Statistical Data Mining
MATH 4720
Lecture 3

Gaurav Gupta

GEH A401

Data Understanding

1. Data Acquisition
2. Data Extraction
3. Data Description
4. Data Assessment
5. Data Profiling
6. Data Cleansing
7. Data Transformation
8. Data Imputation
9. Data Weighting and Balancing
10. Data Filtering and Smoothing
11. Data Abstraction
12. Data Reduction
13. Data Sampling
14. Data Discretization
15. Data Derivation

10. Data Abstraction

- Data preparation for data mining may include some very complex rearrangements of your data set.

- Data abstraction is **the reduction of a particular body of data to a simplified representation of the whole.**

- It allows us to create our own user defined data types (using the class construct) and then define variables (i.e objects) of those new data types.

Abstractions can be classified into four groups (Lavrac et al., 2000):

1. **Qualitative abstraction:** A numeric expression is mapped to a qualitative expression.

   Example: Compared to that of others, customers with ages between 13 and 19 could be abstracted as a value of 1 to a variable “teenager,” while others are abstracted to a value of 0.

2. **Generalization abstraction:** An instance of an occurrence is mapped to its class.

   Example: Compared to non-Asian, listings of “Chinese,” “Japanese,” and “Korean” in the Race variable could be abstracted to 1 in the Asian variable, while others are abstracted to a value of 0.

3. Definitional abstraction: An instance in which one data element from one conceptual category is mapped to its counterpart in another conceptual category.

Example: when combining data sets from different sources for an analysis of customer demand among African-Americans, you might want to map “Caucasian” in a demographic data set and “White Anglo-Saxon Protestant” in a sociological data set to a separate variable of “Non-black.”

4. Temporal abstraction: It is not commonly used.

Without using time-series algorithms (e.g., ARIMA), you can create temporal abstractions and model on them. Some tools (e.g., KXEN) provide a facility for creating these variables. Sometimes, they are called lag variables because the response modeled lags behind the causes of the response. The DVD contains a Perl program you can modify; it will permit you to create temporal abstractions from your time-series data.

12. Data Reduction

Data reduction includes three general processes:

- Reduction of dimensionality (number of variables)
- Reduction of cases (records)—Data Sampling
- Discretization of values

13. Data Sampling

In data mining, data sampling serves four purposes:

1. **It can reduce the number of data cases submitted to the modeling algorithm.**

Example: Random Sampling

2. **It can help you select only those cases in which the response patterns are relatively homogeneous**.

Example: Partitioning

After partitioning, you can randomly select cases within each defined partition. Such a sampling is called **stratified random sampling.** The partitions are the "strata" that are sampled separately.

3. **It can help you balance the occurrence of rare events for analysis by machine learning tools.**

- An unbalanced data set is one in which one category of the target variable is relatively rare compared to the other ones. Balancing the data set involves sampling the rare categories more than average (oversampling) or sampling the common categories less often (undersampling)

- Finally, simple random sampling can be used to divide the data set into three data sets for analysis:

A. **Training set:** These cases are randomly selected for use in training the model.

B. **Testing set:** These cases are used to assess the predictability of the model, before refining the model or adding model enhancements.

C. **Validation set:** These cases are used to test the final performance of the model after all modeling is done

14. Data Discretization

You can convert a continuous numeric variable into a series of categories by assigned subranges of the value range to a group of new variables.

- For example, a variable ranging from 1–100 could be discretized (converted into discrete values) by dividing the range in four subranges (bins): 0–25, 26–50, 51–75, and 76–100.
- In the binning process, each value in the range of a variable is replaced by a bin number. Many data mining packages have binning facilities to create these subranges automatically.
- This process reduces noise in the data.

- **Example:** Credit scores are created using bins, in which bin boundaries are tuned and engineered to maximize the predictive power of the credit scoring model.

Discretize by Binning

(RapidMiner Studio Core)

RapidMiner Studio is an efficient solution for the analysts who need to visualize and understand complex data.

https://docs.rapidminer.com/latest/studio/operators/cleansing/binning/discretize_by_bins.html

15. Data Derivation

Assignment or Derivation of the Target Variable

- The target variable can be selected from among the existing variables in the data set.

