

Participating in the Fragile Families Challenge Activity

Matthew Salganik, Ian Lundberg, Alex Kindel, Sara McLanahan,
and people from around the world

Summer Institutes in Computational Social Science
June 21, 2019

Fragile Families Challenge is supported by the Russell Sage Foundation. 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.



4,242 families

12,942 features
birth to age 9

6 outcomes
age 15

Training

Leaderboard

Holdout

Introducing the outcome variables

GPA¹

¹Learn more at <http://www.fragilefamilieschallenge.org/gpa/>

GPA¹

How do kids beat the odds academically?

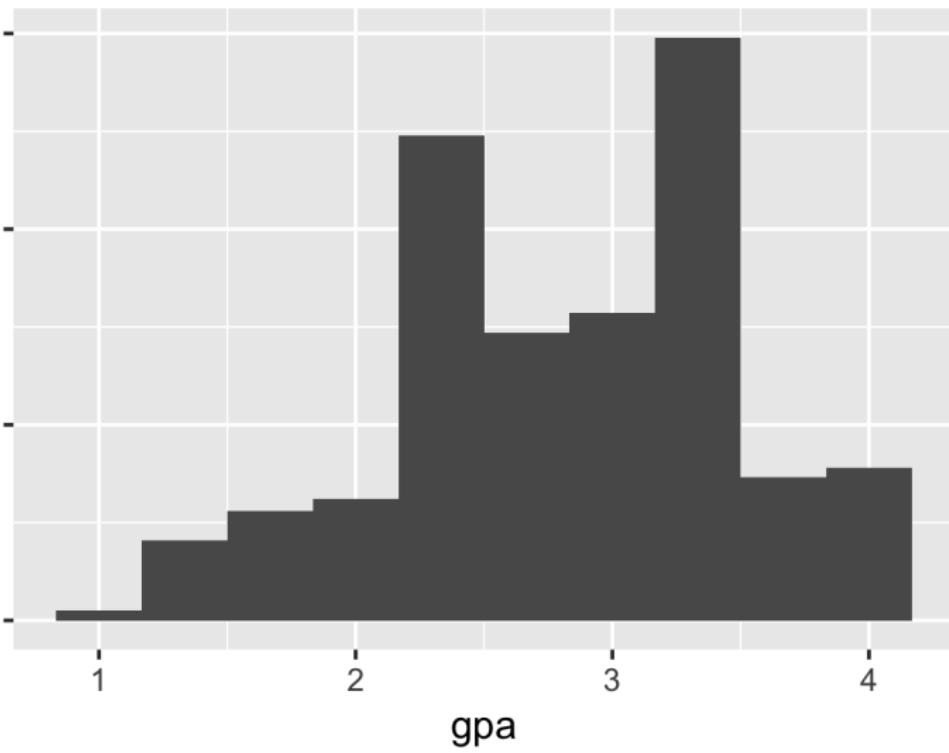
¹Learn more at <http://www.fragilefamilieschallenge.org/gpa/>

GPA²

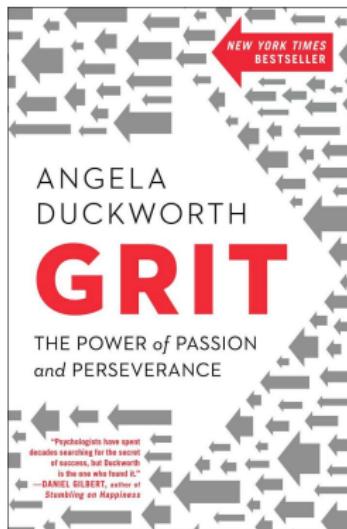
B20. At the {most recent grading period/last grading period in the spring} what was your grade in ...

	A	B	C	D OR LOWE R	NO GRADE OR PASS/FAIL	REF	DK	N/A HOMESCHOoled
B20A English or language arts? ..	1	2	3	4	5	-1	-2	7 → GO TO B22A
B20B Math?	1	2	3	4	5	-1	-2	7 → GO TO B22A
B20C History or social studies? ..	1	2	3	4	5	-1	-2	7 → GO TO B22A
B20D Science?	1	2	3	4	5	-1	-2	7 → GO TO B22A

²This variable is reverse-coded in the data file so that higher values represent higher GPAs.

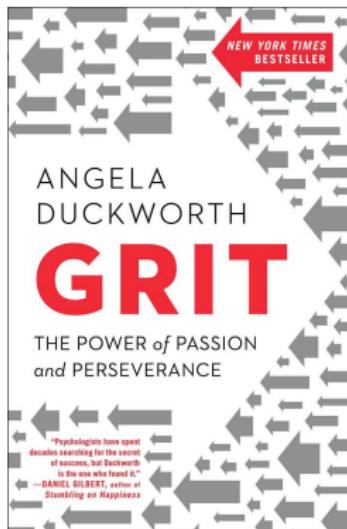


“Grit” predicts success, possibly more than IQ.³



³Learn more at <http://www.fragilefamilieschallenge.org/grit/>

“Grit” predicts success, possibly more than IQ.³



What makes some kids gritty?

³Learn more at <http://www.fragilefamilieschallenge.org/grit/>

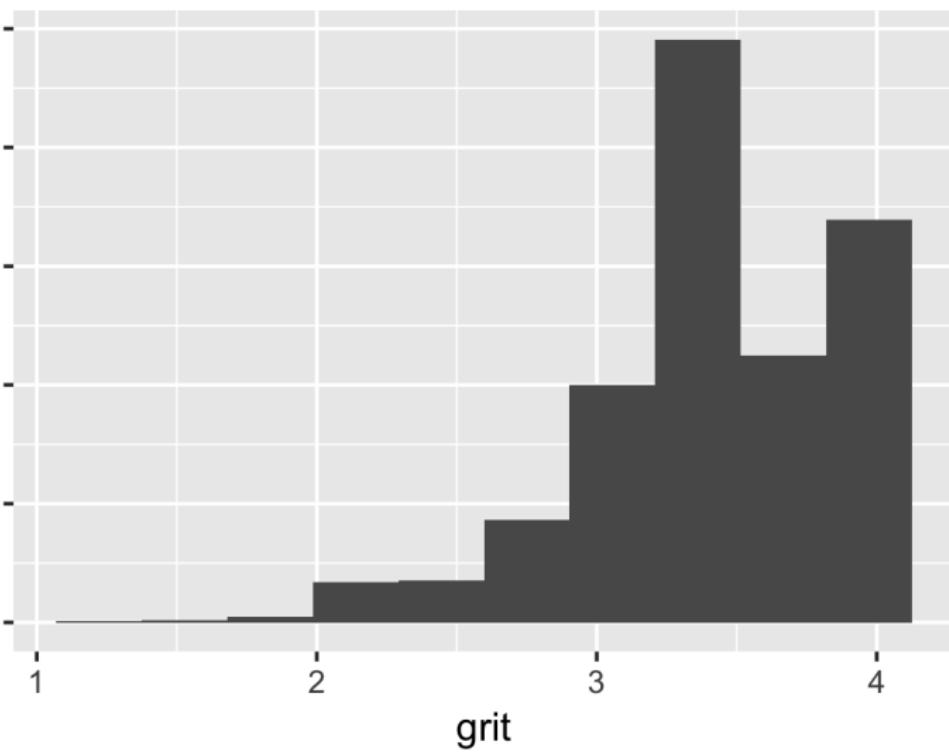
Grit⁴

- D2. Thinking about how you have behaved or felt during the past four weeks, please tell me whether you strongly agree, somewhat agree, somewhat disagree, or strongly disagree with the following statements.

