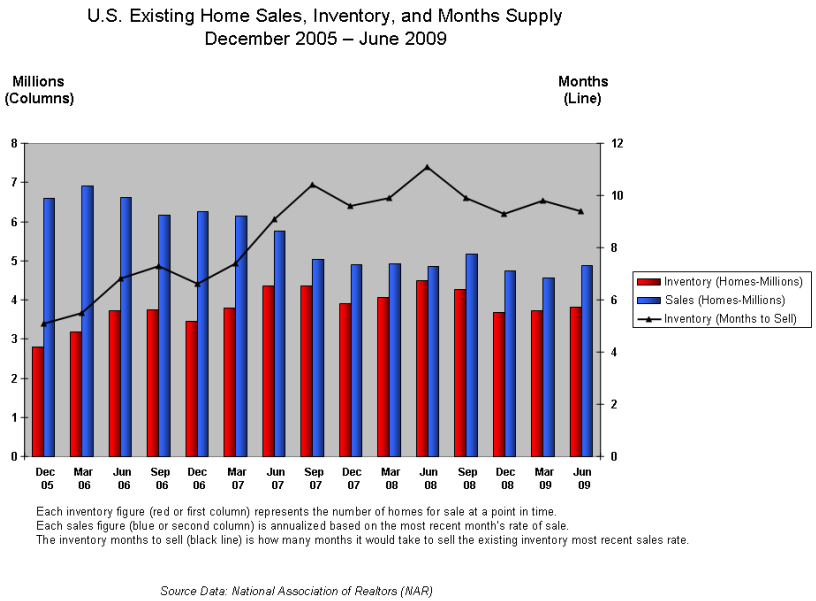
Data Mining Overview

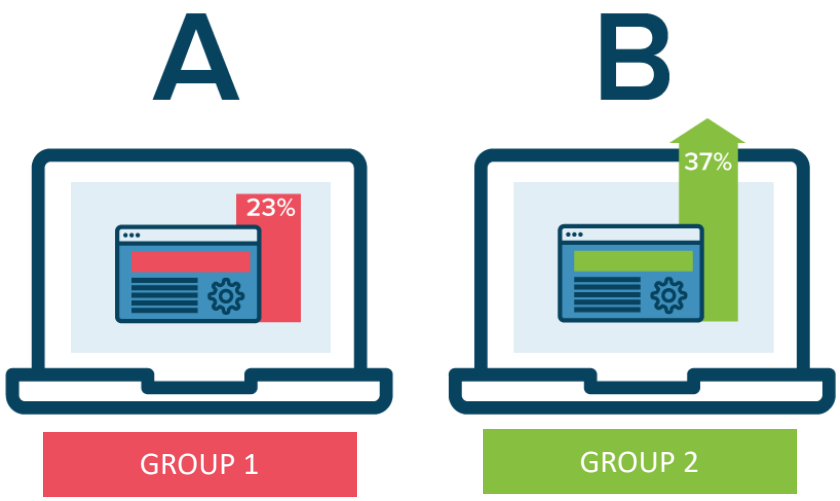
Descriptive Analytics

- Aims to answer *"What happened?"*
- Goal is to provide **insight** into the past
- Uses data aggregation and data mining to make past data **interpretable**
- The vast majority of the statistics we use fall into this category



Diagnostic Analytics

- Aims to answer *"Why did it happen?"*
- Goal is to obtain in-depth insight to a particular problem
- Uses similar techniques as descriptive analytics, but involves measuring historical data against other data
- Many *comparison* studies fall into this category



