Reproducible Research: What Have We Learned in 20 Years?

Roger D. Peng
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Beuhler-Martin Keynote
May 2021

Reproducible Research: A Retrospective

Annual Review of Public Health

Vol. 42:- (Volume publication date April 2021)
Review in Advance first posted online on January 19, 2021. (Changes may still occur before final publication.)
https://doi.org/10.1146/annurev-publhealth-012420-105110

Roger D. Peng and Stephanie C. Hicks

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA; email: rdpeng@jhu.edu, shicks19@jhu.edu

Abstract

Advances in computing technology have spurred two extraordinary phenomena in science: large-scale and high-throughput data collection coupled with the creation and implementation of complex statistical algorithms for data analysis. These two phenomena have brought about tremendous advances in scientific discovery but have raised two serious concerns. The complexity of modern data analyses raises questions about the reproducibility of the analyses, meaning the ability of independent analysts to recreate the results claimed by the original authors using the original data and analysis techniques. Reproducibility is typically thwarted by a lack of availability of the original data and computer code. A more general concern is the replicability of scientific findings, which concerns the frequency with which scientific claims are confirmed by completely independent investigations. Although reproducibility and replicability are related, they focus on different aspects of scientific progress. In this review, we discuss the origins of reproducible research, characterize the current status of reproducibility in public health research, and connect reproducibility to current concerns about replicability of scientific findings. Finally, we describe a path forward for improving both the reproducibility and replicability of public health research in the future.

Expected final online publication date for the *Annual Review of Public Health*, Volume 42 is April 2021. Please see http://www.annualreviews.org/page/journal/pubdates for revised estimates.

About Me

- Indoor air pollution and health
- Panel studies in vulnerable groups (COPD, asthma)
- Environmental interventions and clinical trials
- Longitudinal data, causal inference, mediation

About Me

- Outdoor air pollution and health
- Air pollution epidemiology
- Ambient air quality standards
- Time series, spatial statistics, hierarchical modeling

What is Reproducibility?

- **Reproducible research** - independently recreating original numerical and other results from a publication using the *same dataset* and (ideally) the same code.

- **Replication** - independently obtaining similar or consistent results using new data and similar approaches/ methods

- Barba (2018) found no agreement on the definitions of replication and reproducibility across many fields

Borrowing Ideas From Open Source Software

American Journal of Epidemiology

Commentary

Reproducible Epidemiologic R

Roger D. Peng, Francesca Dominici,

From the Biostatistics Department, Johns H

Received for publication November 4, 2005;

TABLE 1. Criteria for reproducible epidemiologic research

Research component	Requirement
Data	Analytical data set is available.
Methods	Computer code underlying figures, tables, and other principal results is made available in a human-readable form. In addition, the software environment necessary to execute that code is available.
Documentation	Adequate documentation of the computer code, software environment, and analytical data set is available to enable others to repeat the analyses and to conduct other similar ones.
Distribution	Standard methods of distribution are used for others to access the software, data, and documentation.

Peng et al. 2006

Goals of Reproducibility

- Communicating the details of an investigation
- Increase trust in the data analysis
- Provide tools for learning about data analysis

Goals

- Usable / Transferable software
- Sharing of data

By-Products

Yes, but....


PNAS
Proceedings of the National Academy of Sciences of the United States of America

Keyword, Author, or DOI

Advanc

Home | Articles | Front Matter | News | Podcasts | Authors

NEW RESEARCH IN Physical Sciences | Social Sciences | Biological Sciences

OPINION

Check for updates

Opinion: Reproducible research can still be wrong: Adopting a prevention approach

Jeffrey T. Leek and Roger D. Peng

PNAS February 10, 2015 112 (6) 1645-1646; https://doi.org/10.1073/pnas.1421412111

Article | Figures & SI | Info & Metrics | PDF

Article Alerts
Email Article
Citation Tools
© Request Permissions

Share
Tweet
Like 0
Mendeley

PNAS
Tumor cell heterogeneity

Leek & Peng 2015

Reproducibility Limitations

- An extension of the traditional model of publication (paper + data + software vs. paper)
- Reproducible analyses allow us to uncover problems but still **long after they occur**
- Not useful for **preventing** the release of poor quality data analysis
- Data privacy is an increasingly important consideration

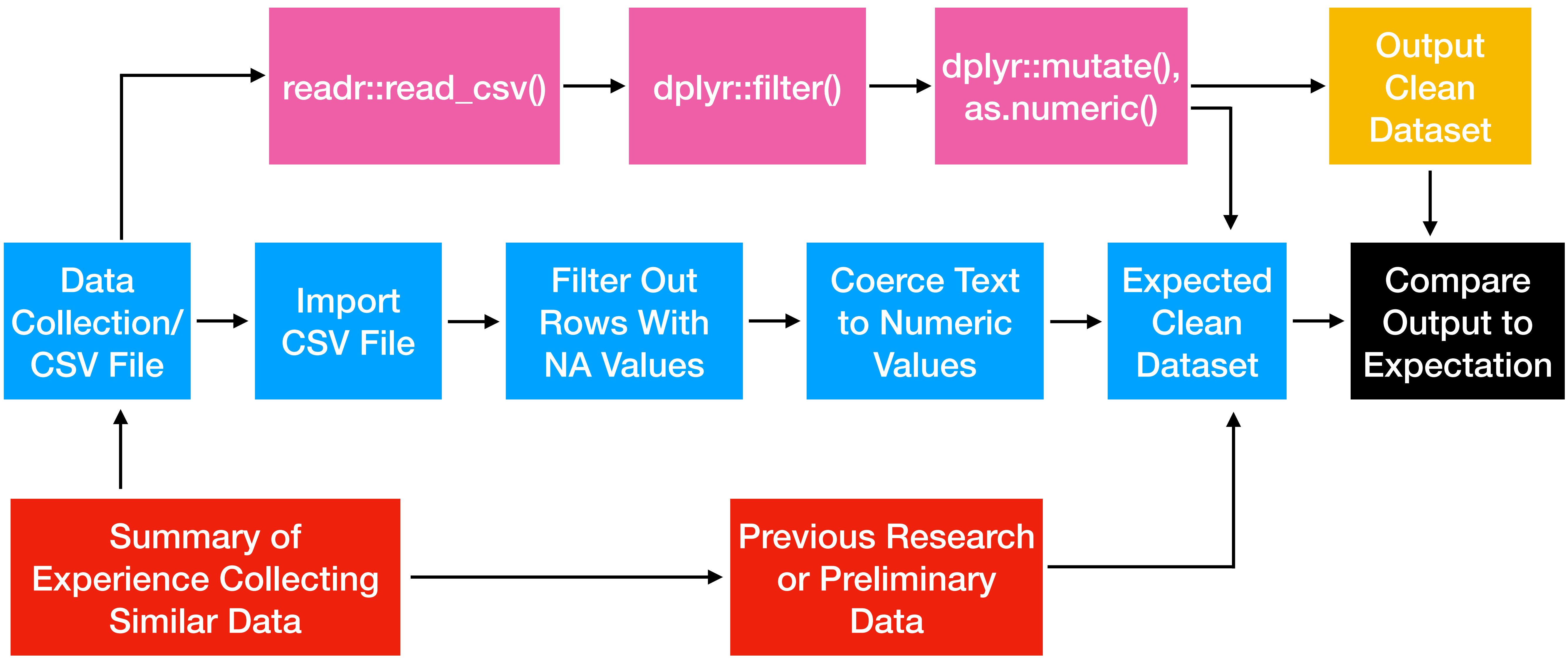
Why is Reproducibility Important?

- Data analyses can produce two important outcomes:
  - Results that are *unexpected* - a deviation/anomaly - **Why**?
  - Results that are *as-expected* - **What if**?
- Without code or data
  - We cannot explain *why* a given result occurred without details of the underlying systems that produced the results
  - We cannot improve future data analyses and prevent mistakes or errors
- But....

Example: Data Cleaning System

- Read CSV file
- Remove rows containing missing values
- Coerce text to numeric values
- Output clean dataset
- **CHECK**: Count number of rows in clean dataset

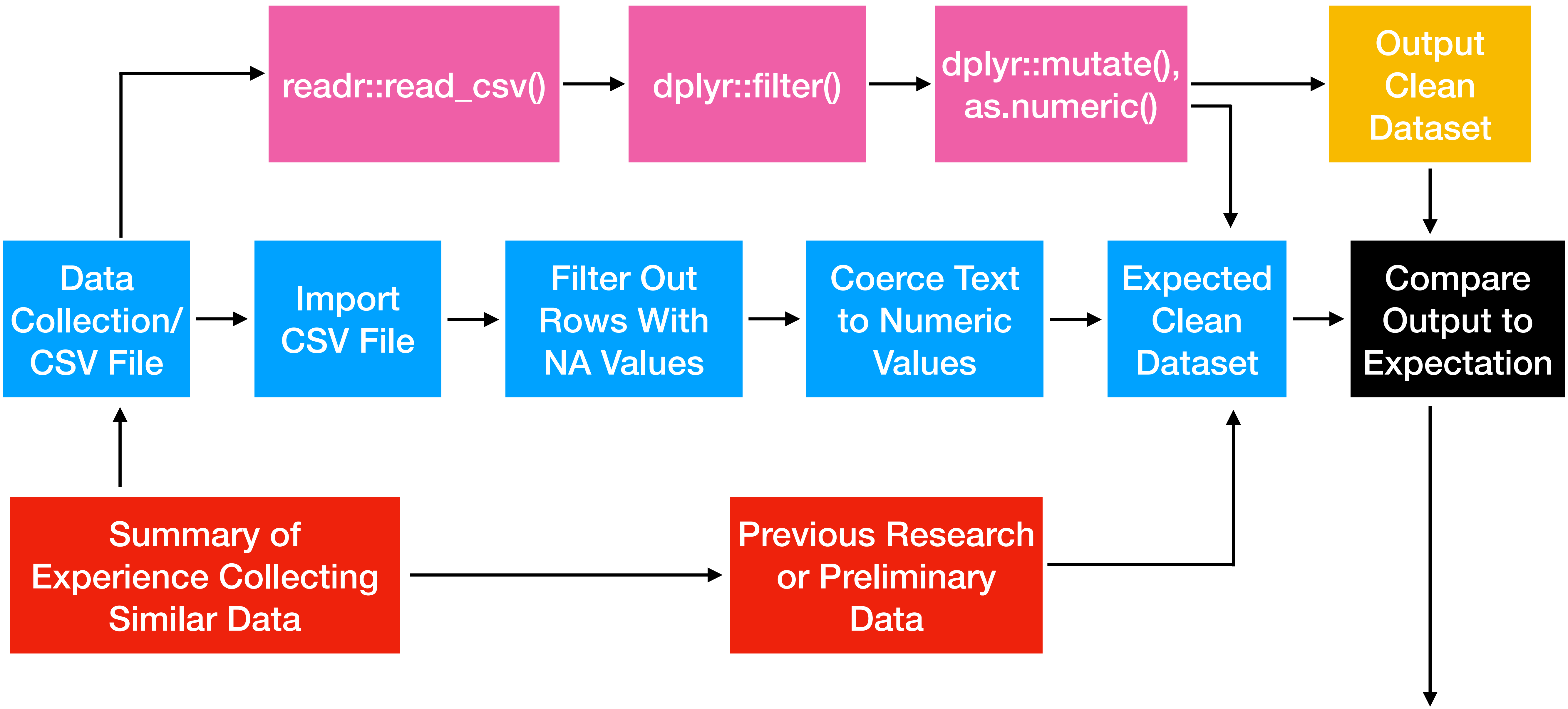
A Data Cleaning System



Developing Expectations

- Expectation: 10% of rows have missing values
- Original dataset 100 rows --> Clean dataset 90 rows
    - Detailed knowledge of messiness of data collection; 10% missing is common
    - Previous experience with measurements
    - Software system unreliable; data gets corrupted
- **What if number of rows in clean dataset is 15?**

A Data Cleaning System



What if # rows in clean dataset is 15?

