Data Analyses: A New Deal

Stephanie C. Hicks
Roger D. Peng

Homework 1 available!

https://classroom.github.com/a/mM8AuMDF

(Will also be posted on Slack after class today)

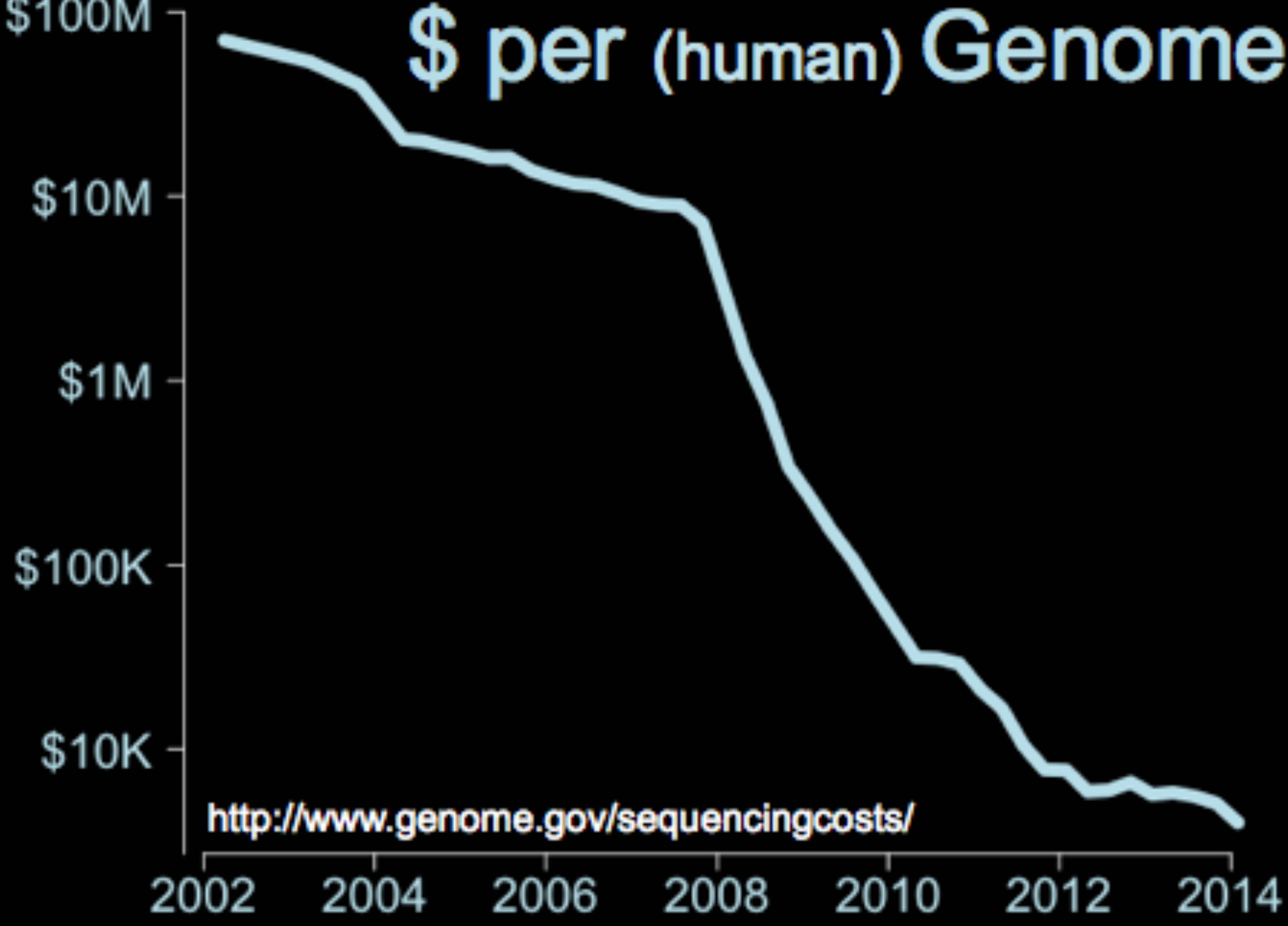
Protecting Health, Saving Lives —

Millions ↑ at a Time

(of data points)

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

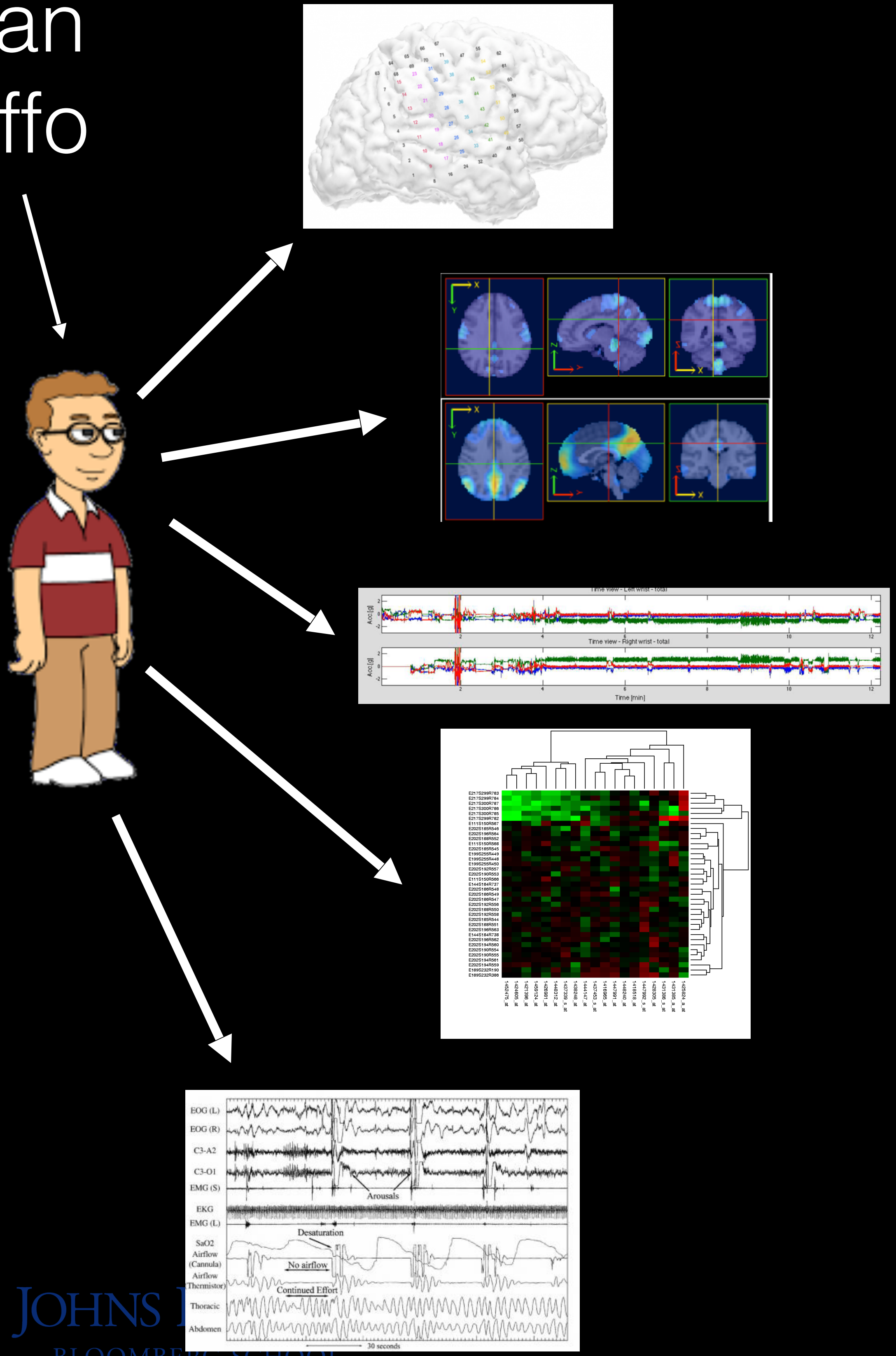
The Measurement Revolution…



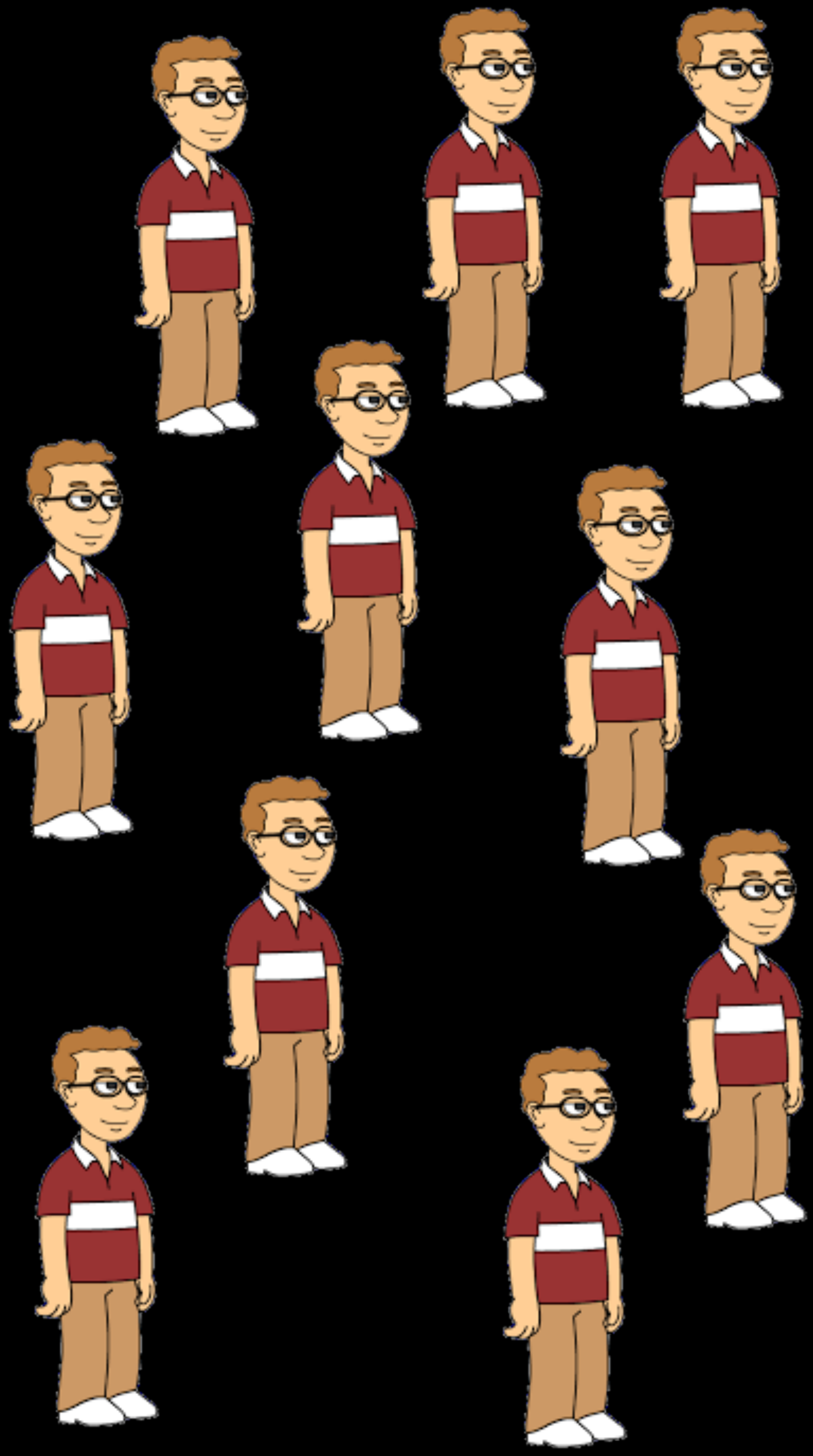
http://www.genome.gov/sequencingcosts/

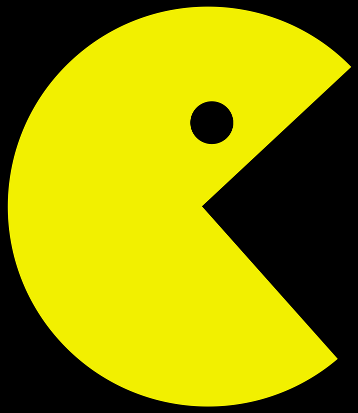
The Measurement Revolution…

Brian Caffo



Multiplied!



