STATISTICAL HORIZONS

Machine Learning with Text Data

Christopher Bail, Ph.D.

Upcoming Seminar:
April 26-27, 2019, Philadelphia, Pennsylvania

Welcome to Text as Data

Table of Contents

Chris Bail
Duke University
www.chrisbail.net

Introductions

Welcome to the Statistical Horizons "Text as Data" course. It will be my honor to introduce you to a range of topics over the next two days that include screenscraping, application programming interfaces, basic text analysis, dictionary-based text analysis, topic modeling, and text networks.

Schedule

I will aim to guide us through the aforementioned topics according to the schedule below:

Thursday (Collecting Text Data)

Time	Activity
9-9:15	Introductions/Housekeeping
9:15-10	Introduction to Text as Data
10-10:15	Break
10:15-11:15	Screen-Scraping Part I
11:15-11:30	Break
11:30-12	Screen-Scraping Part II
12-1:15	Lunch
1:15-2:15	Application Programming Interfaces Part I
2:15-2:30	Break
2:30-4	Application Programming Interfaces Part II (Breaks as needed)

Friday (Analyzing Text Data)

Time	Activity
9-10:00	Basic Text Analysis
10-10:15	Break
10:15-11:15	Dictionary-Based Text Analysis
11:15-11:30	Break
11:30-12:00	Topic Modeling
12-1:15	Lunch
1:15-2:15	Structural Topic Modeling
2:15-2:30	Break
2:30-4:00	Text Networks (Breaks as needed)

A note about pedagogy

In this course, you will encounter a number of coding techniques that fall far outside the scope of conventional statistical training. Though it may be tempting to approach this subject in the same way you might aim to learn about the latest techniques for causal inference or missing data, I am going to encourage you to focus on "learning how to learn" instead. This is because the field of text analysis is a) rapidly changing; and b) emphatically "open source", meaning that many of the techniques you are going to learn about weave together different types of coding techniques and different software packages that make it virtually certain that some part of the code I present in this course will not work for another case study, or will not work years from now.

In order to maximize the benefit of this class, we will also work through messy, real world examples in order to simulate various situations you may encounter (as, for example, when none of the official documentation of an R package helps you figure out how to intepret an error message).

Though I will be lecturing quite a bit in order to fit in as much material as possible, I have also developed in-class exercises for you to try in order to ensure that you are retaining the material. I also encourage you to interrupt me whenever you like if you have a question that you think will be of concern to the entire class. If you have an issue that is unlikely to be of interest to the entire class (such as an operating system issue), please find me during one of the breaks or during lunch.

Accessing the course materials

All course materials will be linked from my GitHub site which I will write on the wall of our classroom on the first day. In addition to the printed versions, you may find it helpful to open up the html versions as well, which include hyperlinks that may guide you through our case studies and examples more seemlessly.

Feedback

I want to monitor what you think about this class and whether or how it could be improved as much as possible. I invite you to submit anonymous feedback at the link below that I will monitor during all course breaks, lunches, and at the end of each day:
https://goo.gl/forms/HOyhFcAcKXGHq2403

An Introduction to Text As Data

Table of Contents

Chris Bail
Duke University
www.chrisbail.net

Introduction

This is the first in a series of tutorials I've created about collected data from web-based sources such as Facebook or Twitter and analyzing such data using a range of new techniques for automated text analysis. Before we proceed to the technical aspects of these techniques, I want to give you some sense of where the came from

A Brief History of Text as Data

The field of automated text analysis has been around for more than half a century, yet it has evolved very rapidly in recent years because of recent advances in the field of natural language processing. Perhaps the first person to propose the idea of quantifying patterns in text was Harold Laswell, who famously used this approach to study WWI propoganda. "We may classify references into categories," wrote Laswell in 1938, " according to the understanding which prevails among those who are accustomed to the symbols. References used in interviews may be quantified by counting the number of references which fall into each category during a selected period of time (or per thousand words uttered)." Lasswell was ahead of his time. In 1935- and at the age of 21-Laswell was developing methods that tracked the association between word utterances and physiological reactions (e.g. pulse rate, electrical conductivity of the skin, and blood pressure).