PROBE: Thinking about the past four weeks, do you strongly agree, somewhat agree, somewhat disagree, or strongly disagree with this statement?

	STRONGLY AGREE	SOMEWHAT AGREE	SOMEWHAT DISAGREE	STRONGLY DISAGREE	REF	DK
D2I. I keep at my schoolwork until I am done with it.....	1	2	3	4	-1	-2
D2K. Once I make a plan to get something done, I stick to it.....	1	2	3	4	-1	-2
D2M. I finish whatever I begin.....	1	2	3	4	-1	-2
D2V. I am a hard worker	1	2	3	4	-1	-2

⁴This variable is reverse-coded in the data file so that higher values represent more grit.



Material hardship⁵

⁵Learn more at

<http://www.fragilefamilieschallenge.org/material-hardship/>

Material hardship⁵

What unmeasured predictors are associated with families unexpectedly escaping severe deprivation?

⁵Learn more at

<http://www.fragilefamilieschallenge.org/material-hardship/>

Material hardship⁵

What unmeasured predictors are associated with families unexpectedly escaping severe deprivation?

What sends families unexpectedly into deep poverty?

⁵Learn more at

<http://www.fragilefamilieschallenge.org/material-hardship/>

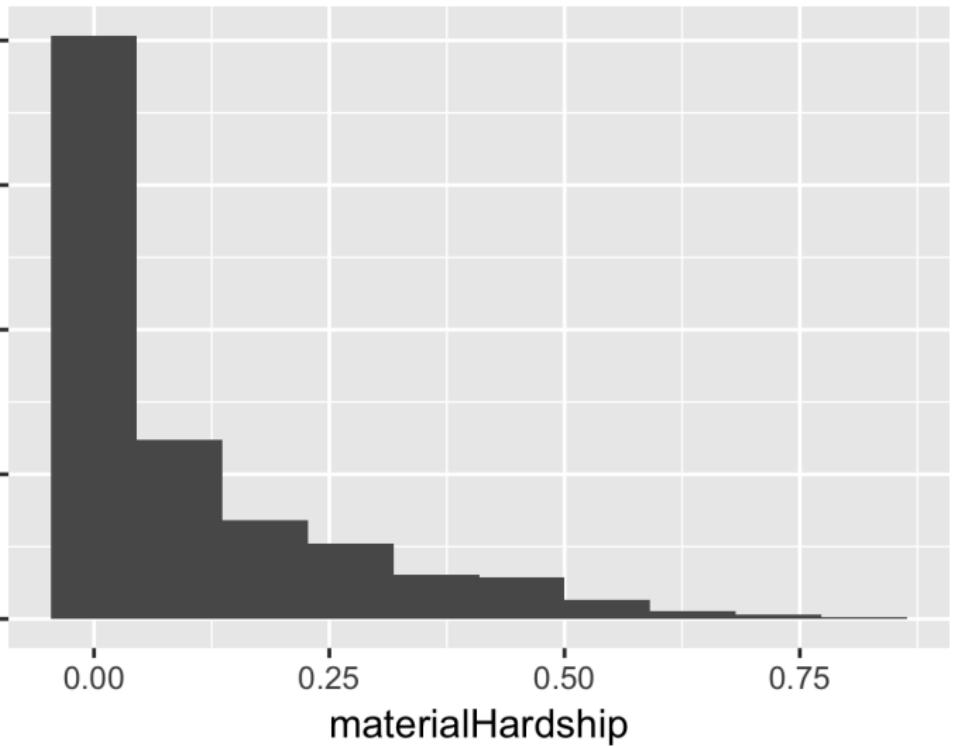
Material hardship

We are also interested in some of the problems that families face making ends meet. In the past twelve months, did you do any of the following because there wasn't enough money?

		YES	NO	REF	DK
J37.	In the past twelve months, did you receive free food or meals?	1	2	-1	-2
J38.	In the past twelve months, were you ever hungry, but didn't eat because you couldn't afford enough food?	1	2	-1	-2
J39.	In the past twelve months, did you ever not pay the full amount of rent or mortgage payments?	1	2	-1	-2
J40.	In the past twelve months, were you evicted from your home or apartment for not paying the rent or mortgage?	1	2	-1	-2
J41.	In the past twelve months, did you not pay the full amount of gas, oil, or electricity bill?	1	2	-1	-2
J42.	In the past twelve months, was your gas or electric services ever turned off, or the heating oil company did not deliver oil, because there wasn't enough money to pay the bills?	1	2	-1	-2
J43.	In the past twelve months, did you borrow money from friends or family to help pay bills?	1	2	-1	-2
J44.	In the past twelve months, did you move in with other people even for a little while because of financial problems?	1	2	-1	-2

Material hardship

J45.	In the past twelve months, did you stay at a shelter, in an abandoned building, an automobile or any other place not meant for regular housing, even for one night?	1	2	-1	-2
J46.	In the past twelve months, was there anyone in your household who needed to see a doctor or go to the hospital but couldn't go because of the cost?	1	2	-1	-2
J47.	In the past twelve months, was your telephone service (mobile or land line) cancelled or disconnected by the telephone company because there wasn't enough money to pay the bill?	1	2	-1	-2



Eviction⁶

⁶Learn more at <http://www.fragilefamilieschallenge.org/eviction/>

⁷Note: You will just create propensity scores for eviction given background variables; causal inference comes in the second stage of the Challenge when outcomes are measured several years from now.

