Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges

Cynthia Rudin, Alina Barnett

Duke University

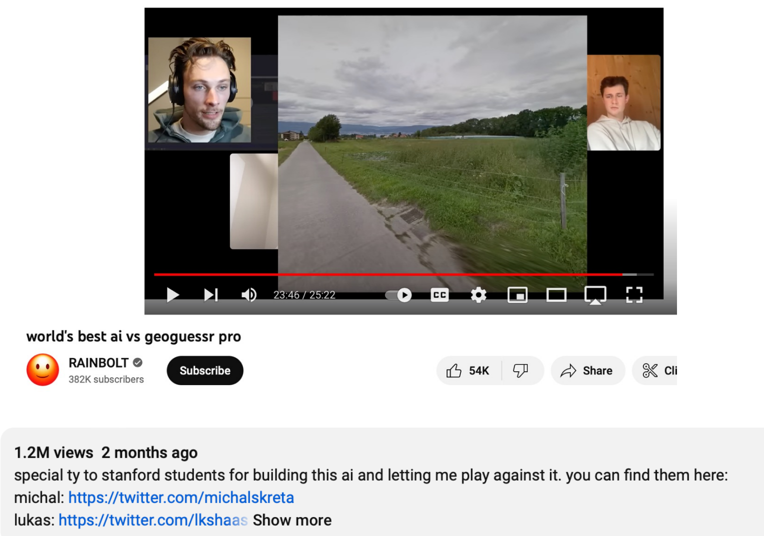
with Chaofan Chen, Zhi Chen, Haiyang Huang,

Lesia Semenova, Chudi Zhong

- A <u>*black box machine learning*</u> model is a formula that is either too complicated for any human to understand, or proprietary, so that one cannot understand its inner workings.

Black box models

- Are hard to troubleshoot while designing them
  - “Does the model often predict the right answer for the wrong reason?”



(a) Attention attribution map for an image in Canada.

Black box models

- Are hard to troubleshoot while designing them
  - “Does the model often predict the right answer for the wrong reason?”
- Are hard to troubleshoot in practice
  - “Will this model predict accurately for my current patient?”
  - “Could a typo in the inputs have led to this prediction?”
- Are hard to evaluate with respect to bias and fairness
  - “Does this model depend on a variable I don’t want it to?”
- Are hard to “explain”
  - Most “explanations” are flawed or incomplete. They often disagree with each other.
  - Makes the problem worse by providing false/misleading characterizations.
  - Adds unnecessary authority to the black box
  - Replacing the black box is almost always the better option.

Black box models turn computer-aided decisions into automated decisions.

- A *black box machine learning* model is a formula that is either too complicated for any human to understand, or proprietary, so that one cannot understand its inner workings.

- An *interpretable machine learning model* obeys a domain-specific set of constraints to allow it (or its predictions, or the data) to be more easily understood by humans. *These constraints can differ dramatically depending on the domain.*

- There's a spectrum.

Computer Science > Machine Learning

[Submitted on 20 Mar 2021]

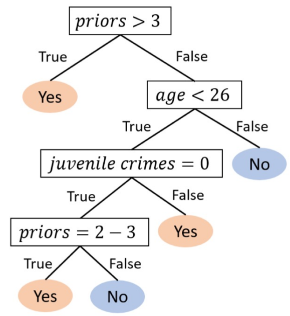
Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, Chudi Zhong

Interpretability in machine learning (ML) is crucial for high stakes decisions and troubleshooting. In this work, we provide fundamental principles for interpretable ML, and dispel common misunderstandings that dilute the importance of this crucial topic. We also identify 10 technical challenge areas in interpretable machine learning and provide history and background on each problem. Some of these problems are classically important, and some are recent problems that have arisen in the last few years. These problems are: (1) Optimizing sparse logical models such as decision trees; (2) Optimization of scoring systems; (3) Placing constraints into generalized additive models to encourage sparsity and better interpretability; (4) Modern case-based reasoning, including neural networks and matching for causal inference; (5) Complete supervised disentanglement of neural networks; (6) Complete or even partial unsupervised disentanglement of neural networks; (7) Dimensionality reduction for data visualization; (8) Machine learning models that can incorporate physics and other generative or causal constraints; (9) Characterization of the "Rashomon set" of good models; and (10) Interpretable reinforcement learning. This survey is suitable as a starting point for statisticians and computer scientists interested in working in interpretable machine learning.

10+ Grand Challenges

1. Sparse Logical Models: Decision Trees, Decision Lists, and Decision Sets



(complex interactions, multiclass tabular data, no addition required)

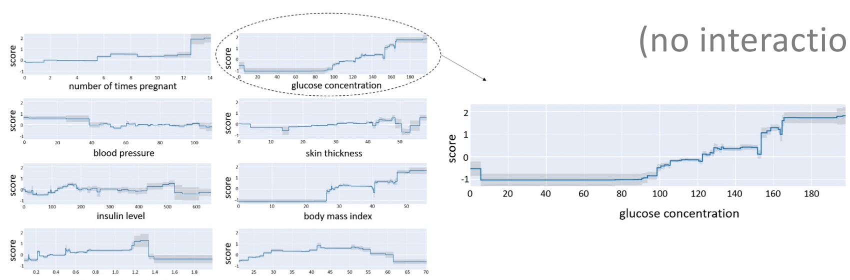
2. Scoring Systems

Patient screens positive for obstructive sleep apnea if Score >1			
1.	age ≥ 60	4 points	. . . . . .
2.	hypertension	4 points	$+$ . . . . . .
3.	body mass index ≥ 30	2 points	$+$ . . . . . .
4.	body mass index ≥ 40	2 points	$+$ . . . . . .
5.	female	-6 points	$+$ . . . . . .
	Add points from row 1-6	Score	$=$ . . . . . .

(no interactions, 2-class tabular data)

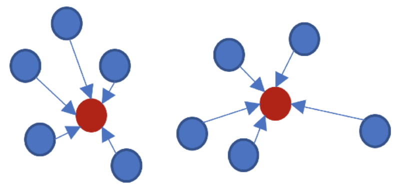
(no interactions, 2-class tabular data)

3. Generalized Additive Models

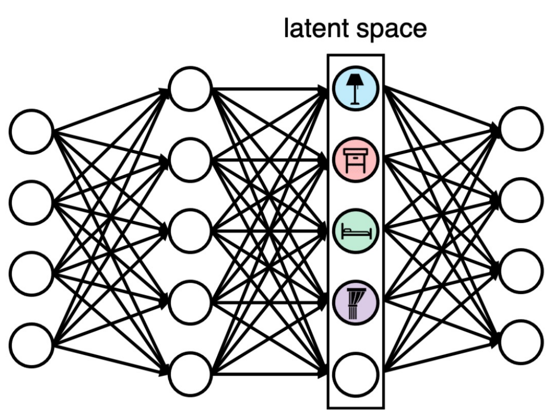


10+ Grand Challenges

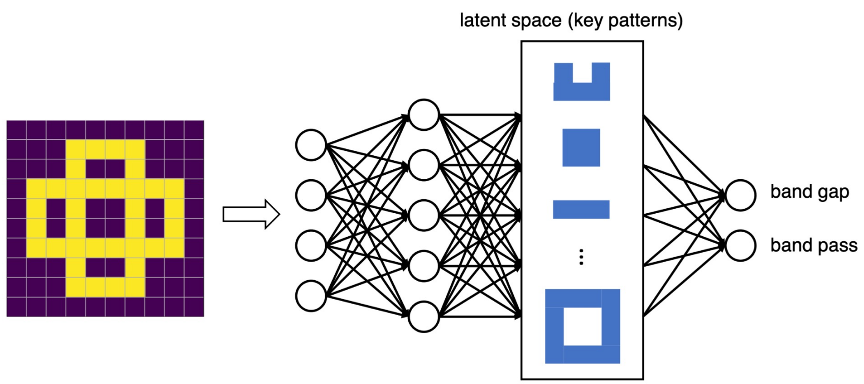
4. Case-Based Reasoning



5. Complete Supervised Disentanglement of Neural Networks



6. Unsupervised Disentanglement of Neural Networks

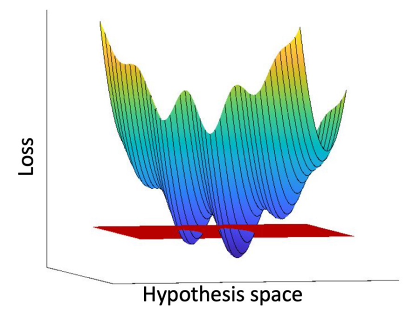


10+ Grand Challenges

7. Dimension Reduction for Data Visualization

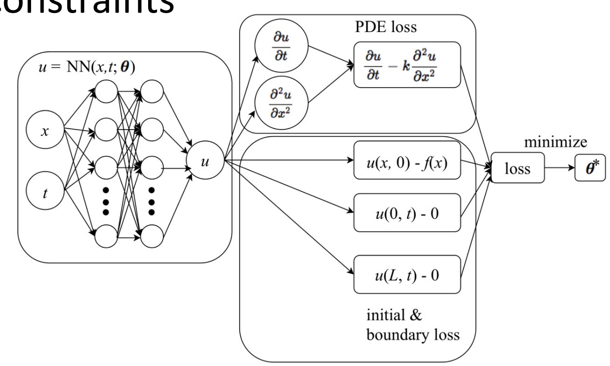
MNIST →

9. Characterization of the “Rashomon” set of good models

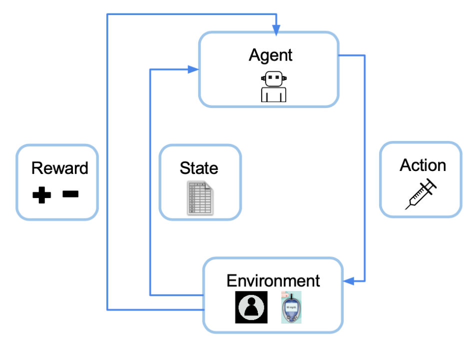


(a)

