Measuring the predictability of life outcomes with a scientific mass collaboration

Ian Lundberg
Princeton University

Berkeley Good Data Seminar
12 May 2020

This research is supported by the Russell Sage Foundation. We are grateful to the members of the Board of Advisors of the Fragile Families Challenge: Jeanne Brooks-Gunn, Kathryn J. Edin, Barbara E. Engelhardt, Irwin Garfinkel, Moritz Hardt, Dean Knox, Nicholas Lemann, Karen Levy, Sara McLanahan, Arvind Narayanan, Timothy J. Nelson, Matthew Salganik, Brandon Stewart, & Duncan Watts.
Replication materials are on Dataverse: https://doi.org/10.7910/DVN/CXSECU.
Source for these slides: www.github.com/fragilefamilieschallenge.

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,ii], 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]

This research is supported by the Russell Sage Foundation. We are grateful to the members of the Board of Advisors of the Fragile Families Challenge: Jeanne Brooks-Gunn, Kathryn J. Edin, Barbara E. Engelhardt, Irwin Garfinkel, Moritz Hardt, Dean Knox, Nicholas Lemann, Karen Levy, Sara McLanahan, Arvind Narayanan, Timothy J. Nelson, Matthew Salganik, Brandon Stewart, & Duncan Watts.
Replication materials are on Dataverse: https://doi.org/10.7910/DVN/CXSECU.
Source for these slides: www.github.com/fragilefamilieschallenge.

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,ii], 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]

This research is supported by the Russell Sage Foundation. We are grateful to the members of the Board of Advisors of the Fragile Families Challenge: Jeanne Brooks-Gunn, Kathryn J. Edin, Barbara E. Engelhardt, Irwin Garfinkel, Moritz Hardt, Dean Knox, Nicholas Lemann, Karen Levy, Sara McLanahan, Arvind Narayanan, Timothy J. Nelson, Matthew Salganik, Brandon Stewart, & Duncan Watts.
Replication materials are on Dataverse: https://doi.org/10.7910/DVN/CXSECU.
Source for these slides: www.github.com/fragilefamilieschallenge.

Life course

Early
experiences

Later
outcomes

Life course

Early experiences → Later outcomes

Standard social science practice

– Describe social patterns
– Theorize important factors
– Estimate causal effects

Life course

Early experiences → Later outcomes

Standard social science practice

– Describe social patterns
– Theorize important factors
– Estimate causal effects

Open questions

Life course

Early experiences → Later outcomes

Standard social science practice

- Describe social patterns
- Theorize important factors
- Estimate causal effects

Open questions

- Can we predict individual life outcomes?

Life course

Early experiences → Later outcomes

Standard social science practice

- Describe social patterns
- Theorize important factors
- Estimate causal effects

Open questions

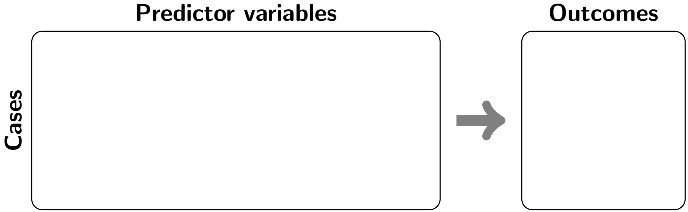
- Can we predict individual life outcomes?
- Credibility

Standard social science practice

- Describe social patterns
- Theorize important factors
- Estimate causal effects

Open questions

- Can we predict individual life outcomes?
- Credibility

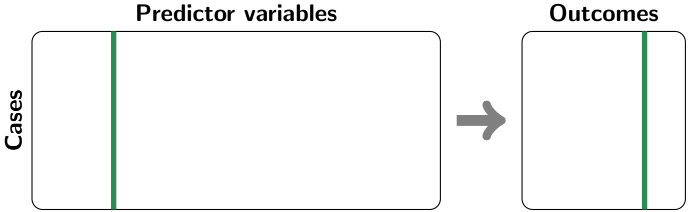


Standard social science practice

- Describe social patterns
- Theorize important factors
- Estimate causal effects

Open questions

- Can we predict individual life outcomes?
- Credibility

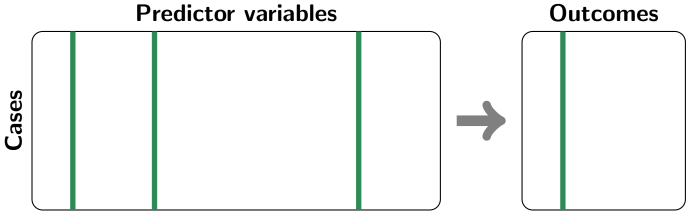


Standard social science practice

- Describe social patterns
- Theorize important factors
- Estimate causal effects

