Systematic Approaches to Diagnosing Data Analytic Problems

Roger D. Peng
Department of Statistics and Data Sciences
University of Texas, Austin

Rice University
October 2022

I know of only two ways to make money in business: One is to bundle; the other is to unbundle.

–Jim Barksdale, Co-Founder, Netscape, Inc.

Data Analysis "Bundle"

- Data analysis often presented as (ideally) a "whole experience" from question to results
- Taught through bundling of science, statistical methods, computational approaches, and real data.
- Focus on experiential learning and working on real data.
- Data science: The ultimate bundle.

Data Analysis "Bundle"

- There can be challenges to presenting data analysis as a bundle of activities
- The "whole data analysis" experience can be heterogeneous for larger groups of students
- Lessons learned with certain datasets may not be appropriate to generalize to other problems
- Can be difficult to reinforce specific skills unless the dataset is narrowly tailored -- can lead to "toy datasets"

Unbundling The Bundle

- What are the core skills of data analysis that are useful across a broad range of problems?

- Can we unbundle these skills from "whole data analytic work" for the purpose of practice and refinement?

- Can we build tools / methodologies to assist us with executing these skills on real problems?

Data Analysis Iteration

- Various "iterative cycles" have been proposed such as the investigative cycle (PPDAC) and the interrogative cycle (Wild & Pfannkuch, 2007)

- Data analysis is a sense-making process that iterates between mental models and data (Grolemund & Wickham, 2014)

- Relatively little statistical theory regarding how to operationalize these models in real data analysis situations

- How can we isolate generalizable skills from these iterative models of data analysis?

Data Analysis Iteration

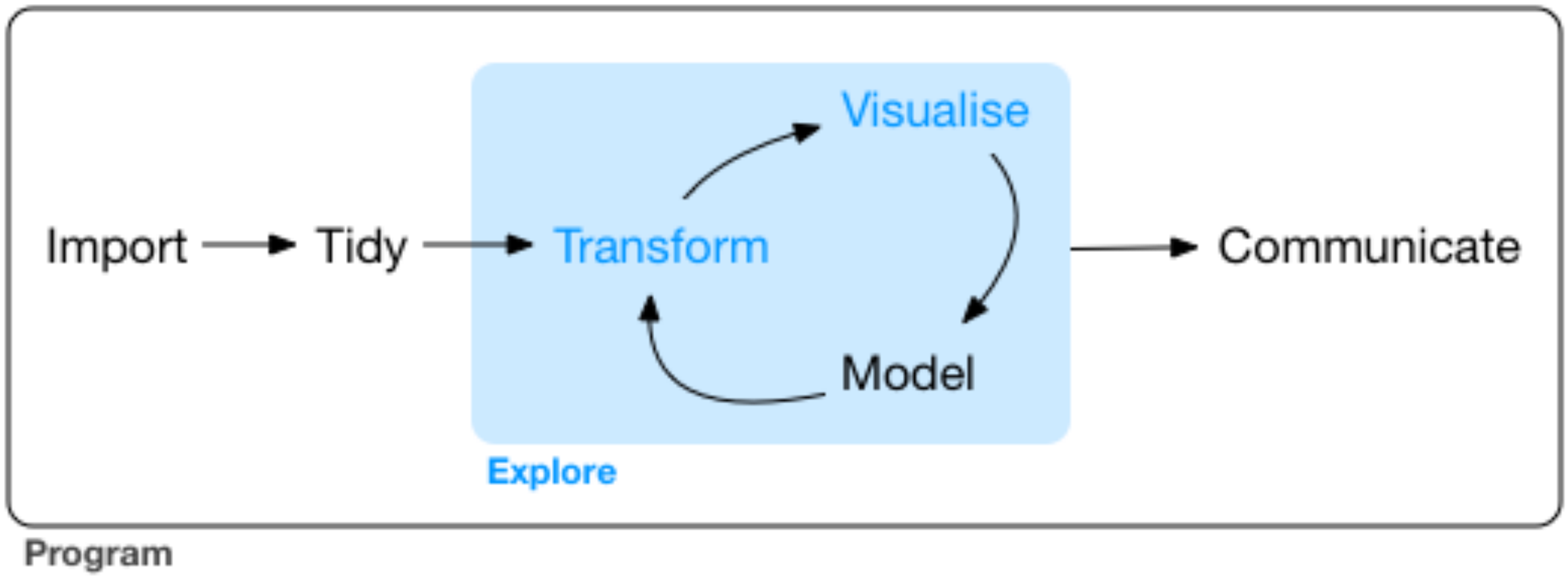
- Can we build a **model for analytic iteration** that can be easily operationalized / generalized across a range of data science problems?

- Can aspects of the iterative process be **isolated and targeted** for teaching / training?

- Can we **build case studies** that target specific skills in a homogeneous manner?

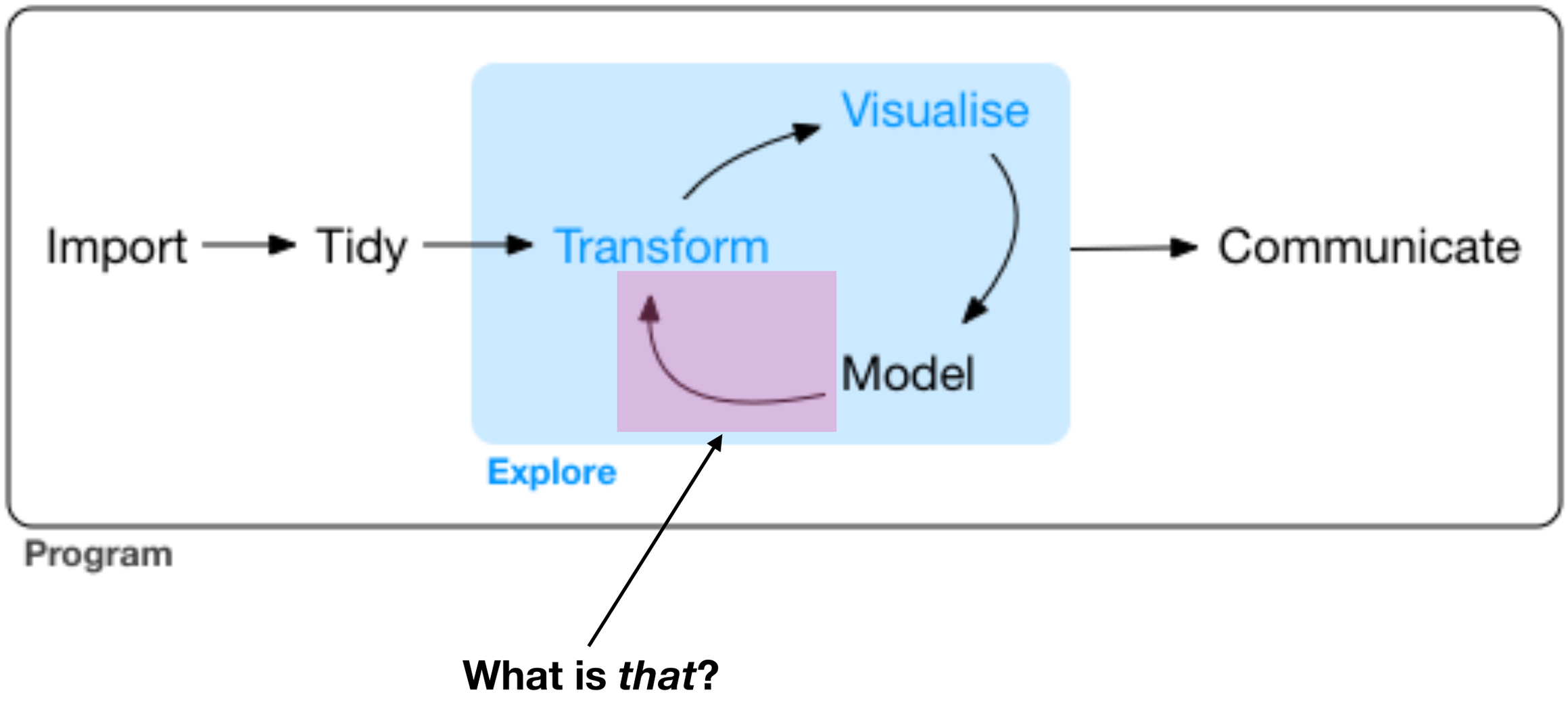
- Can we incorporate the **analyst's perspective** that drives data analytic choices and decisions?

Data Analysis (?)



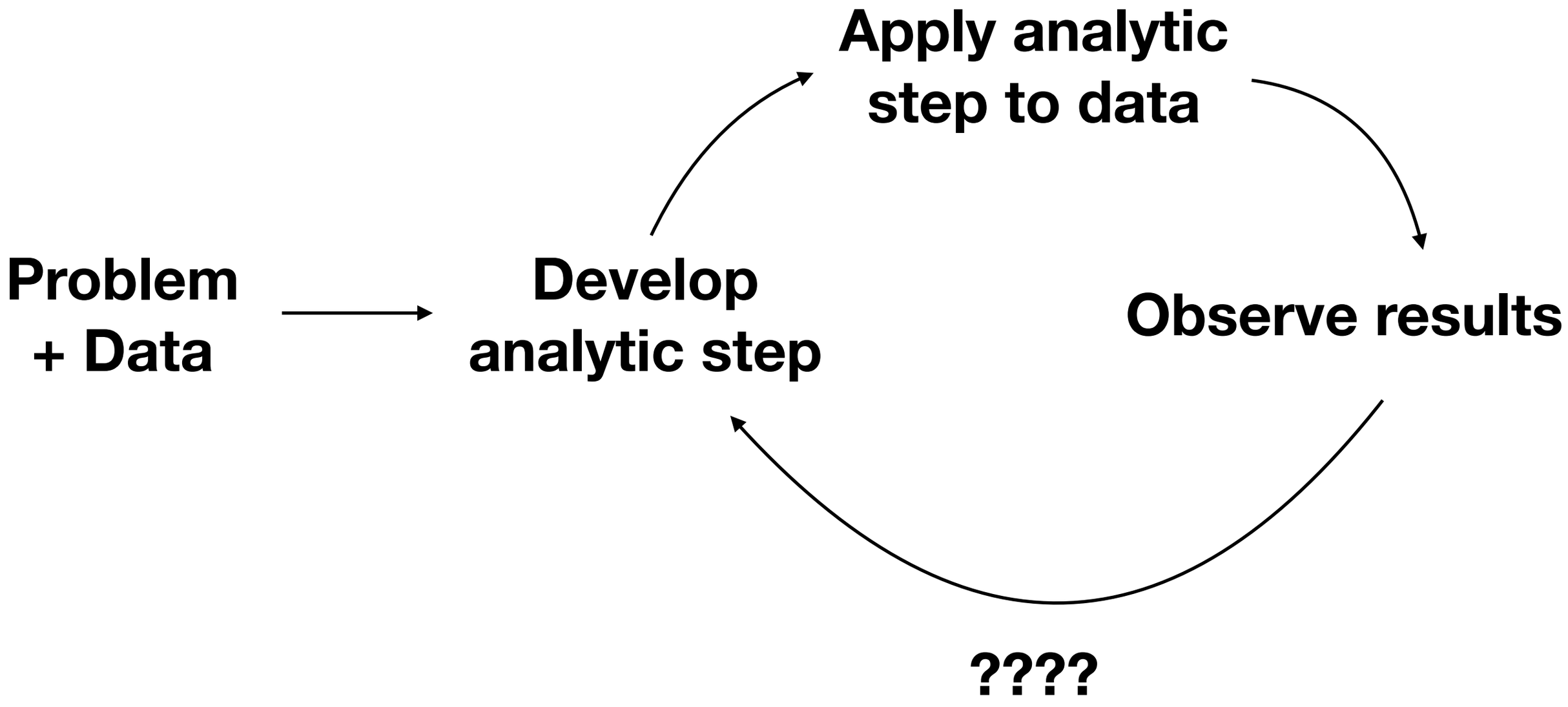
Wickham, *R for Data Science*

Data Analysis (?)

