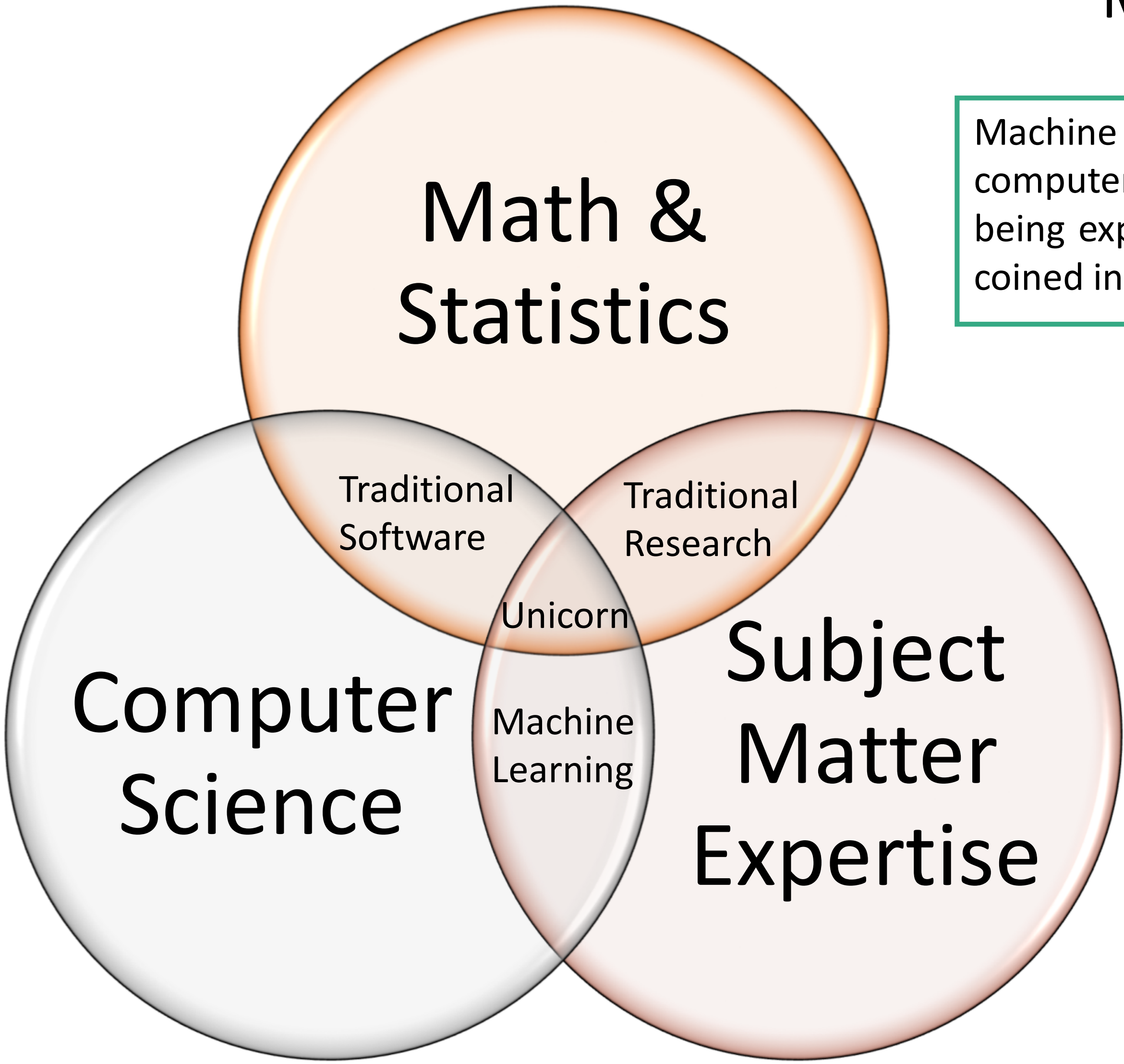
Machine Learning

Machine Learning
Introduction

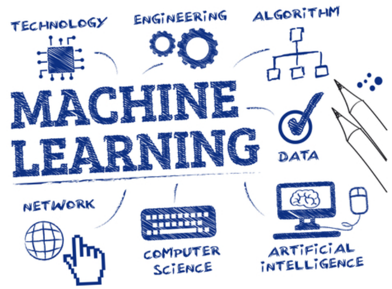
Machine Learning Introduction



Machine learning is a field of computer science that gives computer systems the ability to "**learn**" with data, without being explicitly programmed. The name Machine learning was coined in 1959 by Arthur Samuel.

Machine Learning Introduction

- ML is a system which can do automatic acquisition and integration of knowledge.
- It is that branch of artificial intelligence that deals with the construction of systems that can learn from data
- Develop methods that can automatically detect patterns in data, and then to use these patterns to predict future data
- Machine learning can predict the future based on the past
- Computer programs that automatically improve their performance through experience



Why Machine Learning?

- Automatically adapt and customize to individual users.
  - Personalized news, mail filters, movie/book recommendation
- Discover new knowledge from huge amount of data
  - Market analysis
- Perform repetitive monotonous tasks of humans which require intelligence and experience
  - Recognize signatures or handwritten characters
  - Driving a car, flying a plane
- Rapidly changing phenomenon
  - Credit scoring, financial modeling, diagnosis, fraud detection
- No human experts industrial/manufacturing control, mass spectrometer analysis, drug design

Machine Learning

Concepts & Dimensions of Machine Learning

Concepts of Learning

- Learning = Improve Task “**T**” with respect to performance measure “**P**” based on experience “**E**”

- **Example: Spam Filtering**
  - T: Identify Spam emails
  - P: % of Spam emails filtered correctly, % of non-Spam emails that were filtered incorrectly (false positives)
  - E: Database of emails labelled manually by users

- **A checkers learning problem:**
  - **Task T:** playing checkers
  - **Performance measure P:** percent of games won against opponents
  - **Training experience E:** playing practice games against itself

  We can specify many learning problems in this fashion, such as learning to recognize handwritten words, or learning to drive a robotic automobile autonomously

Concepts of Learning

- **Example: Signature matching**
  - T: Determine if signature belongs to correct person
  - P: % of signatures that were correctly matched, % of valid signatures that were incorrectly labelled as not matching
  - E: Database of signatures known to be of that person

Dimensions Of Learning Systems

- **What is Dimension of Learning?**
  - Dimensions of Learning is a comprehensive model that uses what researchers and theorists know about learning to define the learning process. There are five types of thinking – what we call the five dimensions of learning- are essential to successful learning.

- **Dimension 1: Attitudes and Perceptions**
  - Attitudes and perceptions affect students’ abilities to learn. For example, if students view the classroom as an unsafe and disorderly place, they will likely learn little there.

Dimensions Of Learning Systems

- **Dimension 2: Acquire and Integrate Knowledge**
  - When students are learning new information, they must be guided in relating the new knowledge to what they already know, organising that information, and then making it part of their long-term memory.

