

Lecture 1

Instructor : Pramod Gupta aka PG

What to expect

- About the course
 - About me
 - How this course has been put together
 - Rubric
- This course is not a regular course with a regular textbook. This is a course:
 - To familiarize you with the developing field of predictive analytics and machine learning
 - To equip you with useful knowledge and methods
 - To inspire your research interests
 - To connect your knowledge to real world problems
- This course is a fun class to enjoy

About me

- HELLO
 - My name is Pramod Gupta
- PhD in Electrical and Computer Engineering, from McMaster University, Canada
- Experience: Academics, NASA, EMC, GE, VISA, Startups etc.
- Teaching Data Science and Machine Learning Related Courses including R and Python
- Independent data consultant

Philosophy

- Sharing what I've Learned
- Goal: " Learning to self-Learn
- It can sometimes take time (hours, days) to figure out how to do something
- I would like this class to be as open and interactive
- Don't be shy or hesitant
- Knowledge flow is bi-direction.
- Ideas are important than age. Just because someone is junior does not mean they don't deserve respect and cooperation

We're all exploring and figuring out. Just share what you've learned.

Course Logistics

Who is this course for?

- Anyone who is interested in:
 - Helping companies make decisions aided by data
 - Refreshing some theory learned in school, but with a practical focus
 - Getting up to speed with new Open-Source tools and libraries
 - Curious about the new technology
- What is missing:
 - Lack of real-world analysis experiment, outside of work domain
 - Balance between theory and practice
 - Toolkit with which to explore

Goals and objectives

- The overall goal of this class is to introduce you to the discipline of predictive analytics, a science of understanding and analyzing data and machine learning algorithms for various tasks such *as prediction, classification, regression, clustering* etc. This class is designed to provide you with the tools you need for solving real world problems using statistics and machine learning algorithms.
- *How to achieve above goal:* We plan to achieve these goals by introducing you to the relevant statistical knowledge, how to use Python to perform these tasks and engage in solving problems, analysis through homework, discussion, and project.

Cont'd

- At the end of the course:
 - Feel inspired to work on and learn more about Machine Learning
 - Understand how various machine learning algorithms work
 - Look at a real-world problem and see if machine learning is appropriate for the given problem at hand.
 - If so, identify what type of algorithm might be applicable
 - Implement them and hopefully their variants and improvements on your own

Course Structure

- The course has two parts
 - Lectures
 - Assignments and project (done in groups)
- Lecture slides and recordings will be available on the course web page

Homework and Exams

- 3-4 homework assignments
 - Permission for late submission should be obtained in advance
 - Throughout the course, you are encouraged to discuss issues in class including homework problems with your peers or me (but everyone should submit his/her own work and no cheating)
- No midterm; no final exam

Project

- **Aim:** Turning machine learning techniques you learn in class to become your strength in dealing with real world problems
- The project involves analysis of the data, implementation of ML algorithm, preparation of a report and presentation of the results during the last week of the class. The project will be done in groups of 2~3 students. If you already working on a research project in the area of interest, you are encouraged to use dataset/topic from your research provided you make some extra effort for the class.
- Detailed instructions for the project will be posted later

Grading

- Grading
 - Homework 40%
 - Project 60%
- Work Load
 - You are expected to put in 8-10 hrs. of work outside of class. A few of you will do well with less time than this, and a few of you will need more.

Course description

- Will introduce
 - Basic concepts of predictive analytics
 - Basic of various machine learning algorithms
 - Implementation of algorithms using Python and
 - How to use and write your own code
- Hands-on experience
- Report and presentation for final project

Textbooks

- Machine Learning, Tom M. Mitchell
- Pattern Recognition and Machine Learning, Christopher .M. Bishop
- Data Mining: Practical Machine Learning Tools and Techniques
- An Introduction to Statistical Learning with Applications in R, R.G. James, D.Witten, T. Hastie and R. Tibshirani
- Machine Learning in Action, Peter Harrington
- Machine Learning with R , Brett Lantz
- Introduction to Machine Learning in Cloud with Python, P. Gupta and Naresh K. Sehgal

