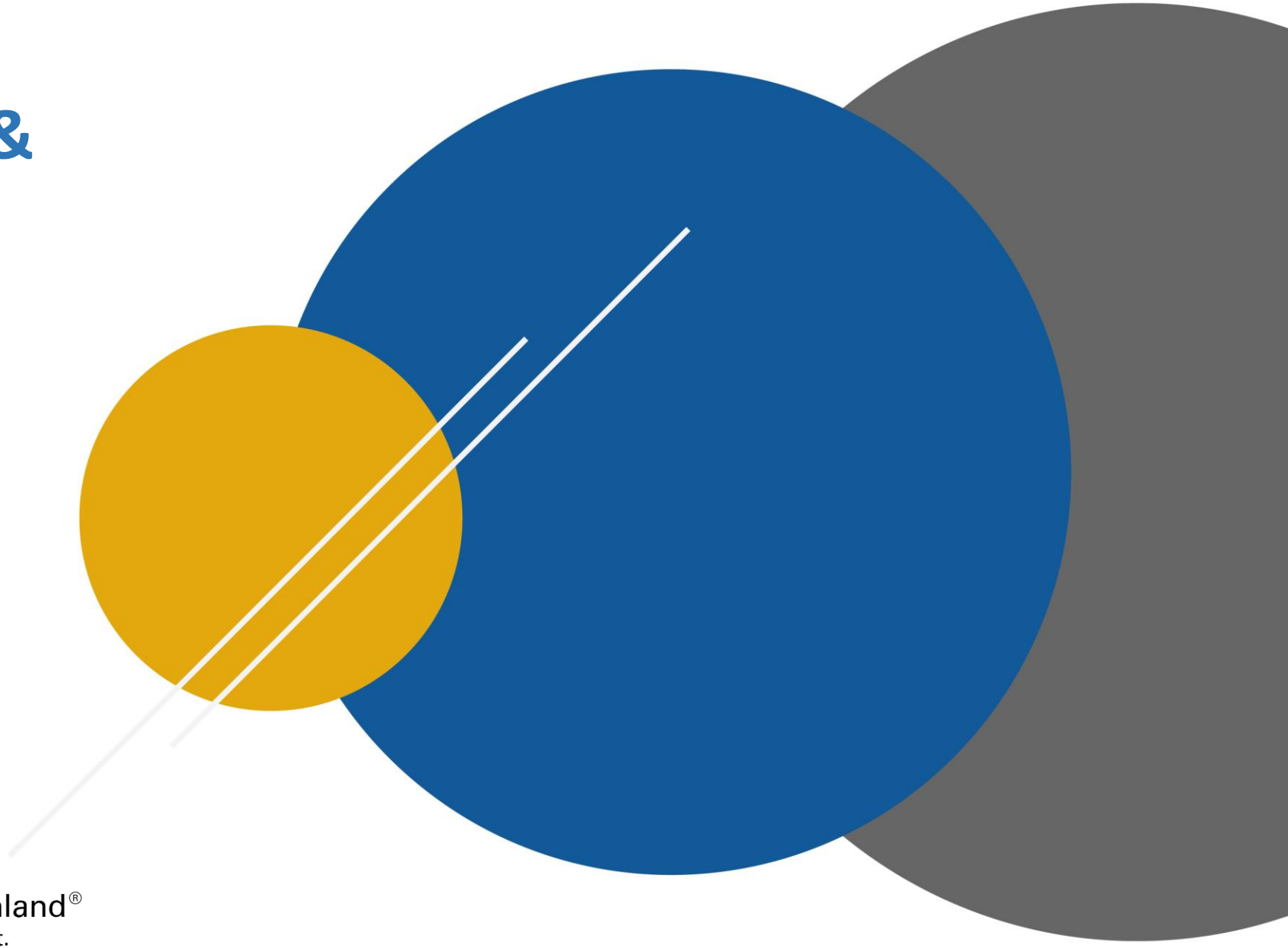
Machine Learning &
Predictive Analytics



Telkom University

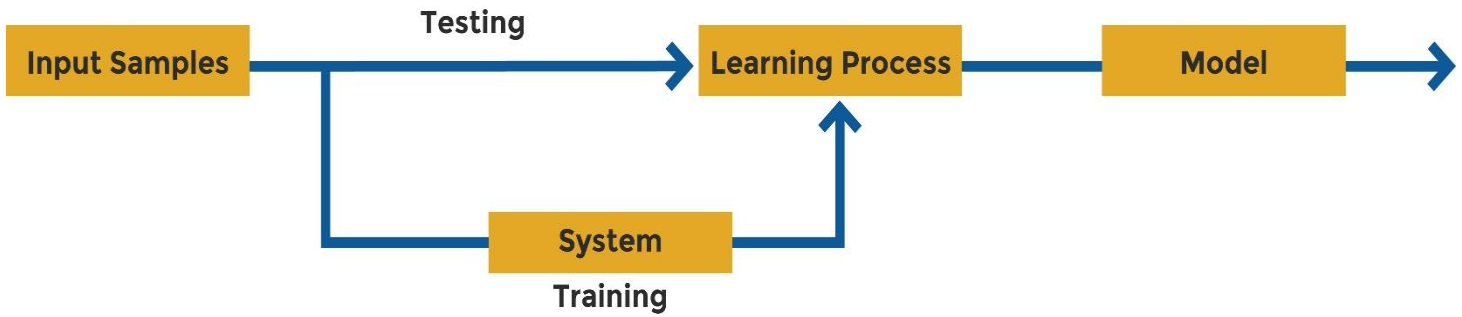
TÜVRheinland®
Precisely Right.

Agenda

- **Machine Learning**
- Learning Illustration
- Supervised vs Unsupervised Learning
- Reinforcement Learning
- Predictive Analytics
- Model Evaluation

