Fragile Families CHALLENGE

Onur Varol
@onurvarol

Center for Complex Network Research
Northeastern University

Detection of Social Bots

- Behavior of social bots have more regular patterns
- Interactions and user activities has more granular data
- Feature engineering is possible and important aspect of the methodology
- Closed environment and most of the interactions are trackable

Social Scientists ⟷ Data Scientists

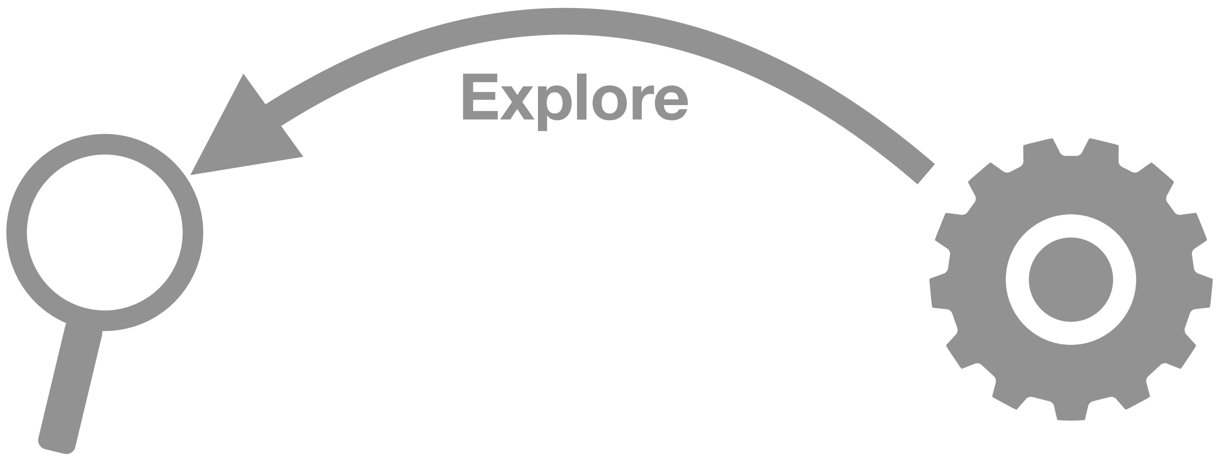
$$\hat{\beta} \quad \& \quad \hat{y}$$

Social Scientists $\longleftrightarrow$ Data Scientists

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

Social Scientists $\longleftrightarrow$ Data Scientists

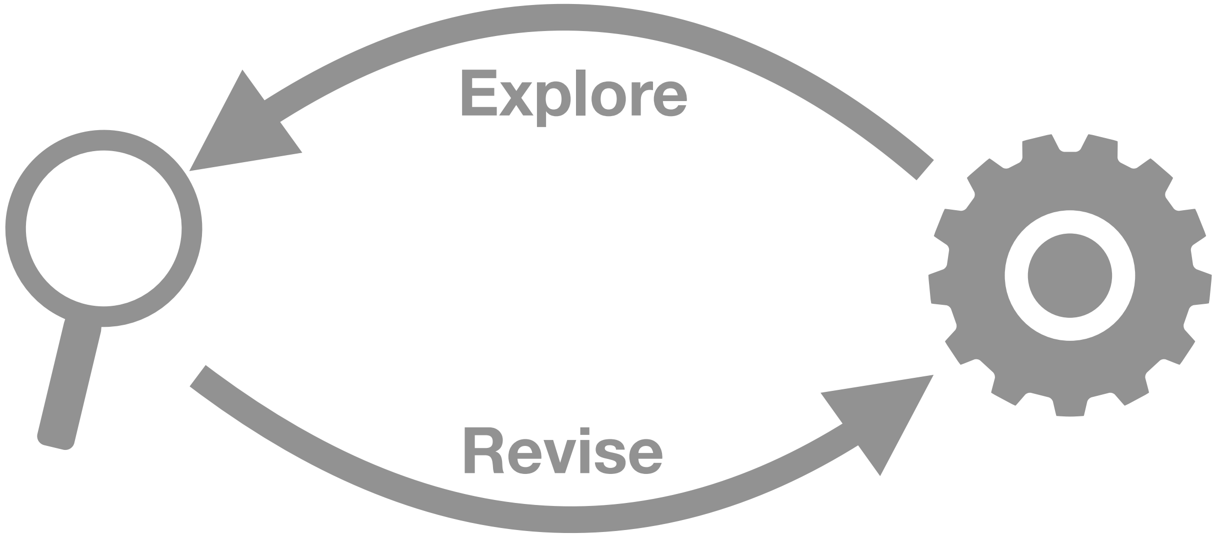
$$\hat{\beta} \quad \& \quad \hat{y}$$



Explore

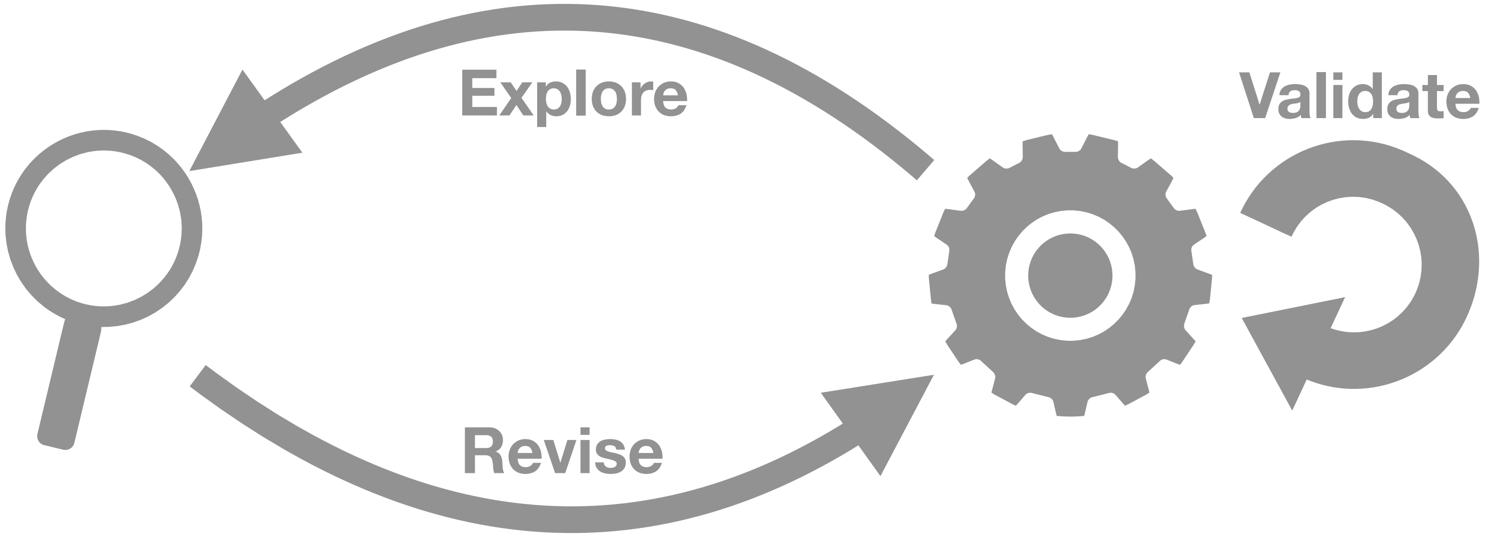
Social Scientists ⟷ Data Scientists

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



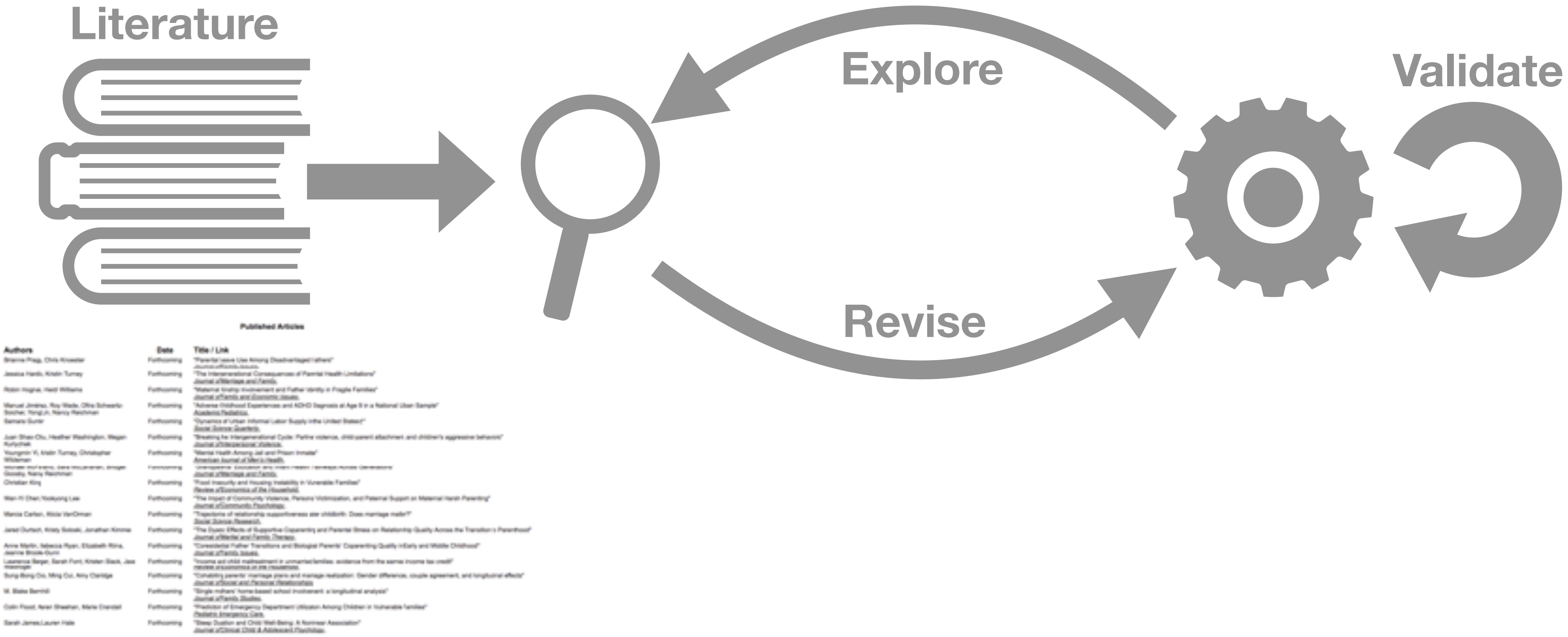
Social Scientists ⟷ Data Scientists

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



Social Scientists ⟷ Data Scientists

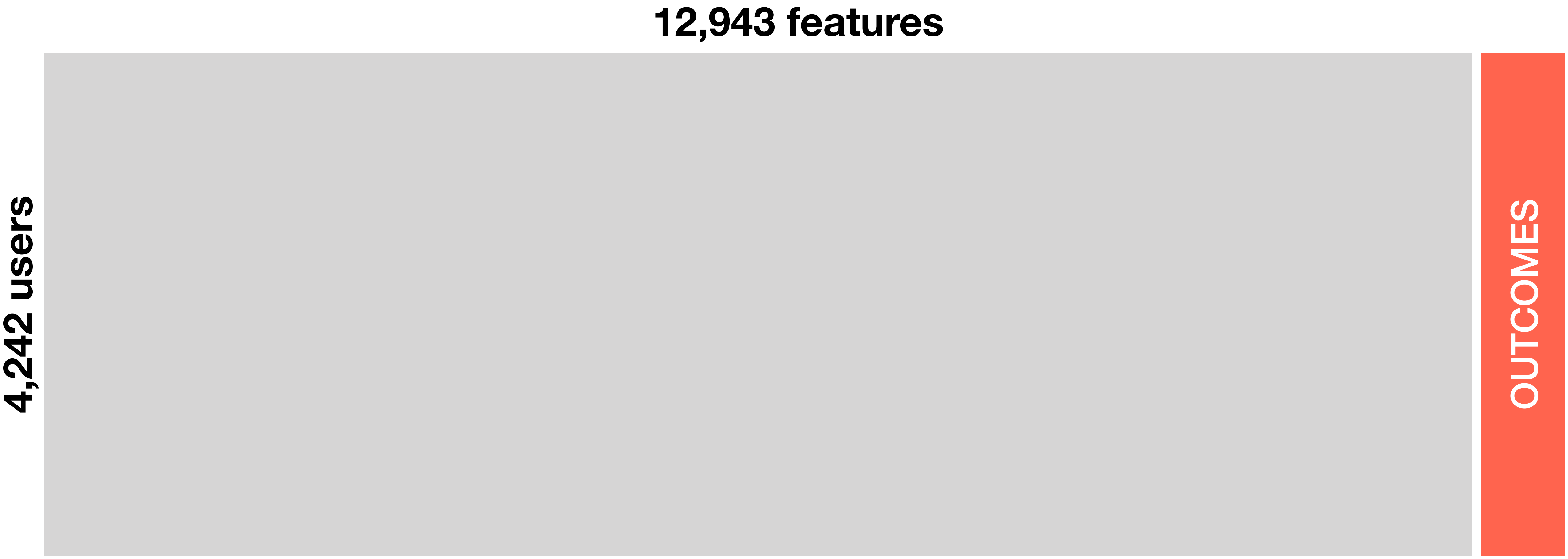
$$\hat{\beta} \quad \& \quad \hat{y}$$



Published Articles

Authors	Date	Title / Link
Brianne Pragg, Chris Knoester	Forthcoming	"Parental Leave Use Among Disadvantaged Fathers" Journal of Family Issues.
Jessica Hardie, Kristin Turney	Forthcoming	"The Intergenerational Consequences of Parental Health Limitations" Journal of Marriage and Family.
Robin Hognas, Heidi Williams	Forthcoming	"Maternal Kinship Involvement and Father Identity in Fragile Families" Journal of Family and Economic Issues.
Manuel Jiménez, Roy Wade, Ofira Schwartz-Soicher, Yong Lin, Nancy Reichman	Forthcoming	"Adverse Childhood Experiences and ADHD Diagnosis at Age 9 in a National Urban Sample" Academic Pediatrics.
Samara Gunter	Forthcoming	"Dynamics of Urban Informal Labor Supply in the United States?" Social Science Quarterly.
Juan Shao-Chiu, Heather Washington, Megan Kurlychek	Forthcoming	"Breaking the Intergenerational Cycle: Partner violence, child-parent attachment, and children's aggressive behaviors" Journal of Interpersonal Violence.
