JOHNS HOPKINS
WHITING SCHOOL *of* ENGINEERING

EN.540.635
Software Carpentry

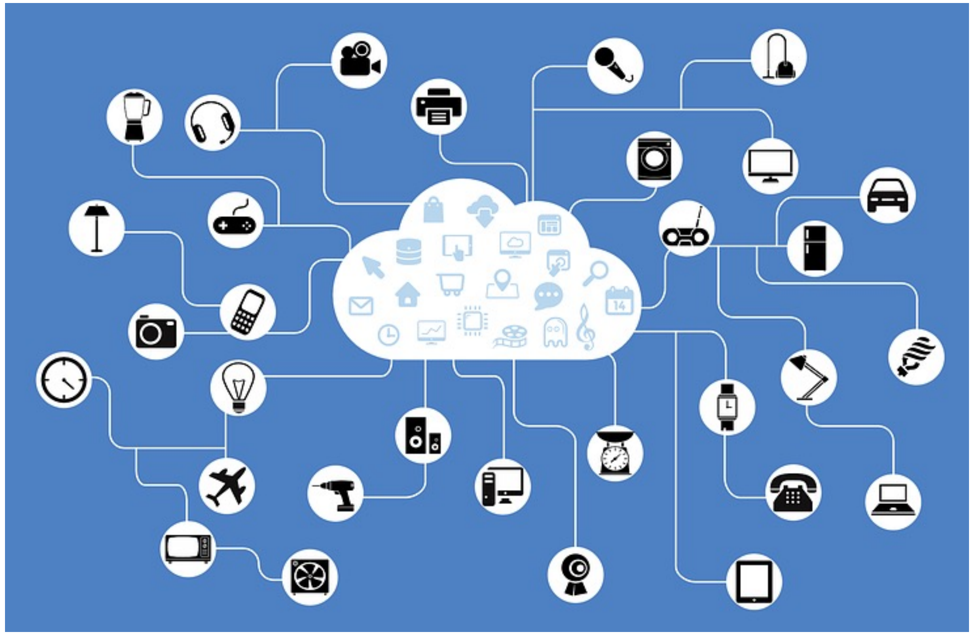
Lecture 12
Data Mining, APIs, and an Introduction to Machine Learning

What is Data Mining?

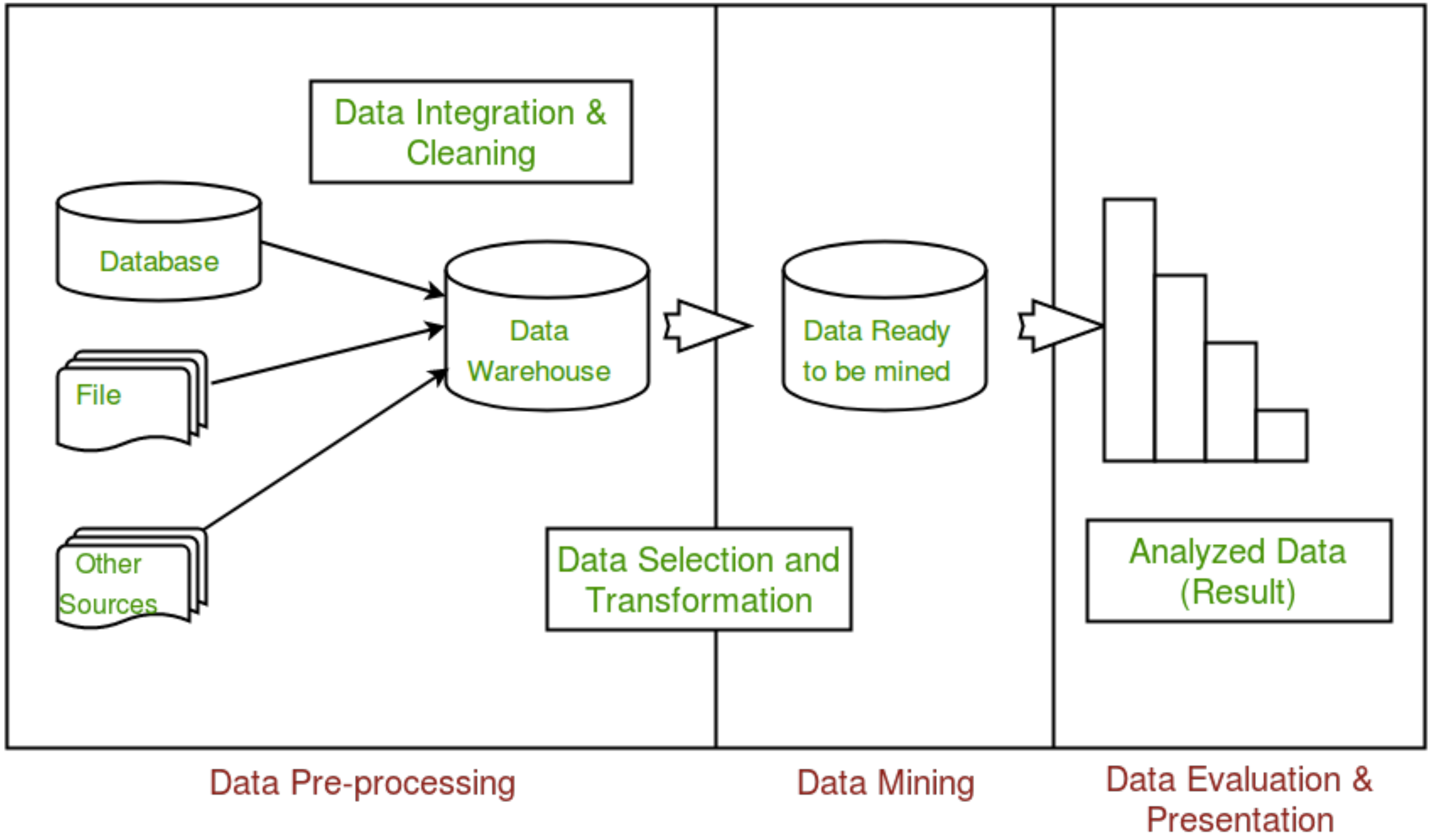
- Data mining is defined as the exploration and analysis of large quantities of data (an intersection of statistics, machine learning, and artificial intelligence).

- The goal is to extract meaningful conclusions and find patterns that exist in data.

- Applications:
  - Consumer research and marketing
  - Finance
  - Healthcare
  - Telecommunications
  - Manufacturing and engineering
  - Bioinformatics



Data Mining Processes



Common Classes of Tasks

- Anomaly Detection:
  - identification of unusual data
- Association Rule Learning:
  - searching for relationships between variables
- Clustering:
  - discovering groups and structures that are similar
- Classification:
  - applying a classifier to data
- Regression:
  - find a function that can be used to model data
- Summarization:
  - provide a more compact representation of a data set

Fayyad et al., From Data Mining to Knowledge Discovery in Databases, AI Magazine (1996).

Example – Sentiment Analysis

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

- Automated process used to identify positive, negative, and/or neutral opinions from text.

- Examples of text include social media posts, product reviews, survey responses, etc.

Example – Sentiment Analysis

- Here, we propose a data mining scenario:
    - Twitter is a widely-used social network where users give their opinions and thoughts on a broad range of topics.
    - There are millions of tweets posts sent out per day – meaning that there is a large amount of data generated and stored on Twitter's servers.
- Given a certain keyword, we parse through a collection of tweets containing that keyword and then we want to find out which other terms appear most frequently and perform sentiment analysis on the given keyword.

Example – Sentiment Analysis

- Here, we propose a data mining scenario:
    - ~~Twitter~~ X is a widely-used social network where users give their opinions and thoughts on a broad range of topics.
    - There are millions of ~~tweets~~ posts sent out per day – meaning that there is a large amount of data generated and stored on ~~Twitter's~~ X's servers.
- Given a certain keyword, we parse through a collection of tweets containing that keyword and then we want to find out which other terms appear most frequently and perform sentiment analysis on the given keyword.

How do we Access Tweets?

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

- Application Programming Interface (API):
  - A set of functions and procedures (protocol) allowing the creation of applications that access the features or data of an operating system, application, or other service.
  - Abstract the underlying implementation so that the client only has access to necessary functions and nothing else.
- Many companies have APIs for interfacing with their own codebase. For example, X has APIs for streaming and searching for posts.

Example – Sentiment Analysis

- Pointwise Mutual Information – measure of how associated two words are:

$$\mathrm{PMI}(word_1, word_2) = \log_2\left(\frac{\mathrm{p}(word_1 \,\&\, word_2)}{\mathrm{p}(word_1)\,\mathrm{p}(word_2)}\right)$$

- Semantic Orientation – the difference between a word's associations with positive and negative words:

$$SO(w) = \sum_{w\prime \in V^+} PMI(w, w') - \sum_{w\prime \in V^-} PMI(w, w')$$

  - V+ corresponds to a list of positive words.
  - V- corresponds to a list of negative words.

Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

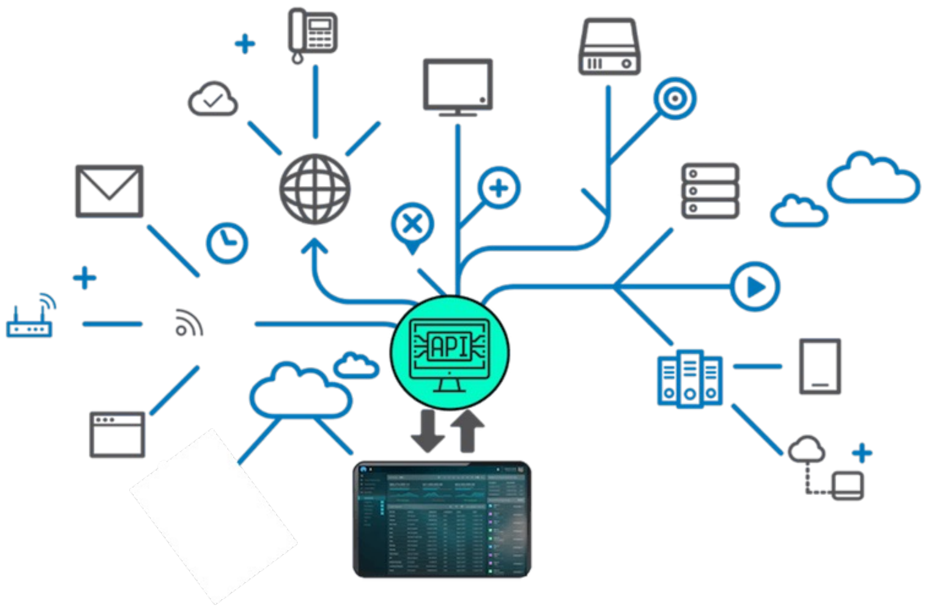
Peter D. Turney
Institute for Information Technology
National Research Council of Canada
Ottawa, Ontario, Canada, K1A 0R6
peter.turney@nrc.ca

Application Programming Interfaces

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

- APIs allow for different software/applications to communicate with one another (programming, internet, operating systems).

- Common examples:
  - Ridesharing (Uber, Lyft)
  - Application development (iOS, Android)
  - Mobile weather applications
  - Web development (embedding applications)



- Internal vs. External APIs

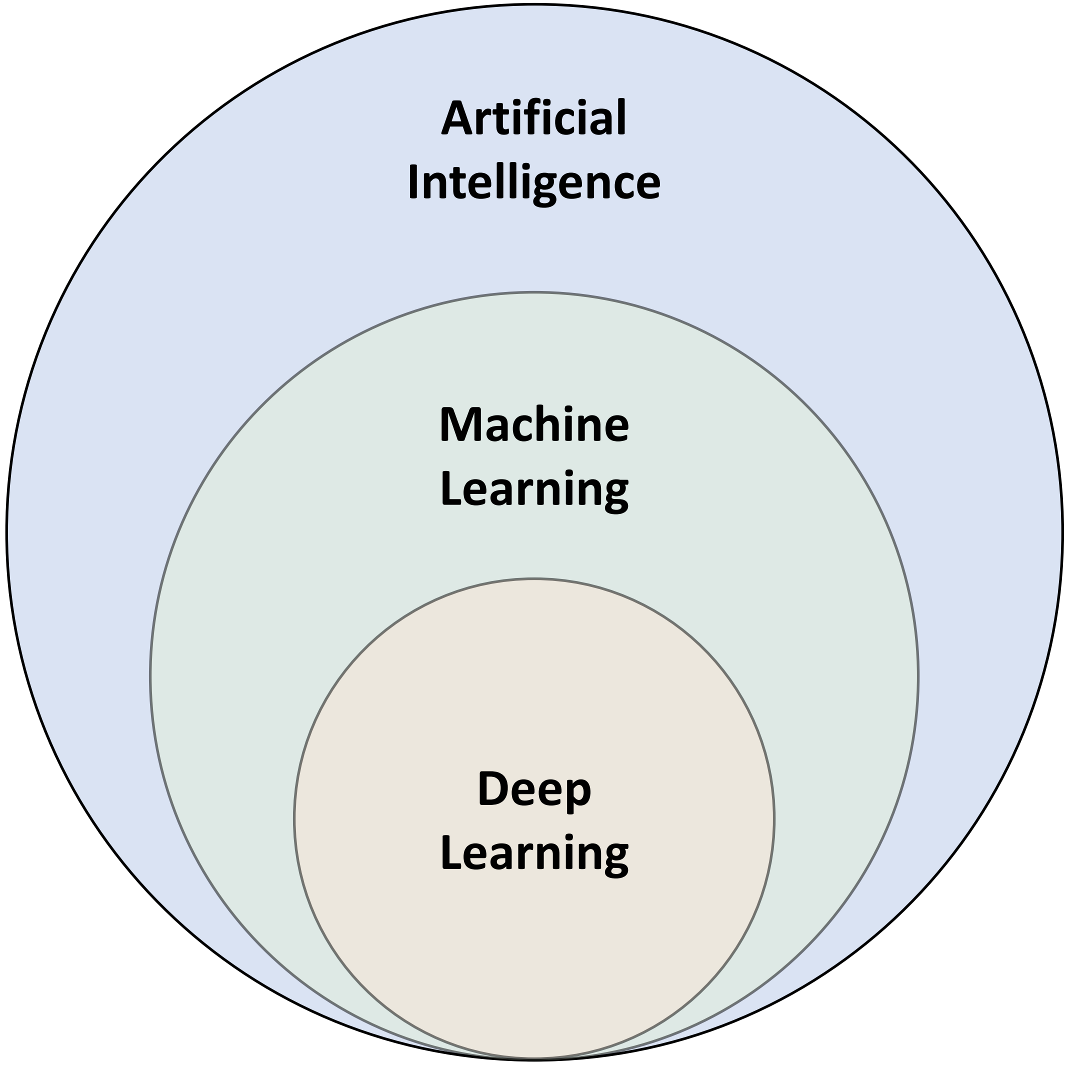
Simple Example of Using APIs

Summary

- With data mining, we can learn from large amounts of data and gain insights on patterns that occur in the data.

