The Fragile Families Challenge

Predictability of family and child well-being in adolescence

Matthew J. Salganik, Ian Lundberg, Alex Kindel, Sara S. McLanahan,
and participants in the Fragile Families Challenge

Princeton University
(with collaborators from many institutions)

Aug. 12, 2018
Annual Meeting of the American Sociological Association

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, & Duncan Watts. Source for these slides: www.github.com/fragilefamilieschallenge.

The Fragile Families Challenge

Predictability of family and child well-being in adolescence

Matthew J. Salganik, Ian Lundberg, Alex Kindel, Sara S. McLanahan,
and participants in the Fragile Families Challenge

Princeton University
(with collaborators from many institutions)

Aug. 12, 2018
Annual Meeting of the American Sociological Association

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, & Duncan Watts. Source for these slides: www.github.com/fragilefamilieschallenge.

The Fragile Families Challenge

↓

Predictability of family and child well-being in adolescence

Matthew J. Salganik, Ian Lundberg, Alex Kindel, Sara S. McLanahan,
and participants in the Fragile Families Challenge

Princeton University
(with collaborators from many institutions)

Aug. 12, 2018
Annual Meeting of the American Sociological Association

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, & Duncan Watts. Source for these slides: www.github.com/fragilefamilieschallenge.

Mobility research can be framed as a **prediction** task.

$$Y = \mathrm{E}\left(Y \mid \vec{X}\right) + \epsilon$$

Mobility research can be framed as a **prediction** task.

$$Y = \mathrm{E}\left(Y \mid \vec{X}\right) + \epsilon$$

Attainment

Mobility research can be framed as a **prediction** task.

$$Y = \mathrm{E}\left(Y \mid \vec{X}\right) + \epsilon$$

Attainment

- Academic achievement
- Occupation
- Income

Mobility research can be framed as a **prediction** task.

$$Y = \underbrace{\mathrm{E}\left(Y \mid \vec{X}\right)}_{\textbf{Predictable component}} + \epsilon$$

Attainment

- Academic achievement
- Occupation
- Income

Mobility research can be framed as a **prediction** task.

$$\underbrace{Y}_{\text{Attainment}} = \underbrace{\mathrm{E}\left(Y \mid \vec{X}\right)}_{\text{Predictable component}} + \epsilon$$

Attainment

– Academic achievement

– Occupation

– Income

Predictable component

– Life chances

– Social rigidity

– Stability

Mobility research can be framed as a **prediction** task.

$$Y = \underbrace{\beta_1 X_1 + \beta_2 X_2} + \epsilon$$

$\swarrow$

Attainment

- Academic achievement
- Occupation
- Income

$\downarrow$

Predictable component

- Life chances
- Social rigidity
- Stability

Mobility research can be framed as a **prediction** task.

$$Y = \underbrace{\mathrm{E}\left(Y \mid \vec{X}\right)}_{\text{Predictable component}} + \epsilon$$

Attainment

- Academic achievement
- Occupation
- Income

Predictable component

- Life chances
- Social rigidity
- Stability

Mobility research can be framed as a **prediction** task.

$$\underset{\text{Attainment}}{Y} = \underbrace{\mathrm{E}\left(Y \mid \vec{X}\right)}_{\text{Predictable component}} + \underset{\text{Unpredictable component}}{\epsilon}$$

Attainment

- Academic achievement
- Occupation
- Income

Predictable component

- Life chances
- Social rigidity
- Stability

Unpredictable component

Mobility research can be framed as a **prediction** task.

$$Y = \underbrace{\mathrm{E}\left(Y \mid \vec{X}\right)} + \epsilon$$

Attainment

- Academic achievement
- Occupation
- Income

Predictable component

- Life chances
- Social rigidity
- Stability

Unpredictable component

- Mobility
- Social fluidity
- Volatility

Mobility research can be framed as a **prediction** task.

$$\underbrace{Y}_{\textbf{Attainment}} = \underbrace{\mathrm{E}\left(Y \mid \vec{X}\right)}_{\textbf{Predictable component}} + \underbrace{\epsilon}_{\textbf{Unpredictable component}}$$

Puzzle: Theories focus on the predictable component

Mobility research can be framed as a **prediction** task.

$$Y = \underbrace{\mathrm{E}\left(Y \mid \vec{X}\right)}_{\textbf{Predictable component}} + \epsilon$$

Attainment

Unpredictable component

Puzzle: Theories focus on the predictable component but empirically the unpredictable component dominates.

