Exploratory Data Analysis

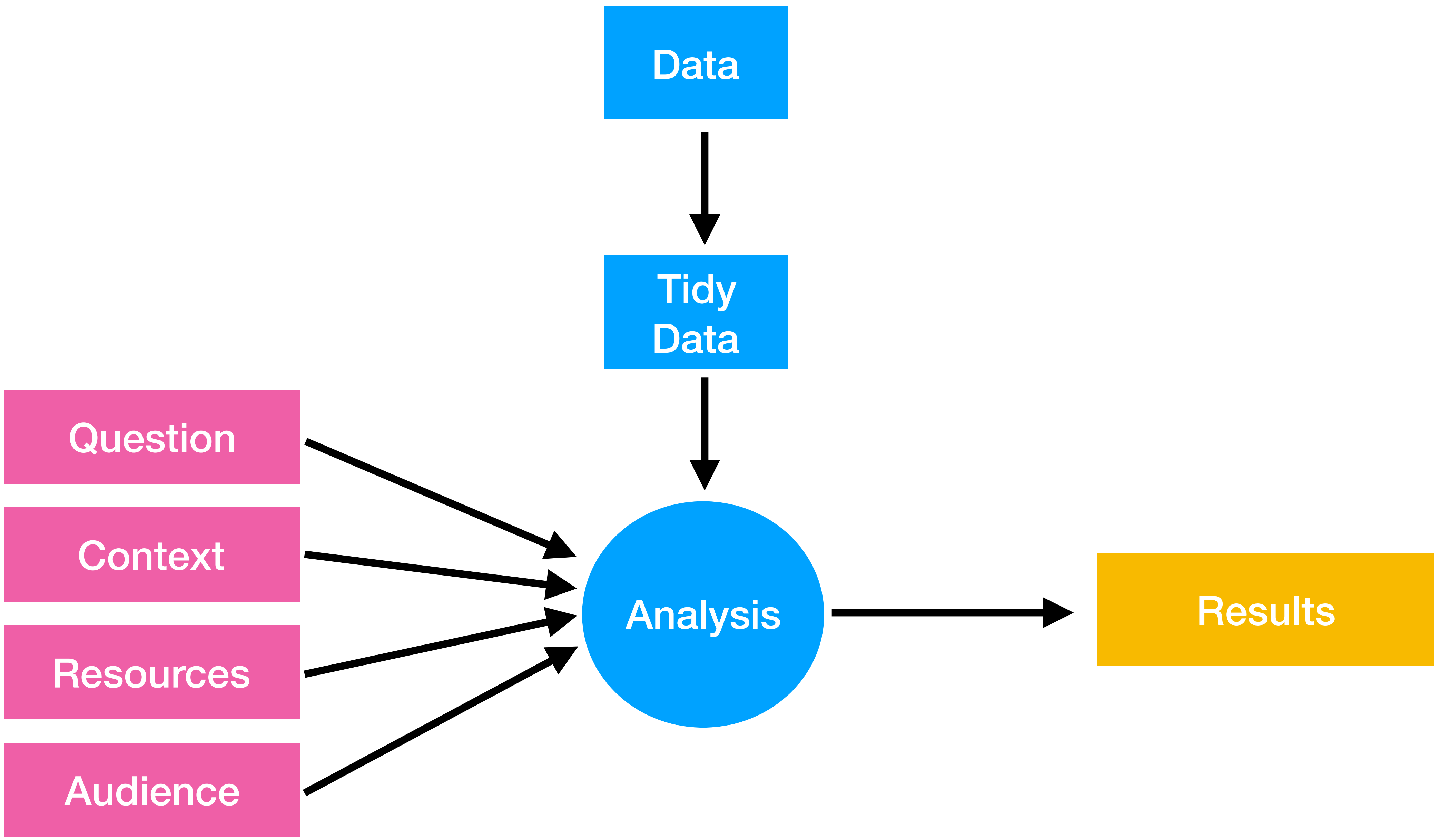
Roger D. Peng
Stephanie C. Hicks

Advanced Data Science
Term 1
2019

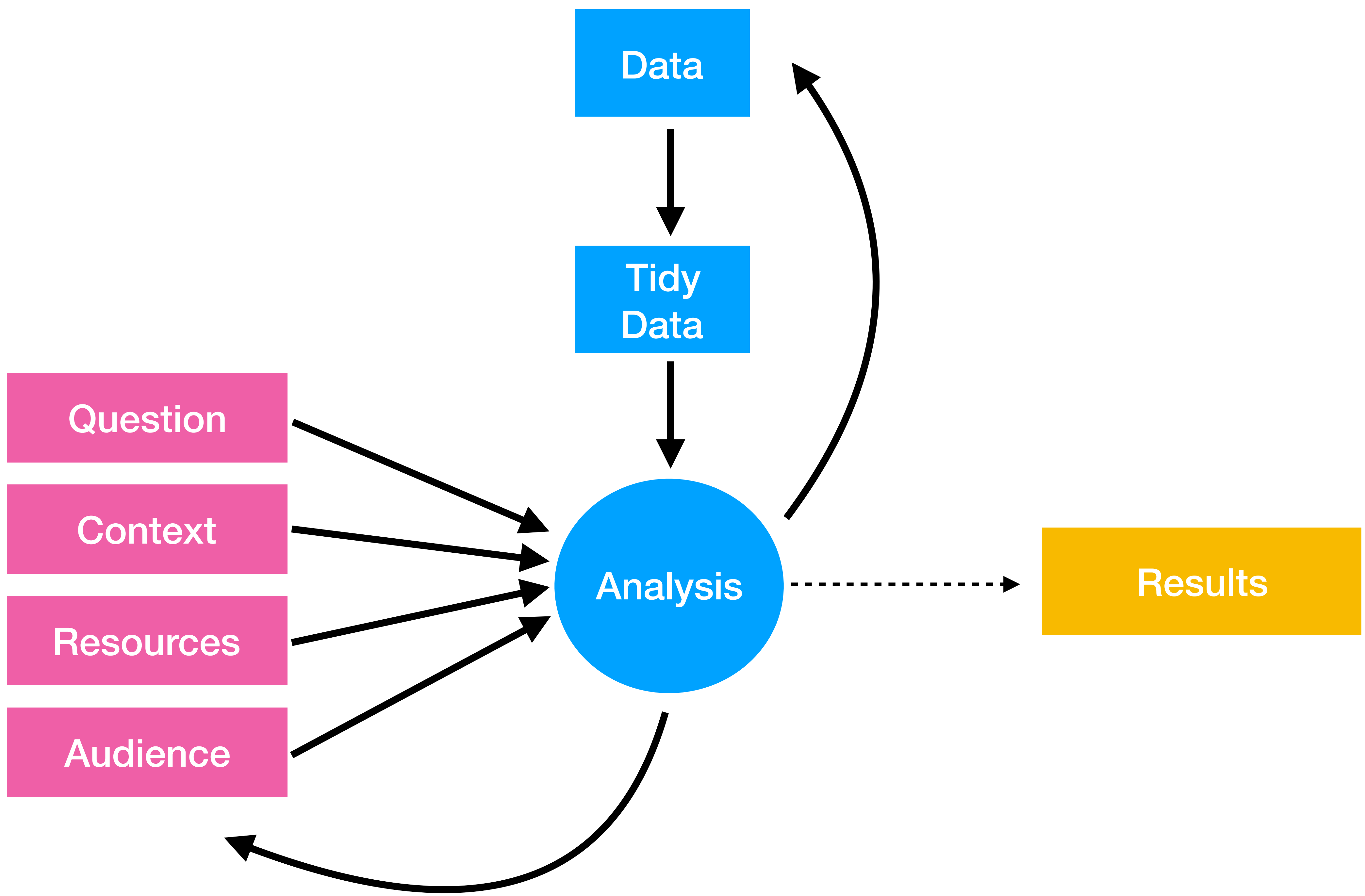
“Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.”

–John Tukey, "The Future of Data Analysis",
Annals of Mathematical Statistics*, 1962*

