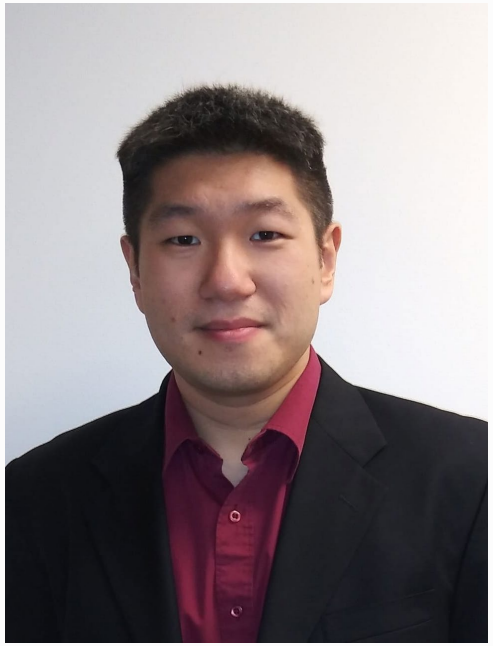
Data Science & Machine Learning Fundamentals

CFI

Instructors – Data Science & Machine Learning Fundamentals

Seb studied Economics and Finance with a focus on Econometrics. He has spent the last 15 years working across Energy, Finance and Retail delivering analytical and automation solutions. Seb is passionate about business intelligence and data analysis and loves to find simple ways to explain things.

Lester has led a startup fintech data science team and has held VP roles in quantitative analysis and strategy at Wells Fargo. He has held various data consulting roles in other Fortune 500 companies and is and expert in the application of machine learning techniques.



Seb Taylor

VP of Business Intelligence & Data Analysis

Lester Leong

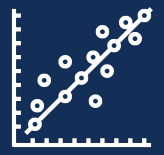
Data Science Professional & Consultant

Learning Objectives

Distinguish data science from business intelligence

Outline the data science and machine learning process

Describe basic data science terms, roles, skills, and applications



Distinguish popular models used in data science and machine learning

Identify data preparation steps that will help you explore your data, remove errors, and structure it to be easier to work with

Evaluate model results

What is Data Science?

What is Data Science

Data science is all about creating data driven insights that help us deal with uncertainty.

Which type of customer is most likely to buy our product?

What type of market regime are we entering?

When will we run out of warehouse stock?

How many ice creams should we order given the forecast next week?

Data science can be easily confused with **business intelligence (BI)**.

BI is generally **backwards looking** (descriptive).

Data science uses past observations **to make predictions, estimations and decisions about the future**.

Example Questions – BI or Data Science?

Start-up
What proportion of our crowd-sourced investors invested $200 or less?

BUSINESS INTELLIGENCE

Bank
What proportion of our loans were issued to at risk customers?

BUSINESS INTELLIGENCE

Store Chain
What proportion of our Q1 forecasted sales come from the Pet Food category?

BUSINESS INTELLIGENCE

Example Questions – BI or Data Science?

Financial Institution
Based on past transactions, which of these new transactions are likely fraudulent?

DATA SCIENCE

Manufacturing Company
Based on sensor data, when is this critical machine component likely to wear out?

DATA SCIENCE

E-commerce Company
Based on sales data, which of our high-value customers are most likely to leave?

DATA SCIENCE

Types of Analysis

Descriptive

Provides a view of the facts of **who**, **where**, **when**, **how many,** and **what** exactly happened?

Predictive

Provides a probable state of the **future** or an **unknown variable**.

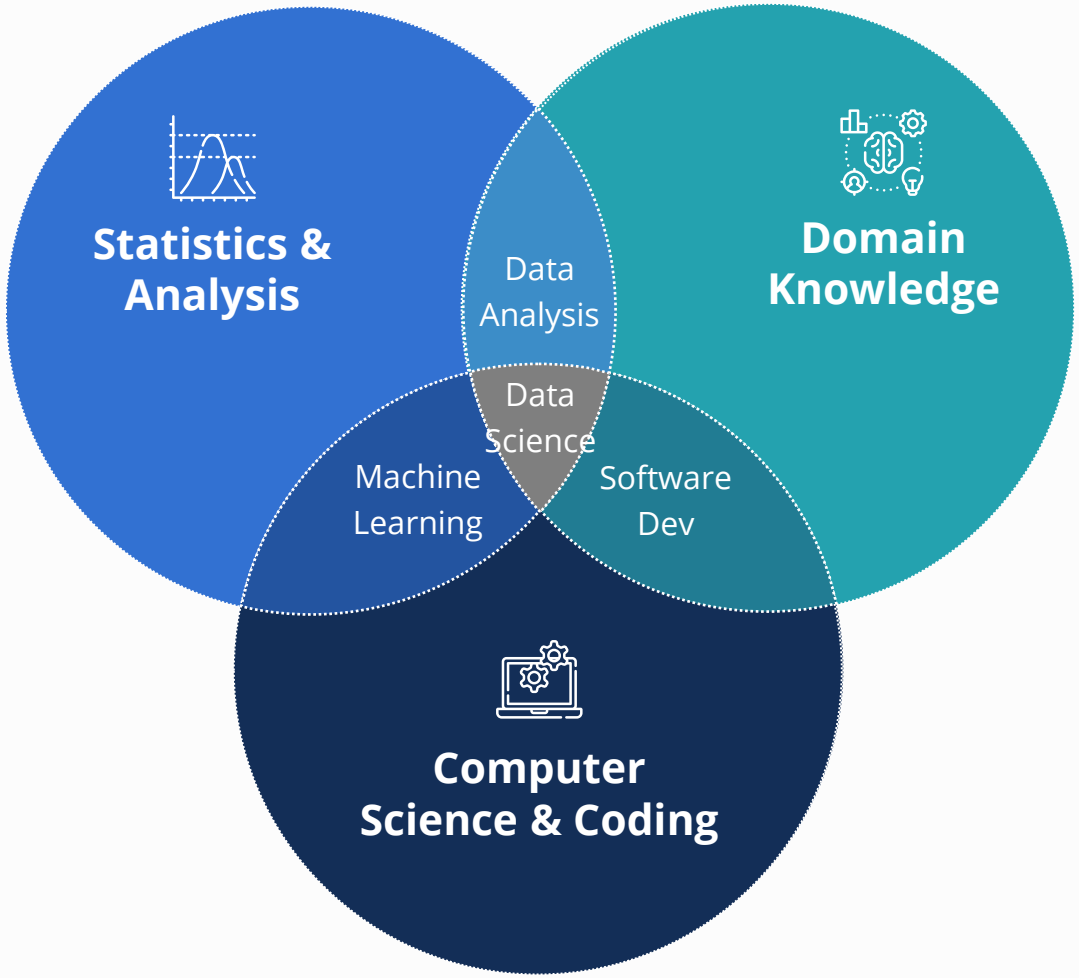
Diagnostic

Provides an analysis to tell us **why** something is happening—**what was the leading cause**?

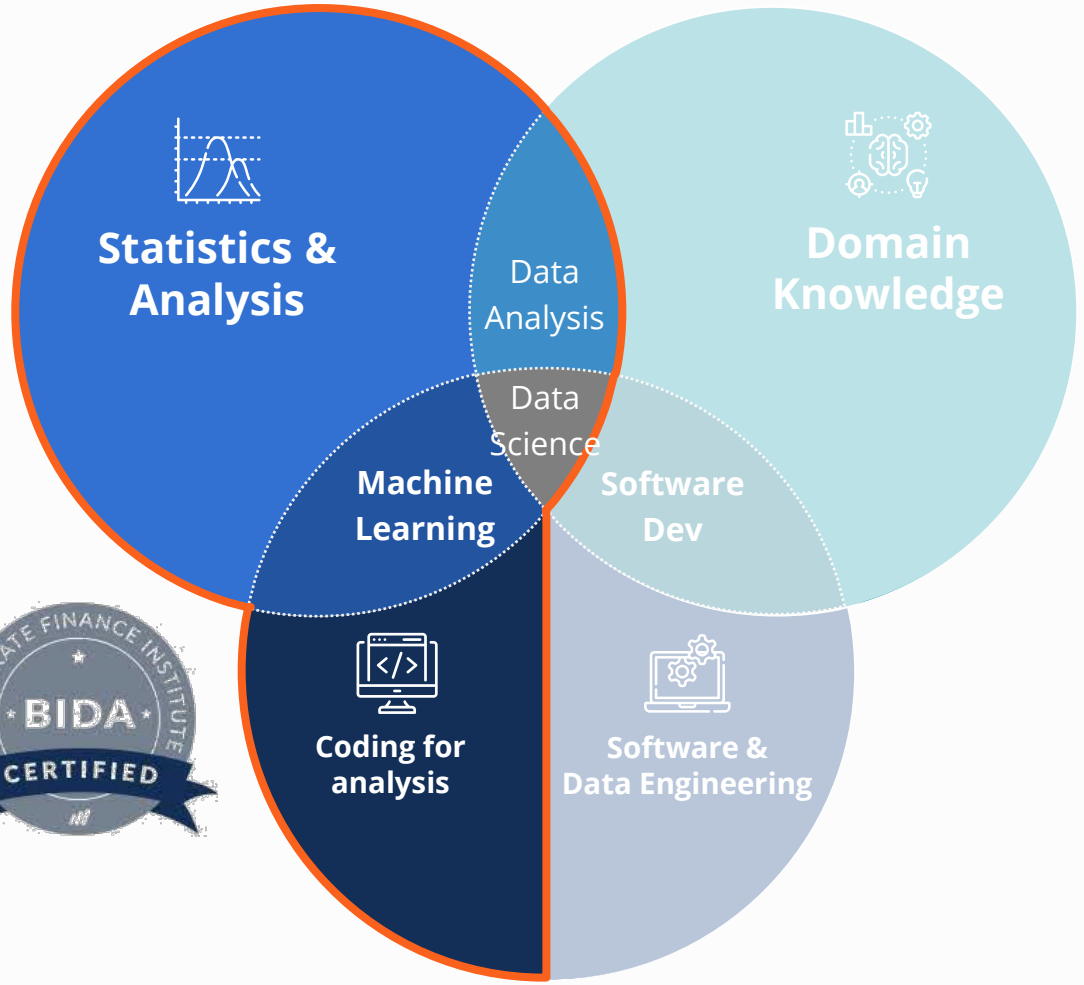
Prescriptive

Provides the **best course of action** in order to achieve a given outcome.

Data Science Skills



Two Types of Computer Science

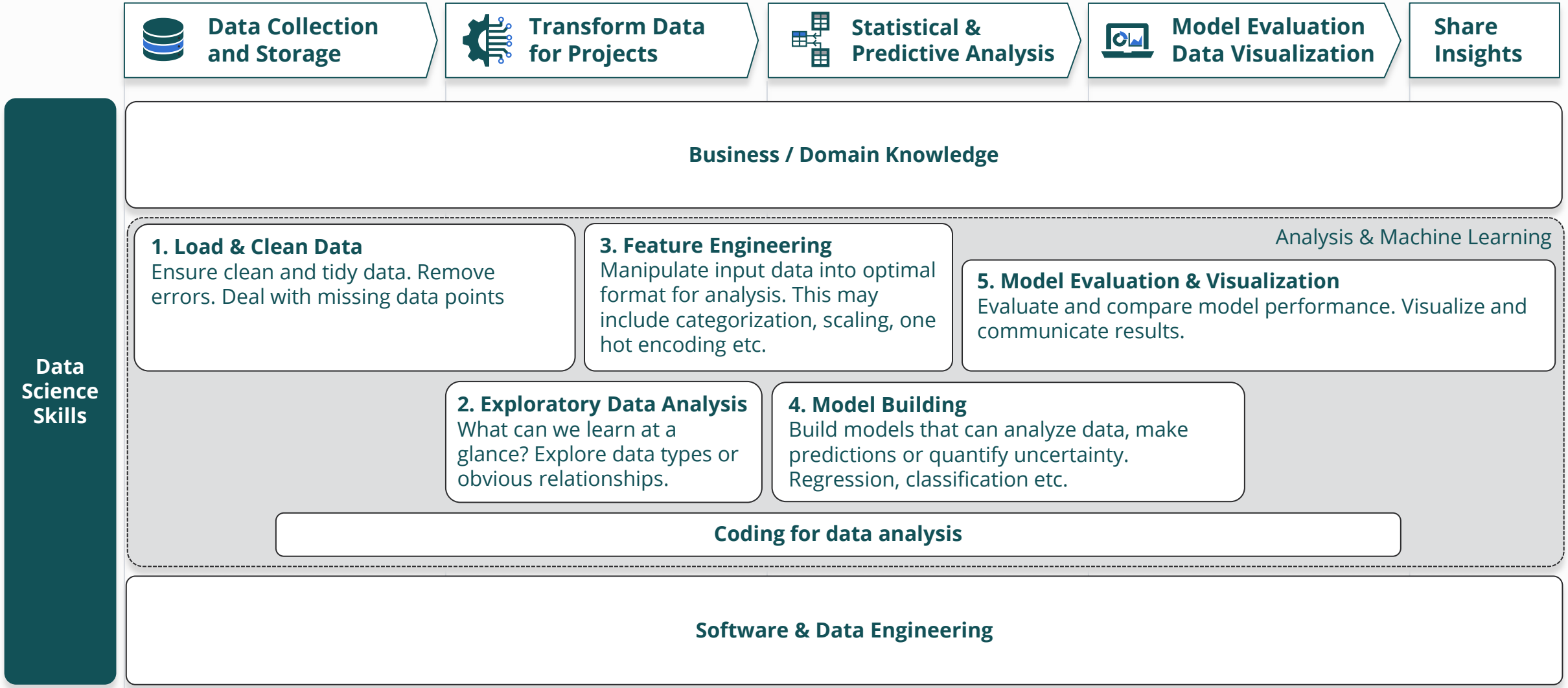


The Data Science Process

	Definition	Scenario
Data Collection and Storage	**Capture information**, ensure quality of data, and **store data** into database.	**Collect transaction data**, user-names, credit history. **Identify past fraud**.
Transform Data for Projects	**Optimize data** for the project we're working on and **select features of interest**.	**Combine or manipulate datasets**, filtering out items, or adjust formatting.
Statistical & Predictive Analysis	**Build models and algorithms** that spot patterns in our data.	**Train model** to identify the leading indicators of fraudulent transactions.
Model Evaluation Data Visualization	**Test** how well is the model performing? **Present results** using data visualization.	**Which model is best** at identifying fraudulent transactions? **Optimize** for business objectives.
Share Insights	**Share dashboards** and reports to business users for use in decision making and **deploy our models** into operations.	**Share real time information** identifying risky transactions.

Machine Learning Skills

There is some **specific terminology** used when discussing the Machine Learning process.



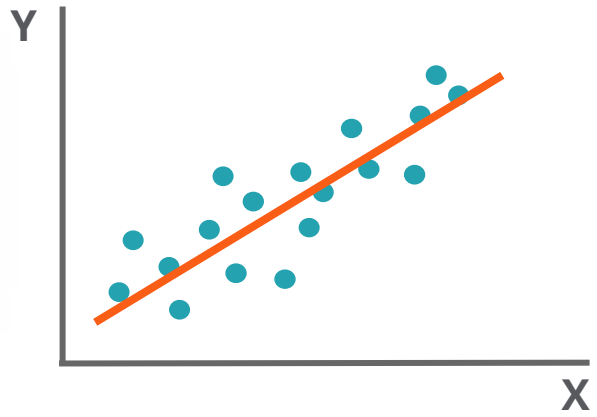
Types of Machine Learning

Supervised Machine Learning

- More common in business to answer **pre-defined questions**.
- **Predict a target variable** based on input data.
- Once model is trained on example data, predictions can be made on new data.
- Ensemble models are **combinations of other models**.

Input Data (Features) | **Target Data**

Income	Credit Score	Age	Default Loan
$56k	755	43	No
$38k	682	22	Yes
$120,000	731	38	No
$65,000	595	54	Yes
$52,00	784	68	No

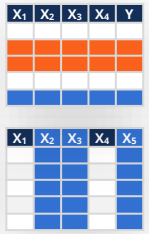


Classification Problems
Which one? What category? True or false?

Regression Problems
How much? How many?

Unsupervised Machine Learning

- No specific question in mind
- Point us in the right direction



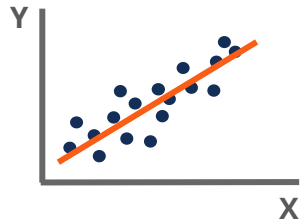
Clustering Problems

Variable Reduction

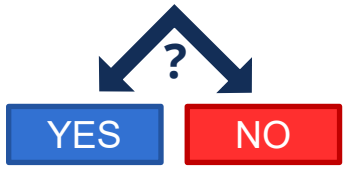
- Identifies the most important features in a dataset.

Types of Data Science Models

Supervised Machine Learning



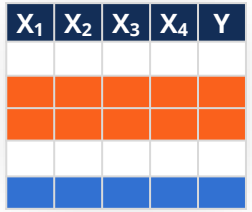
Regression



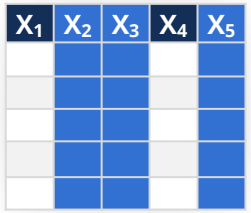
Classification

Variations: Imputation & Time Series Regression

Unsupervised Machine Learning



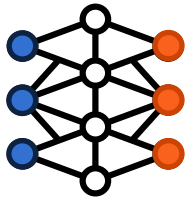
Clustering



Variable Reduction

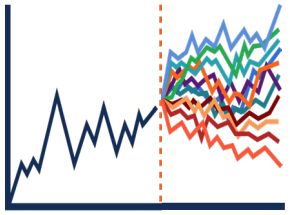
Other Machine Learning Models

Reinforcement Learning



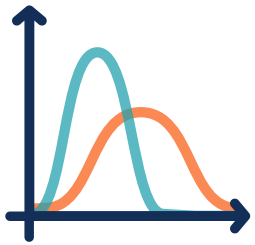
Neural Networks & Deep Learning

Other Data Science Models



Monte Carlo Simulation

Rule Based Models



Other Statistical Models

Priorities for Business Leaders

Model evaluation is the most important part of the data science process from a leadership perspective.

Data Scientists, Analysts & Engineers

Solve technical challenges and select the best analytical approach.

+

Business Leaders

Provide insight on business objectives, project goals and business costs.

=

Successful Data Science Project

Projects and models that deliver targeted value to the business.

Business leaders and data science teams should work closely to align **priorities**, **objectives** and **measures of success**.

Business leaders need a **basic understanding of model outputs**, and their impact on decision making.