Youngmin Yi, Kristin Turney, Christopher Wildeman	Forthcoming	"Mental Health Among Jail and Prison Inmates" American Journal of Men's Health.
Michael McFarland, Sara McLanahan, Bridget Goosby, Nancy Reichman	Forthcoming	"Grandparents' Education and Infant Health: Pathways Across Generations" Journal of Marriage and Family.
Christian King	Forthcoming	"Food Insecurity and Housing Instability in Vulnerable Families" Review of Economics of the Household.
Wan-Yi Chen, Yookyong Lee	Forthcoming	"The Impact of Community Violence, Partners Victimization, and Paternal Support on Maternal Harsh Parenting" Journal of Community Psychology.
Marcia Carlson, Alicia VanOrman	Forthcoming	"Trajectories of relationship supportiveness after childbirth: Does marriage matter?" Social Science Research.
Jared Durtschi, Kristy Soloski, Jonathan Kimmes	Forthcoming	"The Dyadic Effects of Supportive Coparenting and Parental Stress on Relationship Quality Across the Transition to Parenthood" Journal of Marital and Family Therapy.
Anne Martin, Rebecca Ryan, Elizabeth Riina, Jeanne Brooks-Gunn	Forthcoming	"Coresidential Father Transitions and Biological Parents' Coparenting Quality in Early and Middle Childhood" Journal of Family Issues.
Lawrence Berger, Sarah Font, Kristen Slack, Jane Waldfogel	Forthcoming	"Income and child maltreatment in unmarried families: evidence from the earned income tax credit" Review of Economics of the Household.
Sung-Bong Cho, Ming Cui, Amy Claridge	Forthcoming	"Cohabiting parents' marriage plans and marriage realization: Gender differences, couple agreement, and longitudinal effects" Journal of Social and Personal Relationships.
M. Blake Berryhill	Forthcoming	"Single mothers' home-based school involvement: a longitudinal analysis" Journal of Family Studies.
Colin Flood, Karen Sheehan, Marie Crandall	Forthcoming	"Predictors of Emergency Department Utilization Among Children in Vulnerable Families" Pediatric Emergency Care.
Sarah James, Lauren Hale	Forthcoming	"Sleep Duration and Child Well-Being: A Nonlinear Association" Journal of Clinical Child & Adolescent Psychology.

