DATA4ALL

INSTRUCTOR GUIDE

INSTRUCTOR GUIDE

Version 2024-2 (September 14, 2024)

2022-24

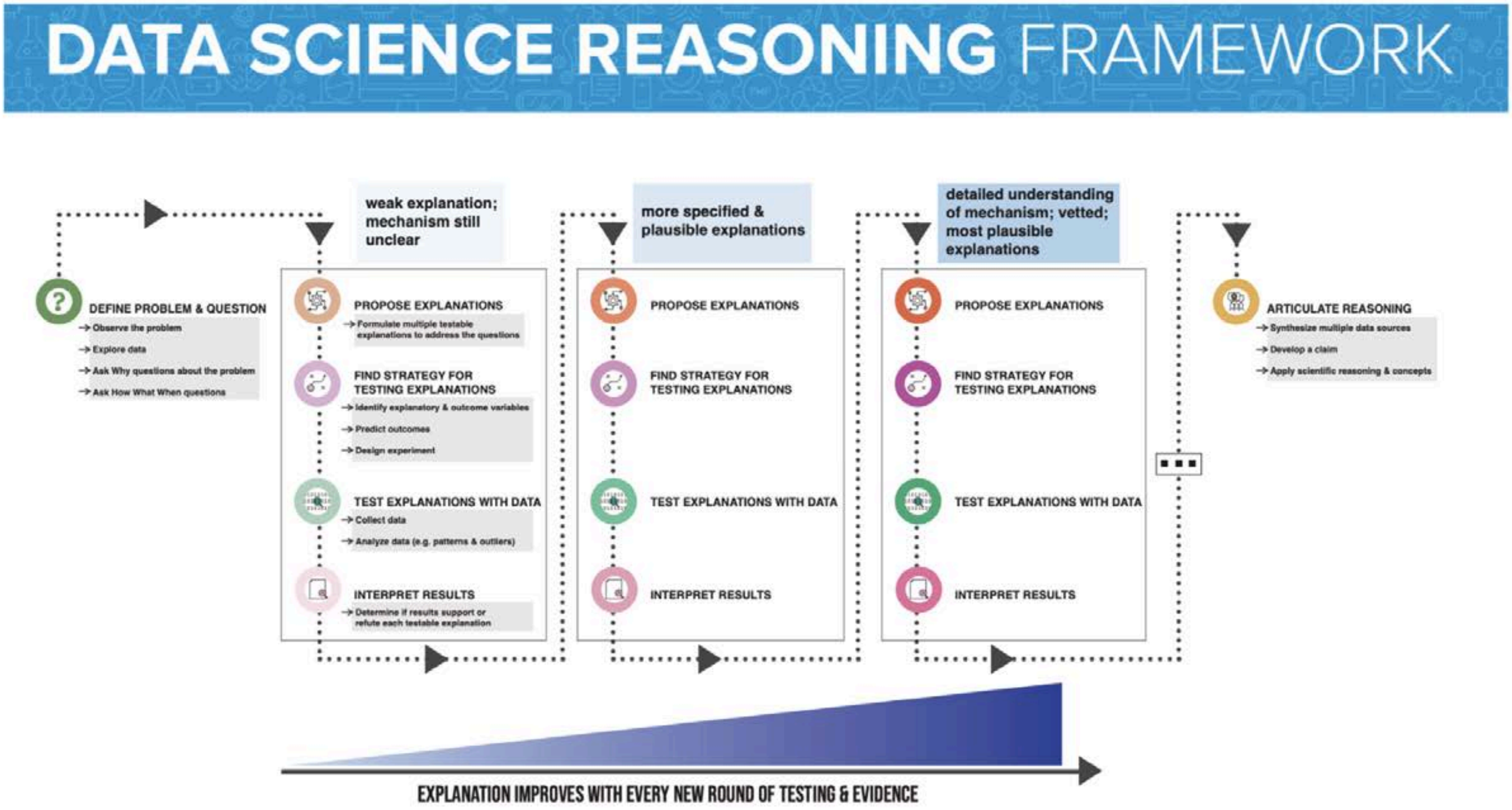
Authors:
John Domyancich, Argonne National Laboratory
Bethany Frank, Argonne National Laboratory
Julia Koschinsky, Center for Spatial Data Science, University of Chicago
Evelyn Campbell, Data Science Institute, University of Chicago

Reviewed by Damaris Hernandez

Curriculum Overviews

Curriculum Overview

The Data Science Reasoning Framework


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

The Data Science Reasoning Framework (DSRF) drives the design of the curriculum and is made explicit in the instructional sequence so that students are aware of the process involved in working with data to answer a question or solve a problem. The DSRF is not linear and should not be viewed as a "recipe" for doing data science. However, it does embody the overall progression of using data to answer questions and solve problems through describing a phenomenon, identifying interesting patterns or outliers, proposing explanations for these patterns and testing these explanations against each other as one moves toward a claim that is supported by evidence and reasoning. This process is iterative: With each new cycle of evidence, explanations become more specific and plausible.

The DSRF is meant to provide structure to doing data science in a way that encourages the discovery of unexpected patterns and developing explanations for these patterns as well as addressing some of its pitfalls that plague data science such as confirmation bias and focus on expected patterns.

In professional practice, data science is much more complex, being more iterative, requiring domain knowledge, intuition and a host of other elements that are only built through years of experience. For the purposes of this program, elements that are beyond scope or unreasonable due to time constraints have either been delivered through direct instruction or avoided altogether. The two investigations presented here have been highly simplified in order to focus on the key concepts and skills that are most useful and -accessible to the high school student who has little to no experience working with authentic data in this way.

Case Studies

Case studies provide students an authentic way to engage problem solving with data. Whereas Project Based Learning focuses on an applied student project, case studies in this workshop are also applied projects but they place particular emphasis on reflection of the problem-solving process. Our case study methodology requires students to first identify a problem and define a driving question and subsequently research relevant topics around the question. By nature, case studies have multiple solutions and therefore align well with data-driven problems in that answers/explanations are not absolute, but one gradually builds confidence in one over others as the investigation proceeds.

Students will be exposed to the DSRF through the investigation of two related, yet unique, case studies. The first case study models the process of problem exploration and experimentation using data in the context of the John Snow cholera investigation from mid-19th century London. This case study deals with relatively small datasets, is well-researched and has a definitive solution. While the path forward is more uncertain in open data science investigations, the cholera case study serves as an exemplar for how data can be used to determine that an explanation is more plausible than others.

The second case study provides a modern context where students can leverage their personal knowledge by investigating the disparate impacts of Covid-19 in Chicago. Besides being more student-driven, this case study requires different methods and perspectives and is conducive to a wider variety of research paths and claims.

Instructional Approach

Spark Activities

Spark activities are short, low-risk activities designed to introduce students to the key concepts of the main lesson in an intuitive and playful way. The goals of the spark activities are to 1) help transition students from outside activities to the day's lesson, 2) elicit prior knowledge, 3) serve as a formative assessment to help guide the instruction of the main lesson, 4) build an intuitive understanding of the technical concepts covered in class, and 5) to be a fun activity to help build a sense of classroom community. Spark activities are diverse in their form and approach to achieving student engagement.

Google Colab and Jupyter/Colab Notebooks

The notebooks are largely independent work where students will be able to develop their computational thinking, coding and data science reasoning skills concurrently. Each notebook has a specific goal that the students are working towards and provides the opportunity to directly work with the data on a practical level. If they have questions, they can work with their small groups and their near-peer mentor.

The notebooks will use the "Use, Modify, Create" approach to coding. This means that the notebooks will often already have functioning code. Depending on the task, students simply need to run a cell of code (Use) or change (Modify) individual elements like variables and parameters of functions and view the output. As the workshop progresses and the students become more familiar with the code, students will be asked to write (Create) more of their code from scratch. This can include pulling code from other notebooks.

Interspersed through the notebooks are "Journal" prompts where students are asked to check understanding, reflect and predict. The journal prompts are an essential part of the notebook experience and students should be held accountable to record their thoughts in these spaces. This is valuable from a learning perspective and it enhances the notebook's value to the student beyond the workshop as they can refer back to their thinking.

Classroom Structure

Data4All was originally structured to facilitate project-based learning through group collaboration. The program consists of 30 students divided into 6 groups. Each group had an assigned mentor that would help lead them through notebooks and activities. While this structure is not necessary to implement the curriculum, students have been shown to benefit from the help of near-peer mentors with a bit more experience in data science. Depending on your classroom size, adjustments can be made. If it is not possible to facilitate the aforementioned structure in your classroom environment, consider prioritizing group work for activities and discussions. The notebooks are structured such that students can work on them individually, in pairs, or in small groups.

Small Group Discussion

Small group discussions are an essential part of the curriculum as they are often used to enhance student voice, wrestle with challenging concepts, and solidify understanding. They are also a forum where students feel more comfortable to ask about coding bugs and gaps in understanding. Mentors should look to create a welcoming environment in their small groups where all voices are heard and respected. The following guidance can be helpful in achieving this:

- □ **Make everyone's voice heard:** Often, one or two students can consistently dominate the conversation. This can put others in a state of deference which stifles the collective creativity and functioning of the group. Solution: Actively call on everyone for their thoughts. Even if a student doesn’t speak up or raise their hand, it doesn't mean they don’t have something to share. The following questions are low risk for the students and encourage discourse:

- “Did anyone have a similar question to that?”
- “Can anyone add onto this idea?”
- “Who has a different way of thinking about this?”
- “Who can summarize some of the ideas we’ve heard today?”
- “What do you think of that idea?

□ **Questions are key:** Questions are much more valuable to the students than answers. Questions give them ownership of the learning. Questioning is a bit of an art as it is a combination of essential questions and “on the spot” questions. Essential questions are bolded in the Instructor Guide. These are scripted and designed for a specific learning purpose. Often a discussion is started and driven by one of these questions. However, as the conversation ensues, you will need to be responsive to what the students say and develop other questions “on the spot”. While there is some improvisation to this, there are some simple questions that you can have at the ready for certain purposes:

- **Elicit:** get an idea of what they are thinking:
  - “What are your ideas about X?”
  - “Can someone restate the question we are trying to answer?”
- **Probe or Clarify:** make sure you understand what is being said and push students to articulate their thoughts clearly:
  - “Can you say more about that?”
  - “What do you mean when you say X?”
  - “What made you think of that?”

□ **Listen and be patient:** It is natural to fill the void of silence with your own thoughts. We often want to avoid these awkward silences. However, students (and adults) have often been trained to “wait you out”. This results in you answering your own question.

□ **Solution:** There is no harm in randomly choosing a student to share their thoughts. While it may feel like you are putting them on the spot, if you continue to work to create a welcoming environment in your small group, the students will feel less anxious and actually show respect for their perspective.

□ **Accommodate:** In situations where students may face challenges in actively participating in discussions, several strategies can be employed to facilitate their engagement. One effective approach is to provide students with a list of discussion questions in advance, enabling them to prepare thoughtful responses. Additionally, employing scaffolded questions and offering alternative ways for sharing their thoughts can create a more inclusive environment, encouraging students to express themselves comfortably. Another valuable technique is to initiate discussions with a 'think-pair-share' activity, which often elicits better and more productive contributions.

Large Group Discussion

These are opportunities for the whole class to tie everything together, e.g. in the “town hall” capstone experience in the last week where students articulate their evidence-based arguments to “public health decision makers”.

Lectures

Lectures are presentations to the whole class to introduce and explain key concepts of a class, such as the difference between correlation and causation or between proxy variables (such as demographic characteristics) and explanatory mechanisms.

Guest Speakers

Guest speakers give students an overview of how the workshop content is related to college and career options. E.g., in 2022, two PhD biologists walked students through innovative case studies of how they used data science to solve problems that save lives and improve health. Other speakers explained their own career trajectories to make the steps for choosing college majors and careers more transparent. They also overviewed where and how to access college and career information.

Case 1: Cholera in London

Lesson 1 – Define the Problem & Ask the Right Questions

Overview

Big Idea:

Observing a problem and asking the right questions are the key steps in beginning to answer a question or solve a problem. The questions greatly influence the direction of the investigation.

Lesson Summary:

Asking testable questions is done by making careful observations of a phenomenon and describing what we do not understand about it. Good questions ask why something happens (“why are some people getting sick and not others?”), and the answers should be empirically testable, meaning we can use evidence and reasoning to answer them (“are some people getting sick because they are drinking something that others are not drinking?”). When we carefully observe a problem or phenomenon, we break it down into units of information, known as “data” (“what are the characteristics and behaviors of people who are getting sick vs. not?”). That is, we decompose the phenomenon into its components and choose appropriate measures of these parts (“how do we measure if someone is sick vs healthy? Which parts of the body are affected vs not?”). Choosing how we describe something with data is determined by what we intend to use that data for (“how can we prevent people from getting sick? How can someone who is sick become healthy?”).

Goals:

- □ **DSRF:** Learn to apply scientific reasoning to define and address relevant questions, including identifying questions that can be assessed with evidence and alternative causal explanations, and using data visualization and statistics to assess which of these explanations is more plausible.
- □ **Computational Thinking:** Break a phenomenon down into individual pieces of data that can then be analyzed quantitatively.
- □ **Coding:** Learn the basics of Python programming, including data types, variables, lists, lists of lists, Pandas DataFrames, and debugging errors.

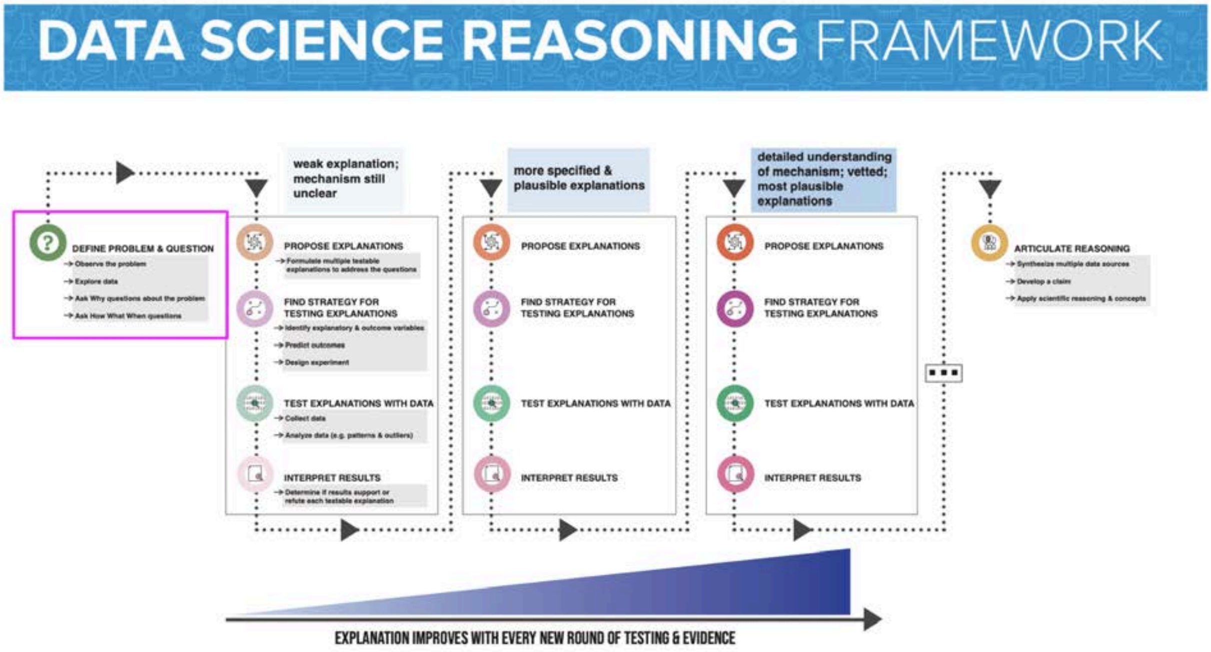
Activity Summary:

- □ **Spark (20 minutes):** Students are introduced to a medical mystery and then must ask testable scientific questions in order to progress the case to its solution. Students will be introduced to the case of beriberi and then will ask questions and be rewarded with clues for those questions that qualify as “testable” questions. Students will be rewarded with additional information to help solve the mystery. Students will then reflect on which questions were successful and which were not to summarize the characteristics of a good scientific question.

- □ **Notebook (40 minutes):** Students see how a problem, in this case cholera killing people in mid-19th century London, can be represented through data. By examining an excerpt from John Snow’s actual notes, the students learn how to break down a phenomenon into parts that can be measured with data and represent it as structured data in a computer. In doing so, they learn about fundamental coding concepts such as variables, data types, lists and arrays. Ultimately, they will represent John Snow’s data from the beginning of the notebook in a Pandas DataFrame.

- □ **Discussion (30 minutes):** Instructors role play as mid-19th century Londoners. Students ask questions of each in order to gain a sense of what was known and not known about cholera at the time as well as a sense of the popular beliefs about its causes. The students collect this information on post-its. Strategy involves asking several questions of each Londoner in order to develop a more complete picture. The activity concludes with the formulation of a research question. Students reflect back on the spark activity in the development of this question.

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
weak explanation; mechanism still unclear
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
more specified & plausible explanations
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
detailed understanding of mechanism; vetted; most plausible explanations
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

- □ **Lesson 1: Beriberi:** DEFINE PROBLEM & QUESTION - Observe the problem; Ask Why questions; Ask How What When questions

□ **Lesson 1 Pre-work: Introduction to Jupyter/Colab Notebooks**

□ **Lesson 1 Notebook: Seeing the Problem as Data:** DEFINE PROBLEM & QUESTION - Observe the problem; Explore data

□ **Lesson 1 Class Discussion: Interview the Londoners:** DEFINE PROBLEM & QUESTION - Ask Why questions; Ask How What When questions

Lesson 1 Spark: Beriberi

(20 minutes)

Big Idea	Goals	Resources
The scientific process starts with observations and then progresses based on what type of questions are asked and answered.	□ Identify the characteristics of a (testable) scientific question. □ Ask (testable) scientific questions to solve a medical mystery	□ Slides to introduce the spark activity. □ Envelopes with pieces of the medical puzzle to hand out to students. (6 copies, 1 for each group) □ Post-it notes for students to write proposed questions
ACTIVITY SUMMARY: Working in groups with their mentors, students are introduced to a medical mystery and then must ask testable scientific questions in order to progress the case to its solution. Students will be introduced to the case of Beriberi and then will ask questions and be rewarded with clues for those questions that qualify as “testable” questions. Students will be rewarded with additional information to help solve the mystery. Students will then reflect on which questions were successful and which were not to summarize the characteristics of a good scientific question.		

Preparation:

Before class, make copies of the “Beriberi Information for Envelopes” for each team, cut the pieces of the mystery apart and put them into an envelope. The mentor will hold on to the envelopes and hand out pieces to the team at the appropriate time. Gather other materials (post-it notes, large piece of paper for discussion at end, marker).

Roles:

An instructor will lead the main discussions. Mentors will assist with students bringing their questions to them and determining if they move the case forward; additional team members can go around and check in on groups while they are discussing what questions to put forth. (See tables below for examples of questions if a group gets stuck.)

□ **Instructor:** Explain to students that they will play the role of an investigator to solve a medical mystery. To do so, they will be introduced to the mysterious case of beriberi and will need to ask questions that will move the case forward. Encourage students to ask the types of questions they imagine an investigator would ask.

□ In their groups, students write down questions on post-its and ask the mentor. If the mentor thinks the question will move the case forward, the students are rewarded with “another piece of the puzzle”. If the question is too vague, the mentor will have the students try again. The mentor should keep all post-it note questions for an activity at the end. As questions are collected, sort them into two piles: "testable" questions and "non-testable" questions.

□ **Instructor:** Begin by introducing students to the case of beriberi:

It is 1897 and people are dying in Java, an island in Indonesia or the Dutch East Indies. They all seemed to share the same severe symptoms, beginning with overall muscle weakness, and loss of appetite, and eventually, they suffered paralysis and eventually death by heart failure. This disease was called "beriberi" by the indigenous people. This was a word from their native language that meant "I cannot, I cannot."

□ Remind students that in order to get more information they will need to ask specific follow-up questions. Encourage students to put themselves into the role of a scientist or investigator who collects empirical evidence to understand the cause of a problem. Then ask the types of questions they imagine someone in those professions would ask. Engage students in a discussion to explore the distinctions between questions posed by scientists and investigators versus other types of inquiries. Guide them in recognizing that scientific questions are falsifiable and rely on empirical evidence derived from observations, measurements, and experiments, rather than relying on speculation. In other words, frame a question in a way that allows for alternative explanations – where you can be proven wrong, instead of asking a rhetorical question where you're expecting to be right. You can also have a short discussion on the type of questions (who, what, why, when, where, and how).

□ Examples of questions:

General Question	Instead: More specific questions
How was it spreading?	Was it contagious or not?
	Is there a part of the body that felt especially sick (e.g. lungs vs. guts)
What was making them sick?	Was it something they were eating or breathing in?
	Who was getting sick and who wasn't?
	When were they getting sick? After eating or drinking or using the bathroom?
	Was it a virus or a bacteria making them sick?
	Did anyone survive? If yes, were they different?
	Who was the first to die and who did they have contact with before?

□ If students are struggling, the mentor can pause for a discussion to reinforce the idea that a good question should lead to a proposed explanation and experiment. Ask the students, "Based on your question, what is your proposed explanation for what might be happening?” Write down all reasonable explanations on the board. Choose one or two favorites and ask students how the scientists might go about testing these explanations. Ask students what they would expect the results of the test to be (their predictions) if their proposed explanations were correct.

□ **Mentors:** Look for the students to submit at least 3 good/testable questions. In order to decide if a question is “good/testable”, ask yourself, “Can an experiment be done to answer this question?” If yes, it is a “good” question. The idea here is for students to ask “testable” questions. However, do not reveal that this is the criteria being used. This is for the students to figure out!

□ Once students have submitted 3 testable questions, hand out the envelope with the next bit of information:

Scientists thought the disease might be caused by blood-borne bacteria. (After all, since the discovery of bacteria, almost all previously unknown diseases were attributed to a bacterial infection.) They decided to prove that a bacterium was the culprit by conducting an experiment. They used chickens as their trial subject. They injected a group of chickens with the blood from a patient who had beriberi and then to prove that the blood carried the “bacterium that caused the disease” they injected another group of chickens with saline or simple salt solution. Well, it turned out that both groups had chickens that got beriberi! So back to the starting board they went.

□ Continue to have students ask questions. Encourage students to consider what questions worked well last time and make improvements upon their questions this time around if necessary. Now the questions should be more focused, but still look for ones that can be potentially answered through experiment.

□ Examples of questions:

General question	Instead: More specific questions
What else could it be?	What has changed in their environment?
	Were all the chickens getting sick?
Was it something in the environment?	Was it something they were eating?
	Was it something they were drinking?
	Who was not getting sick?
	When were they getting sick?
	Did anyone recover from it and what did they have in common?

□ Once students have submitted 3 more successful questions, hand out the envelope with the next bit of information:

One of the scientists who had been sent to work on this mystery was a Dutch physician and pathologist named Dr. Christiaan Eijkman. One day, as he walked around the hospital compound, he observed his surroundings. He noticed that the cook fed every one of the patients the staple diet of the nation: polished white rice. Polished rice is wild, brown rice with the husk or outer layer rubbed off so that its color is white. It was the rice of choice of the middle class of the Indonesian people. He also noticed that the hospital staff fed the chickens (that would eventually be the chicken soup for the patients) wild brown rice. White rice was more expensive than brown rice, so the chickens were usually fed brown rice.

□ Continue to have students ask questions. Encourage students to consider what questions worked well last time and make improvements upon their questions this time around if necessary.

□ Again, look for questions that can be investigated through experiment.

□ Example of questions:

General questions	Instead: More specific questions
Was it something the chickens were eating that was making them sick?	Did chickens fed brown rice get sick?
	Did chickens fed white rice get sick?
	Which chickens got sick…those fed white rice or brown rice?
Was it something the people were eating that made them sick?	Did all the patients get fed both rice and soup?
	Did the patients who were only fed rice get sick?
	Did patients who were fed only soup get sick?

□ Once students have submitted 3 more successful questions, hand out the envelope with the final bit of information:

Dr. Eijkman realized that what the chickens were being fed was an important observation and thought that maybe the wild brown rice contained something that the white rice did not. So he conducted another experiment. He divided the chickens once again into two separate groups. He fed one group of chickens only white rice and the other group only wild brown rice. Then he watched and waited.

It turned out that the chickens that had been fed wild brown rice did not get sick at all, but the chickens that had been fed the polished or white rice became weak, lost their appetite and eventually died from beriberi. Eureka, the case was solved!

As Dr. Eijkman and others continued to research this interesting case, they found that polished white rice lacked thiamine, a vitamin necessary for good health. This was actually the first "vital amine" or vitamin to be discovered. It is also called vitamin B1.

We've now known for more than a hundred years that brown rice is more nutritious than white rice. But most Asian cultures associate eating white rice with prosperity and eating brown rice with bad luck. Most rice is still milled or polished, both in Asia and elsewhere. In Europe and America both *white rice and brown rice are consumed, but mostly white. In fact, some white rice is chemically fortified to add back the B vitamins. In 1929, Eijkman and Hopkins were awarded the Nobel Prize for Physiology or Medicine for this discovery.*

□ As a wrap up activity, the mentor organizes the submitted questions in two groups: “scientific questions” and “non-scientific questions” on the table for the students to see. The mentor then asks the students to find what the scientific questions have in common. The goal of this discussion is the realization that scientific questions are testable (can be investigated through experimentation).

Scientific questions	**Non-scientific questions**
Are testable: □ Are about facts (will X reduce Y?), not values (do you prefer chocolate or vanilla?) □ Can be investigated using an experiment □ Generate multiple possible explanations (hypotheses) that are falsifiable (can be proven false) □ Can be measured (e.g. using a variable to measure an outcome like a disease)	□ Have known answers (e.g. are google-able for a number, word, or short statement) □ Cannot be tested or measured (“does God exist?”) □ Elicits an opinion or value statement (e.g. “do you like the color blue?” or “is this the right thing to do?”)

Pre-work: Introduction to Jupyter/Colab Notebooks

(10 minutes)

Big Idea	Goals	Resources
Google Colaboratory (Colab) and Jupyter Notebooks are interactive tools for learning Python	□ To become familiar with Python and the Jupyter/Colab environment	Introductory Notebooks (Pick 1) □ Colab □ Jupyter
ACTIVITY SUMMARY: Students will familiarize themselves with the functionalities of Jupyter and Colab notebooks (depending on instructor preference), which will be essential for navigating notebooks for the rest of the curriculum.		

Facilitation Notes:

The instructor may wish to guide students through the Jupyter/Colab notebooks or allow students to work through them independently. This short activity will not be comprehensive for all the ways they can use a notebook, but will allow for a basic understanding so that they can approach the first notebook with confidence.

Lesson 1 Notebook: Seeing the Problem as Data

(40 minutes)

BIG IDEA	GOALS	RESOURCES
Information collected or recorded from the real world is called data. This data can be organized and stored in a computer.	□ Break down a phenomenon into measurable parts, organize real world data, and store it in a computer.	□ Notebook 1

ACTIVITY SUMMARY: Students see how a problem, in this case cholera killing people in mid-19th century London, can be represented through data. By examining an excerpt from John Snow's actual notes, the students learn how to break a phenomenon down into data components and represent it as structured data in a computer. In doing so, they learn about fundamental coding concepts such as variables, data types, lists and arrays. Ultimately, they will represent John Snow's data from the beginning of the notebook in a Pandas DataFrame.

Here is an example of data relevant to students: Many of us listen to music streaming services that use algorithms to predict our musical preferences based on the songs that we listen to. Spotify generates a playlist for each of its users based on these preferences, and in order to do this, it uses the songs that a user listens to as data. The catalog of songs that we listen to can be measured in various ways, such as quantifying the number of songs that we listen to of a specific genre or keeping track of what artists we tend to listen to. The phenomena - our musical preference, is represented as data - the songs we listened to, which then is used to generate a playlist that may or may not include some of the songs.

FACILITATION NOTES:

- □ The students are not expected to understand the details of the code. For example, the dictionary cell showing John Snow's data is meant to show how the historical data can be represented in code. The syntax is not important, but the students should be able to draw connections to the data and the image at the beginning of the notebook and what is shown in this particular cell.

- □ As this is the students' first encounter with a Jupter/Colab notebook, they will need particular guidance and prompting for the following:
  - ○ Running a cell
  - ○ How the order of running cells influences the value of variables
  - ○ The difference between markdown and code cells
  - ○ Comments in code cell (demarcated with a #)
  - ○ Where they are expected to make changes to the code cells
  - ○ How to edit a markdown cell (double click)

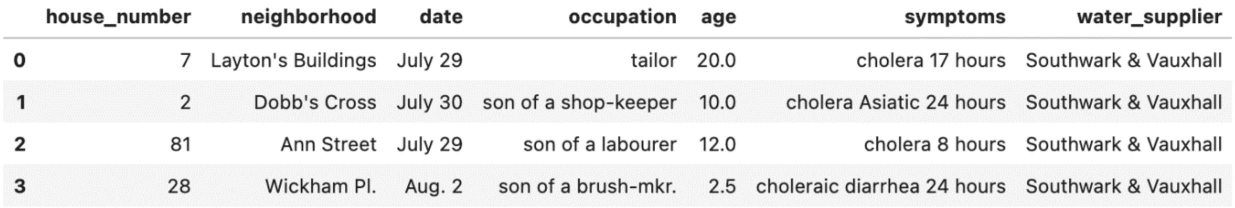
- □ If a student somehow “breaks” their code by deleting something or changing some part of the code that they weren’t supposed to, they can “Revert Notebook to Checkpoint” to return the notebook to a state that was working.
    - ○ Encourage the students to frequently save their notebook as they work through it to save their “good” progress.