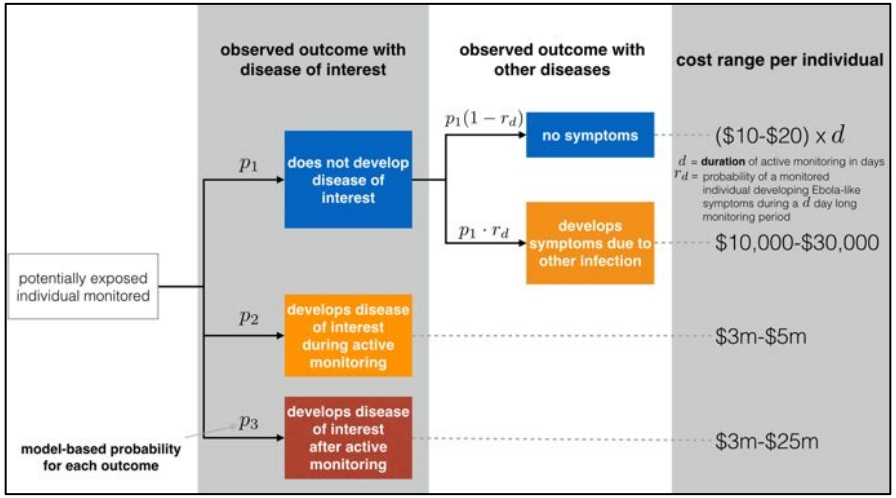
But…be very careful about drawing ANY conclusions on cause-and-effect relationships!

Predictive Analytics

- Aims to answer *"What will happen?"*
- Goal is to estimate the **likelihood** of a future outcome
  - Note that no statistical algorithm can "predict" the future with 100% certainty
- Uses statistical models and forecast techniques to identify trends in past data and infer the future

Prescriptive Analytics

- Aims to answer *"What should we do?"*
- Uses optimization and simulation algorithms to evaluate possible outcomes
- Goal is to provide guidance for future actions to obtain specified outcome
- Relatively new field that evaluates "multiple futures" so organizations can assess a number of possible outcomes based on actions

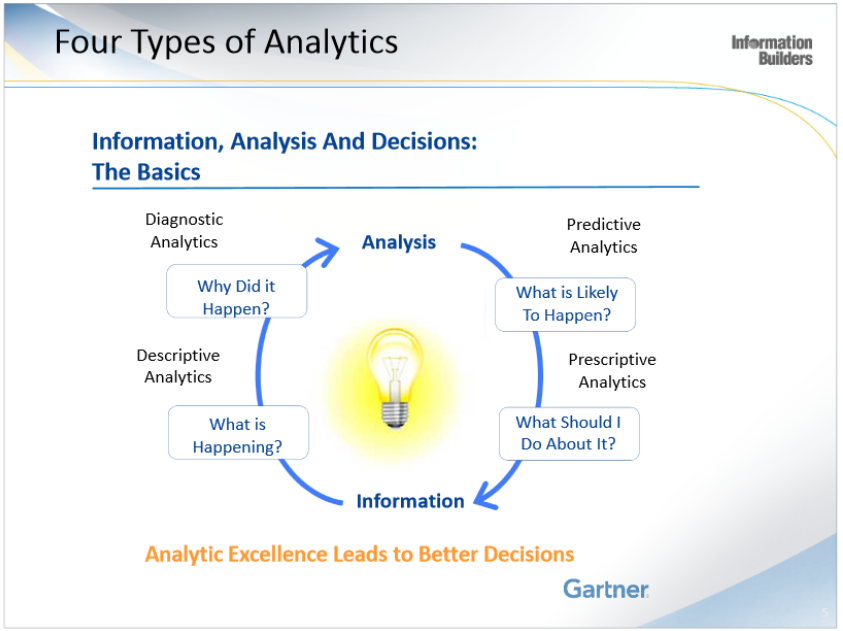


Reich, N. G., Lessler, J., Varma, J. K. & Vora, N. M. Quantifying the Risk and Cost of Active Monitoring for Infectious Diseases. Scientific Reports **8**, 1093 (2018).

Types of Analytics

- No single type of analytics is better than another
  - Each answers a different set of questions
  - Each co-exists with and builds on top of the others
  - A robust analytic environment would require a mix of different types of analytics in order for businesses to gain a holistic view of the market and compete effectively

Regardless of type, the key goal of ALL analytics is to reduce complex datasets into **actionable** intelligence that supports **decision-making** and organizational processes



Source: Michael Corcoran, Sr. Vice President & CMO. "The Five Types of Analytics." http://docplayer.net/985643-The-five-types-of-analytics-michael-corcoran-sr-vice-president-cmo.html

Question

Where do you begin with analyzing data?

