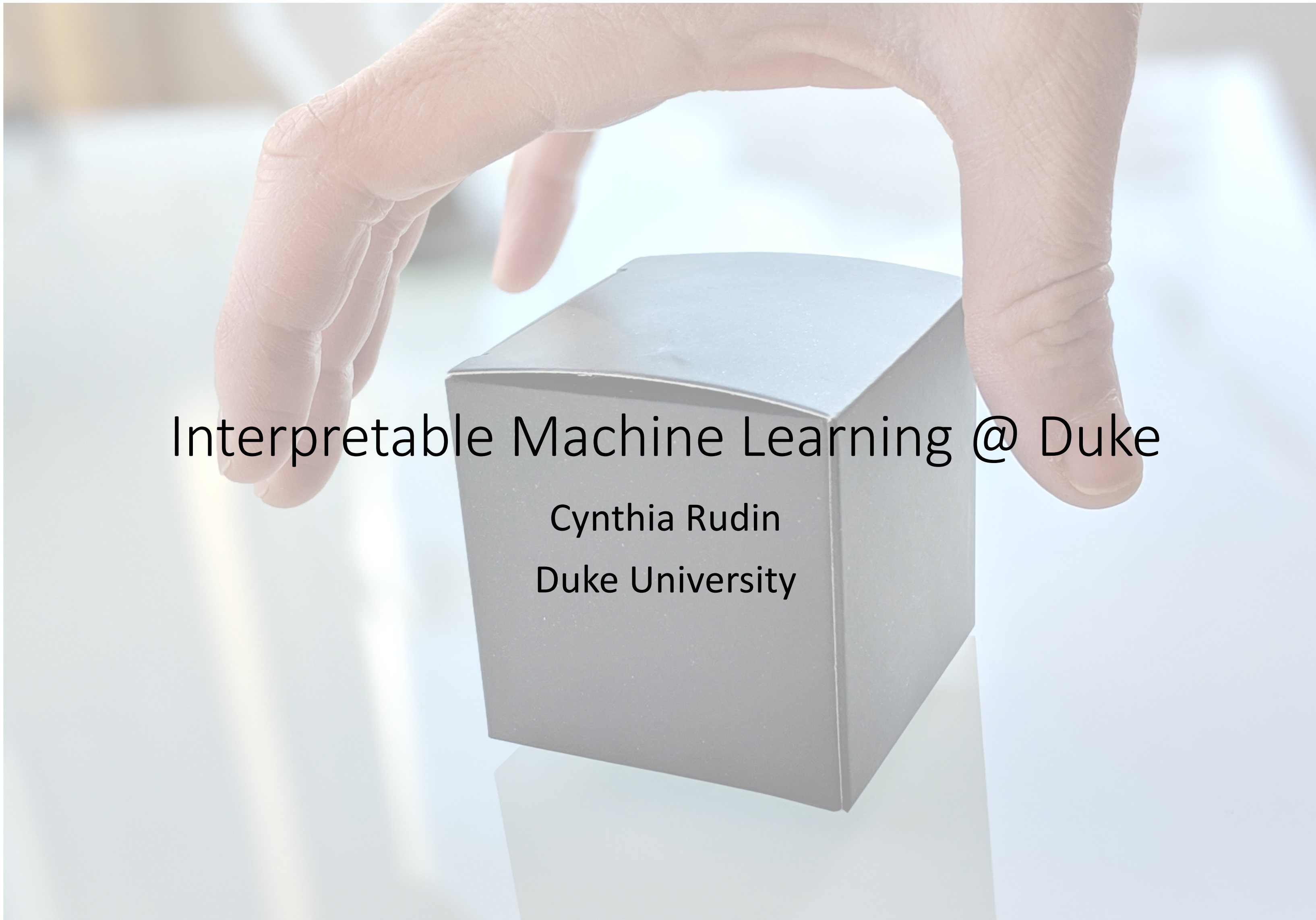
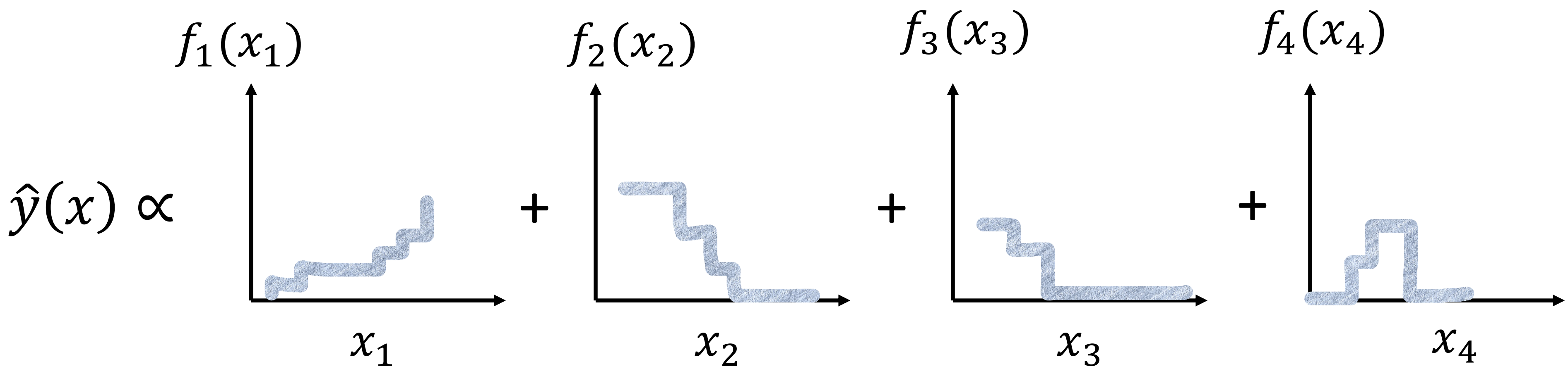


Interpretable Machine Learning @ Duke

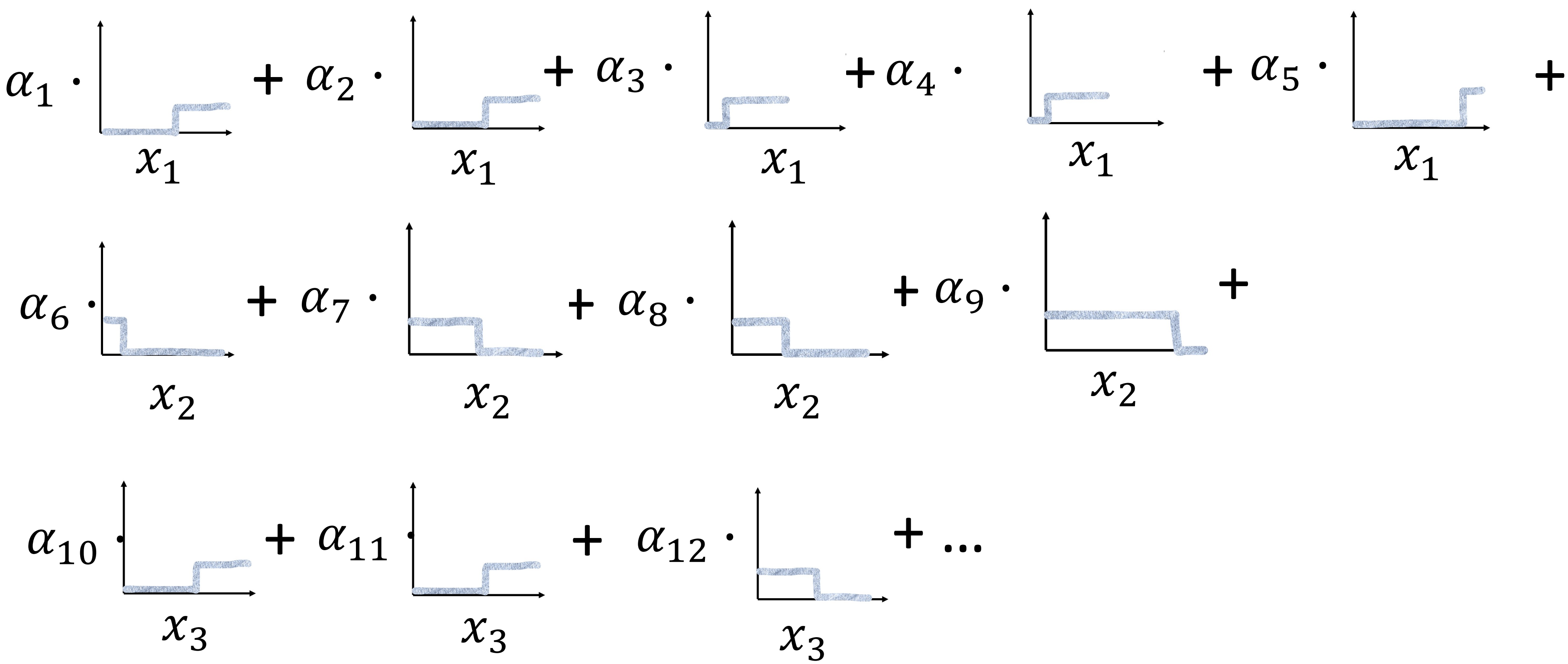
Cynthia Rudin
Duke University

In this talk, I will discuss interpretable models that are small enough to fit on a powerpoint slide.

