Questions and Context in Data Science

Roger D. Peng
Stephanie C. Hicks

Advanced Data Science
Term 1
2019

Message of the Day

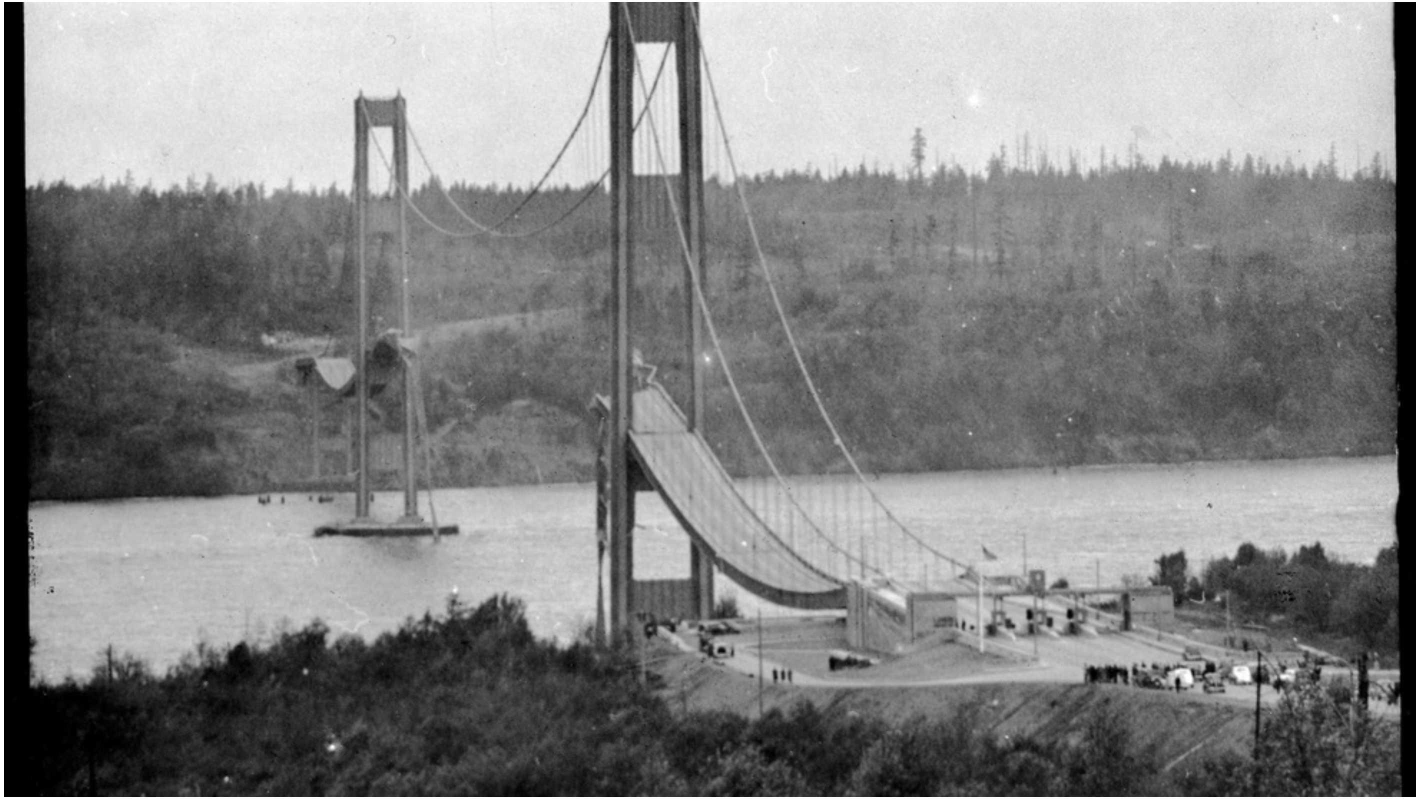
- Context plays a critical role in conducting and interpreting data analysis
- Matching an analysis to its question is a key aspect of determining analysis quality
- Knowing the type of question being asked can avoid common mistakes and help evaluate success

What is Data Analysis?

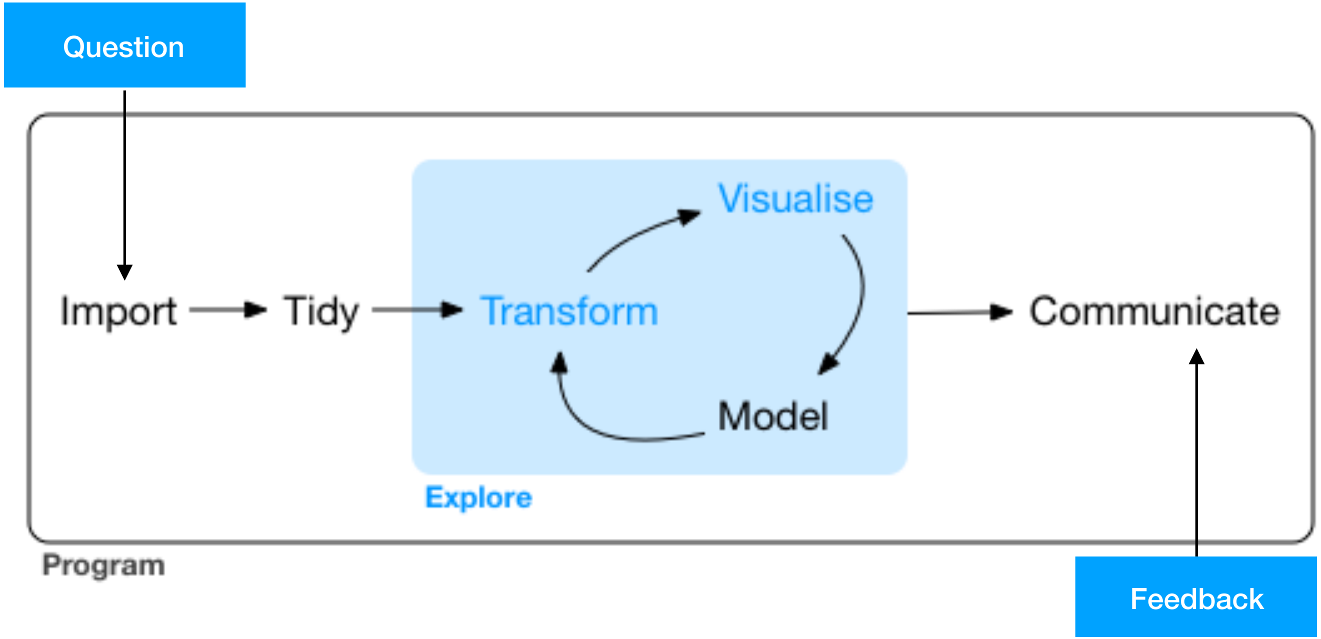
What is Data Analysis?

- Does not occur naturally
- Must be *designed* to be useful
- Must follow basic structural principles (or else collapse)
- Solution can take many forms

What is Data Analysis?