Example 1: Model of equipment failure

Target variable: presence or absence of a failure date in the data record.

.

Example 2:

Customer attrition model:

Target variable: Month in which customer phone usage declined at least 70% over the previous two billing periods.

This variable was derived by comparing the usage of all customers for each month in the time-series data set with the usage two billing periods in the past.

The billing period of this cellular phone company was every two months, so the usage four months previous to each month was used as the value of comparison.

Most often, the target variable must be derived following some heuristic (logical rule). The simplest version of an attrition target variable in that cellular phone company would have been to identify the month in which the service was discontinued. Insurance companies define attrition in that manner also.

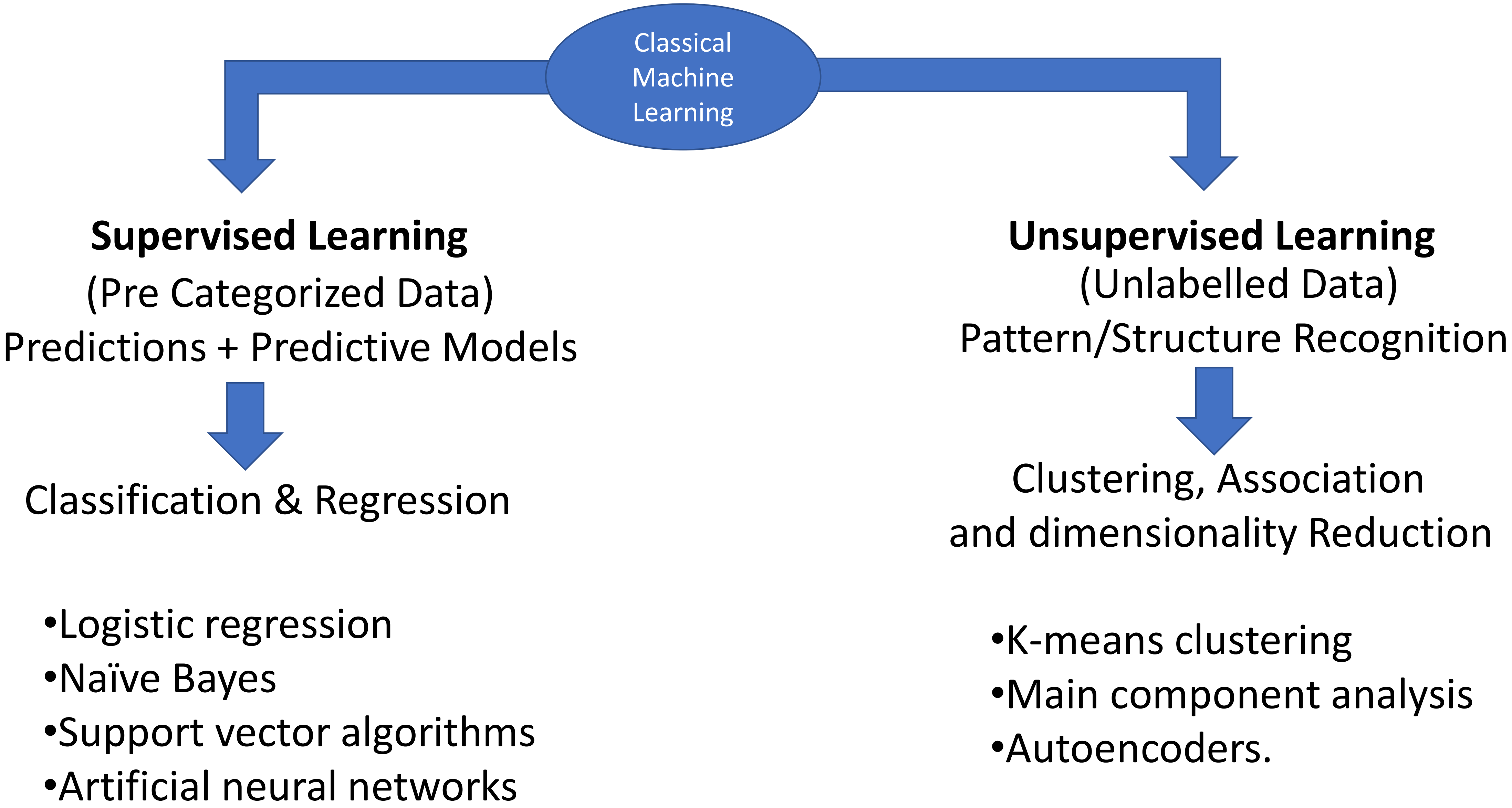
