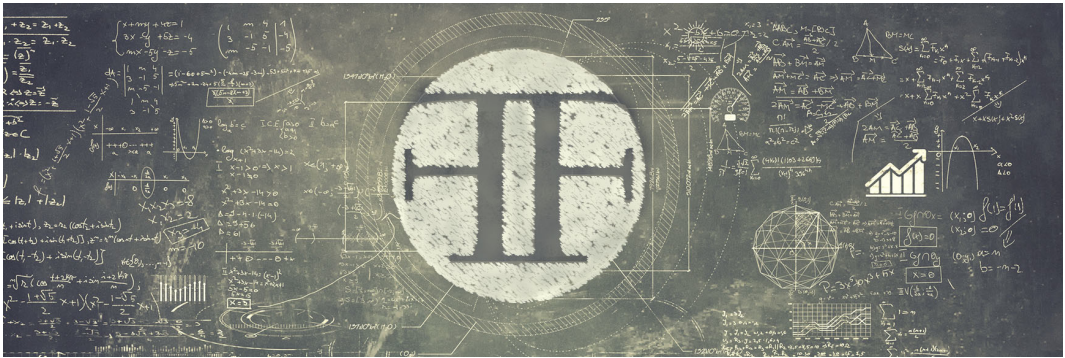
Class slides for Tuesday, November 10: Predictability of life trajectories

Matthew J. Salganik

COS 597E/SOC 555 Limits to prediction
Fall 2020, Princeton University



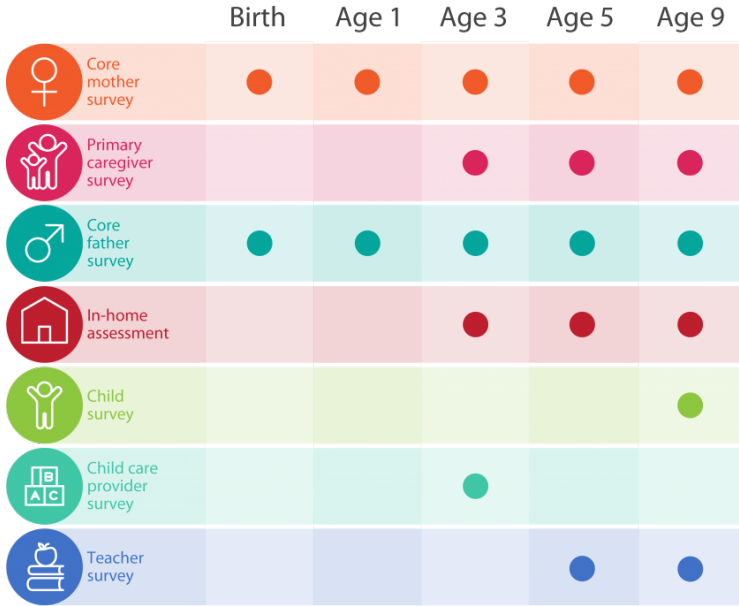
Fragile Families Challenge

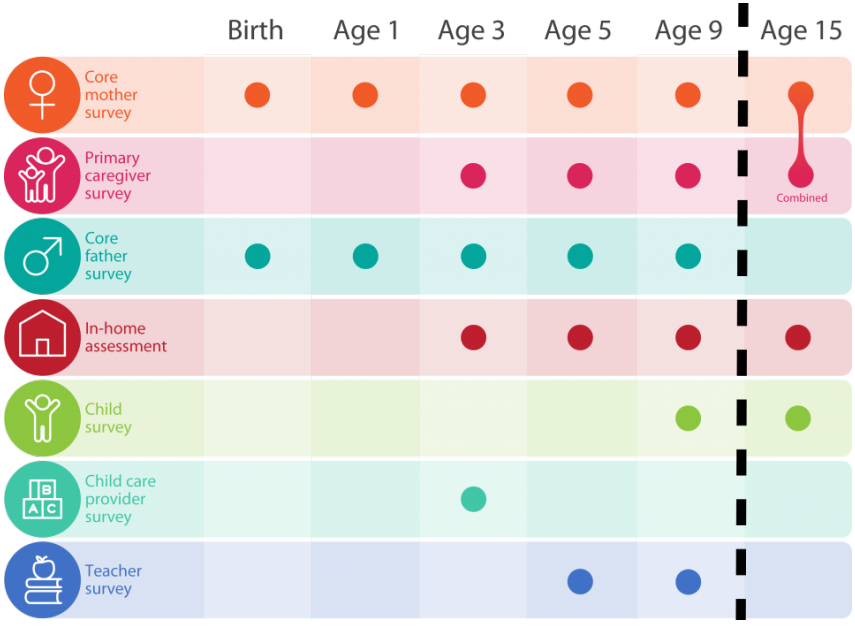
$$\hat{y} \quad \& \quad \hat{\beta}$$

Mullainathan and Spiess (2017)

