[Introduction to mass collaboration], [Human computation], [Open call], [Distributed data collection], [Fragile Families Challenge]

Matthew J. Salganik
Department of Sociology
Princeton University

1) Introduction
2) Observing behavior
3) Asking questions
4) Running experiments
5) Mass collaboration
6) Ethics
7) The future

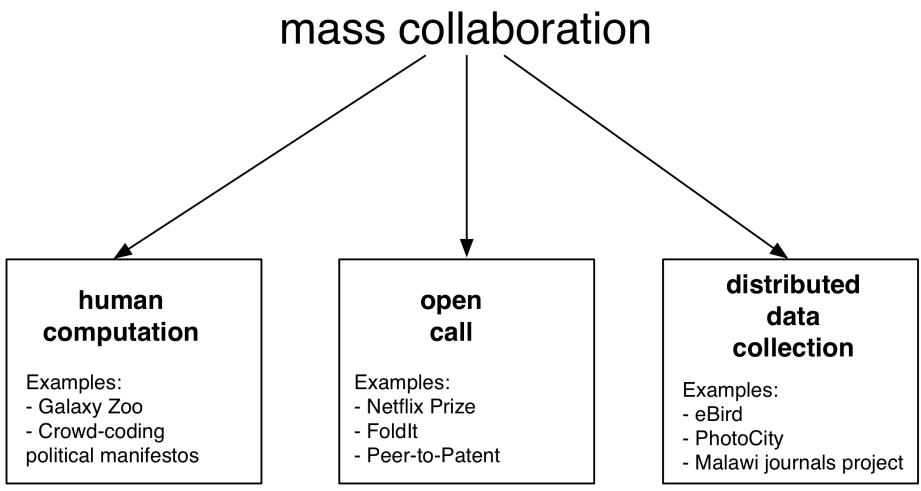


Fig 5.4 (Salganik 2018)

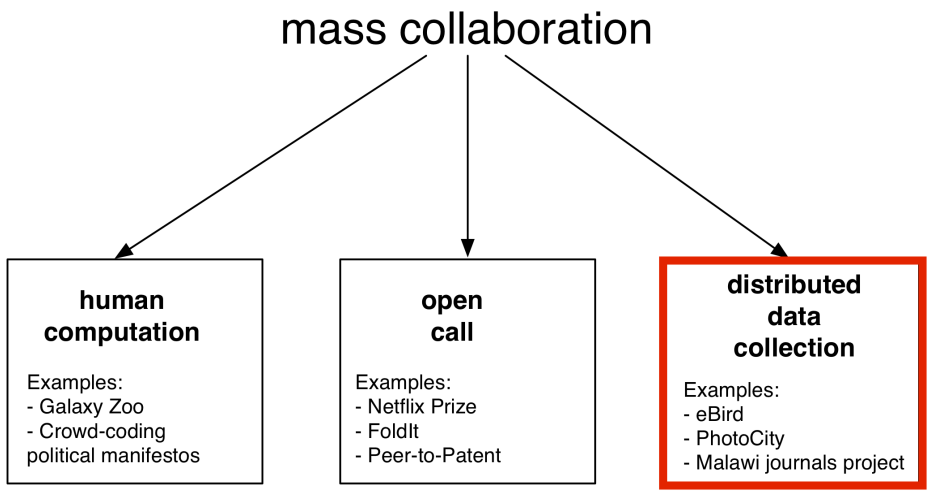


Fig 5.4 (Salganik 2018)

Distributed data collection:

- people can be where the researchers can't

Distributed data collection:

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

Distributed data collection:

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

Claims:

- Distributed data collection is possible for real research

Claims:

- Distributed data collection is possible for real research
- Sampling and data quality concerns are not insurmountable

Claims:

- Distributed data collection is possible for real research
- Sampling and data quality concerns are not insurmountable
- Distributed data collection can produce different—not just cheaper—data

eBird

eBird

- Builds on a long tradition in ornithology

eBird

- Builds on a long tradition in ornithology
- Takes advantage of “work” that is already happening anyway

eBird

- Builds on a long tradition in ornithology
- Takes advantage of "work" that is already happening anyway
- Huge amounts of data over wide geographic scale: 250,000 participants who have submitted more than 260 million bird sightings

Spatiotemporal Variation in Avian Migration Phenology: Citizen Science Reveals Effects of Climate Change

