Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration

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

Department of Sociology & Center for Information Technology Policy, Princeton University

March 4, 2020
Institute for Advanced Study, Workshop on Machine Learning, Theory, and Method in the Social Sciences

Funding for FFCWS provided by NICHD (R01HD36916, R01HD39135, R01HD40421) and a consortium of private foundations, including the Robert Wood Johnson Foundation. Funding for FFC provided by the Russell Sage Foundation, NSF, & the Overdeck Fund. FFC Board of Advisors: Jeanne Brooks-Gunn, Kathryn Edin, Barbara Engelhardt, Irwin Garfinkel, Moritz Hardt, Dean Knox, Nicholas Lemann, Karen Levy, Sara McLanahan, Arvind Narayanan, Timothy Nelson, Matthew Salganik, Brandon Stewart & Duncan Watts.
An overly simple view of stratification research.

\[ Y = E \left( Y \mid \vec{X} \right) + \epsilon \]
An overly simple view of stratification research.

\[ Y = E \left( Y \mid \vec{X} \right) + \epsilon \]

Attainment
An overly simple view of stratification research.

\[ Y = E(Y | \vec{X}) + \epsilon \]

- **Attainment**
  - Academic achievement
  - Occupation
  - Income

Theories focus on the predictable component, but empirically the unpredictable component dominates.
An overly simple view of stratification research.

\[ Y = E \left( \frac{Y}{X} \right) + \epsilon \]

\*Attainment\*
- Academic achievement
- Occupation
- Income

Predictable component

Theories focus on the predictable component, but empirically the unpredictable component dominates.
An overly simple view of stratification research.

\[ Y = E \left( Y \mid \vec{X} \right) + \epsilon \]

- Attainment
  - Academic achievement
  - Occupation
  - Income

Theories focus on the predictable component, but empirically the unpredictable component dominates.
An overly simple view of stratification research.

\[ Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon \]

- **Attainment**
  - Academic achievement
  - Occupation
  - Income

Predictable component
An overly simple view of stratification research.

\[ Y = \mathbb{E} \left( Y \mid \vec{X} \right) + \epsilon \]

- **Attainment**
  - Academic achievement
  - Occupation
  - Income

Theories focus on the predictable component, but empirically the unpredictable component dominates.
An overly simple view of stratification research.

\[ Y = E \left( Y \mid \vec{X} \right) + \epsilon \]

Theories focus on the predictable component, but empirically the unpredictable component dominates.

- Attainment
  - Academic achievement
  - Occupation
  - Income

Unpredictable component
An overly simple view of stratification research.

\[ Y = E \left( Y \mid \vec{X} \right) + \epsilon \]

- **Attainment**
  - Academic achievement
  - Occupation
  - Income

Predictable component

Unpredictable component
An overly simple view of stratification research.

\[ Y = E \left( Y \mid \vec{X} \right) + \epsilon \]

Attainment

\[ \Downarrow \]

Predictable component

Unpredictable component

Theories focus on the predictable component, but empirically the unpredictable component dominates.
Mullainathan and Spiess (2017):
http://dx.doi.org/10.1257/jep.31.2.87
Why should we care about the predictability of social outcomes?
Why should we care about the predictability of social outcomes?

▶ Policy reasons

Can an Algorithm Tell When Kids Are in Danger?

Child protective agencies are haunted when they fail to save kids. Pittsburgh officials believe a new data analysis program is helping them make better judgment calls.

By DAN HURLEY JAN. 2, 2018
Why should we care about the predictability of social outcomes?

- **Policy reasons**

- **Scientific reasons**
Why should we care about the predictability of social outcomes?

- **Policy reasons**

- **Scientific reasons**
  - Basic social fact
  - Discovery
Birth cohort panel study
≈ 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
<table>
<thead>
<tr>
<th></th>
<th>Birth</th>
<th>Age 1</th>
<th>Age 3</th>
<th>Age 5</th>
<th>Age 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core mother survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary caregiver survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core father survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-home assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child care provider survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Birth to age 9
12,942 variables

4,242 families

Information about child and family

Age 15
6 variables

Training
Leaderboard
Holdout

Background data

Outcome data
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

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

More on privacy and ethics audit:
https://arxiv.org/abs/1809.00103
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?