Eviction⁶

Does housing eviction **cause** worse outcomes as kids transition to adulthood?⁷

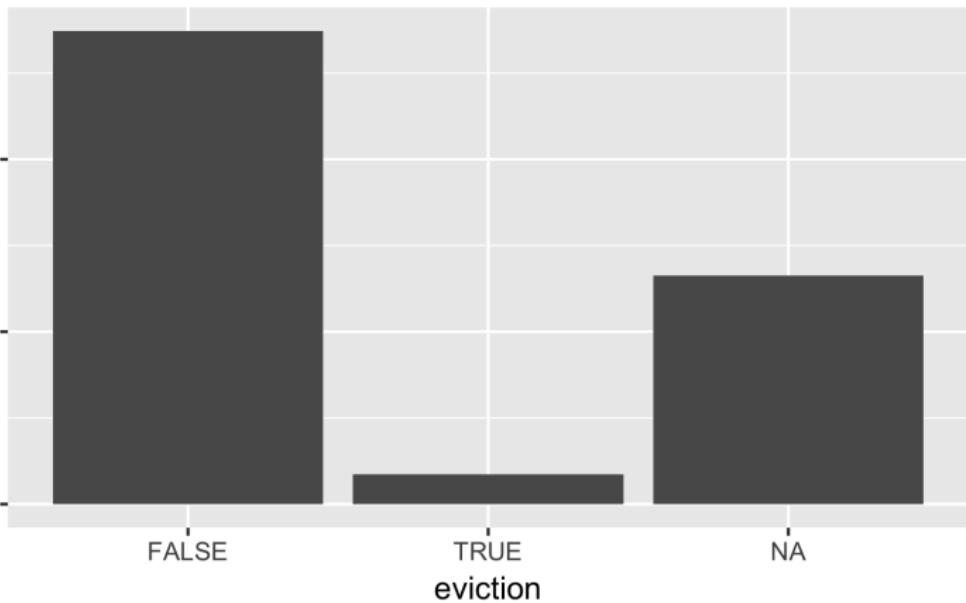
⁶Learn more at <http://www.fragilefamilieschallenge.org/eviction/>

⁷Note: You will just create propensity scores for eviction given background variables; causal inference comes in the second stage of the Challenge when outcomes are measured several years from now.

Eviction

J51. Since {MONTH AND YEAR COHORT CITY FIELDDED IN YR 9}, were you evicted from your home or apartment for not paying the rent or mortgage?

YES	1
NO	2
REFUSED	-1
DON'T KNOW	-2



Caregiver layoff⁸

⁸Learn more at <http://www.fragilefamilieschallenge.org/layoff/>

⁹Note: You will just create propensity scores for caregiver layoff given background variables; causal inference comes in the second stage of the Challenge when outcomes are measured several years from now.

Caregiver layoff⁸

Does layoff of a caregiver **cause** collateral damage for kids?⁹

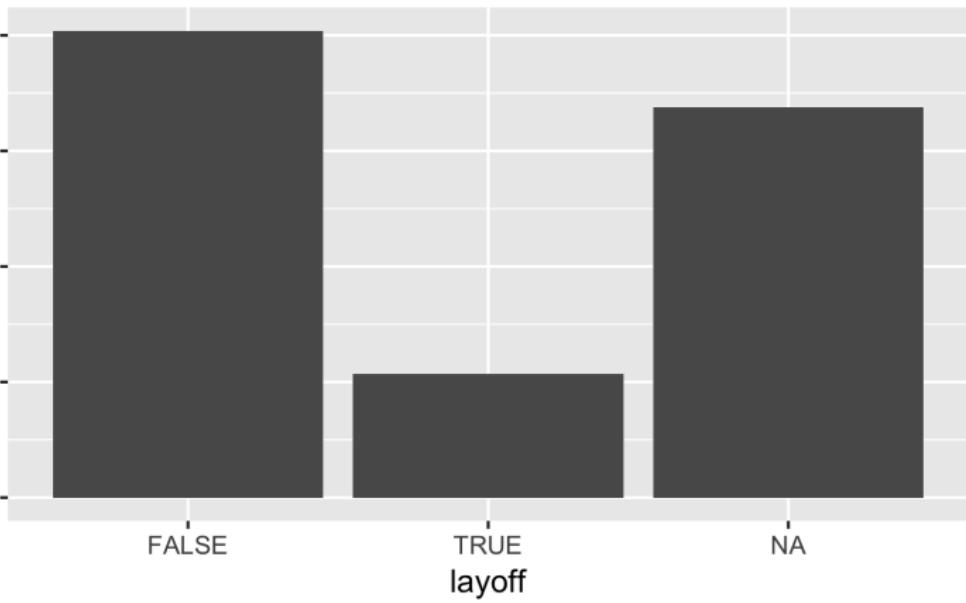
⁸Learn more at <http://www.fragilefamilieschallenge.org/layoff/>

⁹Note: You will just create propensity scores for caregiver layoff given background variables; causal inference comes in the second stage of the Challenge when outcomes are measured several years from now.

Caregiver layoff

K13. Since {MONTH AND YEAR COHORT CITY FIELDDED IN YR 9}, have you been laid off from your employer for any time?

YES	1
NO.....	2
REFUSED.....	-1
DON'T KNOW	-2



Job training¹⁰

¹⁰Learn more at

<http://www.fragilefamilieschallenge.org/job-training/>

¹¹Note: You will just create propensity scores for job training given background variables; causal inference comes in the second stage of the Challenge when outcomes are measured several years from now.

Job training¹⁰

Does job training for a caregiver **cause** collateral benefits for children?¹¹

¹⁰Learn more at

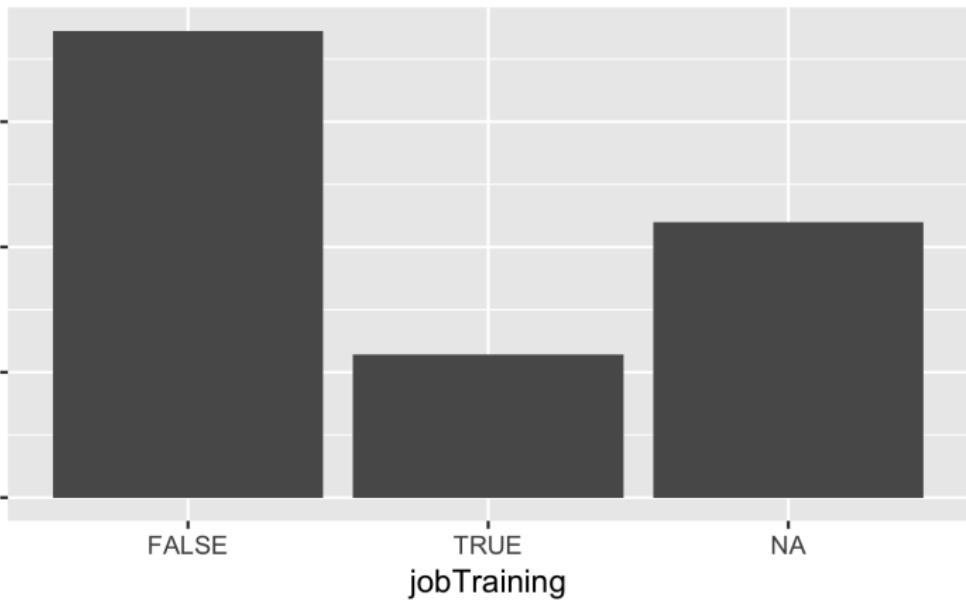
<http://www.fragilefamilieschallenge.org/job-training/>

¹¹Note: You will just create propensity scores for job training given background variables; causal inference comes in the second stage of the Challenge when outcomes are measured several years from now.

Caregiver job training

K4. Since {MONTH AND YEAR COHORT CITY FIELDDED IN YR 9}, have you taken any classes to improve your job skills, such as computer training or literacy classes?

