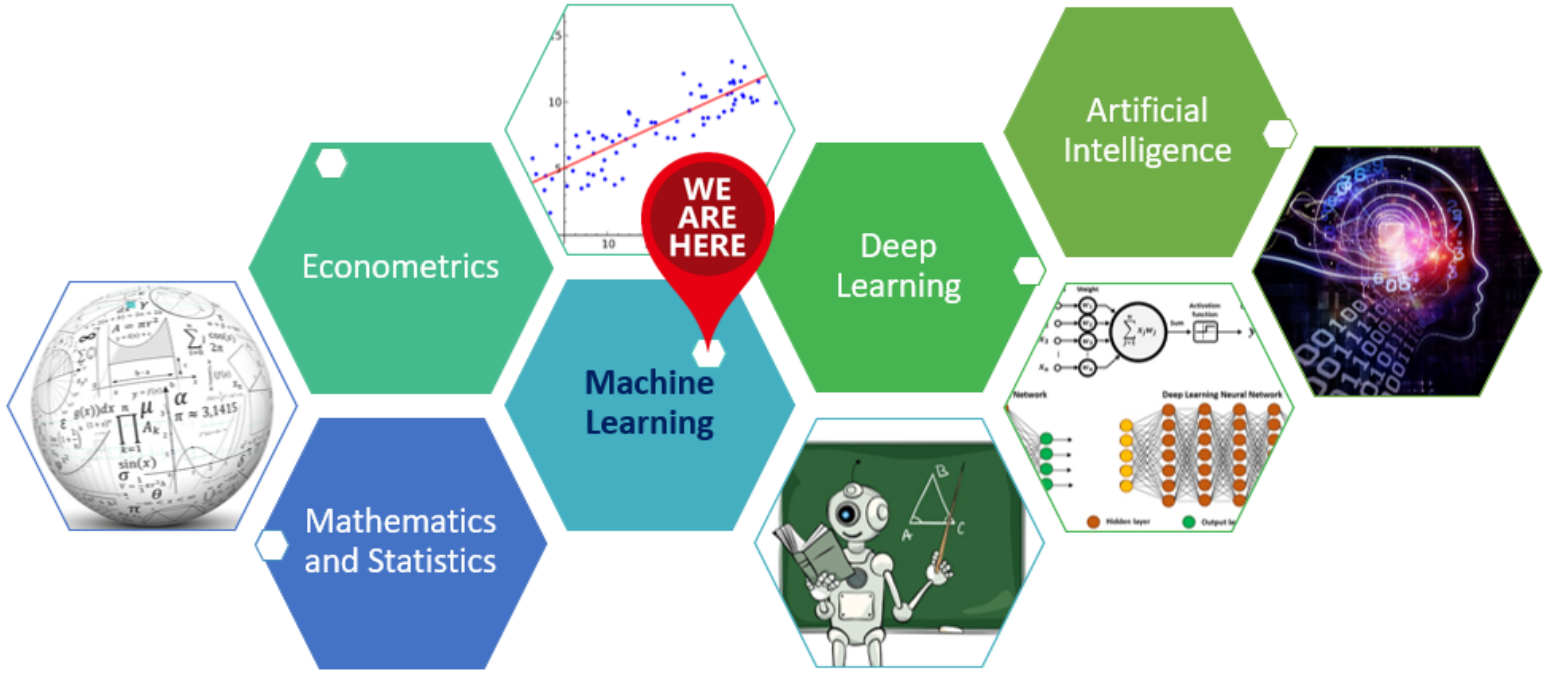
Module 1
Introduction to Machine Learning



Part 1 –Machine Learning vs Statistical Learning
The BIG picture

Prof. Pedram Jahangiry

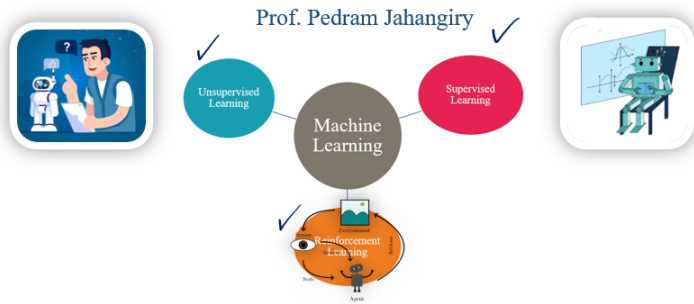
	Statistical Learning	Machine Learning
Focus	Hypothesis testing & interpretability	Predictive accuracy
Driver	Math, theory, hypothesis	Fitting data
Data size	Any reasonable set	Big data
Data type	Structured	Structured, unstructured, semi-structured
Dimensions / scalability	Mostly low dimensional data	High dimensional data
Model choice	Parameter significance & in-sample goodness of fit	Cross-validation of predictive accuracy on partitions of data
Interpretability	High	Low
Strength	Understand **causal** relationship & behavior	Prediction (forecasting and nowcasting)