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

Before I show the results, let’s vote . . .
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?

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

Before I show the results, let’s vote . . . .
Material Hardship: 0.23
GPA: 0.19
Grit: 0.06
Eviction: 0.06
Job Training: 0.05
Layoff: 0.03
Is this even better than a benchmark model?
Green line: 4 variable linear regression model
Eviction  

Job training  

Layoff  

Density of predicted probability of event
What can we learn looking at the all the predictions?
Squared error predicting material Hardship

units.to.plot: all, beat.baseline: TRUE, all.legal: FALSE, y−sort: mse.unit.outcome, x−sort: mse.account.outcome

Tue Feb 5 16:15:07 2019
What do these results mean?
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
How can we expand our understanding?
How can we expand our understanding?

In-depth, semi-structured interviews
systematic longitudinal data collection

predictive modeling

in-depth interviews
What’s next?
Next steps:
- One community paper (including all code and predictions)
Next steps:

- One community paper (including all code and predictions)
- Special issue of *Socius*
  - 12 submitted manuscripts from Challenge participants (all with accompanying code and computing environment)
Next steps:

- One community paper (including all code and predictions)
- Special issue of *Socius*
  - 12 submitted manuscripts from Challenge participants (all with accompanying code and computing environment)
  - 3 papers from our group
Next steps:

▶ One community paper (including all code and predictions)
▶ Special issue of *Socius*
  ▶ 12 submitted manuscripts from Challenge participants (all with accompanying code and computing environment)
  ▶ 3 papers from our group
    ▶ “Privacy, ethics, and data access: A case study of the Fragile Families Challenge” by Lundberg, Narayanan, Levy, & Salganik,
Next steps:

- One community paper (including all code and predictions)
- Special issue of *Socius*
  - 12 submitted manuscripts from Challenge participants (all with accompanying code and computing environment)
  - 3 papers from our group
    - “Privacy, ethics, and data access: A case study of the Fragile Families Challenge” by Lundberg, Narayanan, Levy, & Salganik,
    - “Improving metadata infrastructure for complex surveys: Insights from the Fragile Families Challenge” by Kindel, Catena, Hartshorne, Jaeger, Koffman, McLanahan, Phillips, Rouhani, Vinh, & Salganik,
Next steps:

- One community paper (including all code and predictions)
- Special issue of *Socius*
  - 12 submitted manuscripts from Challenge participants (all with accompanying code and computing environment)
  - 3 papers from our group
    - “Privacy, ethics, and data access: A case study of the Fragile Families Challenge” by Lundberg, Narayanan, Levy, & Salganik,
    - “Improving metadata infrastructure for complex surveys: Insights from the Fragile Families Challenge” by Kindel, Catena, Hartshorne, Jaeger, Koffman, McLanahan, Phillips, Rouhani, Vinh, & Salganik,
    - “Successes and struggles with computational reproducibility in the Fragile Families Challenge” by Liu & Salganik
Mass collaborations with longitudinal surveys are an ideal way to study predictability of life trajectories:
Mass collaborations with longitudinal surveys are an ideal way to study predictability of life trajectories:

- Dozens already happening all over the world with interesting similarities and differences
Mass collaborations with longitudinal surveys are an ideal way to study predictability of life trajectories:

- Dozens already happening all over the world with interesting similarities and differences
- Data collected with informed consent under well-developed ethical frameworks
Mass collaborations with longitudinal surveys are an ideal way to study predictability of life trajectories:

- Dozens already happening all over the world with interesting similarities and differences
- Data collected with informed consent under well-developed ethical frameworks
- Already strong community around each survey
Mass collaborations with longitudinal surveys are an ideal way to study predictability of life trajectories:

- Dozens already happening all over the world with interesting similarities and differences
- Data collected with informed consent under well-developed ethical frameworks
- Already strong community around each survey
- Collaboration yields a credible estimate of the best possible predictive performance
Mass collaborations with longitudinal surveys are an ideal way to study predictability of life trajectories:

- Dozens already happening all over the world with interesting similarities and differences
- Data collected with informed consent under well-developed ethical frameworks
- Already strong community around each survey
- Collaboration yields a credible estimate of the best possible predictive performance
- Code from a single Challenge can be repurposed to create many simulated Challenges
\[ \hat{y} \ & \ \hat{\beta} \]