

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Administrative Information

- Instructor:
 - Ragesh Jaiswal
 - *Email:* ragesh@iitdabudhabi.ac.ae
 - *Office:* ADM.2.117

- Grading scheme:
 - Quizzes: 20%
 - Minor Exam: 20%
 - Major Exam: 30%
 - Lab assignments/projects: 28%
 - Class participation/attendance: 2%

- Grading scheme:
 - Quizzes: 20%
 - Minor Exam: 20%
 - Major Exam: 30%
 - Lab assignments/projects: 28%
 - Class participation/attendance: 2%
- Important points:
 - Homework will not be graded and so you are not supposed to submit the homework.
 - There will be a quiz based on the material of the homework sheet (*as well as lab*).
 - Lab projects/assignments will be evaluated through a viva. We will ask you to write and execute code in front of us.
- Unfair means:
 - Copying from one another, copying from the internet or LLMs.

Administrative Information

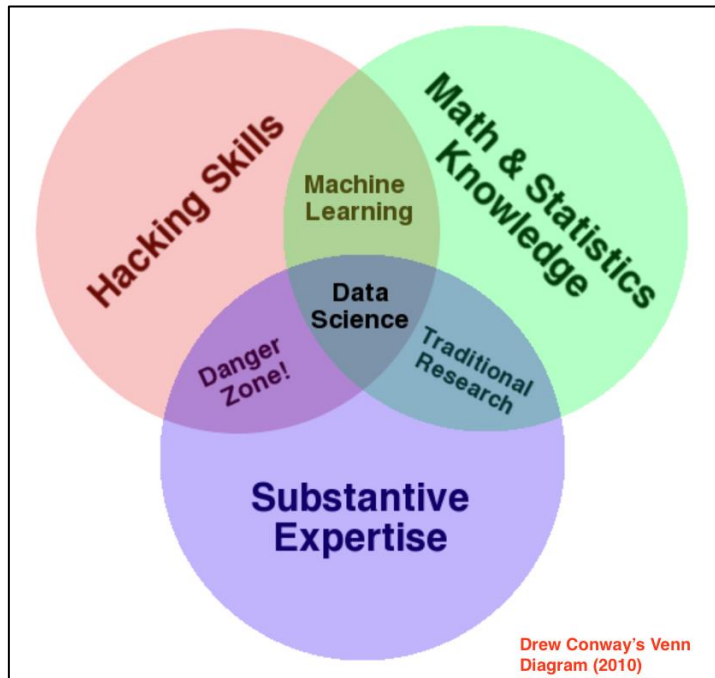
- Books: *(We will use a collection of sources most of which are available on the internet.)*
 - [LDS] Learning Data Science, *by Sam Lau, Joey Gonzalez, and Deb Nolan.*
 - [CIT] Computational and Inferential Thinking: The Foundations of Data Science *by Ani Adhikari, John DeNero, David Wagner.*
 - [ML] Machine Learning *by Tom Mitchell.*
 - [ISL] An Introduction to Statistical Learning *by Gareth James, Daniela Witten, Trevor Hastie, and Rob Tibshirani.*
- Gradescope: A paperless grading system. You will get added to the class roster and will receive an email regarding this on your official IITD-AD address. Please follow the registration instructions.
- Course webpage: <http://www.cse.iitd.ac.in/~rjaiswal/Teaching/2026/AAIL100>.
 - The site will contain course information, references, homework/tutorial problems. Please check this page regularly.