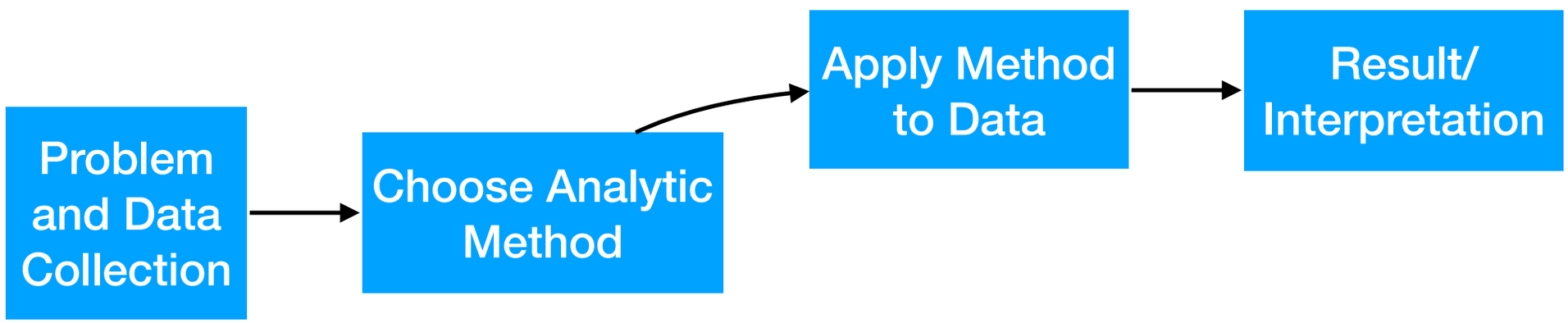


Wickham, *R for Data Science*

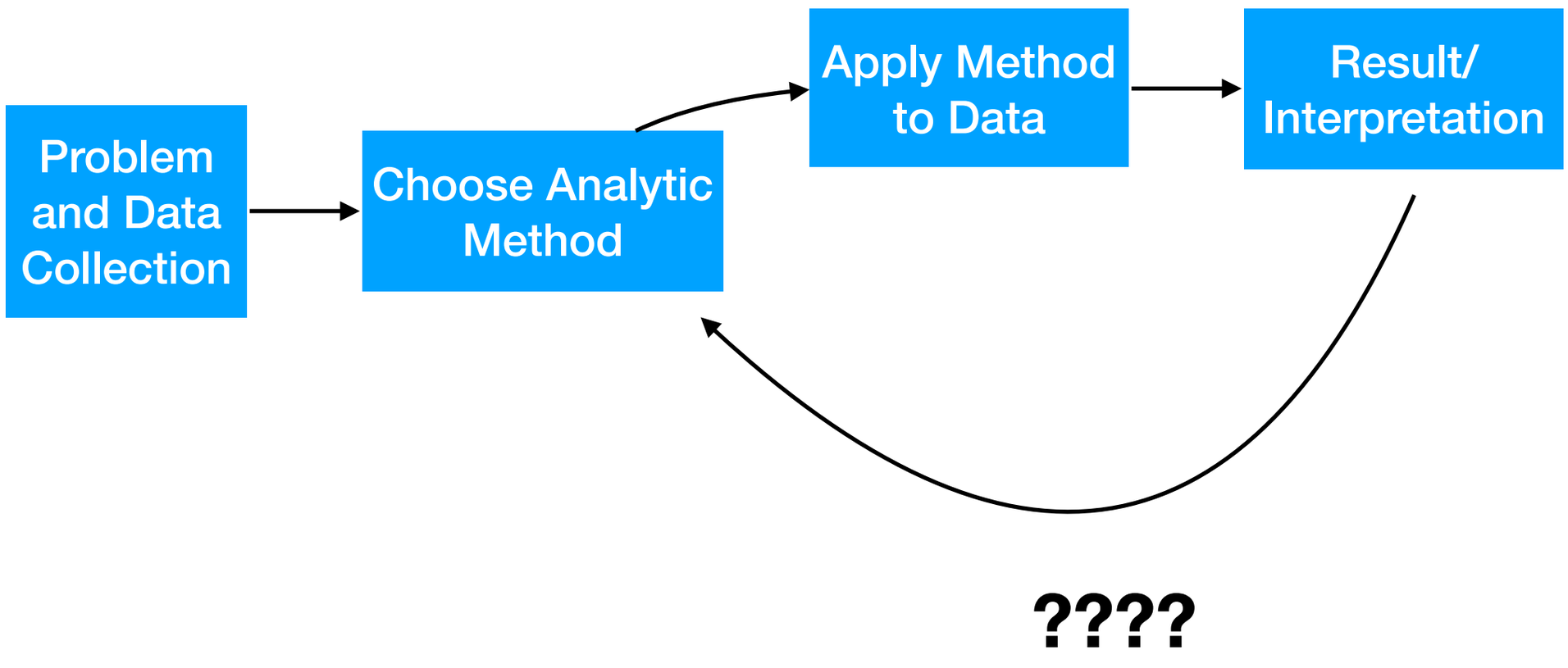
Analytic Iteration (?)



