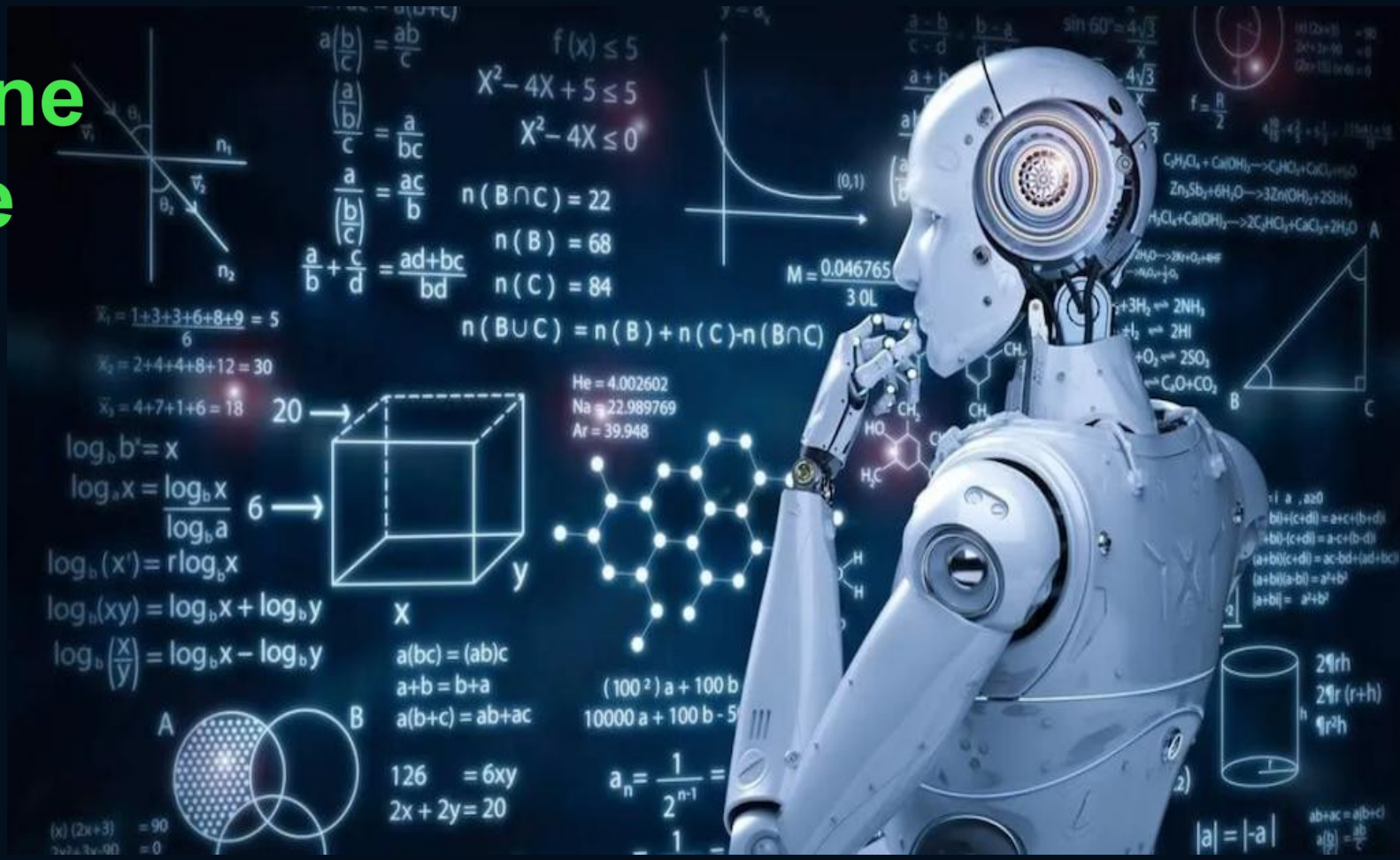
NBS Nottingham Business School

Principles of Machine Learning in Finance

1. Types of Machine Learning | Feature Engineering

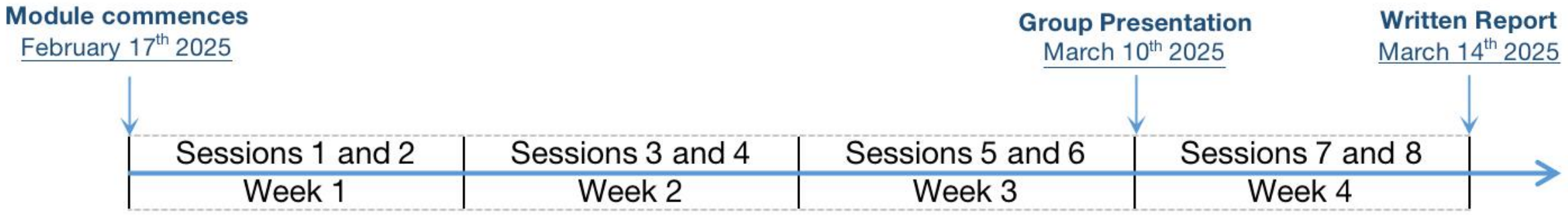


Learning Outcomes

- Main types of machine learning (ML)
- Workflow structure in machine learning (ML): PACE
- Recommendation Systems
- Python Toolbeit for machine learning (ML)
- Feature Engineering in machine learning (ML)
- **Coding Activity 1**: Feature Engineering with Python || [European Bank Data Modelling]

Module Overview

- Format: 4-hour sessions/workshops
- Timeline: February 17th 2025 - March 14th 2025
- Module team: Dr. Olga Khon, Lecturer in digital finance at NBS
- Experiential learning: Coding activities during each session
- Module timeline:



Final Assessment

1. Group Presentation:
   - Assessement date: Session 7 (March 10th 2025, 11:00 PM) || Week 4
   - Area: Machine learning models in finance
   - Focus: The comparison of two machine learning models in finance
2. Written Report (3000 words):
   - Submission deadline: March 14th 2025, 11:00 PM || Week 4
   - Area: Machine learning models in finance
   - Route: Theoretical or emprical report

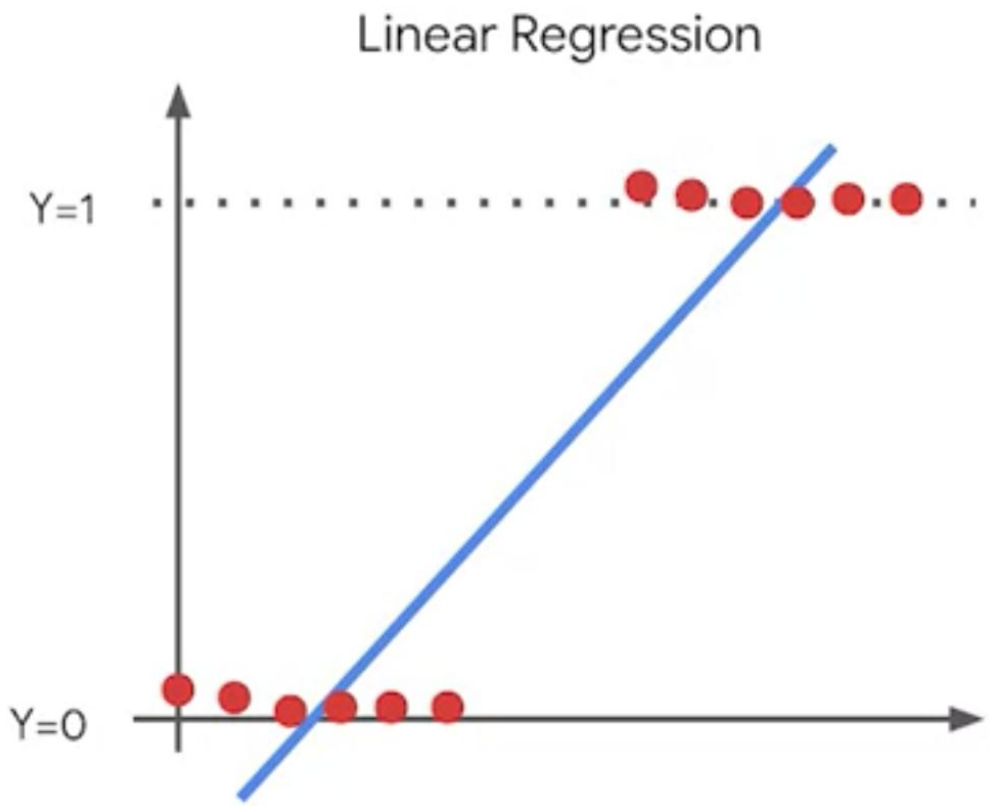
Machine Learning Discoveries

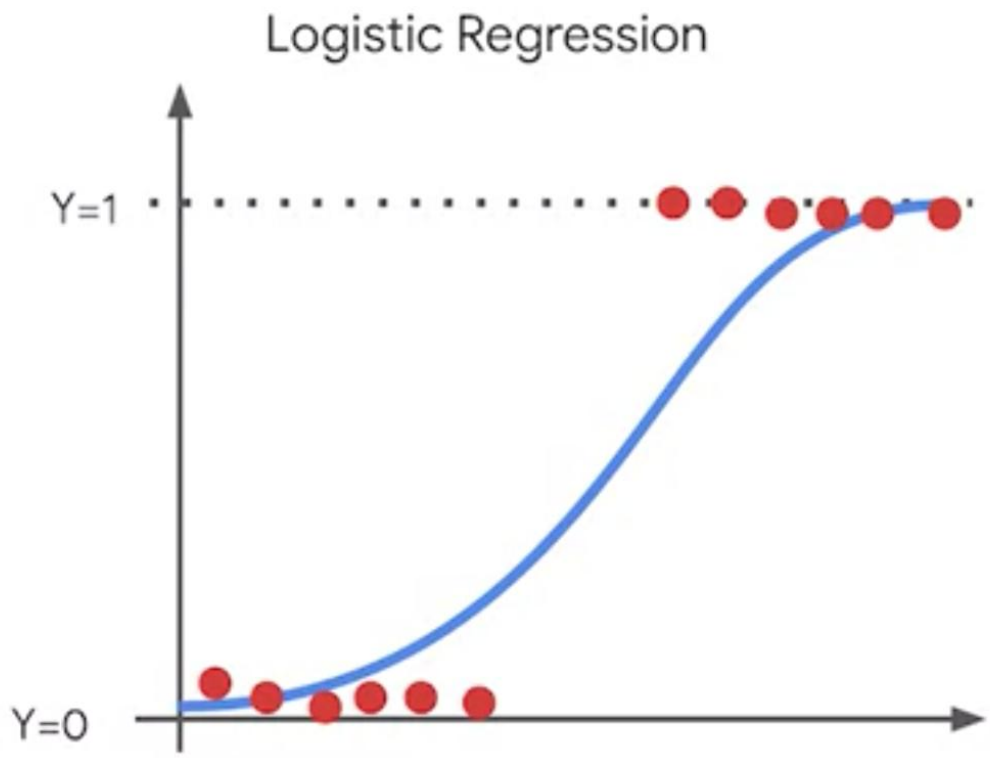
The 2024 Nobel Prize for Physics was awarded to

John J. Hopfield
(Princeton University, NJ, USA)
and
Geoffrey Hinton
(University of Toronto, Canada)