Understanding Features

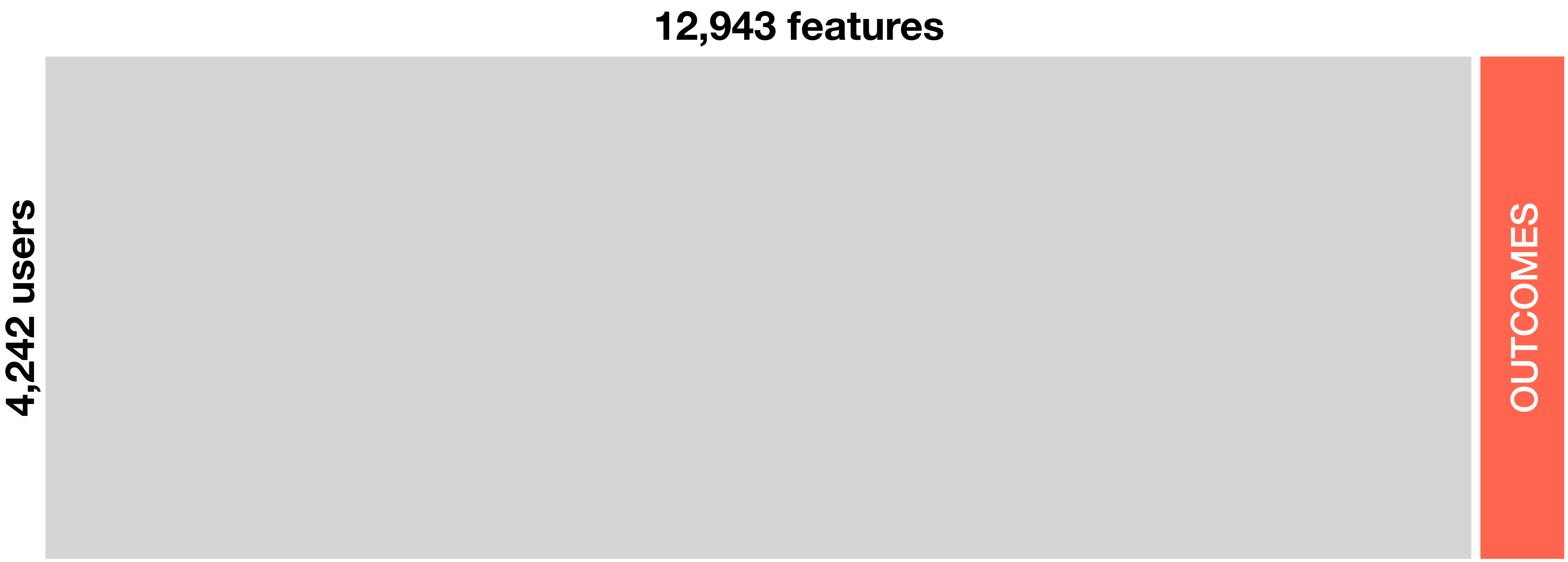


12,943 features

4,242 users

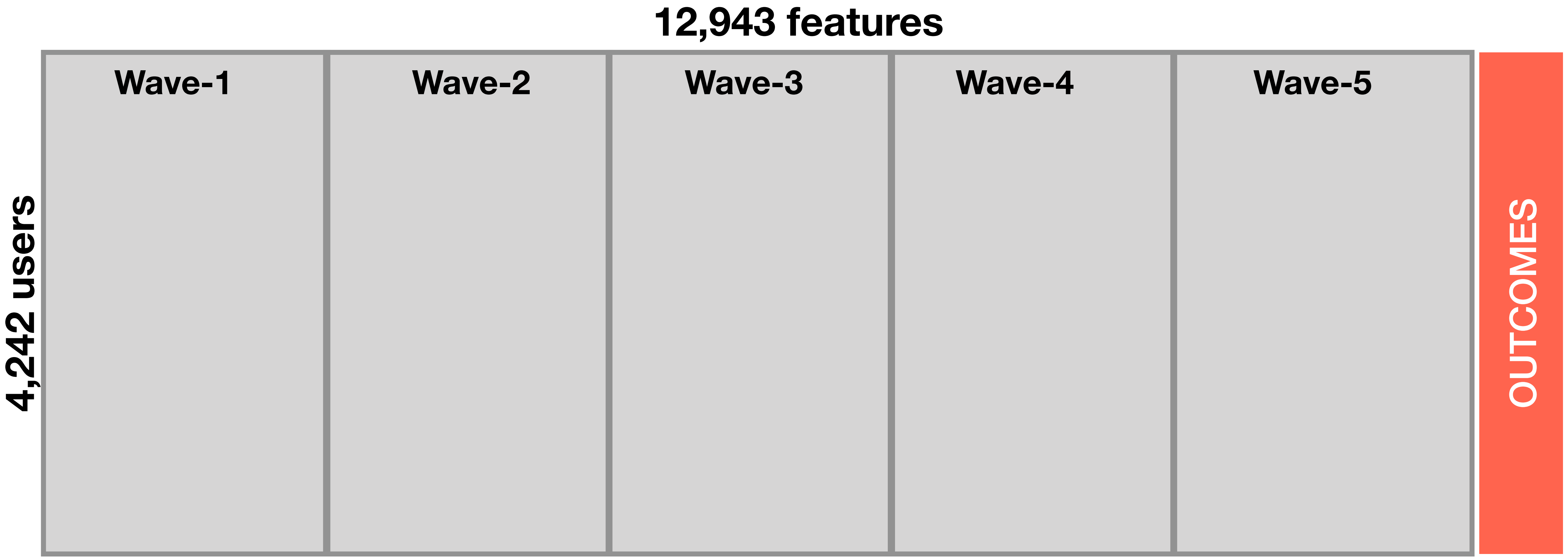
OUTCOMES

Understanding Features



Number of features are much higher than training data

Understanding Features



12,943 features

Wave-1

Wave-2

Wave-3

Wave-4

Wave-5

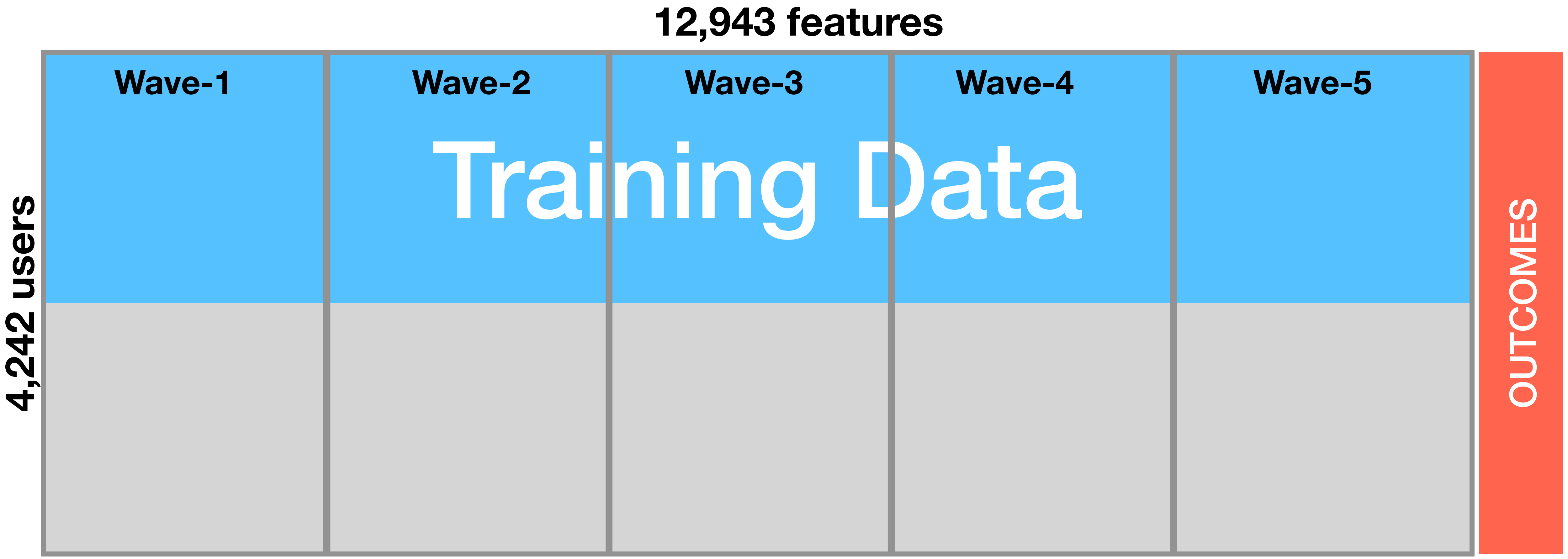
4,242 users

OUTCOMES

Number of features are much higher than training data

Collecting data in different waves

Understanding Features



12,943 features

Wave-1 Wave-2 Wave-3 Wave-4 Wave-5

Training Data

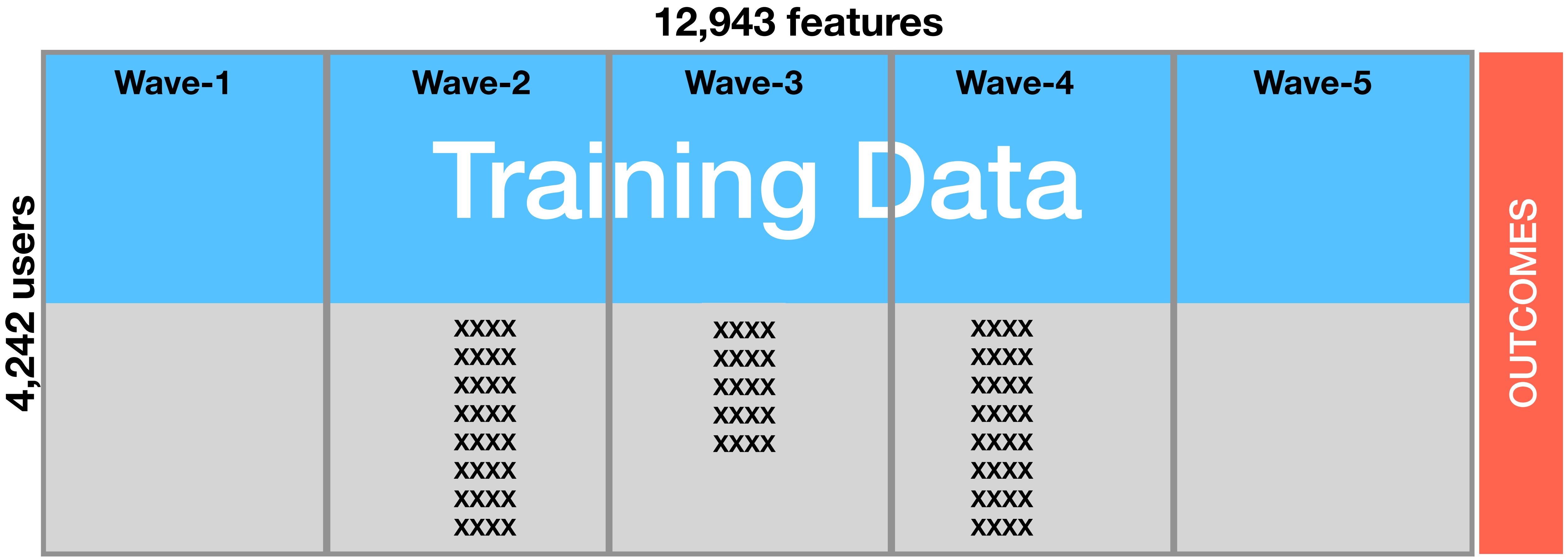
4,242 users

OUTCOMES

