THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Praxis Development for Artificial Intelligence

Professor Hamza F. Alsarhan

SEAS 8599 – DA1

Lecture 8

June 15, 2024

Agenda

1. Main Takeaways from HW #1 – HW #3
2. Data
3. Data Analytics
4. Praxis Review
5. Research Methodology
6. Results & Conclusion

Main Takeaways from HW #1 – HW #3

1. Concise and clear title
2. Scope of work that can be accomplished within a year of research
3. Quantifiable problem statement
4. Thesis statement that makes a claim with a deliverable
5. Research questions that help develop your deliverable and get you closer to a solution for the issue in your problem statement
6. Research hypotheses that are testable and make claims to answer your research questions
7. Annotated Bibliography that is relevant to your topic

Data – Data Types

There are two main types of data:

- **Categorical (Qualitative)** – tends to be descriptive using names or labels for each data point.
  - **Examples:** colors, brand names, yes/no answers, education levels, etc.
- **Numerical (Quantitative)** – consists of numbers that quantify each data point.
  - **Examples:** age, test scores, number of cars, etc.

Data – Qualitative Data

Qualitative data subtypes:

- **Nominal** – categorizing the data points into separate unique categories with no specific order.
  - **Examples:** colors, brand names, yes/no answers, etc.
- **Ordinal** – categorizing the data points into categories with a specific order.
  - **Examples:** customer satisfaction levels, education levels, etc.

Data – Quantitative Data

Quantitative data subtypes:

- **Discrete** – numerical data points that can be finitely countable.
  - **Examples:** number of students in class, number of employees at a company, etc.
- **Continuous** – numerical data points that can be measurable.
  - **Examples:** time, height, temperature, etc.

Data – Sources

Some common sources of datasets used in academic research are:

- **Databases** – a structured format of storing and organizing both numerical and descriptive datasets.
- **Graphics** – a visual format of providing insights, such as navigational, transportation lines, trend lines, scatter plots, etc.
- **Literature** – in the form of journal articles (best practices), handbooks (step-by-step instructions), timetables (simulations), and previous findings.
- **People** – in the form of surveys (opinions) or filling in data gaps (based on expert judgement).

Data – Common Issues

Common issues with datasets include, but are not limited to, the following:

- **Data access** – not gaining approval to use a dataset for your research or not being allowed to publish the dataset/findings as part of your praxis.
- **Data imbalance** – wide gaps in representation of different classes in the dataset leading to poor analysis results and performance.
- **Data incompleteness** – various values missing from different columns and rows making it difficult to use those features or records.
- **Data gaps** – significant parts of the data missing, which is especially problematic with datasets collected over a period of time.
- **Data inconsistency** – data pooled from different sources representing similar scenarios in materially different ways.
- **Data scale** – wide gaps in scale between different features in the dataset leading to unfair representation of larger scale features over smaller scale features.

Data – Common Issues for ML Models

Common issues with datasets for ML models include, but are not limited to, the following:

- **Insufficient Quantity of Training Data –** In general, it takes a lot of data for most machine learning algorithms to work properly (thousands, millions+ of training examples).
- **Nonrepresentative Training Data –** To generalize well, your training data should be representative of the new cases you want to generalize to.
- **Poor-Quality Data –** Difficult to detect underlying patterns when data is full of errors, outliers, and noise.
- **Irrelevant Features –** Characteristics that have little or no connection to the target variable can cause issues such as reduced prediction accuracy, increased model complexity, higher computational costs, and introduced bias or noise.

Sources: Hands-On Machine Learning, Ch1, p25-7

Data – Common Operations

Common operations on datasets include, but are not limited to, the following:

- **Data imputation** – replacing missing data points in the dataset with substitute values.
- **Data balancing** – using majority class under-sampling and minority class over-sampling to lead to a balanced dataset.
- **Data deduplication** – identifying and removing duplicate records from the dataset to avoid over representation of certain observations.
- **Data reshaping** – modifying the structure of the dataset using stacking or pivoting to make it more suitable for conducting analyses.
- **Data normalization** – scaling the values of various columns in the dataset to use a common range.

Data Analytics – Types

Main types of data analytics:

- **Descriptive** – comprehensive, accurate, and live data. Effective visualization of the data.
- **Diagnostic** – drilling down to discover the root cause of a specific pattern or phenomenon.
- **Predictive** – using consistent historical values and patterns to forecast future trends.
- **Prescriptive** – using various forecasts to determine the various outcomes and lead to selecting the optimal choice/decision.

Data Analytics – Graphical Representation

Graphical representation of multiple variables can be accomplished in various ways depending on the data types:

- **Line graphs** – used to track the changes of a variable over the course of another variable (usually time).
- **Scatter plots** – used with two quantitative variables to visualize the relationship between the two variables.
- **Box plots** – used to provide summary statistics of numerical variables, such as min, max, median, first quartile, and third quartile.
- **Contingency tables** – used to display the frequencies of multiple categorical variables. Useful for probability applications.
- **Confusion matrices –** used to visualize the performance of a machine learning classification model. It compares the number of actual instances to predicted instances of each class in the dataset.

Data Analytics – Exploratory Data Analysis

Exploratory data analysis (EDA) is a crucial first step to learn about your dataset. Useful Python commands for EDA on a data frame called *df* are:

- **df.info()** – displays information about a data frame, such as the index data type, columns, non-null values, and memory consumption.
- **df.describe()** – provides a basic statistical summary about numerical columns in the dataset, such as mean, standard deviation, min, max, etc.
- **df.count()** – provides the count of non-NA values in each column in the data frame.
- **df.head()** – returns the top n rows of the data frame.
- **df.nunique()** – returns the number of unique values in either the columns or the rows.
- **df.isna()** – returns a Boolean (True/False) for whether a given value in the data frame is NA or not.

Source: *Pandas.dataframe#*. pandas.DataFrame - pandas 2.0.3 documentation. (n.d.). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

Data Analytics – Exploratory Data Analysis

Additional Python commands for EDA are:

- **df.shape** – provides the number of rows and columns in the data frame.
- **df.size** – provides the total number of data points in the data frame.
- **df.columns** – provides the names of the columns in the data frame.
- **df.index** – provides the names of the rows in the data frame.
- **df.dtypes** – provides the data type of each column in the data frame.
- **df.ndim** – provides number of axes in the data frame (1 for Series and 2 for DataFrame).

