

NINE THINGS TO CONSIDER TO NOT RUIN PEOPLE'S LIVES WITH DATA SCIENCE

adapted from Thomas Donoghue

1. THE QUESTION
2. THE IMPLICATIONS
3. THE DATA
4. INFORMED CONSENT
5. PRIVACY
6. EVALUATION
7. ANALYSIS
8. TRANSPARENCY & APPEAL
9. CONTINUOUS MONITORING

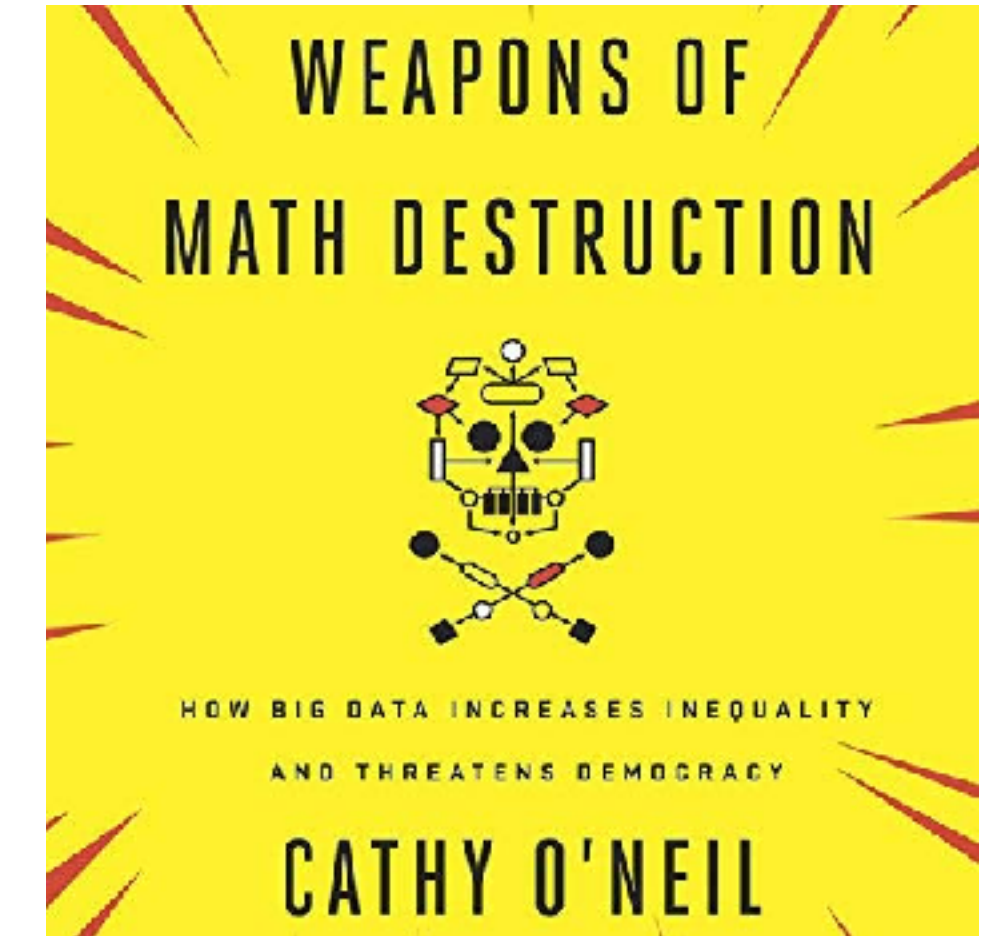
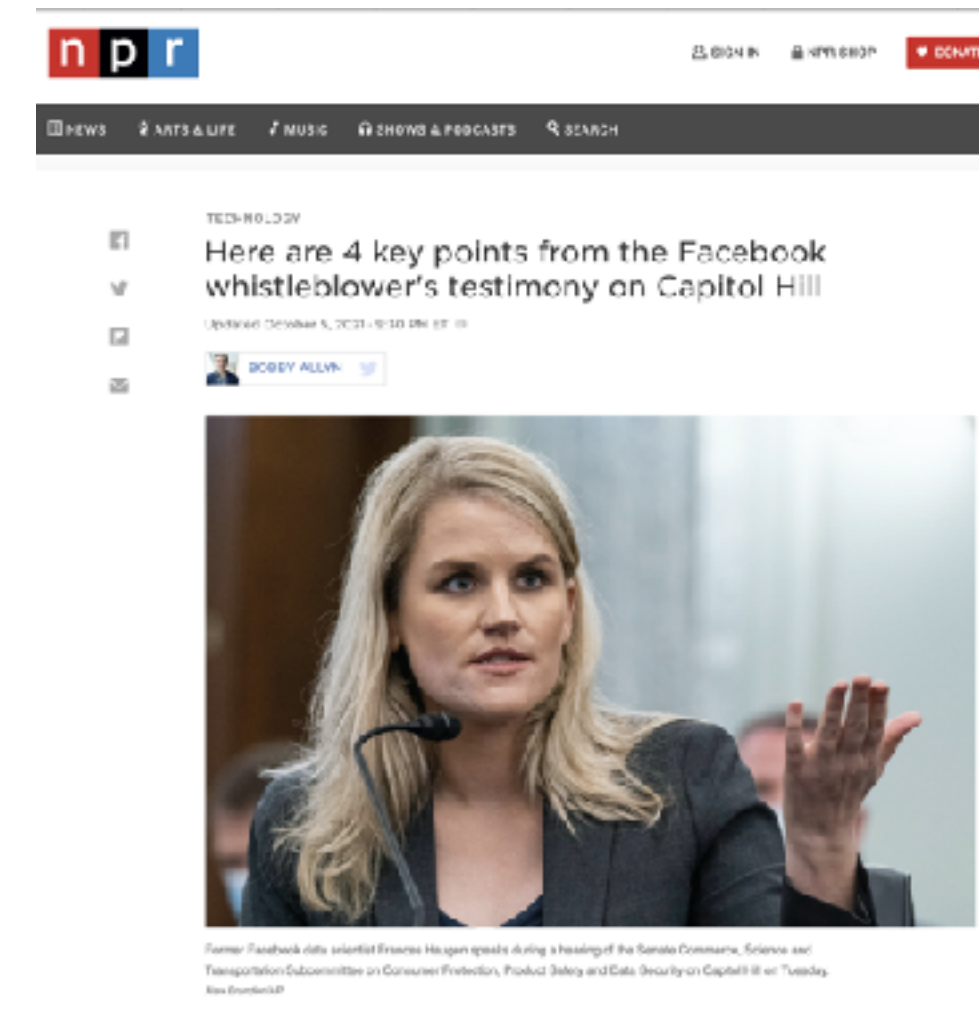
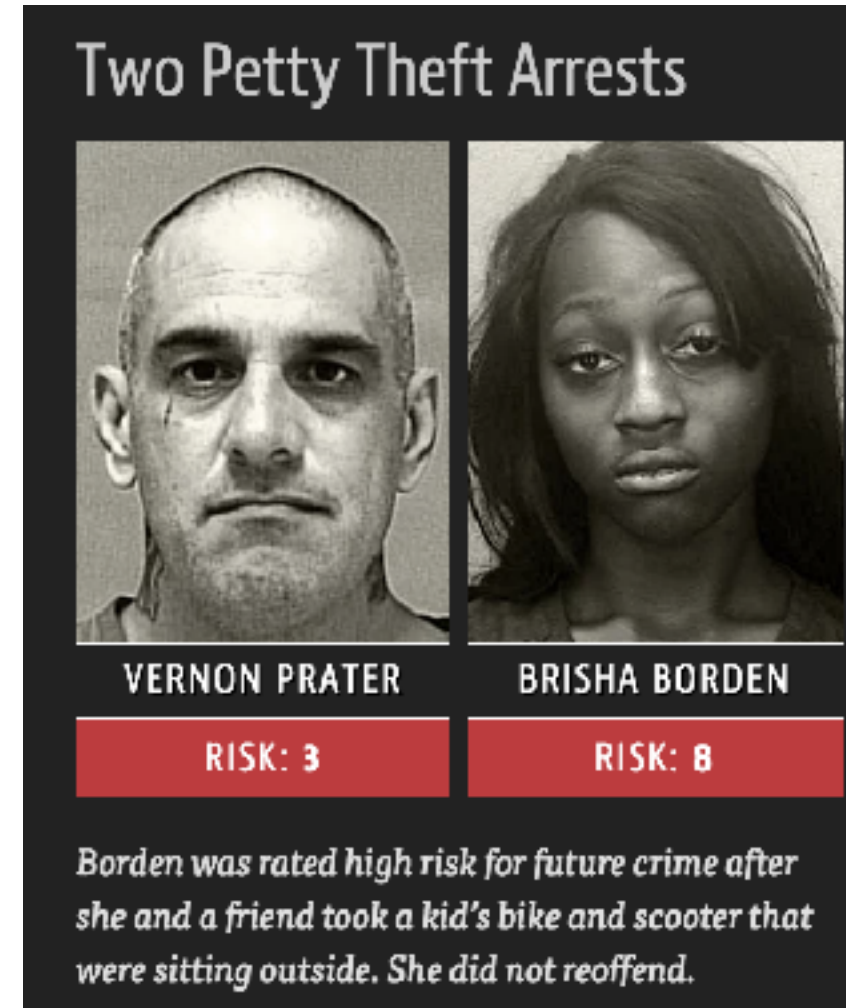


