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

Course Reminders

- Due Today
  - Q2
  - Github ID**
  - Group signup** we will be doing this on Tuesday at some point so you may have an extra 24 hours but get it done ASAP
- Due Wednesday
  - A1
- Due Friday
  - D2

Today's Learning Objectives:

- How to think about the data science process
- Demonstrate ability to move from a general question to a data science question

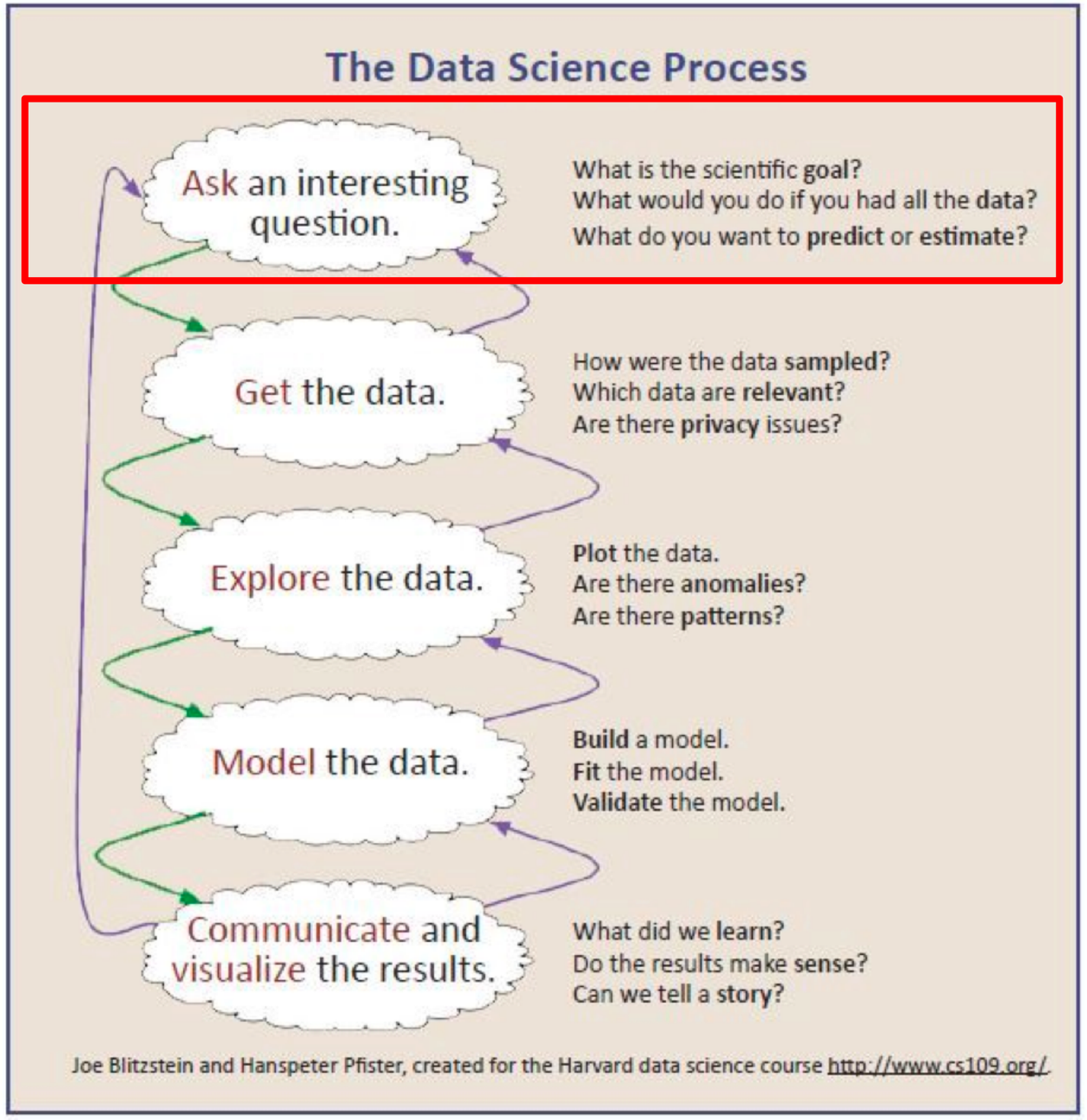
Nature of a data scientist

- data-driven
- care about answers
- analyze data to discover something about how the world works
- know that each analysis is just a different viewpoint, trying to make sense of a complex whole that can’t easily be perceived
- care about whether the results make sense, because they care about it means
- are comfortable with the idea that data have errors
- are comfortable with the idea that there’s more than one way to analyze the same
- know nothing is ever completely true or false in science
- know that you can still learn something and make decisions in spite of these uncertainties
- cares about communicating these subtleties as well as the results themselves

