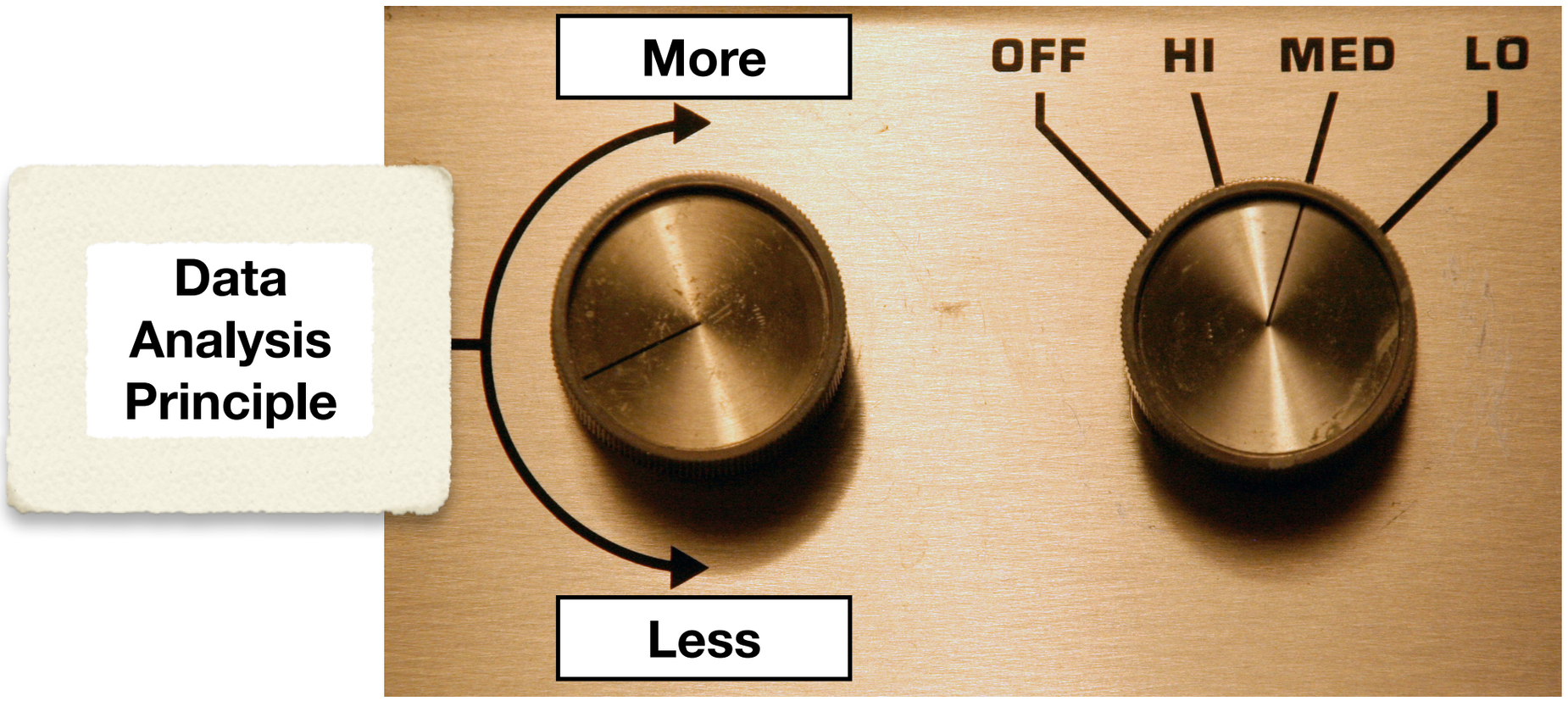
Design Principles for Data Analysis

Stephanie C. Hicks
Roger D. Peng

Johns Hopkins Bloomberg School of Public Health



What is a data analysis?

What is a data analysis?