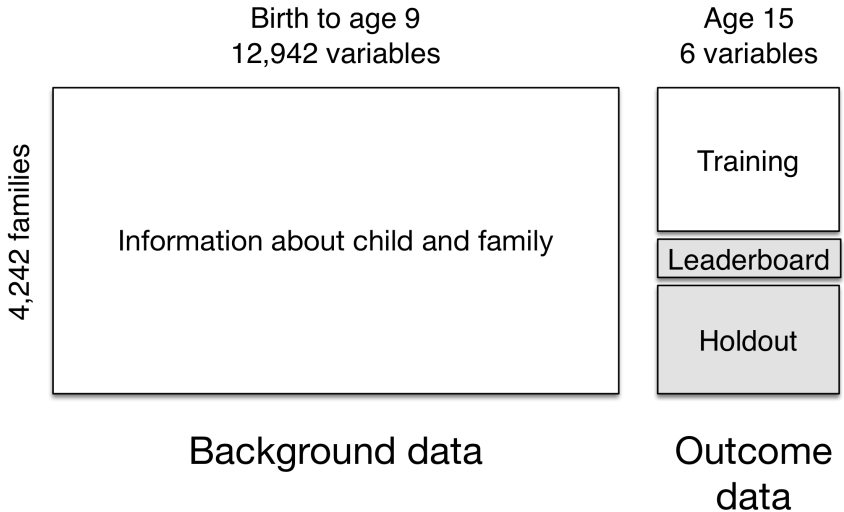
Fragile Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

- Birth cohort panel study
- $\approx$ 5,000 children born in 20 U.S. cities with an over-sample of non-marital births
- Followed from birth through age 15
- Already used in hundreds of papers and dozens of dissertations


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


Birth to age 9
12,942 features
Age 15
1,500 features
4,242 families



This is not inferring the present or forecasting. What is it?

Outcomes

- Child: GPA (continuous), Grit (continuous)
- Household: Eviction (binary), Material hardship (continuous)
- Primary care giver: Job training (binary), Job loss (binary)

457 researchers applied to participate. Many worked in interdisciplinary teams. Goal: Make a prediction that minimizes mean square error on the hold-out set

$$MSE_{\text{holdout}} = \frac{\sum_{i \in \text{holdout}} (\hat{y}_i - y_i)^2}{n_{holdout}}$$

More on privacy and ethics audit: Lundberg et al. (2019)

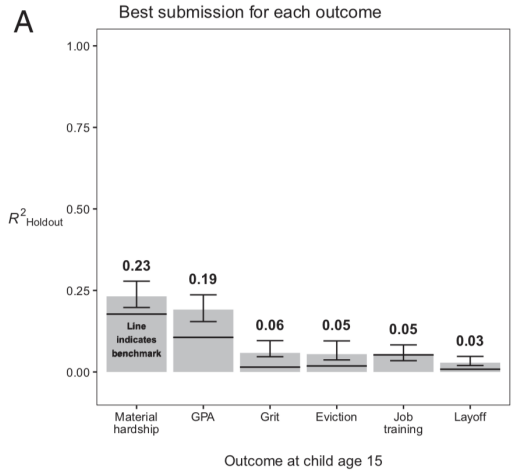
Using a large, high-quality social science dataset collected since birth and modern machine learning methods, how accurately can we predict outcomes from children, parents, and families?

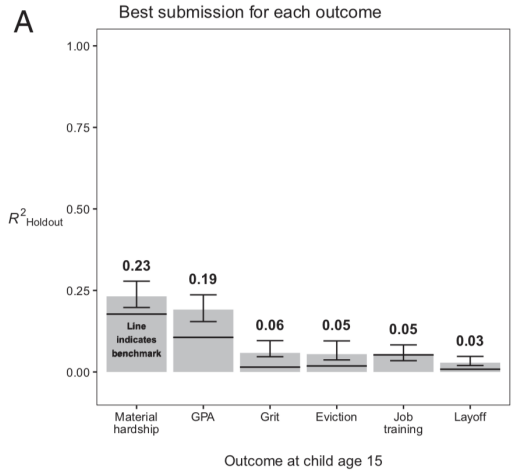
$$R^2_{holdout} = 1 - \frac{\sum_{i \in \text{holdout}} (\hat{y}_i - y_i)^2}{\sum_{i \in \text{holdout}} (\bar{y}_{train} - y_i)^2}$$

Using a large, high-quality social science dataset collected since birth and modern machine learning methods, how accurately can we predict outcomes from children, parents, and families?