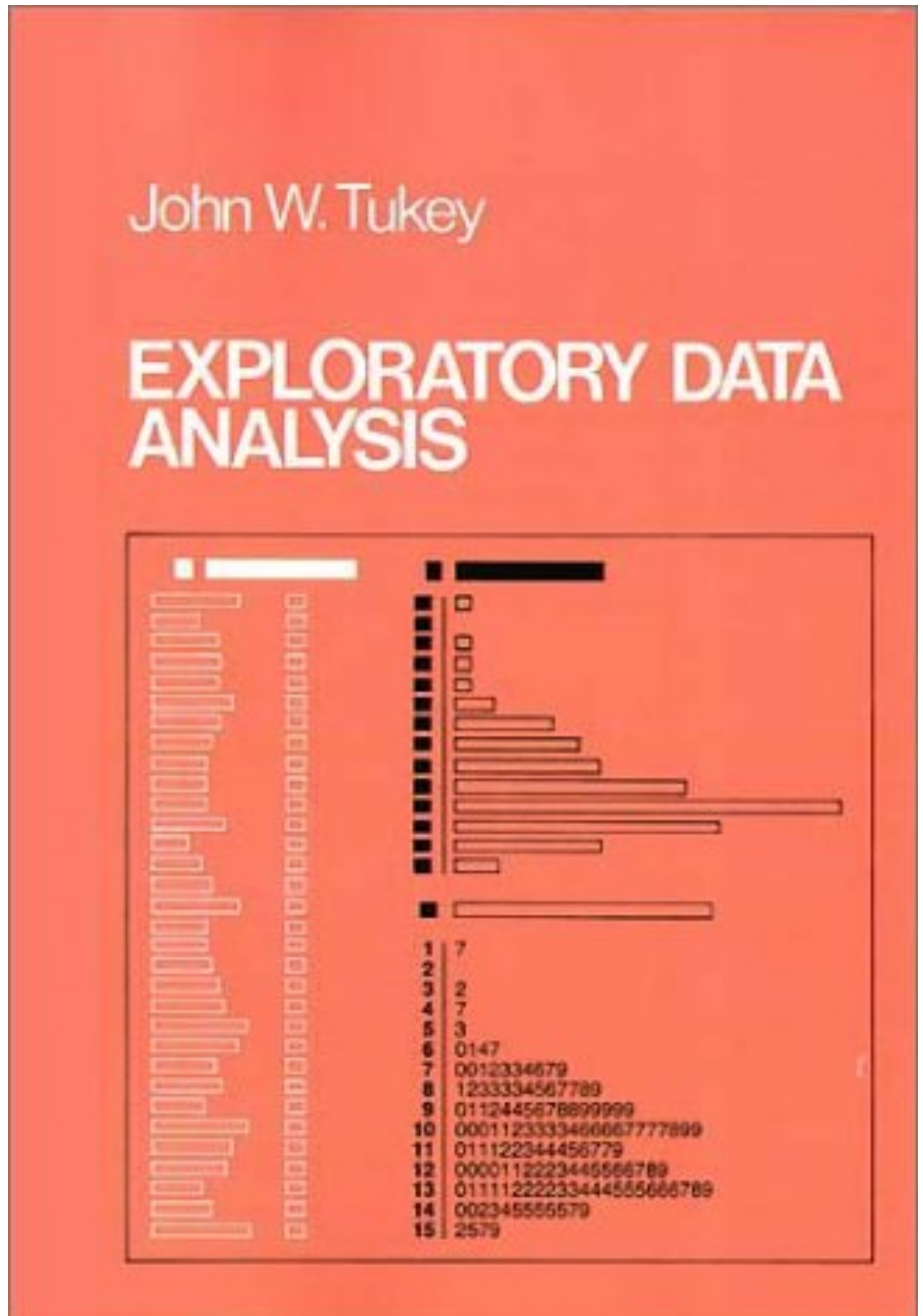
Data Analysis in a Nutshell



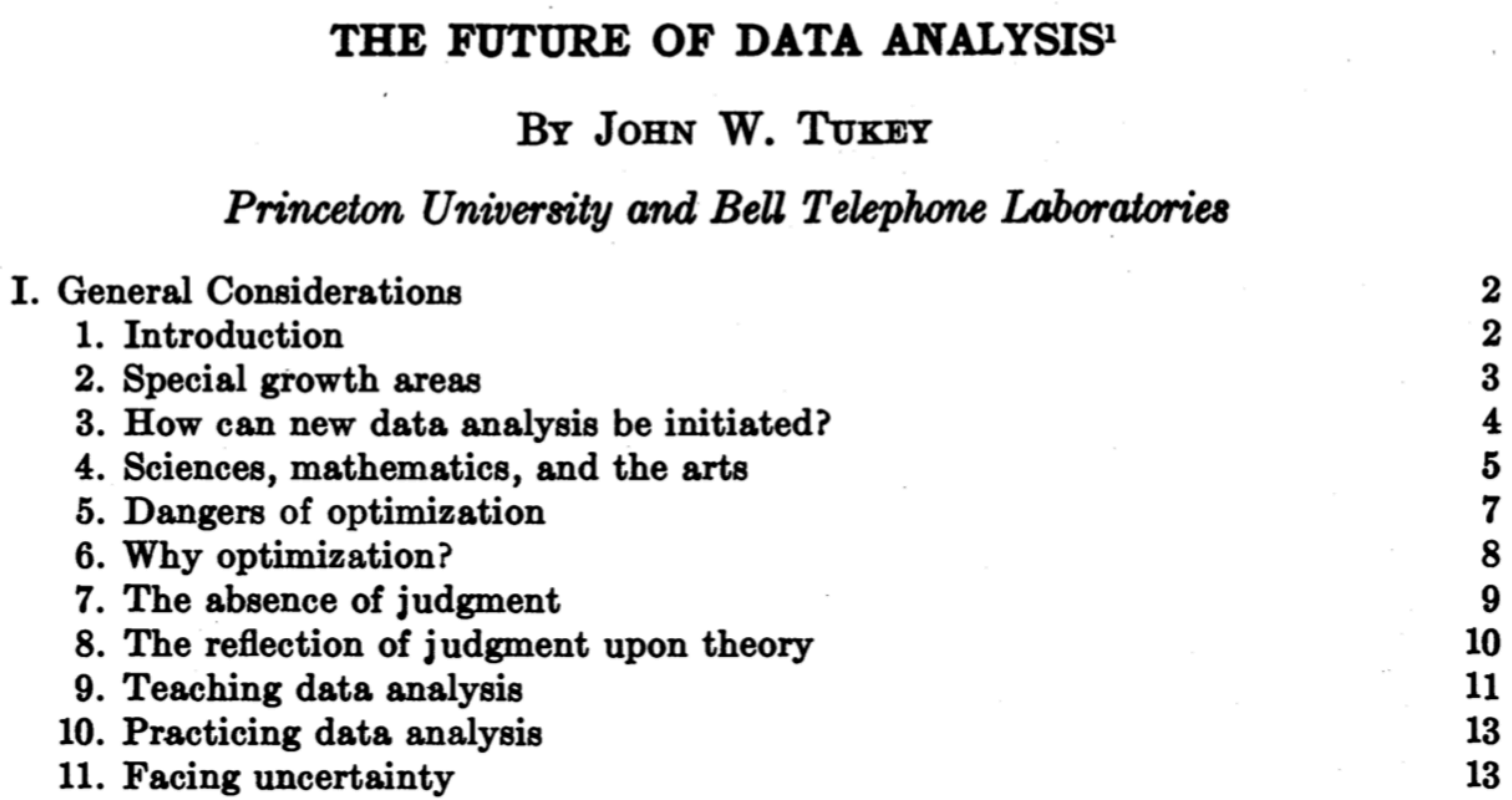
Data Analysis in a Nutshell



Tukey-sian Data Analysis



Tukey-sian Data Analysis


THE FUTURE OF DATA ANALYSIS[1]

BY JOHN W. TUKEY

Princeton University and Bell Telephone Laboratories

John Tukey, "The Future of Data Analysis", *Annals of Mathematical Statistics*, 1962

