Data Science and Machine Learning with the Watson Data Platform

IBM Cloud

Technical Evangelist Team

Session Overview

Description	The goal of this session is to **familiarize participants with the Watson Data Platform; specifically the Data Science Experience and Watson Machine Learning**. This will be one within the context of analyzing customer churn.
Audience	Intended for individuals seeking to develop a **basic understanding of data science** and machine learning.
Pre-requisites	Pre-requisite skills: • Business Intelligence • Conditioning and management of business data • Familiarity with basic statistics

The pre-reqs are recommended but not requires. This course and the lab can be relevant to all parties

Session Objectives

Upon completion of this session, you should be able to:

- Execute a notebook in the Data Science Experience

- Deploy a Machine Learning Model

- Understand the Tools, Technology, and Processes involved in Data Science and Machine Learning

© IBM 2018

- Feel free to ask questions through out
- The point of today is to challenge you, teach you much and have fun!

Section 1

Introduction to Data Science

Objectives

Upon completion of this section, you should be able to:

- Define Analytics
- Differentiate Between Data, Information, and Knowledge
- Discuss Data Science & the Role of the Data Scientist
- Describe the Data Science Methodology
- List Examples of Tools & Technology Used in Data Science

Analytics

A DEFINITION

*"Analytics are the quantifiable **informational** inputs that use **past data** to identify possible trends that may provi[...] **valuable insight** for future action."* [2]
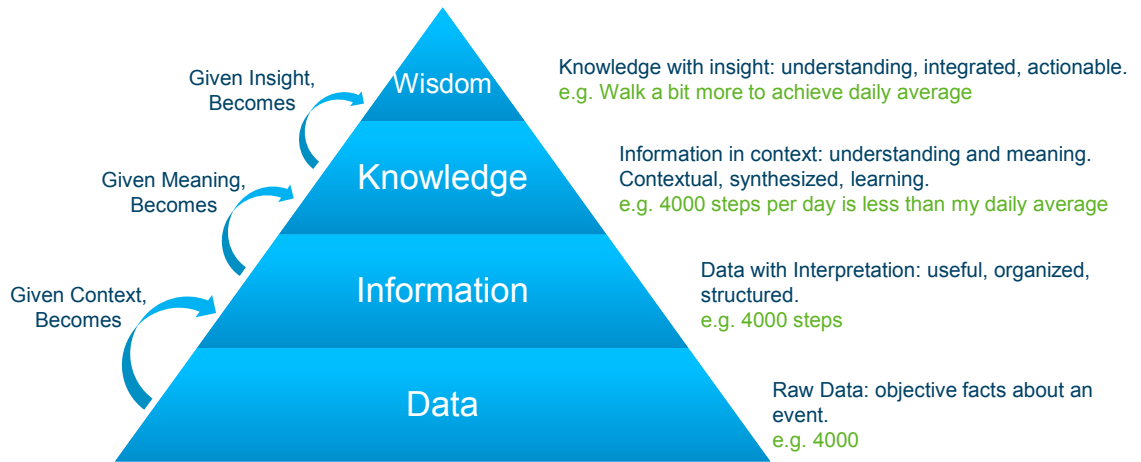

Pixabay. Analytics Word Cloud [1]

[1] Image. Analytics Word Cloud. https://pixabay.com/en/analytics-business-resources-wordle-1368293/

[2] Frolio, Louis. "Foundations of Data Science and Analytics" Week 6, Topic note 1 (2014) Brandeis University, Fall Semester, 2014

Analytics

DATA'S JOURNEY TO VALUE

Given Insight, Becomes → Wisdom

Knowledge with insight: understanding, integrated, actionable.
e.g. Walk a bit more to achieve daily average

Given Meaning, Becomes → Knowledge

Information in context: understanding and meaning.
Contextual, synthesized, learning.
e.g. 4000 steps per day is less than my daily average

Given Context, Becomes → Information

Data with Interpretation: useful, organized, structured.
e.g. 4000 steps

Data

Raw Data: objective facts about an event.
e.g. 4000

Charles Sturt University. Data and Knowledge[1]

© IBM 2018

[1] Information. Charles Sturt University.
https://www.csu.edu.au/division/dit/eal/portfolios/information

Analytics

DATA STRUCTURE TYPES

Structured	Defined data type, format, and structure	Transactional Data
Semi-Structured	Textual data with a discernable pattern, enabling parsing	Self describing XML with schema
Quasi-Structured	Textual data with erratic data formats, can be formatted with effort, tools, and time	Clickstream data
Unstructured	Data that has no inherent structure and is usually stored in different file types	PDF, Excel, JPG

© IBM 2018

"The graphic shows different types of data structures, with 80-90% of the future data growth coming from non-structured data types (semi, quasi and unstructured).

Although the image shows four different, separate types of data, in reality, these can be mixed together at times. For instance, you may have a classic RDBMS storing call logs for a software support call center. In this case, you may have typical structured data such as date/time stamps, machine types, problem type, operating system, which were probably entered by the support desk person from a pull-down menu GUI.

In addition, you will likely have unstructured or semi-structured data, such as free form call log information, taken from an email ticket of the problem or an actual phone call description of a technical problem and a solution. The most salient information is often hidden in there. Another possibility would be voice logs or audio transcripts of the actual call that might be associated with the structured data. Until recently, most analysts would **NOT** be able to analyze the most common and highly structured data in this call log history RDBMS, since the mining of the textual information is very labor intensive and could not be easily automated." [1]

[1] Dietrich, D. Heller, B. Yang, B. (2015). Data Science and Big Data Analytics. Wiley.

Data Science

A DEFINITION

*"**Data science**, also known as data-driven science, is an interdisciplinary field about scientific methods, processes, and systems to **extract knowledge or insights from data** in various forms, either structured or unstructured."*[1]

* Term first used in publication in 1974
* Introduced as an independent discipline in 2001

[1] https://en.wikipedia.org/wiki/Data_science

Data Scientist

MODERN DAY UNICORNS

- **Quantitative**
  - Skilled in mathematics or statistics

- **Curious & Creative**
  - Passionate about finding creative ways to solve problems and portray information

Tveten,J . Data Science 101 [1]

- **Skeptical**
  - Must be able to examine their own work critically

- **Technical**
  - Aptitude for software engineering, programming, and machine learning

- **Communicative & Collaborative**
  - Strong verbal and written skills. Must be able to articulate business value and collaborate with others.

© IBM 2018

"Harvard Business Review called the modern data scientist the "sexiest job of the 21st century". Companies are sitting on mountains of data assets, without the proper skills and tools to translate the data into business value. Enter the modern data scientist. These difficult-to-find individuals have the complete skill set needed to take one of the world's most mispriced assets – data -- and use it to improve the economics of the firm.