Open questions

- Can we predict individual life outcomes?
- Credibility



Life course

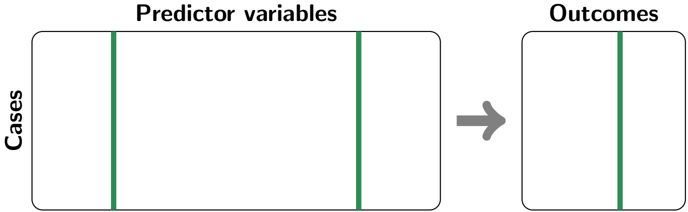
Early experiences → Later outcomes

Standard social science practice

- Describe social patterns
- Theorize important factors
- Estimate causal effects

Open questions

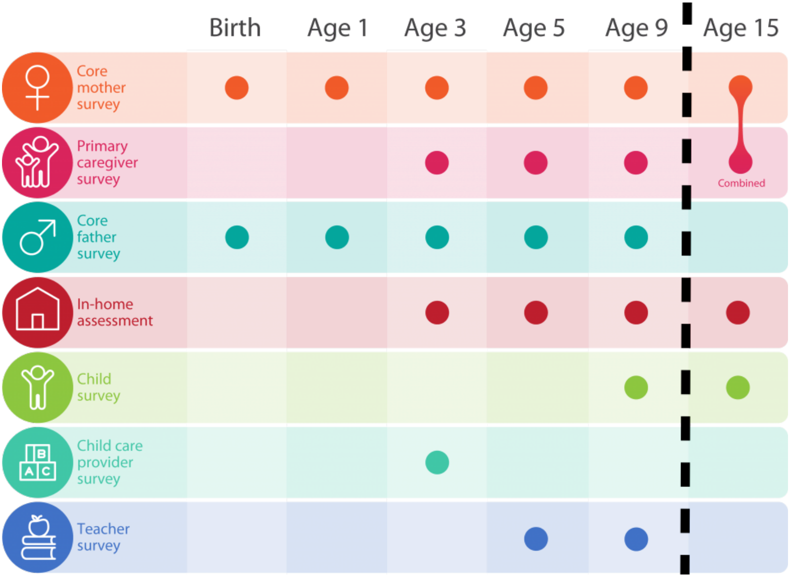
- Can we predict individual life outcomes?
- Credibility



Fragile Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

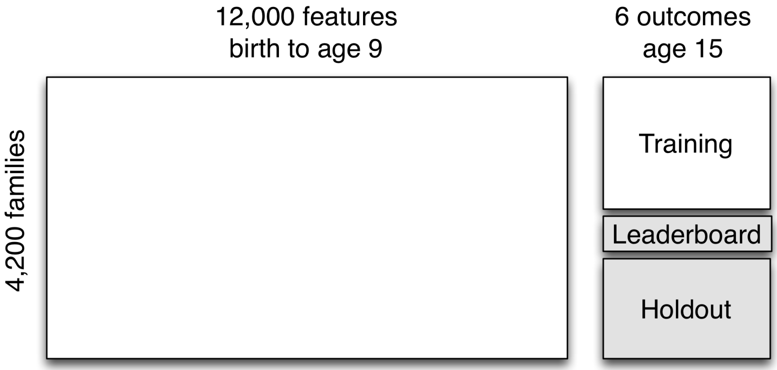
- Birth cohort panel study
- $\approx$ 5,000 children born in 20 U.S. cities
- Followed from birth through age 15

	Birth	Age 1	Age 3	Age 5	Age 9
Core mother survey	●	●	●	●	●
Primary caregiver survey			●	●	●
Core father survey	●	●	●	●	●
In-home assessment			●	●	●
Child survey					●
Child care provider survey			●		
Teacher survey				●	●


Birth
Age 1
Age 3
Age 5
Age 9
Age 15
Core mother survey
Primary caregiver survey
Combined
Core father survey
In-home assessment
Child survey
Child care provider survey
Teacher survey

Six age 15 outcomes:

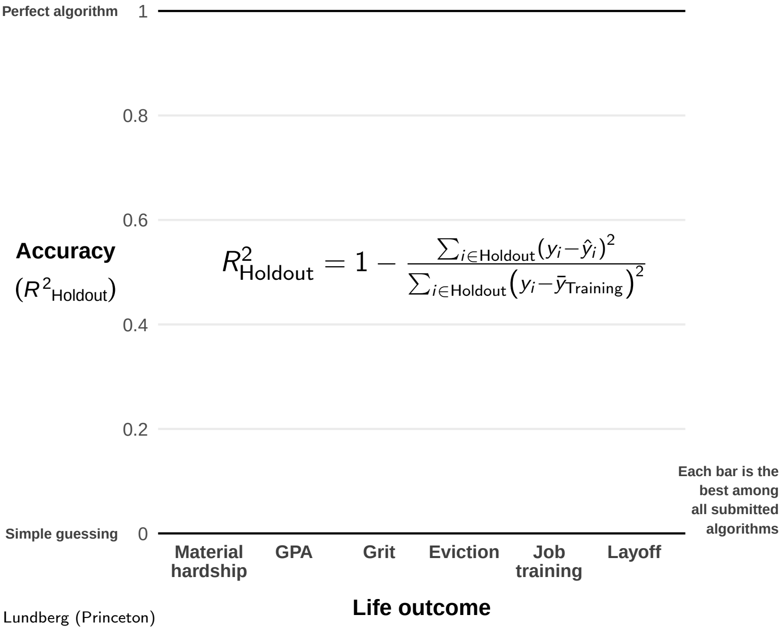
- GPA
- Material Hardship
- Grit
- Evicted
- Job training
- Job loss


