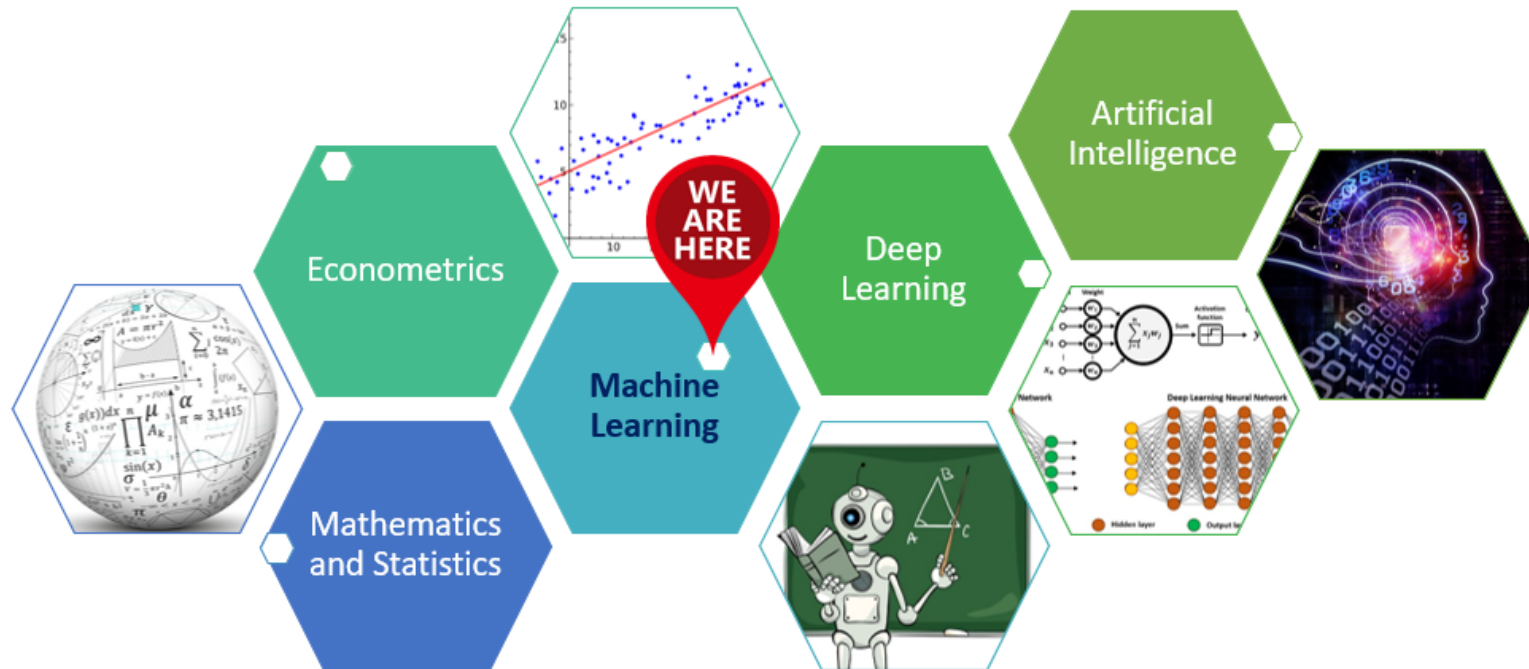
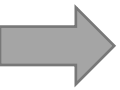


Module 1- Part I

Welcome to the magic world of Machine Learning





Big picture: Econometrics vs Machine Learning



What are we trying to do as a researcher?



Solve real world problems, right?



Is there a theory?

What is the relationship between

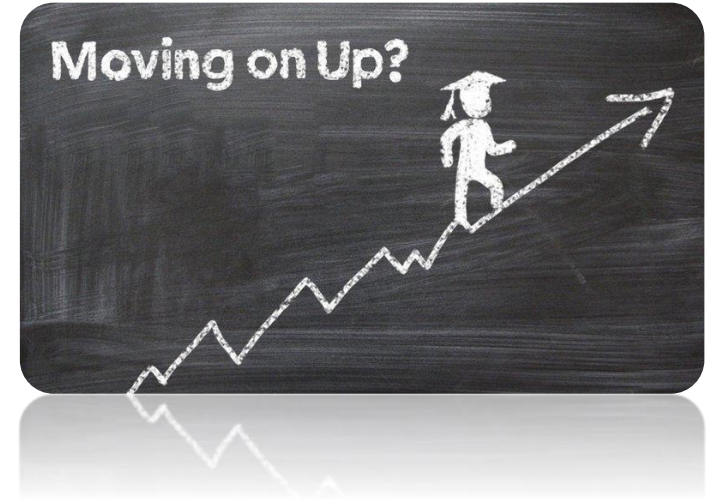
- Sales and advertisement / R&D expenditure / seasonality / industry / ... ?
- Quantity demanded and price / income / technology / price of competitors / ... ?
- Wage and education/ age/ gender/ experience/ ...?

