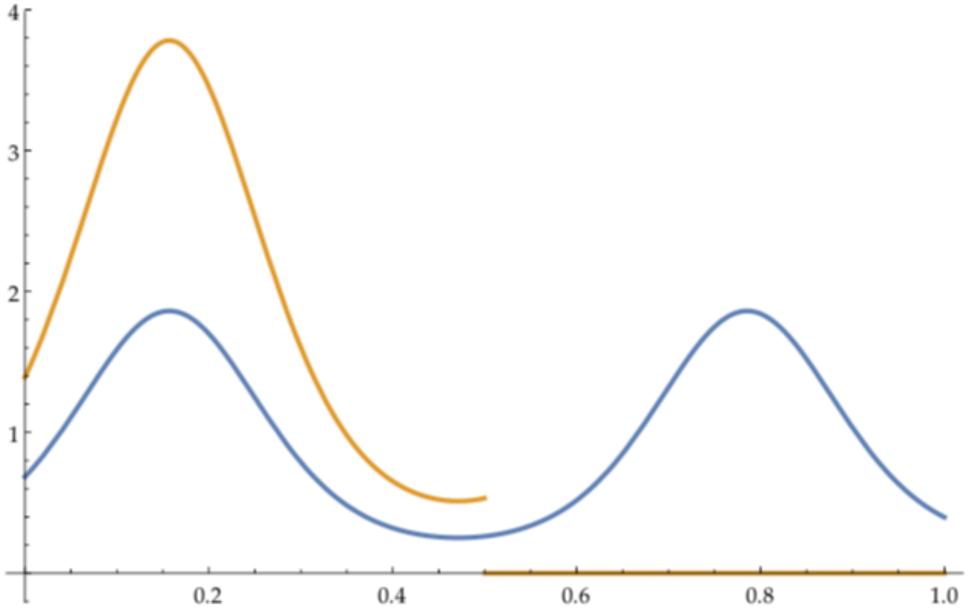
Learning from Biased Data

Constantinos Daskalakis

EECS & CSAIL, MIT

Amuse Bouche



$$f(x) = \frac{e^{\sin 10x}}{\int_0^1 e^{\sin 10x} dx}$$

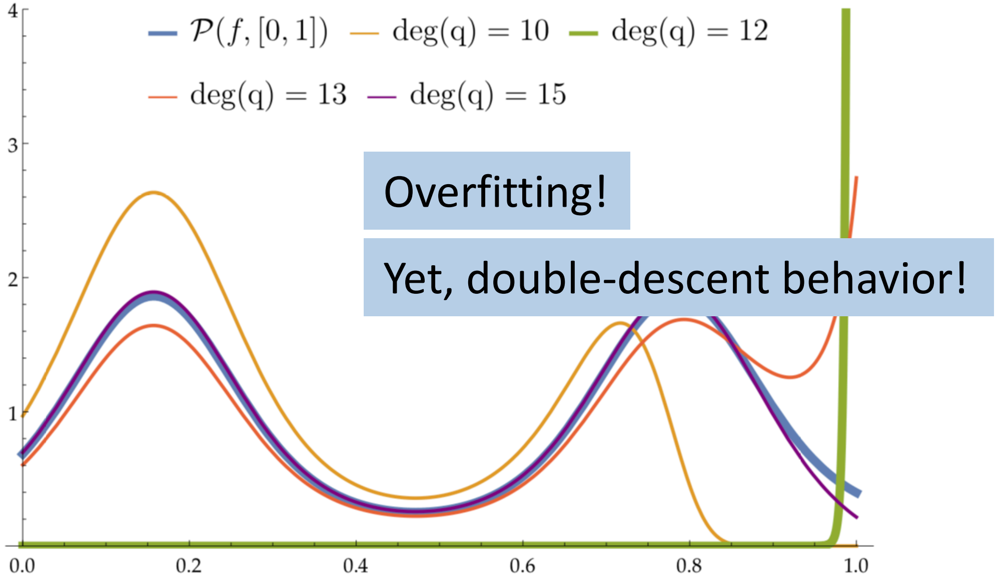
$$g(x) = \frac{e^{\sin 10x}}{\int_0^{1/2} e^{\sin 10x} dx}$$

(conditional of $f(x)$ on $[0,0.5]$)

Experiment: Take large sample $S \subseteq [0,0.5]^N$ from $g(x)$; do MLE to fit most likely density $\frac{e^{q(x)}}{\int_0^{1/2} e^{q(x)} dx}$, where q is some polynomial.

Question: How well does fitted polynomial ***extrapolate***?

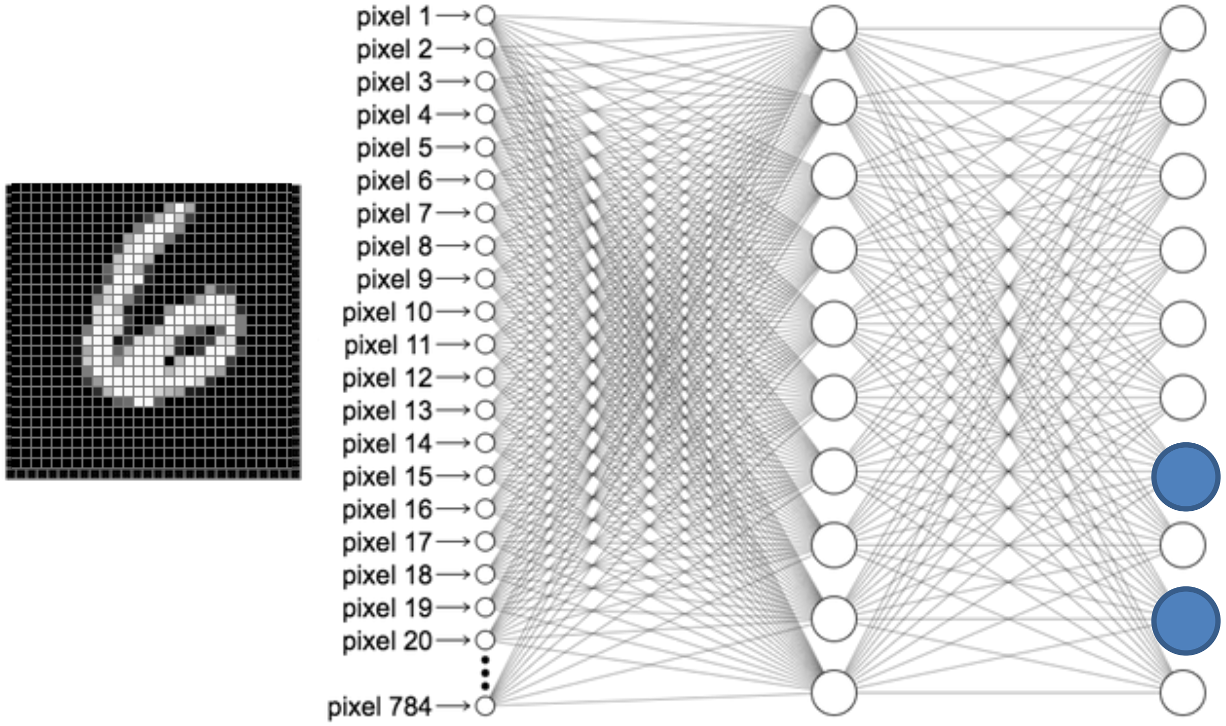
- compare $\frac{e^{q(x)}}{\int_0^1 e^{q(x)} dx}$ to $f(x)$



Overfitting!

Yet, double-descent behavior!

