# Predictive Analytics Foundations

Lecture 1

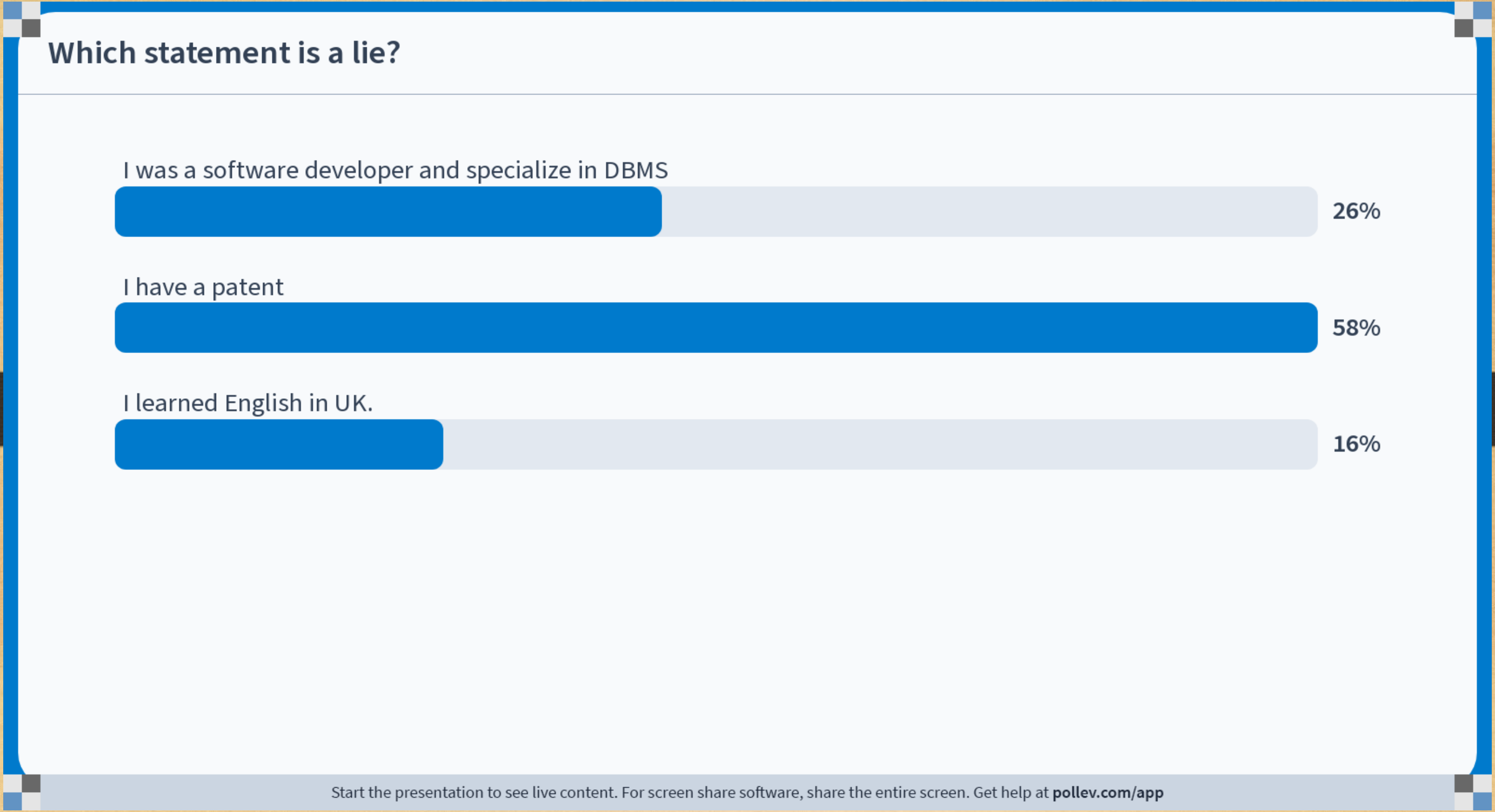
Lecture 1

- **Intro**
- What is predictive analytics?
- Data Science Lifecycle
- What will you learn in this class?
- Demo

# Introduction

# Meet with your instructor

# Which statement is a lie?



Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Lecture 1

- Intro
- **What is predictive analytics?**
- Data Science Lifecycle
- What will you learn in this class?
- Demo

# Technology Trends

2020s ● ?