Derivation of New Predictor Variables

- New variables can be created from a combination of existing variables.
- **Example: Given:** latitude and longitude
- **New variable:** Distance to Store
- Equations for calculating distance on the surface of the earth between two pairs of latitude-longitude coordinates.

$$= \mathrm{ACOS}(\mathrm{SIN}(\mathrm{Lat1}) * \mathrm{SIN}(\mathrm{Lat2}) + \mathrm{COS}(\mathrm{Lat1}) * \mathrm{COS}(\mathrm{Lat2}) * \\ \mathrm{COS}(\mathrm{Lon2} - \mathrm{Lon1})) * 3934.31$$

Output is the distance in miles between the two points

Supervised and Unsupervised Machine Learning



Classical Machine Learning

Supervised Learning

(Pre Categorized Data)

Predictions + Predictive Models

Classification & Regression

- Logistic regression
- Naïve Bayes
- Support vector algorithms
- Artificial neural networks

Unsupervised Learning

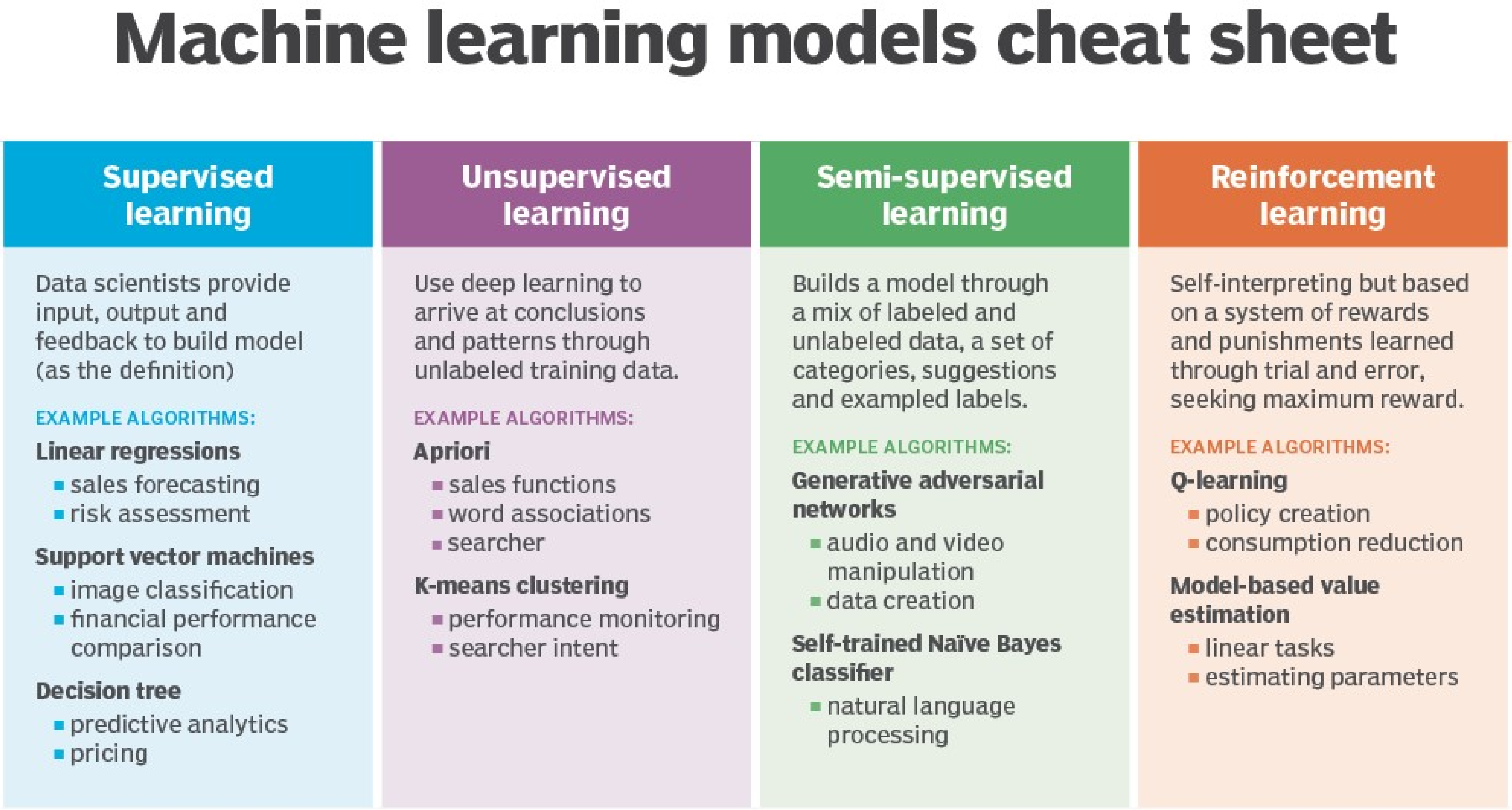
(Unlabelled Data)

