



# **DATA SCIENCE & MACHINE LEARNING WORKSHOP**

Week 2 - Understanding Data

# INDUSTRY APPLICATIONS MASTERCLASS

**ARE YOU LOOKING FOR AN INTERNSHIP OR  
PLACEMENT?**

**WONDERING HOW TO START APPLYING OR  
LOOKING FOR SOME TIPS THAT COULD  
ULTIMATELY BOOST YOUR APPLICATIONS?**

**THIS EVENT IS GOING TO HAPPEN ON  
WEDNESDAY 29TH OCTOBER AT THE MURRAY  
LEARNING CENTRE (THE BUILDING OPPOSITE  
THE COMPUTER SCIENCE BUILDING) IN UG09.**

**LIMITED AVAILABILITY!!!  
SCAN THE QR CODE TO SECURE YOUR SPOT!!!**



<https://luma.com/sx1nfjll>

**TALK FROM PREVIOUS INTERNS AT:**



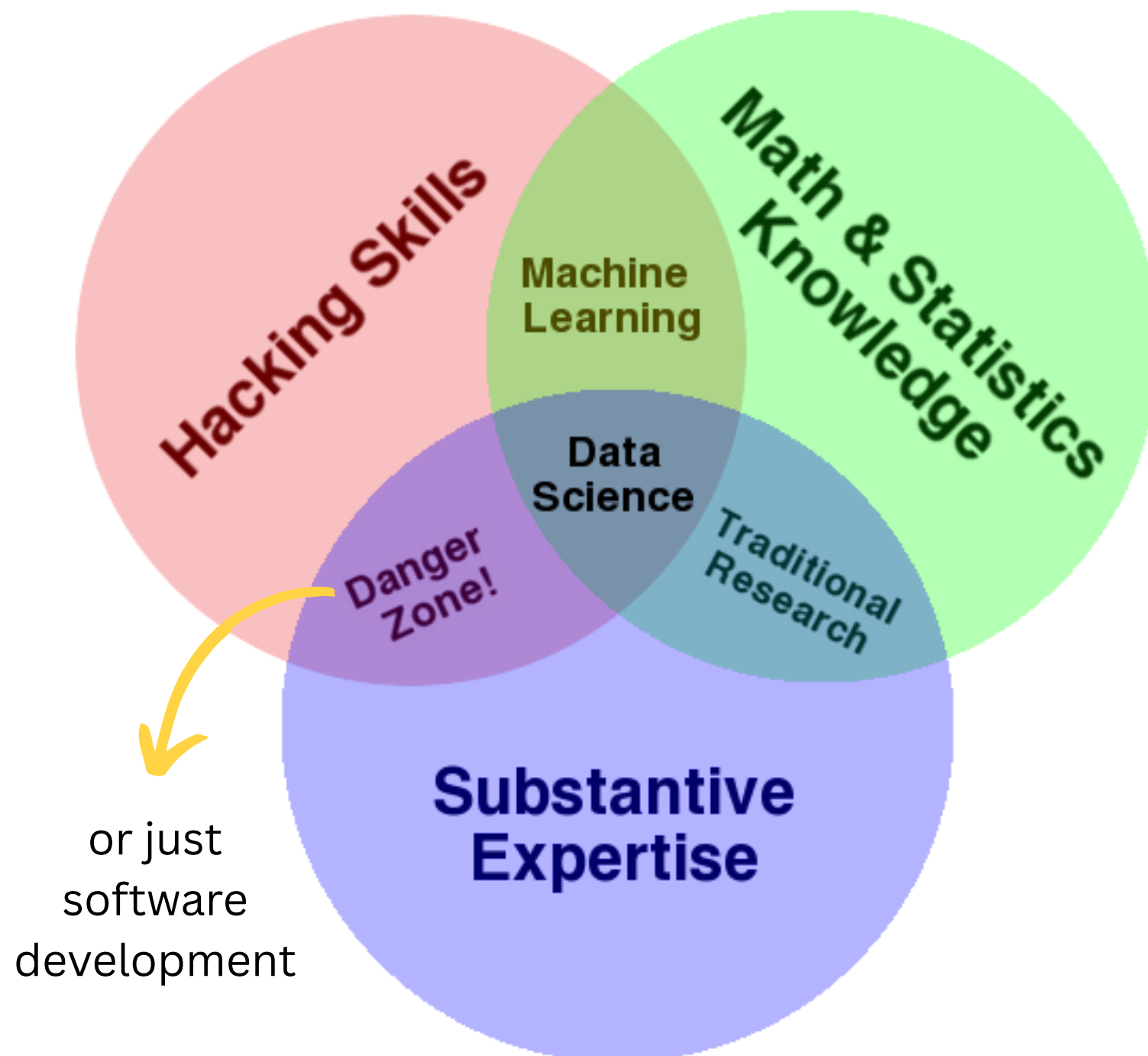
## Week 2 topic:

- 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!