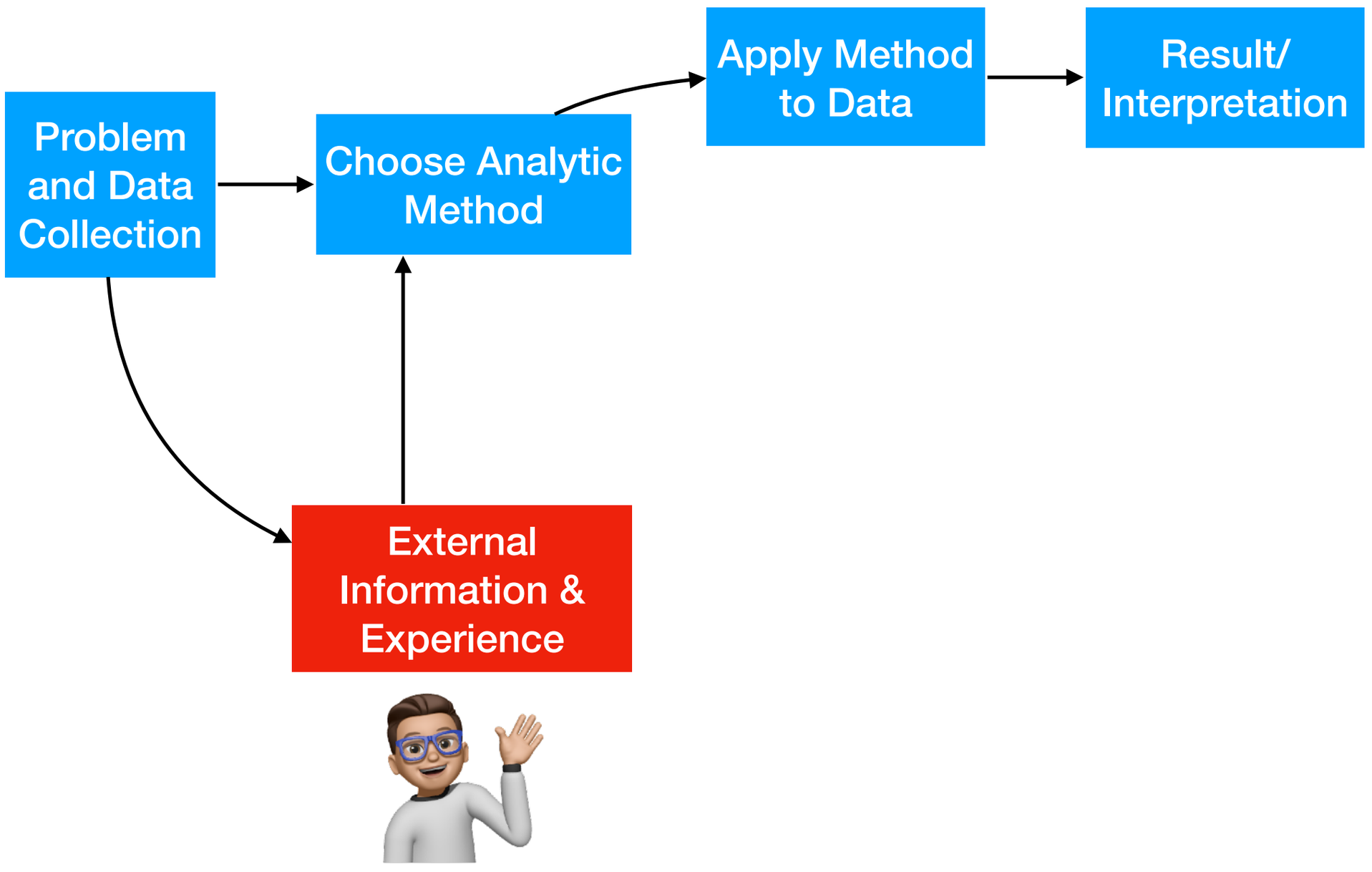
Model for Analytic Step



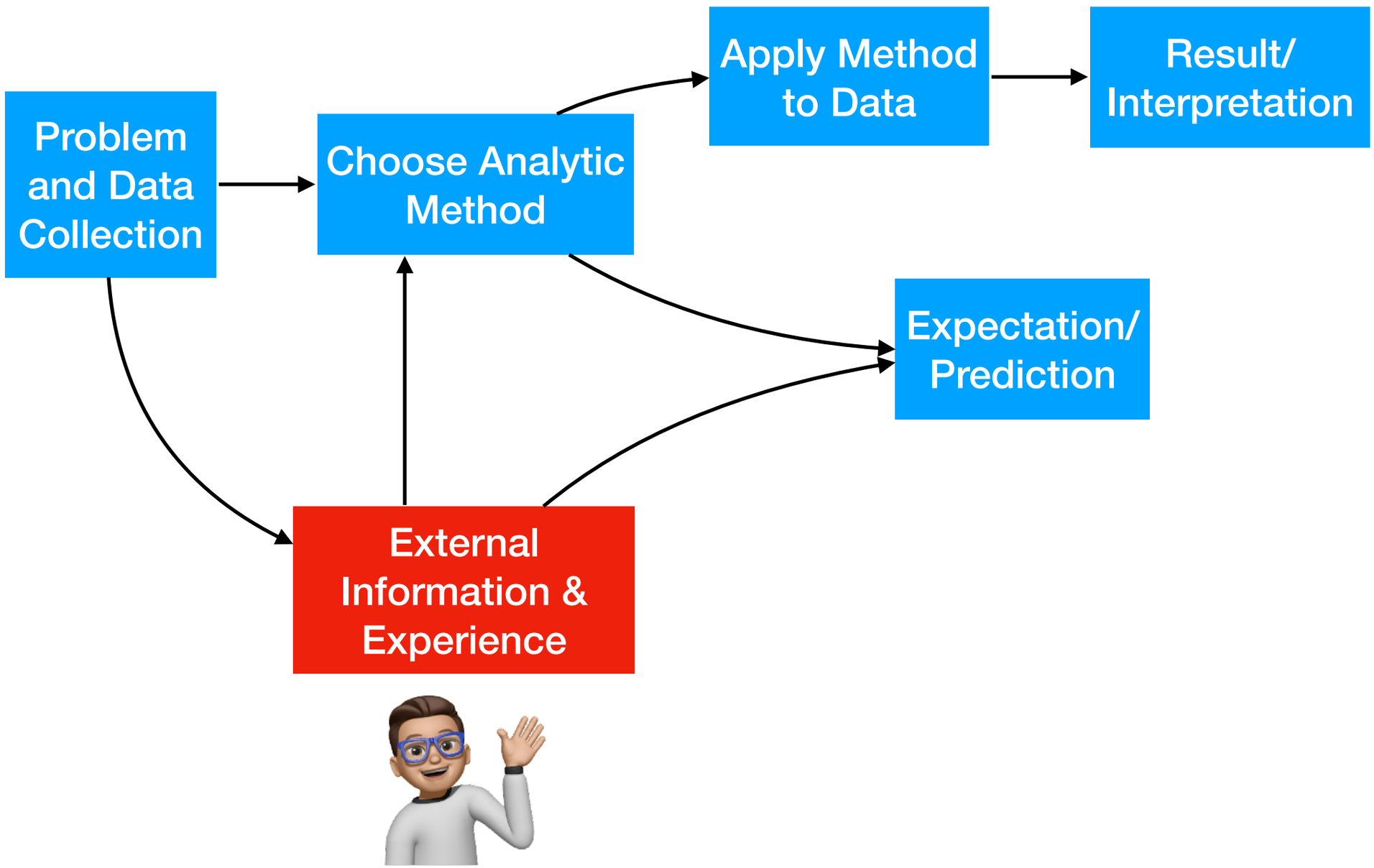
Model for Analytic Step



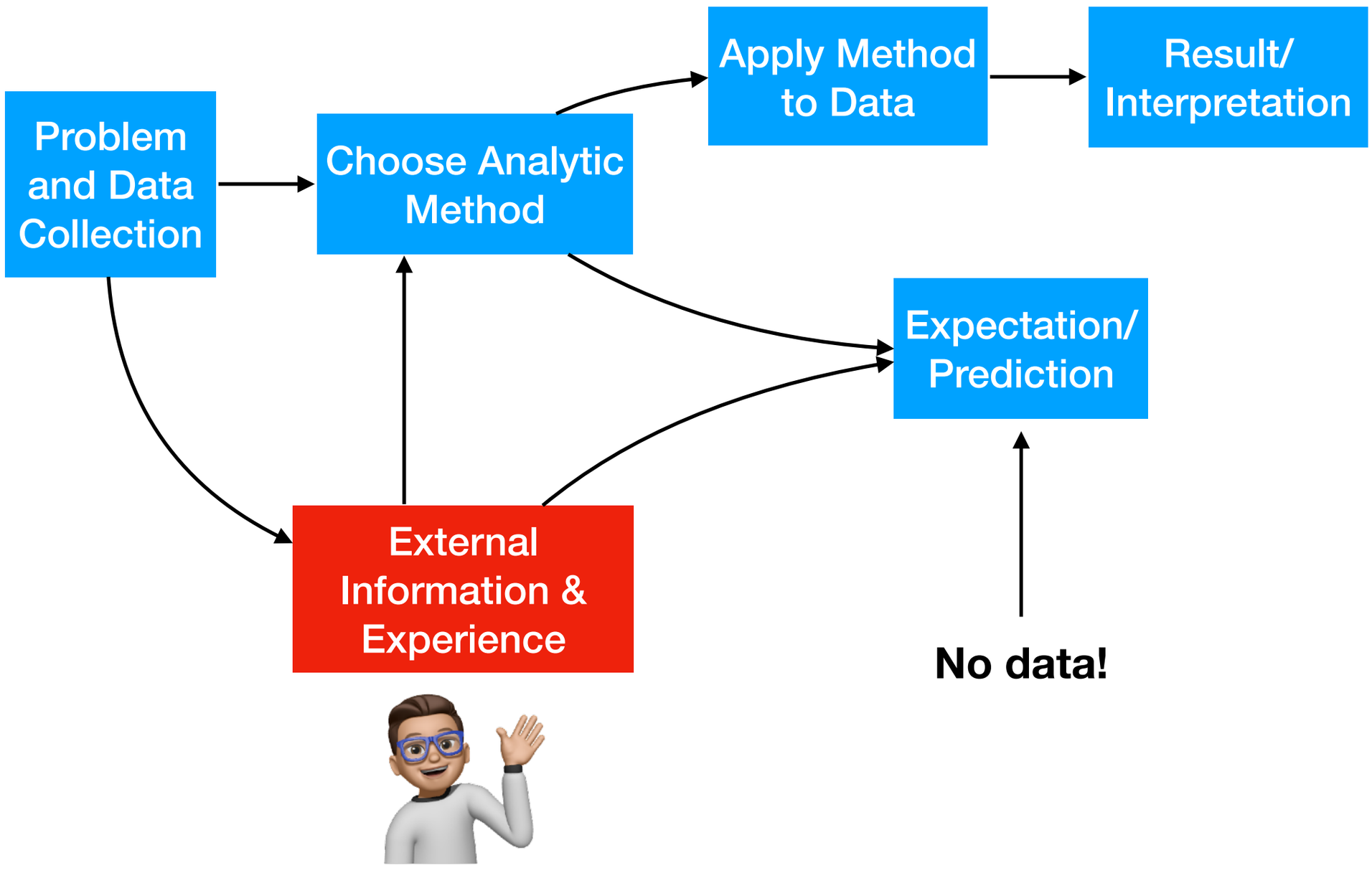
Model for Analytic Step



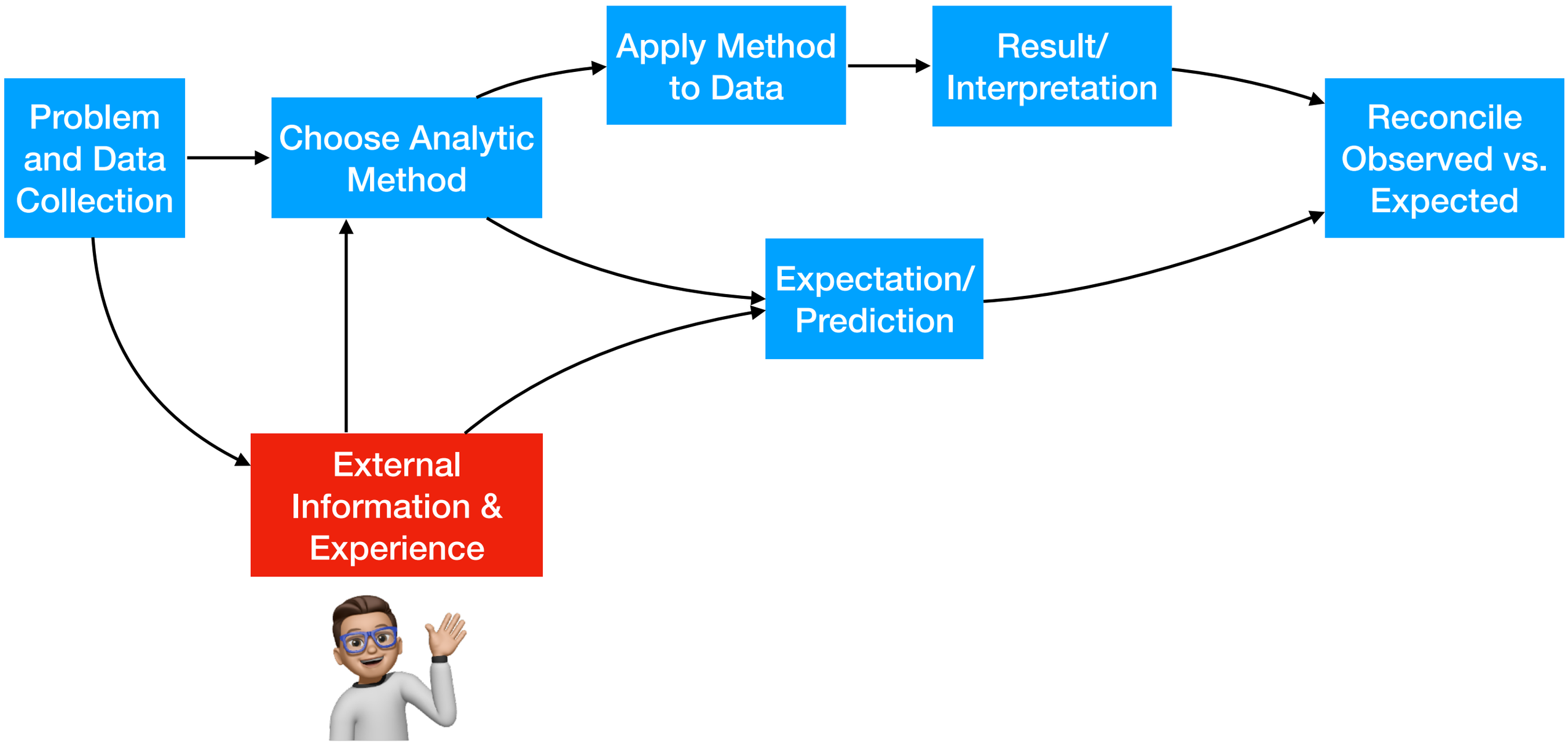
Model for Analytic Step



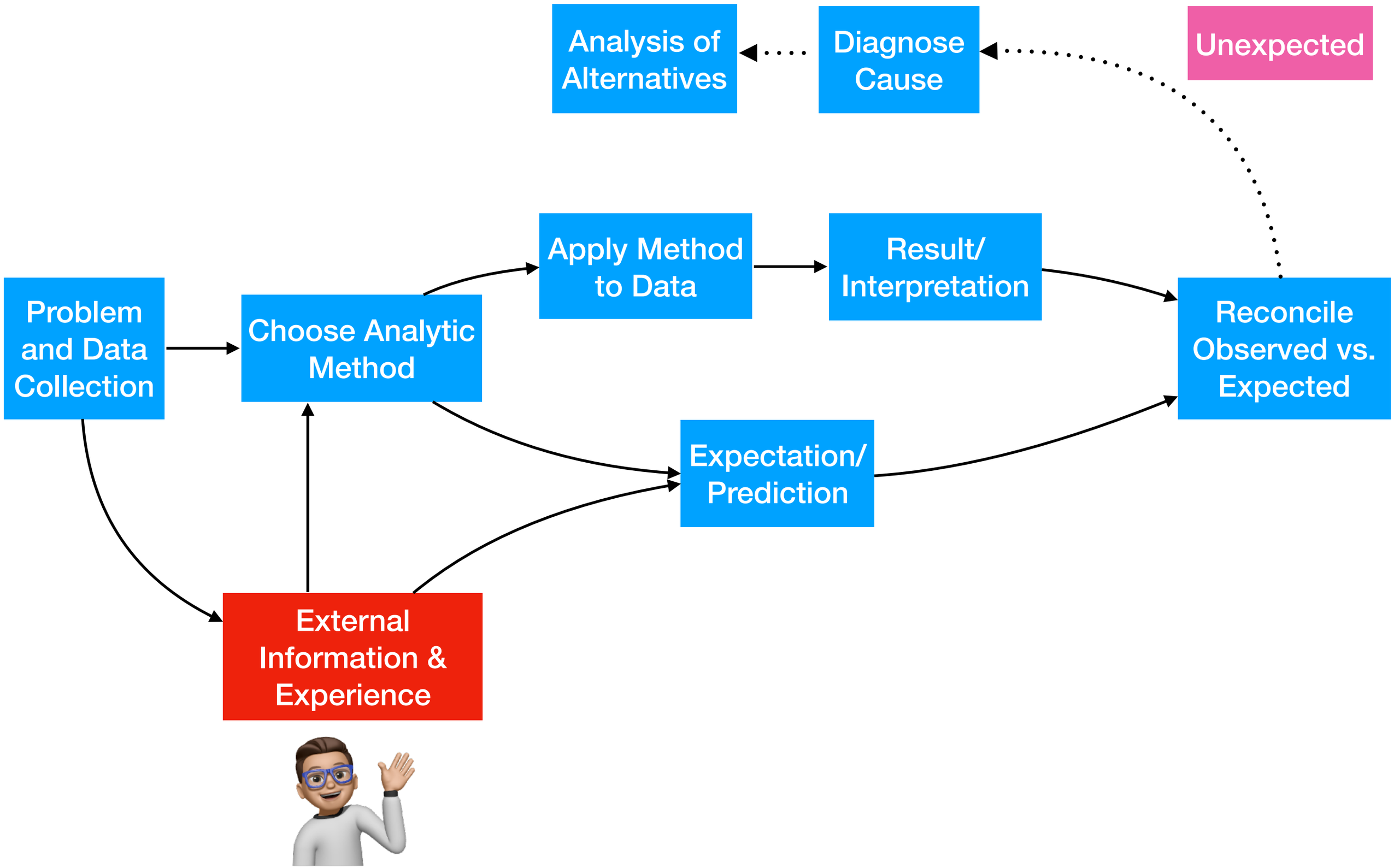
Model for Analytic Step



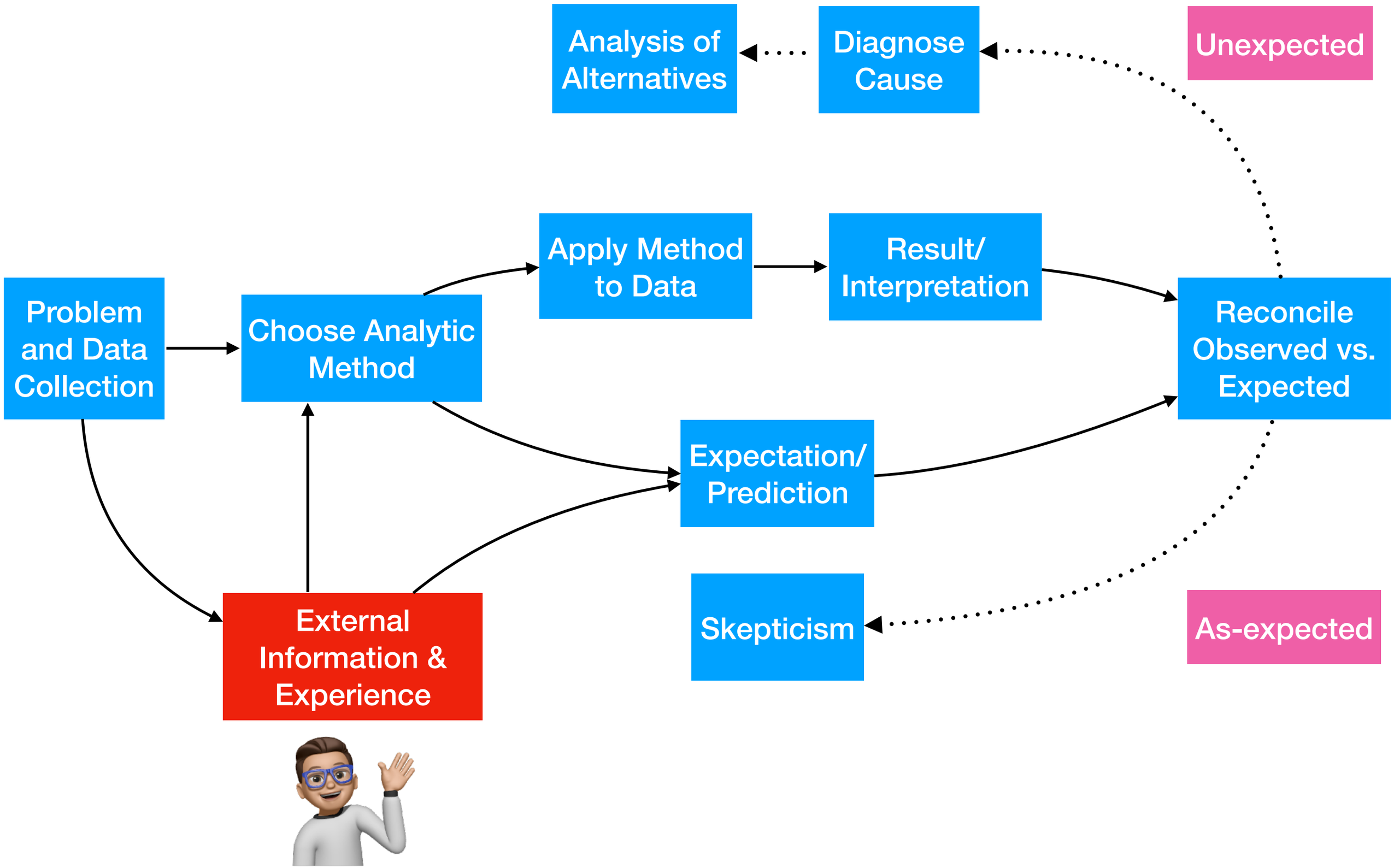
Model for Analytic Step



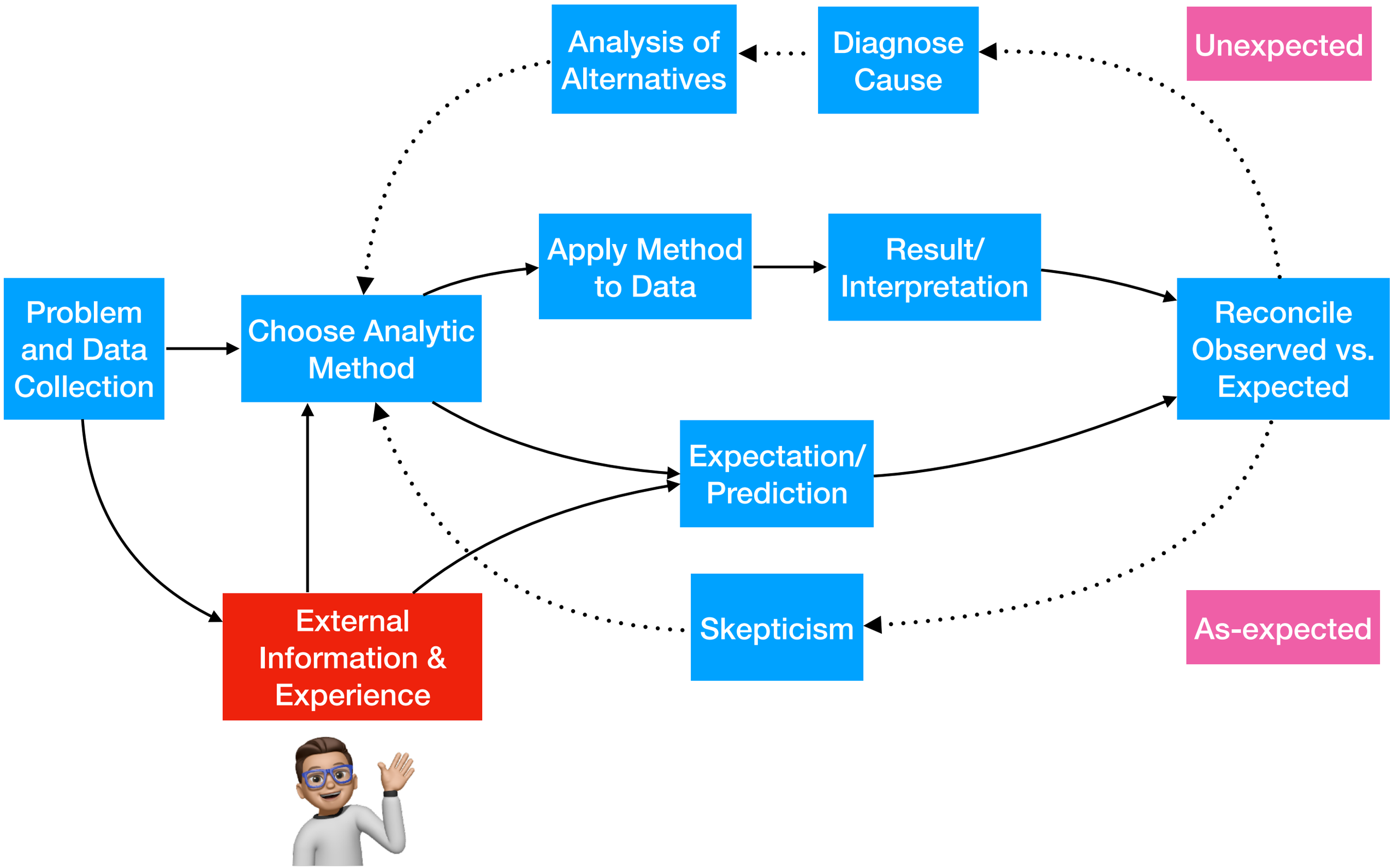
Model for Analytic Step



Model for Analytic Step