- **Dimension 3: Extend and Refine Knowledge**
  - Learning does not stop with acquiring and integrating knowledge. Learners develop in-depth understanding through the process of extending and refining their knowledge

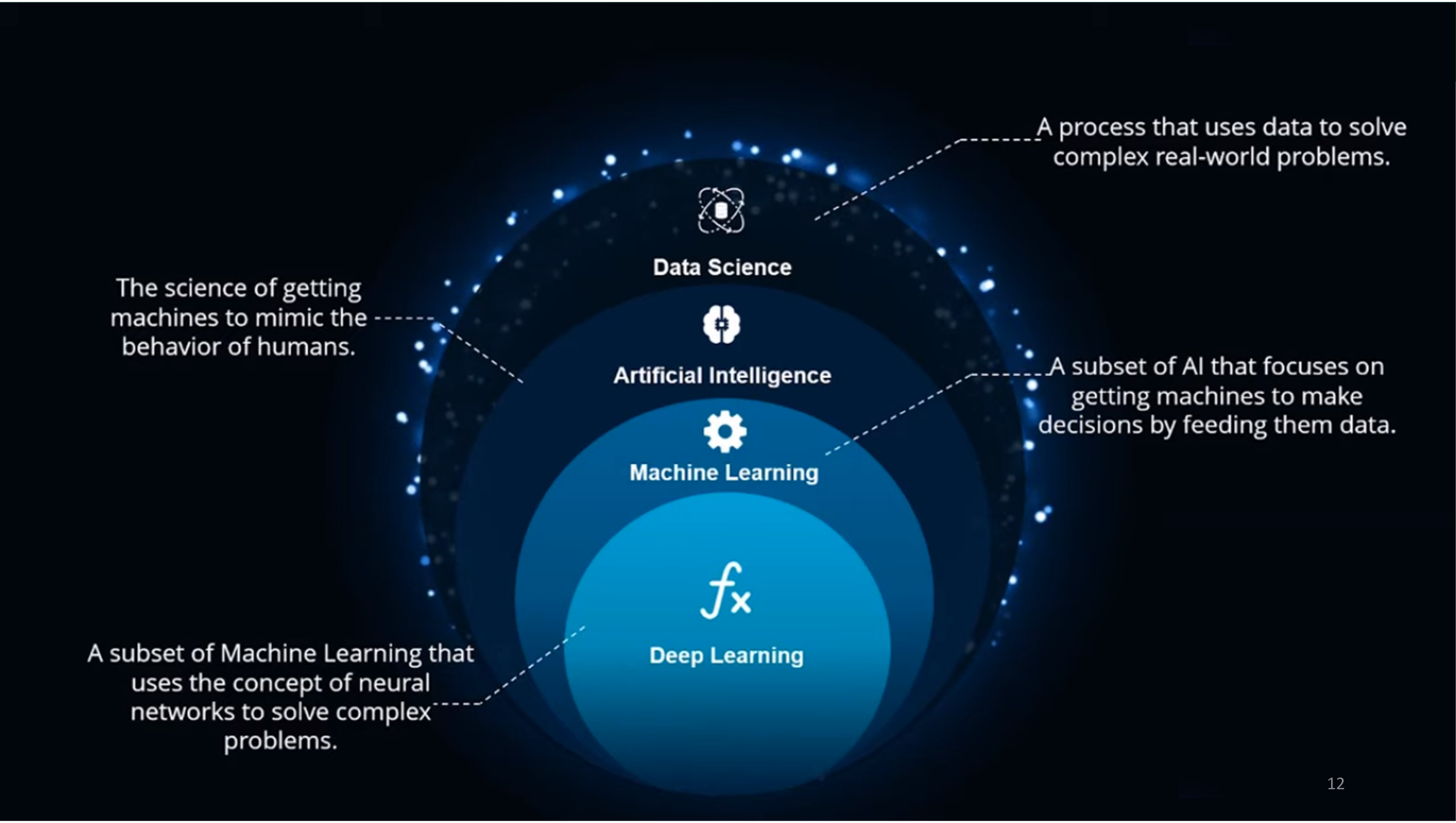
Dimensions Of Learning Systems

- **Dimension 4: Use Knowledge Meaningfully**
  - The most effective learning occurs when we use knowledge to perform meaningful tasks. For example, we might initially learn about tennis rackets by talking to a friend or reading a magazine article about them.

- **Dimension 5: Habits of Mind**
  - The most effective learners have developed powerful habits of mind that enable them to think critically, think creatively, and regulate their behaviour.

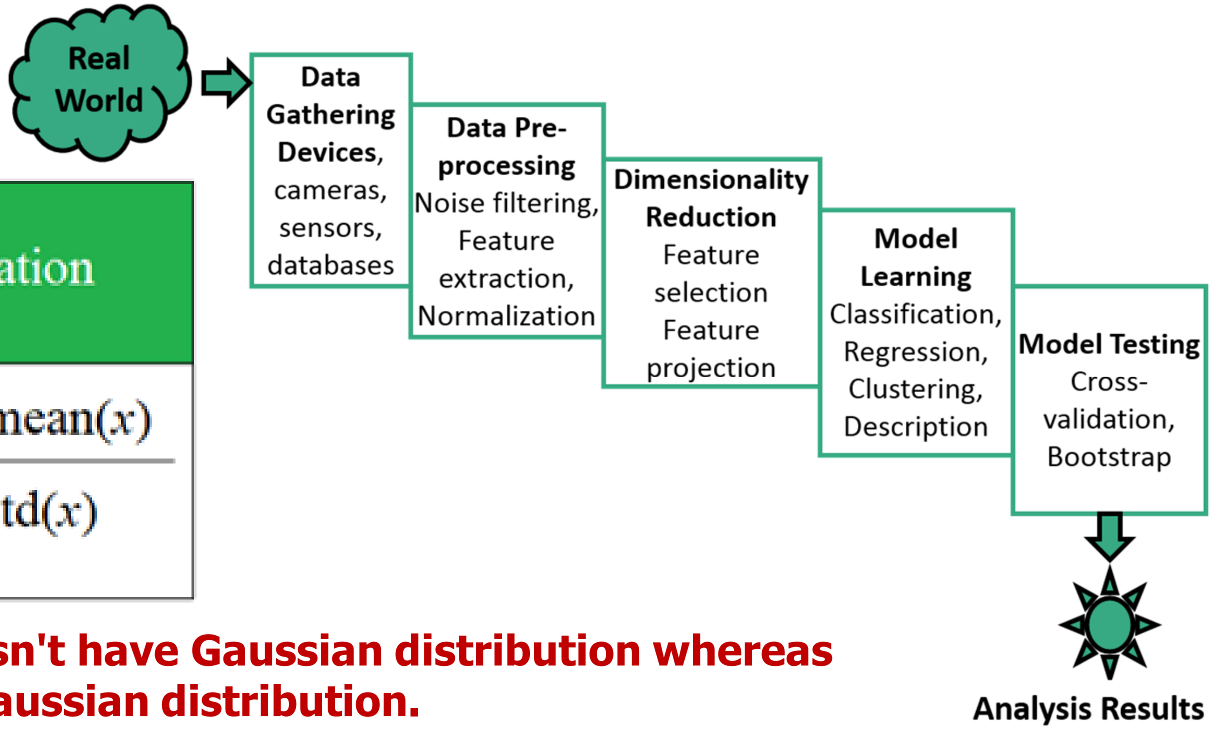
Machine Learning

The Learning Process



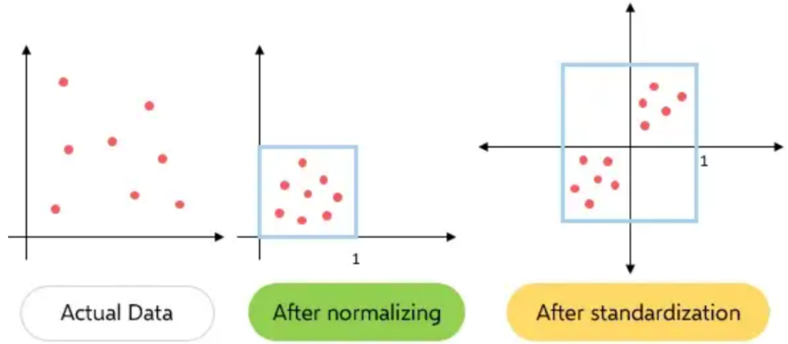
A process that uses data to solve complex real-world problems.
Data Science
The science of getting machines to mimic the behavior of humans.
Artificial Intelligence
A subset of AI that focuses on getting machines to make decisions by feeding them data.
Machine Learning
fx
Deep Learning
A subset of Machine Learning that uses the concept of neural networks to solve complex problems.

