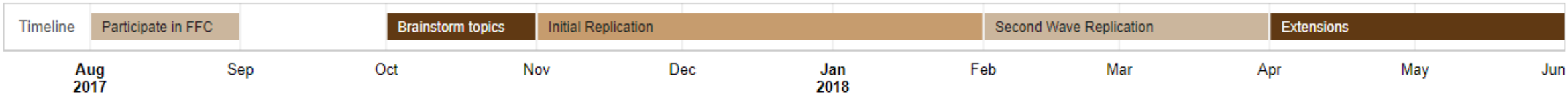
Computational Reproducibility in the Fragile Families Challenge

DAVID M. LIU

ADVISED BY: MATTHEW J. SALGANIK

Overview of the Presentation

1. **Background:** Initial Involvement
2. **Methodology:** Replication Stack
3. **Results:** The Guidelines
4. **Analysis:** Costs of Reproducibility
5. **Extension:** Ensemble and Causal Inference



The Issue: Reproducibility Crisis

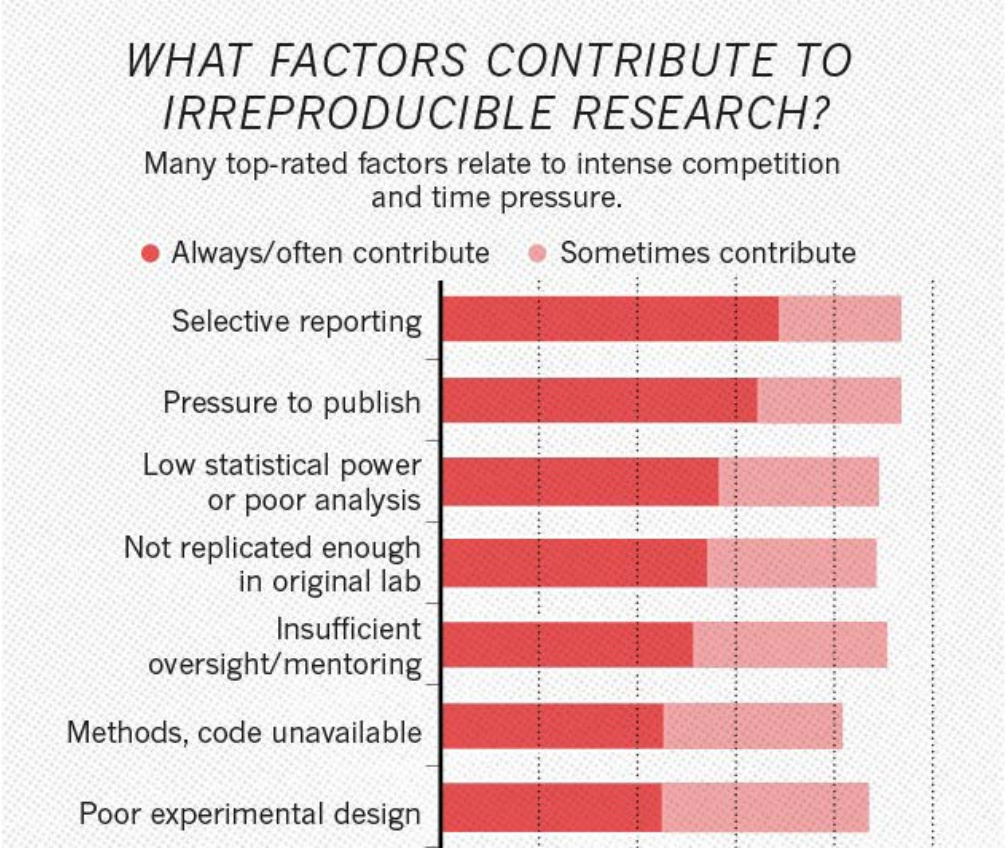
Replicable results are key to scientific findings

Nature survey finds that 70% of respondents have attempted and failed to replicate a study.