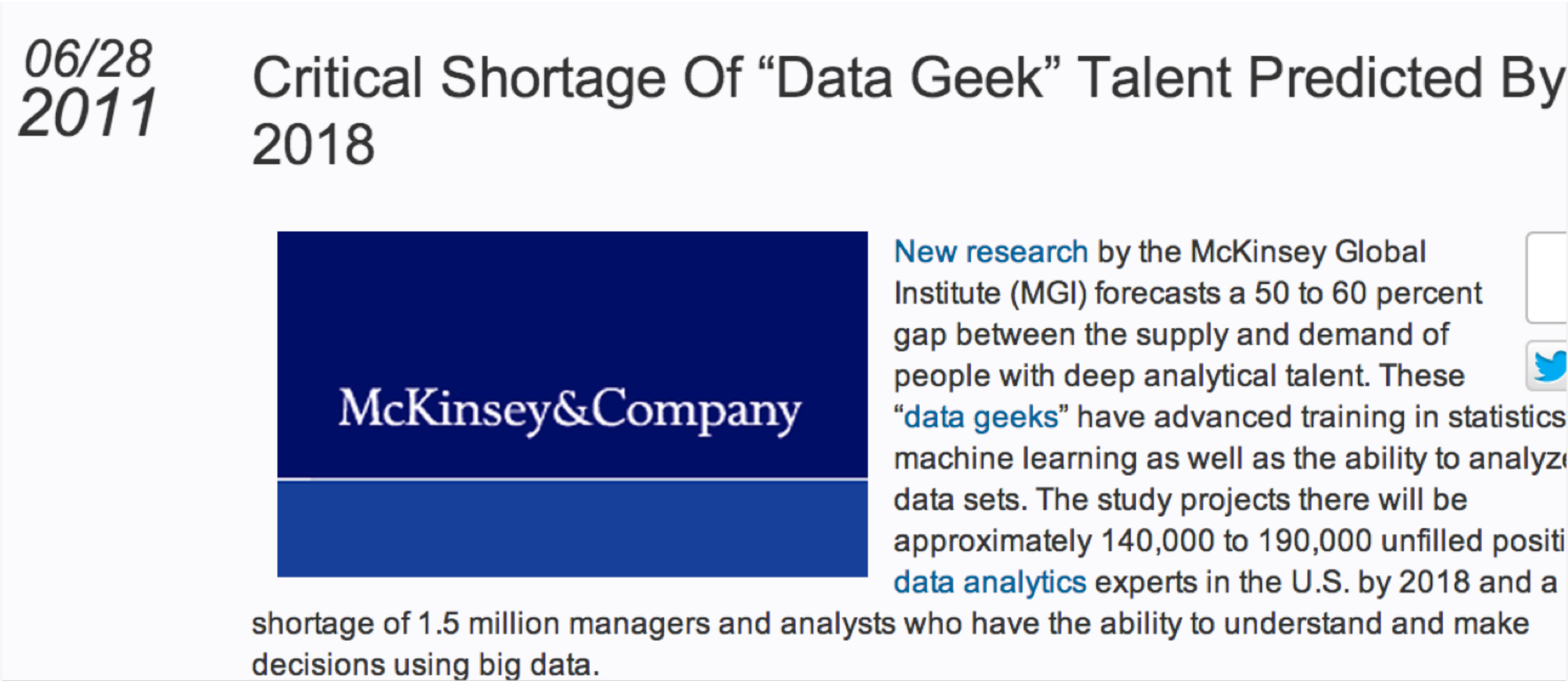


Data are Eating the World*

(but analysis hasn’t yet)

*see “Software is Eating the World” by Marc Andreessen

Demand for Data Science



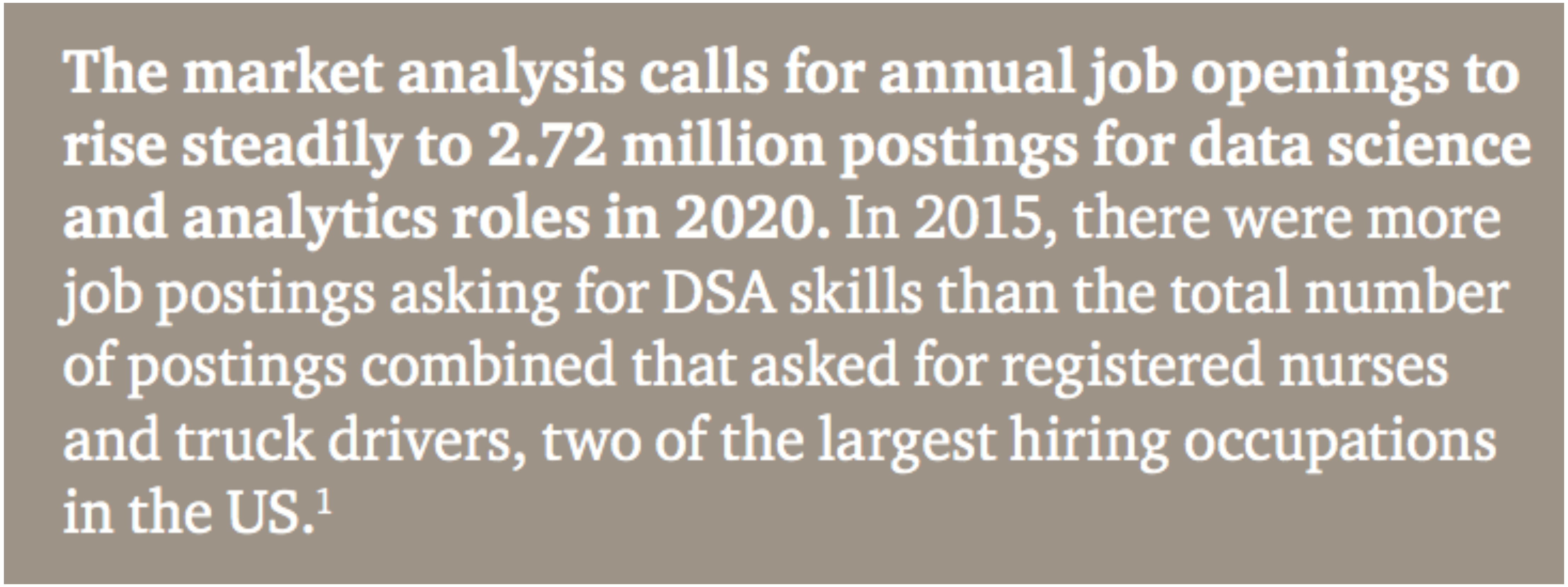
06/28
2011

Critical Shortage Of “Data Geek” Talent Predicted By 2018

McKinsey&Company

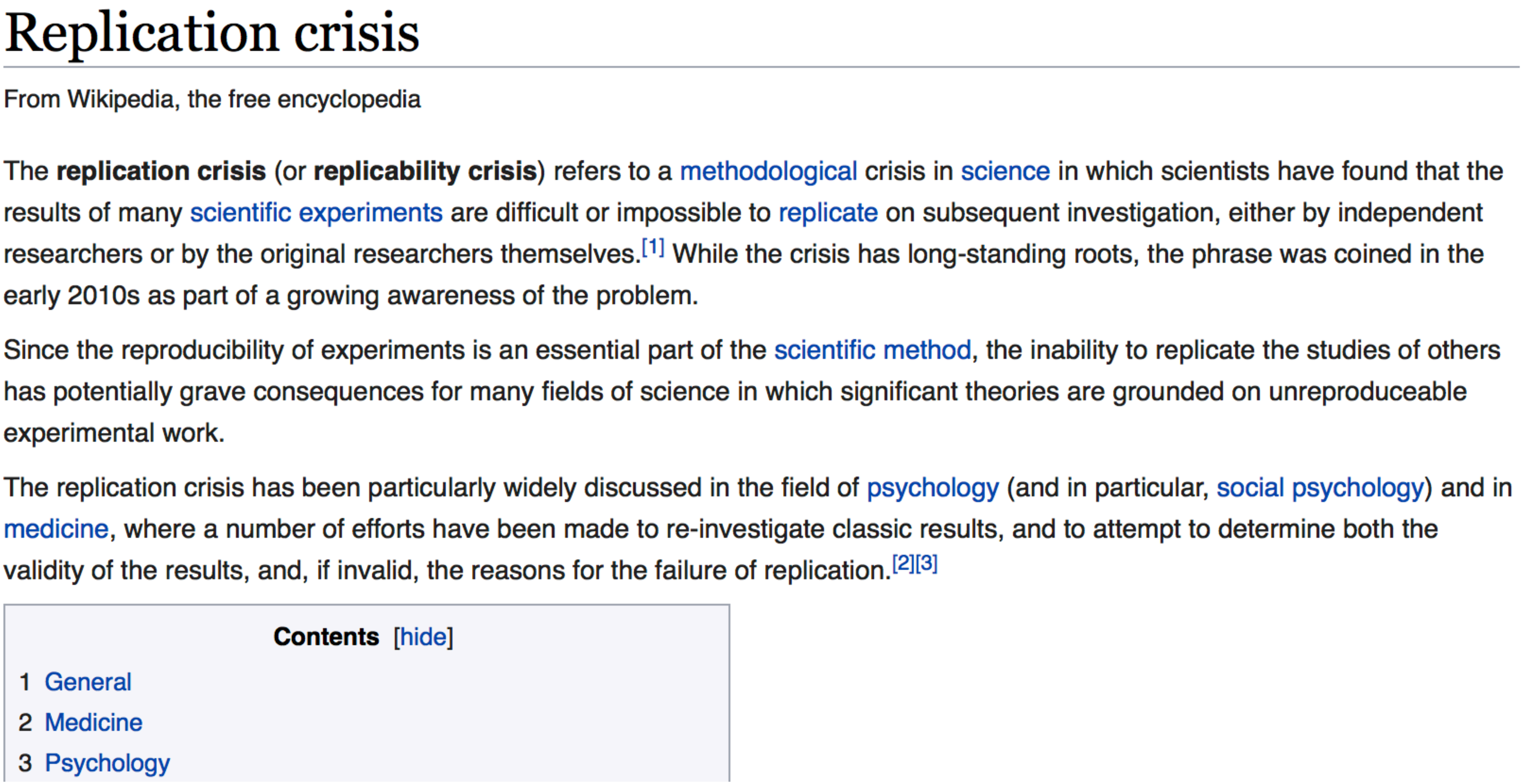
New research by the McKinsey Global Institute (MGI) forecasts a 50 to 60 percent gap between the supply and demand of people with deep analytical talent. These “data geeks” have advanced training in statistics machine learning as well as the ability to analyz data sets. The study projects there will be approximately 140,000 to 190,000 unfilled positi data analytics experts in the U.S. by 2018 and a shortage of 1.5 million managers and analysts who have the ability to understand and make decisions using big data.

Demand for Data Science



The market analysis calls for annual job openings to rise steadily to 2.72 million postings for data science and analytics roles in 2020. In 2015, there were more job postings asking for DSA skills than the total number of postings combined that asked for registered nurses and truck drivers, two of the largest hiring occupations in the US.[1]

Investing in America's Data Science and Analytics Talent: The Case for Action,
PricewaterhouseCoopers



Replication crisis

From Wikipedia, the free encyclopedia

The **replication crisis** (or **replicability crisis**) refers to a methodological crisis in science in which scientists have found that the results of many scientific experiments are difficult or impossible to replicate on subsequent investigation, either by independent researchers or by the original researchers themselves.[1] While the crisis has long-standing roots, the phrase was coined in the early 2010s as part of a growing awareness of the problem.

Since the reproducibility of experiments is an essential part of the scientific method, the inability to replicate the studies of others has potentially grave consequences for many fields of science in which significant theories are grounded on unreproduceable experimental work.

The replication crisis has been particularly widely discussed in the field of psychology (and in particular, social psychology) and in medicine, where a number of efforts have been made to re-investigate classic results, and to attempt to determine both the validity of the results, and, if invalid, the reasons for the failure of replication.[2][3]

Contents [hide]

Challenges

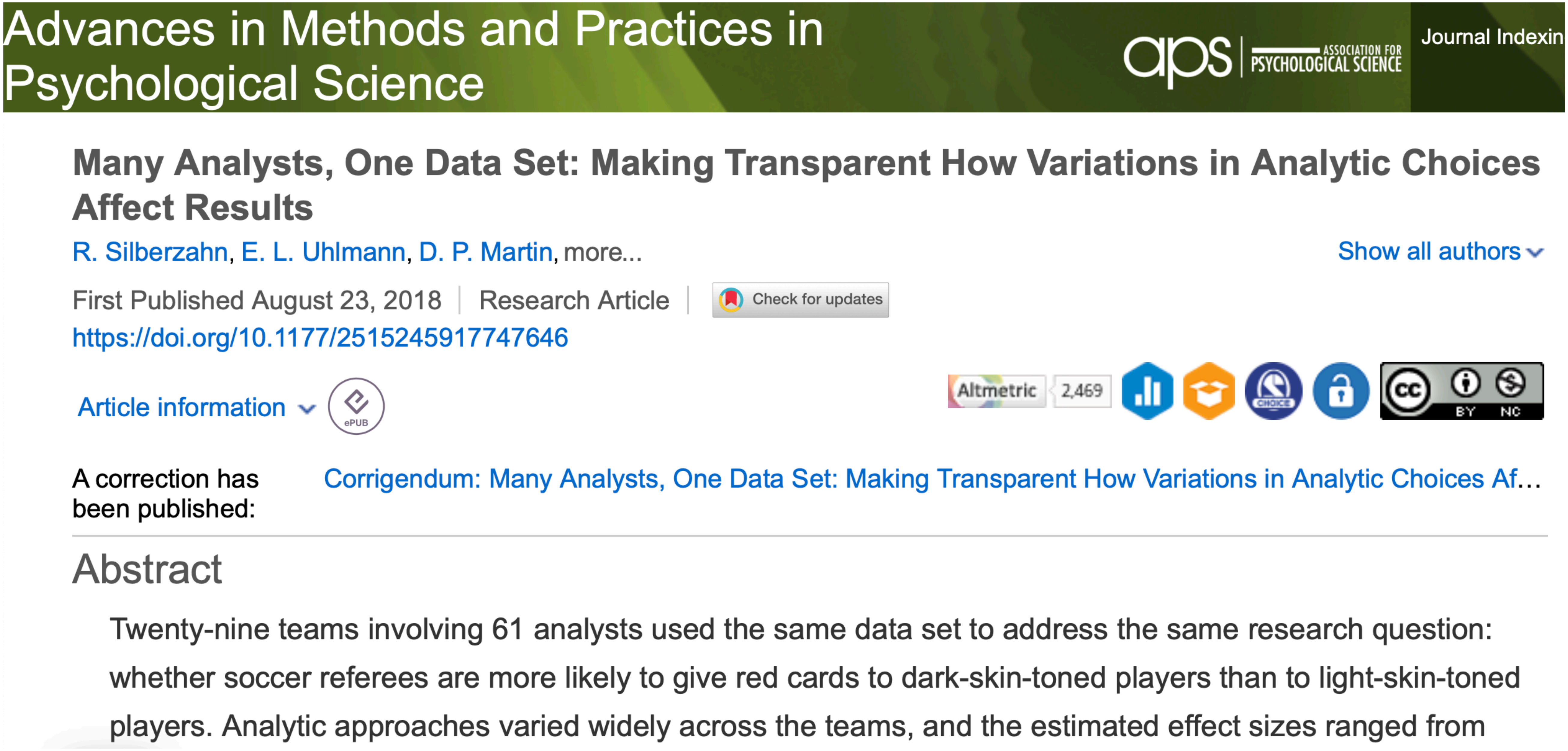
- Data analysis is everywhere because data are everywhere
- Everyone is a data analyst, whether they like it or not!
- Training for data analysis is essential but there is limited bandwidth
- Our understanding of the data analysis process is fundamentally narrow

Questions

- What is a data analysis?
- What are differences between analyses?
- What is a successful data analysis?