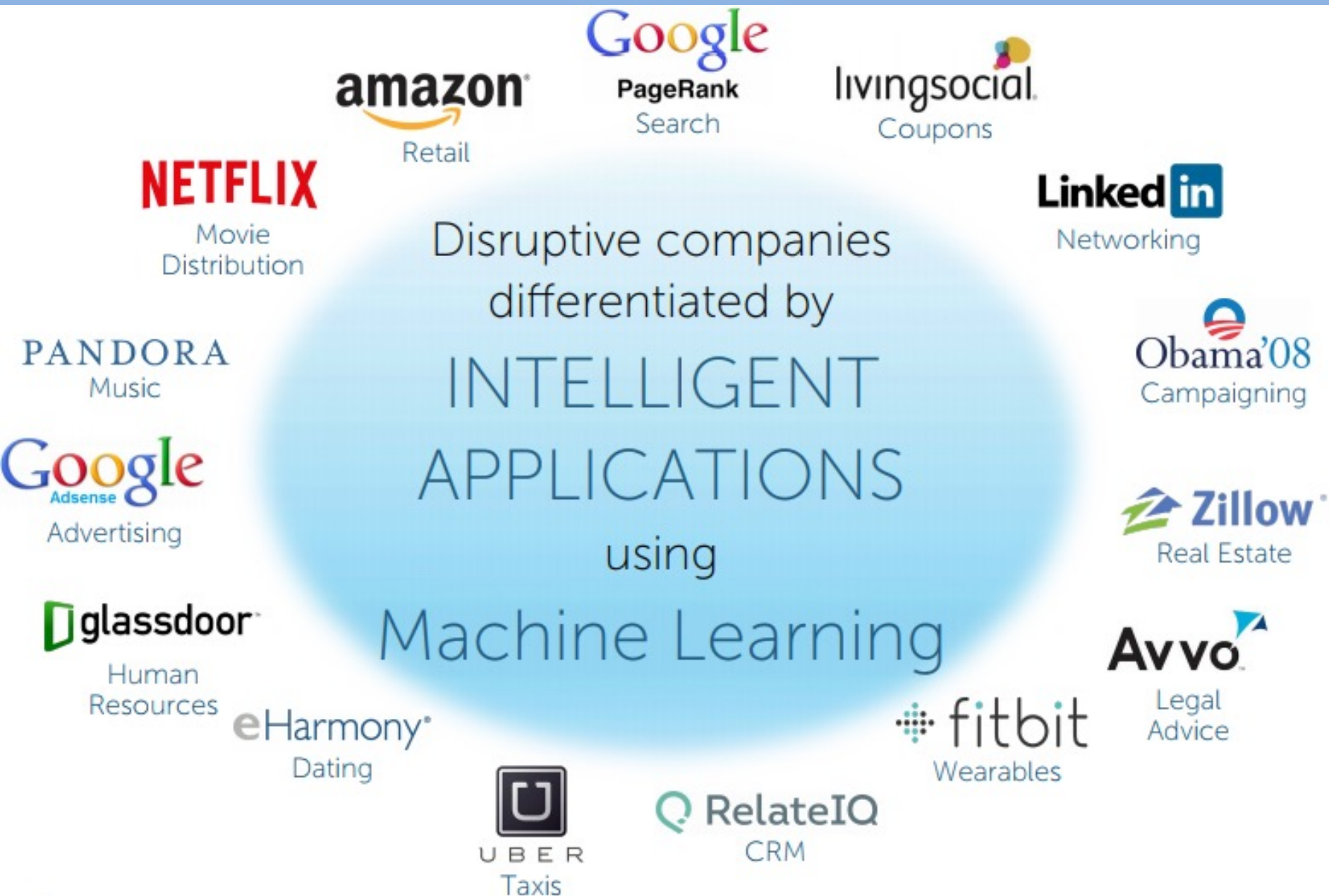
Topics to cover

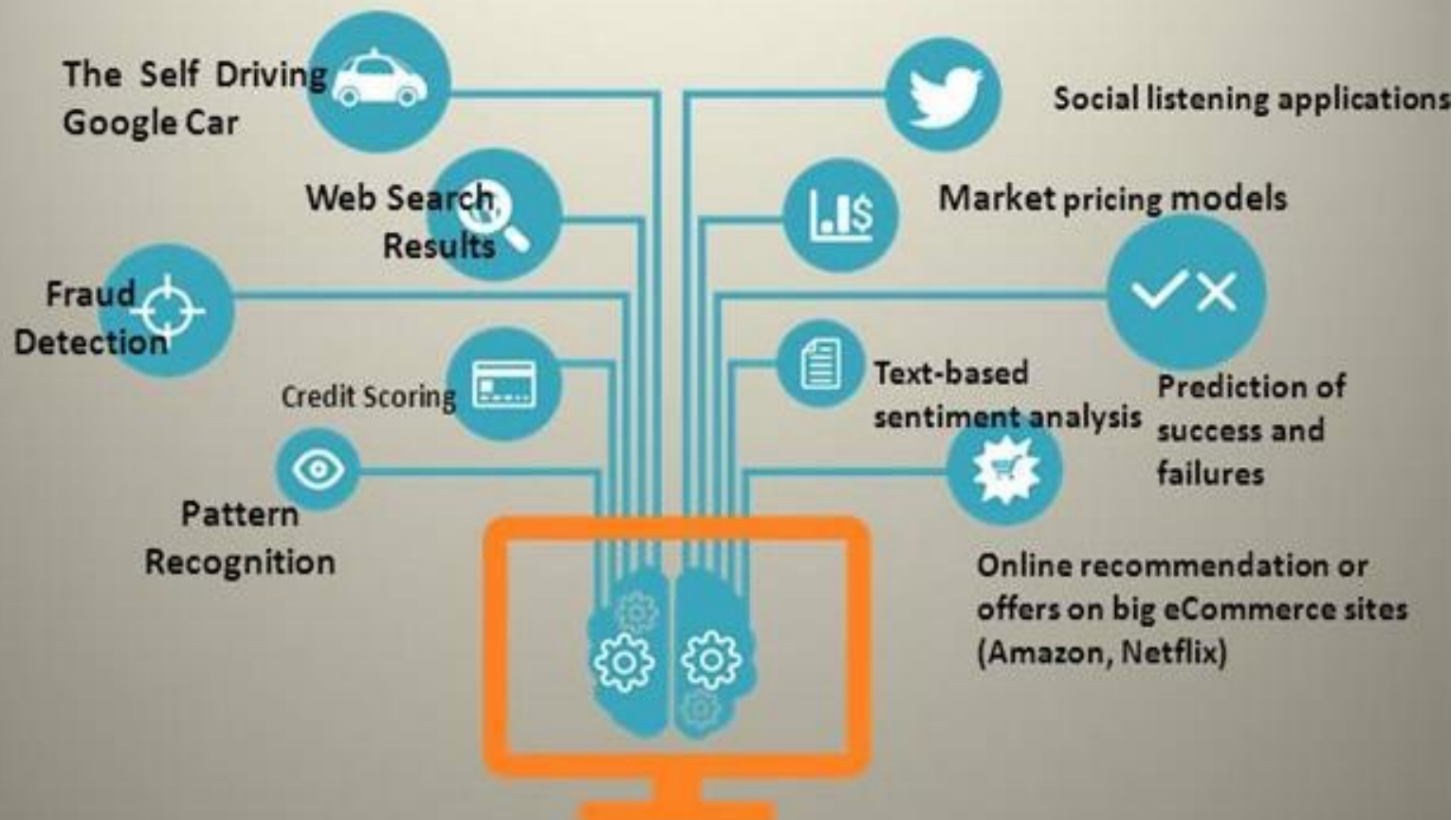
- Introduction
- Data pre-processing
- Classification (supervised learning)
- Forecasting: regression methods
- Clustering (unsupervised learning)
- Evaluating and improving model performance
- Overfitting and regularization
- Dimensionality Reduction
- Applying machine learning: guidance and practical issues

Predictive Analytics

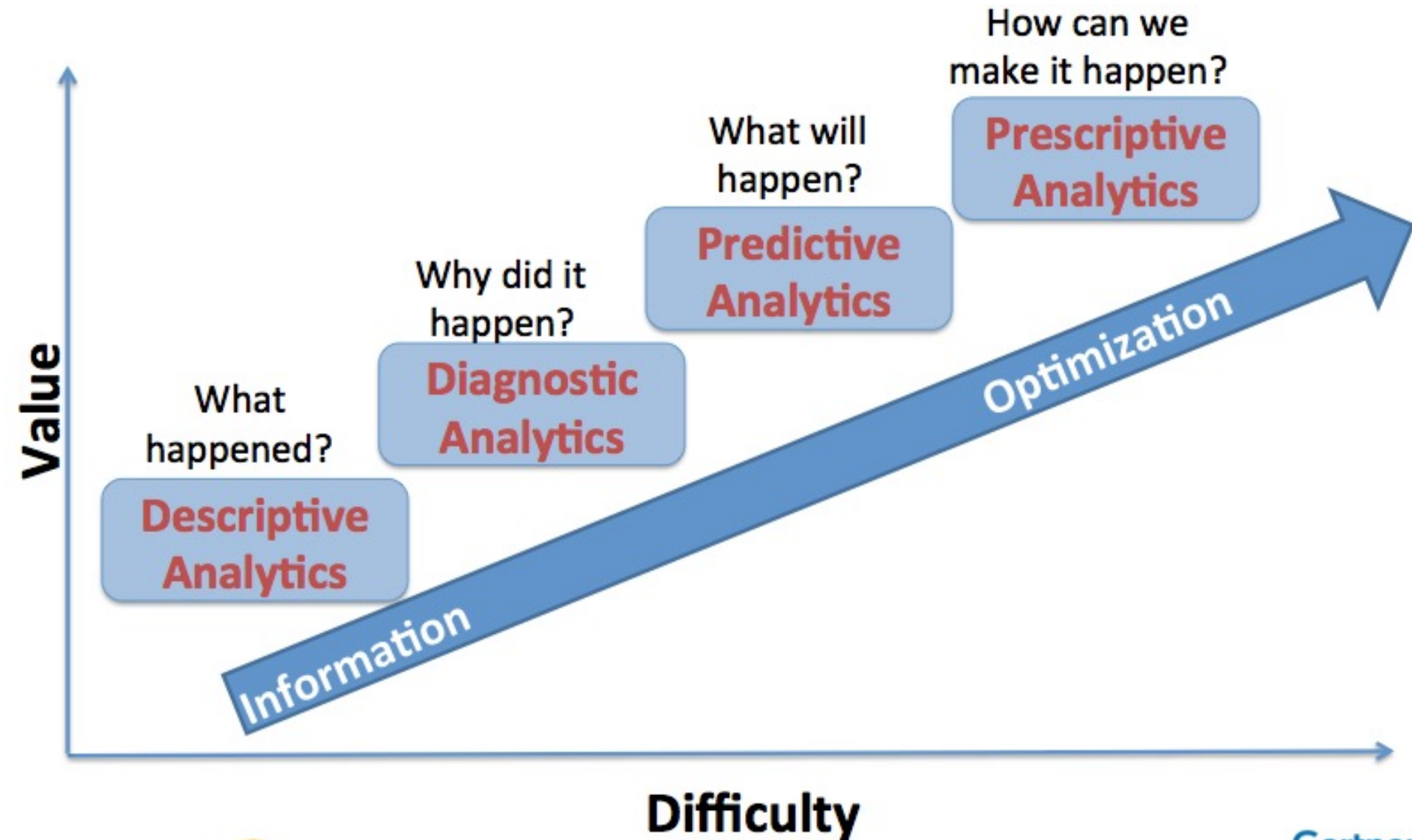
Motivation

- Lots of data is being collected and warehoused
 - Web data, social networking, e-commerce
 - Bank/credit card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is strong
- Data collected and stored at enormous speed (GB/hour)
- Traditional techniques infeasible for raw data
- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all.
 - Web data, social networking, e-commerce
 - Bank/credit card transactions