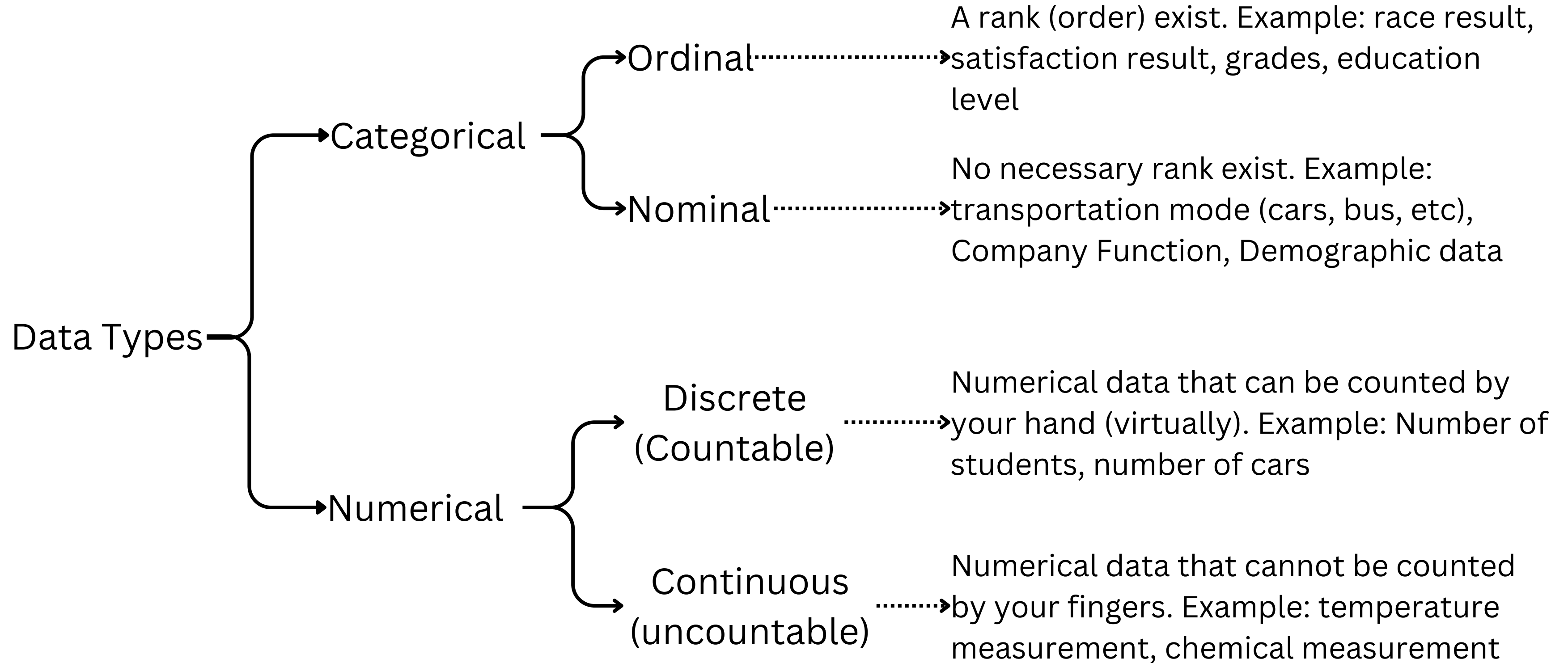
Full agenda this semester:  
**<https://bit.ly/DataScienceAlgosoc>**

