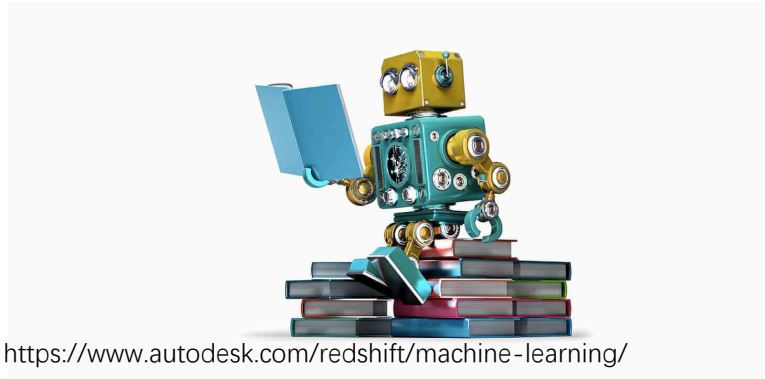
Introduction to Machine Learning for Family Research

NCFR Workshop

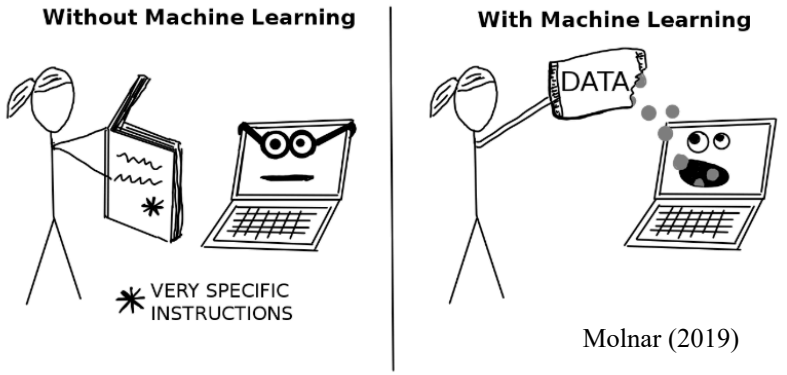
Xiaoran Sun

Departments of Pediatrics & Communication, Stanford University

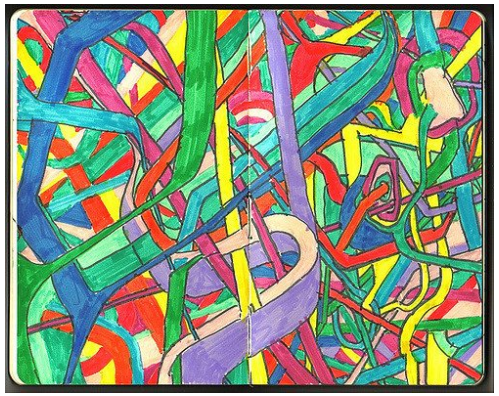
Stanford Data Science Scholars Program



https://www.autodesk.com/redshift/machine-learning/



Molnar (2019)



"Maze" by Danny Glix (2006)

Workshop Outline

- Background/Motivation
- What is machine learning?
- Apply machine learning in family research: An example
- Interactive exercise: Brainstorm about how you can use machine learning in your research
- A demonstration in Jupyter notebook (Python)

Questions are welcome anytime in chat!

Feedback would be very much appreciated!!
https://stanforduniversity.qualtrics.com/jfe/form/SV_9EK6XvsgUzHWzZz

Background/Motivation

- In this digital era, with booming “big” data and access to computational power, the utility of machine learning (ML) methods is gaining attention in social sciences
- Including in developmental science (e.g., Brandmaier et al., 0217; Brick et al., 2017; Puterman et al., 2020; Rosenberg et al., 2018 ; Whelan et al., 2014), such as for
  - processing raw big data (e.g., video and image; Gilmore et al., 2016)
  - predicting future developmental outcomes (e.g., adolescent alcohol misuse, mortality; Whelan et al., 2014; Puterman et al., 2020)
  - selecting important features from large-scale data for optimal study design (Brick et al., 2017)
  - discovering developmental patterns (e.g., growth curve for terminal well-being; Brandmaier et al., 2017)

Background/Motivation

- ML still rarely seen in family studies, but has been emerging recently
  - e.g., The Fragile Families Challenge (Salganik et al., 2020)
  - "Applications of Artificial Intelligence Methodologies to Behavioral and Social Science"(Robila & Robila, 2020, *JCFS*)
  - "Adolescent Family Experiences Predict Young Adult Educational Attainment: A Data-Based Cross-Study Synthesis With Machine Learning" (Sun, Ram, & McHale, 2020)

  - And lot more research to come…

What is Machine Learning?

- A brief definition (by MIT Technology Review):

  ML algorithms

  - find and apply **patterns** in data;
  - use **statistics** to find patterns in massive amounts of **data**

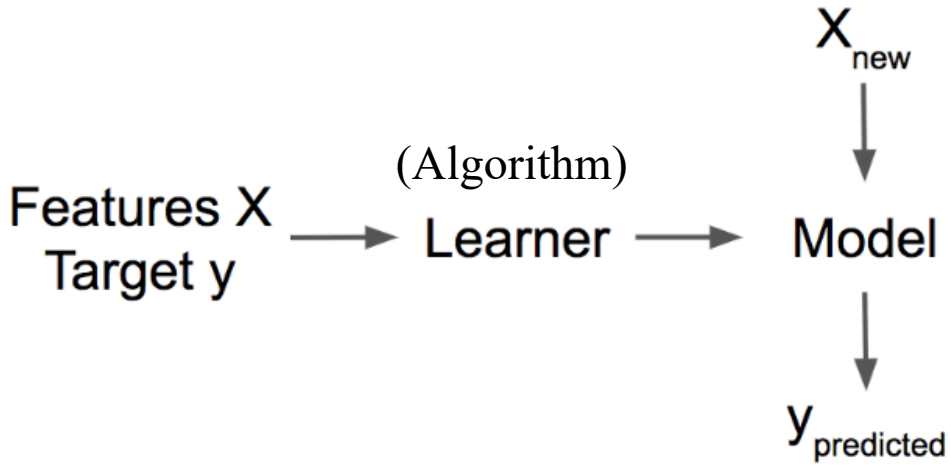
- Different from traditional hypothesis-testing methods in family research, ML is
  - data-driven (v. hypothesis-driven), and
  - exploratory (v. confirmatory)