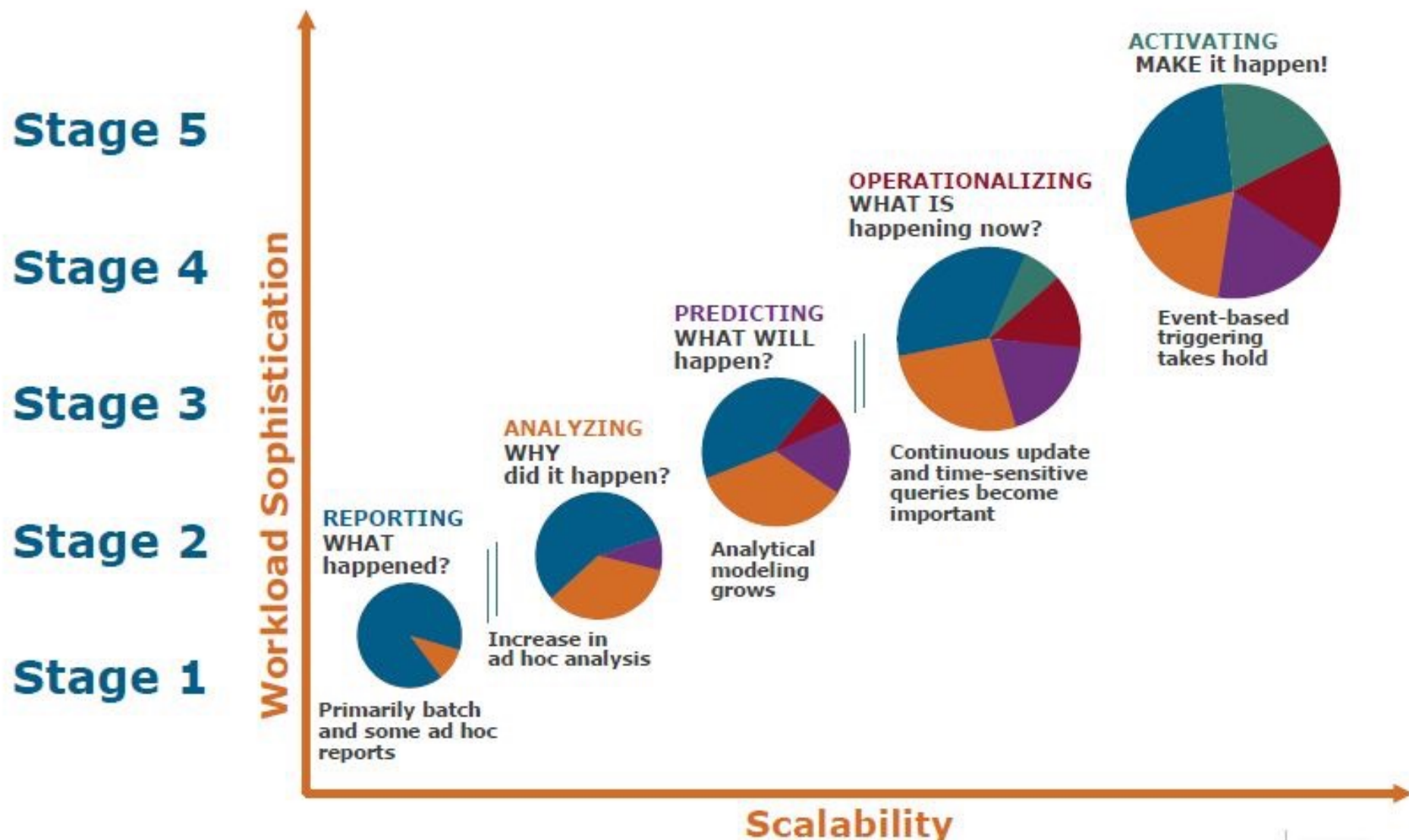




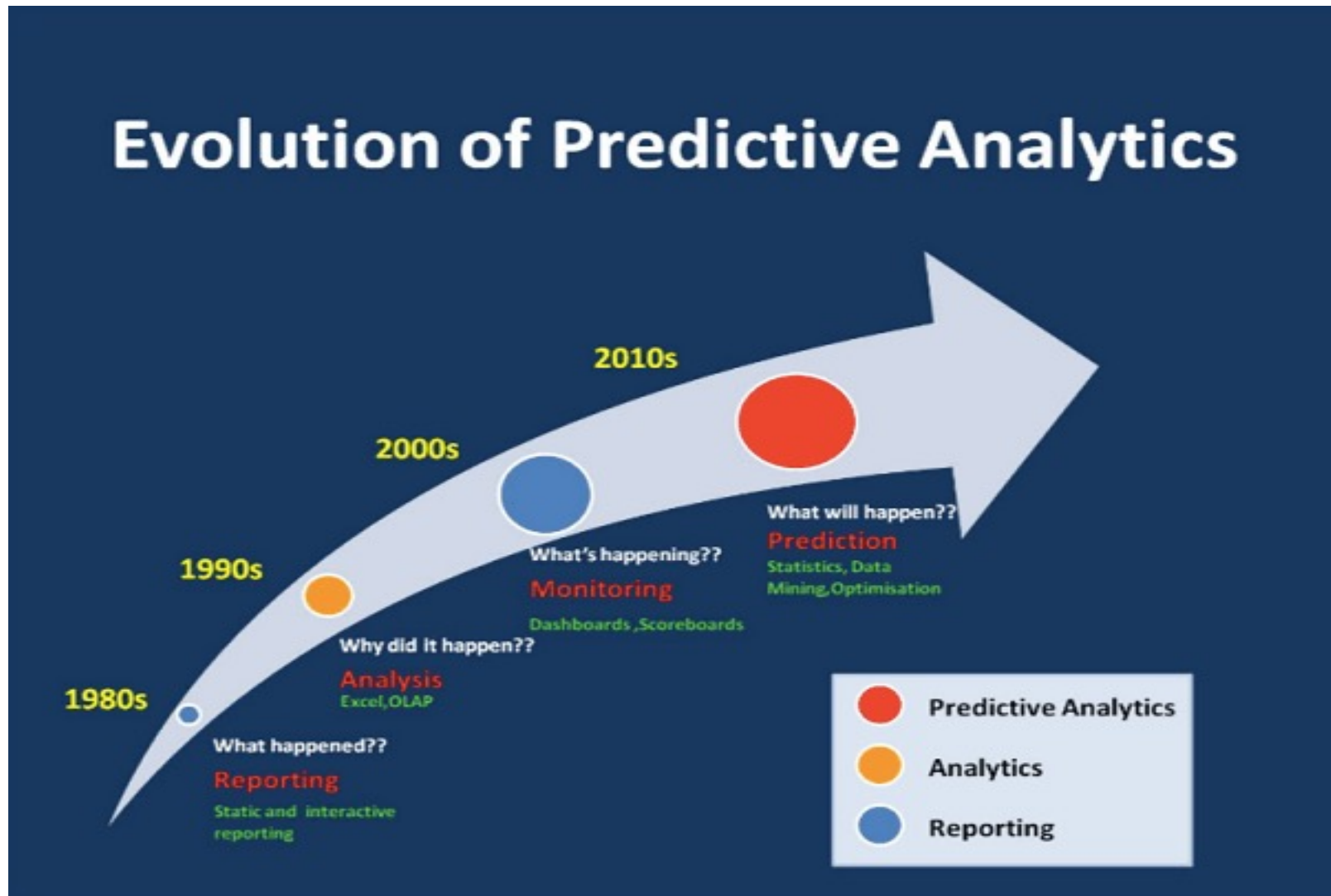
Analytics capabilities framework



Analytics Capabilities Framework



Analytics Capabilities Framework



- **What is Predictive Analytics**
 - Predictive analytics is an area of statistical analysis that deals with extracting information from data and uses it to predict future values and behavior patterns (Wikipedia)
 - The core predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences and exploiting it to predict future outcomes.
 - Predictive analytics uses historical and current data combined with techniques such as advanced statistics and machine learning to model unknown future events.
 - It is generally defined as learning from past collective experience of an organization to make better decisions in the future using data science and machine learning.
- **Other Definitions**
 - Predictive Analytics is emerging as a game-changer. Instead of looking backward to analyze “what happened?” predictive analytics help executives answer “what’s next?” and “what should we do about it?” (Forbes Magazine, April 1, 2010)
 - Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends (searchcrm.com)

Predictive analytics allow for identifying patterns contained in data to assess risks or opportunities for business, addressing important business questions like: which machine needs maintenance? Which product should I recommend now? And who's in danger of going into cardiac arrest? Essentially, you can use predictive analytics to forecast confidence levels of events based on very defined conditions and parameters.

- **Understanding Predictive Analytics**

Let us take an example of a certain organization that wants to know what it will be its profit after a few years in the business given the current trends in sales, the customer base in different locations, etc. Predictive analytics will use the variables given and using techniques such as data mining, artificial intelligence would predict the future profit or any other factor that the organization is interested in.

Why should we use Predictive Analytics

Predictive analytics goes beyond simple descriptive analytics which are the basics that most companies utilize today. Descriptive analytics can only tell the business what has happened. To predict and uncover insights about the future of the business, you need predictive analytics. Those insights can prove extremely valuable in reducing risks, optimizing operations, and increasing profits. Even better, predictive analytics help businesses solve complex problems with predictive models and find new opportunities for business success.

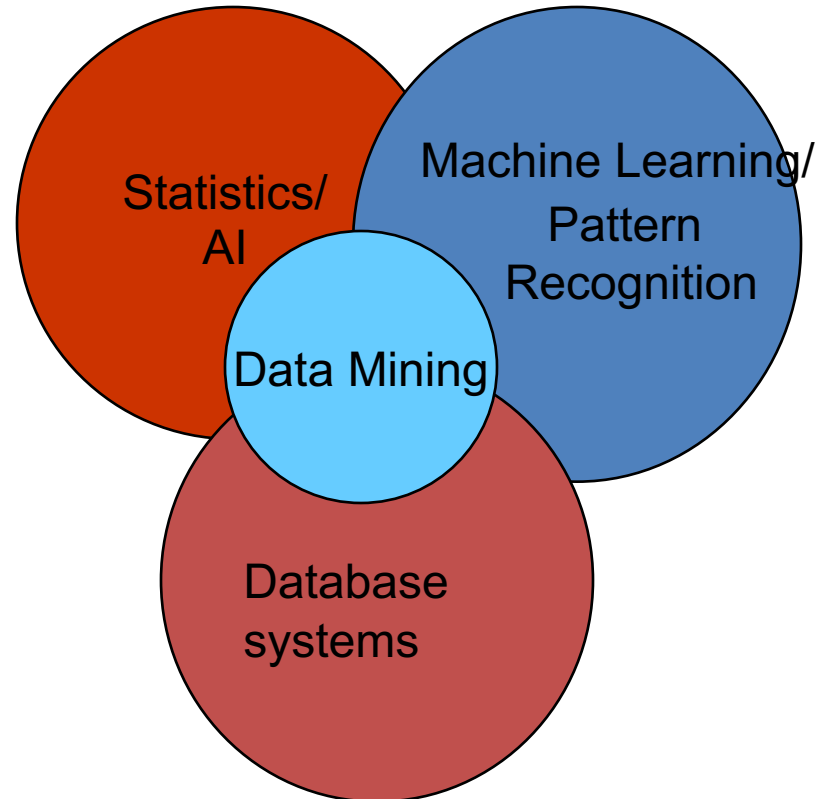
It analyzes a huge volume of data to show many key points in a business, which helps the organization in understanding its areas of strength and weakness. It helps identify future patterns, which can be very useful for an organization in understanding customer needs better, improving their marketing, etc. In a competitive and complex environment, it simplifies the tasks by providing automation, such as keeping two different teams in sync by notifying each other about the status of the other.

