Machine Learning for Public Policy
Week 3: Prediction within a Population

Brandon Stewart[1]

Princeton

February 10–12, 2025

[1]Huge thanks to Ian Lundberg and Matt Salganik for materials around the Fragile Families Challenge and Matt Salganik for slides from the Summer Institute in Computational Social Science.

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - discovery
  - measurement
  - regression + non-linear basis functions

Where We've Been and Where We're Going...

- Last Week
  - discovery
  - measurement
  - regression + non-linear basis functions
- This Week
  - preview of three types of prediction
  - prediction within population
  - logistic regression, GAMs, decision trees

Where We've Been and Where We're Going...

- Last Week
  - discovery
  - measurement
  - regression + non-linear basis functions
- This Week
  - preview of three types of prediction
  - prediction within population
  - logistic regression, GAMs, decision trees
- Next Week
  - prediction in a new population
  - random forest and other ensembles

Where We've Been and Where We're Going...

- Last Week
  - discovery
  - measurement
  - regression + non-linear basis functions
- This Week
  - preview of three types of prediction
  - prediction within population
  - logistic regression, GAMs, decision trees
- Next Week
  - prediction in a new population
  - random forest and other ensembles
- Long Run
  - target tasks $\rightarrow$ data sources $\rightarrow$ guiding values

Admin

Admin

- Proposal Signup and Procedure

Admin

- Proposal Signup and Procedure
- Reconsidering Content + Methods Material

Admin

- Proposal Signup and Procedure
- Reconsidering Content + Methods Material
- Minor Assignment Reminder (and shift on when released)

1. Preview Three Types of Prediction
2. Prediction for Enhanced Measurement
3. Fragile Families Challenge
4. Logistic Regression
5. Generalized Additive Models
6. Decision Trees

1 Preview Three Types of Prediction

2 Prediction for Enhanced Measurement

3 Fragile Families Challenge

4 Logistic Regression

5 Generalized Additive Models

6 Decision Trees

Prediction

Prediction

- Prediction as a word is overloaded—it had many distinct meanings.

Prediction

- Prediction as a word is overloaded—it had many distinct meanings.
- We have been learning different techniques for supervised learning—taking training examples (pairs of features and labeled outputs) and estimating a prediction function $\hat{f}()$ such that

$$Y \approx \hat{f}(X)$$

Prediction

- Prediction as a word is overloaded—it had many distinct meanings.
- We have been learning different techniques for supervised learning—taking training examples (pairs of features and labeled outputs) and estimating a prediction function $\hat{f}()$ such that

$$Y \approx \hat{f}(X)$$

- Can we feed in any set of training examples we want?

Prediction

- Prediction as a word is overloaded—it had many distinct meanings.
- We have been learning different techniques for supervised learning—taking training examples (pairs of features and labeled outputs) and estimating a prediction function $\hat{f}()$ such that

$$Y \approx \hat{f}(X)$$

- Can we feed in any set of training examples we want?
- Ideally we want a random sample from the population where we want to do prediction.

Three Prediction Tasks of Increasing Difficulty

Three Prediction Tasks of Increasing Difficulty

1) Prediction in the Same Population

Three Prediction Tasks of Increasing Difficulty

1) Prediction in the Same Population
2) Prediction in a Different (Real) Population

Three Prediction Tasks of Increasing Difficulty

1) Prediction in the Same Population
2) Prediction in a Different (Real) Population
3) Prediction in a Counterfactual Population

A Concrete Example

A Concrete Example

- Have you ever thought about how worldwide mortality is actually collected?

A Concrete Example

- Have you ever thought about how worldwide mortality is actually collected?
- Estimates suggest that a majority of the 60 million people who die annually do so without medical certification of the cause of death (particularly in low and middle income countries).

A Concrete Example

- Have you ever thought about how worldwide mortality is actually collected?
- Estimates suggest that a majority of the 60 million people who die annually do so without medical certification of the cause of death (particularly in low and middle income countries).
- One mechanism for assessing causes of death is through verbal autopsies.

Verbal Autopsy

Verbal Autopsy

- Core Idea: ask family member about hundreds of symptoms of recently deceased. Use symptoms to guess the cause of death.

Verbal Autopsy

- Core Idea: ask family member about hundreds of symptoms of recently deceased. Use symptoms to guess the cause of death.
- This is a prediction problem: we have some symptoms X and we want to predict the cause of death Y.

Verbal Autopsy

- Core Idea: ask family member about hundreds of symptoms of recently deceased. Use symptoms to guess the cause of death.
- This is a prediction problem: we have some symptoms X and we want to predict the cause of death Y.
- How could we obtain labeled data? Plenty of cases in major hospitals where doctors certify the cause of death.

Verbal Autopsy

- Core Idea: ask family member about hundreds of symptoms of recently deceased. Use symptoms to guess the cause of death.
- This is a prediction problem: we have some symptoms X and we want to predict the cause of death Y.
- How could we obtain labeled data? Plenty of cases in major hospitals where doctors certify the cause of death.
- Using the standard technique we make a train/test split and we evaluate our predictive ability. What does this tell us?


OPD

- Where you can easily collect labeled data is different from where you need to make the predictions.

- Where you can easily collect labeled data is different from where you need to make the predictions.
- Challenges:
  - different common causes of death

- Where you can easily collect labeled data is different from where you need to make the predictions.
- Challenges:
  - different common causes of death
  - different presentations of disease

- Where you can easily collect labeled data is different from where you need to make the predictions.
- Challenges:
  - different common causes of death
  - different presentations of disease
  - different types of data collection.

- Where you can easily collect labeled data is different from where you need to make the predictions.
- Challenges:
  - different common causes of death
  - different presentations of disease
  - different types of data collection.
- Major variations across sites and datasets.

Example Prediction Tasks of Increasing Difficulty

1) Prediction in the Same Population

2) Prediction in a Different (Real) Population

3) Prediction in a Counterfactual Population

Example Prediction Tasks of Increasing Difficulty

1) Prediction in the Same Population
   predict cause of death in major hospitals
2) Prediction in a Different (Real) Population

3) Prediction in a Counterfactual Population

Example Prediction Tasks of Increasing Difficulty

1) Prediction in the Same Population
   predict cause of death in major hospitals
2) Prediction in a Different (Real) Population
   predict cause of death in rural hospitals
3) Prediction in a Counterfactual Population

Example Prediction Tasks of Increasing Difficulty

1) Prediction in the Same Population
   predict cause of death in major hospitals
2) Prediction in a Different (Real) Population
   predict cause of death in rural hospitals
3) Prediction in a Counterfactual Population
   predict cause of death if we intervened

Lighter Example: Kickstarter

Lighter Example: Kickstarter

Consider the following questions:

- Will the project be funded based on the description?

Lighter Example: Kickstarter

Consider the following questions:

- Will the project be funded based on the description?
- Would the project be funded on a competitor platform or in the future?

Lighter Example: Kickstarter

Consider the following questions:

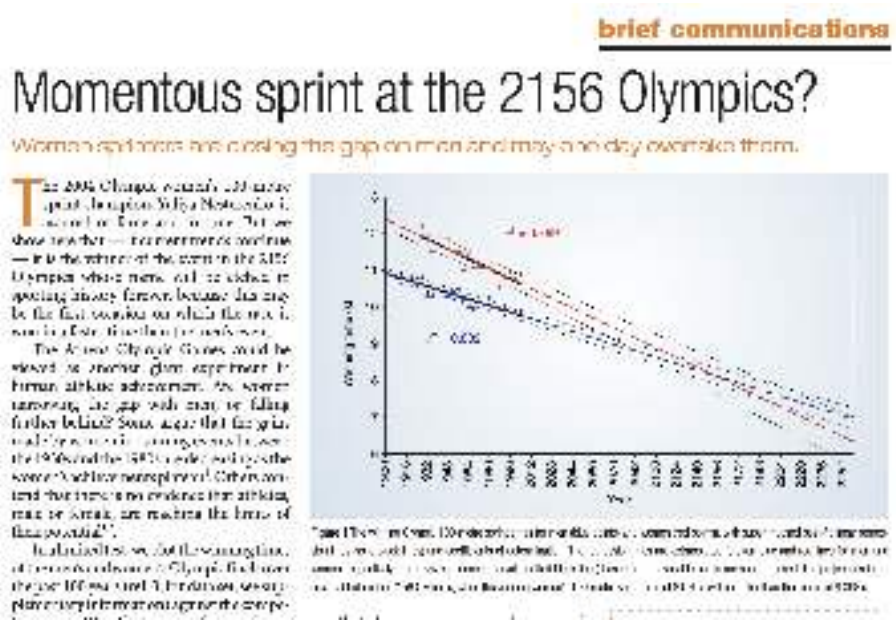
- Will the project be funded based on the description?
- Would the project be funded on a competitor platform or in the future?
- If I changed the description to say "featuring faster *and simpler* gameplay" would it be more likely to fund?

Why is It Harder to Predict in a Different Population?

Why is It Harder to Predict in a Different Population?

- Model Dependence:
  you can't measure $p(Y|X)$ within some regions of X.
  Interpolation is hard, extrapolation is harder.

Example: Tatem, et al. Sprinters Data

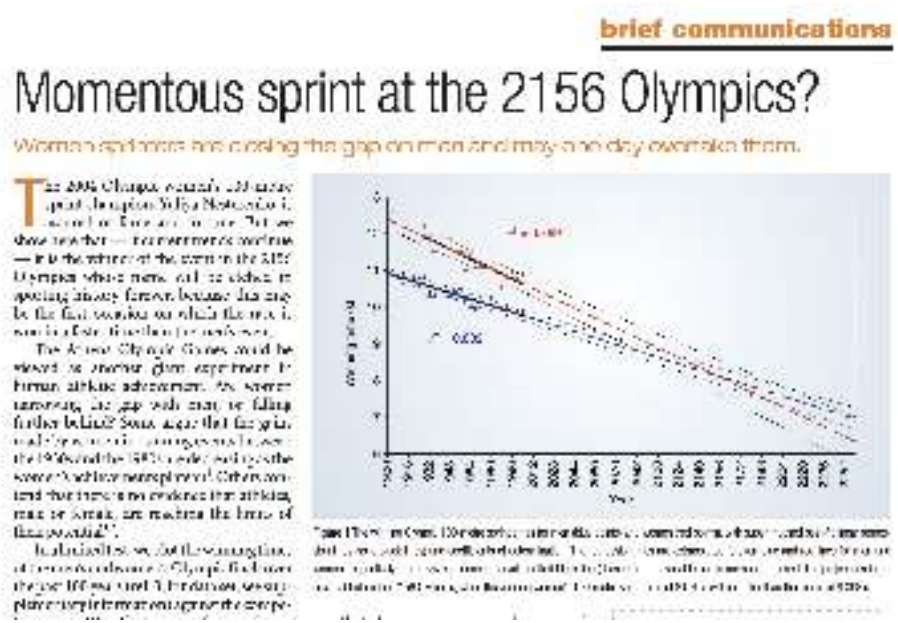


brief communications

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

Example: Tatem, et al. Sprinters Data



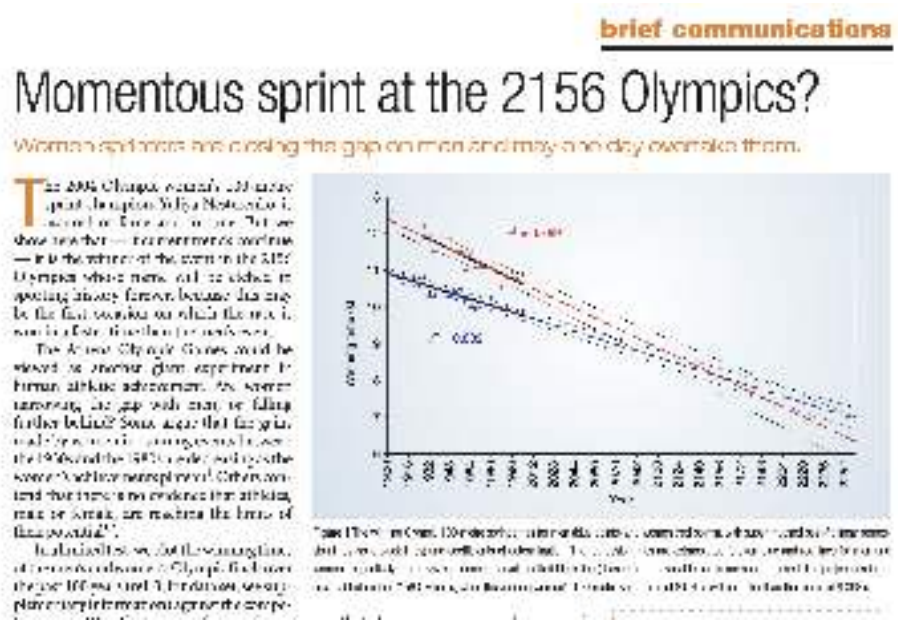
brief communications

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

In a 2004 *Nature* article, Tatem et al. use linear regression to conclude that in the year 2156 the winner of the women's Olympic 100 meter sprint may likely have a faster time than the winner of the men's Olympic 100 meter sprint.

Example: Tatem, et al. Sprinters Data



brief communications

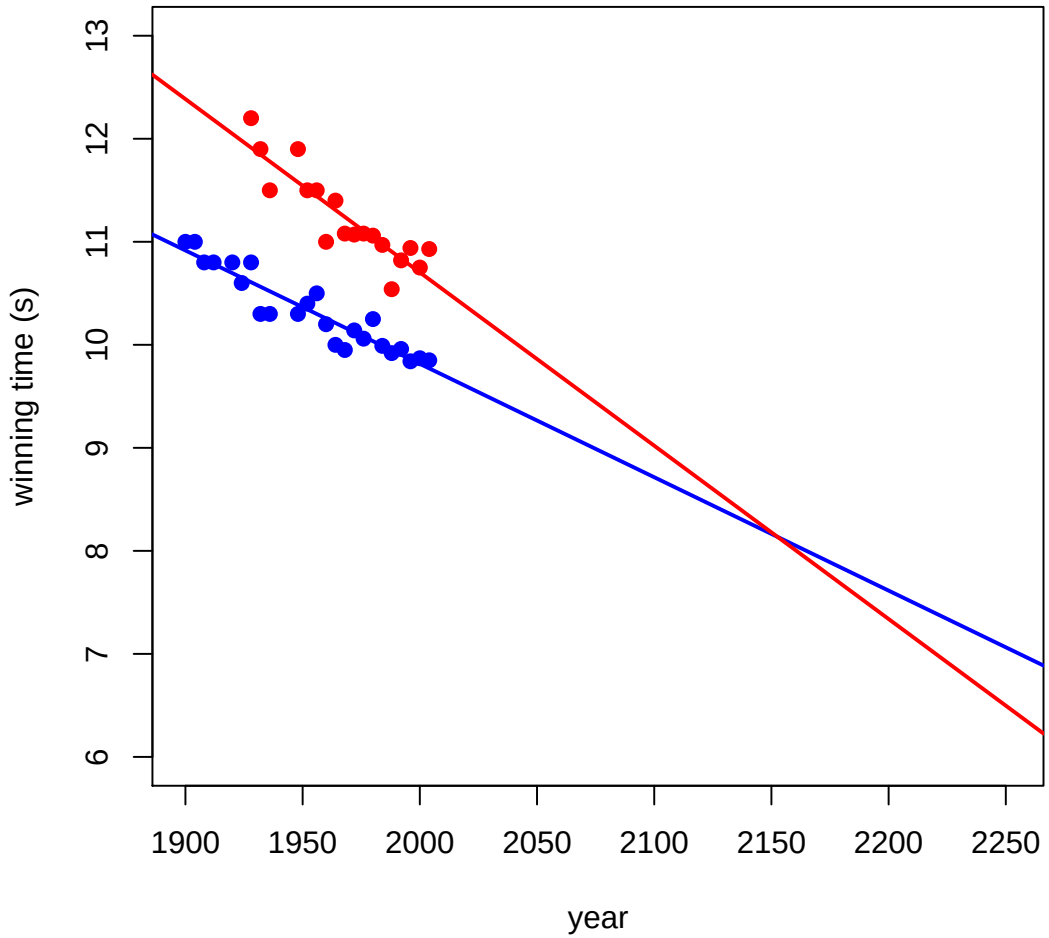
Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

In a 2004 *Nature* article, Tatem et al. use linear regression to conclude that in the year 2156 the winner of the women's Olympic 100 meter sprint may likely have a faster time than the winner of the men's Olympic 100 meter sprint.

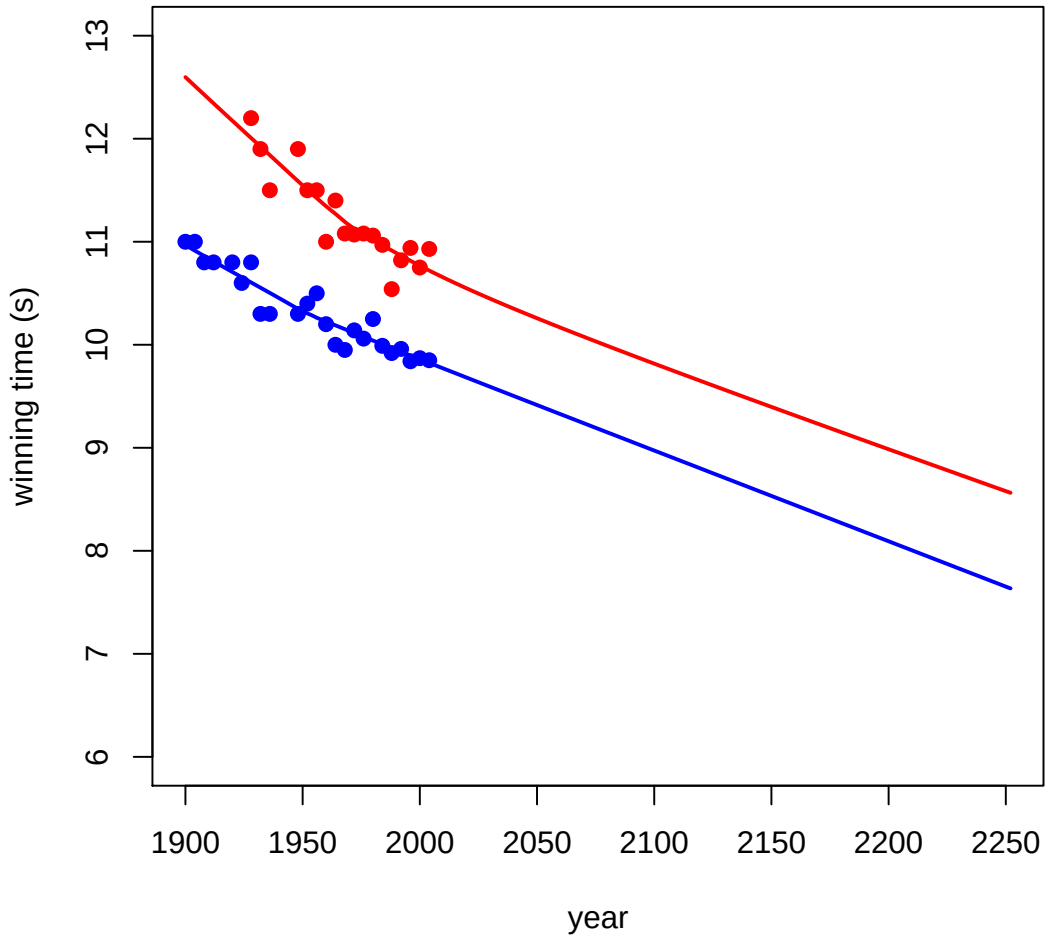
Using data from 1900 to 2004, they fit linear regression models of the winning 100 meter time on year for both men and women. They then use the estimates from these models to extrapolate 152 years into the future.

Tatem et al. Extrapolation

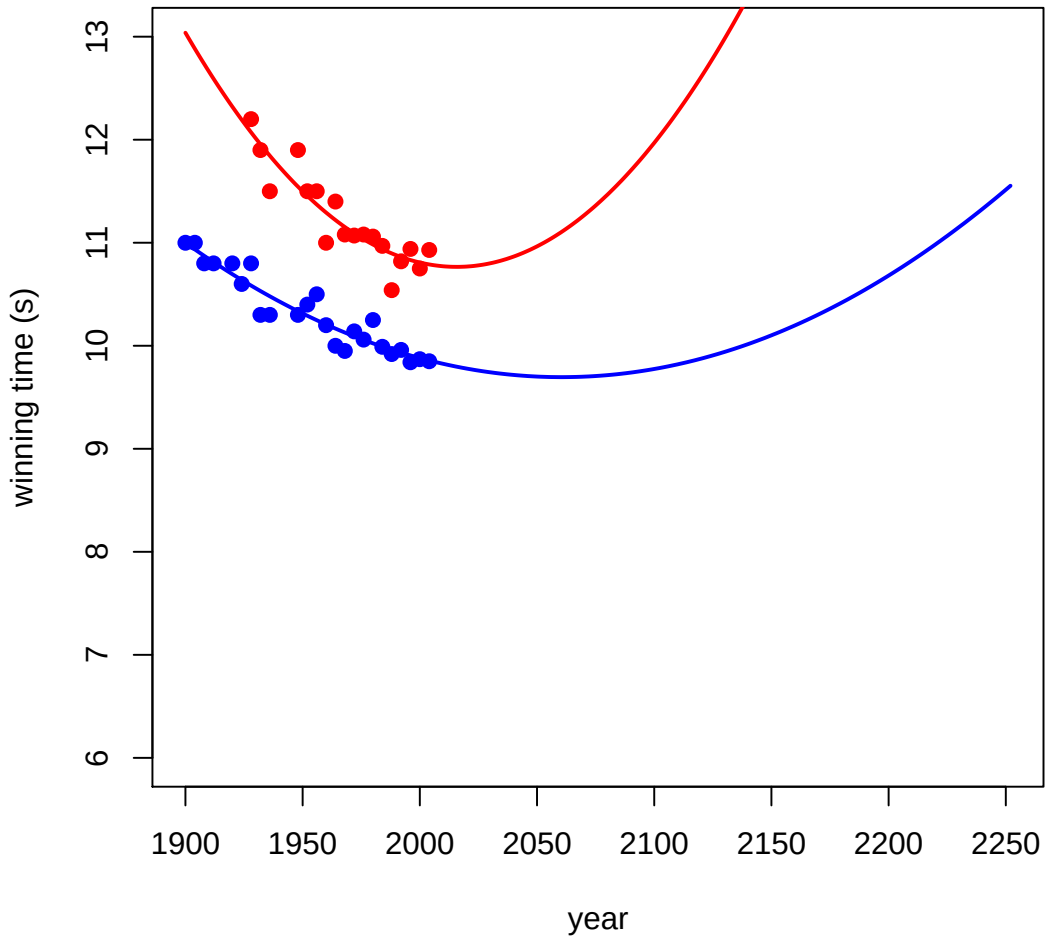


Tatem et al.'s predictions. Men's times are in blue, women's times are in red.

Alternate Models Fit Well, Yield Different Predictions



Alternate Models Fit Well, Yield Different Predictions



Why is It Harder to Predict in a Different Population?