8. Machine learning models that incorporate physics and other generative or causal constraints

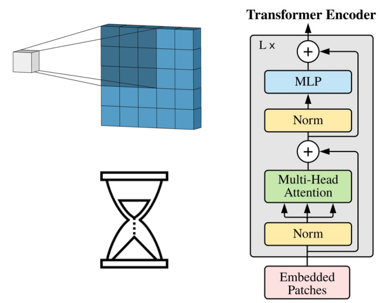


10. Interpretable Reinforcement Learning



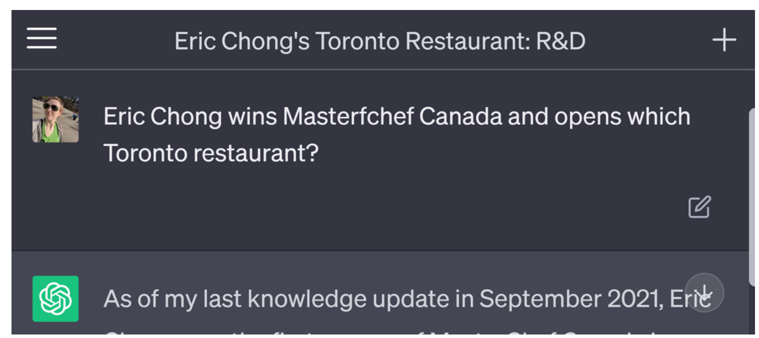
10+ Grand Challenges

11. Explanations styles that generalize



12. Explanations for Generative AI

13. Interpretability for NLP



arXiv.org > cs > arXiv:2103.11251

Search...

Help | Advanced S

Computer Science > Machine Learning

[Submitted on 20 Mar 2021]

Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, Chudi Zhong

Interpretability in machine learning (ML) is crucial for high stakes decisions and troubleshooting. In this work, we provide fundamental principles for interpretable ML, and dispel common misunderstandings that dilute the importance of this crucial topic. We also identify 10 technical challenge areas in interpretable machine learning and provide history and background on each problem. Some of these problems are classically important, and some are recent problems that have arisen in the last few years. These problems are: (1) Optimizing sparse logical models such as decision trees; (2) Optimization of scoring systems; (3) Placing constraints into generalized additive models to encourage sparsity and better interpretability; (4) Modern case-based reasoning, including neural networks and matching for causal inference; (5) Complete supervised disentanglement of neural networks; (6) Complete or even partial unsupervised disentanglement of neural networks; (7) Dimensionality reduction for data visualization; (8) Machine learning models that can incorporate physics and other generative or causal constraints; (9) Characterization of the "Rashomon set" of good models; and (10) Interpretable reinforcement learning. This survey is suitable as a starting point for statisticians and computer scientists interested in working in interpretable machine learning.

Principle 1

- Interpretable ML models are constrained.

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_i \text{Loss}(f, z_i) + C \cdot \text{InterpretabilityPenalty}(f), \quad \text{subject to} \quad \text{InterpretabilityConstraint}(f),$$

soft

"be sparse if it doesn't sacrifice accuracy"

hard

"be sparse"

- Should we rigorously/comprehensively/completely define interpretability in machine learning?

- Perhaps should rigorously define “predictive performance” first.
  - Accuracy, weighted accuracy, precision, average precision, precision@N, recall, recall@N, DCG, NCDG, AUC, partial AUC, mean-time-to-failure, exponential loss, logistic loss, …

- Better to ask what is often relevant: sparsity, linear or logical reasoning, visual comparisons, 1d or 2d functions, monotonicity, decomposability into sub-models, …

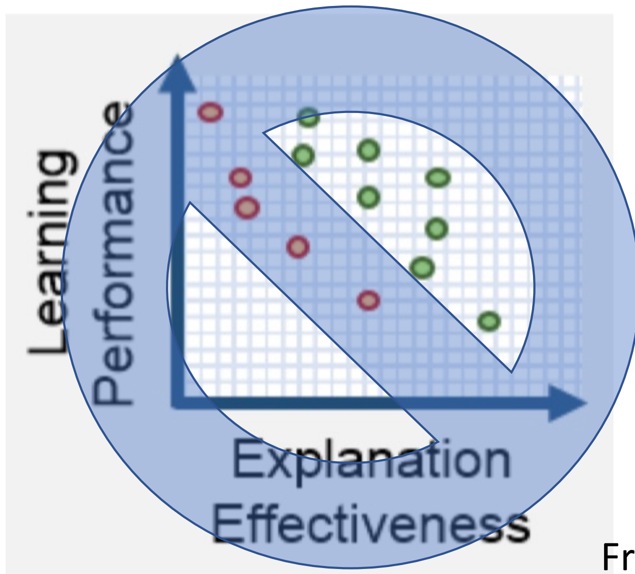
Principle 2

Despite common rhetoric, interpretable models do not necessarily create or enable trust -- they could also enable *distrust*.

They permit a decision of trust, rather than trust itself.

Principle 3

- Interpretability versus accuracy is, in general, a false dichotomy in machine learning.



From the DARPA XAI BAA, 2016

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

232

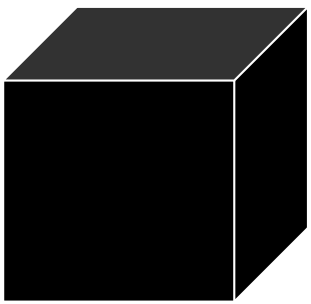
Glenn Rodriguez was denied parole because of a miscalculated “COMPAS” score.

How accurate is COMPAS? Data from Florida can tell us…

COMPAS vs. CORELS

COMPAS: (Correctional Offender Management Profiling for Alternative Sanctions)

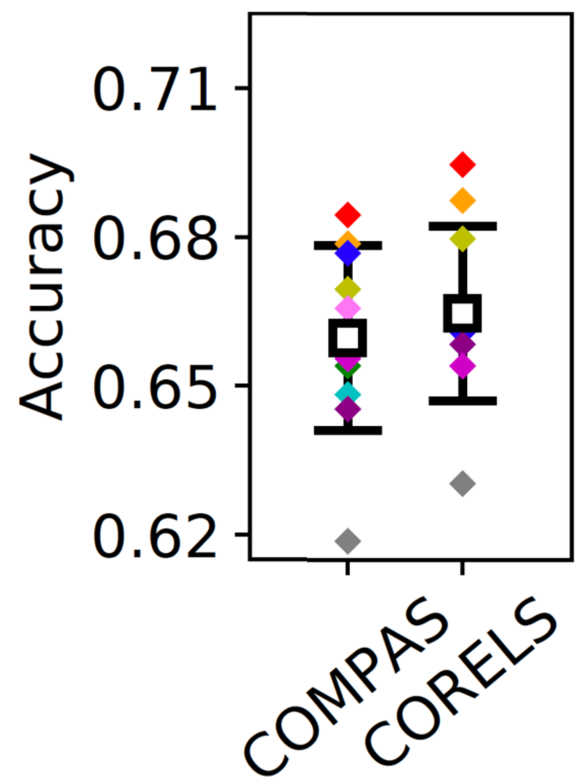
CORELS: (Certifiably Optimal RulE ListS, with Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, and Margo Seltzer, KDD 2017 & JMLR 2018)

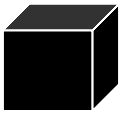


Here is the machine learning model:

If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

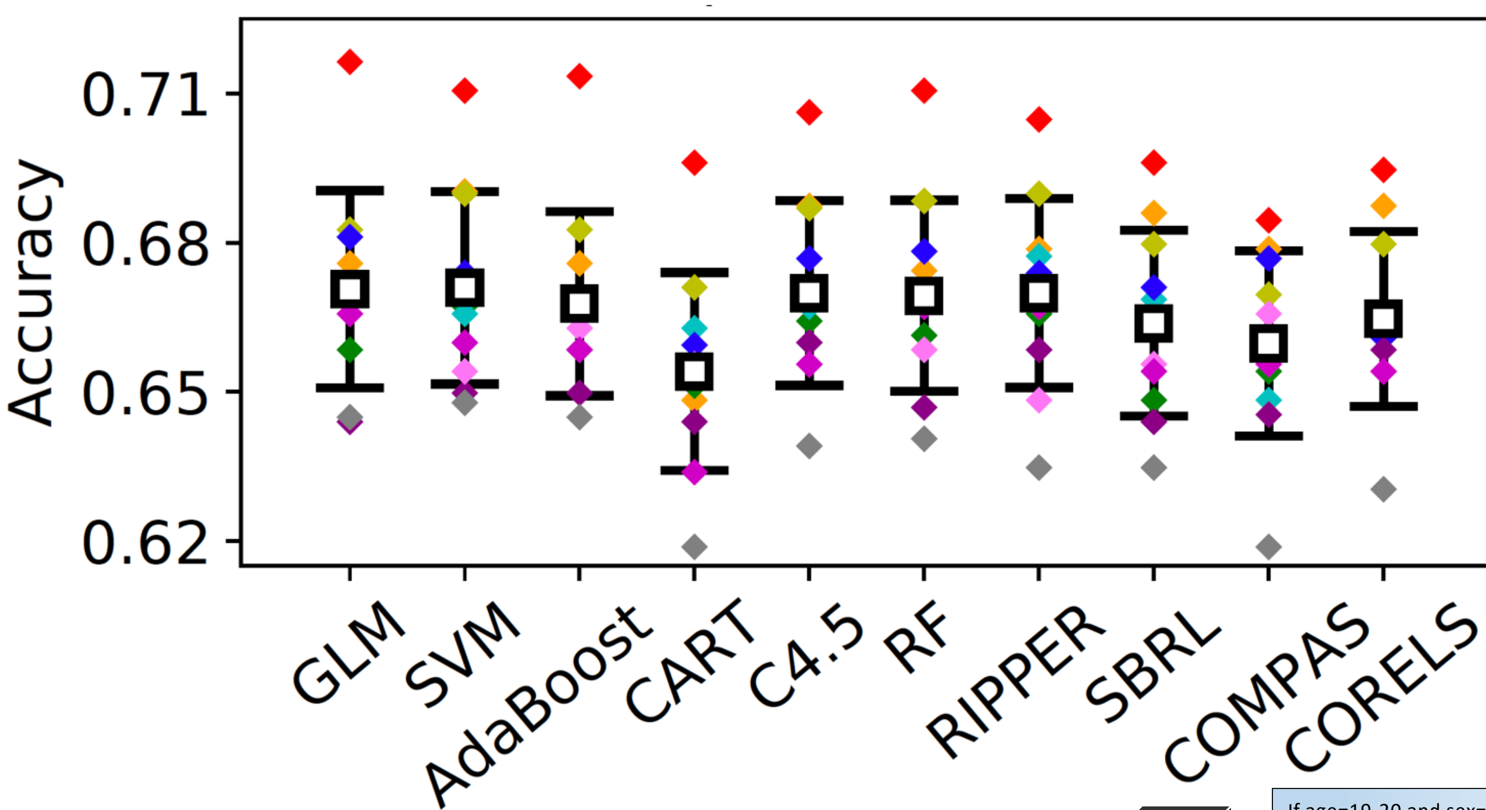
Prediction of re-arrest within 2 years

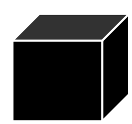




If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

Prediction of re-arrest within 2 years

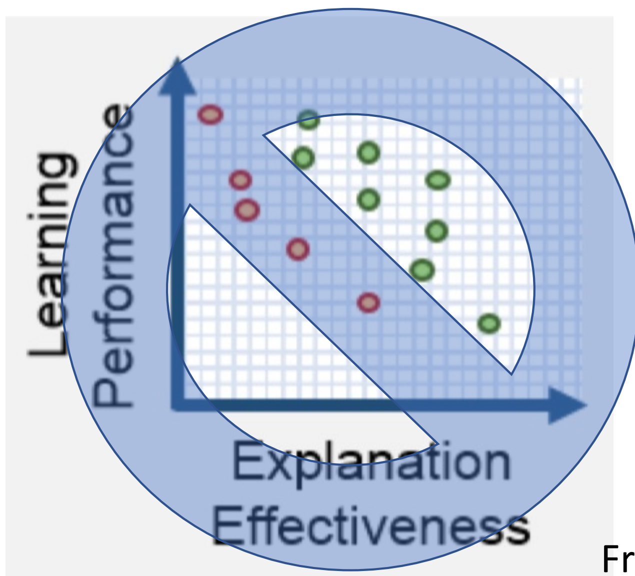




If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

Principle 3

- Interpretability versus accuracy is, in general, a false dichotomy in machine learning.

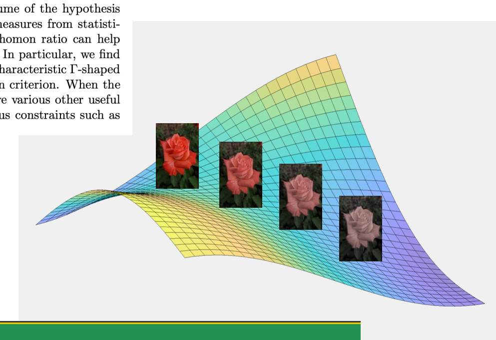


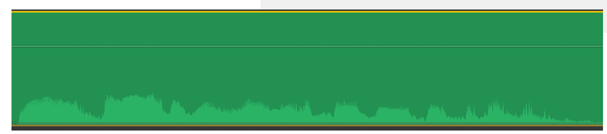
From the DARPA XAI BAA, 2016

Problem spectrum

age 45
congestive heart failure? yes
takes aspirin
smoking? no
gender M
exercise? yes
allergies? no
number of past strokes 2
diabetes? yes

The *Rashomon effect* occurs when many different explanations exist for the same phenomenon. In machine learning, Leo Breiman used this term to characterize problems where many accurate-but-different models exist to describe the same data. In this work, we study how the Rashomon effect can be useful for understanding the relationship between training and test performance, and the possibility that simple-yet-accurate models exist for many problems. We consider the *Rashomon set*—the set of almost-equally-accurate models for a given problem—and study its properties and the types of models it could contain. We present the *Rashomon ratio* as a new measure related to simplicity of model classes, which is the ratio of the volume of the set of accurate models to the volume of the hypothesis space; the Rashomon ratio is different from standard complexity measures from statistical learning theory. For a hierarchy of hypothesis spaces, the Rashomon ratio can help modelers to navigate the trade-off between simplicity and accuracy. In particular, we find empirically that a plot of empirical risk vs. Rashomon ratio forms a characteristic Γ-shaped *Rashomon curve*, whose elbow seems to be a reliable model selection criterion. When the Rashomon set is large, models that are accurate—but that also have various other useful properties—can often be obtained. These models might obey various constraints such as interpretability, fairness, or monotonicity.





Tabular: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

Problem spectrum

Very sparse models (trees, scoring systems)

With minor pre-processing, all methods have similar performance

Neural networks

Tabular: All features are interpretable

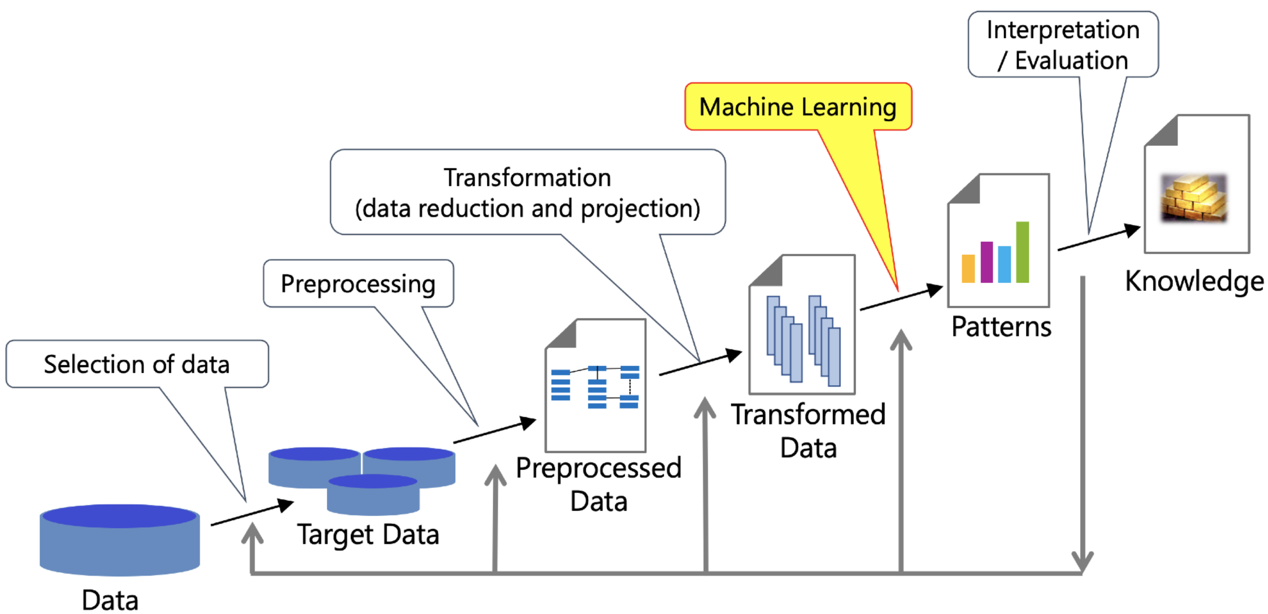
- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

Principle 4

- As part of the full data science process, one should expect both the performance metric and interpretability metric to be iteratively refined.



KDD Process, adapted from Fayyad et al., 1996

Principle 5

- For high stakes decisions, interpretable models should be used, if possible, rather than "explained" black box models.

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

There has been an increasing trend in healthcare and criminal justice to leverage machine learning (ML) for high-stakes prediction applications that deeply impact human lives. Many of not. There is a spectrum between fully transparent models (where we understand how all the variables are jointly related to each other) and models that are lightly constrained in model form (such as models

- Black box models *still* force you to trust the dataset.
- Double trouble: Forces you to rely on two models instead of one.

  Those models necessarily disagree with each other

  - An explanation that is right 90% of the time is wrong 10% of the time.
- The explanations are not really explanations, they don't use the same variables.

(Propublica scandal: They said COMPAS depends on age, criminal history, and *race*. But their analysis is wrong.)

- If you can produce an interpretable model, why explain black boxes? Do you really want to extend the authority of the black box?

Note

- LIME, SHAP, and Grad-CAM are tools that explain black box models. Not needed for interpretable models.

arXiv.org > cs > arXiv:2103.11251

Search...