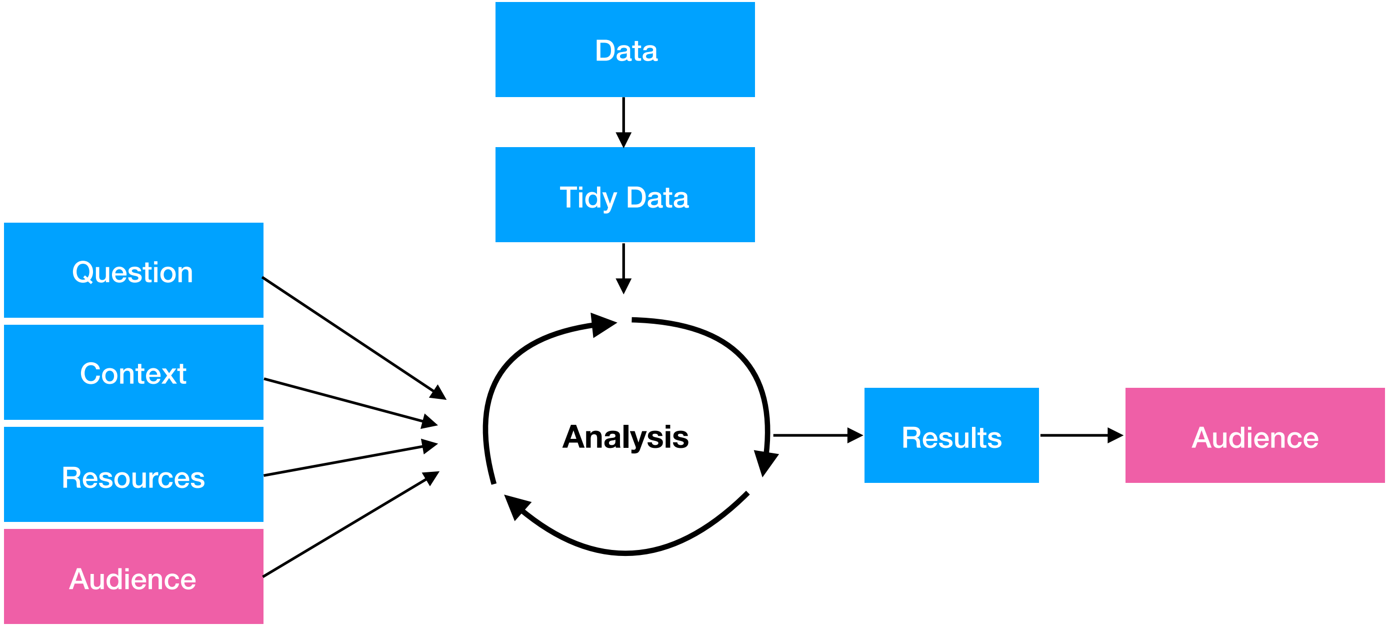


Data Analysis

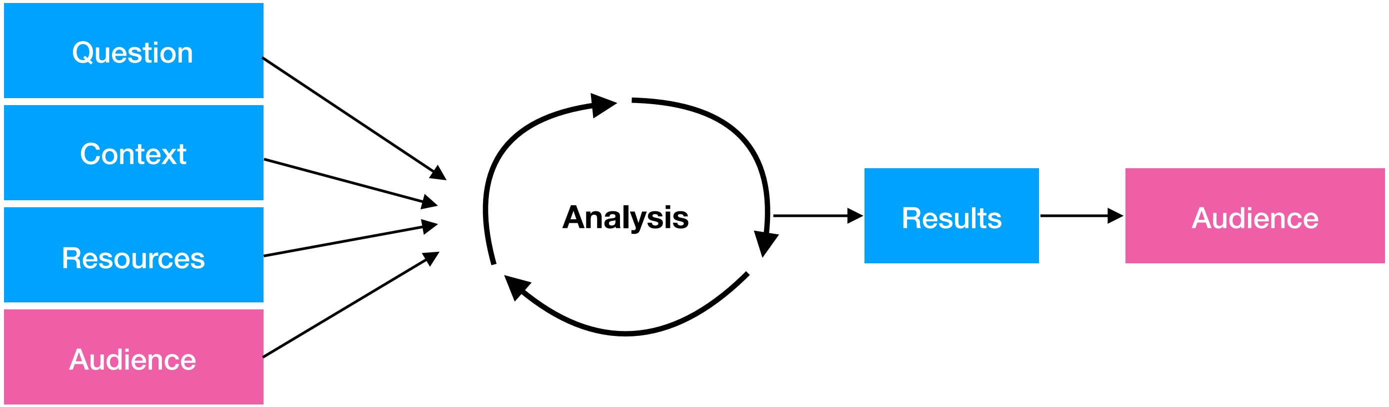


Wickham, *R for Data Science*

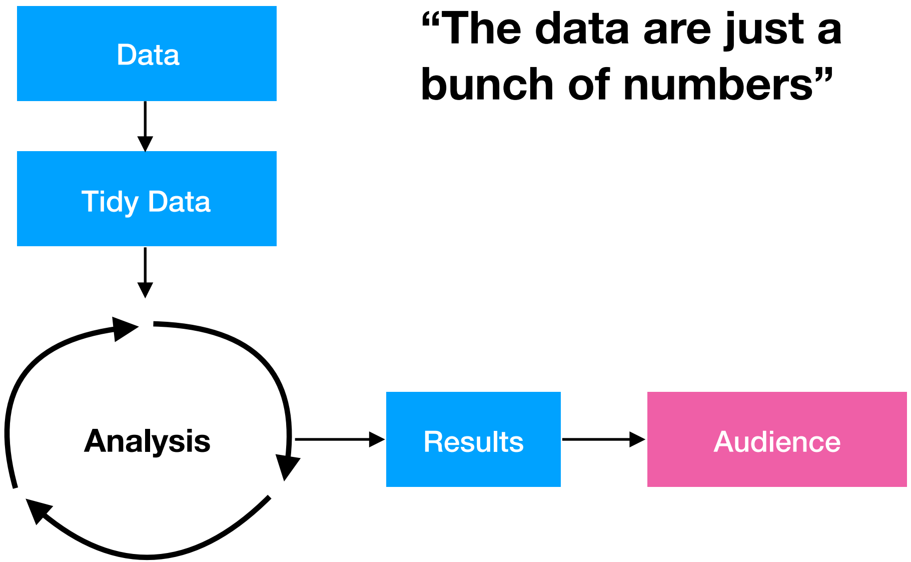
Data Analysis (revised)



Data Analysis (revised)



Data Analysis (revised)

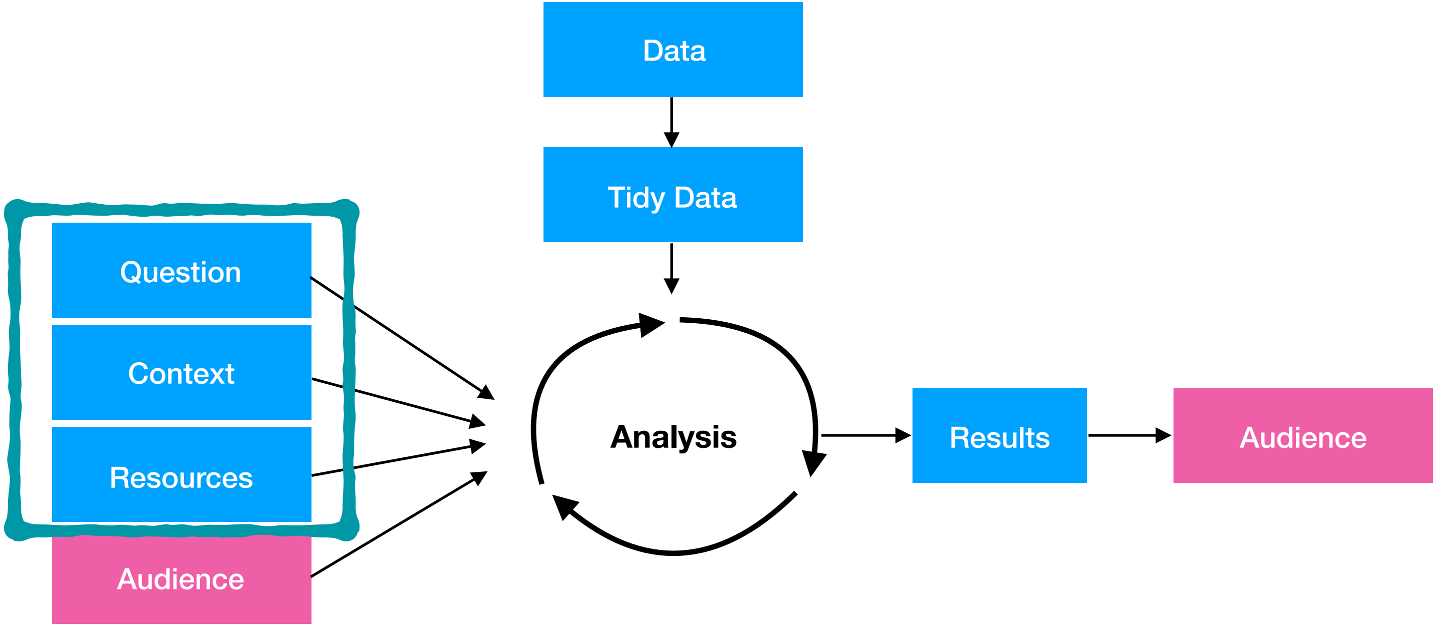


“The data are just a bunch of numbers”

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)