- Model Dependence:
  you can't measure $Y|X$ within some regions of X. Interpolation is hard, extrapolation is harder.

Why is It Harder to Predict in a Different Population?

- Model Dependence:
  you can't measure $Y|X$ within some regions of X. Interpolation is hard, extrapolation is harder.
- Drift:
  the distribution of $Y|X$ is different in a different population.

Why is It Harder to Predict in a Different Population?

- Model Dependence:
  you can't measure $Y|X$ within some regions of X. Interpolation is hard, extrapolation is harder.
- Drift:
  the distribution of $Y|X$ is different in a different population.
- Unknowable:
  in counterfactual settings we can never get to observe the true answer.

Why Predicting People is Not Like Predicting Cats

Why Predicting People is Not Like Predicting Cats

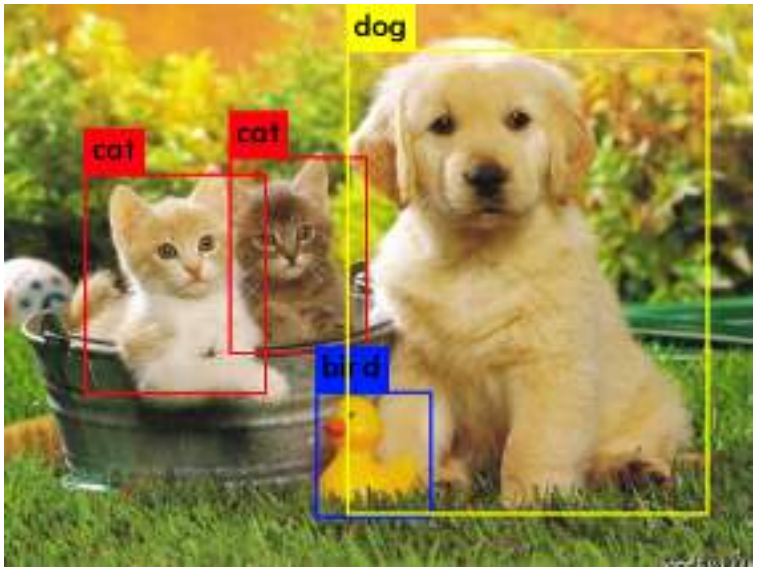


Image Credit: The Internet

Why Predicting People is Not Like Predicting Cats

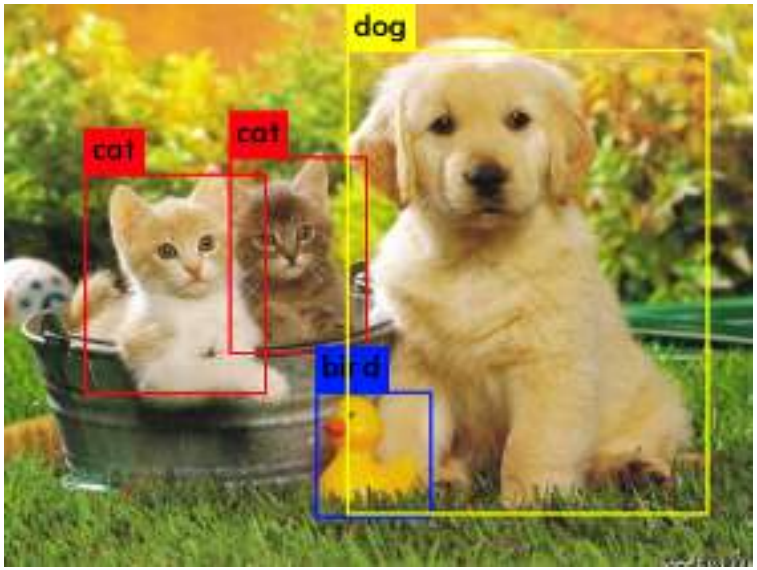


Image Credit: The Internet

1) Populations vs. Individuals:

haystack vs. the hay

Why Predicting People is Not Like Predicting Cats

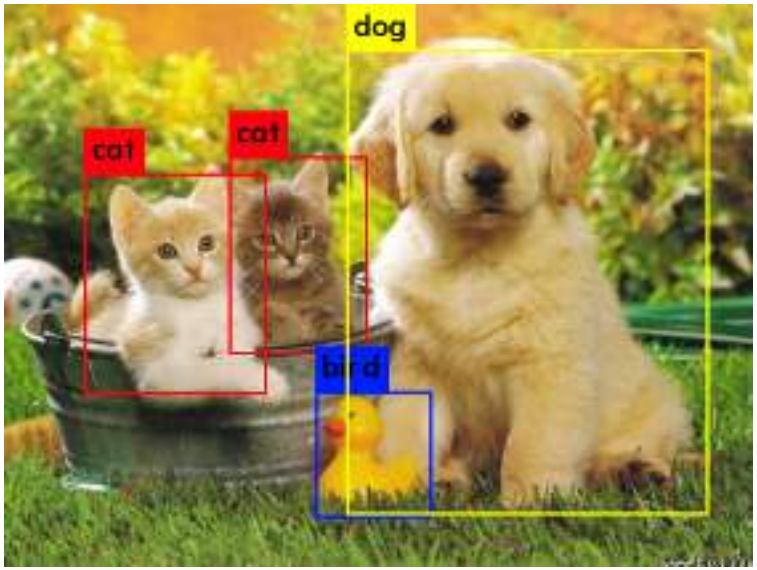


Image Credit: The Internet

1) Populations vs. Individuals:

   haystack vs. the hay

2) Limits to Prediction:

   settling for a conditional distribution

Why Predicting People is Not Like Predicting Cats

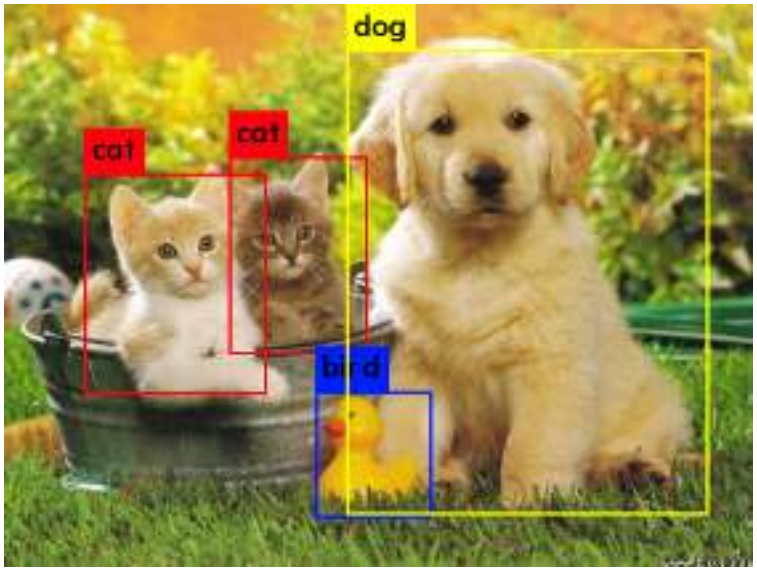


Image Credit: The Internet

1) Populations vs. Individuals:

   haystack vs. the hay

2) Limits to Prediction:

   settling for a conditional distribution

3) Adversarial:

   people try to game the system

1. Preview Three Types of Prediction
2. Prediction for Enhanced Measurement
3. Fragile Families Challenge
4. Logistic Regression
5. Generalized Additive Models
6. Decision Trees

1 Preview Three Types of Prediction

2 Prediction for Enhanced Measurement

3 Fragile Families Challenge

4 Logistic Regression

5 Generalized Additive Models

6 Decision Trees

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1]* Gabriel Cadamuro,[2] Robert On[3]

Blumenstock, Cadamuro, and On (2015)