(a) Three samples in criminal ID photo set S_c .



(b) Three samples in non-criminal ID photo set S_n .

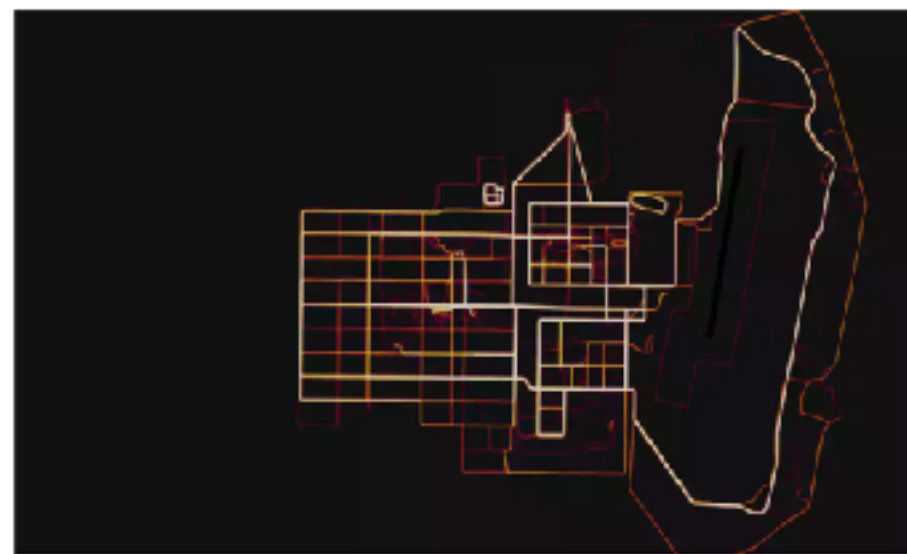
Figure 1. Sample ID photos in our data set.



Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

● Latest: Strava suggests military users 'opt out' of heatmap as row deepens



▲ A military base in Helmand Province, Afghanistan with route taken by joggers highlighted by Strava. Photograph: Strava Heatmap

@AncientMayra
"a raccoon astronaut with the cosmos reflecting on the glass of his helmet dreaming of the stars"
@OpenAI DALL-E 2



Prompt: ceo;
Date: April 6, 2022



Prompt: a photo of a personal assistant;
Date: April 1, 2022



ON PERPETUATING INEQUALITY

- Data & Algorithms can & will entrench social disparities
- Errors and bias typically target the disenfranchised
- The combination of damage, scale, and opacity can be incredibly destructive
- They can introduce feedback in such a way as to enact self-fulfilling prophecies

PUTTING IT ALL TOGETHER (GOOD)

- well-posed question that you know something about
- have considered implications of work
- adequate data, covering population of interest, with known and manageable biases
- allowed to use the data
- have de-identified data, stored securely
- defined metrics for success, objectively measured
- if suggesting causality, have actually established causality
- model is understandable, has procedure for appeal
- will monitor system for changes, have way & plan to update

HOW TO BE BAD WITH DATA SCIENCE

- ill-posed question you know nothing about
- don't consider implications
- haphazardly collected biased data
- didn't check or are not allowed to use data for this purpose
- un-anonymized, identifiable data, stored insecurely
- no clear metric for success (meh, it 'seems to work')
- present spurious correlations as meaningful
- model is a black box, no method for appeal in place
- no monitoring, no way to identify biases or update model

Data science questions

Jason G. Fleischer, Ph.D.