- Can we extend data mining techniques further?

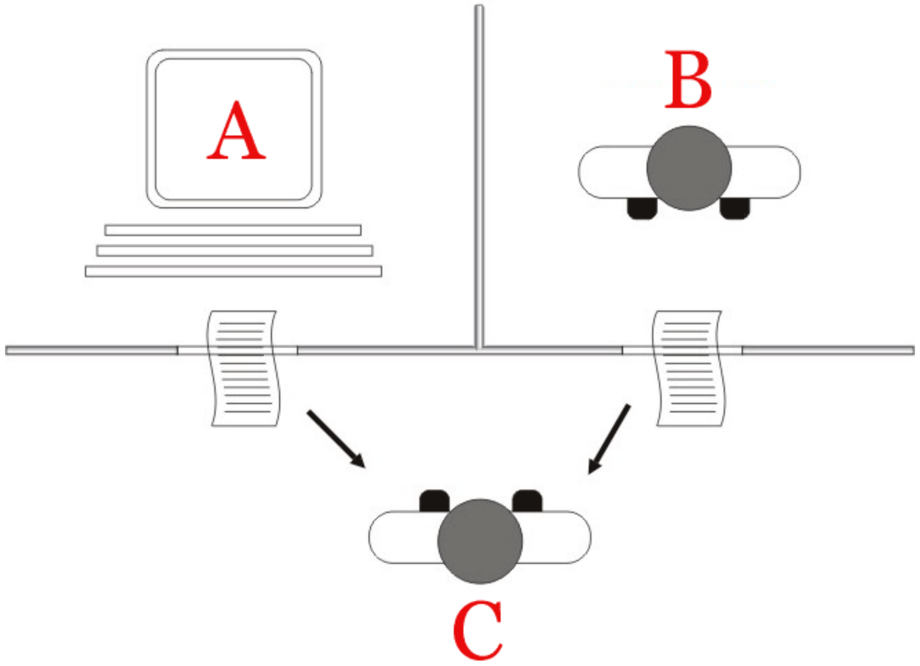
AI, ML, Deep Learning

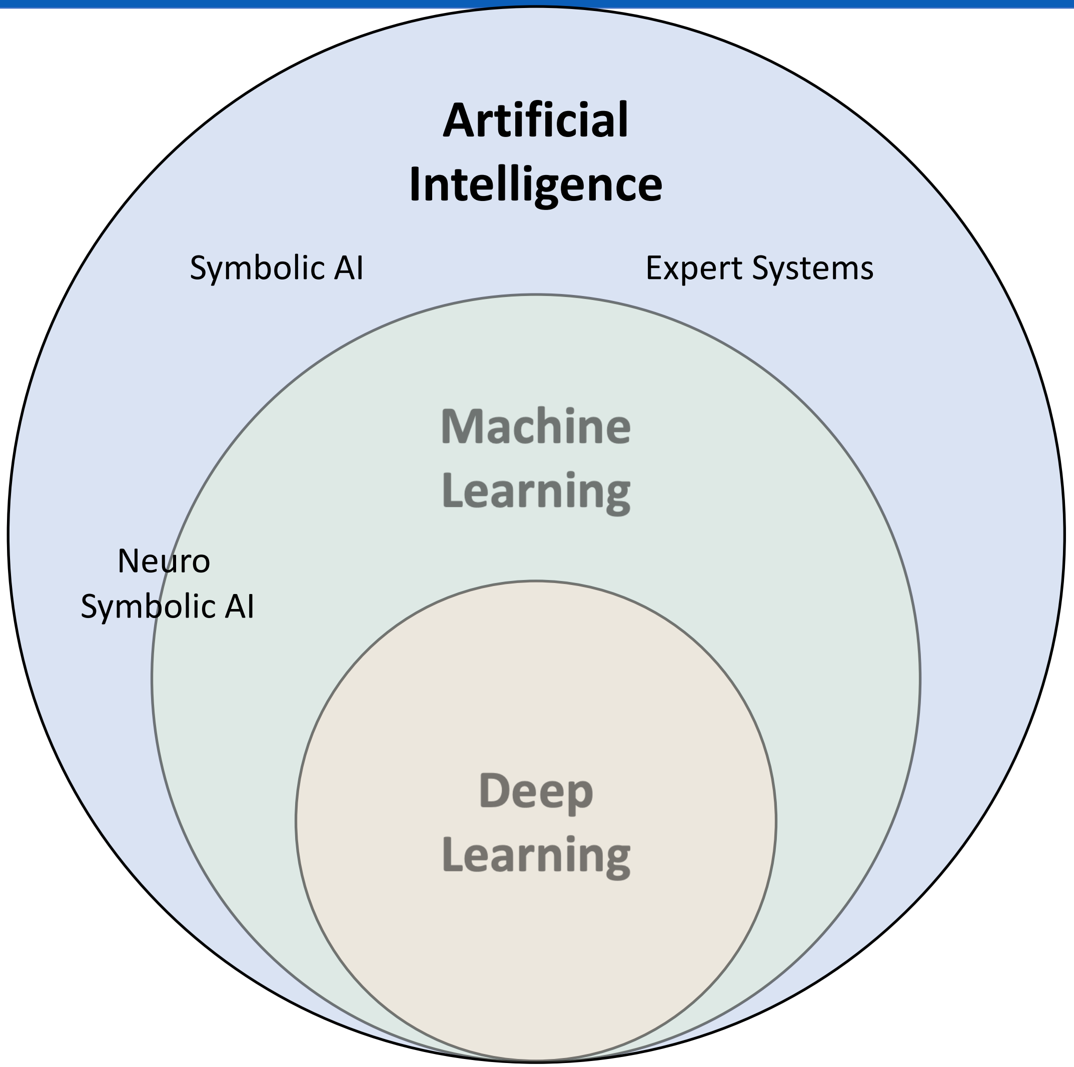


Artificial Intelligence

Theory and development of computer systems to perform tasks requiring human intelligence [1]

Can machines think ? [1]





[1] Turing, A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind* **LIX**, 433–460 (1950).
[2] Russell, S. J., Norvig, P. & Davis, E. *Artificial intelligence: a modern approach*. (Prentice Hall, 2010).
[3] Garnelo, M. & Shanahan, M. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences* **29**, 17–23 (2019).

Artificial Intelligence

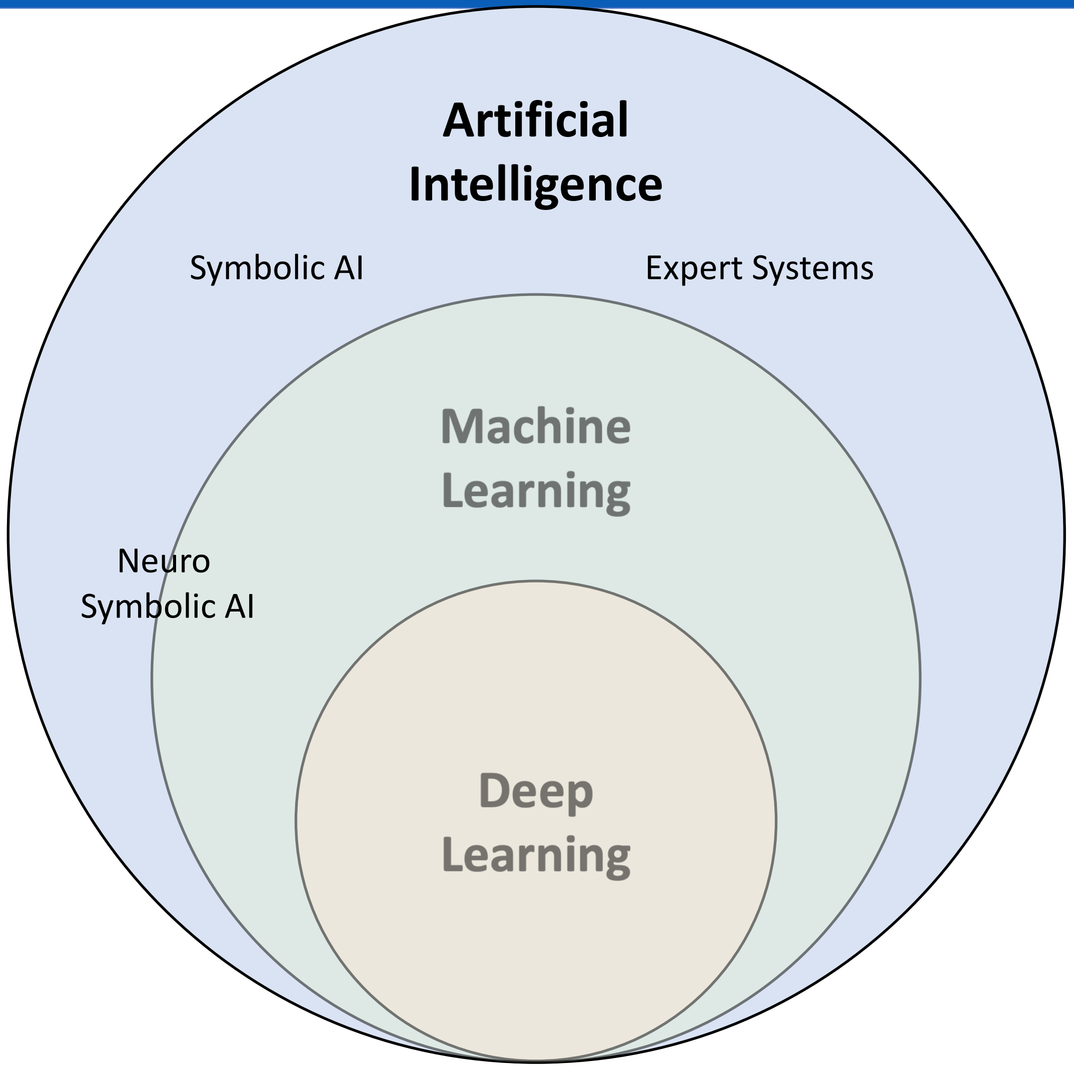
Theory and development of computer systems to perform tasks requiring human intelligence [1]

Can machines think ? [1]

Symbolic AI : Collection of methods that carry out logical reasoning by symbolic representations of problems. [2]

Expert Systems : Computer programs that emulate human decision making through large bodies of conditional statements (Rule based systems)

Neuro-symbolic AI : Combines strength of symbolic AI with deep learning architectures.[3]



[1] Turing, A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind* **LIX**, 433–460 (1950).
[2] Russell, S. J., Norvig, P. & Davis, E. *Artificial intelligence: a modern approach*. (Prentice Hall, 2010).
[3] Garnelo, M. & Shanahan, M. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences* **29**, 17–23 (2019).

Machine Learning

Use and development of computer systems that are able to learn and adapt without following explicit instructions. (Oxford dictionary)

IMPORTANT ML LINGO

Feature : Measurable property

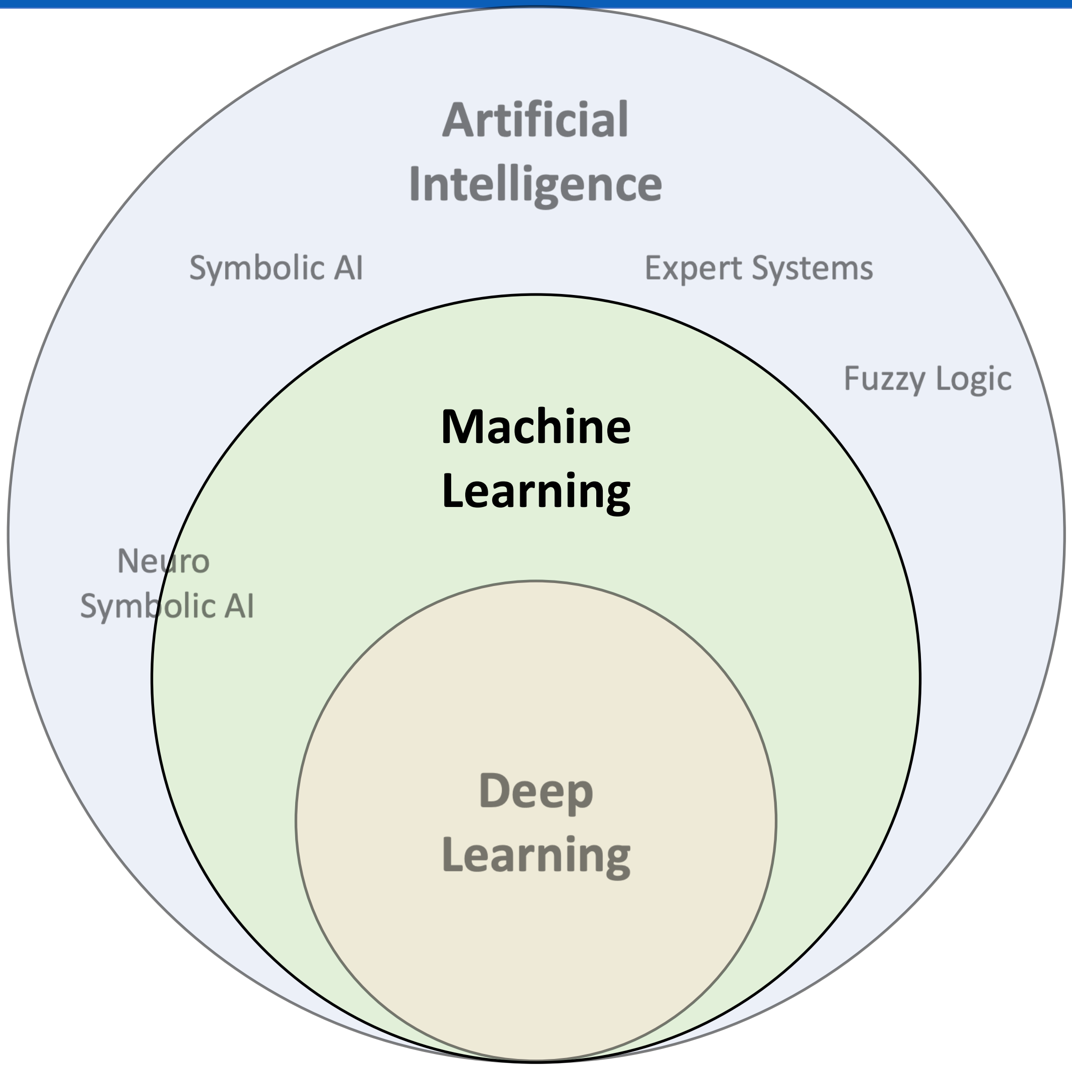
Labels : Value associated with a feature or set of features

Dataset : Set of features and labels.

