Data Science and Machine Learning with the Watson Data Platform

Carly Kizorek, Dan de Grazia
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

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

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

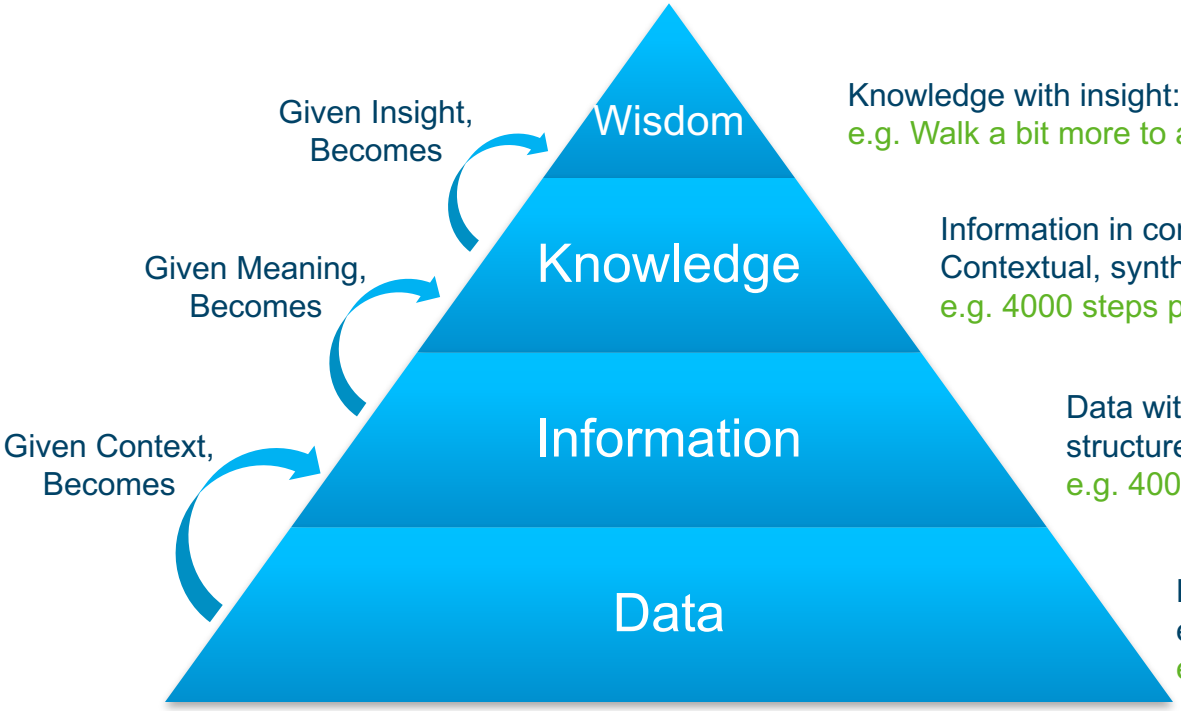
A DEFINITION

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

Pixabay. Analytics Word Cloud (1)

Analytics

DATA'S JOURNEY TO VALUE



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

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

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

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

Charles Sturt University. Data and Knowledge(1)

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

Data Science

A DEFINITION

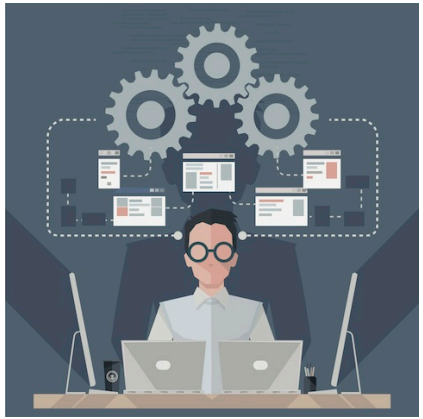
*"**Data science**, also known as data-driven science, is an interdisciplinary field about scientific methods, processes, and systems to **extract knowledge or insights from data** in various forms, either structured or unstructured."*(1)