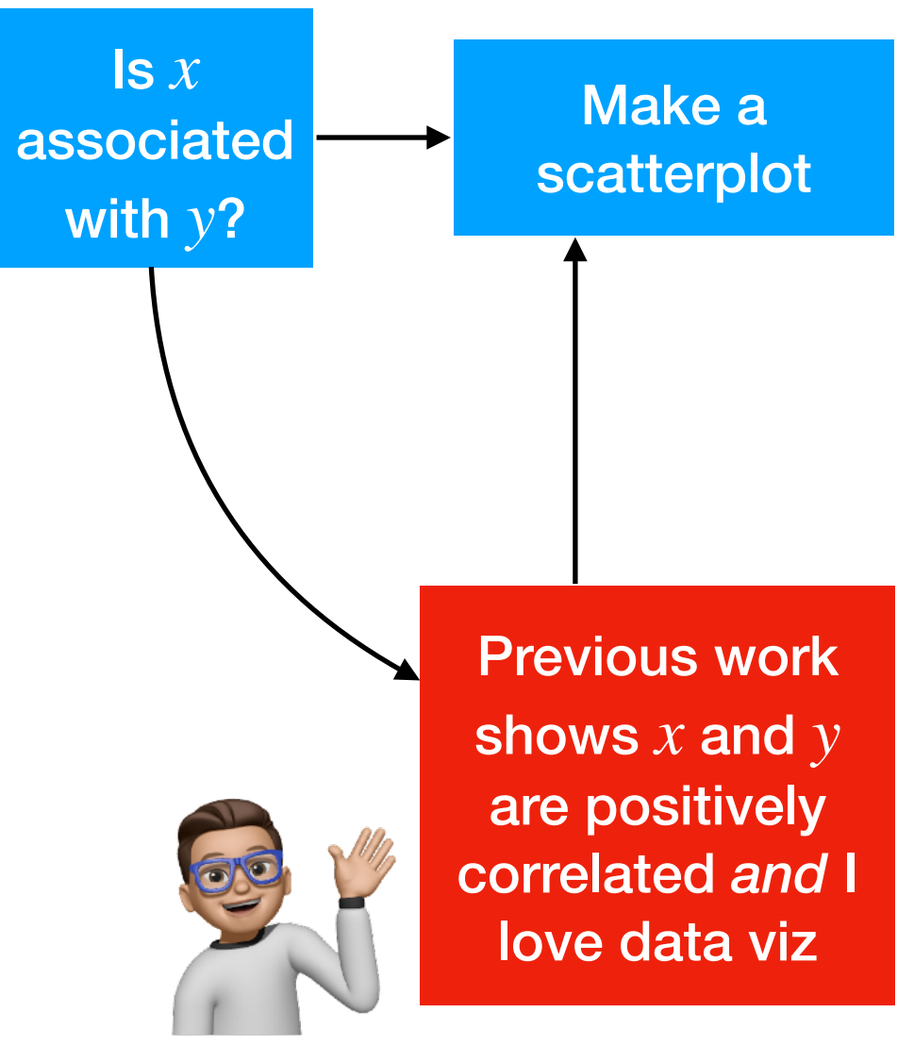
Model for Analytic Step



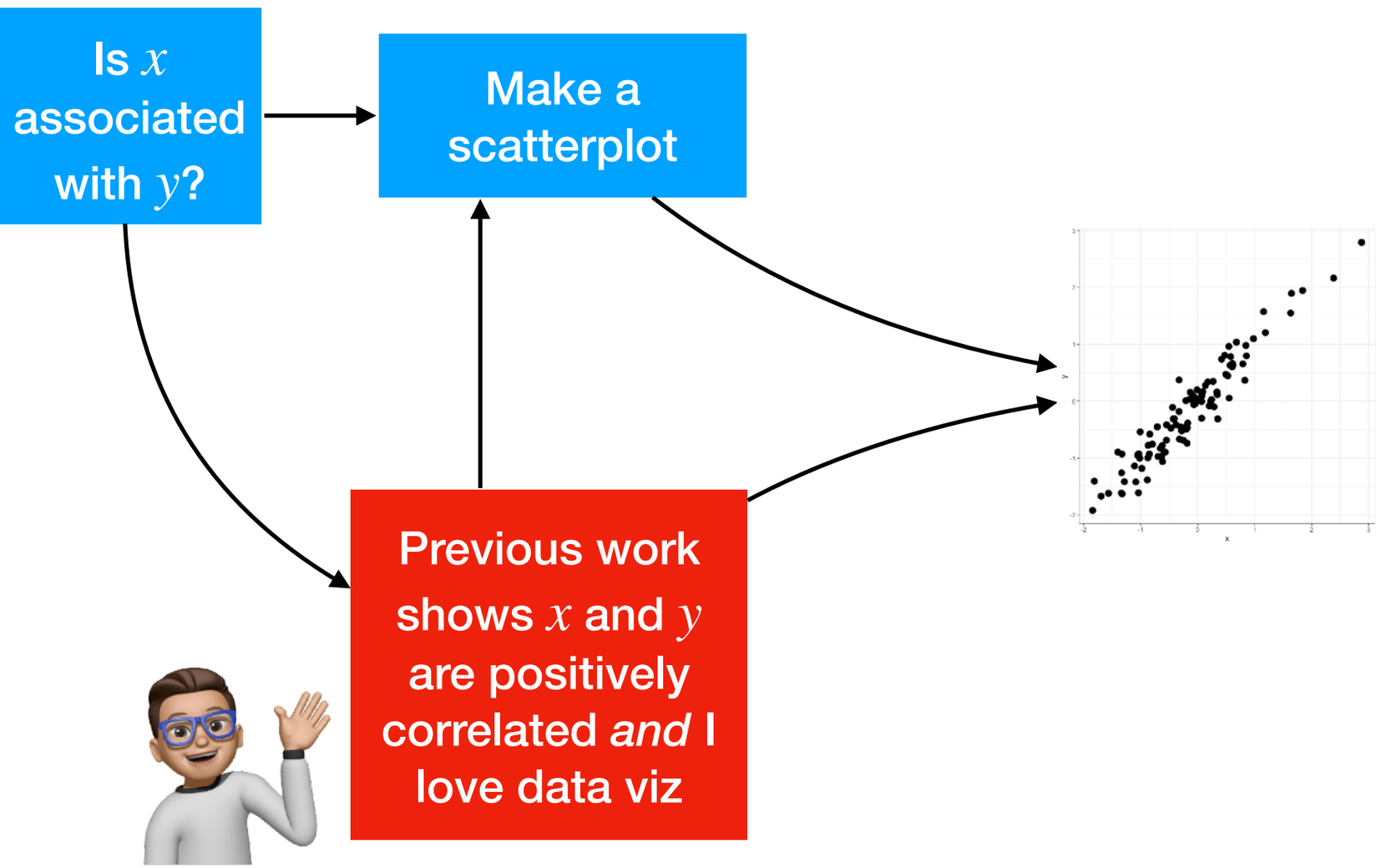
Model for Analytic Step

- Analytic iteration is composed of multiple **analytic steps**
- The output of an analytic step is a **comparison** between the observed result and the expected result
- ➡️ Building the analysis and choosing the methods requires **subject matter knowledge** and **design thinking**
- ⬅️ Working through unexpected results requires **system knowledge** and **diagnostic thinking**