Asst. Teaching Professor

Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu



@jasongfleischer

<https://jgfleischer.com>

Today's Learning Objectives:

Explain the data science process

Demonstrate ability to move from a
general question to a data science
question

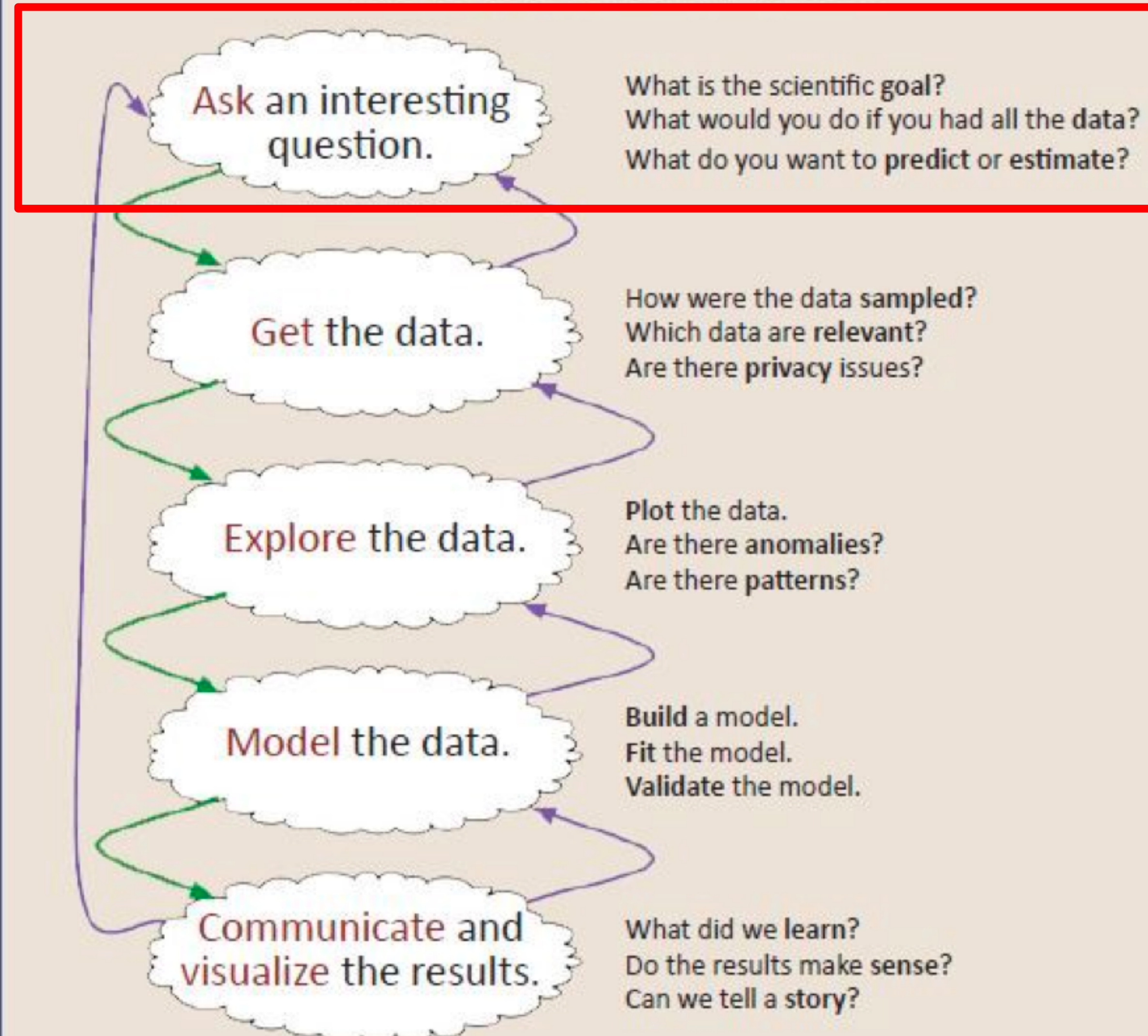
Nature of a data scientist

- data-driven.
- care about answers. They analyze data to discover something about how the world works.
- care about whether the results make sense, because they care about what the answers mean.
- are comfortable with the idea that data have errors.
- know nothing is ever completely true or false in science, while everything is either true or false in computer science or mathematics.