Training set : Set of features and labels the model is trained on.

Test set : Set of features and labels that we use to evaluate the model predictions.

Hyperparameters : Model parameters



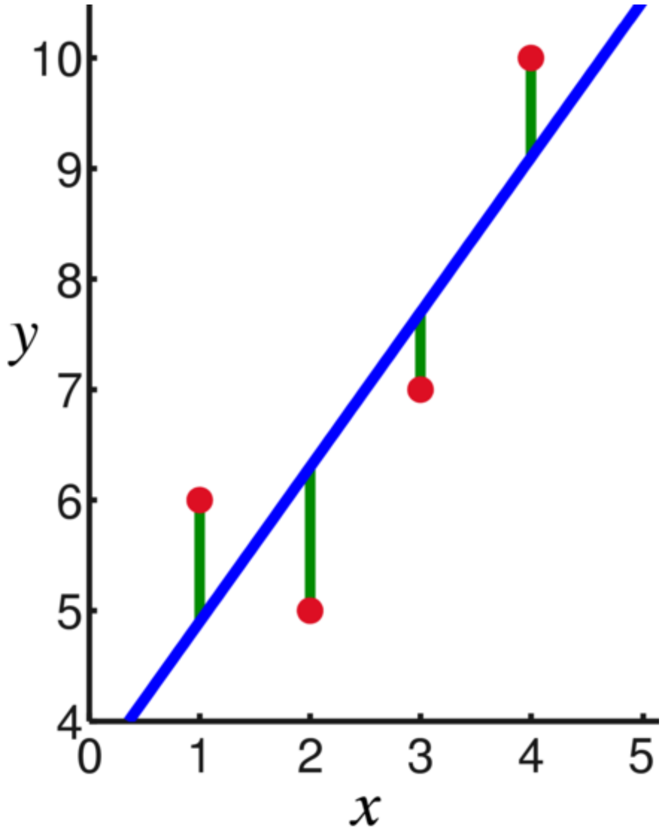
Linear Regression

The goal:

- Given sampled points, can we find a best-fit line?

The process:

- Using two parameters (m and b), find the line that minimizes the deviation.



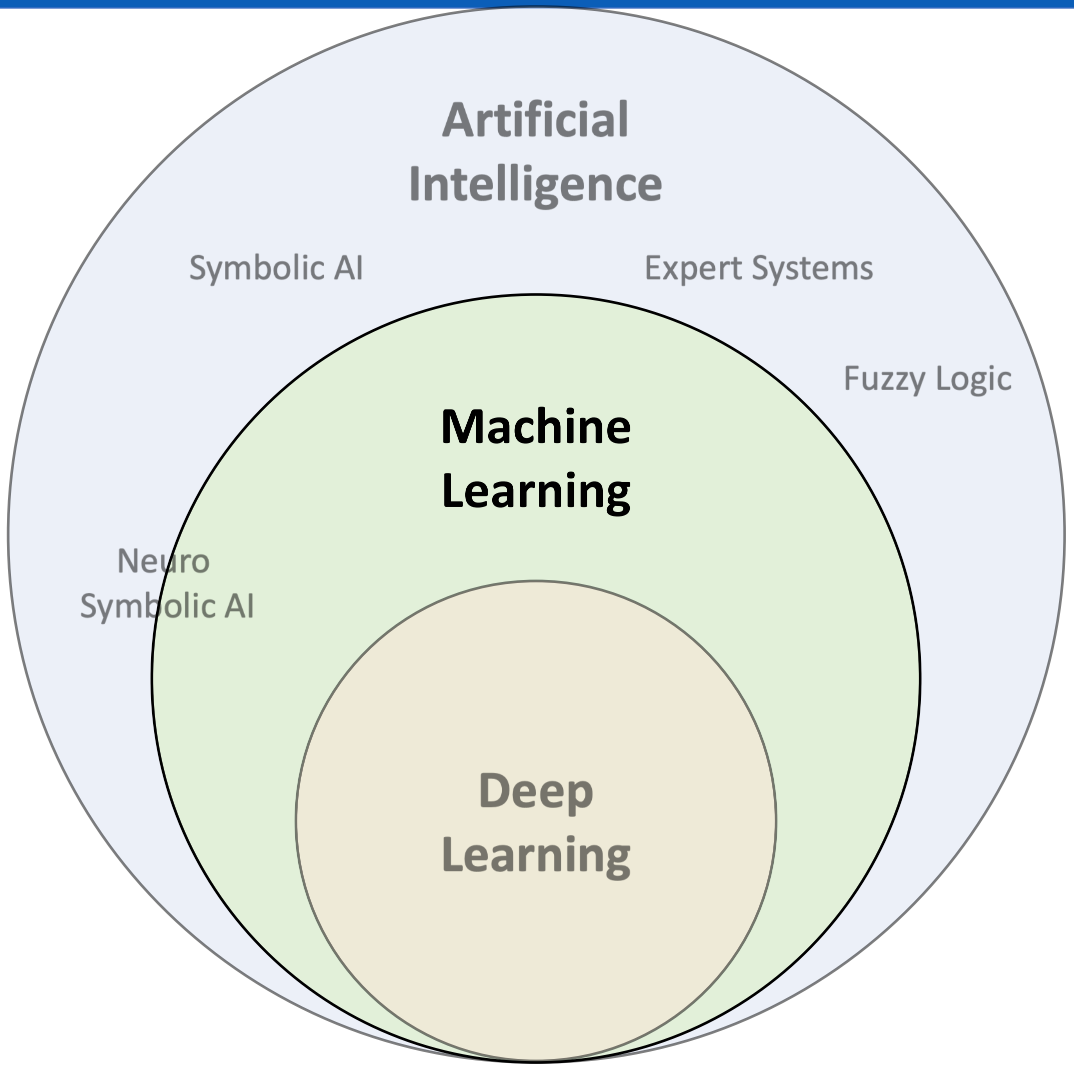
Numerical Data!

Task : Predict the number of fruits sold

	Fruit (Feature 1)	Day of week (Feature 2)	Price (in $) (Feature 3)	Num sold (Label)
Training Set	'apple'	'Monday'	1	20
	'banana'	'Monday'	0.30	20
	'apple'	'Tuesday'	1	10
Test Set	'pear'	'Wednesday'	1.5	5
	'oranges'	'Thursday'	1.5	10

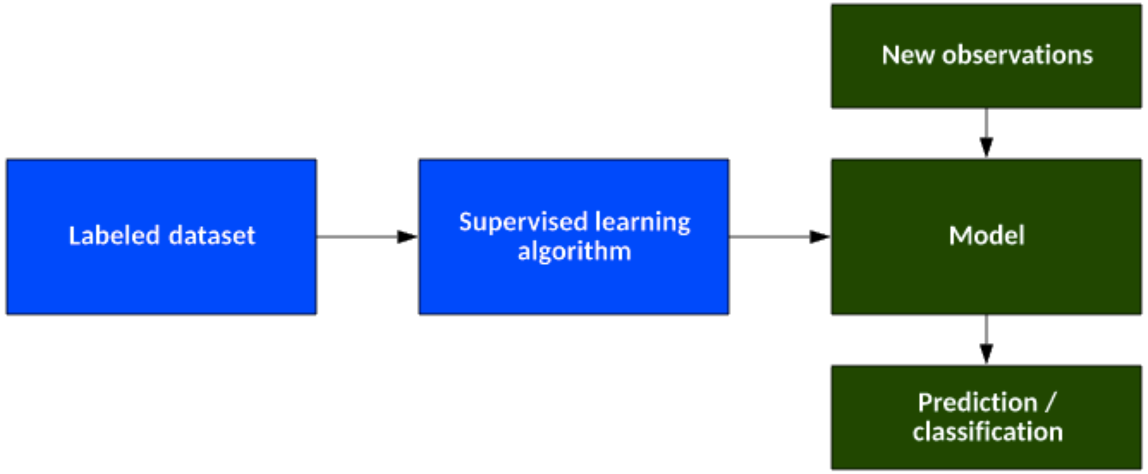
Note :

- A machine learning algorithm can only work with numerical data.
- Need to convert all the categories of fruits and days of weeks into numerical vectors (One hot encoding)



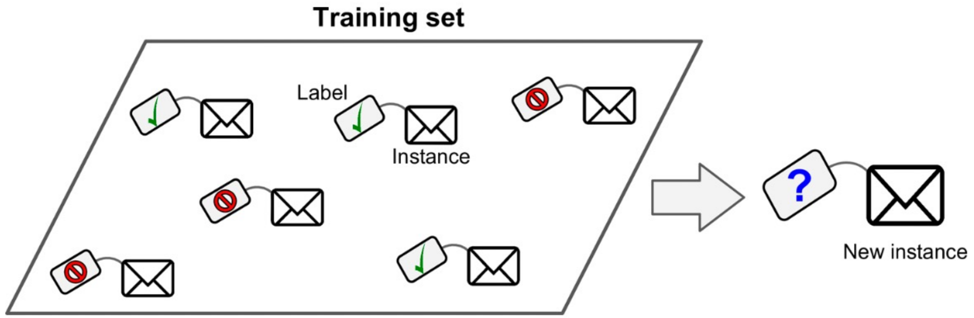
Supervised Learning

- We have a dataset with inputs and desired outputs.
- We train a model to predict the output from the input (normally using optimization techniques to minimize an error function).
- We use a different data set to test the model and see how accurate it is.