Image Credit: Salganik 2017 *Bit by Bit*

Image Credit: Salganik 2017 *Bit by Bit*

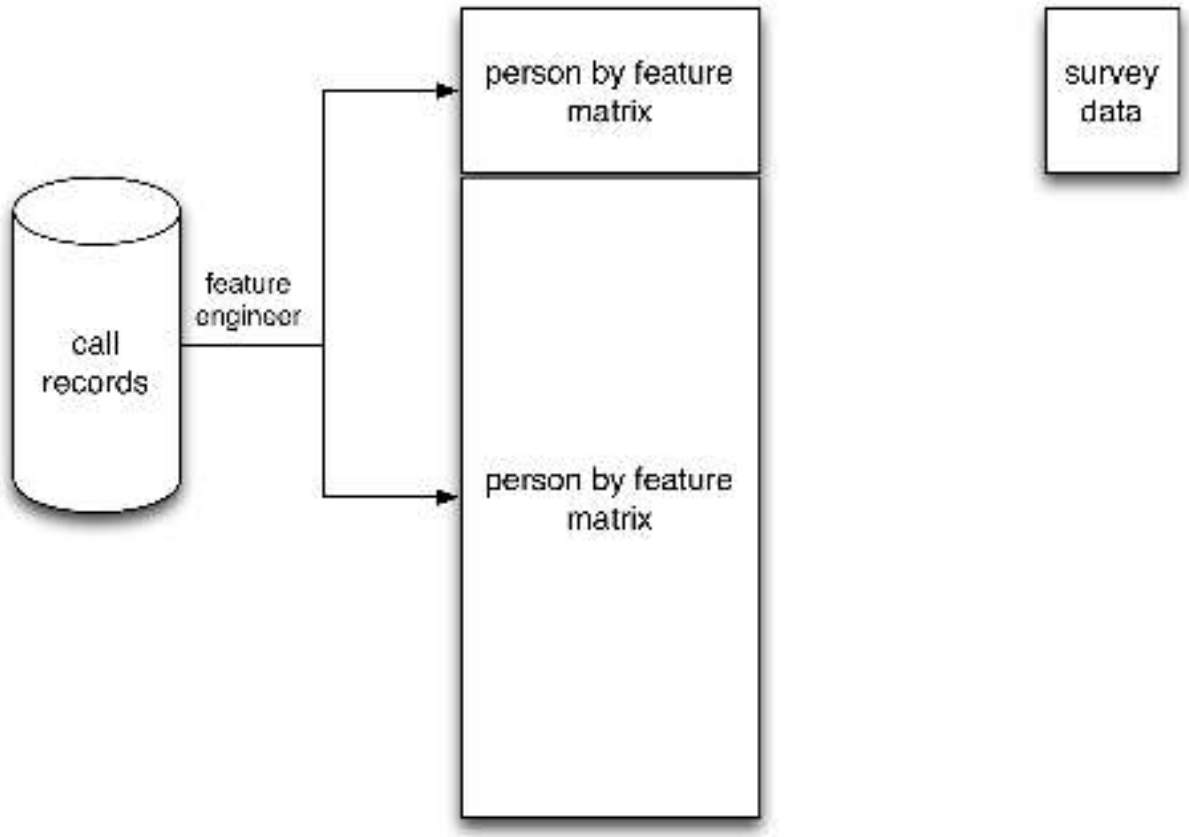


Image Credit: Salganik 2017 *Bit by Bit*

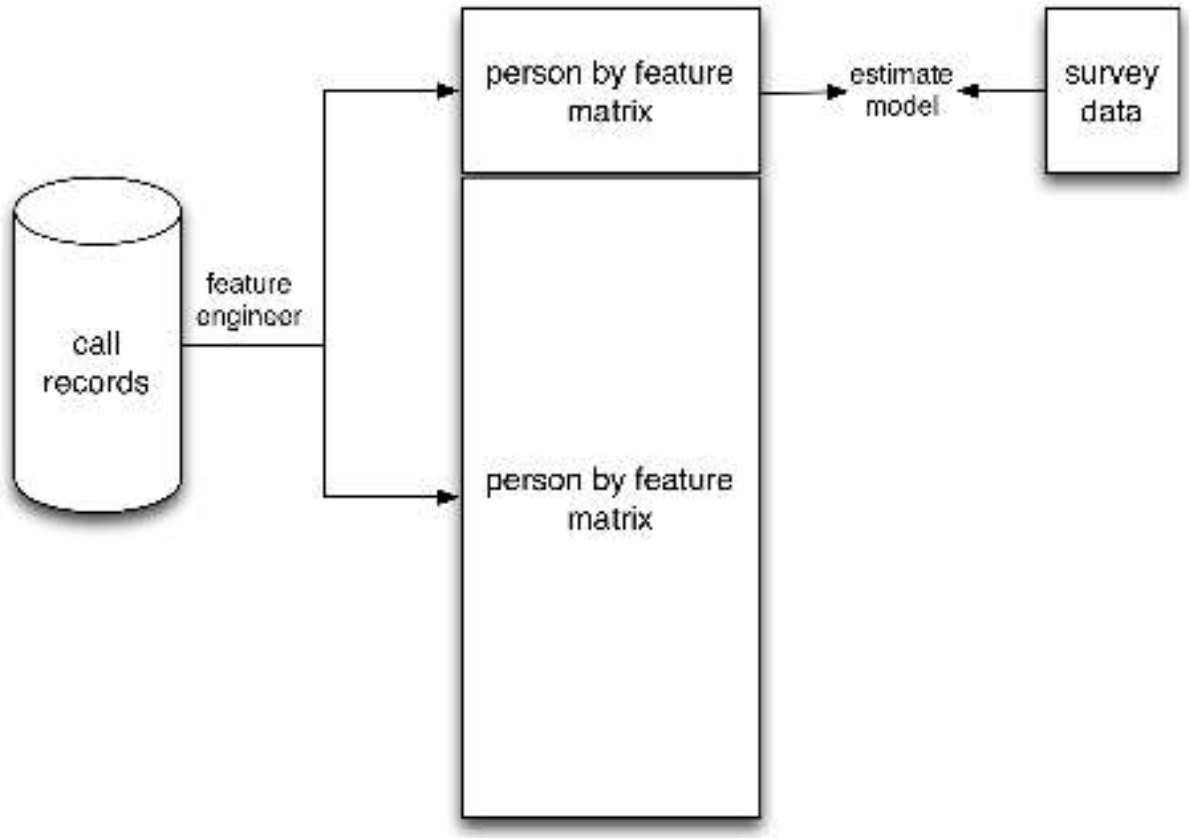


Image Credit: Salganik 2017 *Bit by Bit*

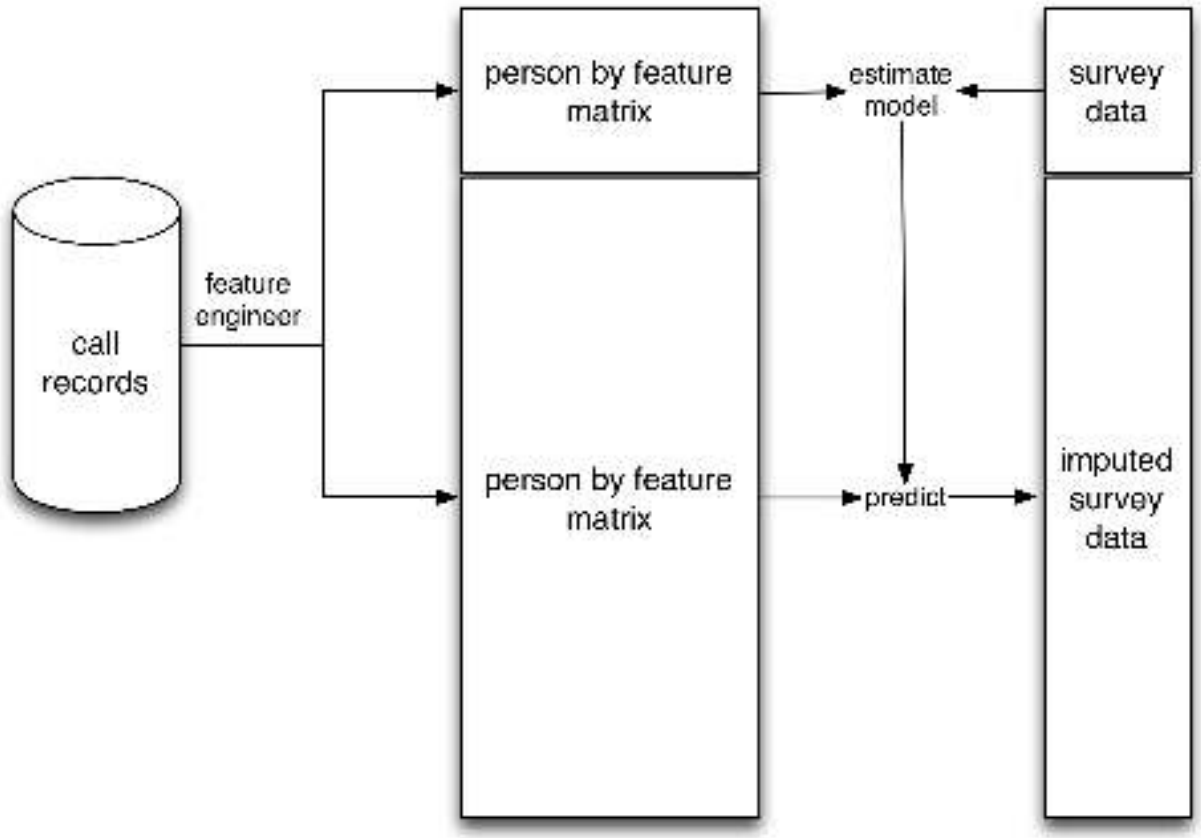


Image Credit: Salganik 2017 *Bit by Bit*

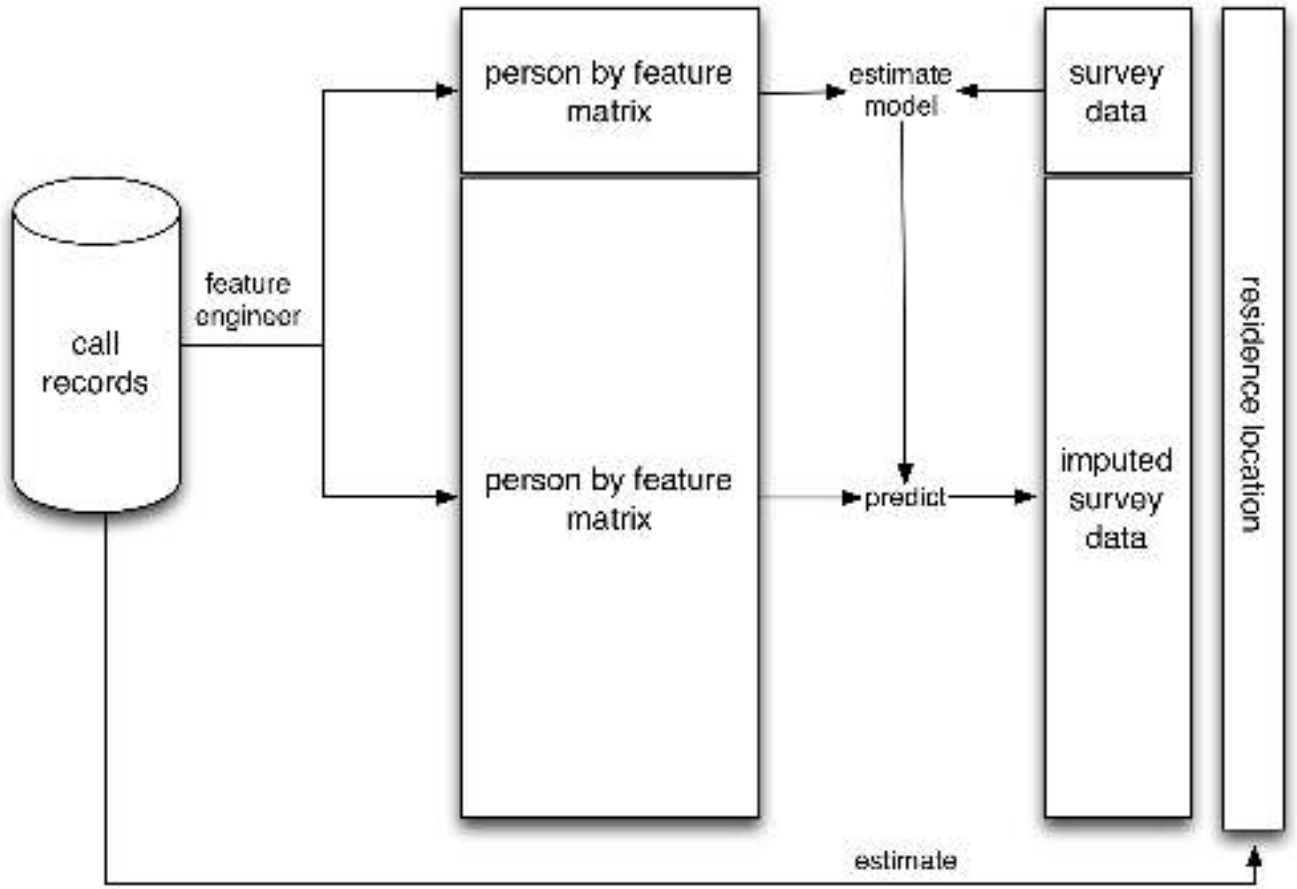
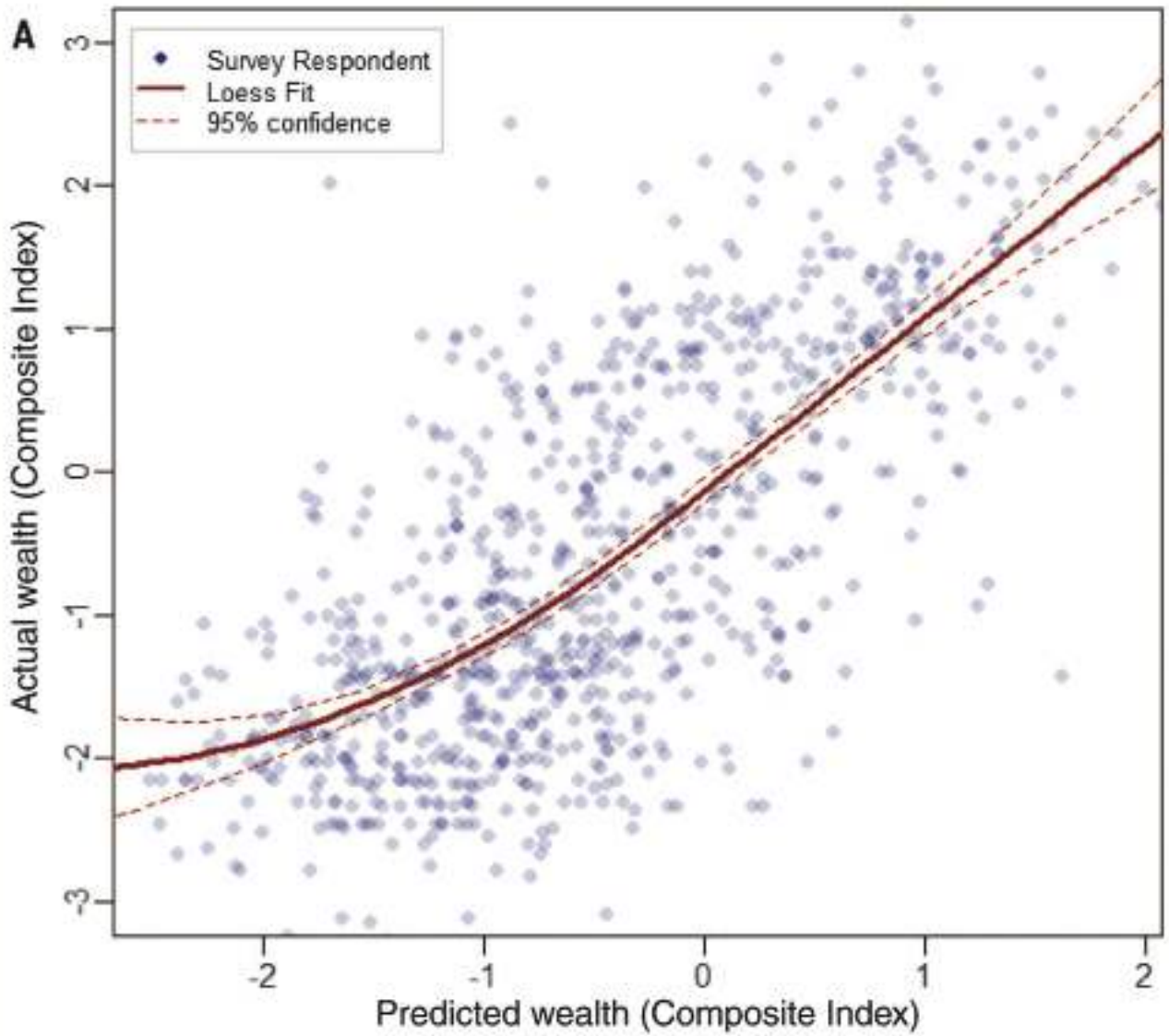
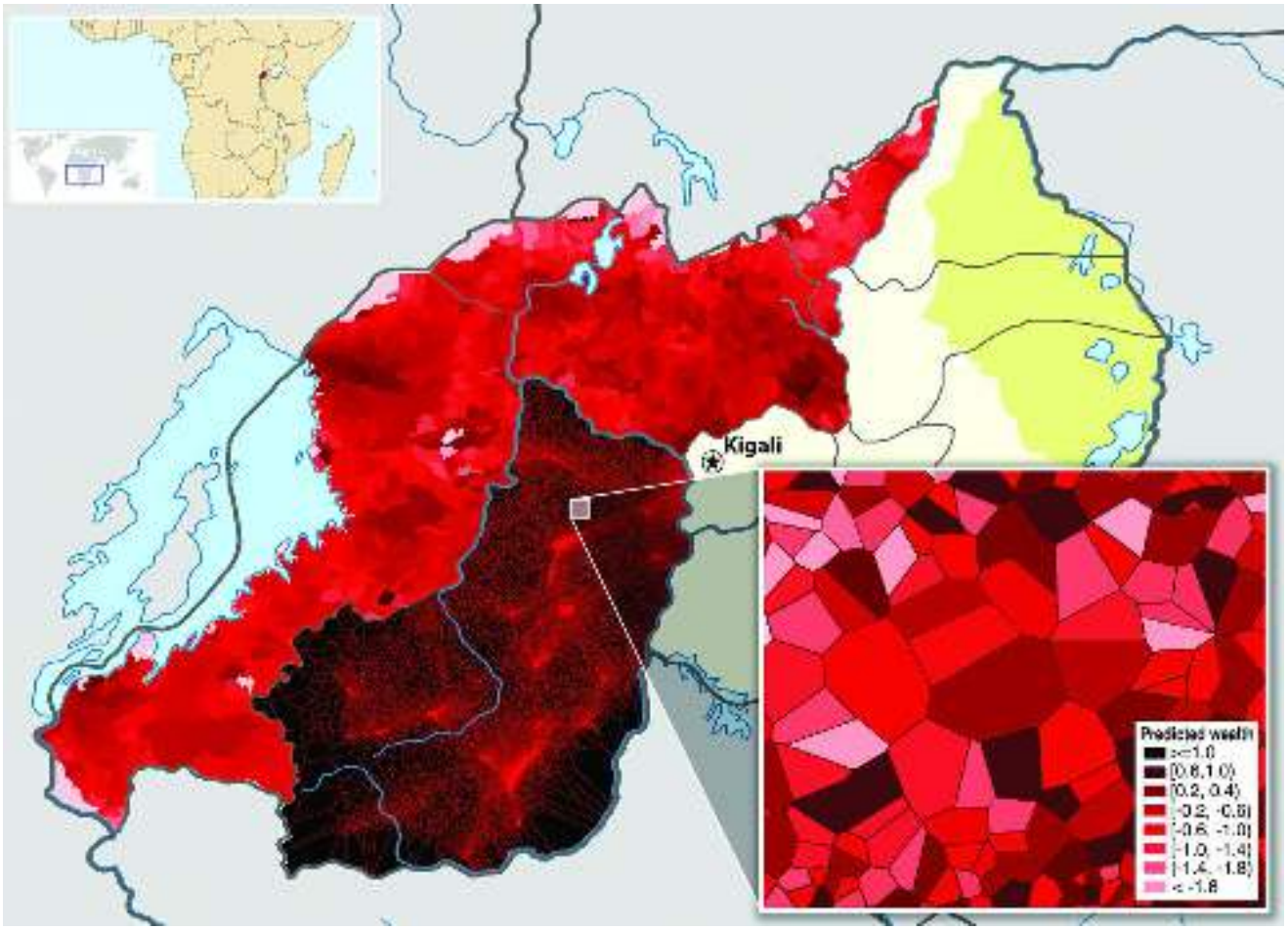
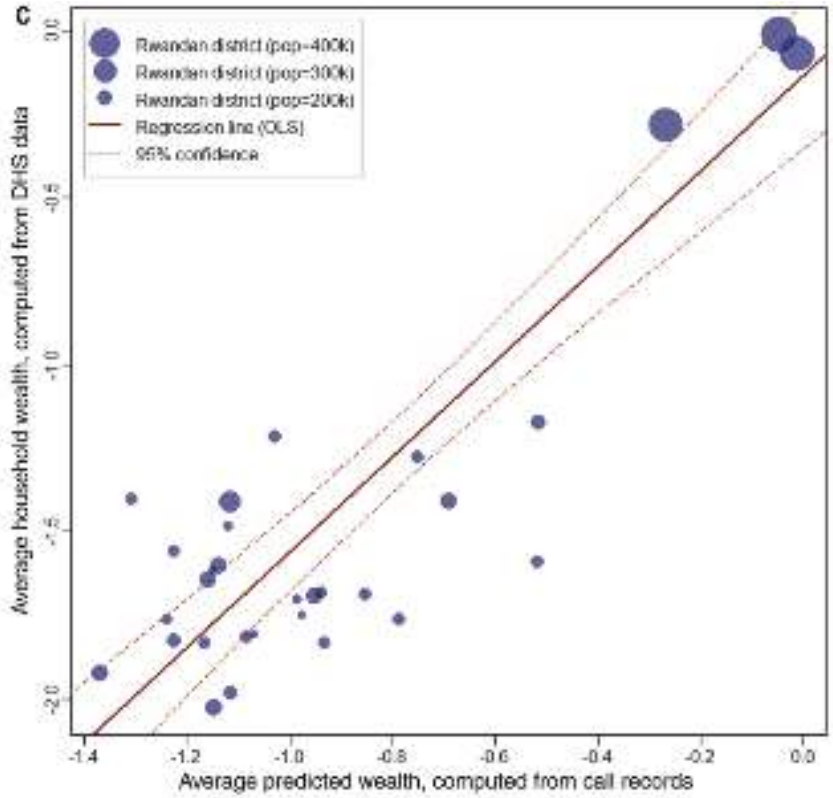


Image Credit: Salganik 2017 *Bit by Bit*

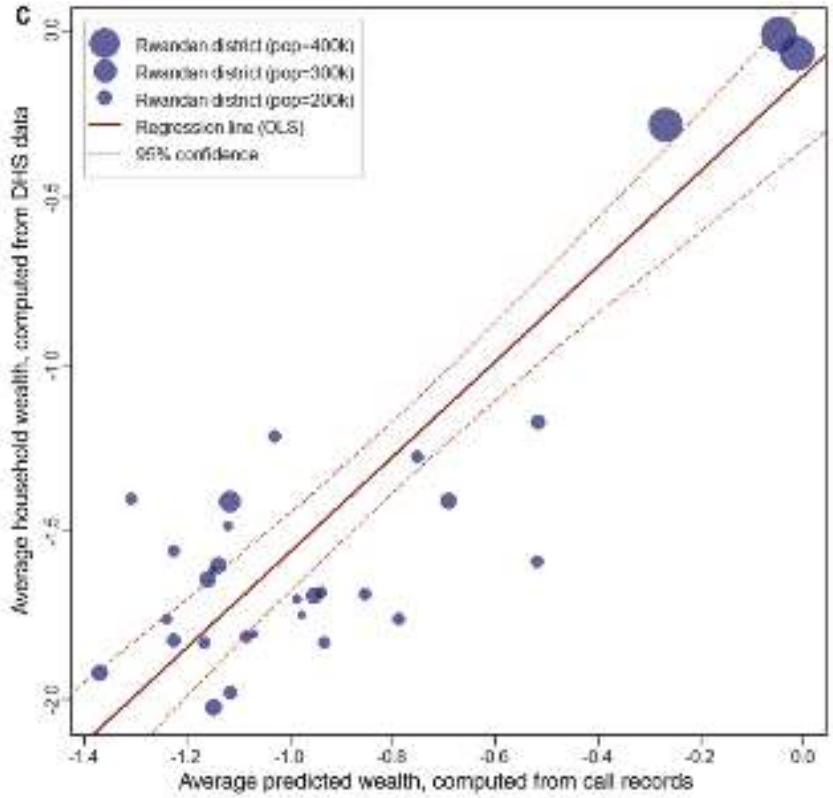

A
Survey Respondent
Loess Fit
95% confidence
Actual wealth (Composite Index)
Predicted wealth (Composite Index)
-3
-2
-1
0
1
2
3


Kigali
Predicted wealth
x<=1.0
(0.6, 1.0)
(0.2, 0.4)
(-0.2, -0.6)
(-0.6, -1.0)
(-1.0, -1.4)
(-1.4, -1.8)
x < -1.8

Comparing to Demographics and Health Study



Comparing to Demographics and Health Study



- 10 times faster
- 50 times cheaper

Limitations

Limitations

- Likely need a new random sample each time you would repeat (still faster and cheaper!)

Limitations

- Likely need a new random sample each time you would repeat (still faster and cheaper!)
- Only 856 people in the original study and they had been previously used in a predictive study (overfitting?)

Limitations

- Likely need a new random sample each time you would repeat (still faster and cheaper!)
- Only 856 people in the original study and they had been previously used in a predictive study (overfitting?)
- Results with DHS are correlated but wealth predictions are shifted and still with non-trivial error. Whether it is sufficient depends on the task!

Limitations

- Likely need a new random sample each time you would repeat (still faster and cheaper!)
- Only 856 people in the original study and they had been previously used in a predictive study (overfitting?)
- Results with DHS are correlated but wealth predictions are shifted and still with non-trivial error. Whether it is sufficient depends on the task!
- Were the original people really a random sample? (they had to pick up the phone and agree to answer questions)

Concluding Thoughts on Example

Concluding Thoughts on Example

- This is the same core strategy as used with text as data or image classification work in the social sciences.

Concluding Thoughts on Example

- This is the same core strategy as used with text as data or image classification work in the social sciences.
- This example is what Salganik calls Amplified Asking (more later!)

Concluding Thoughts on Example

- This is the same core strategy as used with text as data or image classification work in the social sciences.
- This example is what Salganik calls Amplified Asking (more later!)
- The key is that our test set only let's us assess performance if people we want to predict on come from same population as the original training data.

1. Preview Three Types of Prediction
2. Prediction for Enhanced Measurement
3. Fragile Families Challenge
4. Logistic Regression
5. Generalized Additive Models
6. Decision Trees

1 Preview Three Types of Prediction

2 Prediction for Enhanced Measurement

3 Fragile Families Challenge

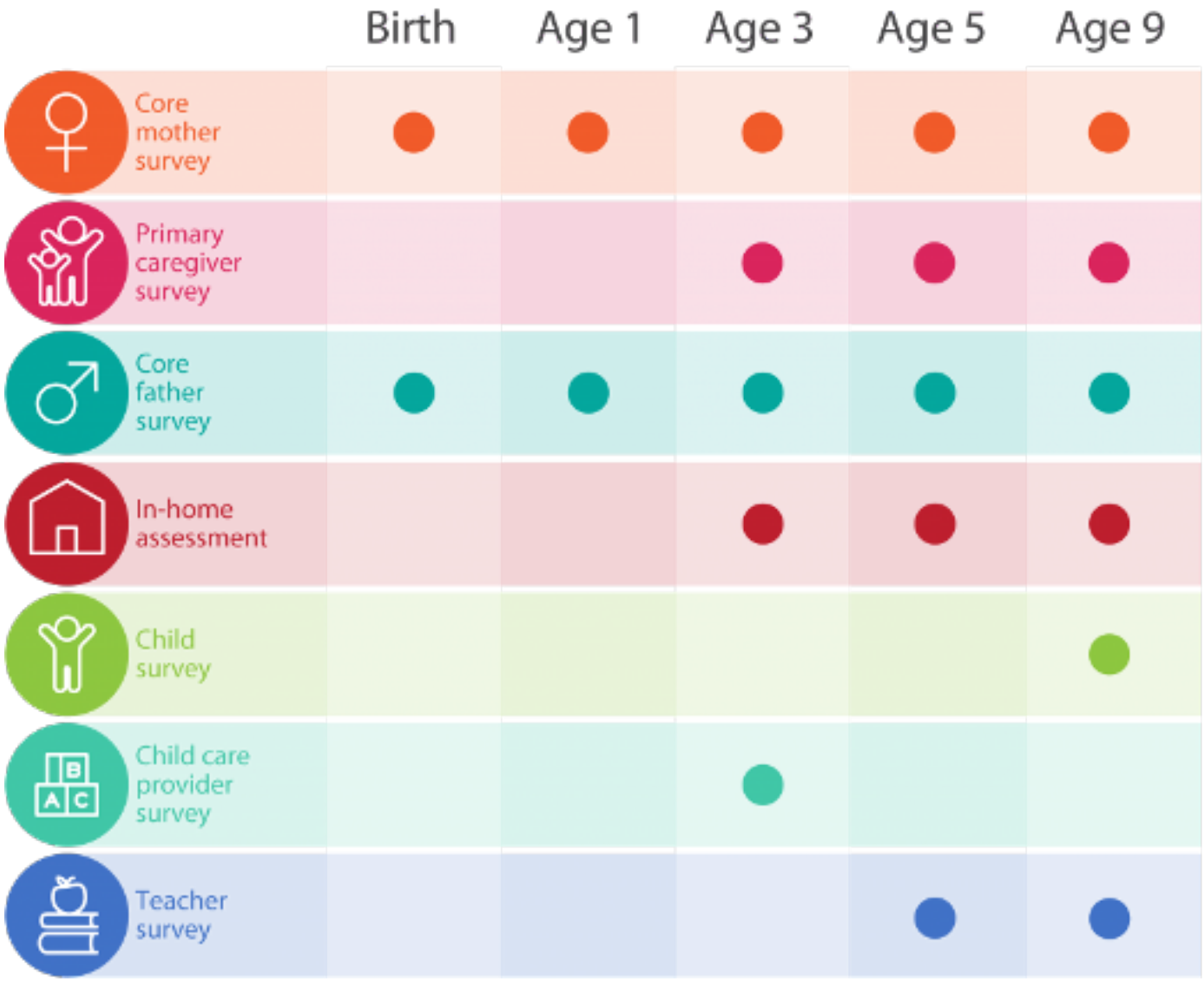
4 Logistic Regression

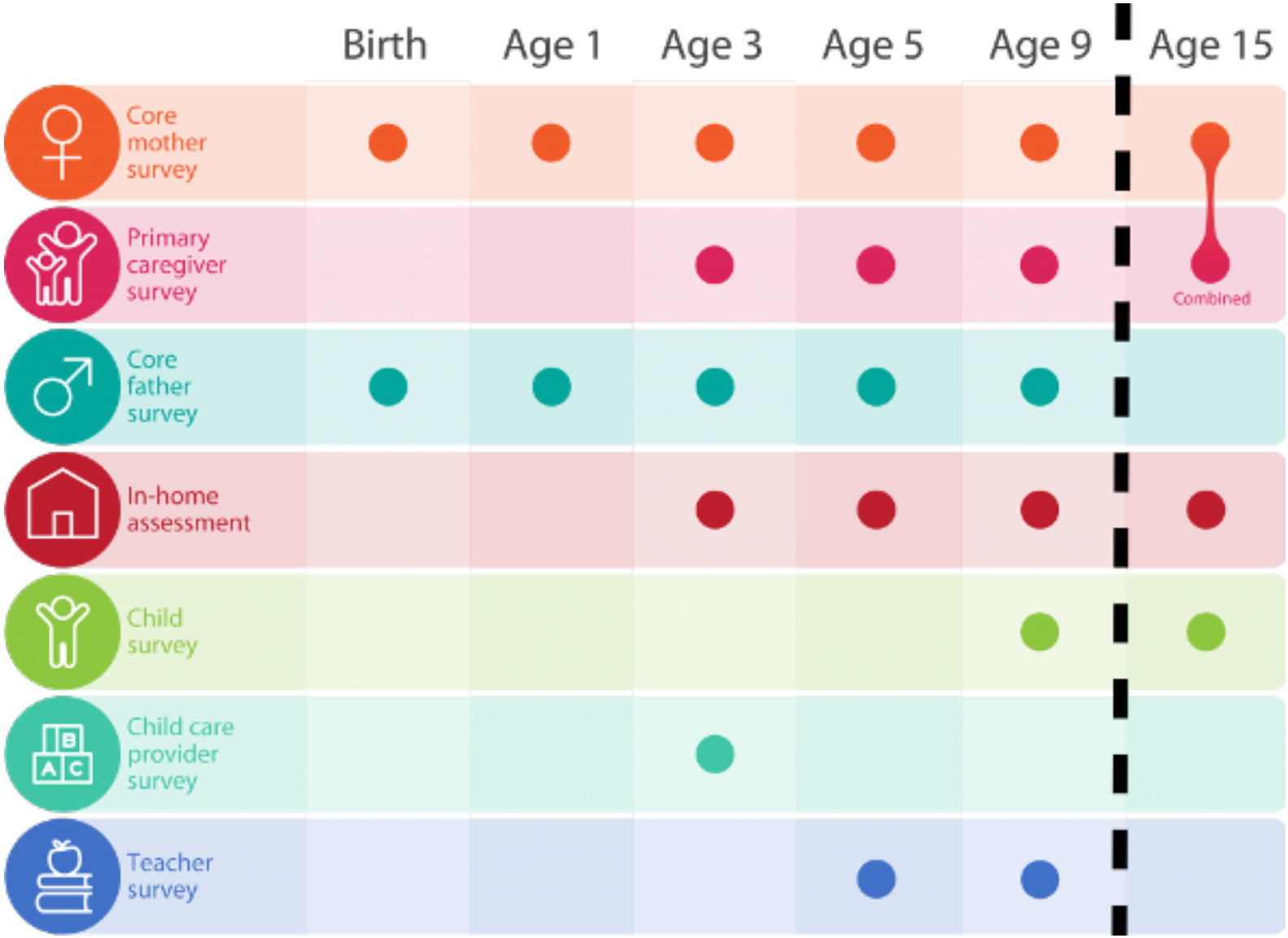
5 Generalized Additive Models

6 Decision Trees

Fragile Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

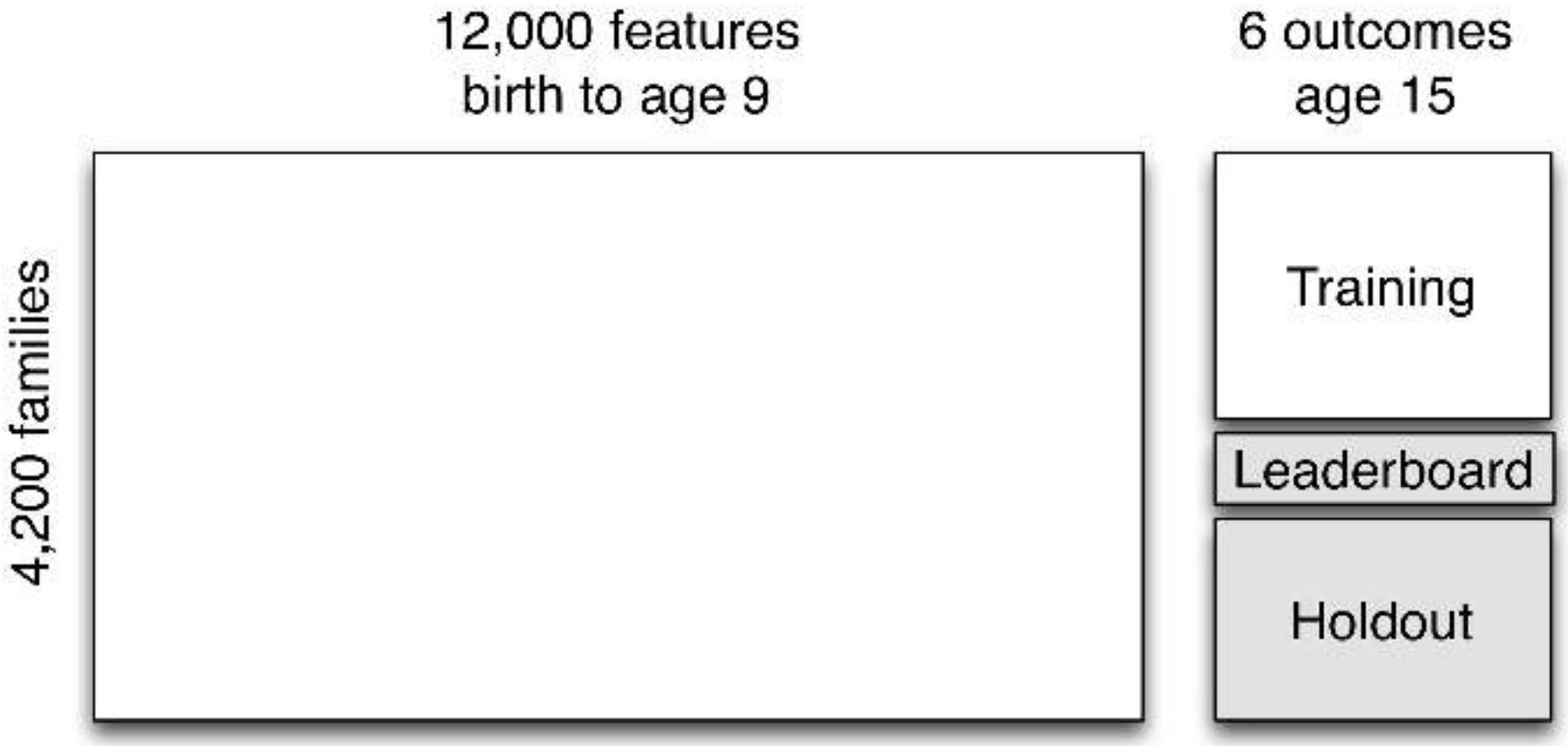
- Birth cohort panel study
- $\approx$ 5,000 children born in 20 U.S. cities
- Followed from birth through age 15


