UE17MC613
Machine Learning

Dr. S Thenmozhi

Course Outline

- 3 credit - Autonomy course
- 50 – 60% practice sessions
- ISA – 60
  - ISA1 – Quiz - 20 Marks, Practical – 20 Marks --- scaled to 20 Marks
  - ISA2 – Quiz – 20 Marks , Practical – 20 Marks --- scaled to 20 Marks
  - Datathon – 10 Marks
  - Assignment – 5 Marks - 20 Marks
  - App Development – 5 Marks
- ESA – 40
  - Theory – 40 Marks
  - Practical – 60 marks - 40 Marks

Agenda

- Unit 1 – Machine Learning
- Unit 2 – Support Vector Machines
- Unit 3 – Decision Trees
- Unit 4 – Artificial Neural Networks
- Unit-5 – ML application Development

"Go ahead and think that
I'm not really thinking. I thought
you would think that."

"Sweetheart, my neural net
predicts that you and I are
98.9% compatible.
Will you be my Valentine?"

Reading Resources

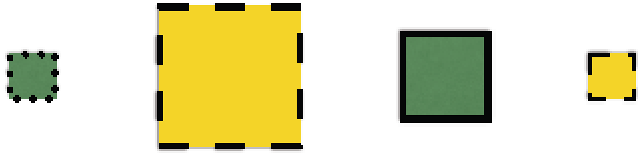
- Tom Mitchell, Machine Learning, McGraw Hill Publication, 2013
- Sebastian Raschk, Python Machine Learning, Packt Publishing, 2015
- Jake Vander Plas, Python Data Science Handbook, O'Reilly media, 2016
- Samir Madhavan, Mastering Python for Data Science, Packt Publishing, 2015
- Willi Richert, Luis Pedro Coelho, Building Machine Learning Systems with Python, 1st Edition, Packt Publishing, 2013
- Any online-material

A Few Quotes

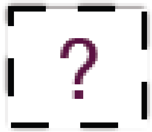
- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)
- “Machine learning is the next Internet” (Tony Tether, Director, DARPA)
- Machine learning is the hot new thing” (John Hennessy, President, Stanford)
- “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)
- “Machine learning is going to result in a real revolution” (Greg Papadopoulos, CTO, Sun)
- “Machine learning is today’s discontinuity” (Jerry Yang, CEO, Yahoo)

What is Machine Learning?

- Construct model by using algorithms and learn from data
- Use models for prediction
- More information --→ High Performance
- Previous solutions --→ Experience
- Eg: Label squares: size and edge -→ color



- Earlier Observations
- Task for Computer is label unseen square

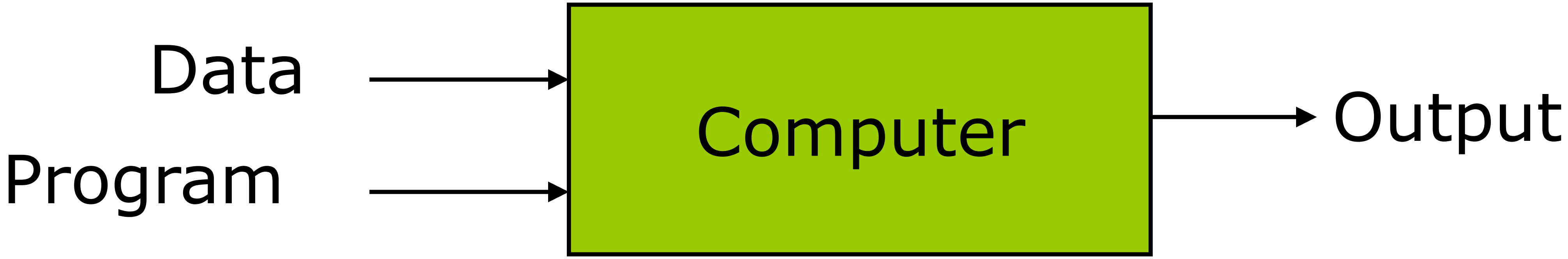


- Result: right or wrong
- Goal: Building models for prediction

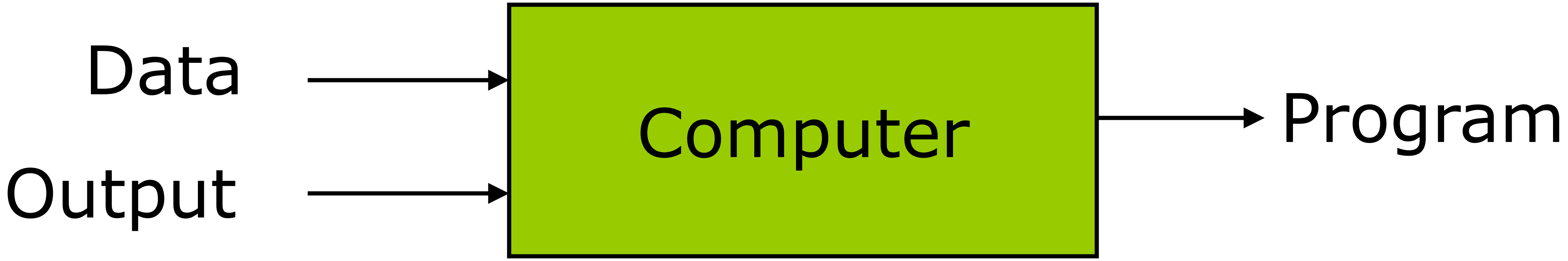
? Machine Learning by large

- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!

Traditional Programming



Machine Learning



Magic?

No, more like gardening

- **Seeds** = Data
- **Nutrients** = Algorithms
- **Gardener** = You
- **Plants** = Programs

Formulation

Estimated Function/model

Well–Posed Learning Problem

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T as measured by P, improves the experience E
- Eg: checkers learning problem
  - Task T : Playing checkers
  - Performance measure P: % of games won against opponent
  - Training Experience E : Playing practice games against itself

What is not ML?

- Just statistical analysis is not ML
  - Determining most occurring colour
  - Calculating average size

What is a Model and Algorithm?

- Algorithm is a set of steps performed in order
- A model is any sort of function that has the predictive power
- Models: Regression models, Classification models and Mixed models

Terminology

- Report – a static object with no predictive power
- Function – An object that has some kind of processing power, likely sits inside a model
- Model – A complex object that takes an input parameter and gives an output function
- Equation – A mathematical representation of a function. Sometimes a mathematical model
- Algorithm – A set of steps that are passed into a model for calculation or processing

ML Components

- Every machine learning algorithm has three components:
  - **Representation**
  - **Evaluation**
  - **Optimization**

Representation

- Decision trees
- Sets of rules / Logic programs
- Instances
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles
- Etc.

Evaluation

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
- Etc.