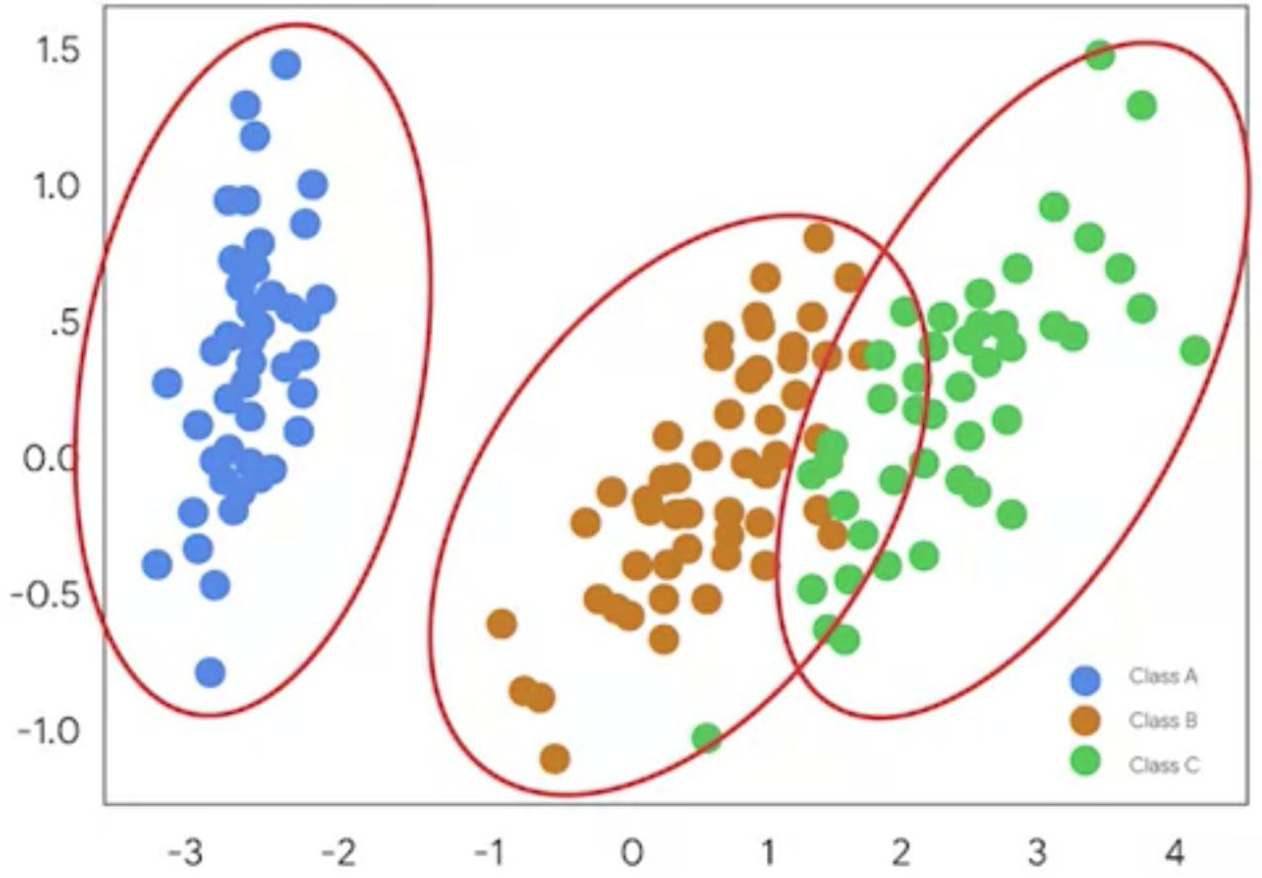
"for foundational discoveries and inventions that enable machine learning with artificial neural networks"

Data Analysis in Finance: Case 1

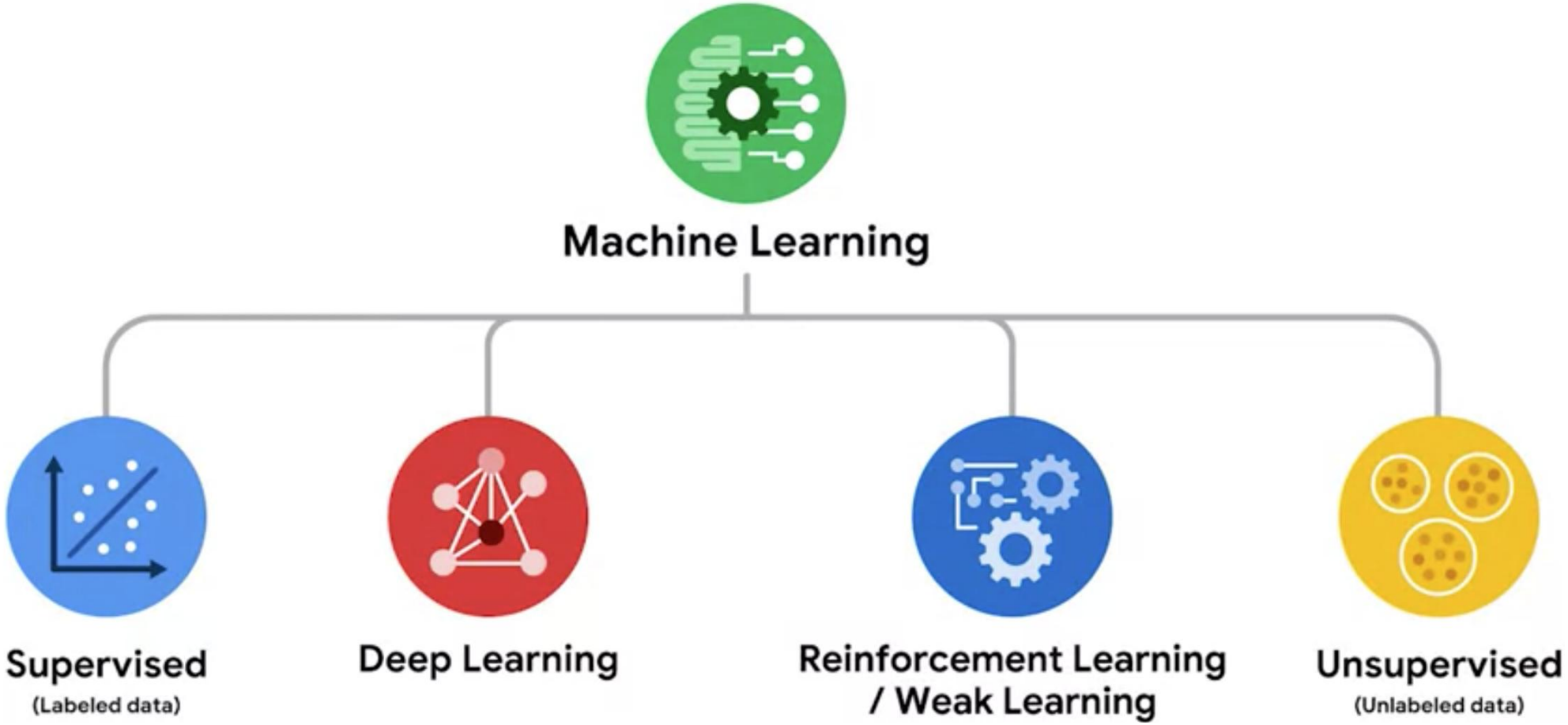




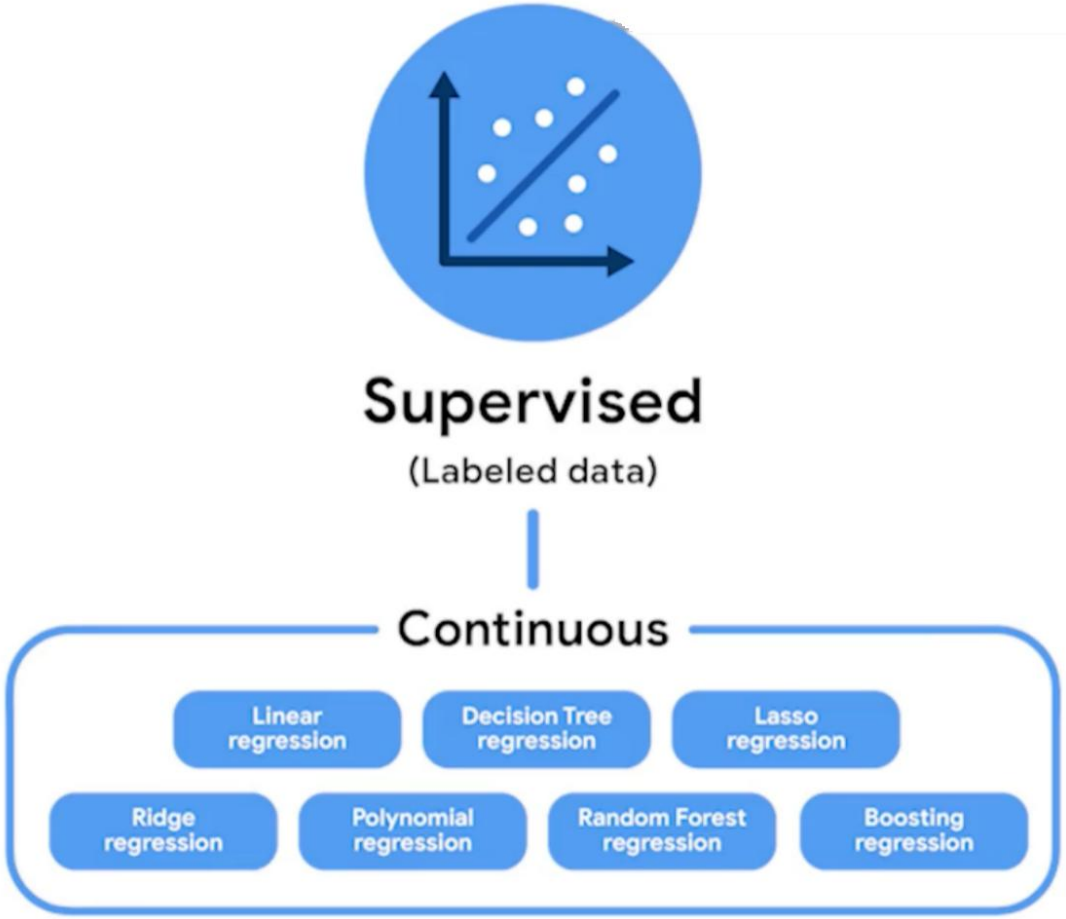
Data Analysis in Finance: Case 2



Main Types of Machine Learning (ML)



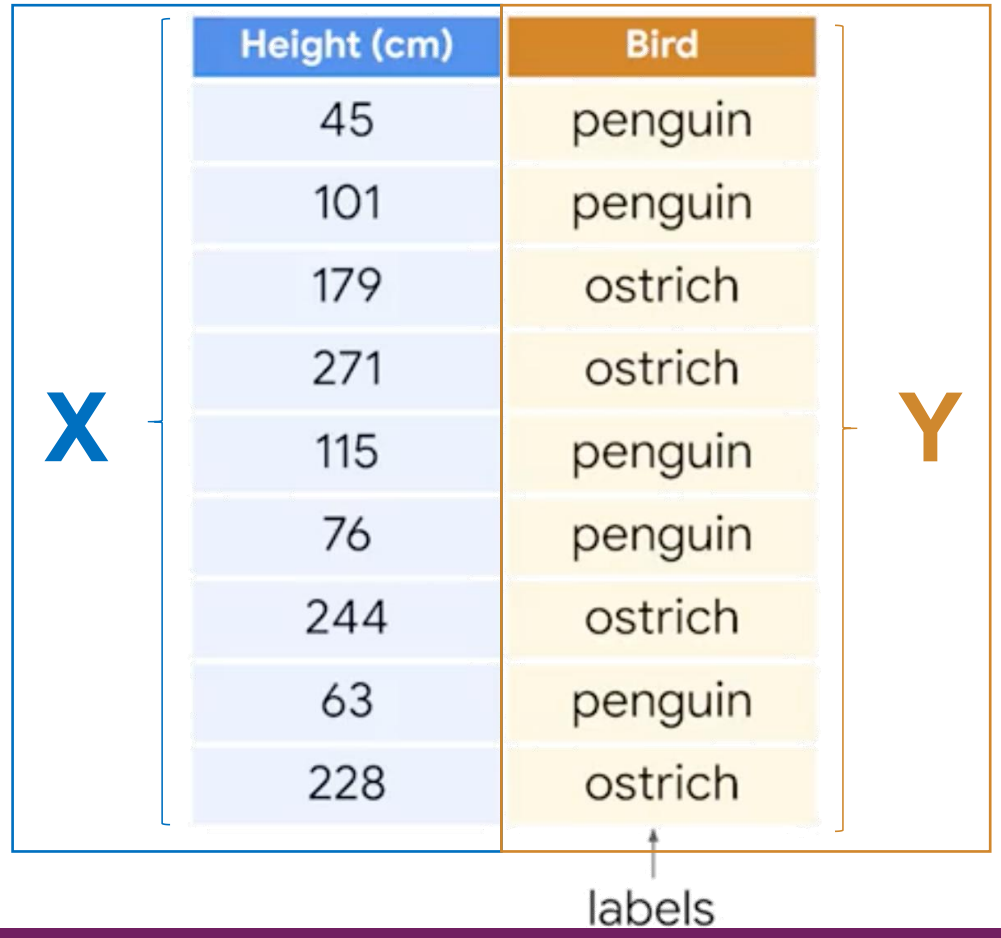
Supervised Machine Learning (ML)



Supervised machine learning uses labeled datasets to train algorithms to classify or predict outcomes.