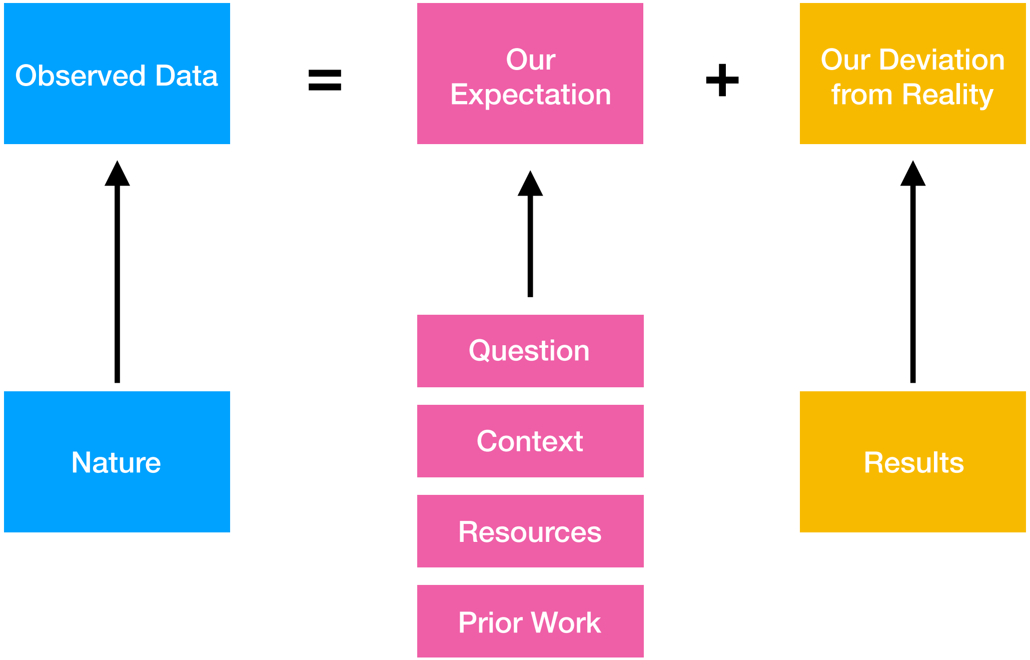


Data Analysis Expectations

Observed Data = Our Expectation + Our Deviation from Reality

© JamPress

Data Analysis Expectations



Observed Data = Our Expectation + Our Deviation from Reality

Nature

Question

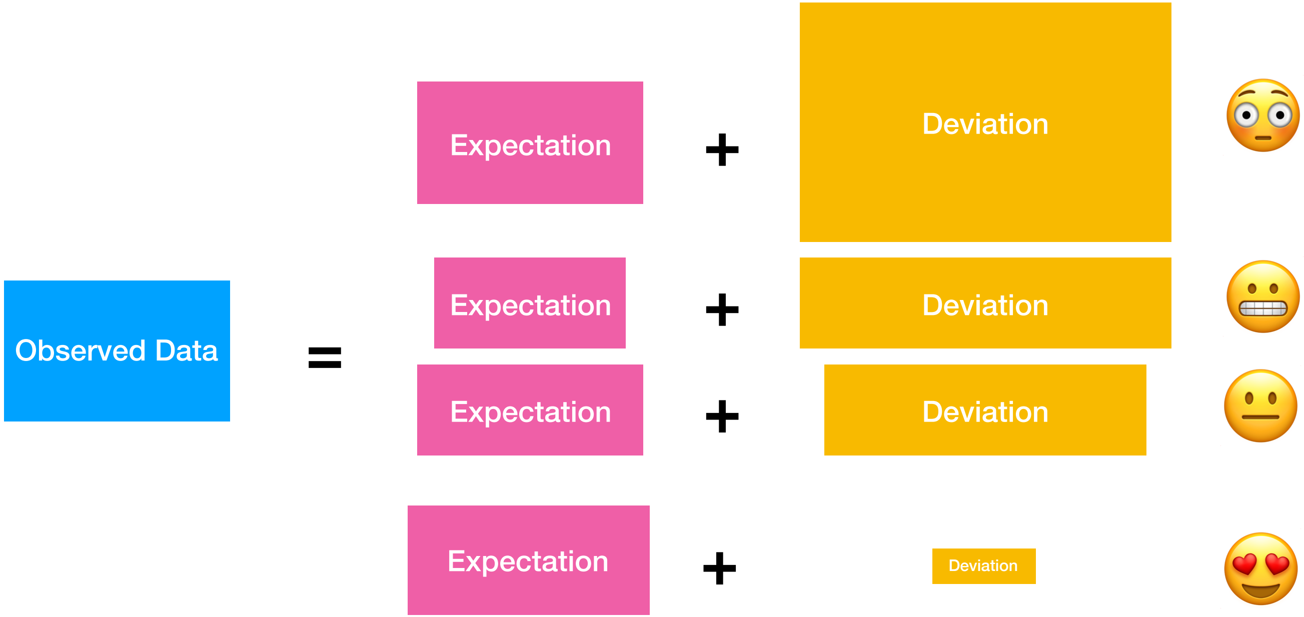
Context

Resources

Prior Work

Results

Data Analysis Expectations



Observed Data =

Expectation + Deviation 😳

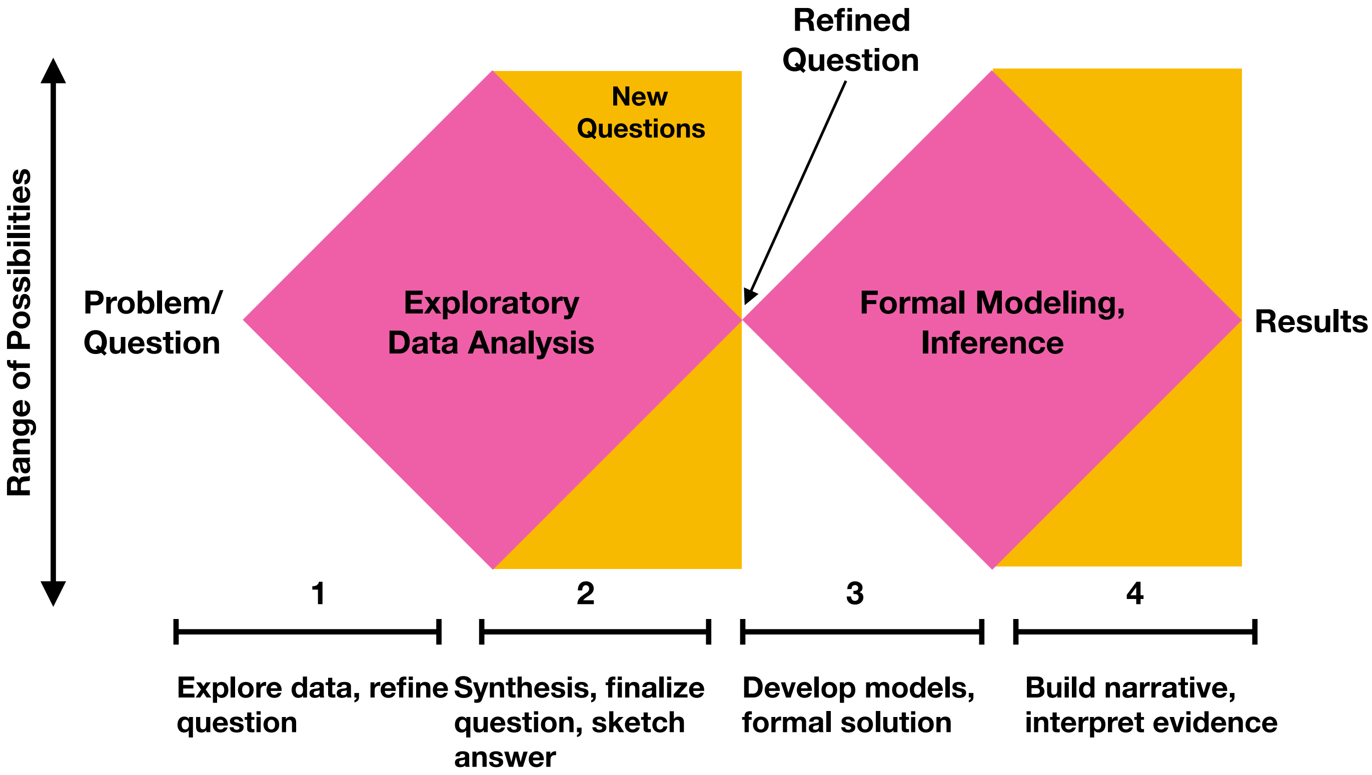
Expectation + Deviation 😬

Expectation + Deviation 😐

Expectation + Deviation 😍

Level of “surprise”

Phases of Data Analysis



Range of Possibilities

Problem/ Question

Exploratory Data Analysis

New Questions

Refined Question

Formal Modeling, Inference

Results

1

Explore data, refine question

2

Synthesis, finalize question, sketch answer

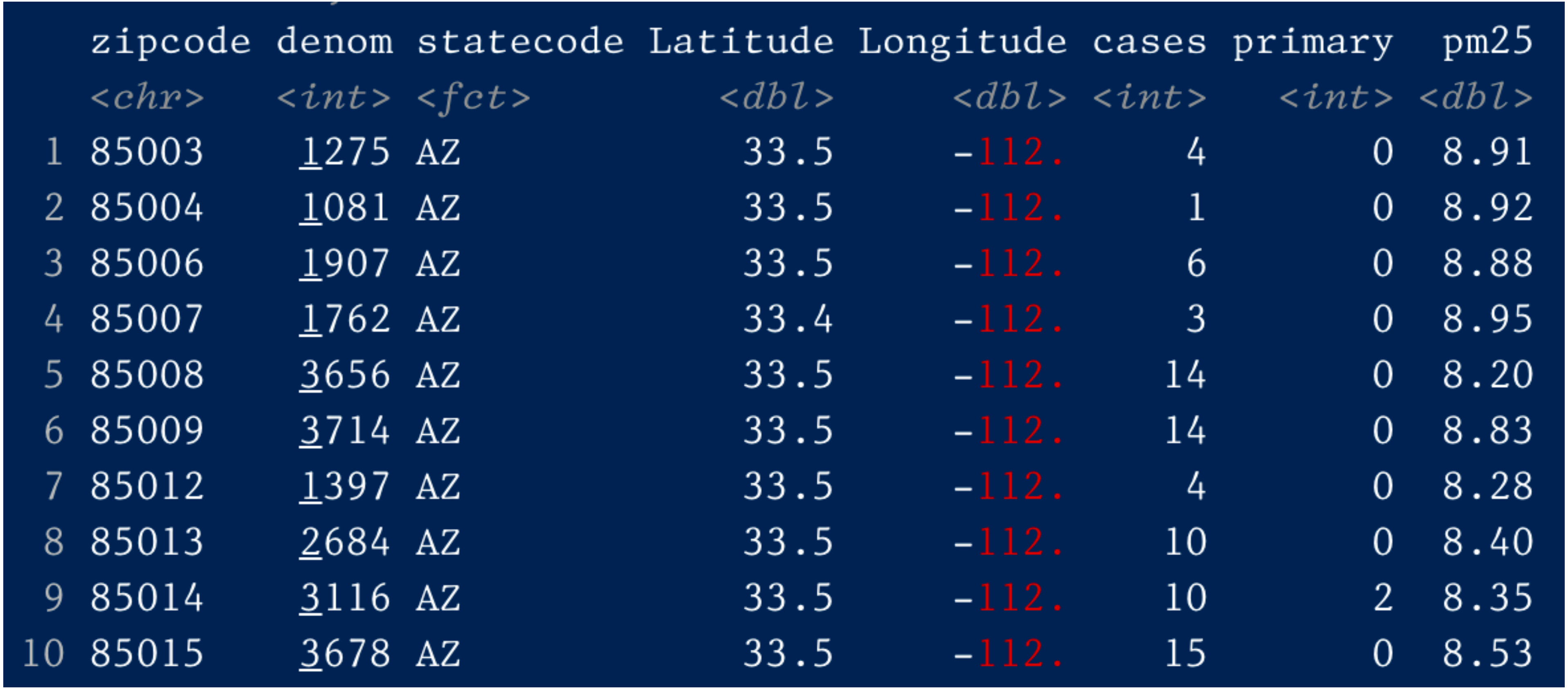
3

Develop models, formal solution

4

Build narrative, interpret evidence

Pulmonary Hypertension in California and Arizona, 2010



```
   zipcode denom statecode Latitude Longitude cases primary  pm25
   <chr>   <int> <fct>        <dbl>     <dbl> <int>   <int> <dbl>
 1 85003    1275 AZ            33.5     -112.     4       0  8.91
 2 85004    1081 AZ            33.5     -112.     1       0  8.92
 3 85006    1907 AZ            33.5     -112.     6       0  8.88
 4 85007    1762 AZ            33.4     -112.     3       0  8.95
 5 85008    3656 AZ            33.5     -112.    14       0  8.20
 6 85009    3714 AZ            33.5     -112.    14       0  8.83
 7 85012    1397 AZ            33.5     -112.     4       0  8.28
 8 85013    2684 AZ            33.5     -112.    10       0  8.40
 9 85014    3116 AZ            33.5     -112.    10       2  8.35
10 85015    3678 AZ            33.5     -112.    15       0  8.53
```

Pulmonary Hypertension in California and Arizona, 2010

- Annual Medicare outpatient/hospital events
- Air pollution levels, emissions, sources
- Weather data
- Census, SES data
- Road network, satellite surface info

Questions

Data Analysis 1

- We fit a Poisson linear regression model with PH cases as the outcome
- Only air pollution and variables correlated with air pollution—temperature, SES vars, road network — were included as predictors
- The coefficient (log-relative risk) for the air pollution variable was 0.003 (95% CI: 0.001, 0.005)

What question is being asked? Evaluate the quality.

