Brief Overview of Lecture 2

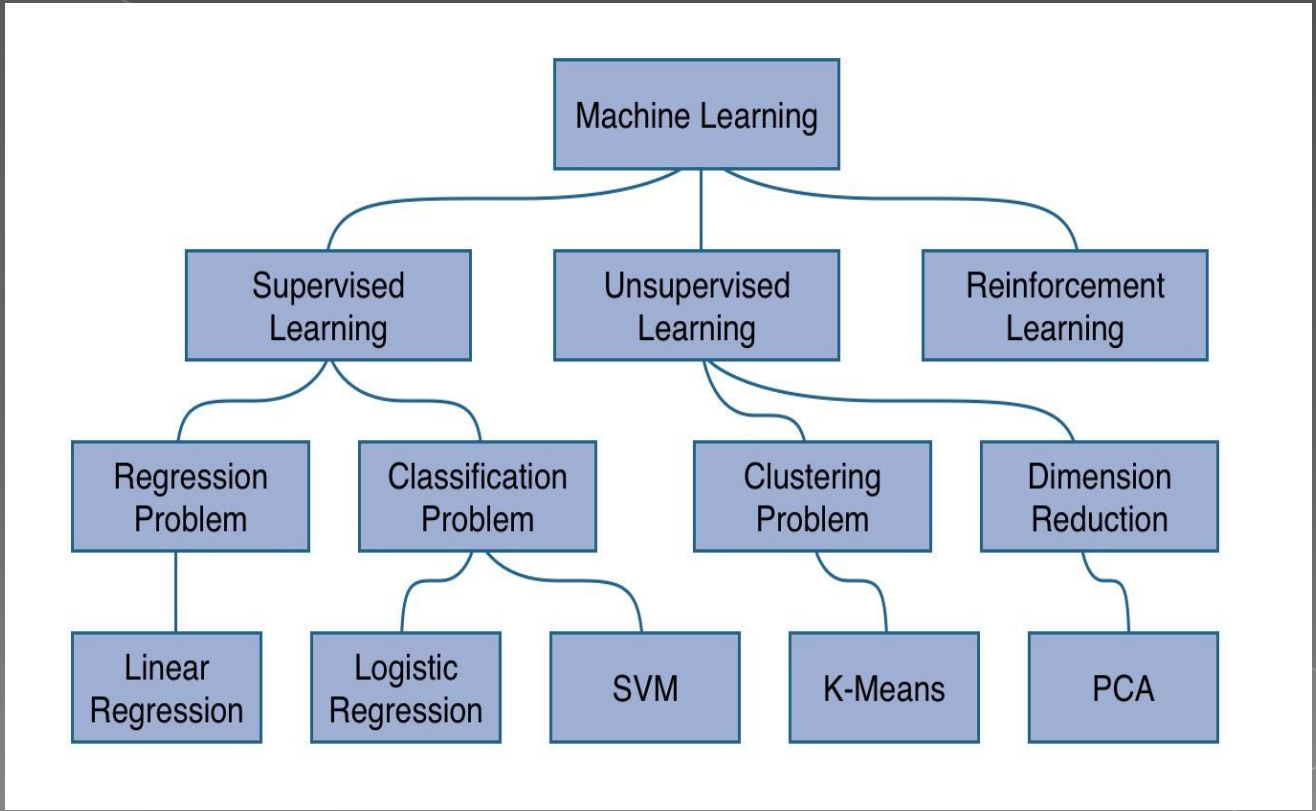
- Basic data mining tasks
- Data mining development
- Data mining issues

https://www.menti.com/5o5g9ow31r

Outline of Today's lecture

- Why Machine Learning?
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Machine Learning



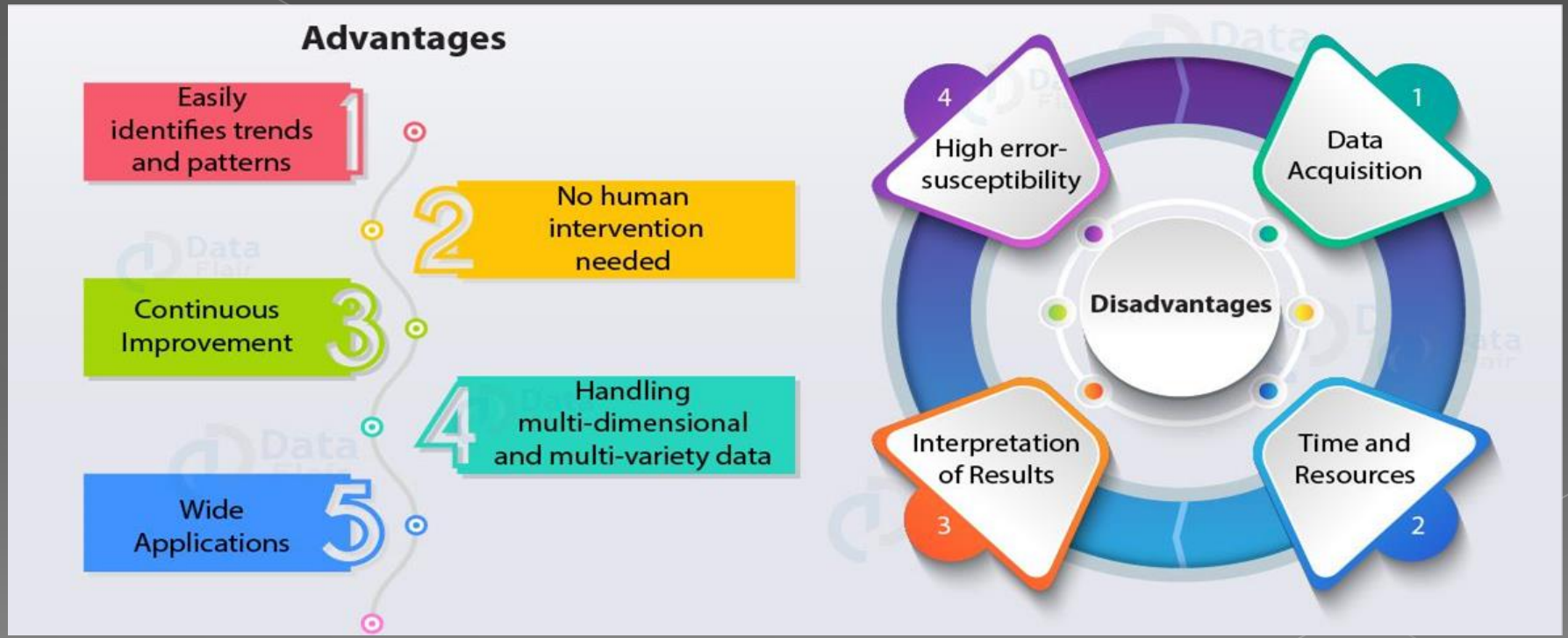


Image Courtesy: https://data-flair.training/blogs/advantages-and-disadvantages-of-machine-learning