Optimization

- Combinatorial optimization
  - E.g.: Greedy search
- Convex optimization
  - E.g.: Gradient descent
- Constrained optimization
  - E.g.: Linear programming

Modelling Limitations

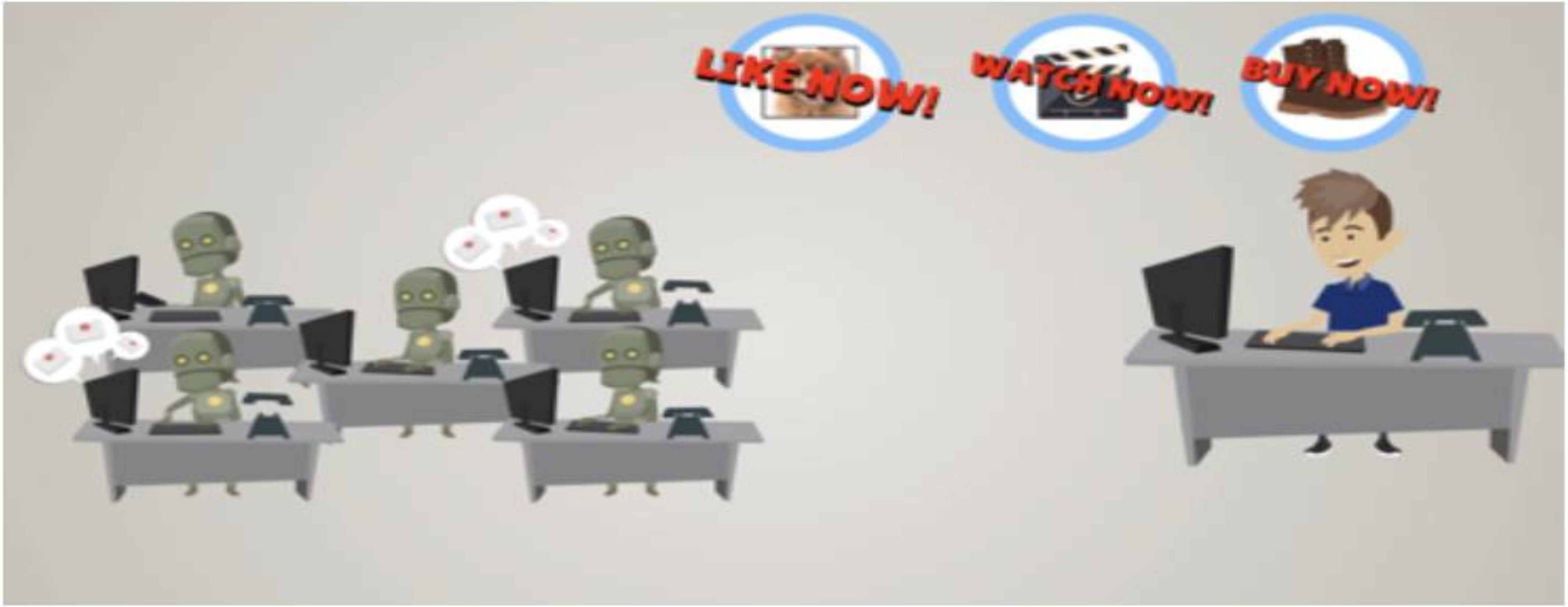
- A model is simplified picture of reality
- Models are just approximation of the universe that we are studying
- All models are bound to errors
- When new features are added the model has to be redesigned, re-evaluated or re-implemented to fit the observations
- A model might be limited by computational speed

Growth of ML

- ML is preferred for
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology

Recommendation Systems

- On Netflix, Amazon, and Facebook, everything that is recommended to you depends on your search activity, likes, and previous behaviour.
- Amazon has such amazing machine learning algorithms in place that it can predict with high certainty what you'll buy and when you'll buy it. The company even owns a patent for "anticipatory shipping", a system that ships a product to the nearest warehouse so you can order and receive your item on the same day


LIKE NOW!
WATCH NOW!
BUY NOW!

Algorithmic Stock Trading

- Random behaviour, ever-changing data, and a variety of factors — from political to judicial — that are far away from traditional finance.
- While financiers cannot predict much of that behaviour, machine learning algorithms can — and they respond to changes in the market much faster than a human.

Autonomous Vehicles

Many more…

- You can predict if an employee will stay with your company or leave.
- You can decide if a customer is worth your time, if they'll likely buy from a competitor, or not buy at all.
- You can optimize processes, predict sales, and discover hidden opportunities.
- You can predict who will be on thrones…

Understanding the Evolution

- Statistics
  - more theory-based
  - more focused on testing hypotheses
- Machine learning
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

Statistics and Computation in modelling

- Basis of mathematics and statistics
- To run machine learning models – mathematics is not much required
- Model tuning, hunt for bugs, assess model limitations – mathematics is required
- Statistics required to determine the training data
- It is not advisable to take the entire data and test it, b'coz, the model has already learnt with the data
- So, split into train and test set

How do have Train and Test? - Sampling methods

- Random Sampling: The data is picked randomly from the dataset.
- Stratified Sampling - The data is separated into mutually exclusive groups called strata and then simple random sampling is done on each stratum.
- Cluster sample – similar to stratified sampling but picking the entire strata randomly instead of doing a simple random sample in the strata.
- When samples were spread out geographically or spatially, perform cluster sample.

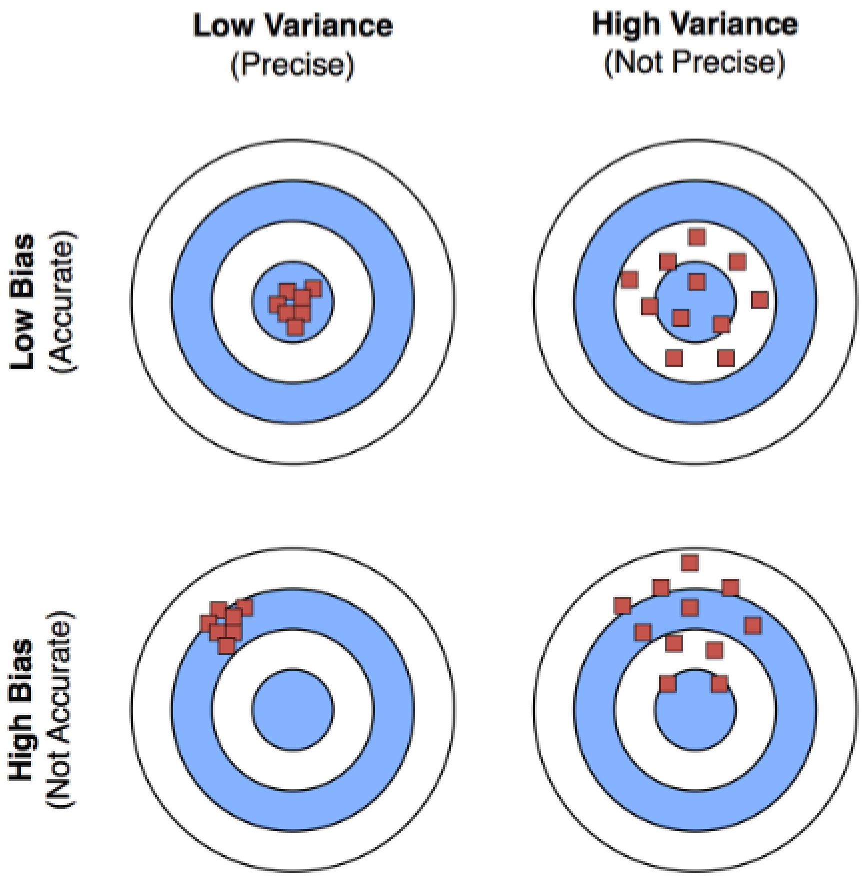
Bias and Variance

- Sampling Bias –Distribution of the sample taken doesn't match with the population from which they are drawing
- Sampling Variation – extent to which the sample statistic differs from the population
- Bias and variance with sampling can be represented in 4 ways
- Low bias, low variance – Best-scenario. samples are pretty well representative of the population
- High bias, low variance – samples are consistent, but not reflecting the population