Particularly significant in Stats / ML

- Data is large and noisy
- Algorithms are stochastic
- Findings dependent on hardware

Active, unsolved issue plaguing science



The Issue: Reproducibility Crisis

The Fragile Families Challenge

Basic Facts:

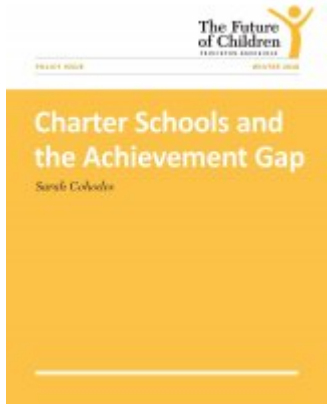
Longitudinal study behind *Future of our Children*

Kaggle-style predictive modeling challenge

- February - July 2017

Attracted submissions from social *and* data science

- COS 424



The Fragile Families Challenge

Basic Facts:

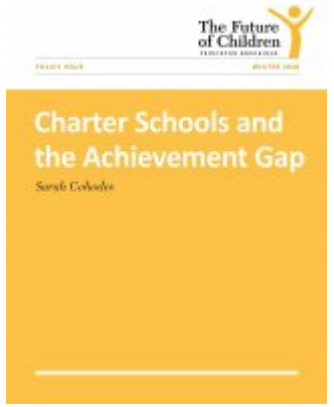
Longitudinal study behind *Future of our Children*

Kaggle-style predictive modeling challenge

- February - July 2017

Attracted submissions from social *and* data science

- COS 424



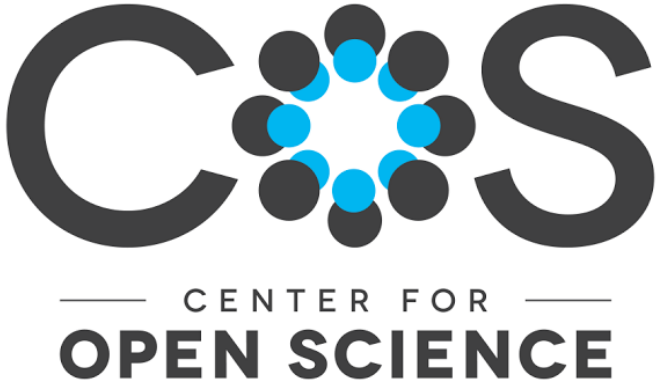
Main Question:

Can modern machine learning and statistical models be used for social prediction?

Socius and Computational Reproducibility

FFC submissions → special edition of social-science journal *Socius*

Setting a new standard for reproducibility in the social sciences



Enhance the Challenge's mission of collaboration

Effective contribution from computer science undergraduate

The Models

Feature Selection:

Lasso
Cross validation
Domain knowledge
None

The Models

Feature Selection:

Lasso
Cross validation
Domain knowledge
None

Imputation:

Amelia (R Package)
MICE
Naïve methods
Deletion

The Models

Feature Selection:

Lasso
Cross validation
Domain knowledge
None

Imputation:

Amelia (R Package)
MICE
Naïve methods
Deletion

Modeling:

Random Forest
GLM
Decision Trees
Neural Network

Docker

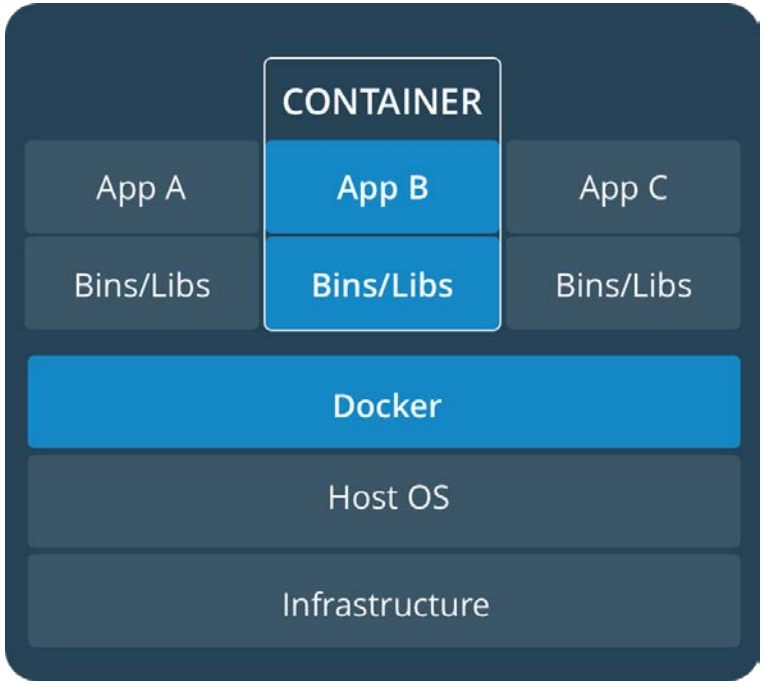
Software tool → research

Pros:

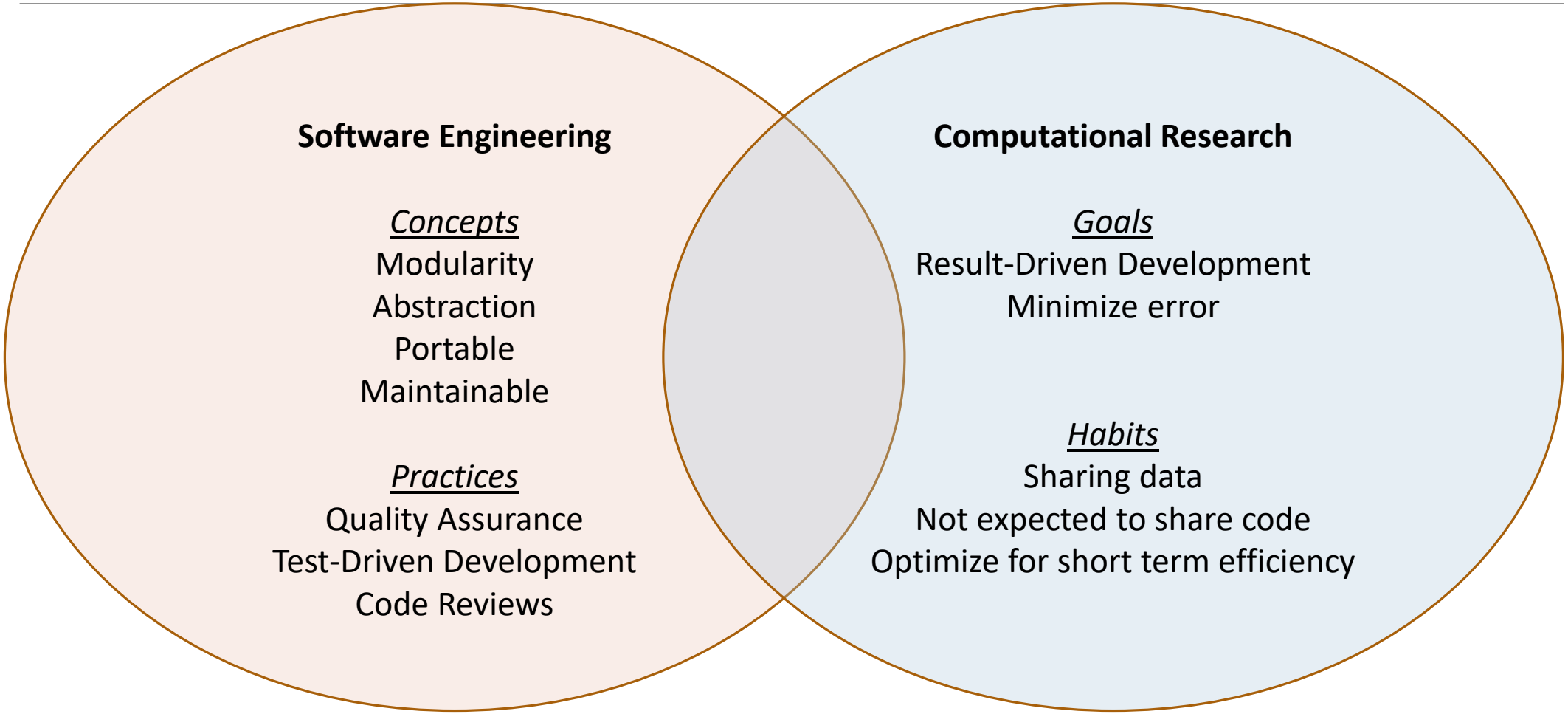
- Lighter weight than virtual machines
- Create a standardized environment
- Strong open source community

