[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

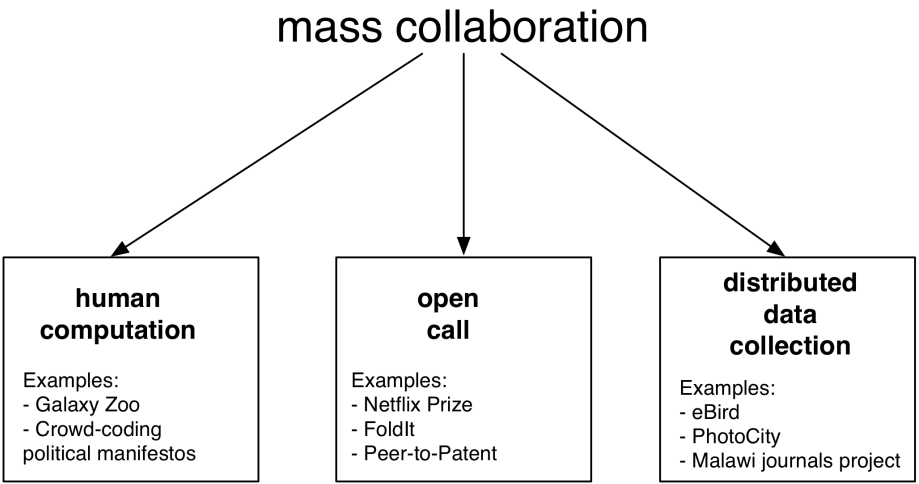


Fig 5.4 (Salganik 2018)

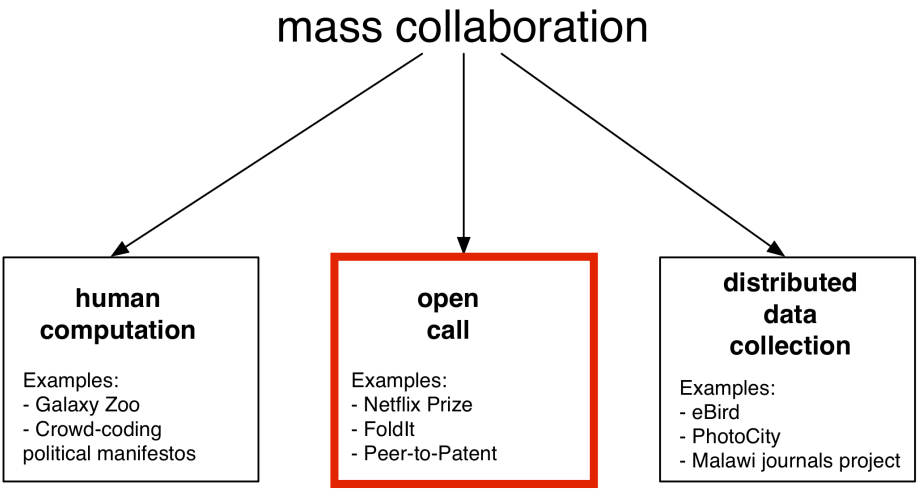


Fig 5.4 (Salganik 2018)

Open call:

- Problems where solutions are easier to check than generate

Open call:

- Problems where solutions are easier to check than generate
- Enables easy and fair evaluation

Open call:

- Problems where solutions are easier to check than generate
- Enables easy and fair evaluation
- Taking the best submission (not a combination of submissions)

Open call:

- Problems where solutions are easier to check than generate
- Enables easy and fair evaluation
- Taking the best submission (not a combination of submissions)
- Participants require specialized skills


NETFLIX
Netflix Prize
Home
Rules
Leaderboard
Register
Update
Submit
Download
NETFLIX
Browse
Recommendations
Friends
Queue
Buy DVDs
Movies For You
You really liked it...
Welcome!
The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.
Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.
You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.
Good luck and thanks for helping!
FAQ | Forum | Netflix Home
© 1997-2009 Netflix, Inc. All rights reserved.

Netflix Prize

- Released a **training set** of ∼100 million ratings
- Held-back a **test set** of ∼1.5 million ratings

	Movie 1	Movie 2	Movie 3	...	Movie 20,000
User 1	2	5			?
User 2		?	2		3
User 3					4
⋮					
User 500,000	?		2		1

Goal is clear, but path is not clear.