Data Analysis 2

- We fit a Poisson linear regression model with PH cases as the outcome and all of the other available variables as predictors
- The the regression procedure indicated that temperature, living near a major road, and living in a high poverty zip code were statistically significant (after a multiple-testing adjustment)

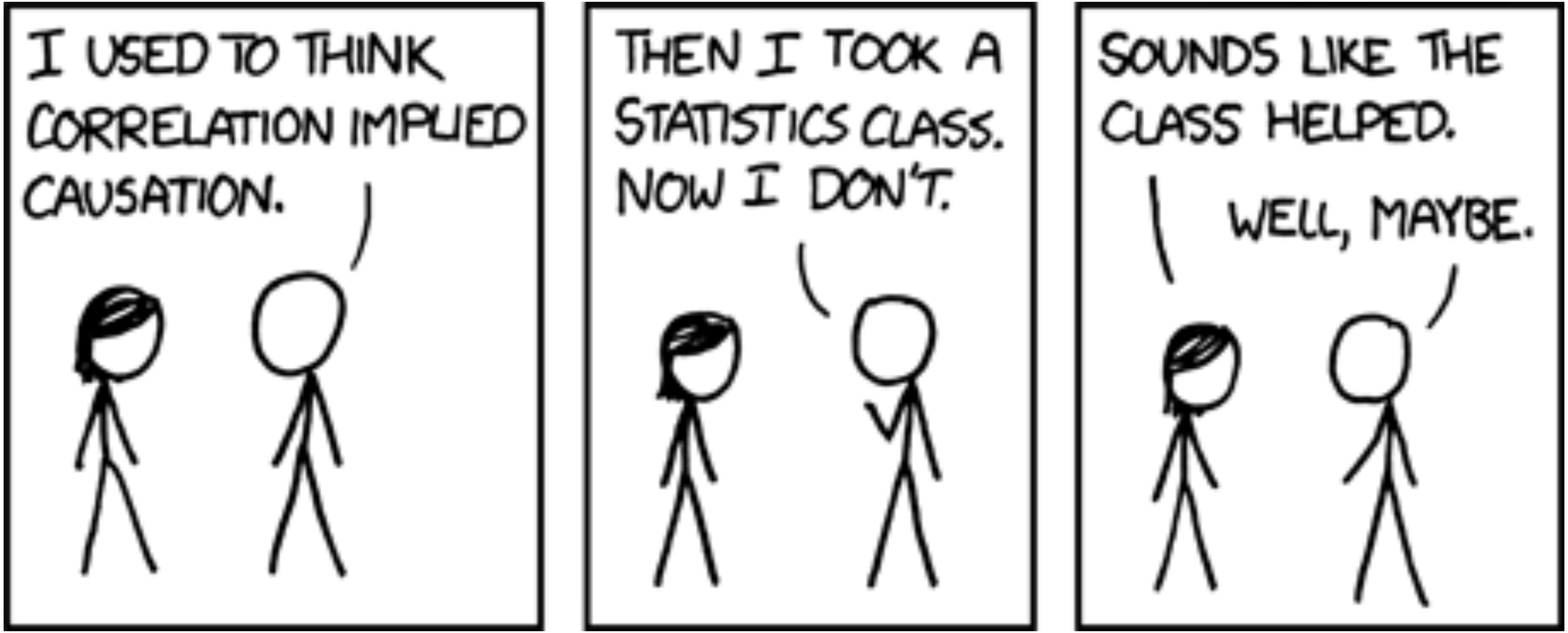
Data Analysis 3

- We used a linear lasso regression model with PH cases as the outcome and all other variables as predictors
- We chose the lasso penalty via 10-fold cross validation
- The non-zero coefficients in the model included air pollution levels, temperature, living in a high poverty zip code, and % people with at least HS education

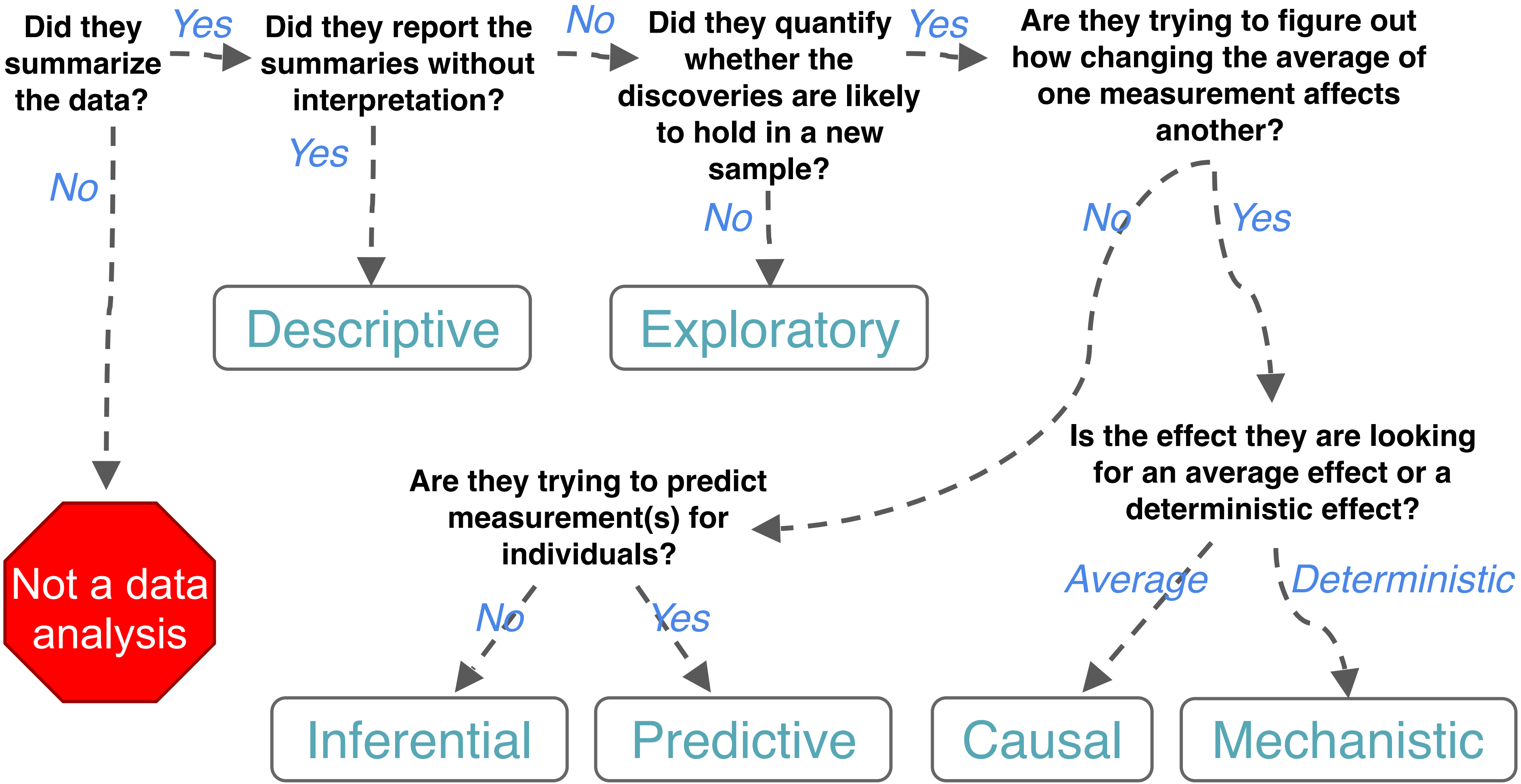
Know Your Question

Know Your Question

- Matching questions with analyses is a key part of the data analysis task
- There are different **types** of questions
- Mistaking different types of questions can result in common errors


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

https://xkcd.com/552/



Did they summarize the data?

Did they report the summaries without interpretation?

Did they quantify whether the discoveries are likely to hold in a new sample?

Are they trying to figure out how changing the average of one measurement affects another?

Descriptive

Exploratory

Is the effect they are looking for an average effect or a deterministic effect?

Are they trying to predict measurement(s) for individuals?