Steps to Learn Machine Learning

Image Courtesy: https://data-flair.training/blogs/best-way-to-learn-machine-learning/

Steps to Learn Machine Learning

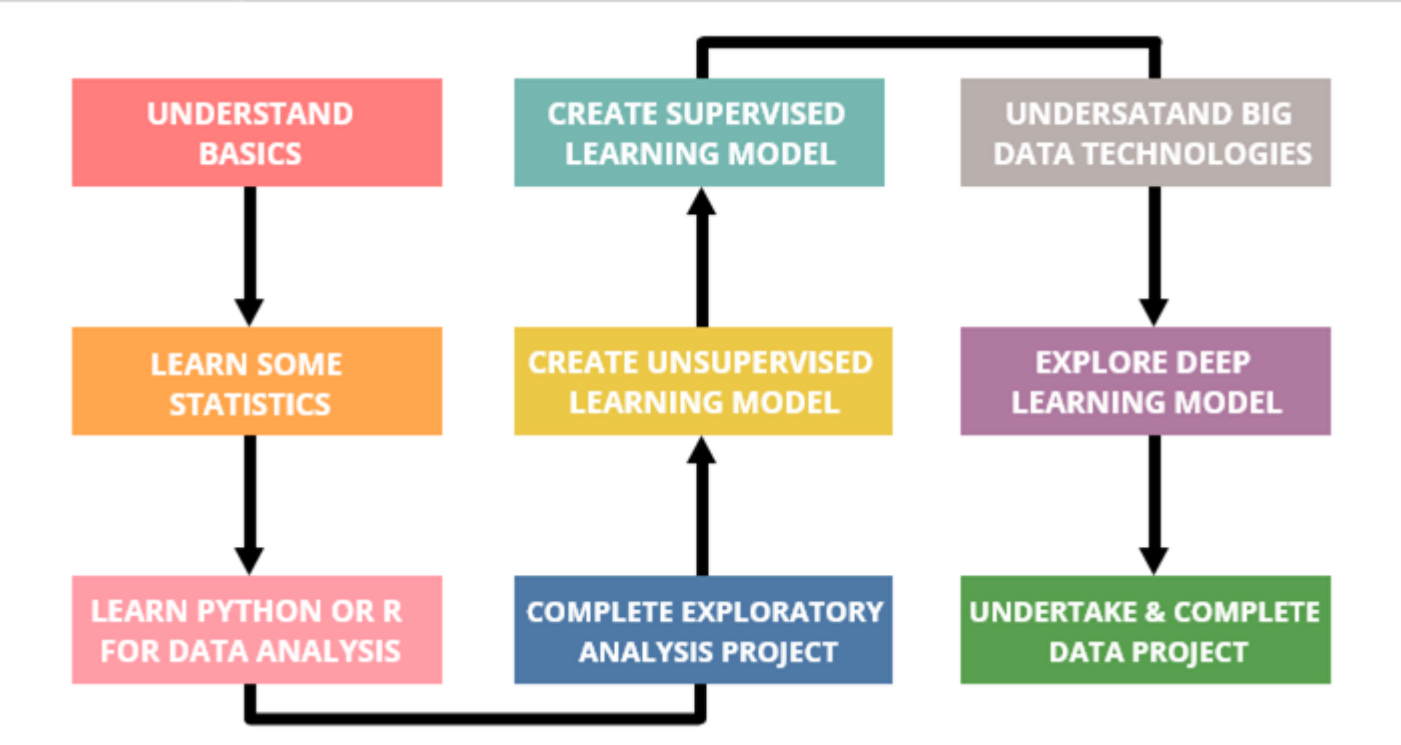


Image Courtesy of https://skillsmatter.com/blog/machine-learning-expert/

End to End Machine Learning Workflow

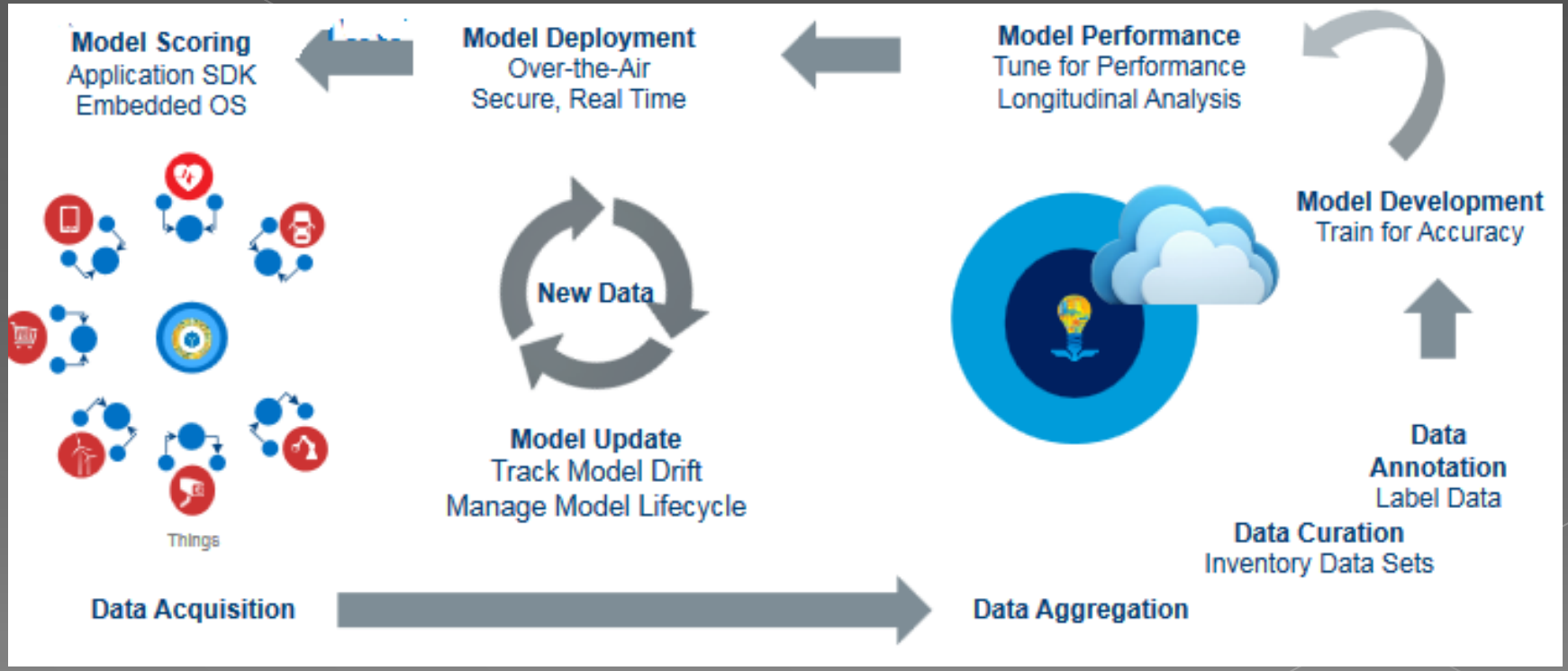


Image Courtesy: : Intel's Machine Learning Strategy slides by Gary Paek, HPC Marketing Manager, Intel Americas HPC User Forum, Tucson, AZ April 12, 2016

Why Machine Learning now?

Image Courtesy : Intel's Machine Learning Strategy slides by Gary Paek, HPC Marketing Manager, Intel Americas HPC User Forum, Tucson, AZ April 12, 2016

Why Machine Learning now?