2010s ● Data Industry

- ➢ Collect and sell information

2000s ● Internet Industry

- ➢ Online retailers and services

1990s ● Software Industry

- ➢ Sold computer software

1980s ● Hardware Industry

- ➢ Sold computers

# Predictive Analytics

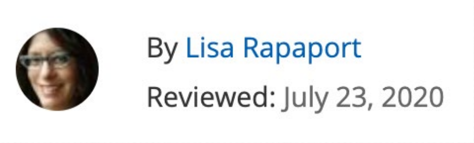
- **Exploration**
  - Identifying patterns in data
  - Uses visualizations
- **Inference**
  - Using data to draw reliable conclusions about the world
  - Uses statistics
- **Prediction**
  - Making informed guesses about unobserved data
  - Uses machine learning

# Some (broad) questions we might try to answer with predictive analytics

- In which markets should we focus our advertising campaign?
- Should I send my kids to daycare?
- Is the world getting better or worse?
- What areas of the world are at higher risks for climate change impact in 10 years? 20?
- What should we eat to avoid dying early of heart disease?
- Do immigrants from poor countries have a positive or negative impact on the economy?

# Why Predictive Analytics Matters

# Regularly Eating Chocolate Is Linked to 8 Percent Lower Heart Attack Risk



By Lisa Rapaport

Reviewed: July 23, 2020

Report on a July 2020 article in the European Journal of Preventive Cardiology

https://journals.sagepub.com/doi/full/10.1177/2047487320936787

https://www.everydayhealth.com/diet-nutrition/eating-chocolate-regularly-linked-to-lower-heart-attack-risk/

# Observation

- **individuals**, study subjects, participants, units
  - *336,289 US, Swedish, and Australian adults in several studies*
- **treatment**
  - *chocolate consumption*
- **outcome**
  - *heart disease*

# The First Question

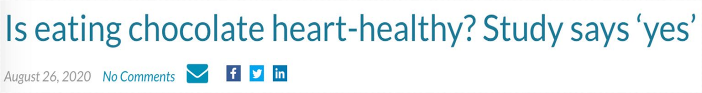
Is there any relation between chocolate consumption and heart disease?

- **association**
  - any relation
  - link

Answer: Yes, because those who ate chocolate had less heart disease.

Other headlines about the same article:

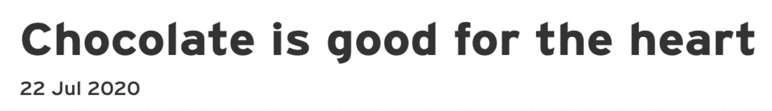
# A Stronger Link?



Is eating chocolate heart-healthy? Study says ‘yes’

August 26, 2020 No Comments

Family Safety and Health, National Safety Council



**Chocolate is good for the heart**

22 Jul 2020

European Society of Cardiology Press Release

https://www.escardio.org/The-ESC/Press-Office/Press-releases/Chocolate-is-good-for-the-heart

https://www.safetyandhealthmagazine.com/articles/20257-is-eating-chocolate-heart-healthy-study-says-yes

# The Next Question

Does chocolate consumption lead to a reduction in heart disease?

- **causality**

This question is often harder to answer.

“Dr. Alice Lichtenstein, an American Heart Association volunteer and professor of nutrition science and policy at Tufts University, was more skeptical of the findings.”

