Data Science Questions

C. Alex Simpkins Jr., Ph.D

RDPRobotics LLC,

UC San Diego, Department of Cognitive Science

rdprobotics@gmail.com

csimpkinsjr@ucsd.edu

Lectures :http://casimpkinsjr.radiantdolphinpress.com/pages/cogs108_ss1_23/index.html

Plan for this lecture

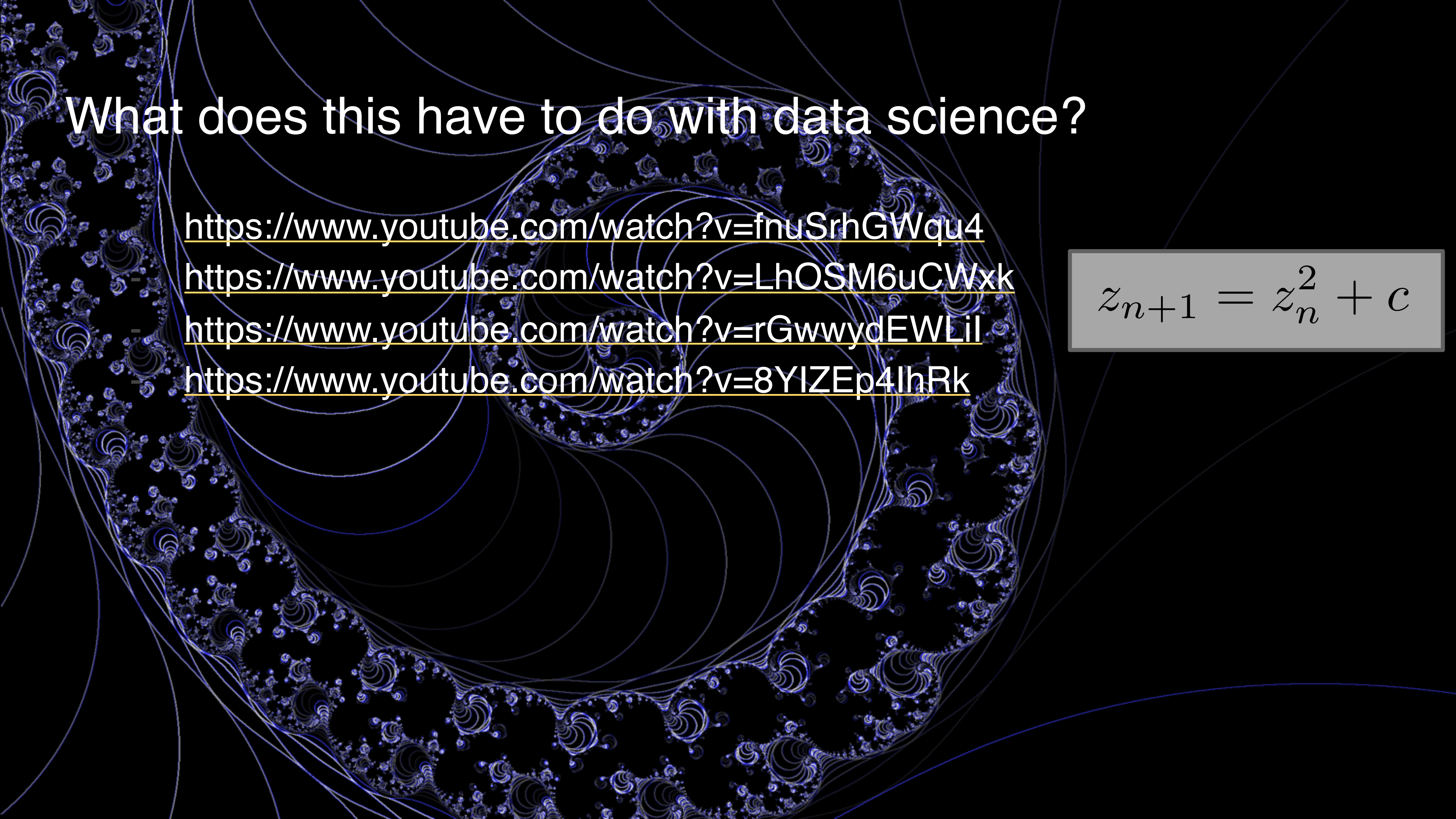
- Announcements
- Data science questions
- Hypothesis testing and the null hypothesis
- Causality

D3 topics

- Starting on exploratory data analysis
- Visualization (part 1)
- Please excuse the topics

A word of caution first…

- WARNING - if you have a tendency toward epilepsy, do not watch these videos as they contain patterns of color and light
- Otherwise enjoy!