- Does not occur naturally

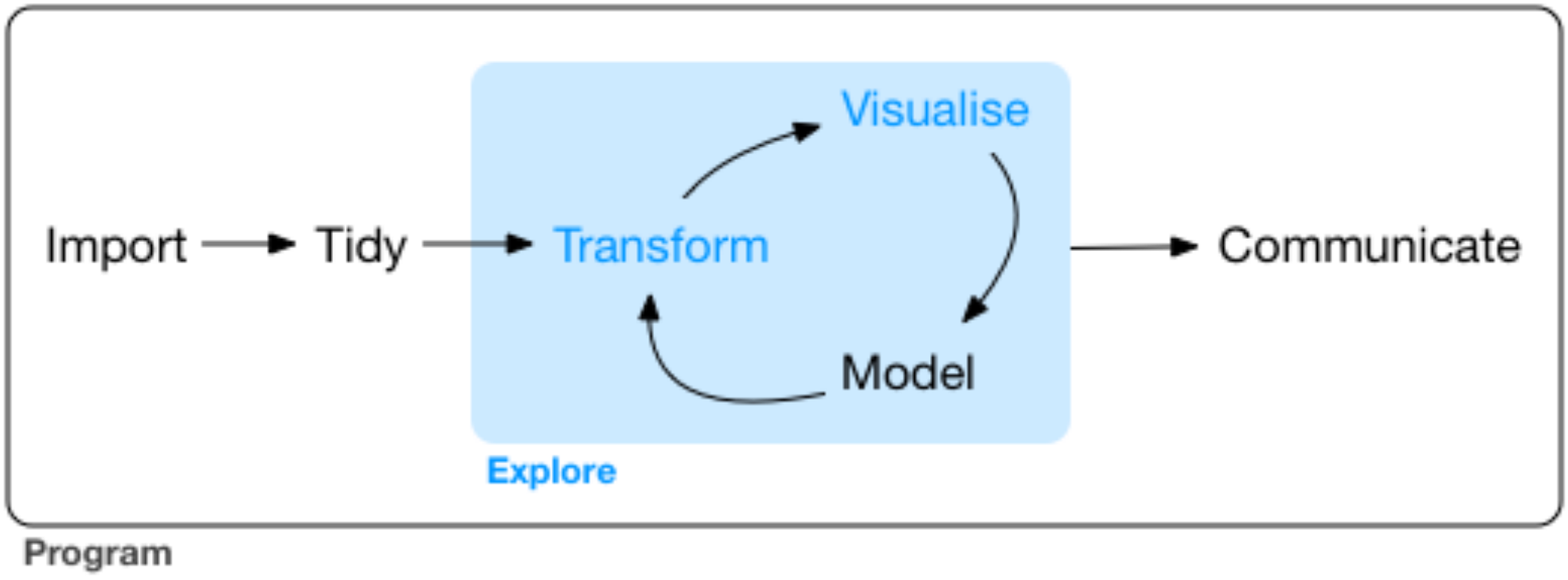
- Must be **designed** to be useful

- Solution can take many forms

- Must follow basic **structural principles** (or else collapse)

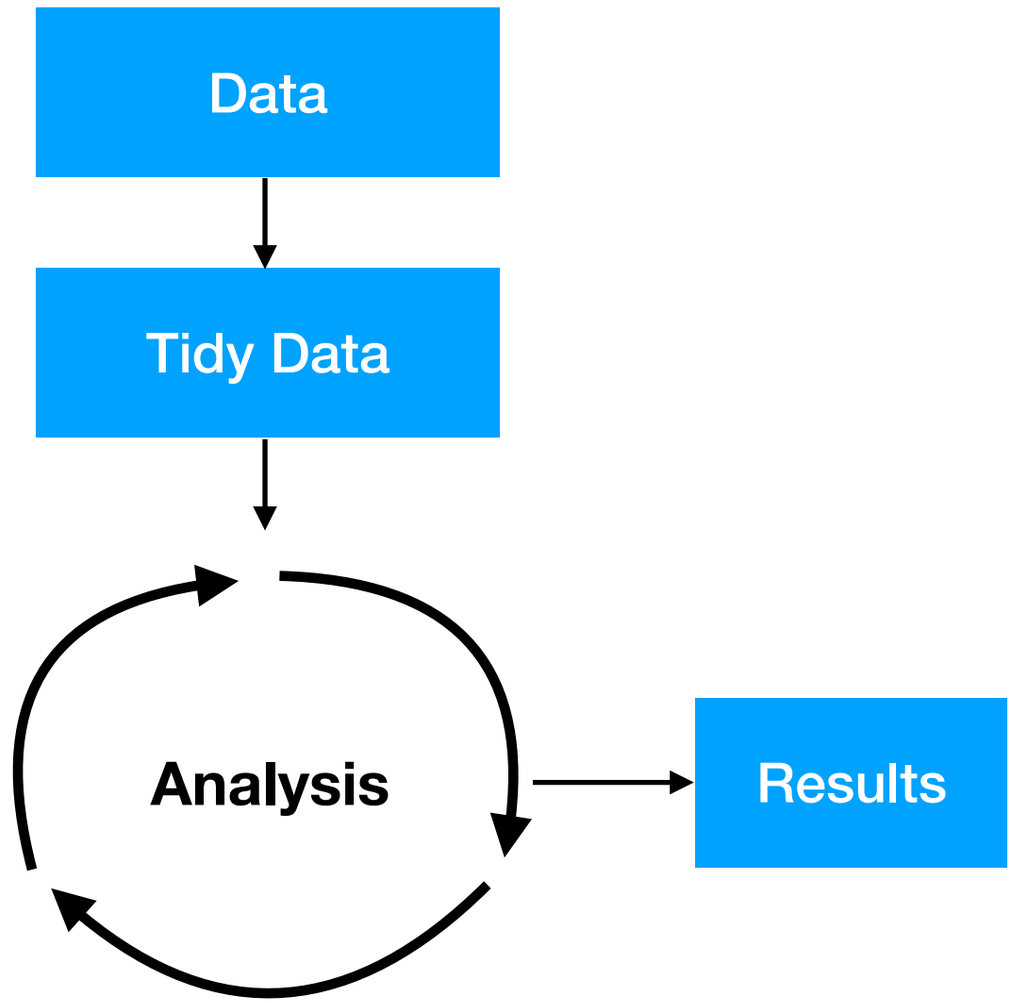
What is a data analysis?