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

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

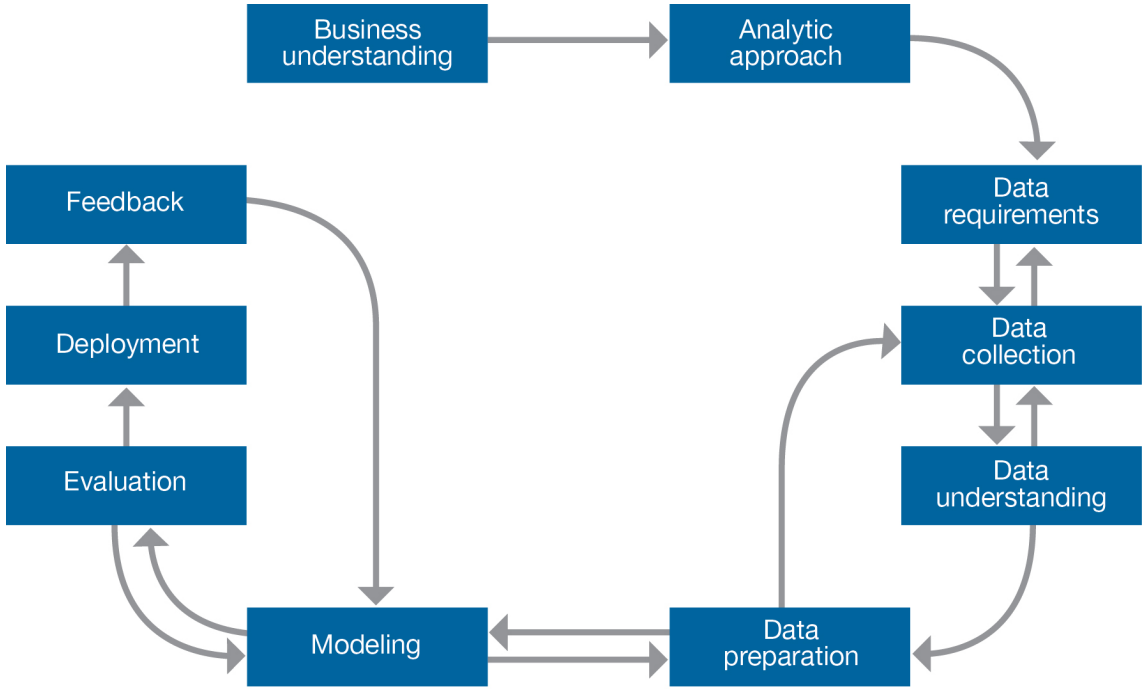
- Communicative & Collaborative
  - Strong verbal and written skills. Must be able to articulate business value and collaborate with others.

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

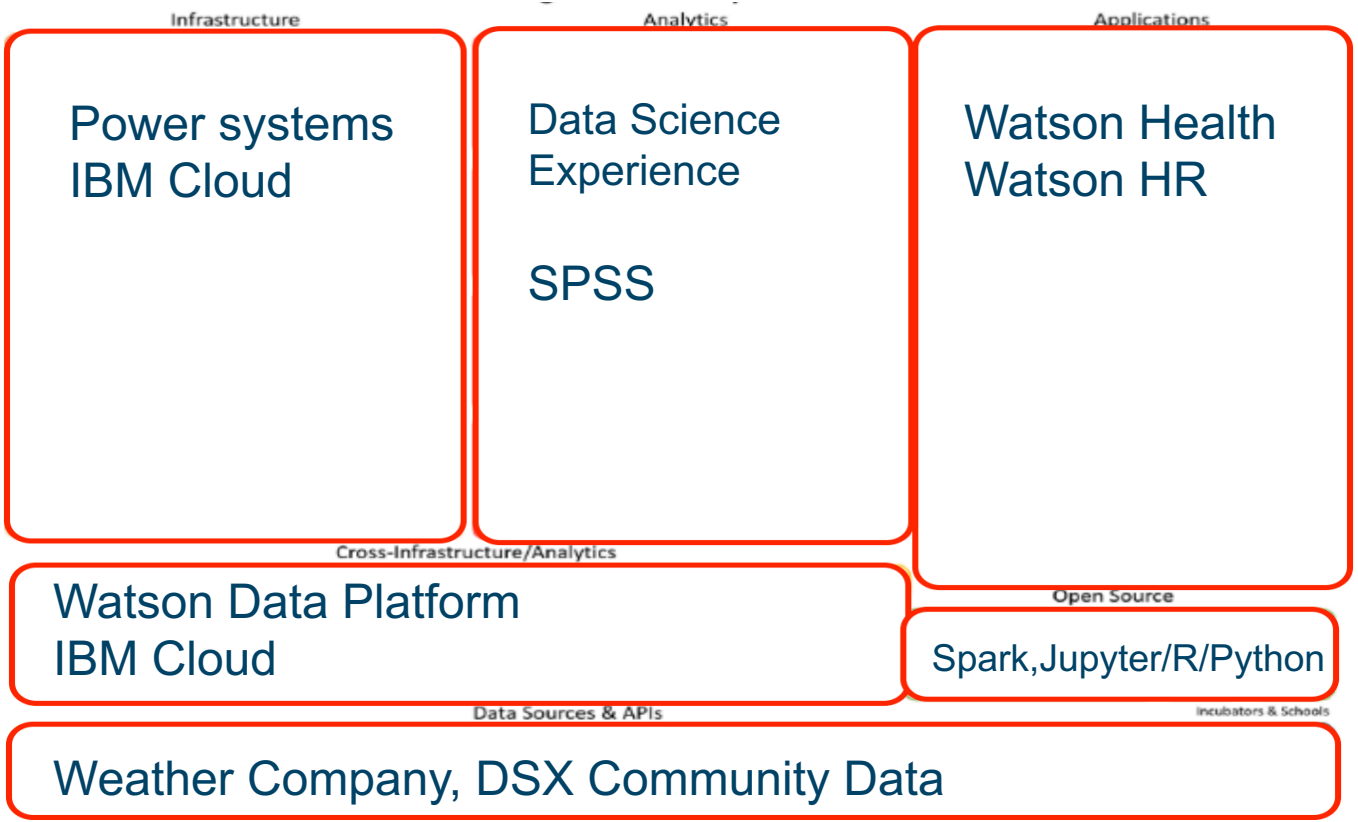
Data Science Methodology

METHODOLOGY DIAGRAM



Tools for Data Science

Tools for Data Science



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

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)