➔ A simple example

- Quantifying wage components! (is there a theory?)
- What are the drivers:
 - Education, age, experience, IQ, ...
 - Ethnicity, race, gender, ...
 - Industry, location, working hours, ...
- Let's build a model (**assuming** a linear functional form!)

$$wage = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 exper + \beta_4 IQ + \dots + \beta_k hours + u$$

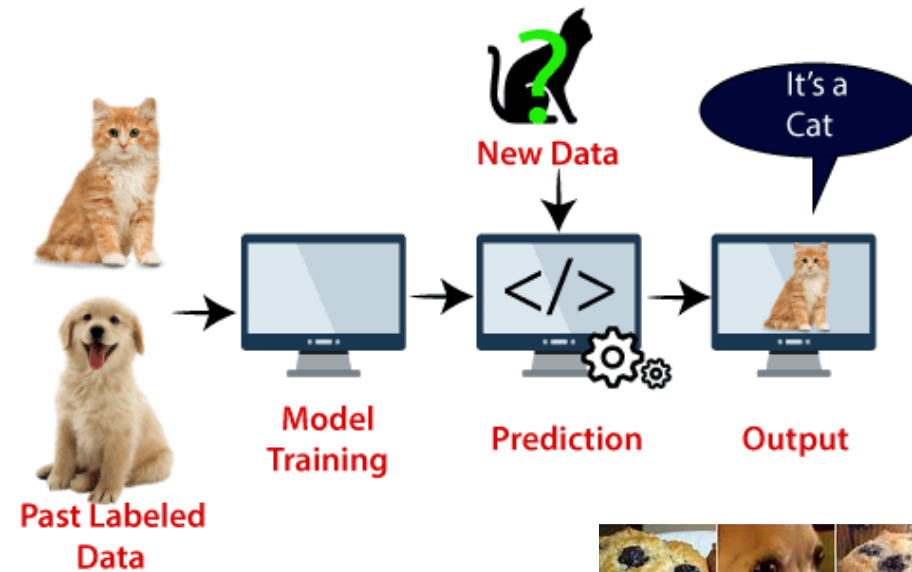
- Can you **interpret** this model? Do you care about the interpretability?
- Can you make **predictions** using your model?
- Can you make this functional form more flexible? What are the caveats?





A different example

- Cat vs dog classification problem (image classification)



- Do you really care about **interpretability** of the model here?
- What about accuracy of your **predictions**?





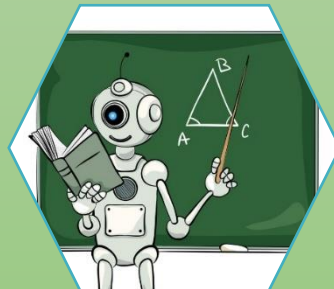
Artificial intelligence vs Machine learning vs Deep learning

Artificial intelligence: Any technique which enables machines to mimic human behavior



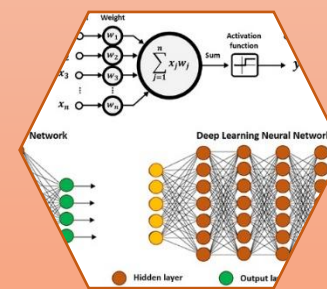
1950's

Machine Learning: Subset of AI that enables computers to learn from data. the model is trained with a set of algorithms



1980's

Deep Learning: Subset of ML that extract patterns from data using neural networks.



2010's



Statistical learning vs machine learning

	Statistical Learning	Machine Learning / Deep Learning
Focus	Hypothesis testing & interpretability	Predictive accuracy and extracting complex patterns
Driver	Math, theory, hypothesis	Fitting data
Data size	Any reasonable set	Big data
Data type	Structured	Structured, unstructured, semi-structured
Dimensions / scalability	Mostly low dimensional data	High dimensional data
Strength	Understand causal relationship & behavior	Prediction (forecasting and nowcasting)
Interpretability	High	Medium to Low



