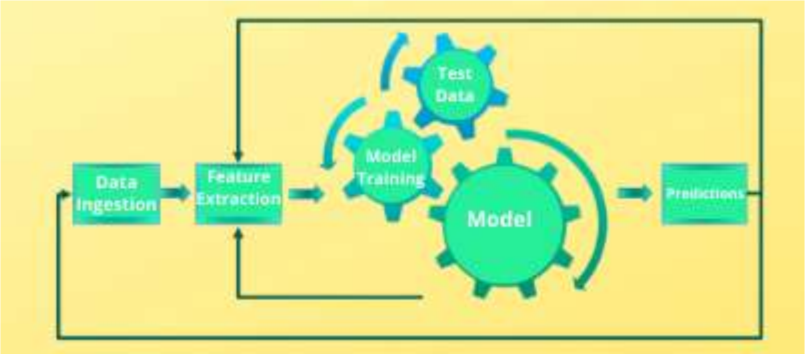
Arbaz Ali

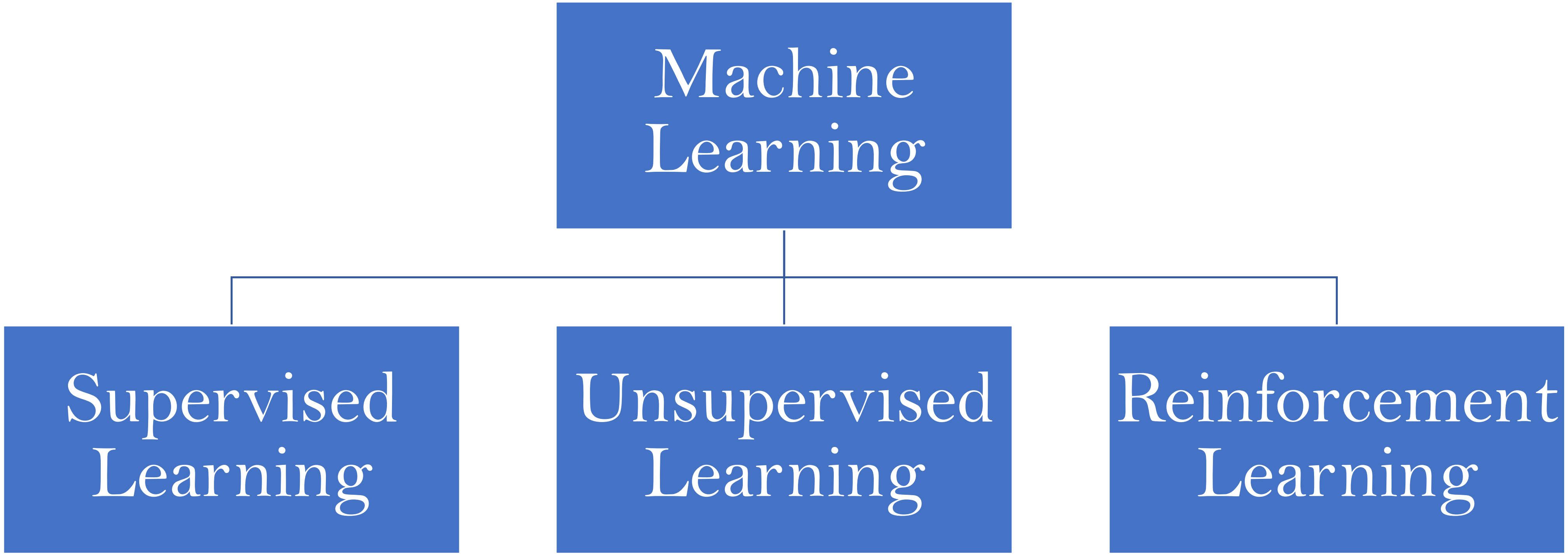
Machine Learning Pipeline



Created by Arbaz Ali

arbazali9299@gmail.com

Types of Learning



Machine Learning

Supervised Learning

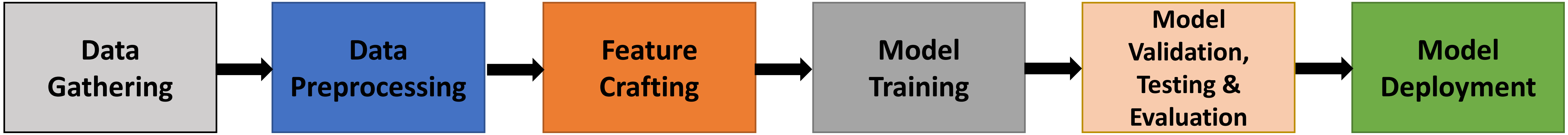
Unsupervised Learning

Reinforcement Learning

Machine Learning Pipeline

- A machine learning pipeline is:
    - A way to codify and automate the workflow it takes to produce a machine learning model.
- Machine learning pipelines consist of multiple sequential steps that perform operations from data extraction and preprocessing to model training and deployment.

Classic Machine Learning Pipeline



Data Preprocessing: Why?

- The majority of the real-world datasets for machine learning are highly susceptible to be missing, inconsistent, and noisy due to their heterogeneous origin.
- Applying machine learning algorithms on this noisy data would not give quality results as they would fail to identify patterns effectively. Data Processing is, therefore, important to improve the overall data quality.
- Duplicate or missing values may give an incorrect view of the overall statistics of data.
- Outliers and inconsistent data points often tend to disturb the model's overall learning, leading to false predictions.
- Quality decisions must be based on quality data. Data Preprocessing is important to get this quality data, without which it would just be a *Garbage In*, *Garbage Out* scenario.

