DATA SCIENCE & MACHINE LEARNING WORKSHOP

Week 2 - Understanding Data

AGENDA

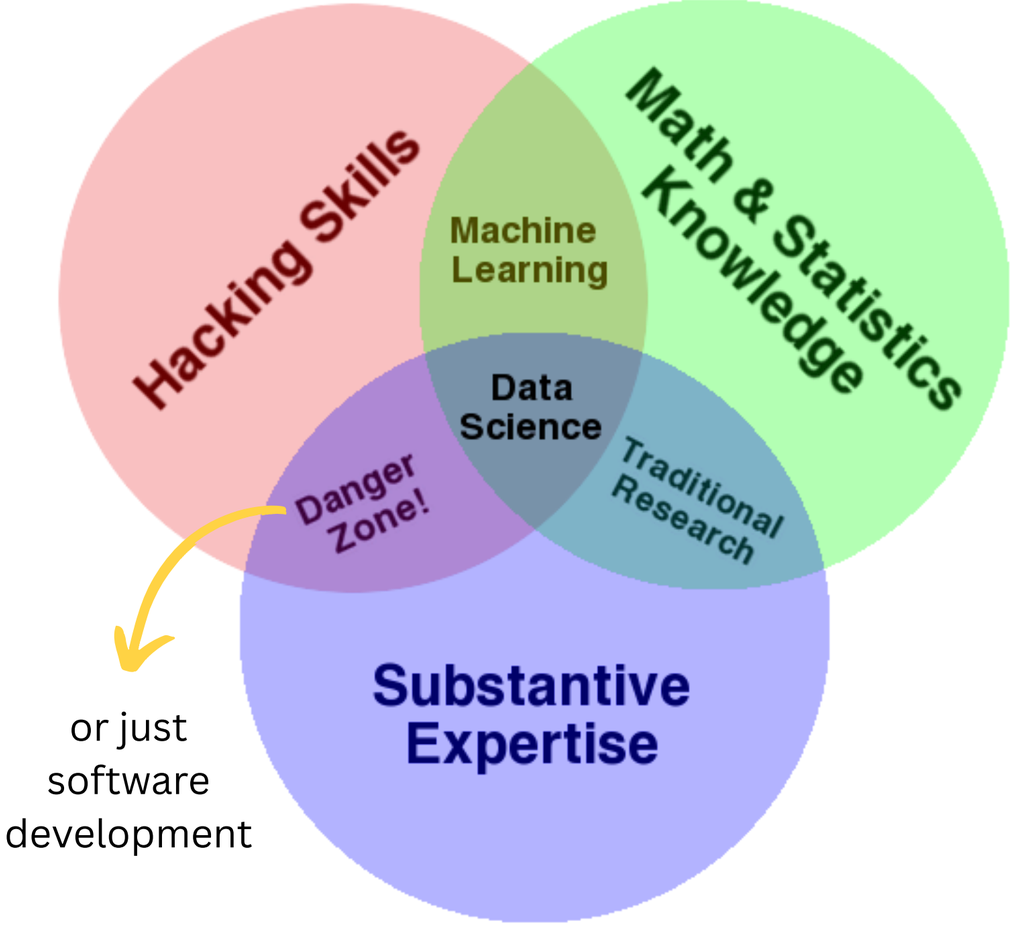
Week 2 topic:

- Recap week 1
- Types of data
- structured vs. unstructured data
- Common data sources
- missing data, outliers, inconsistencies
- basic data wrangling - extracting features
- Intro to Pandas

We are also planning to have a
Quantitative Finance workshop and
a Datathon next semester!