Assessments:

- □ Near-peer mentors will check the code of the students in their small group

- □ The following questions can be asked to assess understanding as the students work through the notebook:
    - ○ What is the data type of that piece of data?
    - ○ How many variables do you see in that cell?
    - ○ What data is contained in that list?
    - ○ Where are the lists in the DataFrame?

- □ Completion of the notebook should result in a Pandas DataFrame with four entries.
    - ○ The existing code will only print 3 entries.
    - ○ In addition to making the person_list_3 variable, they will have to add this list to the person_2d_list variable as well.



	house_number	neighborhood	date	occupation	age	symptoms	water_supplier
0	7	Layton's Buildings	July 29	tailor	20.0	cholera 17 hours	Southwark & Vauxhall
1	2	Dobb's Cross	July 30	son of a shop-keeper	10.0	cholera Asiatic 24 hours	Southwark & Vauxhall
2	81	Ann Street	July 29	son of a labourer	12.0	cholera 8 hours	Southwark & Vauxhall
3	28	Wickham Pl.	Aug. 2	son of a brush-mkr.	2.5	choleraic diarrhea 24 hours	Southwark & Vauxhall

Lesson 1 Class Discussion: Interview the Londoners

(30 minutes)

Big Idea	Goals	Resources
Asking the right questions is the key step in beginning to answer a question or solve a problem. The questions greatly influence the direction of the investigation.	□ Gather a diverse set of information that provides context for the problem. □ Brainstorm a variety of questions about the problem.	□ Londoner role playing cards for instructors (also see slides from lesson 3). □ Post-its and large paper to construct KWL chart □ See 1-page overview of cholera theories for mechanisms □ Slide with DSRF

ACTIVITY SUMMARY: Mentors role play as mid-19th century Londoners. Each Londoner will argue for one of the predominant theories of how cholera was spread. Rotating between Londoners, students, in their groups, ask questions of each in order to gain an understanding of the theory, a sense of what was known and not known about cholera at the time and what life was like for different people during this time. The students record what they know, want to know and have learned before and after these interviews. This process guides them in the development and refinement of strong questions to drive their investigation.

Preparation:

In their conversations with the student groups, each "Londoner" will advocate for one of the following airborne theories while also providing a broader context of what life was like in London during the mid-19th century (the #s in the 1-pager reflect when they were first introduced historically; excluding the first theory):

□ **Airborne-Contagious:** People infect each other through the air (students will find this familiar because of COVID: greater distancing would reduce risk). (#3 in cholera theories 1-pager)

□ **Generally Airborne-Noncontagious:** People get cholera by inhaling generally polluted air (source is somewhere from earth's surface: hard to avoid unless you leave the area). (#2 in cholera theories 1-pager)

□ **Locally Airborne-Noncontagious:** People get sick by inhaling locally polluted air (e.g. coming from sewage through gullies, or pest field: avoid inhaling air from sewers or move away from pest field). (#5 in cholera theories 1-pager)

□ **Airborne-Elevation:** People get cholera after inhaling polluted air that settles in low elevation areas (move out of low-elevation areas). (#6 in cholera theories 1-pager)

□ **Airborne-Poverty:** People get cholera after inhaling polluted air but also infect each other in severely overcrowded quarters with unsanitary conditions (improved sanitary conditions would reduce disease risk). (#4 in cholera theories 1-pager)

None of the Londoners argue for the waterborne theory since John Snow developed this in contradiction to the predominant airborne theories (and in the capstone experience, students will try to convince the Londoners that their theories are wrong and vice versa):

□ **Waterborne/Foodborne-Contagious:** People infect each other by ingesting food or drink that was contaminated by someone who had cholera (e.g. soiled hands from changing diapers that are then used for cooking). Cholera risk could be reduced through better sanitation and avoiding contaminated water sources. (#7 in cholera theories 1-pager)

□ Give each mentor (up to 6) a role-playing card. If there are fewer than 6 mentors, prioritize these four theories: Airborne-Contagious, Locally-Airborne Noncontagious, Waterborne/Foodborne-Contagious and Airborne-Poverty theories over Airborne-Elevation (Farr). The mentors will need to study their card in advance of the activity and can supplement the information on the card with research into the specifics of what life was like for their particular role.

□ Each mentor leads his/her group in the construction of a KWL chart. A **KWL** is a curriculum mapping tool that illustrates what students already **K**now about the problem, what additional information students **W**ant to know and charts learning over time by asking students what they have **L**earned over the unit. Each group will construct a KWL and revisit and revise it as they progress through the case study. Make sure each group keeps their KWL chart for Lesson 2.

□ Mentors should begin by making three columns on a large piece of paper and label the columns as follows:

Figure 1. Setup of the Class KWL Charts.

What do you **KNOW?**	What do you **WANT** to know?	What have you **LEARNED?**

□ Mentors should lead their groups in a quick brainstorm of what they **know** and what they **want** to know about the problem: *Cholera is killing the people of London*.

□ At this point, the students may not know much about the problem, but encourage them to clearly state what they can definitively say is true and avoid statements that may be vague or possibly inaccurate. For students who may still not know much about cholera, giving them another short explanation of the disease may be helpful.

□ Record these statements in the "What do you KNOW" column.

□ The "want to know" is especially important as this seeds the development of questions that can be investigated. Remind the students of the beriberi activity in guiding the formulation of these questions. Instructors may want to share the following table with students as a reminder:

Scientific questions	Non-scientific questions
Are testable: □ Are about facts (will X reduce Y?), not values (do you prefer chocolate or vanilla?) □ Can be investigated using an experiment □ Generate multiple possible explanations (hypotheses) that are falsifiable (can be proven false) □ Can be measured (e.g. using a variable to measure an outcome like a disease)	□ Have known answers (e.g. are google-able for a number, word, or short statement) □ Cannot be tested or measured ("does God exist?") □ Elicits an opinion or value statement (e.g. "do you like the color blue?" or "is this the right thing to do?")

□ Record these questions in the "What do you WANT to know" column.

□ Explain to the students that while data will be critical in their study of cholera, there is also a lot of important information to be gained by talking to people that are experiencing the problem themselves. Unfortunately, we can't travel back to 19th century London, but we can bring it to us!

□ Explain that each group will have the opportunity to talk with each Londoner, played by a mentor. Their responsibility is to interview each Londoner and gather as much useful information as possible. Give students time to organize what questions they want to begin with. Instructors can check in on the groups during this time to assess if any groups need more help crafting their questions.

□ Have the Londoners spread out around the edges of the room.

□ Each group should go to a different Londoner. It does not matter which one. They will get three minutes to speak to each one. Encourage the students to ask questions strategically and remind them to focus on their "Want to know" questions.

□ Set a timer for 3 minutes and allow the interviews to begin.

Guidance for Londoners:

- ○ First introduce yourself (name and title only). It is not your responsibility to provide the information on your card. The groups should be asking questions that elicit your knowledge and perspective.
- ○ Pay attention to the type of question asked. If it can be answered with a word or short statement, leave it at that. Do not attempt to fill in any gaps or infer what the student is "really asking".
- ○ Do your best to answer but avoid speaking beyond what would be reasonably known by that person (e.g., the mayor of London would not know the variety of treatments people tried at the time for cholera. However, the doctor would.)
- ○ If a question elicits bias or opinion, it is an opportunity to provide less useful information. Here is a place you can "editorialize".
- ○ It's perfectly OK to say, "I don't know".

□ Students should record what they learn in their notebooks.

□ After 3 minutes, signal the groups to rotate to the next Londoner and reset the timer for another 3 minutes.

□ Repeat until the groups have each interviewed all of the Londoners.

□ Groups can now return to their tables.

□ Mentors lead their group in the revision of their KWL chart.

□ First focus on adding statements to the "What have you LEARNED" column. If a "LEARNED" statement answers a question in the "WANT" column, cross out the question.

□ Ensure that the students are conservative in what they have learned from the Londoners. It is possible that some statements could be opinions or anecdotal.

□ Also add to the "KNOW" column, but again, make sure the students only provide statements that can be considered factual.

□ Finally, record any new questions they have in the "WANT" column. Again, encourage the generation of testable questions.

Wrap-Up:

The lead instructor will refer to the Data Science Reasoning Framework image (on the last slide), pointing out the “Define Problem & Question” phase, specifically how we have engaged with the problem in multiple ways (through data, by talking with people) to better understand the problem and begin formulating the question that will drive our investigation.

However, defining the question is so important that we must explore additional data before committing to a single question. Therefore, we will next be learning ways that we can see, describe and change data that reveal more about the problem and allow us to formulate a clear question.

Lesson 2 – Explore Data

Big Idea:

Exploring the problem (phenomenon) through data allows us to observe it in a way that reveals interesting patterns that are the foundation of proposed explanations.

Lesson Summary:

Often, new data must be created using existing data in order to make meaningful comparisons. The variable of interest is a particular piece of data that best describes the outcome associated with the problem. Interesting patterns and unexpected variations in the variable of interest cause us to ask questions about the mechanisms at work that might cause the outcome, leading to proposed explanations.

Goals:

□ **DSRF:** Understand the need to normalize data and be able to identify variables of interest.

□ **Computational Thinking:** Identifying outliers.

□ **Coding:** Element wise arithmetic operations, DataFrame manipulations.

Overview

Lesson 2a: Seeing the Problem in the Data

□ **Spark (20 minutes):** Students consider multiple variables of interest in making the claim, "Who is the GOAT: LeBron or Jordan? In doing so, they see the challenges of choosing one variable of interest over others when making a claim and how different variables can lead to different claims.

□ **Notebook (30 min):** Students normalize mortality data from 19th century London in order to identify cholera trends by year. They are introduced to the concept of an outcome variable (in this case, mortality rate) and reflect on the value of normalization.

Lesson 2b: Finding Patterns

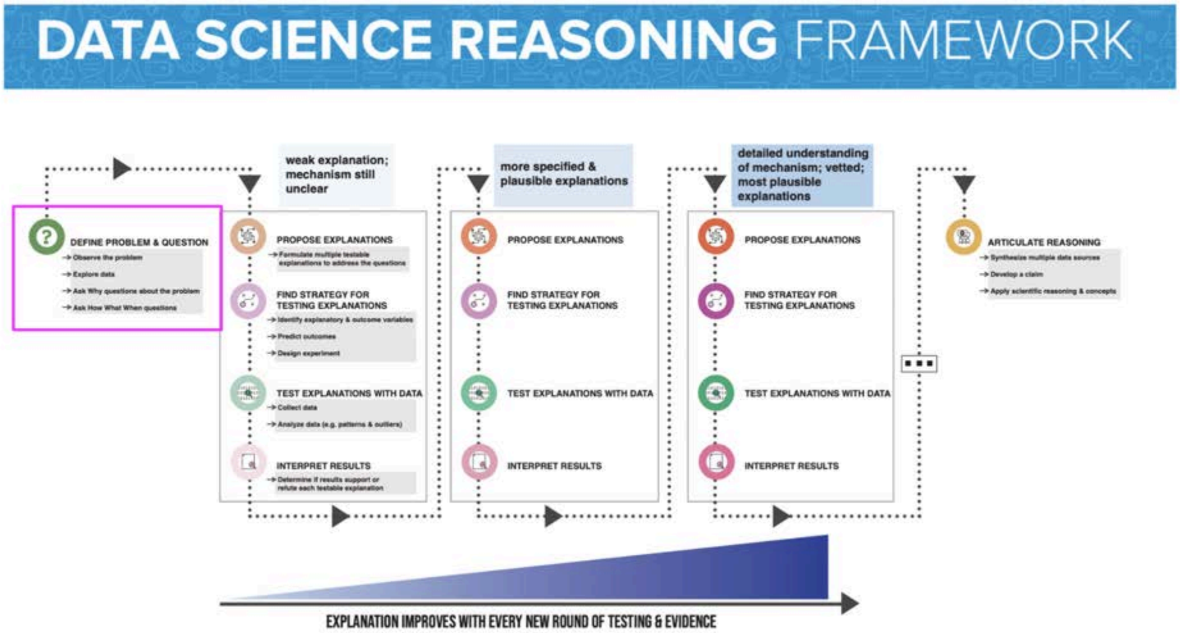
□ **Spark:** Students will evaluate their ability to determine a random pattern from a true pattern. In this activity one instructor leaves the room and students write down two lists of 40 coin-toss results: "heads, tails, tails, heads...," the first generated by students sequentially calling out "heads" or "tails," trying to simulate random coin flips and the second by actually flipping coins. The instructor returns and has to guess which is random and which is simulated random. The class will then discuss how truly random does not feel random and ways to detect if something is random or not (truly random tends to generate longer sequences of heads or tails – students often think of random patterns as alternating heads and tails, which is how the instructor can often tell the difference).

- □ **Alternative Spark at UChicago:** Students will play a game developed at the University of Chicago’s Weston Game Lab (Infection City).

- □ **Notebook (30 min):** Students gain familiarity with common Pandas DataFrame manipulations including selecting columns, creating new ones with generated data and summarizing the data with column statistics by investigating how mortality rates varied spatially across London in 1849.

- □ **Discussion (10 min):** Considering the new information on the spatial patterns of cholera, groups revisit their explanations and update.

Reference to DSRF:



DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

- □ **Lesson 2a: Spark: LeBron vs. Jordan:** EXPLORE DATA I – Seeing the Problem in the Data

- □ **Lesson 2a Notebook:** EXPLORE DATA I – Seeing the Problem in the Data

- □ **Lesson 2b Spark (Is it Random? or Infection City):** EXPLORE DATA I – Finding Patterns

- □ **Lesson 2b Notebook:** EXPLORE DATA II – Explore Spatial Patterns of Cholera

- □ **Lesson 2b Class Discussion:** DEFINE PROBLEM & QUESTION: Finalizing the Question

Lesson 2a – Explore Data I: Seeing the Problem in the Data

(20 min)

Lesson 2a Spark: LeBron vs. Jordan

Big Idea	Goals	Resources
Students investigate normalization of data as well as variables of interest (different measures/ variables of an outcome).	□ Understand the need to normalize data and be able to understand the challenges of measuring an outcome with different variables of interest when making a claim and how different measures/ variables can lead to different claims.	□ Slides to introduce the activity

ACTIVITY SUMMARY: Students consider multiple measures (variables of interest) in answering the question , “Who is the GOAT: LeBron or Jordan? In stating a claim as their answer, they see the challenges of choosing one measure over others when making a claim and how different measures/variables can lead to different claims. This activity was originally created by skewthescript.org.

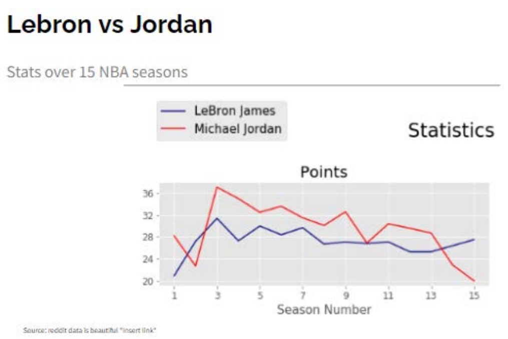
Preparation:

Pull up slides: Measuring an Outcome of Interest.

Roles:

An instructor will lead the main discussion.

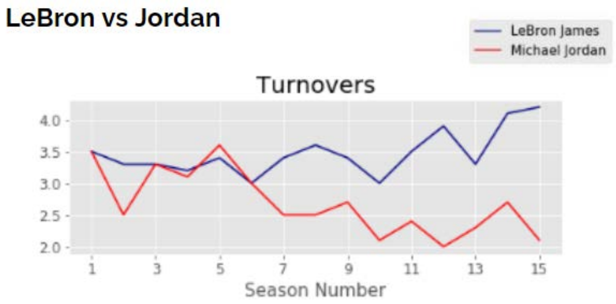
- □ Take a poll and get student opinions on who they think is the better player and why. Then say “let’s see what the data says”
- □ Go over the results of the poll and ask students to defend their choice.
- □ Lead/facilitate discussion on this graph. Who is the best [according to this graph]? Michael Jordan. Expect students to bring up other stats that are reflected here. Ask what makes a player The GOAT? What MORE data do we need to answer this question? Another way to ask this question is, what other measures of being a great player



Lebron vs Jordan

Stats over 15 NBA seasons

Source: reddit data is beautiful "insert link" (**variables of interest)** should we look at (e.g., turn overs, rebounds, etc.)?



□ Take another poll after looking at the graph.

□ Ask, “What about according to this graph?”

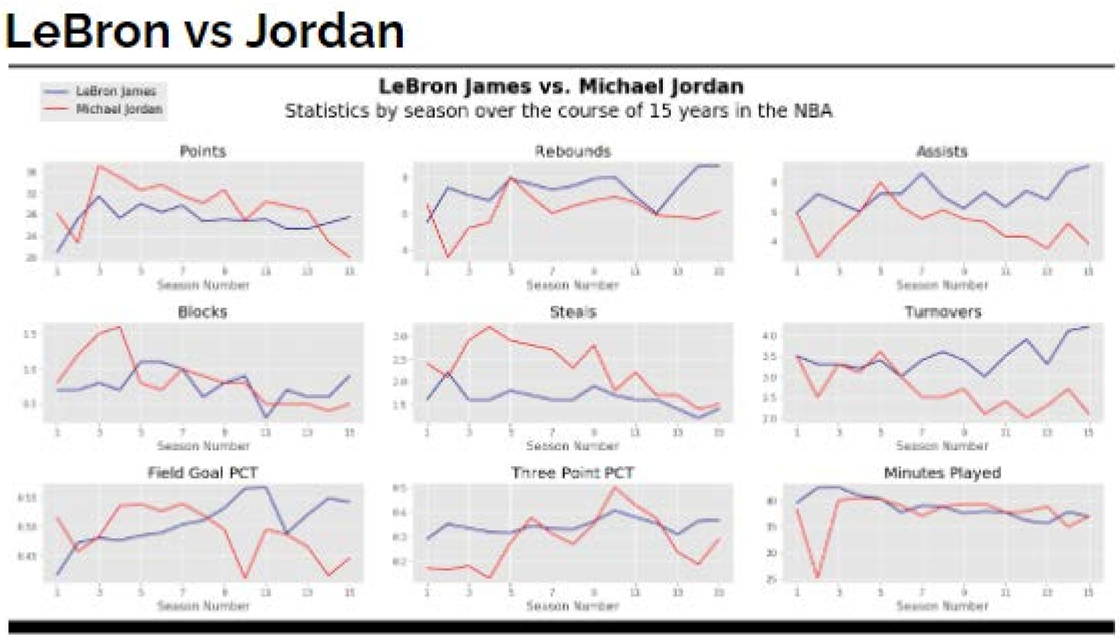
Lead/facilitate the discussion of this graph. Who is the best [according to this graph]? Lebron James?

Expect students to bring up other stats that are reflected here.

Ask what makes a player the GOAT? What MORE data do we need to answer this question?

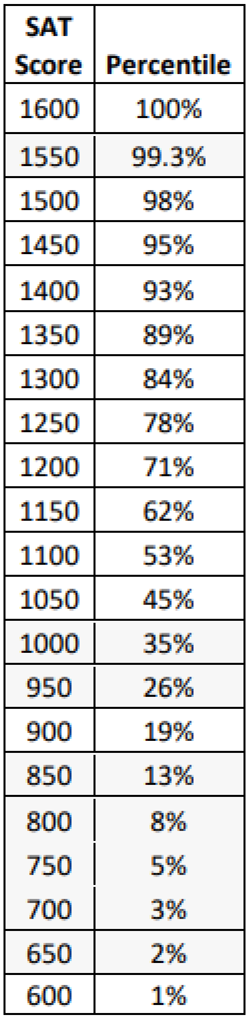
□ Poll again.

□ “What about when we consider all of this data?” Why do we need to look at more data to answer this question? Why can’t we just base it on one or two player stats? Explain the connection between considering these other stats or variables of interest and the idea of normalizing data. Give students a few minutes to analyze the data and then have them state their claim.

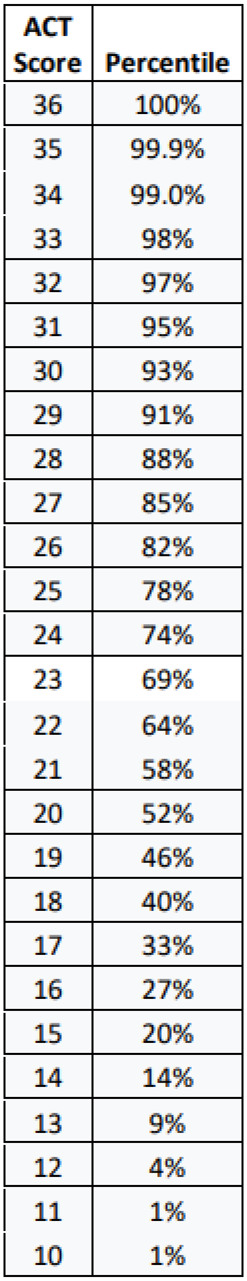


IF YOU HAVE TIME TO CONTINUE THE CONVERSATION:

□ Another example is comparing ACT and SAT scores. Which is better, a SAT of 1050 (out of 1600) or an ACT of 23 (out of 36)? How would we compare the two? We need to look at percentiles since the tests use different scales and scoring systems. (May also be an opportunity to discuss the merits of test-optional admission policies being implemented by many higher education institutions.)



SAT Score	Percentile
1600	100%
1550	99.3%
1500	98%
1450	95%
1400	93%
1350	89%
1300	84%
1250	78%
1200	71%
1150	62%
1100	53%
1050	45%
1000	35%
950	26%
900	19%
850	13%
800	8%
750	5%
700	3%
650	2%
600	1%



ACT Score	Percentile
36	100%
35	99.9%
34	99.0%
33	98%
32	97%
31	95%
30	93%
29	91%
28	88%
27	85%
26	82%
25	78%
24	74%
23	69%
22	64%
21	58%
20	52%
19	46%
18	40%
17	33%
16	27%
15	20%
14	14%
13	9%
12	4%
11	1%
10	1%

Lesson 2a Notebook – Seeing the Problem in the Data

(30 Minutes)

Big Idea	Goals	Resources
In order to make comparisons between groups, variables must be developed that measure these groups on the same scale.	□ Generate new data in the form of a normalized outcome variable.	□ Notebook 2a
ACTIVITY SUMMARY: Students normalize mortality data from 19th century London in order to identify years that experienced a cholera outbreak. They then gain familiarity with common Pandas DataFrame manipulations including selecting columns, creating new ones with generated data and summarizing the data with column statistics by investigating how mortality rates varied spatially across London in 1849.		

Even though this notebook does not require the students to do a lot of coding, it introduces very important concepts, --normalization and outcome variables–. It is important to pay attention to the following:

□ Make sure students are taking the time to record their thoughts in the reflections. Encourage reflection by repeating or rephrasing the prompts to the reflections and having one-on-one discussions about them.

□ The formula for the death rate may be confusing to some students. It may be helpful to explain that fractions are just another way to show division.

“Talking points” that can be brought up to enhance the experience for students:

□ The normalization calculation is done “elementwise” in the DataFrame, meaning the calculation is repeated row by row with new numbers for population and deaths each time. This process is not evident to students, but it is worth pointing out this “under the hood” look at the computation.

□ The normalization code cell toward the end of the notebook shows the power and efficiency of using computers to do data science: we can provide one calculation and the computer does it for every row in the DataFrame. It may not seem like much for a DataFrame with 15 rows, but we could use the exact same code cell to do this calculation for a million rows.

Assessments:

□ The following questions can be asked to assess understanding as the students work through the notebook:

- What is the data type of that piece of data?
- How many variables do you see in that cell?
- What data is contained in that list?
- Where are the lists in the DataFrame?

Lesson 2b – Explore Data II: Finding Patterns

Lesson 2b Spark: Is It Random? (20 Minutes)

Big Idea	Goals	Resources
People underestimate the frequency of apparent patterns produced by randomness, leading to overperception of spurious patterns much more frequently than people account for.	Students will be able to: □ Become aware of their tendency to see patterns when patterns do not actually exist.	□ Coins for half the students □ Slides to introduce the activity □ Article of how Spotify made their random shuffles less random **Extension Resources** □ Website for discussion points on randomness: How Spotify's shuffle algorithm works □ "A Very Lucky Wind" story from Radiolab's podcast on stochasticity
ACTIVITY SUMMARY: Students will evaluate their ability to determine a random pattern from a true pattern. In this activity one instructor leaves the room and students write down two lists of 40 coin-toss results: "heads, tails, tails, heads...," the first generated by students sequentially calling out "heads" or "tails," trying to simulate random coin flips and the second by actually flipping coins. The instructor returns and has to guess which is random and which is simulated random. The class will then discuss how truly random does not feel random and ways to detect if something is random or not (truly random tends to generate longer sequences of heads or tails – students often think of random patterns as alternating heads and tails, which is how the instructor can often tell the difference).		

Preparation:

Have coins to hand out to students; preview Google sheet for data entry; preview article, website and possibly podcast; pull up slides for the lesson.

Roles:

An instructor will lead the discussion; the instructor will then leave the room and one of the team members will lead the students in the coin flipping/ calling out activity. Another instructor will record data on the Google sheet. The instructor will then return and continue leading the discussion.

□ We will begin this spark activity by telling students that we will investigate what random means by doing a quick activity.

□ Have one instructor leave the room. Break the class into two groups. Have one group write down a list of 40 simulated coin-toss results by going down the row of students and having each student call out heads or tails (do NOT have students write down their answer ahead of time, as this defeats the point of the activity). Due to class size, you may want to do two rounds. Have the second group of students actually flip a coin and then go down the row and call out what they got. Again, due to class size, you may want to do two rounds. Before they begin, the instructor or assistant can ask the two groups "**Who here is good at lying**?" Whichever group answers first is the simulation coin flip group. Have an assistant keep track of what everyone calls out in both groups and record here on a Google sheet (sheet 1 is example data, sheet 2 is for the class data). Once both groups have gone, display the data.

□ Have the instructor return and guess which is random and which is simulated random. Usually the data with the longer run lengths indicate random numbers.

□ The class will then discuss how truly random does not feel random. Talking points are that random numbers often give you things that look like patterns. Also, coincidence is hard for us to accept. We usually want to impart some sort of meaning or reason.

□ Share the article, "Spotify Made Random Shuffle Less Random on Purpose – Here's Why" with students and discuss.

□ The instructor can also use examples from everyday life, such as:

  □ spotting animals in the clouds;

□ and thinking the following indicates a special connection:

  □ Running into an acquaintance while traveling far away from where you both live;

  □ Running into someone when you were just thinking about them the previous day;

  □ Dreaming an event and then something similar happening within the next month,

  □ Finding that two people in your class share a birthday.

□ Instructor can ask, "**Why do you think we do this**?" Maybe it is because people are evolutionarily disposed to over-perceive patterns because the cost of missing a real pattern is typically higher than the cost of mistaking randomness for a pattern.

□ Instructor can ask, "**What can we do to avoid this pitfall with our own data**?" The instructor can mention that "knowing is half the battle". Also, statistical analysis such as chi-squared tests (is there a correlation between two variables?") and p-values (is that correlation random?) can also help us.

IF YOU HAVE TIME TO CONTINUE THE CONVERSATION:

□ Use the website How Spotify's shuffle algorithm works to discuss What Randomness Looks like to drive home the idea that random patterns can have more repeat patterns than we expect.

□ If time permits, the instructor can share with them the "A Very Lucky Wind" story from Radiolab's podcast on randomness.

Lesson 2b Spark – Explore Data II: Finding Patterns

Alternative Lesson 2b Spark: Infection City (45 Minutes)

> *This spark activity is an alternative to the aforementioned "Is It Random?" activity. Materials for this activity are provided by Ashlyn Sparrow, Assistant Director of the Weston Game Lab at the University of Chicago.*

Big Idea	Goals	Resources
Show how disease spreads through a city and how health experts (epidemiologists) have to work together to stop spread.	Students will be able to: □ Intuitively solve a spatial optimization problem by placing health clinics far enough apart to minimize disease spread (instead of clustering them all in one area)	□ For UChicago: Infection City Games: one game per table (developed and provided by guest Ashlyn Sparrow at UChicago)
ACTIVITY SUMMARY: Four to six players per table: one player is disease/infection, the others are epidemiologists. As the disease tries to spread, the epidemiologists try to contain its spread by placing clinics in neighborhoods. Clustering clinics in one location increases spread while evenly spreading them out across the city reduces spread.		

Setup:

Four to six players: one player is Disease/Infection, the others are epidemiologists:

□ Open board (Hexacago)

□ Disease picks cards from the deck (doesn't share with others):

- ○ **Five players:** pick seven spread cards.
- ○ **Six players:** pick six spread cards.

□ Black cubes are infection.

□ White houses = clinics.

□ Each epidemiologist picks one clinic and places it on the map.

□ Infection places one black cube for each of the epidemiologists.

Two phases: disease phase, epidemiologist phase:

□ **Disease objection:** get as many black cubes as possible on the board. Disease is spreading and trying to overwhelm epidemiologists.

