Introduction to Big Data

Analytic Model Development Considerations

YORK
UNIVERSITÉ
UNIVERSITY

Review Previous Lesson

Review Concepts from Day 3

- During Day 3 you learned to:
- Identify, interpret and correctly apply categories of measurement data
  - Ordinal
  - Nominal
  - Interval
  - Ratio
- Interpret descriptive statistics to quantitatively describe the shape of a variable
  - Central tendency measures
  - Dispersion measures
  - Skewness and Kurtosis “height” measures
- Understand and describe bivariate data set according to:
  - Scatter plots.
  - Linear relationships
  - Correlations from a scatter plot
  - Correlation coefficient
- Understand basic Data Exploratory methods to develop Probability Distributions
  - Probability Density
  - Data Relationships
  - Data Frequency Distributions
- Apply filtering techniques and query from multiple data tables from Pandas/Python

Day 4 – New Topics Introduced

The following major topics are discussed this class.

- Analytic models
- Statistical modeling
- Machine Learning
- Data Mining
- CRISP-DM
- Iterative vs Waterfall Approaches
- Analytic and Data Science Teams

Learning Objectives for Day 4

Day 4 - Learning Objectives

During Day 4 you will learn to:

- Describe a range of different analytical modeling techniques
  - Statistical modeling
  - Data mining
  - Machine learning
- Describe the Model Development Process CRISP-DM
  - Business Problem Framing
  - Data Collection and Evaluation
  - Data Preparation
  - Model Building
  - Model Evaluation
  - Model Deployment
- Differentiate and recommend a suitable development process
  - Iterative
  - Waterfall
- Identify the structure and role of Programs, Projects and Services in terms of model development
  - Analytic team composition and structure
- Build and evaluate linear regression models using Python

Some Ways To Approach Models

1. Positivism- objectivity of scientific analysis and testing hypotheses to build knowledge and understanding
2. Humanistic- people create subjective worlds in their minds- behavior understood only by a methodology that penetrates the subjectivity
3. Structuralists- cannot explain observed pattern by examining pattern itself. But rather establish theories to explain development of societal conditions within which people must act

Role of Statistics

- Room in all the above interpretations for quantitative analysis.
- But increasingly both quantitative and qualitative analysis are important
- Qualitative analysis involves?
- Statistics and measurement are used commonly in our lives
  - A. Making home purchase decisions
  - B. Setting up investments
  - C. Weather variations are expressed as probabilities

How Do Data Analysts Use Statistics?

1. Describe and summarize data
2. Make generalizations concerning complex spatial patterns
3. Estimate likelihoods of outcomes for events at particular location(s)
4. Use sample data to make inferences about a larger set of data (a population)
5. Learn whether actual pattern matches an expected or theoretical
6. Wish to compare or associate (correlate) patterns of distributions

Formulating the Research Process

1. Problem Identification
2. Develop Questions to Investigate
3. Collect and Prepare Data
4. Process descriptive data (maps, graphics)>>>>> Reach conclusions
5. Formulate Hypothesis >>>>> Collect and Prepare Sample Data
6. Test Hypothesis>>Evaluate Hypothesis
7. Develop Model, Law, or Theory

What Are Models?

- Abstractions of the real world
- Simplified versions of reality
- Easier to examine scaled down and simplified structures in attempt to understand
- Iconic models- look like what they represent (
- Analogue models- one property used to represent another
- Symbolic models- equations

Basic Terms and Concepts

- Data element- basic element of information which we measure
- Data Set- groups of data (commuting sheds of industries)
- Observations-Cases-Individuals- elements of phenomena under study
- Variable- property or characteristics of each observation that can be measured, classified or counted
- Values may vary among set of observations: rainfall, per capita income, years of schooling

Geographic Data

1. What sources of data are available?
2. Which methods of data collections should be used?
3. What type of data will be collected and then analyzed statistically?