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



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…

- answerable with data it is possible to collect
- unambiguous in meaning
- specifically describing exactly what data/metrics/analysis are required to answer the question
- big enough to be interesting, small enough to be accomplishable

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 through 2021?

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

Refining a question

https://forms.gle/haxpaaVJcjAQ9Yy86

Here is a much too vague question:

What racial disparities exist in policing?

Let’s refine it. Which of the following would be the best possible data science question?

Nailing down the right question: student success (a skit)

Too-vague question: Who does well at university?

Improving:

•

Your turn to nail down the right question: $an Diego Co$t$ (think for 3 min, groups of 2-4 for 5 min, share for 5 min)

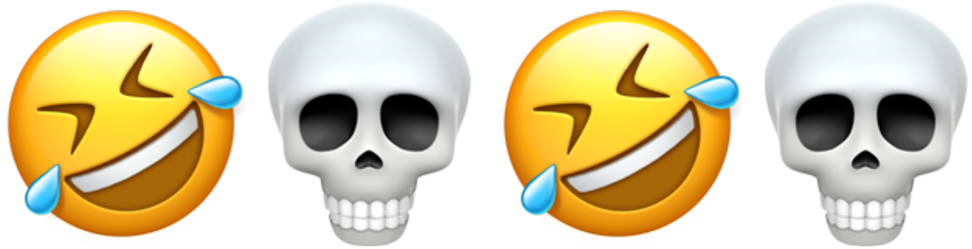
Too-vague question: Why is it so expensive to live in San Diego?

•

Are these good enough questions? What could be improved?