# 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**







# STRUCTURED VS UNSTRUCTURED

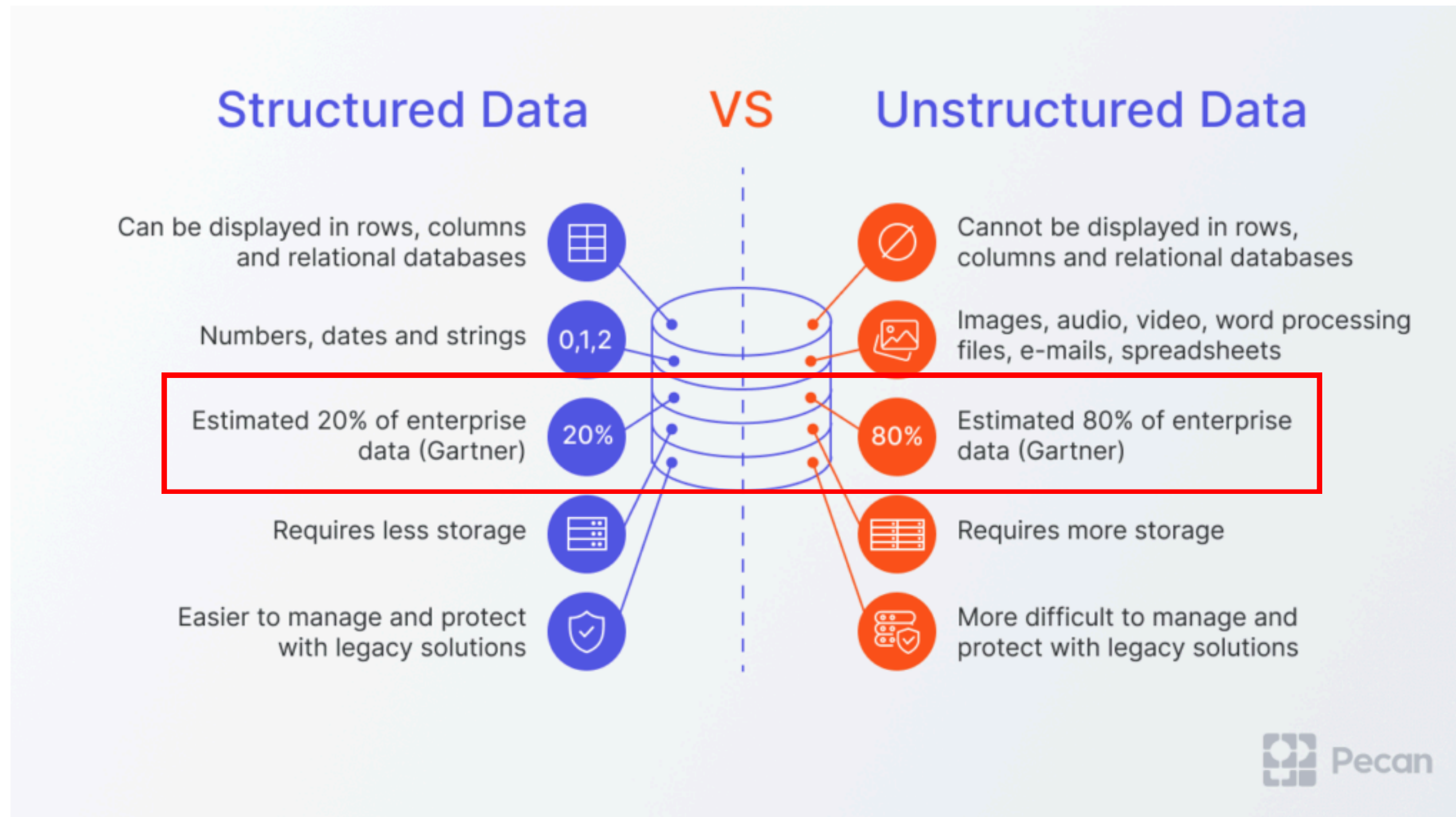


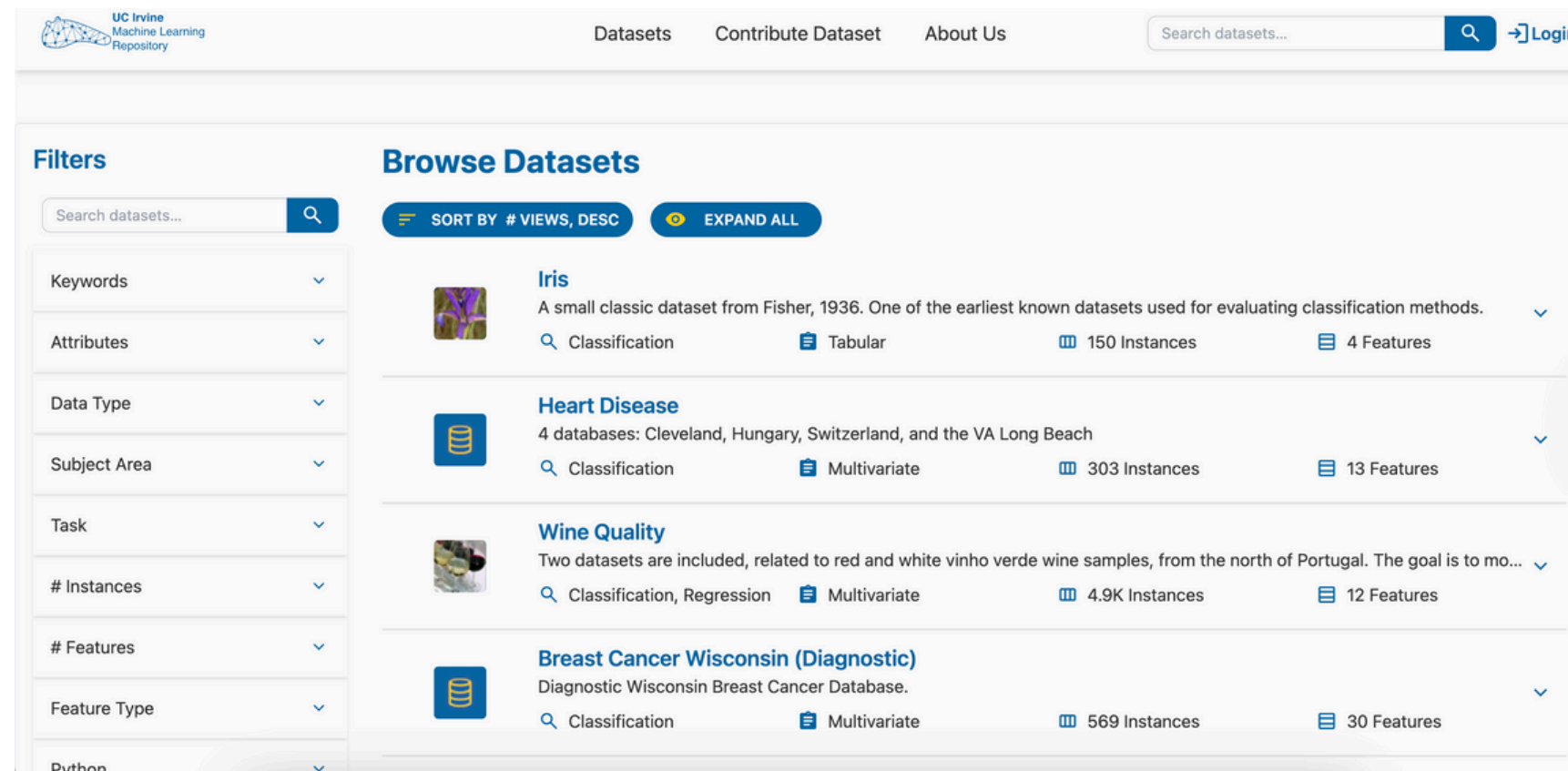
Image Source: <https://www.pecan.ai/blog/what-is-structured-data/>

**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).

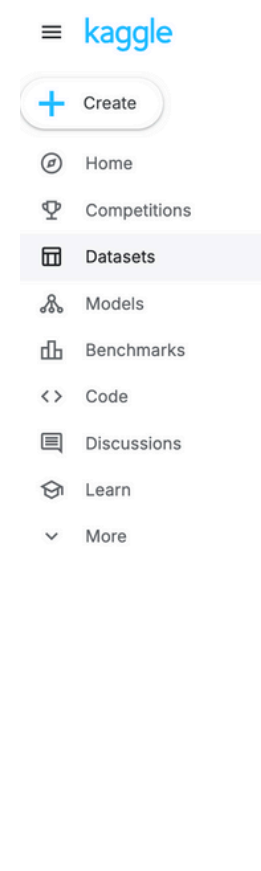
\*Keep in mind, 80% of data on the business sector come in the form of unstructured data.