Source: *Pandas.dataframe#*. pandas.DataFrame - pandas 2.0.3 documentation. (n.d.). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

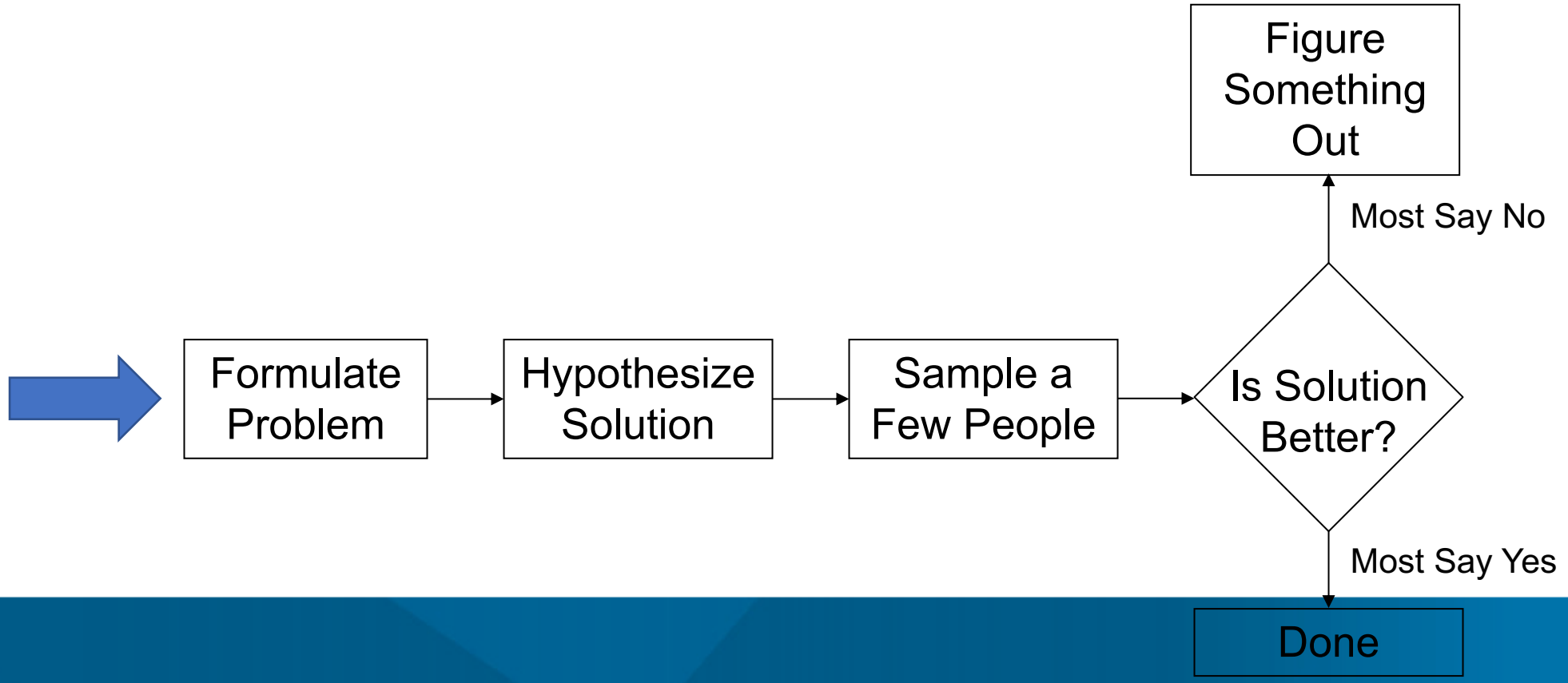
Data Analytics – Data Preprocessing

Data preprocessing is an important step to perform in order to prepare the dataset for further analytics after EDA.

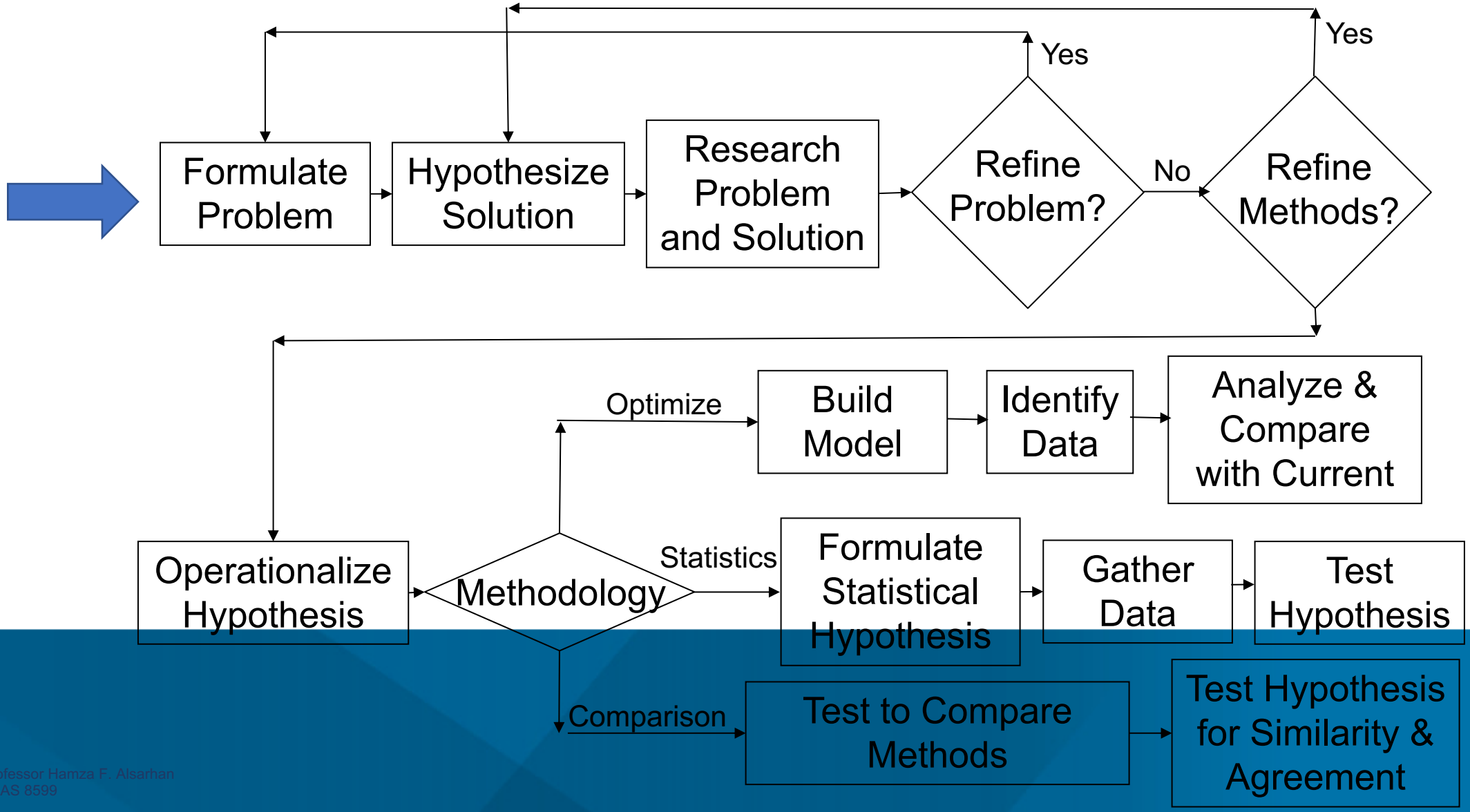
- **df.fillna()** – replaces missing values in the data frame with a specified value.
- **df.dropna()** – drops rows or columns that have a certain number of missing values.
- **df.drop()** – drops specific columns or rows from the data frame.
- **df.interpolate()** – populates NaN values in the data frame with values generated using a certain interpolation method.
- **df.merge()** – merges multiple data frames together.