Limitations of Econometrics/Structured ML

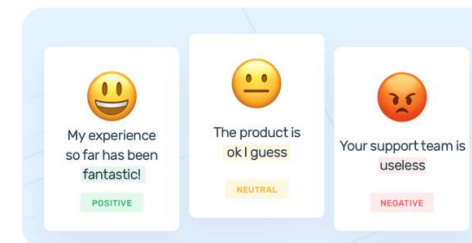
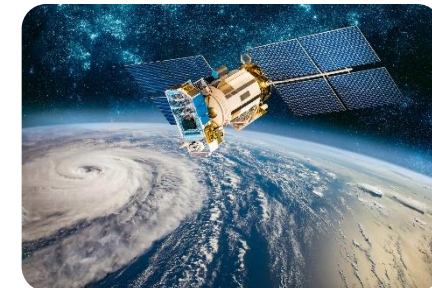
Econometrics/structured ML can **only** handle structured data (tabular data)!

Structured Data

	A	B	C	D
1	Date	Account	Transaction Type	Amount
2	2017-01-12	123	Credit	6089.78
3	2017-01-12	123	Fee	9.99
4	2017-01-12	456	Debit	1997
5	2017-01-12	123	Debit	20996.12
6	2017-01-13	123	Debit	17
7	2017-01-13	123	Debit	914.36
8	2017-01-14	789	Credit	11314
9	2017-01-14	789	Fee	9.99
10	2017-01-14	456	Debit	15247.89
11	2017-01-14	123	Debit	671.28
12	2017-01-15	456	Credit	5072.1
13	2017-01-15	456	Fee	9.99
14	2017-01-16	456	Debit	5109.07
15	2017-01-19	123	Credit	482.01



Unstructured Data
(everything else!!)



➔ A more complex example

Stock price prediction \$\$\$

- What are the classical drivers:
 - Company's fundamentals (balance sheet, income statement, cash flow statement)
 - Competitors (comparing multiples)
 - Technical analysis!
 - Seasonality (holidays, months, days, ...)

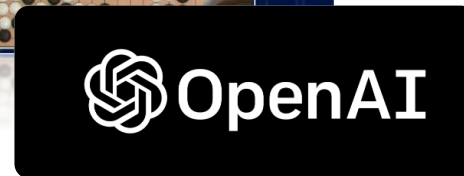


What else?

- Market sentiment (news, tweets, blogger opinions, conference calls, ...)
- Satellite images from parking lots!

Why should I learn it?

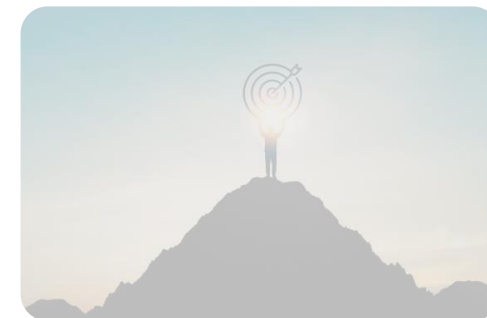
- It's a bid deal, it is **everywhere!**
- Better career opportunities
- Hedge against next recession





Class Modules

- Module 1- Introduction to Machine Learning
- Module 2- Setting up Machine Learning Environment
- Module 3- Linear Regression (Econometrics approach)
- Module 4- Machine Learning Fundamentals
- Module 5- Linear Regression (Machine Learning approach)
- Module 6- Penalized Regression (Ridge, LASSO, Elastic Net)
- Module 7- Logistic Regression
- Module 8- K-Nearest Neighbors (KNN)
- Module 9- Classification and Regression Trees (CART)
- Module 10- Bagging and Boosting
- Module 11- Dimensionality Reduction (PCA)
- Module 12- Clustering (KMeans – Hierarchical)



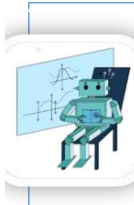
Module 1- Part II

What is Machine Learning?



Supervised

- Regression
- Classification



Unsupervised

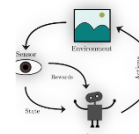
- Clustering
- Anomaly detection
- Dimensionality reduction