What does this have to do with data science?

- https://www.youtube.com/watch?v=fnuSrhGWqu4
- https://www.youtube.com/watch?v=LhOSM6uCWxk
- https://www.youtube.com/watch?v=rGwwydEWLiI
- https://www.youtube.com/watch?v=8YIZEp4IhRk

$$\boxed{z_{n+1} = z_n^2 + c}$$

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,***
- ***Makes clear what exactly is being measured***.

Nature of data scientists

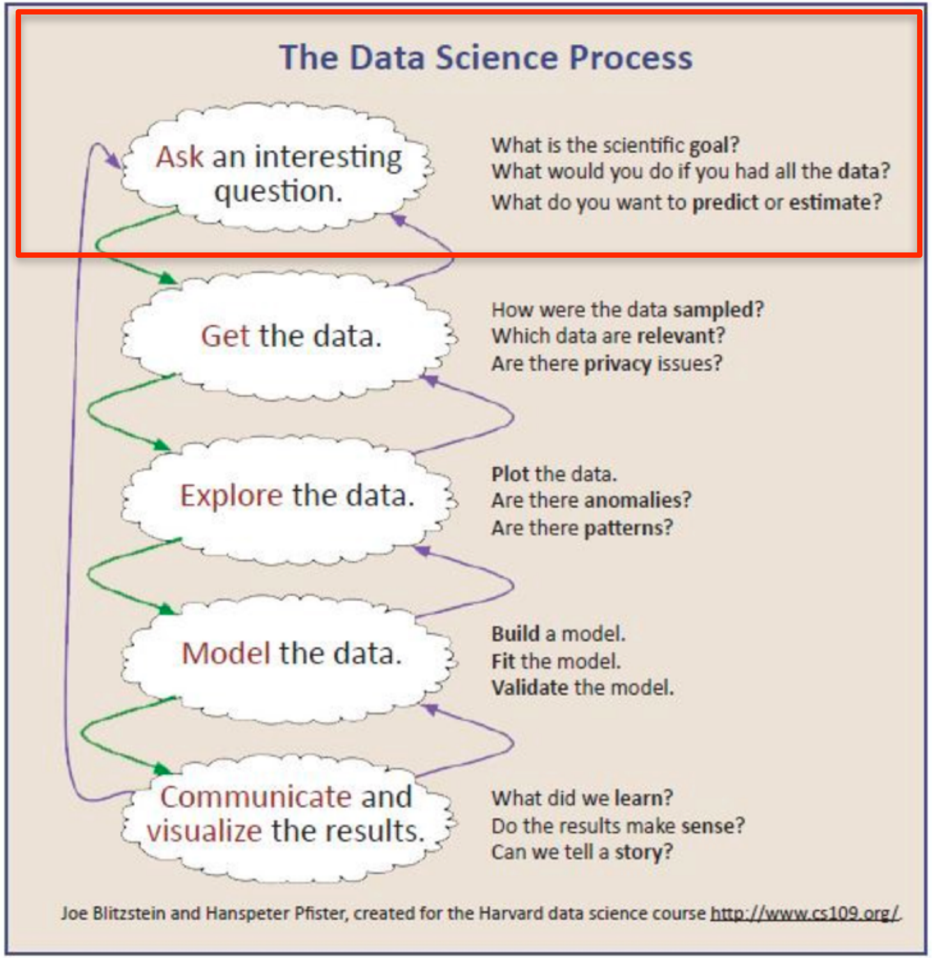
- Data-driven.
- Care about answers. They analyze data to discover something about how the world works.
- Care about whether the results make sense, because they care about what the answers mean.
- Are comfortable with the idea that data have errors.
- Know nothing is ever completely true or false in science

adapted from Chris Keown

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
  - This is the quality of a great scientist/engineer in general (ability to take in multi scale views of the world)

adapted from Chris Keown



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/.