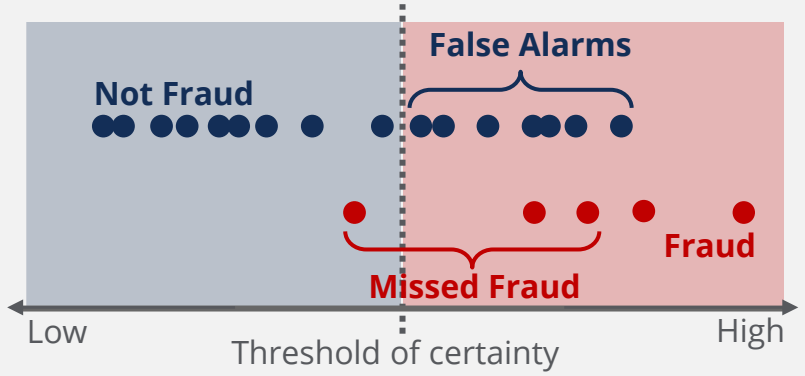
Model Objectives

Why are we building this model and what outcome are we targeting?

Identify Fraudulent Transactions
Using transaction data

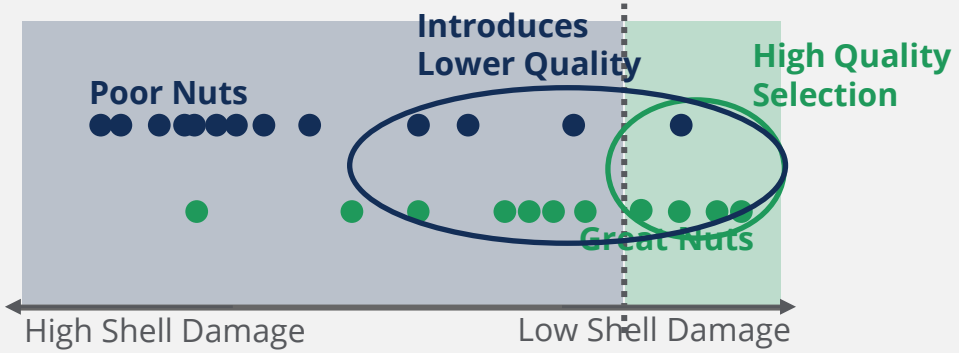
$

Why?
- Reduce the workload on human investigators?
- Meet a regulated level of fraud detection?
- Fulfill an ambitious claim by our marketing department?



Automated Nut Filtering
Using laser scans of nuts

Possible Objectives
- Are minimum standards enforced by food laws?
- Is quality dependent on crop?
- Perhaps standards are assessed versus competitors?



The business must **quantify the cost of false alarms** and the **benefits of correct identification**.

Model Limitations

No data science model can be 100% accurate.

As we chase higher accuracy, the greater the marginal cost of time and resources needed to achieve it.

How good is good enough?

Business knowledge helps leaders **balance results with resources.**

What is the current cost of doing this process manually?

What is the $ cost of improving this process?

What is the $ cost of our data science resources?

Does the Data Science team have the resources to complete all these projects well?

Allocate all resources to a single data science project?

OR

Improve 3 processes by 15%

Business leaders should work closely with DS teams to ensure **expectations, objectives and resources are aligned**.

Evaluation Metrics

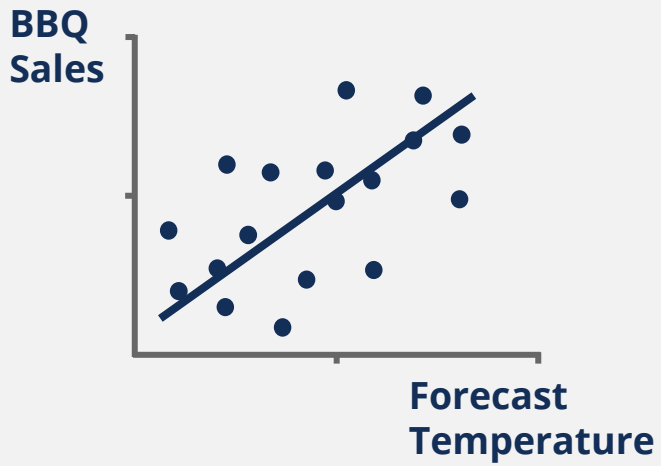
Success metrics help us measure how well models are meeting our expectations.

False Positives and False Negatives

Actual \ Prediction	NOT SPAM (-)	SPAM (+)
NOT SPAM (-)	True Negative	False Positive
SPAM (+)	False Negative	True Positive

Different business scenarios require a **different balance of results**.

Regression Coefficients



Relationship: Every +1 degree = 20 BBQ Sales

Fit: Does our line explain all the variation?

Leaders should **understand the basics of model evaluation**, to help challenge and discuss outcomes.

Priorities for Business Leaders Summary

Basic knowledge of model evaluation is essential for productive collaboration.

Model Objectives

Aligned on what we are trying to achieve?

Model Limitations

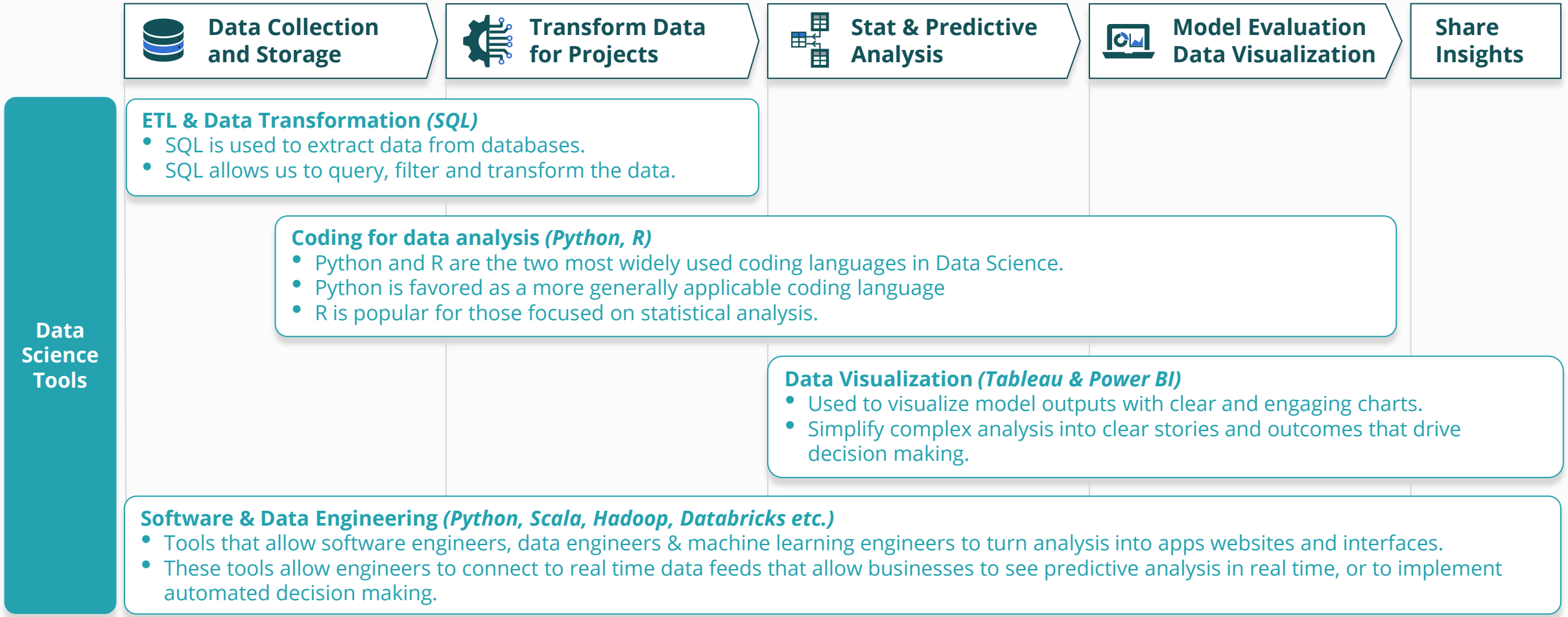
When to stop & how to distribute resources.

Evaluation Metrics

Implications of predictions, and whether they meet our objectives.

Data Science Tools

Data Science tools have each evolved for a specific purpose. Notice how a number of these tools are shared with Business Intelligence.



Data Science Roles

Data Science roles can be highly specialized or very general. It's always important to read the job description

Data Science Roles

Data Architect
- Creates the data strategy inc. how, where, when, and what data is stored

Data Analyst
- Focuses on the analysis that we all are most familiar with
- Sourcing data, formulas, data models, pivot tables, and visuals
- Understand the business well and search for insights in data

Data Engineer / SQL Developer
- One of the more technical roles in BI
- Ensure data quality & availability of data and queries
- Ensure that analysts have what they need to do their job

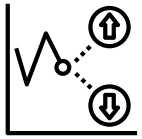
Data Visualization Specialist
- Turn insights into meaningful visuals that drive action
- Understand how the business works and how people think
- Communication skills are vital to their success

Database Admin (DBA)
- Acts as a caretaker and gatekeeper for a database
- Security, access, changes, and performance

Data Scientist *(more focused on coding for analysis)*
- Proficient in advanced statistical methods and coding, which is used to create analysis and predictions from data
- Their coding tends to be focused on analyzing the data itself

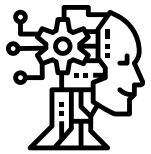
Machine Learning Engineer (*more focused on Software and Data Engineering*)
- Integrate analysis and predictive models into a real-world systems, apps or websites.
- Link models with automated data feeds that often update in real time.
- Will likely know several coding languages and have a highly technical skillset

What about Artificial Intelligence (AI)?



Machine learning is the process by which computers learn from data and make inferences or predictions.

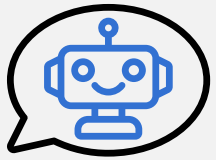
Machine learning is a **subset** of artificial intelligence.



Artificial intelligence is when a computer replicates human thinking or abilities.

Artificial intelligence is a **broader** subject than machine learning.

- **Autonomous cars** evaluate surroundings
- Make decisions and take action, just **like a human**



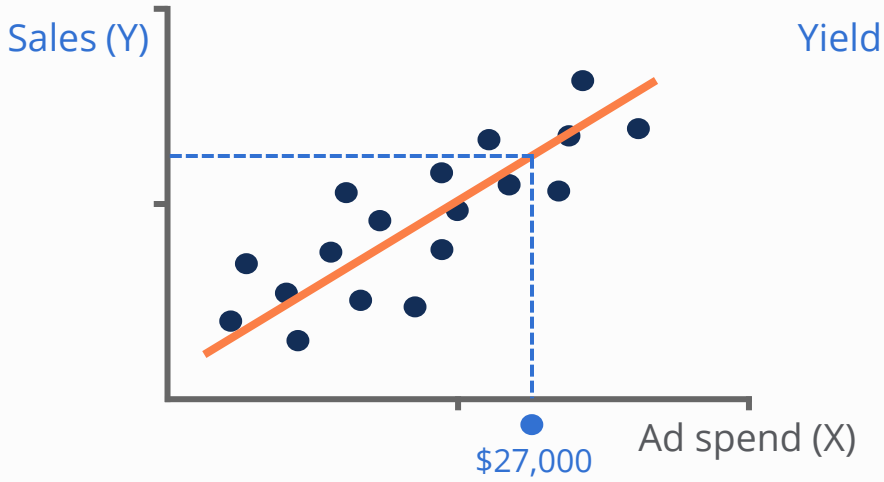
- **Chat bots** evaluate message contents
- Reply with useful links, or tips, just **like a human**

- **Trading algorithms** follow statistical rules
- Buy and sell stocks, just **like a human**

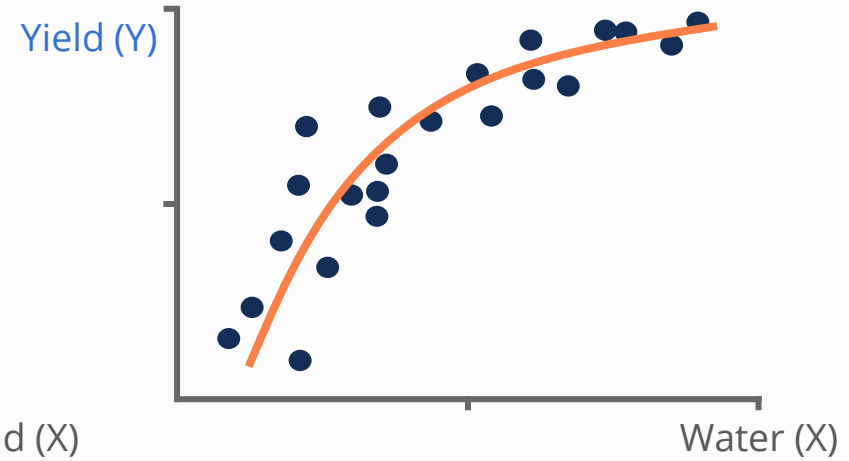
Regression Basics

Regression – Theory & Business Objectives

The goal of regression is to assess the relationship between one or more input variables (X) and a continuous output variable (Y).



How **sales** is affected by **ad spend**.



Crop yield verse **water**.

Our **line of best fit** allows us to make predictions about the value of the target variable in a given scenario.

Regression - Terminology

Predictor Variable(s)	Target Variable
Input Variable(s)	Output Variable
Independent Variable(s)	Dependent Variable
Explanatory Variable(s)	Response Variable
X (X1, X2, X3 ...)	Y

Linear Regression

Linear Regression is the simplest form of Regression Analysis.

$$y = mx + c$$

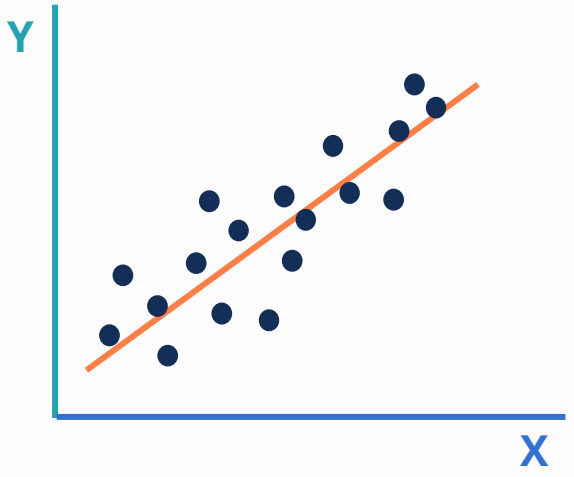
Y is the Output (target) variable

X is the Input variable

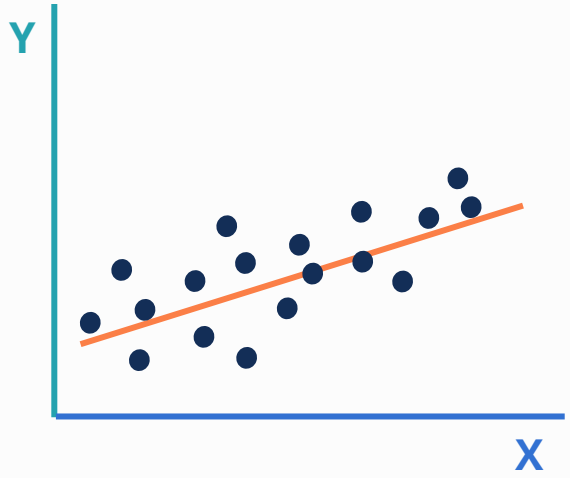
M is the Coefficient value (**slope**)

C is the Intercept (of the Y axis)

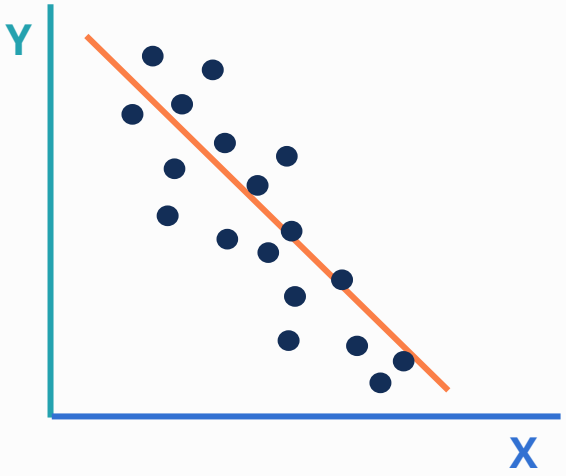
Strong positive



Weak positive



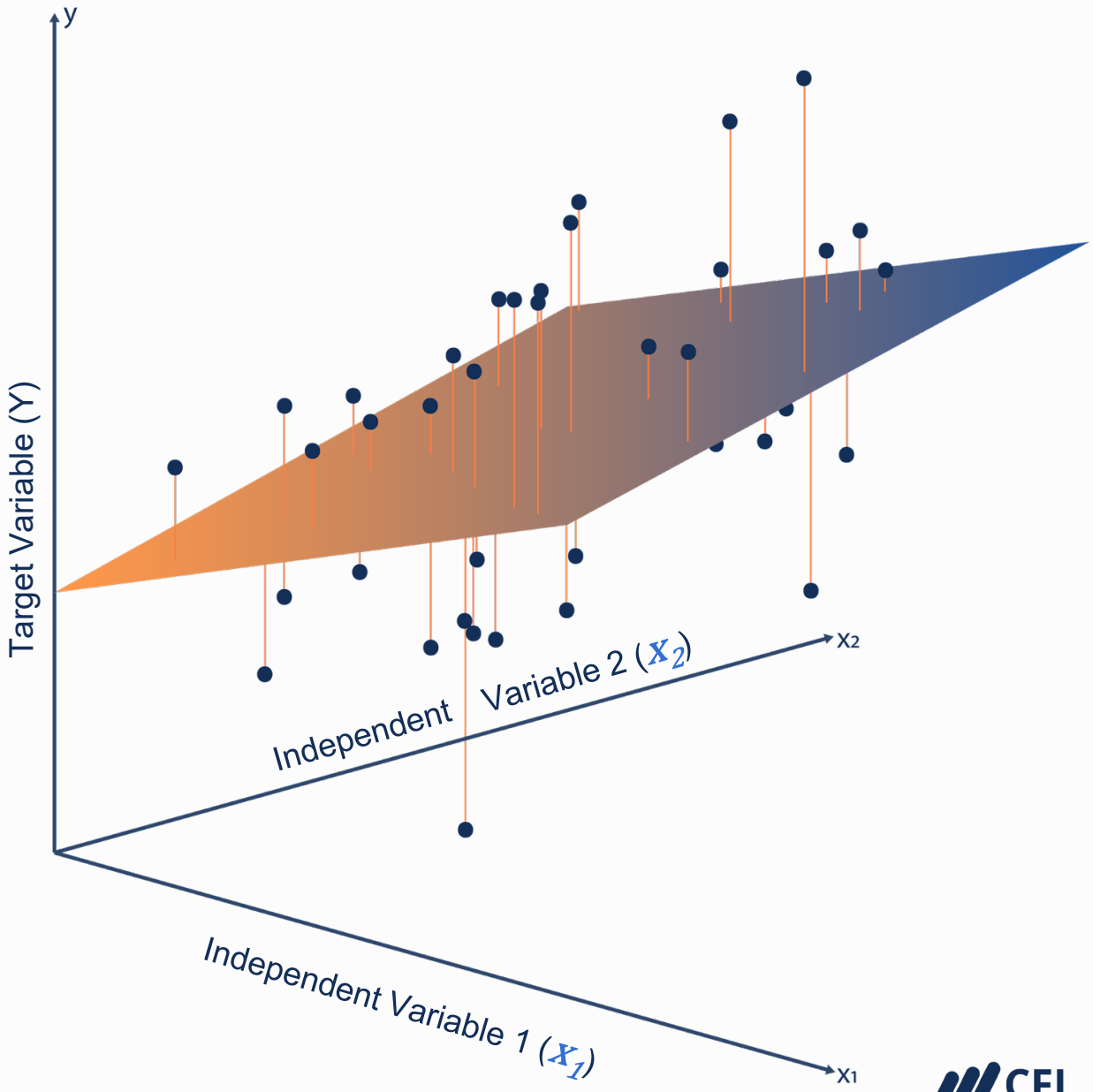
Negative



Multiple Linear Regression?