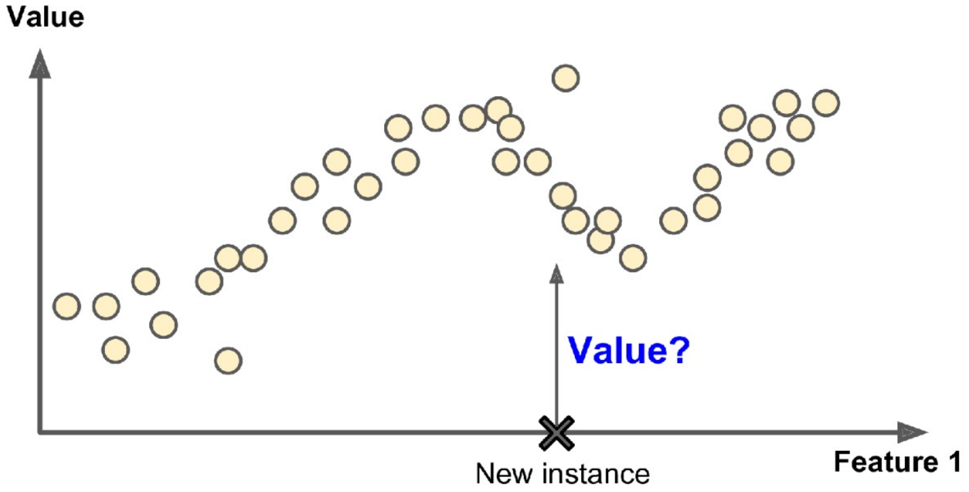


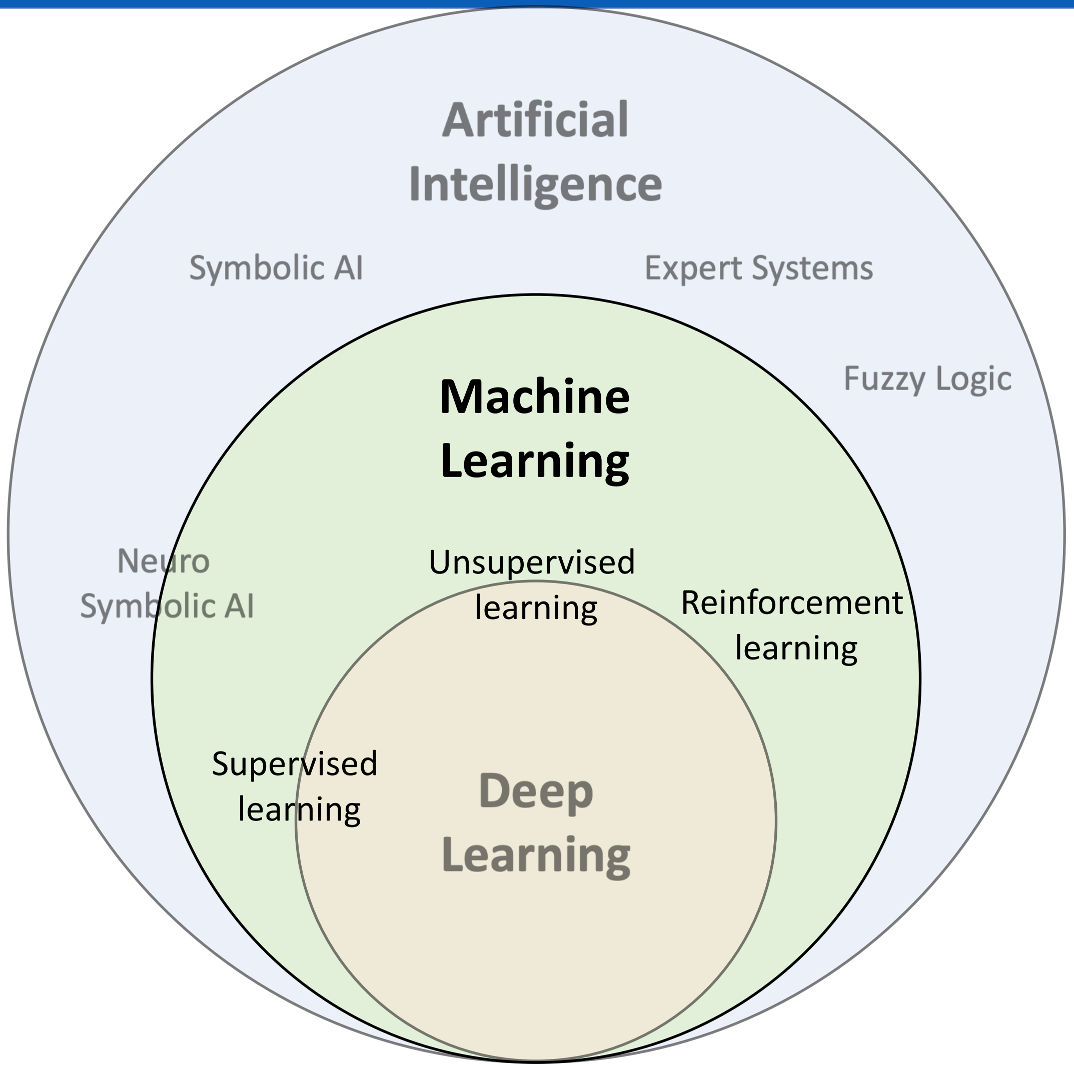
Supervised Learning

Classification



Regression



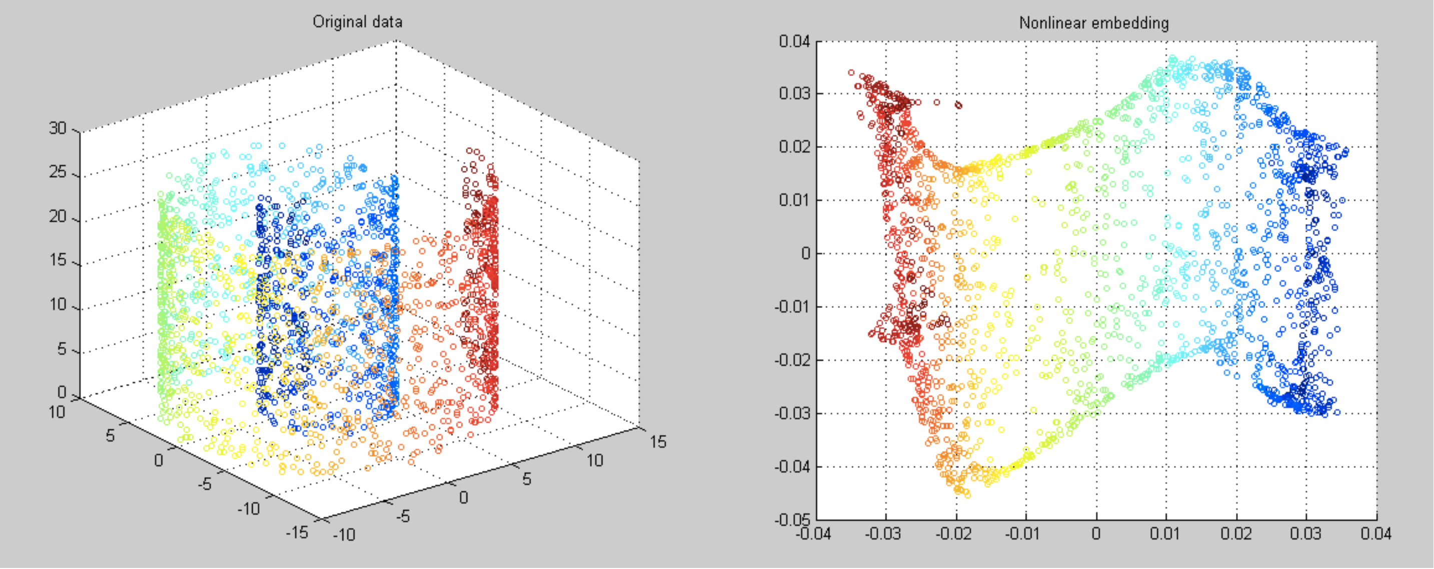


[1] Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 1150.

Unsupervised Learning

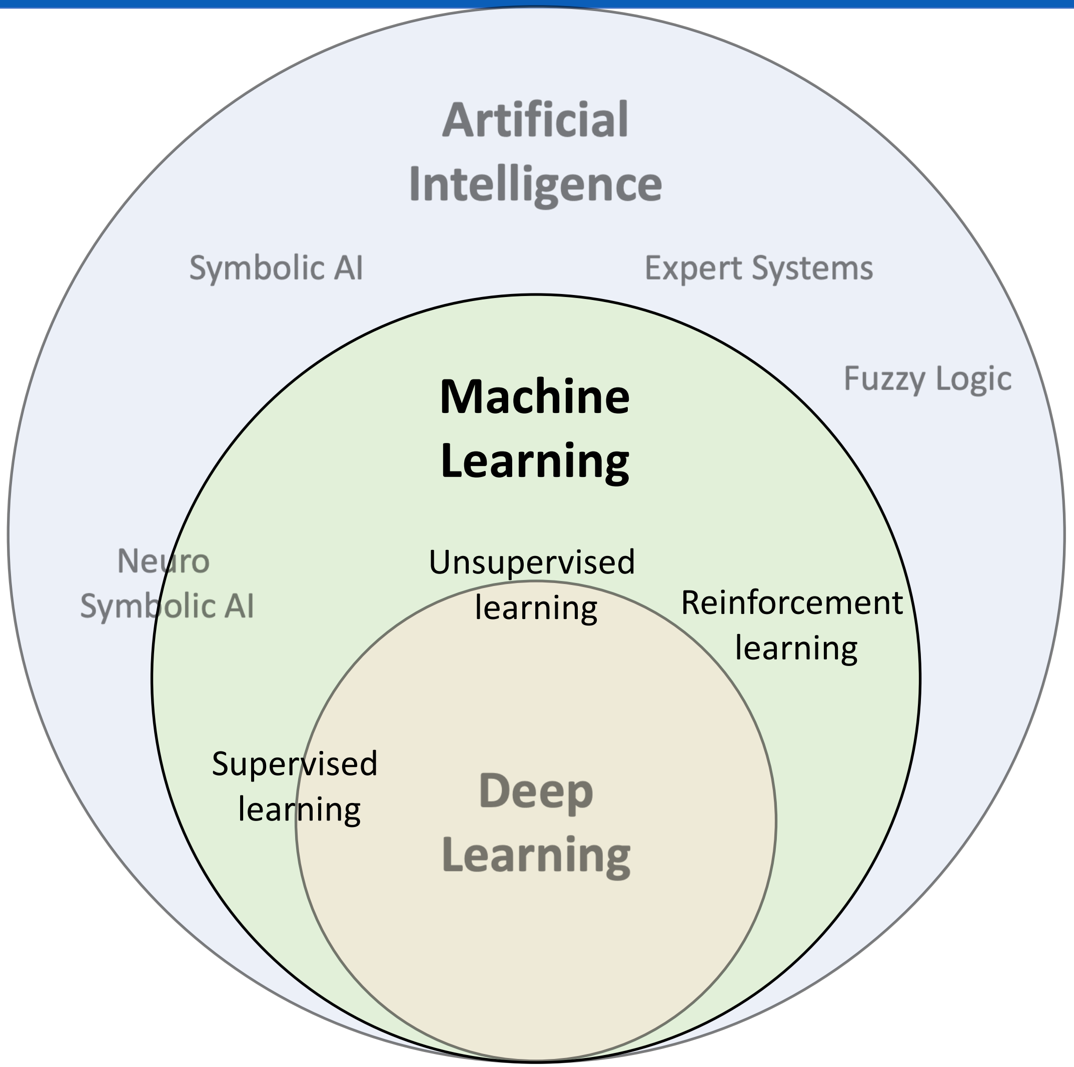
- We have a data set with only inputs.

- The algorithm finds patterns in the data and reacts accordingly.

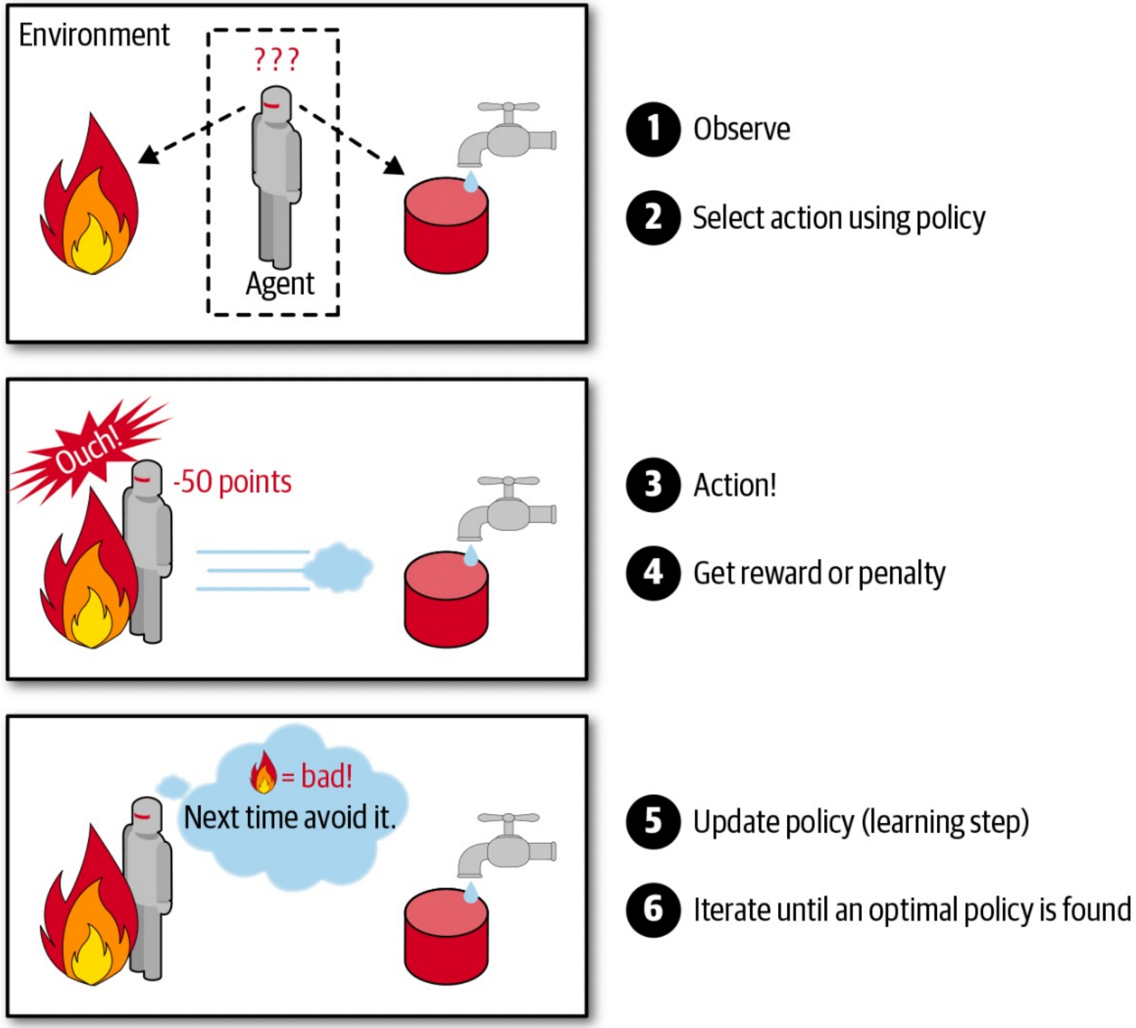


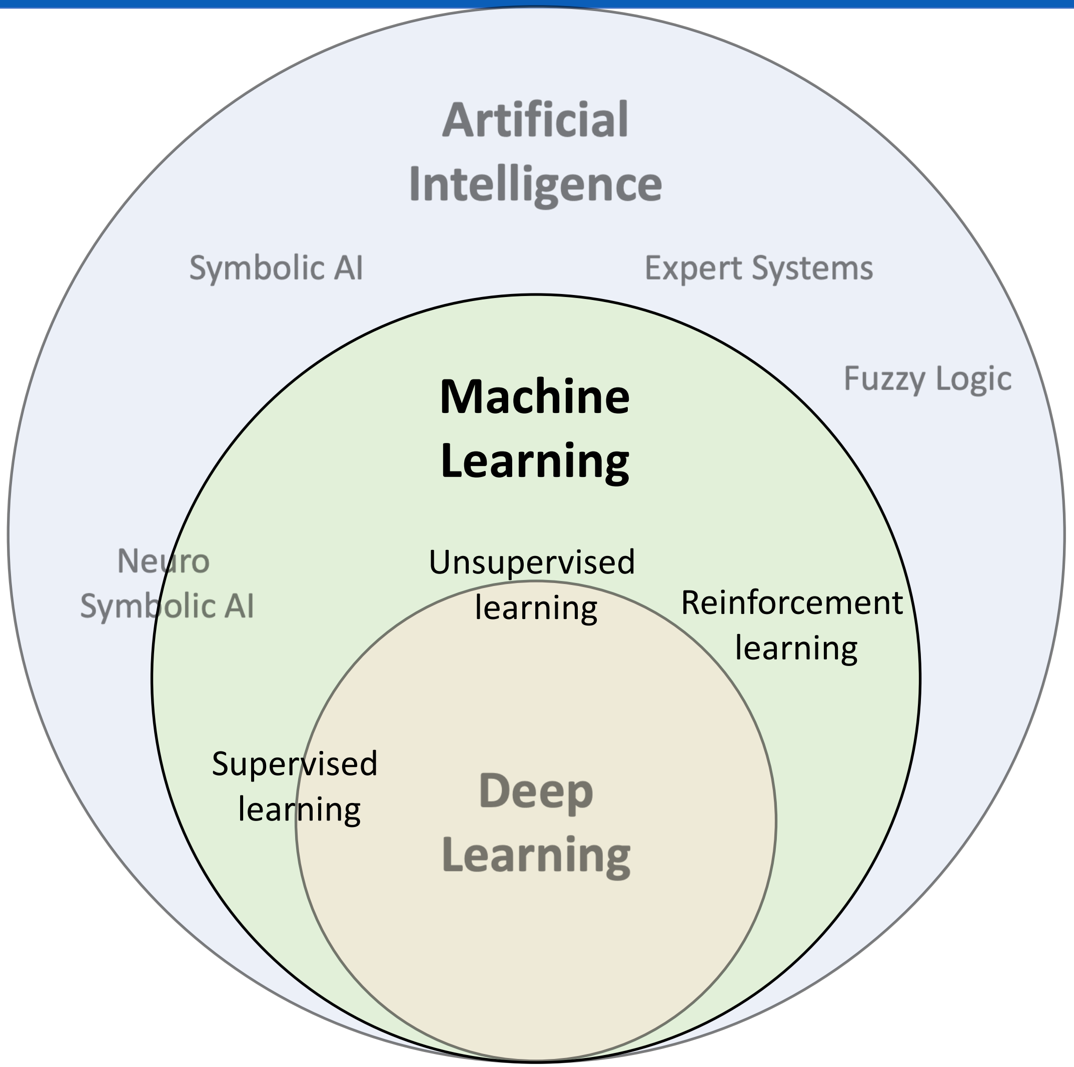
- Examples:
  - Clustering analysis
  - Principal component analysis