How to start Predictive Analytics

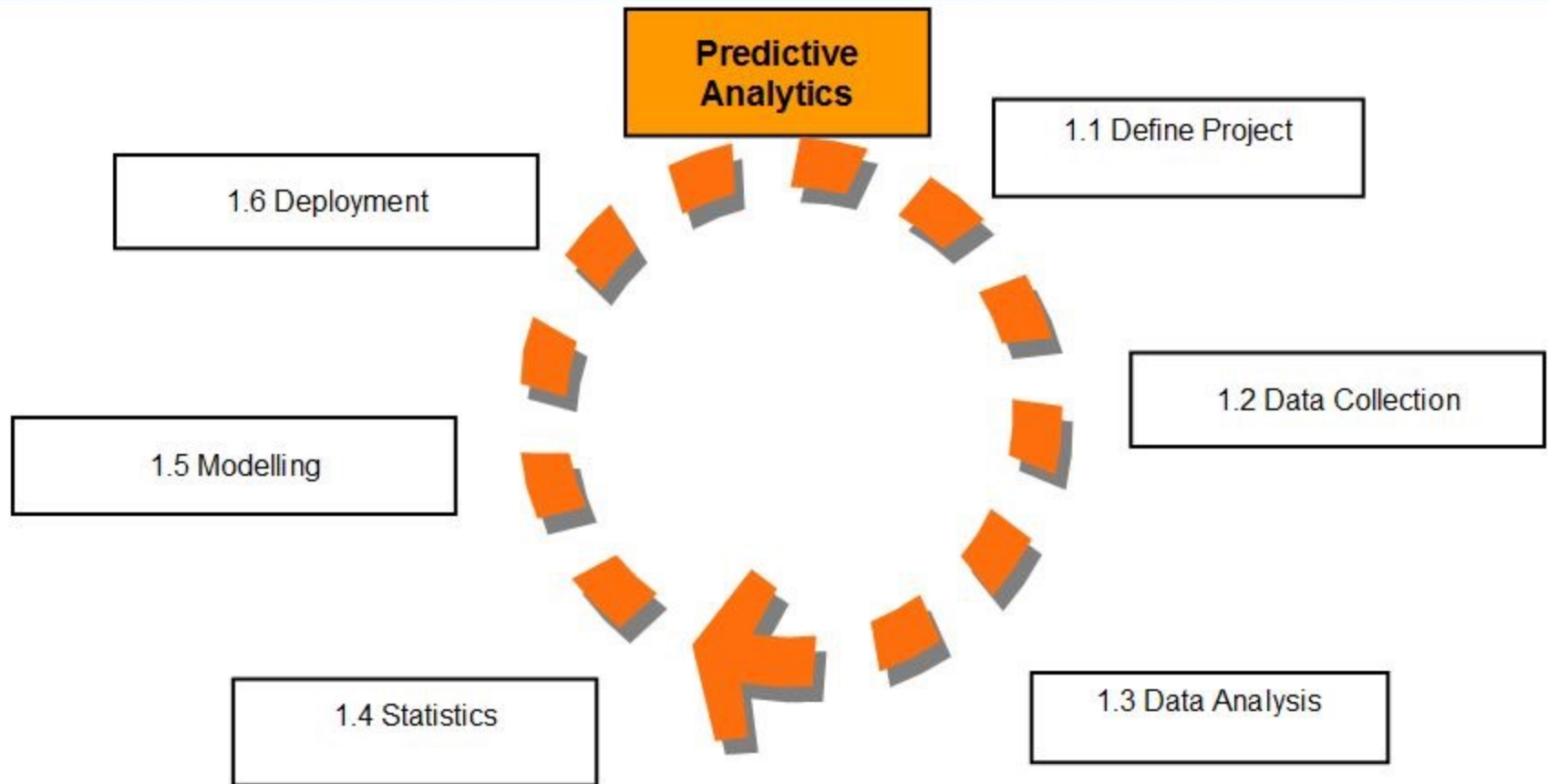
- While getting started with predictive analytics is not so easy, when done right, it can prove extremely beneficial to companies that are religiously invested and committed to the project. The ideal approach would be, to begin with, a limited-scale pilot project in a critical business area. This way, you would be able to limit the start-up costs and reduce the time frame before which the money can start pouring in. Once the predictive model is successful, it will only require fine-tuning to grind out actionable insights in the years to come.

Origins of predictive analytics

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
 - Enormity of data
 - High dimensionality of data
 - Heterogeneous, distributed nature of data



What do predictive modelers do?



How do Predictive Analytics Work?

Predictive analytics rely heavily on machine learning (ML). ML is a combination of statistics and computer science that is used to create models by processing data with algorithms. These models can recognize trends and patterns in data that are generally deeper in sophistication than just visual data discovery methods alone. Using data from diverse sources (for example, the Internet of Things (IoT), sensors, social media, and an array of devices), machine learning processes that data through sophisticated algorithms and builds models for identifying and solving a problem and making predictions.

Predictive analytics also rely on data science, which is a more encompassing concept than just ML. Data science combines statistics, computer science, and application-specific domain knowledge to solve a problem. In a business setting, it combines machine learning methods with business data, processes, and domain expertise to solve a business problem. Basically, it provides predictive insights to decision makers.

What is Machine Learning

- Machine Learning is the most popular technique of predicting the future or classifying information to help people in making necessary decisions.
- It is very hard to write programs that solve problems like recognizing a face.
- We don't know what program to write because we don't know how our brain does it.
- Even if we had a good idea about how to do it, the program might be horrendously complicated.
- Instead of writing a program by hand, we collect lots of examples that specify the correct output for a given input.

What is Machine Learning

- A machine learning algorithm then takes these examples and produces a program that does the job.
- The program produced by the learning algorithm may look very different from a typical hand-written program. It may contain millions of numbers.
- As it trains over the examples, again and again, it is able to identify patterns in order to make predictions about the future.
- If we do it right, the program works for new cases as well as the ones we trained it on.

What is Machine Learning

With the help of Machine Learning, we can develop intelligent systems that are capable of taking decisions on an autonomous basis. These algorithms learn from the past instances of data through statistical analysis and pattern matching. Then, based on the learned data, it provides us with the predicted results.

Machine Learning has opened up a vast potential for data science applications. Machine Learning combines computer science, mathematics, and statistics. Statistics is essential for drawing inferences from the data.

Mathematics is useful for developing machine learning models and finally, computer science is used for implementing algorithms.

What is Machine Learning

Building models is not enough. You must also optimize and tune the model appropriately so that it provides you with accurate results. Optimization techniques involve tuning the hyperparameters to reach an optimum result.

What is Machine Learning

- Let's start with a very "old" attempt at a definition by Arthur Samuel, an IBM pioneer:
- "Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed."
- A good attempt, but many questions remain unanswered. Almost 40 years later, in 1998, Tom Mitchell shaped a "well-posed learning problem" as follows:
- "Well posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E ."
- There is another way of putting it: In traditional heuristic decision-making algorithms, the programmers set the rules according to which the decisions are made. With machine learning, this is done independently by the program without interference from human beings!

What is Machine Learning

- Machine Learning
 - Study of algorithms that
 - improve their performance
 - at some task
 - with experience
- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
 - Solve the optimization problem
 - Representing and evaluating the model for inference

What is Machine Learning

- Machine Learning:
 - Designing algorithms that can learn patterns from historical data, in other words, ML is the field of development of computer algorithms for transforming data into intelligent action
 - Approach: human supplies training examples, the machine learn
 - Example: Show the machine with the data having two classes (C1 and C2) and let it learn to predict if the new data belong to C1 or C2.
- Machine Learning primarily uses the statistically based approach.
 - The statistical model helps to uncover the process which generated the data
- Most desirable property is the *generalization*, i.e., model should generalize well on the new/unseen data

Generalization

- The real aim of supervised learning is to do well on test data that is not known during learning.
- Choosing the values for the parameters that minimize the loss function on the training data is not necessarily the best policy.
- We want the learning machine to model the true regularities in the data and to ignore the noise in the data.
 - But the learning machine does not know which regularities are real and which are accidental quirks of the particular set of training examples we happen to pick.
- So how can we be sure that the machine will generalize correctly to new data?

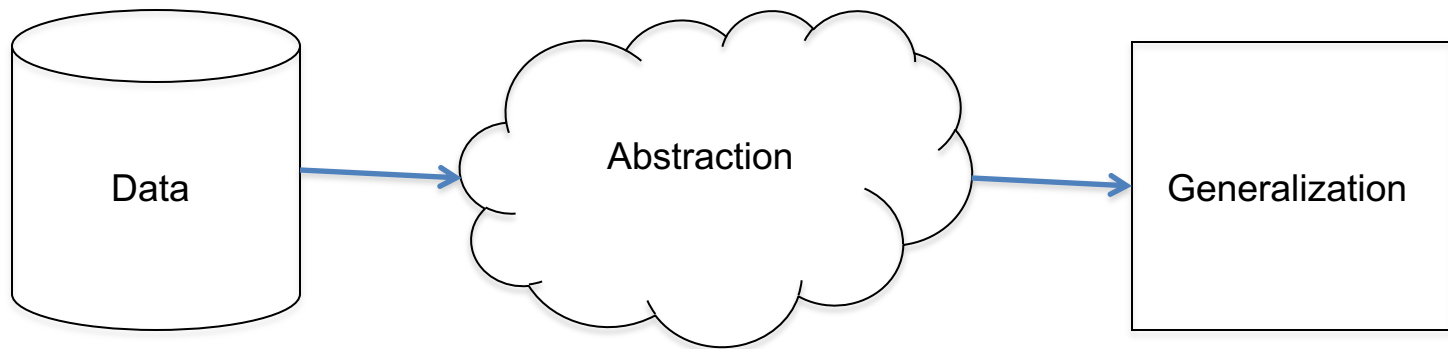
Classical Tasks for ML

- Classification:
 - Mining patterns that can classify future (new) data into known classes
- Clustering/grouping:
 - Identify a set of similar groups in the data
- Prediction/Regression
 - Predict the future value/behavior based on the past history
- Association rule mining:
 - Mining any rule of the form $X \rightarrow Y$, where X and Y are sets of data items, e.g., apple, orange \rightarrow fruits
- Anomaly detection:
 - Discover the outliers/most significant changes in data