Harold Laswell, Pioneer of Quantitative Text Analysis

But Laswell was but the first in a long line of pioneers in the field of quantitative text analysis that hailed from many different fields— from sociology to computer sience and linguistics. The table below provides an overview of some of the major milestones in the field.

Timeline of Quantitative Text Analysis

Time	Activity
1934	Laswell Produces first Key-Word Count
1934	Vygotsky Produces first Quantitative Narrative Analaysis
1950	Gottschalk Uses Content Analysis to Track Freudian Themes

1950 Turin Applies AI to text

1952 Bereleson Publishes First Textbook on Content Analysis

1954 First Automatic Translation of Text (Georgetown Experiment)

1963 Msteller and Wallace analyze Federalist Papers

1965 Tomashevsky Further Formalizes Quantitative Narrative Analysis

1966 Stone and Bales use mainframe computer to measure psychometric properties of text at RAND

1980 Decline of Chomskyean Formalism/Field of Natural Language Processing is Born

1980 Machine Learning is Applied to Natural Language Processing

1981 Weintraub counts parts of speech

1985 Schrodt Introduces Auomated Event Coding

1986 Pennebaker develops Linguistic Inquiry Word Count

1989 Franzosi brings Quantitative Narrative Analysis to Social Science

1998 First Topic Models Developed

1998 Mohr conducts first Quantitative Analysis of Worldviews

1999 Bearman et al. apply Network Methods to Narratives

2001 Blei et al. develop Latent Dirichlet Allocation

2003 MALLET created

2005 Quinn et al use analyze political speeches using topic models

2010 King/Hopkins develop unsupervised classifier

2010 Tools for Text Workshop at Washington

What I find particularly remarkable about this (probably incomplete) timeline is that it covers scholars in at least seven different fields. Though social scientists made some of the earliest contributions to the field, computer scientists and linguists have exerted considerable influence in recent decades. The intellectual diversification of the field also coincided with the tremendous outgrowth of text-based data via the internet and other sources, as I describe in the following section.

The Text Data Explosion

The past decade has witnessed an increasingly voluminous amount of text-based data that is produced on the internet which describes human behavior and other objects of scholarly inquiry. As the figure below shows, recent decades have not only witnessed an increase in the amount of text based data, but also increased computing power which is increasingly necessary to analyze it. Together, these two shifts hold the potential to significantly expand the scope of research in many different fields.

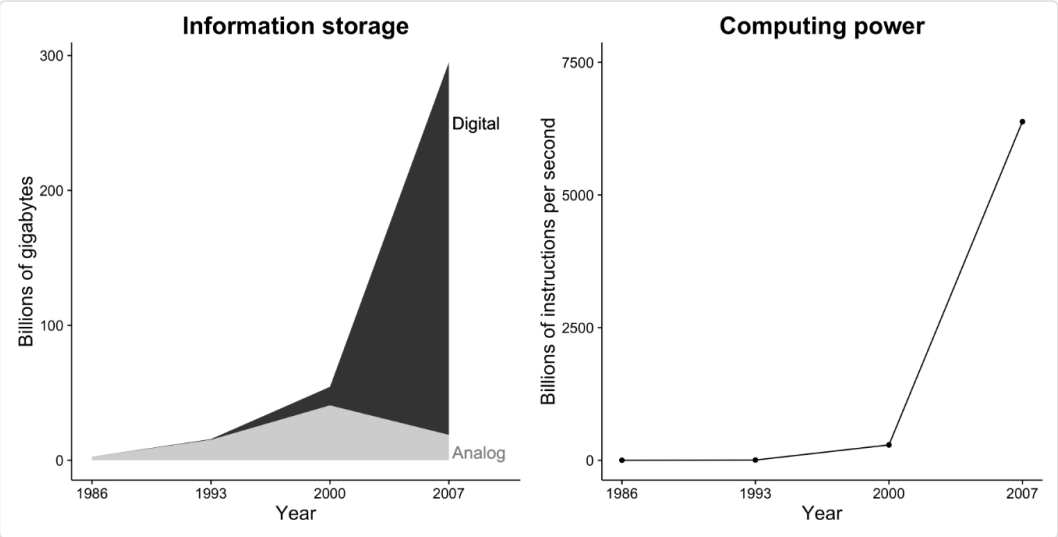


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital (Hilbert and López 2011). These changes create incredible opportunities for social researchers.