Full agenda this semester:
https://bit.ly/DataScienceAlgosoc

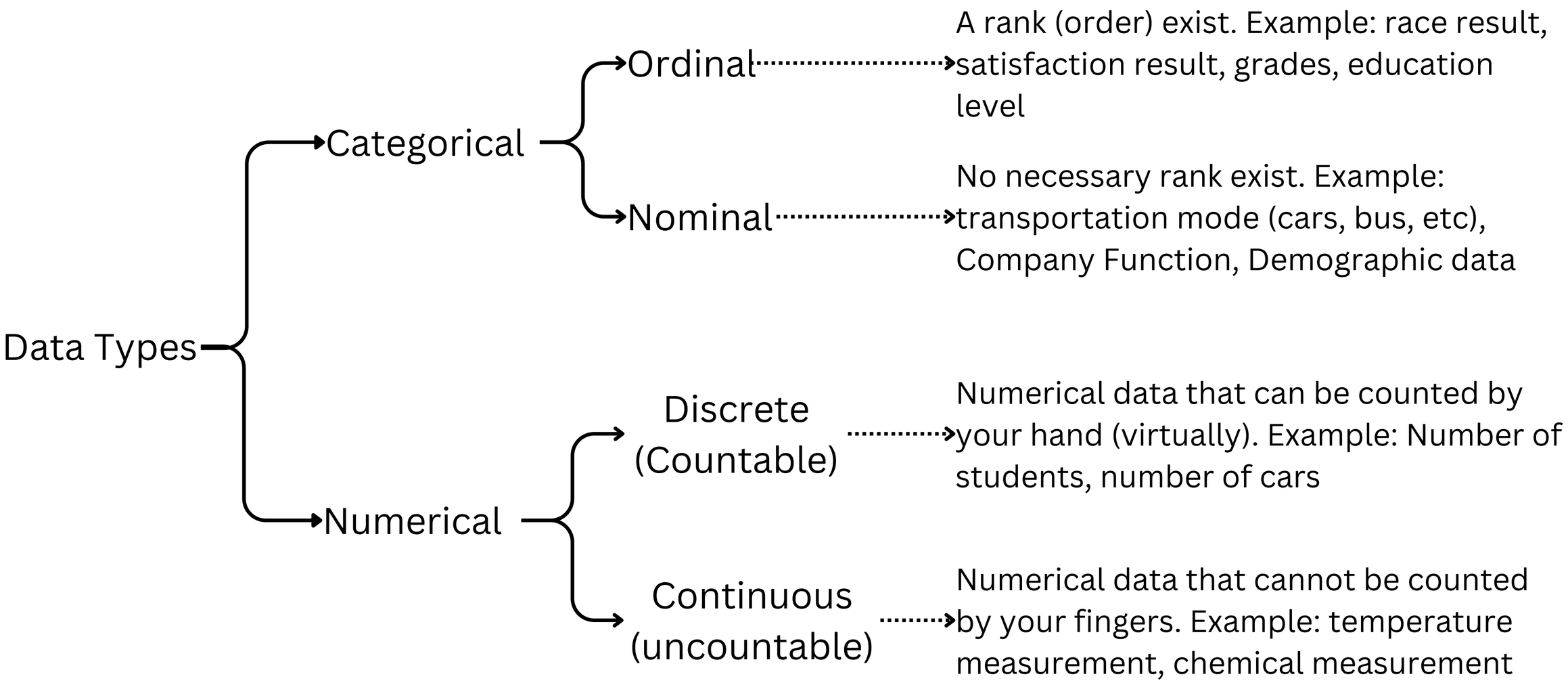
WEEK 1 RECAP



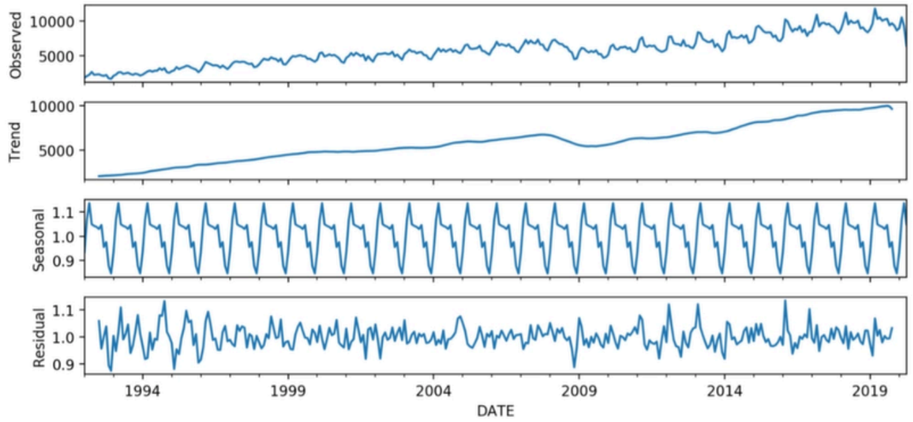
The first important step in the data science workflow is to ask the right question.

Find the right problem to be solved

DATA TYPES


Data Types
Categorical
Ordinal
A rank (order) exist. Example: race result, satisfaction result, grades, education level
Nominal
No necessary rank exist. Example: transportation mode (cars, bus, etc), Company Function, Demographic data
Numerical
Discrete (Countable)
Numerical data that can be counted by your hand (virtually). Example: Number of students, number of cars
Continuous (uncountable)
Numerical data that cannot be counted by your fingers. Example: temperature measurement, chemical measurement

(OTHER) DATA TYPES



Time Series Data: If you can draw information from the data based on their time sequence, they are time series data. Example: Temperature along the year (seasonality in temperature), Audio signal

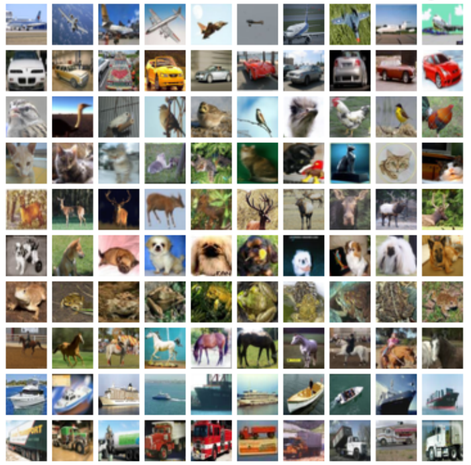


Image Data: A lot of current AI applications support the processing of images. Various information on images can be extracted (e.g., edges, colours, texture) and can be feeded to a processing agent for further use.

