[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

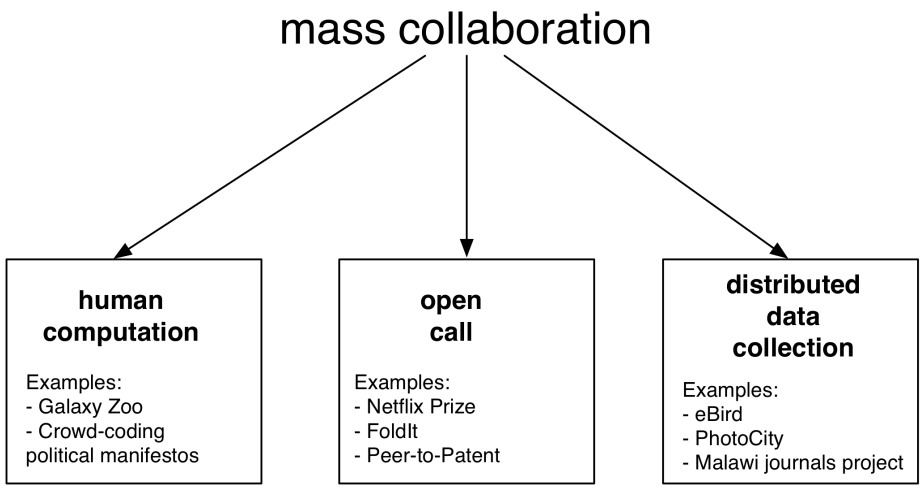


Fig 5.4 (Salganik 2018)

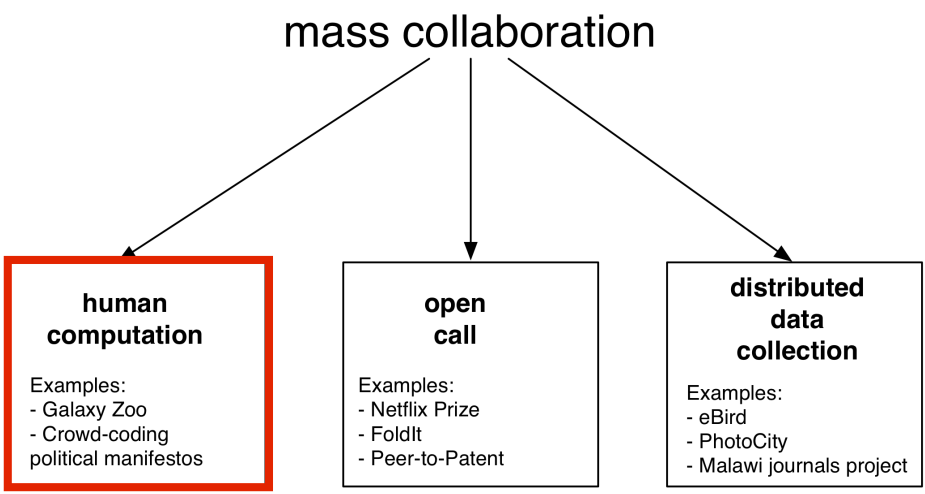


Fig 5.4 (Salganik 2018)

Human computation:

- Easy task, big scale problems where humans better than computers

Human computation:

- Easy task, big scale problems where humans better than computers
- Split-apply-combine strategy

Human computation:

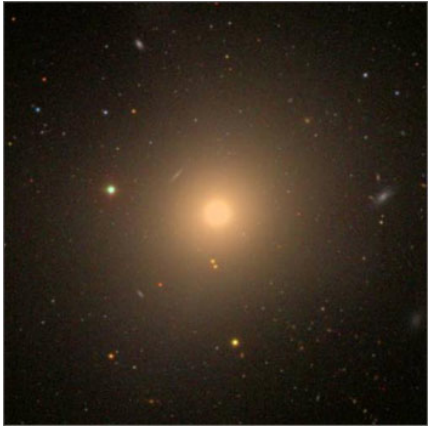
- Easy task, big scale problems where humans better than computers
- Split-apply-combine strategy
- Human effort can be magnified with supervised learning

Human computation:

- Easy task, big scale problems where humans better than computers
- Split-apply-combine strategy
- Human effort can be magnified with supervised learning
- Increasingly important as we move from numeric survey data to working with text, images, movies, and audio.

Galaxy Zoo

Astronomers are interested in understanding the relationship between the shape and color of galaxies



(a) Elliptical

(b) Spiral

Galaxy Zoo

Needed hand-classified galaxies so Schawninski worked seven, 12 hour days to classify 50,000 galaxies

The Astrophysical Journal Supplement Series, 173:512–523, 2007 December
© 2007. The American Astronomical Society. All rights reserved. Printed in U.S.A.