Data Preprocessing

- Missing Data
  - Ignore the missing data
  - Remove the missing data
  - Manually fill by some value
  - Fill by some computed value
    - Fill by mean
    - Fill randomly from a distribution such as Gaussian or Standard
    - Fill using machine learning algorithms such as KMeans

Data Preprocessing

- Noise Removal
  - Noise reduction via outlier detection and removal
  - Binning & Discretization
  - Clustering
  - Manually Removing
- Undesired features' removal
  - Undesired features are also considered noise.
- Noise Addition

Data Preprocessing

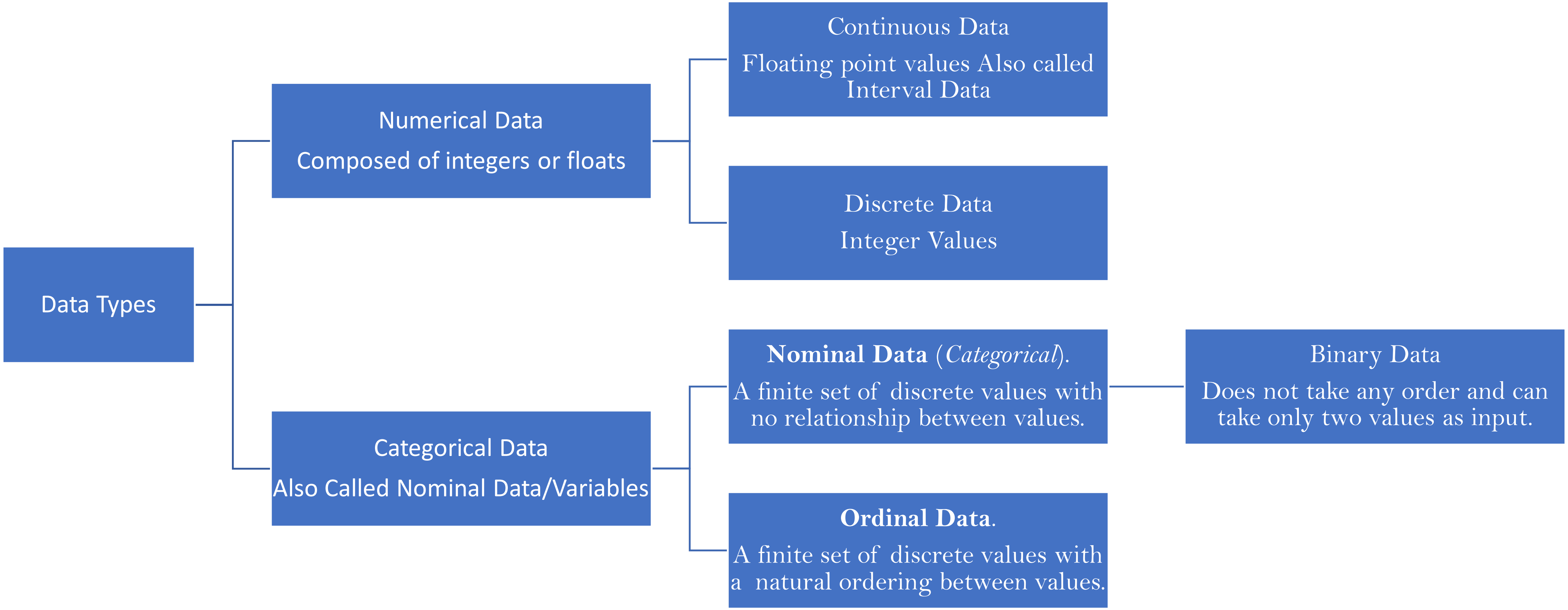
- Balancing of Imbalanced Datasets:
  - Resampling
  - Up sampling
  - Down sampling
  - Synthetic Samples addition
- Resizing (cropping, making docs of a certain length etc.)
- Data Encoding
  - (whole section coming up on next slide)

Data Encoding

- ML algorithms require all input and output variables to be numeric.
- Before feeding the data to ML models, they need to be converted into numerical variables.
- The data is divided into two types: Structured and Unstructured
- Structured Data is further divided into Numeric and Categorical.
- Some of the common structured data types that are used in Machine Learning and Data Science point of view are:
    - Continuous Data (Interval Data)
    - Discrete Data
    - Nominal Data
    - Ordinal Data
    - Binary Data

Data Encoding

Types of Data



Data Types

Numerical Data

Composed of integers or floats

Continuous Data

Floating point values Also called Interval Data

Discrete Data

Integer Values

Categorical Data

Also Called Nominal Data/Variables

Nominal Data (*Categorical*).

A finite set of discrete values with no relationship between values.

Binary Data

Does not take any order and can take only two values as input.

Ordinal Data.

A finite set of discrete values with a natural ordering between values.

Data Encoding

Types of Data

- Numerical data, as its name suggests, involves features that are only composed of numbers, such as integers or floating-point values.
- Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set. Categorical variables are often called nominal.
  - A ***"pet"*** variable with the values: *"dog"* and *"cat*".
  - A ***"color"*** variable with the values: *"red*", *"green*", and *"blue*".
  - A ***"place"*** variable with the values: *"first*", *"second*", and *"third*".
- Here, each value represents a different category. Some categories may have a natural relationship to each other, such as a natural ordering.
  - The ***"place"*** variable above does have a natural ordering of values. This type of categorical variable is called an **ordinal variable** because the values can be ordered or ranked.
  - The ***"color"*** variable has multiple values but there is no natural order or relationship between them. It is a **nominal variable**.
  - The ***"pet"*** variable holds only two types of values and the values do not have any relationship among one another. It is a **binary variable**.