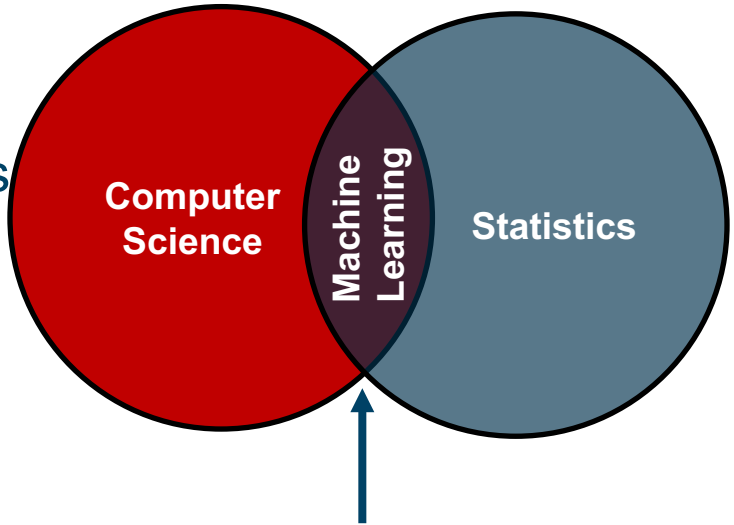
"Machine Learning is the field of study that gives computers the ability to ***learn without being explicitly programmed****."*

Arthur Samuel, Artificial Intelligence Pioneer – IBM Corporation.

Overview of Machine Learning

ML IS THE NATURAL INTERSECTION OF TWO DISCIPLINES

How we build machines that solve problems.



What conclusions can be inferred from data.

How do we get computers to program themselves.

Types of Data in Machine Learning

Labeled

Cat

Hot Wings

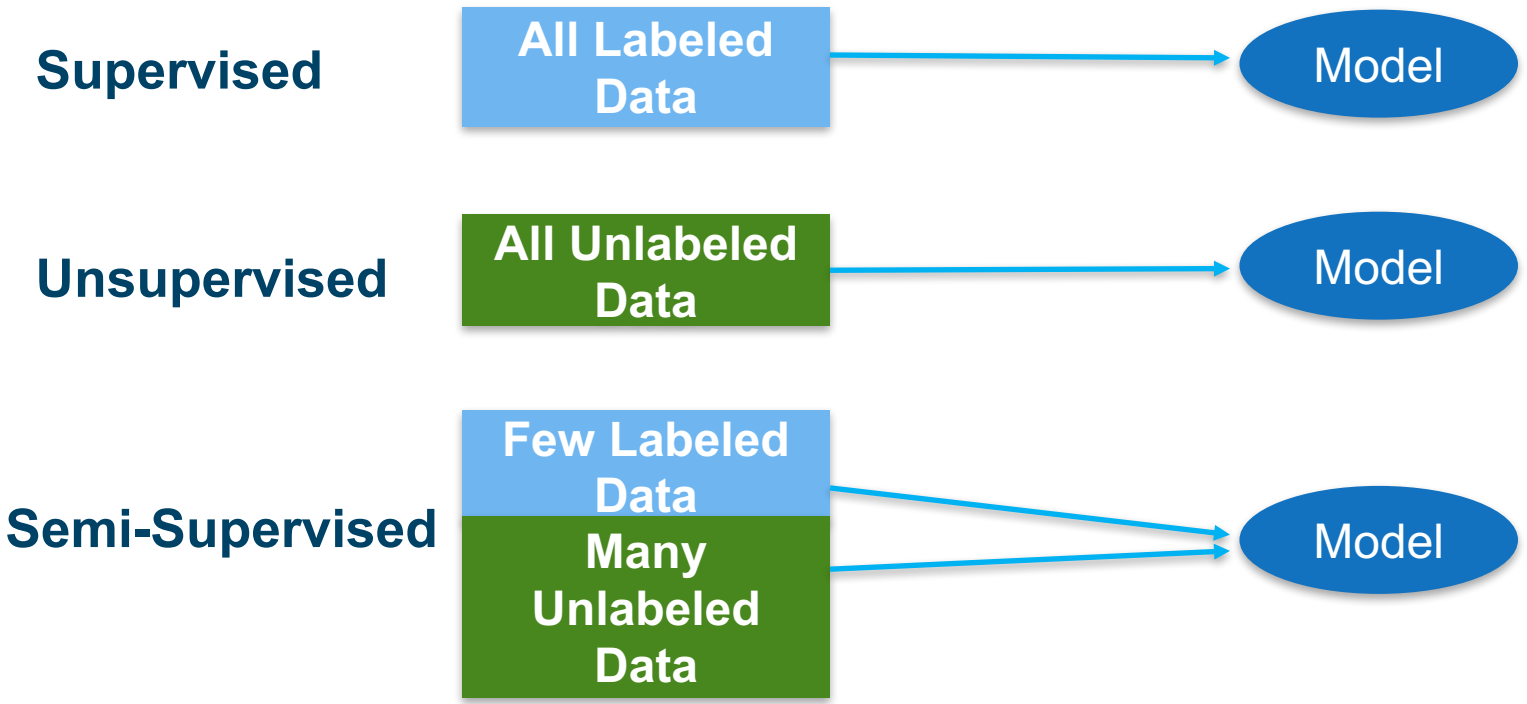
Unlabeled

?