12,000 features
birth to age 9
6 outcomes
age 15
4,200 families
Training
Leaderboard
Holdout

441 registered participants

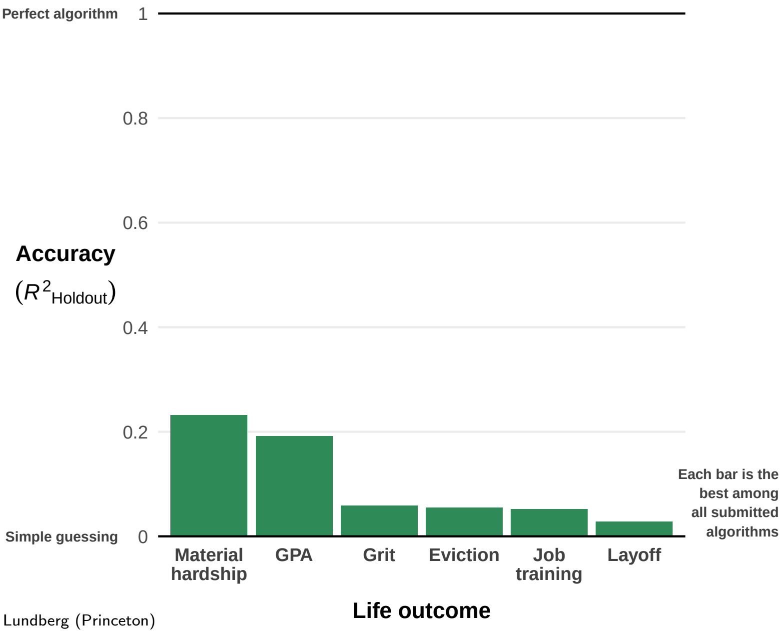
- social scientists and data scientists
- undergraduates, grad students, and professionals
- many working in teams

How did they do?

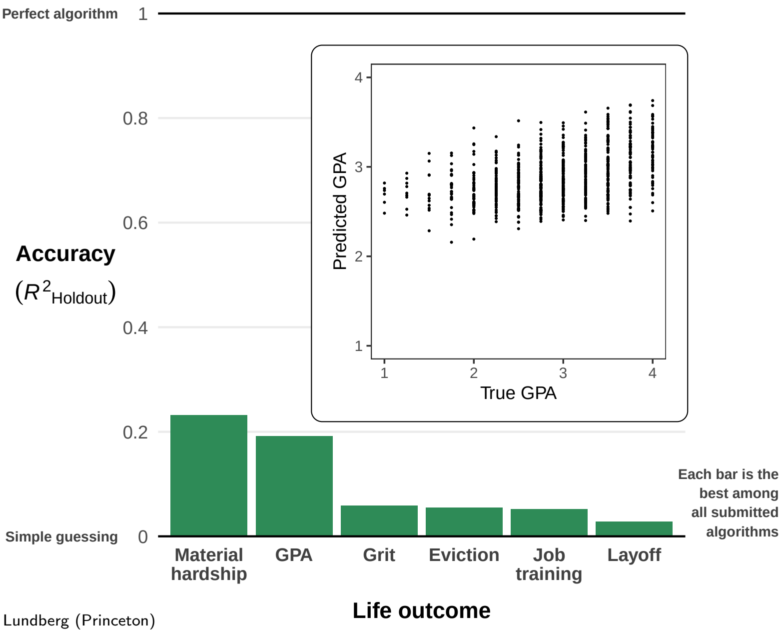


$$R^2_{\text{Holdout}} = 1 - \frac{\sum_{i \in \text{Holdout}} (y_i - \hat{y}_i)^2}{\sum_{i \in \text{Holdout}} \left(y_i - \bar{y}_{\text{Training}}\right)^2}$$