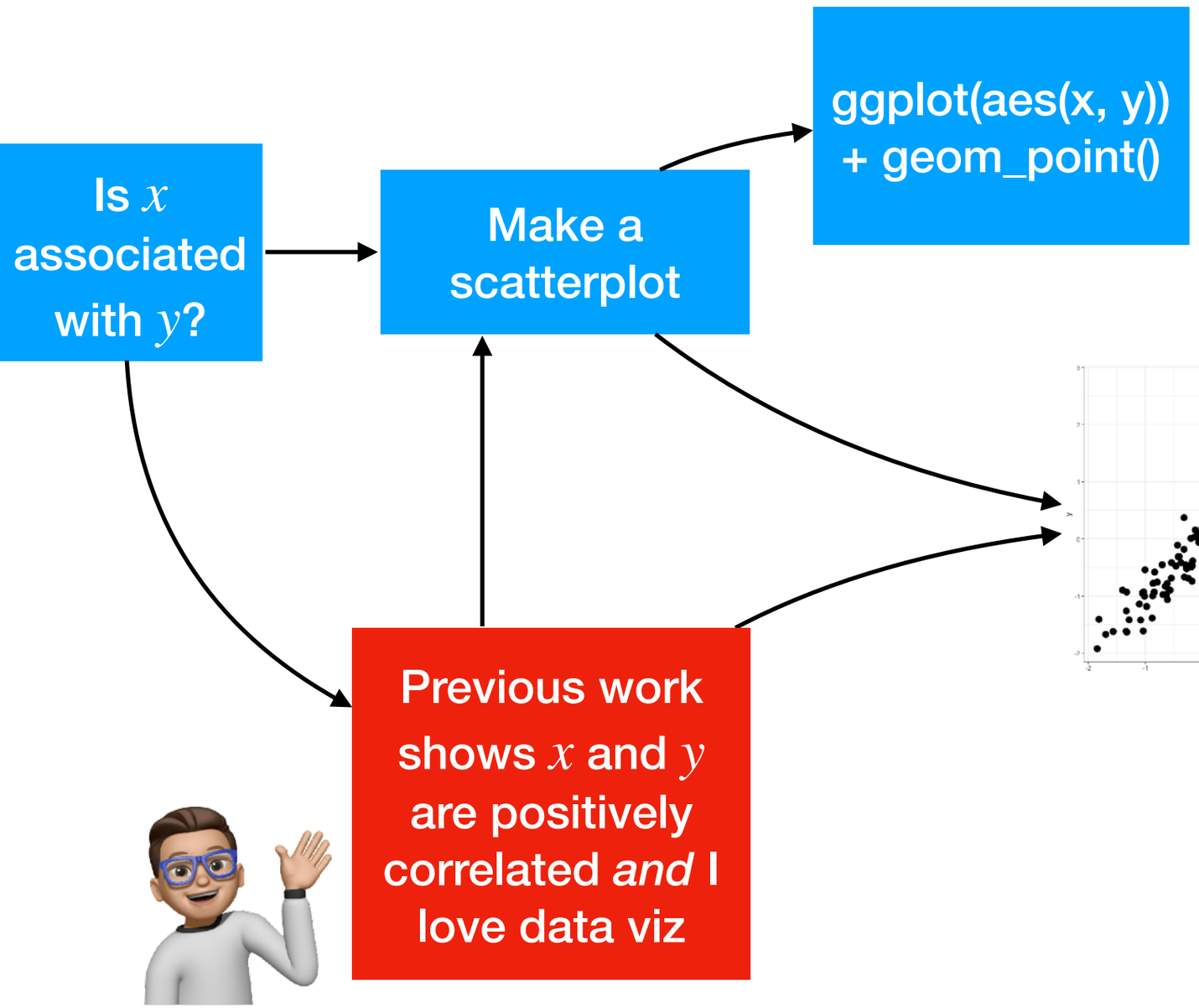
Model for Analytic Step



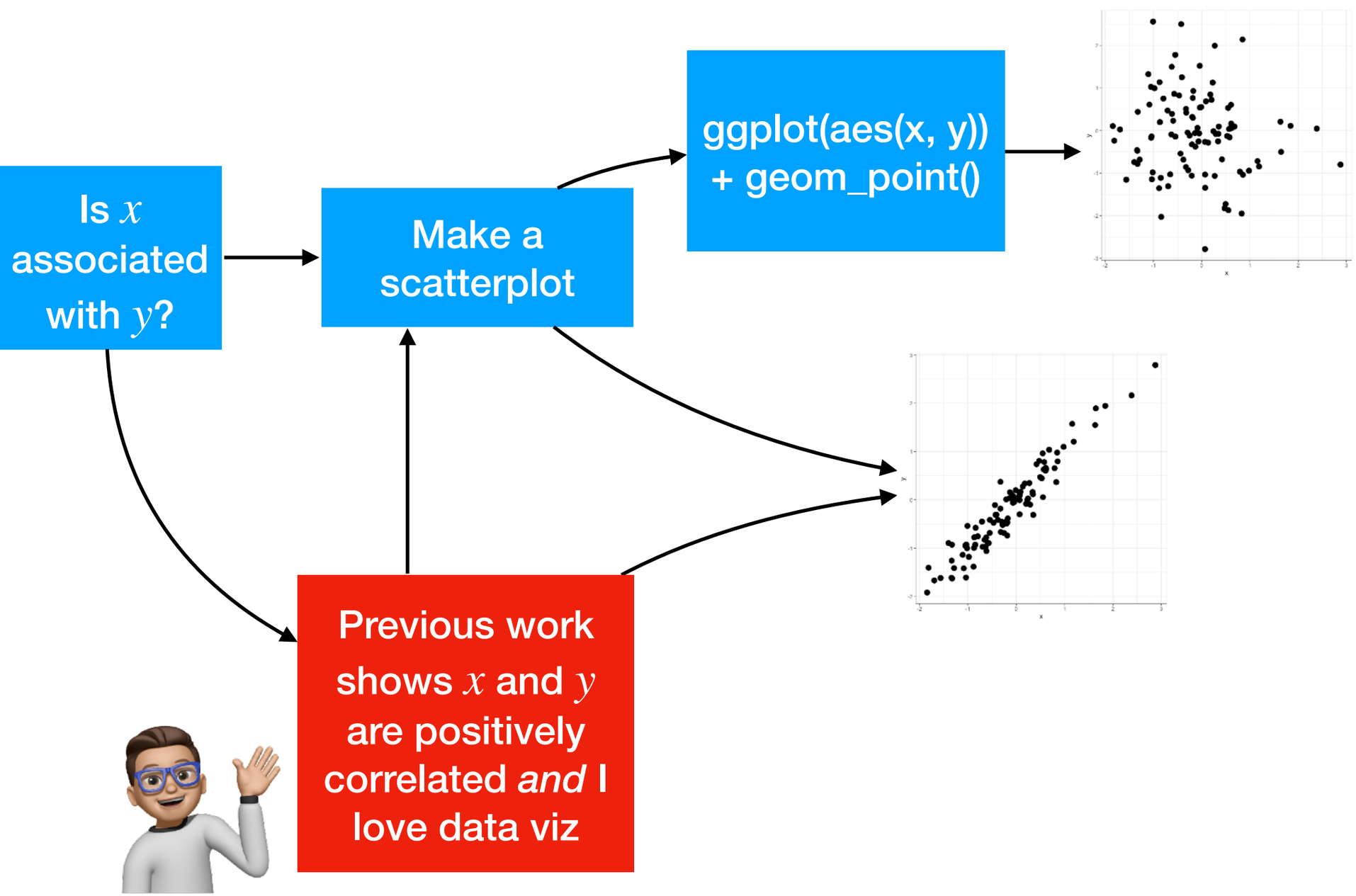
Model for Analytic Step



Model for Analytic Step



Model for Analytic Step

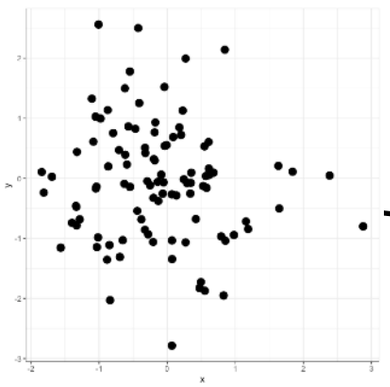


Model for Analytic Step

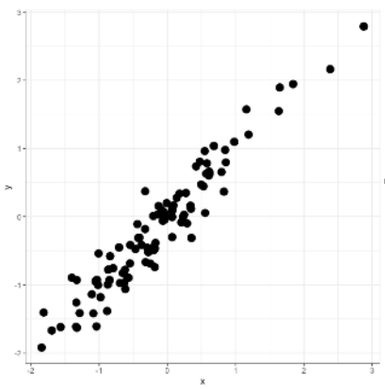
Is x associated with y?

Make a scatterplot

ggplot(aes(x, y)) + geom_point()



Reconcile Observed vs. Expected



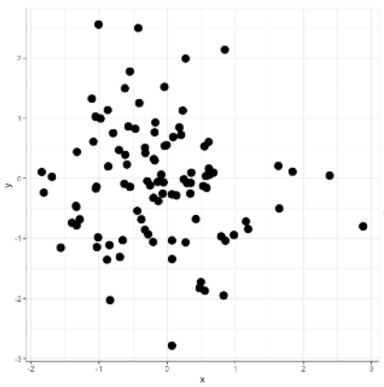
Previous work shows x and y are positively correlated *and* I love data viz

Model for Analytic Step

Is x associated with y?

Make a scatterplot

ggplot(aes(x, y)) + geom_point()

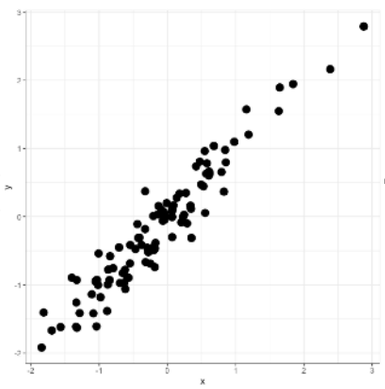


Reconcile Observed vs. Expected

Unexpected

Diagnose Cause

Analysis of Alternatives



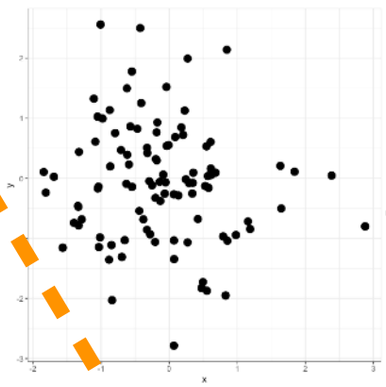
Previous work shows x and y are positively correlated *and* I love data viz

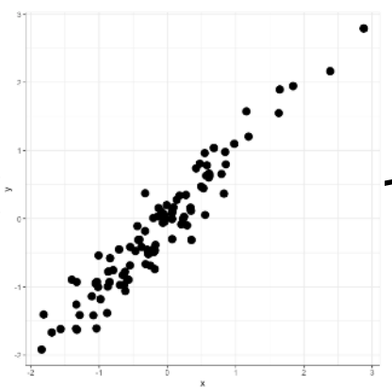
Model for Analytic Step

Is x associated with y?