Introduction

- *What is data science?*

- The definition is yet to stabilize. We will try to word it based on popular opinions:

“Data Science is an interdisciplinary area that uses a combination of tools and techniques from Computer Science and Statistics to draw useful insights from the data”

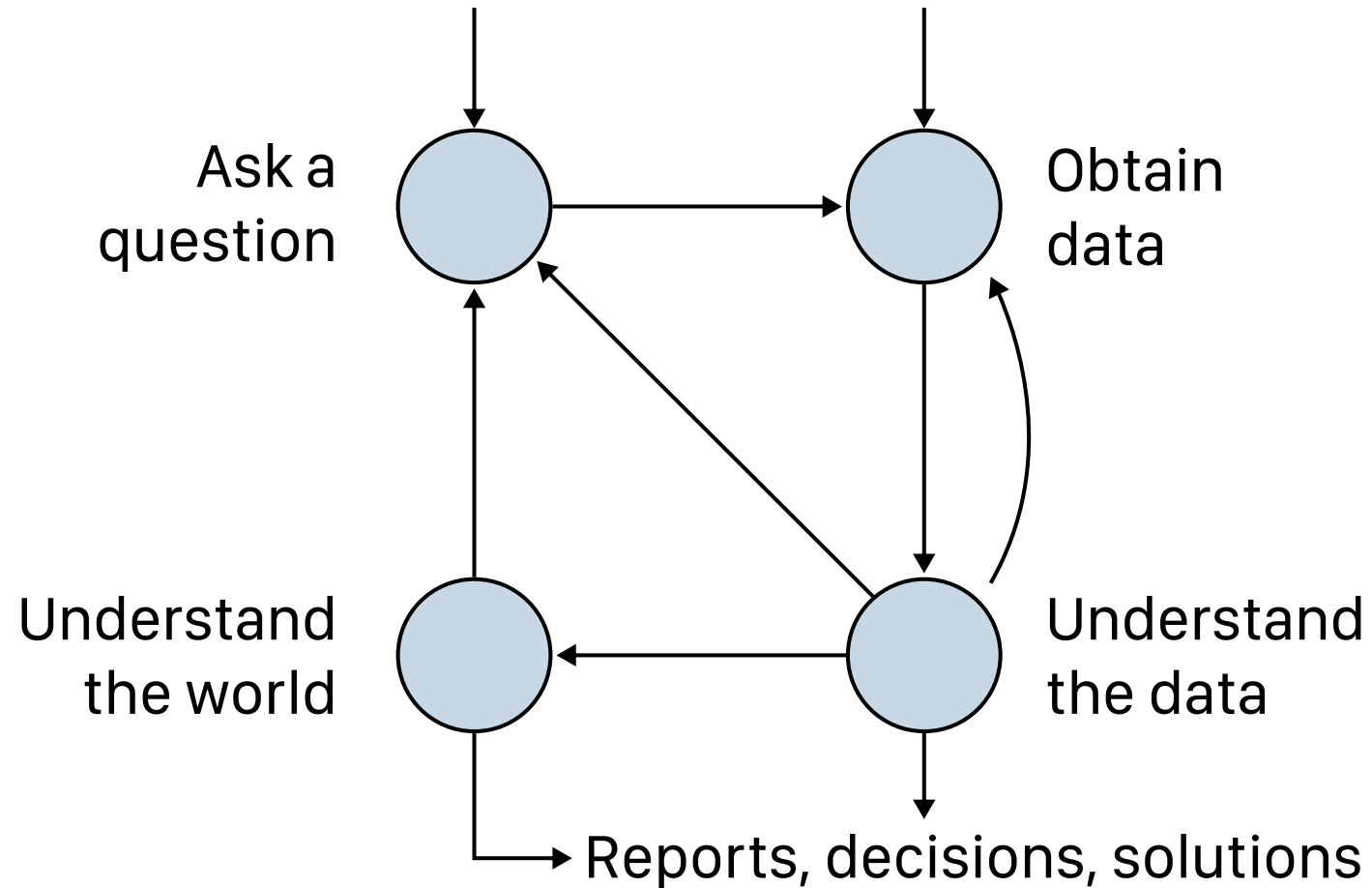


- *Data Science is about drawing useful conclusions from large and diverse data sets through **exploration**, **prediction**, and **inference**.*
 - *Data science extends the field of statistics by taking full advantage of computing, data visualization, machine learning, optimization, and access to information.*

- Course goals:
 - Prepare for courses on Machine Learning, Statistics, and Data Management.
 - Learn real-world tools and techniques.
 - Practice applying the concepts to real problems.