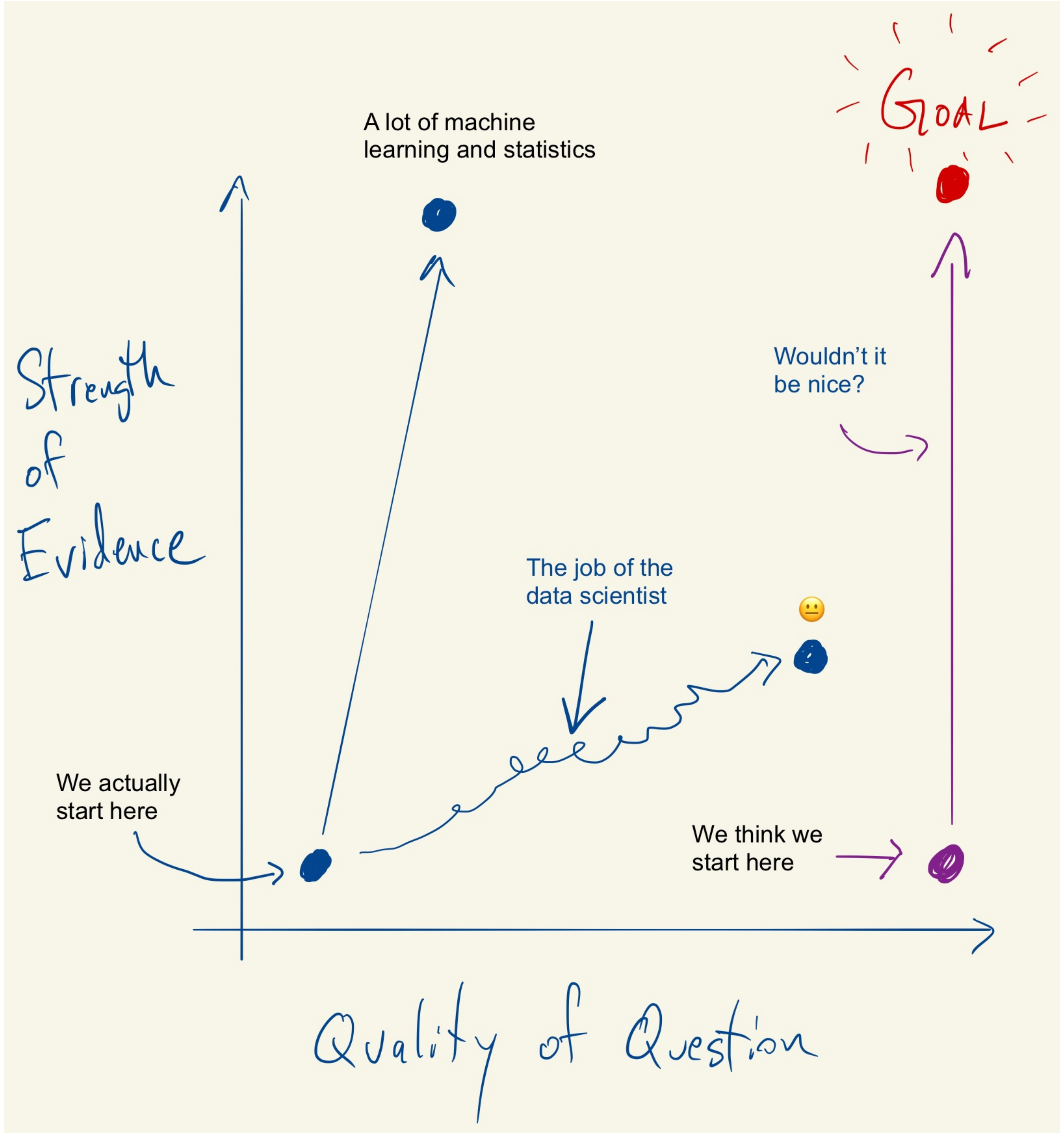
Tukey-sian Data Analysis

- Data analysis must seek for scope and usefulness rather than security
- Data analysis must be willing to err moderately often in order that inadequate evidence shall more often *suggest* the right answer
- Data analysis must use mathematical argument and mathematical results as bases for judgment rather than as bases for proof or stamps of validity
- "These points are meant to be taken seriously."

Tukey-sian Data Analysis

- (a1') Recognition of problem
- (a1'') One technique used
- (a2) Competing techniques used
- (a3) Rough comparisons of efficacy
- (a4) Comparison in terms of precise (and thereby inadequate) criterion
- (a5') Optimization in terms of a precise, and similarly inadequate criterion
- (a5'') Comparison in terms of several criteria

Tukey-sian Data Analysis



https://simplystatistics.org/2019/04/17/tukey-design-thinking-and-better-questions/

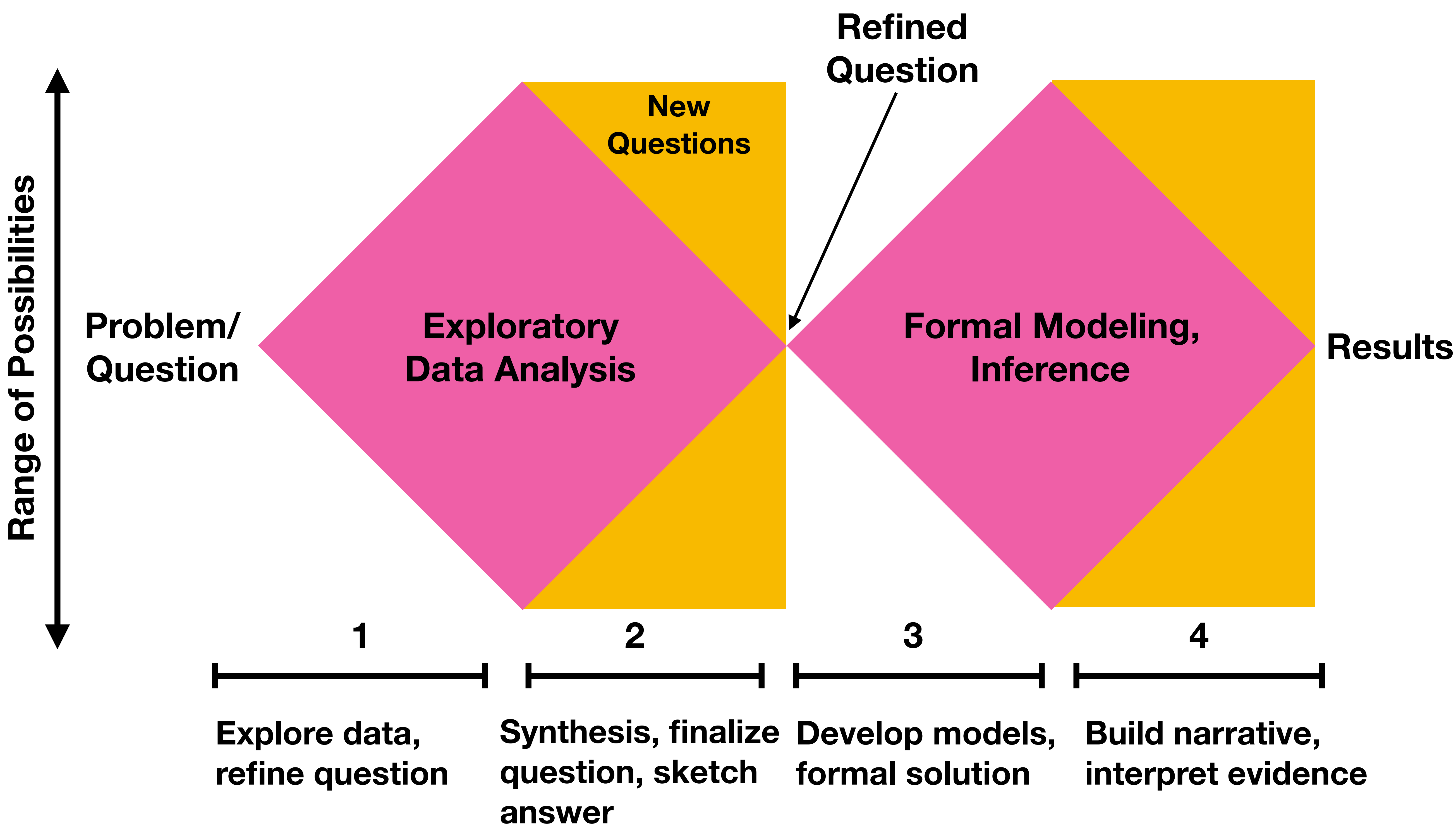
“In my experience when a moderately good solution to a problem has been found, it is seldom worth while to spend much time trying to convert this to the 'best' solution. The time is much better spent in real research.”

–George Kimball, "A critique of operations research,"
J. Wash. Acad. Sci, 1958

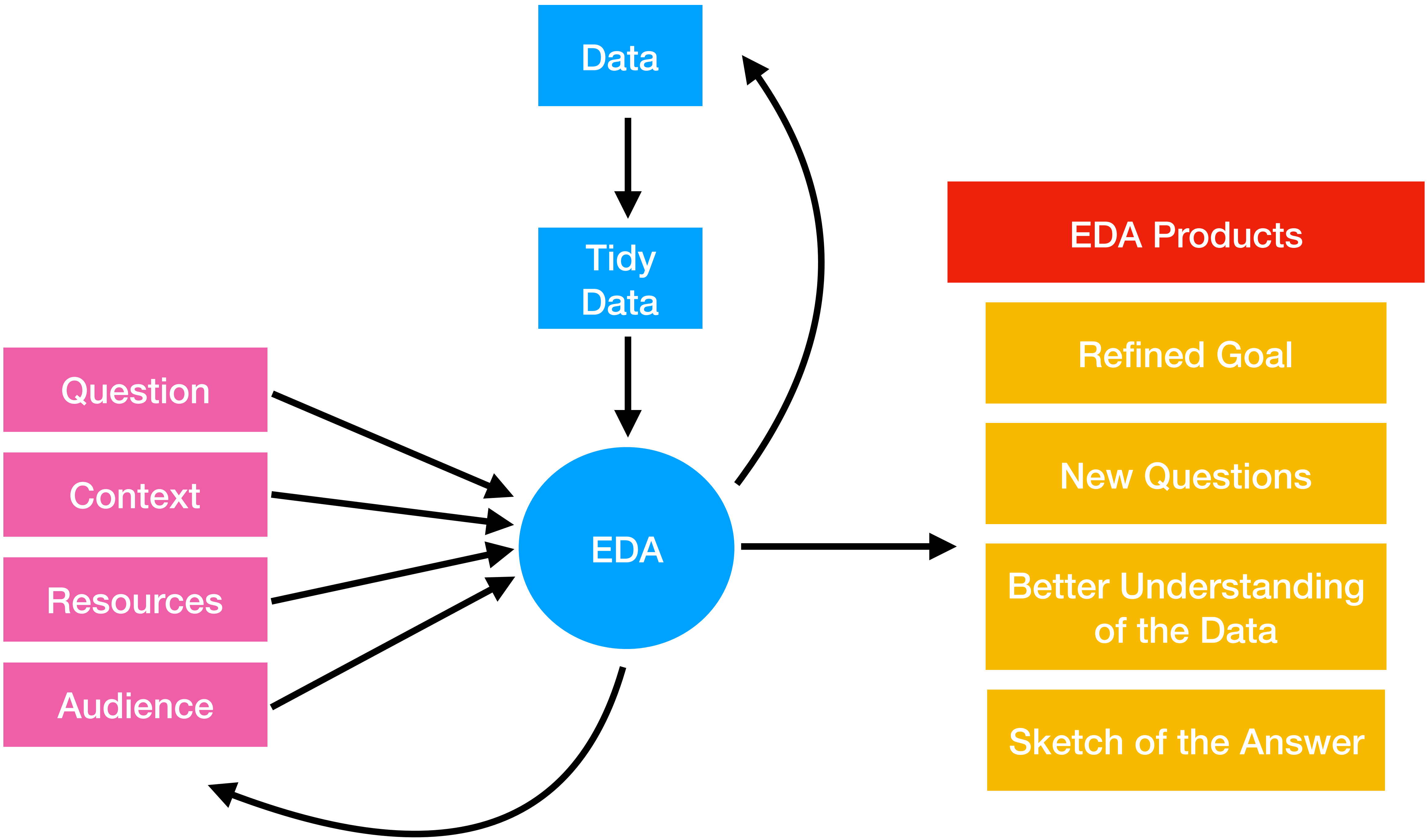
Questions to Resolve

- Do we have the right question?
- Do we have the right data?
- Can we sketch the solution?