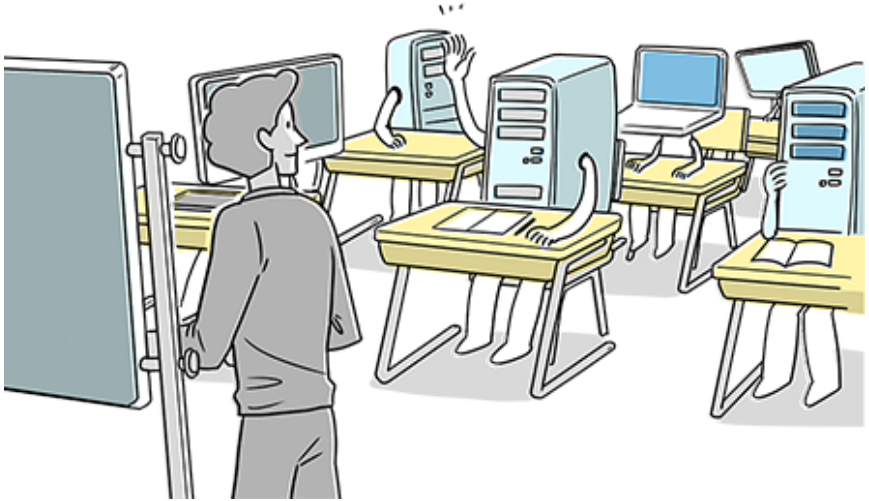
?

?

Main Categories of ML



Supervised Learning



Supervised Learning[2]

- There is a teacher
- Correct classes of training data are known
- Output is a model, or rule set

Supervised Learning

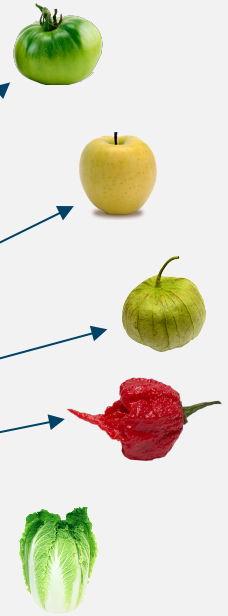
WHAT DO YOU SEE?

Supervised Learning

Features (Input)

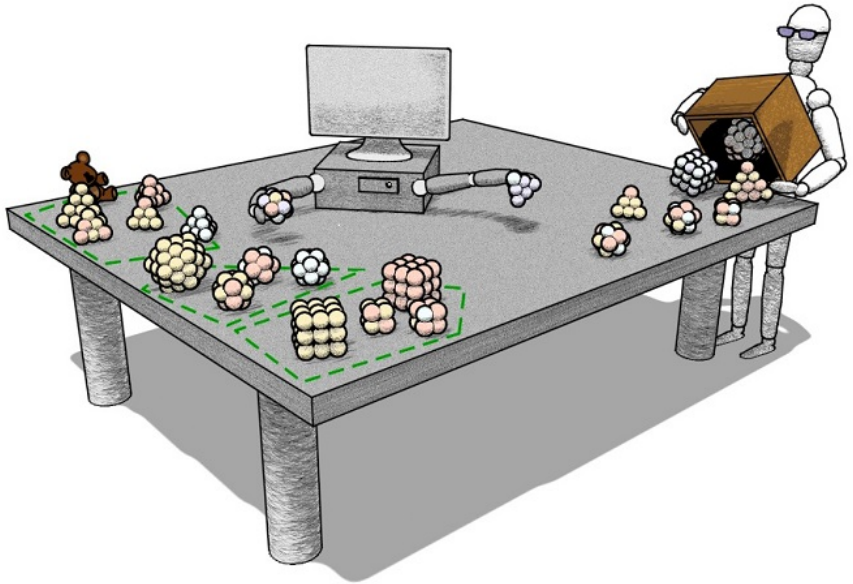
Label (Output)

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)