Example 1: Supervised Machine Learning

Labeled data



Task:

To predict the type of species (Y) based on their height (X)

Unsupervised Machine Learning (ML)

Unsupervised machine learning uses algorithms to analyze and cluster unlabeled datasets.

Example 2: Unsupervised Machine Learning

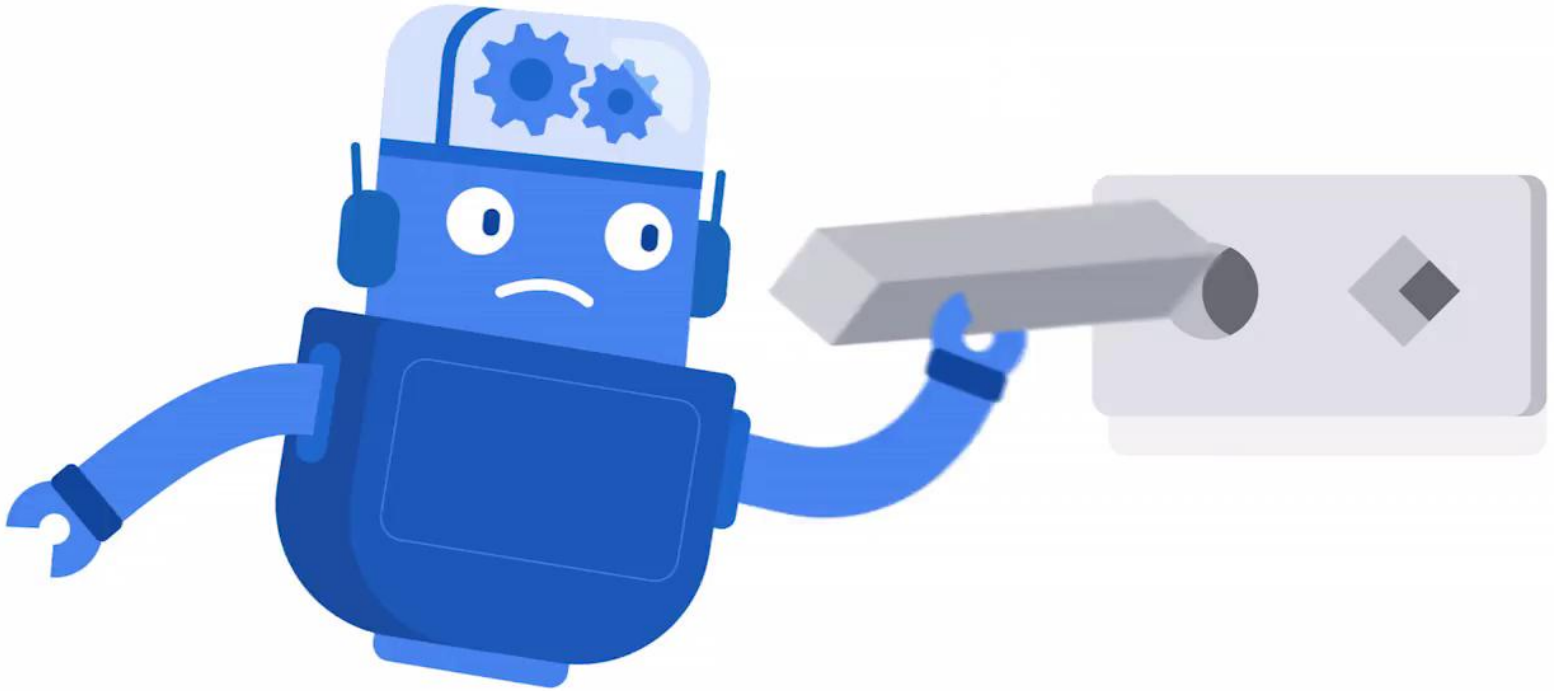
Unlabeled data

Height (cm)	Bird
45	
101	
179	
271	
115	
76	
244	
63	
228	

Task:

To group the species by their similarity based on patterns detected by the model

Reinforcement Learning

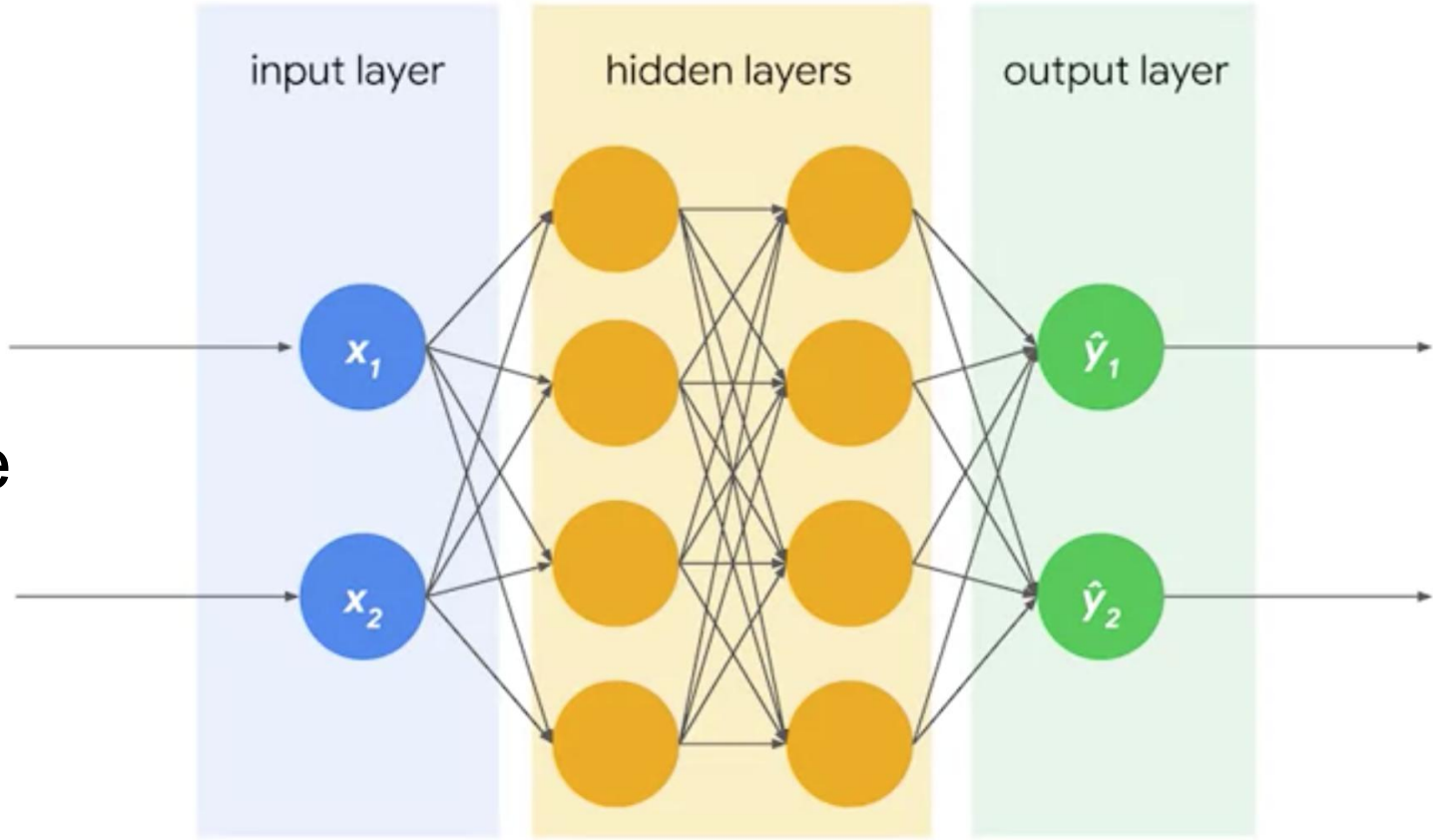


Reinforcement learning is often used in robotics and is based on rewarding or punishing a computer's behaviors.

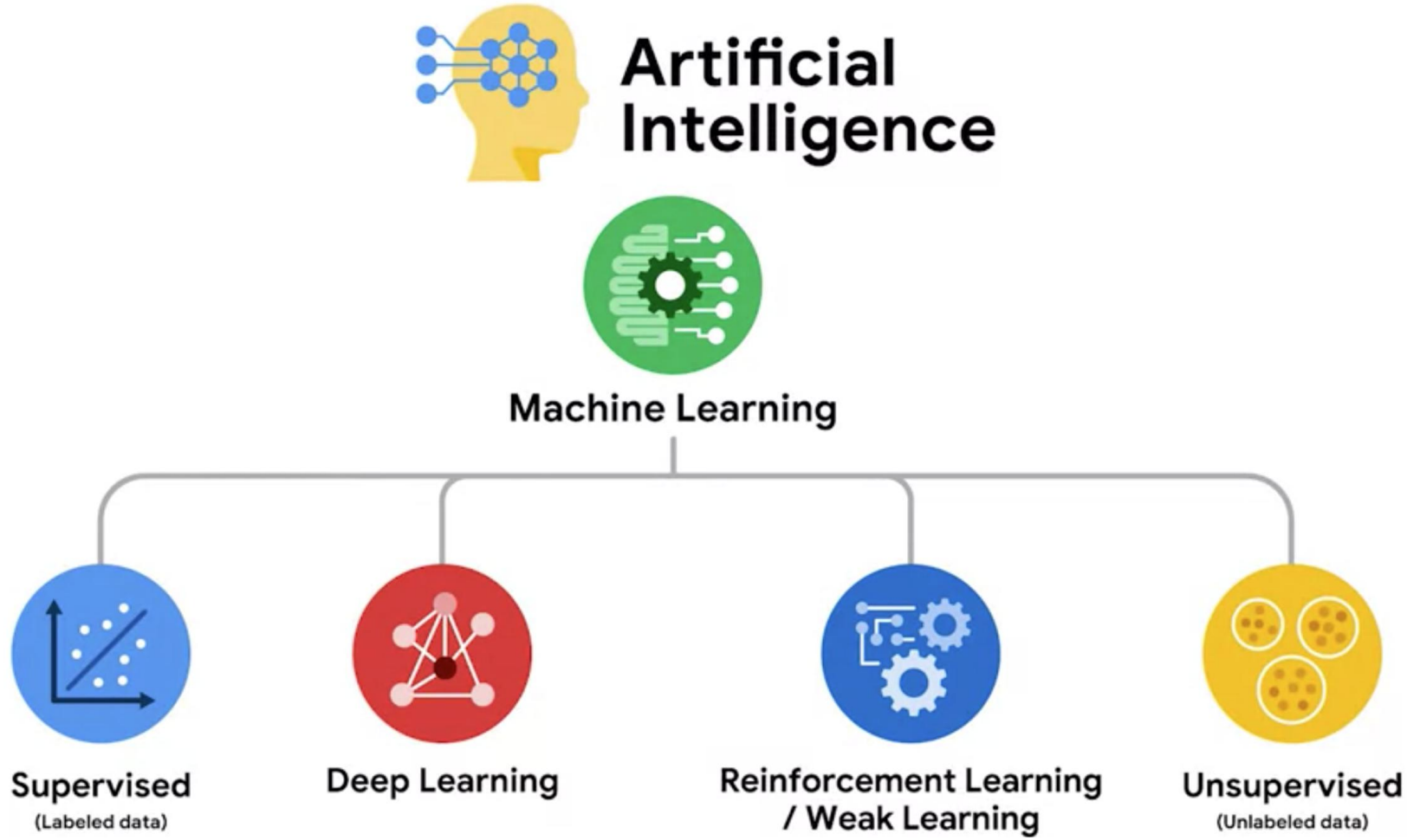
Deep Learning

Deep learning models are made of layers of interconnected nodes.