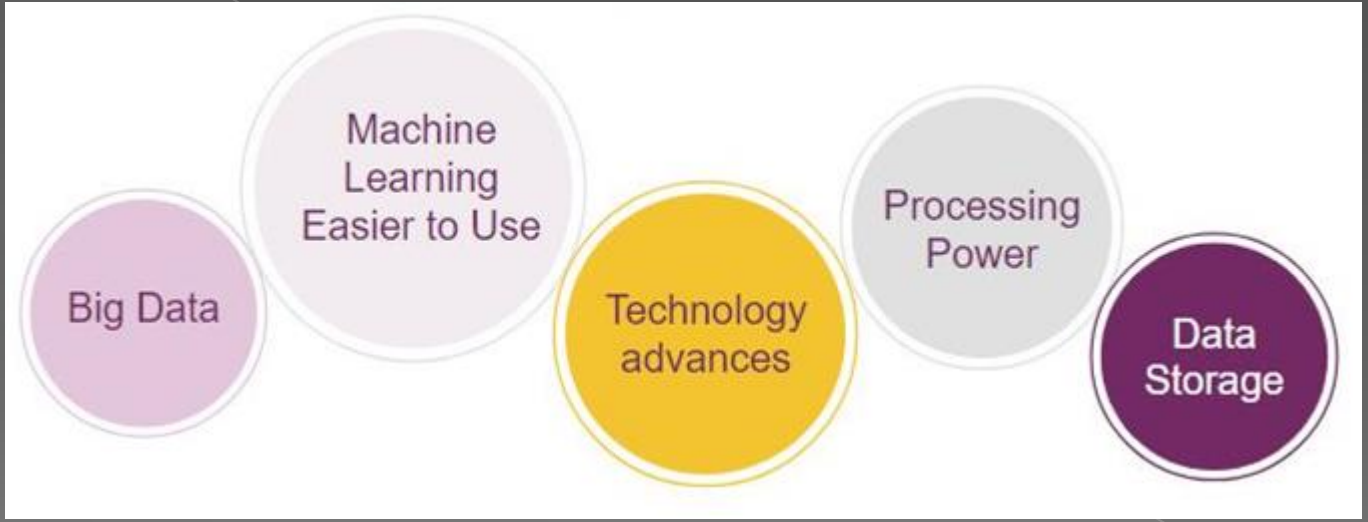


Image Courtesy: :https://www.learnovatecentre.org/machine-learning-in-edtech/

Taxonomic Foundation

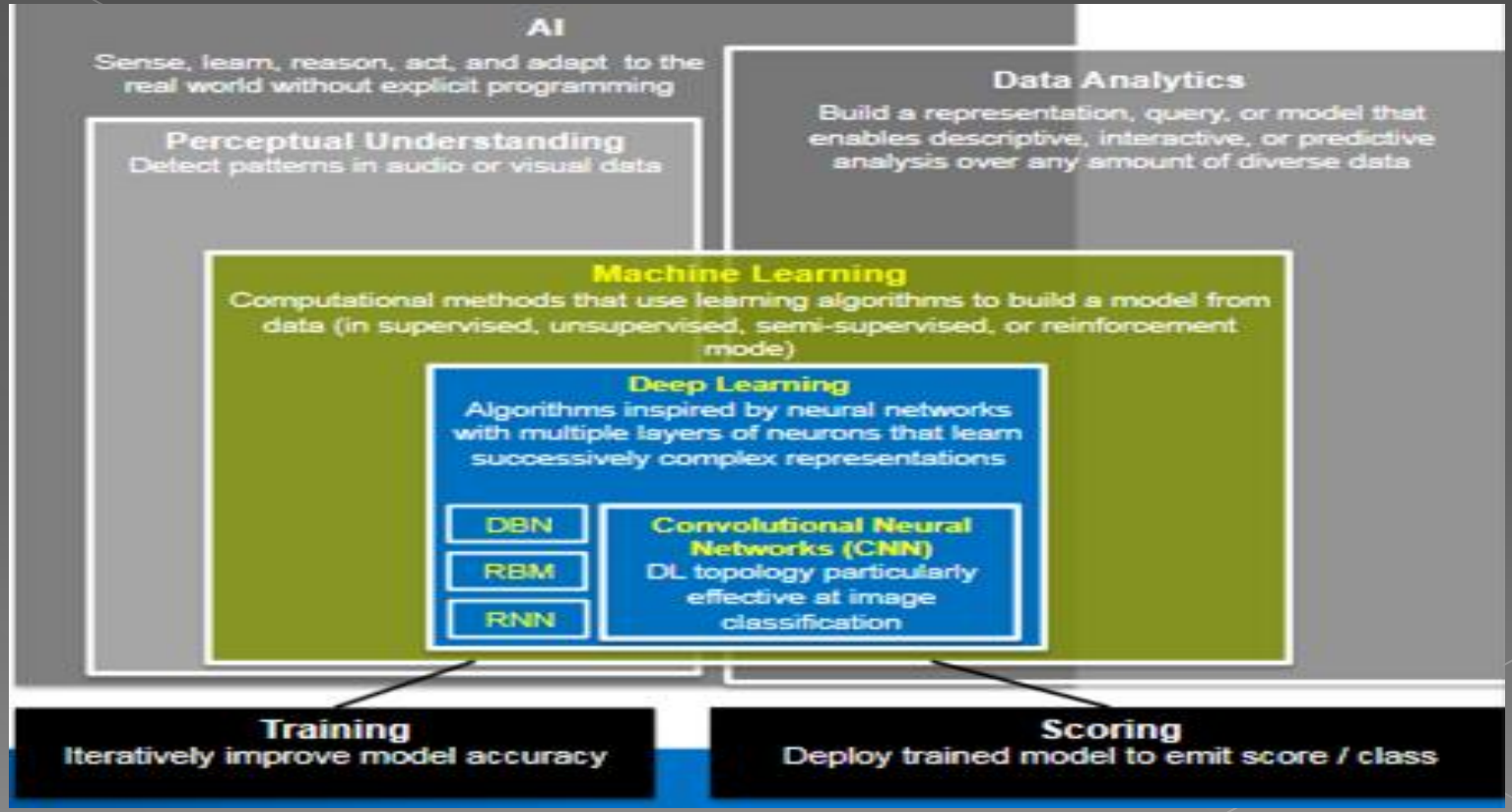


Image Courtesy: Intel's Machine Learning Strategy slides by Gary Paek, HPC Marketing Manager, Intel Americas HPC User Forum, Tucson, AZ April 12, 2016

Machine Learning Applicability

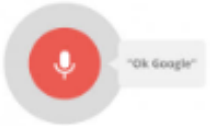
	Application	Model Type
IMAGENET	Object Localization and Image Classification	Convolutional Neural Networks (CNN), Support Vector Machines
NETFLIX RECOMMENDED TV	Collaborative Filtering, Recommendation Engines, Inputting Missing Interactions	Restricted Boltzmann Machines (RBM), ALS
	Anomaly Detection	Clustering, Decision Trees
	Forecasting or prediction of time-series and sequences like speech and video	Recurrent Neural Networks (RNN), Long-short Term Memory (LSTM), Hidden Markov Models
	Click Through Rate (CTR) Prediction	Logistic Regression
	State-Action Learning, Decision Making	Deep Q Networks (Reinforcement Learning)

Image Courtesy: Intel's Machine Learning Strategy slides by Gary Paek, HPC Marketing Manager, Intel Americas HPC User Forum, Tucson, AZ April 12, 2016