Machine Learning

- A branch of artificial intelligence, concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data.
- As intelligence requires knowledge, it is necessary for the computers to acquire knowledge.
- Learning System Model



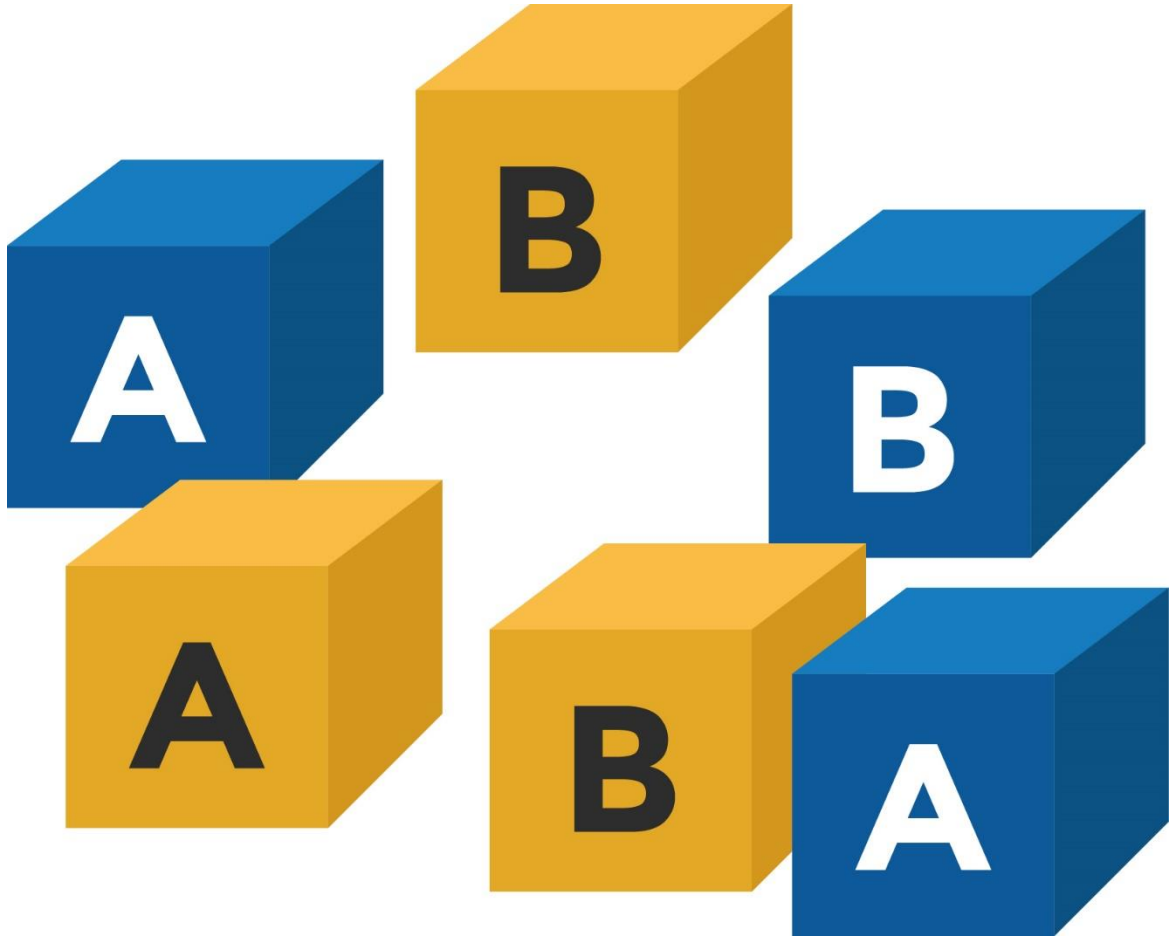
Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Agenda