CRISP-DM:
A Framework for Data Analysis

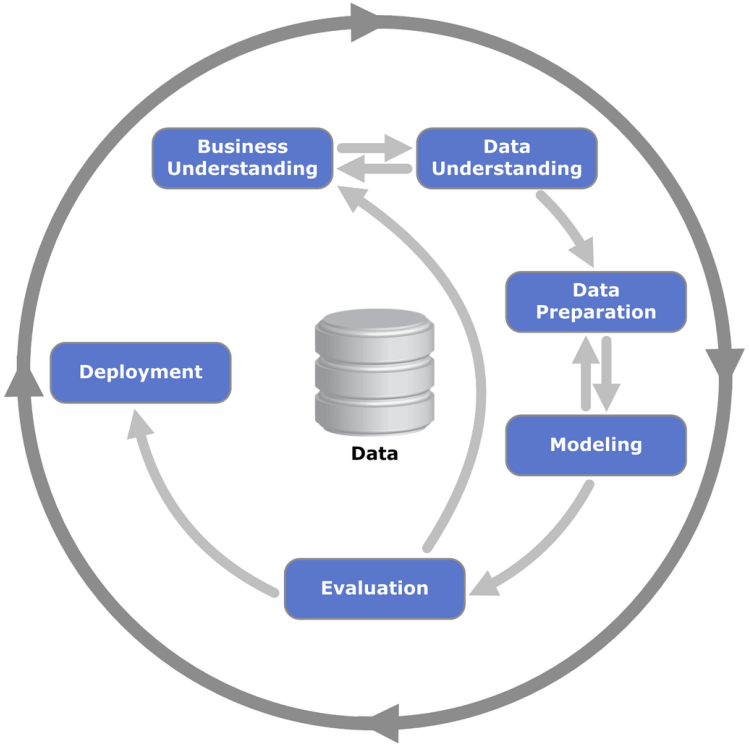


Image Source: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

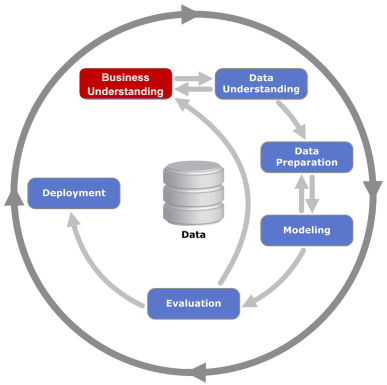
- Acronym for CRoss-Industry Standard Process for Data Mining
- Provides a complete *blueprint* for conducting a data analysis project
- Breaks down data projects into six iterative phases
  - Note the cyclical nature of the framework
  - This is because what is learned leads to more focused business questions, which triggers the next round of analytics

Where Did CRISP-DM Come From?

- In the early 1990's, data-mining (as the data science market was known) was still in its infancy
- As the 90's progressed, interest in data-mining and its capabilities to give companies a competitive advantage became rapidly grew
- In recognition of the need to have industry standards, in 1996, a consortium—funded by the European Commission—was formed
  - The consortium included four leading organizations in data-mining: Daimler-Benz, Integral Solutions Ltd. (ISL), NCR, and OHRA
  - Input from 200+ data-mining users and tool/service providers was gathered over the next several years
    - Hence why the first three letters of CRISP-DM stand for **CR**oss-**I**ndustry
  - The first version of CRISP-DM was released in 2000

Image Source: clipart-library.com

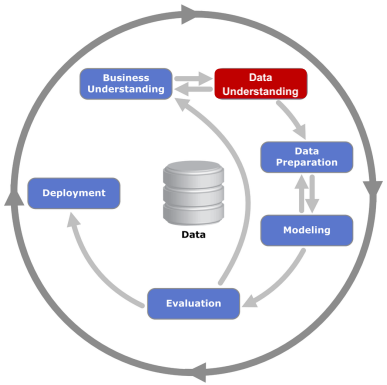
Phase 1: Business Understanding



- Most important phase of any data mining project
- Three major goals:
  - Establish project objectives
    - Understand background, or the “why” for the project
    - Clarify goals for the project
    - Establish criteria for measuring success
    - Assess needed resources, limitations, constraints, and risk of project
  - Define the data project
    - Translate project objectives into data science goals
    - Convert success criteria into data science terms
  - Develop preliminary plan to achieve the objectives
    - Outline specific steps and proposed timeline for project
    - Because not everything always goes perfectly (it would not be research if it did), have some “backup” ideas or strategies available