Diagnosing the Problem

- Source of the unexpected outcome can arise from code, statistics, or science
- Detailed knowledge is required even for a "simple" data cleaning operation
- Possible follow-up action may vary widely depending on the root cause
- Code alone is likely to be sufficient to diagnose the problem

Diagnosing the Problem

- An extra step can be added to the data cleaning sub-system that **generates an error message** if the cleaned dataset has fewer than a certain number of rows.
- **Call the data collection team** to see if there were any recent problems in the latest batch of data.
- A protocol can be put in place where the data collection team **messages the data scientist** if a future batch of data has greater than expected missing observations.

Representing the Analysis

- Code only gives a picture into a single "system" of the analysis
- Problems / failures / unexpected outcomes can originate elsewhere, beyond the code
- Successfully executing an analysis is only one goal
- Understanding an analysis and its sensitivities is also important

Representing Data Analysis

- What is the best way to present the details of a data analysis? Code? Or....
- What goals are we trying to achieve?
- Should we have multiple representations?
- How can we communicate "what we have learned" about data analysis?

How do we describe music?


```
Capo: 1st fret

Intro F#m D A E    x2

F#m      D     A                   E
   Don't know, Don't know if I can do this on my own
F#m       D   A       E
   Why do you have to leave, me
F#m   D         A                 E
   It seems, I'm losing something deep inside of me
F#m     D   A    E
   Hold on, onto me

A
Now I see
D    E
Now I see

Chorus:

F#m       D          E
Everybody hurts some days
F#m        D   E
Its okay to be afraid
F#m
Everybody hurts
D
Everybody screams
A                  E
Everybody feels this way
F#m                D
And its okay
A        E
La di da di da
   F#m    D    A    E
Its okay
```

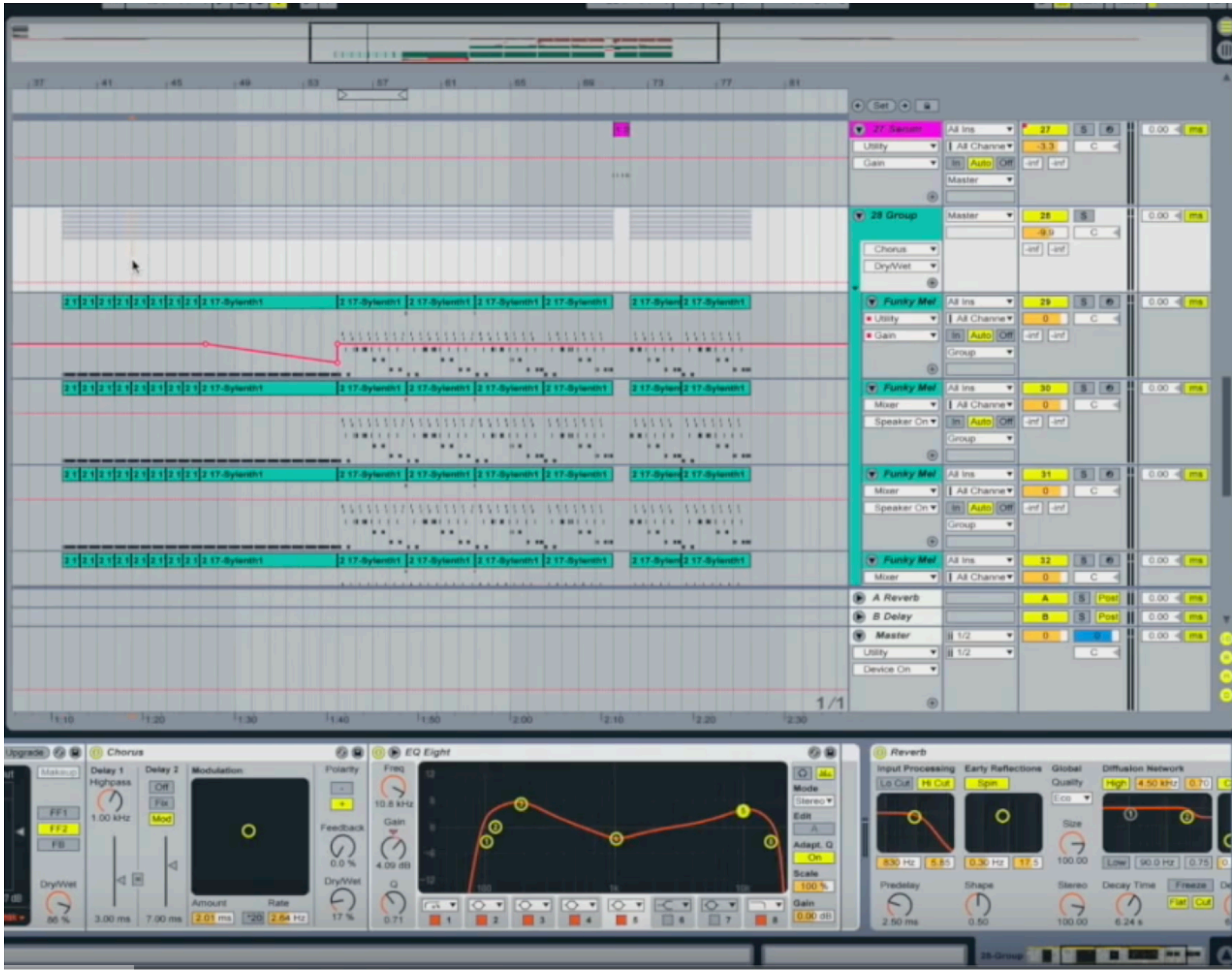
JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH


JOHN COLTRANE
GIANT STEPS


CD
GIANT STEPS
BY JOHN COLTRAN
C VERSION
FAST SWING
TO CODA
SOLOS (11 CHORUSES)
CODA

Music "Code"

Laidback Luke



JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Caroline Shaw, *Valencia*

A Balancing Act

- Each musical representation balances trade-offs for performers and composers
  - Complexity, tool-dependence
  - length, compactness, longevity
  - dependence on external knowledge
  - abstraction
- Data analytic representations make similar trade-offs

Computing the Mean

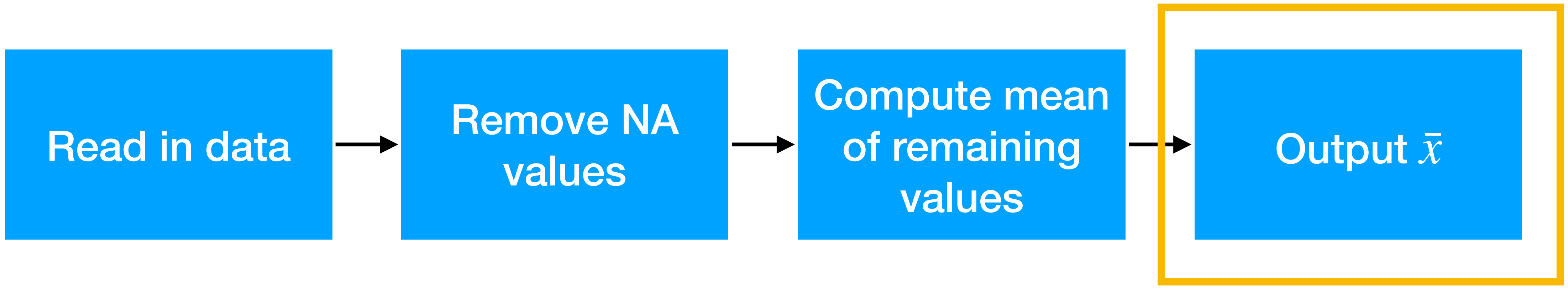
What is the average level of $PM_{2.5}$ air pollution in Baltimore City?

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()
```

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

Anomaly Set

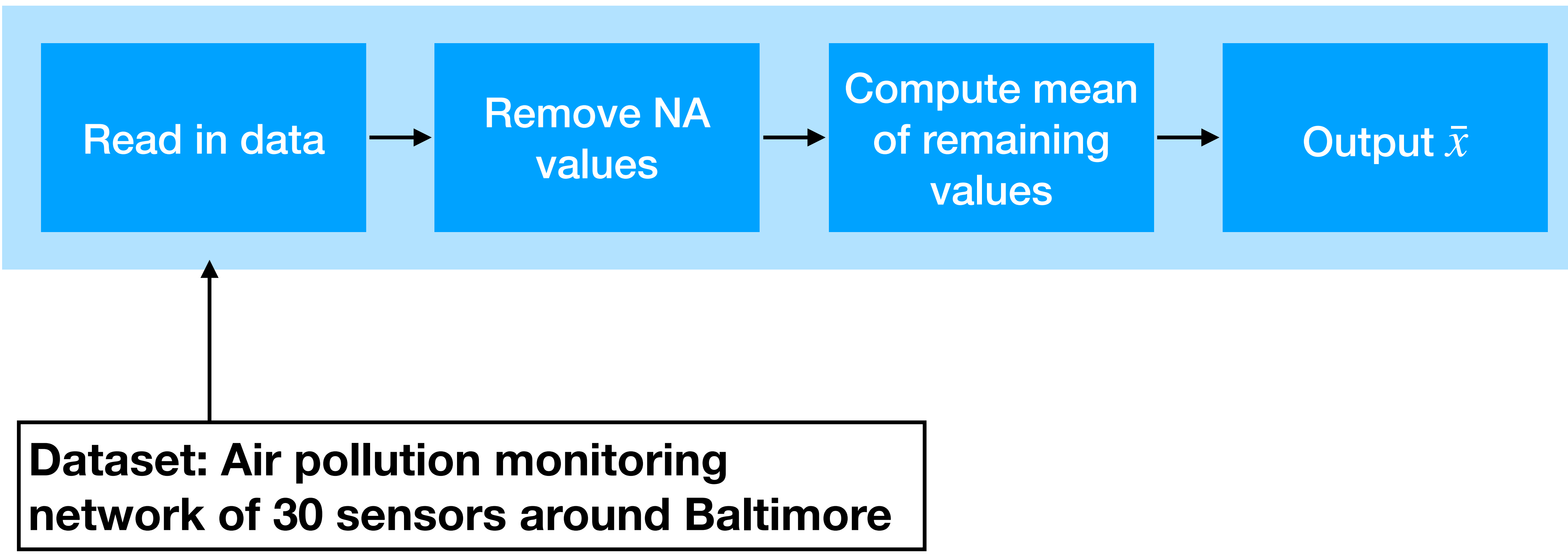


- This system has a single **output**: $\bar{x}$

- 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



Observation $\bar{x} = 25$

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

How did this happen?

Data Analytic System

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()
```

