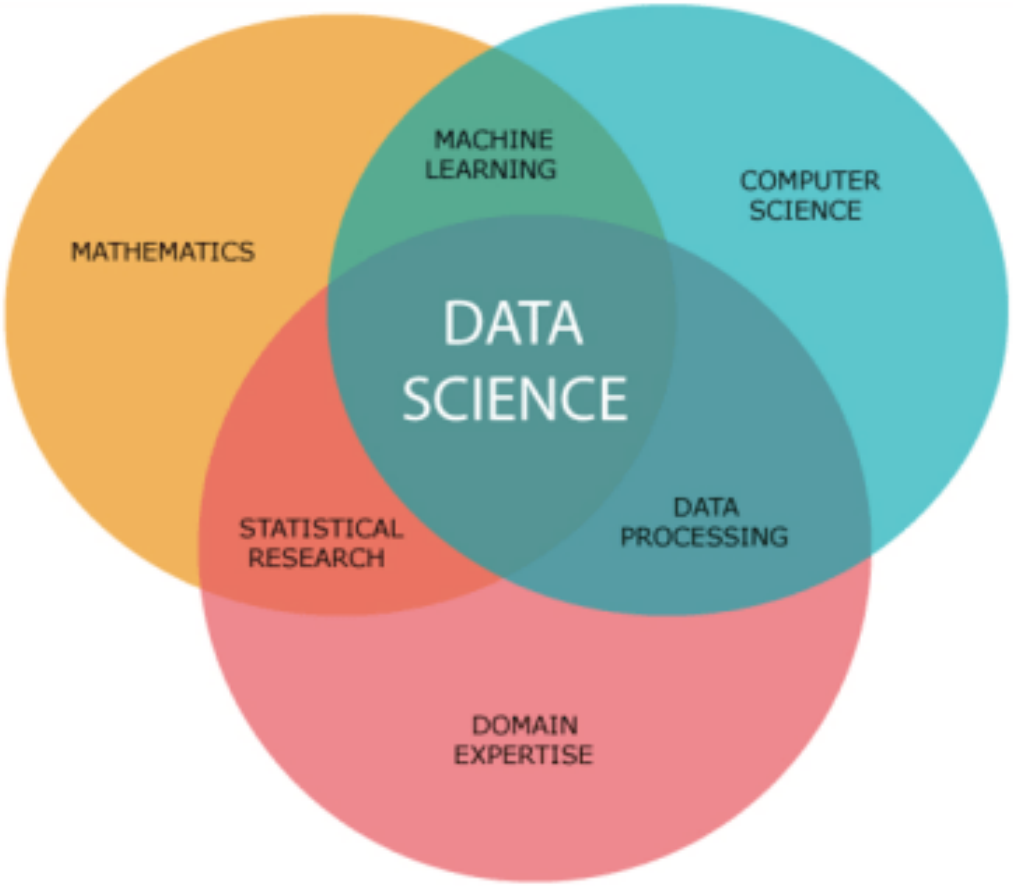
4243/5243: Applied Data Science

Lecture 01: End – to – End Data Science Project

What is Data Science?

- **Data Science** is a field of study that combines domain expertise, programming skills, and knowledge of mathematics & statistics to extract meaningful insights from data.



What is Data Science?

- A data science lifecycle involves various roles, tools, and processes, which enables analysts to glean actionable insights.

- Typically, a data science project undergoes the following stages: **data ingestion**, **data storage and data processing**, **data analysis**, **communication**.

Data Ingestion

- The lifecycle begins with the data collection – both raw **structured** and **unstructured** data from all relevant sources using a variety of methods.

- These methods can include **manual entry**, **web scraping**, and **real-time streaming** data from systems and devices.

Data Storage and Data Processing

- This stage covers taking the raw data and putting it in a form that can be used.

- This includes **data warehousing**, **data cleansing**, **data processing**, **data architecture**.

Data Analysis

- Here, data scientists conduct analysis to examine biases, patterns, ranges, and distributions of values within the data.
- It involves **exploratory/confirmatory analysis**, **predictive analysis**, **data mining**, **machine learning**, **text mining** etc.

Communication

- Finally, insights are presented as reports and other data visualizations that makes the insights and their impact easier to understand.

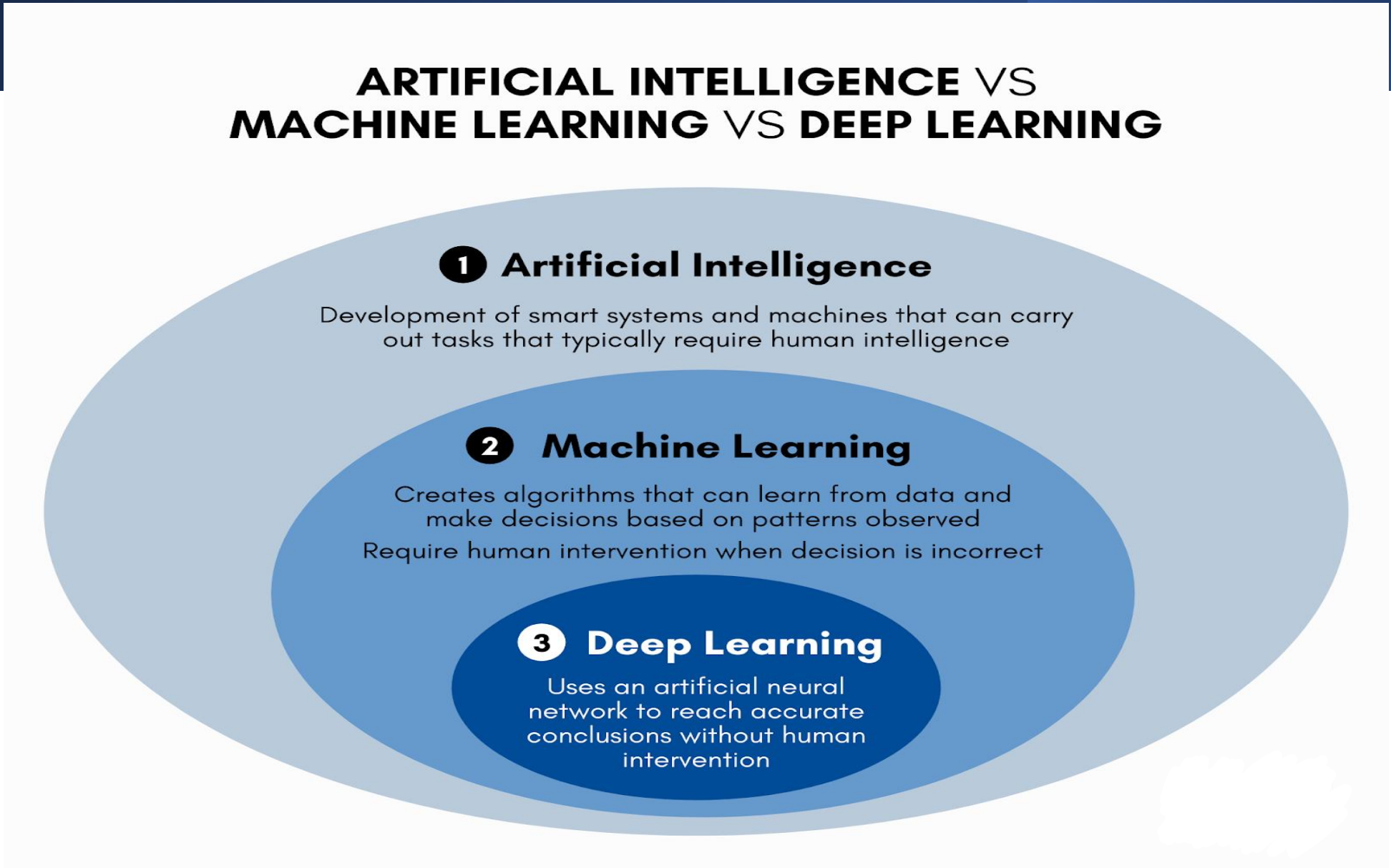
What is Machine Learning?