Mobility research can be framed as a **prediction** task.

$$Y = \underbrace{\mathrm{E}\left(Y \mid \vec{X}\right)} + \epsilon$$

Attainment

Predictable component

Unpredictable component

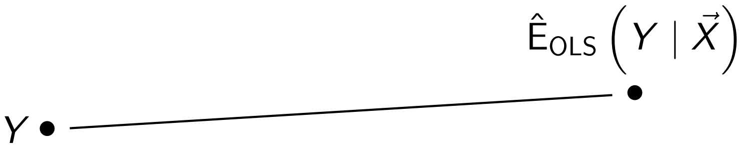
Puzzle: Theories focus on the predictable component but empirically the unpredictable component dominates.

Candidate explanation: Modeling errors

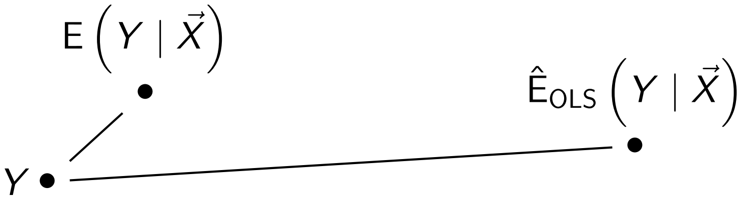
$$\hat{\mathrm{E}}\left(Y \mid \vec{X}\right) \neq \mathrm{E}\left(Y \mid \vec{X}\right).$$

Modeling errors

Modeling errors

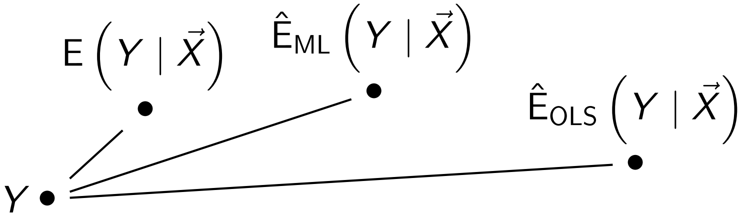


Modeling errors

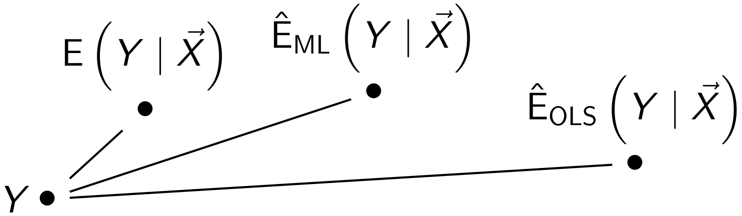


Modeling errors can be minimized by **machine learning**.

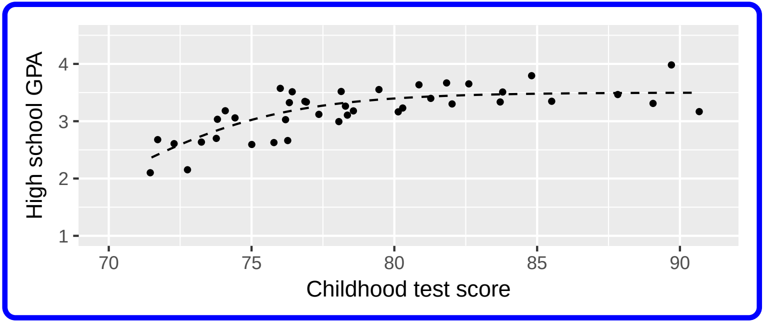
Modeling errors can be minimized by **machine learning**.