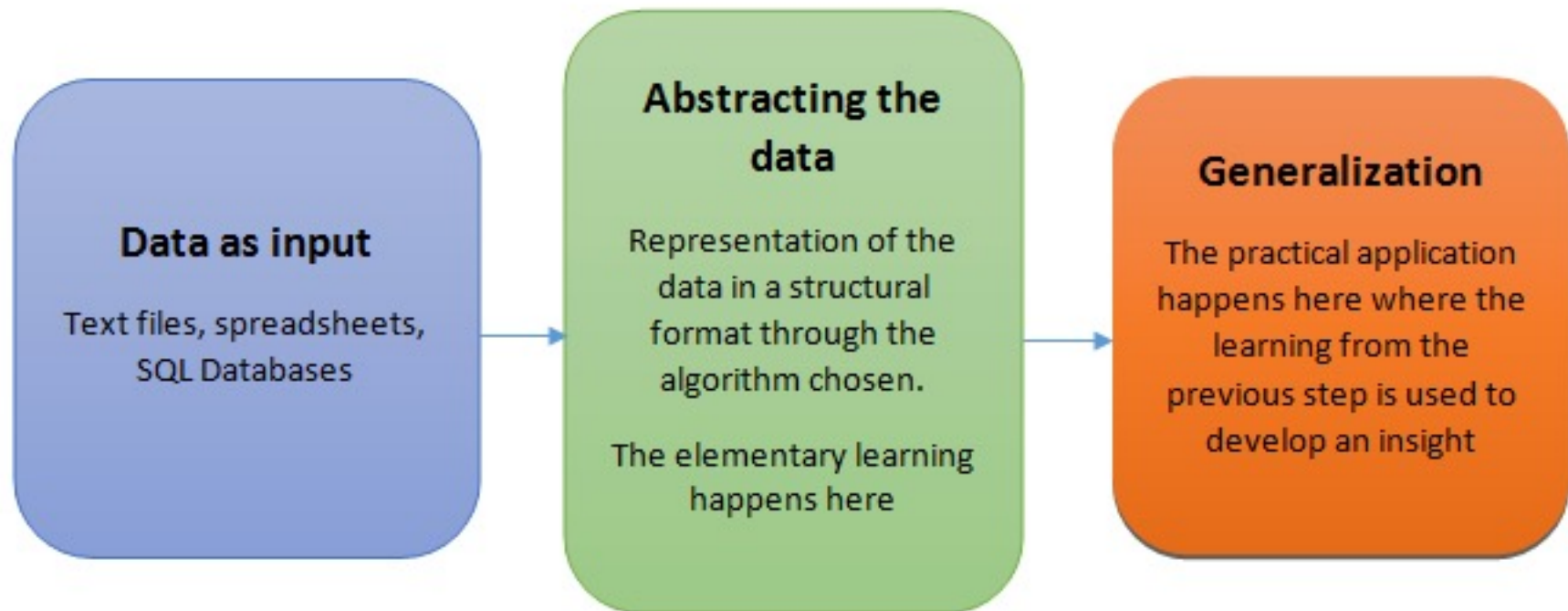
How do machines learn?

- A commonly cited formal definition of machine learning, proposed by computer scientist *Tom M. Mitchell*, says that a machine is said to learn if it is able to take experience and utilize it such that its performance improves up on similar experiences in the future. His definition is fairly exact, yet says little about how machine learning techniques actually learn to transform data into actionable knowledge.
- Regardless of whether the learner is a human or a machine, the basic learning process is similar. It can be divided into three components as follows:
 - **Data input:** It utilizes observation, memory storage, and recall to provide a factual basis for further reasoning.
 - **Abstraction:** It involves the translation of data into broader representations.
 - **Generalization:** It uses abstracted data to form a basis for action.

How do machines learn?



How do machines learn?



What We Talk About When We Talk About “Learning”

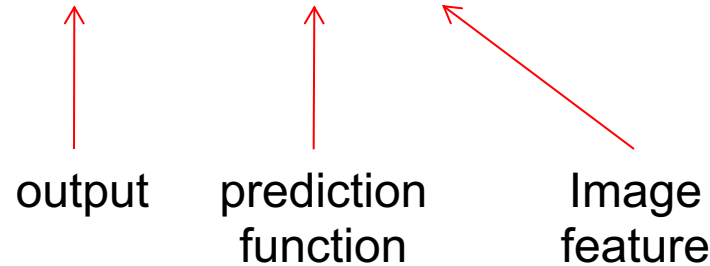
- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:
People who bought “Milk ” also bought “Bread ”
- Build a model that is *a good and useful approximation* to the data.
- These machine learning algorithms use the patterns contained in the training data to perform **classification** and **future predictions**. Whenever any new input is introduced to the **ML model**, it applies its learned patterns over the new data to **make future predictions**. Based on the final accuracy, one can **optimize** their models using various **standardized approaches**.

What We Talk About When We Talk About “Learning”

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:
People who bought “Milk ” also bought “Bread ”
- Build a model that is *a good and useful approximation* to the data.

The machine learning framework

$$y = f(\mathbf{x})$$



- **Training:** given a *training* set of labeled examples $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set
- **Testing:** apply f to a never before seen *test example* \mathbf{x} and output the predicted value $y = f(\mathbf{x})$

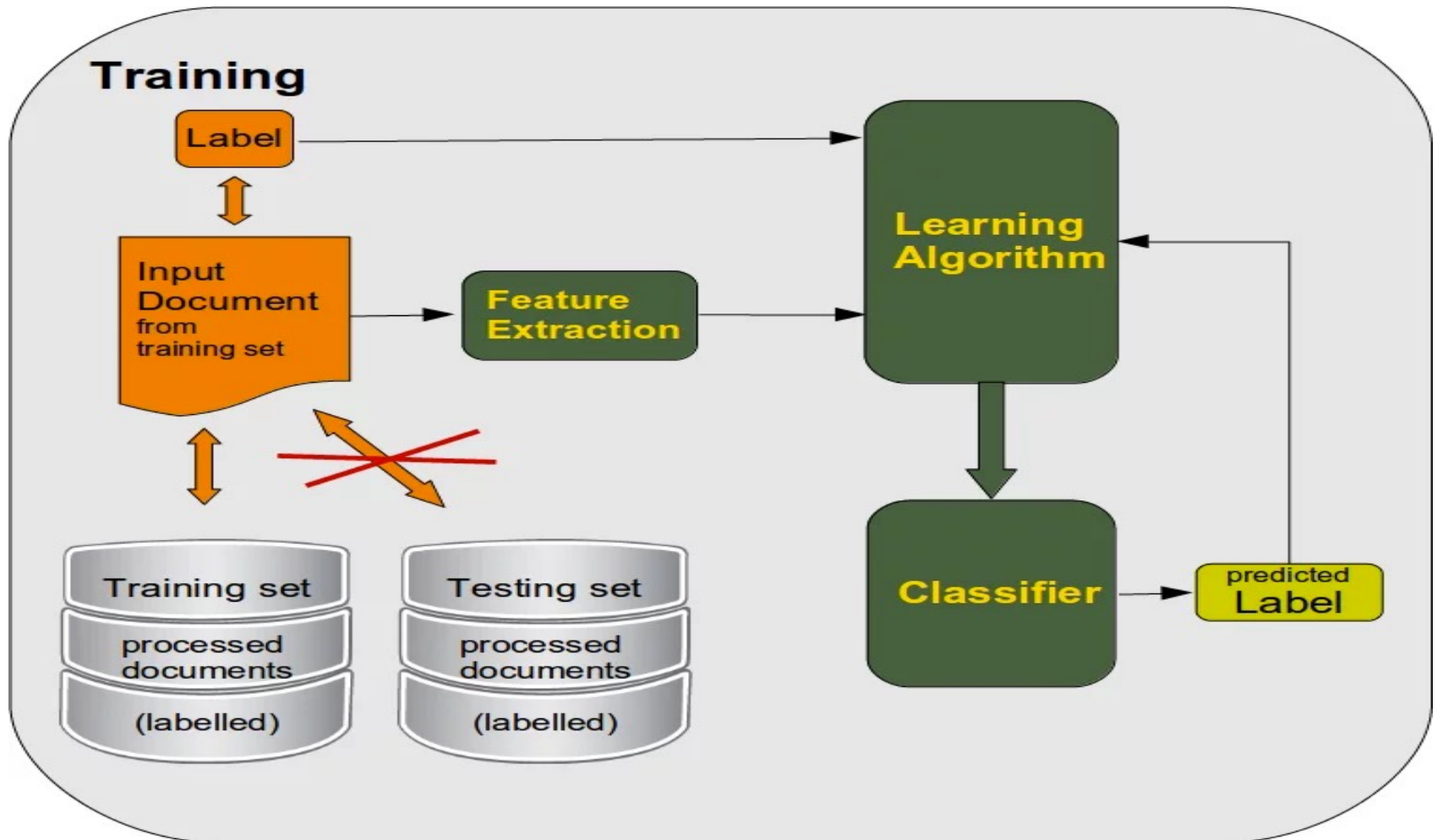
Steps to apply Machine Learning

- **Training a model on the data:** By the time the data has been prepared for analysis, you are likely to have a sense of what you are hoping to learn from the data. The specific machine learning task will inform the selection of an appropriate algorithm, and the algorithm will represent the data in the form of a model.
- **Evaluating model performance:** Because each machine learning model results in a biased solution to the learning problem, it is important to evaluate how well the algorithm learned from its experience. Depending on the type of model used, you might be able to evaluate the accuracy of the model using a test dataset, or you may need to develop measures of performance specific to the intended application.
- **Improving model performance:** If better performance is needed, it becomes necessary to utilize more advanced strategies to augment the performance of the model. Sometimes, it may be necessary to switch to a different type of model altogether. You may need to supplement your data with additional data, or perform additional preparatory work as in step two of this process.

