Bit by Bit: Social Research in the Digital Age

Matthew J. Salganik
Department of Sociology
Princeton University

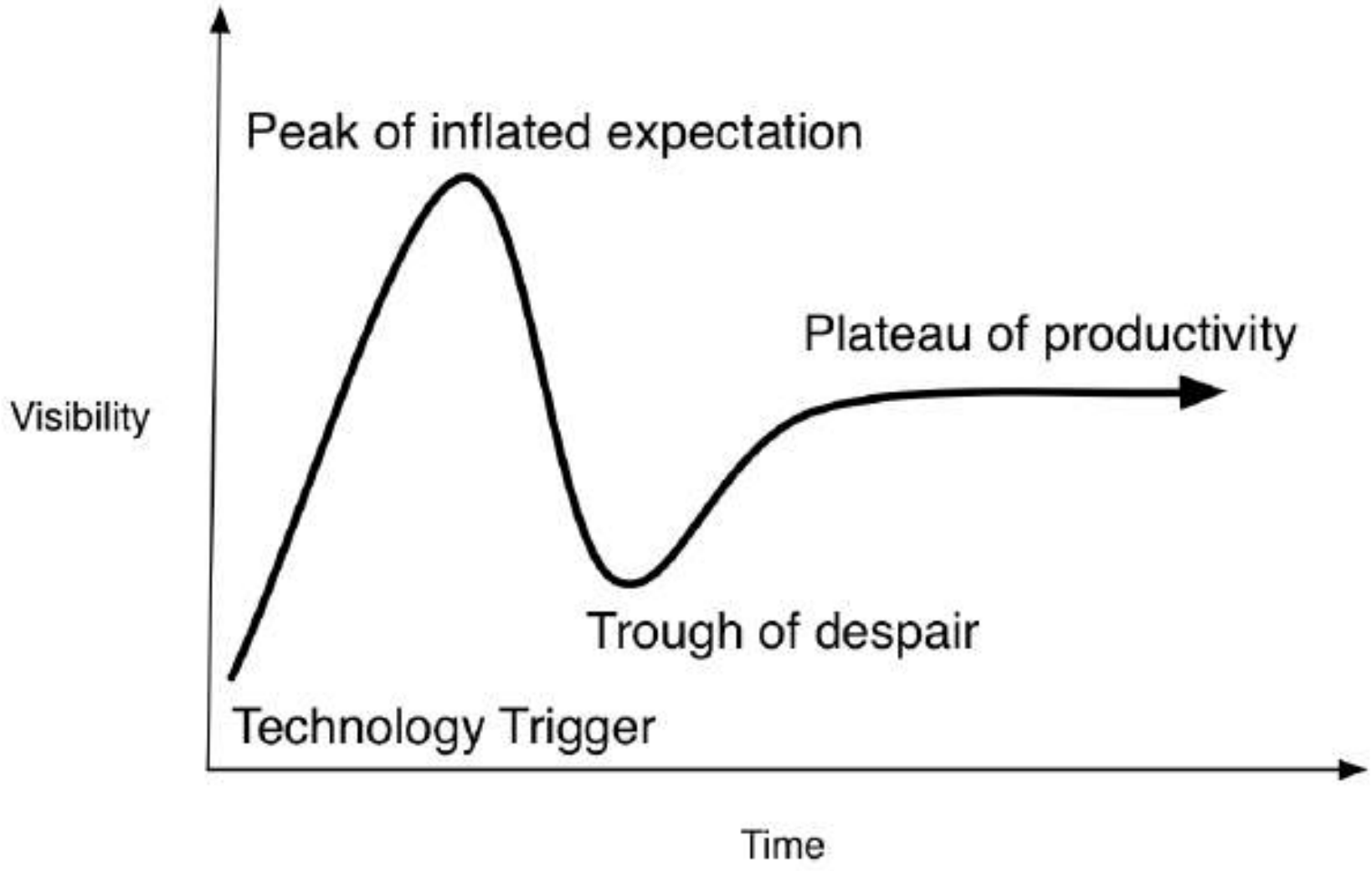
Behave Lab
University of Milan
March 21, 2019


BIT
SOCIAL RESEARCH
BY
in the DIGITAL AGE
BIT
MATTHEW J. SALGANIK

Isn't computational social science a fad?

Isn't computational social science a fad?

No



https://commons.wikimedia.org/wiki/File:Gartner_Hype_Cycle.svg

Social Scientists $\longleftrightarrow$ Data Scientists


R. Mutt
1917


R. MUTT
1917


R. MUTT
1917

Readymades

Custommades

Readymades

+

Custommades

https://commons.wikimedia.org/wiki/File:Duchamp_Fountaine.jpg
https://commons.wikimedia.org/wiki/File:%27David%27_by_Michelangelo_JBU0001.JPG

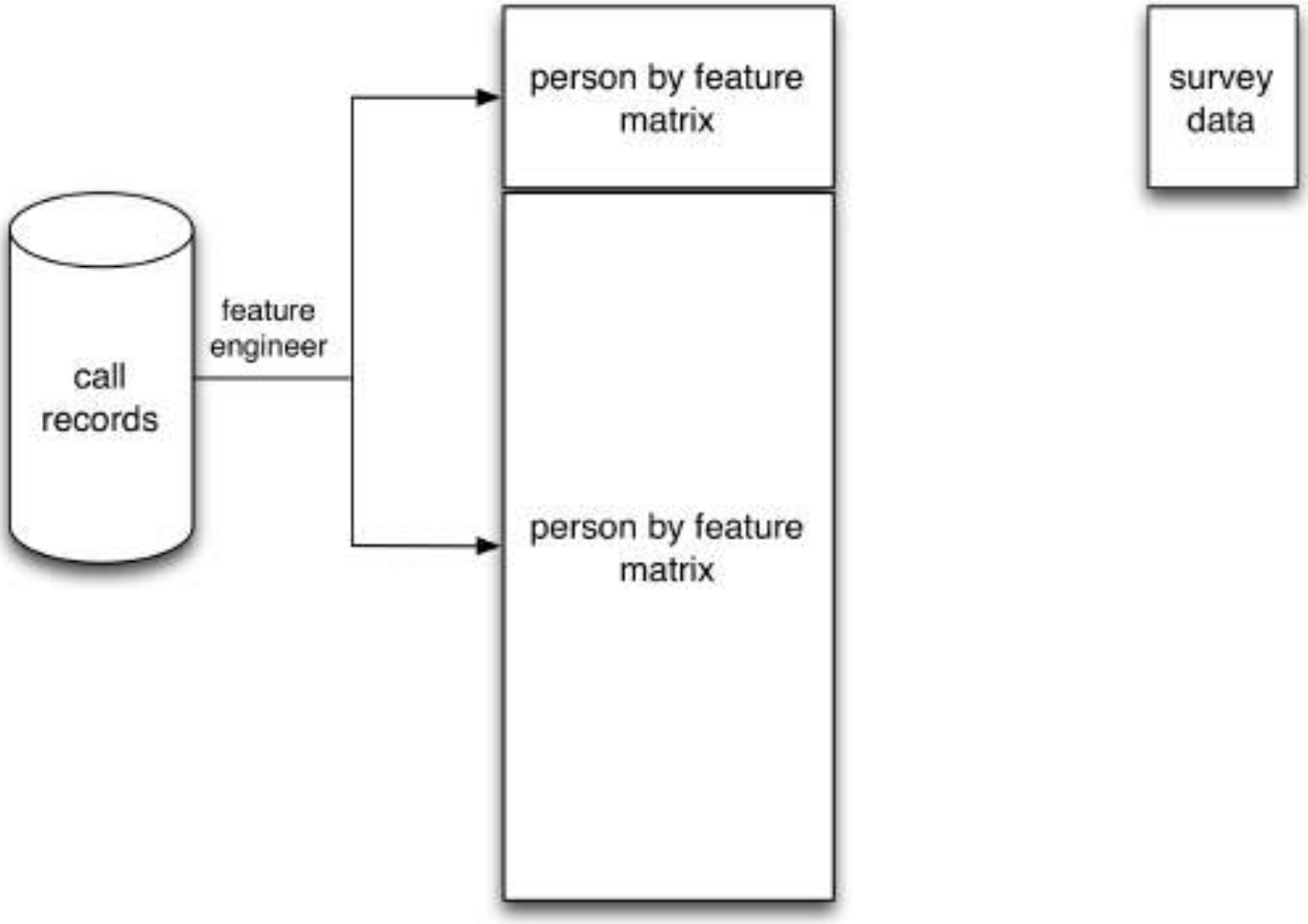
Predicting poverty and wealth from mobile phone metadata

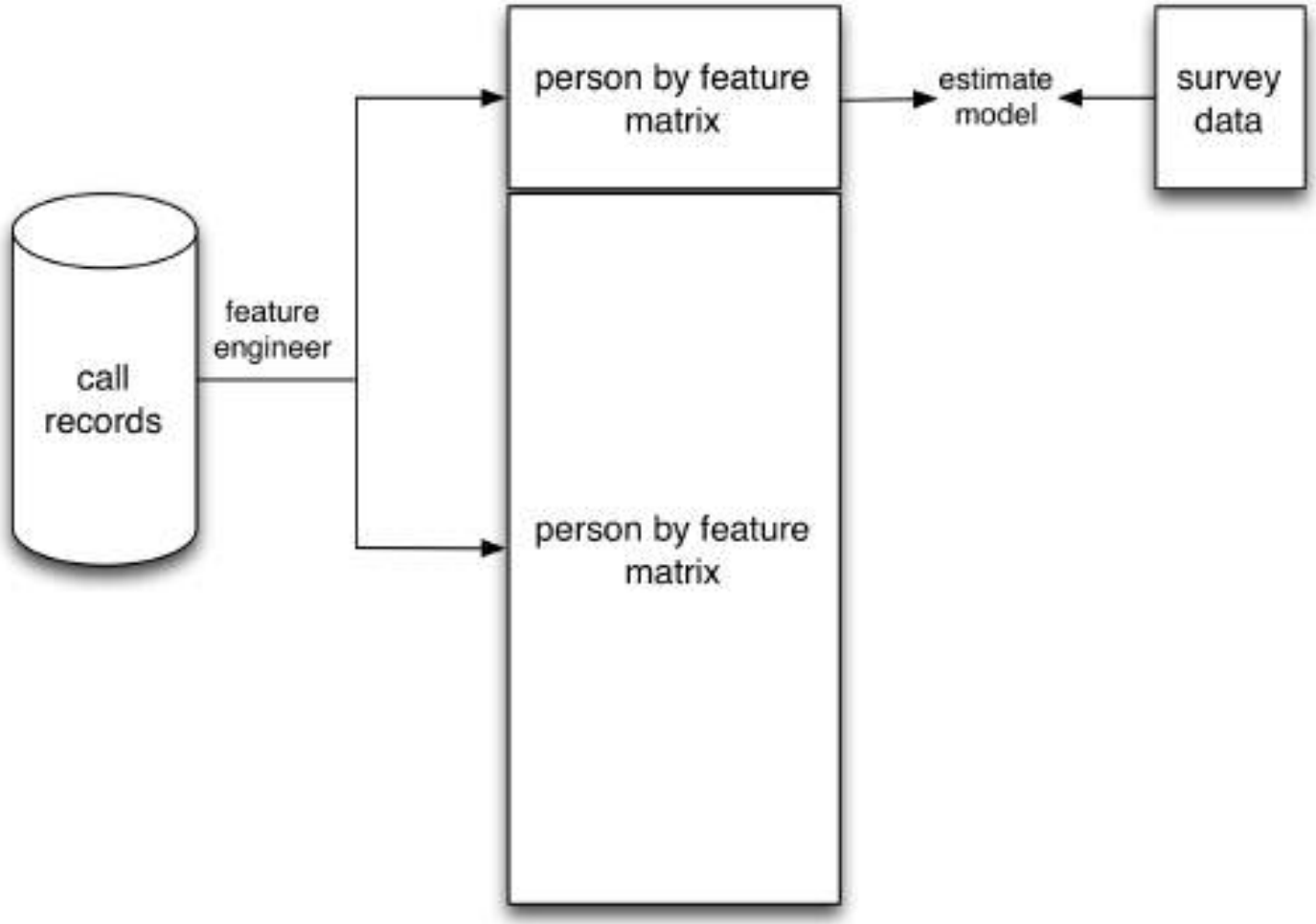
Joshua Blumenstock,[1*] Gabriel Cadamuro,[2] Robert On[3]