Strengths of Digital Trace Data

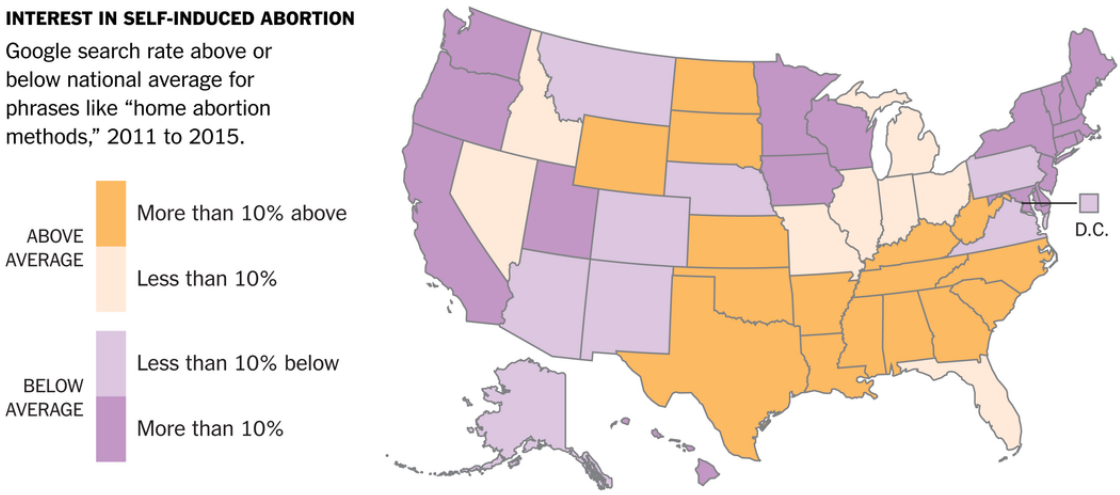
I will begin by discussing some of the positive aspects of digital trace data, and then move on to some of the challenges. In so doing, I draw upon Matt Salganik's Book *Bit by Bit* which I highly recommend not only for a more detailed discussion of digital trace data, but the nascent field of computational social science more broadly.

Always On

One of the most attractive features of digital trace data is that it is continuously collected, unlike surveys which usually only provide a brief snapshot of the social world. As the image below indicates, social media can occasionally provide a glimpse of major events such as protests, revolutions, or stock market surges, *as they unfold*.