Nature of a GREAT data scientist

- Conscientious, works using proven and understood methods, triple checks things
- Yet is open to new methods and creative at finding solutions (just checks them thoroughly!)
- Methodical
- Yet after working down in the details, takes a step back and questions the big picture

The Data Science Process



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://www.cs109.org/>

If I had an hour to solve a problem and my life depended on it, I would use the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than five minutes. —Einstein

Data Science questions should...

- Be specific
- Be answerable with data
- Specify what's being measured



What makes a question a good question?

Specifying what you're going to measure is important

Examples of poor questions that leave wiggle room for useless answers:

- What can my data tell me about my business?
- What should I do?
- How can I increase my profits?

Examples of good questions where the answer is impossible to avoid:

- How many Model 3s will Tesla sell in San Diego during the third quarter?
- How many students will apply for admission to UCSD in 2030?
- How many students should UCSD admit in 2030 for a target class size of 50,000?

Working toward a strong
data science question

Nailing down the right question: politics

Too-vague question: What impacts politics in America?

Improving: Does pop culture have an impact on American politics?

... Do American TV shows have an impact on American politics?

... Does South Park affect American politics?

... Is there a relationship between words in South Park episodes and American politics?

... Is there a relationship between the sentiment of political words in South Park and American politics?

... Is there a relationship between the sentiment of political words in South Park and America's presidential approval rating?

Nailing down the right question: cause of death

Too-vague question: What gets attention in the news?

Improving: Do terrorist attacks get reported too much?

... Is there a relationship between the number of people who die relative to the amount of media attention a story gets?

... What causes of death are over reported in the news relative to CDC death data? Underreported?

... Is there a relationship over time between cause of death terms in the *NYT*, The Guardian, and Google trends data relative to data from the CDC?

*do you think asking the question above would give different results on data up to 2019 vs a dataset that includes 2020?

Nailing down the right question: policing

Too-vague question: Why isn't police response time always the same?

Improving: How can we improve police response time?

... Do crime levels and time of day affect response time?

... Where should police cars be stationed, accounting for crime levels and time of day, to make police response times equitable?

... Where should police cars be stationed, accounting for crime levels and time of day, to make police response times equitable throughout San Diego?

Nailing down the right question: housing costs

Too-vague question: Why are housing costs so high in San Diego

Improving:

-

Nailing down the right question: environment

Too-vague question: What did the pandemic change about our environmental problems?

Improving:



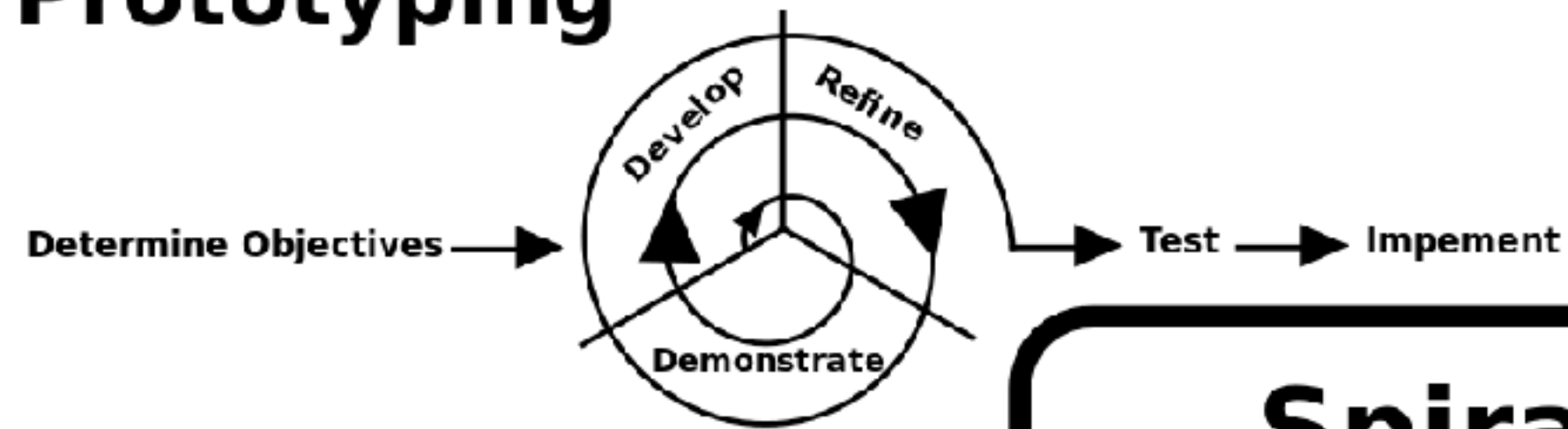
I don't need to define a question... the boss/customer gives me the question!