□ **Epidemiologist phase:** Every player has to take a turn. Each player can spend action point to remove disease from board,

Each epidemiologist has four action points (16 total points to use together):

□ Two action points to put down a clinic–allows you to remove diseases from a certain neighborhood.

□ Can fortify clinic to expand range/network to treat more patients and spend action points–start to produce vaccines.

□ Herd immunity–clusters of clinics that are operating in a way that they are vaccinating as much of the population as possible.

□ Strategy game–lay clinics as quickly as possible–dominant strategy.

□ Play first time without guidance, then tell players they need to cooperate to win game.

□ **Goal of game:** show how disease spreads through a city and how epidemiologists have to work together to stop spread.

Lesson 2b Notebook – Explore Spatial Patterns of Cholera

(30 min)

Big Idea	Goals	Resources
Interesting patterns and unexpected variations in the outcome variable cause us to ask questions about the mechanisms at work.	□ Use basic DataFrame manipulations to summarize, generate and filter (select) data. □ Apply DataFrame manipulations to explore patterns and outliers in data.	□ Notebook 2b
ACTIVITY SUMMARY: By applying basic Pandas DataFrame manipulations, students investigate how mortality rates varied across London by district in 1849. They then group districts by region (North, Central, etc.) to expose trends in these variations to see that South London was disproportionately impacted by cholera.		

This notebook is the first to require students to write a fair amount of code. Issues of syntax will be a stumbling block for many students. As a result, much of the time they will be attempting to find errors in their code and decipher error messages. The following tips will assist in helping students find their own errors as opposed to relying on the instructor.

□ Don't be intimidated by the size and complexity of the error messages: Most errors are caused by simple "typos": an unclosed parentheses or bracket, a misspelled variable.

□ When a string is used, it must be put in quotes " " or apostrophes ' '. It doesn't matter which you use.

□ Look at the end of the error message to find out what is wrong.

□ Deciphering error messages takes practice.

The syntax students see here will include:

□ Variable assignments

□ Referencing columns

□ `DataFrame['column_name'].method()` notation

□ Double brackets when selecting multiple columns

Helping students navigate errors:

- o Walk through the process with the student as opposed to finding the error yourself and pointing it out (Teach them to fish instead of giving them a fish!)
- o Look at the error first and then share your experience: “That error usually means there is a typo in a variable name.” Then leave it up to the student to do the rest.

□ When there are multiple lines of code in a cell, “comment out” lines one by one to find the error.

□ If a student is done early, enlist their help for others that may be stuck on an error or are struggling.

Lesson 2b Class Discussion – Finalizing the Question

(10 min)

BIG IDEA	GOALS	RESOURCES
Interesting patterns and unexpected variations in the outcome variable cause us to ask questions about the mechanisms at work.	□ Narrow focus of questions based on new data and finalize a research question.	□ KWL from Lesson 1 □ What Makes a Good Question?
ACTIVITY SUMMARY: Students add to and edit their previous brainstorm of questions by considering the presence of spatial patterns and previously acquired information.		

□ Ask the students, **"Do you notice any connection between a district's region of London (north, south, etc.) and mortality rate?"** From the DataFrame manipulations in the notebook, students should notice that the south districts had noticeably higher death rates.

□ At this point, ask **"What kind of questions come to mind after seeing this data?"** Facilitate adding these questions to the brainstorm (KWL chart) from Lesson 1. This is a brainstorm as well, but it is ok to remove or edit some of their questions from before. This is part of the narrowing process and is natural considering the new information.

Note: It is natural to jump to explanations before asking testable questions. If/when this happens, note that it's good that some people are thinking ahead, but we will need to develop some good questions first and look at the data more closely before we propose any explanations.

□ At this point, the group needs to identify a single scientific question. This question could come from the brainstorm, could be a synthesis of several questions or a completely new question. Remind them of a few guiding principles of good scientific questions:

- It can be investigated or answered by experimenting to distinguish between alternatives (it is testable).
- We can collect and analyze data as part of this experimentation.
- It is a causal question (will help us discern between why and why not).
- Starts out general but allows us to move to more specific as we add more details to our understanding.
- It is not a simple yes/no question.
- It is about an unresolved puzzle, i.e. not "google-able"

□ Also, mention that when trying to solve a mystery or puzzle, like we are doing here, a “why” question is a good choice.

□ Try to guide the group to develop a question that is sufficiently broad. Students may tend to think very specifically, i.e., “Why does lower population density result in higher mortality?” but this will put them in the trap of “confirmation bias”. In other words, they will look to prove the link between population density and mortality and dismiss other potential hypotheses.

□ An instructor will close this activity with the whole class by focusing on a common question: **“Why is cholera killing some but not others?”** This question focuses on a comparison between people who died vs. people who survived to prepare students to think in terms of impact and control groups. Try to guide your group to a question like this in your discussion, but do not force it.

□ Some students may think this question is too generic, but the discussion will focus on how this simple question has the key characteristics of a good scientific question. Most notably, it is a causal question (why?) with a focus on an impact (killed) and control group (not killed) that could be tested with an experiment.

Lesson 3 – Propose Explanations

Overview

Big Idea:

Good research questions are testable. Testable questions can be systematically investigated through data, which can help us begin to think about how variables relate to one another. Through data exploration, we can begin to propose explanations for how one variable may impact another, which can later be validated through experimentation.

Lesson Summary:

Proposed explanations should include mechanisms that could generate the observed patterns in the outcome variable of interest. Multiple alternative explanations that are distinguishable from each other should be proposed.

Goals:

□ **DSRF**: Begin formulating testable explanations for phenomenon based on observations and data

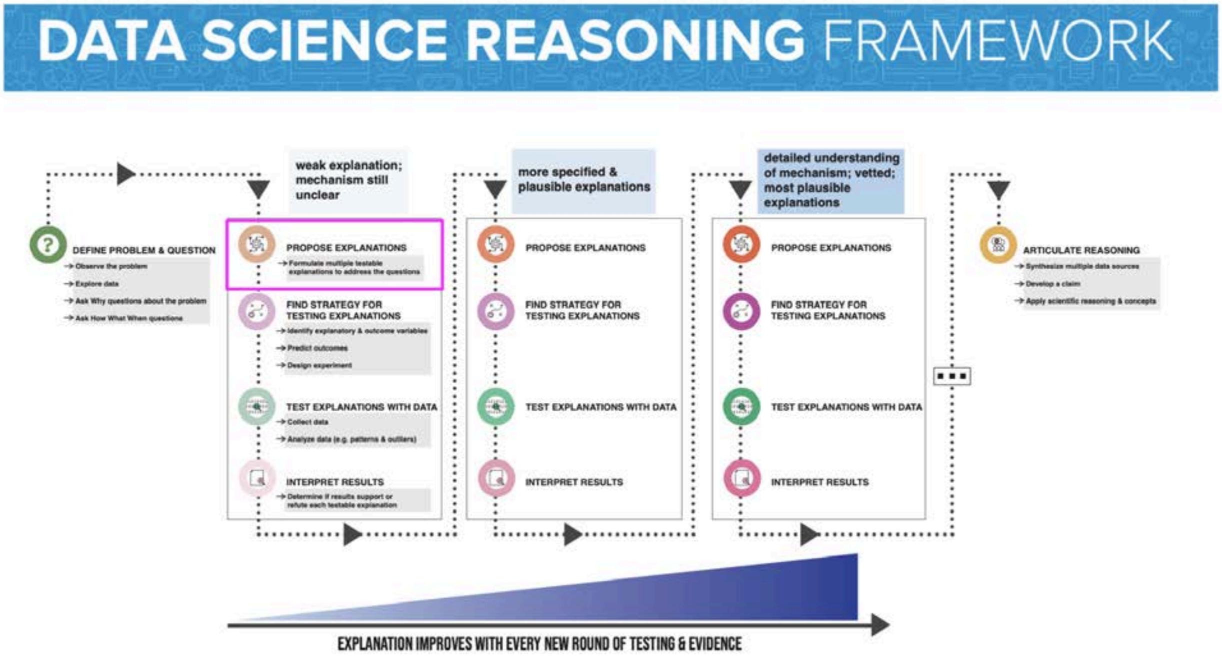
□ **Computational Thinking**: Perform statistical analysis to understand relationships between variables

□ **Coding**: Learn functions and methods for grouping, aggregating data, and performing linear regression

Activity Summary:

1. **Spark (30 min): Correlations and Linear Regression -** Students will collect data to see if there is any correlation between shoe length and height. The class will then have a discussion about correlation coefficients and the relationship between correlation and causation.
2. **Notebook (30 min): Exploring the Data II** Students work through a notebook to create and explore explanatory variables and their various correlations to the outcome variable (mortality rate). They summarize these correlations and use this to seed a discussion around assuming potential causation between each explanatory variable and the outcome variable. In doing so, students are proposing potential causative mechanisms that form the basis of hypotheses.
3. **Lecture (20 min): Developing Proposed Explanations -** A short lecture that details the characteristics of good hypotheses (proposed explanations). Examples of contemporary hypotheses are given, and students identify their faults. Groups then formulate their own hypotheses based on this guidance and share out. Finally, the instructor introduces John Snow and explains the 2 predominant hypotheses for cholera in London at the time: waterborne (ingesting contaminated food or drink) and airborne (inhaling polluted air). The different variations of the airborne theories were represented by the 6 Londoners in the beginning.

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

□ **Lesson 3 Spark: Correlations and Linear Regression:** preparation for PROPOSE EXPLANATIONS I – Collect data and explore it with correlations and linear regression

□ **Lesson 3 Notebook: Exploring the Data II:** PROPOSE EXPLANATIONS I – outcome and explanatory variables

□ **Lesson 3 Lecture: Developing Proposed Explanations:** PROPOSE EXPLANATIONS I

Lesson 3 Spark: Correlations and Linear Regression

(30 min)

Big Idea	Goals	Resources
An observed relationship between two variables may be described as a correlation and can be used to explore potential causal relationships.	□ Collect data and construct a scatterplot representing that data. □ Approximate a line of best fit through the data points and describe the relative linearity of the data with the correlation coefficient. □ Distinguish between correlation and causation.	□ Slides to introduce the activity. □ Measuring tapes. □ Google Sheet: Relationship between height and shoe length □ Linear regression simulation
ACTIVITY SUMMARY: Students will collect data to see if there is any correlation between shoe length and height. The class will then have a discussion about correlation coefficients and the relationship between correlation and causation.		

Preparation:

Each group will need a measuring tape and access to the shared Google Sheet.

Roles:

The instructor will lead discussion; Mentors will assist students in collecting and recording data.

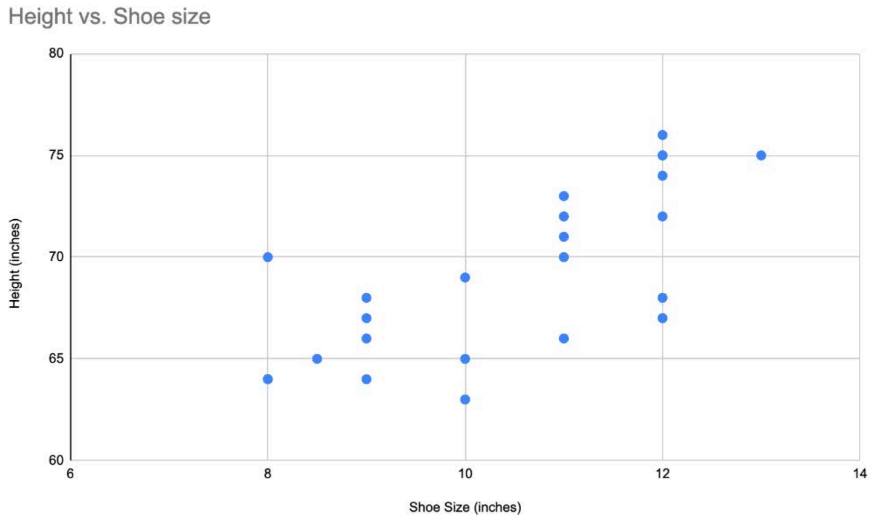
□ Inform students that as a class we will be investigating the idea of correlation. To begin, they will be looking at two variables (shoe length and height) to evaluate if there is any correlation between the variables.

□ Upload a copy of Relationship between height and shoe length, make it editable for everyone, and give students the link to access & edit the Google sheet. Alternatively, the instructor could do the data entry or assign a student to do data entry.

□ Have students determine their height in inches and shoe length in inches (do not use shoe size as the scale is different for men and women shoes) using the supplied tape measures. Students can help each other make the measurements.

□ After taking the measurements, students should report the results to the mentor who will record the values in the spreadsheet.

□ The resulting graph will look something like the graph below. The class data should show a positive correlation between height and shoe length.



□ This is a good opportunity to explain this visualization and to define correlation. For example:

- ○ Height and shoe length are what are called variables. Variables are observations that can vary between instances. In this case, each student is an observation, and height is the variable that can differ from person to person.
- ○ This graph is called a **scatterplot** and is used to visualize the relationship between the two variables.
- ○ Each dot (made by combining the two variables (x = shoe length, y = height) on the scatterplot represents a person.
- ○ The scatterplot allows us to see how all the people compare to each other in one graph.
- ○ Correlation means there is a relationship or pattern between the values of two variables. It could be positive (as x increases, y tends to increase) or negative (as x increases, y tends to decrease) or no relationship.

□ The instructor can ask: Based on the graph, does it appear that there **is a relationship between these two variables? If so, how would you describe it? What conclusions, if any, can you draw from this graph?** It looks like the taller you are, the more likely you are to have bigger feet? This is a positive correlation.

□ Display the Phet simulation and select “Height vs. Shoe Size from the drop down menu at the top.

□ Explain that data scientists often need to describe correlations (relationships) mathematically, not just in words. Point out that the points in the scatterplot form a relatively straight line. From algebra, we know that straight lines have the formula: y=mx+b.

□ Ask for a volunteer to come up to the computer and use the sliders on the right side of the Phet to draw their "best fit line". In other words, a line that best shows the relationship.

□ As the volunteer does this, ask them to explain their method. What strategy are they using? Most likely they will be trying to "split" the points so about half are on one side of the line and the other half on the other side.

□ Point out the resulting equation for the best-fit line (printed on website).

□ When they have finished, point out that a straight line that goes through every point is impossible. Therefore, we have some "error". Explain that we can see these errors by showing the distance between each point and the line by checking the "Residuals" box.

□ However, because some of the errors are positive (above the line) and others are negative (below the line), we square each error, so they don't cancel out in the next step of summing them. Otherwise, the positive and negative errors would add up to zero.

□ Check the "Squared Residuals" box to show this.

□ Point out the area of the squares can be added up to describe the overall error of the line which is shown in the "sum" graph.

□ Explain that our goal in creating a best-fit line should be to minimize the sum of these errors. Fortunately, there are mathematical methods to minimize the squared residuals without requiring us to "eyeball" a best-fit line.

□ Click on the "Best-Fit Line" box to show this calculated line and ask, "How did our volunteer's line compare to the best fit line?"

□ Students should see that the sum of squared residuals is slightly less than the volunteer's.

□ Lastly, point out that unless the points fall on an exactly straight line (which rarely happens in the real world), there will always be some error. We can describe the size of this error with a single number.

□ While the sum of the squared residuals is a single number, we want to be able to have a universal system that compares different best-fit lines on the same scale. Ask the students what this technique is called. Answer: normalization.

□ Explain that the method used to do this normalizes the errors to account for the line's slope. This measurement is called the correlation coefficient or "R-squared" (it compares the sum of the squared errors to the difference of the y-values of each point with the average y value of all the points)

□ For instance, say we measured the shoe length and height of a different group of students. We would get a different graph, with a different line of best fit, which would have a different slope and different error. In order to compare the results between that group and this group we would need to normalize. To do this, we use what is called a correlation coefficient or "R-squared." We won't dive into how the R-square is calculated, but basically it takes into account the residuals, the squared residuals, and the slope of the line of best fit.

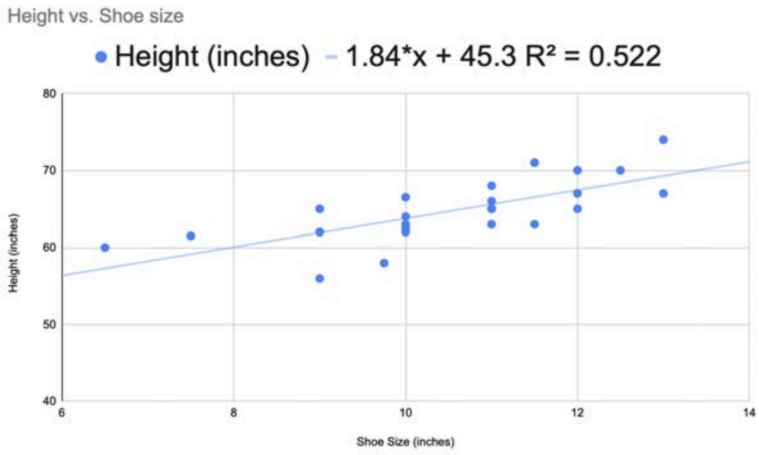
□ Explain that the way to interpret this value is on a scale of 0 to 1. 1 means the data points fall exactly on the best-line, 0 means the data points are randomly scattered and do not follow a line-shaped trend at all. In real data science, the R-squared value will be somewhere in between 0 and 1.

□ There is no "hard and fast" rule for a minimum R-squared for two variables to be "strongly correlated" or "weakly correlated" or "not correlated". For example, in a physics experiment where the environment is closely controlled (a weight on a spring in a lab), an R-squared of 0.85 may not be that great, but in the social sciences where humans behaving in the real world are often the subject, a 0.85 could be a really strong correlation. Therefore, you have to consider what you are studying to interpret the R-squared value.

□ Introduce students to the idea of causation. Causation means that one event causes another event to occur by saying, "Does there appear to be a correlation between shoe length and height?" Answer: Yes, somewhat.

□ Ask, "Does this mean being taller causes you to have bigger shoes/feet or having bigger shoes/feet cause you to be taller?"

□ Field thoughts. Students may propose a variety of ideas related to the idea that body type determines many aspects about a person's shape. There will be a bit of confusion around if shoe length affects height or the opposite. This "chicken vs. the egg" dilemma is natural and highlights the idea that correlations often oversimplify relationships and are easy to interpret as one variable causing the other. However, if two variables are correlated, even if very strongly, we can't say that one causes the other, even though there is a possibility that one may be at least partially causing the other.



□ Display the slide that shows the relationship between shoe length and height and its corresponding R-squared value that indicates a relatively strong correlation. An example is shown on the right.

□ Explain that even though we cannot assume the correlation means causation, looking for correlations is often a good first step in looking to possible causations. We just have to be very careful!

□ Explain that the next notebook will allow them to do this type of exploration around cholera.

Lesson 3 Notebook: Explore the Data II

(30 min)

Big Idea	Goals	Resources
The exploration of mathematical relationships between variables that may be causally linked is a key initial step in the formulation of hypotheses.	□ Generate explanatory variables from existing data. □ Use analysis of DataFrames to identify correlations between explanatory and outcome variables.	□ Notebook 3 □ Large paper and markers
ACTIVITY SUMMARY: Students work through a notebook to create and explore explanatory variables and their various correlations to the outcome variable (mortality rate). They summarize these correlations and use this to seed a discussion around attaching potential causation between each explanatory variable and the outcome variable. In doing so, students are proposing potential causative mechanisms that form the basis of hypotheses.		

Key Talking Points:

□ The approach used here is different from what the students are used to: one explanatory variable and one outcome variable. When doing data science, we are trying to establish causative relationships. They are not already known as is typical with scientific experiments in school.

□ Help students distinguish between the **outcome variable** (dependent variable), in this case mortality rate, and **explanatory variables** (independent variables) and model the use of these terms.

□ The section where students need to calculate the mortality and population density for the south region is a good opportunity to show how to **copy and paste cells** as well as use the **find and replace** feature (copy the cell for “central”, paste it as a new cell, use *find and replace* to change all of the “central” to “south”).

□ There are four explanatory variables that can be obtained from the data: Two explanatory variables are already in the DataFrame (Elevation, Average value of house) but the other two can be calculated:

- ○ Population density = Population / Area.
- ○ Crowded housing = Population / Houses, Inhabited.

Stumbling Blocks:

□ The column headings must be typed exactly as they appear in the DataFrame.

□ The syntax for filtering is tricky. Make sure the students refer to the examples provided in the markdown.

Facilitation Notes:

Encourage breaks for small group discussions throughout this activity to allow for collaboration and to break up the longer, code-heavy sections. If students are struggling with what code to use, encourage them to review the code they have learned so far.

Small Group Discussion (Reserve last 15 minutes to do the following)

The key outcome of this notebook is the summary table of explanatory variables at the end. While students can work through this independently, it is important to reserve the last 10 minutes to get everyone on the same page.

Given the length of this notebook, students may be at various states of completion and may not have analyzed all 4 explanatory variables. However, make sure all students have done at least 2 before moving ahead.

Population density: no correlation.

□ **Average value of house:** negative correlation (weak)

□ **Elevation:** negative correlation (strong)

□ **Crowded housing:** negative correlation (weak)

□ Ask each student to share out an explanatory variable that they analyzed and share results. Use other students to verify results.

□ Record the results of explanatory variable analysis from the table on large paper (same format as above).

□ Students should record or make edits in their markdown cells as you do this.

□ Ask, **"Which explanatory variables are most related to cholera mortality rates?"**

□ Ask, **"Why do you think these relationships exist?"**

Follow up questions:

- ○ Is there something going on? i.e., Is there a mechanism at work?
- ○ Could it be a misleading (spurious) correlation?
- ○ Is there other data that you wished was in the dataset?

Lesson 3 Lecture: Develop Proposed Explanations

(20 min)

Big Idea	Goals	Resources
Hypotheses (proposed explanations) point to explanatory mechanisms, are measurable and provide distinguishable predictions.	□ Formulate hypotheses that demonstrate the characteristics of a good proposed explanation.	□ Slides: 6 theories of the Londoners
ACTIVITY SUMMARY: A short lecture that details the characteristics of good hypotheses (proposed explanations). Examples of contemporary hypotheses are given and students identify their faults. Groups then formulate their own hypotheses based on this guidance and share out. Finally, the instructor introduces John Snow and explains the 2 predominant hypotheses for cholera in London at the time: waterborne and airborne.		

Lecture (15 minutes): Led by Instructor

□ **Slide 1:** Display the “Court for King Cholera” image. **Ask them if it has any additional meaning now.** Hopefully, the students see that because the south districts were poor, this was tied, along with other prejudices and stereotypes, to being the reason why these districts were being impacted so heavily by cholera. (Most specifically, it was meant to imply a link to simple population density, which we have seen in the data as not being true)

□ Explain that public perceptions are rarely driven by objectivity and data. However, we can get closer to the truth by navigating the data and testing ideas in a systematic way. This begins with the creation of competing hypotheses.

□ **Slide 2:** Remind students that a hypothesis is often framed as an “educated guess.” A better definition is a “proposed explanation.” Good hypotheses are:

- Mechanistic (cause and effect)
- Measurable (testable/can be investigated through experimentation)
- Distinguishable (lead to different predictions). This allows them to be proven wrong.

□ **Mechanisms:** Variables should be related to a mechanism that connects A and B. Mechanisms are an important starting place because they affect how we will measure and test the hypothesis. Point out that a characteristic of people is not the same as a mechanism. For example, tying an outcome to race may show strong correlation, but race is a characteristic of people and does not describe a mechanism.

□ **Measurable:** Need to be measurable so we can falsify or reject.

□ **Distinguishable:** Hypotheses should be distinct from each other in that they lead to different predictions (e.g. a disease is spread through polluted air vs contaminated food). How do you know if they are not different enough? Their predicted outcomes are the same or similar. Or more simply, a hypothesis should allow us to make distinct predictions.

□ **Slides 3-8:** One by one, have students identify the fault of each of the six theories presented by the Londoners using the three options: 1. Characteristics of people, not a mechanism 2. Can't measure 3. Can't make predictions (not distinguishable).

Small Group Discussion (10 minutes): Led by mentors:

□ In their small groups, students should develop 2-3 hypotheses for the driving question: "Why is cholera killing some but not others?".

□ At this point, students can move to more specific questions. Specific questions are the ones that seed the hypotheses.

□ Record the hypotheses to share out.

Lecture (10 minutes): Led by Instructors:

□ Have groups quickly share out their hypotheses.

□ **Slide 9:** Introduce John Snow. Being a doctor, he had seen many people suffer and die from cholera. It primarily affected their gastrointestinal system (guts), leading to extreme dehydration and eventually death. Therefore, he was convinced that it had something to do with what people were eating or drinking. Therefore, he focused on the water supplies of the people of London. However, it was hard for him to convince people to change their beliefs, but he knew he could do it with the right data.

□ **Slide 10:** Explain the primary theories present at the time: airborne and waterborne. Explain that, as they saw earlier, there were variants of these theories at the time, much like the different ones developed in small groups.

□ **Slide 11:** Explain that the process we have been using to try to solve a problem is a way of thinking used by data scientists, and scientists in general. We have named it the "Data Science Reasoning Framework"

□ **As you point out the sections of the Framework:** So far, we have identified a problem and asked questions. We observed that cholera is killing people, explored the data to find patterns, and asked questions about it. After this, we proposed some possible explanations that attempt to answer these questions. Moving forward, we are going to test the two main hypotheses (airborne and waterborne) by seeing how well they predict outcomes in different scenarios. Their ability to make accurate predictions will help us assess our confidence in each.

□ If time permits, engage students in small group discussions about how the framework may differ and be similar to other classroom investigations they have done in the past: The typical high school framework highlights the same steps (from question to hypothesis, analysis, results) but it is often unclear why you want to run multiple rounds of analysis (to improve your explanations through multiple rounds of testing against evidence). This may help students see the value of iterative reasoning based on data.

□ **Finally, explain the challenge:** We need to be objective and see what the data tells us, and not let our modern science knowledge bias our work. We need to put ourselves in John Snow's shoes. Even though he had a theory (waterborne) which he thought was stronger than the airborne one, it was not a popular one, and in order to change minds, he would have to develop a convincing argument with data.

Lesson 4 – Find Strategy for Testing Explanations I

Overview

Big Idea:

Hypothesis testing with data and experimentation allows us to test competing explanations side by side. To set this up, we need to establish a strategy, or research design, to examine multiple explanations.

Lesson Summary:

Lesson 4 utilizes a spark activity about home court advantage in the NBA during the Covid lockdown to demonstrate the testing of a null and alternative hypothesis against each other. Following this, the Broad Street Pump Outbreak is used as a case study within the larger cholera case study to test two hypotheses (waterborne and airborne) against each other.

Goals:

□ **DSRF**: Be able to use statistical methods to test competing hypotheses in a way that enables direct comparison.

□ **Computational Thinking**: Run simulations to compare the null hypothesis to the alternative

□ **Coding**: converting continuous data to categorical data, visualizing data in a bar graph.

Activity Summary:

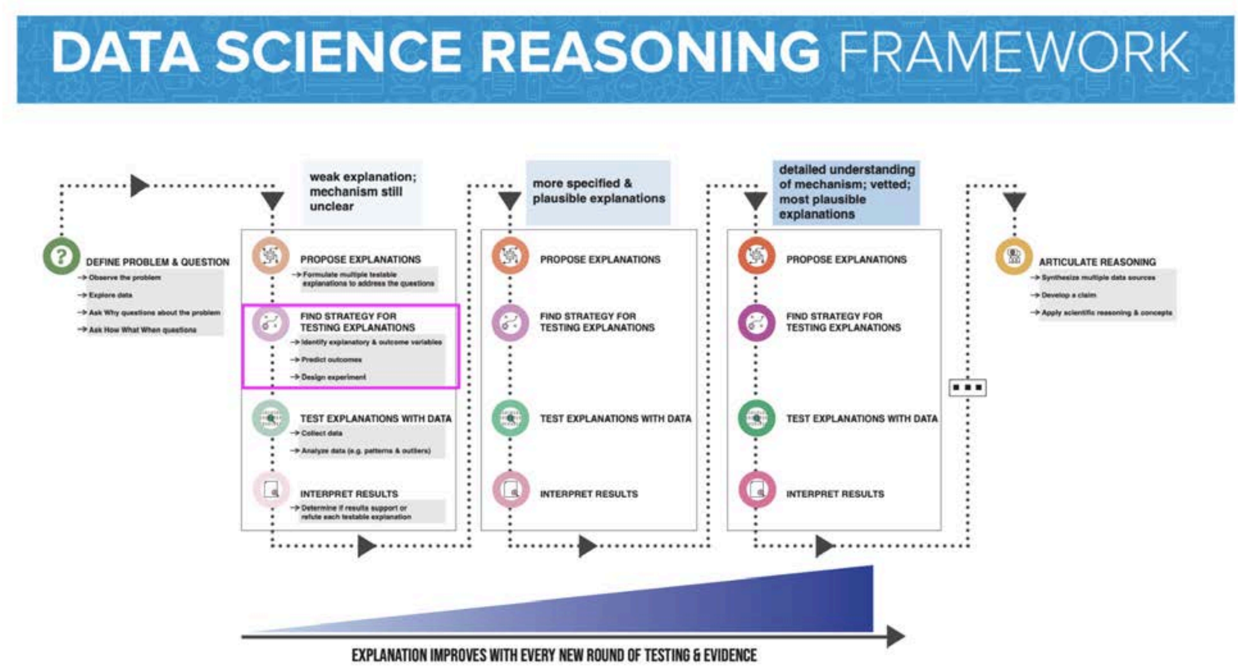
□ **Spark (20 min): Home Court Advantage -** Adapted from Skew the Script. The activity finishes with the groups articulating the null hypothesis for their proposed alternative explanation (hypothesis) created at the end of Lesson 3.