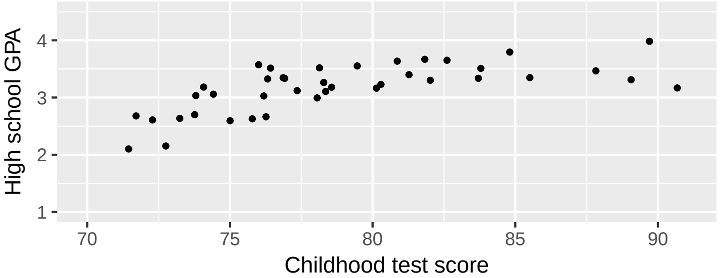
Modeling errors can be minimized by **machine learning**.



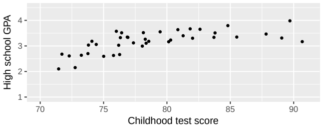
How much does predictability improve when we utilize this **untapped modeling potential**?



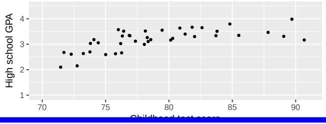
True predictability: $R^2 = 0.54$

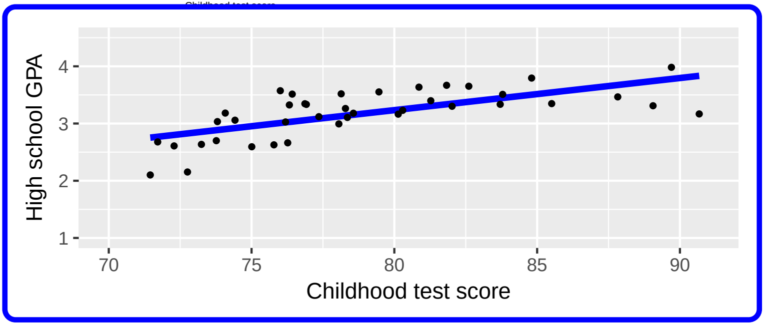

High school GPA
Childhood test score
1
2
3
4
70
75
80
85
90

R^2



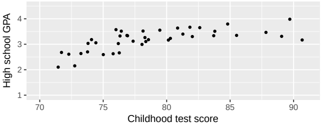
R^2


High school GPA
Childhood test score


High school GPA
4
3
2
1
70
75
80
85
90
Childhood test score

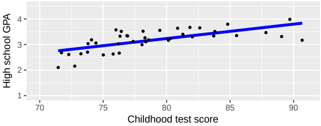
$$R^2 = 0.45$$

R^2

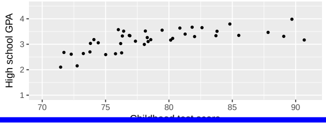


0.45

(estimated)

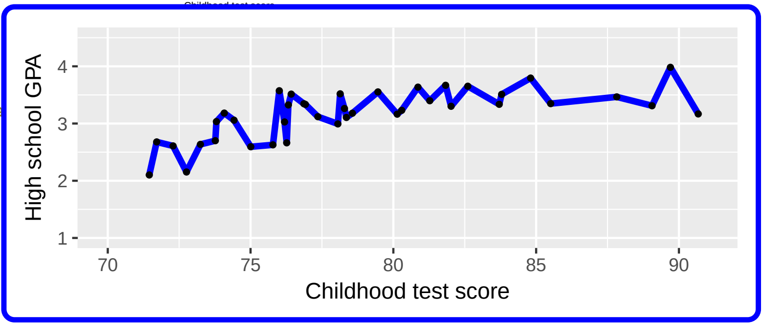


R^2



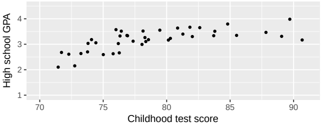
0.4

(estima



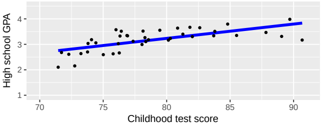
$$R^2 = 1.00$$

R^2



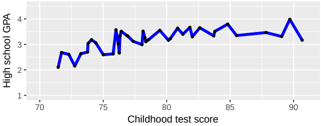
0.45

(estimated)

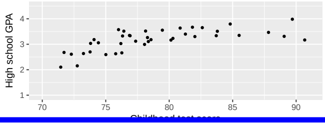


1.00

(estimated)