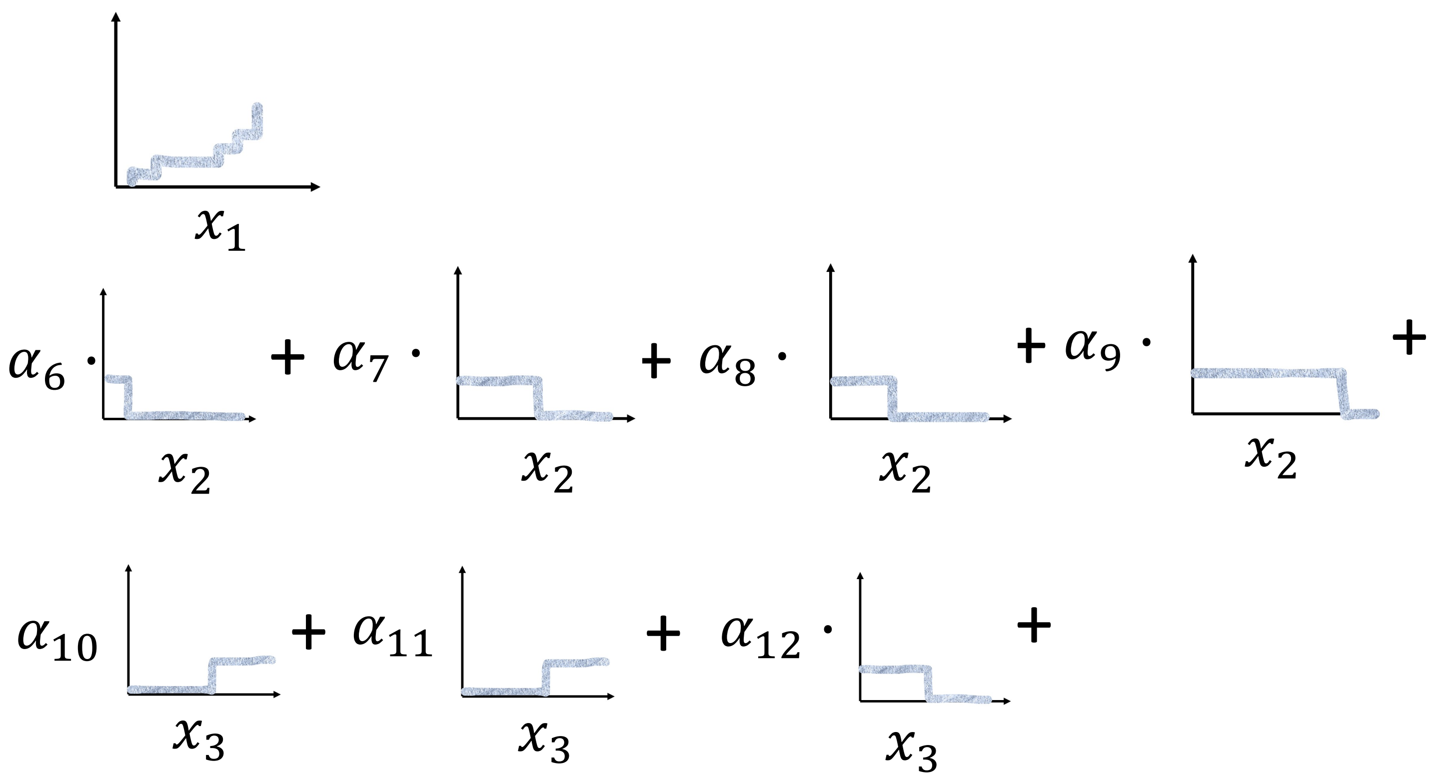
Sparse generalized additive models



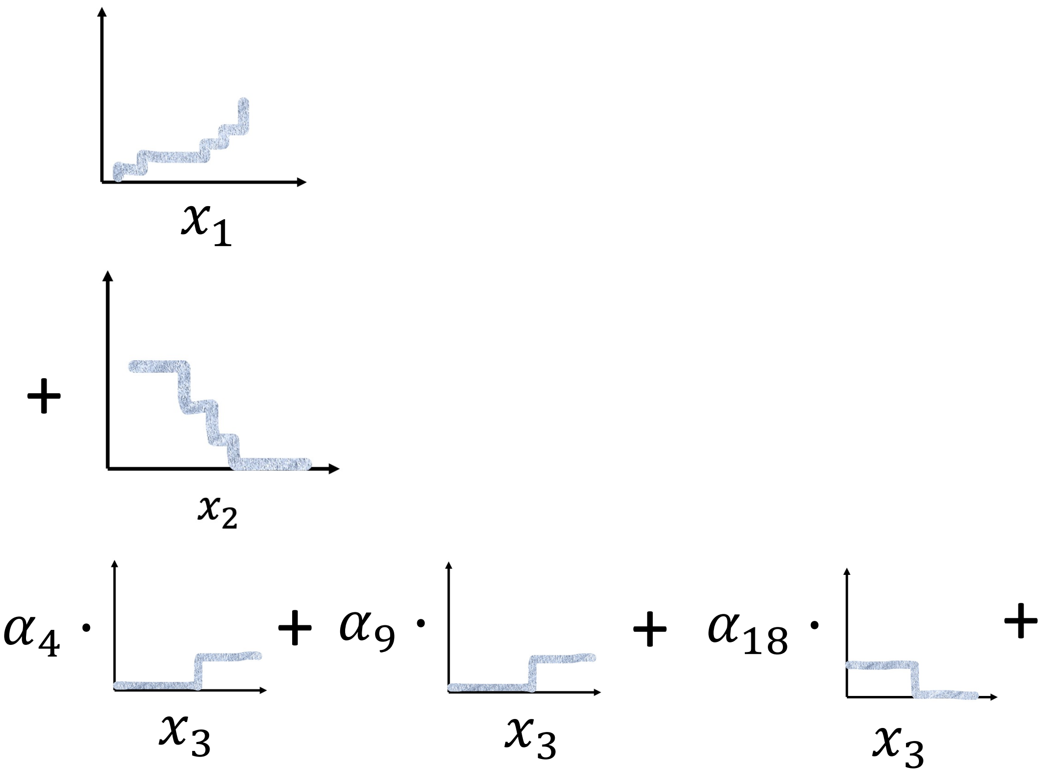
Sparse generalized additive models



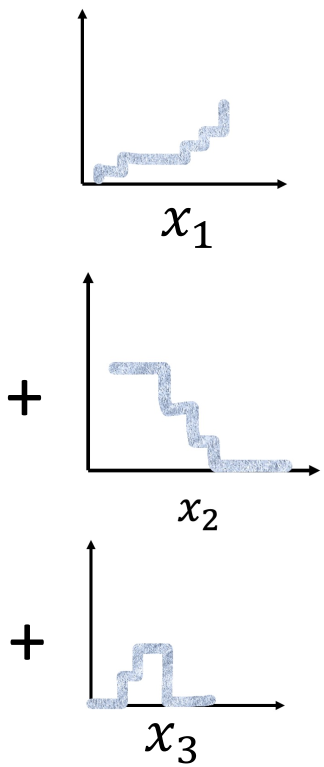
Sparse generalized additive models



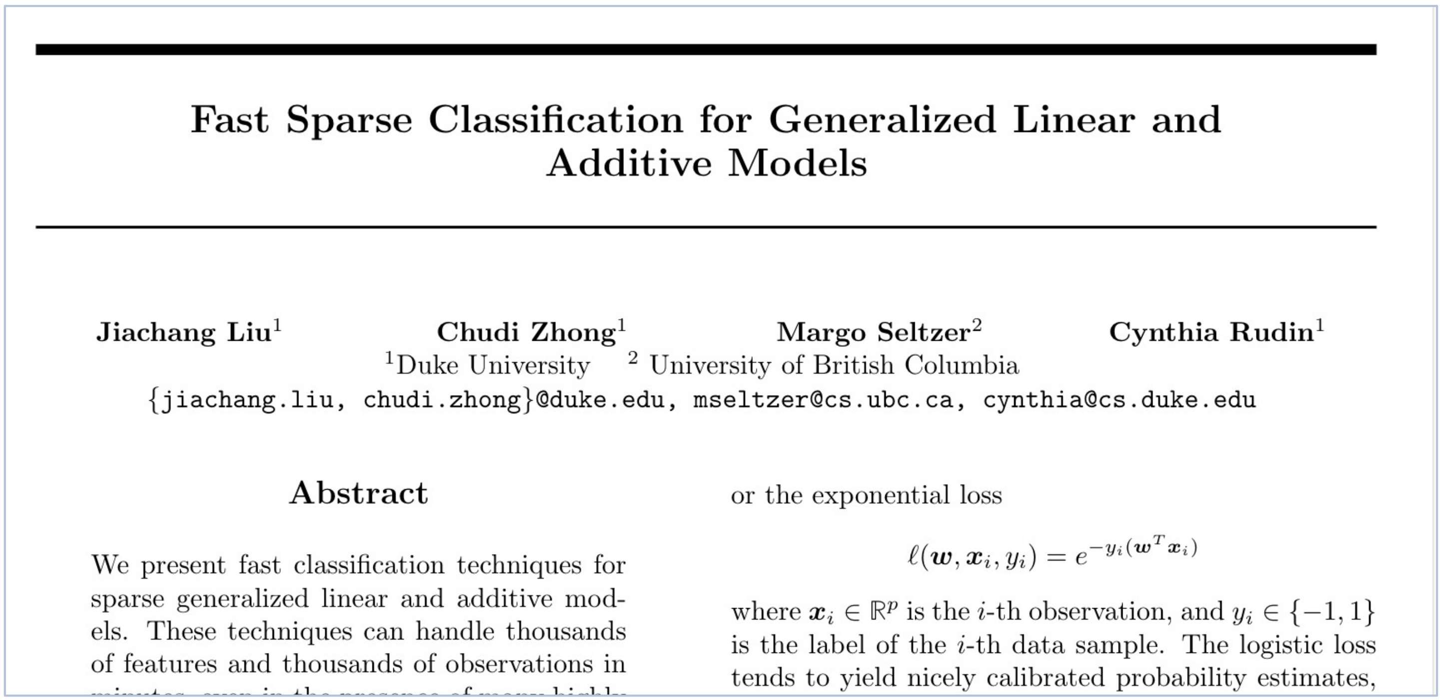
Sparse generalized additive models



Sparse generalized additive models



Sparse generalized additive models



Fast Sparse Classification for Generalized Linear and Additive Models

Jiachang Liu[1] **Chudi Zhong**[1] **Margo Seltzer**[2] **Cynthia Rudin**[1]
[1]Duke University [2] University of British Columbia
{jiachang.liu, chudi.zhong}@duke.edu, mseltzer@cs.ubc.ca, cynthia@cs.duke.edu

Abstract

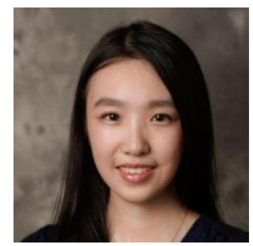
We present fast classification techniques for sparse generalized linear and additive models. These techniques can handle thousands of features and thousands of observations in

or the exponential loss

$$\ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) = e^{-y_i(\boldsymbol{w}^T \boldsymbol{x}_i)}$$

where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the i-th observation, and $y_i \in \{-1, 1\}$ is the label of the i-th data sample. The logistic loss tends to yield nicely calibrated probability estimates,

Jiachang Liu



Chudi Zhong

Margo Seltzer

AISTATS, 2022

Sparse Logistic Regression

$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) + \lambda_0 \|\boldsymbol{w}\|_0$$

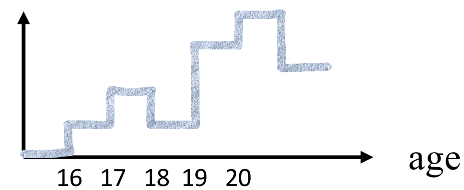
where $\ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) = \log\left(1 + e^{-y_i(\boldsymbol{w}^T \boldsymbol{x}_i)}\right)$

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

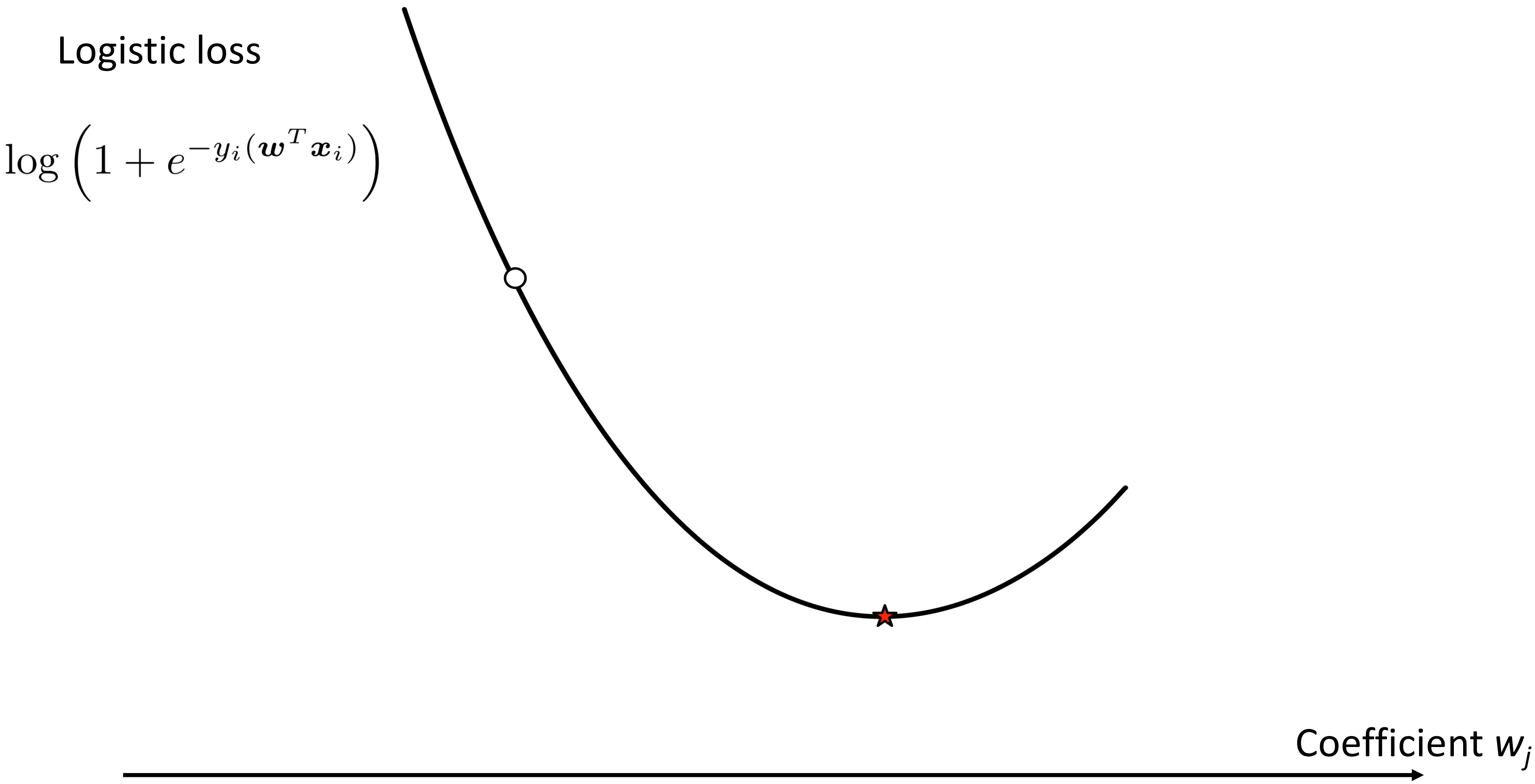
$$\hat{P}_{\text{logistic}}(y = 1|\boldsymbol{x}) = \frac{e^{f(\boldsymbol{x})}}{1+e^{f(\boldsymbol{x})}}$$

age $\longrightarrow$ age≥16, age≥17,…, age≥90

contribution to f

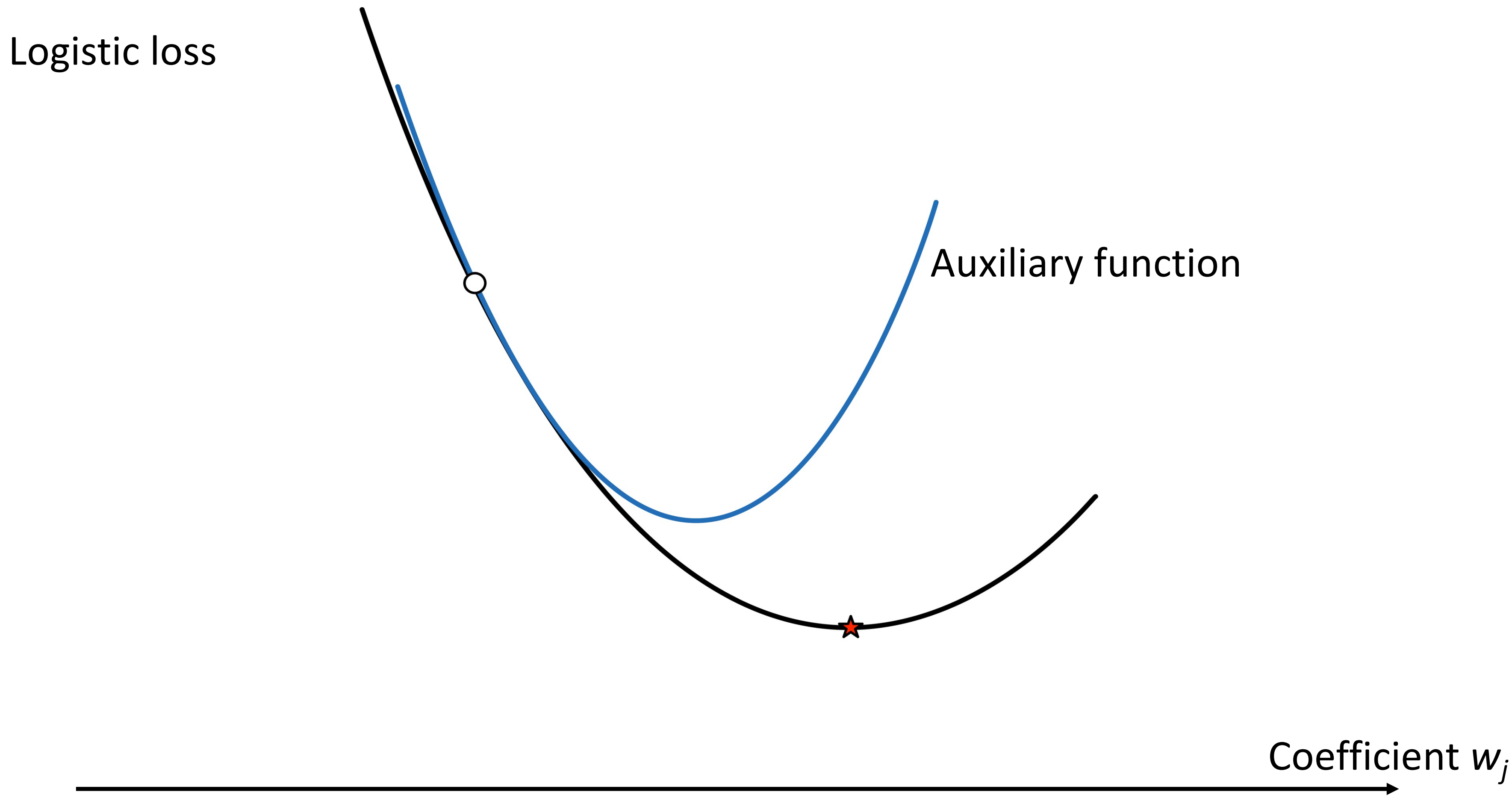


- coordinate descent + bounds (often setting coeffs to 0)
- search over subsets of features

