Plan:

1. Discuss the properties of a good data science question
2. Walk through examples of data science question formation

Data Science Questions

Shannon E. Ellis, Ph.D
UC San Diego

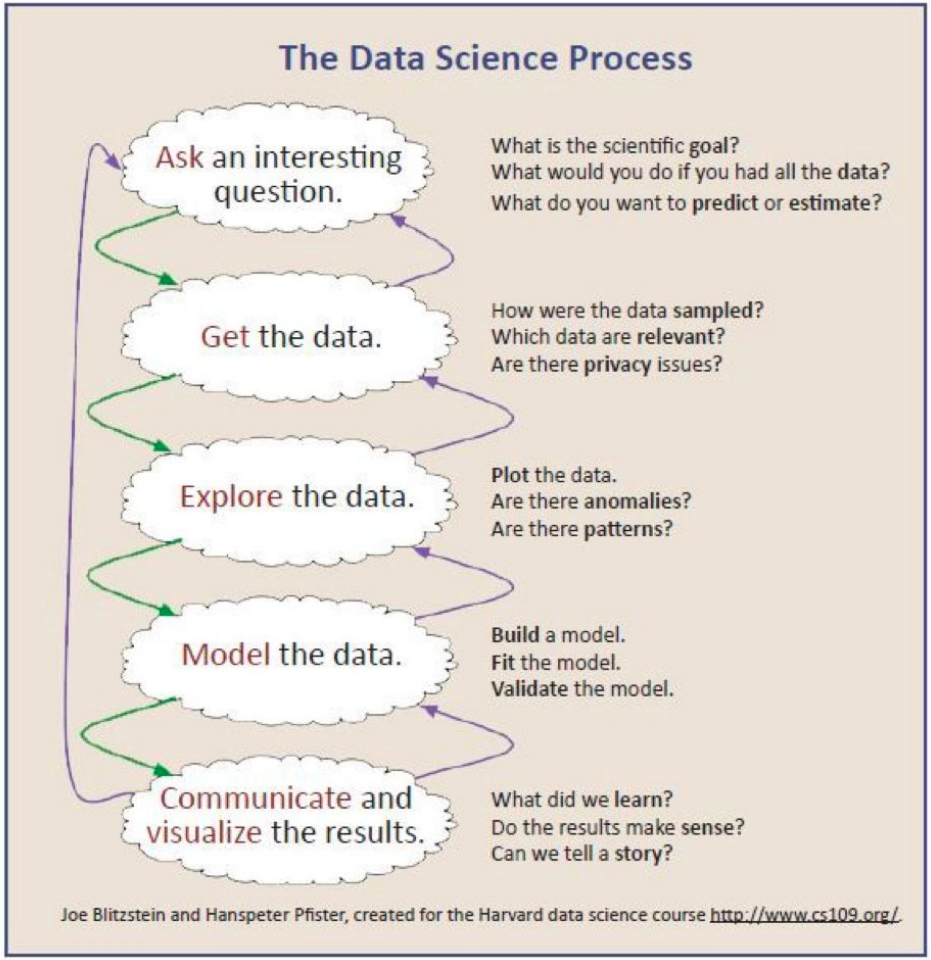
Department of Cognitive Science
sellis@ucsd.edu

Formulating Data Science Questions

When you and your group sit down to figure out what you're going to do for your final project in this class, you'll have to formulate a strong question - one that is specific, can be answered with data, and makes clear what exactly is being measured.

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.


The Data Science Process
Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?
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?

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?

https://github.com/COGS108/FinalProjects-Wi18/blob/master/071-FinalProject.ipynb

Nailing down the right question: flight delays

Too-vague question: Why are the flights I take never on time?

Improving: What causes delays during travel?

...Do certain airports have more flight delays than others?

...Does the likelihood of a flight delay depend upon where the flight starts or ends?

...Does the likelihood of a flight delay, cancellation, or diversion depend upon where the flight starts or ends?

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?

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