R^2

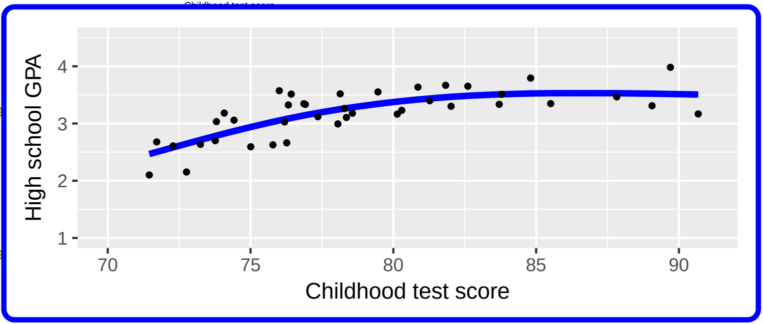


0.4

(estima

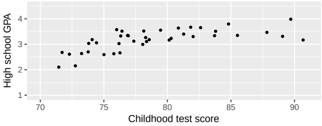
1.0

(estima



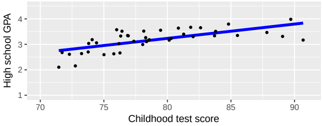
$$R^2 = 0.62$$

R^2



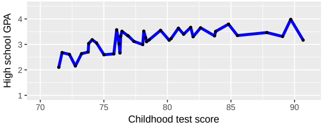
0.45

(estimated)



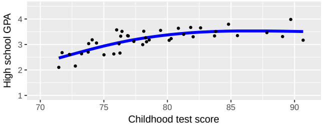
1.00

(estimated)

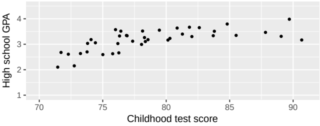


0.62

(estimated)

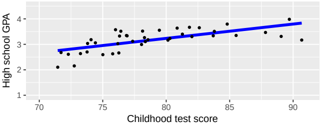


R^2



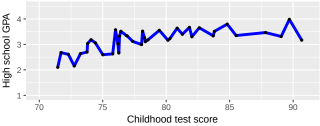
0.45

(estimated)



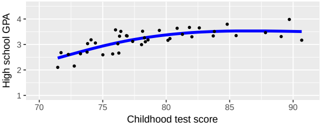
1.00

(estimated)

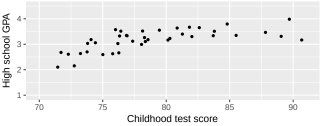


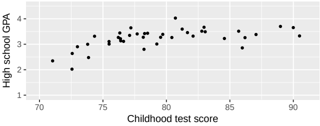
0.62

(estimated)



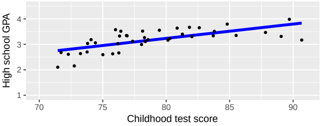
R^2_{Train} Train Holdout R^2_{Holdout}





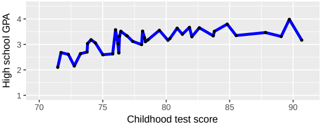
0.45

(estimated)



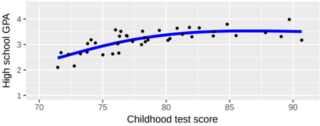
1.00

(estimated)

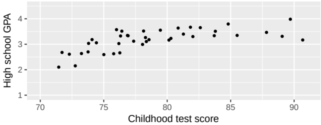


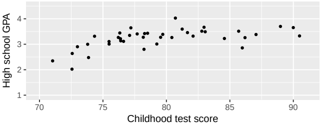
0.62

(estimated)

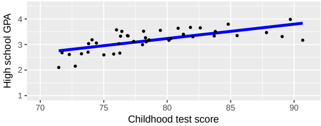


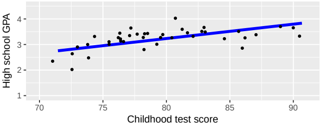
R^2_{Train} | Train | Holdout | R^2_{Holdout}





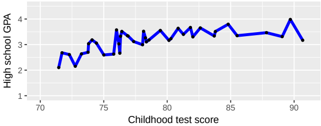
0.45
(estimated)



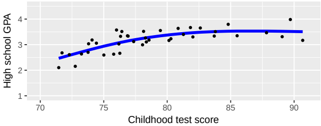


0.30
(estimated)