Not a data analysis

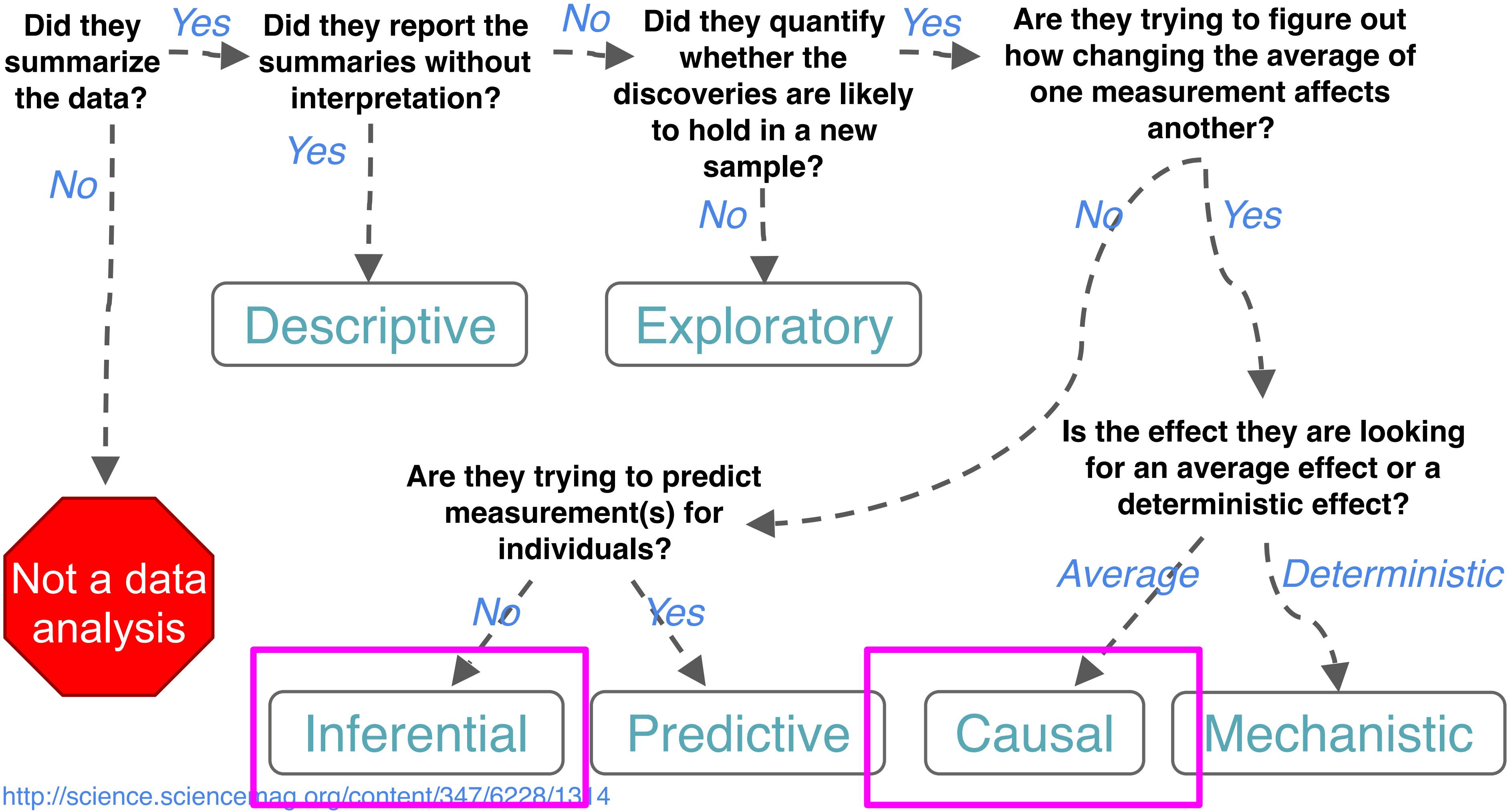
Inferential

Predictive

Causal

Mechanistic

http://science.sciencemag.org/content/347/6228/1314



Did they summarize the data?

Yes

No

Did they report the summaries without interpretation?

Yes

No

Did they quantify whether the discoveries are likely to hold in a new sample?

Yes

No

Are they trying to figure out how changing the average of one measurement affects another?

No

Yes

Is the effect they are looking for an average effect or a deterministic effect?

Average

Deterministic

Are they trying to predict measurement(s) for individuals?

No

Yes

Not a data analysis

Descriptive

Exploratory

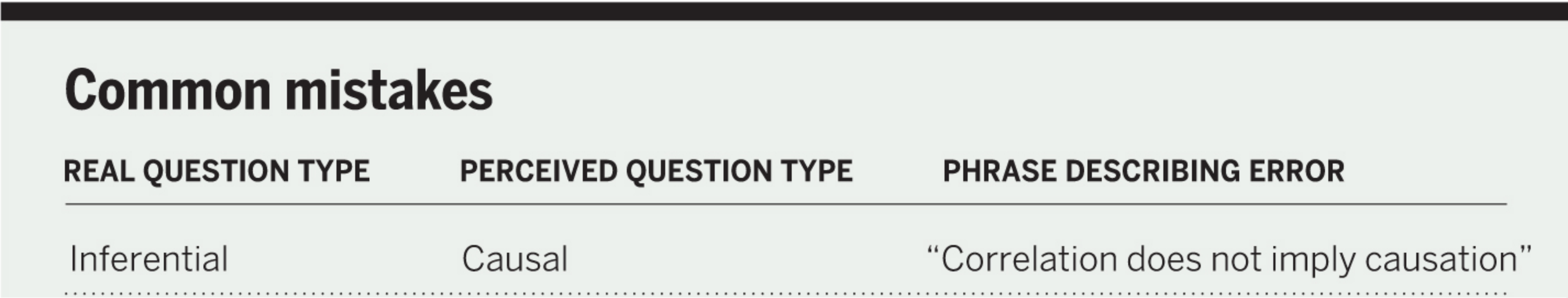
Inferential

Predictive

Causal

Mechanistic

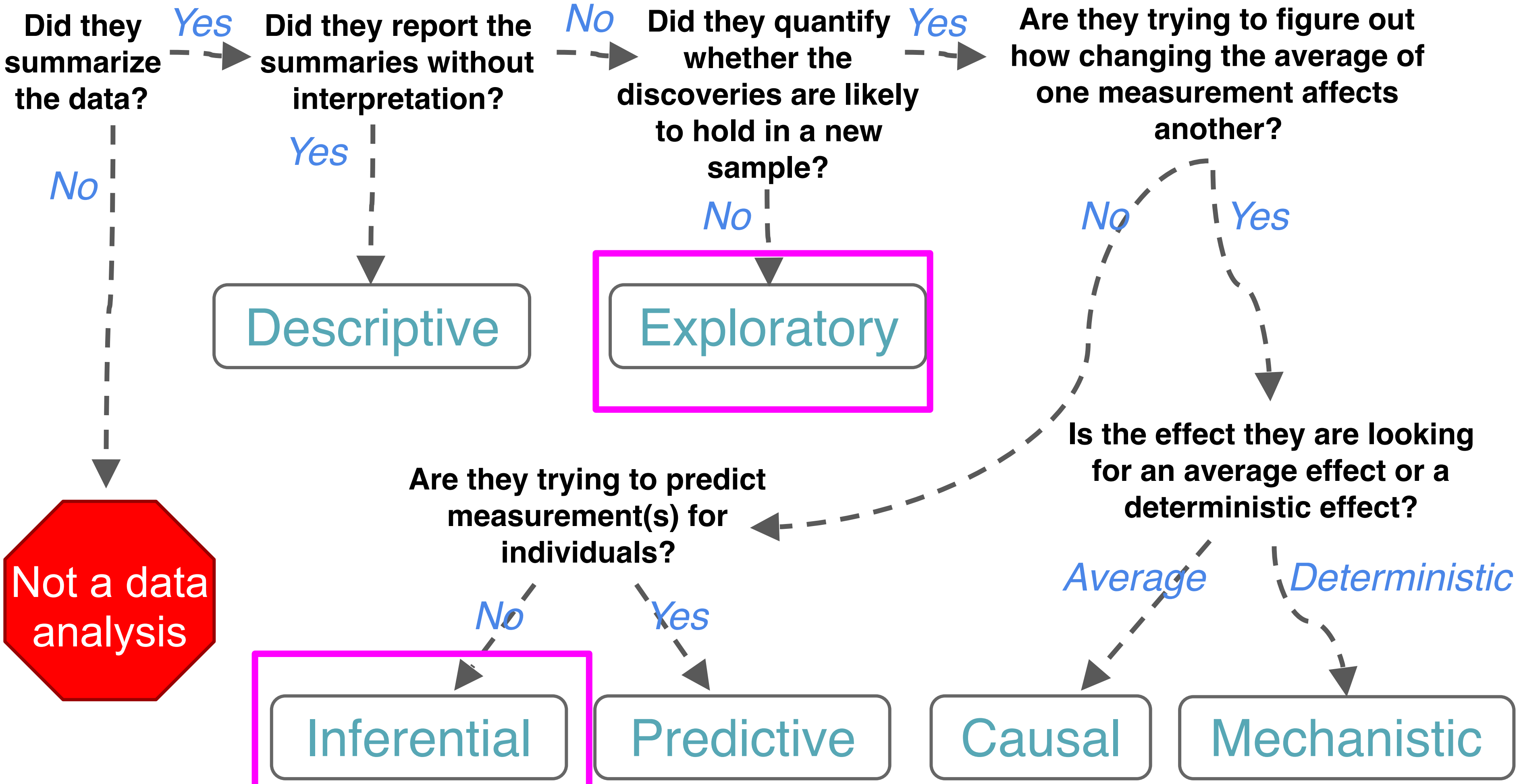
http://science.sciencemag.org/content/347/6228/1314



Common mistakes

REAL QUESTION TYPE	PERCEIVED QUESTION TYPE	PHRASE DESCRIBING ERROR
Inferential	Causal	“Correlation does not imply causation”

http://science.sciencemag.org/content/347/6228/1314



Did they summarize the data?

Yes

Did they report the summaries without interpretation?

No

Did they quantify whether the discoveries are likely to hold in a new sample?

Yes

Are they trying to figure out how changing the average of one measurement affects another?

No

Yes

No

No

Yes

Descriptive

Exploratory

Is the effect they are looking for an average effect or a deterministic effect?

Are they trying to predict measurement(s) for individuals?