Non-Reactive

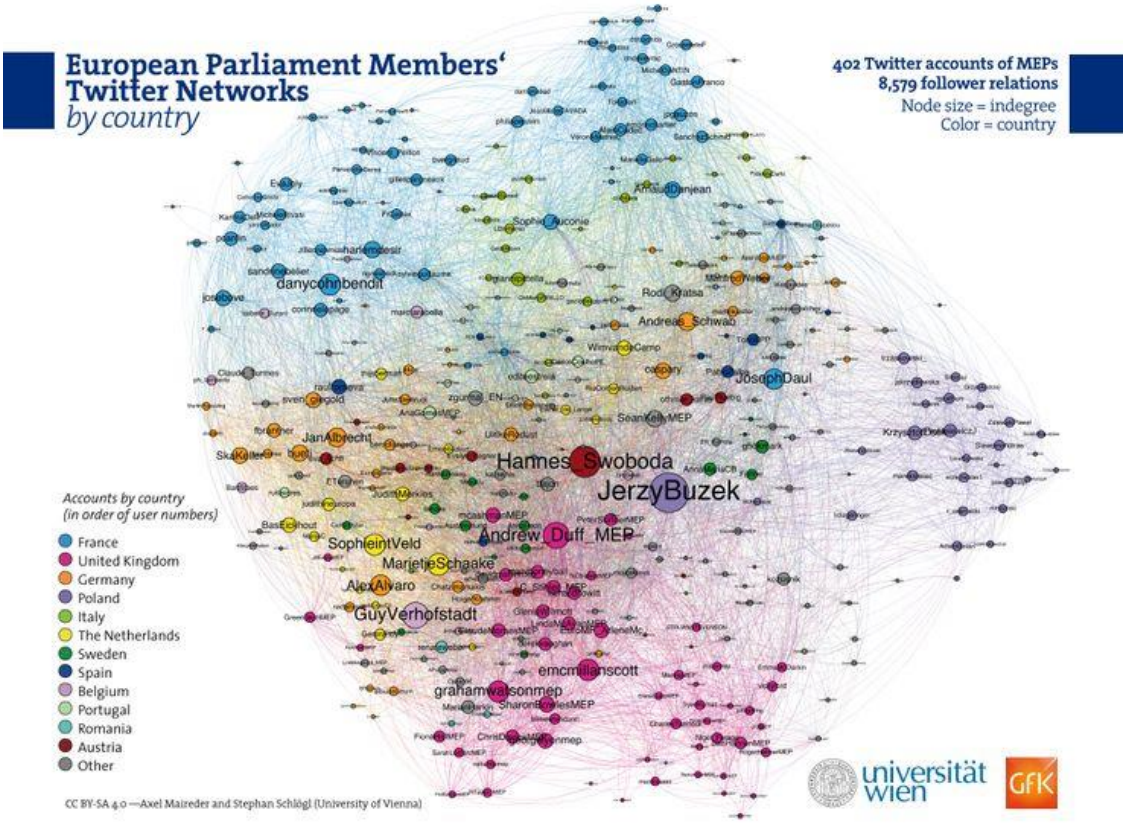
Another important advantage of digital trace data is that it is non-reactive, or not produced via interaction between researchers and those who they study. In some cases, this may lead to significant reductions in social desirability bias or other forms of interviewer effects. Consider, for example, the use of Google search data to study self-induced abortion (see figure below)



Captures Social relationships

Digital trace data are also somewhat unusual in that they often describe social relationships. Whereas conventional survey techniques usually only measure

characteristics of individual subjects, for example, digital trace data can often be used to measure social relationships such as the network of European politicians pictured below.



Weaknesses of Digital Trace Data

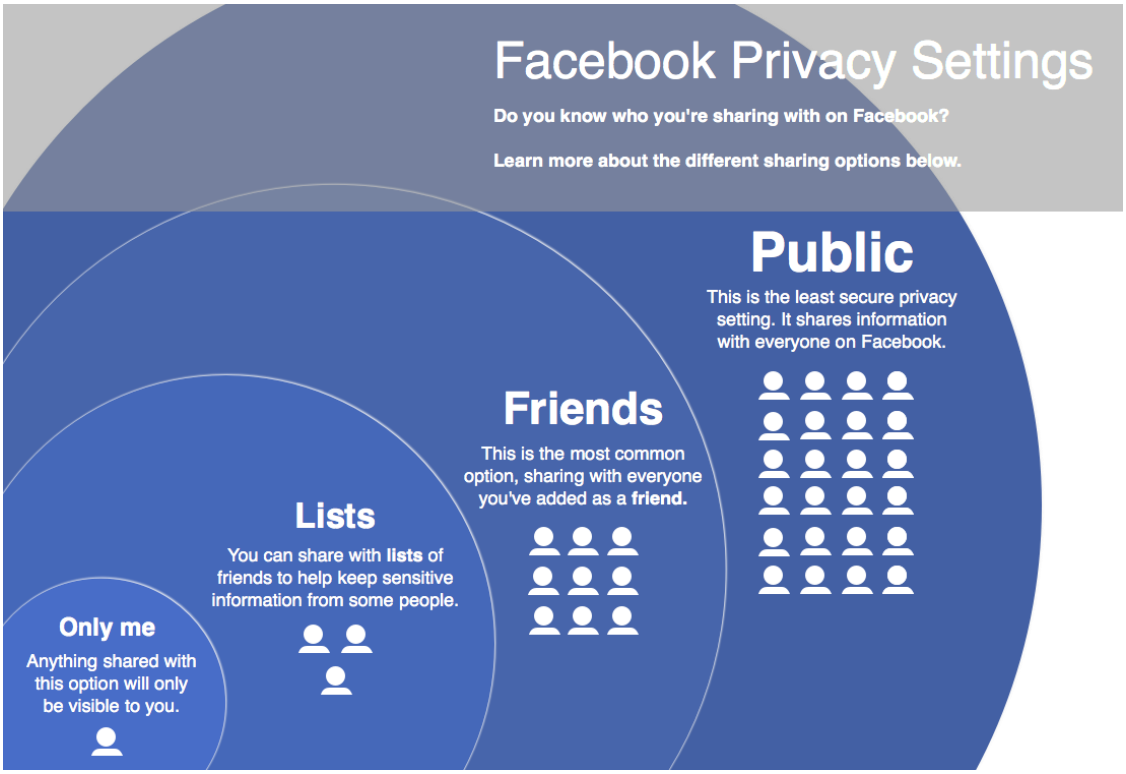
Despite the considerable advantages of digital trace data described above, they also create a range of challenges for empirical observation and causal inference.