Every data analyst makes analytic choices

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



Advances in Methods and Practices in Psychological Science

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

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

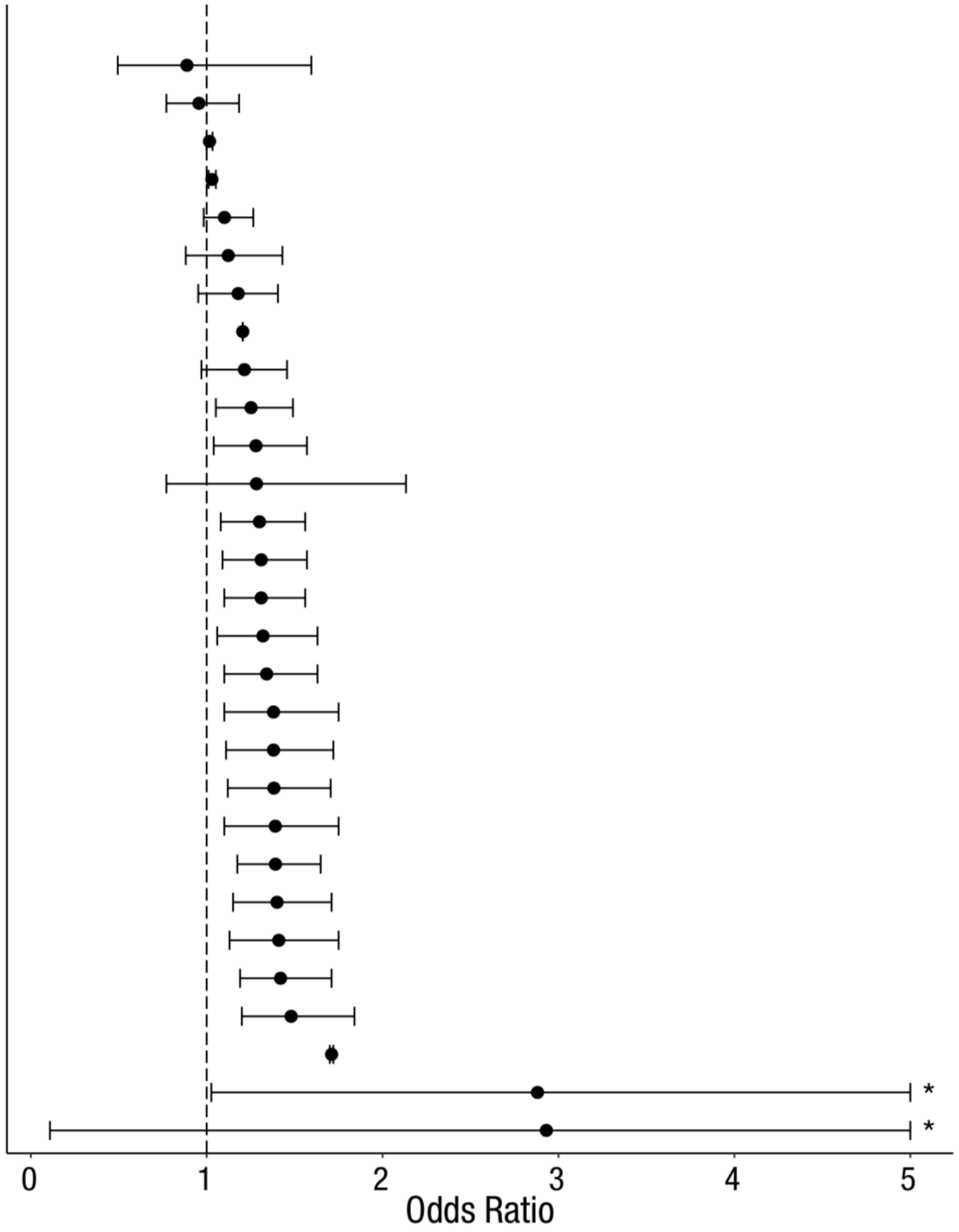
First Published August 23, 2018 | Research Article |
https://doi.org/10.1177/2515245917747646

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

Team	Analytic Approach	Odds Ratio
12	Zero-Inflated Poisson Regression	0.89
17	Bayesian Logistic Regression	0.96
15	Hierarchical Log-Linear Modeling	1.02
10	Multilevel Regression and Logistic Regression	1.03
18	Hierarchical Bayes Model	1.10
31	Logistic Regression	1.12
1	OLS Regression With Robust Standard Errors, Logistic Regression	1.18
4	Spearman Correlation	1.21
14	WLS Regression With Clustered Standard Errors	1.21
11	Multiple Linear Regression	1.25
30	Clustered Robust Binomial Logistic Regression	1.28
6	Linear Probability Model	1.28
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	1.30
3	Multilevel Logistic Regression Using Bayesian Inference	1.31
23	Mixed-Model Logistic Regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear Probability Model, Logistic Regression	1.34
5	Generalized Linear Mixed Models	1.38
24	Multilevel Logistic Regression	1.38
28	Mixed-Effects Logistic Regression	1.38
32	Generalized Linear Models for Binary Data	1.39
8	Negative Binomial Regression With a Log Link	1.39
20	Cross-Classified Multilevel Negative Binomial Model	1.40
13	Poisson Multilevel Modeling	1.41
25	Multilevel Logistic Binomial Regression	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	1.48
7	Dirichlet-Process Bayesian Clustering	1.71
21	Tobit Regression	2.88
27	Poisson Regression	2.93



Point estimates (smallest effect size at top) and 95% CIs for the effect of soccer players' skin tone on the number of red cards awarded by referees

Twenty teams (69%) found a statistically significant positive effect, and 9 teams (31%) did not observe a significant relationship

Neither analysts' prior beliefs about the effect of interest nor their level of expertise readily explained the variation in the outcomes of the analyses.

Peer ratings of the quality of the analyses also did not account for the variability.

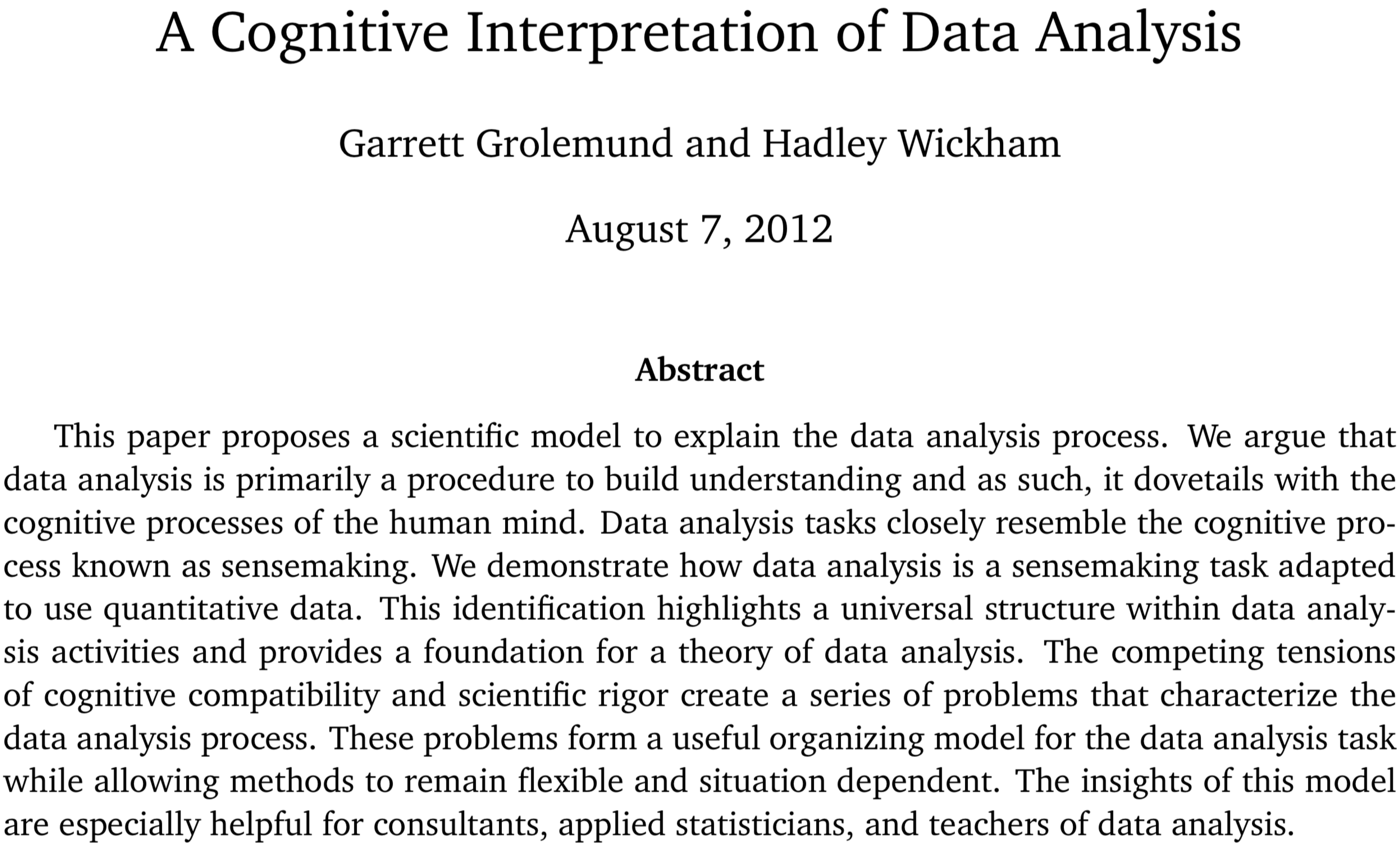
So how can we understand how data analyst make these analytic choices?

Incorporating Statistical Expertise Into Software (1991)

“Throughout American or even global industry, there is much advocacy of statistical process control and of understanding processes. **Statisticians have a process they espouse but do not know anything about**. It is the process of putting together many tiny pieces, the **process called data analysis**, and **is not really understood**.”

Daryl Pregibon, *The Future of Statistical Software*, NRC Report 1991

https://vita.had.co.nz/papers/sensemaking.pdf



A Cognitive Interpretation of Data Analysis

Garrett Grolemund and Hadley Wickham

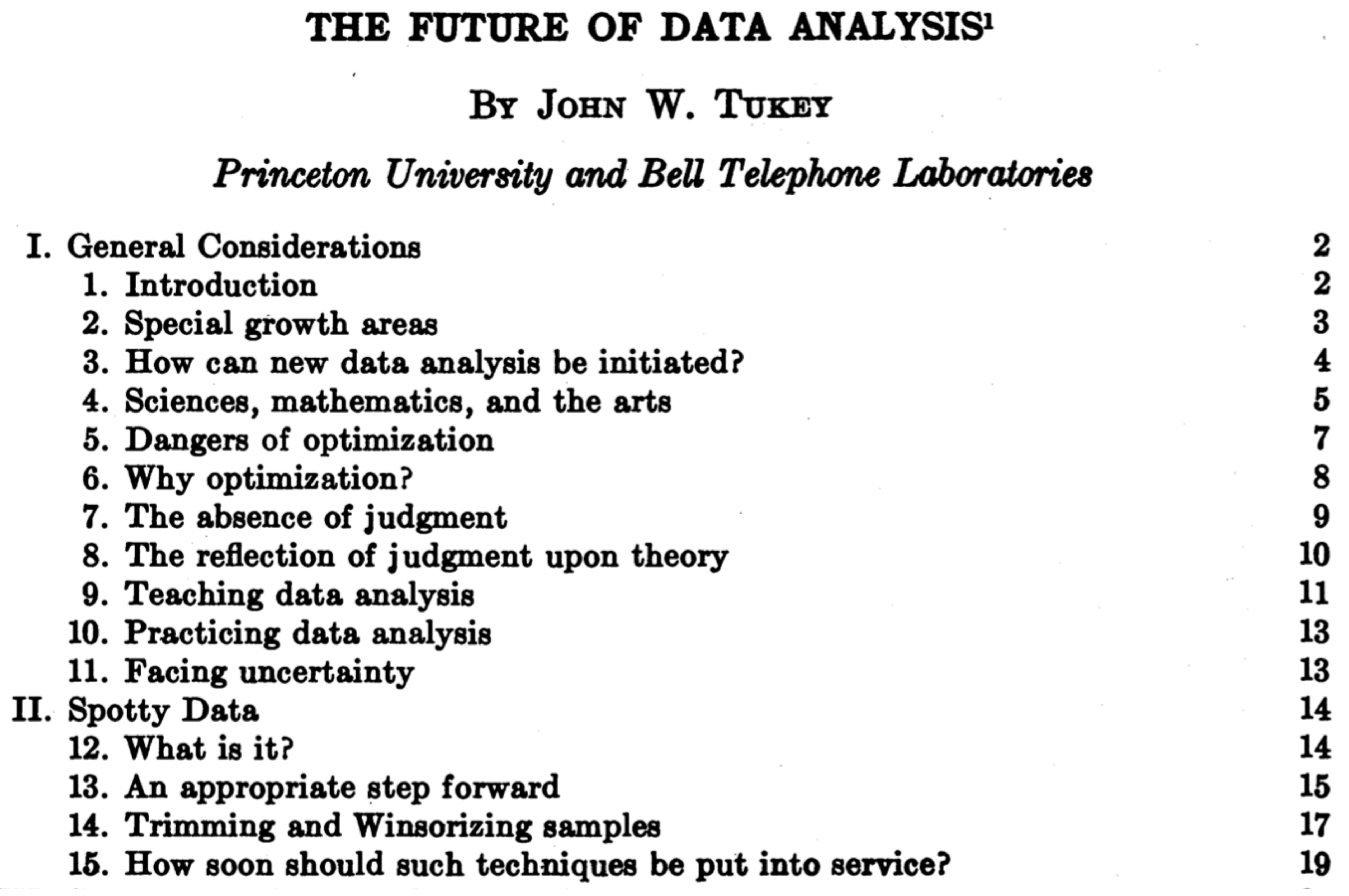
August 7, 2012

Abstract

This paper proposes a scientific model to explain the data analysis process. We argue that data analysis is primarily a procedure to build understanding and as such, it dovetails with the cognitive processes of the human mind. Data analysis tasks closely resemble the cognitive process known as sensemaking. We demonstrate how data analysis is a sensemaking task adapted to use quantitative data. This identification highlights a universal structure within data analysis activities and provides a foundation for a theory of data analysis. The competing tensions of cognitive compatibility and scientific rigor create a series of problems that characterize the data analysis process. These problems form a useful organizing model for the data analysis task while allowing methods to remain flexible and situation dependent. The insights of this model are especially helpful for consultants, applied statisticians, and teachers of data analysis.