Source: *Pandas.dataframe#*. pandas.DataFrame - pandas 2.0.3 documentation. (n.d.). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

Praxis Review – What A Praxis Is NOT



Praxis Review – What A Praxis Could Be



Praxis Review – Abstract

- Single most important parts of your praxis that captures the essence of your work
- Describes your praxis without having to read the entire document
- Most readers decide to read or pass the praxis based on what is in the abstract
- 200 to 350 words long
- It includes:
  - Your problem statement
  - The significance of your research
  - Your Thesis Statement
  - Your methodology
  - Your major findings and conclusions
- Does **NOT** include citations

Praxis Review – Introduction

- What is the Praxis About? **Problem Statement**
- Why is this important? **Background and motivation**
- What are you going to do? **Methodology**
- Why are you taking this approach? **Brief Literature Influence**
- Why/how is this going to be better than what exists or what has been done? **Thesis Statement**
- What will the Praxis look like? **Summary of what is to come**

Praxis Review – Literature Review

The purpose of the literature review chapter is to:

- Show that you understand the problem and the research that has been done to date.
  - The problem may have appeared in similar but not identical circumstances.
  - The problem may have occurred in different industries.
  - There may be many different types of solutions.
- Show that you understand the methodology you have selected and that it can be applied to your problem.

Praxis Review – Literature Review

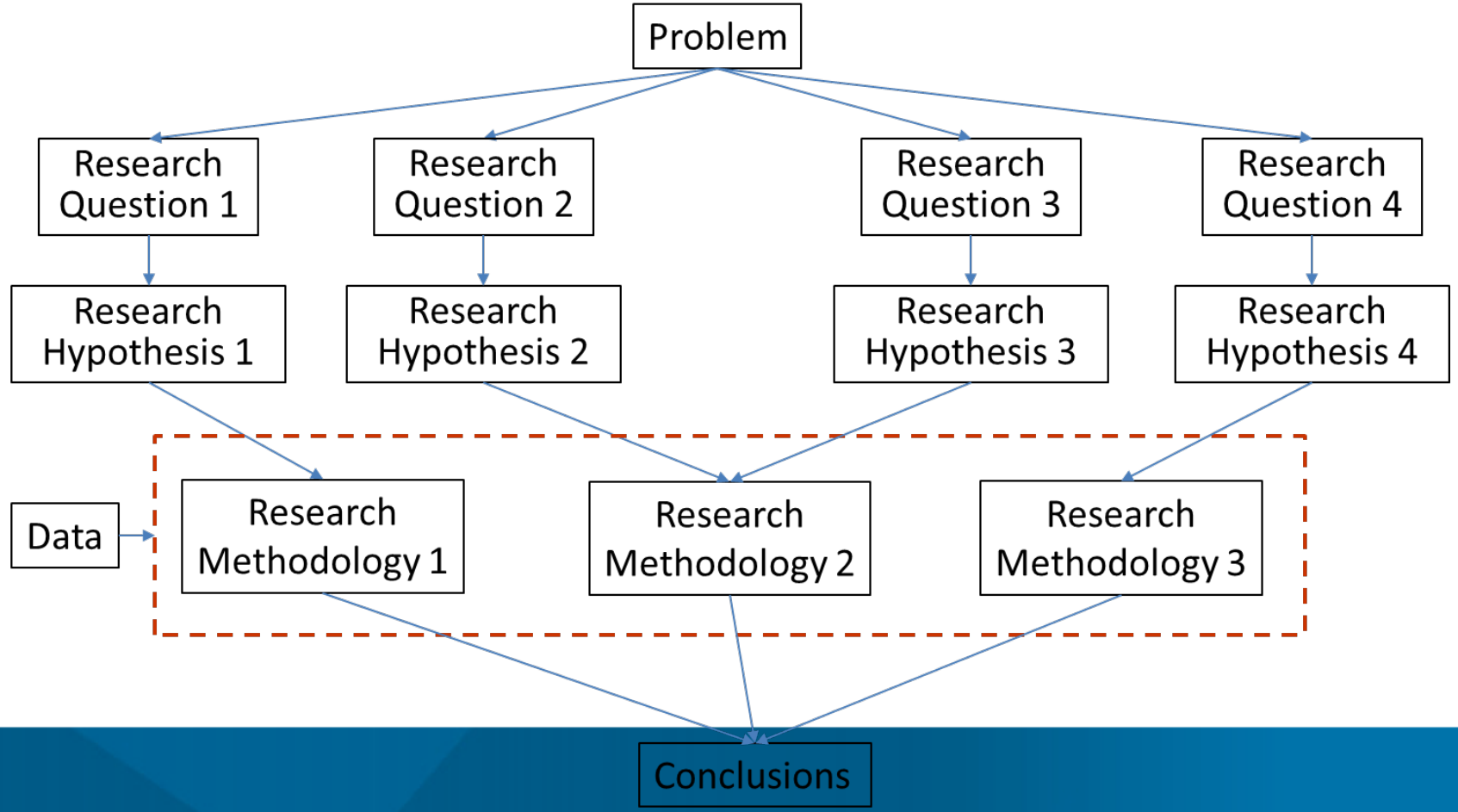
The literature review chapter provides:

- A thoughtful and critical comparison and contrast of the available work provided in the literature that is directly related to the problem at hand.
- A solid overview of the methodology used, including its assumptions for use, its limitations, and any applications it was used in that are similar to the problem at hand.

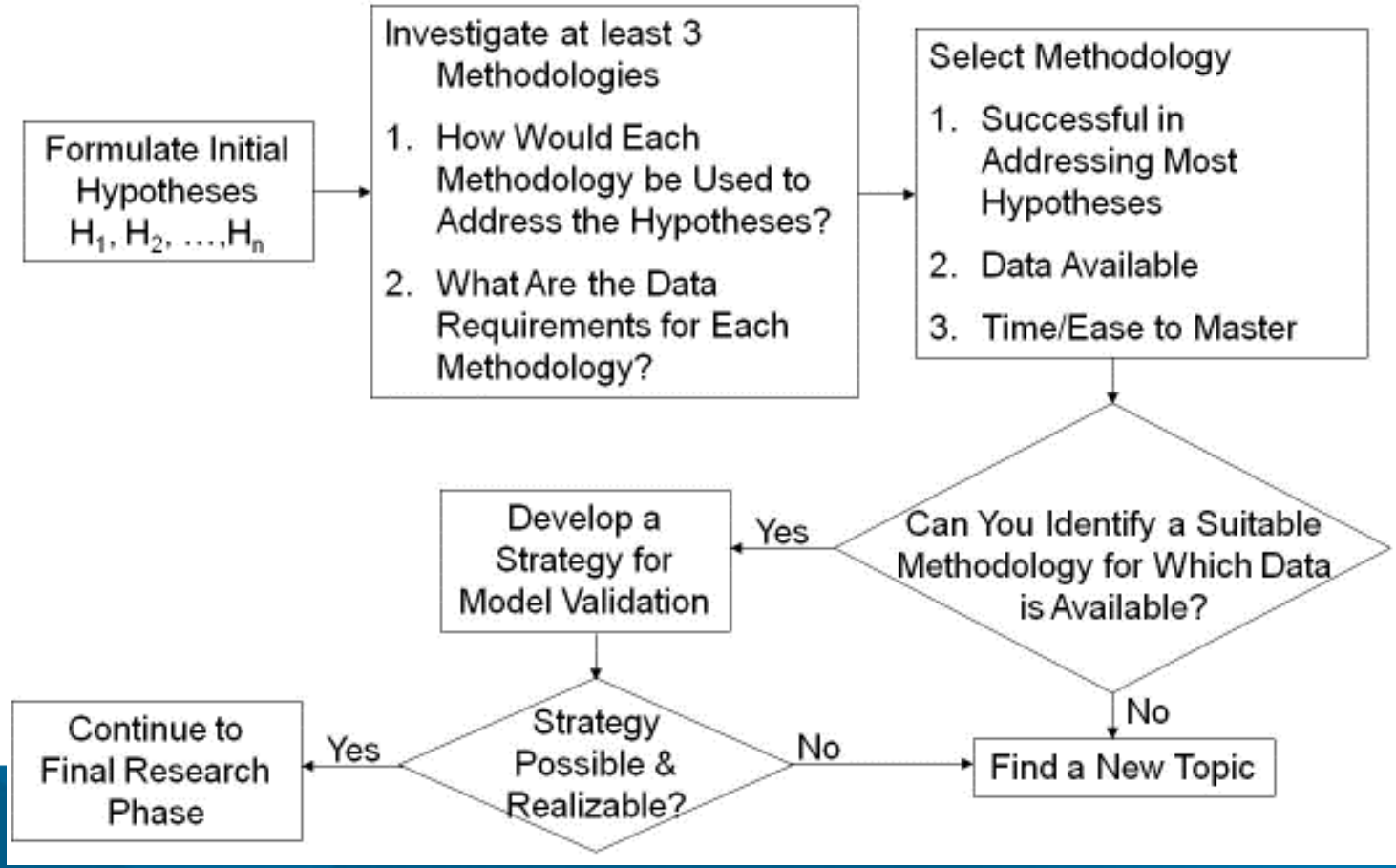
Research Methodology

- This chapter should contain a brief overview of the methodology capturing:
  - Assumptions
  - Ease of use
  - Limitations
  - Required adaptations
- This section should convey your complete understanding of the methodology you use

Research Methodology – Why Is It Needed?



Research Methodology – How Is It Selected?



Research Methodology – Methodology Map

- How does the methodology relate to the research questions and hypotheses?