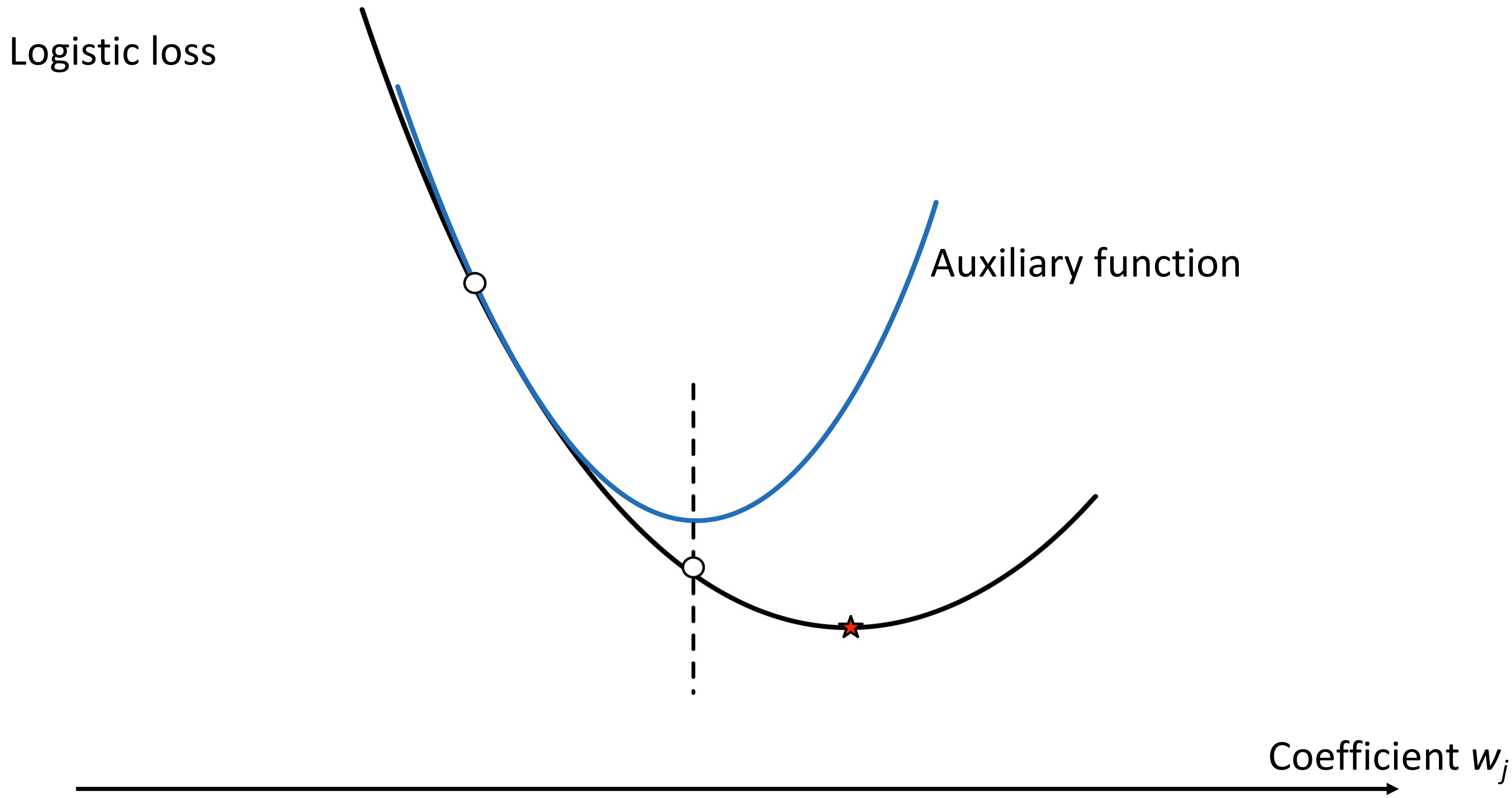


Andrei Patrascu and Ion Necoara. Random coordinate descent methods for l0 regularized convex optimization. IEEE Transactions on Automatic Control, 2015

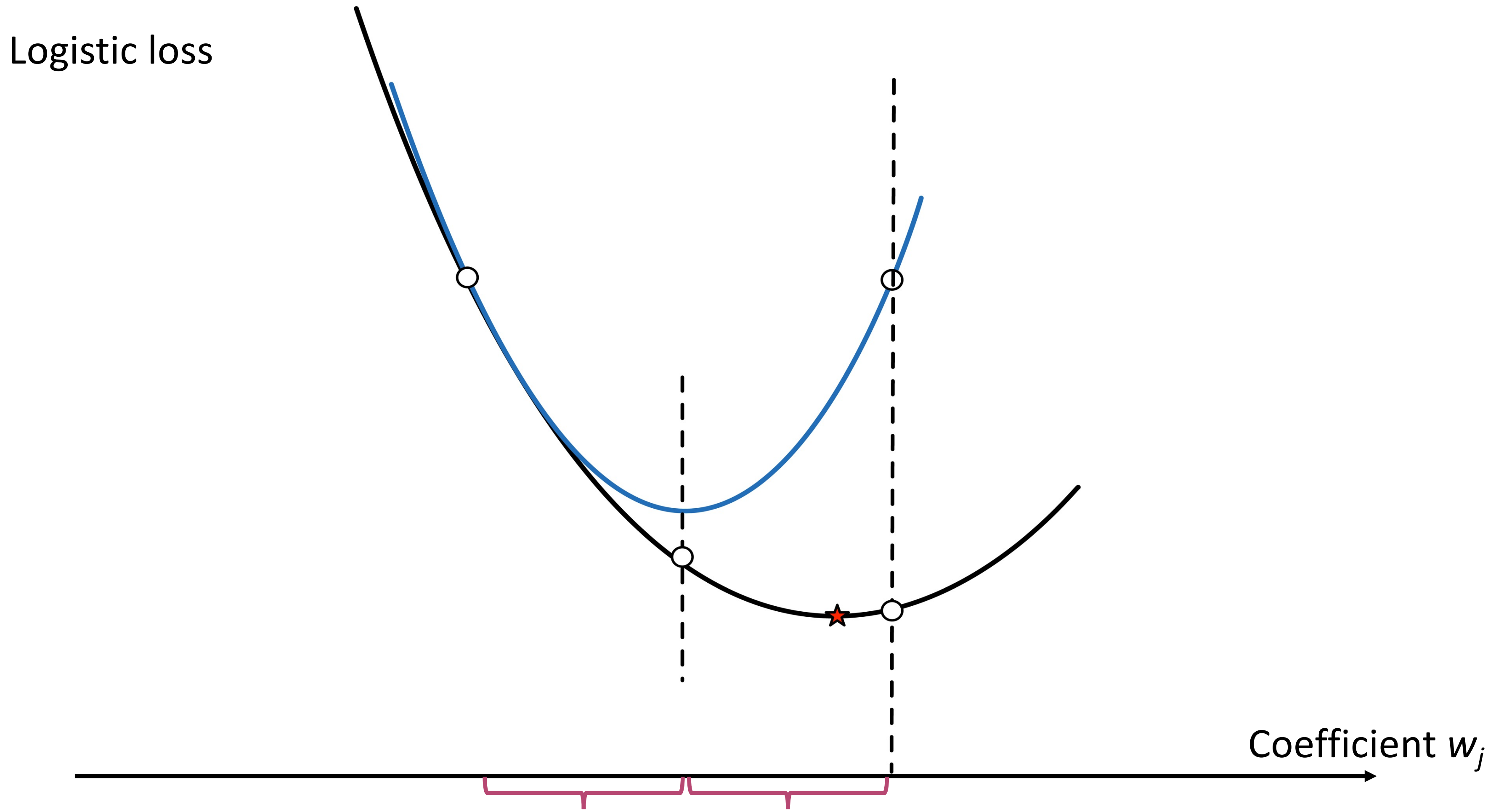
Antoine Dedieu, Hussein Hazimeh, and Rahul Mazumder. Learning sparse classifiers: Continuous and mixed integer optimization perspectives. Journal of Machine Learning Research, 2021



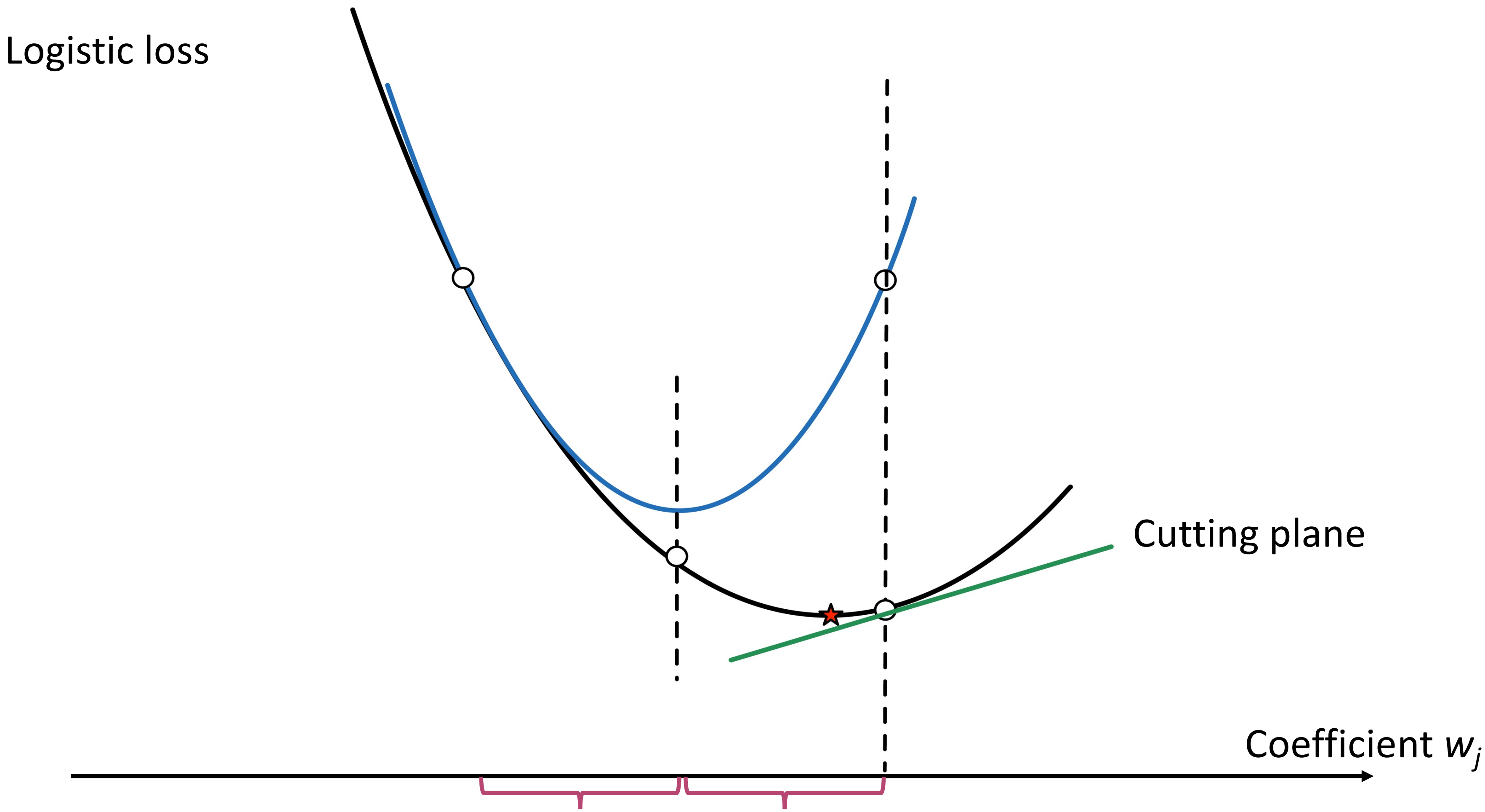
Andrei Patrascu and Ion Necoara. Random coordinate descent methods for l0 regularized convex optimization. IEEE Transactions on Automatic Control, 2015
Antoine Dedieu, Hussein Hazimeh, and Rahul Mazumder. Learning sparse classifiers: Continuous and mixed integer optimization perspectives. Journal of Machine Learning Research, 2021