Incomplete

Though much is often made of the size and scale of digital trace data that can be collected, newcomers to the field are often surprised about the amount of data that are often missing or incomplete. Consider, for example, a study of bullying behavior on social media— many of the most abusive posts that might be of interest to a researcher are often removed by Facebook before one might attempt to study them.

Inaccessible

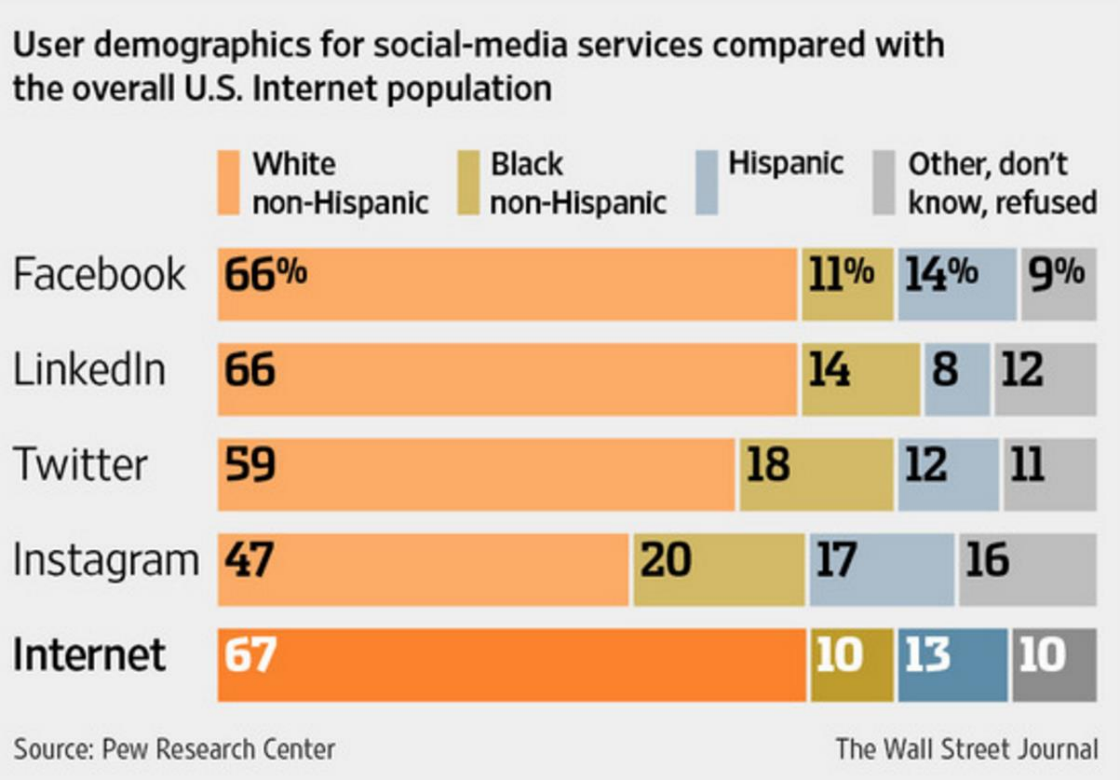
An even more formidable challenge is that data are often inaccessible. Though Twitter provides a massive amount of publicly available data, the vast majority of data generated on Facebook is private. Though some Facebook pages such as "fan pages" have default public settings, the vast majority of Facebook users set their default privacy settings in a manner that only allows people to access their data if they are affiliated with each other as "friends."