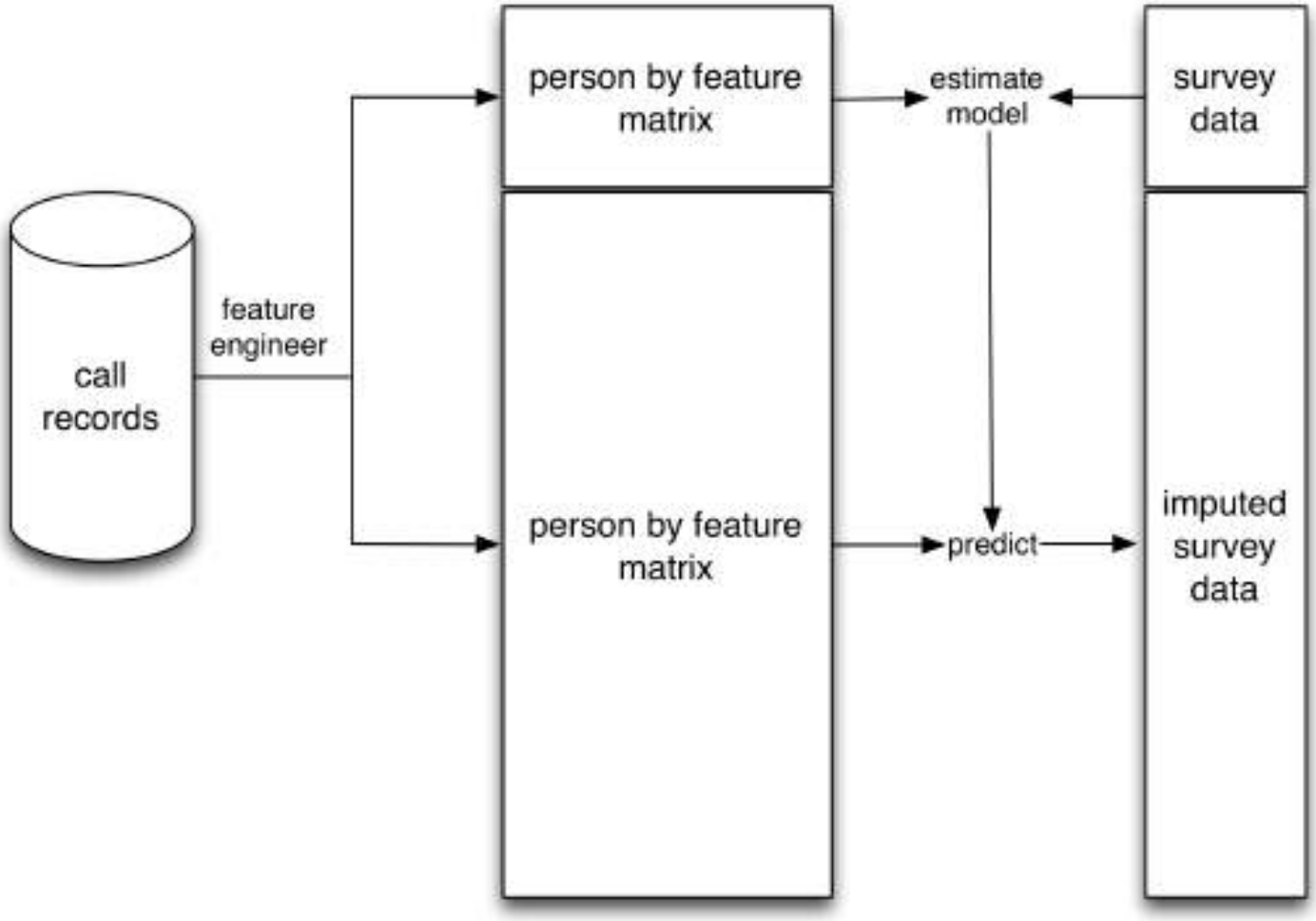
https://doi.org/10.1126/science.aac4420

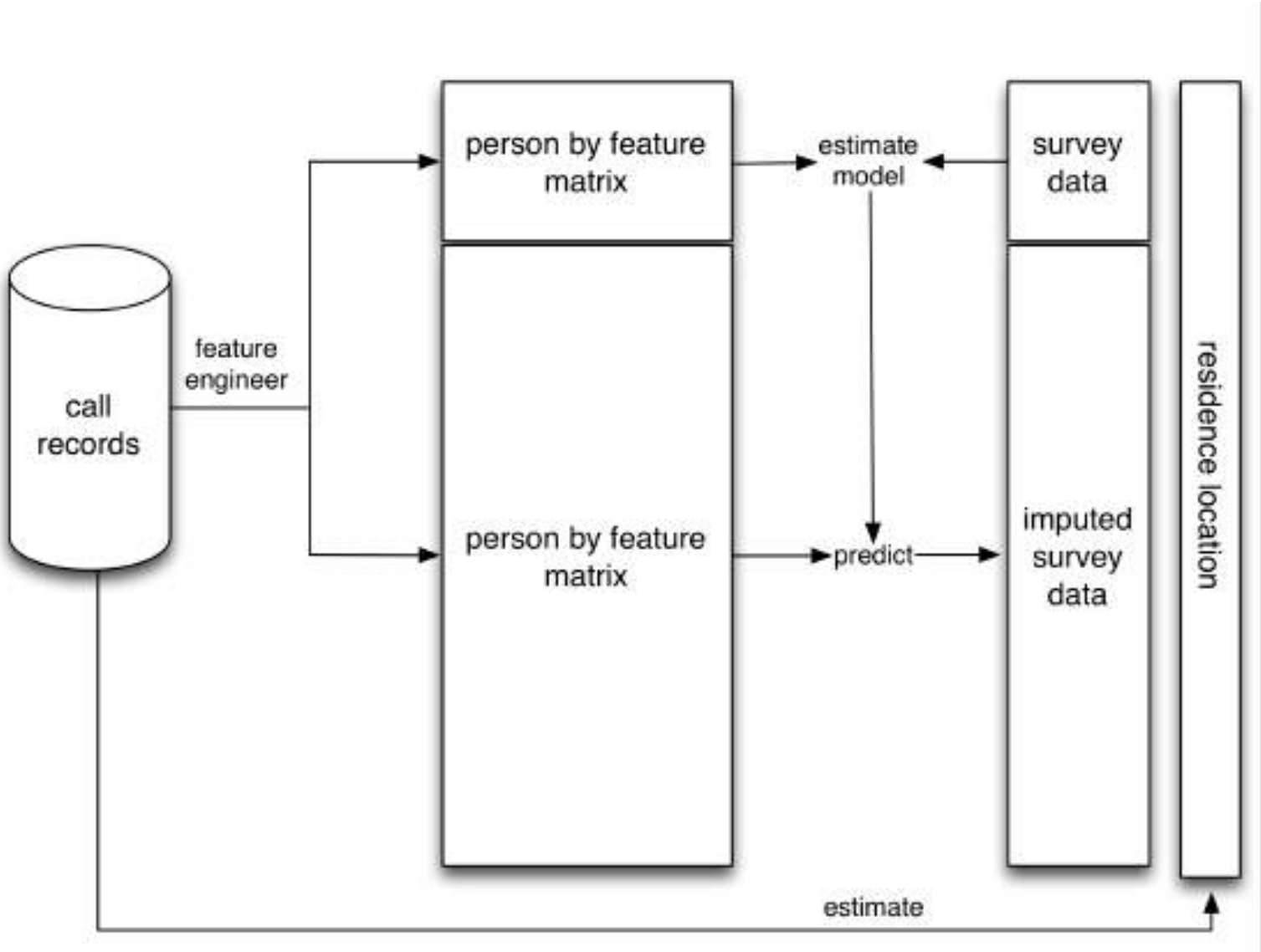

call
records

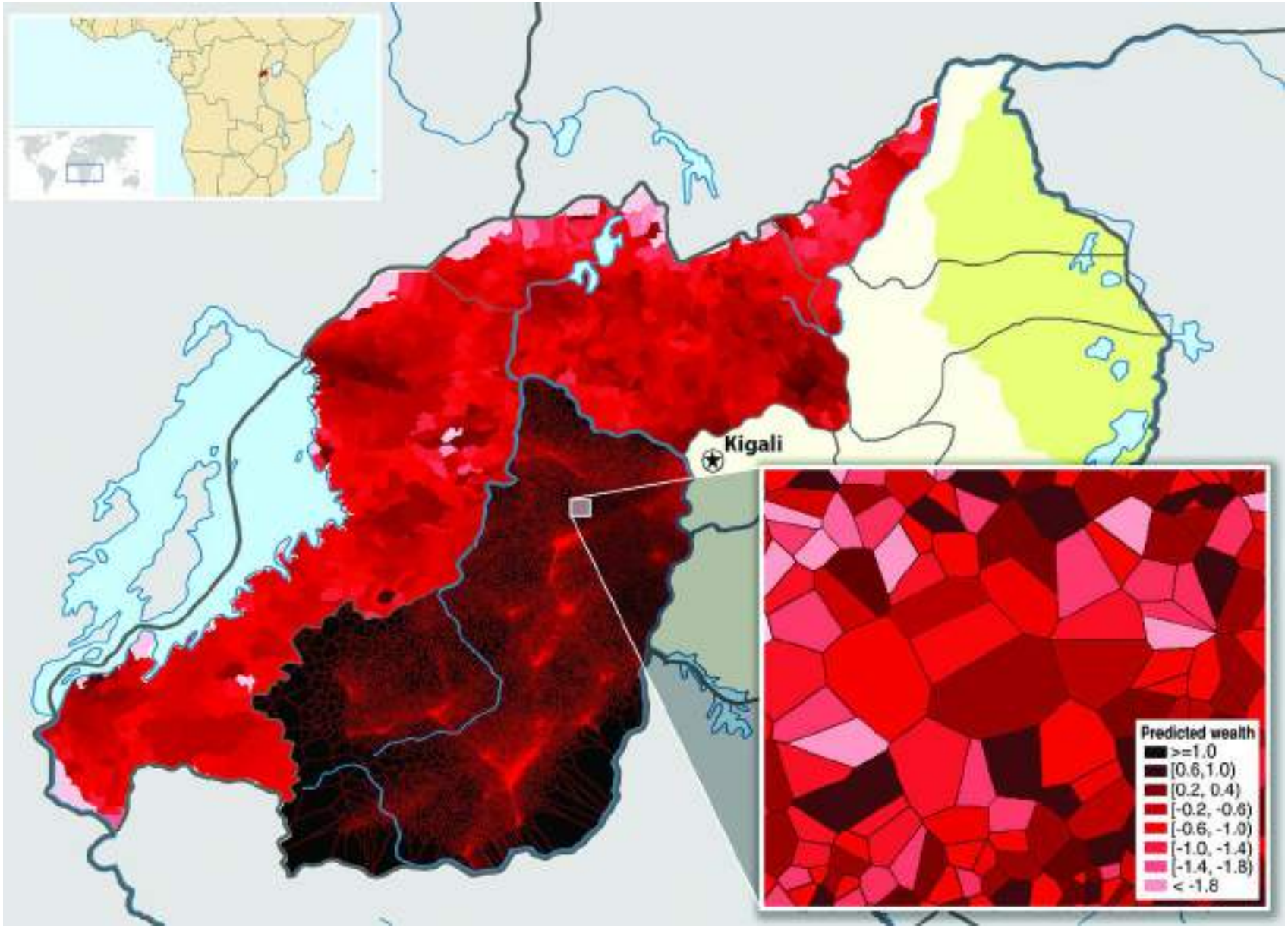
survey
data


person by feature matrix
survey data
feature engineer
call records
person by feature matrix

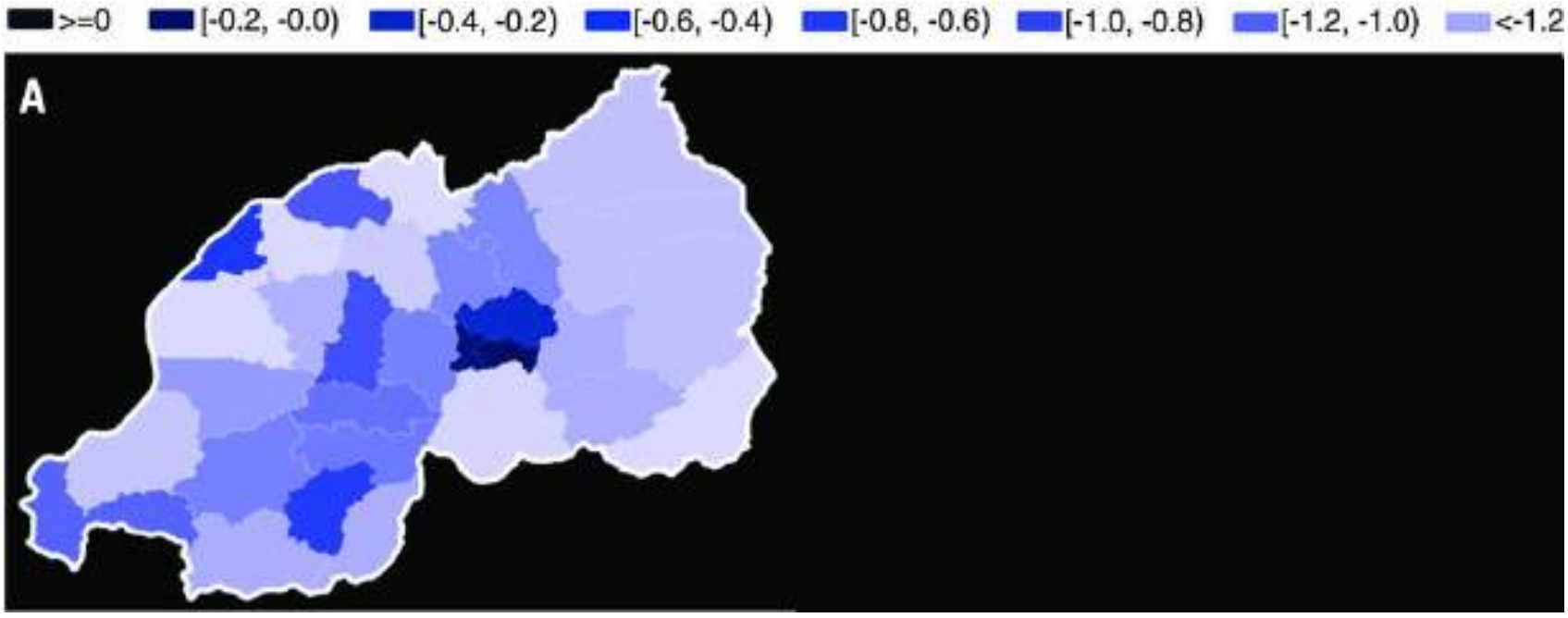

person by feature matrix
estimate model
survey data
call records
feature engineer
person by feature matrix


call
records
feature
engineer
person by feature
matrix
person by feature
matrix
estimate
model
predict
survey
data
imputed
survey
data

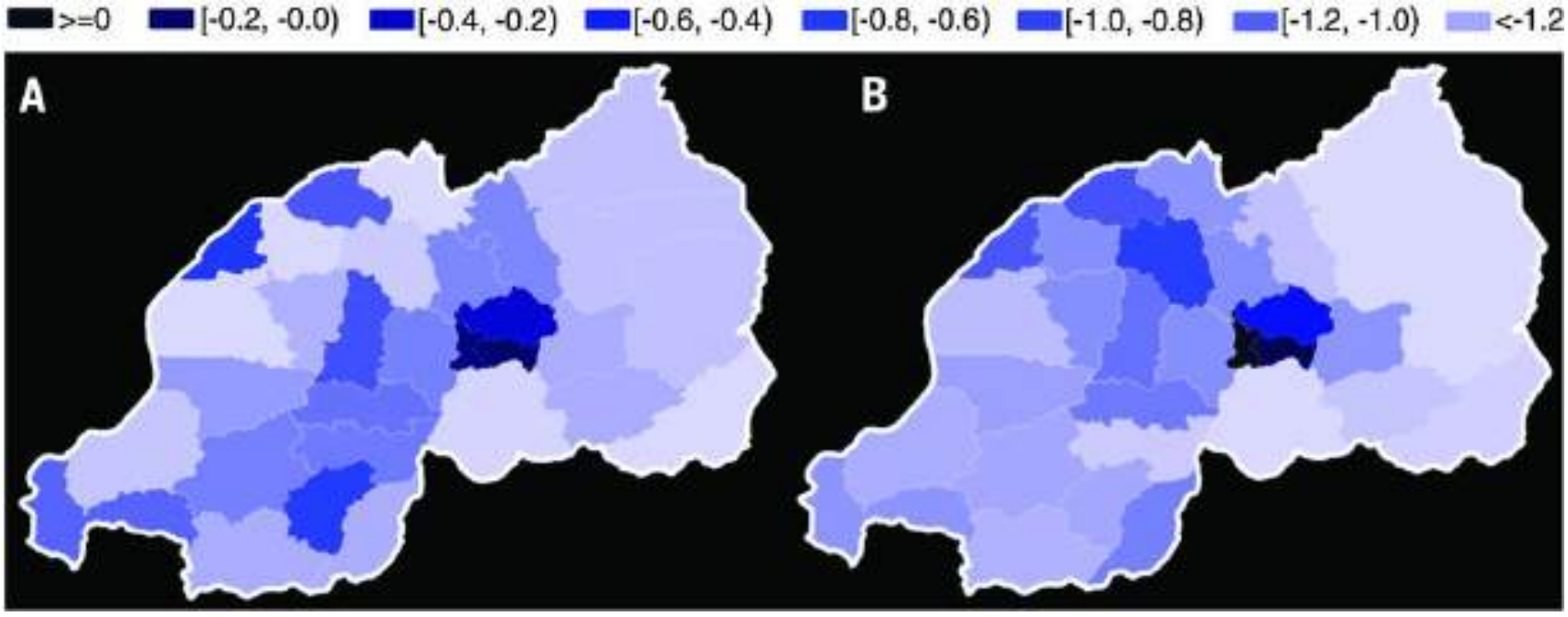

call records
feature engineer
person by feature matrix
person by feature matrix
estimate model
predict
survey data
imputed survey data
residence location
estimate



Blumenstock et al. (2015), Figure 2

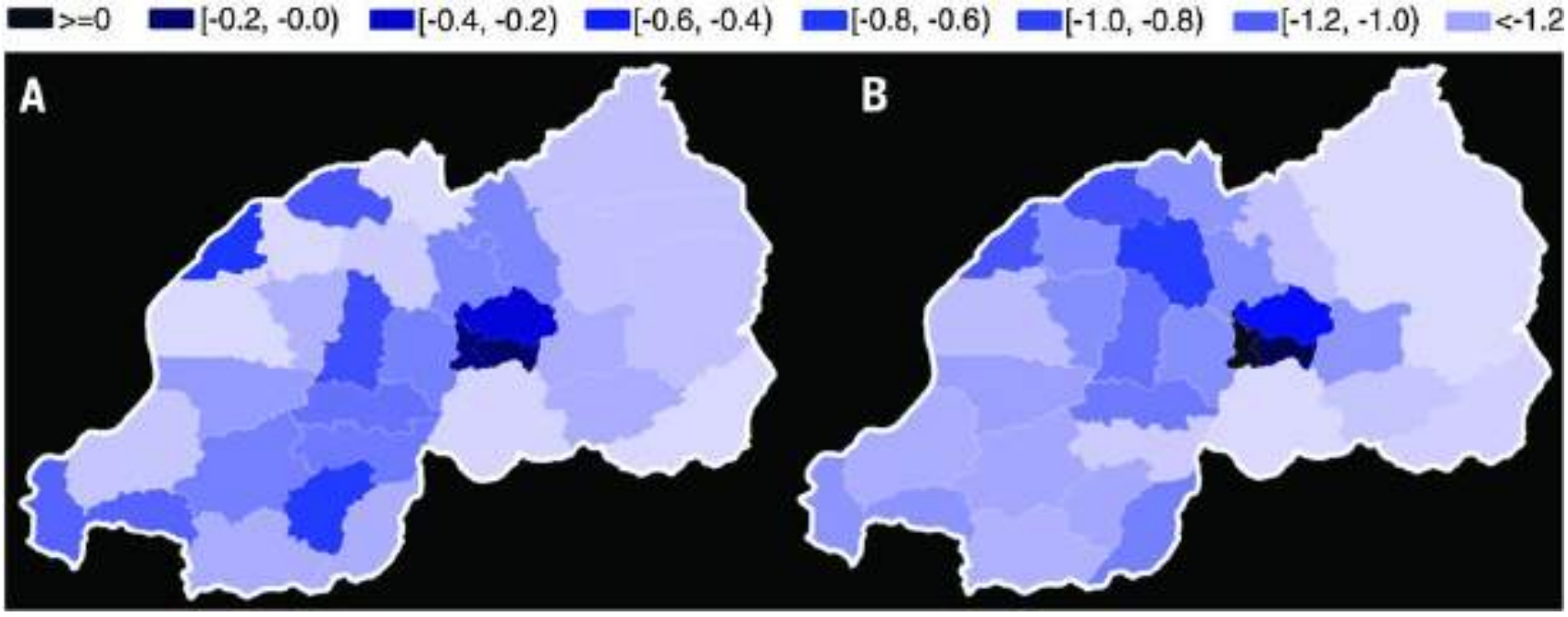


Readymade + Custommade



Readymade + Custommade

Custommade



Readymade + Custommade

Custommade

- 10 times faster
- 50 times cheaper

Blumenstock et al. (2015), Figure 3

Readymades

+

Custommades

https://commons.wikimedia.org/wiki/File:Duchamp_Fountaine.jpg
https://commons.wikimedia.org/wiki/File:%27David%27_by_Michelangelo_JBU0001.JPG

Created by Kedron Rhodes

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

Mullainathan and Spiess (2017):
http://dx.doi.org/10.1257/jep.31.2.87

Fragile Families Challenge

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

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 and 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, and Duncan Watts.

