

Survey research in the digital age

Bernhard Clemm von Hohenberg
Department of Computational Social Science
GESIS

Summer Institutes in Computational Social Science
July 28, 2023

Schedule

- ▶ 9.00-9.45 Introduction & total error survey framework
- ▶ 9.45-10.15 Probability and non-probability sampling
- ▶ Coffee break
- ▶ 10.30-11.00 Computer-administered interviewing
- ▶ 11.00-11.30 Linking surveys to big data
- ▶ 11:30-13:00 Intro and begin group exercise
- ▶ Lunch (or Eisbach plunge)
- ▶ 14:00-15:45 Continue group exercise

Credits

These materials build heavily on Matthew Salganik's 2019 SICSS class as well as Chapter 3 of "Bit by Bit: Social Research in the Digital Age".

	Sampling	Interviews	Data environment
1st era	Area probability	Face-to-face	Stand-alone
2nd era	Random digital dial probability	Telephone	Stand-alone
3rd era	Non-probability	Computer-administered	Linked

Will big data kill surveys?



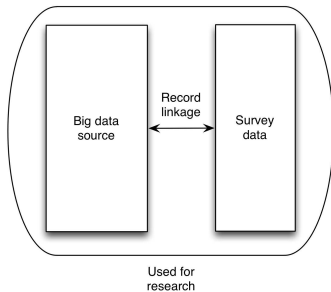
<http://schlitterblog.com/wp-content/uploads/2014/05/peanutbutterlover.jpg>



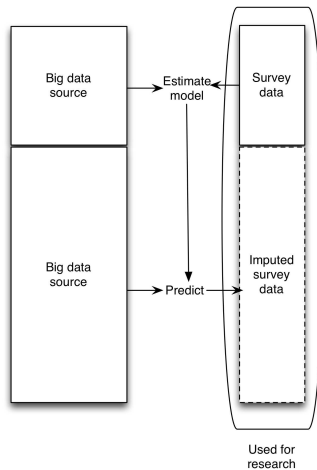
<http://schlitterblog.com/wp-content/uploads/2014/05/peanutbutterlover.jpg>

Overview

Enriched asking



Amplified asking



Enriched asking

Survey data builds context around a big data source (or vice versa) that contains some important measurements but lack others.

:



AMERICAN JOURNAL
of POLITICAL SCIENCE

(Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets

Andrew M. Guess

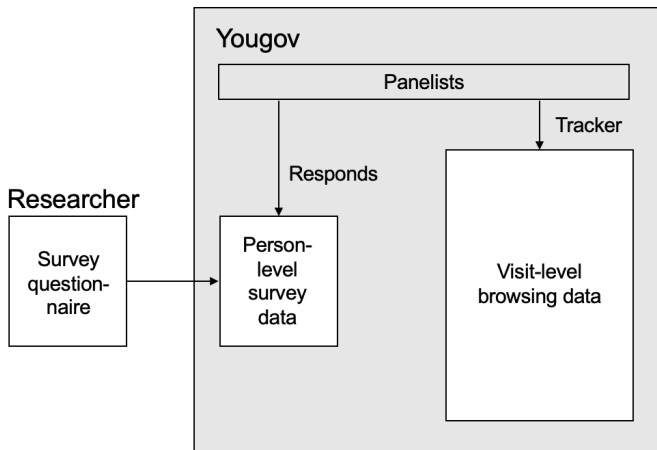
Princeton University

Abstract: *Does the internet facilitate selective exposure to politically congenial content? To answer this question, I introduce and validate large-N behavioral data on Americans' online media consumption in both 2015 and 2016. I then construct a simple measure of media diet slant and use machine classification to identify individual articles related to news about politics. I find that most people across the political spectrum have relatively moderate media diets, about a quarter of which consist of mainstream news websites and portals. Quantifying the similarity of Democrats' and Republicans' media diets, I find nearly 65% overlap in the two groups' distributions in 2015 and roughly 50% in 2016. An exception to this picture is a small group of partisans who drive a disproportionate amount of traffic to ideologically slanted websites. If online "echo chambers" exist, they are a reality for relatively few people who may nonetheless exert disproportionate influence and visibility.*