Semi-supervised



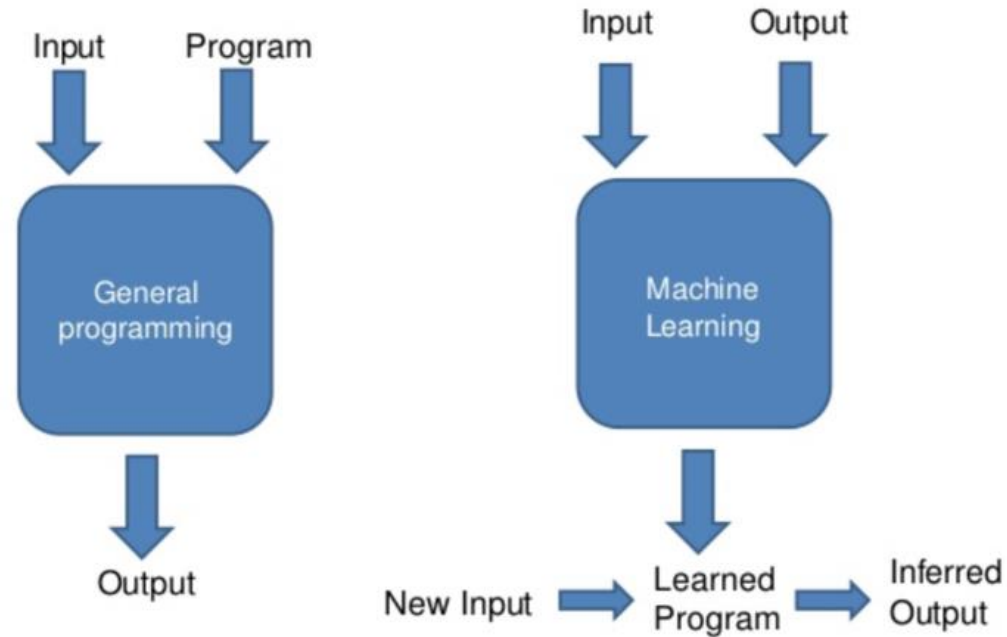
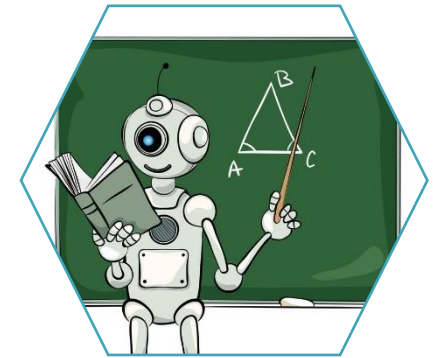
Self-supervised



Reinforcement Learning



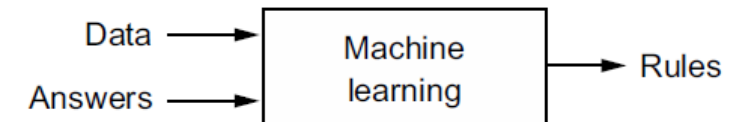
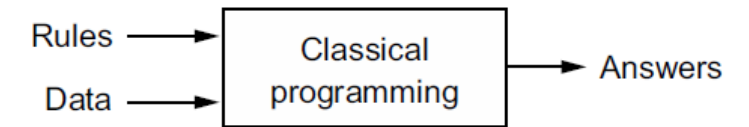
General programming vs Machine learning

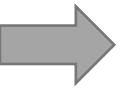


- Machine Learning: Involves **automated detection of meaningful patterns** in data and apply the pattern

➔ What is Machine Learning?

- A machine learning system is **trained** (with algorithms) rather than explicitly **programmed**.
- Machine Learning is a subset of AI that enables computers to **learn** from data.
- ML involves automated detection of meaningful **patterns** in data and apply the pattern to make **predictions** on **unseen data**!
- This is done by **minimizing** the loss on the training data.
- The goal is to **maximize** the performance on the unseen data.





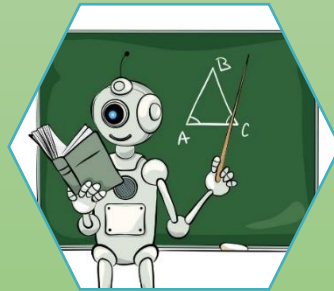
Artificial intelligence vs Machine learning vs Deep learning

Artificial intelligence: Any technique which enables machines to mimic human behavior



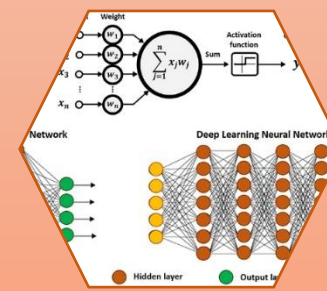
1950's

Machine Learning: Subset of AI that enables computers to learn from data. the model is trained with a set of algorithms



1980's

Deep Learning: Subset of ML that extract patterns from data using neural networks.