An overly simple view of stratification research.

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

An overly simple view of stratification research.

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

Attainment

An overly simple view of stratification research.

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

Attainment

- Academic achievement
- Occupation
- Income

An overly simple view of stratification research.

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

Attainment

- Academic achievement
- Occupation
- Income

An overly simple view of stratification research.

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

Attainment

- Academic achievement
- Occupation
- Income

An overly simple view of stratification research.

$$Y = \underbrace{\beta_1 X_1 + \beta_2 X_2}_{\text{Predictable component}} + \epsilon$$

Attainment

- Academic achievement
- Occupation
- Income

An overly simple view of stratification research.

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

Attainment

- Academic achievement
- Occupation
- Income

An overly simple view of stratification research.

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

Attainment

- Academic achievement
- Occupation
- Income

Unpredictable component

An overly simple view of stratification research.

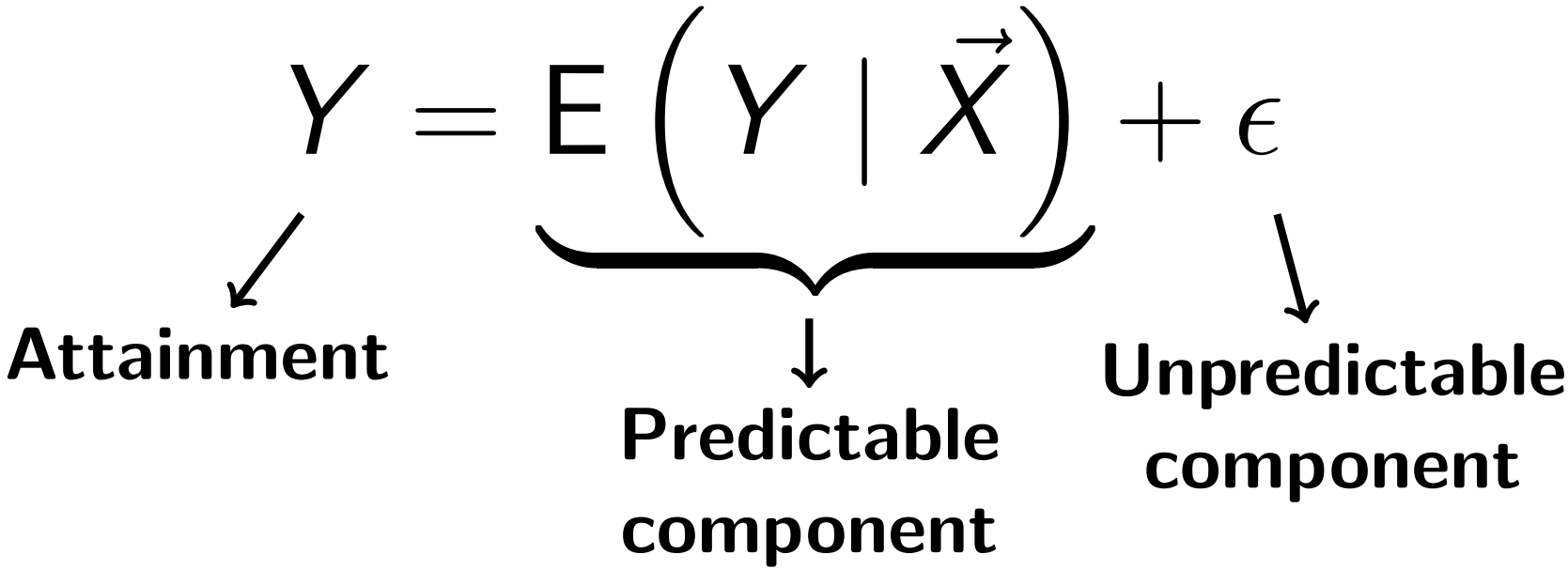
$$Y = \underbrace{\mathrm{E}\left(Y \mid \vec{X}\right)}_{\text{Predictable component}} + \epsilon$$

Attainment

- Academic achievement
- Occupation
- Income

Unpredictable component

An overly simple view of stratification research.



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

Why should we care about the predictability of social outcomes?

Why should we care about the predictability of social outcomes?

- Scientific reasons

Why should we care about the predictability of social outcomes?

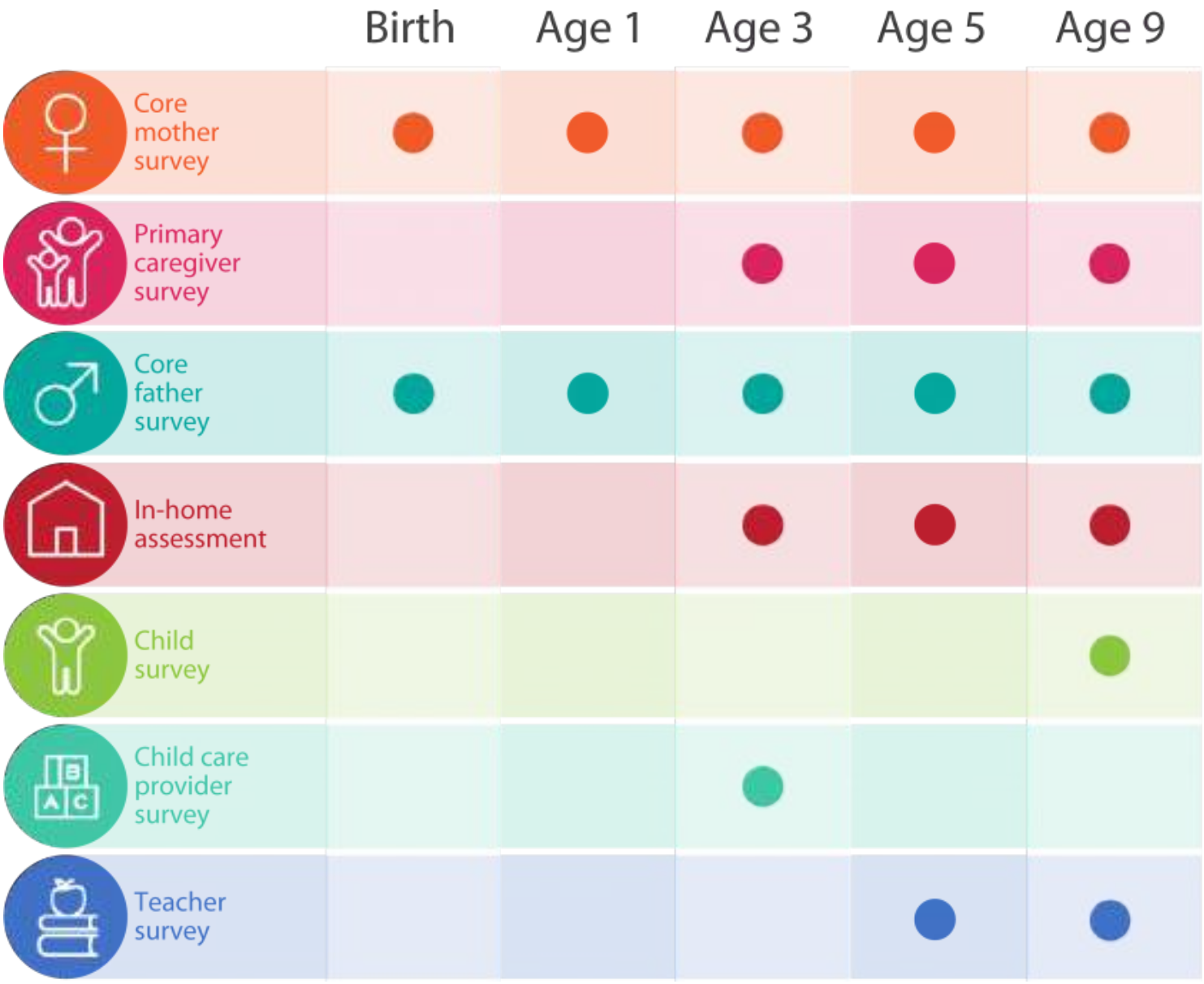
- Scientific reasons
  - Basic social fact
  - Discovery

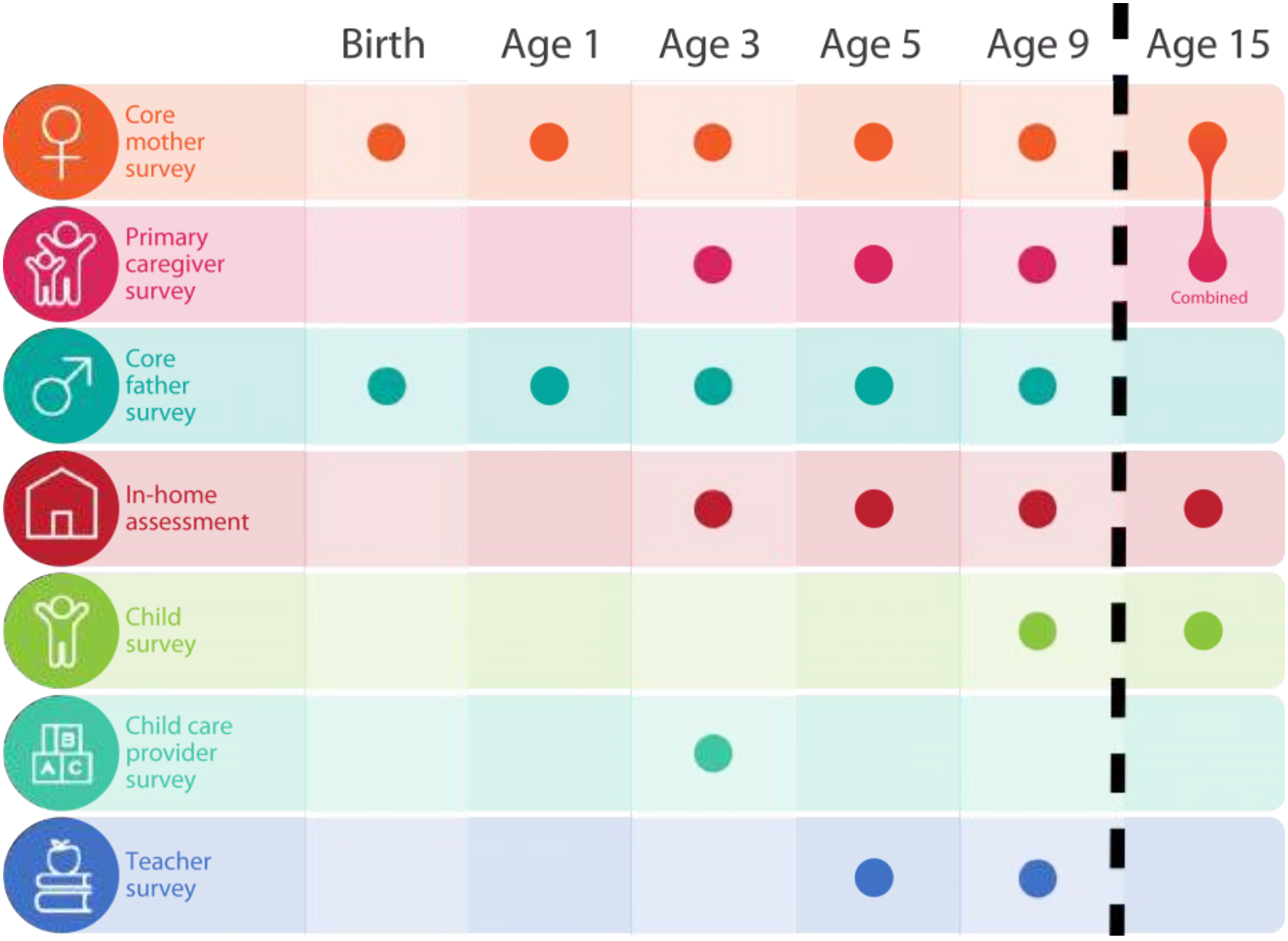
Why should we care about the predictability of social outcomes?

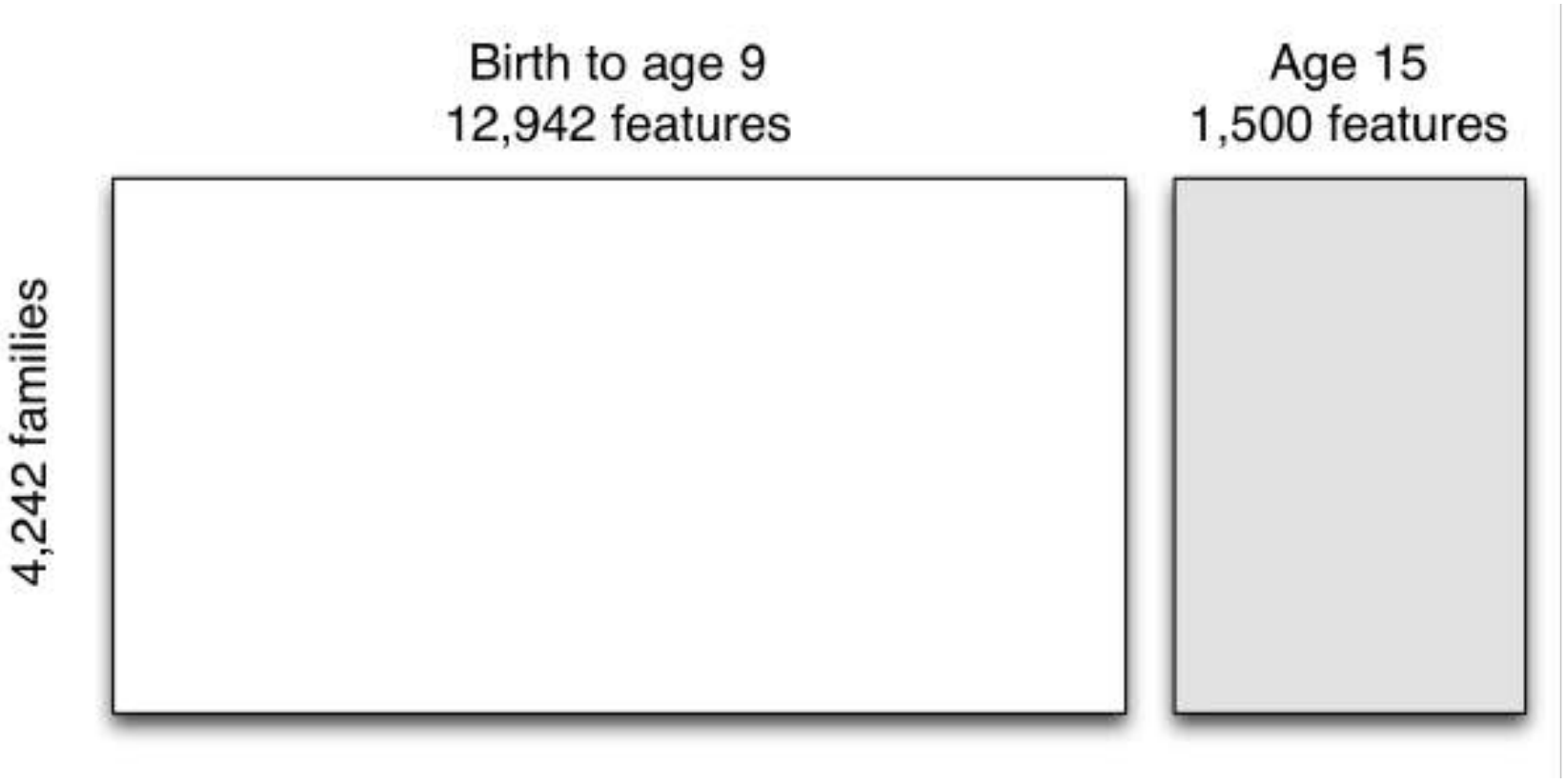
- Scientific reasons
  - Basic social fact
  - Discovery
- Policy reasons