Here we see some of the skills that people in the Data Science field bring to the office everyday. The data scientist has a strong background in mathematics and statistics, along with the programming and database expertise needed to build maintainable applications. As data volumes continue to grow, there is also as emphasis on the need for expertise in distributed computing and understanding algorithm computational complexity. And like any good companion, the data scientist is an excellent communicator. They can articulate the stories in the data and how data products can be used to improve business processes.

It's also interesting to note their education – this likely surprises a number of people that data science isn't a collection of PhDs – not at all

The term "data science" like "big data," is not precisely defined but can be described by its principal characteristics and functions:
- Quantitative
  - Skilled in mathematics or statistics
- Curious & Creative
  - Passionate about finding creative ways to solve problems and portray information

- Communicative & Collaborative
  - Strong verbal and written skills. Must be able to articulate business value and collaborate with others.
- Skeptical
  - Must be able to examine their own work critically
- Technical
  - Aptitude for software engineering, programming, and machine learning" [2]

[1] Image Credit: http://www.builtinla.com/2014/09/03/data-science-101-what-data-scientist-does-and-how-become-one

[2] Brandon MacKenzie. (2015). Building Data Science Teams. https://w3-connections.ibm.com/files/app#/file/843f860b-7e68-4d36-a5c1-44a8d6982820

Data Science Team

Data Engineer	Data ingestion pipelines
Data Scientist	Wrangling, exploring, and hacking data
Quantitative Analyst	R&D advanced mathematical algorithms
Data Analyst	Test hypothesis, creates data driven reports
Front-end Developer	Develop end-user applications
Business Expert	Identify business opportunities

"Before we can think about building a data science team, we need to understand what data science is really all about. It's about solving problems. Everything we do in data science is rooted in this underlying methodology for data science that you see on the right-hand side of the screen. This methodology provides a guiding strategy, regardless of the technologies, data volumes, approaches, or human resources.

Building a successful data product requires the concatenation of several skill sets. As you can see in the data science methodology, there are several phases in which collaboration between various roles is required. The romantic idea of a single person having the complete skillset to build data products is actually possible – these people do exist -- but they are very rare.

In practice, several people work on a team (it could be a cross-functional team) to build a data product. These people bring a variety of skill sets to the table – the deeper programming types such as data engineers and front-end developers bring consumability to data science. The mathematics types use statistical algorithms to find patterns in data. And throughout the process, everything needs to be aligned with driving business outcomes – guided by the eye of domain expertise.

Here I've discretized the typical clustering of skill sets that you see in a high-performing data science team into 6 categories: the Data Engineer, Data Scientist, Quantitative ("Quant") Analyst, Data Analyst, Front-end Developer, and Business Expert. Now, we'll look at how each of these roles compare and what these groups of people actually do.

Background: Please read the "Foundational Methodology for Data Science" white paper and/or take the equivalent data science methodology course on BigDataUniversity.com."

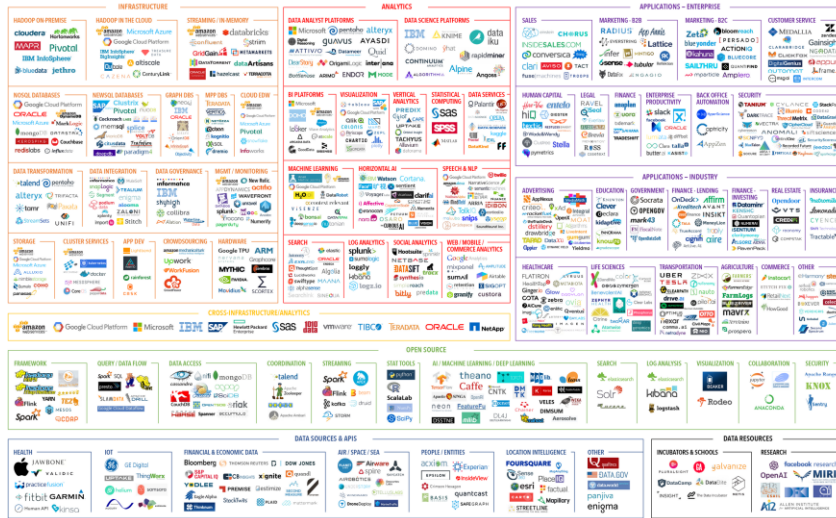# Data Science Methodology

METHODOLOGY DIAGRAM
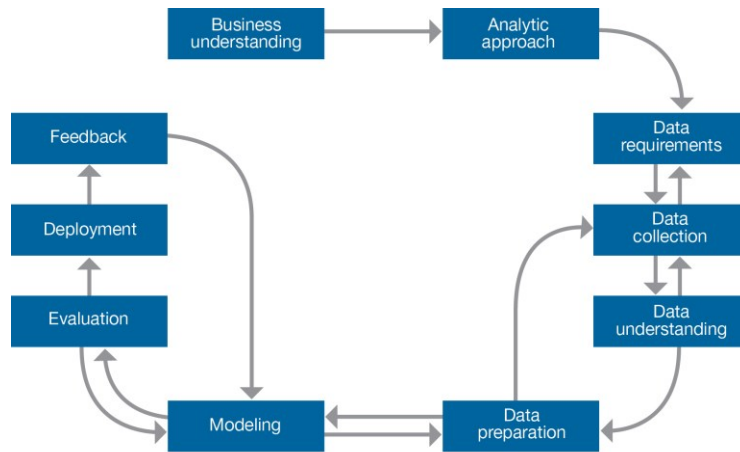
© IBM 2018

"Such a methodology is illustrated in this figure. It bears some similarities to recognized methodologies for data mining, but it emphasizes several of the new practices in data science such as the use of very large data volumes, the incorporation of text analytics into predictive modeling and the automation of some processes.

Looking at this diagram, we immediately spot two outstanding features of our methodology:

• First, it is highly iterative, meaning that we typically iterative between stages throughout the problem-solving process.