Cons

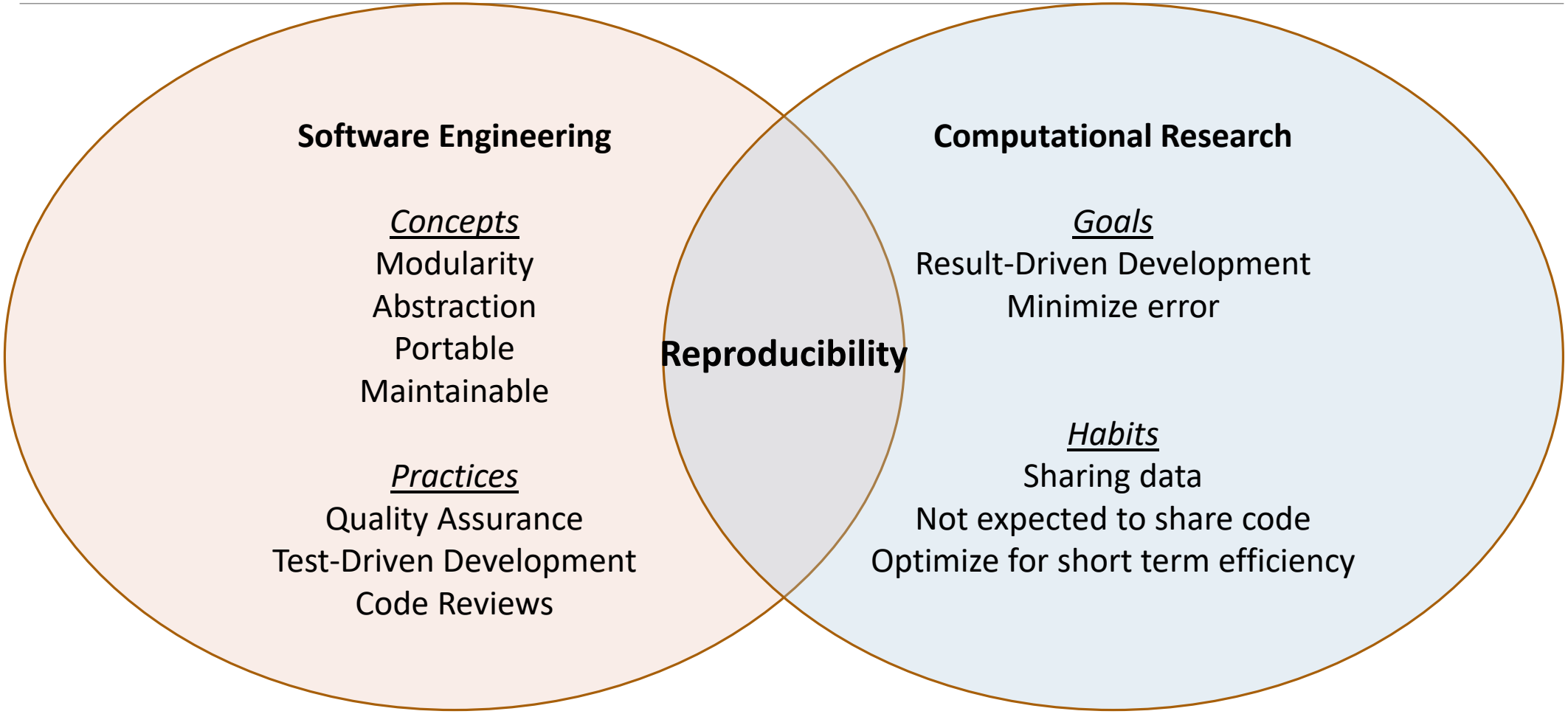
- Additional layer of between researcher and the code