- Machine Learning
- **Learning Illustration**
- Supervised vs Unsupervised Learning
- Reinforcement Learning
- Predictive Analytics
- Model Evaluation

Learning
Illustration



Two Possible Solutions

Agenda

- Machine Learning
- Learning Illustration
- **Supervised vs Unsupervised Learning**
- Reinforcement Learning
- Predictive Analytics
- Model Evaluation

Supervised Vs Unsupervised

- Task performed
  - Classification
  - Pattern Recognition
- NN model :
  - Perceptron
  - Feed-forward NN

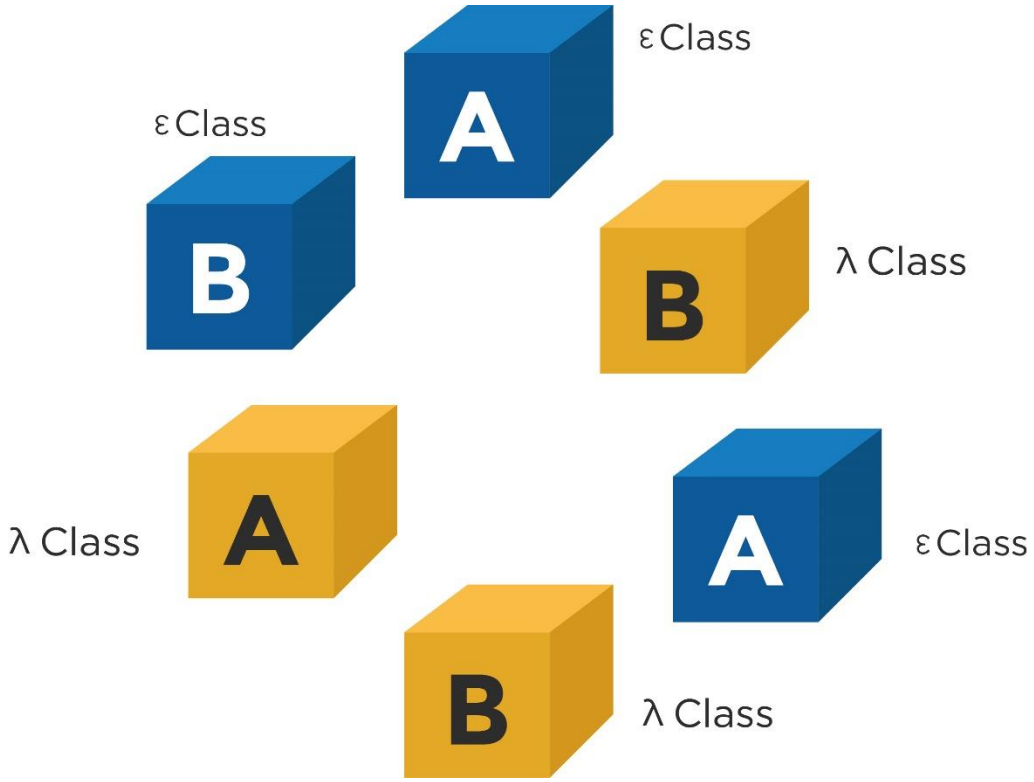
"What is the class of this data point?"

- Task performed
  - Clustering
- NN model :
  - Self Organizing Maps

"What groupings exist in this data?"

"How is each data point related to the data set as a whole?"

Supervised Learning



- It is based on a labeled training set.
- The class of each piece of data in training set is known.
- Class labels are pre-determined and provided in the training phase.

Supervised Learning

- Prediction methods are commonly referred to as supervised learning. Supervised methods are thought to attempt the discovery of the relationships between input attributes and a target attribute.
- A training set is given and the objective is to form a description that can be used to predict unseen examples.

