Day 3: Artificial intelligence for policy-making

The promise and perils of big data

Simon Munzert
Hertie School

Table of contents

1. What is big data?
2. The big data paradox
3. Garbage in, garbage out
4. Opportunities of big data for the public good

What is big data?

What is big data?

Hertie School

The five V's of big data

VOLUME

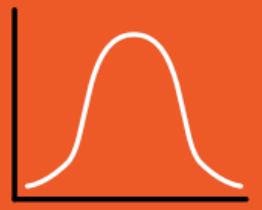
The scale of data.

VARIETY

Data comes in different forms. Structured data are easily searchable like spreadsheets. Unstructured data include disorganized information like tweets and video.

VELOCITY

The speed of data processing. These days a great deal of data are available in real time, such as social media posts.



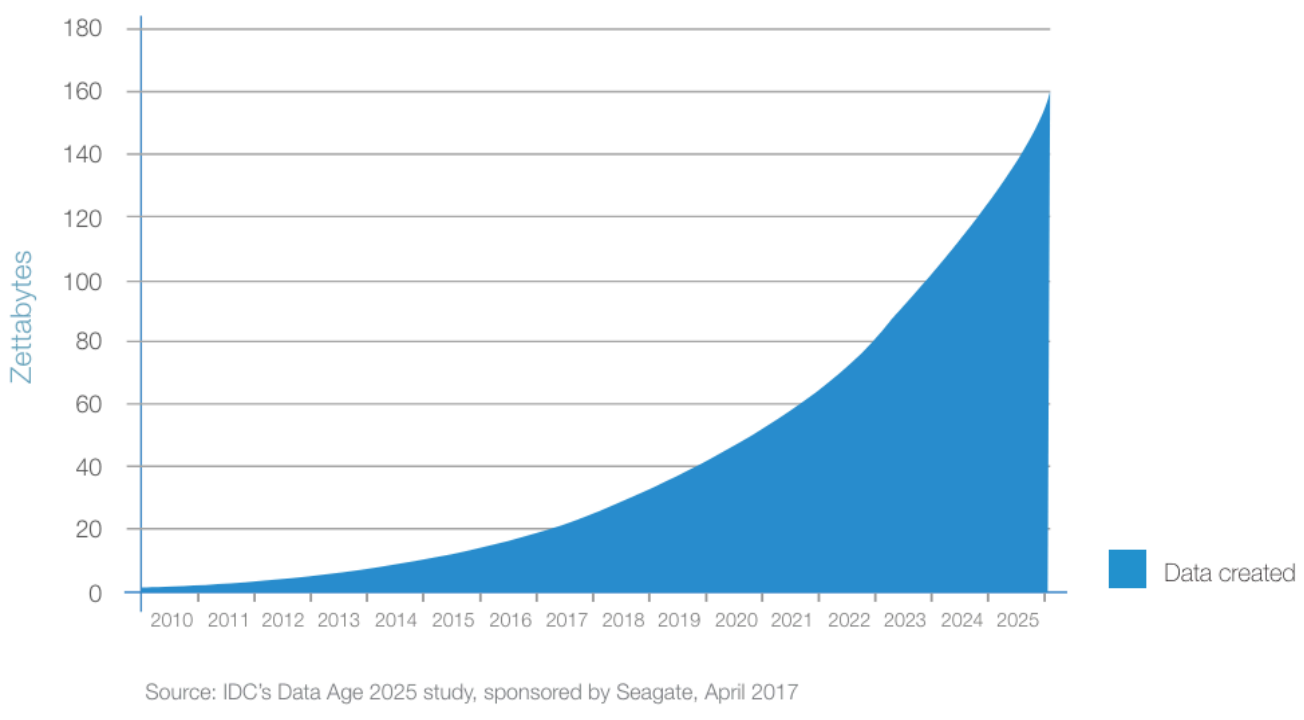
VARIABILITY

How spread out the data are.

VERACITY

The accuracy of data provides confidence during research.

Massive amounts of data around us[1]



`Source` Reinsel et al., 2017, "Data Age 2025"

`Source` eDiscoveryToday, LTMG

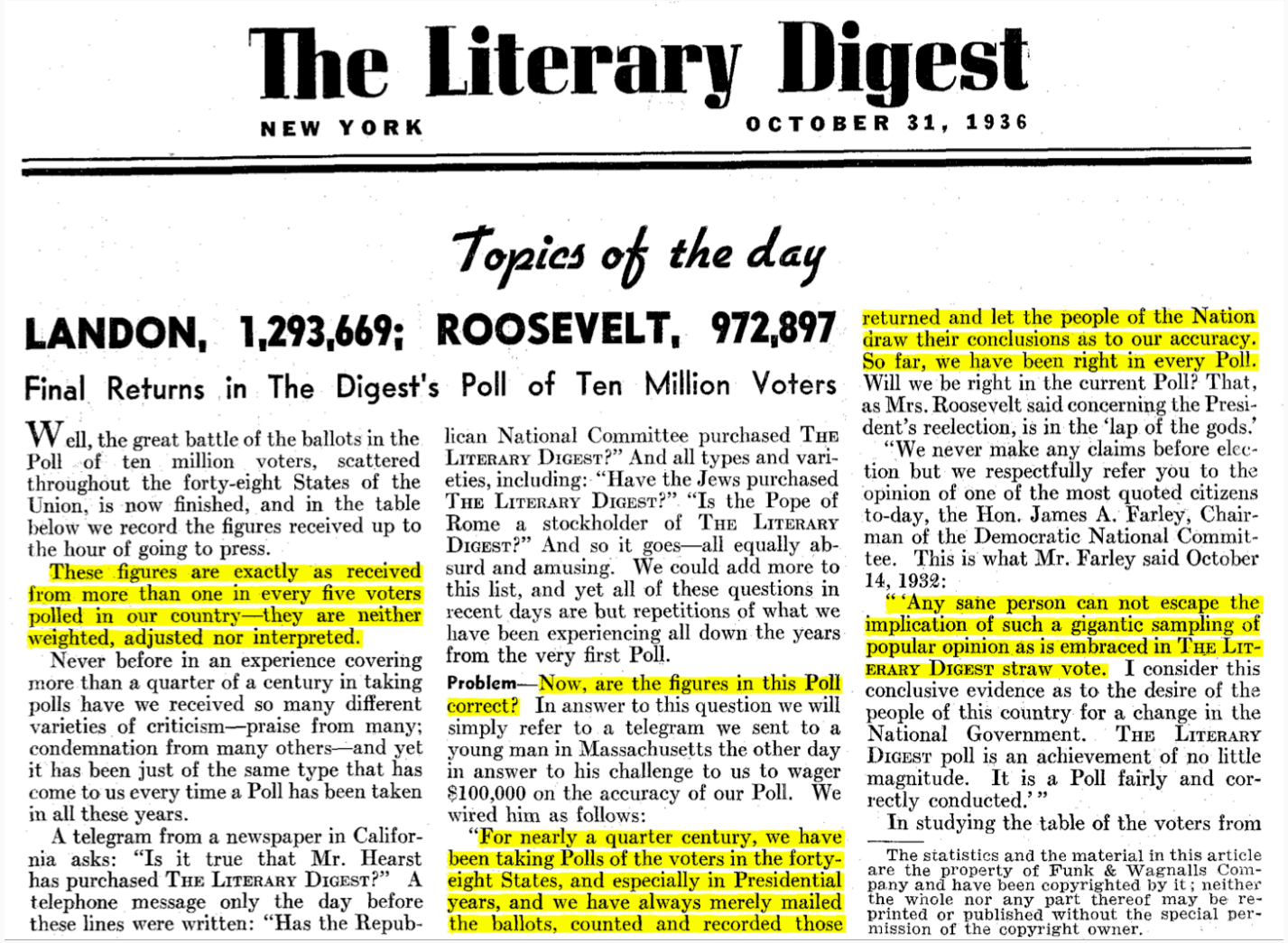
[1]Don't take any of those numbers too seriously. Those are very difficult to measure and numbers you find online are at times differing by oders of magnitude. Also, they are a moving target.