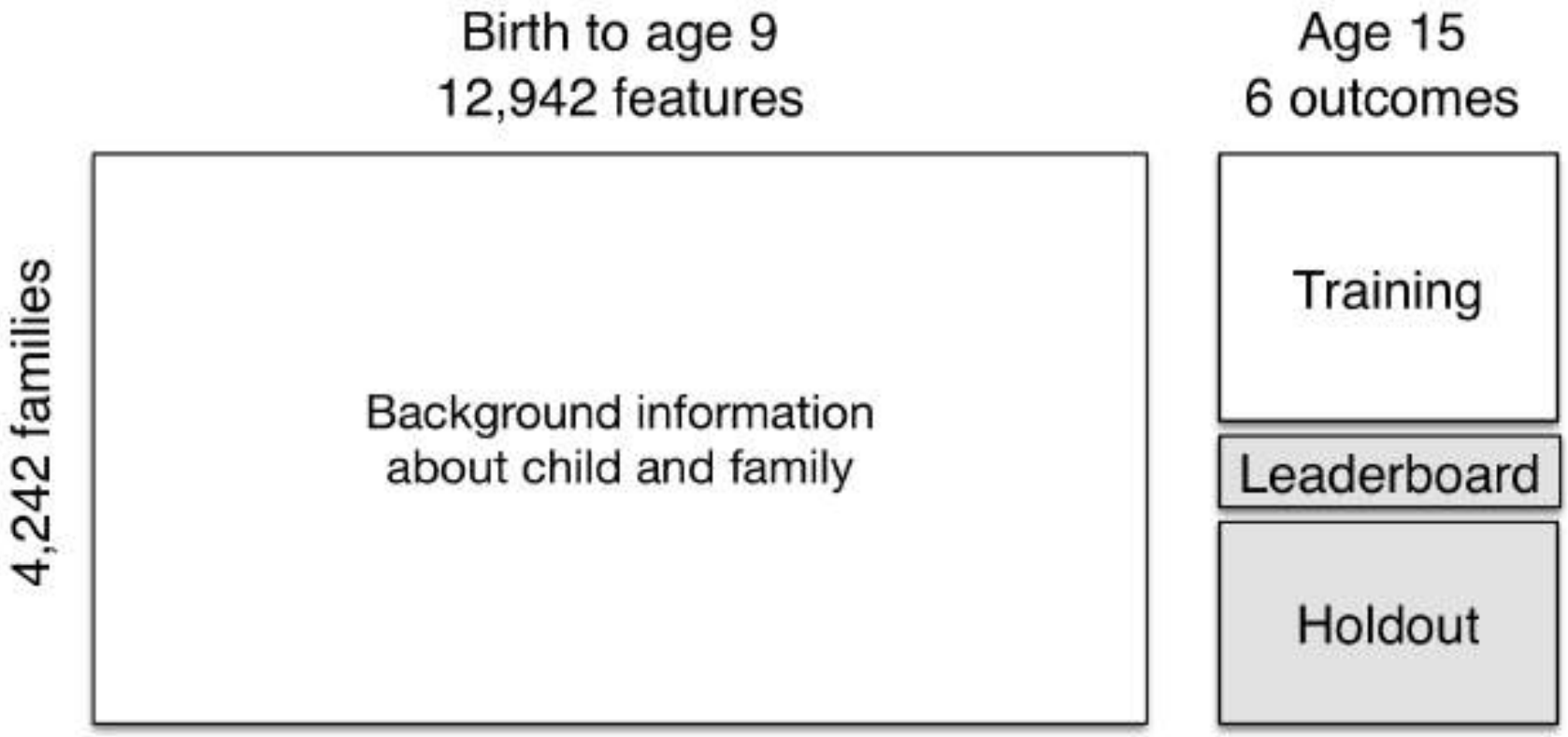
Fragile Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

- 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


Birth to age 9
12,942 features
Age 15
6 outcomes
4,242 families
Background information
about child and family
Training
Leaderboard
Holdout

Outcomes

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

459 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:
`https://arxiv.org/abs/1809.00103`

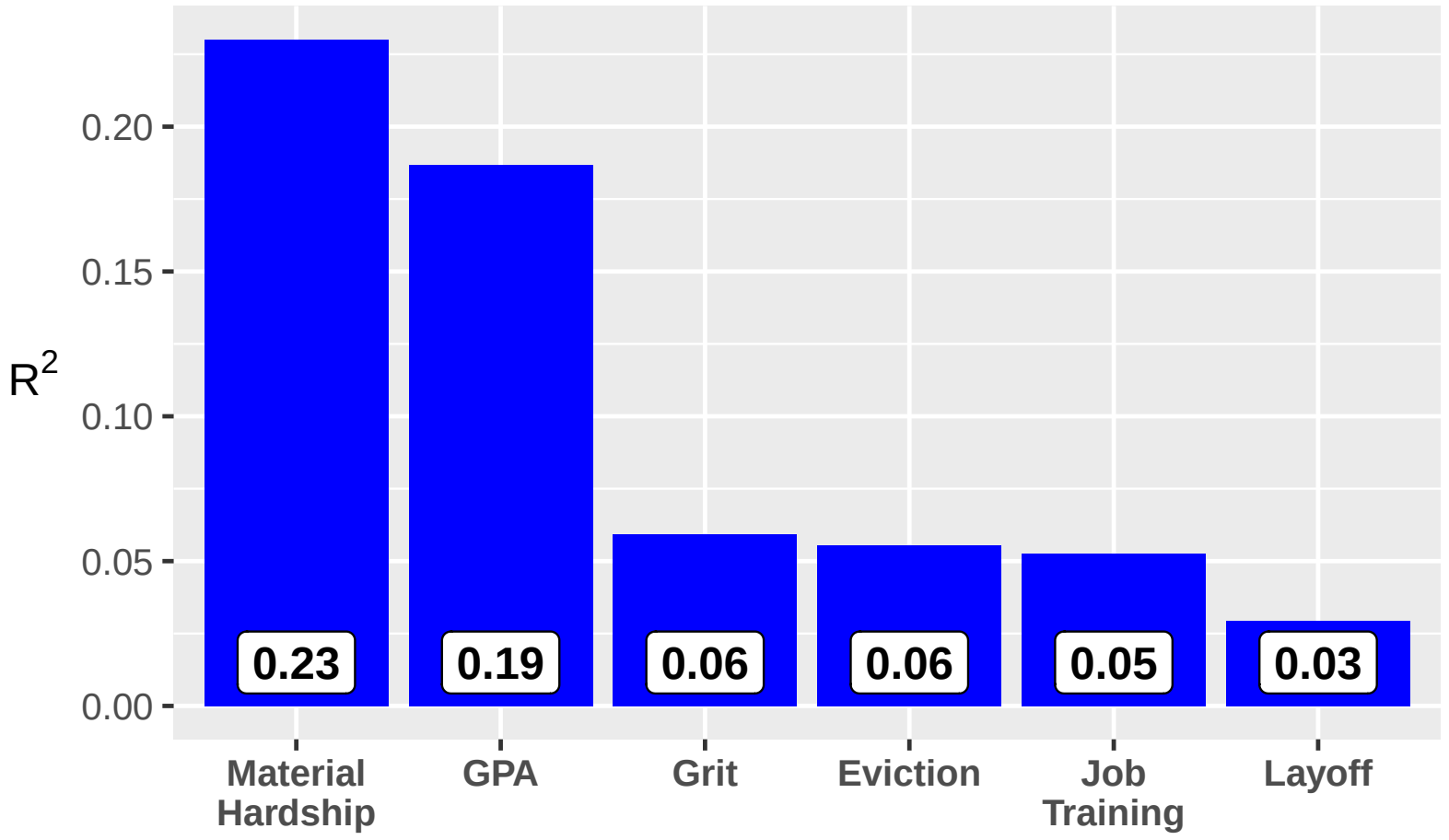
Using a large, high-quality social science dataset collected since birth and modern machine learning methods, how accurately can we predict outcomes from children, parents, and families?

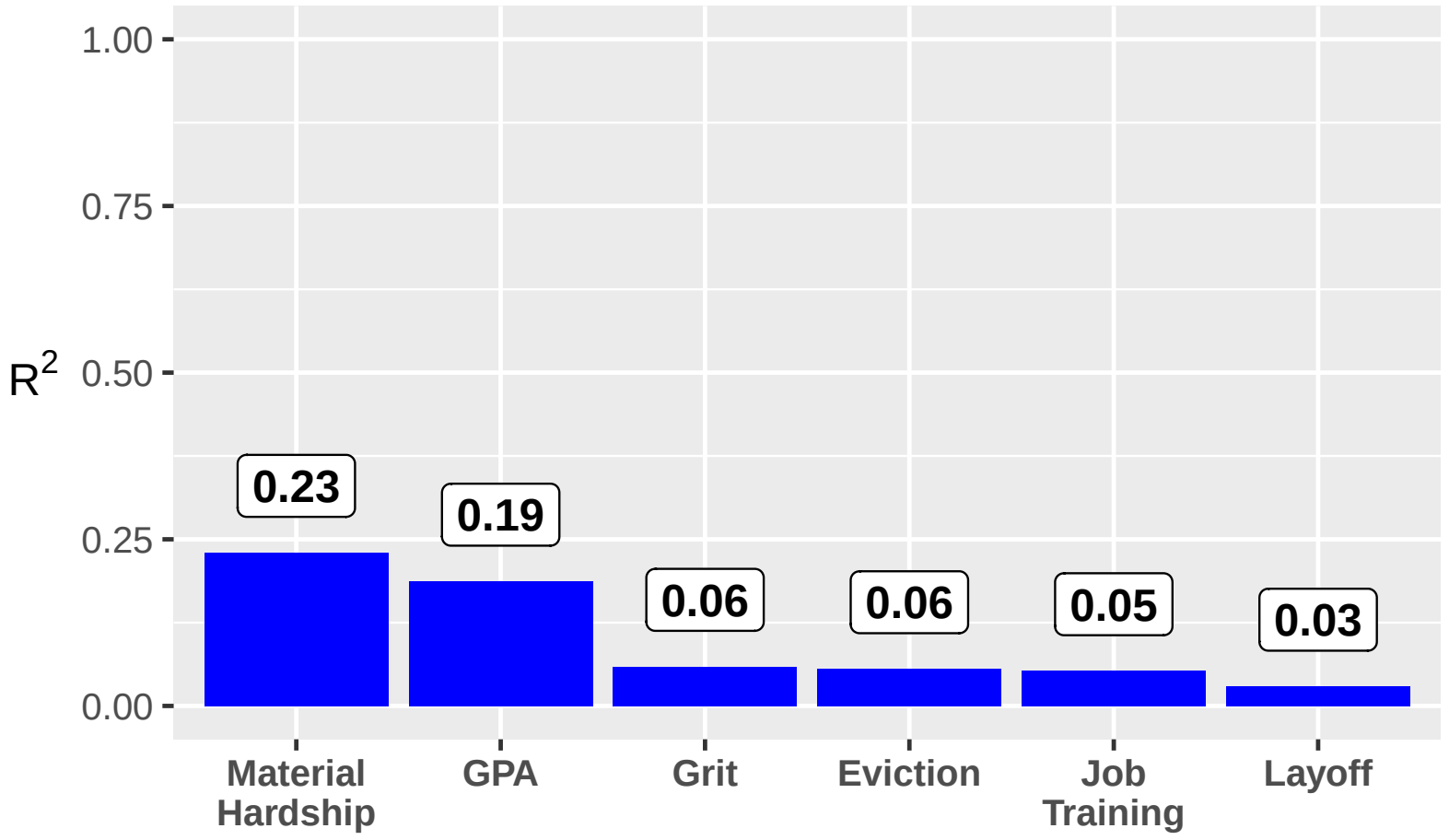
$$R^2_{holdout} = 1 - \frac{\sum_{i \in \text{holdout}} (\hat{y}_i - y_i)^2}{\sum_{i \in \text{holdout}} (\bar{y}_{train} - y_i)^2}$$

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?

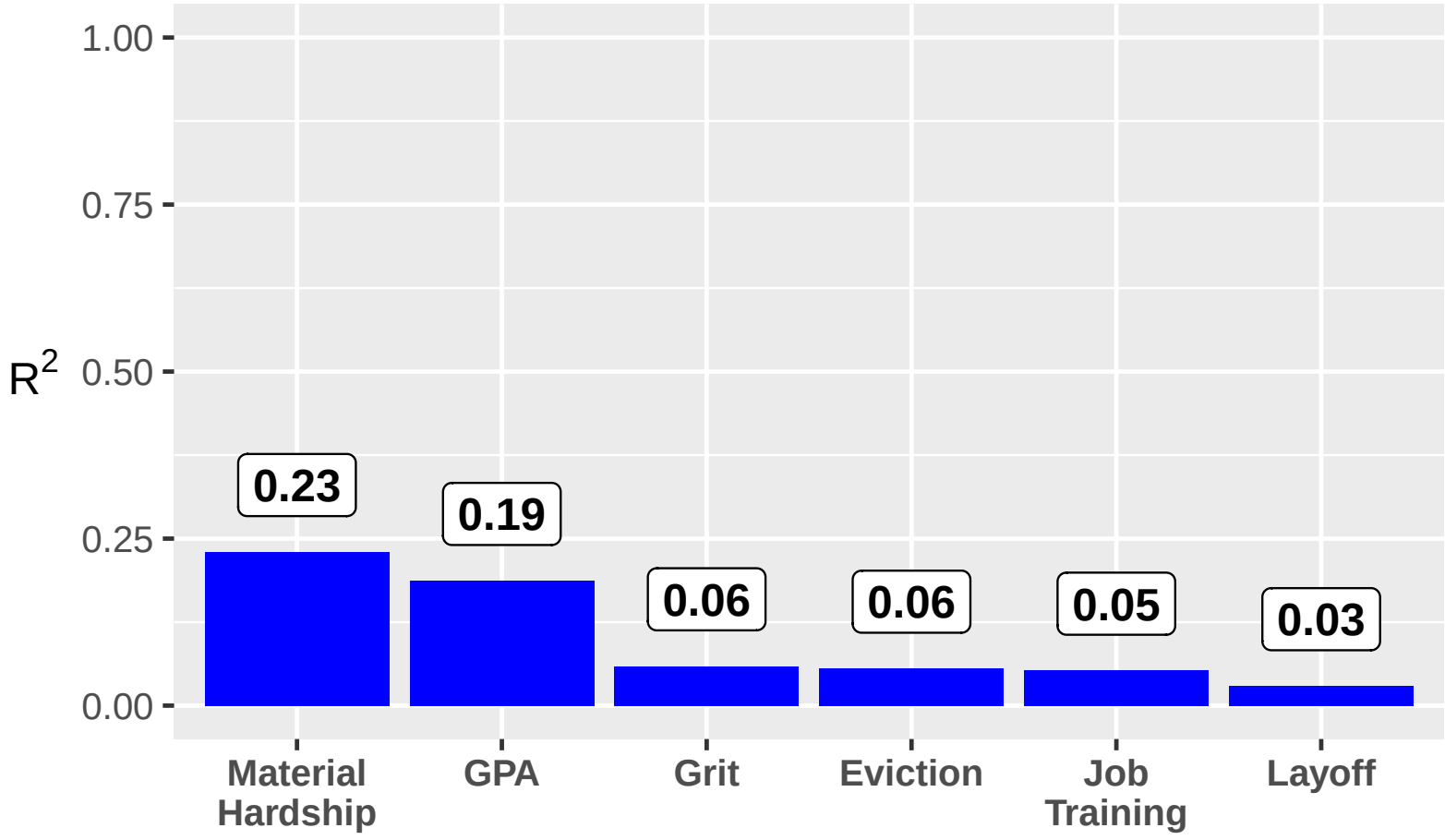
$$R^2_{holdout} = 1 - \frac{\sum_{i \in \text{holdout}} (\hat{y}_i - y_i)^2}{\sum_{i \in \text{holdout}} (\bar{y}_{train} - y_i)^2}$$

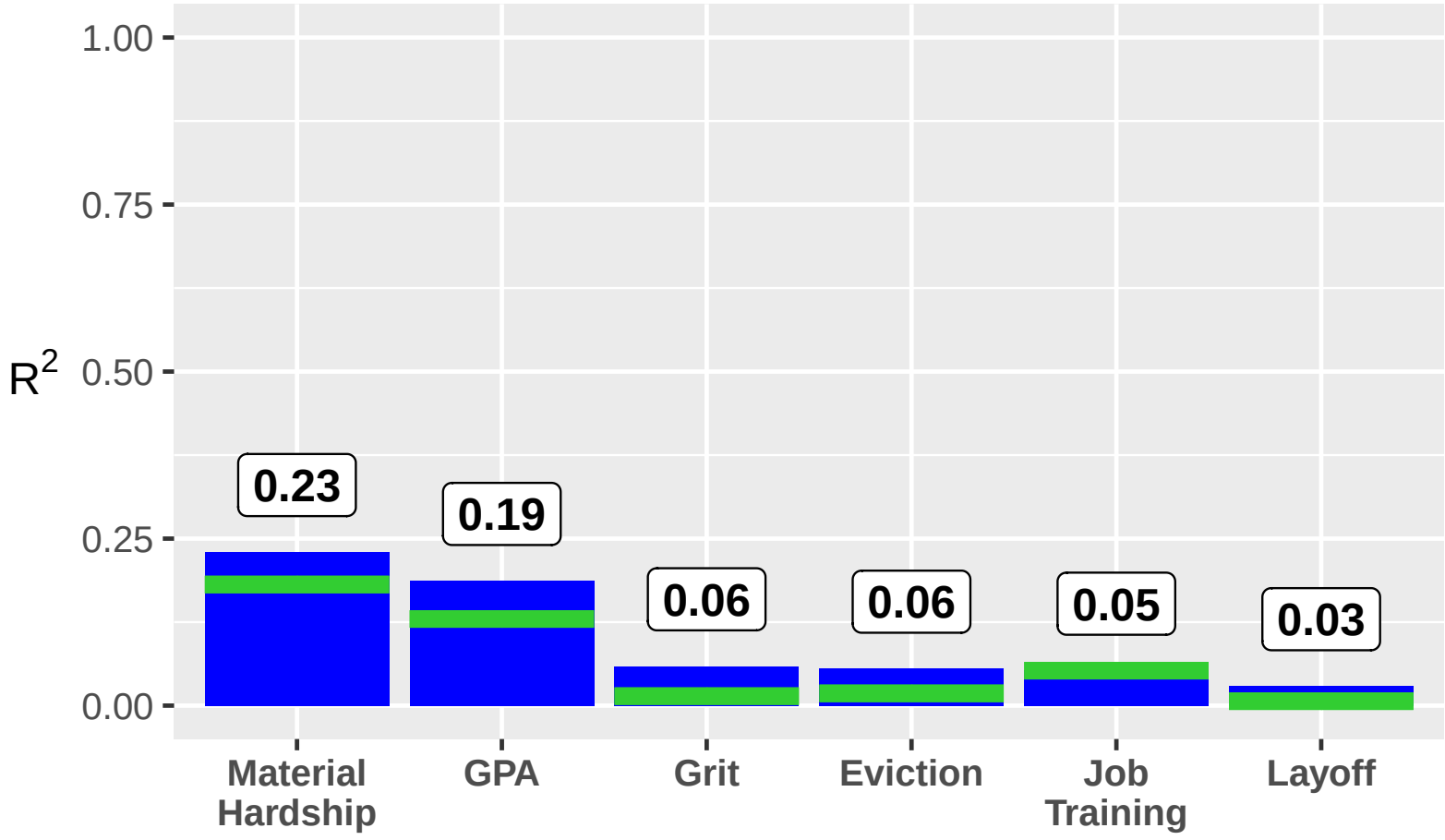
Before I show the results, let's vote . . . .