Pattern/Structure Recognition

Clustering, Association and dimensionality Reduction

- K-means clustering
- Main component analysis
- Autoencoders.

- **Supervised Learning:** Develop predictive model based on both input and output data.
- **Unsupervised Learning:** Group and interpret data based only on input



Machine learning models cheat sheet

Supervised learning

Data scientists provide input, output and feedback to build model (as the definition)

EXAMPLE ALGORITHMS:

Linear regressions

- sales forecasting
- risk assessment

Support vector machines

- image classification
- financial performance comparison

Decision tree

- predictive analytics
- pricing

Unsupervised learning

Use deep learning to arrive at conclusions and patterns through unlabeled training data.

EXAMPLE ALGORITHMS:

Apriori

- sales functions
- word associations
- searcher

K-means clustering

- performance monitoring
- searcher intent

Semi-supervised learning

Builds a model through a mix of labeled and unlabeled data, a set of categories, suggestions and exampled labels.

EXAMPLE ALGORITHMS:

Generative adversarial networks

- audio and video manipulation
- data creation

Self-trained Naïve Bayes classifier

- natural language processing

Reinforcement learning

Self-interpreting but based on a system of rewards and punishments learned through trial and error, seeking maximum reward.

EXAMPLE ALGORITHMS:

Q-learning

- policy creation
- consumption reduction

Model-based value estimation

- linear tasks
- estimating parameters

https://searchenterpriseai.techtarget.com/definition/unsupervised-learning

What is supervised learning?

Supervised learning is a machine learning approach that’s defined by its use of labeled datasets. These datasets are designed to train or “supervise” algorithms into classifying data or predicting outcomes accurately.

Using labeled inputs and outputs, the model can measure its accuracy and learn over time.

Classification

- Classification is the operation of separating various entities into several classes.
- These Classes can be defined by business rules, class boundaries, or some mathematical function.
- In the real world, supervised learning algorithms can be used to classify spam in a separate folder from your inbox.
- **Linear classifiers, support vector machines, decision trees and random forest** are all common types of classification algorithms.

Regression

- It is another type of **supervised learning** method that uses an algorithm to understand the relationship between dependent and independent variables.
- Regression models are helpful for predicting numerical values based on different data points, such as sales revenue projections for a given business.
- Some popular regression algorithms are **linear regression, logistic regression and polynomial regression.**

What is unsupervised learning?

- Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets.
- These algorithms discover hidden patterns in data without the need for human intervention (hence, they are “unsupervised”).
- Unsupervised learning models are used for three main tasks: clustering, association and dimensionality reduction:

Clustering

Clustering is a data mining technique for grouping unlabeled data based on their similarities or differences.