The age of big data - big trends

1. Proliferation of **human-generated data** at scale, in particular in the digital sphere
2. Use of **new data types**: Text, video, digital traces
3. **Computational and storage costs** fallen dramatically
4. Mainstreaming of **machine learning** and **AI** technologies
5. (Limited) **democratization of access** to large data resources
6. **Embracement of computational methods** development across many disciplines
7. **Shift in research avant-garde** from academia to industry

The big data paradox

An early big data disaster



The Literary Digest
NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: "Is it true that Mr. Hearst has purchased THE LITERARY DIGEST?" A telephone message only the day before these lines were written: "Has the Republican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased THE LITERARY DIGEST?" "Is the Pope of Rome a stockholder of THE LITERARY DIGEST?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

Problem—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager $100,000 on the accuracy of our Poll. We wired him as follows:

"For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

"'Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in THE LITERARY DIGEST straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. THE LITERARY DIGEST poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.'"

In studying the table of the voters from

Literary Digest's 1936 election poll

Background

- The weekly magazine *Literary Digest* had correctly predicted the outcomes of all presidential elections between 1920 and 1932 using straw polls.
- Their 1936 poll of 10m voters indicated that Republican candidate Alfred Landon would be the overwhelming winner.

Aftermath

- The poll was a disaster: Landon lost in a landslide to Franklin D. Roosevelt, who carried 46 out of 48 states and won 60.8% of the popular vote.
- The outcome was correctly predicted by George Gallup with sample of 50k people.
- The magazine went bankrupt in 1938.

Literary Digest's 1936 election poll

Hertie School

The Literary Digest
NEW YORK NOVEMBER 14, 1936

Topics of the day

WHAT WENT WRONG WITH THE POLLS?

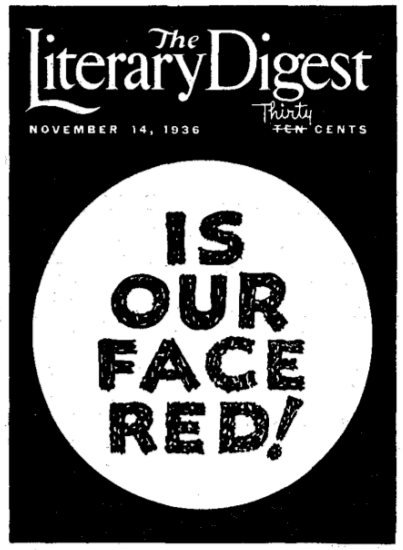
None of Straw Votes Got Exactly the Right Answer—Why?

In 1920, 1924, 1928 and 1932, THE LITERARY DIGEST Polls were right. Not only right in the sense that they showed the winner; they forecast the *actual popular vote* with such a small percentage of error (less than 1 per cent. in 1932) that newspapers and individuals everywhere heaped such phrases as "uncannily accurate" and "amazingly right" upon us.

Four years ago, when the Poll was running his way, our very good friend Jim Farley was saying that "no sane person could escape the implication" of a sampling "so fairly and correctly conducted."

Well, this year we used precisely the same method that had scored four bull's-eyes in four previous tries. And we were far from correct. Why? We ask that question in all sincerity, because *we want to know*.

"Reasons"—Oh, we've been flooded with "reasons." Hosts of people who feel they have learned more about polling in a few months than we have learned in more than a score of years have told us just where we were off. Hundreds of astute "second-guessers" have assured us by tele-



The following telegram was received by The Literary Digest: "With full and sympathetic appreciation of the rather tough spot you now

out of the 30,811 who voted returned ballots to us showing a division of 53.32 per cent. to 44.67 per cent. in favor of Mr. Landon. What was the actual result? It was 56.93 per cent. for Mr. Roosevelt, 41.17 per cent. for the Kansan.

In Chicago, the 100,929 voters who returned ballots to us showed a division of 48.63 per cent. to 47.56 per cent. in favor of Mr. Landon. The 1,672,175 who voted in the actual election gave the President 65.24 per cent., to 32.26 per cent. for the Republican candidate.

What happened? Why did only one in five voters in Chicago to whom THE DIGEST sent ballots take the trouble to reply? And why was there a preponderance of Republicans in the one-fifth that did reply? Your guess is as good as ours. We'll go into it a little more later. The important thing in all the above is that all this conjecture about our "not reaching certain strata" simply will not hold water.

Hoover Voters—Now for another "explanation" dinned into our ears: "You got too many Hoover voters in your sample."

Well, the fact is that we've *always* got too big a sampling of Republican voters. That was true in 1920, in 1924, in 1928, and even in 1932, when we *over*estimated the Roosevelt popular vote by three-quarters of 1 per cent.

Literary Digest's 1936 election poll

Anatomy of a debacle

1. **Sampling frame**: (1) own readers, (2) registered automobile owners, (3) registered telephone users
2. **Data collection**: everyone was mailed a mock ballot and asked to return marked ballot
3. **Response rate**: 2.4m out of 10m

Selection bias as a consequence of **coverage** and **nonresponse bias**: overrepresentation of wealthier individuals with a preference for Landon

`Source` Peverill Squire, 1988, Public Opinion Quarterly

Table 1. 1936 Presidential Vote by Car and Telephone Ownership (in Percent)

Presidential Vote	Car & Phone	Car, No Phone	Phone, No Car	Neither
Roosevelt	55	68	69	79
Landon	45	30	30	19
Other	1	2	0	2
Total *N*	946	447	236	657

SOURCE: American Institute of Public Opinion, 28 May 1937.

Table 2. Presidential Vote by Receiving *Literary Digest* Straw Vote Ballot or Not (in Percent)

Presidential Vote	Received Poll	Not Receive Poll	Do Not Know
Roosevelt	55	71	73
Landon	44	27	25
Other	1	1	3
Total *N*	780	1339	149

SOURCE: American Institute of Public Opinion, 28 May 1937.

Table 3. Presidential Vote by Returning or Not Returning Straw Vote Ballot (in Percent)

Presidential Vote	Did Return	Did Not Return	Do Not Know
Roosevelt	48	69	56
Landon	51	30	40
Other	1	1	4
Total *N*	493	288	48

SOURCE: American Institute of Public Opinion, 28 May 1937.