The learning Process



Normalization	Standardization
$x_{\text{norm}} = \dfrac{x - \min(x)}{\max(x) - \min(x)}$	$x_{\text{stand}} = \dfrac{x - \text{mean}(x)}{\text{Std}(x)}$

- **Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.**
- **Normalization scales in a range of [0, 1] or [-1, 1]. Standardization is not bounded by range.**
- **Normalization is highly affected by outliers. Standardization is slightly affected by outliers.**
- **Normalization is considered when the algorithms do not make assumptions about the data distribution. Standardization is used when algorithms make assumptions about the data distribution.**



Normalization	Standardization
$x_{\text{norm}} = \dfrac{x - \min(x)}{\max(x) - \min(x)}$	$x_{\text{stand}} = \dfrac{x - \text{mean}(x)}{\text{Std}(x)}$

The learning Process

Gaussian distribution

- Gaussian or normal distribution, 1D

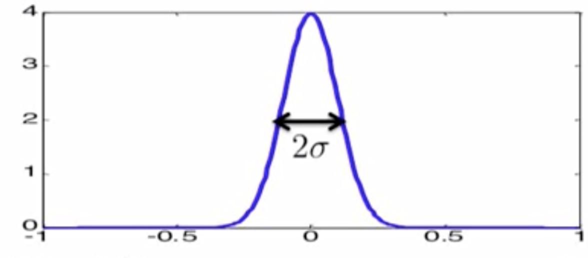
$$\mathcal{N}(x\ ;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}(x-\mu)^2/\sigma^2\right]$$

  - Parameters: mean μ, variance σ^2 (standard deviation σ)

Maximum Likelihood estimates

$$\hat{\mu} = \frac{1}{N}\sum_i x^{(i)}$$

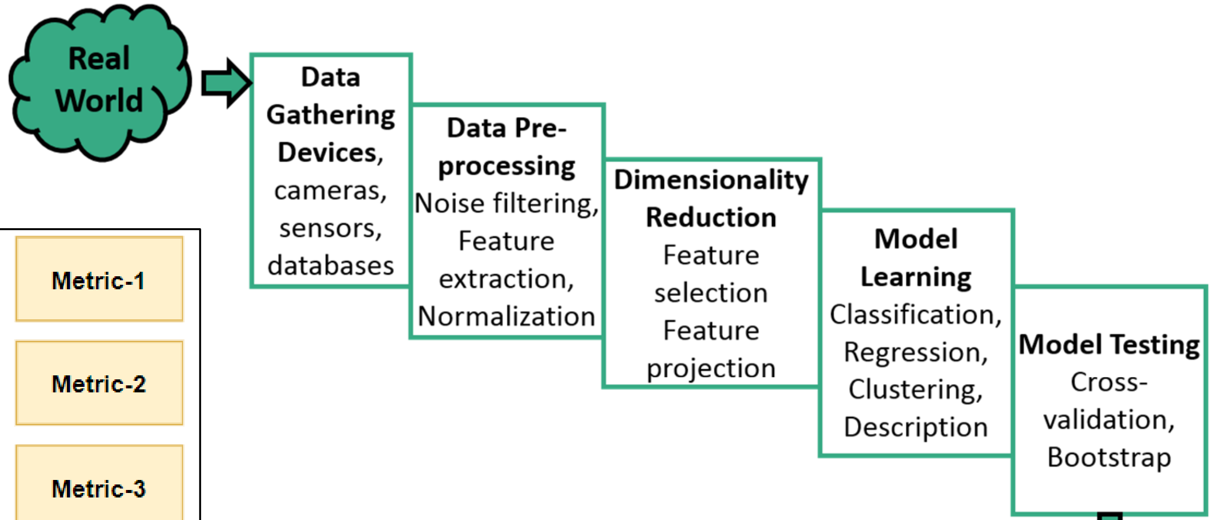
$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x^{(i)} - \hat{\mu})^2$$

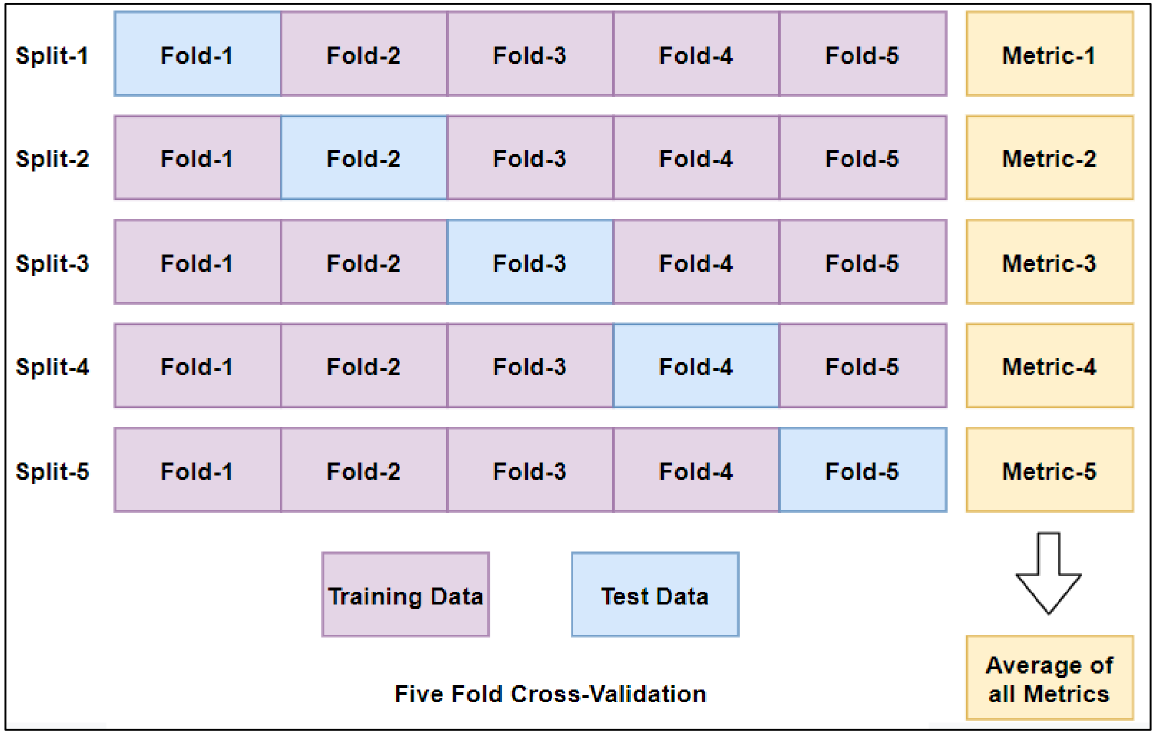


(c) Alexander Ihler 2

- ❑ **Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.**
- ❑ **Normalization scales in a range of [0, 1] or [-1, 1]. Standardization is not bounded by range.**
- ❑ **Normalization is highly affected by outliers. Standardization is slightly affected by outliers.**
- ❑ **Normalization is considered when the algorithms do not make assumptions about the data distribution. Standardization is used when algorithms make assumptions about the data distribution.**

The learning Process





Bootstrap aggregating, also called bagging, is a **machine learning** ensemble meta-algorithm designed to improve the stability and accuracy of **machine learning** algorithms used in statistical classification and regression.