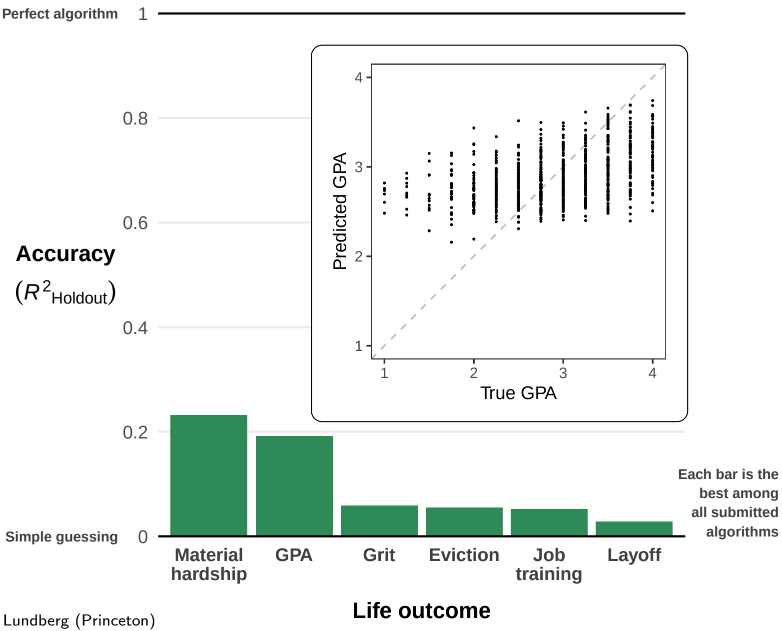
Best algorithms were not very accurate



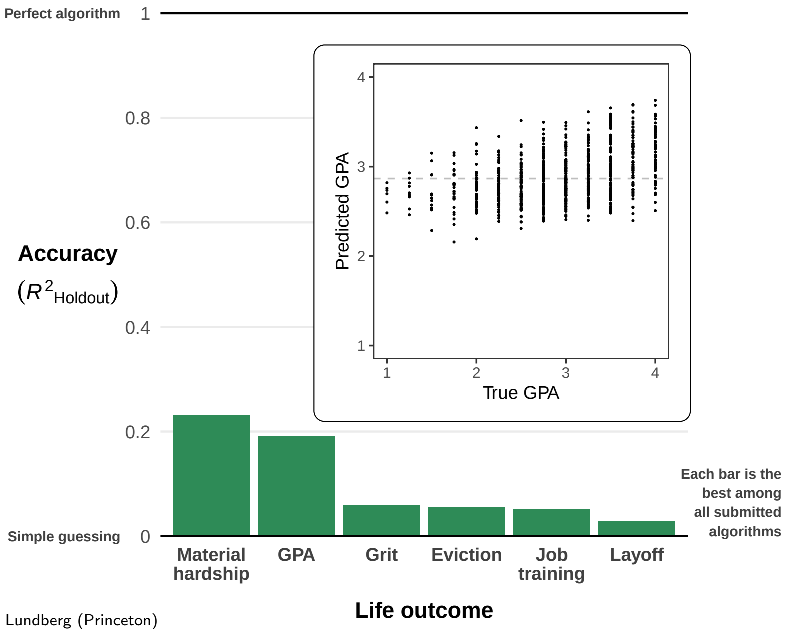
Best algorithms were not very accurate



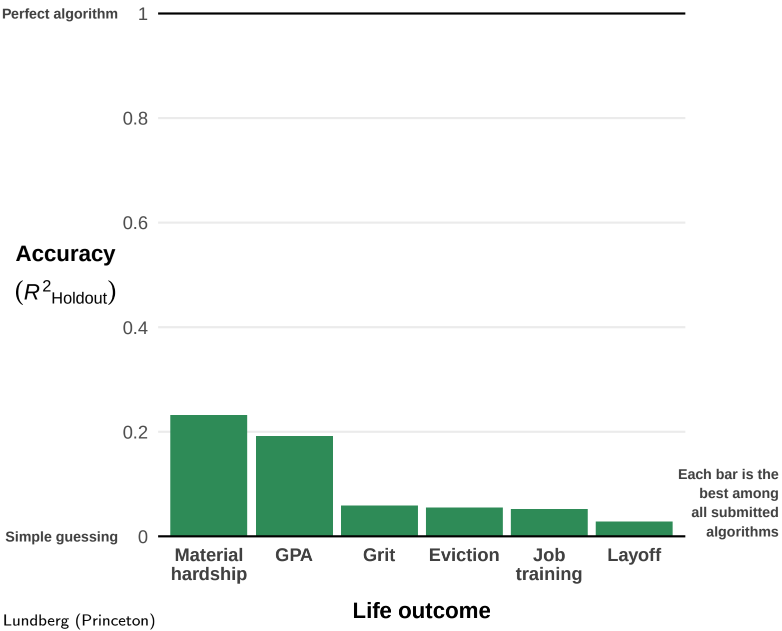
Best algorithms were not very accurate



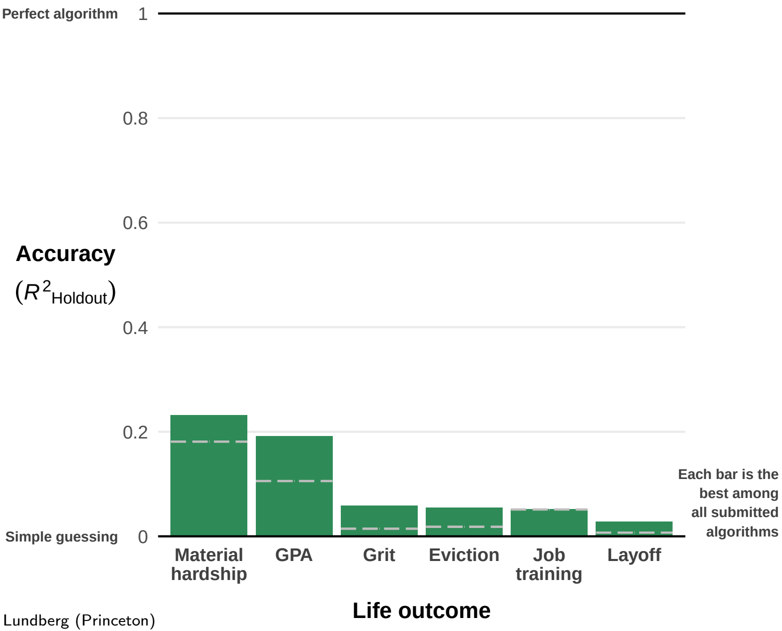
Best algorithms were not very accurate



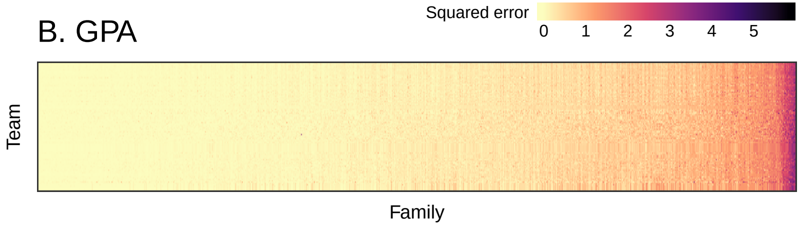
Best algorithms were not very accurate



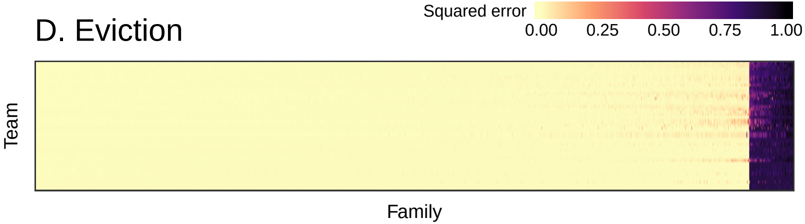
Best algorithms were not very accurate



B. GPA



D. Eviction



What do these results mean for policy and for science?

For **policymakers**,

For **policymakers**,

– Do not assume that predictive algorithms are accurate

For **policymakers**,

– Do not assume that predictive algorithms are accurate

– Transparent evaluation of any algorithm is needed

For **policymakers**,

– Do not assume that predictive algorithms are accurate

– Transparent evaluation of any algorithm is needed

– Complex models may not outperform simple models

For **scientists**: A paradox of prediction and understanding

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

...yet the very same data did not yield accurate predictions

↗

(our result)

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

...yet the very same data did not yield accurate predictions

↗

(our result)

Resolutions:

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

...yet the very same data did not yield accurate predictions

↗

(our result)

Resolutions:

1. If understanding implies an ability to predict, then **we do not understand** the life course.

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

...yet the very same data did not yield accurate predictions

↗

(our result)

Resolutions:

1. If understanding implies an ability to predict, then **we do not understand** the life course.
2. Perhaps we have **understanding despite poor prediction**.
   - Aggregate description
   - Causal effects

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

...yet the very same data did not yield accurate predictions

↗

(our result)

Resolutions:

1. If understanding implies an ability to predict, then **we do not understand** the life course.
2. Perhaps we have **understanding despite poor prediction**.
   - Aggregate description
   - Causal effects
3. Our understanding is correct but **incomplete**. We need theories that point toward poor prediction