Types of Machine Learning

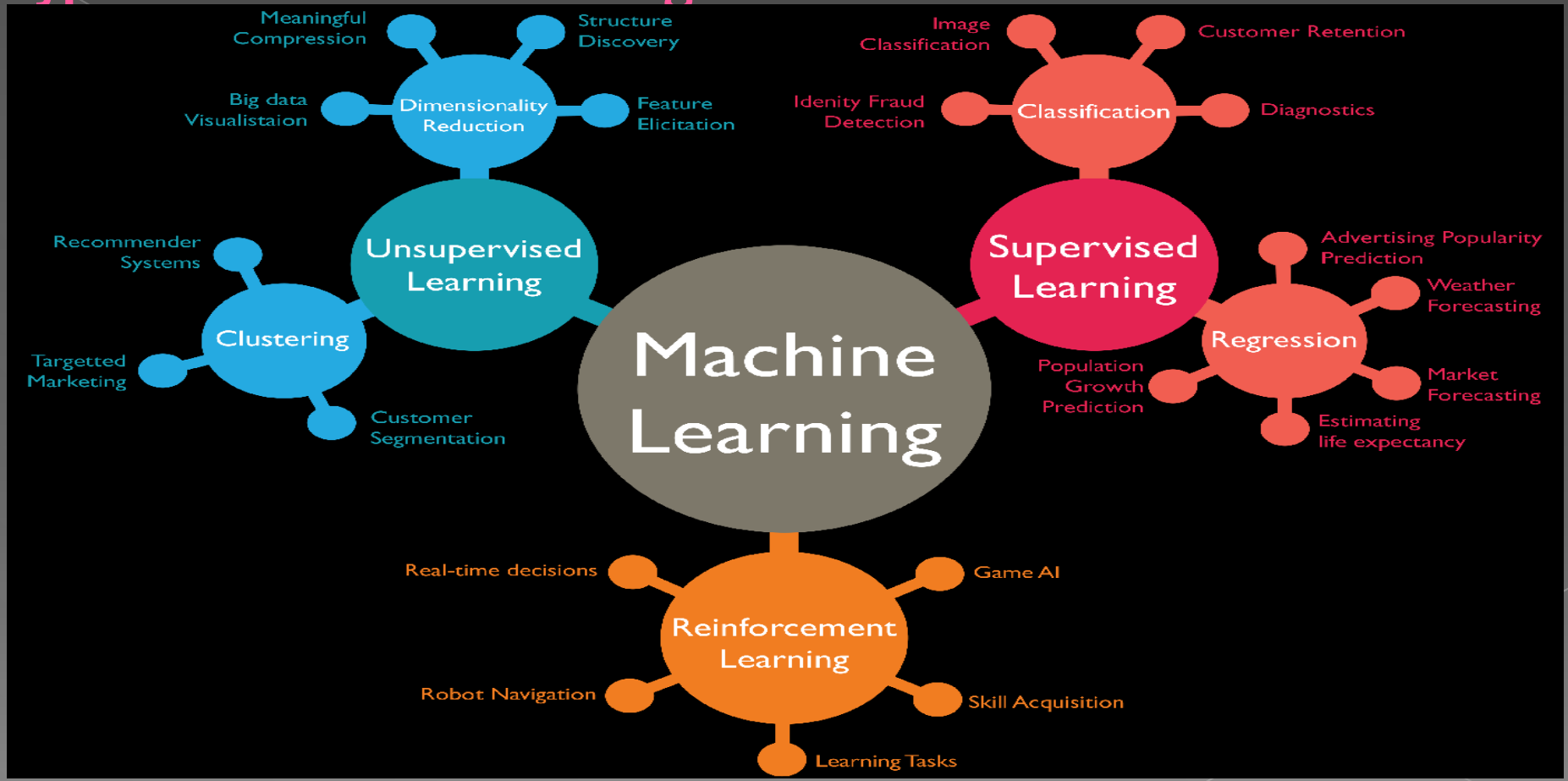


Image Courtesy: https://artificialintelligence.oodles.io/blogs/intro-to-machine-learning/

What is Machine learning?

Machine learning is a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!).

Machine learning Notations

- *D* **training set**,
- *N* is the number of training examples.
- $\mathbf{X}_i$ training input *is a D*-dimensional vector.
- These are called **features**, **attributes** or **covariates**.

Eg:

- X_i=Email message
  X_i[To ,From,Bcc,cc,content]
- X_i =Image
  X_i[color,size,shape]

$\mathbf{Y}_i$: *output variable or Response variable* . **categorical** or **nominal**

Supervised Learning

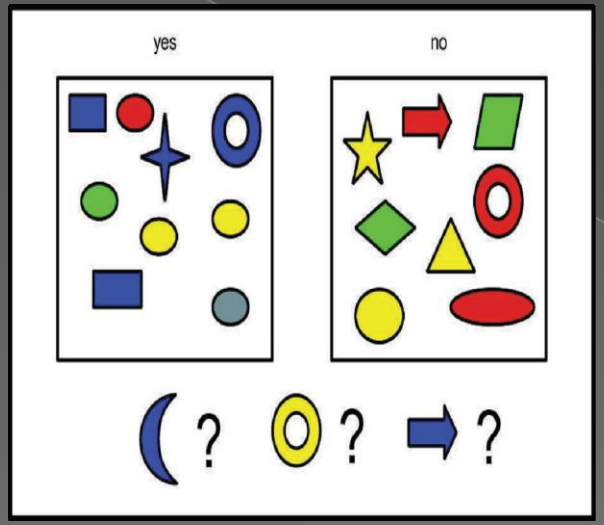
- The goal is to learn a mapping from inputs **x** to outputs *y*, given a labelled set of input-output pairs.

- This kind of learning is possible when inputs and the outputs are clearly identified, and algorithms are trained using labelled examples.

Supervised learning-Classification

