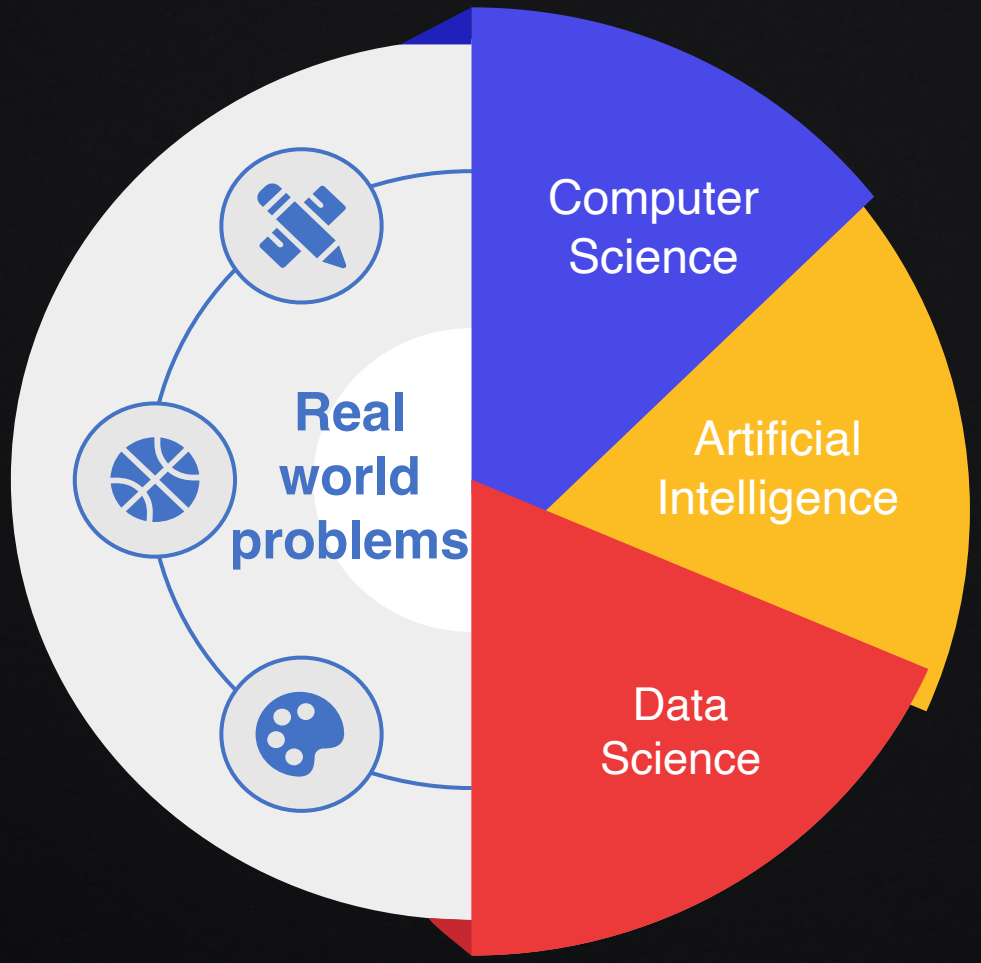

What are the differences between computer science, AI, and data science?


What are the differences between computer science, AI, and data science?

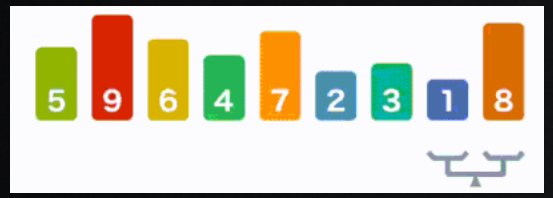

What are the differences between computer science, AI, and data science?

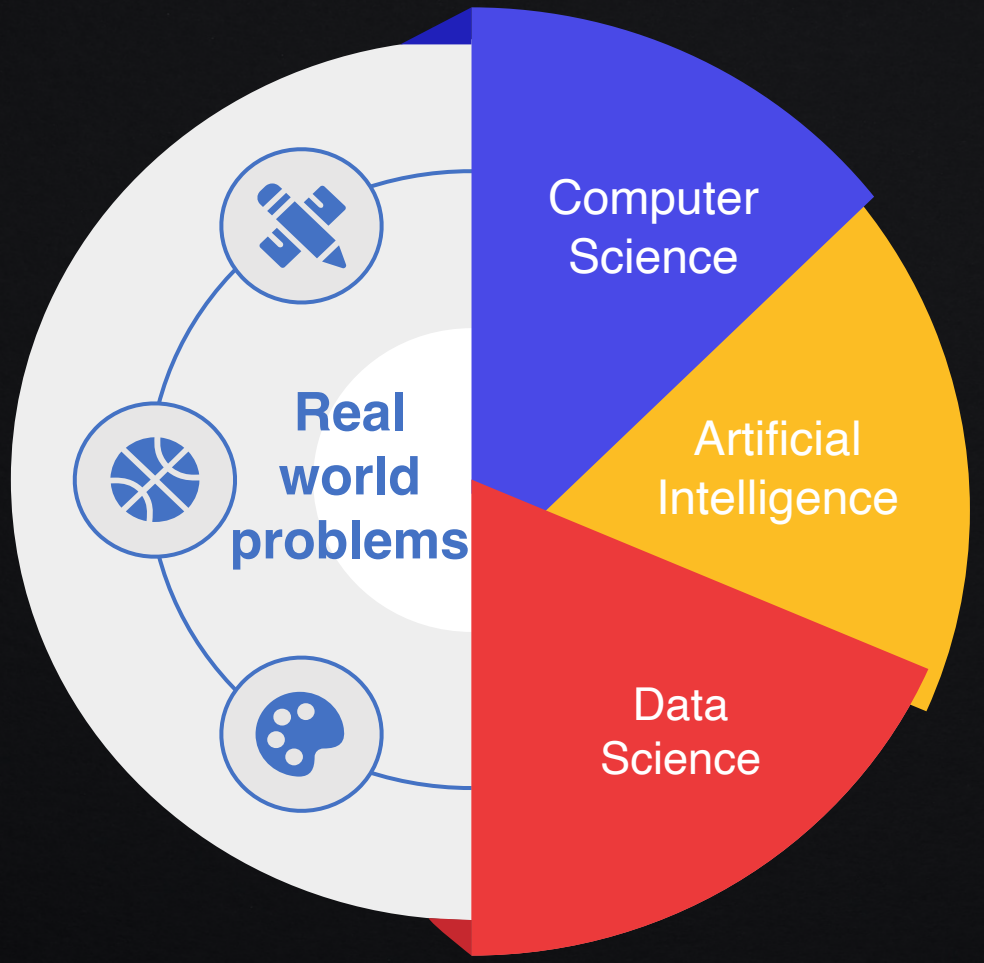

WHY
?
Computer
Science
Artificial
Intelligence
Data
Science


Real world problems
Computer Science
Artificial Intelligence
Data Science

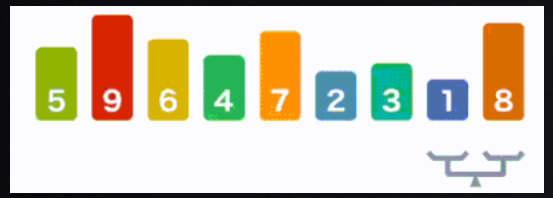

Real world problems
Computer Science
Artificial Intelligence
Data Science

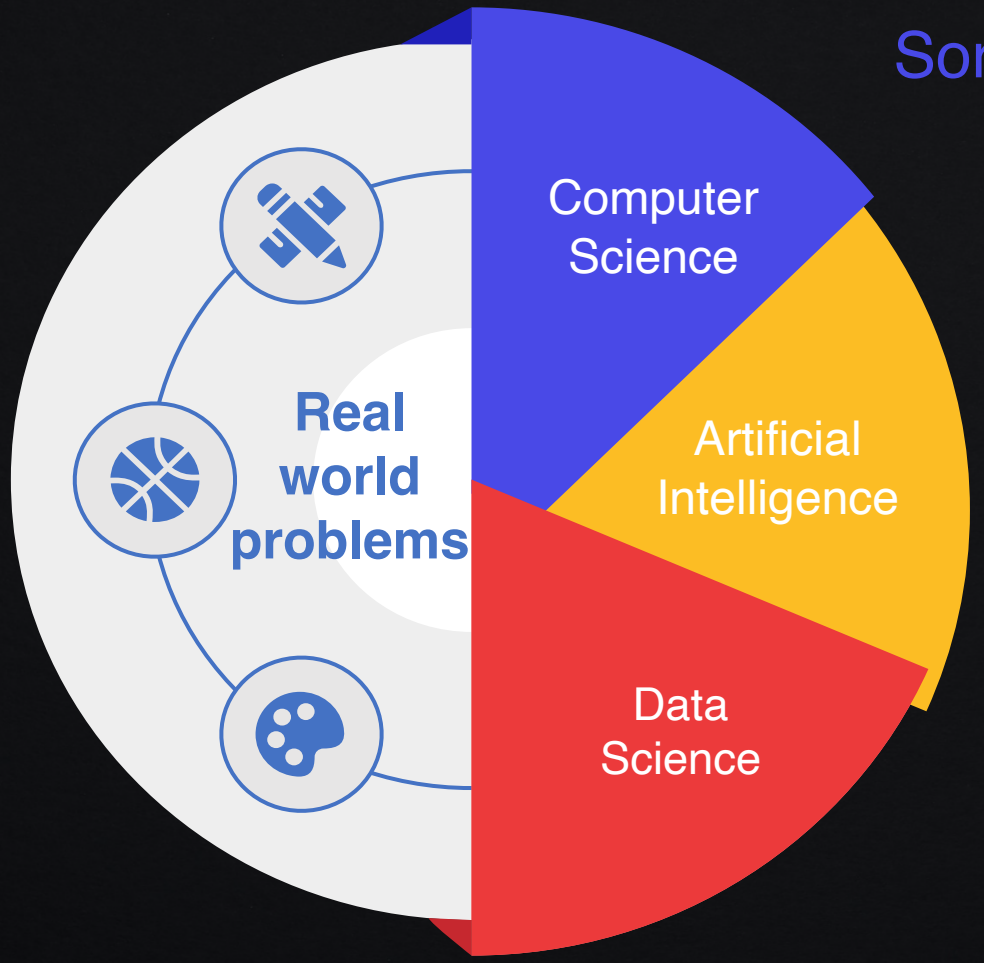
Sorting Problem



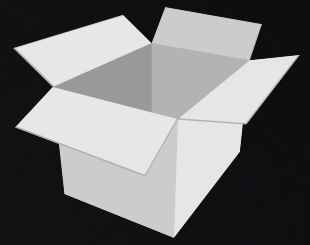


Sorting Problem

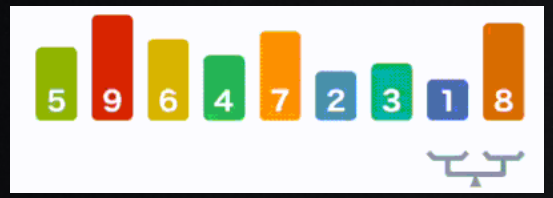




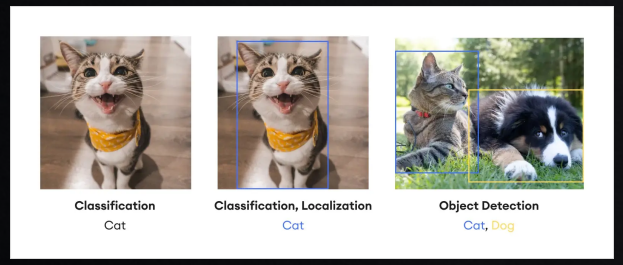
Sorting algorithm

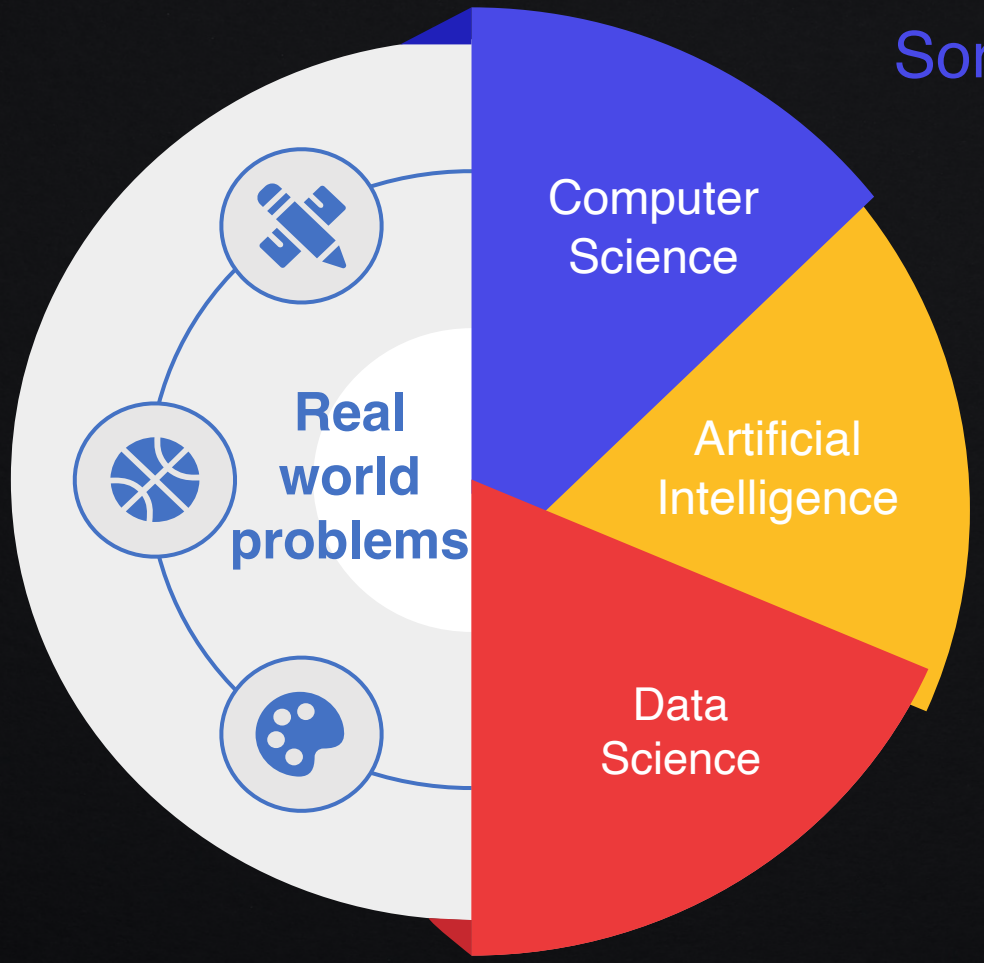


Sorting Problem

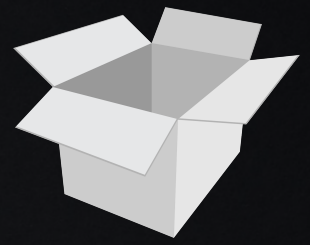


Image

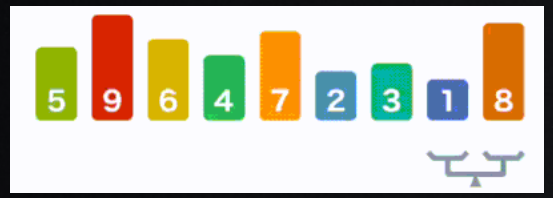




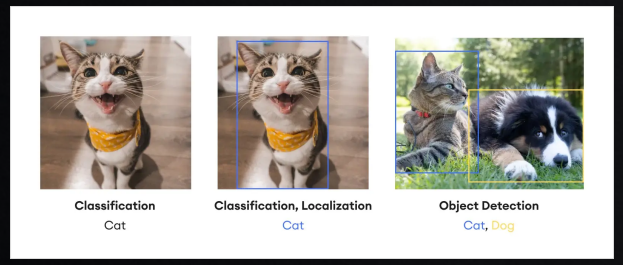
Sorting algorithm

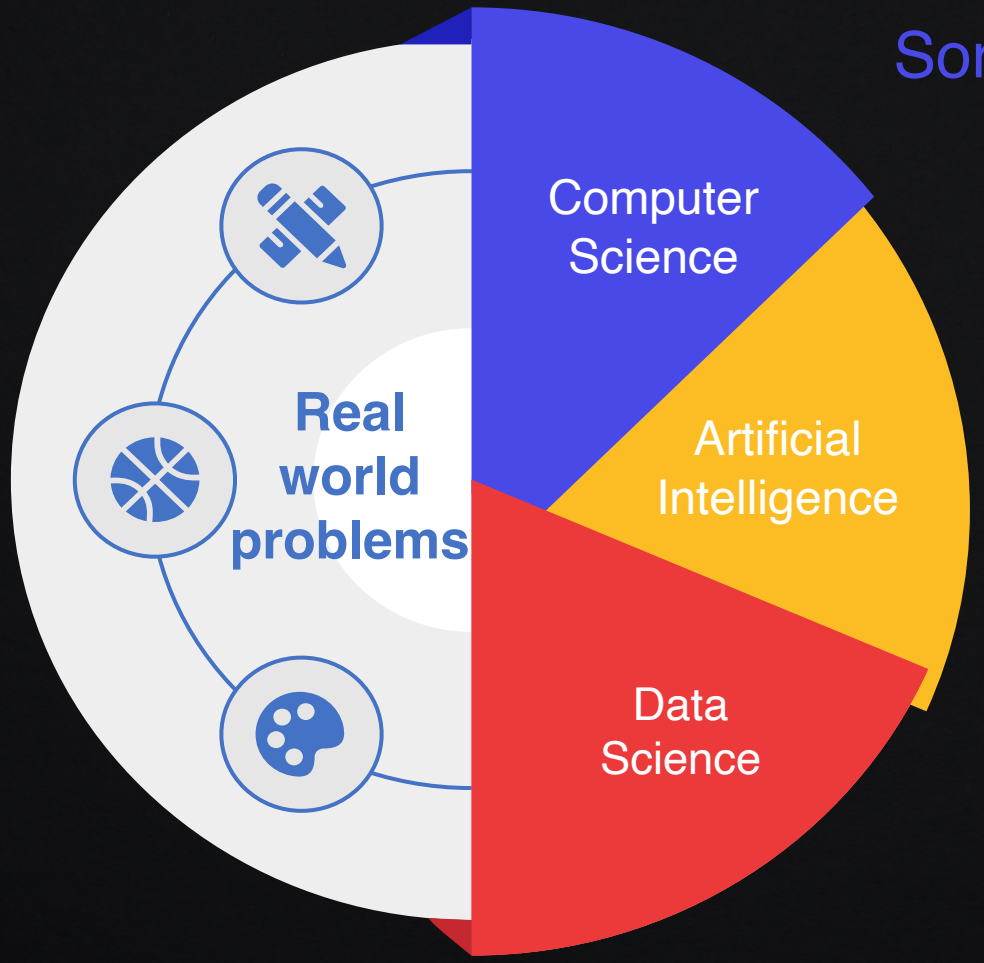


Sorting Problem



Image





Sorting algorithm

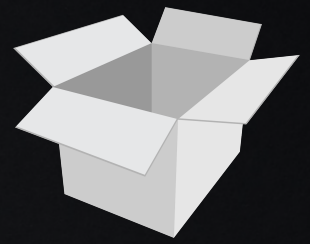
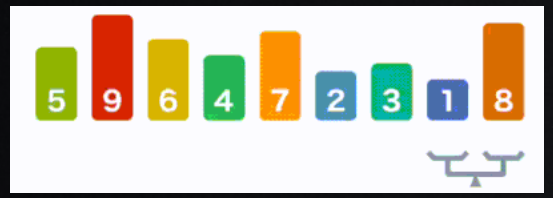
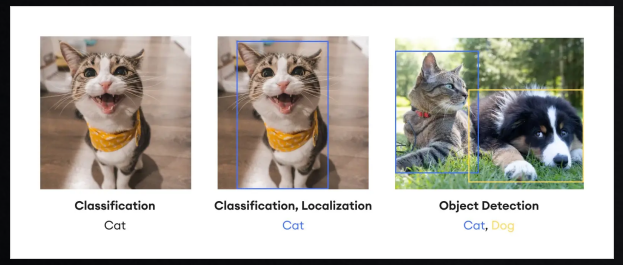


Image classification/ object detection

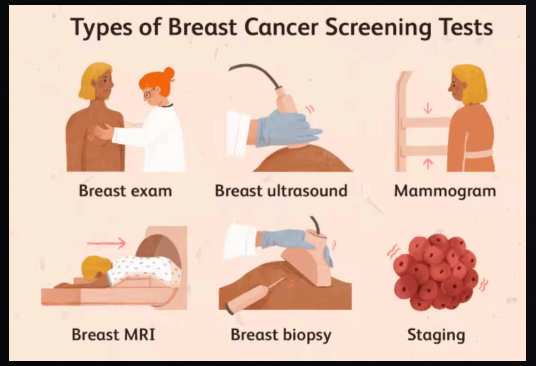
Sorting Problem

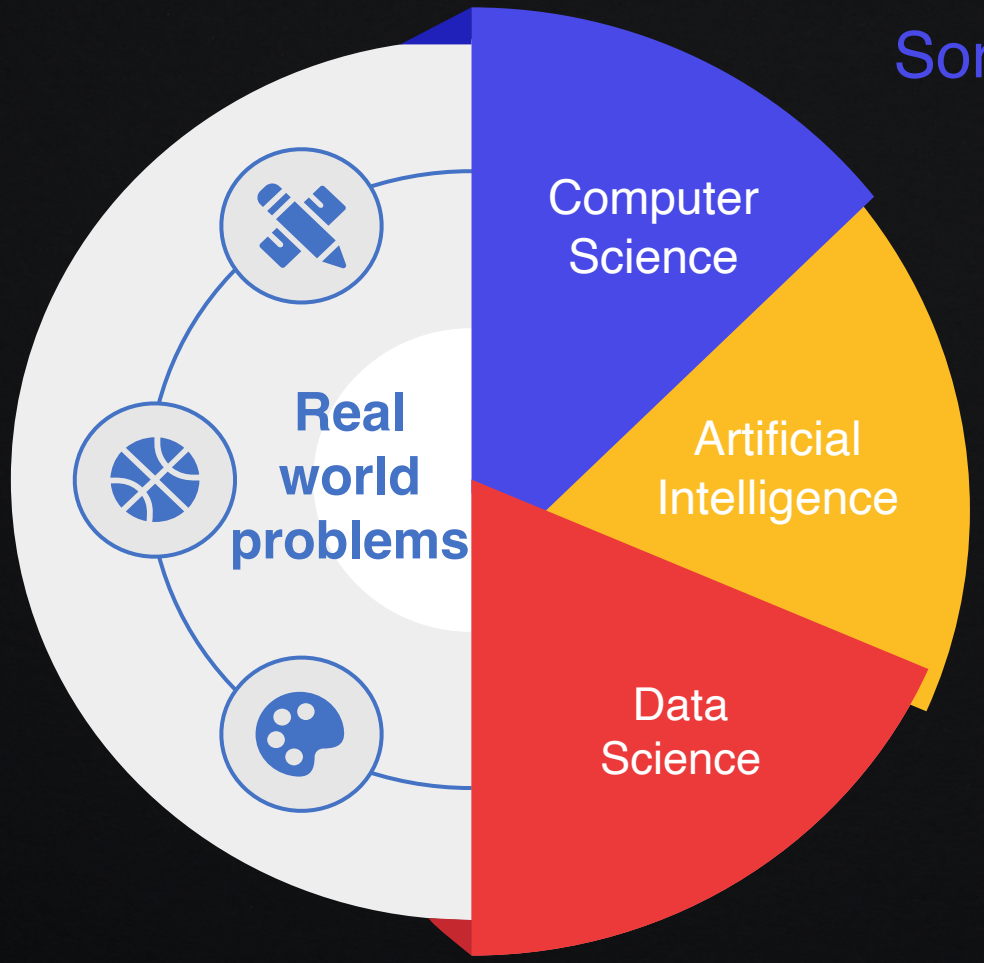


Image



Breast cancer detection





Sorting algorithm

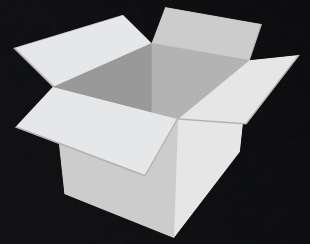
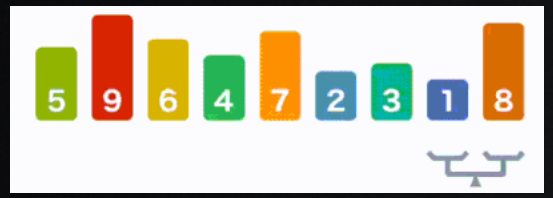
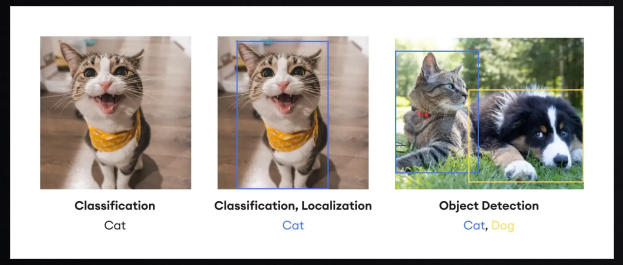


Image classification/ object detection

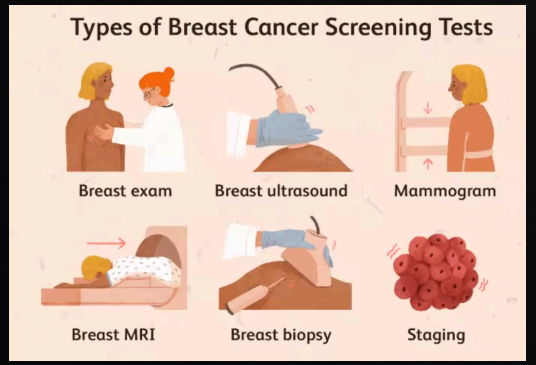
Sorting Problem

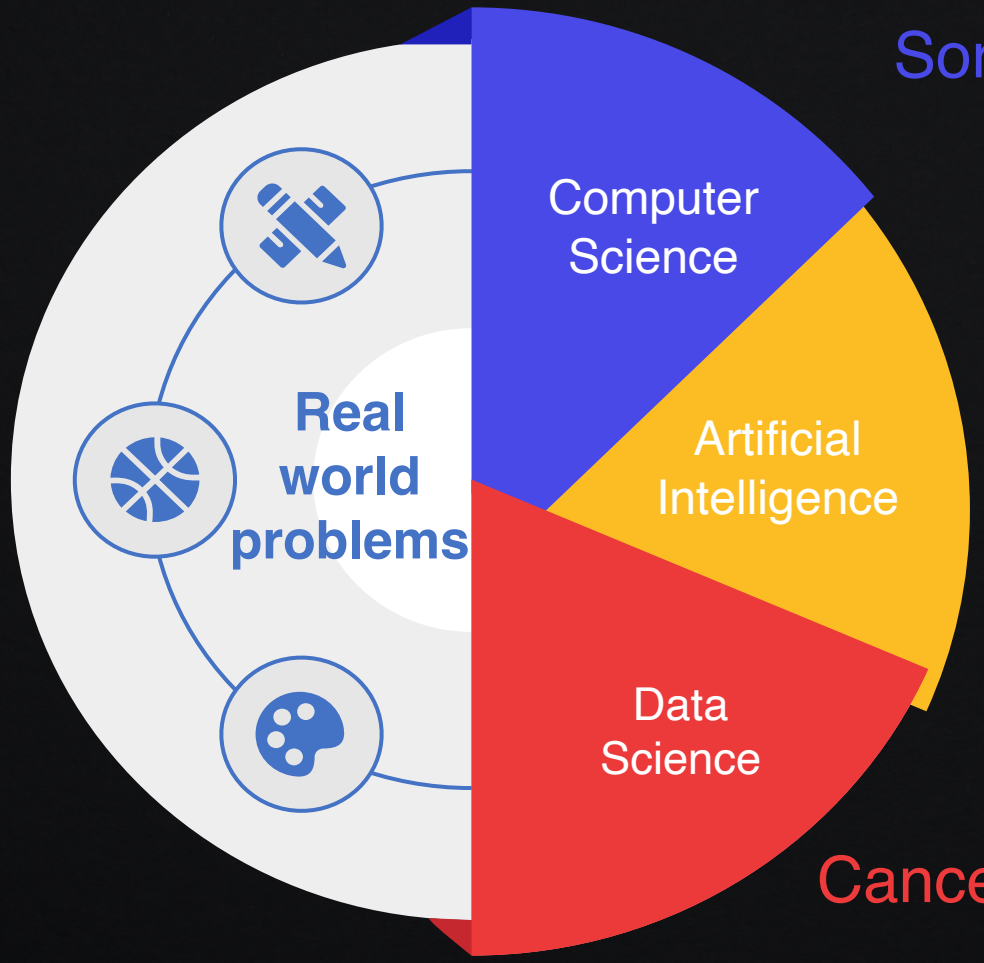


Image



Breast cancer detection





Sorting algorithm

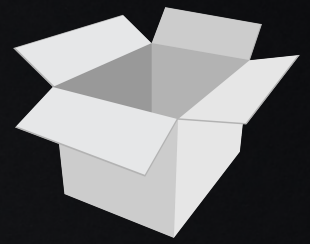
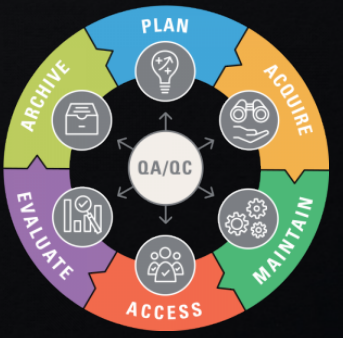
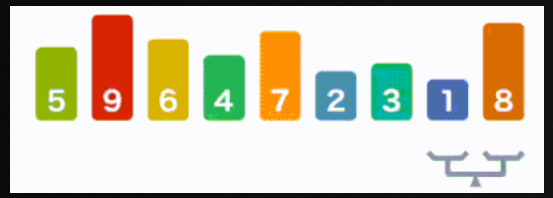


Image classification/ object detection

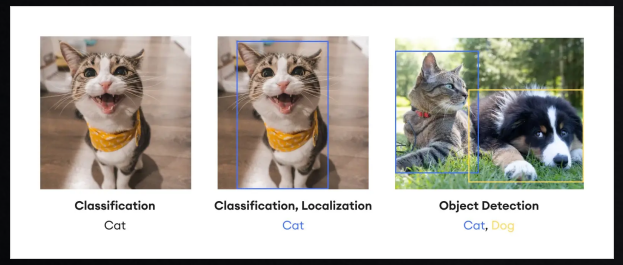
Cancer detection



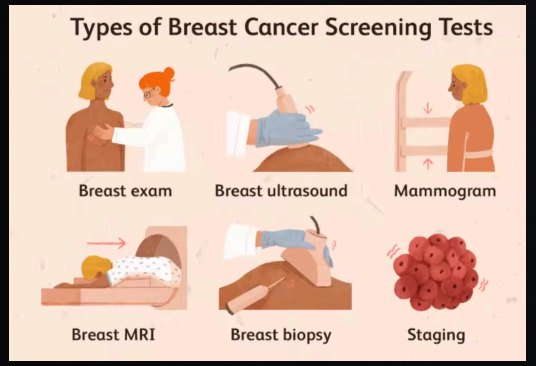
Sorting Problem

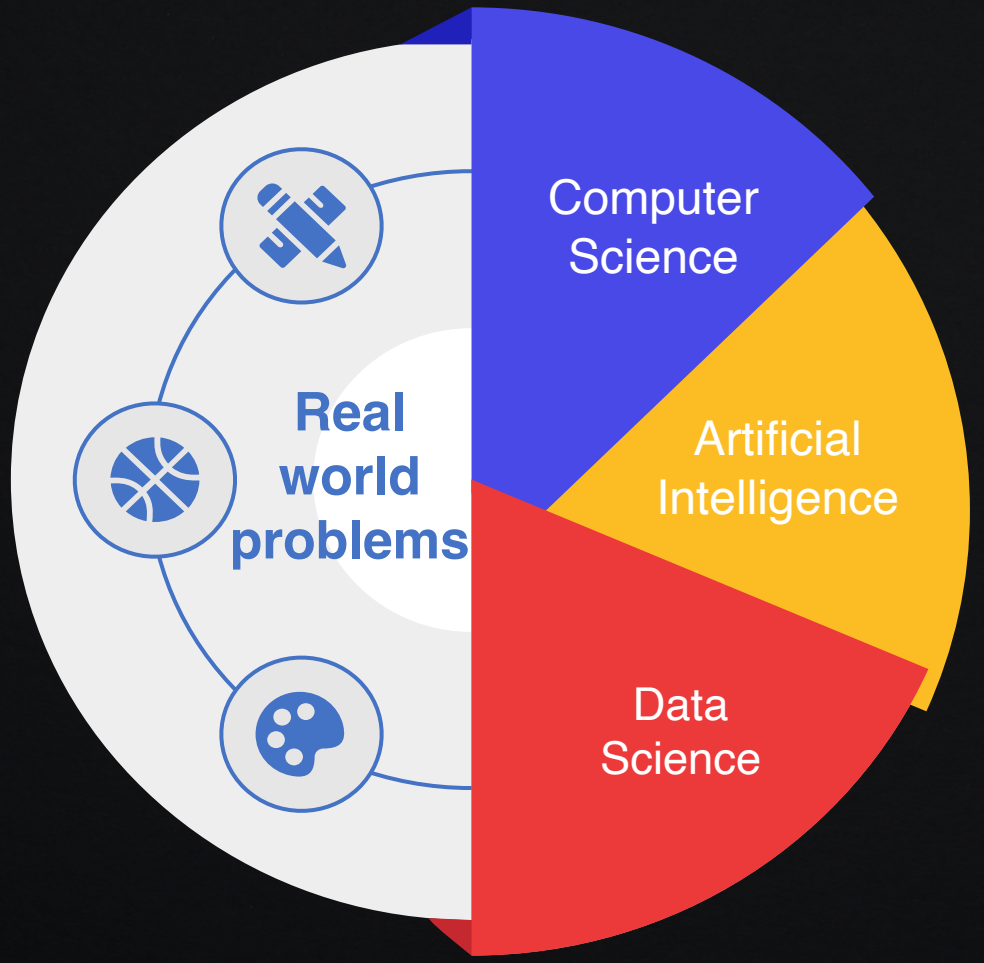


Image

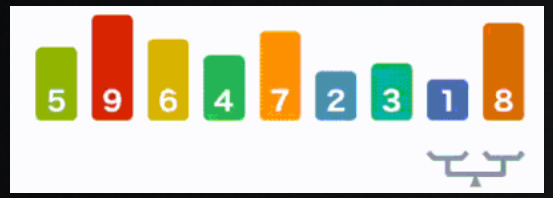


Breast cancer detection

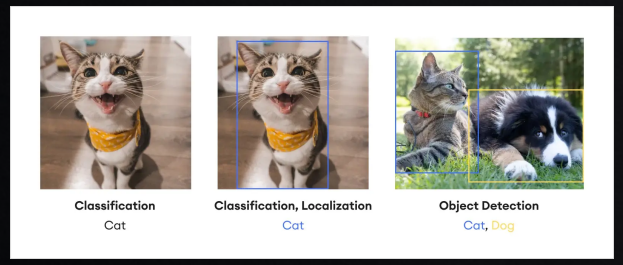




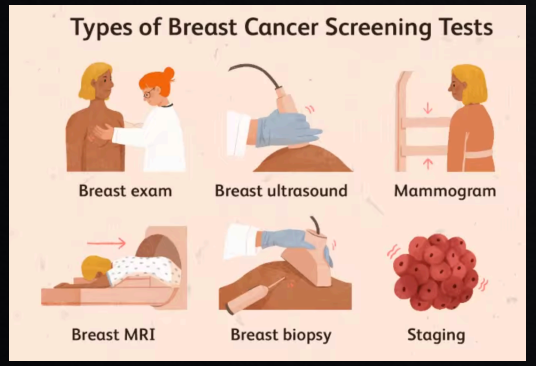
Sorting Problem

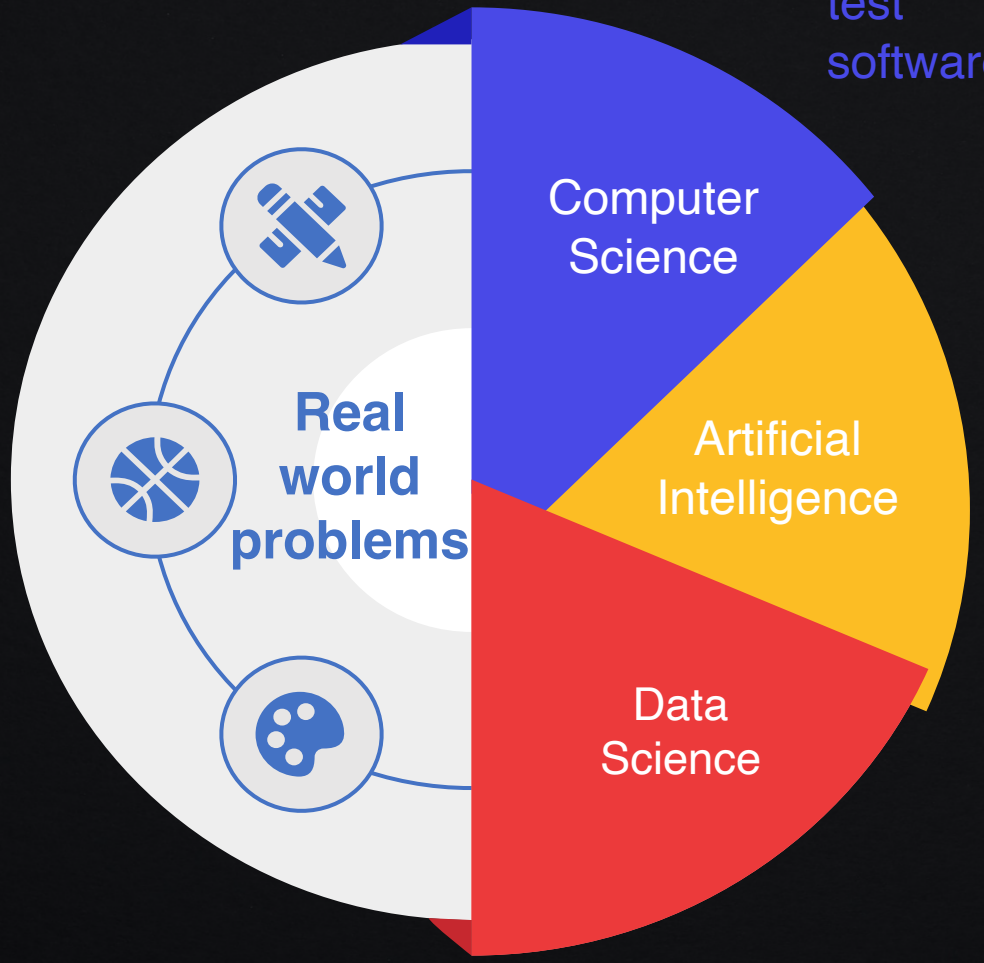


Image



Breast cancer detection

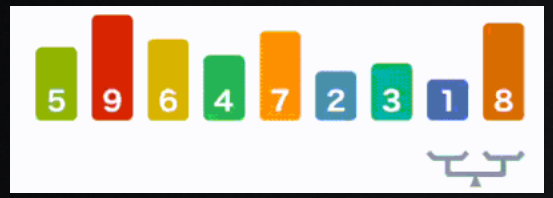




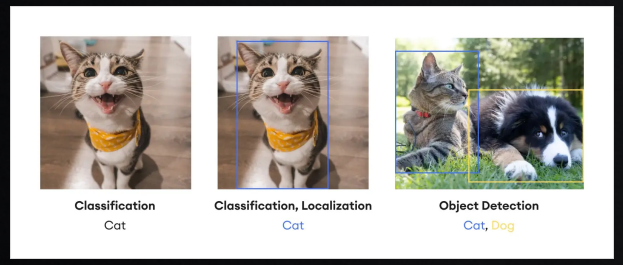
algorithm
code
test
software

- IDEs
- Python, Java, C++
- Visual Studio
- Microsoft Azure

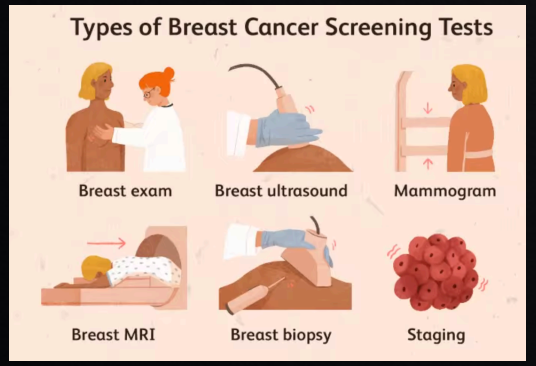
Sorting Problem

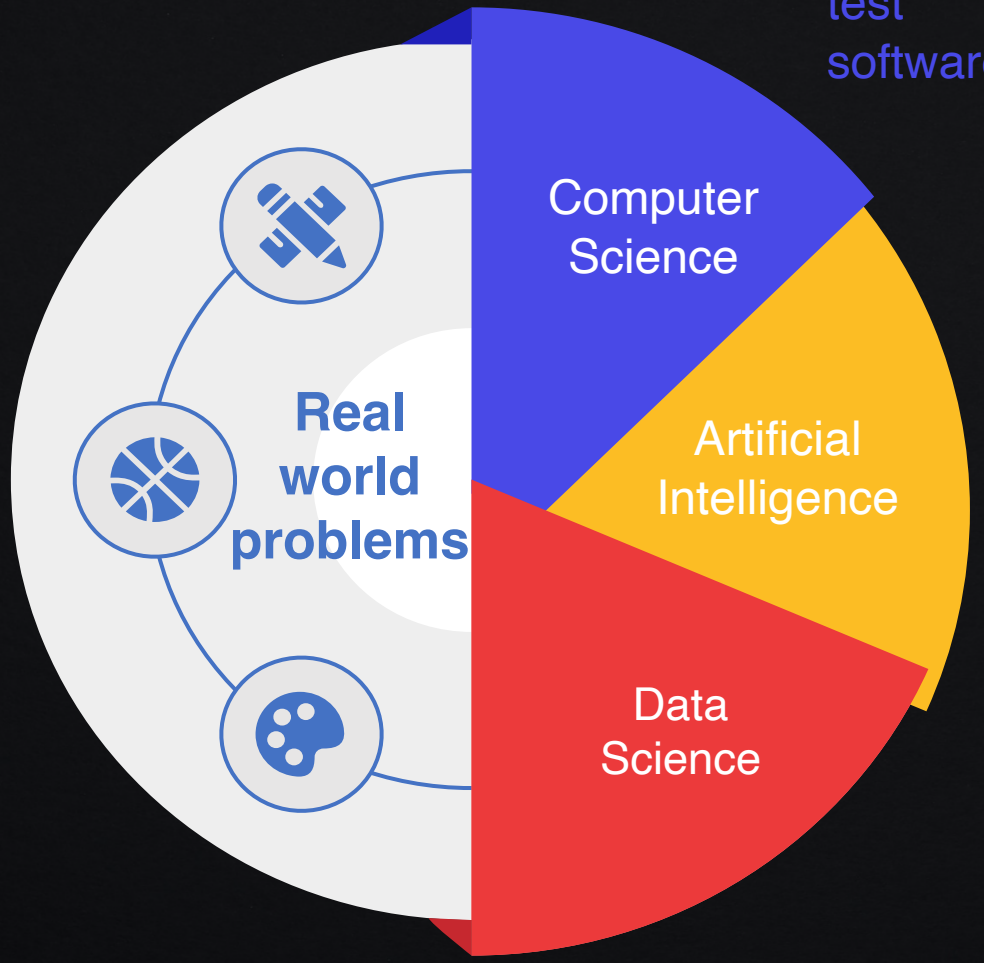


Image



Breast cancer detection





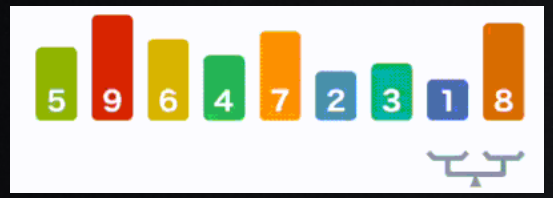
algorithm
code
test
software

- IDEs
- Python, Java, C++
- Visual Studio
- Microsoft Azure

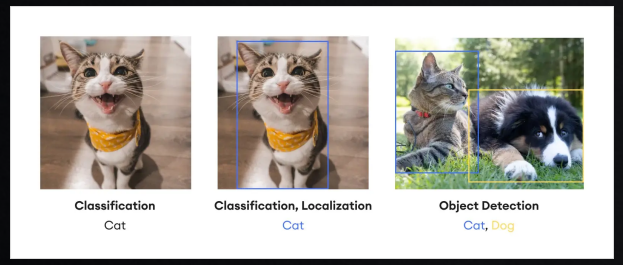
models (ML, DL)
feature engineering
train
learn-from-data