Netflix Prize

- Open it up to the world and offer a prize of $1,000,000
- 44,014 valid submissions from 5,169 different teams
- Fortunately, they were easy to check

$$RMSE = \sqrt{\frac{\sum_i \sum_j (\widehat{r_{ij}} - r_{ij})^2}{n}}$$

Netflix Prize

Netflix Update: Try This at Home

Netflix Update: Try This at Home

sifter.org/~simon/journal/20061211.html

[<< | Prev | Index | Next | >>]

Monday, December 11, 2006

Netflix Update: Try This at Home

[Followup to this]

Ok, so here's where I tell all about how I (now we) got to be tied for third place on the netflix prize. And I don't mean a sordid tale of computing in the jungle, but rather the actual math and methods. So yes, after reading this post, you too should be able to rank in the top ten or so.

Ur... yesterday's top ten anyway.

My first disclaimer is that our last submission which tied for third place was only actually good enough for ninth place or so. It landed where it did because, just for giggles and grins, we blended results (50/50) with Jetrays who had a similar score to us at the time.

Second, my friend Vincent has been manning the runs on his desktop machines, diligently fine tuning and squeezing out every last bit of performance possible with whatever controls I could give him (not to mention learning python so he could write scripts to blend submissions and whatnot). In short, almost all my progress since my last post has been due to other people. In the meantime I've implemented a handful of failed attempts at improving the performance, plus one or two minorly successful ones which I'll get to.

Netflix Prize

So, in other words, if we take the rank-40 singular value decomposition of the 8.5B matrix, we have the best (least error) approximation we can within the limits of our user-movie-rating model. I.e., the SVD has found our "best" generalizations for us. Pretty neat, eh?

Only problem is, we don't have 8.5B entries, we have 100M entries and 8.4B empty cells. Ok, there's another problem too, which is that computing the SVD of ginormous matrices is... well, no fun. Unless you're into that sort of thing.

But, just because there are five hundred really complicated ways of computing singular value decompositions in the literature doesn't mean there isn't a really simple way too: Just take the derivative of the approximation error and follow it. This has the added bonus that we can choose to simply ignore the unknown error on the 8.4B empty slots.

So, yeah, you mathy guys are rolling your eyes right now as it dawns on you how short the path was.

Sifter aka Simon Funk

Netflix Prize

So, in other words, if we take the rank-40 singular value decomposition of the 8.5B matrix, we have the best (least error) approximation we can within the limits of our user-movie-rating model. I.e., the SVD has found our "best" generalizations for us. Pretty neat, eh?

Only problem is, we don't have 8.5B entries, we have 100M entries and 8.4B empty cells. Ok, there's another problem too, which is that computing the SVD of ginormous matrices is... well, no fun. Unless you're into that sort of thing.

But, just because there are five hundred really complicated ways of computing singular value decompositions in the literature doesn't mean there isn't a really simple way too: Just take the derivative of the approximation error and follow it. This has the added bonus that we can choose to simply ignore the unknown error on the 8.4B empty slots.

So, yeah, you mathy guys are rolling your eyes right now as it dawns on you how short the path was.

Sifter aka Simon Funk

Moved him instantly into fourth place, and was later used by all serious competitors.

