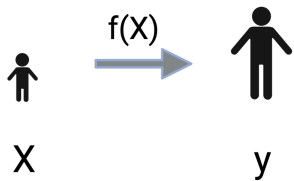
HW2: Fragile Families Challenge

COS424/524 precept

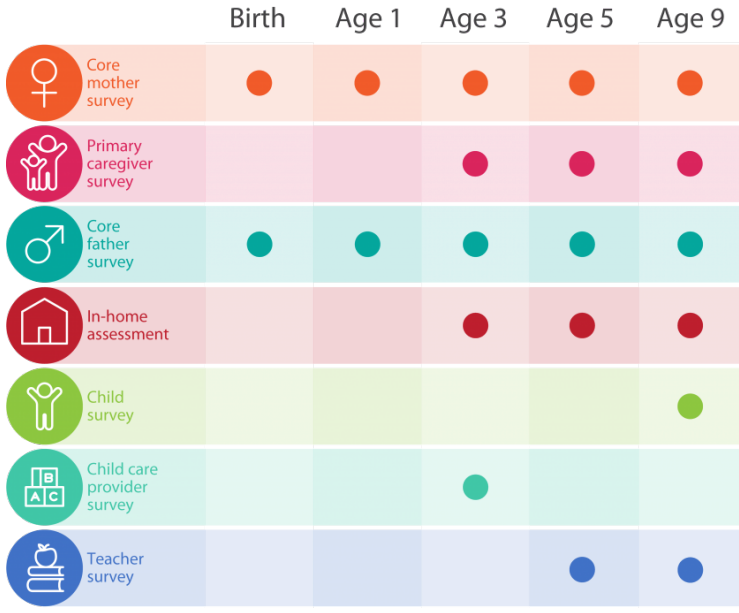
Slides adapted from Prof. Matt Salganik's SICSS materials.

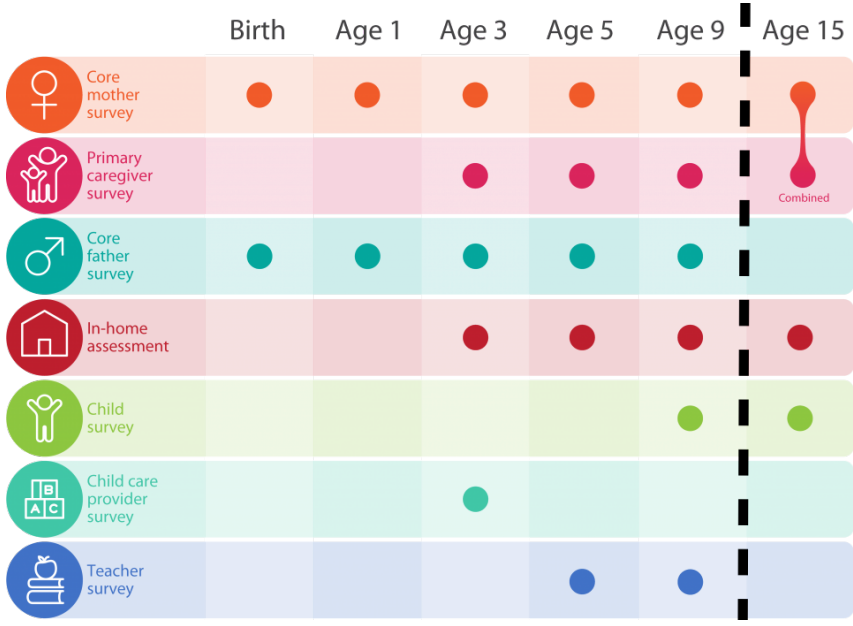
Given data about a young child's circumstances, are outcomes later in life predictable?



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
- "Fragile" because they're at a greater risk of breaking up and living in poverty than more traditional families


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

Example questions

Mother: How many days per week do you read stories to your child?

Father: How many days per week do you hug or show physical affection to your child?

Child: On average, how much time do you spend hanging out with friends on a weekday?

Life course researchers have:

- described social patterns
- theorized important factors
- estimated causal effects

Life course researchers have:

- described social patterns
- theorized important factors
- estimated causal effects

How well can we predict individual life outcomes?