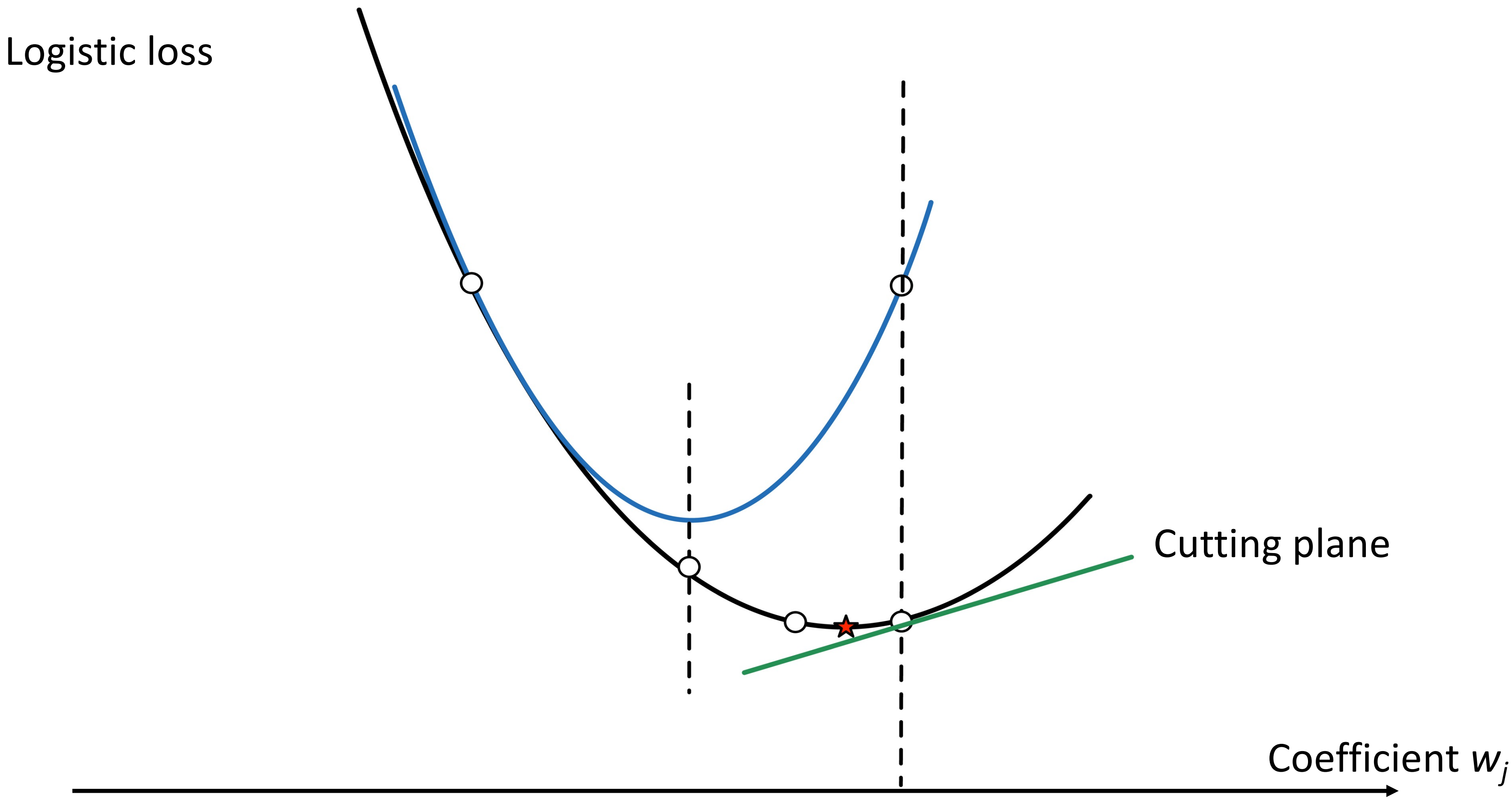


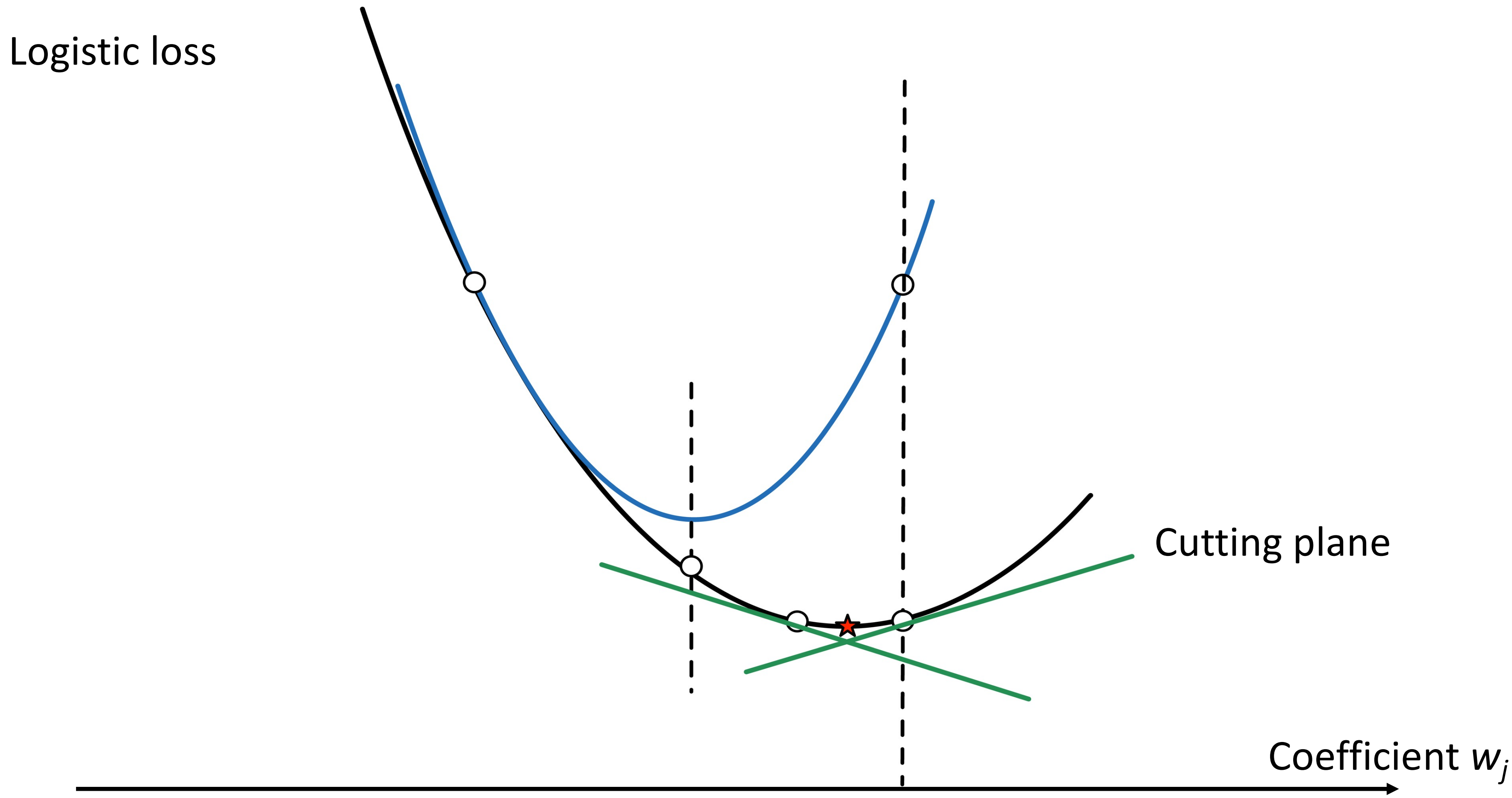
Andrei Patrascu and Ion Necoara. Random coordinate descent methods for l0 regularized convex optimization. IEEE Transactions on Automatic Control, 2015
Antoine Dedieu, Hussein Hazimeh, and Rahul Mazumder. Learning sparse classifiers: Continuous and mixed integer optimization perspectives. Journal of Machine Learning Research, 2021

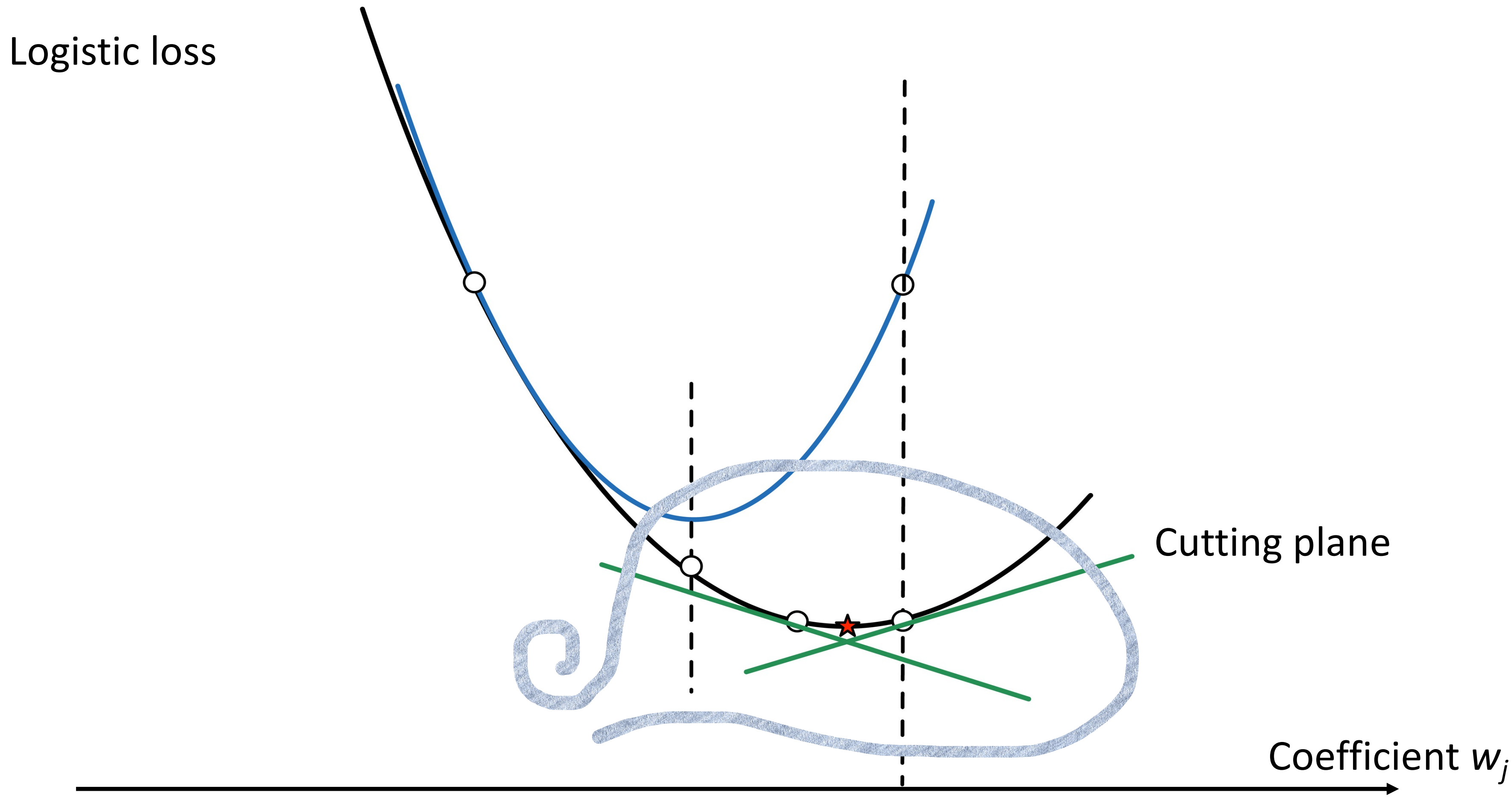


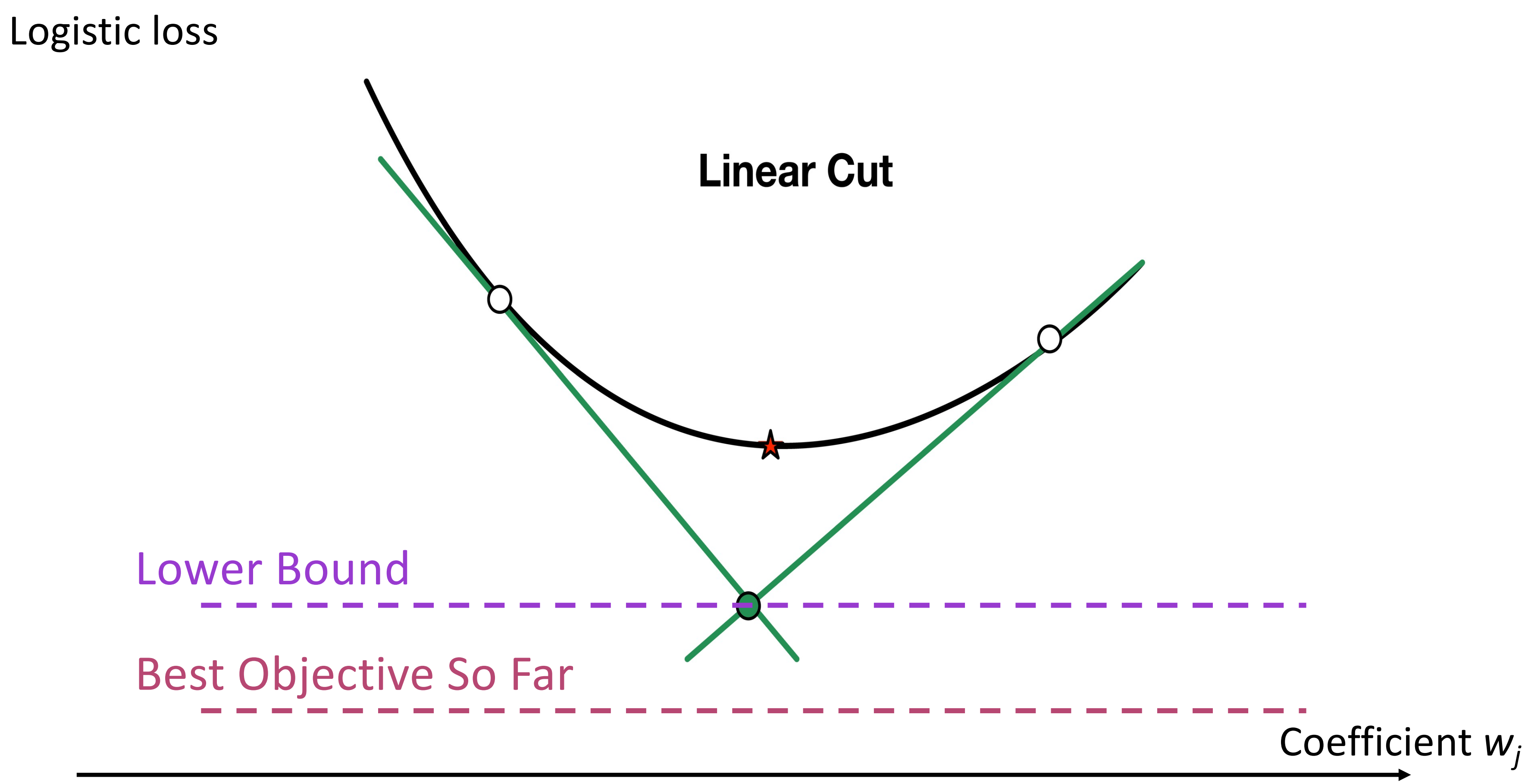
Logistic loss
Coefficient w_j

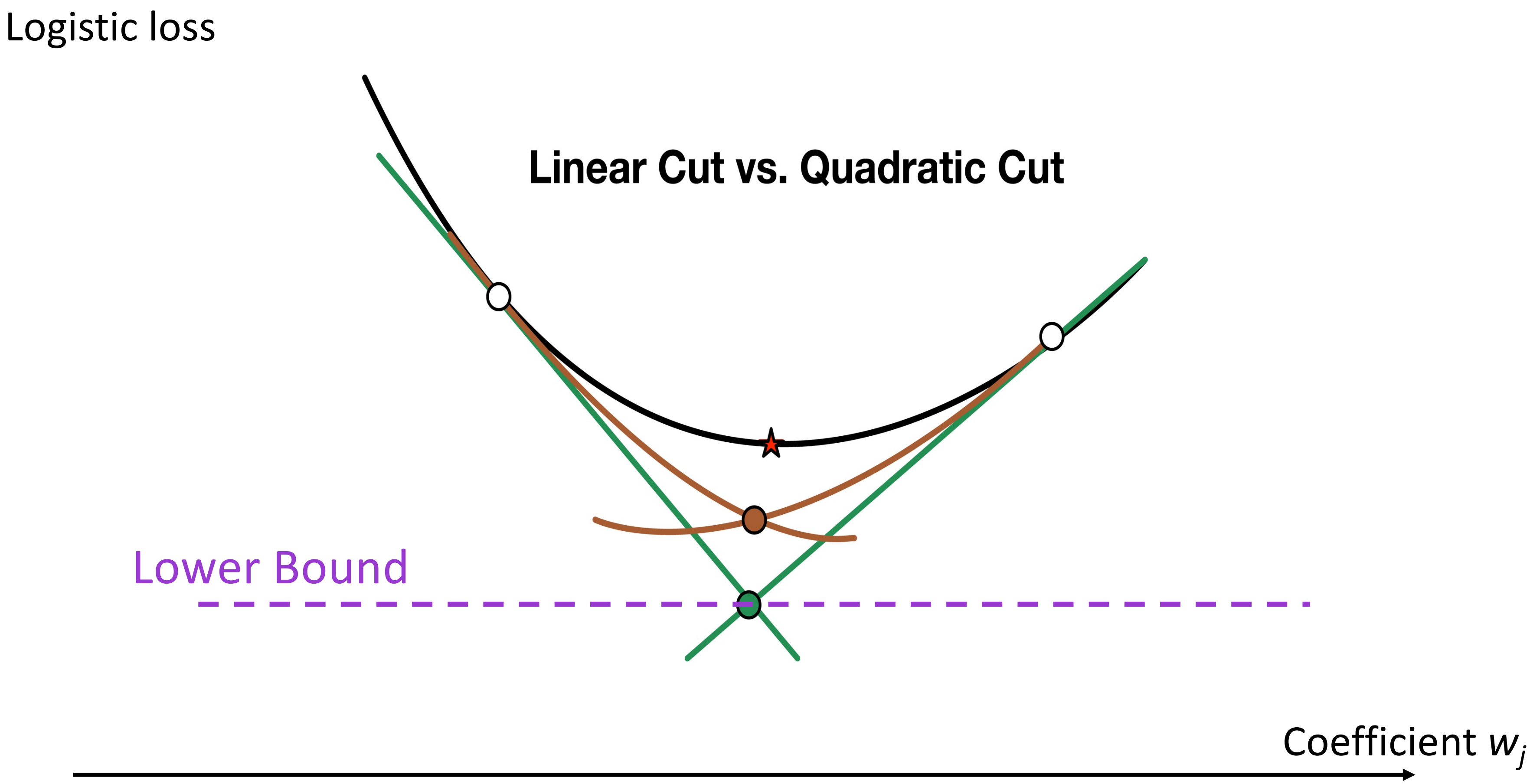

Logistic loss
Cutting plane
Coefficient w_j


Logistic loss
Cutting plane
Coefficient w_j


Logistic loss
Cutting plane
Coefficient w_j


Logistic loss
Cutting plane
Coefficient w_j


Logistic loss
Linear Cut
Lower Bound
Best Objective So Far
Coefficient w_j


Logistic loss
Linear Cut vs. Quadratic Cut
Lower Bound
Coefficient w_j

Sparse Logistic Regression

$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) + \lambda_0 \|\boldsymbol{w}\|_0$$

$$\text{where } \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) = \log\left(1 + e^{-y_i(\boldsymbol{w}^T \boldsymbol{x}_i)}\right)$$