The learning Process

Bagging

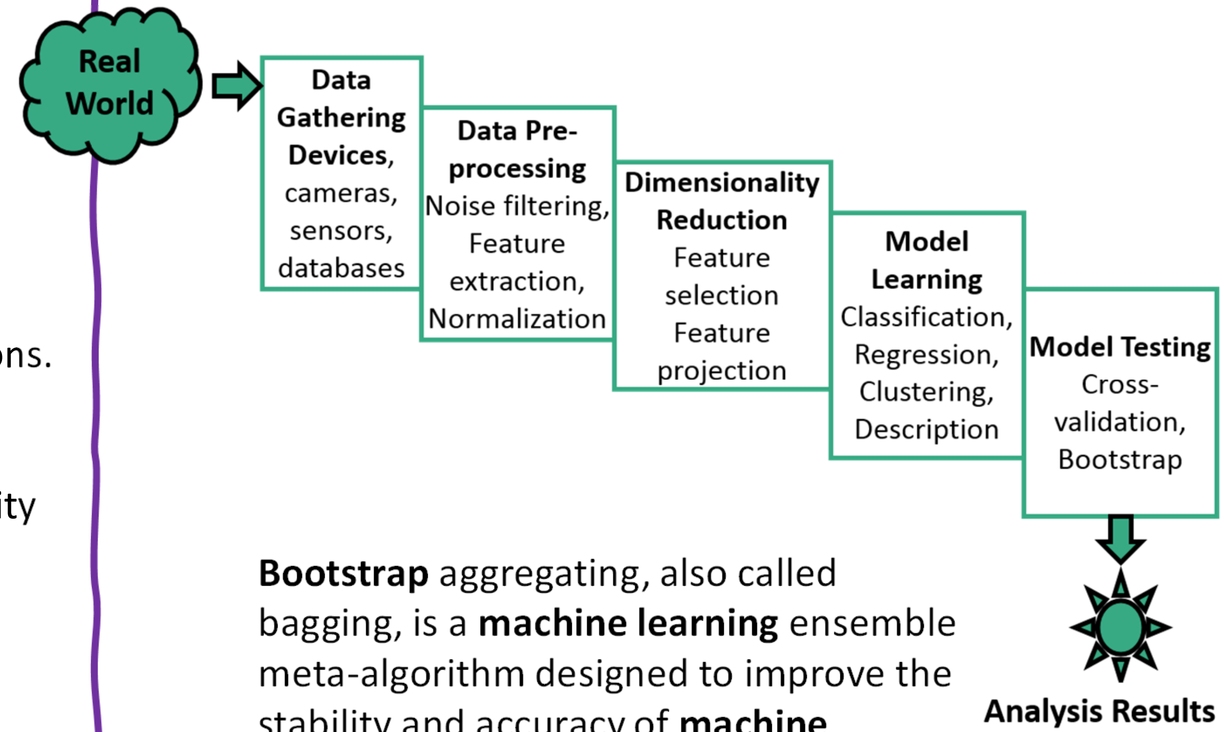
Bagging attempts to reduce the chance overfitting complex models.

- It trains a large number of "strong" learners in parallel.
- A strong learner is a model that's relatively unconstrained.
- Bagging then combines all the strong learners together in order to "smooth out" their predictions.
- Example – Random Forest Classifier.

Boosting

Boosting attempts to improve the predictive flexibility of simple models.

- It trains a large number of "weak" learners in sequence.
- A weak learner is a constrained model (i.e., you could limit the max depth of each decision tree).
- Each one in the sequence focuses on learning from the mistakes of the one before it.
- Boosting then combines all the weak learners into a single strong learner.
- Example – Boosted Tree Classifier



Bootstrap aggregating, also called bagging, is a **machine learning** ensemble meta-algorithm designed to improve the stability and accuracy of **machine learning** algorithms used in statistical classification and regression.

Hyper Parameters and Model Parameters

What are Hyper Parameters?

So far, we've been casually talking about "tuning" models. When we talk of tuning models, we specifically mean tuning hyper-parameters.

There are two types of parameters in machine learning algorithms. The key distinction is that model parameters can be learned directly from the training data while hyper-parameters cannot. Hyper-parameters express "higher-level" structural settings for algorithms. They are decided before fitting the model because they can't be learned from the data.

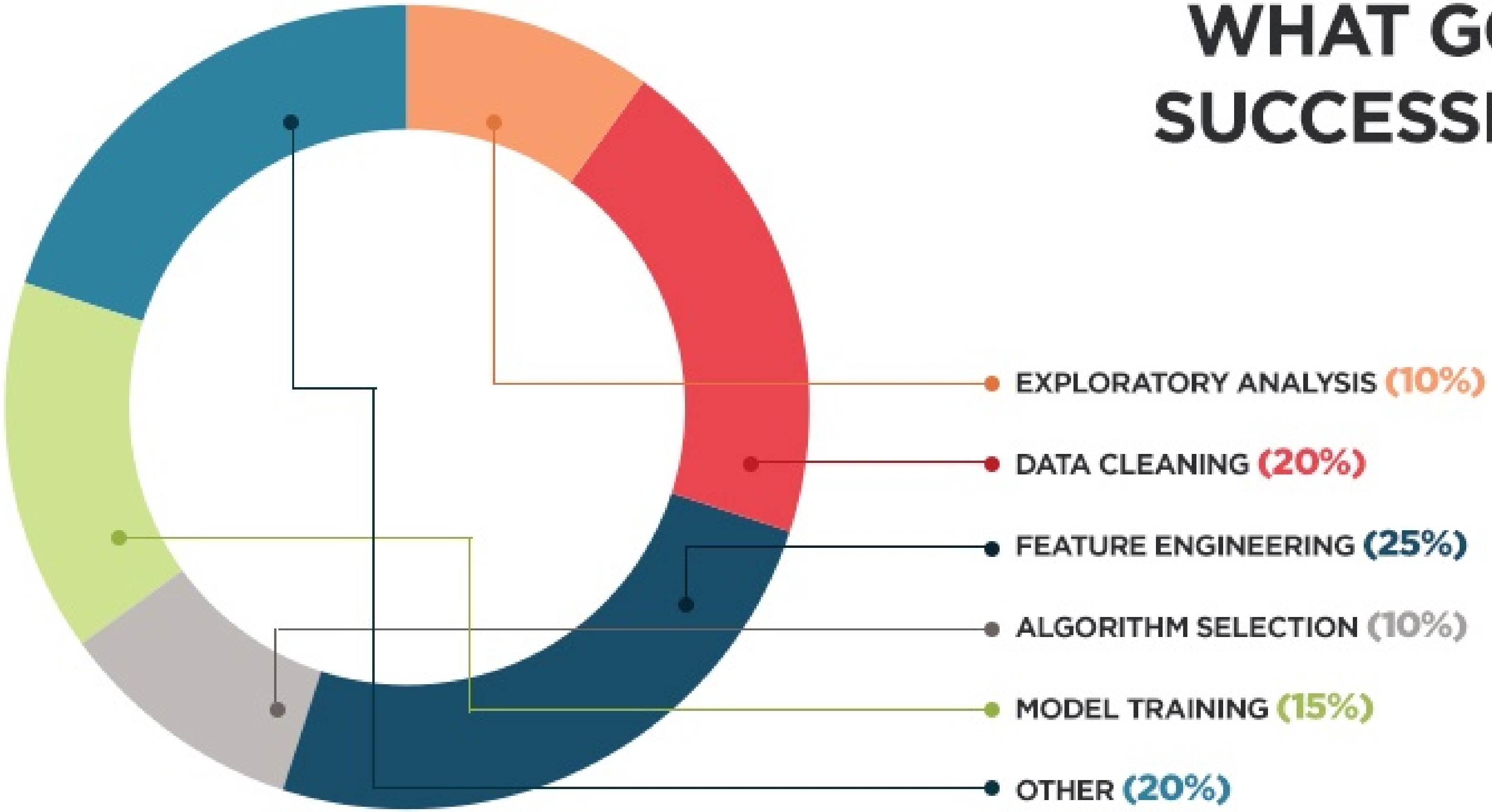
- **As Example:** **Strength of the penalty used in Regularized Regression**
- **As Example:** **The number of trees to include in a Random Forest**