Data Encoding

Types of Data

Age	Height	Gender	Pet	Class Position
19	5.0	M	Dog	First
27	5.7	M	Cat	Second
18	6.9	F	Rabbit	Third
33	5.4	F	Parrot	Fourth
Numeric Discrete	**Numeric Continuous**	**Categorical (Nominal) Binary**	**Categorical Nominal**	**Categorical Ordinal**

Data Encoding

Types of Data

- **Continuous Data:**
  - The data is classified into numeric data and can take any value in a specified interval. These types of data are also called interval, float or numeric data. Mean and standard deviation are the arithmetic operations that can be performed on continuous data. The statistical operations such as Pearson correlation coefficient and t-test and F-test can also be applied to this data to gain insights into it.
    - **Examples:** Height or Weight of an individual, Rate of Interest on loans, etc.

Data Encoding

Types of Data

- **Discrete Data:**
  - The data which can take on only integer values are said to be discrete data. This type of data is usually used in counting the number of occurrences of the event. The discrete data cannot take on floating or decimal values.
    - **Examples:** Student count in a class, Colour count in a Rainbow.

Data Encoding

Types of Data

- **Nominal Data:**
  - Nominal data can be categorized into categorical that has no explicit ordering associated with it. The nominal data are plainly used as labelled data. No statistical operations such as calculation of mean, median or standard deviation can be performed on the nominal data as performing such statistical operations on such data doesn’t imply anything insightful.
    - **Examples:** States in a country, zip codes of areas.

Data Encoding

Types of Data

- **Ordinal Data:**
  - The data which has an explicit ordering associated with it is known as Ordinal data. The ordinal data is also a type of categorical data and has a specific definitive order with it. Calculations like frequency distribution, percentage of total calculation and other non-parametric statistics with ordinal data. However, mean calculation, standard deviation calculation and other parametric tests of statistics makes no sense for this kind of data.
    - **Examples:** Ratings for a restaurant (e.g. very good, good, bad, very bad), Level of Education of an individual (e.g. Doctorate, Post Graduate, Undergraduate), etc.

Data Encoding

Types of Data

- **Binary Data:**
  - This is a special case of Nominal data which does not take any order and can take only two values as input. The kind of operations that can be performed on this data are the same as that are performed on the nominal data.
    - **Examples:** Gender (male or female), Fraudulent transaction (Yes or No), Cancerous Cell (True or False).

Data Encoding

- Converting **Numerical Variable** to **Ordinal Variable**
  - A numerical variable can be converted to an ordinal variable by dividing the range of the numerical variable into **bins** and assigning values to each bin. For example, a numerical variable between 1 and 10 can be divided into an ordinal variable with 5 labels with an ordinal relationship: 1-2, 3-4, 5-6, 7-8, 9-10. **This is called discretization**.

Data Encoding

Encoding Schemes

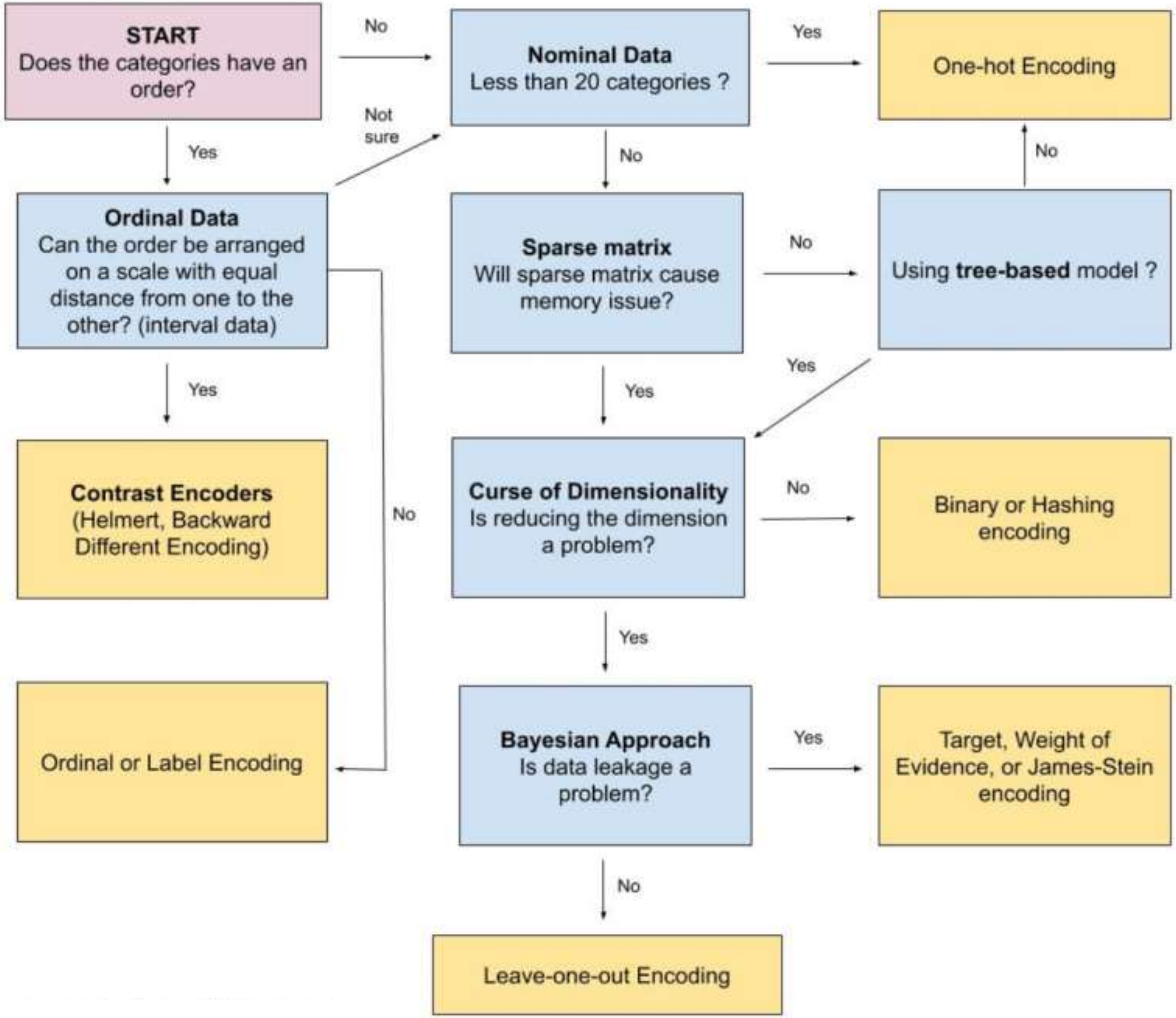
- Different Data Encoding Schemes:
  - Integer Encoding / Label encoding
  - One-hot Encoding
  - Dummy Variable Encoding