Number of features are much higher than training data

Collecting data in different waves

Understanding Features

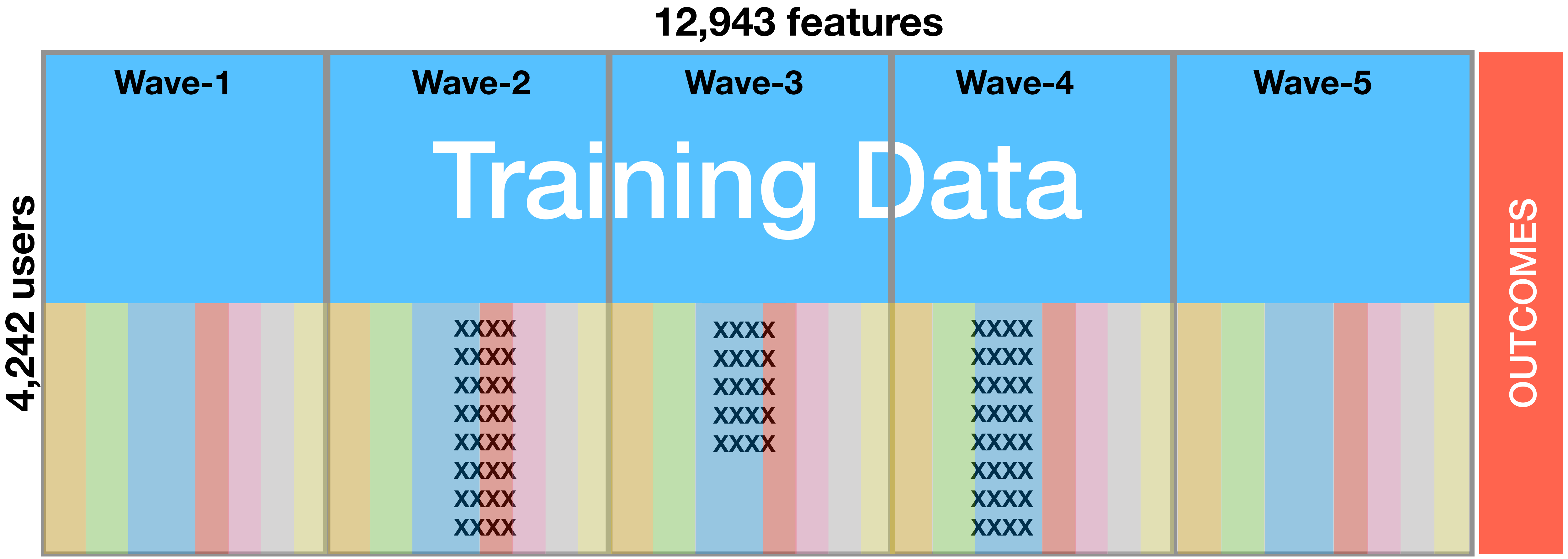


Number of features are much higher than training data

Collecting data in different waves

Systematically occurring missing values

Understanding Features



Number of features are much higher than training data

Collecting data in different waves

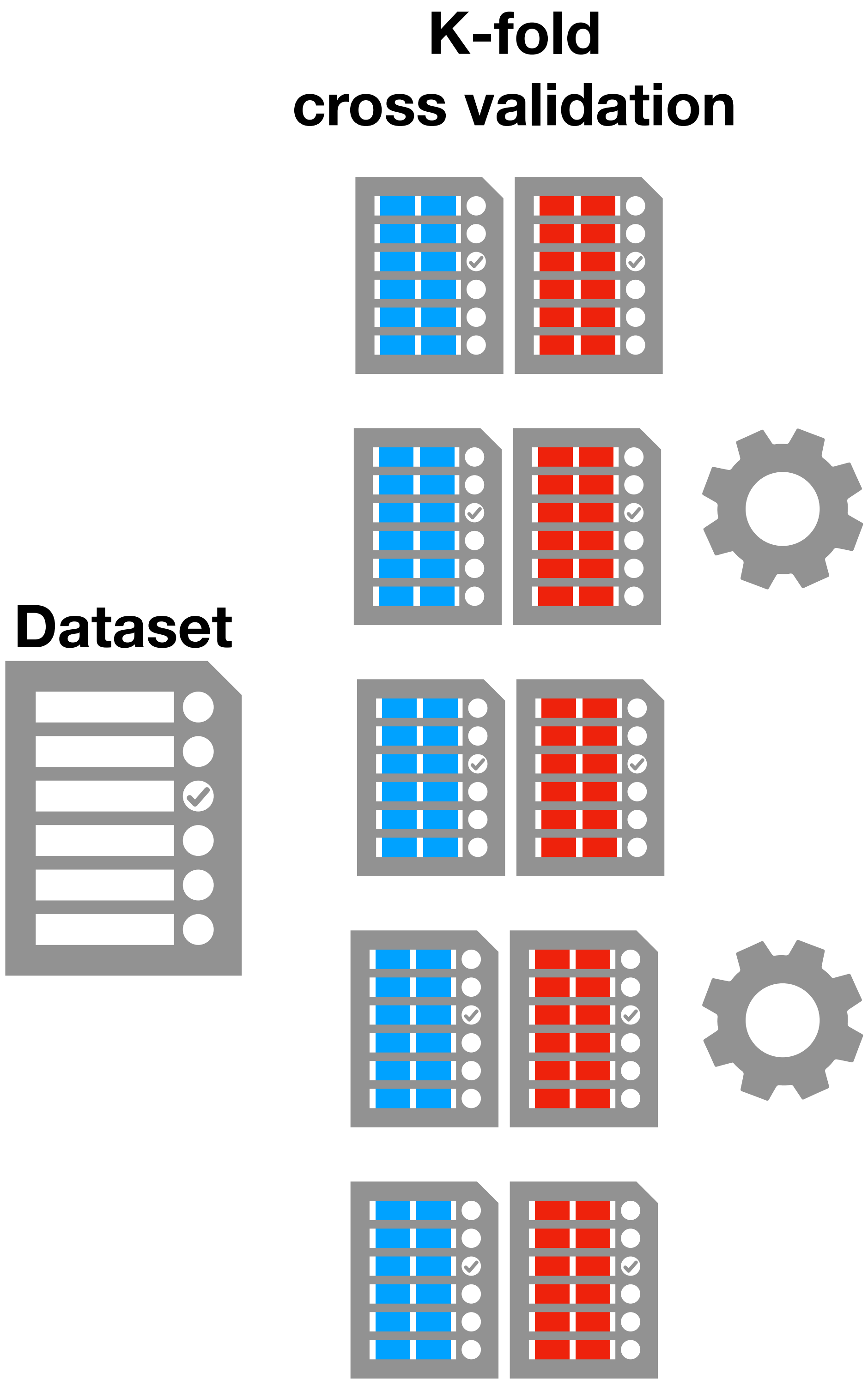
Systematically occurring missing values

Categories by different survey respondents: father, mother, kid, teacher, etc.

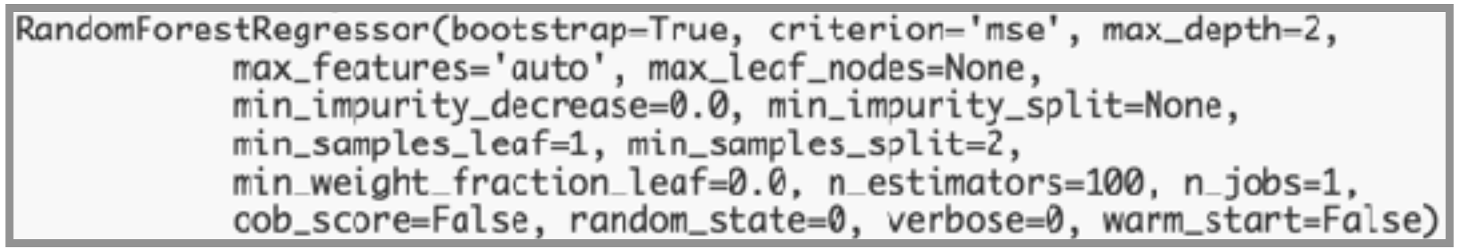
Feature filtering

- Removing features that has missing values