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

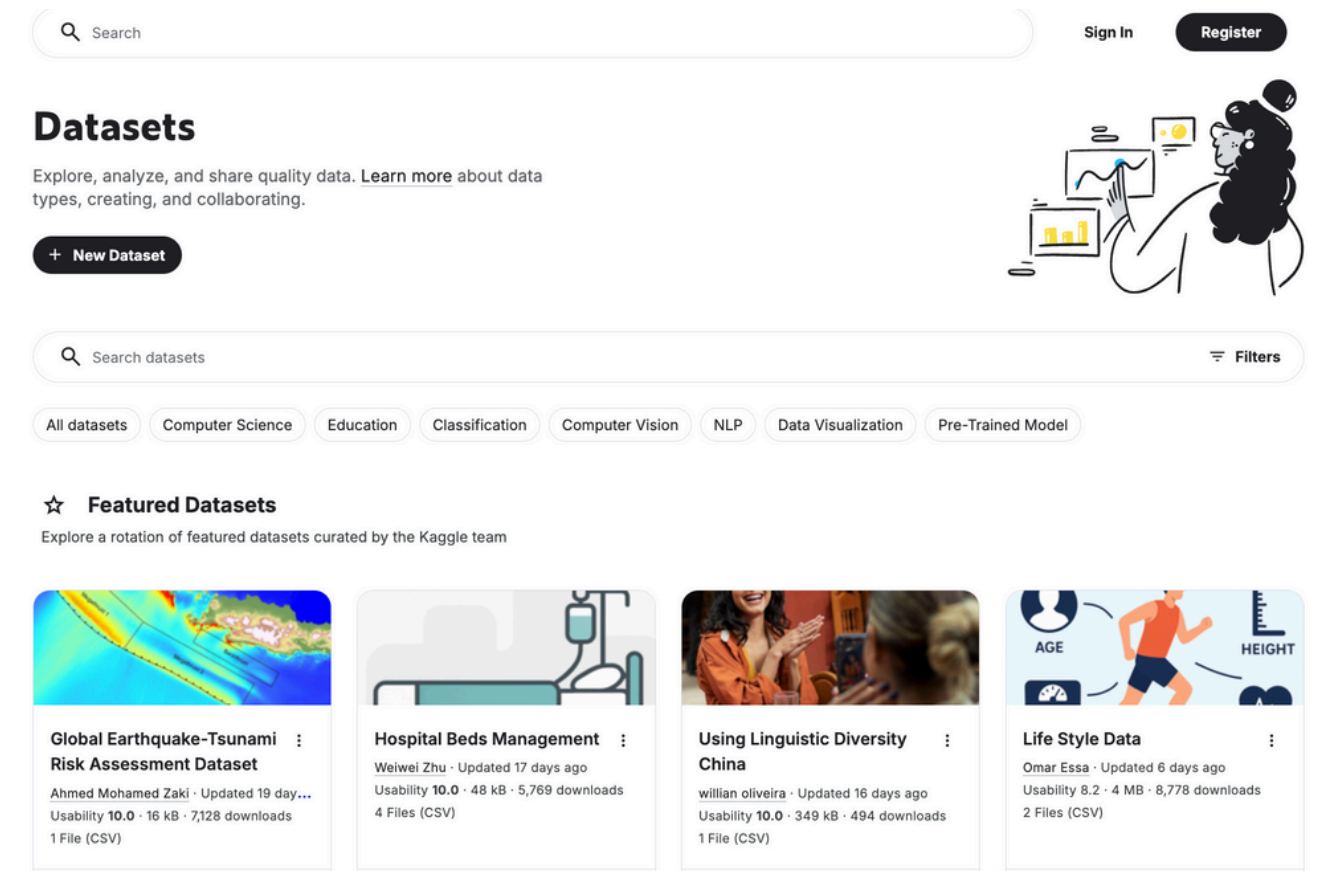
# DATA SOURCES



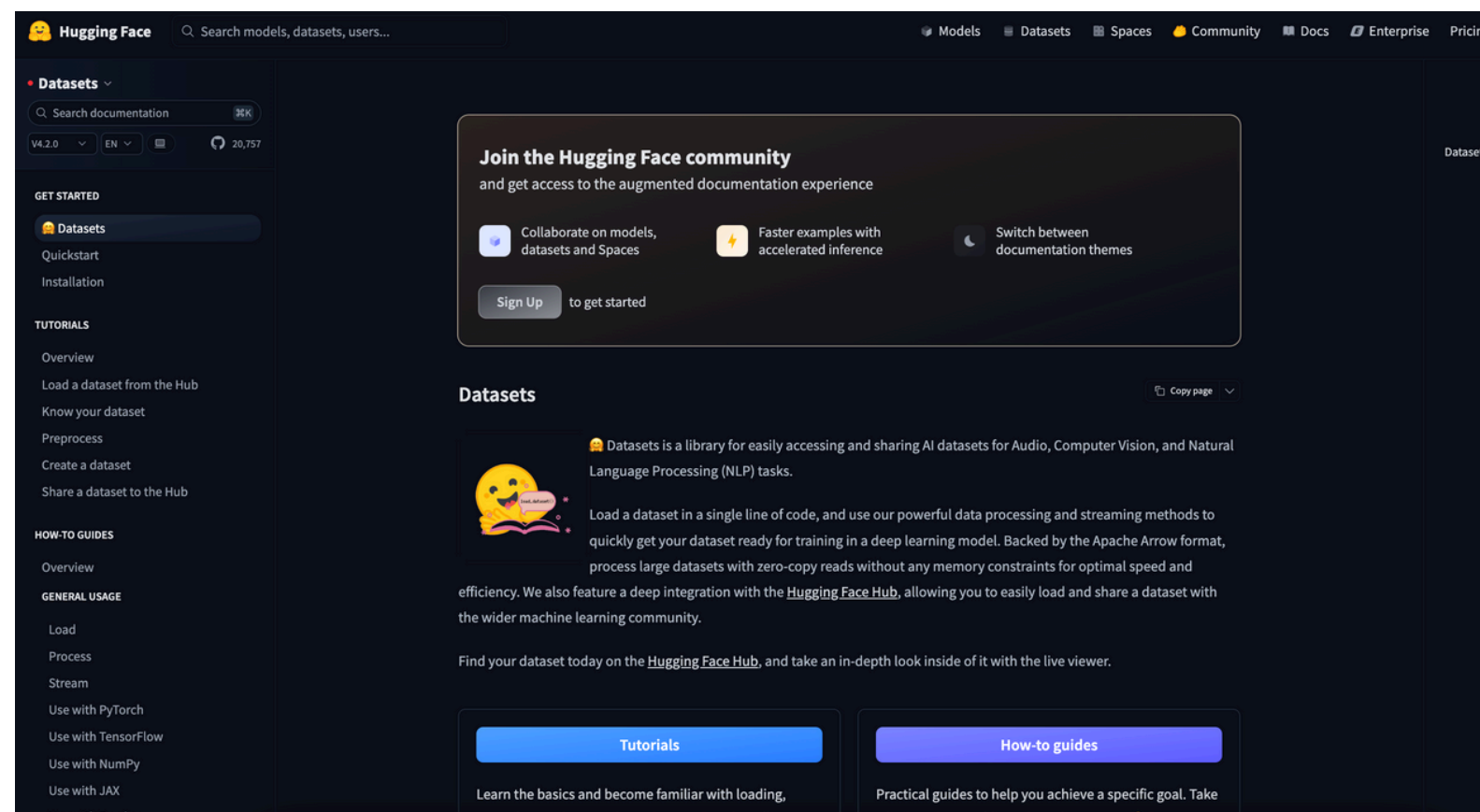
The screenshot shows the UC Irvine Machine Learning Repository website. It features a navigation bar with 'Datasets', 'Contribute Dataset', and 'About Us'. A search bar is present with the text 'Search datasets...'. On the left, there are filters for Keywords, Attributes, Data Type, Subject Area, Task, # Instances, # Features, Feature Type, and Python. The main section, 'Browse Datasets', lists several datasets: Iris, Heart Disease, Wine Quality, and Breast Cancer Wisconsin (Diagnostic). Each dataset entry includes a brief description, classification type, tabular status, number of instances, and number of features.