Introduction: Data Science lifecycle

- Lifecycle of a data science project: *There various versions... we'll follow [LDS]*



Introduction: Causality

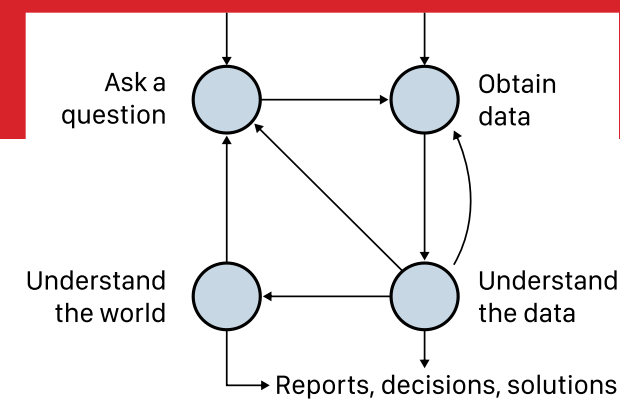
- The main challenge of data science is to draw robust conclusions about the world using incomplete information. *Inferential thinking* is one of the main ideas.
- A key aspect of inferential thinking is assigning cause to effect. Questions such as "*is coffee good for you?*", "*does coffee cause cancer?*" This is known as *causality*.
- Remaining discussion today: Key concepts in establishing causality

Introduction: Causality

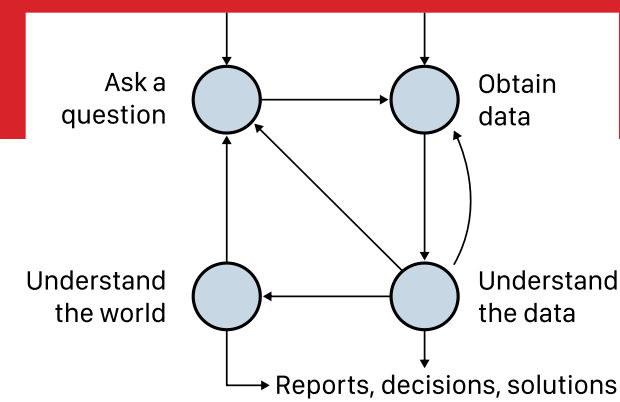
- *Observational study*: A study in which we make conclusions based on data we have observed but had no hand in generating.
- In data science, many such studies involve observations on a group of individuals, a factor of interest called a *treatment*, and an *outcome* measured on each individual.
- The fundamental question: “*Does the treatment have an effect on the outcome?*”
 - Any relation between the treatment and the outcome is called an *association*.
 - If the treatment causes the outcome to occur, then the association is *causal*.
- The establishment of causality often takes place in two stages. First, an association is observed. Next, a more careful analysis leads to a decision about causality.

Introduction: Causality

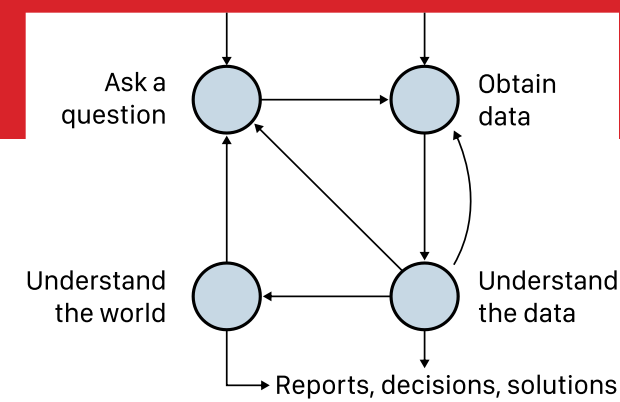
- Case study: *Cholera outbreaks in 1950's London*
- London in the 1850's:
 - Wealthiest city as well as poorest in many parts
 - Charles Dickens
 - **Cholera**
- Leading theory for Cholera: **Miasma** (*invisible poisonous particles arising out of decaying matter and manifesting as bad smell*)
 - Solution: Hold a sweet-smelling thing to the nose
- The data analyst in this case study was a miasma theory skeptic by the name **John Snow**.