- | | |
|------------------|-----|
| YES | .1 |
| NO | .2 |
| REFUSED | -.1 |
| DON'T KNOW | -.2 |



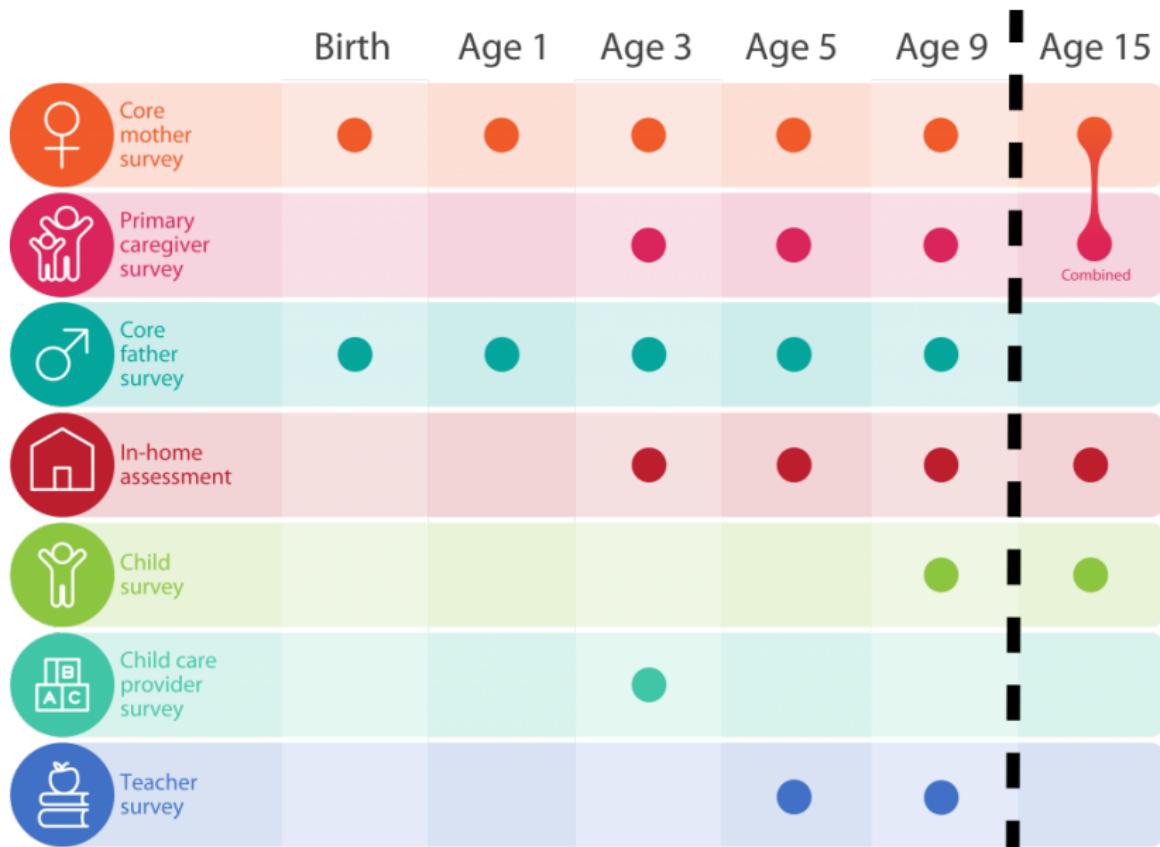
Introducing the data

The Fragile Families and Child Wellbeing Study is a dataset of real people who have selflessly opened up their lives to us for the last 15 years so that their experiences can contribute to scientific research. By participating in the Fragile Families Challenge, you become a collaborator in this project. It is of the utmost importance that you respect the families in the data by using what they have told us responsibly.

- ▶ Before you have access to the data, you will sign a data use agreement

- ▶ Before you have access to the data, you will sign a data use agreement
- ▶ After this activity you will delete the data from your computer

- ▶ Before you have access to the data, you will sign a data use agreement
- ▶ After this activity you will delete the data from your computer
- ▶ If you want to keep working with the data afterwards, you can apply for access through the Fragile Families website



RStudio Source Editor

fsf x Filter

	fsf1	fsf1a	fsf2	fsf3	fsf3a	fsf3b	fsf3b1	fsf4	fsf4a	fsf4b	fsf5	fsf6	fsf7a	fsf7b	fsf7c	fsf8a1	fsf8b1	fsf8c1	fsf8a2	fsf8b2	fsf8c2	fsf8a3
1	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
2	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
3	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
4	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
5	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
6	2	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	2	2	2	2	-6	-6	1	12	600	
7	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
8	1	3	2	-6.00	-6.00	-6.000000	-3	196.26903	1	2	2	2	2	2	2	-6	-6	2	-6	-6	-6	
9	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
10	1	1	1	-6.00	-6.00	-6.000000	-3	1310.50458	-10	2	2	2	2	2	2	-6	-6	2	-6	-6	-6	
11	1	5	4	149734....	11087.07	1.271596	-3	-6.00000	-6	-6	-6	-6	1	2	2	-6	-6	2	-6	-6	-6	
12	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
13	1	2	1	-6.00	-6.00	-6.000000	-3	1024.43600	3	2	2	2	2	2	2	-6	-6	2	-6	-6	-6	
14	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
15	1	5	3	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	2	-6	2	2	2	-6	-6	2	-6	-6	-6	
16	2	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	2	2	2	-6	-6	2	-6	-6	-6	
17	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
18	1	3	1	-6.00	-6.00	-6.000000	-3	877.82490	1	2	2	2	1	1	2	1	3	176	2	-6	-6	
19	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
20	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
21	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
22	1	1	1	-6.00	-6.00	-6.000000	-3	720.26305	1	2	2	2	1	2	2	-6	-6	2	-6	-6	-6	

Showing 1 to 24 of 4,242 entries



- ▶ A team of people¹² has spent months making the data easier to use.

¹²Alexander Kindel, Vineet Bansal, Kristin Catena, Thomas Hartshorne, Kate Jaeger, Dawn Koffman, Sara McLanahan, Maya Phillips, Shiva Rouhani, Ryan Vinh, Matthew Salganik.

- ▶ A team of people¹² has spent months making the data easier to use.
- ▶ Is it possible to do better than last time? Or, is there a fundamental limit with this data and this task?

¹²Alexander Kindel, Vineet Bansal, Kristin Catena, Thomas Hartshorne, Kate Jaeger, Dawn Koffman, Sara McLanahan, Maya Phillips, Shiva Rouhani, Ryan Vinh, Matthew Salganik.

- ▶ A team of people¹² has spent months making the data easier to use.
- ▶ Is it possible to do better than last time? Or, is there a fundamental limit with this data and this task?
- ▶ Let's see if it improves performance,

¹²Alexander Kindel, Vineet Bansal, Kristin Catena, Thomas Hartshorne, Kate Jaeger, Dawn Koffman, Sara McLanahan, Maya Phillips, Shiva Rouhani, Ryan Vinh, Matthew Salganik.