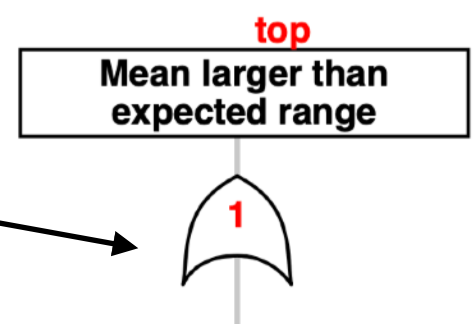
Observation $\bar{x} = 25$

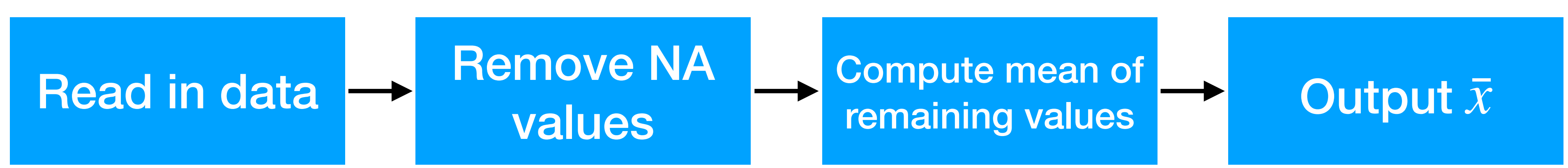
Expectation $\bar{x} \in [8,12]$

Code doesn't really tell us what went wrong!