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

adapted from Chris Keown

Hypothesis testing

- *Cannot prove hypothesis*
- *Can only reject or fail to reject null hypothesis*
- *Why?*

Hypothesis testing

- *Cannot prove hypothesis*
- *Can only reject or fail to reject null hypothesis*
- *Why?*
  - There is always the possibility that there is an underlying variable, effect correlation, connection, direction of connection etc. that might be really affecting things causally which we are not modeling
  - Un-modeled dynamics

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?

adapted from Chris Keown

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: movies & the economy

Too-vague question: Do movies affect the stock market?

Improving: Does a Marvel or a DC blockbuster movie affect the stock market significantly?

...Does a Marvel or DC studio blockbuster affect the major stock market indexes?

…Does the Movie studio blockbuster type (Marvel vs. DC) 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 dominant blockbuster action movie (Marvel or DC)? Is there a significant difference in the percent change in stock prices between a Marvel hit and a DC hit movie?

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?

https://github.com/COGS108/FinalProjects-Wi21/blob/main/FinalProject_group084.ipynb

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?

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

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

Improving:?

-Why are housing costs so high in La Jolla San Diego?

-Is there a correlation between median salary and housing costs in La Jolla?

-How is the median income of adults over 30 related to rent prices in La Jolla? (us census, transparent ca, redfin databases)

-Is there a correlation between proximity to the beach and housing prices in San Diego/La Jolla?

-How does education in San Diego affect housing costs?

-How does proximity to UCSD campus affect housing costs?

(can get comparative data by adding other cities - San Diego vs. other cities in US)

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:?

-Was there a significant change in air quality after lockdown procedures due to the pandemic were implemented

> -Was there a significant change in air quality in CA for the month of march to may 2020 due to (during?) COVID 19 lockdown procedures etc.

-How did the # of k95 masks sold on amazon since 2020 affect the polluting rates (street pollution) in SD county?

-how did the mandated quarantine affect the levels of plastic waste?

-what was the increase in AQI related to the # of remote jobs in La Jolla in 2019 vs. 2020

Nailing down the right question: environment

Refining a hypothesis

A hypothesis should be

- Narrow
- Very specific
- ***<u>Not</u>*** include a conclusion or interpretation
- Consist of a research and null hypothesis
- Remember we are trying to reject or fail to reject the null, which basically says we either
  - ***'didn't find anything' or***
  - ***'failed to <u>not</u> find anything'***

Developing a hypothesis - overview readings to review

- https://www.scribbr.com/statistics/hypothesis-testing/
- https://opentext.wsu.edu/carriecuttler/chapter/developing-a-hypothesis/#:~:text=A%20researcher%20begins%20with%20a,prediction%20is%20called%20a%20hypothesis.
- https://www.skillsyouneed.com/num/hypotheses-testing.html
- https://www.nedarc.org/statisticalhelp/advancedstatisticaltopics/hypothesisTesting.html
- https://www.youtube.com/watch?v=joNb67F1UbY

Hypothesis : Simplicity, narrowness

- KISS principle
- Boiled down to the essence of the relationship you are testing
    - **Null** is the thing being tested, **Alternative** is everything else
- Research/Alternative and Null are opposites
    - H_0 - Null Hypothesis
    - H_a or H_1 - Research/Alternative Hypothesis

Examples of research/null hypotheses

Example 10.2: Hypotheses with One Sample of One Categorical Variable

About 10% of the human population is left-handed. Suppose a researcher at Penn State speculates that students in the College of Arts and Architecture are more likely to be left-handed than people found in the general population. We only have one sample since we will be comparing a population proportion based on a sample value to a known population value.

- **Research Question**: *Are artists more likely to be left-handed than people found in the general population?*
- **Response Variable**: Classification of the student as either right-handed or left-handed

State Null and Alternative Hypotheses

- **Null Hypothesis**: Students in the College of Arts and Architecture are no more likely to be left-handed than people in the general population (population percent of left-handed students in the College of Art and Architecture = 10% or $p = .10$).
- **Alternative Hypothesis**: Students in the College of Arts and Architecture are more likely to be left-handed than people in the general population (population percent of left-handed students in the College of Arts and Architecture > 10% or $p > .10$). This is a one-sided alternative hypothesis.

- https://online.stat.psu.edu/stat100/lesson/10/10.1