• Second, there is no endpoint, meaning that the process remains in play as long the problem is relevant."[1]

[1] John B. Rollins. (2015) Foundational Data Science Methodology. https://w3-connections.ibm.com/files/app#/file/ac78b93f-9fe8-465c-a054-2409184bb861

Tools for Data Science

Tools for Data Science

Infrastructure	Analytics	Applications
Power systems IBM Cloud	Data Science Experience SPSS	Watson Health Watson HR

Cross-Infrastructure/Analytics

Watson Data Platform
IBM Cloud

Open Source

Spark,Jupyter/R/Python

Data Sources & APIs   **Incubators & Schools**

Weather Company, DSX Community Data
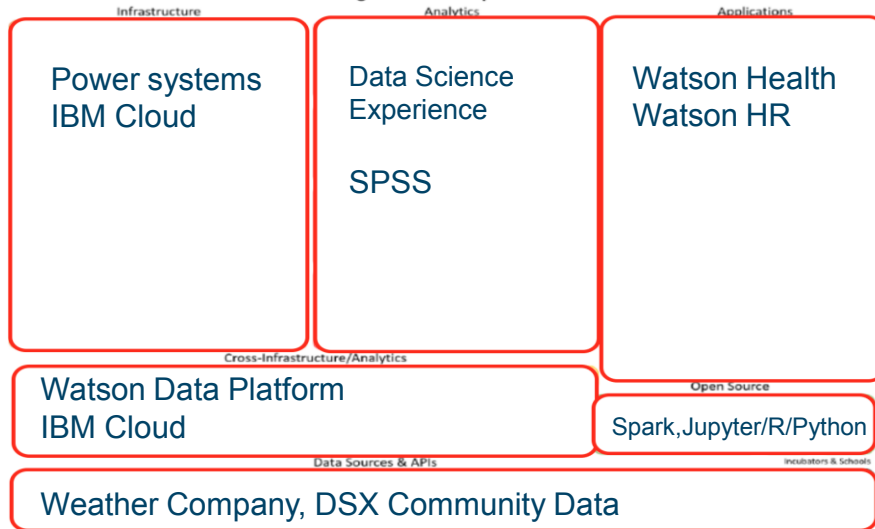
© IBM 2018

OOO GIRLL YESSSL, love it!

Section 1: Summary

Key points covered in this section:

- Relationship Between Data, Information, and Knowledge
- Key Characteristics of Data Science, and the Data Scientist
- Data Science Methodology
- Tools & Technology Used in Data Science by Data Scientists

Section 2
Introduction to Machine Learning

Objectives

Upon completion of this module, you should be able to:

- Define Machine Learning
- Differentiate Between Labeled and Un-labeled Data
- Describe the Main Categories of Machine Learning

This module provides a brief introduction to machine learning including its definition. Key terminologies that distinguish between the various categories of machine learning are discussed along with definitions of the various types of data used in machine learning.

Overview of Machine Learning

Arthur Samuel demonstrating his Checkers program on the IBM 701 computer in 1956.

TERM FIRST COINED BY AN IBMer (circa 1950's)

"Machine Learning is the field of study that gives computers the ability to **learn without being explicitly programmed**.*"*

Arthur Samuel, Artificial Intelligence Pioneer – IBM Corporation.

© IBM 2018

Let's first take an overview of where machine learning from and what is machine learning. The term "machine learning" was first introduced in the early 1950's by an IBM researcher named Arthur Samuel[1].

Arthur was an early pioneer of artificial intelligence and gaming; in fact, his early research was focused on teaching a computer to play checkers[2]. As you know that Alpha-go beat the No1 Go-player now. And that's how you can see machine learning migrated from there to here.

Let see what he had said about machine learning –

"Machine learning is as its name, it is a machine is learning based on sets of data that it recognized." [1]

[1] IBM Journal of Research and Development "Some studies in machine learning using the game of checkers." Volume 44 Issue 1.2

http://ieeexplore.ieee.org/document/5389202/?reload=true&arnumber=5389202

[2] Solving the Game of Checkers. Schaeffer, J. Lake, R. (1996).
http://library.msri.org/books/Book29/files/schaeffer.pdf

Overview of Machine Learning

ML IS THE NATURAL INTERSECTION OF TWO DISCIPLINES

How we build machines that solve problems.

Computer Science

Machine Learning

Statistics

What conclusions can be inferred from data.

How do we get computers to program themselves.
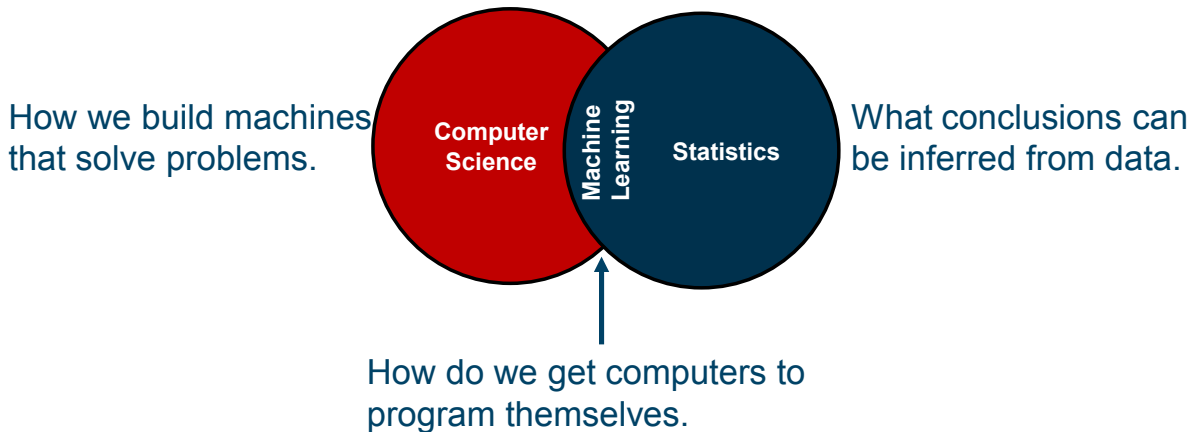
© IBM 2018

<u>"Machine Learning is a natural outgrowth of the intersection of the following two disciplines".</u>

Computer science: We might say the definition of Computer Science is "How can we build machines that solve problems, and which problems are inherently tractable/intractable?"

Statistics: The question that largely defines Statistics might be "What can be inferred from data plus a set of modeling assumptions, and with what reliability?"

And when computer science and statistics come together, there is where the machine learning. It combines with both features, and also let the computers program themselves. So machine learning means to capture useful information from data by teaching the computers program themselves by some computational architectures or algorithms.

And actually machine learning is the foundation of AL and deep learning. That's you can see how important it is.