Text Data: Similarly to image data, texts can be regarded as data. However, similar to image data too, text data needs further processing for information to be extracted (ngrams, TF-IDF)

STRUCTURED VS UNSTRUCTURED

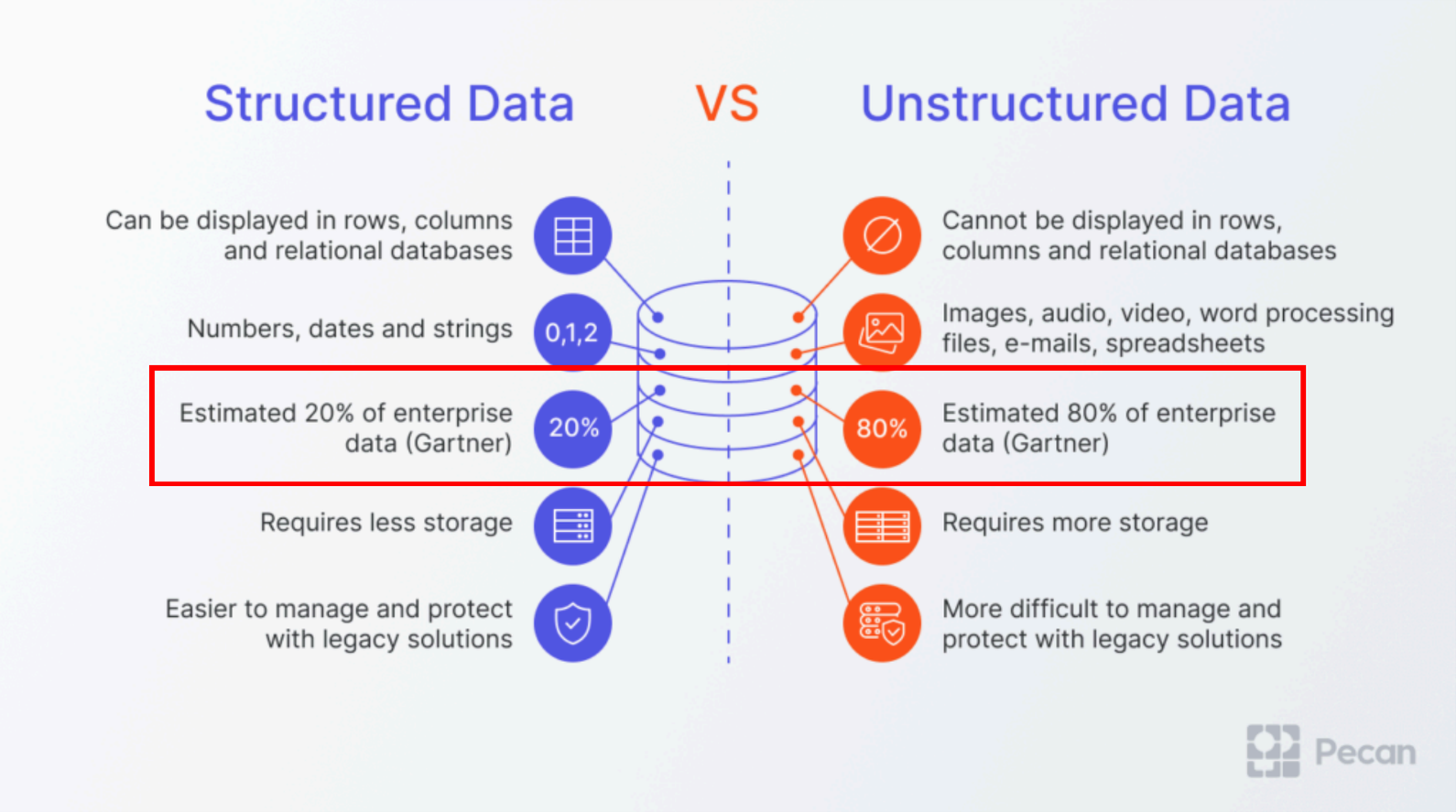


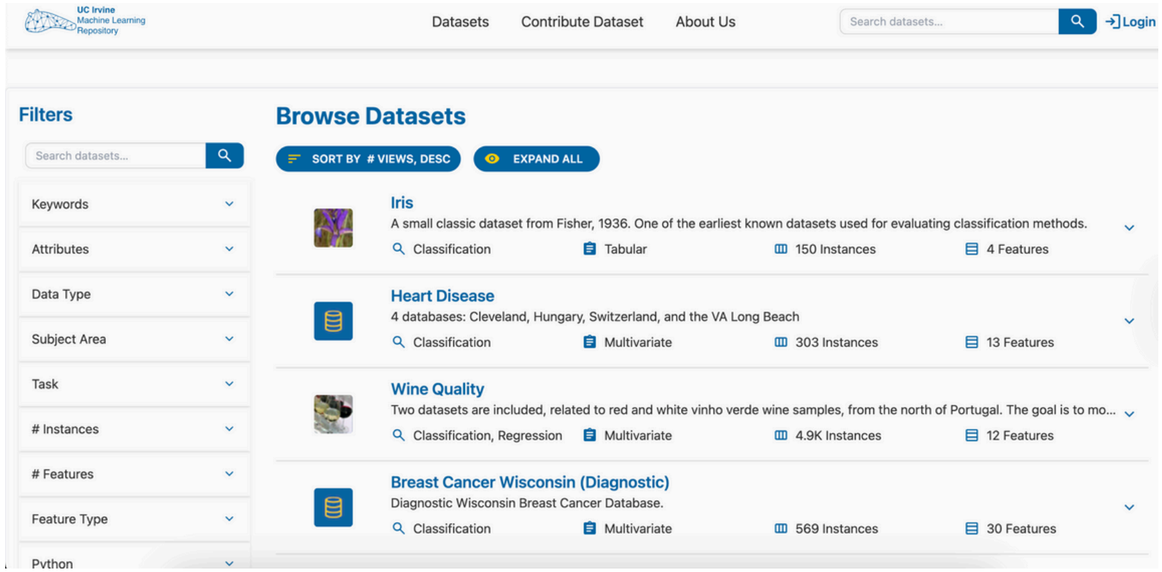
Image Source: https://www.pecan.ai/blog/what-is-structured-data/