We should care about the predictability of social outcomes

We should care about the predictability of social outcomes

- Scientific reasons

We should care about the predictability of social outcomes

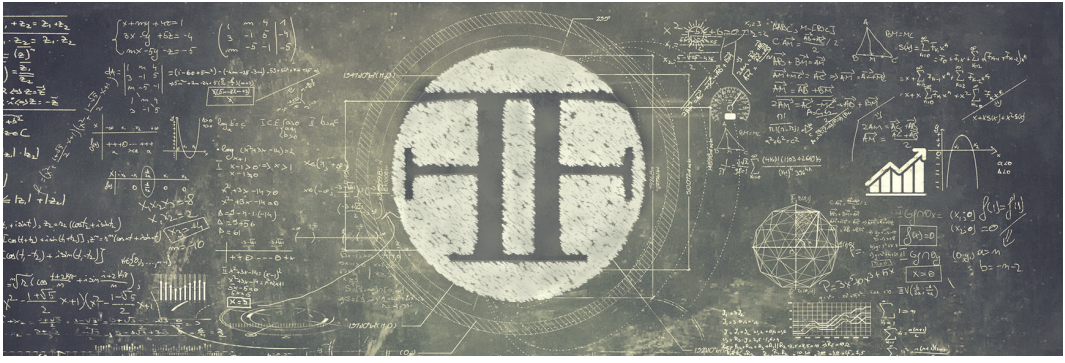
- Scientific reasons
  - Basic social fact

We should care about the predictability of social outcomes

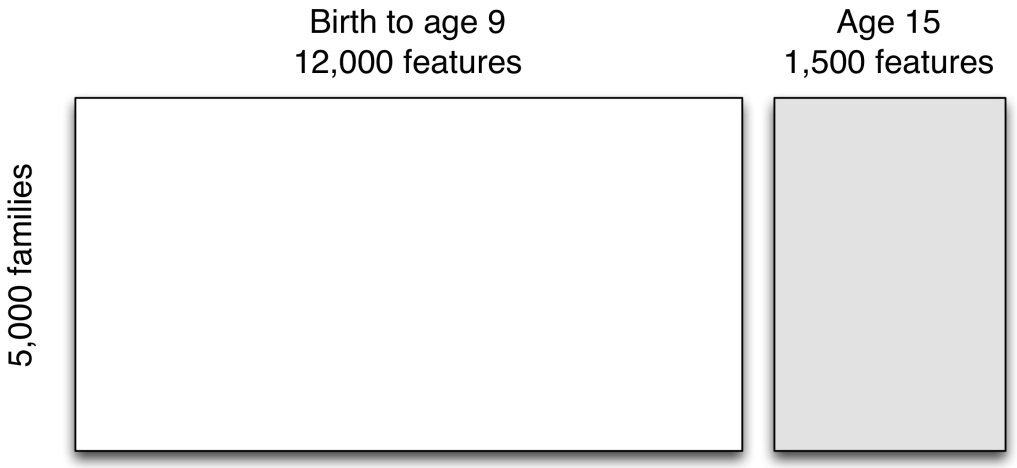
- Scientific reasons
  - Basic social fact
  - Discovery

We should care about the predictability of social outcomes

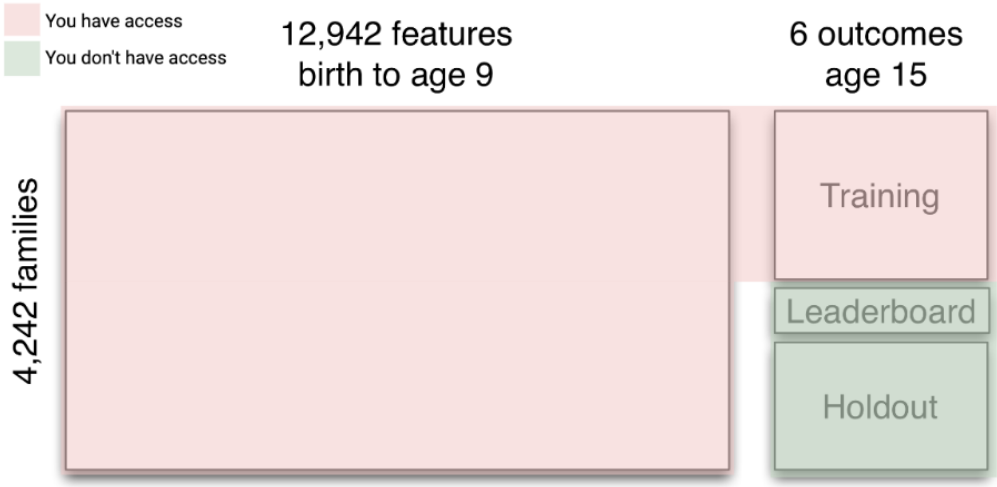
- Scientific reasons
  - Basic social fact
  - Discovery
- Policy reasons



