

Human-in-the-Loop Approaches to the Fragile Families Challenge

Fragile Families Team

Part 1

Ridhi Kashyap, Allie Morgan, Kivan Polimis, Adaner Usmani

30 June 2017

Motivation

- ▶ Problem: High dimensional data with few observations
- ▶ Goals:
 - ▶ Increase predictive accuracy
 - ▶ Give back to the Fragile Families community
 - ▶ Develop machine-facilitated meta-analysis
- ▶ Approach: Human-in-the-loop feature engineering
 - ▶ Harness the substantive knowledge within the Fragile Families research community
 - ▶ But how?

First approach: Text mining

Convert 634 Fragile Family articles from PDF to plain text

- ▶ Goal: Understand which variables are associated with challenge outcomes
- ▶ Methods: Machines or people? or both?
 - ▶ Scrape regression tables
 - ▶ All variables?
 - ▶ Only variables that were significant?
 - ▶ Scrape text
 - ▶ Entire paper?
 - ▶ Abstract?

Example table

Table 1
Means and Standard Deviations for Demographic and Key Variables (N = 2,096)

Variables	Mean	SD
F age at Y9 (years)	37.26	7.16
F household income at Y9 (thousands)	57.03	58.04
	<i>n</i>	%
<i>F education</i>		
Less than high school	383	18.3
High school or equiv.	583	27.8
Some college	762	36.4
College or higher	368	17.6
<i>F ethnicity</i>		
White, non-Hispanic	475	22.7
Black, non-Hispanic	1012	48.3
Hispanic	530	25.3
Other	79	3.8
<i>F thought about M having an abortion</i>		
Yes	351	16.8
No	1739	83.2
<i>M-F Relationship status at Y9</i>		
Married	838	40.2
Cohabiting	243	11.6
Dating, non-cohabiting	46	2.2
Nonromantic	959	46.0
<i>Sex of focal child</i>		
Boy	1100	52.5
Girl	996	47.5
<i>Child reported closeness to F at Y9</i>		
Not very close	288	13.9
Fairly close	167	8.0
Quite close	384	18.5
Extremely close	1240	59.6

Note: M = Mother, F = Father.

Figure 1: pdf table

Example scraped table

```
In [6]: test_table = convert_pdf_to_table("/Users/allisonmorgan/Desktop/ff_pdfs/Adamsons -- A Longitudinal  
Investigation of Mothers and Fathers Initial Fathering Identities and Later Father-Child Relation  
ship Quality.pdf")
```

```
In [7]: test_table
```

Out[7]:

	F education	Unnamed: 1	Unnamed: 2
0	Less than high school	383	18.3
1	High school or equiv.	583	27.8
2	Some college	762	36.4
3	College or higher	368	17.6
4	F ethnicity	NaN	NaN
5	White, non-Hispanic	475	22.7
6	Black, non-Hispanic	1012	48.3
7	Hispanic	530	25.3
8	Other	79	3.8
9	F thought about M having an abortion	NaN	NaN
10	Yes	351	16.8
11	No	1739	83.2
12	M-F Relationship status at Y9	NaN	NaN
13	Married	838	40.2
14	Cohabiting	243	11.6
15	Dating, non-cohabiting	46	2.2
16	Nonromantic	959	46.0
17	Sex of focal child	NaN	NaN
18	Boy	1100	52.5

Figure 2: parsed table

KARI ADAMSONS^{Ph.D.*}

A LONGITUDINAL INVESTIGATION OF MOTHERS' AND FATHERS' INITIAL FATHERING IDENTITIES AND LATER FATHER-CHILD RELATIONSHIP QUALITY

Children benefit from high quality relationships with their fathers in a number of ways. However, little is known about the origins of father-child relationships. Here, identity theory and data from the Fragile Families dataset are used to investigate associations between mothers' and fathers' fathering identities at the time of the child's birth and nine years later, and the father-child relationship as reported by children at age nine. Neither mothers' nor fathers' role identity standards at birth were associated with father-child relationship quality, but greater father status centrality and not having considered abortion were associated with better father-child relationships. The association between abortion consideration and relationship quality was mediated by whether parents were romantically involved at Year 9. Implications for theory, policy, and practice are discussed.

Keywords: *father-child relationships, identity theory, father identity, fatherhood*

Figure 3: example abstract

Text mining abstracts

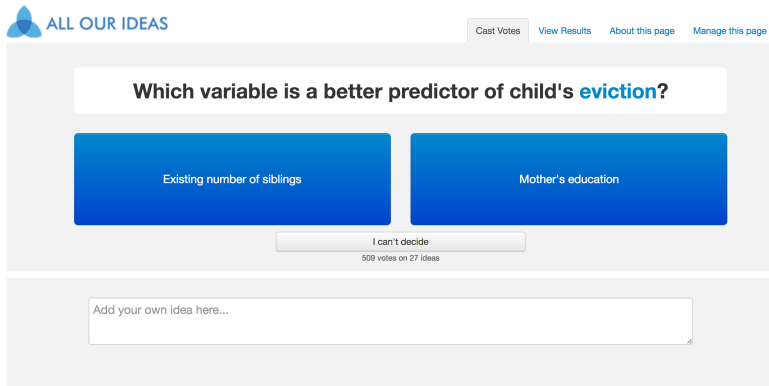
- ▶ Use Amazon Mechanical Turk (MTurk) to discover dependent variables and outcomes
- ▶ Pivot
 - ▶ Tables are messy
 - ▶ Focal relationships not immediately clear from variables in tables
 - ▶ Discovering cause and effect from abstracts is a difficult task for Turkers (or machines)