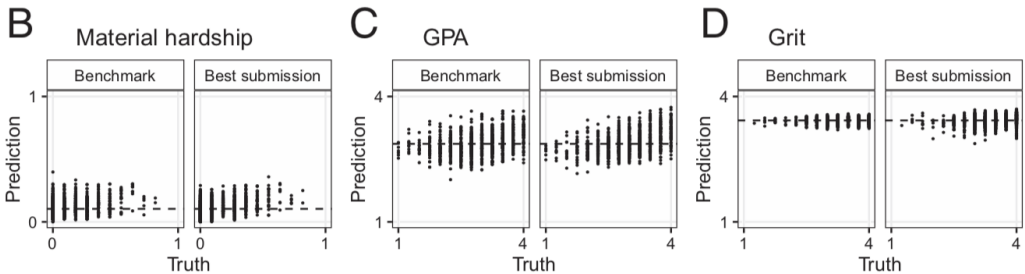
$$R^2_{holdout} = 1 - \frac{\sum_{i \in \text{holdout}} (\hat{y}_i - y_i)^2}{\sum_{i \in \text{holdout}} (\bar{y}_{train} - y_i)^2}$$

Before I show the results, let's vote . . . . (Also notice the switch from absolute to relative metric)

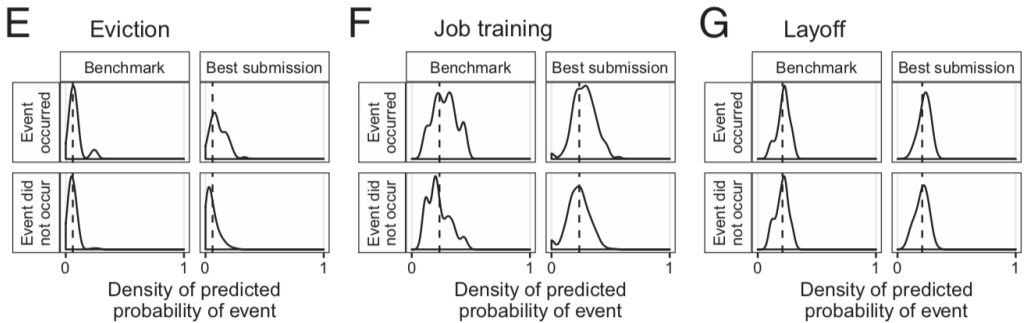

A
Best submission for each outcome
1.00
0.75
0.50
0.25
0.00
$R^2_{Holdout}$
0.23
0.19
0.06
0.05
0.05
0.03
Line indicates benchmark
Material hardship
GPA
Grit
Eviction
Job training
Layoff
Outcome at child age 15



What is the right y-scale here?

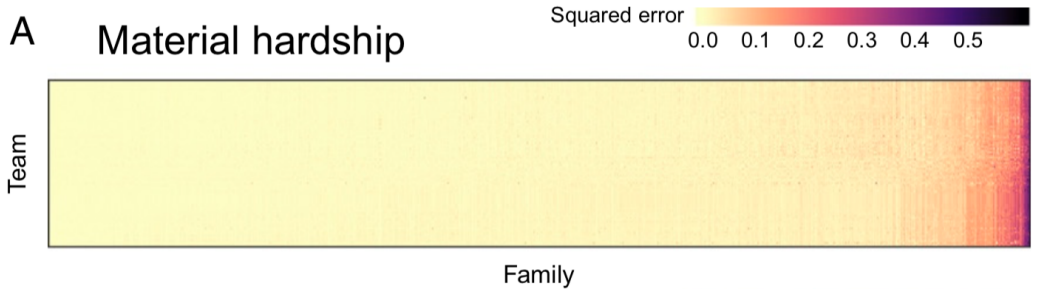


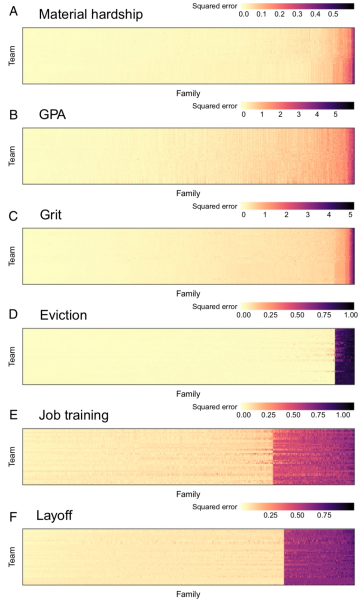
- Best submission is like the mean of the training data with a bit of signal, not like the truth with a bit of noise
- Best submission is qualitatively similar to the benchmark model