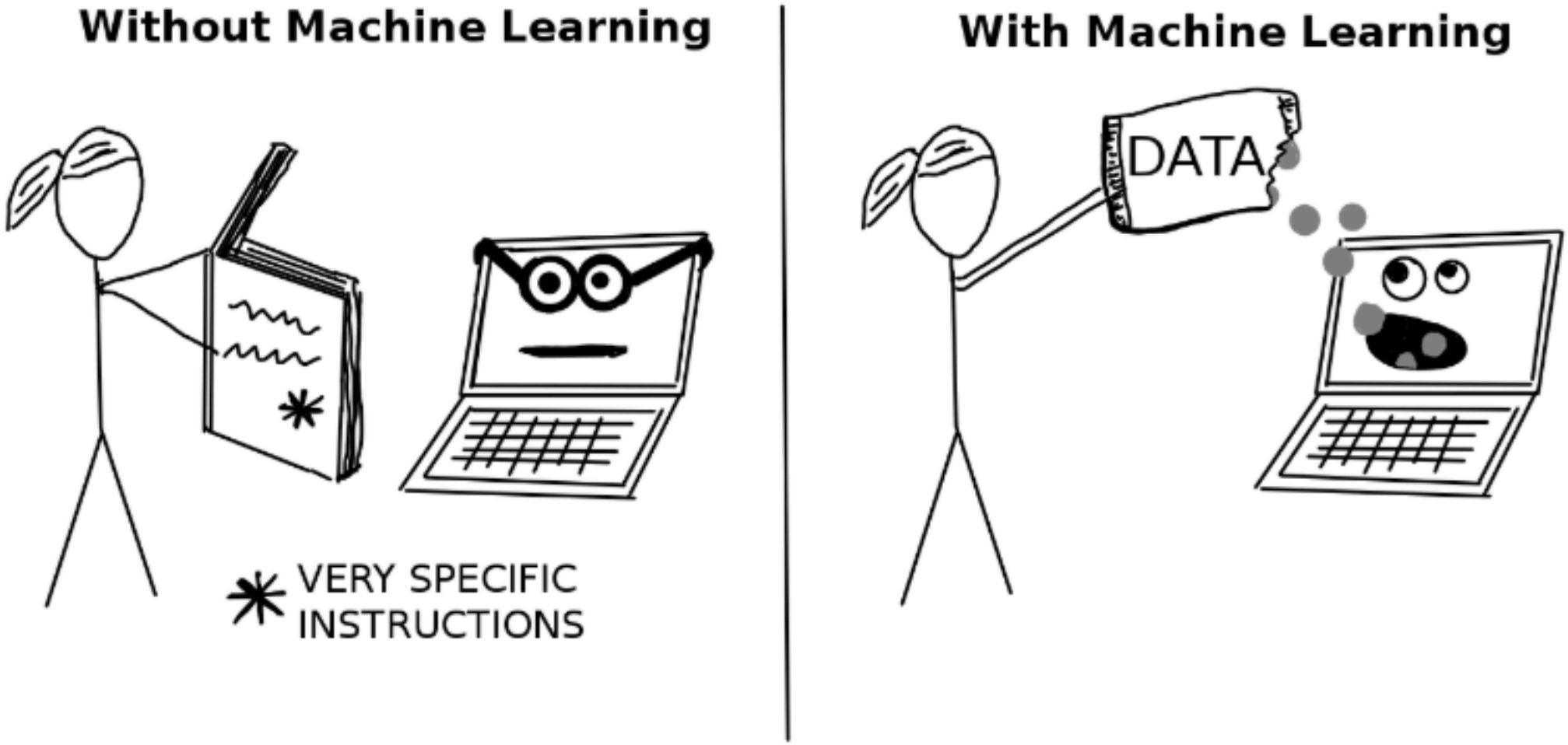
What is Machine Learning?

- Basic types of ML:
  - **supervised learning**
  - unsupervised learning (e.g., cluster analysis)
  - reinforcement learning (e.g., basis of Google AlphaGo)

Supervised Learning Approach

- Protocol: Given data, predictors and an outcome variable, build predictive models that maximize prediction performance (e.g., accuracy, minimal error).
- Data-driven, exploratory; typically no pre-determined associations
  - Allow a large number of predictors
  - (Some algorithms) Allow nonlinearities and interactions

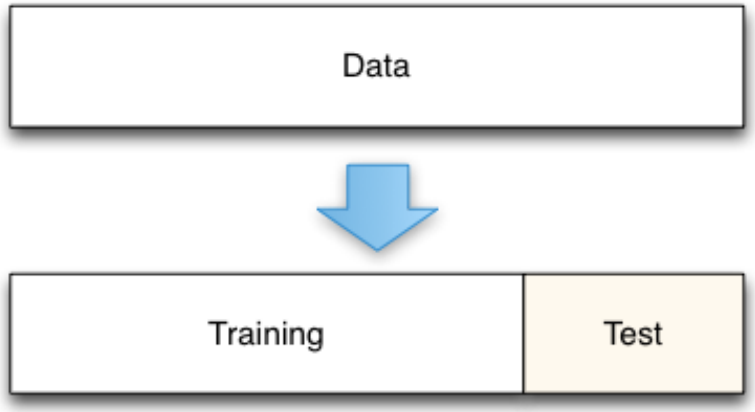




Molnar, Christoph (2019).