Not a data analysis

Average

Deterministic

No

Yes

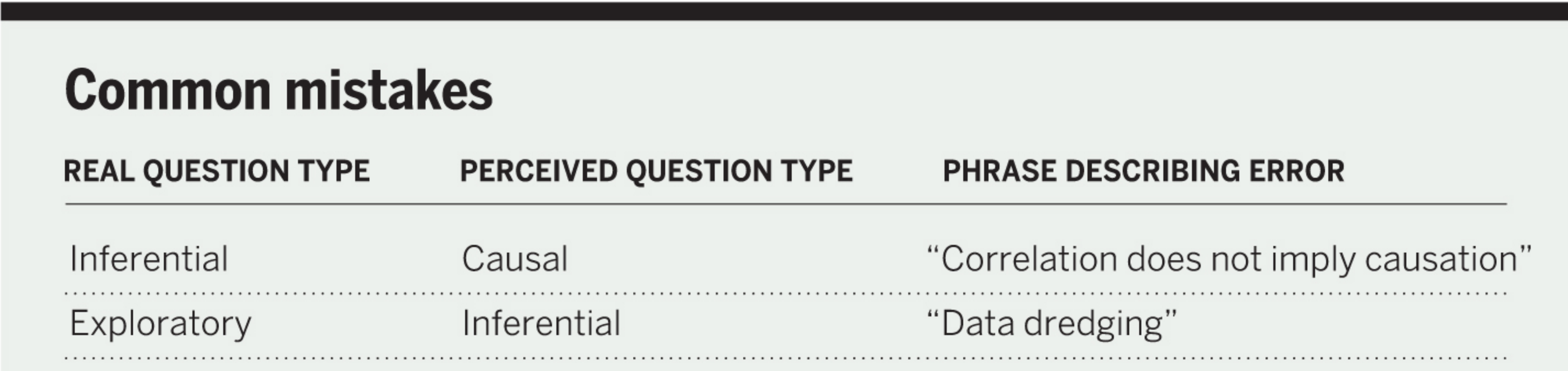
Inferential

Predictive

Causal

Mechanistic

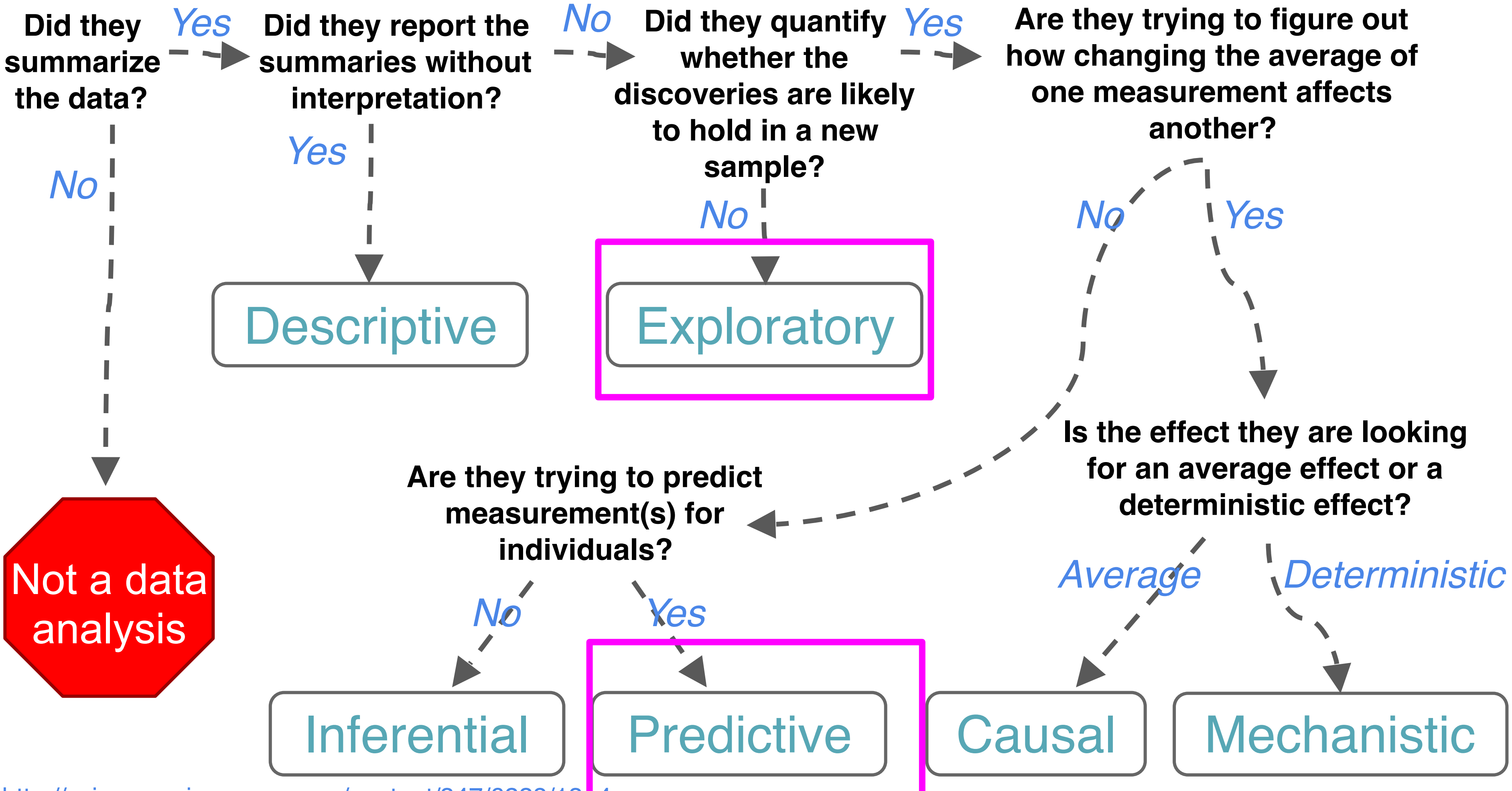
http://science.sciencemag.org/content/347/6228/1314



Common mistakes

REAL QUESTION TYPE	PERCEIVED QUESTION TYPE	PHRASE DESCRIBING ERROR
Inferential	Causal	“Correlation does not imply causation”
Exploratory	Inferential	“Data dredging”

http://science.sciencemag.org/content/347/6228/1314



Did they summarize the data?

Yes

Did they report the summaries without interpretation?

No

Did they quantify whether the discoveries are likely to hold in a new sample?

Yes

Are they trying to figure out how changing the average of one measurement affects another?

No

Yes

No

Not a data analysis

Descriptive

No

Exploratory

No

Yes

Is the effect they are looking for an average effect or a deterministic effect?

Are they trying to predict measurement(s) for individuals?

Average

Deterministic

No

Yes

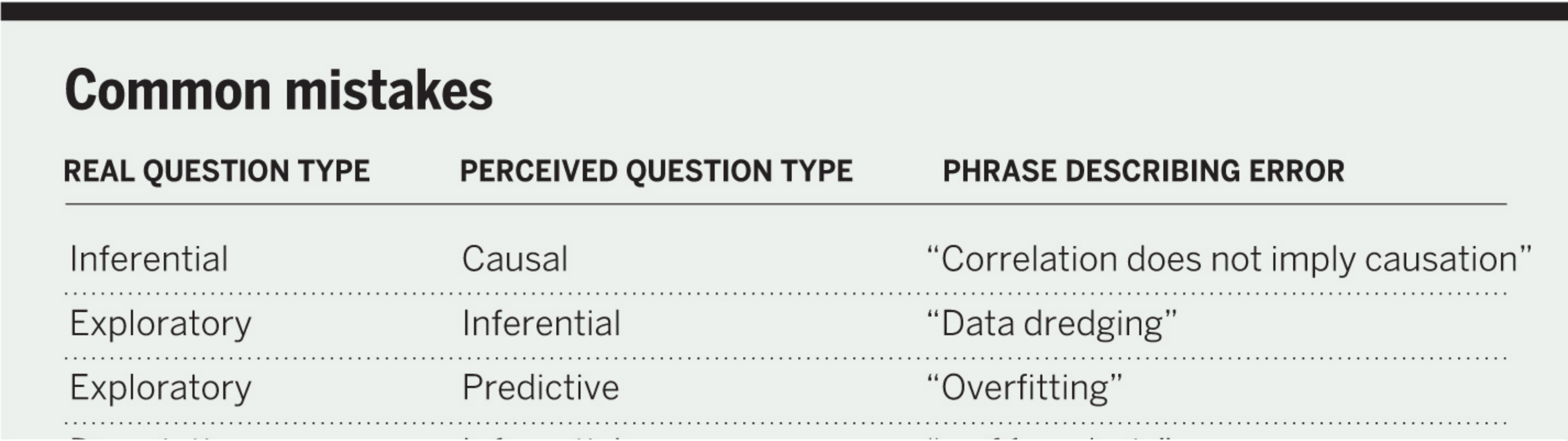
Inferential

Predictive

Causal

Mechanistic

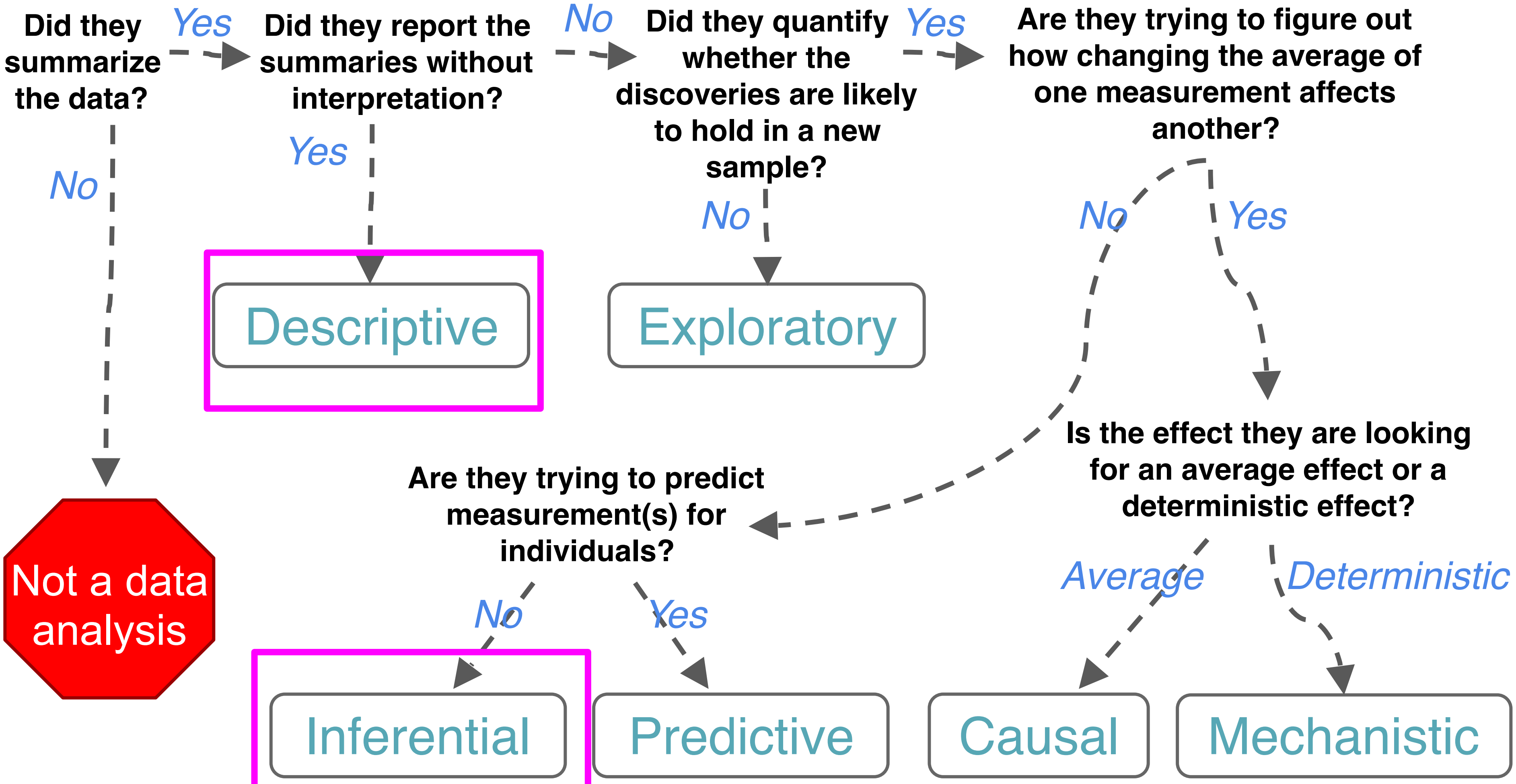
http://science.sciencemag.org/content/347/6228/1314