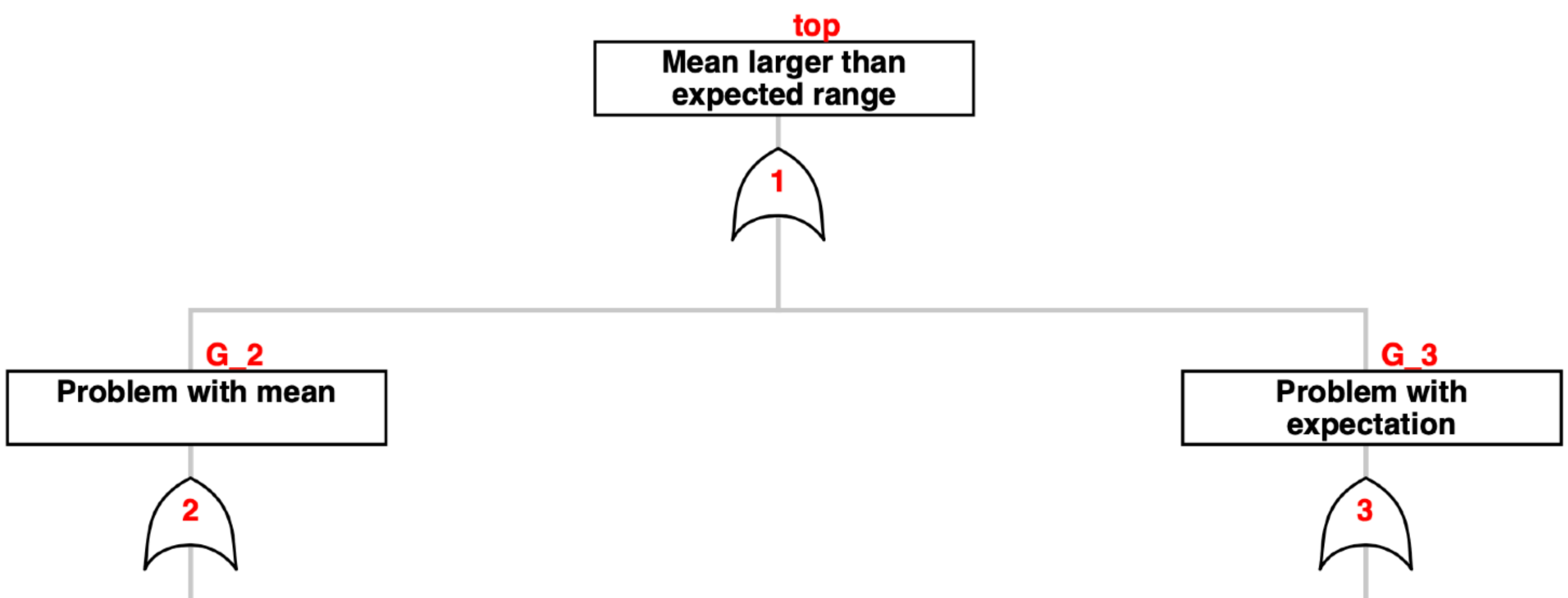
Anomaly Diagnosis

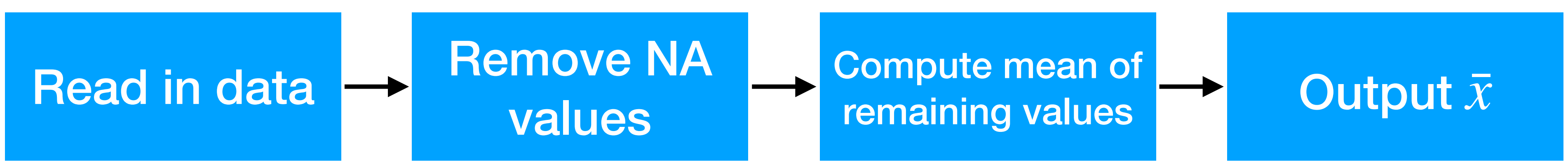
Logical OR gate



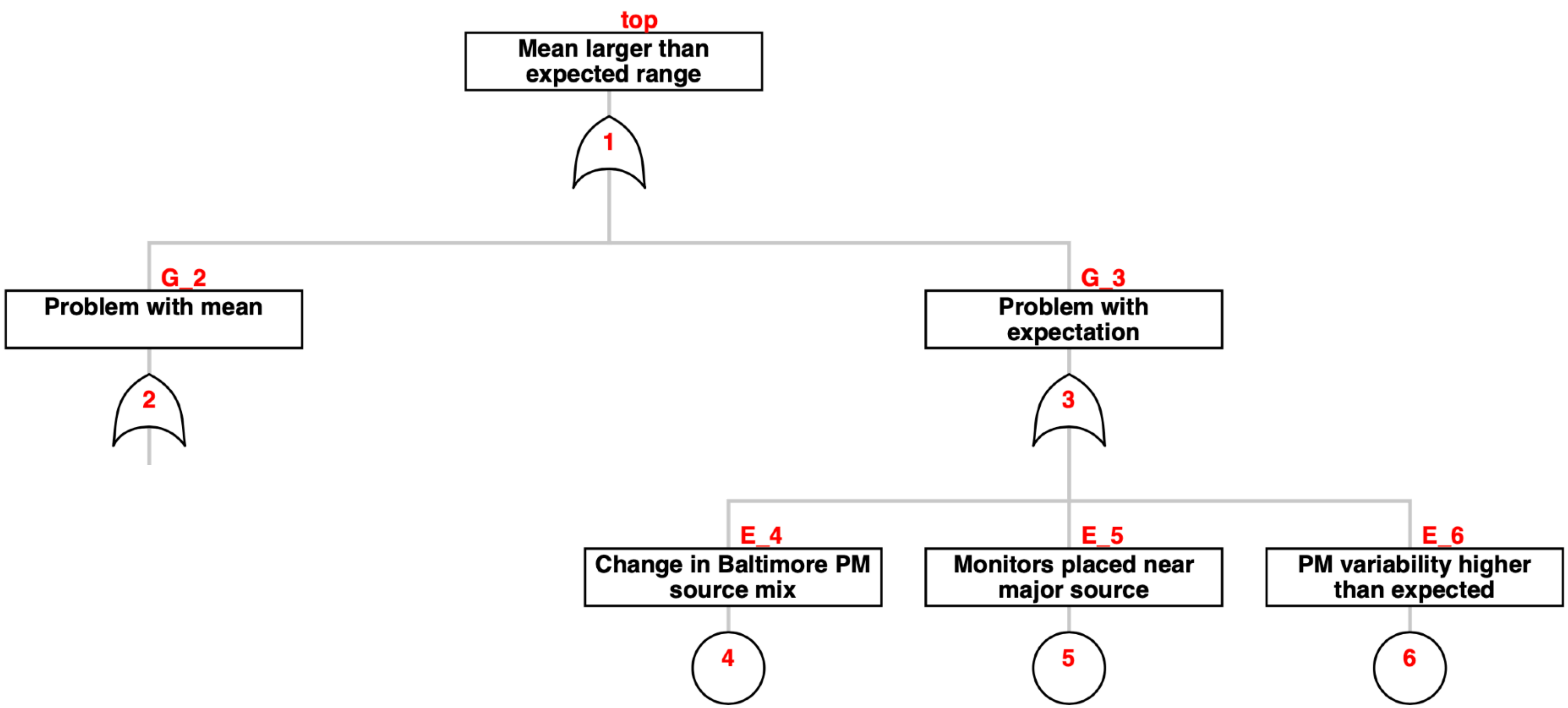


Anomaly Diagnosis

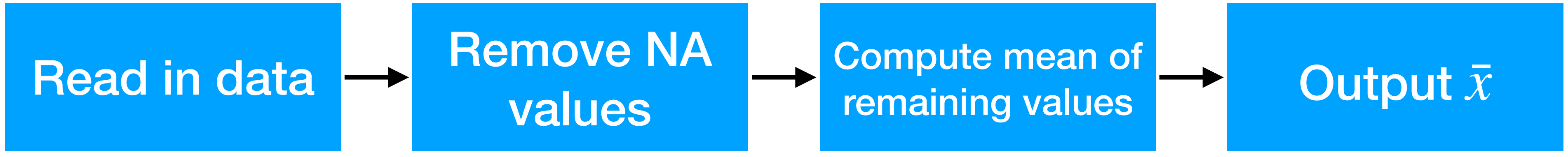




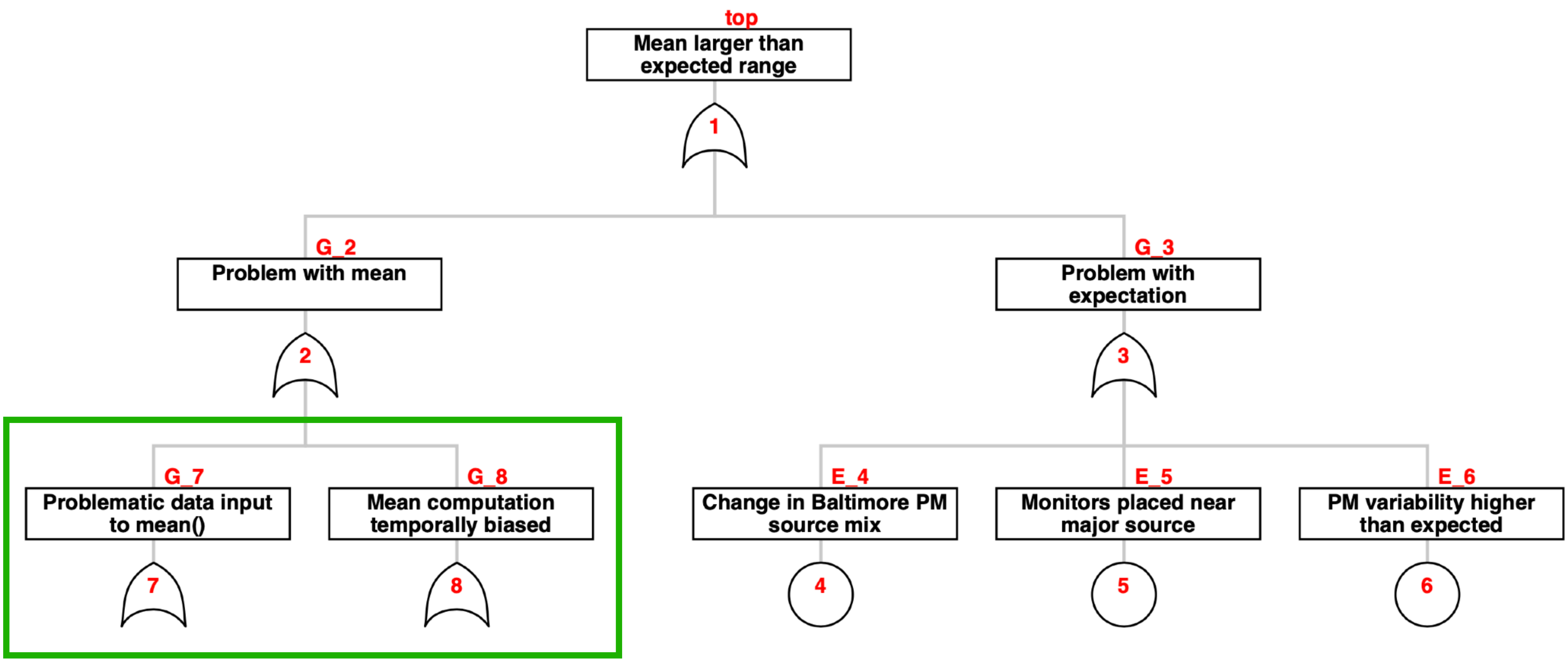
Anomaly Diagnosis



Misunderstanding of process that generates data



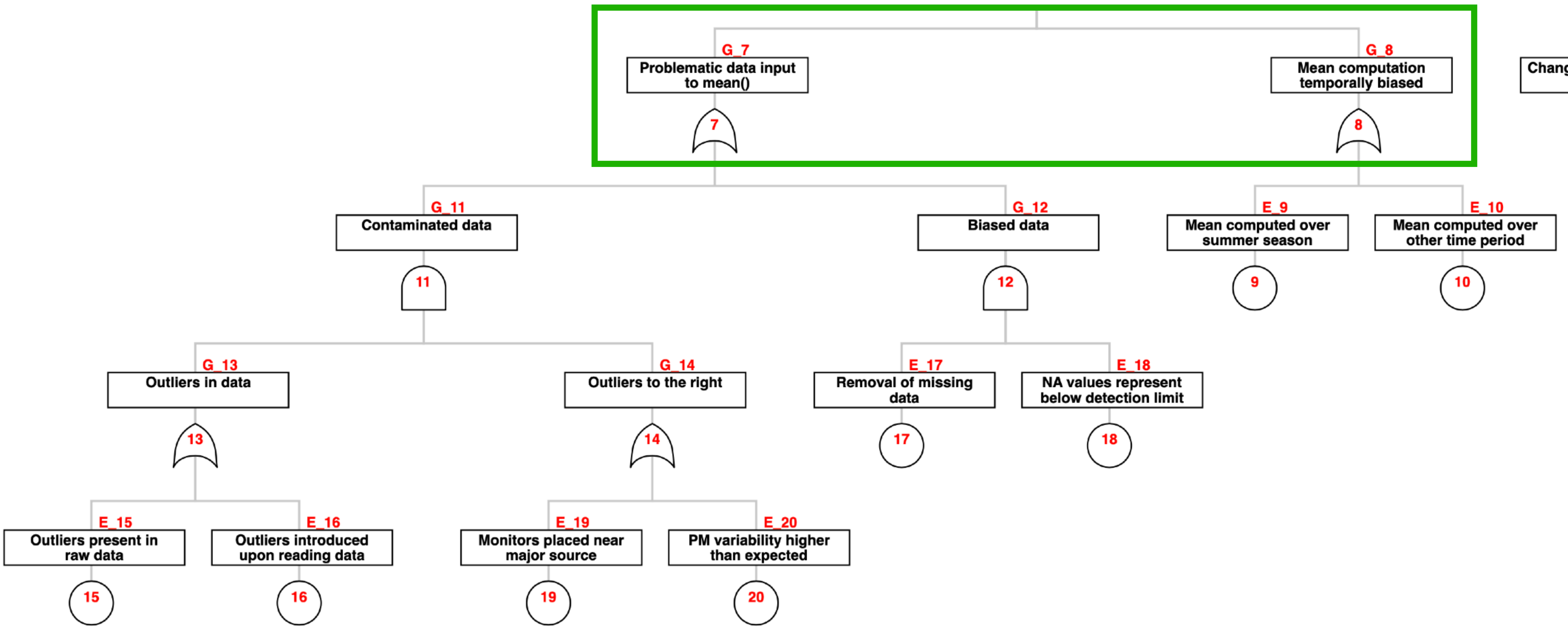
Anomaly Diagnosis



Misunderstanding of process that generates data

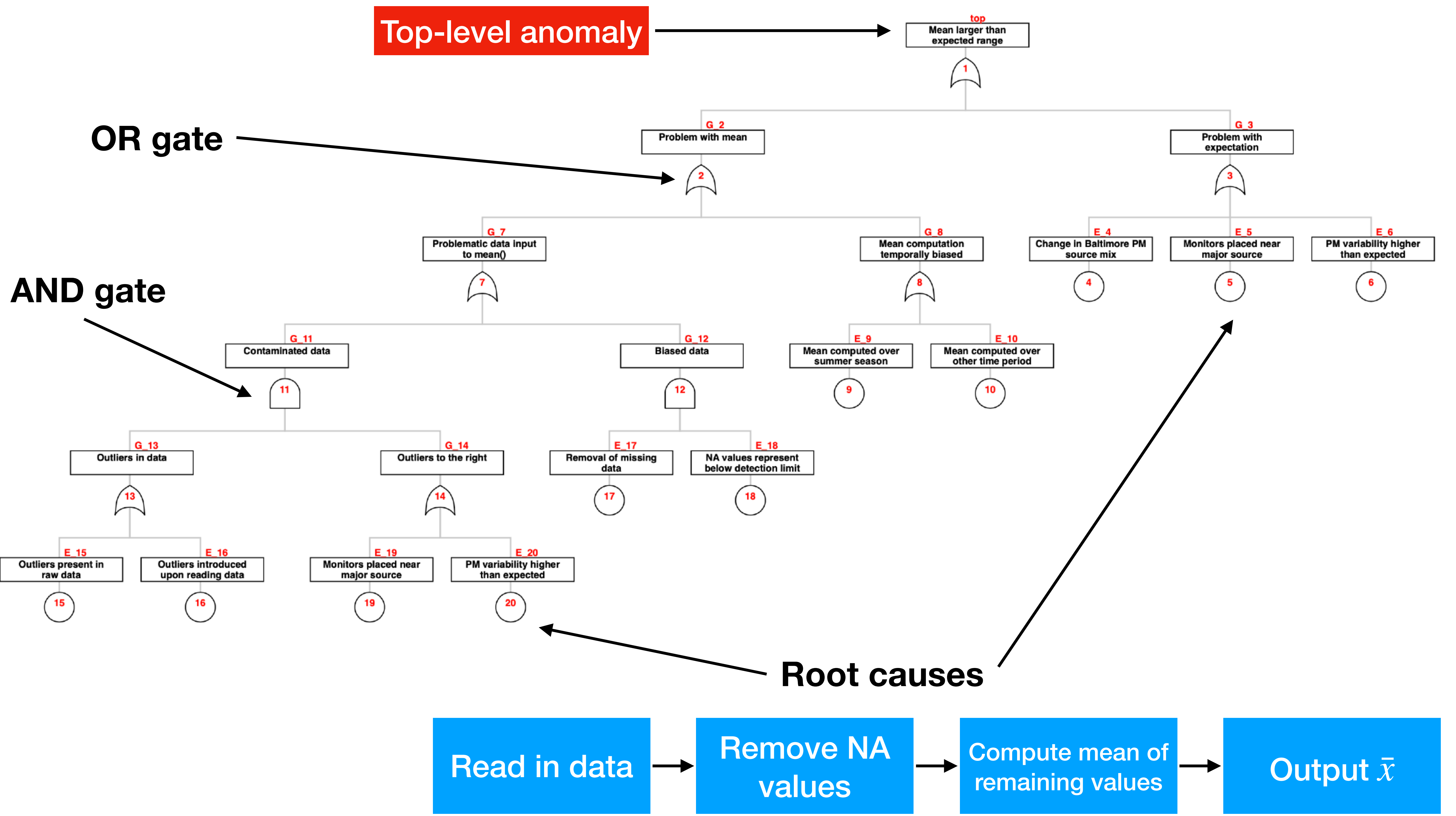
Read in data → Remove NA values → Compute mean of remaining values → Output $\bar{x}$