- Low bias, high variance – samples are not consistent, but some reflect the population
- High bias, high variance – the samples are little more consistent, but not likely to represent the population
- A simple random sample is one way to control the bias and variance.
- Here, when you select values from your data at random such that every row has an equal chance of being selected
- **Sampling Error – caused by skewness of the variable**

Bias and variance

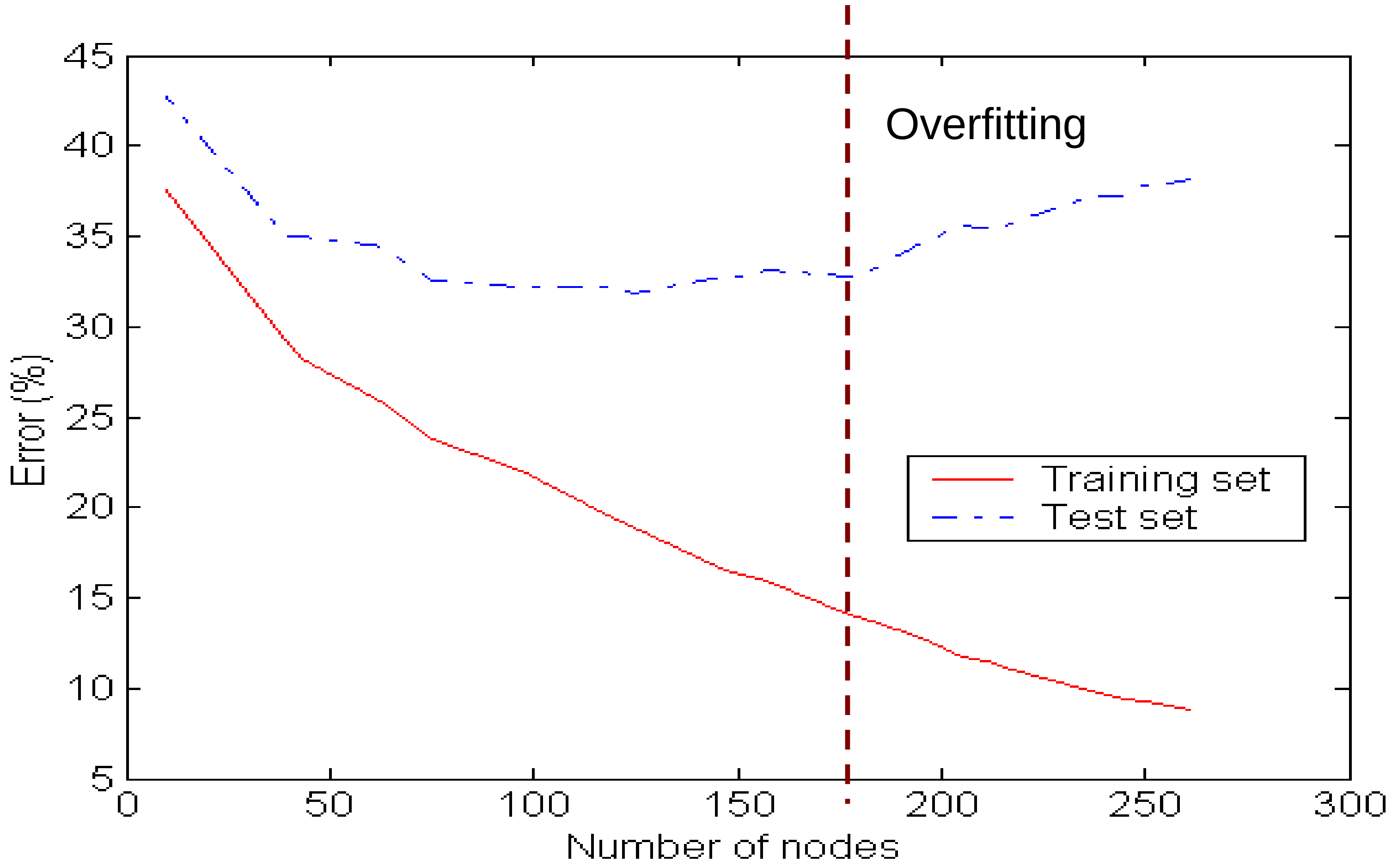
- Bias- how far the model prediction is from the correct value
- Variance – how much predictions vary from realization



Overfitting and Underfitting

- **Overfitting:**
  - Refers to a model that models the training data too well.
  - Given a model space H, a specific model $h \in H$ is said to overfit the training data if there exists some alternative model $h' \in H$, such that h has smaller error than h' over the training examples, but h' has smaller error than h over the entire distribution of instances i.e, your machine recognition is worse
  - The model is too complex, which creates a noise in the model.
- **Underfitting:**
  - The model is too simple, so that both training and test errors are large