Birth
Age 1
Age 3
Age 5
Age 9
Core mother survey
Primary caregiver survey
Core father survey
In-home assessment
Child survey
Child care provider survey
Teacher survey


Birth
Age 1
Age 3
Age 5
Age 9
Age 15
Core mother survey
Primary caregiver survey
Combined
Core father survey
In-home assessment
Child survey
Child care provider survey
Teacher survey

Six age 15 outcomes:

- GPA
- Material Hardship
- Grit
- Evicted
- Job training
- Job loss


12,000 features
birth to age 9
6 outcomes
age 15
4,200 families
Training
Leaderboard
Holdout

Missing Data Abounds

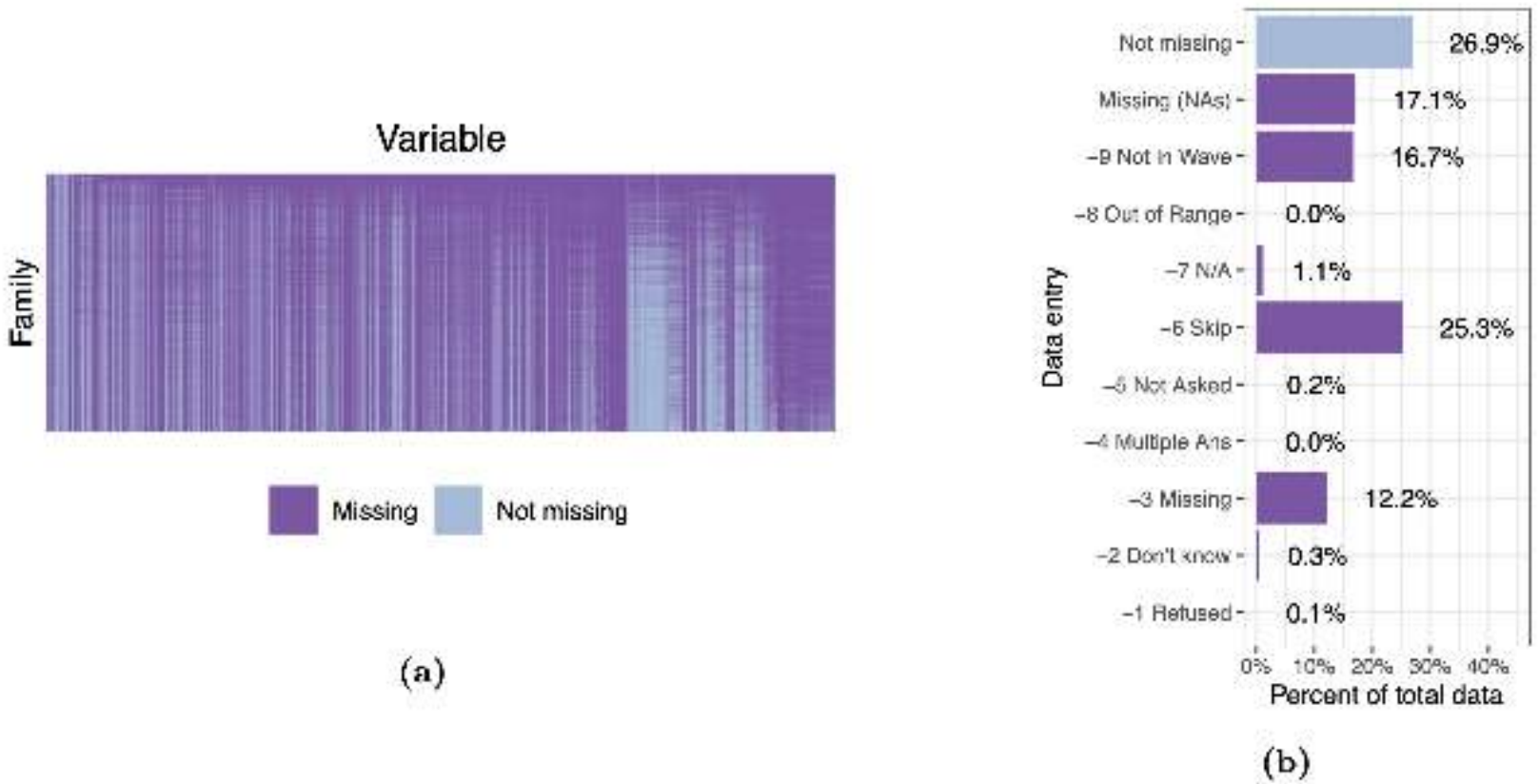
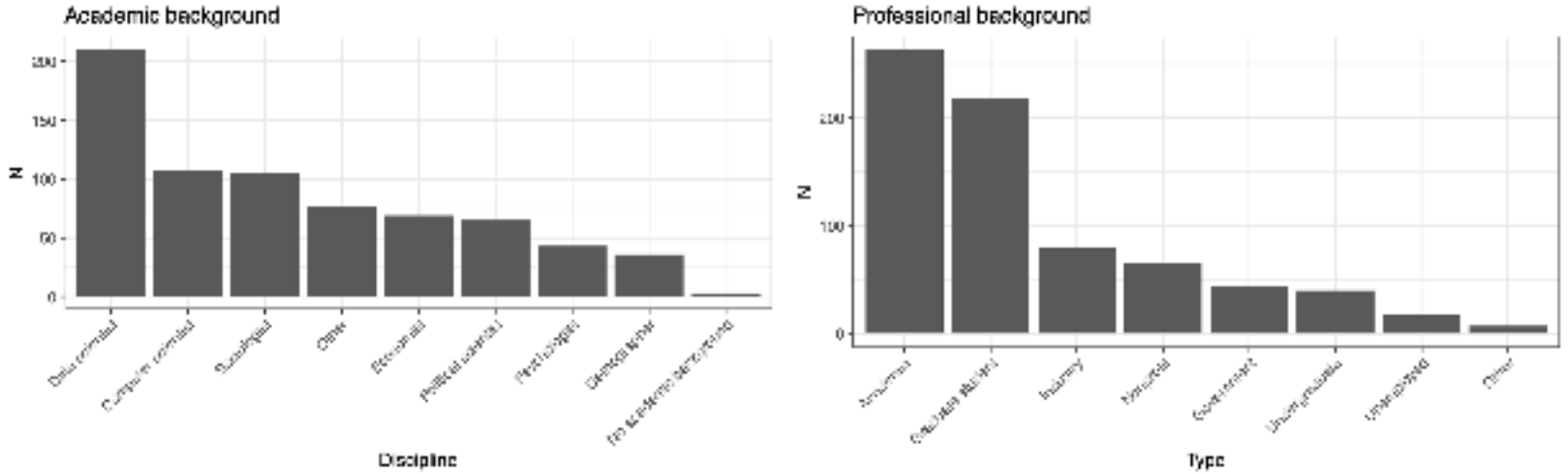


Fig. S1. Missing entries in the Fragile Families Challenge background dataset. The Fragile Families Challenge background dataset had 4,242 rows and 12,942 columns (plus an ID number). Of the approximately 55 million distinct data entries, about 73% were missing. There were many types of missing entries.



(a) (b)

Fig. S3. Self-reported disciplinary affiliation and professional background of applicants. Each applicant could select as many of these as applied. See [18] for a copy of the application.

441 registered participants

- social scientists and data scientists
- undergraduates, grad students, and professionals
- many working in teams

Many Different Algorithms

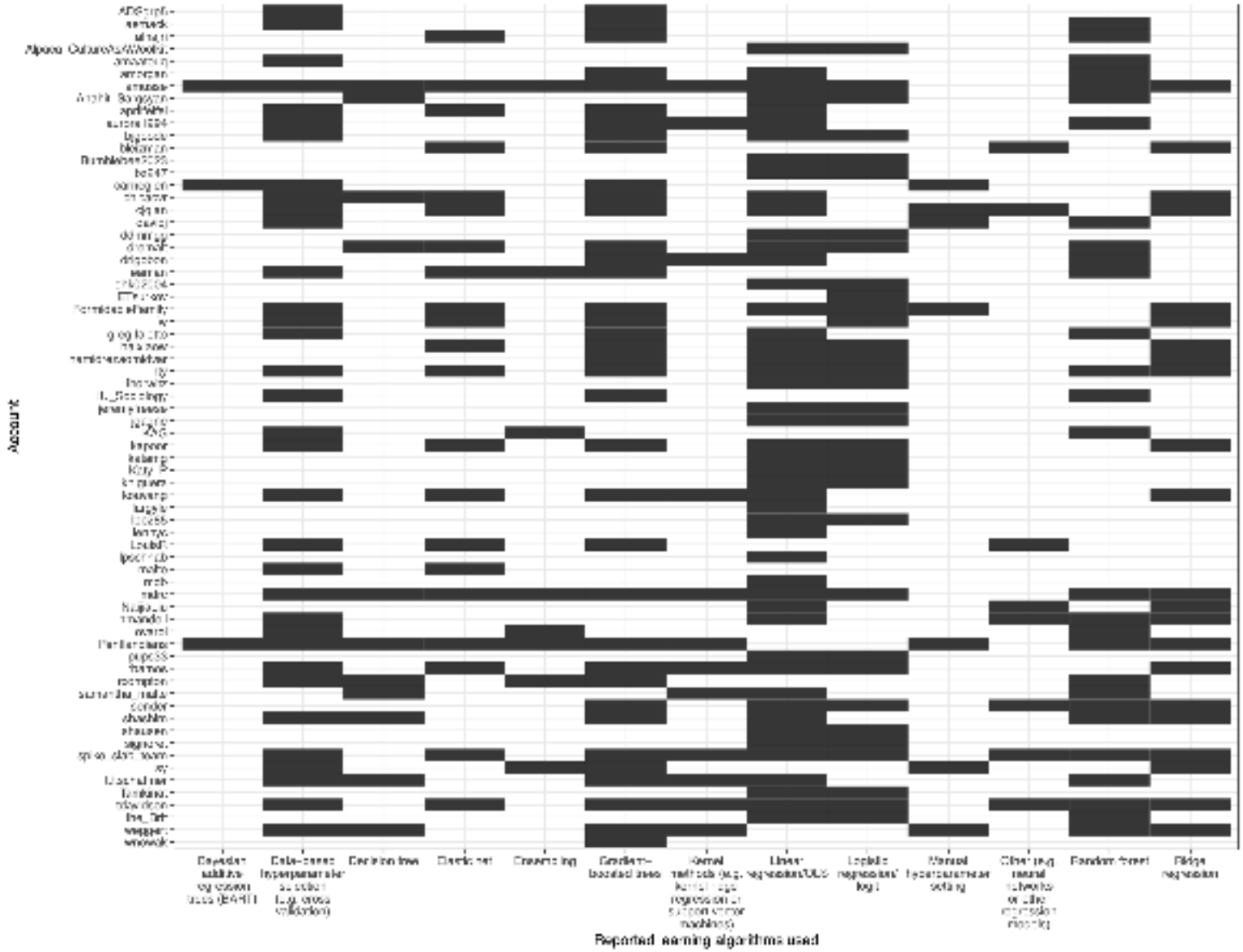
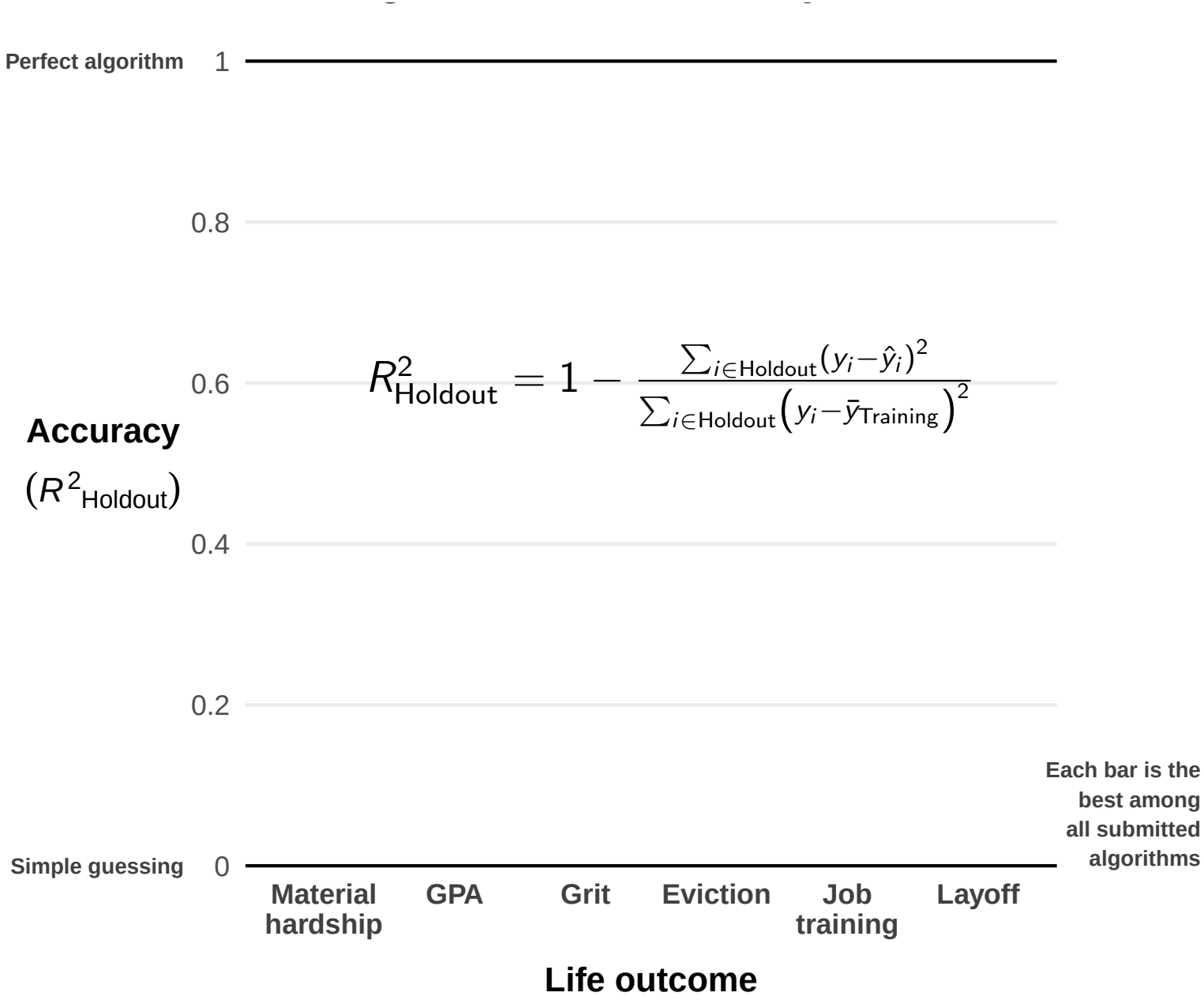
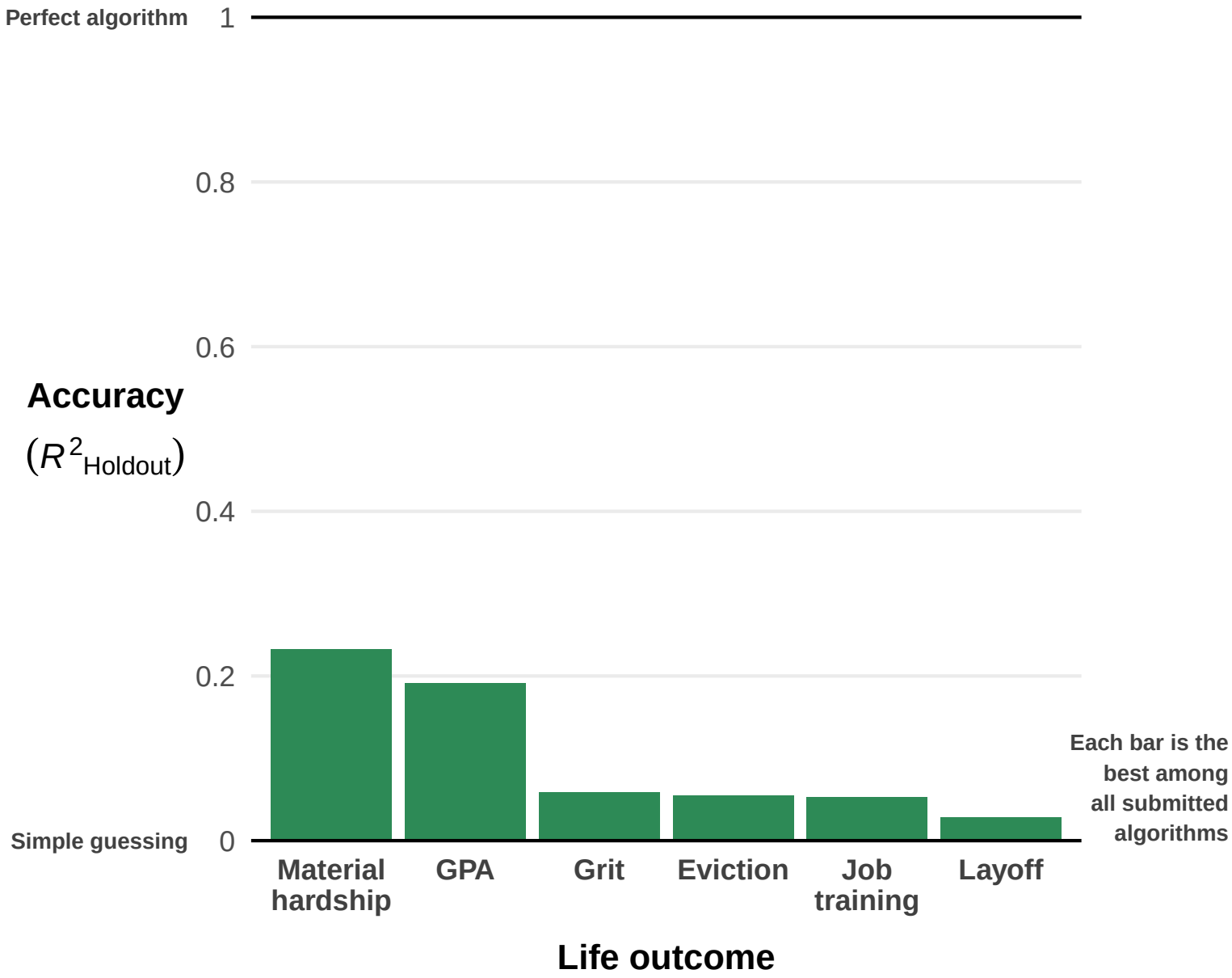


Fig. S21. Participant reports of learning algorithms used.

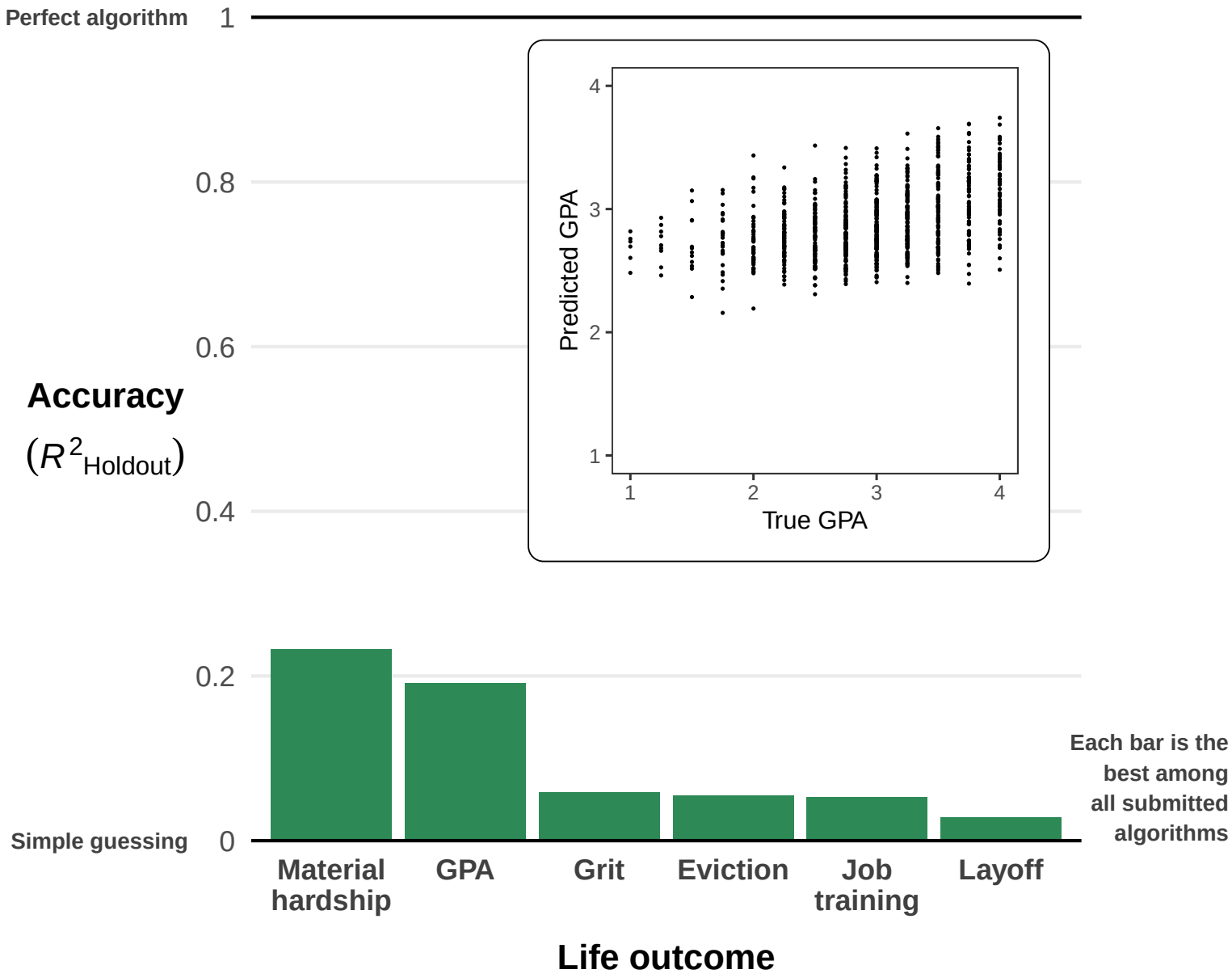
How did they do?