Phases of Data Analysis



EDA Process



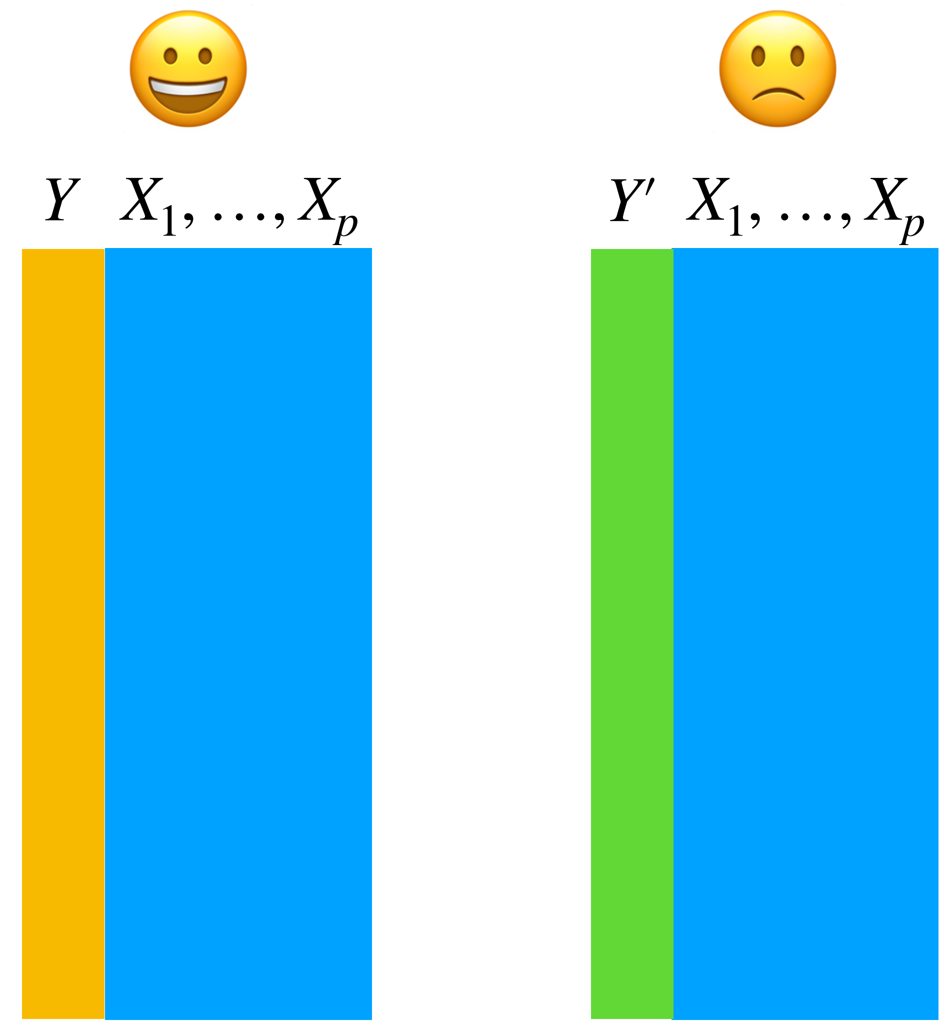
Data
Tidy Data
Question
Context
Resources
Audience
EDA
EDA Products
Refined Goal
New Questions
Better Understanding of the Data
Sketch of the Answer

Do We Have the Right Question?

- Too **vague**
  - Unwieldy analysis
- Too **specific**
  - We don’t have that particular kind of data
  - Affected population is too small
- Does not lead to a **decision** or **intervention**
  - Relevance?

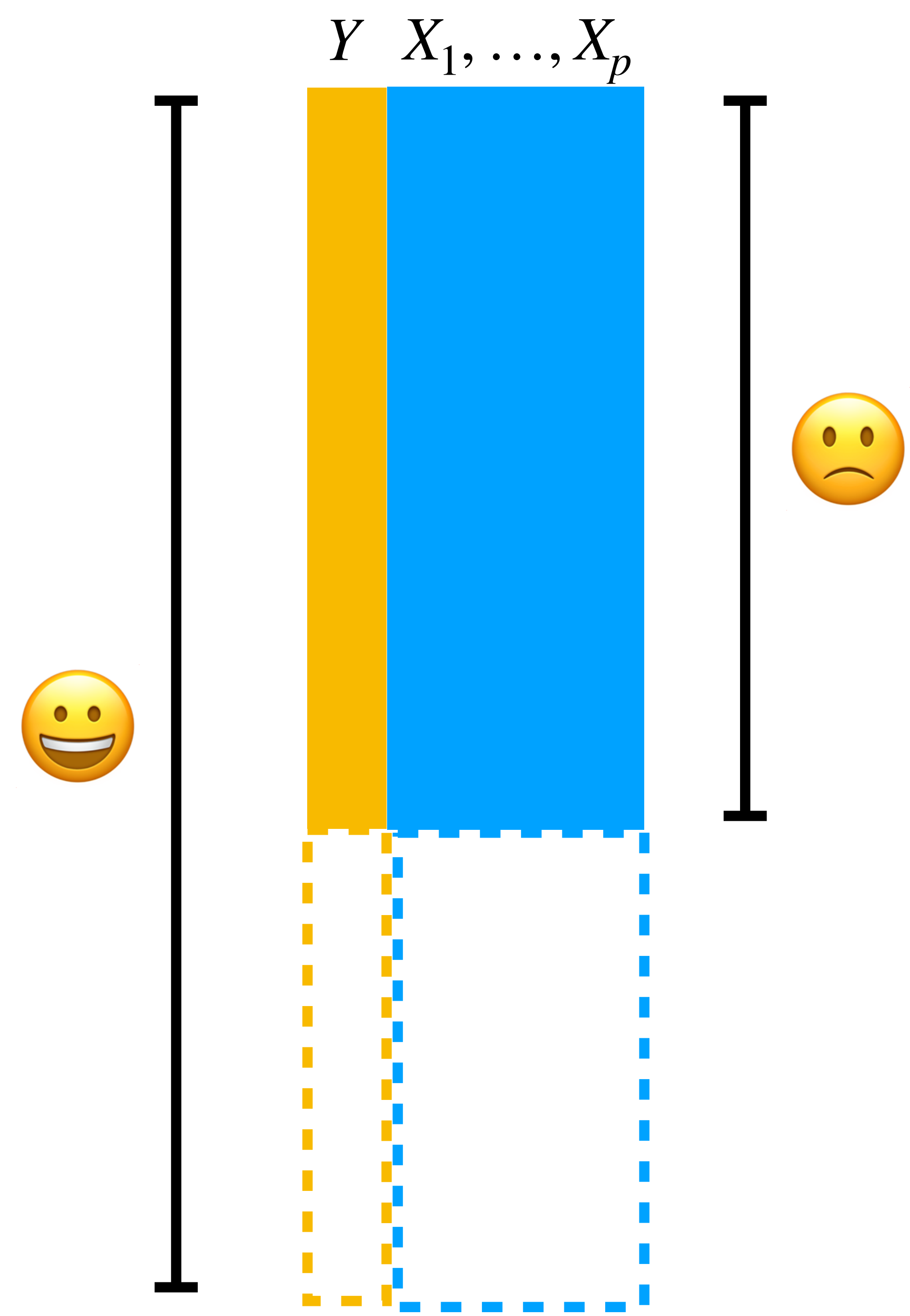
Do We Have the Right Data?