****

  - Binary Encoding
  - Frequency Encoding
  - Hash Encoding
  - Helmert Encoding
  - Using Bayesian Encoders

Data Encoding

Encoding Schemes



START
Does the categories have an order?

Nominal Data
Less than 20 categories ?

One-hot Encoding

Ordinal Data
Can the order be arranged on a scale with equal distance from one to the other? (interval data)

Sparse matrix
Will sparse matrix cause memory issue?

Using **tree-based** model ?

Contrast Encoders
(Helmert, Backward Different Encoding)

Curse of Dimensionality
Is reducing the dimension a problem?

Binary or Hashing encoding

Ordinal or Label Encoding

Bayesian Approach
Is data leakage a problem?

Target, Weight of Evidence, or James-Stein encoding

Leave-one-out Encoding

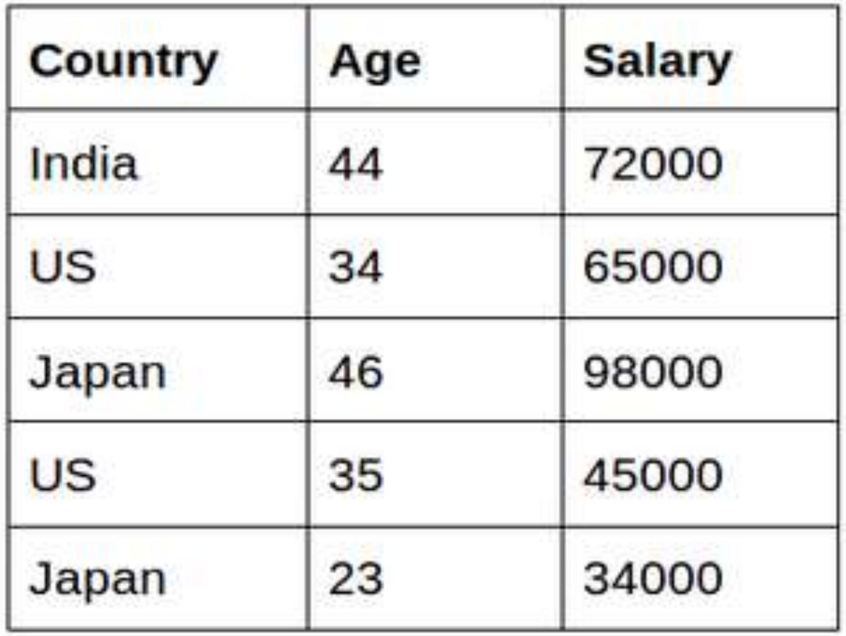
Data Encoding

- Some algorithms can work with categorical data directly. For example, a decision tree can be learned directly from categorical data with no data transform required (this depends on the specific implementation).
- Many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric.
- In general, this is mostly a constraint of the efficient implementation of machine learning algorithms rather than hard limitations on the algorithms themselves.
- Some implementations of machine learning algorithms require all data to be numerical. For example, scikit-learn has this requirement.
- This means that categorical data must be converted to a numerical form. If the categorical variable is an output variable, you may also want to convert predictions by the model back into a categorical form in order to present them or use them in some application.