- Implementing text based feature filtering based on:
    - Keywords in the survey text
    - Wave number
    - Survey respondent
- Type of data: continuous vs. discrete, number of unique value etc.

Model building

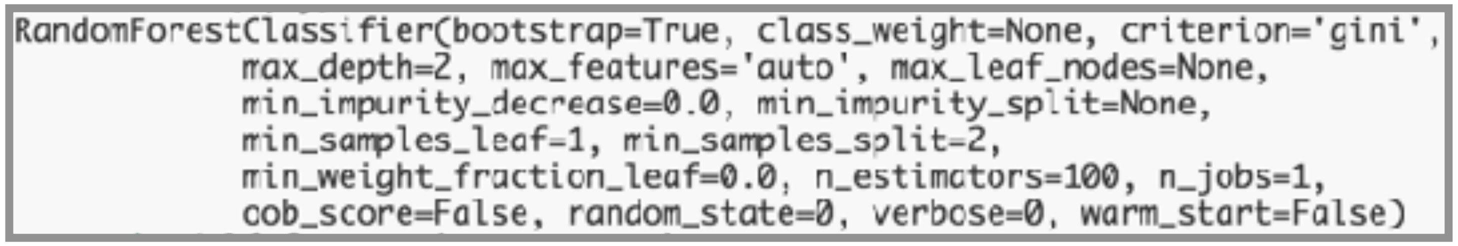


Grit, GPA, Material Hardship

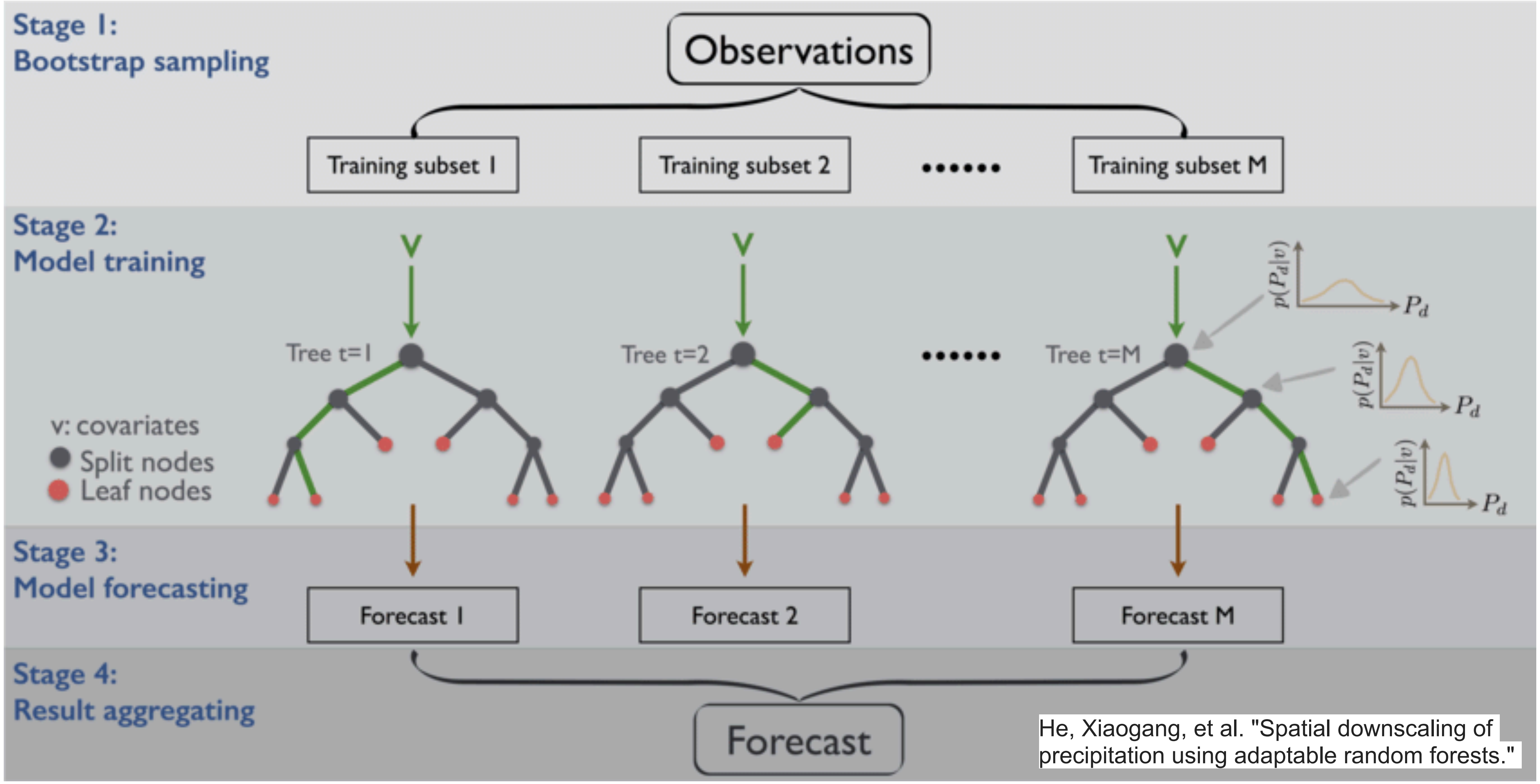


```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=2,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_decrease=0.0, min_impurity_split=None,
           min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=1,
           oob_score=False, random_state=0, verbose=0, warm_start=False)
```