Types of Data

- Primary Data- acquired directly from original source

1. Information collected in the field
2. Usually very time consuming
3. Involves decision about a sample design so representative data may be obtained

Types of Data

- Secondary Data (or Archival Data)

1. Usually collected by some organization (United Nations, U S Bureau of Census)

2. Often easily accessible- hardcopy or CD rom

3. Less time consuming but also more limiting

4. Often need to inspect historical records and archives for diaries, oral histories, official reports in order to develop a picture of problem

Characteristics of Data – Explicit/Implicit Quality

1. Some data are explicitly spatial- locations are directly analyzed
2. Other data implicitly spatial- data represents places but locations themselves are not analyzed (population sizes of towns)

Measurement Concepts

1.Precision- level of exactness associated with measurement (rain gauge to inches or fractions of inches)

2. Accuracy- extent of system wide bias in measurement process

3. Validity- if scientific concept is complex expressing "true" or "appropriate" meaning of the concept through measurement may be difficult (levels of poverty, economic well being, environmental quality)

4. Reliability- changes in spatial patterns are analyzed over time must ask about consistency and stability of data

Types of Statistical Analysis

- Descriptive Statistics- concise numerical or quantitative summaries of the characteristics of a variable or data set (e.g. mean, standard deviation, etc)
- Inferential Statistics- here we wish to make generalizations about a statistical population (total set of information or data under investigation) based on the information from a sample
- Sample- typical or representative or unbiased subset of the broader, larger more complete statistical population

Analytic Modeling

Analytic Modeling Outline

- Analytic Models
  - Variable Names
  - Terminology
  - Functionality
- Statistical Modeling
  - Univariate
  - Bivariate
  - Multivariate
- Machine Learning
  - Supervised Learning
  - Unsupervised Learning
- Data Mining
  - Regression Models
  - Decision Trees

Analytic Models

Analytic Models

Description

Models are representations of reality
They are based on simplifications and assumptions

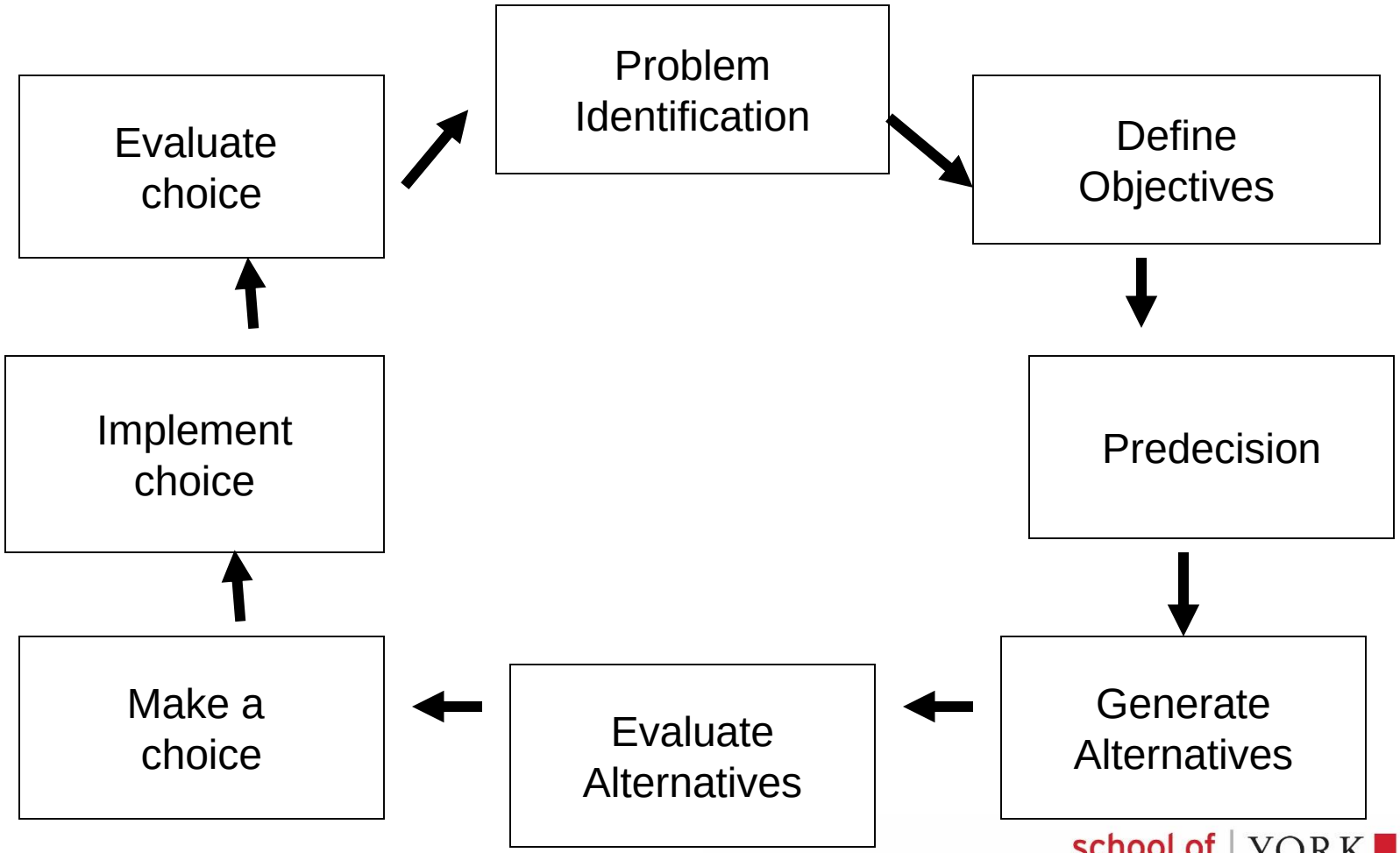
"All models are wrong. However some are more useful than others"