- **Machine Learning** (**ML**) is a core-area and, undoubtedly, one of the most exciting subsets of **Artificial Intelligence** (**AI**) which focuses on the use of *data* and *algorithms* to imitate the way that humans learn, gradually improving its accuracy.

- ML involves finding insightful information without being told where to look. Instead, it does this by leveraging algorithms that learn from data in an iterative process.

- When exposed to new data, these algorithms learn, grow, change, and develop by themselves.

What is Machine Learning?

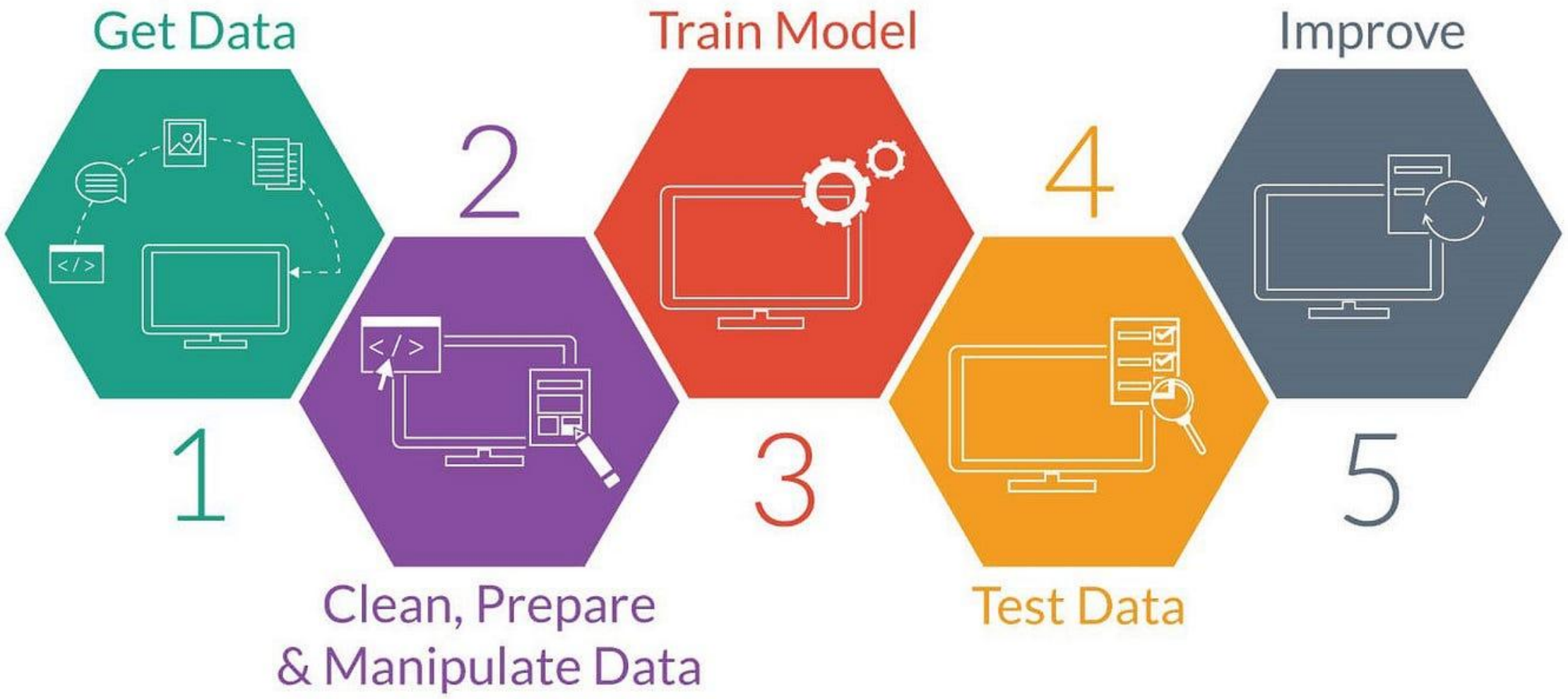


What is Machine Learning?

Machine Learning algorithms are widely applicable across many industries. For instance,

- **Recommendation engines** are used by e-commerce, social media, and news organizations to suggest content based on a customer's past behavior.
- Machine Learning and **Machine Vision** are a critical component of self-driving cars, helping them navigate the roads safely.
- In **Healthcare**, ML is used to diagnose and suggest treatment plans.

End – to – End ML Project



Problem Definition & Project Goals

Problem Definition & Project Goals

- **Define the business or research problem**: a clear understanding of the problem is crucial to guide the project. This involves identifying the specific questions or challenges that need to be addressed and framing it in a way that data science techniques can solve.

- **Specify objectives and success criteria**: establish measurable goals to determine project success. For instance, success might be defined as achieving at least 85% accuracy in prediction.

- **Examples**: forecasting sales for an e-commerce platform, detecting fraudulent transactions in financial data, or optimizing delivery routes for a logistics company.