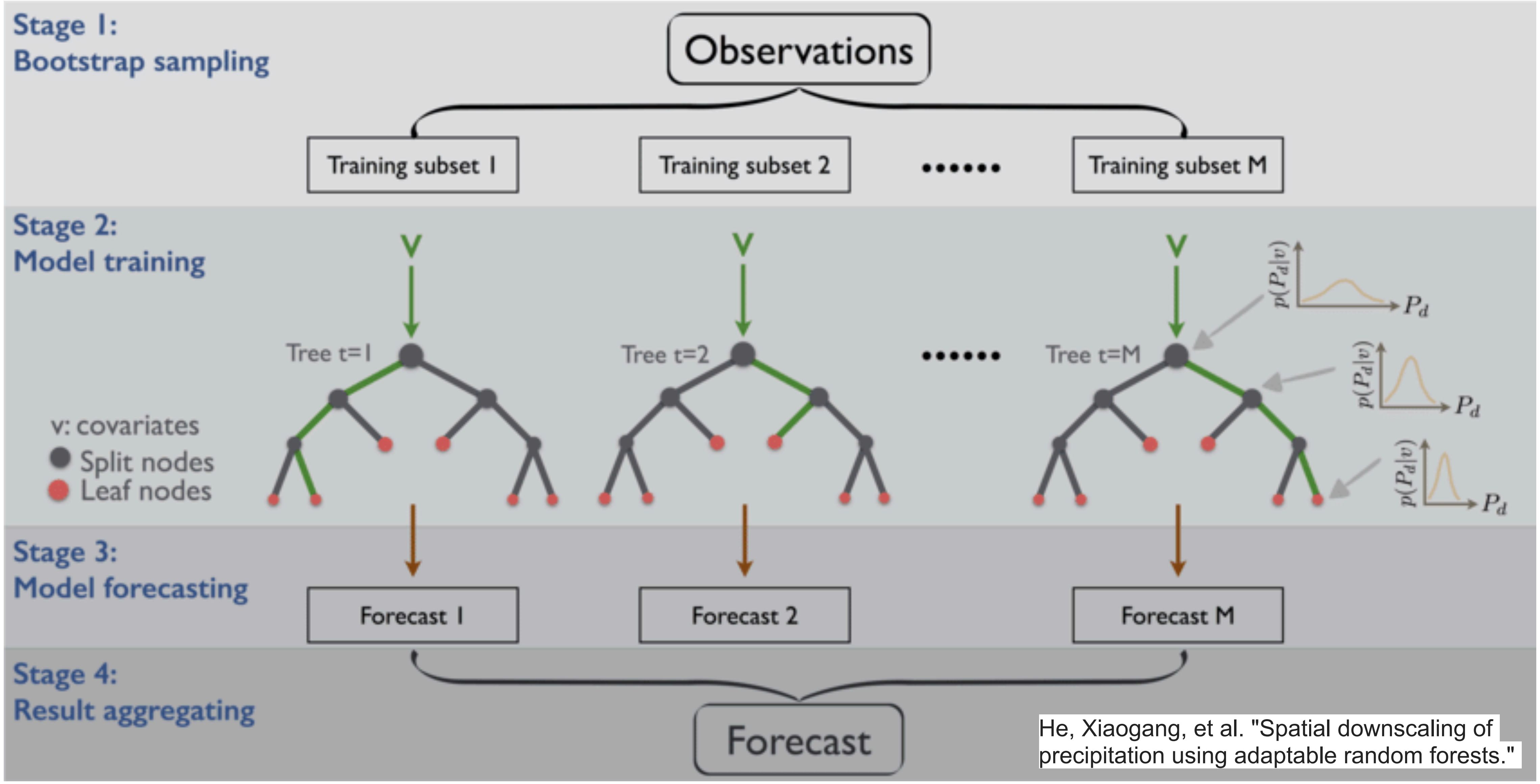
Eviction, Layoff, Job Training



```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=2, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=1,
            oob_score=False, random_state=0, verbose=0, warm_start=False)
```



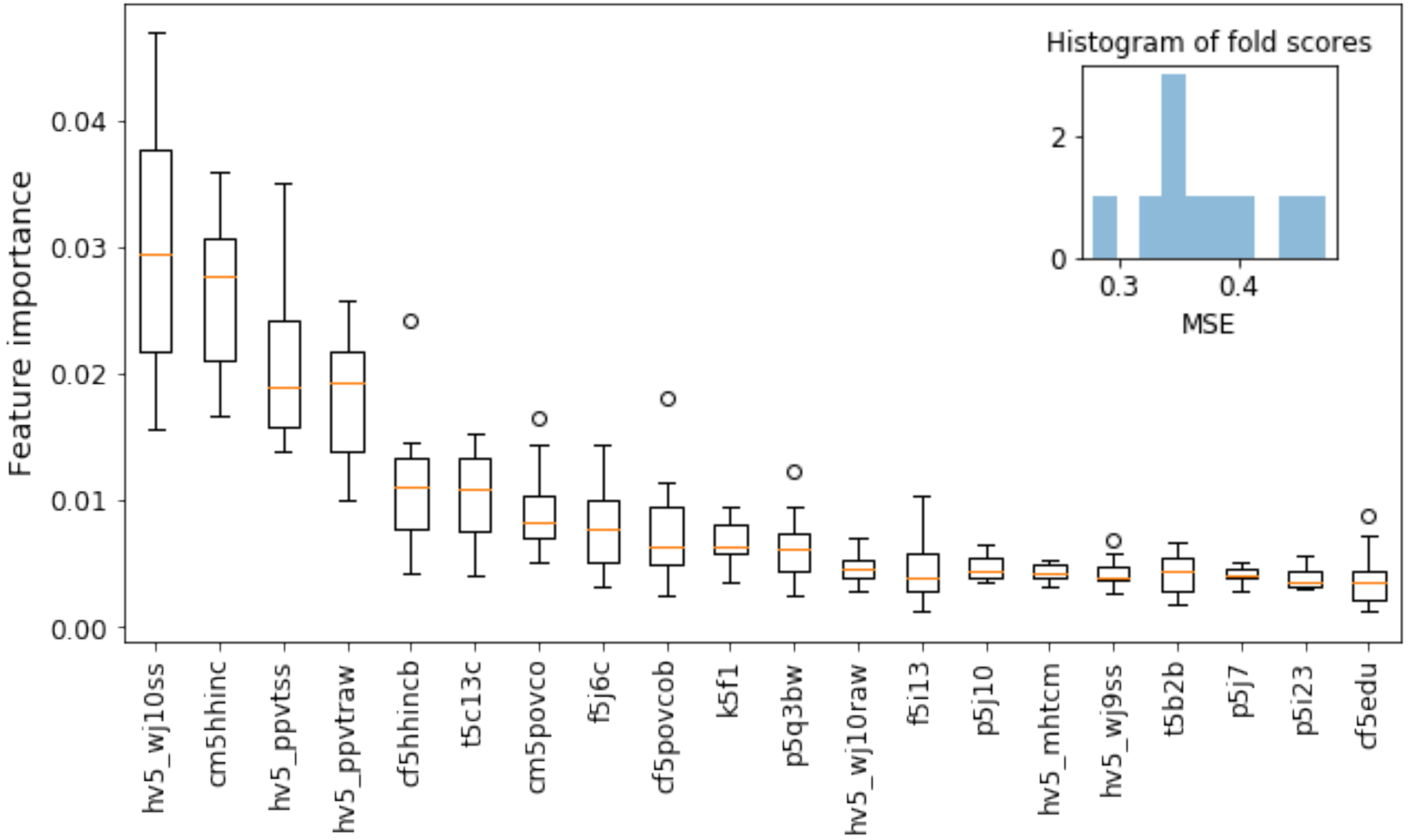
He, Xiaogang, et al. "Spatial downscaling of precipitation using adaptable random forests."