Supervised Learning Approach

- Protocol: Given data, predictors and an outcome variable, build predictive models that maximize prediction performance (e.g., accuracy, minimal error).
- Data-driven, exploratory; typically no pre-determined associations
  - Allow a large number of predictors
  - (Some algorithms) Allow nonlinearities and interactions
- "Hold-out" data, or cross-validation: test a model's prediction performance with "unknown" cases
- Regularization: prevents over-fitting



Supervised Learning Approach: Algorithms

- The predicted outcome can be binary/categorical, or continuous variable
- Algorithms can serve as classification, or regression
- Some commonly used algorithms:
  - Regularized logistic regression; Ridge/Lasso regression
  - Linear discriminant analysis
  - Decision trees/Classification and regression trees
  - Support vector machine
  - Ensemble methods, such as random forests
  - …

Supervised Learning Approach: Algorithms

- Regularization: For example, L2 cost function for regularized logistic regression

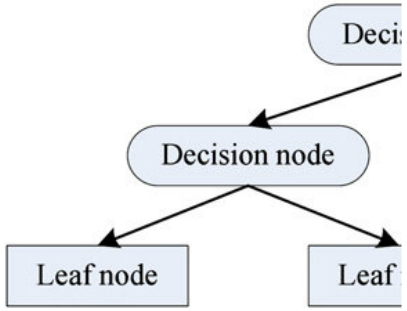
$$\widehat{\boldsymbol{\beta}} = \min_{\beta} \; -LL(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) + \lambda \, R(\boldsymbol{\beta}).$$

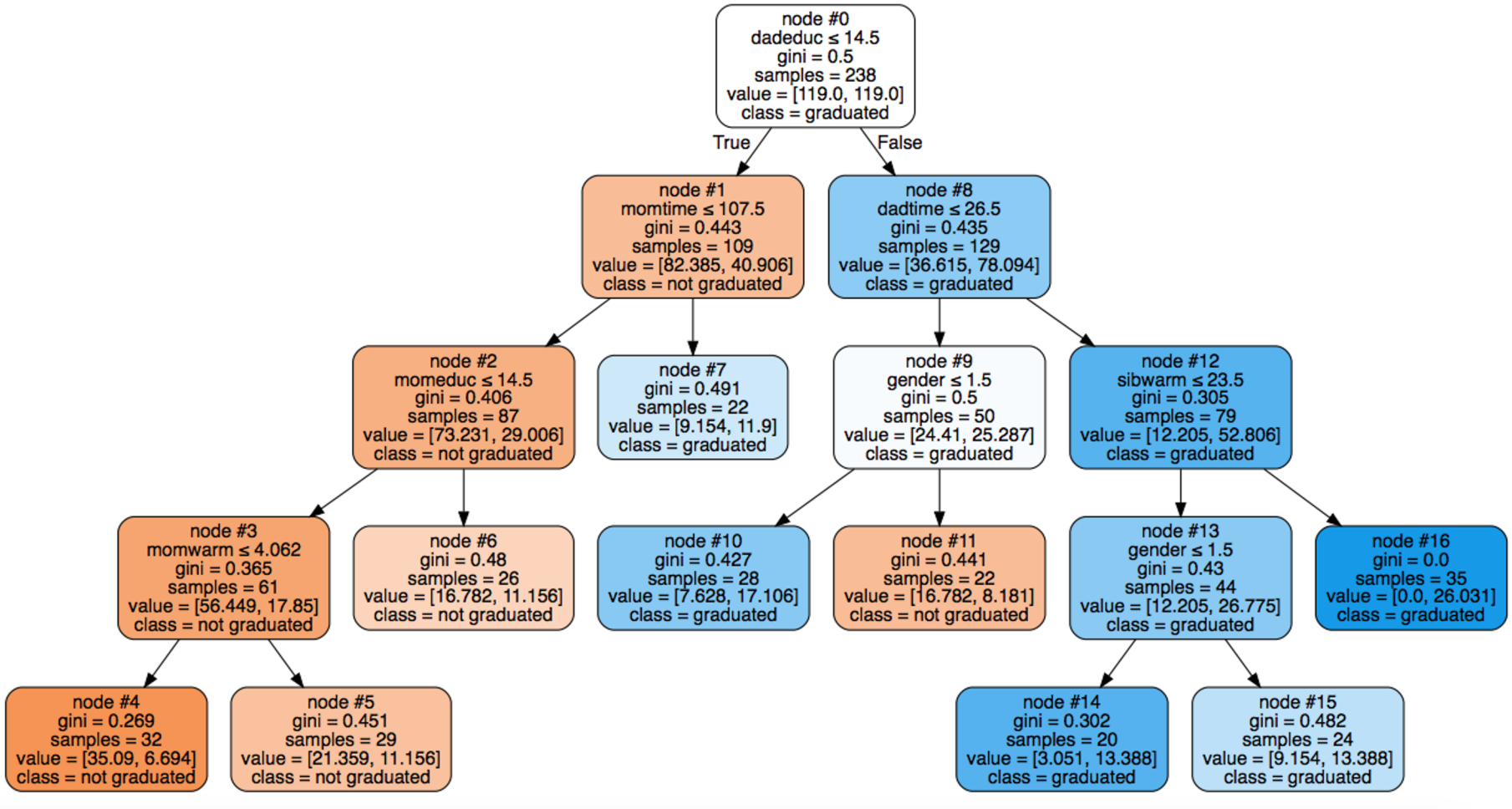
$$R(\boldsymbol{\beta}) = \frac{1}{2} \| \boldsymbol{\beta} \|_2^2 = \frac{1}{2} \sum_{i=0}^{n} \beta_i^2,$$