Phase 2: Data Understanding

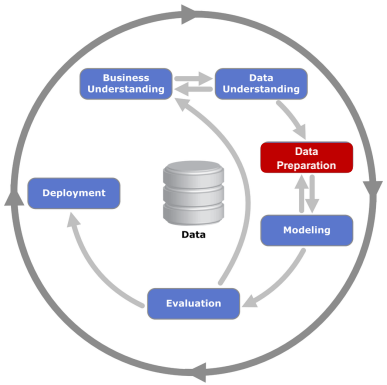
- This phase consists of four major steps:



  - Obtain data
    - May have to integrate if data coming from multiple sources
  - Evaluate data
    - Determine if quality and features sufficient for meeting project objectives
  - Conduct basic descriptive analysis
    - Summarize initial findings
    - Determine if there is a need to revisit Business Understanding phase to revise project definition
  - Verify data quality
    - Identify issues with missing data, plausibility of values, consistency in meaning, etc.

Phase 3: Data Preparation

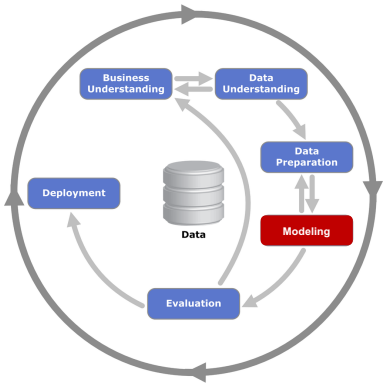
- This phase involves cleaning and organizing data to construct the final dataset that will be analyzed



- Major steps include:
  - Selecting/filtering records and specific attributes, or features, to use for analysis
  - Cleaning data to resolve data quality issues
  - Converting or deriving new data attributes as needed
  - Integrating data from multiple sources create a single input for analysis
  - Formatting data

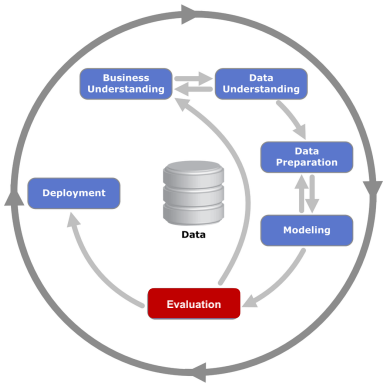
DO NOT SKIMP ON THE DATA PREPARATION WORK!
Your results downstream are only as good as what you do here.

Phase 4: Modeling



- This is the “fun” part
  - This is what everyone thinks about when you say “AI” or “predictive modeling”
- There are several major steps here:
  - Choose analytical model or technique to perform
    - Decision tree? Neural network? Chi-square test? T-Test? Regression?
  - Design test criteria for model
    - What measures will be used to evaluate performance and how?
  - Build the model
    - Exactly what it says…
  - Assess model performance
    - Based on domain knowledge, data mining success criteria, test design, etc.
- Note that this step is iterative with the previous phase; based on the results, it may be necessary to go back to the data preparation phase