Non-Representative

Another core challenge facing those who wish to work with digital trace data is that a random sample of Facebook or Twitter users is not representative of the broader

population of the United States, or most other countries. The figure below presents some data from the Wall Street Journal on user demographics of several social media sites that demonstrates significant differences by platform according to race. On the other hand, usage of Facebook has become so widespread that some readers might be surprised to see how much more representative it has become of the U.S. public in recent years.

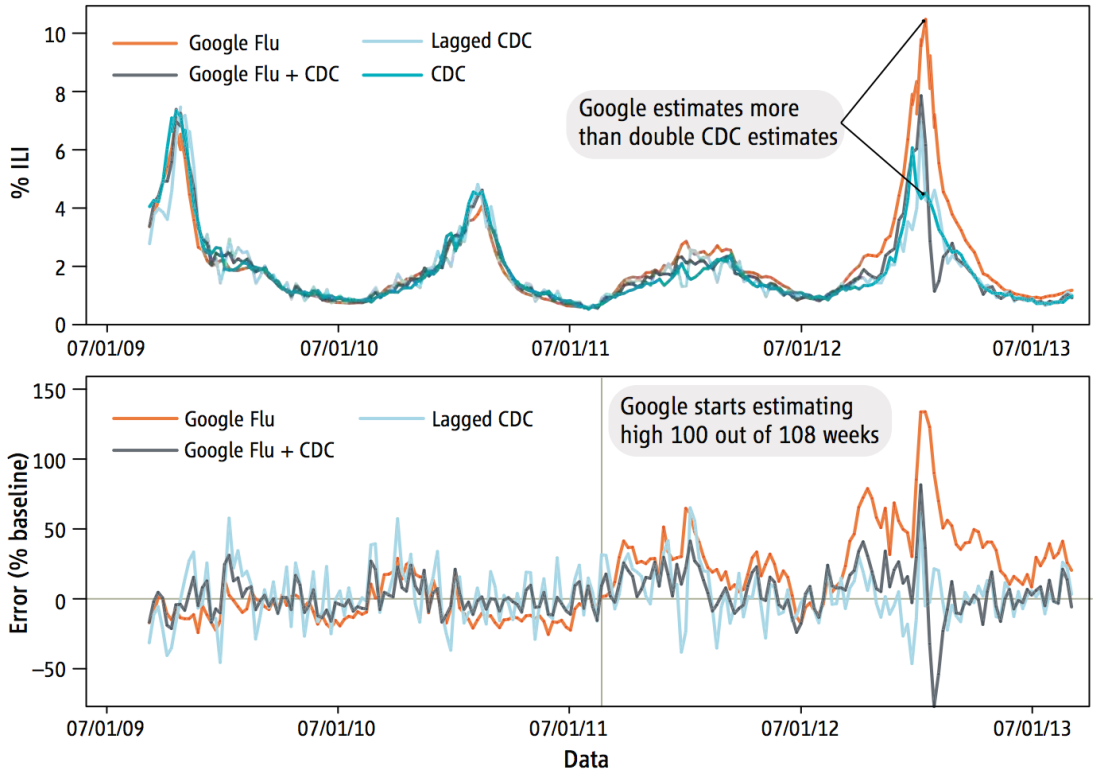


Drift

MySpace was once the largest social media site in the world, according to some analysts. It is now resides in the graveyard of internet history like so many other websites. This raises the risk of "drift" in digital trace data– platforms not only shift in their overall popularity (which of course has important implications for their representativeness), but also according to who uses them and why. Though Facebook was once the most popular platform for U.S. undergraduates, many have shifted to Instagram or Snapchat—possibly in reaction to the uptick of Facebook usage by their parents' generation :)

Algorithmic Counfounding

Sometimes, digital trace data that appear to describe human behavior actually reflect changes in the way humans interact with algorithms. One popular example of this is the "parable of Google Flu." Google Flu was once a popular tool that allowed users to estimate the prevalence of influenza using Google search data. The tool was so accurate that some suggested it should displace official surveys from the Centers for Disease Control (CDC). Yet in early 2013, Google estimates were far higher than those from the CDC. Researchers later discovered that estimates of influenza had been inflated by google advertising links about the flu that people were clicking on that had appeared in their web browsers after they searched for information about symptoms of the common cold. This is sometimes refered to as "blue-team" dynamics.