- One input variable is rarely enough to make predictions about a target variable.
- **Multiple linear regression** allows us to predict a target variable using **multiple independent variables**
- When we have two independent variables, we are fitting a plane to the data instead of a straight line

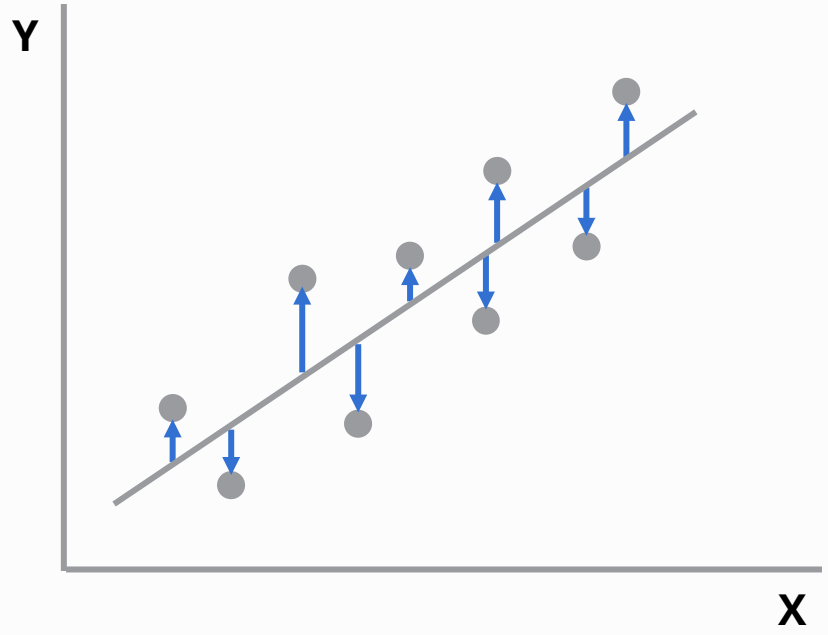
$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$



The Optimal Line of Best Fit

How do we decide where exactly the line of best fit sits? Generally, we try to **minimize the size of errors** in our predictions.

Errors represent the amount by which the target variable is different from the value predicted.



Other Metrics	Sensitive to Outliers
Sum of **Squared** Error (SSE)	Yes
Mean **Squared** Error (MSE)	Yes
Root Mean **Squared** Error (RMSE)	Yes

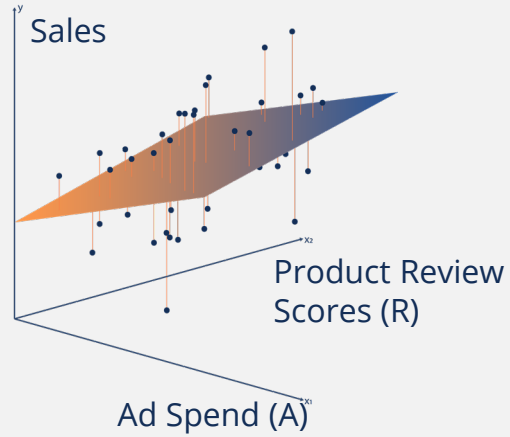
Other Metrics	Sensitive to Outliers
Sum **Absolute** Error (SAE)	No
Mean **Absolute** Error (MAE)	No

The most common approach **minimizes the squared errors,** and is know as **Ordinary Least Squares**

Interpreting Coefficients

Regression coefficients help us understand the interaction between variables

Marketing Scenario



$$Sales = 2{,}000 + 1.3\boldsymbol{A} + 0.11\mathbf{R}$$

Where **A is Ad Spend** and **R is Product review scores**

The **coefficient for Ad Spend** is 80

We can say that for every unit increase in ad spend, **sales will increase by 1.3**

The **coefficient for Sunlight** is 7

We can say that for every unit increase in Product Review Scores, **sales will increase by 0.11**

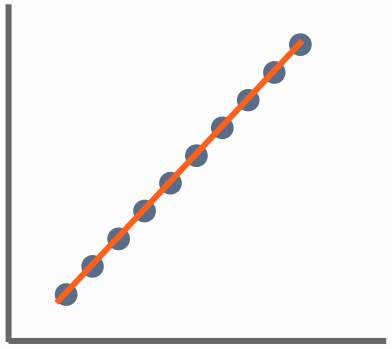
P-values help us understand if our **coefficients are statistically significant**.

Regression Metrics

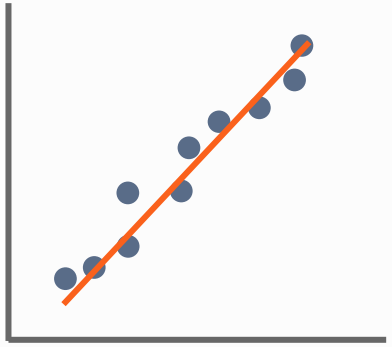
Coefficient of Determination (R^2) is one of the most used metrics to evaluate regression models.

R^2 measures how close the data are to the fitted regression line. In other words, how much of the variability in Y is explained by changes in X.

$R^2 = 1$



$R^2 = 0.86$



$R^2 = 0$

Higher R^2 indicates better fit of the model, and therefore smaller errors.

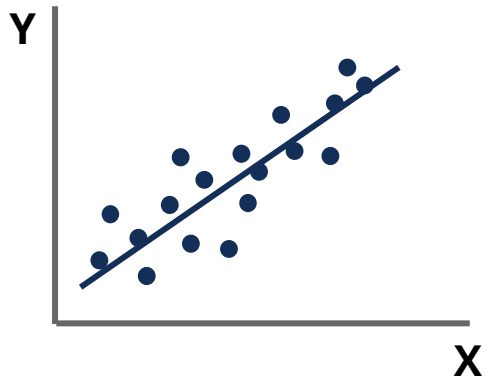
R^2 can sometimes be biased, so a related measure called **Adjusted R^2** can also be used.

Training and Testing

Models need to be tested on new data, before we allow them to make real world decisions.

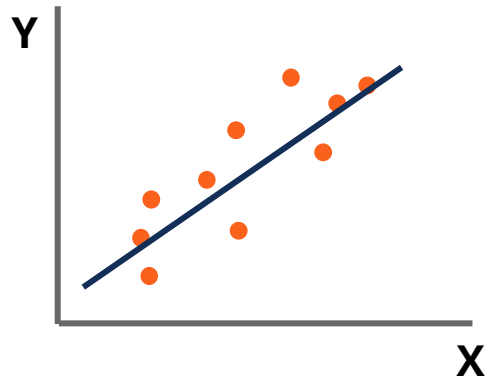
Approx 80% of sample data is used to teach (train) the model.

Approx 20% of the data is used to test how well the model performs.



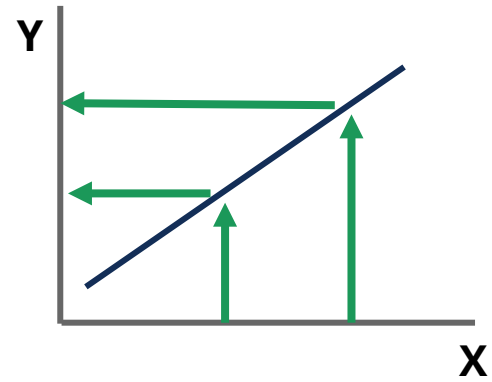
Training Data

Used to teach the model what the relationship looks like



Testing Data

Used to test model performance on new data



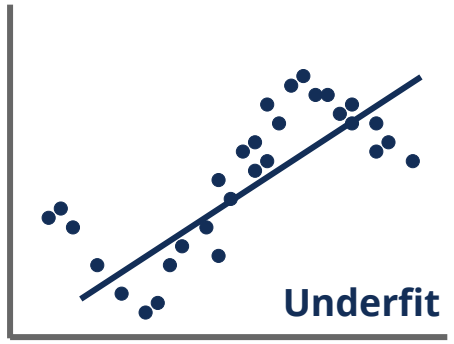
Real World Data

Used to make real world predictions and decisions

It is important that models are tested on **data they have never seen before**.

Underfitting vs. Overfitting

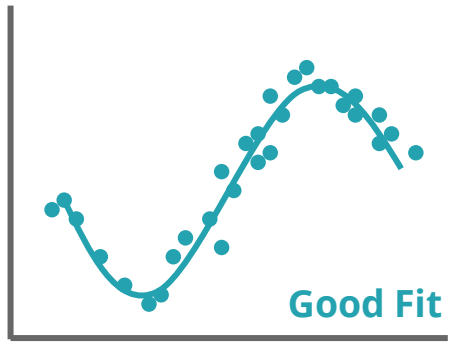
Summary: Overfitting and underfitting describe how well (or poorly) a model fits the training data. We can use performance metrics on the training and testing data to test for these scenarios.



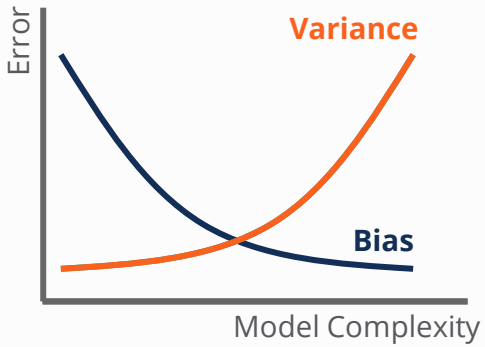
Over generalizes the trend

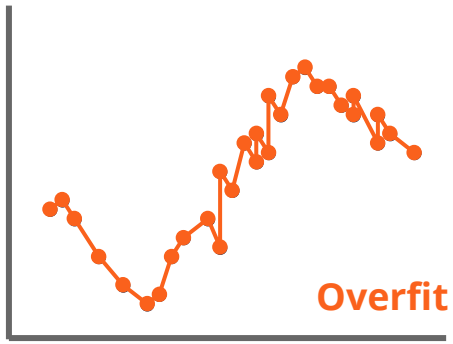
Cause: **Model too simple**

High Bias Error
Bias is the **error created by oversimplification.**



Bias vs Variance Tradeoff





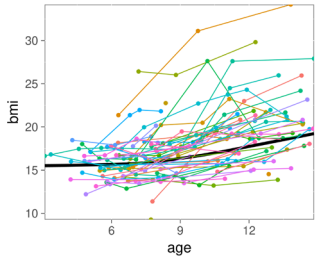
Learns a relationship that its **too specific to this sample**

Cause: **Model too complex**

High Variance Error
Variance is the extent to which the model has focused **too much on the randomness in the training data**.

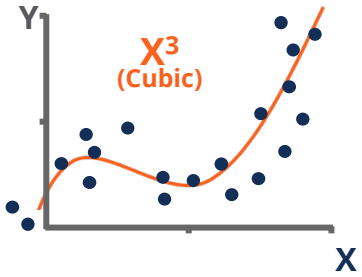
Other Regression Techniques

Repeated Measure Regression



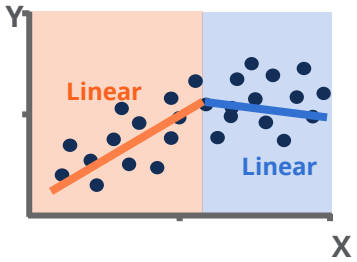
Values sampled repeatedly. Common in medicine.

Polynomial Regression



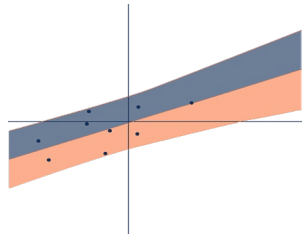
Useful for smooth, non linear relationships

Segmented Regression



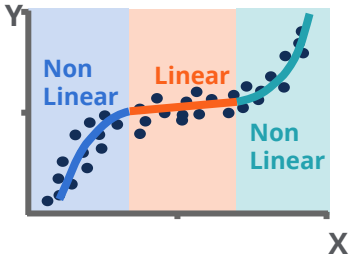
Best fit may differ by sample region

Bayesian Regression



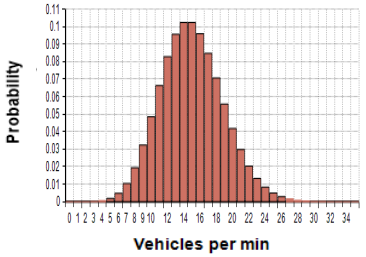
Provides a level of certainty with outputs

Generalized Additive Models



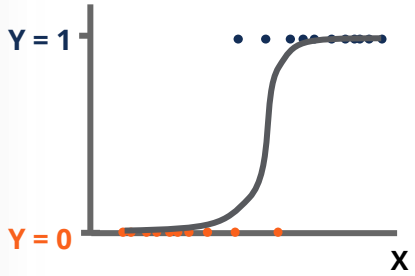
Combines different regression techniques

Poisson Regression



Used to model counts of something

Logistic Regression

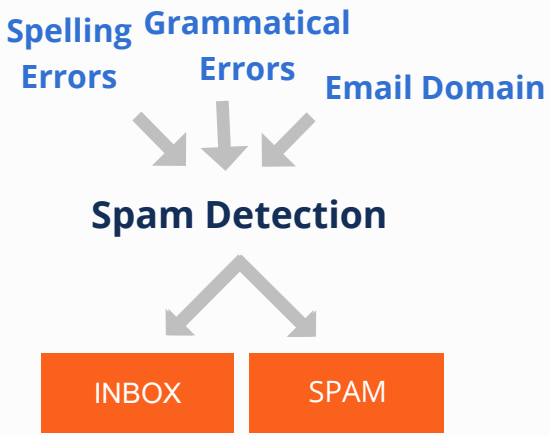


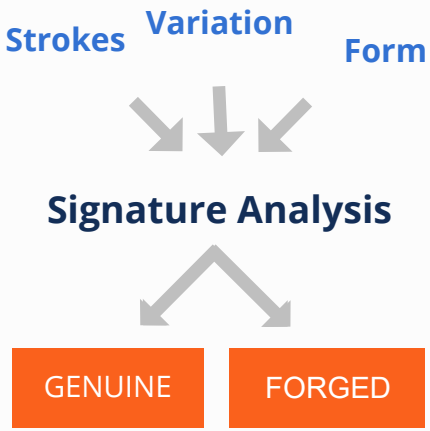
Typically used for classification

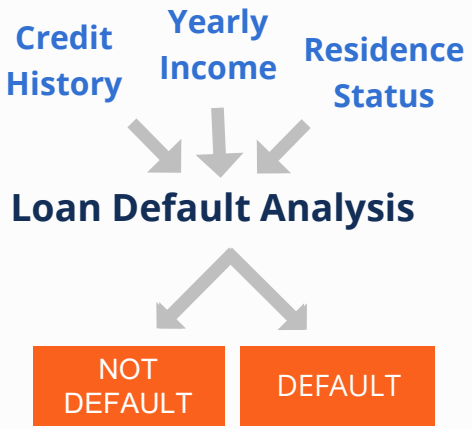
Classification Basics

What Is Classification?

- A classification problem involves classifying data into different labels/categories.
- The target categories should be **discrete variables**
- Predictions are made using one or more **input variables**.







Theory & Business Objectives

Binary

- Classification tasks that have *two class labels*
- Outcomes must be ONE of the two classes

Use Case	Output Classes
Tumour diagnosis	**Malignant** OR **Benign**
Email Spam Detection	**Spam** OR **Not Spam**

Multi-Class

- Has *more than two class labels*
- Outcomes must be ONE of a range of classes

Use Case	Output Classes
Tumour diagnosis	**Malignant** OR **Benign** OR **Premalignant**
Customer Prediction	**Will Buy** OR **Will Not Buy** OR **Insufficient Data**

Multi-label

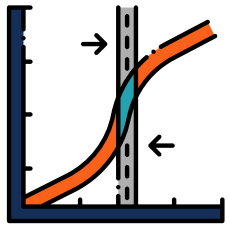
- Has **two or more** class labels
- Outcome can be ***ONE or MORE*** of the class labels

Use Case	Output Labels
Social Tag Optimization	**#MachineLearning** AND / OR **#DataScience** AND / OR **#DataAnalysisJobs**

Classification Algorithms

In the rest of this course, we'll explore the most common classification algorithms.

Logistic Regression



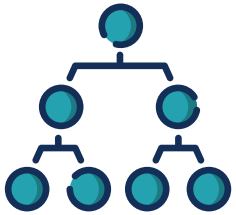
Uses regression principles to achieve separation between discrete classes

Naïve Bayes

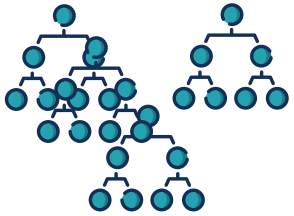
KNN

SVM

Decision Trees



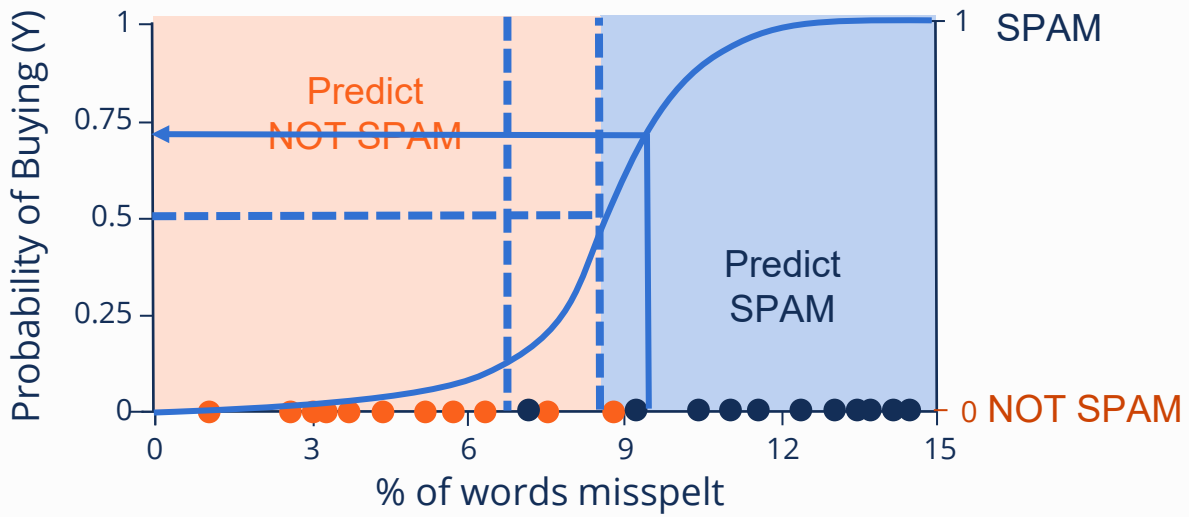
Random Forest



Once we understand each technique, we'll compare and contrast the benefits, and outputs.

