CS410: Principles and Techniques of Data Science

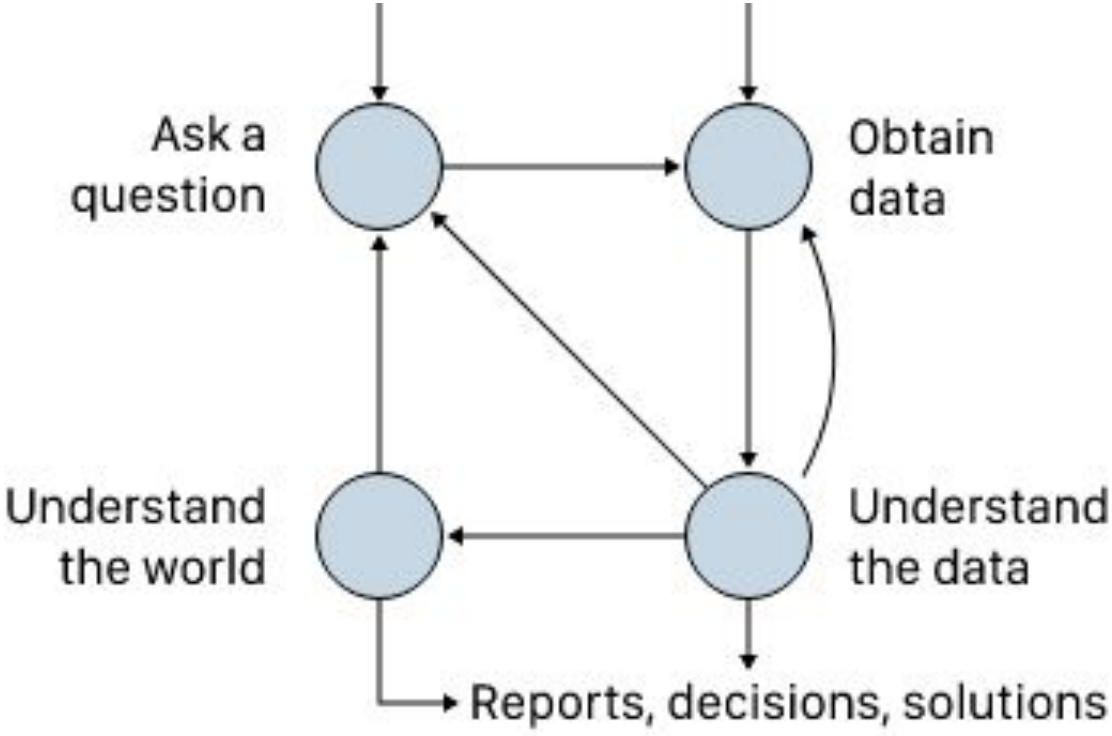
Module 2: The Data Science Lifecycle

Data Science Life Cycle

- Data science is a rapidly evolving field.
- Data science helps assess whether a vaccine works, filter out spam from our email inboxes, and advise urban planners where to build new housing.

Data Science Life Cycle

Data science lifecycle is split into four stages: asking a question, obtaining data, understanding the data, and understanding the world.



Asking a Question

- Asking questions lies at the heart of data science because different kinds of questions require different kinds of analyses.
- For example, “How have house prices changed over time?” is very different from
  “How will this new law affect house prices?”.
- Understanding our research question tells us what data we need, the patterns to look for, and how we should interpret our results.
- In this course, we focus on three broad categories of questions: exploratory, inferential, and predictive.

Exploratory Questions

- **Exploratory questions** aim to find out information about the data that we have.
- For example, we can use environmental data to ask: have average global temperatures risen in the past 40 years?
- The key part of an exploratory question is that it aims to summarize and interpret trends in the data without quantifying whether these trends will hold in data that we don't have.
- "How many people voted in the last election?" is an exploratory question.
  "How many people will vote in the next election?" is not an exploratory question.

Inferential Questions

- **Inferential questions** quantify whether trends found in our data will hold in unseen data.
- Let's say we have data from a sample of hospitals across the US.

"Whether air pollution is correlated with lung disease for the individuals in our sample"

- this is an exploratory question.

"Whether air pollution is correlated with lung disease for the entire US"

- this is an inferential question, since we're using our sample to *infer* a correlation for the entire US.

Predictive Questions

- **Predictive questions**, like inferential questions, aim to quantify trends for unseen data.
- While inferential questions look for trends in the population, predictive questions aim to make predictions for individuals.
- An inferential question could ask: "What factors increase voter turnout in the US?"

  A predictive question could ask: "Given a person's income and education, how likely are they to vote?"

We often change and refine our research questions. Each time we do so, it's important to consider what kind of question we want to answer.

Asking a Question

For a more detailed breakdown of the types of research questions, see (Leek and Peng, 2015).