Supervised Learning

Problem

- Classification
  - The domain of the target attribute is finite and categorical.
  - A classifier must assign a class to a unseen example.
- Regression
  - The target attribute is formed by infinite values.
  - To fit a model to learn the output target attribute as a function of input attributes.
- Time Series Analysis
  - Making predictions in time.

Unsupervised Learning

- Input : set of patterns P, from n-dimensional space S, but little/no information about their classification, evaluation, interesting features, etc.

  It must learn these by itself! :)

- Tasks:
  - Clustering - Group patterns based on similarity
  - Vector Quantization - Fully divide up S into a small set of regions (defined by codebook vectors) that also helps cluster P.
  - Feature Extraction - Reduce dimensionality of S by removing unimportant features (i.e. those that do not help in clustering P)

Unsupervised Learning

- There is no supervisor and only input data is available.
- The aim is now to find regularities, irregularities, relationships, similarities and associations in the input.

Unsupervised Learning

Problems:

- Clustering
- Association Rules
- Pattern Mining
  - It is adopted as amore general term than frequent pattern mining or association mining.
- Outlier Detection
  - It is the process of finding data examples with behaviours that are very different from the expectation (outliers or anomalies).

Agenda

- Machine Learning
- Learning Illustration
- Supervised vs Unsupervised Learning
- **Reinforcement Learning**
- Predictive Analytics
- Model Evaluation

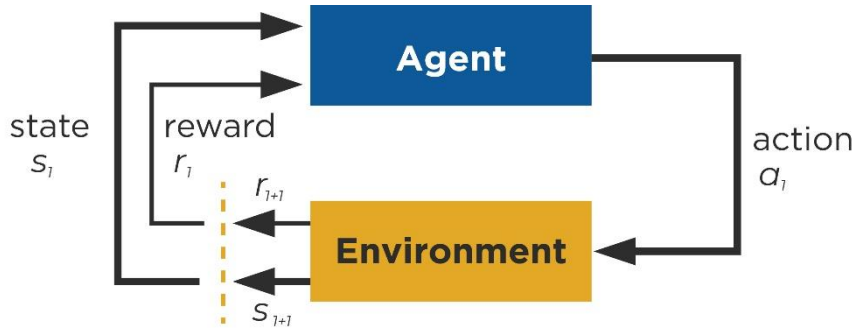
Background

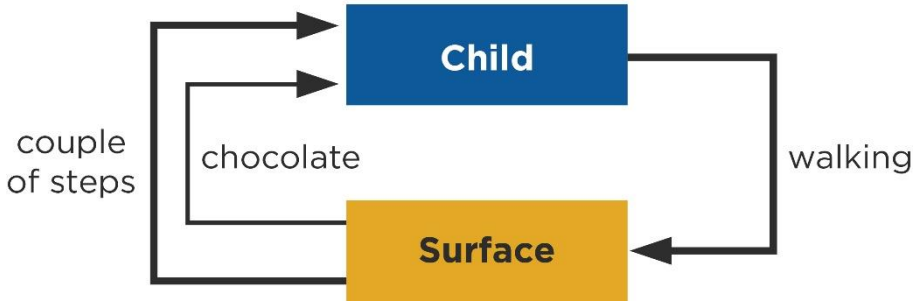
- How to learn a new Skill
- Learning and intelligence
- Interaction with environment
- Goal-oriented learning
- Agent – Environment interactions
- Activities
  - What to do
  - How to map situations to actions
  - Process positive and negative rewards

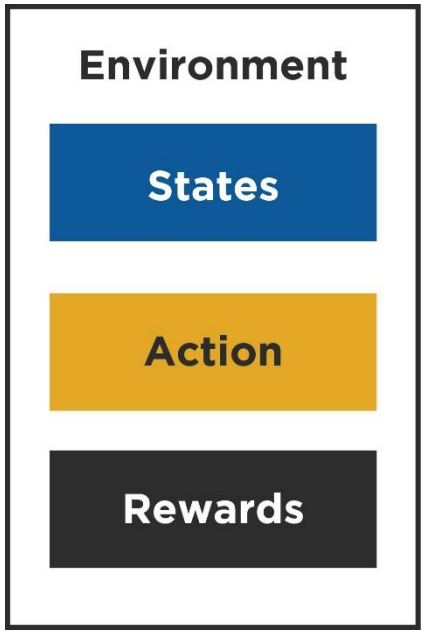
Reinforcement Learning

- The Analogy
  - A child learns to walk
  - The child is an agent trying to manipulate the environment
  - The child is taking actions (state 1, state 2, state 3, and so on)
  - Positive rewards when able to walk
  - Negative rewards when not able to walk