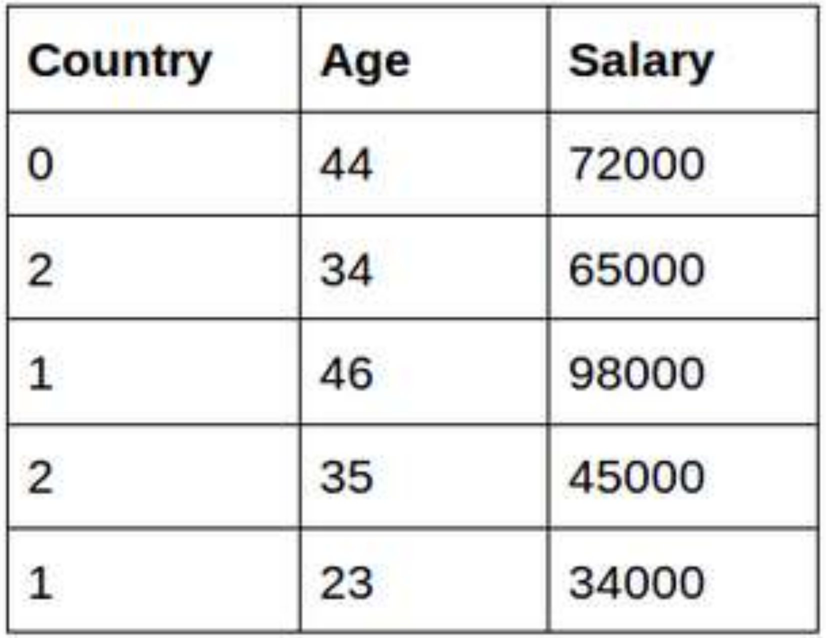
Data Encoding

Encoding Schemes

- Integer Encoding/Label Encoding



Country	Age	Salary
India	44	72000
US	34	65000
Japan	46	98000
US	35	45000
Japan	23	34000

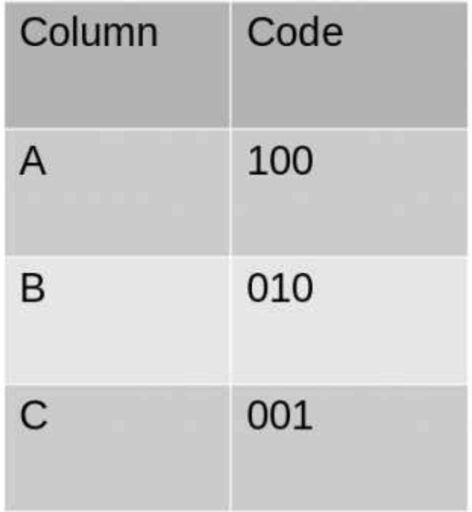


Country	Age	Salary
0	44	72000
2	34	65000
1	46	98000
2	35	45000
1	23	34000

Data Encoding

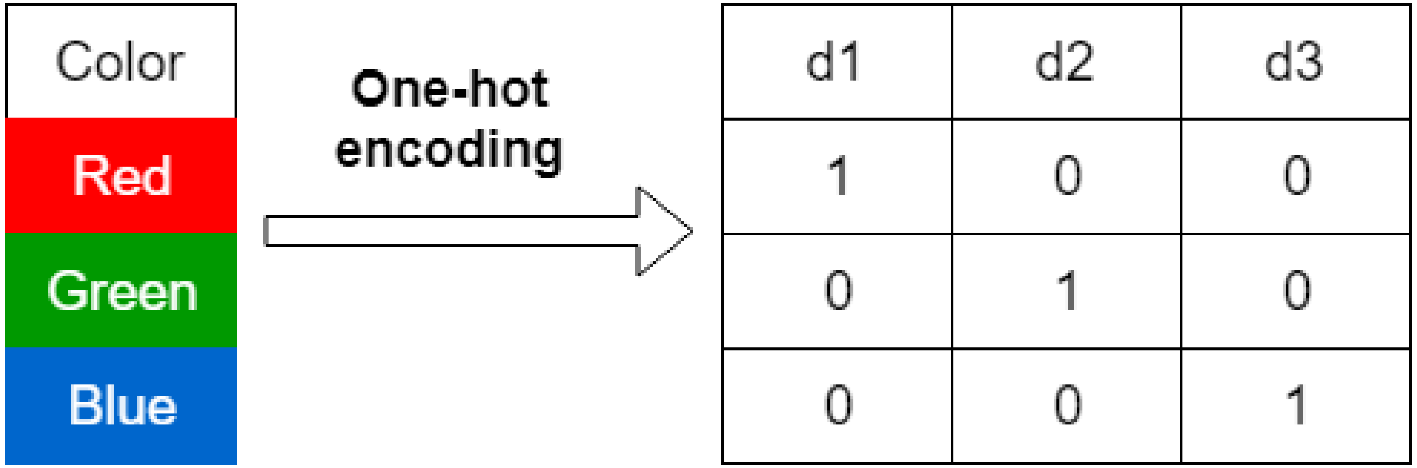
Encoding Schemes

- One Hot Encoding
  - In one-hot encoding, we create a new set of dummy (binary) variables that is equal to the number of categories (k) in the variable.