The data analysis process is characterized as a sensemaking task whereby theories or expectations are set and then compared to reality (data) — any difference are further examined and then theories are modified

Problems with this approach
Process is not observable from outsiders (aka not the analyst)



THE FUTURE OF DATA ANALYSIS[1]

BY JOHN W. TUKEY

Princeton University and Bell Telephone Laboratories

Tukey, *Annals of Statistics*, 1962

We would teach [data analysis] like biochemistry, with emphasis on **what we have learned**…with relegation of all question of detailed methods to the “laboratory work”. All study of detailed proofs…or comparisons of ways of presentation would belong in “the laboratory” rather than “in class”.

J. W. Tukey, “The Future of Data Analysis,” 1962

The (New) Future of Data Analysis

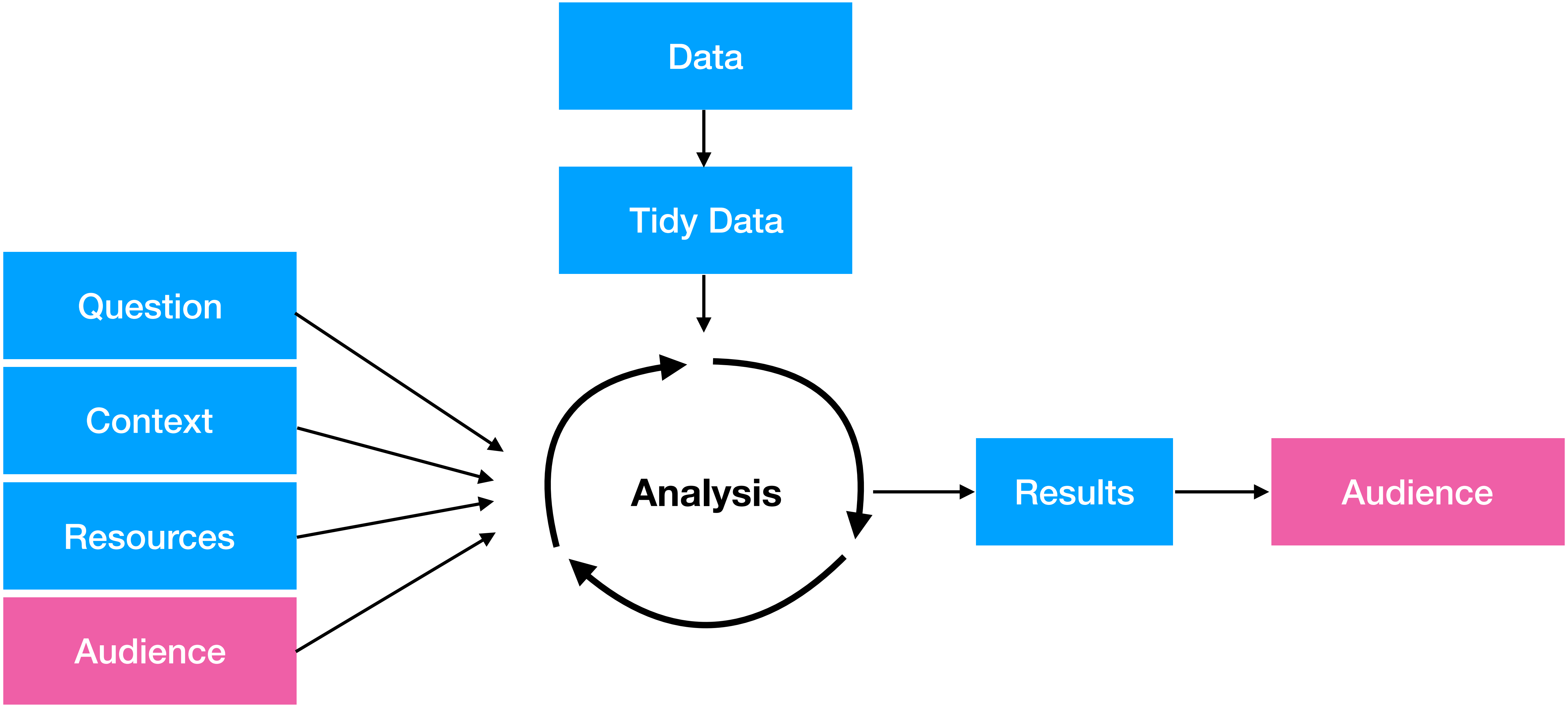
- We need to draw a **bigger picture** of the data analysis process (and own the training)

- Define what **success** and **failure** mean in data analysis

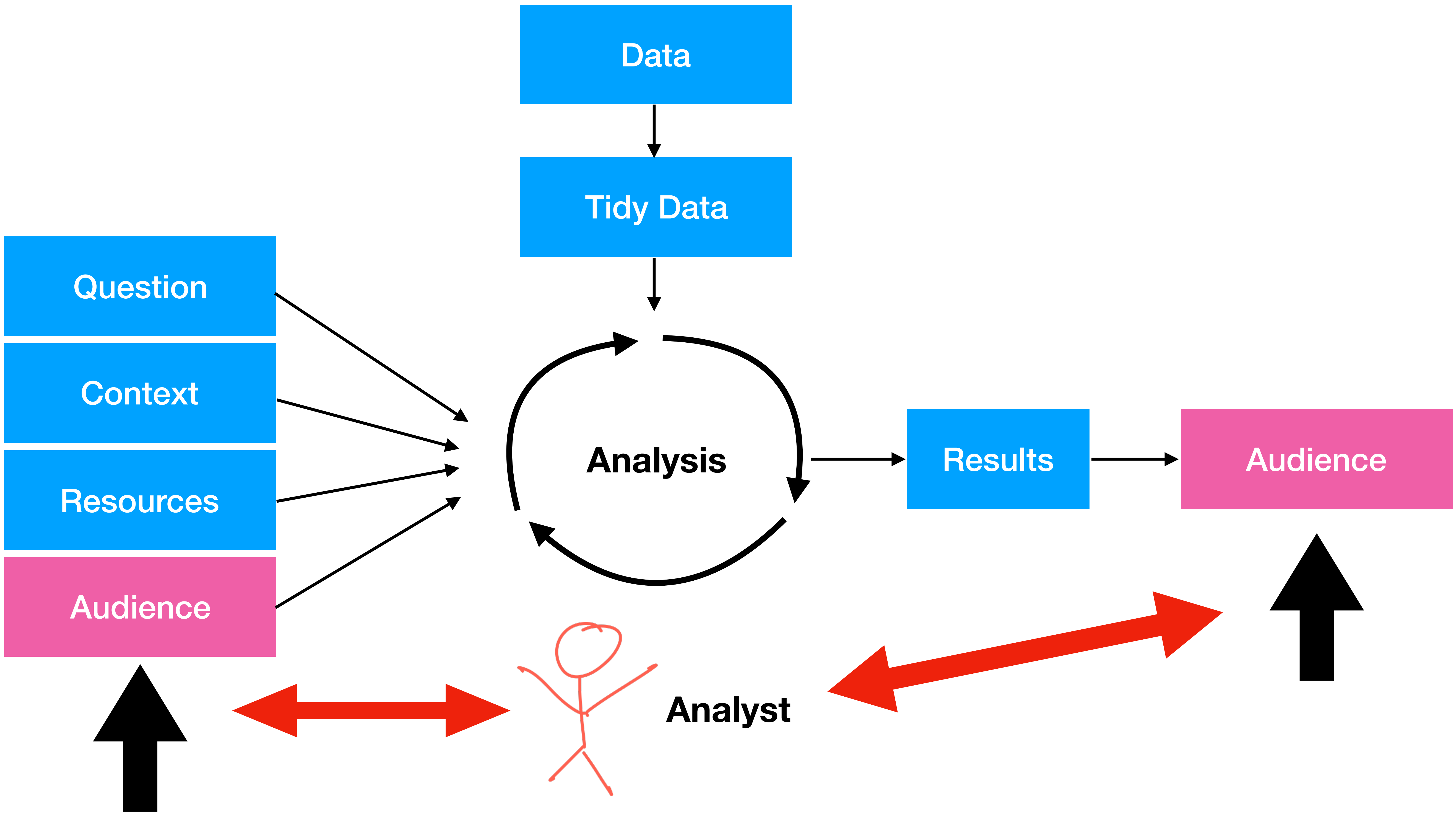
- **Design** data analyses to be successful

- **Learn** the correct lessons from failure

Data Analysis (revised)



Data Analysis (revised)


Data

Tidy Data

Question

Context

Resources

Audience

Analysis

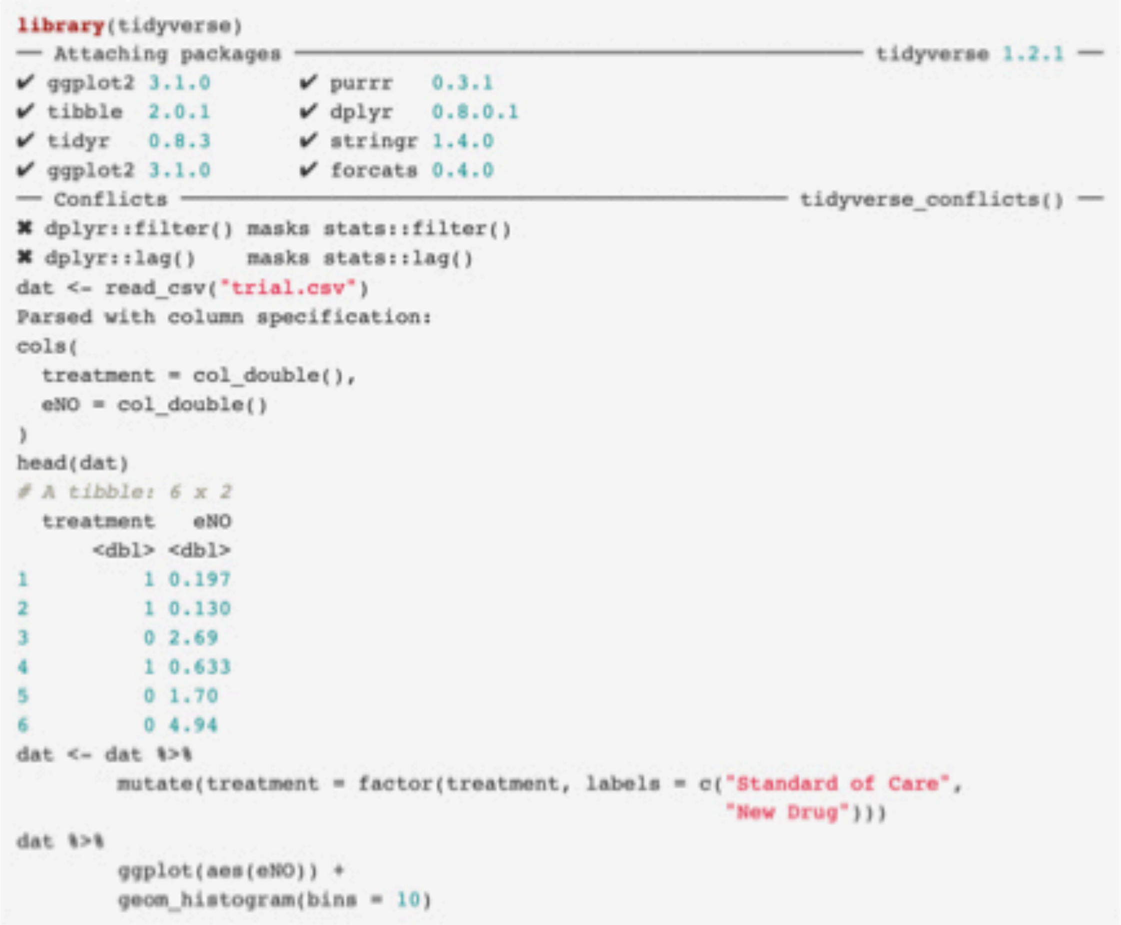
Results

Audience

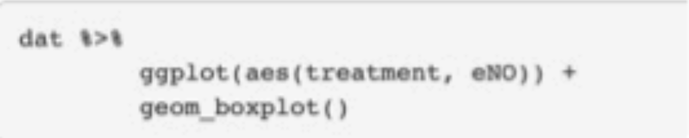
Analyst

Analytic container, analytic product, analytic presentation

Code

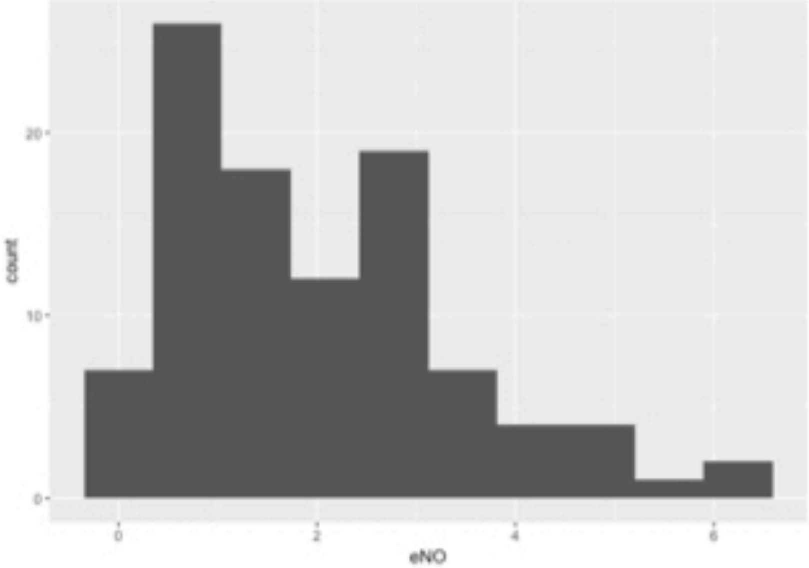


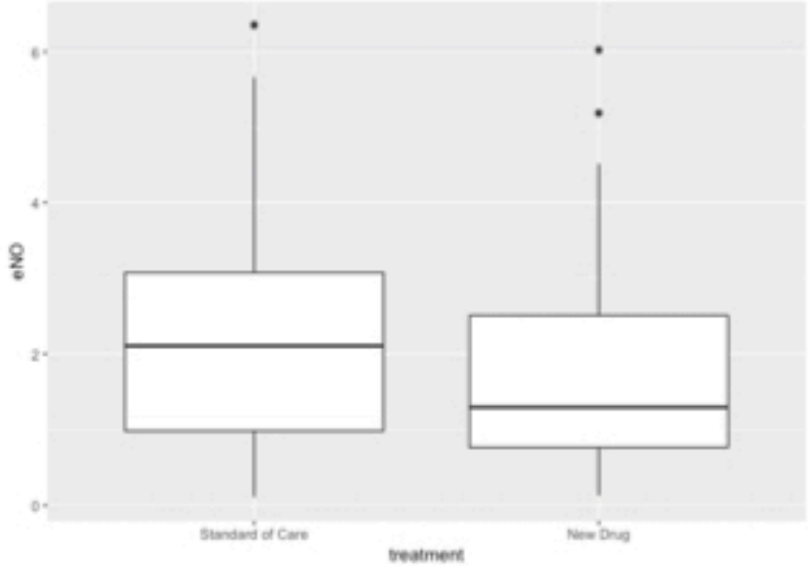
```
library(tidyverse)
── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
✔ ggplot2 3.1.0     ✔ purrr   0.3.1
✔ tibble  2.0.1     ✔ dplyr   0.8.0.1
✔ tidyr   0.8.3     ✔ stringr 1.4.0
✔ ggplot2 3.1.0     ✔ forcats 0.4.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
dat <- read_csv("trial.csv")
Parsed with column specification:
cols(
  treatment = col_double(),
  eNO = col_double()
)
head(dat)
# A tibble: 6 x 2
  treatment   eNO
      <dbl> <dbl>
1         1 0.197
2         1 0.130
3         0 2.69
4         1 0.633
5         0 1.70
6         0 4.94
dat <- dat %>%
        mutate(treatment = factor(treatment, labels = c("Standard of Care",
                                                        "New Drug")))
dat %>%
        ggplot(aes(eNO)) +
        geom_histogram(bins = 10)
```



