Methodology, Statistics and Pitfalls

Data Science and Society Statistics Lecture 2

Matthieu Brinkhuis

October ~~9~~ 16, 2018

Utrecht University, Information and Computing Sciences
m.j.s.brinkhuis@uu.nl

Introduction

Tutorial on R/Spark

How was the tutorial?

Please reach out if you encountered problems!

Topics today

Today discuss the paper by D Lazer et al. (2014a). “The Parable of Google Flu: Traps in Big Data Analysis”. In: *Science* 343.6176, pp. 1203–1205. DOI: 10.1126/science.1248506, with discussion by Broniatowski, Paul, and Dredze (2014) and Lazer et al. (2014b)

Intended learning outcomes

By the end of this lecture, you'll be able to:

- Recognize that methodology for data science is crucial (1)
- Understand different types of pitfalls (2)
- Apply the principles in your own research (3)
- Analyze potential traps (4,5)

Bloom's Taxonomy:

1. Remember
2. Understand
3. Apply
4. Analyze
5. Evaluate
6. Create

Table of Contents

Parable of the Google Flu paper

1. Big Data Hubris
2. Algorithm Dynamics
3. Transparency, Granularity, and All-Data

Research on whether search or social media can predict x has become commonplace and is often put in sharp contrast with traditional methods and hypotheses. (Lazer et al., 2014a)

Big Data Hubris

Hubris

- *quantity of data does not mean that one can ignore foundational issues of measurement, construct validity and reliability*
- overfitting
- flu vs winter detector
- specific model misfit
- CDC data does better
- dynamically recalibration needed

Hubris and pitfalls

- *quantity of data does not mean that one can ignore foundational issues of measurement, construct validity and reliability* (garbage i&o, data quality)
- overfitting (spurious correlations, causality)
- flu vs winter detector (third variable, causality)
- specific model misfit (inference)
- CDC data does better (what do you predict against: baseline)
- dynamically recalibration needed (inference, complexity)

Algorithm Dynamics

Dynamics

What do you think this chapter is mostly about?

Dynamics and pitfalls

What do you think this chapter is mostly about?

Search algorithm itself changes

- Complexity due to changes in algorithm
- Replication problems (even using Google Correlate)
- Changes on commercial aspects and suggesting searches using trends

Transparency, Granularity, and All-Data

Why this paper?

The GFT parable is important as a case study where we can learn critical lessons as we move forward in the age of big data analysis.

- Transparency and Replicability
- Use Big Data to Understand the Unknown (granularity)
- Study the Algorithm
- It’s Not Just About Size of the Data

Closing statement

Instead, traditional "small data" often offer information that is not contained (or containable) in big data [...]. Instead of focusing on a "big data revolution," perhaps it is time we were focused on an "all data revolution," where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world.

Closing

Reaction 1

Reply by Broniatowski, Paul, and Dredze (2014)

Many limitation overcome:

- By including Twitter data (replicability, overfitting, construct validity, granularity, and temporal confounds)
- Separation of awareness and infection on Twitter
- Now we are doing great (great correlation, externally validated)
- Do not generalize GFT to big data analyses

Reaction 2

Reply by Lazer et al. (2014b)

- Not all problems occur in the entire field
- Problems with data quality (changes by all sorts of parties) remains
- Still related to Zombies
- Build strong collaboration to *learn* from such data

Intended learning outcomes recap

By the end of this lecture, you'll be able to:

- Recognize that methodology for data science is crucial (1)
- Understand different types of pitfalls (2)
- Apply the principles in your own research (3)
- Analyze potential traps (4,5)

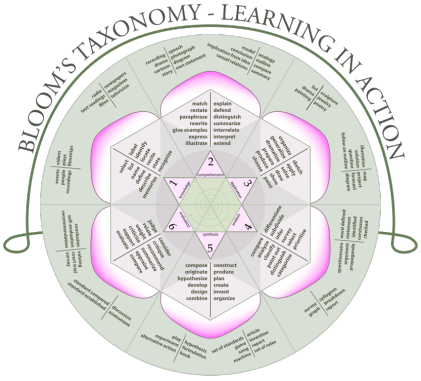

Figure 1: Bloom's Taxonomy (image from Wikipedia).

Thank you.

References i

- Broniatowski, DA et al. (2014). “Twitter: Big data opportunities”. In: *Science* 345.6193, p. 148. DOI: 10.1126/science.345.6193.148-a (cit. on pp. 4, 18).
- Lazer, D et al. (2014a). “The Parable of Google Flu: Traps in Big Data Analysis”. In: *Science* 343.6176, pp. 1203–1205. DOI: 10.1126/science.1248506 (cit. on pp. 4, 7).
- – (2014b). “Twitter: Big data opportunities - Response”. In: *Science* 345.6193, pp. 148–149. DOI: 10.1126/science.345.6193.148-b (cit. on pp. 4, 19).