Fragile Families Challenge


Birth to age 9
12,000 features
Age 15
1,500 features
5,000 families


12,942 features
birth to age 9
6 outcomes
age 15
4,242 families
Training
Leaderboard
Holdout


You have access
You don't have access
12,942 features
birth to age 9
6 outcomes
age 15
4,242 families
Training
Leaderboard
Holdout

Covariates

- $\sim 13,000$ features from questionnaires

Outcomes

- GPA
- Grit
- Eviction
- Material hardship
- Job training
- Job loss

Outcomes

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

Measuring the predictability of life outcomes with a scientific mass collaboration

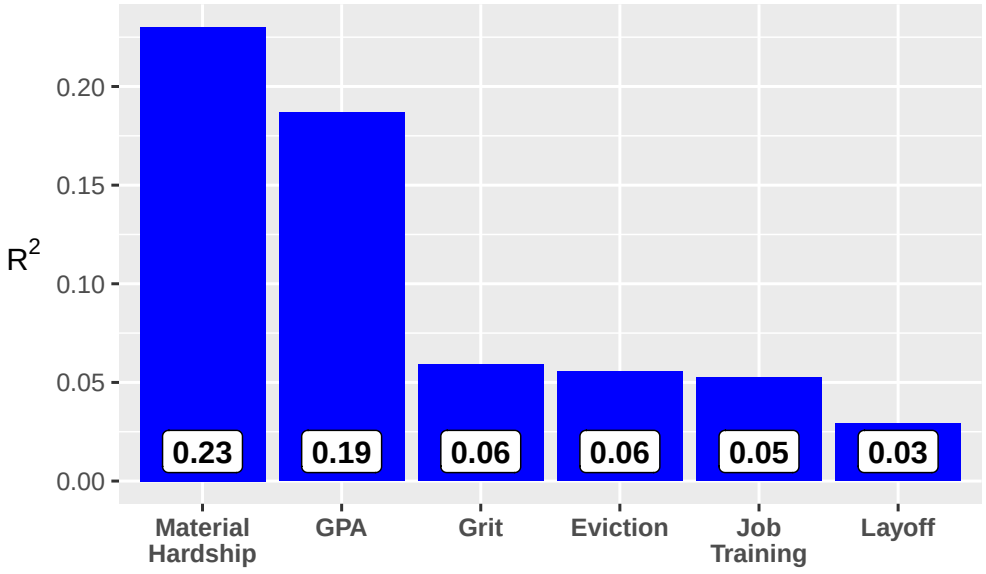
Matthew J. Salganik[a,1], Ian Lundberg[a], Alexander T. Kindel[a], Caitlin E. Ahearn[b], Khaled Al-Ghoneim[c], Abdullah Almaatouq[d,e], Drew M. Altschul[f], Jennie E. Brand[b,g], Nicole Bohme Carnegie[h], Ryan James Compton[i], Debanjan Datta[j], Thomas Davidson[k], Anna Filippova[l], Connor Gilroy[m], 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]