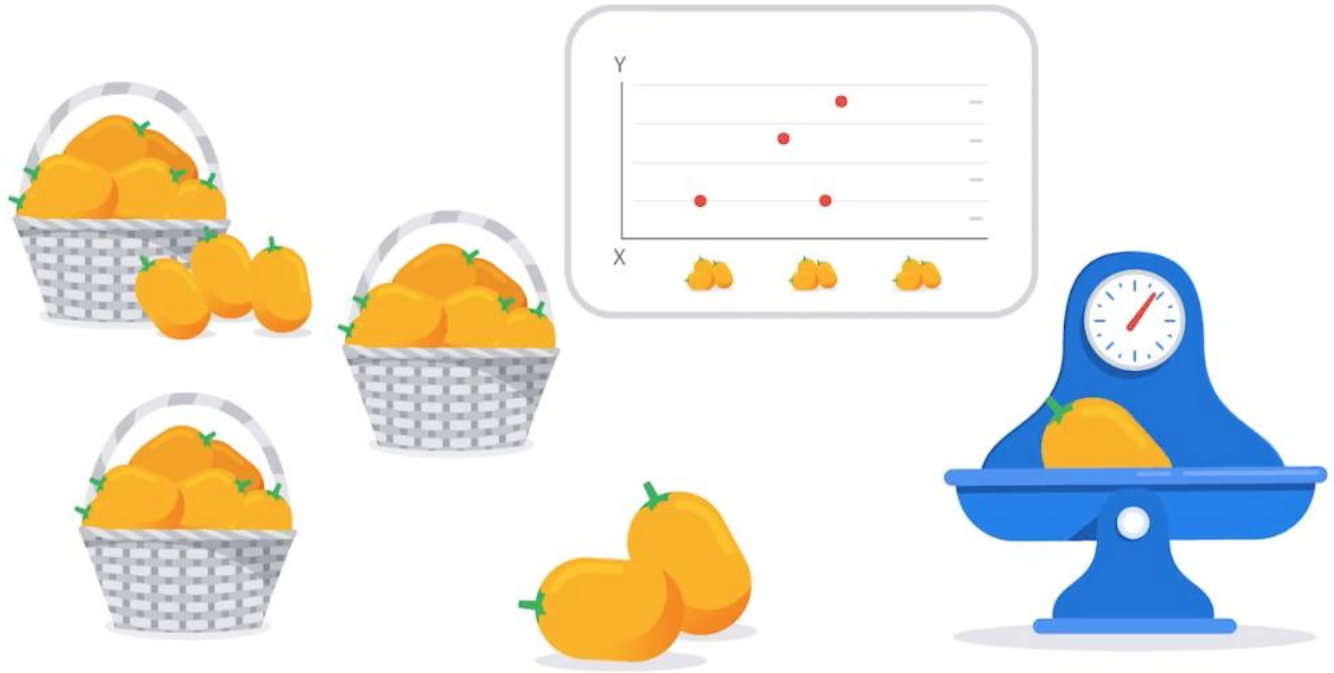
Neural networks are the underlying technology in deep learning.



Artificial Intelligence and Machine Learning

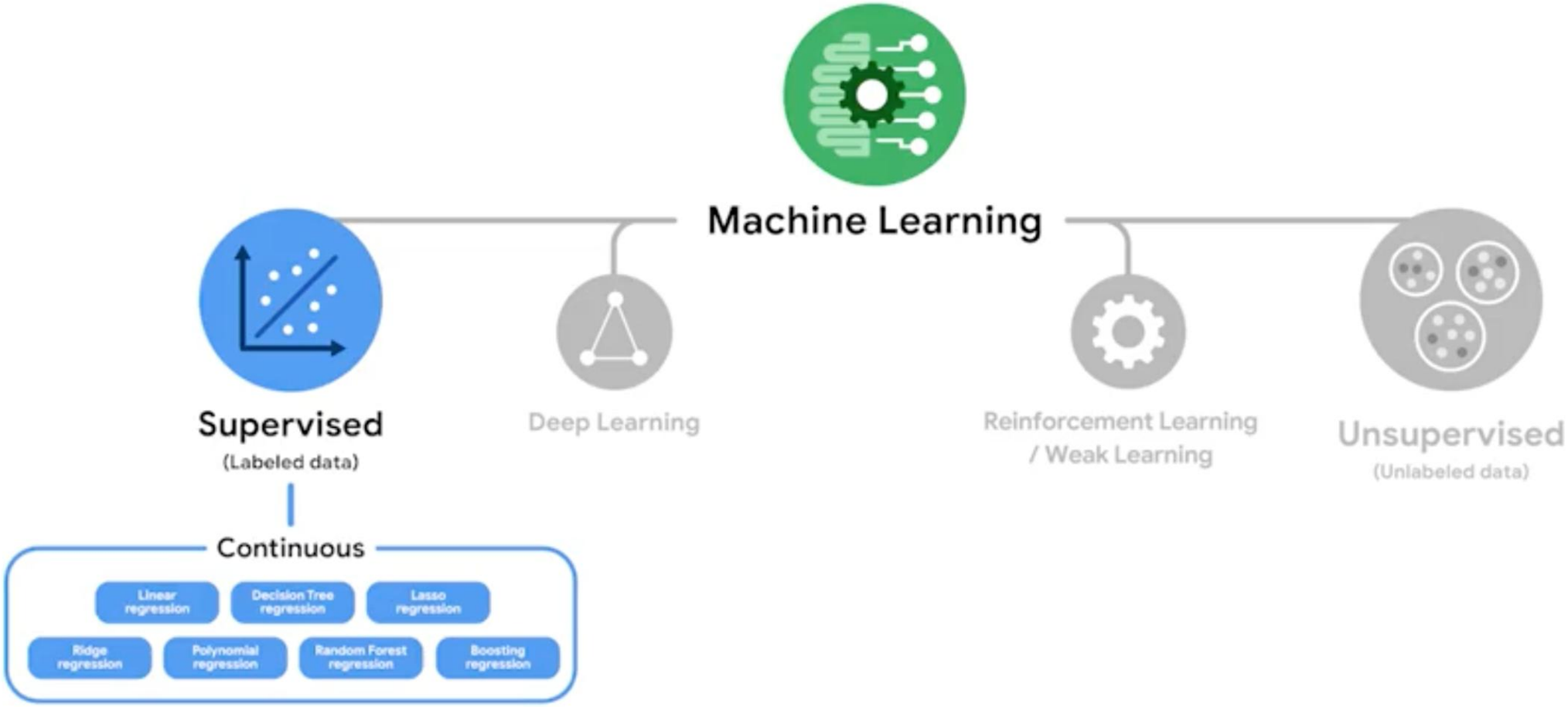


Categorical vs Numerical Features

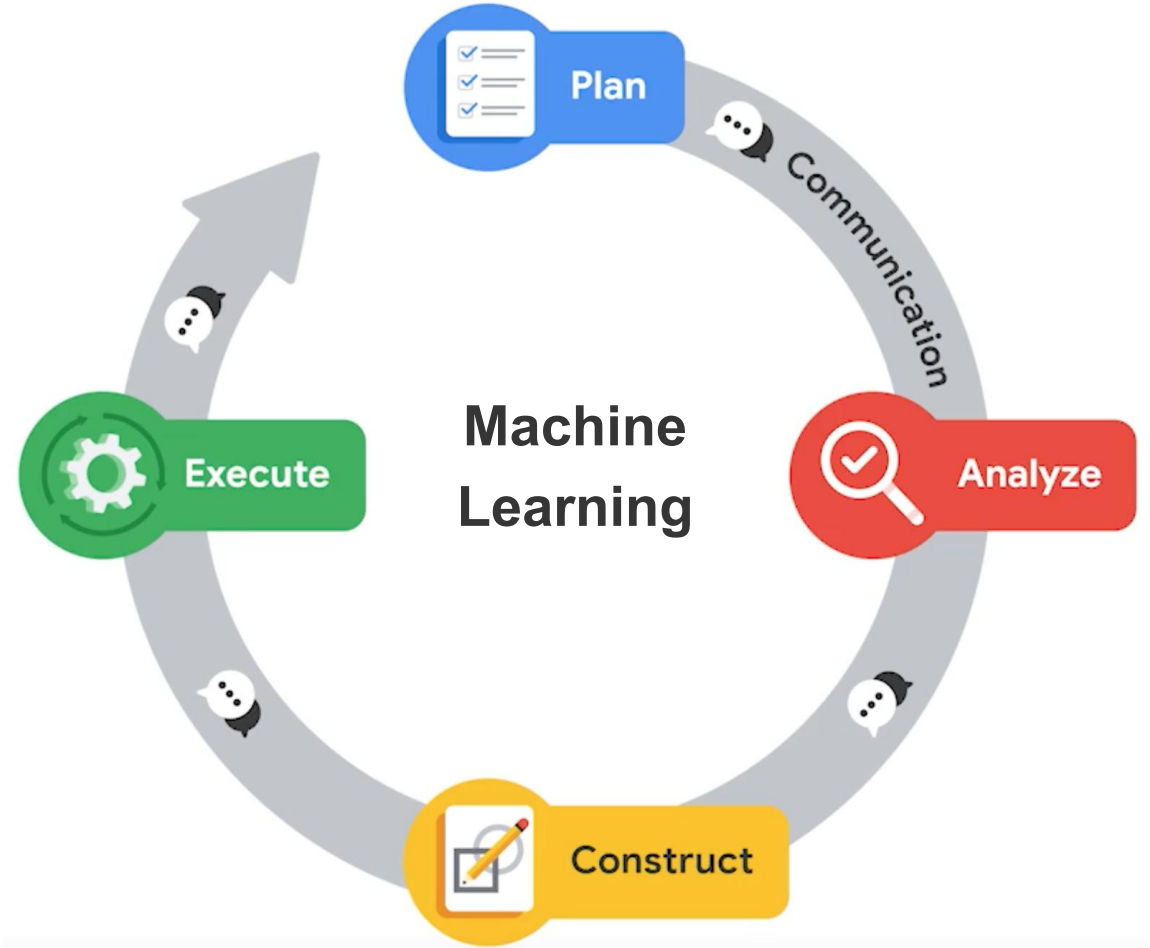


Continious features are feautures that can take on an infinite and uncountable set of values