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

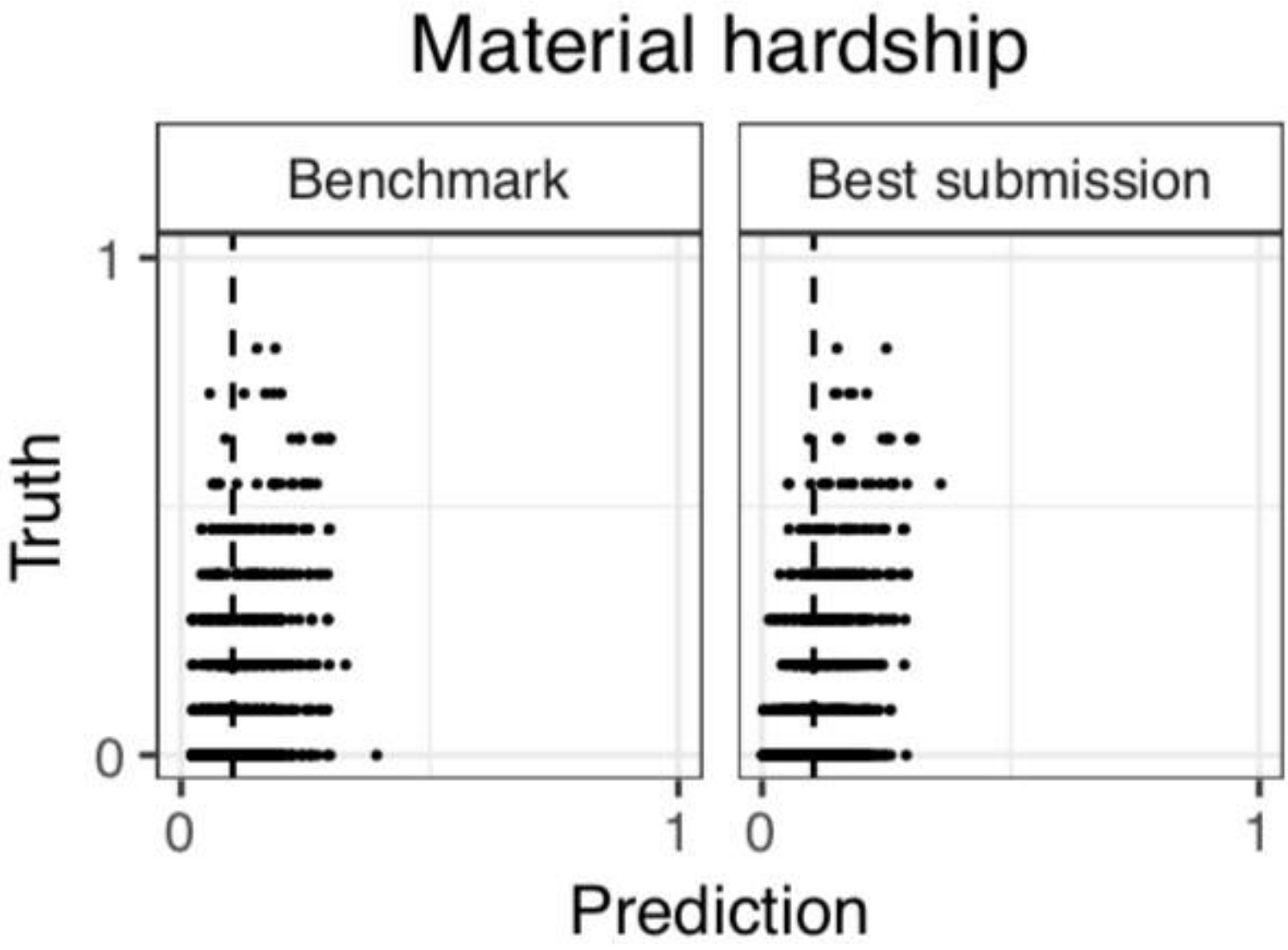

1.00
0.75
0.50
0.25
0.00
R^2
0.23
0.19
0.06
0.06
0.05
0.03
Material Hardship
GPA
Grit
Eviction
Job Training
Layoff

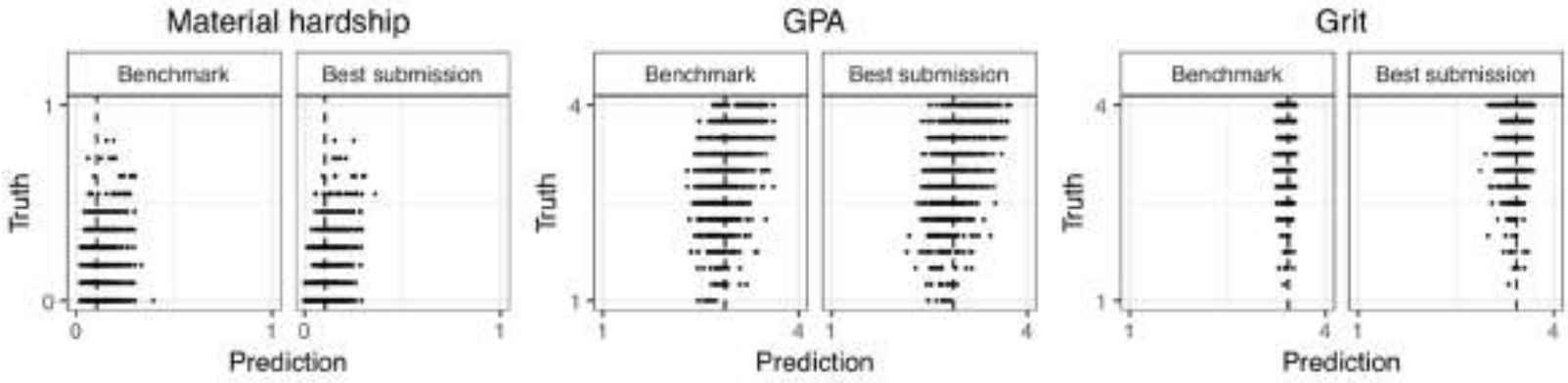
Is this even better than a benchmark model?

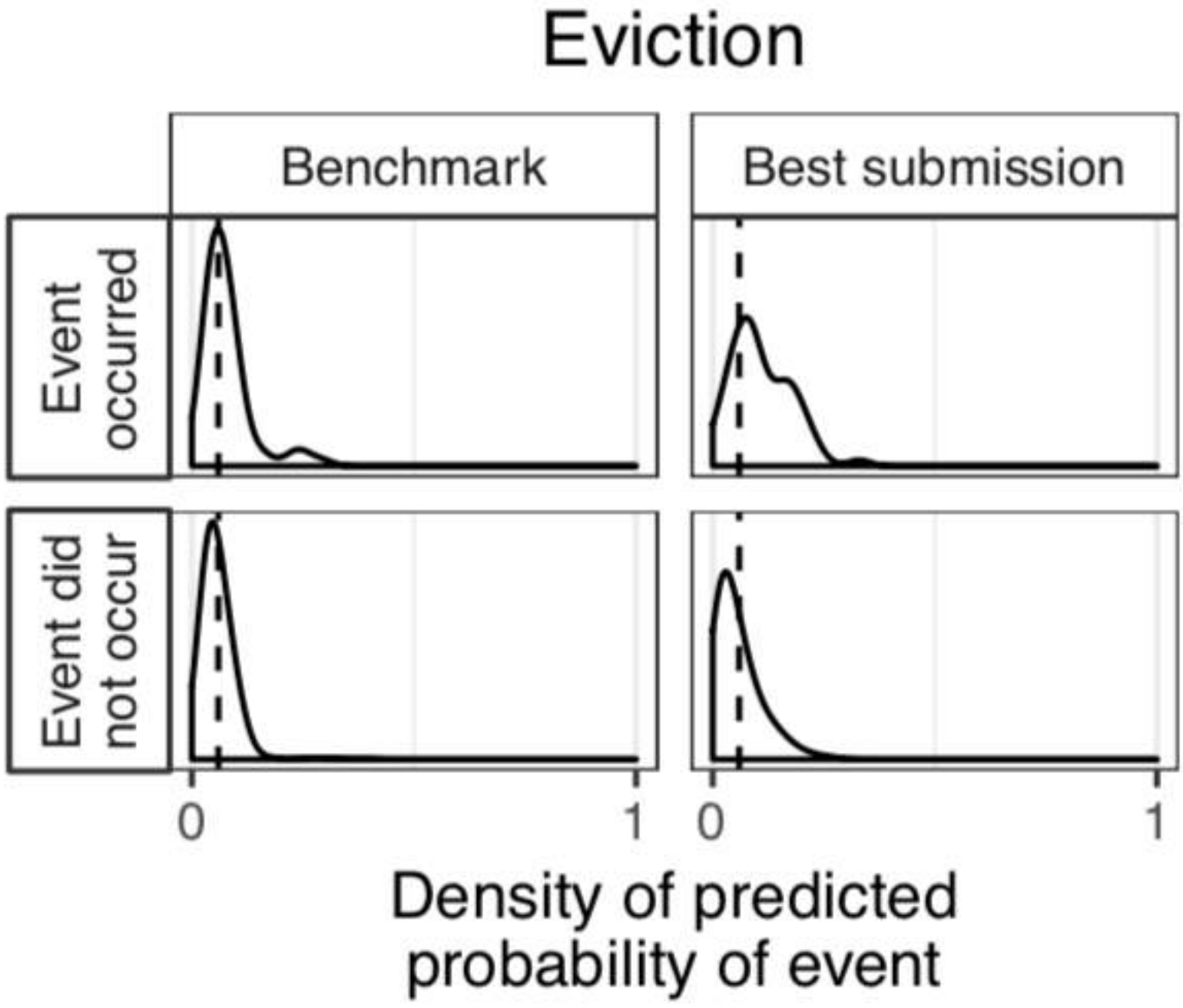

1.00
0.75
0.50
0.25
0.00
R^2
0.23
0.19
0.06
0.06
0.05
0.03
Material Hardship
GPA
Grit
Eviction
Job Training
Layoff