Detecting Overfitting



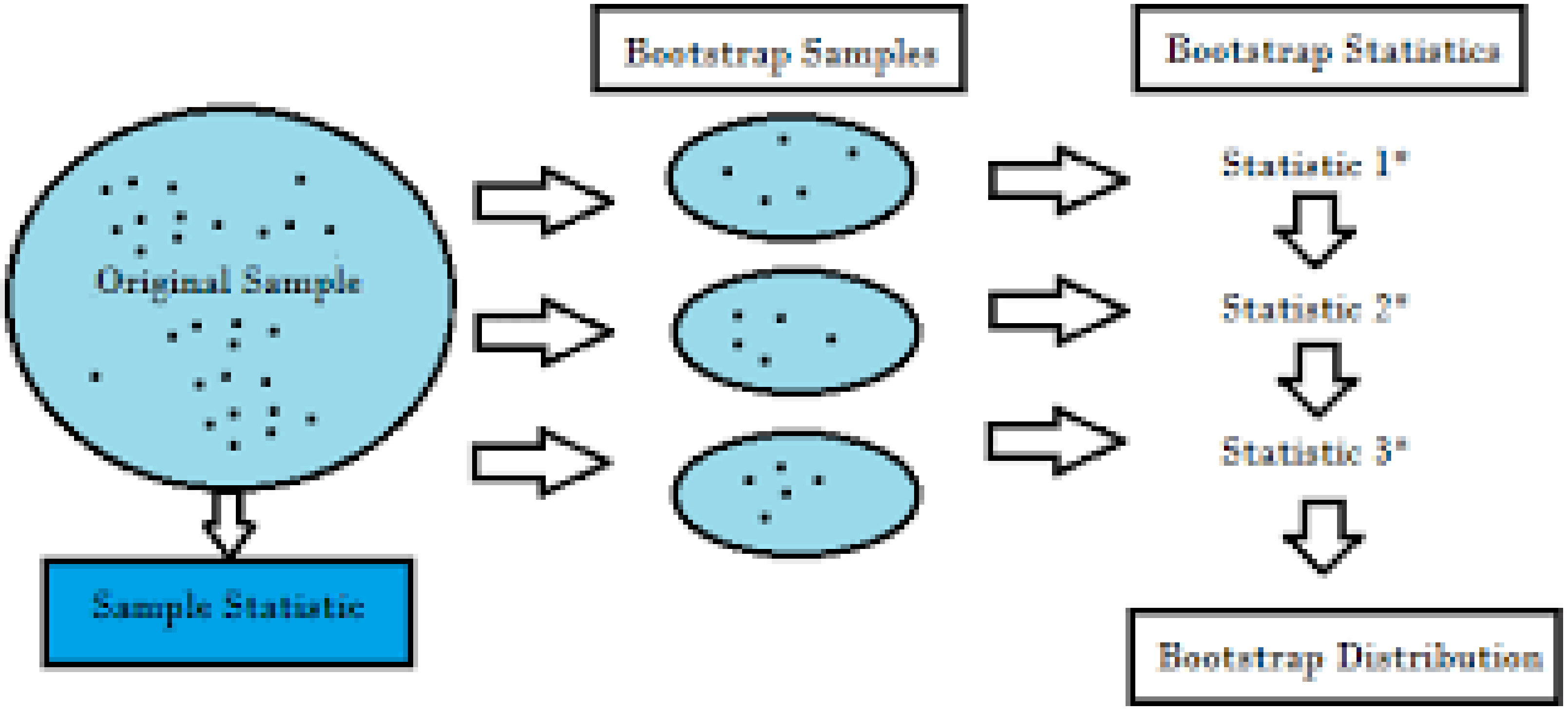
Detecting Overfitting

- If our model does much better on the training set than on the test set, then we’re likely **overfitting**.
- **How to prevent?**
  - **Cross validation**

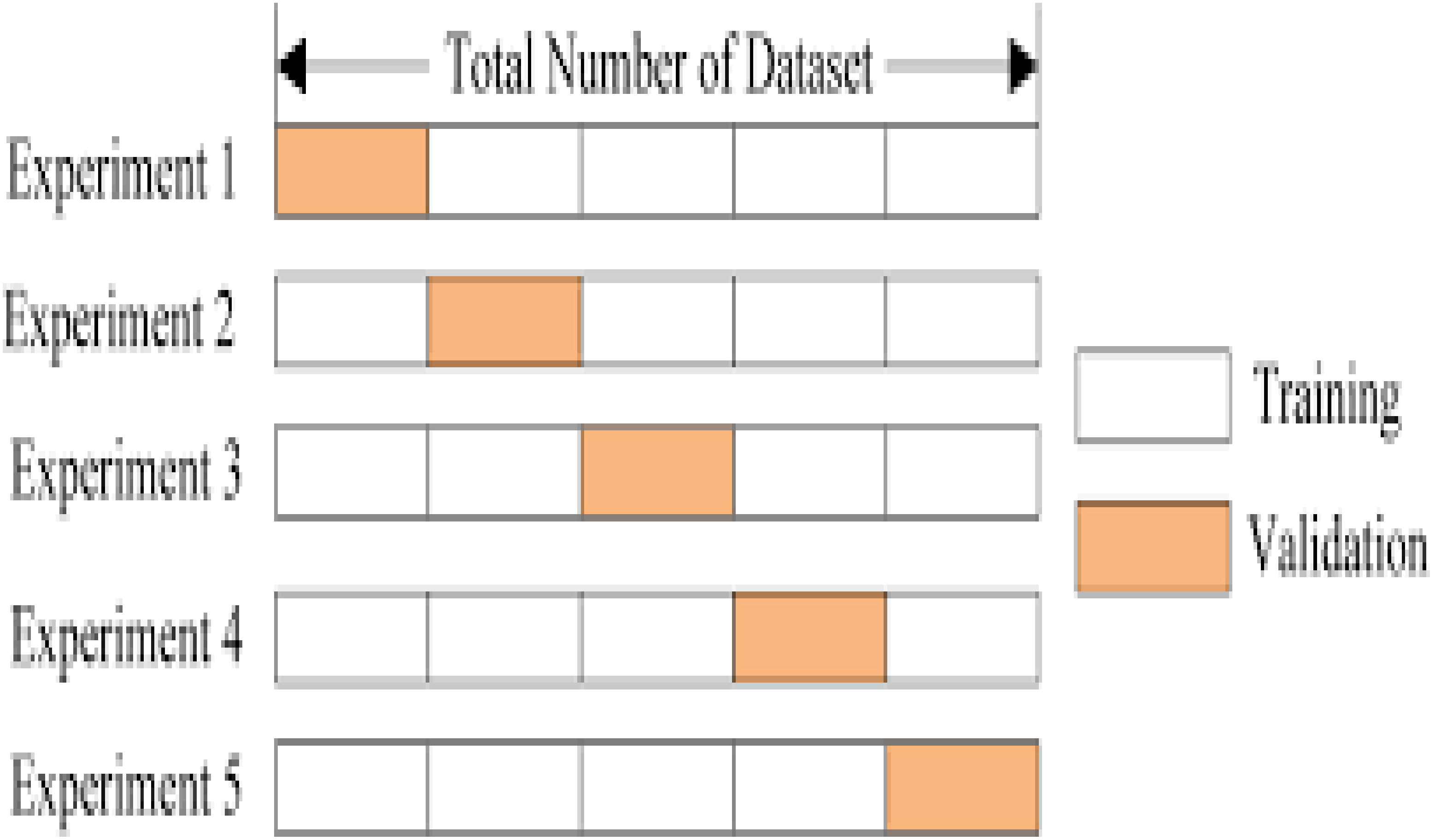
Resampling methods

- Resampling is a methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter.
- Each new subsample from the original data sample is used to estimate the population parameter.
- Two commonly used resampling methods that you may encounter are k-fold cross-validation and the bootstrap.
    - **Bootstrap**. Samples are drawn from the dataset with replacement (allowing the same sample to appear more than once in the sample), where those instances not drawn into the data sample may be used for the test set.
    - **k-fold Cross-Validation**. A dataset is partitioned into k groups, where each group is given the opportunity of being used as a held out test set leaving the remaining groups as the training set.