Visualizing Logistic Regression (II)

Logistic Regression probabilities are estimated using **one or more input variables**



Logistic Regression uses a curved line to summarize our observed data points

The logistic regression line generates probabilities **between 0 and 1**.

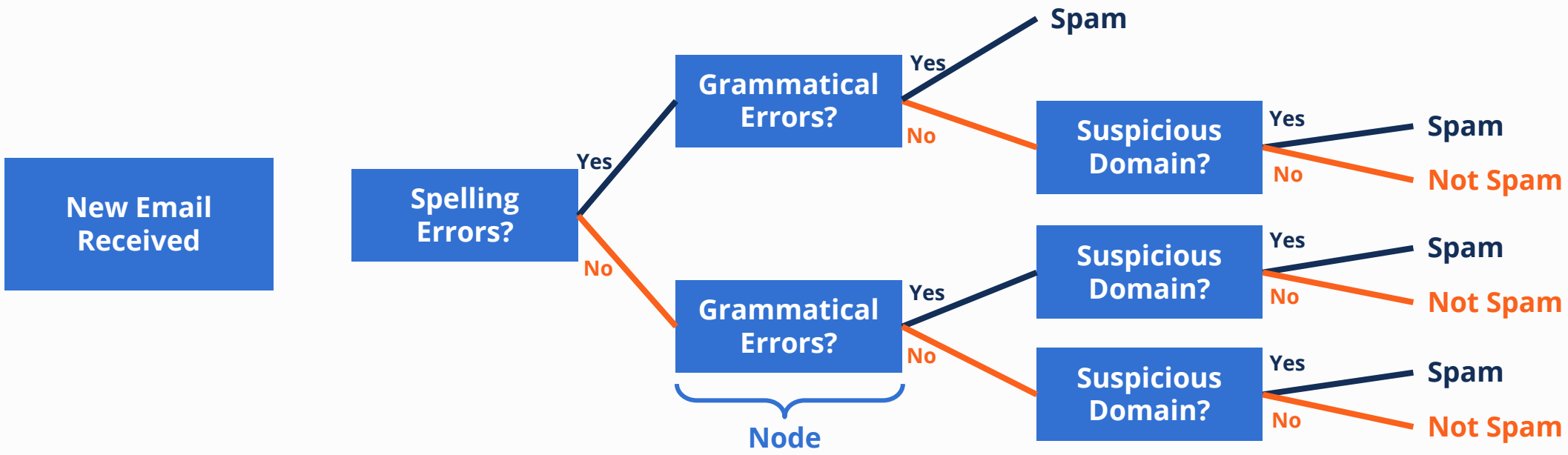
Decision Tree

The **decision tree algorithm** can be used to **predict both categorical or numeric outcomes**.

Use Case
Email Spam Detection

Use Case	Output Classes
Email Spam Detection	**Spam** OR **Not Spam**

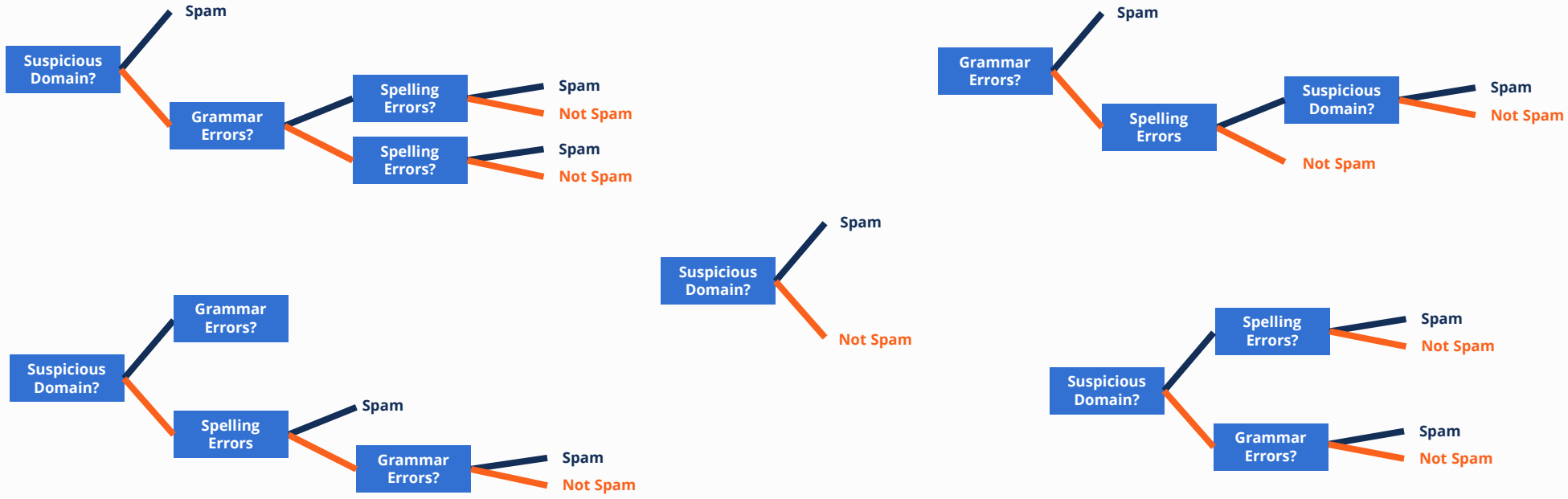
Output Classes
Spam OR **Not Spam**



Decision Tree

We can change a number of model parameters in order to optimize model performance.

Parameters include: **Number of nodes, minimum group size,**

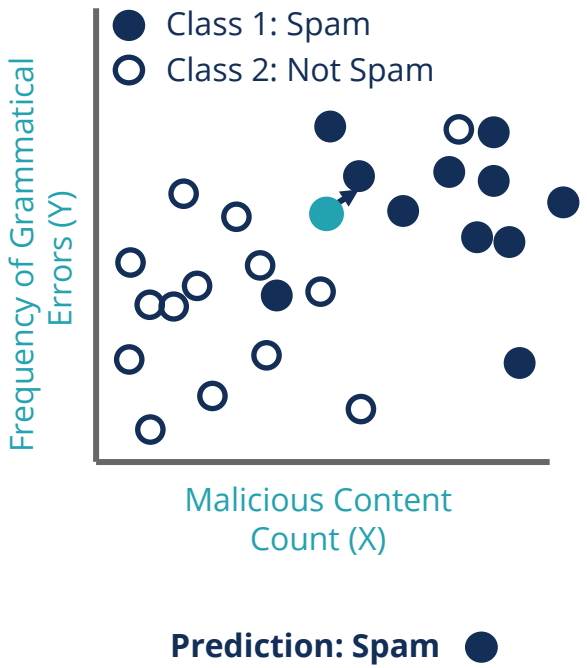


We choose the model which best separates the two classes, in this case **Spam** and **Not Spam**.

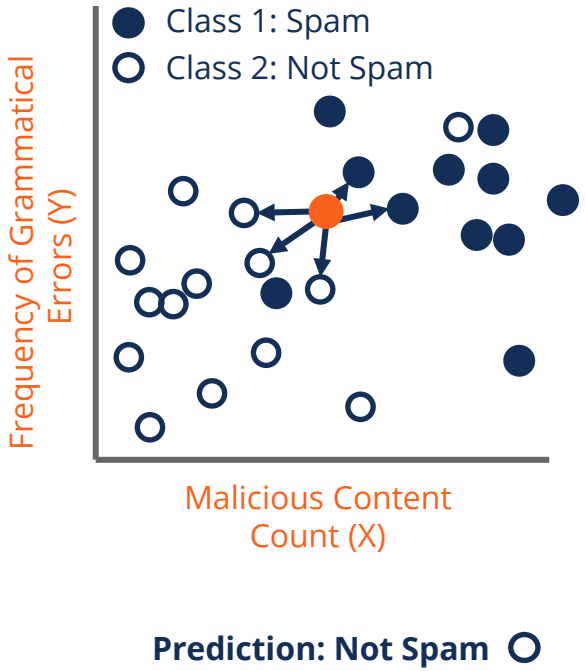
K-Nearest Neighbours

KNN assigns output classes based on the most similar observations in our sample space.

1 Nearest Neighbour



5 Nearest Neighbours

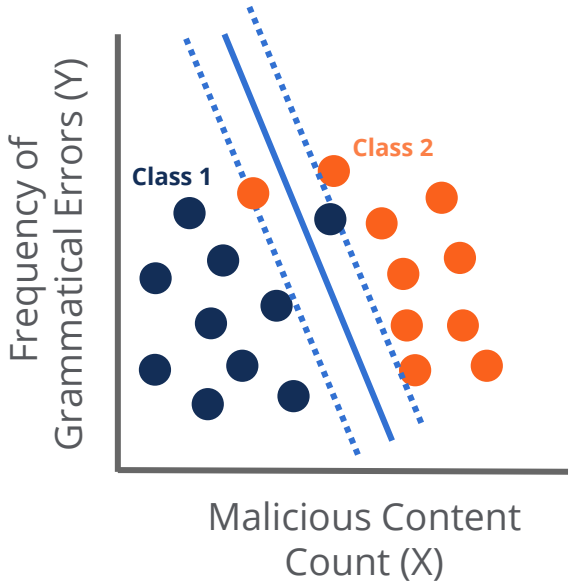


- The **lower the number K**, the **more specific** our model becomes to this dataset, potentially **overfitting**.
- The **higher the number K**, the **more generalized** the model becomes, potentially **underfitting**.
- We need to find a **good balance**, to give us good performance in a sensible sample space.

Support Vector Machines

SVM models try to **maximize the separation or margin** between classes in the sample space.

Support Vector Classifier with Two Permitted Outliers



Support Vector Machines are an extension of **Support Vector Classifiers**.

Support Vector Classifiers

- Maximize the margin between classes using a **linear decision boundary**.
- By **allowing outliers**, we make the model **less sensitive to the training data**.

Support Vector Machines

- SVMs are more advanced as they allow us to model **non-linear decision boundaries**.
- SVMs use Kernel Functions to **hypothetically transform the data**, which simplifies the math.

Naïve Bayes

Naïve Bayes is a probabilistic model based on **Bayes theorem** which calculates conditional probabilities.

Conditional probability is the probability of one event, given the probabilities of other events.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where A is the ***hypothesis*** or ***outcome variable,*** and B is the ***evidence*** *or* ***features***

Did you know? The Naïve refers to the models assumption that all input variables are independent.

For example, when we observe a phrase in an potential SPAM email, we might ask:

Known or expected **probability of SPAM** emails

Likelihood of observing PHRASE in a SPAM email

Likelihood of observing PHRASE in a NOT SPAM email

Known or expected **probability of NOT SPAM** emails

We can **manipulate some values, known as priors** if we have **additional knowledge**.

Gaussian Naïve Bayes

The Gaussian Naïve Bayes is an extension of Naïve Bayes, and is used to model **normally distributed variables**.

Showing Likelihood of a Twitter Post going Viral

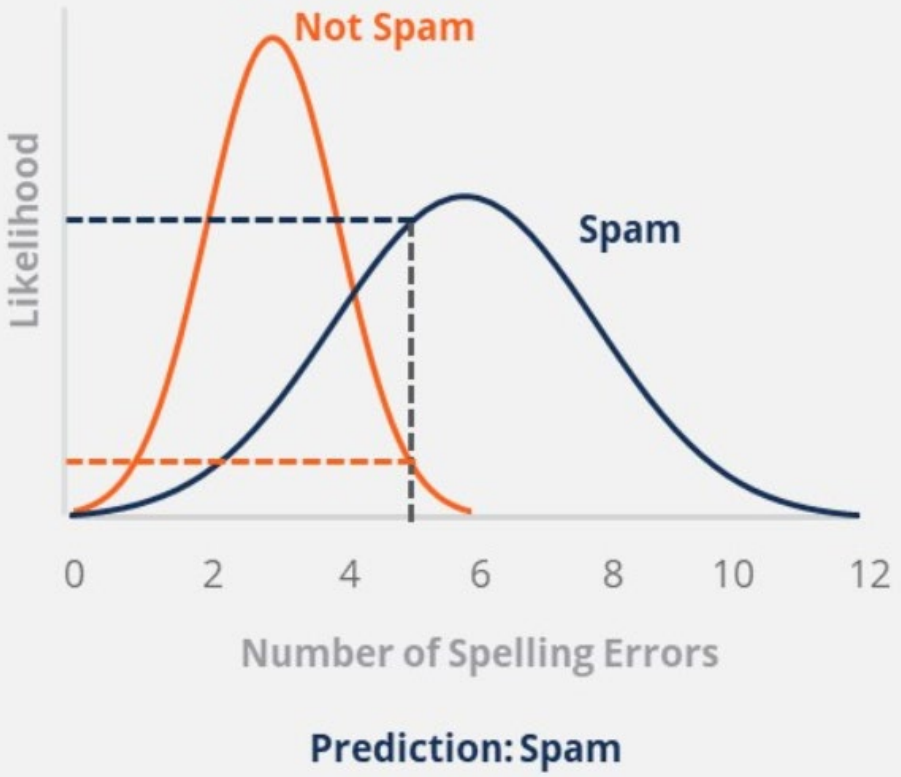
We plot our sample data and observe two very different but overlapping distributions.

Spams emails tend to have a **higher number of spelling errors**.

Not Spam emails tend to have a **lower number of spelling errors**.

For a **new email, with 5 spelling errors**:

Is it more likely that the email has 5 spelling errors if it came from the **Spam Distribution** or the **Not Spam Distribution**.

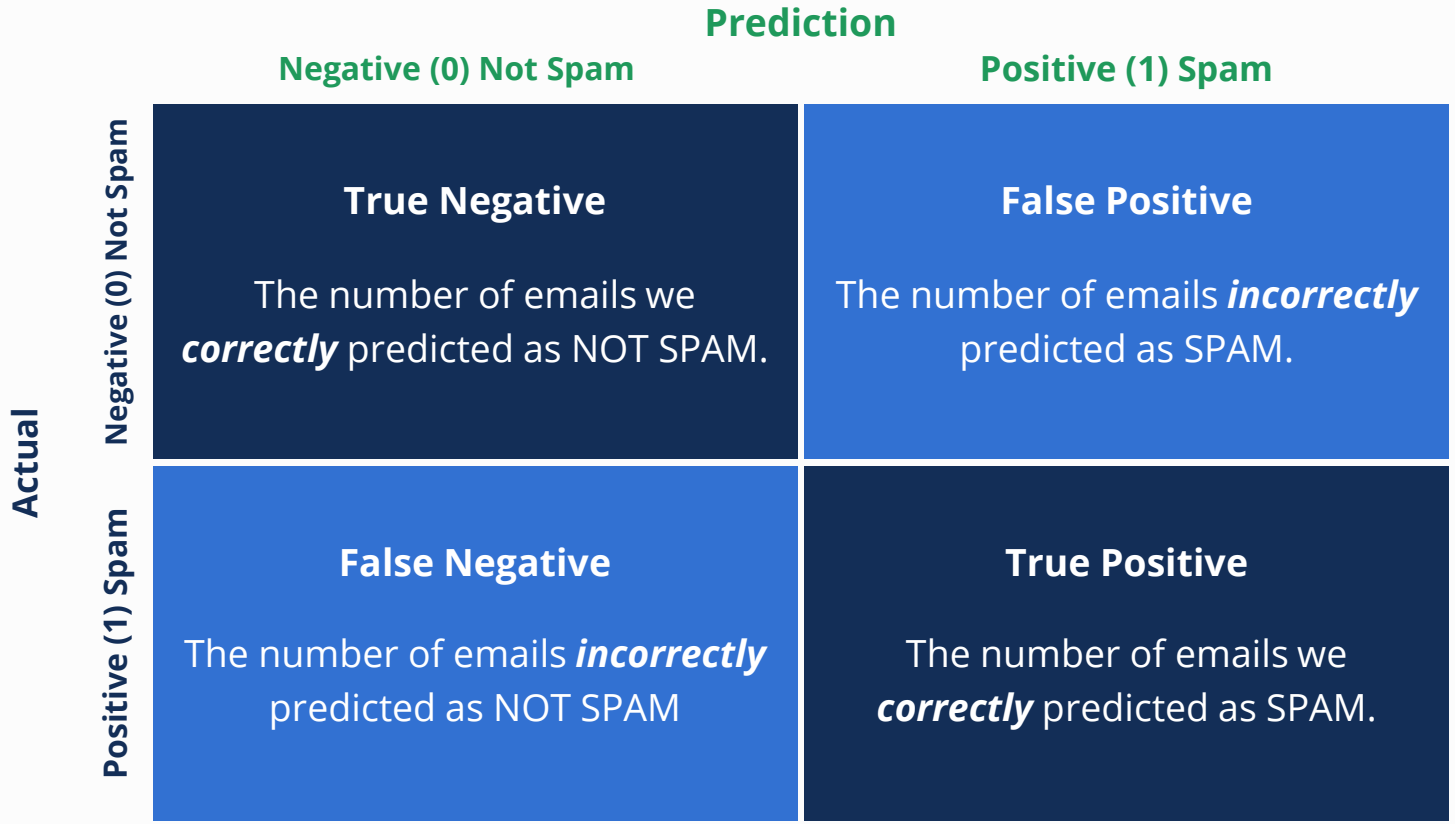


The Gaussian Naïve Bayes **also captures prior probabilities**, allowing us to capture additional information.