Unsupervised Learning



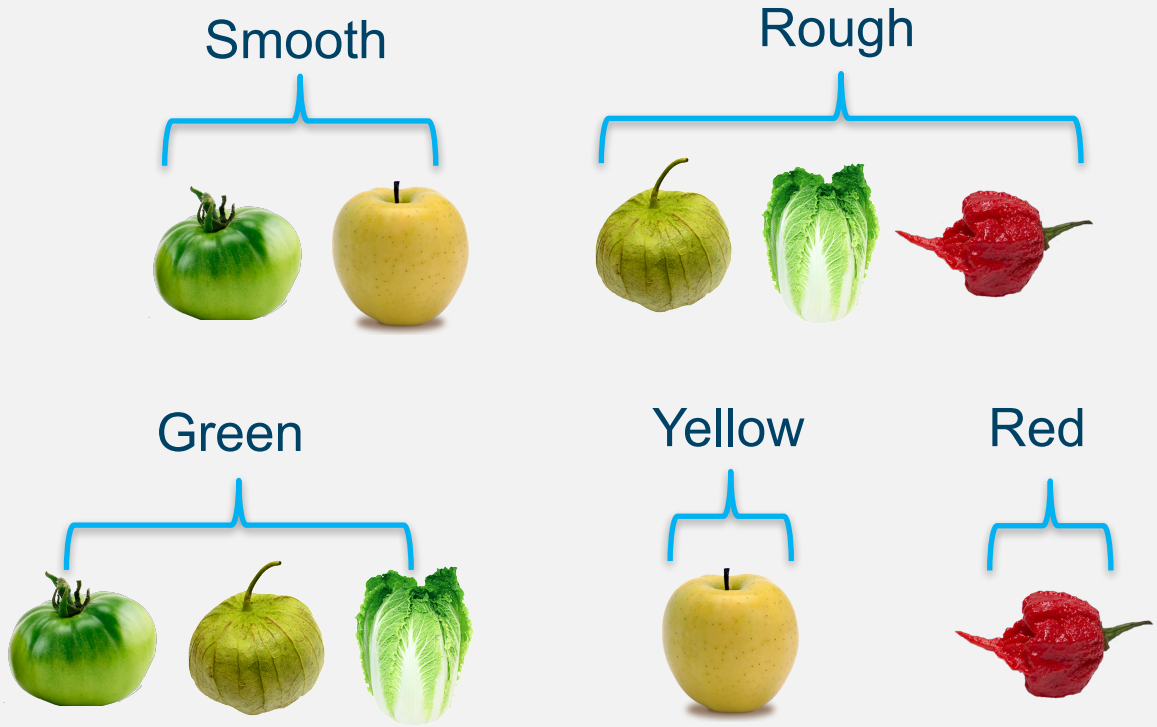
- There is **NO** teacher
- Correct classes of training data are **NOT** known
- Output is natural meaningful groupings

Unsupervised Learning

WHAT DOES A MACHINE SEE?

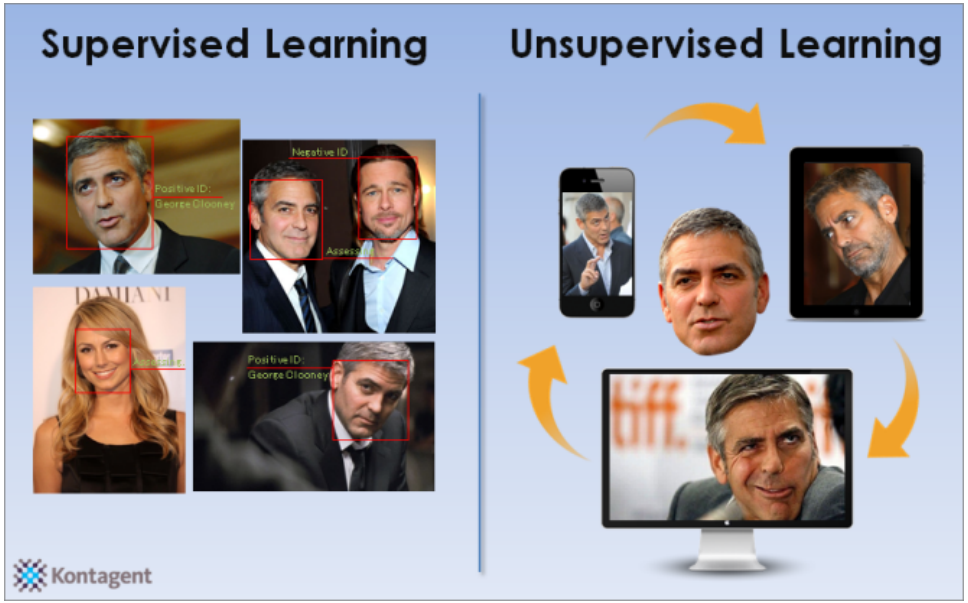
Unsupervised Learning

GROUP AND ASSOCIATE



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

Section 2: Summary

Key points covered in this module:

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

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

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.