THE EFFECT OF ENVIRONMENT ON THE ULTRAVIOLET COLOR-MAGNITUDE RELATION OF EARLY-TYPE GALAXIES

K. Schawinski,[1] S. Kaviraj,[1] S. Khochfar,[1] S.-J. Yoon,[2,1] S. K. Yi,[2,1,] J.-M. Deharveng,[3] A. Boselli,[3] T. Barlow,[4] T. Conrow,[4] K. Forster,[4] P. G. Friedman,[4] D. C. Martin,[4] P. Morrissey,[4] S. Neff,[5] D. Schiminovich,[6] M. Seibert,[4] T. Small,[4] T. Wyder,[4] L. Bianchi,[7] J. Donas,[3] T. Heckman,[7] Y.-W. Lee,[2] B. Madore,[8] B. Milliard,[3] R. M. Rich,[9] and A. Szalay[7]

Received 2005 November 2; accepted 2005 December 28

Galaxy Zoo

Needed hand-classified galaxies so Schawninski worked seven, 12 hour days to classify 50,000 galaxies

The Astrophysical Journal Supplement Series, 173:512–523, 2007 December
© 2007. The American Astronomical Society. All rights reserved. Printed in U.S.A.

THE EFFECT OF ENVIRONMENT ON THE ULTRAVIOLET COLOR-MAGNITUDE RELATION OF EARLY-TYPE GALAXIES

K. Schawinski,[1] S. Kaviraj,[1] S. Khochfar,[1] S.-J. Yoon,[2,1] S. K. Yi,[2,1,] J.-M. Deharveng,[3] A. Boselli,[3] T. Barlow,[4] T. Conrow,[4] K. Forster,[4] P. G. Friedman,[4] D. C. Martin,[4] P. Morrissey,[4] S. Neff,[5] D. Schiminovich,[6] M. Seibert,[4] T. Small,[4] T. Wyder,[4] L. Bianchi,[7] J. Donas,[3] T. Heckman,[7] Y.-W. Lee,[2] B. Madore,[8] B. Milliard,[3] R. M. Rich,[9] and A. Szalay[7]

Received 2005 November 2; accepted 2005 December 28

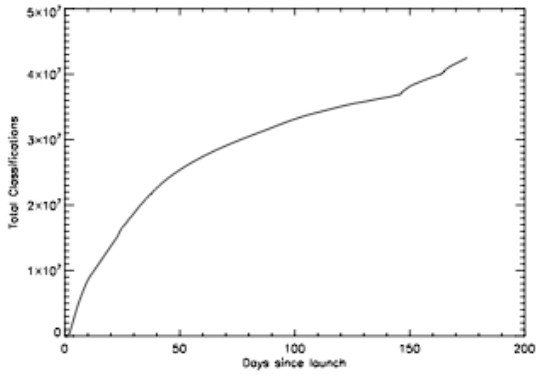
Only 5% of the $\sim$ 1 million galaxies in the Sloan Digital Sky Survey. A new approach was needed . . . .


GALAXY ZOO.org
Welcome
Home
The Science
How to Take Part
Galaxy Analysis
Forum
Press & News
FAQ
Links
Contact Us
Login
Register
Galaxy Tutorial
Galaxy Analysis
Galaxy Analysis
Welcome to Galaxy Zoo's view of the Universe. If you're here you should already have seen the Tutorial, but feel free to go and remind yourself. There's no need to agonise for too long over any one image, just make your best guess in each case.
Galaxy Ref:
588010880371851294
Choose the Galaxy Profile by clicking the buttons below
CLOCK
ANTI
EDGE ON/ UNCLEAR
SPIRAL GALAXY
ELLIPTICAL GALAXY
STAR / CONTAMIN
MERGERS
Show Grid Overlay on the next Image

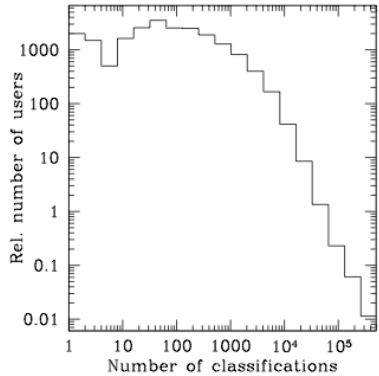
Galaxy Zoo

- Volunteers had a $\sim$5 minute training and passed a quiz
- Categorized as many or as few galaxies as they wished
- Much of the recruiting happened through the media

Galaxy Zoo



(a) Classifications over time



(b) Classifications per user

Galaxy Zoo

40 million classification to a consensus labels (Lintott et al., 2011)

1. Cleaning
   - only the first classification that a volunteer made of a specific galaxy was used in the analysis
   - anyone who classified more than 2 galaxies more than 5 times each had all their classifications discarded

Galaxy Zoo

40 million classification to a consensus labels (Lintott et al., 2011)

1. Cleaning
   - only the first classification that a volunteer made of a specific galaxy was used in the analysis
   - anyone who classified more than 2 galaxies more than 5 times each had all their classifications discarded
2. De-biasing
   - bias to classify far away spiral galaxies as elliptical galaxies (Bamford et al., 2009)

Galaxy Zoo

40 million classification to a consensus labels (Lintott et al., 2011)

1. Cleaning
   - only the first classification that a volunteer made of a specific galaxy was used in the analysis
   - anyone who classified more than 2 galaxies more than 5 times each had all their classifications discarded
2. De-biasing
   - bias to classify far away spiral galaxies as elliptical galaxies (Bamford et al., 2009)
3. Combining ($\sim$40 classifications per galaxy)
   - use classifier/classification matrix to upweight good classifiers

Produces data comparable in quality to expert coders (Lintott et al. 2011), but at much greater scale

Galaxy Zoo

From millions to billions to trillions . . . .