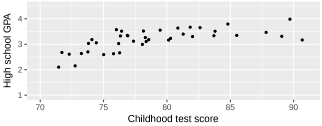
1.00
(estimated)

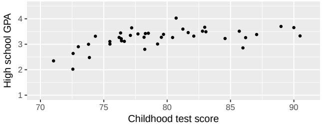


0.62
(estimated)

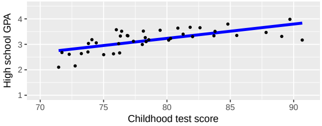


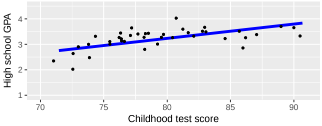
R^2_{Train}	Train	Holdout	R^2_{Holdout}





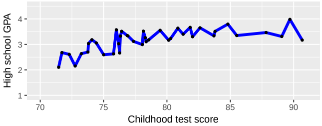
0.45
(estimated)

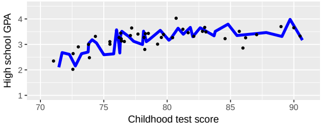




0.30
(estimated)

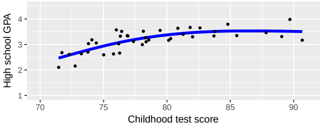
1.00
(estimated)



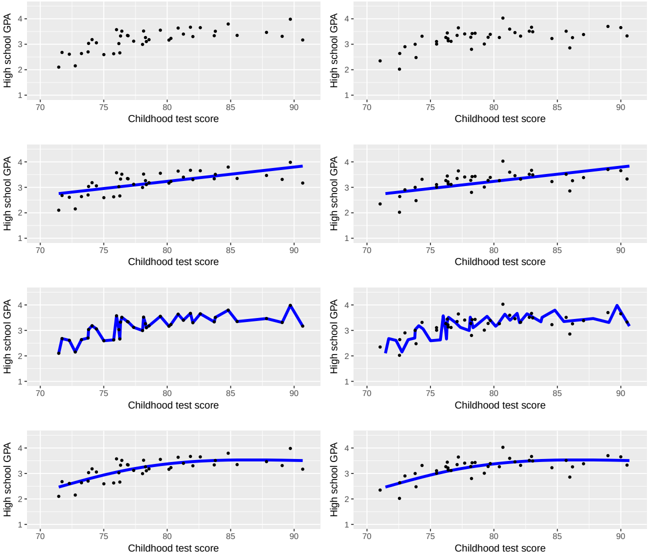


0.13
(estimated)

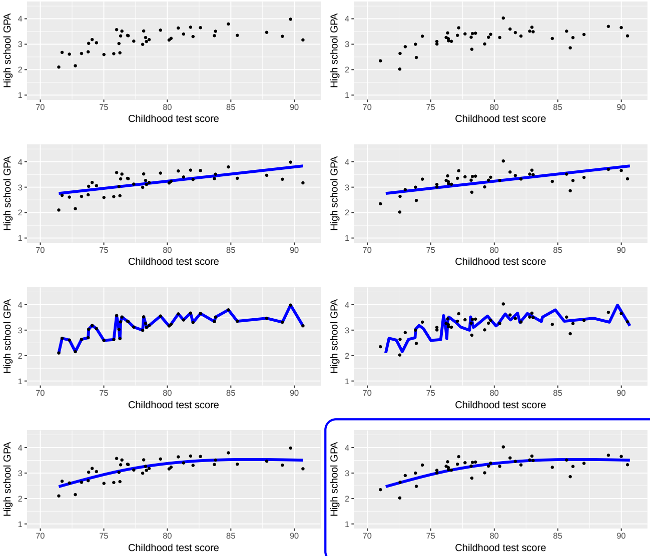
0.62
(estimated)



R^2_{Train}	Train	Holdout	R^2_{Holdout}
0.45 (estimated)			0.30 (estimated)
1.00 (estimated)			0.13 (estimated)
0.62 (estimated)			0.52 (estimated)



R^2_{Train}	Train	Holdout	R^2_{Holdout}
0.45 (estimated)			0.30 (estimated)
1.00 (estimated)			0.13 (estimated)
0.62 (estimated)			0.52 (estimated)