Market Watch

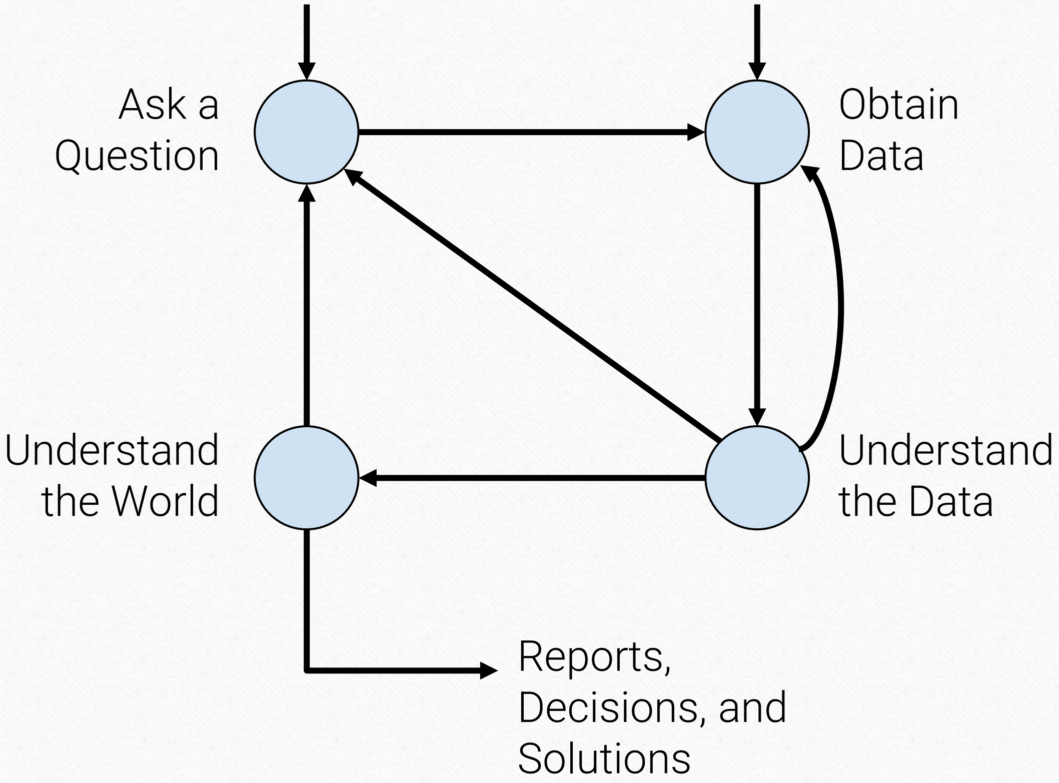
# Association ?
# Causation?

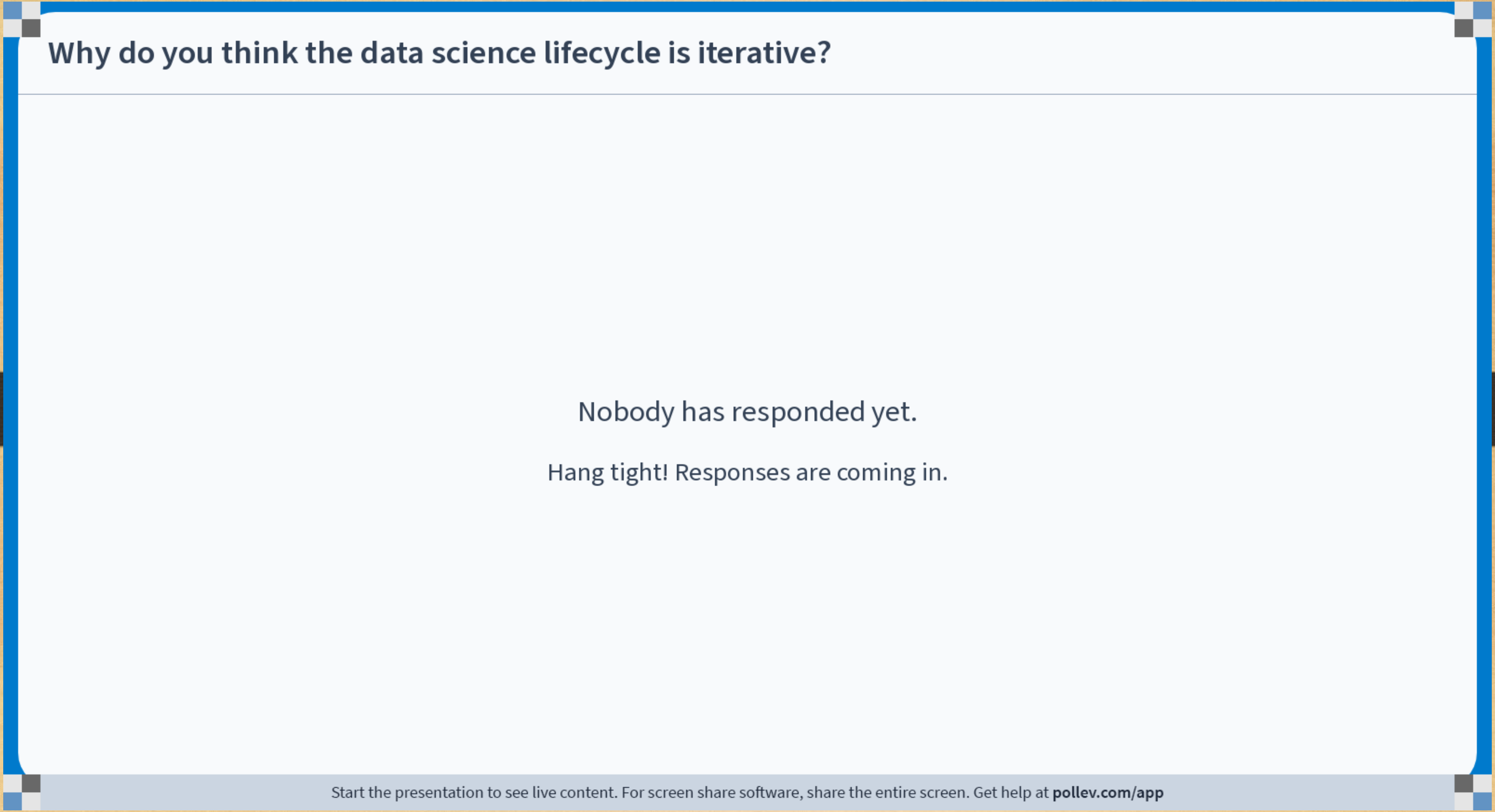
Prediction?