Get Data

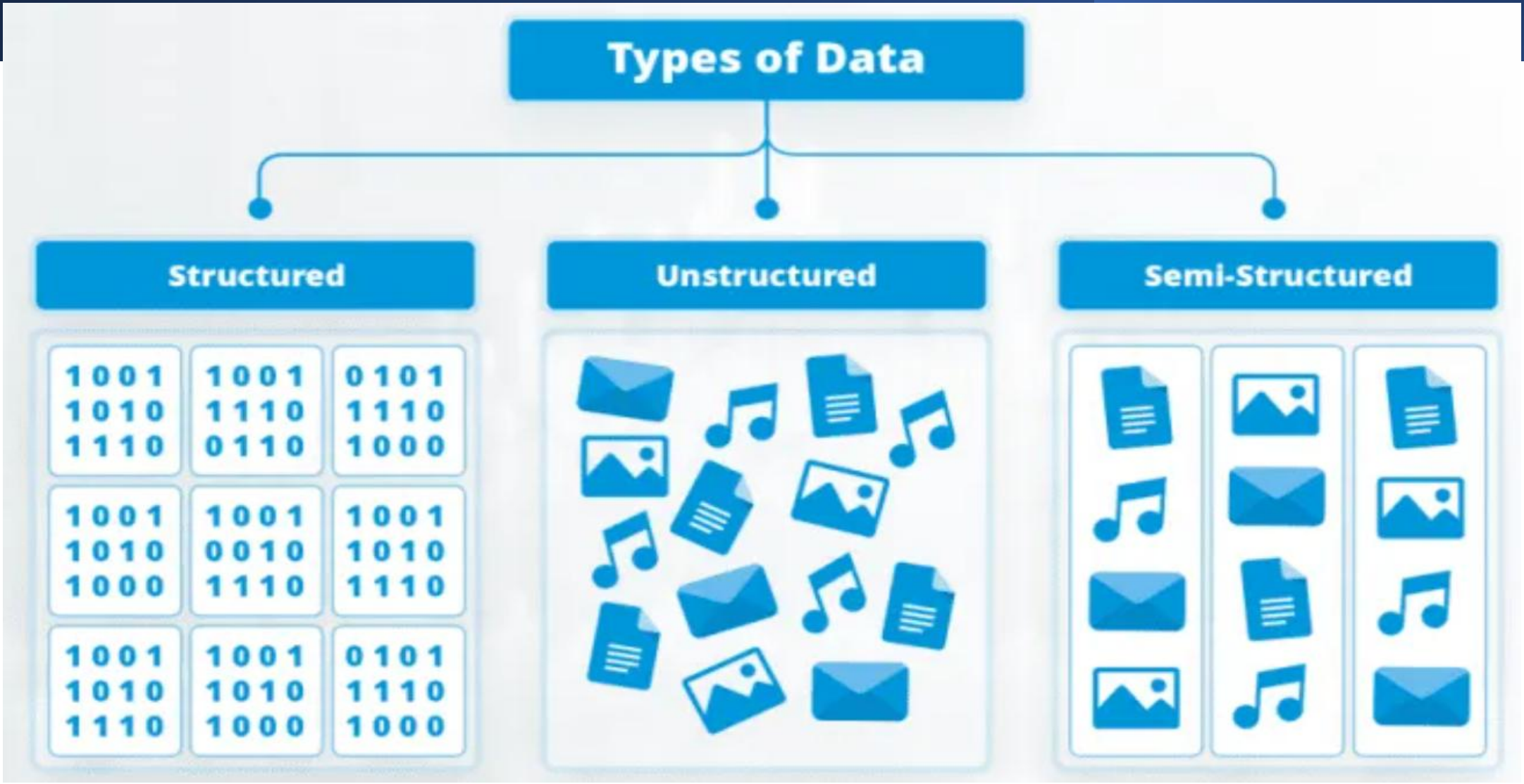
Data

- **Data** is a crucial component in the field of Machine Learning. It refers to the set of observations or measurements that can be used to train an ML model.

- The quality and quantity of data available for training and testing play an important role in determining the performance of a machine-learning model.

- Data can be in various forms such as numerical, categorical, or time-series, and can come from various sources such as databases, spreadsheets, or APIs.

Data Types


Types of Data
Structured
Unstructured
Semi-Structured
1001 1010 1110
1001 1110 0110
0101 1110 1000
1001 1010 1000
1001 0010 1110
1001 1010 1110
1001 1010 1110
1001 1010 1000
0101 1110 1000

Structured Data

- **Structured Data** – highly organized and formatted data stored in databases. Examples – relational databases, excel spreadsheets, tabular data, and more.

Unstructured Data

- **<u>Unstructured Data</u>** – Data that lacks a predefined structures and is not easily organized. Examples – text documents, images, videos.

Data Collection

- **Identifying data sources**: determining where and how to obtain the required data is a critical first step. Data sources might include structured databases, APIs provided by service, or even user-generated surveys.

- **Methods of data extraction**: data extraction techniques vary depending on the source. Structured databases require SQL queries, APIs often involve HTTP requests, and web scraping might be necessary for unstructured web data.

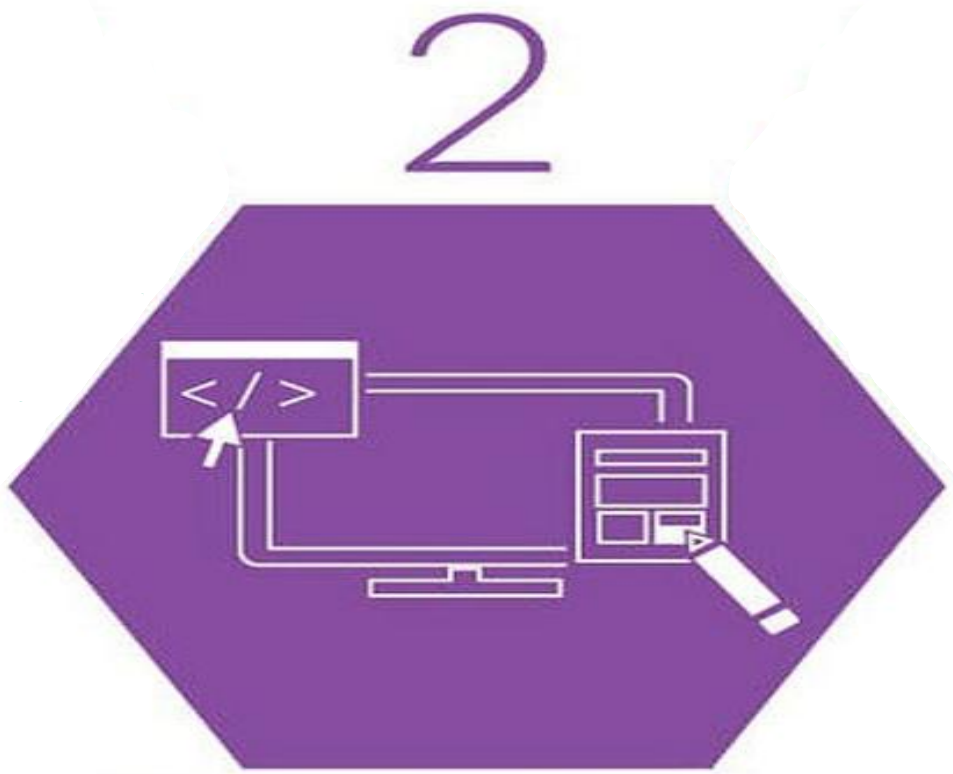
- For example, Python libraries such as ***requests*** and ***BeatifulSoup*** are commonly used for web scraping tasks.

Data Collection

Below are some data repositories that can be used for data acquisition:

- Kaggle Repository
- UCI ML Repository
- Government Data Repository
- GEO (Gene Expression) Data Repository

Data Preparation



Data Preparation

- **Stages Included**: data cleaning, exploratory data analysis (EDA), and feature engineering are essential steps for preparing raw data into a usable format for modeling.

- **Objectives**: ensure data is clean, informative, and optimized for machine learning models.

- **Importance**: Quality of data preparation often determines the success of the project.

Data Preparation: Data Cleaning and Preprocessing

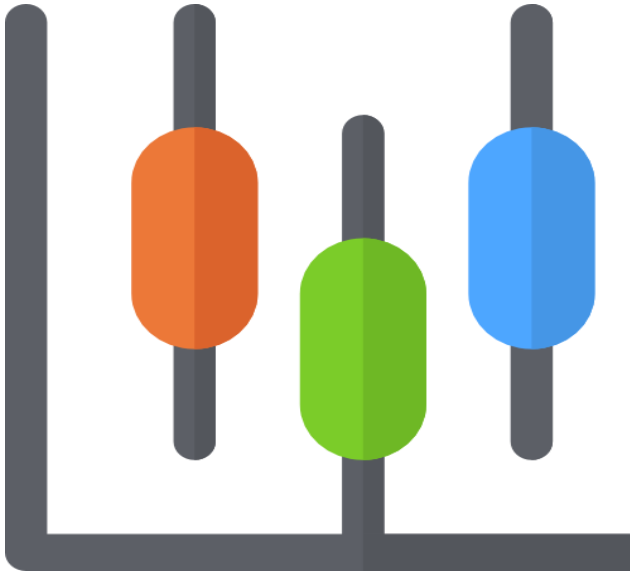
- **Handle missing data**: missing values can distort analyses and model performance. Techniques include filling missing values with mean, median, or using advanced imputation methods like KNN.