□ **Notebook (30 min):** A Jupyter/Colab notebook that allows students to simulate the results of a number of games where there is no home court advantage. Students visualize the results through a bar graph.

□ **Whole Class Activity (20 min):** Predict the Broad Street Outbreak: Students will make predictions of the spatial distribution of cholera cases for both the airborne and waterborne theories by annotating two paper maps of the Soho neighborhood in London. They will then see the actual spatial distribution of deaths and summarize how the data supports or does not support the airborne and waterborne theories.

□ **Lecture & Spark Activity (30 min):** Discussion of two-way tables (also called contingency tables) and how they may be used to organize data for downstream analysis.

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

□ **Lesson 4a Spark: Home Court Advantage: Correlations and Linear Regression:** FIND STRATEGY FOR TESTING EXPLANATIONS I – null and alternative hypothesis

□ **Lesson 4 Notebook: Simulating the Null:** FIND STRATEGY FOR TESTING EXPLANATIONS I

□ **Lesson 4 Whole Class Activity**: PREDICT OUTCOMES

□ Spark Activity: Two-way Tables: DESIGN EXPERIMENT

Lesson 4 Spark: Home Court Advantage

(20 min)

BIG IDEA	GOALS	RESOURCES
To introduce the concept of a null hypothesis and demonstrate how it is falsifiable and can be used to determine if the data occurred by chance.	□ Interpret data and determine whether to accept or reject the null hypothesis.	□ Spark slides
ACTIVITY SUMMARY: Adapted from Skew the Script. A Jupter/Colab notebook that allows students to simulate the results of a number of games where there is no home court advantage. Students visualize the results through a bar graph.		

Lecture: Introduction

□ Explain the following:

- **Slide 2:** Many of us are familiar with the idea of “home court advantage” where the home team has a better chance of winning because of the support of their fans and playing in a familiar place.
- **Slides 3 and 4:** When the Covid pandemic hit, the NBA was in the middle of their season. In order to finish the season, the league devised a unique solution.
- **Slide 5:** Test and quarantine the players, coaches and staff at Disney World in a “bubble” where no one was allowed to enter or leave.
- **Slide 6:** You would think that because the teams were playing on a neutral court, there would be no home court advantage in the bubble.
- **Slides 7 and 8:** However, the NBA attempted to create an artificial home court advantage by randomly selecting one team as the home team and displaying virtual fans and cheerleaders for the home team and playing sounds from the home team’s arena.

□ **Slide 9:** Ask, “Do you think the home team had an advantage in the bubble and was more likely to win?” Field predictions and explanations.

- Encourage debate as some students will suggest that the virtual fans, cheerleaders and sounds would provide a clear home court advantage, while others will suggest that these attempts to add fake elements of a home court were not enough to make a difference.
- Point out that there are two hypotheses/proposed explanations competing here:
  - There was not a home court advantage causing neither team to have an advantage (no effect).

- There was a home court advantage for the home team causing the home team to win more (some effect).

□ **Slide 10 and 11:** Explain that the NBA played 88 games this way. Of these 88 games, the “home” team won 49 (56%) of them. The “visiting” team only won 39 (44%) of the games.

□ Ask again, “Now do you think the “home” team had an advantage in the bubble?” It is likely that more students than before will think there is a home court advantage. Probe reasoning in order to make evident arguments for both hypotheses. Emphasize that we have two competing hypotheses:

□ **Slide 13:** There was no true home team advantage in the bubble. Home teams won more often by chance alone.

□ **Slide 14:** There was a home team advantage in the bubble. Home teams won more often because they had a true advantage.

□ **Slide 15:** Explain that we will do an analysis to determine which hypothesis is more likely true, but the way we approach competing hypotheses is to first assume one is true and see if there is enough evidence to reject it in favor of the other hypothesis. The hypothesis we assume to be true is the one where the effect we are studying, here the artificial home court advantage, **does not** have an effect.

□ Ask, “Which hypothesis should we assume to be true?”

□ **Slide 16:** Answer: There was not a home court advantage.

□ **Slide 17:** Explain that we call this assumption the **null hypothesis** and we assume it to be true until there is evidence against it. It’s sometimes called the “dull” hypothesis because it describes the situation where nothing new or interesting is happening but this is often easier to assume than to guess the size of an effect. The other hypothesis is called the **alternative hypothesis.** This is our still unproven alternative explanation.

□ **Slide 18:** While we often think of the alternative hypothesis first, our analysis does not try to prove the alternative hypothesis. *Rather, we are seeing if there is enough evidence to “reject” the null hypothesis which would suggest that there is “support” for the alternative hypothesis.* We never use the word “prove” when testing hypotheses since there could be evidence we are missing. Initially, this may be a difficult concept for students. It is worth taking the time to reinforce its importance in data-driven investigations.

Note: Can also be described this way:

Null hypothesis: Explanatory variable does not affect outcome variable.

Alternative hypothesis: Explanatory variable affects outcome variable.

□ **Slide 19:** Explain that we will begin our analysis with a focus on the null hypothesis. Ask, “In a world where there is no home court advantage, what proportion (percentage) of games would we expect home teams to win?”

□ **Slide 20:** Answer: About 50% of the time.

□ **Slide 21:** Remind the students of what was observed in the actual NBA bubble: The home team won 49 of 88 games. This equals 56% of the games.

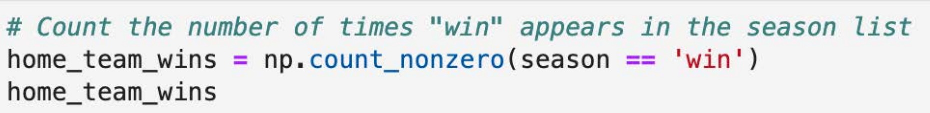
□ **Slide 22:** Explain that our goal will be to determine how likely it is for home teams to win 56% of the time if there is no home court advantage.

□ **Slide 23:** This is the same as deciding which team wins by flipping a coin. We can use computers to simulate 88 games played in a world where there is no home court advantage. Simulations are very powerful tools in data science and science in general because they allow us to predict outcomes in situations we can design -- in this case, a basketball season where 88 games are played when there is no home court advantage. This will allow us to answer the question, “In a world where there is no home court advantage, what are the chances that the home teams win 56% or more of the 88 games played?”

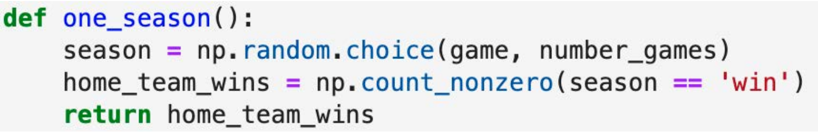
Lesson 4 Notebook: Simulating the Null

(30 min)

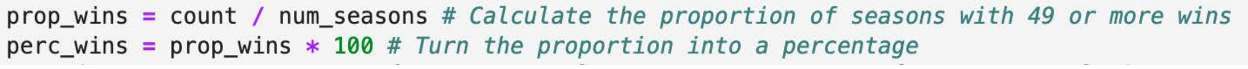
BIG IDEA	GOALS	RESOURCES
Run the simulations to determine how likely an observed outcome (49 wins out of 88 games) is and use statistics to decide if the null can be rejected or not.	□ Compare the number of actual to simulated wins □ Understand p-values as a statistic that tells the difference	□ Notebook 4
ACTIVITY SUMMARY: The activity starts with the groups articulating the null hypothesis for their proposed alternative explanation (hypothesis) created at the end of Lesson 3. Students will then run simulations to determine how many of the simulated games were won “at home.” They then use a p-value to determine if the observed number is unlikely (it it not, hence the null cannot be rejected).		

1. 

```
# Count the number of times "win" appears in the season list
home_team_wins = np.count_nonzero(season == 'win')
home_team_wins
```

2. 

```
def one_season():
    season = np.random.choice(game, number_games)
    home_team_wins = np.count_nonzero(season == 'win')
    return home_team_wins
```

3. 

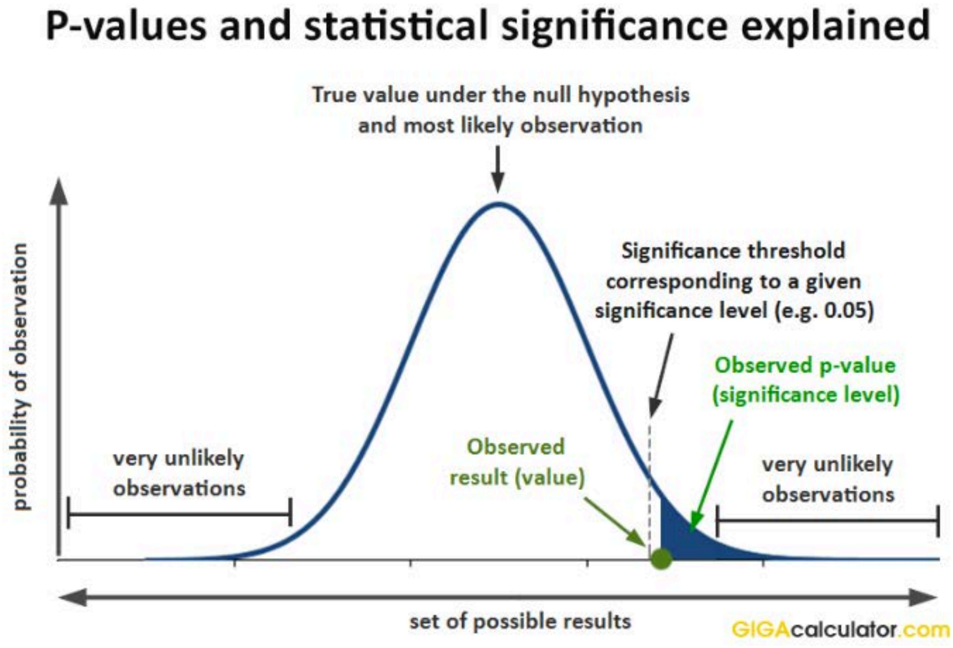
```
prop_wins = count / num_seasons # Calculate the proportion of seasons with 49 or more wins
perc_wins = prop_wins * 100 # Turn the proportion into a percentage
```

Small Group Discussion

□ Ask the students to share out their responses to the last reflection in the notebook: Is there enough evidence to reject the null hypothesis?

□ Field student responses and guide them through their reasoning. It may help to restate the reflection question as “Is the home team winning 49 of the 88 games unusual enough for us to reject the null hypothesis?”

□ The group should conclude that winning 49 of the games in a world where there is no home court advantage is not that unusual. In the simulation, 15% (answers will vary by a percent or two) of the seasons had the home team winning 49 or more of the games. This is not that uncommon (~1 in 8 chance). Therefore, we cannot reject the null hypothesis.

4. Explain:
   - If the p-value is low, what we observed in the real world is unlikely under the null hypothesis and we can reject the null hypothesis. This means we DO have convincing evidence to support the alternative hypothesis.
   - If the p-value is high, what we observe is not that unlikely/unusual under the null hypothesis and we cannot reject the null hypothesis. This means that we do NOT have convincing evidence to support the alternative hypothesis.
   - 
P-values and statistical significance explained
True value under the null hypothesis and most likely observation
probability of observation
Significance threshold corresponding to a given significance level (e.g. 0.05)
Observed p-value (significance level)
very unlikely observations
Observed result (value)
very unlikely observations
set of possible results
GIGAcalculator.com

□ Explain that ~15% of seasons in the simulation had 49 or more home team wins. Finish by asking, "How low would this percentage have to be to convince us that there is a home team advantage?" In other words, "How low does the p-value have to be in order to reject the null hypothesis?" Field responses but do not offer any answers.

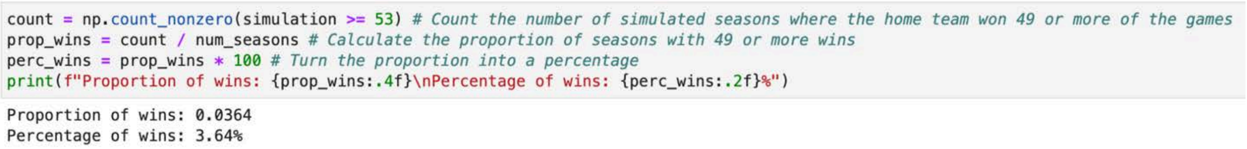
□ **Provide the answer:** A common p-value that is used is 0.05 or 5%. In other words, if what we observe has a less than 5% chance of being caused by random chance, we have evidence that we can reject the null hypothesis and support the alternative.

□ Therefore, based on the p-value of our analysis, there is not enough evidence to reject the null hypothesis and we cannot say there was a home court advantage in the NBA bubble.

Note: Language is very important when making such statements:
   - We cannot "accept" the null hypothesis because we are already assuming it is true.
   - We can "reject the null hypothesis" if the p-value is sufficiently low.

    - If the p-value is low enough, there is “evidence to support the alternative hypothesis”, but we cannot “prove” the alternative hypothesis.

- □ The students will likely want to know how many home team wins would correspond to a p-value of 0.05. Encourage them to go back into the notebook and investigate this by changing the last cell to see how many home teams wins produces a p-value < 0.05:



```
count = np.count_nonzero(simulation >= 53) # Count the number of simulated seasons where the home team won 49 or more of the games
prop_wins = count / num_seasons # Calculate the proportion of seasons with 49 or more wins
perc_wins = prop_wins * 100 # Turn the proportion into a percentage
print(f"Proportion of wins: {prop_wins:.4f}\nPercentage of wins: {perc_wins:.2f}%")
```

```
Proportion of wins: 0.0364
Percentage of wins: 3.64%
```

Lesson 4 Whole Group Activity: The Broad Street Pump

(20 min)

Lesson 4 Whole Class: Annotate Broad St. Basemap with Predictions

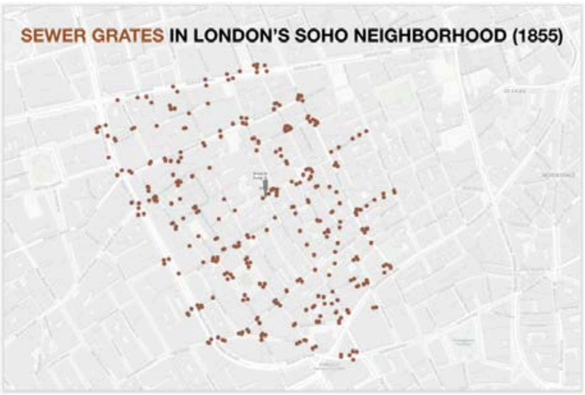
Big Idea	Goals	Resources
Hypotheses can be used to make predictions of outcomes. Different hypotheses should result in distinguishable/ different predictions.	□ Using established hypotheses, make predictions of outcomes.	● Soho neighborhood poster maps: sewer grates & pumps (original maps FYI) ● Pins to put on each map ● Paper ● Pen ● Web-based mapping software: https://kepler.gl/demo Data for mapping (kepler): GeoJSON files: ● pumps.geojson ● sewergrates_ventilators.geojson ● deaths_by_bldg.geojson ● deaths_nd_by_house.geojson
Activity Summary: Small groups will make predictions of the spatial distribution of cholera cases for both the airborne and waterborne theories by annotating two maps of Soho (one with the Broad St pump; the other with the locations of sewers). They will then see the actual spatial distribution of deaths and summarize how the data supports or does not support the airborne and waterborne theories.		

Part 1:

□ Explain that there was a significant outbreak of cholera in the Soho neighborhood of London in 1854.

We can use this outbreak to test the two theories (cholera transmitted through water vs through air) by predicting the spatial patterns for each one and comparing the predictions to what actually happened.

□ Bring the whole class to the wall with two large maps of Soho. Split the class into two groups: One for each map. Ask the students to describe what they see. They should point out the presence of streets, a pump and many sewer grates.



BROAD ST PUMP IN LONDON'S SOHO NEIGHBORHOOD (1855)

□ Ask, **"How can this map be used to show the predicted outcomes from the two hypotheses?"** If the students struggle to come up with responses, point out that we have 2 mechanisms for the spread of cholera. What are they? Answer: drinking contaminated water (pump) and inhaling polluted air (sewers for airborne).

□ Provide the students each with a couple of pins and have them each place them on the pump or sewer map where they think deaths would occur if the waterborne theory was correct. Encourage students to work collectively: it does not make sense that all the pins would be on the pump. Just as there are natural variations in the data, so should there be in their predictions.

□ Repeat with pins for the airborne theory (near sewers for the sewer map / could be anywhere for pump map).

□ To finish, ask students to compare the two maps: "Do the hypotheses result in distinguishable (different) predictions?" They should note that yes, the pattern of pins on each map are notably different.

□ Next, let's see how these predicted patterns fit with the real patterns in London.

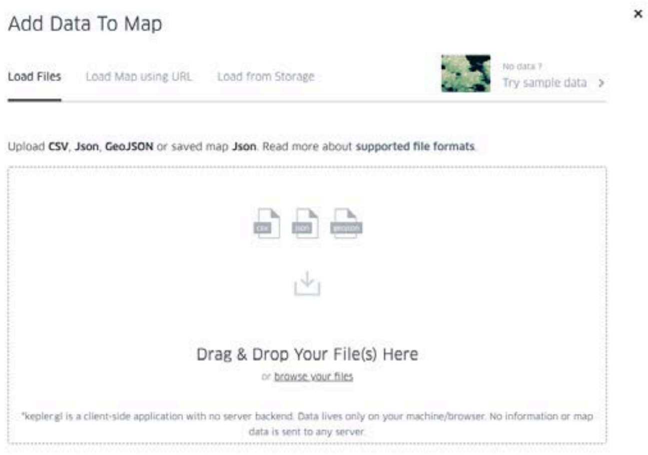
Part 2: Lecture (Can also be done independently by students if time allows)

Mapping (Kepler)

□ Explain that John Snow and his colleagues mapped the deaths of people during the Soho outbreak. While they recorded the locations of victims on a map by hand, we can view the same data digitally:

□ Navigate to the kepler.gl site (https://kepler.gl/demo).

□ Upload the pumps.geojson and sewergrates_ventilators.geojson files by dragging and dropping (or via the 'browse your files' button). Explain that the resulting visualization is the same as the paper map.



Add Data To Map

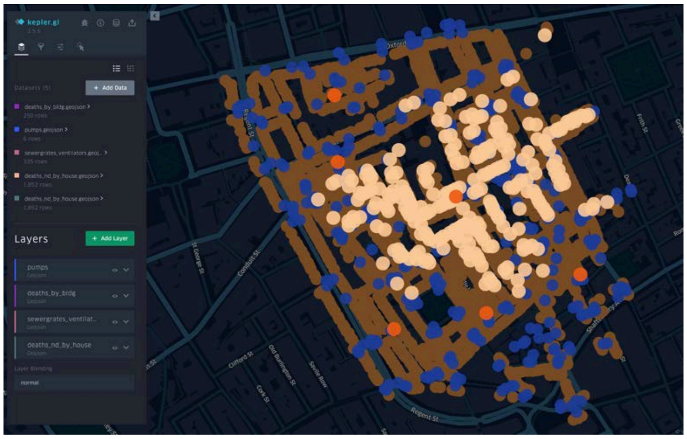
Upload **CSV**, **Json**, **GeoJSON** or saved map **Json**. Read more about **supported file formats**.

Drag & Drop Your File(s) Here

or browse your files

*kepler.gl is a client-side application with no server backend. Data lives only on your machine/browser. No information or map data is sent to any server.

□ Repeat for deaths_by_bldg.geojson (only deaths) and deaths_nd_by_house.geojson (all houses; note that this file contains the number of deaths from 0-18 in each house; see data documentation).



Datasets (5)

- deaths_by_bldg.geojson > 250 rows
- pumps.geojson > 6 rows
- sewergrates_ventilators.geoj... > 325 rows
- deaths_nd_by_house.geojson > 1,852 rows
- deaths_nd_by_house.geojson > 1,852 rows

Layers

- pumps Geojson
- deaths_by_bldg Geojson
- sewergrates_ventilat... Geojson
- deaths_nd_by_house Geojson

Layer Blending: normal

□ If students are doing independently, otherwise, demonstrate: Allow the students a few minutes to explore the map. They may also change the order and color of the layers to better see the data. It is helpful for all houses (deaths_nd_by_house.geojson) to be set as the bottom layer (see image above).

□ After the students have explored the data, have them pause and ask, **"Which patterns do you see that either support or do not support each of the theories?"** Have the students stick to objective observations of only the data in the map as opposed to trying to "fill in the gaps."

□ Chart responses of the "Hypothesis Analysis" on a piece of paper in the following format (example observations provided, but make sure the "Does not support" are similar to the ones below):

AIRBORNE	WATERBORNE
Inhalation of a poison given off by dead or contaminated organic matter like sewage which enters the body through the lungs and poisons the blood (breathing in polluted air).	Ingestion of “excretions of the sick” which contain a living organism which infects the gastrointestinal system (guts), e.g. by preparing food after changing diapers of infected infants without washing hands.
Supports □ Deaths appear to be clustered around open sewer grates (because people breathed in polluted air from the sewers)	**Supports** □ Deaths appeared to be centered around the Broad Street Pump (since people drank contaminated water from the pump)
Does not support □ Deaths do not appear centered on sewer grates.	**Does not support** □ Many deaths are far away from the Broad Street Pump, closer to other pumps.

□ Conclude with the question, “**We can see patterns in these visualizations, but how can we quantitatively test the 2 hypotheses using the location of deaths in order to objectively assess which hypothesis made better predictions of the outbreak?”** This does not need to be answered but can be used to preview the analysis done in the notebook.

Lesson 4 Spark Activity: Two-way Tables

(30 min)

Big Idea	Goals	Resources
Students will be introduced to two-way tables and learn what they are and why they are used.	□ Use two-way tables to determine if two variables are correlated.	□ Slide show □ Freakonomics: Correlation vs. Causality Video □ Google form survey □ Student responses from survey (template and example)
Activity Summary: Students will begin by filling out an “either/or” survey. They will then be introduced to two-way tables, what they are, and why they are used. Next, students will act out the construction of a two-way table using their responses to two questions from the survey.		

Lecture:

□ **Slide 2:** Show the video clip from Freakonomics. Afterwards, explain that we have seen that correlation is a valuable tool for exploring data and developing hypotheses, but we have to be careful moving forward in our search for the answer to the question: “Why is cholera killing some people and not others?”

□ **Slide 3:** Explain that our use of scatter plots is a good way to identify potential correlations between two variables, which can help us propose explanations. However, if we create a scatter plot and find a nice-looking correlation, our work is far from done. We need to think deeply and perform experiments that test our proposed explanations, and we need to do so without bias. In other words, we should not assume a proposed explanation is correct.

□ If time permits, this could be a good opportunity to revisit the Data Science Framework. Have students make the connection that they are using scatter plots to explore the data in order to formulate a testable explanation, analyze the results, and then decide if they should accept or reject their null hypothesis. It reinforces that they are conducting multiple experiments in order to solve the larger question.

□ As seen with the NBA bubble activity, we need a way of testing these hypotheses and calculating the chance that the connection that we are seeing could actually be due to random chance (coincidence).

□ **Slide 4:** Show difference between quantitative and categorical variables.

- Quantitative data is numerical. We have seen a lot of quantitative data: years, population, deaths, etc. Because quantitative data is numerical, it can be put in a certain order (e.g. from least to greatest).
- Categorical data is categories or group labels (like favorite sports team, ethnicity, month of the year) that do not need to have a certain order.

□ **Slides 5-6:** Use the examples to check understanding. Students will often get tricked by ZIP code. Just because data is numerical does not mean it is quantitative. ZIP codes are numbers but they do not have a specific order. They are simply labels and therefore, categorical.

□ Ask, “In the NBA bubble example, were we examining the relationship between quantitative or categorical data?” Answer: Both. Number of wins in a season (quantitative) and home team or away team (categorical).

□ Explain that we often need to see if two categorical variables are related. A classic example of this is: people who smoke vs not and people who got lung cancer vs not.

□ Explain that we will be exploring how to analyze the relationship between categorical variables, but we first need some data to analyze. We will be gathering data about each of you.

□ Provide the Google survey and have students complete. Instruct students that they must choose the one in each pairing that they identify with the most or prefer.

□ After all students complete the survey, ask the students to brainstorm pairs of variables from the survey that they think may be associated (related). This is also an opportunity to probe their rationale and highlight the perceived mechanisms. At this point, it is not necessary for students to articulate a mechanism. In fact, the variables in the survey do not have any obvious causal mechanisms. However, it is fine to ask, “Why do you think those two are related?”

□ Explain that we will be using their responses to two of the questions: Chicago side (Southside or Northside) and preferred baseball team (White Sox or Cubs) to investigate.

□ Ask, “What relationship do you think there might be between these two variables?”

□ In Chicago, the White Sox are the “Southside” baseball team and the Cubs are the “Northside” team. Therefore, a likely response will be that the side where you live is associated with your preferred baseball team.

□ Ask, “In this case what are our explanatory/independent and outcome/dependent variables:

- **Explanatory/Independent:** Side.
- **Outcome/Dependent:** Team.

□ **Slide 7:** Explain that we will use something called a **two-way table**, also known as a **contingency table** or **crosstab** to compare the variables.

- ○ Two-way tables allow us to make a two-way comparison – here Southside vs Northside and Sox vs Cubs. Put differently, they allow us to compare the outcome in one group (Sox vs Cubs) to that of another group (city side), contingent on a condition. The condition is the potential explanation that you want to explore. In this case, the condition is the Chicago side.
- ○ They can then be later used to determine if the two variables are correlated (is there a link between the two variables or are they independent?)

	Southside	Northside	
White Sox	Individuals who are Southside and White Sox fans	Individuals who are Northside and White Sox fans	Total White Sox fans
Cubs	Individuals who are Southside and Cubs fans	Individuals who are Northside and Cubs fans	Total Cubs fans
	Total Southside	**Total Northside**	**Grand Total**

- □ Discuss the table with students:
    - o For instance, look at the sample two-way table. Ask what two variables are being compared? What do you think the relationship is between these two variables? Does one variable “explain” the other? If so, which one? What question could this table help us answer? (Is a person’s preferred baseball team influenced by (contingent upon) whether they identify as Southside or Northside?)
    - o Explain that the “outcome variable” is the variable on the left side or in the rows. (In this case preferred baseball team). The “explanatory variable” is on the top of the table or in the columns. (In this case, Chicago side). Some students may be familiar with the terms dependent and independent variables, in which case the dependent variable would be the “outcome variable” and the independent variable would be “explanatory variable”.
    - o So, with this table we are asking if there is a relationship between the side you identify with and your preferred Chicago baseball team. That could then lead to the question “Does your identity with the Southside or Northside cause you to be a White Sox or Cubs fan?” — **which this table can NOT tell us**. It just allows us to explore if the two variables are associated or not.
    - o Notice that the table displays *groups within groups*…for example Northsiders that are White Sox fans versus Northsiders that are Cubs fans.
- □ Explain that we will use their responses from the survey to see if there is evidence to support the hypothesis: **The side a person identifies with is associated with their preferred baseball team.** More specifically, you’re more likely to be associated with the team that’s close to where you live.

Activity Setup:

□ Create an open space in the room where you can mark out a two-way table on the floor. Alternatively, you may wish to use a whiteboard.

□ Mark off a 2-row, 3-column grid on the floor with tape.

□ Place sheets of paper with the variable labels as shown:

	SOUTHSIDE	NORTHSIDE	
White Sox			All White Sox
Cubs			All Cubs

□ Provide the students with name tags. They are to write their preferred baseball team that they chose for the survey on the nametags and put them on their shirts.

□ Using the Google form spreadsheet, populate the student data. Refer to the template and example spreadsheets as a reference. The “Template” tab will automatically pull the responses for side and team and construct the null and observed two-way tables for the **teachers’ reference.**

□ Have the students go to the “All White Sox” or “All Cubs” square depending on their name tag.

□ Tell the students to remember which Chicago side they selected in the survey but not to reveal it to anyone. Explain that you (the teacher) will sort them into 4 groups on the two-way tables by being the “dull null”.

□ Ask, “Under what assumption will the “dull null” be sorting students?”

- ○ Answer: The dull null is going to assume there is not an association between side and team, i.e. that there are equal numbers in each quadrant.

□ Use the spreadsheet as a reference to quickly do the following.

Note: Use the following script exactly. The order and language is key to the rationale for constructing a two-way table under the null hypothesis. The numbers used below are referencing the “Example” tab. Replace them with what was obtained in the “Template” tab. Values are rounded off to the whole number for simplicity.

- “I see that we have 28 students, so 7 students go in each quarter.”

□ Explain that we have constructed a two-way table for what we would expect if the null hypothesis is true. In other words, the number of people in each of the groups within groups if there was no association between side and team.

□ In order to test the null hypothesis, we will compare the table to what we actually observed.

□ Ask the students to note how many people are in each square (7).

□ Ask the students to move to the square that represents their answer to the survey.