Part 2 – What is Machine Learning?
The definition

Prof. Pedram Jahangiry

Part 3 – Different types of machine learning Algorithms

Prof. Pedram Jahangiry



Part 1 –Machine Learning vs Statistical Learning
The BIG picture

Prof. Pedram Jahangiry

	Statistical Learning	**Machine Learning**
Focus	Hypothesis testing & interpretability	Predictive accuracy
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
Model choice	Parameter significance & in-sample goodness of fit	Cross-validation of predictive accuracy on partitions of data
Interpretability	High	Low
Strength	Understand **causal** relationship & behavior	Prediction (forecasting and nowcasting)

JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Pedram Jahangiry

Big picture

What are we trying to do as a researcher or a practitioner?

Solve real world problems, right? Either making inference or predictions.

Is there a theory?

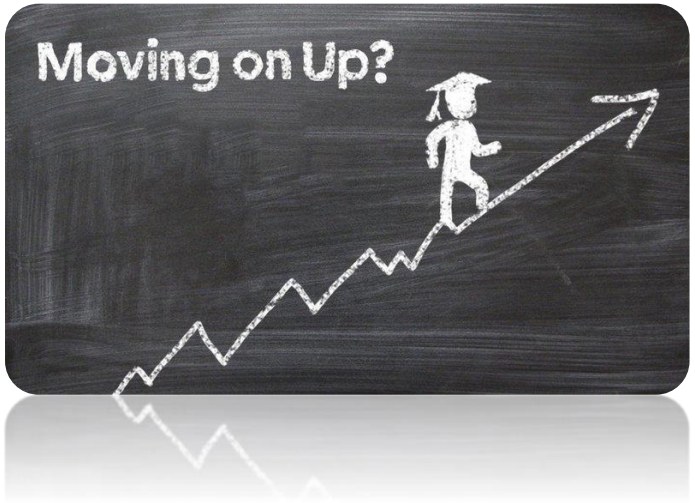
What is the relationship between

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

Jon M. Huntsman School of Business
UtahStateUniversity

Pedram Jahangiry

A simple example



Quantifying wage components?

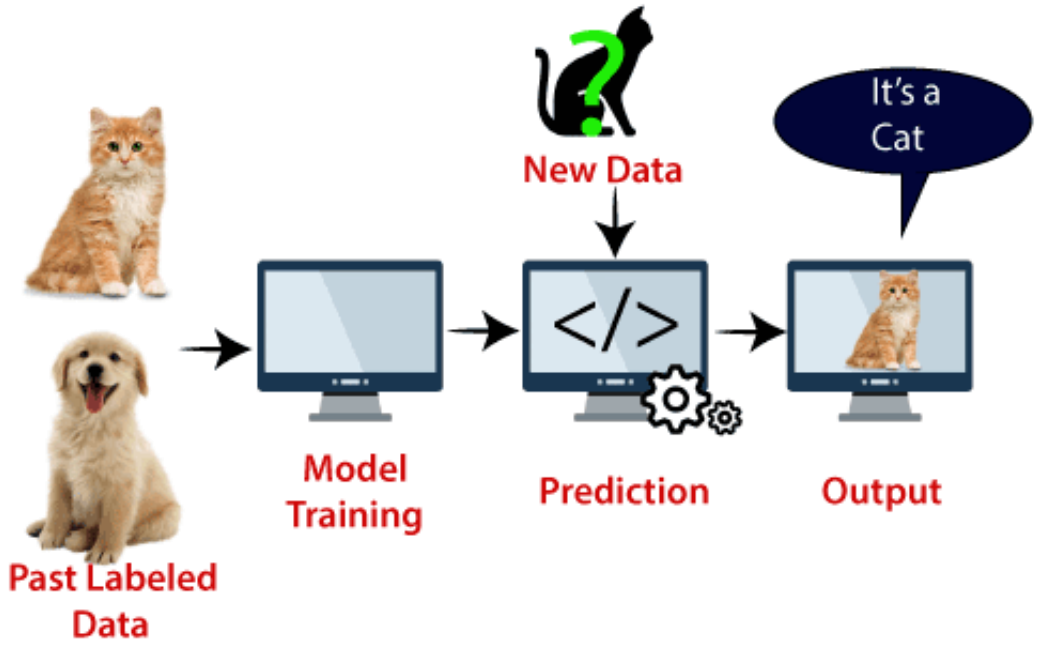
- What are the drivers:
  - Education, age, experience, IQ, …
  - Ethnicity, race, gender, …
  - Industry, location, working hours, …
- Let's build a model (assuming a linear functional form!)