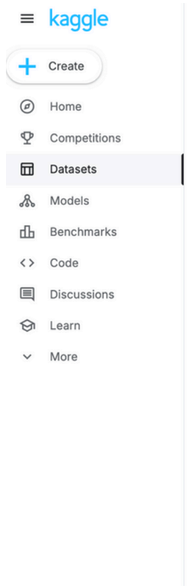
Texts and Images can be regarded as unstructured data, alongside audio (signal) data and measurement data; they are data that essentially need processing to become structured data (that fits in rows and columns).

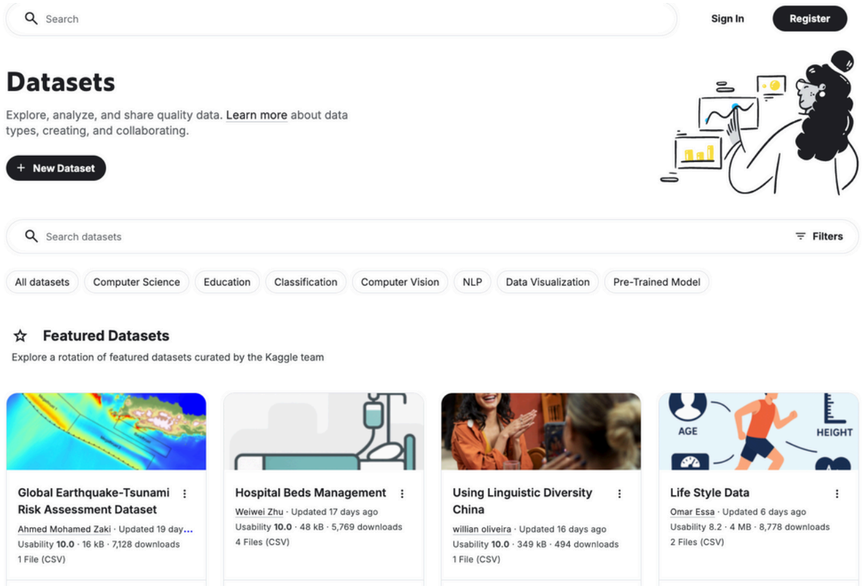
*Keep in mind, 80% of data on the business sector come in the form of unstructured data.