Take surveys with a pinch of salt

A modern big data polling disaster

Hertie School

Unrepresentative big surveys significantly overestimated US vaccine uptake

https://doi.org/10.1038/s41586-021-04198-4

Received: 18 June 2021

Accepted: 29 October 2021

Published online: 8 December 2021

Check for updates

Valerie C. Bradley[1,6], Shiro Kuriwaki[2,6], Michael Isakov[3], Dino Sejdinovic[1], Xiao-Li Meng[4] & Seth Flaxman[5]✉

Surveys are a crucial tool for understanding public opinion and behaviour, and their accuracy depends on maintaining statistical representativeness of their target populations by minimizing biases from all sources. Increasing data size shrinks confidence intervals but magnifies the effect of survey bias: an instance of the Big Data Paradox[1]. Here we demonstrate this paradox in estimates of first-dose COVID-19 vaccine uptake in US adults from 9 January to 19 May 2021 from two large surveys: Delphi–Facebook[2,3] (about 250,000 responses per week) and Census Household Pulse[4] (about 75,000 every two weeks). In May 2021, Delphi–Facebook overestimated uptake by 17 percentage points (14–20 percentage points with 5% benchmark imprecision) and Census Household Pulse by 14 (11–17 percentage points with 5% benchmark imprecision), compared to a retroactively updated benchmark the Centers for Disease Control and Prevention published on 26 May 2021. Moreover, their large sample sizes led to miniscule margins of error on the incorrect estimates. By contrast, an Axios–Ipsos online panel[5] with about 1,000 responses per week following survey research best practices[6] provided reliable estimates and uncertainty quantification. We decompose observed error using a recent analytic framework[1] to explain the inaccuracy in the three surveys. We then analyse the implications for vaccine hesitancy and willingness. We show how a survey of 250,000 respondents can produce an estimate of the population mean that is no more accurate than an estimate from a simple random sample of size 10. Our central message is that data quality matters more than data quantity, and that compensating the former with the latter is a mathematically provable losing proposition.

Table 1 | Comparison of survey designs

	Axios-Ipsos	Census Household Pulse	Delphi-Facebook
Recruitment mode	Address-based mail sample to Ipsos KnowledgePanel	SMS and email	Facebook Newsfeed
Interview mode	Online	Online	Online
Average size	1,000/wave	75,000/wave	250,000/week
Sampling frame	Ipsos KnowledgePanel; internet/tablets provided to ~5% of panelists who lack home internet	Census Bureau's Master Address File (individuals for whom email / phone contact information is available)	Facebook active users
Vaccine uptake question	"Do you personally know anyone who has already received the COVID-19 vaccine?"	"Have you received a COVID-19 vaccine?"	"Have you had a COVID-19 vaccination?"
Vaccine uptake definition	"Yes, I have received the vaccine"	"Yes"	"Yes"
Other vaccine uptake response options	"Yes, a member of my immediate family", "Yes, someone else", "No"	"No"	"No", "I don't know"
Weighting variables	Gender by age, race, education, Census region, metropolitan status, household income, partisanship.	Education by age by sex by state, race/ethnicity by age by sex by state, household size	Stage 1: age, gender "other attributes which we have found in the past to correlate with survey outcomes" to FAUB; Stage 2: state by age by gender

Comparison of key design choices across the Axios–Ipsos, Census Household Pulse and Delphi–Facebook studies. All surveys target the US adult population. See Extended Data Table 1 for additional comparisons and Methods for additional implementation details.

A modern big data polling disaster

Hertie School

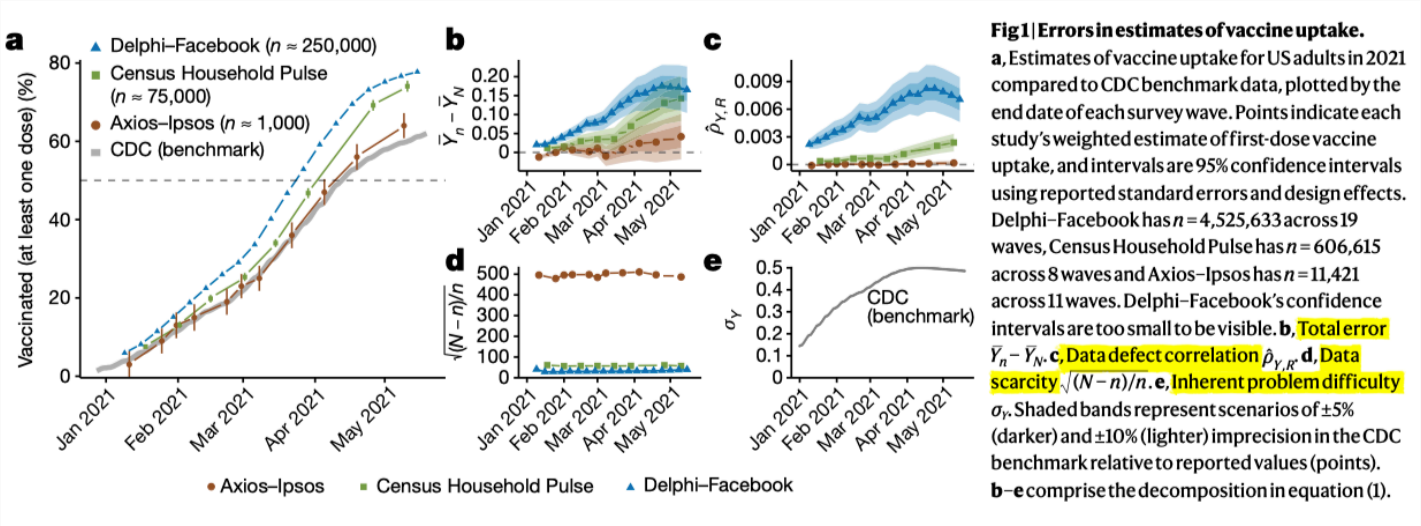


Fig 1 | Errors in estimates of vaccine uptake. **a**, Estimates of vaccine uptake for US adults in 2021 compared to CDC benchmark data, plotted by the end date of each survey wave. Points indicate each study's weighted estimate of first-dose vaccine uptake, and intervals are 95% confidence intervals using reported standard errors and design effects. Delphi–Facebook has $n = 4{,}525{,}633$ across 19 waves, Census Household Pulse has $n = 606{,}615$ across 8 waves and Axios–Ipsos has $n = 11{,}421$ across 11 waves. Delphi–Facebook's confidence intervals are too small to be visible. **b**, Total error $\bar{Y}_n - \bar{Y}_N$. **c**, Data defect correlation $\hat{\rho}_{Y,R}$. **d**, Data scarcity $\sqrt{(N-n)/n}$. **e**, Inherent problem difficulty σ_Y. Shaded bands represent scenarios of ±5% (darker) and ±10% (lighter) imprecision in the CDC benchmark relative to reported values (points). **b**–**e** comprise the decomposition in equation (1).