Phase 5: Evaluation

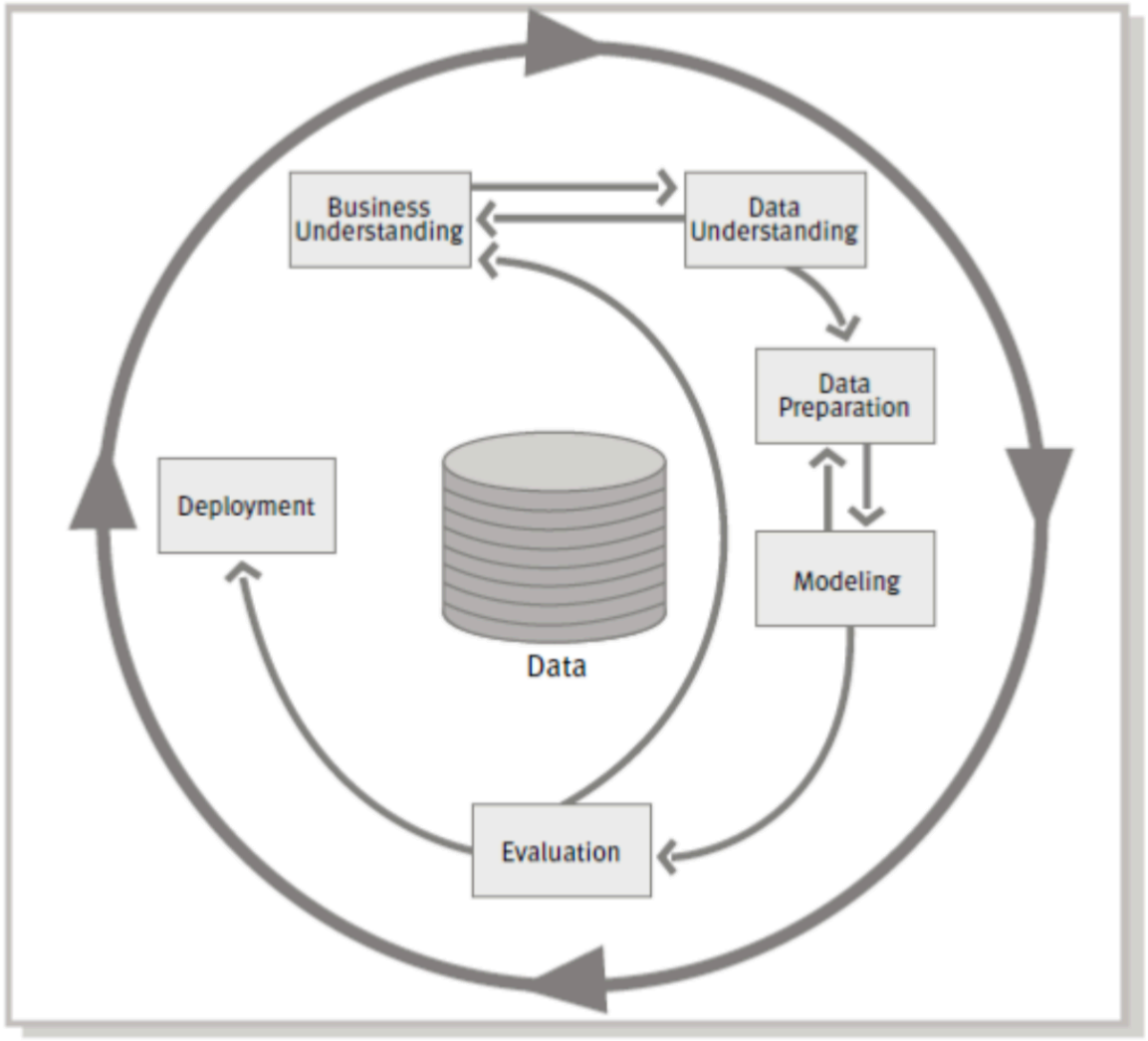


- The goal of this phase is to make sure that the model meets business objectives prior to deployment
- Key steps are:
  - Evaluate results to determine degree to which the model meets business objectives
    - This is different from the evaluation done in the modeling phase, which focuses on accuracy of the results
  - Review process to make sure that no important factor or step was missed during model development
  - Determine whether to deploy, or go back and initiate further iterations

Phase 6: Deployment

- Model creation is generally not the end goal of a data project; need to figure out how to "deploy"
  - Remember…you're not creating the models just for yourself
- Deployment can take many forms, depending on project requirements
  - Can be as simple as generating a report, or as complex as implementing a process across the organization
- More complex deployments will involve yet another set of processes related to the Project Management Life Cycle
  - Initiation: already completed
  - Planning: what will deployment look like?
  - Monitoring: ensure that processes are working and results are being used correctly
  - Report: document deliverables, and summarize/organize results
  - Review: Assess failures and successes; document lessons learned

Cross-Industry Standard Process for Data Mining (CRISP-DM)



Data Understanding

- Collecting the data.
- Describing the data.
- Exploring the data.
- Verifying the data quality.

This step is the classic case of

Extract, Transform, Load (**ETL**)

Data Preparation

- Selecting the data.
- Cleaning the data.
- Constructing the data.
- Integrating the data.
- Formatting the data.

Modeling

- Selecting a modeling technique.
- Generating a test design.
- Building a model.
- Assessing a model.