Research Hypothesis 1	Research Hypothesis 2	Research Hypothesis 3

Research Methodology – Methodology Map

Methodology 1

Research Hypothesis 1 | Research Hypothesis 2 | Research Hypothesis 3

Research Methodology – Methodology Map

Methodology 1

Methodology 2

Research Hypothesis 1

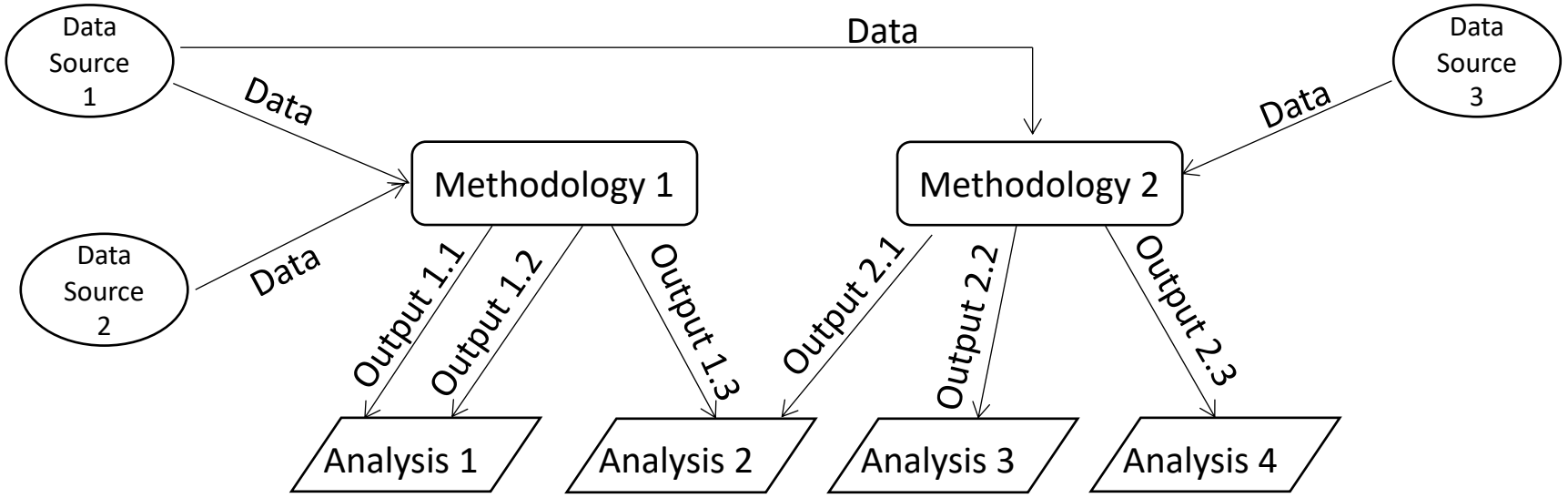
Research Hypothesis 2

Research Hypothesis 3

Research Methodology – Methodology Map

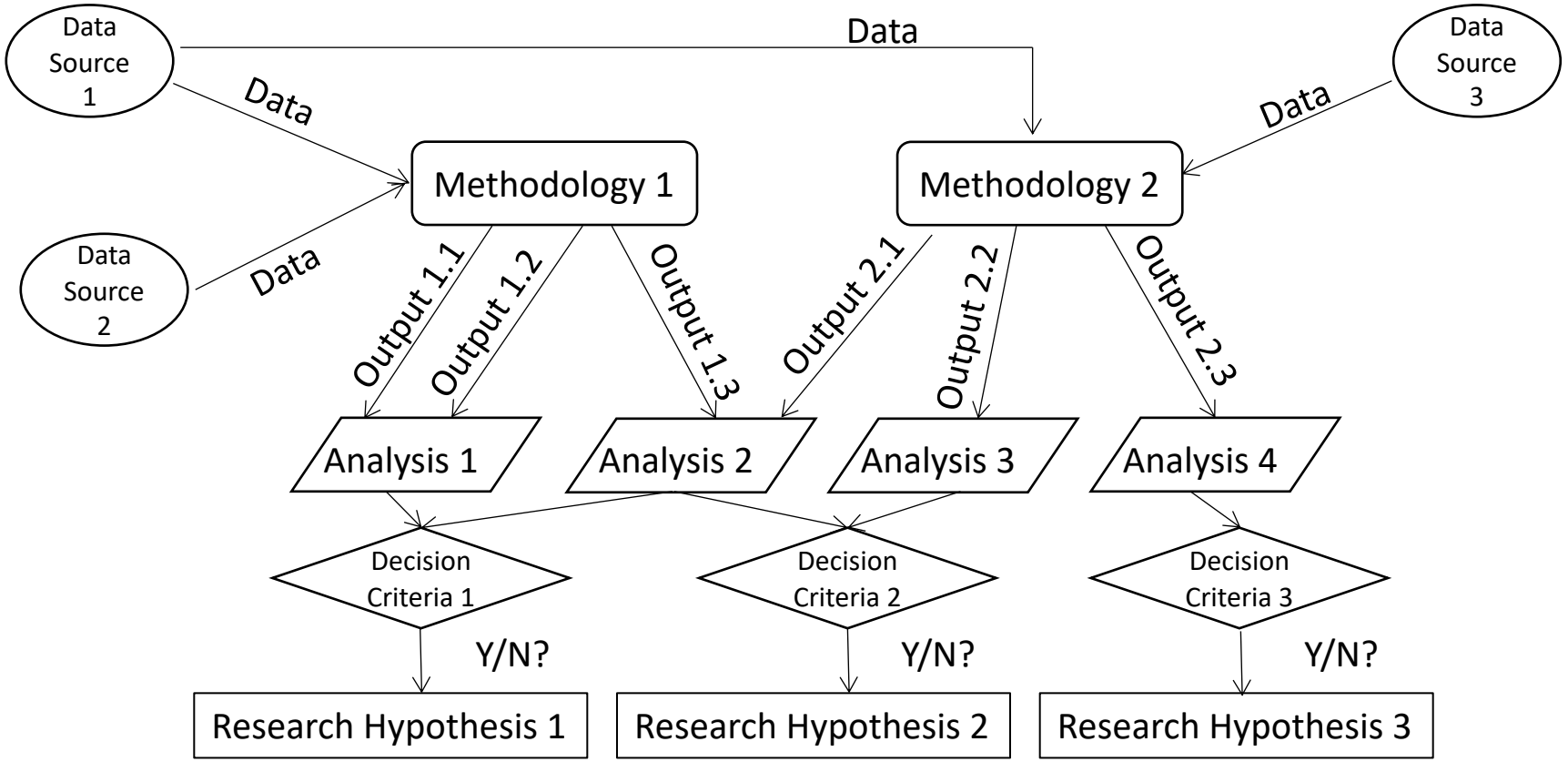
Research Hypothesis 1 | Research Hypothesis 2 | Research Hypothesis 3