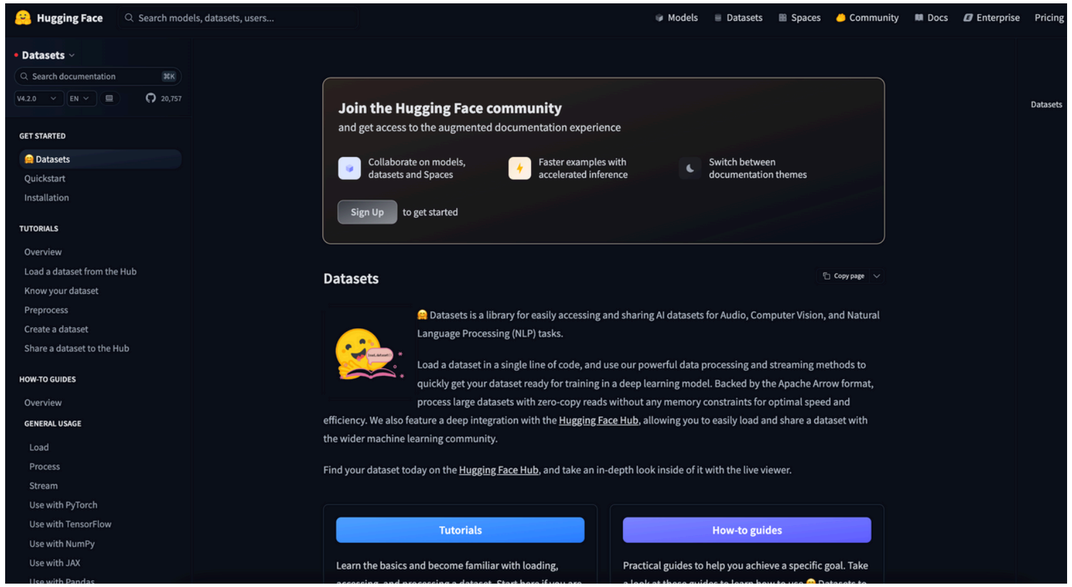
Therefore, it is important to know what features or information to extract from the data!

DATA SOURCES









for practice or experimental purposes, it is good to use data from **Kaggle, UCI Repository, or HuggingFace**. Manual lookup on **research articles** can also be performed.

However, for making a **production-ready model**, these types of data are **far from sufficient**. Although some research article data may suffice, most are not.

DATA SOURCES

Why can’t we just build a model from Kaggle datasets?

It mostly concerns the epistemic uncertainty (our lack of knowledge of the data). We don’t really know how the ‘nature’ of the data is. Therefore, we would usually have a form of **ML-OPS** workflow for production-scale models

