The Bay Area Consortium of Population Researchers (BACPOP)

Sponsored by: U.C. Berkeley Department of Demography&The Berkeley Population Center
http://www.popcenter.berkeley.edu

November 21, 2014

PRESENTING

1 Emmanuel Letouzé
Director & Co-founder of the Data Pop Alliance on Big Data and Development, Ph.D. Candidate in Demography, UC Berkeley

“Big Data and Demo-Economic Research: Applications and Implications from the Case of Human Mobility Analysis.”

2 Sharad Goel
Assistant Professor, Management Science and Engineering, Stanford University

"Forecasting Elections with Non-Representative Polls."

3 Tapan Parikh
Assistant Professor, UC Berkeley School of Information

"Data in the First Mile."

4 Maya Petersen
Assistant Professor of Biostatistics and Epidemiology, Berkeley School of Public Health

"From Data to Impact: Drawing Causal Inferences from Big Data."

5 Emilio Zagheni
Assistant Professor, Department of Sociology, University of Washington, Seattle

Session Organizer and Discussant

Population Research in the Age of Big Data

Workshop background

The Bay Area Colloquium on Population (BACPOP) hosts noted demographers and scholars from related fields, bringing together specialists interested in demography from the entire Bay Area. We hold three BACPOP events each academic year, at the Berkeley Faculty Club. These are half-day gatherings, starting with an informal buffet lunch followed by presentations, moderated discussion, and Q&A from attendees.

BACPOP hosted its first re-designed, workshop-style seminar on November 21, 2014, taking up a topic that is very much of the moment – Big Data – through the lens of several disciplines, considering what it implies, and how it may be informed by, demography and population research. The seminar was a great success, with the need for a waitlist just a few days after the opening of registration.

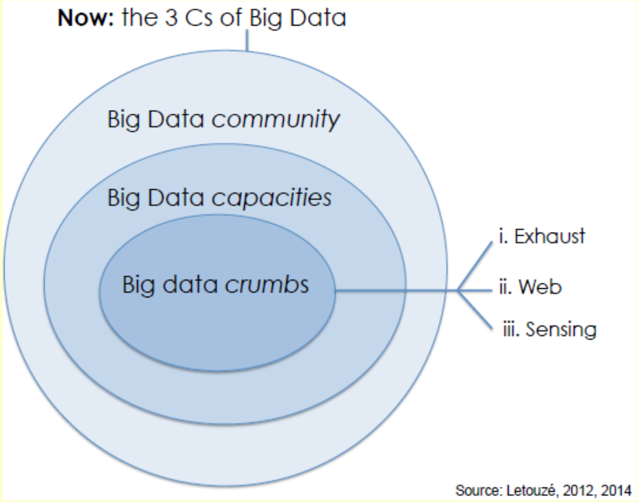
General Themes & Common Threads

Brief overview of the field

The “data revolution” and the advent of big data have paved the way for new possibilities in understanding social change and human behavior on a global scale. Under examination in this workshop was the question of how a data-driven approach aids in the construction of social and demographic facts, with data serving as a mechanism through which theories can be tested and new theories developed in an effort to inform social policy and problems of societal importance.

Emmanuel Letouzé of U.C. Berkeley Demography provided a general overview of the field, asserting that big data as a research ecosystem can be descriptive, predictive and/or prescriptive.Though we have been talking about big data for the last 30+ years, big data as we understand it today is about five years old, representing the next frontier for innovation, competition and productivity.

Letouzé characterizes data as having three main characteristics, as represented by the illustration bellow.



Source: Letouzé, 2012, 2014

Multiple presenters directly and indirectly touched on the necessity to consider and model the bias inherent in this method of data collection. It is necessary to consider theoretical drivers behind the data collected, in addition to balancing the feedback loops between theory and data. A range of skills and disciplinary approaches is needed to ensure that the variables present in the data represent the diversity of the target population, ensuring a representative sample.

Applications & Implications

Researchers need to consider the applications as well as the ethical implications of their research. There is an underlying tradeoff between generating new information with low confidence and potentially foregoing innovation by replicating existing knowledge. The ethics of big data are an emerging field. A question raised in the workshop was the potential to inadvertently perpetuate existing inequalities, as the technological skill gap within countries is still large, risking the creation of a new global digital divide.

Tangible applications of Big Data discussed include: crime & mobility (Letouzé), Xbox surveys & elections (Goel), air quality and commuting (Parikh), and geolocated data & migration (Zagheni).

Tradeoffs & Innovations

A key area of discussion was the ongoing debate in the field regarding representative sampling. Sharad Goel of Stanford University focused on techniques to deal with non-representative samples, grounding model design in theory in order to ensure scientific merit. His work highlights the importance of considering the demographic attributes of the respondents in non-representative samples and making statistical inferences conditional on the model. Workshop participants additionally proposed that online polls are particularly vulnerable to "hijacking" for political purposes.