- Question 1: *Is cholera caused by miasma?*
- Data 1: Map of cholera deaths
- Observation 1: *While entire households were wiped out by cholera, the people in neighboring houses sometimes remained completely unaffected.*
- Understanding 1: *There was no compelling association between bad smells and the incidence of cholera.*



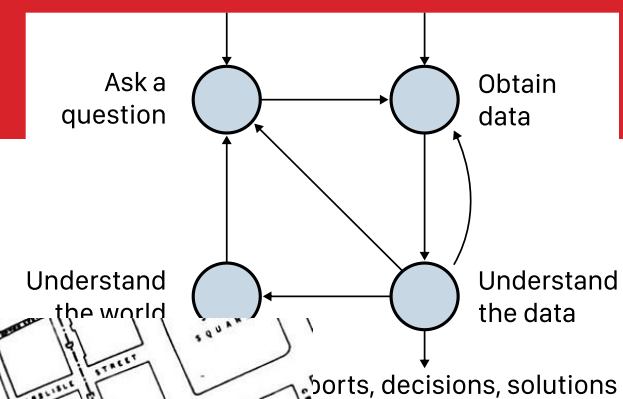
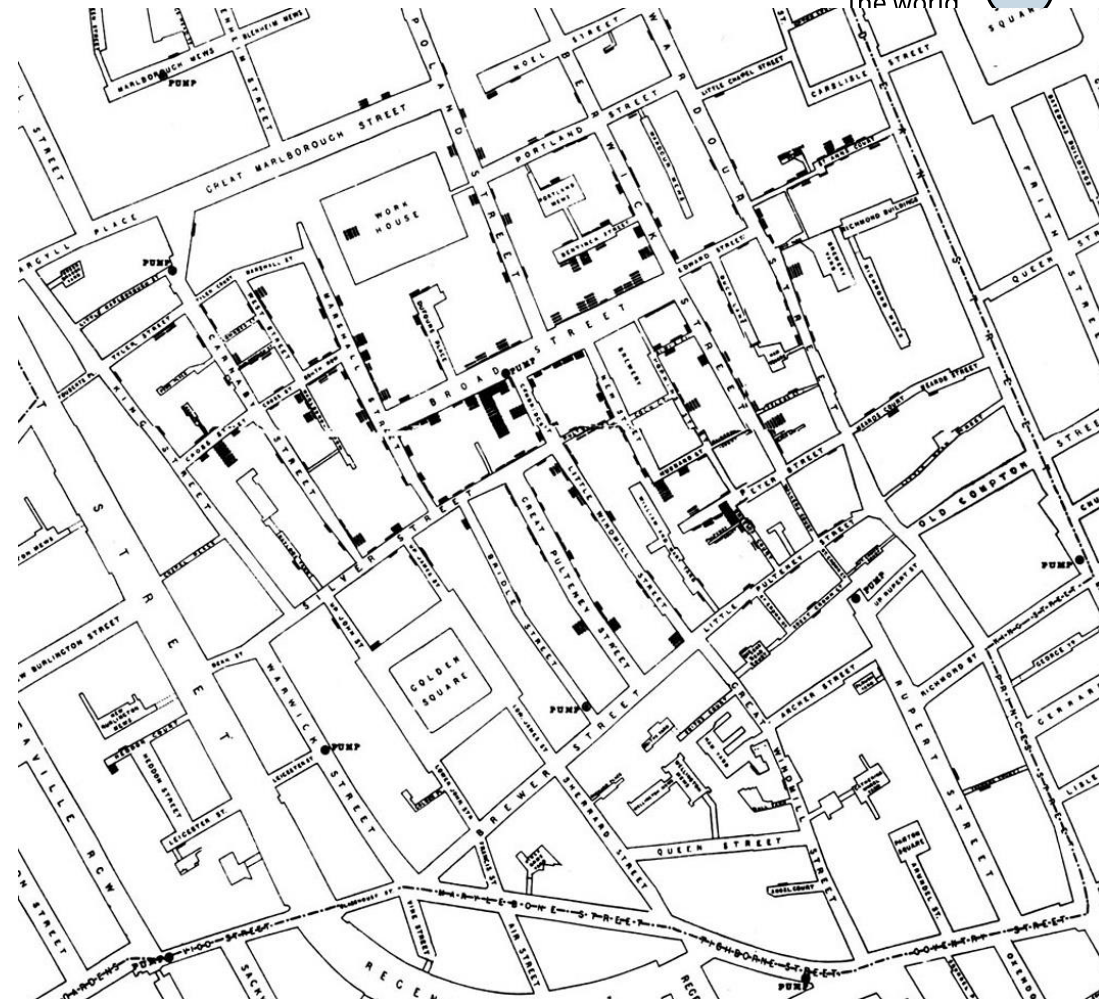
- Question 2: *What causes Cholera?*
- Data 2: Symptoms
- Observation 2: *The onset of the disease almost always involved vomiting and diarrhea.*
- Understanding 2: *The infection was carried by something people ate or drank, not by the air that they breathed.*