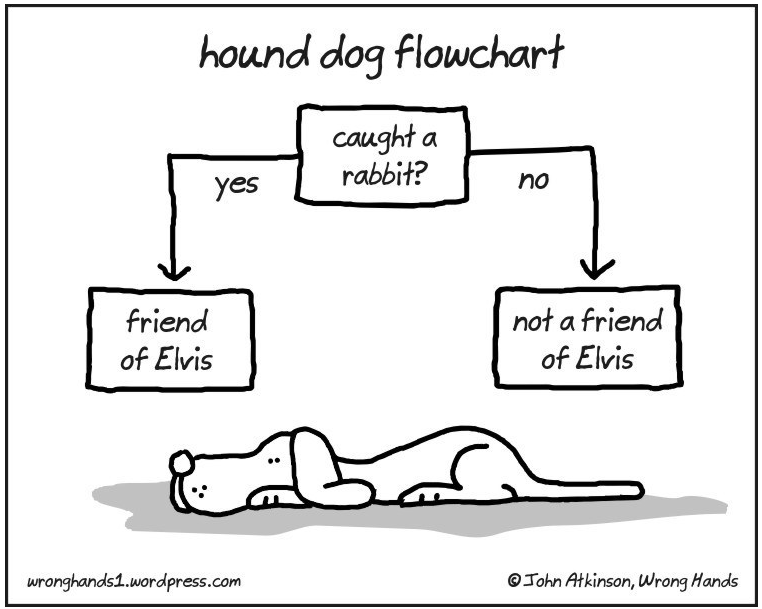
PREDICT

Color	Texture	Shape	Avg Size (in)	Sweet	Object
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
Yellow	Smooth	Oblong	8	Yes	?

Make predictions for a new sample described by its attributes.

Classification

IDENTIFY GROUP MEMBERSHIP

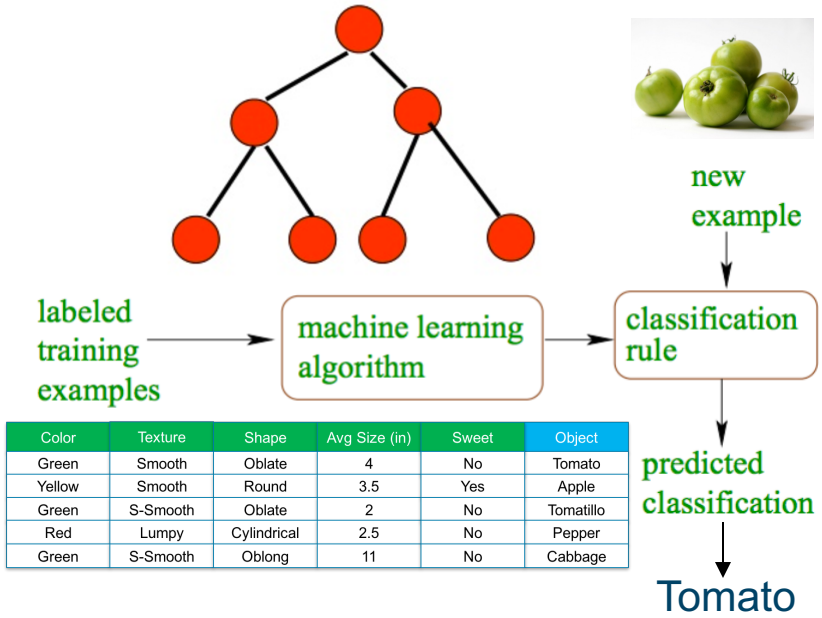


A Friend of Elvis[1]

- Supervised learning
- Predict class from observations
  - “friend of Elvis”
  - “not a friend of Elvis”
- Response variable is categorical and unordered
- Binary and nominal data

Classification

TECHNIQUES



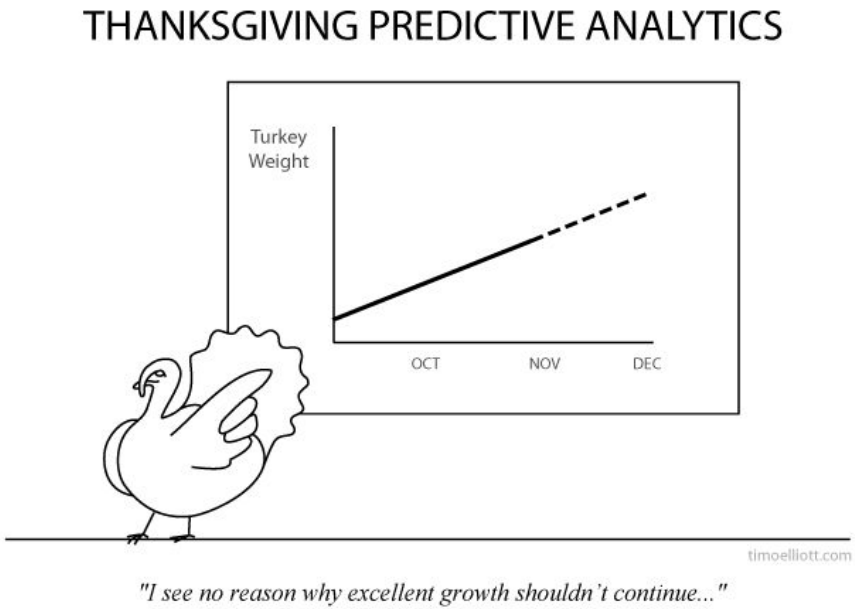
Color	Texture	Shape	Avg Size (in)	Sweet	Object
Green	Smooth	Oblate	4	No	Tomato
Yellow	Smooth	Round	3.5	Yes	Apple
Green	S-Smooth	Oblate	2	No	Tomatillo
Red	Lumpy	Cylindrical	2.5	No	Pepper
Green	S-Smooth	Oblong	11	No	Cabbage