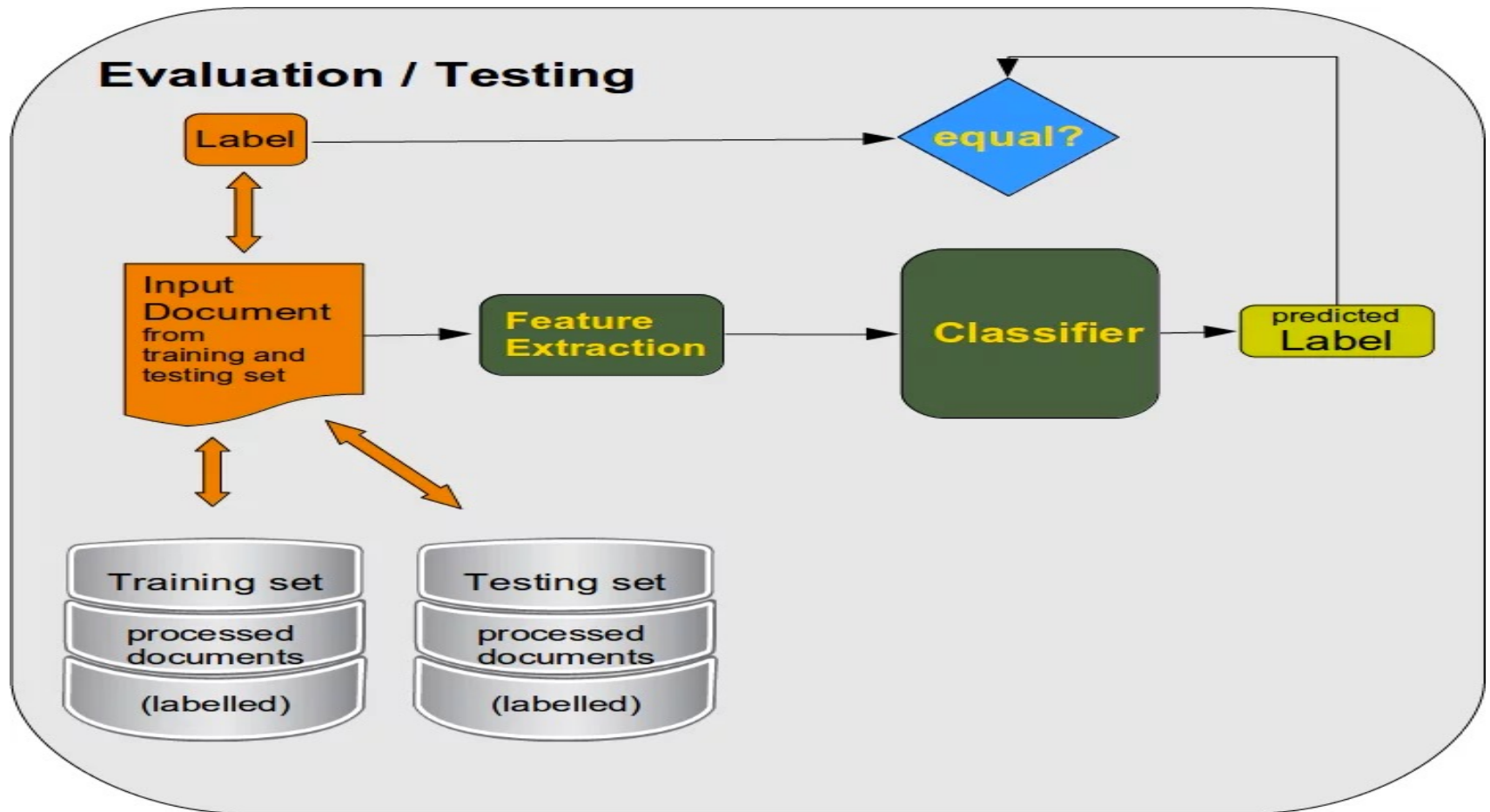
Steps to apply Machine Learning

- **Deployment:** After the above steps are completed and if the model appears to be performing satisfactorily, it can be deployed for its intended task. The successes and failures of the deployed model might even provide additional data to train the next generation of your model.

Steps to apply Machine Learning

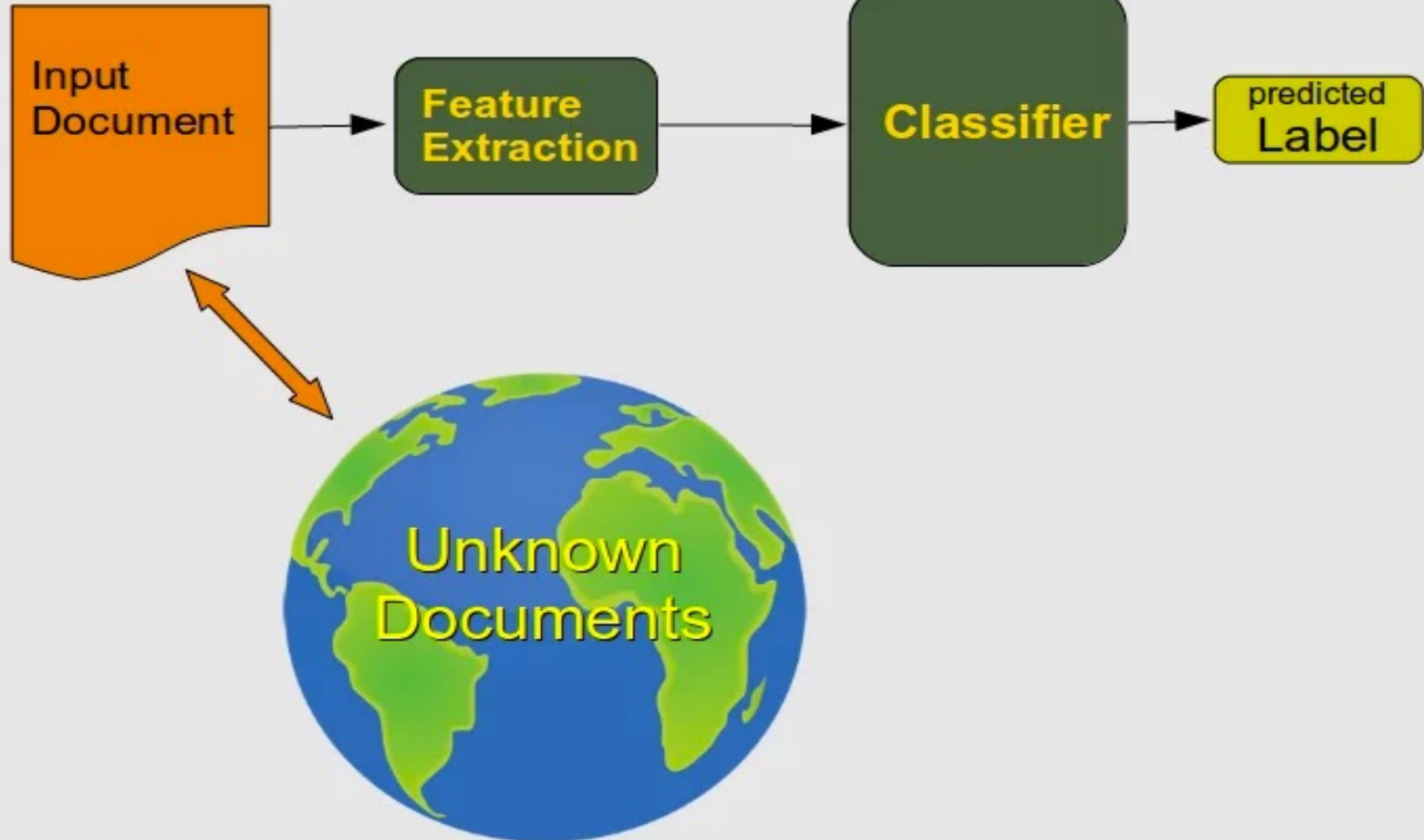


Steps to apply Machine Learning



Steps to apply Machine Learning

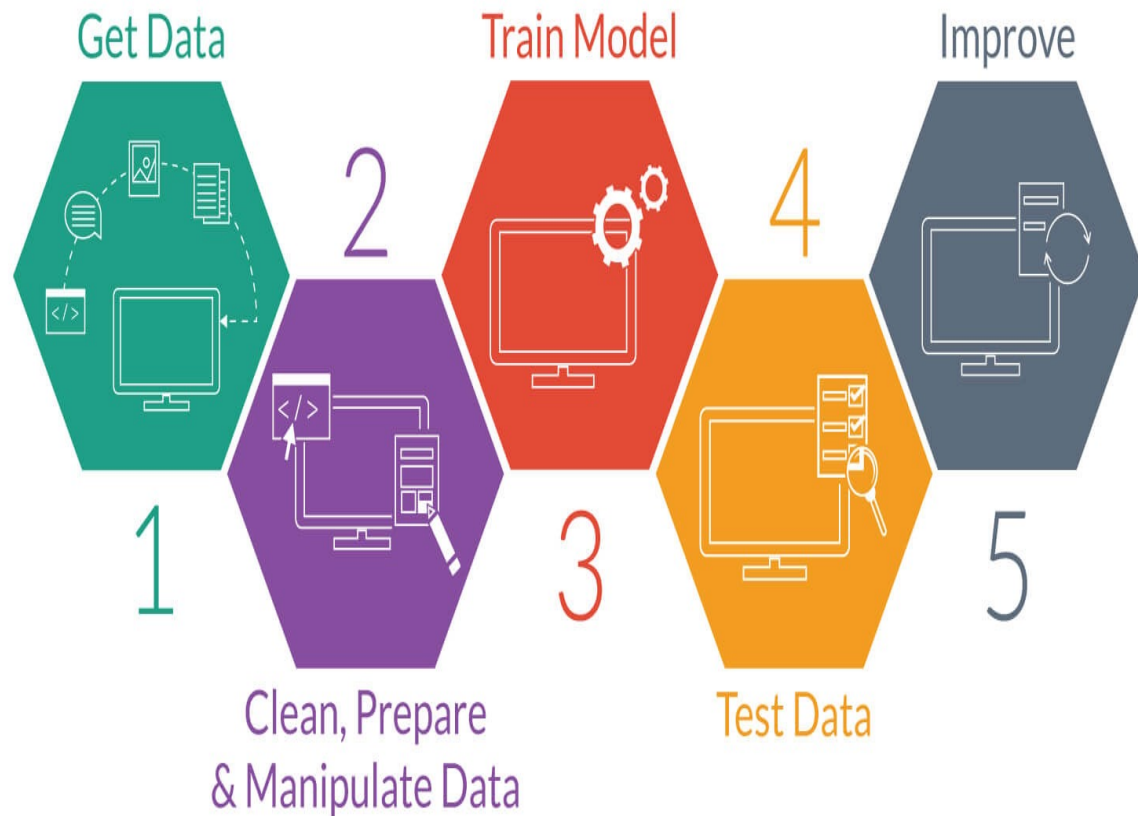
Prediction



Steps in developing ML application

- Collect data
- Prepare the input data
- Train the algorithm
- Test the algorithm
- Use it and improve

Steps to Predictive Modelling



Types of Machine Learning

Machine Learning

```
graph TD; ML[Machine Learning] --> S[Supervised]; ML --> U[Unsupervised]; ML --> R[Reinforcement]; S --> S1[•Regression]; S --> S2[•Classification]; S --> S3[•Ranking]; U --> U1[•Clustering]; U --> U2[•Association]; U --> U3[•Segmentation]; R --> R1[•Control]; R --> R2[•Recommendation systems]; R --> R3[•Reward system];
```

Supervised

- Regression
- Classification
- Ranking

Unsupervised

- Clustering
- Association
- Segmentation

Reinforcement

- Control
- Recommendation systems
- Reward system

Supervised Learning

Supervised learning is that the machine learning task of learning a function that maps an input to an output supported example input-output pairs.