Confusion Matrix

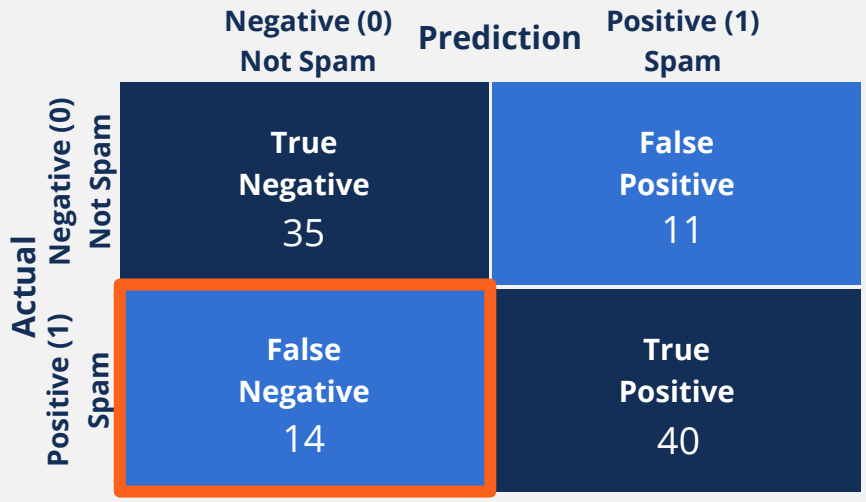
The confusion matrix helps us compare the **predictions we control**, vs the **actual outcomes that we don't**.

It can help us understand the **quality of our predictions**, or the **trade-offs** we must make.



Confusion Matrix

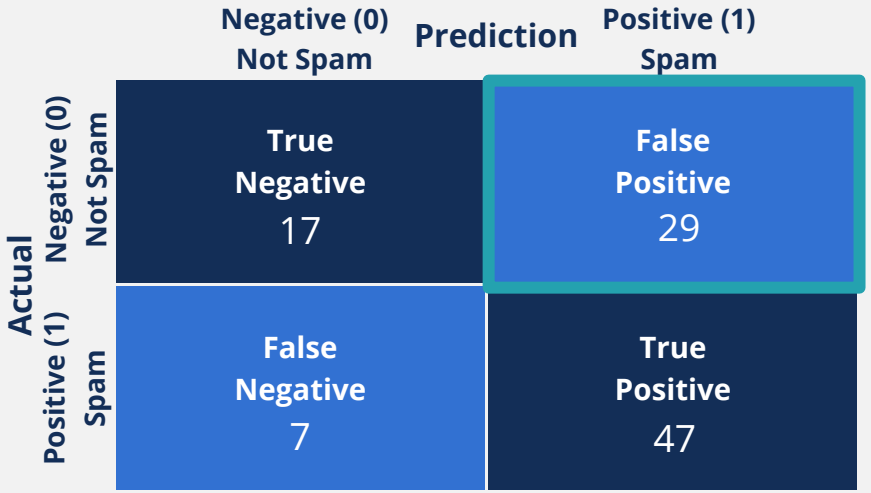
Suppose we predict roughly 50% of emails are SPAM, based on several input variables..



- Overall, **75 (40 + 35)** out of 100 predictions were correct.
- But there are **14 Spam Emails that we didn't detect**.

What if we want to increase the number of actual Spam emails that we detect? We previously missed 14 of them!

Now, roughly 70% of emails are marked as SPAM.



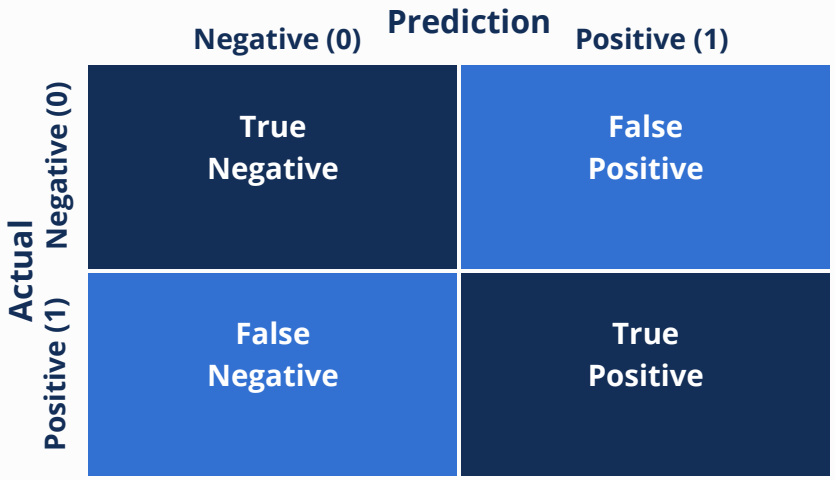
- There are now only 7 missed SPAM emails! Success!
- But now we created **18 more false alarms**. This may get annoying for users as they have to search in their junk.

Our model cannot be perfect, and the confusion matrix helps us **understand the trade offs**.

Understanding Trade Offs

The confusion matrix helps us understand trade offs.

But how do we decide which outcome to favor?



It depends entirely on the situation!

False Negatives are undesirable in disease detection. We cannot afford to miss bad outcomes.

False Negatives are undesirable in Fire Alarms. But if we flag too many, will people start to ignore them?

False Positives are undesirable in recommender systems. Imagine if Netflix constantly suggests shows you weren't interested in.

We must be clear on **what we want to achieve**, and the **costs and benefits of each type of error**.

Evaluation Metrics

There are **several metrics and techniques** that can help summarize the observations in the confusion matrix:

1. **Accuracy = (TN + TP) / Total Predictions**

   Describes what proportion of predictions were correct (may not always be the best indicator of performance).

2. **Precision = TP / (TP + FP)**

   How good are the positive predictions? Out of those predicted positive, how many were actually positive?

3. **Recall = TP / (TP + FN)**

   Describes what proportion of the actual positive cases were correctly identified.

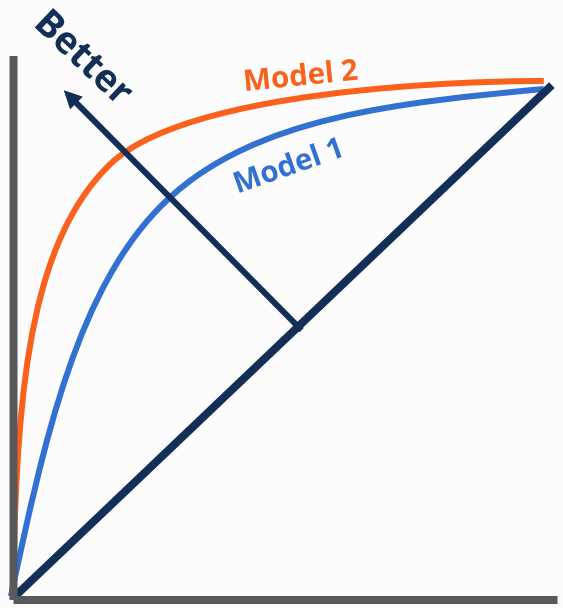
4. **F1 Score = 2 * [(Precision*Recall) / (Precision+Recall)]**

   Provides a balance between precision and recall.

The ROC Curve

Helps us visualize and compare the performance of models.

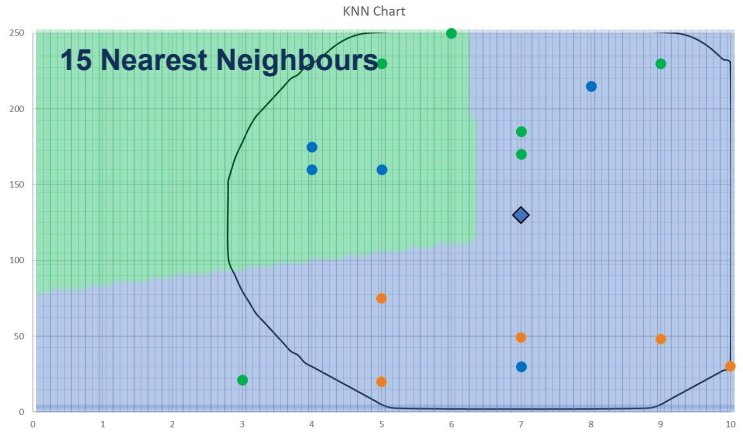
The model with the biggest lift is best.



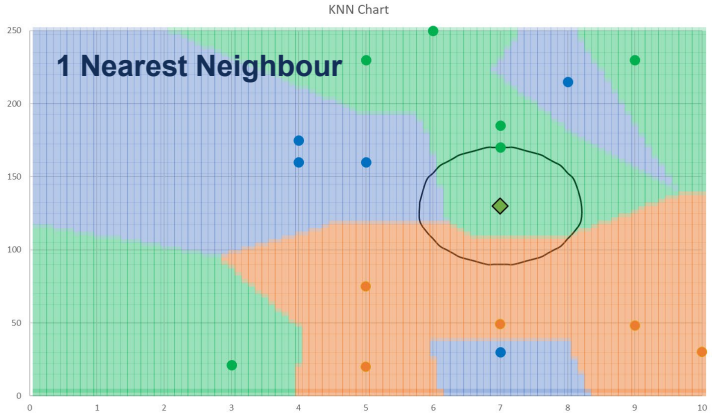
Overfitting Vs Underfitting

Underfitting and Overfitting can help us describe classification model outputs too.

Underfitting means the model **under generalises the data**



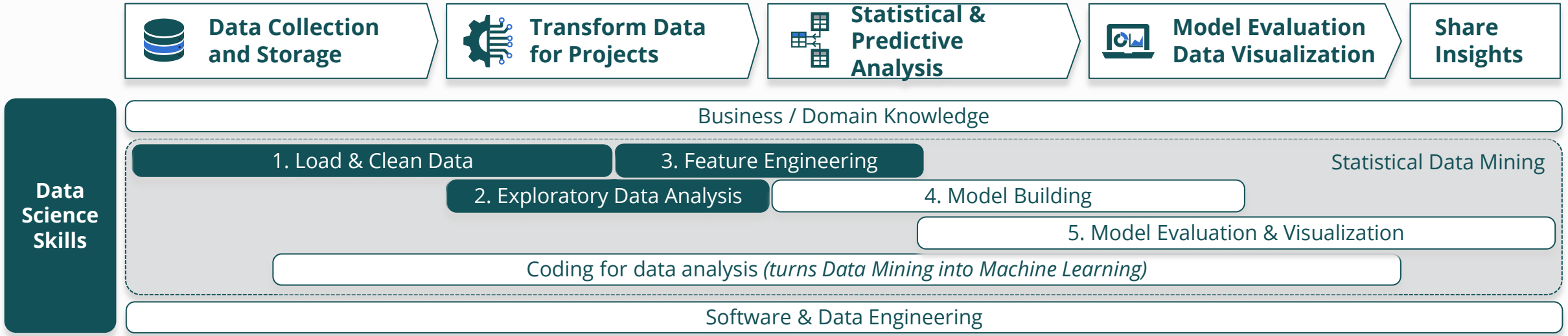
Overfitting means the model learns the training data too well, and misses the more general relationship.



We must **find a balance** to ensure our model **performs according to our evaluation metrics**.

Data Preparation

Data Preparation



Data preparation happens before model building.

We can generalize data prep into three main areas:

1. Load & Clean Data
2. Exploratory Data Analysis
3. Feature Engineering

Basic Dataset Terminology

A few key terms to get you started...

Application ID	Age (18-90)	Credit Rating	Income	Credit Approved
1	25	697	25,000	YES
2	15	527	13,000	NO
3	19	658	23,000	YES
4	65	738	49,000	YES
5	72	538	32,000	NO
6	26	243	9,000	NO
7	186	999	25,000	NO

ROW / OBSERVATION

Feature: Used as inputs to calculations in models or machine learning algorithms.

Target: The variable of interest, that we are trying to predict, estimate or model.

Unique ID: Uniquely identifies each row.

Row: Each row represents a **single observation.**

Data Cleansing – Identifying Errors

Data Cleansing refers to the process of identifying and **dealing with errors or inconsistencies** in our data.

This may include:

Incorrect

Missing

Duplicated

Irrelevant

Errors

The **solutions for each type of error are similar**.

- Remove the feature
- Remove the observation
- Estimate the correct value

Data Cleansing

Incorrect Data

Incorrect data **may be obvious**, or it **may be more ambiguous**.

It is important that we investigate all errors to understand why they occur and how they should or should not affect our analysis.

Applicant dataset for a $1,000 credit card…

Application ID	Age (18-90)	Credit Rating	Income	Credit Approved
1	25	697	25,000	YES
2	15	527	13,000	NO
3	19	658	23,000	YES
4	65	738	49,000	YES
5	72	538	32,000	NO
6	26	243	9,000	NO
7	186	999	25,000	NO

Some errors are **clearly incorrect**, like a customer age of 186.

Some errors are **suspicious**, like a credit score of 999.

Some errors are **ambiguous**, like an applicant age of 15.

Data Cleansing

Missing Data

Missing values **raise many questions** about our data.

Again, **we need to understand the process** through which the data was collected.

Commercial Real Estate Mortgage Data

Borrower ID	Approved Loan Amount	Credit Outstanding	Annual Repayment	Remaining Term (years)
672	10,000,000	NA	700,000	15
72	150,000,000	20,000,000	2,100,000	10
34	3,000,000	1,000,000	200,000	--
41	1,500,000	2,000,000	130,000	12
35	2,500,000	1,000,000	120,000	8
12	600,000	Null	35,000	18
726	700,000	100,000	15,000	7

We must understand **what each missing value type represents**.

We may need to speak to our **data engineer**, or another relevant stakeholder.

Data Cleansing

Duplicated Data

Duplicated data can cause bias in our analysis.

We should **investigate the reason** for the duplicated data.

CSV data...

Row ID	Month	Year	Sales	Costs
1	Jan	2020	150	100
2	Feb	2020	200	120
3	Mar	2020	220	130
1	Jan	2020	150	100
2	Feb	2020	200	120

Two out of three rows **seem to be duplicated.**

Data Cleansing

Irrelevant Data

Incorrect data is simply not useful in the scenario.

Irrelevant data **does not help us make predictions.**

Home Insurance Application Data

Application ID	House Value	Fire Risk	Phone Number	Address
1	300,000	5	604-121-22A	**3424** WR5
2	220,000	3	604-422-22B	**32** NS6
3	250,000	1	604-783-22C	**4532** BR68
4	180,000	0	604-123-22D	**3838** GV38
5	210,000	0	604-993-22E	**1010** P98
6	125,000	3	604-192-22F	**23** RR21
7	175,000	1	604-193-22G	**38** WR38

Sometimes, an **entire column can be removed**.

Sometimes, relevant data is **combined** with irrelevant data.

Data Cleansing

Errors

Errors might be caused by **any number of reasons.**

We should be aware of them and understand **how and why they occur.**

Birth Certificate Registrations

Application ID	Date of Birth	Country	Weight	
1	23/08/2081	ENGLAND	8lbs	
2	24/11/1999	CANADA	9lbs	
3	23/01/2014	+1	8lbs	
4	15/04/1987	FRANCE	7lbs	
5	28/02/1994	SLOVAKIA	6lbs	
6	07/05/1996	IRAN	ERR##	
7	18/11/2005	NIGERIA	8lbs	

Typos, incorrect data types and failed calculations all lead to errors.

Data Cleansing – Solutions

House Fire Risk Survey

ID	Age	Suburb	House Material	Cooking Fuel
1	25	Forestville	Wood	Gas
2	Null	Parksville	Brick	Gas
3	43	Forestville	Wood	Gas
4	75	Parksville	Wood	Gas
5	37	Central	Thatch	Gas
6	Null	Parksville	Wood	Elec
7	Null	Parksville	Wood	Gas
8	65	Central	Brick	Elec
9	0	ERR	Null	NA
10	43	Central	Wood	Gas
…	…	…	…	…

Some rows can **simply be removed**.

Imputation can help us fill in missing values.

Using **average** age is very simplistic!

Summary of age by suburb

Suburb	Avg Dataset Age	Avg Census Age
Central	29	28
Forestville	37	35
Parksville	58	61

The average **age in each suburb is very different**.

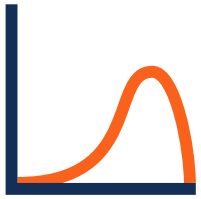
We are already in the realms of **prediction and estimation**.

Exploratory Data Analysis (EDA)

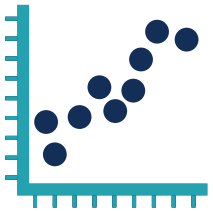
EDA helps us gain **initial insights** from our data, that **help us implement our model**.

What types of data are present?

How can we describe the data in each feature?



Are there any obvious relationships?



EDA – Data Types

Understanding the **data types** helps us **understand limitations and challenges** we may encounter.

Continuous

Continuous features allow us to measure **amounts** or **points along a scale**.

Can be measured on a scale or timeline.

1.7

Might include numbers, or a timeline of dates.

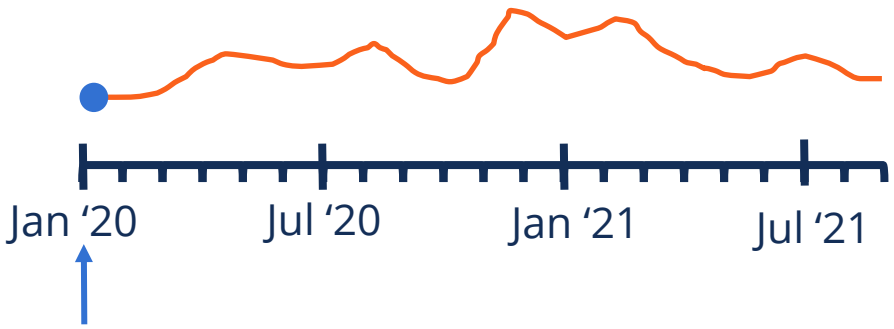
Continuous variables are the easiest to work with.

Continuous numeric

- Speed
- Temperature
- Income
- Age

- Dates

Continuous dates



EDA – Data Types

Understanding the **data types** helps us **understand limitations and challenges** we may encounter.

Categorical

Categorical features tell us **which bucket** a data point falls into.

Unordered Categorical Features

- Unemployed
- Part Time
- Full Time
- Student

No specified order

Ordinal Categorical Features

1. Small
2. Medium
3. Large
4. X Large

A logical order exists

Binary Features

1. True
0. False

Two buckets

EDA – Data Type Exceptions

Some data type examples can be confusing...

Numbers are **not always continuous**

How many customers visited the store?

How many transactions were made?

1 **326** **4,237**

We cannot have 0.5 of a customer, so the scale is not strictly continuous.