Machine Learning Map: Continuous Features

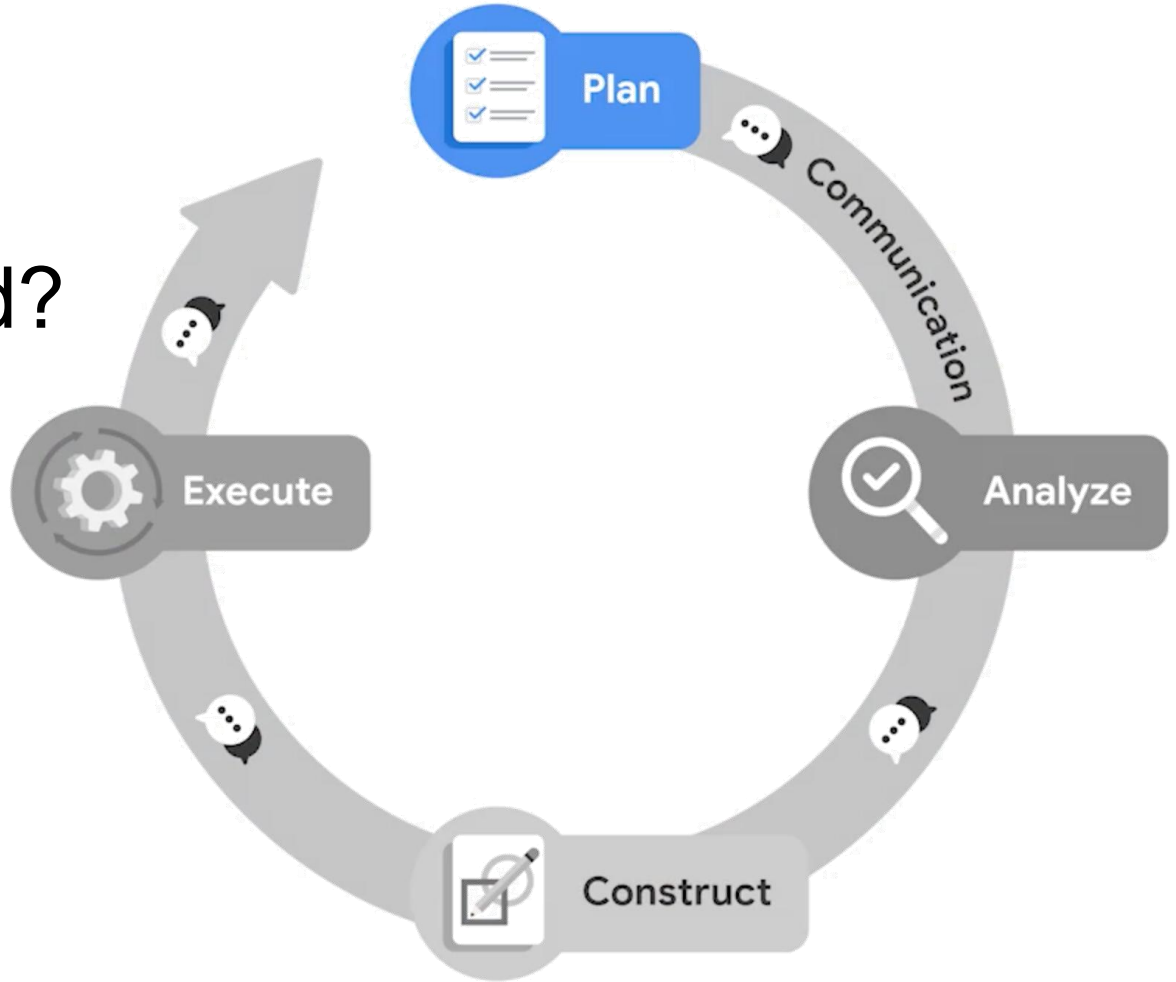


Workflow Structure: PACE in Machine Learning



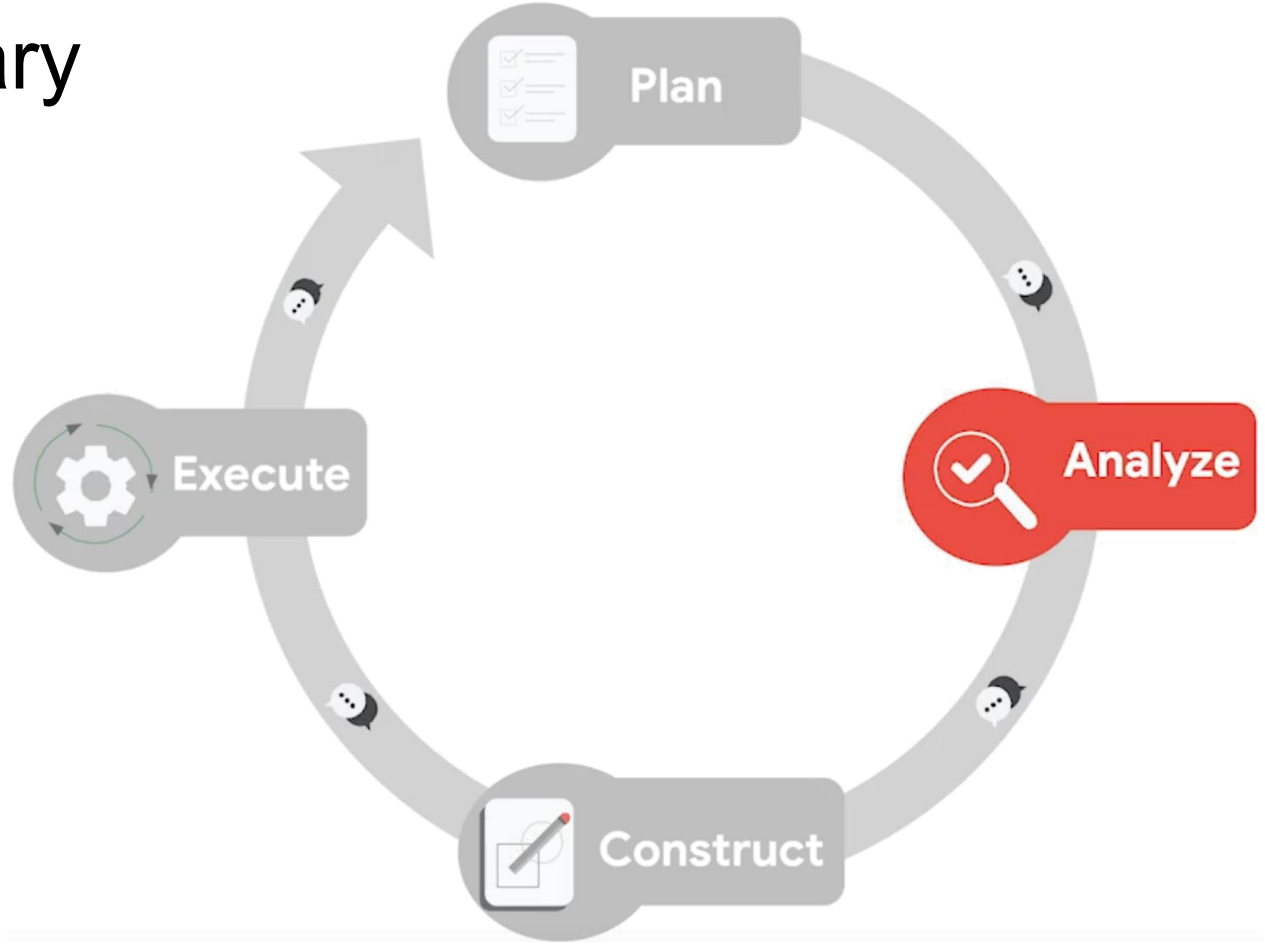
PACE: Plan

- What are the goals of the project?
- What strategies will be needed?
- What will be the business or operational impacts of this plan?



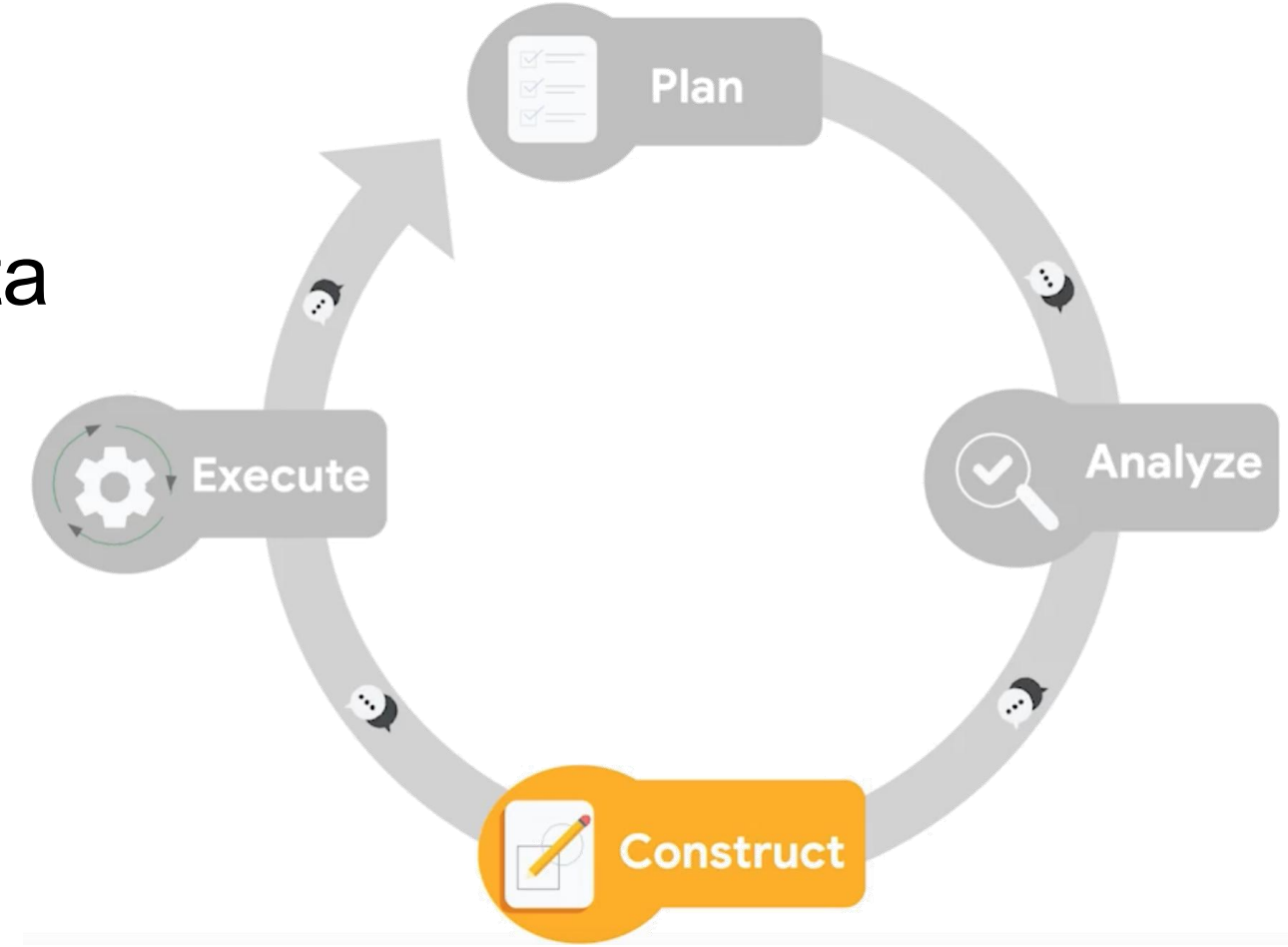
PACE: Analyze

- Acquire the data from primary and secondary sources
- Clean, reorganize, and transform data for analysis
- Engage in EDA
- Work with stakeholders



