Analysis using Mass Collaboration

Dr. Richard W. Evans

November 12, 14, 2018

Outline

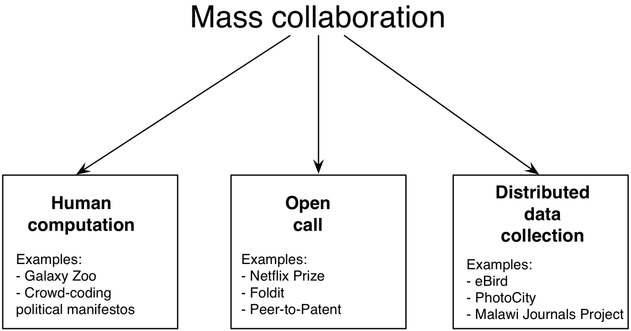
Data and solution collaboration

1. Human computation
2. Open calls
3. Distributed data collection

Research collaboration

4. Collaboration in research, theory, and coding

Mass collaboration examples


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

- Not many examples in social science
- Low hanging fruit available to be picked
- Collaborative research (GitHub) follows same principles

Human computation

Human computation great if ...

"I could solve this problem if I had 1,000 (unskilled) research assistants."

- Interesting dynamic between skilled and unskilled

Split-apply-combine

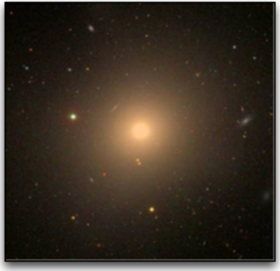
- Decompose big project into simple pieces
- Send many pieces to many workers
- Incorporate redundancy
- Combine carefully
- Computer-assisted human computation system

Galaxy Zoo: Origins

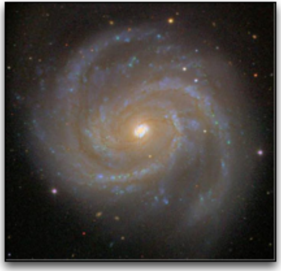
- Kevin Schawinski, Graduate student, Oxford, 2007
  - Graduate student opportunity cost of time is lower
  - Graduate students usually have more computational/tech skills than older full professors
  - Breakthroughs often come while discussing problem with colleague(s) at “pub”
  - Currently full professor at ETH Zurich (University of Zurich)

- Hypothesis: Galaxy color and shape had more complex relationship
  - Previous wisdom: spiral shape = blue color, elliptical shape = red color
  - Grad student solution: Classify 50,000 galaxies in 7 12-hour days

Galaxy shape and color



Elliptical galaxy



Spiral galaxy

- Problem: Sloan Digital Sky Survey had over 1 million

Galaxy Zoo: Importance of doing some yourself

Note the importance of this first stage

What is value of Schawinski classifying his own 50,000 galaxies?

- Sense of whether categories are complete
- Sense of how much time each task takes
- Sense of the skills needed to classify
- Better ability to train non-experts

Galaxy Zoo: Solution 1

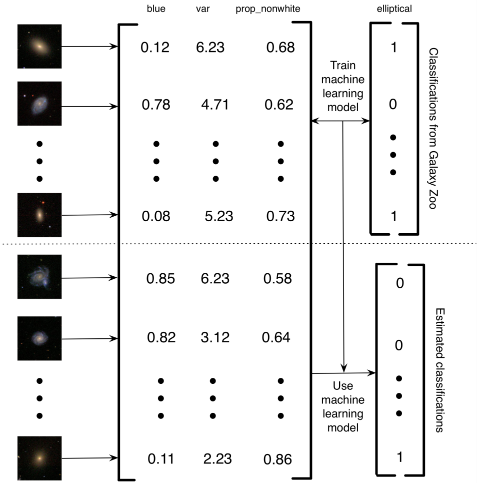
- Solution 1
  - While sitting in pub in Oxford with Chris Lintott
  - Website where volunteers classify galaxies (shape, color)
  - Website included initial 10-minute training
  - If volunteers got 11/15, proceed to classification
  - 100,000 volunteers, 40 million galaxy classifications (1 million galaxies)
  - Most classifications from a small group (fat head)
  - Few classifications from a large group (long tail)
  - Redundancy
    - How implement redundancy to ensure data quality?
    - What volunteers lacked in training, they made up for with redundancy
    - Remove “bogus” repeat classifiers
    - Adjust/remove systematic biases (show later)
    - Weighted consensus classification