- Data are **proxies** for the key variables
- **Insufficient data** to make reasonable inferences or predictions
- Missing variables that might be **confounders, modifiers**
- **Missing data** prevents complete analysis
- Data with **errors** affects can increase bias, uncertainty



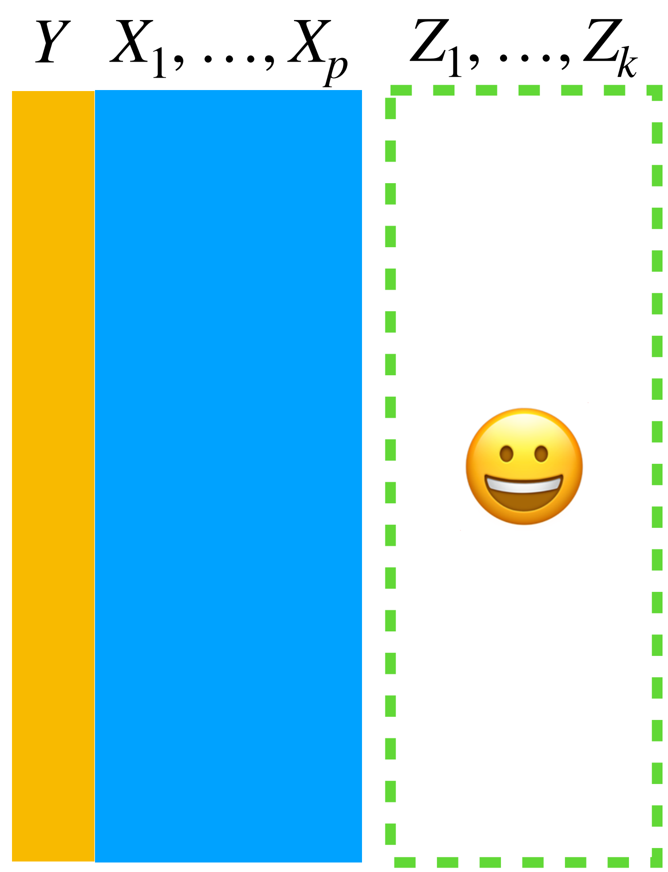
Do We Have the Right Data?

- Data are **proxies** for the key variables
- **Insufficient data** to make reasonable inferences or predictions
- Missing variables that might be **confounders, modifiers**
- **Missing data** prevents complete analysis
- Data with **errors** affects can increase bias, uncertainty



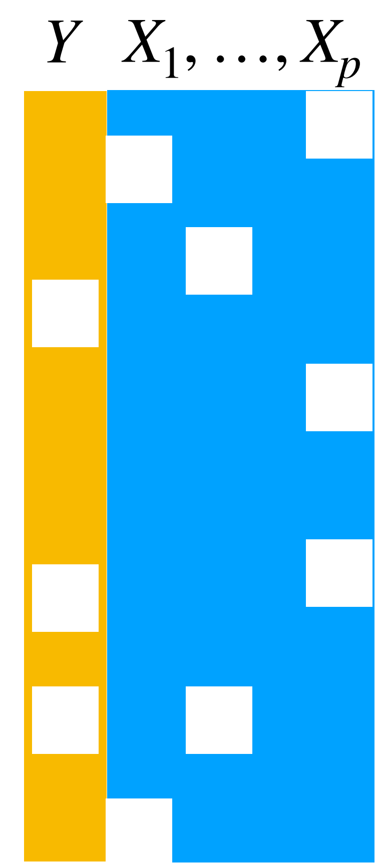
Do We Have the Right Data?

- Data are **proxies** for the key variables
- **Insufficient data** to make reasonable inferences or predictions
- Missing variables that might be **confounders, modifiers**
- **Missing data** prevents complete analysis
- Data with **errors** affects can increase bias, uncertainty



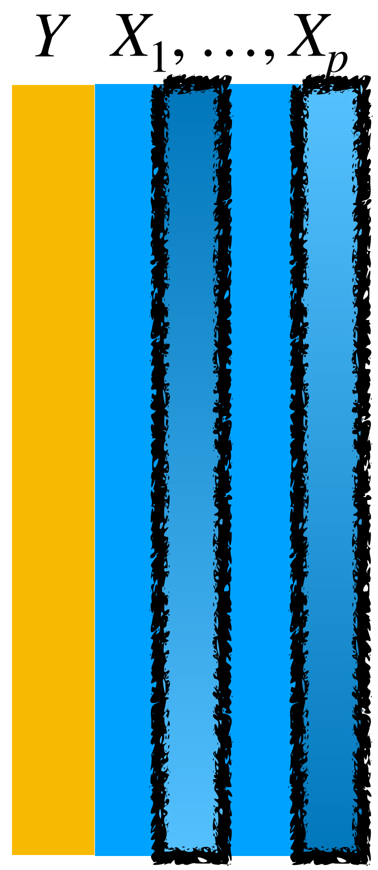
Do We Have the Right Data?

- Data are **proxies** for the key variables
- **Insufficient data** to make reasonable inferences or predictions
- Missing variables that might be **confounders, modifiers**
- **Missing data** prevents complete analysis
- Data with **errors** affects can increase bias, uncertainty



Do We Have the Right Data?

- Data are **proxies** for the key variables
- **Insufficient data** to make reasonable inferences or predictions
- Missing variables that might be **confounders**, **modifiers**
- **Missing data** prevents complete analysis
- Data with **errors** affects can increase bias, uncertainty

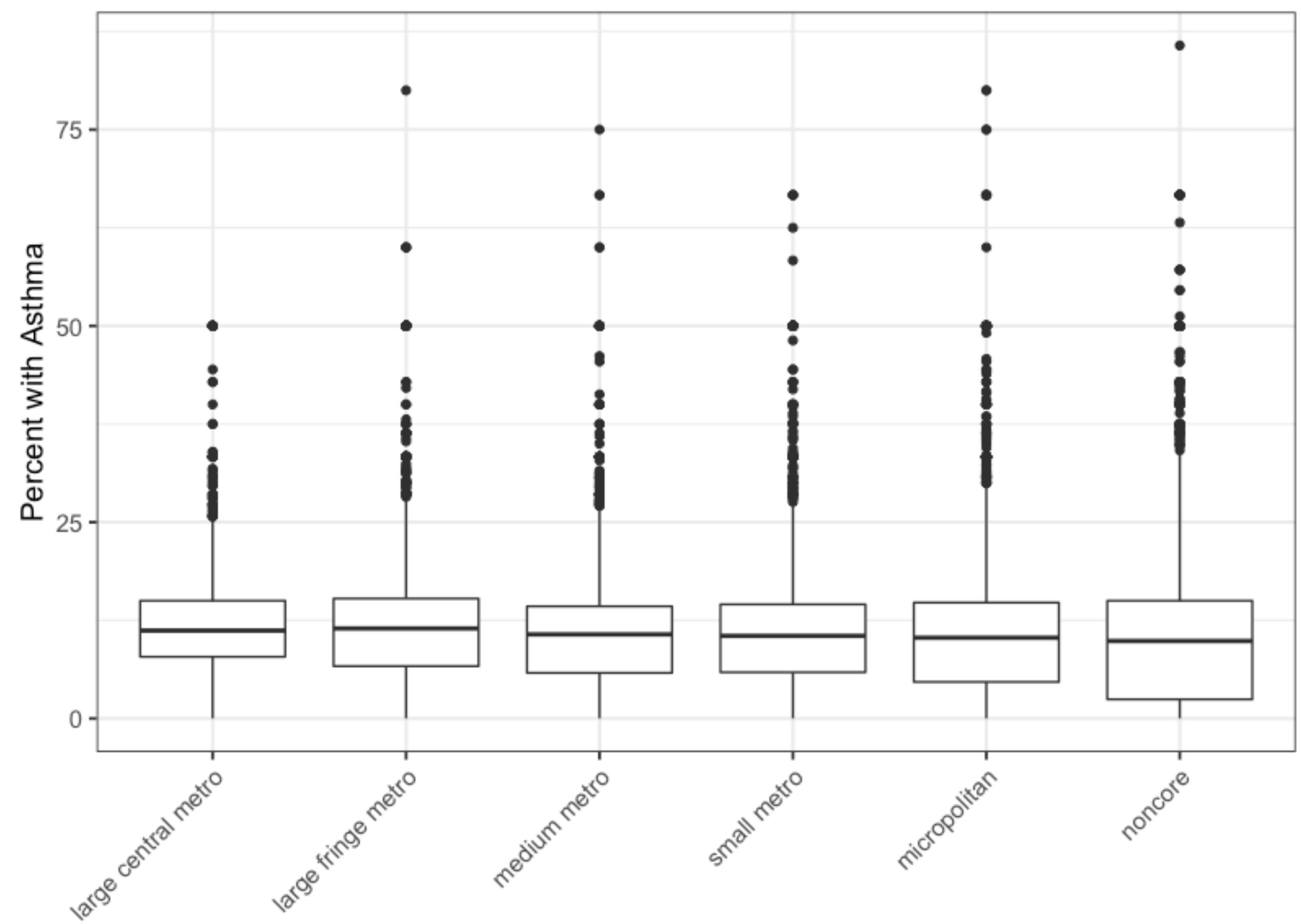


Can We Sketch the Solution?
("Lo-Fi" Model)