## Lecture 1

- Intro
- What is predictive analytics?
- **Data Science Lifecycle**
- What will you learn in this class?
- Demo

## Data Science Lifecycle





## Why do you think the data science lifecycle is iterative?

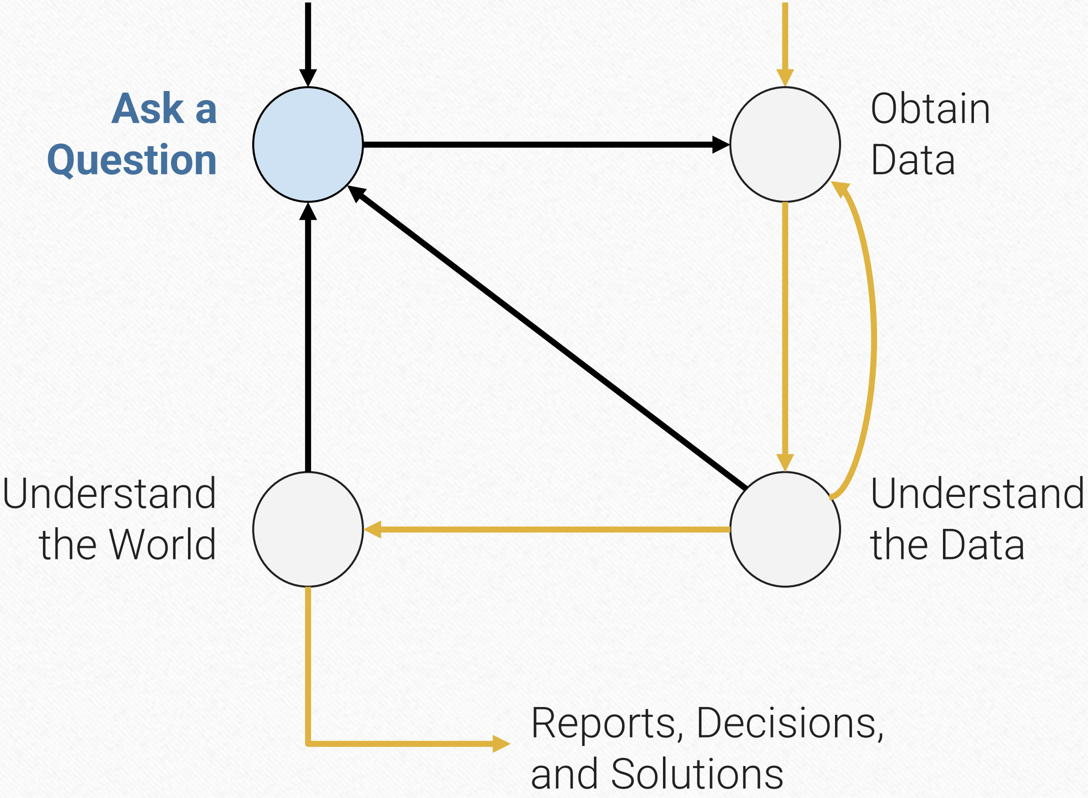
Nobody has responded yet.