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

$$\hat{P}_{\text{logistic}}(y = 1|\boldsymbol{x}) = \frac{e^{f(\boldsymbol{x})}}{1+e^{f(\boldsymbol{x})}}$$

- coordinate descent + bounds (often setting coeffs to 0)
- search over subsets of features

Sparse Exponential Loss Classification

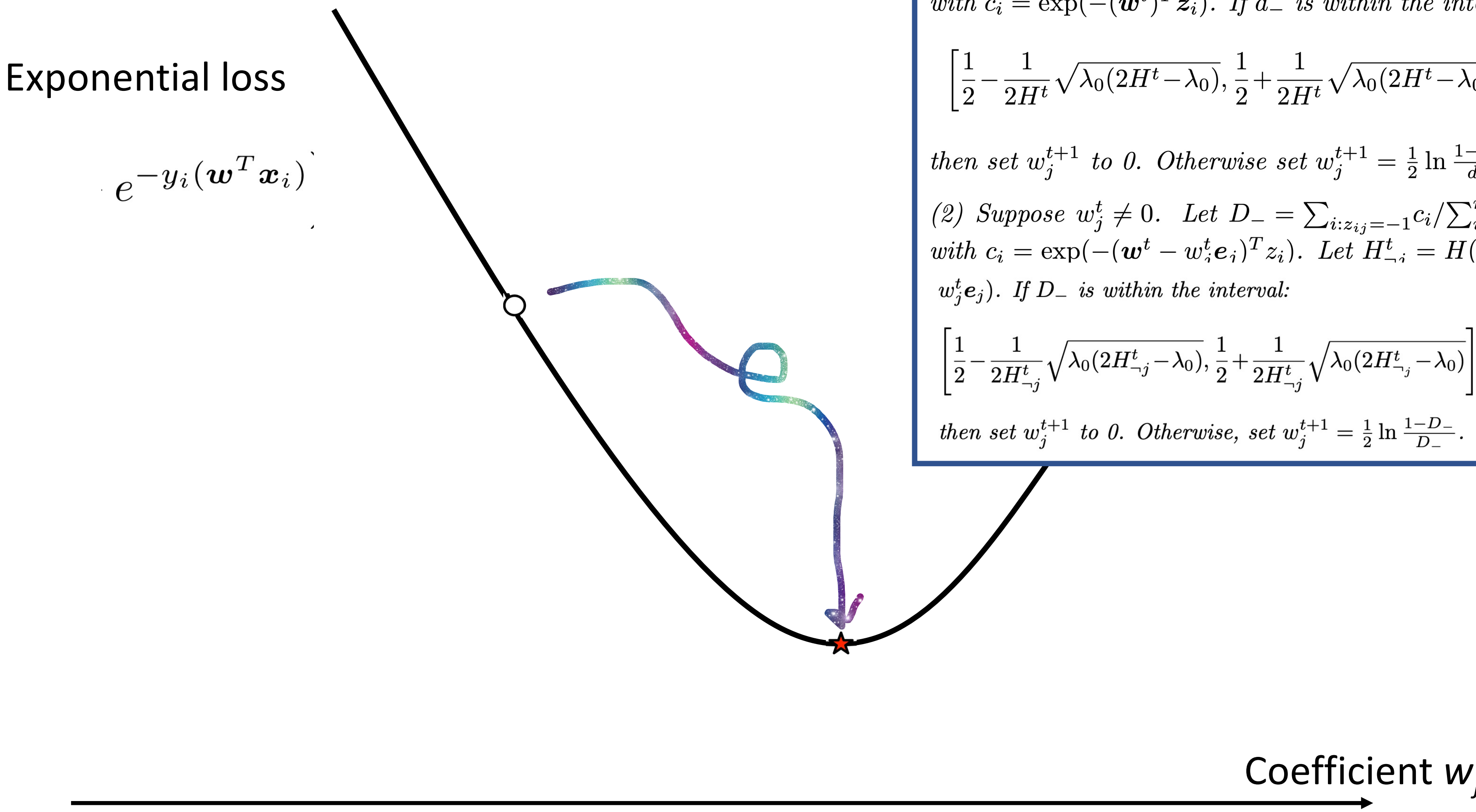
$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) + \lambda_0 \|\boldsymbol{w}\|_0$$

$$\text{where } \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) = e^{-y_i(\boldsymbol{w}^T \boldsymbol{x}_i)}$$

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

$$\hat{P}_{\text{exp loss}}(y = 1|\boldsymbol{x}) = \frac{e^{2f(\boldsymbol{x})}}{1+e^{2f(\boldsymbol{x})}}$$

- coordinate descent + bounds (often setting coeffs to 0)
- search over subsets of features



(1) Suppose $w_j^t = 0$. *Let* $d_- = \sum_{i:z_{ij}=-1} c_i / \sum_{i=1}^n c_i$, *with* $c_i = \exp(-(\boldsymbol{w}^t)^T \boldsymbol{z}_i)$. *If* d_- *is within the interval:*

$$\left[\frac{1}{2} - \frac{1}{2H^t}\sqrt{\lambda_0(2H^t-\lambda_0)}, \frac{1}{2}+\frac{1}{2H^t}\sqrt{\lambda_0(2H^t-\lambda_0)}\right],$$

then set w_j^{t+1} *to 0. Otherwise set* $w_j^{t+1} = \frac{1}{2}\ln\frac{1-d_-}{d_-}$.

(2) Suppose $w_j^t \neq 0$. *Let* $D_- = \sum_{i:z_{ij}=-1} c_i / \sum_{i=1}^n c_i$, *with* $c_i = \exp(-(\boldsymbol{w}^t - w_j^t\boldsymbol{e}_j)^T z_i)$. *Let* $H_{\neg j}^t = H(\boldsymbol{w}^t - w_j^t\boldsymbol{e}_j)$. *If* D_- *is within the interval:*

$$\left[\frac{1}{2} - \frac{1}{2H_{\neg j}^t}\sqrt{\lambda_0(2H_{\neg j}^t-\lambda_0)}, \frac{1}{2}+\frac{1}{2H_{\neg j}^t}\sqrt{\lambda_0(2H_{\neg j}^t-\lambda_0)}\right],$$

then set w_j^{t+1} *to 0. Otherwise, set* $w_j^{t+1} = \frac{1}{2}\ln\frac{1-D_-}{D_-}$.

Sparse Exponential Loss Classification

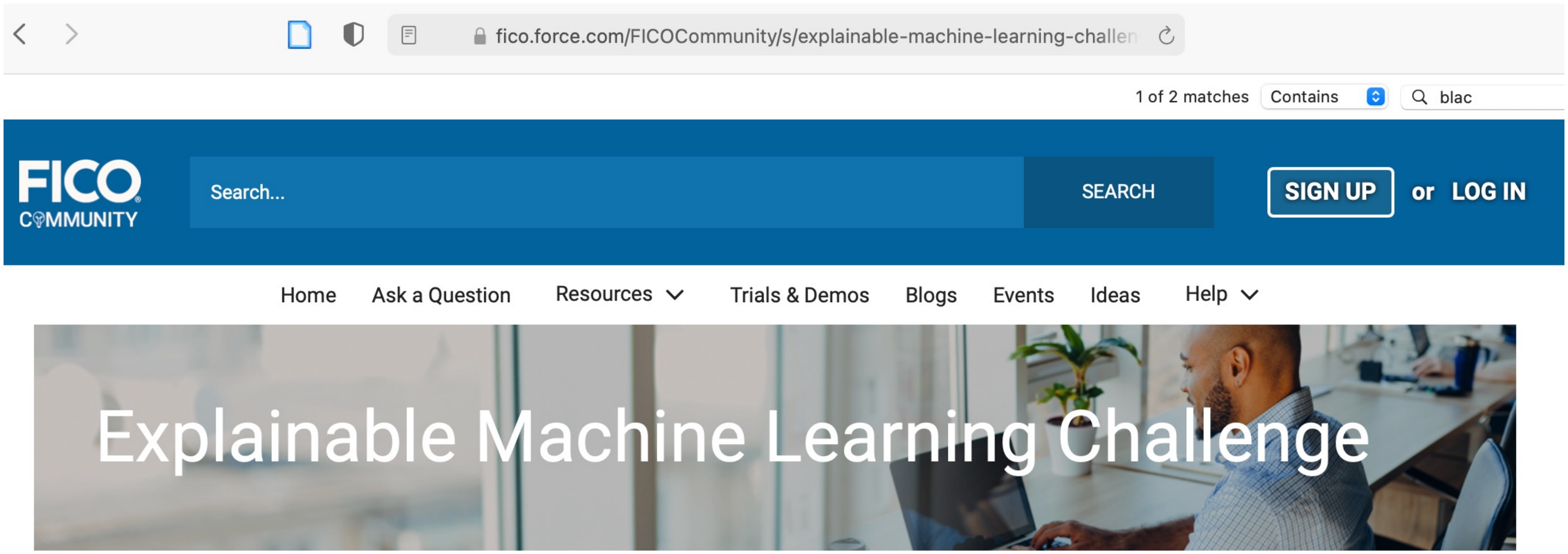
$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) + \lambda_0 \|\boldsymbol{w}\|_0$$

$$\text{where } \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) = e^{-y_i(\boldsymbol{w}^T \boldsymbol{x}_i)}$$

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

$$\hat{P}_{\text{exp loss}}(y = 1|\boldsymbol{x}) = \frac{e^{2f(\boldsymbol{x})}}{1+e^{2f(\boldsymbol{x})}}$$

- coordinate descent + bounds (often setting coeffs to 0)
- search over subsets of features



Home Equity Line of Credit (HELOC) Dataset

This competition focuses on an anonymized dataset of Home Equity Line of Credit (HELOC) applications made by real homeowners. A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price). The customers in this dataset have requested a credit line in the range of $5,000 - $150,000. The fundamental task is to use the information about the applicant in their credit report to predict whether they will repay their HELOC account within 2 years. This prediction is then used to decide whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended.

About the data

- ~10K loan applicants
- Factors:
  - External Risk Estimate
  - Months Since Oldest Trade Open
  - Months Since Most Recent Trade Open
  - Average Months In File
  - Number of Satisfactory Trades
  - Number Trades 60+ Ever
  - Number Trades 90+ Ever
  - Number of Total Trades
  - Number Trades Open In Last 12 Months
  - Percent Trades Never Delinquent
  - Months Since Most Recent Delinquency
  - Max Delinquency / Public Records Last 12 Months
  - Max Delinquency Ever
  - Percent Installment Trades
  - Net Fraction of Installment Burden
  - Number of Installment Trades with Balance
  - Months Since Most Recent Inquiry excluding 7 days
  - Number of Inquiries in Last 6 Months
  - Number of Inquiries in Last 6 Months excluding 7 days.
  - Net Fraction Revolving Burden. (Revolving balance divided by credit limit.)
  - Number Revolving Trades with Balance
  - Number Bank/Natl Trades with high utilization ratio
  - Percent of Trades with Balance

Best black box accuracy
(boosted decision trees) 73%

Best black box AUC
(2-layer neural network) .80

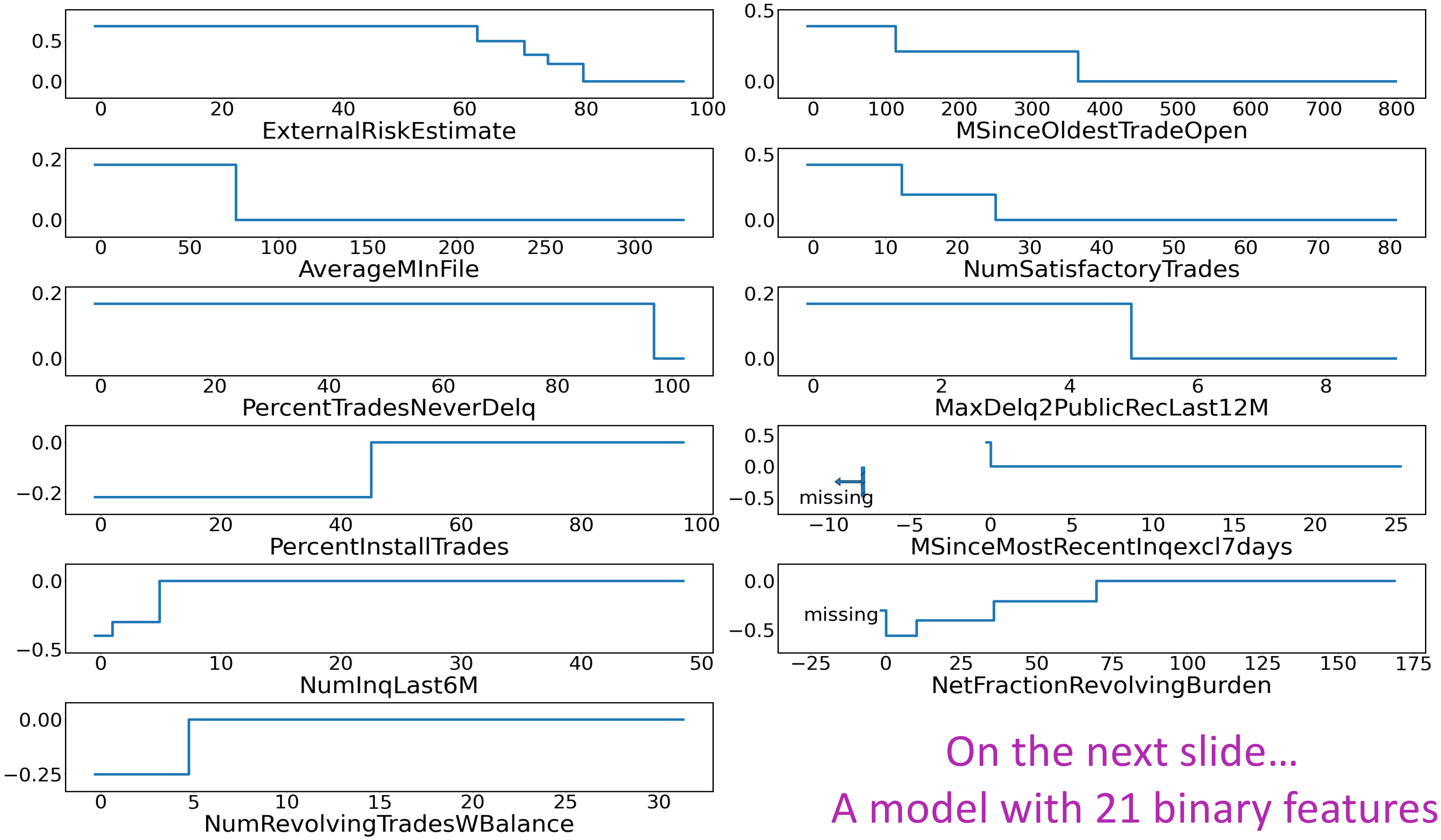
This dataset → 1917 binary features

LogRegQuad-L0 takes < 1 min
Exp-L0 takes < 20 sec

Train/Test Accuracy: 73.05±0.28, 72.35±1.24
Train/Test AUC: 80.32±0.25, 79.11±1.03