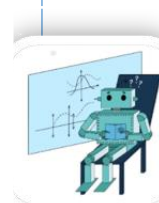
2010's

→ Types of Machine Learning



Supervised

- Regression
- Classification



Unsupervised

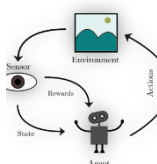
- Clustering
- Anomaly detection
- Dimensionality reduction



Semi-supervised



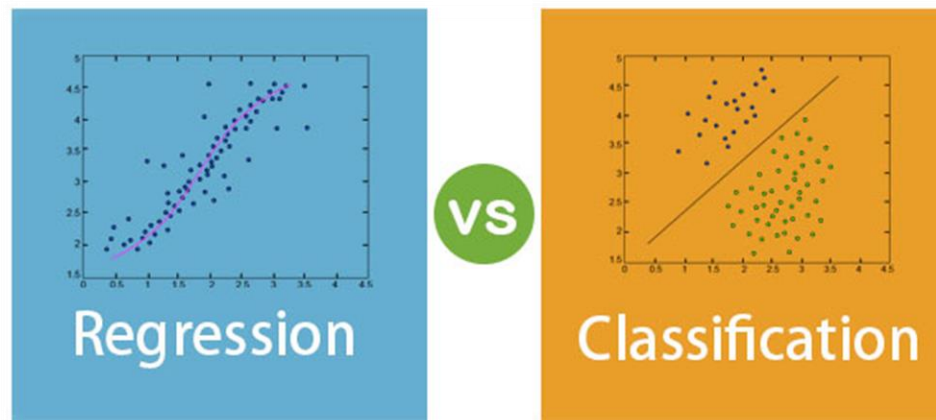
Self-supervised

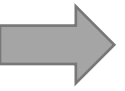


Reinforcement Learning

→ Supervised Learning

- Supervised learning is a type of machine learning where the algorithm is **trained on labeled data**.
- The data is labeled, meaning that the data has been **tagged with the correct output**.
- The model is then able to learn **the relationship** between the input data and the corresponding output labels and can make predictions on new data.
- **Regression:**
 - Predicting housing price
 - Predicting stock market returns
- **Classification:**
 - Generating buy, sell, hold signals.
 - Predicting credit default rate.



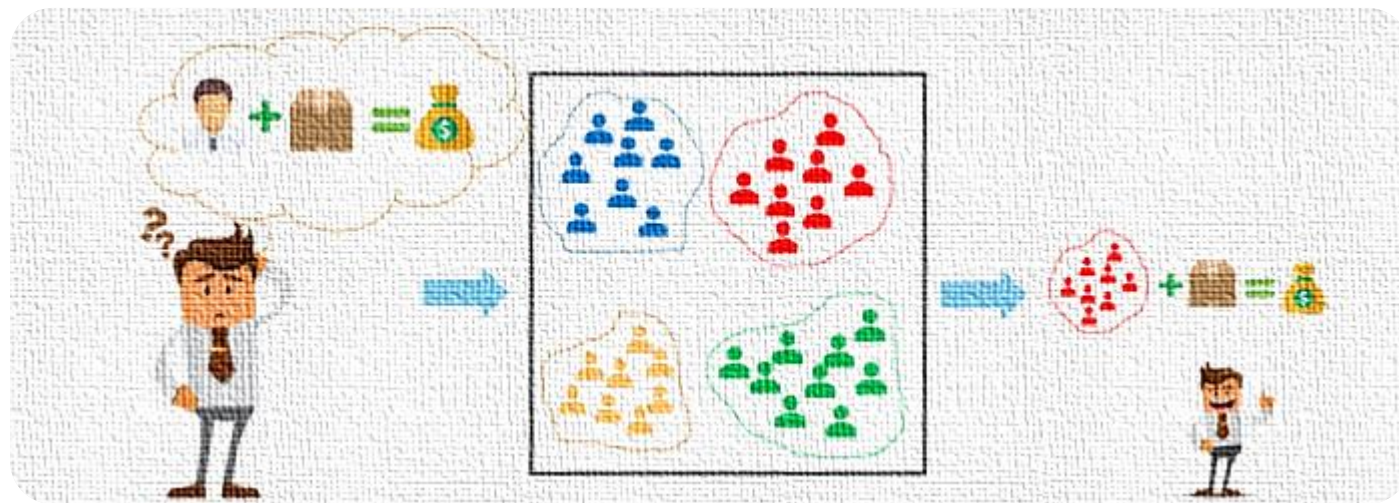


Class exercise

- Can you think of an example of Regression? Classification?