Evaluation

- Evaluating the results.
- Reviewing the process.
- Determining the next step

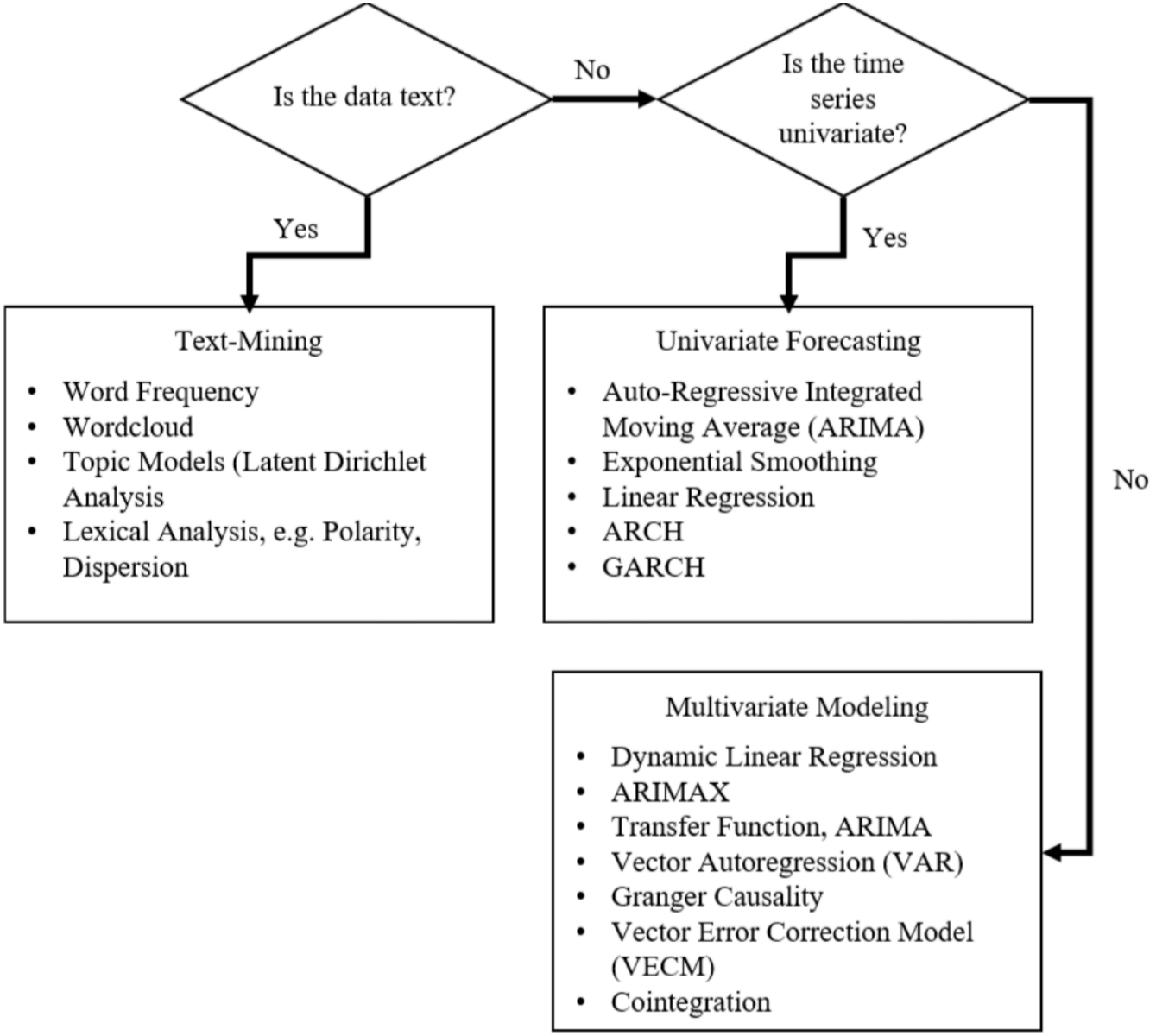
Deployment

- Deploying the plan.
- Monitoring and maintaining the plan.
- Producing the final report.
- Reviewing the project.