He, Xiaogang, et al. "Spatial downscaling of precipitation using adaptable random forests."

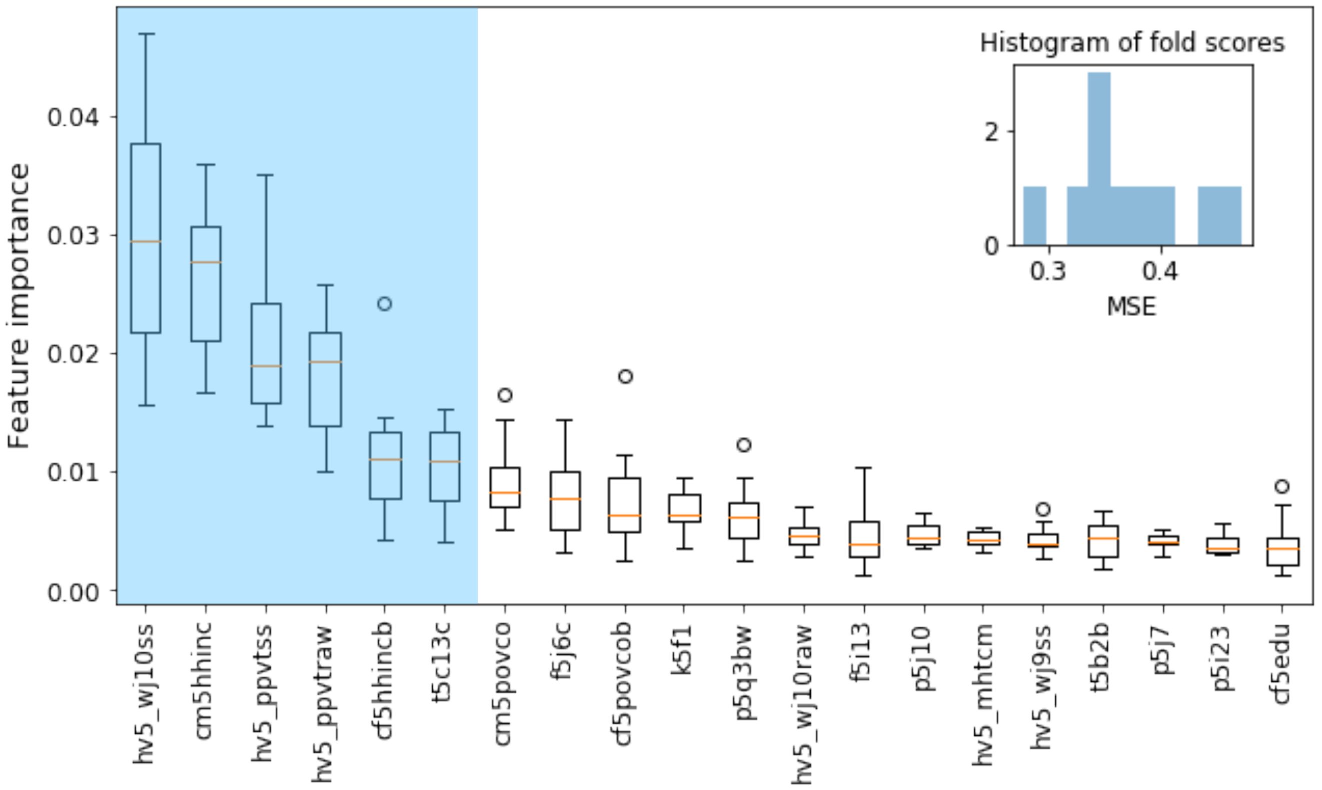
$$\boxed{\begin{array}{l}\hat{y}=\hat{f}_1(x)\\ \hat{y}=\hat{f}_2(x)\\ \hat{y}=\hat{f}_3(x)\end{array}} \longrightarrow \overset{\text{Community model}}{\hat{w}_1\hat{f}_1(x)+\hat{w}_2\hat{f}_2(x)+\hat{w}_3\hat{f}_3(x)}$$

Feature Importance Scores



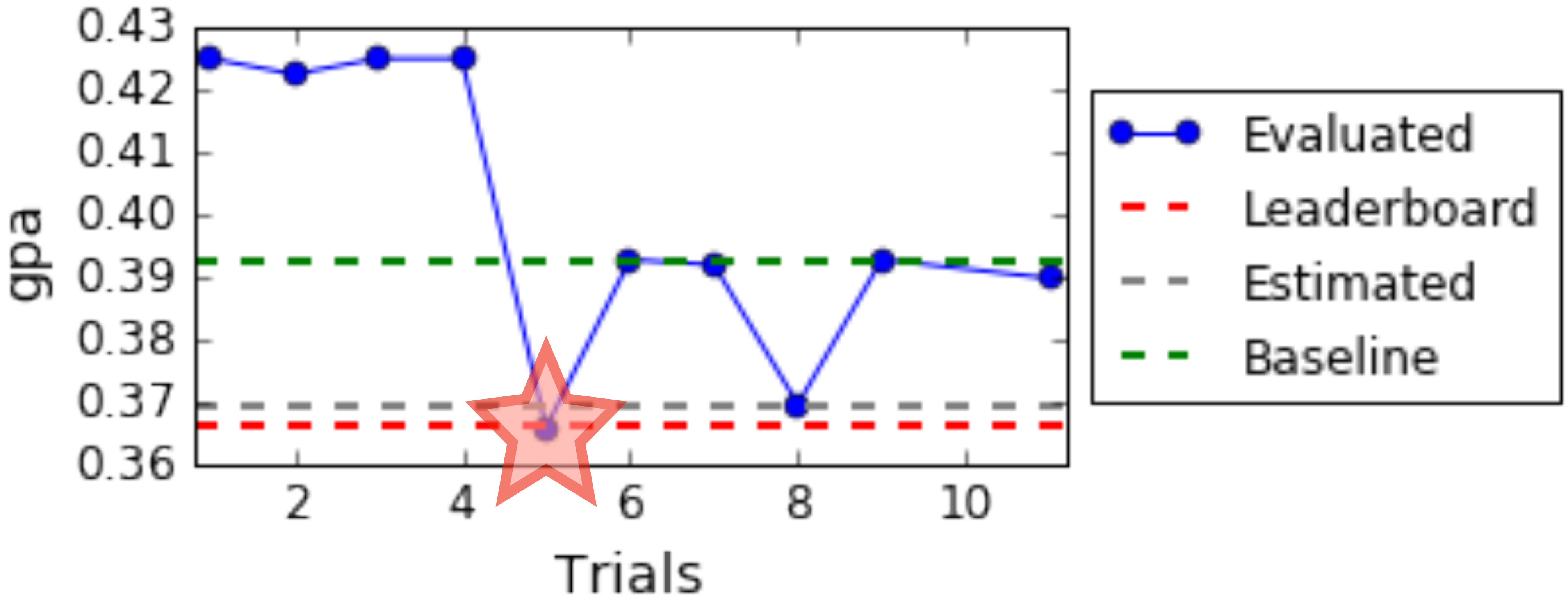
Important features for GPA prediction

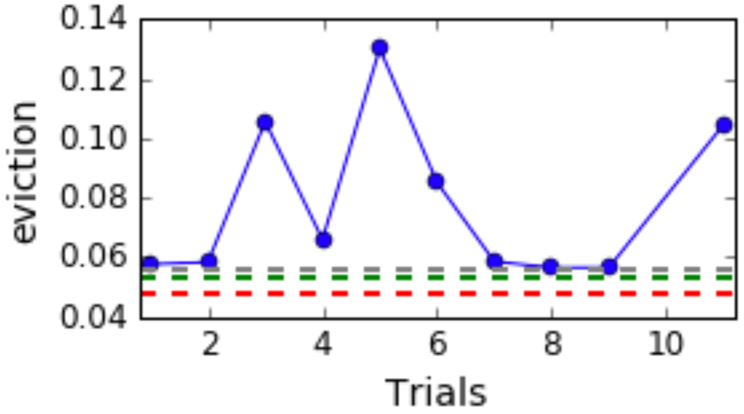
Feature Name	Description
hv5_wj10ss	Woodcock Johnson Test 10 standard score
cm5hhinc	Constructed - Mother’s Household income
hv5_ppvtss	PPVT standard score
hv5_ppvtraw	PPVT raw score
cf5hhincb	Constructed - Household income mother report for married/cohab if available
t5c13c	c13C. Child’s mathematical skills