Research Methodology – Methodology Map

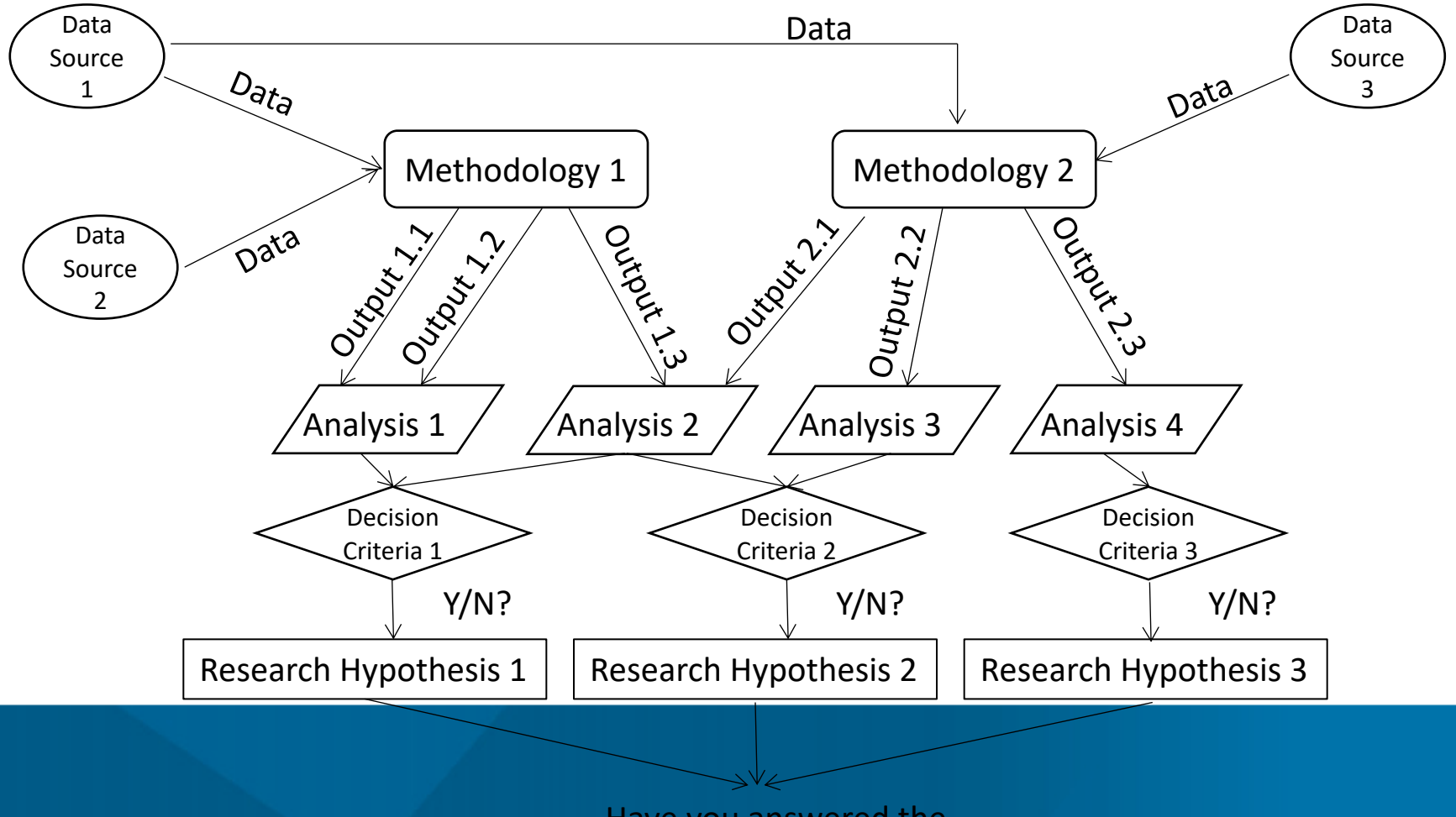


Research Hypothesis 1	Research Hypothesis 2	Research Hypothesis 3

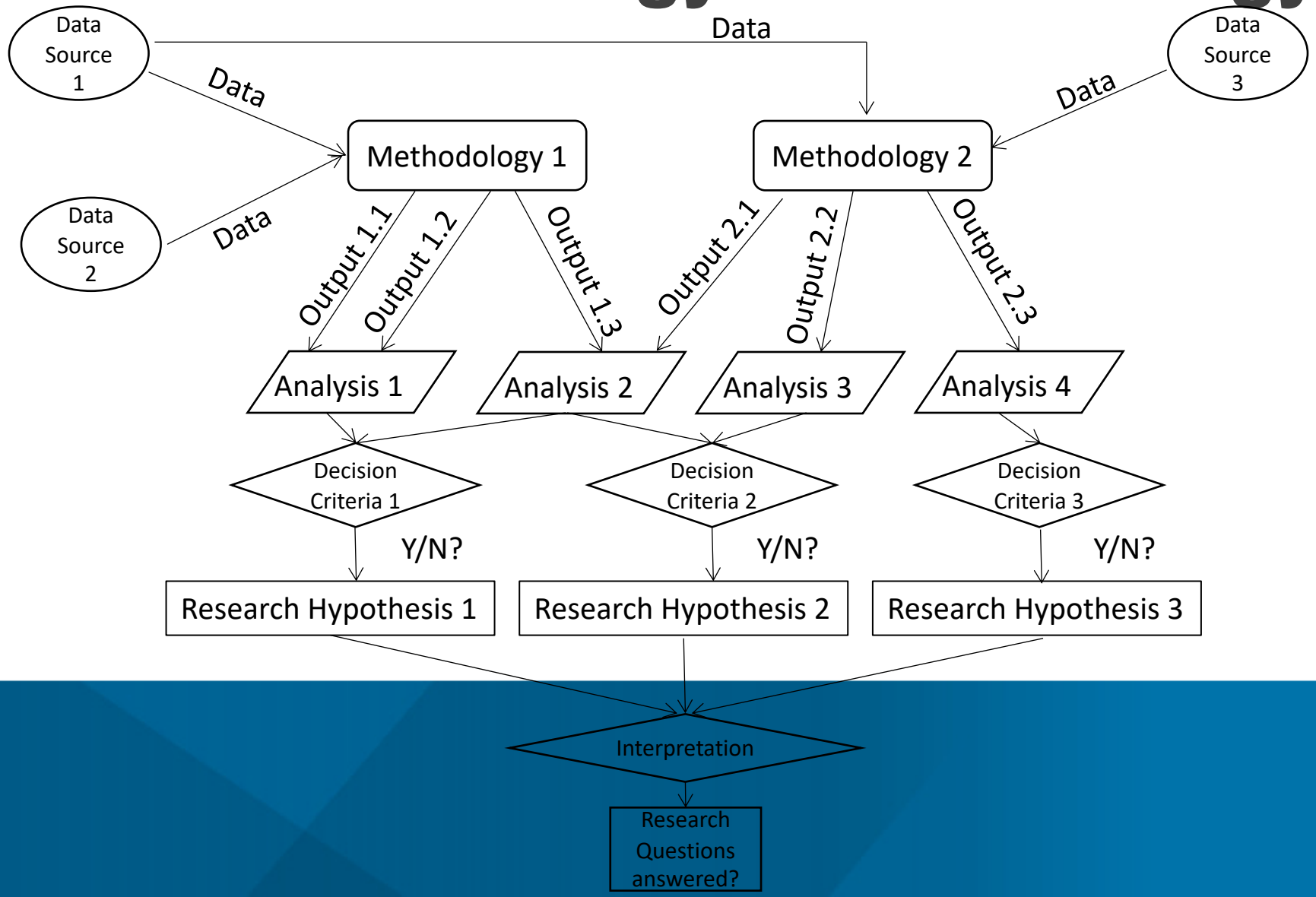
Research Methodology – Methodology Map