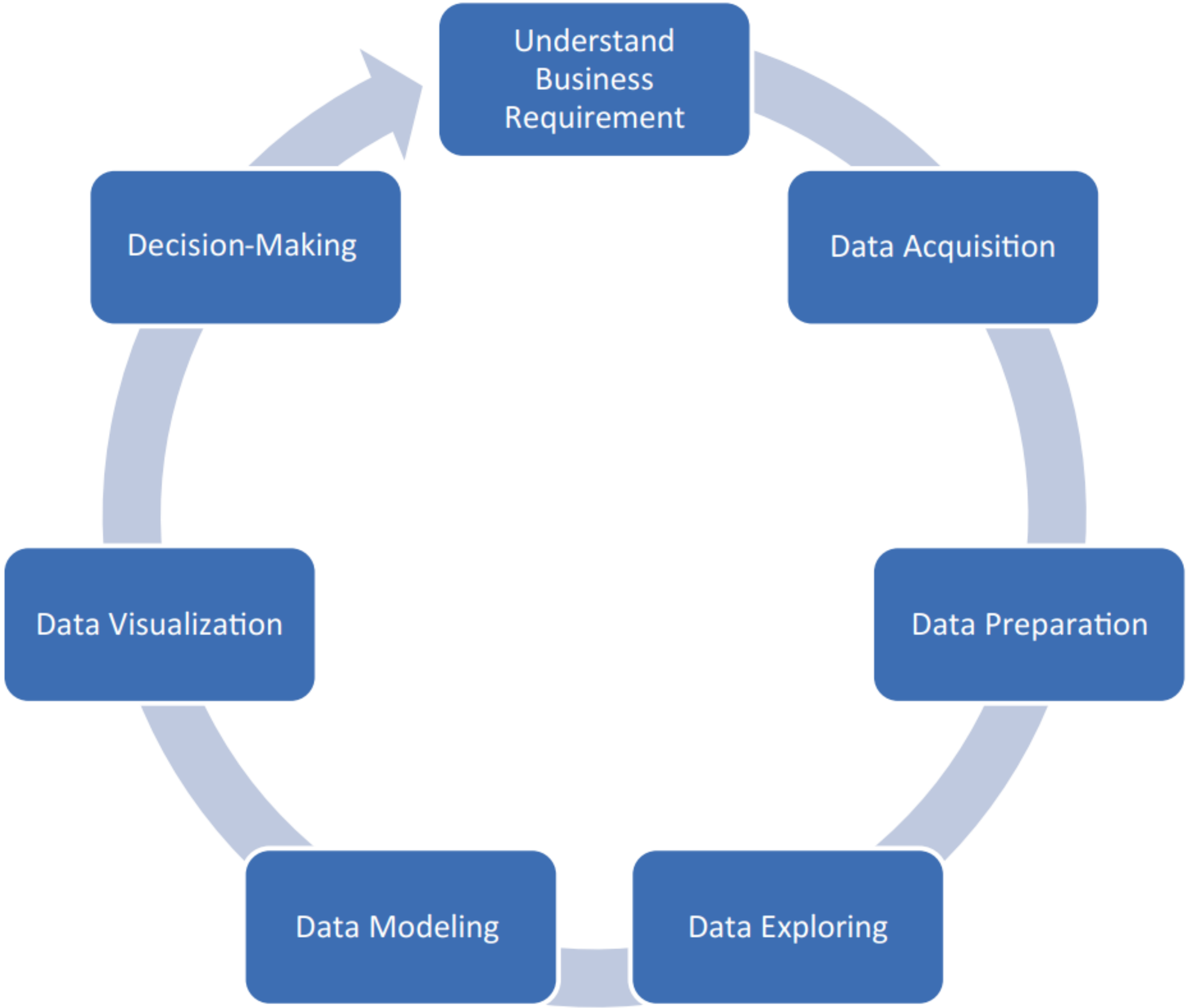
What are Model Parameters?

Model parameters are learned attributes that define individual models. They can be learned directly from the training data.

- **As Example:** **Regression coefficients**
- **As Example:** **Decision Tree split locations**

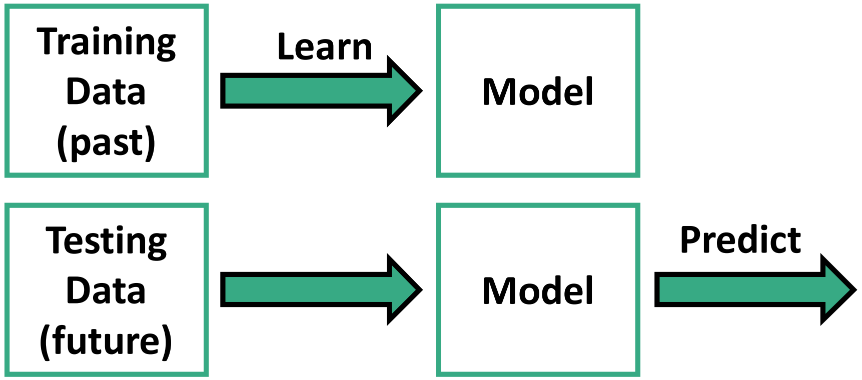
WHAT GOES INTO A SUCCESSFUL MODEL





Data science project stages

Machine Learning Steps



➢ **Steps:**

- Gather Data from various sources
- Clean data to have homogeneity
- Build model (select the right Machine Learning algorithm)
- Gather insights from the model's results
- Visualize – transform results into visual graphs

Performance Evaluation

- Randomly split examples into training set U and also test set V.
- Use training set to learn a hypothesis H.
- Measure % of V correctly classified by H.
- Repeat for different random splits and average results.

Problems - Overfitting & Underfitting

- **Overfitting** – the model learns the training set too well. It overfits to the training set and performs poorly on the test set

- **Underfitting** – when the model is too simple, both training and test errors are large

Machine Learning

Categorization of Machine Learning

Categorization of Machine Learning

Machine Learning

Supervised

- Task Driven, labelled (Regression/Classification).
- Input variables X and output variable y. y = f(X).
- Algorithm learns from training data set. Iteratively make predictions which are corrected (teaching).
- Learning stops when performance reaches acceptable level.