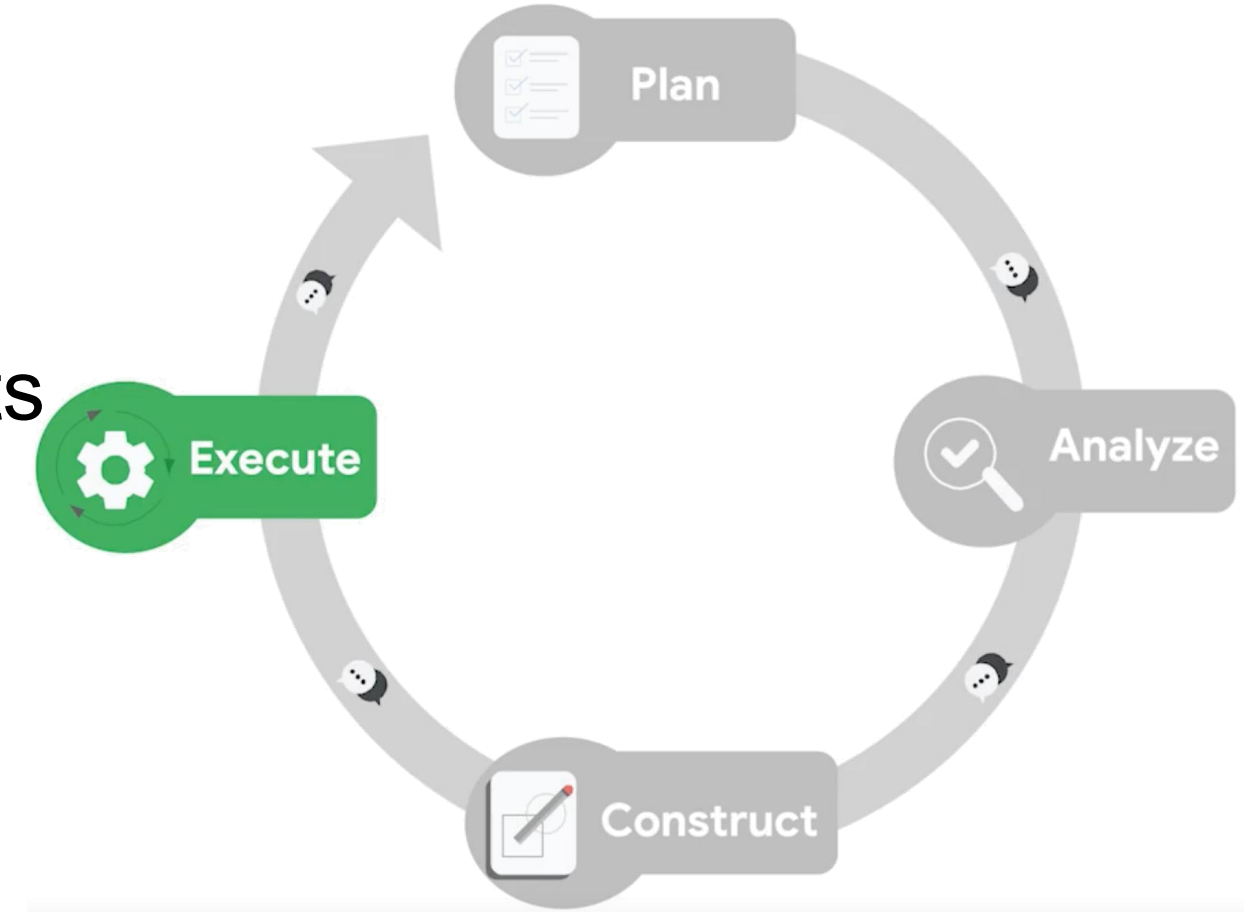
PACE: Construct

- Build and revise machine learning models
- Uncover relationships in data
- Apply statistical inferences about data relationship



PACE: Execute

- Present findings to internal and external stakeholders
- Answer questions
- Consider differing viewpoints
- Present recommendations based on the data



Categorical and Discrete variables

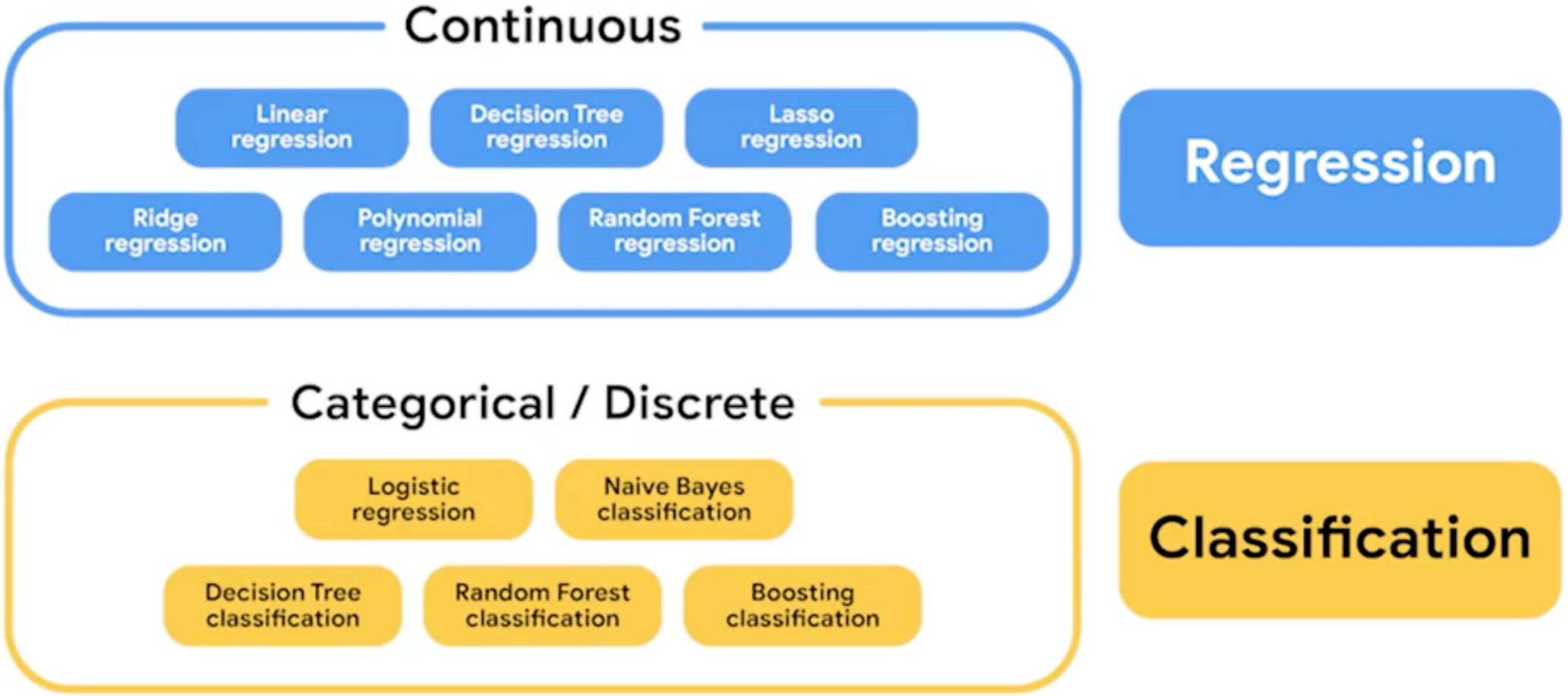
Continious variables are variables that can take on an infinite and uncountable set of values

Categorical and discrete variables are not continious by nature.

Categorical variables are variables that containes a finite number of groups or categories

Discrete features are features with a countable number of values between any two values

ML Map: Categorical / Discrete Features



Machine learning in everyday life

Recommendation systems

Unsupervised learning techniques that use unlabeled data to offer relevant suggestions to users.

The goal is to quantify how similar one thing is to another.

Content-based filtering

Comparisons are made based on attributes of content

Example 3. Recomendation System

A.

Song	Beat	Key	BPM	Piano?	Acoustic guitar?
A	rock	F maj	74	yes	no
B	reggaeton	D min	100	no	yes
C	rock	B ♭ maj	72	yes	no

B.

Song	Beat	Key	BPM	Piano?	Acoustic guitar?
A	rock	F maj	74	yes	no
B	reggaeton	D min	100	no	yes
C	rock	B ♭ maj	72	yes	no

Conten-Based Filtering: Benefits and Drawbacks

Benefits	Drawbacks
+ Easy to understand recommends what user likes; + Doesn't need info from other users; + Can map users and items in the same place	- Always recommends more of the same requires manual input of attributes; - Can't recommend across content type

Example 4. Collaborative Filtering

Collaborative Filtering:

Comparisons are made based on who else liked the content

1.

	Cookies & cream	Chocolate	Fudge brownie	Strawberry	Raspberry
David	4	5	?	1	1
Sanjay	1	2	1	5	5
Amy	5	5	5	1	2
Sara	2	1	1	5	5

2.

	Cookies & cream	Chocolate	Fudge brownie	Strawberry	Raspberry
David	4	5	?	1	1
Sanjay	1	2	1	5	5
Amy	5	5	5	1	2
Sara	2	1	1	5	5

3.

	Cookies & cream	Chocolate	Fudge brownie	Strawberry	Raspberry
David	4	5	5	1	1
Sanjay	1	2	1	5	5
Amy	5	5	5	1	2
Sara	2	1	1	5	5

4.

	Cookies & cream	Chocolate	Fudge brownie	Strawberry	Raspberry
David	4	5	5	1	1
Sanjay	1	2	1	5	5
Amy	5	5	5	1	2
Sara	2	1	1	5	5

Example 4. Collaborative Filtering (2)

Collaborative Filtering:

Comparisons are made based on who else liked the content

4.

David	4	5	5	1	1
Sanjay	1	2	1	5	5
Amy	5	5	5	1	2
Sara	2	1	1	5	5

Collaborative Filtering: Benefits and Drawbacks

Benefits	Drawbacks
+ Can recommend across content type; + Can find hidden correlations in data; + Doesn't require manual mapping	- Requires LOTS of data to work; - Requires lots of data from each user; - Data is sparse

Jupiter Notebook IDE: Advantages

Integrated development envinronment (IDE) is a piece of software with an interface to write, run, and test a piece of code

Jupiter Notebook IDE:

- Uses code to tell a story
- Ability to embed non-code element
- Can be exported to PDF or deployed as html quite easily

Python in Machine Learning

Python is not just a language used in data science.

It's a flexible, general-purpose language that can be used for web development, automation, cryptography, and other tasks.

Types of Python files:

- Python file .py
- Python notebook .ipynb

Advantages: Python Script vs Notebook

Python Script:

- Better for programs incorporating several files
- Easier test and debug

Python Notebook:

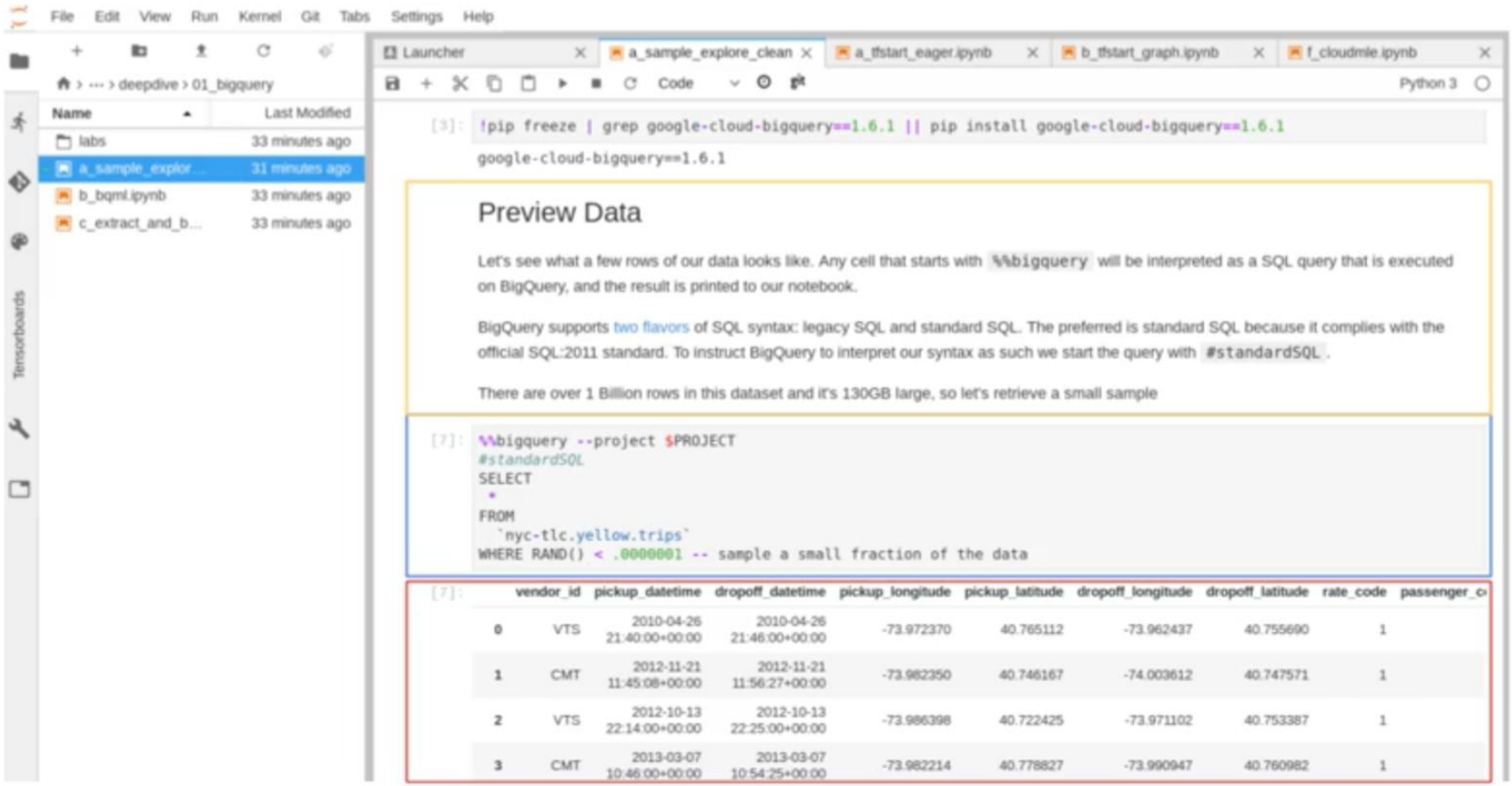
- Uses code to tell a story
- Ability to embeded non-code element
- Can be exported to PDF or deployed as html quite easily