- especially important to exploratory analysis

Supervised Learning Approach: Algorithms

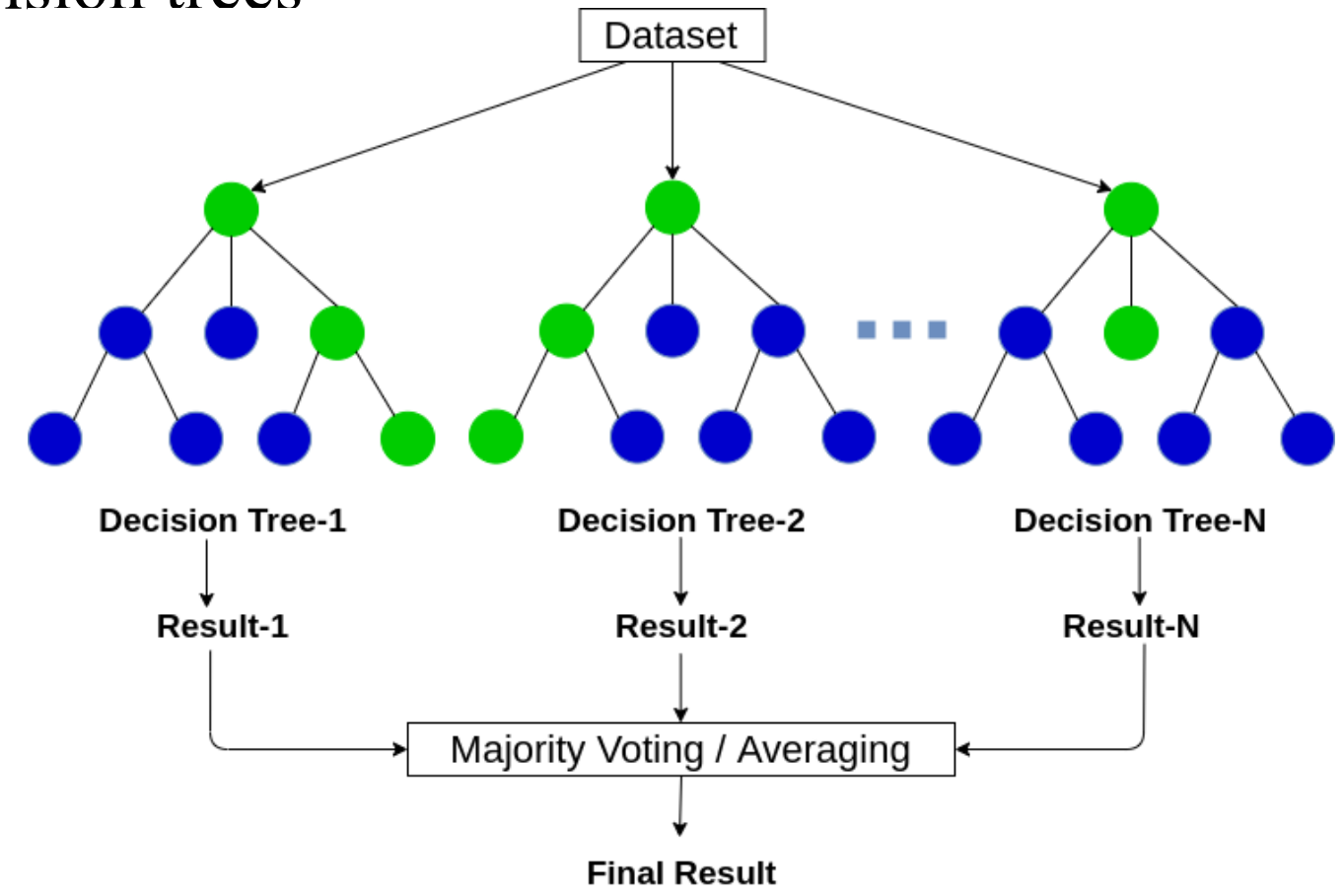
- Decision tree





Supervised Learning Approach: Algorithms

- Random forests: Ensemble of decision trees
- “the wisdom of the crowd”



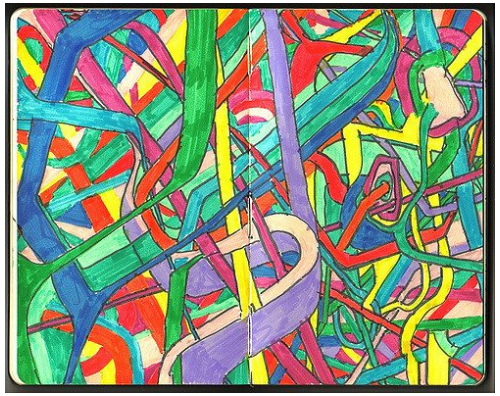
Supervised Learning Approach

- Beyond predicting, many additional useful functions, e.g.
  - **Feature importance & Feature selection:** Select the most important predictors of the outcome
  - **Partial dependence plots (PDP):** Show the relation between one or more predictors and the predicted outcome, holding other predictors constant

Supervised Learning Approach: Analyses

- Very handy to run nowadays, with easy-to-implement packages and modules
- Examples widely used:
  - Python ‘scikit-learn’
  - R ‘caret’

ML for Family Research



"Maze" by Danny Glix (2006)

- Family systems:
  - are complex
  - include numerous factors to examine, such as
    - multiple family relationships: mother-child, father-child, inter-parental, sibling, intergenerational, …
    - multiple aspects of relationships: conflict, warmth, autonomy granting, time, …
    - multiple categories of resources and contexts: education, income, public assistance, employment status, job prestige, heath-related behaviors, sociocultural context, schools and peers, …
  - there could also be interactions among these factors to be explored
- We are trying really hard to be hypothesis-driven, but theory does not necessarily grant us specific hypotheses targeted at specific variables.