For **scientists**: The value of mass collaboration

For **scientists**: The value of mass collaboration

A credible estimate of predictability

For **scientists**: The value of mass collaboration

A credible estimate of predictability

The common task framework does not end
with choosing the winner

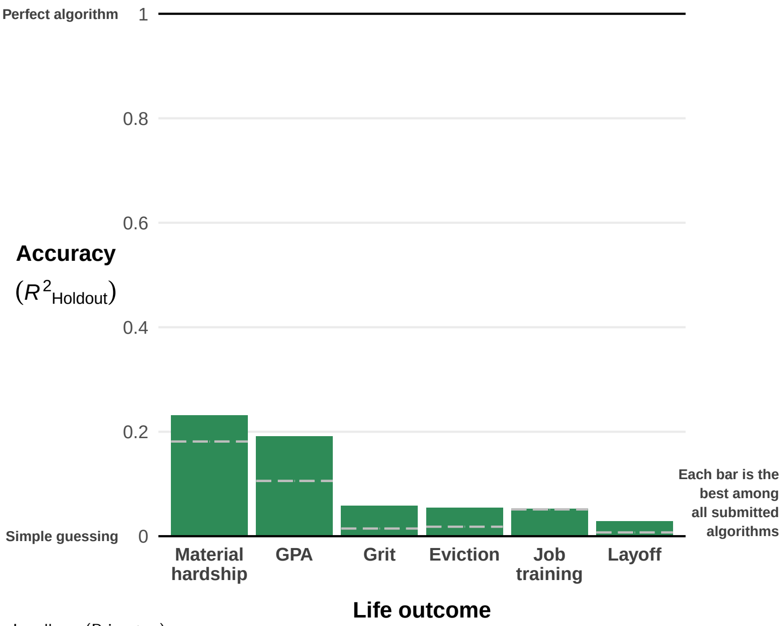
For **scientists**: The value of mass collaboration

A credible estimate of predictability

The common task framework does not end with choosing the winner

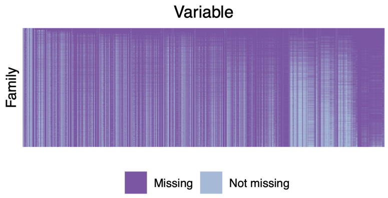
There may be other problems we can solve better collectively than individually

Best algorithms were not very accurate

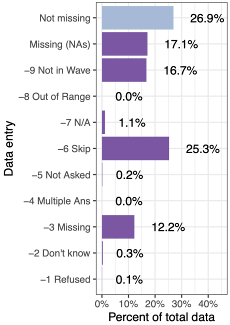


Appendix

Appendix



(a)



(b)

Fig. S1. Missing entries in the Fragile Families Challenge background dataset. The Fragile Families Challenge background dataset had 4,242 rows and 12,942 columns (plus an ID number). Of the approximately 55 million distinct data entries, about 73% were missing. There were many types of missing entries.

Appendix

Age 15 outcome	Age 9 questions	Response values	Reporter	How aggregated
GPA	At the most recent grading period, what was your grade in 1. English or language arts? 2. Math? 3. History or social studies? 4. Science?	1. A 2. B 3. C 4. D or lower	Child	Reverse-coded and averaged. Marked NA if any item missing due to no grade, pass/fail, refusal, don't know, homeschooled, or not interviewed.
Grit	Thinking about how you have behaved or felt during the past four weeks, please tell me whether you strongly agree, somewhat agree, somewhat disagree, or strongly disagree with the following statements. 1. I keep at my schoolwork until I am done with it. 2. Once I make a plan to get something done, I stick to it. 3. I finish whatever I begin. 4. I am a hard worker.	1. Strongly agree 2. Somewhat agree 3. Somewhat disagree 4. Strongly disagree	Child	Reverse-coded and averaged. Marked NA if any item missing due to refusal, don't know, or not interviewed.
Material hardship	We are also interested in some of the problems families have making ends meet. In the past twelve months, did you do any of the following because there wasn't enough money? 1. Did you receive free food or meals? 2. Were you ever hungry, but didn't eat because you couldn't afford enough food? 3. Did you ever not pay the full amount of rent or mortgage payments? 4. Were you evicted from your home or apartment for not paying the rent or mortgage? 5. Did you not pay the full amount of gas, oil, or electricity bill 6. Was your gas or electric services ever turned off, or the heating oil company did not deliver oil, because there wasn't enough money to pay the bills? 7. Did you borrow money from friends or family to help pay bills? 8. Did you move in with other people even for a little while because of financial problems? 9. Did you stay at a shelter, in an abandoned building, an automobile or any other place not meant for regular housing, even for one night? 10. Was there anyone in your household who needed to see a doctor or go to the hospital but couldn't go because of the cost? 11. Was your telephone service (mobile or land line) cancelled or disconnected by the telephone company because there wasn't enough money to pay the bill?	0. Event did not occur 1. Event occurred	Child's primary caregiver	Averaged. Marked NA if any response missing due to refusal, don't know, or not interviewed.
Eviction	1. In the past twelve months, were you evicted from your home or apartment for not paying the rent or mortgage? 2. (If no above:) Since [month and year of interview at approximately child age 9], were you evicted from your home or apartment for not paying the rent or mortgage?	0. No 1. Yes	Child's primary caregiver	If no to both questions, 0. If yes to either question, 1. Marked NA if missing due to refusal, don't know, or not interviewed.
Layoff	Since [month and year of interview at approximately child age 9], have you been laid off from your employer for any time?	0. No 1. Yes	Child's primary caregiver	Marked NA if missing due to refusal, don't know, or not interviewed.
Job training	Since [month and year of interview at approximately child age 9], have you taken any classes to improve your job skills, such as computer training or literacy classes?	0. No 1. Yes	Child's primary caregiver	Marked NA if missing due to refusal, don't know, or not interviewed.