Research Methodology – Methodology Map



Research Methodology – Methodology Map



Research Methodology – Evaluation

Are Data Sources COMPLETELY Defined?

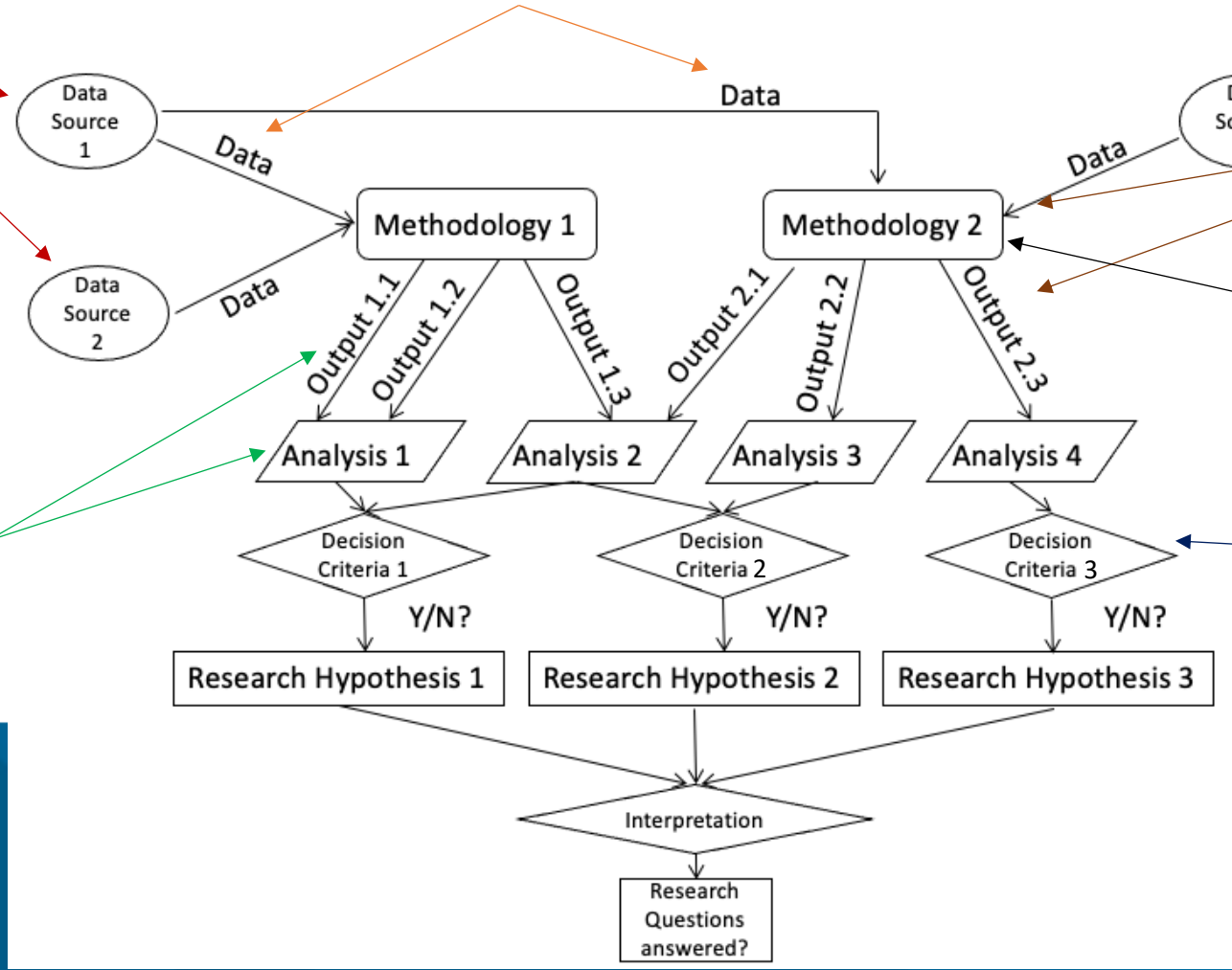
Are Data Types CLEARLY Specified?

Are All Data Flows Clear?