"Whereas Statistics has focused primarily on what conclusions can be inferred from data, Machine Learning incorporates additional questions about what computational architectures and algorithms can be used to most effectively capture, store, index, retrieve and merge these data, how multiple learning subtasks can be orchestrated in a larger system, and questions of computational tractability. " (1)

[1] Mitchel, T. (2006). The Discipline of Machine Learning. Carnegie Mellon University. http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf

Types of Data in Machine Learning
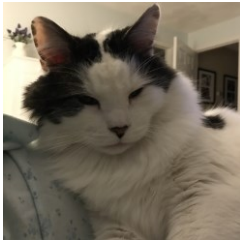
Labeled

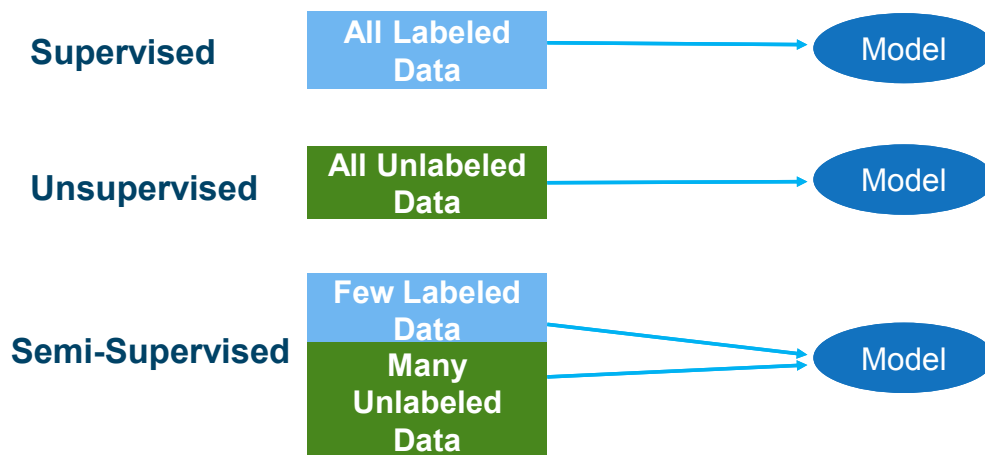Cat

Hot
Wings

Unlabeled

?   ?

?

© IBM 2018

When working with machine learning algorithms, the data it uses is broken down into two camps, labeled, and unlabeled.

Labeled data refers to data that has been identified and assigned a label. Examples include the identification of a disease for a given patient, identifying that an image is that of a cat, or an indication that a person who applies for credit will be a credit risk.

Unlabeled data is the opposite, there exists no clear label or indication. A photograph, where you do not recognize the faces and where there is no name associated with it is just one example.

Main Categories of ML

Supervised	All Labeled Data	→ Model
Unsupervised	All Unlabeled Data	→ Model
Semi-Supervised	Few Labeled Data / Many Unlabeled Data	→ Model

© IBM 2018

The main categories of machine learning include supervised learning, unsupervised learning, and semi-supervised learning.

One defining characteristic of supervised learning is that it requires all labeled data to create a model: We call this training a model. Given a set of features (variables) with a known (correct) output, when applied to machine learning approach (algorithm) we are able to train the algorithm to make informed decisions about data that is not labeled.

With unsupervised learning, we don't set out to train the model; instead, the objective is to unearth meaningful groupings within the data. In essence, the desired outcome of an unsupervised learning process to identify which data logically fits together. You would say "this data looks more like this data than that data."

Semi-supervised data is the intersection of both approaches. The best example can be seen on Facebook, when you upload a picture for the first time Facebook book will ask who is in the picture. You manually assign the correct label "this is Louis" and then, when you upload another picture that includes Louis, Facebook will automatically know who it is.

Supervised Learning

Supervised Learning[2]

- There is a teacher

- Correct classes of training data are known

- Output is a model, or rule set

Taking a closer look at supervised learning we say that "Supervised Learning is the machine learning task of inferring a function from labeled training data." [1] The training data is referred to as the teacher because it has the information necessary to teach the machine learning task. The output of a supervised machine learning task is a function which can be used to map known inputs to a new output. Functions can be rule-sets (Decision trees), algebraic equations (linear regression), or other complex "black box" structures like Random Forests.

[1] Mehryar, M. Afshin Rostamizadeh, Ameet Talwalkar (2012) Foundations of Machine Learning, The MIT Press

[2] Photo via https://www.lexalytics.com/images/extra/machinelearning.png

Supervised Learning

WHAT DO YOU SEE?

You can think of the way how you know them when you are young… your mother told you this is an apple. And then after a few times training, you know that something looks like this and smell and taste like this is an apple. So these things are recognized because over time we were taught to recognize them by knowing it in different ways. And those are all associated with its names.

Supervised Learning

Features (Input)   Label (Output)

Color	Texture	Shape	Avg Size (in)	Sweet	Object
Green	Smooth	Oblate	4	No	Tomato
Yellow	Smooth	Round	3.5	Yes	Apple
Green	Rough	Oblate	2	No	Tomatillo
Red	Lumpy	Cylindrical	2.5	No	Chili
Green	Rough	Oblong	11	No	Cabbage

Training Data
(Teacher)
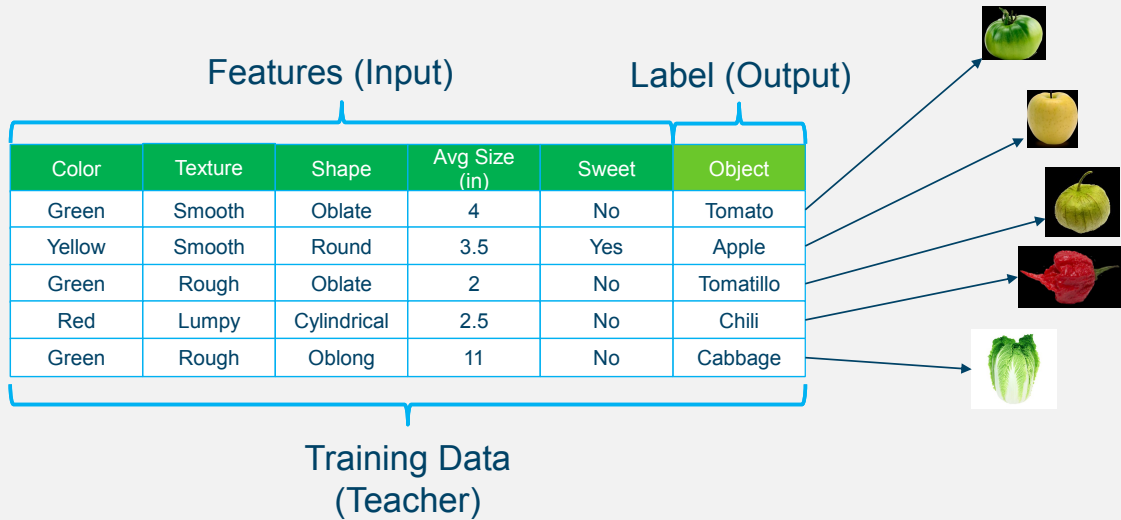
The image above demonstrates one interpretation of the fruits and vegetables into meaningful rows and columns of data, labeled data. This data represents a training data set for several labeled fruits and vegetables. Albeit small, this data set is fully capable of helping identify whether or not a newly introduced vegetable or fruit is one of the six items listed in the data set.