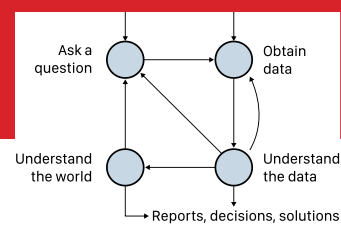


- Question 3: *What causes Cholera?*
- Data 3: *At the end of August 1854, cholera struck in the overcrowded Soho district of London. As the deaths mounted, Snow recorded them diligently, using a method that went on to become standard in the study of how diseases spread: he drew a map. On a street map of the district, he recorded the location of each death.*
- Observation 3: *The deaths were roughly clustered around the Broad Street pump.*
- Understanding 3: *There is an association between the Broad street pump location and the Cholera outbreak.*

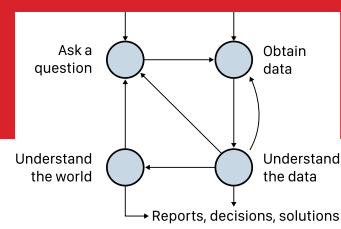
Introduction: Causality

- Question 3: *What causes Cholera?*
- Data 3: *At the end of August 1854, cholera struck in the overcrowded Soho district of London. As the deaths mounted, Snow recorded them diligently, using a method that went on to become standard in the study of how diseases spread: he drew a map. On a street map of the district, he recorded the location of each death.*
- Observation 3: *The deaths were roughly clustered around the Broad Street pump.*
- Understanding 3: *There is an association between the Broad street pump location and the Cholera outbreak.*





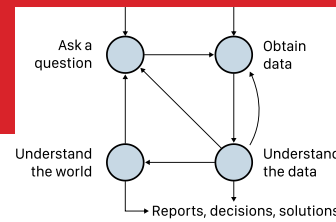
- Question 4: *Is the water from broad street pump the cause of cholera?*
- Data + observation 4:
 - *There were deaths in houses that were nearer the Rupert Street pump than the Broad Street pump. Though the Rupert Street pump was closer as the crow flies, it was less convenient to get to because of dead ends and the layout of the streets. The residents in those houses used the Broad Street pump instead.*
 - *There were no deaths in two blocks just east of the pump. That was the location of the Lion Brewery, where the workers drank what they brewed. If they wanted water, the brewery had its own well.*
 - *There were scattered deaths in houses several blocks away from the Broad Street pump. Those were children who drank from the Broad Street pump on their way to school. The pump's water was known to be cool and refreshing.*
 - *There were two isolated deaths in the leafy and genteel Hampstead area, quite far from Soho. Snow was puzzled by these until he learned that the deceased were Mrs. Susannah Eley, who had once lived in Broad Street, and her niece. Mrs. Eley had water from the Broad Street pump delivered to her in Hampstead every day. She liked its taste.*
- Understanding 4: *Water from the Broad street pump was causing cholera.*