Galaxy Zoo

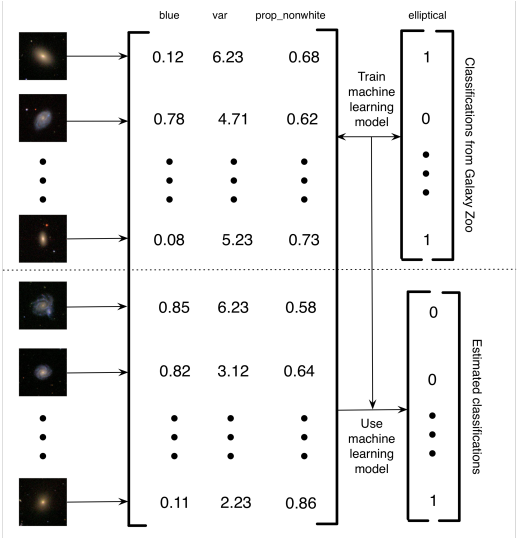


Fig 5.4 (Salganik 2018), inspired by Banerji et al. (2010)

https://www.zooniverse.org/

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*

Benoit et al. (2016)

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,

Benoit et al. (2016)

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

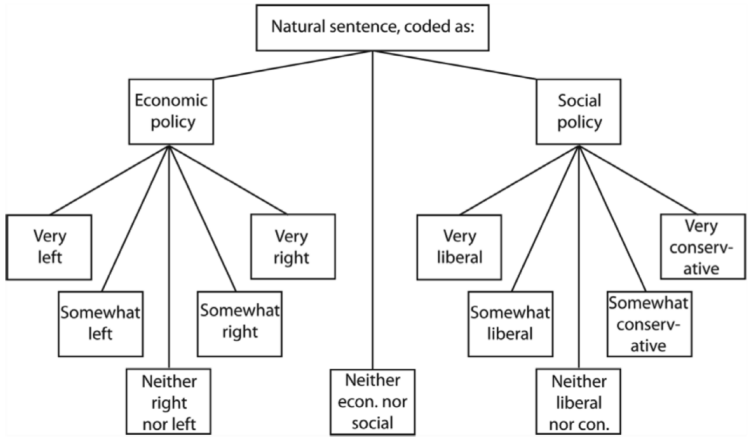
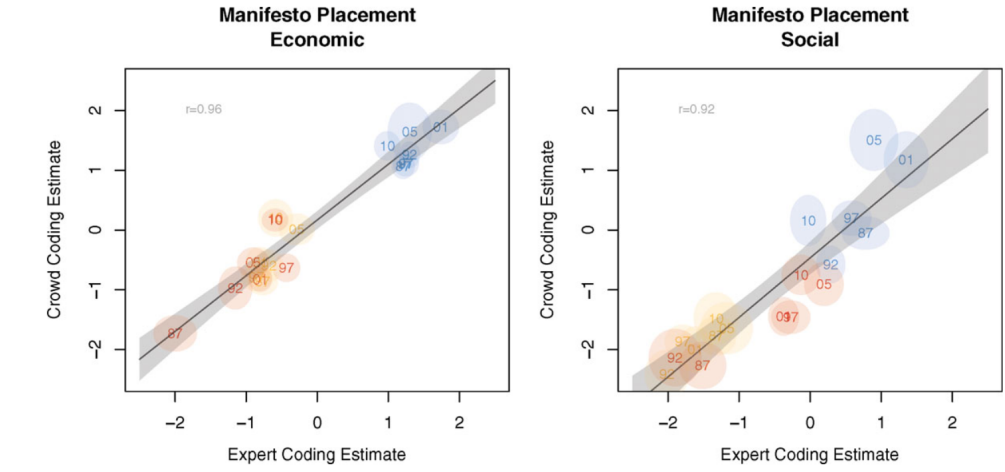


FIGURE 3. Expert and Crowd-sourced Estimates of Economic and Social Policy Positions



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

Wrapping up:

- Easy task, big scale problems where humans better than computers

Wrapping up:

- Easy task, big scale problems where humans better than computers
- Split-apply-combine strategy

Wrapping up:

- Easy task, big scale problems where humans better than computers
- Split-apply-combine strategy
- Human effort can be magnified with supervised learning

Wrapping up:

- Easy task, big scale problems where humans better than computers
- Split-apply-combine strategy
- Human effort can be magnified with supervised learning
- Increasingly important as we move from numeric survey data to working with text, images, movies, and audio.

What to read next:

- *Human computation* (Law and von Ahn, 2011)
- reCAPTCHA (von Ahn et al. 2008)
- Background about Amazon Mechanical Turk: Bohannon 2016

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

Matthew J. Salganik
Department of Sociology
Princeton University