- The goal is to learn a mapping from inputs **x** to outputs y, where $y \in \{1, \ldots, C\}$, with C being the number of classes.
- If $C = 2$, this is called **binary classification** (in which case we often assume $y \in \{0, 1\}$);
- if $C > 2$, this is called **multiclass classification**.
- If the **class labels are not mutually exclusive** (e.g., somebody may be classified as tall and strong), its **multi-label classification**

Example



D features (attributes)

N cases

Color	Shape	Size (cm)	Label
Blue	Square	10	1
Red	Ellipse	2.4	1
Red	Ellipse	20.7	0

Example

Binary Classification

- Spam
- Not spam

Multiclass Classification

- Dog
- Cat
- Horse
- Fish
- Bird
- ...

Multi-label Classification

- Dog
- Cat
- Horse
- Fish
- Bird
- ...

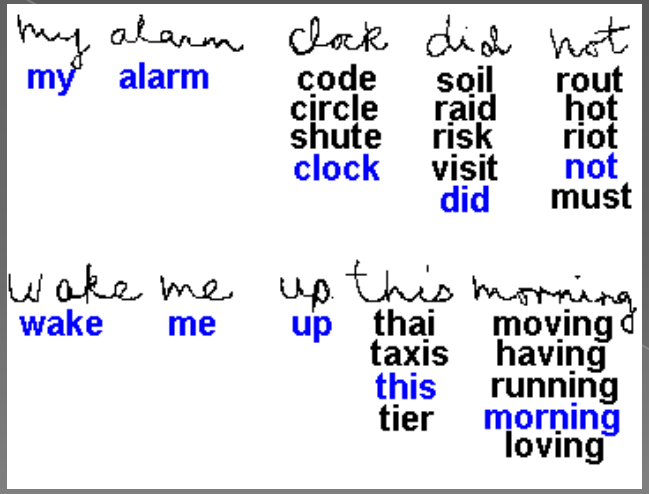
Real time applications-Supervised Learning
Document classification

Image Courtesy:https://www.pericent.com/products/docedge-dms/hot-features/classification/

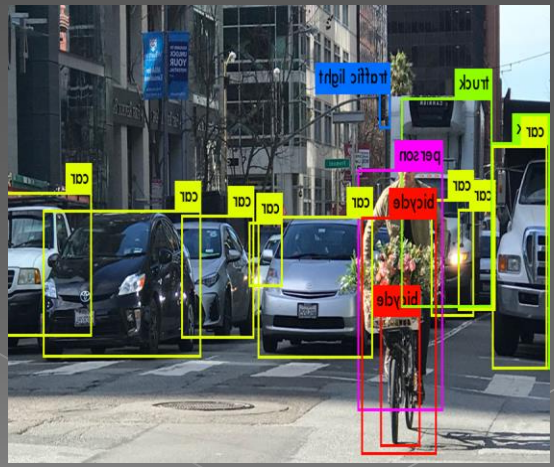
Real time applications-Supervised Learning