- Question 4: *Is the water from broad street pump the cause of cholera?*
- Data + observation 4: Data to rule out anomalies
- Understanding 4: *Water from the Broad street pump was causing cholera.*

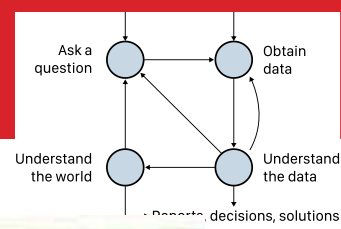


- *Snow used his map to convince local authorities to remove the **handle of the Broad Street pump**.*
- *Later it was discovered that a cesspit that was just a few feet away from the well of the Broad Street pump had been leaking into the well. Thus the pump's water was contaminated by sewage from the houses of cholera victims.*

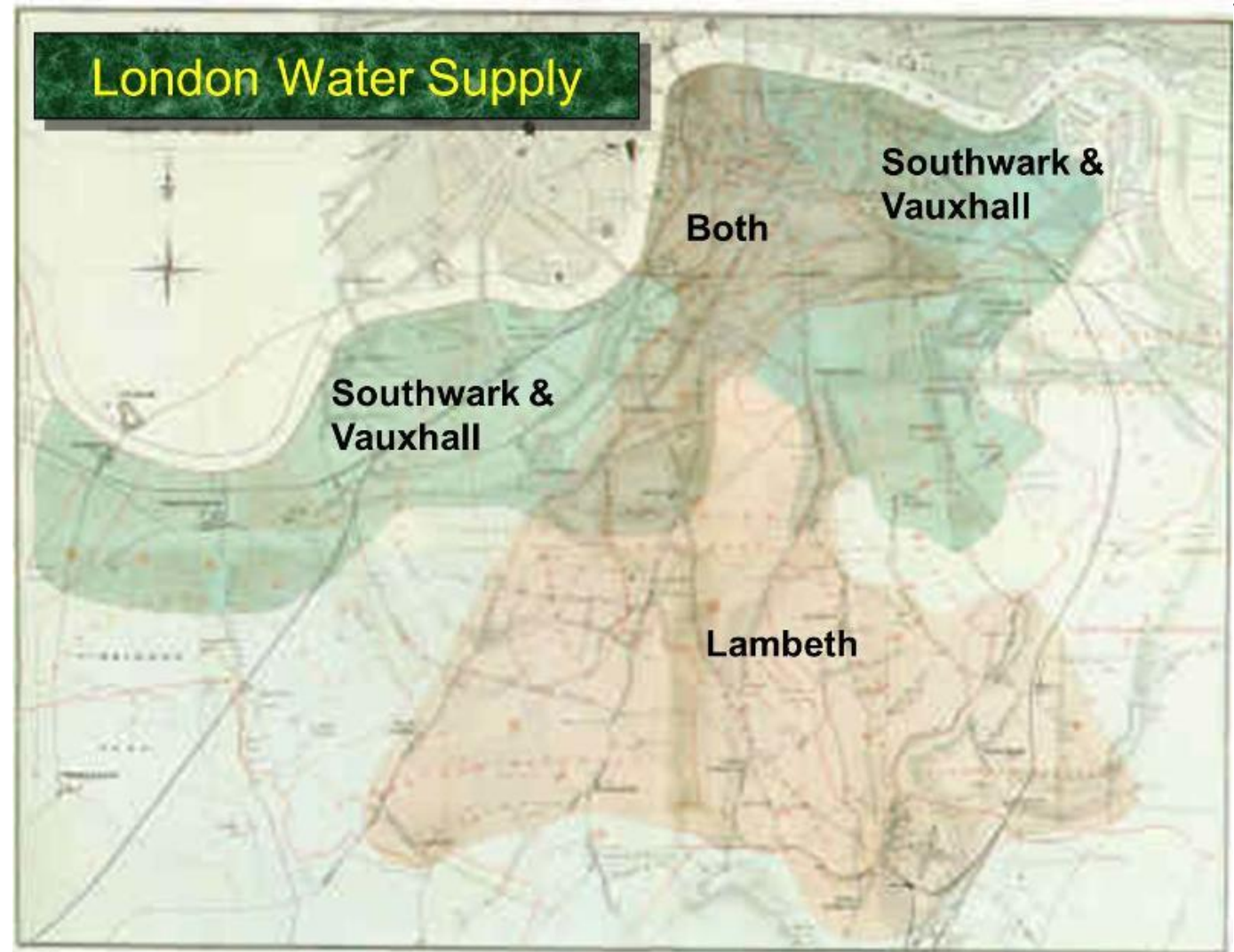


- Question 5: *Is the sewage-tainted water the cause of cholera?*
- This time we would like to obtain a robust conclusion.
- Method of comparison: We use comparison to identify an association between a treatment and an outcome. We compare the outcomes of a group of individuals who got the treatment (the *treatment group*) to the outcomes of a group who did not (the *control group*).
- Data 5: *Map of data on cholera deaths in an area of London that was served by two water companies. The Lambeth water company drew its water upriver from where sewage was discharged into the River Thames. Its water was relatively clean. But the Southwark and Vauxhall (S&V) company drew its water below the sewage discharge, and thus its supply was contaminated.*

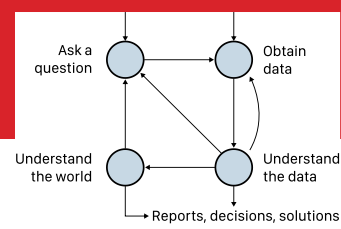
Introduction: Causality



- Question 5: *Is the sewage-tainted water the cause of cholera?*
- This time we would like to obtain a robust conclusion.
- **Method of comparison**: We use comparison to identify an association between a treatment and an outcome. We compare the outcomes of a group of individuals who got the treatment (the **treatment group**) to the outcomes of a group who did not (the **control group**).