Column	Code
A	100
B	010
C	001

One- Hot Coding



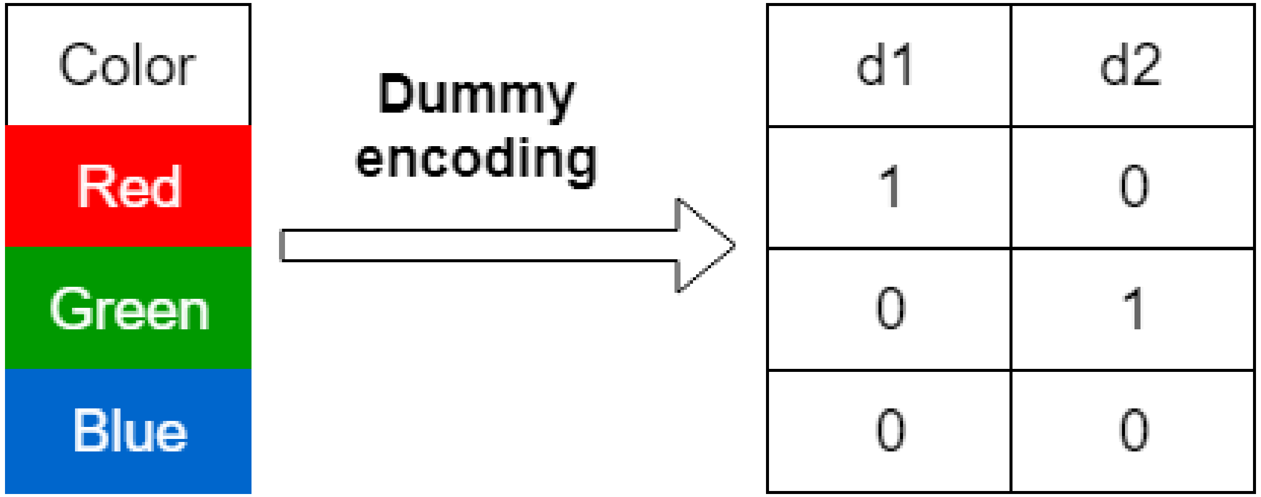
Color
Red
Green
Blue

d1	d2	d3
1	0	0
0	1	0
0	0	1

Data Encoding

Encoding Schemes

- Dummy Encoding
  - Dummy encoding also uses dummy (binary) variables. Instead of creating a number of dummy variables that is equal to the number of categories (k) in the variable, dummy encoding uses k-1 dummy variables.



Color
Red
Green
Blue

d1	d2
1	0
0	1
0	0

Feature Crafting

- Feature Selection/Extraction
  - Finding most influential features in the dataset which heavily influence the prediction/outcome.
  - All of the features we find in the dataset might not be useful in building a machine learning model to make the necessary prediction. Using some of the features might even make the predictions worse. So, feature selection plays a huge role in building a machine learning model.
    - Algos: IG, MI, Chi-Square
  - Some measures to select features on their basis: correlation, p-value
  - Extracting features from the data on the basis of certain criteria such as how many times a feature is occurring.
    - Algos: TF, TF-IDF, PCA
- Handcrafted Features Vs Learned Features
  - Hand crafted: features that are manually engineered by the data scientist.
  - Learned features: are ones that are automatically obtained from a deep learning algorithm

Model Training, Testing & Evaluation

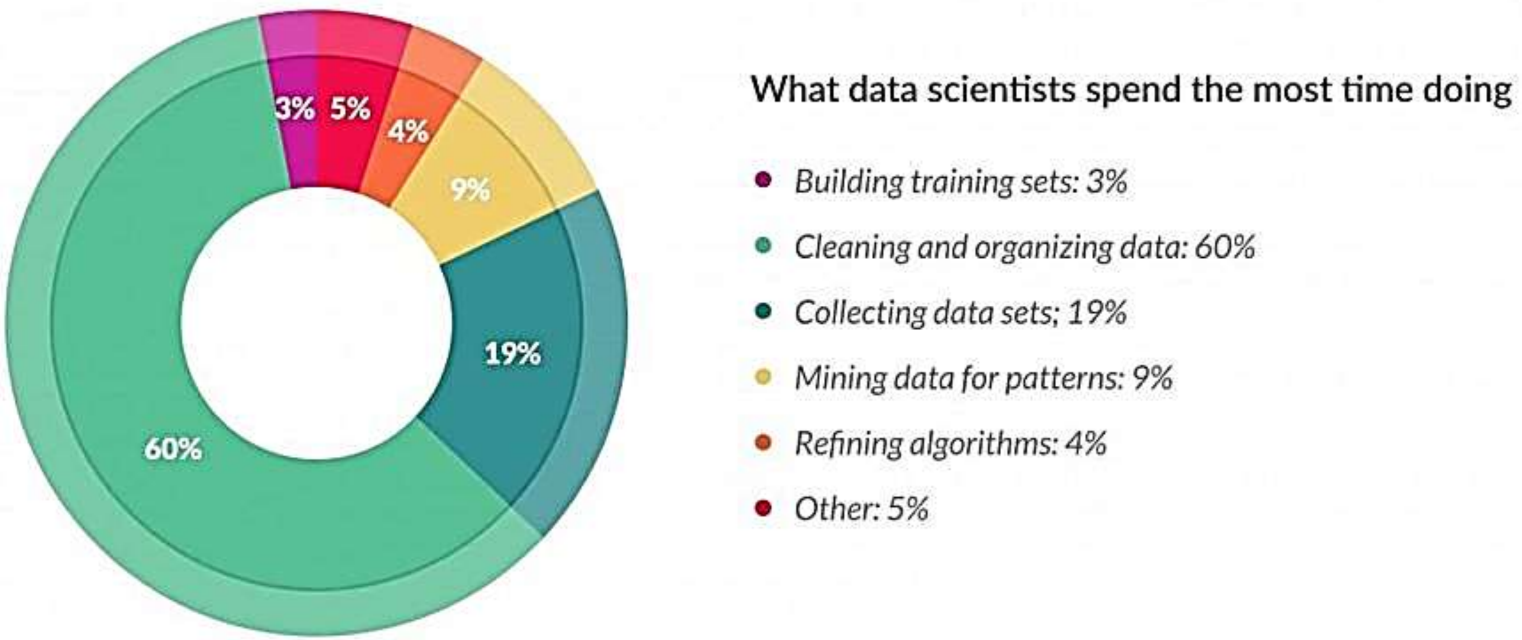
- Model Training: Learning through data in iterative process.
- It is carried out in iterative cycles called epochs.
- Common approaches to train a neural network are:
  - Full-Batch (Weights are updated after one epoch)
  - Mini-Batch (Weights get updated several times over an epoch)
  - Online Learning (Weights are updated after each example)
- Testing:
  - Train, Validation & Test Split
  - K-Fold Testing
  - Leave-on-out Testing
- Evaluation:
  - Confusion Matrix : TP, FP, TN, FN
  - Precision, Recall, F1-Measure, Accuracy, Kappa Score, etc.
- Further Reading Encoding Schemes: Link