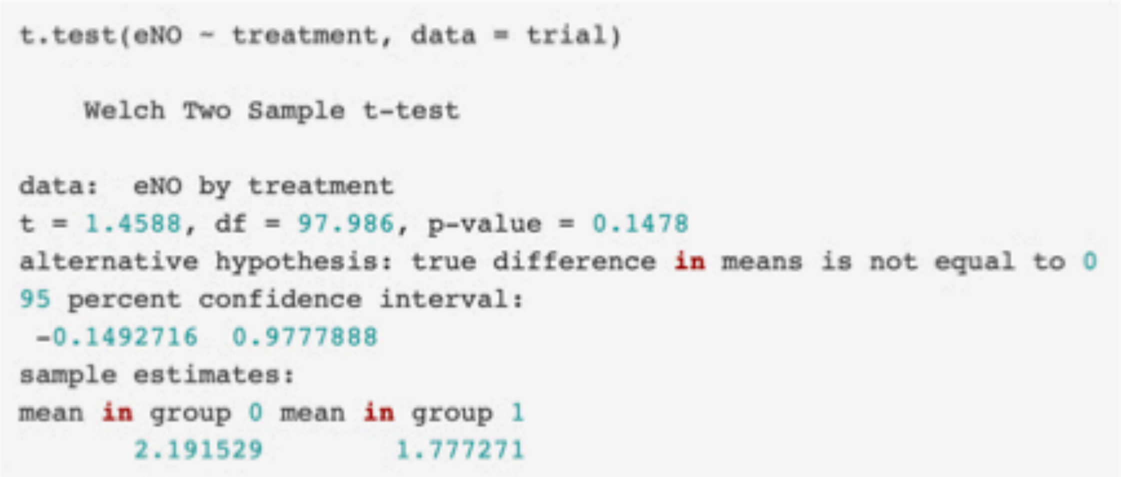
```
dat %>%
        ggplot(aes(treatment, eNO)) +
        geom_boxplot()
```

Executed Code





Model output

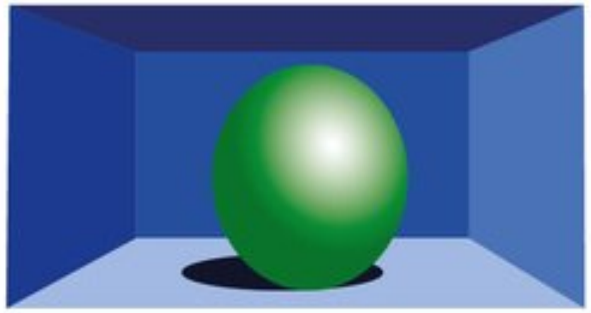


```
t.test(eNO ~ treatment, data = trial)

    Welch Two Sample t-test

data:  eNO by treatment
t = 1.4588, df = 97.986, p-value = 0.1478
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1492716  0.9777888
sample estimates:
mean in group 0 mean in group 1
       2.191529        1.777271
```

Presentation

ELEMENTS OF ART

The elements of art are the building blocks used by artists to create a work of art.

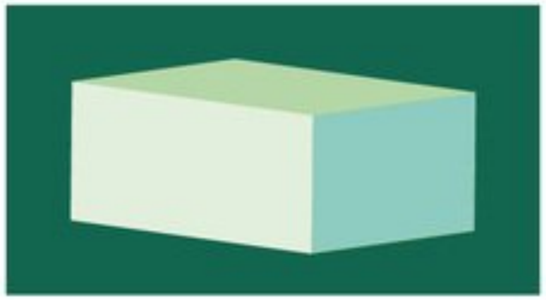


SPACE

Space is the area between and around objects. The space around objects is often called negative space; negative space has shape. Space can also refer to the feeling of depth. Real space is three dimensional; in visual art, when we create the feeling or illusion of depth, we call it space.

FORM

Forms are three-dimensional shapes expressing length, width, and depth. Balls, cylinders, boxes, and pyramids are forms.



SHAPE

Shape is a closed line. Shapes can be geometric, like squares and circles; or organic, like free-form or natural shapes. Shapes are flat and can express length and width.

LINE

A line is a mark with greater length than width. Lines can be horizontal, vertical, or diagonal; straight or curved; thick or thin.

Texture is the surface quality that can be seen and felt. Textures can be rough or smooth, soft or hard. Textures do not always feel the way they look; for example, a drawing of a porcupine may look prickly, but if you touch the drawing, the paper is still smooth.

TEXTURE

COLOR

Color is light reflected off of objects. Color has three main characteristics: ***hue*** (the main property of color, what differentiates colors), ***value*** (how light or dark it is), and ***intensity*** (how bright or dull it is).

- White is pure light; black is the absence of light.

- ***Primary colors*** are the only true colors (red, blue, and yellow). All other colors are mixes of primary colors.

- ***Secondary colors*** are two primary colors mixed together (green, orange, violet).

- ***Complementary colors*** are located directly accross from each on the color wheel. Complementary pairs contrast because they share no common colors. For example, red and green are complements, because green is made of blue and yellow. When complementary colors are mixed together, they neutralize each other to make brown.

**Value can also be categorized as an element of art.*

What are the *elements* of a data analysis (= fundamental components of a data analysis used by the data analyst)?

What are the *elements* of a data analysis (= fundamental components of a data analysis used by the data analyst)?

- Code
- Code comments (human readable; symbol)
- Narrative text (human readable)
- Data visualization (e.g. plot)
- Narrative diagram (e.g. flowchart; not necessarily with data)
- Summary statistics
- Table (summaries, but ordered int a row/column format)
- Statistical model or computational algorithm

What is a Successful Data Analysis?

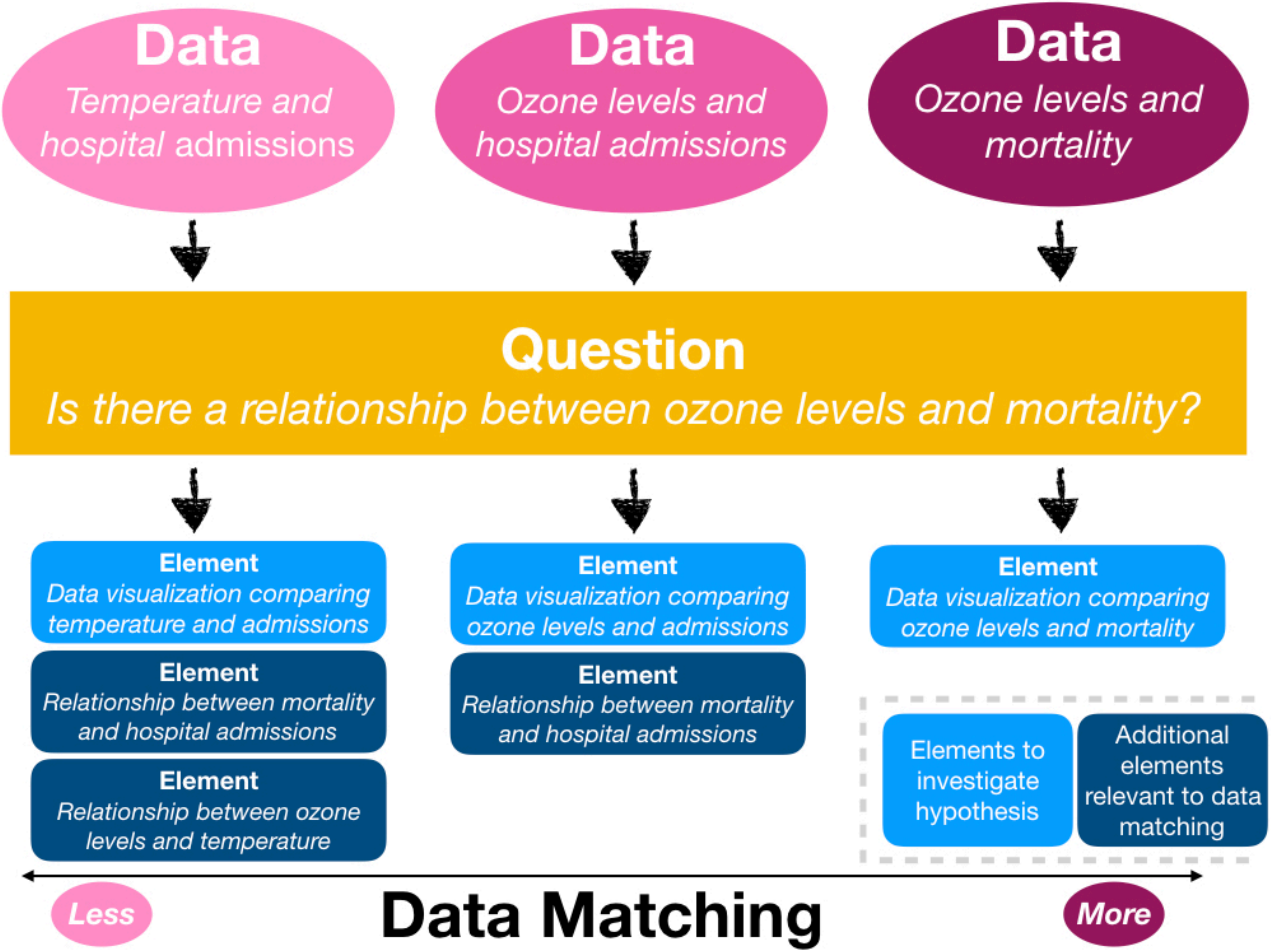
- Data analyses must be created / designed
- Design thinking
  - Identify the problem —> Exploratory data analysis
  - Build the solution —> Modeling, uncertainty, narrative
- Audience has wants and needs

What is a Successful 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
- A successful analysis is one in which the audience **accepts** what the analyst as done and **agrees on the weighting of principles** chosen by the analyst

Data Analysis Principles

- Reflect qualities of a data analysis
- Serve to guide the development of an analysis
- Observable from the output of the data analysis
- Can be roughly measured / weighted (low, medium, high)
- Should be objectively measured



Data
Temperature and hospital admissions

Data
Ozone levels and hospital admissions

Data
Ozone levels and mortality

Question
Is there a relationship between ozone levels and mortality?

Element
Data visualization comparing temperature and admissions

Element
Relationship between mortality and hospital admissions

Element
Relationship between ozone levels and temperature

Element
Data visualization comparing ozone levels and admissions

Element
Relationship between mortality and hospital admissions

Element
Data visualization comparing ozone levels and mortality

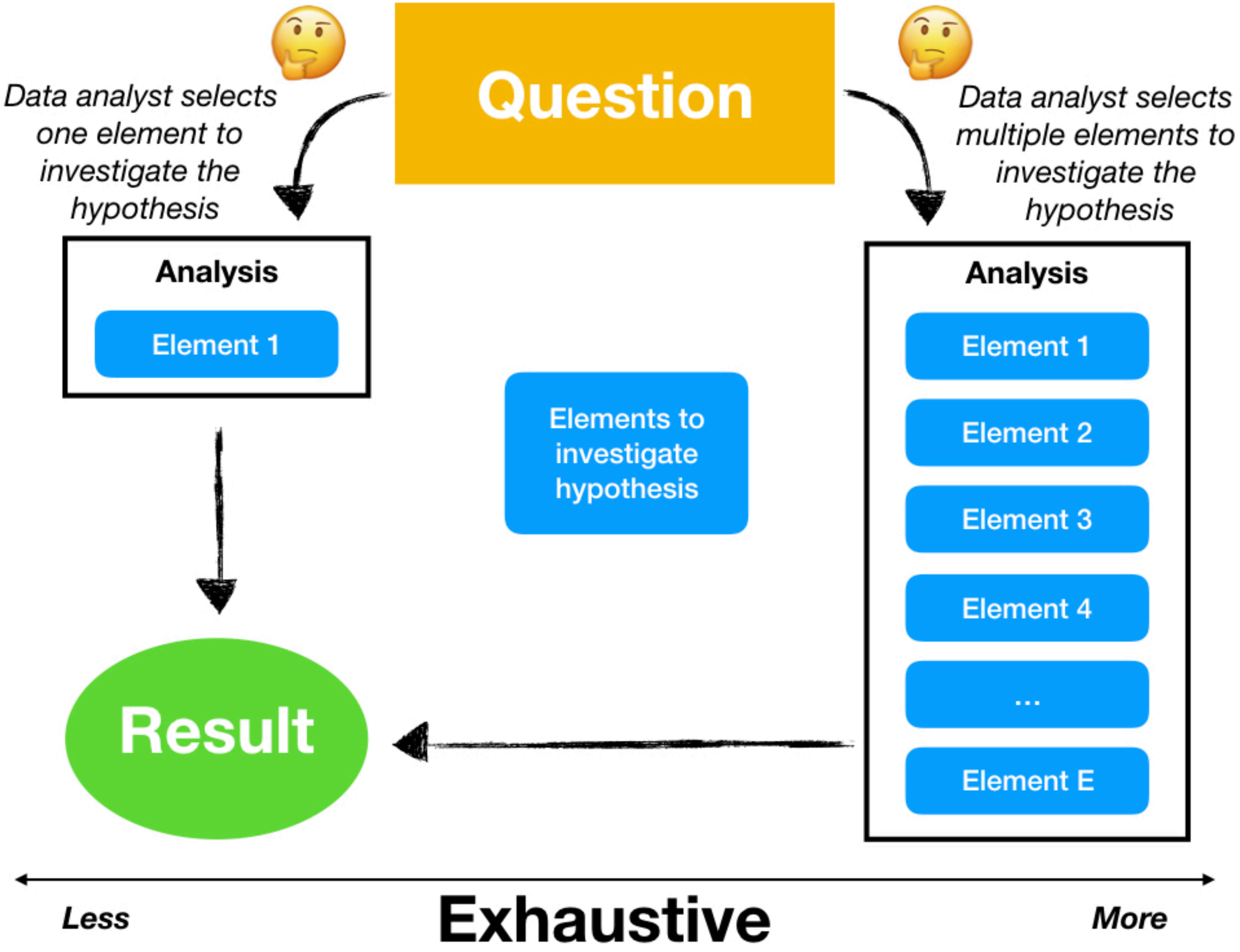
Elements to investigate hypothesis

Additional elements relevant to data matching

Less

Data Matching

More



Question

Data analyst selects one element to investigate the hypothesis

Data analyst selects multiple elements to investigate the hypothesis

Analysis

Element 1

Analysis

Element 1

Element 2

Element 3

Element 4

...