Illustration of a data analysis workflow

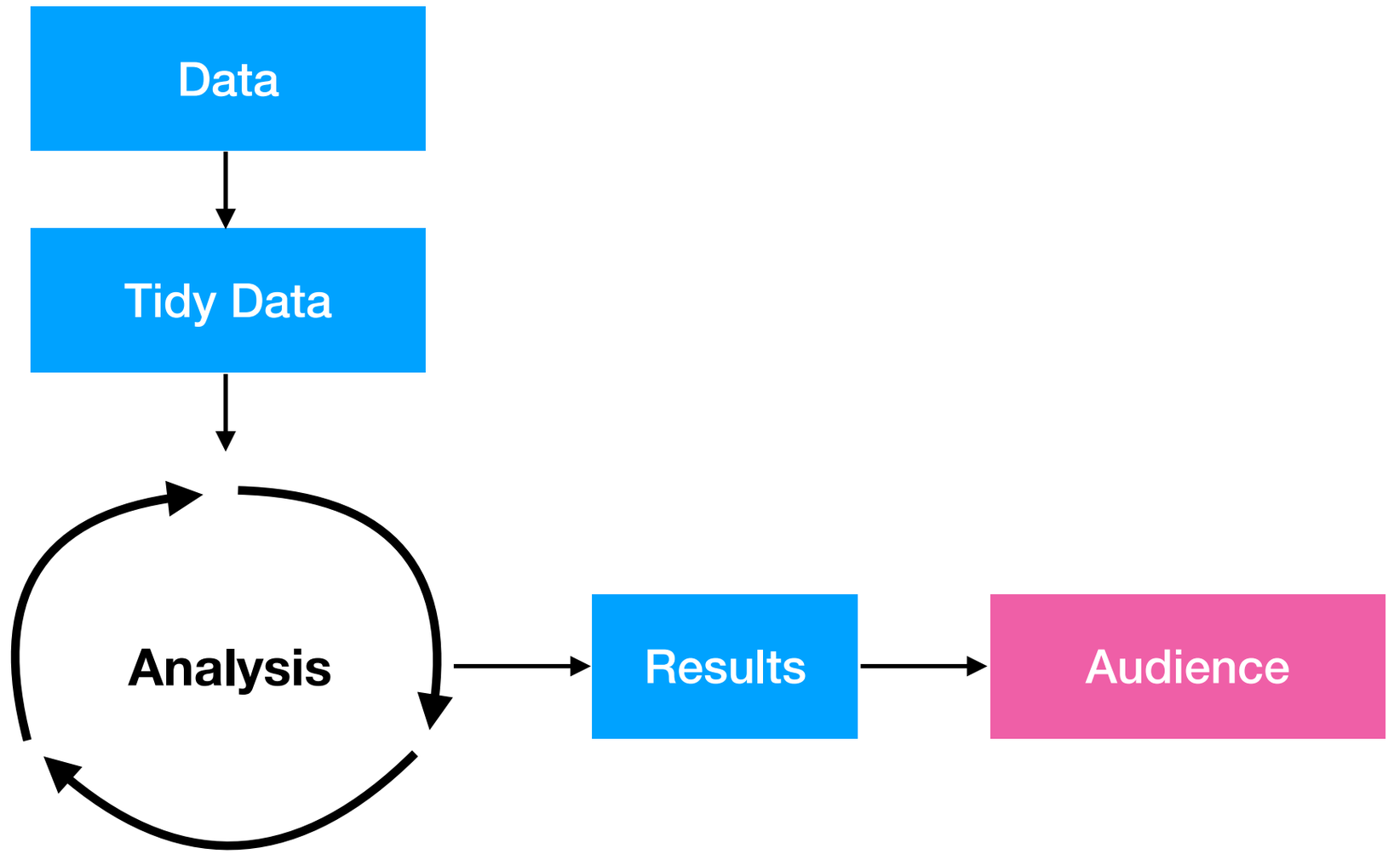


Wickham, *R for Data Science*

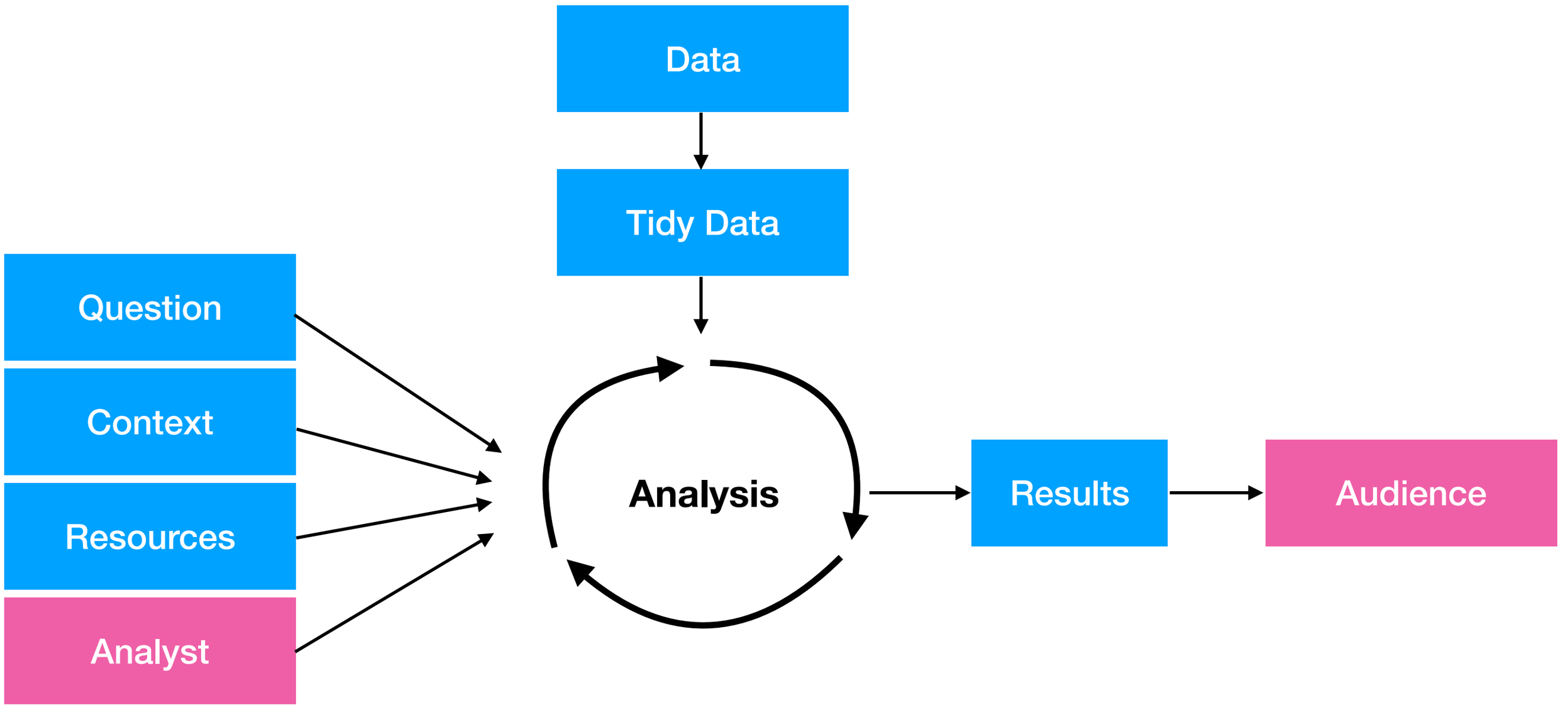
Data Analysis (revised)