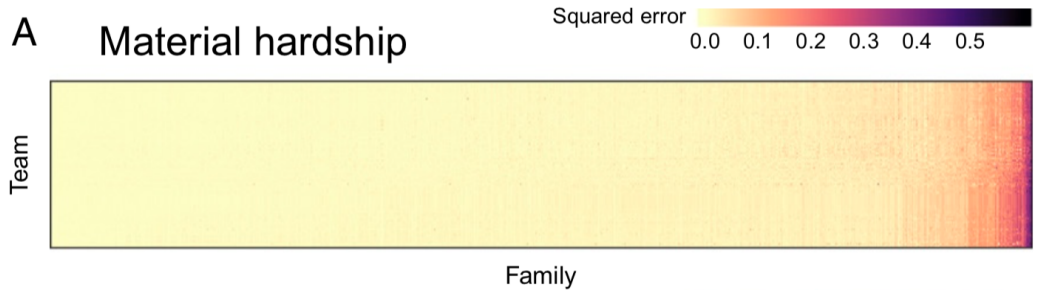


- Best submission is like the mean of the training data with a bit of signal, not like the truth with a bit of noise
- Best submission is qualitatively similar to the benchmark model

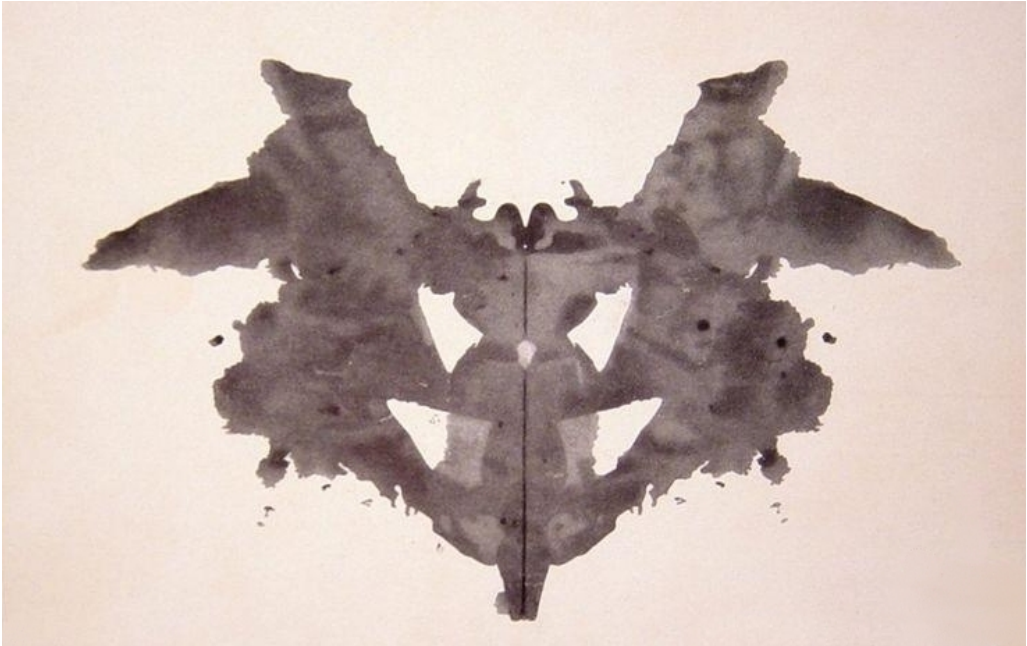
We can learn a lot by looking at all the valid submissions, not just the best


A
Material hardship
Squared error
0.0 0.1 0.2 0.3 0.4 0.5
Team
Family


A
Material hardship
Squared error
0.0 0.1 0.2 0.3 0.4 0.5
Team
Family
B
GPA
Squared error
0 1 2 3 4 5
Team
Family
C
Grit
Squared error
0 1 2 3 4 5
Team
Family
D
Eviction
Squared error
0.00 0.25 0.50 0.75 1.00
Team
Family
E
Job training
Squared error
0.00 0.25 0.50 0.75 1.00
Team
Family
F
Layoff
Squared error
0.25 0.50 0.75
Team
Family