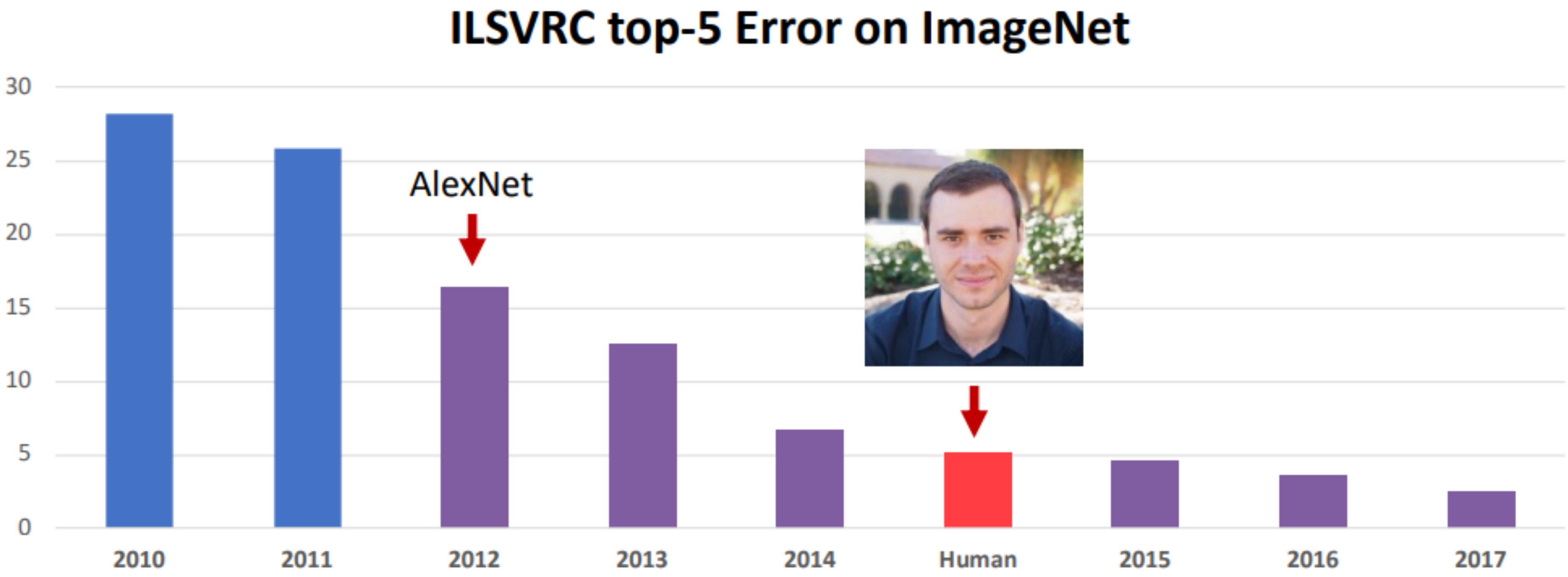
Machine Learning Predictions



Machine Learning Pipeline:

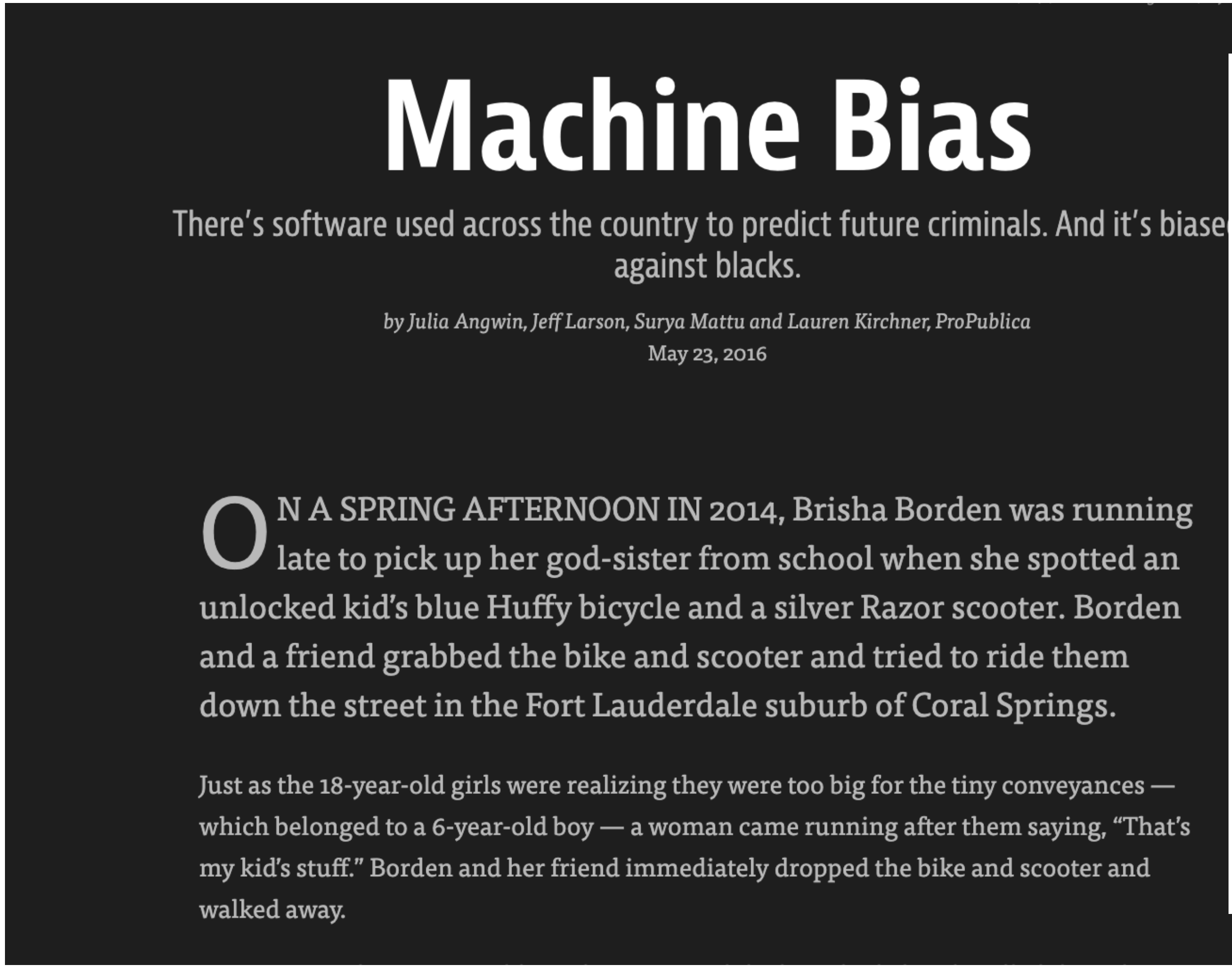
- Collect relevant data; partition into train set and validation set;
- Train/choose hyperparameters
- Deploy

Machine Learning Predictions on Steroids



But what do these results really mean?

Propublica AI Bias Article



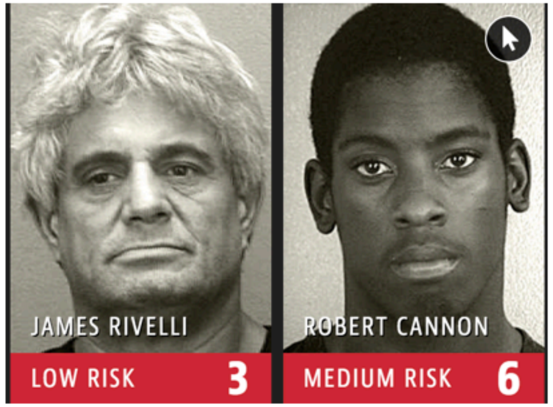
Machine Bias

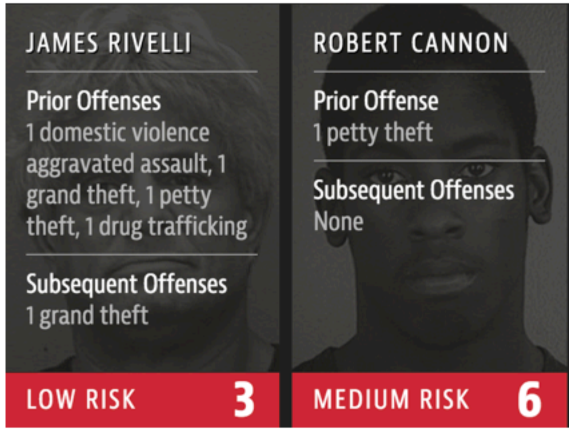
There's software used across the country to predict future criminals. And it's biase against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.





JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

LOW RISK 3

ROBERT CANNON

Prior Offense
1 petty theft

Subsequent Offenses
None

MEDIUM RISK 6

Executive Summary

❑ **Selection bias in data collection**

⇒ **prediction bias (a.k.a. “ML bias”)**

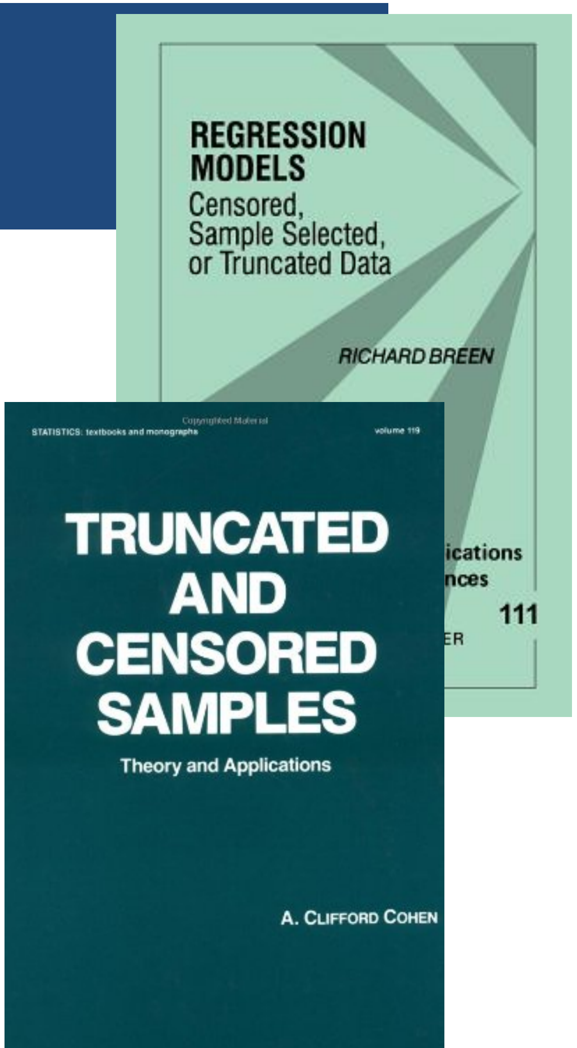
❑ **Goals:** decrease bias, by developing statistical methods robust to **censored and truncated samples**

Truncation: samples falling outside of “observation window” are hidden and their count is also hidden

Censoring: ditto, but count of hidden data is provided

Why Censoring/Truncation?

- ○ limitations of measurement devices
- ○ limitations of data collection
  - ○ experimental design, ethical or privacy considerations,…



REGRESSION MODELS
Censored, Sample Selected, or Truncated Data
RICHARD BREEN
ications
nces
111
ER

STATISTICS: textbooks and monographs
volume 119
TRUNCATED AND CENSORED SAMPLES
Theory and Applications
A. CLIFFORD COHEN

❑ physics

❑ economics

❑ social sciences

❑ clinical studies

Motivating Example: IQ vs Income

Goal: Relationship of IQ to Income for *low-skill workers* **[Wolfle&Smith'56, Hause'71]**

- *"low skill"* = paid under, say, $10/hour

Natural Approach: survey families whose income is less than 1.5 times the poverty line; collect data $(x_i, y_i)_i$ where

- x_i: (IQ, Training, Education,…) of individual i
- y_i: earnings of individual i

Regression: fit some model, e.g. $y = \theta^T x + \varepsilon$

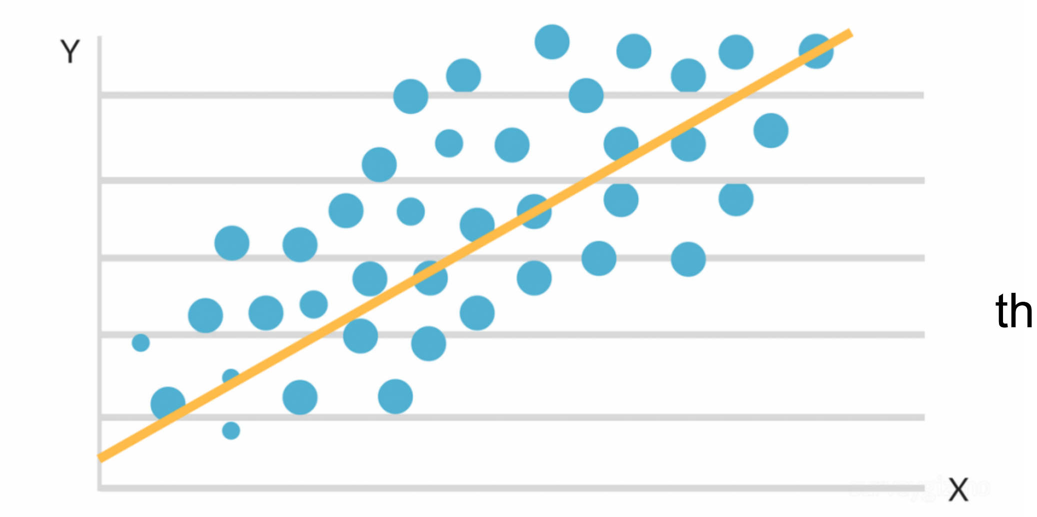
Obvious Issue: **thresholding incomes may introduce bias**

- it does, as shown by **[Hausman-Wise'76]** debunking prior results which had claimed that effects of education are strong, while of IQ are not

What Goes Wrong in Presence of Truncation?

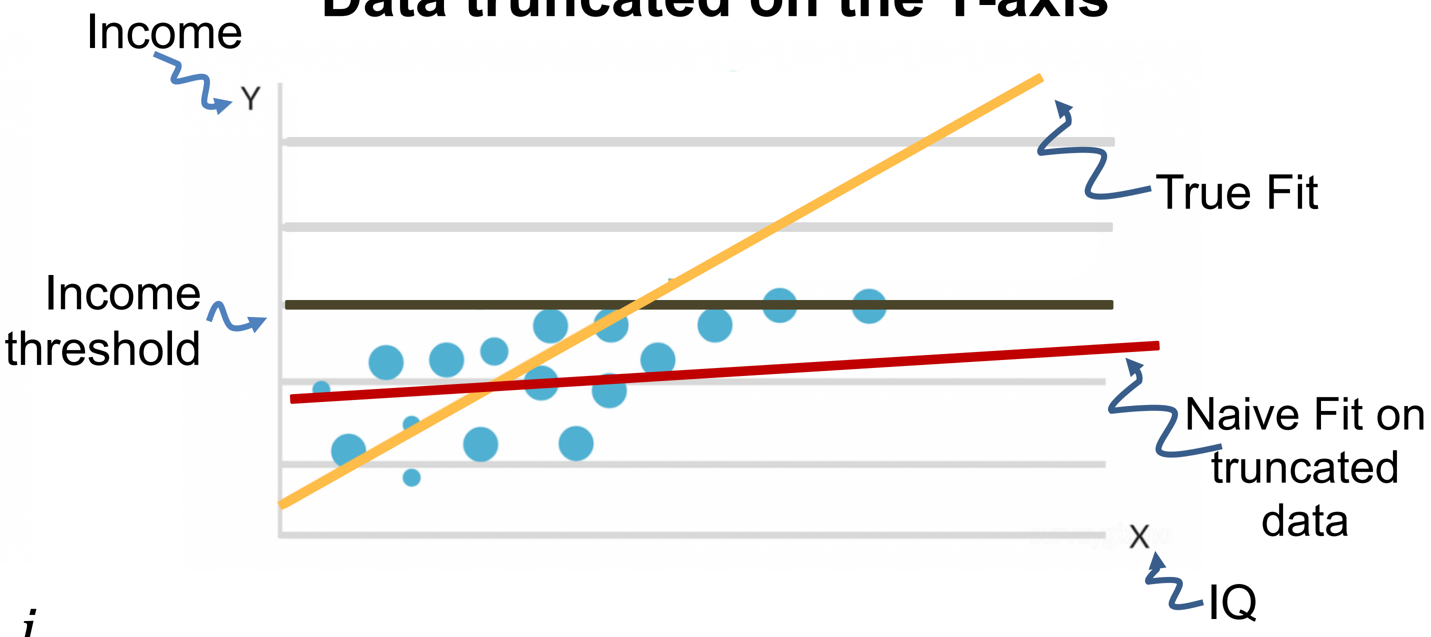
Mental Picture:

Vanilla Linear Regression



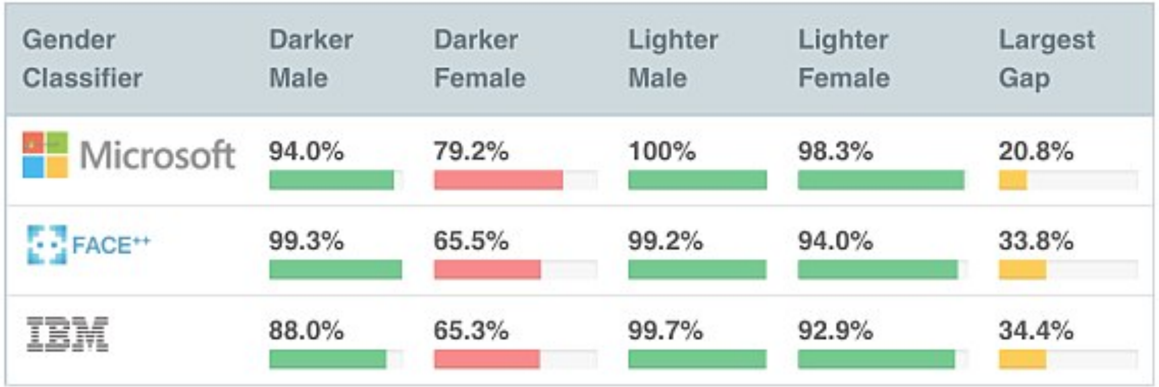
Assumed truth: $y_i = \theta \cdot x_i + \varepsilon_i$, for all i