Perfect algorithm
1
0.8
0.6
0.4
0.2
Simple guessing
0
$R^2_{\text{Holdout}} = 1 - \frac{\sum_{i \in \text{Holdout}} (y_i - \hat{y}_i)^2}{\sum_{i \in \text{Holdout}} \left(y_i - \bar{y}_{\text{Training}}\right)^2}$
Accuracy
(R^2_{Holdout})
Each bar is the best among all submitted algorithms
Material hardship
GPA
Grit
Eviction
Job training
Layoff
Life outcome

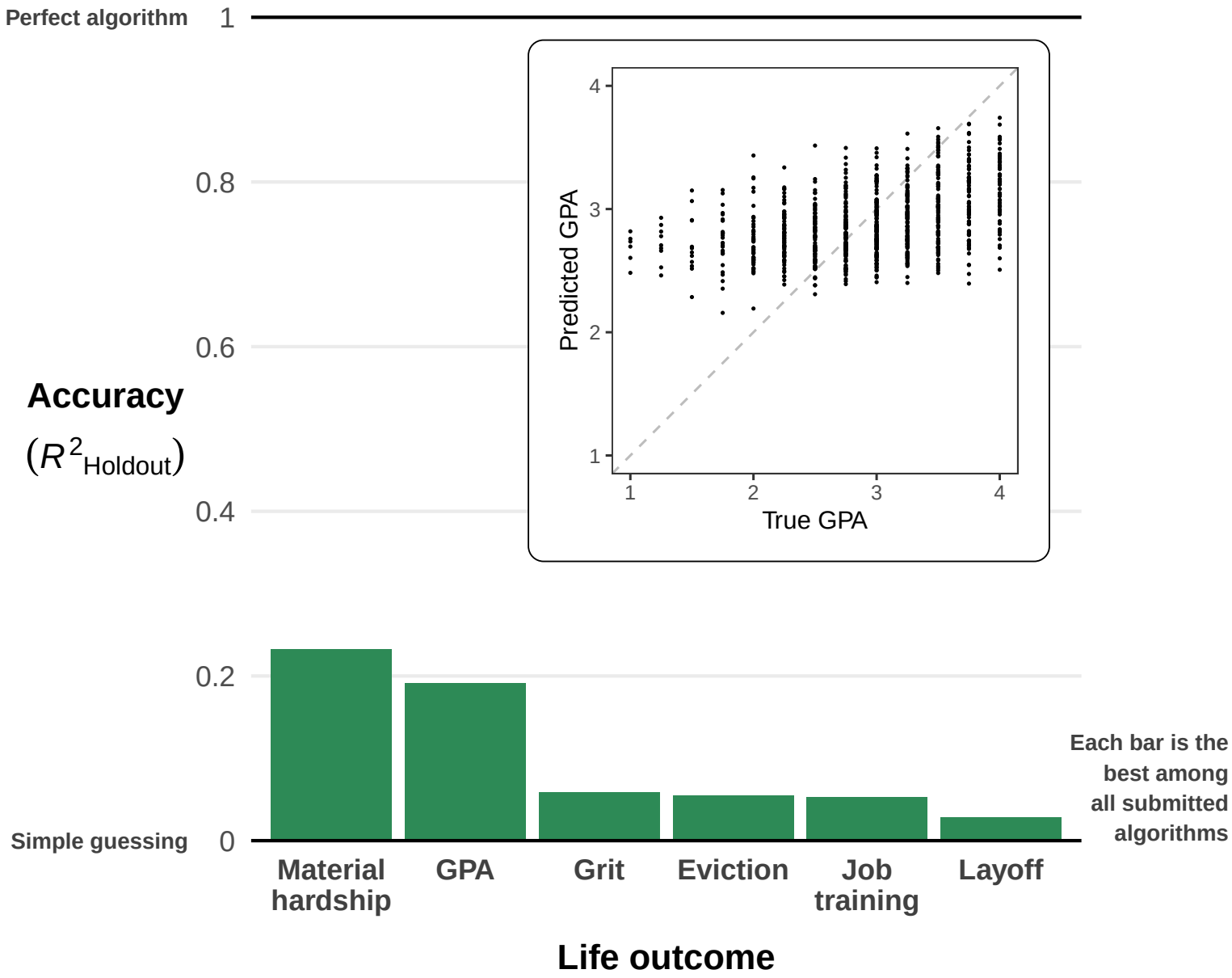
Best algorithms were not very accurate



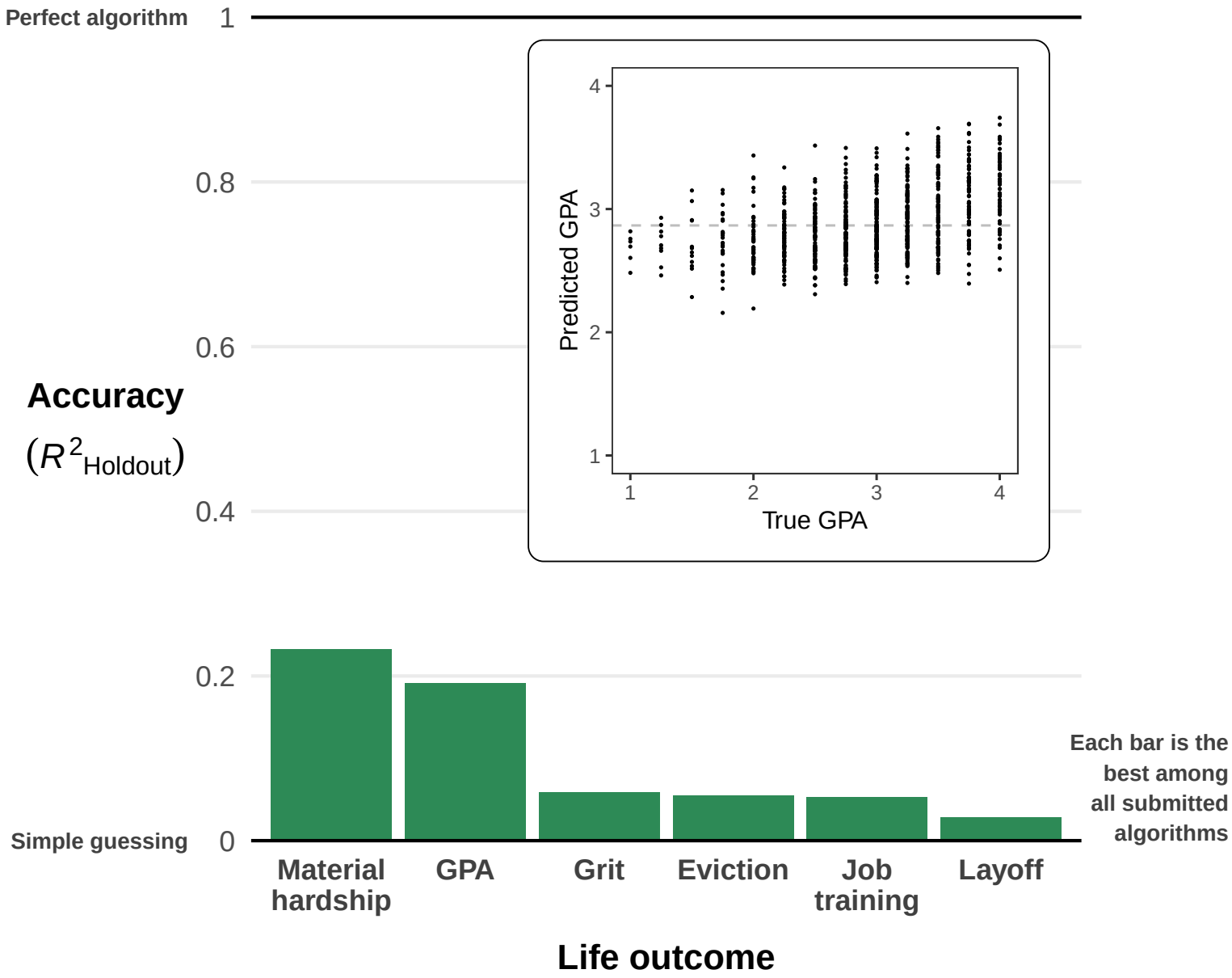
Best algorithms were not very accurate



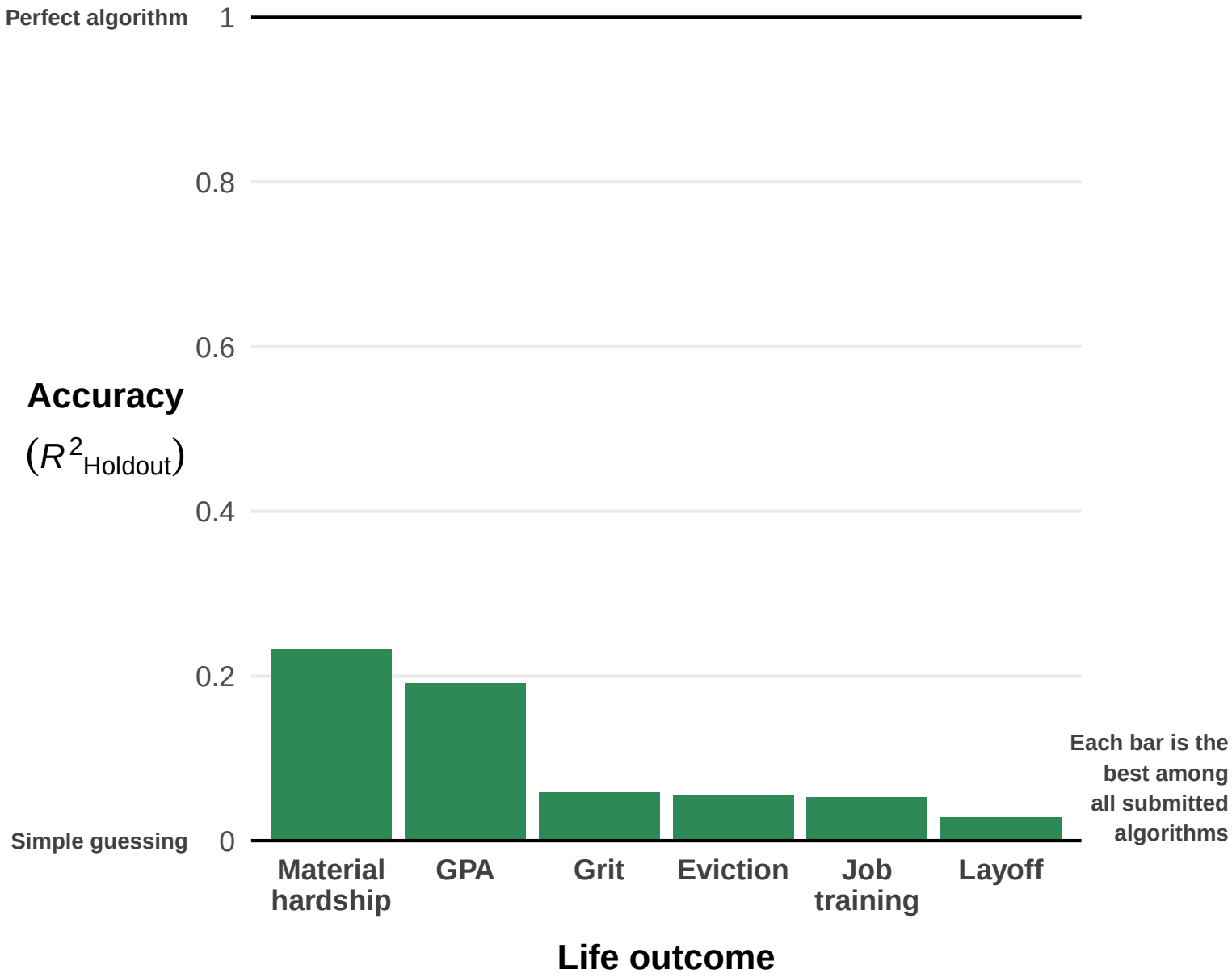
Best algorithms were not very accurate

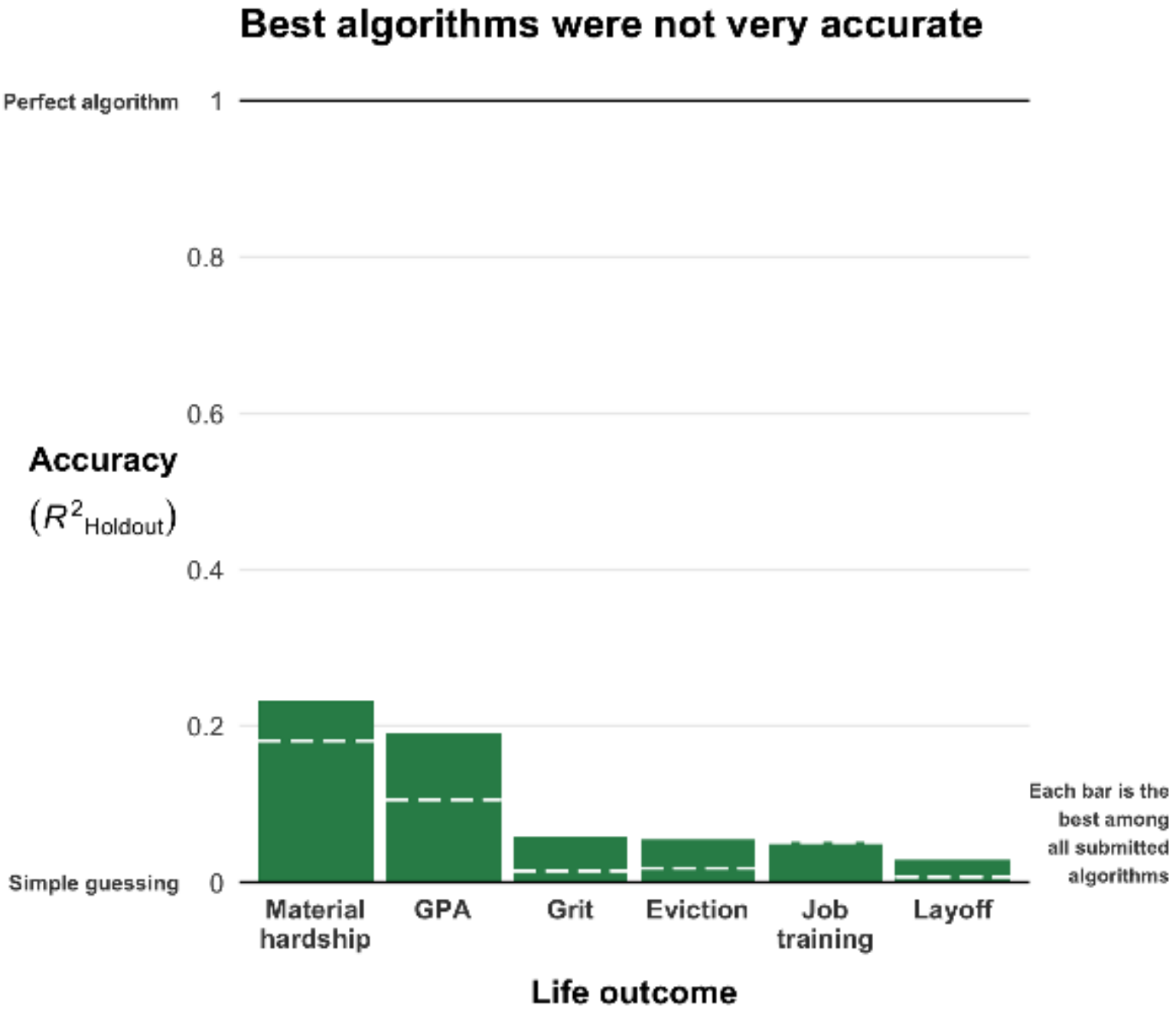


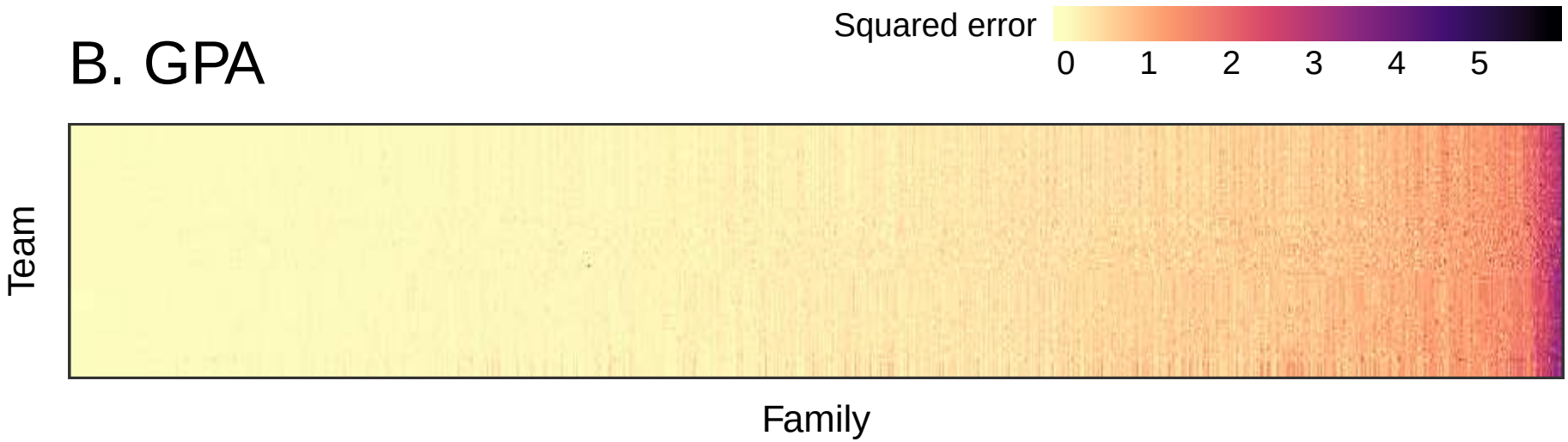
Best algorithms were not very accurate

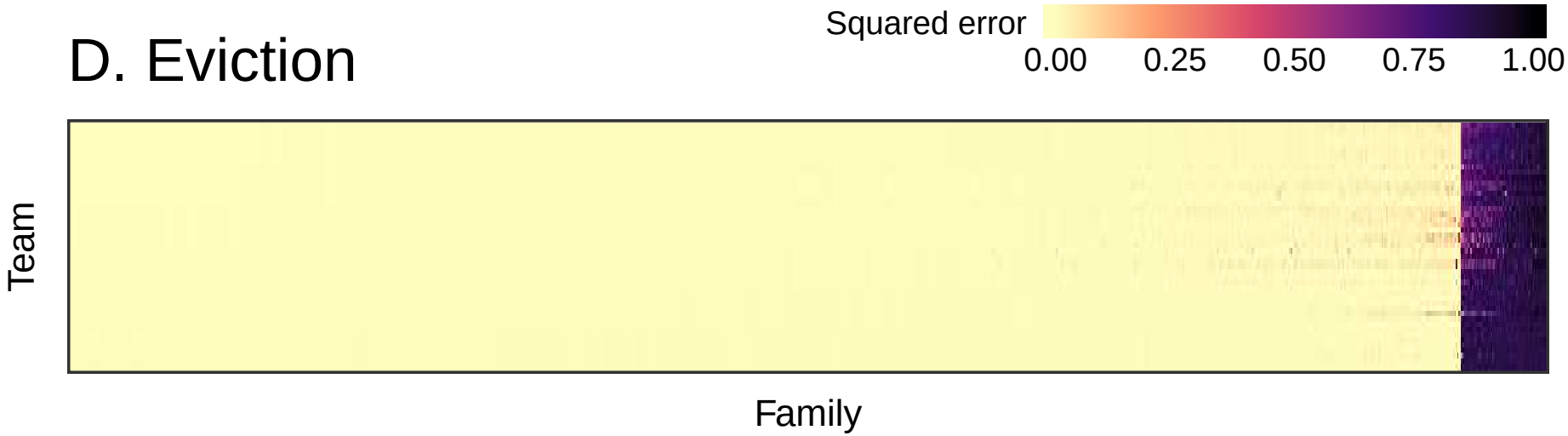


Best algorithms were not very accurate




Best algorithms were not very accurate
Perfect algorithm
1
0.8
0.6
Accuracy
($R^2_{Holdout}$)
0.4
0.2
Simple guessing
0
Each bar is the best among all submitted algorithms
Material hardship
GPA
Grit
Eviction
Job training
Layoff
Life outcome


B. GPA
Squared error
0
1
2
3
4
5
Team
Family


D. Eviction
Squared error
0.00
0.25
0.50
0.75
1.00
Team
Family

What do these results mean for policy and for science?

For **policymakers**,

For **policymakers**,

– Do not assume that predictive algorithms are accurate

For **policymakers**,

– Do not assume that predictive algorithms are accurate

– Transparent evaluation of any algorithm is needed

For **policymakers**,

– Do not assume that predictive algorithms are accurate

– Transparent evaluation of any algorithm is needed

– Complex models may not outperform simple models

For **scientists**: A paradox of prediction and understanding

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

...yet the very same data did not yield accurate predictions

↗

(the result)

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

...yet the very same data did not yield accurate predictions

↗

(the result)

Resolutions:

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

...yet the very same data did not yield accurate predictions

↗

(the result)

Resolutions:

1. If understanding implies an ability to predict, then **we do not understand** the life course.

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

...yet the very same data did not yield accurate predictions

↗

(the result)

Resolutions:

1. If understanding implies an ability to predict, then **we do not understand** the life course.
2. Perhaps we have **understanding despite poor prediction**.
   - Aggregate description
   - Causal effects

For **scientists**: A paradox of prediction and understanding

These data have generated understanding...

↖

(hundreds of papers)

...yet the very same data did not yield accurate predictions

↗

(the result)

Resolutions:

1. If understanding implies an ability to predict, then **we do not understand** the life course.
2. Perhaps we have **understanding despite poor prediction**.
   - Aggregate description
   - Causal effects
3. Our understanding is correct but **incomplete**. We need theories that point toward poor prediction

Was the Challenge Hard or Easy?

Was the Challenge Hard or Easy?

- Hard

Was the Challenge Hard or Easy?

- Hard
  - measurements likely noisy for a number of variables

Was the Challenge Hard or Easy?

- Hard
  - measurements likely noisy for a number of variables
  - age 9 to age 15 is a big gap

Was the Challenge Hard or Easy?

- Hard
  - measurements likely noisy for a number of variables
  - age 9 to age 15 is a big gap
  - life outcomes are a tough thing to predict!

Was the Challenge Hard or Easy?

- Hard
  - measurements likely noisy for a number of variables
  - age 9 to age 15 is a big gap
  - life outcomes are a tough thing to predict!
  - tons of missing data

Was the Challenge Hard or Easy?

- Hard
  - measurements likely noisy for a number of variables
  - age 9 to age 15 is a big gap
  - life outcomes are a tough thing to predict!
  - tons of missing data
- Easy

Was the Challenge Hard or Easy?

- Hard
  - measurements likely noisy for a number of variables
  - age 9 to age 15 is a big gap
  - life outcomes are a tough thing to predict!
  - tons of missing data
- Easy
  - training and test set come from the same distribution by construction.

Was the Challenge Hard or Easy?