- Data 5: *Map of data on cholera deaths in an area of London that was served by two water companies. The Lambeth water company drew its water upriver from where sewage was discharged into the River Thames. Its water was relatively clean. But the Southwark and Vauxhall (S&V) company drew its water below the sewage discharge, and thus its supply was contaminated.*



Introduction: Causality

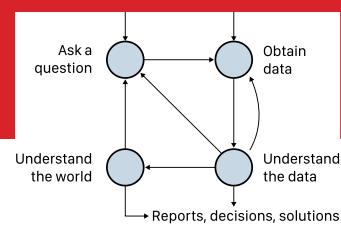


- Question 5: *Is the sewage-tainted water the cause of cholera?*
- Method of comparison: We use comparison to identify an association between a treatment and an outcome. We compare the outcomes of a group of individuals who got the treatment (the **treatment group**) to the outcomes of a group who did not (the **control group**).
- Data 5: *Map of data on cholera deaths in an area of London that was served by two water companies. The Lambeth water company drew its water upriver from where sewage was discharged into the River Thames. Its water was relatively clean. But the Southwark and Vauxhall (S&V) company drew its water below the sewage discharge, and thus its supply was contaminated.*

Supply Area	Number of houses	cholera deaths	deaths per 10,000 houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

- Comparison groups:
 - Treatment group: people in S&V houses
 - Control group: people in Lambeth houses
- *Is this data sufficient to make a conclusion? If not, what is the issue?*

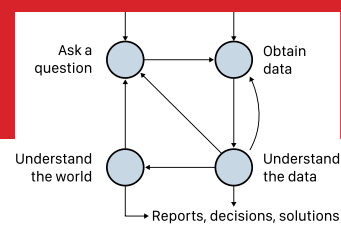
Introduction: Causality



Supply Area	Number of houses	cholera deaths	deaths per 10,000 houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