Anomaly Diagnosis



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

Fault Tree


Top-level anomaly
OR gate
AND gate
Root causes
top
Mean larger than expected range
1
G_2
Problem with mean
2
G_3
Problem with expectation
3
G_7
Problematic data input to mean()
7
G_8
Mean computation temporally biased
8
E_4
Change in Baltimore PM source mix
4
E_5
Monitors placed near major source
5
E_6
PM variability higher than expected
6
G_11
Contaminated data
11
G_12
Biased data
12
E_9
Mean computed over summer season
9
E_10
Mean computed over other time period
10
G_13
Outliers in data
13
G_14
Outliers to the right
14
E_17
Removal of missing data
17
E_18
NA values represent below detection limit
18
E_15
Outliers present in raw data
15
E_16
Outliers introduced upon reading data
16
E_19
Monitors placed near major source
19
E_20
PM variability higher than expected
20
Read in data
Remove NA values
Compute mean of remaining values
Output $\bar{x}$

Representing Data Analysis

- A data analysis is the interaction of multiple systems: scientific, analytic, and software
- There is value in describing a data analysis in terms of its unexpected outcomes to gain visibility into all three systems
- Readers can understand how/why results might deviate from expectations without having to pore over code
- Fault tree can highlight weaknesses in the analytic design
- Key assumptions may be violated and deserve checking

Case Study

ARTICLES

• Retracted •

nature medicine

Genomic signatures to guide the use of chemotherapeutics

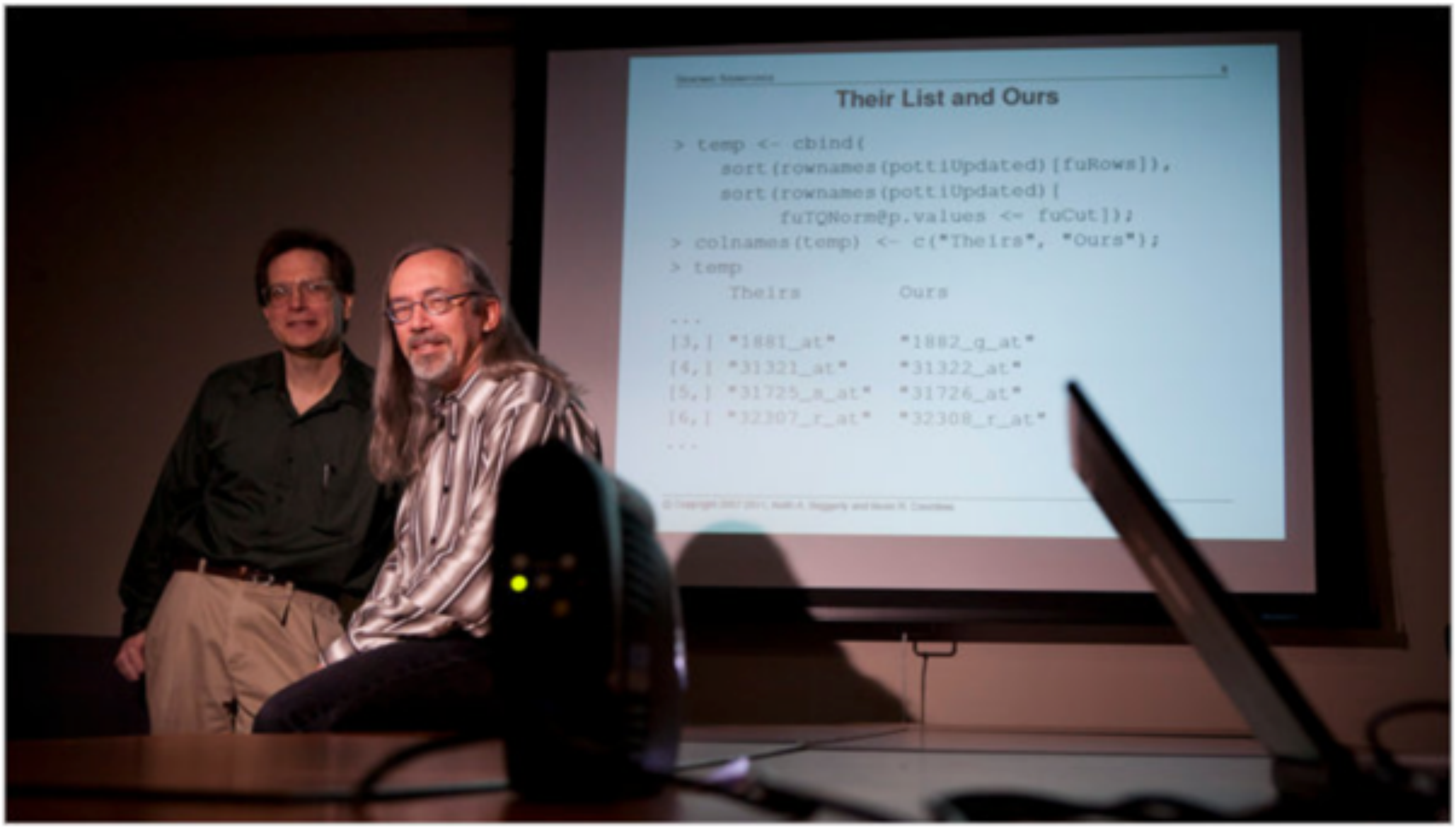
Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

Findings Not Reproducible

- Keith Baggerly and Kevin Coombes attempted to reproduce the findings but could not
- Many basic data data management problems were eventually reverse engineered
    - Off-by-one row mismatches
    - Switching of outcome labels (sensitive/resistant)
    - Duplication of observations
    - Genes cited but not found on arrays
- Evidence of incompetence and fraud

“Front Page” Biostatisticians

How Bright Promise in Cancer Testing Fell Apart



Michael Stravato for The New York Times

New York Times

The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1309–1334
DOI: 10.1214/09-AOAS291

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

BY KEITH A. BAGGERLY[1] AND KEVIN R. COOMBES[2]

University of Texas

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

Institute of Medicine Committee

REPORT BRIEF MARCH 2012

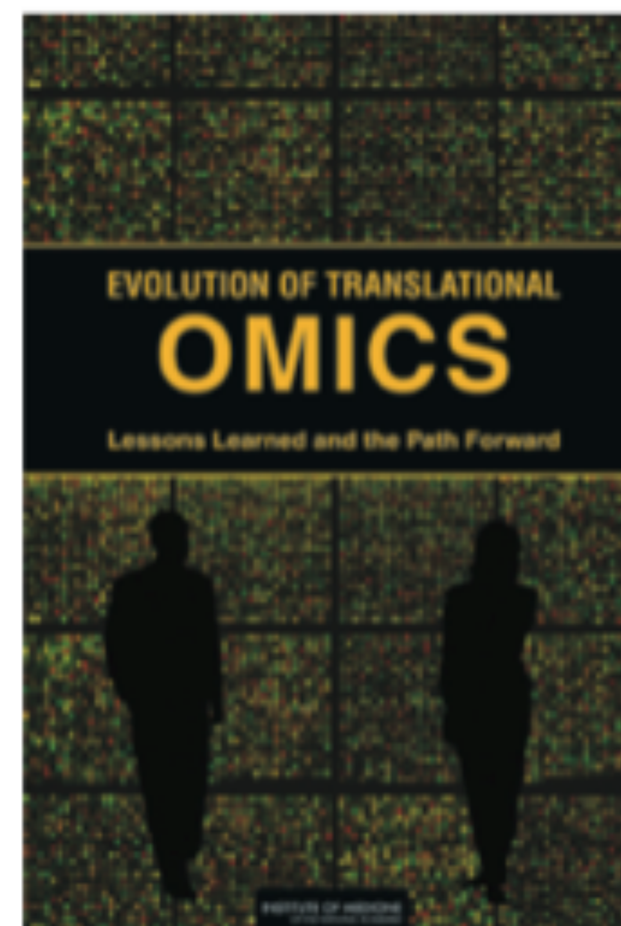
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

Advising the nation • Improving health

For more information visit www.iom.edu/translationalomics

Evolution of Translational Omics

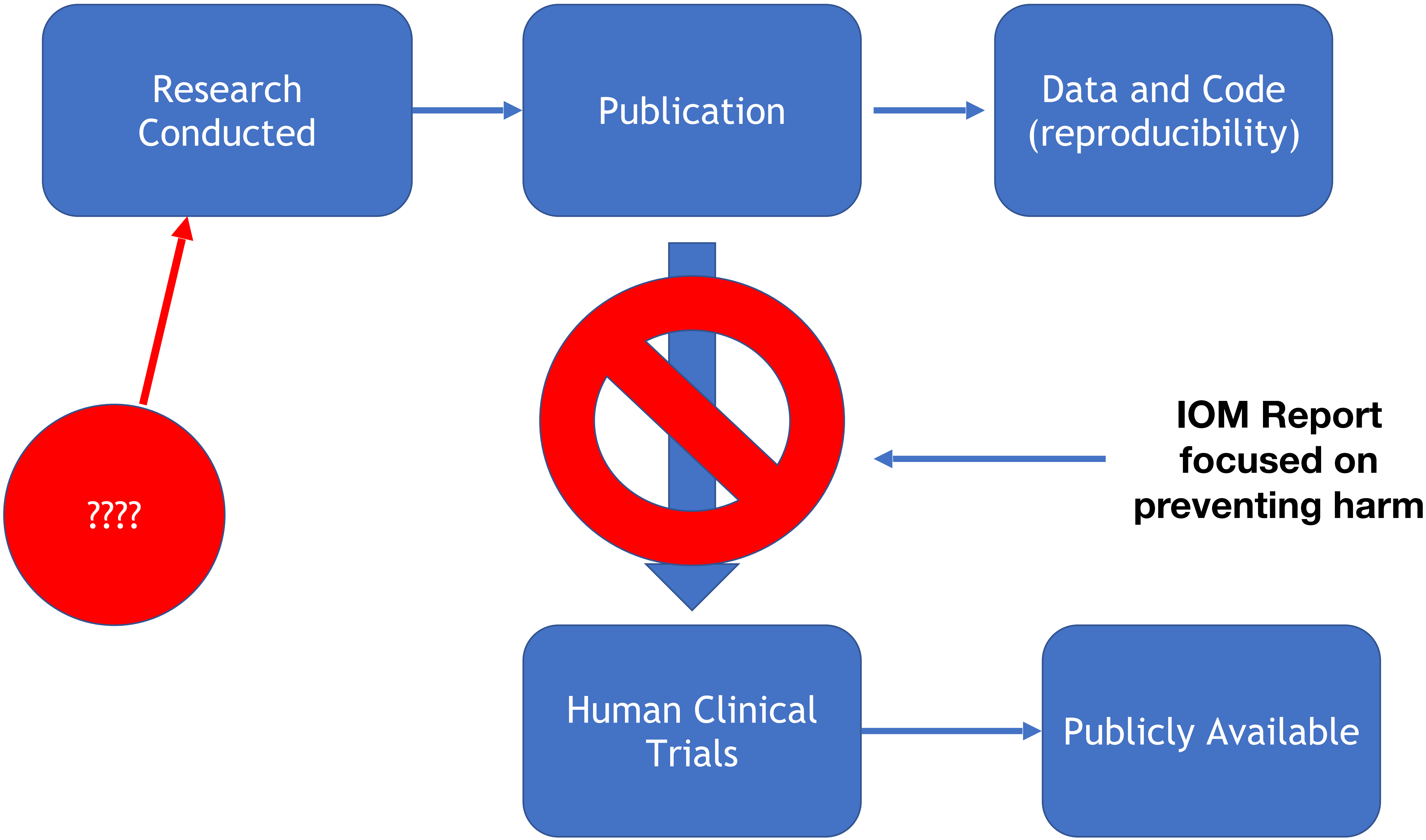
Lessons Learned and the Path Forward



The IOM Report

- Data/metadata used to develop test should be made publicly available
- The **computer code** and fully specified computational procedures used or development of the omics-based test should be made available
- Ideally, the computer code that is released will **encompass all of the steps** of computational analysis, including all data preprocessing steps
- A strong call for reproducibility and transparency