The screenshot shows the Kaggle website sidebar. It includes a 'Create' button and a list of navigation links: Home, Competitions, Datasets, Models, Benchmarks, Code, Discussions, Learn, and More.



The screenshot shows the Kaggle Datasets page. It features a search bar, a 'Sign In' button, and a 'Register' button. The main heading is 'Datasets', followed by a description: 'Explore, analyze, and share quality data. Learn more about data types, creating, and collaborating.' Below this is a '+ New Dataset' button. A section titled 'Featured Datasets' lists several datasets: Global Earthquake-Tsunami Risk Assessment Dataset, Hospital Beds Management, Using Linguistic Diversity China, and Life Style Data. Each dataset entry includes a thumbnail image, title, author, and download statistics.



The screenshot shows the Hugging Face Datasets page. It features a navigation bar with 'Models', 'Datasets', 'Spaces', 'Community', 'Docs', 'Enterprise', and 'Pricing'. The main heading is 'Datasets', followed by a description: 'Datasets is a library for easily accessing and sharing AI datasets for Audio, Computer Vision, and Natural Language Processing (NLP) tasks.' Below this is a section titled 'Join the Hugging Face community' with a 'Sign Up' button. The page also includes a 'Get started' section with links to 'Quickstart' and 'Installation', and a 'Tutorials' section with links to 'Overview', 'Load a dataset from the Hub', 'Know your dataset', 'Preprocess', 'Create a dataset', and 'Share a dataset to the Hub'.