Jupiter Notebook IDE: Review

Markup

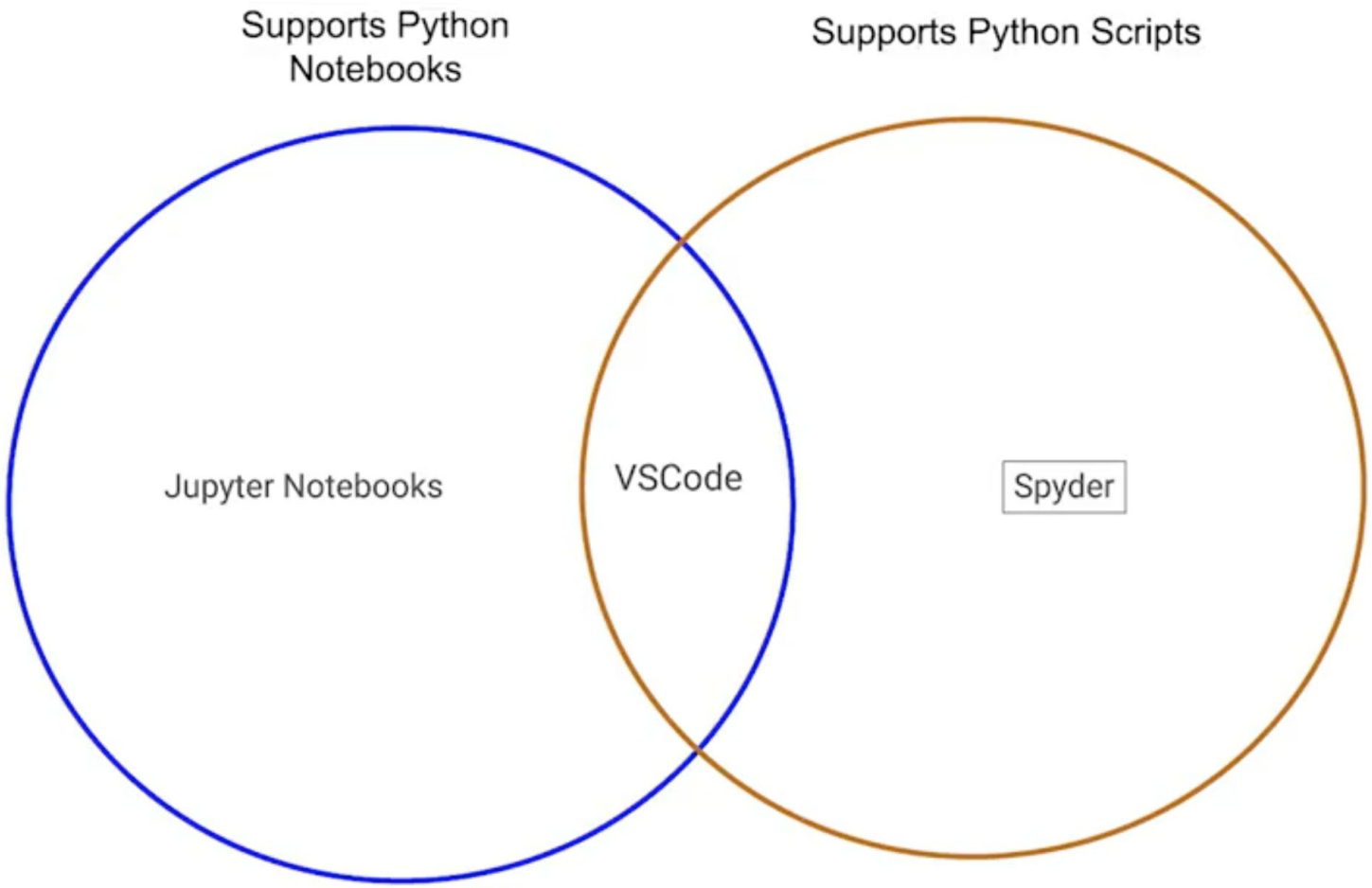
Code

Output



A typical notebook contains code, charts, and explanations.

Example 5. IDEs and Python File Types

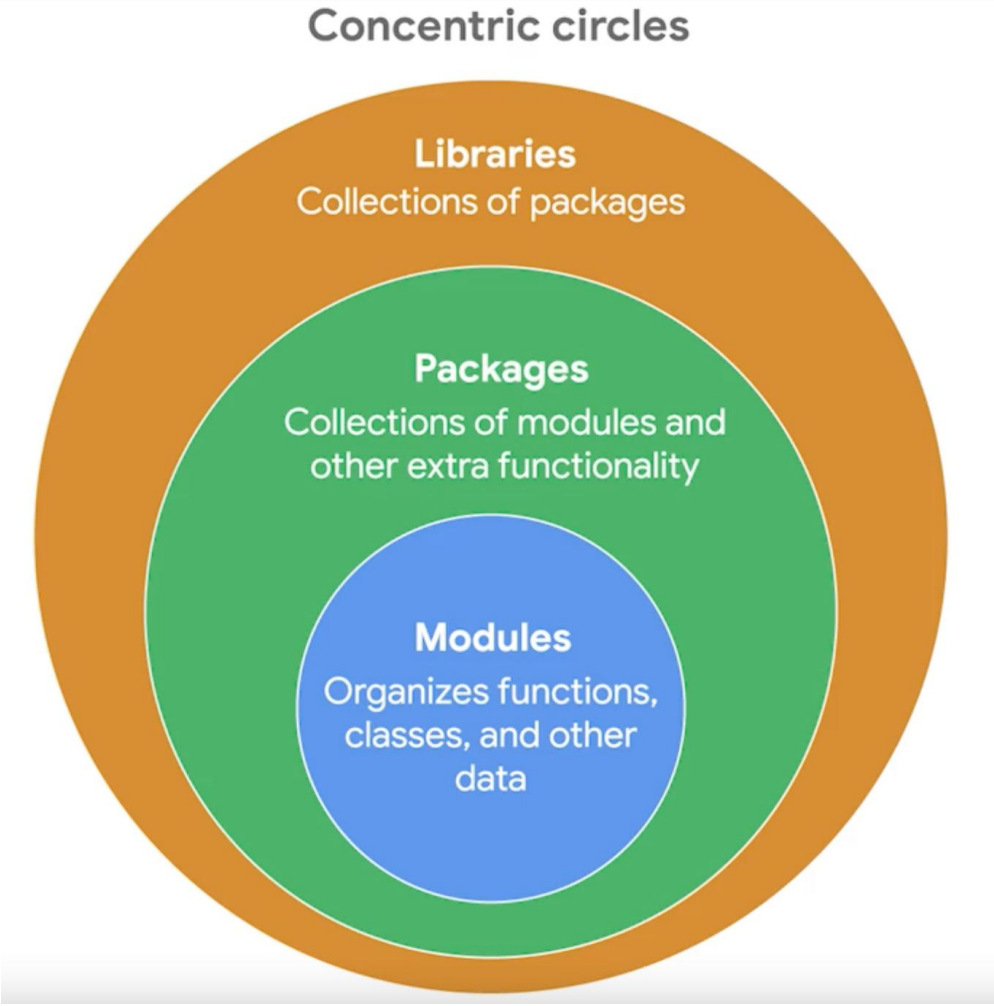


Python Packages, Libraries and Modules

```
import numpy as np
import pandas as pd

from sklearn.model_selection import train_test_split

from imblearn.over_sampling import RandomOverSampler
```



Python Packages

Operational packages: load, structure and prepare a data set for further analysis,

e.g. Pandas, NumPy, and Sumpy Packages.

Data visualization packages: create the perfect plots and graphs based on the needs of a project,

e.g. Matplotlib, Seaborn, Plotly Packages.

Machine learning library, *scikit-learn*, enables you to build a variety of model types (unsupervised & supervised)

Feature Engineering

Feature engineering is the process of using practical, statistical and data science knowledge to select, transform, or extract characteristics, properties, and atributes from raw data.

Continious variables are variables with values obtained by measurement; as a result, they can take on an infinite and uncountable set of values.

Categorical variables contain a finite number of groups, categories, or countable numerical values.

Categories of Feature Engineering

1. **Selection**: Select the features in the data that contribute the most to predicting your response variable;
2. **Transformation**: Modifying existing features in a way that improves accuracy when training the model;
3. **Extraction**: Taking multiple feautures to create a new one that would improve the accuracy of the algorithm.

Example 6. Feature Engineering: Selection

Outlook	Temp	Humidity	Windy	Play Football
Rainy	t > 80	High	False	No
Rainy	t > 80	High	True	No
Overcast	t > 80	High	False	Yes
Sunny	70 ≤ t ≤ 80	High	False	Yes
Sunny	t < 70	Normal	False	Yes
Sunny	t < 70	Normal	True	No
Overcast	t < 70	Normal	True	Yes
Rainy	70 ≤ t ≤ 80	High	False	No
Rainy	t < 70	Normal	False	Yes
Sunny	70 ≤ t ≤ 80	Normal	False	Yes
Rainy	70 ≤ t ≤ 80	Normal	True	Yes
Overcast	70 ≤ t ≤ 80	High	True	Yes
Overcast	t > 80	Normal	False	Yes
Sunny	70 ≤ t ≤ 80	High	True	No

Example 6. Feature Engineering: Selection (2)

Outlook	Temp	Humidity	Windy	Play Football
Rainy	t > 80	High	False	No
Rainy	t > 80	High	True	No
Overcast	t > 80	High	False	Yes
Sunny	70 ≤ t ≤ 80	High	False	Yes
Sunny	t < 70	Normal	False	Yes
Sunny	t < 70	Normal	True	No
Overcast	t < 70	Normal	True	Yes
Rainy	70 ≤ t ≤ 80	High	False	No
Rainy	t < 70	Normal	False	Yes
Sunny	70 ≤ t ≤ 80	Normal	False	Yes
Rainy	70 ≤ t ≤ 80	Normal	True	Yes
Overcast	70 ≤ t ≤ 80	High	True	Yes
Overcast	t > 80	Normal	False	Yes
Sunny	70 ≤ t ≤ 80	High	True	No

Example 6. Feature Engineering: Transformation

Outlook	Temp	Humidity	Windy	Play Football
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Example 7. Fraud Detection