Make a scatterplot

ggplot(aes(x, y)) + geom_point()





Reconcile Observed vs. Expected

Unexpected

Diagnose Cause

Analysis of Alternatives

Problem with coding?

Problem with input data?

Misinterpreted previous work?

Data more variable than expected?

x and y not actually correlated?

Previous work shows x and y are positively correlated *and* I love data viz

Model for Analytic Step

- The model poses a question at the mid-point of an analytic step and forces an action
- Unexpected outcomes have data that directly contradict expectations
- As-expected outcomes are less directly challenging and therefore require skepticism (a separate topic!)
- Analyst's knowledge, biases, intuition, background are incorporated into the expected outcome

Unbundling Diagnostic Methods

- Diagnostic methods are critical for figuring out the "next step" in an *iterative* analysis process.
    - Do we update our mental model or do we question the data themselves? Or the software? Or something else?
- Requires understanding of the entire *system* that lead to the result
    - Different from producing the result itself, which can rely on abstractions and APIs
- Need to consider multiple hypotheses/explanations for a result and assemble the evidence for/against them
- Does not necessarily require data to practice

Building Case Studies for Anomaly Scenarios

- Create scenarios/case studies for students to read and diagnose
- Expectations are clearly stated (and standardized across students)
- The observed result is designed to be outside of expectations
- No data are used in these exercises; no computing required.
- Goals:
    - Identify possible root causes of unexpected outcomes
    - Develop evidence for/against each cause
    - Propose next steps for investigating the anomaly

Case Format

- A brief introduction and description of the data analysis problem along with some rationale for the statistical methods being applied to the data

- A natural language or code description of the data analysis plan being applied to the data

- A description of the **expected** outcome from the analysis

- A description of the **observed** outcome from the analysis

Example #1

Example #1

- Data $x_1, \ldots, x_n$ are collected (assumed to all be > 0) and log-transformed to create $z_i = \log x_i$. We then compute the mean $\bar{z}$.

Example #1

- Data $x_1, \ldots, x_n$ are collected (assumed to all be > 0) and log-transformed to create $z_i = \log x_i$. We then compute the mean $\bar{z}$.

- R code is given for the computation.

Example #1

- Data $x_1, \ldots, x_n$ are collected (assumed to all be > 0) and log-transformed to create $z_i = \log x_i$. We then compute the mean $\bar{z}$.

- R code is given for the computation.

- Expected outcome: $\bar{z} \in (-\infty, \infty)$

Example #1

- Data $x_1, \ldots, x_n$ are collected (assumed to all be > 0) and log-transformed to create $z_i = \log x_i$. We then compute the mean $\bar{z}$.

- R code is given for the computation.

- Expected outcome: $\bar{z} \in (-\infty, \infty)$

- Observed outcome: $\bar{z} =$ `NaN`

Diagnosis

- Most students identified possible problems with the raw data as a root cause -- most likely way a NaN could appear in the output is if the data contained negative values and the log-transform resulted in NaN

- "...the mean function is unlikely to cause an issue. Therefore, the input of the mean function must have not been numeric."

- Another possibility mentioned by a few students was that there was a NaN value in the raw data that was passed-through by the log function to the mean() function.

- Some students noted that missing values or outliers are sometimes coded as -99 and so the log transform would result in NaN.

- "...I would not expect negative entries for 'x' unless there is some systematic error. Since 'x' here is our raw data, I would ask the collaborators if there is are any potential sources of a systematic error...that would result in negative data entries."

Misdiagnoses

- Many thought that there might be a problem reading in the data as character instead of numeric and that the mean of a character vector would result in NaN (the mean() function in fact produces NA in this case)

- Others thought that there might be zeros in the dataset, which might lead to the log() function producing NaN (the log() function produces -Inf in this case, for which the mean will be calculated as -Inf)

- One student was not clear on the behavior of the mean() function and so developed a number of test cases to determine what kinds of input would generate a NaN output.

Example #2

- Simple linear regression fit to (x_i, y_i) data where there are some missing data in both x and y. Interest is in the slope coefficient β_x. Preliminary data suggest standard deviation of the errors is 5.
- R code for conducting the analysis is presented
- Expected outcome is that $\hat{\beta}_x \in [0,2]$
- Observed outcome is that $\hat{\beta}_x = -3$

Diagnoses

- Most students noted that the unexpected outcome could be a result of incorrect expectations for $\hat{\beta}_x$
- High variability in the data (possibly due to the reduced sample size from removing missing data) could cause a negative value of $\hat{\beta}_x$ to be in the expected range
- Some students noted that if the missing data were not missing completely at random then their removal could alter the estimation of the regression coefficient
- “Human error” causes such as
    - sorting data by a single column in Microsoft Excel without sorting corresponding columns,
    - regressing x on y instead of y on x,
    - extracting the intercept coefficient from the model instead of the slope.

Meta Diagnosis

- Students from different backgrounds (epidemiology vs. biostatistics) appeared to identify different causes of the unexpected data analytic results in the case studies

- Students' personal experiences affected their prioritization of root causes

- Focused on observing if students could identify a potential cause and then specify a follow-up action in order to investigate. Presumably, if the root cause was incorrectly diagnosed, the follow-up action would eventually reveal that.

- Recommendations for follow up were often vague (e.g. "plot the data") and students needed to be encouraged to give specifics

Formalizing Diagnosis

- Can we build a general process / tool that can be useful for diagnosing data analytic anomalies?

- Requirements:
  - A systematic approach that can be taught in a "theoretical" manner (as opposed to experiential)
  - A structured approach that focuses on likely causes and discourages wild guessing
  - Provide alternative hypotheses to investigate in "next steps" of an analytic iteration

Example: Diagnosing An Anomaly

Question

What is the average level of $PM_{2.5}$ air pollution in Baltimore City?

Output

$$\bar{x} = 25$$

Example: Diagnosing An Anomaly

Question

What is the average level of $PM_{2.5}$ air pollution in Baltimore City?

Output

$$\bar{x} = 25$$

Expected outcome

$$\bar{x} \in [8,12]$$

Example: Diagnosing An Anomaly

Question

What is the average level of $PM_{2.5}$ air pollution in Baltimore City?

Output

Expected outcome

$\bar{x} = 25$

$\bar{x} \in [8,12]$

Anomaly

Example: Diagnosing An Anomaly

Question

What is the average level of $PM_{2.5}$ air pollution in Baltimore City?

What is the system that lead to this anomaly?

Output

$$\bar{x} = 25$$

Anomaly

Expected outcome

$$\bar{x} \in [8,12]$$

Data Analytic System

Question

Statistical Theory

Analytic System

Context

Expected outcome

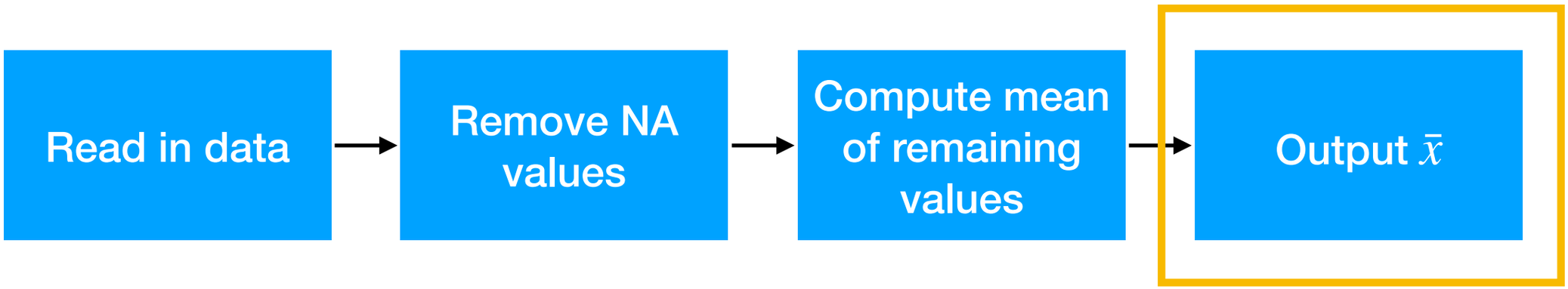
What is the average level of $PM_{2.5}$ air pollution in Baltimore City?

Read in data → Remove NA values → Compute mean of remaining values → Output $\bar{x}$

```
library(tidyverse)
read_csv("dataset.csv") |>
  filter(!is.na(PM)) |>
  summarize(avg = mean(PM)) |>
  print()
```

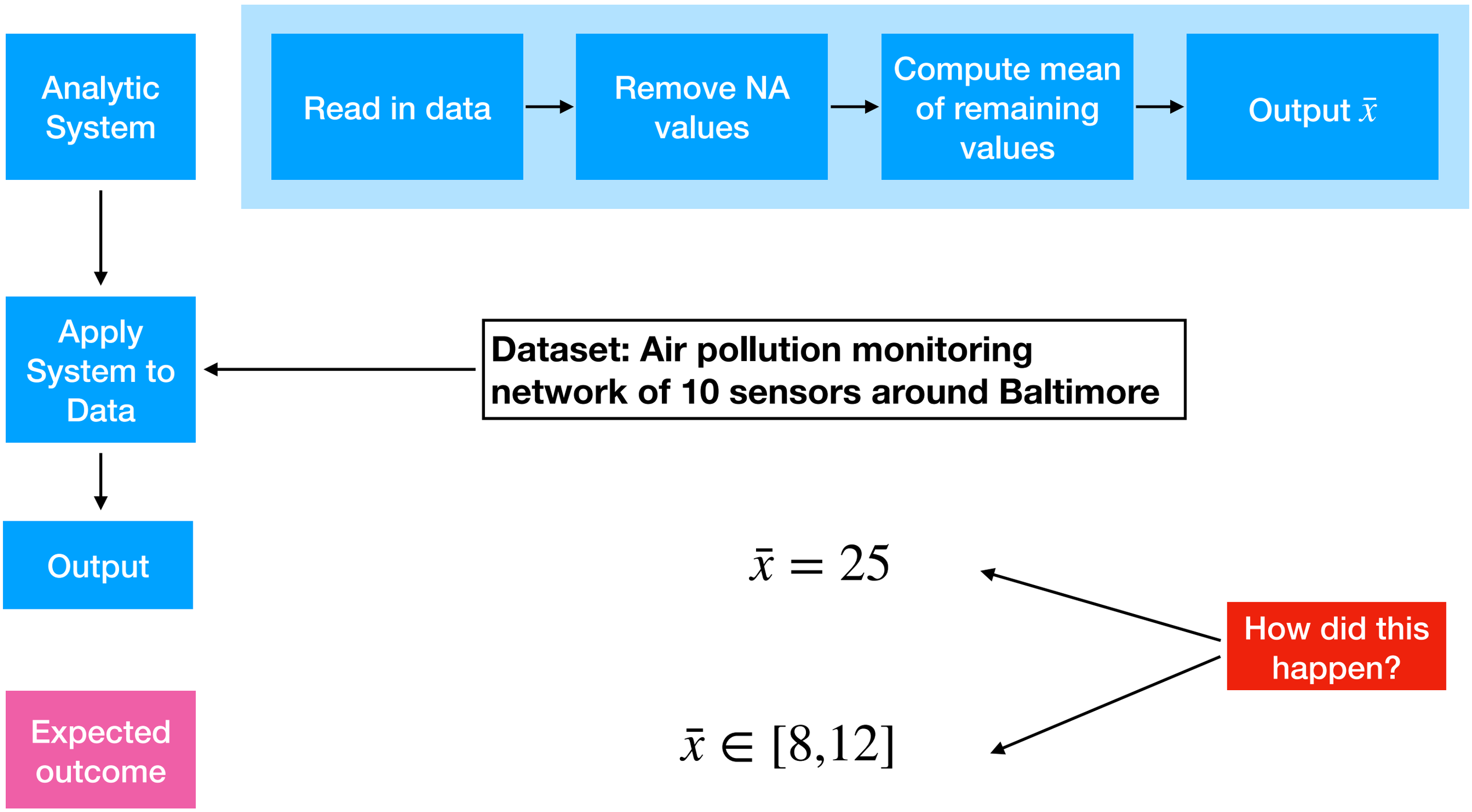
$$\bar{x} \in [8,12]$$

Anomaly Set



- This system has a single **output**: $\bar{x}$
- The **set of expected outcomes** is $[8,12]$
- The **anomaly set** of the system is the set of possible values of $\bar{x}$ that would be considered anomalies if they were observed