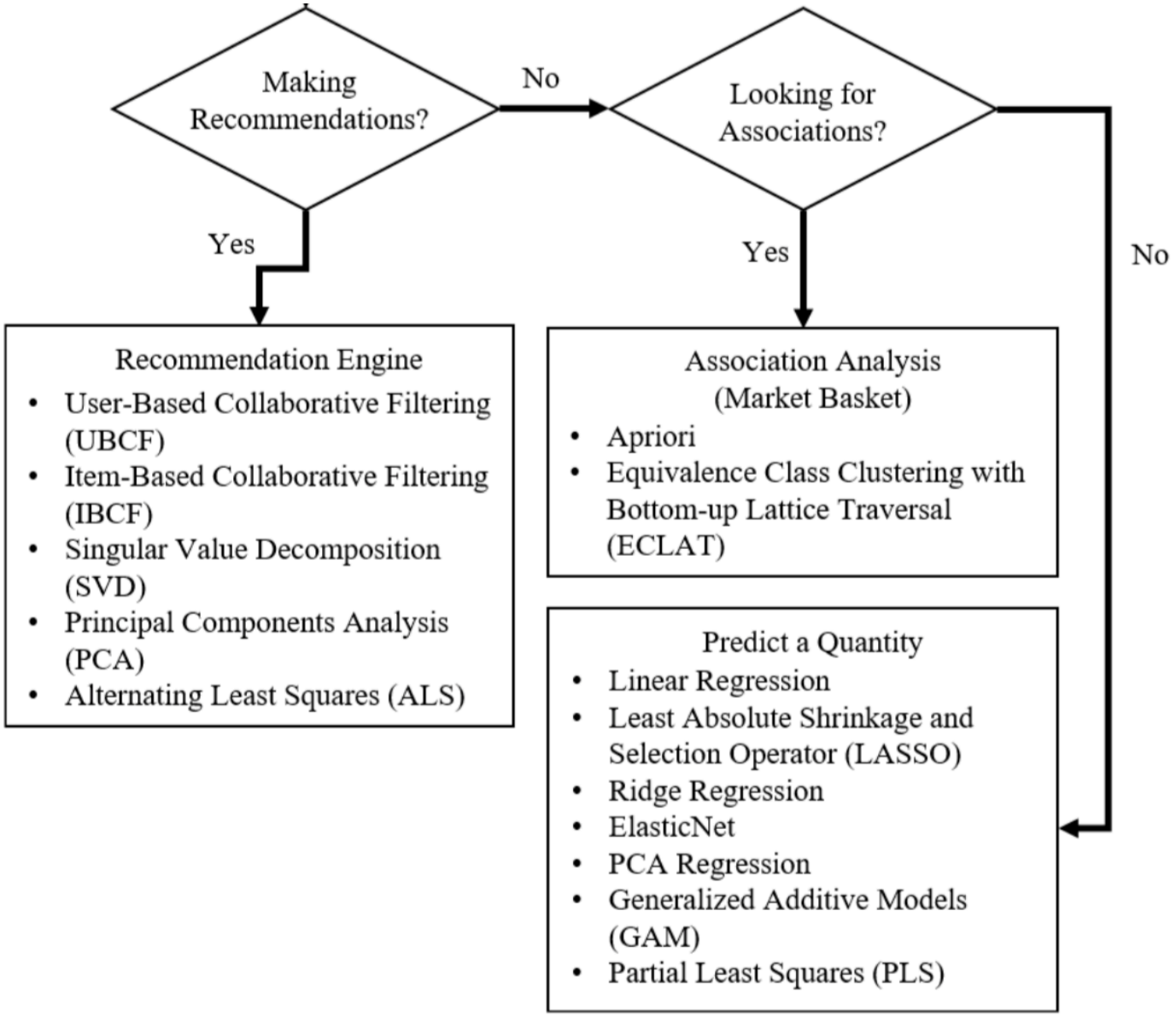
Question

Where do we begin with modeling?

Text data ? Time Series?



Predicting a Category?



No labels?

Clustering

- Hierarchical
- K-Means
- Partition Around Medoids
- Self-Organizing Map (SOM)
- Fuzzy Clustering
- DBSCAN

Labeled data?

Classification

- Logistic Regression
- Linear Discriminant Analysis (LDA)
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Neural Networks/Deep Learning
- Decision Trees
- Random Forest
- Gradient Boosting
- Naïve Bayes
- Survival Analysis