Ensembling didn't improve things much. :(



Researchers must reconcile an “understanding/prediction” paradox

Researchers must reconcile an “understanding/prediction” paradox

- We don’t understand much

Researchers must reconcile an “understanding/prediction” paradox

- We don’t understand much
- Prediction is not a good measure of understanding

Researchers must reconcile an “understanding/prediction” paradox

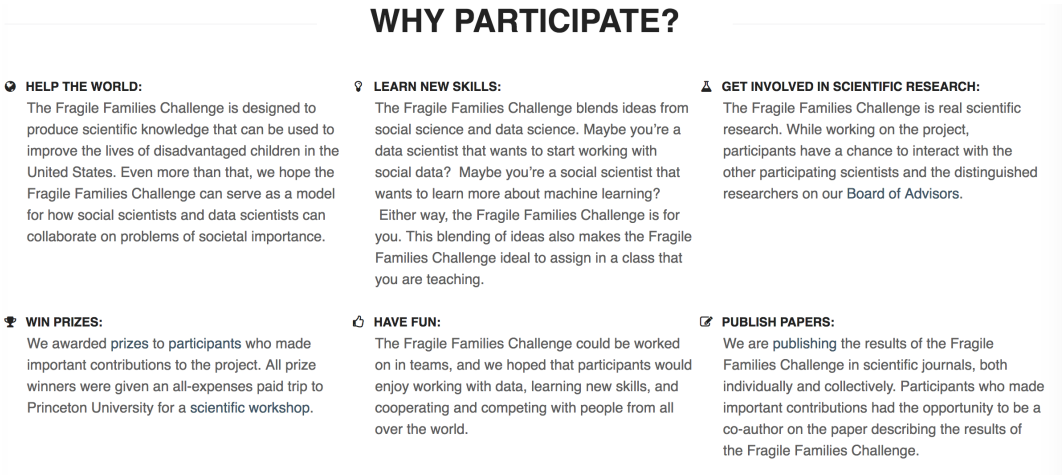
- We don’t understand much
- Prediction is not a good measure of understanding
- Our current understanding is correct but incomplete

Was the Fragile Families Challenge a success or failure?

Behind the scenes: Designing and running a mass collaboration

Too hard, don't do it

Use a mix of approaches to get folks to participate



WHY PARTICIPATE?

HELP THE WORLD:
The Fragile Families Challenge is designed to produce scientific knowledge that can be used to improve the lives of disadvantaged children in the United States. Even more than that, we hope the Fragile Families Challenge can serve as a model for how social scientists and data scientists can collaborate on problems of societal importance.

LEARN NEW SKILLS:
The Fragile Families Challenge blends ideas from social science and data science. Maybe you're a data scientist that wants to start working with social data? Maybe you're a social scientist that wants to learn more about machine learning? Either way, the Fragile Families Challenge is for you. This blending of ideas also makes the Fragile Families Challenge ideal to assign in a class that you are teaching.

GET INVOLVED IN SCIENTIFIC RESEARCH:
The Fragile Families Challenge is real scientific research. While working on the project, participants have a chance to interact with the other participating scientists and the distinguished researchers on our Board of Advisors.

WIN PRIZES:
We awarded prizes to participants who made important contributions to the project. All prize winners were given an all-expenses paid trip to Princeton University for a scientific workshop.

HAVE FUN:
The Fragile Families Challenge could be worked on in teams, and we hoped that participants would enjoy working with data, learning new skills, and cooperating and competing with people from all over the world.

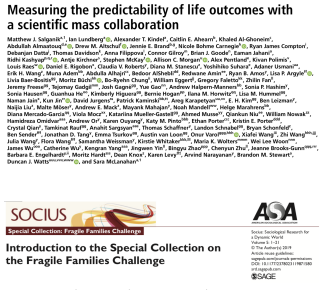
PUBLISH PAPERS:
We are publishing the results of the Fragile Families Challenge in scientific journals, both individually and collectively. Participants who made important contributions had the opportunity to be a co-author on the paper describing the results of the Fragile Families Challenge.

“Hack” existing scientific reward mechanisms: papers and prizes.

Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew J. Salganik[1*], Ian Lundberg[2†], Alexander T. Kindel[3], Caitlin E. Ahearn[4], Khaled Al-Ghoneim[5], Abdullah Almaatouq[6,7], Drew M. Altschul[8], Jennie E. Brand[9,10], Nicole Bohme Carnegie[11], Ryan James Compton[1], Debanjan Datta[1], Thomas Davidson[1], Anna Filippova[1], Connor Gilroy[12], Brian J. Goode[1], Eaman Jahani[13], Rishi Kashyap[14,15], Antje Kirchner[1], Stephen McKay[16], Allison C. Morgan[17], Alex Pentland[6], Kivan Polimis[1], Louis Raes[18], Daniel E. Rigobon[6], Claudia V. Roberts[1], Diana M. Stanescu[1], Yoshihiko Suhara[6], Adaner Usmani[19], Erik H. Wang[6], Muna Adem[20], Abdulla Alhajri[21], Bedoor AlShebli[20], Redhwan Amin[20], Ryan B. Amos[1], Lisa P. Argyle[22], Livia Bare-Bonitz[23], Moritz Bächt[24], Bo-Ryehn Chung[1], William Eggert[1], Gregory Faletto[25], Zhilin Fan[1], Jeremy Freese[26], Tejomay Gadgil[27], Josh Gagnon[28], Yue Gao[29], Andrew Halpern-Manners[30], Sonia P. Hashim[1], Sonia Hausen[20], Guanhua He[31], Kimberly Higuera[28], Bernie Hogan[32], Ilana M. Horwitz[33], Lisa M. Hummel[34], Naman Jain[6], Kun Jin[35], Donald Jurgens[36], Patrick Kaminski[24,37], Areg Karapetyan[38,39], E. H. Kim[40], Ben Leizerman[1], Naijia Liu[1], Malte Möser[1], Andrew E. Mack[1], Mayanka Mahajan[1], Noah Mandell[39,41], Helge Marahrens[39], Diana Mercado-Garcia[42], Viola Mocz[43], Katerina Mueller-Gastell[44], Ahmed Musse[45], Qiankun Niu[46], William Nowak[47], Hamidreza Omidvar[48], Andrew Or[1], Karen Ouyang[1], Katy M. Pinto[49], Ethan Porter[22], Kristin E. Porter[50], Crystal Qian[6], Tamkinat Rauf[49], Anahit Sargsyan[39], Thomas Schaffner[1], Landon Schnabel[49], Bryan Schonfeld[1], Ben Sender[51], Jonathan D. Tang[1], Emma Tsurkov[49], Austin van Loon[52], Onur Varol[53,54], Xiafei Wang[55], Zhi Wang[56,57], Julia Wang[1], Flora Wang[58], Samantha Weissman[1], Kirstie Whitaker[59,60], Maria K. Wolters[61,62], Wan Lee Woon[63,64], James Wu[1,65], Catherine Wu[1], Kengran Yang[1,66], Jingwen Yin[1], Bingyu Zhao[1,67], Chenyun Zhu[1], Jeanne Brooks-Gunn[1,68,69], Barbara E. Engelhardt[2,70], Moritz Hardt[71], Dean Knox[1], Karen Levy[72], Arvind Narayanan[2], Brandon M. Stewart[1], Duncan J. Watts[6,73,74,75,76,77], and Sara McLanahan[1,3]