Is There a Good Selection of Methodologies?

Is it Clear How The Methodology Results in Answering the Research Questions?

Are the Criteria For Answering Research Questions Specified?



Research Methodology – Recommend. Texts

GENERAL RESEARCH METHODOLOGIES

Cooper, D. and Schindler, P., *Business Research Methods,* McGraw-Hill

GRAPHICAL METHODS

- **Influence Diagrams**

  Clemen, R. and Reilly, T., *Making Hard Decisions with Decision Tools,* Duxbury Thompson Learning

- **Bayesian Belief Nets**

  Koski, T. and Noble, J., *Bayesian Networks: An Introduction,* Wiley Software

Research Methodology – Recommend. Texts

SIMULATION

- **Agent Based**

  Wilensky, U. and Rand W., *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo,* MIT Press

- **Systems Dynamics**

  Sternma, J., *Business Dynamics: Systems Thinking and Modeling for a Complex World*, McGraw-Hill

- **MonteCarlo**

  Law, A., *Simulation Modeling and Analysis*, McGraw-Hill

Research Methodology – Recommend. Texts

DECISION ANALYSIS

- **Multi-Criteria**

  Ishizaka, A. and Nemery, P., *Multi-criteria Decision Analysis: Methods and Software*, Wiley

- **Decision Trees and Influence Diagrams**

  Clemen, R. and Reilly, T., *Making Hard Decisions with Decision Tools*, Duxbury Thompson Learning

- **Expert Judgment**

  Cooke, R., *Experts in Uncertainty*, Oxford University Press.

  O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D. and Oakley J., *Uncertain Judgements: Eliciting Experts' Probabilities*, Wiley.

  Meyer, M. and Booker, J., *Eliciting and Analyzing Expert Judgment: A Practical Guide*, SIAM.

Research Methodology – Recommend. Texts

STATISTICAL METHODS

- **Regression**
  Chaterjee, S. and Price, B., *Regression Analysis by Example*, Wiley

  O'Connell, B., *Linear Statistical Models: An Applied Approach*, Duxbury Press
- **Logistic Regression**
  Hosmer, D., Lemeshow, S., and Sturdivant, R., *Applied Logistic Regression*, Wiley
- **Cluster Analysis**
  Afifi, A., May, S., and Clark, V., *Practical Multivariate Analysis*, CRC Press
- **Discriminant Analysis**
  Afifi, A., May, S., and Clark, V., *Practical Multivariate Analysis*, CRC Press
- **Factor Analysis**
  Afifi, A., May, S., and Clark, V., *Practical Multivariate Analysis*, CRC Press

Research Methodology – Recommend. Texts

STATISTICAL METHODS

- **Structural Equation Modeling**
  Schumacker, R. and Lomax, R., *A Beginner's Guide to Structural Equation Modeling*, Routledge
- **Forecasting Techniques**
  Hanke, J. and Wichern, D., *Business Forecasting*, Pearson
- **Bayesian Statistics**
  Stone, J., *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*, Sebtel Press
  Gelman, A. and Carlin, J., *Bayesian Data Analysis*, Chapman & Hall
- **Machine Learning**
  Flach, P., *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, Cambridge

Research Methodology – Recommend. Texts

MATHEMATICAL METHODS

Hillier, F. and Hillier, M., *Introduction to Management Science*, McGraw-Hill

SURVEYS

Buckingham, A. and Saunders, P., *The Survey Methods Workbook: From Design to Analysis*, Polity Press

Results

- Lead the reader through your story and to your findings and conclusions using graphics as an enhancement.
- Do not just use graphics without explaining what they represent and how they support your conclusions.
- Cover the findings and insights from the EDA step, such as mean, standard deviation, minimum, maximum values, and distributions for each numerical feature in the dataset.
- Compare the results of the various methods examined and provide the final performance metrics of each method.

Praxis Structure – Results

- This chapter demonstrates the output of the steps described in the methodology chapter. This chapter highlights the results accomplished after each step of the methodology followed in the praxis, but it does not discuss how those results relate to the research questions or hypotheses. This chapter includes, but is not limited to, the following:
  - **Descriptive Statistics:** covers the findings and insights from the EDA step, such as mean, standard deviation, minimum, maximum values, and distributions for each numerical feature in the dataset.
  - **Visuals:** charts, tables, and any other visual representations of the work conducted as part of the praxis.
  - **Summary of key findings:** comparing results of various methods examined, final performance metrics of a model, etc.

Discussion & Conclusion

- Summary of what you did
- What was accomplished?
- What was not accomplished?
- What are the limitations of your findings?
- Relate your findings to your research
  - You must close the loop on this
- What were you not able to accomplish?
  - Why?
  - What is left for others to do?

Praxis Structure – Discussion & Conclusion

- This chapter outlines how the findings of the study are connected to the research questions and hypotheses. It also draws upon the literature review for any comparison against previous solutions developed to a similar problem. Sections covered in this chapter include, but are not limited to, the following:
  - **Research Hypotheses:** discusses whether the results of the praxis, covered in chapter 4, confirm and prove each hypothesis.
  - **Conclusions:** encapsulates the main points that the researcher intends for the reader to retain.
  - **Contributions to the body of knowledge:** covers the novel contributions from this research to the body of knowledge.
  - **Limitations and future work:** cover items outside the scope of the praxis that could be worked on in future research.

HW #4 – Final Praxis Proposal

- Using the template provided on Blackboard, provide the requested information about the dataset(s) used in your praxis.
- Paste in your slides from HW #3 where it is stated.
- Be sure to color any updates you made to your slides from HW #3 in red.
- Be sure to follow the template as is without making any changes to its structure or organization.
- Be sure to name your submitted assignment as follows: **lastName_firstName_HW#4_SEAS_8599.pptx**

Next Steps

- Come to office hours with any questions you may have: **Wednesday and Thursday** this week and next, 7:00 pm – 8:30 pm ET.
- Work on your HW #4 (Final Praxis Proposal) and submit it by **11:59 pm ET on Thursday, June 20th, 2024.**
- Present your Final Praxis Proposal slides during class time on **June 22nd and June 29th**.

Thank you!