Machine learning provides
a principled framework
for **model selection**

Machine learning provides a principled framework for **model selection** $\longrightarrow$ Predictive performance in a **held-out sample**

Machine learning provides
a principled framework
for **model selection**

→

Predictive performance
in a
held-out sample

Social science
defines the problem

Machine learning provides a principled framework for **model selection** → Predictive performance in a **held-out sample**

Social science **defines the problem** → Machine learning finds an **optimal solution**

Machine learning provides a principled framework for **model selection** → Predictive performance in a **held-out sample**

Social science **defines the problem** → Machine learning finds an **optimal solution**

First example

1 predictor

40 observations

3 participants

Machine learning provides a principled framework for **model selection** → Predictive performance in a **held-out sample**

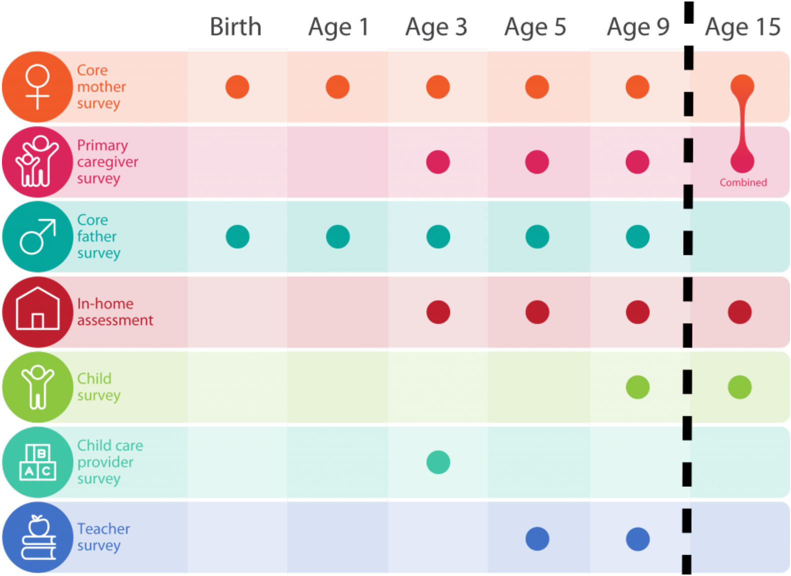
Social science **defines the problem** → Machine learning finds an **optimal solution**

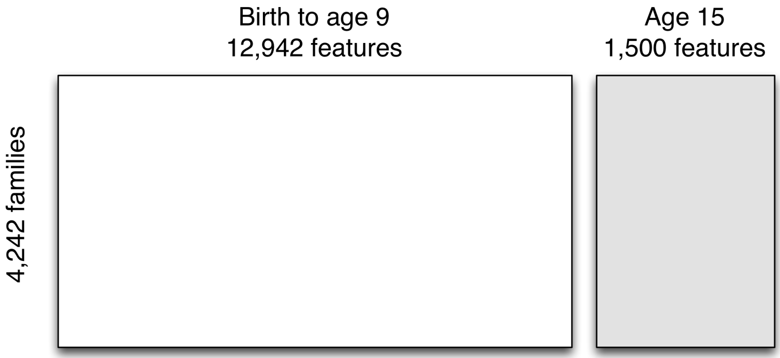
First example		**Fragile Families Challenge**
1 predictor	→	12,942 predictors
40 observations	→	2,121 observations
3 participants	→	441 participants

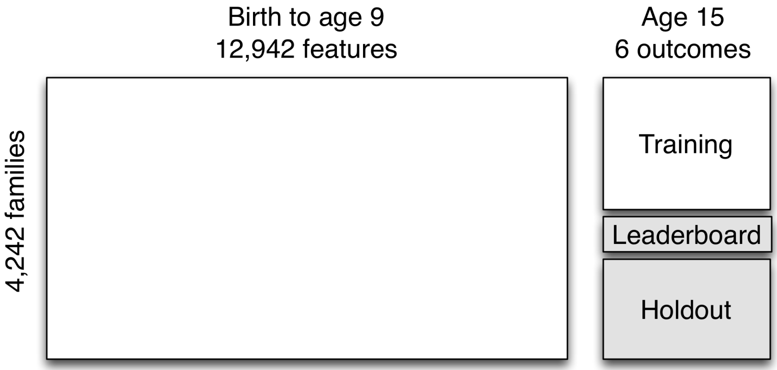
Fragile Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