Unsupervised

- Data Driven, unlabeled (Clustering)
- No correct answers, no teacher, no labelling
- Algorithm looks for patterns in data.

Reinforcement

- Reward feedback based
- Algorithm learns to react to an environment. Automatically determine ideal behavior within a context.

Data Set for Supervised Learning

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
5	2	3.5	1	versicolor
5.9	3	4.2	1.5	versicolor
6.2	3.4	5.4	2.3	virginica
5.9	3	5.1	1.8	virginica

Data Set for Supervised Learning

The most obvious visual difference between cupcakes and muffins is, of course, the frosting. Cupcakes are topped with creamy, delicious frosting. Instead, muffins may have a sugared top or a very thin glaze. Usually the fillings inside the muffins add enough excitement to the baked good, so there might not be anything on top.

Type	Flour	Milk	Sugar	Butter	Egg	Baking Powder	Vanilla	Salt
Muffin	55	28	3	7	5	2	0	0
Muffin	47	24	12	6	9	1	0	0
Muffin	47	23	18	6	4	1	0	0
Muffin	45	11	17	17	8	1	0	0
Muffin	50	25	12	6	5	2	1	0
Cupcake	42	18	25	9	5	1	0	0
Cupcake	36	14	21	14	11	2	1	0
Cupcake	38	15	31	8	6	1	1	0
Cupcake	36	16	24	12	9	1	1	0
Cupcake	34	17	23	11	13	0	1	0

Data Set for Unsupervised Learning

coffee/ tea	sau- ces	confec- tionary	puddings/ deserts	fro- zen	razor blades	spi- ces	jam- jelly	cold- drinks	deter- gent	tiss- ues	toilet clearner
1	1	0	1	1	1	1	0	0	0	0	1
0	0	0	0	1	0	0	0	0	0	0	1
1	0	0	0	1	0	0	1	0	0	0	0
0	0	0	0	1	0	0	1	0	0	1	0
1	0	1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	1	0	1	1	1	0
0	0	0	0	0	0	1	0	1	1	1	0
0	1	0	0	0	1	0	0	0	0	0	1
0	0	0	1	0	1	1	0	0	0	0	1
0	1	0	1	1	0	0	1	0	0	0	0
0	0	1	1	1	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	1	0	0	1	0	1	0
0	0	0	0	0	0	0	0	1	0	0	0

Machine Learning Coordinates

	Supervised Learning	**Unsupervised Learning**
Discrete	**Classification or Categorization**	**Clustering**
Continuous	**Regression**	**Dimensionality Reduction**

Machine Learning Coordinates

	Supervised Learning	Unsupervised Learning
Discrete	**Classification or Categorization**	**Clustering**
Continuous	**Regression**	**Dimensionality Reduction**

In statistics, **machine learning**, and information theory, **dimensionality reduction** or **dimension reduction** is the process of **reducing** the number of random variables under consideration by obtaining a set of principal variables. It can be divided into feature selection and feature extraction

Classification & Clustering

- **Classification** is the problem of identifying to which of a set of categories (classes) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

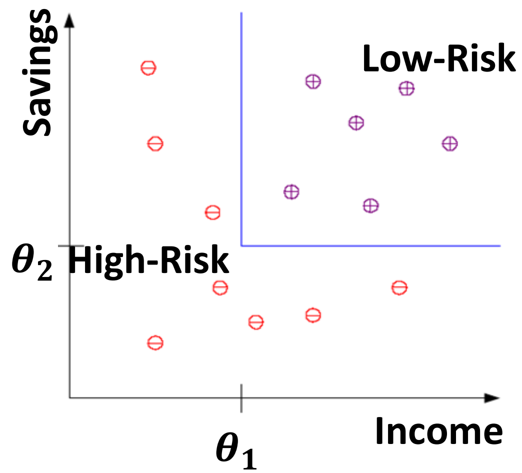
- Classification is considered an instance of *supervised learning*, i.e. learning where a training set of correctly identified observations are available.

- The corresponding *unsupervised* procedure is known as **clustering**, and involves grouping data into categories based on a measure of similarity or distance.

Classification & Clustering

- **Example:** This is an example of a *classification* problem where there are two classes: low-risk and high-risk customers. The information about a customer makes up the *input* to the classifier whose task is to assign the input to one of the two classes. After training with the past data, a classification rule learned may be of the form

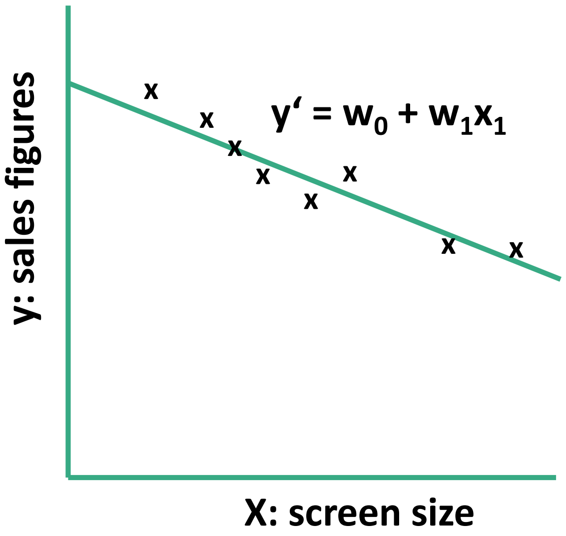
IF Income > θ1 AND Savings > θ2 THEN
Low-Risk
ELSE
High-Risk



Regression

- A regression problem is when the output variable is a real value, such as “dollars” or “weight” instead of a class.
- Estimate the relationship between a dependent variable and one or more independent variables (or 'predictors’)

Sales figures for a television model can depend on several factors like screen size, display type, brand, resolution, technology etc.
Here, we consider just one attribute, screen size and plot the corresponding prices
x: screen size
y: sales figures
We try to find the relation (function) that best matches these values



Dimensionality Reduction

- **Dimensionality reduction** is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

- Feature selection approaches try to find a subset of the original variables (also called features or attributes). It is about choosing some of features based on some statistical score.