Data truncated on the Y-axis

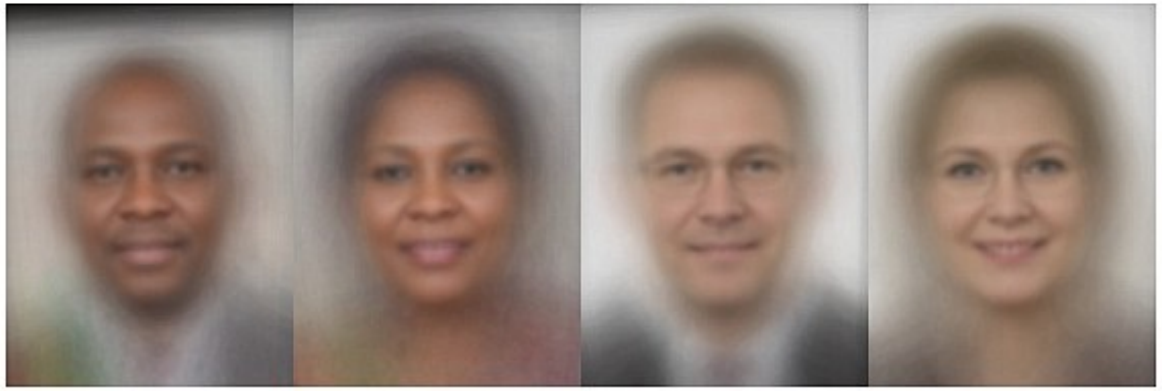


Supervised learning, with y-truncated data → Biased Models

Motivating Example 2: Gender Classification



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



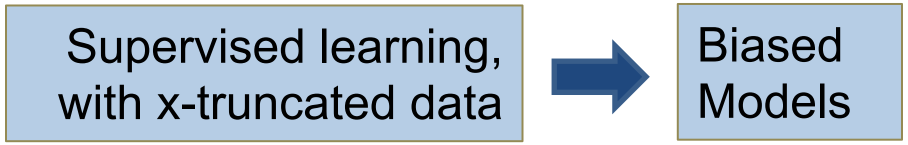
© MIT Media Lab

[Buolamwini, Gebru, FAT 2018]

Explanation: Training data contains more faces that are of lighter skin tone, male gender, Caucasian

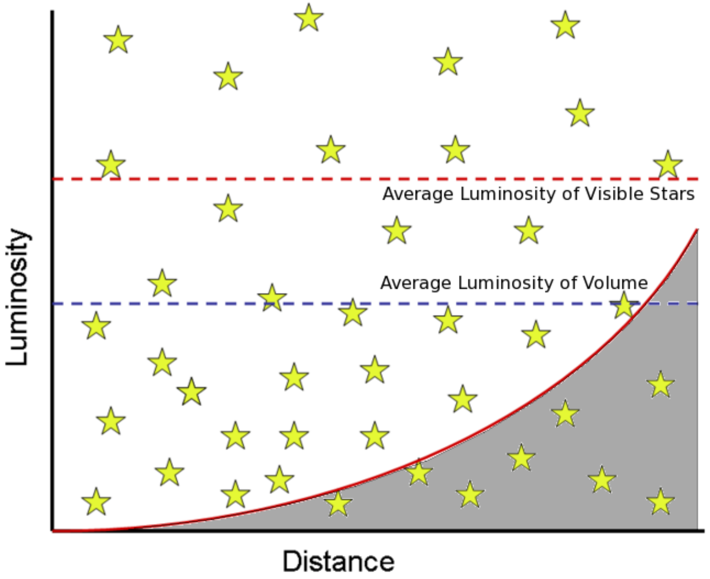
⇒ Training loss of gender classifier pays less attention to faces that are of darker skin tone, female gender, non-Caucasian

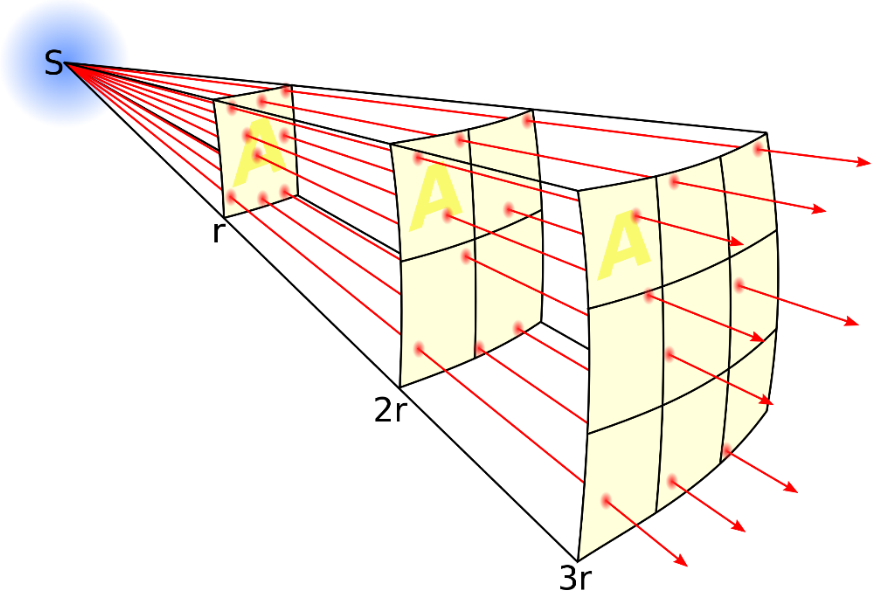
⇒ Test loss on faces that are of darker skin tone, female gender, non-Caucasian is worse



Supervised learning, with x-truncated data → Biased Models

Motivating Example 3: Malmquist Bias



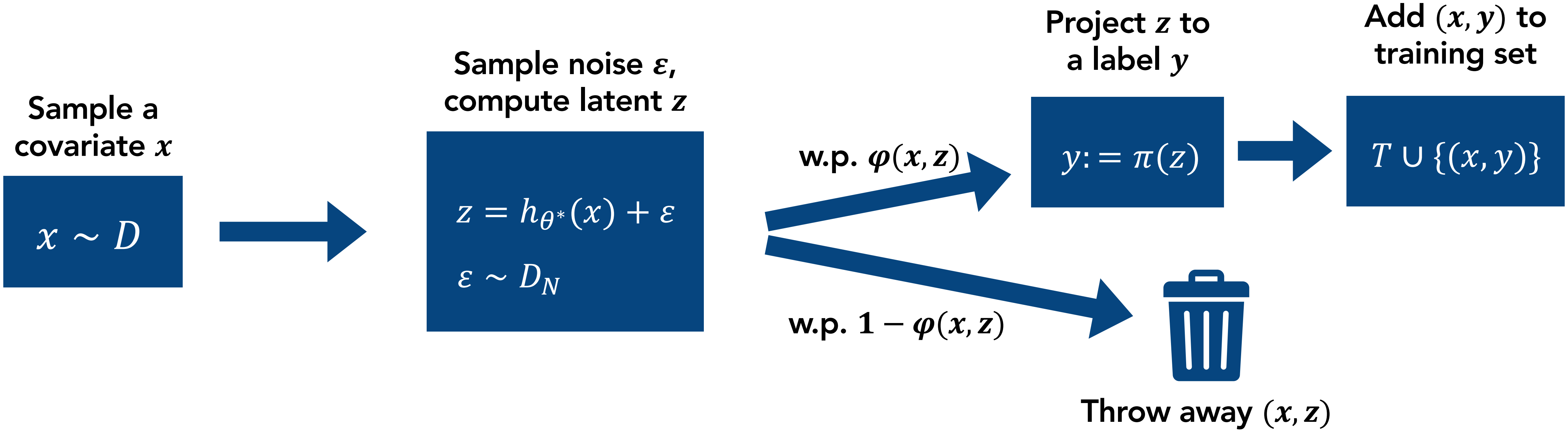


Since light dims with distance, brightness limited surveys of the sky suffer from the cut-off of fainter objects at larger distances

⇒ false trend of increasing intrinsic brightness, and other related quantities, with distance