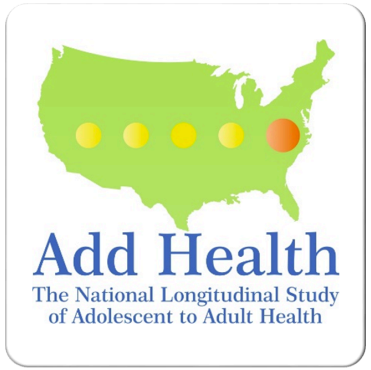
ML for Family Research

- ML can be used for:
  - Taking together a comprehensive set of family predictors, to predict an important family or developmental outcome:
    - And evaluate prediction performance of the set of predictors
  - Selecting the most important family factors among these predictors
  - Exploring interaction patterns

ML for Family Research: An Example

- Sun, X., Ram, N., & McHale, S. M. (2020). Adolescent family experiences predict young adult educational attainment: A data-based cross-study synthesis with machine learning. *Journal of Child and Family Studies, 29,* 2770-2785.

Using high-dimensional, large-scale data to synthesize family experiences as predictors of achievement



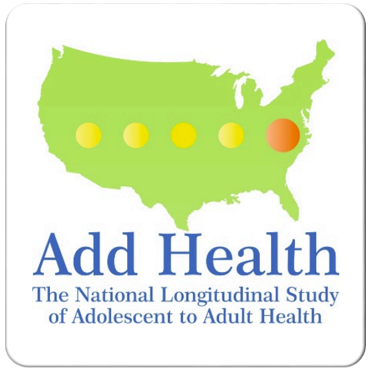
Nationally representative, large sample (~N = 6,000 for public dataset; Wave 1-5):
Measures a variety domains of development (including family experiences and academic/educational achievement) across adolescence & young adulthood

53 Family Experience Variables (Wave I; 1994-1995; Grade 7-12) Predict:
Educational Attainment (College Enrollment & Graduation; Wave IV; 2008-09)

Category	Constructs
Family and adolescent demographic characteristics	Resident mother/father presence
	Resident with two biological parents
	Household size
	Number of siblings
	Parent age
	Birth order
	Biological sex
	Adolescent age
Family socioeconomic characteristics	Mother/Father education levels
	Mother/Father occupational prestige
	Family income (log)
	Parent receiving public assistance
	Family receive welfare (3-item)
	Parent economic hardship
Family and parent-adolescent relationship characteristics	Family social support (4-item)
	Shared dinner with parents
	Intergenerational closure
	Mother/father-adolescent relationship quality (5-item)
	Mother/father-adolescent shared activities (10-item)
	Parental control (7-item)
	Mother/father supervision (3-item)

Category	Constructs
Parental involvement with education	Mother/Father involvement with schoolwork (3-item)
	Parent in parent-teacher association
	Parent in school fund-raising
	Parent met teachers
	Mother/Father educational expectations (2-item)
Family sociocultural characteristics	Mother/Father nativity
	Parent religiosity (2-item)
	English as home language
	Adolescent nativity
	Adolescent race/ethnicity
Family health resources and behaviors	Parent health
	Parent smoking
	Mother/Father alcoholic
	Mother/Father obese
	Mother/Father disabled
	Smoker(s) in household
	Illegal drugs in household
	Family access to medical care

Using high-dimensional, large-scale data to synthesize family experiences as predictors of achievement



RQ1: **How accurately** can a comprehensive set of adolescent family experiences predict young adult educational attainment?

RQ2: Which family experience factors are **key predictors** of young adult educational attainment?

RQ3: What complex patterns, including **nonlinearities and interactions** involving this set of family factors, merit further examination?

Method: *Machine Learning Approach*

Algorithms: Regularized logistic regression; Random forests

[RQ1] Training and testing: stratified 5-fold nested cross-validation

[RQ2] Feature selection: Feature importance & Recursive feature elimination

[RQ3] Nonlinearity and interactions: Partial dependence plots

Model Training & Testing

- Model training—model tuning
  - Regularized logistic regression: λ
  - Random forest classifier: combination of number of trees and maximum depth
  - Finding parameters for highest accuracy
- Model testing: Prediction performance
  - Accuracy
  - AUC (Area under ROC curve)
- Nested cross-validation
  - Outer loop– model training & testing (5-fold)
  - Inner loop– model tuning (5-fold)

An illustration of nested 5-fold cross-validation implemented in this study (Figure crafting in reference to Raschka, 2013-2019).