Bootstrapping



Cross Validation

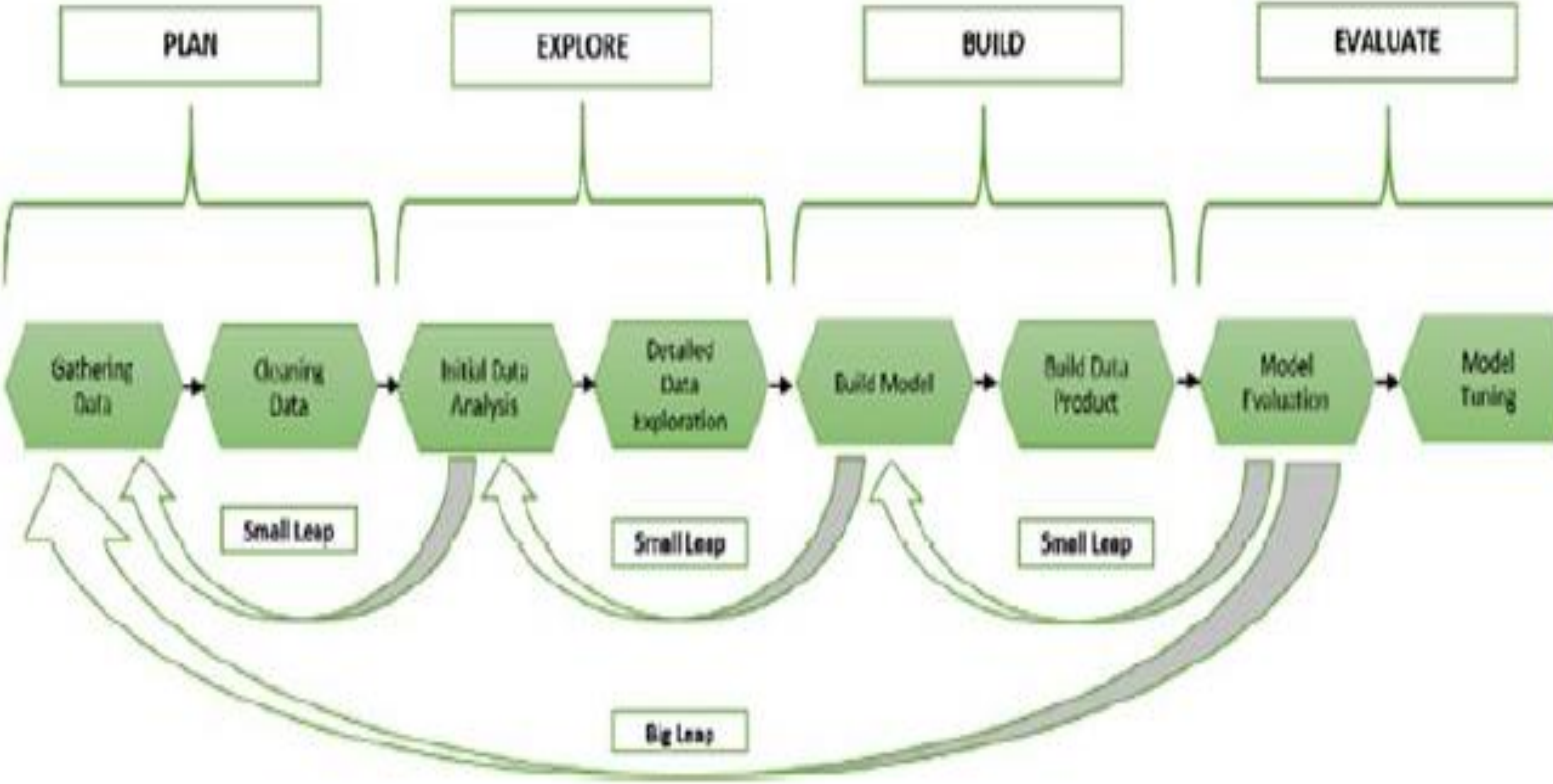


Summary

- To split into train and test – use sampling techniques
- To improve learning on the training set – use resampling techniques

Machine Learning Process Flow

- Plan
- Explore
- Build
- Evaluate


PLAN
EXPLORE
BUILD
EVALUATE
Gathering Data
Cleaning Data
Initial Data Analysis
Detailed Data Exploration
Build Model
Build Data Product
Model Evaluation
Model Tuning
Small Leap
Small Leap
Small Leap
Big Leap

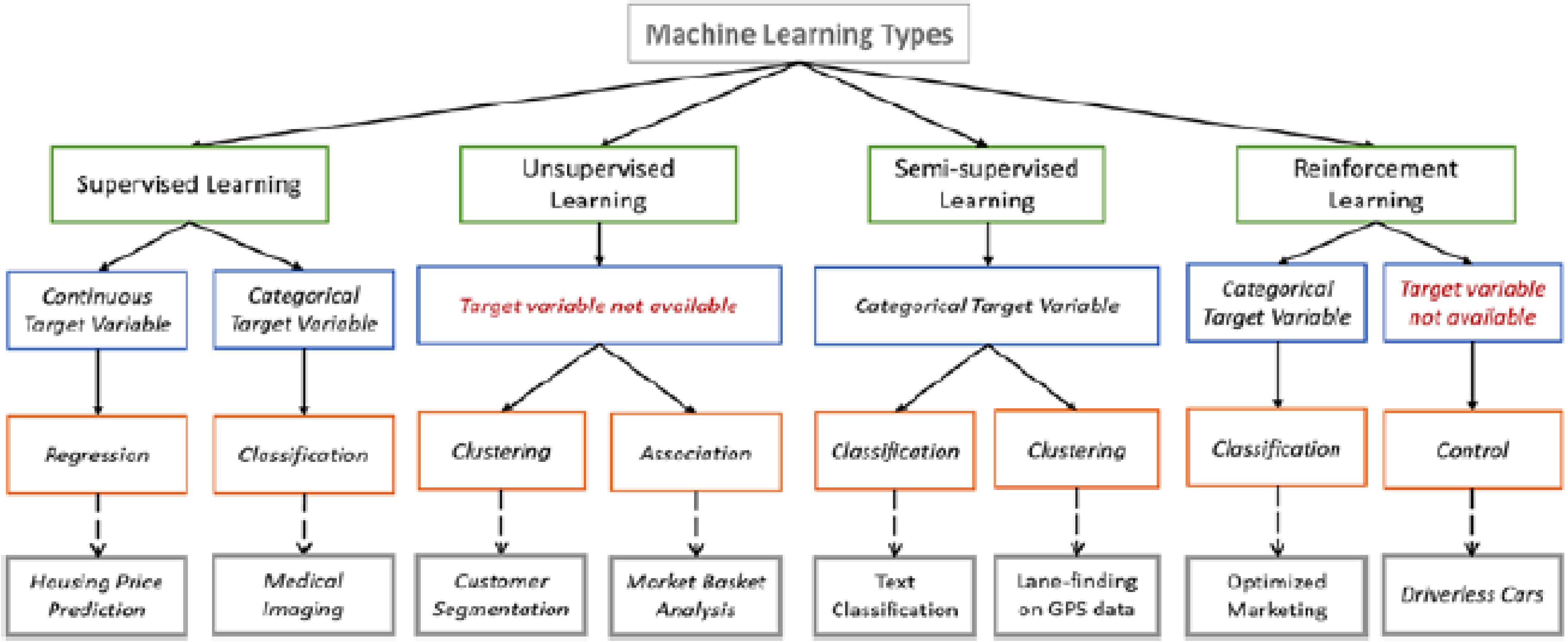
- Plan
  - Understanding the requirements, identifying every data source available
  - Gathering data is the core
  - Data cleaning - maintaining the integrity and veracity of the final outputs of the analysis and model building
- Explore
  - Simple statistical analysis - to identify possibilities, insights, scope, hidden patterns, challenges and errors
  - Creating hypothesis, Identifying sampling techniques