Unsupervised learning, with truncated data → Biased Models

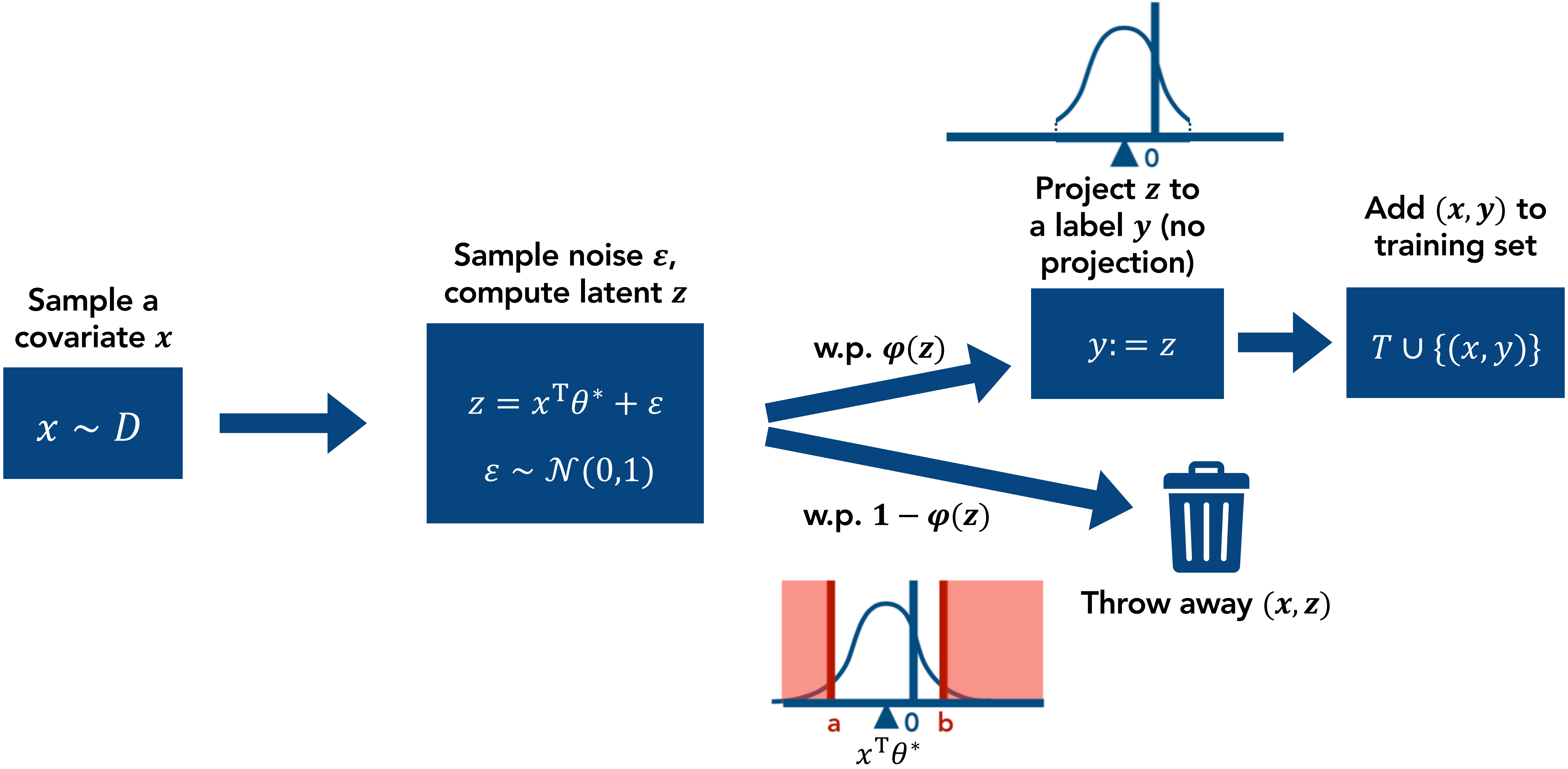
Truncated Regression/Classification Framework



Challenge: Estimate θ^* using training set T produced as above

(φ is either known or from parametric family)

e.g. Truncated Linear Regression



Sample a covariate x

$x \sim D$

Sample noise ε, compute latent z

$z = x^{\mathrm{T}}\theta^* + \varepsilon$

$\varepsilon \sim \mathcal{N}(0,1)$

w.p. $\varphi(z)$

Project z to a label y (no projection)

$y := z$

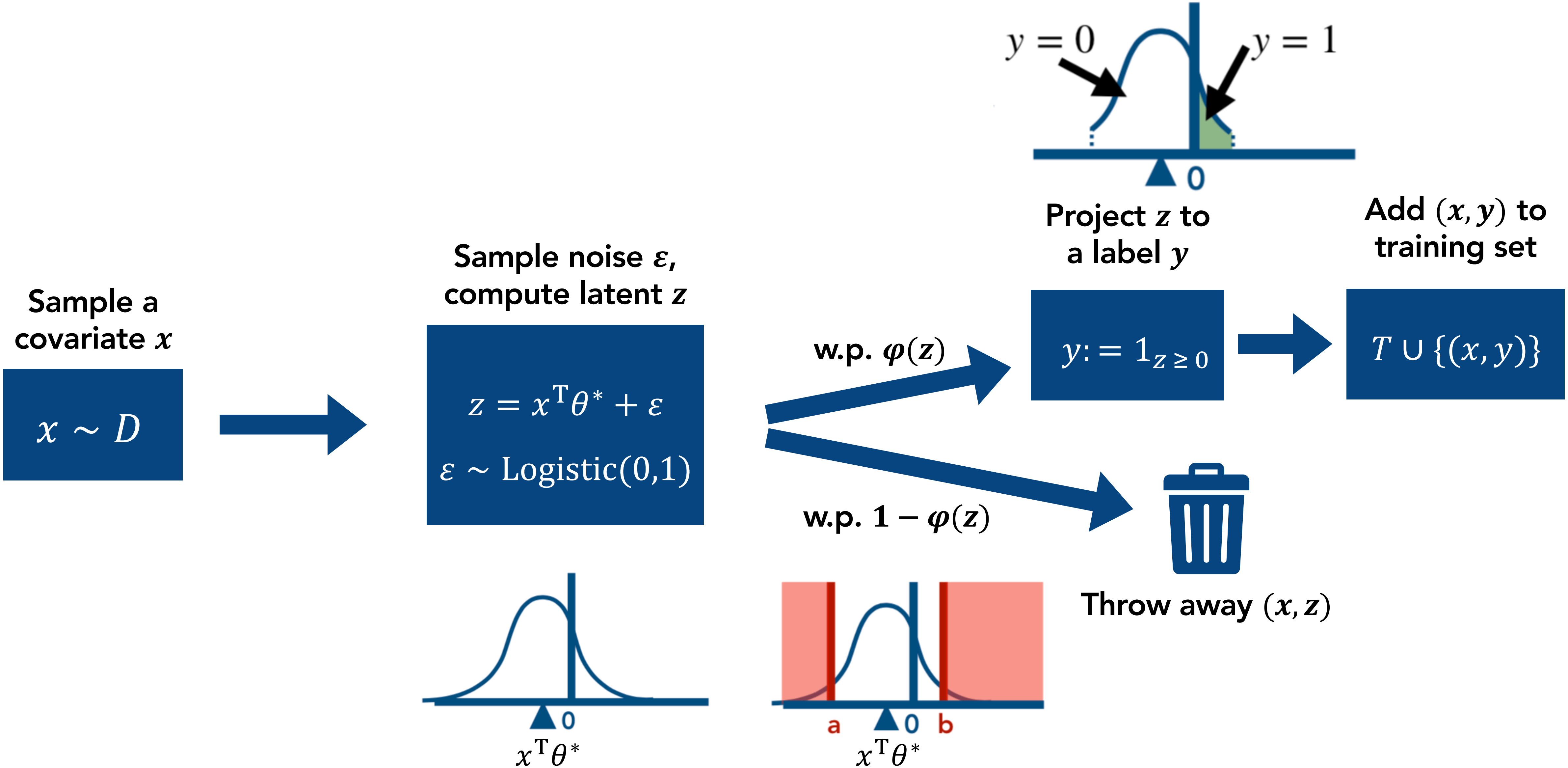
Add (x, y) to training set

$T \cup \{(x, y)\}$

w.p. $1 - \varphi(z)$

Throw away (x, z)