Unsupervised Learning
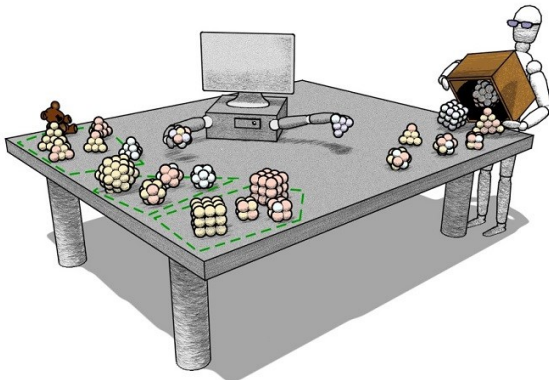
- There is **NO** teacher

- Correct classes of training data are **NOT** known

- Output is natural meaningful groupings

© IBM 2018

With unsupervised learning only the features (variables) are considered. There is no training data to train a model, the correct classifications of the data is not known, and the output is not a function but instead a natural meaningful grouping. Because labeled data is not used, there is no mechanism to evaluate the accuracy of the output. This is a key differentiator from supervised learning.

Unsupervised Learning

WHAT DOES A MACHINE SEE?

© IBM 2018

Unlike the case of supervised learning, where the fruits and vegetable are labeled, unsupervised learning is to figure out meaningful and shared features of the fruits and vegetables. Suppose that you don't recognize those things. You don't know it is an apple, or pear or lettuce. But that doesn't matter. The system doesn't expect me to know its name. Just the next time I see it again, I will know that is the same thing I I've see today. So how do we have the ability to know it/ to group it as the same thing we see today? By Category or by group

Unsupervised Learning

GROUP AND ASSOCIATE
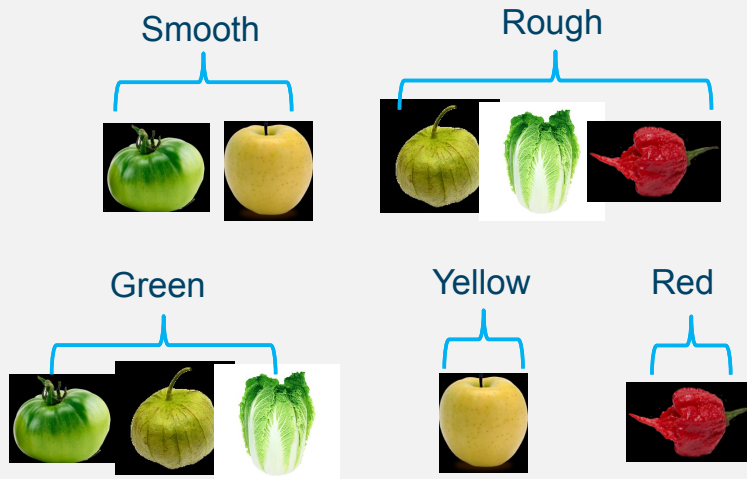
© IBM 2018

An unsupervised machine learning task will attempt to group objects in a meaningful way. In our example, it is not unreasonable to expect groupings by color, shape, or tactile feel. This example is overly simplistic but this may not be the case with all unsupervised machine learning tasks. Often, this approach to machine learning is used to unearth patterns in the data that may not be obvious to the naked eye or other forms of analytics.

Semi-supervised Learning

SOME PRIOR KNOWLEDGE
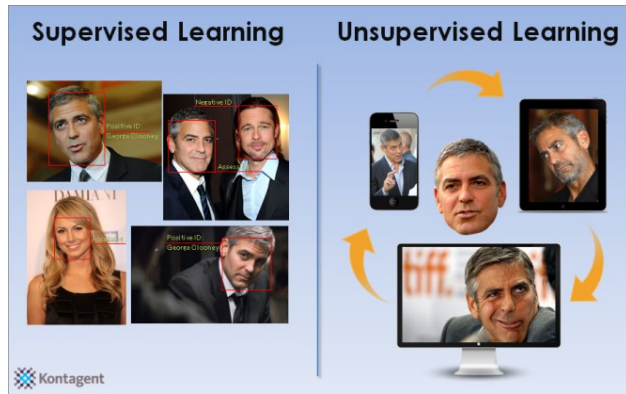
- There is a teaching assistant
- Large amount of data
- Correct classes of training data are known for a small subset
- Output is natural "meaningful" groupings

Semi-supervised learning is the intersection of supervised and unsupervised learning. Using a small training data set along with a supervised machine learning task, a model is created to help predict, or label, data that is not labeled. In the example above, data about images of the actor George Clooney are used to train a model which can be used in conjunction with unsupervised machine learning techniques to identify George Clooney in a other images, some of which may be distorted images of this likeness.

(Another example of Facebook The best example is Facebook. If you upload a picture for the first time Facebook book will ask who is in the picture. You manually assign the correct label "this is Ashley" and then, when you upload another picture that includes Ashley, Facebook will automatically know who it is.)

Section 2: Summary

Key points covered in this module:

- Machine Learning Terminology
- Labeled and Un-Labeled Data
- Main Categories of Machine Learning

In this module, you learned key machine learning terminology, the main categories of machine learning, and the two types of data used in machine learning; labeled, and unlabeled.

Section 3

Approaches to Machine Learning

Objectives

Upon completion of this section, you should be able to:

- Explain the Main Goals of Machine Learning
- Discuss Various Approaches to Machine Learning
- Describe When to Use One Approach Over Another

This module provides a brief introduction to the main objectives of machine learning along with the most common approaches to machine learning. Key terminologies that distinguish between "describing" and "predicting" are discussed along with several common approaches such as classification, regression, clustering, plus others.

Main Goals of Machine Learning

DESCRIBE

Y = golf if rain=no and day=Saturday

Help to understand the relationship between the inputs and the output.

PREDICT

Color	Texture	Shape	Avg Size (in)	Sweet	Object

Make predictions for a new sample described by its attributes.

Machine learning can be described as having two main goals:

Describe: Machine learning is an important tool to help explain, or unearth hidden patterns in complex data. These patterns represent real phenomena, and with the help of machine learning can be transformed into mathematical or rule set representations that are meaningful.

Predict: With predictions, you start with a set of input variables from which you predict an output. An example of prediction includes predicting the sale price of a home based on a data set that includes the selling prices of home in its vicinity, interest rates, and demographic information about the buyer and seller.