Decision Trees

- Decision Trees
- Logistical Regression
- Naïve Bayes
- Random Forests

Regression

ESTIMATE OR PREDICT A RESPONSE

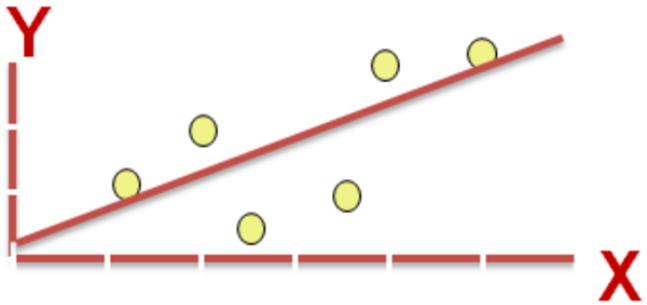


Thanksgiving Turkey. (1)

- Supervised learning
- Predict value from observations
- Response variable is numeric value, or probability of class

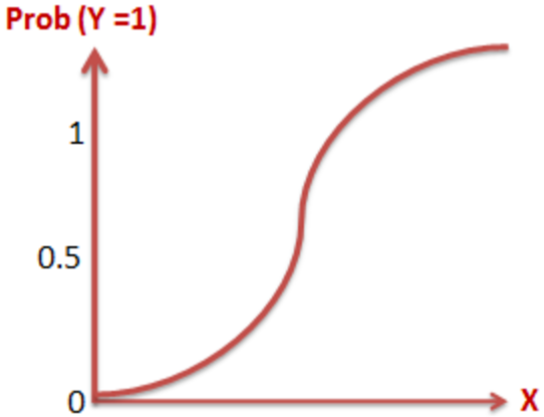
Regression

TECHNIQUES



Linear Regression

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



Logistic Regression

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

Flavors of Regresssion(1)

Clustering

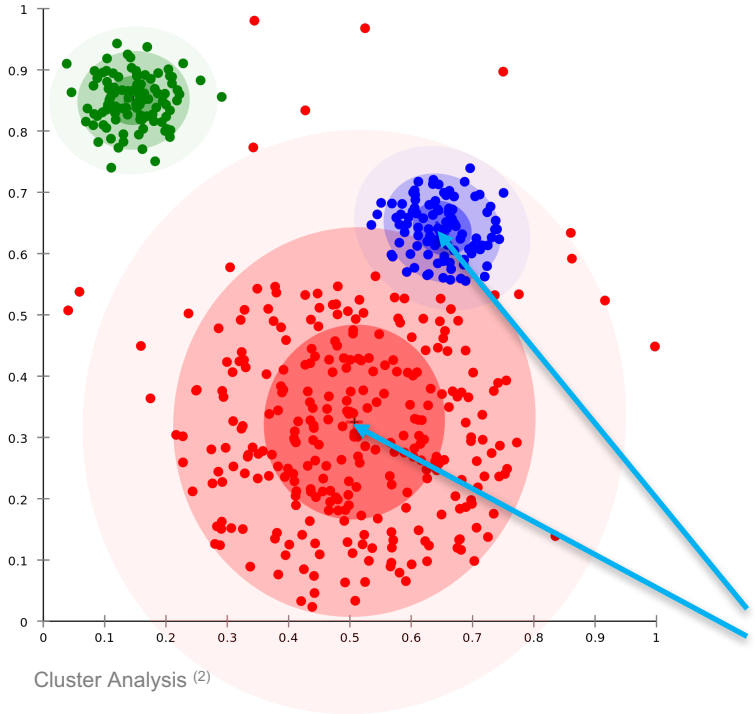
ORGANIZE DATA INTO GROUPS OF MAXIMUM COMMONALITY

Cluster Analysis [1]

- Unsupervised learning
- Describe
- Used for exploratory mining
- Output is meaningful grouping based on similarity

Clustering

TECHNIQUES



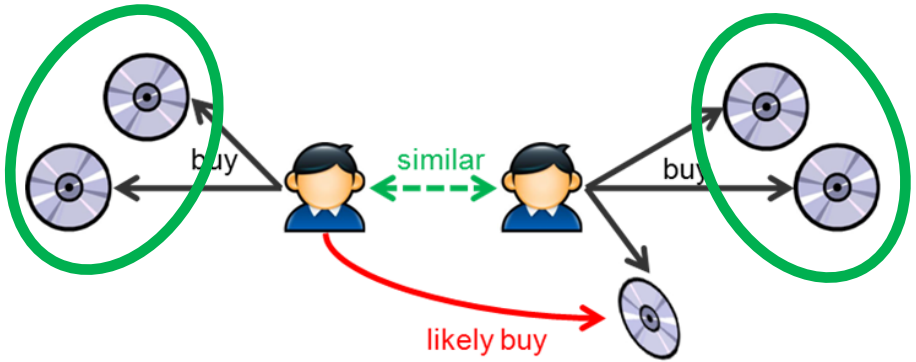
K-Means

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

To which centroid is a point associated?