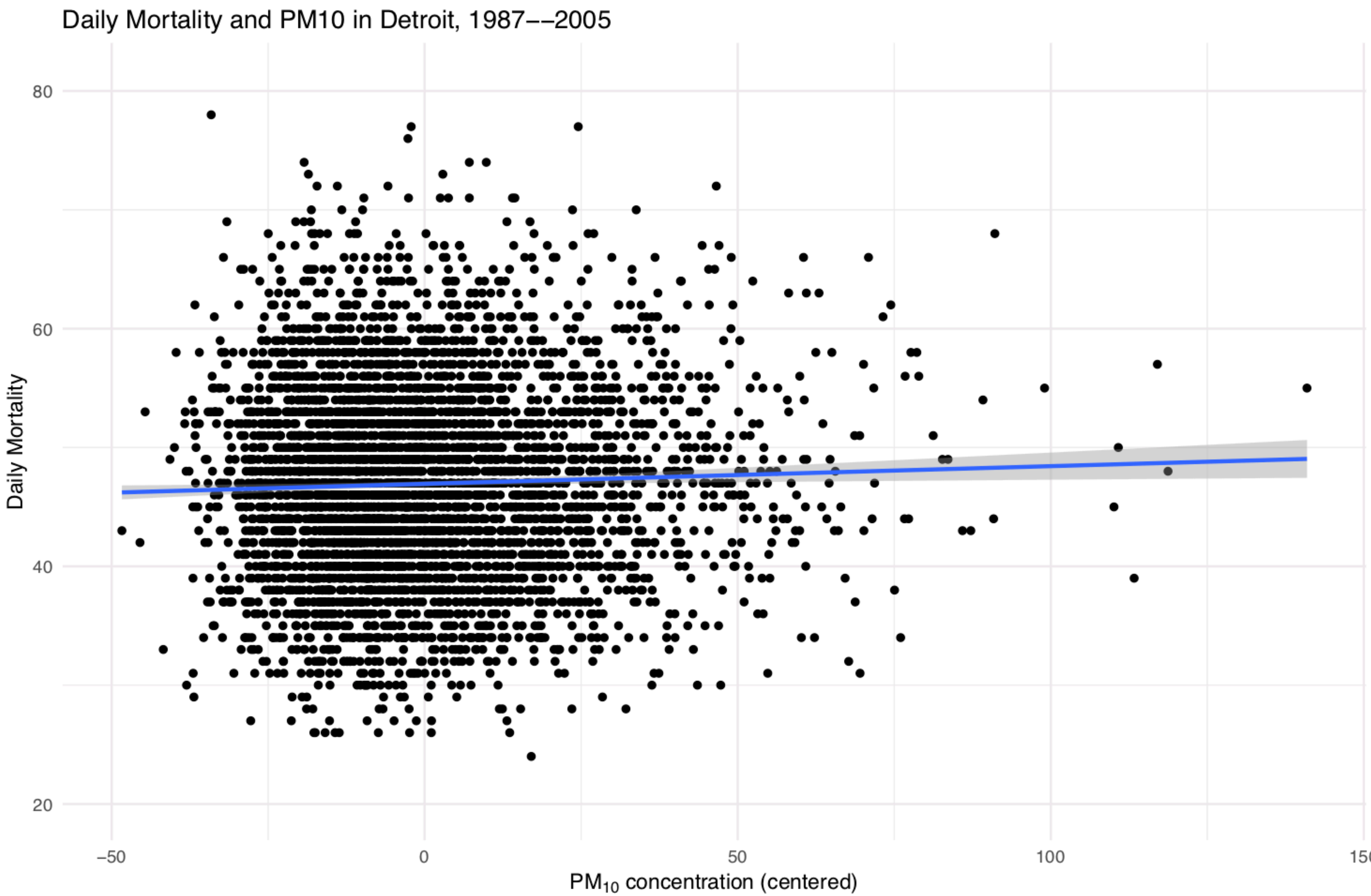
Can We Sketch the Solution?

- Is there any **signal** in the data?
- A **picture**, table, or figure that tells us 80% of the answer
- A simplified **model** that indicates predictive power
- Further work will test the **sensitivity** and **robustness** of our solution
- The sketch will almost **never be seen** by outsiders

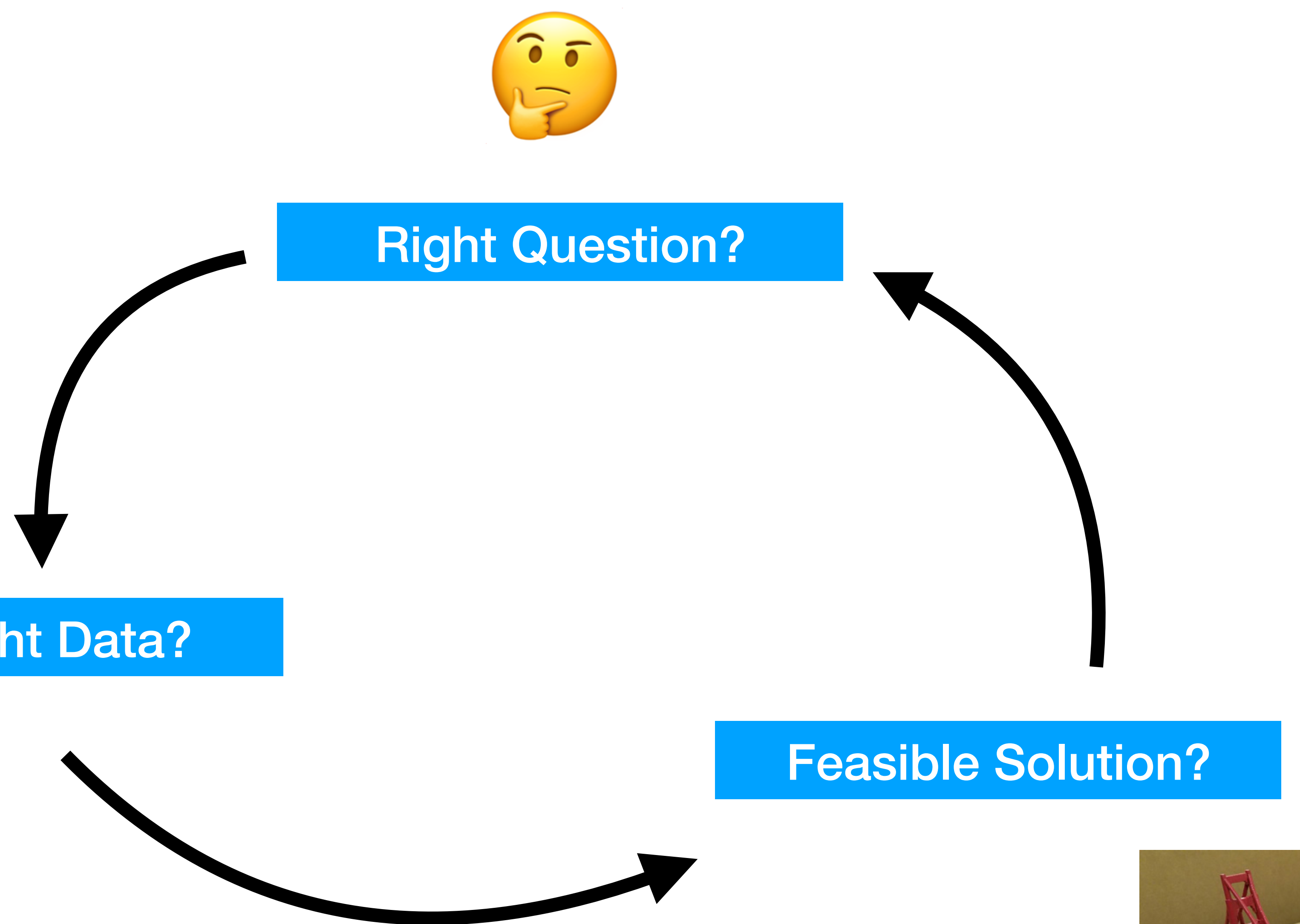
“Inner City” Asthma?

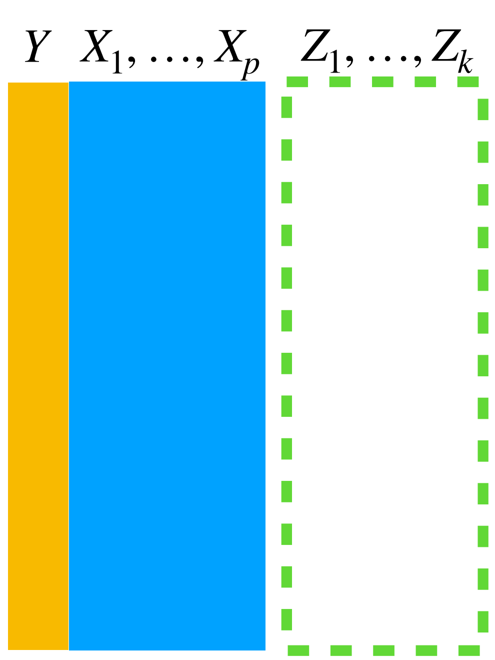


Can We Sketch the Solution?



EDA Macro Cycle





EDA Epicycle

"The value of a plot is that it allows us to see what we never expected to see."