Help | Advanced S

Computer Science > Machine Learning

[Submitted on 20 Mar 2021]

Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, Chudi Zhong

Interpretability in machine learning (ML) is crucial for high stakes decisions and troubleshooting. In this work, we provide fundamental principles for interpretable ML, and dispel common misunderstandings that dilute the importance of this crucial topic. We also identify 10 technical challenge areas in interpretable machine learning and provide history and background on each problem. Some of these problems are classically important, and some are recent problems that have arisen in the last few years. These problems are: (1) Optimizing sparse logical models such as decision trees; (2) Optimization of scoring systems; (3) Placing constraints into generalized additive models to encourage sparsity and better interpretability; (4) Modern case-based reasoning, including neural networks and matching for causal inference; (5) Complete supervised disentanglement of neural networks; (6) Complete or even partial unsupervised disentanglement of neural networks; (7) Dimensionality reduction for data visualization; (8) Machine learning models that can incorporate physics and other generative or causal constraints; (9) Characterization of the "Rashomon set" of good models; and (10) Interpretable reinforcement learning. This survey is suitable as a starting point for statisticians and computer scientists interested in working in interpretable machine learning.

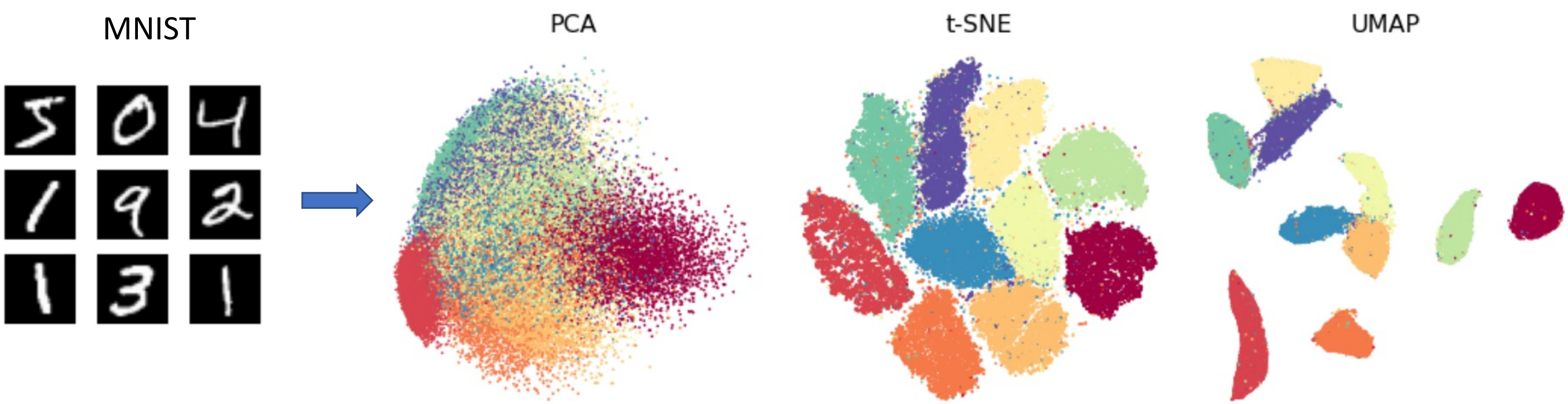
Note: I will focus on topics that I know well, because I work on them.

Start with exploratory data analysis.

7. Dimension reduction for data visualization

$$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \ldots \longrightarrow \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \ldots$$

d dimensions — 2 or 3 dimensions



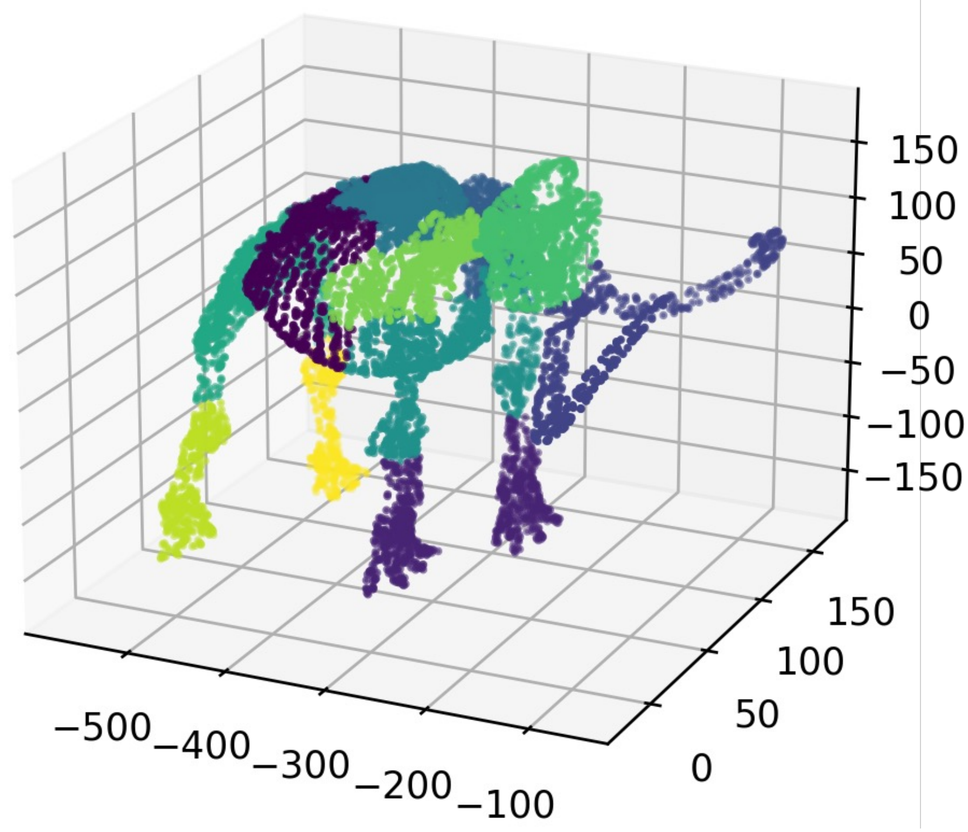
7. Dimension reduction for data visualization

Dimension reduction methods:

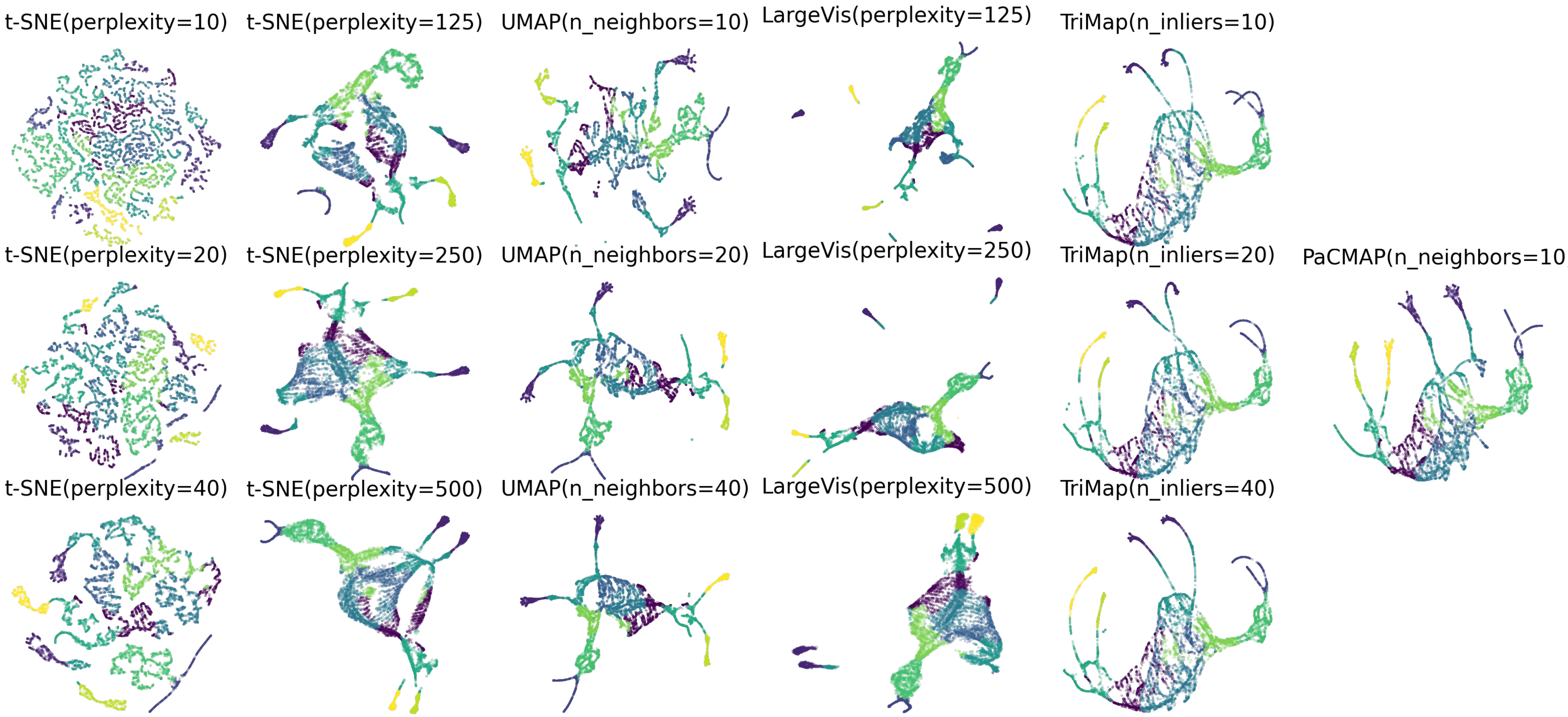
- can illuminate patterns in high dimensional data
- used often in biology
- PCA is the quintessential DR algorithm