- Sickit-learn
- Pytorch
- Import Transformer

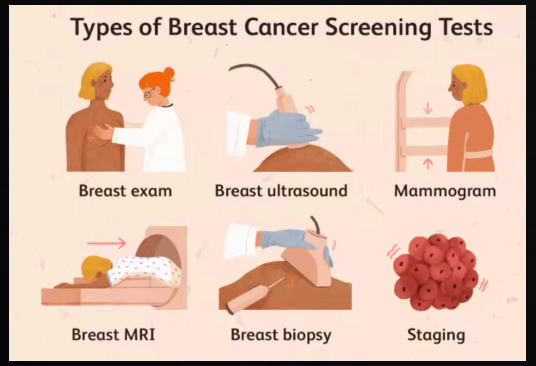
Sorting Problem

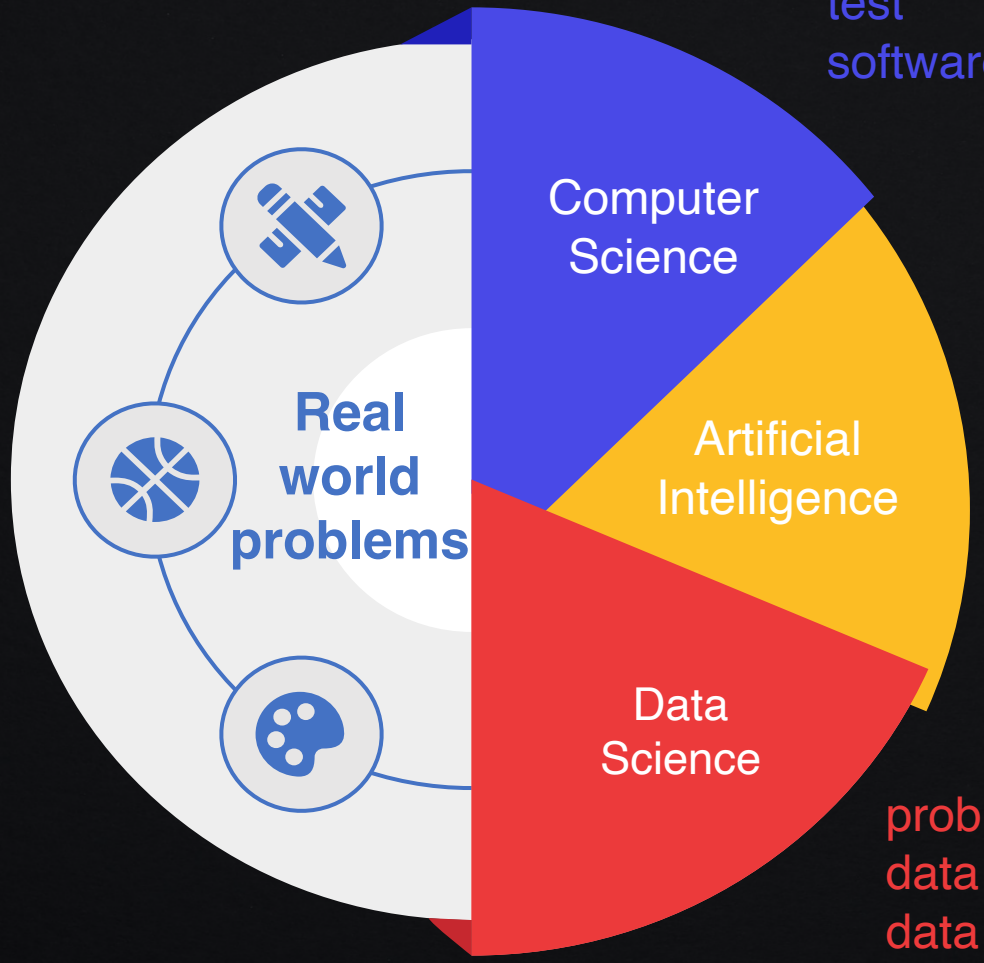


Image



Breast cancer detection





algorithm
code
test
software

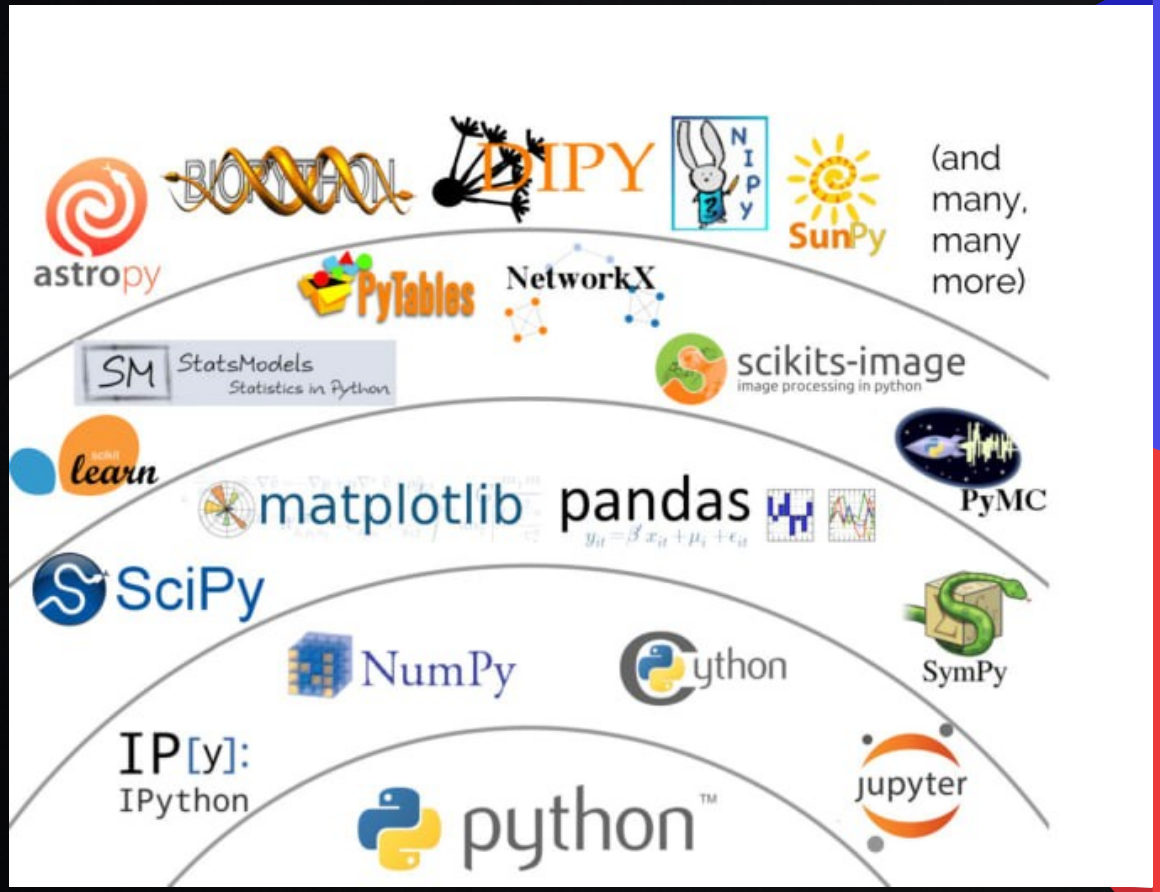
- IDEs
- Python, Java, C++
- Visual Studio
- Microsoft Azure

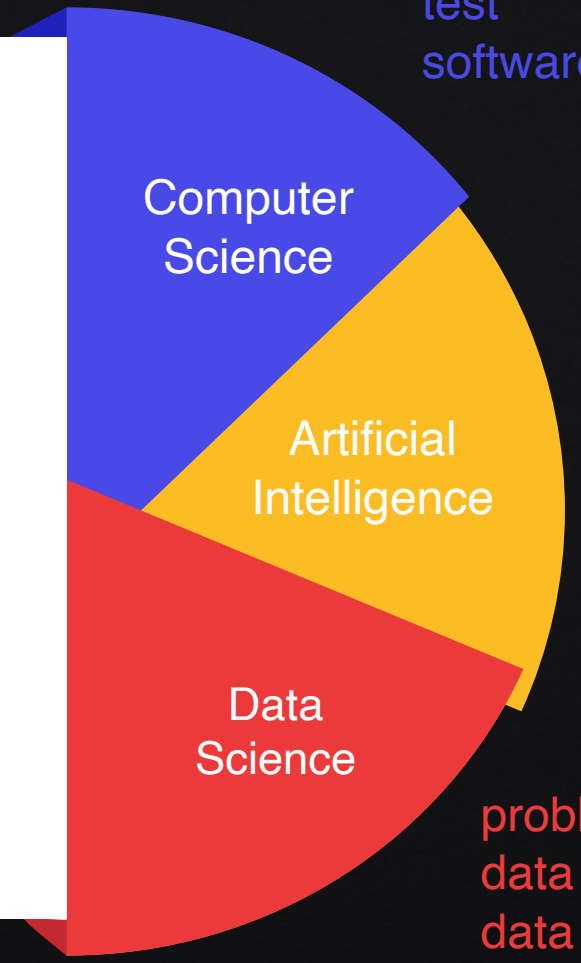
models (ML, DL)
feature engineering
train
learn-from-data

- Sickit-learn
- Pytorch
- Import Transformer

problem?
data acquisition
data exploration
data preparation
CS/AI/Stat methods
loop back

- SQL
- R, SAS
- Python
- Tableau
- Jupyter Notebook
- PySpark





algorithm
code
test
software

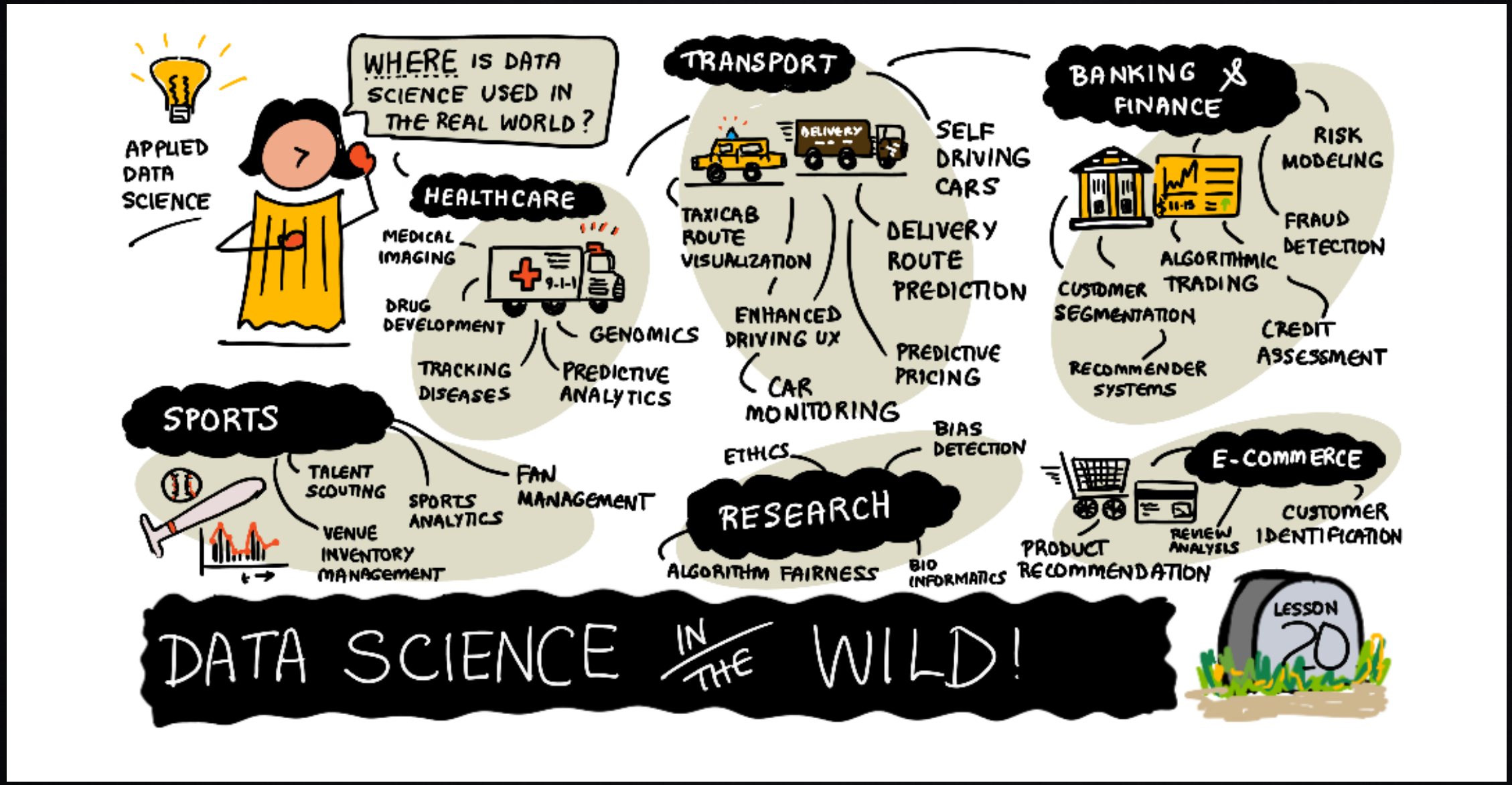
- IDEs
- Python, Java, C++
- Visual Studio
- Microsoft Azure

models (ML, DL)
feature engineering
train
learn-from-data

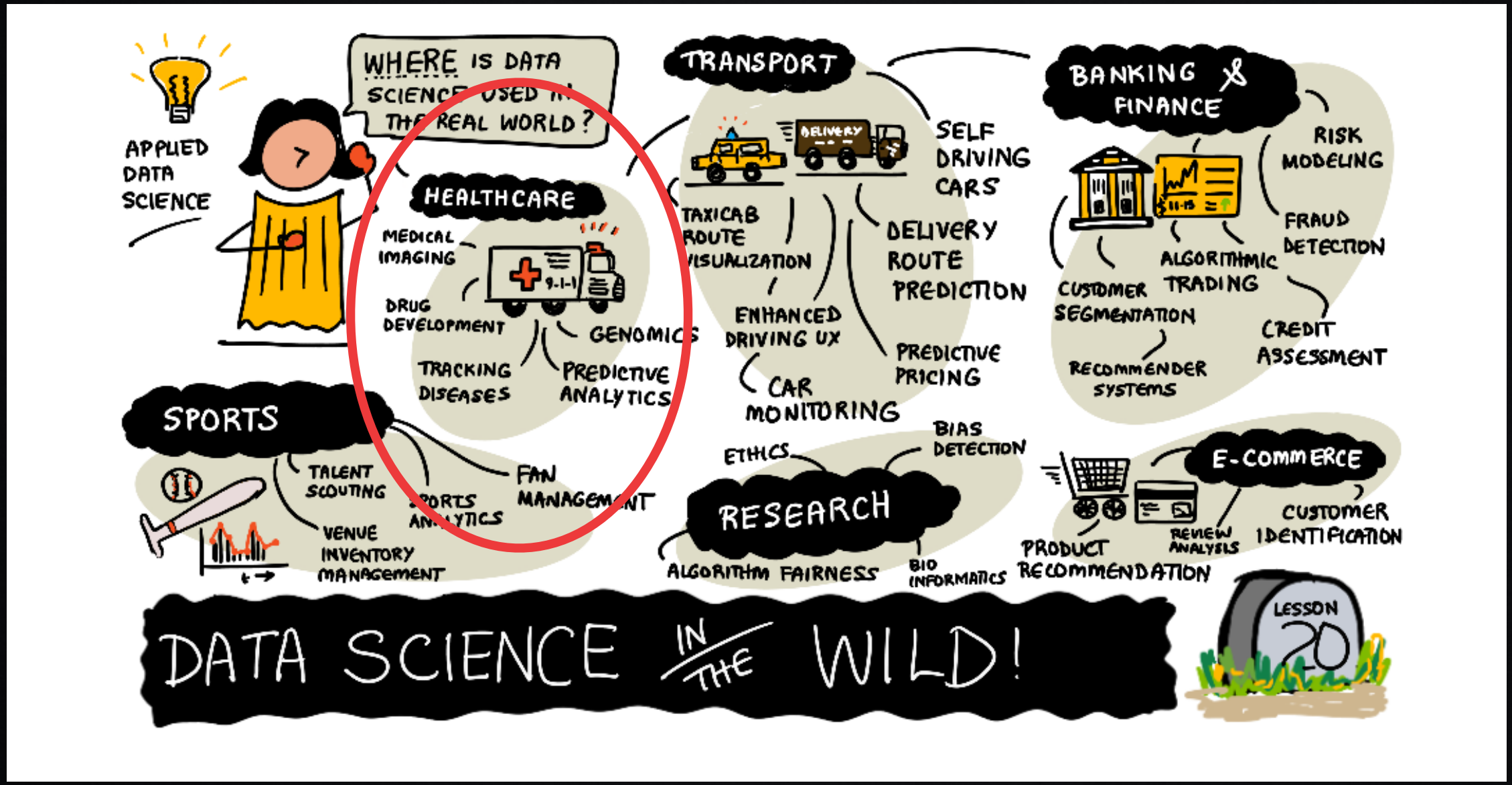
- Sickit-learn
- Pytorch
- Import Transformer

problem?
data acquisition
data exploration
data preparation
CS/AI/Stat methods
loop back

- SQL
- R, SAS
- Python
- Tableau
- Jupyter Notebook
- PySpark



Data Science In The Real World - *Sketchnote by [@nitya](#)*

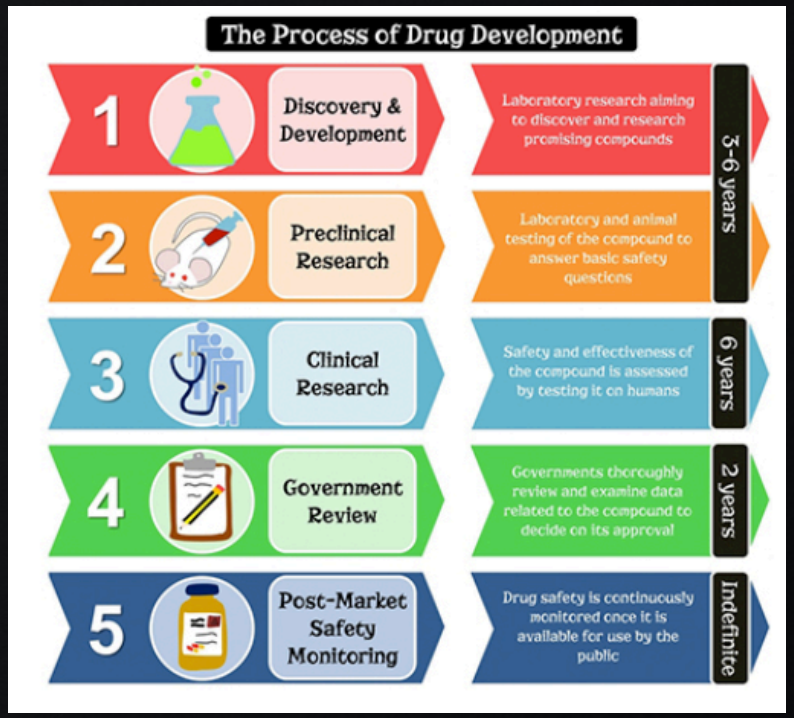


Data Science In The Real World - *Sketchnote by [@nitya]*

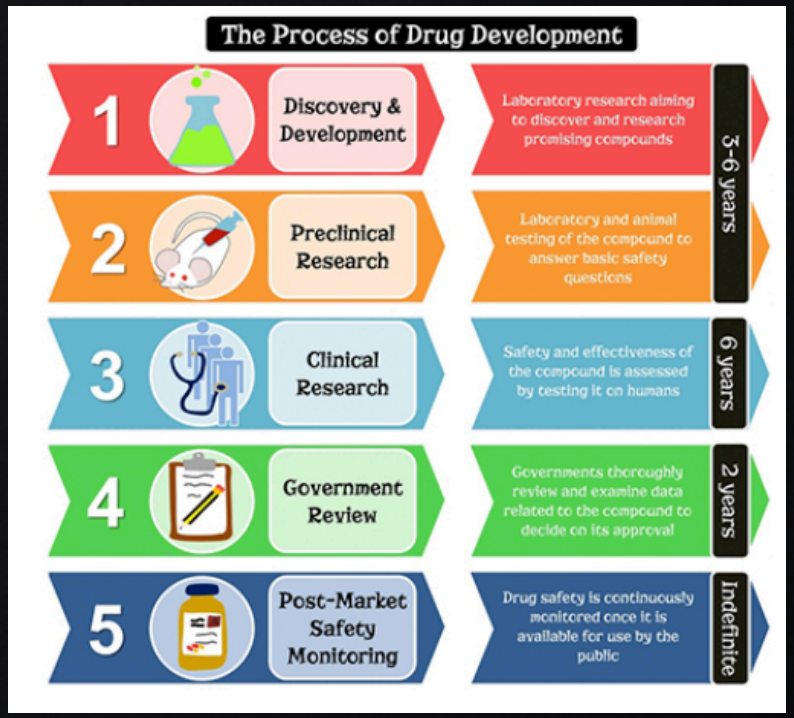
Data Science for Healthcare

Data Science for Healthcare

Data Science for Healthcare



Data Science for Healthcare




Data science life cycle?


Data science life cycle?


Data science life cycle?

DOI:10.1145/3360646

A cycle that traces ways to define the landscape of data science.

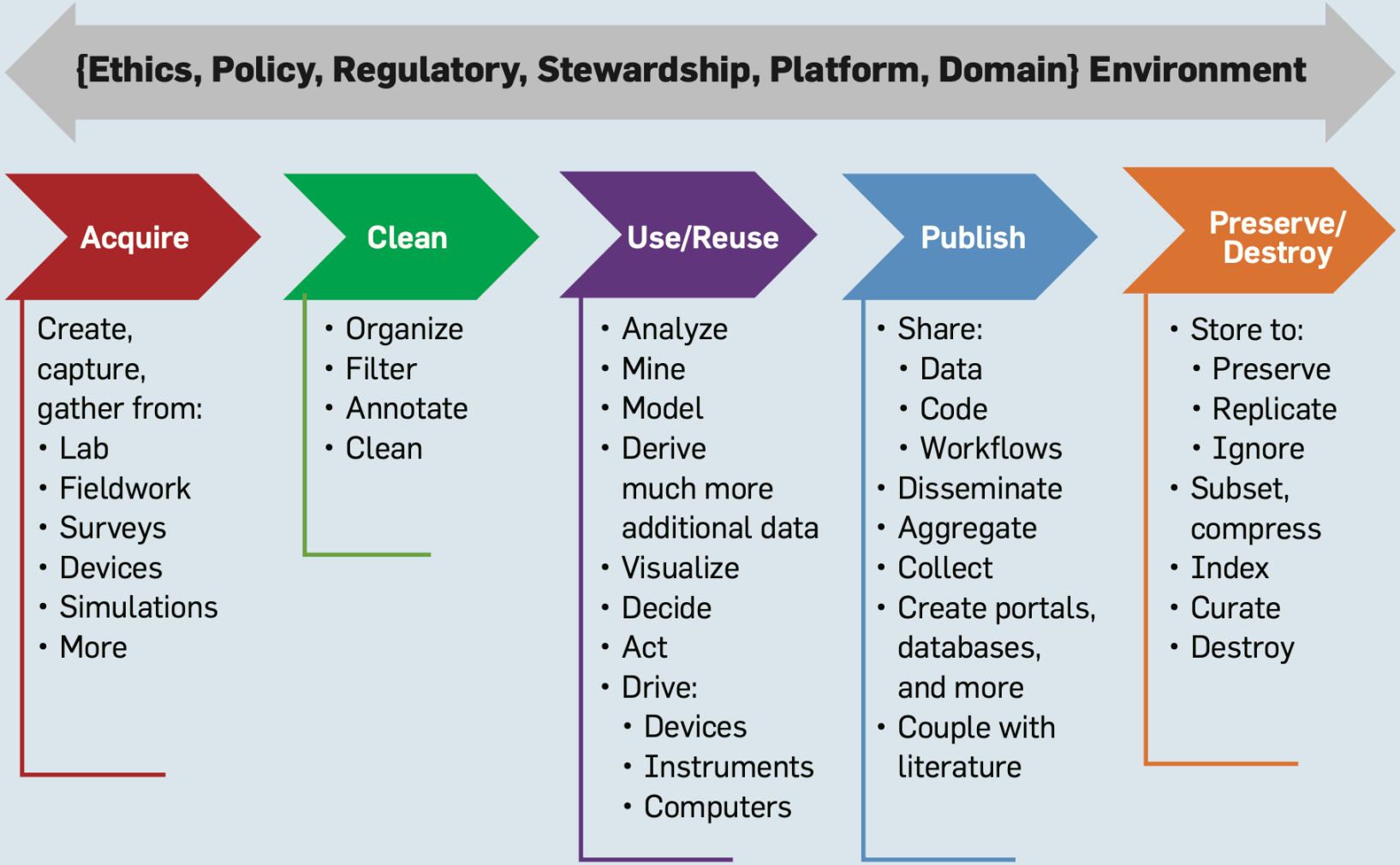
BY VICTORIA STODDEN

The Data Science Life Cycle:

A Disciplined Approach to Advancing Data Science as a Science

THE EDUCATION AND research enterprise is leveraging opportunities to accelerate science and discovery offered by computational and data-enabled technologies, often broadly referred to as data science. Ten years ago, we wrote that an "accurate image [of a scientific researcher] depicts a computer jockey working at all hours to launch experiments on computer servers."[8] Since then, the use of data and computation has exploded in academic and industry research, and interest in data science is widespread in universities and institutions. Two key questions emerge for the research enterprise: How to train

Figure 1. Example of a data life cycle and surrounding data ecosystem (reprinted with permission).[1]



An extended data science life cycle

DOI:10.1145/3360646

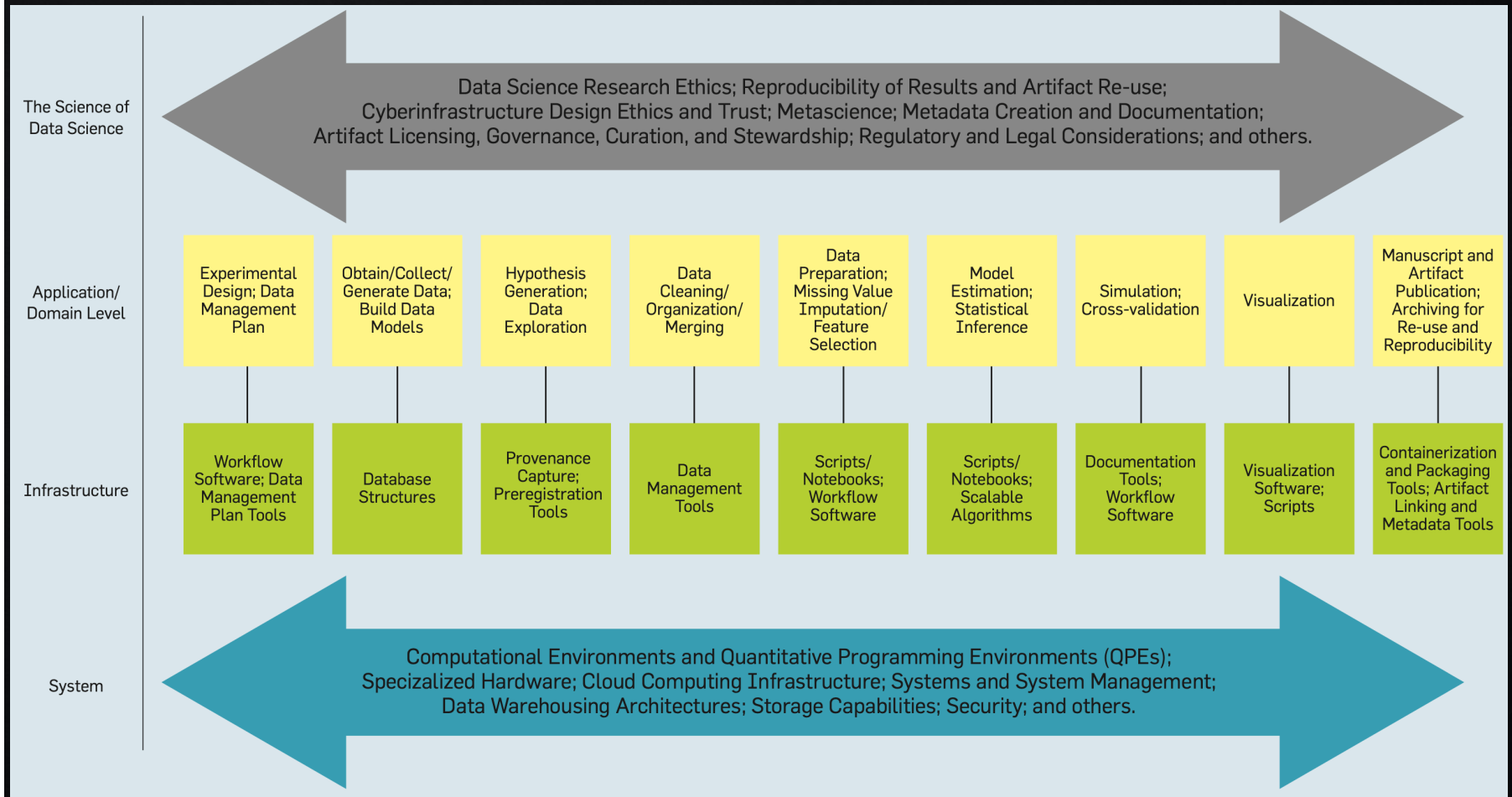
A cycle that traces ways to define the landscape of data science.

BY VICTORIA STODDEN

The Data Science Life Cycle:

A Disciplined Approach to Advancing Data Science as a Science

THE EDUCATION AND research enterprise is leveraging opportunities to accelerate science and discovery offered by computational and data-enabled technologies, often broadly referred to as data science. Ten years ago, we wrote that an "accurate image [of a scientific researcher] depicts a computer jockey working at all hours to launch experiments on computer servers."[8] Since then, the use of data and computation has exploded in academic and industry research, and interest in data science is widespread in universities and institutions. Two key questions emerge for the research enterprise: How to train



Data Science Life Cycle Step	Possible (Existing) Courses	
Experimental design	▸ Introduction to Probability ▸ Introduction to Statistics ▸ Design of Experiments (including Human Subjects and Informed Consent)	CS, Stat
Obtaining data	▸ Experimental Methodology ▸ Introduction to Databases ▸ Introduction to SQL, noSQL ▸ Sensor Integration and Control	D S
Data exploration	▸ Introduction to R ▸ Introduction to python ▸ Graphics and Data Visualization ▸ Introduction to Statistics	DS
Databases and data structures including cleaning/organizing	▸ Introduction to Database Systems ▸ Introduction to SQL, noSQL ▸ Natural Language Processing (NLP)	DS AI
Software engineering	▸ Python, R, C, C++, Julia ▸ Distributed Systems, MapReduce ▸ Software Testing	CS DS AI
Feature selection	▸ Statistical Learning ▸ Domain-specific courses, for example, Bioinformatics for Transcriptomics; Brain Imaging in Cognitive Neuroscience Research	DS AI CMA
Model estimation	▸ Mathematics (Probability, Linear Algebra, Calculus, Real Analysis) ▸ Applied Statistics ▸ Machine Learning ▸ Data Mining ▸ Deep Learning ▸ Scalable Algorithms ▸ Statistical Decision Theory	DS AI CMA
Simulation and cross-validation	▸ Fundamentals of Numerical Methods ▸ Introduction to Computer Modeling and Simulation ▸ Statistical Learning	CS DS
Visualization	▸ Information Visualization ▸ Scientific Visualization and Graphics ▸ [Domain specific courses such as Learning ArcGIS; Spatial Data Visualization]	DS CMA
Publication/Archiving	▸ Introduction to Information ▸ Data Archiving and FAIR Data ▸ Scientific Report Writing ▸ Research Data Management ▸ Open Access and Scholarly Communication ▸ Digital Libraries and Preservation	CS DS
Overarching topics	▸ Ethics for Scientists ▸ Data Privacy ▸ National and International Regulatory Trends in Data Protection	

Data acquisition

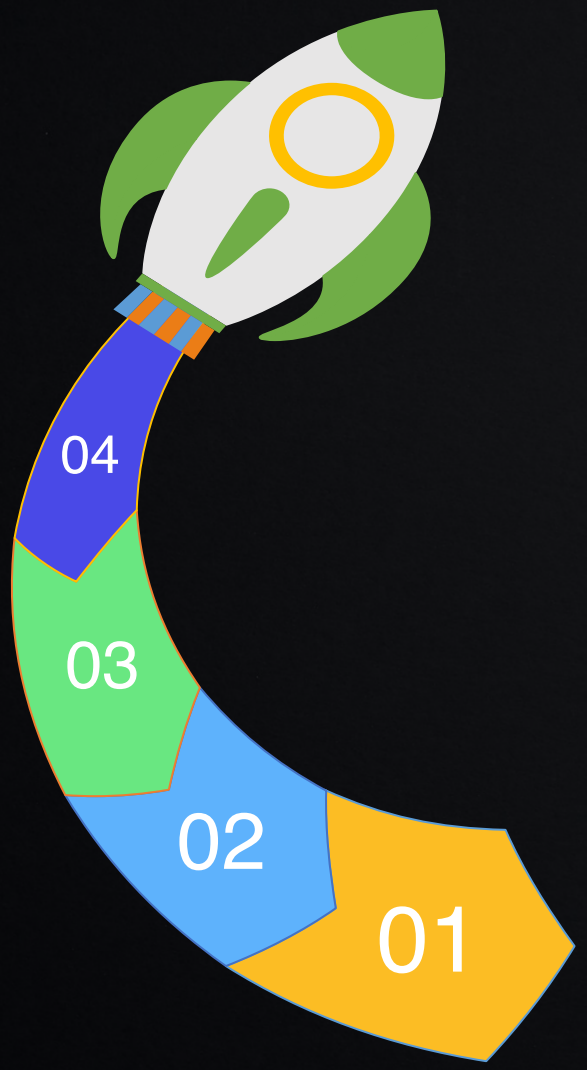
Exploratory data analysis

Data preparation

Data visualization

Advanced topics:
- data-centric AI

Data Science Life Cycle Step	Possible (Existing) Courses	
Experimental design	▸ Introduction to Probability ▸ Introduction to Statistics ▸ Design of Experiments (including Human Subjects and Informed Consent)	CS, Stat
Obtaining data	▸ Experimental Methodology ▸ Introduction to Databases ▸ Introduction to SQL, noSQL ▸ Sensor Integration and Control	D S
Data exploration	▸ Introduction to R ▸ Introduction to python ▸ Graphics and Data Visualization ▸ Introduction to Statistics	DS
Databases and data structures including cleaning/organizing	▸ Introduction to Database Systems ▸ Introduction to SQL, noSQL ▸ Natural Language Processing (NLP)	DS AI
Software engineering	▸ Python, R, C, C++, Julia ▸ Distributed Systems, MapReduce ▸ Software Testing	CS DS AI
Feature selection	▸ Statistical Learning ▸ Domain-specific courses, for example, Bioinformatics for Transcriptomics; Brain Imaging in Cognitive Neuroscience Research	DS AI CMA
Model estimation	▸ Mathematics (Probability, Linear Algebra, Calculus, Real Analysis) ▸ Applied Statistics ▸ Machine Learning ▸ Data Mining ▸ Deep Learning ▸ Scalable Algorithms ▸ Statistical Decision Theory	DS AI CMA
Simulation and cross-validation	▸ Fundamentals of Numerical Methods ▸ Introduction to Computer Modeling and Simulation ▸ Statistical Learning	CS DS
Visualization	▸ Information Visualization ▸ Scientific Visualization and Graphics ▸ [Domain specific courses such as Learning ArcGIS; Spatial Data Visualization]	DS CMA
Publication/Archiving	▸ Introduction to Information ▸ Data Archiving and FAIR Data ▸ Scientific Report Writing ▸ Research Data Management ▸ Open Access and Scholarly Communication ▸ Digital Libraries and Preservation	CS DS
Overarching topics	▸ Ethics for Scientists ▸ Data Privacy ▸ National and International Regulatory Trends in Data Protection	



01
Data Acquisition

- image data
- text data
- tabular data

02
Data Understanding

- exploratory data analysis
- data visualization

03
Data Preparation

- data transformation
- data cleaning
- entity resolution
- … …

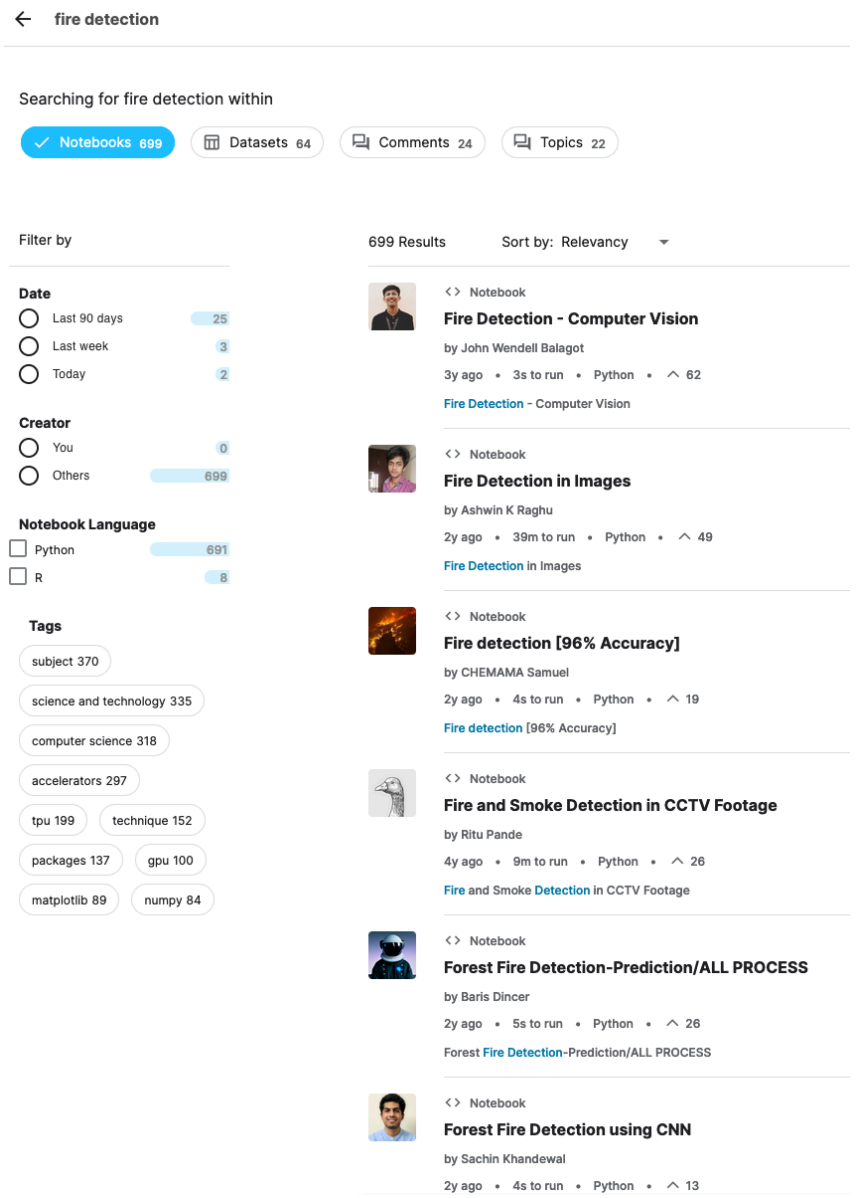
04
Advanced Topics

- data-centric AI
- HPC for LLMs
- … …

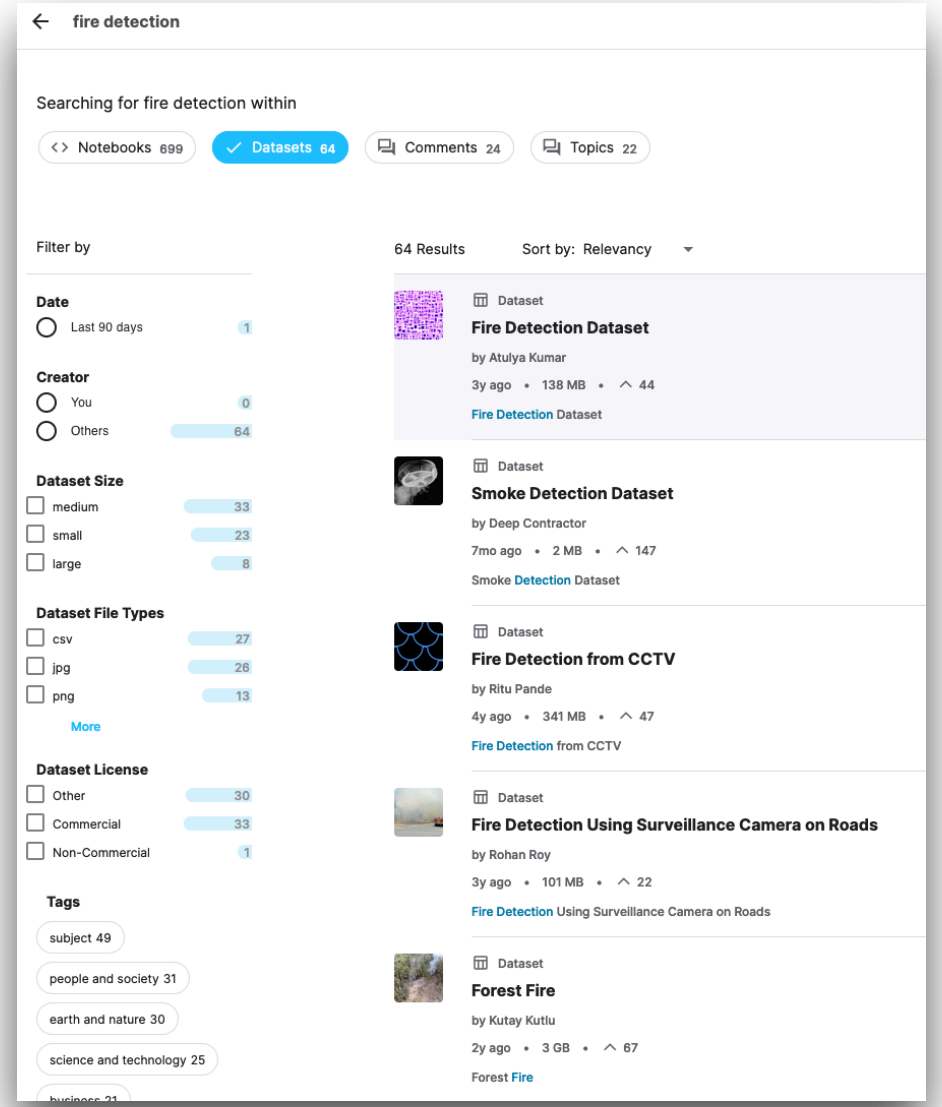
Classify fires as: {Ground, Surface, Crown, No}

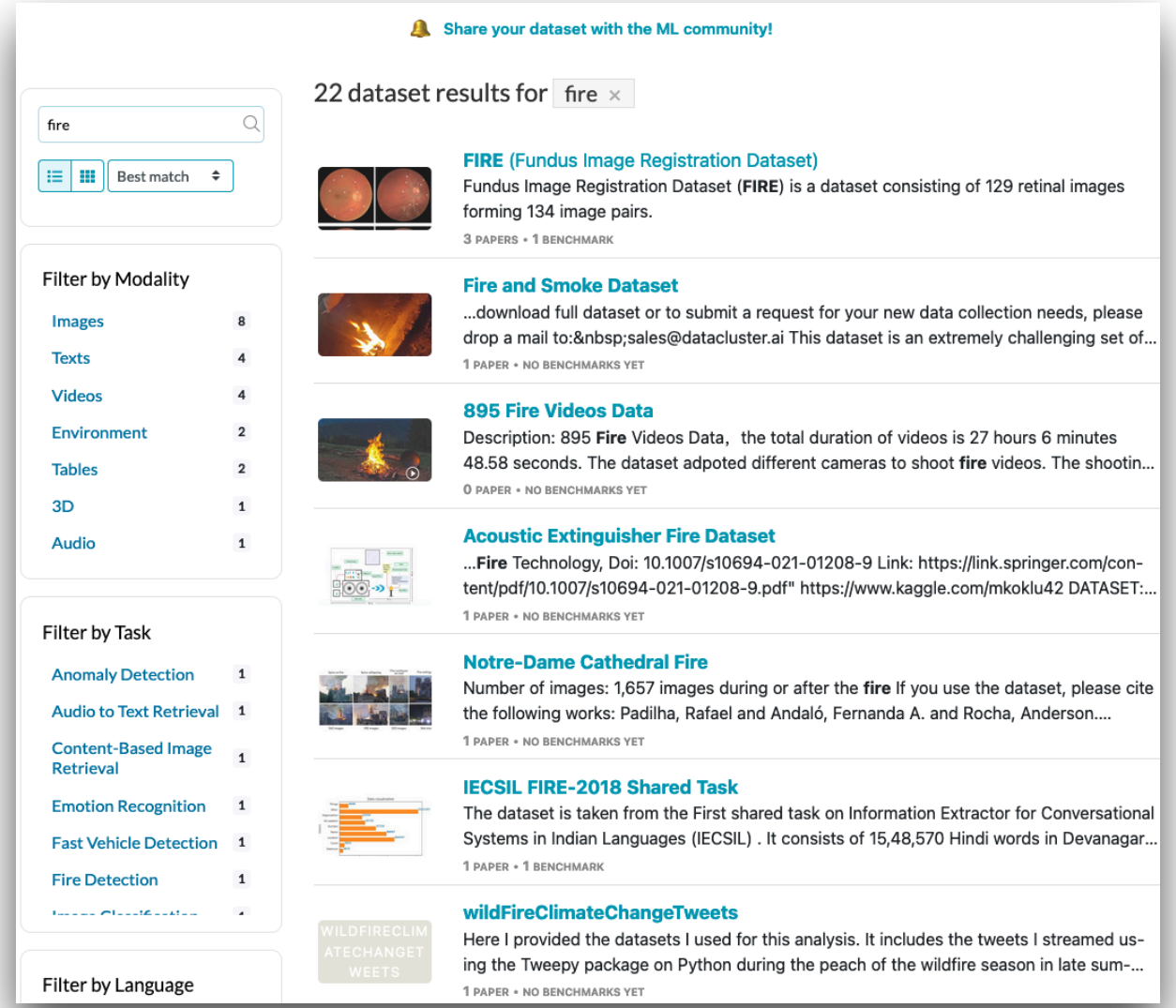
(**a**) Fire image samples.