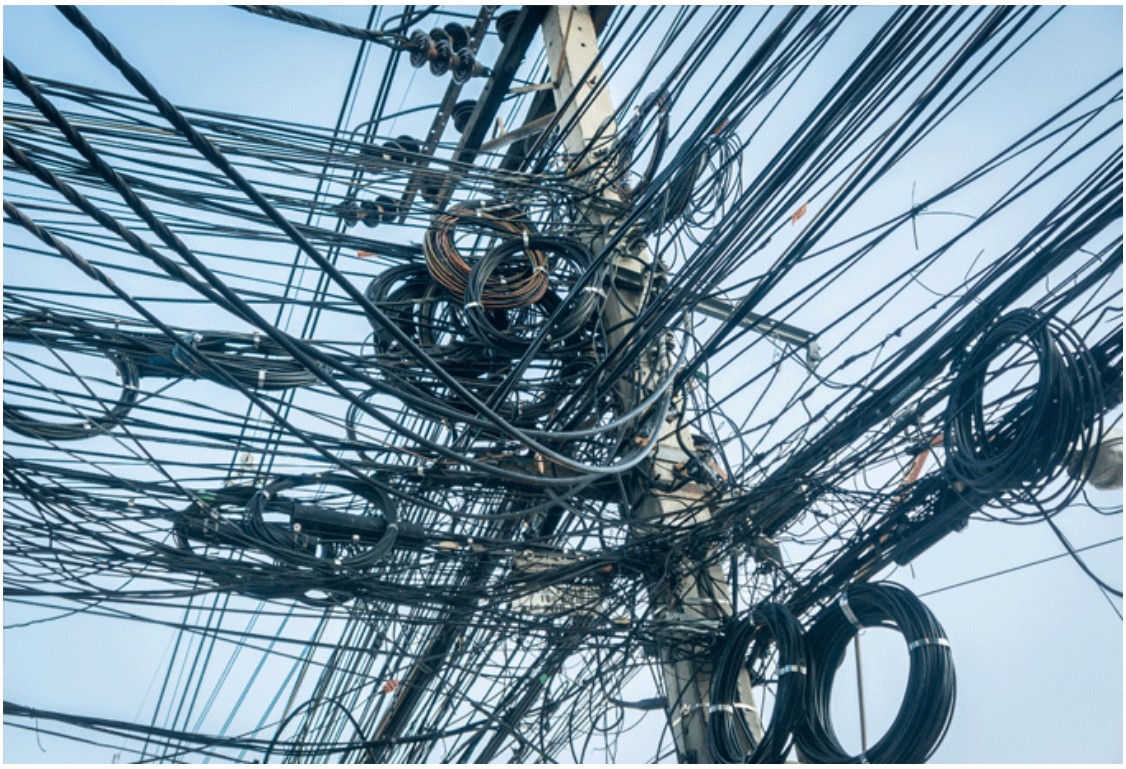


GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (**Bottom**) Error [as a percentage {[Non-CDC estmate)−(CDC estimate)]/(CDC) estimate)}. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

Unstructured

Digital Trace data are also often very messy. Newcomers to the field often assume that because data are generated in digital format that they are well structured, easily

searchable, and quickly transposable across different formats. As we shall see in future tutorials, this is most often not true— a recent New York Times article indicated that data scientists spend upwards of 80% of their time cleaning data!



Sensitive

Digital trace data are also often very sensitive. Recent events involving Facebook and the Cambridge Analytica Political Consulting Firm underscore the dangers of unfettered access to large troves of digital trace data, but there were many more—arguably more invasive—data breaches long before this recent event. One such incident, pictured below, involved European researchers who mined data from the internet dating site OK Cupid and then publicly released their data online.