On the next slide…
A model with 21 binary features

Generalized Additive Model on the FICO Dataset

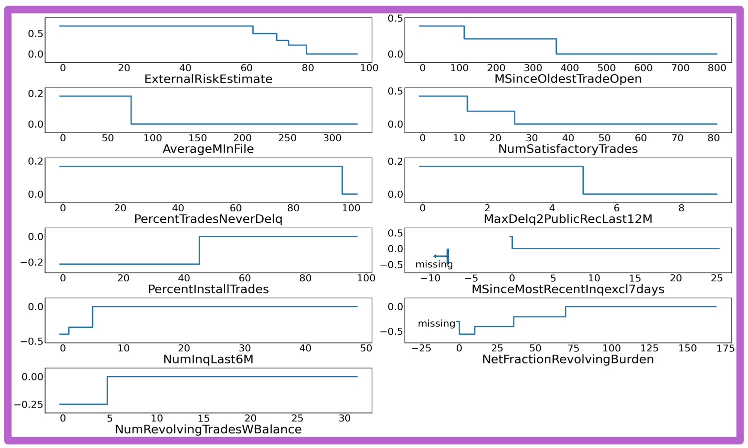


Train/Test Accuracy: 73.05, 72.28
Train/Test AUC: 80.34, 78.99

On the next slide…
A model with 21 binary features created in 3.85 seconds

Even on challenging benchmark datasets,
interpretable models' accuracy = black box accuracy.

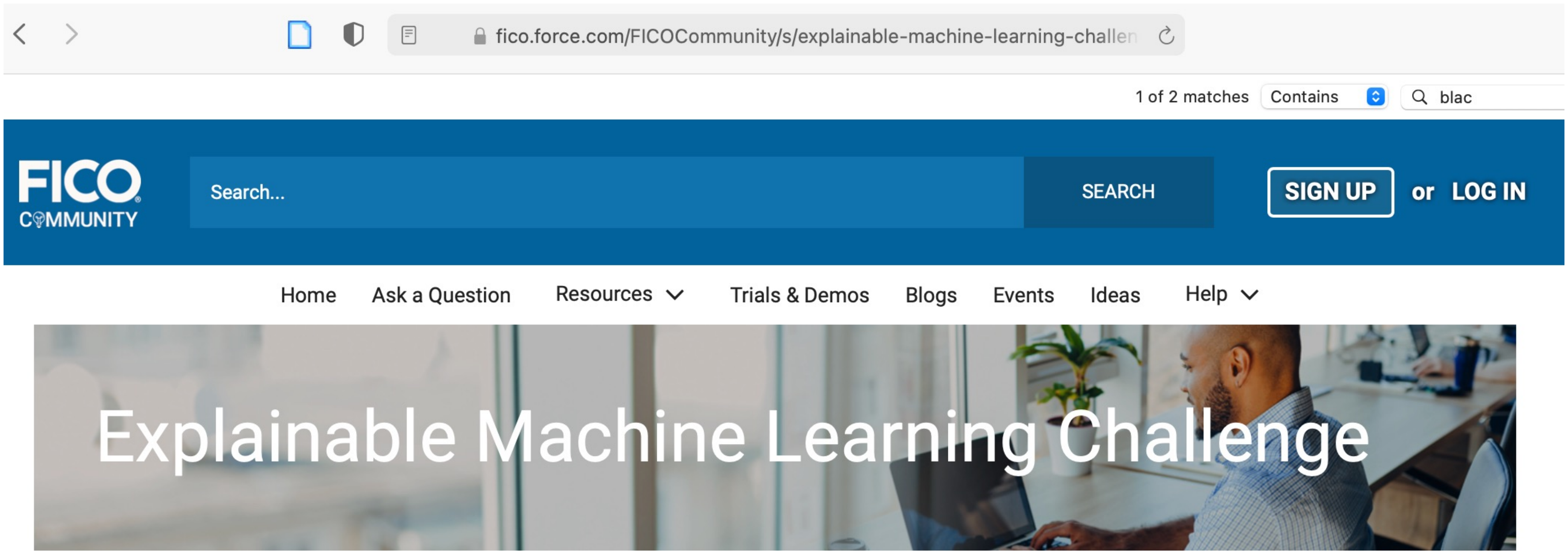
Sparse generalized additive models



3.85 sec

What’s next?

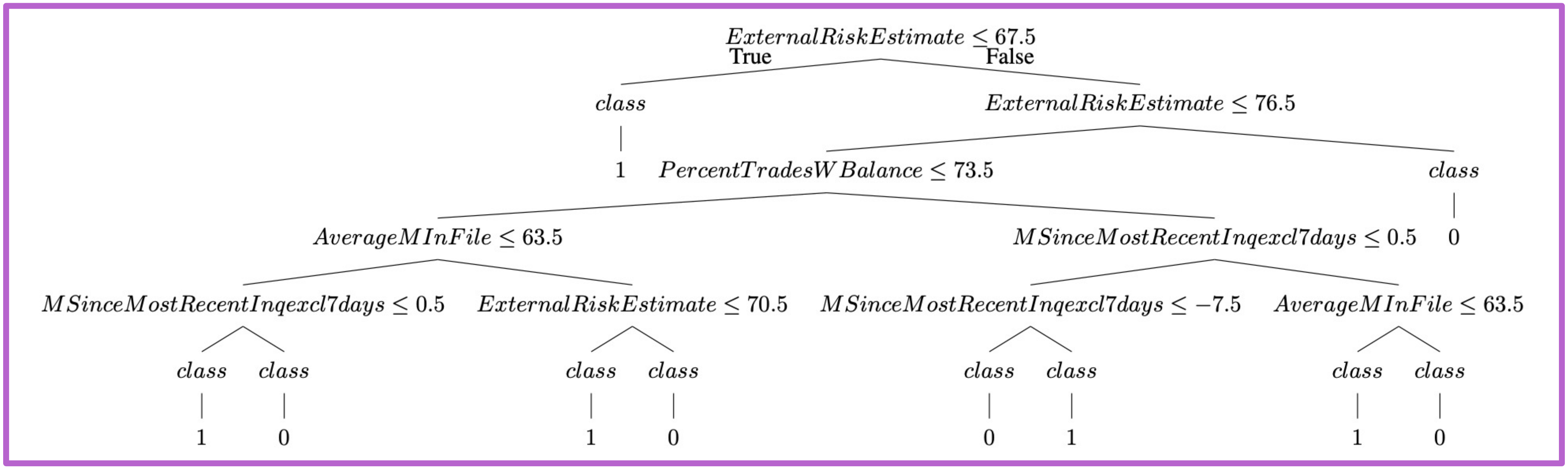
- Already heard about GAMs twice today (Rich & I).
- Decision trees! You’ve heard about those today too. (Chudi)



Home Equity Line of Credit (HELOC) Dataset

This competition focuses on an anonymized dataset of Home Equity Line of Credit (HELOC) applications made by real homeowners. A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price). The customers in this dataset have requested a credit line in the range of $5,000 - $150,000. The fundamental task is to use the information about the applicant in their credit report to predict whether they will repay their HELOC account within 2 years. This prediction is then used to decide whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended.

Explainable ML Challenge (FICO dataset) tree:



GOSDT+Guessing:

- 10K data points, >1900 binary features
- depth limit 5
- training accuracy 72% (best black box is 73%), test accuracy 71.7%
- 10 leaves
- ~8.1 sec

What’s next?

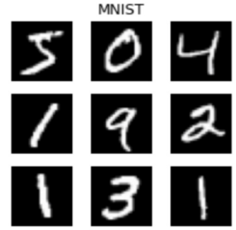
- Already heard about GAMs twice today (Me & Rich).
- Decision trees! You’ve heard about those today too. (Chudi)
- Interpretable neural networks?   (Zhi)
- Exploratory data analysis through dimension reduction

Visualizing data using t-SNE
L Maaten, G **Hinton** - Journal of machine learning research, 2008 - jmlr.org
We present a new technique called" t-SNE" that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map. The technique is a variation of Stochastic Neighbor Embedding (Hinton and Roweis, 2002) that is much easier to optimize …

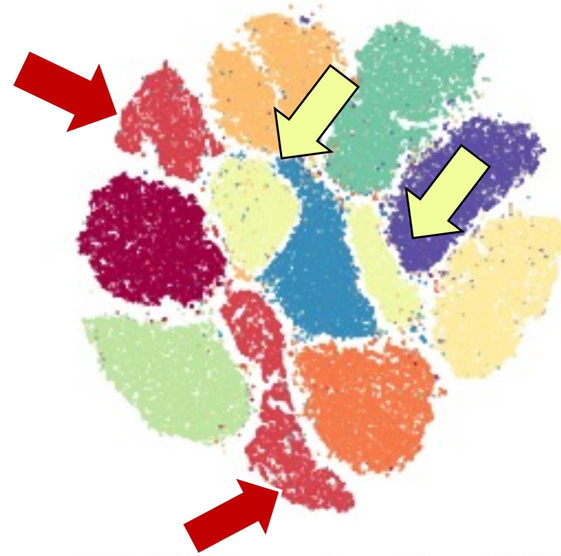
t-SNE is a dimension reduction algorithm.

Input: high-dimensional data
Output: low-dimensional data that preserves…
- the graph structure?
- local neighborhoods?
- global structure?

MNIST



t-SNE on MNIST



Visualizing data using t-SNE
L Maaten, G **Hinton** - Journal of machine learning research, 2008 - jmlr.org
We present a new technique called" t-SNE" that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map. The technique is a variation of Stochastic Neighbor Embedding (Hinton and Roweis, 2002) that is much easier to optimize …
☆ 99 Cited by 13929 Related articles All 44 versions »

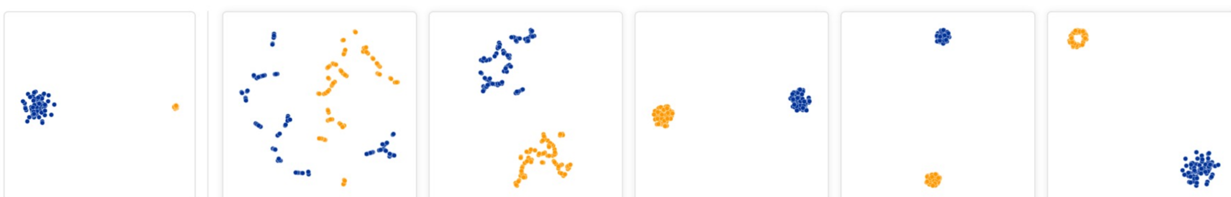
How to Use t-SNE Effectively

MARTIN WATTENBERG Google Brain | FERNANDA VIÉGAS Google Brain | IAN JOHNSON Google Cloud | Oct. 13 2016

1. Those hyperparameters really matter

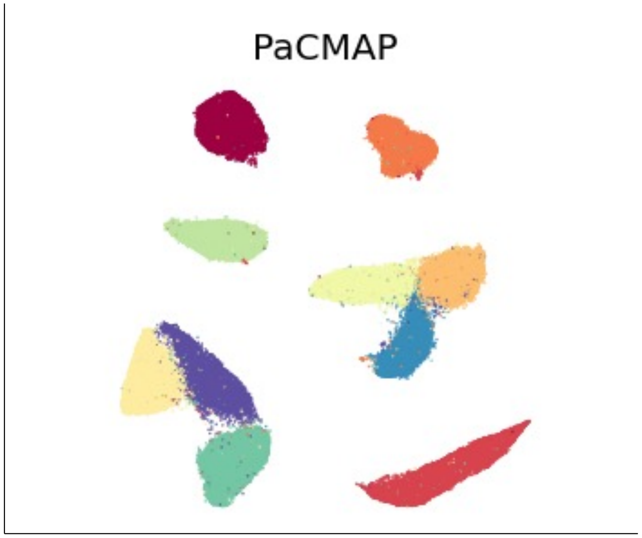
Original | Perplexity: 2 Step: 5,000 | Perplexity: 5 Step: 5,000 | Perplexity: 30 Step: 5,000 | Perplexity: 50 Step: 5,000 | Perplexity: 100 Step: 5,000

2. Cluster sizes in a t-SNE plot mean nothing



Original | Perplexity: 2 Step: 5,000 | Perplexity: 5 Step: 5,000 | Perplexity: 30 Step: 5,000 | Perplexity: 50 Step: 5,000 | Perplexity: 100 Step: 5,000

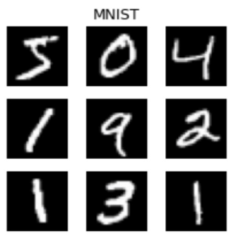
Visualizing data using t-SNE
L Maaten, G **Hinton** - Journal of machine learning research, 2008 - jmlr.org
We present a new technique called" t-SNE" that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map. The technique is a variation of Stochastic Neighbor Embedding (Hinton and Roweis, 2002) that is much easier to optimize …



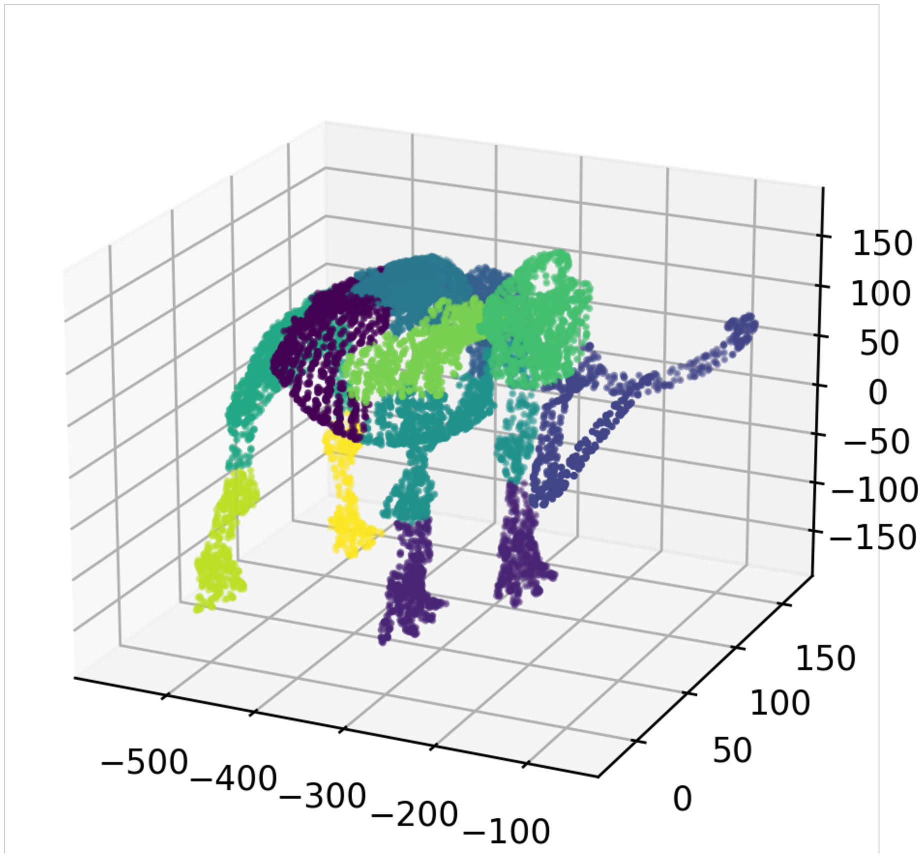
t-SNE is a dimension reduction algorithm.