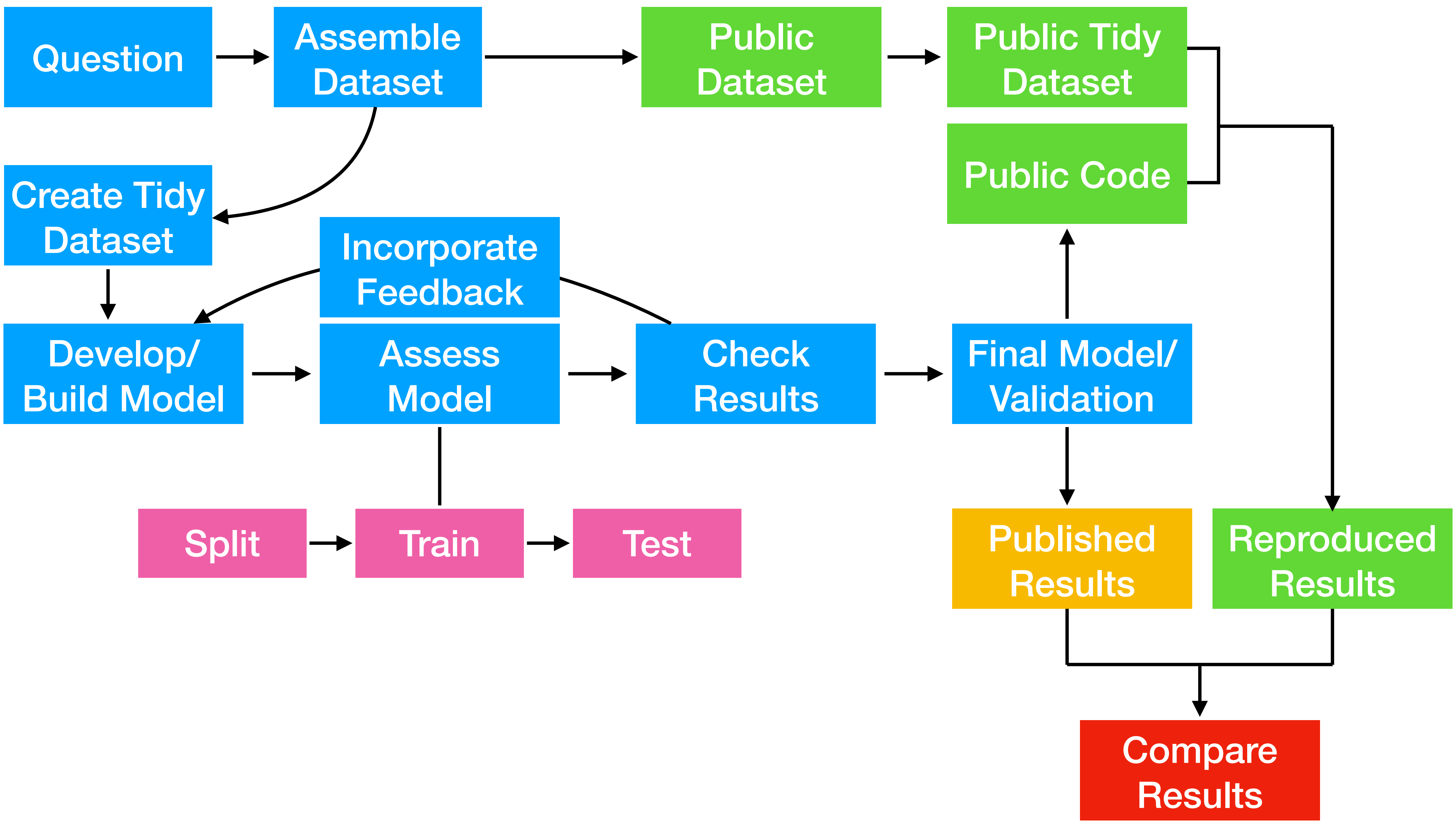
Where to Intervene?



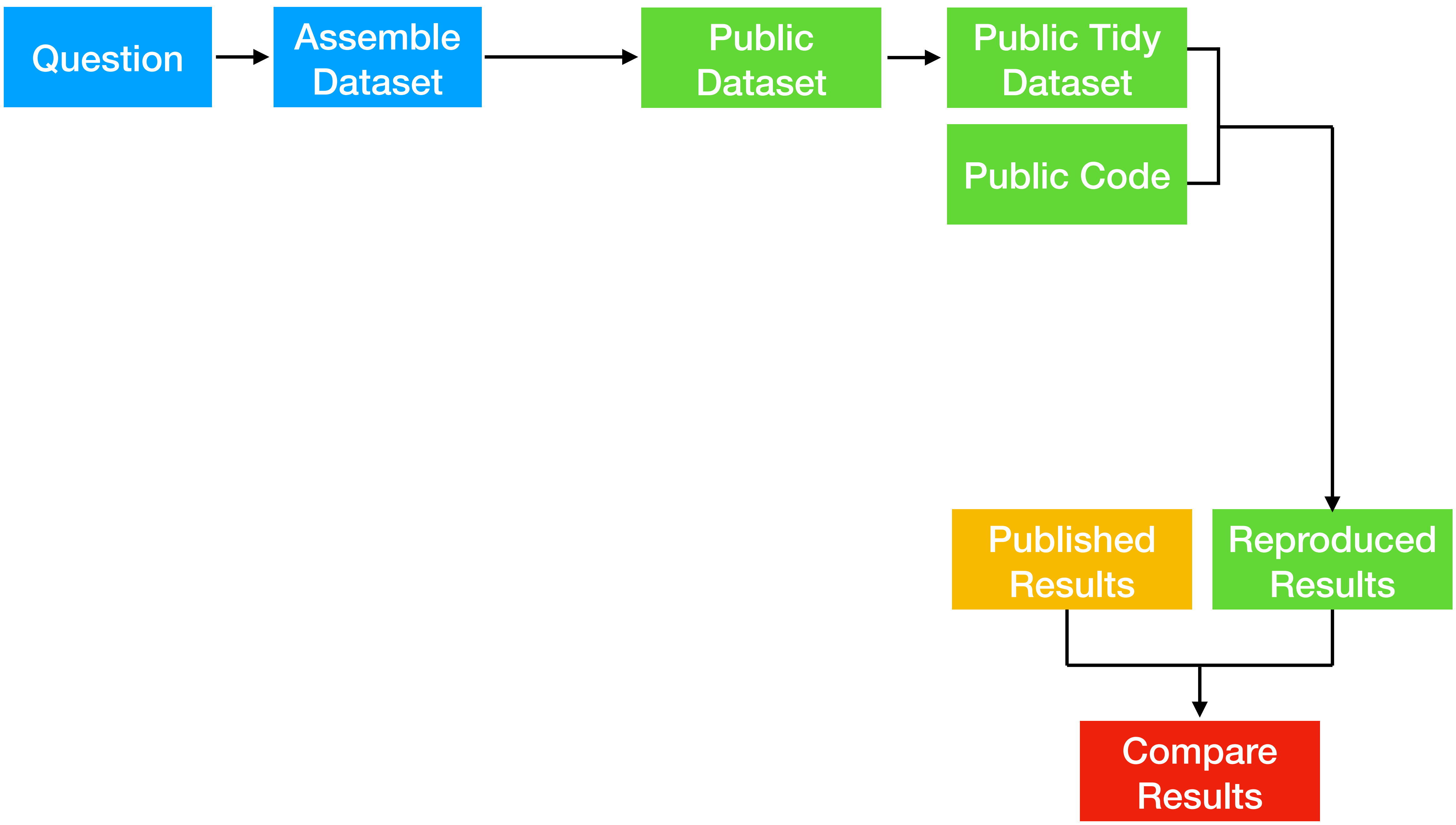
Lessons?