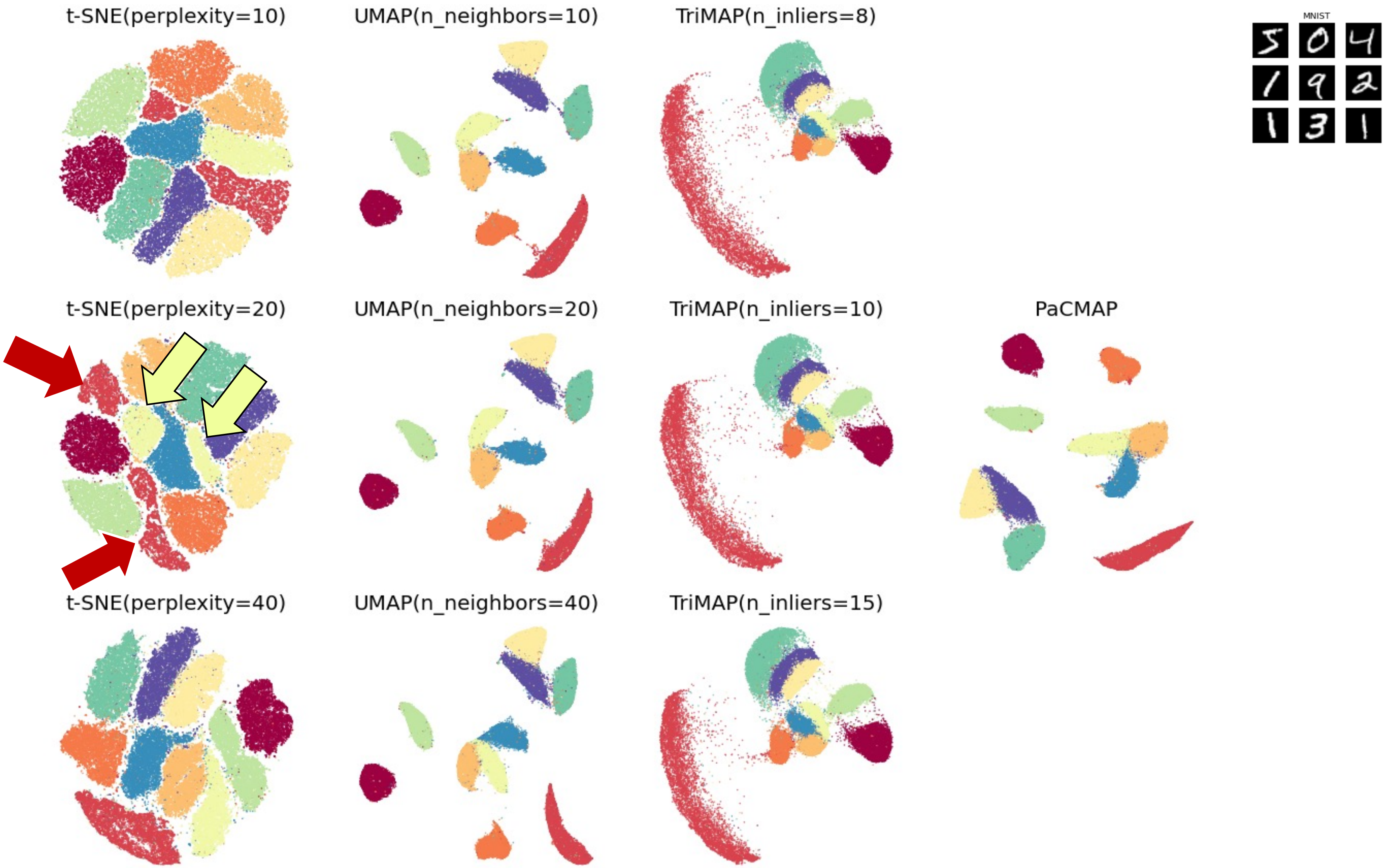
- unsupervised, so no ground truth
- sometimes wildly different results between methods
- dimension reduction plots often lack global structure

Preserve the Mammoth!



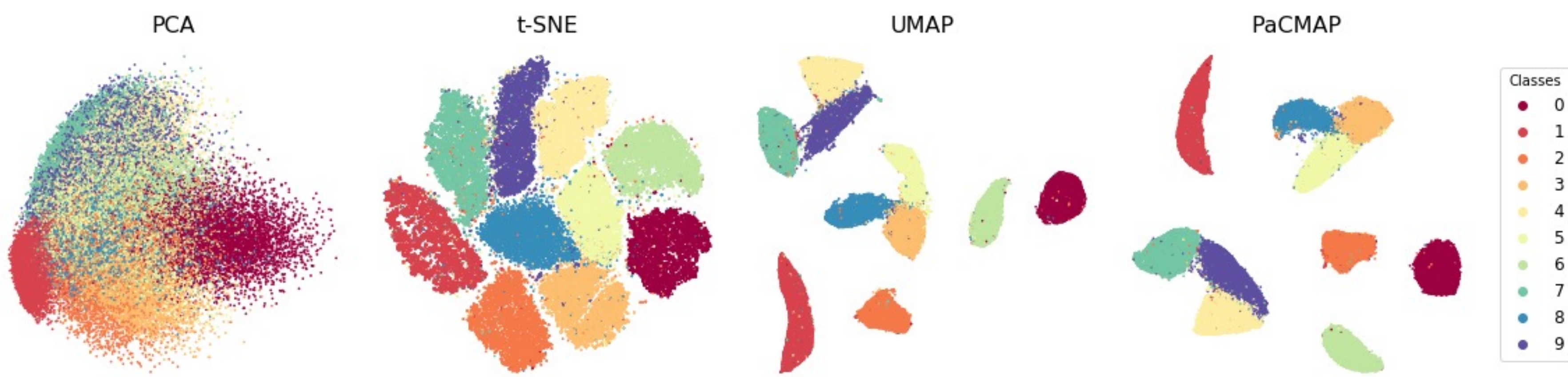
Task: 3d to 2d.


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


t-SNE(perplexity=10)
UMAP(n_neighbors=10)
TriMAP(n_inliers=8)
MNIST
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)

Local vs Global

- Local structure: local neighborhood graph, nearest neighbors
- Global structure: relationships between clusters, respect relative distances between points in high-dimensional space.



(mainly global) (mainly local) (mainly local) (both, actually)

Global Methods

- PCA (Pearson, 1901)
- MDS (Torgerson, 1952)

:

Local Methods

- LLE (Roweis and Saul, 2000),
- Isomap (Tenenbaum et al., 2000)
- Hessian Local Linear Embedding (Donoho and Grimes, 2003)
- Laplacian Eigenmaps (Belkin and Niyogi, 2001)
- Stochastic Neighborhood Embedding (SNE) (Hinton and Roweis, 2003)
- t-SNE (van der Maaten and Hinton, 2008)
- LargeVis (Tang et al., 2016)
- UMAP (McInnes et al., 2018)

:

- PacMAP is both local and global.

Preserve distances, not neighborhoods

Crowding problem

Preserve neighborhoods

Global Methods

- PCA (Pearson, 1901)
- MDS (Torgerson, 1952)

:

Local Methods

- LLE (Roweis and Saul, 2000),
- Isomap (Tenenbaum et al., 2000)
- Hessian Local Linear Embedding (Donoho
- Laplacian Eigenmaps
- Stochastic Neighbor Roweis, 2003)
- t-SNE (van der Maat
- LargeVis (Tang et al.,
- UMAP (McInnes et al., 2018)

:

- PacMAP is both local and glo

Article | Open Access | Published: 28 November 2019

The art of using t-SNE for single-cell transcriptomics

Dmitry Kobak & Philipp Berens

Nature Communications **10**, Article number: 5416 (2019) | Cite this article

How to Use t-SNE Effectively

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

arXiv.org > cs > arXiv:1708.03229

Search...

Help | Advanced S

Computer Science > Artificial Intelligence

[Submitted on 10 Aug 2017]

Automatic Selection of t-SNE Perplexity

Yanshuai Cao, Luyu Wang

t-Distributed Stoch
dimensionality redu
hyperparameter tha

Automated optimal parameters for T-distributed stochastic neighbor embedding improve visualization and allow analysis of large datasets

October 2018

DOI: 10.1101/451690

Project: Automated Analysis of Flow Cytometry Multidimensional Datasets

Authors:

Anna C Belkina Boston University

Christopher O. Ciccolella

Rina Anno

Josef Spidlen

JMLR

Home Page

Papers

Submissions

News

Editorial Board

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; 22(201):1−73, 2021.