Second approach: Consult the experts

Wiki Survey

- ▶ Reach out to authors of Fragile Families challenge (expert opinion) to improve model feature selection
- ▶ Experts
 - ▶ Email every author in the Fragile Families Database
 - ▶ Used MTurk to find author's:
 - ▶ email
 - ▶ affiliation
 - ▶ discipline
 - ▶ some Turkers gave the author's current position (e.g., Assistant Professor)
- ▶ Create weights from Wiki Surveys to inform machine learning model priors

Wiki Survey



The image shows a web interface for a survey titled "ALL OUR IDEAS". At the top right, there are four links: "Cast Votes", "View Results", "About this page", and "Manage this page". The main question is "Which variable is a better predictor of child's **eviction**?". Below the question are three buttons: "Existing number of siblings", "Mother's education", and "I can't decide". Below the "I can't decide" button, it says "509 votes on 27 ideas". At the bottom, there is a text input field with the placeholder "Add your own idea here..." and a small icon of a person in the bottom right corner of the field.

ALL OUR IDEAS

[Cast Votes](#) [View Results](#) [About this page](#) [Manage this page](#)

Which variable is a better predictor of child's **eviction?**

Existing number of siblings

Mother's education

I can't decide

509 votes on 27 ideas

Add your own idea here...

Figure 4: Wiki Survey

Wiki Survey

outcomes	votes	users
eviction	431	13
GPA	270	11
grit	105	2
job training	0	0
layoffs	0	0
material hardship	531	19

- Low response rate and not all outcomes were answered (summer of auto-reply)

Third approach: Consult the masses

- ▶ Discover unmeasured and important factors with collective wisdom
- ▶ You guessed it: create a Wiki Survey for MTurk

outcomes	votes	users
eviction	3692	113
GPA	5062	135
grit	4355	110
job training	4305	127
layoffs	3821	115
material hardship	5486	127

- ▶ Compare and select features with the combined knowledge of experts and the public

User generated ideas

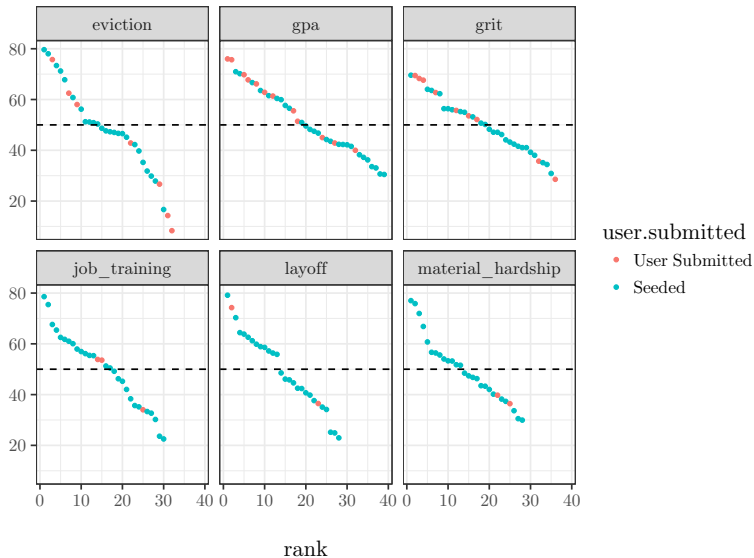


Figure 5: material hardship plot

Dot plots

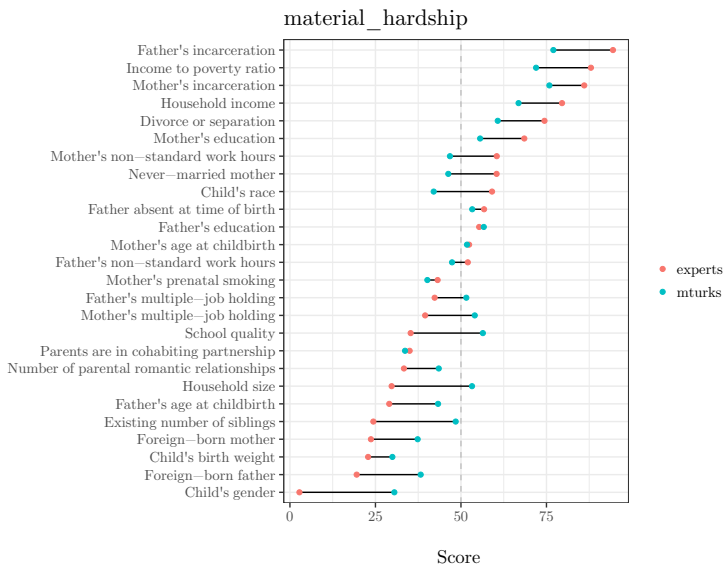


Figure 6: material hardship plot

Possible improvements

- ▶ Survey distribution
- ▶ Mapping ideas to variables
 - ▶ Consult Fragile Families administrators (e.g. does IQ map to PPVT score or Woodcock-Johnson)?
 - ▶ Idea moderation