Classification

IDENTIFY GROUP MEMBERSHIP

hound dog flowchart

caught a rabbit?

yes   no

friend of Elvis   not a friend of Elvis
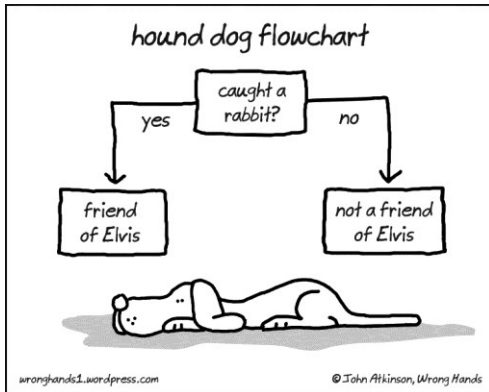
wronghands1.wordpress.com   © John Atkinson, Wrong Hands

A Friend of Elvis[1]

© IBM 2018

- Supervised learning

- Predict class from observations
  o "friend of Elvis"
  o "not a friend of Elvis"

- Response variable is categorical and unordered

- Binary and nominal data

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition. [2]

[1]. Image: Hound Dog. http://guff.com/follow-the-fun-with-these-funny-flow-charts

[2]. Statistical Classification. https://en.wikipedia.org/wiki/Statistical_classification

Classification

TECHNIQUES
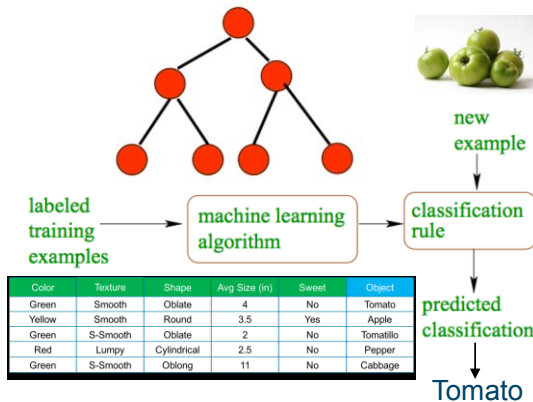
Decision Trees

- Decision Trees

- Logistical Regression

- Naïve Bayes

- Random Forests

© IBM 2018

"Considered to be the most popular machine learning algorithm today, decision trees are easy to interpret, easy to operationalize, and easy to visualize. When a desired outcome is a series of (yes/no) questions, decision trees offer an excellent approach to machine learning.

Reasons to choose:
- Accommodates inputs of any type (numerical, categorical)
- Handles highly correlated independent variables
- Known existence of non-linear (complex) independent variables
- Computationally efficient

Caveats:
- Sensitive to small changes in training data
- Target variable dependent on many independent variables
- Deep trees may signal overfitting of training data" [1]

[1] Diertrich, D. Heller B, Yang, B. (2015). Data Science and Big Data Analytics. Indianapolis: Wiley

Regression

## ESTIMATE OR PREDICT A RESPONSE

THANKSGIVING PREDICTIVE ANALYTICS

"I see no reason why excellent growth shouldn't continue..."

Thanksgiving Turkey. [1]

timoelliott.com
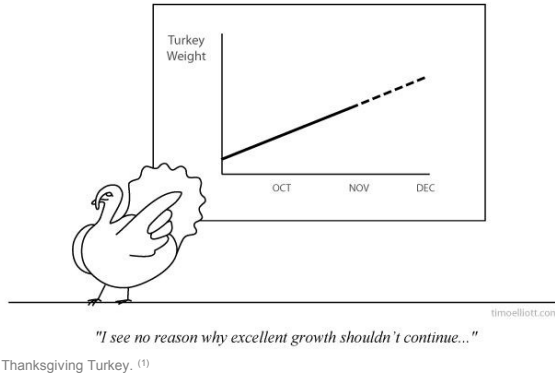
• Supervised learning

• Predict value from observations

• Response variable is numeric value, or probability of class

© IBM 2018

"The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).

Specifically, regression analysis helps one understand how the value of the dependent variable (also referred to as outcome) changes when any one of the independent variables changes (also referred as drivers), while the other independent variables are held fixed. Regression analysis estimates the conditional expectation of the dependent variable given the independent variables — that is, the mean value of the dependent variable when the independent variables are held fixed.

Some example questions are :
• I want to predict the lifetime value of this customer and understand what drives LTV. What drives the LTV higher or lower?
• I want to predict the probability that this loan will default and understand what drives default.

Regression focuses on the relationship between the outputs and the inputs. It also provides a model that has some explanatory value, in addition to predicting outcomes.

Reasons to choose:
• Robust to redundant variables, correlated variables.
• Explanatory value, relative impact of each variable on the outcome.

Caveats:
- Assumes that each variable affects the outcome linearly and additively (e.g. Incomes with a wide dynamic range)
- Can't handle variables that affect the outcome in a discontinuous way. (e.g. Step function)
- Doesn't work well with discrete drivers that have a lot of distinct values. (e.g. Zip codes)" [2]

[1] Illustration. Thanksgiving Turkey.
https://www.pinterest.com/mackenziecorp/predictive-analytics/

[2] Diertrich, D. Heller B, Yang, B. (2015). Data Science and Big Data Analytics. Indianapolis: Wiley

Regression

TECHNIQUES
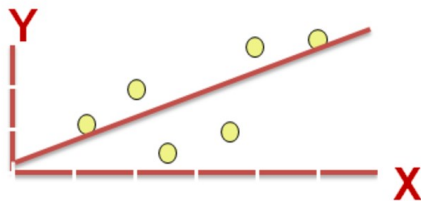
Linear Regression

How much will sales increase (Y) per unit increase of ad expense (X)

Flavors of Regresssion[1]
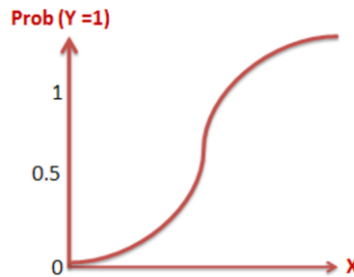
Logistic Regression

What is the change in log odds ratio (Y) per unit increase of ad expense (X)

© IBM 2018

"Regression focuses on the relationship between an outcome and its input variables:

• Doesn't just predict the outcome, it understands how changes in independent variables affect the outcome.

Outcomes can be continuous or discrete,
• When discrete (Logistic Regression) the probability that the outcome will occur is the output.
• When continuous, the predicted value of an outcome is the output.

Examples of the application of Linear Regression include:
• Predicting household income as a function of education, age, and gender.
• House price as a function of median home price in the neighborhood, square footage, number of rooms, etc.

Example of the application of Logistic Regression include:
• Approval or denial of a mortgage
• Customer will purchase or not purchase from website
• Likelihood the New England Patriots will win the next Super bowl"[2]

[1] Images. Difference Between Regressions. http://www.listendata.com/2014/11/difference-between-linear-regression.html

[2] Diertrich, D. Heller B, Yang, B. (2015). Data Science and Big Data Analytics. Indianapolis: Wiley

Clustering

ORGANIZE DATA INTO GROUPS OF MAXIMUM COMMONALITY

Cluster Analysis [1]

© IBM 2018

- Unsupervised learning

- Describe

- Used for exploratory mining

- Output is meaningful grouping based on similarity

"Clustering is a method often used for exploratory analysis of the data. There are no "predictions" of any values done with clustering just finding the similarity between the data and grouping them into clusters The notion of similarities can be explained with the following examples:

Consider questions such as:
1. How do I group these documents by topic?
2. How do I perform customer segmentation to allow for targeted or special marketing programs?