Feature Selection

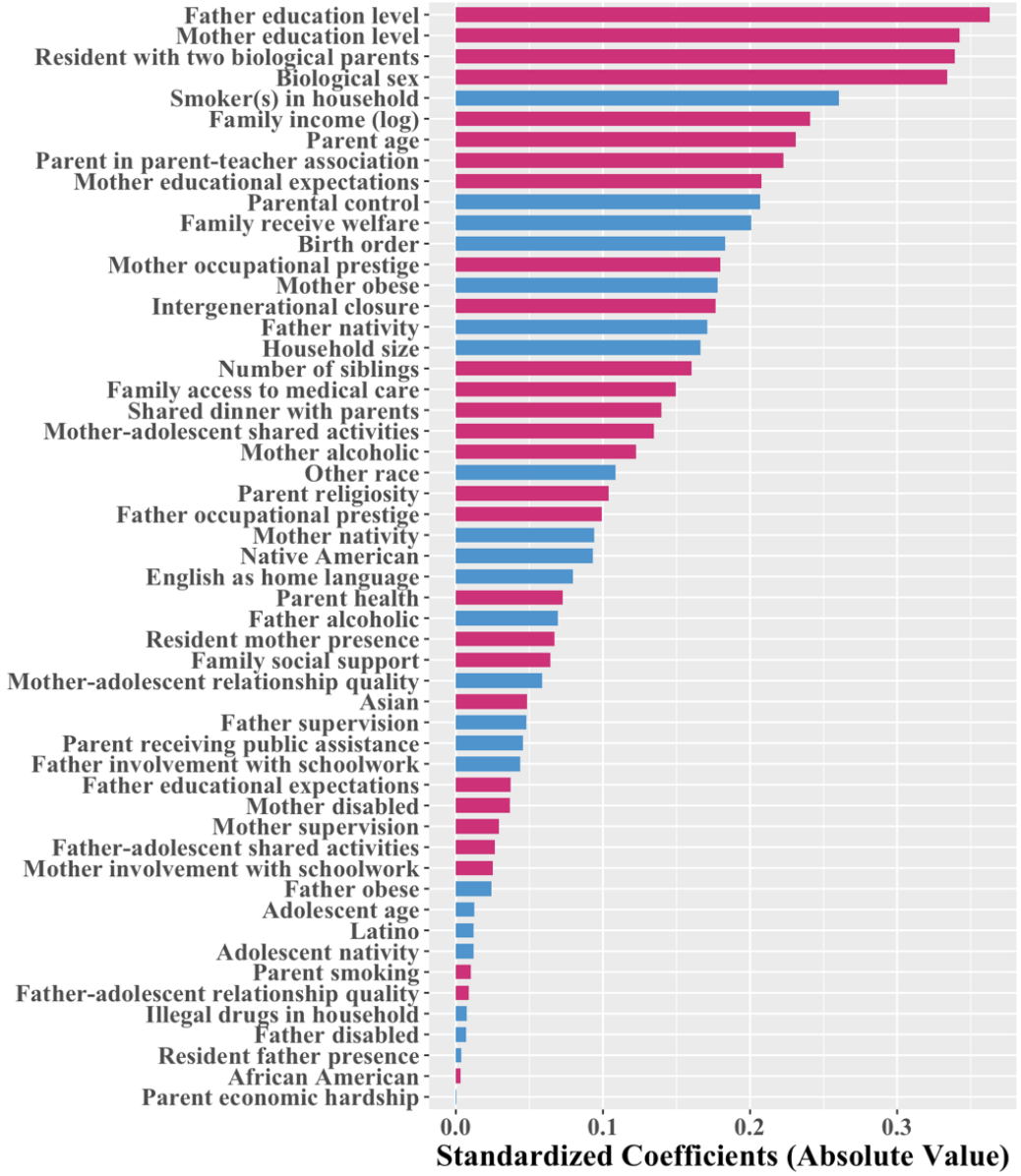
- Recursive feature elimination:
  - Recursively eliminate the least important feature, check reduction in prediction accuracy
  - Using the outer 5-fold loop for cross-validation
- Follow up with partial dependence plots
  - For random forest model interpretation
  - 2D: main effects
  - 3D: two-way interactions

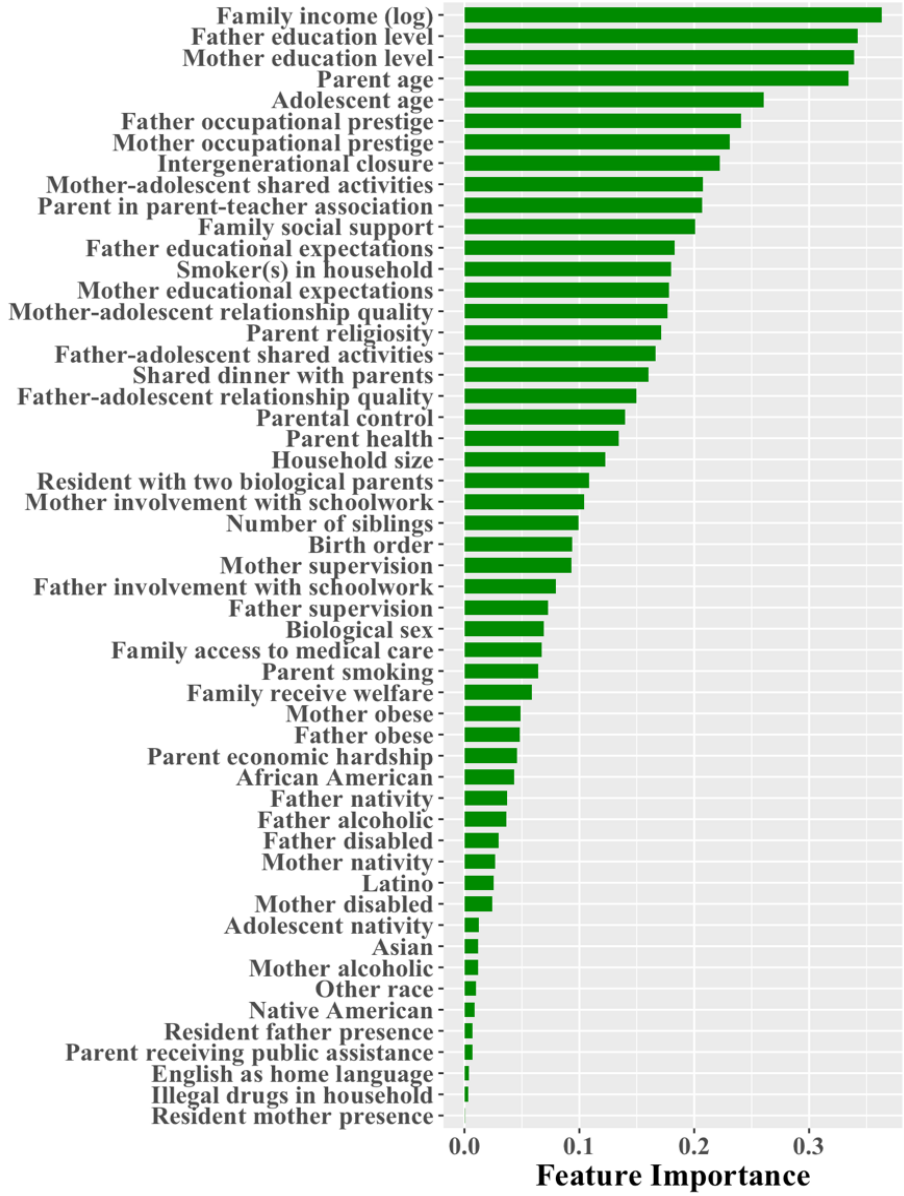
RQ1: **How accurately** can a comprehensive set of adolescent family experiences predict young adult educational attainment?

