Correlation

definition

- If $r = -1$, then there is a perfect negative linear relationship between x and y.
- If $r = 1$, then there is a perfect positive linear relationship between x and y.
- If $r = 0$, then there is no linear relationship between x and y.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Data Science Questions

C. Alex Simpkins Jr., Ph.D
UC San Diego, RDPRobotics LLC

Department of Cognitive Science
rdprobotics@gmail.com
csimpkinsjr@ucsd.edu

Lectures : https://github.com/COGS108/Lectures-Wi23

Today's learning objectives:

- Explain the data science process
- Demonstrate ability to move from a general question to a data science question

Formulating Data Science Questions

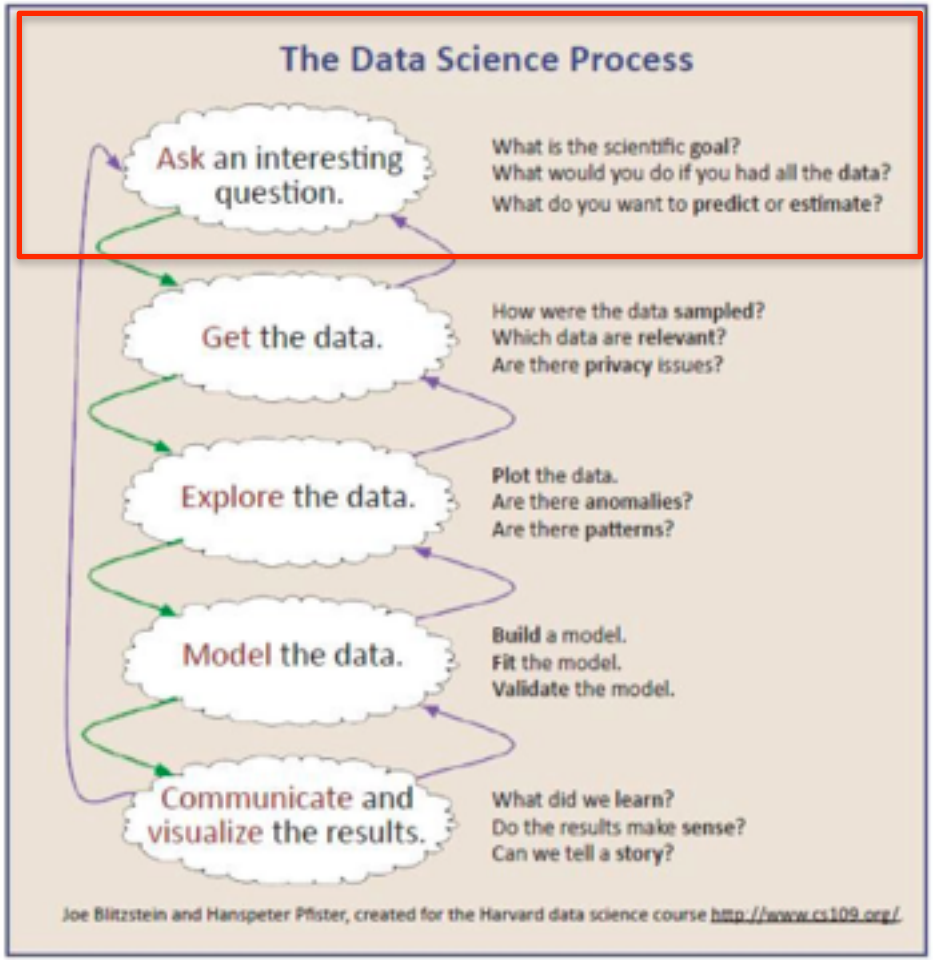
When you and your group sit down to figure out what you're going to do for your final project in this class, you'll have to formulate a strong question - one that is specific, can be answered with data, and makes clear what exactly is being measured.

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.

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

Hypothesis testing

-Cannot prove hypothesis

-Can only reject or fail to reject null hypothesis

-Why?

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: politics & the economy

Too-vague question: Does the President affect the stock market?

Improving: Does the political party of the US President affect the stock market?

...Does the political party of the US President affect the major stock market indexes?

...Does the political party of the US President affect the growth of major stock market indexes?

...Is there a significant change in the growth of the major U.S. stock market indexes caused by the political party of a United State's President? Is there a significant difference in the percent change in stock prices between Democratic presidential terms and Republican presidential terms?

Nailing down the right question: education

Too-vague question: How has COVID-19 impacted students?

Improving: How has COVID-19 impacted university students' education?

… Do students' grades and how they rate their classes differ pre- and during remote learning, due to COVID-19?

… At UCSD, is there a difference between students' grades and how they rate their classes before COVID-19 and during remote learning, due to COVID-19?

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?

Nailing down the right question: Housing

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

Improving:?

Nailing down the right question: Housing

Nailing down the right question: environment

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

Improving:?

Nailing down the right question: environment