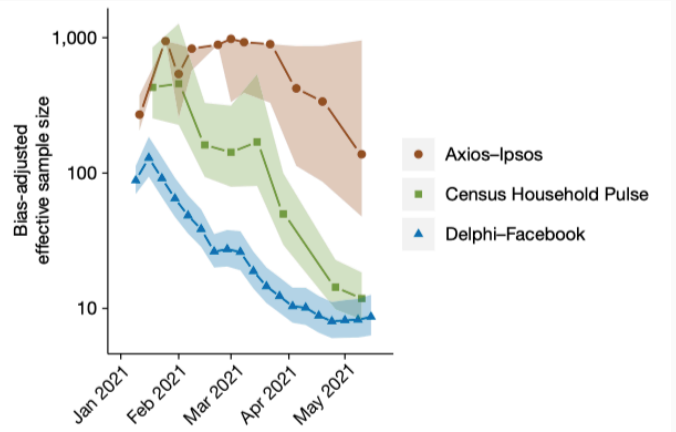


Fig 2 | Bias-adjusted effective sample size. An estimate's bias-adjusted effective sample size (different from the classic Kish effective sample size) is the size of a simple random sample that would have the same MSE as the observed estimate. Effective sample sizes are shown here on the $\log_{10}$ scale. The original sample size was $n = 4{,}525{,}633$ across 19 waves for Delphi–Facebook, $n = 606{,}615$ across 8 waves for Census Household Pulse and $n = 11{,}421$ across 11 waves for Axios–Ipsos. Shaded bands represent scenarios of ±5% benchmark imprecision in the CDC benchmark.

The big data paradox

> "[W]hen biased samples are large, they are doubly misleading: they produce confidence intervals with incorrect centres and substantially underestimated widths. This is the **Big Data Paradox**: the bigger the data, the surer we fool ourselves when we fail to account for bias in data collection."

Bradley et al. (2021), *Nature*

> "[T]he 'bigness' of such Big Data (for population inferences) should be measured by the relative size f = n/N of the sample to the population, not by the absolute size n of the sample."

Xiao-Li Meng (2018), *The Annals of Applied Statistics*

Finite population correction to the rescue?

- Intuition tells us that when the sample size becomes big relative to the population size, we should be less likely to err.
- That's true, but the gain is relatively slow.
- For instance, the finite population correction factor for the standard error of a quantity of interest given by $\sqrt{\frac{N-n}{N-1}}$.

Example

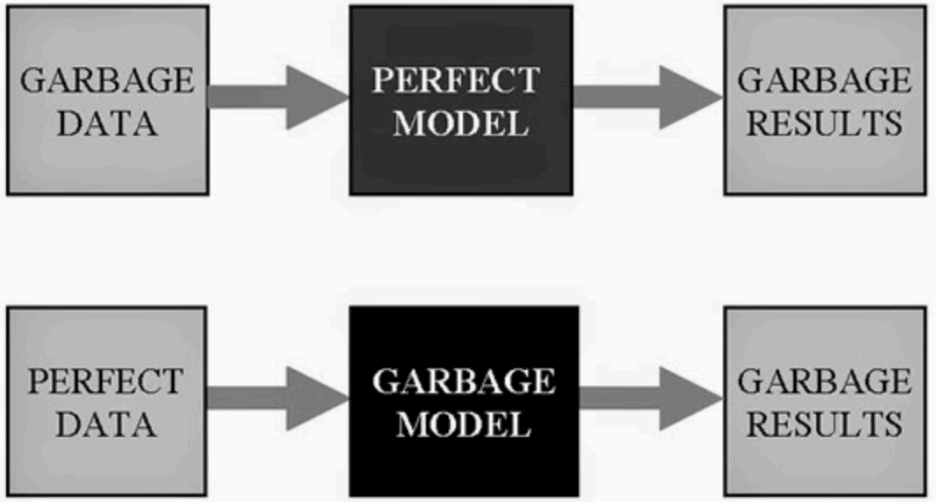
- We run a survey of 100k people in a population of 3.7m.
- For $N = 3.7m$ and $n = 100k$, this is $\sqrt{0.973}$.

Garbage in, garbage out

Garbage in, garbage out

THE GIGO principle

- The quality of information coming out of a model (e.g., predictions) cannot be better than the quality of information that went in.
- The principle is particularly relevant in the context of big data, where data quality is often poor.
- This is particularly relevant in the context of machine learning, where models can be very complex and opaque.



Google Flu Trends - the big promise

nature Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities[2]. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza[3,4]. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

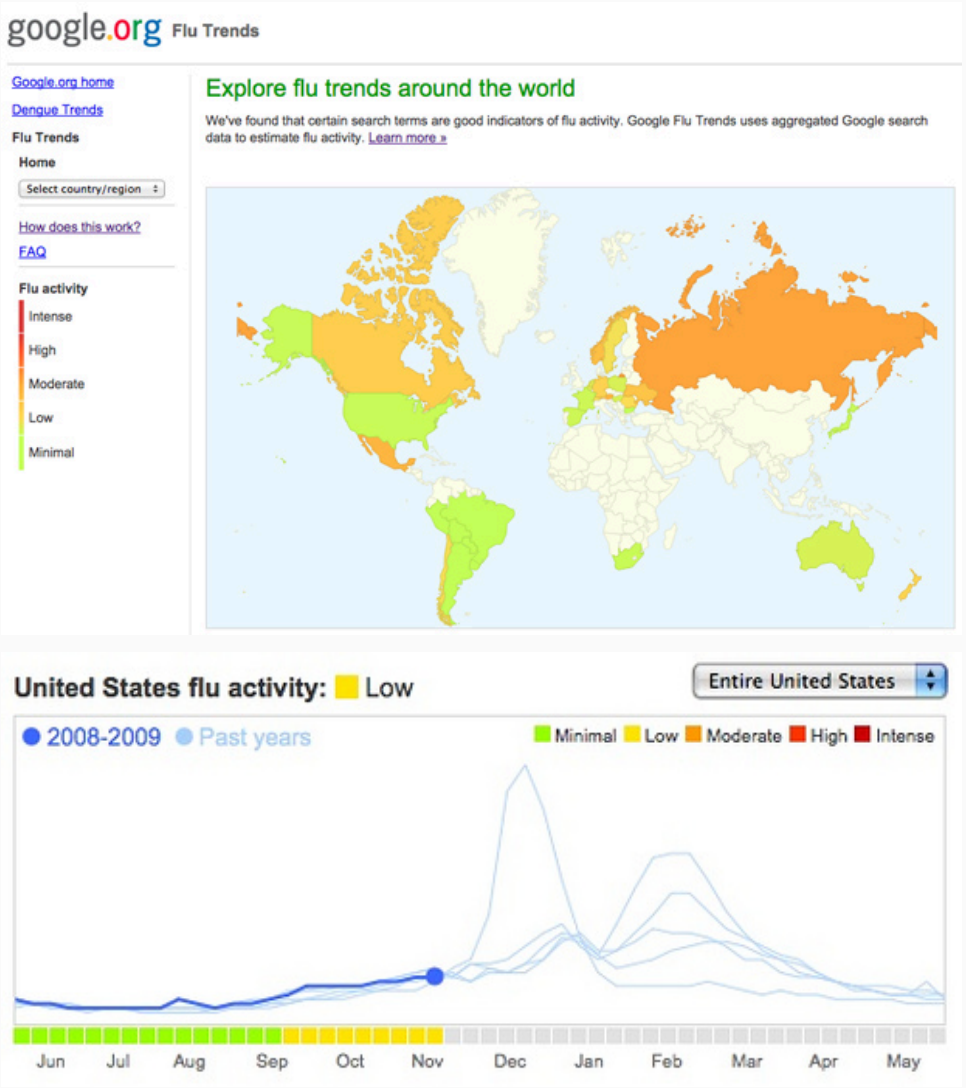
We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query: $\text{logit}(I(t)) = \alpha\text{logit}(Q(t)) + \varepsilon$, where $I(t)$ is the percentage of ILI physician visits, $Q(t)$ is the ILI-related query fraction at time t, α is the multiplicative coefficient, and ε is the error term. $\text{logit}(p)$ is simply $\ln(p/(1-p))$.