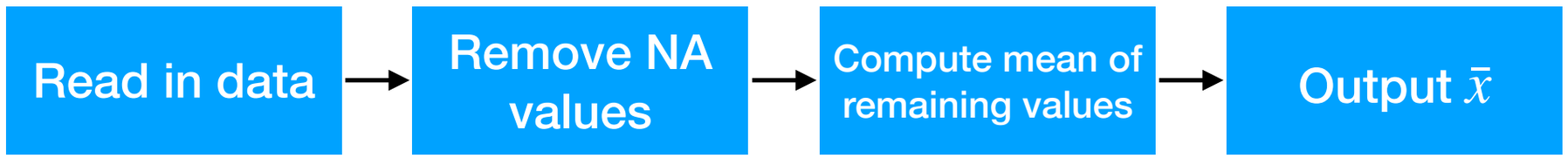
Data Analysis Outcomes



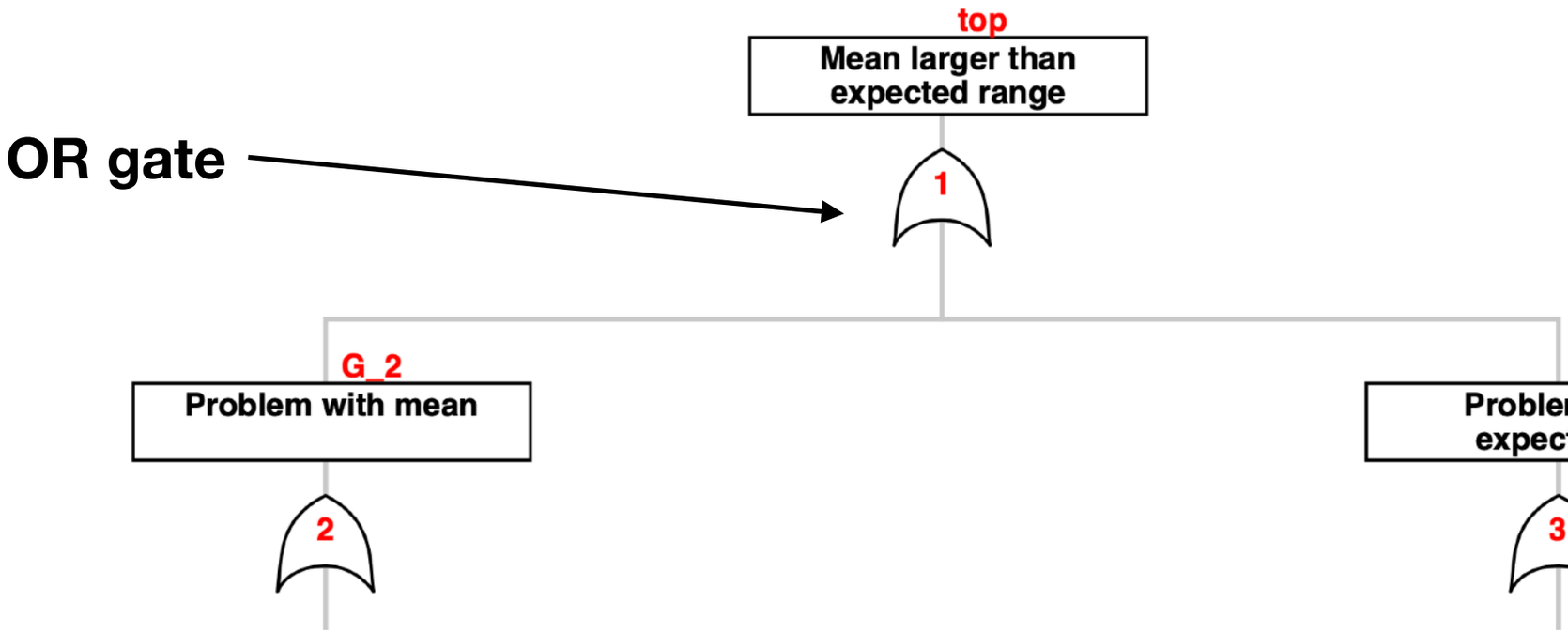
Data Analysis Anomaly

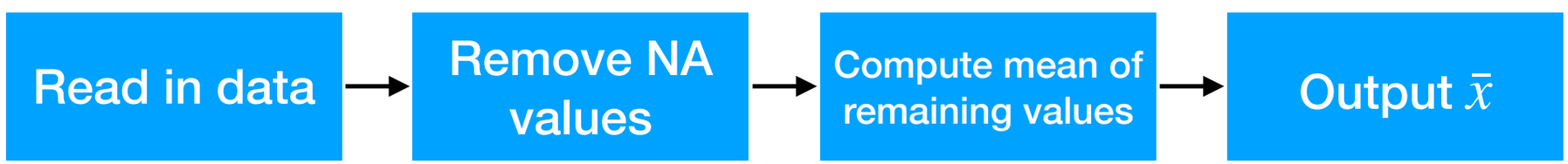
- An **anomaly** is any unexpected outcome that is stated as:
    - *What* was the output that was anomalous?
    - *How* did the output deviate from expectations?
    - *When* did the anomaly occur?
- "$\bar{x}$ *is outside the expected interval* $[8,12]$ *when applied to the monitoring network data.*"
- How can we explain the anomaly? What process can we follow?

Anomaly Diagnosis

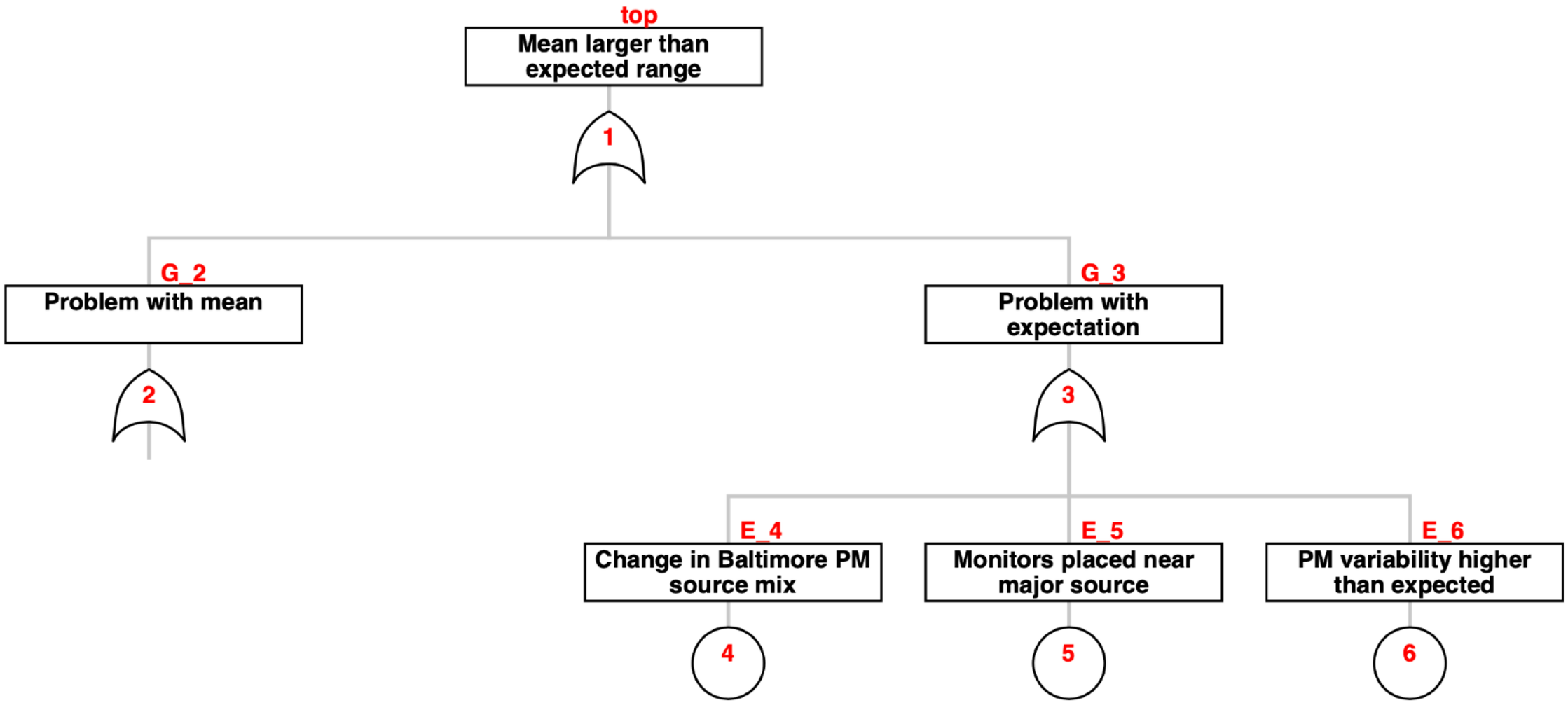


Anomaly Diagnosis





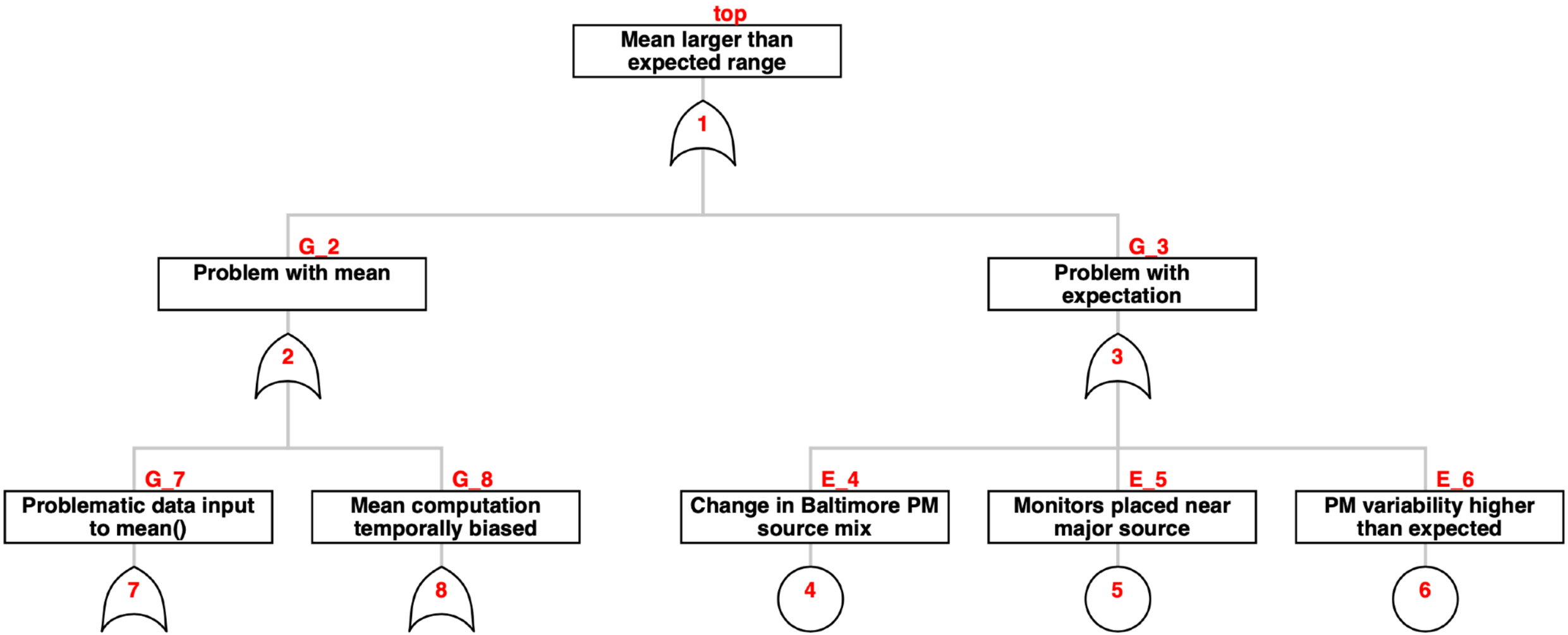
Anomaly Diagnosis



Misunderstanding of process that generates data

Read in data → Remove NA values → Compute mean of remaining values → Output $\bar{x}$