Hang tight! Responses are coming in.

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

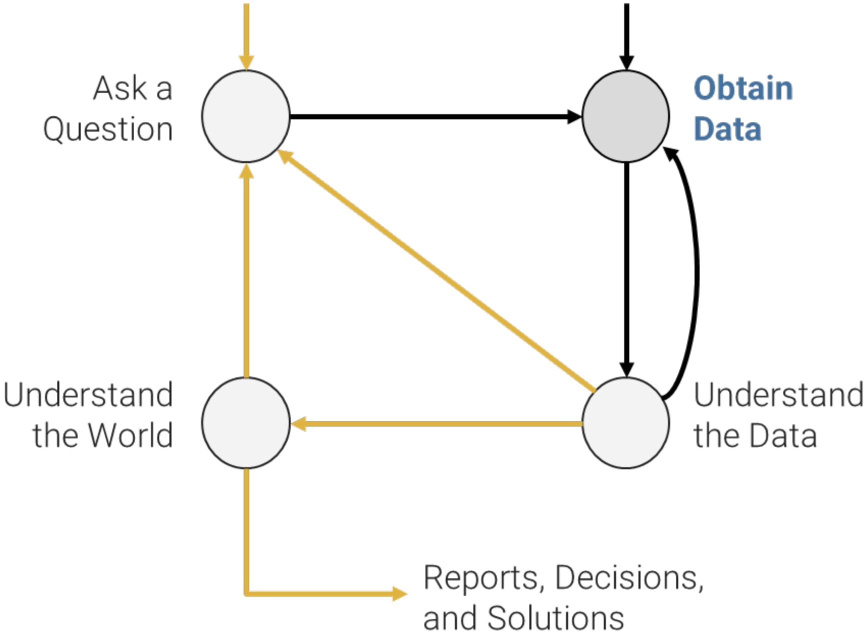
## 1. Question/Problem Formulation

- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics for success?



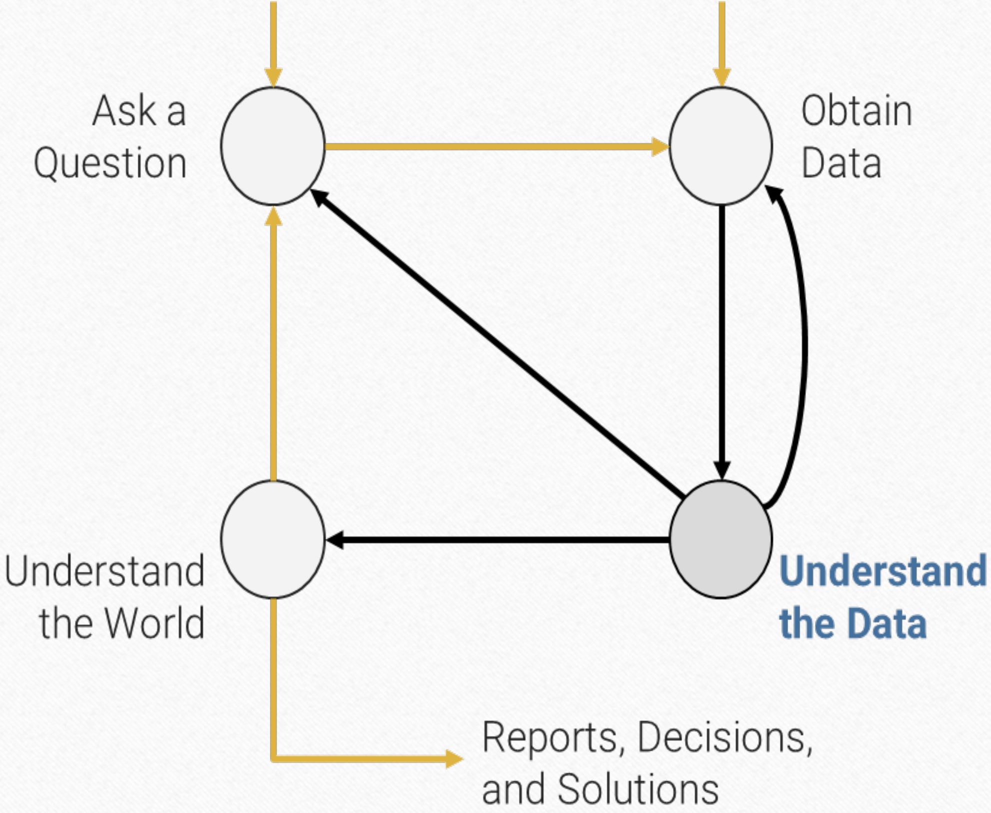
## 2. Data Acquisition and Cleaning

- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?