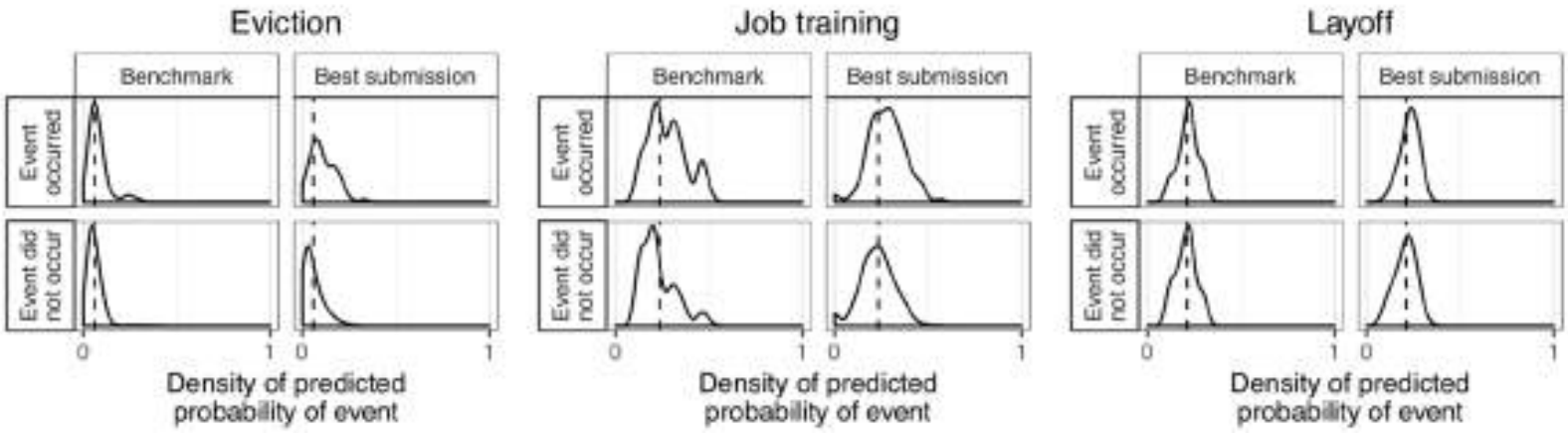


Green line: 4 variable linear regression model

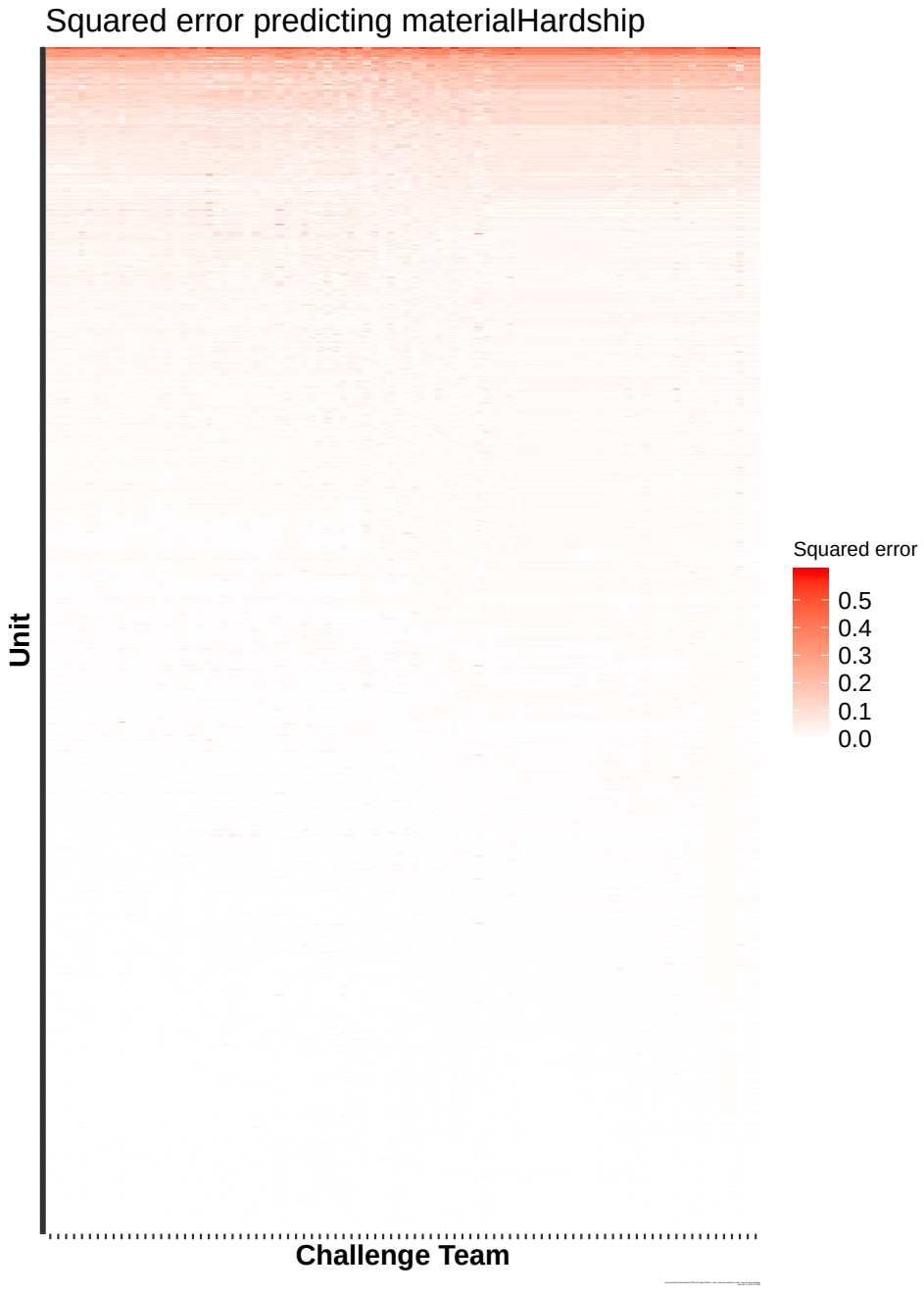

Material hardship
Benchmark
Best submission
Truth
1
0
0
1
0
1
Prediction

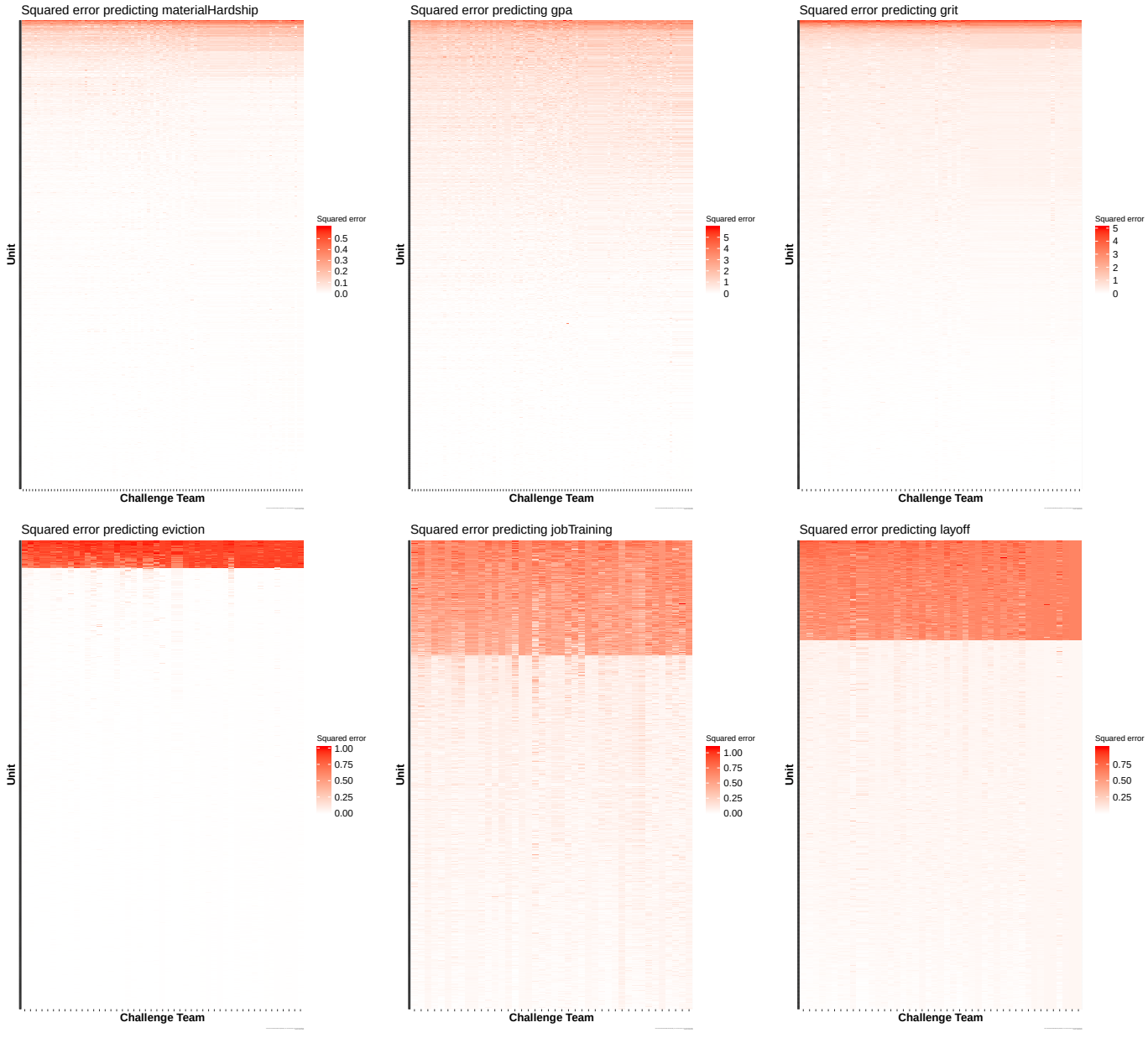

Material hardship
Benchmark
Best submission
Truth
Prediction
1
0
0
1
0
1
GPA
Benchmark
Best submission
Truth
Prediction
4
1
1
4
1
4
Grit
Benchmark
Best submission
Truth
Prediction
4
1
1
4
1
4


Eviction
Benchmark
Best submission
Event occurred
Event did not occur
0
1
0
1
Density of predicted probability of event


Eviction
Job training
Layoff
Benchmark
Best submission
Event occurred
Event did not occur
0
1
Density of predicted probability of event

What can we learn looking at the all the predictions?


Squared error predicting materialHardship
Unit
Challenge Team
Squared error
0.5
0.4
0.3
0.2
0.1
0.0


Squared error predicting materialHardship
Unit
Challenge Team
Squared error
0.5
0.4
0.3
0.2
0.1
0.0
Squared error predicting gpa
Unit
Challenge Team
Squared error
5
4
3
2
1
0
Squared error predicting grit
Unit
Challenge Team
Squared error
5
4
3
2
1
0
Squared error predicting eviction
Unit
Challenge Team
Squared error
1.00
0.75
0.50
0.25
0.00
Squared error predicting jobTraining
Unit
Challenge Team
Squared error
1.00
0.75
0.50
0.25
0.00
Squared error predicting layoff
Unit
Challenge Team
Squared error
0.75
0.50
0.25

Next questions:

- Is it possible to get better predictive performance for this data and prediction task?

Next questions:

- Is it possible to get better predictive performance for this data and prediction task?
- Why is the unpredictability so high even using modern machine learning methods and what many social scientists would consider to be large and high-quality data?

Why is the unpredictability so high even using modern machine learning methods and what many social scientists would consider to be large and high-quality data?

- Not enough cases

Why is the unpredictability so high even using modern machine learning methods and what many social scientists would consider to be large and high-quality data?

- Not enough cases
- Measurement error in existing variables (particularly outcomes)

Why is the unpredictability so high even using modern machine learning methods and what many social scientists would consider to be large and high-quality data?

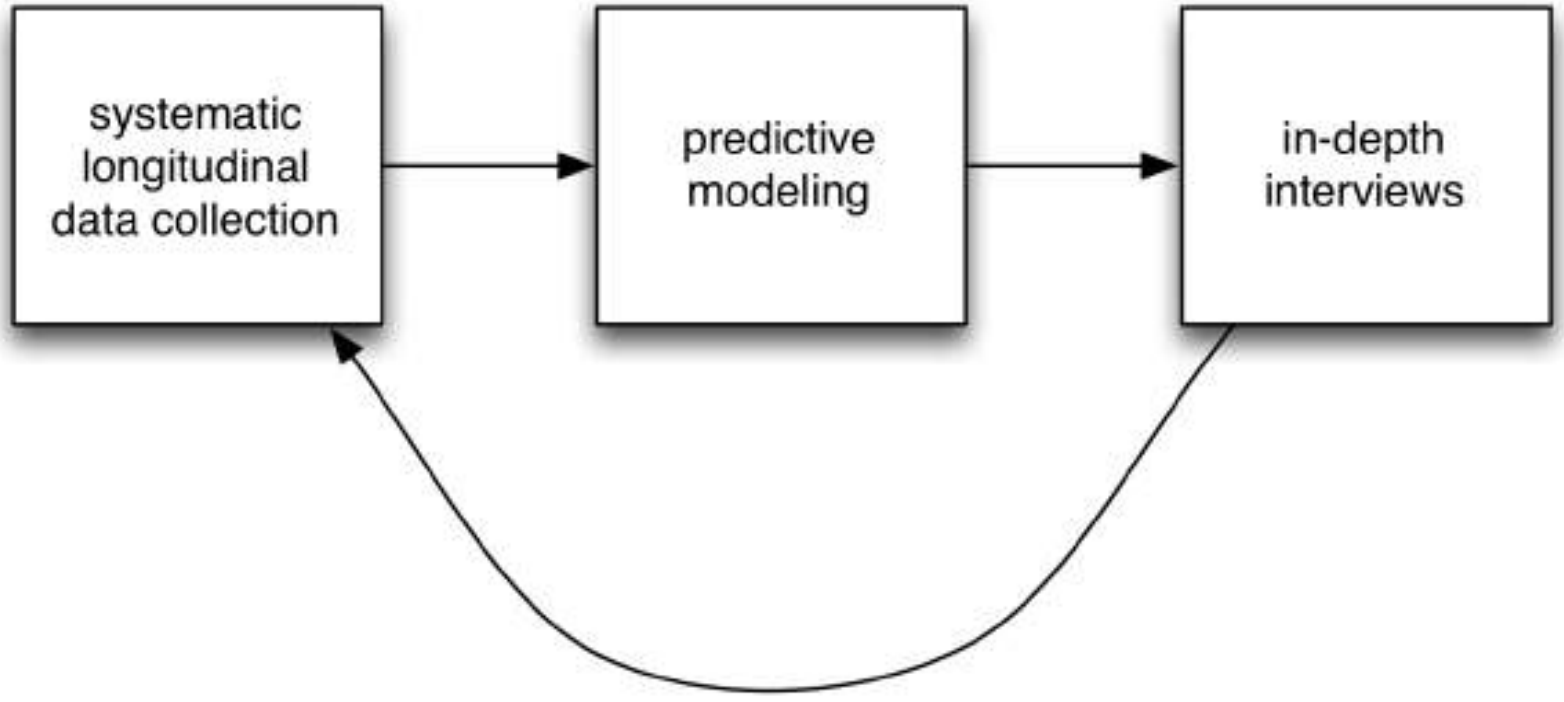
- Not enough cases
- Measurement error in existing variables (particularly outcomes)
- Important unmeasured variables

Why is the unpredictability so high even using modern machine learning methods and what many social scientists would consider to be large and high-quality data?

- Not enough cases
- Measurement error in existing variables (particularly outcomes)
- Important unmeasured variables

How can we learn about important measurement error and unmeasured variables?

In-depth 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, `https://arxiv.org/abs/1809.00103`

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, `https://arxiv.org/abs/1809.00103`
    - "Improving metadata infrastructure for complex surveys: Insights from the Fragile Families Challenge" by Kindel, Catena, Hartshorne, Jaeger, Koffman, McLanahan, Phillips, Rouhani, Vinh, & Salganik, `https://osf.io/93ywg/`

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, `https://arxiv.org/abs/1809.00103`
    - "Improving metadata infrastructure for complex surveys: Insights from the Fragile Families Challenge" by Kindel, Catena, Hartshorne, Jaeger, Koffman, McLanahan, Phillips, Rouhani, Vinh, & Salganik, `https://osf.io/93ywg/`
    - "Successes and struggles with computational reproducibility in the Fragile Families Challenge" by Liu & Salganik, `https://osf.io/preprints/socarxiv/g3pdb/`

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

Mullainathan and Spiess (2017):
http://dx.doi.org/10.1257/jep.31.2.87

- Read: `http://www.bitbybitbook.com`
- Teach: `http://www.bitbybitbook.com/en/teaching/`
  (and Italian version coming soon from il Mulino)

How typical is this result?

- 6 year gap between end of background data and outcome

How typical is this result?

- 6 year gap between end of background data and outcome
- large social disruption—the Great Recession—between end of background data and outcome

How typical is this result?

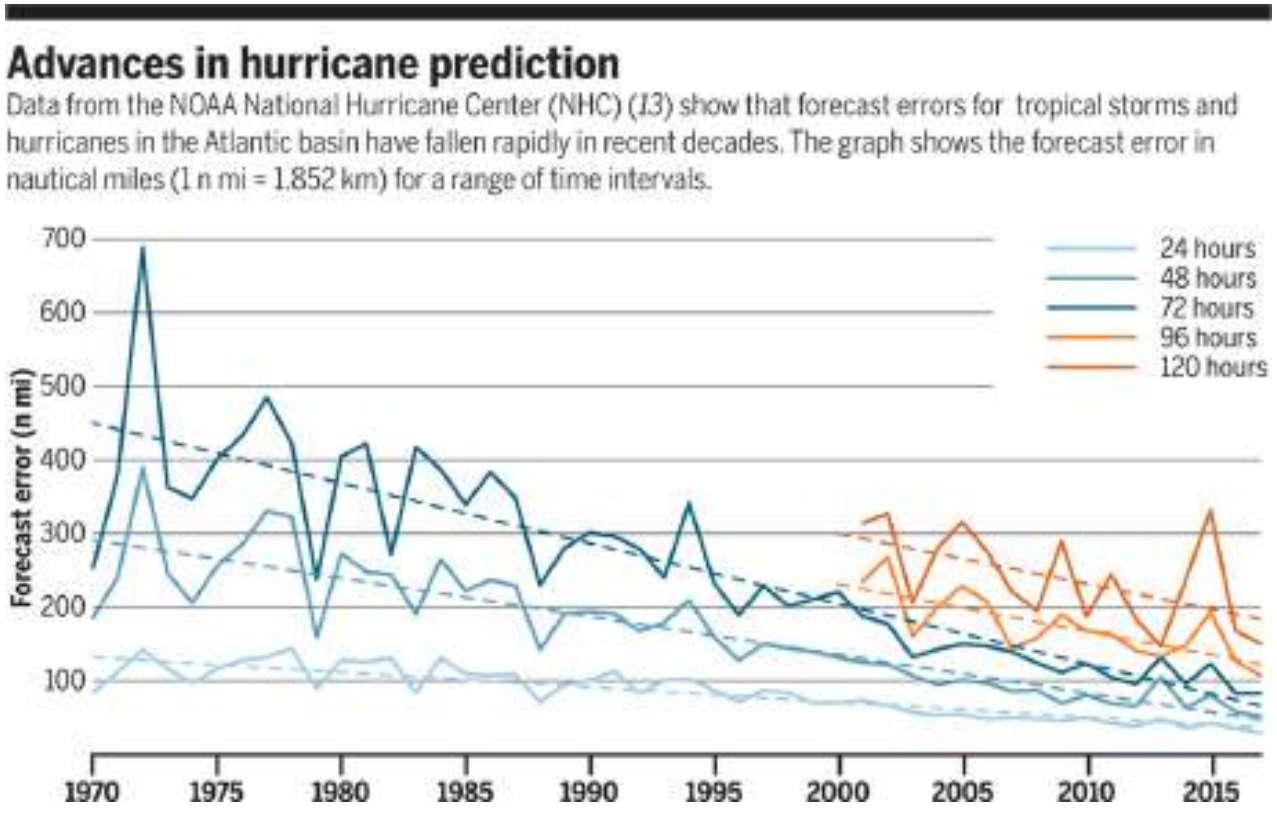
- 6 year gap between end of background data and outcome
- large social disruption—the Great Recession—between end of background data and outcome
- the sample design of the Fragile Families study

How typical is this result?