“Hack” existing scientific reward mechanisms: papers and prizes.



Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew J. Salganik[a,1], Ian Lundberg[b], Alexander T. Kindel[c], Caitlin E. Ahearn[d], Khaled Al-Ghoneim[e], Abdulah Almaatouq[a,f], Drew M. Altschul[g], Jennie E. Brand[h,i], Nicole Bohme Carnegie[j], Ryan James Compton[c], Debanjan Datta[c], Thomas Davidson[k], Anna Filippova[l], Connor Gilroy[m], Brian J. Goode[n], Eaman Jahani[o], Rishi Kashyap[p,q], Antje Kirchner[r], Stephen McKay[s], Allison C. Morgan[t], Alex Pentland[f], Kivan Polimis[u], Louis Raes[v], Daniel E. Rigobon[w], Claudia V. Roberts[x], Diana M. Stanescu[y], Yoshihiko Suhara[z], Adaner Usmani[aa], Erik H. Wang[f], Muna Adem[ab], Abdulla Alhajri[ac], Bedoor AlShebli[ad], Redhwane Amin[ae], Ryan B. Amor[c], Lisa P. Argyle[af], Livia Bae-Bonilla[ag], Moritz Bächt[ah], Bo-Ryehn Chung[c], William Eggert[c], Gregory Faletto[ai], Zhilin Fan[c], Jeremy Freese[aj], Tejomay Gadgil[ak], Josh Gagné[al], Yue Gao[am], Andrew Halpern-Manners[an], Sonia P. Hashim[c], Sonia Hausen[ao], Guanhua He[ap], Kimberly Higuera[aq], Bernie Hogan[ar], Ilana M. Horwitz[as], Lisa M. Hummel[at], Naman Jain[c], Kun Jin[au], David Jurgens[av], Patrick Kaminski[ah,aw], Areg Karapetyan[ax,ay], E. H. Kim[az], Ben Leizerman[c], Naijia Liu[c], Malte Möser[c], Andrew E. Mack[c], Mayanka Mahajan[c], Noah Mandell[ay,c], Helge Marahrens[ah], Diana Mercado-Garcia[ah], Viola Mocz[ah], Katarina Mueller-Gastell[ah], Ahmed Musa[ah], Qiankun Niu[ah], William Nowak[ah], Hamidreza Omidvar[ah], Andrew Or[c], Karen Ouyang[c], Katy M. Pinto[ah], Ethan Porter[ah], Kristin E. Porter[ah], Crystal Qian[c], Tamkinat Rauf[ah], Anahit Sargsyan[ah,ay], Thomas Schaffner[c], Landon Schnabel[ah], Bryan Schonfeld[c], Ben Sender[ah], Jonathan D. Tang[c], Emma Tsurkov[ah], Austin van Loon[ah], Onur Varol[ah,ba], Xiafei Wang[c], Zhi Wang[bb,c], Julia Wang[c], Flora Wang[ah], Samantha Weissman[c], Kirstie Whitaker[ah,bc], Maria K. Wolters[ah,bd], Wei Lee Woon[ah,be], James Wu[ah], Catherine Wu[c], Kengran Yang[ah], Jingwen Yin[c], Bingyu Zhao[ah], Chenyue Zhu[c], Jeanne Brooks-Gunn[bb,ah,c], Barbara E. Engelhardt[c,f], Moritz Hardt[bf], Dean Knox[c], Karen Levy[bg], Arvind Narayanan[c], Brandon M. Stewart[c], Duncan J. Watts[ah,bb,bh,bi], and Sara McLanahan[a,1]