- ▶ A team of people¹² has spent months making the data easier to use.
- ▶ Is it possible to do better than last time? Or, is there a fundamental limit with this data and this task?
- ▶ Let's see if it improves performance, and let's see if you can help make this easier for future researchers.

¹²Alexander Kindel, Vineet Bansal, Kristin Catena, Thomas Hartshorne, Kate Jaeger, Dawn Koffman, Sara McLanahan, Maya Phillips, Shiva Rouhani, Ryan Vinh, Matthew Salganik.

How do I know what these variables are?

The screenshot shows a web browser window with the title bar "FFCWS Metadata | Variables". The address bar contains the URL "browse.fragilefamiliesmetadata.org/variables". The main content area features the "Fragile Families & Child Wellbeing Study" logo with "PRINCETON | COLUMBIA" below it. To the right of the logo is a grid of nine small photographs depicting various family and child scenarios. Below the logo is a navigation bar with links: "FFCWS Metadata", "Browse variables", "Download metadata", "Feedback", and "About FFCWS". The main section is titled "Search variables" and includes a search input field and a "Search »" button. A "Filter" sidebar on the left contains three dropdown menus: "Topic", "Wave", and "Respondent".

http:

//metadata.fragilefamilies.princeton.edu/variables

Introducing cm1relf

[metadata.fragilefamilies.princeton.edu/variables/
cm1relf](http://metadata.fragilefamilies.princeton.edu/variables/cm1relf)

You can filter variables by:

- ▶ Topic
- ▶ Wave
- ▶ Respondent
- ▶ Variable Type (e.g., continuous, ordered categorical, unordered categorical, etc).

Why cm1relf?

Why cm1relf?

Response type	Respondent	Wave	Leaf
(blank): questionnaire c: constructed	m - mother f - father p - primary caregiver k - child t - teacher h - home o - observations n - non parental caregiver d - child care center r - family care center u - post center observations q - couple	1 - baseline 2 - year 1 3 - year 3 4 - year 5 5 - year 9 6 - year 15	(letter): survey section + (number): question number OR (string): Constructed variable ID OR (string): national or city weight

Want direct access to the metadata?

Want direct access to the metadata? No problem

- ▶ API: `api.metadata.fragilefamilies.princeton.edu`
- ▶ Python package:
`github.com/fragilefamilieschallenge/ffmetadata-py`
- ▶ R package:
`github.com/fragilefamilieschallenge/ffmetadata`
- ▶ Raw metadata:
`api.fragilefamiliesmetadata.org/get_metadata`

FF Fragile Families & Child Wellbeing Study PRINCETON | COLUMBIA



Search... 

Home About People Publications Data and Documentation Contact

A red arrow points from the text above to the "Data and Documentation" menu item.

Spotlight on FFCWS and Incarceration Research

Recent publications using the Fragile Families & Child Wellbeing Study sample & available citations of the effects

General Documentation

Baseline

Year 1

[Links to](#)Year 3 [documentation](#)
for each wave

Year 5

Year 9

Year 15

FAQ

Data and Documentation

Data

Data are free to download from Princeton University's [Office of Population Research \(OPR\) data archive](#).

Currently, there are five waves of publicly available data including baseline and Year 1, Year 3, Year 5, and Year 9 follow-ups. In order to protect the confidentiality of survey respondents, geographic identifiers, medical records data, contextual data (i.e., census tract characteristics), macroeconomic indicators, and genetic biomarkers are not available in the public use data files. Researchers may apply for these data via a [restricted use contract](#).

Documentation

General Documentation

Baseline

Year 1

Year 3

Year 5

Year 9

Year 15

FAQ

Data Alerts

Contract Data

Year 9

The Year 9 follow-up wave of data collection took place from 2007 to 2010, which makes the data useful for researchers interested in the effects of the Great Recession on children and families. It is different from previous waves because the home visit was integral to the wave procedures. In previous waves, we conducted core interviews before proceeding to the in-home components. At year 9, our initial interview was with the child's primary caregiver (usually the mother) and we scheduled a home visit at the time of that initial interview. As part of the home visit, we interviewed focal children for the first time. We attempted teacher surveys through the mail. Similar to previous waves, we have core interviews with mothers and fathers. Restricted Data at this wave include [census tract characteristics](#) of mother and father residences, [macroeconomic indicators](#), administrative data on children's [school characteristics](#), and [genetic](#) data from saliva samples from the mother and focal child.

PRIMARY CAREGIVER

[Primary Caregiver Survey](#)

[Primary Caregiver Self-Administered](#)

SCALES

[Scales documentation](#)

MOTHER

[Questionnaire](#)

[Codebook](#)

**Each survey has
a questionnaire
and
a codebook**

FATHER

[Questionnaire](#)

[Codebook](#)

CHILD

[Child Survey](#)

[Home Visit Workbook](#)

[Interviewer Observations](#)

[Codebook](#)

TEACHER

[Questionnaire](#)

[Codebook](#)

Questionnaire:

BOX A3A2

IF PCG=BIOFATHER IN THE PCG IDENTIFIER IN THE SCREENER,
GO TO A3C.

ELSE IF PCG= NON-PARENT AND RELATIONSHIP = MATERNAL
GRANDPARENT(S), PATERNAL GRANDPARENT(S), OTHER
RELATIVES OR FRIEND IN THE PCG IDENTIFIER IN THE
SCREENER, GO TO A3B1A.

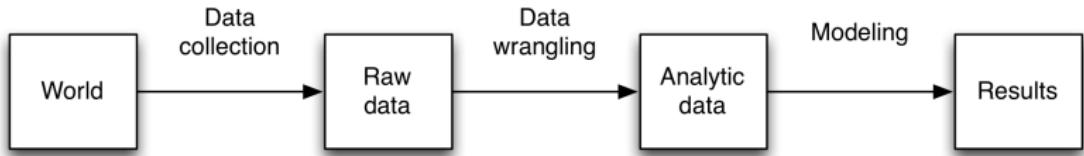
ELSE IF PCG=NON-PARENT AND RELATIONSHIP=FOSTER CARE
IN THE PCG IDENTIFIER IN THE SCREENER, GO TO A3B.

ELSE IF PCG=NON-PARENT AND RELATIONSHIP = OTHER,
SPECIFY IN THE PCG IDENTIFIER IN THE SCREENER, GO TO
A3B1A.

ELSE IF PCG= "NOT MOTHER" IN THE PCG IDENTIFIER GO TO
A3D.

A3B. Are {CHILD}'s foster parents related to you?