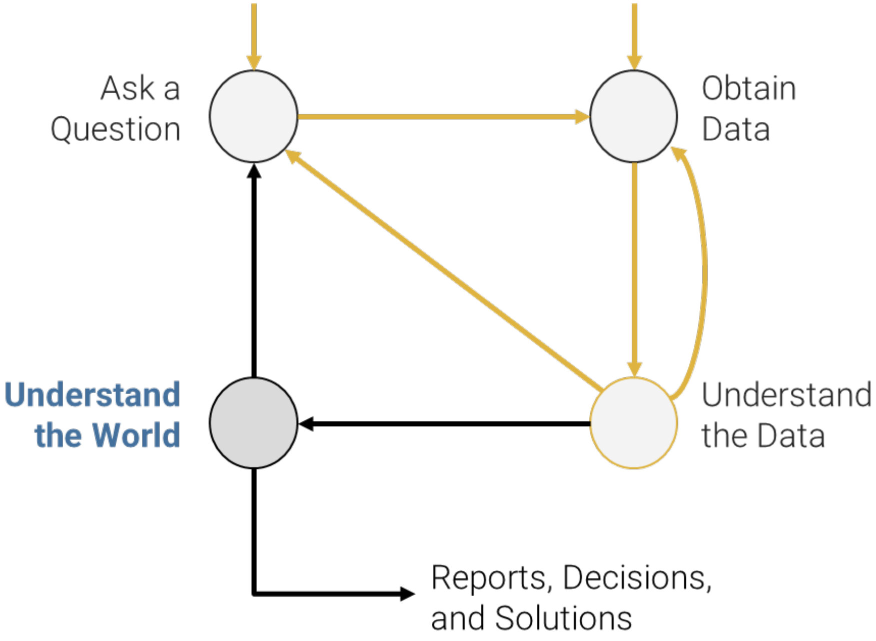
# 3. Exploratory Data Analysis & Visualization

- How is our data organized and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?



# 4. Prediction and Inference

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



## Lecture 1

- Intro
- What is predictive analytics?
- Data Science Lifecycle
- **What will you learn in this class?**
- Demo

# Course Content

# Textbook

# Assessments

- Lab exercises
- Quiz
- Term Exams
- ML Assignments
- Project

# Collaboration

Asking questions is greatly encouraged

- Discuss questions with each other (except exams)
- Submit lab assignments individually

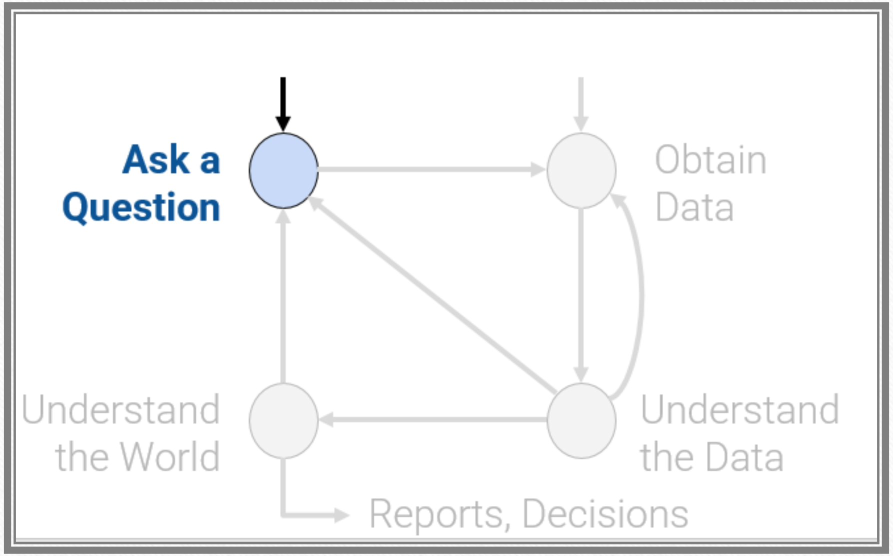
The Limits of collaboration

- Don't share solutions with each other
- Copying or other dishonesty will result in severe penalties