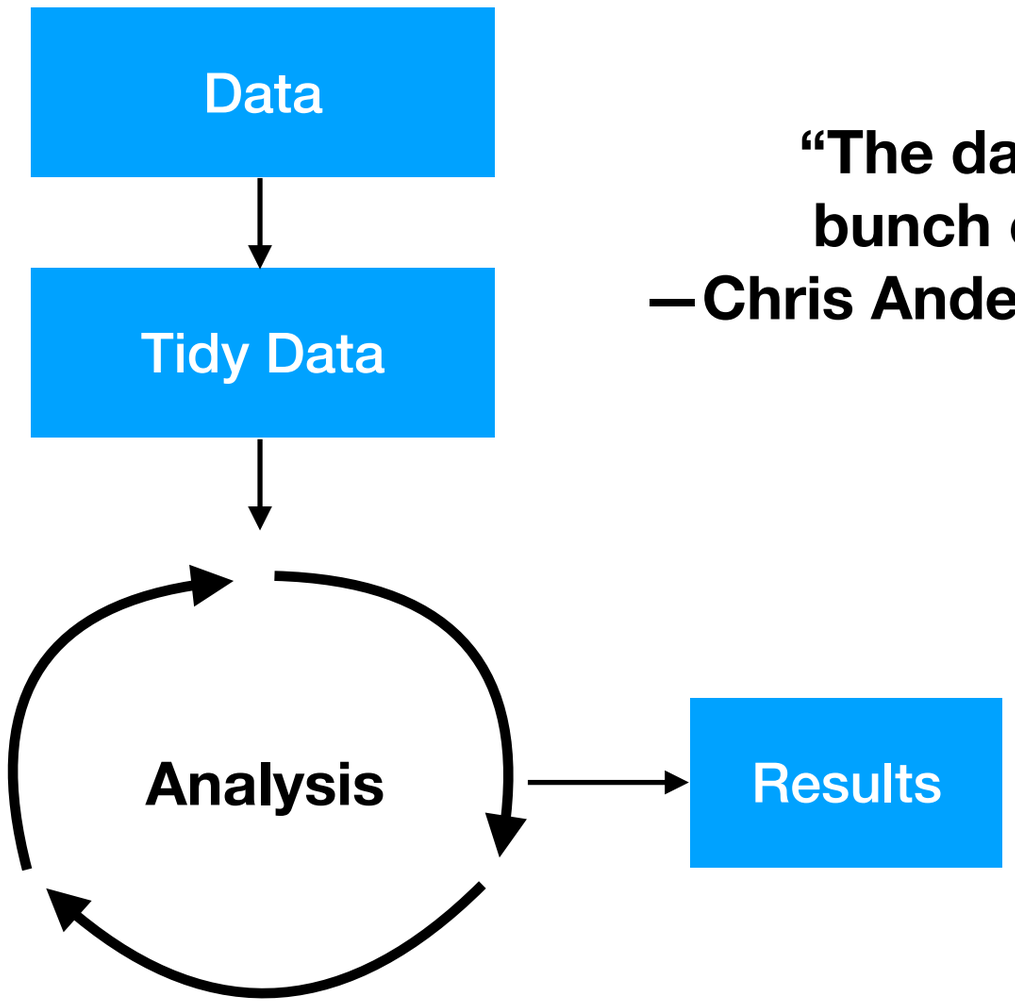
Data Analysis (revised)



Data Analysis (revised)



Data Analysis (revised)