- Build
  - Is it necessary to build a model ?
  - ML algorithms are like a black-box. The output is difficult to interpret.
  - Does your model satisfies your question?
  - See the potentiality of building a data product
- Evaluate
  - You cannot build a powerful ML model in one iteration
  - Evaluate the models goodness and further fine-tune the model

Types of Machine Learning

Supervised Learning	Unsupervised Learning	Reinforcement Learning
• Classification • Regression **Mixed Methods**	• Clustering • Association Mining	• Decision Process • Reward System • Recommendation Systems
Use labelled training data to learn the mapping function from the input variables (X) to the output variable (Y)	Possess only the input variables (X) but no corresponding output variables. It uses unlabelled training data to model	Decide the best next action based on its current state, by learning behaviours that will maximize the reward

Detailed



Machine Learning Types
Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Reinforcement Learning
Continuous Target Variable
Categorical Target Variable
Target variable not available
Categorical Target Variable
Categorical Target Variable
Target variable not available
Regression
Classification
Clustering
Association
Classification
Clustering
Classification
Control
Housing Price Prediction
Medical Imaging
Customer Segmentation
Market Basket Analysis
Text Classification
Lane-finding on GPS data
Optimized Marketing
Driverless Cars

Supervised Learning

Find: function $\hat{f}$ which can be used to assign a **class** or **value** to **unseen observations.**

Given: a set of **labeled** observations

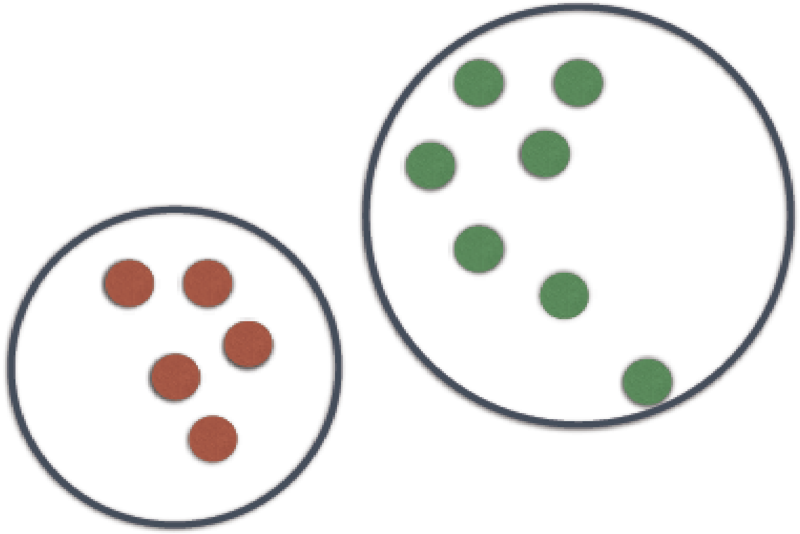
Supervised Learning

Unsupervised Learning

- **Labeling** can be tedious, often done by humans
- Some **techniques** don't require **labeled** data
- **Unsupervised Learning**
  - **Clustering**: find groups observation that are similar
  - Does **not** require **labeled observations**

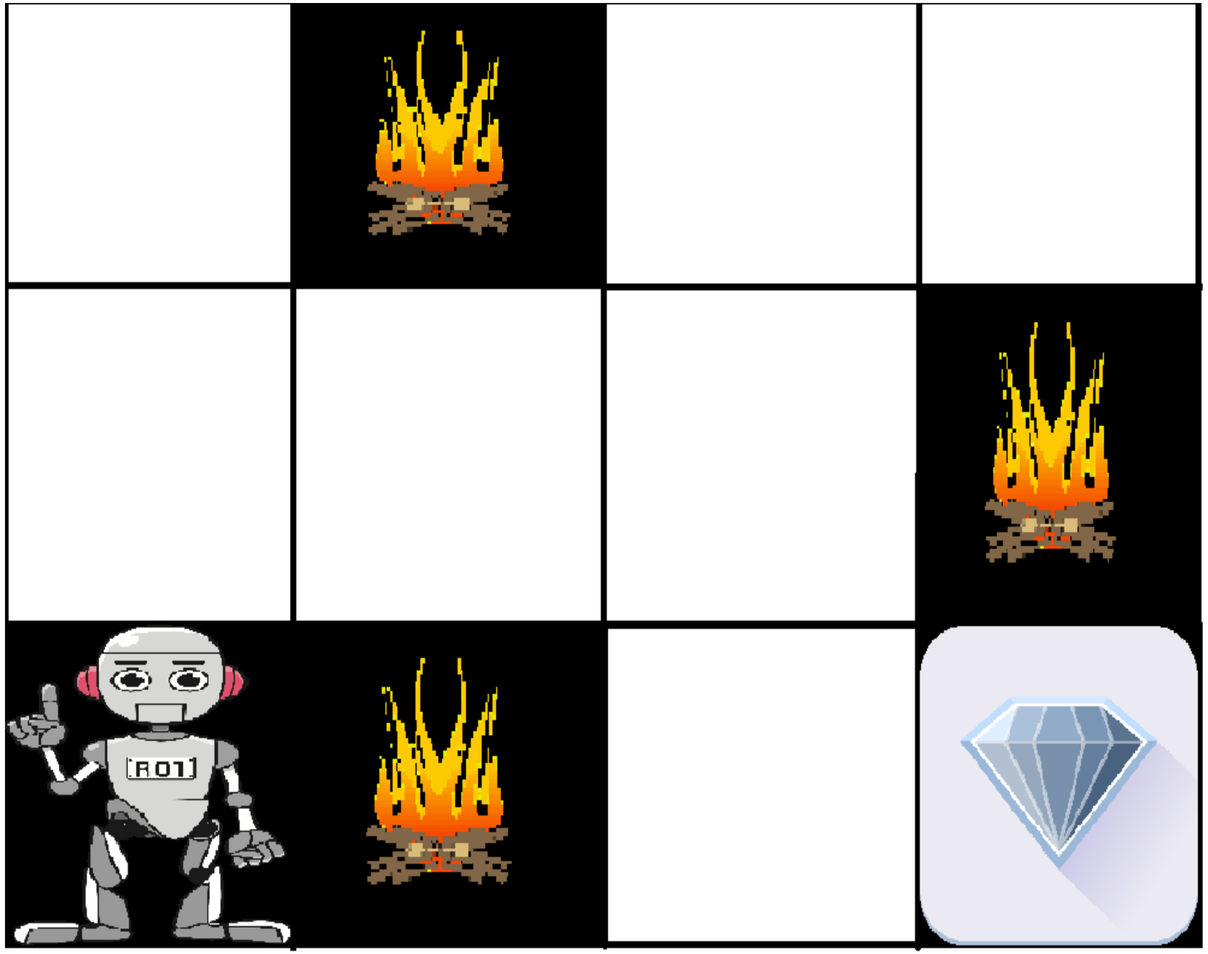
Semi-Supervised Learning

- A lot of **unlabeled observations**
- A few **labeled**
- Group similar observations using **clustering**
- Use **clustering** information and **classes** of **labeled observations** to **assign a class** to unlabelled observations
- More **labeled observations** for **supervised learning**