<https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12589>

Researcher

Survey
question-
naire



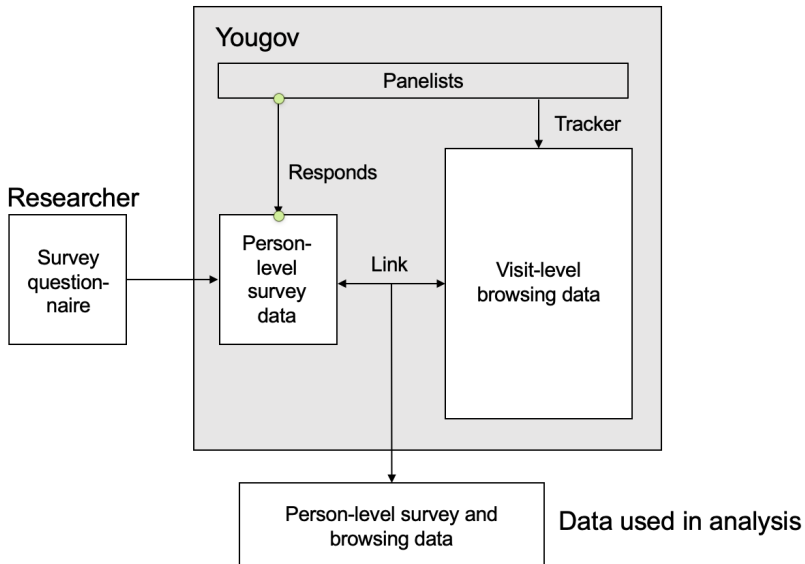
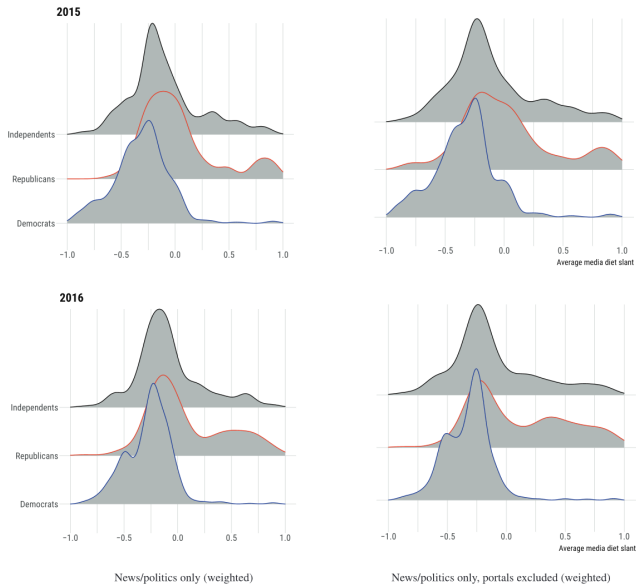



FIGURE 1 Americans' Online Media Diets by Partisanship



Another rising approach: data donations

 Python application **passing** code style black

› OSD2F: Open Source Data Donation Framework

Goal

Use OSD2F to run your own Data Donation service. The aim of this project is to facilitate scientists to collect data donations, by providing an easy-to-use web-based data donation platform. Here, scientists can instruct participants in their research to upload data exports from major online platforms (generally based on participants rights to their own data under GDPR).

Amplified asking

Using a predictive model to combine survey data from a few people with a big data source from many people.

ECONOMICS

Predicting poverty and wealth from mobile phone metadata

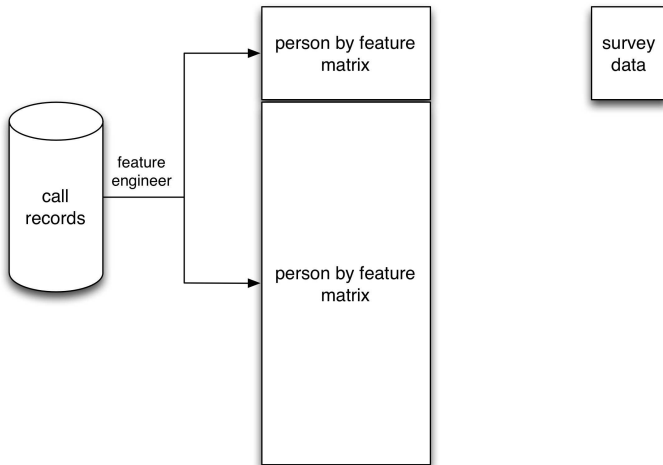
Joshua Blumenstock,^{1*} Gabriel Cadamuro,² Robert On³