Abstract

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 goal is to understand what aspects of DR methods are important for preserving both local and global structure: it is difficult to design

Search or jump to... / Pulls Issues Codespaces Marketplace Explore

YingfanWang / PaCMAP Public

Unwatch 10 | Fork 34 | Starred 272

*Winner of the **2023 John M. Chambers Statistical Software Award** from the American Statistical Association

communications biology

Explore content | About the journal | Publish with us

nature > communications biology > articles > article

Article | Open Access | Published: 19 July 2022

Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization

Haiyang Huang, Yingfan Wang, Cynthia Rudin & Edward P. Browne

Communications Biology **5**, Article number: 719 (2022) | Cite this article

154 Accesses | **2** Altmetric | Metrics

Algorithm	Graph components and Loss function
t-SNE (van der Maaten & Hinton, 2008)	Graph components: Edges (i,j) $\text{Loss}_{i,j}^{\text{t-SNE}} = p_{ij} \log \frac{p_{ij}}{q_{ij}}$, where $q_{ij} = \frac{\left(1+\|\mathbf{y}_i-\mathbf{y}_j\|^2\right)^{-1}}{\sum_{k\neq l}(1+\|\mathbf{y}_k-\mathbf{y}_l\|^2)^{-1}}$ where p_{ij} is a function of $\mathbf{x}_i$, $\mathbf{x}_j$ and other $\mathbf{x}_\ell$'s.
UMAP (McInnes et al., 2018)	Graph components: Edges (i,j) $\text{Loss}_{i,j}^{\text{UMAP}} = \begin{cases} \bar{w}_{i,j} \log \left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1} & i,j \text{ neighbors} \\ (1-\bar{w}_{i,j}) \log \left(1 - \left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1}\right) & \text{otherwise,} \end{cases}$ where $\bar{w}_{i,j}$ is a function of $\mathbf{x}_i$, $\mathbf{x}_j$ and nearby $\mathbf{x}_\ell$'s.
TriMAP (Amid & Warmuth, 2019)	Graph components: Triplets (i,j,k) where $\text{Distance}_{i,j} \leq \text{Distance}_{i,k}$ $\text{Loss}_{i,j,k}^{\text{TM}} = \omega_{i,j,k} \frac{s(\mathbf{y}_i,\mathbf{y}_k)}{s(\mathbf{y}_i,\mathbf{y}_j)+s(\mathbf{y}_i,\mathbf{y}_k)}$, where $s(\mathbf{y}_i,\mathbf{y}_j) = \left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}$ and $\omega_{i,j,k}$ is a function of $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_k$ and nearby points.

Hard to understand what's important here…

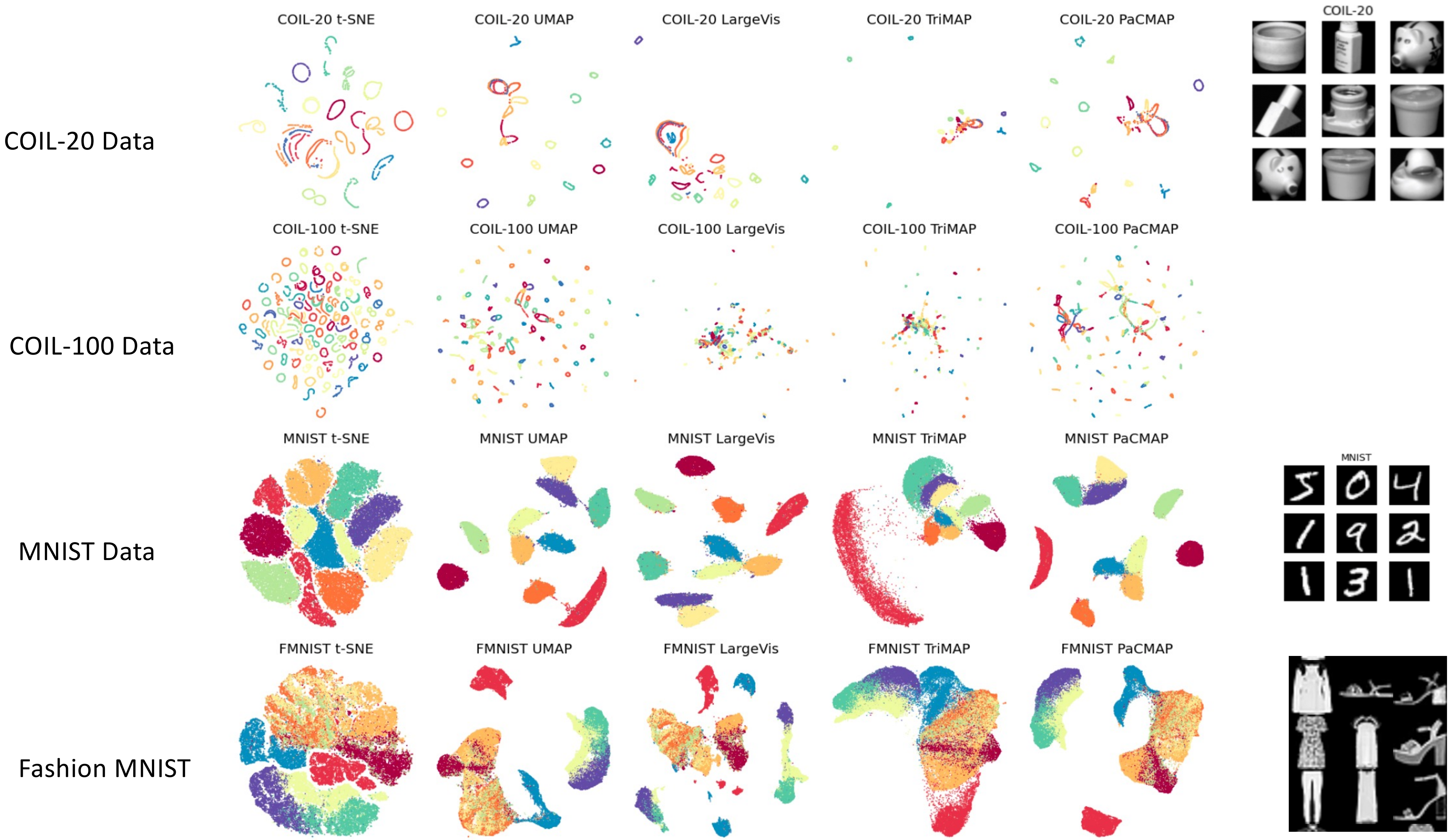
$$\sum_{\text{Subset of graph components } \{i\}} \text{Weight}^{\mathbf{X}}(\mathcal{C}_i^H) \cdot \text{Loss}^{\mathbf{Y}}(\mathcal{C}_i^L)$$

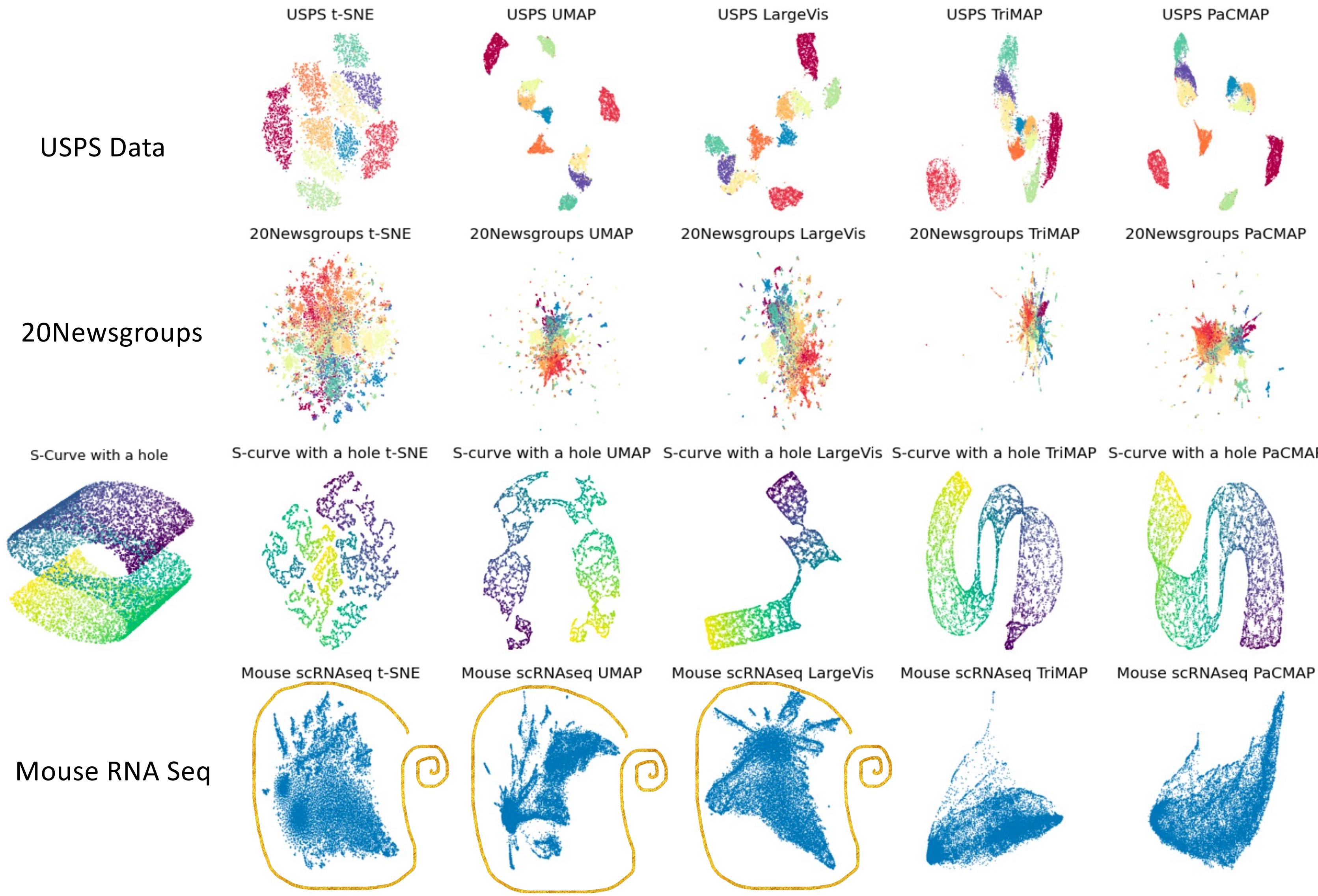
After a huge amount of experimentation, we found that:

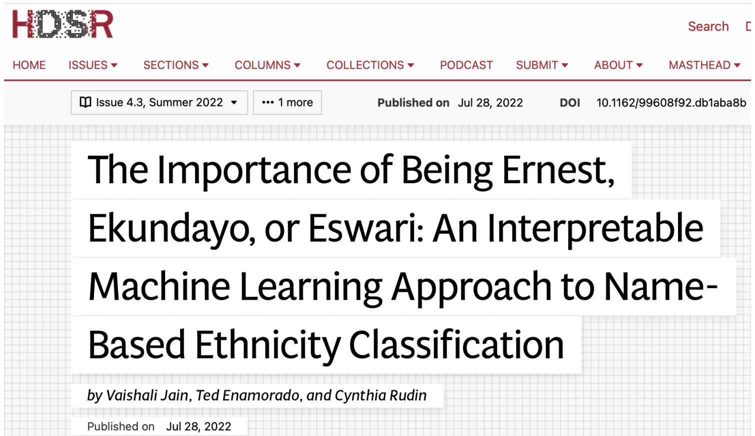
- Certain specific properties of the loss function are important for local structure.
- The choice of which graph components to exert forces on is important for global structure.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, Yaron Shaposhnik. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization. Journal of Machine Learning Research, 2021.