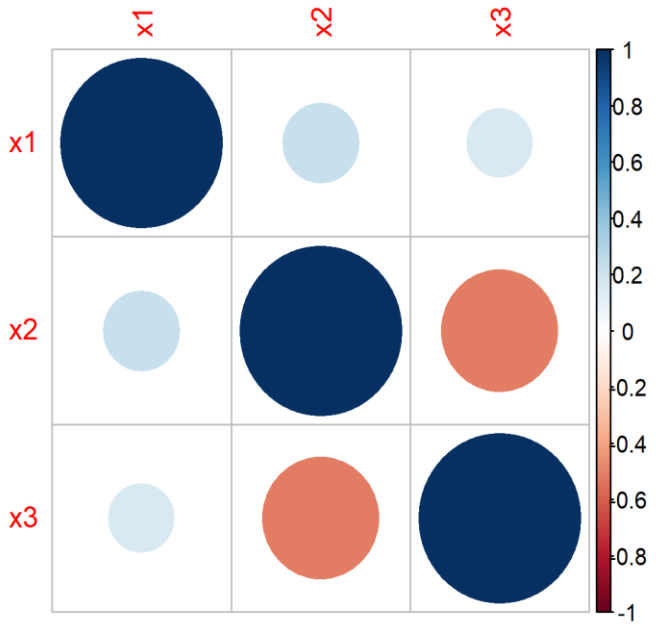
- **Remove duplicates**: duplicate entries can inflate dataset size and introduce biases. Identifying and removing duplicates ensures dataset integrity.

- **Correct data types and inconsistencies**: mismatched data types (e.g., numerical values stored as strings) can lead to errors during analysis.

Data Preparation: Exploratory Data Analysis (EDA)

- **Visualize data to understand patterns**: visualization tools help uncover trends, relationships, and outliers.
- **Summarize key statistics**: descriptive statistics provide a numerical overview of the dataset.