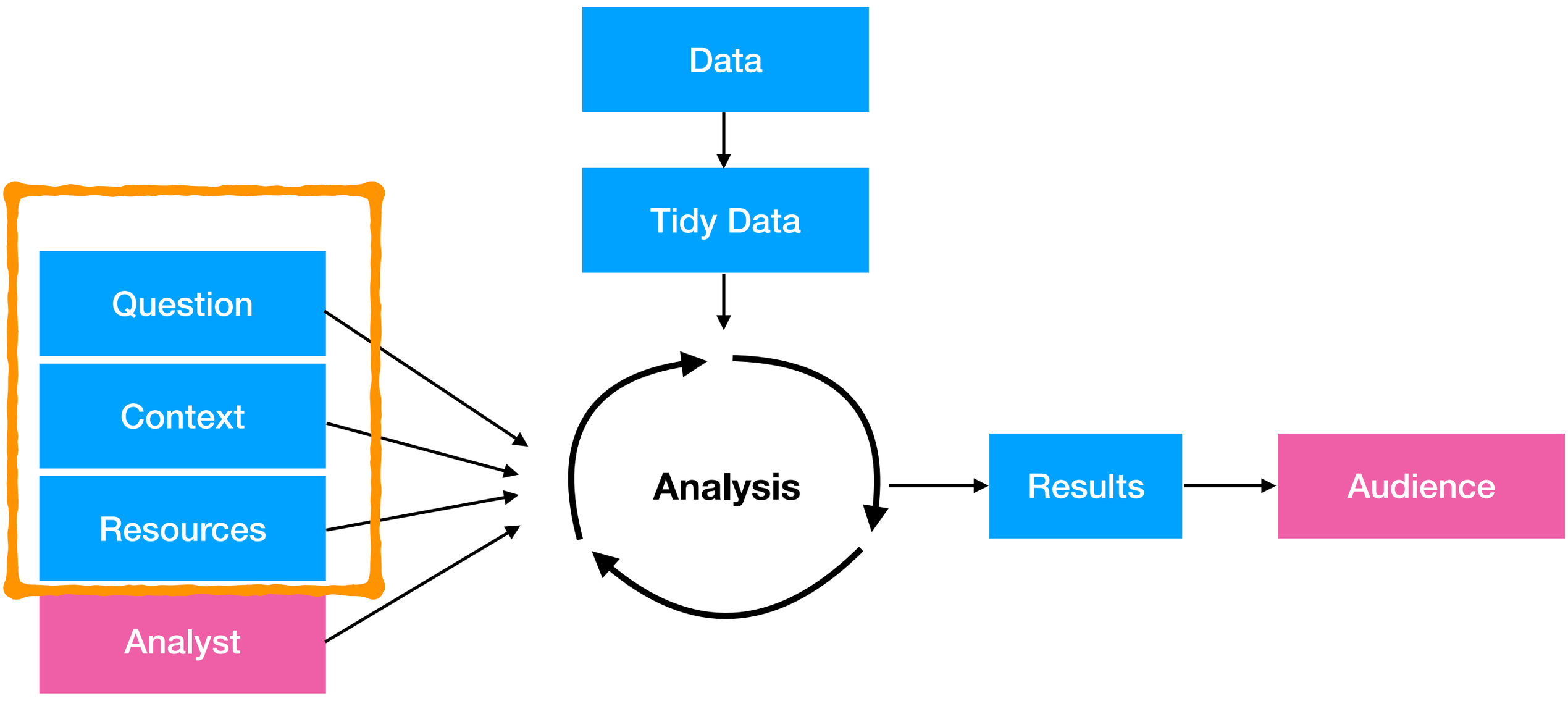


“The data are just a bunch of numbers”
—Chris Anderson (2008) *Wired*

“Do numbers ever speak for themselves? The short answer: no. The longer answer: no”

–*Catherine D’Ignazio & Lauren Klein from* Data Feminism

Data Analysis (revised)



Data Analysis Expectations

Observed data	=	Our expectation	+	Our deviation from reality

Data Analysis Expectations

Observed data = Our expectation + Our deviation from reality

Nature

Question

Context

Resources

Prior work

Results

Data Analysis Expectations

Expectation

Deviation

Observed data = Expectation + Deviation

Expectation + Deviation

Expectation + Deviation

Creating a data analysis

Statistical Thinking

Every data analyst makes analytic choices

- Methods / Approaches / Models
- Algorithms
- Tools
- Languages
- Integrated Developer Environments
- Workflows