Reinforcement Learning (RL) Model







Popular Area
for RL

- Self Driving Cars
- Gaming
- Robotics
- Recommendation Systems
- Advertising and Marketing

RL
Methods

- Q-Learning
- MDPs and Bellman Equations
- Dynamic Programming: Model-Based RL, Policy Iteration and Value Iteration
- Deep Q Learning
- Policy Gradient Methods
- SARSA

Example:
Taxi Problem

Software

- Python, Open AI Gym Platform

Problem:

- Correct drop-off location
- Find minimum time possible to drop off
- Obey rules

Algorithm

- Q-Learning
- Neural Network

Q-Learning

$$Q(state, action) \leftarrow (1-\alpha)Q(state, action) + \alpha\left(reward + \gamma \max_a Q(next\ state, all\ actions)\right)$$

Steps:

1. Initialize the Q-table by all zeros.
2. Start exploring actions: For each state, select any one among all possible actions for the current state (S).
3. Travel to the next state (S') as a result of that action (a).
4. For all possible actions from the state (S') select the one with the highest Q-value.
5. Update Q-table values using the equation.
6. Set the next state as the current state.
7. If goal state is reached, then end and repeat the process.

Using Source Code

Requirement

- Anaconda installation
- Jupyter Notebook
- OpenAI Gym package installation

How to use

1. Run jupyter notebook via Anaconda Prompt
2. Open Solving taxi problem using Gym.ipynb and run

Other Learning Paradigms

- Multi-label Classification
  - Each instance is associated not with a class, but instead with a subset of them.
- Semi-supervised Learning
  - It is concerned with the design of models in the presence of both labeled and unlabeled data.
  - Semi-supervised classification and Semi-supervised clustering.
  - Relationship with Active Learning.
- Subgroup Discovery
  - It is formed as the result of the hybridization between classification and association mining.
  - They aim to extract interesting rules with respect to a target attribute.