	Regularized logistic regression				Random forests			
	λ	Accuracy	AUC		Maximum depth	No. of trees	Accuracy	AUC
Predicting college graduation								
CV1	50	79.58%	0.8421		17	700	77.67%	0.8287
CV2	500	78.51%	0.8271		15	700	78.19%	0.8272
CV3	10	81.46%	0.8627		18	800	82.21%	0.8561
CV4	10	77.98%	0.8386		11	200	78.71%	0.8255
CV5	10	77.99%	0.8193		11	100	78.56%	0.8070
Mean	**-**	**79.10%**	**0.8379**		**-**	**-**	**79.07%**	**0.8289**

Note. CV = Cross-validation (outer loop).
Chance level of accuracy = 55.62% for college graduation.
Chance level of AUC = 0.50.

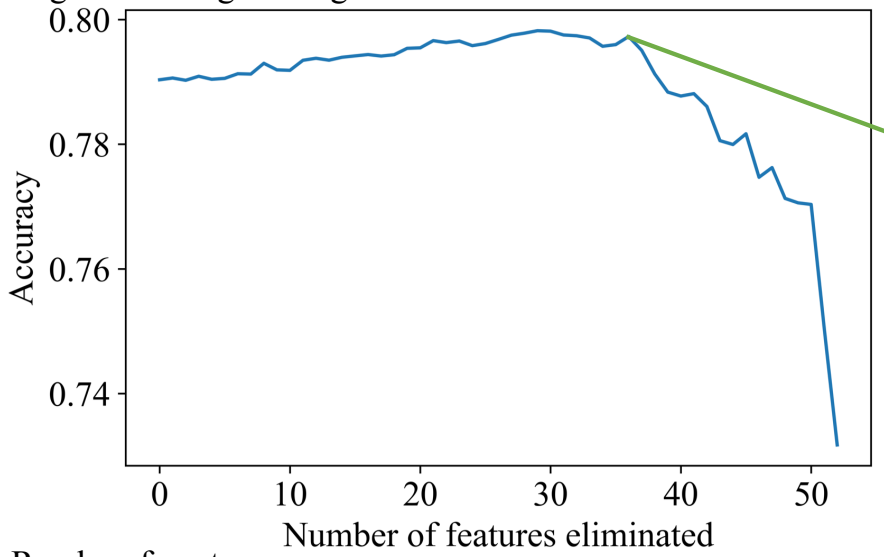
RQ2: Which family experience factors are **key predictors** of young adult educational attainment?





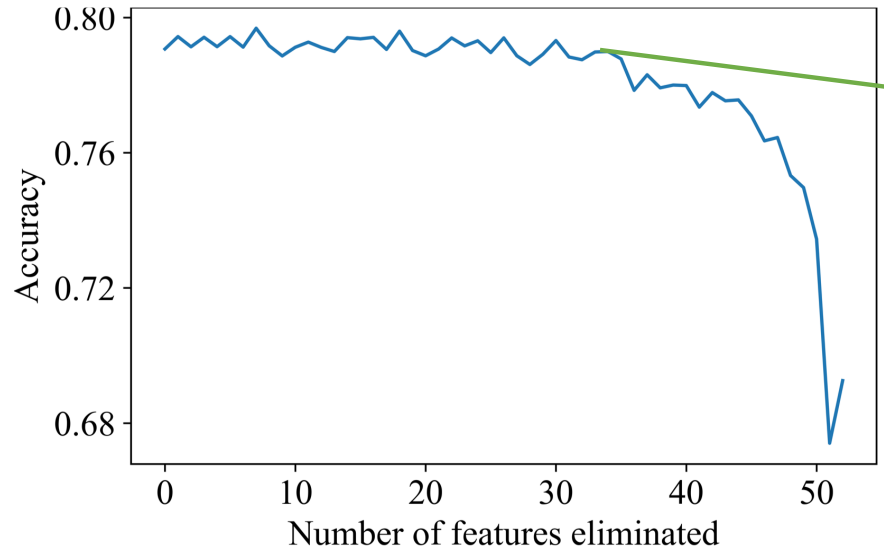
RQ2: Which family experience factors are **key predictors** of young adult educational attainment?

Regularized logistic regression



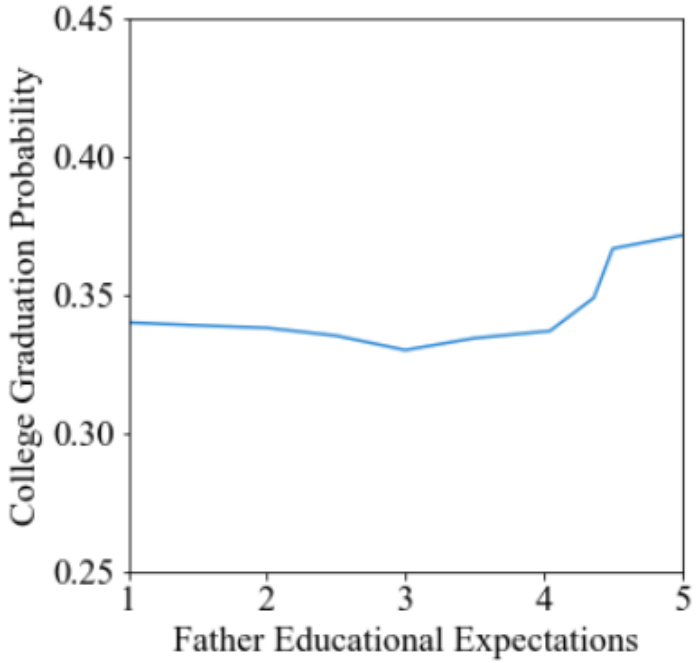
Accuracy = 79.73%
17 features:
Mother education, **Family income**, **Father education**,
Smoker(s) in household, Resident with two biological parents,
Biological sex, **Parent in parent-teacher association**,
Mother educational expectations, Family receive welfare,
Father nativity, **Intergenerational closure**, Mother obese, **Parent age**,
Birth order, **Mother occupational prestige**, **Parental control**,
Shared dinner with parents