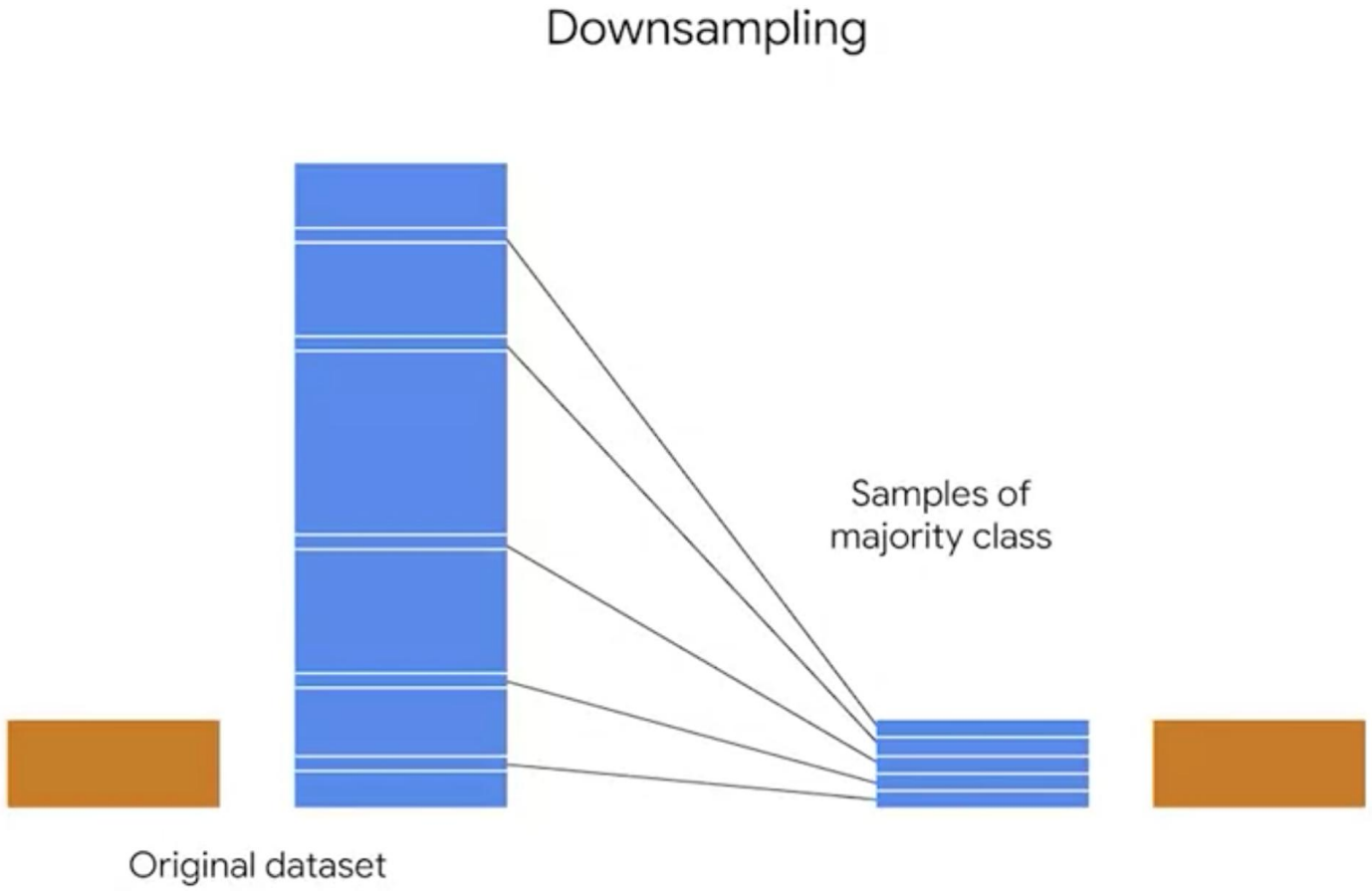
Content: Millions of examples of nonfraudulent transactions and only a few 1000 examples of actual fraudulent transactions.

Problem: Class imbalance is a situation when a dataset has a predictor variable that contains more instances of one outcome than another

Solution techniques: Downsampling and **Upsampling**

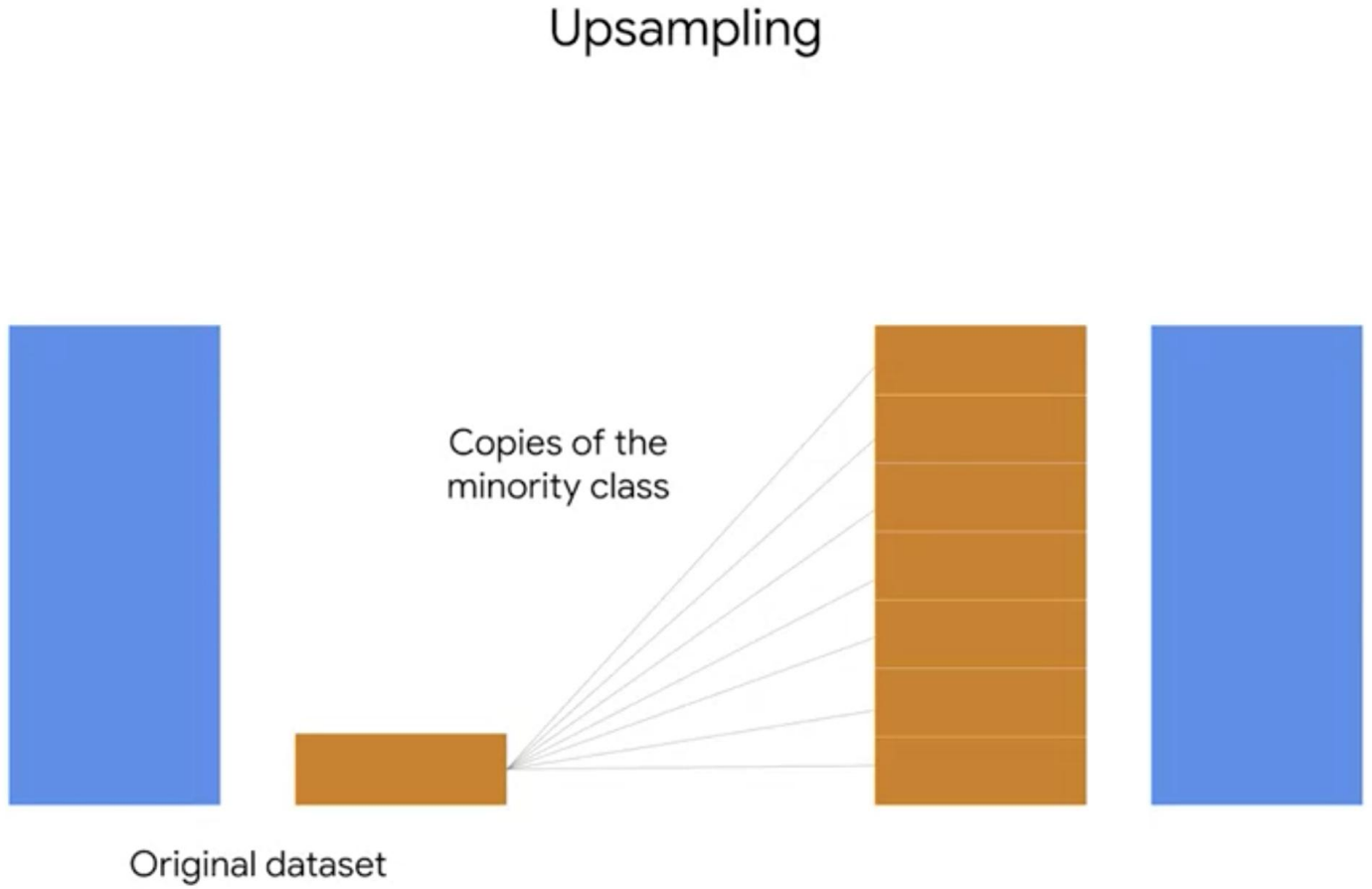
Downsampling

Downsampling involves altering the majority class by using less of the original dataset to produce

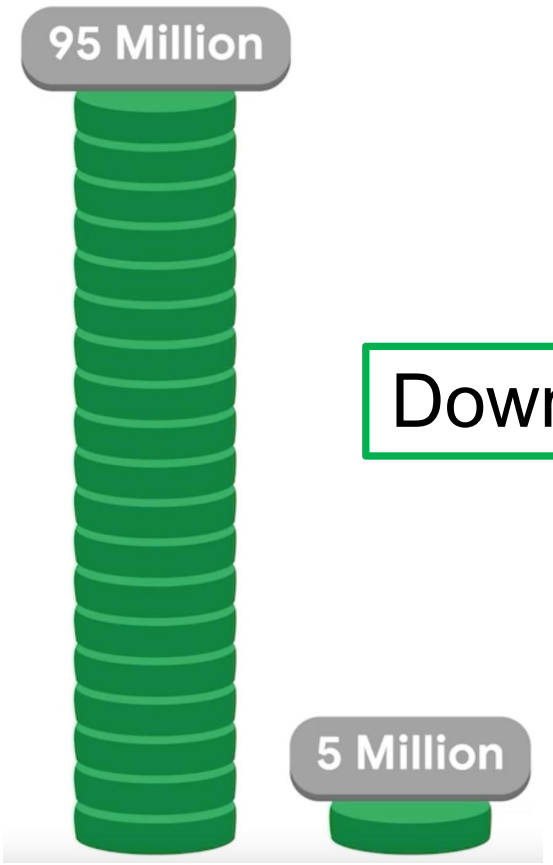


Upsampling

Upsampling artificially increases the frequency of the minority class

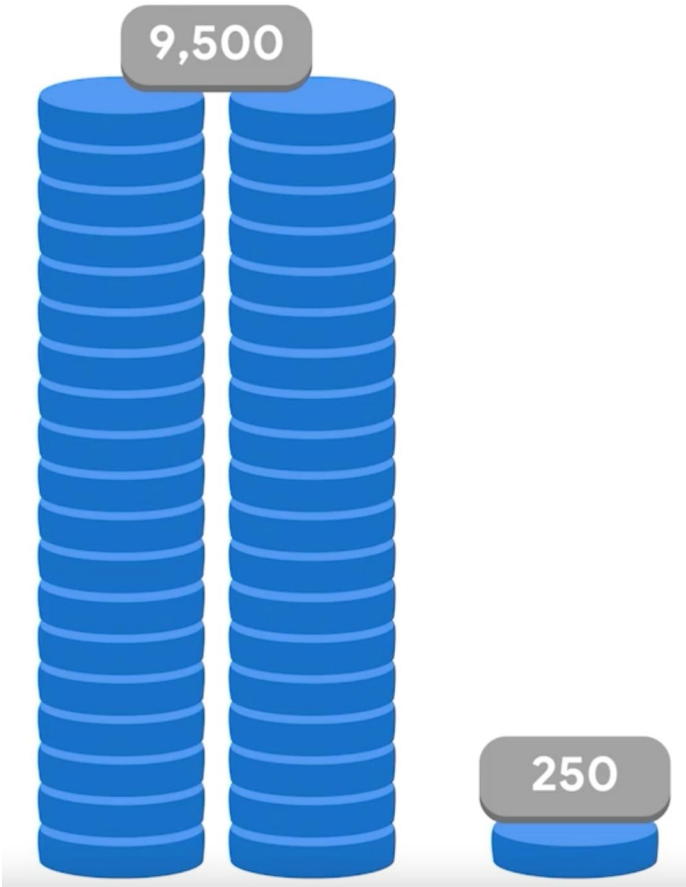


Downsampling vs Upsampling



Downsampling

Upsampling



Coding Activity 1. Feature Engineering with Python

Lab 1. Feature Engineering with Python || European Bank Data Modelling

Steps to follow:

1. Upload the following files from the module learning room:
   - Jupiter notebook "Lab1_Feature_engineering_with_Python.ipynb"
   - Dataset csv-file "Bank_Modelling.csv"
   - Instructions pdf-file "Lab1_Dataset_Description.pdf"
2. Follow along in the Jupiter notebook

Coding Activity 1. Feature Engineering with Python: Definitions

Customer churn is a business term that describes how many and at what rate cutomers stop using a product or service, or stop doing business with a company altogether

Q&A

Thank you!