Introduction to mass collaboration

Matthew J. Salganik
Department of Sociology
Princeton University

Summer Institute in Computational Social Science
June 21, 2019

The Summer Institute in Computational Social Science is supported by grants from the Russell Sage Foundation and the Alfred P. Sloan Foundation.

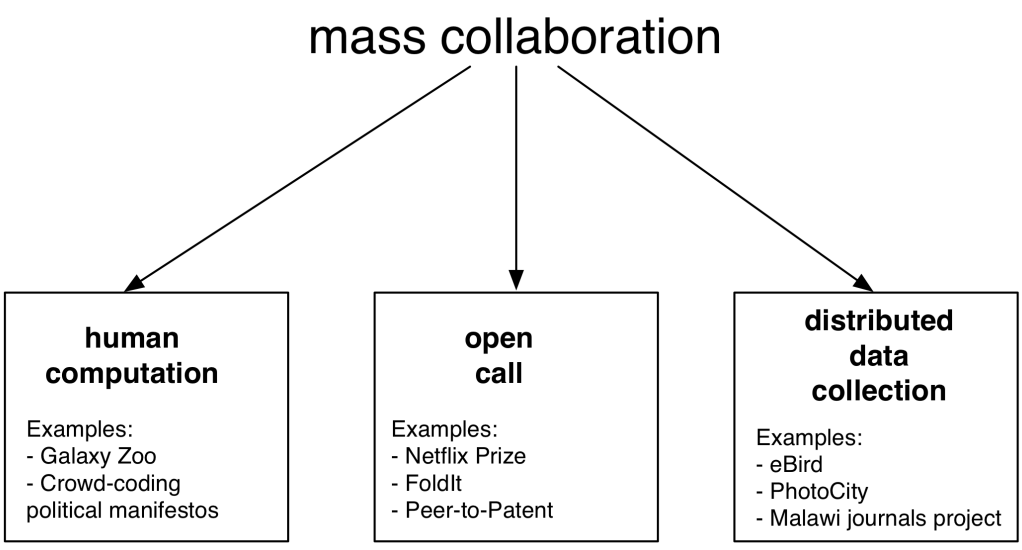
Day schedule HERE

- 9:15 - 9:45: Introduction to mass collaboration
- 9:45 - 10:15 Quick and dirty introduction to applied predictive modeling
- 10:15 - 10:30 Coffee break
- 10:30 - 11:30 Introduction to the Fragile Families Challenge
- 11:30 - 12:30 Working on the Fragile Families Challenge (*Not open to public/No livestream*)
- 12:30 - 1:30 Lunch
- 1:30 - 3:30 Fragile Families Challenge (*Not open to public/No livestream*)
- 3:30 - 3:45 Discussion of the Fragile Families Challenge (*Not open to public/No livestream*)
- 3:45 - 4:00 Break
- 4:00 - 5:30 Guest speaker: Annie Liang
- 6:00 - 7:30 Dinner & Discussion (*Not open to public/No livestream*)

- Observing behavior
- Asking questions
- Running experiments
- Creating mass collaboration

Mass collaboration combines ideas from

- crowdsourcing
- citizen science
- collective intelligence


mass collaboration
human computation
Examples:
- Galaxy Zoo
- Crowd-coding political manifestos
open call
Examples:
- Netflix Prize
- FoldIt
- Peer-to-Patent
distributed data collection
Examples:
- eBird
- PhotoCity
- Malawi journals project

Guiding idea:
Collaborators not cogs (ornithology and astronomy are examples)

- Is this really research?

- Is this really research?
- Does this enable new research?

- Is this perfect?

- Is this perfect?
- Is this better than we can do without mass collaboration?

- Is this impossible?

- Is this impossible?
- Is this possible?

An honest assessment:

An honest assessment: As far as I can tell, most mass collaborations fail

- Human computation
- Open call
- Distributed data collection

- Easy task, big scale

- Easy task, big scale
- Humans better than computers

- Easy task, big scale
- Humans better than computers
- Can be combined with supervised learning

- Easy task, big scale
- Humans better than computers
- Can be combined with supervised learning
- Increasingly important as we move from numeric survey data to working with text, images, movies, audio, etc.

Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data

KENNETH BENOIT *London School of Economics and Trinity College*
DREW CONWAY *New York University*
BENJAMIN E. LAUDERDALE *London School of Economics and Political Science*
MICHAEL LAVER *New York University*
SLAVA MIKHAYLOV *University College London*

http://dx.doi.org/10.1017/S0003055416000058

Empirical social science often relies on data that are not observed in the field, but are transformed into quantitative variables by expert researchers who analyze and interpret qualitative raw sources. While generally considered the most valid way to produce data, this expert-driven process is inherently difficult to replicate or to assess on grounds of reliability. Using crowd-sourcing to distribute text for reading and interpretation by massive numbers of nonexperts, we generate results comparable to those using experts to read and interpret the same texts, but do so far more quickly and flexibly. Crucially,

http://dx.doi.org/10.1017/S0003055416000058

Here's a piece of the manifesto of the Labor Party in the United Kingdom from 2010:

> *"Millions of people working in our public services embody the best values of Britain, helping empower people to make the most of their own lives while protecting them from the risks they should not have to bear on their own. Just as we need to be bolder about the role of government in making markets work fairly, we also need to be bold reformers of government."*

FIGURE 1. Hierarchical Coding Scheme for Two Policy Domains with Ordinal Positioning