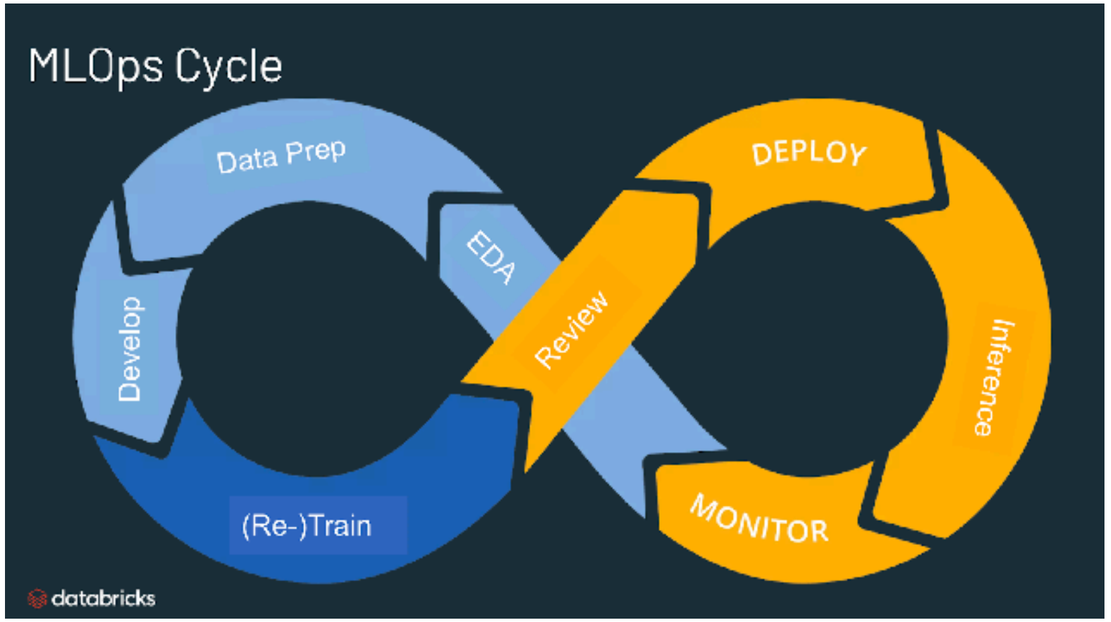


Image Source: https://www.databricks.com/glossary/mlops

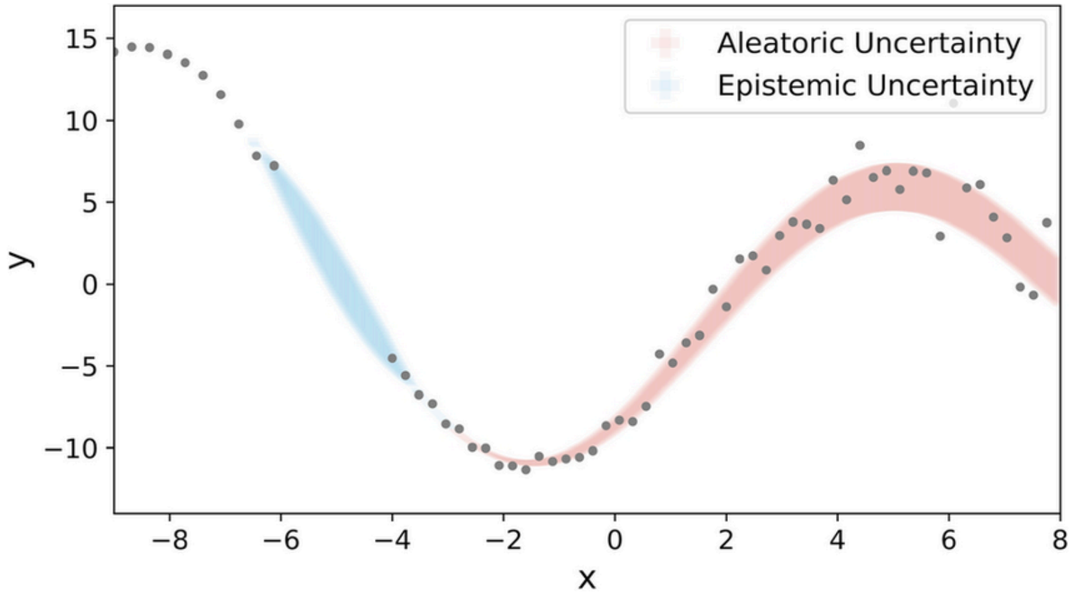


Image Source: https://www.researchgate.net/figure/An-illustration-of-the-difference-between-aleatoric-and-epistemic-uncertainties-The-dots_fig1_368244388

ML-OPS

This workflow would help data to be representative to our business need. If a change to the enviornment happens, we would be able to monitor the deployed model and how well its perfromance is.

DATA SOURCES

In real life practice, data can come from many different sources and be in many forms, depending on where you are working:

Energy Sector- Sensors and measurement
Financial Sector-Market, News
Political Sector-News, Surveys, Polls
Scientific Sector-Lab experiments
Social Sector-Observation, Surveys

MISSING DATA, OUTLIERS, INCONSISTENCIES

ALGOSOC UOB

Problems with acquired data

In real life cases, a data scientist will most likely handle ‘dirty’ data sets after performing data acquisition. Problems would include:

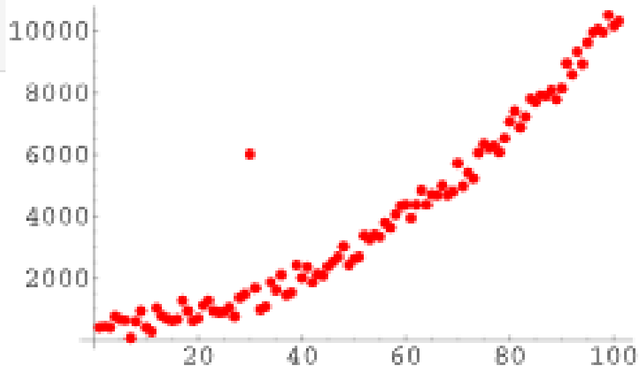
- Missing values
- Outliers
- Type and Value Inconsistencies

```
In [1]: import pandas as pd

        dataset = pd.read_csv("C:/Users/Admin/Desktop/Blog/Missing values/data.csv")
        dataset
```

Out[1]:

	Height	Weight	Country	Place	Number of days	Some column
0	12.0	35.0	India	Bengaluru	1.0	NaN
1	NaN	36.0	US	New York	2.0	NaN
2	13.0	32.0	UK	London	NaN	NaN
3	15.0	NaN	France	Paris	4.0	NaN
4	16.0	39.0	US	California	5.0	12.0
5	NaN	NaN	NaN	Mumbai	NaN	NaN
6	NaN	NaN	NaN	NaN	6.0	NaN




```
TechSupport: ['No' 'Yes']
StreamingTV: ['No' 'Yes']
StreamingMovies: ['No' 'Yes']
Contract: ['Monthly' 'One year' 'Two year']
PaperlessBilling: ['Yes' 'No']
PaymentMethod: ['Manual' 'Bank transfer (automatic)' 'Credit card (automati
TotalCharges: ['29.85' '1889.5' '108.15' ... '346.45' '306.6' '6844.5']
Churn: ['No' 'Yes']
```