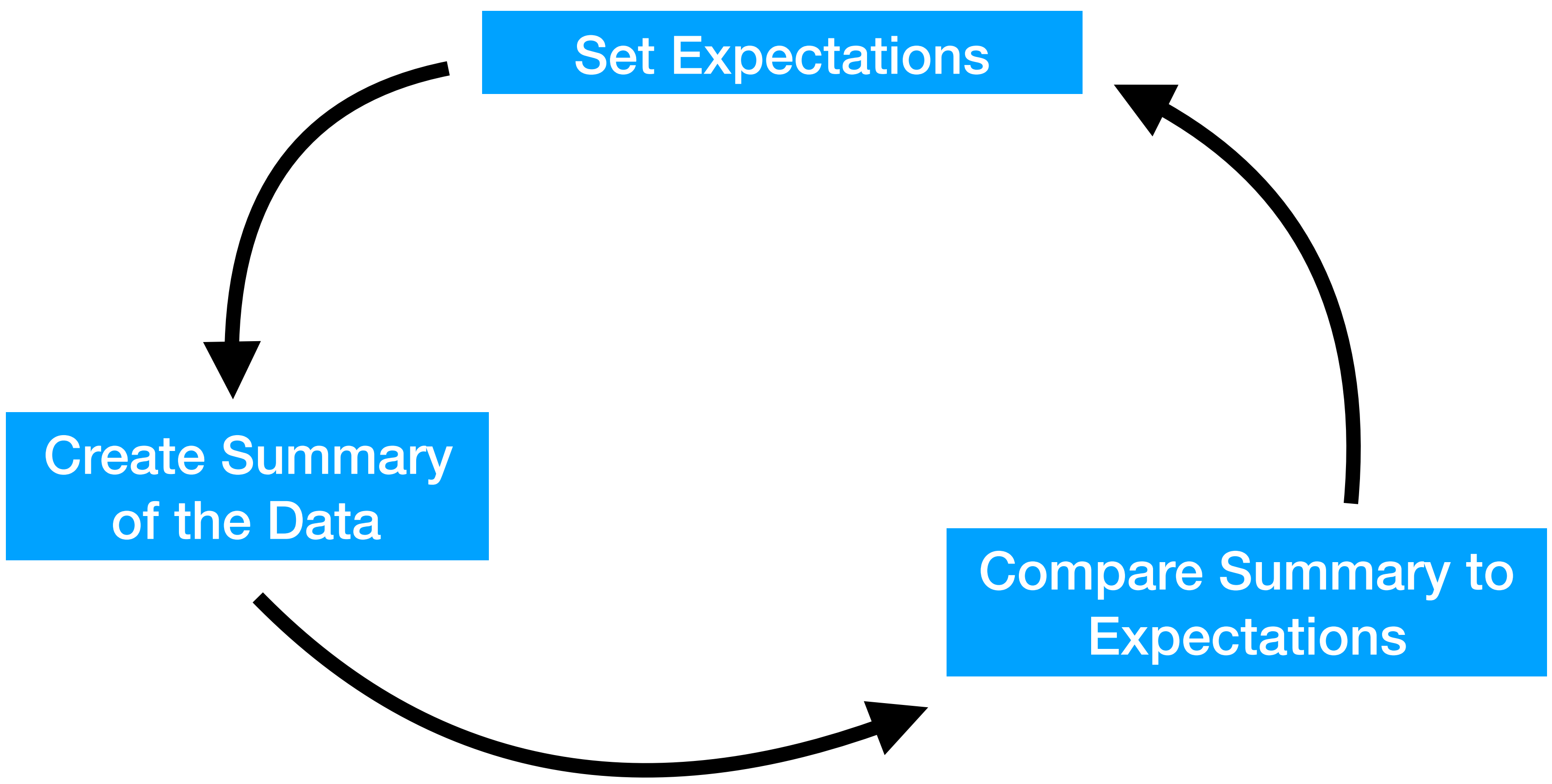
–*John Tukey,* Exploratory Data Analysis

EDA Epicycle

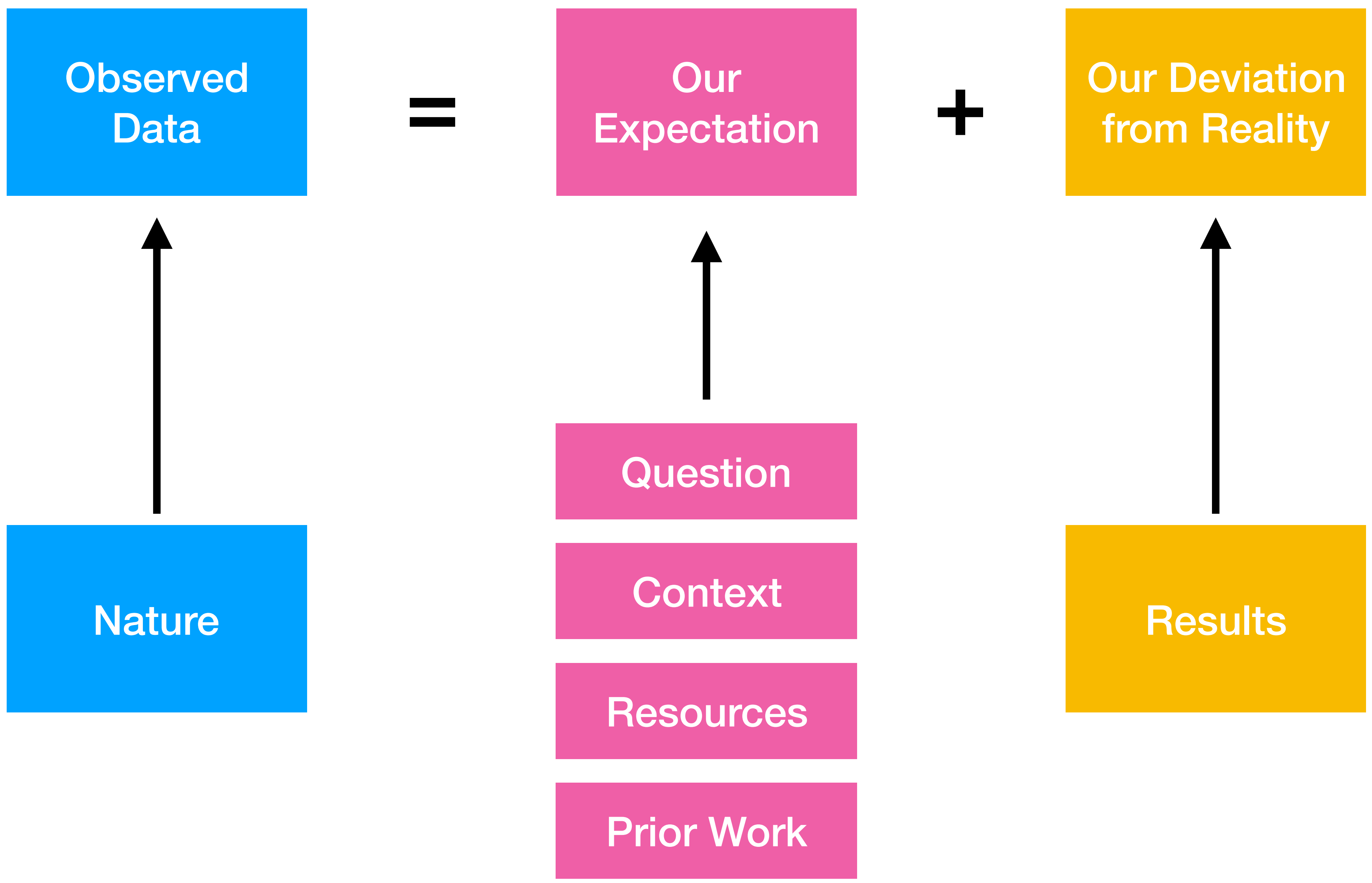
"The value of a plot is that it allows us to **see** what we never **expected to see**."

–*John Tukey,* Exploratory Data Analysis

EDA Epicycle



Expectations vs. Reality



Prevalence of Asthma Amongst Medicaid Enrollees

Age Category

	(5,8]	(8,11]	(11,14]	(14,17]	(17,20]
asian	12.4	10.2	6.9	5.0	3.7
black	15.7	14.5	12.0	10.3	9.5
hispanic	13.9	12.7	10.4	8.7	7.0
other	14.5	13.6	11.5	9.6	8.4
white	12.0	11.3	10.3	9.5	8.7

% of People with Asthma in Medicaid, 2009—2010

Prevalence of Asthma Amongst Medicaid Enrollees

Age Category

	(5,8]	(8,11]	(11,14]	(14,17]	(17,20]
asian	12.4	10.2	6.9	5.0	3.7
black	15.7	14.5	12.0	10.3	9.5
hispanic	13.9	12.7	10.4	8.7	7.0
other	14.5	13.6	11.5	9.6	8.4
white	12.0	11.3	10.3	9.5	8.7

Observed % = $\underbrace{\text{Age + Race}}_{\text{Expectation}}$ + $\underbrace{\text{residual}}_{\text{Deviation}}$

Expectation

Deviation

% of People with Asthma in Medicaid, 2009—2010

Creating More Data With Median Polish

EDA Pre-Flight Check List

- Check the packaging
- Look at the top and bottom of your data
- Check your "n"s
- Validate with at least one external data source
- Make a plot
- Try the easy solution first
- Follow up

Check the Packaging

- What can you learn about the dataset *before* looking directly at the data?
- Check rows and columns
- Check metadata; are all variables there that you expected?
- Are all metadata present?