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

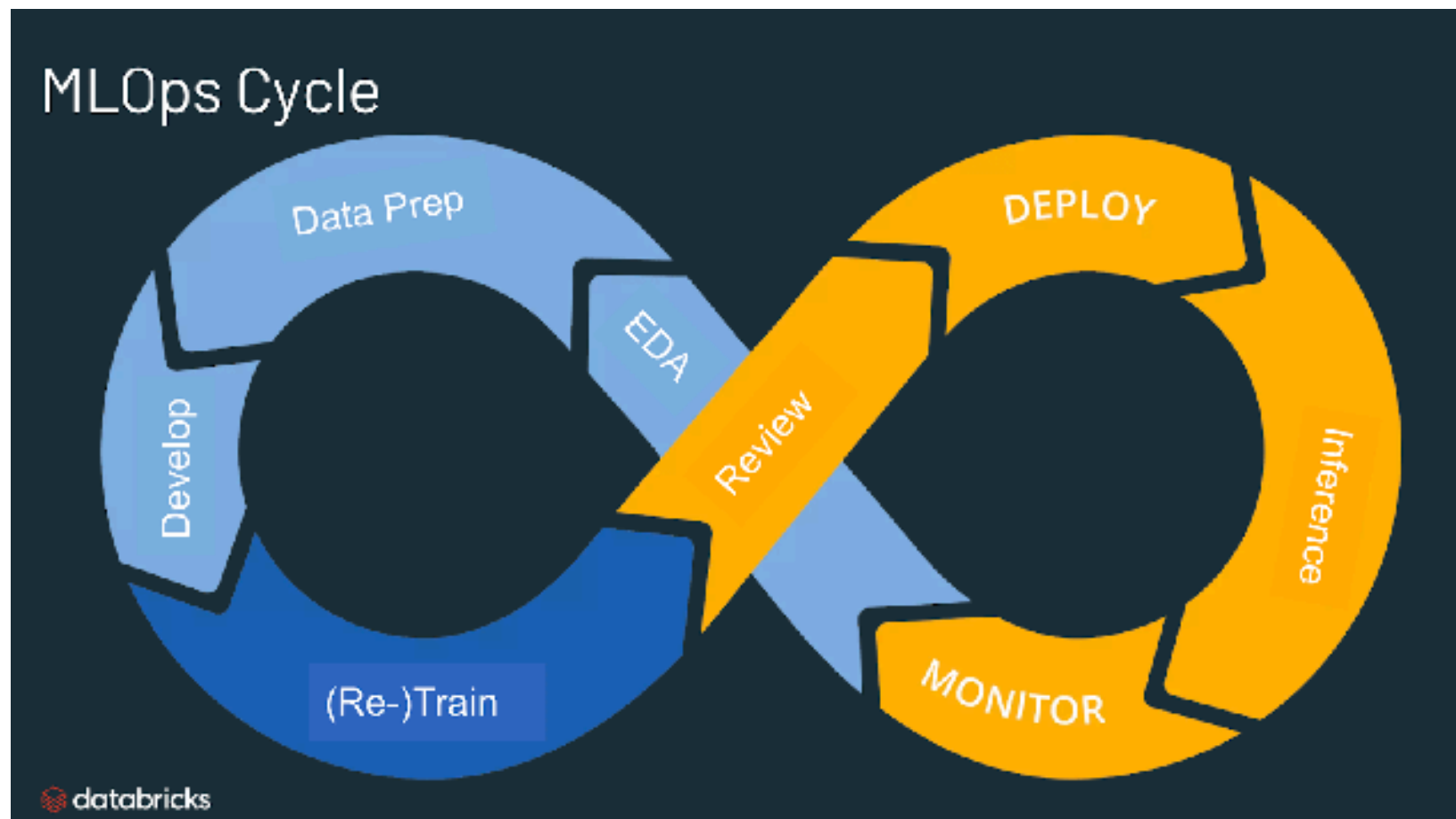


Image Source: <https://www.databricks.com/glossary/mlops>

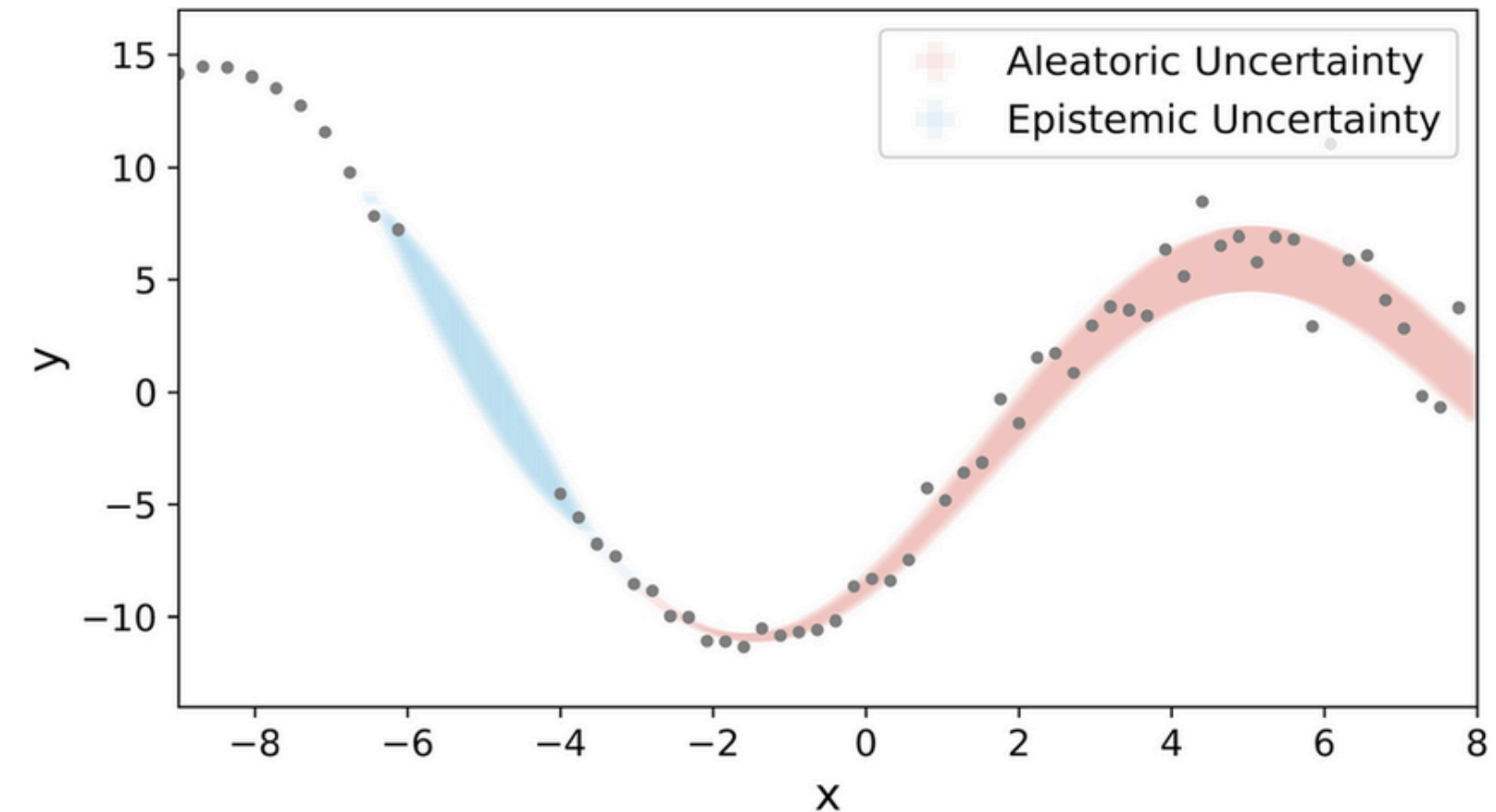


Image Source: [https://www.researchgate.net/figure/An-illustration-of-the-difference-between-aleatoric-and-epistemic-uncertainties-The-dots\\_fig1\\_368244388](https://www.researchgate.net/figure/An-illustration-of-the-difference-between-aleatoric-and-epistemic-uncertainties-The-dots_fig1_368244388)