Publicly available historical data from the CDC's US Influenza

`Source` Ginsberg et al., 2009, Nature

Google Flu Trends - the big promise

Hertie School



Google Flu Trends - the failure

Hertie School

POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

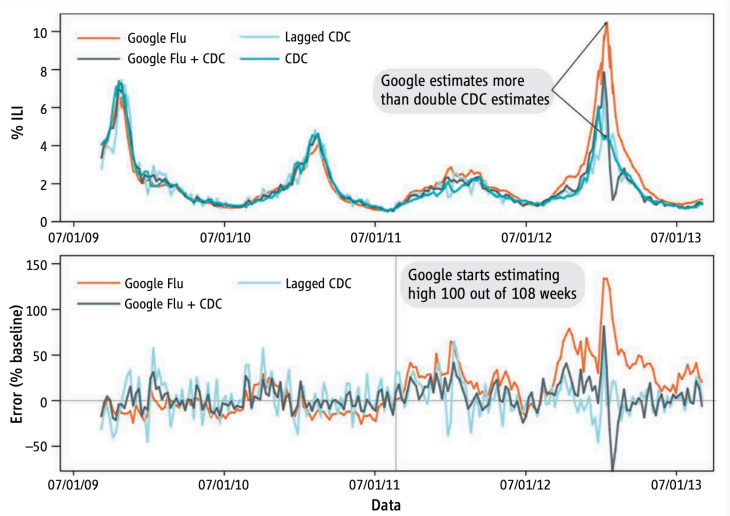
David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can

run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.



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.

`Source` Lazer et al., 2014, Science

Automated Inference on Criminality using Face Images

Xiaolin Wu
Shanghai Jiao Tong University
xwu510@gmail.com

Xi Zhang
Shanghai Jiao Tong University
zhangxi_19930818@sjtu.edu.cn

Abstract

We study, for the first time, automated inference on criminality based solely on still face images. Via supervised machine learning, we build four classifiers (logistic regression, KNN, SVM, CNN) using facial images of 1856 real persons controlled for race, gender, age and facial expressions, nearly half of whom were convicted criminals, for discriminating between criminals and non-criminals. All four classifiers perform consistently well and produce evidence for the validity of automated face-induced inference on criminality, despite the historical controversy surrounding the topic. Also, we find some discriminating structural features for predicting criminality, such as lip curvature, eye inner corner distance, and the so-called nose-mouth angle. Above all, the most important discovery of this research is that criminal and non-criminal face images populate two quite distinctive manifolds. The variation among criminal faces is significantly greater than that of the non-criminal faces. The two manifolds consisting of criminal and non-criminal faces appear to be concentric, with the non-criminal manifold lying in the kernel with a smaller span, exhibiting a law of normality for faces of non-criminals. In other words, the faces of general law-biding public have a greater degree of resemblance compared with the faces of criminals, or criminals have a higher degree of dissimilarity in facial appearance than normal people.

people share the belief that the face alone suffices to reveal innate traits of a person. Aristotle in his famous work Prior Analytics asserted, "It is possible to infer character from features, if it is granted that the body and the soul are changed together by the natural affections". Psychologists have known, for as long as a millennium, the human tendency of inferring innate traits and social attributes (e.g., the trustworthiness, dominance) of a person from his/her facial appearance, and a robust consensus of individuals' inferences . These are the facts found through numerous studies [2, 32, 4, 5, 9, 20, 21, 27, 25].

Independent of the validity of pedestrian belief in the (pseudo)science of physiognomy, a tantalizing question naturally arises: what facial features influence average Joe's impulsive and yet consensual judgments on social attributes of a non-acquaintance member of their own specie? Attempting to answer the question, Todorov and Oosterhof proposed a data-driven statistical modeling method to find visual determinants of social attributes by asking human subjects to score four percepts: dominance, attractiveness, trustworthiness, and extroversion, based on first impression of static face images [26]. This method can synthesize a representative (average) face image for a set of input face images scored closely on any of the four aforementioned social percepts. The ranking of these synthesized face images by subjective scores (e.g., from least to most trustworthy looking) apparently agrees with the intuition of most people.

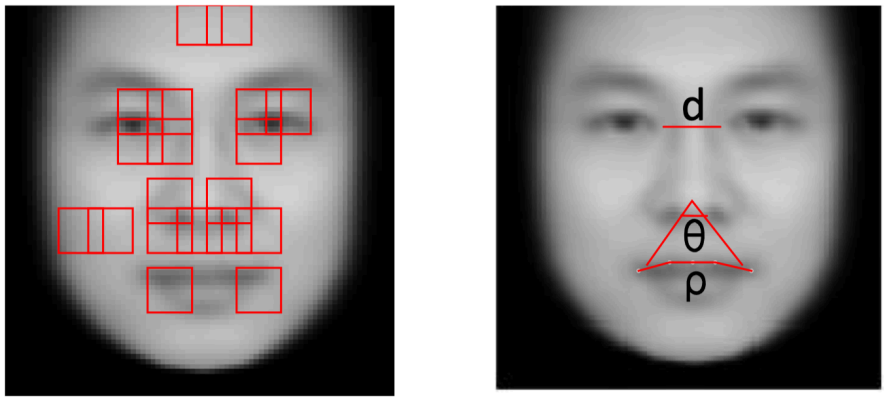
Detecting criminality with face images

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Detecting criminality with face images



(a) (b)

Figure 4. (a) FGM results; (b) Three discriminative features ρ, d and θ.