Look at the Top and Bottom

- Okay, now you can look at the data
- Check the first few rows
- Check the *last* few rows; make sure all rows were read properly and there's no crud at the end
- Time/Date data often sorted; make sure all dates/ times are in appropriate range

ABC: Always be Counting

- Count various aspects of your dataset
- Compare counts with landmarks
- Number of subjects (unique IDs), number of visits per subject, number of locations, number of missing observations, etc.
- Always be counting at every phase ("checking mindset")

Validate With At Least 1 External Source

- Compare your data to something outside the dataset
- Even a single number/summary statistic comparison can be useful
- Compare your measurements to another similar measurement to check that they're correlated
- Get external upper/lower bounds
- Ex: number of people should exceed total population
- Ex: Check for negative values when they should be positive

Make a Plot

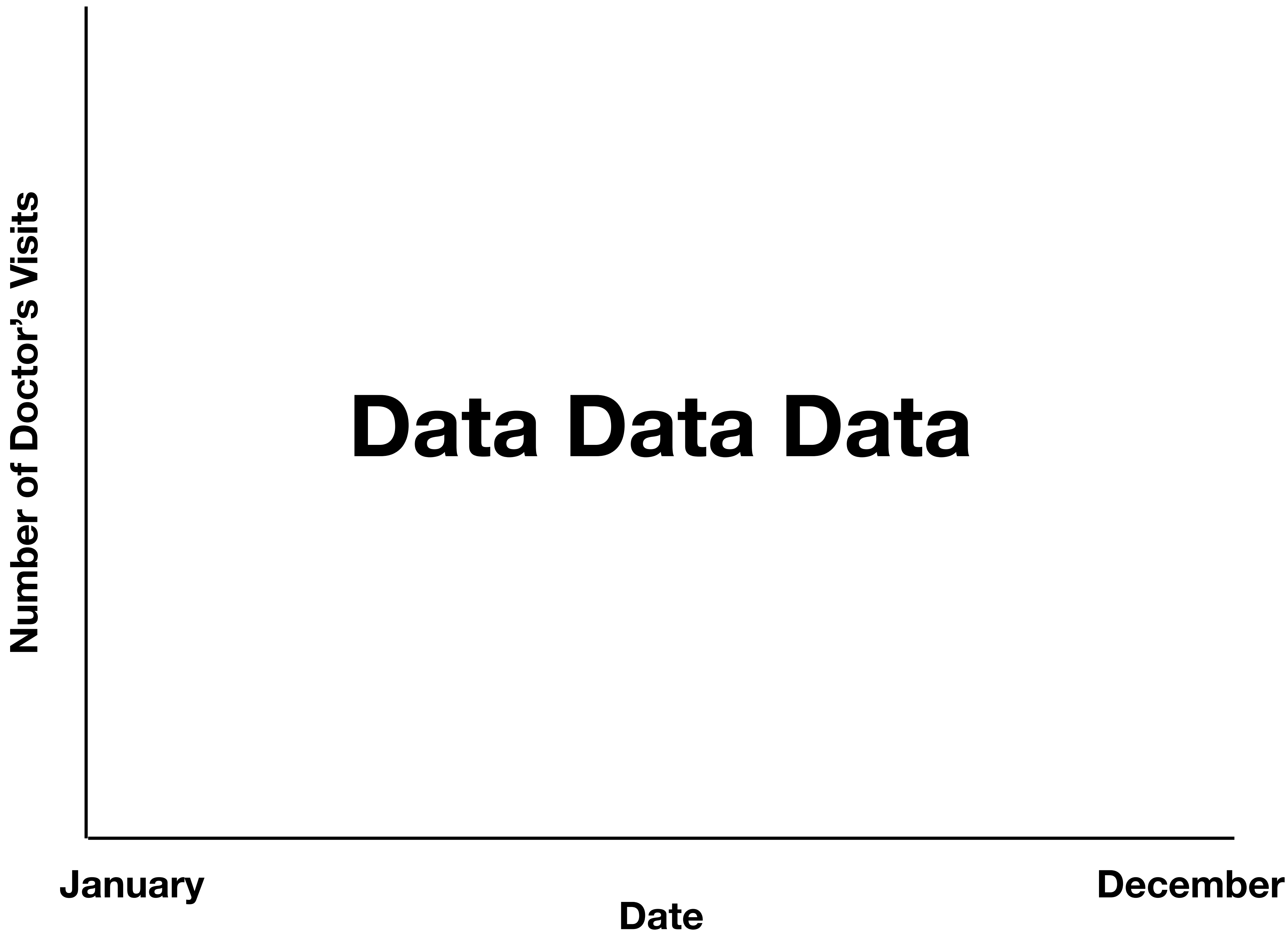
- Plots show expectations *and deviations* from those expectations (i.e. distribution mean and outliers)
- Tables generally only show summaries, not deviations; also everything on the same “scale”
- Draw a “fake plot” first

Try the Easy Solution

- First step in building a **primary model**
- Build *prima facie* evidence
- Basic argument, without nuance (that comes later)
- Maybe just one plot (or table)

Follow Up

- Do you have the right question?
- Do you have the right data?
- Do you need other data?
- Could you sketch the solution?
- Is there signal in the data?



Daily doctor's visits for asthma amongst Medicaid enrollees in Maryland, 2010

Exploring Data With Models

“All models are wrong, but some are useful.”

–George Box

Exploring With Models

- Models represent a formalization of our expectations
- Models can tell us about the *unobserved population*
- Whether a model fits well depends on the **question**

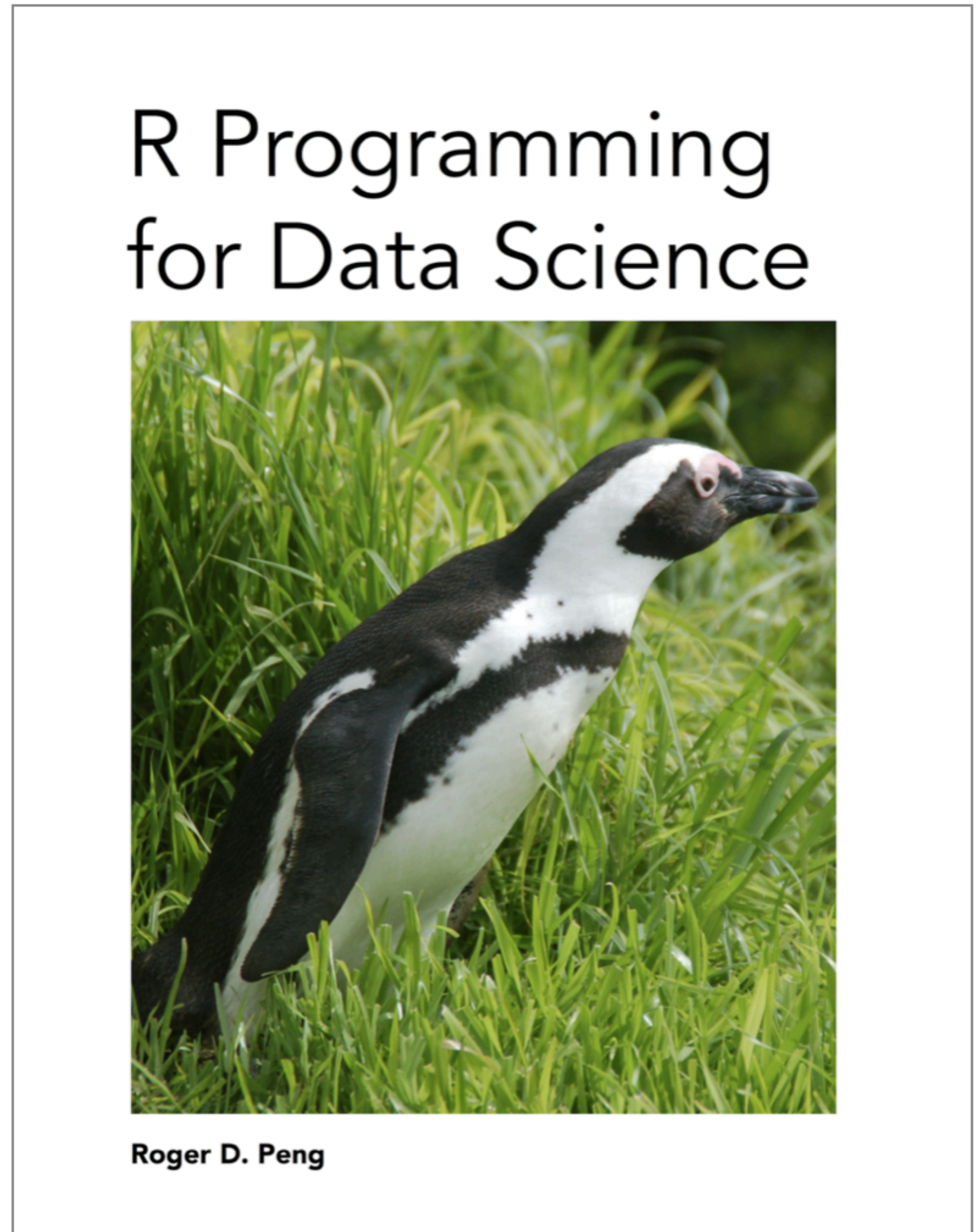
Selling A Book

R Programming for Data Science

Roger D. Peng

This book brings the fundamentals of R programming to you, using the same material developed as part of the industry-leading Johns Hopkins Data Science Specialization. The skills taught in this book will lay the foundation for you to begin your journey learning data science. Printed copies of this book are available through Lulu.

Table Of Contents



LAST UPDATED ON 2019-05-07

Selling A Book

Interested in this book? Show your support by saying what you'd like to pay for it!

NAME

EMAIL

☐ Share email with author (optional)

I'D BUY IT FOR...

$

Notify Me When This Is Published

Selling A Book

- What is the question? What is the goal? (hint: 💰)
- Do we have the right data?
- Can we sketch a solution?

Data Data Data