e.g. Truncated Logistic Regression



Sample a covariate x

$x \sim D$

Sample noise ε, compute latent z

$z = x^{\mathrm{T}}\theta^* + \varepsilon$

$\varepsilon \sim \mathrm{Logistic}(0,1)$

w.p. $\varphi(z)$

Project z to a label y

$y := 1_{z \geq 0}$

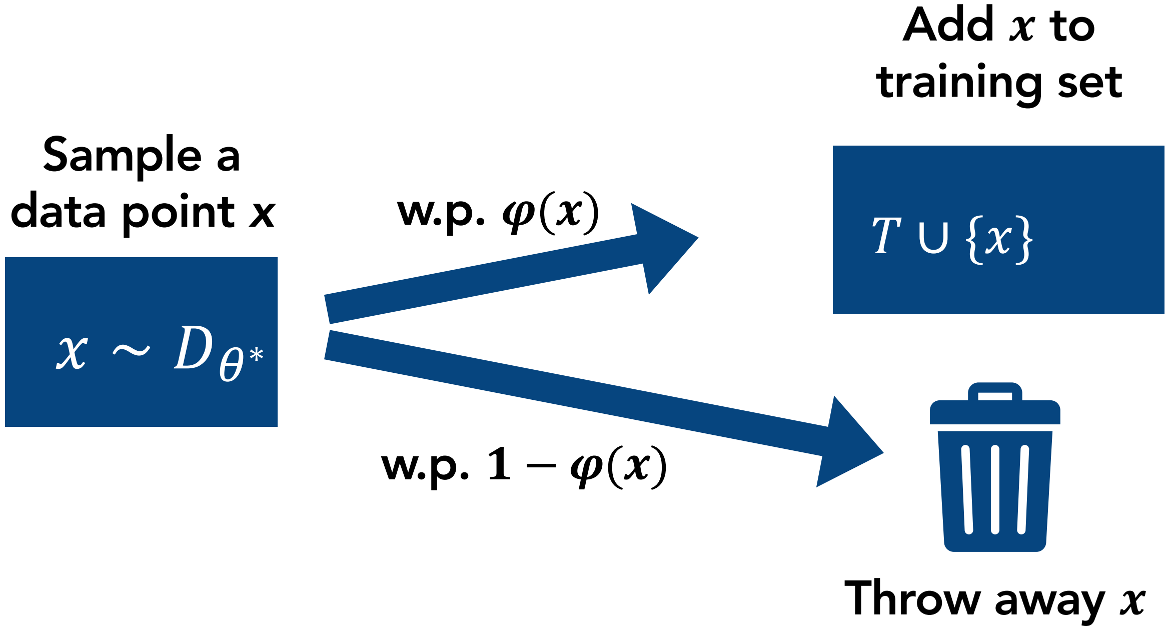
Add (x, y) to training set

$T \cup \{(x, y)\}$

w.p. $1 - \varphi(z)$

Throw away (x, z)

Truncated Density Estimation Framework



Challenge: Estimate θ^* using training set T produced as above

(φ is either known or from parametric family)

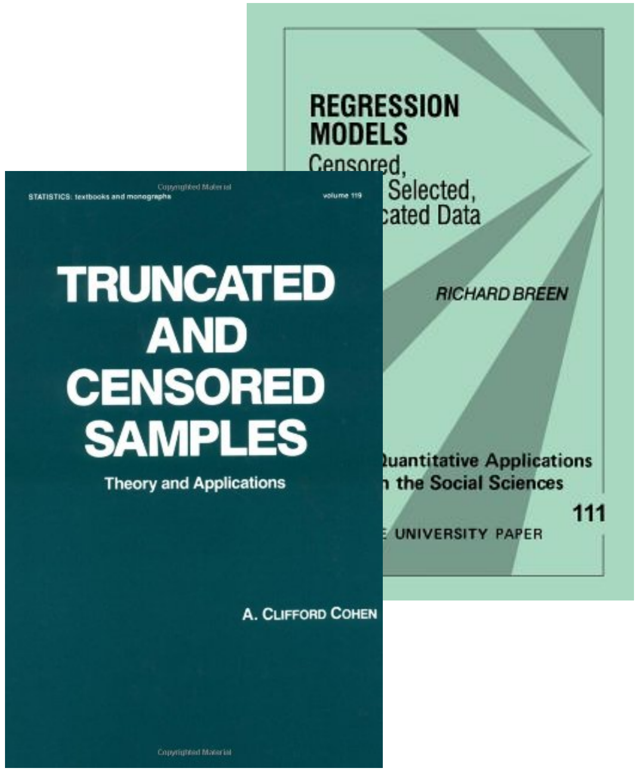
Censored/Truncated Statistics

How to train unbiased models from censored/truncated samples?

- Studied in Statistics/Econometrics since at least **[Bernoulli 1760]**

  [Galton 1897], [Pearson 1902], [Pearson, Lee 1908], [Lee 1914], [Fisher 1931], [Hotelling 1948, [Tukey 1949], [Tobin 1958], [Amemiya 1973], [Hausman, Wise 1976], [Breen 1996], [Hajivassiliou-McFadden'97], [Balakrishnan, Cramer 2014], Limited Dependent Variables models, Method of Simulated Scores, GHK Algorithm

- Intimately related to domain adaptation in Machine Learning



REGRESSION MODELS

Censored, Selected, ...cated Data

RICHARD BREEN

...uantitative Applications ...n the Social Sciences

111

...E UNIVERSITY PAPER

STATISTICS: textbooks and monographs

volume 119

TRUNCATED AND CENSORED SAMPLES

Theory and Applications

A. CLIFFORD COHEN

Challenges:

#parameters/dimension

- Error rates: $\frac{\mathrm{Bad}(d)}{\sqrt{n}}$

#biased samples

- Computationally inefficient algorithms

Recent work **[w/ Gouleakis, Ilyas, Kontonis, Rohatgi, Tzamos, Zampetakis in FOCS'18, COLT'19, AISTATS'20, in progress]**

- Computationally and Statistically efficient algorithms; arbitrary truncation sets
- truncated linear/logistic/probit regression, compressed sensing, (non-parametric) density estimation
  - e.g. rates for linear regression $O\left(\sqrt{d/n}\right)$
  - e.g. rates for compressed sensing $O\left(\sqrt{k\log d\,/n}\right)$

Censored/Truncated Statistics

How to train unbiased models from censored/truncated samples?

- Studied in Statistics/Econometrics since at least [Bernoulli 1760]

  [Galton 1897], [Pearson 1902], [Pearson, Lee 1908], [Lee 1914], [Fisher 1931], [Hotelling 1948, [Tukey 1949], [Tobin 1958], [Amemiya 1973], [Hausman, Wise 1976], [Breen 1996], [Hajivassilio

  Limited Dependent Variables m

- Intimately related to doma

Challenges:

#

- Error rates: $\frac{\text{Bad}(d)}{\sqrt{n}}$

#

- Computationally inefficient

Recent work [w/ Gouleakis, Ilyas, progress]

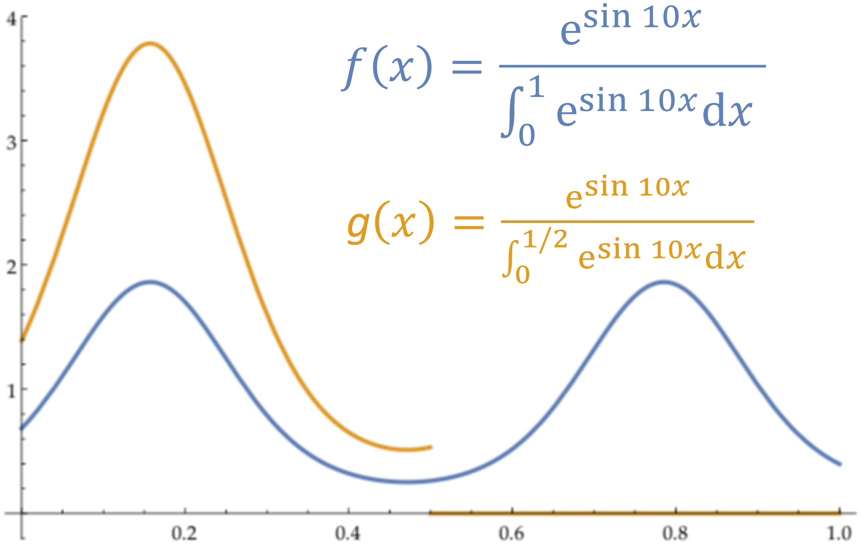
- Computationally and Statistica
- truncated linear/logistic/probit regression, compressed sensing, (non-parametric) density estimation
  - e.g. rates for linear regression $O\left(\sqrt{d/n}\right)$
  - e.g. rates for compressed sensing $O\left(\sqrt{k\log d\,/n}\right)$

Why now?