Table S3. Outcome variables measured at child age 15.

Appendix

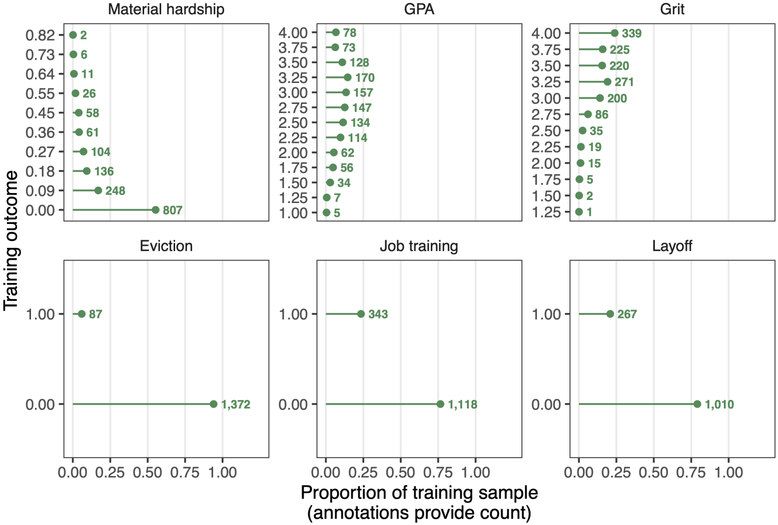


Fig. S2. Distribution of outcomes in the training set. The number of missing cases for each outcome varied (Table S4) and are excluded here.

Appendix

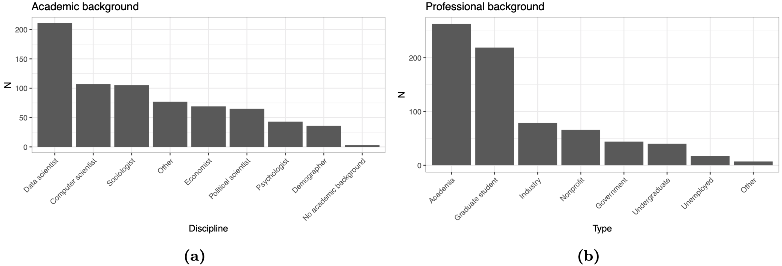


Fig. S3. Self-reported disciplinary affiliation and professional background of applicants. Each applicant could select as many of these as applied. See [18] for a copy of the application.

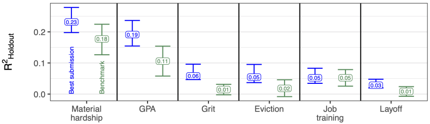
Appendix

Restriction	Material hardship	GPA	Grit	Eviction	Job training	Layoff
Valid submissions	160	160	160	160	160	160
Submissions different from the mean of the training data	122	128	121	111	117	112
Qualifying submissions (better than the mean of the training data)	92	98	65	48	42	42

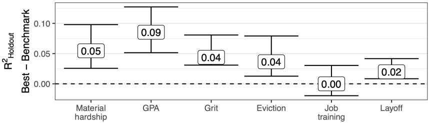
Table S5. Submission restrictions. Many of our analyses use either the full set of 160 valid submissions, the restricted set of submissions for which at least one prediction was at least 10^{-4} away from the mean of the training data, or the restricted set of qualifying submissions, which are the submissions that were more accurate than the mean of the training data.

Appendix

A) Performance of benchmark and best submissions.



B) Absolute improvement of the best over the benchmark.



C) Proportion of unpredicted component closed by best.

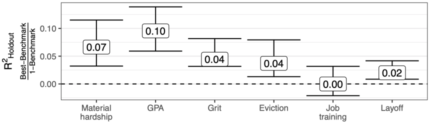


Fig. S6. Maximum $R^2_{\textbf{Holdout}}$ relative to benchmark models. Horizontal lines indicate the value that would be realized if the best submission and the benchmark had equal performance.

Appendix

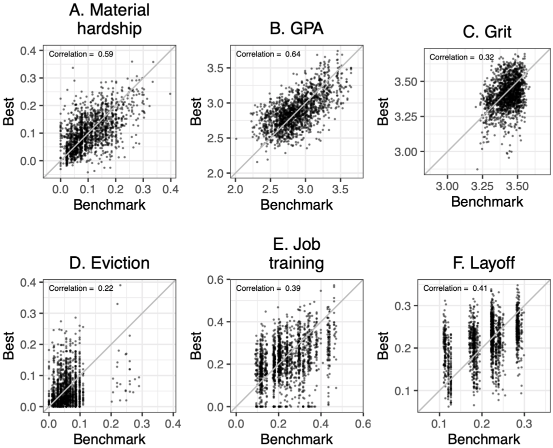


Fig. S8. Comparison of individual-level predictions for the best submissions and benchmark models. Although the best submissions and benchmark models have similar R^2_{Holdout}, they do not have equivalent individual-level predictions.