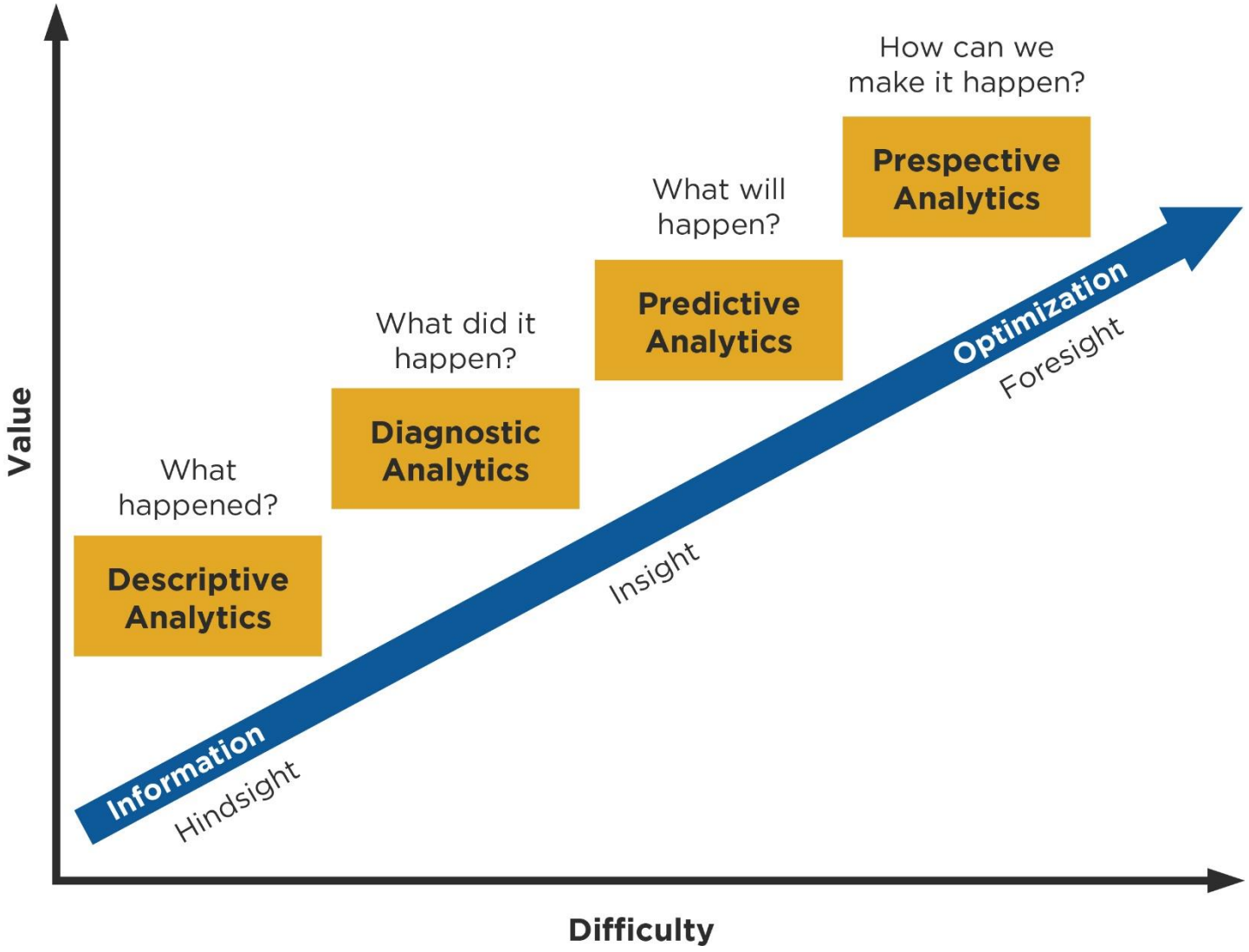
Other Learning
Paradigms

- Transfer Learning
  - Aims to extract the knowledge from one or more source tasks and apply the knowledge to a target task.
  - The so-called data shift problem is closely related.
- Imbalanced Learning
  - A classification problem where the data has exceptional distribution on the target attribute.
  - The number of examples representing the class of interest is much lower than that of the other classes.
- Multi-instance Learning
  - imposed restrictions on models in which each example consists of a bag of instances instead of an unique instance.

Other Learning Paradigms

- Data Stream Learning
  - When all data is not available at a specific moment, it is necessary to develop learning algorithms that treat the input as a continuous data stream.
  - Each instance can be inspected only once and must then be discarded to make room for subsequent instances.

Analytics
Maturity



Agenda

- Machine Learning
- Learning Illustration
- Supervised vs Unsupervised Learning
- Reinforcement Learning
- **Predictive Analytics**
- Model Evaluation

What is
Predictive Analytics

Predictive analytics is the practice of extracting insights from the existing data set with the help data mining, statistical modelling and machine. learning techniques and using it to predict unobserved/unknown events.

- Identifying cause-effect relationships across the variables from the historical data.
- Discovering hidden insights and patterns with the help of data mining techniques.
- Apply observed patterns to unknowns in the Past, Present or Future.

What is
Predictive Analytics

- Area of statistics around capturing relationships between explanatory variables and predicted variables from past occurences and using it for prediction.
- Data mining, automatically discovering nterseting patterns in data.
- Can apply to unknowns in the Past, Present or Future.
- Accuracy and usability vary upon level of analysis and quality of assumptions.
- Improve decision making by enabling learning from past experience.
- Waaaaaaaaay too much data and too many variables to manually analyze or use traditional statistics techniques.
- Traditional analytics are statistical methods fail due to complex non-linear and multi variable combinations.
- Strategic competitive advantages.

Predictive Analytics
Usage (Example)

Function	Description	Example Companies
Supply Chain	Simulate and optimize supply chain flows; reduce inventory and stock-outs.	Dell, Wal-Mart, Amazon, UPS
Customer Selection	Identify customers with greatest profit potential; retain their loyalty.	Harrah's, Capital One, Barclays
Pricing	Identify the price that will maximize yield, or profit.	Progressive, Marriott
Human Capital	Select the best employees for particular tasks or jobs, at particular compensation levels.	New England Patriots, Oakland A's, Boston Red Sox
Product and Service Quality	Detect Quality problems early and minimize them.	Honda, Intel
Financial Performance	Better understand the drivers of financial performance and the effects of nonfinancial factors	MCI, Verizon
Research and Development	Improve quality, efficiacy, and, where applicable, safety of products and services.	Novartis, Amazon, Yahoo
Fraud and Crime	Identity fraud, criminal activity, tax evasion and stolen credit cards	Bank of America, IRS, FBI