- George Box (famous statistician and modeller)

General Analytical Model of Decision Making

- **Conception of decision making as a series of analytical steps**
- **The model focuses on two aspects:**
  - **identifying the problem**
  - **implementing the solution**

Steps ion the Analytical Model



Programmed vs Non-Programmed Decisions

- Programmed decisions are highly routine decisions usually made by lower level workers alone. (Time to re-order toner for the copy machine)
- Non-programmed decisions -- Decisions about a novel problem. There is no set course of action. (Identifying the cause of some rare disease)

Certain vs Uncertain Decisions

- The degree of risk involved in making the decision. The means the continuum ranges from
  - complete uncertainty (no risk) to
  - complete uncertainty (high risk)
- The point is to minimize risk by obtaining information

Rational-Economic Model

- A model in which decision makers consider all possible solutions before selecting the optimal one. Presumes
- We are entirely rational and logical
- We have complete and perfect information
- We can process all this information

Bounded Rationality Model

- A model that realizes that humans have a limited view of the problem, do not have perfect information or all the alternatives
- We accept a solution that is good enough
- This is called satisficing or settling for the decision that meets the criteria but may not meet them optimally

Availability Heuristic

- **The tendency for people to base decisions on information that is easily accessed**
- **Which is riskier?**
- **A. Driving a car on a 400 mile trip**
- **B. Flying on a 400 mile trip on a commercial flight**

Availability Heuristic

You are traveling in the Middle East. Which is the greater worry?

- **Being hurt in an auto accident**
- **Being hurt in a terrorist attack**

Representativeness Heuristic

- The tendency to perceive other in stereotypical ways if they appear to be typical of the category to which they belong

Mark is finishing his MBA at a prestigious university. He is very interested in the arts and at one time considered a career as a musician. Mark is most likely to take a job

- A. In the management of the arts
- B. With a management consulting firm

Other Cognitive Biases

- Anchoring -- People develop estimates based on initial information. When that information turns out to be wrong, people make adjustments to their decisions but not enough to overcome the impact of the initial piece of information.
- Confirmation bias -- We tend to focus only on evidence that supports our decisions.
- Overconfidence -- Because of confirmation bias, we are more confident than we are correct.

Analytic Modeling Techniques

Analytic Modeling Techniques

Statistical Modeling

Statistical Models

- category of mathematical model
- assumptions about the generation of sample data and similar data from a larger population
- represent the data-generating process
- describes a set of probability distributions that are assumed to approximate the distribution for a particular data set
- specified by equations that relate one or more random variables and possibly other non-random variables
- form the foundation of statistical inference

Source - Wikipedia

Analytic Modeling Techniques

Statistical Modeling

Univariate Analysis - Review

- Estimate the shape of a variable's distribution (central tendency & dispersion)
- Can be interpreted as a probability function
- Useful for generating probability estimates of future values
- Useful for carrying various statistical tests

Multivariate Analysis

- Estimate correlation values between pairs of variables
- Regression methods to find the best fit between a dependent variable and set of independent variables
- Regression models can be
  - Simple or Multiple
  - Linear or Non-Linear
  - Continuous output or Binary output

Analytic Modeling

Machine Learning

- field of computer science
- gives computers the ability to learn without being explicitly programmed
- Arthur Samuel coined the term "Machine Learning" in 1959 while at IBM
- evolved from the study of pattern recognition and computational learning theory in artificial intelligence
- explores the study and construction of algorithms that can learn from and make predictions on data
- sometimes conflated with data mining where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning
- in he field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction;
- allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data