YES	1
NO	2
REFUSED.....	-1
DON'T KNOW.....	-2



Building a submission

Submissions include:

1. Predictions
2. Code
3. Narrative explanation

Submission preparation instructions:

www.fragilefamilieschallenge.org/upload-your-contribution/

Get on the leaderboard

← → ⌂ codalab.fragilefamilieschallenge.org/#results

 Fragile Families CHALLENGE

Help Sign Up Sign In



Results							
#	User	GPA ▲	Grit ▲	Material hardship ▲	Eviction ▲	Layoff ▲	Job training ▲
1	wjlei1990	0.36854 (1)	0.21896 (3)	0.02436 (1)	0.05341 (7)	0.17435 (5)	0.20224 (3)
2	OldDriver.ffc	0.37099 (2)	0.22979 (18)	0.02471 (2)	0.05341 (7)	0.17435 (5)	0.20224 (3)
3	yjpeng	0.37120 (3)	0.21759 (2)	0.02493 (3)	0.05223 (2)	0.17048 (1)	0.20169 (2)
4	hamidrezaomidvar	0.37136 (4)	0.22191 (13)	0.02523 (5)	0.05227 (3)	0.18784 (7)	0.21409 (7)
5	t.f.schaffner	0.37143 (5)	0.21755 (1)	0.02499 (4)	0.05660 (8)	0.22453 (9)	0.27736 (9)
6	andrewor	0.37143 (5)	0.21755 (1)	0.02499 (4)	0.06038 (10)	0.26792 (13)	0.30755 (13)
7	pc12	0.37583 (6)	6.18762 (29)	0.03536 (24)	0.94340 (18)	0.77547 (19)	0.72264 (18)
8	mannyg	0.37789 (7)	0.21997 (7)	0.02880 (17)	0.05341 (7)	0.17435 (5)	0.20224 (3)
9	ppz	0.37810 (8)	0.23896 (19)	0.02859 (14)	0.12830 (16)	0.30755 (15)	0.36981 (16)
10	lazs	0.38407 (9)	0.22054 (9)	0.02877 (16)	0.05660 (8)	0.22453 (9)	0.27736 (9)
11	mloyola	0.38644 (10)	0.21969 (4)	0.02880 (17)	0.05341 (7)	0.17435 (5)	0.20224 (3)
12	agalle	0.38846 (11)	0.22137 (11)	0.02740 (8)	0.05341 (7)	0.17435 (5)	0.20224 (3)
13	weggert	0.38868 (12)	0.24682 (21)	0.02546 (6)	0.05660 (8)	0.24528 (12)	0.29245 (11)
14	ieremvfreese	0.39077 (13)	0.22060 (10)	0.02803 (12)	0.05295 (5)	0.17379 (4)	0.20132 (1)

Powered by Codalab v0.1.1

Advice

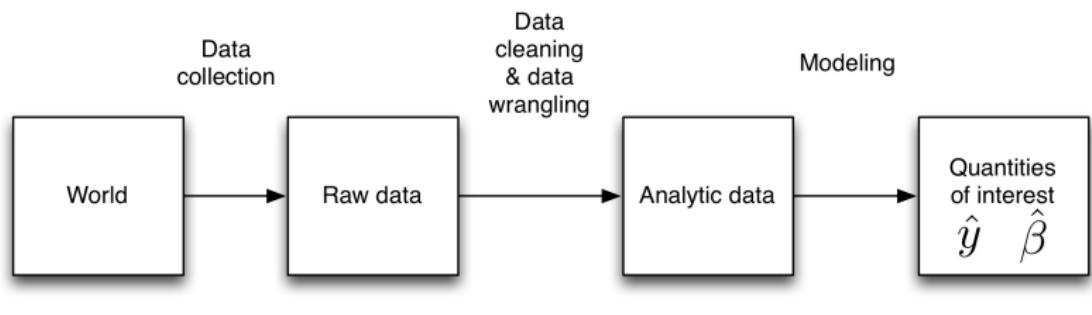
Most good approaches will likely involve

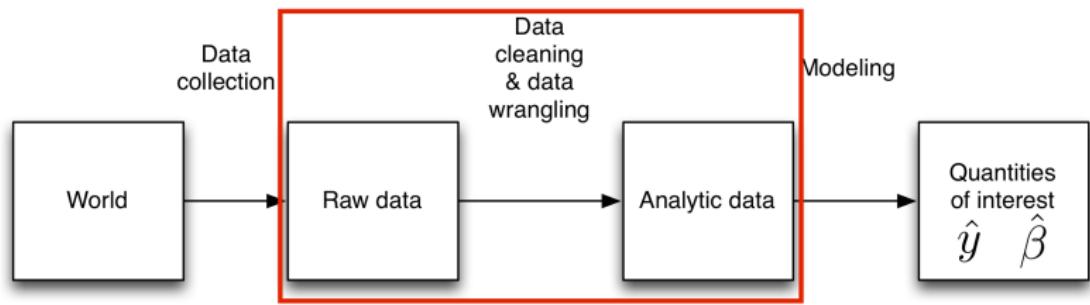
Most good approaches will likely involve

- ▶ careful data preparation

Most good approaches will likely involve

- ▶ careful data preparation
- ▶ flexible models that avoid overfitting





RStudio Source Editor

fsf x Filter

	fsf1	fsf1a	fsf2	fsf3	fsf3a	fsf3b	fsf3b1	fsf4	fsf4a	fsf4b	fsf5	fsf6	fsf7a	fsf7b	fsf7c	fsf8a1	fsf8b1	fsf8c1	fsf8a2	fsf8b2	fsf8c2	fsf8a3
1	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
2	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
3	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
4	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
5	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
6	2	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	2	2	2	2	-6	-6	1	12	600	
7	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
8	1	3	2	-6.00	-6.00	-6.000000	-3	196.26903	1	2	2	2	2	2	2	2	-6	-6	2	-6	-6	
9	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
10	1	1	1	-6.00	-6.00	-6.000000	-3	1310.50458	-10	2	2	2	2	2	2	2	-6	-6	2	-6	-6	
11	1	5	4	149734....	11087.07	1.271596	-3	-6.00000	-6	-6	-6	-6	1	2	2	-6	-6	2	-6	-6	-6	
12	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
13	1	2	1	-6.00	-6.00	-6.000000	-3	1024.43600	3	2	2	2	2	2	2	-6	-6	2	-6	-6	-6	
14	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
15	1	5	3	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	2	-6	2	2	2	-6	-6	2	-6	-6	-6	
16	2	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	2	2	2	-6	-6	2	-6	-6	-6	
17	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
18	1	3	1	-6.00	-6.00	-6.000000	-3	877.82490	1	2	2	2	1	1	2	1	3	176	2	-6	-6	
19	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
20	-6	-6	-6	-6.00	-6.00	-6.000000	-3	-6.00000	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	
21	-9	-9	-9	-9.00	-9.00	-9.000000	-3	-9.00000	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
22	1	1	1	-6.00	-6.00	-6.000000	-3	720.26305	1	2	2	2	1	2	2	-6	-6	2	-6	-6	-6	

Showing 1 to 24 of 4,242 entries

F1. Now I would like to ask you some questions about your housing situation. Have you moved {since {DATE OF LAST INTERVIEW}/in the last four years}?