Initial Challenges



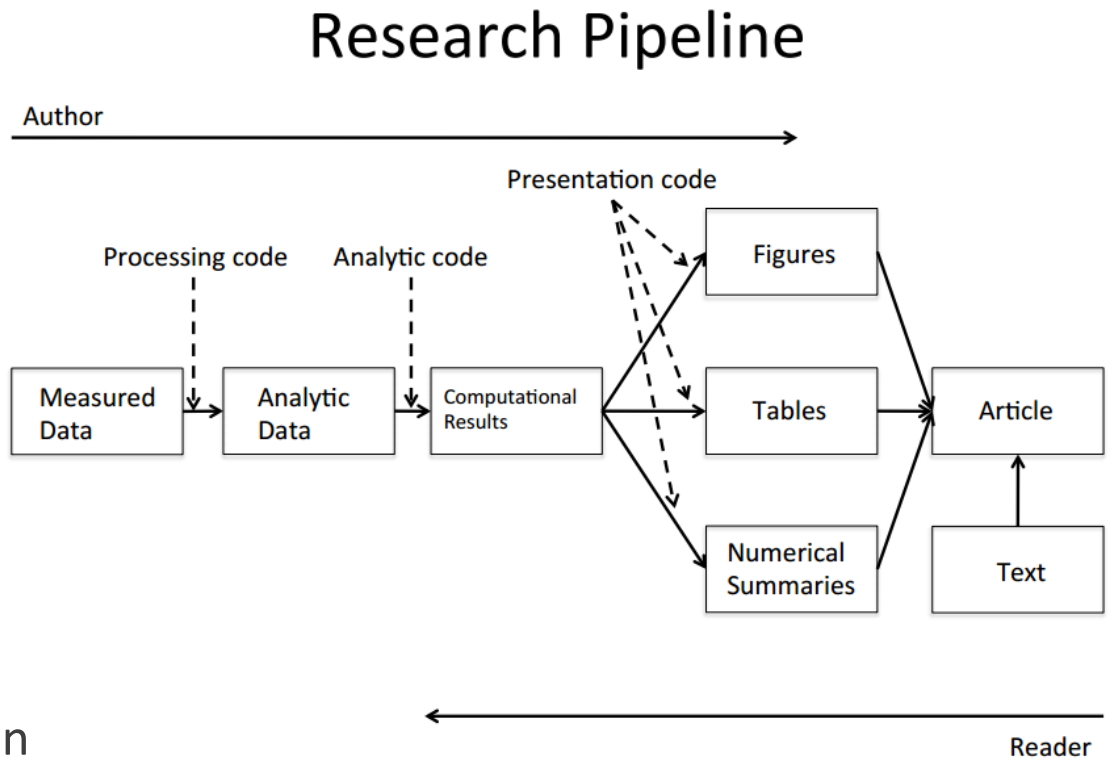
Initial Challenges



The Guidelines

A research pipeline:

- Modular
- Continuous
- Automated



Two recommendations:

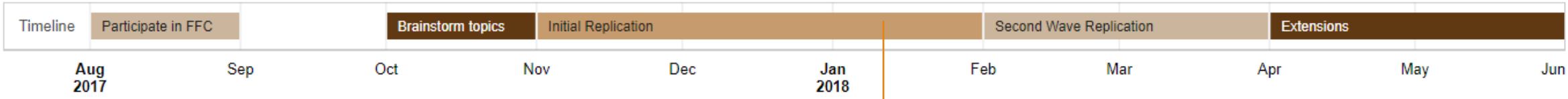
1. A run-all executable
2. Standardized directory structure

Supported by third-parties such as CodeOcean

Implementation

January 2018: sent a reproducibility memo to all authors as part of the peer review process

January – March: authors given two months to implement the guidelines.