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

- Can you **interpret** this model? Do you care about the interpretability?
- Can you make **predictions** using your model?

A different example

- Cat vs dog classification problem (image recognition)



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

Statistical learning vs machine learning

	Statistical Learning	**Machine Learning**
Focus	Hypothesis testing & interpretability	Predictive accuracy
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
Model choice	Parameter significance & in-sample goodness of fit	Cross-validation of predictive accuracy on partitions of data
Interpretability	High	Low
Strength	Understand **causal** relationship & behavior	Prediction (forecasting and nowcasting)

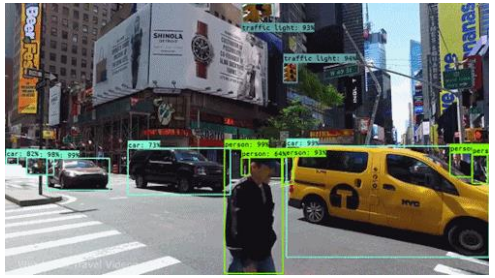
Limitations of Econometrics/Structured ML

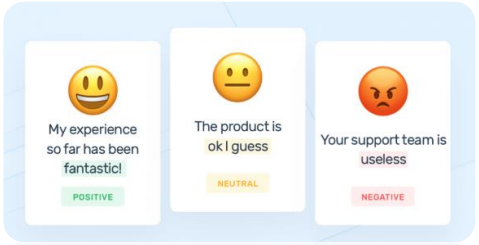
Econometrics/structured ML can only handle structured data (tabular data)!

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!!)





Jon M. Huntsman School of Business
UtahStateUniversity

Pedram Jahangiry

A more complex example

Stock price prediction $$$

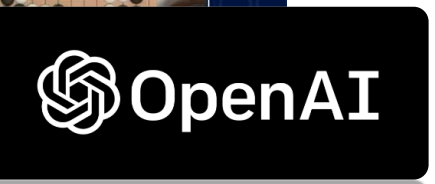
- What are the classical drivers:
  - Company's fundamentals (balance sheet, income statement, cash flow statement)
  - Competitors (comparing multiples)
  - Technical analysis!
  - Seasonality (holidays, months, days, …)

What else?

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

Why should I learn it?

- It’s a big deal, deep learning is everywhere!
- Better career opportunities
- Hedge against next recession



Part 2 – What is Machine Learning? The definition

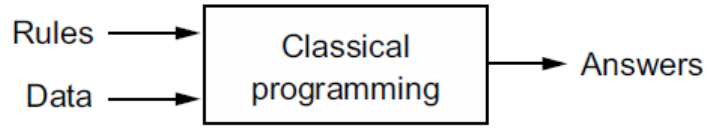
Prof. Pedram Jahangiry

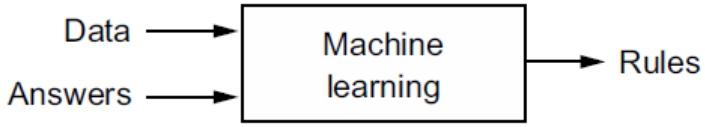
JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Pedram Jahangiry

What is Machine Learning?