Find the Model



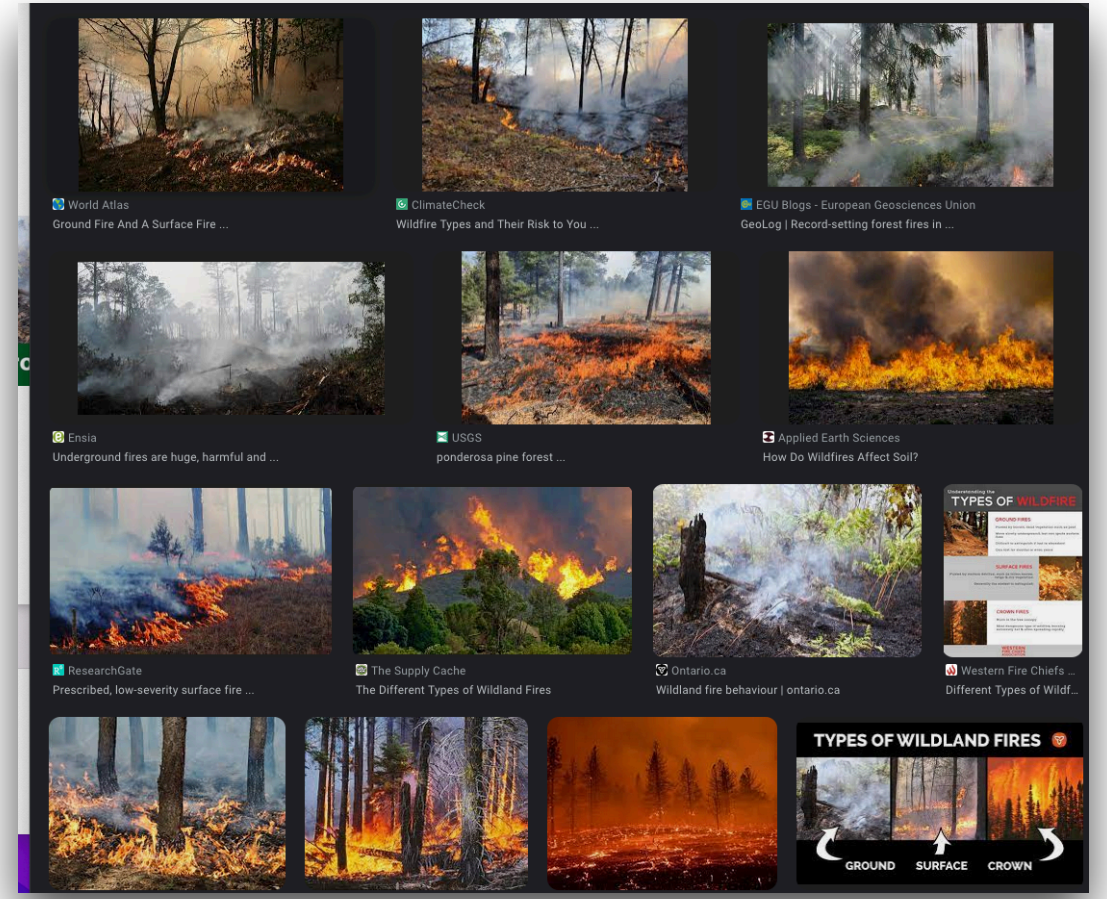
Find Datasets





Find Datasets

Google image search

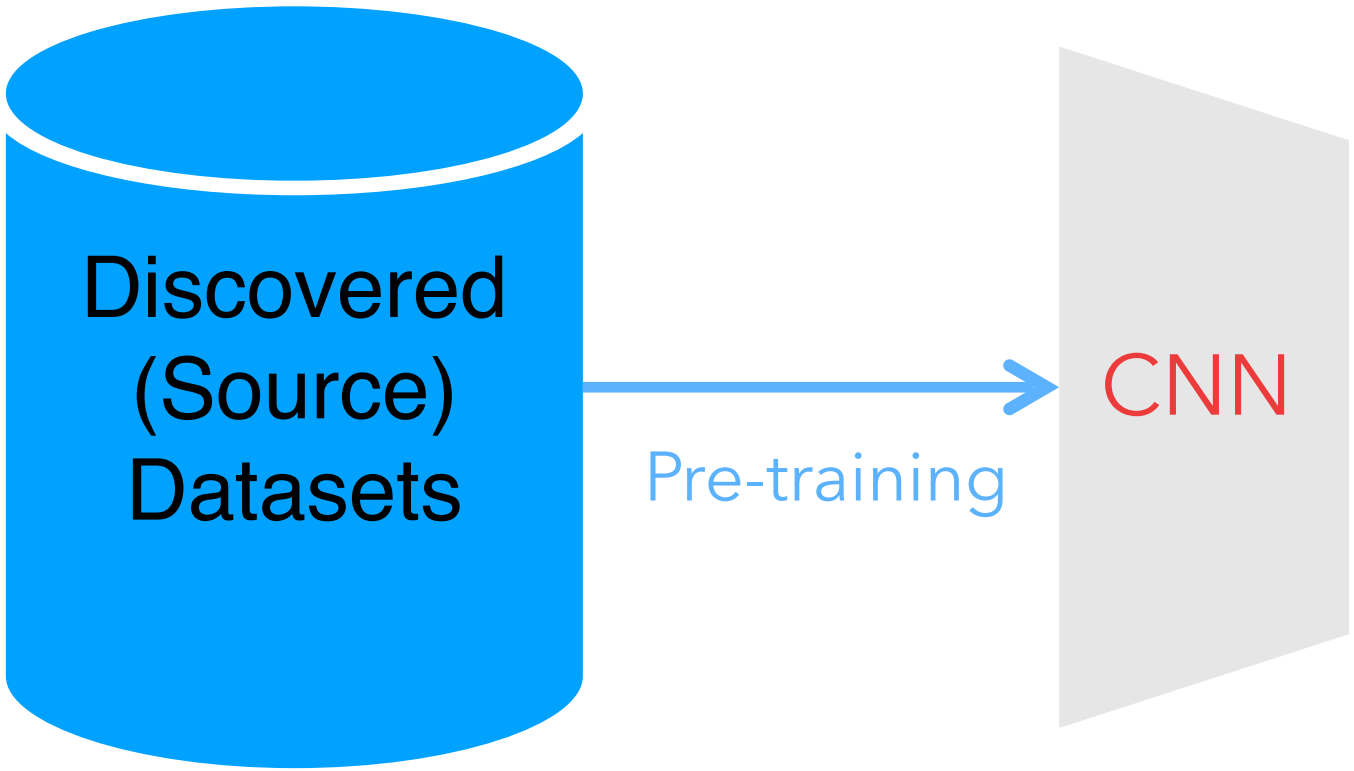


Google search: “Ground fire”

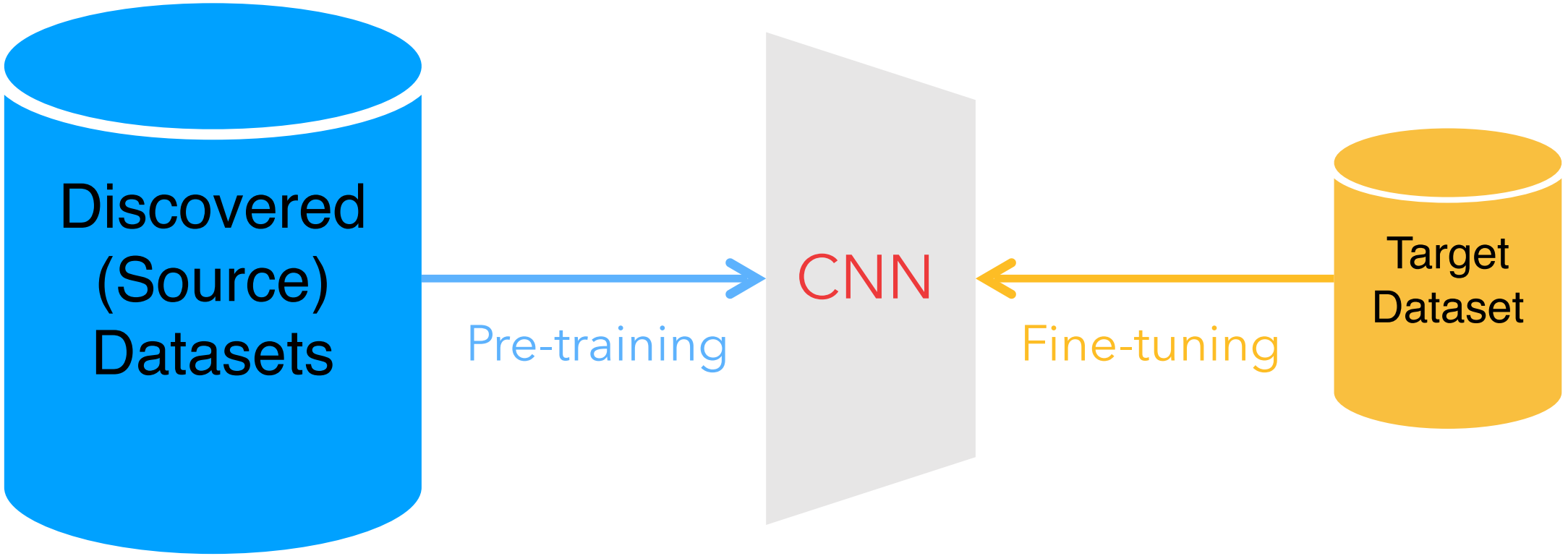
Google search: “Crown fire”

Transfer Learning & Data

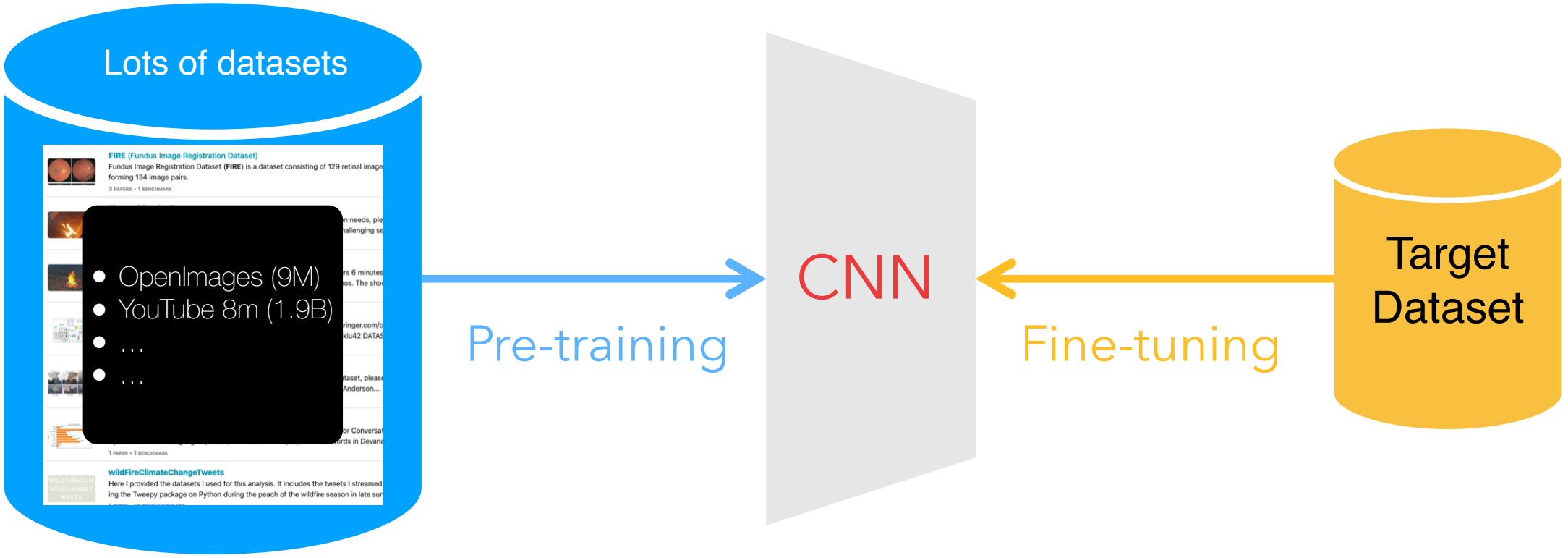
Transfer Learning & Data



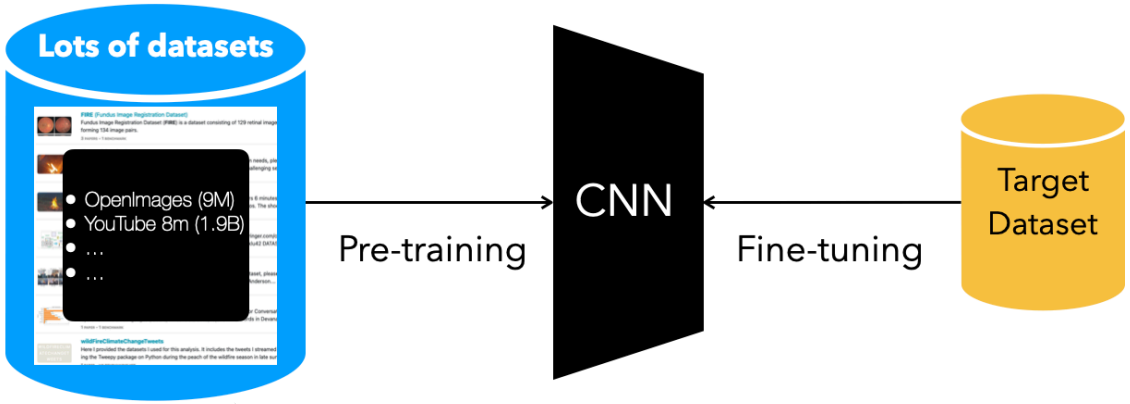
Transfer Learning & Data



Transfer Learning & Data

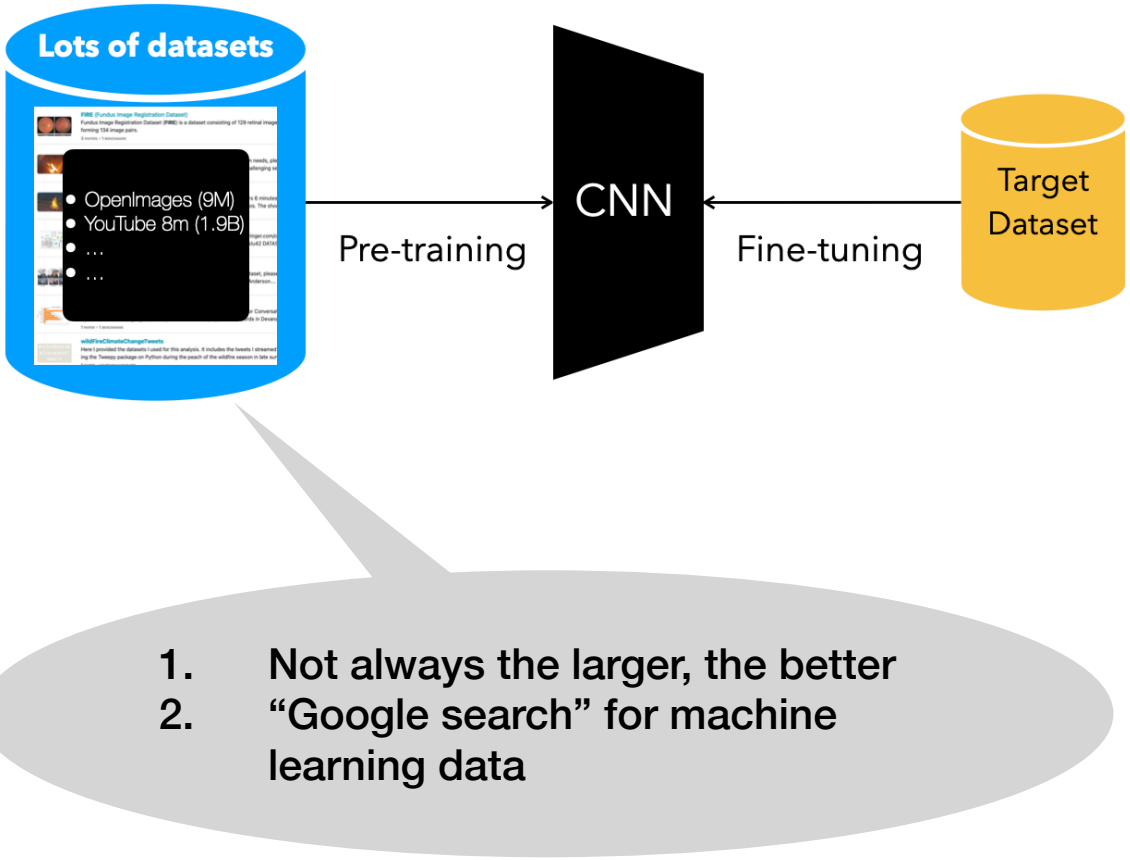


Question: which data to use for pre-training



1. Not always the larger, the better
2. “Google search” for machine learning data

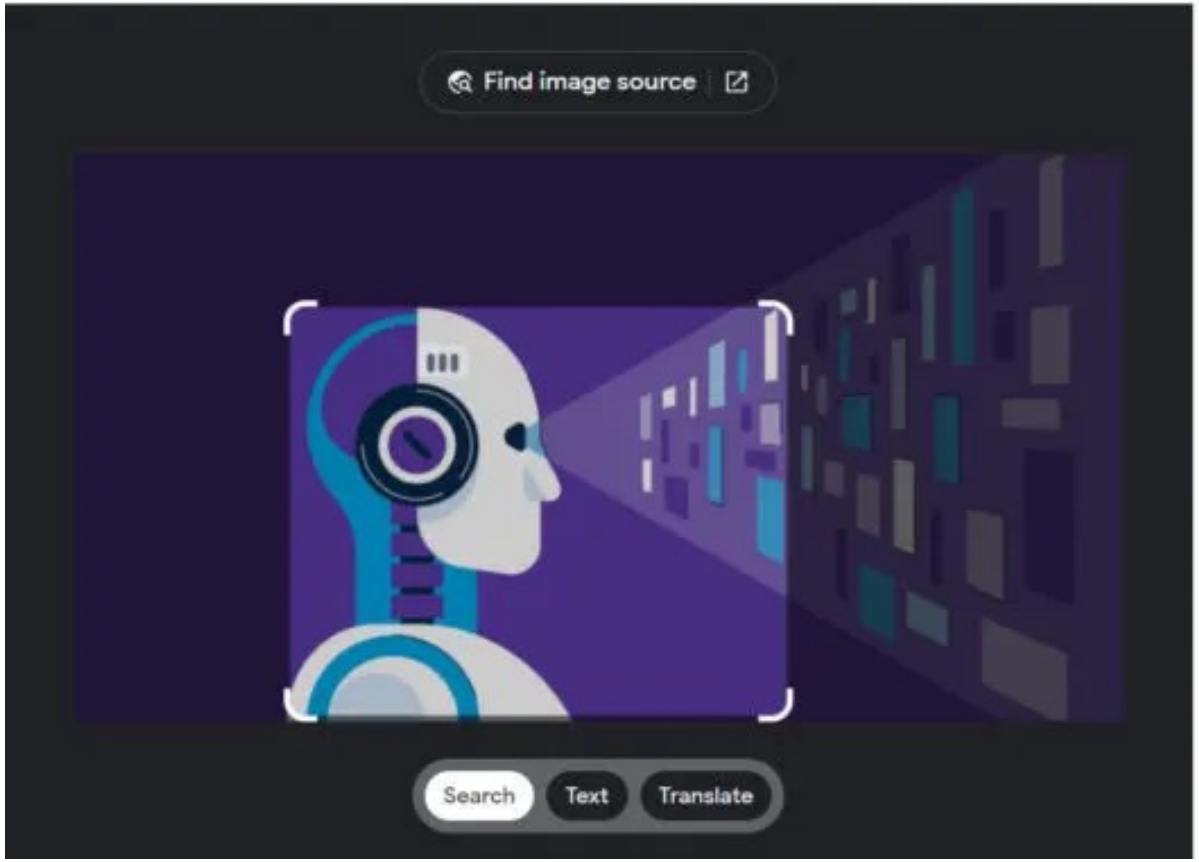
Question: which data to use for pre-training

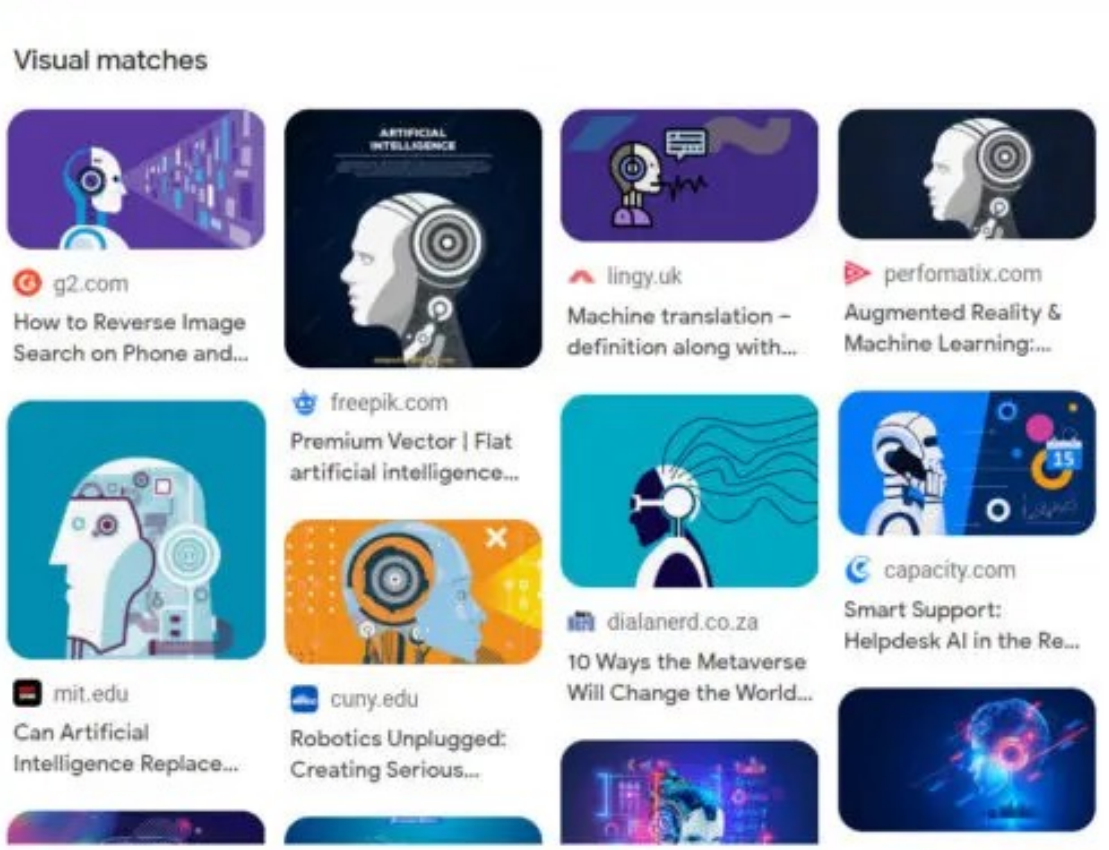


1. Not always the larger, the better
2. "Google search" for machine learning data

1. I have a small dataset
2. Find what data is useful for me
3. Pretrain and finetrune on my small dataset

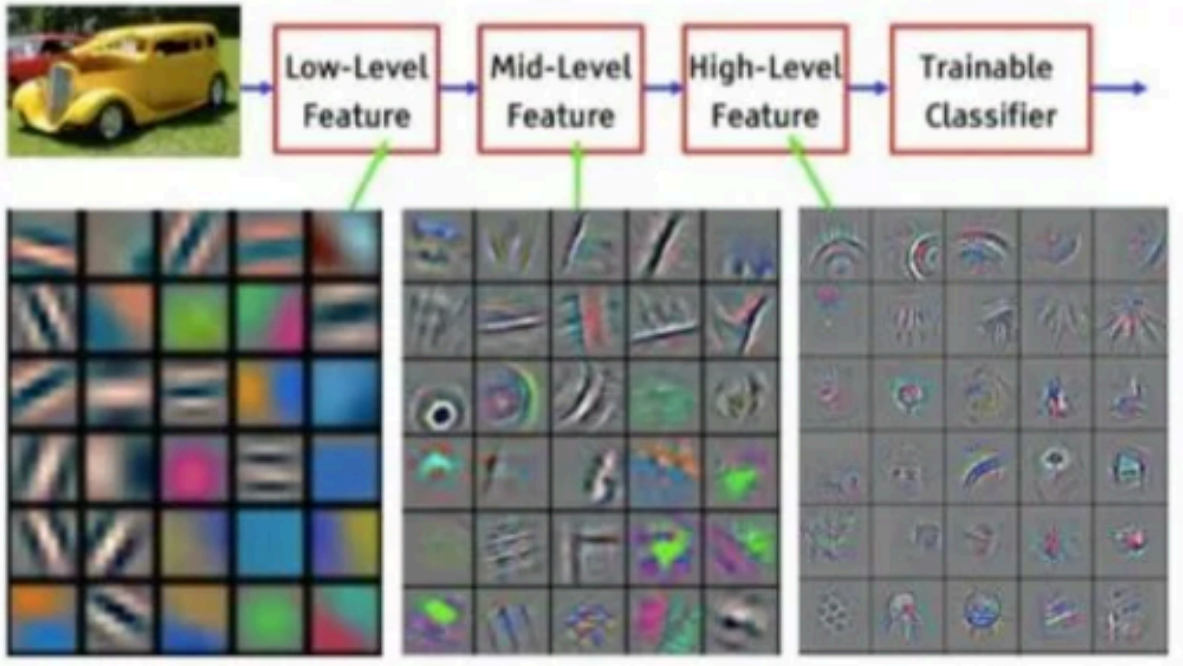
Reverse Image Search





Source of picture [link]

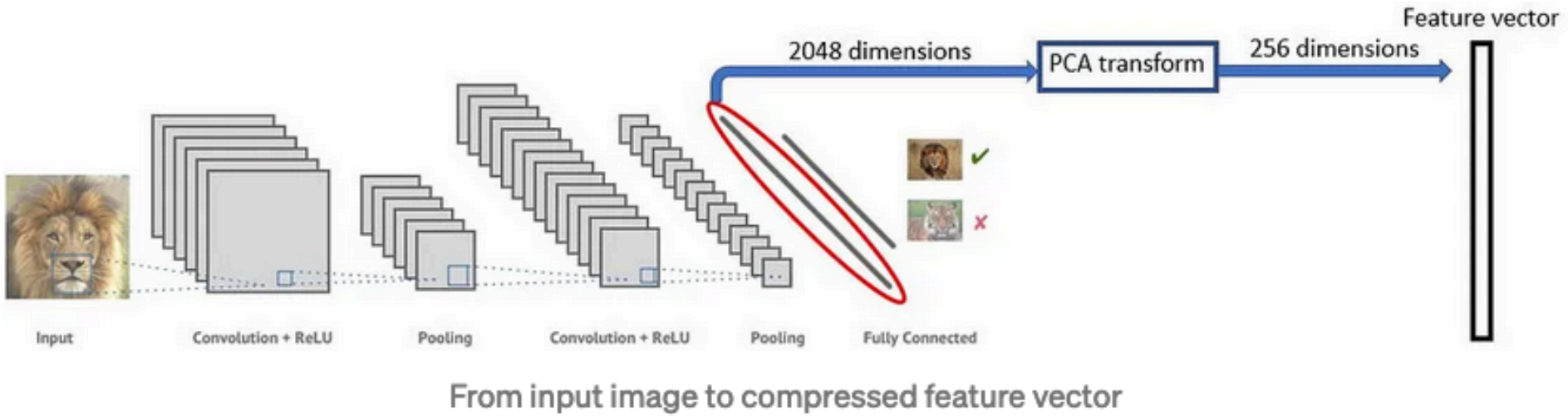
Reverse Image Search



From image to feature extraction. Example of features that the filters in a convolution layer look for at different levels in a network. The deeper into the network (higher level), the more complex the features are. Source: Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." Computer Vision–ECCV 2014. Springer International Publishing, 2014. 818–833

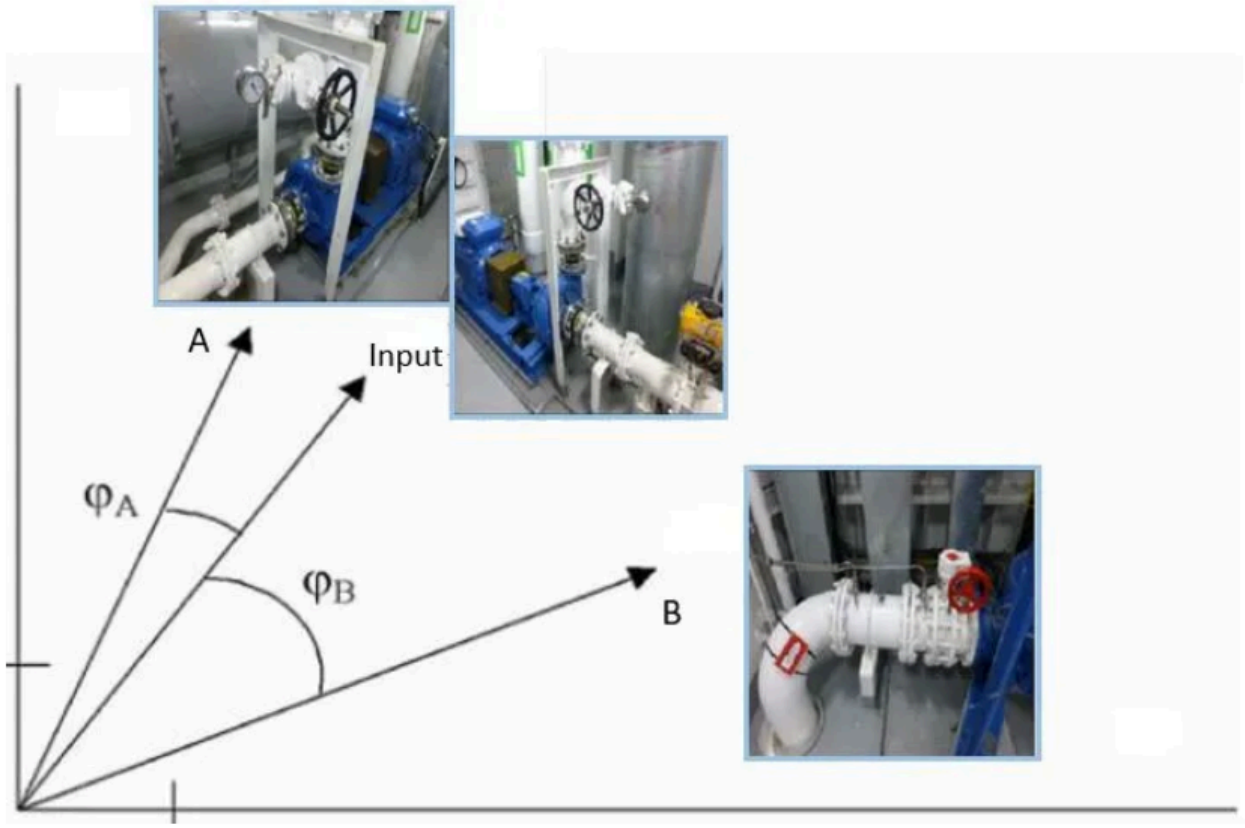
Source of picture [link]

Reverse Image Search: Image2Vec



Source of picture [link]

Reverse Image Search: Cosine Similarity



Cosine similarity as a metric to compare feature vectors

Source of picture [link]

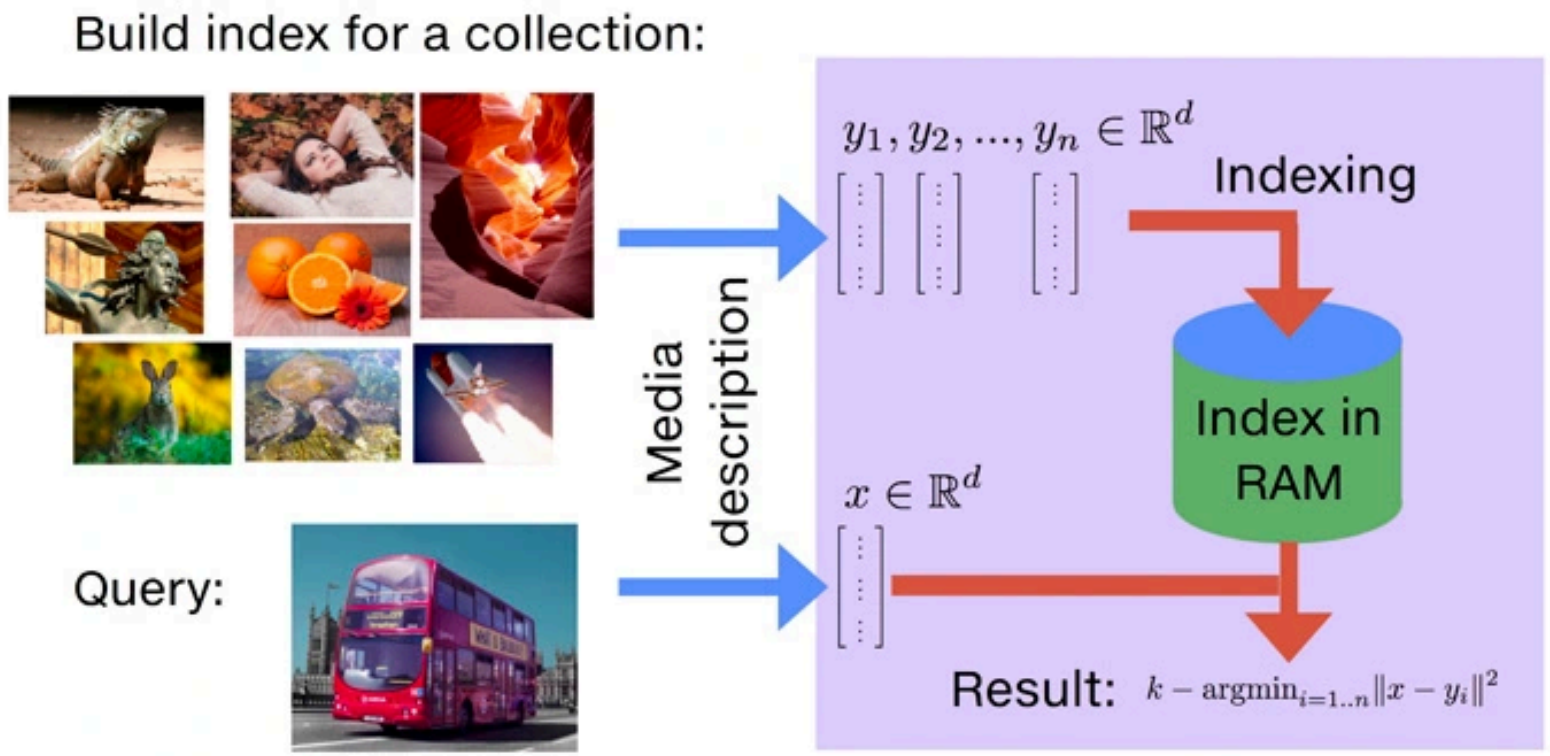
Reverse Image Search: Clustering

From feature vector to clustering. Image source : Visualizing Data using t-SNE, research paper by Laurens van der Maaten and Geoffrey Hinton

Source of picture [link]

Reverse Image Search: Index a Collection

Meta Faiss [link]



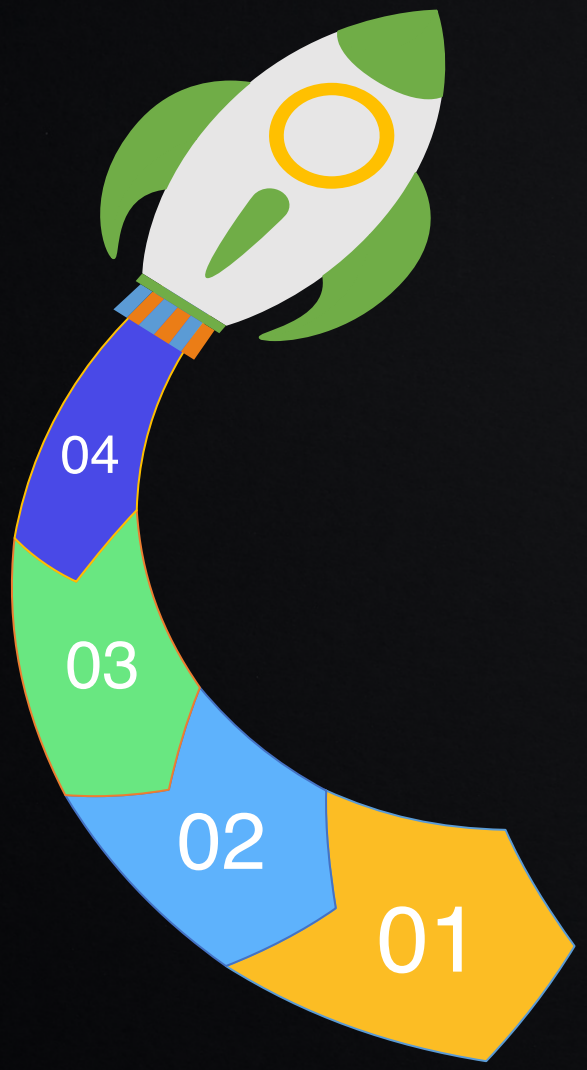
Source of picture [link]

Landinglens @ LandingAI, a data-centric AI company founded by Andrew Ng

Landinglens @ LandingAI, a data-centric AI company founded by Andrew Ng

Landinglens @ LandingAI, a data-centric AI company founded by Andrew Ng

enc(picture) != enc(rotate_180(picture))



01
Data Acquisition

- image data
- tabular
- text data

02
Data Understanding

- exploratory data analysis
- data visualization

03
Data Preparation

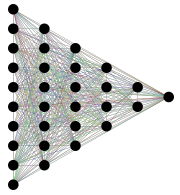
- data transformation
- data cleaning
- entity resolution
- … …

04
Advanced Topics

- data-centric AI
- HPC for LLMs
- … …

Not enough train data

City	Year	Area	Security	Price
Kolkata	2009	710	No	3,200,000
Kolkata	2013	770	No	3,850,000
Kolkata	2007	935	No	2,524,000
Kolkata	2006	973	Yes	3,611,000



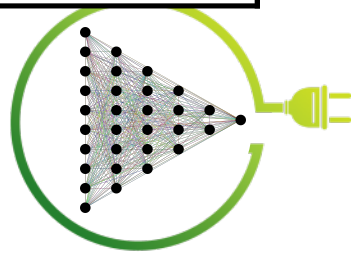
City	Year	Area	Security	Price
Kolkata	2017	350	No	?
Kolkata	2019	465	Yes	?
Kolkata	2015	572	No	?
Kolkata	2012	655	Yes	?
Kolkata	2012	735	No	?
Kolkata	2017	881	Yes	?
Kolkata	2011	1123	Yes	?
Kolkata	2014	1210	Yes	?