Foldit

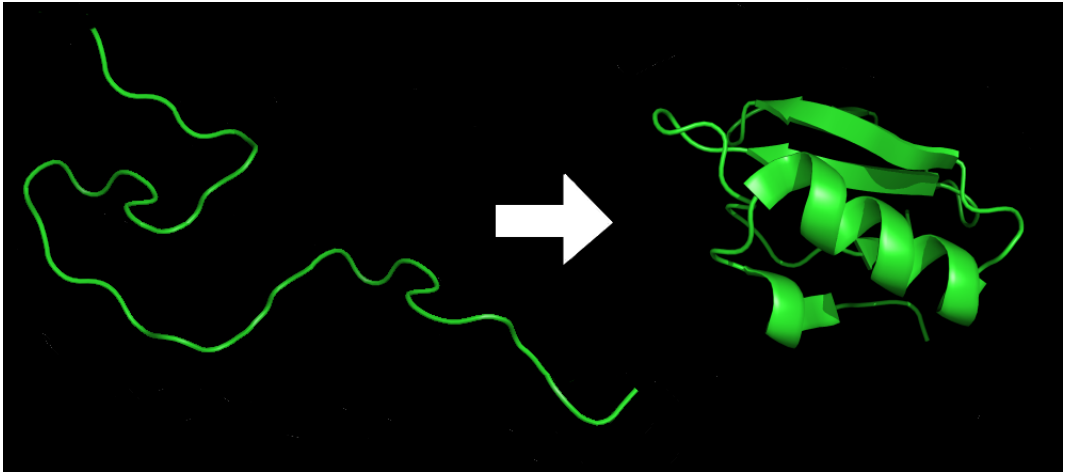
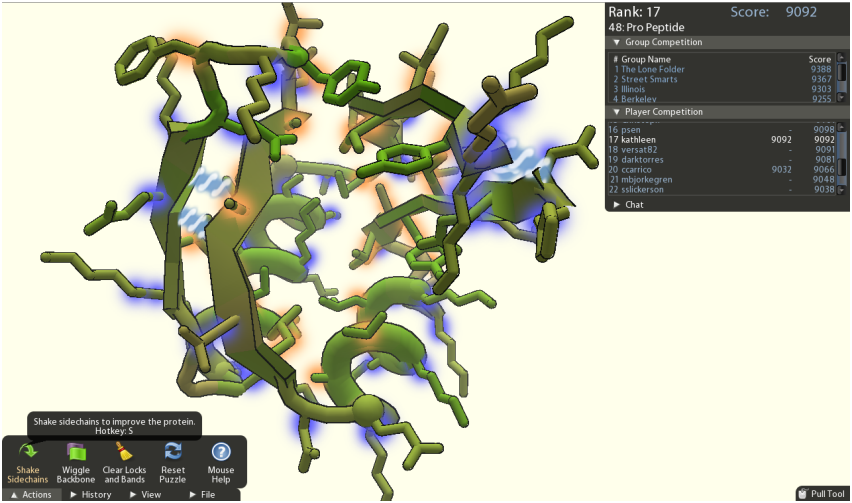
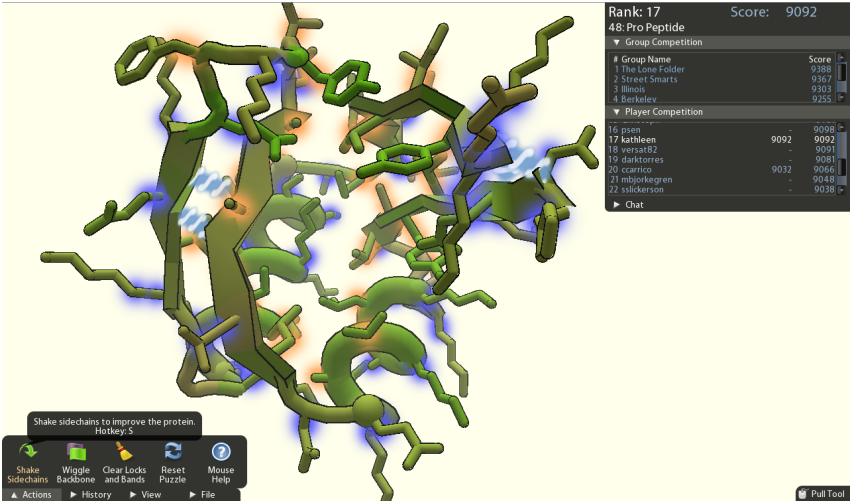


Fig 5.7 (Salganik 2018)

FoldIt


Rank: 17
Score: 9092
48: Pro Peptide
Group Competition
Group Name Score
1 The Lone Folder 9388
2 Street Smarts 9367
3 Illinois 9303
4 Berkelev 9255
Player Competition
16 psen - 9098
17 kathleen 9092 9092
18 versat82 - 9091
19 darktorres - 9081
20 ccarrico 9032 9066
21 mbjorkegren - 9048
22 sslickerson - 9038
Chat
Shake sidechains to improve the protein.
Hotkey: S
Shake Sidechains
Wiggle Backbone
Clear Locks and Bands
Reset Puzzle
Mouse Help
Actions
History
View
File
Pull Tool

FoldIt



Gamers outperformed best known computational algorithms on 5 out of 10 proteins of unknown structure (Cooper et al., 2010)