Email spam filtering

Handwriting recognition



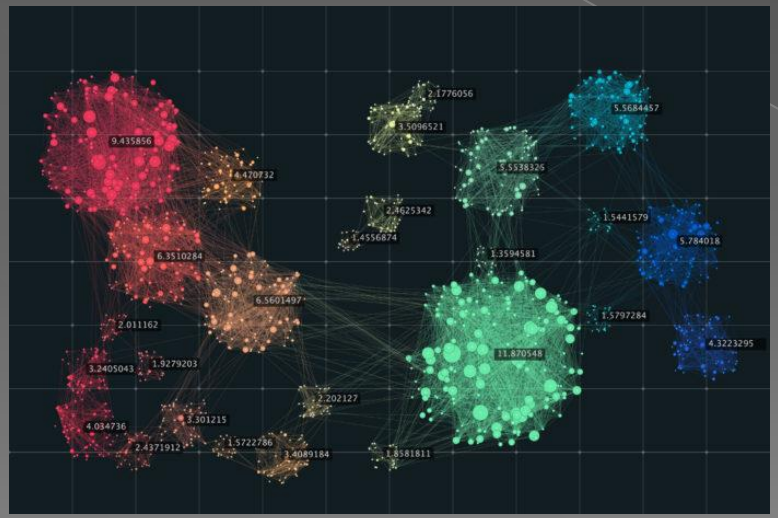
Object Detection



Unsupervised Learning

Descriptive or **unsupervised learning**

For the given inputs, and the goal is to find “interesting patterns” in the data. This is sometimes called **knowledge**



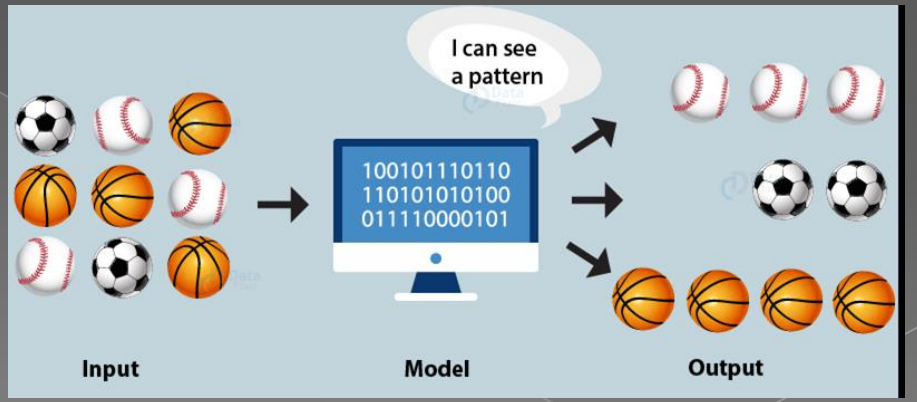


Image Courtesy:https://data-flair.training/blogs/clustering-in-machine-learning/

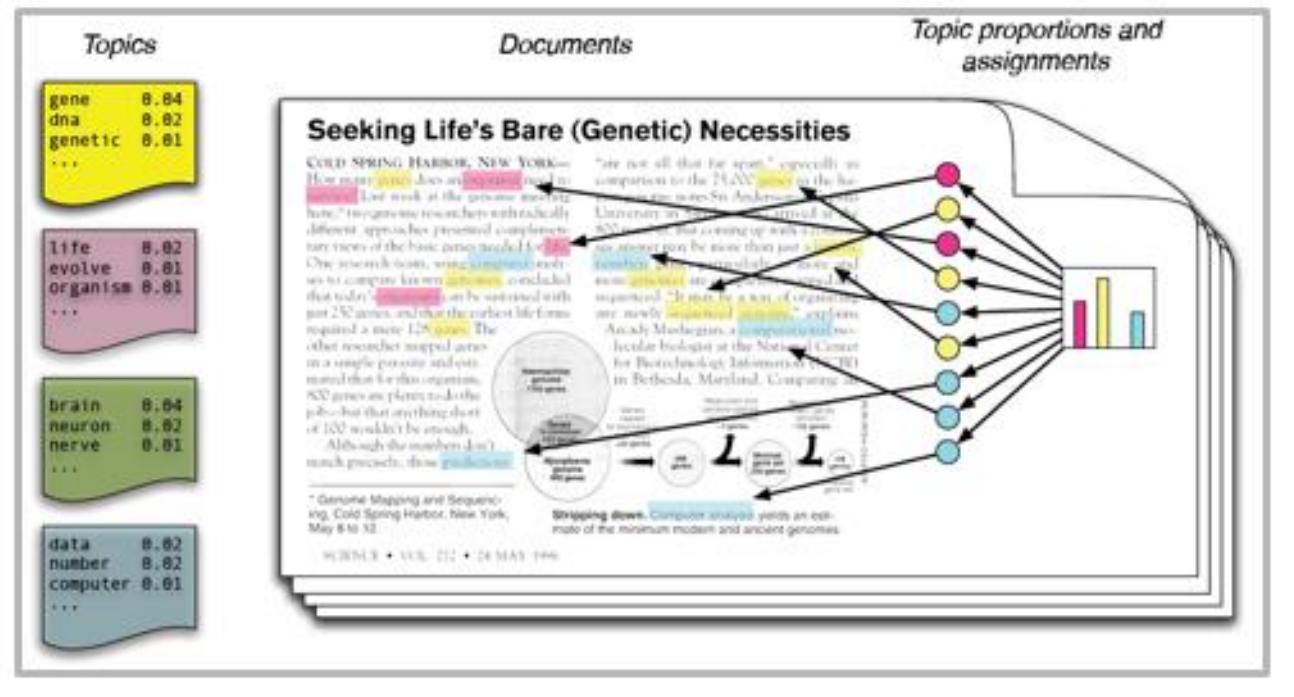
Unsupervised Learning

- Learn from the data - no labels
- Discover **"interesting structure"** in the data - **knowledge discovery**
- Does not require a human expert to label data

Applications

- Discovering **clusters**
- Discovering **latent** factors
  - Topic model / Theme/ Essence of the data
  - Dimensionality reduction
- Discovering graph structures
- Matrix completion - image imputation
- Association mining