Image Source: https://www.datasciencesmachinelearning.com/2018/11/handling-missing-values-in-python.html

Image Source: https://mathworld.wolfram.com/Outlier.html

MISSING DATA, OUTLIERS, INCONSISTENCIES

Problems with each on a data science perspective

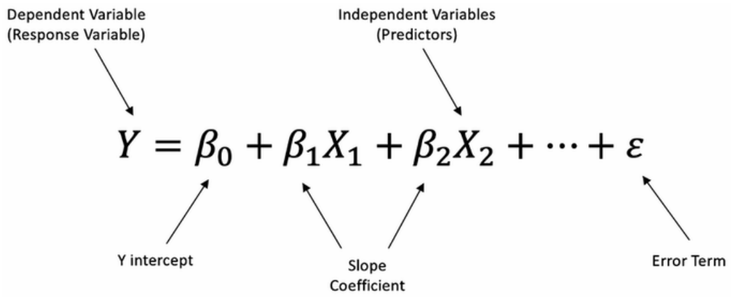


Image Source: https://www.researchgate.net/figure/Linear-regression-equation_fig1_373123252

Say, we are predicting something with a linear function...

If Y : House price, X1: number of floors, X2: year built

When the house was not registered with the year it was built, the function cannot be calculated... Hence, no prediction

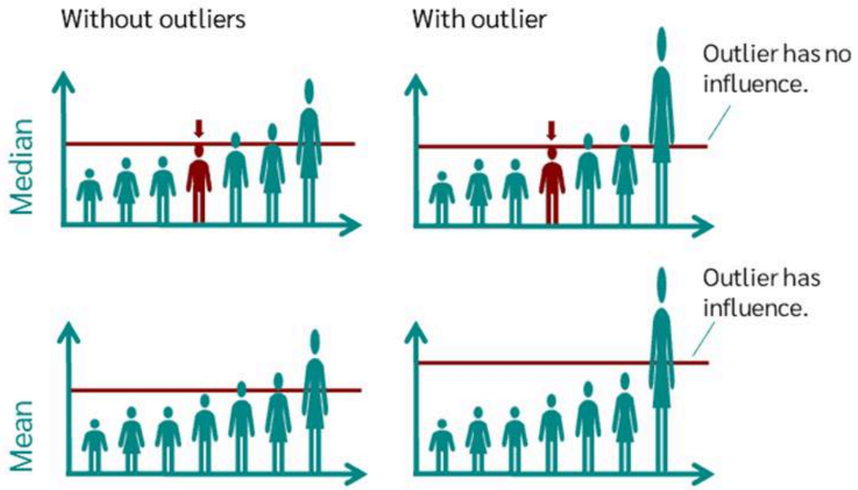


Image Source: https://simbi.in/understanding-central-tendency-mean-median-mode-and-standard-deviation/

Mean average can be heavily affected with the presence of outliers

Median can be a good centrality measure, but it may pose a problem with a really skewed distribution and really small sample size

```
TechSupport: ['No' 'Yes']
StreamingTV: ['No' 'Yes']
StreamingMovies: ['No' 'Yes']
Contract: ['Monthly' 'One year' 'Two year']
PaperlessBilling: ['Yes' 'No']
PaymentMethod: ['Manual' 'Bank transfer (automatic)' 'Credit card (automatic)']
TotalCharges: ['29.85' '1889.5' '108.15' ... '346.45' '306.6' '6844.5']
Churn: ['No' 'Yes']
```

If we are building a machine learning model...

and we input the wrong data type, it may treat the wrongly-typed numericals as categorical data

If we are making a dashboard...

A proper histogram may not be able to be computed as the platform would treat the data as strings.