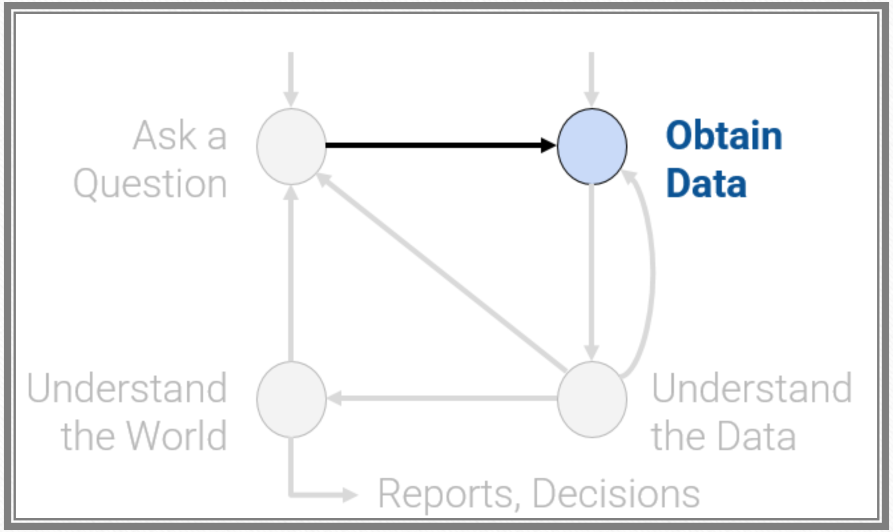
## Lecture 1

- Intro
- What is predictive analytics?
- Data Science Lifecycle
- What will you learn in this class?
- **Demo**

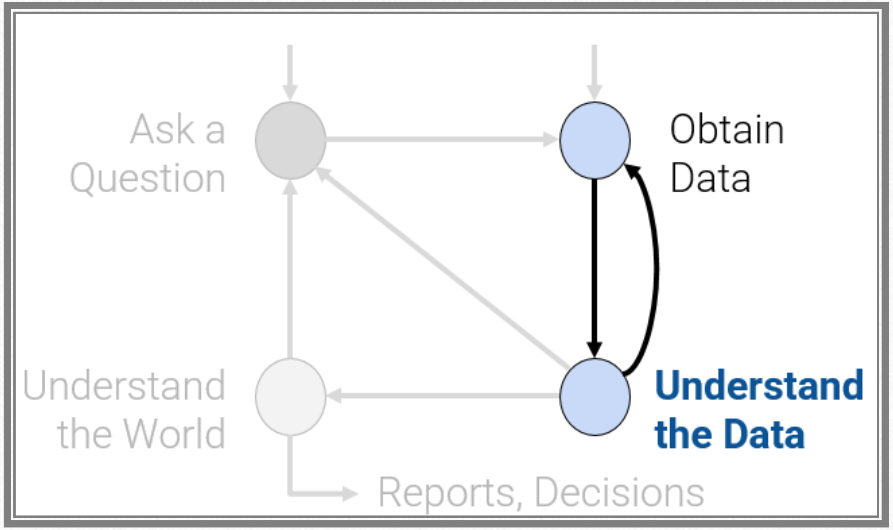
# Ask a Question: Who are you?