## ML-OPS

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

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

## 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

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

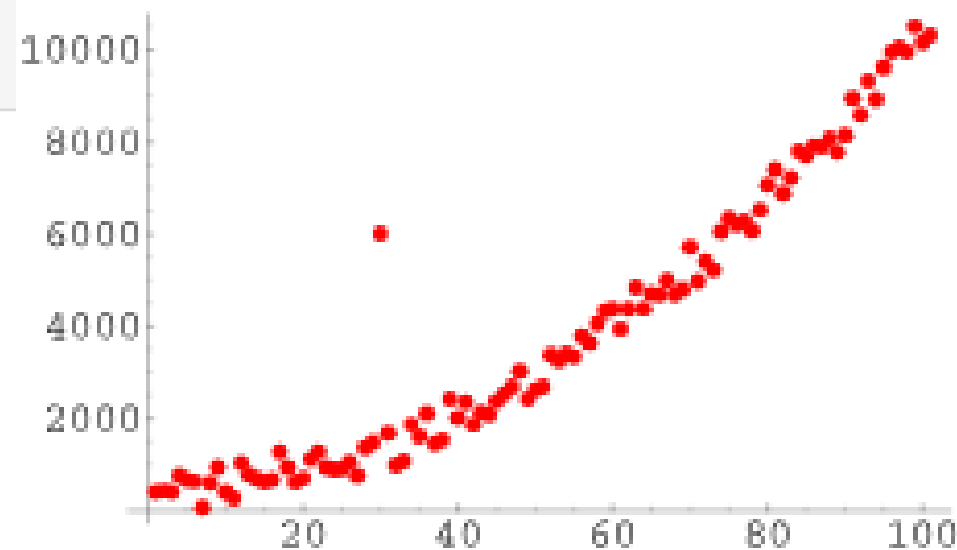


Image Source: <https://mathworld.wolfram.com/Outlier.html>

```
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']
```

# MISSING DATA, OUTLIERS, INCONSISTENCIES

## Problems with each on a data science perspective

Dependent Variable (Response Variable)

Independent Variables (Predictors)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

Y intercept

Slope Coefficient

Error Term

Image Source: [https://www.researchgate.net/figure/Linear-regression-equation\\_fig1\\_373123252](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

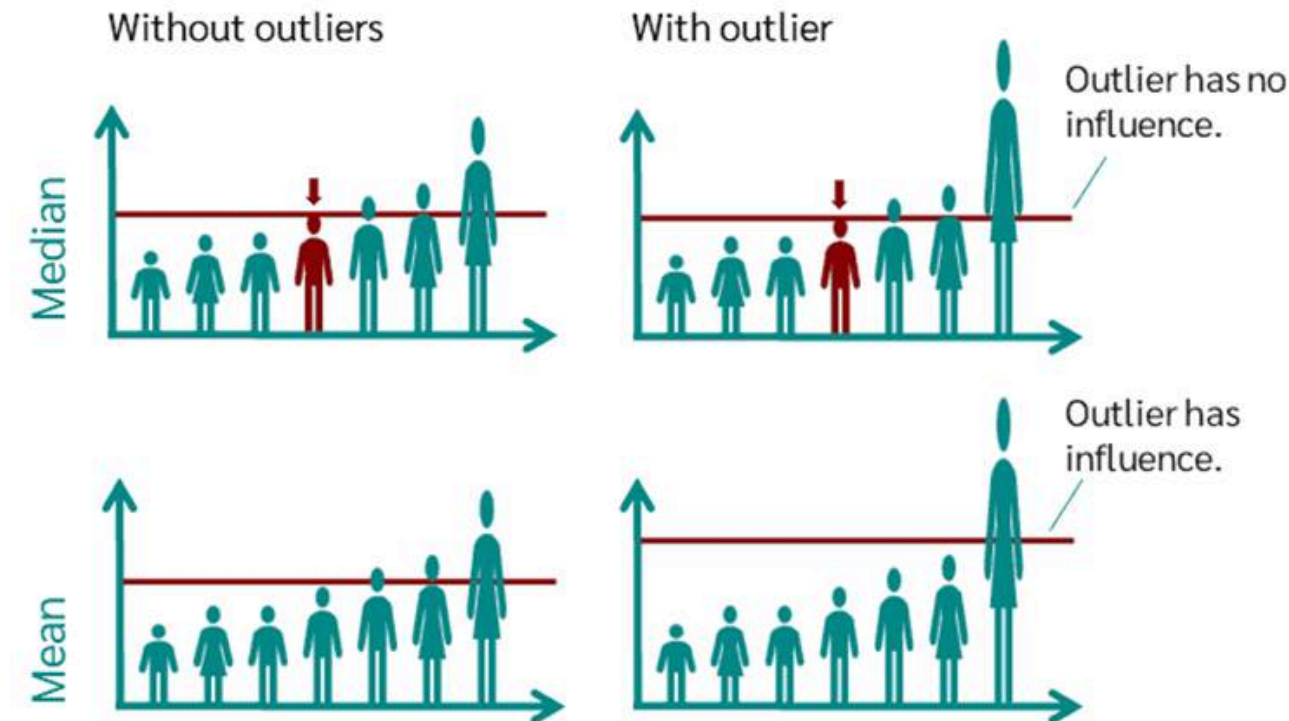


Image Source: <https://simbi.in/understanding-central-tendency-mean-median-mode-and-standard-deviation/>

**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.



## 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 imputation 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 ( $X \rightarrow \log(X)$ )
  - However it only works for right-skewed data