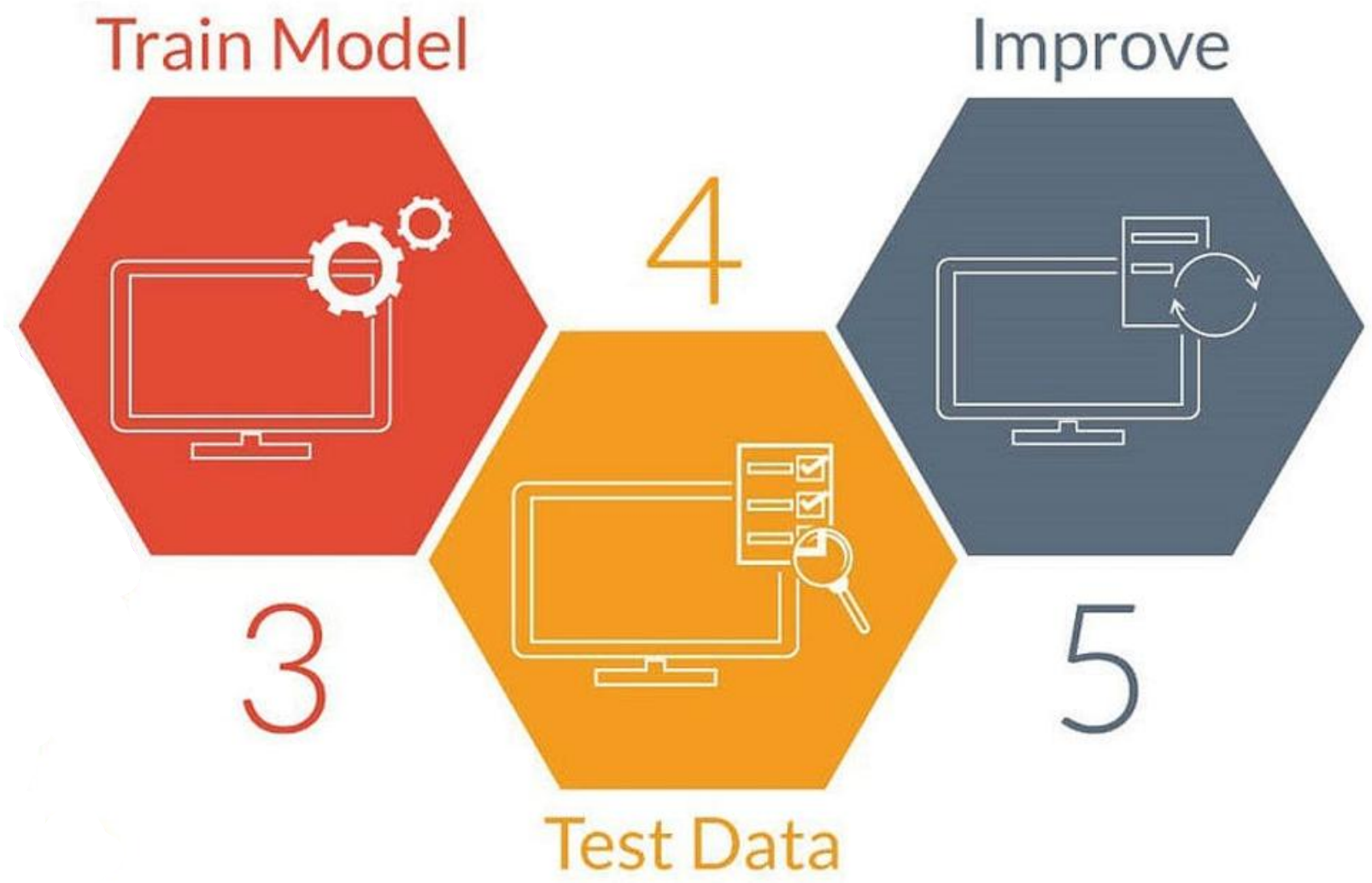




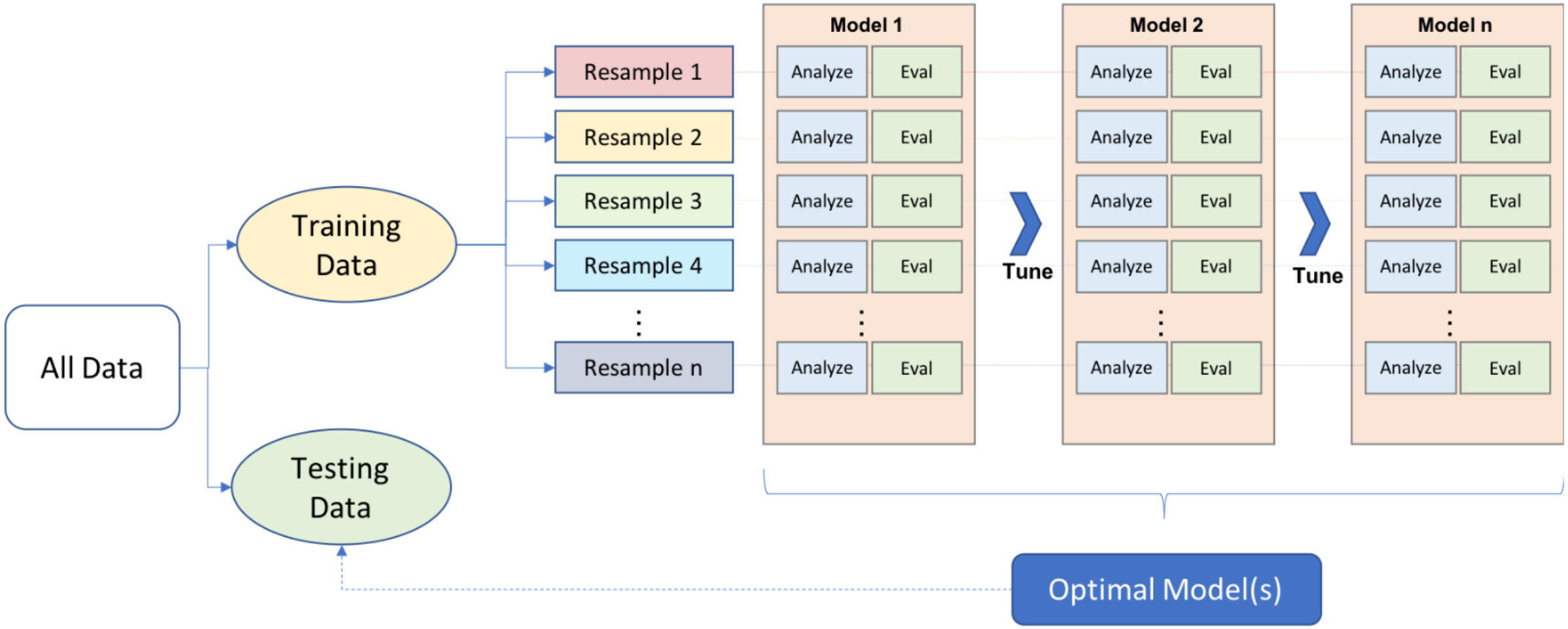
Data Preparation: Feature Engineering

- **Create new features from existing data**: new features can better represent the underlying patterns.
- **Select important features**: reducing the number of features minimizes model complexity. Techniques such as correlation analysis can identify significant features.
- **Examples**: creating interaction terms, binning continuous variables like ages into age groups (e.g., 18-25, 26-35) etc.

Model Building



Model Building



Machine Learning Algorithms

- As you will see later in the semester, there are so many algorithms that it can be feel overwhelming when algorithm names are thrown around and you are expected to just know what they are and where they fit.

- Therefore, to make it more straightforward and less overwhelming, we will categorize the algorithms you may come across in the field.

- We will do so by grouping these algorithms by their **learning style** and their **similarity in form or function**.

Machine Learning Algorithms: Learning Styles

- Classical ML is often categorized by how an algorithm learns to become more accurate in its predictions.

- There are four basic types of ML: **supervised learning**, **unsupervised learning**, **semi-supervised learning**, and **reinforcement learning**.

- The type of algorithm data scientists choose depends on the nature of the data. Many of the algorithms and techniques aren't limited to just one of the primary ML types listed above.

Machine Learning Algorithms: Supervised Learning

- A **predictive model** is used for tasks that involve the prediction of a given **output** (aka **target**, **response**, Y) using other **variables** (aka **features**, **predictors**, X) in the data set.

- The learning algorithm in a predictive model attempts to discover and model the relationships among the target variable (the variable being predicted) and the other features.

- For instance, predicting the home sales price using home attributes, predicting the risk of readmission using patient attributes and symptoms, etc.

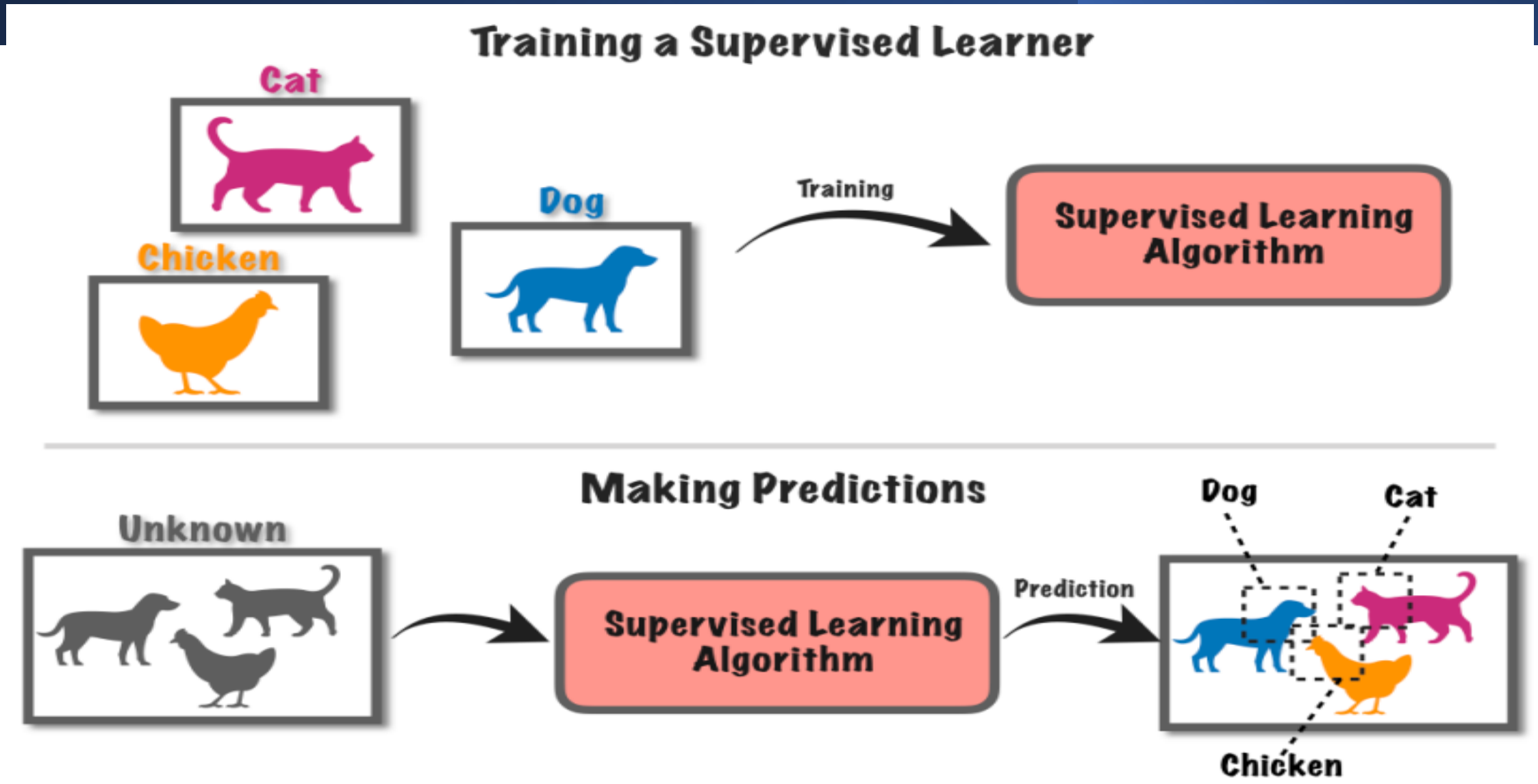
Machine Learning Algorithms: Supervised Learning

- Such predictive modeling examples describe what is known as **supervised learning**.

- The *supervision* refers to the fact that the target values provide a supervisory role, which indicates to the **learner** (the algorithm) the task it needs to learn. In other words, in supervised learning, we use **known** or **labeled** data (the target/output is known).

- Specifically, given a set of data, the learning algorithm attempts to optimized a **function** (**algorithm steps**) to find the combination of feature values that results in a predicted value that is as close to the actual target output as possible.