For example, K-means clustering algorithms assign similar data points into groups, where the K value represents the size of the grouping and granularity.

This technique is helpful for market segmentation, image compression, etc.

Association

Association is another type of unsupervised learning method that uses different rules to find relationships between variables in a given dataset.

These methods are frequently used for market basket analysis and recommendation engines, along the lines of “Customers Who Bought This Item Also Bought” recommendations.

Dimensionality reduction

Dimensionality reduction is a learning technique used when the number of features (or dimensions) in a given dataset is too high. It reduces the number of data inputs to a manageable size while also preserving the data integrity. Often, this technique is used in the preprocessing data stage, such as when autoencoders remove noise from visual data to improve picture quality.

The main difference between supervised and unsupervised learning: Labeled data

- Supervised Learning uses labeled input and output data, while an unsupervised learning algorithm does not.
- In supervised learning, the algorithm “learns” from the training dataset by iteratively making predictions on the data and **adjusting for the correct answer.**
- While supervised learning models tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately.

Example

- **Supervised learning** model can predict how long your commute will be based on the time of day, weather conditions and so on. But first, you’ll have to train it to know that rainy weather extends the driving time.
- **Unsupervised learning** model can identify that online shoppers often purchase groups of products at the same time. However, a data analyst would need to validate that it makes sense for a recommendation engine to group baby clothes with an order of diapers, applesauce and sippy cups.

Goals:

- In supervised learning, the goal is to predict outcomes for new data. You know up front the type of results to expect.

- With an unsupervised learning algorithm, the goal is to get insights from large volumes of new data. The machine learning itself determines what is different or interesting from the dataset

Applications:

- **Applications**: Supervised learning models are ideal for spam detection, sentiment analysis, weather forecasting and pricing predictions, among other things.
- In contrast, unsupervised learning is a great fit for anomaly detection, recommendation engines, customer personas and medical imaging.