YES 1

NO 2 ➔ GO TO F7

REFUSED..... -1 ➔ GO TO F7

DON'T KNOW..... -2 ➔ GO TO F7

BOX F1

IF THE FATHER IS MARRIED TO OR LIVING WITH THE MOTHER
(A4 = 1 OR 4), GO TO F24.
ELSE, GO TO F1.

- F1. Now I would like to ask you some questions about your housing situation. Have you moved {since {DATE OF LAST INTERVIEW}/in the last four years}?

YES 1

NO 2 ➡ GO TO F7

REFUSED..... -1 ➡ GO TO F7

DON'T KNOW..... -2 ➡ GO TO F7

F1A. How many times have you moved {since {DATE OF LAST INTERVIEW}/in the last four years}?

ENTER NUMBER OF MOVES

OR

REFUSED -1

DON'T KNOW -2

F2. What is your current housing situation? Please stop me when I read the statement that describes your situation. Do you . . .

CODE ONE

- Rent your own apartment or house, 1► GO TO F4
- Live with family or friends who rent and you contribute part of the rent, 2► GO TO F4
- Live with family or friends who rent but you do not pay rent, 3 ► GO TO F5
- Own your own home, 4► GO TO F3
- Live in a house or condo owned by another family member, 5► GO TO F3
- Live in temporary housing or a group shelter, or 6► GO TO F7
- Do you live in some other housing arrangement? (SPECIFY) 91► GO TO F3

HALFWAY HOUSE/TREATMENT

- FACILITY 8► GO TO F7
- JAIL/PRISON 9► GO TO F7
- ON THE STREET, HOMELESS 10► GO TO F7
- REFUSED -1► GO TO F4
- DON'T KNOW -2► GO TO F4

F3. Approximately, how much do you think {you/they} could sell this home for today?
PROBE FOR APPROXIMATE AMOUNT.

\$, , ,

ENTER AMOUNT

OR

REFUSED..... -1

DON'T KNOW..... -2

F3A. Approximately, how much do {you/they} owe on this house?

\$, , ,

ENTER AMOUNT

OR

REFUSED..... -1

DON'T KNOW..... -2

F3B1. Whose name is on the mortgage for this house?

CODE ONE

FATHER'S NAME ONLY 1► GO TO BOX F3C

MOTHER'S OR CURRENT PARTNER'S NAME
ONLY 2► GO TO BOX F3C

BOTH FATHER'S NAME AND MOTHER'S
OR CURRENT PARTNER'S NAMES 3► GO TO BOX F3C

FAMILY MEMBER(S) ON THE
MOTHER'S OR CURRENT PARTNER'S SIDE 4► GO TO BOX F3C

FAMILY MEMBER(S) ON THE FATHER'S
SIDE 5► GO TO BOX F3C

OTHER (SPECIFY)..... 91► GO TO BOX F3C

REFUSED..... -1► GO TO BOX F3C

DON'T KNOW..... -2► GO TO BOX F3C

BOX F3C

IF RESPONDENT LIVES IN A HOUSE OR CONDO OWNED BY
ANOTHER FAMILY MEMBER (F2=5), GO TO F4.

ELSE, GO TO F7.

F2. What is your current housing situation? Please stop me when I read the statement that describes your situation. Do you . . .

CODE ONE

- Rent your own apartment or house, 1► GO TO F4
- Live with family or friends who rent and you contribute part of the rent, 2► GO TO F4
- Live with family or friends who rent but you do not pay rent, 3 ► GO TO F5
- Own your own home, 4► GO TO F3
- Live in a house or condo owned by another family member, 5► GO TO F3
- Live in temporary housing or a group shelter, or 6► GO TO F7
- Do you live in some other housing arrangement? (SPECIFY) 91► GO TO F3

HALFWAY HOUSE/TREATMENT

- FACILITY 8► GO TO F7
- JAIL/PRISON 9► GO TO F7
- ON THE STREET, HOMELESS 10► GO TO F7
- REFUSED -1► GO TO F4
- DON'T KNOW -2► GO TO F4

This process is time consuming.

The New York Times

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights



Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.

Peter DaSilva for The New York Times

<https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html>

This process is [error prone](#).

Correction: “Her Support, His Support: Money, Masculinity, and Marital Infidelity”

American Sociological Review
80(3):469–95

This process is [error prone](#).

I made several errors related to the coding of missing data in my June 2015 *ASR* article, “Her Support, His Support: Money, Masculinity, and Marital Infidelity.” Upon discov-

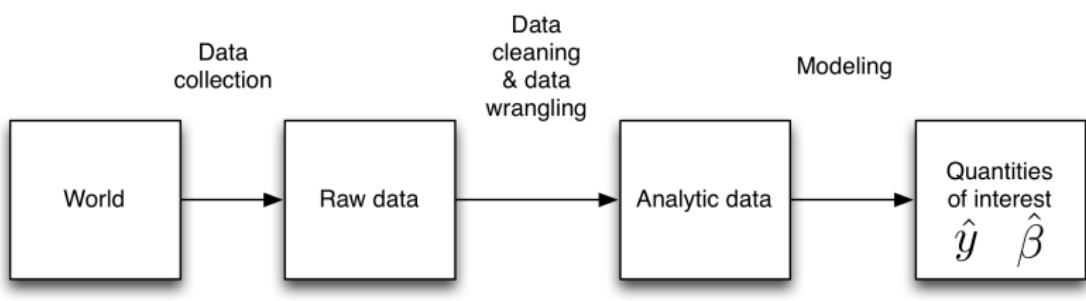
This process is **frustrating**.



<https://www.talk2solicitors.co.uk/blog/limits-frustration-contract/>

Now imagine this **times 100**.

Now imagine this **times 100**. We think this data preparation step is a critical barrier to applying machine learning methods to complex survey data.



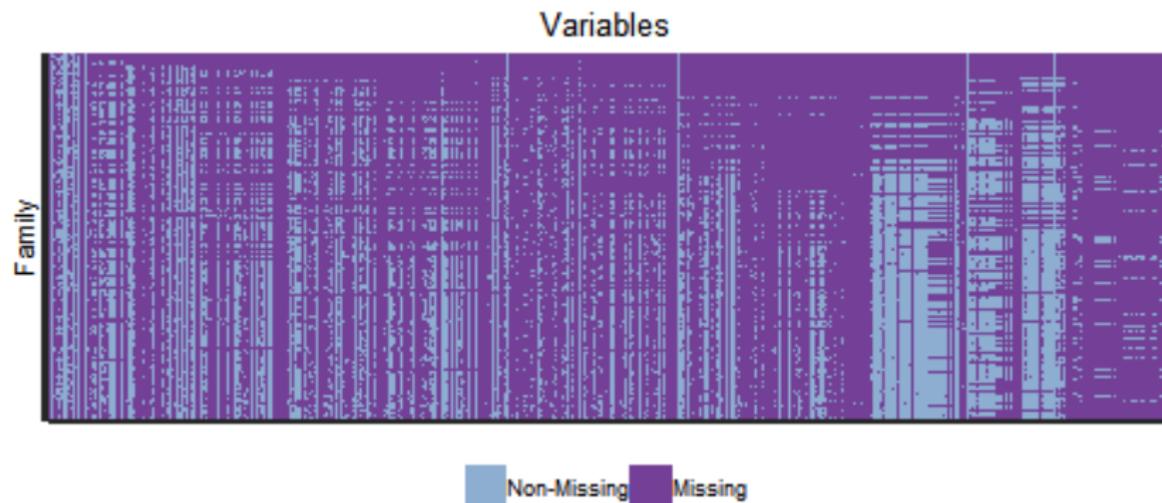
What if this data was as easy to use as your iPhone?