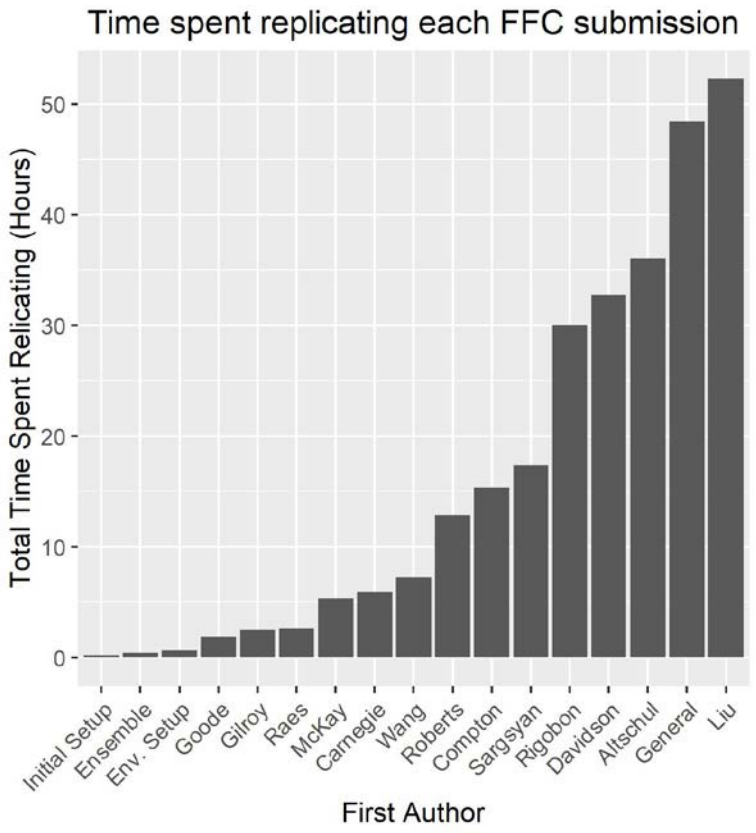


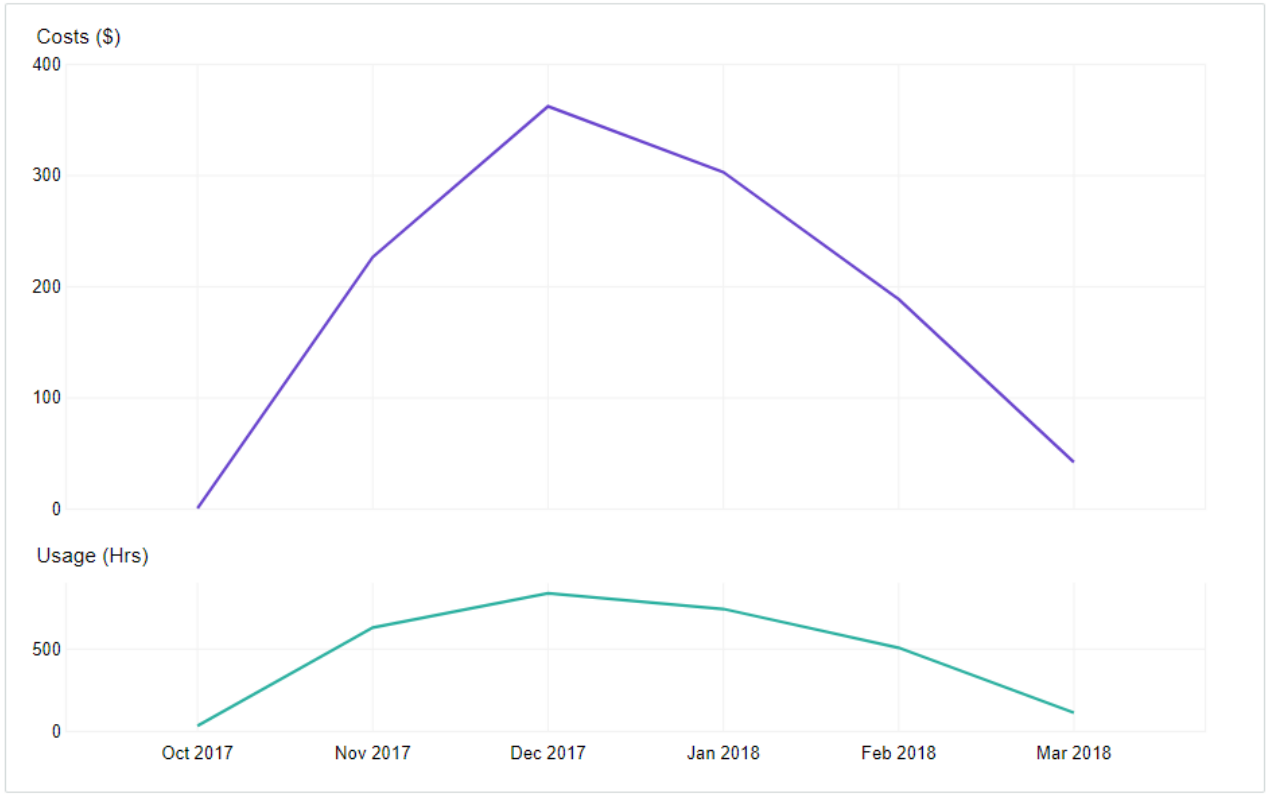
Reproducibility guidelines

Improvement

1. Authors have addressed bugs in the code
2. Submissions have been standardized in format
3. The models are executable and customizable.

Cost of Reproducibility





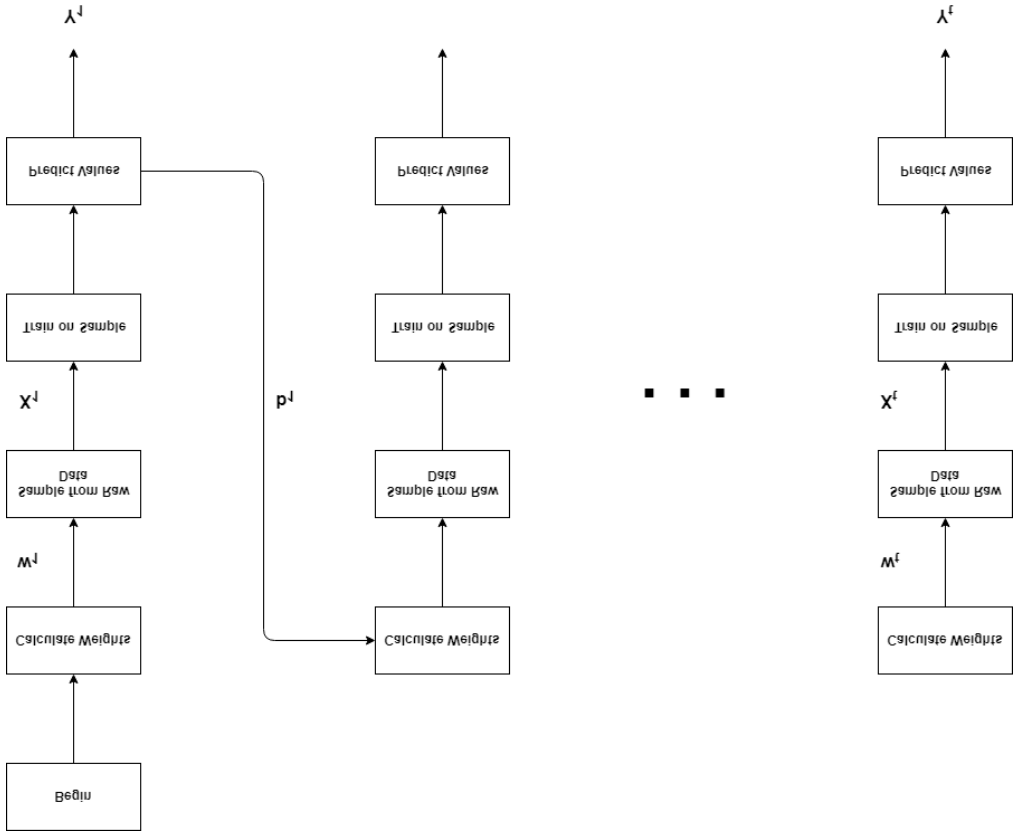
Final Steps and Future Work

Executing the models allows for applications:

- Ensembling
- Causal inference

Formalize guidelines into a research paper

Step in the direction of setting the norm.



Acknowledgements

Matthew J. Salganik

The Fragile Families and Child Wellbeing Study:

Sara McLanahan, Ian Lundberg, and Alex Kindel