Types of ML

Types of Machine Learning Techniques

- The machine learning algorithms which we will be covering are
  - Supervised learning algorithms
  - Unsupervised learning algorithms

Models for Supervised Learning

- When the response variable is numerical, predictive modeling is called Regression.

- When the response variable is nominal / categorical, predictive modeling is called Classification. The values of the response variable can be considered as "class labels" in this case.

Examples: Supervised Learning

- Naïve Bayes

- K-NN

- Decision Trees

- Regression Models

- Neural Nets

- Support Vector Machines

Partitioning in Supervised Learning

- In Supervised Learning, we partition the data

- We typically deal with two or three partitions:
    - a training set,
    - a validation set,
    - and sometimes an additional test set.

Training Partition

- Typically the largest partition
- Contains the data used to build the various models we are examining
- Generally used to develop multiple models.

Validation Partition

- Used to assess the performance of each model so that you can compare models and pick the best one.

- This partition is used for internally verifying the performance of the models

- Important for measuring the goodness of fit

Test Partition

- Used if we need to assess the performance of the chosen model with new data

- Used to overcome the overfitting problem

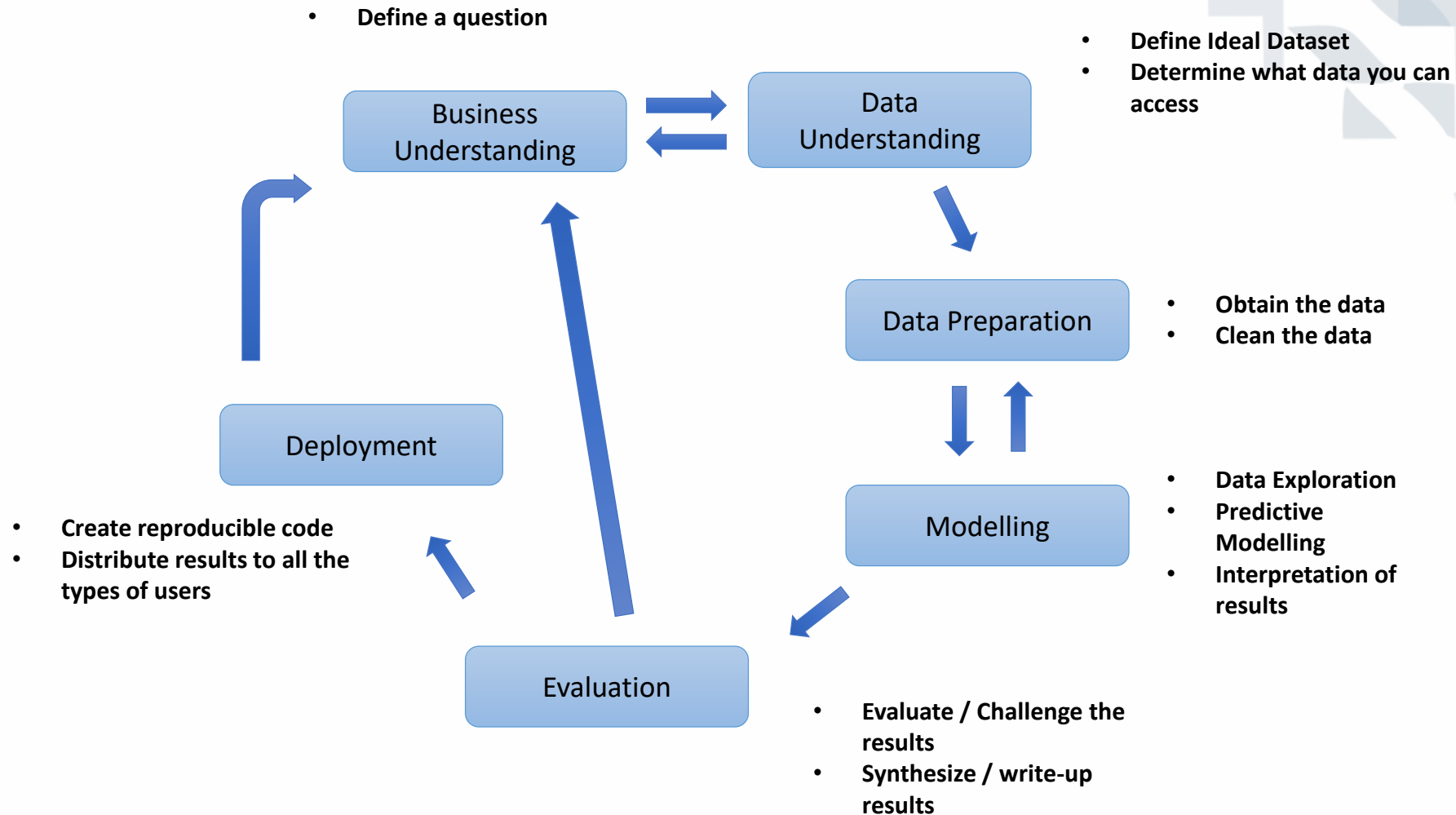# Supervised Learning Process with 2 partitions
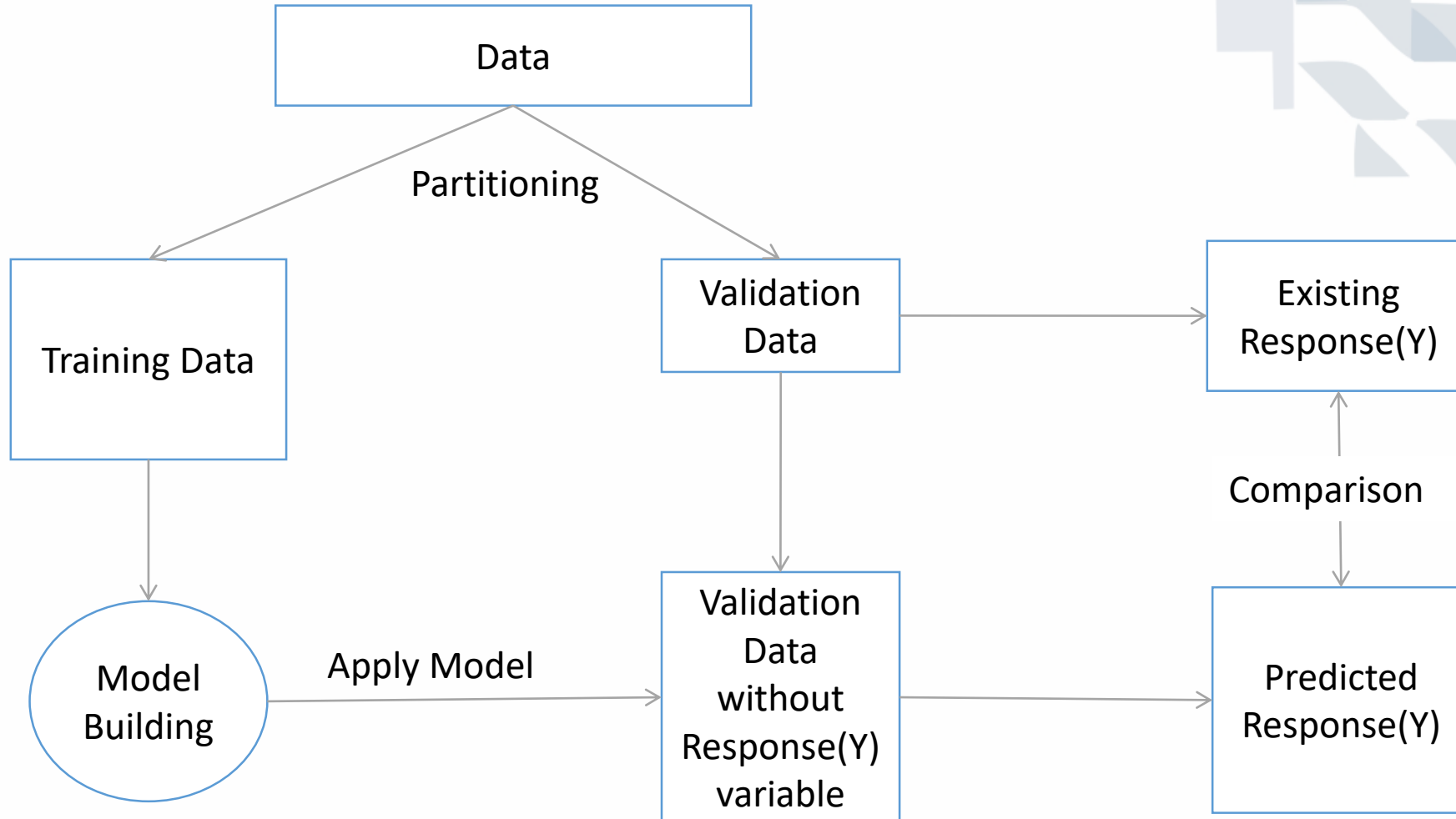
Unsupervised Learning