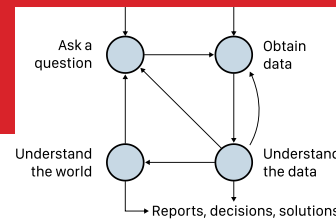
- Comparison groups:
 - Treatment group: people in S&V houses
 - Control group: people in Lambeth houses
- *Is this data sufficient to make a conclusion? If not, what is the issue?*
 - In an observational study, if the treatment and control groups differ in ways other than the treatment, it is difficult to make conclusions about causality.
 - An underlying difference between the two groups (other than the treatment) is called a **confounding** factor, because it might confound you (that is, mess you up) when you try to reach a conclusion.
 - Example: Studies in the 1960's showed that coffee drinkers had higher rates of lung cancer than those who did not drink coffee. The analysis contained a confounding factor—smoking. In those days, coffee drinkers were also likely to have been smokers, and smoking does cause lung cancer.

Introduction: Causality



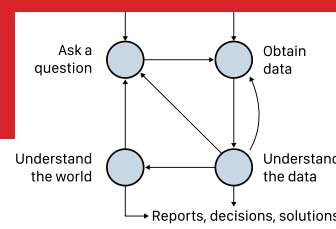
Supply Area	Number of houses	cholera deaths	deaths per 10,000 houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

- In order to establish whether it was the water supply that was causing cholera, Snow had to compare two groups that were similar to each other in all but one aspect—their water supply. Only then would he be able to ascribe the differences in their outcomes to the water supply. If the two groups had been different in some other way as well, it would have been difficult to point the finger at the water supply as the source of the disease. For example, if the treatment group consisted of factory workers and the control group did not, then differences between the outcomes in the two groups could have been due to the water supply, or to factory work, or both. The final picture would have been much more fuzzy.*



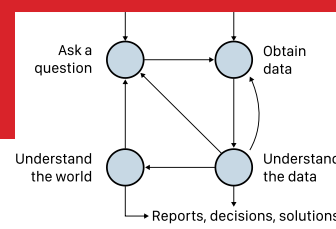
- Comparison groups:
 - Treatment group
 - Control group
- In an observational study, if the treatment and control groups differ in ways other than the treatment, it is difficult to make conclusions about causality.
- An underlying difference between the two groups (other than the treatment) is called a **confounding** factor, because it might confound you (that is, mess you up) when you try to reach a conclusion.
- *Is there a way to avoid confounding variables?*

Introduction: Causality



- Comparison groups:
 - Treatment group
 - Control group
- In an observational study, if the treatment and control groups differ in ways other than the treatment, it is difficult to make conclusions about causality.
- An underlying difference between the two groups (other than the treatment) is called a **confounding** factor, because it might confound you (that is, mess you up) when you try to reach a conclusion.
- *Is there a way to avoid confounding variables?*
 - **Randomization**: An excellent way to avoid confounding is to assign individuals to the treatment and control groups at random and then administer the treatment to those who were assigned to the treatment group. Randomization keeps the two groups similar apart from the treatment. This is known as randomized controlled experiment, also known as a **randomized controlled trial (RCT)**.

Introduction: Causality



- Comparison groups:
 - Treatment group
 - Control group
- In an observational study, if the treatment and control groups differ in ways other than the treatment, it is difficult to make conclusions about causality.
- An underlying difference between the two groups (other than the treatment) is called a **confounding** factor, because it might confound you (that is, mess you up) when you try to reach a conclusion.
- *Is there a way to avoid confounding variables?*
 - **Randomization**: An excellent way to avoid confounding is to assign individuals to the treatment and control groups at random and then administer the treatment to those who were assigned to the treatment group. Randomization keeps the two groups similar apart from the treatment. This is known as randomized controlled experiment, also known as a **randomized controlled trial (RCT)**.
 - Sometimes, people's responses in an experiment are influenced by their knowing which group they are in. So, you might want to run a blind experiment in which individuals do not know whether they are in the treatment group or the control group. To make this work, you will have to give the control group a **placebo**, which is something that looks exactly like the treatment but in fact has no effect.

End