Model Deployment

- Saving the ML model:
  - Pickle converts a python object to a bitstream and allows it to be stored to disk and reloaded at a later time. It provides a good format to store machine learning models provided that their intended applications are also built in python.
  - ONNX the Open Neural Network Exchange format, is an open format that supports the storing and porting of predictive model across libraries and languages. Most deep learning libraries support it and Scikit-Learn also has a library extension to convert their model to ONNX's format.
- Machine Learning Model as a Micro-Service
  - Kubernetes - Kubeflow
  - Docker

ML Pipeline – A Temporal View

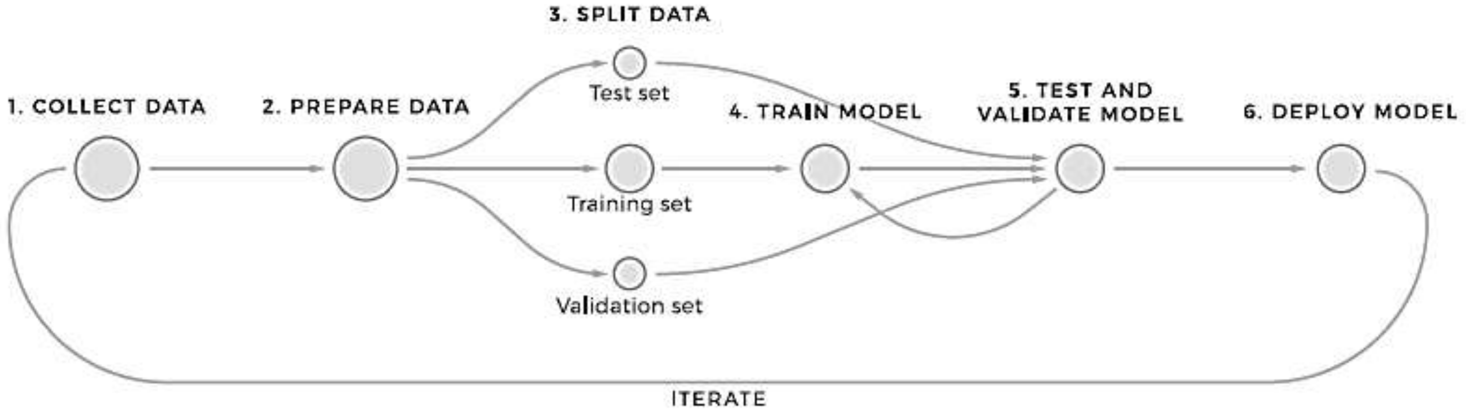
- Data preparation accounts for about 80% of the work of data scientists



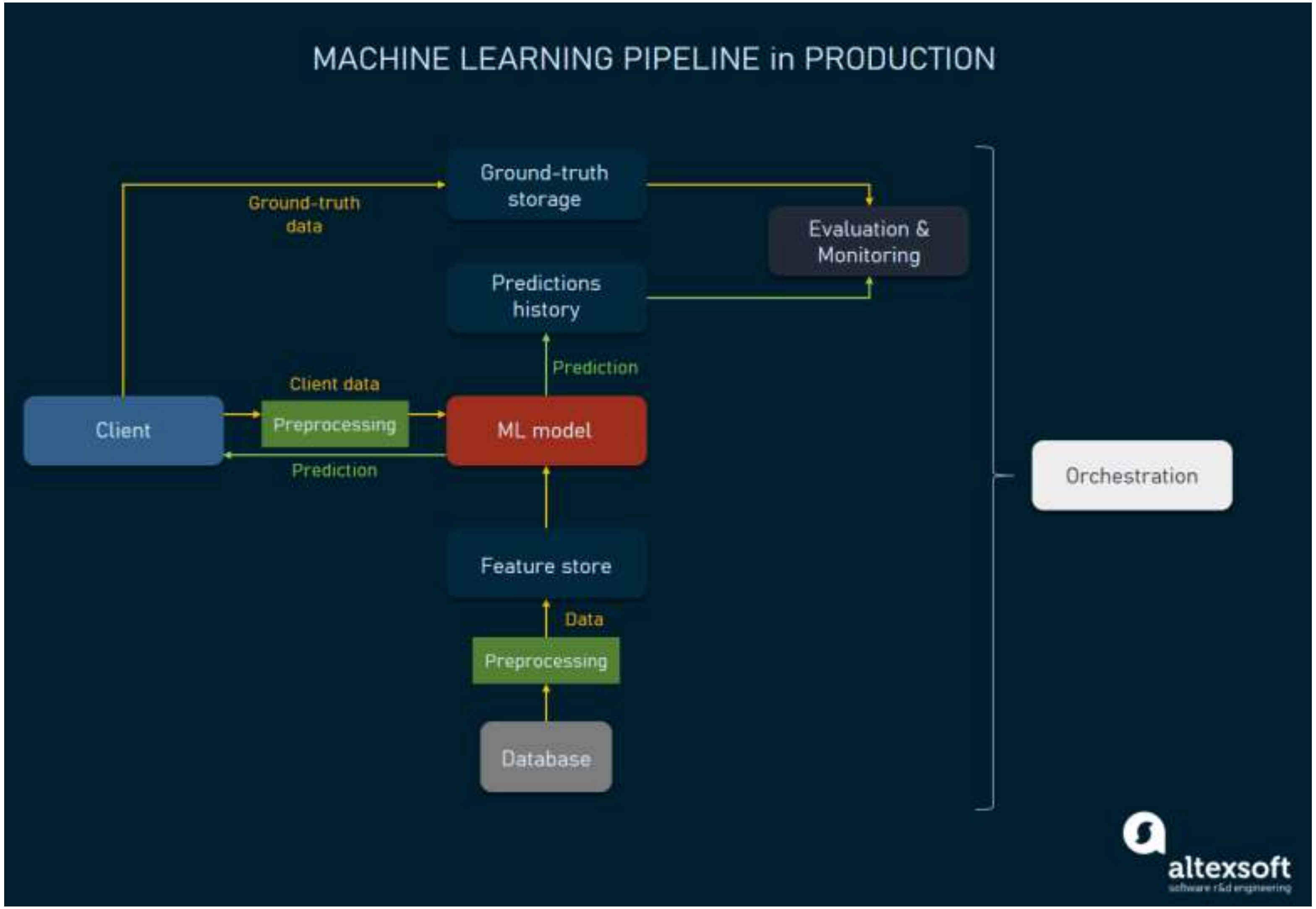
What data scientists spend the most time doing