- ***Mathematics***: concentration/anti-concentration of measure [Carbery-Wright'01]
- ***Machine Learning/Optimization***: stochastic gradient descent
- ***Hardware***: gradient descent based algorithms exportable to Deep Neural Network models

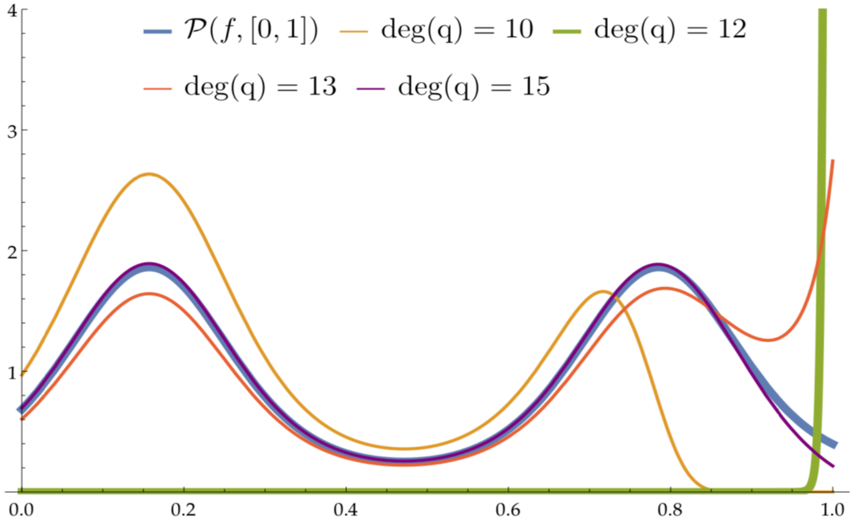
When Does Extrapolation Work?

(an impressionistic picture)



$$f(x) = \frac{e^{\sin 10x}}{\int_0^1 e^{\sin 10x} dx}$$

$$g(x) = \frac{e^{\sin 10x}}{\int_0^{1/2} e^{\sin 10x} dx}$$



Experiment: Take large sample $S \subseteq [0{,}0.5]^N$ from $g(x)$; do MLE to fit most likely density $\frac{e^{q(x)}}{\int_0^{1/2} e^{q(x)} dx}$, where q is some polynomial.

Question: How well does fitted polynomial ***extrapolate***?

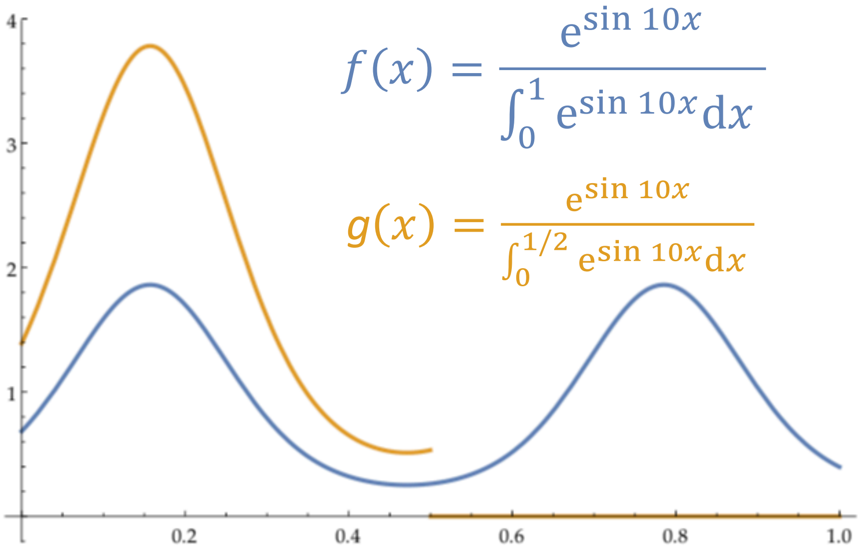
- compare $\frac{e^{q(x)}}{\int_0^1 e^{q(x)} dx}$ to $f(x)$

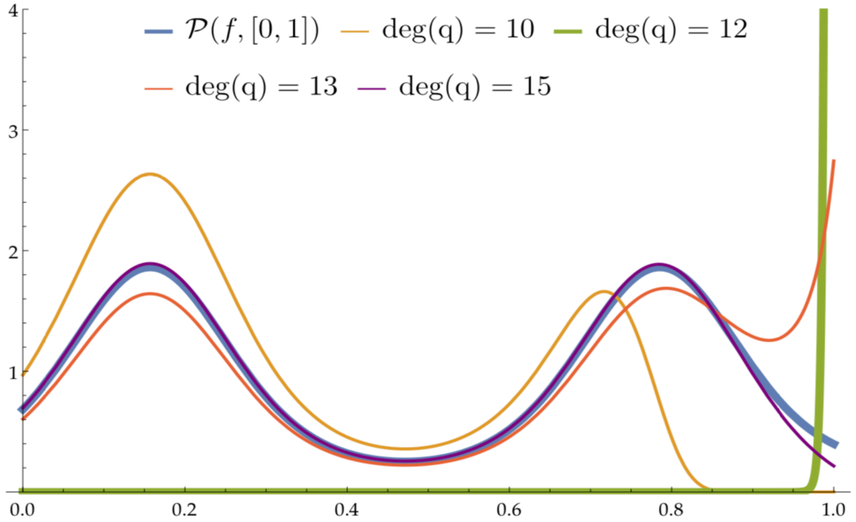
Overfitting!

Yet, double-descent behavior!

When Does Extrapolation Work?

(an impressionistic picture)





Theorem: Suppose P, Q are distributions over $[0,1]^d$, whose log-densities are polynomials of degree k. Suppose $S \subseteq [0,1]^d$ has $vol(S) \geq \alpha$. Then:

$$\left(\frac{d}{\alpha}\right)^{-O(k)} \leq \frac{TV(P,Q)}{TV(P_S,Q_S)} \leq \left(\frac{d}{\alpha}\right)^{O(k)}$$

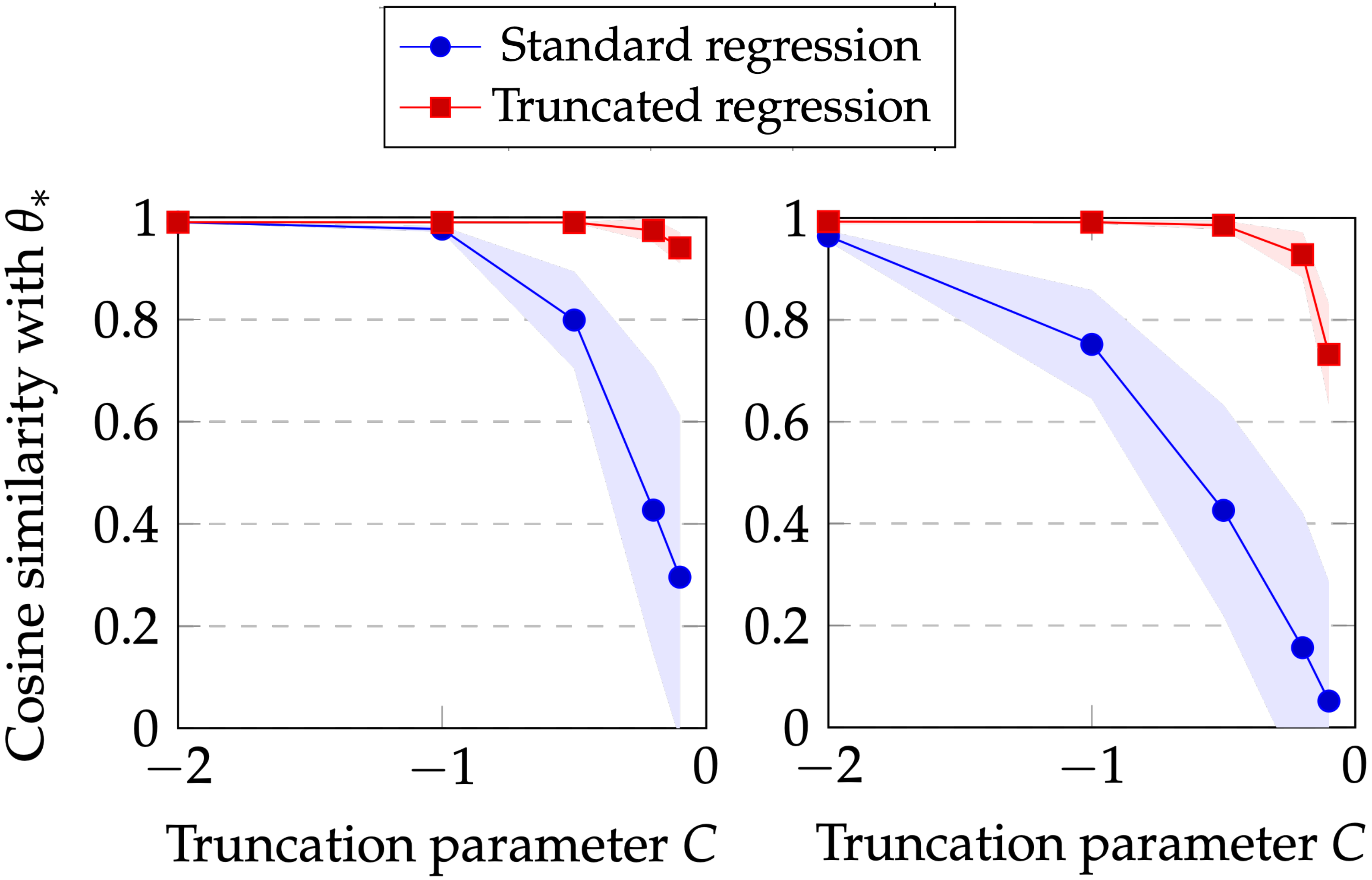
Implication: If P, Q are far in their whole domain, their conditionals can't appear too close.