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



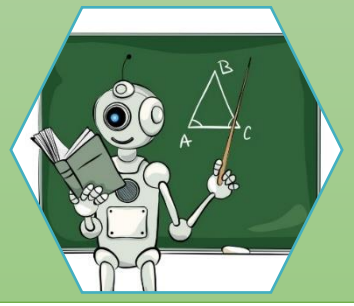


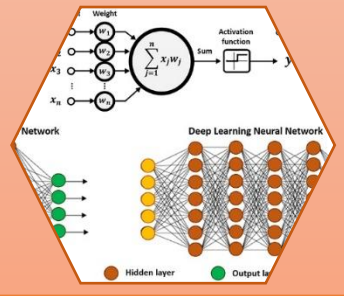
Artificial intelligence vs Machine learning vs Deep learning

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

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

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

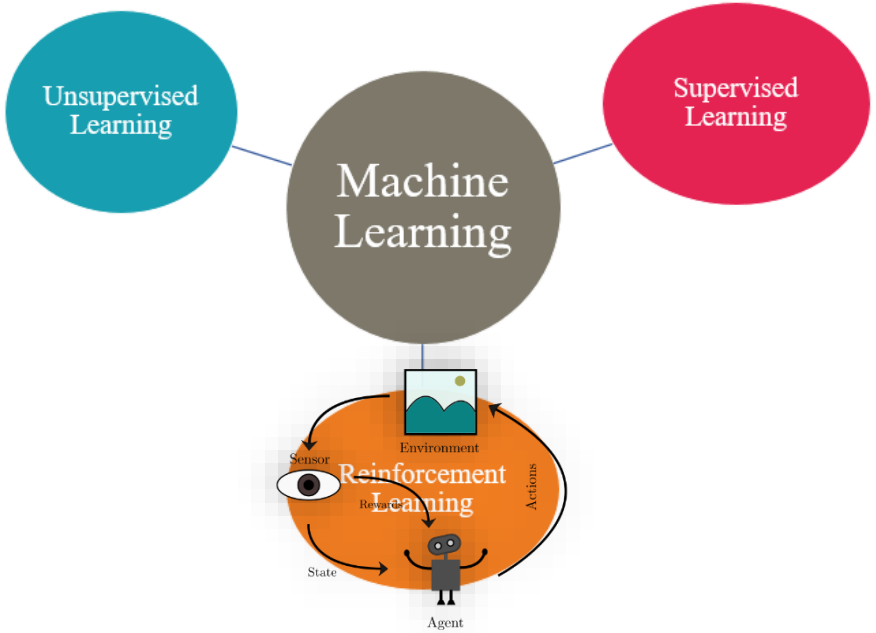


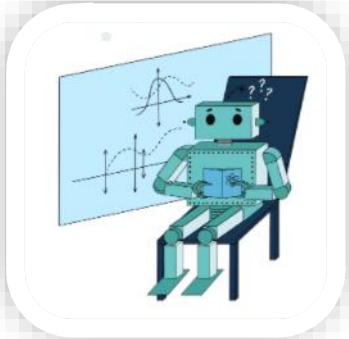


1950's 1980's 2010's

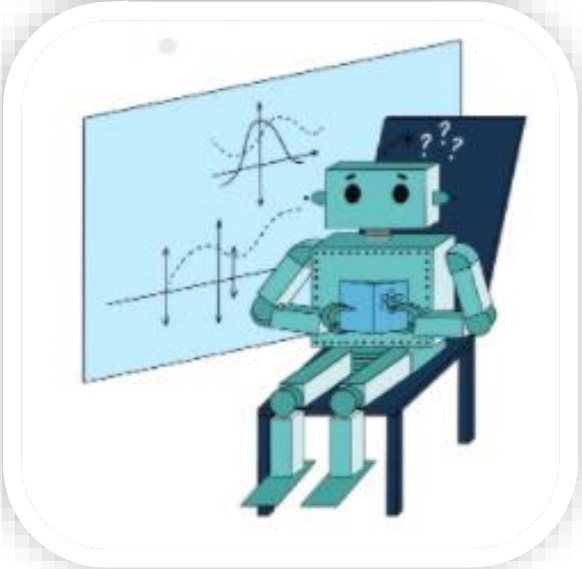
Part 3 – Different types of machine learning Algorithms