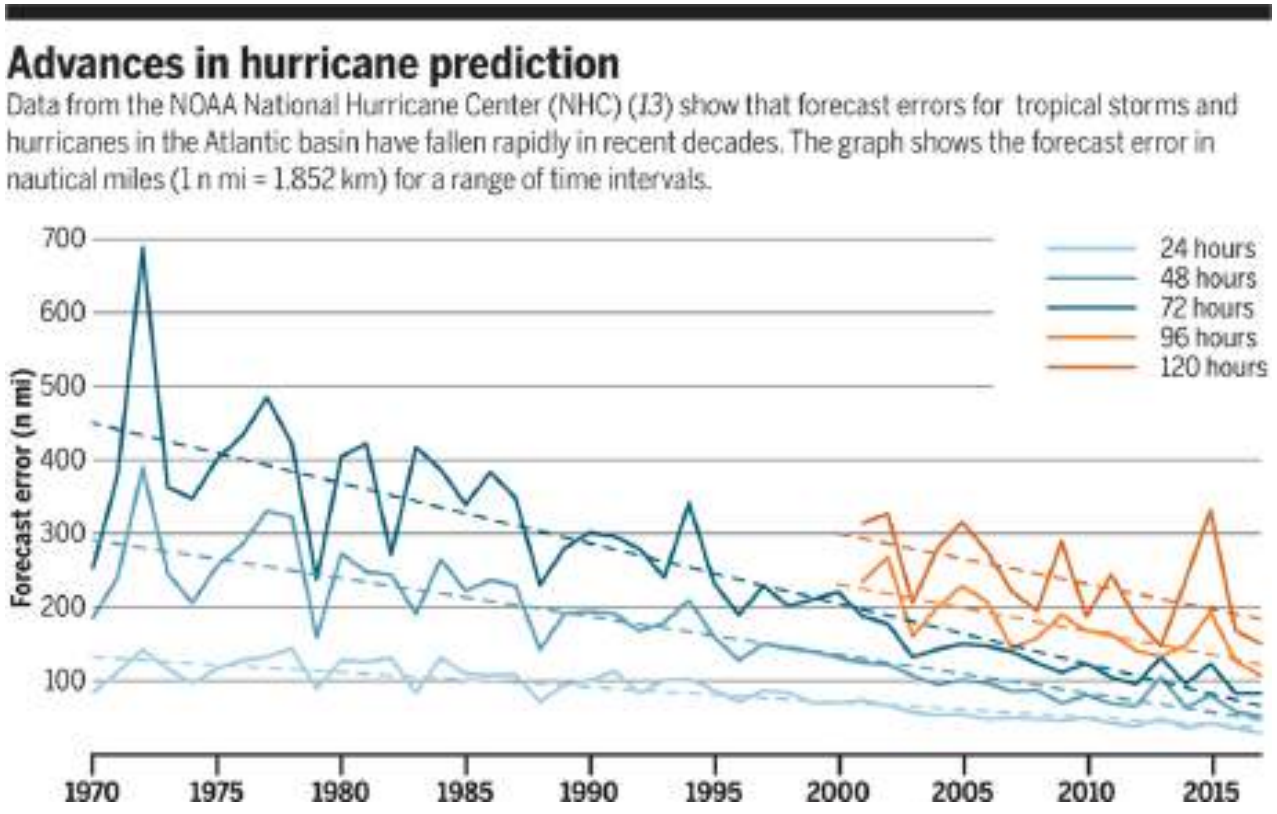
- 6 year gap between end of background data and outcome
- large social disruption—the Great Recession—between end of background data and outcome
- the sample design of the Fragile Families study
- outcomes measured when child was 15
- outcomes are at a relatively narrow point in time rather than average over a longer time period (e.g., grades last semester vs grades in high school)

What is the ideal C. elegans for social science prediction problems?

What is the ideal C. elegans for social science prediction problems?

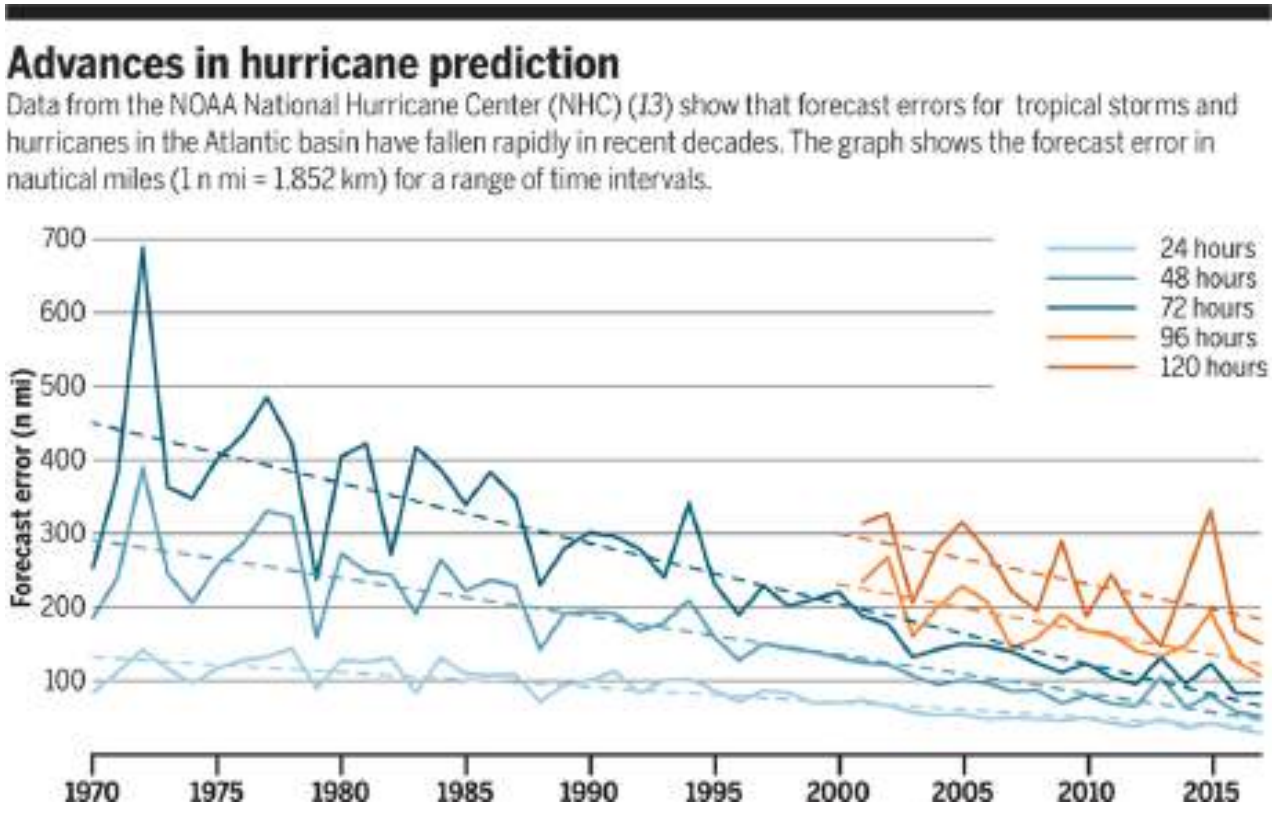


What is the ideal C. elegans for social science prediction problems?



- Can we do this?

What is the ideal C. elegans for social science prediction problems?



- Can we do this?
- Should we do this?

A proposal to focus on longitudinal social surveys:

A proposal to focus on longitudinal social surveys:

- Dozens already happening all over the world with interesting similarities and differences

A proposal to focus on longitudinal social surveys:

- Dozens already happening all over the world with interesting similarities and differences
- Already strong community around each survey

A proposal to focus on longitudinal social surveys:

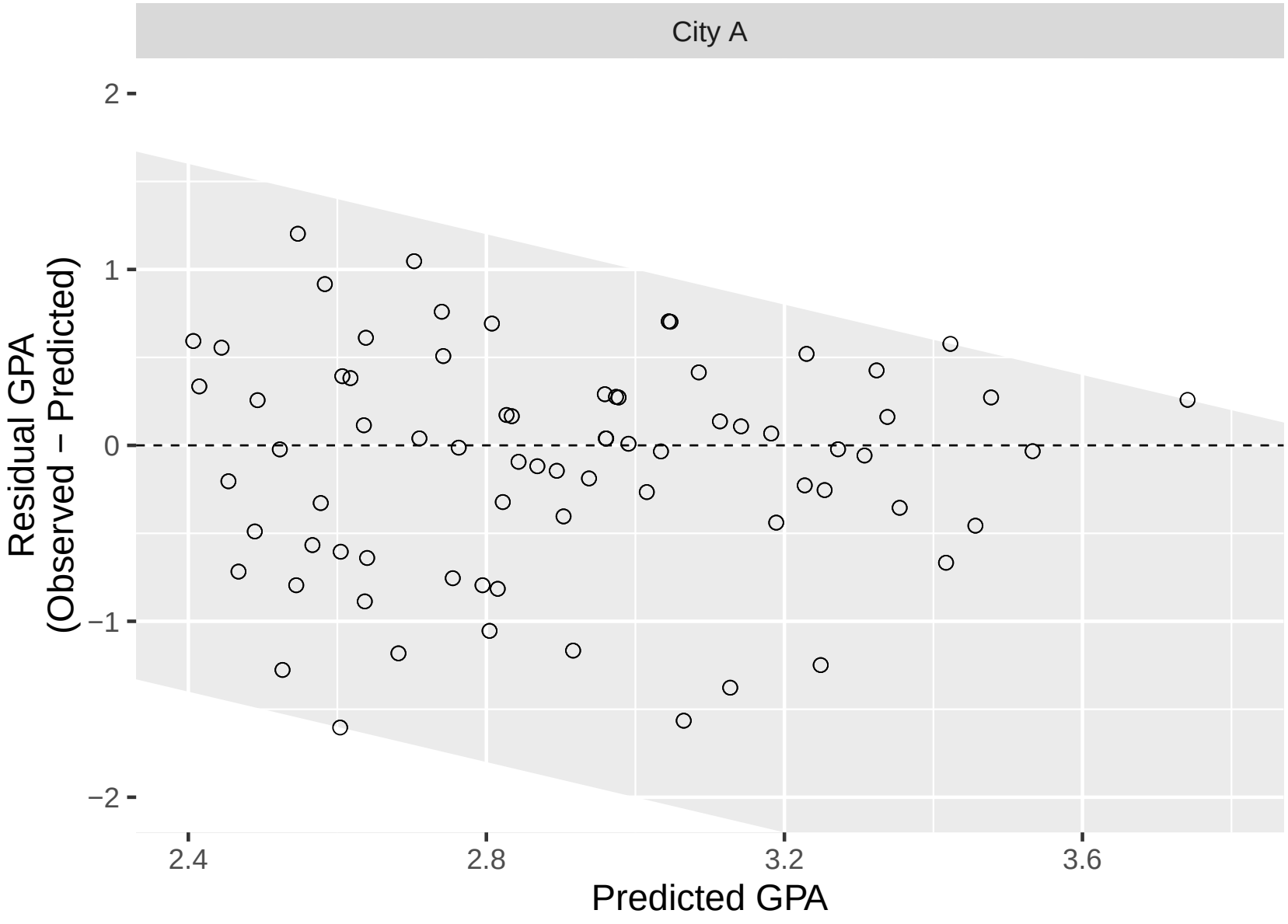
- Dozens already happening all over the world with interesting similarities and differences
- Already strong community around each survey
- Code from a single Challenge can be repurposed to create many simulated Challenges

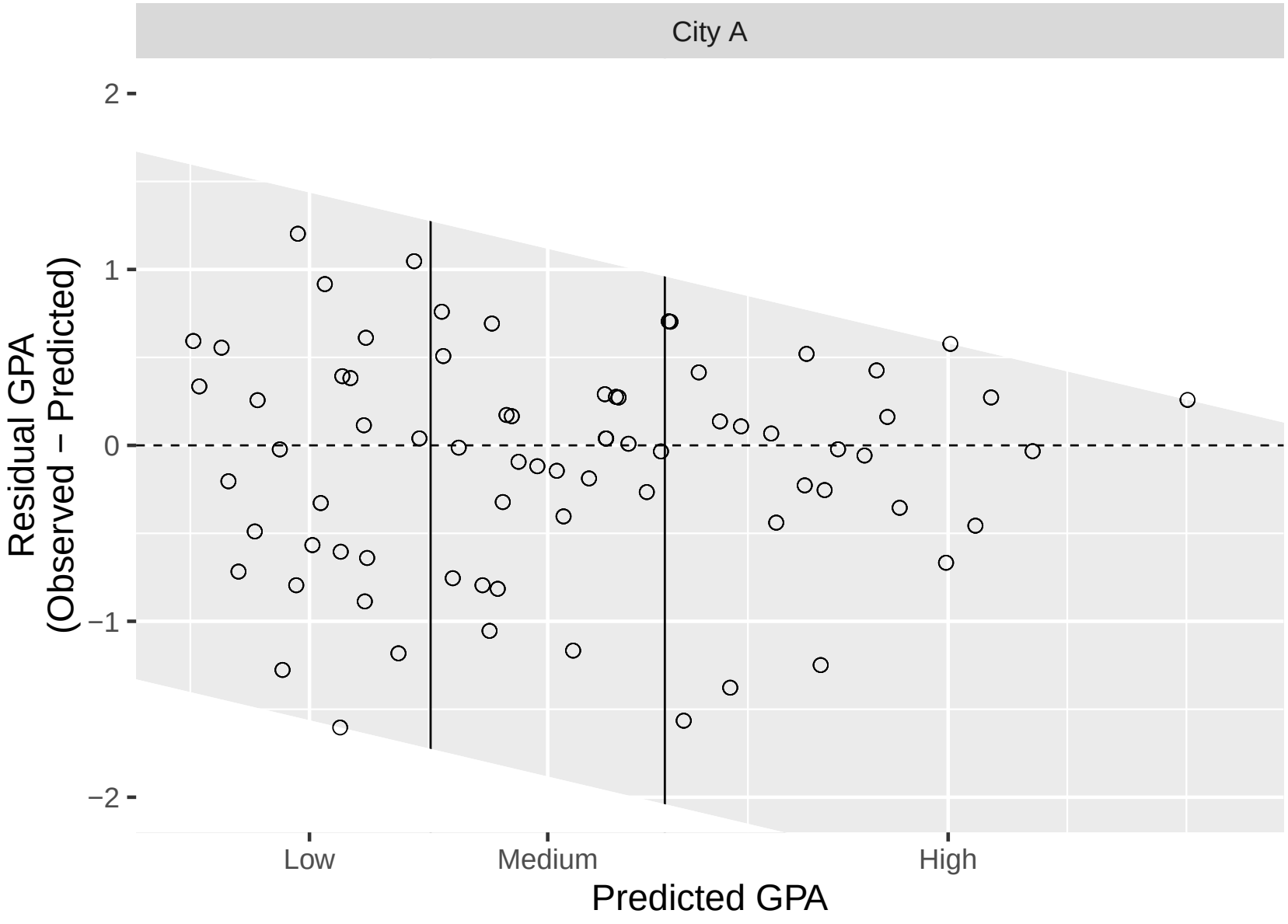
A proposal to focus on longitudinal social surveys:

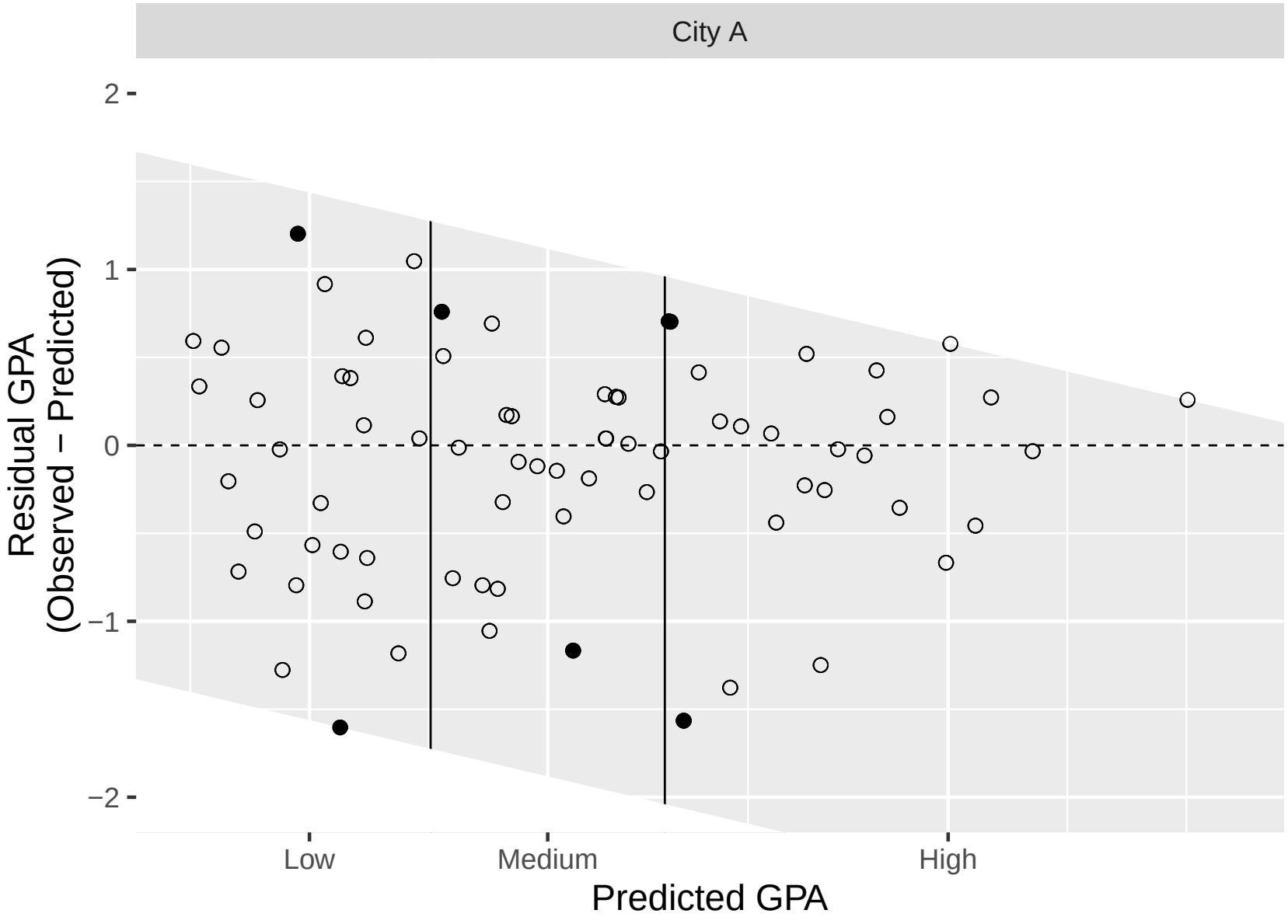
- Dozens already happening all over the world with interesting similarities and differences
- Already strong community around each survey
- Code from a single Challenge can be repurposed to create many simulated Challenges
- Data collected with informed consent under well-developed ethical frameworks