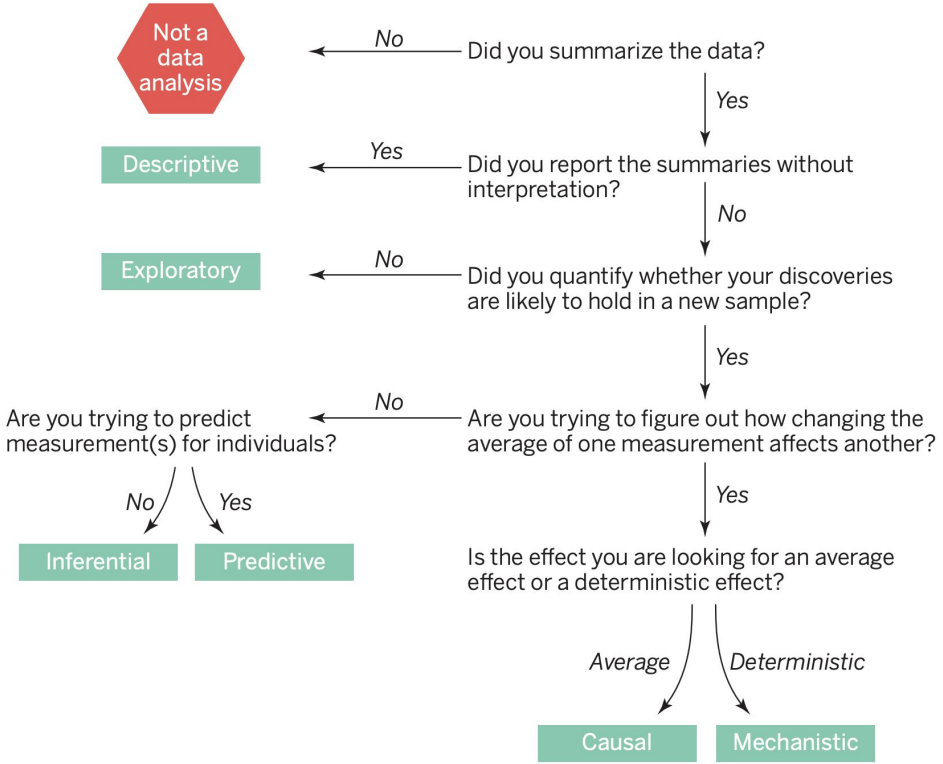
STATISTICS

What is the question?

Mistaking the type of question being considered is the most common error in data analysis

By **Jeffery T. Leek** *and* **Roger D. Peng**

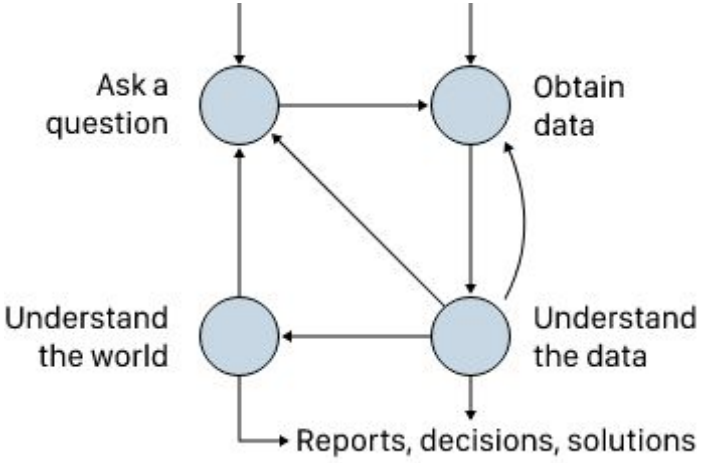
Data analysis flowchart



Obtaining Data

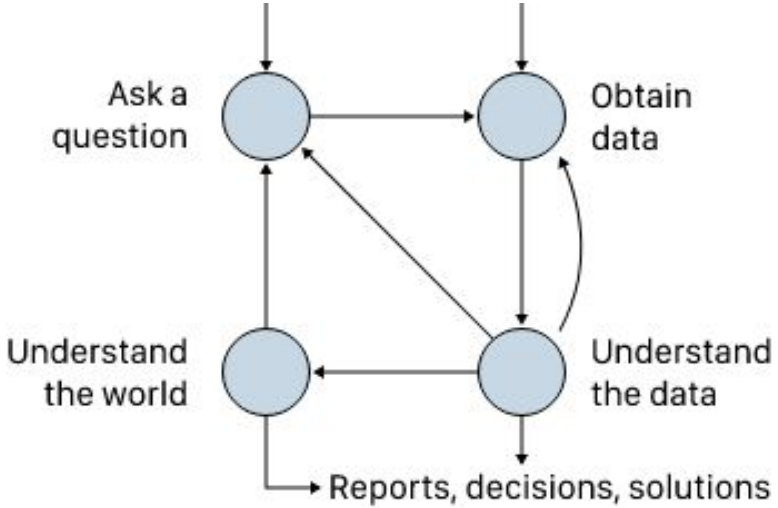
- In this step of the data science lifecycle, we obtain our data and understand how the data were collected.
- Data analyses can begin with asking a question (the previous stage) or with obtaining data (this stage).
- Data are expensive and hard to gather: define a precise research question first and then collect the exact data we need to answer the question.
- Data are cheap and easily accessed: start an analysis by obtaining data, exploring it, and then asking research questions.

  Eg: online data sources; the Twitter website lets people quickly download millions of data points.



Understanding the Data

- After obtaining data, we want to understand the data we have.
- *Exploratory data analysis*: create plots to uncover interesting patterns and summarize the data visually. Look for problems in the data: missing values, weird values, or other anomalies that we need to account for.
- This stage is highly iterative. Understanding the data can lead to any of the other stages in the data science lifecycle. As we understand the data more, we often revise our research questions, or realize that we need to get data from a different source.



Understanding the Data

- When our research questions are purely exploratory, we are only concerned about patterns in the data. In these cases, our analysis can end at this stage of the life cycle.
- When our research questions are inferential or predictive, however, we proceed to the next stage of the life cycle: understanding the world.

Understanding the World

- Draw conclusions about our larger population, and sometimes even the world.

- To draw conclusions about the population, we use use models like linear or logistic regression. This stage is relevant when our research questions are inferential or predictive.

Our goal is to quantify how well we think the trends we find in our sample can generalize.

THANK YOU!