□ Explain that our two-way table now shows the observed data. The key question is “How different does it look compared to the null hypothesis table?”

□ Allow the students to make this assessment. Encourage them to focus on the number of people in each square, not who those people are. Do they see any squares that are noticeably different? Which one(s)?

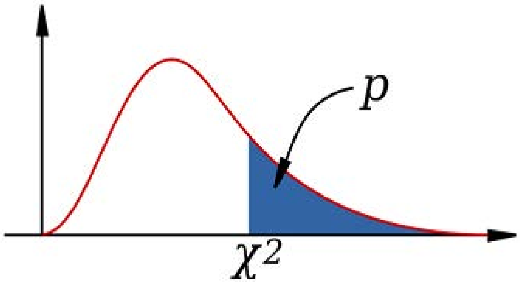
□ Ask, “Based on what you see, does it look like the actual results are different enough to support the alternative hypothesis (that side is associated with team preference)?” Field responses which will vary depending on the data.

□ Return the students to their seats and display the spreadsheet “Template” tab on the screen where they can see the counts for the null and observed two-way tables. Again, ask them to digest the data by asking, “Is the difference between the two tables beyond what you would expect from random chance alone?”

□ Remind the students that data scientists prefer mathematical methods to answer questions like these. In the NBA bubble example, we compared what was observed to what we would expect to see if each game was determined by what amounts to a coin flip.

□ Explain that the method used to determine if the difference between the null and observed data is called a “chi-squared test.” The template shows the squared difference between each of the four groups in the null and observed tables, and the sum of those values, known as the “chi-squared test statistic”, abbreviated χ^2.

□ Explain that the chi-squared test, like what we did for the NBA bubble, calculates a p-value that tells us the chance that the difference we see between the null and observed is due to random chance. The calculation itself is beyond the scope of this lesson, but you can illustrate this by drawing a chi-squared distribution on the board similar to the following:



□ Point out the p-value calculated on the spreadsheet. In the small groups, have students discuss what the p-value suggests. Mentors, use this discussion in order to check understanding and clear up misconceptions.

Lesson 5 – Analyzing Results I

Overview

Big Idea:

Comparing the predicted outcomes to the results of the data experiment provides valuable feedback about the accuracy of the current state of each proposed explanation, driving revision and refinement of the proposed explanation.

Lesson Summary:

Lesson 5 allows students to consider the results of the hypothesis testing in Lesson 4 in order to make edits to their Hypothesis Analysis charts. In doing so, they establish that the waterborne theory is consistent with the data while the airborne theory is not. However, as a follow-up, new data is introduced about the time component of the Broad Street outbreak that shows the need for further experimentation.

Goals:

□ **DSRF**: Be able to interpret the results of hypothesis testing.

□ **Computational Thinking**: Be able to assess competing hypotheses based on the results of a data experiment.

□ **Coding**: Understand how experiments with data are conducted; interpret the chi-square statistic, and create data visualizations.

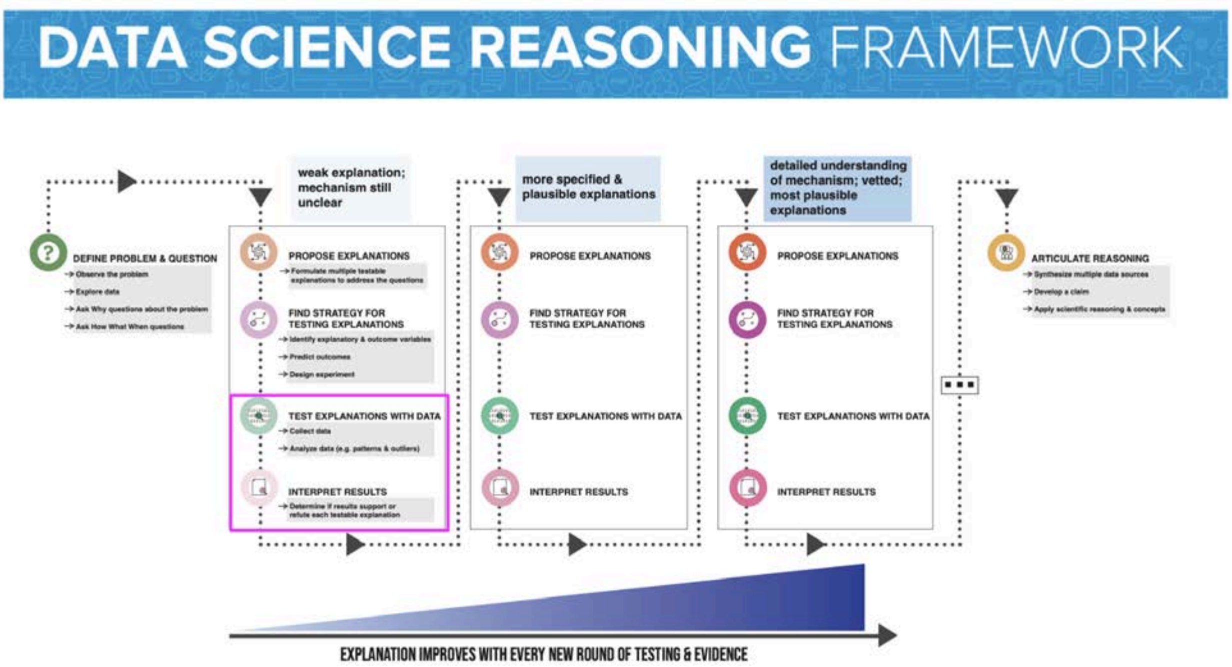
Activity Summary:

□ **Notebook (45 min):** Testing Explanations I: Students use two-way tables and hypothesis testing to compare the airborne and waterborne theories. This involves first calculating the counts in each subgroup, constructing the two-way tables and performing a chi-square analysis for statistical significance. Finally, students visualize the results with a bar graph and reflect on its meaning.

□ **Small group discussion (20 min):** Small groups review the results of the two-way table analysis from the previous notebook. These results are added to the Hypothesis Analysis sheet and the two hypotheses are updated to account for results of the data experiment that did not support them.

□ **Lecture (15 min):** A short lecture summarizes the Broad Street outbreak, showing John Snow’s as well as a modern analysis of the data. Finally, the epidemiological curve for the outbreak is considered for the impact of time on the investigation.

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

□ **Lesson 5a Notebook 5:** TEST EXPLANATIONS I

□ **Lesson 5: Small Group Discussion: Analyze Results of Data Experiment:** INTERPRET RESULTS I

□ **Lesson 5 Lecture: The Broad St. Pump Handle:** TEST EXPLANATIONS & INTERPRET RESULTS I

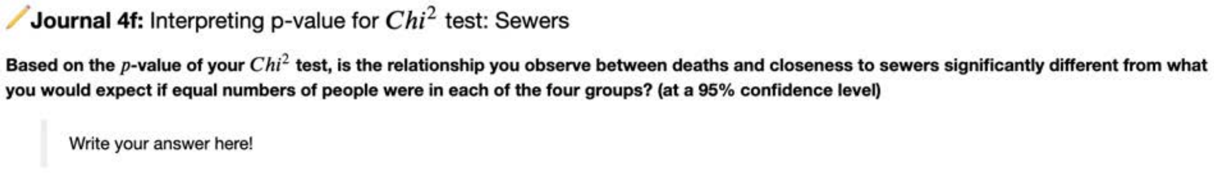
Lesson 5 Notebook 5: Testing Explanations I

(45 min)

Big Idea	Goals	Resources
Data experiments allow us to test multiple explanations side-by-side and gather data on their predictive power.	□ Construct two-way tables and visualizations to measure the statistical significance of the relationship between the outcome variable and an explanatory variable.	□ Notebook 5
ACTIVITY SUMMARY: In this notebook, students use two-way tables and hypothesis testing to compare the airborne and waterborne theories. This involves first calculating the counts in each subgroup, constructing the two-way tables and performing a chi-square analysis for statistical significance. Finally, students visualize the results with a bar graph and reflect on its meaning.		

Journal Stop Points:

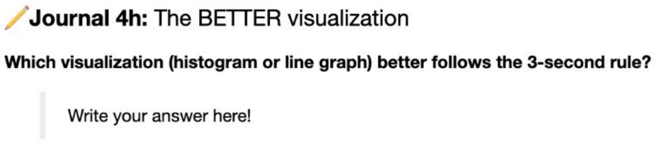
While all journal entries are important for the students to reflect on, for the following, stop the group once all have completed each one to discuss for 2-3 minutes. This will help the students to not lose sight of the "big picture" of what they are doing.



Journal 4f: Interpreting p-value for Chi^2 test: Sewers

Based on the p-value of your Chi^2 test, is the relationship you observe between deaths and closeness to sewers significantly different from what you would expect if equal numbers of people were in each of the four groups? (at a 95% confidence level)

> Write your answer here!



Journal 4h: The BETTER visualization

Which visualization (histogram or line graph) better follows the 3-second rule?

> Write your answer here!

Key Talking Points:

□ p-values of <0.05 suggest that the pattern we are seeing has < 5% chance of being random.

□ A p-value < 0.05 gives us a 95% confidence that what we are seeing is “statistically significant”. But we could still be wrong 5% of the time, i.e. we could think that the proposed mechanism is significant while it is actually random.

□ The “cutoff” (distance from sewers or pump) that is chosen to separate the data into two groups is up to you, but it does have an impact on the results of the analysis for statistical significance.

  ○ For sewers, the cutoff is 40 feet (12.2 meters)

  ○ Using 12.2 meters for the pumps will not result in clear visualizations as there are relatively few people that live that close to the Broad Street Pump (BSP).

  ○ The students can choose the cutoff for the pumps, but it should be greater than sewers because the pumps are more spread out. Encourage the students to look at the map and decide the cutoff that would make sense. Allow them to wrestle with this decision, but hint at what makes a person near or far from the pump. Likely, they will consider the closeness to other pumps. That is, if another pump other than BSP is closer, that would be considered “far” (100-200 feet produce obvious differences in the 2 groups).

□ The line graph of the two-way table can be interpreted as follows:

  ○ If there is not a significant difference between the groups, the lines will show similar slopes.

  ○ If there is a significant difference, the line will show different slopes. They could both still have positive or negative slopes, but if one is positive and the other negative, the difference is even greater.

Coding Advice:

□ Students actually don’t have to do a lot of writing of code in this notebook. The “Investigating the Broad Street Pump” section should use code from the “Investigating the Sewers” section with small changes.

□ Encourage the students to use copy and paste to do the BSP investigation. They can then use the **find and replace** feature of the notebook to change “sewer” to “pump”.

Lesson 5 Small Group Discussion: Analyze Results of Data Experiment

(15 min)

Big Idea	Goals	Resources
The results of data experiments can be used to assess the plausibility of hypotheses and provide information to revise them.	□ Use information obtained in previous analyses to revise the hypotheses.	□ "Hypothesis Analysis" chart from Lesson 4 □ Notebook 5
ACTIVITY SUMMARY: Small groups review the results of the two-way table analysis from the previous notebook. These results are added to the Hypothesis Analysis sheet and the two hypotheses are edited to account for results of the data experiment that did not support them.		

Mentors:

Lead a discussion to summarize the results of the test done in the notebook. Use the following questions (answers in bold) to do so. As a follow-up to each question, ask whether the response supports or does not support the airborne or waterborne theory. Record the answer to the question in the appropriate square on the Hypothesis Analysis sheet developed in Lesson 4.

- □ Was there a significant difference in the number of deaths between those living near sewers and those living far from them? **No, p-value = 0.29 > 0.05**
- □ Which theory does this support or not support? **Does not support airborne.**
- □ Was there a significant difference in the number of deaths between those living near the Broad Street Pump and those living far from them? **Yes, p-value ≈ 0.**
- □ Which theory does this support or not support? **Supports waterborne.**

AIRBORNE	WATERBORNE
Inhalation of a poison given off by dead or contaminated organic matter like sewage which enters the body through the lungs and poisons the blood.	Ingestion of "excretions of the sick" which contain a living organism that infects the gastrointestinal system (guts).
Supports □ Deaths appear to spread out from a central point	**Supports** □ Deaths appeared to be centered around the Broad Street Pump. □ Deaths do not radiate from the pump symmetrically. □ **Statistically significant larger numbers of deaths in close walking distance to the Broad Street Pump**
Does not support □ Deaths do not appear centered on a single sewer grate. □ **No statistically significant increase in deaths for those living close to sewers**	**Does not support** □ There are deaths far away from the Broad Street Pump, closer to other pumps.

□ At this point, explain that as new information comes in, we can revise the hypotheses. Allow the students to review the information not supporting each hypothesis and to change the hypothesis to account for this contradictory information. This is a good time to once again revisit the Data Science Reasoning Framework.

- ○ For the airborne theory, gas emanating from the sewers does not appear to be associated with more nearby deaths. Ask, **"How could people still believe that inhaling polluted air is the cause of cholera?"** Responses may vary but try to draw out the idea that the gasses could be coming from contaminated water. Edit the airborne hypothesis statement to include this.
- ○ For waterborne, ask **"If the Broad Street Pump's water is the source of the outbreak, how are people that live closer to other pumps getting cholera?"** This is much easier to explain: People are moving around and may get their water from multiple pumps – and BSP is not the only source of cholera. The waterborne hypothesis should not need to be edited.

Lesson 5 Lecture: The Broad St. Pump Handle

(5 min)

Big Idea	Goals	Resources
Single data experiments are not sufficient to prove/disprove hypotheses.	□ Understand how multiple data sources and experiments strengthen an argument.	□ Lesson_5_Lecture slides
ACTIVITY SUMMARY: A short lecture giving a summary of the Broad Street outbreak, showing John Snow's as well as a modern analysis of the data. Finally, the epidemiological curve for the outbreak is considered for the impact of time on the investigation.		

Slide 1: Choosing to record cholera deaths for each house on a map of the Soho neighborhood is one of the first and most famous spatial data visualizations in public health.

Slide 2: The creator of this map was John Snow and it clearly shows that deaths centered around the Broad Street Pump. This is an excellent example of a visualization that very clearly communicates a message.

Slide 3: However, in some versions of this map, John Snow drew a strangely shaped outline around the pump. **"What do you think he meant to show?"** Responses will likely point out that it circles most of the deaths. However, it does not include all of the deaths. Point out that it is not circular. As a hint, mention that there were several other pumps in the neighborhood.

Answer: The Broad Street Pump is the closest pump by walking distance for the houses within the outline. It is not a circle because people must walk by street to get to the pump.

Slide 4: Explain that Snow's data convinced city officials to stop usage of the Broad Street Pump by removing its handle on September 8, 1854. Ask, **"Did this end the outbreak?".** It is natural for students to think, yes, it did because the outbreak effectively ended a week or two after the handle was removed, but the point of this graph is to consider our perspective of the outbreak from not just a spatial perspective, but a time (temporal) one as well.

This graph is known as an "epidemic curve" and it shows the progression of the outbreak with deaths over time, with a decline in deaths coinciding with the removal of the pump. From this data, it can be argued that the outbreak was already in decline even before the handle was removed. In addition to the epidemic curve, people were either leaving or avoiding the area of the outbreak even before the Broad Street pump was identified as the source. The news of many deaths in the area was enough data for people to naturally protect themselves through their behaviors. In other words, the outbreak would have likely ended even if the handle had not been removed.

Ask, does this suggest the waterborne theory was not correct? This is not a right/wrong question but is meant to encourage students to see the need to consider other information, such as time, in testing and analyzing these theories.

To close out the lecture, explain that John Snow's map did not suddenly convince everyone of the waterborne theory. In science, there are no "mic drop" moments. More data from more situations will need to be considered. This is another opportunity to show students the Framework and how this step fits in.

Slide 5: As we start to accumulate more and more data, we will be able to gain more knowledge that allows us to refine our proposed explanation (or hypothesis), and eventually this will lead us to articulating reasoning and developing a claim.

Lesson 6 – Testing Explanations II

Overview

Big Idea:

Further testing explanations with new data is essential to expanding the body of evidence needed to promote certain explanations over alternatives.

Lesson Summary:

Lesson 6 conducts another experiment using data from the South London Experiment where mortality rates were recorded for individuals receiving their water from two different water companies: before and after one of the companies started sourcing its water from upstream where sewage entered the river (treatment), while the other company continued to source dirty water (control). The students are first introduced to the concept of a difference-in-difference research design through a carnival-type game. This is then formalized through a lecture and notebook using a two-way table and chi-square analysis.

Goals:

□ **DSRF**: Be able to design and conduct an experiment that compares a treatment and control group where the treatment group has undergone an intervention during the course of the experiment.

□ **Computational Thinking**: Practice statistical significance and scientific visualization to determine if contaminated water is a plausible explanation for cholera deaths

□ **Coding**: Create a two-way table, generate a p-value and visualize the results

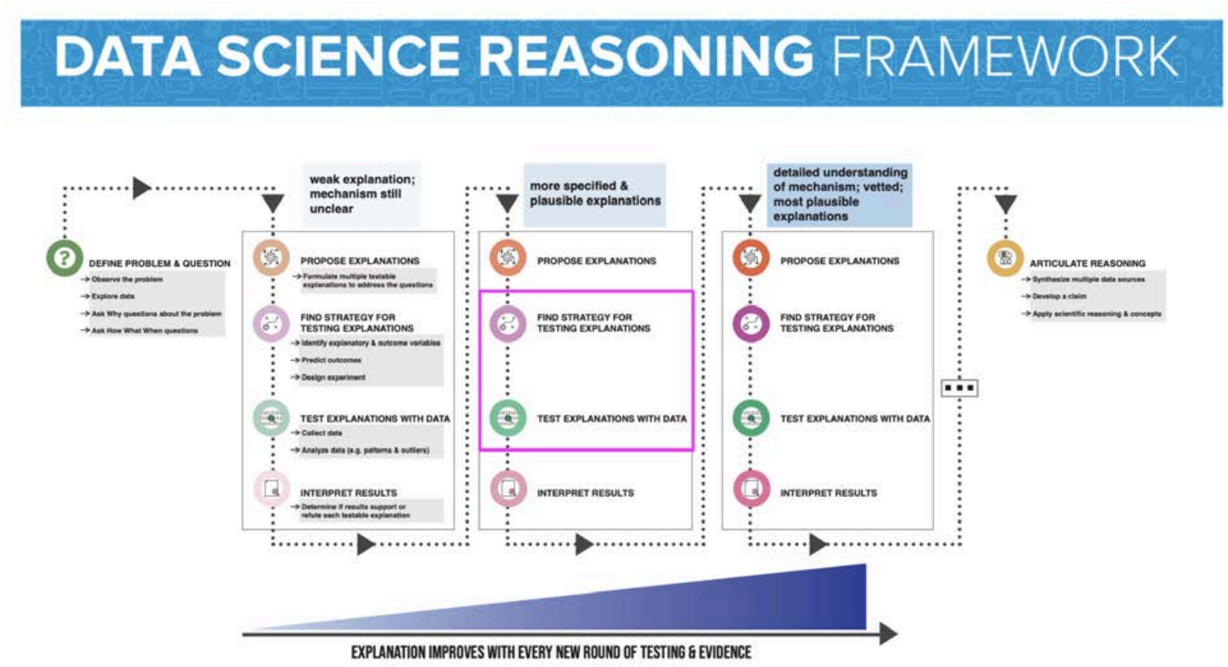
Activity Summary:

□ **Spark (20 min):** Students will play a table version of the classic carnival game (milk bottles) to explore the idea of a difference-in-difference research design. Students will compete in two teams, going through two rounds to knock down as many cans as they can with bean bags. Their scores are collected. For round two of the game some cans will be swapped out and replaced by cans with magnets on their bottoms without their knowledge (the treatment). The instructor will continue to collect data.

□ **Lecture (20 min):** Difference-in-Difference Primer: A brief lecture discussing experimental design as it applies to the South London experiment. Parallels are made to the carnival game played in the spark activity and how two-way tables with a time component can be used to implement a difference-in-differences research design to further test the waterborne theory.

□ **Notebook (45 min):** Students are introduced to the South London Experiment through a brief lecture and prepare for the notebook by discussing why the scenario of the two water suppliers in this region were conducive to a natural data experiment. The students then analyze the characteristics of the candidate groups to use as the experimental and control groups in order to identify groups that are best to compare.

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

- □ **Lesson 6 Spark: Difference-in-Difference Intuition:** FIND STRATEGY FOR TESTING EXPLANATIONS II
- □ **Lesson 6 Lecture: Difference-in-Difference Primer:** FIND STRATEGY FOR TESTING EXPLANATIONS II
- □ **Lesson 6b Notebook: The South London Experiment:** TEST EXPLANATIONS WITH DATA II

Lesson 6 Spark: Difference-in-Difference Design

(20 min)

Big Idea	Goals	Resources
To explore the idea of difference-in-difference model in a fun way (is there a difference in an outcome pre-post and treatment-control?)	□ Explain the concept of difference-in-difference. □ Understand what a treatment is.	□ Slides to introduce the spark activity. □ 3 sets of the bean bag toss game*, magnets, metal board, data collection sheets □ Data table and graph for activity □ Candy as a prize for students. □ Masking tape for line □ For mentors: how DID design works

ACTIVITY SUMMARY: Students will play a table version of the classic carnival game (milk bottles) to explore the idea of a difference-in-difference research design. Students will compete in two teams, going through two rounds to knock down as many cans as they can. Their scores are collected. For round two of the game some cans will be swapped out and replaced by cans with magnets on their bottoms without their knowledge (the treatment). The instructor will continue to collect data.

**If you are ordering this, here is a version of the game. You can also just have 2 games and put th magnets under one of them.*

Preparation:

Set up the bean bag game and place a line of tape on the floor.

Roles:

Instructor will lead discussion; All hands for the game; TAs will swap out (on the sly) regular cans for cans with magnets:

□ Explain to students that they will be playing a tabletop bean bag toss game, where there is a stack of tin cans and the students must stand behind a line and toss a bean bag into the cans to try to knock down as many as they can. Let them know that we will be using this activity to investigate a statistical concept, but that we will discuss the specifics later.

□ Break students into two teams (may be fun to have students come up with silly team names) and let them know that they will be competing against each other for a small prize (candy). When teams cheer for their members and are invested, the impact of the intervention is more fun.

□ Instruct students to take turns throwing the bean bags to try to knock over the cans. If the instructor would like, the teams can have a couple of minutes to practice before they begin the data collection.

□ Have two mentors at the two tables set up the game on top of the magnetic boards. Come up with an appropriate excuse for why we are using the boards…sound dampening, flat surface, protect the surface, etc.

□ After each student throws, two other mentors, one for each team, records the number of cans knocked down in the spreadsheet. Alternatively, the instructor can record the results.

□ Have mentors restack the cans between students. Do NOT allow students to restack themselves.

□ For group A, half way through, the group of students have the mentor swap out the cans and replace them with the magnetic cans. (TAs should do the substitutions on the sly so that students are unaware that the substitution has taken place).

□ The students will then continue with the rest of the round.

□ Have the students return to their seats and display the graph of the data from the spreadsheet.

□ The instructor can ask questions such as Why do we use line graphs? What information do they help us visualize? Hopefully students know that they help us see trends in our data. Then ask What trends do we notice about this data? Who won? Give the winning team half the candy.

□ Was one team winning the entire time? What happened here? (pointing to the dip in Team A's performance in round 2) Why do you think that happened?

□ Instructors can then let students in on the secret that Team A, unbeknownst to them, actually received a treatment that Team B was not exposed to (the weighted cans). Give the losing team the other half of the candy since the game was rigged.

□ Instructors can then discuss the idea of difference-in-difference and show students on the graph the difference-in-difference (slide provided for discussion).

□ After explaining the concept, have students either look at the graph that they created or the sample graph provided in the slide show and identify the following:

  ○ Treatment group - group A

  ○ Control group - group B

  ○ Intervention: magnets

  - ○ Bean bag count before and after the intervention (time trend for both groups)

□ Instructors can explain how by examining the trends in our data we were able to pinpoint a difference in the trends between the two groups (difference in difference). This is a clue that the treatment had an impact.

Lesson 6: Difference-in-Difference Primer

(20 min)

Big Idea	Goals	Resources
Testing the proposed explanations with new data is essential to expand the body of evidence needed to find certain explanations more plausible than their alternatives.	□ Design an experiment, in theory, that compares groups where one group receives a treatment at some point during the experiment.	□ Lesson 6 Slides □ Poster paper: DiD Prediction sheet
ACTIVITY SUMMARY: A brief lecture discussing experimental design as it applies to the South London experiment. Parallels are made to the baseball game and the carnival game played in the spark activities and how two-way tables with a time component can be used to implement a difference-in-differences research design to further test the waterborne theory.		

□ **Instructors may want to begin by reminding students of key points from previous lectures and why more data strengthens an argument (see DSRF).**

□ **Slide 2:** So at this point, we've gathered some evidence to assess the plausibility of the waterborne and airborne theories in mediating the spread of cholera. Now, we are refining our hypothesis by examining evidence from a natural experiment, which will allow us to assess the plausibility of the waterborne theory specifically.

□ **Slide 3:** London had several water companies that supplied water to residents throughout the entire city.

□ **Slide 4:** When the cholera outbreak of 1849 occurred in South London, there were three companies that supplied water to the area: Southwark, Vauxhall, and Lambeth water companies. Southwark and Vauxhall were actually part of a merger and managed by the same parent company, while Lambeth was separate. With this in mind, we will explore a natural experiment that occurred in South London at the time of the outbreak.

□ **Slide 6:** Explain that John Snow chose to use South London as a place to do a data experiment to further test the waterborne theory against airborne theories. Ask, **"Why do you think South London a good place to perform this experiment?"** (remember higher death rates in the South from one of their earlier notebooks)

□ **Slide 7:** Show the slide and explain each of the bullets. Explain that South London was an ideal place to observe the impact of water on death rates as part of a natural experiment.

□ **Slide 8:** Remind students of what a treatment group, control group, outcome variable and intervention is. Ask the students to identify each of these for the South London Experiment.

□ **Slides 9-13:** Reveal each answer as it is discussed.

□ **Slide 14:** State that because the water source for the Lambeth water company changed in 1854, we can apply a Difference-in-Difference analysis where we examine the outcome variable for both groups before and after this change.

□ **Slide 15:** Remind students of how two-way tables work using the Baseball (White Sox vs Cubs) example from before.

□ **Slide 16:** Emphasize that the proposed explanation, in this baseball case, White Sox vs Cubs, defines the two groups to be compared.

□ **Slide 17:** Link this back to the bean bag can game.

□ **Slide 18**: Explain that when we integrate time and a change that occurs during the experiment (a before and after), we can use a Difference-in-Difference analysis of the two-way table. Then, ask **"What would the two-way table look like for the South London Experiment?"** Field responses, refer back to Slide 14 if needed.

□ **Slide 19:** Reveal the answer. Show the two-way tables with values for 1849 calculated from the previous notebook. Remind the students that before performing a data experiment, predictions of the results for each hypothesis must be done. Explain that since we have 2 competing theories, we need to predict the results for 2 tables. Each table assumes that its hypothesis is correct (waterborne would predict a drop of deaths in predominant Lambeth areas while for airborne the difference in water suppliers would not make a difference).

Lesson 6 Notebook: The South London Experiment

(45 min)

Big Idea	Goals	Resources
Good experiments compare groups that are very similar in all aspects except that one is subject to a treatment while the other is not.	□ Articulate the design of the South London Experiment.	□ Lesson 6 Slides □ Notebook 6

Activity Summary: Students prepare for the notebook by discussing why the scenario of the two water suppliers in this region was conducive to a natural data experiment. The students then analyze the characteristics of the candidate groups to use as the experimental and control group in order to identify groups that are best to compare.

Notebook Notes

Tasks:

□ Divide the subdistricts into two groups:

  ○ Subdistricts that got more than 50% of their water from Lambeth.

  ○ Subdistricts that got more than 50% of their water from Southwark & Vauxhall.

□ Determine the total number of cholera deaths in 1849 for each of the two groups.

□ Calculate the death rate per 1,000 for each group.

□ Calculate the total number of deaths in 1854 for each of the two groups (Mostly Lambeth and Mostly Southwark & Vauxhall).

□ Create a two-way table with the total deaths for each group in 1849 and 1854.

□ Determine if there is a statistically significant difference between the two groups.

□ Visualize the differences in the two groups with a line plot.

Lesson 7 – Analyzing Results II

Overview

Big Idea:

Considering the results of new experiments as well as previous ones builds a body of knowledge that determines the confidence in each explanation.

Lesson Summary:

Lesson 7 is a short lesson where students analyze and discuss the results from the South London Experiment, further strengthening their confidence in the waterborne theory.

