Data science questions

Jason G. Fleischer, Ph.D.
Asst. Teaching Professor
Department of Cognitive Science, UC San Diego
jfleischer@ucsd.edu


@jasongfleischer

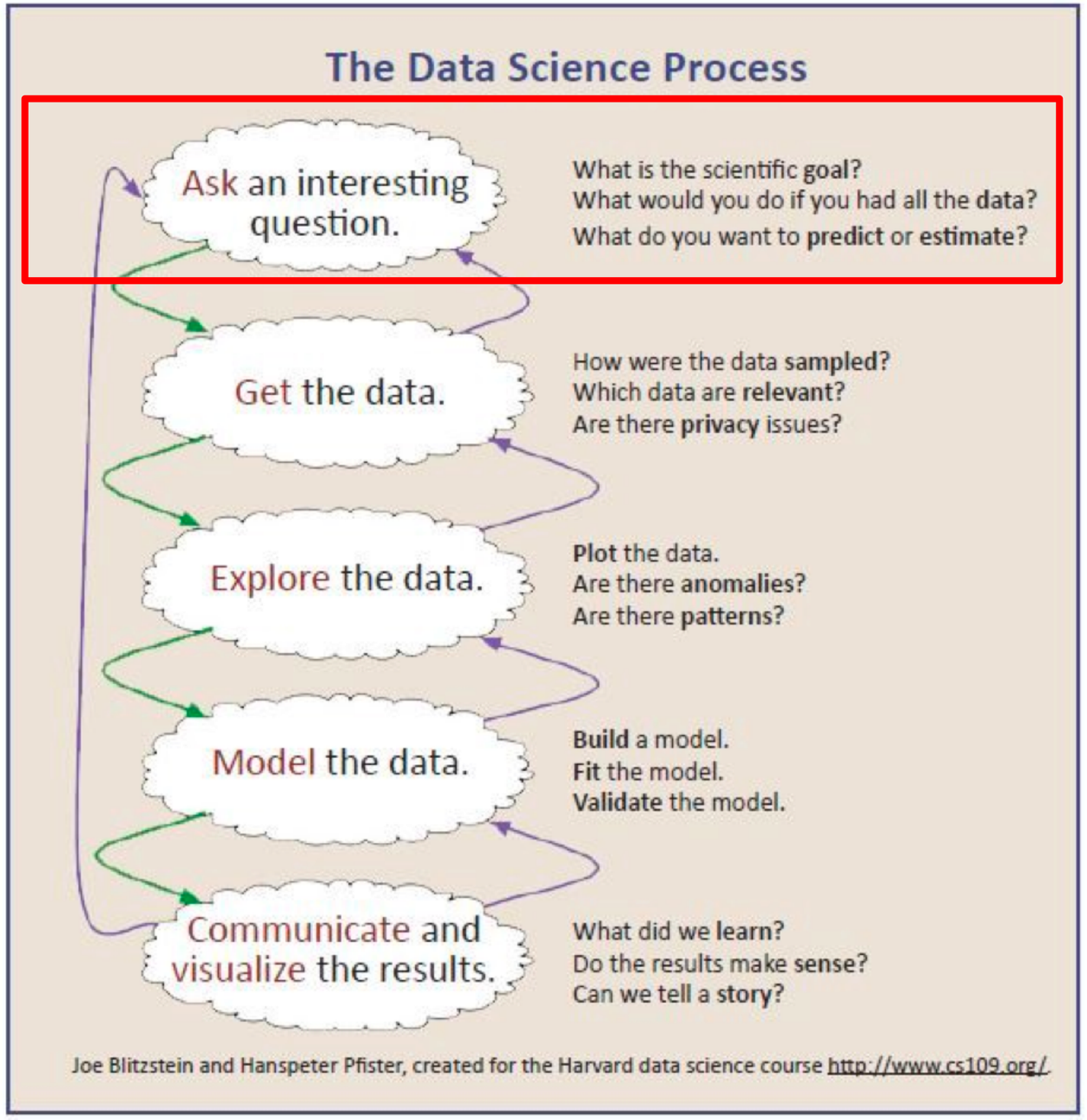
https://jgfleischer.com

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.



adapted from Chris Keown

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?

https://github.com/COGS108/FinalProjects-Wi18/blob/master/010-FinalProject.ipynb

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?

https://github.com/COGS108/FinalProjects-Sp17/blob/master/029-FinalProject.ipynb

Nailing down the right question: housing costs

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

Improving:

•

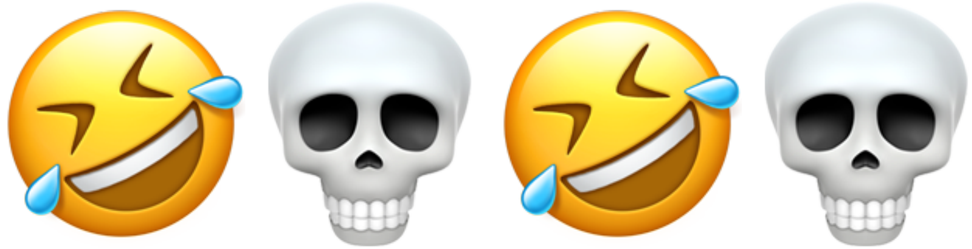
Nailing down the right question: environment

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

Improving:

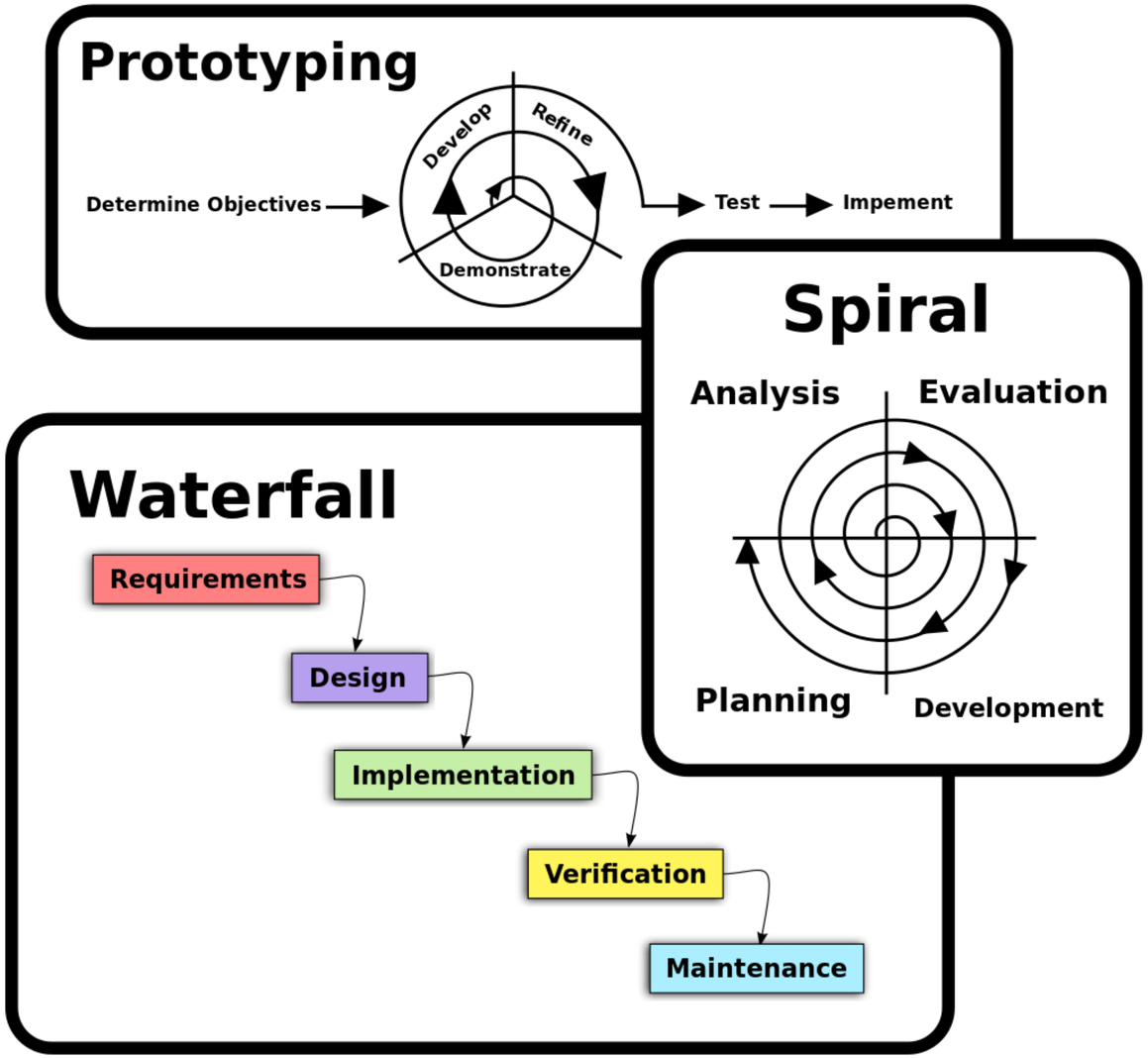
-

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



What happens next?

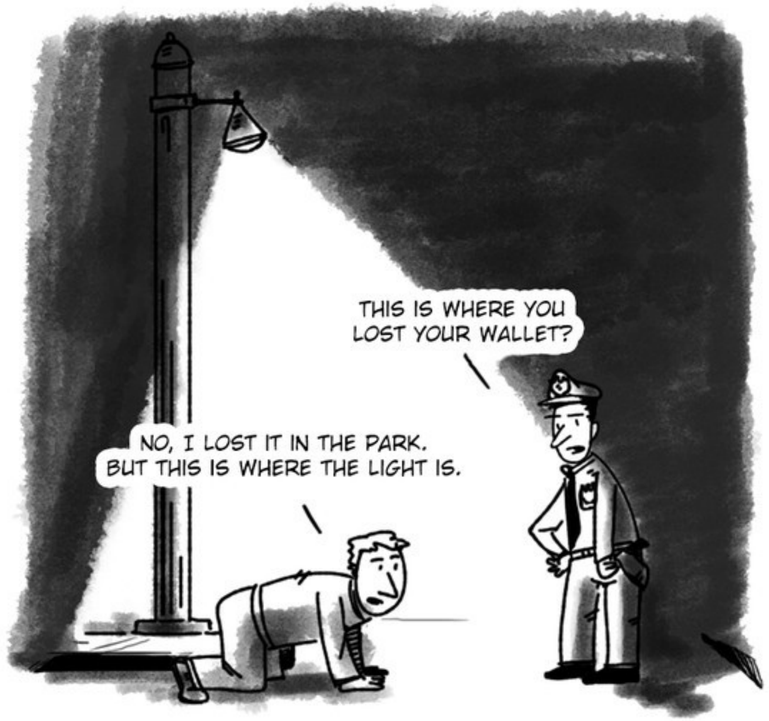
After the question is defined, should it become a project?

- 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”