Associations

DISCOVER STRONG RULES ALONG SOME MEASURE



Association Rules [1]

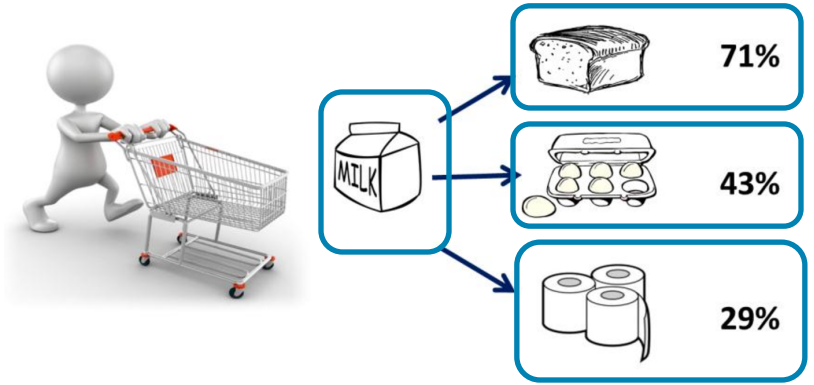
- Supports labeled and unlabeled data
- Describes
- Data mining in large databases
- First introduced to find relationships in POS systems: Market Baskets

Associations

TECHNIQUES

Transaction 1	🍎 🍺 ⚪ 🍗
Transaction 2	🍎 🍺 ⚪
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 ⚪ 🍗
Transaction 6	🍼 🍺 ⚪
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

Apriori

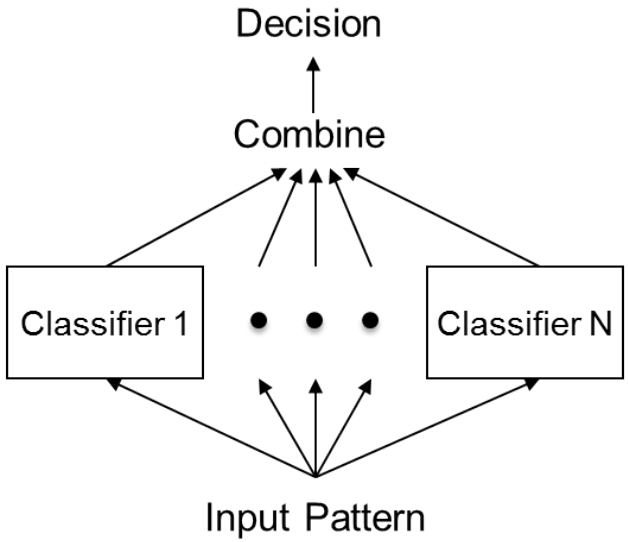


Of transactions that included milk:

- 71% included bread
- 43% included eggs
- 29% included toilet paper

Ensembles

COMBINING MULTIPLE LEARNERS



Ensemble Methods[1]

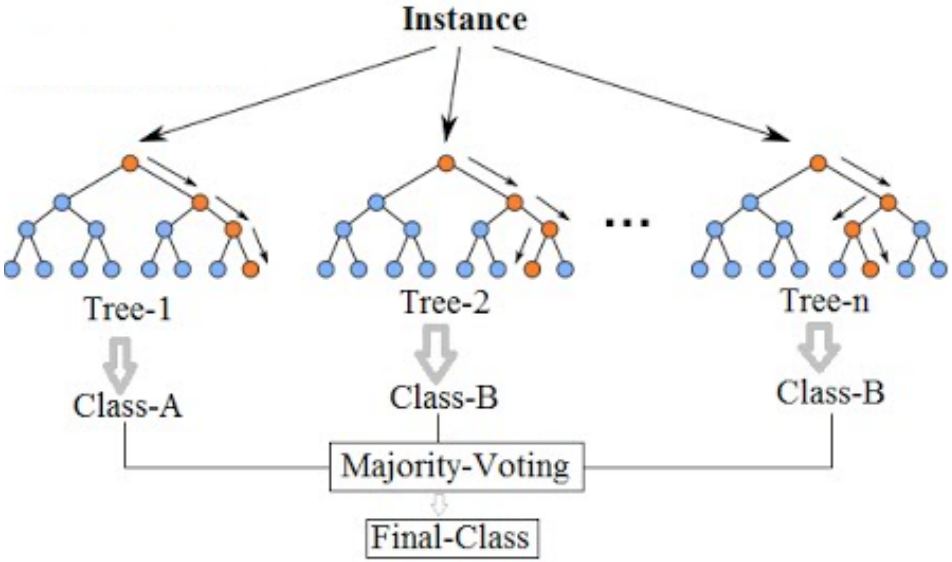
- Supervised learning
- Weak learners independently trained
- Combined predictions produce strong learner

NETFLIX

Winning the Netflix Prize [2]

Ensembles

TECHNIQUES



Random Forests[1]

Random Forests

- Random Forests
- Bootstrap Aggregation
- Stacked Aggregation

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

Key points covered in this section:

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

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.