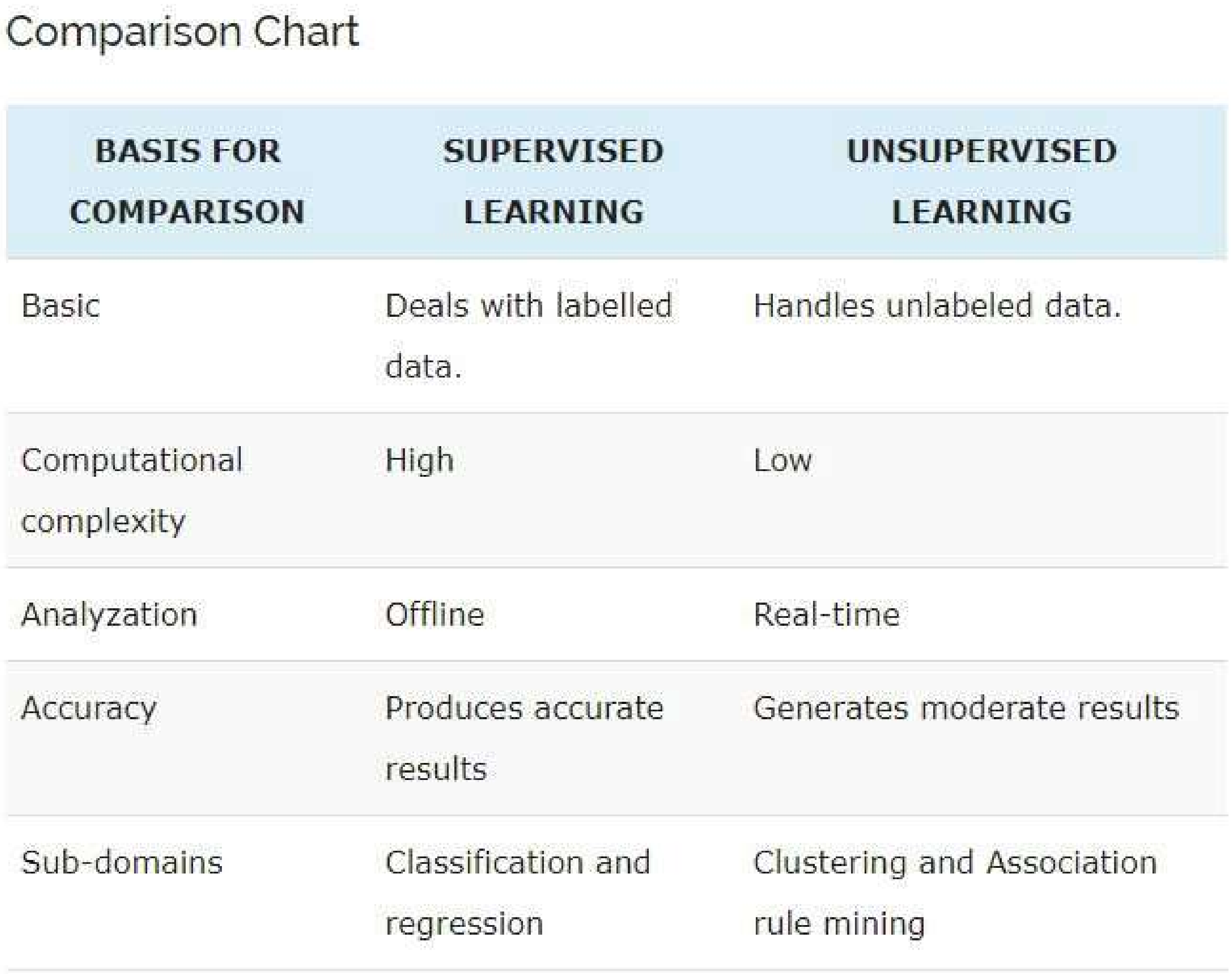
•

Complexity:

- **Complexity:** Supervised learning is a simple method for machine learning, typically calculated through the use of programs like R or Python.
- In unsupervised learning, you need powerful tools for working with large amounts of unclassified data.
- Unsupervised learning models are computationally complex because they need a large training set to produce intended outcomes.

Drawbacks:

Drawbacks: Supervised learning models can be time-consuming to train, and the labels for input and output variables require expertise. Meanwhile, unsupervised learning methods can have wildly inaccurate results unless you have human intervention to validate the output variables.



Comparison Chart

BASIS FOR COMPARISON	SUPERVISED LEARNING	UNSUPERVISED LEARNING
Basic	Deals with labelled data.	Handles unlabeled data.
Computational complexity	High	Low
Analyzation	Offline	Real-time
Accuracy	Produces accurate results	Generates moderate results
Sub-domains	Classification and regression	Clustering and Association rule mining

Which is best for you?

- Choosing the right approach for your situation depends on how your data scientists assess the structure and volume of your data, as well as the use case. To make your decision, be sure to do the following:
- **Evaluate your input data:** Is it labeled or unlabeled data? Do you have experts that can support additional labeling?
- **Define your goals:** Do you have a recurring, well-defined problem to solve? Or will the algorithm need to predict new problems?
- **Review your options for algorithms:** Are there algorithms with the same dimensionality you need (number of features, attributes or characteristics)? Can they support your data volume and structure?

Classifying big data

- Classifying big data can be a real challenge in supervised learning, but the results are highly accurate and trustworthy.
- In contrast, unsupervised learning can handle large volumes of data in real time. But, there’s a lack of transparency into how data is clustered and a higher risk of inaccurate results. This is where semi-supervised learning comes in.

Semi-supervised learning: The best of both worlds

- **Semi-supervised learning** is a happy medium, where you use a training dataset with both labeled and unlabeled data. It’s particularly useful when it’s difficult to extract relevant features from data — and when you have a high volume of data.
- Semi-supervised learning is ideal for medical images, where a small amount of training data can lead to a significant improvement in accuracy.
- For example, a radiologist can label a small subset of CT scans for tumors or diseases so the machine can more accurately predict which patients might require more medical attention.

Thanks