DATA WRANGLING

There are several ways to handle missing values:

- Drop the row; **Problem:** what if there are lots of missing rows?
- Fill in with median/centrality; **Problem:** same as before, a lot of rows would have the same value
- Use ML-based imputation; **Caveat:** Make sure the method is correct
- Another way is to combine using centrality or ML-based impuation with **domain knowledge**
  - Example: Partition data into several groups (e.g., age groups) and use centrality or perform ML-based imputation

There are several ways to handle outliers:

- Drop the row; **Problem:** what if there are a handful row of outliers?
- Use logarithmic transformation (X → log(X))
  - However it only works for right-skewed data

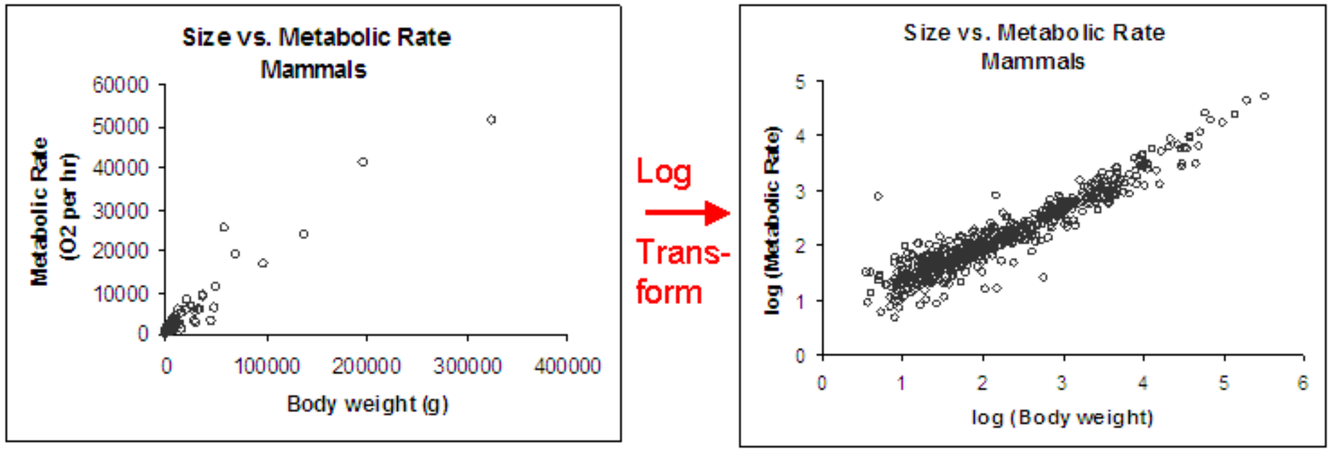


Image Source: https://mathbench.umd.edu/modules/misc_scaling/page07.htm

Ways to handle inconsistent data?

- Perform manual cleaning
  - e.g., identify and change a mis-typed column (string → integer)

DATA WRANGLING

Feature Engineering

Feature engineering can include the action of adding new information by using domain knowledge of the existing data.

This can also include the act of changing the ‘form’ of the data to be friendlier for the machine learning model to learn from.

Example:

If we have a column of passengers, **age** (number), **names** (string), and **travelling with family** (boolean [0 or 1]), we can extract:

- First name and last name columns
- (Big bet) If they are travelling with a family, we can get the size of the family by matching the last name columns and the condition of travelling with a family
- (Another bet) we can add a column ‘travelling with kids’ and ‘number of kids’ by matching the age, lastname, and travelling with family column
- Missing Data can also be imputed, e.g., put the value of ‘1’ to the passenger that is unknown in terms of travelling with family but has the same last name to another passenger

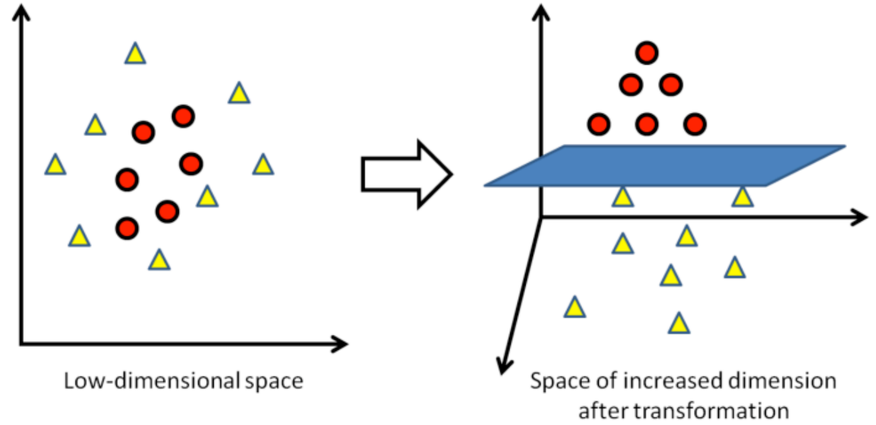


Image Source: https://www.researchgate.net/figure/Kernel-trick-By-transforming-the-original-space-left-into-a-space-of-increased_fig1_305284381

Name	Age	Travelling with Family
David Stone	35	1
Julia Stone	33	1
Mary Stone	8	NaN
Andrew Garfield	20	0

HANDS ON

ALGOSOC
UOB

Colab link provided in live workshop session