# CS410: Principles and Techniques of Data Science

## Module 3: Questions and Data Scope

# Introduction

- Data scientists use data to answer questions
- Quality of the data collection process can significantly impact the strength of conclusions we draw from an analysis, and the decisions we make.
- Ideally, we aim for data to be representative of the phenomenon we are studying.
- If our data are not representative of the object of our study, then our conclusions can be limited, possibly misleading, or even wrong.

Quality data is important

# Big Data and New Opportunities

- When we have large amounts of administrative data or expansive digital traces, it can be tempting to treat them as more definitive than data collected from traditional smaller research studies. We might even consider these large datasets as a replacement for scientific studies or essentially a census.
- One well-known example is the Google Flu Trends tracking system.

Big data is not always good.

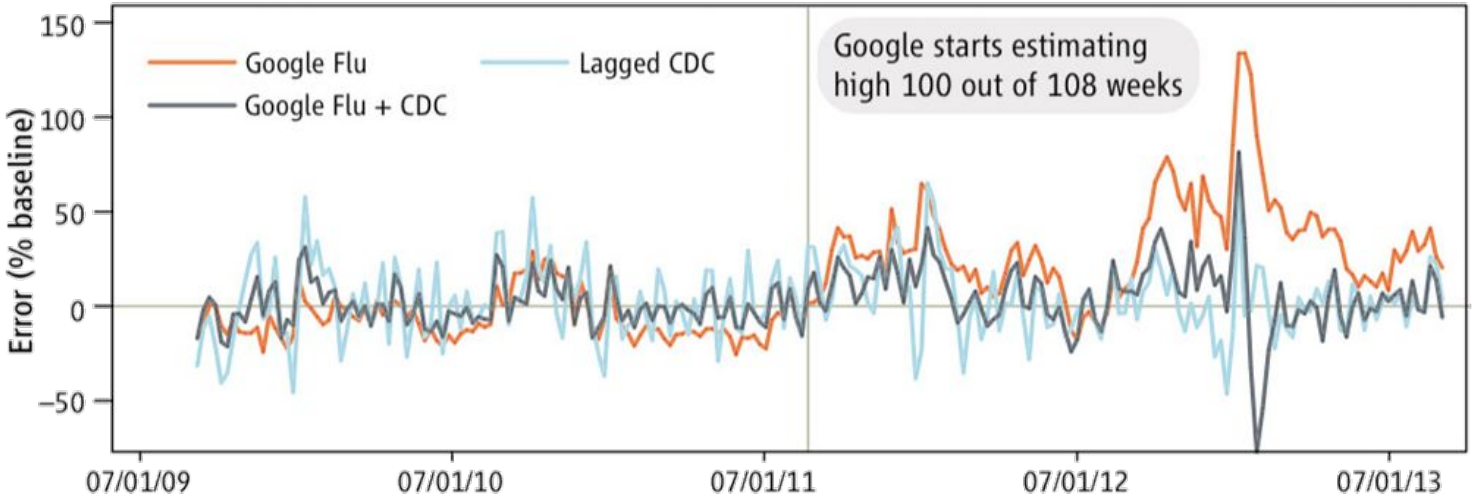
# Example: Google Flu Trends

- In 2007, researchers found that counting the searches people made for flu-related terms could accurately estimate the number of flu cases. It made headlines, and helped make researchers excited about the possibilities of big data. However, GFT did not live up to expectations and was abandoned in 2015.

# Example: Google Flu Trends

- **What went wrong with GFT?** It used millions of digital traces from online queries for terms related to influenza to predict flu activity. Despite initial success, in the 2011–2012 flu season, Google's data scientists found that the GFT was not a substitute for more traditionally collected data from the Centers for Disease Control (CDC) surveillance reports, collected from laboratories across the United States.
- In comparison, GFT overestimated the CDC numbers for 100 out of 108 weeks.

  Week after week, GFT came in too high for the cases of influenza, even though it was based on big data.



# Example: Google Flu Trends

- Data scientists found that the GFT was not a substitute for more traditionally collected data from the CDC.
- A simple model built from past CDC reports that used 3-week-old CDC data and seasonal trends did a better job of predicting flu prevalence than GFT. This does not mean that big data captured from online activity is useless.
- Researchers have shown that the combination of GFT data with CDC data can substantially improve on both GFT predictions and the CDC-based model (Lazer, 2015).

# Instruments and Protocols

***Instrument***: used to take the measurements

***Protocol***: procedure for taking measurements

**Example**: For a survey, the instrument is typically a questionnaire that an individual in the sample answers. The protocol for a survey includes how the sample is chosen, how nonrespondents are followed up, interviewer training, protections for confidentiality, etc.

Good instruments and protocols are important to all kinds of data collection.

# Summary: Questions and Data Scope

It is important to answer the following questions:

- **Who collected the data?**
- **Why were the data collected?**
- **When were the data collected?**
- **Where were the data collected?**
- **What methods were used to select samples/take measurements?**
- **What instruments were used and how were they calibrated?**

Answers to these questions give you valuable insights as to how much trust you can place in your findings.

THANK YOU!