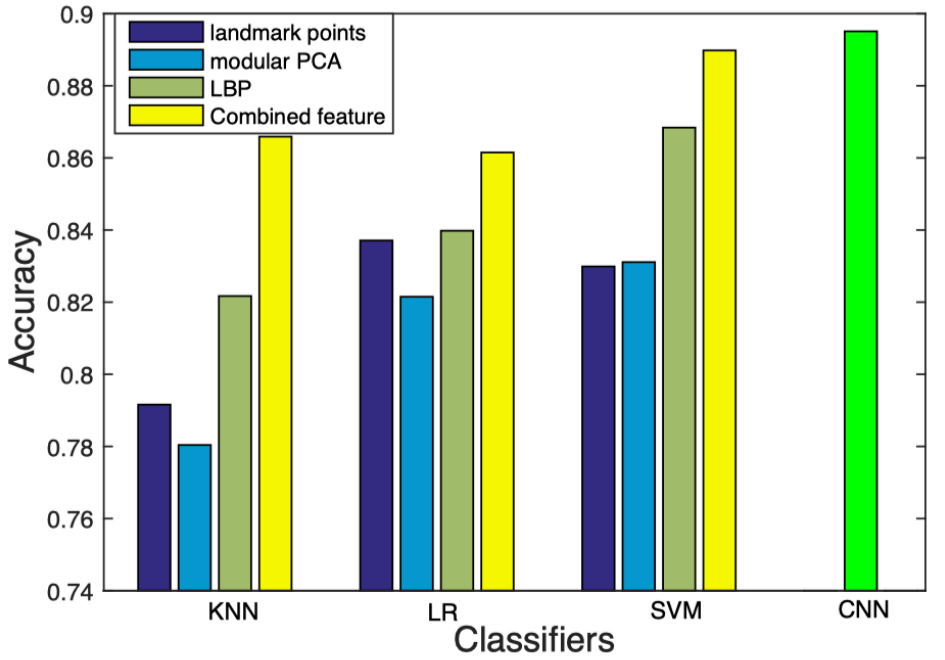
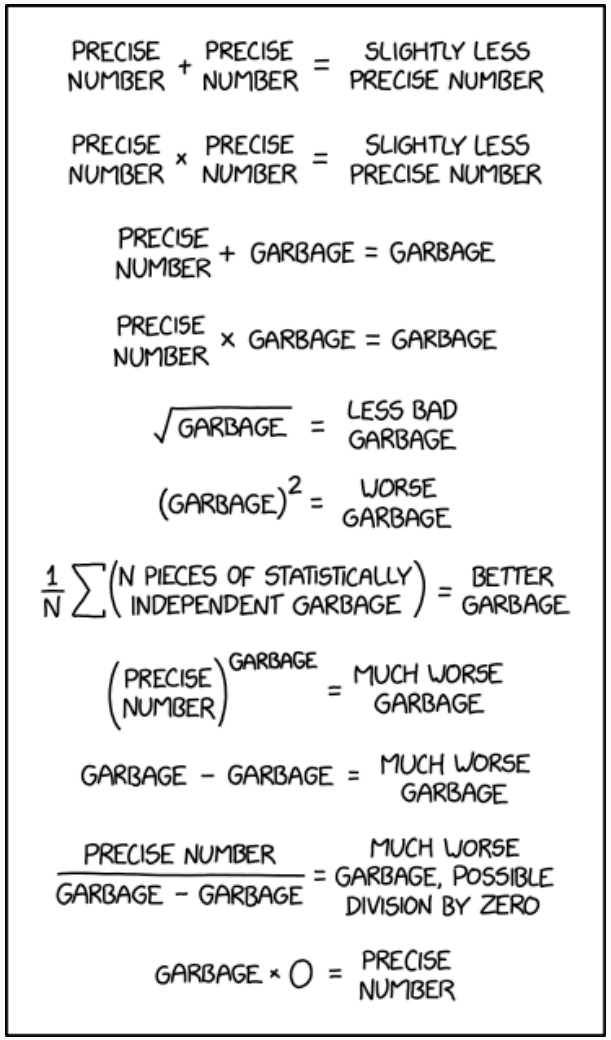


Figure 2. Accuracy of all four classifiers in all thirteen cases.

Garbage in, garbage out: lessons learned

1. **Measurement** and **selection issues** are still key in Big Data analytics
2. Don't trust **measures** that haven't been **properly validated**
3. Look out for what goes into a model (the **input**: cases, variables/features)
4. Look for proper **out-of-sample validation** of models
5. Beware of **uncritical use of online/social media** as a data source



`Source` XKCD, 2295, (explained here)

Opportunities of big data for the public good

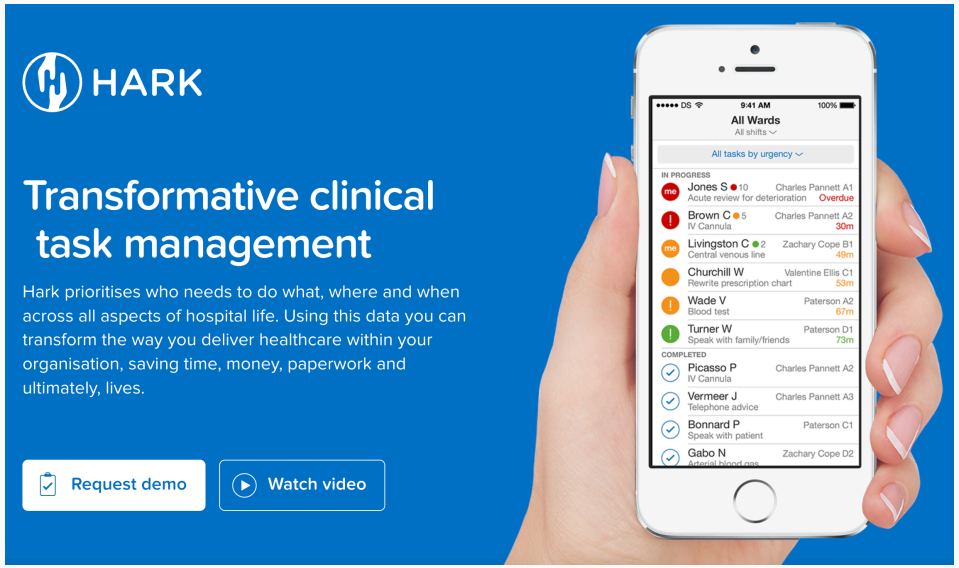
The age of big data – the age of dystopia?

Palantir

Cambridge Analytica

Boost public services with AI for individualized health care

Hertie School



JOURNAL OF MEDICAL INTERNET RESEARCH Patel et al

Original Paper

Interprofessional Communication of Clinicians Using a Mobile Phone App: A Randomized Crossover Trial Using Simulated Patients

Bhavesh Patel[1], BSc, MB BS, MRCS; Maximilian Johnston[2], MB BCh, MRCS, PhD; Natalie Cookson[3], MSc, MBBS; Dominic King[2,4], MBChB, MRCS, PhD; Sonal Arora[2], MBBS, MRCS, PhD; Ara Darzi[1,2], FACS, FRCS, MD

[1]Department of Surgery and Cancer, St Mary's Campus, Imperial College London, London, United Kingdom
[2]Imperial Patient Safety Translational Research Centre, Department of Surgery and Cancer, Imperial College London, London, United Kingdom
[3]Academic Surgical Unit, St Marys Hospital, Imperial College Healthcare NHS Trust, London, United Kingdom
[4]Google DeepMind, London, United Kingdom

`More information` Google Deepmind, Data breach

Boost public services with AI for individualized health care

Hertie School

The Guardian

Google 'betrays patient trust' with DeepMind Health move

Royal Free breached UK data law in 1.6m patient deal with Google's DeepMind

BUSINESS INSIDER DEUTSCHLAND

Google's London AI powerhouse has set up a new healthcare division and acquired a medical app called Hark

CNBC

Google DeepMind patient data deal with UK health service illegal, watchdog says

sky news

Google received 1.6 million NHS patients' data on an 'inappropriate legal basis'

NewScientist

Revealed: Google AI has access to huge haul of NHS patient data

Did Google's NHS patient data deal need ethical approval?

`Source` Climate Change AI

Tackling Climate Change with Machine Learning

DAVID ROLNICK, McGill University and Mila - Quebec AI Institute
PRIYA L. DONTI, Carnegie Mellon University
LYNN H. KAACK, Hertie School and ETH Zürich

Climate change is one of the greatest challenges facing humanity, and we, as machine learning (ML) experts, may wonder how we can help. Here we describe how ML can be a powerful tool in reducing greenhouse gas emissions and helping society adapt to a changing climate. From smart grids to disaster management, we identify high impact problems where existing gaps can be filled by ML, in collaboration with other fields. Our recommendations encompass exciting research questions as well as promising business opportunities. We call on the ML community to join the global effort against climate change.