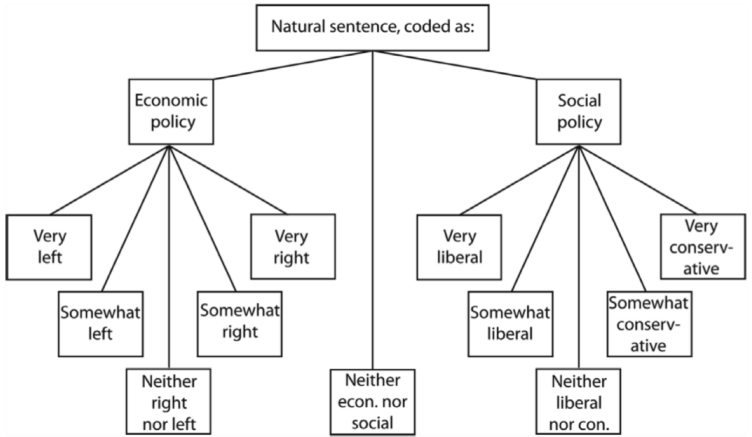
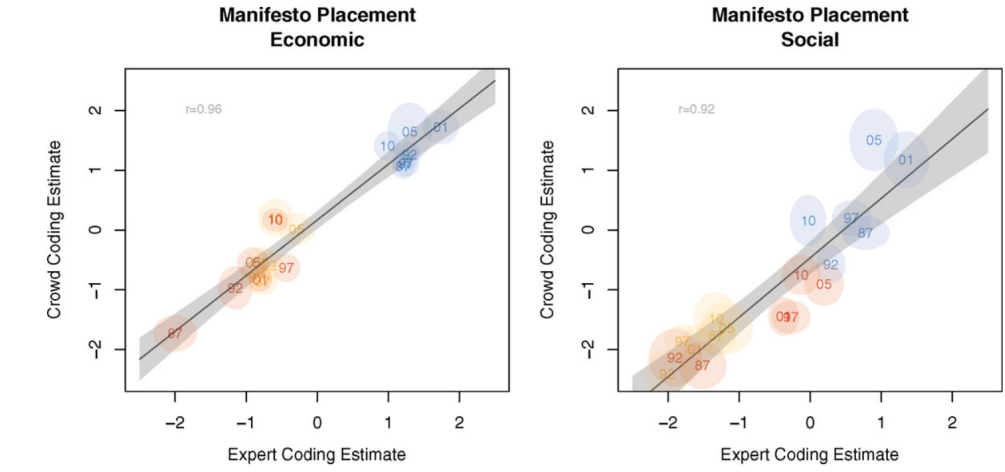


FIGURE 3. Expert and Crowd-sourced Estimates of Economic and Social Policy Positions



What I like about Benoit et al (2016)

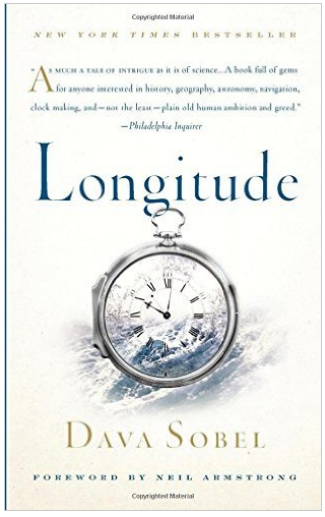
- Better not cheaper

What I like about Benoit et al (2016)

- Better not cheaper
- Experts are a bug not a feature

Questions?

- Human computation
- Open call
- Distributed data collection


Copyrighted Material
NEW YORK TIMES BESTSELLER
"As much a tale of intrigue as it is of science...A book full of gems for anyone interested in history, geography, astronomy, navigation, clock making, and—not the least—plain old human ambition and greed."
—Philadelphia Inquirer
Longitude
DAVA SOBEL
FOREWORD BY NEIL ARMSTRONG
Copyrighted Material

Solutions are easier to check than to generate

You will participate in an open call in a few moments

Questions?

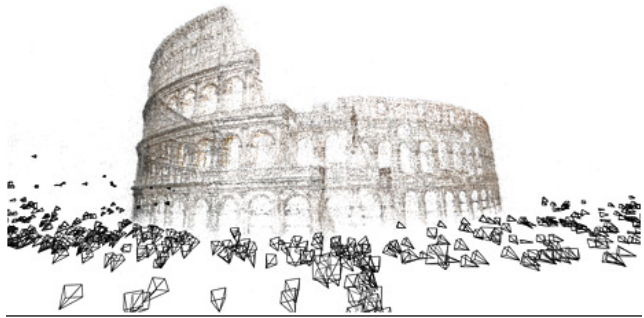
- Human computation
- Open call
- Distributed data collection

- people can be where the researchers can't

- people can be where the researchers can't
- scale that researcher cannot match

- people can be where the researchers can't
- scale that researcher cannot match
- sometimes hard to separate from human computation

Tuite et al. (2011) "PhotoCity: Training Experts at Large-scale Image Acquisition Through a Competitive Game" *CHI*: http://dx.doi.org/10.1145/1978942.1979146



Rome in a Day (Agarwal et al., 2009)

Two campuses: University of Washington and Cornell University

Over 2 months, 100,000 photos submitted by 45 players

(a) Lewis Hall (UW)

(b) Sage Chapel (Cornell)

(c) Uris Library (Cornell)

PhotoCity

Beautiful design solves lots of problems

PhotoCity

Beautiful design solves lots of problems

- data collection is standardized because of cameras

PhotoCity

Beautiful design solves lots of problems

- data collection is standardized because of cameras
- verification is automatic by comparison with nearby images

PhotoCity

Beautiful design solves lots of problems

- data collection is standardized because of cameras
- verification is automatic by comparison with nearby images
- game points are assigned based on the value of data, trains people to collect more valuable data

Questions about distributed data collection or mass collaboration?