The definition of "similarity" is specific to the problem domain. We are defining similarity as those data points with the same "topic" tag or customers who can be profiled in to a same "age group/income/gender" or a "purchase pattern"." [2]

[1] http://www.kdnuggets.com/2015/12/more-data-science-humor-cartoons.html

[2] Diertrich, D. Heller B, Yang, B. (2015). Data Science and Big Data Analytics. Indianapolis: Wiley

Clustering

TECHNIQUES
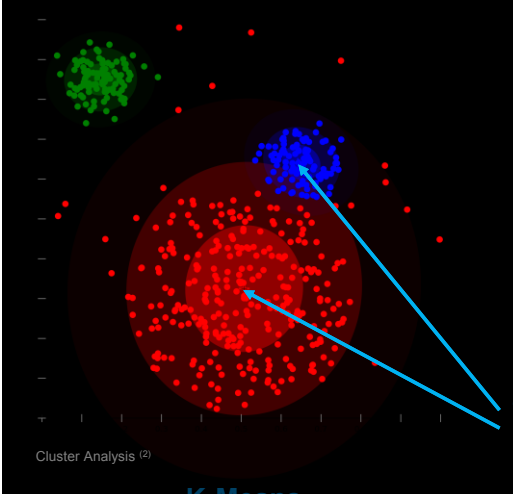
- K-Means
- K-Medians
- Expectation-Maximization
- Hierarchical Clustering

To which centroid is a point associated?

Cluster Analysis [2]

K-Means

© IBM 2018

"K-means clustering is easy to implement and it produces concise output. It is easy to assign new data to the existing clusters by determining which centroid the new data point is closest to it. K refers to the number of clusters being selected.

However, K-means works only on the numerical data and does not handle categorical variables. It is sensitive to the initial guess on the centroids. It is important that the variables must be all measured on similar or compatible scales. If you measure the living space of a house in square feet, the cost of the house in thousands of dollars (that is, 1 unit is $1000), and then you change the cost of the house to dollars (so one unit is $1), then the clusters may change. K should be decided ahead of the modeling process. Wrong guesses for K may lead to improper clustering.

Reasons to choose:
- Easily assign data to new clusters
- Concise output

Caveats:
- Sensitive to initialization
- Variables should be measured on similar or compatible scales
- K: The number of clusters must be known or decided beforehand" [1]

[1] Diertrich, D. Heller B, Yang, B. (2015). Data Science and Big Data Analytics. Indianapolis: Wiley

[2] Data Clustering. https://en.wikipedia.org/wiki/Cluster_analysis

Associations

DISCOVER STRONG RULES ALONG SOME MEASURE

Association Rules [1]
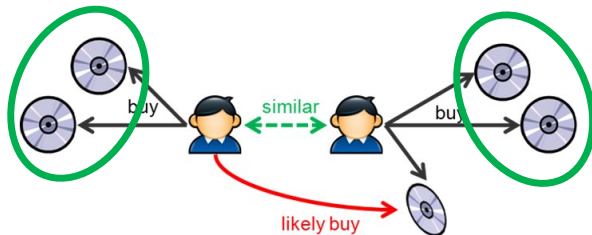
- Supports labeled and unlabeled data
- Describes
- Data mining in large databases
- First introduced to find relationships in POS systems: Market Baskets

© IBM 2018

"Association Rules is another unsupervised learning method. There is no "prediction" performed but is used to discover relationships within the data.

The example questions are:
- Which of my products tend to be purchased together?
- What will other people who are like this person or product tend to buy/watch or click on for other products we may have to offer?
- Advertisements on X could be targeted at buyers who purchase Y

Associations Rules are intended to identify strong rules discovered in data using some measure of interestingness." [2]

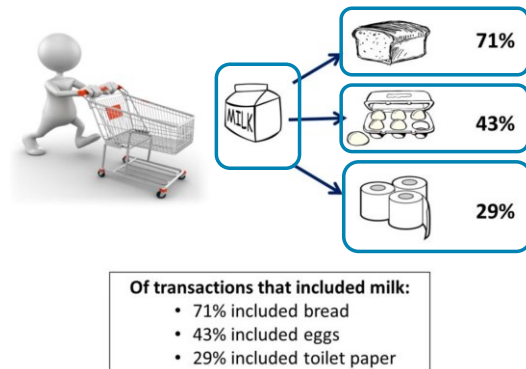[1] Image. Associations. http://horicky.blogspot.com/2011/09/

[2] Diertrich, D. Heller B, Yang, B. (2015). Data Science and Big Data Analytics. Indianapolis: Wiley

Associations

TECHNIQUES

Apriori

Of transactions that included milk:
- 71% included bread
- 43% included eggs
- 29% included toilet paper

© IBM 2018

"Apriori algorithm uses the notion of Frequent Itemset. As the name implies, frequent itemsets are a set of items "L" that appear together "often enough". The term "often enough" is formally defined with a support criterion where the support is defined as the percentage of transactions that contain "L".

For example:
If we define L as a itemset {shoes, purses} and we define our "support" as 50%. If 50% of the transactions have this itemset, then we say the L is a "frequent itemset". It is apparent that if 50% of itemsets have {shoes,purses} in them, then at least 50% of the transactions will have either {shoes} or {purses} in them. This is an Apriori property, which states that any subset of a frequent itemset is also frequent. Apriori property provides the basis for the Apriori algorithm that we will detail in the subsequent slides." [2]

[1] http://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

[2] Diertrich, D. Heller B, Yang, B. (2015). Data Science and Big Data Analytics. Indianapolis: Wiley

Ensembles

COMBINING MULTIPLE LEARNERS

Decision

Combine

Classifier 1 ● ● ● Classifier N

Input Pattern
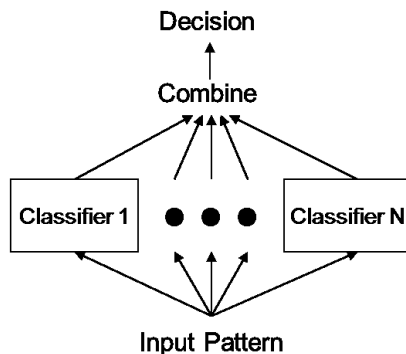
Ensemble Methods[1]

- Supervised learning

- Weak learners independently trained

- Combined predictions produce strong learner

NETFLIX

Winning the Netflix Prize [2]

© IBM 2018

"Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods. In the popular Netflix Competition (2), the winner used an ensemble method to implement a powerful collaborative filtering algorithm.