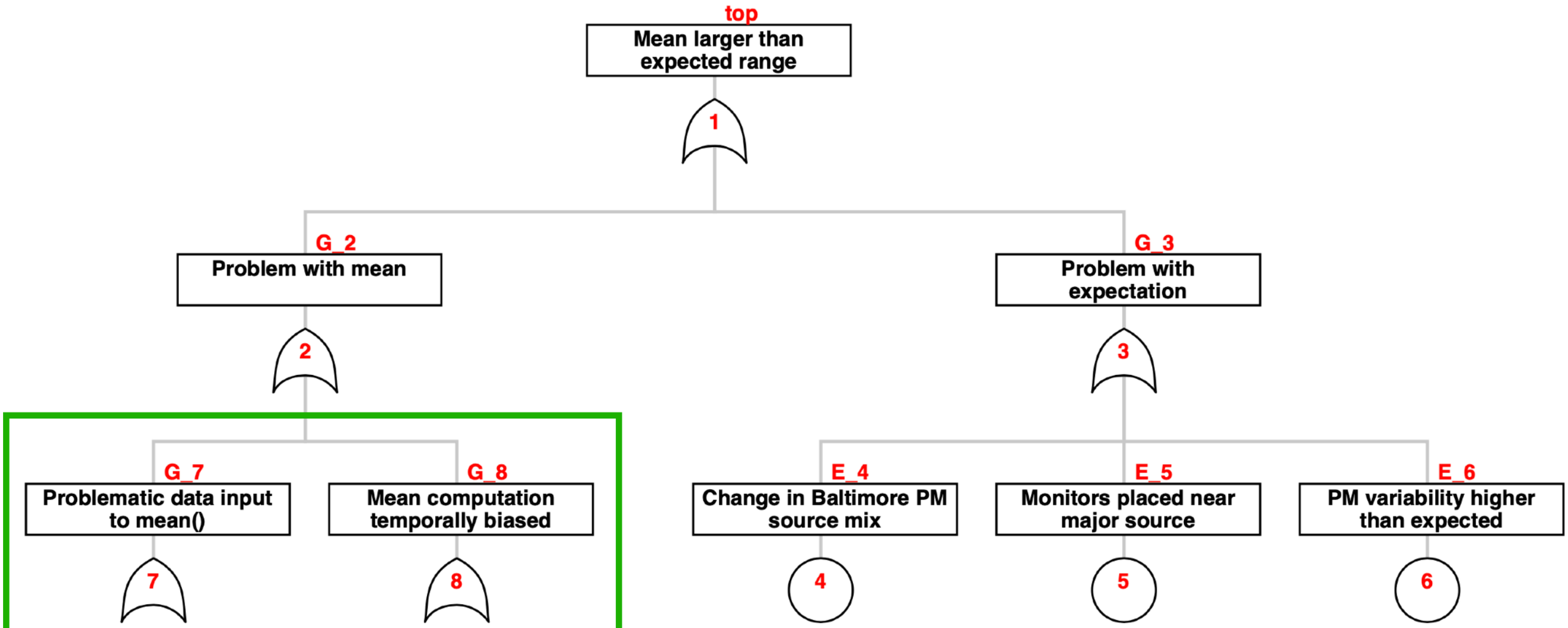
Anomaly Diagnosis



Misunderstanding of process that generates data

Read in data → Remove NA values → Compute mean of remaining values → Output $\bar{x}$

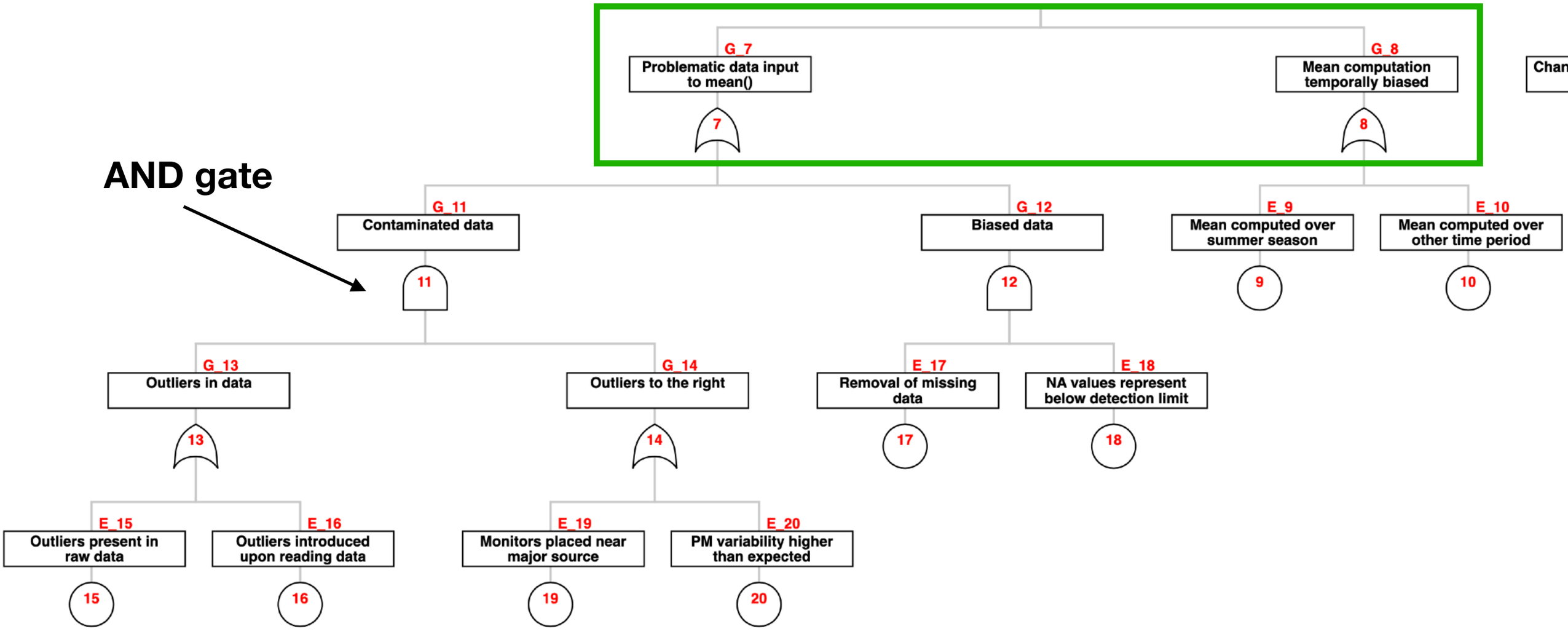
Anomaly Diagnosis



Misunderstanding of process that generates data

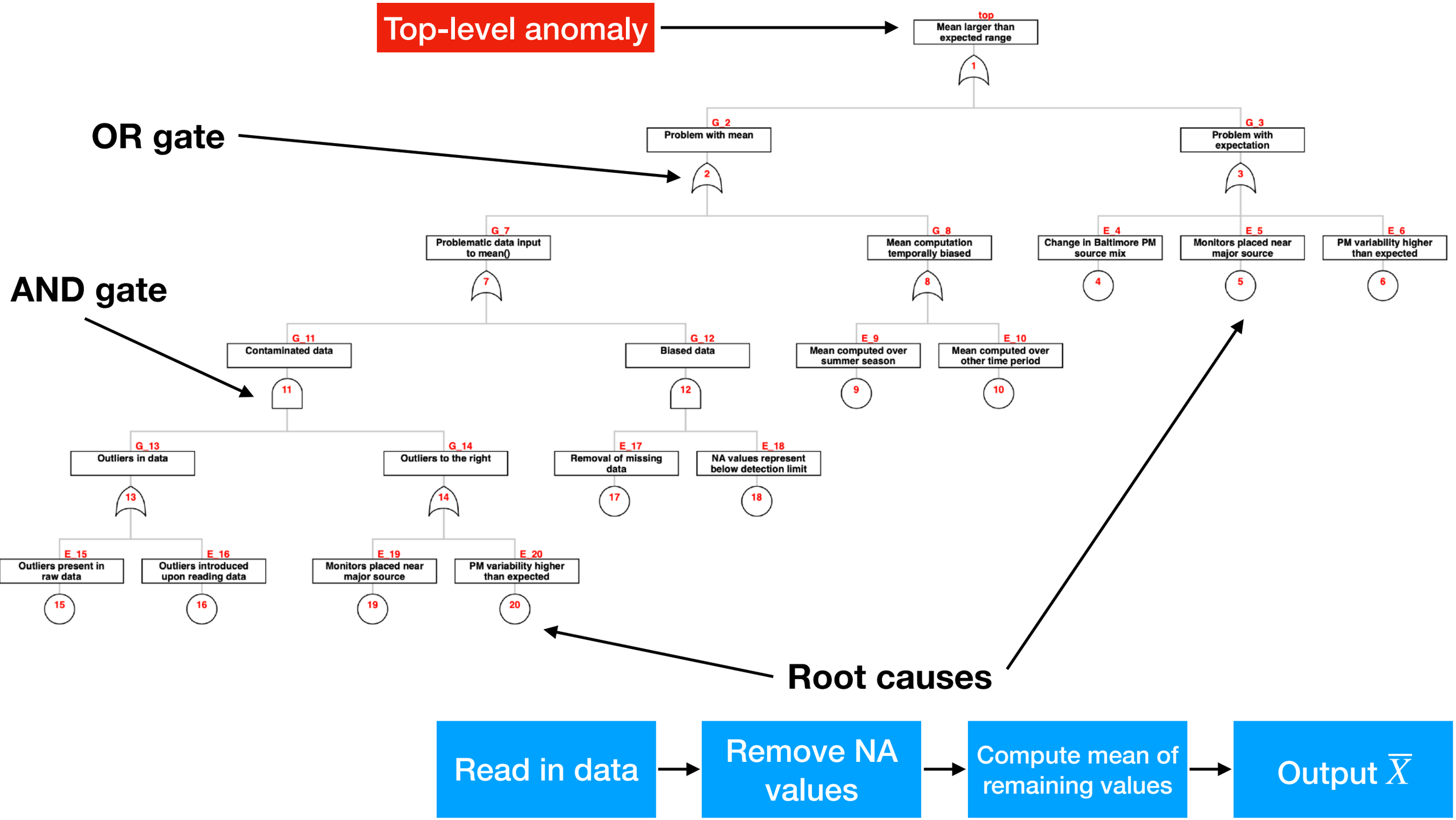
Read in data → Remove NA values → Compute mean of remaining values → Output $\bar{x}$

Anomaly Diagnosis



Read in data → Remove NA values → Compute mean of remaining values → Output $\overline{X}$

Fault Tree



Read in data → Remove NA values → Compute mean of remaining values → Output $\overline{X}$

Fault Trees

- Fault trees are commonly used tools in systems engineering to diagnose anomalies or outright failures

- Can be used on a pre-mortem or post-mortem basis

- We can apply fault trees to "analytic methods systems" to evaluate strengths/weaknesses of system design before applying them to data

- Forces analysts to consider potential outcomes and their causes

- Provides a set of hypotheses (root causes) to explore in the dataset

Fault Trees

- Fault trees provide a systematic, compact, and generalizable method for characterizing a key data analytic process.

- Forces "pre-registration" of set of expected outcomes

- A record of system operation on real-world data

- A roadmap for evaluating data analytic skill; identifies weaknesses in knowledge of underlying systems.

- Hypothesis: Width and depth related to diversity of data analytic experience

Summary

- Data analysis regularly has to explain *unexpected* outcomes

- Diagnosis is a critical skill for driving the *iterative* process of data analysis

- We can *unbundle* diagnostic problems from the rest of data analysis and practice them directly (even without data!)

- Diagnosis is often driven by intuition and prior experience but we should encourage students to engage in *evidence gathering* when evaluating root causes

- Diagnosis requires counterfactual thinking that can be challenging for novices

Annual Review of Statistics and Its Application

Perspective on Data Science

Roger D. Peng[1] and Hilary S. Parker[2]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA; email: rdpeng@jhu.edu

[2]Independent Consultant, San Francisco, California 94102, USA

Roger Peng
UT Austin

Hilary Parker

JOURNAL OF STATISTICS AND DATA SCIENCE EDUCATION
2021, VOL. 29, NO. 3, 267–276
https://doi.org/10.1080/26939169.2021.1971586

Taylor & Francis
Taylor & Francis Group

Diagnosing Data Analytic Problems in the Classroom

Roger D. Peng, Athena Chen, Eric Bridgeford, Jeffrey T. Leek, and Stephanie C. Hicks

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Roger Peng
UT Austin

Athena Chen
Johns Hopkins

Eric Bridgeford
Johns Hopkins

Stephanie Hicks
Johns Hopkins

Jeff Leek
Fred Hutch
Cancer Center