Unsupervised Learning



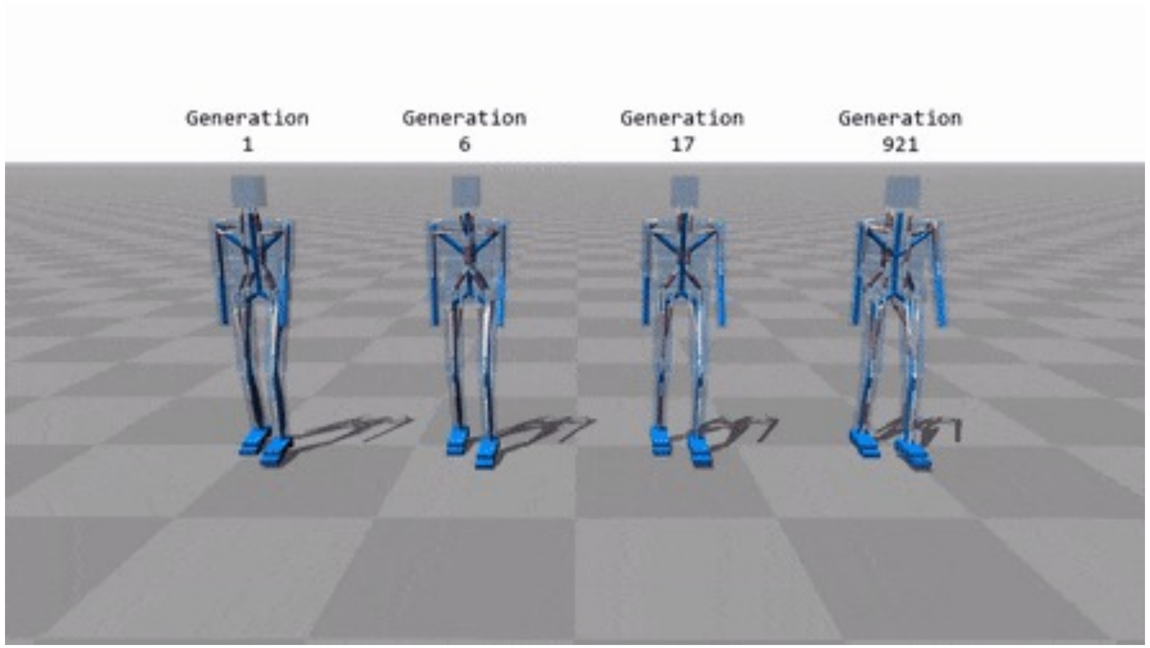
Reinforcement Learning

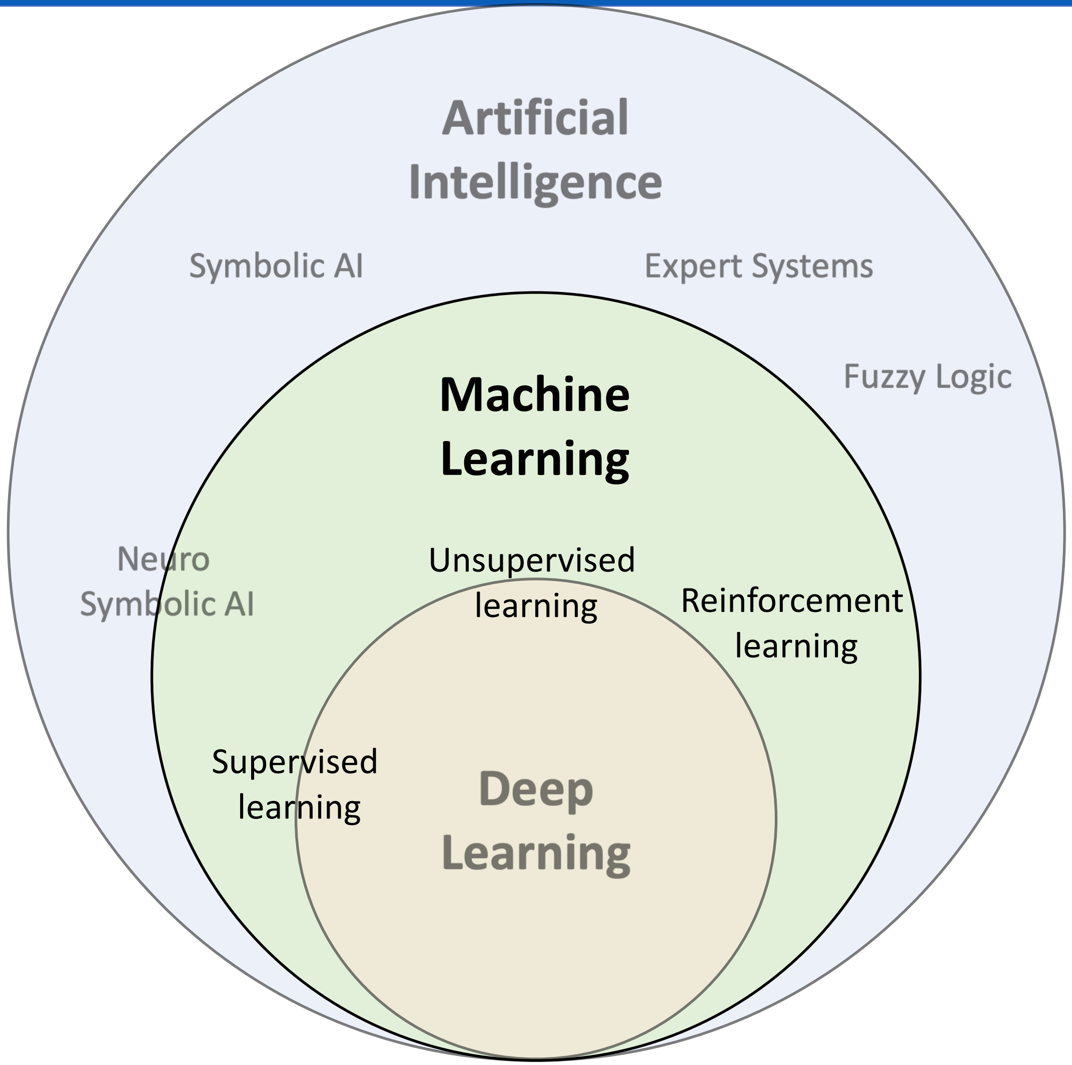




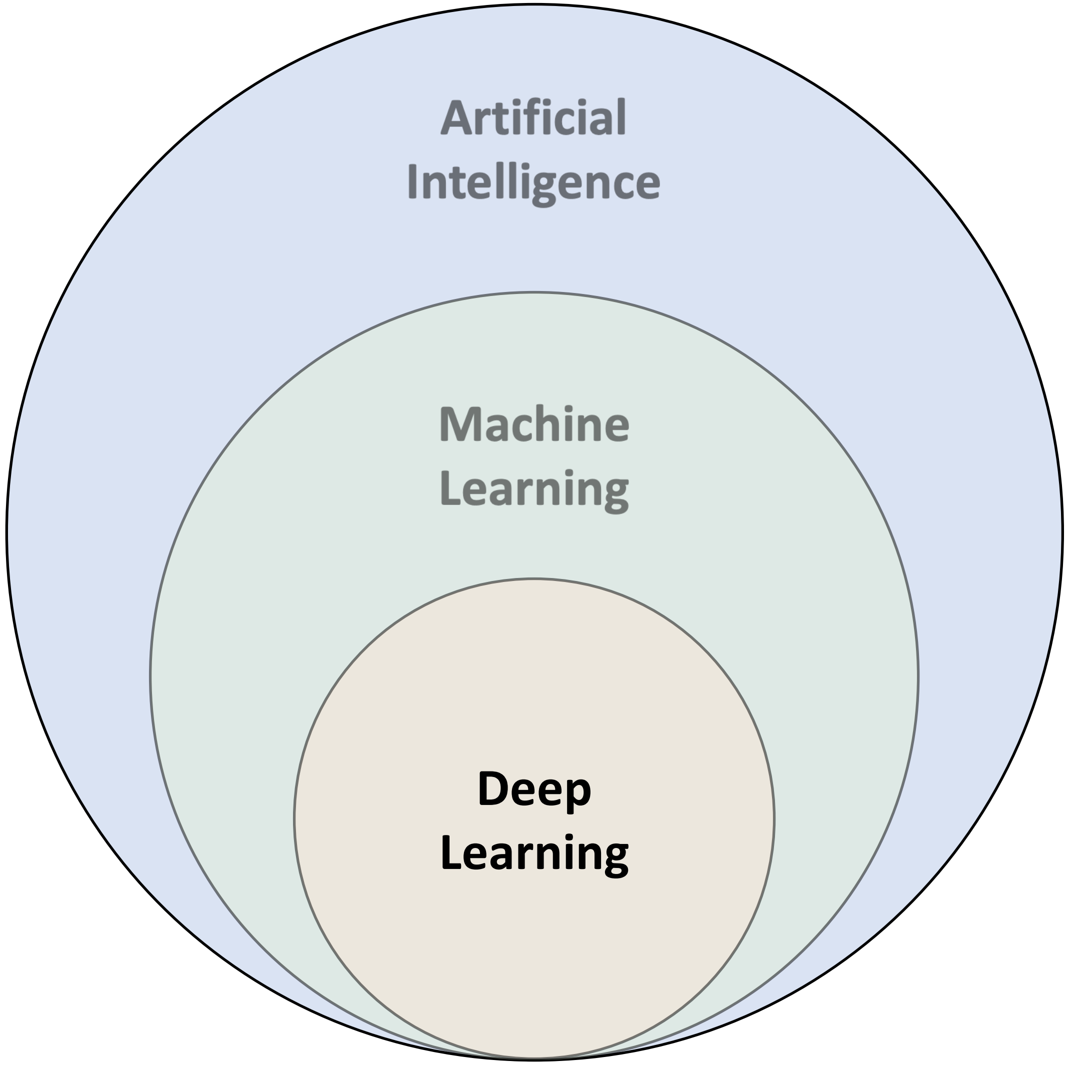
[1] Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 1150.

Unsupervised Learning





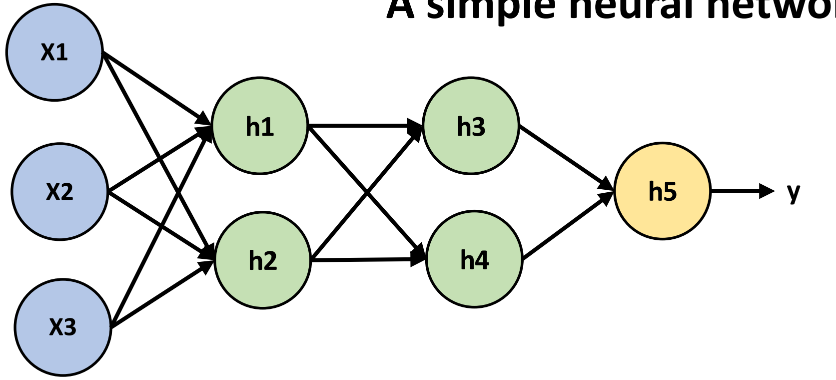
Deep Learning



Deep Learning

A subset of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher level of features from data (Oxford dictionary)