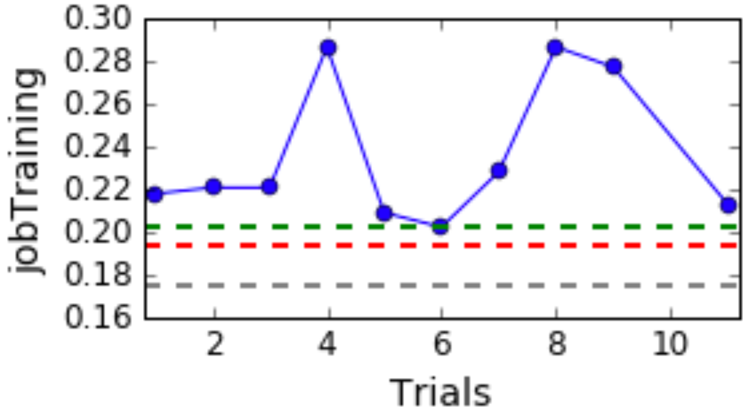


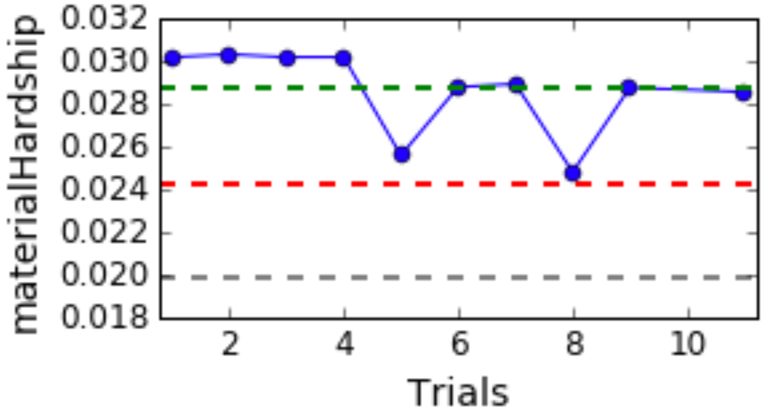
There are 5 other variables more important to predict **GPA** than **child’s own mathematical skills**

Apr 9 ← → Apr 26


gpa
0.43
0.42
0.41
0.40
0.39
0.38
0.37
0.36
2
4
6
8
10
Trials
Evaluated
Leaderboard
Estimated
Baseline

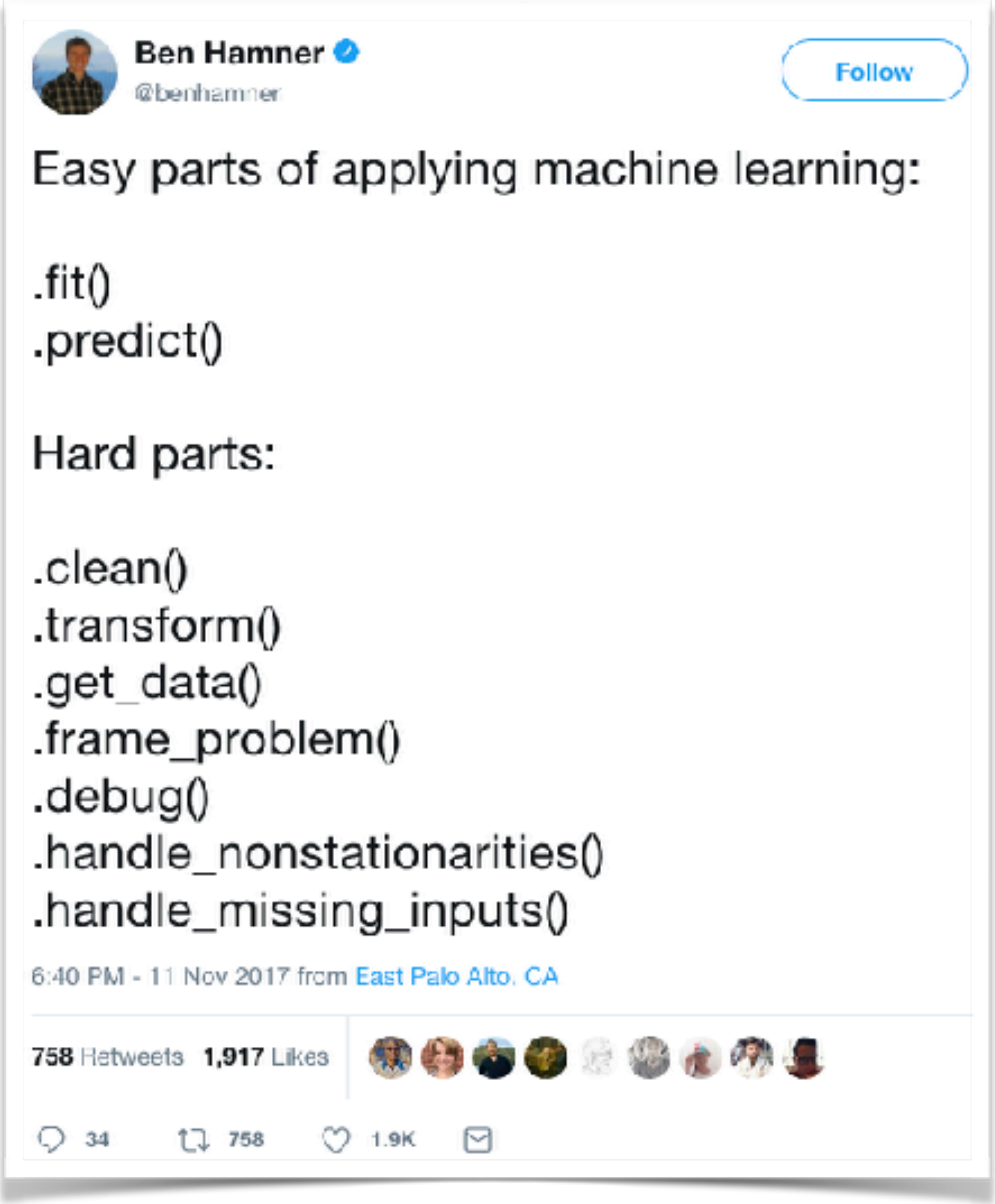

eviction
0.14
0.12
0.10
0.08
0.06
0.04
2
4
6
8
10
Trials


jobTraining
0.30
0.28
0.26
0.24
0.22
0.20
0.18
0.16
2
4
6
8
10
Trials


materialHardship
0.032
0.030
0.028
0.026
0.024
0.022
0.020
0.018
2
4
6
8
10
Trials

What could I do differently?

- Handling missing values
  - Imputation on missing values
  - Computing propensity scores for common responses (filled w/ negative values)
- Better understanding features
  - Clustering beyond main categories (m, f, pc, k, etc.)
  - Topical categorization instead of filtering by keywords
- Taking time into account and analyze waves together
- Building hypothesis and models using published work that use FFC dataset



Ben Hamner
@benhamner

Easy parts of applying machine learning:

```
.fit()
.predict()
```

Hard parts:

```
.clean()
.transform()
.get_data()
.frame_problem()
.debug()
.handle_nonstationarities()
.handle_missing_inputs()
```

6:40 PM - 11 Nov 2017 from East Palo Alto, CA

758 Retweets 1,917 Likes

34 758 1.9K

Discussions and Questions