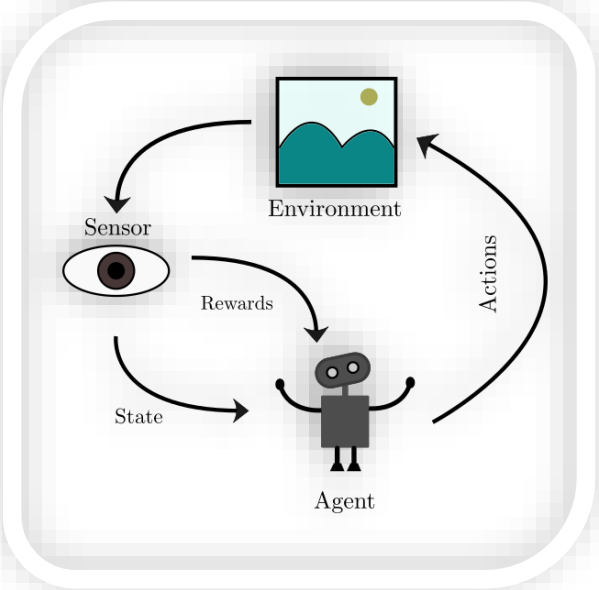
Prof. Pedram Jahangiry





Types of Machine Learning Algorithms





Supervised Learning

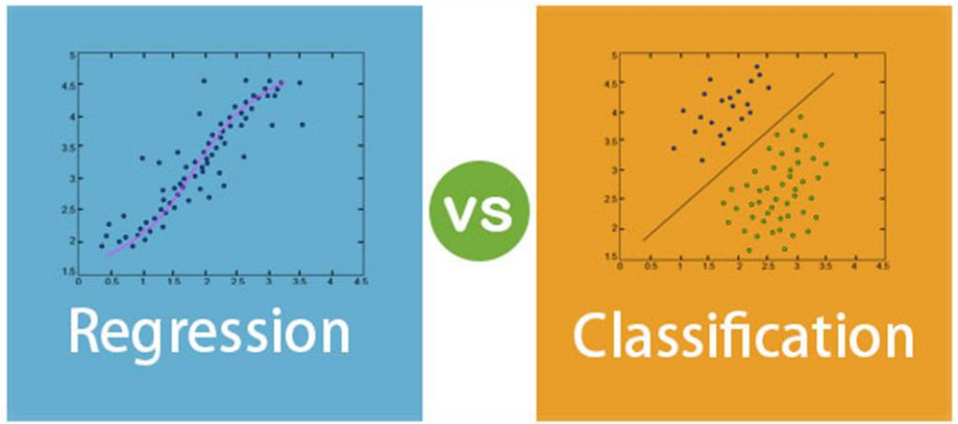
- In **supervised learning**, computers learn to model relationships based on train data. In supervised learning, inputs and outputs are labeled for the algorithm. After learning the pattern, the trained algorithms are used to predict outcomes for test data.

- **Regression:**

1. Predicting stock market returns
2. Predicting housing prices
3. ….

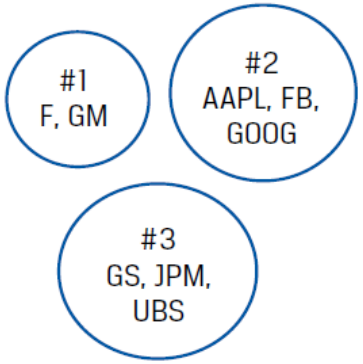
- **Classification:**

1. Generating buy, sell, hold signals.
2. Estimating the likelihood of a successful M&A or IPO
3. Predicting credit default rate.
4. Classification on winning and losing funds or ETFs
5. …



Unsupervised Learning

- In **unsupervised learning**, computers are trained on unlabeled train data without any guidance. The goal is to discover the underlying patterns and find groups of samples that behave similarly. Examples:

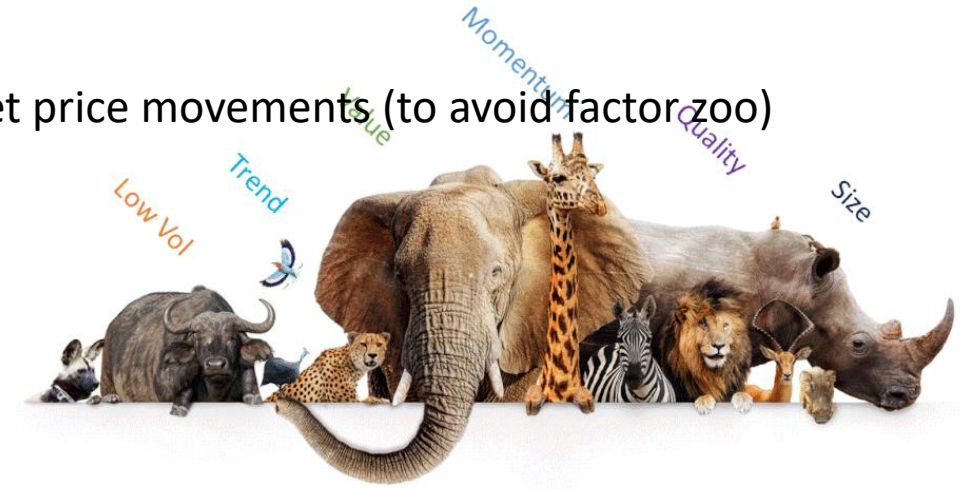


- **Clustering:**

1. Grouping companies into peer groups based on some non-standard characteristics like financial statement data or corporate characteristics rather than sectors or countries.
2. Client profiling and asset allocation
3. Portfolio diversification and stock selection based on co-movements similarities

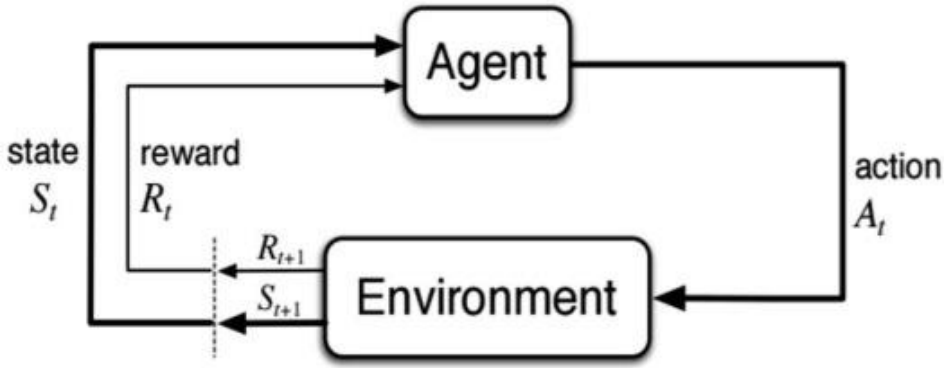
- **Dimensionality Reduction:**

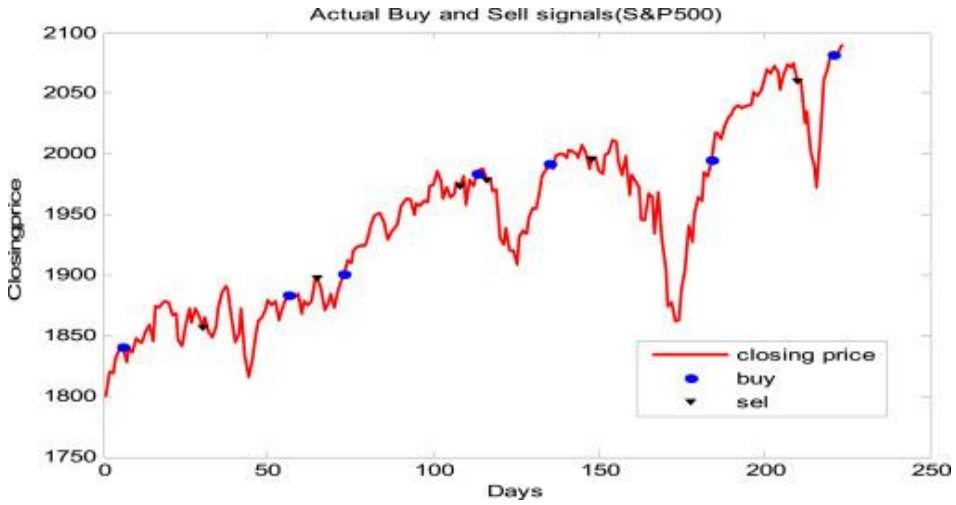
1. Identify the most predictive factors underlying asset price movements (to avoid factor zoo)