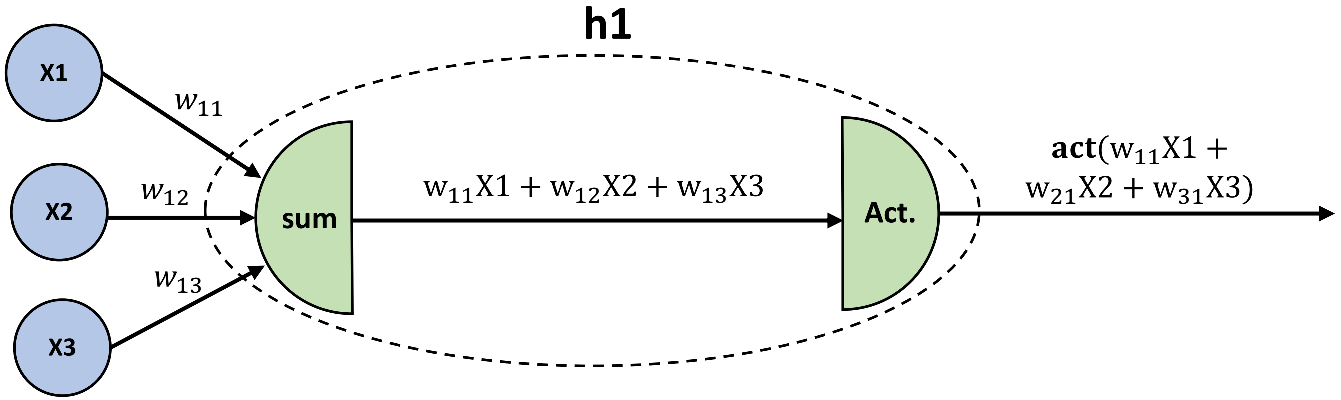
A simple neural network

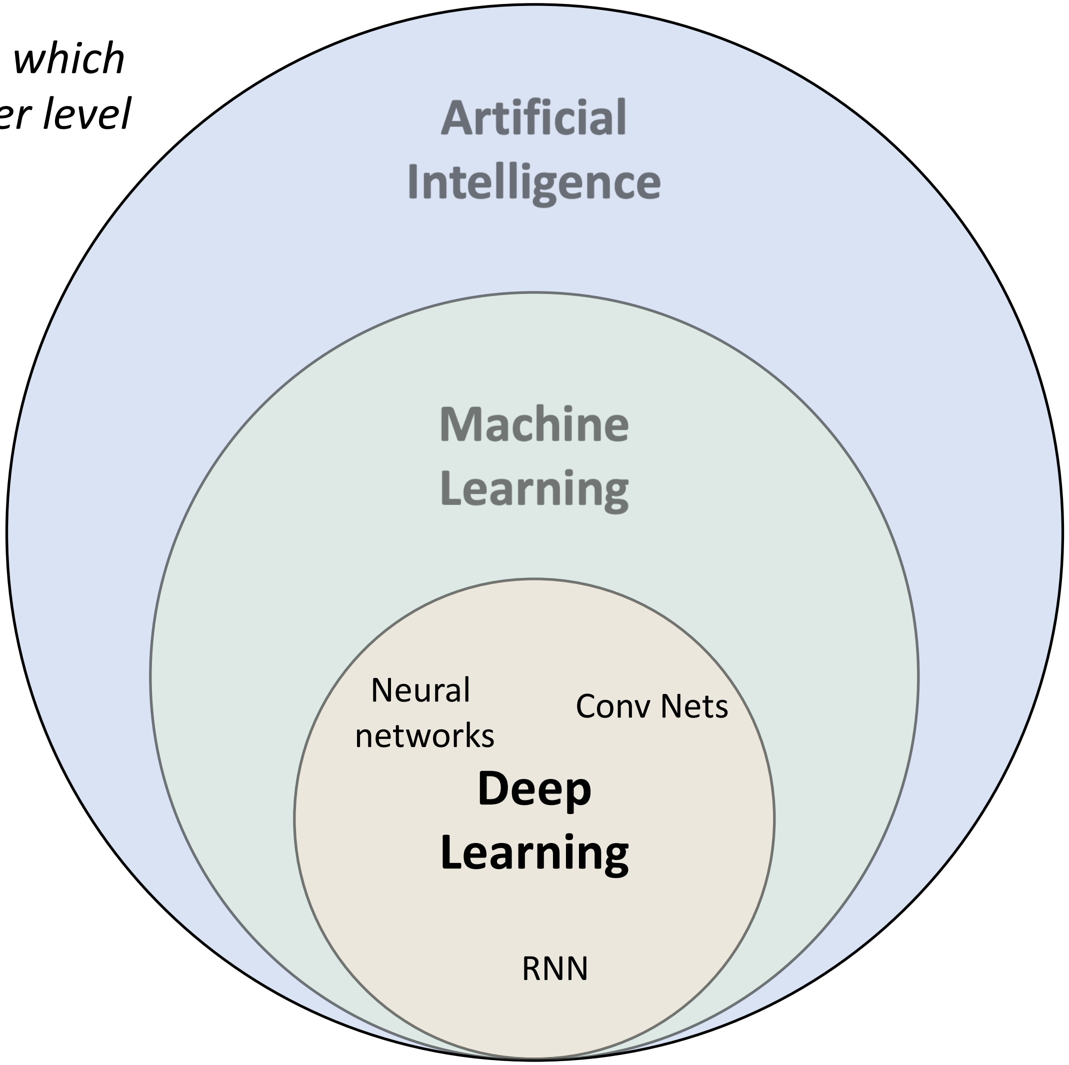


X1, X2, X3 are features

h1, h2, h3, h4 are called hidden nodes

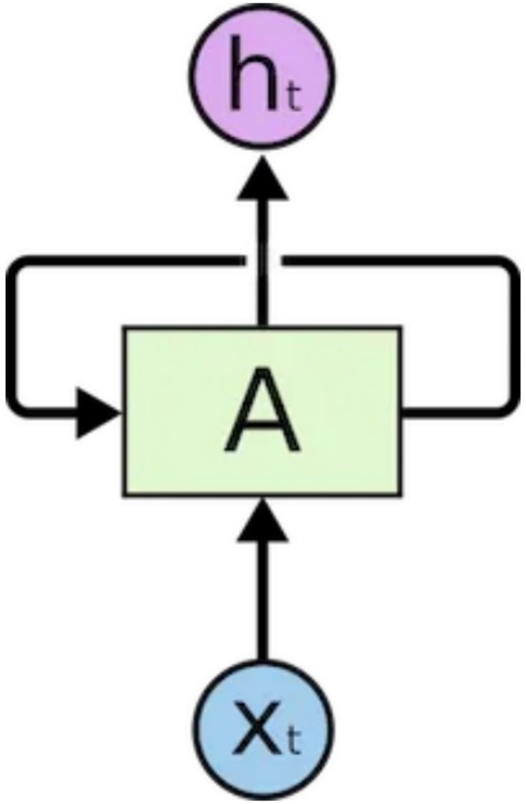
y is label



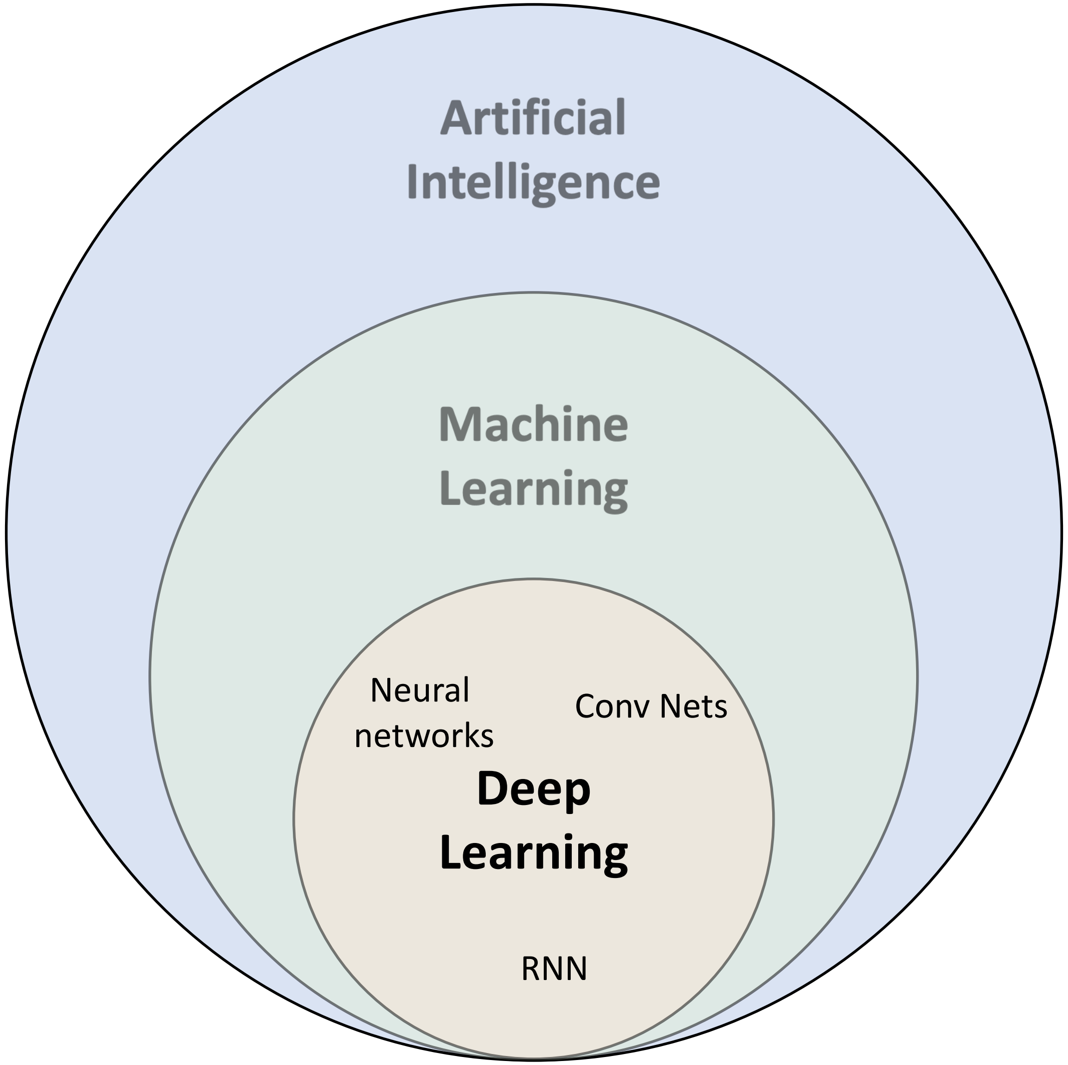


Training a neural network is simply learning the appropriate weights

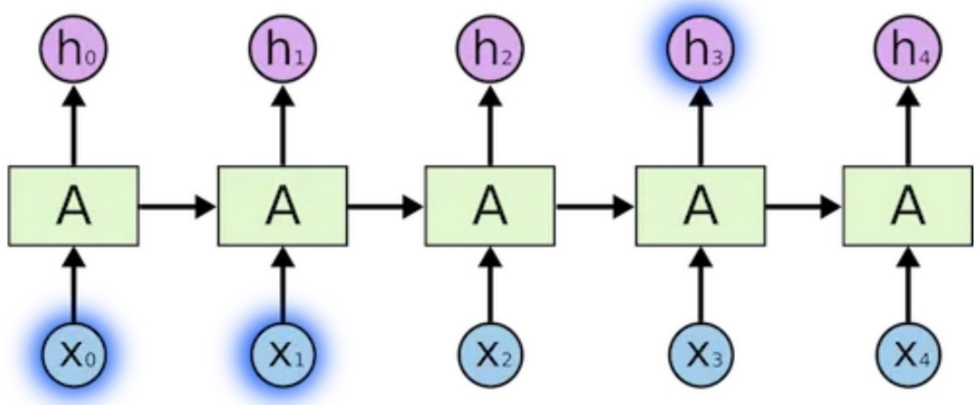
Recurrent Neural Net