Creating a data analysis

Statistical Thinking

Design Thinking

Advances in Methods and Practices in Psychological Science

aps | ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Journal Indexin

Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results

R. Silberzahn, E. L. Uhlmann, D. P. Martin, more...

Show all authors

First Published August 23, 2018 | Research Article | Check for updates

https://doi.org/10.1177/2515245917747646

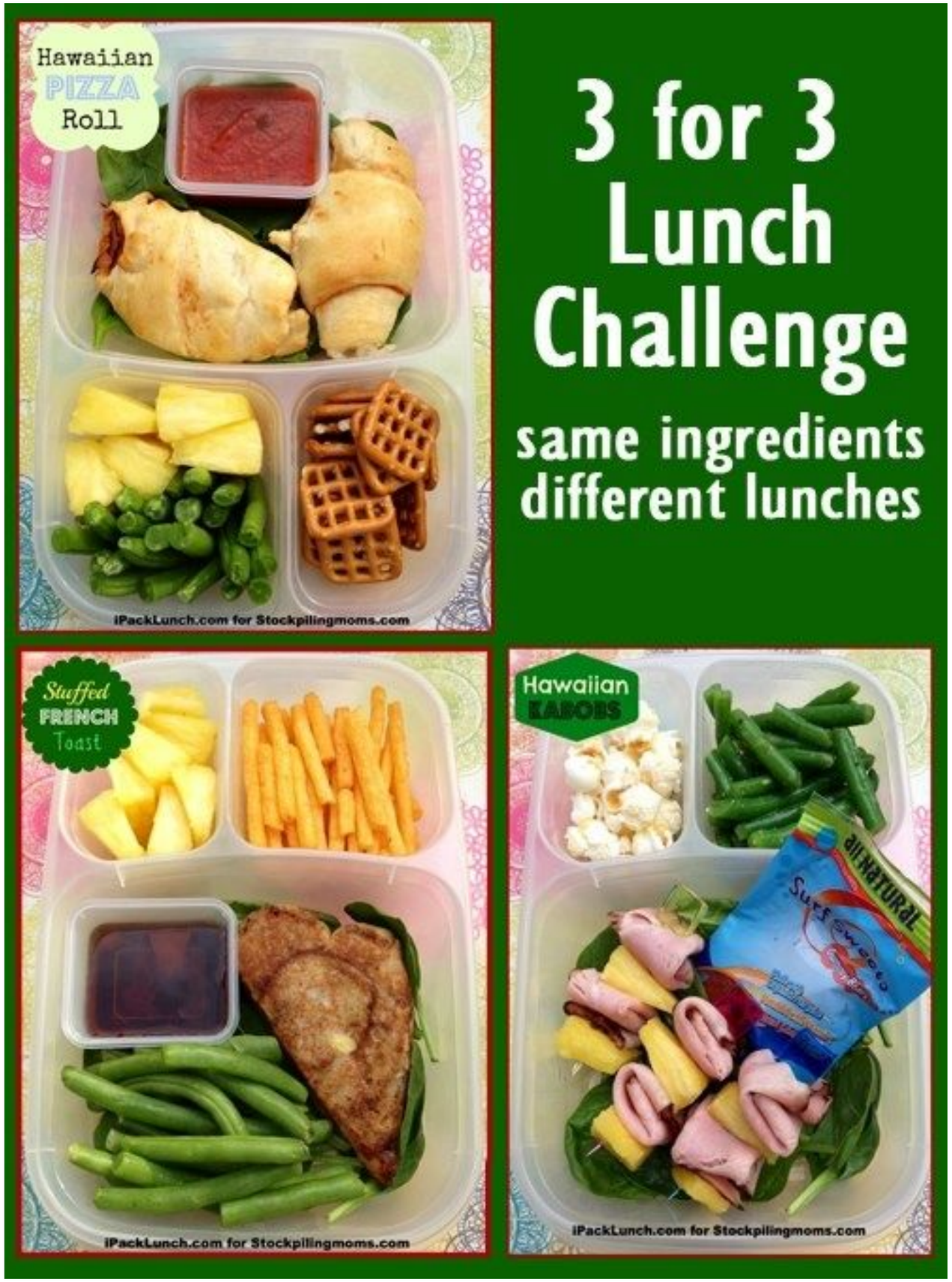
A correction has been published: Corrigendum: Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Af…

Abstract

Twenty-nine teams involving 61 analysts used the same data set to address the same research question: whether soccer referees are more likely to give red cards to dark-skin-toned players than to light-skin-toned players. Analytic approaches varied widely across the teams, and the estimated effect sizes ranged from

Same ingredients

Different meals



Pinterest

Question

“Is X associated with Y?”

Data analytic elements to investigate a question

e.g. a plot, a correlation coefficient

Question

Data analytic elements to investigate a question

Data analyst selects one element to investigate the question

Question

Analysis

Element 1

Data analytic elements to investigate a question

Data analyst selects one element to investigate the question

Question

Analysis

Element 1

Data analytic elements to investigate a question

Result

Question

Data analyst selects one element to investigate the question

Data analyst selects multiple elements to investigate the question

Analysis

Element 1

Analysis

Element 1

Element 2

Element 3

Element 4

...

Element E

Data analytic elements to investigate a question

Result

Question

Data analyst selects one element to investigate the question

Analysis

Element 1

Data analyst selects multiple elements to investigate the question

Analysis

Element 1

Element 2

Element 3

Element 4

...