OKCUPID

Researchers Caused an Uproar By Publishing Data From 70,000 OkCupid Users

Robert Hackett
May 18, 2016

Positivity-Bias

Finally, digital trace data often have performative dimensions. Many people do not report negative information about themselves online precisely because they know that their friends, colleagues— or other people they do not know— may be watching them. This creates another common form of bias in social media research.

Exploring Text-Based Datasets

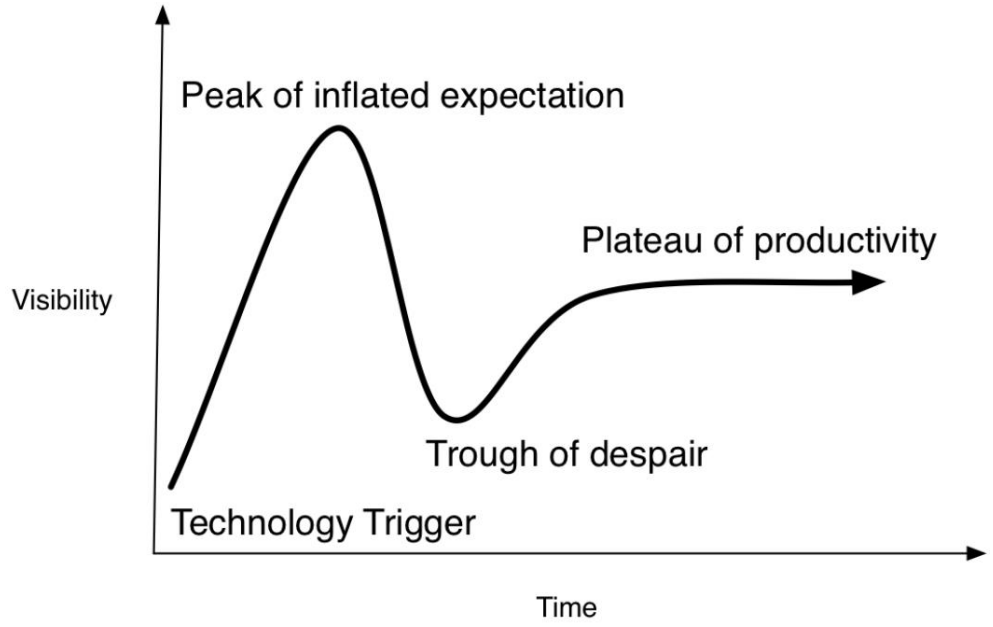
Before we proceed to analyze the data we collected in previous tutorials from Facebook and Twitter, it is worth noting that there are numerous high-quality text datasets that are publicly available on the internet. These include very large corpora from the New York Times, Google, Wikipedia, Reuters, and many other sites). I curate a crowd-sourced list of promising datasets here.

Now YOU Try It

1) Pair up with your neighbor.
2) Pick a dataset from the list in the previous section—or another one that you are hoping to analyze after this course;
3) Identify at least three strengths and weaknesses of the dataset drawing upon this introduction, or other sources.

The Future of Digital Trace Data

After reviewing so many of the negatives of digital trace data, you may be questioning whether its strengths outweigh its weaknesses. I am overall optimistic about the future of digital trace data research because it is in its infancy. As Salganik (2016) writes, the field has recently experienced a Gartner hype cycle (see figure below). In my opinion, reaching the "plateau of productivity" will most likely require hybrid approaches that combine digital trace data analysis alongside more conventional modes of research such as survey analysis. I've written at length about this issue elsewhere, and specifically the potential of using app technology to integrate digital trace data collection with survey research.



https://commons.wikimedia.org/wiki/File:Gartner_Hype_Cycle.svg

Screen-Scraping in R

Table of Contents

Chris Bail
Duke University
www.chrisbail.net

What is Screen-Scraping?

Screenscraping refers to the process of automatically extracting data from web pages, and often a long list of websites that cannot be mined by hand. As the figure below illustrates, a typical screenscraping program a) loads the name of a web-page to be scraped from a list of webpages; b) downloads the website in a format such as HTML or XML; c) finds some piece of information desired by the author of the code; and d) places that information in a convenient format such as a "data frame" (which is R speak for a dataset). Screenscraping can also be used to download other types of content as well, however, such as audio-visual content.