CrossMark

OPEN ACCESS

RECEIVED
17 August 2023

REVISED
17 November 2023

ACCEPTED FOR PUBLICATION
1 December 2023

PUBLISHED
19 January 2024

PAPER

A scalable system to measure contrail formation on a per-flight basis

Scott Geraedts[1], Erica Brand[1], Thomas R Dean[2], Sebastian Eastham[3], Carl Elkin[1], Zebediah Engberg[2], Ulrike Hager[1], Ian Langmore[1], Kevin McCloskey[1], Joe Yue-Hei Ng[1], John C Platt[1], Tharun Sankar[1], Aaron Sarna[1], Marc Shapiro[2] and Nita Goyal[1]

[1] Google, Mountain View, CA, United States of America
[2] Breakthrough Energy, Kirkland, WA, United States of America
[3] Laboratory for Aviation and the Environment, Massachusetts Institute of Technology, United States of America

E-mail: geraedts@google.com

Keywords: climate change, aviation, remote sensing

Abstract

Persistent contrails make up a large fraction of aviation's contribution to global warming. We describe a scalable, automated detection and matching (ADM) system to determine from satellite data whether a flight has made a persistent contrail. The ADM system compares flight segments to contrails detected by a computer vision algorithm running on images from the GOES-16 Advanced Baseline Imager. We develop a flight matching algorithm and use it to label each flight segment as a match or non-match. We perform this analysis on 1.6 million flight segments. The result is an analysis of which flights make persistent contrails several orders of magnitude larger than any previous work. We assess the agreement between our labels and available prediction models based on weather forecasts. Shifting air traffic to avoid regions of contrail formation has been proposed as a possible mitigation with the potential for very low cost/ton-CO2e. Our findings suggest that imperfections in these prediction models increase this cost/ton by about an order of magnitude. Contrail avoidance is a cost-effective climate change mitigation even with this factor taken into account, but our results quantify the need for more accurate contrail prediction methods and establish a benchmark for future development.

Source Geraedts et al., 2024

LLMs for more efficient administration?

Hertie School

LLM Based Multi-Agent Generation of Semi-structured Documents from Semantic Templates in the Public Administration Domain

Emanuele Musumeci[0009−0004−2359−5032][1], Michele Brienza[1][0009−0000−1549−9500], Vincenzo Suriani[1][0000−0003−1199−8358], Daniele Nardi[1][0000−0001−6606−200X], and Domenico Daniele Bloisi[2][0000−0003−0339−8651]

[1] Dept. of Computer, Control, and Management Engineering Sapienza University of Rome, Rome (Italy), {lastname}@diag.uniroma1.it
[2] UNINT Univeristy, Via Cristoforo Colombo, 200 - 00147 Rome (Italy), domenico.bloisi@unint.eu

Abstract. In the last years' digitalization process, the creation and management of documents in various domains, particularly in Public Administration (PA), have become increasingly complex and diverse. This complexity arises from the need to handle a wide range of document types, often characterized by semi-structured forms. Semi-structured documents present a fixed set of data without a fixed format. As a consequence, a template-based solution cannot be used, as understanding a document requires the extraction of the data structure. The recent introduction of Large Language Models (LLMs) has enabled the creation of customized text output satisfying user requests. In this work, we propose a novel approach that combines the LLMs with prompt engineering and multi-agent systems for generating new documents compliant with a desired structure.The main contribution of this work concerns replacing the commonly used manual prompting with a task description generated by semantic retrieval from an LLM. The potential of this approach is demonstrated through a series of experiments and case studies, showcasing its effectiveness in real-world PA scenarios.

Keywords: Human-Centered AI · Public Administration · Task optimization

Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs

Alejandro Peña[1][0000−0001−6907−5826], Aythami Morales[1][0000−0002−7268−4785], Julian Fierrez[1][0000−0002−6343−5656], Ignacio Serna[1][0000−0003−3527−4071], Javier Ortega-Garcia[1][0000−0003−0557−1948], Iñigo Puente[2], Jorge Córdova[2], Gonzalo Córdova[2]

[1] BiDA - Lab, Universidad Autónoma de Madrid (UAM), Madrid 28049, Spain
[2] VINCES Consulting, Madrid 28010, Spain

Abstract. The analysis of public affairs documents is crucial for citizens as it promotes transparency, accountability, and informed decision-making. It allows citizens to understand government policies, participate in public discourse, and hold representatives accountable. This is crucial, and sometimes a matter of life or death, for companies whose operation depend on certain regulations. Large Language Models (LLMs) have the potential to greatly enhance the analysis of public affairs documents by effectively processing and understanding the complex language used in such documents. In this work, we analyze the performance of LLMs in classifying public affairs documents. As a natural multi-label task, the classification of these documents presents important challenges. In this work, we use a regex-powered tool to collect a database of public affairs documents with more than 33K samples and 22.5M tokens. Our experiments assess the performance of 4 different Spanish LLMs to classify up to 30 different topics in the data in different configurations. The results shows that LLMs can be of great use to process domain-specific documents, such as those in the domain of public affairs.

Keywords: Domain Adaptation · Public Affairs · Topic Classification · Natural Language Processing · Document Understanding · LLM

RESEARCH ARTICLE

PAR Public Administration Review — ASPA

"Chat-Up": The role of competition in street-level bureaucrats' willingness to break technological rules and use generative pre-trained transformers (GPTs)

Neomi Frisch-Aviram | Gabriela Spanghero Lotta | Luciana Jordão de Carvalho

Getulio Vargas Foundation, São Paulo, Brazil

Correspondence
Neomi Frisch-Aviram, Vargas Foundation, Avenida 9 de julho, 2029, Bela Vista, São Paulo, SP 01313-902, Brazil.
Email: neomi.frisch82@gmail.com

Abstract
Organizations worldwide are concerned about workers using generative pretrained transformers (GPTs), which can generate human-like text in seconds at work. These organizations are setting rules on how and when to use GPTs. This article focuses on street-level bureaucrats' (SLBs) intentions to use GPTs even if their public organization does not allow its use (tech rule-breaking). Based on a mixed-methods exploratory design, using focus groups ($N = 14$) and a survey experiment ($N = 279$), we demonstrate that SLBs intend to break the rules and use GPTs when their competitors from the private sector have access to artificial intelligence (AI) tools. We discuss these findings in the context of hybrid forms of public management and the Promethean moment of GPTs.

Evidence for practice
- Regulating the use of AI by street-level bureaucrats' (SLBs) in public organizations is a growing challenge.
- SLBs have mixed feelings toward using GPTs in their work. On the one hand, they feel that GPTs can make their work more efficient, while on the other hand, they do not have the resources to learn how to use this tool well.
- However, SLBs are willing to use GPTs even if the organization does not allow their use under some circumstances. Public institutions should recognize this tendency.
- Competition with street-level professional colleagues from the private sector on resources and reputation mobilizes tech rule-breaking intentions among SLBs. When SLBs know that street-level colleagues from the private sector have access to AI tools, they are more willing to use GPTs even if the organization does not allow their use.

Analysis of Research on Artificial Intelligence in Public Administration: Literature Review and Textual Analysis

Nejc Lamovšek
Educational Research Institute, Slovenia
nejc.lamovsek@pei.si
https://orcid.org/0000-0003-3528-0527

Received: 28. 8. 2023
Revised: 25. 10. 2023
Accepted: 20. 11. 2023
Published: 30. 11. 2023

ABSTRACT

Purpose: This study aims to investigate how analysing academic research through digital tools can improve our understanding of the applications, functions, and challenges related to the use of advanced artificial technologies (AI) in public administration.