Learn a good model: mission impossible

Need more relative good data

Train

City	Year	Area	Security	Price
Kolkata	2009	710	No	3,200,000
Kolkata	2013	770	No	3,850,000
Kolkata	2007	935	No	2,524,000
Kolkata	2006	973	Yes	3,611,000



Test

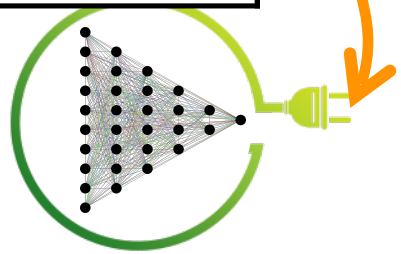
City	Year	Area	Security	Price
Kolkata	2017	350	No	?
Kolkata	2019	465	Yes	?
Kolkata	2015	572	No	?
Kolkata	2012	655	Yes	?
Kolkata	2012	735	No	?
Kolkata	2017	881	Yes	?
Kolkata	2011	1123	Yes	?
Kolkata	2014	1210	Yes	?

Need more relative good data

Model charging

Train

City	Year	Area	Security	Price
Kolkata	2009	710	No	3,200,000
Kolkata	2013	770	No	3,850,000
Kolkata	2007	935	No	2,524,000
Kolkata	2006	973	Yes	3,611,000



Test

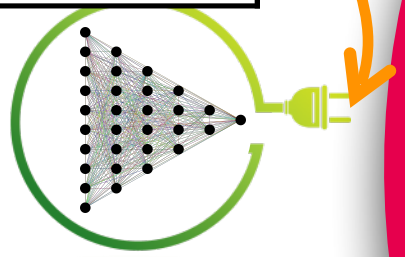
City	Year	Area	Security	Price
Kolkata	2017	350	No	?
Kolkata	2019	465	Yes	?
Kolkata	2015	572	No	?
Kolkata	2012	655	Yes	?
Kolkata	2012	735	No	?
Kolkata	2017	881	Yes	?
Kolkata	2011	1123	Yes	?
Kolkata	2014	1210	Yes	?

Need more relative good data

Model charging

Train

City	Year	Area	Security	Price
Kolkata	2009	710	No	3,200,000
Kolkata	2013	770	No	3,850,000
Kolkata	2007	935	No	2,524,000
Kolkata	2006	973	Yes	3,611,000



Test

City	Year	Area	Security	Price
Kolkata	2017	350	No	?
Kolkata	2019	465	Yes	?
Kolkata	2015	572	No	?
Kolkata	2012	655	Yes	?
Kolkata	2012	735	No	?
Kolkata	2017	881	Yes	?
Kolkata	2011	1123	Yes	?
Kolkata	2014	1210	Yes	?

Data lakes

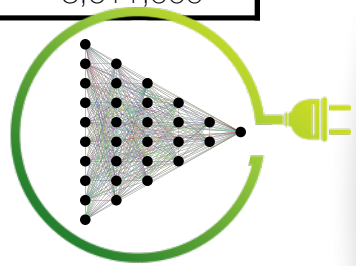
Enterprise data

Online data repos

Data marketplaces

Dataset Discovery

City	Year	Area	Security	Price
Kolkata	2009	710	No	3,200,000
Kolkata	2013	770	No	3,850,000
Kolkata	2007	935	No	2,524,000
Kolkata	2006	973	Yes	3,611,000



City	Year	Area	Security	Price
Kolkata	2017	350	No	?
Kolkata	2019	465	Yes	?
Kolkata	2015	572	No	?
Kolkata	2012	655	Yes	?
Kolkata	2012	735	No	?
Kolkata	2017	881	Yes	?
Kolkata	2011	1123	Yes	?
Kolkata	2014	1210	Yes	?

Candidate datasets

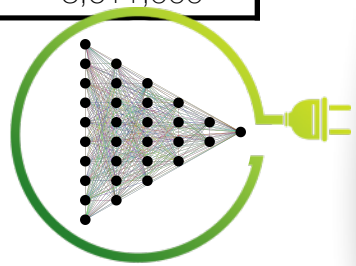
	City	Year	Area	Security	Swimming Pool	Garage	Price
r1	Bangalore	2017	1210	Yes	Yes	Yes	5,700,000
r2	Bangalore	2018	3340	Yes	Yes	No	30,000,000
r3	Bangalore	2016	2502	Yes	Yes	Yes	20,000,000
r4	Bangalore	2009	2293	Yes	Yes	No	9,630,000

	City	Year	Area	Security	Resale	Garden	Gym	Price
s1	Mumbai	2019	425	Yes	Yes	No	No	5,500,000
s2	Mumbai	2013	720	No	Yes	No	No	4,850,000
s3	Mumbai	2018	1060	Yes	Yes	Yes	Yes	11,000,000
s4	Mumbai	2016	1680	Yes	Yes	No	Yes	15,000,000

	City	Year	Area	24*7 Security	Intercom	Price
t1	Delhi	2007	385	No	No	3,300,000
t2	Delhi	2009	435	No	Yes	2,500,000
t3	Delhi	2014	600	No	No	12,500,000
t4	Delhi	2004	900	No	No	5,800,000

Dataset Discovery

City	Year	Area	Security	Price
Kolkata	2009	710	No	3,200,000
Kolkata	2013	770	No	3,850,000
Kolkata	2007	935	No	2,524,000
Kolkata	2006	973	Yes	3,611,000



City	Year	Area	Security	Price
Kolkata	2017	350	No	?
Kolkata	2019	465	Yes	?
Kolkata	2015	572	No	?
Kolkata	2012	655	Yes	?
Kolkata	2012	735	No	?
Kolkata	2017	881	Yes	?
Kolkata	2011	1123	Yes	?
Kolkata	2014	1210	Yes	?

Candidate datasets

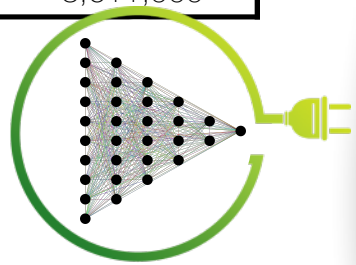
	City	Year	Area	Security	Swimming Pool	Garage	Price
r1	Bangalore	2017	1210	Yes	Yes	Yes	5,700,000
r2	Bangalore	2018	3340	Yes	Yes	No	30,000,000
r3	Bangalore	2016	2502	Yes	Yes	Yes	20,000,000
r4	Bangalore	2009	2293	Yes	Yes	No	9,630,000

	City	Year	Area	Security	Resale	Garden	Gym	Price
s1	Mumbai	2019	425	Yes	Yes	No	No	5,500,000
s2	Mumbai	2013	720	No	Yes	No	No	4,850,000
s3	Mumbai	2018	1060	Yes	Yes	Yes	Yes	11,000,000
s4	Mumbai	2016	1680	Yes	Yes	No	Yes	15,000,000

	City	Year	Area	24*7 Security	Intercom	Price
t1	Delhi	2007	385	No	No	3,300,000
t2	Delhi	2009	435	No	Yes	2,500,000
t3	Delhi	2014	600	No	No	12,500,000
t4	Delhi	2004	900	No	No	5,800,000

Dataset Discovery

City	Year	Area	Security	Price
Kolkata	2009	710	No	3,200,000
Kolkata	2013	770	No	3,850,000
Kolkata	2007	935	No	2,524,000
Kolkata	2006	973	Yes	3,611,000



City	Year	Area	Security	Price
Kolkata	2017	350	No	?
Kolkata	2019	465	Yes	?
Kolkata	2015	572	No	?
Kolkata	2012	655	Yes	?
Kolkata	2012	735	No	?
Kolkata	2017	881	Yes	?
Kolkata	2011	1123	Yes	?
Kolkata	2014	1210	Yes	?

Candidate datasets

Heterogeneous

	City	Year	Area	Security	Swimming Pool	Garage	Price
r1	Bangalore	2017	1210	Yes		Yes	
r2	Bangalore	2018	3340		Yes	No	
r3	Bangalore	2016	2502				
r4	Bangalore	2009	2293		Yes	No	9,630,000

	City	Year	Area	Price
s1	Mumbai	2019	425	
s2	Mumbai	2013	720	
s3	Mumbai	2018	1060	
s4	Mumbai	2016	1680	15,000,000

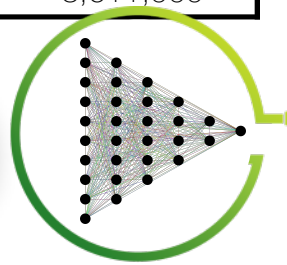
	City	Area	24*7 Security	Intercom	Price
t1		385	No	No	3,300,000
t2	2009	435	No	Yes	2,500,000
t3	Delhi	2014	600	No	No	12,500,000
t4	Delhi	2004	900	No	No	5,800,000

Where is relative good data

Dataset Discovery

City	Year	Area	Security	Price
Kolkata	2009	710	No	3,200,000
Kolkata	2013	770	No	3,850,000
Kolkata	2007	935	No	2,524,000
Kolkata	2006	973	Yes	3,611,000

Model charging



City	Year	Area	Security	Price
Kolkata	2017	350	No	?
Kolkata	2019	465	Yes	?
Kolkata	2015	572	No	?
Kolkata	2012	655	Yes	?
Kolkata	2012	735	No	?
Kolkata	2017	881	Yes	?
Kolkata	2011	1123	Yes	?
Kolkata	2014	1210	Yes	?

Candidate datasets

Heterogeneous

	City	Year	Area	Security	Swimming Pool	Garage	Price
r1	Bangalore	2017	1210	Yes		Yes	
r2	Bangalore	2018	3340		Yes	No	
r3	Bangalore	2016	2502				
r4	Bangalore	2009	2293		Yes	No	9,630,000

	City	Year	Area	Price
s1	Mumbai	2019	425	
s2	Mumbai	2013	720	
s3	Mumbai	2018	1060	
s4	Mumbai	2016	1680	15,000,000

	City	Year	Area	24*7 Security	Intercom	Price
t1			385	No	No	3,300,000
t2		2009	435	No	Yes	2,500,000
t3	Delhi	2014	600	No	No	12,500,000
t4	Delhi	2004	900	No	No	5,800,000

Where is relative good data

Selective Data Acquisition in the Wild for Model Charging

Chengliang Chai
Tsinghua University
Beijing, China
ccl@mail.tsinghua.edu.cn

Jiabin Liu
Tsinghua University
Beijing, China
liujb19@mails.tsinghua.edu.cn

Nan Tang
QCRI
Doha, Qatar
ntang@hbku.edu.qa

Guoliang Li
Tsinghua University
Beijing, China
liguoliang@tsinghua.edu.cn

Yuyu Luo
Tsinghua University
Beijing, China
luoyy18@mails.tsinghua.edu.cn

ABSTRACT

The lack of sufficient labeled data is a key bottleneck for practitioners in many real-world supervised machine learning (ML) tasks. In this paper, we study a new problem, namely *selective data acquisition in the wild for model charging*: given a supervised ML task and data in the wild (*e.g.*, enterprise data warehouses, online data repositories, data markets, and so on), the problem is to select labeled data points from the data in the wild as additional train data that

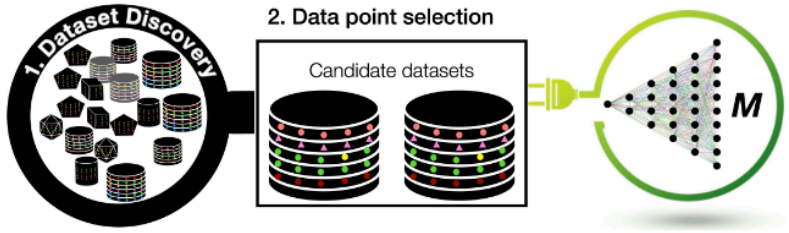
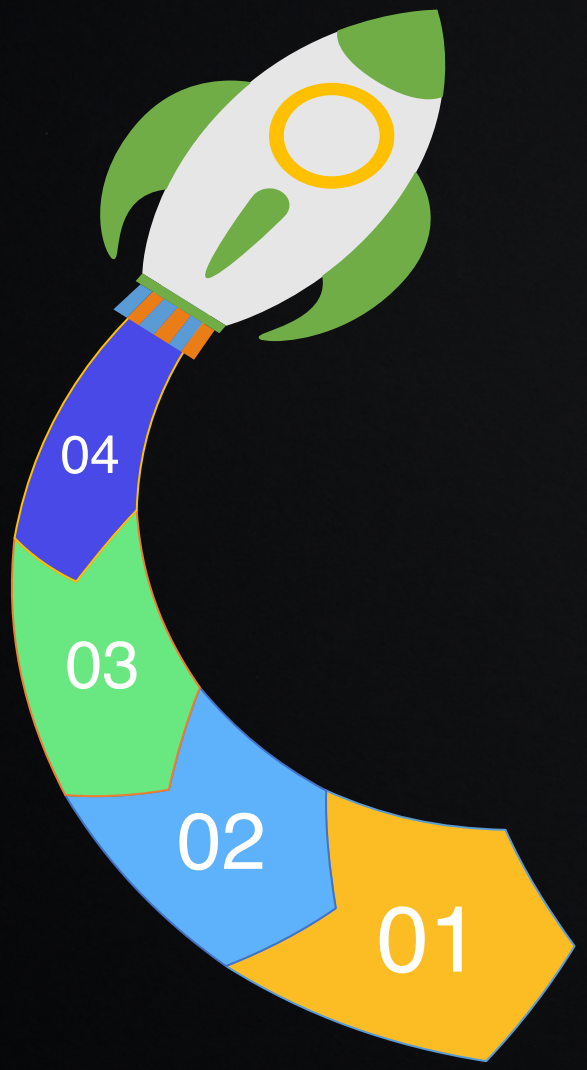


Figure 1: Selective data acquisition for model charging.



01
Data Acquisition

- image data
- tabular
- text data

02
Data Understanding

- exploratory data analysis
- data visualization

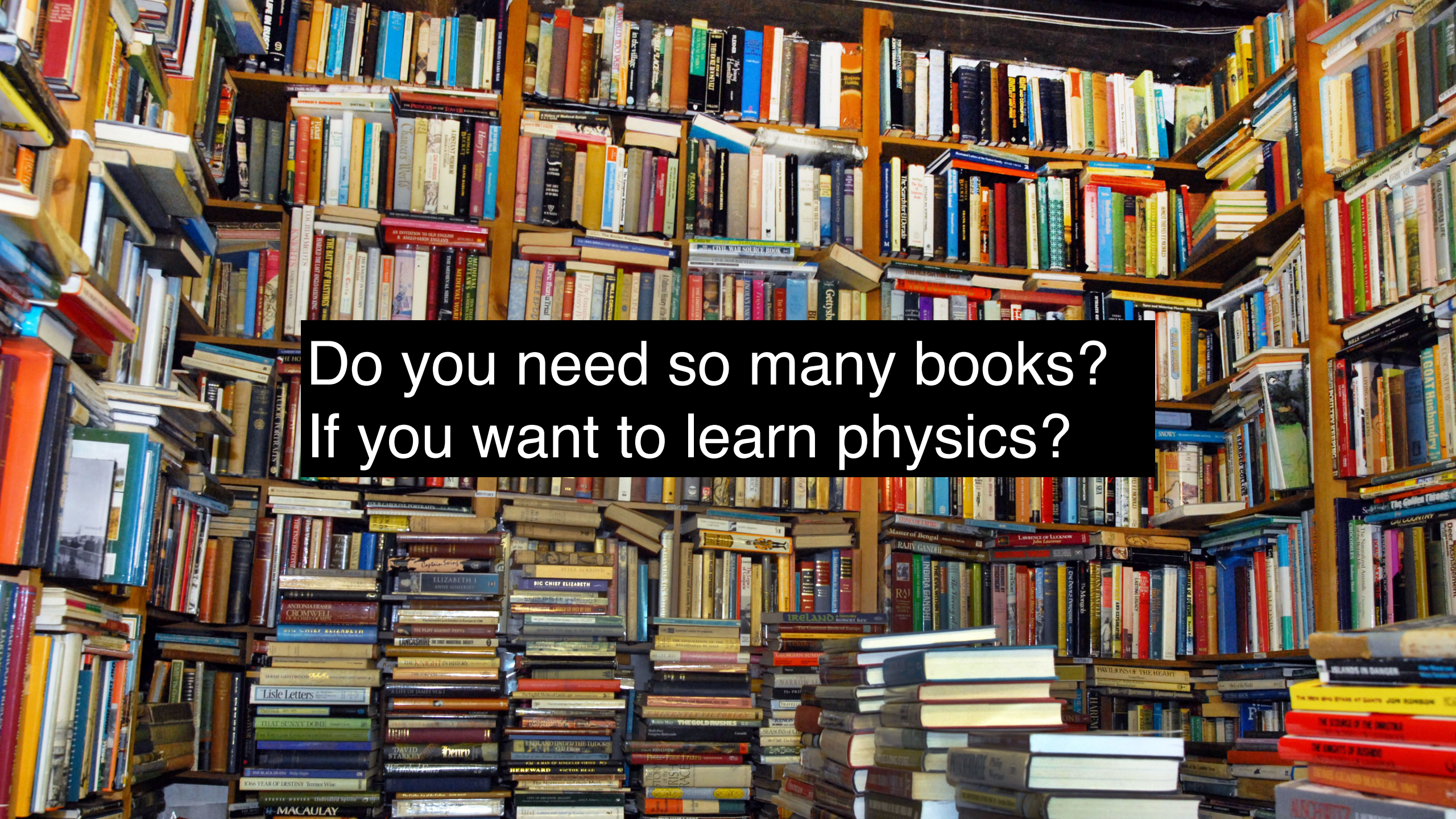
03
Data Preparation

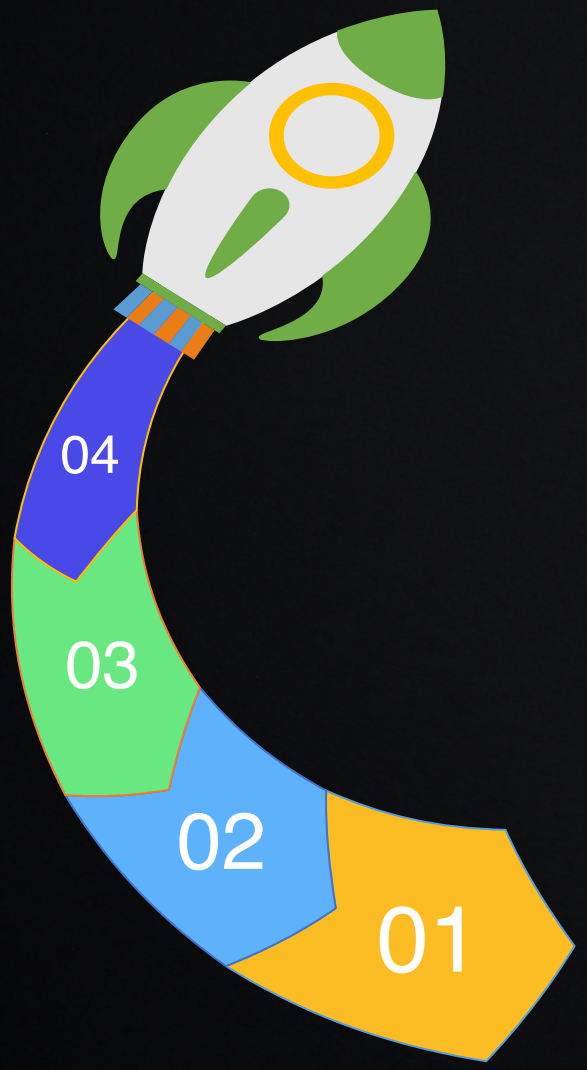
- data transformation
- data cleaning
- entity resolution
- … …

04
Advanced Topics

- data-centric AI
- HPC for LLMs
- … …


CIVIL WAR SOURCE BOOK
Gettysburg July 1
BLOODY HILL
THE HOUSE OF TUDOR
AN INVITATION TO OLD ENGLISH & ANGLO-SAXON ENGLAND
THE BATTLE OF HASTINGS
Native American
WASHINGTON
AFRICA
SNOWY
Master of Bengal
RAJIV GANDHI
INDIRA GANDHI
ELIZABETH I
BIG CHIEF ELIZABETH
CROMWELL
IRELAND
Lisle Letters
THAT SUNNY DOME
THE GOLD RUSHES
MACAULAY
PAVILIONS OF THE HEART
1066 YEAR OF DESTINY


Do you need so many books?
If you want to learn physics?



01
Data Acquisition

- image data
- tabular
- text data

02
Data Understanding

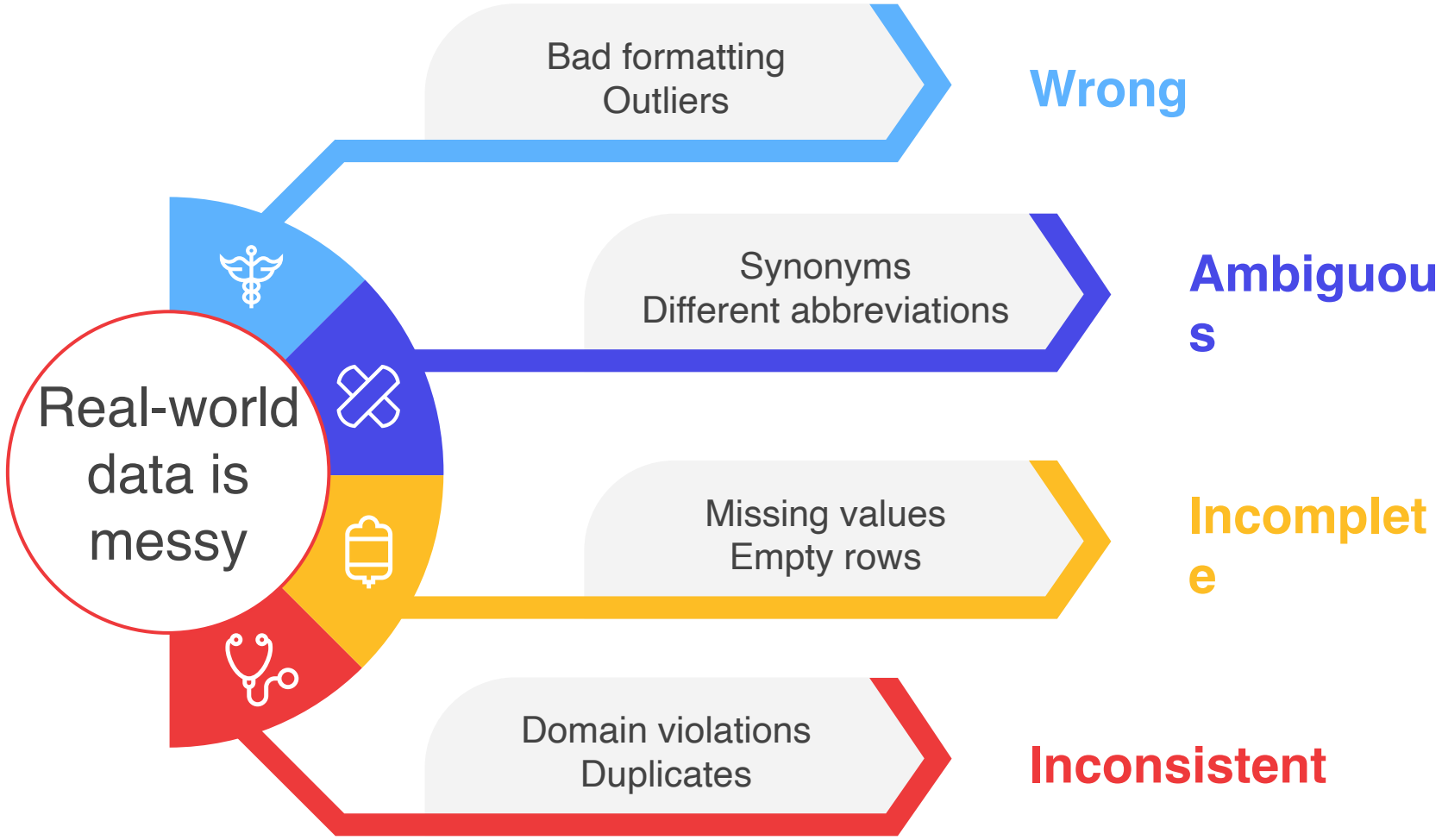
- exploratory data analysis
- data visualization

03
Data Preparation

- data transformation
- data cleaning
- entity resolution
- … …

04
Advanced Topics

- data-centric AI
- HPC for LLMs
- … …

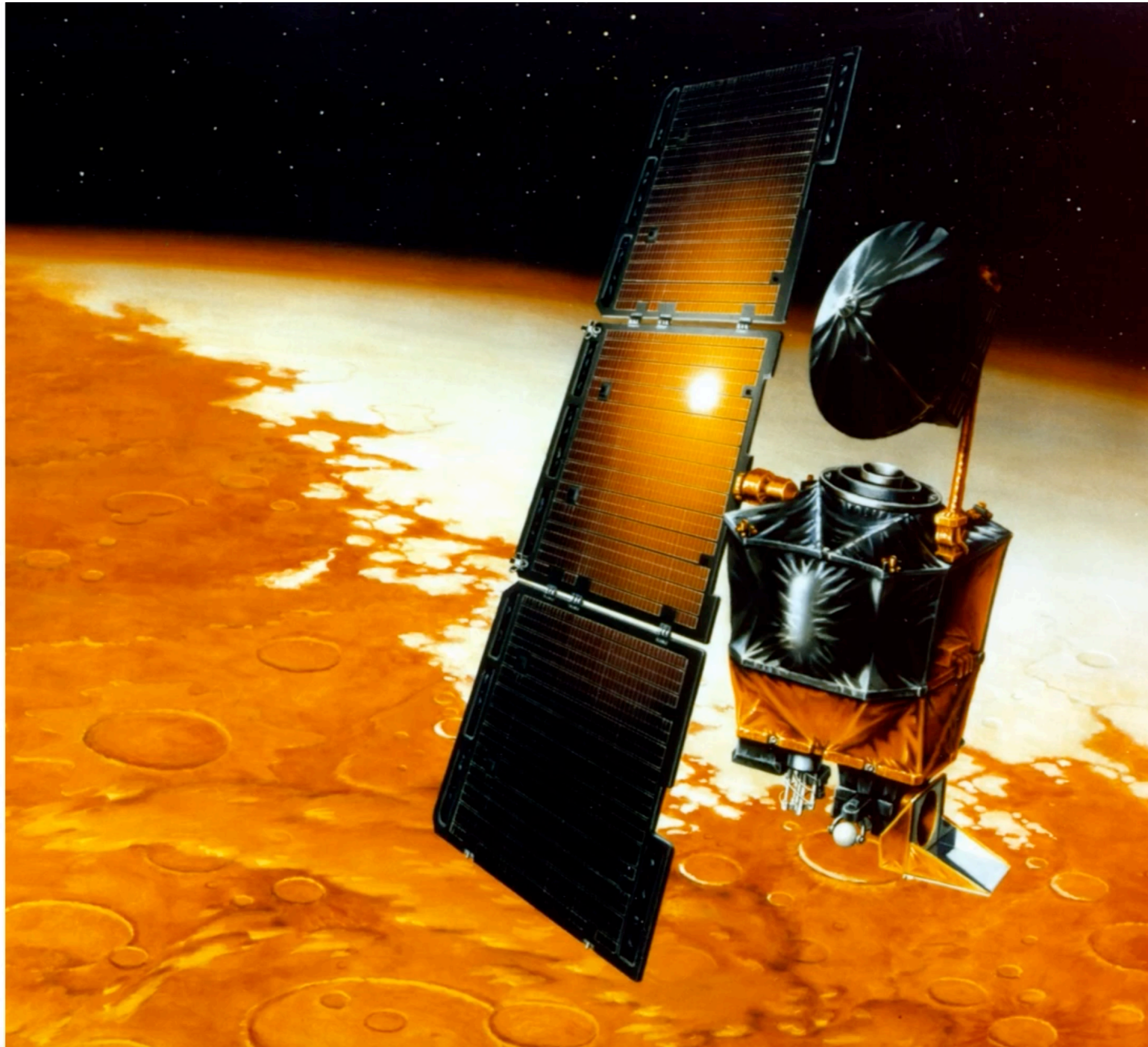

Real-world data is messy
Bad formatting
Outliers
Wrong
Synonyms
Different abbreviations
Ambiguous
Missing values
Empty rows
Incomplete
Domain violations
Duplicates
Inconsistent

linical research, the error rates
be between 2.3% to 26.

linical research, the error rates
be between 2.3% to 26.

NASA Lost a Spacecraft Due to a Metric Math Mistake

In 1999, NASA lost a $193M Mars orbiter because an engineering team failed to convert measurements from English to metric units.



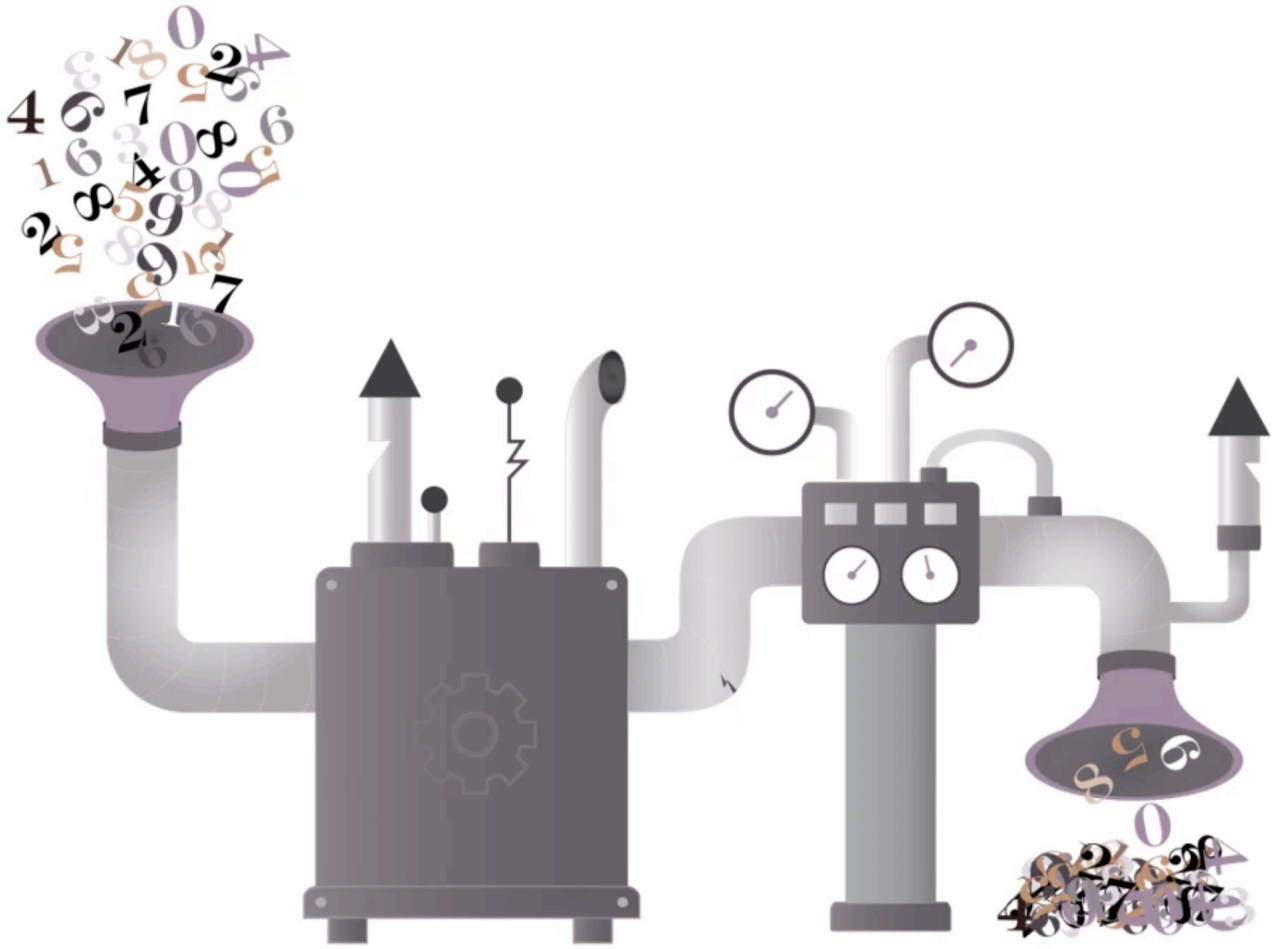

NASA-JPL


NASA-JPL


NASA-JPL

Data Exploration

- Data exploration should be done before any modeling work
- Many modeling tools do not work on messy data
  - E.g., a modeling approach can fail if it encounters a non-numeric value for a numeric variable
- Models based on messy data can be misleading
  - Garbage in, garbage out



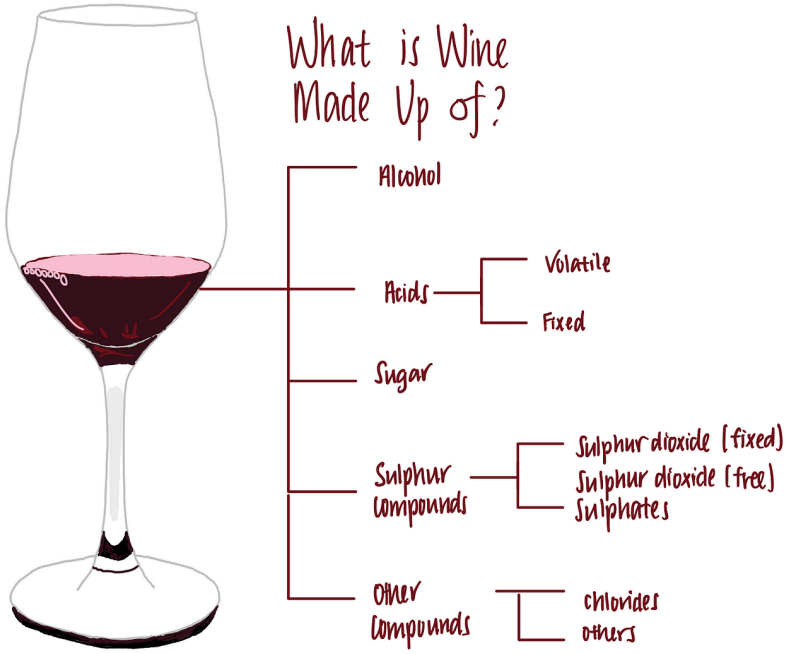
Purposes of Data Exploration

- To gain intuition about the data
- To conduct sanity checks: are the data in the right format and on the right scale?
- To find out where data is missing or if there are outliers
- To summarize the data.

Wine quality (UCI ML data repo)

- from the north of Portugal

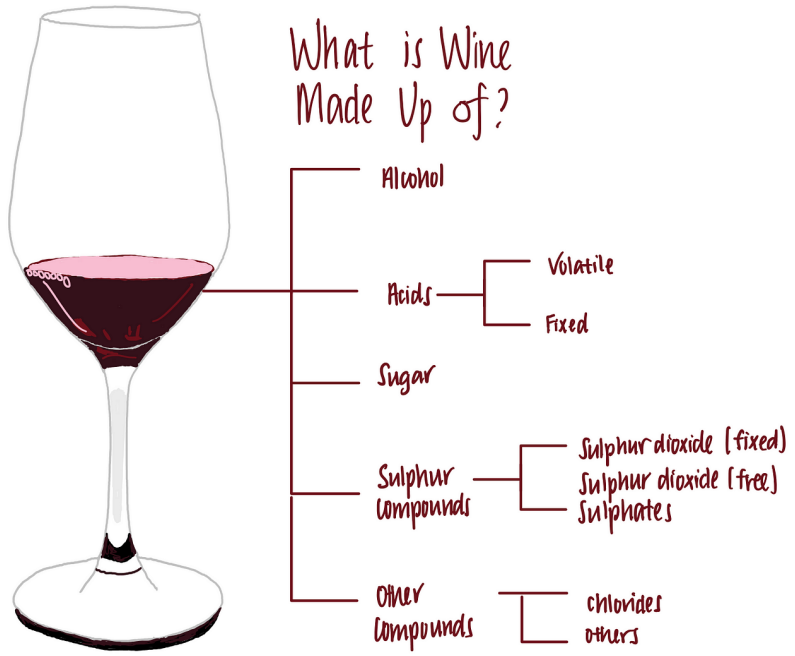
- goal: model wine quality based on physicochemical test



Wine quality (UCI ML data repo)

- from the north of Portugal

- goal: model wine quality based on physicochemical test



winequality-red.csv

```
"fixed acidity";"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5
7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5
11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.4;0.66;0;1.8;0.075;13;40;0.9978;3.51;0.56;9.4;5
7.9;0.6;0.06;1.6;0.069;15;59;0.9964;3.3;0.46;9.4;5
7.3;0.65;0;1.2;0.065;15;21;0.9946;3.39;0.47;10;7
7.8;0.58;0.02;2;0.073;9;18;0.9968;3.36;0.57;9.5;7
7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5
6.7;0.58;0.08;1.8;0.097;15;65;0.9959;3.28;0.54;9.2;5
7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5
5.6;0.615;0;1.6;0.089;16;59;0.9943;3.58;0.52;9.9;5
7.8;0.61;0.29;1.6;0.114;9;29;0.9974;3.26;1.56;9.1;5
8.9;0.62;0.18;3.8;0.176;52;145;0.9986;3.16;0.88;9.2;5
8.9;0.62;0.19;3.9;0.17;51;148;0.9986;3.17;0.93;9.2;5
8.5;0.28;0.56;1.8;0.092;35;103;0.9969;3.3;0.75;10.5;7
8.1;0.56;0.28;1.7;0.368;16;56;0.9968;3.11;1.28;9.3;5
7.4;0.59;0.08;4.4;0.086;6;29;0.9974;3.38;0.5;9;4
7.9;0.32;0.51;1.8;0.341;17;56;0.9969;3.04;1.08;9.2;6
8.9;0.22;0.48;1.8;0.077;29;60;0.9968;3.39;0.53;9.4;6
7.6;0.39;0.31;2.3;0.082;23;71;0.9982;3.52;0.65;9.7;5
7.9;0.43;0.21;1.6;0.106;10;37;0.9966;3.17;0.91;9.5;5
8.5;0.49;0.11;2.3;0.084;9;67;0.9968;3.17;0.53;9.4;5
6.9;0.4;0.14;2.4;0.085;21;40;0.9968;3.43;0.63;9.7;6
6.3;0.39;0.16;1.4;0.08;11;23;0.9955;3.34;0.56;9.3;5
7.6;0.41;0.24;1.8;0.08;4;11;0.9962;3.28;0.59;9.5;5
7.9;0.43;0.21;1.6;0.106;10;37;0.9966;3.17;0.91;9.5;5
7.1;0.71;0;1.9;0.08;14;35;0.9972;3.47;0.55;9.4;5
7.8;0.645;0;2;0.082;8;16;0.9964;3.38;0.59;9.8;6
6.7;0.675;0.07;2.4;0.089;17;82;0.9958;3.35;0.54;10.1;5
6.9;0.685;0;2.5;0.105;22;37;0.9966;3.46;0.57;10.6;6
8.3;0.655;0.12;2.3;0.083;15;113;0.9966;3.17;0.66;9.8;5
6.9;0.605;0.12;10.7;0.073;40;83;0.9993;3.45;0.52;9.4;6
5.2;0.32;0.25;1.8;0.103;13;50;0.9957;3.38;0.55;9.2;5
7.8;0.645;0;5.5;0.086;5;18;0.9986;3.4;0.55;9.6;6
7.8;0.6;0.14;2.4;0.086;3;15;0.9975;3.42;0.6;10.8;6
8.1;0.38;0.28;2.1;0.066;13;30;0.9968;3.23;0.73;9.7;7
5.7;1.13;0.09;1.5;0.172;7;19;0.994;3.5;0.48;9.8;4
7.3;0.45;0.36;5.9;0.074;12;87;0.9978;3.33;0.83;10.5;5
7.3;0.45;0.36;5.9;0.074;12;87;0.9978;3.33;0.83;10.5;5
8.8;0.61;0.3;2.8;0.088;17;46;0.9976;3.26;0.51;9.3;4
7.5;0.49;0.2;2.6;0.332;8;14;0.9968;3.21;0.9;10.5;6
8.1;0.66;0.22;2.2;0.069;9;23;0.9968;3.3;1.2;10.3;5
6.8;0.67;0.02;1.8;0.05;5;11;0.9962;3.48;0.52;9.5;5
4.6;0.52;0.15;2.1;0.054;8;65;0.9934;3.9;0.56;13.1;4
7.7;0.935;0.43;2.2;0.114;22;114;0.997;3.25;0.73;9.2;5
8.7;0.29;0.52;1.6;0.113;12;37;0.9969;3.25;0.58;9.5;5
6.4;0.4;0.23;1.6;0.066;5;12;0.9958;3.34;0.56;9.2;5
5.6;0.31;0.37;1.4;0.074;12;96;0.9954;3.32;0.58;9.2;5
8.8;0.66;0.26;1.7;0.074;4;23;0.9971;3.15;0.74;9.2;5
6.6;0.52;0.04;2.2;0.069;8;15;0.9956;3.4;0.63;9.4;6
6.6;0.5;0.04;2.1;0.068;6;14;0.9955;3.39;0.64;9.4;6
8.6;0.38;0.36;3;0.081;30;119;0.997;3.2;0.56;9.4;5
7.6;0.51;0.15;2.8;0.11;33;73;0.9955;3.17;0.63;10.2;6
7.7;0.62;0.04;3.8;0.084;25;45;0.9978;3.34;0.53;9.5;5
10.2;0.42;0.57;3.4;0.07;4;10;0.9971;3.04;0.63;9.6;5
7.5;0.63;0.12;5.1;0.111;50;110;0.9983;3.26;0.77;9.4;5
7.8;0.59;0.18;2.3;0.076;17;54;0.9975;3.43;0.59;10;5
7.3;0.39;0.31;2.4;0.074;9;46;0.9962;3.41;0.54;9.4;6
8.8;0.4;0.4;2.2;0.079;19;52;0.998;3.44;0.64;9.2;5
7.7;0.69;0.49;1.8;0.115;20;112;0.9968;3.21;0.71;9.3;5
7.5;0.52;0.16;1.9;0.085;12;35;0.9968;3.38;0.62;9.5;7
7;0.735;0.05;2;0.081;13;54;0.9966;3.39;0.57;9.8;5
7.2;0.725;0.05;4.65;0.086;4;11;0.9962;3.41;0.39;10.9;5
7.2;0.725;0.05;4.65;0.086;4;11;0.9962;3.41;0.39;10.9;5
```

Take a closer look at the data

In [2]:
```
df = pd.read_csv('winequality-white.csv',sep=';')
df.head()
```

Out[2]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Take a closer look at the data

In [2]:
```
df = pd.read_csv('winequality-white.csv',sep=';')
df.head()
```

Out[2]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

The number of rows and columns

In [3]:
```
df.shape
```

Out[3]: (4898, 12)

Data types, contain null values or not

```
In [5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
fixed acidity           4898 non-null float64
volatile acidity        4898 non-null float64
citric acid             4898 non-null float64
residual sugar          4898 non-null float64
chlorides               4898 non-null float64
free sulfur dioxide     4898 non-null float64
total sulfur dioxide    4898 non-null float64
density                 4898 non-null float64
pH                      4898 non-null float64
sulphates               4898 non-null float64
alcohol                 4898 non-null float64
quality                 4898 non-null int64
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```

Data types, contain null values or not

```
In [5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
fixed acidity           4898 non-null float64
volatile acidity        4898 non-null float64
citric acid             4898 non-null float64
residual sugar          4898 non-null float64
chlorides               4898 non-null float64
free sulfur dioxide     4898 non-null float64
total sulfur dioxide    4898 non-null float64
density                 4898 non-null float64
pH                      4898 non-null float64
sulphates               4898 non-null float64
alcohol                 4898 non-null float64
quality                 4898 non-null int64
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```

Data types, contain null values or not

```
In [5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
fixed acidity           4898 non-null float64
volatile acidity        4898 non-null float64
citric acid             4898 non-null float64
residual sugar          4898 non-null float64
chlorides               4898 non-null float64
free sulfur dioxide     4898 non-null float64
total sulfur dioxide    4898 non-null float64
density                 4898 non-null float64
pH                      4898 non-null float64
sulphates               4898 non-null float64
alcohol                 4898 non-null float64
quality                 4898 non-null int64
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```

The describe() function in pandas is very handy in getting various summary statistics