- 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


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


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

Six age 15 outcomes:

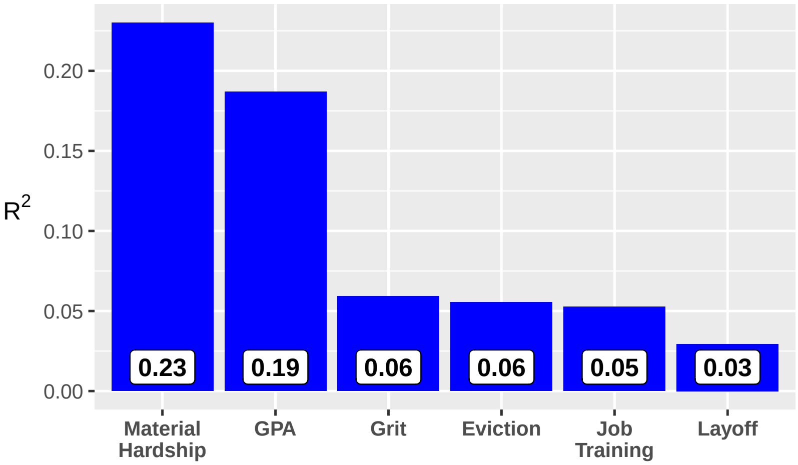
- GPA
- Material Hardship
- Grit
- Evicted
- Job training
- Job loss

441 registered participants

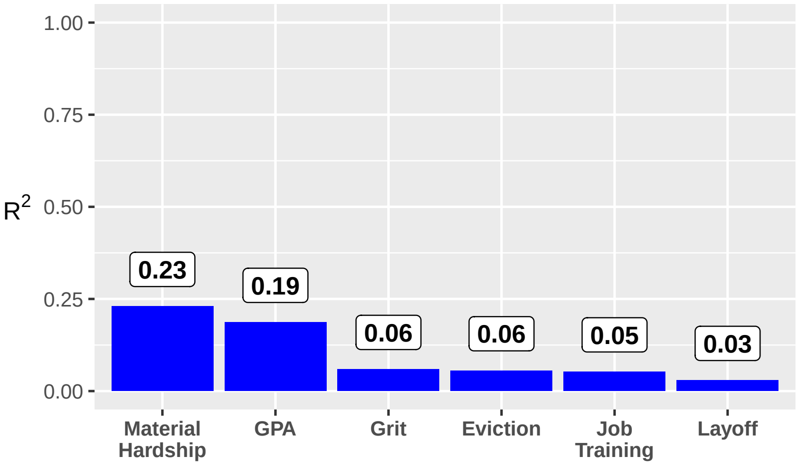
- social scientists and data scientists
- undergraduates, grad students, and professionals
- many working in teams

How did they do?

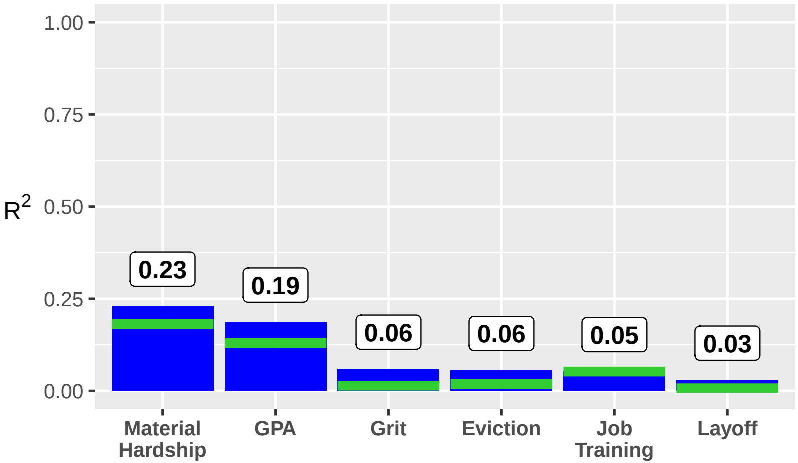
Before I show you, let's vote . . .