Fragile Families Challenge

Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew J. Salganik[a,1], Ian Lundberg[a], Alexander T. Kindel[a], Caitlin E. Ahearn[b], Khaled Al-Ghoneim[c], Abdullah Almaatouq[d,e], Drew M. Altschul[f], Jennie E. Brand[b,g], Nicole Bohme Carnegie[h], Ryan James Compton[i], Debanjan Datta[j], Thomas Davidson[k], Anna Filippova[l], Connor Gilroy[m], Brian J. Goode[n], Eaman Jahani[o], Ridhi Kashyap[p,q,r], Antje Kirchner[s], Stephen McKay[t], Allison C. Morgan[u], Alex Pentland[e], Kivan Polimis[v], Louis Raes[w], Daniel E. Rigobon[x], Claudia V. Roberts[y], Diana M. Stanescu[z], Yoshihiko Suhara[e], Adaner Usmani[aa], Erik H. Wang[z], Muna Adem[bb], Abdulla Alhajri[cc], Bedoor AlShebli[dd], Redwane Amin[ee], Ryan B. Amos[y], Lisa P. Argyle[ff], Livia Baer-Bositis[gg], Moritz Büchi[hh], Bo-Ryehn Chung[ii], William Eggert[jj], Gregory Faletto[kk], Zhilin Fan[ll], Jeremy Freese[gg], Tejomay Gadgil[mm], Josh Gagné[gg], Yue Gao[nn], Andrew Halpern-Manners[bb], Sonia P. Hashim[y], Sonia Hausen[gg], Guanhua He[oo], Kimberly Higuera[gg], Bernie Hogan[pp], Ilana M. Horwitz[qq], Lisa M. Hummel[gg], Naman Jain[x], Kun Jin[rr], David Jurgens[ss], Patrick Kaminski[bb,tt], Areg Karapetyan[uu,vv], E. H. Kim[gg], Ben Leizman[y], Naijia Liu[z], Malte Möser[y], Andrew E. Mack[z], Mayank Mahajan[y], Noah Mandell[ww], Helge Marahrens[bb], Diana Mercado-Garcia[qq], Viola Mocz[xx], Katariina Mueller-Gastell[gg], Ahmed Musse[yy], Qiankun Niu[ee], William Nowak[zz], Hamidreza Omidvar[aaa], Andrew Or[y], Karen Ouyang[y], Katy M. Pinto[bbb], Ethan Porter[ccc], Kristin E. Porter[ddd], Crystal Qian[y], Tamkinat Rauf[gg], Anahit Sargsyan[eee], Thomas Schaffner[y], Landon Schnabel[gg], Bryan Schonfeld[z], Ben Sender[fff], Jonathan D. Tang[y], Emma Tsurkov[gg], Austin van Loon[gg], Onur Varol[ggg,hhh], Xiafei Wang[iii], Zhi Wang[hhh,jjj], Julia Wang[y], Flora Wang[fff], Samantha Weissman[y], Kirstie Whitaker[kkk,lll], Maria K. Wolters[mmm], Wei Lee Woon[nnn], James Wu[ooo], Catherine Wu[y], Kengran Yang[aaa], Jingwen Yin[ll], Bingyu Zhao[ppp], Chenyun Zhu[ll], Jeanne Brooks-Gunn[qqq,rrr], Barbara E. Engelhardt[y,iii], Moritz Hardt[sss], Dean Knox[z], Karen Levy[ttt], Arvind Narayanan[y], Brandon M. Stewart[a], Duncan J. Watts[uuu,vvv,www], and Sara McLanahan[a,1]

(Salganik et al., 2020)

Wrapping-up:

- Problems where solutions are easier to check than generate

Wrapping-up:

- Problems where solutions are easier to check than generate
- Enables easy and fair evaluation

Wrapping-up:

- Problems where solutions are easier to check than generate
- Enables easy and fair evaluation
- Taking the best submission (not a combination of submissions)

Wrapping-up:

- Problems where solutions are easier to check than generate
- Enables easy and fair evaluation
- Taking the best submission (not a combination of submissions)
- Participants require specialized skills

What to read next:

- *Longitude*, Sobel (1996)
- "Statistical Significance of the Netflix Challenge", Feurverger et al. (2012)

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

Matthew J. Salganik
Department of Sociology
Princeton University