- Hard
  - measurements likely noisy for a number of variables
  - age 9 to age 15 is a big gap
  - life outcomes are a tough thing to predict!
  - tons of missing data
- Easy
  - training and test set come from the same distribution by construction.
  - easier than the policy-relevant task (no age 15 data, predicting to different cohorts)

1. Preview Three Types of Prediction
2. Prediction for Enhanced Measurement
3. Fragile Families Challenge
4. Logistic Regression
5. Generalized Additive Models
6. Decision Trees

1 Preview Three Types of Prediction

2 Prediction for Enhanced Measurement

3 Fragile Families Challenge

4 Logistic Regression

5 Generalized Additive Models

6 Decision Trees

Last Time: Linear Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} \sum_i (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j$$

What Happens with Binary Variables?

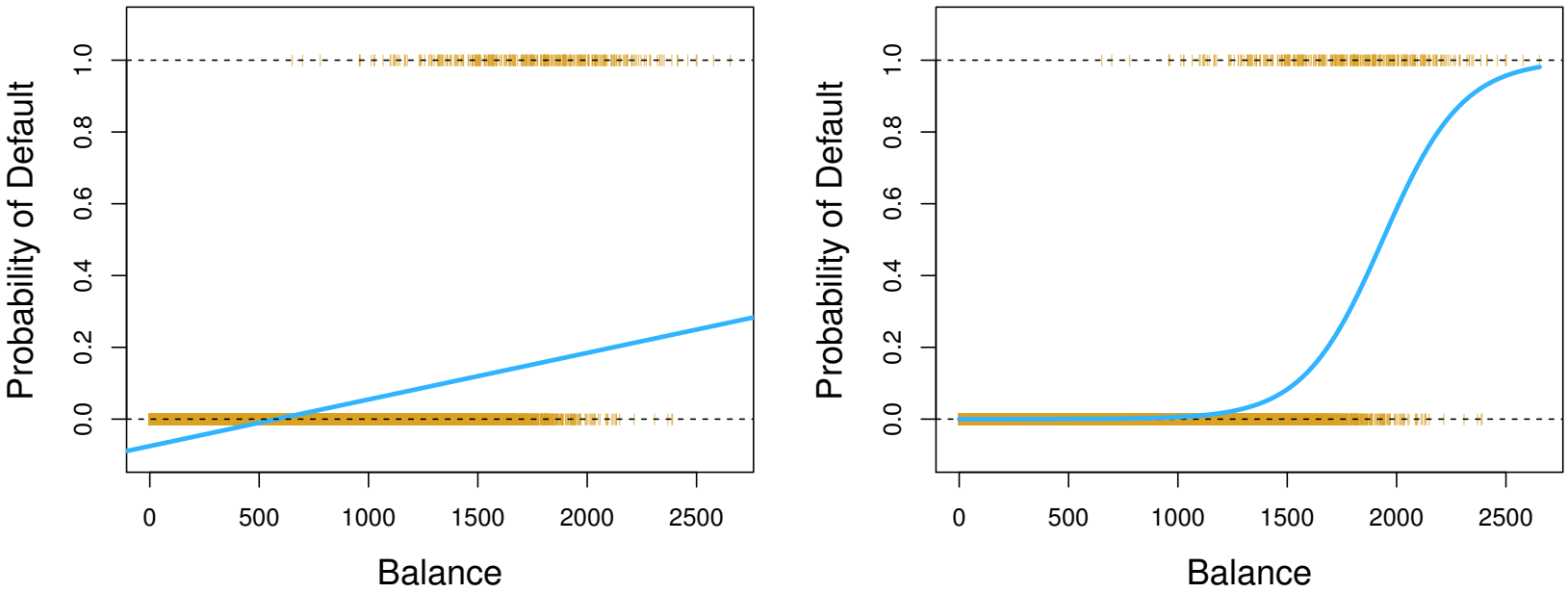


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

Logistic Regression

Logistic Regression

- Core Problem: for unbounded inputs, linear prediction $\hat{y}_i$ can always span $-\infty$ to ∞

Logistic Regression

- Core Problem: for unbounded inputs, linear prediction $\hat{y}_i$ can always span $-\infty$ to ∞
- Key idea of logistic regression $\rightarrow$ take the linear form and squish it down to run the range of 0 to 1.

Logistic Regression

- Core Problem: for unbounded inputs, linear prediction $\hat{y}_i$ can always span $-\infty$ to ∞
- Key idea of logistic regression $\rightarrow$ take the linear form and squish it down to run the range of 0 to 1.
- Two equivalent ways of thinking about this:

Logistic Regression

- Core Problem: for unbounded inputs, linear prediction $\hat{y}_i$ can always span $-\infty$ to ∞
- Key idea of logistic regression $\rightarrow$ take the linear form and squish it down to run the range of 0 to 1.
- Two equivalent ways of thinking about this:
  1) linear model of log odds

Logistic Regression

- Core Problem: for unbounded inputs, linear prediction $\hat{y}_i$ can always span $-\infty$ to ∞
- Key idea of logistic regression $\rightarrow$ take the linear form and squish it down to run the range of 0 to 1.
- Two equivalent ways of thinking about this:
  1) linear model of log odds
  2) project line into 0 to 1 space using a convenient function

(1) Log Odds Version

(1) Log Odds Version

- We want something that ranges $-\infty$ to $+\infty$

(1) Log Odds Version

- We want something that ranges $-\infty$ to $+\infty$
- Denote probability $p(Y_i = 1|X_i) = \pi_i$. π_i ranges [0,1].

(1) Log Odds Version

- We want something that ranges $-\infty$ to $+\infty$
- Denote probability $p(Y_i = 1|X_i) = \pi_i$. π_i ranges [0,1].
- The odds are $\frac{\pi_i}{1-\pi_i}$ and range $(0, \infty)$

(1) Log Odds Version

- We want something that ranges $-\infty$ to $+\infty$
- Denote probability $p(Y_i = 1|X_i) = \pi_i$. π_i ranges [0,1].
- The odds are $\frac{\pi_i}{1-\pi_i}$ and range $(0, \infty)$
- The log function takes a value that is in range $(0, \infty)$ and converts to $(-\infty, \infty)$

(1) Log Odds Version

- We want something that ranges $-\infty$ to $+\infty$
- Denote probability $p(Y_i = 1|X_i) = \pi_i$. π_i ranges [0,1].
- The odds are $\frac{\pi_i}{1-\pi_i}$ and range $(0, \infty)$
- The log function takes a value that is in range $(0, \infty)$ and converts to $(-\infty, \infty)$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$$

$$\pi_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j)}}$$

(2) Convenient Function Version

(2) Convenient Function Version

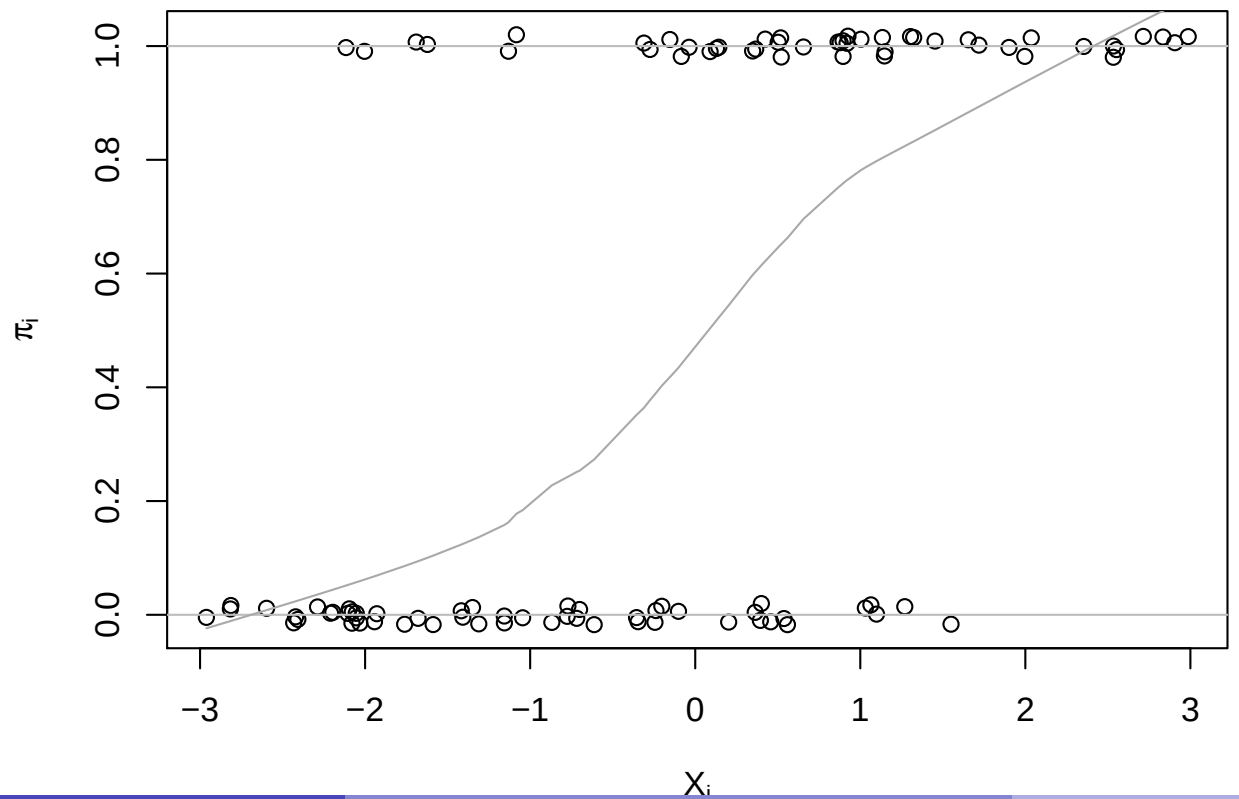
- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.

(2) Convenient Function Version

- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.
- Different functions lead to (slightly) different models.

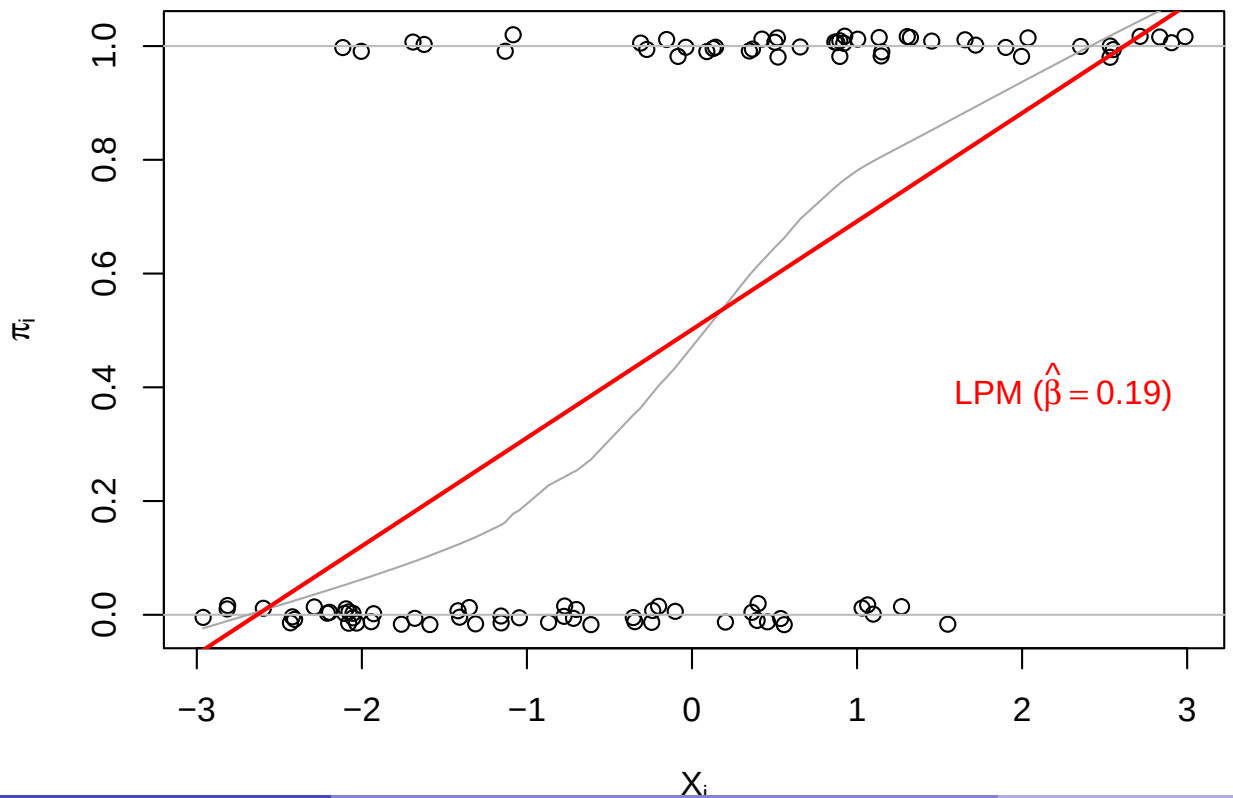
(2) Convenient Function Version

- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.
- Different functions lead to (slightly) different models.



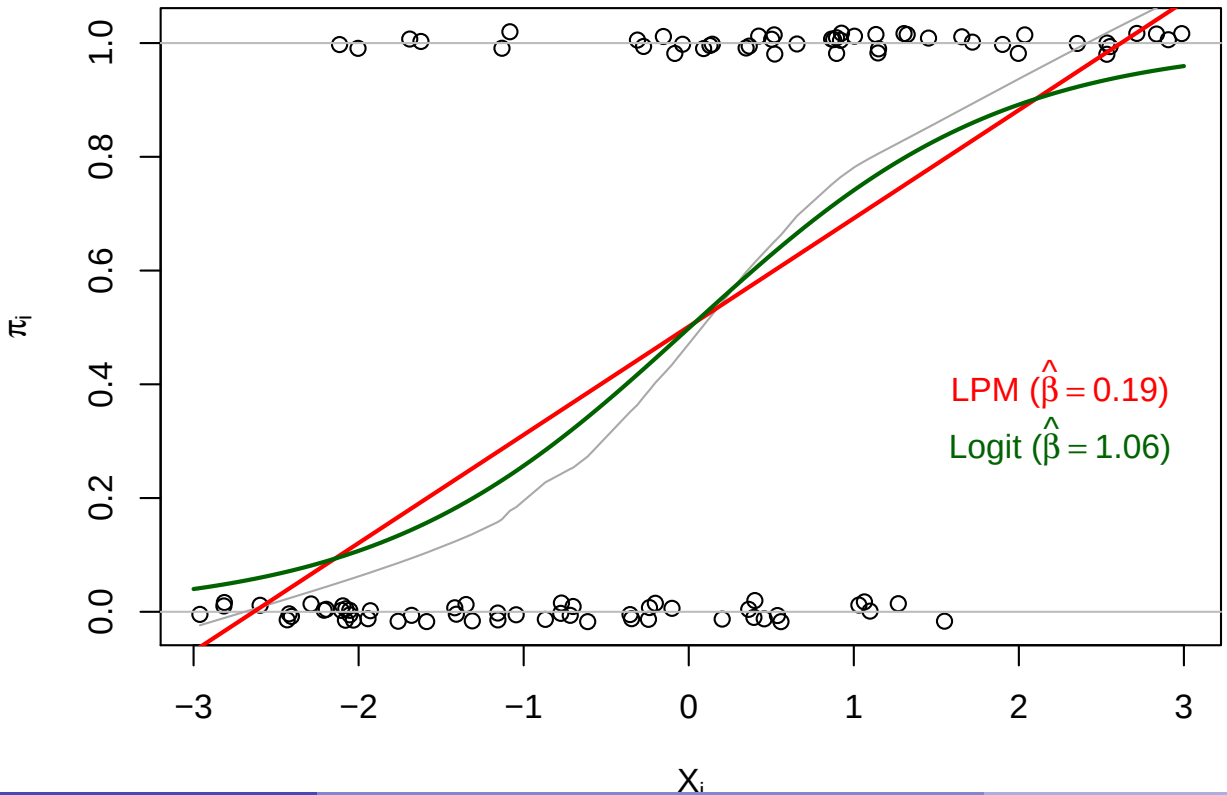
(2) Convenient Function Version

- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.
- Different functions lead to (slightly) different models.



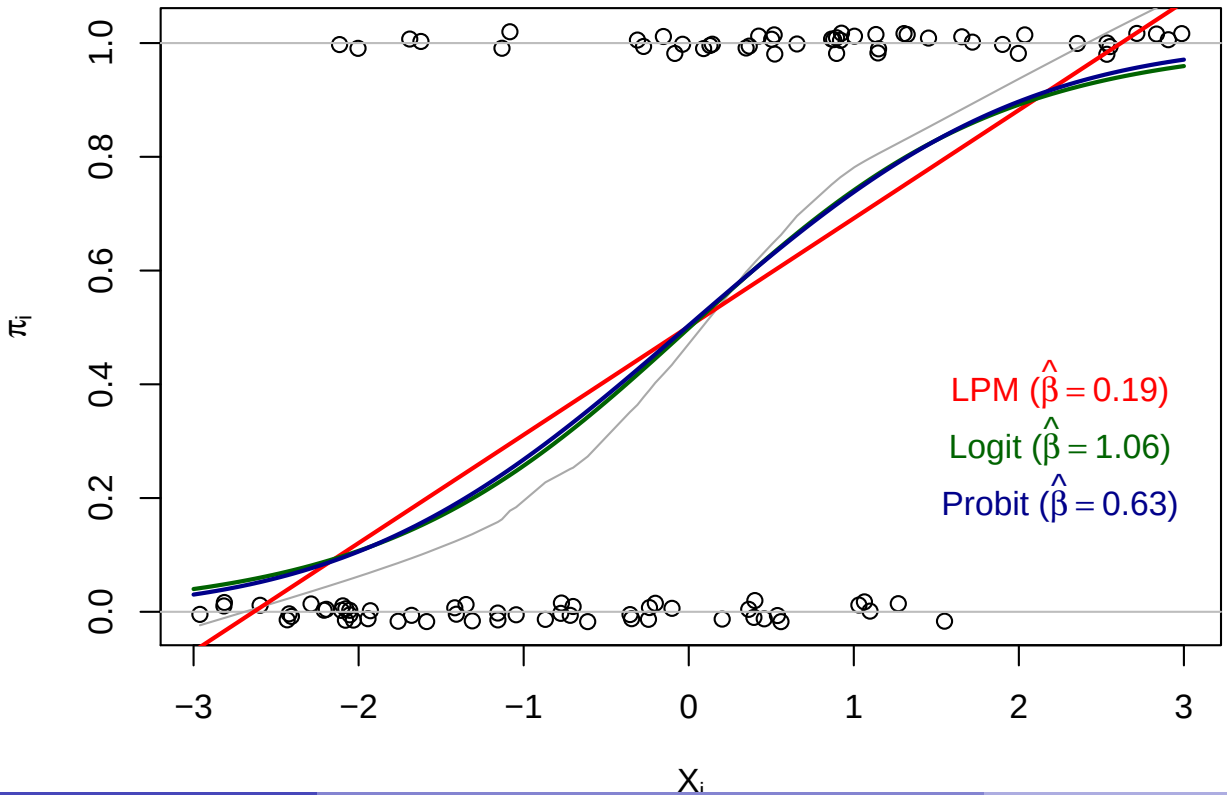
(2) Convenient Function Version

- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.
- Different functions lead to (slightly) different models.



(2) Convenient Function Version

- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.
- Different functions lead to (slightly) different models.



Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$

Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$
- To get the loss function we will use maximum likelihood estimation.

Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$
- To get the loss function we will use maximum likelihood estimation.
- Core Idea: we are modeling the probability of seeing the data we observe. We want that probability to be as high as possible.

Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$
- To get the loss function we will use maximum likelihood estimation.
- Core Idea: we are modeling the probability of seeing the data we observe. We want that probability to be as high as possible.
- In the case of categorical data this leads to what is called the cross entropy loss in machine learning.

Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$
- To get the loss function we will use maximum likelihood estimation.
- Core Idea: we are modeling the probability of seeing the data we observe. We want that probability to be as high as possible.
- In the case of categorical data this leads to what is called the cross entropy loss in machine learning.

$$\arg\max \prod_{Y_i=1} \pi_i \prod_{Y_i=0} (1 - \pi_i)$$

Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$
- To get the loss function we will use maximum likelihood estimation.
- Core Idea: we are modeling the probability of seeing the data we observe. We want that probability to be as high as possible.
- In the case of categorical data this leads to what is called the cross entropy loss in machine learning.

$$\arg\max \quad \prod_{Y_i=1} \pi_i \prod_{Y_i=0} (1-\pi_i)$$

$$\arg\min \quad -\left(\sum_{Y_i=1} \log \pi_i\right) - \left(\sum_{Y_i=0} \log (1-\pi_i)\right)$$

Logistic Regression

$$\hat{\beta} = \underset{\tilde{\beta}}{\arg\min} - \left(\sum_{Y_i=1} \log \pi_i \right) - \left(\sum_{Y_i=0} \log \left(1 - \pi_i\right) \right) \tag{1}$$

$$\pi_i = \frac{1}{1 + e^{-(\tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j} \tilde{\beta}_j)}} \tag{2}$$

Logistic Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} - \left(\sum_{Y_i=1} \log \pi_i \right) - \left(\sum_{Y_i=0} \log \left(1 - \pi_i\right) \right) \tag{1}$$

$$\pi_i = \frac{1}{1 + e^{-(\tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j)}} \tag{2}$$

Key Notes:

- Just as linear regression always makes a line, this will always make a line in log-odds (or more generally an S-shaped curve).

Logistic Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} - \left(\sum_{Y_i=1} \log \pi_i \right) - \left(\sum_{Y_i=0} \log (1 - \pi_i) \right) \tag{1}$$

$$\pi_i = \frac{1}{1 + e^{-(\tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j} \tilde{\beta}_j)}} \tag{2}$$

Key Notes:

- Just as linear regression always makes a line, this will always make a line in log-odds (or more generally an S-shaped curve).
- We can generalize this in the same way we did for linear regression by making the line more flexible with basis functions.

Logistic Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} -\left(\sum_{Y_i=1} \log \pi_i\right) - \left(\sum_{Y_i=0} \log\left(1-\pi_i\right)\right) \tag{1}$$

$$\pi_i = \frac{1}{1+e^{-(\tilde{\beta}_0+\sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j)}} \tag{2}$$

Key Notes:

- Just as linear regression always makes a line, this will always make a line in log-odds (or more generally an S-shaped curve).
- We can generalize this in the same way we did for linear regression by making the line more flexible with basis functions.
- Predictions are in the form of probability that the outcome is 1.

Generalization to Many Outcome Categories

- This strategy works in the general setting of more than two outcome categories.
  - Social Science Name: multinomial logistic regression
  - Machine Learning Name: softmax classifier

Generalization to Many Outcome Categories

- This strategy works in the general setting of more than two outcome categories.
  - Social Science Name: multinomial logistic regression
  - Machine Learning Name: softmax classifier
- New setup:

$$\pi_{ik} = \Pr(Y_i = k \mid X_i) = \frac{e^{\beta_{0,k} + \sum_{j=1}^{p} X_{i,j}\beta_{j,k}}}{\sum_{k'=1}^{K} e^{\beta_{0,k'} + \sum_{j=1}^{p} X_{i,j}\beta_{j,k'}}}$$

Generalization to Many Outcome Categories

- This strategy works in the general setting of more than two outcome categories.
  - Social Science Name: multinomial logistic regression
  - Machine Learning Name: softmax classifier
- New setup:

$$\pi_{ik} = \Pr(Y_i = k \mid X_i) = \frac{e^{\beta_{0,k} + \sum_{j=1}^{p} X_{i,j}\beta_{j,k}}}{\sum_{k'=1}^{K} e^{\beta_{0,k'} + \sum_{j=1}^{p} X_{i,j}\beta_{j,k'}}}$$

- Predictions give a probability of falling in each outcome category.