Element E

Data analytic elements to investigate a question

Result

Less **Exhaustive** ***More***

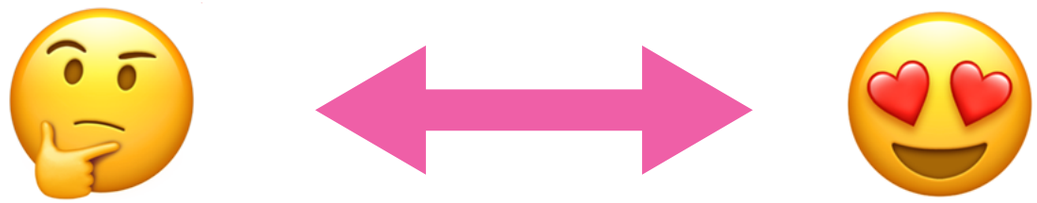
Data Analysis Alignment

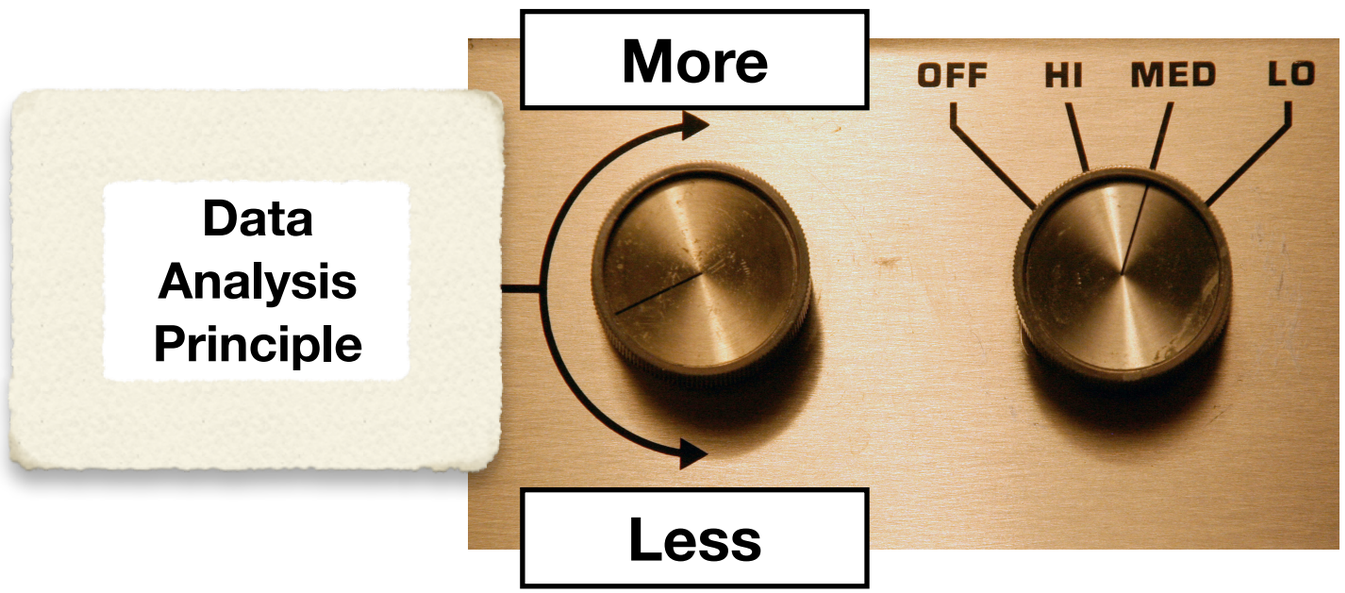
Exhaustive

Analyst

Audience

Exhaustive





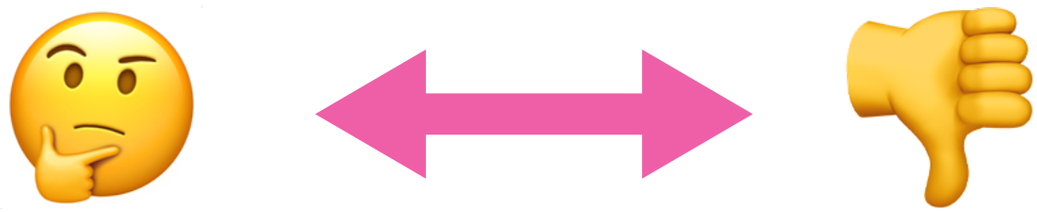
Data Analysis Misalignment

Exhaustive

Analyst

Audience

Exhaustive



Data Analysis Principles

- Reflect objective qualities of a data analysis
- Inclusion or exclusion of principles ≠ judgment or assessment of quality
- Characteristics can be highly influenced by outside constraints or resources (e.g. time or budget)
- A data analyst assigns different weights of the principles which leads to different data analyses (all addressing the same question)

What is an Aligned Data Analysis?

- Data analyses must be designed to be useful
- Design thinking
    - Identify the problem —> Exploratory data analysis
    - Build the solution —> Modeling, uncertainty, narrative
- Audience has wants and expectations

What is an Aligned Data Analysis?

- Every analysis has at least two roles
    - **Analyst** - conducts / leads the data analysis
    - **Audience** - reviews / reads / receives the analysis
        - Analyst and audience may be played by the same actor
- Both analyst and audience weigh a set of **principles** that articulate what they think is important about an analysis
- An aligned analysis is one in which the audience **accepts** what the analyst as done and **agrees on the weighting of principles** chosen by the analyst

What is an Aligned Data Analysis?

Principle 1

Principle 2

Principle 3

Principle 4

...

Principle N

Analyst

Audience

Principle 1

Principle 2

Principle 3

Principle 4

...

Principle N

What is an Aligned Data Analysis?

Principle 1

Principle 2

Principle 3

Principle 4

...

Principle N

Analyst

Audience

Principle 1

Principle 2

Principle 3

Principal 4

...

Principle N

Consider

- Analyst *i*
- Audience *j*
- Data analytic principle *k*

Assume we have weights (random) for analyst $\psi_i^{(k)}$ and audience $\alpha_j^{(k)}$, we can write the principle-specific weight difference for a given data analysis as

$$D_{ij}^{(k)} = \psi_i^{(k)} - \alpha_j^{(k)}$$

The overall analyst-audience distance for a given data analysis is then characterized by the collection of distances for the set of *K* principles

$$\mathbf{D}_{ij} = \left(D_{ij}^{(1)}, \ldots, D_{ij}^{(K)} \right)$$

Defining an Aligned Data Analysis

Strong Pairwise Alignment

$$\left\| \mathbf{D}_{ij} \right\|_{\infty} = \max_{k=1,\ldots,K} \left| D_{ij}^{(k)} \right| < \varepsilon$$

Assume the $D_{ij}^{(k)}$ values can never be equal to zero.

The definition of *strong pairwise alignment* requires that the differences are never too large for any given principle.