→ Unsupervised Learning

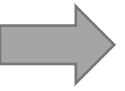
- Unsupervised learning is a type of machine learning where the algorithm is **not given any labeled** training data.
- The goal is to discover the **underlying patterns** and find groups of samples that behave similarly. **Find something interesting!**
- **Clustering**: group similar data points together
 - ✓ Mall customer segmentation
 - ✓ Client profiling and asset allocation



→ Unsupervised Learning

- Unsupervised learning is a type of machine learning where the algorithm is **not given any labeled** training data.
- **Anomaly detection:** Find anomalies (unusual data points)
 - ✓ Fraud detection
- **Dimensionality Reduction:** compress data in lower dimension
 - ✓ Identify the most predictive factors underlying asset price models.





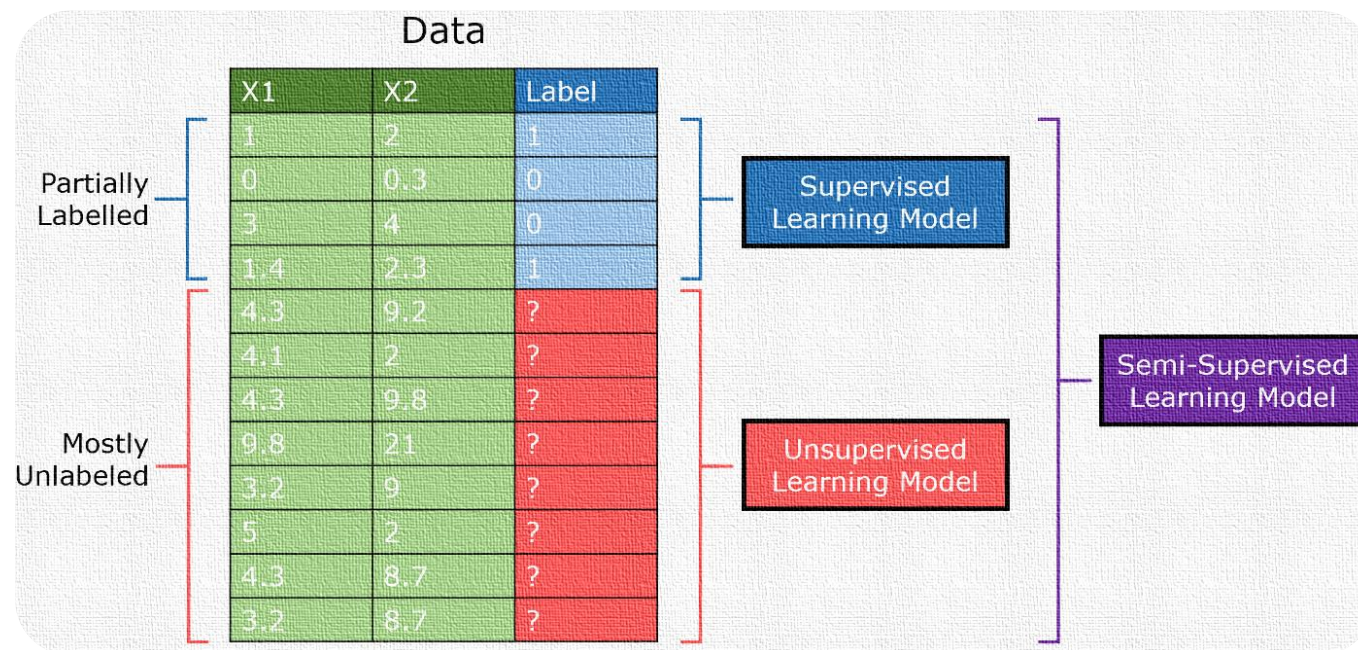
Class exercise

- Can you think of an example of Clustering? Anomaly detection?