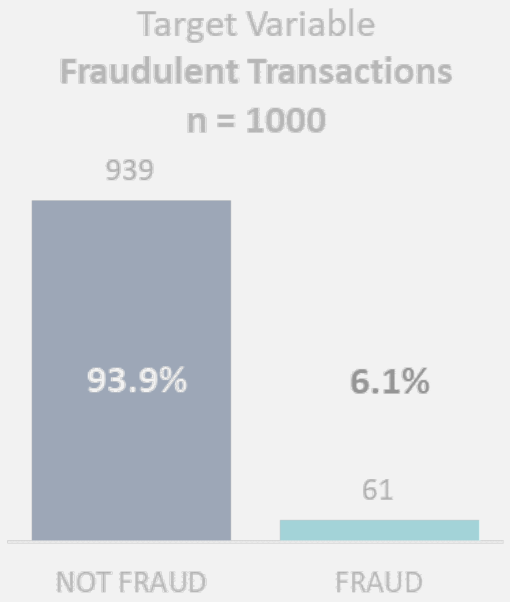
Dates are **not always continuous**

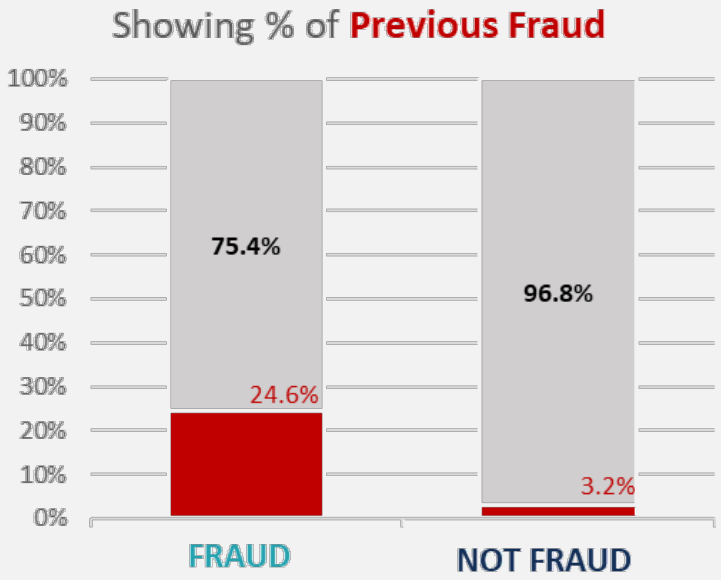
2021 **2022** **2023** **2024**

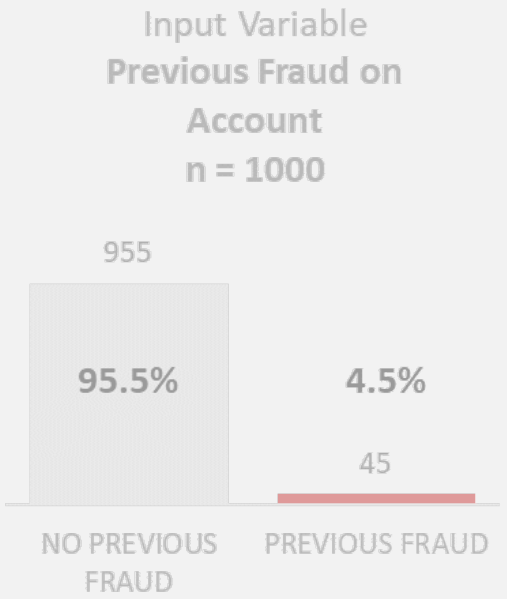
When datapoints **belong in buckets**, they are considered **categorical data**.

EDA – Binary Descriptive Stats

We can use histograms to view the **distribution of our target variable**.







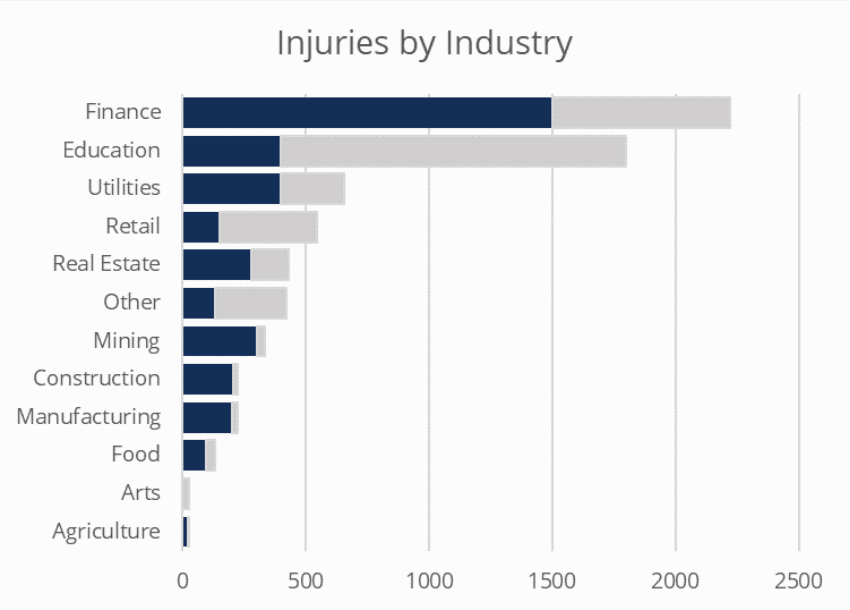
We have **very few FRAUDULENT** transactions. Our sample of data is imbalanced.

Previous Fraud may be a **helpful predictor of fraud**

EDA – Categorical Descriptive Stats

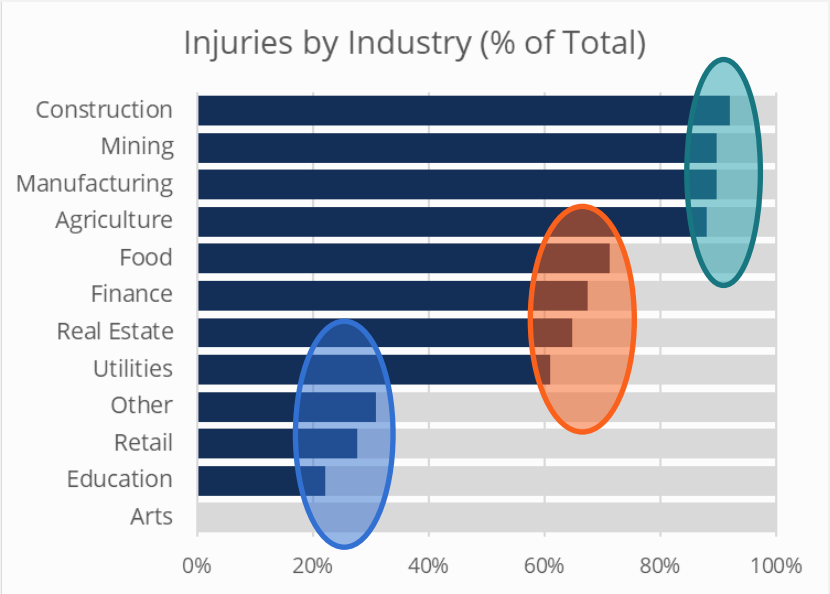
How many categories there are and **how frequently** do they appear?

Predicting Injuries for Insurance



INJURY AT WORK | **NO INJURY**

A large amount of our data is from finance and education

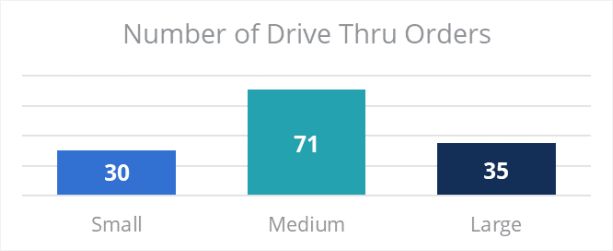


Industry may be a helpful predictor for injuries at work

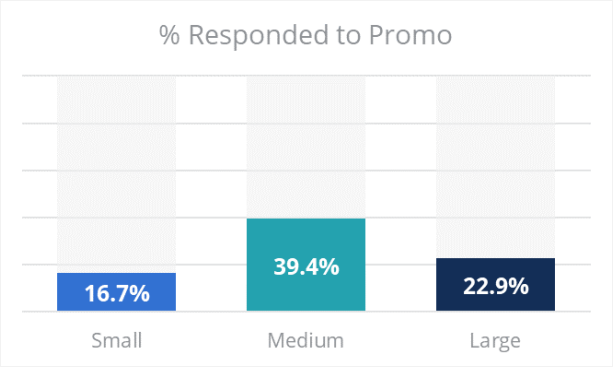
EDA – Descriptive Stats Part 1

Ordered categorical variables are still considered buckets, but they **have a sense of order to them**.

Drive Through Dataset



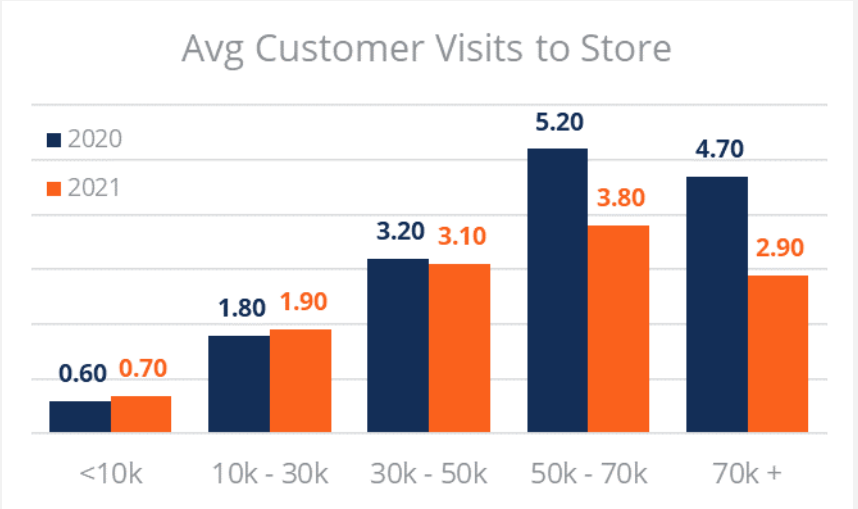
Which group will respond better to health focused promos?



EDA often uncovers patterns we didn't expect to see.

Does this align with **industry expectations?**

Gaming Dataset



Clearly higher income individuals **visit the store more often**.

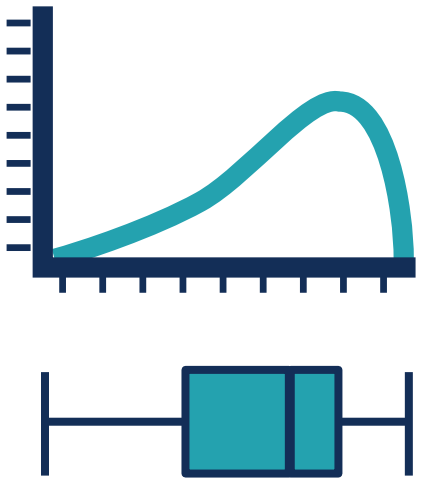
What could **explain the dip** in high income visits in 2021?

Exploratory data analysis **helps us be curious**, ask **questions** and **uncover patterns** in our data.

EDA – Descriptive Stats Part 1

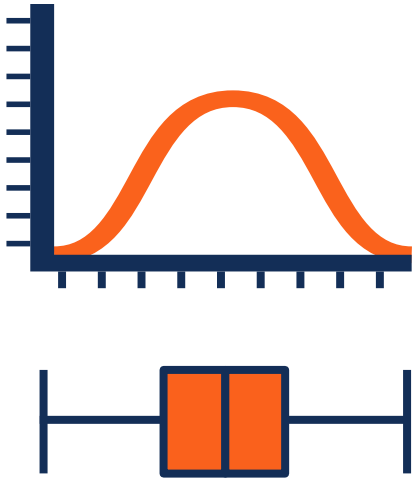
We have a variety of ways to plot and describe continuous variables.

Left Skewed



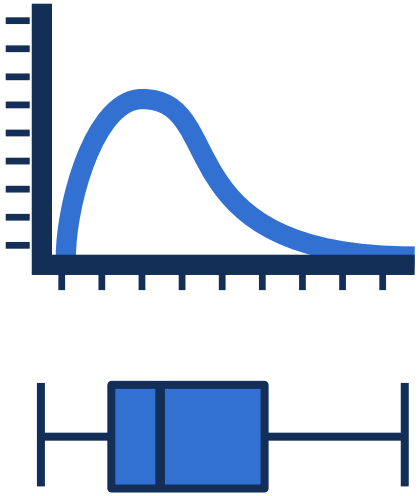
Mean
Standard Deviation
Median
IQ Range

Normal



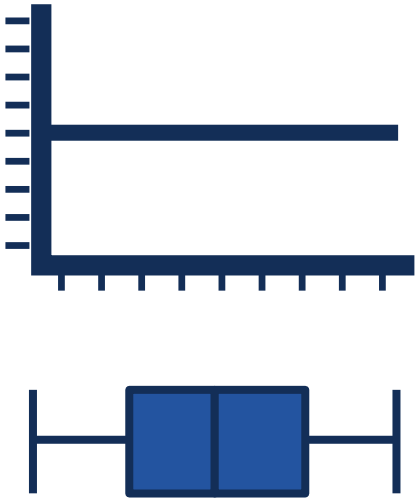
Mean
Standard Deviation
Median
IQ Range

Right Skewed



Mean
Standard Deviation
Median
IQ Range

Uniform

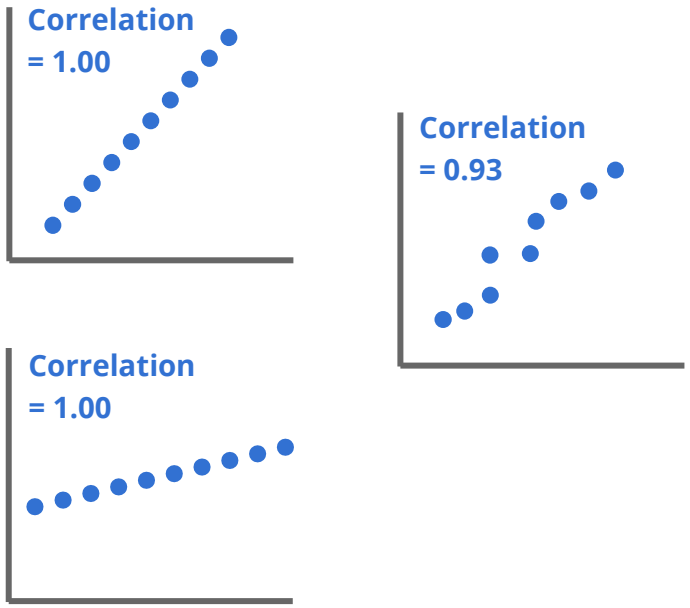


Mean
Standard Deviation
Median
IQ Range

EDA - Correlation

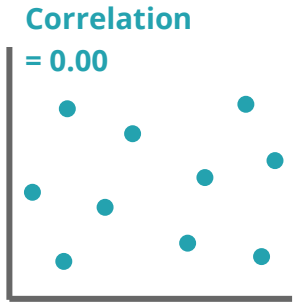
Correlation describes the extent to which one variable moves with another.

Positive Correlations tell us that as one variable increases, the other tends to also.

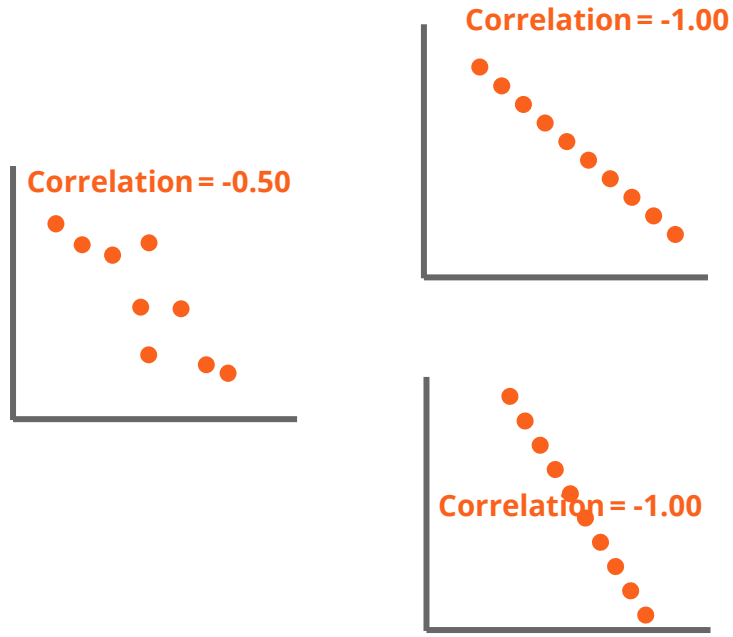


Correlation can have a **max value of 1.**

Zero Correlation tells us that one variable has no impact on the other.



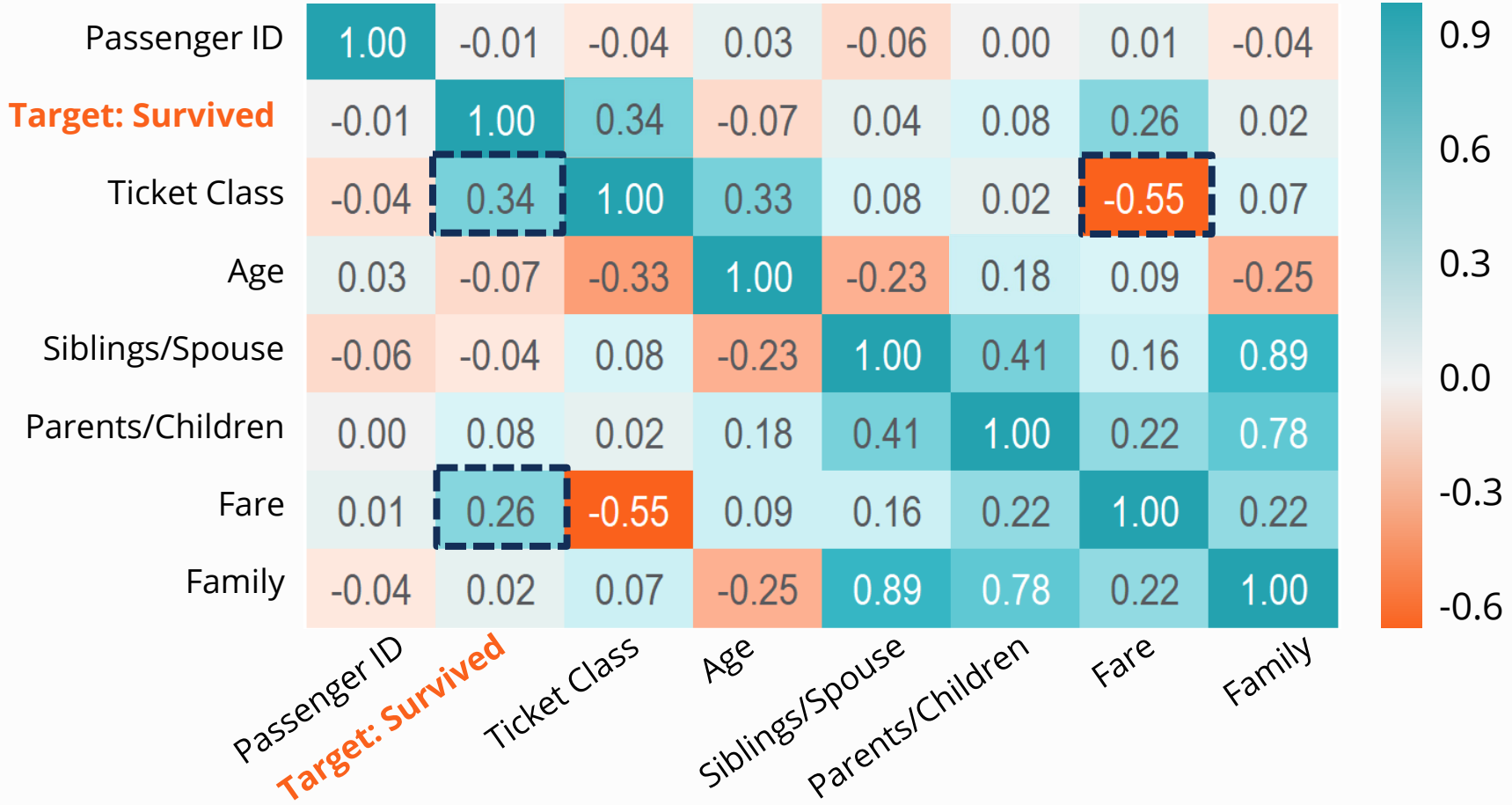
Negative Correlations tell us that as one variable increases, the other decreases.