Reinforcement Learning

- Reinforcement learning is all about making decisions sequentially.
- Current state will be the input for the next state
- In Reinforcement learning decision is dependent, So we give labels to sequences of dependent decisions
- Example: Chess game



Groups of ML Algorithms

	Algorithms
Regression Analysis	Ordinary Least Squares Regression (OLSR)
	Linear Regression
	Logistic Regression
	Stepwise Regression
	Polynomial Regression
	Locally Estimated Scatterplot Smoothing (LOESS)

	Algorithms
Distance-based Algorithms	k-Nearest Neighbor (kNN)
	Learning Vector Quantization (LVQ)
	Self-Organizing Map (SOM)

-contd

	Algorithms
Regularization Algorithms	Ridge Regression
	Least Absolute Shrinkage and Selection Operator (LASSO)
	Elastic Net
	Least-Angle Regression (LARS)

	Algorithms
Decision Tree Algorithms	Classification and Regression Tree (CART)
	Iterative Dichotomiser 3 (ID3)
	C4.5 and C5.0 (different versions of a powerful approach)
	Chi-squared Automatic Interaction Detection (CHAID)
	Random Forest
	Conditional Decision Trees

-contd

	Algorithms
Bayesian Algorithms	Naive Bayes
	Gaussian Naive Bayes
	Multinomial Naive Bayes
	Bayesian Belief Network (BBN)
	Bayesian Network (BN)

	Algorithms
Clustering Algorithms	k-Means
	k-Medians
	Partitioning Around Medoids (PAM)
	Hierarchical Clustering

-contd

	Algorithms
Association Rule Mining Algorithms	Apriori algorithm
	Eclat algorithm
	FP-growth algorithm
	Context Based Rule Mining

	Algorithms
Artificial Neural Network Algorithms	Perceptron
	Back-Propagation
	Hopfield Network
	Radial Basis Function Network (RBFN)

-contd

	Algorithms
Deep Learning Algorithms	Deep Boltzmann Machine (DBM)
	Deep Belief Networks (DBN)
	Convolutional Neural Network (CNN)
	Stacked Auto-Encoders

	Algorithms
Ensemble Algorithms	Boosting
	Bagging
	AdaBoost
	Stacked Generalization (blending)
	Gradient Boosting Machines (GBM)

-contd

	Algorithms
Text Mining	Automatic summarization
	Named entity recognition (NER)
	Optical character recognition (OCR)
	Part-of-speech tagging
	Sentiment analysis
	Speech recognition
	Topic Modeling

	Algorithms
Dimensionality Reduction Algorithms	Principal Component Analysis (PCA)
	Principal Component Regression (PCR)
	Partial Least Squares Regression (PLSR)
	Multidimensional Scaling (MDS)
	Linear Discriminant Analysis (LDA)
	Mixture Discriminant Analysis (MDA)
	Quadratic Discriminant Analysis (QDA)

What is learnt previous??

- Classification – knn, naïve bayes
- Regression – simple linear, multiple linear, logistic
- Clustering – K-means, k-mediods, hierarchical
- Association – apriori

Classification

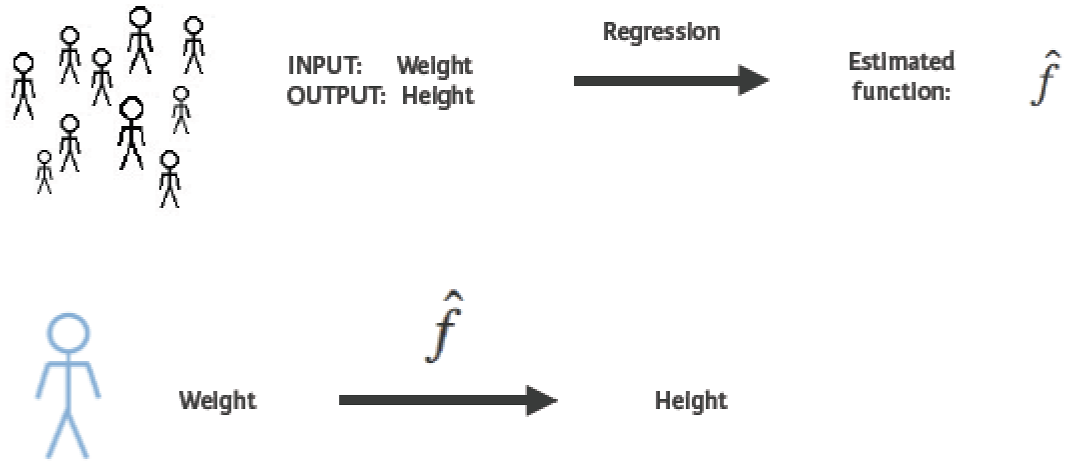
Goal: predict category of new observation

Earlier Observations $\xrightarrow{\text{Estimate}}$ CLASSIFIER

Unseen Data $\xrightarrow{\text{CLASSIFIER}}$ Class

- Important
  - Qualitative Output
  - Predefined classes
- Applications
  - Medical Diagnosis – Sick or not sick
  - Face recognition – human or animal
  - Character recognition – alphabets or numbers
  - Speech recognition

Regression



- Important
  - Quantitative output
  - Previous input-output observations
- Regression Applications
  - Payments -> credit scores
  - Time -> subscriptions
  - Grades -> landing a job
  - Navigating a car – angle of steering wheel

Fitting a **linear** function

$$\text{Height} \approx \beta_0 + \beta_1 \times \text{Weight}$$

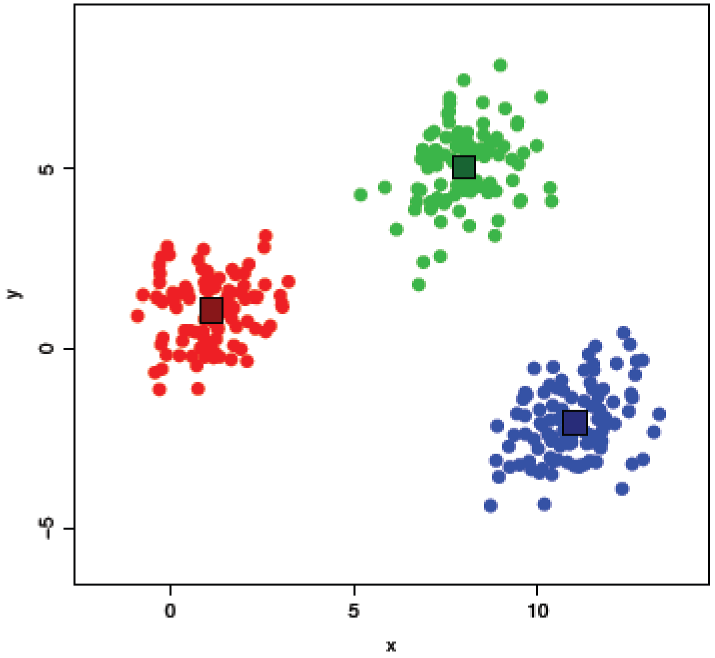
- Predictor: Weight
- Response: Height
- Coefficients: β_0, β_1