Galaxy Zoo: Solution 2

- Solution 2:
  - Stage 1 classified 1 million galaxies
  - New digital sky surveys had 10 billion galaxies
  - Would require 10,000 times more participants

- Teach machine learning model to do what humans do (computer assisted human computation system)
  - Decompose each image into features (feature engineering)
  - Estimate model to map image features to human classifications
  - Use estimated model to predict other classifications (supervised learning)

Galaxy Zoo: Learning model



blue	var	prop_nonwhite	elliptical
0.12	6.23	0.68	1
0.78	4.71	0.62	0
⋮	⋮	⋮	⋮
0.08	5.23	0.73	1
0.85	6.23	0.58	0
0.82	3.12	0.64	0
⋮	⋮	⋮	⋮
0.11	2.23	0.86	1

Train machine learning model
Classifications from Galaxy Zoo
Use machine learning model
Estimated classifications

- Can handle “infinite amounts of data”

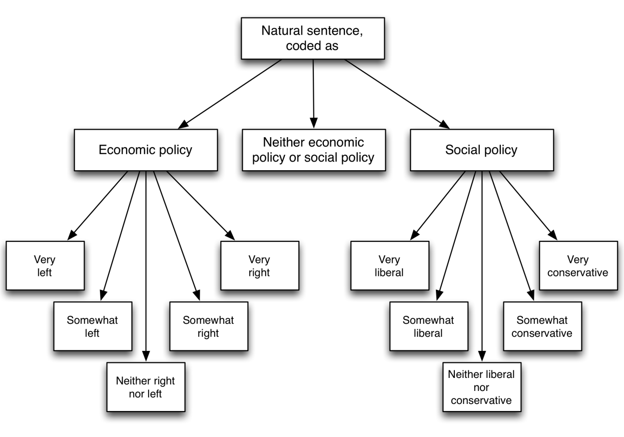
Classifying political manifestos

How would you classify this text?

> *"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."*

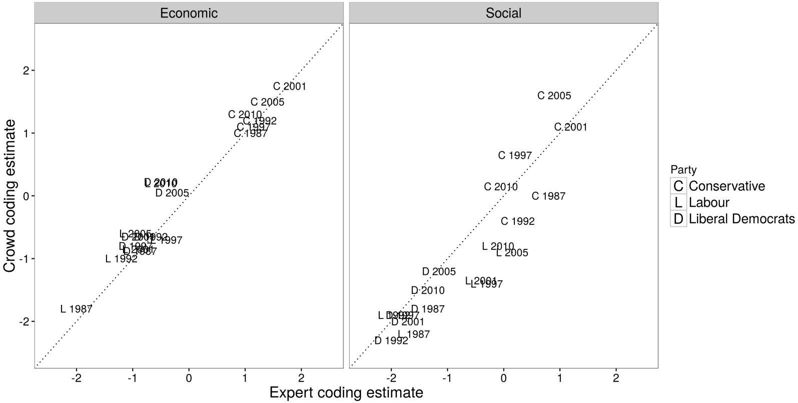
- The Manifesto Project
- 4,000 manifestos, 1,000 parties, 50 countries, each sentence, 56 categories
- Has resulted in over 200 papers

Manifesto Project Classification Scheme



- Use microlabor market (paid) to enlist volunteers

Crowd accuracy vs. experts and bias



- Can see crowd classification bias and how to adjust it

Crowd extensions are easier

- After initial study, wanted immigration classification
    - pro, neutral, anti
- 2010 UK general election
- After 5 hours, had 22,000 responses, $360
- Agreed with expert classifications

Human computation summary

- Best for:
  - Easy task, big scale
  - Tasks not easily solved by computers
  - Tasks can be done by non-experts (although experts OK)
  - Classification is not subjective
    - subjective: “Is this news story biased?”, “Is this bad policy?”
- Augmented and scaled by computer assisted human computation

Open calls

- Best for:
  - Clearly specified goal
  - Solutions are hard
  - Easier to check than to create solutions
  - Solutions may span large classes of models