Appendix

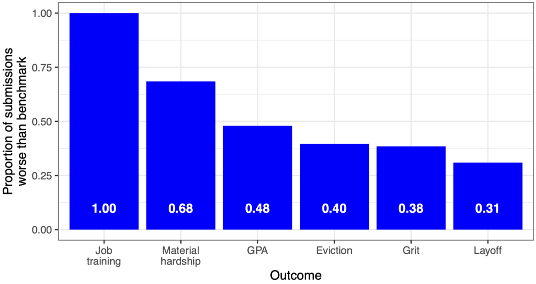


Fig. S9. Many qualifying submissions have worse predictive performance than the benchmark model.

Appendix

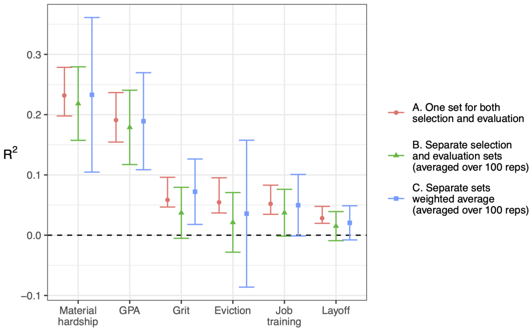


Fig. S11. Maximum R^2_{Holdout} relative to alternative estimators of out-of-sample error. Model (A) is the best-scoring submission to the Challenge, evaluated on the same set used to select it. Error bars approximate the sampling distribution of this estimate by the middle 95% of bootstrap draws, with selection and evaluation performed within each bootstrap draw. Model (B) splits the holdout set into selection and evaluation samples, selects the best model on the selection sample, and evaluates it on the evaluation sample, and averages over 100 repetitions of this procedure as described in Section S2.4. This approach avoids overfitting in the model selection stage. Model (C) uses the same sample splitting procedure as (B) but creates a weighted average of submissions as described in Section S2.5. Confidence intervals in (B) and (C) are analytic as described in Sections S2.4 and S2.5.

Appendix

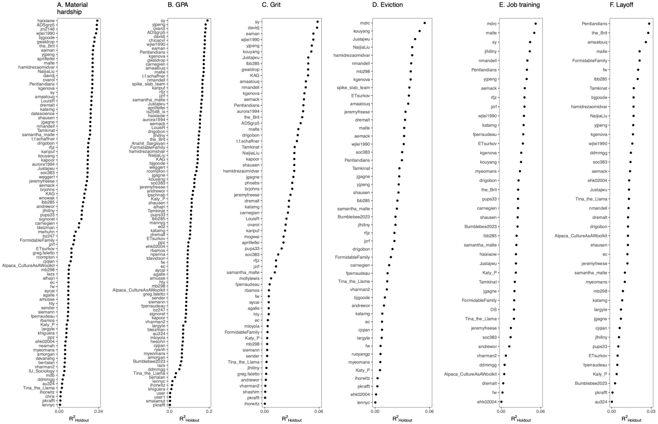


Fig. S15. The R^2_{Holdout} of many submissions are similar. The best model is not enormously better than the next-best model. Figure shows all qualifying submissions.

Appendix

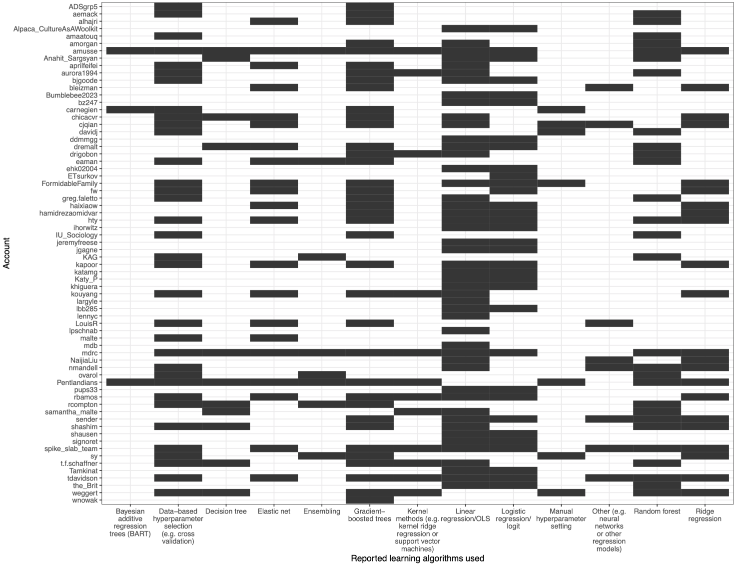


Fig. S21. Participant reports of learning algorithms used.

Ian Lundberg (Princeton)