- Reproducibility
- Expertise and training
- Publication pressure; glamour journals
- Funding, conflicts of interest
- Very little visibility into the system generating results

Model System Diagram



Model System Diagram



New Details Emerge (Jan 2015)

“In raising these concerns, I have nothing to gain and much to lose.”
— Bradford Perez

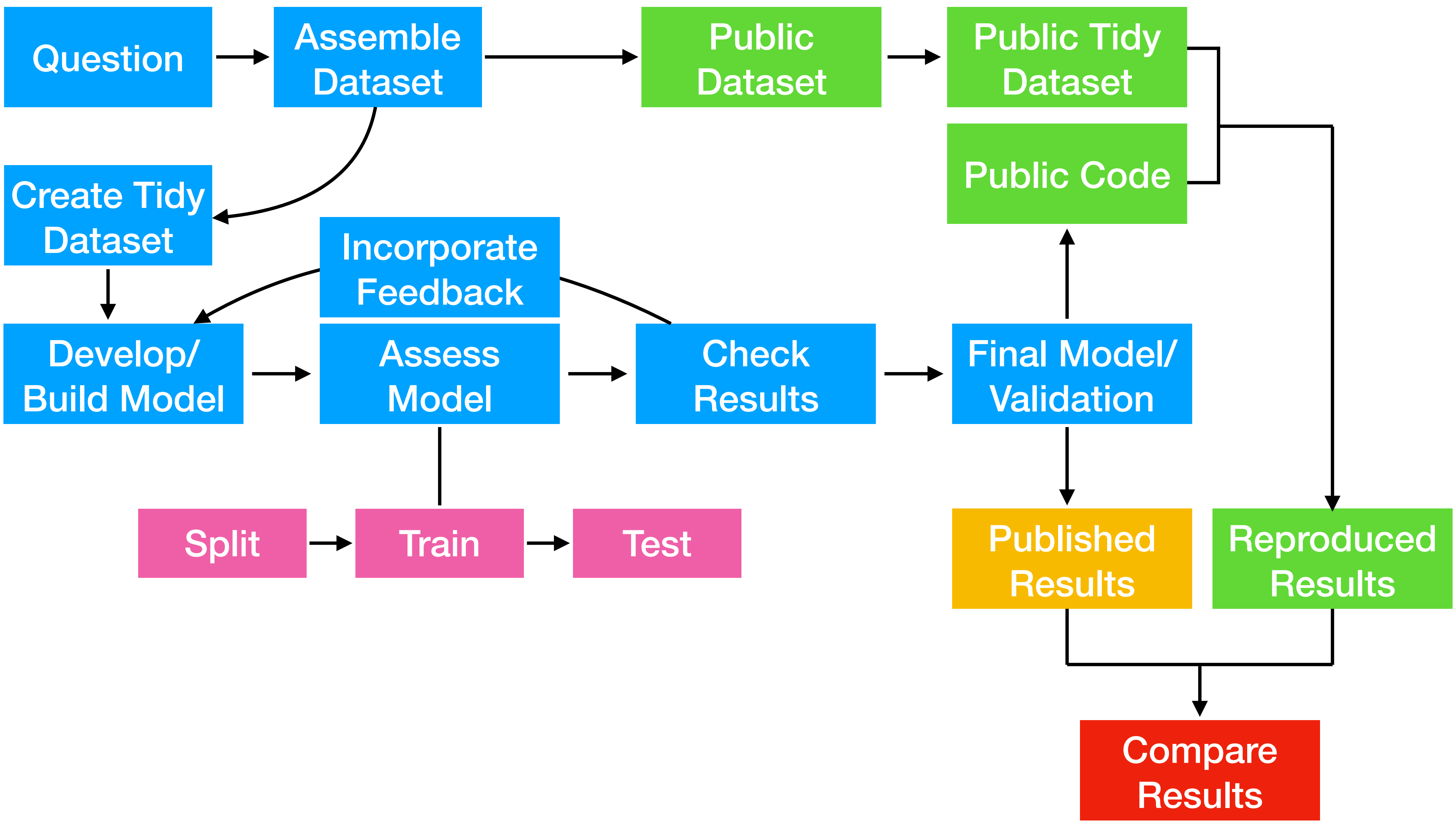
The Cancer Letter, January 2015

The Perez Memo (cont'd)

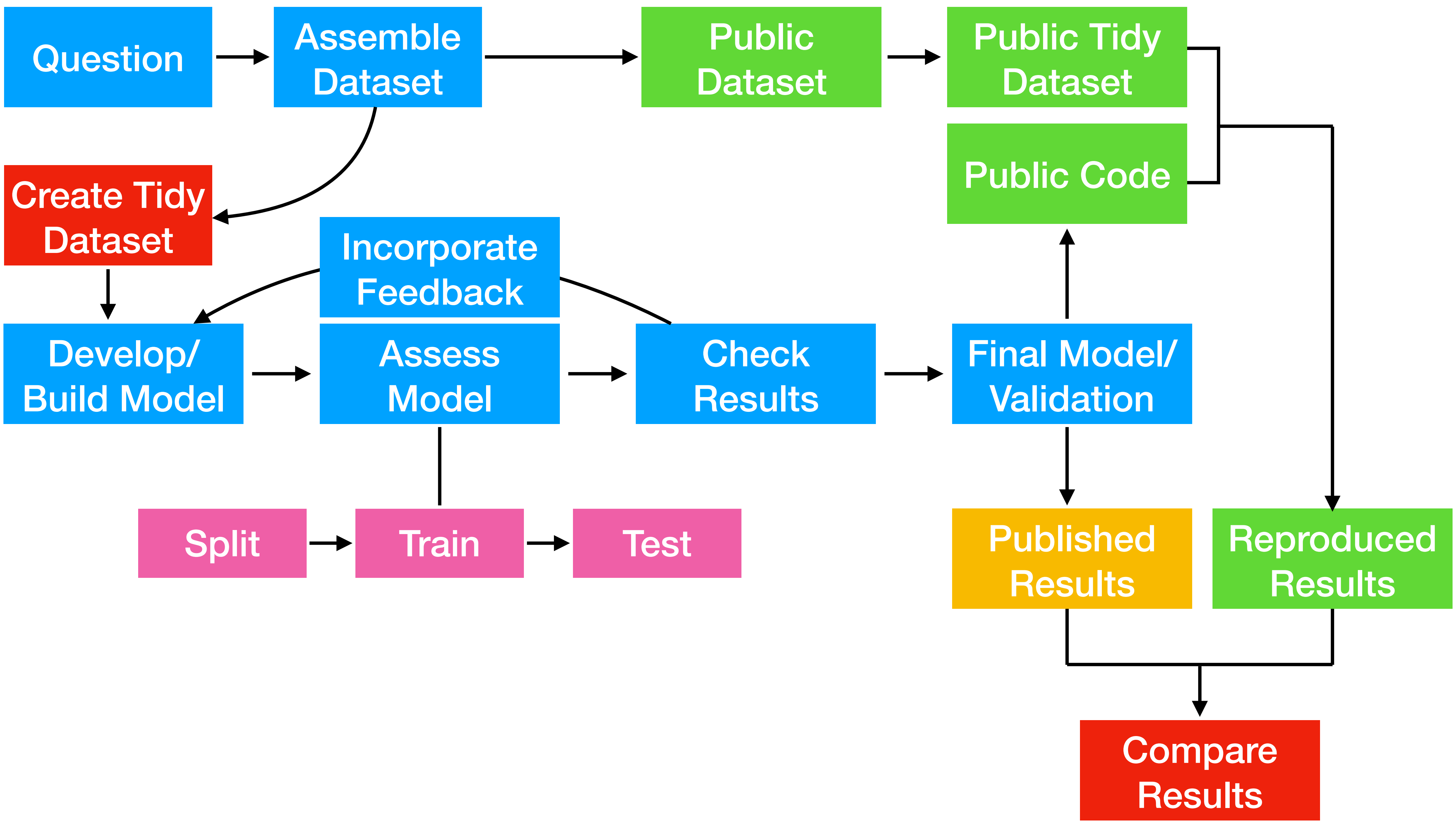
"At this point, I believe that the situation is serious enough that **all further analysis should be stopped** to evaluate what is known about each predictor and it should be reconsidered which are appropriate to continue using and under what circumstances…. I would argue that at this point nothing…should be taken for granted. **All claims of predictor validations should be independently and blindly performed**." [emphasis added]

-Memo from Bradford Perez, April 2008 (The Cancer Letter)