Subtleties to Be Aware Of

Subtleties to Be Aware Of

- If some variable predicts the outcome perfectly in the training set you can get a problem called perfect separation where one or more coefficient pushes towards ∞ or $-\infty$.

Subtleties to Be Aware Of

- If some variable predicts the outcome perfectly in the training set you can get a problem called perfect separation where one or more coefficient pushes towards ∞ or $-\infty$.
- The uncertainty of logistic regression is connected to the rarest class.

Subtleties to Be Aware Of

- If some variable predicts the outcome perfectly in the training set you can get a problem called perfect separation where one or more coefficient pushes towards ∞ or $-\infty$.
- The uncertainty of logistic regression is connected to the rarest class.
- The coefficients do not have the same interpretation as regression coefficients.

Summary of Logistic Regression

Summary of Logistic Regression

- Works for binary or categorical outcomes

Summary of Logistic Regression

- Works for binary or categorical outcomes
- Provides predictions in probabilities

Summary of Logistic Regression

- Works for binary or categorical outcomes
- Provides predictions in probabilities

Going Deeper:

James, Witten, Hastie and Tibshirani (2021) *An Introduction to Statistical Learning with Applications in R*. Chapters 4.0–4.3
Free online at `www.statlearning.com`

1. Preview Three Types of Prediction
2. Prediction for Enhanced Measurement
3. Fragile Families Challenge
4. Logistic Regression
5. Generalized Additive Models
6. Decision Trees

1 Preview Three Types of Prediction

2 Prediction for Enhanced Measurement

3 Fragile Families Challenge

4 Logistic Regression

5 Generalized Additive Models

6 Decision Trees

Generalized Additive Models (GAM)

Recall the linear model,

$$\hat{y}_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \dots$$

Generalized Additive Models (GAM)

Recall the linear model,

$$\hat{y}_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \dots$$

For GAMs, we maintain additivity, but instead of imposing termwise linearity we allow flexible functional forms for each explanatory variable, where $s_1(\cdot), s_2(\cdot),$ and $s_3(\cdot)$ are smooth functions that are estimated from the data:

Generalized Additive Models (GAM)

Recall the linear model,

$$\hat{y}_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \dots$$

For GAMs, we maintain additivity, but instead of imposing termwise linearity we allow flexible functional forms for each explanatory variable, where $s_1(\cdot), s_2(\cdot),$ and $s_3(\cdot)$ are smooth functions that are estimated from the data:

$$\hat{y}_i = \beta_0 + s_1(X_{1i}) + s_2(X_{2i}) + s_3(X_{3i}) + \dots$$

Generalized Additive Models (GAM)

$$\hat{y}_i = \beta_0 + s_1(X_{1i}) + s_2(X_{2i}) + s_3(X_{3i}) + \dots$$

- $s_j(\cdot)$ are usually estimated with locally weighted regression smoothers or cubic smoothing splines (but many approaches are possible)

Generalized Additive Models (GAM)

$$\hat{y}_i = \beta_0 + s_1(X_{1i}) + s_2(X_{2i}) + s_3(X_{3i}) + \dots$$

- $s_j(\cdot)$ are usually estimated with locally weighted regression smoothers or cubic smoothing splines (but many approaches are possible)
- Theory and estimation are somewhat involved, but they are easy to use:
  - `gam.out <- gam(y~s(x1)+s(x2)+x3)`
    `plot(gam.out)`
  - Multiple functions but I recommend `mgcv` package

Generalized Additive Models (GAM)

The GAM approach can be extended to allow interactions ($s_{12}(\cdot)$) between explanatory variables, but this eats up degrees of freedom so you need a lot of data.

$$\hat{y}_i = \beta_0 + s_{12}(X_{1i}, X_{2i}) + s_3(X_{3i})$$

Generalized Additive Models (GAM)

The GAM approach can be extended to allow interactions ($s_{12}(\cdot)$) between explanatory variables, but this eats up degrees of freedom so you need a lot of data.

$$\hat{y}_i = \beta_0 + s_{12}(X_{1i}, X_{2i}) + s_3(X_{3i})$$

It can also be used for hybrid models where we model some variables as linear and others with a flexible function:

$$\hat{y}_i = \beta_0 + \beta_1 X_{1i} + s_2(X_{2i}) + s_3(X_{3i})$$

GAM Fit

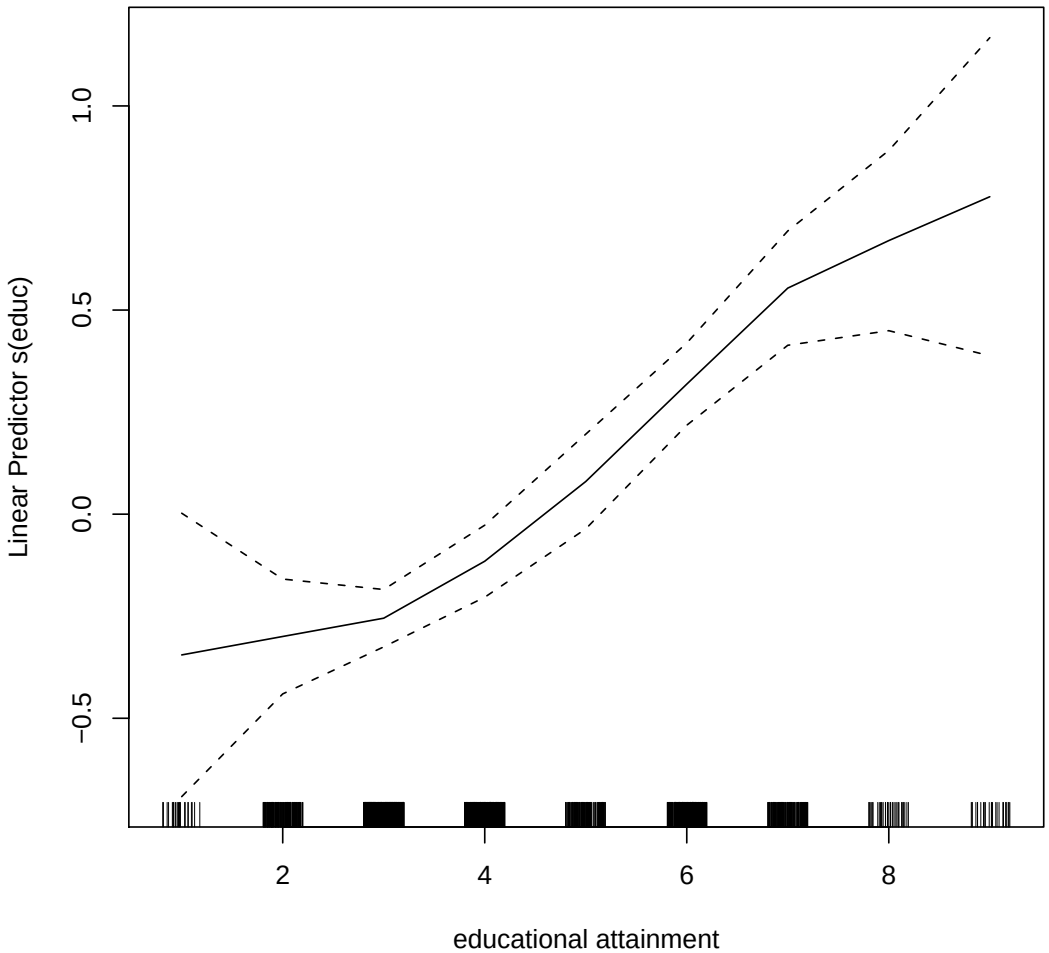


Image Credit: Jens Hainmueller

GAM Fit

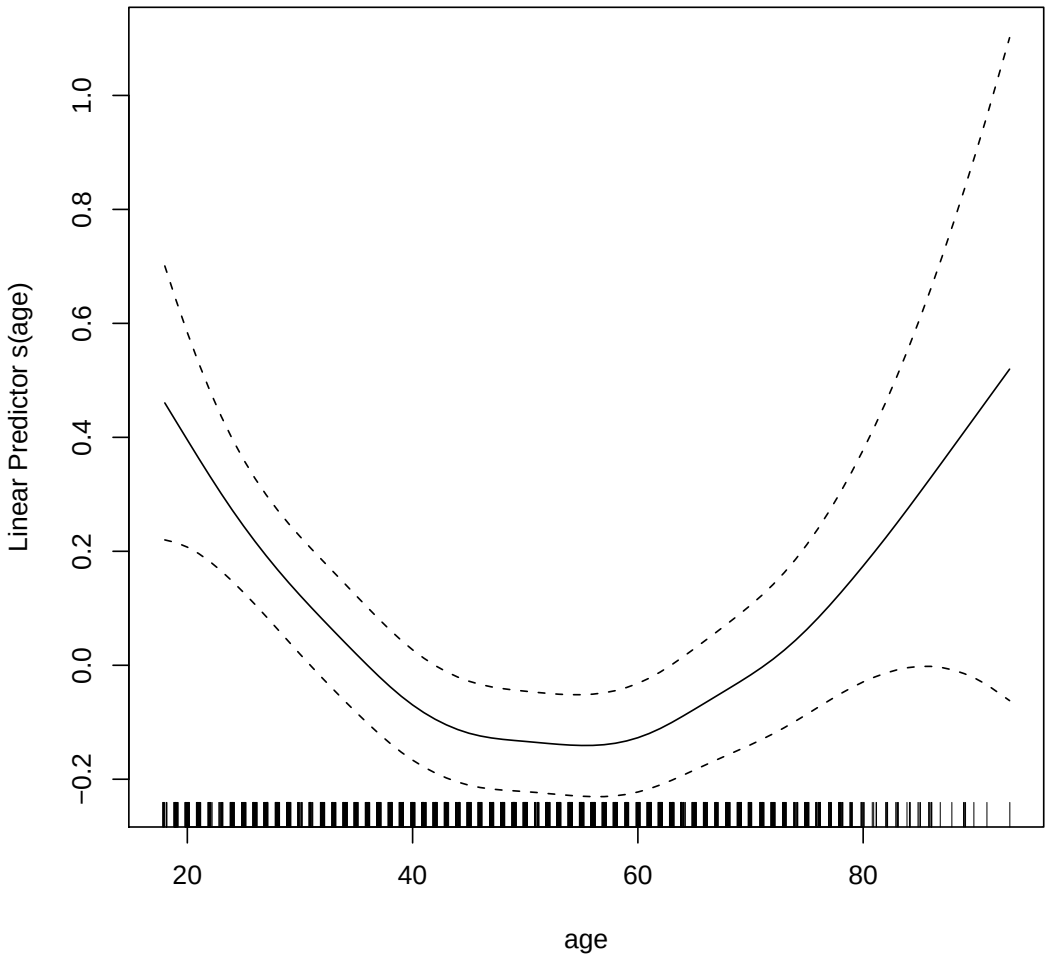


Image Credit: Jens Hainmueller

GAM Fit

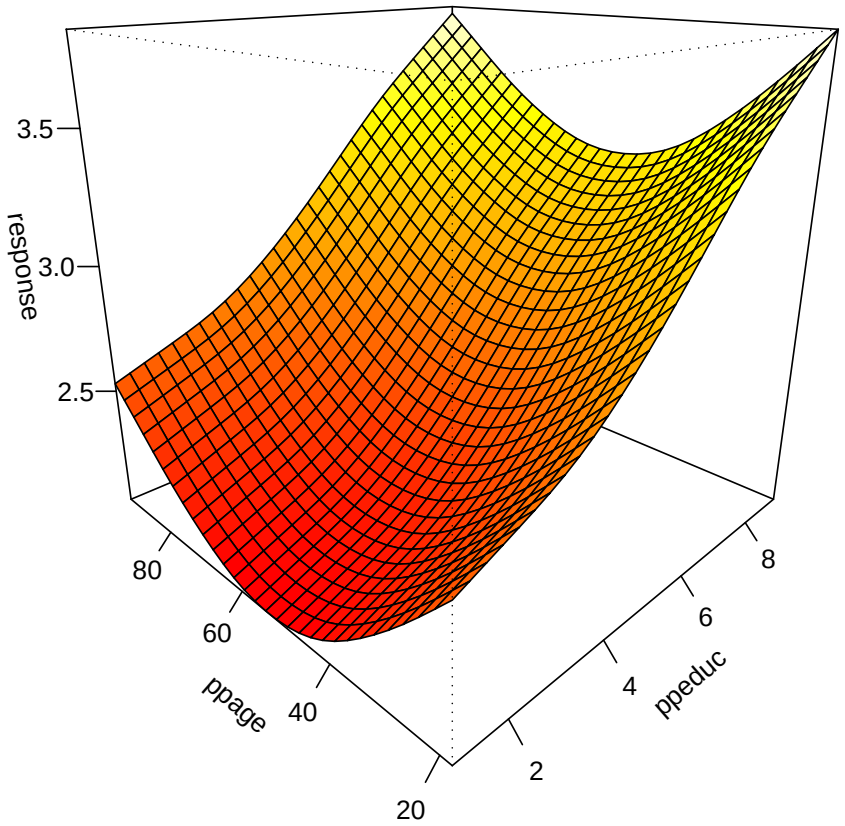


Image Credit: Jens Hainmueller

GAMs using `mgcv`

- There is an avalanche of small, consequential decisions when doing machine learning. Many strategies that seem perfectly feasible on paper work poorly in practice.

GAMs using `mgcv`

- There is an avalanche of small, consequential decisions when doing machine learning. Many strategies that seem perfectly feasible on paper work poorly in practice.
- Simon Wood and colleagues have spent years getting this right for GAMs and implementing in `mgcv`.

GAMs using `mgcv`

- There is an avalanche of small, consequential decisions when doing machine learning. Many strategies that seem perfectly feasible on paper work poorly in practice.
- Simon Wood and colleagues have spent years getting this right for GAMs and implementing in `mgcv`.
- This includes careful choices of tuning parameters, extensions to all kinds of generalized linear models, visualizations, confidence intervals etc.

GAMs using mgcv

- There is an avalanche of small, consequential decisions when doing machine learning. Many strategies that seem perfectly feasible on paper work poorly in practice.
- Simon Wood and colleagues have spent years getting this right for GAMs and implementing in mgcv.
- This includes careful choices of tuning parameters, extensions to all kinds of generalized linear models, visualizations, confidence intervals etc.
- The big limitation is limited interactivity, but in many cases that might be completely fine.

Summary

- GAMs with `mgcv` offer a nice turnkey solution to doing additive smoothing.

Summary

- GAMs with `mgcv` offer a nice turnkey solution to doing additive smoothing.
- Additivity makes it easy to interpret what is going on and `mgcv` comes with nice plotting options.

Summary

- GAMs with `mgcv` offer a nice turnkey solution to doing additive smoothing.
- Additivity makes it easy to interpret what is going on and `mgcv` comes with nice plotting options.
- We have skimmed over many, many details...

Going Deeper:

Wood, Simon (2017) *Generalized Additive Models An Introduction with R, Second Edition*.

1. Preview Three Types of Prediction
2. Prediction for Enhanced Measurement
3. Fragile Families Challenge
4. Logistic Regression
5. Generalized Additive Models
6. Decision Trees