Experiment: Logistic and Probit Regression

Synthetic data

Setup:

- $\theta^* \sim \mathcal{U}([-1,1]^{10})$
- $X_1, \ldots, X_n \sim \mathcal{U}([0,1]^{10})$
- $Z_i := \theta_*^\top X_i + \varepsilon_i$
- $\varepsilon_i \sim \mathcal{N}(0,1)$/Logistic(0,1)
- Truncation: $\varphi(\cdot) = 1_{[C,\infty)}$
- Projection: $Y_i = \mathbf{1}_{Z_i \geq 0}$
  - when $C = 0$ only see positive examples

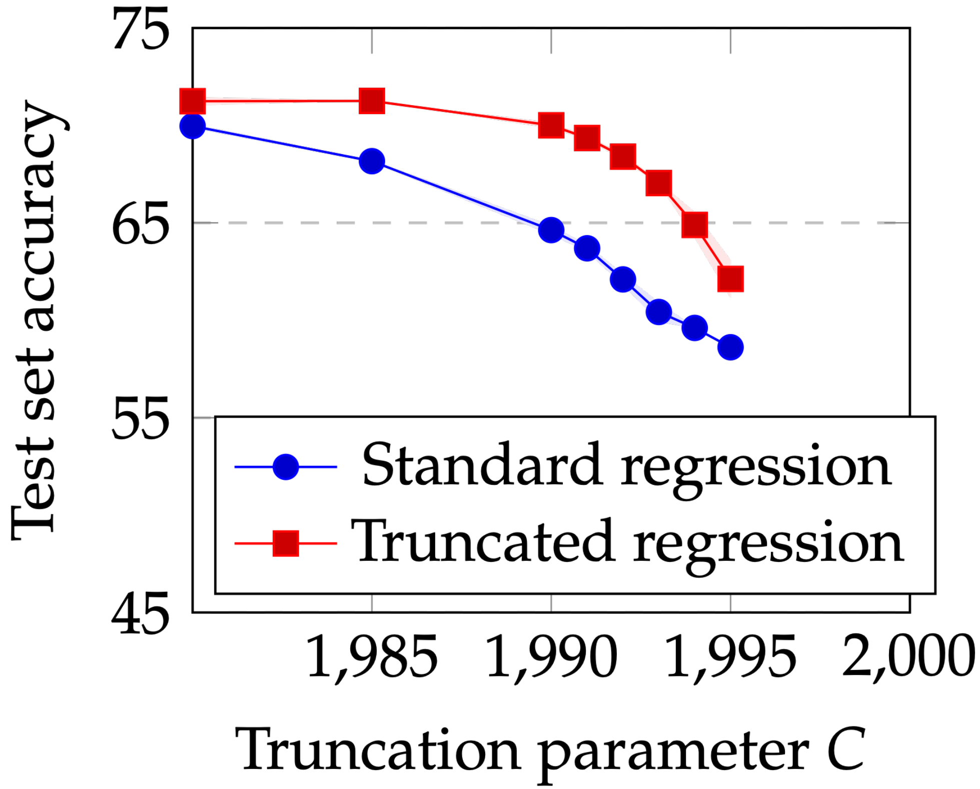


Experiment 2: Logistic Regression

UCI MSD dataset

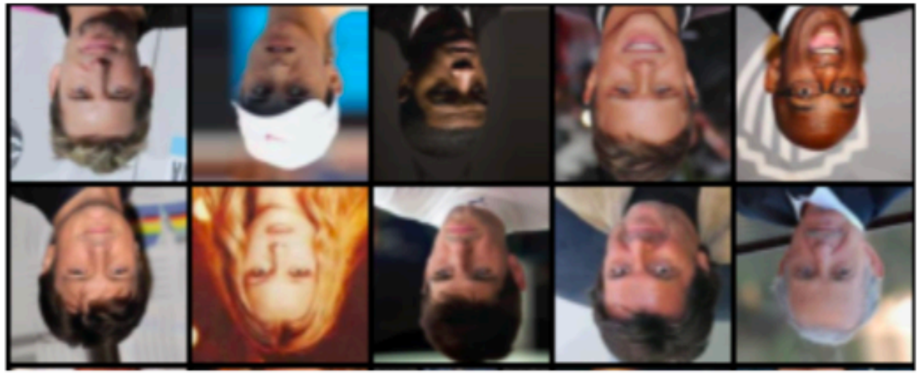
Setup:

- X: song attributes
- Z: year recorded
- Truncation $[C, \infty)$
- Y: recorded before ’96?



Experiment 3: Extreme Domain Adaptation

Train Set



Test Set

Metaphor of settings where support of test set distribution is measure 0 on support of train set distribution

Test Error of Naïve AlexNet Gender Classifier: 55%

Improvement using truncated Statistics: 80%

Conclusions

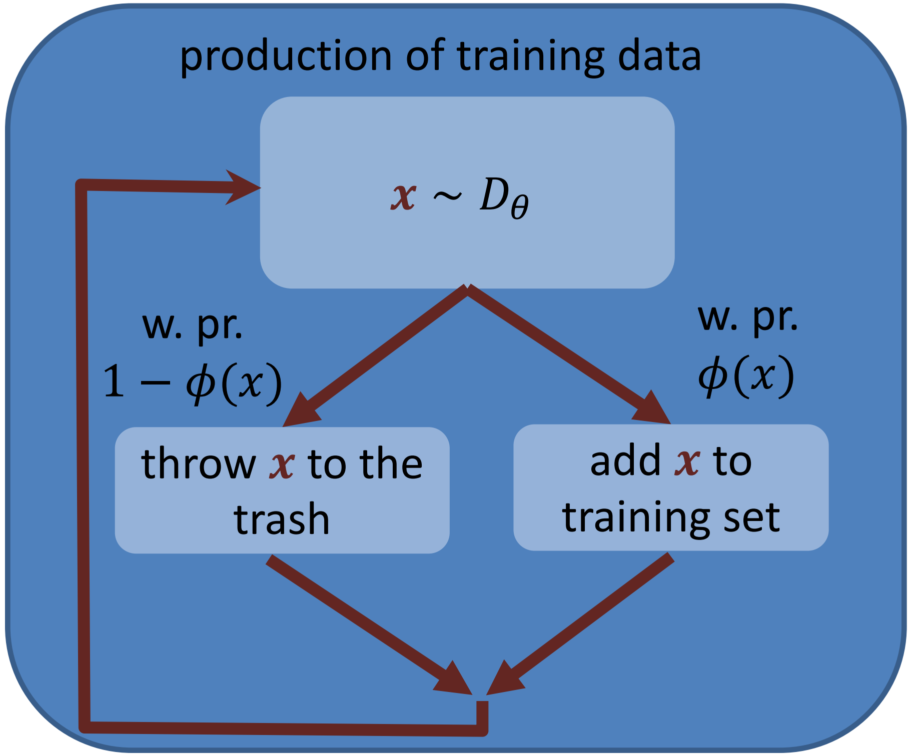
- ❑ **Missing Observations ⇒ prediction bias (a.k.a. “AI bias”)**
- ❑ **Our Work:** decrease bias, by developing statistical methods more robust to **censored and truncated samples**
- ❑ **General Framework:** SGD on Population Log-Likelihood (applies to DNNs)
- ❑ **End-to-end guarantees:** statistical rates and efficient algorithms for several classical problems in Statistics: linear/probit/logistic regression, compressed sensing, non-parametric density estimation
- ❑ **Future work:** push further on reducing parametric assumptions

Thank you!

- Skipped Slides

Censored/Truncated Statistics

Truncated Density Estimation



Goal: Estimate θ using truncated training set