Human-readable → machine-actionable

- ▶ Big team effort: Alexander T. Kindel, Vineet Bansal, Kristin Catena, Tom Hartshorne, Kate Jaeger, Dawn Koffman, Sara McLanahan, Maya Philips, Shiva Rouhani, Ryan Vinh, Matthew J. Salganik

- ▶ Big team effort: Alexander T. Kindel, Vineet Bansal, Kristin Catena, Tom Hartshorne, Kate Jaeger, Dawn Koffman, Sara McLanahan, Maya Philips, Shiva Rouhani, Ryan Vinh, Matthew J. Salganik
- ▶ Key principle: metadata is data
- ▶ see Kindel et al (2018) “Improving metadata infrastructure for complex surveys: Insights from the Fragile Families Challenge

Data preparation



There are a lot of missing entries in the data matrix

Data preparation

Not all missing data is the same

- ▶ -9 Not in wave - Did not participate in survey/data collection component
- ▶ -6 Valid skip - Intentionally not asked question; question does not apply to respondent or response known based on prior information.
- ▶ -2 Don't know - Respondent asked question; responded "Don't Know".
- ▶ -1 Refuse - Respondent asked question; refused to answer question
- ▶ NA also used occasionally

Given that we have only a few hours, I would recommend not worrying about this. In the special issue, we saw some evidence that complex approaches to missing data didn't help much.

Data preparation

You'll want to deal with unordered categorical variables

F3B1. Whose name is on the mortgage for this house?

CODE ONE

FATHER'S NAME ONLY 1► GO TO BOX F3C

MOTHER'S OR CURRENT PARTNER'S NAME
ONLY 2► GO TO BOX F3C

BOTH FATHER'S NAME AND MOTHER'S
OR CURRENT PARTNER'S NAMES 3► GO TO BOX F3C

FAMILY MEMBER(S) ON THE
MOTHER'S OR CURRENT PARTNER'S SIDE 4► GO TO BOX F3C

FAMILY MEMBER(S) ON THE FATHER'S
SIDE 5► GO TO BOX F3C

OTHER (SPECIFY)..... 91► GO TO BOX F3C

REFUSED -1► GO TO BOX F3C

DON'T KNOW -2► GO TO BOX F3C

BOX F3C

IF RESPONDENT LIVES IN A HOUSE OR CONDO OWNED BY
ANOTHER FAMILY MEMBER (F2=5), GO TO F4.

ELSE, GO TO E7

What are the main “recipes” that people used in the Fragile Families Challenge?

- ▶ data preparation
- ▶ feature selection
- ▶ statistical learning

Here are the approaches that authors used for data preparation
(Key ideas relate to missing data and unordered categorical variables):

- ▶ Mean, median, mode imputation
- ▶ Missingness indicators
- ▶ Model-based imputation
- ▶ Imputation based on survey structure
- ▶ Constructed variables
- ▶ Principle component analysis
- ▶ One-hot encoding (dummy variables)
- ▶ Standardization
- ▶ Transformation

Here are the approaches that authors used for feature selection/variable selection (Key ideas relate to automatic vs manual):

- ▶ Prior expertise
- ▶ Study documentation
- ▶ Literature review
- ▶ F-test
- ▶ Mutual information
- ▶ Dropping low-variance features
- ▶ Multiple datasets
- ▶ Own train/test split

Here are the approaches that authors used for statistical learning
(Key ideas are flexibility and over-fitting):

- ▶ general linear model (linear and logistic regression)
- ▶ Tree-based methods (e.g., random forest, gradient boosted trees)
- ▶ Regularization approaches (e.g., LASSO, Ridge, Elastic Net)

What are the main “recipes” that people used in the Fragile Families Challenge?

- ▶ data preparation
- ▶ feature selection
- ▶ statistical learning

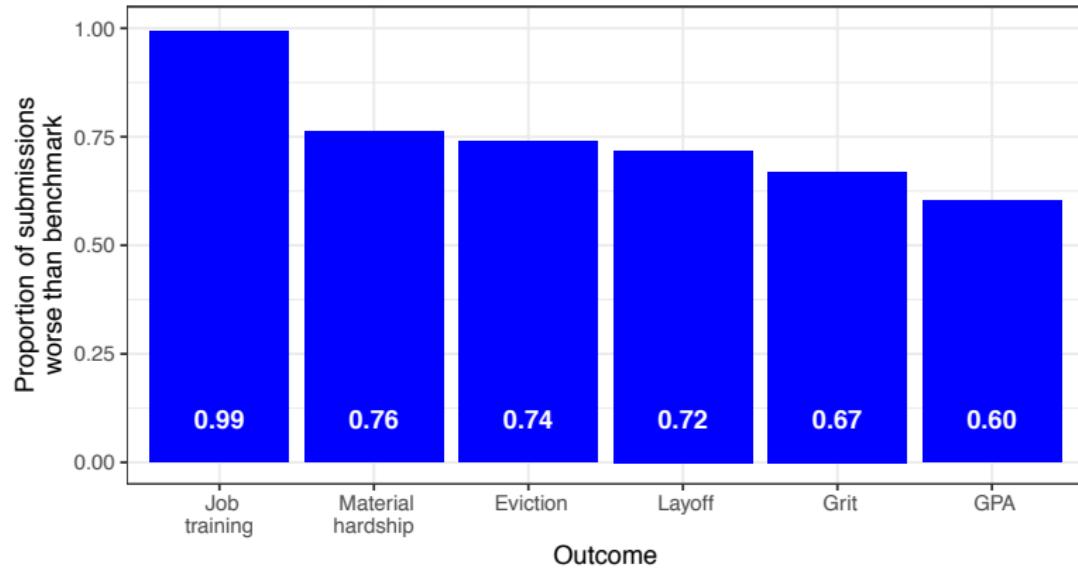
Getting started

Alex has created a template submission zip file that includes:

- ▶ predictions.csv (one prediction per family per outcome)
- ▶ narrative (a text file explaining the approach)
- ▶ baseline.R (code to run the submission)

You can upload it here:

<https://codalab.fragilefamilieschallenge.org/competitions/26>



Good luck and enjoy