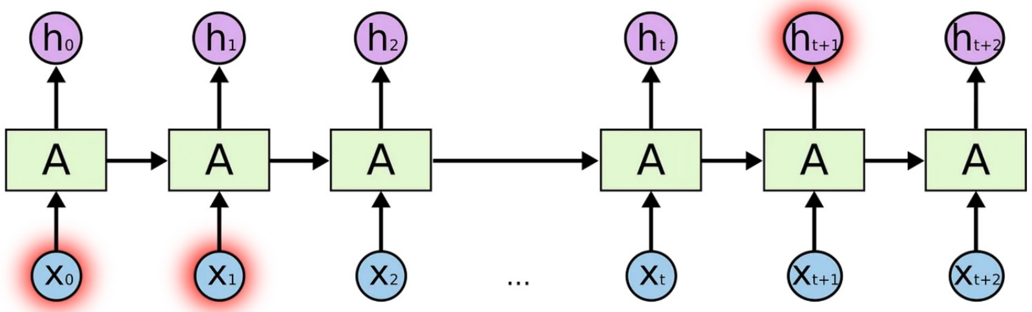


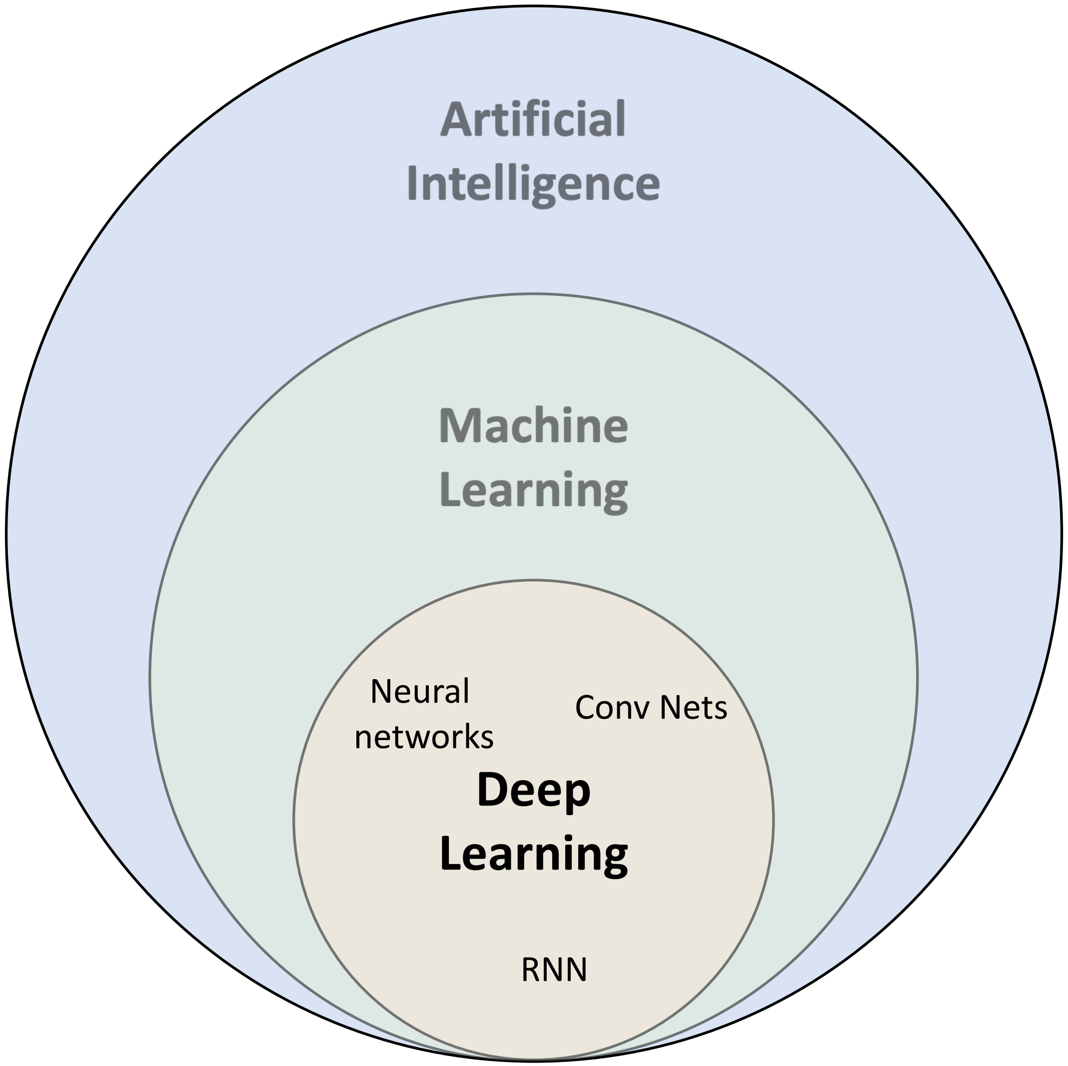
The input is represented as x_t



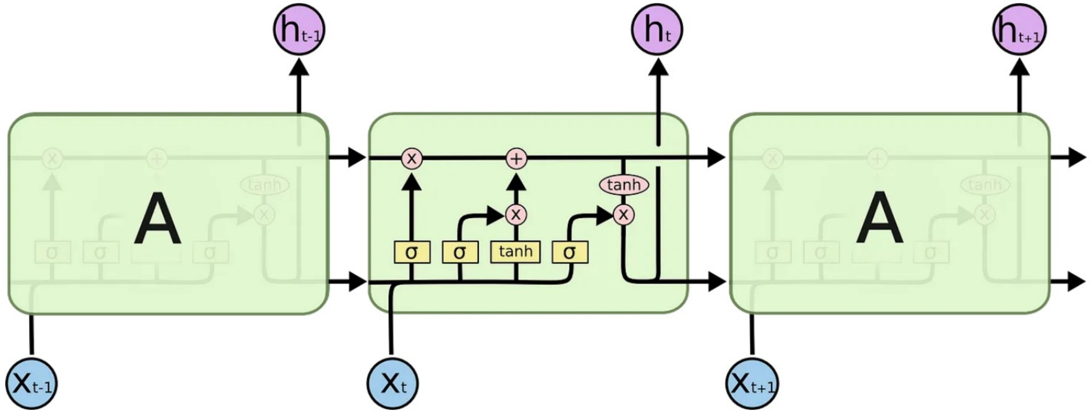
Recurrent Neural Nets … are Inefficient



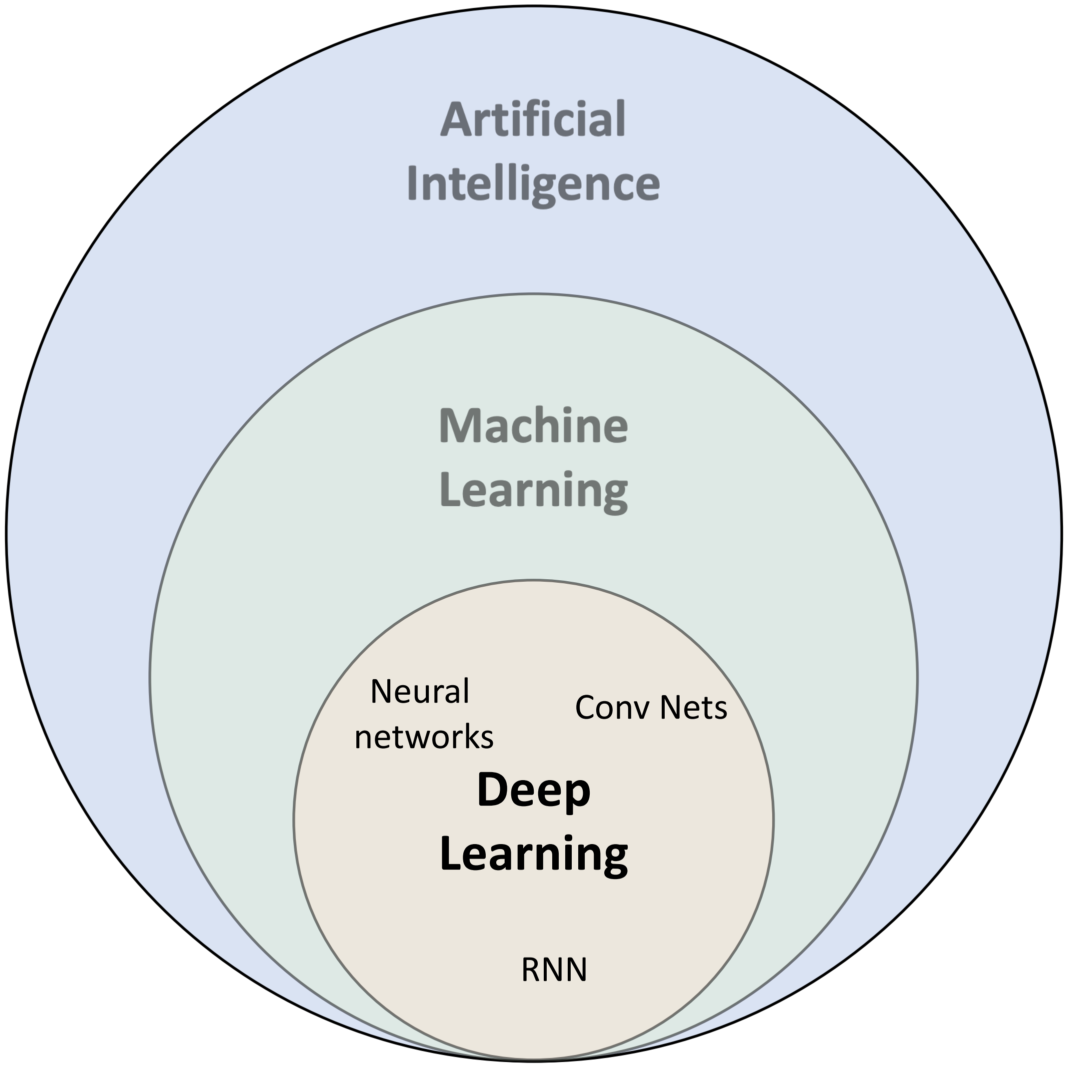




Long-Short Term Memory



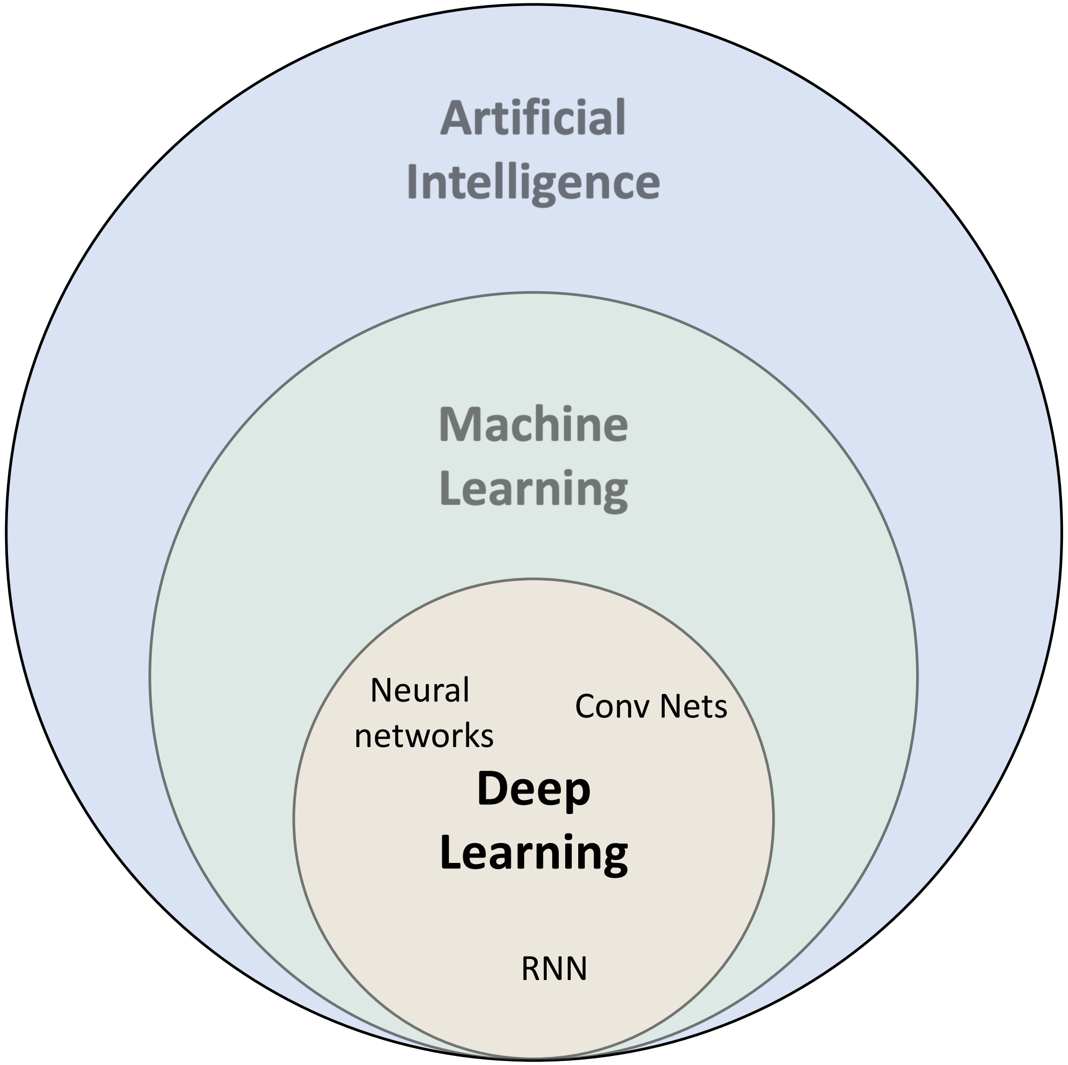
- Cannot parallelize
- No explicit modeling of long & short range dependencies
- Distance between positions is linear