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



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



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

What we learned

Hundreds of teams tried many modeling strategies.
Predictions were poor.
That's the best we could do.

What we learned

Hundreds of teams tried many modeling strategies.
Predictions were poor.
That's the best we could do.

Theory
Focuses on predictability

← **Puzzle** →

Empirics
Unpredictability dominates

What we learned

Hundreds of teams tried many modeling strategies.
Predictions were poor.
That's the best we could do.

Candidate explanation
Untapped modeling potential?

Theory
Focuses on predictability

Empirics
Unpredictability dominates

⟵ Puzzle ⟶

What we learned

Hundreds of teams tried many modeling strategies.
Predictions were poor.
That's the best we could do.

~~**Candidate explanation**~~
~~Untapped modeling potential?~~

Theory
Focuses on predictability

Empirics
Unpredictability dominates

← **Puzzle** →

What we learned

Hundreds of teams tried many modeling strategies.
Predictions were poor.
That's the best we could do.

Theory
Focuses on predictability

Empirics
Unpredictability dominates

⟵ Puzzle ⟶

New candidate explanations

Poor prediction may be attributable to:

New candidate explanations

Poor prediction may be attributable to:

- Measurement error

New candidate explanations

Poor prediction may be attributable to:

- Measurement error
- Data hard to use

New candidate explanations

Poor prediction may be attributable to:

- Measurement error
- Data hard to use $\rightarrow$ metadata (Kindel et al. forthcoming)

New candidate explanations

Poor prediction may be attributable to:

- Measurement error
- Data hard to use $\rightarrow$ metadata (Kindel et al. forthcoming)
- Not enough observations: Estimation error

New candidate explanations

Poor prediction may be attributable to:

- Measurement error
- Data hard to use $\rightarrow$ metadata (Kindel et al. forthcoming)
- Not enough observations: Estimation error
- Unmeasured predictors

New candidate explanations

Poor prediction may be attributable to:

- Measurement error
- Data hard to use $\rightarrow$ metadata (Kindel et al. forthcoming)
- Not enough observations: Estimation error
- Unmeasured predictors $\rightarrow$ qualitative interviews

New candidate explanations

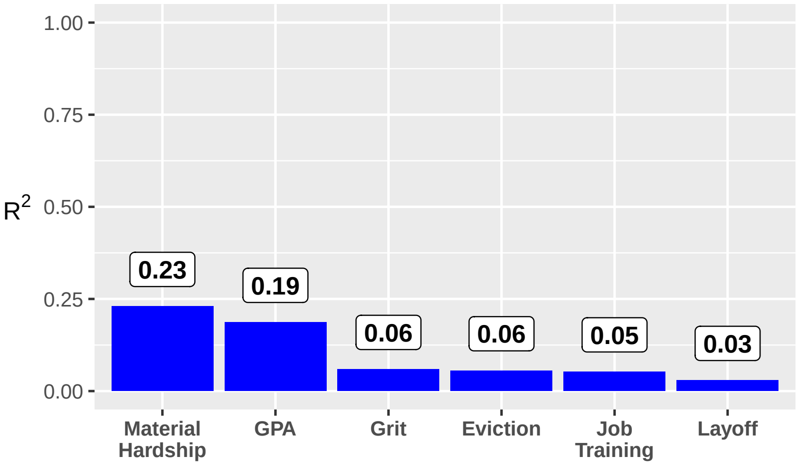
Poor prediction may be attributable to:

- Measurement error
- Data hard to use $\rightarrow$ metadata (Kindel et al. forthcoming)
- Not enough observations: Estimation error
- Unmeasured predictors $\rightarrow$ qualitative interviews
- Unexpected shocks

New candidate explanations

Poor prediction may be attributable to:

- Measurement error
- Data hard to use $\rightarrow$ metadata (Kindel et al. forthcoming)
- Not enough observations: Estimation error
- Unmeasured predictors $\rightarrow$ qualitative interviews
- Unexpected shocks $\rightarrow$ qualitative interviews



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