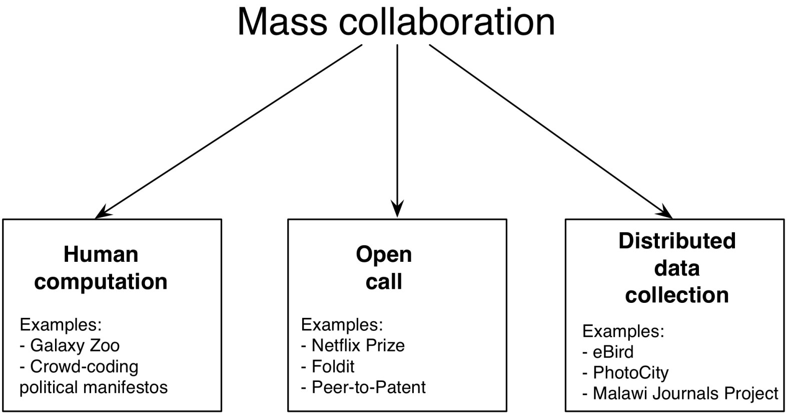
Open calls

- Researcher poses a problem
- Specific evaluation criterion
- Submission guidelines
- It is a horse race

Open calls

- Netflix Prize
- Project Euler
- Foldit
- Optimal tax policy
- Kaggle
- Assignment 6

Mass collaboration examples



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

Netflix prize

- 2000: Netflix launches *Cinematch* system
  - Initially worked poorly
  - Slowly improved at predicting customer preferences
  - Progress plateaued

- 2006: Netflix Prize, open call
  - Give dataset of 100 million ratings, 500,000 customers
  - Winner would be best predictor of 3 million held out ratings

How would you approach solving this problem?

What methods would you use to fit a predictive model on the 100 million ratings?

Netflix ratings data

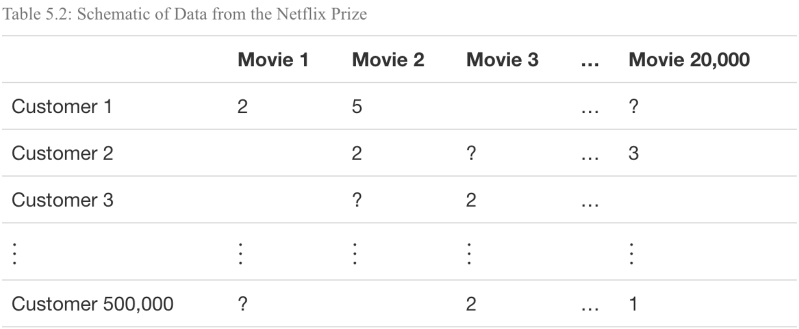


Table 5.2: Schematic of Data from the Netflix Prize

	Movie 1	**Movie 2**	**Movie 3**	**…**	**Movie 20,000**
Customer 1	2	5		…	?
Customer 2		2	?	…	3
Customer 3		?	2	…	
⋮	⋮	⋮	⋮		⋮
Customer 500,000	?		2	…	1

- NBC video gives overview
- AT&T video describes the winners

Netflix prize: Napoleon Dynamite

Hard question

- How would you predict Napoleon Dynamite rating?
- Are there any drawbacks to averaging 800 models?

Netflix prize: cool characteristics

- Submissions came from many groups, multiple fields
- Interactive leaderboard over the period
- Motivation was more than money
  - Simon Funk blog post on using singular value decomposition (SVD)
- Easy to evaluate

Netflix Prize: Simon Funk contribution

“...at one point during the Netflix Prize, someone with the screen name Simon Funk posted on his blog a proposed solution based on a singular value decomposition, an approach from linear algebra that had not been used previously by other participants. Funk’s blog post was simultaneously technical and weirdly informal. Was this blog post describing a good solution or was it a waste of time? Outside of an open call project, the solution might never have received serious evaluation. After all, Simon Funk was not a professor at MIT; he was a software developer who, at the time, was backpacking around New Zealand (Piatetsky 2007). If he had emailed this idea to an engineer at Netflix, it almost certainly would not have been read.”

“Fortunately, because the evaluation criteria were clear and easy to apply, his predicted ratings were evaluated, and it was instantly clear that his approach was very powerful: he rocketed to fourth place in the competition, a tremendous result given that other teams had already been working for months on the problem. In the end, parts of his approach were used by virtually all serious competitors (Bell, Koren, and Volinsky 2010).”