```
In [6]: df.describe()
```

Out[6]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000

The describe() function in pandas is very handy in getting various summary statistics

```
In [6]: df.describe()
```

Out[6]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000


NO. 1
NO. 2
NO. 3
NO. 4
NO. 5
NO. 6
NO. 7
NO. 8
NO. 9
NO. 10

```
In [7]: df.quality.unique()

Out[7]: array([6, 5, 7, 8, 4, 3, 9], dtype=int64)
```

```
In [7]: df.quality.unique()

Out[7]: array([6, 5, 7, 8, 4, 3, 9], dtype=int64)
```

Vote counts?

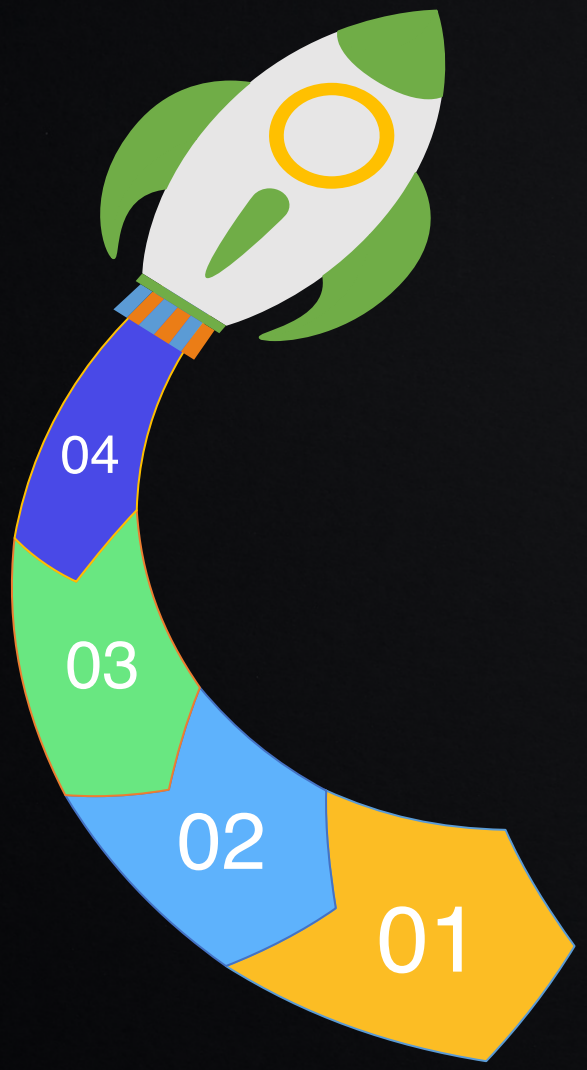
```
In [8]: df.quality.value_counts()

Out[8]: 6    2198
        5    1457
        7     880
        8     175
        4     163
        3      20
        9       5
        Name: quality, dtype: int64
```

The more you get involved

Datasets with similar statistics may look very different

Datasets with similar statistics may look very different

Datasets with similar statistics may look very different



01
Data Acquisition

- image data
- tabular
- text data

02
Data Understanding

- exploratory data analysis
- data visualization

03
Data Preparation

- data transformation
- data cleaning
- entity resolution
- … …

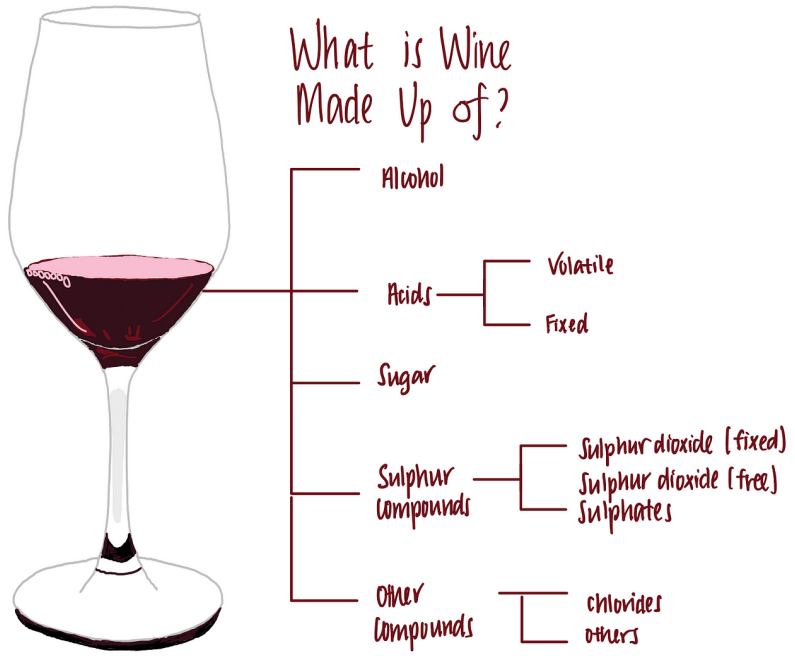
04
Advanced Topics

- data-centric AI
- HPC for LLMs
- … …

Wine quality (UCI ML data repo)

- from the north of Portugal

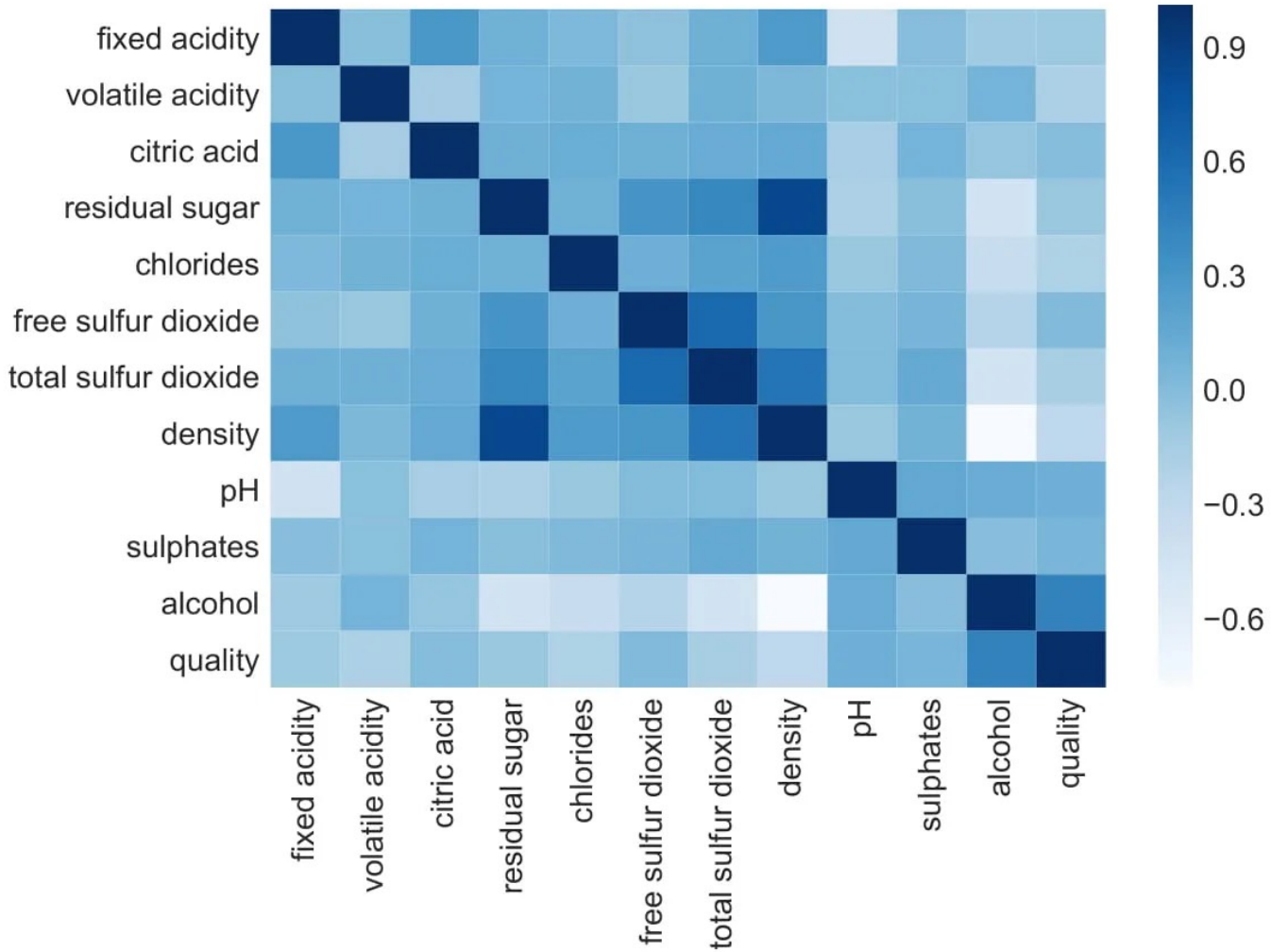
- goal: model wine quality based on physicochemical test



winequality-red.csv

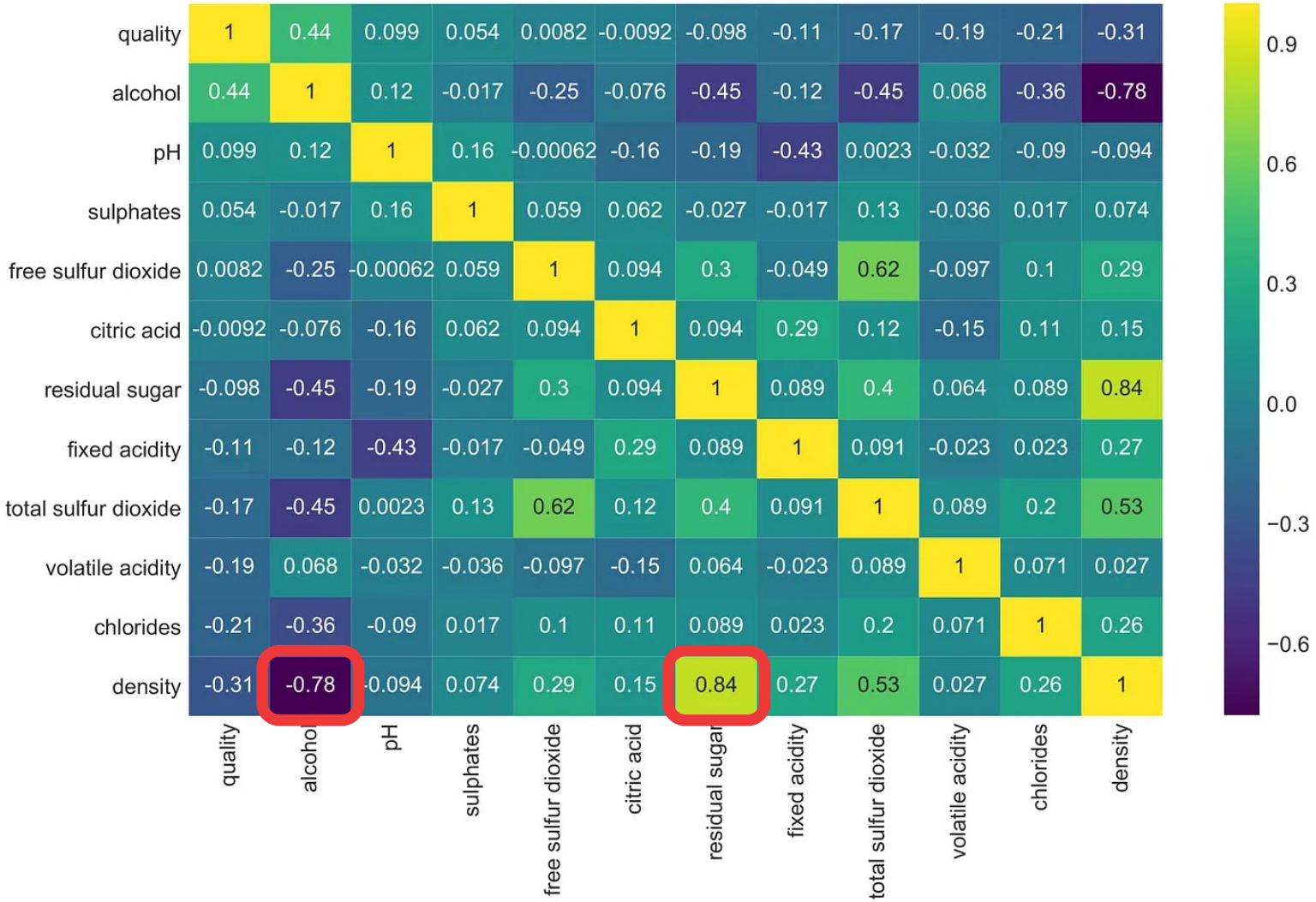
```
"fixed acidity";"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5
7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5
11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.4;0.66;0;1.8;0.075;13;40;0.9978;3.51;0.56;9.4;5
7.9;0.6;0.06;1.6;0.069;15;59;0.9964;3.3;0.46;9.4;5
7.3;0.65;0;1.2;0.065;15;21;0.9946;3.39;0.47;10;7
7.8;0.58;0.02;2;0.073;9;18;0.9968;3.36;0.57;9.5;7
7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5
6.7;0.58;0.08;1.8;0.097;15;65;0.9959;3.28;0.54;9.2;5
7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5
5.6;0.615;0;1.6;0.089;16;59;0.9943;3.58;0.52;9.9;5
7.8;0.61;0.29;1.6;0.114;9;29;0.9974;3.26;1.56;9.1;5
8.9;0.62;0.18;3.8;0.176;52;145;0.9986;3.16;0.88;9.2;5
8.9;0.62;0.19;3.9;0.17;51;148;0.9986;3.17;0.93;9.2;5
8.5;0.28;0.56;1.8;0.092;35;103;0.9969;3.3;0.75;10.5;7
8.1;0.56;0.28;1.7;0.368;16;56;0.9968;3.11;1.28;9.3;5
7.4;0.59;0.08;4.4;0.086;6;29;0.9974;3.38;0.5;9;4
7.9;0.32;0.51;1.8;0.341;17;56;0.9969;3.04;1.08;9.2;6
8.9;0.22;0.48;1.8;0.077;29;60;0.9968;3.39;0.53;9.4;6
7.6;0.39;0.31;2.3;0.082;23;71;0.9982;3.52;0.65;9.7;5
7.9;0.43;0.21;1.6;0.106;10;37;0.9966;3.17;0.91;9.5;5
8.5;0.49;0.11;2.3;0.084;9;67;0.9968;3.17;0.53;9.4;5
6.9;0.4;0.14;2.4;0.085;21;40;0.9968;3.43;0.63;9.7;6
6.3;0.39;0.16;1.4;0.08;11;23;0.9955;3.34;0.56;9.3;5
7.6;0.41;0.24;1.8;0.08;4;11;0.9962;3.28;0.59;9.5;5
7.9;0.43;0.21;1.6;0.106;10;37;0.9966;3.17;0.91;9.5;5
7.1;0.71;0;1.9;0.08;14;35;0.9972;3.47;0.55;9.4;5
7.8;0.645;0;2;0.082;8;16;0.9964;3.38;0.59;9.8;6
6.7;0.675;0.07;2.4;0.089;17;82;0.9958;3.35;0.54;10.1;5
6.9;0.685;0;2.5;0.105;22;37;0.9966;3.46;0.57;10.6;6
8.3;0.655;0.12;2.3;0.083;15;113;0.9966;3.17;0.66;9.8;5
6.9;0.605;0.12;10.7;0.073;40;83;0.9993;3.45;0.52;9.4;6
5.2;0.32;0.25;1.8;0.103;13;50;0.9957;3.38;0.55;9.2;5
7.8;0.645;0;5.5;0.086;5;18;0.9986;3.4;0.55;9.6;6
7.8;0.6;0.14;2.4;0.086;3;15;0.9975;3.42;0.6;10.8;6
8.1;0.38;0.28;2.1;0.066;13;30;0.9968;3.23;0.73;9.7;7
5.7;1.13;0.09;1.5;0.172;7;19;0.994;3.5;0.48;9.8;4
7.3;0.45;0.36;5.9;0.074;12;87;0.9978;3.33;0.83;10.5;5
7.3;0.45;0.36;5.9;0.074;12;87;0.9978;3.33;0.83;10.5;5
8.8;0.61;0.3;2.8;0.088;17;46;0.9976;3.26;0.51;9.3;4
7.5;0.49;0.2;2.6;0.332;8;14;0.9968;3.21;0.9;10.5;6
8.1;0.66;0.22;2.2;0.069;9;23;0.9968;3.3;1.2;10.3;5
6.8;0.67;0.02;1.8;0.05;5;11;0.9962;3.48;0.52;9.5;5
4.6;0.52;0.15;2.1;0.054;8;65;0.9934;3.9;0.56;13.1;4
7.7;0.935;0.43;2.2;0.114;22;114;0.997;3.25;0.73;9.2;5
8.7;0.29;0.52;1.6;0.113;12;37;0.9969;3.25;0.58;9.5;5
6.4;0.4;0.23;1.6;0.066;5;12;0.9958;3.34;0.56;9.2;5
5.6;0.31;0.37;1.4;0.074;12;96;0.9954;3.32;0.58;9.2;5
8.8;0.66;0.26;1.7;0.074;4;23;0.9971;3.15;0.74;9.2;5
6.6;0.52;0.04;2.2;0.069;8;15;0.9956;3.4;0.63;9.4;6
6.6;0.5;0.04;2.1;0.068;6;14;0.9955;3.39;0.64;9.4;6
8.6;0.38;0.36;3;0.081;30;119;0.997;3.2;0.56;9.4;5
7.6;0.51;0.15;2.8;0.11;33;73;0.9955;3.17;0.63;10.2;6
7.7;0.62;0.04;3.8;0.084;25;45;0.9978;3.34;0.53;9.5;5
10.2;0.42;0.57;3.4;0.07;4;10;0.9971;3.04;0.63;9.6;5
7.5;0.63;0.12;5.1;0.111;50;110;0.9983;3.26;0.77;9.4;5
7.8;0.59;0.18;2.3;0.076;17;54;0.9975;3.43;0.59;10;5
7.3;0.39;0.31;2.4;0.074;9;46;0.9962;3.41;0.54;9.4;6
8.8;0.4;0.4;2.2;0.079;19;52;0.998;3.44;0.64;9.2;5
7.7;0.69;0.49;1.8;0.115;20;112;0.9968;3.21;0.71;9.3;5
7.5;0.52;0.16;1.9;0.085;12;35;0.9968;3.38;0.62;9.5;7
7;0.735;0.05;2;0.081;13;54;0.9966;3.39;0.57;9.8;5
7.2;0.725;0.05;4.65;0.086;4;11;0.9962;3.41;0.39;10.9;5
7.2;0.725;0.05;4.65;0.086;4;11;0.9962;3.41;0.39;10.9;5
```

To use linear regression for modeling, it is necessary to remove correlated variables

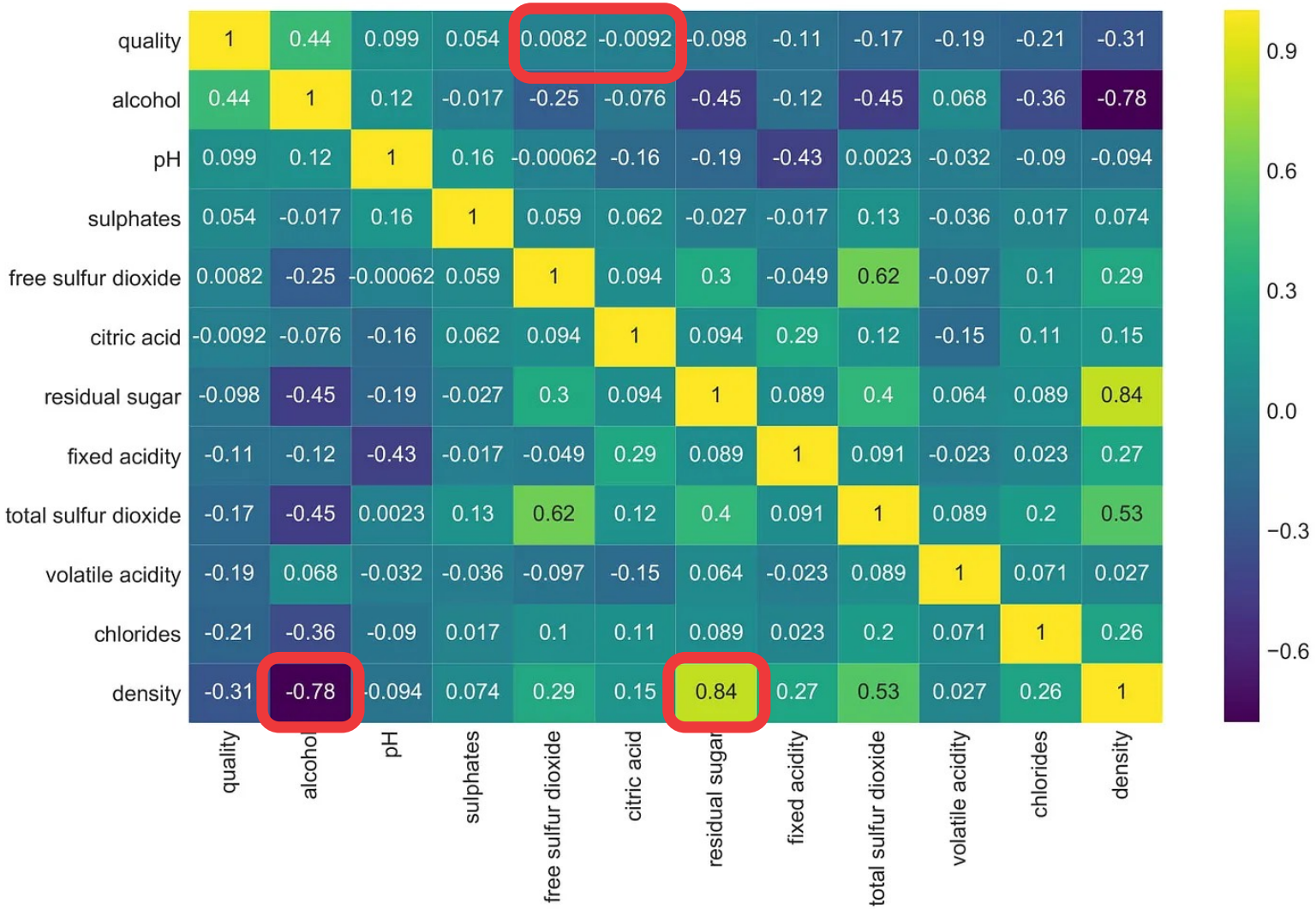


It’s a good practice to remove correlated variables during feature selection

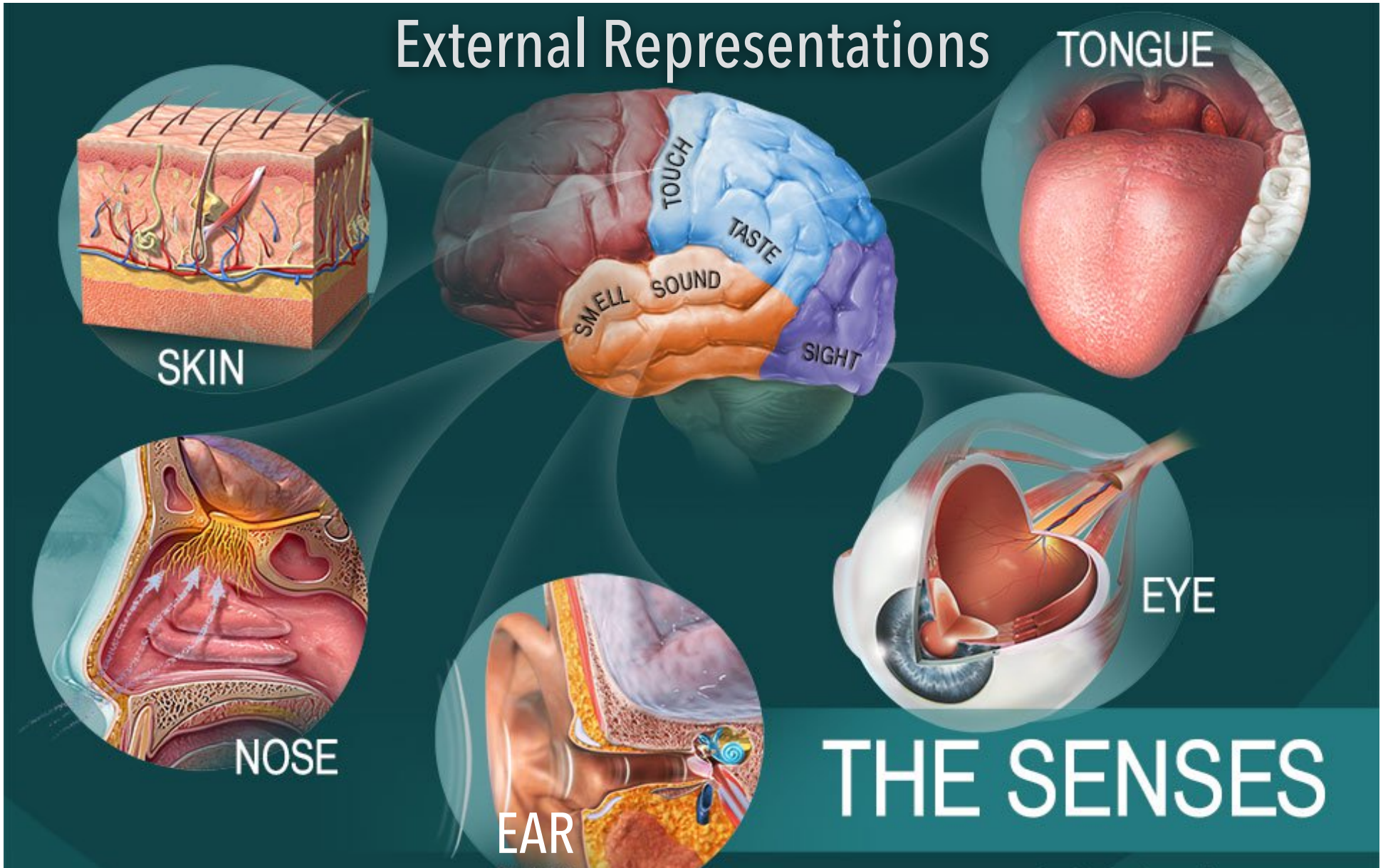
It’s a good practice to remove correlated variables during feature selection



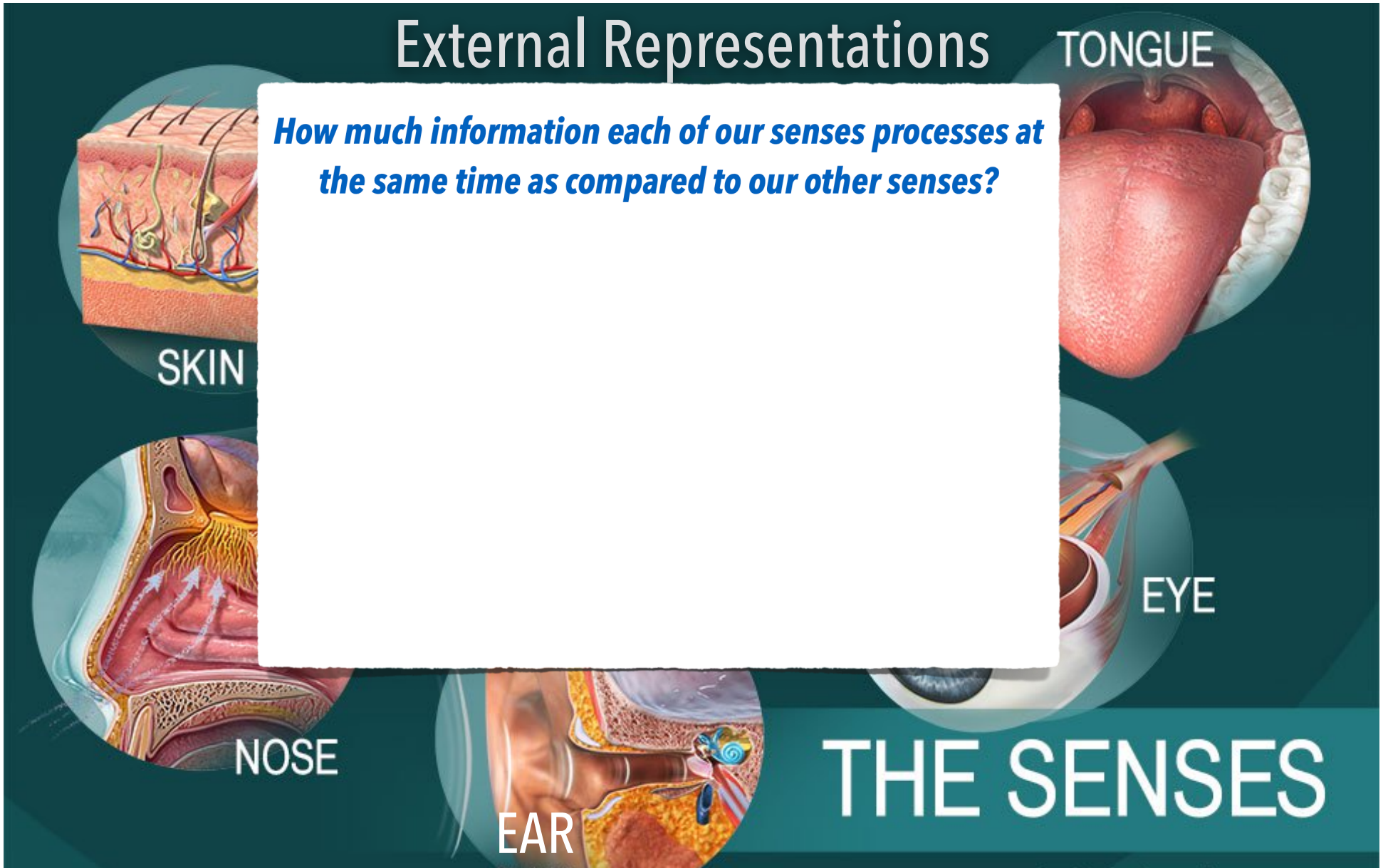
It’s a good practice to remove correlated variables during feature selection



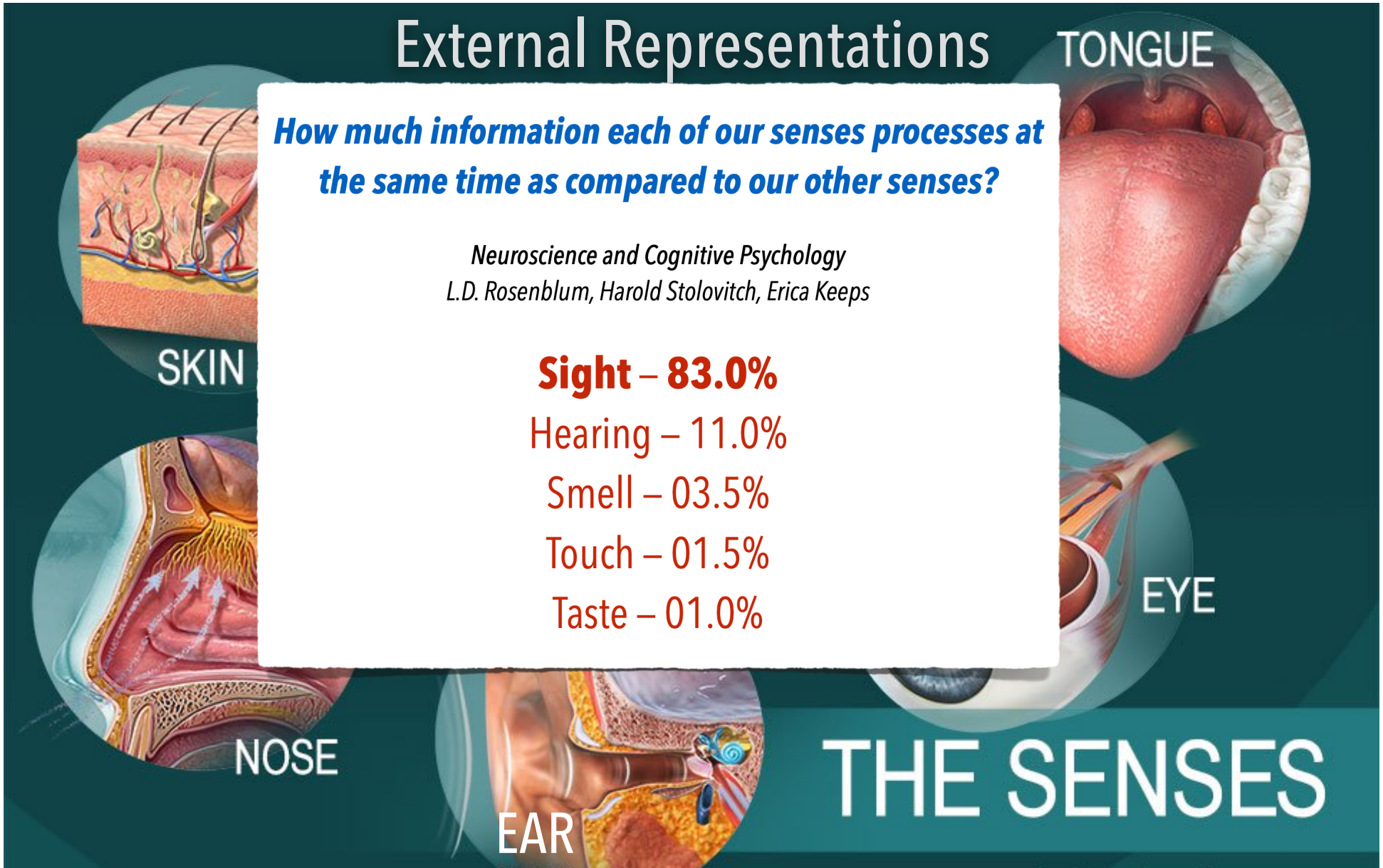
Sight > The Other Senses ?



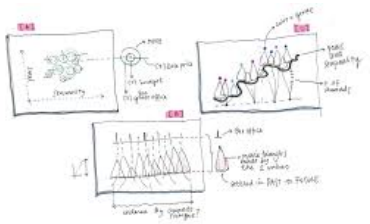
Sight > The Other Senses ?



Sight > The Other Senses ?

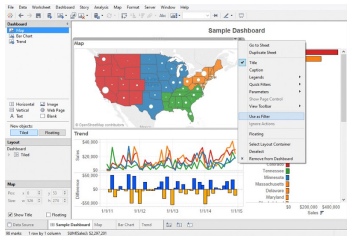


What and how



human

human



machine

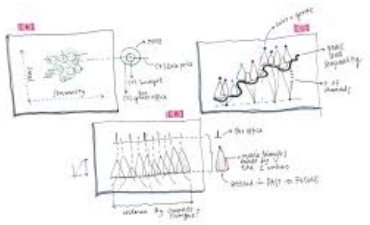
human

machine

X

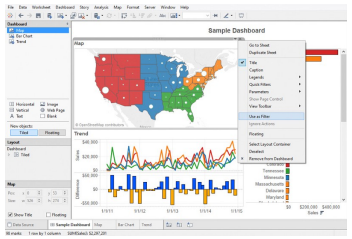
machine

What and how



human

human



machine

human

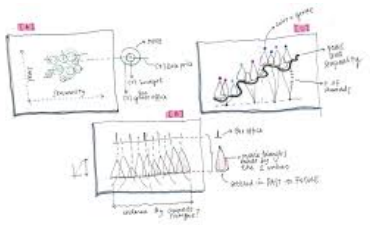
machine

X

machine

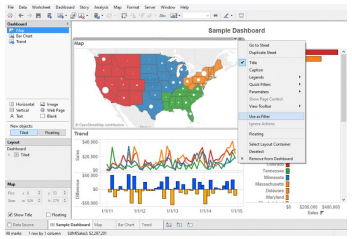
Computer-based visualization systems provide visual representations of **datasets** designed to help **people** carry out **tasks** more **effectively**. *– Tamara Munzner at UBC*

What and how



human

human



machine

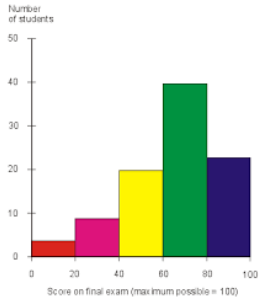
human

machine

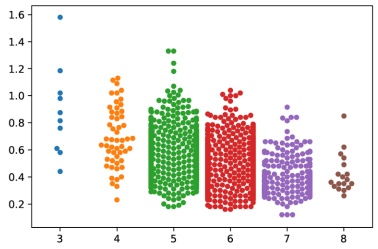
X

machine

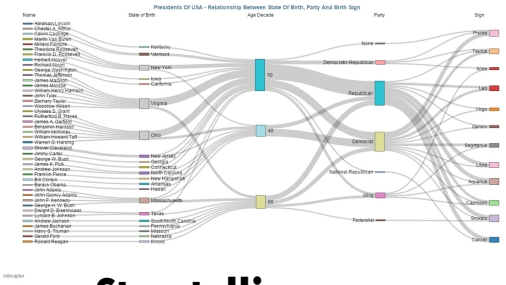
Computer-based visualization systems provide visual representations of **datasets** designed to help **people** carry out **tasks** more **effectively**. *– Tamara Munzner at UBC*



Understanding

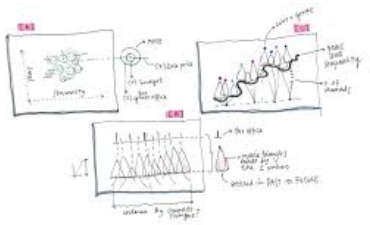


Exploratory



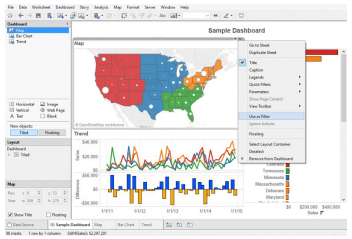
Storytelling

What and how



human

human



machine

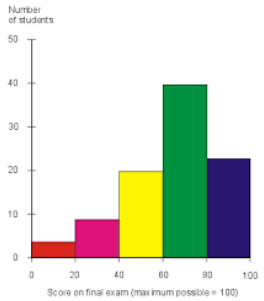
human

machine

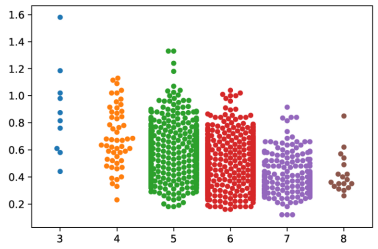
X

machine

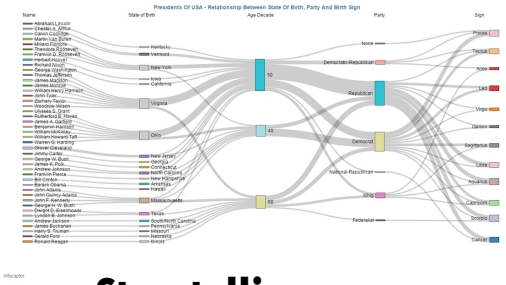
Computer-based visualization systems provide visual representations of **datasets** designed to help **people** carry out **tasks** more **effectively**. *– Tamara Munzner at UBC*



Understanding



Exploratory

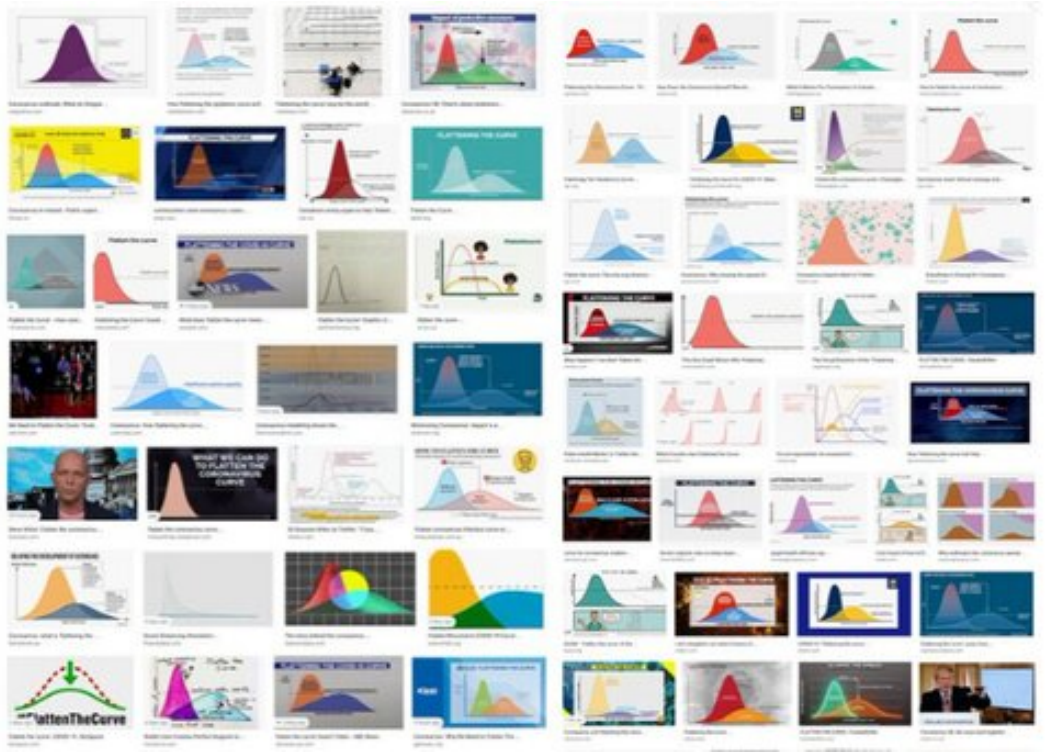


Storytelling

*Making **Human-in-the-loop** Data Analytics (Science) More **Effective***

COVID 19 (2020-)

There have never been so many line charts, bar charts and maps occupying the news, as simple data visualizations have become key to communicating vital information about the coronavirus pandemic to the public.



what you can do if you have learned effective visualization techniques

COVID-19 DashBoard
Tsinghua Database Research Group
Qatar Computing Research Institute

可视化分析 疫情小区查询 疫情分析报告 联系我们

Last Updated at: 2021-11-24 18:54:50 (UTC+8)

较昨日+50	较昨日+25	较昨日-8	较昨日0	较昨日+44
125073	9303	361	5695	116827
累计确诊	现存疑似	现存重症	死亡	治愈

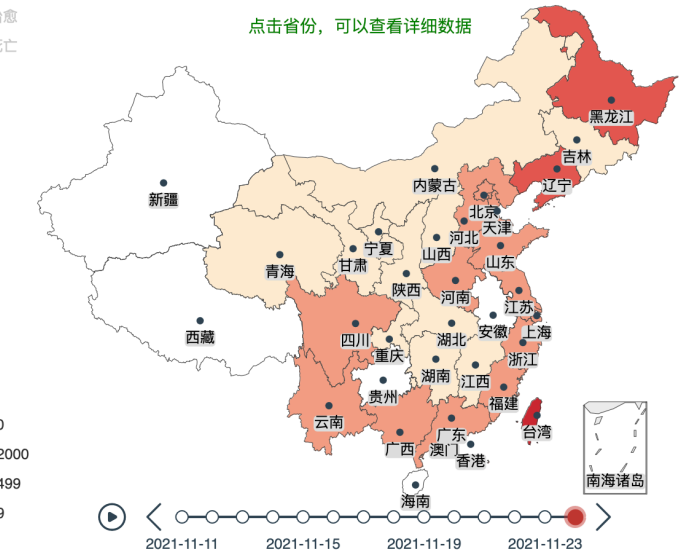
中国-省级 | 中国-县市级 | 世界 | 卡塔尔 | 美国 | 加拿大 | 澳大利亚 | 日本 | 韩国 | 意大利

现存确诊
累计确诊
累计治愈
累计死亡

COVID-19疫情地图（中国）

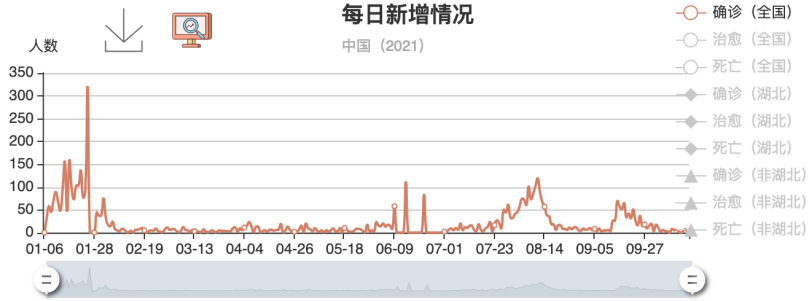
累计确诊: 127501 现存确诊: 3013 累计死亡: 5697 累计治愈: 118791

点击省份，可以查看详细数据

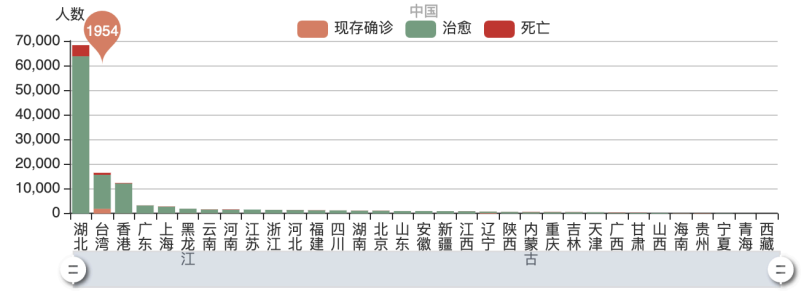


每日新增情况

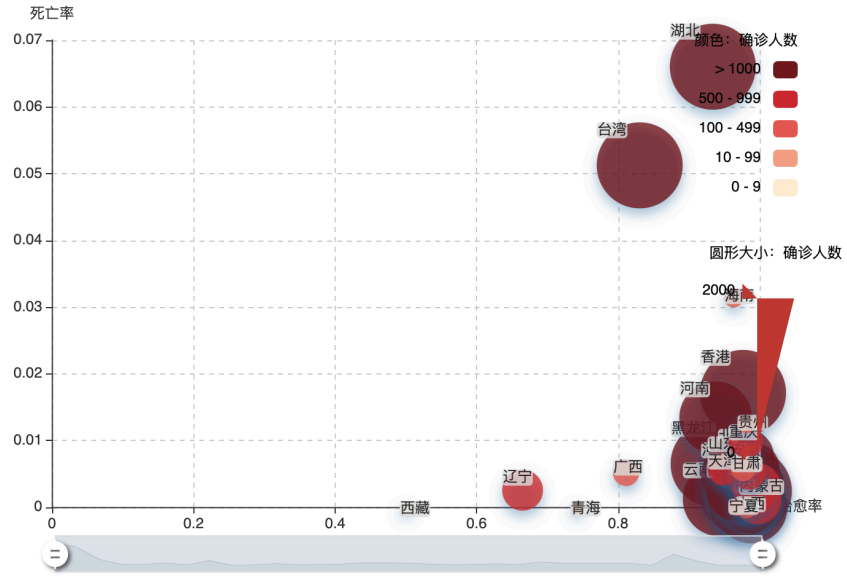
中国（2021）



病患总数排名（中国-下属一级行政区）



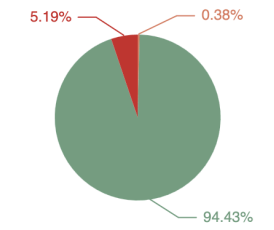
治愈率-死亡率分布（中国-下属一级行政区）



病患类型比例

中国

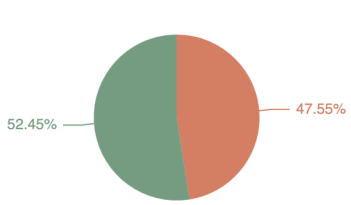
现存确诊 治愈 死亡



病患类型比例

中国

男 女



Yuyu Luo
DSA Thrust

Simple and Fun

Early 2020: I wanted to know what is happening in Qatar about COVID19?

Early 2020: I wanted to know what is happening in Qatar about COVID19?

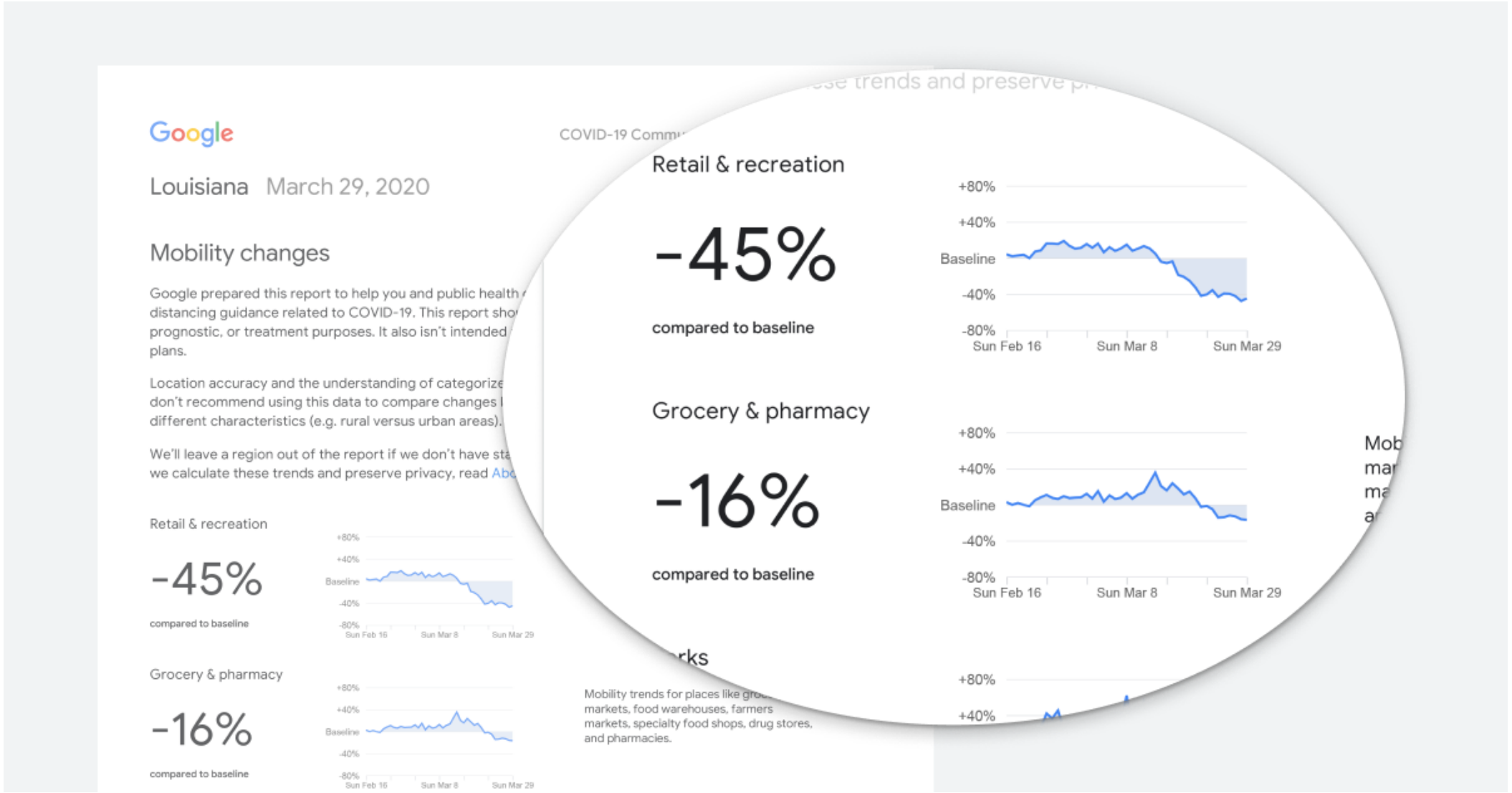
But, there was no good COVID dashboard in Qatar

Early 2020: I wanted to know what is happening in Qatar about COVID19?

But, there was no good COVID dashboard in Qatar

Can I make useful dashboards?

Key: Google Mobility Data



Dash in 20 Minutes

By the end of this tutorial, you will understand the basic building blocks of Dash and you will know how to build this app:

▸ View app

Hello World

Building and launching an app with Dash can be done with just 7 lines of code.

Open a Python IDE on your computer, create an `app.py` file with the code below and install Dash if you haven't done so already. To launch the app, type into your terminal the command `python app.py`. Then, go to the http link.

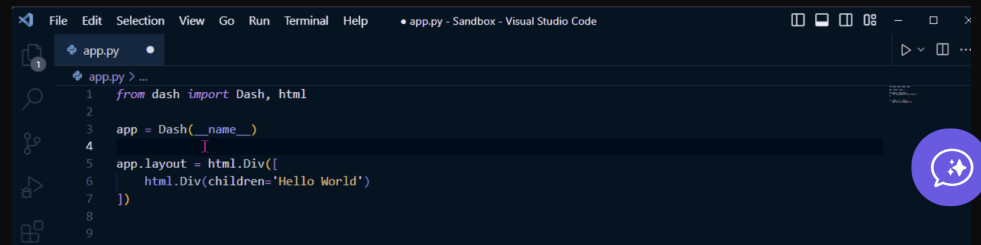
Alternatively, with Dash 2.11 or later, you can run this app and other examples from this documentation in a **Jupyter Notebook**.

The code below creates a very small "Hello World" Dash app.