Methodology: The applied methodology relies on the use of digital tools, specifically Voyant-Tools and Chat Generative Pre-Trained Transformer (GPT-4), for text analysis in conjunction with a selection of scientific literature on artificial intelligence and public administration.

Findings: The results of our study show that researchers equally report advantages and disadvantages of using AI in public administration. Moreover, the research highlights the benefits of using artificial intelligence while emphasising the importance of the ethical and appropriate regulation thereof.

Practical implications: Our innovative approach of developing and using a combined methodology involving specialised digital tools to analyse scientific literature introduces a new dimension to the examination of scientific texts and has the potential to shape public policy in the field of public administration.

Originality: The existing body of research on public administration and artificial intelligence is limited. Our study expands the scientific field by delving into the use of artificial intelligence in public administration.

Automating Government Response to Citizens' Questions: A Large Language Model-Based Question-Answering Guidance Generation System

Keyuan Fang
IPE, Society Hub
The Hong Kong University of Science and Technology(Guangzhou)
Guangzhou, 511455. China
kfang087@connect.hkust-gz.edu.cn

Kewei Xu*
IPE, Society Hub
The Hong Kong University of Science and Technology(Guangzhou)
Guangzhou, 511455, China
* Corresponding author, coreyxu@hkust-gz.edu.cn

***Abstract*-In the era of digital governance, governments are expected to respond to citizens' inquiries timely and effectively. Nevertheless, government agencies find it increasingly difficult to cope with an unprecedented volume of citizens' inquiries on diverse issues. Governments are calling for a more effective and intelligent question-answering (QA) system to generate responses to citizens' inquiries. However, the existing QA systems in governments are primarily operated by human agents, with limited assistance from AI algorithms. The application of advanced large language models (LLMs) in digital governments holds promise to addressing citizens' requests in an automatic and effective manner. LLMs can enhance the efficiency and intelligence of government-citizen interactions, offering nuanced and human-like answers to diverse citizen inquiries. Nevertheless, existing LLMs-based all-purpose GPT application has limited understanding of professional expressions in the government domain and are unable to effectively respond like public officials. This study tries to build a QA guidance system specialized in government affairs based on LLMs and historical citizen question vector databases. After inputting a new question, the system can generate practically effective exemplary responses for government staff to reference when answering citizens' new questions. This system shows better performance than baseline models and improves the efficiency and accuracy of digital governments when answering citizens' questions.**

significant value for both citizens and the government.

Traditional digital government QA systems primarily operate in three approaches [4]. Information retrieval-based systems answer questions by searching for short text snippets in document collections [5]. Knowledge-matching systems respond to user queries by mapping natural language questions to queries on structured databases [5]. The third approach involves applying machine learning and deep learning methods to train models in learning the mapping relationships between questions and answers [6]. While these three approaches can achieve relatively good results in QA systems, they exhibit certain flaws when applied to e-government queries. The first is their relatively low comprehension of some professional and fixed expressions in the public service domain. The second is that their responses often deviate from the user's original intent. These shortcomings become more apparent in the era of digital government. Citizens' online inquiry data is vast and diverse, particularly in China, where regional policy differences are significant [7]. This necessitates strong capabilities in historical data retrieval and efficient, high- quality handling of numerous inquiries by the government. Current digital government QA systems struggle to grasp the contextual

The End of the Policy Analyst? Testing the Capability of Artificial Intelligence to Generate Plausible, Persuasive, and Useful Policy Analysis

MEHRDAD SAFAEI, Canada School of Public Service, Ottawa, Ontario, Canada
JUSTIN LONGO, Johnson Shoyama Graduate School of Public Policy, University of Regina, Regina, Saskatchewan, Canada

Policy advising in government centers on the analysis of public problems and the developing of recommendations for dealing with them. In carrying out this work, policy analysts consult a variety of sources and work to synthesize that body of evidence into useful decision support documents commonly called briefing notes. Advances in natural language processing (NLP) have led to the continuing development of tools that can undertake a similar task. Given a brief prompt, a large language model (LLM) can synthesize information in content databases. This article documents the findings from an experiment that tested whether contemporary NLP technology is capable of producing public policy relevant briefing notes that expert evaluators judge to be useful. The research involved two stages. First, briefing notes were created using three models: NLP generated; human generated; and NLP generated/human edited. Next, two panels of retired senior public servants (with only one panel informed of the use of NLP in the experiment) were asked to judge the briefing notes using a heuristic evaluation rubric. The findings indicate that contemporary NLP tools were not able to, on their own, generate useful policy briefings. However, the feedback from the expert evaluators indicates that automatically generated briefing notes might serve as a useful supplement to the work of human policy analysts. And the speed with which the capabilities of NLP tools are developing, supplemented with access to a larger corpus of previously prepared policy briefings and other policy-relevant material, suggests that the quality of automatically generated briefings may improve significantly in the coming years. The article concludes with reflections on what such improvements might mean for the future practice of policy analysis.

INACIA: Integrating Large Language Models in Brazilian Audit Courts: Opportunities and Challenges

JAYR PEREIRA, NeuralMind.ai, Campinas, Brazil and University of Campinas, Campinas, Brazil
ANDRE ASSUMPCAO, National Center for State Courts (NCSC), Williamsburg, United States and Brazilian Association of Jurimetrics (ABJ), São Paulo, Brazil
JULIO TRECENTI, Terranova Consultoria, São Paulo, Brazil
LUIZ AIROSA, Brazilian Federal Court of Accounts (TCU), Brasília, Brazil
CAIO LENTE, Terranova Consultoria, São Paulo, Brazil
JHONATAN CLÉTO, NeuralMind.ai, Campinas, Brazil and University of Campinas, Campinas, Brazil
GUILHERME DOBINS, NeuralMind.ai, Campinas, Brazil and University of Campinas, Campinas, Brazil
RODRIGO NOGUEIRA, NeuralMind.ai, Campinas, Brazil and University of Campinas, Campinas, Brazil
LUIS MITCHELL, Brazilian Federal Court of Accounts (TCU), Brasília, Brazil
ROBERTO LOTUFO, NeuralMind.ai, Campinas, Brazil and University of Campinas, Campinas, Brazil

This paper introduces INACIA (**I**nstrução **A**ssistida com **I**nteligência **A**rtificial), a groundbreaking system designed to integrate Large Language Models (LLMs) into the operational framework of Brazilian Federal Court of Accounts (TCU). The system automates various stages of case analysis, including basic information extraction, admissibility examination, *Periculum in mora* and *Fumus boni iuris* analyses, and recommendations generation. Through a series of experiments, we demonstrate INACIA's potential in extracting relevant information from case documents, evaluating its legal plausibility, and formulating propositions for judicial decision-making. Utilizing a validation dataset alongside LLMs, our evaluation methodology presents a novel approach to assessing system performance, correlating highly with human judgment. These results underscore INACIA's potential in complex legal task handling while also acknowledging the current limitations. This study discusses possible improvements and the broader implications of applying AI in legal contexts, suggesting that INACIA represents a significant step towards integrating AI in legal systems globally, albeit with cautious optimism grounded in the empirical findings.