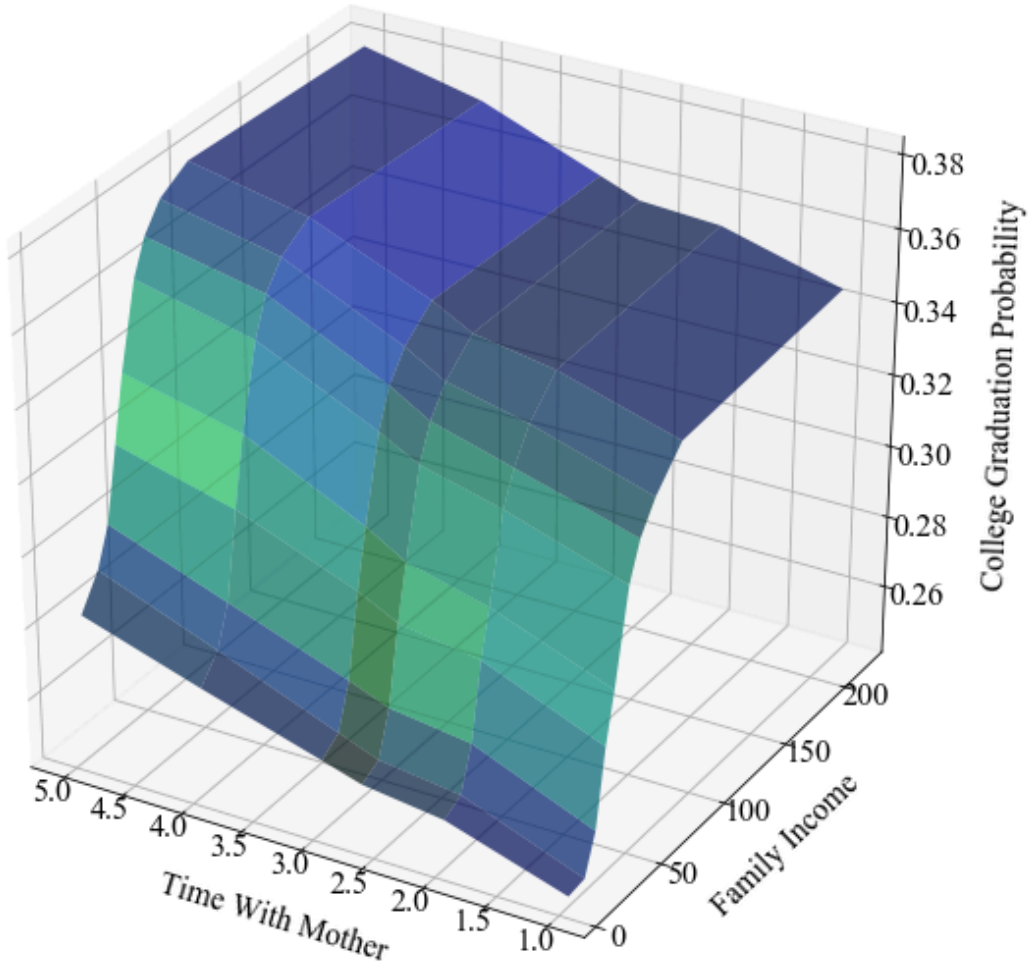
Random forests



Accuracy = 78.99%
19 features:
Family income, Adolescent age, **Father education**, **Parent age**,
Mother education, Family social support,
Mother-adolescent shared activities, **Intergenerational closure**,
Father occupational prestige, Father-adolescent relationship quality,
Mother-adolescent relationship quality, **Mother occupational prestige**,
Father-adolescent shared activities, **Mother educational expectations**,
Parent religiosity, **Shared dinner with parents**,
Father educational expectations**, Parental control**,
Parent in parent-teacher association

RQ3: What complex patterns, including **nonlinearities and interactions** involving this set of family factors, merit further examination?





https://github.com/xiaoransun/ML.family.edu.AddHealth

Study Summary

- Utility of machine learning approach in research on family and youth development outcomes
  - Reveal predictive power with a (relatively) large set of family factors
  - Reveal important factors for particular development outcomes
  - Reveal nonlinear & interaction effects that deserve further examination
- Family experiences in adolescence predict young adult educational achievement
  - 53 adolescent adolescent family features predicted college graduation with ~79% accuracy
  - Key family experiences identified in predicting educational outcomes
  - Nonlinear (e.g., activation point) and interaction effects of family experiences revealed

Machine Learning Approach

Traditional Statistical Approaches
(Especially Hypothesis-Testing)

ML for Family Research: Brainstorming

- Take 5~10 minutes, to think about these questions:
  - The research questions you are interested in answering
    - Don’t hesitate to go big picture!
  - The dataset that you think can answer these questions
    - okay if you are not entirely sure
  - How ML may apply to answer your questions

Feel free to type in the chat box, or unmute yourself and speak up!

Feedback would be very much appreciated!!
https://stanforduniversity.qualtrics.com/jfe/form/SV_9EK6XvsgUzHWzZz

ML for Family Research: Brainstorming

- Principles:
  - No thoughts or questions are “dumb”, or “too bold”
  - Please be constructive and non-judgmental in providing thoughts to one another

ML for Family Research: Demo

- Use ‘scikit-learn’ in Python, with the Jupyter notebook
- Data: Add Health family experience predictors (X) + college graduation (y)