Machine Learning Algorithms: Supervised Learning



Machine Learning Algorithms: Supervised Learning

Most supervised learning problems can be bucketed into one of two categories, **Regression** and **Classification**.

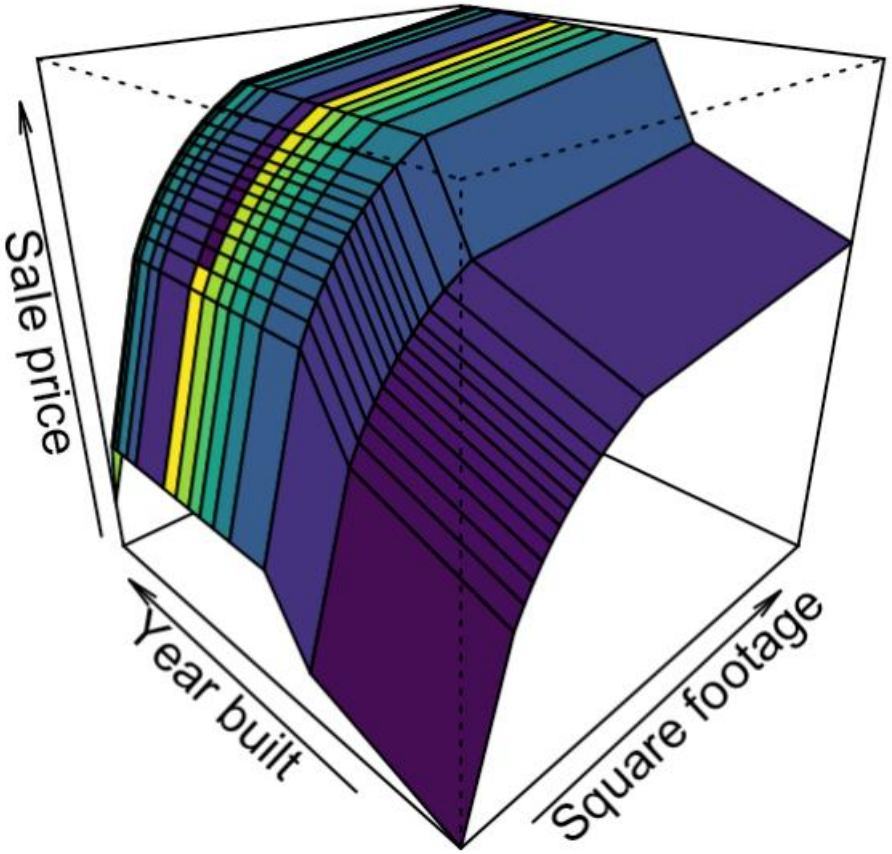
- When the objective of our supervised learning is to **predict a numeric outcome**, we refer to this as a **regression problem**.

- Regression problems revolve around predicting output that falls on a continuum. This means, given the combination of predictor values, the response value could fall anywhere along some continuous spectrum.

Machine Learning Algorithms: Supervised Learning

- The figure on the right illustrates average home sales price as a function of two home feature: year built and total square footage.

- Depending on the combination of these two features, the expected home sales price could fall anywhere along the plane.



Machine Learning Algorithms: Supervised Learning

- When the objective of our supervised learning is to predict a **categorical outcome**, we refer to this as a **classification problem**.

- Classification problems most commonly revolve around predicting a **binary** or **multinomial** response measure.

- For instance, did a customer redeem a coupon (yes/no coded as 1/0)?, did a customer click on our online ad (yes/no), predicting the disease stage based on other features (stage 1, stage 2, stage 3).

Machine Learning Algorithms: Supervised Learning

- The figure on the right illustrates an example of a classification problem with a multinomial response.
- Here the goal is to analyze hand-written digits and predict the numbers written.

Machine Learning Algorithms: Unsupervised Learning

- **Unsupervised learning**, in contrast to supervised learning, includes a set of statistical tools to better understand and describe your data, but performs the analysis without a target variable (**unlabeled data**).

- In essence, unsupervised learning is concerned with **identifying groups** in a data set.

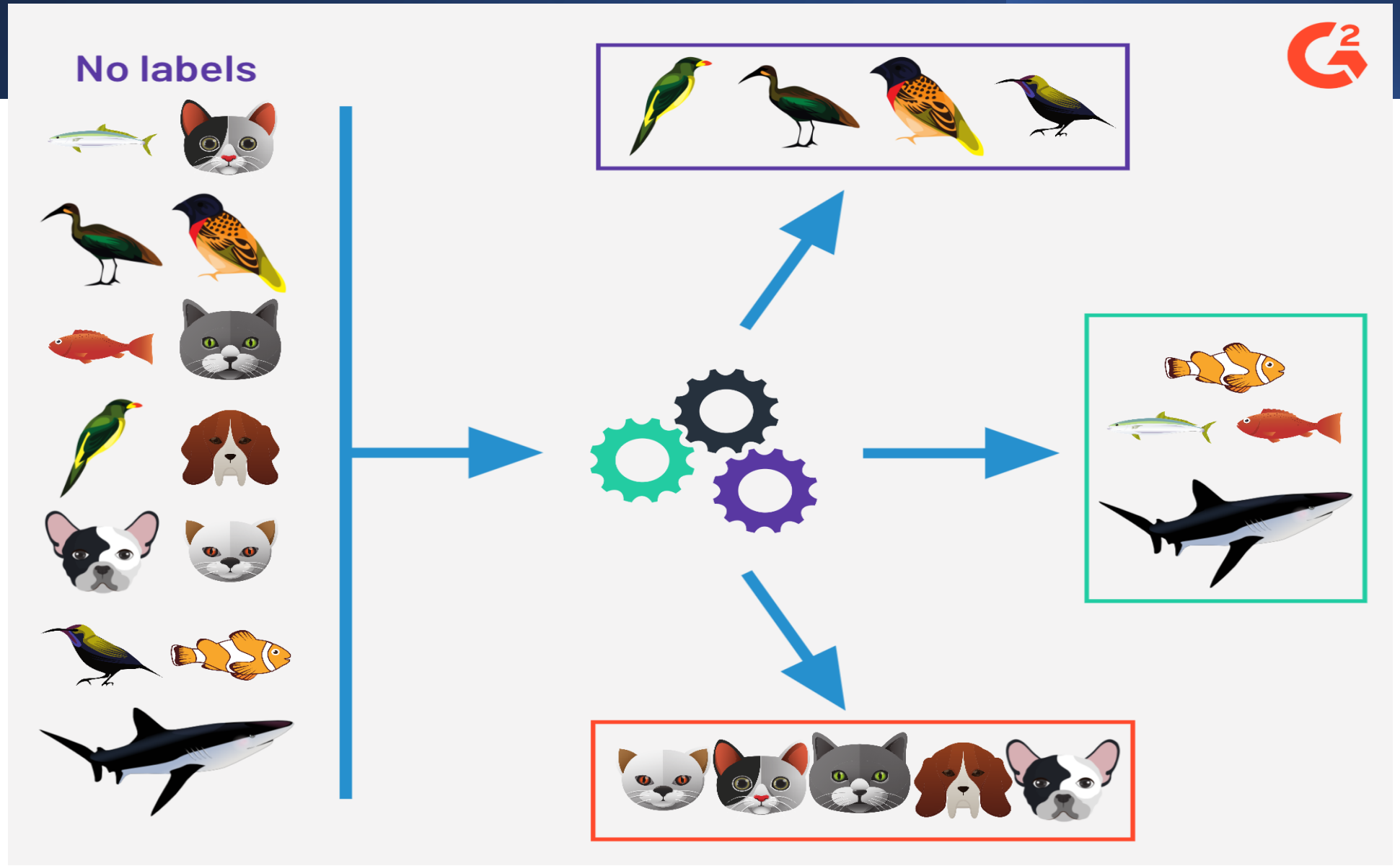
- The groups may be defined as rows (i.e., **clustering**) or the columns (i.e., **dimension reduction**); however, the motive in each case is quite different.