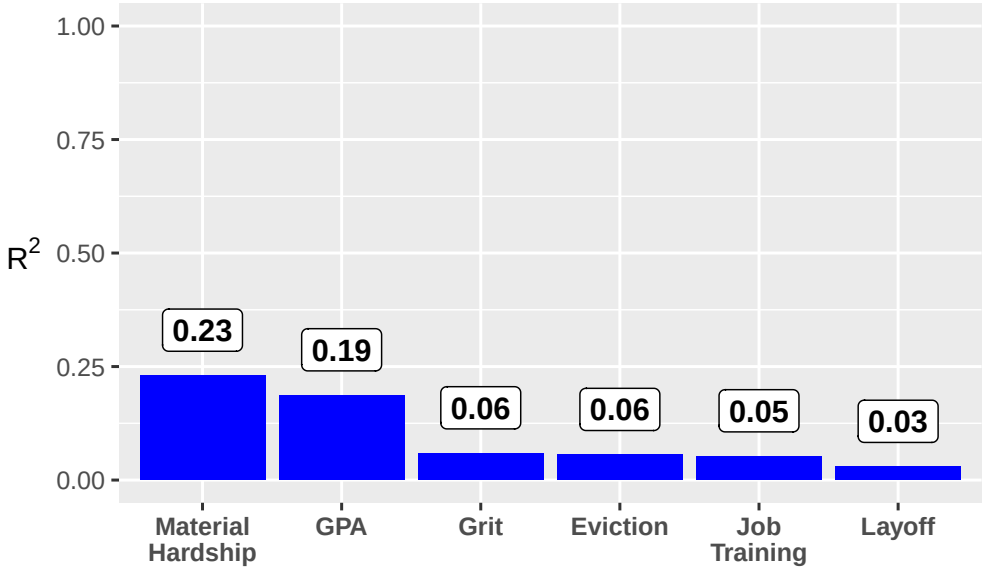
https://doi.org/10.1073/pnas.1915006117

457 researchers applied to participate. Many worked in interdisciplinary teams.

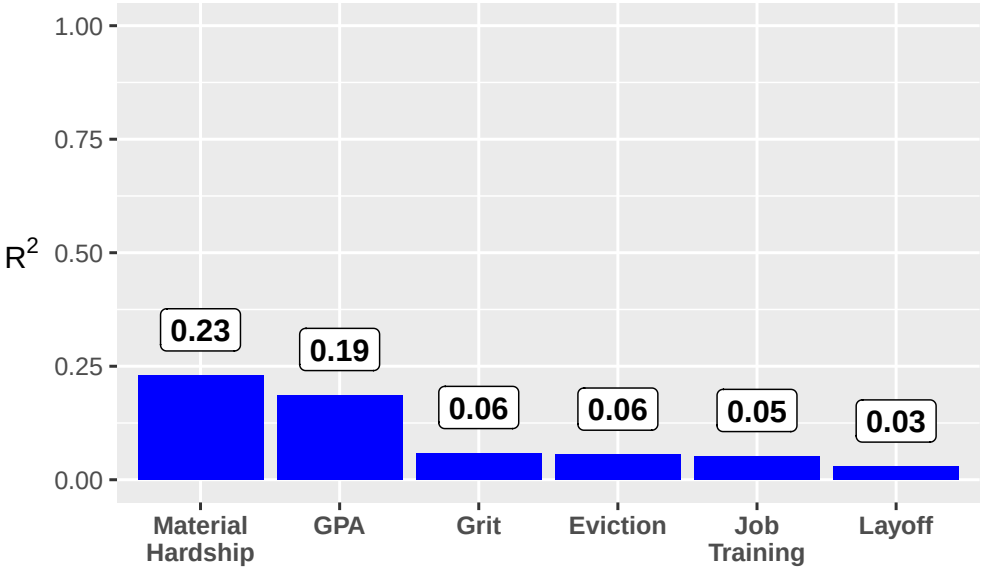
457 researchers applied to participate. Many worked in interdisciplinary teams. Using a large, high-quality social science dataset collected since birth and modern machine learning methods, how accurately can we predict outcomes from children, parents, and families?