SOCIUS
Special Collection: Fragile Families Challenge

ASA
American Sociological Association

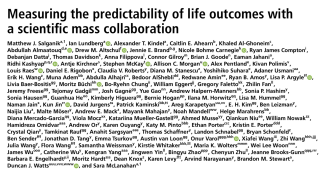
Socius: Sociological Research for a Dynamic World
Volume 5: 1–21
© The Author(s) 2019
Article reuse guidelines
sagepub.com/journals-permissions
DOI: 10.1177/2378023119871580
srd.sagepub.com

SAGE

Introduction to the Special Collection on the Fragile Families Challenge

Matthew J. Salganik[1], Ian Lundberg[2], Alexander T. Kindel[1], and Sara McLanahan[1]

“Hack” existing scientific reward mechanisms: papers and prizes.



Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew J. Salganik[1], Ian Lundberg[2], Alexander T. Kindel[3], Caitlin E. Ahearn[4], Khaled Al-Ghoneim[5], Abdulah Almaatouq[6,7], Drew M. Altschul[8], Jennie E. Brand[9,10], Nicole Bohme Carnegie[11], Ryan James Compton[1], Debanjan Datta[1], Thomas Davidson[2], Anna Filippova[1], Connor Gilroy[12], Brian J. Goode[1], Eaman Jahani[2], Rishi Kashyap[13,14], Antje Kirchner[1], Stephen McKay[15], Allison C. Morgan[16], Alex Pentland[6], Kivan Polimis[1], Louis Raes[17], Daniel E. Rigobon[6], Claudia V. Roberts[1], Diana M. Stanescu[1], Yoshihiko Suhara[6], Adaner Usmani[18], Erik H. Wang[6], Muna Adem[19], Abdulla Alhajri[5], Bedoor AlShebli[20], Redhwane Amin[19], Ryan B. Amos[1], Lisa P. Argyle[16], Livia Bare-Bonitz[20], Moritz Bächt[21], Bo-Ryehn Chung[1], William Eggert[6], Gregory Faletto[22], Zhilin Fan[1], Jeremy Freese[23], Tejomay Gadgil[24], Josh Gagnon[25], Yue Gao[26], Andrew Halpern-Manners[25], Sonia P. Hashim[1], Sonia Hausen[20], Guanhua He[27], Kimberly Higuera[28], Bernie Hogan[29], Ilana M. Horwitz[30], Lisa M. Hummel[20], Naman Jain[6], Kun Jin[31], David Jurgens[32], Patrick Kaminski[24,33], Anup Karapetyan[34,35], E. H. Kim[6], Ben Leizerman[1], Naijia Liu[1], Malte Möser[2], Andrew E. Mack[1], Mayank Mahajan[1], Noah Mandell[24,36], Helge Marahrens[20], Diana Mercado-Garcia[19], Viola Mocz[37], Katarina Mueller-Gastell[20], Ahmed Muse[19], Qiankun Niu[19], William Nowak[38], Hamidreza Omidvar[2,39], Andrew Or[6], Karen Ouyang[1], Katy M. Pinto[20], Ethan Porter[23], Kristin E. Porter[40], Crystal Qian[2], Tamkinat Rauf[20], Anahita Sargsyan[24,36], Thomas Schaffner[1], London Schnabel[20], Bryan Schonfeld[1], Ben Sender[20], Jonathan D. Tang[2], Emma Tsoodle[20], Austin van Loon[20], Onur Varol[20,41,42], Xiafei Wang[43], Zhi Wang[20,44], Julia Wang[2], Flora Wang[20], Samantha Weissman[1], Kirstie Whitaker[20,45], Maria K. Wolters[46,47], Wan Lee Woon[48,49], James Wu[50], Catherine Wu[1], Kengran Yang[2,51], Jingwen Yin[2], Bingyu Zhao[20], Chenyue Zhu[2], Jeanne Brooks-Gunn[20,52,53], Barbara E. Engelhardt[54], Moritz Hardt[55], Dean Knox[1], Karen Levy[23], Arvind Narayanan[2], Brandon M. Stewart[1], Duncan J. Watts[6,56,57,58,59], and Sara McLanahan[1,3]

SOCIUS
Special Collection: Fragile Families Challenge
ASA
Sociological Research for a Dynamic World
Volume 5: 1-21

DOI: 10.1177/2378023119871580
srd.sagepub.com
SAGE

Introduction to the Special Collection on the Fragile Families Challenge

Matthew J. Salganik[1], Ian Lundberg[2], Alexander T. Kindel[3], and Sara McLanahan[1]