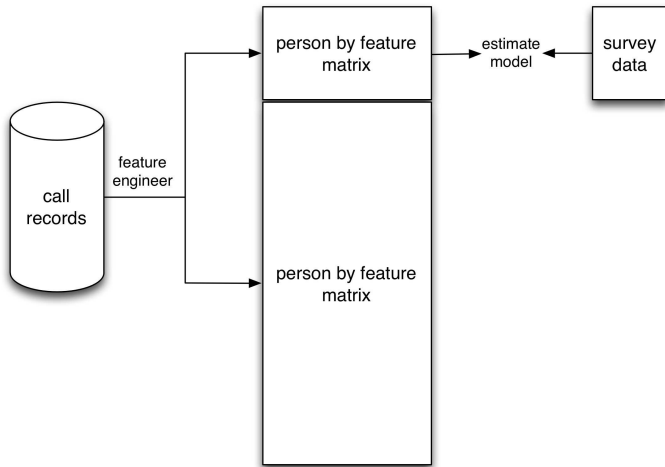
Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer sources of big data exist. We show that an individual's past history of mobile phone use can be used to infer his or her socioeconomic status. Furthermore, we demonstrate that the predicted attributes of millions of individuals can, in turn, accurately reconstruct the distribution of wealth of an entire nation or to infer the asset distribution of microregions composed of just a few households. In resource-constrained environments where censuses and household surveys are rare, this approach creates an option for gathering localized and timely information at a fraction of the cost of traditional methods.

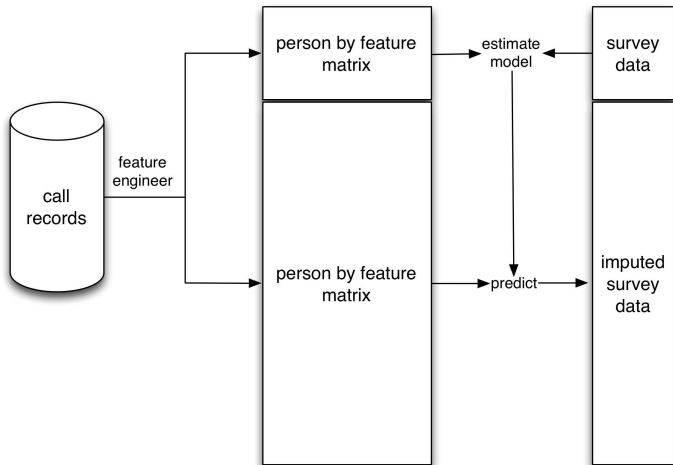
<http://dx.doi.org/10.1126/science.aac4420>