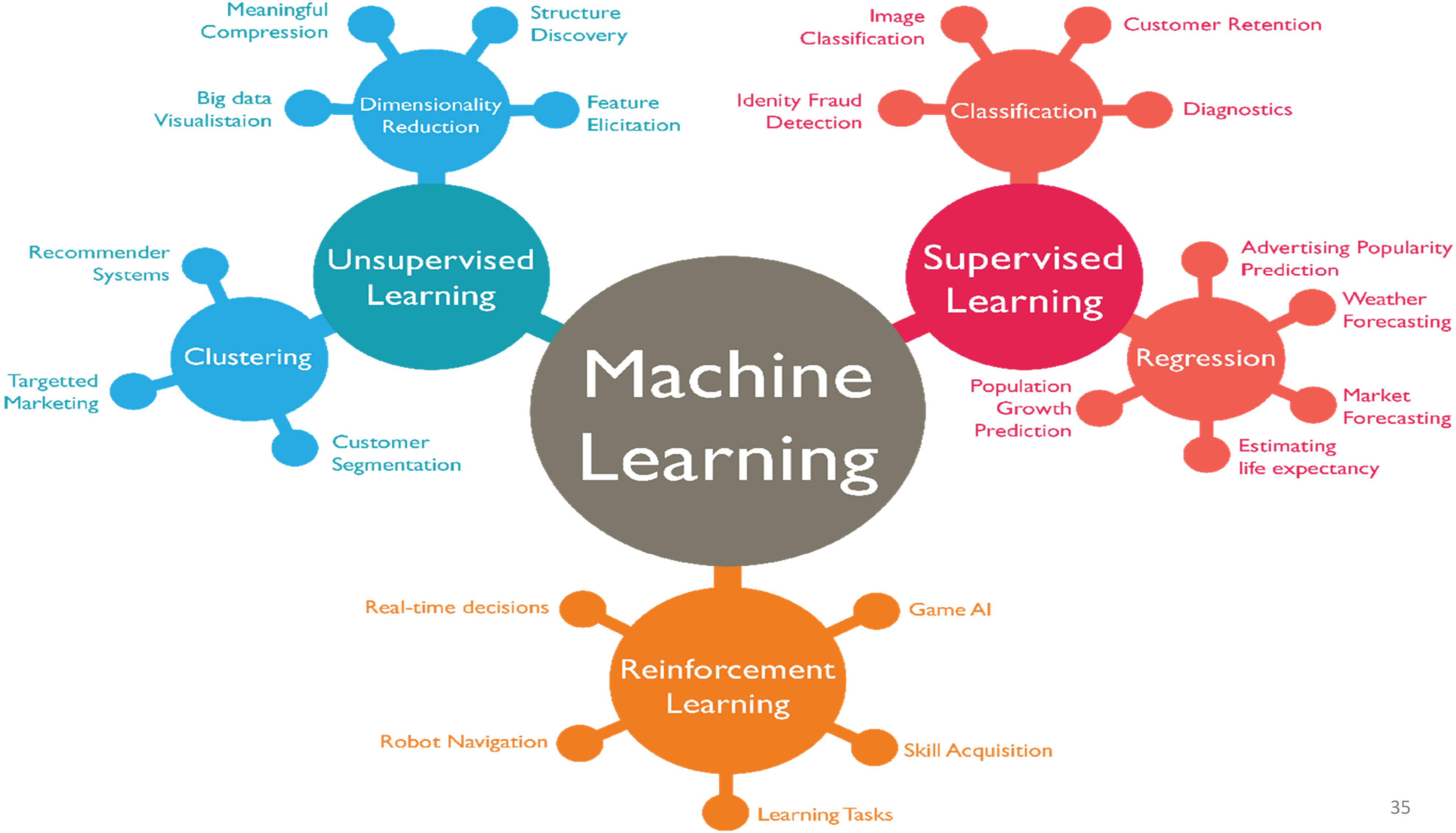
- Feature extraction transforms the data in the high-dimensional space to a space of fewer dimensions. It is using techniques to extract some second layer information from the data e.g. interesting frequencies of a signal using Fourier transform.

- Dimensionality reduction helps in data compression, reduces computation time and removes redundant features.

Machine Learning

Machine Learning Applications

Applications of Machine Learning



Classification Applications

- **Face Recognition**
  - Identify or verify a person from a digital image or a video frame
- **Character Recognition**
- **Spam detection**
- **Medical Diagnosis**
  - Determine which disease or condition explains a person's symptoms and signs.
- **Biometrics**
  - Authentication using physical and/or behavioral characteristics: Face, iris, signature, etc

Regression Applications

- **Economics/Finance:** predict the value of a stock
- **Epidemiology**
  - incidence, distribution, and possible control of diseases and other factors relating to health
- **Car/plane navigation:** angle of the steering wheel, acceleration
- **Temporal trends:** weather over time

Manufacturing & Retail Industries

Manufacturing

- Predictive maintenance or condition monitoring
- Warranty reserve estimation
- Demand forecasting
- Process optimization
- Telematics

Retail

- Predictive inventory planning
- Recommendation engines
- Upsell & cross-channel marketing
- Market segmentation & targeting
- Customer ROI & lifetime value

Healthcare & Life Science & Travel & Hospitality

Healthcare & Life Science

- Alerts & diagnostics from real-time patient data
- Disease identification & risk stratification
- Patient triage optimization
- Proactive health management
- Healthcare provider sentiment analysis

Travel & Hospitality

- Aircraft scheduling
- Dynamic pricing
- Social media – consumer feedback & interaction analysis
- Customer complaint resolution
- Traffic patterns & congestion management

Financial Services & Energy, Feedstock & Utilities

Financial Services

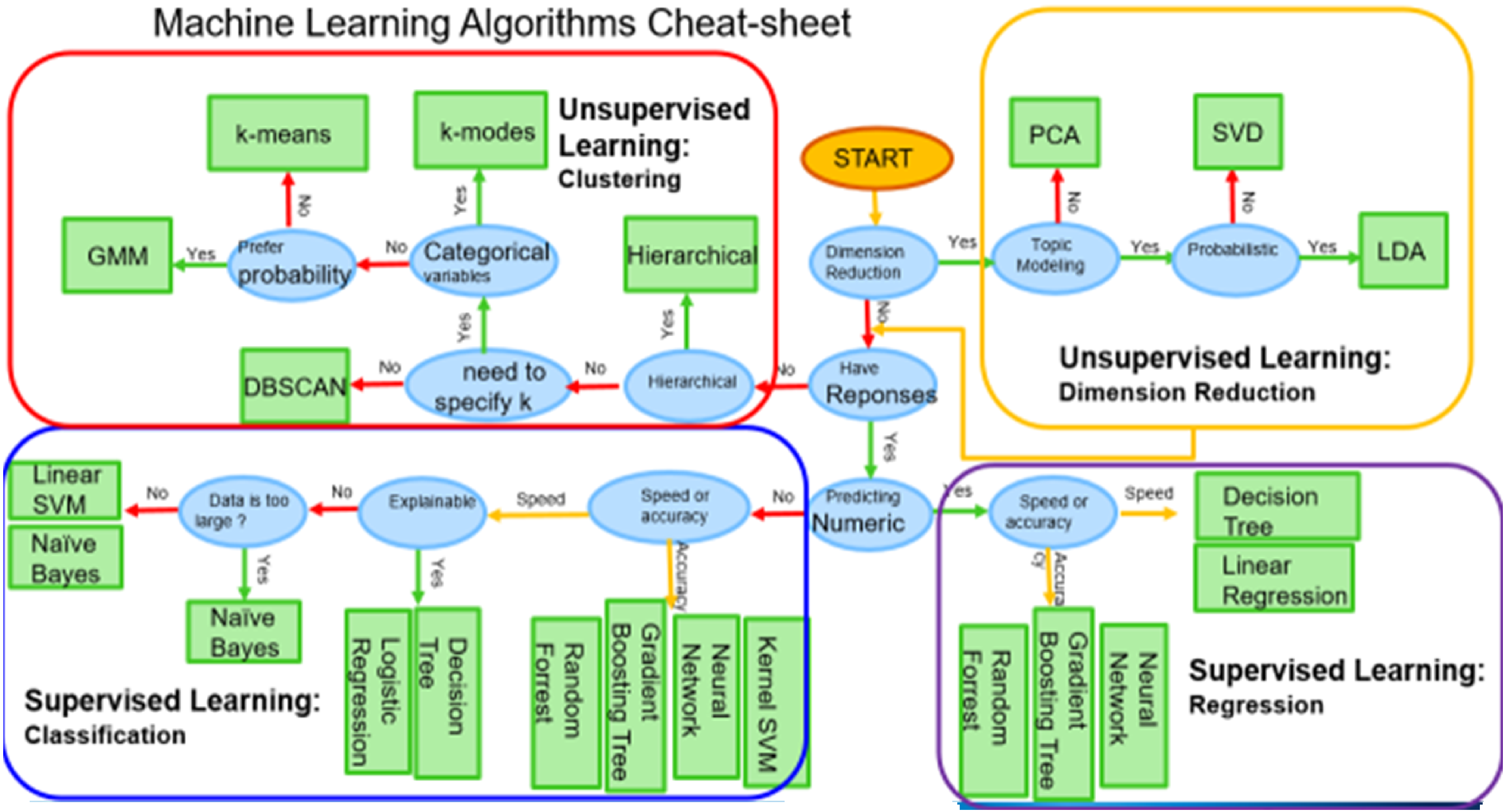
- Risk analytics & regulation
- Customer Segmentation
- Cross-selling & up selling
- Sales & marketing Campaign management
- Credit Worthiness evaluation