Semi-Supervised

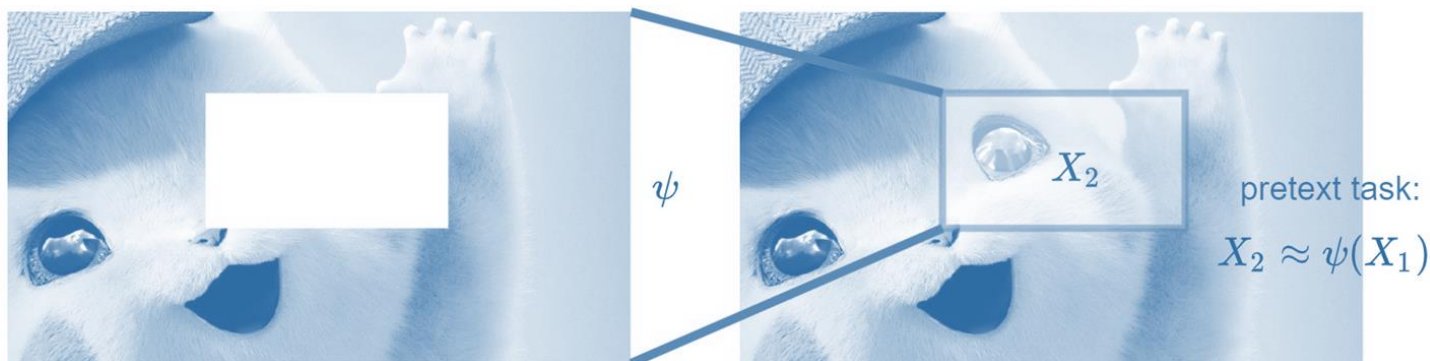
- Semi-supervised learning is a **combination** of supervised and unsupervised learning.
- It's used when you have a large dataset with some **labeled** examples and many **unlabeled** examples.
- The goal is to use the labeled examples to learn a mapping from inputs to outputs, and then use that mapping to make predictions on the unlabeled examples (**pseudo labels**)





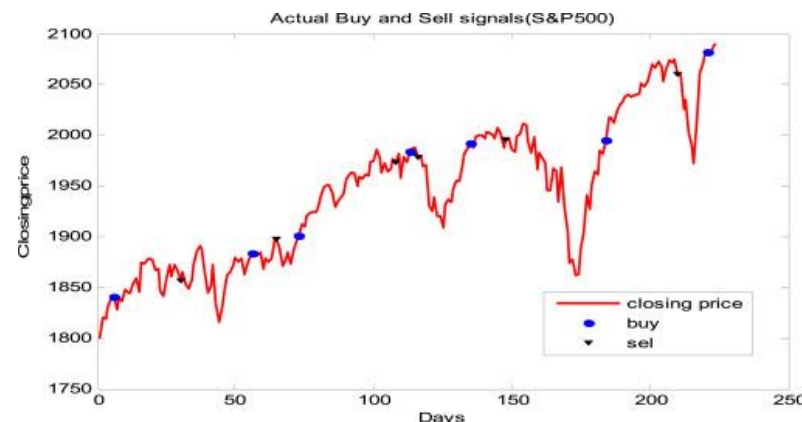
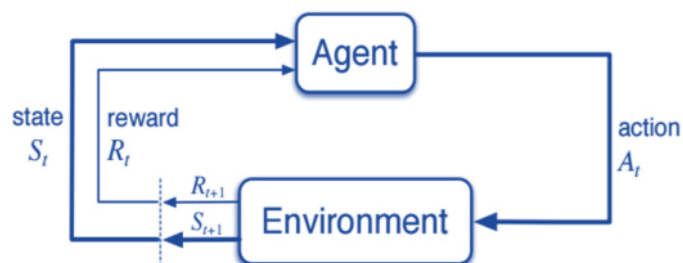
Self-Supervised

- Self-supervised learning is a type of **unsupervised** learning, but it has the property that the learning process is being fed with the data itself (and not a human annotation).
- The **goal** of self-supervised learning is learning useful **representations** from the data (representation learning)
- The model learns a representation of the data by predicting properties of the **input data itself**.
- Example:
 - Predicting missing part of an input (text, image, ...)

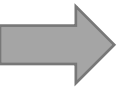


→ Reinforcement Learning

- **Reinforcement learning** is a type of machine learning where an **agent** learns to interact with its **environment** in order to maximize a **reward**.
- The agent receives rewards for performing actions that lead to successful outcomes and learns to repeat successful actions and avoid unsuccessful ones. (**Explore** and **Exploit**)



- Example: a virtual trader (**agent**) who follows certain trading rules (**actions**) in a specific market (**environment**) to maximize its profits (**reward**).