Machine Learning Algorithms: Unsupervised Learning

- The goal of **clustering** is to segment observations into similar groups based on the observed variables. For example, to divide consumers into different homogeneous (similar) groups, a process known as **market segmentation**.

- In **dimension reduction**, we are often concerned with reducing the number of variables in a data set. For instance, classical linear regression models break down in the presence of highly correlated features. Some dimension reduction techniques can be used to reduce the feature set.

Machine Learning Algorithms: Unsupervised Learning



Machine Learning Algorithms: Grouped by Similarity

- Machine learning algorithms are often grouped by similarity in terms of their function (how they work). For example, tree-based methods, neural network inspired methods, and so on.

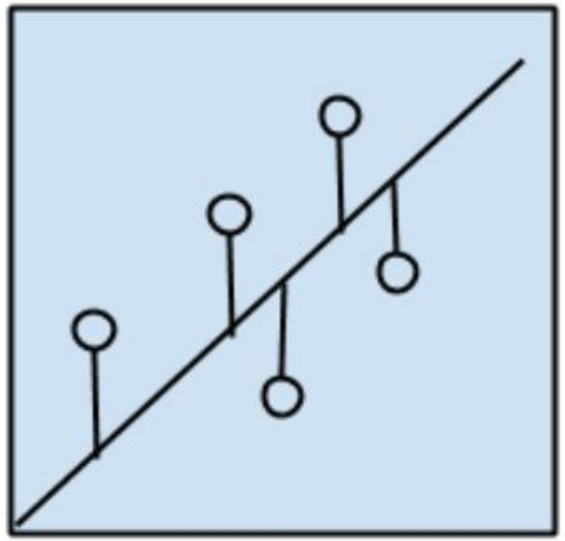
- We will segregate commonly used ML methods into the following groups: **regression algorithms**, **instance-based algorithms**, **regularization algorithms**, **decision tree algorithms**, **bayesian algorithms**, **clustering algorithms**, **dimensionality reduction algorithms, ensemble algorithms.**

Regression Algorithms

Regression is concerned with modeling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model.

The most popular regression algorithms are:

- **Linear Regression**
- **Stepwise Regression**
- **Polynomial Regression**
- **Multivariate Adaptive Regression Splines (MARS)**
- **Locally Estimated Scatterplot Smoothing (LOESS)**



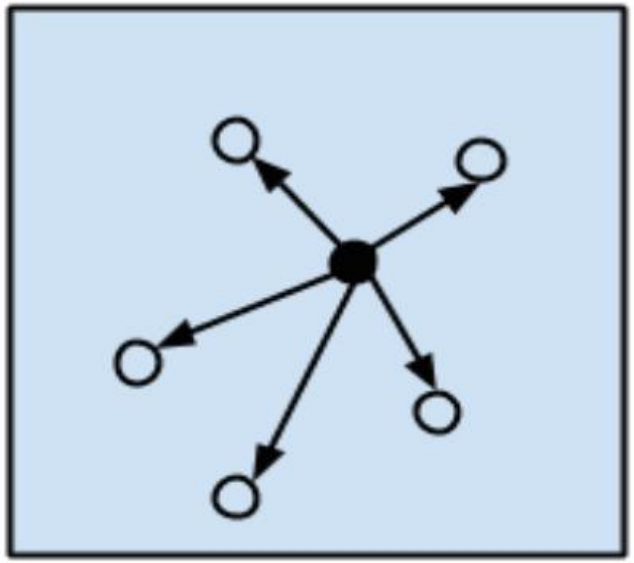
Regression Algorithms

Instance-based Algorithms

Such methods typically build up a database of example data and compare new data to the database using a similarity measure to find the best match and make a prediction.

The most popular instance-based algorithms are:

- **k-Nearest Neighbor (kNN)**
- **Support Vector Machines (SVM)**
- **Learning Vector Quantization (LVQ)**
- **Locally Weighted Learning (LWL)**

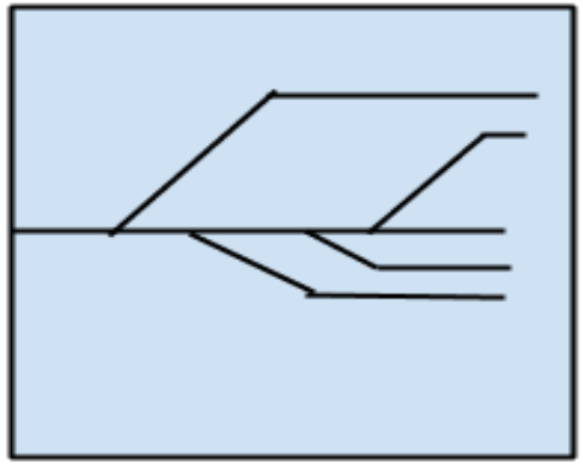


Instance-based Algorithms

Regularization Algorithms

An extension made to another method (typically regression methods) that penalizes models based on their complexity, favoring simpler models that are also better at generalizing.

The most popular regularization algorithms are:



Regularization Algorithms

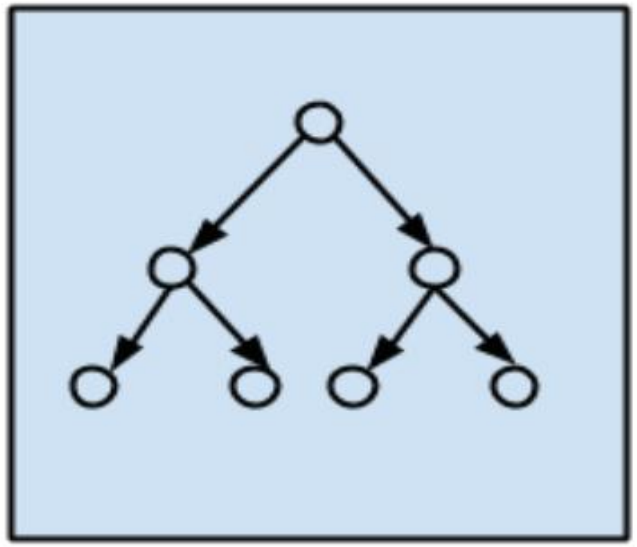
- **Ridge Regression**
- **Elastic Net**
- **Least-Angle Regression(LARS)**
- **Least Absolute Shrinkage and Selection Operator (LASSO)**

Decision Tree Algorithms

Decision tree methods construct a model of decisions made based on actual values of features in the data. Decisions fork in tree structures until a prediction decision is made for a given observation.