Energy, Feedstock & Utilities

- Power usage analytics
- Seismic data processing
- Carbon emissions and trading
- Customer-specific pricing
- Smart grid management
- Energy demand & supply optimization

Machine Learning

Machine Learning Algorithms



Machine Learning Algorithms Cheat-sheet
START
Dimension Reduction
Yes
No
Have Reponses
Yes
No
Predicting Numeric
Yes
No
Unsupervised Learning: Clustering
k-means
k-modes
GMM
Yes
Prefer probability
No
Categorical variables
Yes
No
Hierarchical
Yes
Hierarchical
No
need to specify k
Yes
No
DBSCAN
Unsupervised Learning: Dimension Reduction
PCA
SVD
No
Topic Modeling
Yes
Probabilistic
Yes
LDA
Supervised Learning: Classification
Linear SVM
Naïve Bayes
No
Data is too large ?
Yes
Naïve Bayes
No
Explainable
Yes
Logistic Regression
Decision Tree
Speed
Speed or accuracy
Accuracy
Random Forrest
Gradient Boosting Tree
Neural Network
Kernel SVM
Supervised Learning: Regression
Speed or accuracy
Speed
Decision Tree
Linear Regression
Accuracy
Random Forrest
Gradient Boosting Tree
Neural Network

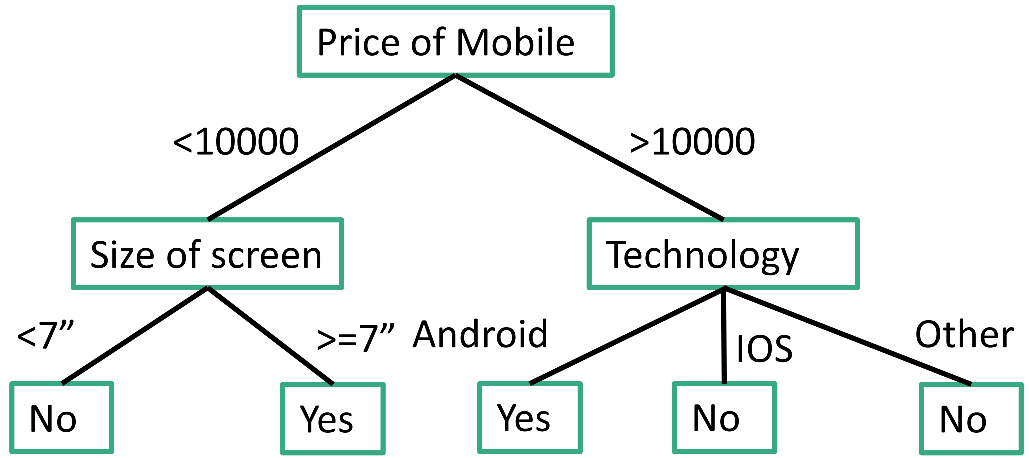
Decision Trees

- A decision tree takes as input an object or situation described by a set of properties and outputs a yes/no **decision**.
- Each **decision node** tests the value of an input attribute.
- **Branches** from the node are all possible values of the attribute
- **Leaf nodes** supply the value (Yes/No) to be returned if that leaf is reached.

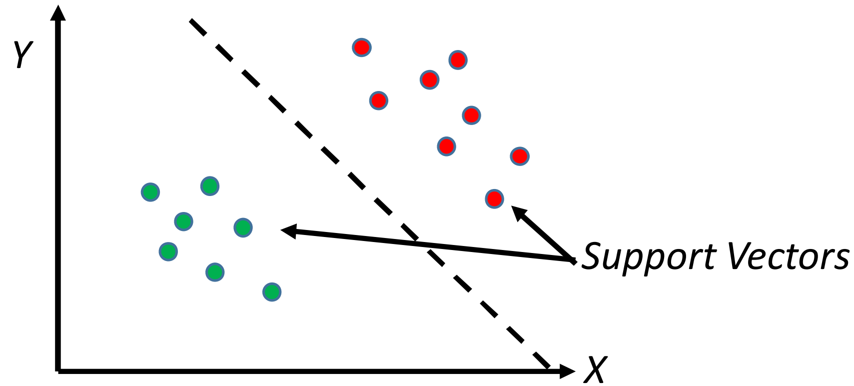
➢ Criteria used to choose the best nodes to build the most precise decision tree:
  - **Entropy -** degree of disorganization in our data. Entropy is 1 when collection has equal no. of positive and negative examples
  - **Information Gain** is used to determine the goodness of a split. The attribute with the most entropy reduction is chosen.

Example: Decision Tree

- **Decision Tree on whether to buy a Mobile Phone or not**



Support Vector Machines (SVM)



- Map data to *higher-dimensional space* where they will be *linearly separable*.
- Algorithm - plot each data item as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate.
- Then perform classification into 2 classes by finding the hyper-plane that differentiates the classes very well
- **Support Vectors** are the co-ordinates of the individual observations. **Support Vector Machine** is a frontier which best segregates the two classes (hyper-plane/ line).

Bayesian Networks

- Compute probability distribution for unknown variables given observed values of other variables.
- Start with a belief, called a **prior**
- Obtain some data and use it to update the belief. The outcome is called a **posterior**.
- Should we obtain even more data, the old posterior becomes a new prior and the cycle repeats.

➢ Obeys **Bayes rule**:
  - **P (A | B) = P (B | A) * P(A) / P(B)**
  - **P (A | B) is conditional probability, how likely is A if B happens?**

K Nearest Neighbor Model (k-NN)

- Idea: Properties of an input x are likely to be similar to those of points in the neighborhood of x
- Find (k) nearest neighbor(s) of x and infer target attribute value(s) of x based on corresponding attribute value(s).
- In k-NN classification, the output is a class membership. An object is assigned to the class most common among its k nearest neighbors.
- In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.
- To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. There can be various types of distance measures like Euclidean, Hamming, Manhattan etc.

Ensemble Learning

- Use multiple models to obtain better predictive performance than could be obtained from any of the individual constituent models

- Boosting – incrementally build an ensemble by training each new model instance to emphasize the training instances that previous instances misclassified

Deep Learning (Neural Networks)

- Subset of machine learning and covers all three paradigms using artificial neural networks (ANNs)

- ANNs are composed of multiple nodes that imitate the biological neurons of the human brain
  - Neurons are connected by links and interact with each other
  - Nodes can take input data and perform simple operations on the data. They pass the results to other neurons.
  - The output at each node is called its ”activation value” or ”node value”
  - Each link is associated with a weight

- ANNs “learn” by altering the link weight values

- Convolutional neural networks are specialized to read images as input, so are used for image recognition