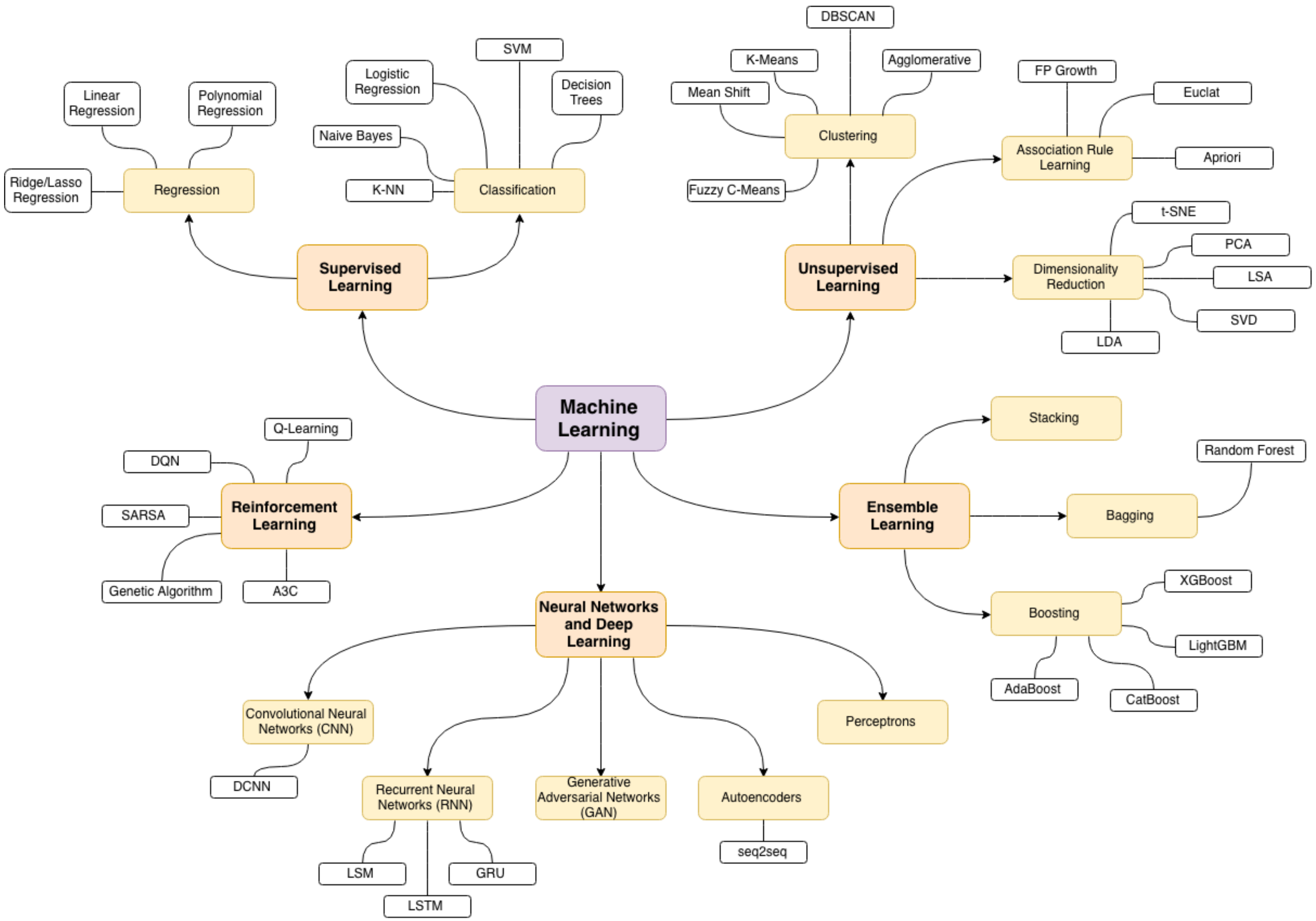
...Attention

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

https://www.youtube.com/watch?v=fjJOgb-E41w&ab_channel=GoogleCloudTech



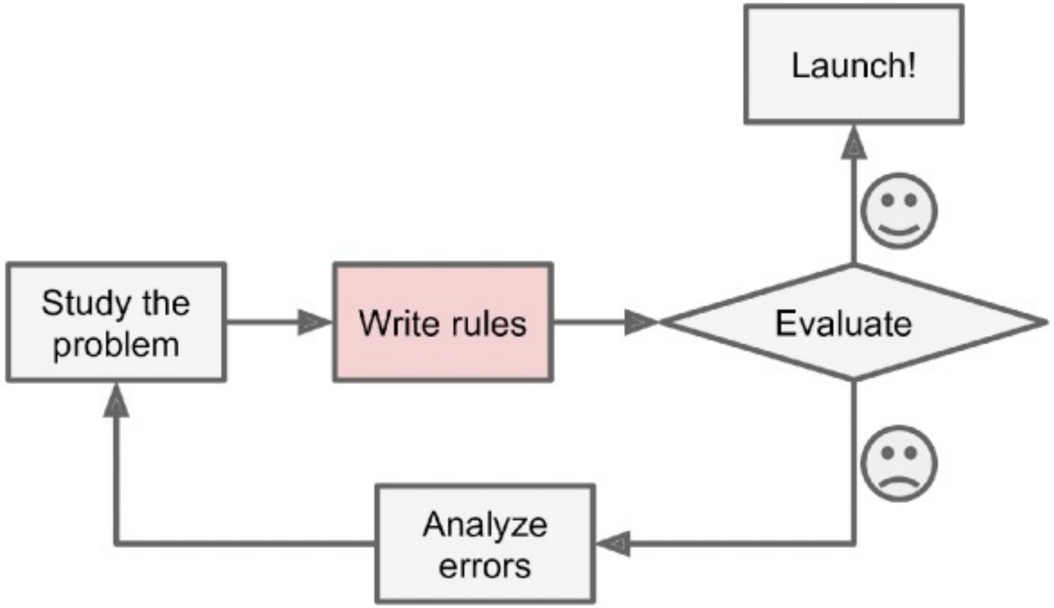
Machine Learning… in a Nut Shell



https://github.com/trekhleb/homemade-machine-learning

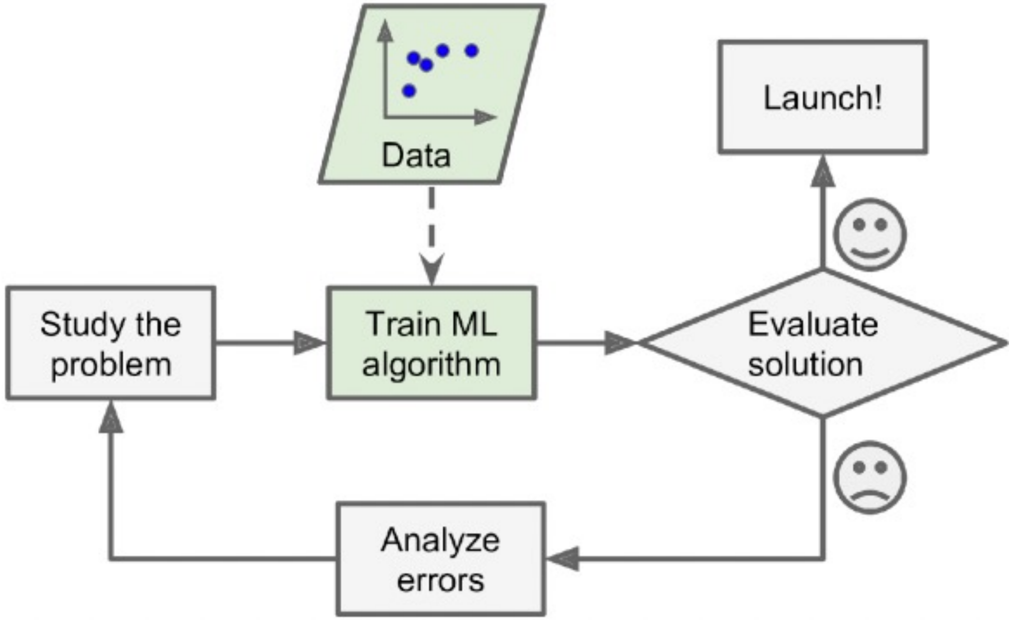
Why Machine Learning is useful ?

Simplified traditional program workflow



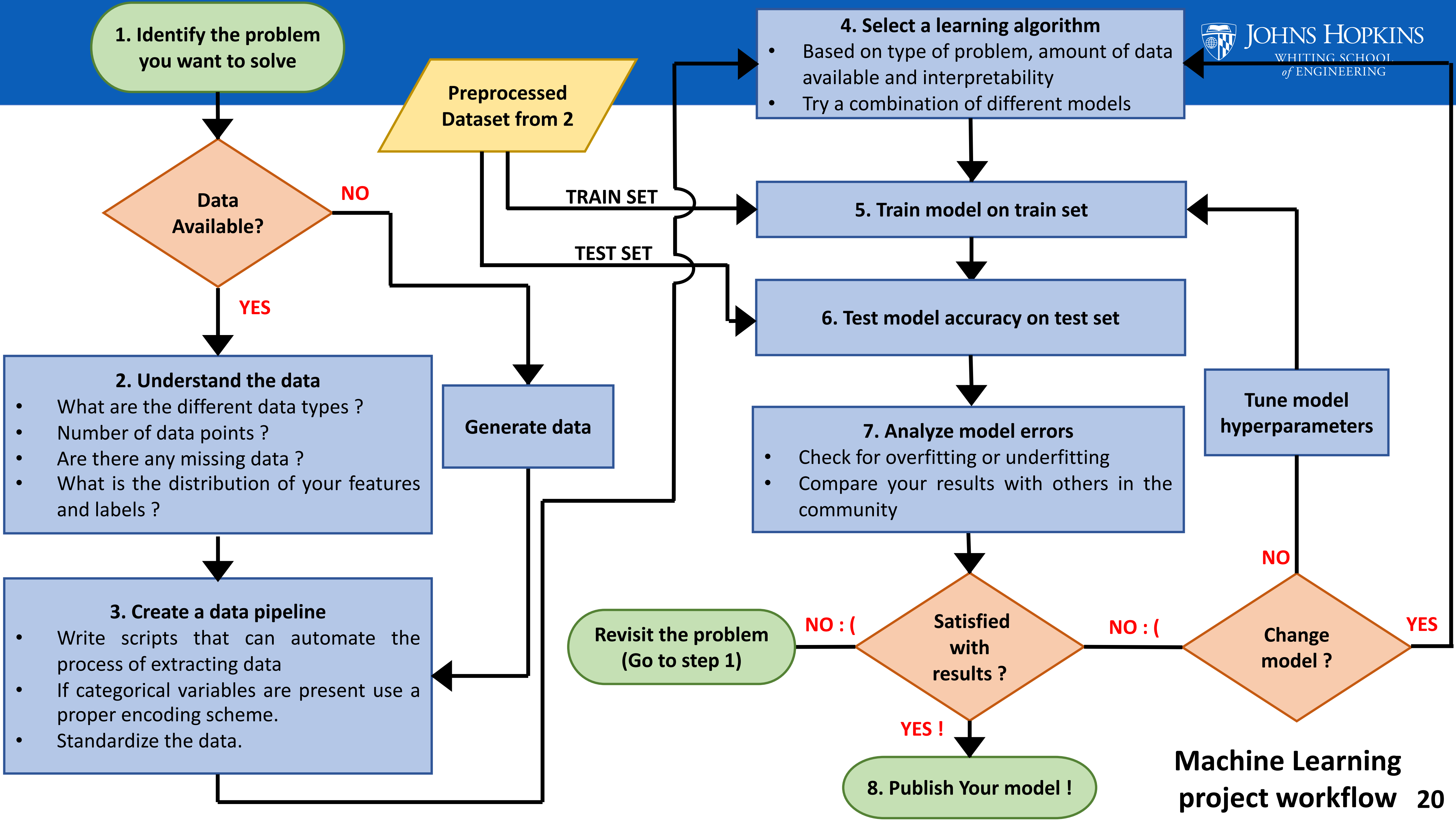
- Requires defining conditional statements
- Building model takes a significant amount of time.
- Must rewrite model code to incorporate new data.
- Cannot solve problems with no analytical or numerical solution.

Simplified machine learning program workflow



- No hard coded rules, model learns patterns from the data.
- Building model is fast due to well documented python libraries
- A well trained model is generalizable to perform a task on new data.
- Can perform complex tasks such as image recognition, language translation and can generate new insights from the data.

[1]. Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 1150.



Machine Learning project workflow

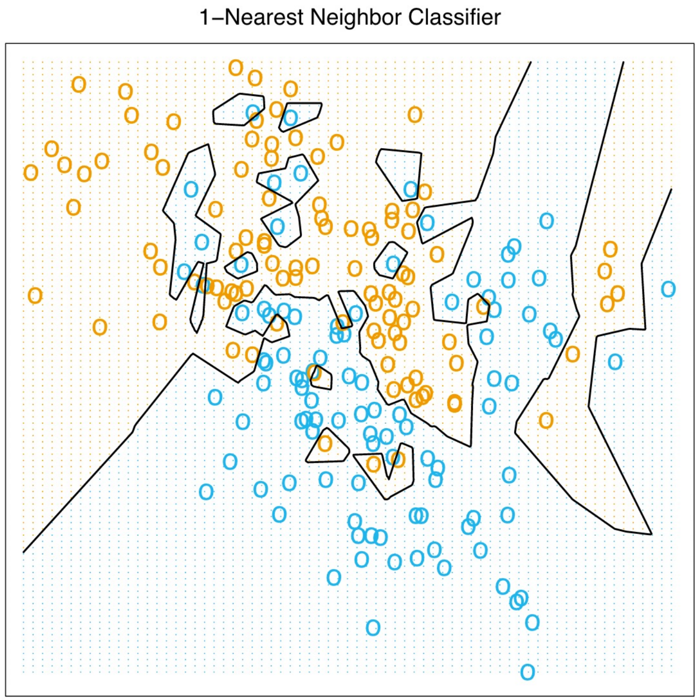
Challenges

Challenges related to data collection

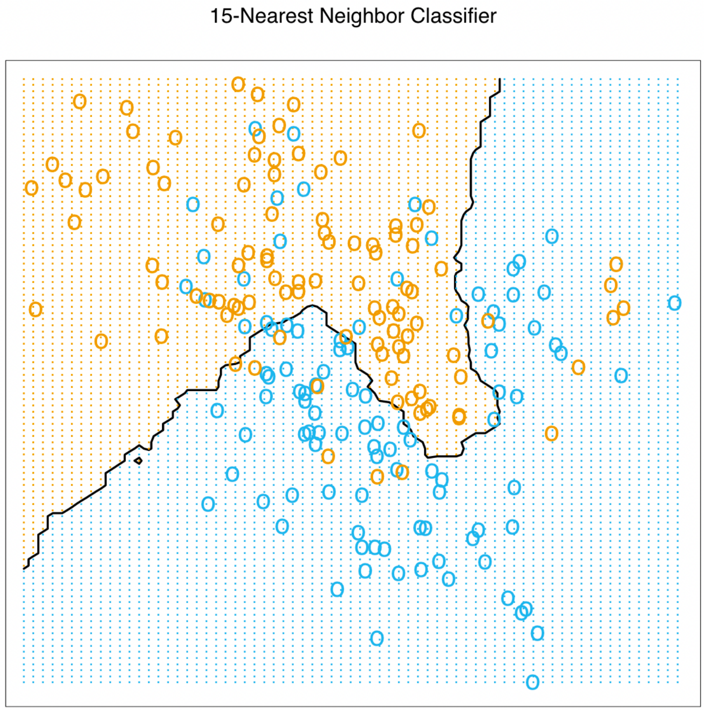
1. Insufficient quantity
2. Non representative
3. Poor quality data
4. Irrelevant features (GIGO)

Challenges related to training machine learning model

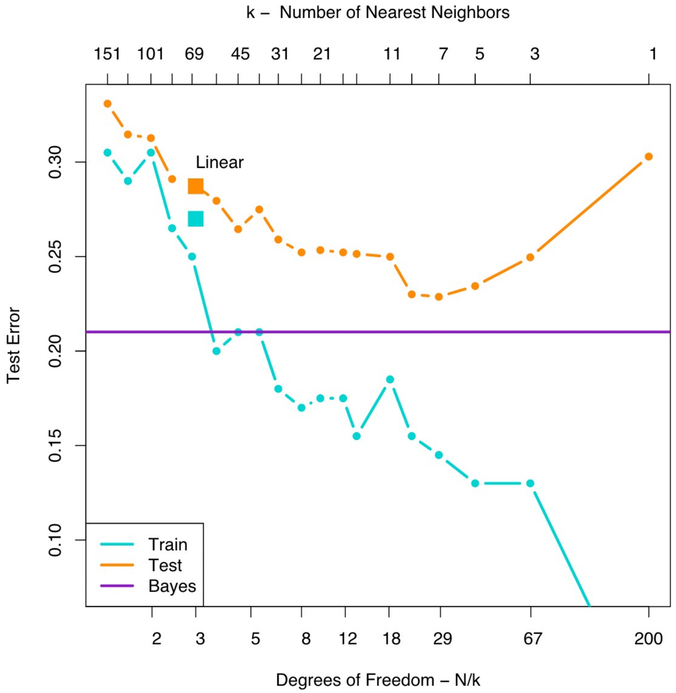
Overfitting on training data



Underfitting on training data



Bias Variance Tradeoff



1. Tibshirani, S. & Friedman, H. Valerie and Patrick Hastie. 764.

Common Machine Learning Modules/Toolboxes

Johns Hopkins
Whiting School of Engineering

- Keras
  - Neural network library.
  - High-level abstractions, easy to use.
- scikit-learn
  - Designed to work with NumPy and SciPy.
  - Has basic regression, classification, and clustering algorithms.
- TensorFlow
  - Google's internal software library for machine learning applications.
  - Similar to Keras, but a bit harder to use.
- Theano
  - Efficient optimization of mathematical expressions using multi-dimensional arrays.
- PyTorch
  - Based on Torch library (developed by Facebook).
  - Used for computer vision and natural language processing.
- Shogun
  - Focus on kernel methods.
  - Originally developed for bioinformatics purposes.

Common Algorithms

Johns Hopkins Whiting School of Engineering

- Linear Regression
- Logistic Regression
- Support Vector Machines
- Decision Trees
- Random Forests
- Naïve Bayes Classification
- Boosting
- Clustering
- Principal Component Analysis
- Singular Value Decomposition
- Independent Component Analysis
- Neural Networks
- Generative Adversarial Networks
- Bayesian Optimization
- Gaussian Process Regression

Resources to start on your ML journey !

Textbooks

1. **Machine Learning : A Probabilistic Approach** by **Kevin P. Murphy**
   - Excellent book that dives right into the math of some common machine learning algorithms
2. **Elements of Statistical Learning** by **Trevor Hastie and Robert Tibshirani and Jerome Friedman**
   - A more advanced level introduction to machine learning. Skips a lot of the derivation parts. (Not for faint hearted !)
3. **Hands on Machine Learning with Scikit-Learn, Keras and Tensorflow** by **Aurelien Geron**
   - Great book to start with ! Guided tutorials and explanations to building and training machine learning models without diving too much into the math.
4. **Deep Learning with Python** by **Francois Chollet**
5. https://www.deeplearningbook.org by **Ian goodfellow and Yoshua Bengio and Aaron Courville**
   - Excellent book to follow for building and training neural networks !

Neural net visualization

Tensorflow documentation

PyTorch documentation

Youtube video links by 3Blue1Brown on Deep Learning

1. But what is a neural network ? | Chapter 1, Deep Learning
2. Gradient Descent, how neural networks learn | Chapter 2, Deep Learning
3. What is backpropagation really doing ? | Chapter 3, Deep Learning
4. Backpropagation calculus | Chapter 4, Deep Learning

Courses at Johns Hopkins for Upper Level Undergraduates

- EN.601.475 (01) Machine Learning -> CSE
- EN.540.405 (01) Machine Learrning -> ChemBE
- EN.601.475 (01) Machine Learning for Medical Applns -> ECE
- EN.540.405 (01) Machine Learning : Deep Learning -> CSE