Real time -applications-Unsupervised-Discovering latent factors



Real time -applications-Unsupervised Learning Discovering Graph Structure

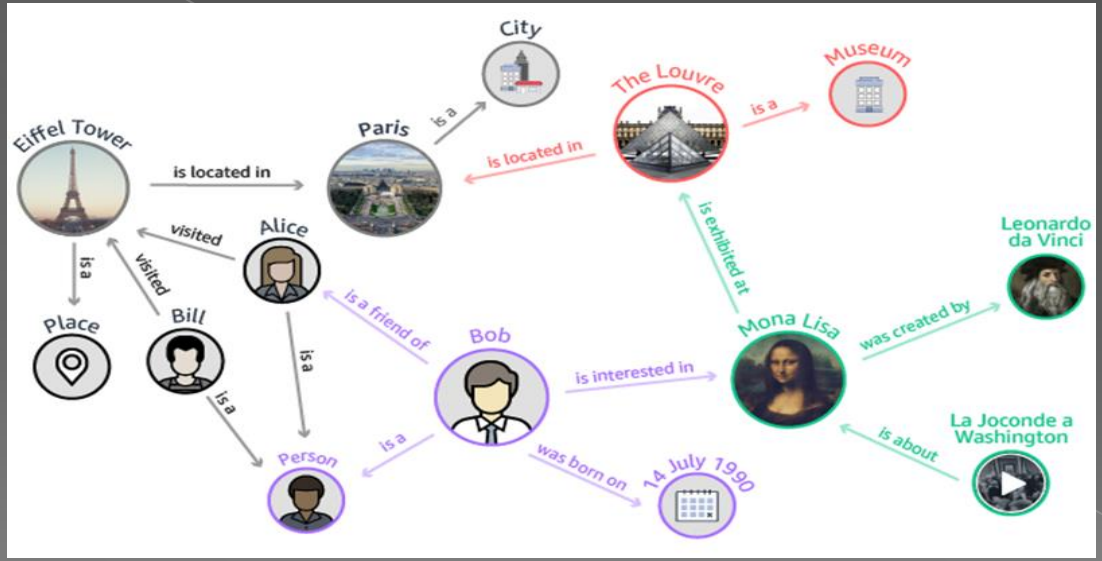
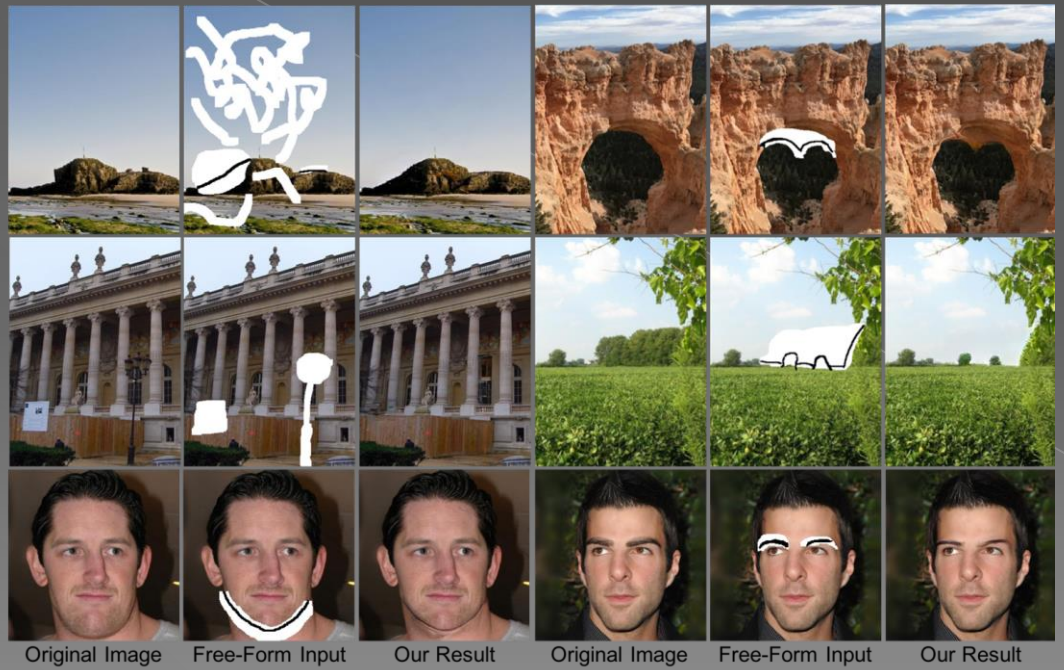


Image Courtesy:https://aws.amazon.com/blogs/apn/exploring-knowledge-graphs-on-amazon-neptune-using-metaphactory/

Real time -applications-Unsupervised Learning-Image Inpainting



Real time -applications-Unsupervised Learning-Collaborative Filtering

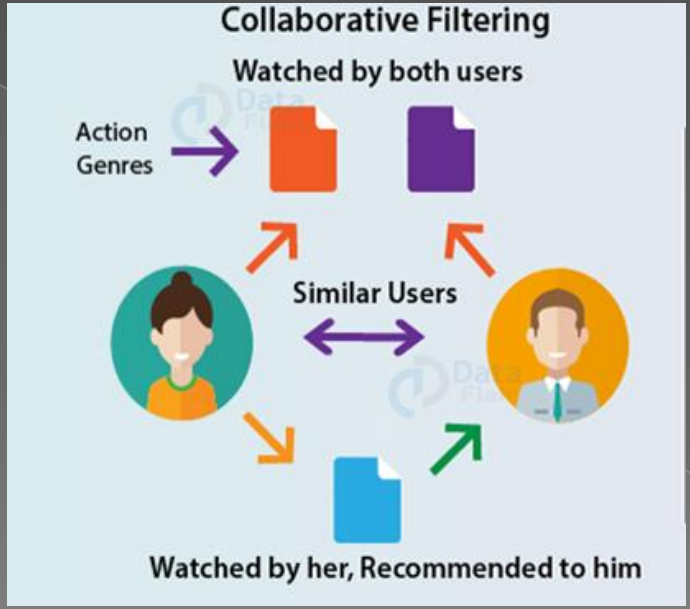


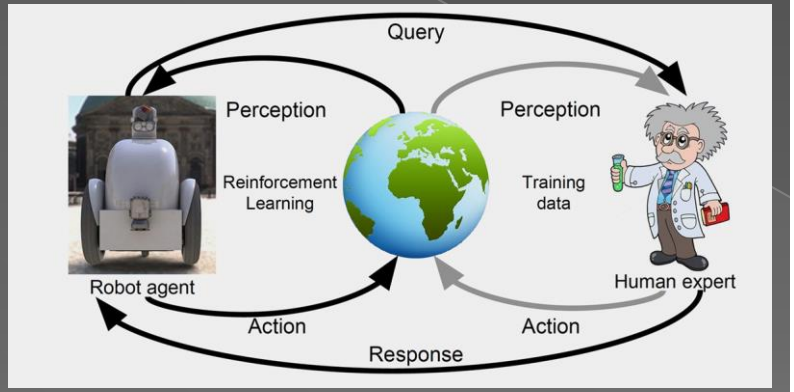
Image Courtesy: https://data-flair.training/blogs/data-science-r-movie-recommendation/

Reinforcement Learning

Computer learns to perform a task through repeated trial-and-error interactions with a dynamic environment.