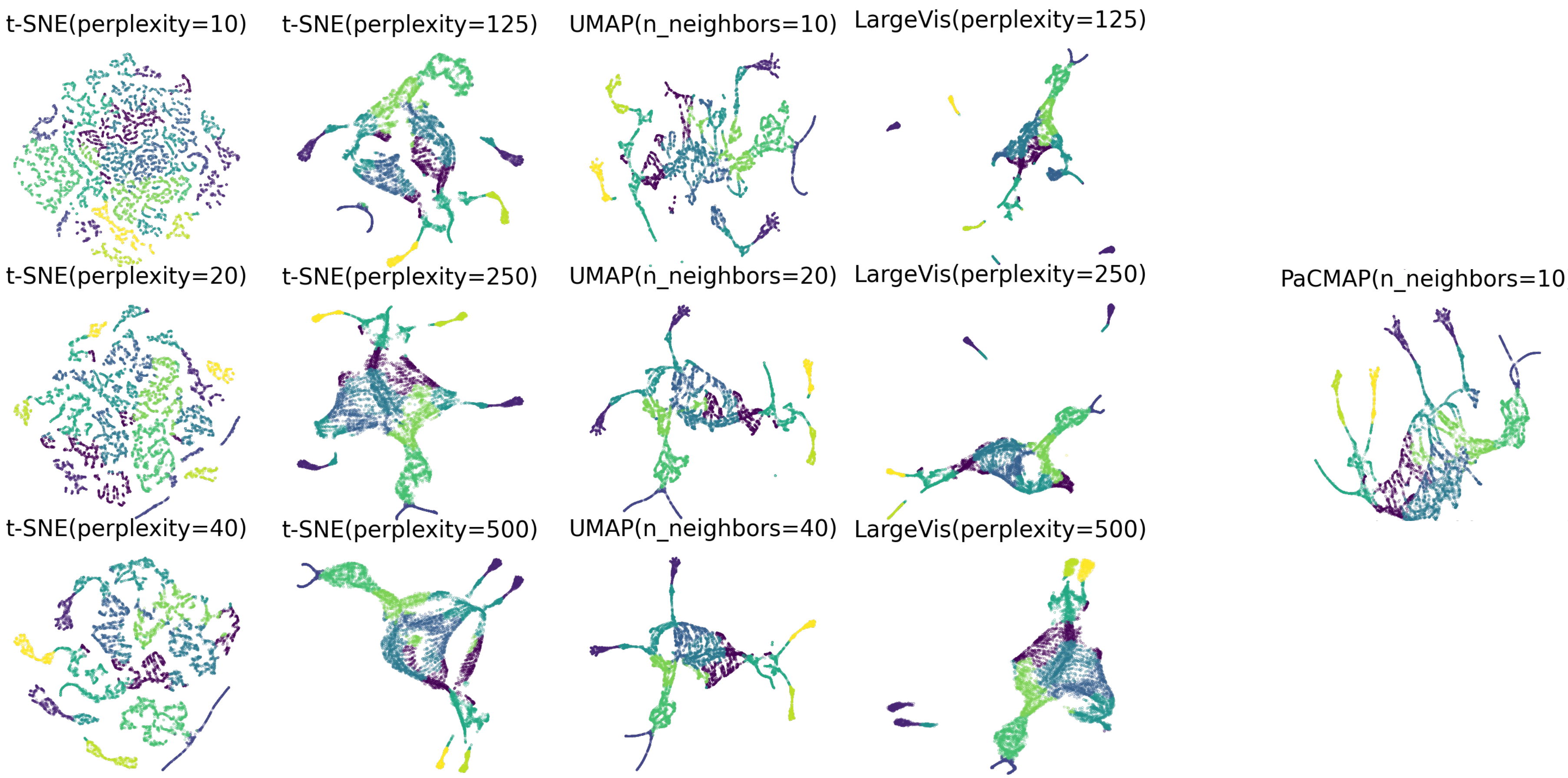
t-SNE on MNIST

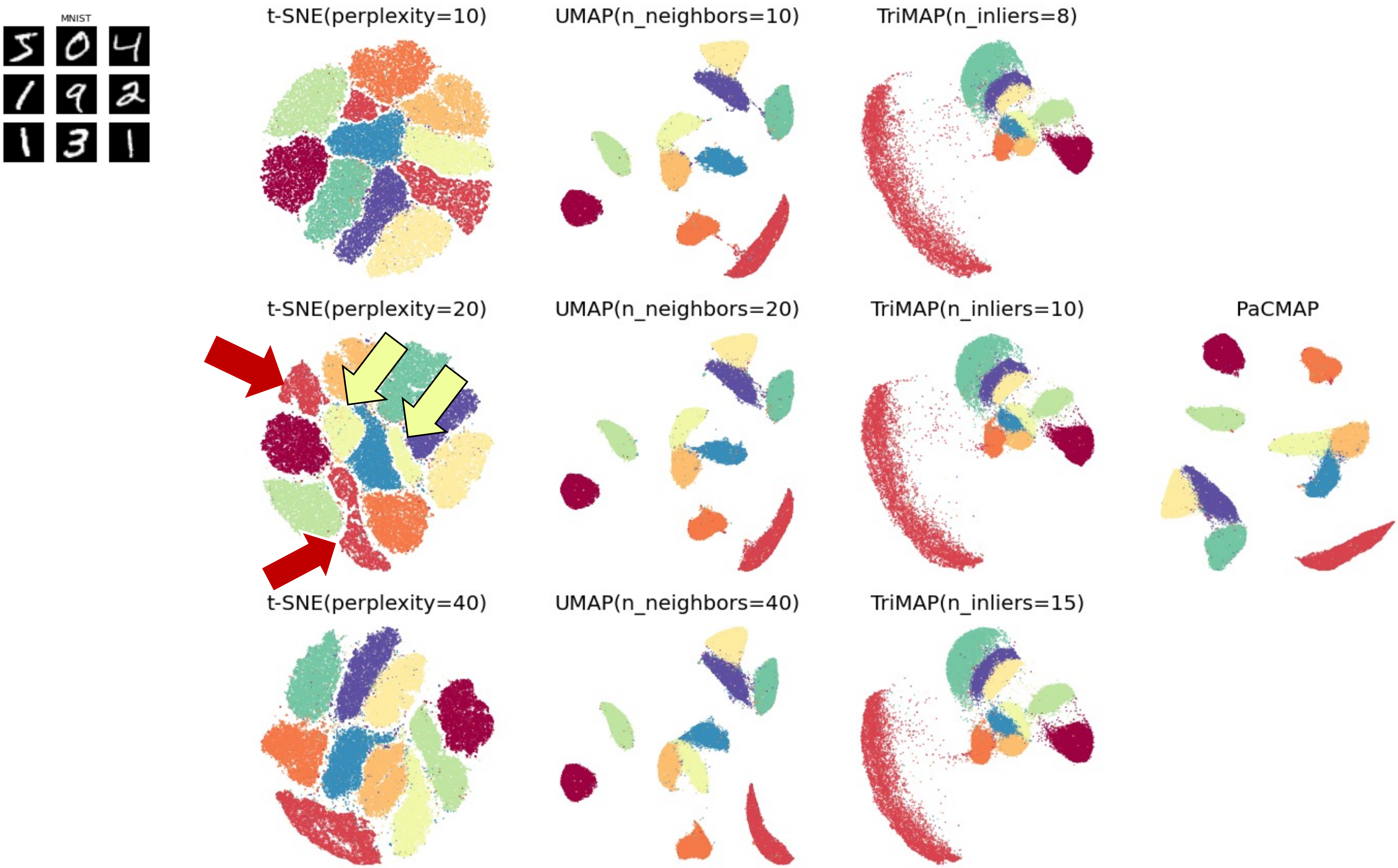
Input: high-dimensional data
Output: low-dimensional data that preserves…
- the graph structure?
- local neighborhoods?
- global structure?



Original Mammoth



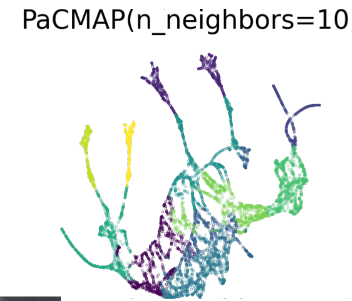

t-SNE(perplexity=10)
t-SNE(perplexity=125)
UMAP(n_neighbors=10)
LargeVis(perplexity=125)
t-SNE(perplexity=20)
t-SNE(perplexity=250)
UMAP(n_neighbors=20)
LargeVis(perplexity=250)
PaCMAP(n_neighbors=10)
t-SNE(perplexity=40)
t-SNE(perplexity=500)
UMAP(n_neighbors=40)
LargeVis(perplexity=500)


MNIST
t-SNE(perplexity=10)
UMAP(n_neighbors=10)
TriMAP(n_inliers=8)
t-SNE(perplexity=20)
UMAP(n_neighbors=20)
TriMAP(n_inliers=10)
PaCMAP
t-SNE(perplexity=40)
UMAP(n_neighbors=40)
TriMAP(n_inliers=15)

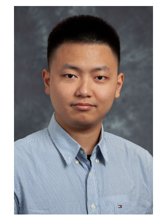
PaCMAP

- Maintains both local and global structure
- Much simpler than t-SNE or UMAP
- More computationally efficient
- Pairwise Controlled Manifold Approximation Projection

…on FICO?