1 Preview Three Types of Prediction

2 Prediction for Enhanced Measurement

3 Fragile Families Challenge

4 Logistic Regression

5 Generalized Additive Models

6 Decision Trees

The Core Idea: Recursive Binary Splitting

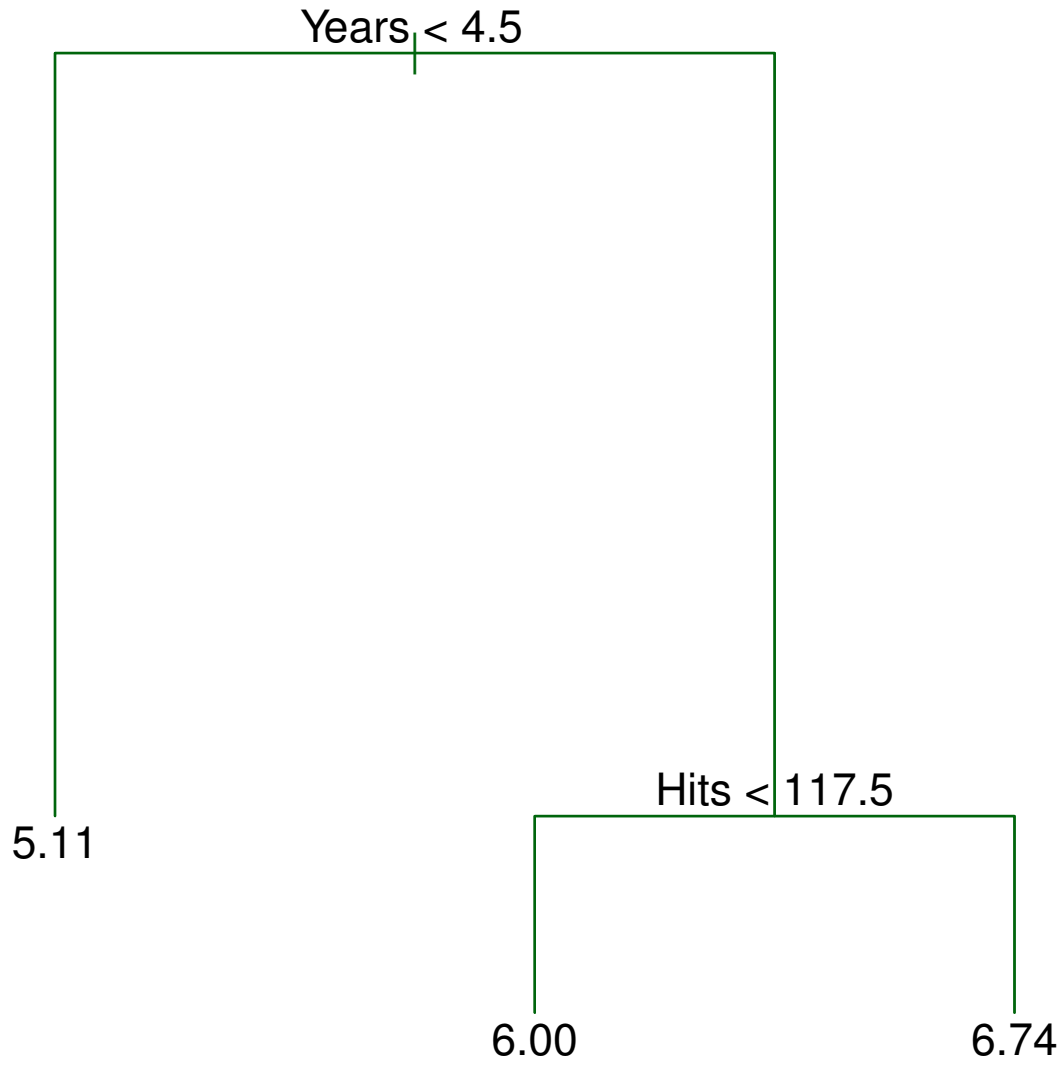


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

The Core Idea: Recursive Binary Splitting

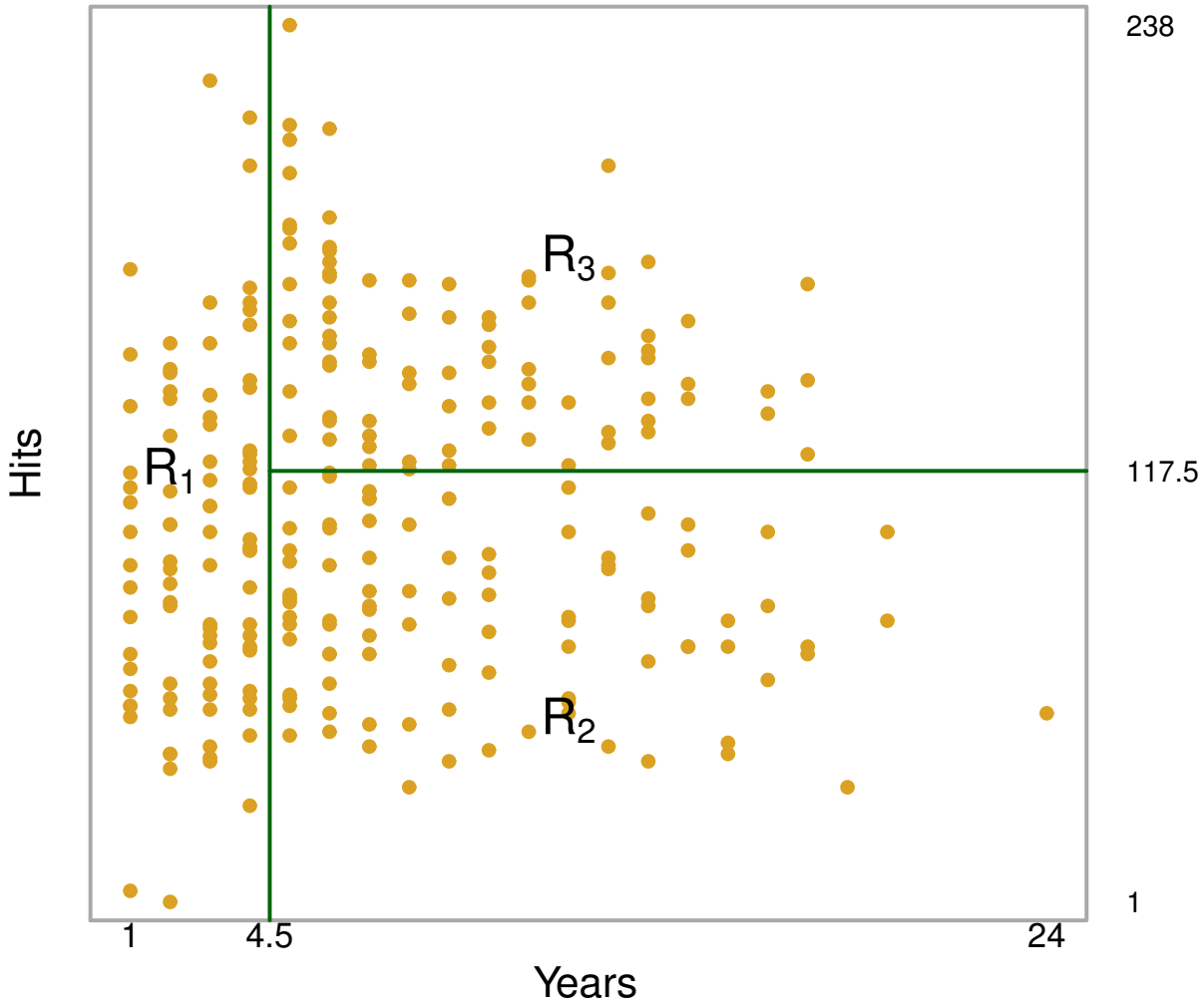


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

Terminology

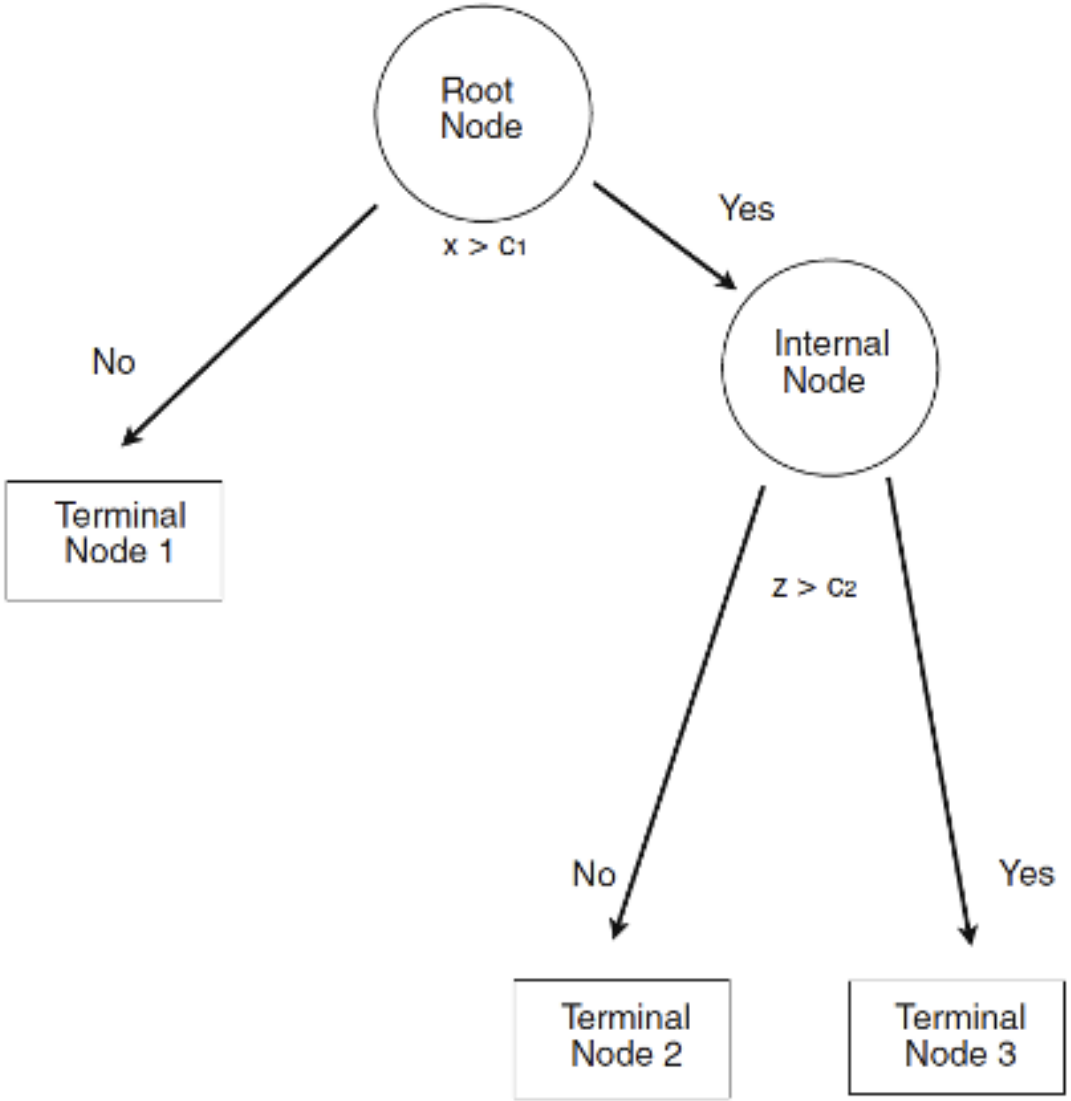


Image Credit: Berk 2020. *Statistical Learning From A Regression Perspective, Third Edition.*

Loss Function and Prediction Function

Loss Function and Prediction Function

For some regions of the data (in practice, rectangles) denoted R_j

Loss Function and Prediction Function

For some regions of the data (in practice, rectangles) denoted R_j

$$\arg\min_{R_j} \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2$$

Loss Function and Prediction Function

For some regions of the data (in practice, rectangles) denoted R_j

$$\arg\min_{R_j} \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2$$

where $\hat{y}_{R_j}$ is the mean of the training data in the region.

Optimization

Optimization

- The optimal solution here is very difficult to find!

Optimization

- The optimal solution here is very difficult to find!
- Instead we use a top-down and greedy approach: recursive binary splitting

Optimization

- The optimal solution here is very difficult to find!
- Instead we use a top-down and greedy approach: recursive binary splitting
- Starting at the top of the tree (the root), we choose the best split at that point.

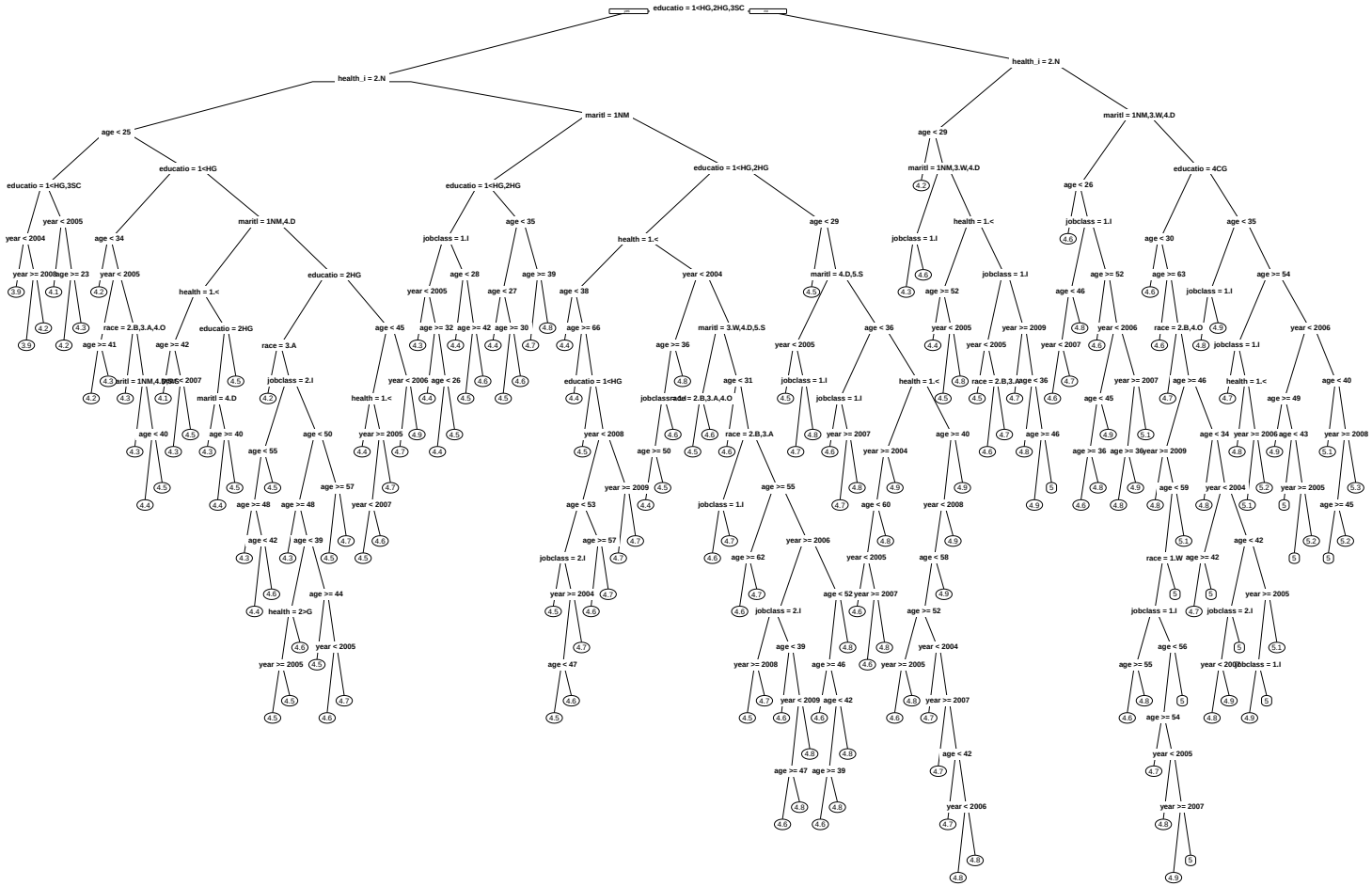
Optimization

- The optimal solution here is very difficult to find!
- Instead we use a top-down and greedy approach: recursive binary splitting
- Starting at the top of the tree (the root), we choose the best split at that point.
- Thus each split is only locally optimal which is the risk of greedy approaches.

Optimization

- The optimal solution here is very difficult to find!
- Instead we use a top-down and greedy approach: recursive binary splitting
- Starting at the top of the tree (the root), we choose the best split at that point.
- Thus each split is only locally optimal which is the risk of greedy approaches.
- We can keep repeating this splitting procedure until every terminal node has some minimum number of observations.

Regression Tree of the Wage Data



Overfitting

- Due to the greedy fitting procedure, trees can have very high-variance and run the risk of overfitting the data.

Overfitting

- Due to the greedy fitting procedure, trees can have very high-variance and run the risk of overfitting the data.
- Many possible ways to control overfitting but the best is grow a big tree and then prune back to a smaller tree.

Overfitting

- Due to the greedy fitting procedure, trees can have very high-variance and run the risk of overfitting the data.
- Many possible ways to control overfitting but the best is grow a big tree and then prune back to a smaller tree.
- A straightforward way to do this is cost complexity pruning.

Overfitting

- Due to the greedy fitting procedure, trees can have very high-variance and run the risk of overfitting the data.
- Many possible ways to control overfitting but the best is grow a big tree and then prune back to a smaller tree.
- A straightforward way to do this is cost complexity pruning.
- For some fixed value of the regularization parameter λ we can identify the subtree such that:

$$\arg\min_{R_m} \sum_{m=1}^{|T|} \sum_{i \in R_m} (y_i - \hat{y}_{R_m})^2 + \lambda |T|$$

where $|T|$ is the number of terminal nodes of the tree.

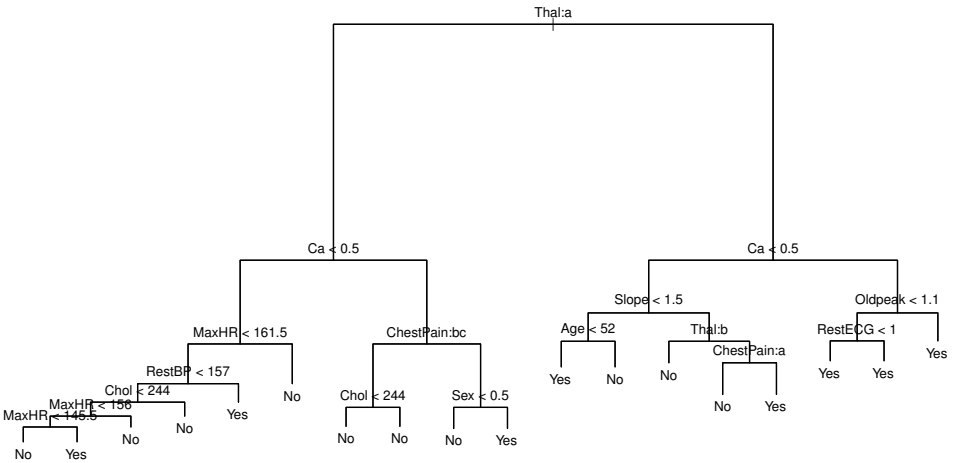
Overfitting

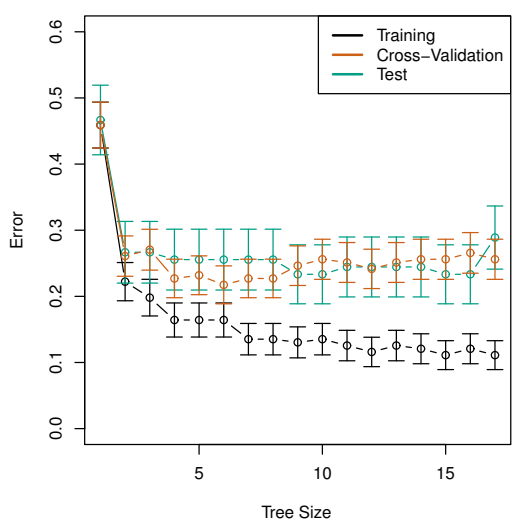
- Due to the greedy fitting procedure, trees can have very high-variance and run the risk of overfitting the data.
- Many possible ways to control overfitting but the best is grow a big tree and then prune back to a smaller tree.
- A straightforward way to do this is cost complexity pruning.
- For some fixed value of the regularization parameter λ we can identify the subtree such that:

$$\arg\min_{R_m} \sum_{m=1}^{|T|} \sum_{i \in R_m} (y_i - \hat{y}_{R_m})^2 + \lambda |T|$$

  where $|T|$ is the number of terminal nodes of the tree.
- λ can be estimated by cross-validation (or with a validation set).

Pruning





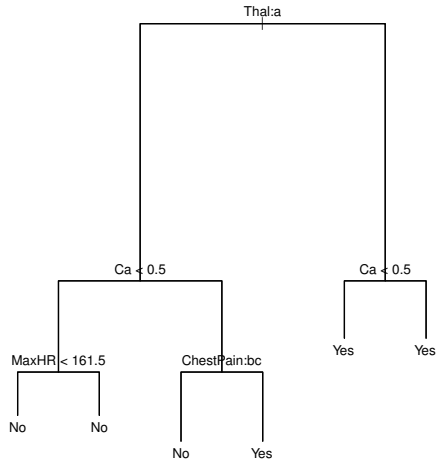
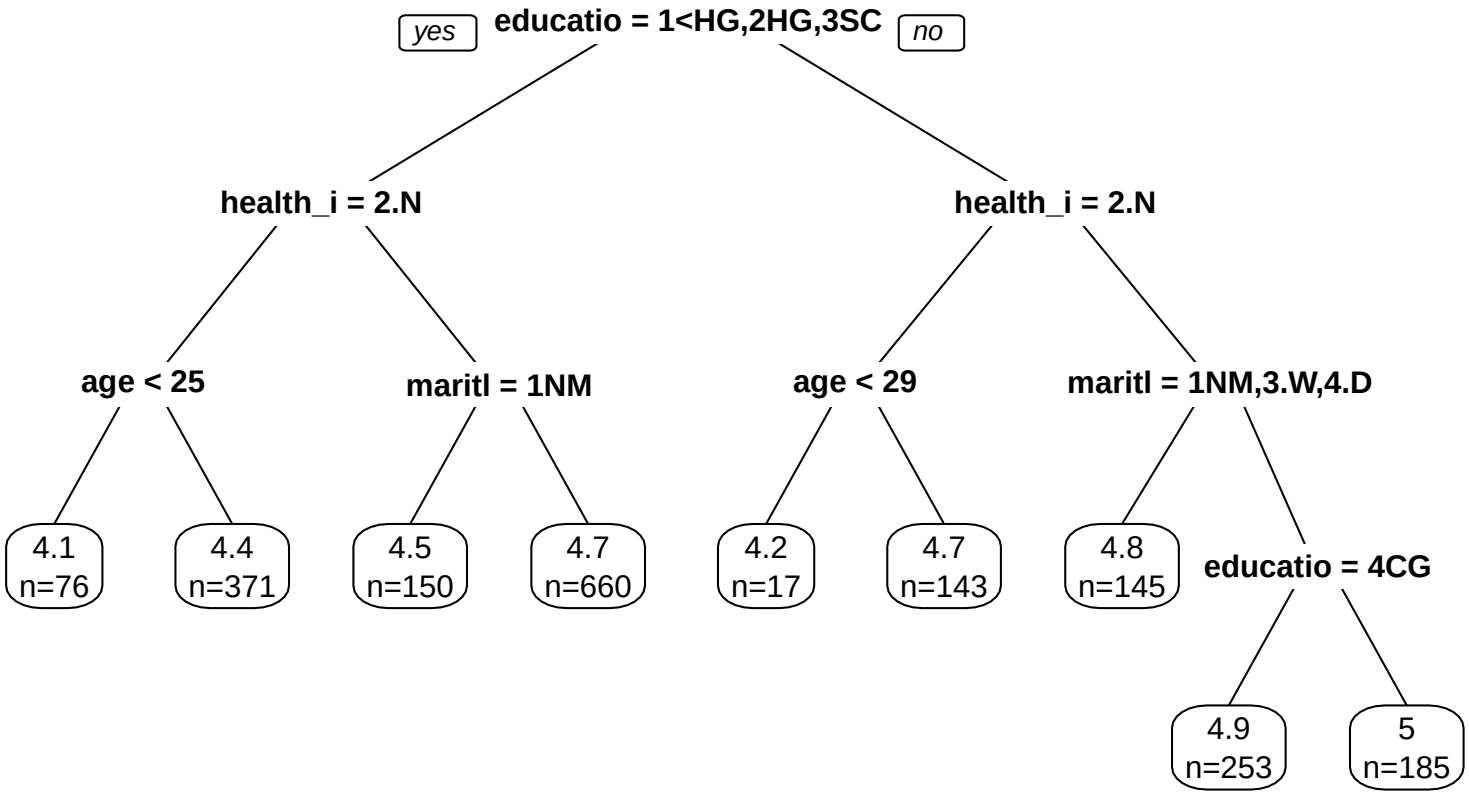


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

Pruned Wage Tree



Revisiting The Loss Function

- In practice the distinction between classification trees and regression trees is the criteria for doing the splitting.

Revisiting The Loss Function

- In practice the distinction between classification trees and regression trees is the criteria for doing the splitting.
- Three major options for K classes

Revisiting The Loss Function

- In practice the distinction between classification trees and regression trees is the criteria for doing the splitting.
- Three major options for K classes
  - Classification Error Rate:
    $1 - \max_k \hat{p}_{mk}$
  - Gini index:
    $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
  - Entropy:
    $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$

Revisiting The Loss Function

- In practice the distinction between classification trees and regression trees is the criteria for doing the splitting.
- Three major options for K classes
  - Classification Error Rate:
    $1 - \max_k \hat{p}_{mk}$
  - Gini index:
    $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
  - Entropy:
    $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$
- Gini Index and Entropy are often used because they encourage node purity—the node contains primarily observations from one class.

Revisiting The Loss Function

- In practice the distinction between classification trees and regression trees is the criteria for doing the splitting.
- Three major options for K classes
  - Classification Error Rate:
    $1 - \max_k \hat{p}_{mk}$
  - Gini index:
    $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
  - Entropy:
    $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$
- Gini Index and Entropy are often used because they encourage node purity—the node contains primarily observations from one class.
- Can in theory use one metric while building and another while pruning.

Implementation

Implementation

- There are many different R packages for building trees!

Implementation

- There are many different R packages for building trees!
- We will use `rpart` because the `rpart.plot` package provides really nice plotting functions!

Implementation

- There are many different R packages for building trees!
- We will use `rpart` because the `rpart.plot` package provides really nice plotting functions!
- Can also print out rules which can be really convenient.

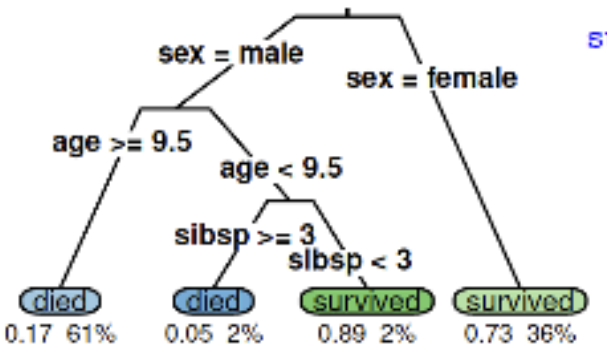
Implementation

- There are many different R packages for building trees!
- We will use `rpart` because the `rpart.plot` package provides really nice plotting functions!
- Can also print out rules which can be really convenient.

Another example is the probability of survival of Titanic passengers:

```
data(ptitanic)
survived <- rpart(survived ~ ., data = ptitanic, cp = .02)
rpart.plot(survived, type = 3, clip.right.labs = FALSE, branch = .3, under = TRUE)
rpart.rules(survived, cover = TRUE)
```

The tree and rules are:




```
survived                                                        cover
    0.05 when sex is   male & age <  9.5 & sibsp >= 3           2%
    0.17 when sex is   male & age >= 9.5                       61%
    0.73 when sex is female                                    36%
    0.89 when sex is   male & age <  9.5 & sibsp <  3           2%
```

Image Credit: Milborrow 2020 Vignette "Plotting rpart trees with the rpart.plot package

Summary

Summary

- Trees are nice because they are:

Summary

- Trees are nice because they are:
  - interpretable especially when pruned

Summary

- Trees are nice because they are:
  - interpretable especially when pruned
  - naturally interactive

Summary

- Trees are nice because they are:
  - interpretable especially when pruned
  - naturally interactive
  - perform automatic feature selection and transformation

Summary

- Trees are nice because they are:
  - interpretable especially when pruned
  - naturally interactive
  - perform automatic feature selection and transformation
  - fast to fit

Summary

- Trees are nice because they are:
  - interpretable especially when pruned
  - naturally interactive
  - perform automatic feature selection and transformation
  - fast to fit
- Can however be highly variable and work best with small number of predictors.

Summary

- Trees are nice because they are:
  - interpretable especially when pruned
  - naturally interactive
  - perform automatic feature selection and transformation
  - fast to fit
- Can however be highly variable and work best with small number of predictors.
- As we will see they are better as inputs to future algorithms than stand-alone algorithms.

Going Deeper:

James, Witten, Hastie and Tibshirani (2021) *An Introduction to Statistical Learning with Applications in R*. Chapters 8.0–8.1
Free online at `www.statlearning.com`

Summary

Going Deeper:

James, Witten, Hastie and Tibshirani (2021) *An Introduction to Statistical Learning with Applications in R*. Chapters 8.0–8.2.2
Free online at `www.statlearning.com`

This Week in Review

This Week in Review

- Prediction within a population
- Prediction for enhanced measurement (Blumenstock et. al)
- Fragile Families Challenge

This Week in Review

- Prediction within a population
- Prediction for enhanced measurement (Blumenstock et. al)
- Fragile Families Challenge

Next week: Prediction in a New Population

- Lazer et al. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science*.
- Teng Ye, Rebecca Johnson, Samantha Fu, Jerica Copeny, Bridgit Donnelly, Alex Freeman, Mirian Lima, Joe Walsh, and Rayid Ghani. 2019. Using machine learning to help vulnerable tenants in New York city. In Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '19).