```
from dash import Dash, html

app = Dash(__name__)

app.layout = html.Div([
    html.Div(children='Hello World')
])

if __name__ == '__main__':
    app.run(debug=True)
```

Hello World

Follow this example gif (using VS Code) if you are not sure how to set up the app:



Dash in 20 Minutes

By the end of this tutorial, you will understand the basic building blocks of Dash and you will know how to build this app:

▸ View app

Hello World

Building and launching an app with Dash can be done with just 7 lines of code.

Open a Python IDE on your computer, create an `app.py` file with the code below and install Dash if you haven't done so already. To launch the app, type into your terminal the command `python app.py`. Then, go to the http link.

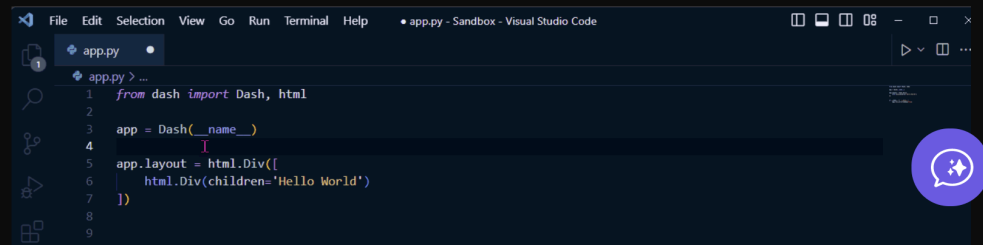
Alternatively, with Dash 2.11 or later, you can run this app and other examples from this documentation in a **Jupyter Notebook**.

The code below creates a very small "Hello World" Dash app.

```
from dash import Dash, html

app = Dash(__name__)

app.layout = html.Div([
    html.Div(children='Hello World')
])