Software engineering methods

Metaphor and tool for data science projects

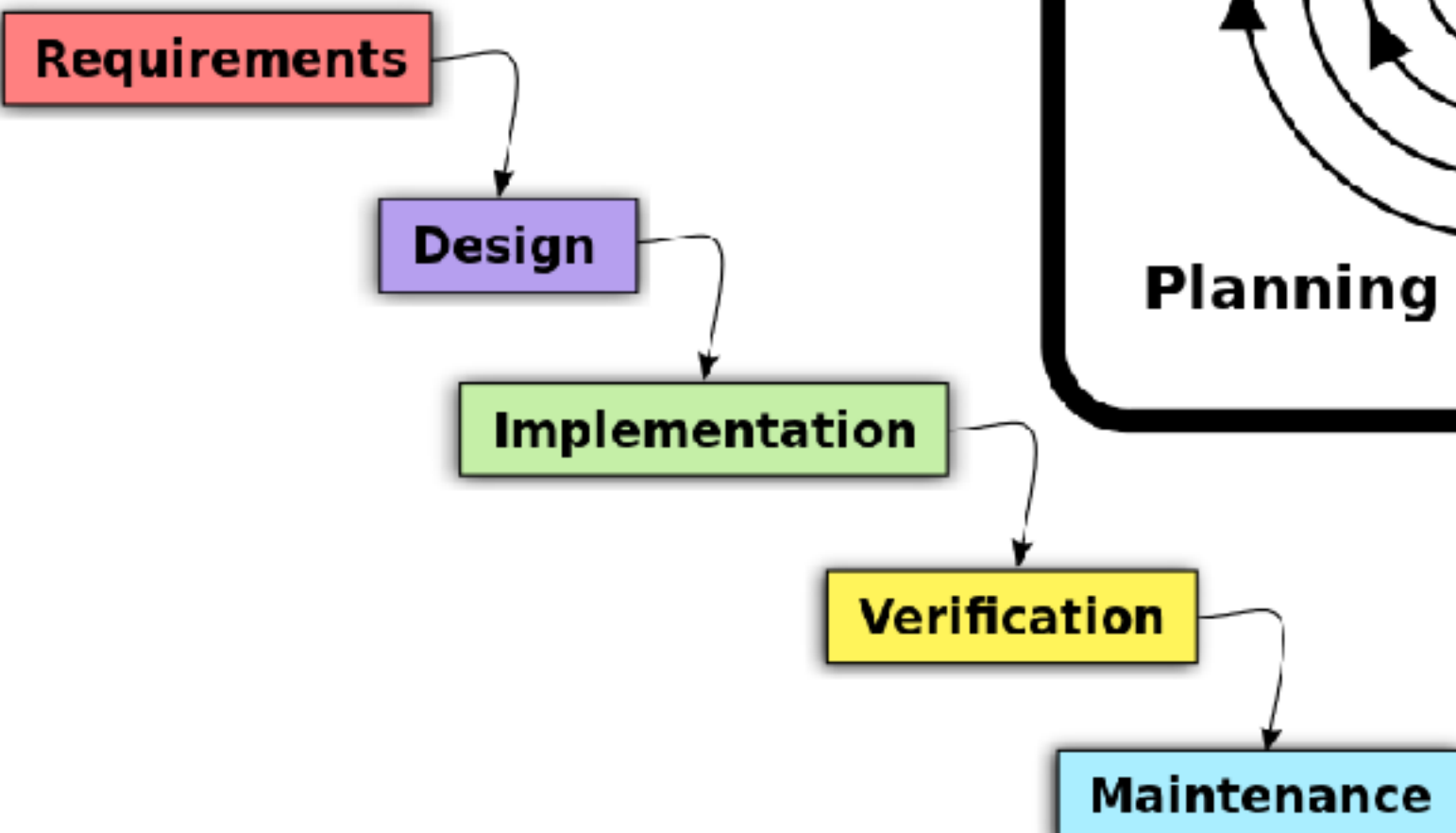
Prototyping



Spiral



Waterfall



What happens next?