The most popular decision tree algorithms are:

- **Classification and Regression Tree (CART)**
- **Iterative Dichotomiser 3 (ID3)**
- **C4.5 and C5.0**
- **Decision Stump**



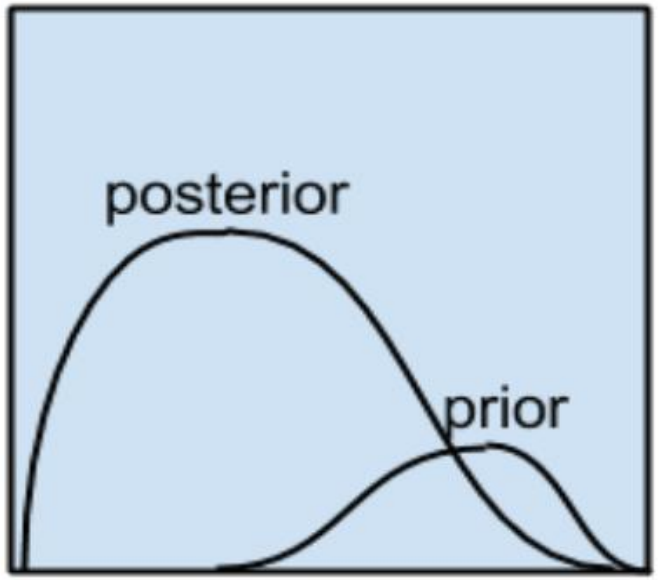
Decision Tree Algorithms

Bayesian Algorithms

Bayesian methods are those that explicitly apply Bayes' Theorem for problems such as classification and regression.

The most popular Bayesian algorithms are:

- **Naïve Bayes**
- **Gaussian Naïve Bayes**
- **Bayesian Belief Network (BBN)**
- **Bayesian Network (BN)**



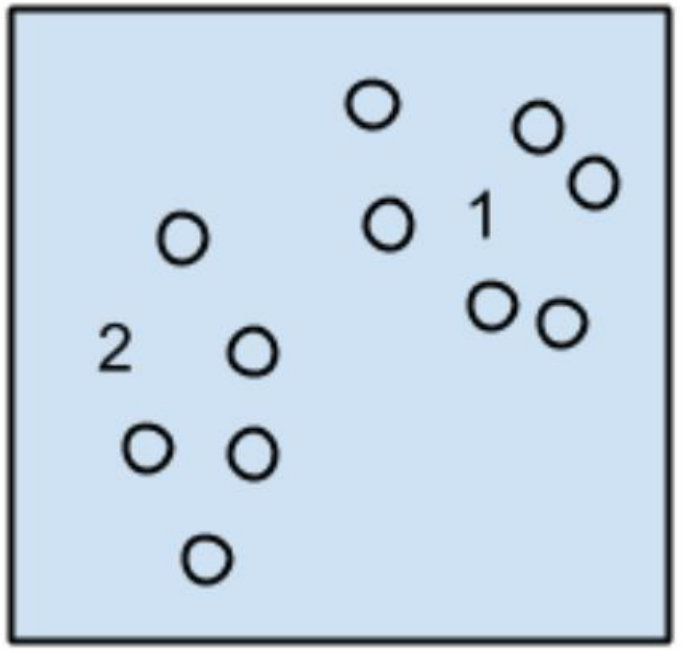
Bayesian Algorithms

Clustering Algorithms

Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchal.

The most popular clustering algorithms are:

- **k-Means**
- **Expectation Maximisation (EM)**
- **Hierarchical Clustering**
- **Model-based Clustering**



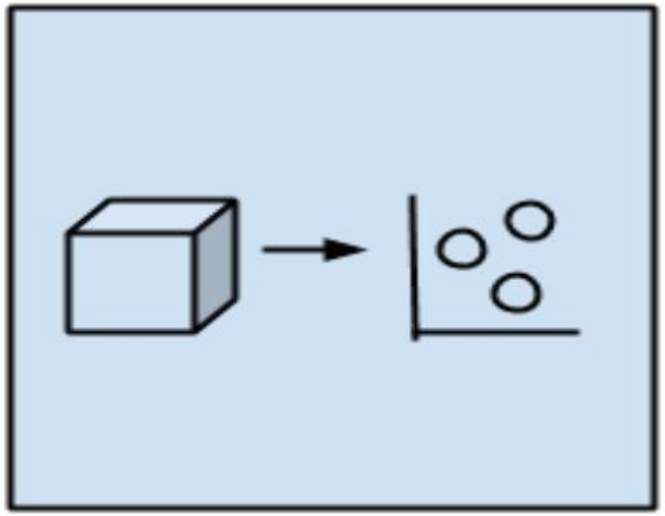
Clustering Algorithms

Dimensionality Reduction Algorithms

This can be useful to visualize dimensional data or to simplify data which can then be used in a supervised learning method.

The most popular dimensionality reduction methods are:

- **Principal Component Analysis (PCA)**
- **Principal Component Regression (PCR)**
- **Partial Least Squares Regression (PLSR)**
- **Linear Discriminant Analysis (LDA)**



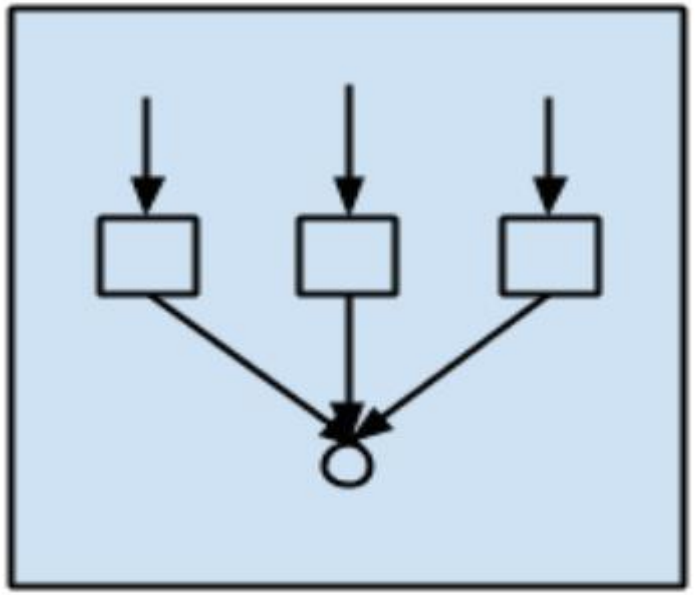
Dimensional Reduction Algorithms

Ensemble Algorithms

Ensemble methods are models composed of multiple weaker models that are independently trained and whose predictions are combined in some way to make the overall prediction.

The most popular ensemble methods are:

- **Boosting**
- **Bootstrapped Aggregation (Bagging)**
- **Gradient Boosting Machines (GBM)**
- **Random Forest**



Ensemble Algorithms

Python and Libraries

- You are encouraged to select an IDE that suits your workflow and preferences. Below are some popular options:

Python and Libraries

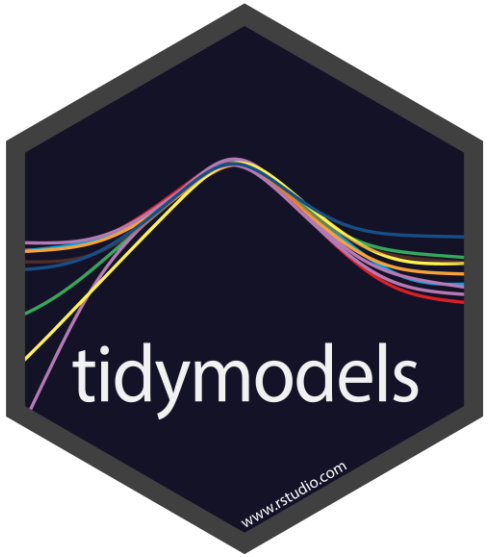
- You will mostly utilize the following libraries:

R and RStudio

- In addition, you will need to install two applications: R and RStudio:

R Packages

- You will mostly utilize the following packages:



Version Control Systems (Git and GitHub)