Element E

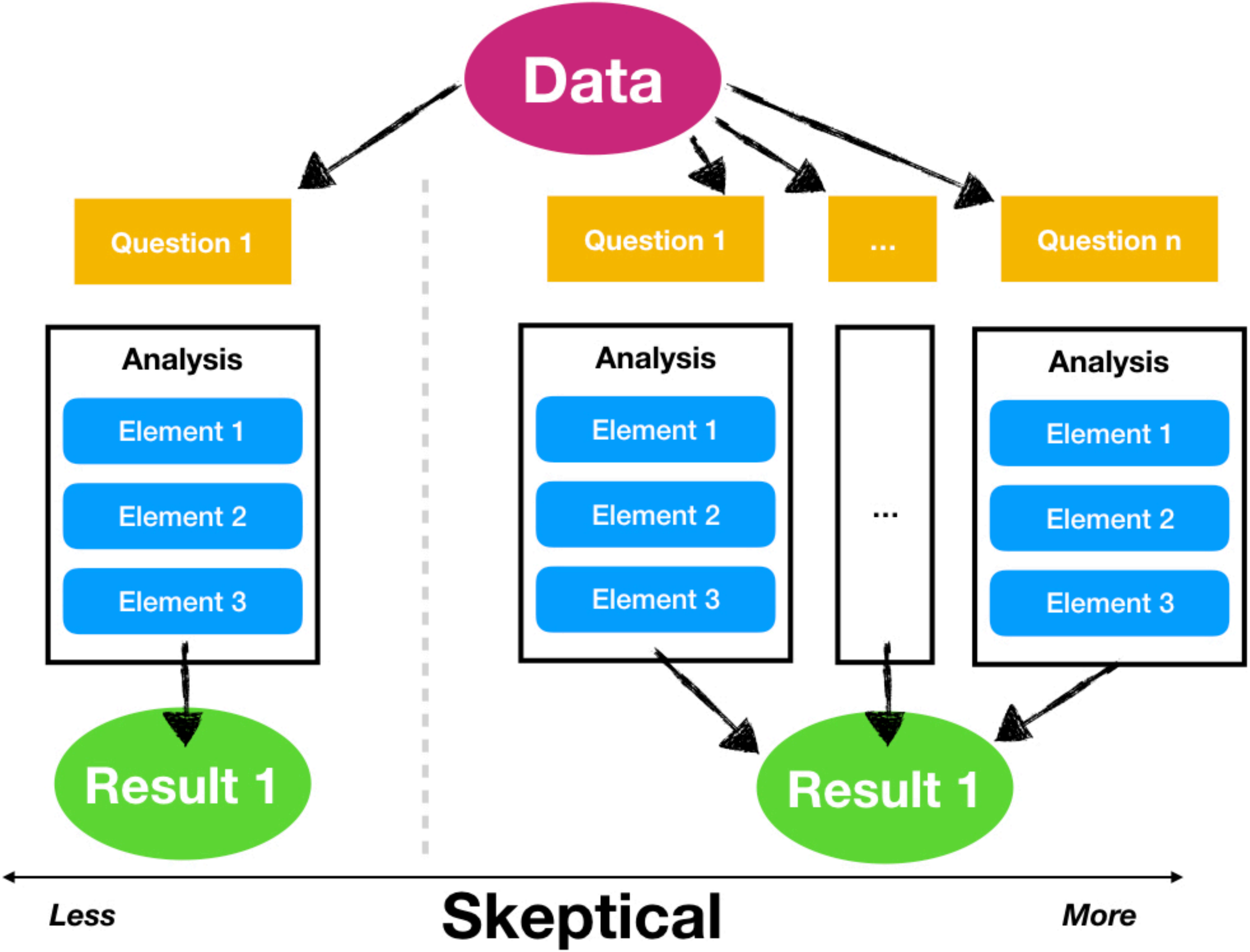
Elements to investigate hypothesis

Result

Less

Exhaustive

More



Data

Question 1

Analysis

Element 1

Element 2

Element 3

Result 1

Question 1

...

Question n

Analysis

Element 1

Element 2

Element 3

...

Analysis

Element 1

Element 2

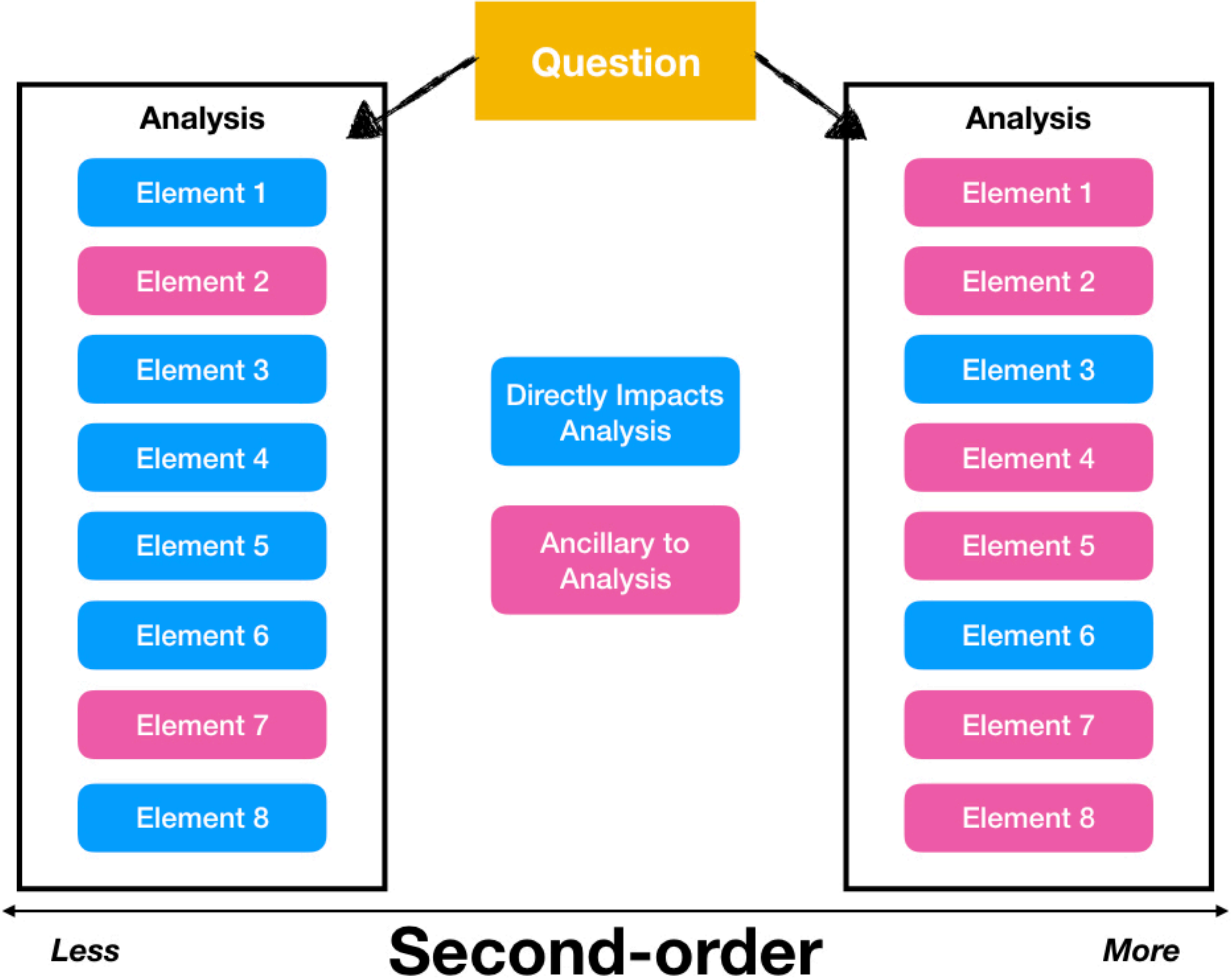
Element 3

Result 1

Less

Skeptical

More



Question

Analysis

Element 1

Element 2

Element 3

Element 4

Element 5

Element 6

Element 7

Element 8

Directly Impacts Analysis

Ancillary to Analysis

Analysis

Element 1

Element 2

Element 3

Element 4

Element 5

Element 6

Element 7

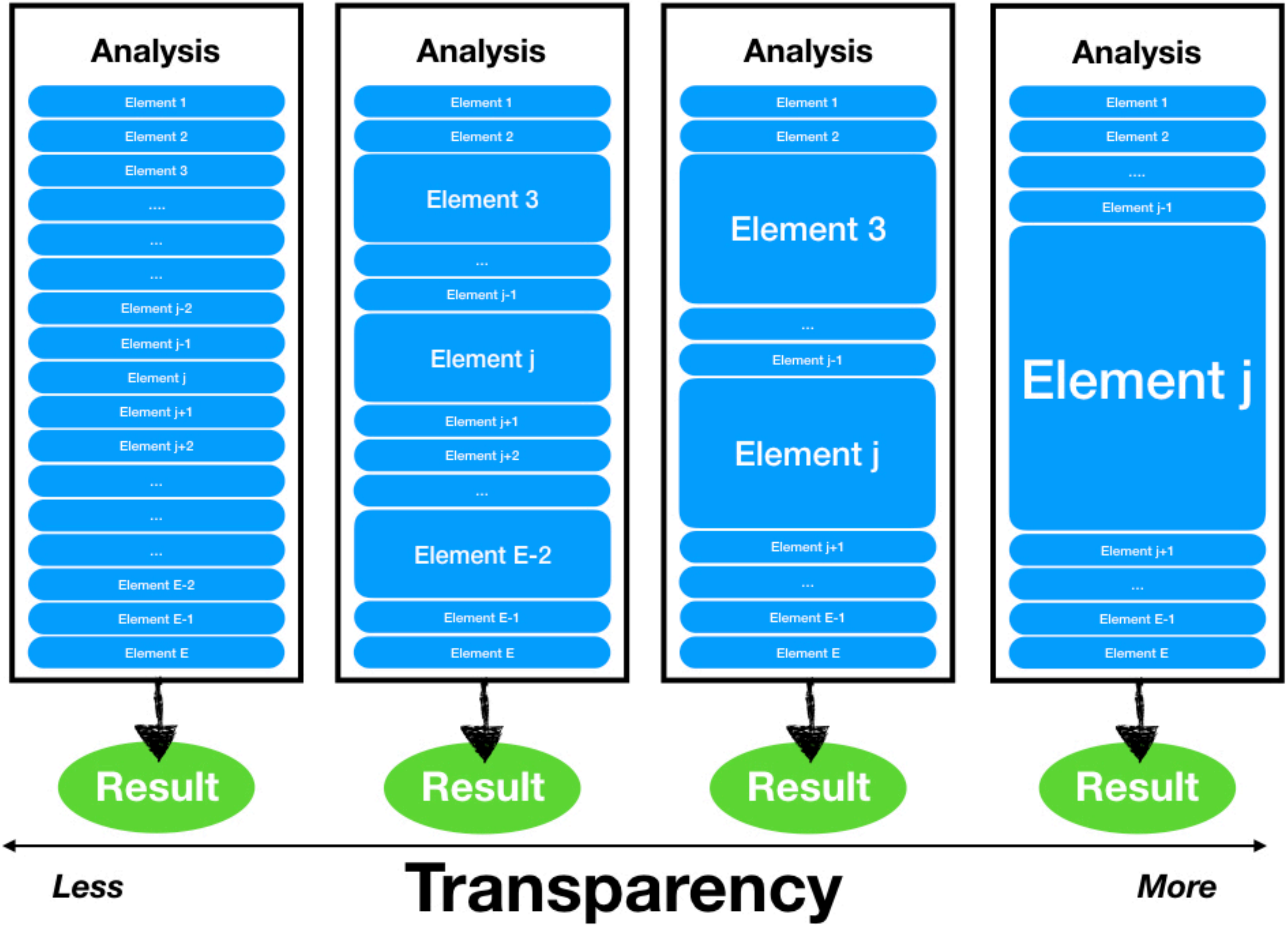
Element 8

Less

Second-order

More

Height of elements represent the amount of variation explained by element(s) that are influential and insightful in connecting evidence from data to key results or conclusions



Analysis

Element 1
Element 2
Element 3
....
...
...
Element j-2
Element j-1
Element j
Element j+1
Element j+2
...
...
...
Element E-2
Element E-1
Element E

Result

Analysis

Element 1
Element 2
Element 3
...
Element j-1
Element j
Element j+1
Element j+2
...
Element E-2
Element E-1
Element E

Result

Analysis

Element 1
Element 2
Element 3
...
Element j-1
Element j
Element j+1
...
Element E-1
Element E

Result

Analysis

Element 1
Element 2
....
Element j-1
Element j
Element j+1
...
Element E-1
Element E

Result

Less **Transparency** *More*

Can a neural network learn to recognize doodling?