Project Euler: non monetary motivation

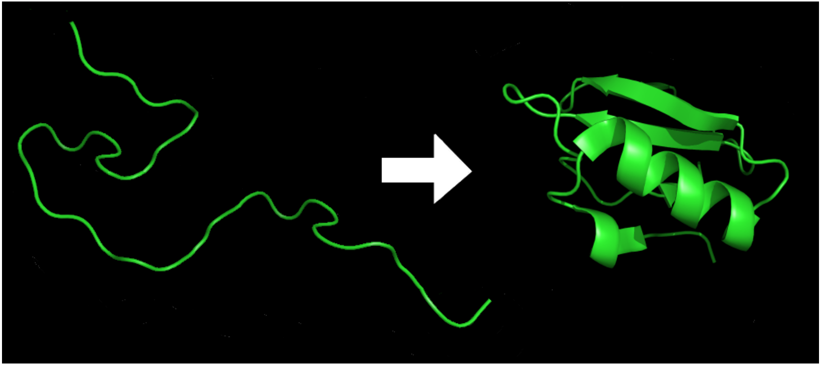
- Project Euler, math problems
- Incentivizes good solutions
- Gives awards for different achievements
- Allows for competition

Could and couldn't

- Could not be used for open call on math problem
- Could be used for open call on solution method improvement

Foldit: Protein folding

- Foldit, protein folding
  - Process where chain of amino acids takes on its shape
  - Tends toward its lowest energy configuration
  - Billions of permutations of how this could happen



Foldit: cool characteristics

- Game can be played by anyone
- Playing game changes energy state of protein
- Lower energy results in higher score
- Lowest energy state is predicted fold pattern

Optimal Tax Policy

Evans, Richard W., Kenneth Judd, and Kramer Quist, “Big Data Techniques as a Solution to Theory Problems,” in *Conquering Big Data with High Performance Computing*, ed. Ritu Arora, Springer (2016)

- HARD:
  - Nonconvex optimization
  - high dimensional control
  - multiobjective programming

TaxBrain, open call game

Can we use TaxBrain to make an open call game of tax policy?

- What would be the choices?
- What would be the criteria?

Kaggle: open call platform

- Kaggle
- Look through Competitions
- Look at completed competitions
- Different topics, different rewards

Open calls summary

- Propose specific question
- Broad solicitation of solutions
- Hard to know what is best method
- Solutions easy to verify
- Pick the best solution

Problem...

Not widely used in social science research

- Open call better for prediction
- Watts (2014) connect explanation (theory) with prediction

Distributed data collection

- Examples: eBird, PhotoCity
- Capture work that is already happening
- Enlist volunteers as data collectors
- Enlist work on a scale otherwise impossible
- Difficult to ensure quality and sampling

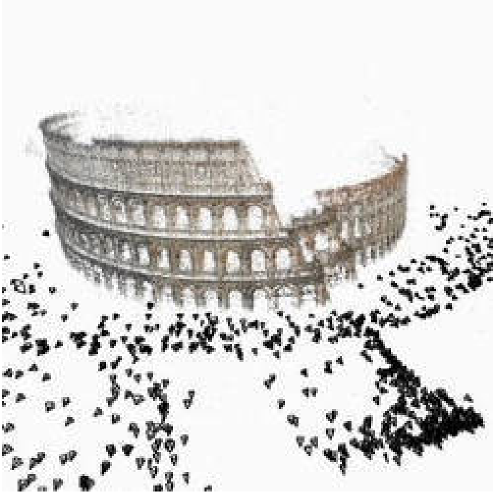
eBird

- Dream: track every bird in real time
- eBird: Birders upload photos of birds
  - Who, where, when, what species, how many, effort (methods)
  - Play Bard Swallow abundance visualization
- eBird researchers use statistical models to correct for noisy, heterogeneous data
- Are eBird data perfect? No.
- For certain questions, is eBird data better than existing ornithology data? Definitely yes!

PhotoCity

- 3D models from 2D images
- Crowd source the image selection
- System protected itself against bad data
    - Reject photos that don’t connect to others
    - Give points to images that contribute the most

PhotoCity and the Coliseum



Mass collaboration summary

1. Motivate participants
   - money, knowledge, community
2. Leverage heterogeneity
   - skill, effort, available time
   - fat head, long tail
3. Focus attention
4. Enable surprise
   - Green pee galaxies, SVD Netflix solution
5. Be ethical
   - Don’t expose people’s info
   - Attribution options
   - textbfYour crowd is made up of people