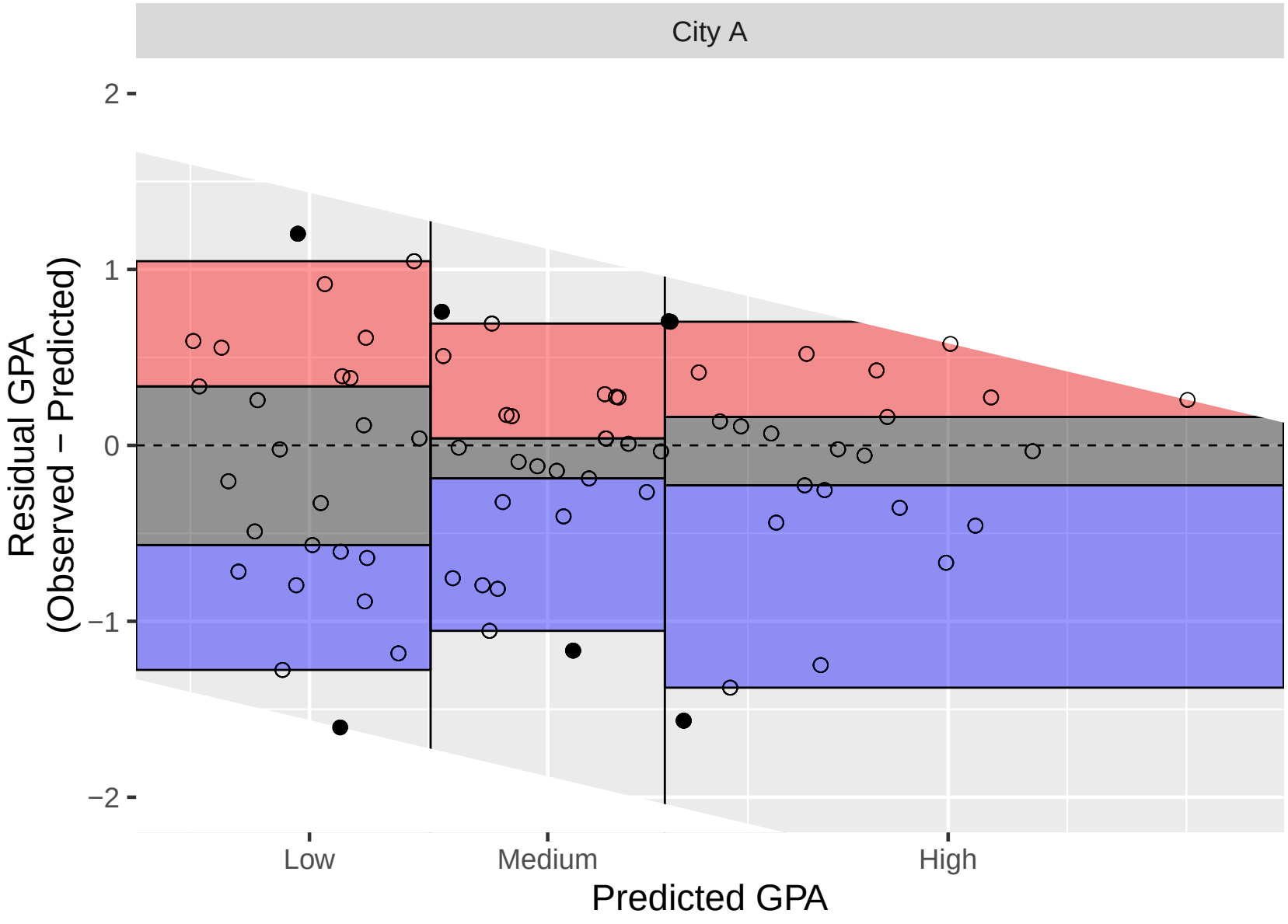
A proposal to focus on longitudinal social surveys:

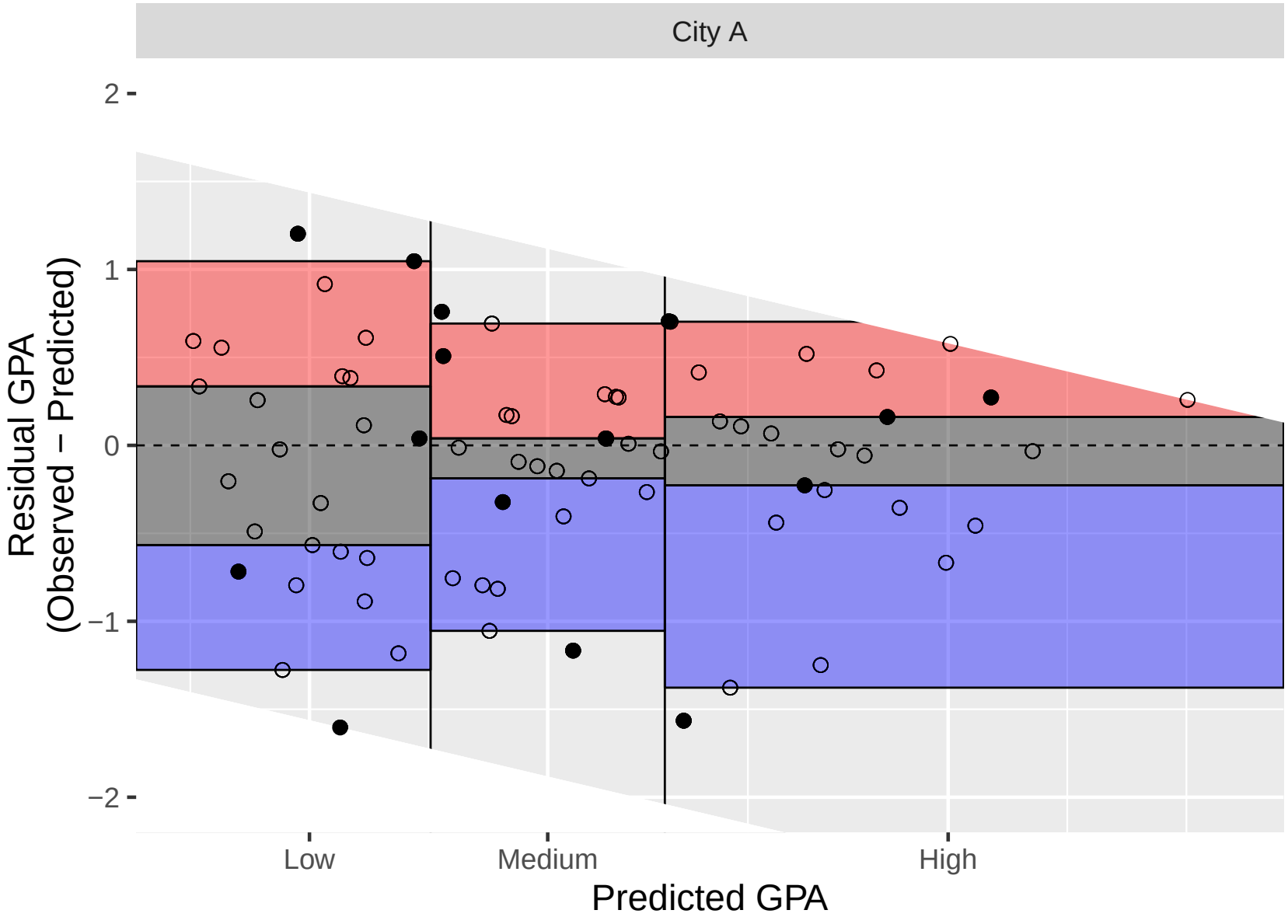
- Dozens already happening all over the world with interesting similarities and differences
- Already strong community around each survey
- Code from a single Challenge can be repurposed to create many simulated Challenges
- Data collected with informed consent under well-developed ethical frameworks
- Likely to spur useful scientific developments

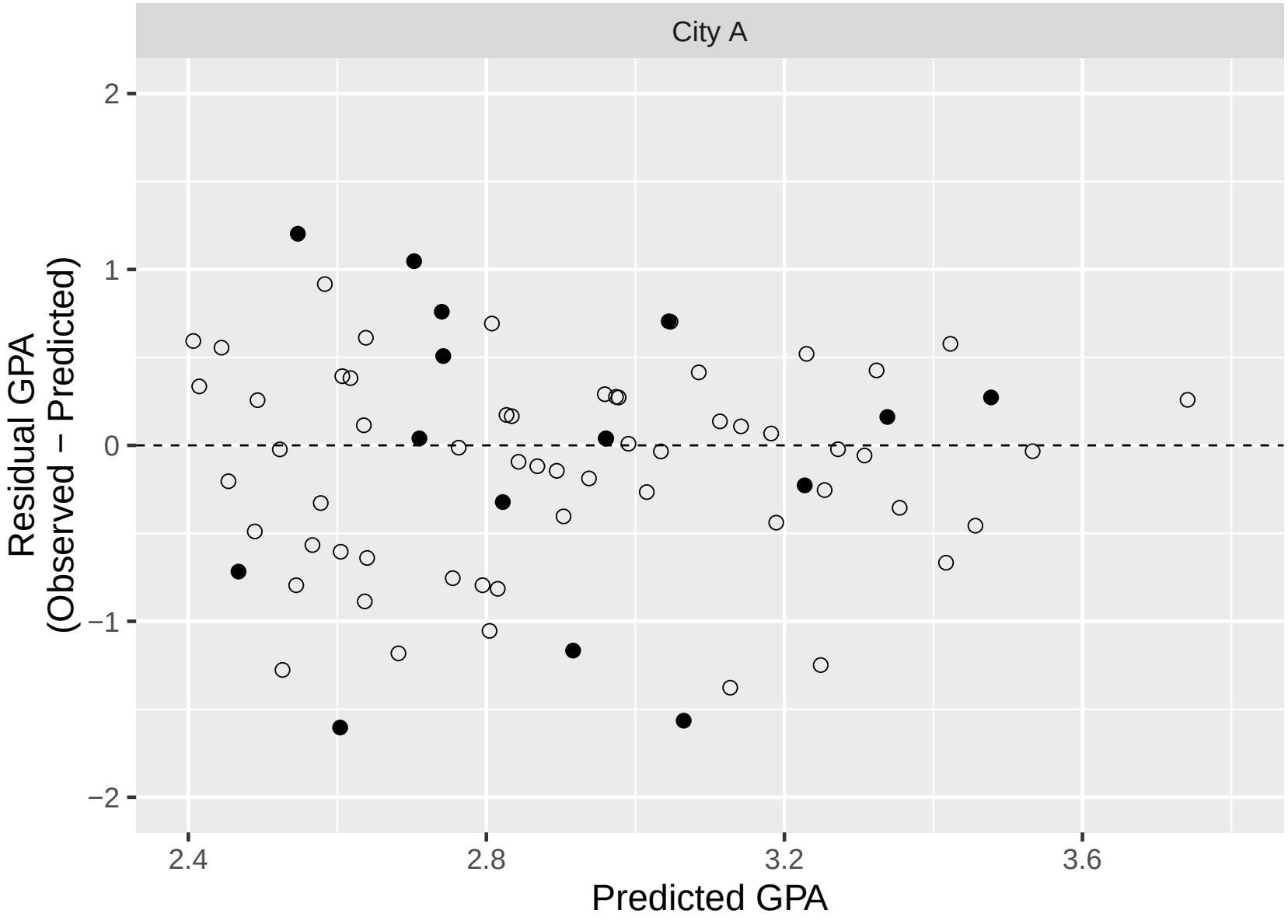

City A
Residual GPA
(Observed – Predicted)
2
1
0
−1
−2
2.4
2.8
3.2
3.6
Predicted GPA

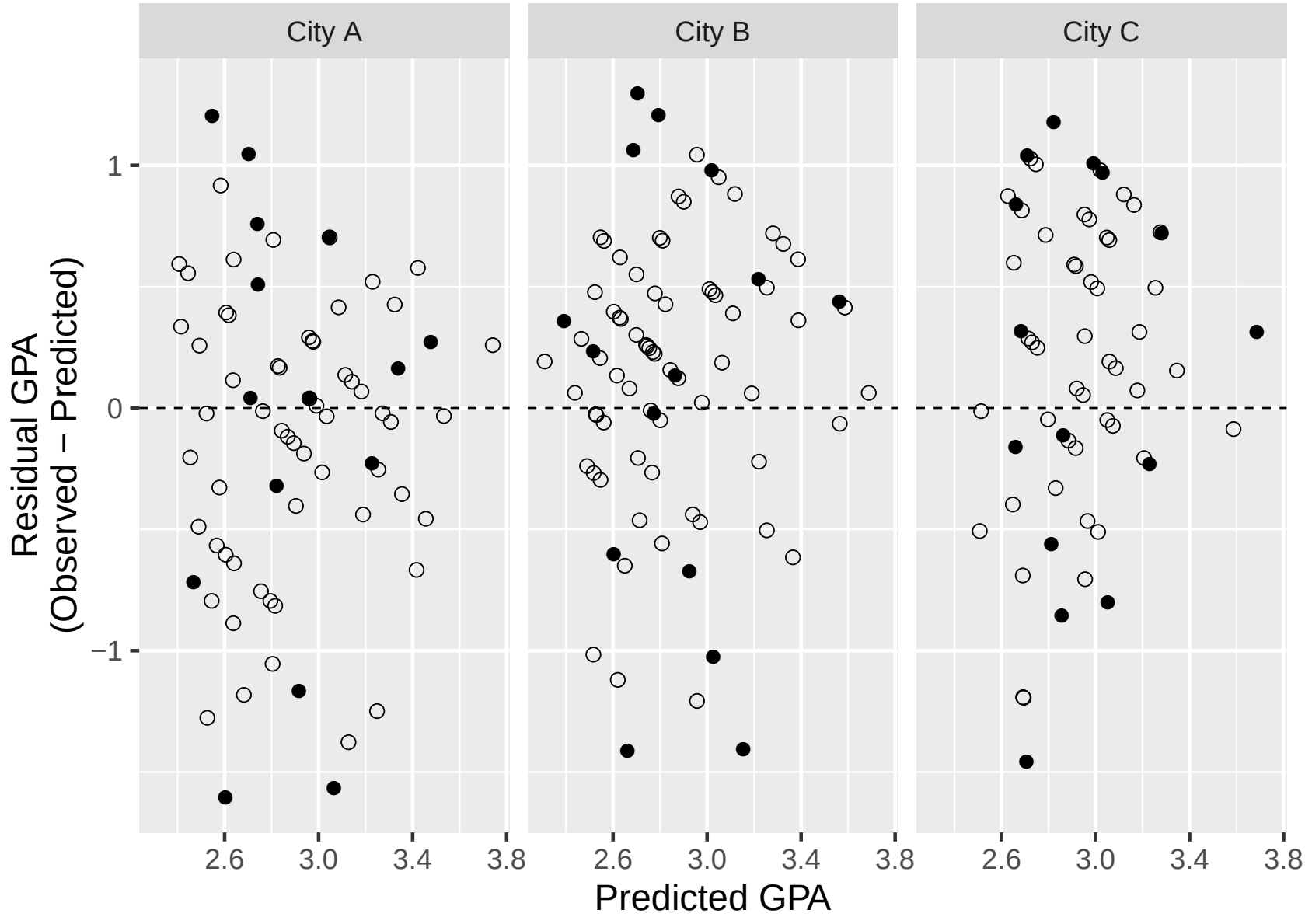

City A
2
1
0
−1
−2
Residual GPA
(Observed − Predicted)
Low
Medium
High
Predicted GPA


City A
2
1
0
−1
−2
Residual GPA
(Observed − Predicted)
Low
Medium
High
Predicted GPA


City A
2
1
0
−1
−2
Residual GPA
(Observed − Predicted)
Low
Medium
High
Predicted GPA


City A
2
1
0
−1
−2
Residual GPA
(Observed − Predicted)
Low
Medium
High
Predicted GPA


City A
Residual GPA
(Observed – Predicted)
2
1
0
−1
−2
2.4
2.8
3.2
3.6
Predicted GPA


City A
City B
City C
Residual GPA
(Observed – Predicted)
1
0
−1
2.6
3.0
3.4
3.8
Predicted GPA