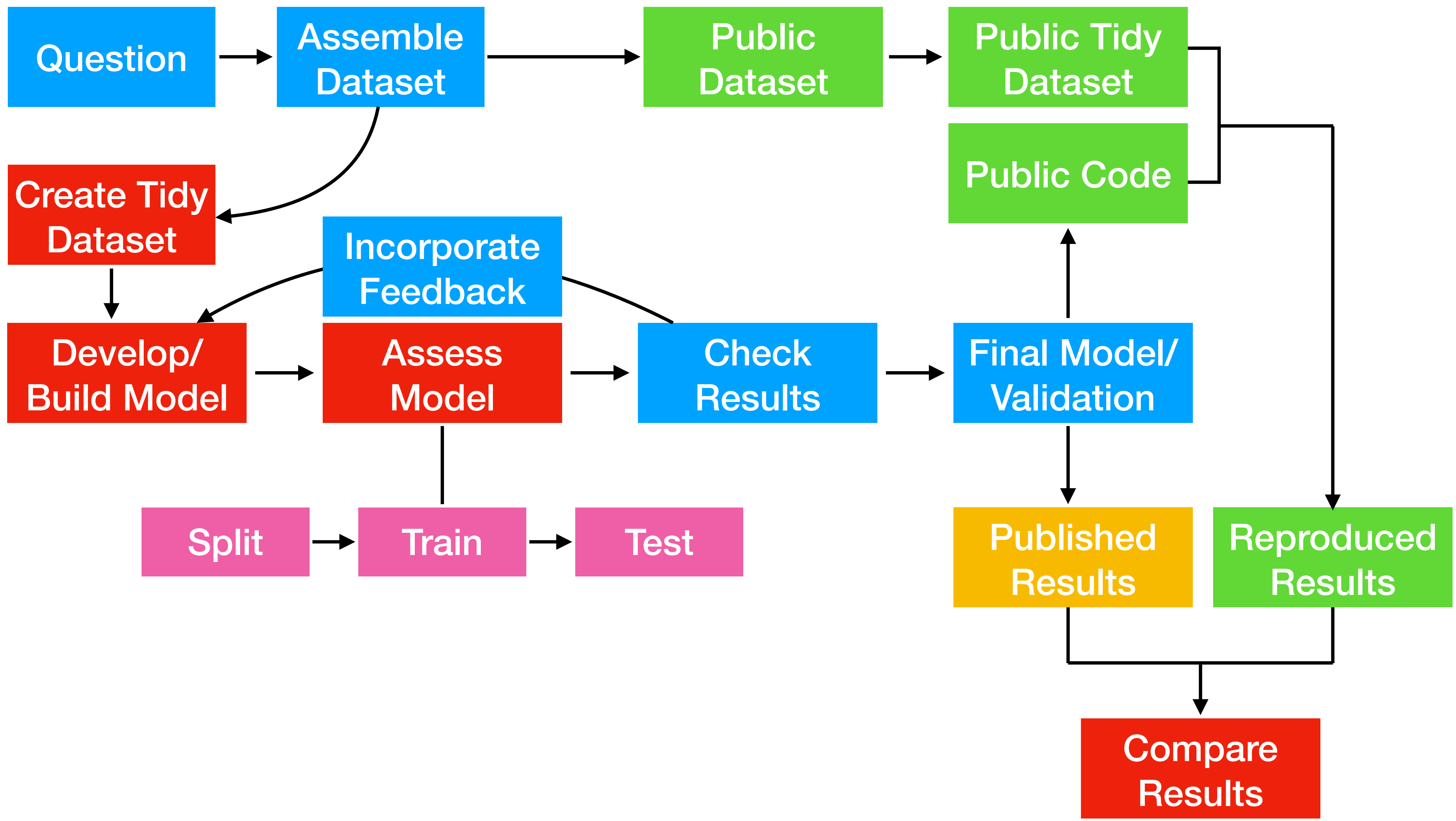
Model System Diagram



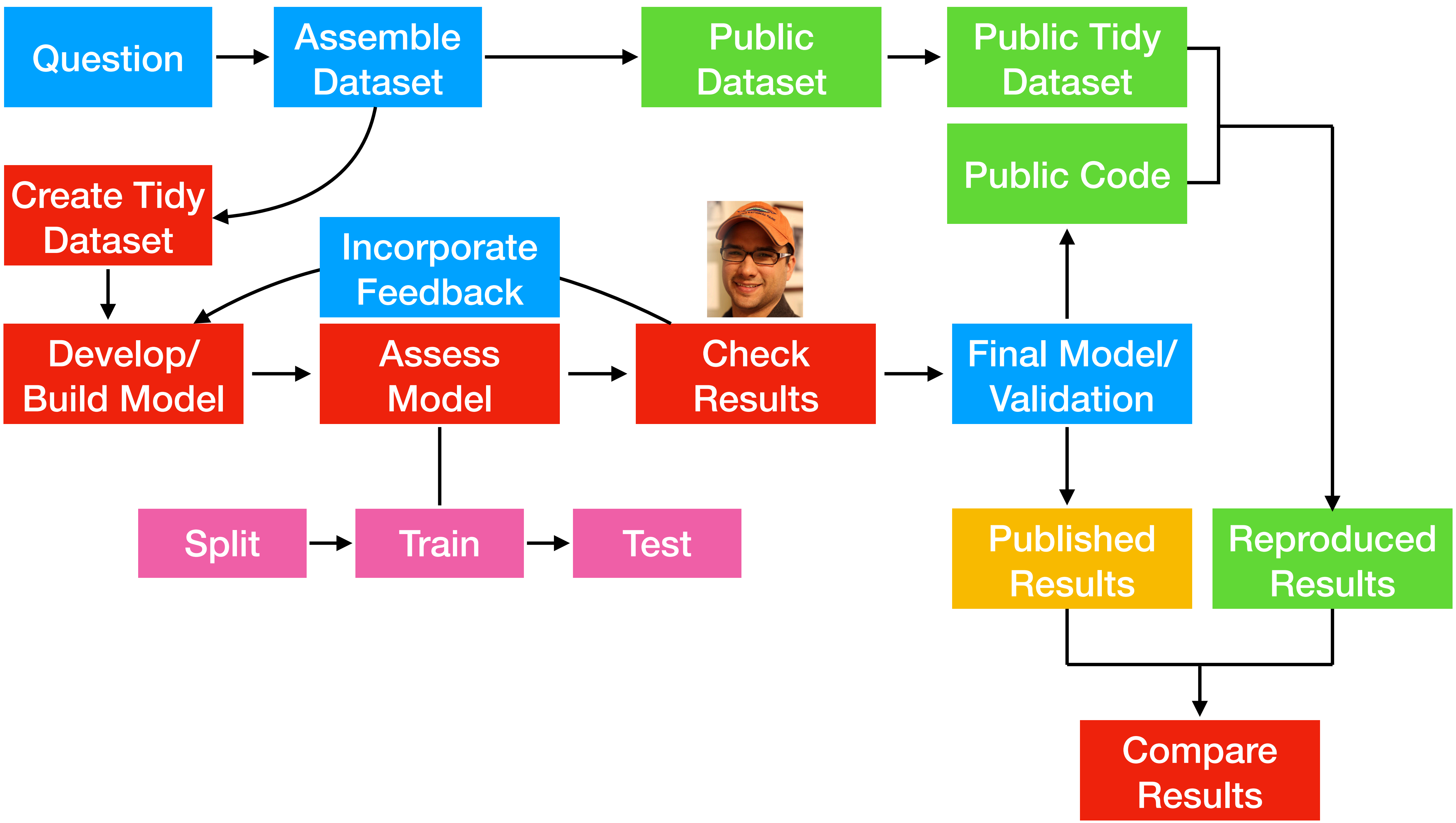
Model System Diagram



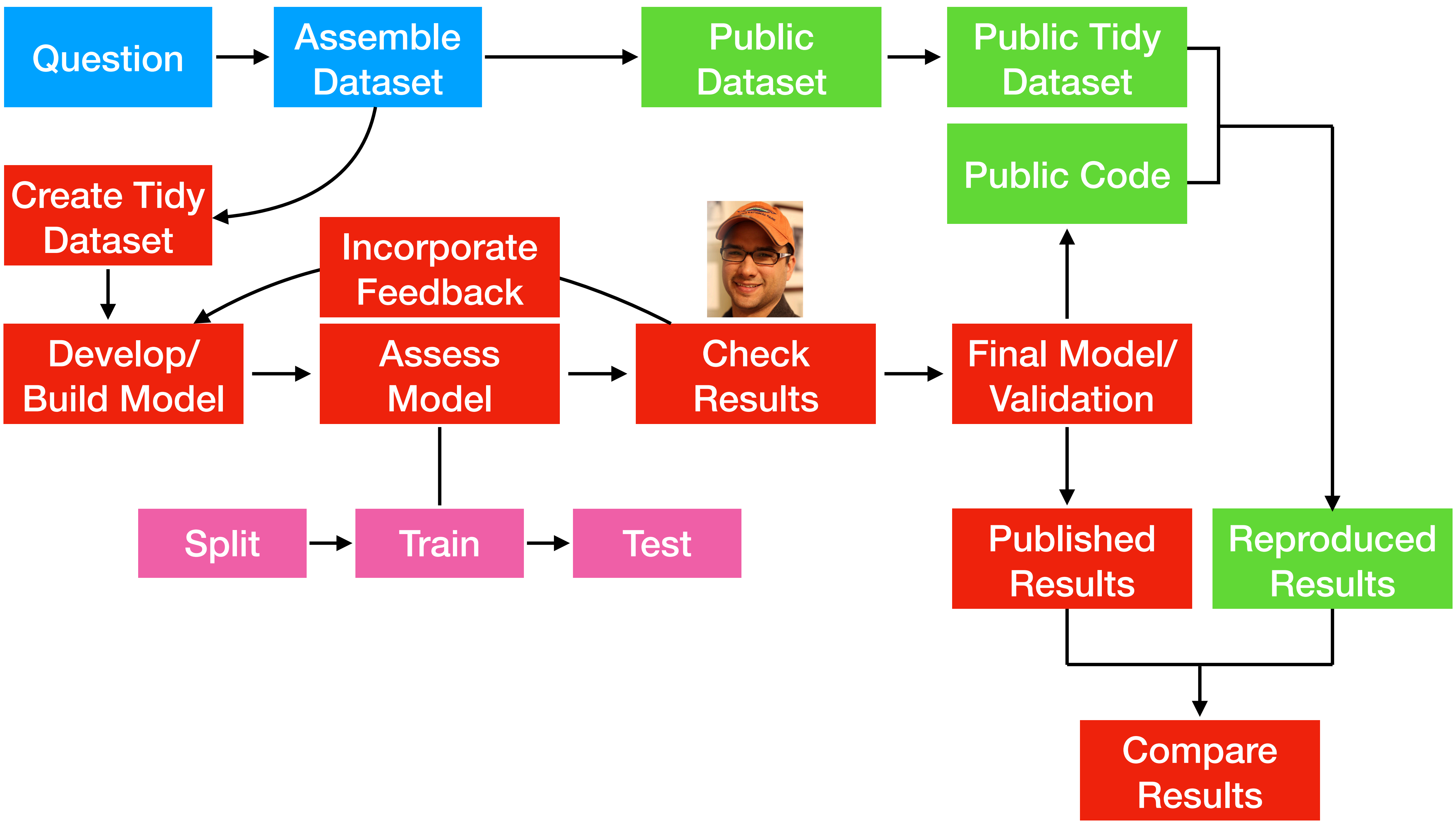
Model System Diagram



Model System Diagram



Model System Diagram



Lessons Learned

- Analyses were not “too complicated” in that there was insufficient expertise; problems were readily recognized
- Lab/institute cultural problems lead to unwillingness to communicate obvious problems
- From the analyst perspective, a breakdown in communication is an early warning sign of potential data analytic problems
- Making published analyses more reproducible likely would not have changed much

Lessons Learned

- Most common problems are simple
- Most simple problems are common
- Lack of reproducibility hides simplicity of errors
- **Recommendations:** Reproducible reporting, better report structure, check for common errors (i.e. "severing data from its associated annotations")
- **Implicit**: Good team structure for constant iteration and improvement

Baggerly & Coombes (2009) ***Ann. Appl. Stat.***

Reproducibility is Key To Improving Data Analysis

- Without code or data
  - We cannot explain *why* a given result occurred by detailing the underlying systems that produced the results
  - We cannot improve future data analyses and avoid mistakes
- But...
  - Without working feedback loops and open communication, **errors may not be prevented**
  - Code alone may be insufficient for diagnosing unexpected or surprising results

Why is Data Analysis Hard?

The Four Jobs of the Data Scientist

Roger Peng 2020/11/24

In 2019 I wrote a post about The Tentpoles of Data Science that tried to distill the key skills of the data scientist. In the post I wrote:

> When I ask myself the question "What is data science?" I tend to think of the following five components. Data science is (1) the application of design thinking to data problems; (2) the creation and management of workflows for transforming and processing data; (3) the negotiation of human relationships to identify context, allocate resources, and characterize audiences for data analysis products; (4) the application of statistical methods to quantify evidence; and (5) the transformation of data analytic information into coherent narratives and stories.
>
> My contention is that if you are a good data scientist, then you are good at all five of the tentpoles of data science. Conversely, if you are good at all five tentpoles, then you'll likely be a good data scientist.

I still feel the same way about these skills but my feeling now is that actually that post made the job of the data scientist seem easier than it is. This is because it wrapped all of these skills into a single job when in reality data science requires being good at **four** jobs. In order to explain what I mean by this, we have to step back and ask a much more fundamental question.

https://simplystatistics.org/2020/11/24/the-four-jobs-of-the-data-scientist/

Data Analytic Iteration

Activity	Role
1. Construct set of **expected outcomes**	Scientist
2. Build/Apply **analytic system** to data and recognize anomalies in output	Statistician
3. Enumerate potential **root causes**	Systems Engineer
4. Make a decision and implement revisions, balancing any **trade-offs**	Politician

Summary

- Reproducibility represents the start of a large-scale iteration where we learn about a data analysis

- Code and data are essential for describing what was done

- New representations of data analysis are needed to more easily diagnose problems in data analyses

- Reproducibility must be coupled with feedback loops, open communication to drive an improvement in data analytic quality