Common mistakes

REAL QUESTION TYPE	PERCEIVED QUESTION TYPE	PHRASE DESCRIBING ERROR
Inferential	Causal	"Correlation does not imply causation"
Exploratory	Inferential	"Data dredging"
Exploratory	Predictive	"Overfitting"

http://science.sciencemag.org/content/347/6228/1314



Did they summarize the data?

Did they report the summaries without interpretation?

Did they quantify whether the discoveries are likely to hold in a new sample?

Are they trying to figure out how changing the average of one measurement affects another?

Descriptive

Exploratory

Is the effect they are looking for an average effect or a deterministic effect?

Are they trying to predict measurement(s) for individuals?

Not a data analysis

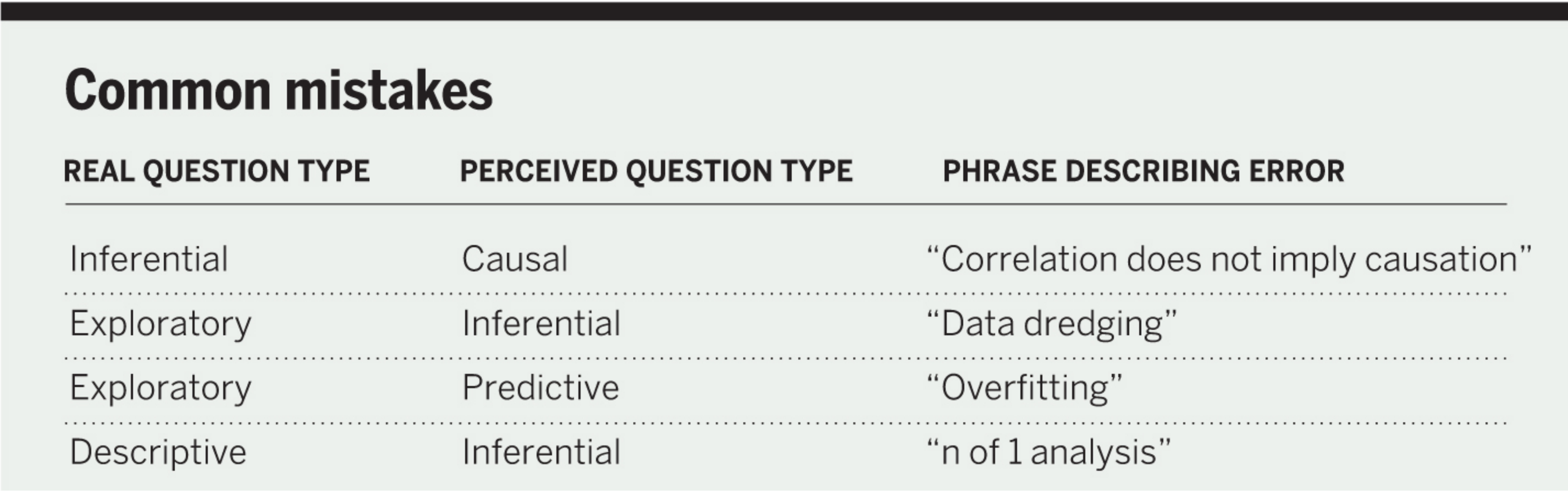
Inferential

Predictive

Causal

Mechanistic

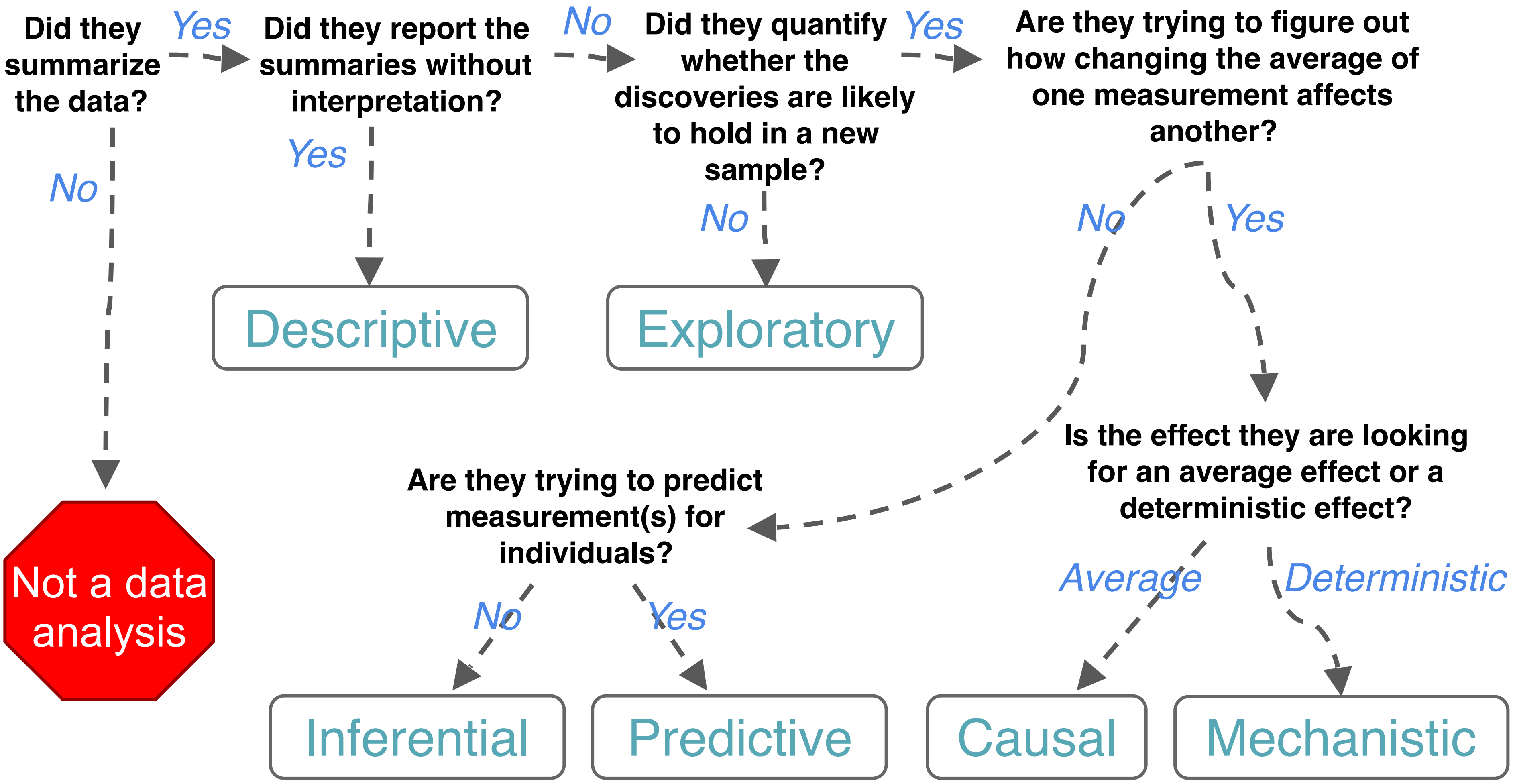
http://science.sciencemag.org/content/347/6228/1314



Common mistakes

REAL QUESTION TYPE	PERCEIVED QUESTION TYPE	PHRASE DESCRIBING ERROR
Inferential	Causal	“Correlation does not imply causation”
Exploratory	Inferential	“Data dredging”
Exploratory	Predictive	“Overfitting”
Descriptive	Inferential	“n of 1 analysis”

http://science.sciencemag.org/content/347/6228/1314



Did they summarize the data?

Yes

Did they report the summaries without interpretation?

No

Did they quantify whether the discoveries are likely to hold in a new sample?

Yes

Are they trying to figure out how changing the average of one measurement affects another?

No

Yes

Descriptive

No

Exploratory

No

Yes

Is the effect they are looking for an average effect or a deterministic effect?

Are they trying to predict measurement(s) for individuals?