SOCIUS
Special Collection: Fragile Families Challenge

Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge

Ian Lundberg[1], Arvind Narayanan[1], Karen Levy[2], and Matthew J. Salganik[1]

Socius: Sociological Research for a Dynamic World
Volume 5: 1–25

DOI: 10.1177/2378023118818023
srd.sagepub.com
SAGE

SOCIUS
Special Collection: Fragile Families Challenge

Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge

Ian Lundberg[1], Arvind Narayanan[1], Karen Levy[2], and Matthew J. Salganik[1]

Socius: Sociological Research for a Dynamic World
Volume 5: 1–25

DOI: 10.1177/2378023118813023
srd.sagepub.com

SOCIUS
Special Collection: Fragile Families Challenge

Improving Metadata Infrastructure for Complex Surveys: Insights from the Fragile Families Challenge

Alexander T. Kindel[1], Vineet Bansal[1], Kristin D. Catena[1], Thomas H. Hartshorne[1], Kate Jaeger[1], Dawn Koffman[1], Sara McLanahan[1], Maya Phillips[1], Shiva Rouhani[1], Ryan Vinh[1], and Matthew J. Salganik[1]

Socius: Sociological Research for a Dynamic World
Volume 5: 1–24

DOI: 10.1177/2378023118817378
srd.sagepub.com

Socius: Sociological Research for a Dynamic World
Volume 5: 1–25

DOI: 10.1177/2378023118813023
srd.sagepub.com

Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge

Ian Lundberg[1], Arvind Narayanan[1], Karen Levy[2], and Matthew J. Salganik[1]

Socius: Sociological Research for a Dynamic World
Volume 5: 1–24

DOI: 10.1177/2378023118817378
srd.sagepub.com

Improving Metadata Infrastructure for Complex Surveys: Insights from the Fragile Families Challenge

Alexander T. Kindel[1], Vineet Bansal[1], Kristin D. Catena[1], Thomas H. Hartshorne[1], Kate Jaeger[1], Dawn Koffman[1], Sara McLanahan[1], Maya Phillips[1], Shiva Rouhani[1], Ryan Vinh[1], and Matthew J. Salganik[1]

Socius: Sociological Research for a Dynamic World
Volume 5: 1–21

DOI: 10.1177/2378023119849803
srd.sagepub.com

Successes and Struggles with Computational Reproducibility: Lessons from the Fragile Families Challenge

David M. Liu[1] and Matthew J. Salganik[1]

Competition vs collaboration

- We can learn from looking at the full set of submissions not just the best

Competition vs collaboration

- We can learn from looking at the full set of submissions not just the best
- We can open source everything to seed future research

Competition vs collaboration

- We can learn from looking at the full set of submissions not just the best
- We can open source everything to seed future research
- Just feels right to me

Break into discussion groups

Getting ready for Thursday

Thoughts about reading the interviews

- Starts simple, gets more personal. This is what happens when you talk to people.

Thoughts about reading the interviews

- Starts simple, gets more personal. This is what happens when you talk to people.
- Year 15 interview GPA refers to 9th grade for both young adults.

Thoughts about reading the interviews

- Starts simple, gets more personal. This is what happens when you talk to people.
- Year 15 interview GPA refers to 9th grade for both young adults.
- Difficulty of calibration.

Thoughts about reading the interviews

- Starts simple, gets more personal. This is what happens when you talk to people.
- Year 15 interview GPA refers to 9th grade for both young adults.
- Difficulty of calibration.
- What is the goal of the reading? Filling out the sheet and helping us make sense of the results from the Challenge

Thoughts about reading the interviews

- Starts simple, gets more personal. This is what happens when you talk to people.
- Year 15 interview GPA refers to 9th grade for both young adults.
- Difficulty of calibration.
- What is the goal of the reading? Filling out the sheet and helping us make sense of the results from the Challenge
- If you don't have access yet, let me know

Thoughts about reading the interviews

- Starts simple, gets more personal. This is what happens when you talk to people.
- Year 15 interview GPA refers to 9th grade for both young adults.
- Difficulty of calibration.
- What is the goal of the reading? Filling out the sheet and helping us make sense of the results from the Challenge
- If you don't have access yet, let me know
- We will have guess on Thursday from the FFCWS data team and the dark matter interview team

Thoughts about reading the interviews

- Starts simple, gets more personal. This is what happens when you talk to people.
- Year 15 interview GPA refers to 9th grade for both young adults.
- Difficulty of calibration.
- What is the goal of the reading? Filling out the sheet and helping us make sense of the results from the Challenge
- If you don't have access yet, let me know
- We will have guess on Thursday from the FFCWS data team and the dark matter interview team

Class slides for Tuesday, November 10: Predictability of life trajectories

Matthew J. Salganik

COS 597E/SOC 555 Limits to prediction
Fall 2020, Princeton University