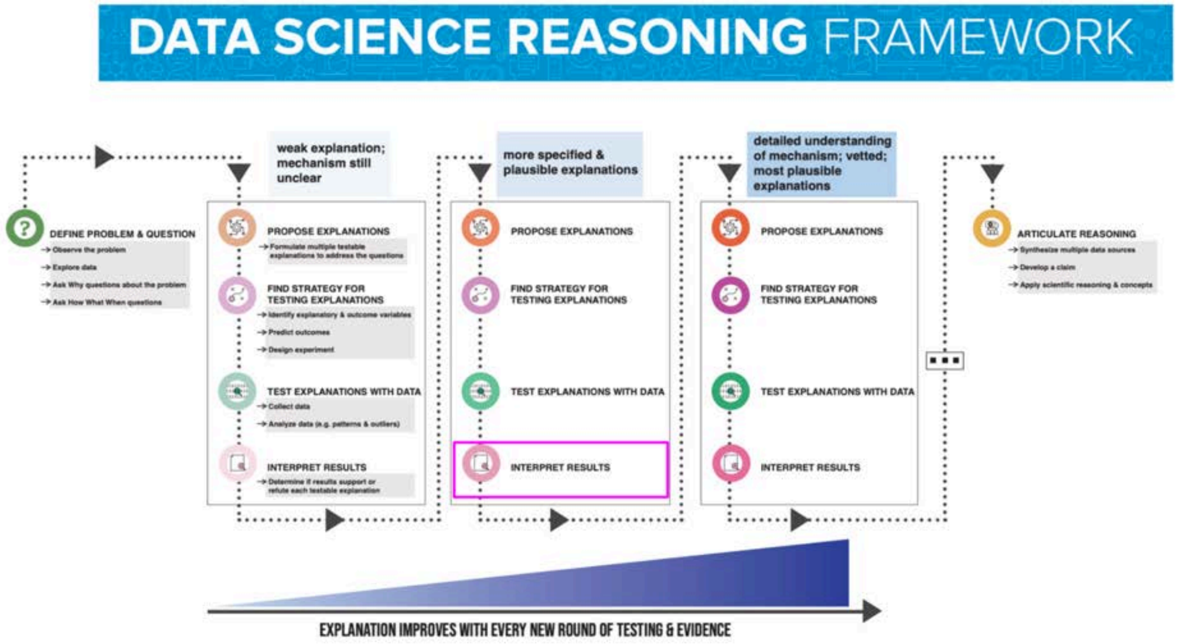
Goals:

- □ **DSRF**: After the initial round of interpreting the results of the Broad Street pump analysis, students now make sense of a new set of results from the South London natural experiment.
- □ **Computational Thinking**: Be able to interpret the results from a difference-in-differences analysis.
- □ **Coding**: Interpreting the results of the notebook in the previous lesson.

Activity Summary:

Small group discussion (15 min): Small groups analyze the results of the South London Experiment and add to their Hypothesis Analysis charts.

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

□ **Lesson 7 Small Group Discussion: Notebook 6 Results –** INTERPRET RESULTS II

Lesson 7 Small Group Discussion: Notebook 6 Results

(15 min)

Big Idea	Goals	Resources
Considering the results of new experiments adds to a body of knowledge that determines the confidence in each explanation.	□ Interpret the results of the difference-in-difference analysis.	□ Notebook 6
ACTIVITY SUMMARY: Small groups analyze the results of the South London Experiment and add to their Hypothesis Analysis charts.		

□ Have the students open Notebook 6 on their devices and go to the visualization they generated at the end of the notebook.

□ Call on students to summarize what this bar graph shows. Do not try to elicit meaning or interpretation. This is simply meant to remind students of the experiment.

□ Explain that we are looking to see if there is a difference between what we observe and what we would expect to see if the intervention (moved water supply) did not have an effect on the outcome (death rate).

□ Explain that the situation where the intervention does not make a difference between the two groups is called the "null hypothesis"

□ Ask, **"What would the bar graph look like if the null hypothesis were true?"** Allow students to discuss amongst themselves for a minute. Then, field responses and come to the conclusion that the Lambeth bar for 1849 would be similar to the Lambeth bar for 1854. This is what is actually observed for the Southwark & Vauxhall groups, indicating 1854 was similar to 1849 and that those subdistricts experienced similar outcomes in both years. This is what we would expect (and what we want) out of the control group.

□ Ask, **"Was the differences between the groups statistically significant?"**

□ Students should look for the p-value in their notebooks and see that it is essentially zero (0.00000) and therefore statistically significant.

□ Ask, **"What can we add to the chart based on the results from the South London Experiment?"** Record responses on the chart. Response should be similar to those highlighted below:

Airborne	Waterborne
Inhalation of a poison given off by dead or contaminated organic matter like sewage which enters the body through the lungs and poisons the blood. After Broad Street Pump: contaminated water can give off the poison gas.	Ingestion of “excretions of the sick” which contain a living organism and which infects the gastrointestinal system (guts).
Supports Deaths appear to spread out from a central point	**Supports** Deaths appeared to be centered around the Broad Street Pump. Deaths do not radiate from the pump symmetrically. Statistically significant increase in deaths near the Broad Street Pump. **Statistically significant difference in death rates between similar groups with different water sources. Rates are lower for areas with clean water than dirty water after supplier moves from dirty to clean source.**
Does not support Deaths do not appear centered on a single sewer grate. No statistically significant increase in deaths for those living close to sewers. **People that share the same air have different death rates and this difference is statistically significant.**	**Does not support** There are deaths far away from the Broad Street Pump, closer to other pumps.

Facilitation Notes:

□ If students are struggling to answer the above questions, giving them a chance to think, pair share may encourage participation.

Lesson 8 – Articulate Reasoning

Overview

Big Idea:

Articulating an argument involves espousing one proposed explanation over alternatives by leveraging the evidence (statistical analysis of data) available and applying sound logic and reasoning to demonstrate higher confidence in one proposed explanation over alternatives.

Lesson Summary:

This is the culminating activity for students. Students will learn the components of a strong scientific argument (tested in multiple rounds and supported by the evidence) and how to distinguish it from unfounded opinions and implausible explanations (rejected by multiple rounds of evidence). Students will then learn how to summarize and organize their arguments in the form of a poster and will practice defending these arguments against counter-arguments.

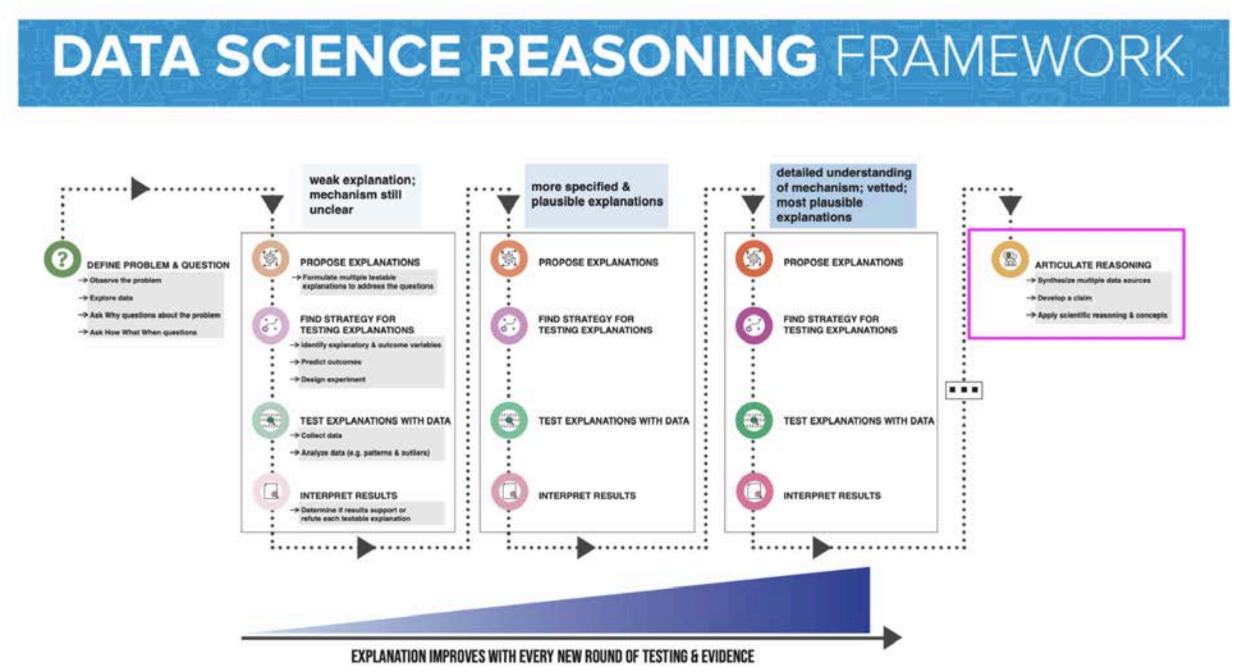
Goals:

- □ **DSRF**: Students will be able to identify the elements of a strong argument, organize and summarize their arguments, and defend their arguments against counter-arguments.
- □ **Computational Thinking**: Synthesis of statistical analysis and data visualization.
- □ **Coding**: No coding but making sense of the programmed results.

Activity Summary:

- ❑ **Lecture (10 min):** A brief lecture distinguishing unfounded opinions from evidence-based claims.
- ❑ **Small Group (30 min)**: Small groups will work together to build a poster that will be used to argue for the waterborne theory over the airborne theory.
- ❑ **Gallery Walk (30 min)**: Students will present their posters and defend their arguments against counter-arguments.

REFERENCE TO DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

- □ **Lesson 8 Lecture: Changing Minds –** ARTICULATE REASONING
- □ **Lesson 8 Small Group: Construct Posters –** ARTICULATE REASONING
- □ **Lesson 8 Gallery Walk: Persuade the Londoners –** ARTICULATE REASONING

Lesson 8 Lecture: Changing Minds

(15 min)

BIG IDEA	GOALS	RESOURCES
Articulating an argument involves espousing one proposed explanation over alternatives by leveraging the evidence (data) available and applying sound logic and reasoning to demonstrate higher confidence in one proposed explanation over alternatives.	□ Distinguish opinions from evidence-based claims. □ Identify the elements of a strong argument	□ Lesson 8 Lecture Slides
ACTIVITY SUMMARY: A brief lecture distinguishing opinions from evidence-based claims (scientific arguments).		

Slide 2: Students have probably used Claim, Evidence, Reasoning in their science classes at school. It is a good way to explain a science idea to someone. What makes opinions different from evidenced claims…

Slide 3: Arguing based on opinions that aren't grounded in evidence or that are based on values ('I like chocolate ice cream better than vanilla') is different from scientific claims. Those are based on evidence and have been tried to be rejected over and over again where scientists try to poke holes in differing explanations in order to arrive at the truth together. These disagreements are needed to make sure countervailing evidence is not overlooked.

Slide 4: Compare and contrast opinions and evidence-based claims. Highlight: looking for countervailing evidence, not just supporting.

Slide 5: Remind the students of our driving question: "Why is cholera killing some but not others?" Ask, "**Remember the answers of the Londoners: Were those opinions or evidence-based claims?**" Field responses but don't provide an answer.

Slide 6: Explain that it depends on the situation. Ask, "**How is 2020s different than 1850s?**" See is students can suggest that we actually know what causes cholera and how it is spread in 2020+, but they didn't in 1854.

Slide 7: Explain that we have an established scientific theory called the "germ theory of disease" that was established in the 1890's, but this theory was not widely accepted in 1854. Ask the question, "**The answers of the Londoners: Were those opinions or evidence-based claims?**" again. Get at the difference between having some evidence for airborne in the 1850s but with multiple rounds of testing in the next 10-20 years there was more evidence against than for airborne theories.

Slide 8: Show the slide.

Slide 9: Detail the parts of a good argument. Bullets 2-4 are the claim evidence reasoning, but what makes it an argument is considering alternative explanations and critiquing them. This may include providing evidence that contradicts them or questioning the reliability of data sources among other.

Slide 10: Explain that we will be transported back to 1854 London and our mentors will revert back to their 19th century selves like in the first week. Each group's task is to construct a poster to use as a visual for arguing for the waterborne theory of cholera over the airborne theory. The Londoners will use their 19th century perspective to argue for the airborne theory.

Slide 11: Provide the directions for constructing the poster. Emphasize the following:

□ Planning: sketch the layout of the poster beforehand. Don't just jump in and start putting things on the paper. The poster is meant to present your argument clearly and concisely. Simplicity is better.

□ The poster is just a visual aid. Your reasoning and explanation of what it says must be verbalized. Develop a plan for how you would like to present your argument.

□ Study the Londoners. Each will come with a different perspective, none of which support the waterborne theory. You will need to counter their argument with data and reasoning that counter their claim. This could include pointing out weaknesses in their argument.

Lesson 8 Small Group Activity: Construct Posters

(30 min)

BIG IDEA	GOALS	RESOURCES
Articulating an argument involves espousing one proposed explanation over alternatives by leveraging the evidence (data) available and applying sound logic and reasoning to demonstrate higher confidence in one proposed explanation over alternatives.	□ Organize and summarize final arguments. □ Create a scientific poster.	□ Printouts of data and visualizations from notebooks □ Londoner role playing cards
ACTIVITY SUMMARY: Small groups will work together to build a poster that will be used to argue for the waterborne theory over the airborne theory.		

This task is not linear as each group will take a different approach to constructing their poster. However, the following guidance will be helpful.

Layout

It is wise to sketch the layout of the poster on a separate sheet of paper before constructing the poster. Encourage the students to “tell a story”. In other words, the argument should naturally flow from top to bottom. Also, clarity is important as a cluttered and disorganized poster is hard to follow. Students may want to use the Framework as a reference to help them organize their posters.Even though the design of the poster is up to the students, you should guide them to include elements that will allow them to use the poster to argue with the Londoners. The key components of the poster include (also reference the print-out of the DSRF here):

□ Clearly state the question/problem.

□ State your claim.

□ Provide data and visuals (evidence)

□ Explain your reasoning.

□ Provide a critique of the alternative explanation(s) including:

  ○ How it may not be supported by the data.

- How the data that supports it is not good.

Visuals

Printouts of key data and visualizations from the case study will be provided to each group, along with the Data Science Reasoning Framework. It is up to the students to determine which ones to use and how to integrate them into the poster.

It also may be helpful to annotate them to point out key elements.

Countering the Airborne Theory

The group needs to be prepared to counter arguments of the other theories or anecdotal evidence.
Each Londoner will come with a unique perspective and angle of argument. Some Londoners will rely on anecdotal evidence, others will have their own data.

It will be helpful for the group to spend some time predicting what evidence and reasoning may be brought in support of the other theories and how to best counter them with their own data.

Lesson 8 Gallery Walk: Persuade the Londoners

(30 min)

BIG IDEA	GOALS	RESOURCES
Articulating an argument not only involves defending it with supporting evidence but also being prepared to remain plausible in the face of countervailing evidence.	❑ Defend an argument against a counter-argument using data. □ Defend an argument against a counter-argument by highlighting weakness in counter-argument.	□ Londoner role playing cards for mentors
ACTIVITY SUMMARY: Students will present their posters and defend their arguments against counter-arguments.		

EACH mentor will take on the perspective of their Londoner from Lesson 1. Prior to this activity, reread the background information on your Londoner as described in the Londoner role playing cards.

Londoners:

□ Londoners each go to a separate group. Begin by reintroducing yourself (name and occupation).

□ Remind the group which of the airborne theory you support.

□ Allow the groups to present their argument.

□ Groups present an "elevator pitch" to argue for the waterborne theory over other theories.

□ Look for opportunities to find holes in the group's argument or point out when they are not being clear.

□ Also look for an opportunity to argue for your theory and challenge the group's argument. This can be done at a lull in the conversation when the topic comes up.

□ Score each group on a scale of 1-5 by giving a point for doing each of the following:

  ○ Waterborne theory is clearly described and includes a mechanism.

  ○ Evidence from the Broad Street Pump data experiment is used.

  ○ Evidence from the South London data experiment is used.

  ○ Londoner's theory is acknowledged.

- Argument by Londoner is countered either using data or highlighting weakness in argument (anecdotal, lack of data to support, flawed reasoning)

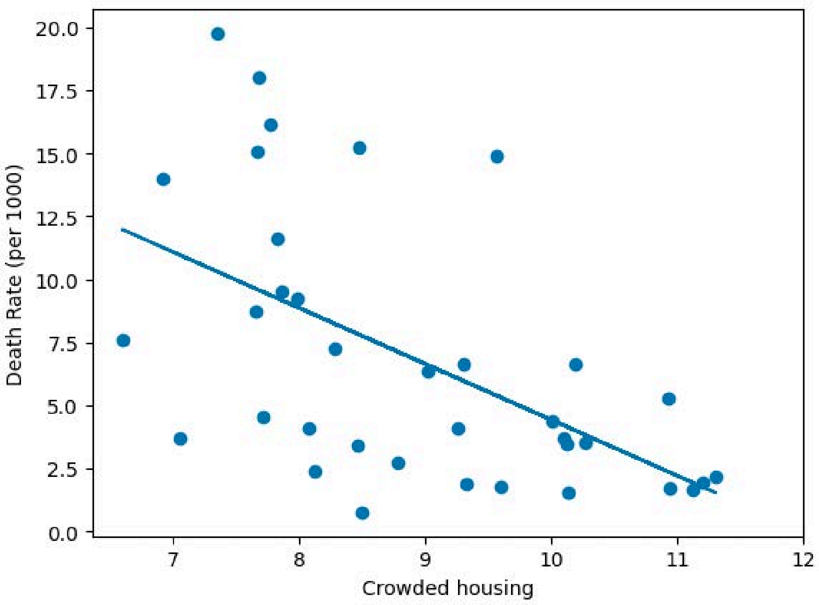
The following are the evidence and reasoning for each Londoner and the corresponding counter argument that is an exemplar response from the students:

Francis Moon, Lord Mayor of London (Airborne-Poverty)

Airborne-Poverty Theory: Under unsanitary conditions and severe overcrowding, the usually non-contagious airborne disease becomes contagious (improved sanitary conditions would reduce disease risk).

Argument: London's population has exploded. Lack of housing has been a major problem. Crowded housing where unclean living conditions exist and polluted air gets trapped must be to blame.

Counter: There is actually a negative correlation between people per house and death rate. The data does not support the mayor's claim.

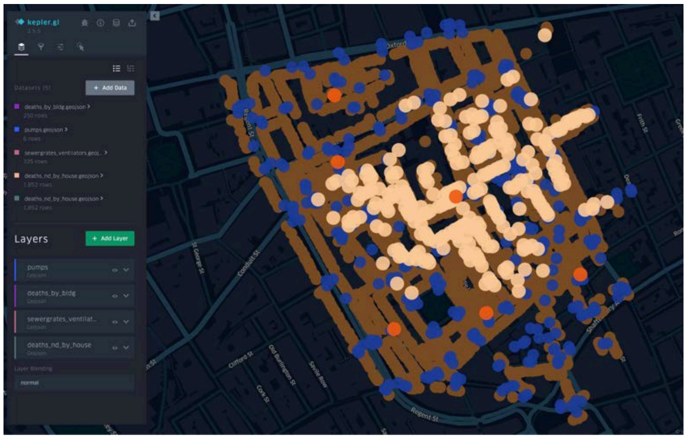


Edwin Chadwick, Sanitation Commissioner

Theory: Locally Airborne-Noncontagious: Locally polluted air (e.g. coming from decomposing bodies of pest field or sewers). People get sick by inhaling locally polluted air (e.g. coming from the pest field: move away from pest field) or by ingesting something contaminated.

Argument: Cholera occurs in areas with polluted air. Smells are evidence of the disease. Cholera is spread through the air and living close to sewers or the pest field increases the chance of getting the disease. In addition, polluted air can contaminate food or drink and people who eat this food can also get sick.

Counter: A spatial analysis showed no relationship between sewer grates and cholera deaths (see below). The South London experiment showed that neighbors who breathed the same air but drank water from different suppliers got sick, which suggests that the mechanism works through sewage in water, not polluted air to water.



Night-Soil Man

Theory: Generally Airborne-Noncontagious: Generally polluted air (source is somewhere from earth's surface)

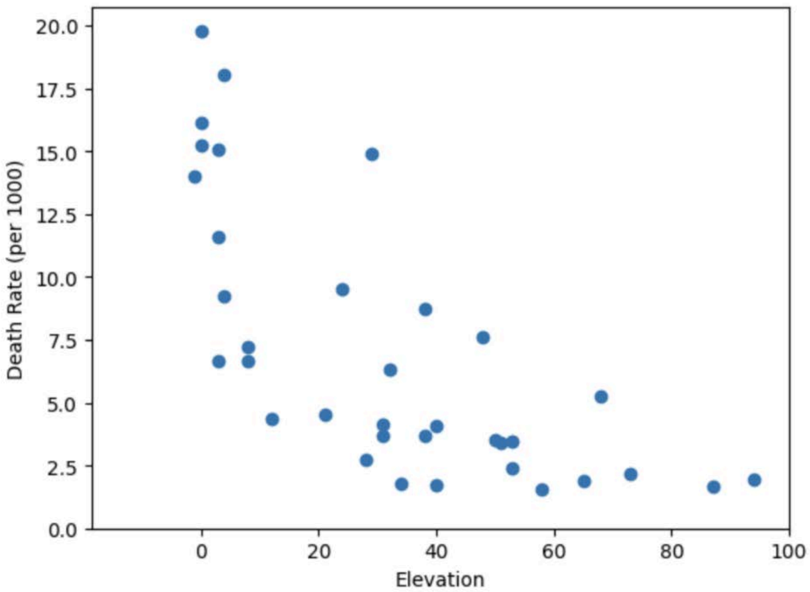
Argument: You don't believe that cholera could be spread by water contaminated with human waste. You work with the stuff, and you and your fellow night-soil people are not at any higher risk of cholera as the general public.

Counter: Students could ask you if you have any data to support your claim (you don't). The claim is based on anecdotal evidence.

William Farr, Registrar General of London

Theory: Airborne-Elevation: Polluted air from water surfaces gets trapped in low elevation areas.

Argument: He has collected lots of data connecting lower elevation to higher cholera to show that polluted air gets trapped in these areas, making it more likely for people to get cholera there. He has found a strong negative correlation between elevation and cholera mortality (i.e. more deaths at lower elevations). Provide the printout of this scatterplot as evidence:



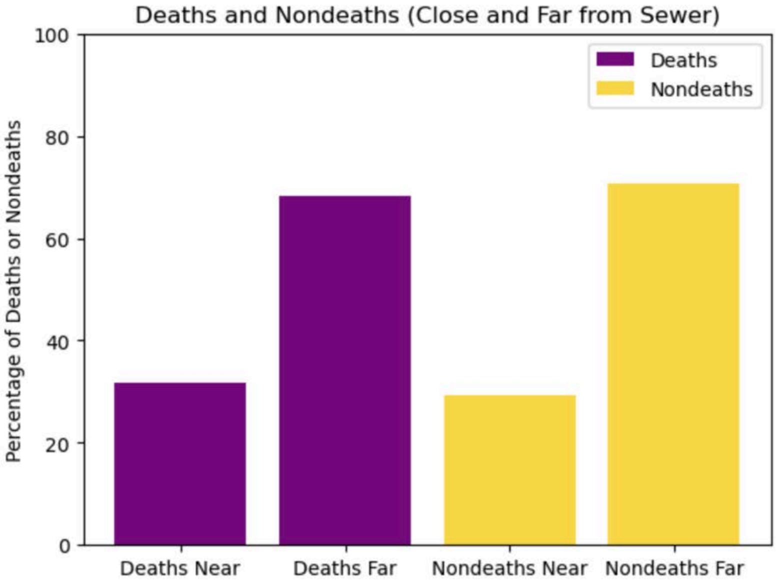
Counter: The south districts were all at a low elevation. The South London experiment shows significant differences in death rates for people living in this region. If elevation was a causal factor, we would expect all people at low elevations to have similar death rates. Elevation is a confounder.

Dr. William Rogers

Theory: Locally Airborne - Non-contagious: People get sick by inhaling locally polluted air (e.g. coming from sewage through gullies: avoid inhaling air from sewers).

Argument: Cholera occurs in areas with poor sanitation. Smells are evidence of the disease. Cholera is spread through the air and living close to sewers increases the chance of getting the disease.

Counter: A distance analysis of the Broad Street outbreak shows that a higher percentage of deaths actually occurred among people that lived farther from sewers than those who lived closer to sewers.

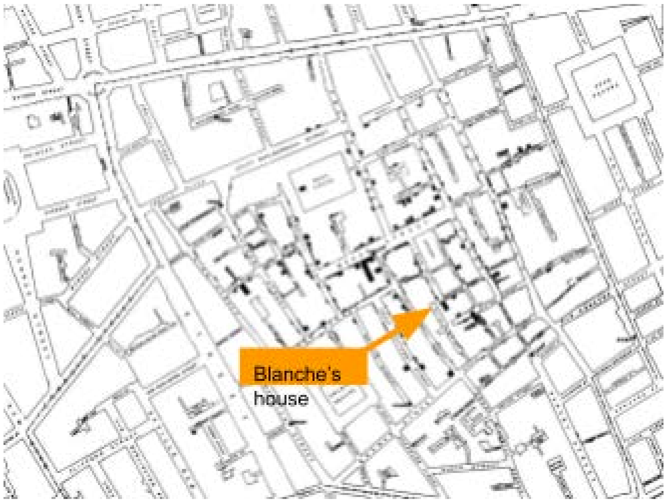


Blanche Smith, Wife of a Brass Finisher

Theory: (Airborne-Contagious): People infect each other through the air.

Argument: She lives in the Soho neighborhood, not far from the Broad Street Pump. She sends her children to get water from the pump but no one in the family got cholera during the Broad Street outbreak.

Counter: If asked to point to a map to show where she lives, Blanche will point to a house that is closer to the Broad Street Pump than other pumps on a straight line, but another pump is closer by walking distance. Her children are not getting the water from the Broad Street Pump. It is faster to walk to a different pump.

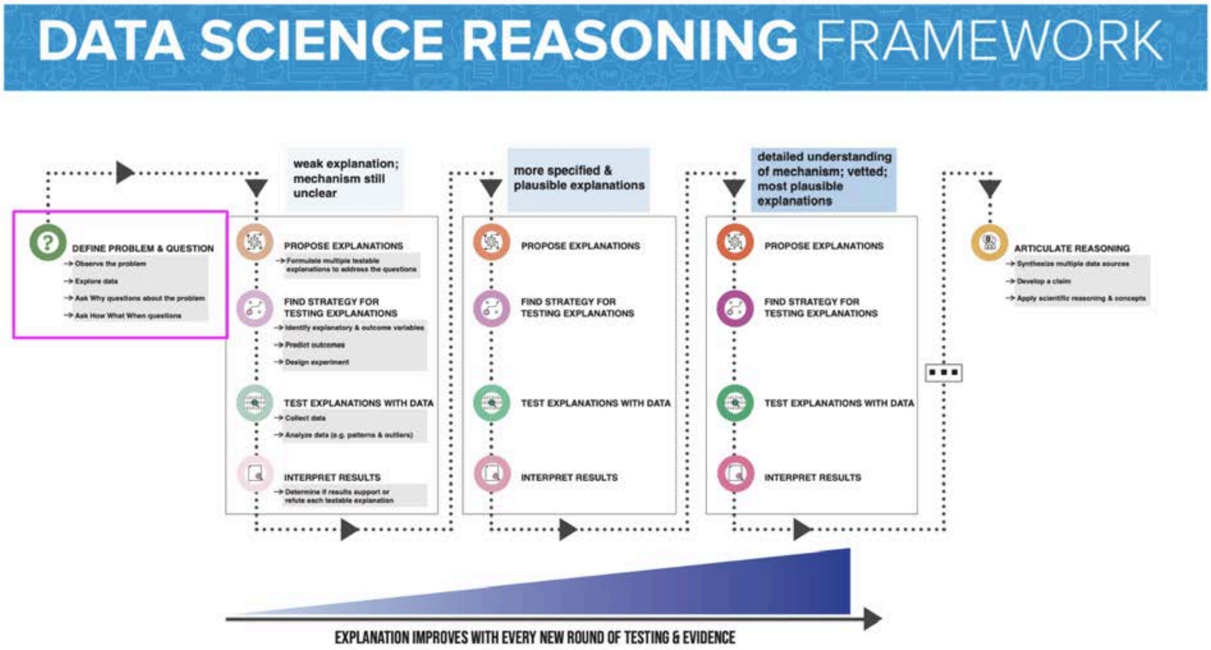


Case 2: Covid in Chicago

Since Case 2 follows the same logic as Case 1, the same learning goals apply in each lesson as specified in Case 1. We therefore do not repeat them for this second case.

Lesson 1 – Define the Problem & Ask the Right Questions

Reference to DSRF:



DATA SCIENCE REASONING FRAMEWORK
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
weak explanation; mechanism still unclear
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
more specified & plausible explanations
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
detailed understanding of mechanism; vetted; most plausible explanations
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

Activity Summary:

- □ **Class Discussion & Lecture (10 min): The Early Days of Covid –** DEFINE THE PROBLEM & QUESTION
- □ **Notebook (30 min): How has Covid Impacted Communities? –** EXPLORE DATA

Lesson 1 Class Discussion: The Early Days of Covid

Big Idea	Goals	Resources
The scientific process starts with observations and then progresses based on what type of questions are asked and answered.	□ Identify the characteristics of a (testable) scientific question. □ Ask (testable) scientific questions to solve a medical mystery	□ Slides
ACTIVITY SUMMARY: An open discussion eliciting students' recall of the impact of Covid early in the pandemic and uncertainties at the time		

- □ **Slide 2:** Ask the class, **"How did your daily life change at the very beginning of the Covid pandemic in March-April of 2020?"** Allow students to share out as much as they feel comfortable.
- □ **Slide 3:** Hand out a post-it to each student. Ask them, **"What was known and not known about Covid with respect to the disease itself (origins, how it spread, symptoms, who it affected more) in the early days of the pandemic?"**
- □ Have them write down one thing that was known or one thing that was not known about Covid in the early days of the pandemic.
- □ Have the students place these post-its in two groups on the wall and remain there for a discussion.
- □ Summarize for the class. If any post-its are incorrectly placed, explain why and move them (e.g., loss of taste/smell as a symptom was not known early in the pandemic).
- □ Point out post-its that describe things that were not known at the time, but are known now. At the least, try to find references to the following:
  - ○ Mode of transmission (droplets).
  - ○ Risk factors for serious illness.
- □ A student may have pointed out that the existence of Covid was put up for debate in the early days of the pandemic. If not, be sure to bring it up at the end. Use this to transition into the notebook.