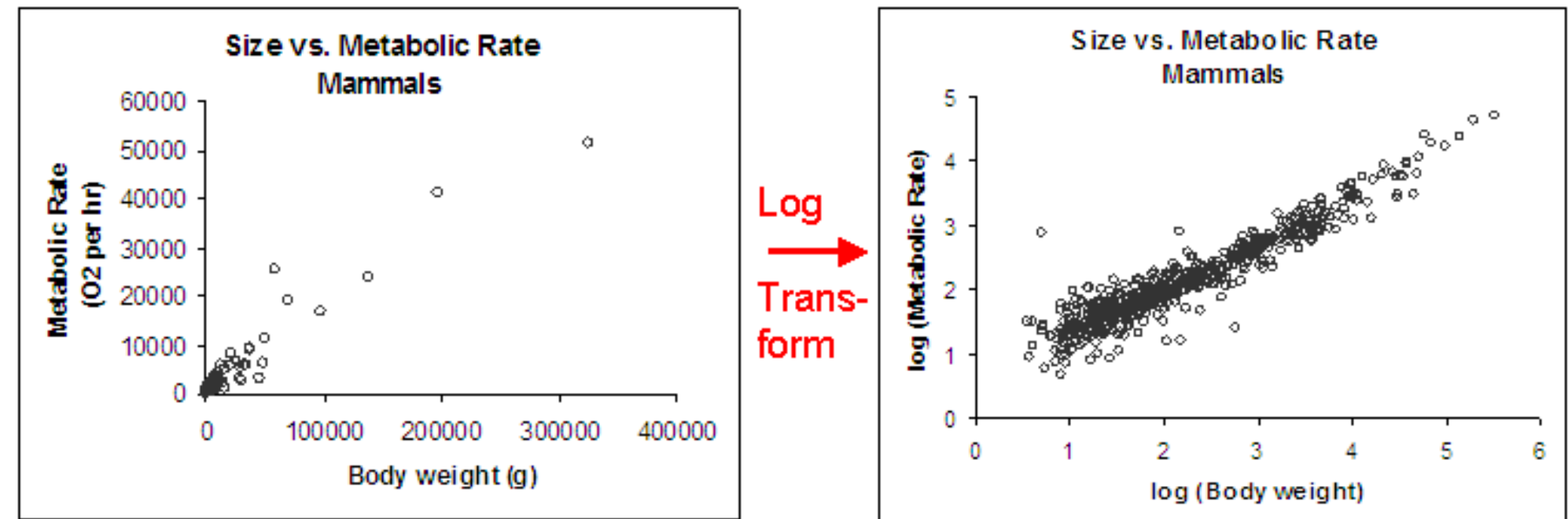


Image Source: [https://mathbench.umd.edu/modules/misc\\_scaling/page07.htm](https://mathbench.umd.edu/modules/misc_scaling/page07.htm)

## Ways to handle inconsistent data?

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

**Imputation: filling in missing values**

## 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

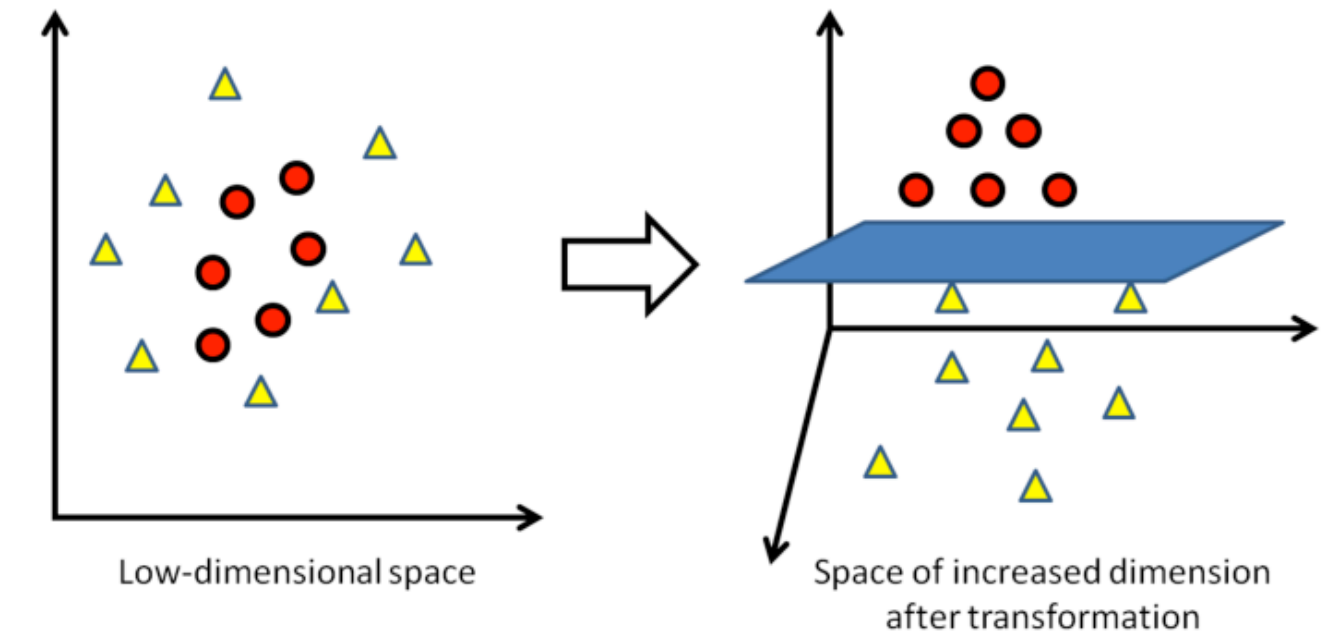


Image Source: [https://www.researchgate.net/figure/Kernel-trick-By-transforming-the-original-space-left-into-a-space-of-increased\\_fig1\\_305284381](https://www.researchgate.net/figure/Kernel-trick-By-transforming-the-original-space-left-into-a-space-of-increased_fig1_305284381)

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

We will be using Notebook

<https://bit.ly/AlgosocWK2Notebook>