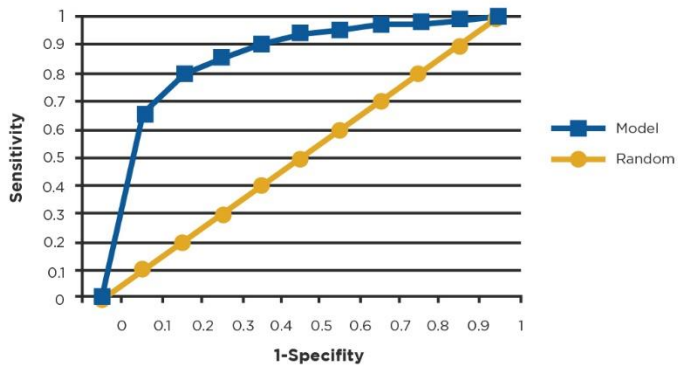
Common Predictive
Analytics Methods

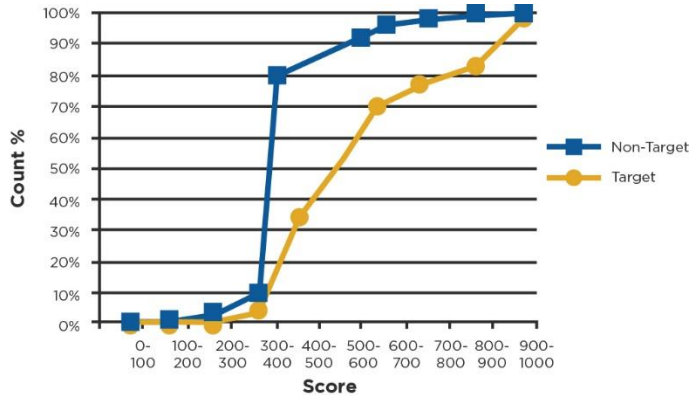
- Regression
  - Predicting output variable using its cause-effect relationship with input variables.
  - OLS Regression, GLM, Random forests, ANN etc.
- Classification
  - Predicting the item class.
  - Decision Tree, Logistic Regression, ANN, SVM, Naïve Bayes classifier etc.
- Time Series Forecasting
  - Predicting future time events given past history.
  - AR, MA, ARIMA, Triple Exponential Smoothing, Holt-Winters etc.

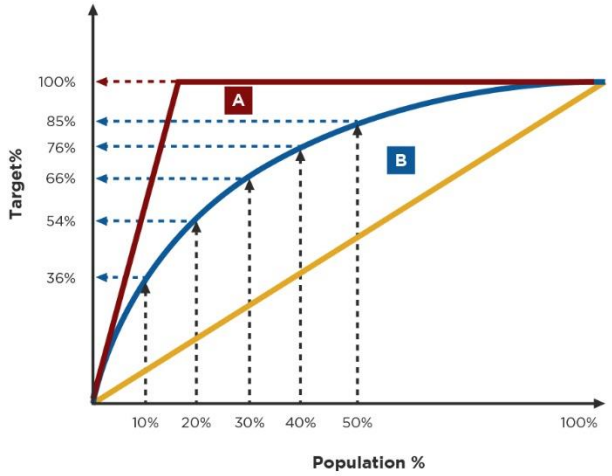
Common Predictive Analytics Methods (contd)

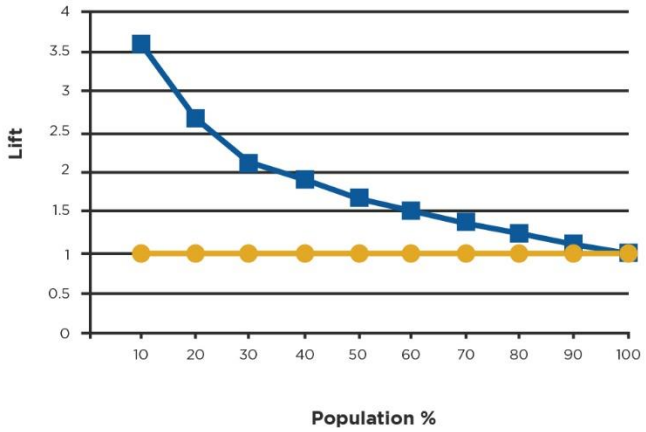
- Association rule mining
  - Mining items occurring together.
  - Apriori Algorithm.
- Clustering
  - Finding natural groups or clusters in the data.
  - K-means, Hierarchical, Spectral, Density based EM algorithm Clustering etc.
- Text mining
  - Model and structure the information content of textual sources,
  - Sentiment Analysis, NLP

Evaluating Predictive Models









- Need to check predictive model's out of sample performance.
- Model Assessment:

  Hit Rate, Gini Coefficient, K-S Chart, Confusion Matrix, ROC Curve, Lift Chart, Gain Chart etc.