Lesson 1 Notebook: Seeing the Problem in the Data

(30 min)

BIG IDEA	GOALS	RESOURCES
The scientific process starts with observations and then progresses based on what type of questions are asked and answered.	□ Observe trends in data pertaining to deaths in the city of Chicago	□ Notebook 7
ACTIVITY SUMMARY: This notebook is a review of several concepts from earlier including data structures, normalization and visualization as applied to the new problem, Covid.		

Part 1:

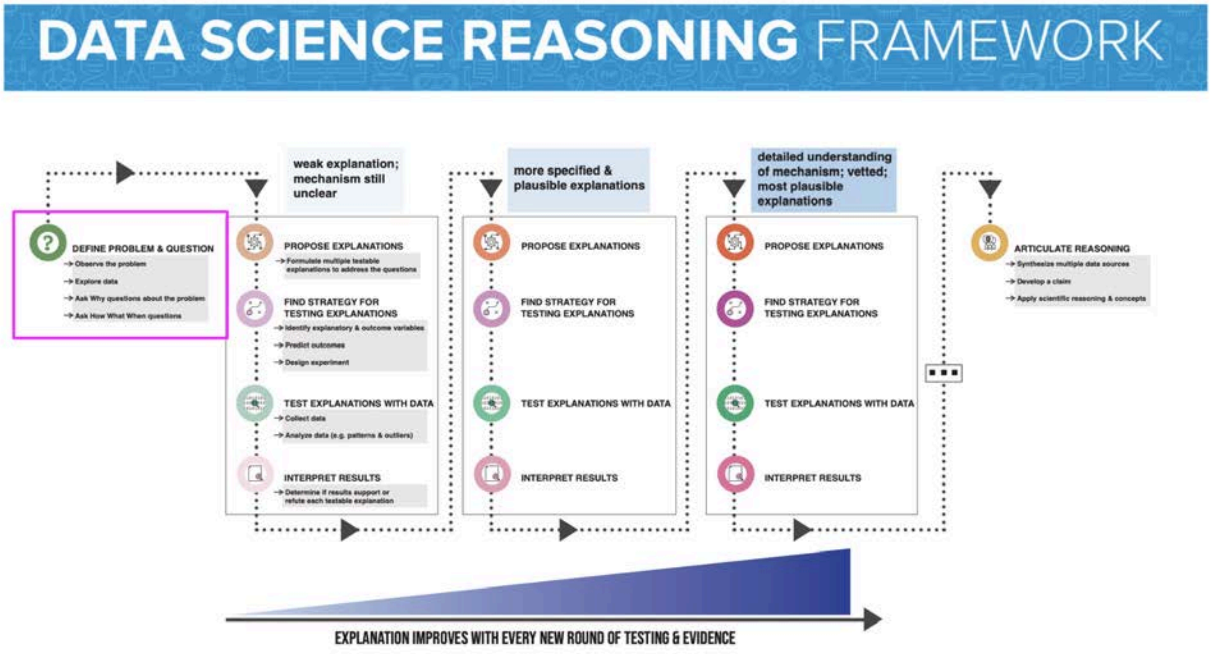
TASKS:

□ Normalize the number of deaths in Chicago by creating a new column: “Deaths per 1000.”

□ Create a line graph with title and labeled axes showing the change in death rates between 2012 and 2021.

Lesson 2 – Explore Data

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
weak explanation; mechanism still unclear
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
more specified & plausible explanations
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
detailed understanding of mechanism; vetted; most plausible explanations
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

Activity Summary:

□ **Lecture & Notebook (15 min):** DEFINE THE PROBLEM & QUESTION

□ **Small Group Activity (30 min): Spatial Patterns of Positivity –** EXPLORE DATA

□ **Small Group Discussion (10 min): How Can a Community be Described with Data? –** EXPLORE DATA

□ **Notebook (30 min): Uncovering Correlations with Covid-19 Positivity Rate** - EXPLORE DATA

Lesson 2 Lecture/Notebook: How has Covid Impacted Communities?

(40 min)

Big Idea:	Goals:	Resources:
Selecting a variable of interest requires considerations of normalization but also the accuracy and reliability of data that is used to generate the variable of interest.	Students will be able to: □ Identify and calculate a variable of interest that is normalized. □ Recognize potential sources of error arising from the data used to calculate the variable of interest.	□ **Slides** □ Notebook 7
Activity Description: Students consider how to use the supplied data (from Kevin Credit's paper) to best describe Covid's impact on communities in Chicago. Students consider the availability and accuracy of data, specifically with respect to testing variations. Case rate and testing rate are calculated from available data. Discussion: Do all communities have the same testing rates? Why the variation? How reliable are the case rates if testing is inconsistent?		

The focus of this activity is to first gain some familiarity with the Covid data with respect to what data is there and which is most appropriate to describing Covid's impact.

□ **Slide 5:** Pose the driving question: *How can we best measure Covid's impact on different communities in Chicago?* Collect responses from the class.

The students will likely suggest death rate after seeing it in the John Snow case study, but encourage them to think about other metrics that are not as obvious or worst-case for people. As students do this, the conversation may switch to the many perspectives of Covid's impact beyond disease, i.e., unemployment, mental health, crime, education. Encourage these perspectives, and point out that "impact" can mean many things, but for the purposes of our investigation, we will focus on the health impact (mortality, infection, hospitalization, etc.).

If the students struggle to come up with ideas, ask them what kind of numbers have been reported in the news. At the very minimum, the list should include:

- Number of deaths or mortality rate.
- Positive tests.
- Hospitalizations.

□ **Slide 6:** Explain Kevin Credit's research into Covid's impact on communities in Chicago during the early months of the pandemic and that the students will be using his data. Tell the students that they are now going to spend a few minutes looking at what data is available to them. Students go into Jupter/Colab and load the Covid dataset. Give them a couple minutes to run a cell that provides a listing of column names. Also included in the Jupter/Colab notebook is a key for the full names of each column abbreviation.

Part 2:

□ **Slide 7:** The task in this part of the notebook are to:

- Load the dataset and explore the available variables.
- Identify variables that could be used to measure Covid's impact.

□ After 5 minutes, bring the students back and ask them, **"What variables in the dataset could be used to measure Covid's impact?"** Students may notice that the dataset does not provide number of deaths, therefore mortality cannot be used as our outcome variable. Allow this conversation to identify the most viable options: `cases` and `tests.`

□ **Slide 8:** "**Which one is best to describe Covid's impact**?" Poll the students and ask them to share why they picked one over the other. Encourage them to list the advantages and disadvantages of each. In order to get ideas flowing ask them: **What has to happen for a person to get recorded in this data?** Lead them in this manner: for someone to be listed as a case, they have to be tested, for someone to be tested, they have to have a reason (symptoms, exposure, etc.).

Part 3:

□ **Slide 9:** Remind the students that we will need to normalize cases and tests in order to compare groups. Have the students go back into the notebook and complete the next section. The tasks are:

- Make a function called "case_rate" that normalizes case rate.
- Generate a new column of data for the week of 4/16 called "case_rate_4_16" using the "case_rate" function.
- Make a function called "test_rate" that normalizes test rate.
- Generate a new column of data for the week of 4/16 called "test_rate_4_16" using the "test_rate" function.

□ **Slide 10:** Have a conversation about the factors that influence whether someone will seek out a test or not. Explain that for someone to be tested, they need access and the desire to be tested. Based on this, ask "**Can we trust these numbers?"** Give the example on the slide to illustrate the dangers of case rate and testing rate as inaccurate outcome variables.

□ Ask students for ideas on how we can use the data we have (`case rate` and `testing rate)` to, as accurately as we can, measure Covid infections in a community. Remind them that a key strategy when working with data is to minimize the effect of variations across the data caused by factors that correlate, but are not causal to the problem. In other words, we know that testing access and the willingness to get tested is not consistent across the city. How can we compensate for variations in testing? This is a challenging question and students may or may not have ideas. Either way, ask them if they have heard of "positivity rate" and what it is. If no one can explain, tell them that it is simply the percentage of tests that come back positive.

□ **Slide 11:** Ask **"How does positivity rate affect differences in testing?"** Guide the group to the understanding that even if a community has a low testing rate (Slide 12), positivity rate will still be high if there are a lot of infections. If an area has a high testing rate (Slide 13), more people will test positive, but it will be balanced out by the larger number of tests.

□ Have the students return to their Jupyter/Colab notebooks. Here they will see some cells that allow them to define a function for calculating positivity rate. All they need to do is fill in the correct variables into the formula. They may find it helpful to look back at Exercise 2 for how we did this with mortality rate. Allow the students to talk through this together.

□ Once they have the function defined, the next cell will allow them to apply it to the dataset and calculate a new column, "`pos_rate`". Congratulate the students on making the "variable of interest" for how we measure Covid's impact across the city.

Notebook tasks:

□ Define a function that calculates positivity rate.

□ Create a new column in the DataFrame for the week of 4/16 called "pos_rate_4_16".

Lesson 2 Small Group Activity: Spatial Patterns of Positivity

(45 min)

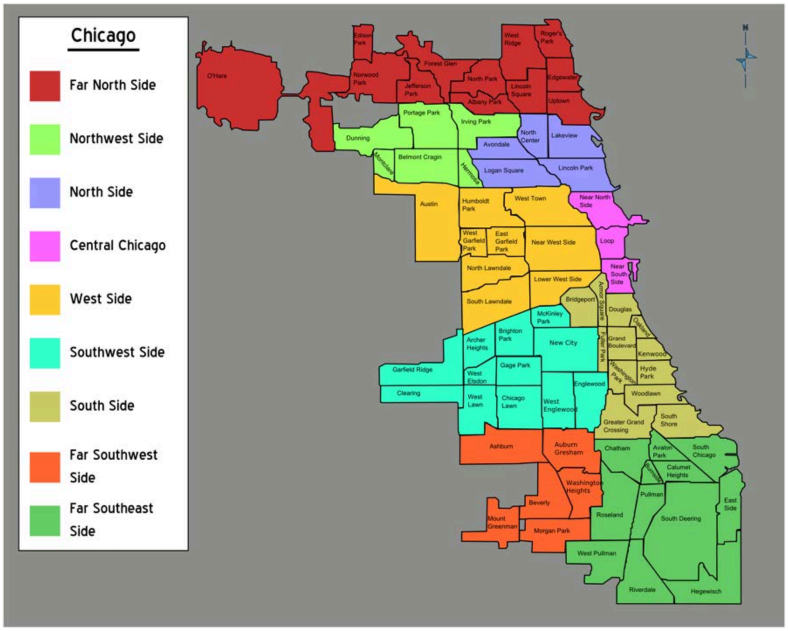
BIG IDEA:	GOALS:	RESOURCES:
Data visualization can be used to identify spatial patterns and outliers in a phenomenon. This, in turn, generates questions that can be tested.	□ Use exploratory spatial data analysis (ESDA) to identify trends/patterns.	□ 20 in x 30 in maps of Chicago covid rates and "zip codes" (already printed) □ Link to spatial data

Activity Description:

1. Students use Kepler.gl to explore trends in Covid's impact on different regions (sides) of Chicago.
2. After confirming that some communities have been impacted more than others by Covid, the instructor guides them to their testable question: "Why has Covid affected some communities more than others?"
3. The focus of this activity is for students to formulate some very rough generalizations about the regional Covid impact patterns in Chicago using exploratory spatial data analysis.

□ Explain to the students that we now have our outcome variable, positivity rate, that describes the health impact of Covid. We can now use it to explore the question, **"How has Covid affected the different 'sides' of Chicago?"**

□ **Slide 15:** Show them a map of the Chicago "sides". They will be familiar with the north side and south side and possible west side (represented as the three stripes on the Chicago flag, but Chicago is generally divided in nine sides:



- □ Ask the students (whole group or small groups), **"Which sides of Chicago do you think have been impacted most by Covid? Why?"** This is an open brainstorm meant to elicit their baseline perspective on the problem.

- □ Explain that they will be using Kepler to do an **Exploratory Spatial Data Analysis** of this question. Guide the students through the following process:
  - o Go to kepler.gl on your browser.
  - o Download the Covid dataset from Google Drive.
  - o cov_chi_with_positivity.geojson.
  - o Expand the layer with the "v" button.
  - o Click the three vertical dots next to "Fill Color."
  - o Change "Color Based On" to pos_rate_4_16.
  - o Click on the color scale and check the "Reversed" toggle.
  - ○ Click on the show legend icon (the last icon in the top right) so students know what the colors mean numerically

- □ Once all students have the appropriate visualization pulled up in Kepler, Have the students apply post-its to each of the 9 sides on the 20 in x 30 in map corresponding to the scale colors in Kepler. For all sides, students will have to take a rough average of the ZCTAs (like zip codes) in that side. Conduct this as a group activity, assigning a side to each group.

□ Once the map is complete, refer them back to the question, **“What sides of Chicago have been impacted most by Covid?”**. Have an open discussion, gathering interpretations from each student. There are no right or wrong answers, but guide the students to see that communities on the west and south sides generally show more impact.

□ **Note:** While the visualization in Kepler would normally suffice to answer the driving question, the variation by ZCTA is probably a little too much. By organizing impact by side, the students should see regional patterns more easily.

□ Ask the students, **“After seeing these patterns, what kind of questions come to mind?”** Encourage questions as opposed to explanations (we are not even close to being there in the process – see where we are in the DSRF above).

□ Finally, remind them about John Snow’s work: he saw unusual patterns in where cholera impacted London, but made sure to ask questions that were **“testable”**. In other words, questions that could be answered through experiment.

Lesson 2 Small Group Discussion: How Can a Community be Described with Data?

(30 min)

Big Idea	Goals	Resources
There are many ways that people can discuss the impact of Covid in Chicago. However, using data to support this discussion allows us to understand the effects of Covid in more objective ways.	□ Students can identify ways to describe communities and Covid's impact on communities using data.	□ **Slides**
ACTIVITY SUMMARY: Groups brainstorm ways that communities can be described with data. Specifically, what metrics and groups of metrics can be used.		

Lead Instructor:

□ **Slide 18:** Show the question: "How can a community and its people be described with data?"

□ Remind the students that data must be measurable, often with a number. Even things that are normally not thought of as numerical, like colors, can be coded as numbers.

□ Explain that they are going to brainstorm as many ways to measure a community as possible in 5 minutes.

Mentors:

□ In their groups, students should use post-its to brainstorm as many ways that a community can be described with data. Do not provide hints, allow the students to think of the many aspects of the community: people, physical space, resources, etc. Encourage creativity.

Lead Instructor:

□ After 5 minutes, ask the groups to sort their post-its into groups. Do not provide any other instructions.

Mentors:

□ Allow the students to sort their metrics into groups. There are no right answers here. How the students organize is up to them.

□ After the sorting is complete, have the students name each group and put that name on a post-it. Examples include: education, income, culture, family, business, recreation, etc.

Lead Instructor:

□ Explain that a very large amount of data is constantly being collected about communities and their people. We are going to use some of that data to try to explain why Covid has impacted some communities more than others.

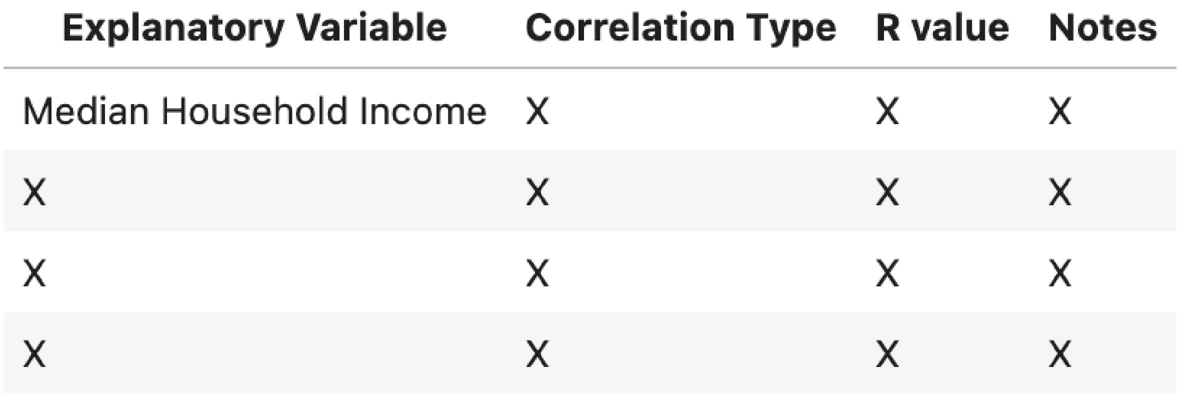
Lesson 2 Notebook: Uncovering Correlations with Covid-19 Positivity Rate

(45 min)

Big Idea	Goals	Resources
By exploring relationships between variables, we can begin to formulate explanations around mechanisms leading to an outcome	□ Determine which explanatory variables correlate with COVID positivity □ Construct scatter plots to visualize our results	□ Notebook 8
ACTIVITY SUMMARY:		

Task:

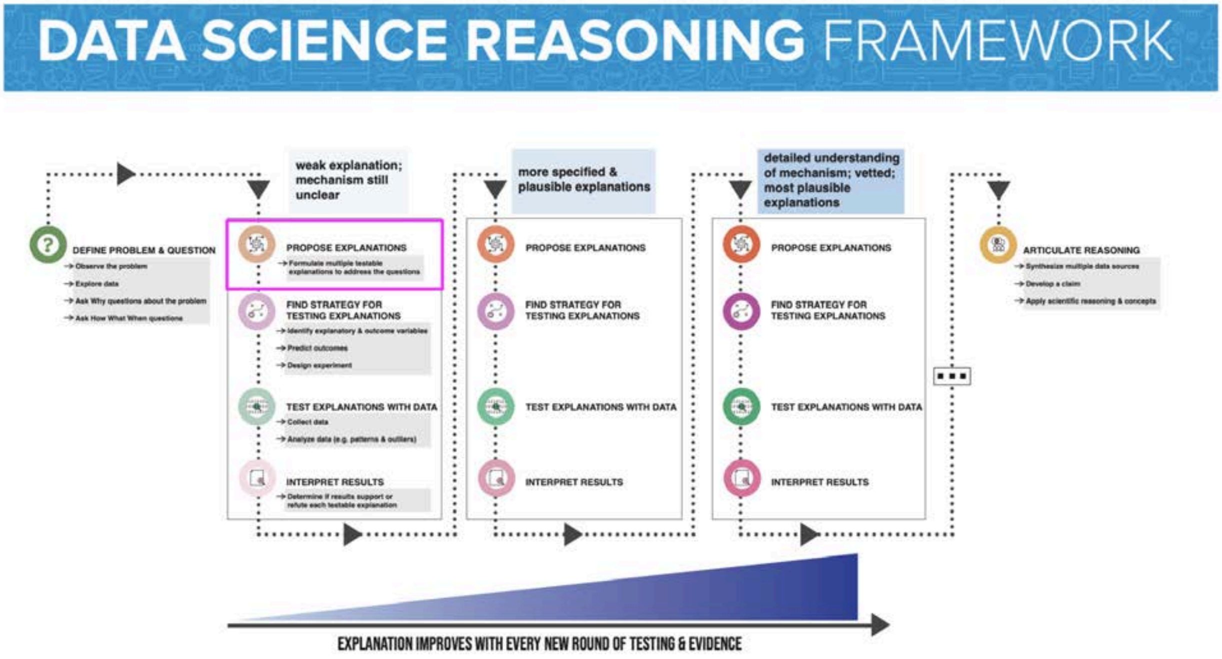
Explore a total of 4 explanatory variables. Record what you learn from each in the following table:



Explanatory Variable	Correlation Type	R value	Notes
Median Household Income	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X

Lesson 3 – Propose Explanations

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

Activity Summary:

□ **Small Group Discussion (30 min): Distinguishing Between Explanatory Variables and Their Proxies –** PROPOSE EXPLANATIONS

Lesson 3 Small Group Discussion: Distinguishing Between Explanatory Variables and Their Proxies

(30 min)

Big Idea	Goals	Resources
Establishing mechanistic proposed explanations requires the selection of explanatory variables that describe mechanistic relationships.	□ Distinguish between explanatory and proxy variables. □ Construct a proposed explanation and identify a corresponding explanatory variable.	□ Cards with variables and descriptions □ Lesson 3 slides
ACTIVITY SUMMARY: After a brief lecture exploring proxy and explanatory variables, small groups review the variables available to them in the Covid dataset. They group the variables into 2 categories: “explanatory variables” and “proxies”. They then develop their proposed explanation for the driving question: “Why has Covid affected some communities more than others?” and select an explanatory variable to test the explanation.		

Introduction (Lead Instructor):

□ **Slide 2:** Show the scatterplot of positivity rate vs. median income and point out that there is a very strong correlation between the two. The p-value is very small.

□ **Slide 3:** Jokingly ask, “Does this mean money protects you from Covid?” Students will respond, “No” but a few may point out that, in a way, having higher income does provide people with certain things that are directly protective of Covid.

□ Ask students to dive deeper into this idea and propose examples of where a higher income carries with it, certain benefits that may protect one from Covid. Field responses.

□ **Slide 4:** After the student brainstorm, show a few examples. Explain that this is far from a complete list. In fact, income, as we all know, affects countless outcomes in our society.

□ **Slide 5:** Explain that median income is what is known as a “proxy” variable. Proxy variables are surface-level variables that often do not have a direct role in the mechanism but can be highly correlated with the outcome. This is because the proxy stands in for, or is connected to, many variables that are directly linked to the mechanism.

□ Proxy variables can be dangerous in that they can be used to make broad generalizations and stereotypes about people that do not get at the root cause of a problem.

□ However, proxy variables can also be used when data for the variables are "under their umbrella". However, data scientists need to be very careful when using proxies and be clear about why and how they are using them.

Small Group (Mentors)

□ Using the cards with some of the variables from the dataset (ZCTA, ZIP, POP and the Covid variables have been removed). As a group, begin sorting them into 2 groups: **proxies** and **explanatory mechanisms**.

□ Remind the students that explanatory variables must be mechanistic. In other words, they would describe a characteristic of people or communities that could increase or decrease either exposure to Covid (eg. transportation to work). Encourage students to think a "level down" for a variable. Ie. Does this variable represent other, more specific things? For the proxy variables, have them think about what mechanisms they might proxy for (see slide 4 for income and slide 8: different scales of explanatory mechanisms)

□ The proxies include median income and race/ethnicity. These variables are broad characteristics of people but require extra steps to link them to a mechanism of infection.

□ Once this is complete, develop a proposed explanation for the driving question, **"Why has Covid affected some communities more than others?".** There may be a variety of ideas. To steer the conversations, refer the students back to Notebook 8 in order to identify strongly correlated explanatory variables.

□ Choose one explanatory variable that best aligns with the proposed explanation. Two challenges may arise:

  ○ The proposed explanation is broad and several, linked variables can be selected. Solution: this is a good sign that the students are thinking systemically. At this point select the best variable for an initial test. Explain that as you dive deeper into the problem, the other variables can be considered.

  ○ The proposed explanation does not fit with any explanatory variables:

  Solution 1: Help the students refine the explanation so that it aligns with one variable (and preferably others, see above). Remind them that we can only test explanations that we have data for.

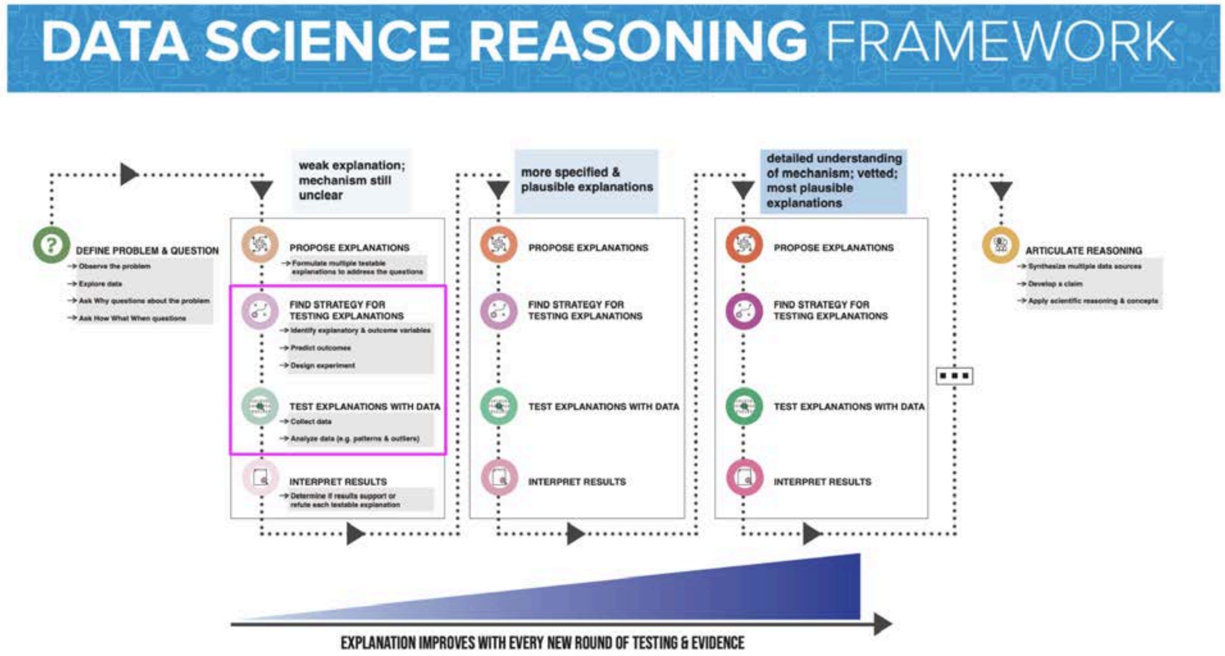
  Solution 2: Use a proxy variable. Proceed with caution. Race/ethnicity and income might be too broad, but age and no insurance can be justified.

□ Students will naturally gravitate to variables with expected positive correlations with positivity rate (eg. percent healthcare service workers), but negative correlations (eg. percent telecommuting) are equally important.

□ On a large piece of paper, write the proposed explanation at the top. Leave plenty of space for editing this statement and adding the results of the data experiments. This will be used in the final activity.

Lesson 4 – Find Strategy for Testing Explanations I

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

Activity Summary:

- □ **Notebook (45 min): Two-way Table Analysis of Explanatory Variable and Positivity –** FIND STRATEGY FOR TESTING EXPLANATIONS & TEST EXPLANATIONS WITH DATA

Lesson 4 Notebook: Contingency Table Analysis of Explanatory Variable and Positivity

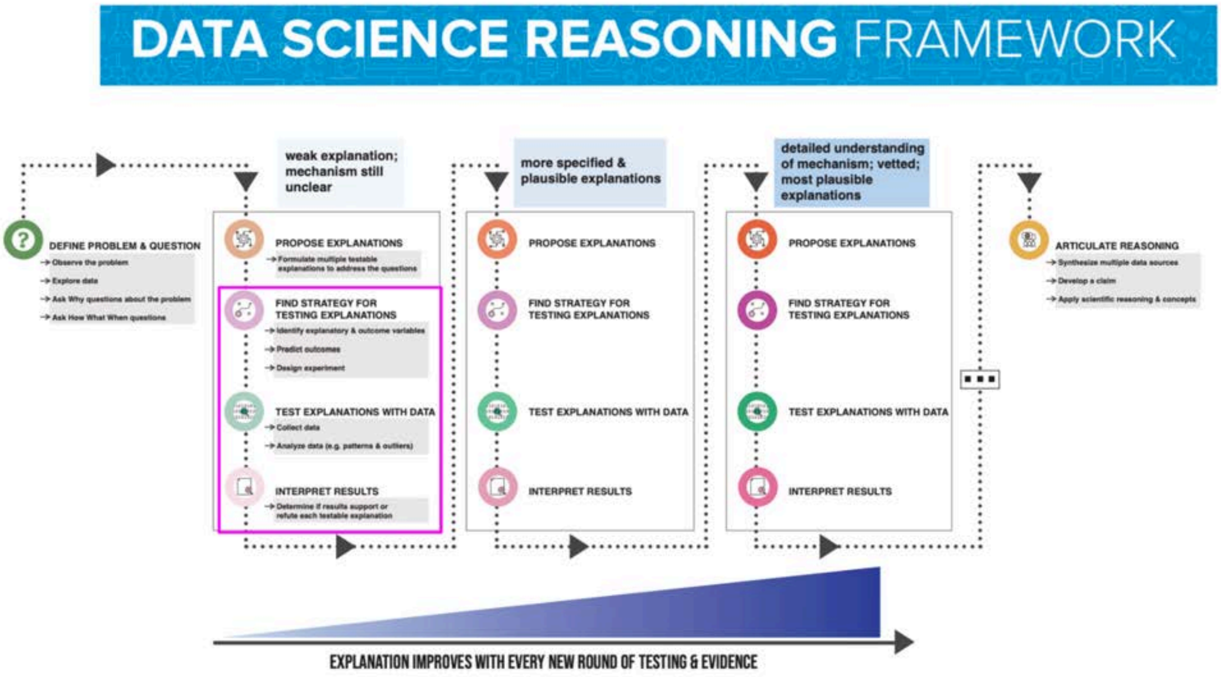
(45 min)

BIG IDEA	GOALS	RESOURCES
Testing proposed explanations requires the statistical analysis of inferred relationships.	□ Assess the relationship between the explanatory variable / the proposed explanation and positivity rate.	□ Notebook 9
ACTIVITY SUMMARY: Students construct and test a two-way table for their explanatory variable.		