In Supervised Learning, the dataset on which we train our model is labeled. There is a clear and distinct mapping of input and output. Based on the example inputs, the model is able to get trained on the instances.

An example of supervised learning is spam filtering.

Based on the labeled data, the model is able to determine if the data is spam or ham. This is an easier form of **training**.

Spam filtering is an example of this type of machine learning algorithm.

Unsupervised Learning

Unsupervised Learning may be a machine learning technique during which the users don't got to supervise the model. Instead, it allows the model to figure on its own to get patterns and knowledge that was previously undetected. It mainly deals with the unlabeled data.

In Unsupervised Learning, there is no labeled data. The algorithm identifies the patterns within the dataset and learns them. The algorithm groups the data into various clusters based on their density. Using it, one can perform visualization on high dimensional data.

One example of this type of Machine learning algorithm is the Principle Component Analysis.

Clustering is another type of Unsupervised Learning where the data is clustered in groups of a similar order. The learning process in Unsupervised Learning is solely on the basis of finding patterns in the data.

Reinforcement Learning

Reinforcement learning is one among three basic machine learning paradigms, alongside supervised learning and unsupervised learning.

Reinforcement Learning is an emerging and most popular type of Machine Learning Algorithm. It is used in various autonomous systems like cars and industrial robotics. The aim of this algorithm is to reach a goal in a dynamic environment. It can reach this goal based on several rewards that are provided to it by the system.

It is most heavily used in programming robots to perform autonomous actions. It is also used in making intelligent self-driving cars.

Let us consider the case of robotic navigation.

Furthermore, the efficiency can be improved with further experimentation with the agent in its environment. This is the main principle behind reinforcement learning.

There are similar sequences of action in a reinforcement learning model.

ML Cheat Sheet

A step-by-step guide for this sheet:

Learning Styles

Regressions

Classification

Clustering

The Curse of Dimensionality

Our * Wildcard * Section

<https://www.datasciencecentral.com/profiles/blogs/the-making-of-a-cheatsheet-emoji-edition>

MACHINE LEARNING IN EMOJI

SUPERVISED




UNSUPERVISED



REINFORCEMENT

SUPERVISED human builds model based on input / output
human input, machine output
human utilizes if satisfactory


UNSUPERVISED human input, machine output
human reward/punish, cycle continues


BASIC REGRESSION



LINEAR `linear_model.LinearRegression()`
Lots of numerical data   

LOGISTIC `linear_model.LogisticRegression()`
Target variable is categorical  or 


CLASSIFICATION


NEURAL NET `neural_network.MLPClassifier()`
Complex relationships. Prone to overfitting
Basically magic. 

K-NN `neighbors.KNeighborsClassifier()`
Group membership based on proximity 

DECISION TREE `tree.DecisionTreeClassifier()`
If/then/else. Non-contiguous data
Can also be regression  

RANDOM FOREST `ensemble.RandomForestClassifier()`
Find best split randomly
Can also be regression    

SVM `svm.SVC()` `svm.LinearSVC()`
Maximum margin classifier. Fundamental
Data Science algorithm 


NAIVE BAYES `GaussianNB()` `MultinomialNB()` `BernoulliNB()`
Updating knowledge step by step with new info 


CLUSTER ANALYSIS


K-MEANS `cluster.KMeans()`
Similar datum into groups
based on centroids 



ANOMALY DETECTION `covariance.EllipticalEnvelope()`
Finding outliers
through grouping    

FEATURE REDUCTION

T-DISTRIBUT STOCHASTIC NEIB EMBEDDING `manifold.TSNE()`
Visualize high dimensional data. Convert
similarity to joint probabilities 

PRINCIPLE COMPONENT ANALYSIS `decomposition.PCA()`
Distill feature space into components that
describe greatest variance 

CANONICAL CORRELATION ANALYSIS `decomposition.CCA()`
Making sense of cross-correlation
matrices 

LINEAR DISCRIMINANT ANALYSIS `lda.LDA()`
Linear combination of features that
separates classes  

OTHER IMPORTANT CONCEPTS

BIAS VARIANCE TRADEOFF  

UNDERFITTING / OVERFITTING 

INERTIA 

ACCURACY FUNCTION $(TP + TN) / (P + N)$ 

PRECISION FUNCTION $TP / (TP + FP)$ 

SPECIFICITY FUNCTION $TN / (FP + TN)$ 

SENSITIVITY FUNCTION $TP / (TP + FN)$



Types of Machine Learning

Supervised Learning



Classification

- Fraud detection
- Email Spam Detection
- Diagnostics
- Image Classification

Regression

- Risk Assessment
- Score Prediction

Unsupervised Learning



Dimensionality Reduction

- Text Mining
- Face Recognition
- Big Data Visualization
- Image Recognition

Clustering

- Biology
- City Planning
- Targetted Marketing

Reinforcement Learning



- Gaming
- Finance Sector
- Manufacturing
- Inventory Management
- Robot Navigation

Deep Learning

- Deep learning (such as deep neural networks (DNN), recurrent neural networks (RNN) or convolution neural networks CNN) is a part of a broader class of ML methods. Learning can be supervised, semi-supervised or unsupervised.
- They have been applied to fields such as computer vision, speech recognition, natural language processing machine translation, bio-informatics, drug design and self-driving cars where they have produced results comparable to human experts.
- Deep learning uses a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input.
- Learn multiple levels of representations that correspond to different levels of abstraction; the levels form a hierarchy of concepts.

Machine Learning vs Deep learning vs AI

Machine Learning

Machine learning may be a method of knowledge analysis that automates analytical model building. It's a branch of AI supported the thought that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Machine Learning is a part of Artificial Intelligence that involves implementing algorithms that are able to learn from the data or previous instances and are able to perform tasks without explicit instructions.

The procedure for learning from the data involves statistical recognition of patterns and fitting the model so as to evaluate the data more accurately and provide us with precise results.

Deep Learning

- Deep learning is a component of a broader family of machine learning methods supported artificial neural networks with representation learning.
- Learning is often supervised, semi-supervised or unsupervised. Deep Learning is a part of Machine Learning that involves the usage of artificial neural networks.
- Deep Learning machine learning algorithms are the most popular choice in many industries due to the ability of neural networks to learn from large data more accurately and provide steadfast results to the user.

Artificial Intelligence

- AI is the greater pool that contains an amalgamation of all the above-discussed technologies. Artificial Intelligence is still under research and involves imparting sentient intelligence to the machines.
- However, Artificial General Intelligence is still far fetched and will require years of research before we can have even a basic version of it.

Machine Learning in Practice

- Machine learning algorithms are only a very small part of using machine learning in practice as a data analyst or data scientist. In practice, the process often looks like:
 - Start Loop
 1. **Understand the domain, prior knowledge and goals.** Talk to domain experts. Often the goals are very unclear. You often have more things to try than you can possibly implement.
 2. **Data integration, selection, cleaning and pre-processing.** This is often the most time-consuming part. It is important to have high quality data. The more data you have, the more it sucks because the data is dirty. Garbage in, garbage out.
 3. **Learning models.** The fun part. This part is very mature. The tools are general.

Machine Learning in Practice

4. Interpreting results. Sometimes it does not matter how the model works as long it delivers results. Other domains require that the model is understandable. You will be challenged by human experts.

5. Consolidating and deploying discovered knowledge. The majority of projects that are successful in the lab are not used in practice. It is very hard to get something used.

- End Loop
- It is not a one-shot process, it is a cycle. You need to run the loop until you get a result that you can use in practice. Also, the data can change, requiring a new loop.

Why Use Machine Learning

- It is important to remember that Machine learning (ML) is not solution to every type of problem in hand. There are cases where solutions can be developed without using ML techniques. For example, you don't need ML if you can determine a target value by using simple rules, computations, or predetermined steps that can be programmed without needing any data driven learning.
- Machine Learning has revolutionized industries like medicine, healthcare, manufacturing, banking, and several other industries. Therefore, Machine Learning has become an essential part of modern industry.
- Data is powerful and in order to harness the power of this data, added by the massive increase in computation power, Machine Learning has added another dimension to the way we perceive information.
- Machine Learning is being utilized everywhere.