- Accompany an example with positive or negative feedback according to the solution the algorithm proposes
- Learning by trial and error
- "how to act or behave when given occasional reward or punishment signals"

Reinforcement Learning Real Time Applications



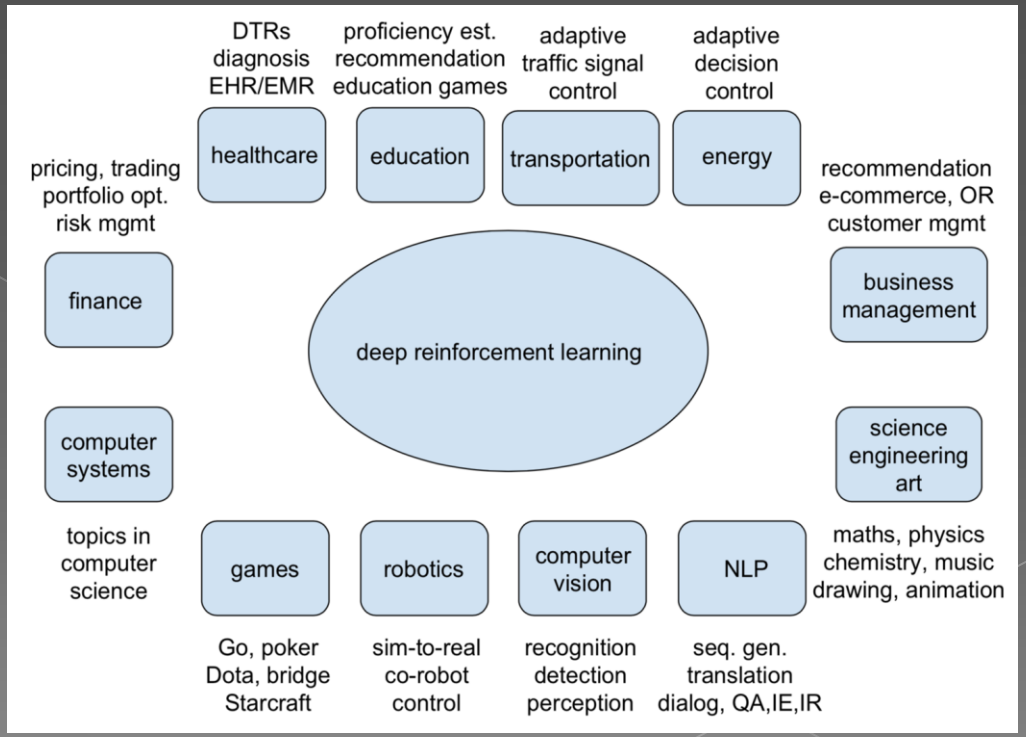


Image Courtesy: https://chatbotsmagazine.com/reinforcement-learning-and-its-practical-applications-8499e60cf751
https://medium.com/@yuxili/rl-applications-73ef685c07eb

Parametric vs Non-parametric models

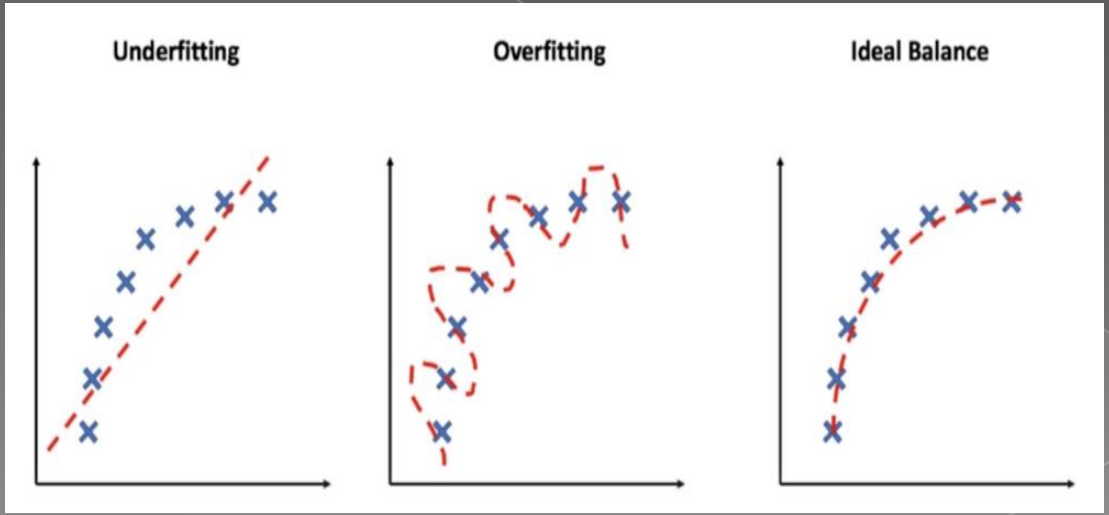
- **Parametric models**
  - Summarizes data with a set of **parameters of fixed size** (independent of the number of training examples) is called a **parametric model**.
  - Assume fixed form for mapping function

    Eg: Naive Bayes, Simple Neural Network, Logistic regression etc

- **Non-parametric models**
  - Non-parametric methods are good when you have a lot of data and no prior knowledge
  - Do not make strong assumptions about the form of the mapping function
  - Eg: KNN, SVM, Decision Tree

Overfitting

- **Modeling error** that occurs when a function is too closely fit to a limited set of data points.
- Happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data



Model Selection

- How should we pick the right one?

Compute the misclassification (error) rate of each method, select with minimum error

  - Training error
  - Generalization error (test error)
  - Validation error
- Cross Validation
  - K-Fold CV
  - Leave One Out Cross Validation (LOOCV)