Some demos


COIL-20 Data
COIL-100 Data
MNIST Data
Fashion MNIST
COIL-20 t-SNE
COIL-20 UMAP
COIL-20 LargeVis
COIL-20 TriMAP
COIL-20 PaCMAP
COIL-20
COIL-100 t-SNE
COIL-100 UMAP
COIL-100 LargeVis
COIL-100 TriMAP
COIL-100 PaCMAP
MNIST t-SNE
MNIST UMAP
MNIST LargeVis
MNIST TriMAP
MNIST PaCMAP
MNIST
FMNIST t-SNE
FMNIST UMAP
FMNIST LargeVis
FMNIST TriMAP
FMNIST PaCMAP


USPS Data
USPS t-SNE
USPS UMAP
USPS LargeVis
USPS TriMAP
USPS PaCMAP
20Newsgroups
20Newsgroups t-SNE
20Newsgroups UMAP
20Newsgroups LargeVis
20Newsgroups TriMAP
20Newsgroups PaCMAP
S-Curve with a hole
S-curve with a hole t-SNE
S-curve with a hole UMAP
S-curve with a hole LargeVis
S-curve with a hole TriMAP
S-curve with a hole PaCMAP
Mouse RNA Seq
Mouse scRNAseq t-SNE
Mouse scRNAseq UMAP
Mouse scRNAseq LargeVis
Mouse scRNAseq TriMAP
Mouse scRNAseq PaCMAP


HDSR

Search

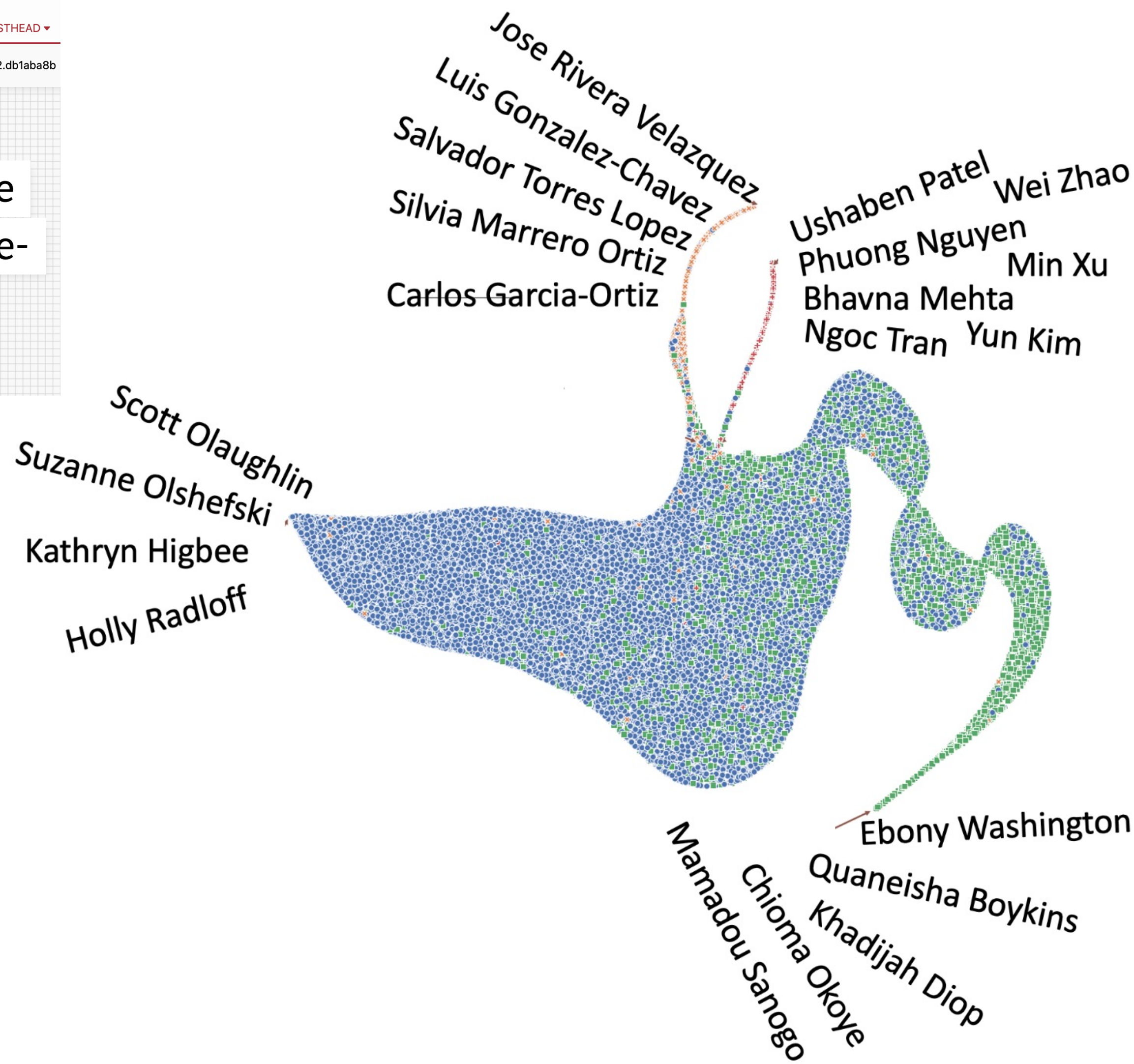
HOME ISSUES SECTIONS COLUMNS COLLECTIONS PODCAST SUBMIT ABOUT MASTHEAD

Issue 4.3, Summer 2022 ··· 1 more Published on Jul 28, 2022 DOI 10.1162/99608f92.db1aba8b

The Importance of Being Ernest, Ekundayo, or Eswari: An Interpretable Machine Learning Approach to Name-Based Ethnicity Classification

by Vaishali Jain, Ted Enamorado, and Cynthia Rudin

Published on Jul 28, 2022



EthnicIA

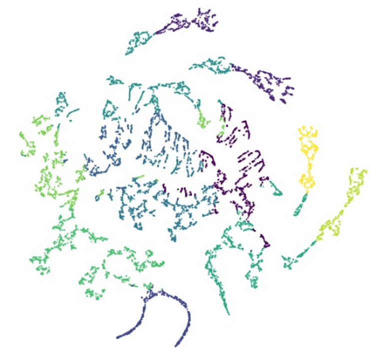
Challenges for DR

- Scalability – Huge datasets
- Global structure isn't perfect (still! Is it possible that there are multiple equally good DR plots?)
- Interacting with DR plots to find out more about the data

Take-Aways on Dimension Reduction

- DR algorithms help you see into high-dimensional data.
- They cannot always be trusted.
- PacMAP takes advantage of separate ways to preserve local and global structure.
- Evaluation metrics for DR are listed in our paper.

t-SNE (Mammoth)

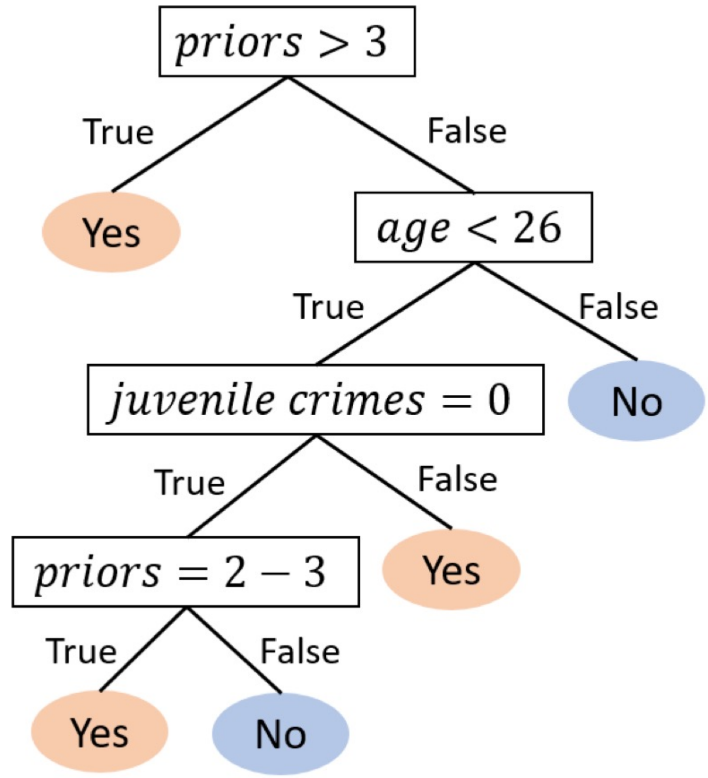


PaCMAP (Mammoth)