Estimate on previous input-output

```
> lm(response ~ predictor)
```

Clustering

- Grouping objects in clusters
  - Similar within cluster
  - Dissimilar between clusters
- Example
  - Grouping Similar animal photos
    - No Labels
    - No right or wrong
    - Plenty possible clusterings
  - Customer segmentation in CRM
  - Image compression: Color quantization



Association

- Finding What goes with what
- Find rules that will predict the occurrence of an item based on the occurrences of other items
- Applications
  - Market basket Analysis
  - Catalog design
  - Cross-marketing

Machine Learning Tasks

- Classification
- Regression
- Clustering

quite similar

What is to be learnt here???? - Mixed Methods

- Tree-based methods
- Random Forests
- Neural Networks
- Support Vector Machines

Tree based models

- Tree is a structure that has nodes and edges
- In a decision tree, at each node we might have a value against which we split in order to gain some insight from the data
- Trees are obtained starting at each field and going down on other fields
- Each time the entropy is calculated and The tree with maximum info gain is selected as model
- The root node will have the major influencing and it goes deep with less influencing
- This arrived decision tree can be used for further prediction

Random forest

- Suppose you have arrived at multiple decision trees and each of the tree is equally important to make decisions, then you build an ensemble classifier, or a forest
- The model is not a single decision tree but a forest
- The more the number of trees, the minimal is an error
- But you can also control the growth of the trees.

Tree based models – Pros and Cons

- Tree are easy to understand and explain
- Easy to use in real world business
- Easy to handle qualitative variables
- But, tree models are non-robust i.e, it is sensitive to the data. If there is a slight change in the data, the trees change completely.
- When can be used? Predict and obtain inference

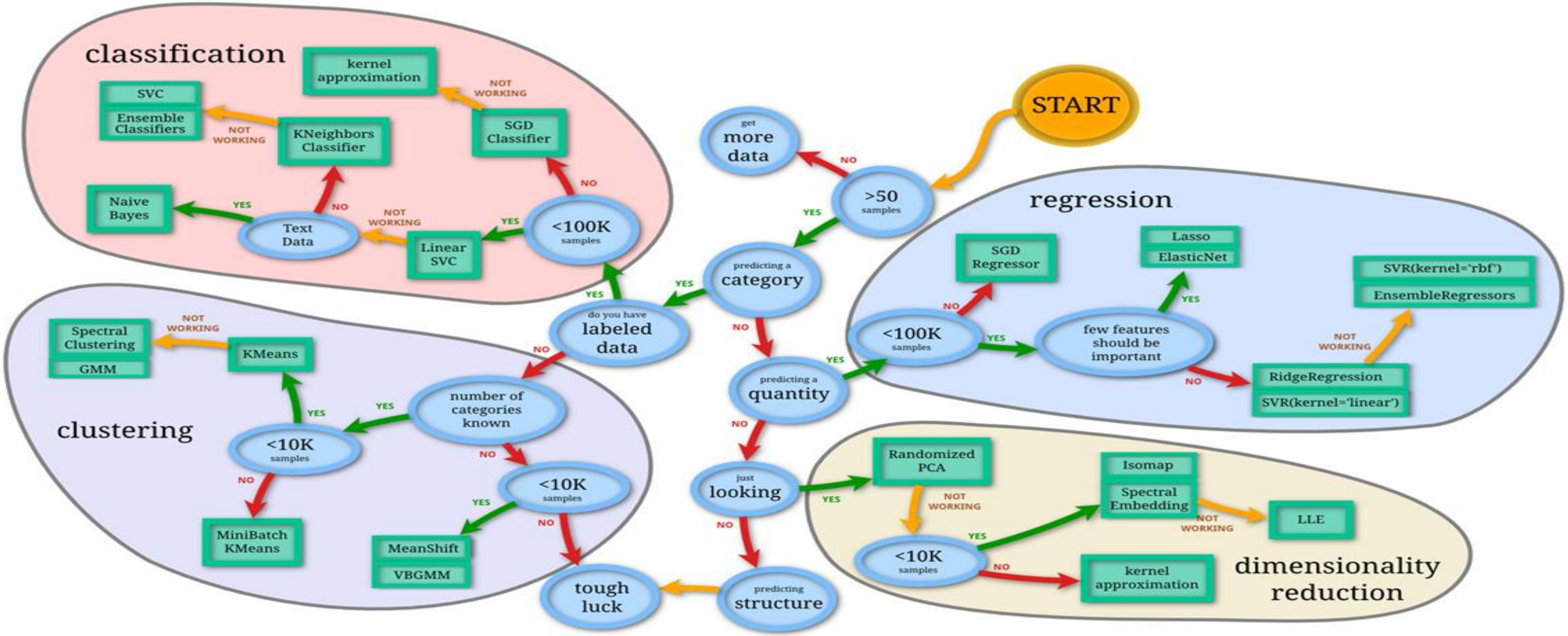
Neural Networks

- For a given list of inputs, a neural network performs a number of processing steps before returning an output
- The complexity of the neural network comes in how many number of processing steps and how complex each particular step might be
- It has the input layer, the hidden or the compute layer and the output layer
- The hidden layer may be simple having one node or complex having multiple nodes
- A single function f(x) is arrived as an output of the model

Support Vector Machines

- SVM is similar to logistic regression
- The objective is to find a plane that can separte the data into different classes
- Suppose we have n features and m observations.
  - If n is greater than m use logistic regressor
  - If m is greater than n then use svm.
- But in either case neural network can be used.

Machine Learning Map



ML Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive: http://kdd.ics.uci.edu/summary.data.application.html
- Statlib: http://lib.stat.cmu.edu/
- Delve: http://www.cs.utoronto.ca/~delve/
- Open Government data: data.gov.in
- Github

Reading resources in Nutshell

- Tree based models - https://www.youtube.com/watch?v=pEjLd5RMjAA
- Svm - https://www.youtube.com/watch?v=Y6RRHw9uN9o
- NN - https://www.youtube.com/watch?v=P2HPcj8lRJE

https://www.youtube.com/watch?v=GQVLl0RqpSs

Machine Learning in Python

- Distribution – Anaconda
- Notebook – Jupyter
- Package – scikit learn [pip/conda install scikit-learn if not installed]
- General format of importing the model
  - from sklearn.family import Model
  - Eg: from sklearn.linear_model import LinearRegression

- Model building
  - Model.fit(X,y) – supervised learning
  - Model.fit(x) – unsupervised learning

  - Model.predict(X_new)
  - Model.predict_proba() – for some models
  - Model.score() – larger score indicating the better fit vlaues between 0 to 1

Automated ML

- Managing the ML workflows
- Many moving parts in ML model that is to be tied together
- Process of tying together is pipeline
- Pipeline has different stages
- Each stage can feed an input to the next stage
- Pipelines are help to do auto ML
- Pipelines is a way to create error free models
- The purpose of the pipeline is to assemble several steps that can be cross-validated together while setting different parameters.

ML Pipeline in Python

- from sklearn.pipeline import Pipeline
- Pipeline(list_of_*steps*)
- Eg: pipe = Pipeline([

```
                    ('minmax', MinMaxScaler()),
                     ('knn', KNeighborsClassifier())
                     ])
```

End of Unit 1