We took two approaches, a statistical approach (see here and here) and a machine learning approach (see here) to build two algorithms to classify food images. Yum!

Figure 1: A Doodle from each Food Class

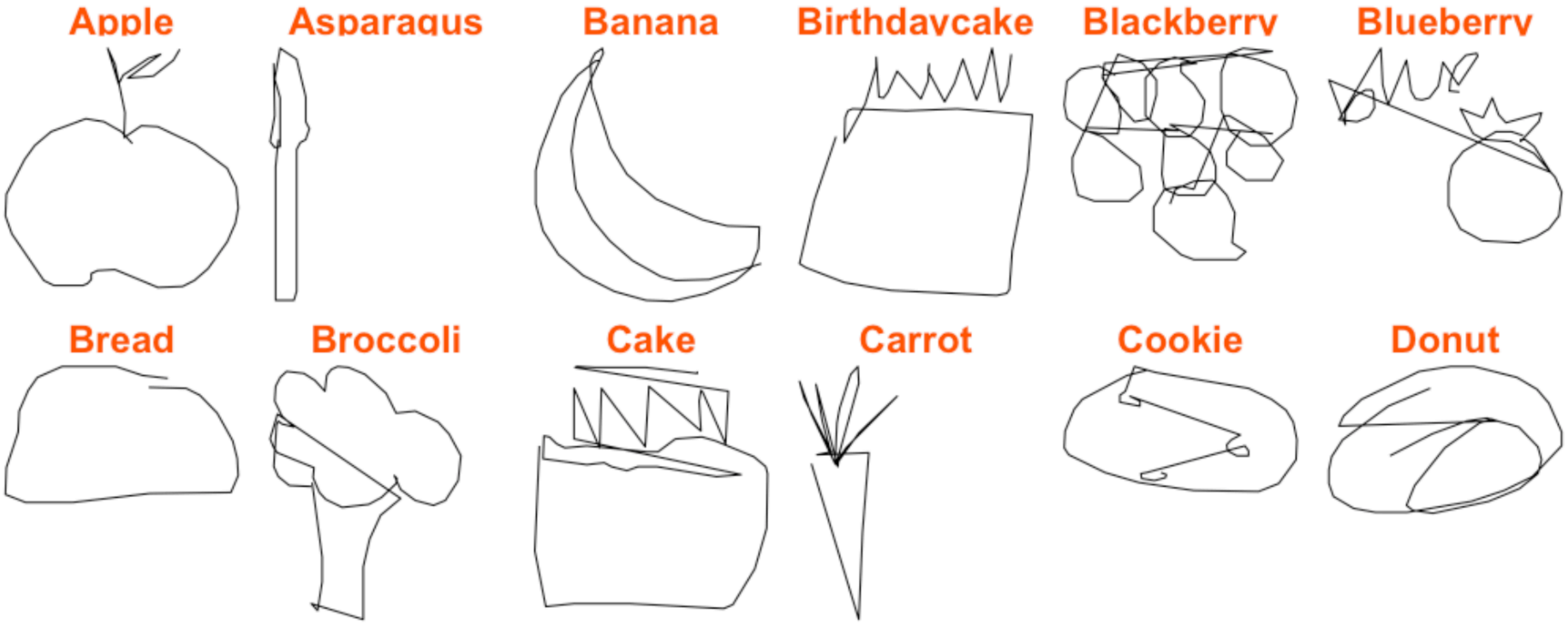
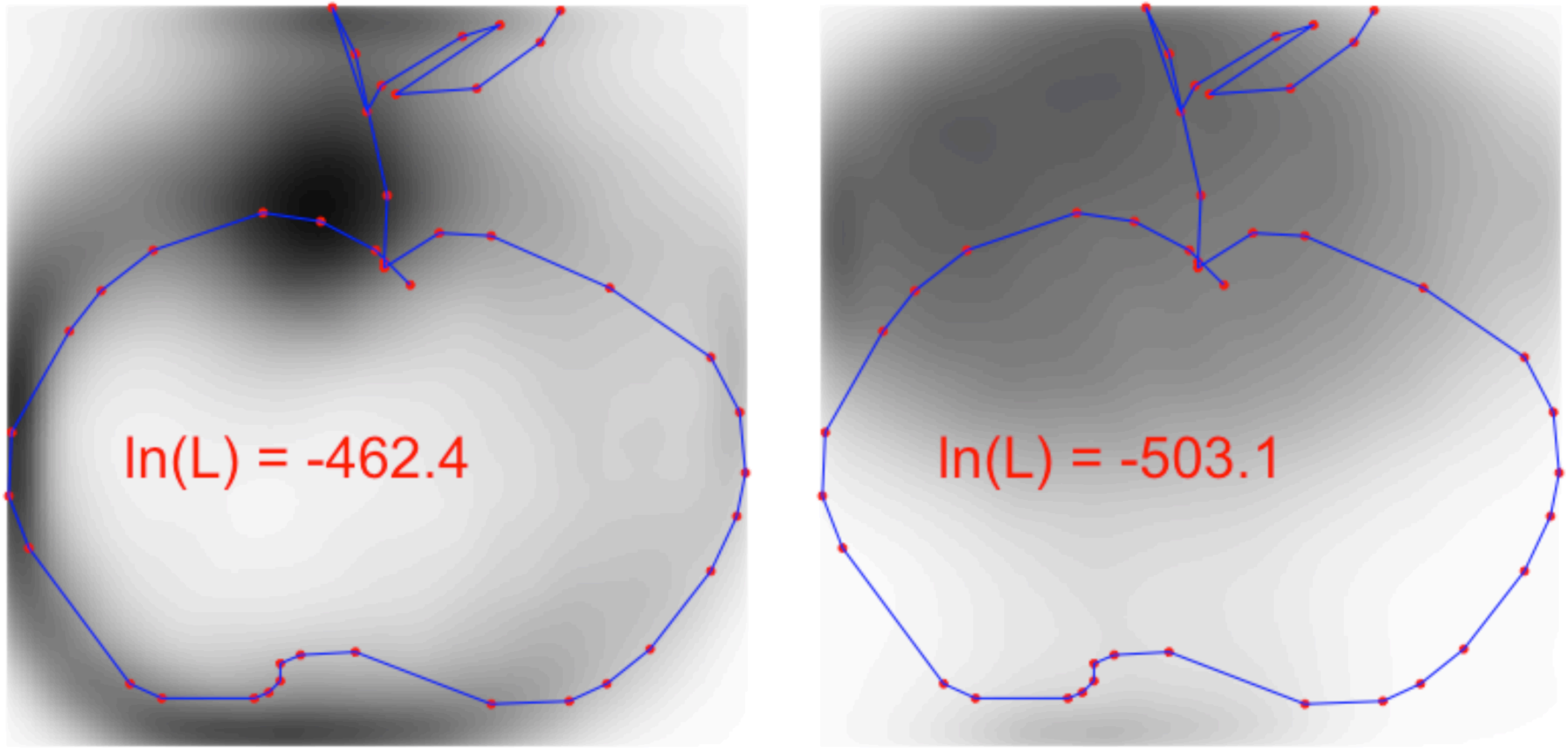
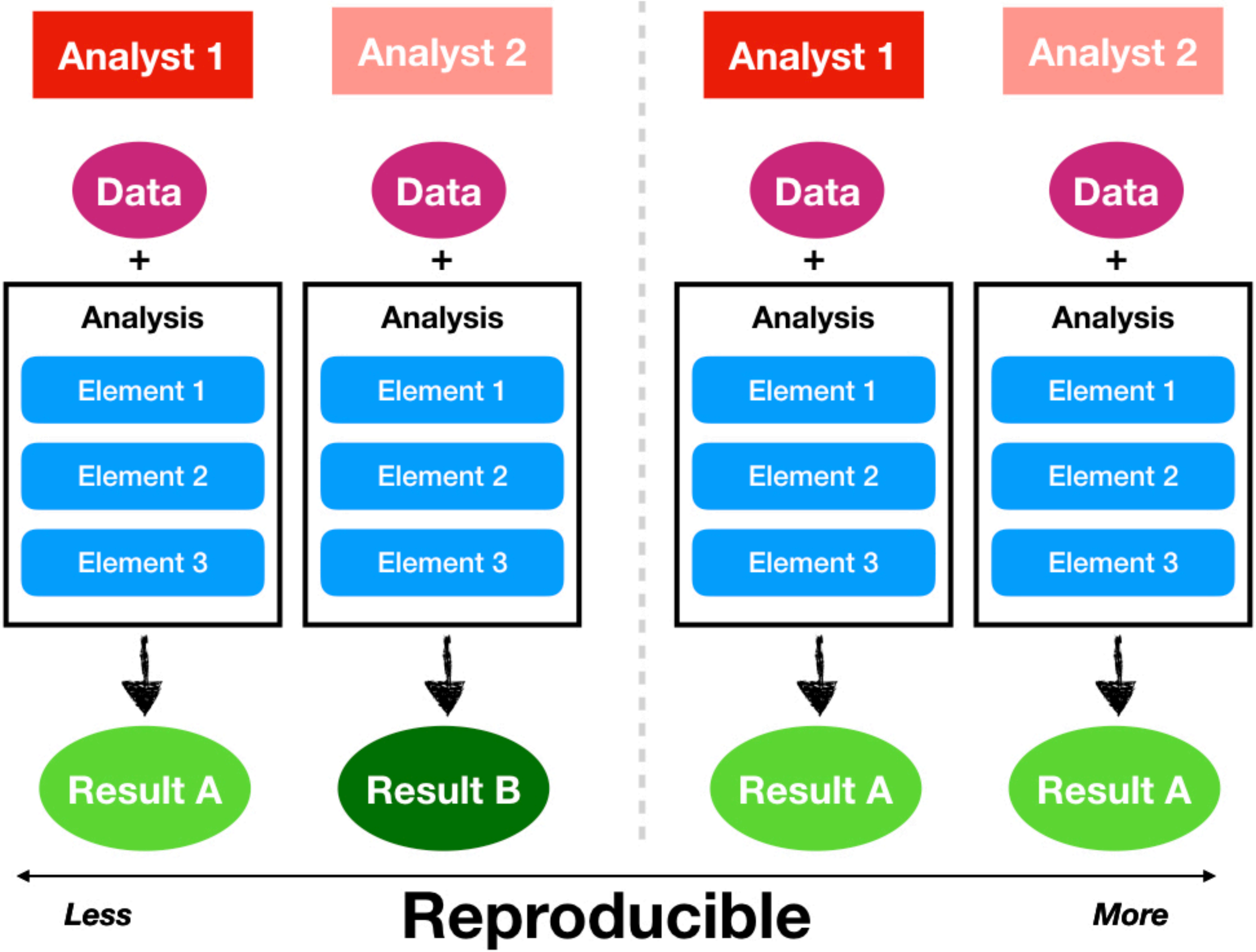


Figure 1: Apple Drawing Overlayed on the Apple and Broccoli Density Estimates





Analyst 1

Analyst 2

Analyst 1

Analyst 2

Data

Data

Data

Data

+

+

+

+

Analysis

Element 1

Element 2

Element 3

Analysis

Element 1

Element 2

Element 3

Analysis

Element 1

Element 2

Element 3

Analysis

Element 1

Element 2

Element 3

Result A

Result B

Result A

Result A

Less

Reproducible

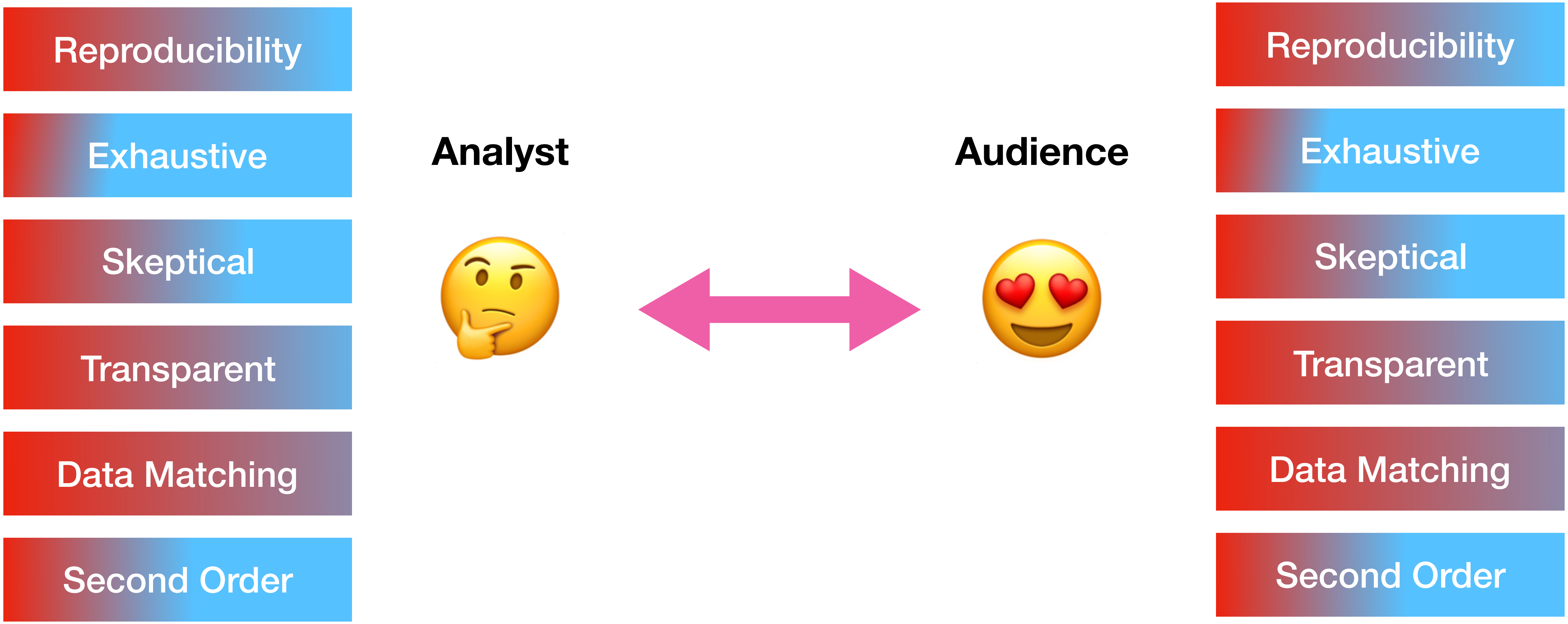
More

What can the elements and principles of data analysis be used for?