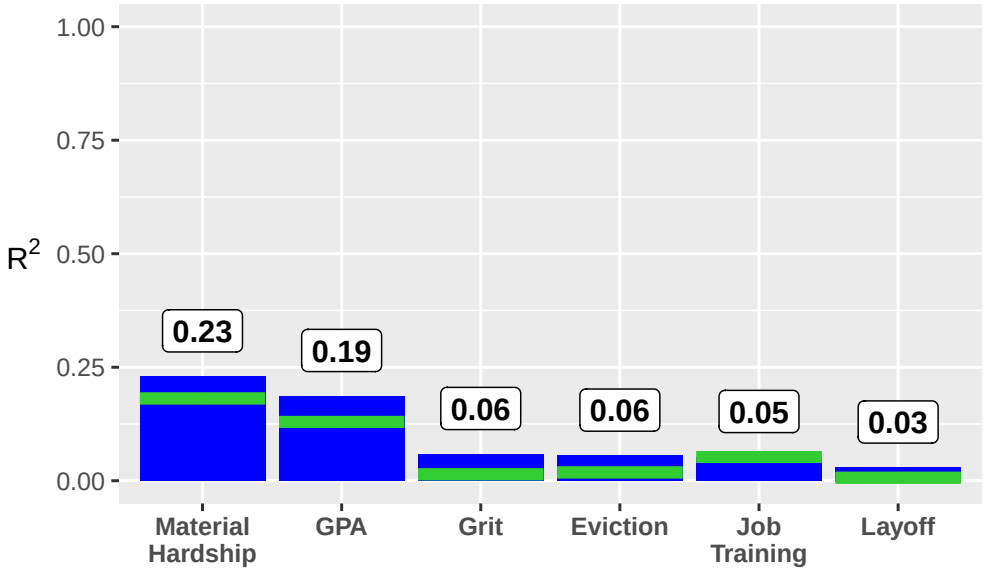
$$R^2_{holdout} = 1 - \frac{\sum_{i \in \text{holdout}} (\hat{y}_i - y_i)^2}{\sum_{i \in \text{holdout}} (\bar{y}_{train} - y_i)^2}$$


R^2
0.20
0.15
0.10
0.05
0.00
0.23
0.19
0.06
0.06
0.05
0.03
Material Hardship
GPA
Grit
Eviction
Job Training
Layoff


1.00
0.75
0.50
0.25
0.00
R^2
0.23
0.19
0.06
0.06
0.05
0.03
Material Hardship
GPA
Grit
Eviction
Job Training
Layoff

Is this better than a simple benchmark model?


1.00
0.75
0.50
0.25
0.00
R^2
0.23
0.19
0.06
0.06
0.05
0.03
Material Hardship
GPA
Grit
Eviction
Job Training
Layoff



Green line: 4 variable regression model

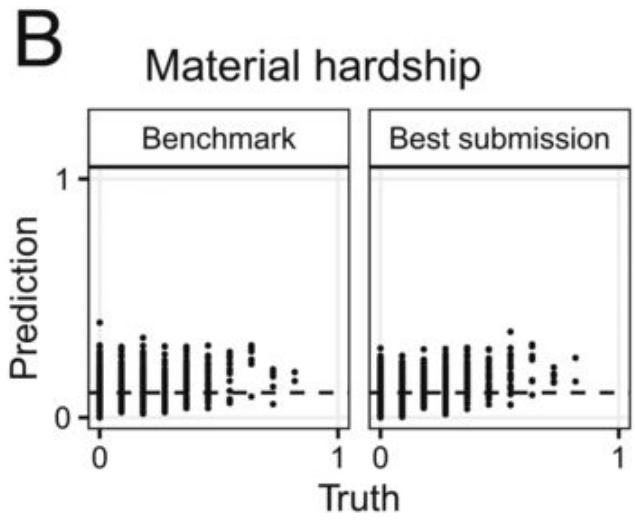


Fig 3, Salganik et al. (2020)

What can we learn looking at all the predictions?

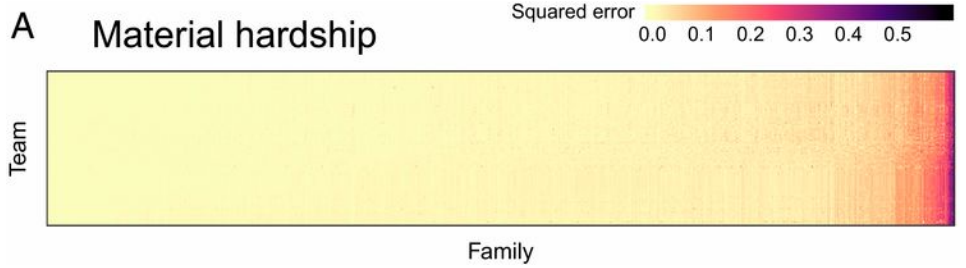


Fig 4, Salganik et al. (2020)

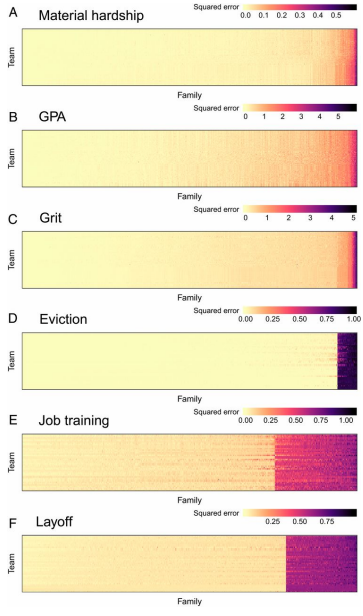


Fig 4, Salganik et al. (2020)

What do these results mean for policy makers?