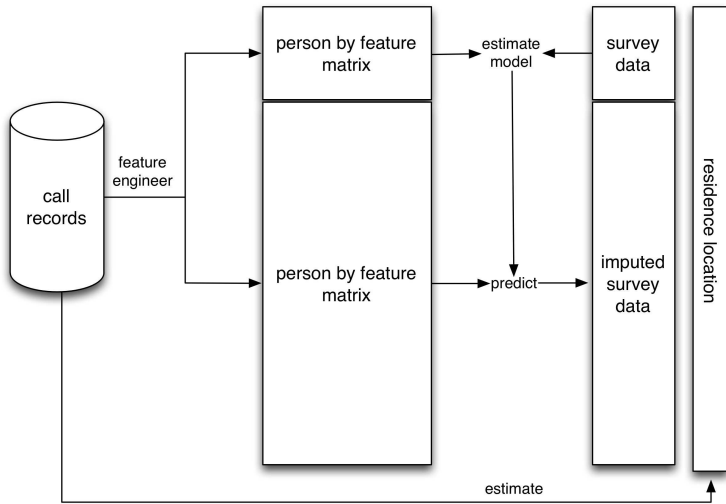


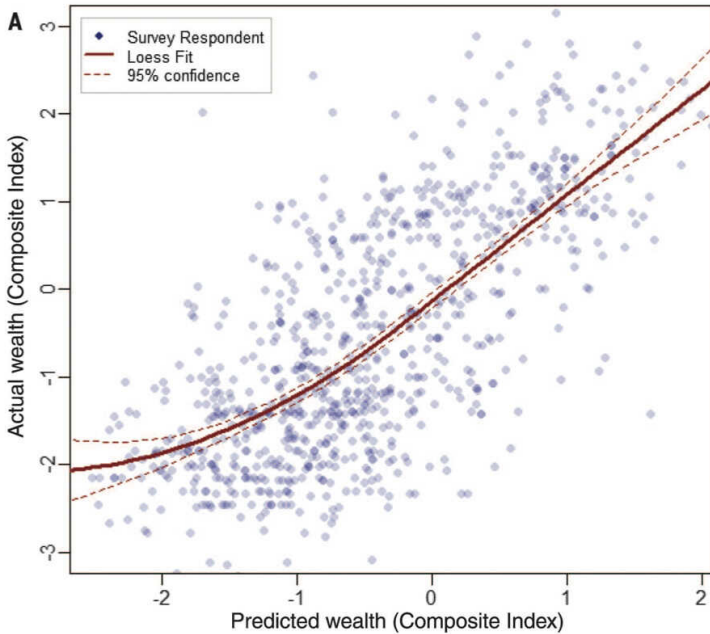


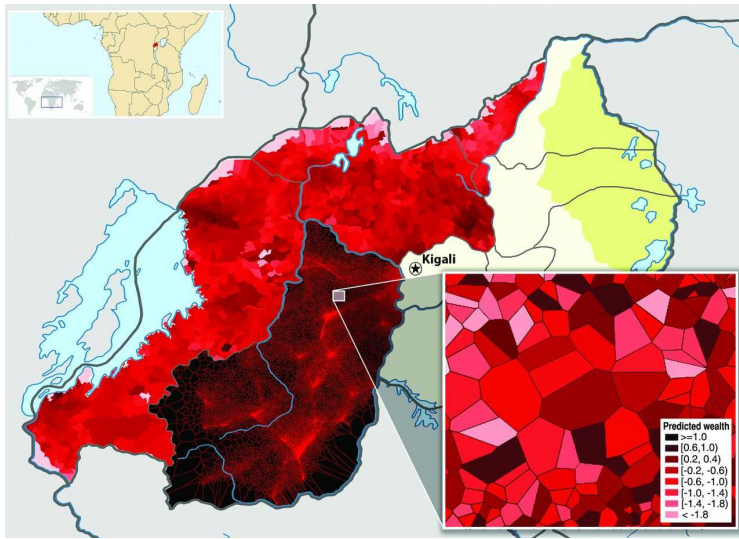


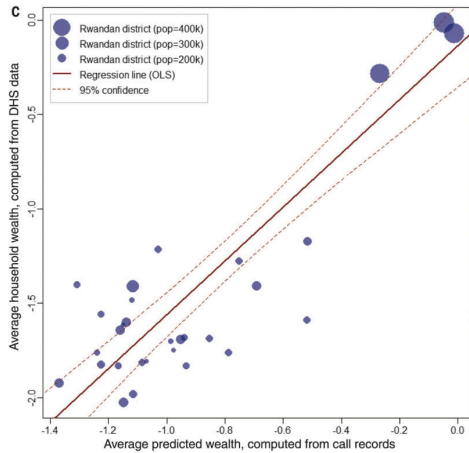


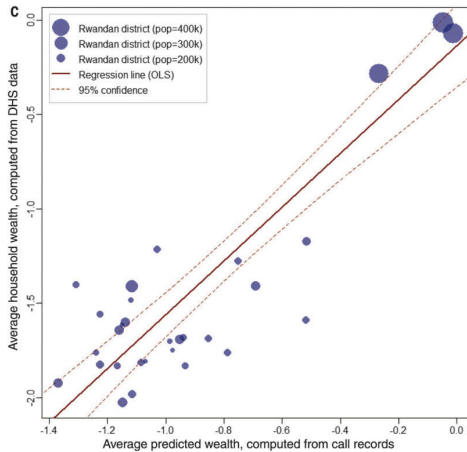












- ▶ 10 times faster
- ▶ 50 times cheaper

Summary

- ▶ Surveys and big data are compliments not substitutes

Summary

- ▶ Surveys and big data are compliments not substitutes
- ▶ Sometime we do “enriched asking” and sometimes “amplified asking” (role of big data source is different in both cases)

Summary

- ▶ Surveys and big data are compliments not substitutes
- ▶ Sometime we do “enriched asking” and sometimes “amplified asking” (role of big data source is different in both cases)
- ▶ The black box of many big-data providers is a challenge for scientists

Questions