Reinforcement Learning

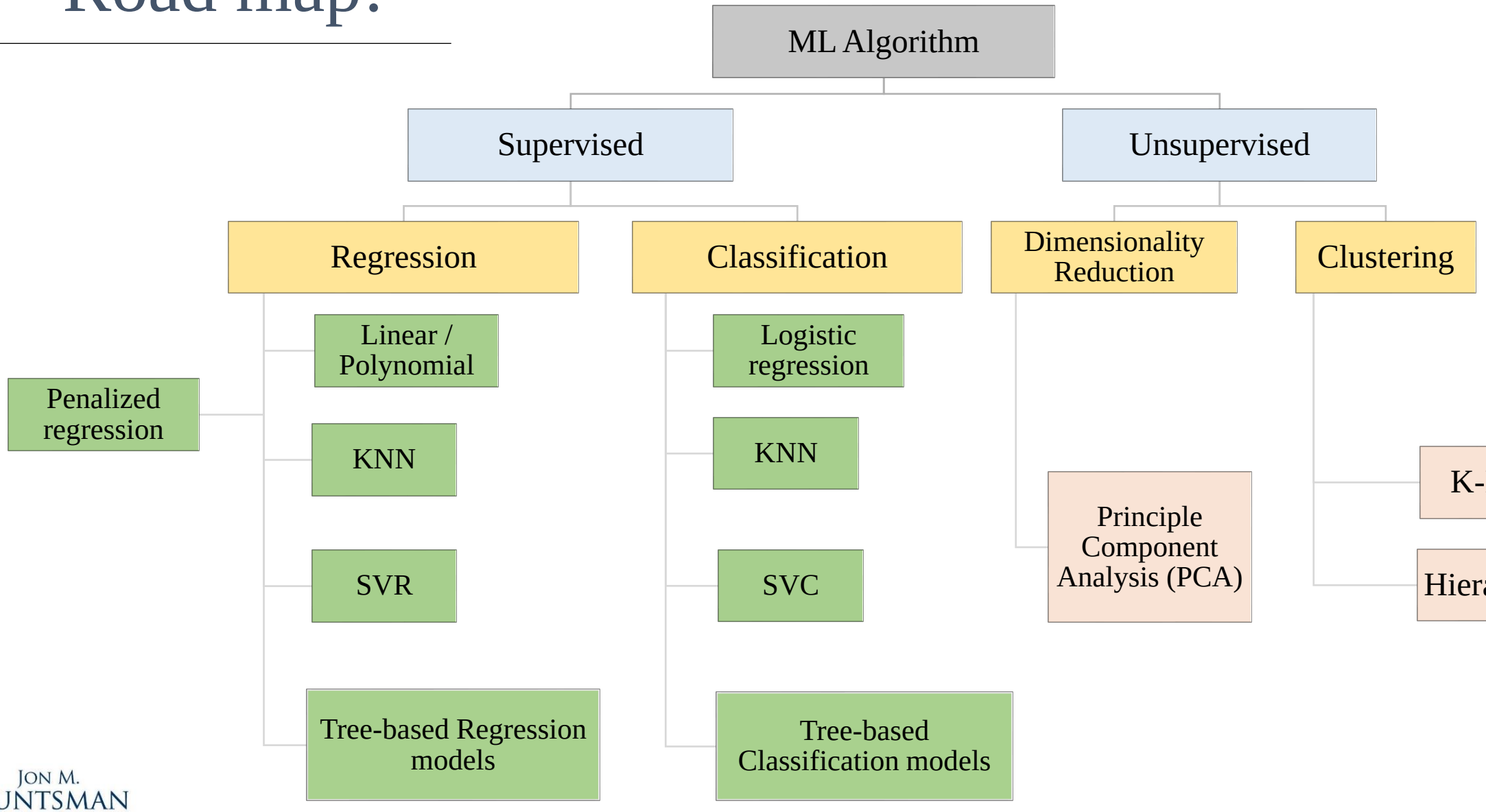
- In **reinforcement learning**, a computer (agent) learns from interacting with its environment by producing actions and discovering rewards. You need to define the environment, actions and the reward system. The machine will then explore and exploit to maximize the reward. The new actions may not be immediately optimal. The learning subsequently occurs through millions of trials and errors.





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

Road map!


ML Algorithm
Supervised
Unsupervised
Regression
Classification
Dimensionality Reduction
Clustering
Linear / Polynomial
Penalized regression
KNN
SVR
Tree-based Regression models
Logistic regression
KNN
SVC
Tree-based Classification models
Principle Component Analysis (PCA)
K-Mean
Hierarchical

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.