Allen H. Hurlbert*, Zhongfei Liang

https://doi.org/10.1371/journal.pone.0031662

eBird

But, data is complex

- Despite input filters that remove clearly incorrect data, the data quality is unclear

eBird

But, data is complex

- Despite input filters that remove clearly incorrect data, the data quality is unclear
- Location of observed birds is based on location of birders

eBird

But, data is complex

- Despite input filters that remove clearly incorrect data, the data quality is unclear
- Location of observed birds is based on location of birders
- Heterogeneity in observer skill and protocol

eBird

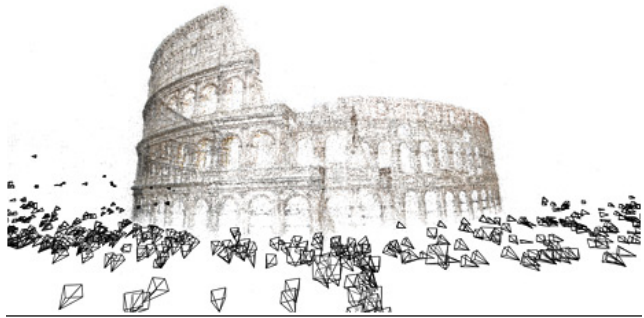
But, data is complex

- Despite input filters that remove clearly incorrect data, the data quality is unclear
- Location of observed birds is based on location of birders
- Heterogeneity in observer skill and protocol

eBird data has been used in more than 200 scientific papers:
https://ebird.org/science/publications

PhotoCity

Tuite et al. (2011) "PhotoCity: Training Experts at Large-scale Image Acquisition Through a Competitive Game" *CHI*.



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

PhotoCity

Two campuses: University of Washington and Cornell University

PhotoCity

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

- 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

PhotoCity: Player strategies

- "[I tried to] approximate the time of day and the lighting that some pictures were taken; this would help prevent rejection by the game. With that said, cloudy days were the best by far when dealing with corners because less contrast helped the game figure out the geometry from my pictures"

PhotoCity: Player strategies

- "[I tried to] approximate the time of day and the lighting that some pictures were taken; this would help prevent rejection by the game. With that said, cloudy days were the best by far when dealing with corners because less contrast helped the game figure out the geometry from my pictures"
- "When it was sunny, I utilized my camera's anti-shake features to allow myself to take photos while walking around a particular zone. This allowed me to take crisp photos while not having to stop my stride. Also bonus: less people stared at me!"

PhotoCity: Player strategies

- "[I tried to] approximate the time of day and the lighting that some pictures were taken; this would help prevent rejection by the game. With that said, cloudy days were the best by far when dealing with corners because less contrast helped the game figure out the geometry from my pictures"
- "When it was sunny, I utilized my camera's anti-shake features to allow myself to take photos while walking around a particular zone. This allowed me to take crisp photos while not having to stop my stride. Also bonus: less people stared at me!"
- "Taking many pictures of one building with 5 megapixel camera, then coming home to submit, sometimes up to 5 gigs on a weekend shoot, was primary photo capture strategy. Organizing photos on external hard drive folders by campus region, building, then face of building provided good hierarchy to structure uploads."

Malawi Journal Project

Project:

- part of Malawi Diffusion and Ideational Change Project
- 22 citizen "journalists" write down all the conversations they hear about AIDS
- 15 year timespan resulting in about 12,000 pages of text

Malawi Journal Project

Project:

- part of Malawi Diffusion and Ideational Change Project
- 22 citizen "journalists" write down all the conversations they hear about AIDS
- 15 year timespan resulting in about 12,000 pages of text

Result:

- "journalists" access very different knowledge than Western researchers and formal surveys

Claims:

- Distributed data collection is possible for real research (eBird)

Claims:

- Distributed data collection is possible for real research (eBird)
- Sampling and data quality concerns are not insurmountable (PhotoCity)

Claims:

- Distributed data collection is possible for real research (eBird)
- Sampling and data quality concerns are not insurmountable (PhotoCity)
- Distributed data collection can produce different—not just cheaper—data (Malawi Journal Project)

What to read next:

- Sullivan, Wood, Iliff, Bonney, Fink, and Kelling. 2009. "eBird: a citizen-based bird observation network in the biological sciences. *Biological Conservation*.
- Kaler, Watkins, and Angotti, 2015. "Making meaning in the time of AIDS: longitudinal narratives from the Malawi Journals Project", *African Journal of AIDS Research*.
- *Bit by Bit*, Section 5.5 Designing your own

[Introduction to mass collaboration], [Human computation], [Open call], [Distributed data collection], [Fragile Families Challenge]

Matthew J. Salganik
Department of Sociology
Princeton University