Task:Create a two-way table and perform a Chi-square test to determine if the relationship between the chosen explanatory variables and outcome variable are statistically significant.

Lesson 5 – Analyzing Results I

Reference to DSRF:



DATA SCIENCE REASONING FRAMEWORK

DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions

weak explanation; mechanism still unclear

PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions

FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment

TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)

INTERPRET RESULTS
→ Determine if results support or refute each testable explanation

more specified & plausible explanations

PROPOSE EXPLANATIONS

FIND STRATEGY FOR TESTING EXPLANATIONS

TEST EXPLANATIONS WITH DATA

INTERPRET RESULTS

detailed understanding of mechanism; vetted; most plausible explanations

PROPOSE EXPLANATIONS

FIND STRATEGY FOR TESTING EXPLANATIONS

TEST EXPLANATIONS WITH DATA

INTERPRET RESULTS

ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts

EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

Activity Summary:

□ **Lecture (15 min): Multiple Causes –** INTERPRET RESULTS I

□ **Small Group Discussion (15 min): Revise Theory to Account for Multiple Causes** - PROPOSE EXPLANATIONS II

Lesson 5 Lecture: Multiple Causes

(15 min)

Big Idea	Goals	Resources
Establishing a robust explanation requires consideration of multiple causal factors.	□ Explore the concept of multiple causes for a single outcome.	□ Lesson 5 Slides
ACTIVITY SUMMARY: A brief lecture exploring the concept of systems thinking and multiple causes.		

□ **Slide 2:** Remind students of the idea of “cause and effect” and the dangers of assuming that “A causes B”. Explain that in many real-life problems, the mechanism is not that simple.

□ **Slide 3:** Rather, there are often multiple causes where several factors affect the outcome. Sometimes these factors are related to each other, sometimes they are not.

□ **Slide 4:** Show the simple example of a boring teacher causing a student to feel sleepy. Ask, **“Is this a reasonable cause-and-effect relationship?”** The students will answer in the affirmative based on much personal experience. Ask, **“Are there potentially other causes to you feeling sleepy?”**

□ Field responses. Students should have a variety of ideas.

□ **Slide 5:** Show the expanded cause and effect to include not getting enough sleep and just eating lunch.

□ **Slide 6:** Explain that while there are multiple causes for the effect/outcome, not all of them have an equal impact. Untangling the amount of impact that a cause has is one of the most challenging things in data science. We won’t go into it here.

Lesson 5 Small Group Discussion: Revise Theory to Account for Multiple Causes

(15 min)

BIG IDEA	GOALS	RESOURCES
The results of data experiments allow us to assess the plausibility of proposed explanations. The resulting assessment guides expansion, revision or refutation of the proposed explanation.	□ Revise or completely change the proposed explanation based on results of the two-way table test. □ Identify additional explanatory variables to further test the proposed explanation.	□ Notebook 9
ACTIVITY SUMMARY: Groups revise their initial explanation to incorporate an additional one or more explanatory variables.		

- □ Review the results of the two-way table from the previous notebook. Specifically, consider the p-value. Record the two-way table and p-value on the large piece of paper.
- □ If the p-value was relatively small ($p < 0.05$), it suggests the proposed explanation could have some validity. Proceed to Step 6. However, for larger p-values, the proposed explanation will need to be refined or even changed completely. Encourage the students to attempt to revise the theory first.
- □ Revising the theory will require students to consider another explanatory variable that may be more important than the one originally chosen. This is an example of the “weight” of the causes. The original variable still may have influence on the outcome, but it may be masking another that has more impact on the outcome.
- □ If the students find that they cannot revise the theory, it may be necessary to discard the original theory and start a new one. In this case, record the reasoning on the large piece of paper and guide the students through the development of a new theory using the relevant steps of Lesson 3.
- □ Starting a new theory may prevent the group from investigating multiple causes. However, explain that there is still value in their decision to discard the original theory and the reasoning and it will be important to share this seeming “failure” later on.
- □ Guide the small group in identifying additional variables that may be connected to the explanatory variable identified in Lesson 3.

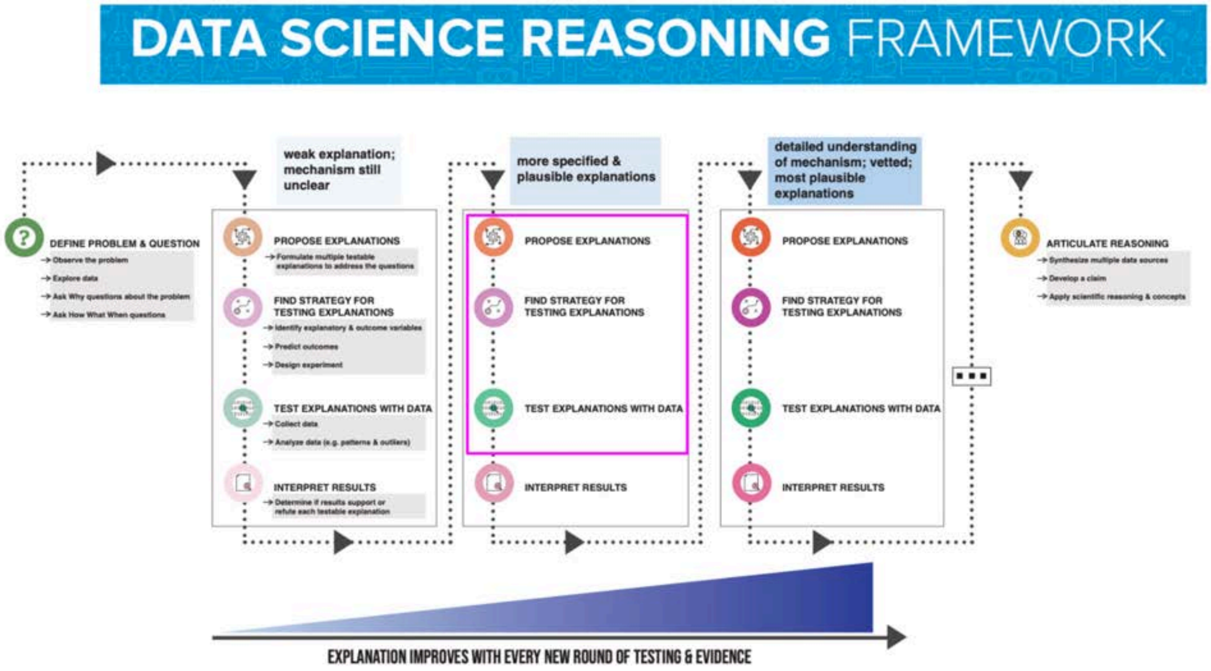
□ Remind the students that they need to stay within the boundaries of their proposed explanation.

Other explanatory variables they bring in should be, at least, tangentially connected. For example, if they believe that higher use of public transportation is increasing exposure and positivity rates, they should select other variables like occupation or food deserts that are connected to the likelihood of using public transportation. Bringing in variables that are not directly linked to either the mechanism or the original explanatory variable suggests a completely different proposed explanation and should be avoided.

□ If necessary, revise the original proposed explanation from Lesson 3 to incorporate these additional variables. Record changes on the large piece of paper.

Lesson 6 – Testing Explanations II

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

Activity Summary:

- □ **Whole Class Discussion (15 min): Present Initial Findings –** PROPOSE EXPLANATIONS, FIND STRATEGY FOR TESTING EXPLANATIONS II
- □ **Lecture: Comparing Explanatory Variables –** TEST EXPLANATIONS WITH DATA II
- □ **Lesson 6 Notebook: Testing & Validating Explanations –** TEST EXPLANATIONS WITH DATA II

Lesson 6 Whole Class Discussion: Present Initial Findings

(15 min)

BIG IDEA	GOALS	RESOURCES
Collaborative efforts in science require regular updates on progress from the individual contributors. This facilitates the collective movement toward consensus.	□ Summarize and assess results of the first data experiment.	□ Notebook 9
ACTIVITY SUMMARY: Each group presents the results of their two-way table analysis of their chosen explanatory variable(s).		

□ **Mentors:** prepare your groups to share out their results for the first data experiment by taking 2-3 minutes discussing the following in your small group. The answers to these questions are what will be presented:

- ○ What was the reasoning for selecting your particular explanatory variable? i.e. What insights from the initial data exploration and personal knowledge drove your initial hypothesis?
- ○ What were the results of the two-way table? Was there a statistically significant relationship between the explanatory variable and positivity rate?
- ○ What is your group's plan moving forward? Back to the drawing board? Bring in more variables?

□ Select a student from your group to present the findings.

□ Each presenter will take 1-2 minutes to share the group’s findings.

□ Other students will also have a chance to ask any questions. Encourage students to think about and discuss how the answers to their questions may be addressed.

Lesson 6 Lecture: Comparing Explanatory Variables

(20 min)

Big Idea	Goals	Resources
The simple bivariate setup between an explanatory and an outcome variable is extended to look at the relationship between multiple explanatory variables.	Students start to think about explanatory mechanisms in more complex ways by considering the relationship between multiple explanatory variables.	□ Lesson 6 Slides
ACTIVITY SUMMARY: A brief lecture describing methods for comparing explanatory variables and how to gain insights from th4ese analyses.		

- □ **Slide 2:** Show the scatterplot of comparing positivity rate and percent food service workers and explain that this is usually how we explore potential relationships: how the explanatory variable (percent food service workers) correlates with the outcome variable (positivity rate). While this may provide insight to a potential relationship, it does not begin to prove that the explanatory variable causes the outcome variable.
- □ **Slide 3:** Using the slide from Lesson 5, remind the students that multiple causes could be at work and we should consider multiple explanatory variables.
- □ **Slide 4:** However, just as important as investigating how each explanatory variable is related to the outcome, we should consider how the explanatory variables are related *to each other*.
- □ Rhetorically ask, “Why is this important?” Because comparing explanatory variables to each other can help us understand more about the complex relationships that exist and we can use reasoning to untangle them and identify which relationships are potentially most important. From this, we can build a better picture of what is actually happening.
- □ **Slide 5:** Explain that we will use the positive correlation between percentage of food service workers and positivity. On the surface, a potential mechanism could be suggested that these workers are being exposed to the virus at their workplaces where they must be around many other people. However, other questions could be asked such as:
- □ **Slide 6:** Is food service a spurious correlation? In other words, is there something else about food service workers that is increasing infection that has nothing to do with their jobs? Explain the hypothesis:

- o Do these workers live in areas that don’t have good access to hospitals?

Ask, **“What variables would we want to compare to test this hypothesis?”** Answer: PERFOOD and WS_5 (hospital accessibility score)

□ **Slide 7:** Show the scatterplot that shows no correlation. Explain that the data does not support the hypothesis.

□ **Slide 8:** Is it something related to working in food service that isn’t as simple as being exposed to more people?

- o Do these workers tend to have less access to health insurance because their jobs are lower paying and/or do not offer good benefits?

□ **Slide 9:** Show the scatterplot that shows a strong correlation. Explain that the data supports the hypothesis, but we should avoid suggesting that it is correct. We will want to test this hypothesis by bringing in other explanatory variables (no health insurance could be spurious and not actually part of the mechanism of getting infected) and using the other methods, beyond scatterplots, like bar graphs, two-way tables and spatial analysis, to name a few.

Lesson 6 Notebook: Testing & Validating Explanations

(30 min)

BIG IDEA	GOALS	RESOURCES
Scientists often explore the relationships between multiple variables to understand mechanisms and strengthen proposed explanations. In the end, this leads to the formation of strong claims supported by a body of evidence.	□ Perform more extensive analysis and construct multiple visuals to build evidence as to factors impacting Covid positivity rates	□ Notebook 10
ACTIVITY SUMMARY: Students use the notebook to conduct analysis of their expanded list of explanatory variables. The notebook provides previously used tools and methods such as scatterplots, bar graphs, and two-way tables.		

As a transition from the previous lecture, explain to the students that investigating and building a strong hypothesis requires considering the problem and data from many angles. Each part of the investigation provides a little more information. As the information builds, a more complete picture of the problem and potential solutions grow.

At this point, we are going to “cut them loose” to investigate the question, **“Why has Covid affected some communities more than others?”** This final notebook gives them access to both the data and the tools we have learned about over the course of the workshop including scatterplots, bar graphs and two-way tables. At this point, it is up to them on how they use the data and tools to build their hypothesis.

However, they should be working on the development of a concise, data-driven answer to this question and ideas on how to solve the problem.

Guidance for using the notebook:

□ Keep the students focused on the proposed hypothesis and systematically testing it with the tools.

It will be tempting for students to perform data experiments rather randomly. In order to mitigate this, ask probing questions such as:

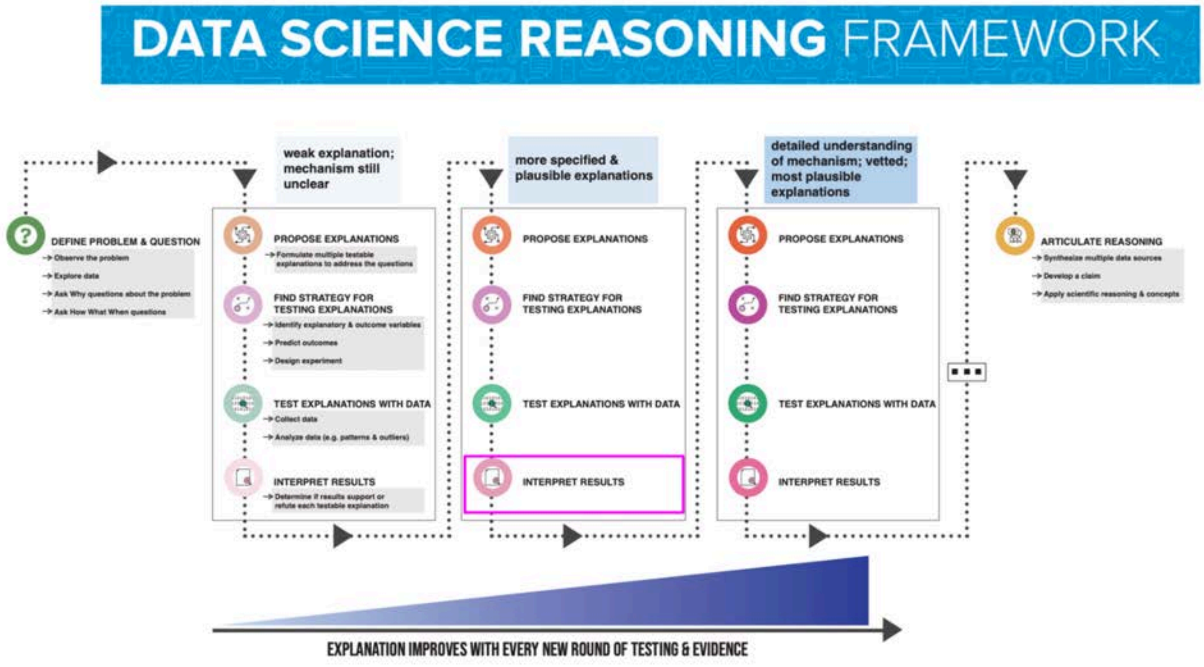
- ○ “What question are you trying to answer here?”
- ○ “What variable(s) are best to investigate this?”
- ○ “Which method (eg. two-way table) is best to perform that data experiment?”

□ Encourage the students to take notes of their observations and insights in markdown cells.

□ Have the students copy and paste methods they use repeatedly so that they have a record of their data experiments and results.

Lesson 7 – Analyzing Results II

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

Activity Summary:

□ **Small Group Discussion (15 min):** INTERPRET RESULTS II

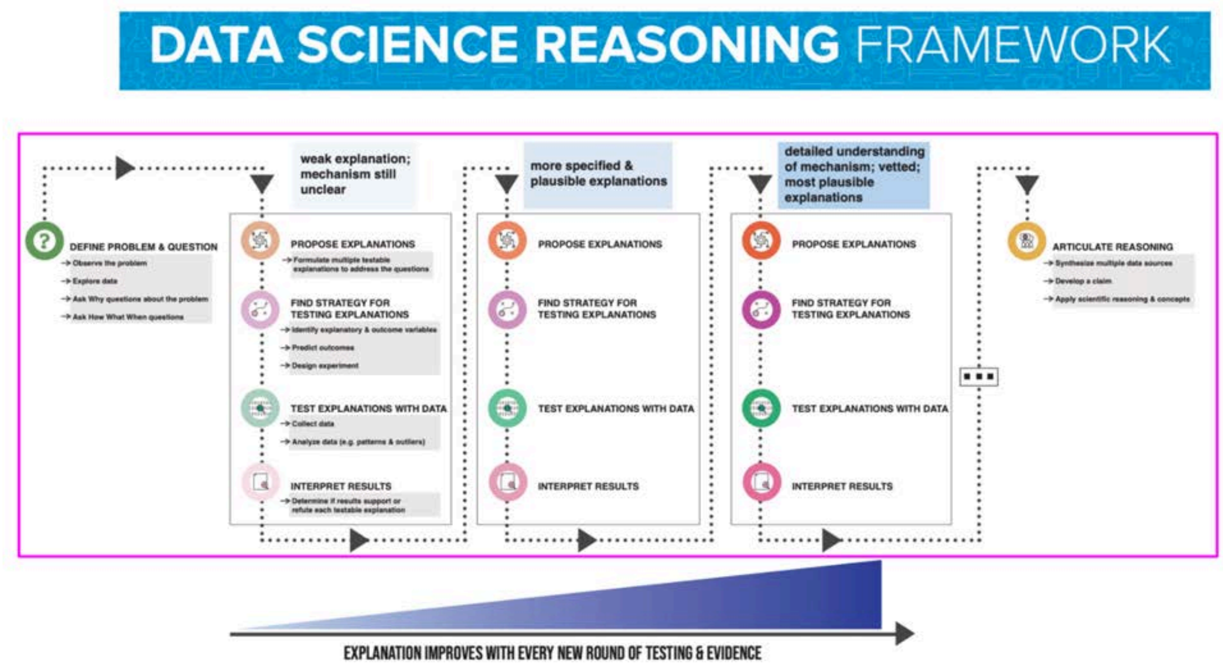
Lesson 7 Small Group Discussion

(15 min)

BIG IDEA	GOALS	RESOURCES
With a body of evidence researchers can begin to understand phenomena and prepare to form a claim.	□ Identify the strongest and most plausible explanatory variables contributing to Covid positivity	□ Notebook 10
ACTIVITY SUMMARY: Students consider the results of their second analysis and determine which explanatory variables and their interactions to include in their final explanation.		

Lesson 8 – Articulate Reasoning

Reference to DSRF:


DATA SCIENCE REASONING FRAMEWORK
weak explanation; mechanism still unclear
more specified & plausible explanations
detailed understanding of mechanism; vetted; most plausible explanations
DEFINE PROBLEM & QUESTION
→ Observe the problem
→ Explore data
→ Ask Why questions about the problem
→ Ask How What When questions
PROPOSE EXPLANATIONS
→ Formulate multiple testable explanations to address the questions
FIND STRATEGY FOR TESTING EXPLANATIONS
→ Identify explanatory & outcome variables
→ Predict outcomes
→ Design experiment
TEST EXPLANATIONS WITH DATA
→ Collect data
→ Analyze data (e.g. patterns & outliers)
INTERPRET RESULTS
→ Determine if results support or refute each testable explanation
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
PROPOSE EXPLANATIONS
FIND STRATEGY FOR TESTING EXPLANATIONS
TEST EXPLANATIONS WITH DATA
INTERPRET RESULTS
ARTICULATE REASONING
→ Synthesize multiple data sources
→ Develop a claim
→ Apply scientific reasoning & concepts
EXPLANATION IMPROVES WITH EVERY NEW ROUND OF TESTING & EVIDENCE

Activity Summary:

- □ **Spark Activity (20 min): Evidence-Based Reasoning and Confirmation Bias –** ARTICULATE REASONING
- □ **Small Group Discussion (20 min): Construct Argument –** ARTICULATE REASONING
- □ **Small Group Activity (45 min):** ARTICULATE REASONING

Lesson 8 Spark: Evidence-Based Reasoning and Confirmation Bias

(20 min)

BIG IDEA	GOALS	RESOURCES
Confirmation bias often leads to flawed evidence-based arguments. Students are often unable to recognize cognitive bias in their own disciplines, and simply describing cognitive bias to students has shown to be insufficient to improve critical thinking. This activity will demonstrate the effect of confirmation bias in the hopes that students can minimize its effects in the future.	Students will be able to: □ recognize cognitive biases that influence their ability to be objective and then understand and use reflective techniques to reduce confirmation bias when evaluating evidence.	□ Lesson 8 Slides to introduce the activity. □ Summary of research on "Is Homework Worth it?" □ Puzzle activity handout

ACTIVITY SUMMARY: Students are asked to write a short evidence-based claim, then they do a confirmation-bias intervention activity, and finally reevaluate their evidence-based claim to see how they can improve upon it by recognizing and addressing their own confirmation bias.

PREPARATION:

Make copies of "Is Homework Worth It" and "Puzzle activity" for students.

ROLES:

Instructor will lead discussion, TAs will hand out sheets, All hands for small group discussions.

□ As an intro activity, ask the students to solve a simple math problem in their heads:

- ○ Take a 1000.
- ○ Add 40.
- ○ Add 1000.
- ○ Add 30.
- ○ Add 1000.

- Add 20.
- Add 1000.
- Add 10.

□ Have the students yell out the answer. Surprisingly, most will answer, “5000”.

□ On the board show the problem to verify that the answer is 4100.

□ Ask the students, “What did this tell us about how your brain works? What does it tend to do?” Responses may include:

- It makes mistakes.
- It struggles when the problem isn’t written down.

But look for the following:

- It likes patterns.
- It looks for shortcuts.

□ Explain that the brain’s tendency to find patterns and take advantage of shortcuts has **advantages**: brain can work very fast and not worry about every little detail, its is efficient and saves energy and **disadvantages**: the brain makes mistakes and is not accurate like a calculator, once the brain sees a pattern or what it thinks is a pattern, it locks on and resists evidence that goes against that pattern.

□ Inform students that this is an activity to help them engage in evidence-based arguments.

□ **Use the slide show** to introduce students to the claim, evidence, reasoning method of forming a scientific argument.

□ Explain that the two examples represent different types of evidence, experiment based and research based. The two examples also represent different disciplines where the method is used, science and social science. It may be worth mentioning that the second example is a socio-scientific problem and **not** a data science problem, so there may not be a “right” answer, but the process is similar and therefore it is worth exploring this example.

□ Next, tell students we will be practicing this method to argue whether or not homework is worth it.

□ Give students a couple of minutes to state their claim, write down their evidence and give their reasoning. They do not have to share these with the group, it is just for their use and we will come back to it at the end of the exercise. Have students use the summary of the research handout, telling them that it will save us some time.

□ A few students may share their examples if they would like and a quick discussion can happen. Ideally a student from each side will share their example. You can tell them that we will come back to this later.

□ Next, tell students that we will look at another way to evaluate our evidence by trying to solve a puzzle.

□ Students pair up with a neighbor and each partner gets a different set of instructions (Instruction Handout is included. Confirmation Bias – Puzzle Activity Handout). Student #1 is the "guesser" whose job it is to guess the "pattern rule" that describes the relationship of a sequence of numbers. This student is given an example sequence that fits the pattern rule: 2, 4, 8, 16. Student #1 can ask their partner if sequences fit the pattern rule or not – as many sequences as they wish – before they have one chance to guess what the pattern rule is. Students should once again fill out the claim, evidence, reasoning sheet. The job of Student #2 (the "teller") is to tell whether the sequence fits the pattern rule or not. Their role consists of saying "yes" or "no." The pattern rule is that the numbers are increasing.

□ **Here is an example exchange:**

- ○ **Guesser:** "Does 3, 6, 12, 24" fit the rule?
- ○ **Teller:** "Yes."
- ○ **Guesser:** "Does 1, 2, 4, 8, 16" fit the rule?
- ○ **Teller:** "Yes."
- ○ **Guesser:** "Is the rule that the numbers are doubling?"
- ○ **Teller:** "No, that is not the rule."

Remember that Guessers can test out as many sequences as they wish, but they only get one guess as to what the pattern rule is.

□ **Basis of discussion for unpacking the activity:** Most students will (incorrectly) guess that the pattern rule is "doubling" after trying just one or two sequences. They might guess that the pattern rule is "increasing" (correct), but it's quite unusual for anyone to ever try to get a negative answer – that is, to try a sequence they think will NOT fit the pattern rule, despite the fact that there's no penalty for testing a sequence that does not fit the pattern rule. No one thinks to intentionally test something they think is wrong! This leads to a broader discussion about the importance of exploring alternative views, including those that you feel are incorrect, and the relevance of testing assumptions as a part of evidenced-based reasoning.

□ **Guide students through a discussion:** How many people guessed the rule wrong? How many sequences did you try before you guessed the rule? Did you try to get a sequence that didn't fit the rule? Why or why not?), then on what it means (Why do people not try things they think are wrong?) and then finally on how we can generalize this behavior (If we know we behave this way, what does it mean for understanding evidence and data? How can we counteract this behavior? What can you do when researching your case study to alleviate this bias? What practices do scientists use to counteract this form of bias?).

***IF YOU HAVE TIME TO CONTINUE THE CONVERSATION:**

□ Now have students revisit their "Is Homework Worth it" argument. Ask them, **"How many of you only used evidence from "your" side of the argument?"** Using what they just learned, ask them how they could improve upon it. Students can think, pair, share. They should answer that they use evidence from the other side and address it. If time permits, go through one or two examples. or have them rewrite the claim, evidence reasoning to reflect what they learned.

□ The take-home message is that students need to always consider the counter-arguments (null hypothesis) and address them in their evidence and reasoning.

Lesson 8 Small Group Discussion: Construct Argument

(20 min)

Big Idea	Goals	Resources
Once researchers have done the work to validate a claim with extensive evidence, they communicate their findings to the broader community to facilitate discourse.	□ Communicate a claim backed by evidence	□ Digital presentation: Laptop, Microsoft Word/Google Slides □ Or poster presentation: Large paper, markers, tape, print-outs of visuals of students' choosing

ACTIVITY SUMMARY: After a brief explanation of the capstone experience, groups construct a digital argument in a notebook (or slides or paper posters) for their proposed explanation. They also prepare for counterarguments.

□ Explain to the students that we will be having a unique guest joining us: a policy maker that has a couple hundred million dollars in Covid relief funds but not a clue on how to spend it. Your task, as an entire class, is to put the different hypotheses on the table, consider them and arrive at a consensus in the form of an answer to the driving problem and way to use the Covid relief funds to help solve the problem.

□ Finalize hypothesis.

□ Gather relevant data and visualizations into notebook (can use end of Notebook 10), slides or poster.

□ Think about solutions to the problem (just consider at this point).

Lesson 8 Small Group Activity: Policy Presentation

(45 min)

Big Idea	Goals	Resources
Once researchers have done the work to validate a claim with extensive evidence, they communicate their findings to the broader community to facilitate discourse.	□ Community claim backed by evidence	□ Digital presentation: laptop □ Poster Presentation: poster
ACTIVITY SUMMARY: Groups present their arguments to a "policy maker" that is role played by an instructor, attempting to advise them on how best to use Covid relief funds.		

Background:

The Policy Maker is responsible for allocating federal Covid relief funds. However, while they have the decision-making power, they do not not have any understanding of the problem, the data or potential solutions.

Role Playing: Instructor:

The Policy Maker should "play dumb" with the students. Act as if you are walking into the room with a big bag of cash but no idea how to use it, other than it has to be used for some sort of Covid relief.

□ **Instructor 1:** Explain to the class that we will be joined by a guest and provide the background provided above.

□ **Instructor 2:** Walk into the room, and introduce yourself and why you are here. Emphasize that you do not know anything about Chicago, its communities or Covid, in general.

□ **Instructor 1:** Get the students started by having them present the problem (Covid is impacting some communities in Chicago more than others. Bring up the Kepler map of positivity rates and call on a student to explain what is being shown and how it conveys the problem.

□ **Instructor 2:** Possible questions:

- ○ What are positivity rates? Why not just map positive cases?
- ○ How are these communities different from each other?

Now that the problem has been introduced, each group will have a chance to present their findings. Each group can log into their Jupyter/Colab notebooks from the presenter’s computer and/or show their posters.

As the presentations proceed, encourage the students to challenge the arguments. The Policy Maker can seed these challenges by asking questions that point out potential inconsistencies in arguments:

□ Not including other related causes (simple A causing B)

□ It could be a spurious correlation (appears that A causes B but it could be C causing B)

□ Making an argument with just a scatterplot: Scatterplots are a good exploratory technique but are not enough to suggest causation. Two-way tables should be used to enhance the argument.