- Source Wikipedia

Analytic Modeling

Supervised Learning

- the task of inferring a function from *labeled training data*
- training data consist of a set of *training examples*.
- each example consists of an input object and a desired output value
- analyzes the training data and produces an inferred function, which can be used for mapping new examples
- will allow for the algorithm to correctly determine the class labels for unseen instances
- requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way

- Source Wikipedia

Analytic Modeling

Unsupervised Learning

- the task of inferring a function to describe hidden structure from "unlabeled" data (a classification or categorization is not included in the observations).
- the examples given to the learner are unlabeled, there is no evaluation of the accuracy of the structure that is output by the relevant algorithm
- one way of distinguishing unsupervised learning from supervised learning

- Source Wikipedia

Analytic Modeling

Data Mining

- computing process of discovering patterns in large data sets
- involves methods at the intersection of machine learning, statistics, and database systems]
- an interdisciplinary subfield of computer science
- overall goal s to extract information from a data set and transform it into an understandable structure for further use.
- involves database and data management aspects, data pre processing, model and inference considerations, post-processing of discovered structures, visualization, and online updating
- term is a misnomer because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data itself
- is the analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining).

- Source Wikipedia

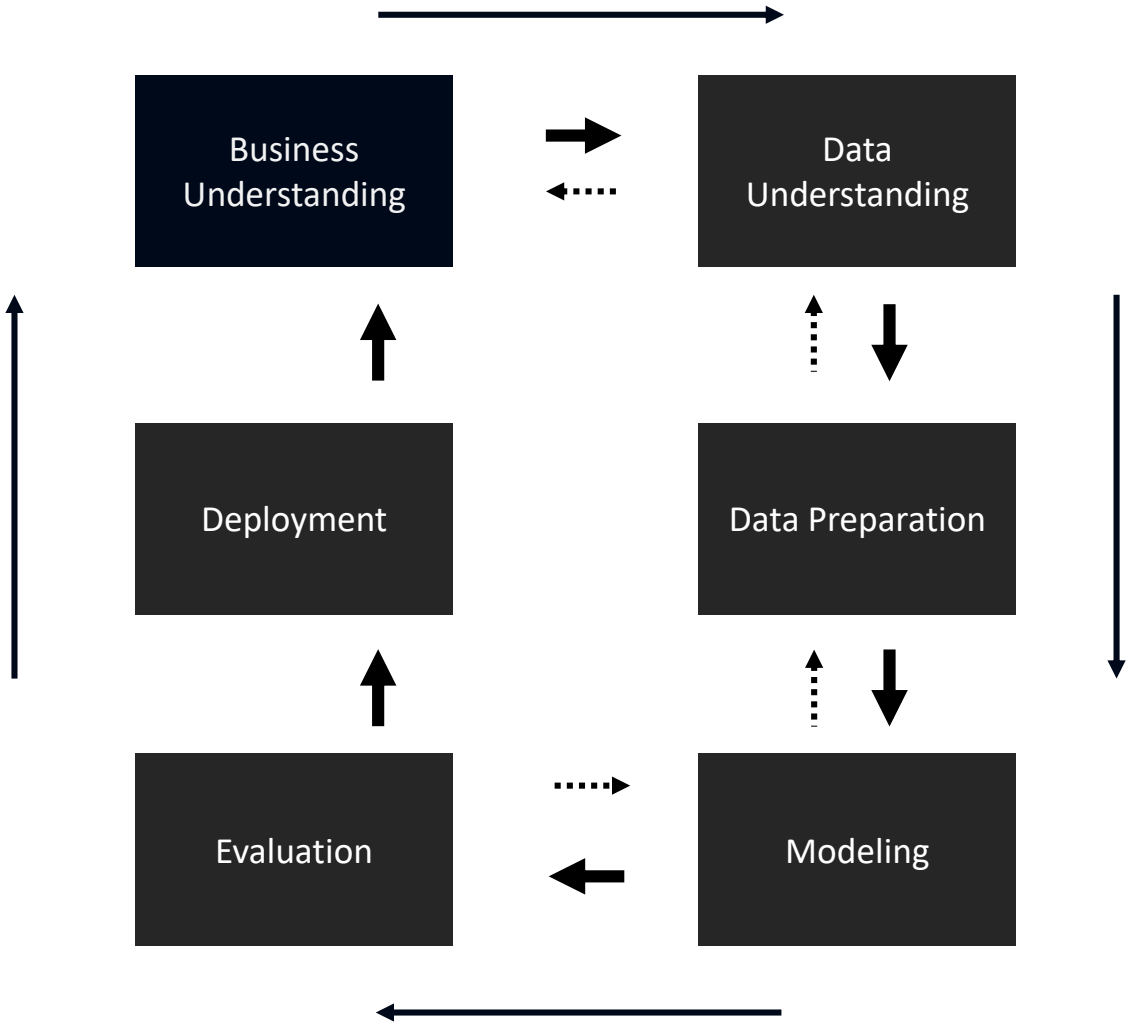
Model Development

Model Development

Topics

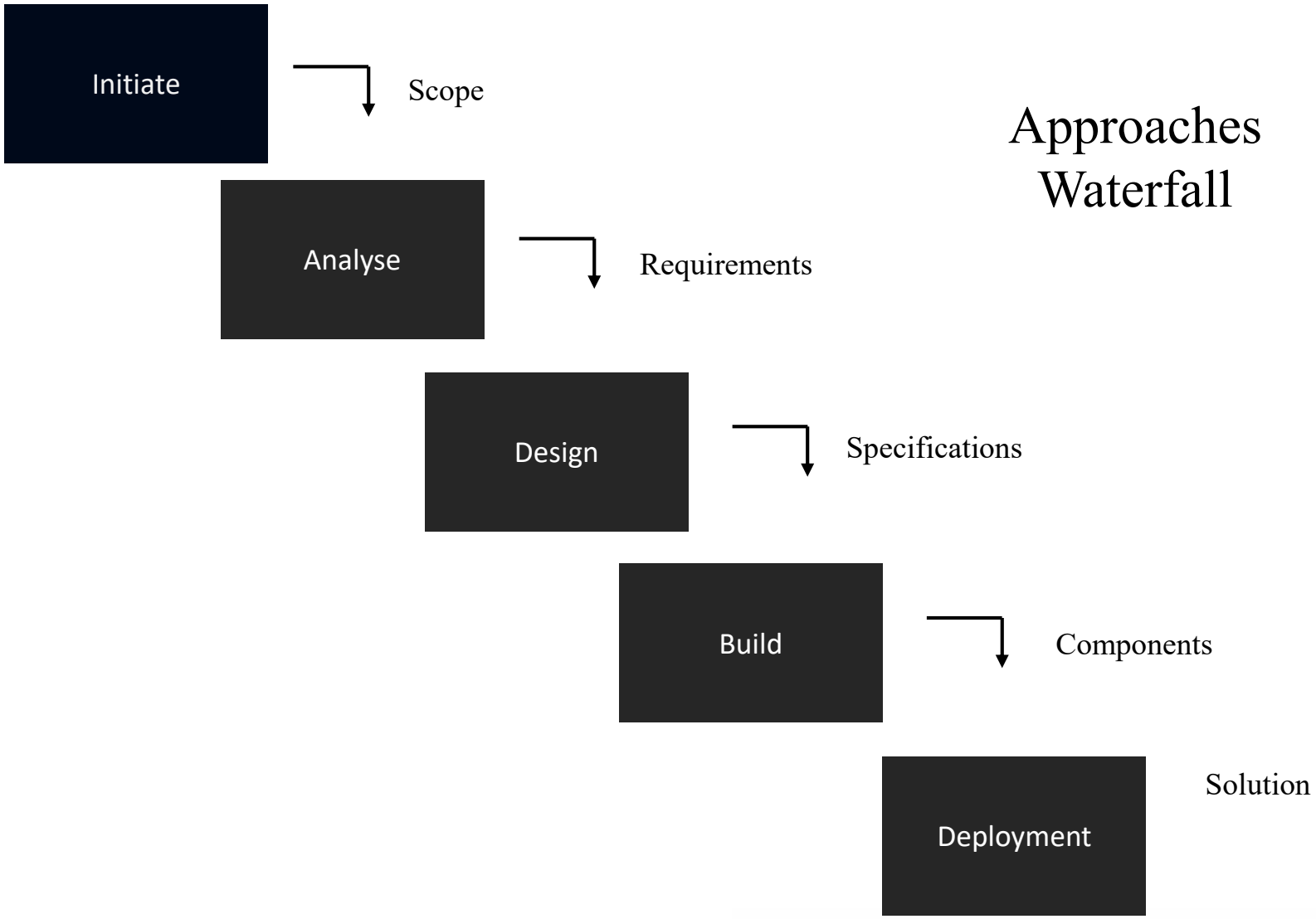
- Methodology
- Approaches
- Delivery Models
- Teams