Why Use Machine Learning

Consider using machine learning when you have a complex task or problem involving a large amount of data and lots of features, but no existing formula or equation. For example, machine learning is a good option if you need to handle situations like:

- Hand written rules and equations are too complex (face recognition)
- We can not write the program ourselves
- We cannot explain how (speech recognition)
- Need customized solutions (spam or not)
- Rules are constantly changing (Fraud detection)
- You cannot scale: ML solutions are effective at handling large-scale problems

Why Use Machine Learning

- Develop systems that can automatically adapt and customize themselves to individual users (personalized news or mail filter)
- Discover new knowledge from large databases (Market Basket analysis)

Why now?

- Lots of available data (especially with the advent of internet, social networking and e-commerce)
- Increasing computational power
- Growing progress in available algorithms and theory
- Increased support from industries

Growth of Machine Learning

- Machine learning is preferred approach to
 - Speech recognition, Natural language processing
 - Computer vision
 - Medical outcomes analysis
 - Robot control
 - Computational biology
 - Self driving cars
- This trend is accelerating
 - Improved machine learning algorithms
 - Improved data capture, networking, faster computers
 - Software too complex to write by hand
 - New sensors / IO devices
 - Demand for self-customization to user, environment

Application of Machine Learning

- Broadly applicable in many domains (e.g., pattern recognition, finance, natural language, computer vision, robotics, manufacturing etc.). Some applications:
- Identify and filter spam messages from e-mail
- Speech/handwriting recognition
- Object detection/recognition
- Predict the outcomes of elections
- Stock market analysis
- Search engines (e.g, Google)
- Credit-card fraud detection
- Webpage clustering (e.g., Google News)
- Recommendation systems (e.g., pandora, amazon, Netflix)

Application from Day to Day Life

Artificial Intelligence is everywhere. Possibility is that you are using it in one way or the other and you don't even realize.

- 1.Virtual Personal Assistants ; Siri, Alexa, and Google are some of the popular examples of virtual personal assistants.
- 2.Predictions while Commuting: Traffic predictions, online transportation Networks (when booking a cab, the app estimates the price of the ride. While sharing these services, how do they minimize the detours?
- 3.Video Surveillance: Monitoring multiple video cameras
- 4.Social Media Services: people you may know, Face Recognition etc.
- 5.Email Spam and Malware Filtering
- 6.Product Recommendation , online fraud detection

Application from Day to Day Life



Finance and Banking

- Credit scoring
- Fraud detection
- Risk analysis
- Client analysis
- Trading exchange forecasting



Travel and Booking

- Demand forecasting
- Price optimization
- Price forecasting (for dynamically changing prices)



Retail and E-commerce

- Demand forecasting
- Price optimization
- Recommendations
- Fraud detection
- Customer segmentation



Healthcare and Life Sciences

- Increase in diagnostic accuracy
- Identifying at-risk patients
- Insurance product cost optimization



Marketing and Sales

- Market and customer segmentation
- Price optimization
- Churn rate analysis
- Customer lifetime value prediction
- Upsell opportunity analysis
- Sentiment analysis in social networks



Other

- Object recognition (photo and video)
- Content recommendations (movies, music, articles and news)
- And more

Software paradigm

Top
Frameworks



Programming
languages



Data Scientist's Toolbox

Predictive Analytics Tools in Market



These are some tools that a data scientist use for data analysis purpose.

- Java, R, Python, Clojure, Haskell, Scala...
- Hadoop, HDFS & MapReduce, Spark, Storm...
- HBase, Pig, Hive, Shark, Impala, Cascalog...
- ETL, Web scrapers, Flume, Sqoop, Hume...
- SQL, RDMS, DW, OLAP...
- Knime, Weka, RapidMiner, Scipy, NumPy, scikit-learn, pandas...
- js, Gephi, ggplot2, Tableau, Flare, Shiny...
- SPSS, Matlab, SAS...
- NoSQL, Mongo DB, Couchbase, Cassandra...
- And Yes! ... MS-Excel : the most used, most underrated DS tool.

Essential Libraries for Machine Learning in Python



Python is often the language of choice for developers who need to apply statistical techniques or data analysis in their work. It is also used by data scientists whose tasks need to be integrated with web apps or production environments.

Python really shines in the field of machine learning. Its combination of consistent syntax, shorter development time and flexibility makes it well-suited to developing sophisticated models and prediction engines that can plug directly into production systems.

One of Python's greatest assets is its extensive set of libraries.

Machine Learning is largely based upon mathematics. Specifically, mathematical optimization, statistics and probability. Python libraries help researchers/developers who are less equipped with developer knowledge to easily “do machine learning”.

Pandas for data extraction preparation

Pandas is a very popular library that provides high-level data structures which are simple to use as well as intuitive. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation

It has many inbuilt methods for grouping, combining data and filtering as well as performing time series analysis.

Pandas can easily fetch data from different sources like SQL databases, CSV, Excel, JSON files and manipulate the data to perform operations on it.

Numpy and Scipy

NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities.

SciPy is a very popular library among Machine Learning enthusiasts as it contains different modules for optimization, linear algebra, integration and statistics.

Matplotlib for data visualization

The best and most sophisticated ML is meaningless if you can't communicate it to other people. So how do you actually turn around value from all this data that you have? How do you inspire your business analysts and tell them “stories” full of “insights”? It is through visualization.

This is where Matplotlib comes to the rescue. It is a standard Python library used by every data scientist for creating plots and graphs.

With enough commands, you can make just about any kind of graph you want with Matplotlib. You can build diverse charts, from histograms and scatterplots to non-Cartesian coordinates graphs. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc.

Matplotlib for data visualization

It supports different GUI backends on all operating systems, and can also export graphics to common vector and graphic formats like PDF, SVG, JPG, PNG, BMP, GIF, etc.

Seaborn is another data visualization library

Seaborn is a popular visualization library that builds on Matplotlib's foundations. It is a higher-level library, meaning it's easier to generate certain kinds of plots, including heat maps, time series, and violin plots.

Most commonly used libraries in ML

Scikit-learn for working with classical ML algorithms



Scikit-learn is one of the most popular ML libraries. It supports many supervised and unsupervised learning algorithms. Examples include linear and logistic regressions, decision trees, clustering, k-means and so on.

It is built on two basic libraries of Python, Numpy and SciPy. It adds a set of algorithms for common machine learning and data mining tasks, including clustering, regression and classification. Even tasks involving pre-processing of data like transforming data, feature selection and ensemble methods can be implemented in a few lines and with ease.

What is Scikit-Learn?

Scikit-learn (or sklearn for short) is a free open-source **machine learning library for Python**. It is designed to cooperate with SciPy and NumPy libraries and simplifies data science techniques in Python with built-in support for popular classification, regression, and clustering machine learning algorithms.

Sklearn serves as a unifying point for many ML tools to work seamlessly together. It also gives data scientists a one-stop-shop toolkit to import, preprocess, plot, and predict data.

Scikit-learn provides tools for:

Regression, including Linear and Logistic Regression

Classification, including K-Nearest Neighbors

Model selection

Clustering, including K-Means and K-Means++

Preprocessing, including Min-Max Normalization

Advantages of Scikit-Learn

Developers and machine learning engineers use Sklearn because:

- It's easy to learn and use.
- It's free and open-source.
- It helps in all aspects and algorithms of Machine Learning, even Deep Learning.
- It's very versatile and powerful.
- Detailed documentation and active community.
- It is the most widely used Machine Learning toolkit.

Challenges in Machine Learning

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data ownership and distribution
- Privacy preservation
- Streaming Data

What do data scientist do?

- Define the question
- Define the ideal data sets
- Determine what data is needed
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction / modeling
- Interpret result
- Challenge result
- Write up the result
- Create reproducible code
- Distribute result to other people

Machine Learning Skills Pyramid v1.0

Will the real
"Data Scientist"
please stand up?

ML
Researcher

Creates
Algorithms

Machine Learning Researcher/Scientist:

- Research novel machine learning problems
- Creates new mathematical models and algorithms
- Publishes papers on research results
- Typically PhD/MA Level: Robotics, Machine Learning, Cognitive Science, Applied Statistics, Engineering, Operations Research, Math, etc.
- Skills: Builds mathematical models, Breaks ground in research, Establish new paradigms, Scientific Formalism, Experiment design

ML
Engineer

Applies Algorithms
Create Solutions

Machine Learning Engineer:

- Solves business/data learning problems
- Creates ML solutions to achieve an organization's objective
- Applies established algorithms
- Uses ML algorithm libraries
- Understands strengths and weaknesses of different algorithms
- Typically BS/MA Level: Computer Science, Math, Other Technical
- Skills: Software Eng. PLUS Data Analysis, ML Algorithm Selection, Cross Validation, Metrics/Scoring, Feature Engineering

Data Engineer

Creates Data-Software Infrastructure

Data Engineer:

- Develops code in support of Machine Learning Solutions
- Data extraction, transformation, scraping, joining, cleaning
- Summary Statistics, counting, sampling on request
- Skills: Platform/DB/Language specific expertise, Performance, Parallel and Distributed Computing, Quality, Reliability, Map/Reduce-Hadoop, VMs/Cloud, SQL/noSQL, Production Scaling etc.



Analytics philosophy

- You have to get your hands dirty
- Keep trying out things
- Download data, or some code, and try to run
- Make small tweaks
- Analysis is both a science and art.
- Understand how the analysis has been put together
- There is no way to know everything. Learning is the answer
 - You learn by observing and practicing

Starting with the end in mind

- Ask yourself these before you start the analysis
- What do I want to present?
- Which graphs will I create? and how many?
- What data will I need
- Where I can the data, i.e., source of the data
- Try a “mock plot” with dummy data
- Does it look like what I want

Summary

- we went through the basics of **machine learning** and how **computing power** has evolved over time to accommodate **advanced machine learning algorithms**.
- Computers are **gaining intelligence** owing to the **data** that is generated in a **vast amount**.
- We went through the **different types** of **machine learning algorithms** and further took a brief look at some of the **popular ML algorithms**.