ML Pipeline – A Software Engineering View

- A machine learning pipeline (or system) is a technical infrastructure used to manage and automate ML processes in the organization.
- The pipeline logic and the number of tools it consists of vary/change depending on the ML needs. But, in any case, the pipeline would provide data engineers with means of managing data for training, organizing models, and managing them on production.

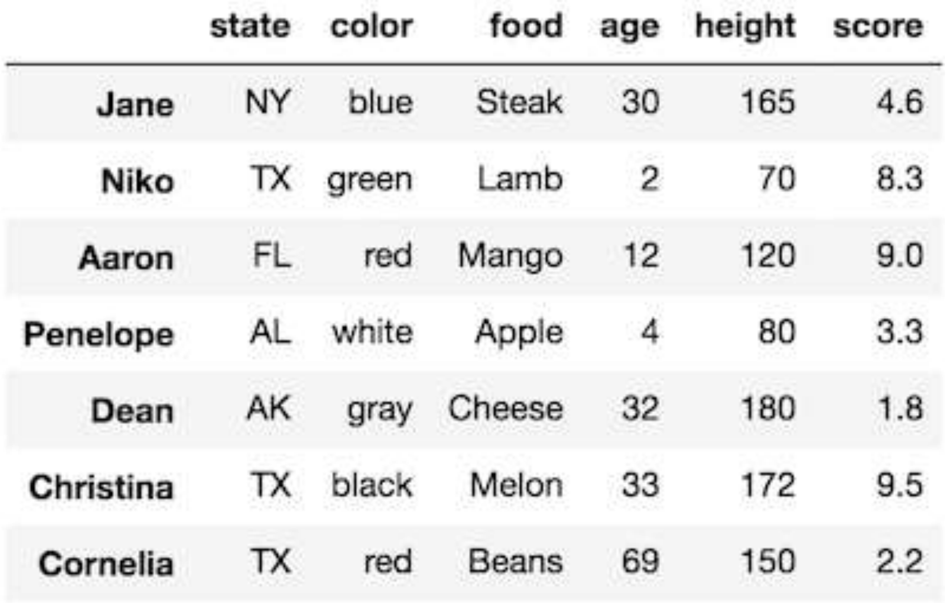


ML Pipeline – A Software Engineering View



Terminologies

- Example
- Ground Truth / Label
- Model (used in two senses)
  - ML Algorithm
  - The trained Model (Structural design and weights and biases)
- Annotating/Labeling the data & annotated / labeled data



	state	color	food	age	height	score
Jane	NY	blue	Steak	30	165	4.6
Niko	TX	green	Lamb	2	70	8.3
Aaron	FL	red	Mango	12	120	9.0
Penelope	AL	white	Apple	4	80	3.3
Dean	AK	gray	Cheese	32	180	1.8
Christina	TX	black	Melon	33	172	9.5
Cornelia	TX	red	Beans	69	150	2.2

Data Pipeline VS ETL Pipeline

- Data Pipeline Is an Umbrella Term of Which ETL Pipelines Are a Subset
- ETL Pipelines Always Involve Transformation
- In the extraction part of the ETL Pipeline, the data is sourced and extracted from different systems like CSVs, web services, social media platforms, CRMs, and other business systems. In the transformation part of the process, the data is then molded into a format that makes reporting easy. Sometimes data cleansing is also a part of this step. In the loading process, the transformed data is loaded into a centralized hub to make it easily accessible for all stakeholders.
- The purpose of the ETL Pipeline is to find the right data, make it ready for reporting, and store it in a place that allows for easy access and analysis. An ETL tool will enable developers to put their focus on logic/rules, instead of having to develop the means for technical implementation. This frees up a lot of time and allows your development team to focus on work that takes the business forward, rather than developing the tools for analysis.