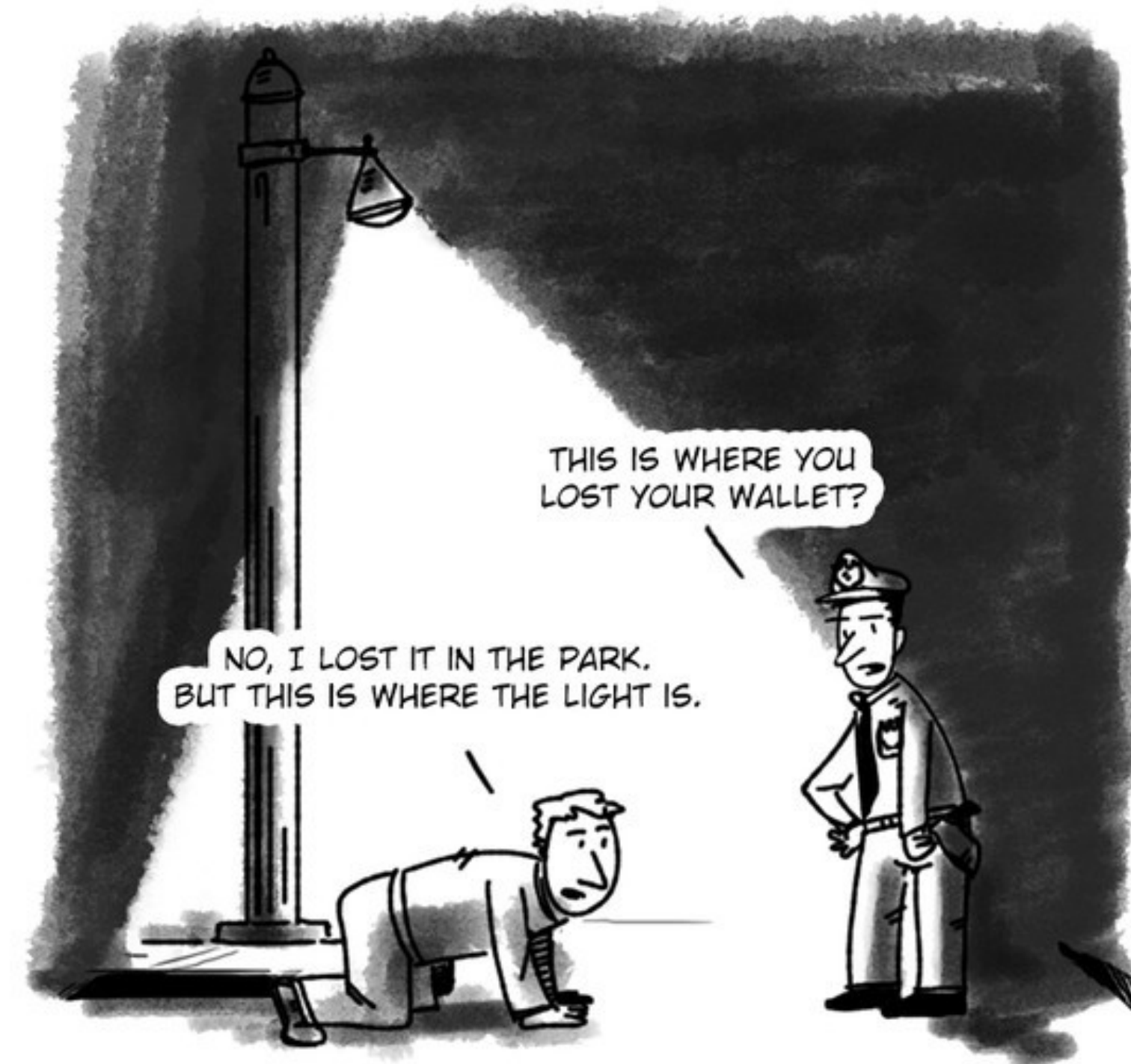
After the question is defined, should it become a project?

- What are the constraints?
- What are the resources available?
- What are the sure costs and benefits?
- What are the potential risks and rewards? (Includes ethical!)
- Can we define a metric to determine the success of the project?

Unanswerable questions worth asking

A well-spec'd question can still be unanswerable

- Often only bits and pieces of the data puzzle are available, options are:
 - Guide the project to (GOOD!) questions that can be answered with the data available
 - Create a new project to gather the data to answer the question (opportunity!)
- Raising an unanswerable can change how people think and react



"The streetlight effect"