One idea: How can you evaluate the quality of an analysis?

Success? Validity? Honesty?

What is a Successful Data Analysis?



Reproducibility

Exhaustive

Skeptical

Transparent

Data Matching

Second Order

Analyst

Audience

Reproducibility

Exhaustive

Skeptical

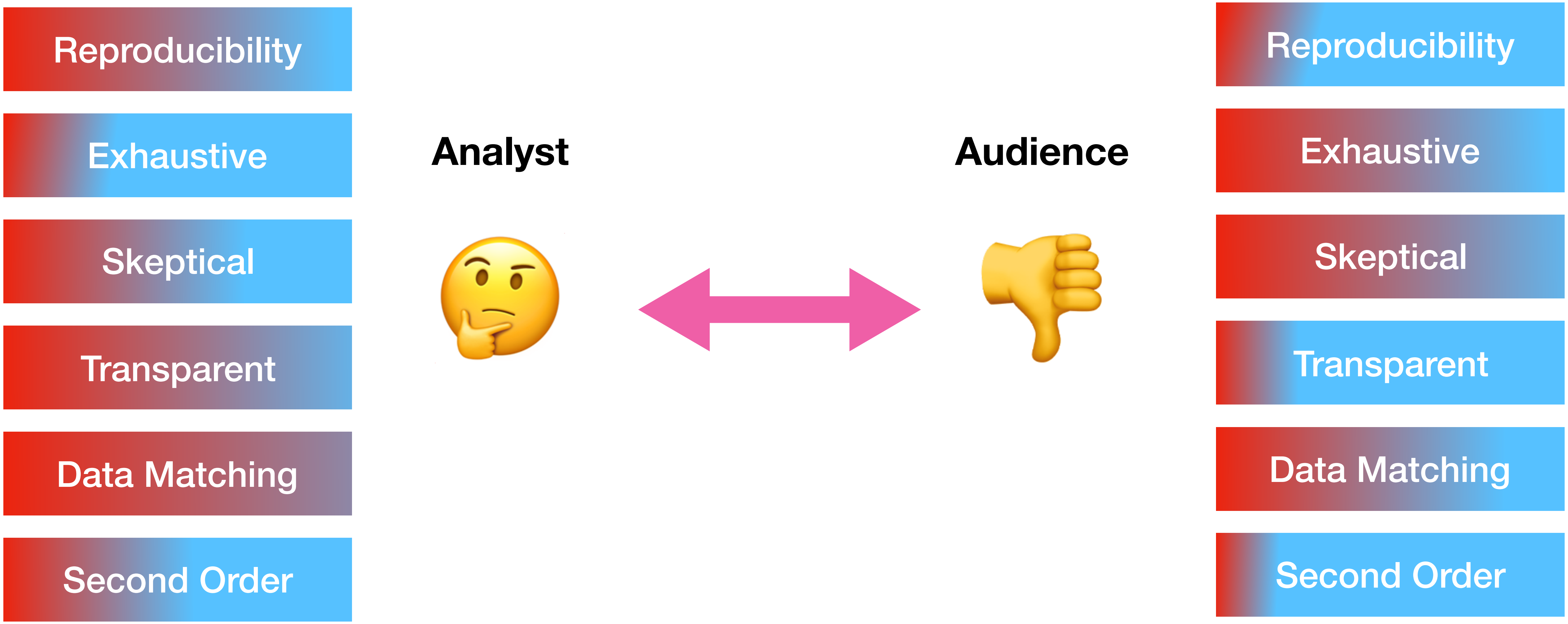
Transparent

Data Matching

Second Order

- Inclusion or exclusion of principles ≠ judgment or assessment of quality
- A data analyst assigns weights to principles to ↑ or ↓ objective characteristics (principles)
- Characteristics can be highly influenced by outside constraints or resources (e.g. time or budget)
- → different weighting of the principles can lead to different data analyses (all addressing the same question)

What is a Successful Data Analysis?



Reproducibility

Exhaustive

Skeptical

Transparent

Data Matching

Second Order

Analyst

Audience

Reproducibility

Exhaustive

Skeptical

Transparent

Data Matching

Second Order

- Inclusion or exclusion of principles ≠ judgment or assessment of quality
- A data analyst assigns weights to principles to ↑ or ↓ objective characteristics (principles)
- Characteristics can be highly influenced by outside constraints or resources (e.g. time or budget)
- → different weighting of the principles can lead to different data analyses (all addressing the same question)

Consider

- analyst *i*
- specific weight assign to principle *k* is $W_i^{(k)}$

Total weight assigned to the analysis by analyst

$$N_i = \sum_{k=1}^{K} W_i^{(k)}$$

Can model the individual principle-specific weights ($W_i^{(k)}$) with the multinomial distribution

$$\mathbf{W}_i = \left(W_i^{(1)}, \ldots, W_i^{(K)}\right) \sim \textbf{Multinomial}\left(N_i; \pi_i^{(1)}, \ldots, \pi_i^{(K)}\right).$$

- $\pi_i^{(k)}$ = probability of analyst *i* assigning weight to principle *k*
- probabilities must sum to 1 across the *K* principles, i.e. $\sum_{k=1}^{K} \pi_i^{(k)} = 1$
    - Reflects the reality that all analysts must decide how to allocate their priorities towards each principle when building a data analysis

For a given principle *k*, we can derive the marginal distribution from the multinomial

$$W_i^{(k)} \sim \textbf{Binomial}(N_i; \pi_i^{(k)})$$

We can then model the $\pi_i^{(k)}$s as

$$\psi_i^{(k)} = \log\left(\frac{\pi_i^{(k)}}{1-\pi_i^{(k)}}\right) = \lambda_{f_i}^{(k)} + \delta_i^{(k)} + \mathbf{x}_i'\beta_i^{(k)}$$

- $\lambda_{f_i}^{(k)}$ is the field-specific mean for principle *k* and analyst *i* in field f_i
- $\delta_i^{(k)}$ is analyst *i*'s deviation from the field-specific mean for principle *k*
- $\mathbf{x}_i$ is a vector of analysis-specific resources and characteristics for the analysis
  - i.e. time, budget, personnel, significance
- $\beta_i^{(k)}$ is a vector of coefficients that indicate how each resource is related to the up-weighting or down-weighting of the k^{th} principle for the analysis
- We consider the analyst deviation $\delta_i^{(k)}$ to be randomly distributed across the set of potential analysts with mean 0 and finite variance

Analogous to the analyst's weights, the weight given to principle *k* by audience member *j* (who is a member of field f_j) can be written as $A_j^{(k)}$ with $N_j = \sum_{k=1}^{K} A_j^{(k)}$ being the total weight given to the analysis.

We similarly model the vector $\mathbf{A}_j = \left(A_j^{(1)}, \ldots, A_j^{(K)}\right)$ as multinomial with total N_j and proportions $\omega_j^{(1)}, \ldots, \omega_j^{(K)}$.

We then similarly model the proportions $\omega_j^{(k)}$ as

$$\alpha_j^{(k)} = \log\left(\frac{\omega_j^{(k)}}{1 - \omega_j^{(k)}}\right) = \lambda_{f_j}^{(k)} + \eta_j^{(k)} + \mathbf{z}_j'\gamma_j^{(k)}$$

- $\lambda_{f_j}^{(k)}$ and $\eta_j^{(k)}$ are the field-specific mean and individual-level deviation for the j^{th} audience member, respectively
- $\mathbf{z}_j$ is the audience's perception of resources available and question significance
- $\gamma_j^{(k)}$ is the audience member's sense of the relationship between a given resource and the weight that should be given to the principle
    - Note that we consider $\eta_j^{(k)}$ to be independent of $\delta_i^{(k)}$ in the analyst's weight model

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

$$D_{ij}^{(k)} = \psi_i^{(k)} - \alpha_j^{(k)} = \left(\lambda_{f_i}^{(k)} - \lambda_{f_j}^{(k)}\right) + \left(\delta_i^{(k)} - \eta_j^{(k)}\right) + \left(\mathbf{x}_i'\beta_i^{(k)} - \mathbf{z}_j'\gamma_j^{(k)}\right)$$

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 a Successful Data Analysis

Strong Pairwise Success

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

Because of the randomness in $\delta_i^{(k)}$ and $\eta_j^{(k)}$, the $D_{ij}^{(k)}$ values can never be equal to zero.

However, the definition of *strong pairwise success* requires that the differences are never too large for any given principle.

Defining a Successful Data Analysis

Weak Pairwise Success

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

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.

From definition of strong pairwise success, if we assume $\delta_i^{(k)}$ and $\eta_j^{(k)}$ are random (with mean 0 and finite variance) and independent, then the principle-specific weight difference has expectation

$$\mathbb{E}\left[D_{ij}^{(k)}\right] = (\lambda_{f_i}^{(k)} - \lambda_{f_j}^{(k)}) + (\mathbf{x}_i'\beta_i^{(k)} - \mathbf{z}_j'\gamma_j^{(k)})$$

(and in general this will be different from 0)

However:

- Analyst *i* may only have general information about the audience member *j*, but may not know specifically who the audience will be
  - Here, the analyst may have info about the population parameters of the audience and want to measure success based on the mean values for the population

Alternative idea:

- Look at the difference in expected values for the weightings for all *K* principles and denote this the *potential* pairwise success of an analysis, because we have not yet observed the audience's principle weighting

Defining a Successful Data Analysis

Potential Pairwise Success

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

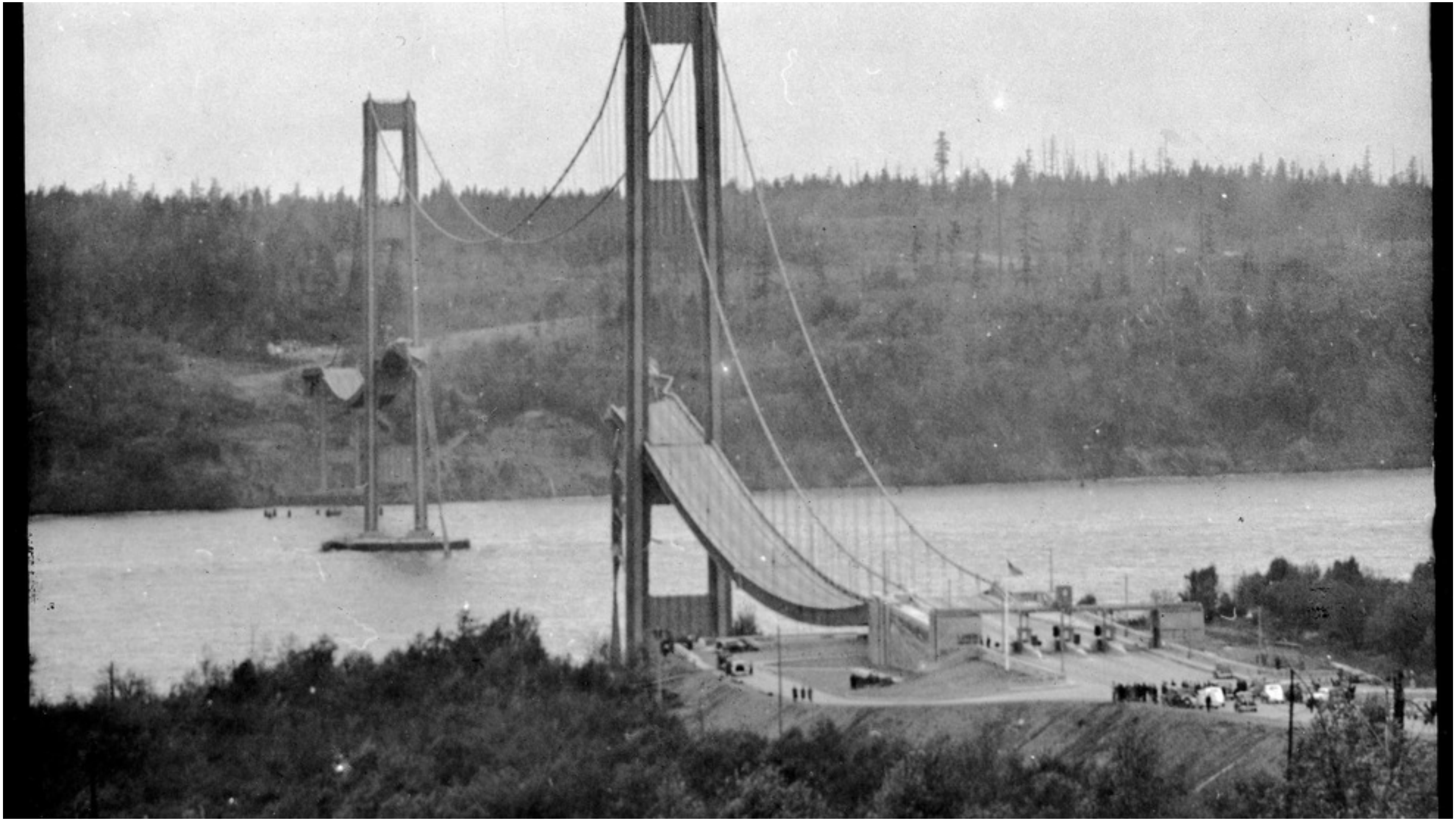
Key distinction between *strong* (or *weak*) pairwise success and *potential* pairwise success is:

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

Hence

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

Give People What They Want?



Designing 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 “fail”

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

How Can We Improve Data Analysis?

- Improvement/advancement of approaches to data analysis depends in part on learning from mistakes and failures
- When analyses fail (or succeed!) the details are often not public or are not recorded for later review
- When information about failures is available we often
  - Draw incorrect lessons because of incomplete information
  - Focus primarily on assigning blame (“user error”)
- Learning what works in data analysis is challenging, requiring either detective work or first hand knowledge.

Do We Always Learn the Right Lessons?

- Analyses of failures in data analysis typically focus on narrow proximate causes or vague high-level “environmental” causes
- Lessons should lead to relevant **interventions**
- For a useful post-mortem we need
  - Detailed *timeline* information about analyses
  - Open and honest discussion from participants

What Makes for a Good Data Analyst?

- The application of **design thinking** to data problems
- The creation and management of **workflows** for transforming and processing data
- The negotiation of **human relationships** to identify context, allocate resources, and characterize audiences for data analysis results
- The application of **statistical methods** to quantify evidence
- The transformation of data analytic information into coherent **narratives and stories**

Thank You!