Agenda

- Machine Learning
- Learning Illustration
- Supervised vs Unsupervised Learning
- Reinforcement Learning
- Predictive Analytics
- **Model Evaluation**

Model
Evaluation

- Metrics for Performance Evaluation
  - How to evaluate the performance of a model (accuracy, etc.)?

- Methods for Performance Evaluation
  - How to obtain reliable estimates (precision, etc.)?

Metrics for
Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	a	b
	Class = No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metrics for
Performance Evaluation

- Most widely-used metric:

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	a (TP)	b (TN)
	Class = No	c (FP)	d (TP)

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Cost-Sensitive Measures

- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

Model
Evaluation

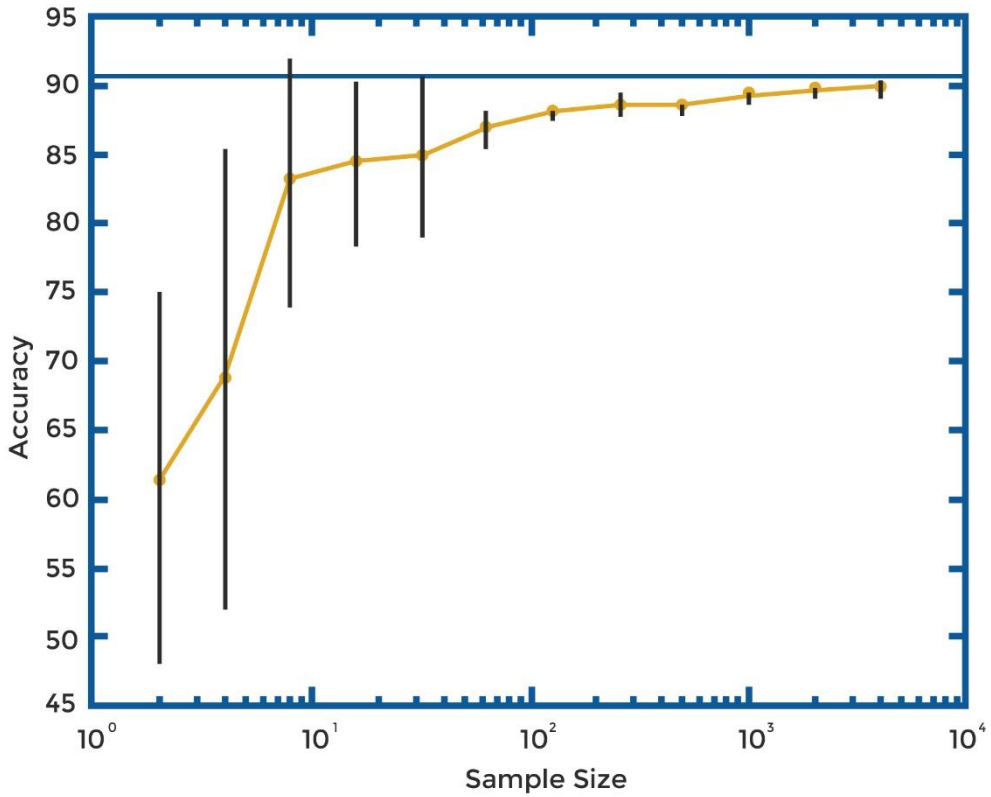
- Metrics for Performance Evaluation
  - How to evaluate the performance of a model?
- Methods for Performance Evaluation
  - How to obtain reliable estimates?
- Methods for Model Comparison
  - How to compare the relative performance among competing models?

Methods for Performance Evaluation

- How to obtain a reliable estimate of performance?

- Performance of a model may depend on other factors besides the learning algorithm:
  - Class distribution
  - Cost of misclassification
  - Size of training and test sets

Learning
Curve



- Learning curve shows how accuracy changes with varying sample size
- Requires a sampling schedule for creating learning curve:
  - Arithmetic sampling (Langley, et al)
  - Geometric sampling (Provost et al)
- Effect of small sample size:
  - Bias in the estimate
  - Variance of estimate

Methods of Estimation

- Holdout
  - Reserve 2/3 for training and 1/3 for testing
- Random Subsampling
  - Repeated holdout
- Cross Validation
  - Partition data into k disjoint subsets
  - k-fold: train on k-1 partitions, test on the remaining one
  - Leave-one-out: k=n
- Stratified Sampling
  - oversampling vs undersampling
- Bootstrap
  - Sampling with replacement