Defining an Aligned Data Analysis

Weak Pairwise Alignment

$$\left\| \mathbf{D}_{ij} \right\|_p = \left(\frac{1}{K} \sum_{k=1}^{K} \left| D_{ij}^{(k)} \right|^p \right)^{1/p} < \varepsilon$$

Here, the analyst and audience may differ slightly wrt how each principle is weighted, but overall differences between analyst and audience must be small.

The choice of *p* here (and hence, the norm) will have an impact on how much deviation is allowed between analyst and audience and how much any single principle may differ.

Different circumstances may require the use of different norms.

Defining an Aligned Data Analysis

Potential Pairwise Alignment

$$\mathbb{E}\left[\mathbf{D}_{ij}\right] = \mathbf{0}$$

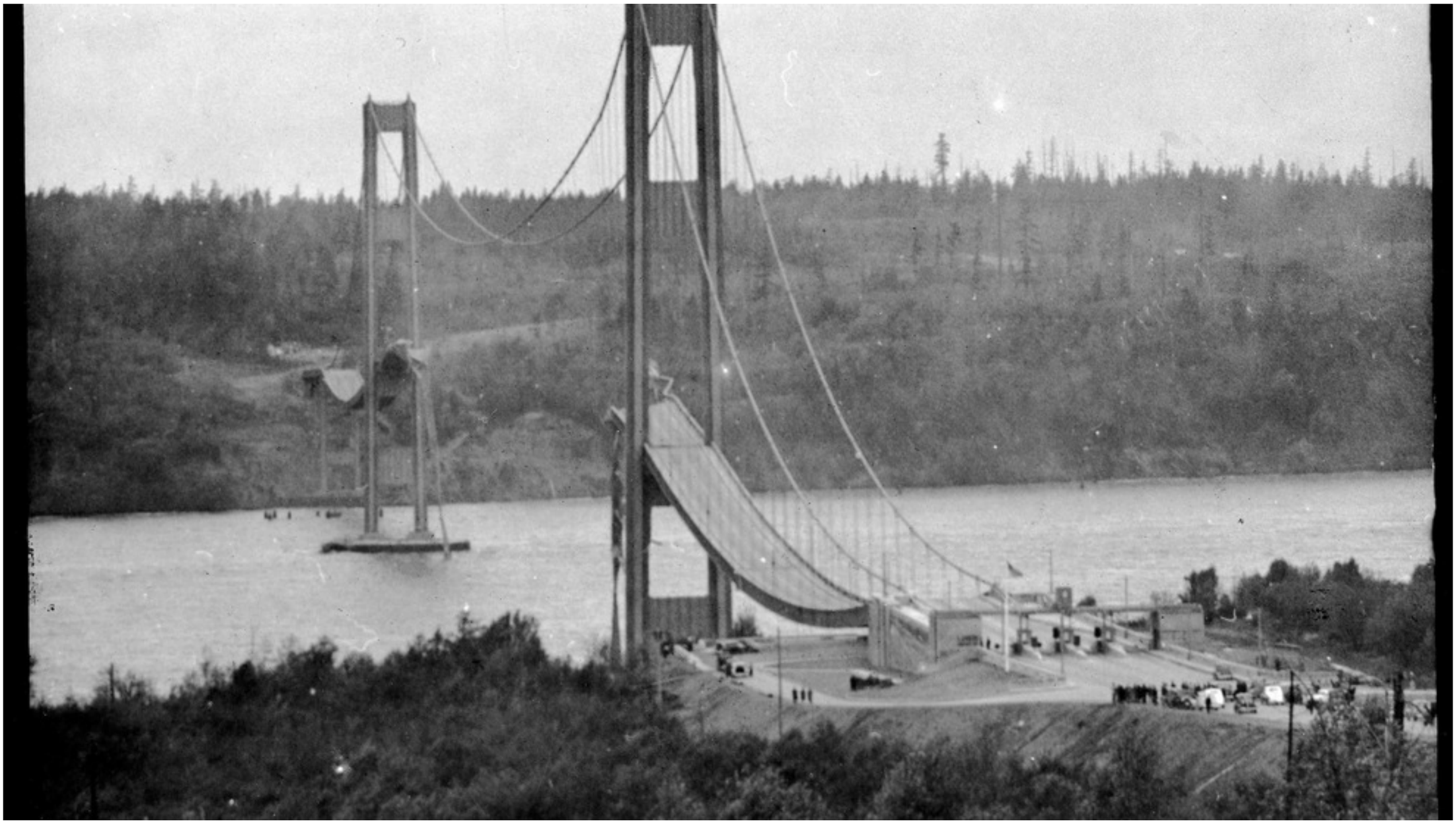
Key distinction between *strong* (or *weak*) pairwise alignment and *potential* pairwise alignment is:

- the former can only be evaluated when analyst and audience meet and a data analysis is presented
- *potential* pairwise alignment can be evaluated before an analyst presents the analysis to the audience

Hence

- *potential* pairwise alignment metric could serve as a target for optimization by the analyst

Give People What They Want?



Aligning Data Analyses

- Data analyses should be designed based on a set of shared principles

- Identifying relevant principles **requires consideration of the audience**

- Even a good analysis can be “misaligned”

- Aligned analyses are different from valid, honest, complete, etc

- Future: use principles as a form of **intervention**

Thank You!

- Design Principles for Data Analysis. https://arxiv.org/abs/2103.05689

- Evaluating the Alignment of a Data Analysis. https://arxiv.org/abs/1904.11907