Not a data analysis

No

Yes

Average

Deterministic

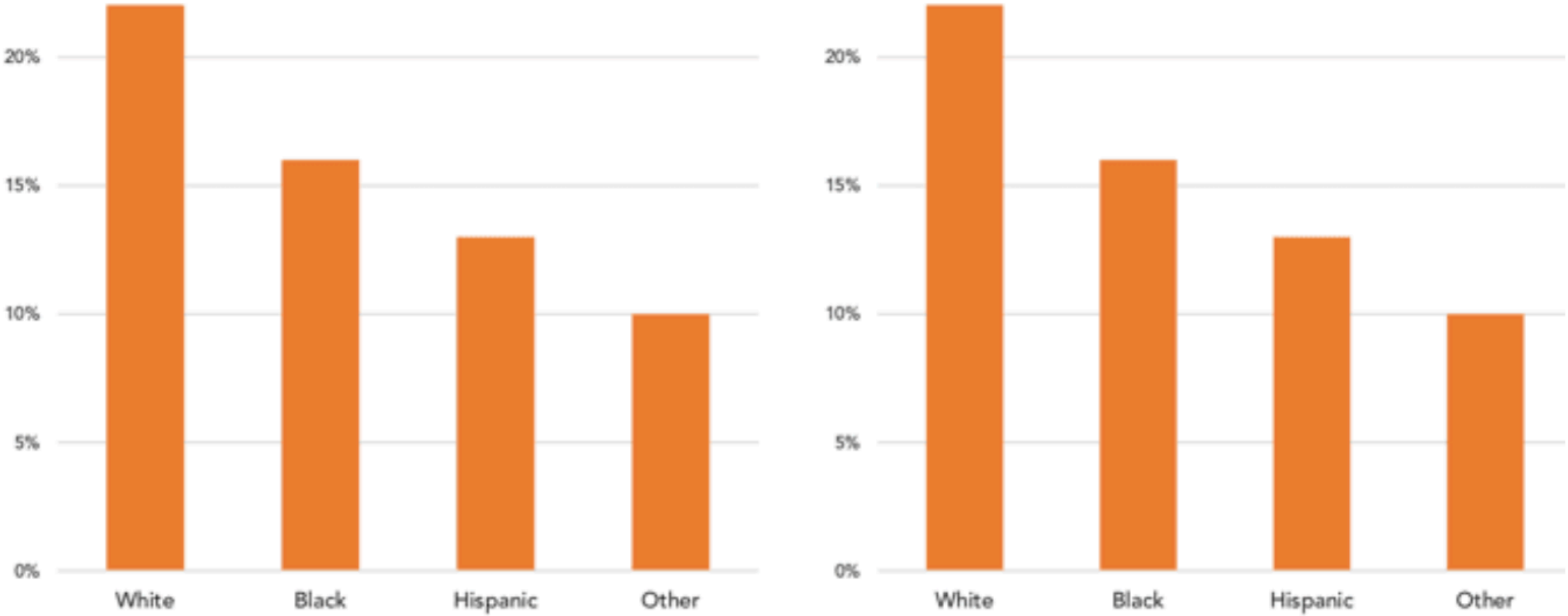
Inferential

Predictive

Causal

Mechanistic

http://science.sciencemag.org/content/347/6228/1314



Two portrayals of the same data. The data is from a study of first-time inmates in NYC jails between 2011-2013 by Fatos Kaba et al. 2015 titled “Disparities in Mental Health Referral and Diagnosis in the New York City Jail Mental Health Service.” American Journal of Public Health 105 (9). American Public Health Association: 1911–16. doi:10.2105/AJPH.2015.302699.
Credit: Graphics by Catherine D'Ignazio
Source: Catherine D’Ignazio

Question Compatibility/Specificity

- Questions can range from vague —> specific
- More specific questions require more specific data
- Balance question specificity with available data
- Some problems can be solved with assumptions

Do We Have the Right Question?

- Too **vague** —> an unwieldy analysis
    - “Does the environment affect health?”
- Too **specific** —> we don’t have that particular kind of data
    - “Does chromium-6 exposure increase doctor’s visits amongst children aged 5 to 8 in Baltimore City?”
- Does not lead to a **decision** or **intervention** —> Relevance?
    - “Does living near a highway increase your exposure to air pollution?”

Question Compatibility

Which is best?
Which is feasible?

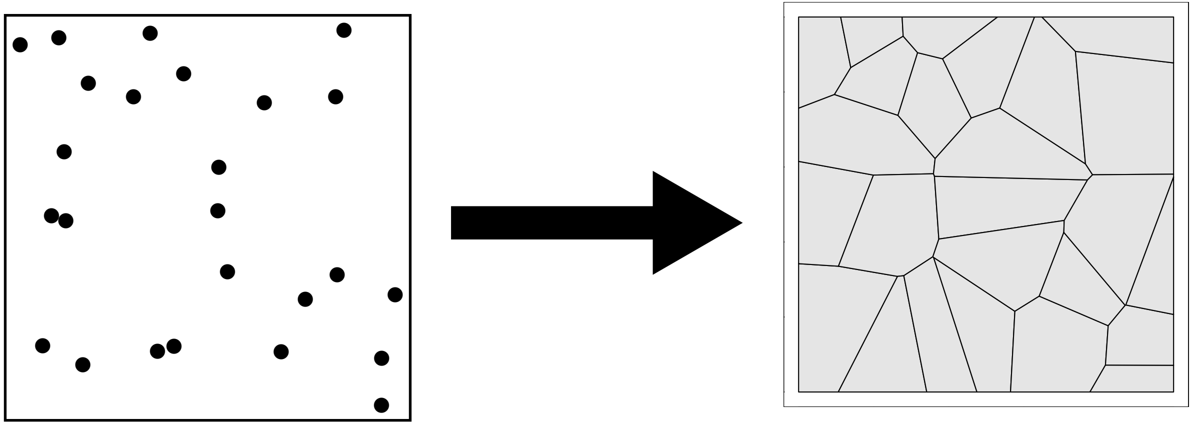
"Does chromium-6 exposure increase doctor's visits amongst children aged 5 to 8 in Baltimore City?"

"Does particulate matter (w/chromium constituents) exposure increase hospital admissions all children in Baltimore City?"

"Does outdoor ozone exposure increase hospital admissions amongst adults in Maryland?"

What considerations should be made?

Example: Change of Support



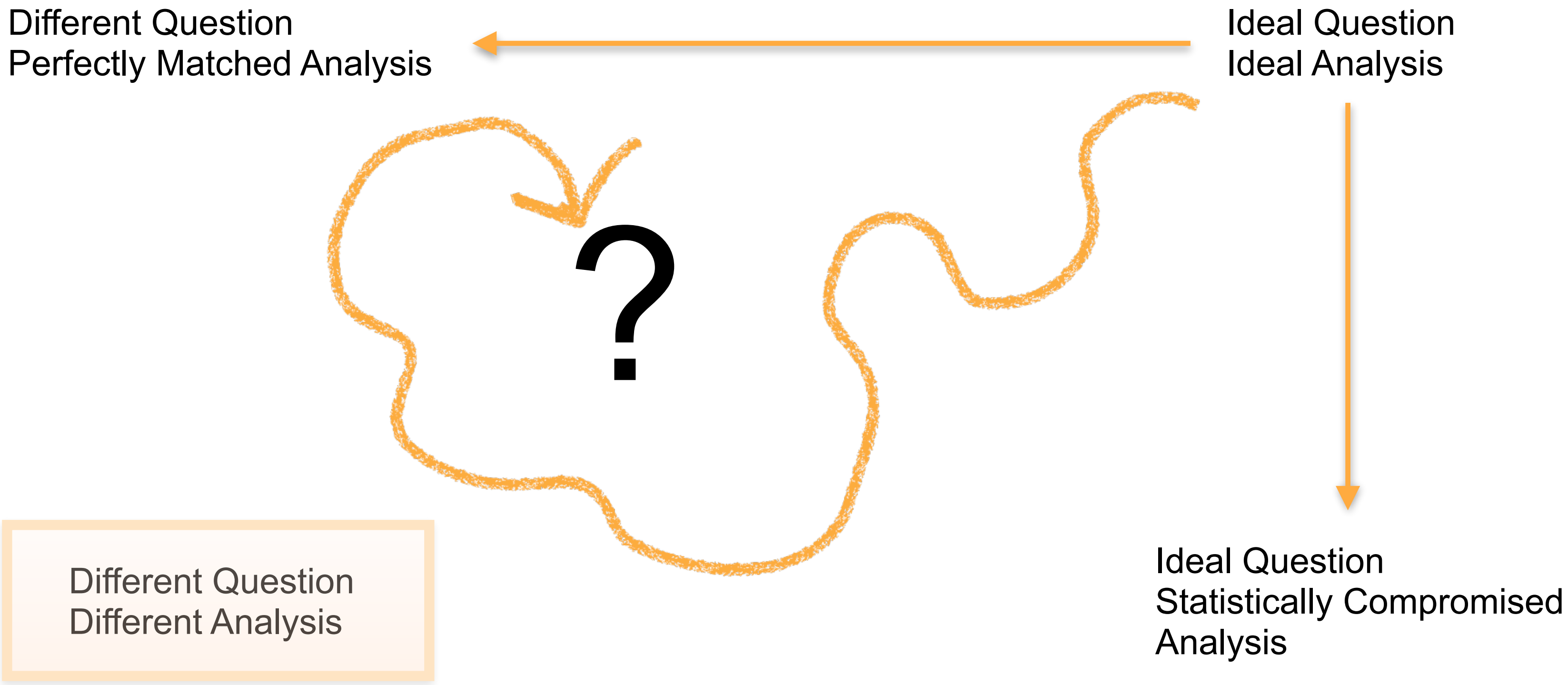
$x(s_i)$ $\Phi(A)$

Possible Solutions

1. Assume points are representative of areas
2. Develop a model to predict a unobserved locations
3. Collect more data
4. Don’t do analysis

What would determine which choice you make?

Question/Analysis Matching



Different Question
Perfectly Matched Analysis

Ideal Question
Ideal Analysis

?

Ideal Question
Statistically Compromised Analysis

Different Question
Different Analysis

Summary

- Questions can vary and formulating the question is a key task
- Navigate the question/analysis space based on a non-universal criteria
- Active engagement at the early stages is needed
- Improperly handling the question can lead to challenges in analysis and interpretation