if __name__ == '__main__':
    app.run(debug=True)
```

```
Hello World
```

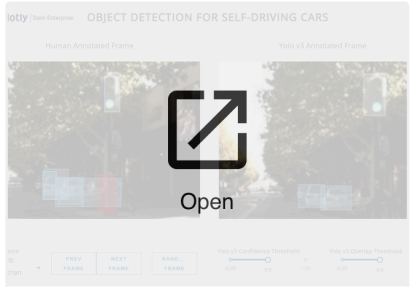
Follow this example gif (using VS Code) if you are not sure how to set up the app:



plotly | Dash

All Apps (104)

Search applications...

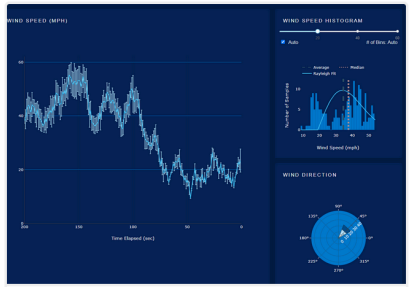


Real-Time Object Detection

Image Processing | AI | ML | (+2)

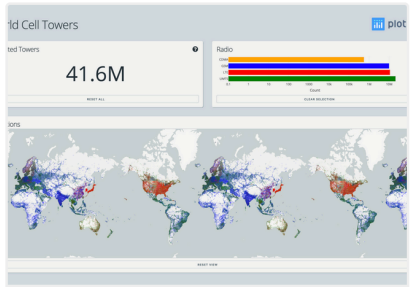
Manufacturing SPC Dash...

Streaming | DAQ | Manufacturing



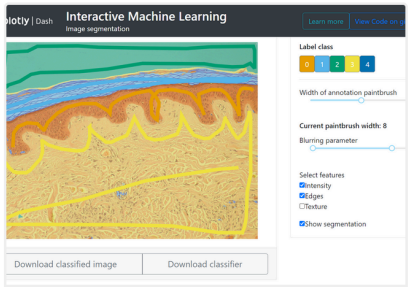
Wind Speed Dashboard f...

Streaming | SQL | Energy



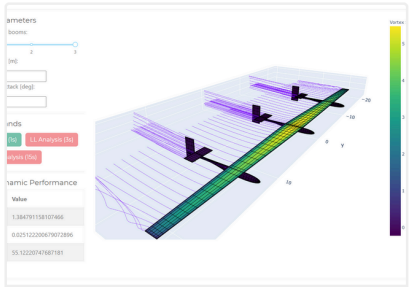
World Cell Towers

Geospatial | Datashader | Dask | Telecom



Interactive Image Segme...

ML | Healthcare | AI | Image Processing | (+1)

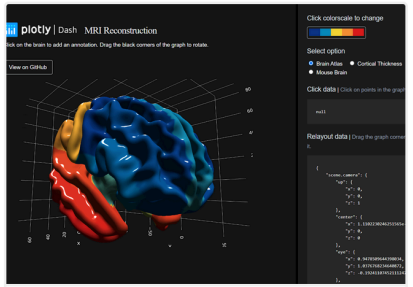


Aircraft CFD

Aerospace | CFD

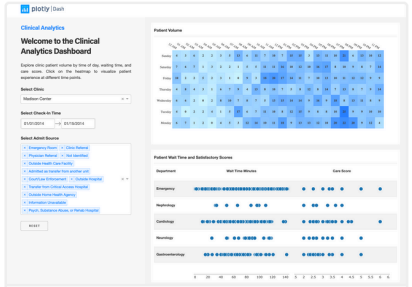
Pivottable Demo

Financial | Insurance



Brain Surface Viewer

Biotechnology | 3d



Clinical Analytics Dashb...

Pharma | Healthcare

Monitoring Qatar Mobility Data

Following the cabinet's decision to lift all COVID-19 restrictions in March, 2023, in Qatar, the data collection has been stopped. When the monitoring is needed again, the data collection will resume.

Last updated at: 2024-03-16 15:50:24

REFRESH DATA

Download HotSpots

Download HotZones

April 1, 2023 → April 1, 2023

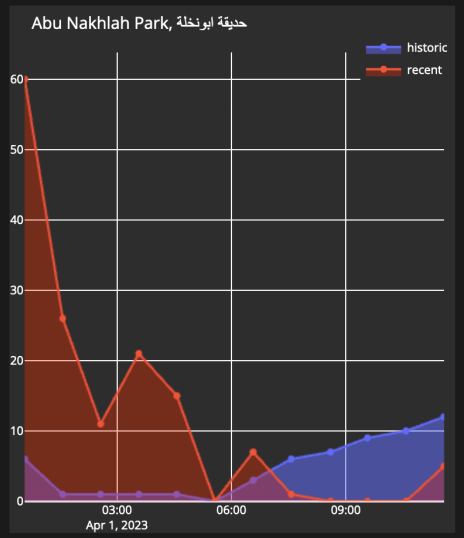
× recreation

Select certain hours

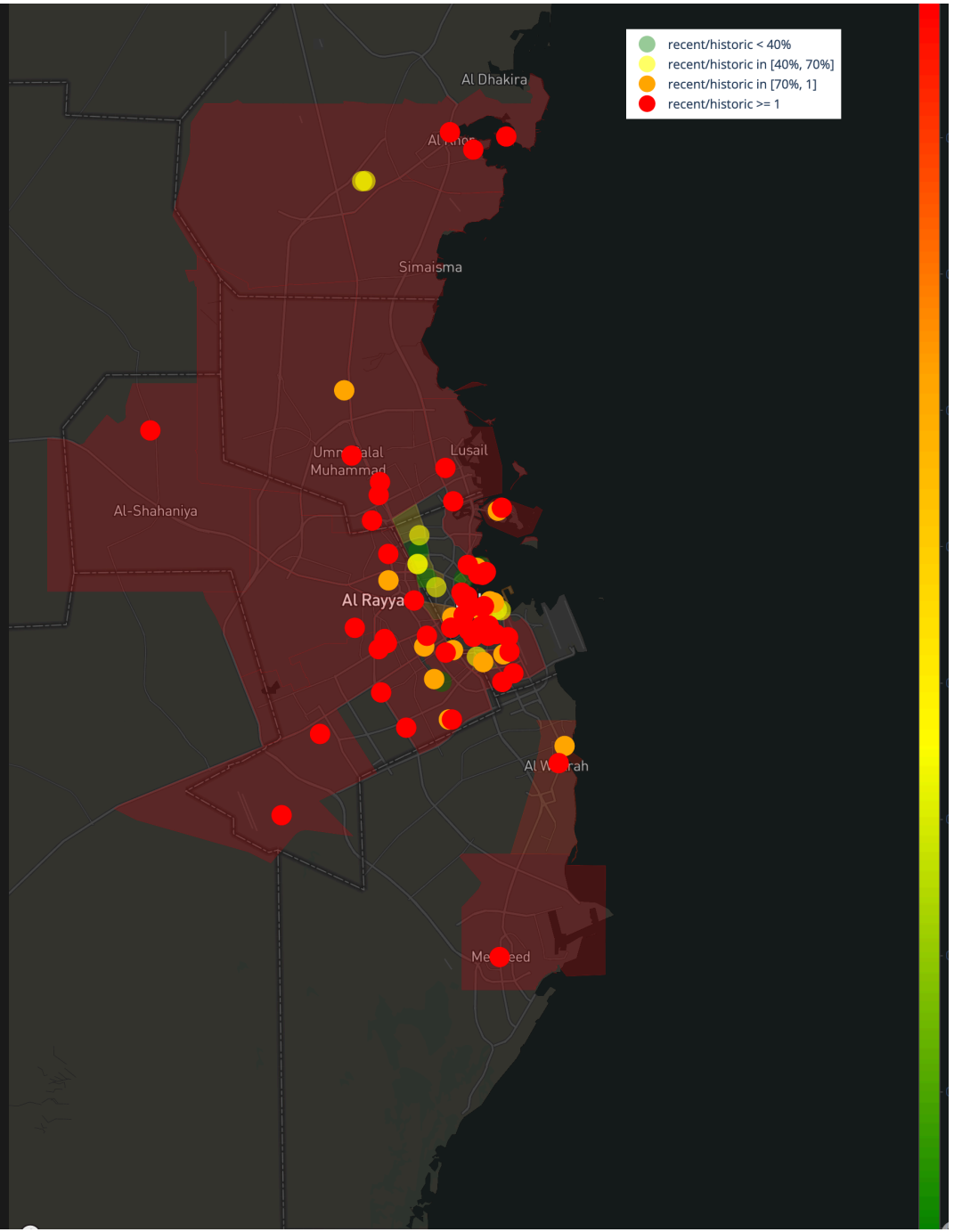
Select or type a place name

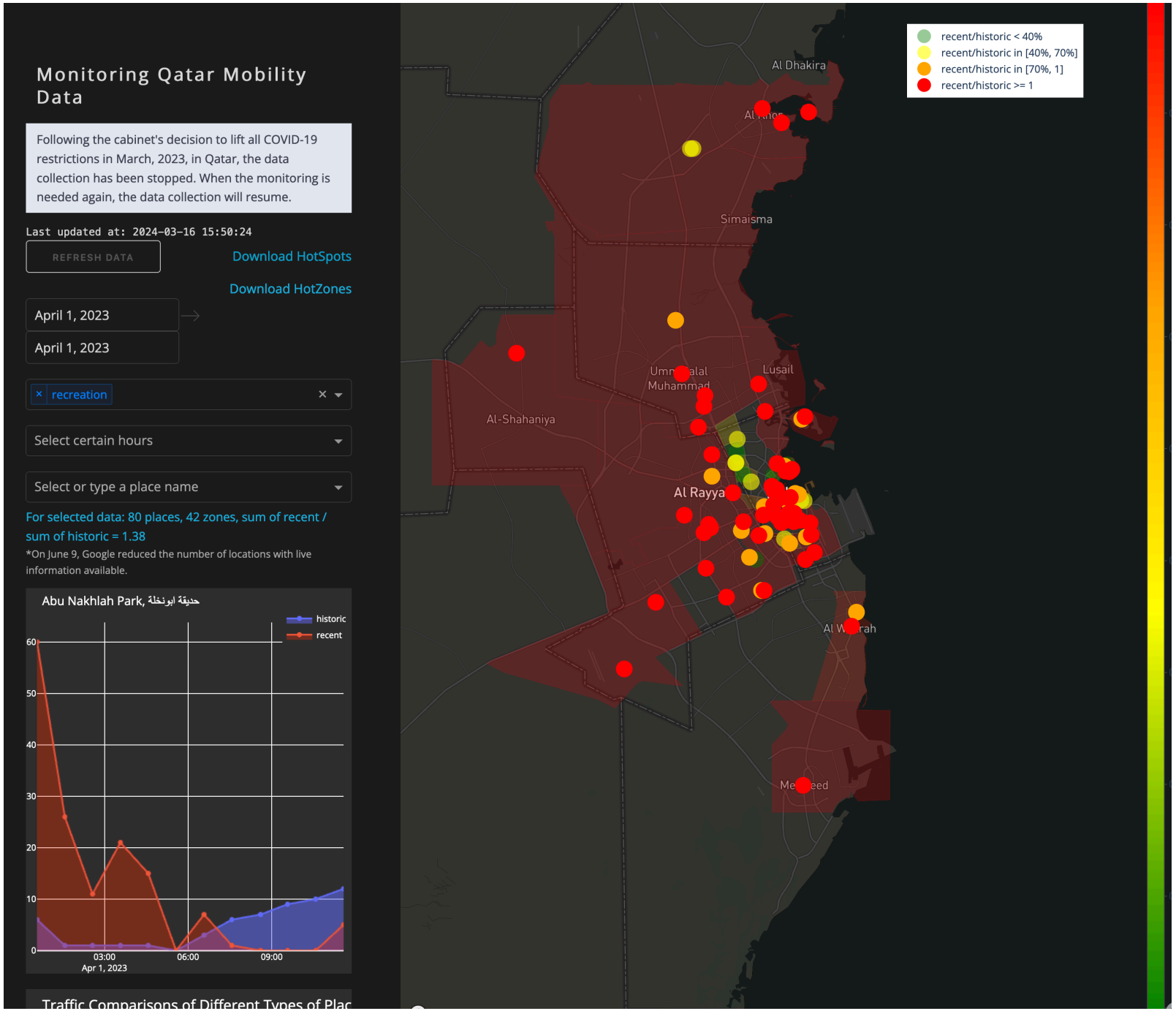
For selected data: 80 places, 42 zones, sum of recent / sum of historic = 1.38

*On June 9, Google reduced the number of locations with live information available.



Traffic Comparisons of Different Types of Plac





In March of 2020, the traffic between **Al Udeid Air Base (US)**

and **QDC**

has **Increased 200%.**

Monitoring Qatar Mobility Data

Following the cabinet's decision to lift all COVID-19 restrictions in March, 2023, in Qatar, the data collection has been stopped. When the monitoring is needed again, the data collection will resume.

Last updated at: 2024-03-16 15:50:24

REFRESH DATA

Download HotSpots

Download HotZones

April 1, 2023 → April 1, 2023

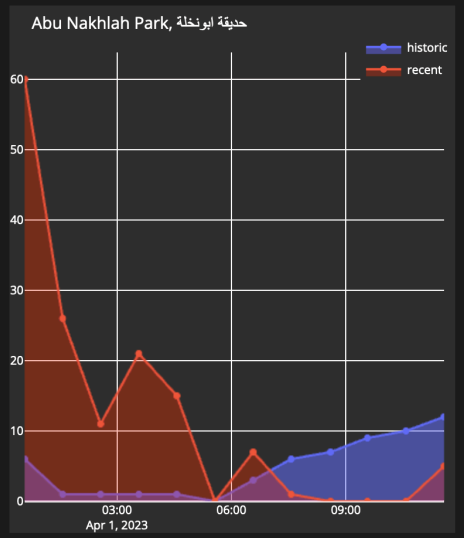
× recreation

Select certain hours

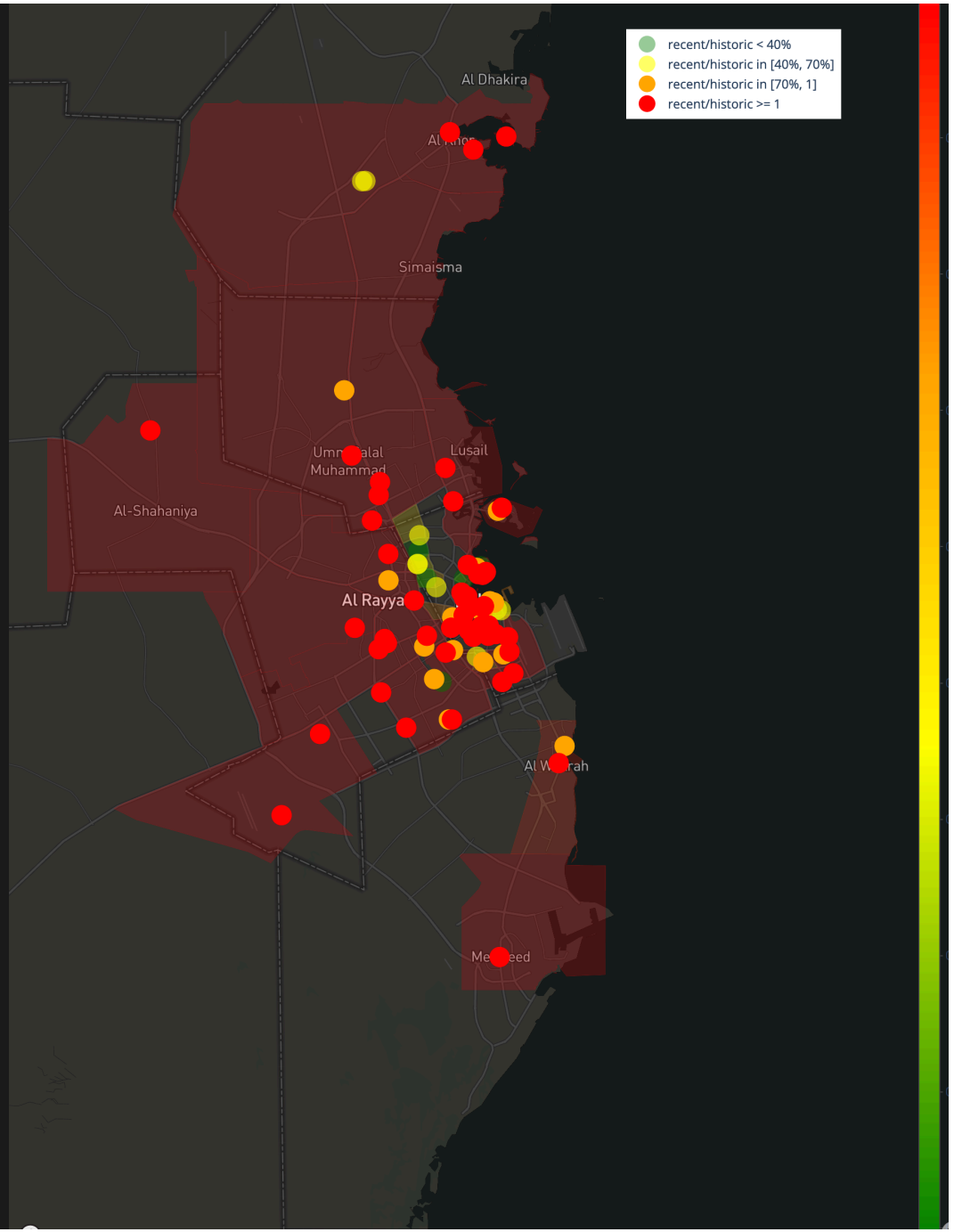
Select or type a place name

For selected data: 80 places, 42 zones, sum of recent / sum of historic = 1.38

*On June 9, Google reduced the number of locations with live information available.



Traffic Comparisons of Different Types of Plac



In March of 2020, the traffic between **Al Udeid Air Base (US)**

and **QDC**

has **Increased 200%.**

Monitoring COVID-19 in Nigeria

COVID-19 is the infectious disease caused by the novel coronavirus. This new virus and disease were unknown to the population before the outbreak began in Wuhan, China, in December 2019. This page provides data, maps and resources about the coronavirus response across Nigeria

If you are showing symptoms of Covid-19 or believe you are a close contact of a confirmed case, call your state helpline here.

NCDC Toll-free Number: **08009700001**
SMS: **0809 955 5577**
WhatsApp: **0708 711 0839**

Confirmed Cases	Deaths	Test Samples	Active Cases
190,983	2,361	2,750,298	10,575
Source: Covid_Cases_by_State	Source: Covid_Cases_by_State	Source: Covid_Cases_by_State	Source: Covid_Cases_by_State

COVID-19 Nigeria Dashboards

This section lists a series of available dashboards, produced by governmental agencies and external organisations, providing data and insights into the COVID-19 situation in Nigeria.

Presidential Task Force Incidence & Response...

Incidence and Response Tracker by the Presidential Task Force on Covid-19

Open Dashboard

Main NBS Dashboard

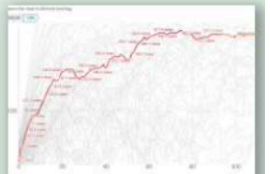
Monitoring cases, deaths, socioeconomic vulnerability & other measures by state

Open Dashboard

QCRI Lagos Mobility Dashboard

Mobility Dashboard using Google Mobility Data

Open Dashboard



Our World In Data Dashboard

Coronavirus pandemic: daily updated research and data from Our World in Data

Open Dashboard

Creative and Great

200 Countries, 200 Years

Data Geographies - v1 - by Gapminder.xlsx
Request more info

View

	geo	name	four_regions	eight_regions	six_regions	members
1	afg	Afghanistan	asia	asia_west	south_asia	g77
2	alb	Albania	europe	europe_east	europe_central_asia	others
3	dza	Algeria	africa	africa_north	middle_east_north_africa	g77
4	and	Andorra	europe	europe_west	europe_central_asia	others
5	ago	Angola	africa	africa_sub_saharan	sub_saharan_africa	g77

Showing 1-5 of 197 rows, 13 columns See all

Switch to column overview

life_expectancy_years.csv
Request more info

View

	country	1800	1801	1802	1803	1804	1805	1806
1	Afghanistan	28.2	28.2	28.2	28.2	28.2	28.2	2
2	Albania	35.4	35.4	35.4	35.4	35.4	35.4	3
3	Algeria	28.8	28.8	28.8	28.8	28.8	28.8	2
4	Andorra	No data.	No data.	No data.	No data.	No data.	No data.	No da
5	Angola	27	27	27	27	27	27	

Showing 1-5 of 187 rows, 302 columns See all

Switch to column overview

income_per_person_gdppercapita_ppp_inflation_adjusted.csv
Request more info

View

	country	1800	1801	1802	1803	1804	1805	1806
1	Afghanistan	603	603	603	603	603	603	
2	Albania	667	667	667	667	667	668	
3	Algeria	715	716	717	718	719	720	
4	Andorra	1200	1200	1200	1200	1210	1210	
5	Angola	618	620	623	626	628	631	

Showing 1-5 of 193 rows, 242 columns See all

Switch to column overview

population_total.csv
Request more info

View

	country	1800	1801	1802	1803	1804	1805	1806
1	Afghanistan	3280000	3280000	3280000	3280000	3280000	3280000	3280
2	Albania	400000	402000	404000	405000	407000	409000	411
3	Algeria	2500000	2510000	2520000	2530000	2540000	2550000	2560
4	Andorra	2650	2650	2650	2650	2650	2650	2
5	Angola	1570000	1570000	1570000	1570000	1570000	1570000	1570

Showing 1-5 of 195 rows, 302 columns See all

Switch to column overview


BBC FOUR

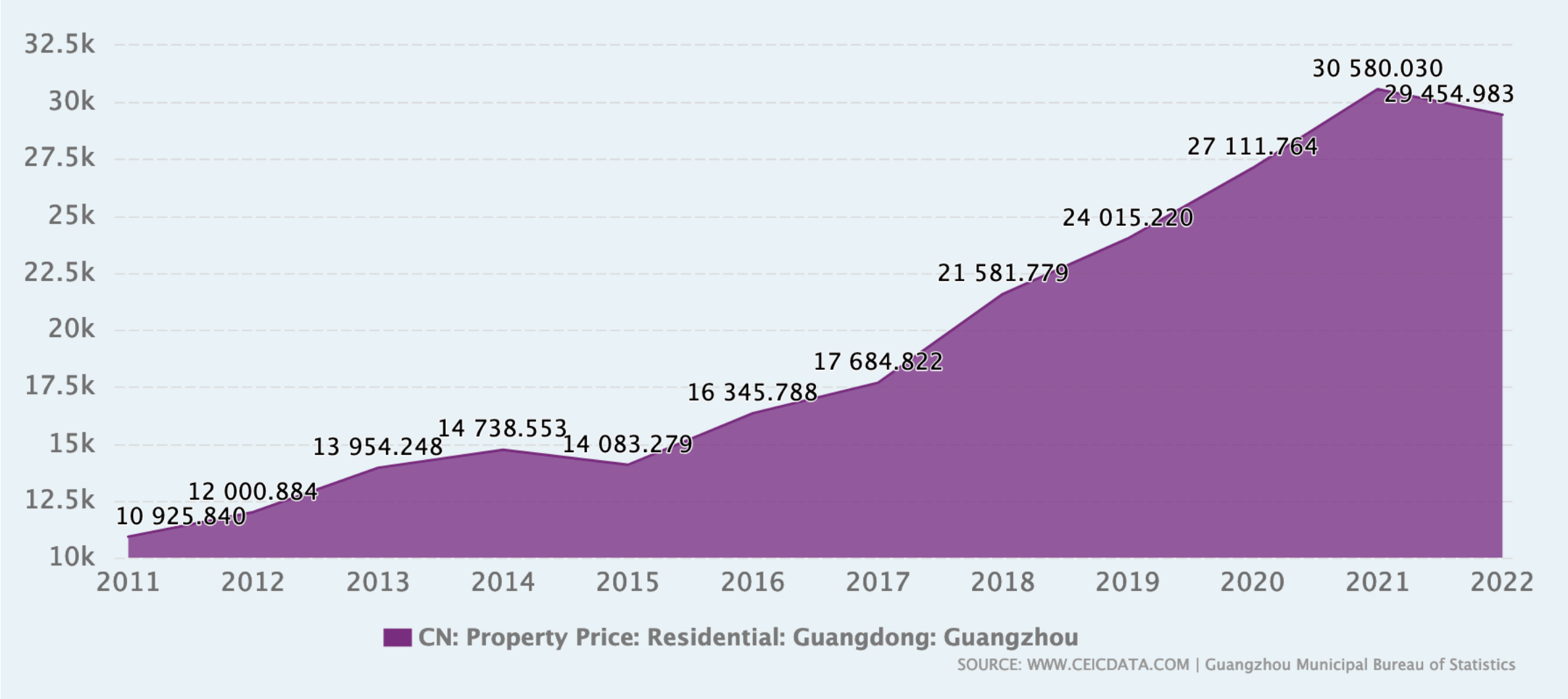

BBC FOUR


BBC FOUR

Sample Project: Make a visualization to do storytelling about

- House price
- Real estate
- Stock price
- Average salary
- Unemployment rate
- Birth rate
- Ageing population
- ……

Screenshot (realistic): can it be easily altered to answer question like ”house price of GZ I 2023”



EXCLUSIVE

PIET & LINDA SP
THE WIZARDING

Severus Snape
NEW Headmaster

HIPPOGRIFF ETIQUETTE

A

Reporters of the Daily Prophet confirmed that Piet and Linda where spotted in the wizarding world. At this point it is not clear how they managed to enter the wizarding world and if they are still present. If you see or meet these muggles please contact the Daily Prophet or the Ministry of Magic. Contact information is available on page 5. More information will follow soon.

T

Weasleys Wizard Wheezes Now Open

EXCLUSIVE

PIET & LINDA SP
THE WIZARDING

Severus Snape
NEW Headmaster

HIPPOGRIFF ETIQUETTE

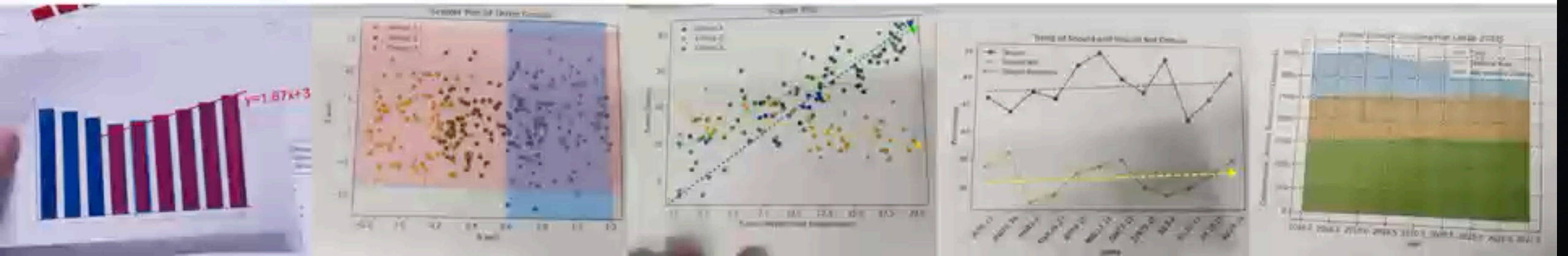
A

Reporters of the Daily Prophet confirmed that Piet and Linda where spotted in the wizarding world. At this point it is not clear how they managed to enter the wizarding world and if they are still present. If you see or meet these muggles please contact the Daily Prophet or the Ministry of Magic. Contact information is available on page 5. More information will follow soon.

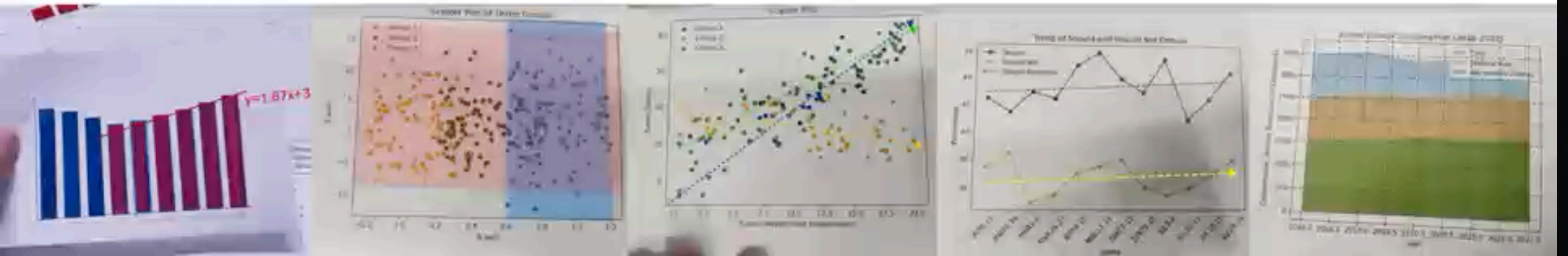
T

Weasleys Wizard Wheezes Now Open

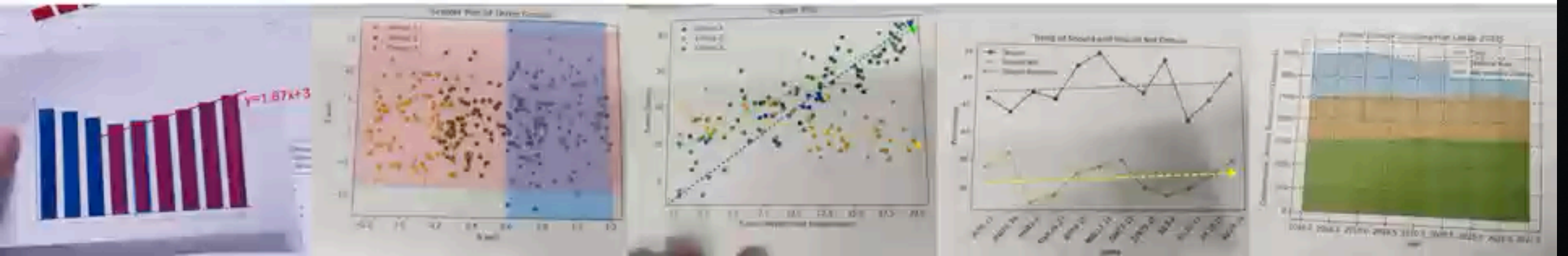
Augmenting Realistic Visualizations with Virtual Overlays

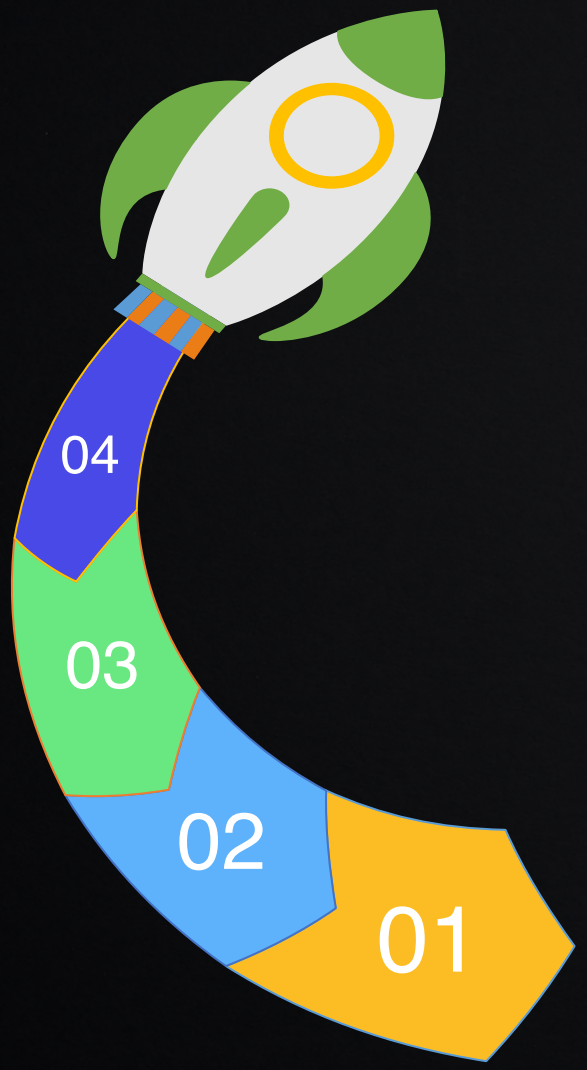


Augmenting Realistic Visualizations with Virtual Overlays



Augmenting Realistic Visualizations with Virtual Overlays





01
Data Acquisition

- image data
- text data
- tabular data

02
Data Understanding

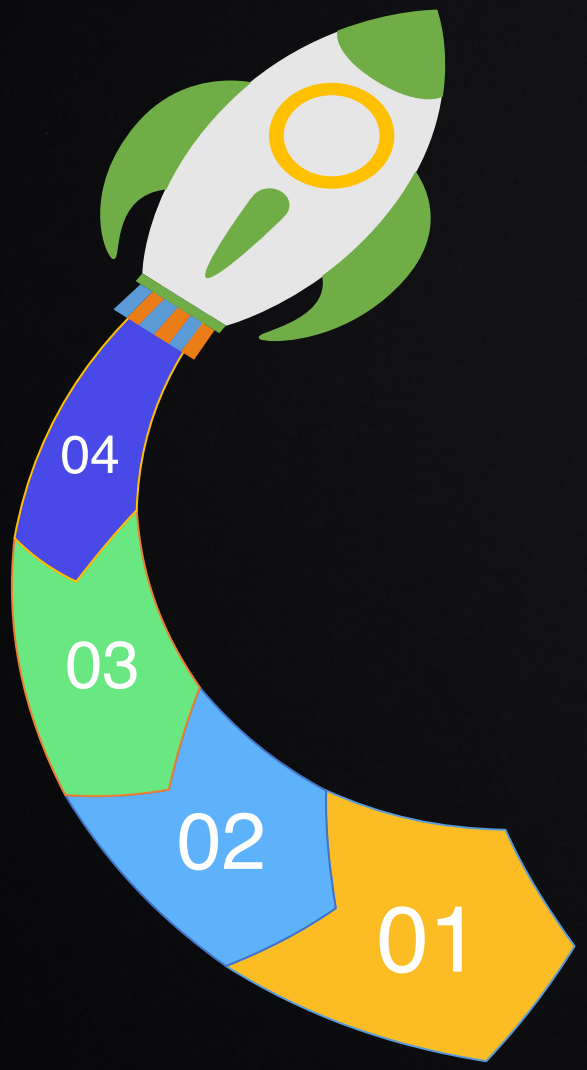
- exploratory data analysis
- data visualization

03
Data Preparation

- data transformation
- data cleaning
- entity resolution
- … …

04
Advanced Topics

- data-centric AI
- HPC for LLMs
- … …



01
Data Acquisition

- image data
- text data
- tabular data

02
Data Understanding

- exploratory data analysis
- data visualization

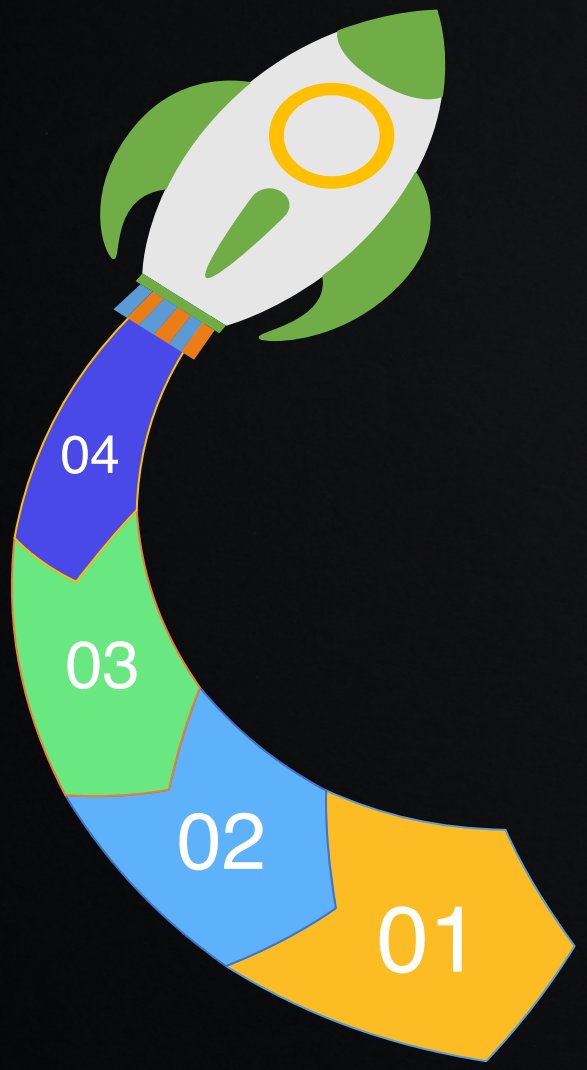
03
Data Preparation

- data transformation
- data cleaning
- entity resolution
- … …

04
Advanced Topics

- data-centric AI
- HPC for LLMs
- … …

Data Science: Unleashing Your Inquisitive Mind



01
Data Acquisition

- image data
- tabular
- text data

02
Data Understanding

- exploratory data analysis
- data visualization

03
Data Preparation

- data transformation
- data cleaning
- entity resolution
- … …

04
Advanced Topics

- data-centric AI
- HPC for LLMs
- … …

Discovery

Cleaning

Integration

Chef

Recipe (algorithm)

L covery

Cleaning

Integration

Food Prep

human-in-the-loop

Chef

Recipe (algorithm)

L covery

Cleaning

Integration

Food Prep

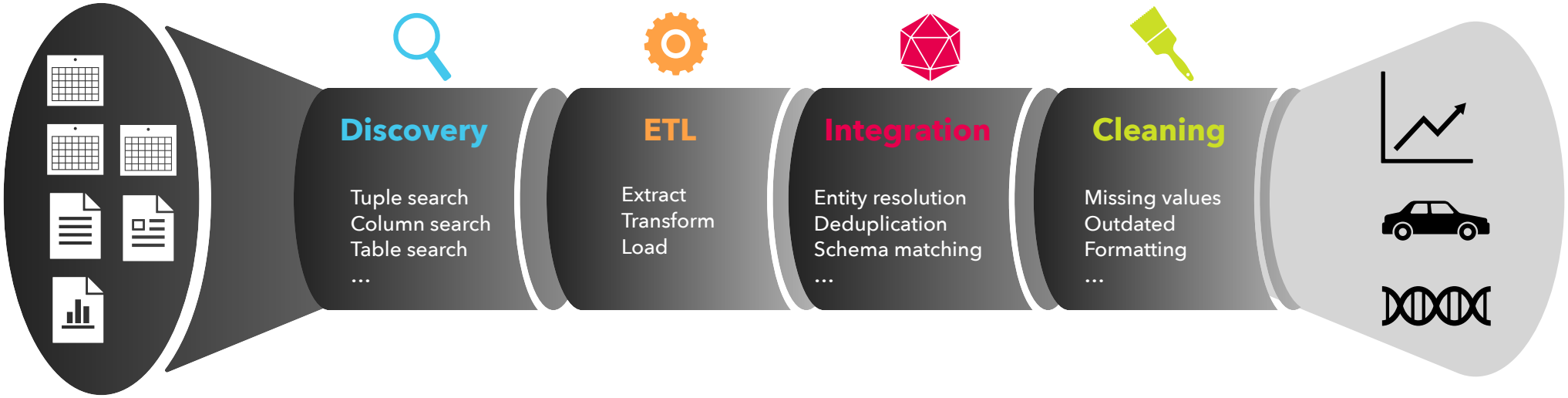
human-in-the-loop

FLOUR

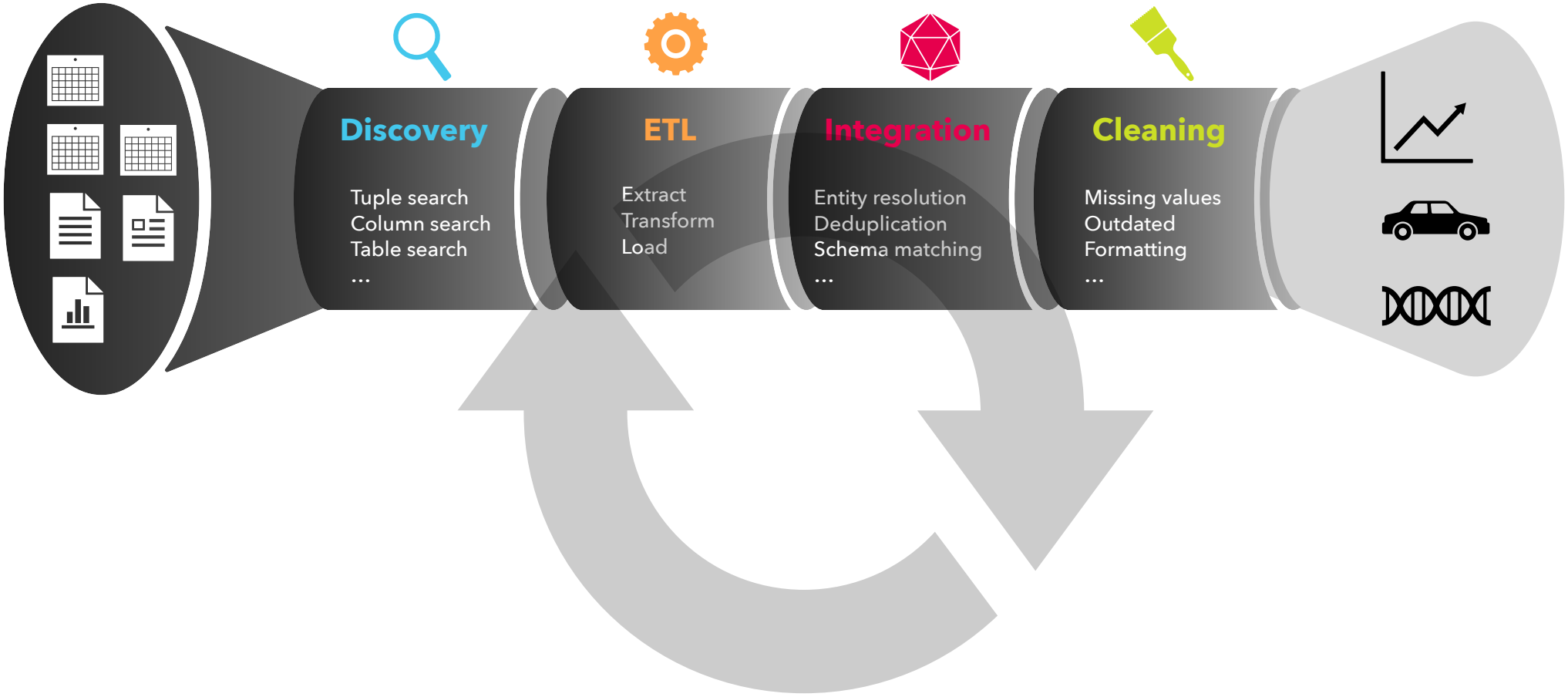
Taste Testing

Recipe (algorithm)

Key techniques

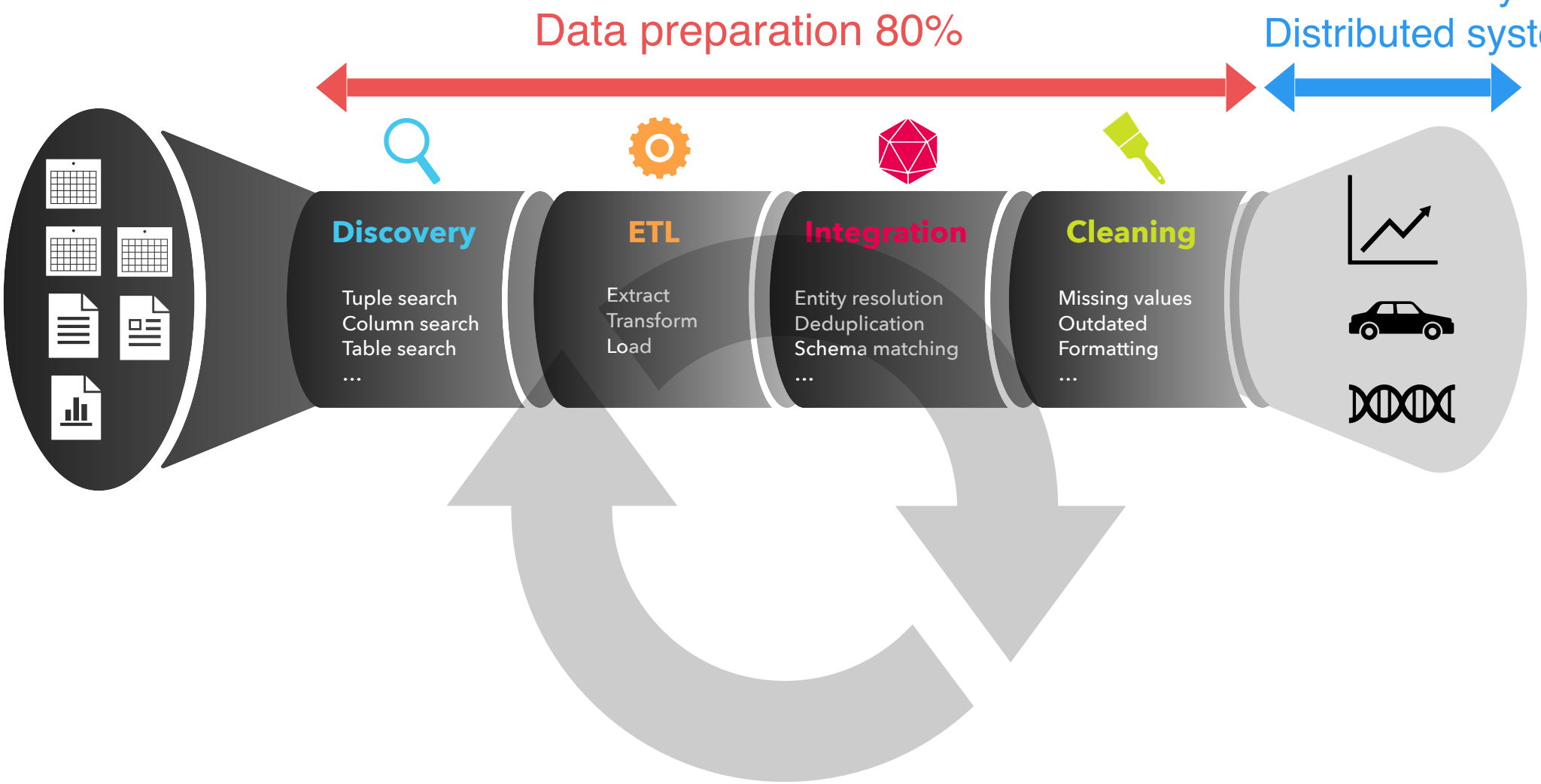


Key techniques



Key techniques

Exploratory data analysis
Data Visualization
Statistics
ML
Network analysis
Distributed systems



Not Fun
But Unavoidable

Data in the la-la land

Here is the data, design your algorithms

SuperGLUE Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_a$/EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

kaggle

Datasets

Explore, analyze, and share quality data. Learn more about data types, creating, and collaborating.

+ New Dataset

Search datasets

Filters

All datasets | Computer Science | Education | Classification | Computer Vision | NLP | Data Visualization | Pre-Trained Model

Trending Datasets

See All

Hourly Electricity Consumption and...

Stefan Comanita · Updated...

Usability **10.0** · 724 kB

1 File (CSV)

6

Box-Office Secrets with TMDB Trends

Shiv_D24Coder · Updated a...

Usability **9.4** · 88 kB

1 File (CSV)

8

House price index (2015 = 100) -...

Sándor Burian · Updated 17...

Usability **10.0** · 3 kB

1 File (CSV)

14

List of Countries by GDP Sector...

Raj Kumar Pandey · Update...

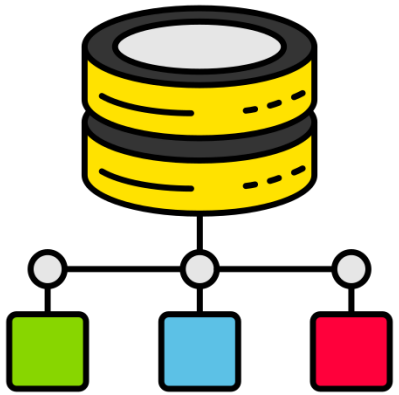
Usability **9.4** · 8 kB

1 File (CSV)

4

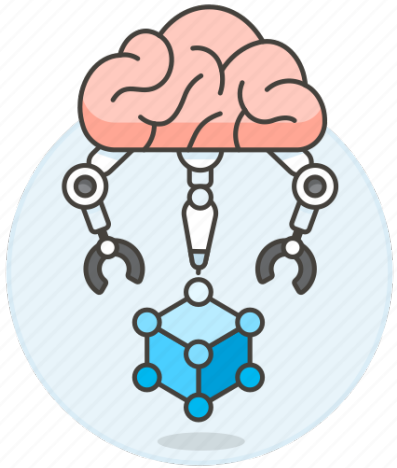
=

Data



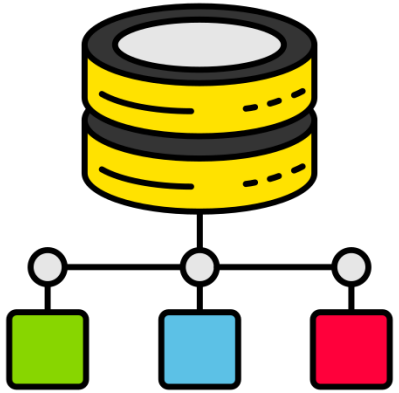
+

Algorithm



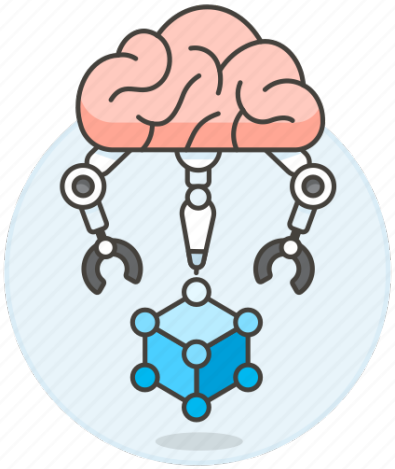
=

Data



+

Algorithm



Garbage In

Garbage Out

IATA	ICAO
KGW	AYKQ
KRX	AYKR
KIE	AYKT
KZF	AYKT
KUQ	AYKU
KVG	AYKV

(a) Uniqueness error

Genus Name	Species
Amphimachairodus	4
Hemimachairodus	1
Lokotunjailurus	1
Megantereon	8
Hemimachairodus	1
Ischyrosmilus	1

(b) Uniqueness error

ID	Awardee
865512	FRANKS, Robert James
865513	BARBER, Alan Leonard
865514	BARROWS, William James
865514	CONNERS, Quentin David
865515	MORLEY, Richard John
865516	CARMODY, David John

(c) FD error

Call sign	City of license	State
WXAV	Chicago	Illinois
WRBC	Chicago	Illinois
WLUW	Chicago	Illinois
WBOR	Brunswick	Maine
WRBC	Lewiston	Maine
WMEB-FM	Orono	Maine
WUPI	Presque Isle	Maine

(d) FD error

Banua	Population
1861	8,011
1871	8.716
1881	9,954
1901	11,895
1911	13,329
1921	11,352
1931	11,709

(e) Outlier error ("8.716" uses "." in place of ",")

Name and surname	Height
Katarina Zec	1.78
Bojana Stevanovic	183
Aleksandra Katanic	175
Jovana Subašić	187
Snezana Bogicevic	177
Kristina Arsenic	189

(f) Outlier error

Author	Director
Joshua Ravetch	Steve Gomer
David Grae	Kevin Doeling
Tom Garrius	Alan Myerson
Sibyl Gardner	James Hayman
Joy Gregory	Kevin Dowling
Antoinette Stella	Rob Morrow

(g) Spelling error

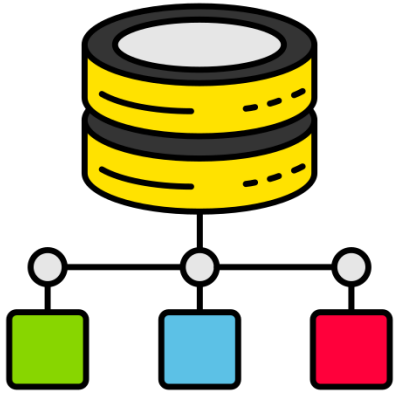
Title	Directed by
"Ratters"	Jet Wilkinson
"Last Seen"	Jet Wilkinson
"No Smoke"	Pino Amenta
"In Harm's Way"	Pino Amanta
"The Hit"	Nicholas Bufalo
"Just Desserts"	Nicholas Bufalo

(h) Spelling error

[link]

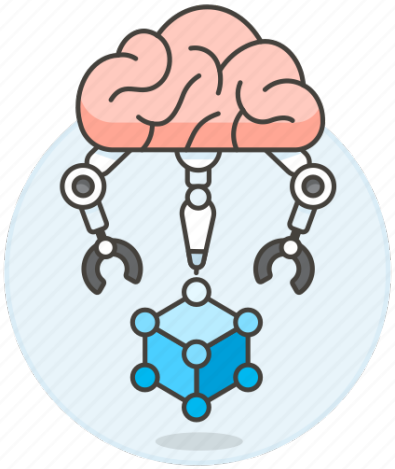
=

Data



+

Algorithm

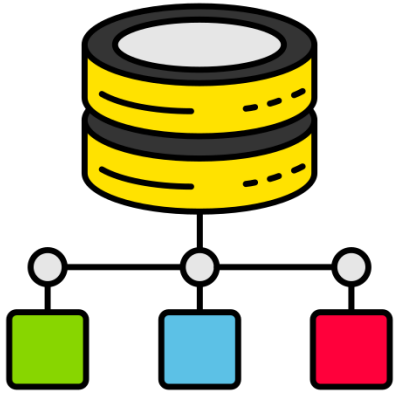


Garbage In

Garbage Out

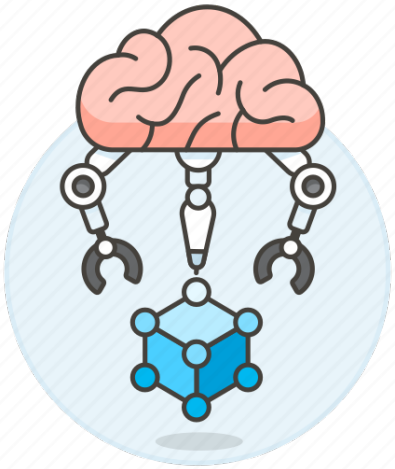
=

Data



+

Algorithm



Error Detection

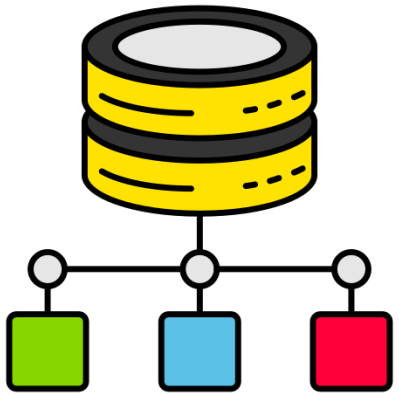
Data Repairing

Garbage In

Garbage Out

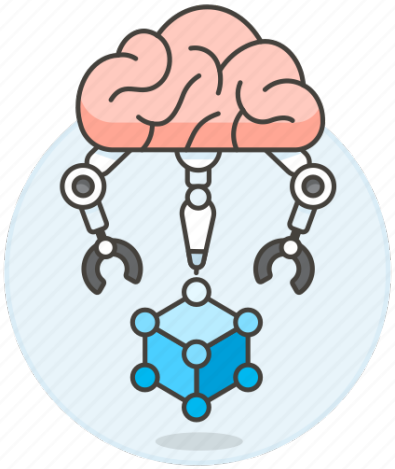
=

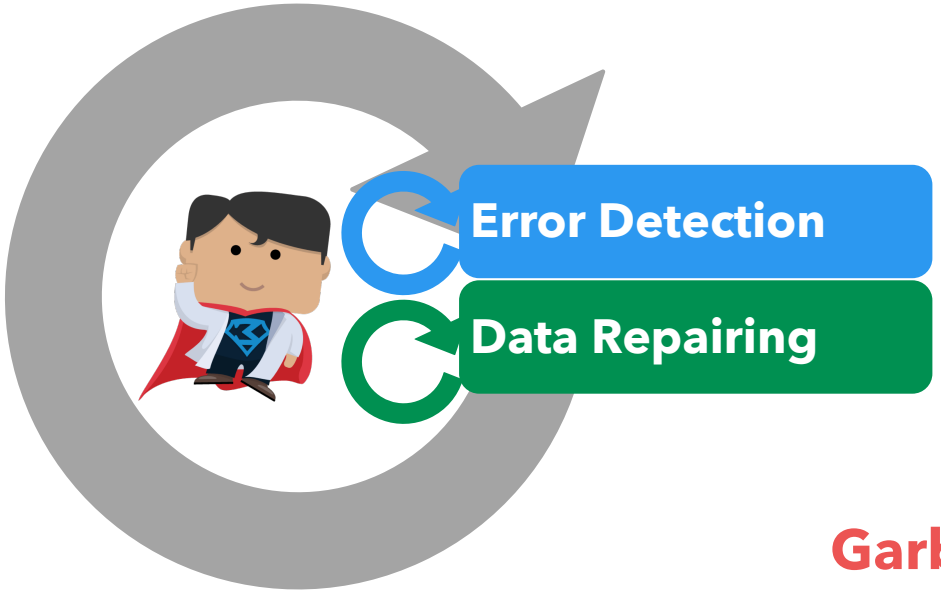
Data



+

Algorithm



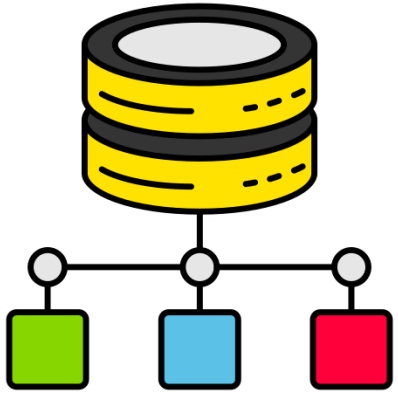


Garbage In

Garbage Out

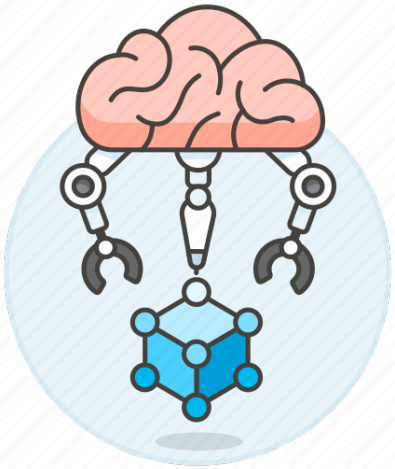
=

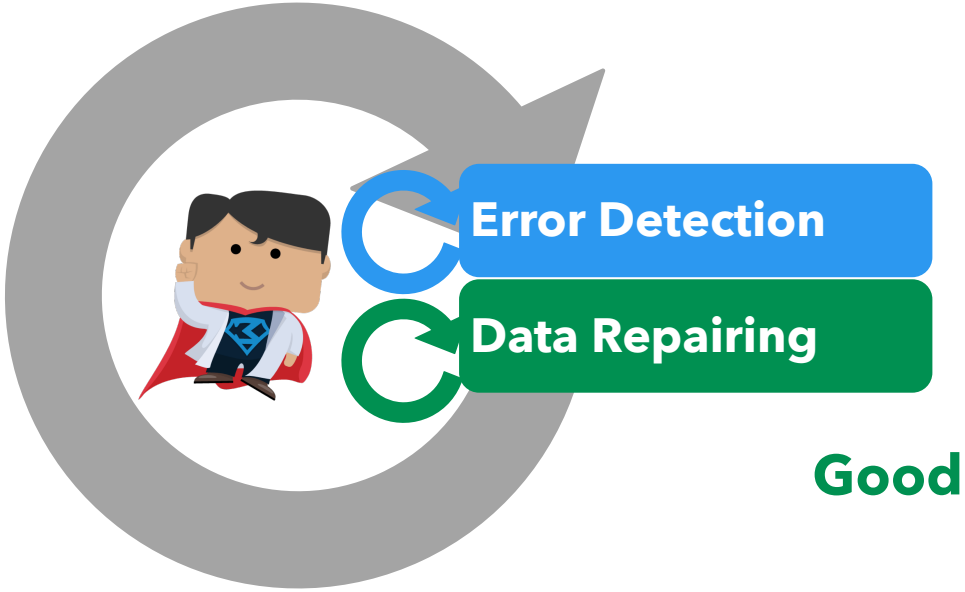
Data



+

Algorithm

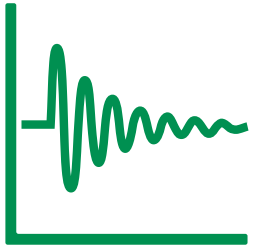




Error Detection

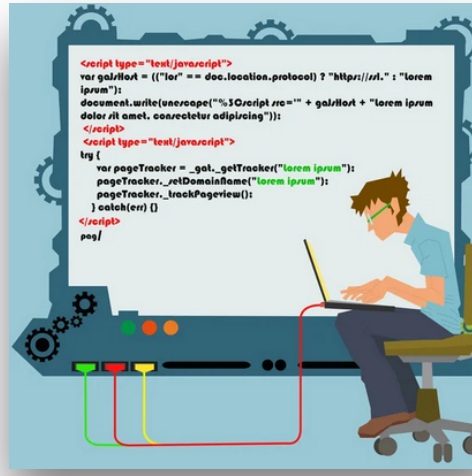
Data Repairing

Good Data



Good Result

Data cleaning (a single table)

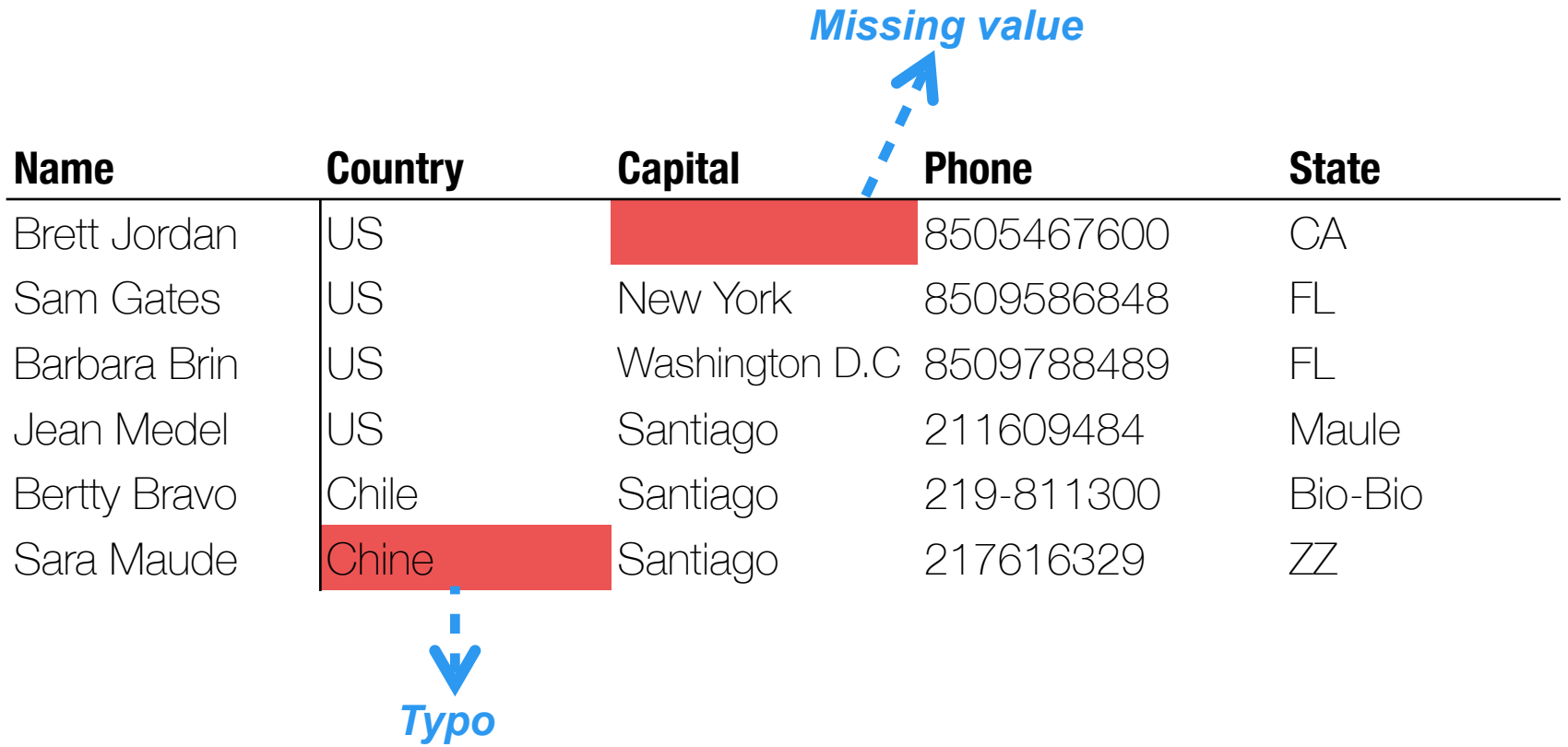


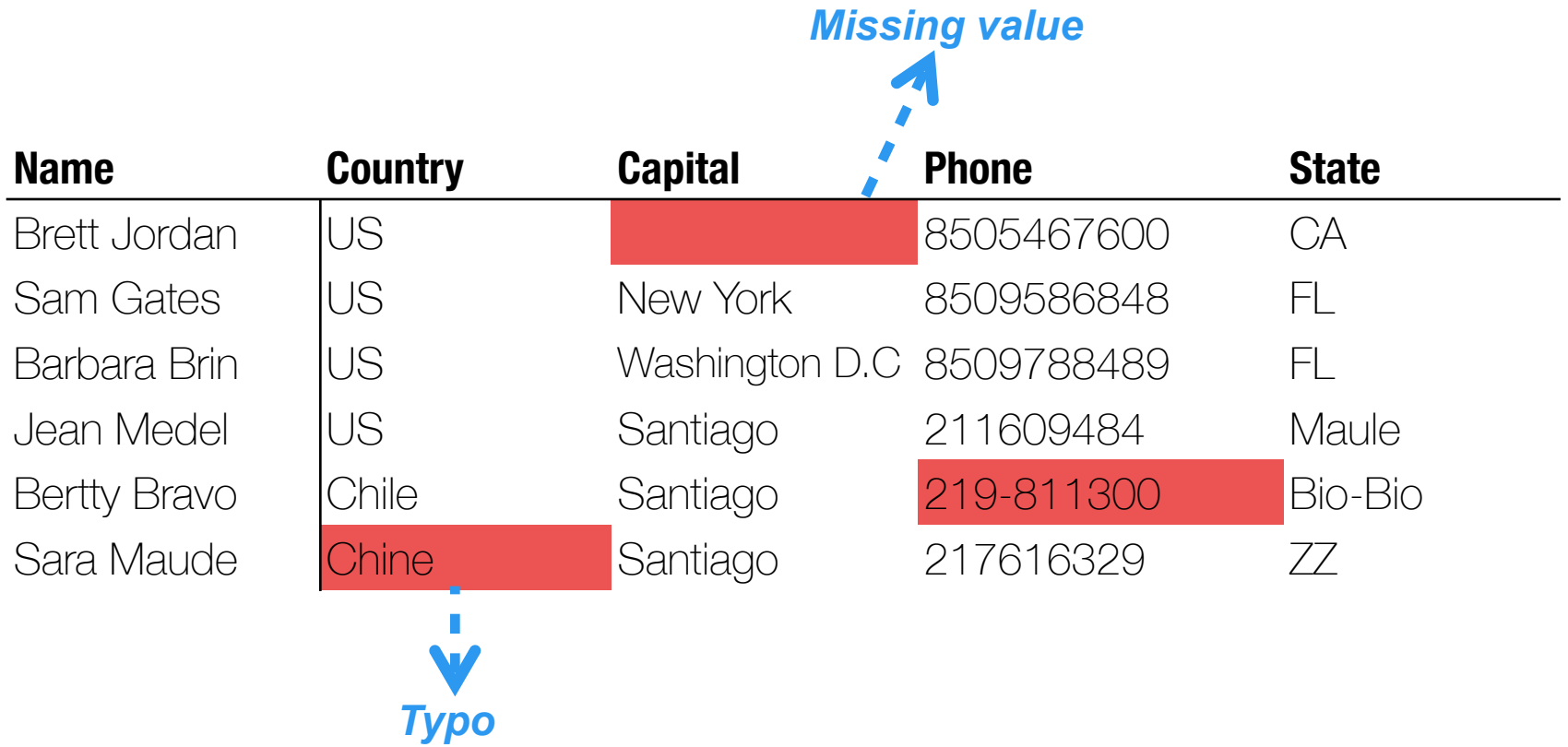
Name	Country	Capital	Phone	State
Brett Jordan	US		8505467600	CA
Sam Gates	US	New York	8509586848	FL
Barbara Brin	US	Washington D.C	8509788489	FL
Jean Medel	US	Santiago	211609484	Maule
Bertty Bravo	Chile	Santiago	219-811300	Bio-Bio
Sara Maude	Chine	Santiago	217616329	ZZ

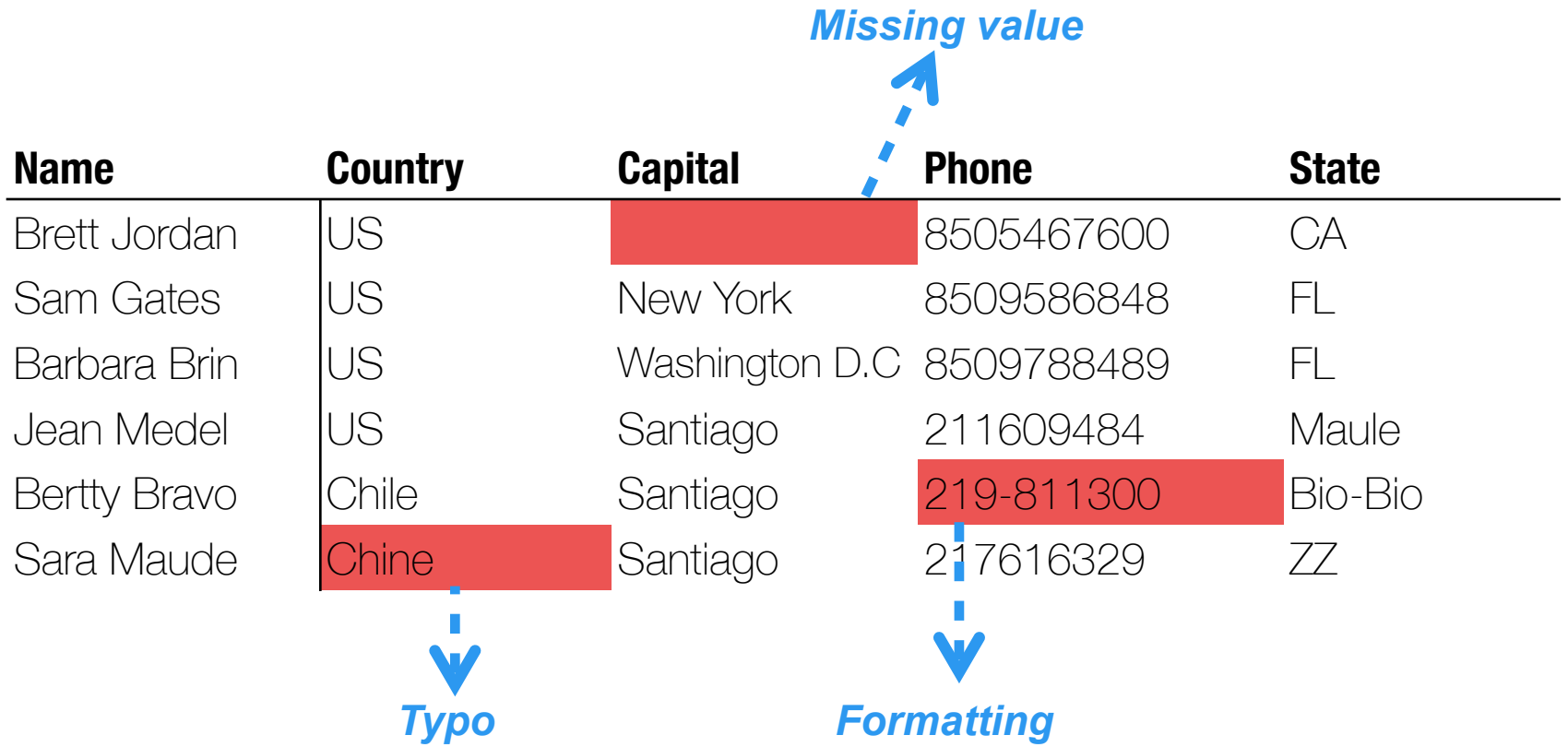
Name	Country	Capital	Phone	State
Brett Jordan	US		8505467600	CA
Sam Gates	US	New York	8509586848	FL
Barbara Brin	US	Washington D.C	8509788489	FL
Jean Medel	US	Santiago	211609484	Maule
Bertty Bravo	Chile	Santiago	219-811300	Bio-Bio
Sara Maude	Chine	Santiago	217616329	ZZ

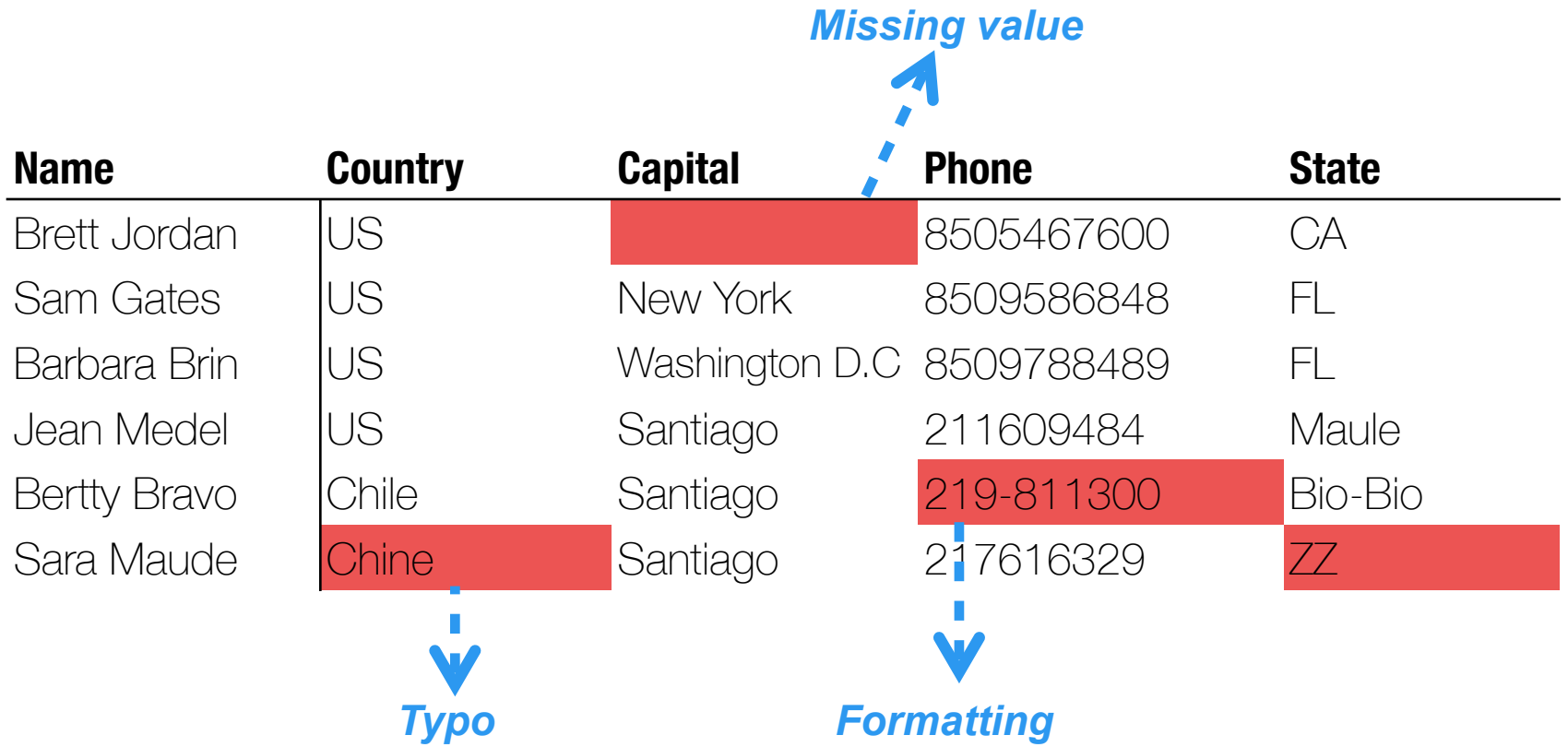
Name	Country	Capital	Phone	State
Brett Jordan	US		8505467600	CA
Sam Gates	US	New York	8509586848	FL
Barbara Brin	US	Washington D.C	8509788489	FL
Jean Medel	US	Santiago	211609484	Maule
Bertty Bravo	Chile	Santiago	219-811300	Bio-Bio
Sara Maude	Chine	Santiago	217616329	ZZ

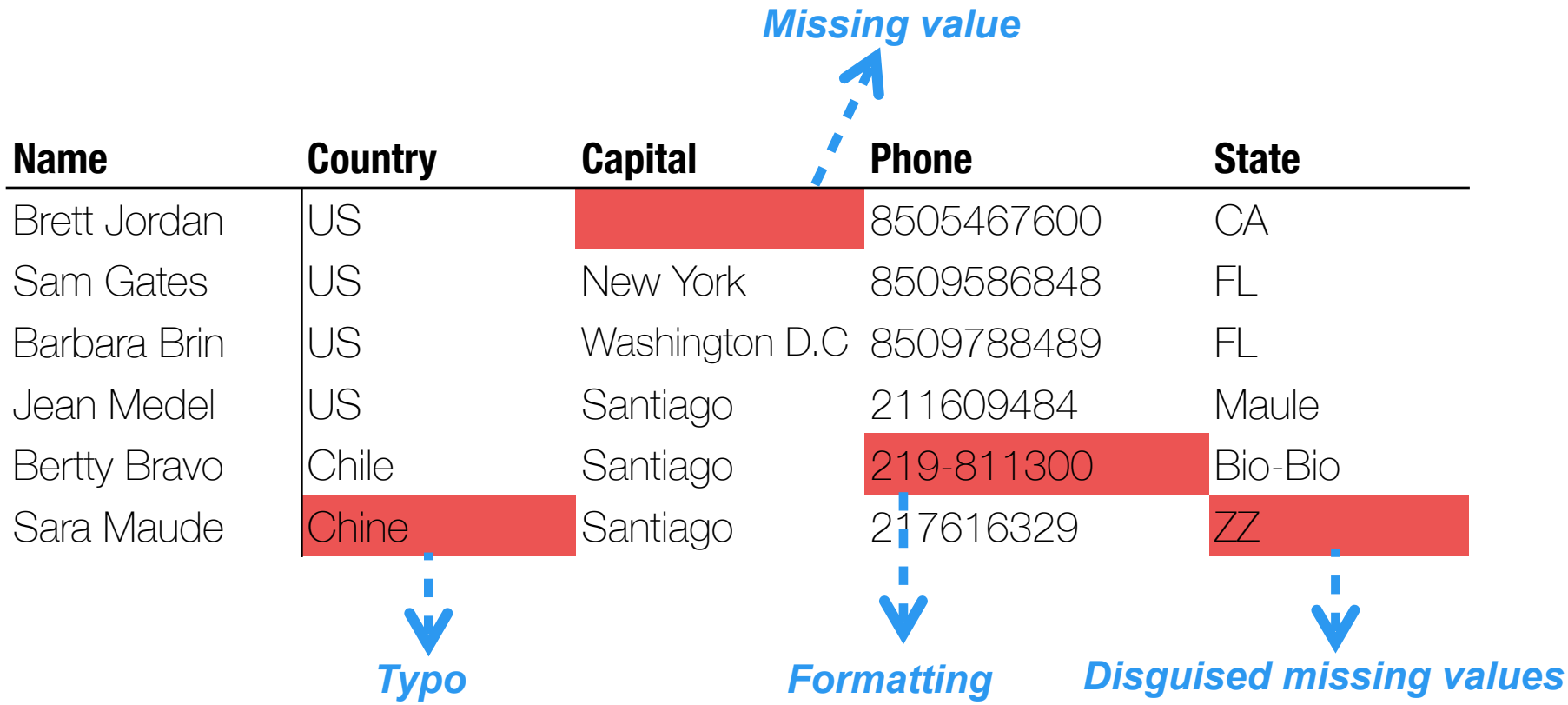
Name	Country	Capital	Phone	State
Brett Jordan	US		8505467600	CA
Sam Gates	US	New York	8509586848	FL
Barbara Brin	US	Washington D.C	8509788489	FL
Jean Medel	US	Santiago	211609484	Maule
Bertty Bravo	Chile	Santiago	219-811300	Bio-Bio
Sara Maude	Chine	Santiago	217616329	ZZ


Missing value
Name
Country
Capital
Phone
State
Brett Jordan
US
8505467600
CA
Sam Gates
US
New York
8509586848
FL
Barbara Brin
US
Washington D.C
8509788489
FL
Jean Medel
US
Santiago
211609484
Maule
Bertty Bravo
Chile
Santiago
219-811300
Bio-Bio
Sara Maude
Chine
Santiago
217616329
ZZ
Typo

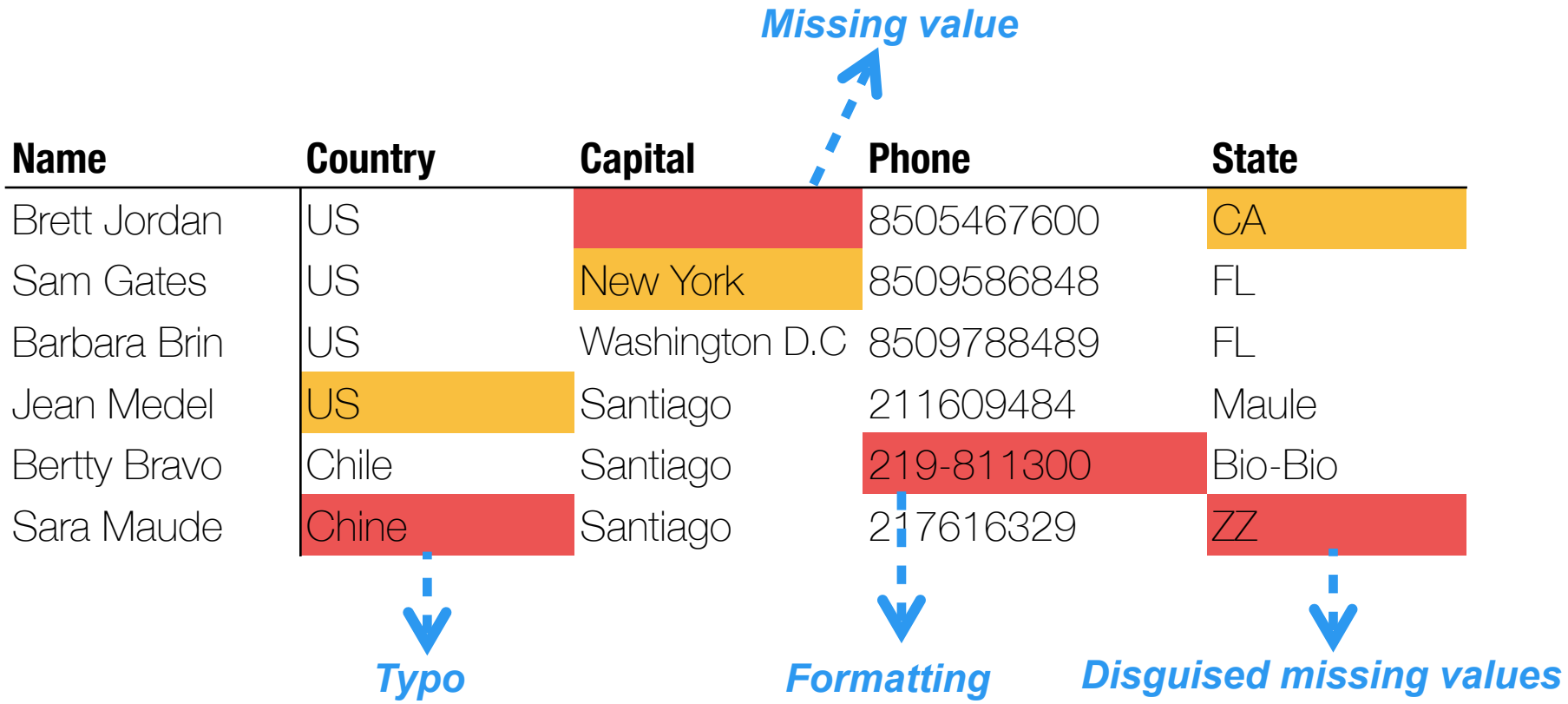

Missing value
Name
Country
Capital
Phone
State
Brett Jordan
US
8505467600
CA
Sam Gates
US
New York
8509586848
FL
Barbara Brin
US
Washington D.C
8509788489
FL
Jean Medel
US
Santiago
211609484
Maule
Bertty Bravo
Chile
Santiago
219-811300
Bio-Bio
Sara Maude
Chine
Santiago
217616329
ZZ
Typo


Missing value
Name
Country
Capital
Phone
State
Brett Jordan
US
8505467600
CA
Sam Gates
US
New York
8509586848
FL
Barbara Brin
US
Washington D.C
8509788489
FL
Jean Medel
US
Santiago
211609484
Maule
Bertty Bravo
Chile
Santiago
219-811300
Bio-Bio
Sara Maude
Chine
Santiago
217616329
ZZ
Typo
Formatting

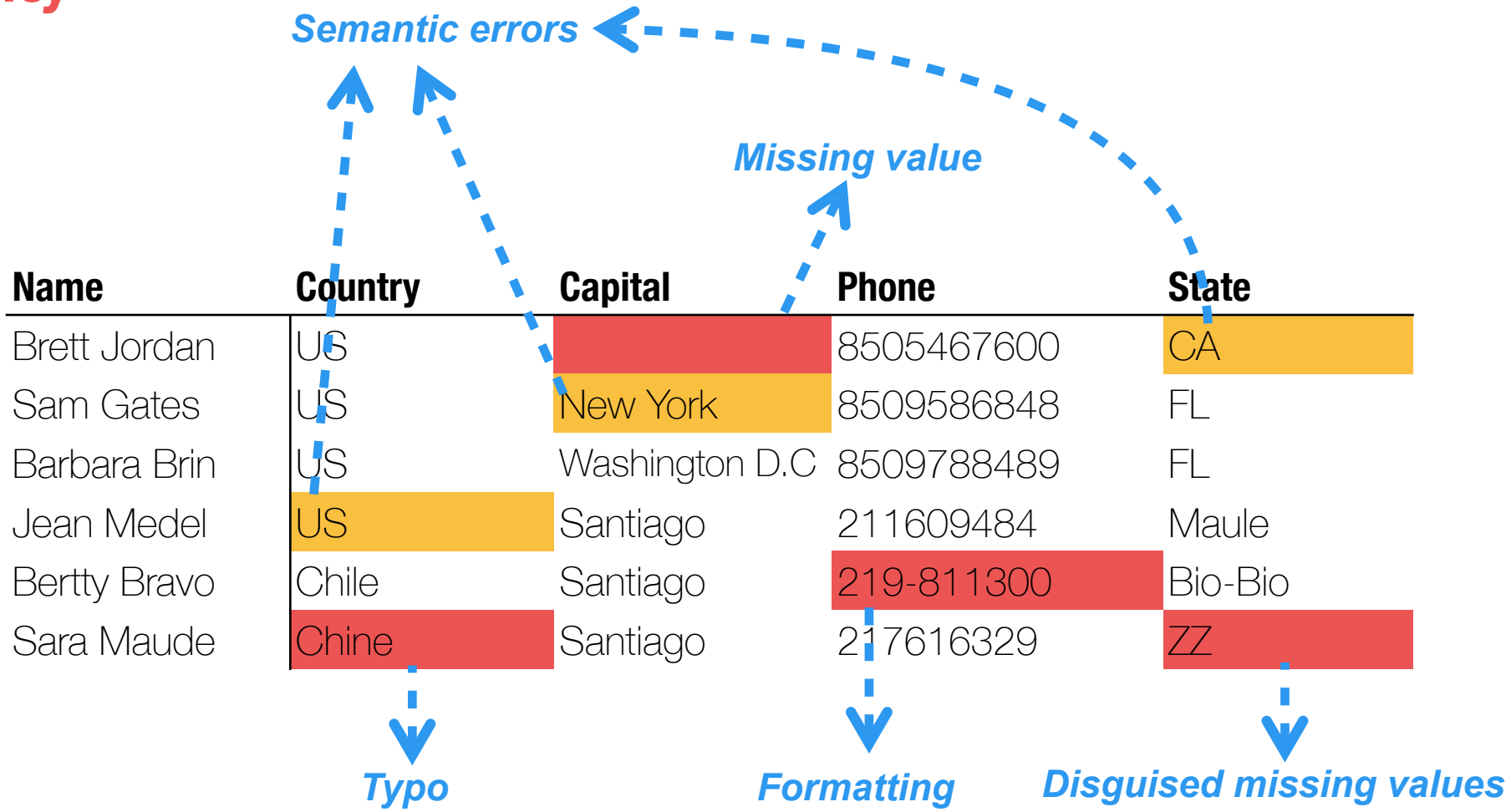

Missing value
Name
Country
Capital
Phone
State
Brett Jordan
US
8505467600
CA
Sam Gates
US
New York
8509586848
FL
Barbara Brin
US
Washington D.C
8509788489
FL
Jean Medel
US
Santiago
211609484
Maule
Bertty Bravo
Chile
Santiago
219-811300
Bio-Bio
Sara Maude
Chine
Santiago
217616329
ZZ
Typo
Formatting


Missing value
Name
Country
Capital
Phone
State
Brett Jordan
US
8505467600
CA
Sam Gates
US
New York
8509586848
FL
Barbara Brin
US
Washington D.C
8509788489
FL
Jean Medel
US
Santiago
211609484
Maule
Bertty Bravo
Chile
Santiago
219-811300
Bio-Bio
Sara Maude
Chine
Santiago
217616329
ZZ
Typo
Formatting
Disguised missing values

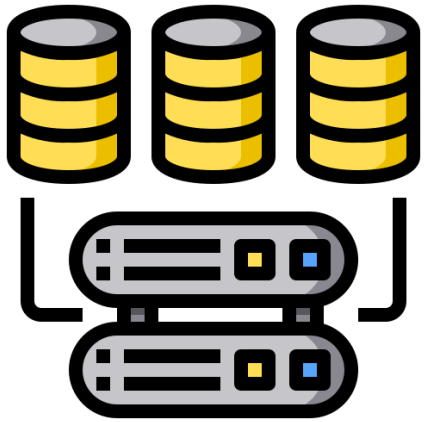
Inconsistency



Inconsistency



Multiple Tables (Data Warehouse)



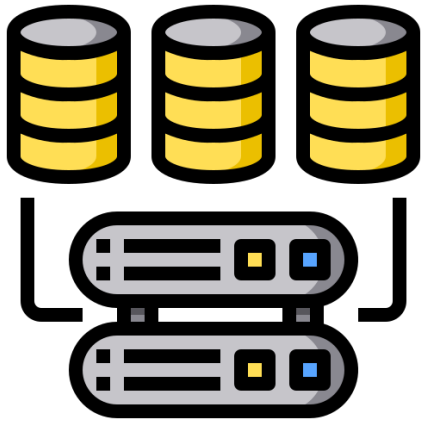
Companies

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View	$210bn
Intl. Business Machines	Armonk, NY	$200bn
Microsoft	Redmond, WA	$250bn

Employees

Company	#-employees
Google	135,301
IBM	297,900
Google	98,771

Multiple Tables (Data Warehouse)



Companies

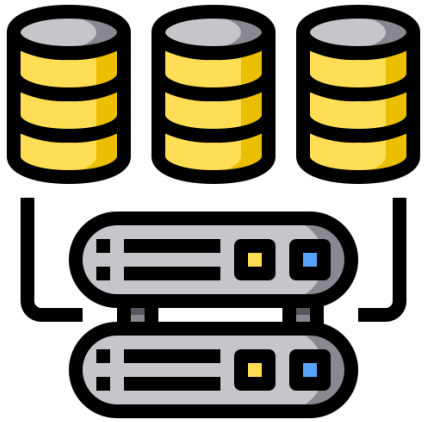
Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View	$210bn
Intl. Business Machines	Armonk, NY	$200bn
Microsoft	Redmond, WA	$250bn

Employees

Company	#-employees
Google	135,301
IBM	297,900
Google	98,771