Correlation can have a **min value of -1.**

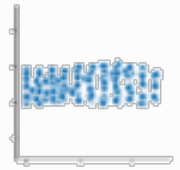
EDA – Correlation Matrix

A **correlation matrix** can be used to evaluate the **relationship between each pair** of variables in the dataset.

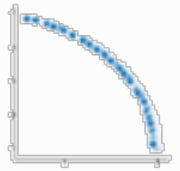


EDA –Scatter Plot Matrix

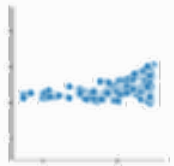
A **scatter plot matrix** can be used to visually inspect patterns and relationships between variables.



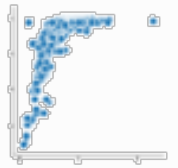
No relationship



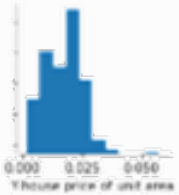
Clear non-linear relationship



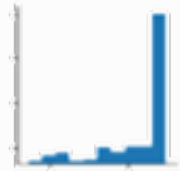
Changing variance



Exponential relationship



Right Skew Distribution



Imbalanced Categories

EDA - Feature Selection

Feature selection: select the related features from the dataset and remove the irrelevant ones.

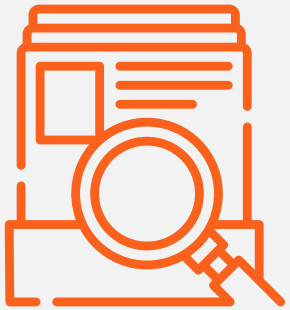
Having **too many features** in a model can result in **overfitting**.

Company Valuation

5000 companies

50 Features (Financial Ratios)

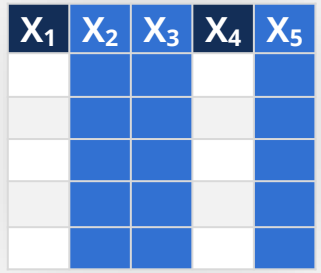
Feature Selection



5000 companies

10 Key Features

In summary, we are **eliminating some columns (features)** from our dataset.



Benefits

- **Reduce processing time**
- **Improve analysis results**

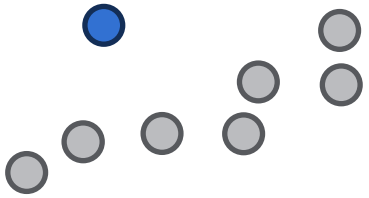
Common Feature Selection methods are **Principal Component Analysis** and **Feature Importance**.

Feature Engineering

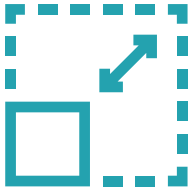
Feature engineering is the process of **modifying the structure or contents of our data** to make it **more suitable for analysis**, or to help **improve the performance** of a model.

Imputation

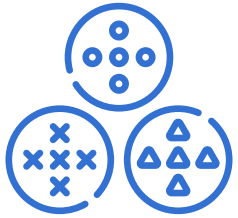
Dealing with Outliers



Scaling



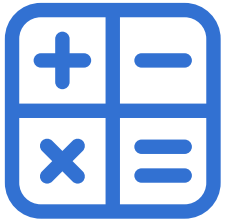
Grouping / Binning



One Hot Encoding

Category
0
1

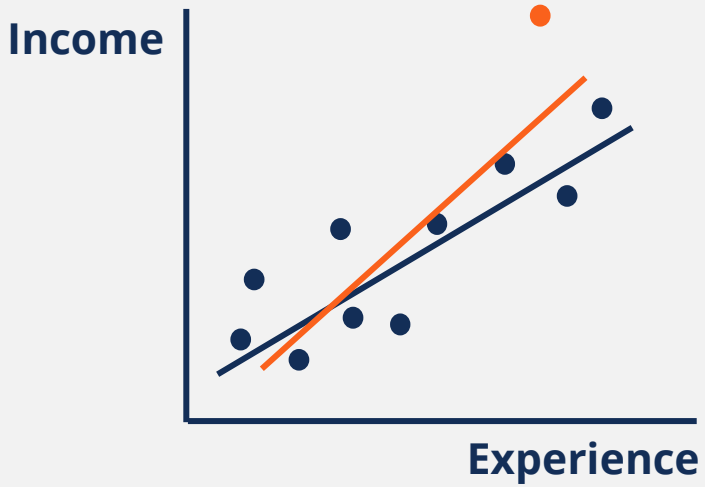
Calculation



Feature Engineering – Dealing with Outliers

Outliers can change the results of our analysis significantly, but there is **no single solution** that works in all scenarios.

An outlier or just a high value?



Potential Questions

- Is there **something we're not capturing** that might be generating these results, or is this truly a random outlier?
- **Beyond what threshold** do we consider an observation to be an outlier?
- **What impact will outliers have** on the particular model we are using?
- Should I change my **analysis method** based on the known presence of outliers?
- Do we want to **include it in our analysis**?

Feature Engineering – Normalization

Features of **significantly different scale** can cause problems for our Machine Learning models.

Normalization (also Min Max scaling) is a form a scaling, which allows us to transform variables onto a **consistent numeric scale**.

Feature values are scaled so that they all sit **between 0 and 1**.

- **1 represents the maximum** value in each column
- **0 represents the minimum** value in each columns
- **0.5 represents halfway** between the two.

The **distribution of values** in each column **does not change**

Income	Credit Score	Age
$56,000	755	43
$38,000	682	22
$120,000	731	38
$65,000	595	54

Income	Credit Score	Age
0.2195	1.0000	0.4565
0.0000	0.5438	0.0000
1.0000	0.8500	0.3478
0.3293	0.0000	0.6957

Feature Engineering – Standardization

Standardization performs a similar role to normalization, rescaling values to a **standardized, comparable scale**.

Differences:

- Typically used for **features with a gaussian distribution.**
- **Rescales a normal distribution** so that it has a mean of 0 and standard deviation of 1.
- Scaled values can be **positive or negative**.
- **No bounds** on the values of the feature.

Feature Engineering – Grouping / Binning

Grouping or binning helps us **simplify our data** to make it more digestible, or **remove some unnecessary detail**.

High Cardinality

Features with **many unique values** are referred to as high cardinality**.**

High cardinality provides **lots of detail**, but results in a **small sample set per category**.

Solutions

It can help to reduce the number of categories:

- **Group** into districts, states or regions
- Keep only large groups and group **remaining observations as 'other'**

Zip Code
22261
23621
25612
23261
25211
22515
26612
22324
22726
25353

Feature Engineering – Binning Numeric Values

We can also apply grouping or **binning to numeric variables.**

Numbers recorded to a high degree of accuracy can lead to overfitting.

Groups help represent a **range of similar values**.

We can even **re-incode numeric values** to each group, **maintaining their ordered** characteristics.

Income	Income Group	Income Class
35,650	30-40k	3
36,230	30-40k	3
84,570	80-90k	8
45,328	40-50k	4
20,303	20-30k	2
150,320	100k+	10
26,330	20-30k	2
62,320	60-70k	6
48,321	40-50k	4
72,320	70-80k	7

Feature Engineering – One Hot Encoding

One Hot Encoding turns **categorical data into binary columns**.

Some ML models (Decision Trees) work well with text categories, but **most require numeric values**.

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Dummy Variable Encoding achieves a similar outcome, instead removing the final column.

Feature Engineering – Calculation

Calculations can help us **extract new information** or **summarize the data** we have available.

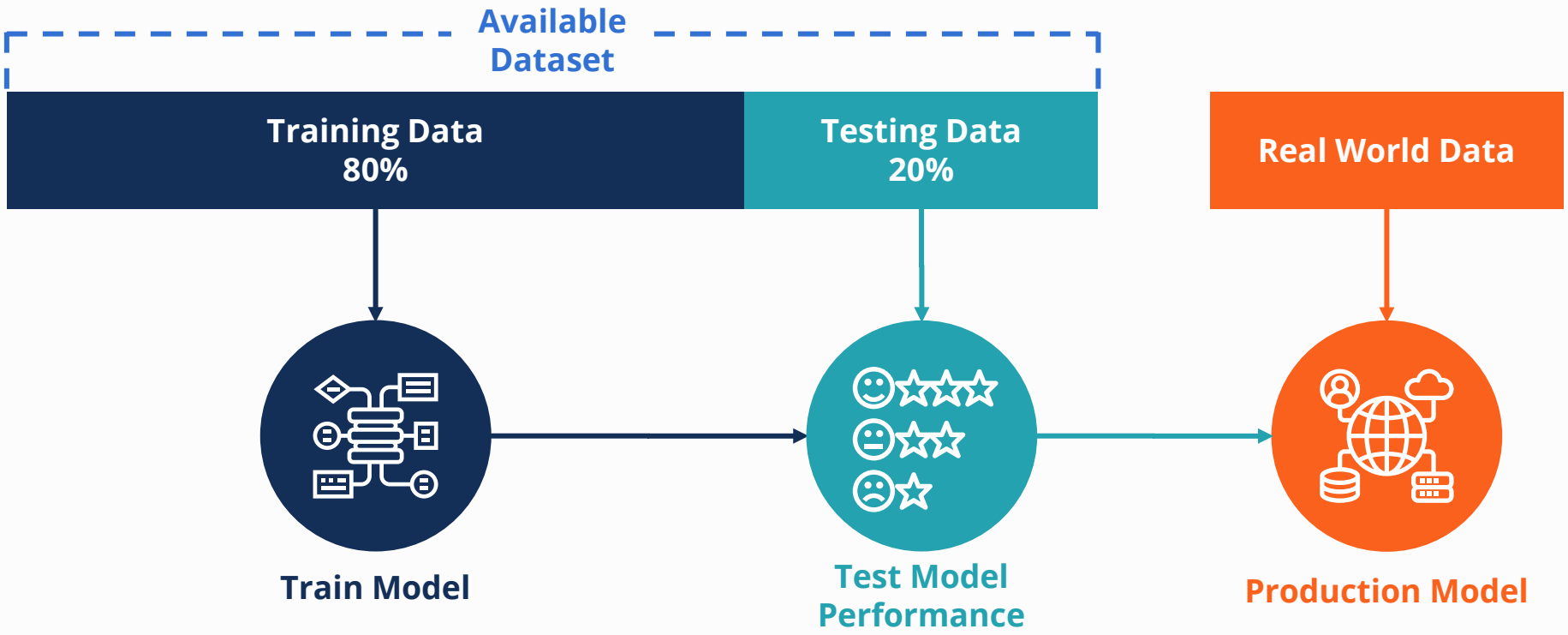
Start Date	First Performance Review	Review Wait (Months)
01 Apr 2020	01 Sep 2020	5
01 Jun 2020	15 Jun 2020	0.5
01 Sep 2020	01 Dec 2020	3
01 Dec 2020	01 Apr 2021	4
01 Feb 2021	01 May 2021	3

Part 1	Part 2	Part 3	Part 4	Avg
30%	20%	45%	50%	36%
80%	60%	70%	80%	73%
90%	100%	60%	85%	84%
40%	60%	40%	60%	50%
50%	55%	50%	80%	59%

The new features are **more suited to machine learning analysis**.

Training & Testing

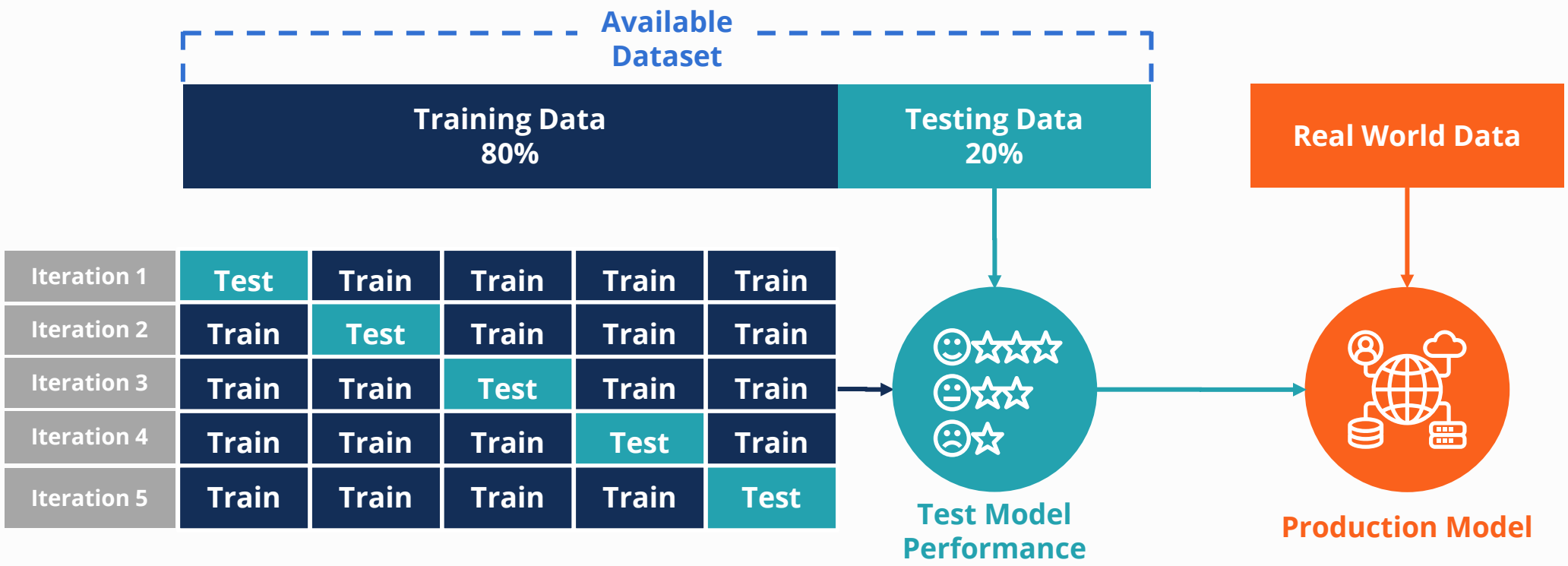
Models need to be tested before we use them to predict real-world outcomes.



Tests must be carried out on new data that the model has **never seen before**.

Training & Testing Variations

K-fold cross-validation goes one step further, by **splitting the training set into segments** to **add additional validation**.



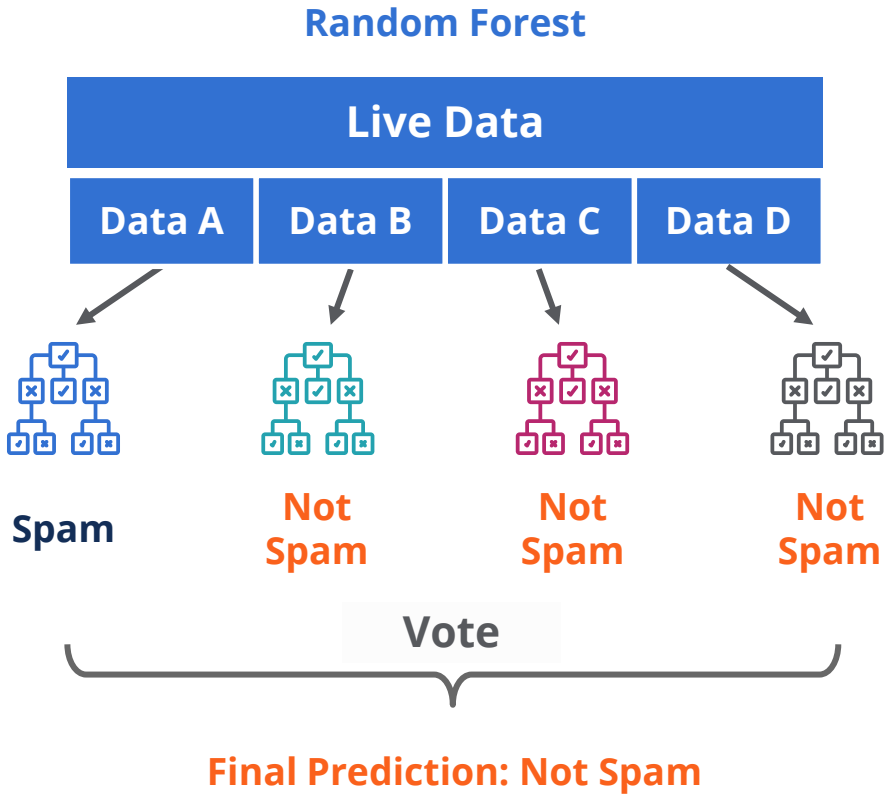
A further technique: **Training, Validation & Test** splits the **dataset into 3 segments**.