- Machine learning is not magic

- Machine learning is not magic
- Transparent evaluation of any algorithm is needed

- Machine learning is not magic
- Transparent evaluation of any algorithm is needed
- Complex models may not outperform simple models

What do these results mean for researchers?

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

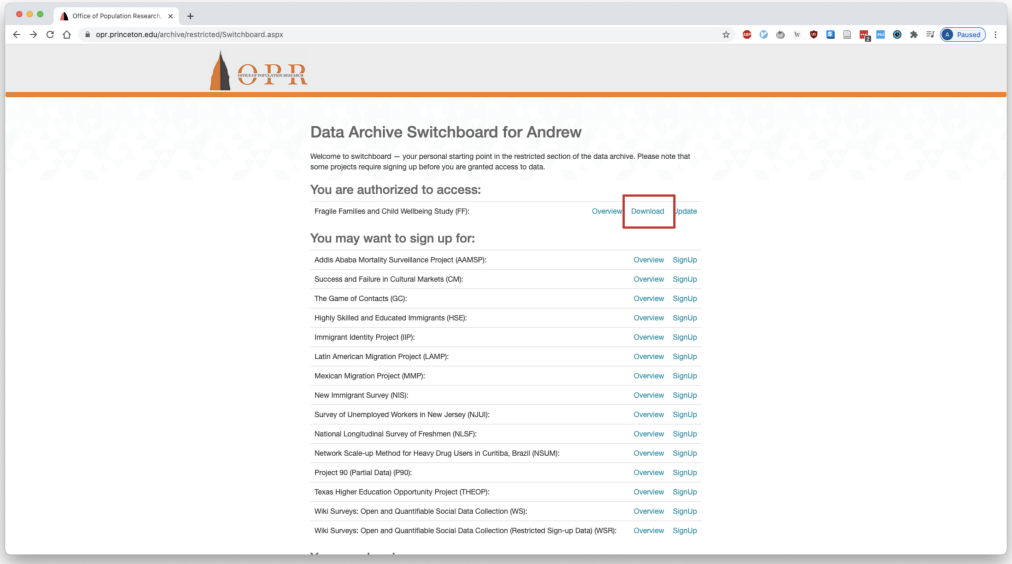
- We don’t understand much
- Prediction is not a good measure of understanding
- Our current understanding is correct but incomplete

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

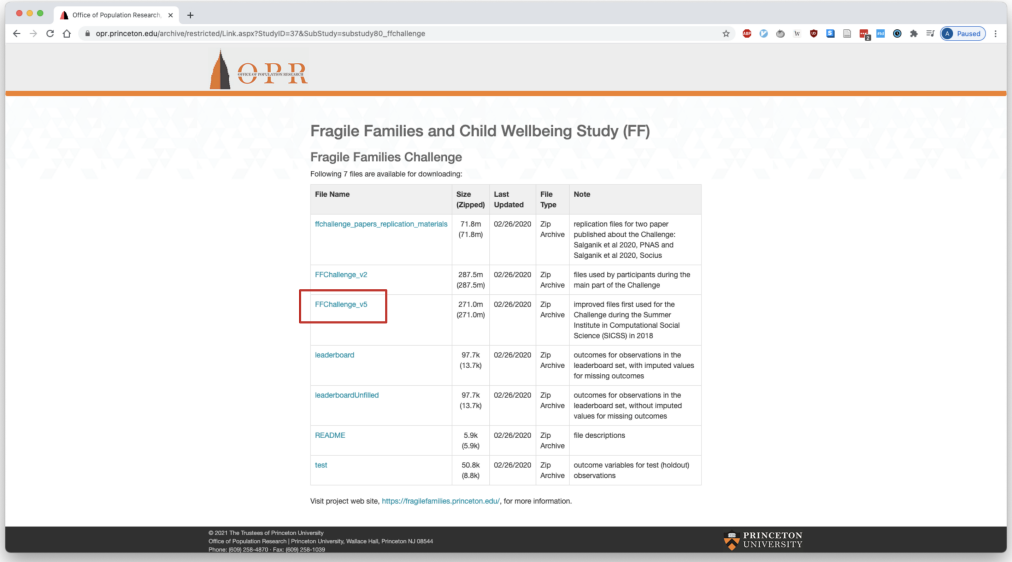
Mullainathan and Spiess (2017)