```
SELECT Company_Name, Address, #-employees
FROM Companies, Employees
WHERE Company_Name = "IBM"
```

Multiple Tables (Data Warehouse)



Companies

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View	$210bn
Intl. Business Machines	Armonk, NY	$200bn
Microsoft	Redmond, WA	$250bn

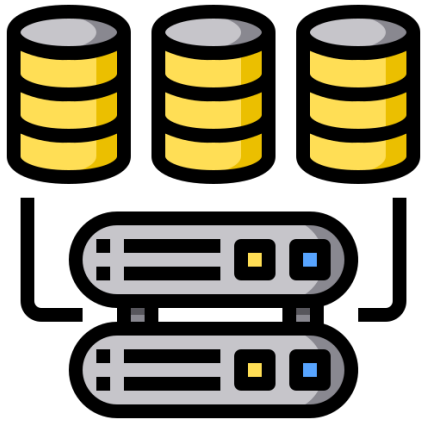
Employees

Company	#-employees
Google	135,301
IBM	297,900
Google	98,771

```
SELECT Company_Name, Address, #-employees
FROM Companies, Employees
WHERE Company_Name = “IBM”
```

→ #-rows: 0

Multiple Tables (Data Warehouse)



Companies

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View	$210bn
Intl. Business Machines	Armonk, NY	$200bn
Microsoft	Redmond, WA	$250bn

Employees

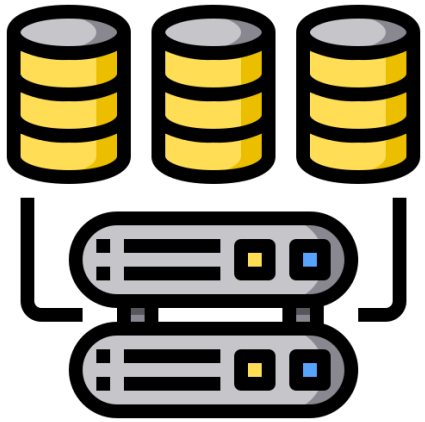
Company	#-employees
Google	135,301
IBM	297,900
Google	98,771

```
SELECT Company_Name, Address, #-employees
FROM Companies, Employees
WHERE Company_Name = "IBM"
```

→ #-rows: 0 →

Problem
Formatting

Multiple Tables (Data Warehouse)



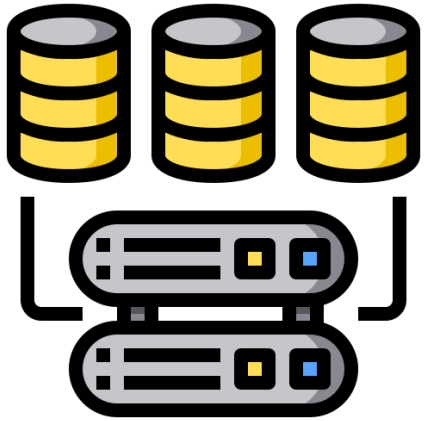
Companies

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View	$210bn
Intl. Business Machines	Armonk, NY	$200bn
Microsoft	Redmond, WA	$250bn

Employees

Company	#-employees
Google	135,301
IBM	297,900
Google	98,771

Multiple Tables (Data Warehouse)



Companies

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View	$210bn
Intl. Business Machines	Armonk, NY	$200bn
Microsoft	Redmond, WA	$250bn

Employees

Company	#-employees
Google	135,301
IBM	297,900
Google	98,771

```
SELECT Company_Name, Address, #-employees
FROM Companies, Employees
WHERE Company_Name = “Google”
```

→ #-rows: 2 →

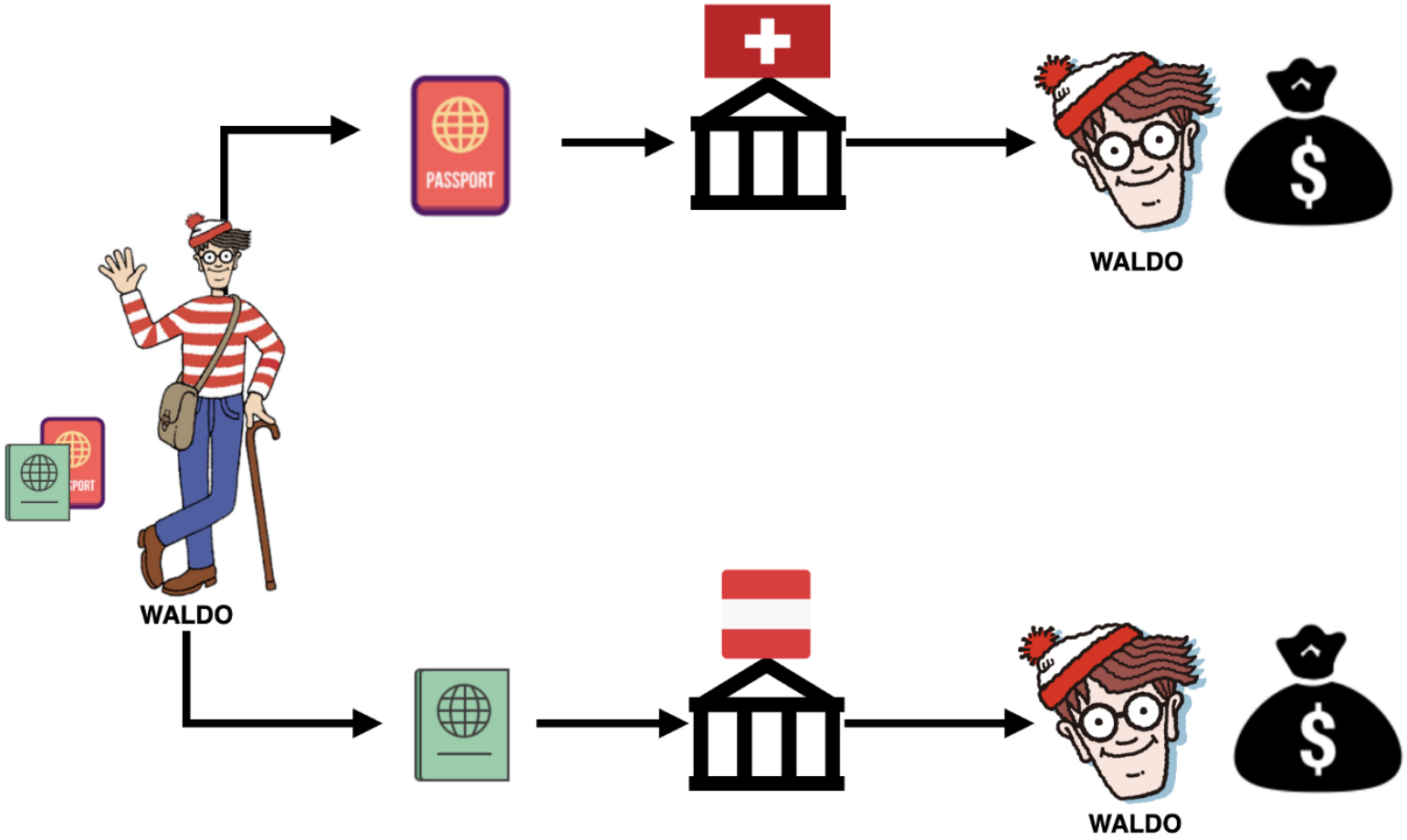
Google	135,301
Google	98,771

Problem
Duplicates

Entity Matching

Entity Matching

Source:[Unit8]



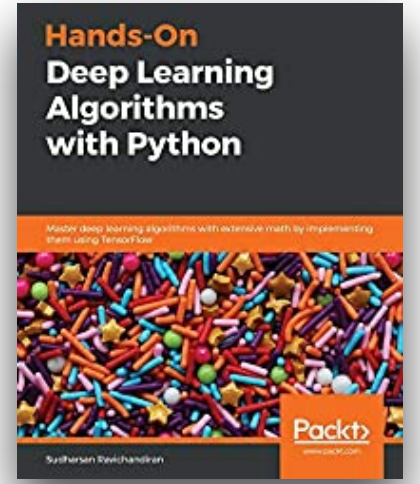
Entity Matching

Source:Unit8

- For example, HSBC was fined **$1.9 billion in 2012** for failure to prevent money laundering by Latin American drug cartels.
- Another example is BNP Paribas, **fined for $8.9 billion in 2015** for violating sanctions against Sudan, Cuba and Iran.
- Hence, *not* being able to find these kind of connections can lead to serious financial and reputation repercussions.
- Such high-profile cases led to more companies adopting entity resolution as part of their business processes.

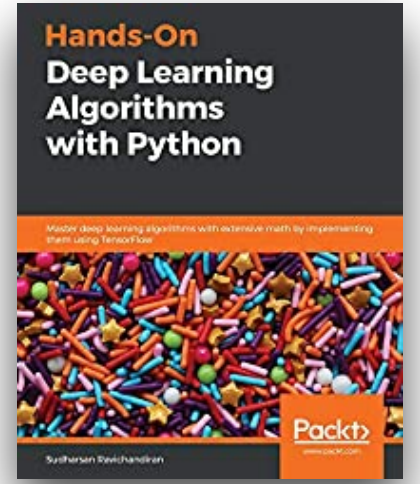
Absolute Good Data vs. Relative Good Data

Absolute good data (traditional data prep):
the “goodness” of the data **is nothing**
about its downstream applications



Absolute Good Data vs. Relative Good Data

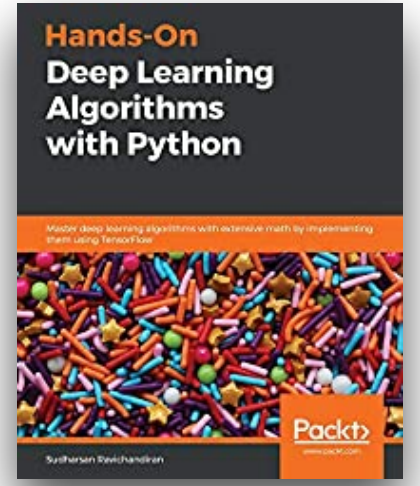
Absolute good data (traditional data prep):
the "goodness" of the data **is nothing**
about its downstream applications



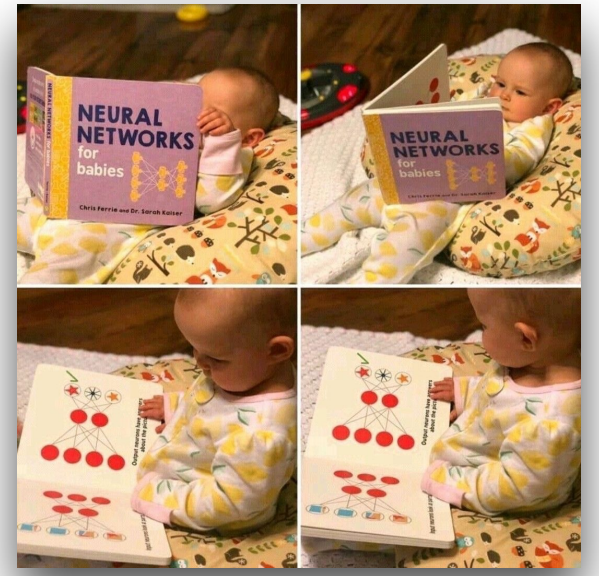
- **Relative good data:**
  the "goodness" of the data **depends**
  on its downstream applications (e.g., **AI**)

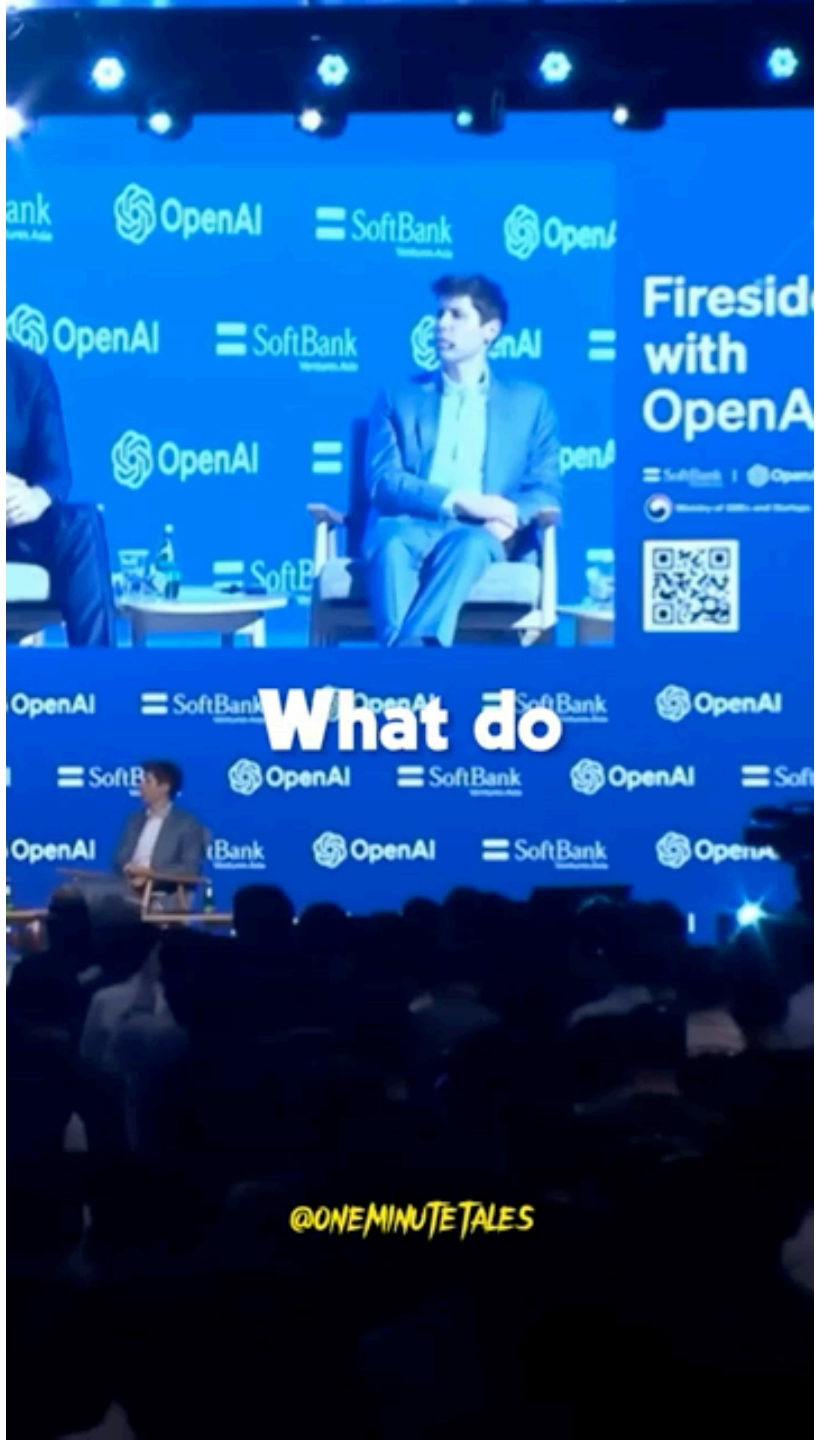
Absolute Good Data vs. Relative Good Data

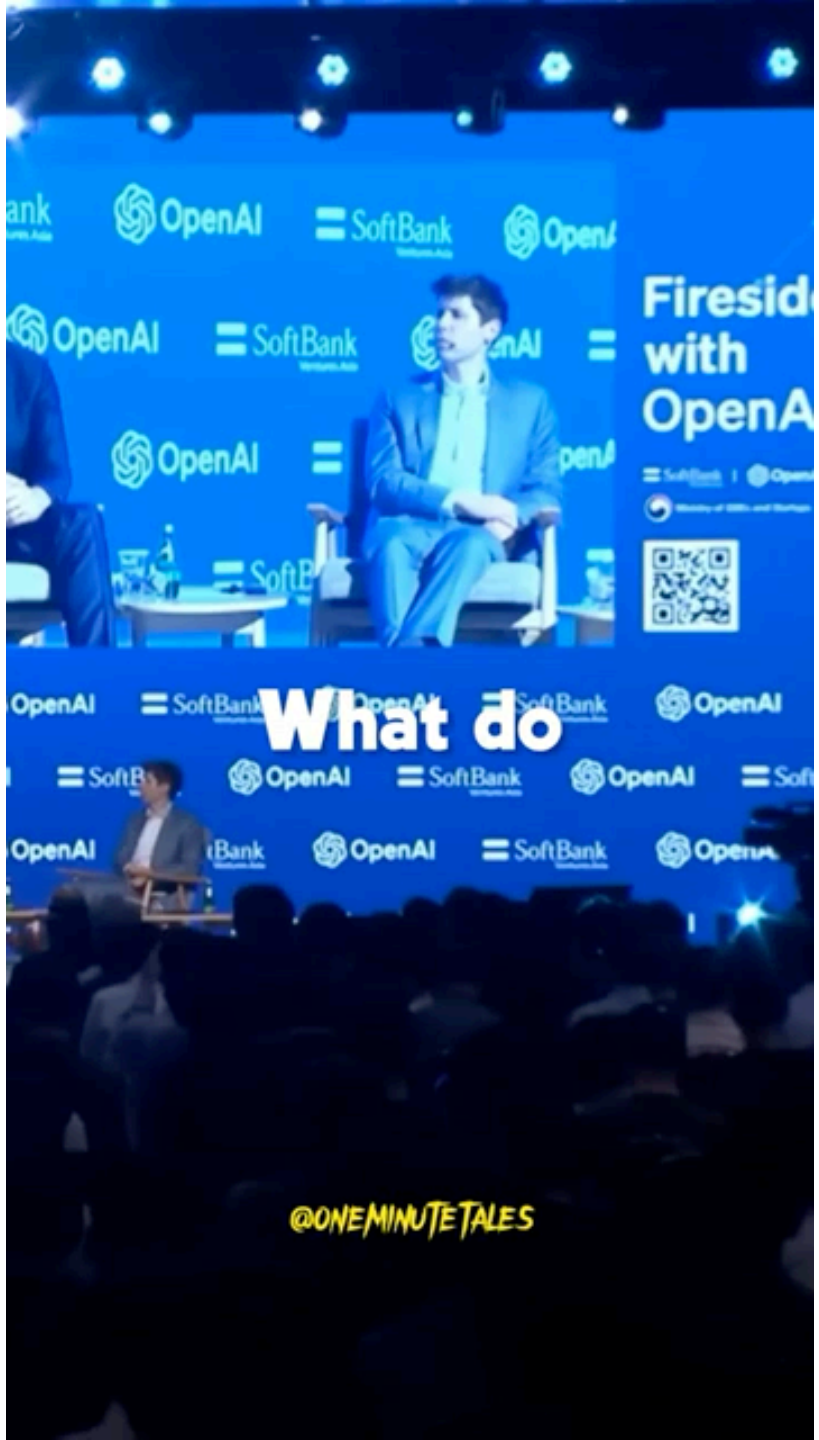
Absolute good data (traditional data prep):
the "goodness" of the data **is nothing**
about its downstream applications

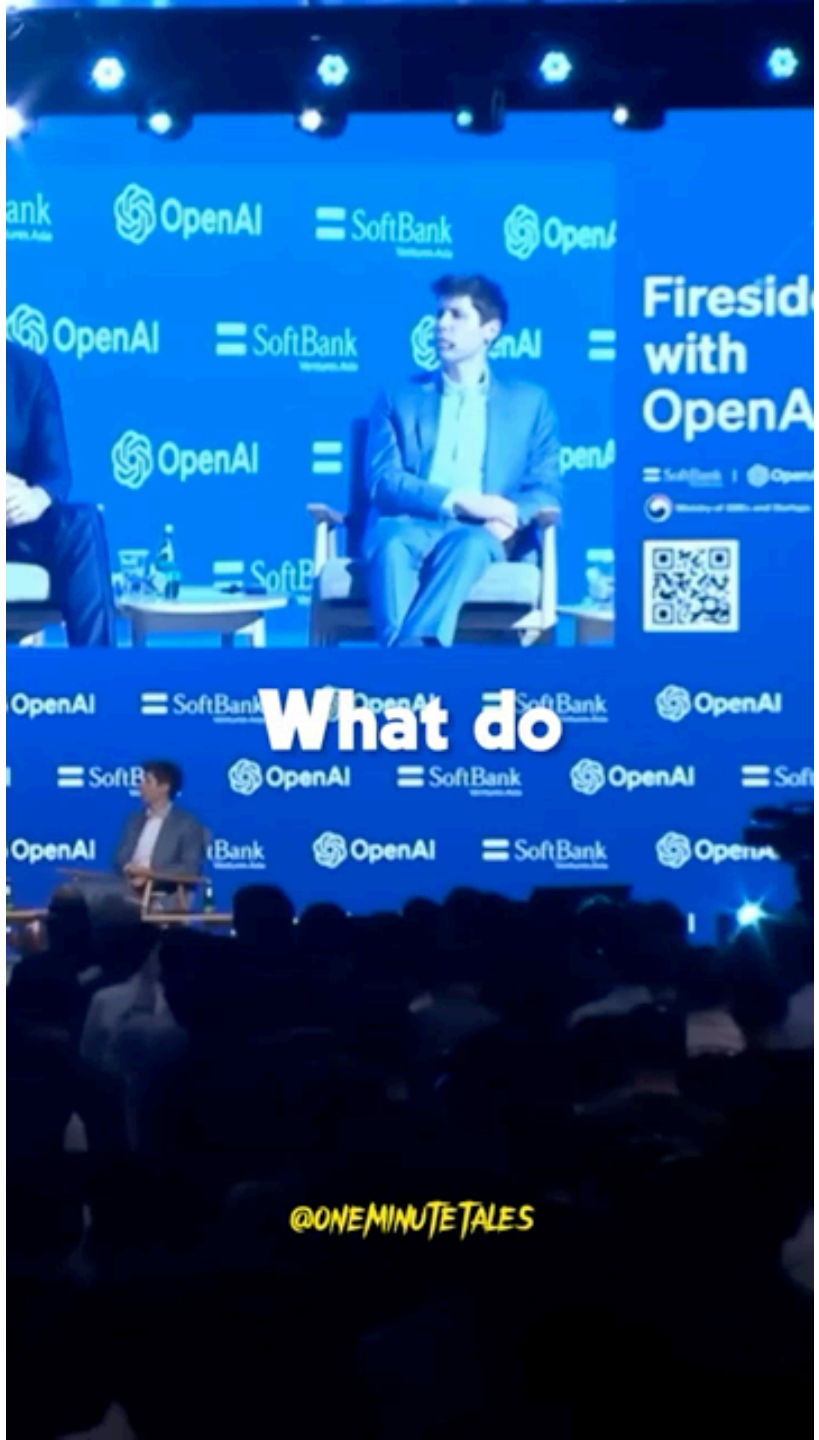


- **Relative good data:**
  the "goodness" of the data **depends**
  on its downstream applications (e.g., **AI**)

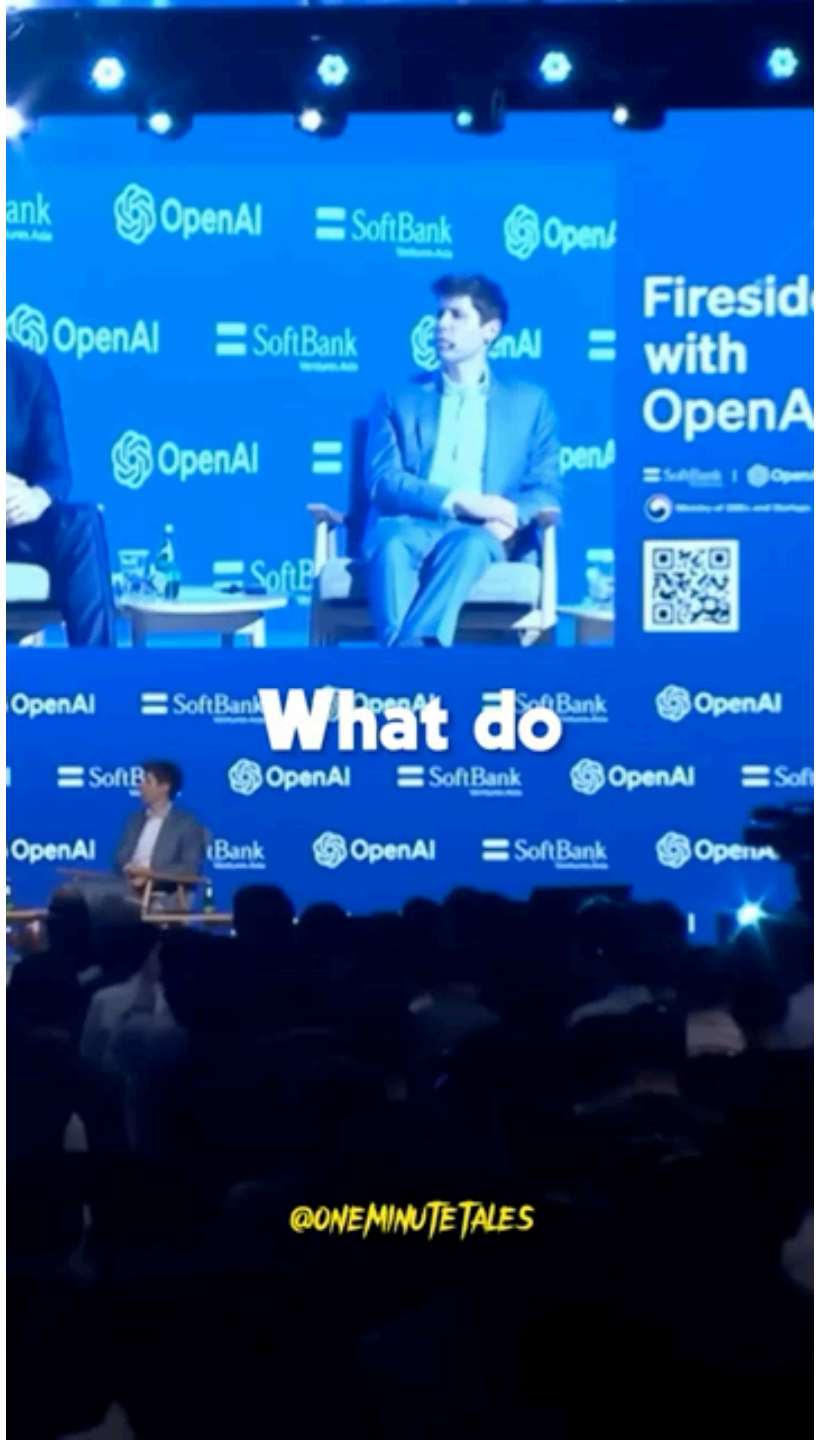



OpenAI
SoftBank
Fireside
with
OpenA
What do
@ONEMINUTETALES


OpenAI
SoftBank
Fireside
with
OpenA
What do
@ONEMINUTETALES



Data Science: Unleashing Your Inquisitive Mind



Data Science: Unleashing Your Inquisitive Mind