Model Development



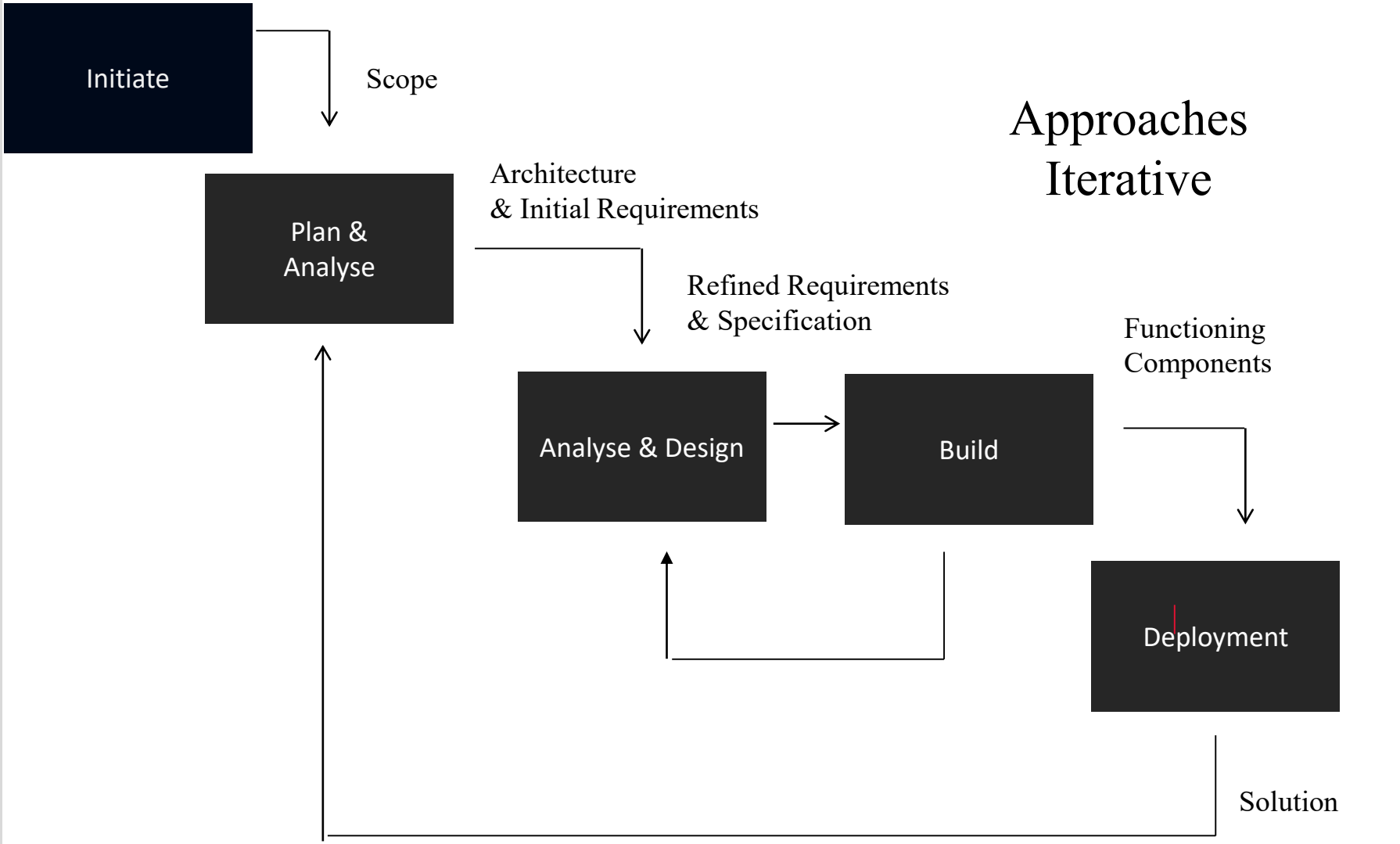
Methodology
CRISP-DM

Model Development

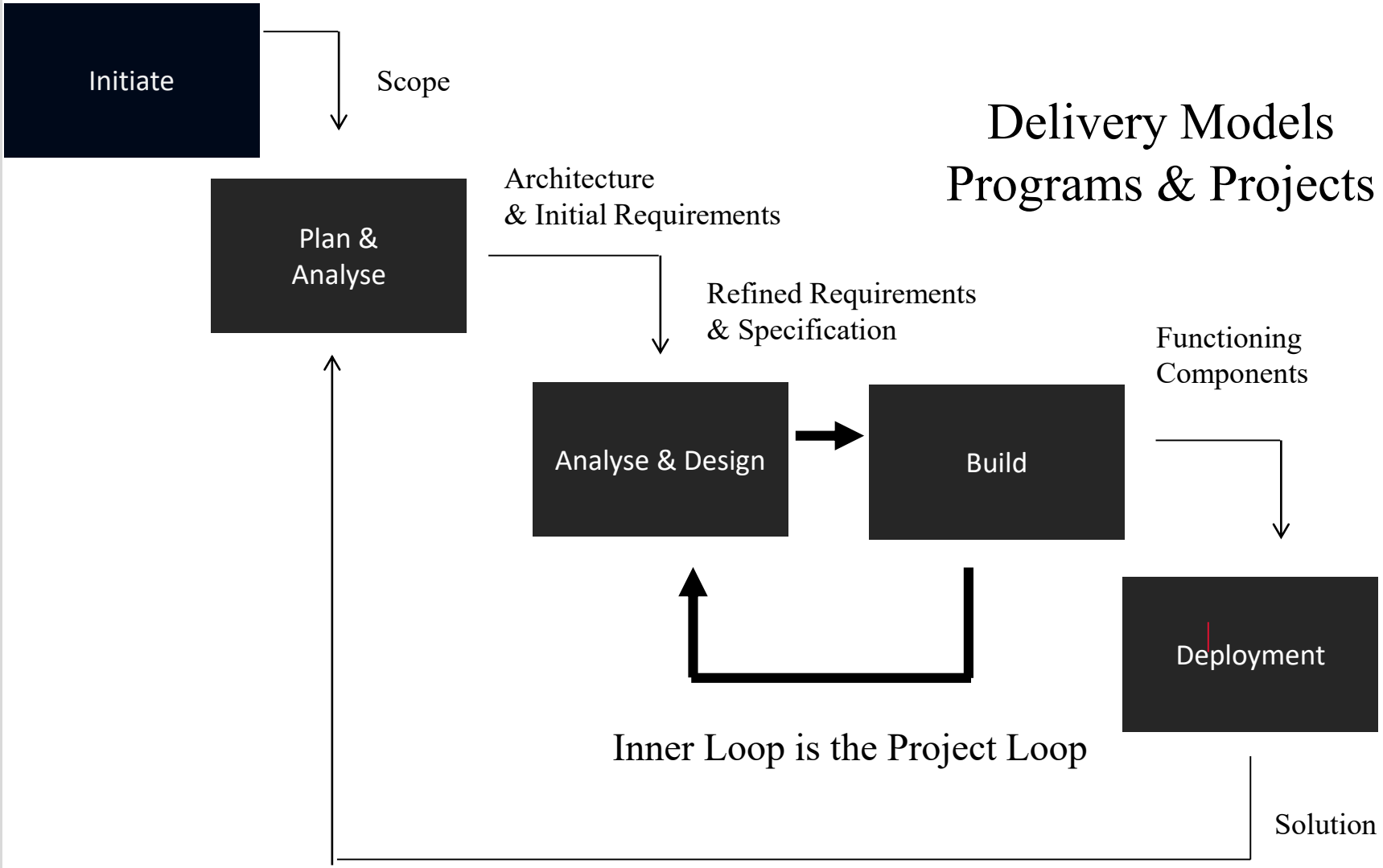


Approaches
Waterfall

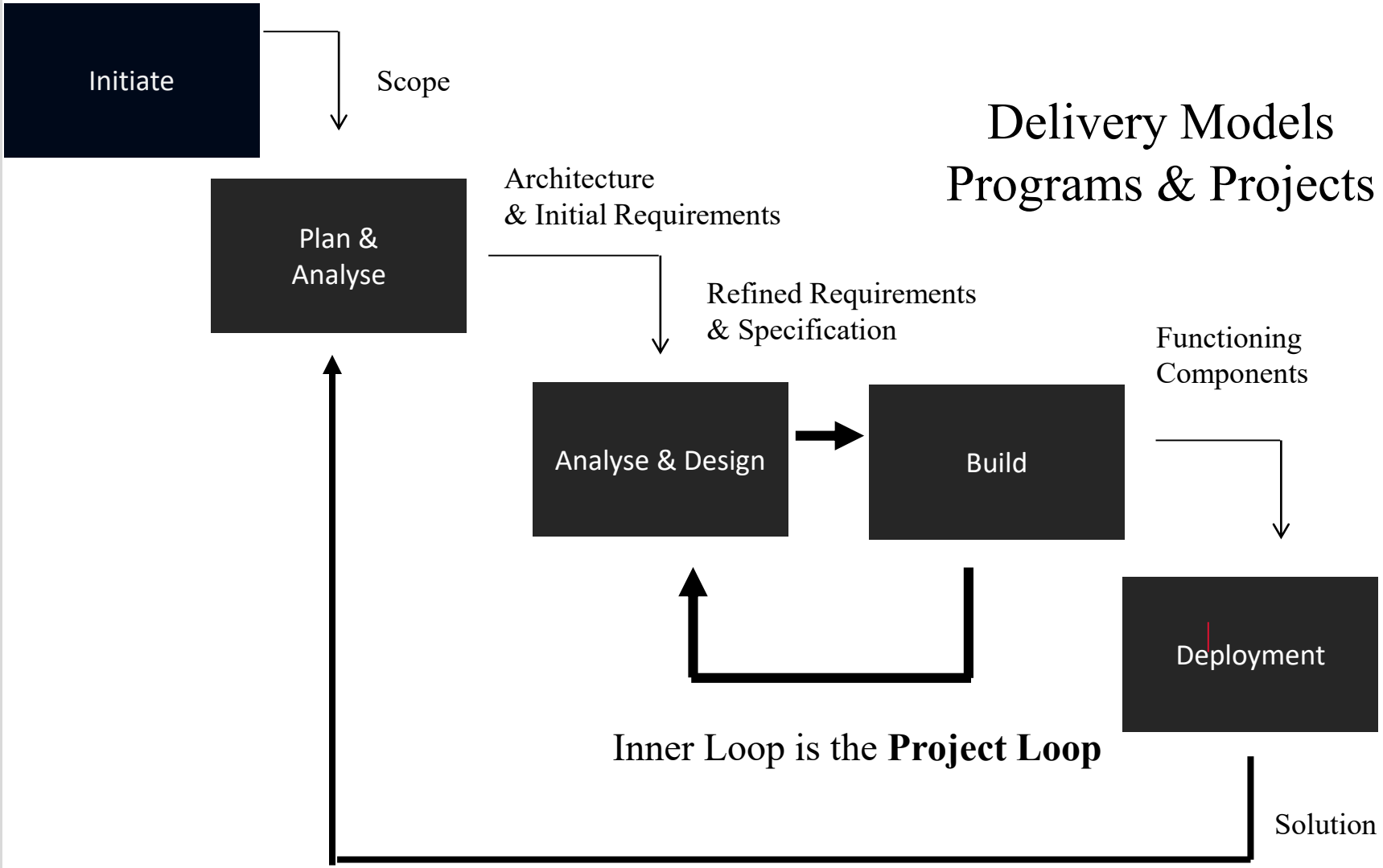
Model Development



Model Development



Model Development



Lesson Review

Lesson Review

Consider the following questions that you should be able to answer by completing Day 4.

- What are some terms used to describe input variables?
- What are some terms used to describe output variables?
- What are some common functions of models?
- What are some core techniques used in statistical modeling?
- What are some differences between supervised learning unsupervised learning?
- What is the purpose of data mining?
- What are the major steps of CRISP-DM?
- Why is the waterfall approach not recommended for analytics projects?

Lesson Summary

Day 4 Lesson Summary

During Day 4 you learned to:

- Describe a range of different analytical modeling techniques
  - Statistical modeling
  - Data mining
  - Machine learning
- Describe the Model Development Process CRISP-DM
  - Business Problem Framing
  - Data Collection and Evaluation
  - Data Preparation
  - Model Building
  - Model Evaluation
  - Model Deployment
- Differentiate and recommend a suitable development process
  - Iterative
  - Waterfall
- Identify the structure and role of Programs, Projects and Services in terms of model development
  - Analytic team composition and structure
- Build and evaluate linear regression models using Python