Accessing the data

- You've already registered through Princeton's OPR
- Log in and download `FFChallenge_v5`


Office of Population Research
opr.princeton.edu/archive/restricted/Switchboard.aspx
Paused
OPR
Data Archive Switchboard for Andrew
Welcome to switchboard — your personal starting point in the restricted section of the data archive. Please note that some projects require signing up before you are granted access to data.
You are authorized to access:
Fragile Families and Child Wellbeing Study (FF): Overview Download Update
You may want to sign up for:
Addis Ababa Mortality Surveillance Project (AAMSP): Overview SignUp
Success and Failure in Cultural Markets (CM): Overview SignUp
The Game of Contacts (GC): Overview SignUp
Highly Skilled and Educated Immigrants (HSE): Overview SignUp
Immigrant Identity Project (IIP): Overview SignUp
Latin American Migration Project (LAMP): Overview SignUp
Mexican Migration Project (MMP): Overview SignUp
New Immigrant Survey (NIS): Overview SignUp
Survey of Unemployed Workers in New Jersey (NJUI): Overview SignUp
National Longitudinal Survey of Freshmen (NLSF): Overview SignUp
Network Scale-up Method for Heavy Drug Users in Curitiba, Brazil (NSUM): Overview SignUp
Project 90 (Partial Data) (P90): Overview SignUp
Texas Higher Education Opportunity Project (THEOP): Overview SignUp
Wiki Surveys: Open and Quantifiable Social Data Collection (WS): Overview SignUp
Wiki Surveys: Open and Quantifiable Social Data Collection (Restricted Sign-up Data) (WSR): Overview SignUp


Office of Population Research
opr.princeton.edu/archive/restricted/link.aspx?studyID=37
Paused
OPR
Fragile Families and Child Wellbeing Study (FF)
Data for this study are organized by sub-studies as follows:
All Waves (all FFCWS public data from Baseline - Year 15)
Baseline
Year 1
Year 3
Year 5
Year 9
Year 15
Fragile Families Challenge
Visit project web site, https://fragilefamilies.princeton.edu/, for more information.
© 2021 The Trustees of Princeton University
Office of Population Research | Princeton University, Wallace Hall, Princeton NJ 08544
Phone: (609) 258-4870 - Fax: (609) 258-1039
PRINCETON UNIVERSITY


Office of Population Research
opr.princeton.edu/archive/restricted/Link.aspx?StudyID=37&SubStudy=substudy80_ffchallenge
Paused
OPR
Fragile Families and Child Wellbeing Study (FF)
Fragile Families Challenge
Following 7 files are available for downloading:
File Name
Size (Zipped)
Last Updated
File Type
Note
ffchallenge_papers_replication_materials
71.8m (71.8m)
02/26/2020
Zip Archive
replication files for two paper published about the Challenge: Salganik et al 2020, PNAS and Salganik et al 2020, Socius
FFChallenge_v2
287.5m (287.5m)
02/26/2020
Zip Archive
files used by participants during the main part of the Challenge
FFChallenge_v5
271.0m (271.0m)
02/26/2020
Zip Archive
improved files first used for the Challenge during the Summer Institute in Computational Social Science (SICSS) in 2018
leaderboard
97.7k (13.7k)
02/26/2020
Zip Archive
outcomes for observations in the leaderboard set, with imputed values for missing outcomes
leaderboardUnfilled
97.7k (13.7k)
02/26/2020
Zip Archive
outcomes for observations in the leaderboard set, without imputed values for missing outcomes
README
5.9k (5.9k)
02/26/2020
Zip Archive
file descriptions
test
50.8k (8.8k)
02/26/2020
Zip Archive
outcome variables for test (holdout) observations
Visit project web site, https://fragilefamilies.princeton.edu/, for more information.
© 2021 The Trustees of Princeton University
Office of Population Research | Princeton University, Wallace Hall, Princeton NJ 08544
Phone: (609) 258-4870 · Fax: (609) 258-1039
PRINCETON UNIVERSITY

Your project

You'll be working with two files in `FFChallenge_v5/`:

- `background.csv` — contains $13,026$ variables for $4,242$ families.
- `train.csv` — contains 6 later-life outcome variables for half $(2,121)$ of the families.