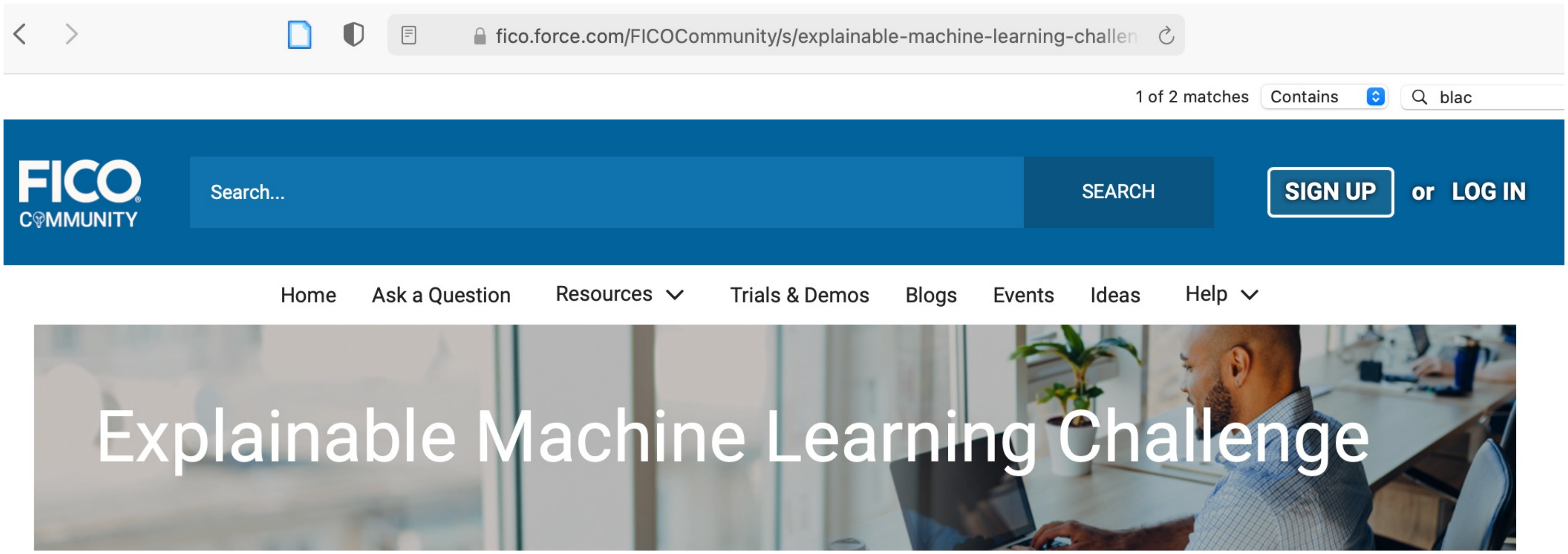


Yingfan Wang
PhD student, Duke



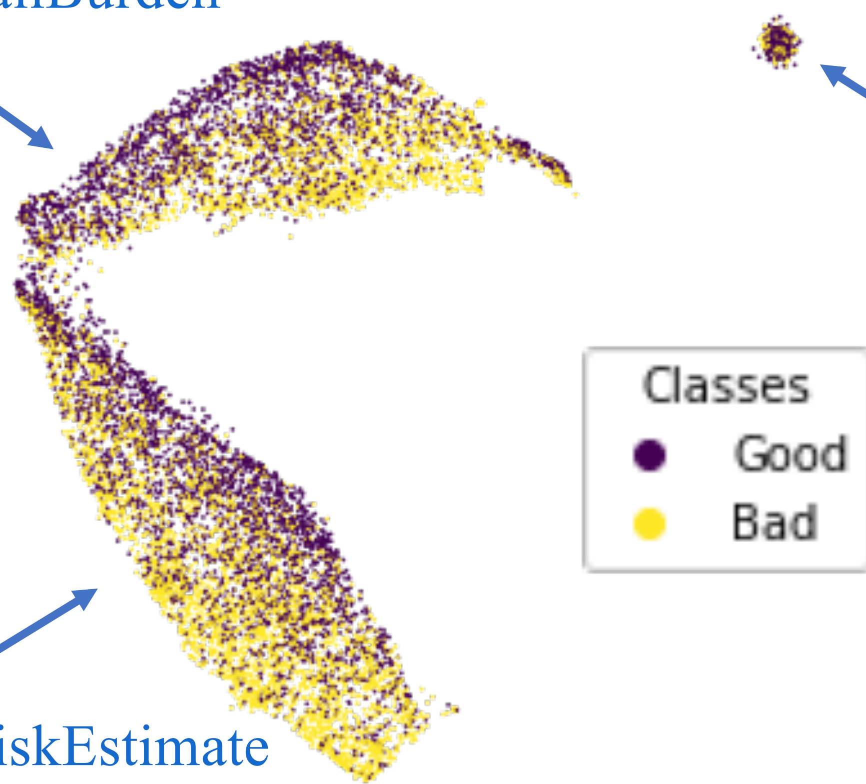
Haiyang Huang
PhD student, Duke

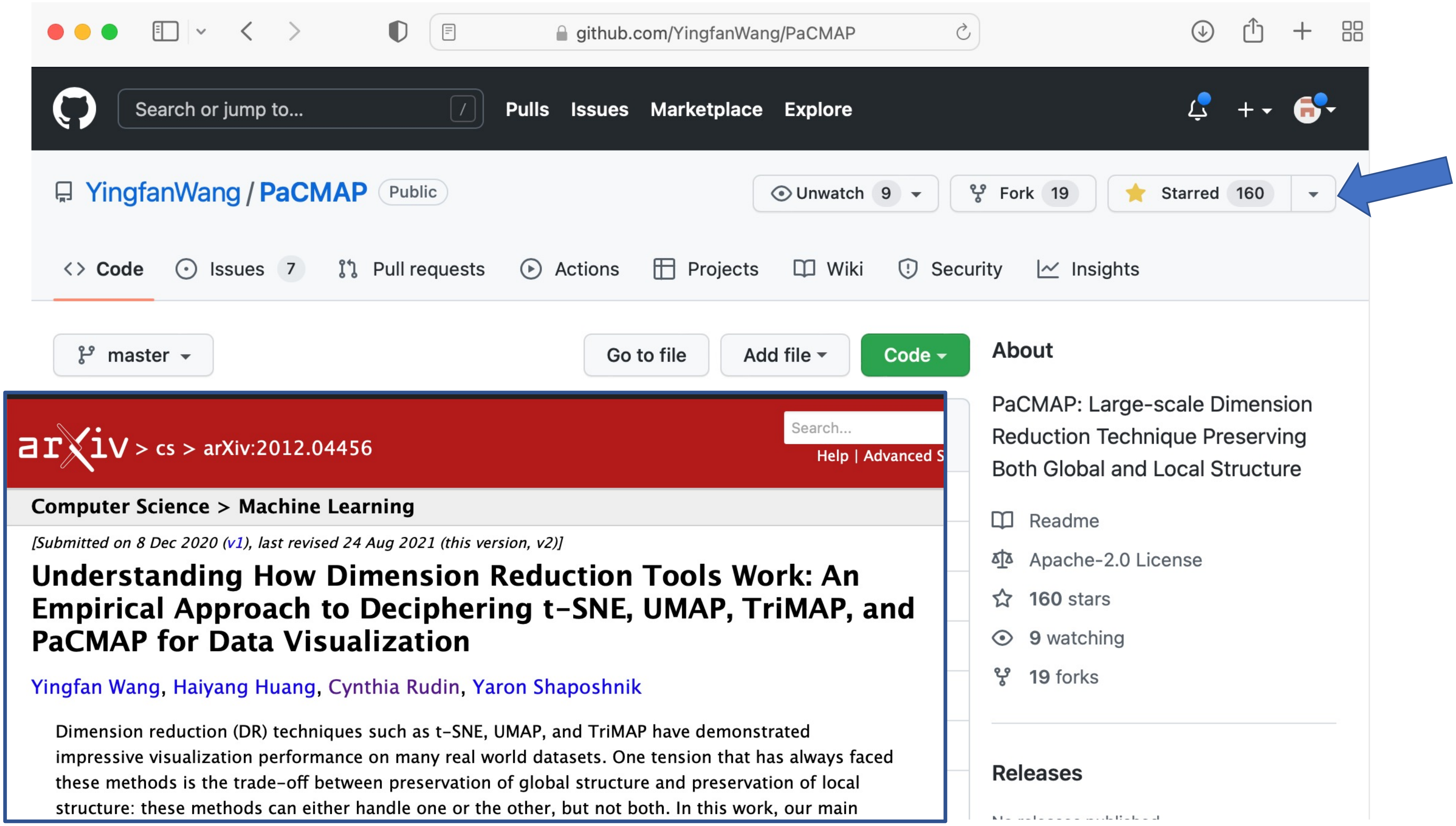
Yaron Shaposhnik
Asst. Prof., U Rochester



Home Equity Line of Credit (HELOC) Dataset

This competition focuses on an anonymized dataset of Home Equity Line of Credit (HELOC) applications made by real homeowners. A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price). The customers in this dataset have requested a credit line in the range of $5,000 - $150,000. The fundamental task is to use the information about the applicant in their credit report to predict whether they will repay their HELOC account within 2 years. This prediction is then used to decide whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended.


higher PercentTradesWBalance,
NetFractionRevolvingBurden and
NetFractionInstallBurden
has missing data
higher ExternalRiskEstimate
Classes
Good
Bad


github.com/YingfanWang/PaCMAP
Search or jump to...
Pulls
Issues
Marketplace
Explore
YingfanWang / PaCMAP
Public
Unwatch 9
Fork 19
Starred 160
Code
Issues 7
Pull requests
Actions
Projects
Wiki
Security
Insights
master
Go to file
Add file
Code
About
PaCMAP: Large-scale Dimension Reduction Technique Preserving Both Global and Local Structure
Readme
Apache-2.0 License
160 stars
9 watching
19 forks
Releases
arXiv > cs > arXiv:2012.04456
Search...
Help | Advanced S
Computer Science > Machine Learning
[Submitted on 8 Dec 2020 (v1), last revised 24 Aug 2021 (this version, v2)]
Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization
Yingfan Wang, Haiyang Huang, Cynthia Rudin, Yaron Shaposhnik
Dimension reduction (DR) techniques such as t-SNE, UMAP, and TriMAP have demonstrated impressive visualization performance on many real world datasets. One tension that has always faced these methods is the trade-off between preservation of global structure and preservation of local structure: these methods can either handle one or the other, but not both. In this work, our main

What’s next?

- Already heard about GAMs twice today (Me & Rich).
- Decision trees! You’ve heard about those today too. (Chudi)
- Interpretable neural networks?   (Zhi)
- Exploratory data analysis through dimension reduction
- Exploratory model analysis through variable importance

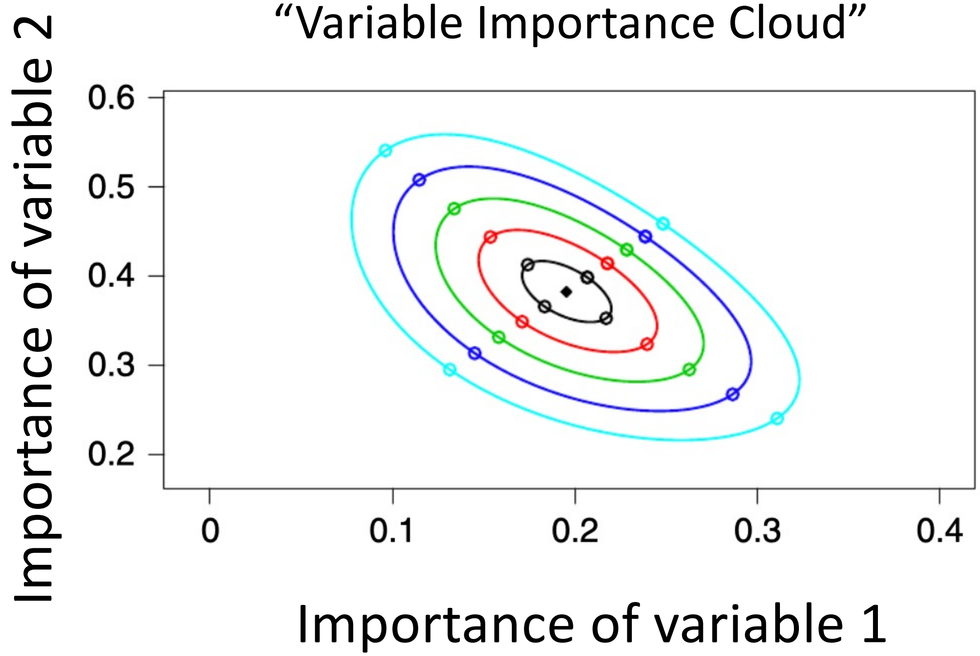
ARTICLES
https://doi.org/10.1038/s42256-020-00264-0
nature machine intelligence

Check for updates

Exploring the cloud of variable importance for the set of all good models

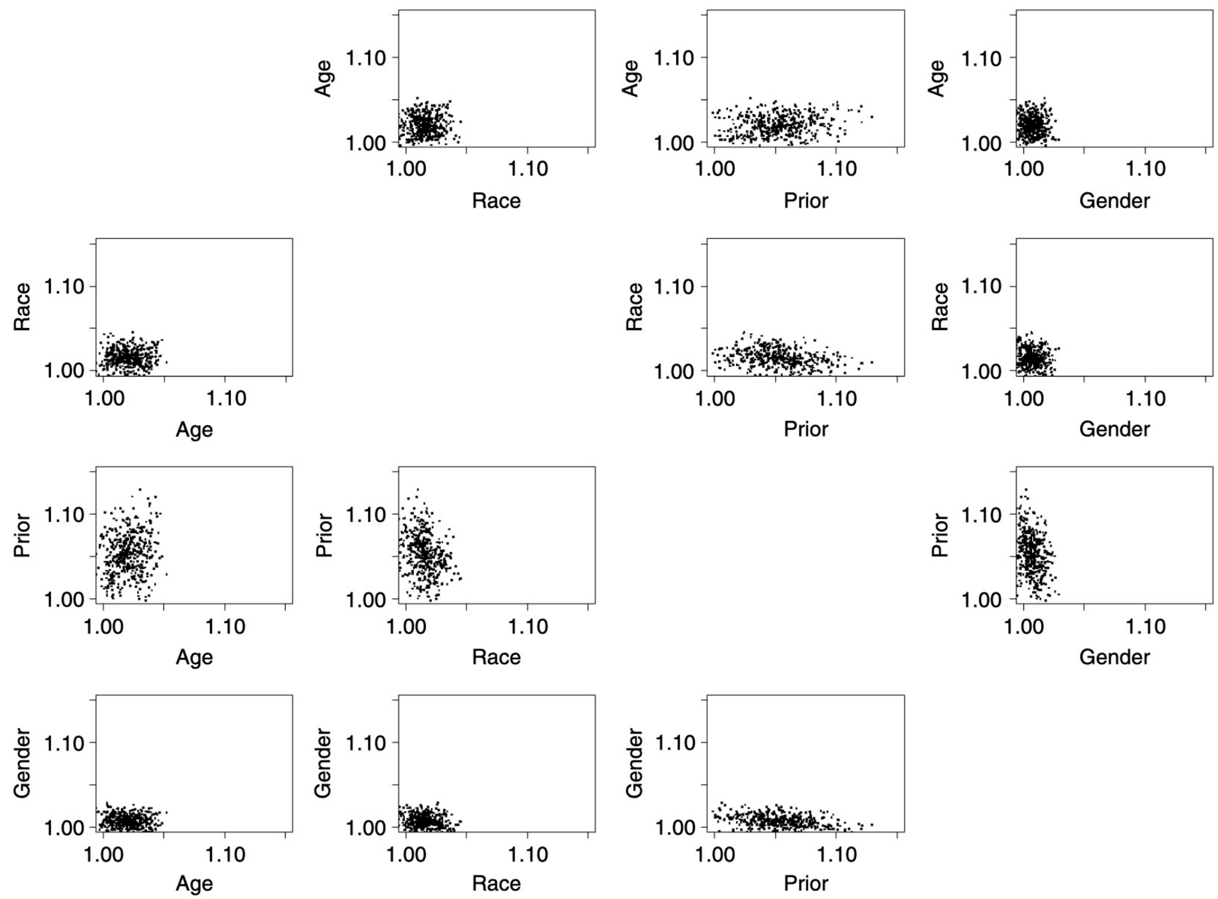
Jiayun Dong [1] and Cynthia Rudin [2]

“Variable Importance Cloud”



ARTICLES

nature machine intelligence



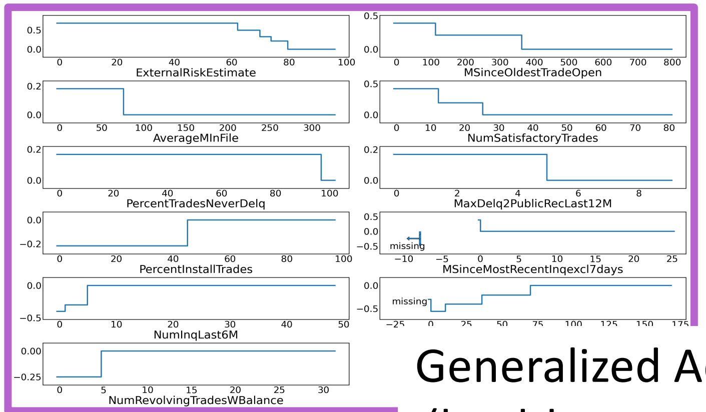
Importance of variable 2

Importance of variable 1

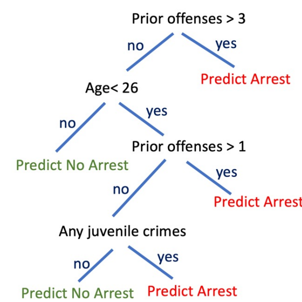
What’s next?

- Already heard about GAMs twice today (Me & Rich).
- Decision trees! You’ve heard about those today too. (Chudi)
- Interpretable neural networks?   (Zhi)
- Exploratory data analysis through dimension reduction
- Exploratory model analysis through variable importance

Interpretable Machine Learning Lab



Generalized Additive Models
(healthcare, criminal justice)

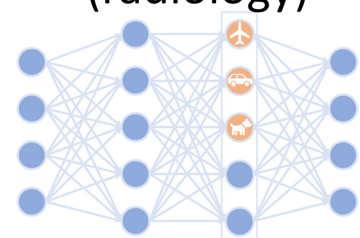


Optimal Sparse
Decision Trees
(materials science)

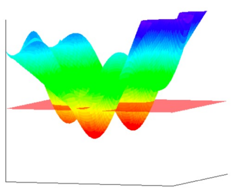
Data Visualization/
Dimension Reduction
(biology)

Thanks

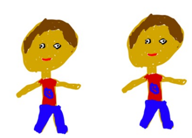
Interpretable Neural Networks for
Computer Vision
(radiology)



Understanding the
Set of Good Models
and Importance of Variables



Almost Exact Matching for Causal Inference
(criminal justice)



Neural Disentanglement