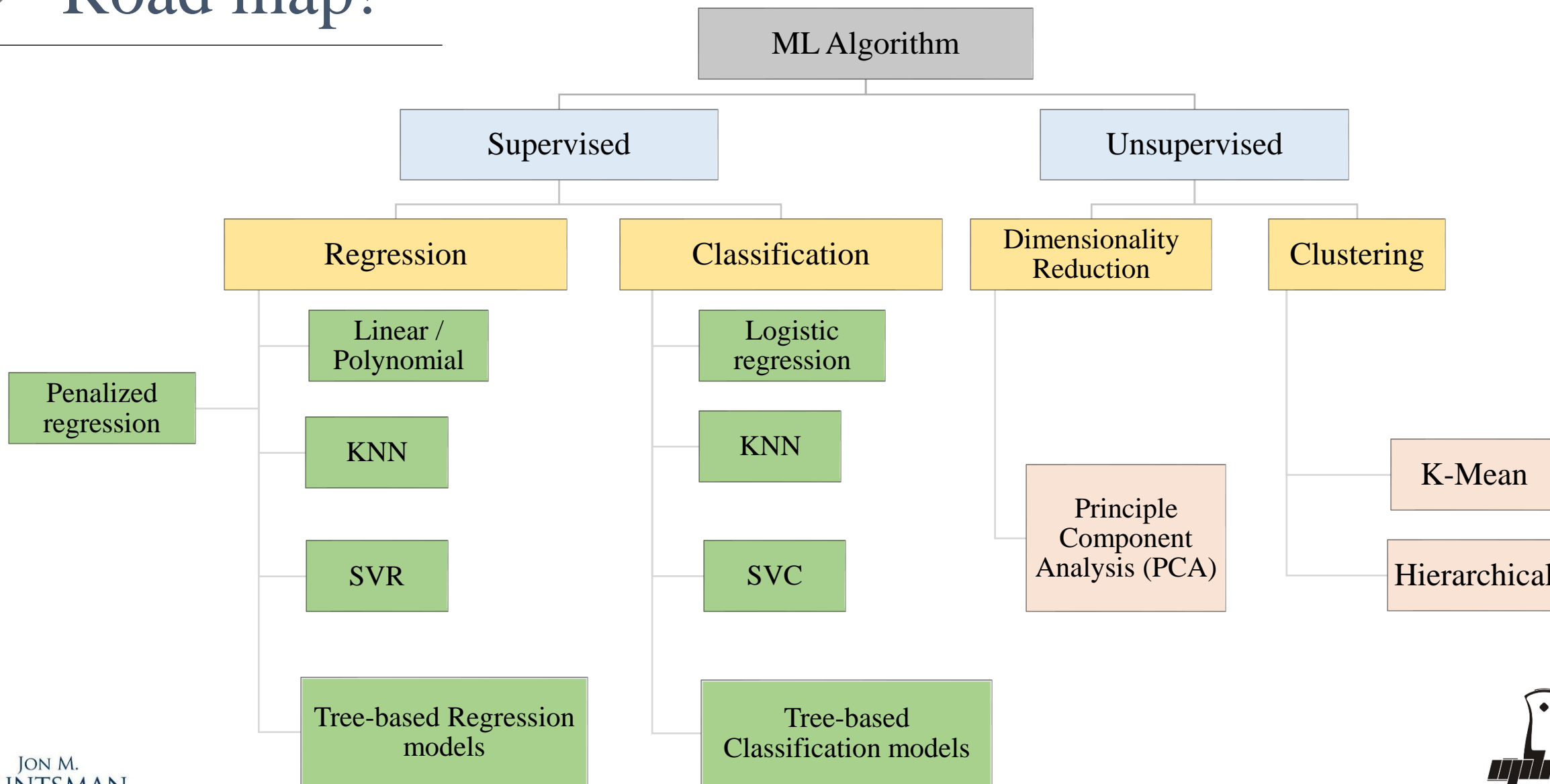


Class exercise

- Can you think of an example of reinforcement learning?



Road map!

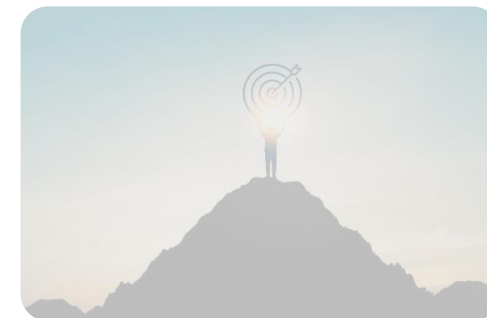


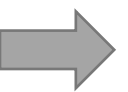


What's on GitHub

✓ Module 1- Introduction to Machine Learning

- Module 2- Setting up Machine Learning Environment
- Module 3- Linear Regression (Econometrics approach)
- Module 4- Machine Learning Fundamentals
- Module 5- Linear Regression (Machine Learning approach)
- Module 6- Penalized Regression (Ridge, LASSO, Elastic Net)
- Module 7- Logistic Regression
- Module 8- K-Nearest Neighbors (KNN)
- Module 9- Classification and Regression Trees (CART)
- Module 10- Bagging and Boosting
- Module 11- Dimensionality Reduction (PCA)
- Module 12- Clustering (KMeans – Hierarchical)





Having said that...

- **Warning:** A ML algorithm will always find a pattern, even if there is none.