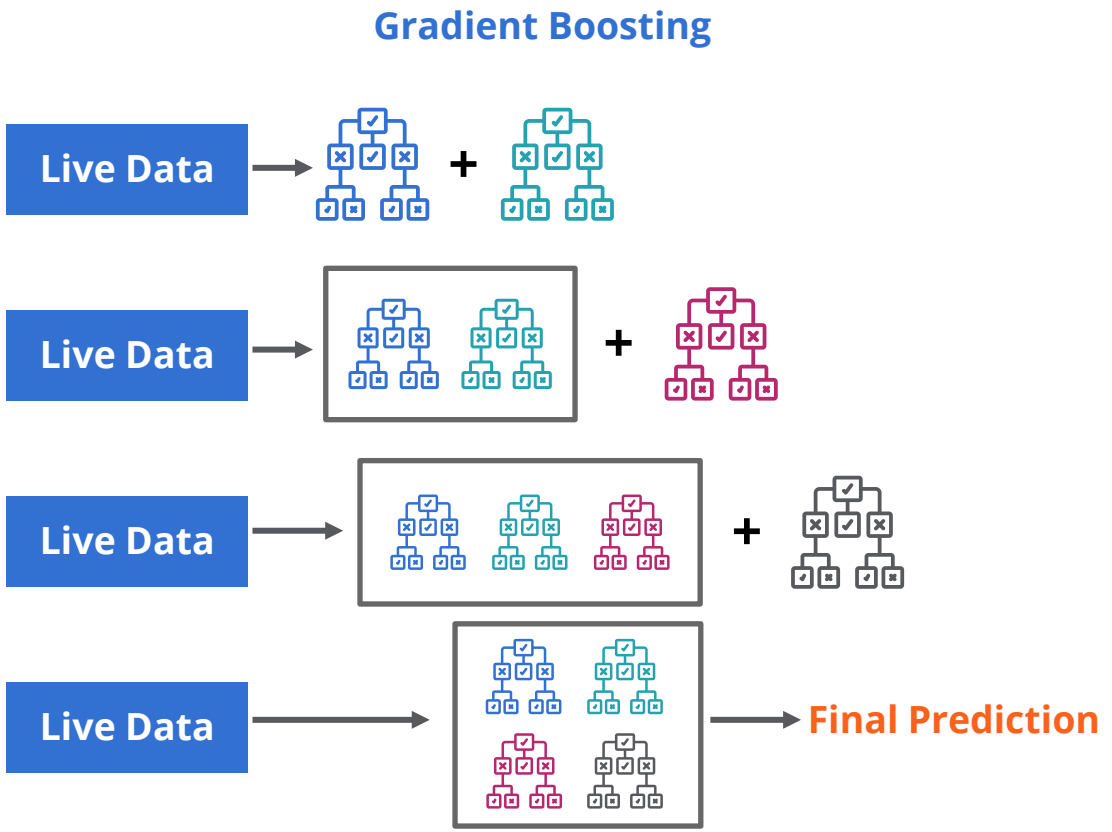
What do they all **have in common**? All techniques are trying to **validate results on new, unseen data**.

Other Data Science Techniques

Ensemble Models

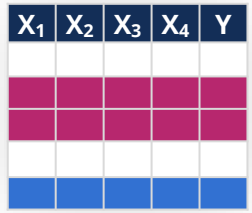
Empirically, ensemble models tend to add **~5% improved performance** over stand-alone machine learning models.





Ensemble models can be any combination of the machine learning algorithms.

Unsupervised Learning - Clustering



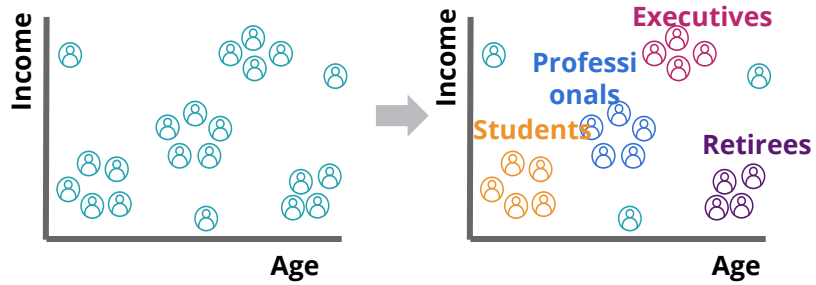
The purpose of clustering is to **group data points** into those with **similar characteristics**.

Importantly, we are **grouping observations (rows)** into clusters

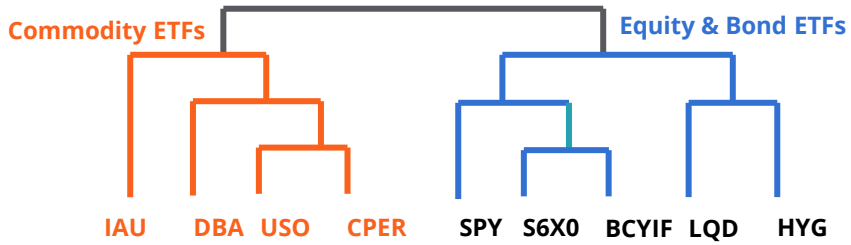
Input: Netflix Viewer Data

Algorithm

K Means Clustering



Hierarchical Clustering



Unsupervised Learning - Variable Reduction

Our earlier example: Company Valuation

5000 companies

50 Features (Financial Ratios)

Feature Selection

5000 companies

10 Key Features

Benefits

- **Reduce processing time**
- **Improve analysis results**

X_1	X_2	X_3	X_4	X_5

Variable Reduction algorithms are designed to reduce the number of features.

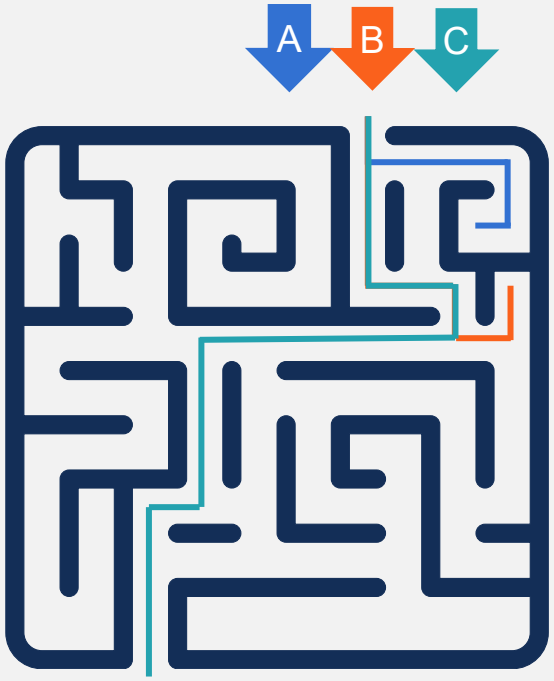
Importantly, we are identifying the **most important columns** in our dataset.

Common methods are **Principal Component Analysis** and **Feature Importance.**

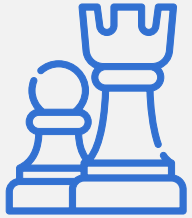
Reinforcement Learning

Reinforcement Learning is where machines **learn how to navigate scenarios through repetition.**

Simplistic Example: Completing a Maze



Complex Example: Playing Chess or Go



Clear rules which the bot (agent) must follow.

Clear outcomes act as a reward (win / lose / draw).

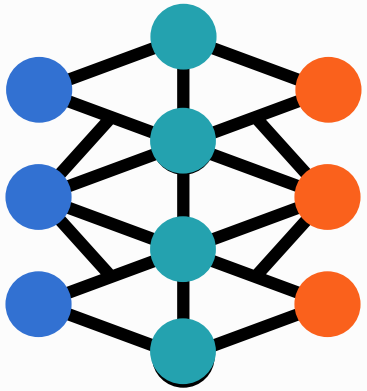
Unknowns including opponent or environment.

Why are computers well suited to Reinforcement Learning?

- **No memory loss**
- **Computers hold no recent memory biases**
- **Computational superiority**

Neural Networks & Deep Learning

Neural Networks are inspired by the structures of **neurons in our brains**. They consist of **nodes organized into layers.**



The **input layer** receives information or data.

The **output layer** is where a prediction is made.

The **hidden layers** are where all the complex math happens.

Deep Learning is an **extension of Neural Networks**, where the model may retrain itself multiple times.

Deep Learning models are **less reliant on humans** and may be able to train themselves.

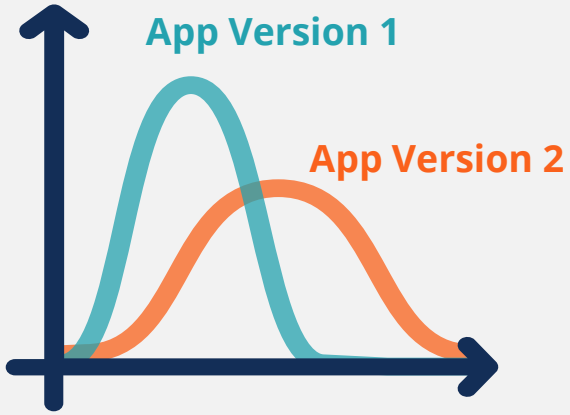
Applications of Neural Networks & Deep Learning

- **Collaborative Filtering**
- **Image Recognition**
- **Anomaly Detection**

Statistical Models

There are many types of other statistical model.

AB Testing a Crypto Trading App



App Version 1: Avg Transaction Value = $50

App Version 2: Avg Transaction Value = $70

Questions to Ask:

We must consider other business factors.

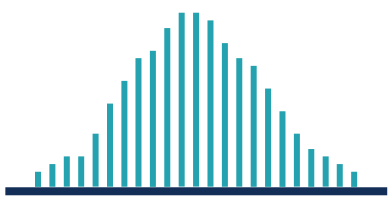
Was it a fair test?

Are there any unmeasured consequences?

Are the findings statistically significant?

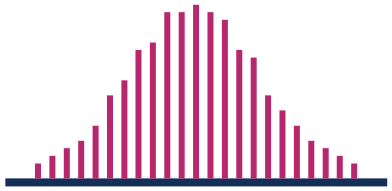
Monte Carlo Simulation

Monte Carlo Simulation is a statistical technique used to **quantify risk** in forecasting models.

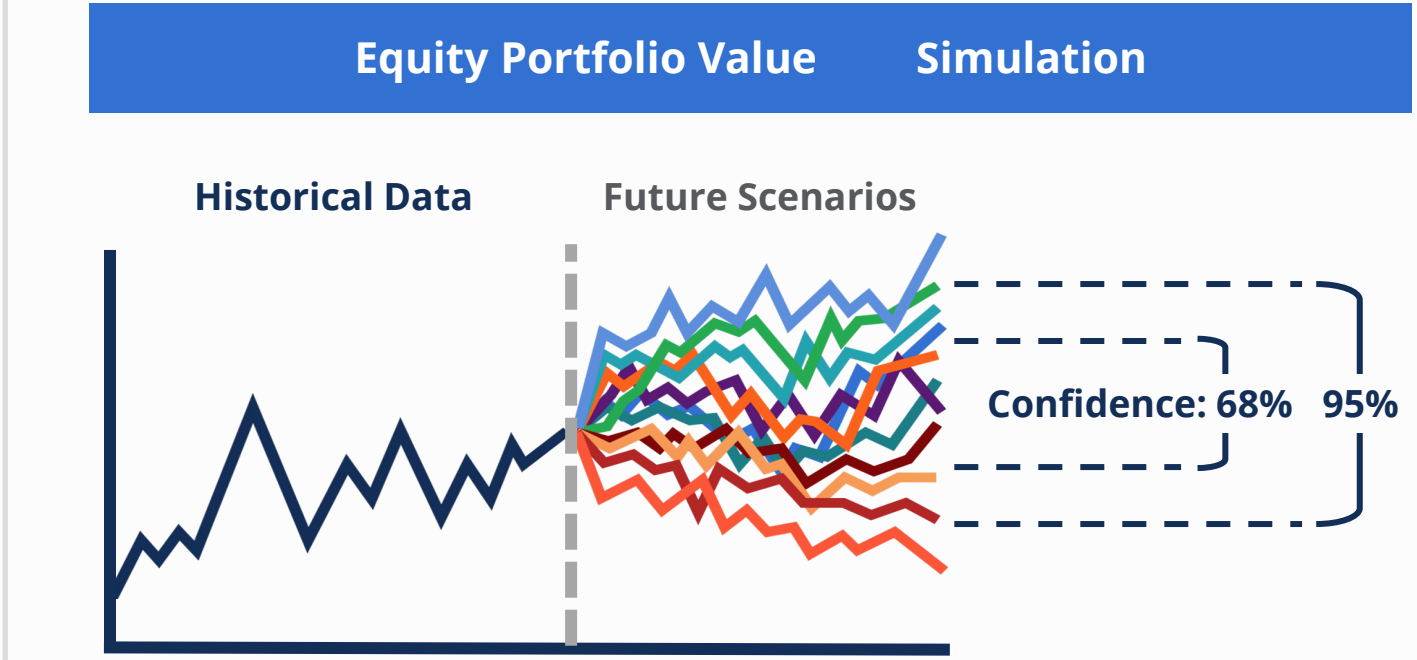


Past Events

Monte Carlo Simulation



Future Events



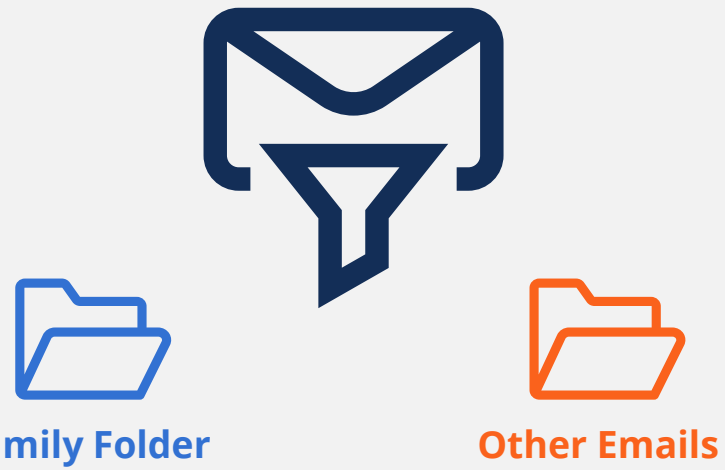
Mean of daily return: 0.13%

Std dev of daily return: 0.014%

Rule Based Models

Rule based models are used to **auto follow rules**, often at speeds that a **human would not be capable**.

Email Sorting



A basic rule-based model.

Algorithmic Trading

Human traders trade by the **minute**

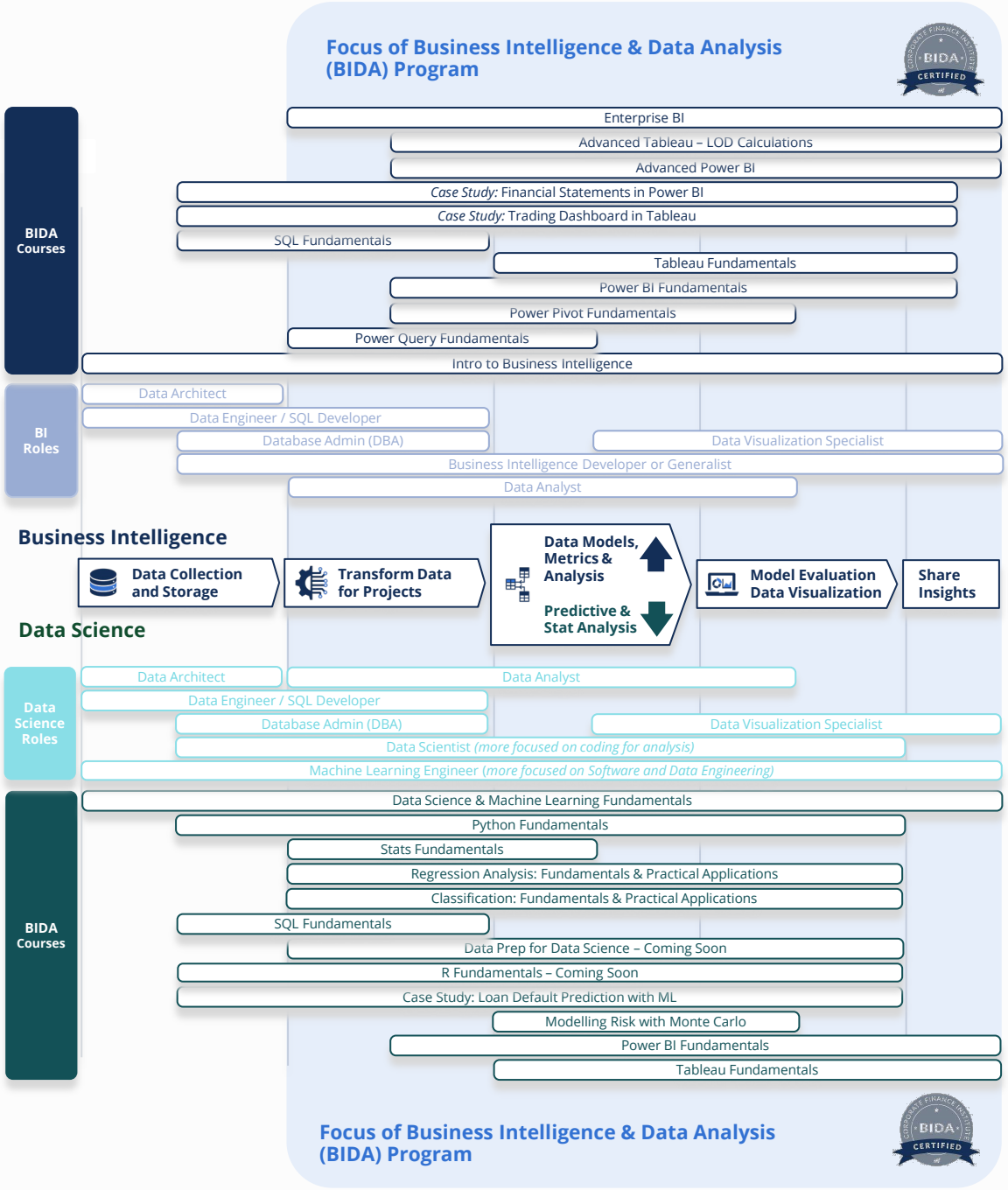
Algorithmic traders trade by the **millisecond**

Bots make trades when certain **conditions are met**.

The basic principles in both scenarios are the same: **Computers are simply following instructions.**

BIDA Syllabus

CFI


Focus of Business Intelligence & Data Analysis (BIDA) Program
BIDA Courses
Enterprise BI
Advanced Tableau – LOD Calculations
Advanced Power BI
Case Study: Financial Statements in Power BI
Case Study: Trading Dashboard in Tableau
SQL Fundamentals
Tableau Fundamentals
Power BI Fundamentals
Power Pivot Fundamentals
Power Query Fundamentals
Intro to Business Intelligence
BI Roles
Data Architect
Data Engineer / SQL Developer
Database Admin (DBA)
Data Visualization Specialist
Business Intelligence Developer or Generalist
Data Analyst
Business Intelligence
Data Collection and Storage
Transform Data for Projects
Data Models, Metrics & Analysis
Predictive & Stat Analysis
Model Evaluation Data Visualization
Share Insights
Data Science
Data Science Roles
Data Architect
Data Analyst
Data Engineer / SQL Developer
Database Admin (DBA)
Data Visualization Specialist
Data Scientist (more focused on coding for analysis)
Machine Learning Engineer (more focused on Software and Data Engineering)
BIDA Courses
Data Science & Machine Learning Fundamentals
Python Fundamentals
Stats Fundamentals
Regression Analysis: Fundamentals & Practical Applications
Classification: Fundamentals & Practical Applications
SQL Fundamentals
Data Prep for Data Science – Coming Soon
R Fundamentals – Coming Soon
Case Study: Loan Default Prediction with ML
Modelling Risk with Monte Carlo
Power BI Fundamentals
Tableau Fundamentals
Focus of Business Intelligence & Data Analysis (BIDA) Program