Goel suggested the necessity for the incorporation of methods for detecting anomalous behavior in online data collection. Participants also addressed the potential for respondent fatigue in online surveys, touching on the imperative to consider respondent schedules and the "fun factor" that can be incorporated in online surveys better than in traditional polls. Goel underscored that this pertinent yet under-studied issue supports the need for increased attention to ensuring that the incorporation of the additional techniques mentioned above still draws high response rates and generates reliable results.

Numerous participants and attendees touched on the potential integration of different approaches from multiple domains such as respondent-driven sampling with a network based approach and propensity score adjustments.

The workshop highlighted tensions and tradeoffs between data collected from traditional surveys and social media. Emilio Zagheni of the University of Washington, Seattle proposed an innovative approach to addressing the above tradeoff by suggesting the collection of "socio-markers" (an analogy of biomarkers), incorporating the collection of social media data from sample survey respondents.

New Frontiers in Data Collection

Tapan Parikh drew attention to harnessing and building on the data literacy of the general population, with a specific focus on youth in order to teach technological skills for community-level advocacy. His innovative model specifically incorporates the communication and translation of findings into policies, highlighting the imperative to consider data as part of the political apparatus of the state. He calls for the incorporation of multiple, community-driven and informed data collection methodologies to fully capture the lived experiences of the respondents.

In a similar vein, Maya Petersen of the University of California, Berkeley also calls for analyses targeting complex policy questions, asserting that "instead of simply describing the status quo, we must learn how to improve it!" Her work focuses on the substantial methodological challenges faced when working on the new frontier of big data as researchers seek to translate complex causal questions into statistical questions that allow them to get the most out of very rich but "messy" data.

Letouzé incorporates cellular technology to collect real-time data, using cell phones as "social sensors". Using a "place-centric perspective" and data-driven approach for predicting crime hotspots, his work incorporates a ground-breaking multi-modal approach, with a specific focus on crime prediction instead of individual profiling. His work describes novel "risk-inducing or risk-reducing" features of geographic areas.

Fresh Perspectives & Contributions of Population Scientists

The workshop discussions highlighted several promising avenues for further research. Some specific areas identified include:

- Conceptual and theoretical frontiers to include the movement from predictive to prescriptive uses of big data.
- Ethical and institutional implications and the need to create a new ethical framework.
- Constructing the "age pyramid of big data" with size and data renewal in mind. More broadly, using demographic methods to understand populations of digital objects.
- The possibility to extend data collected in the past beyond the original hypothesis. This includes the need for repositories of anonymized data.
- The integration of machine learning approaches in predictive models.
- The need to incorporate multiple methodologies of data collection and analysis, to evaluate data quality and address the issue of selection bias.
- The necessity to think critically about the tension between "quick and dirty" research utilizing small bits of big data vs conducting in-depth investigations and research projects.
- The necessity to think critically about the issue of replicability of our research.
- The imperative to consider the ethical issues surrounding data retention: How long should our data live?
- Further investigation of the limits of the power of big data, as currently big data is powerful in some fields while being non-existent in other fields.
- The potential of technology to capture new and novel data.
- The need to address the problem of confidentiality when it comes to securely linking large datasets.

Fresh Perspectives cont.

- Emphasis on the need to develop methods (e.g., the father of the "life table" was a big data scientist if we consider technology available 350 years ago. However, we remember him for his methods, not for the size of the data he used.
- Opportunities to bridge the gap between qualitative and quantitative research using structured data like videos, interviews, text and drawings.

References & Additional Resources

Links:

Center on the Economics and Demography of Aging

http://www.ceda.berkeley.edu

Berkeley Population Center

http://www.popcenter.berkeley.edu/

Berkeley Department of Demography

http://www.demog.berkeley.edu/

Datasets:

http://webscope.sandbox.yahoo.com/catalog.php

https://www.yelp.com/academic_dataset

http://snap.stanford.edu/data/

http://www.d4d.orange.com/

Acknowledgements

BACPOP would like to acknowledge the contributions made to the success of our re-designed workshop-style seminar.
The event was a great success, due in part to the outstanding contributions of the presenters, many who traveled far to join us.

We also wish to gratefully acknowledge the hard work and dedication of Monique Verrier and Pavenpreet Dhanoa for their responsiveness and meeting support.

We would like to give special thanks to Professor Zagheni for his efforts in organizing this event and to Emmanuel Letouzé for contributing his refreshing art work for this report.

Next Workshop

Mortality and Inequality
3/13/2015

Professor Ronald D. Lee, University of California, Berkeley
Session Organizer