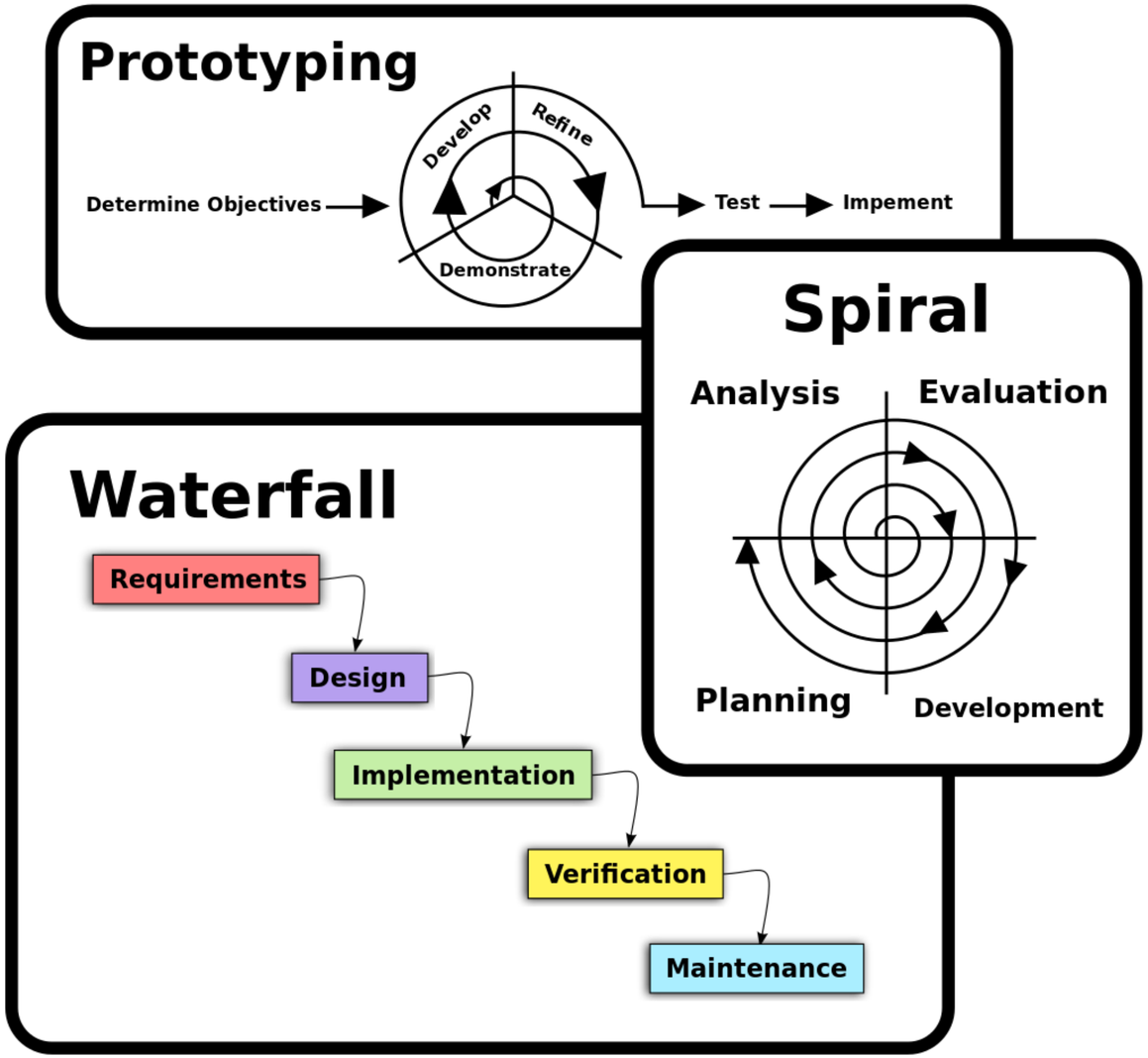
- Does one US political party have a tendency to disproportionately use more negative sentiment on Twitter than the other and if so, what motivates this?
- Is the city's population, game day weather conditions, and this season's win rate sufficient to predict attendance at a professional sports match?
- What effects do demographic factors such as age structure, median household income and racial diversity have in influencing pet (cat and dog) adoption rates per state during 2019?

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?

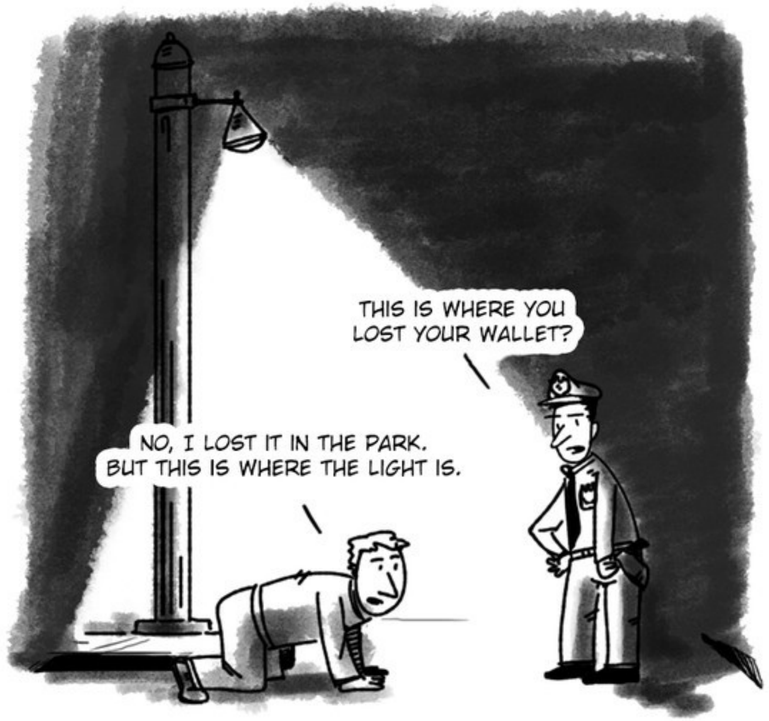
After the question is defined, should it become a project?

- What are the constraints?
- What are the resources available?
- IS THE NECESSARY DATA GETTABLE??
- 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”