# Data Acquisition and Cleaning



# EDA and Visualization

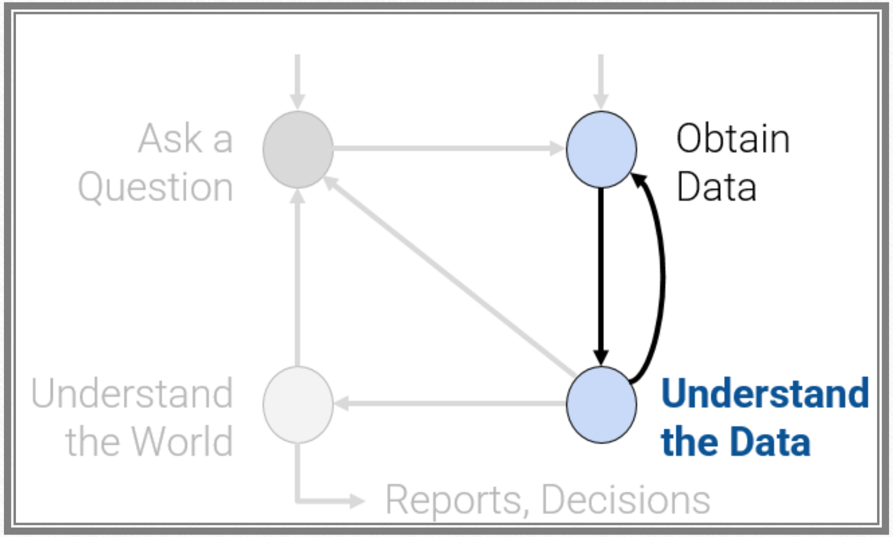


# EDA and Visualization

Population: Data students

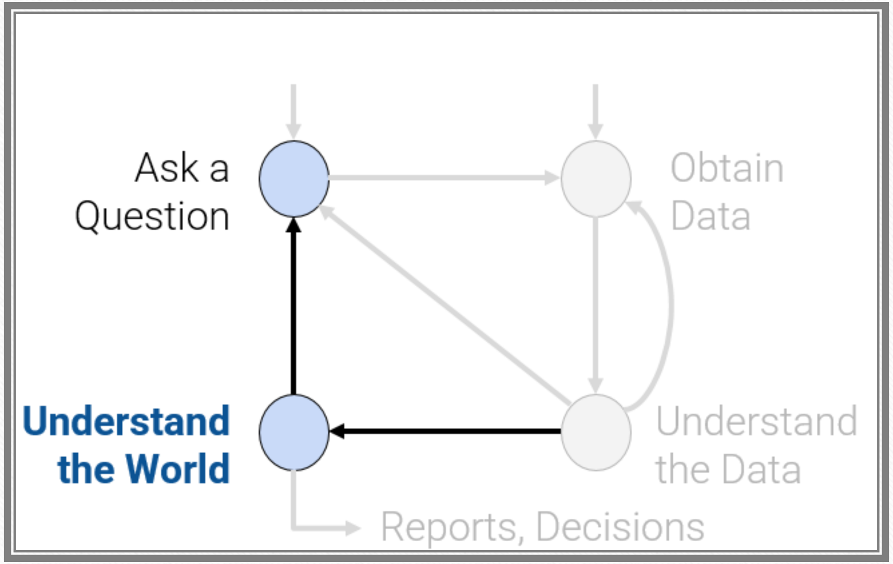
Some sub-questions:

- How many students are in the class?
- What are your majors?
- What year are you?
- Diversity ...?



# What fraction of the students are female?

This is a complex question. Are we asking about **sex** (biological trait) or **gender** (individual, social, cultural identity)?



# What is the gender diversity of all the data classes?

We don’t currently have data to answer this question. We could either:

- Survey the students, or…
- …Use the data we have to estimate the sex of the students as a proxy for gender???*

# Again, but for Baby Names Data

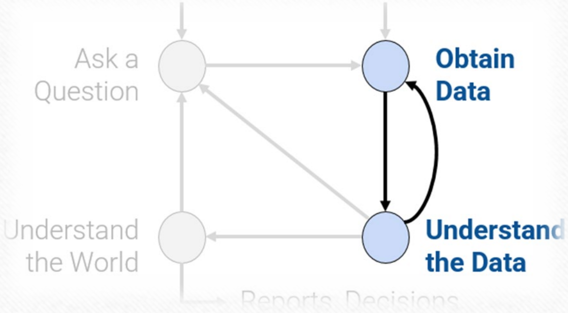
1. Can we estimate a person's sex using their name?
2. Obtain more data: SSN Baby Names

**Discuss**: Based on the description of the SSN data: What are **limitations** of this datasource?
What limitations might it have with respect to our original task?

# Again, but for Baby Names Data

What does each row/column represent?

What can you observe about how U.S. baby names have changed over time?



Reports, Decisions

# Prediction and Inference: Simple Classifier

Let’s use this data to estimate the fraction of female students in the class.

# What are some limitations of our analysis?

Possible limitations:

- U.S. name data, not global data
- Everyone born since 1937
- No “rare” names
- Sex as a proxy for gender

… …

# What’s the point of this demo?

There are many assumptions in data science:

- Whether the data is representative:
  - Of the question being asked
  - Of the world and its implications
- Beliefs/backgrounds of data collectors
- Beliefs/backgrounds of data analysts
- Beliefs/backgrounds of the population