- Unsupervised learning algorithms are those used where there is no outcome variable to predict or classify.

- Association rules, data reduction methods, and clustering techniques are all unsupervised learning methods.

Examples of Unsupervised Learning

- Clustering Techniques
  - Hierarchical
  - K-means
  - K-medoids
- Principal Component Analysis
- Association Rules

Analytics Lifecycle (CRISP-DM)

- **Define a question**

- **Define Ideal Dataset**
- **Determine what data you can access**

```
Business          →          Data
Understanding     ←          Understanding
```

- **Obtain the data**
- **Clean the data**

Data Preparation

Deployment

Modelling

- **Data Exploration**
- **Predictive Modelling**
- **Interpretation of results**

- **Create reproducible code**
- **Distribute results to all the types of users**

Evaluation

- **Evaluate / Challenge the results**
- **Synthesize / write-up results**

Sane's
STATS
Academy of Statistics

Technologies for ML

Desktop Software

- Click and Drag (Menu Driven)
    - ➢ KNIME
    - ➢ RapidMiner
    - ➢ SAS Enterprise Miner
    - ➢ IBM SPSS Modeller

Programming Languages

- **R**
- **Python**
- Julia
- Scala

- An open source project
- Fast on desktop with small sized data
- Add-ins (packages) available for every statistical/ML algorithm in the world
- Has been used since last 2 decades for statistical computing by statistical professionals community
- There are good IDEs available like RStudio, RTVS, R Commander, Tinn-R, STATET(Eclipse plug-in) etc.
- Among IDEs R Studio is most known
- Provides a scope for implementing our own algorithms being an open source language

- An open source project

- Fast on desktop with small sized data

- Add-ins (packages) available for every statistical/ML algorithm in the world

- The statistical aspects of Python have been developed recently

- There are good IDEs available like Spyder(Anaconda Installation), PyCharm etc.

- Provides a scope for implementing our own algorithms being an open source language

Cloud-Based Platform

- Amazon Web Services

- Microsoft Azure

- Google Cloud AI

Large Scale Data Processing Libraries

- Libraries are such kind of modules which are language independent.
- Using libraries, one can code in R / Python / Java
- Well known libraries for ML are
  - Apache Spark
  - h2o (by h2o.ai)
  - TensorFlow (by Google)
  - Theano (by University of Montreal)
  - CNTK (by Microsoft)
- All of the above provide support for GPU-based operations for algorithms in Deep Learning
- The superb feature which these libraries provide is the fast speed that too at relatively low cost.