Voting and averaging are two of the easiest ensemble methods. They are both easy to understand and implement. Voting is used for classification and averaging is used for regression. In both methods, the first step is to create multiple classification/regression models using some training dataset. Each base model can be created using different splits of the same training dataset and same algorithm, or using the same dataset with different algorithms, or any other method." [3]

[1] Image: Ensembles.
https://img.w3cschool.cn/attachments/image/20170604/1496579056599305.jpg

[2] Netflix Prize. http://blog.echen.me/2011/10/24/winning-the-netflix-prize-a-summary/

[3] Demir, N. Ensemble Methods: Elegant Techniques to Product Improved Machine Learning Results. https://www.toptal.com/machine-learning/ensemble-methods-machine-learning

Ensembles

TECHNIQUES

Random Forests(1)

Random Forests

© IBM 2018
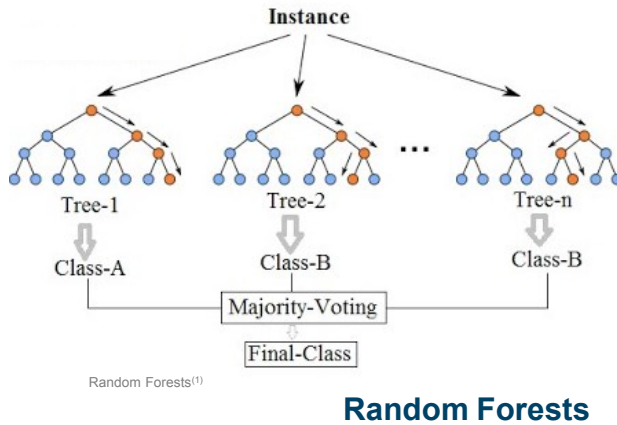
- Random Forests

- Bootstrap Aggregation

- Stacked Aggregation

Random Forests are ensembles of decision trees that leverage the power of a randomized strategy and results in a strong learner. This approach is highly accurate and computationally efficient.

The term "Random Forests" is trademarked by its inventors Leo Breiman and Adele Cutler.

For a given dataset with N observations, each decision tree in the forest is built or trained using a random sample with replacement of size N. This process is referred to as bootstrap aggregation or bagging, for short. Any observations not included in an individual sample is called out of bag (OOB) data. For fairly large N, one would expect about 37% of the observations to be OOB data.

In general, a Random Forests model does not require cross-validation or a separate testing data set to determine an estimate of error, this functionality is part of the Random Forests algorithm. [2]

[1] Image: Random Forests. https://i.ytimg.com/vi/ajTc5y3OqSQ/hqdefault.jpg

[2] Breiman, L. (2001). Random Forests.
https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Method Summary

Method	Type	Goal	Input	Output	Algorithm Examples
Classification	Supervised	Predict class	Binary Nominal	Unordered categorical	Decision tree Logistic regression Naïve bayes Random forests
Regression	Supervised	Predict value Predict probability	Numeric	Numeric Probability	Linear regression Logistic regression
Clustering	Unsupervised	Describe data	Numeric	Grouping by similarity	K-Means K-Medians Expectation-Maximization Hierarchical Clustering
Association	Unsupervised	Describe data	Numeric Categorical	Associated item	Apriori
Ensemble	Supervised	Describe data Predict group Predict value Predict probability	Binary Nominal Numeric	Categorical Numerical Similarity	Random forests Bootstrap aggregation Stacked aggregation

Section 3: Summary

<u>Key points covered in this section:</u>

- Main Goals of Machine Learning
- Various Approaches to Machine Learning
- Applicability of Each Approach

In this module you learned the main goals of machine learning and its various approaches. You also learned how each approach is used and when they are appropriate.

Sections 4

Lab – Predicting Customer Churn

Appendix
Glossary

Glossary

Analytics	The quantifiable informational inputs that use past data to identify possible trends that may provide valuable insight for future action.
Association Rules	Unsupervised learning technique to find similarities in data items based on frequent item sets. It does not predict but is used for data exploration. Most common example is market basket analysis.
Classification	Supervised learning technique that identifies to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.
Clustering	Unsupervised learning technique to find similarities based on the proximity of features in a dataset. It cannot make predictions and is used for data exploration.
Collinearity	Refers to a linear relationship between two or more independent variables (multi-collinearity). When this exists in a data set, it can cause a model to be less accurate.
Correlation	Statistical relationship that involves dependence and is most often used in reference to a linear relationship. Example: Price and Demand.

Glossary (cont.)

Data	Raw objective fact about an event.
Data Science	Interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured.
Dependent Variable	The value of the variable is influenced by other variables. Also referred to as the outcome or target.
Descriptive Statistics	Methods of organizing, summarizing and presenting information about data.
Ensemble	Machine learning technique that combines multiple models (weak learners) to make an overall strong learner. The models may use all the same or different algorithms and the training data may be all the same or resampled. Most common example is Random Forests.
Independent Variable	The value of the variable is not influenced by other variables. These variables may influence dependent variables.

Glossary (cont.)

Inferential Statistics	Methods to determine something about a population from a sample.
Information	Data with interpretation: useful, organized, and structured.
Knowledge	Information with context, it has understanding and meaning.
Kurtosis	A measure of the "tailedness" of the probability distribution of a real-valued random variable.
Labeled Data	Data that has been identified and assigned a label.
Machine Learning	The field of study that gives computers the ability to learn without being explicitly programmed.
Mean	The quotient of the sum of the data points and the number of data points; another name for the average.
Median	When the data are arranged in sorted order, the median is that data point at which 50% of the data points are either less than or greater than that data point; the data point in the middle.

Glossary (cont.)

Mode	The data point that occurs most frequently. There can be more than one mode.
Quasi-Structured Data	Textual data with erratic data formats, can be formatted with effort, tools, and time.
Semi-Structured Data	Textual data with a discernable pattern, enabling parsing. E.g. Self-describing XML with schema.
Semi-Supervised Learning	Category of machine learning that uses few labeled data, and many unlabeled data. There is a teaching assistant, usually involves large amounts of unclassified data with few classified data. The output is a meaningful grouping of data.
Skewness	A measure of the asymmetry a data distribution. A measure of 0 indicates a perfect symmetry.
Structured Data	Defined data type, format, and structure. E.g. Transactional Data.

Glossary (cont.)

Semi-Supervised Learning	Category of machine learning that uses few labeled data, and many unlabeled data. There is a teaching assistant, usually involves large amounts of unclassified data with few classified data. The output is a meaningful grouping of data.
Unlabeled Data	Refers to data that has no clear label or indication of what it is.
Unstructured Data	Data that has no inherent structure and is usually stored in different file types. E.g. PDF, Excel, and JPG.
Unsupervised Learning	Category of machine learning that only uses all unlabeled data. There is no teacher, the correct classes of training data are not known, and the output is a natural meaningful grouping of the data.
Wisdom	Knowledge with insight. Integrated understanding and actionable.