Your project

You'll be working with two files in `FFChallenge_v5/`:

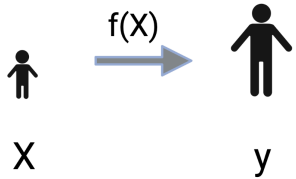
- `background.csv` — contains $13,026$ variables for $4,242$ families. **This is your** X.
- `train.csv` — contains 6 later-life outcome variables for half $(2,121)$ of the families. **This is your** y.

Your project

You'll be working with two files in `FFChallenge_v5/`:

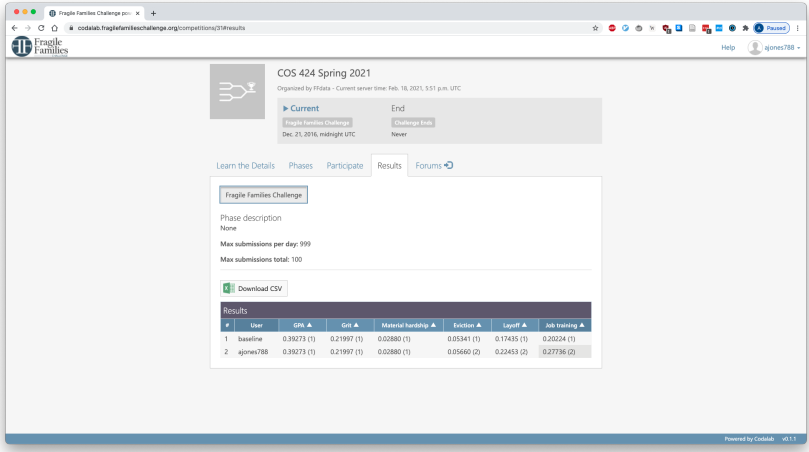
- `background.csv` — contains $13,026$ variables for $4,242$ families. **This is your** X.
- `train.csv` — contains 6 later-life outcome variables for half $(2,121)$ of the families. **This is your** y.

Your goal: Using the training data, build a predictive model to predict y from X.



Leaderboard

The leaderboard will track the performance of each person's/group's model on the test dataset.



Imputing missing data

Use `MissingDataScript.py` to fill in missing values in `background.csv`.

```python
import pandas as pd
import numpy as np

def fillMissing(inputcsv, outputcsv):

    # read input csv - takes time
    df = pd.read_csv(inputcsv, low_memory=False)
    # Fix date bug
    df.cf4fint = ((pd.to_datetime(df.cf4fint) - pd.to_datetime('1960-01-01')) / np.timedelta64(1, 'D')).astype(int)

    # replace NA's with mode
    df = df.fillna(df.mode().iloc[0])
    # if still NA, replace with 1
    df = df.fillna(value=1)
    # replace negative values with 1
    num = df._get_numeric_data()
    num[num < 0] = 1
    # write filled outputcsv
    df.to_csv(outputcsv, index=False)

# Usage:
if __name__ == "__main__":
    from os.path import join as pjoin

    data_dir = "./FFChallenge_v5"

    ## impute the background data
    in_path = pjoin(data_dir, 'background.csv')
    out_path = pjoin(data_dir, 'background_clean.csv')
    fillMissing(in_path, out_path)
```

Important notes

- Submitting the leaderboard requires a very specific format. Read the HW2 doc on Canvas carefully.
- You'll submit predictions for both the *train* and *test* data
- Each family has a unique `challengeID`. Use these to match the samples in `background.csv` with the samples in `train.csv`.

Support-Vector Networks (Cortes, Vapnik 1995)

$\sim$ 15 minutes for discussion

Wrap-up

- HW2: Fragile Families Challenge out today!

Good luck!