Stop here for ≤2 questions

5 principles,
Grand challenges:
- 7 DR
- 1 Logical models

1. Sparse Logical Models: Decision Trees, Decision Lists, and Decision Sets



1. Sparse Logical Models: Decision Trees, Decision Lists, and Decision Sets

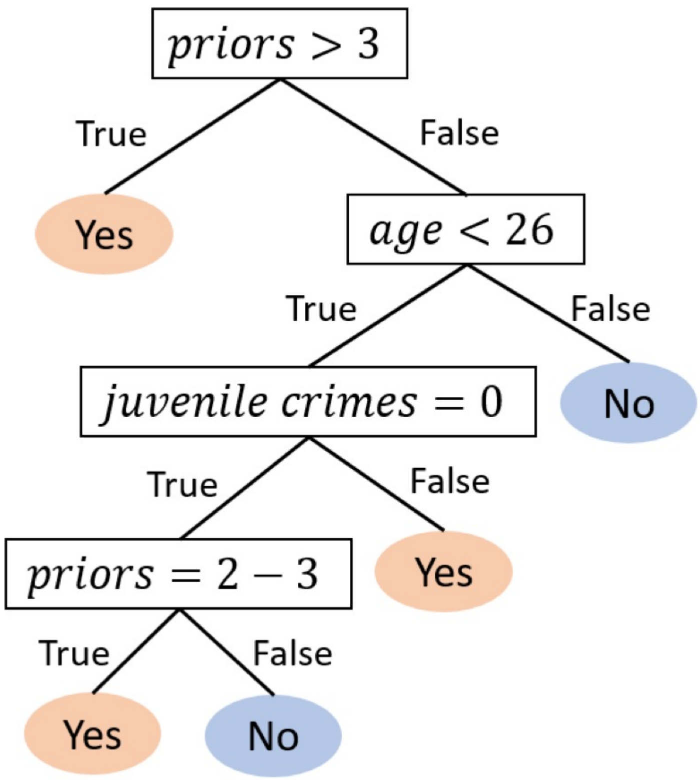
Logical models:

- arose from expert systems, first algorithms ~1960's
- are nonlinear and powerful
- are robust to outliers
- handle missing data well
- easily handle multiclass

- non-smooth
- hard to optimize

If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

CORELS (Angelino et al., JMLR, 2018)



GOSDT (Lin et al. ICML, 2020)

IF user:

goes to coffee houses $\geq$ once per month
AND destination $\neq$ Urgent Place AND Passenger $\neq$ Kids

OR goes to coffee houses $\geq$ once per month
AND Coupon expires in one day

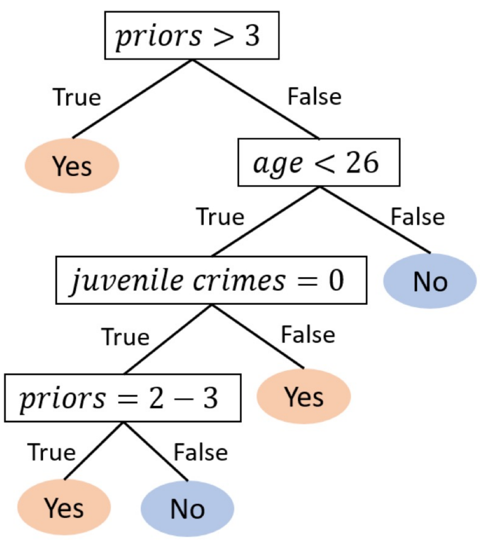
Then predict user will claim the coupon.

Bayesian Rule Sets, Wang et al., JMLR 2017

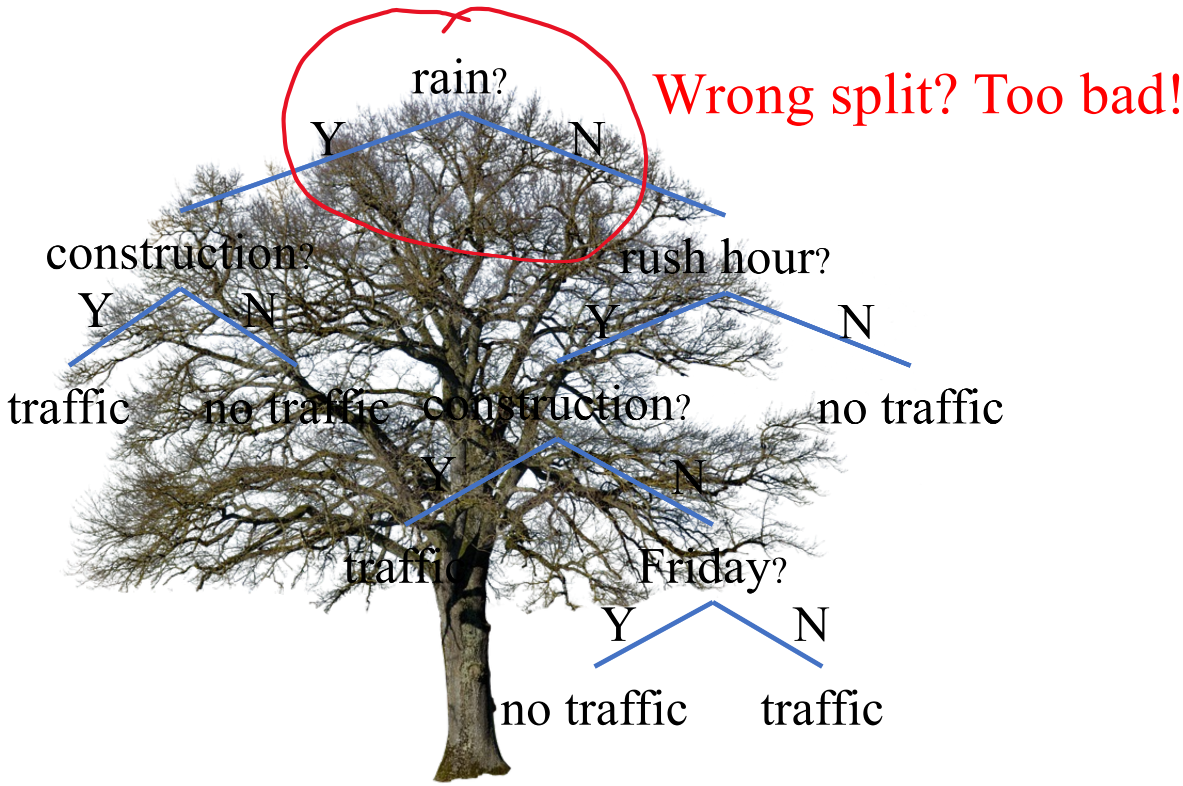
Optimal Sparse Decision Trees

$$\min_{f \in \text{ set of trees}} \frac{1}{n} \sum_i \text{Loss}(f, z_i) + C \cdot \textbf{Number of leaves } (f),$$

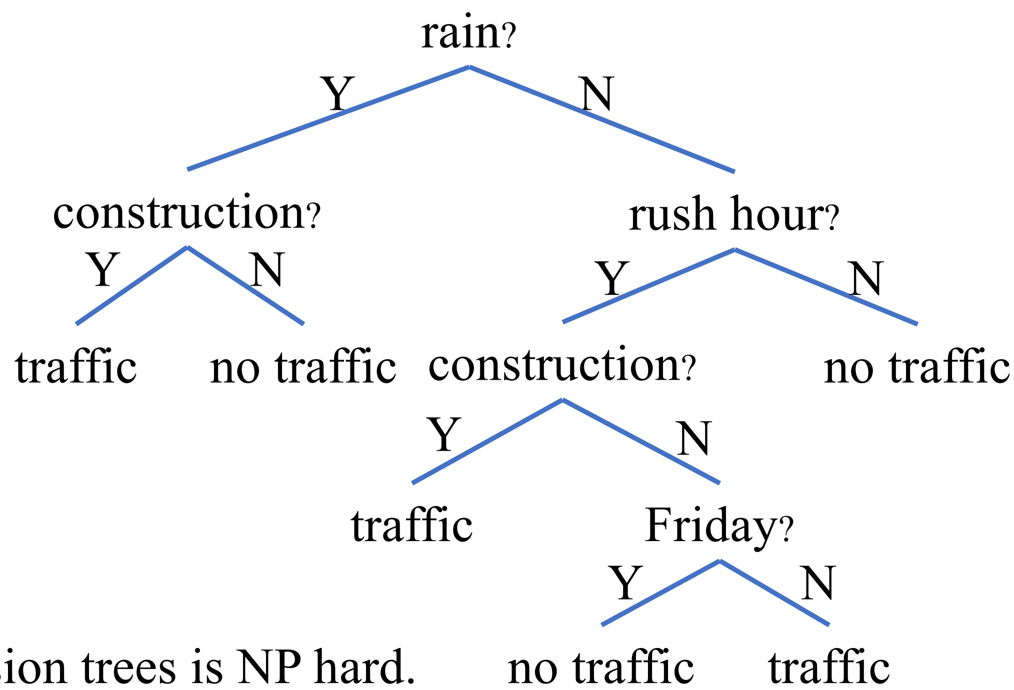
$$\textbf{Depth}(f) \leq D$$



Optimal Sparse Decision Trees



THeta Automatic Interaction Detection (THAID) (Messenger & Mandell, 1972)



Optimal sparse decision trees is NP hard.
Factorial in the number of variables.

Approaches for optimal sparse trees that are not greedy:

- Genetic Programming (e.g., Fan & Gray, 2005, Janikow & Malatkar, 2011), Neural Networks (Zantedeschi et al, 2020), no optimality gap

- Mathematical Programming Solvers, SAT solvers (Bennett mid-1990's,.., Blanquero et al., 2018, 2020, Menickelly et al., 2018; Vilas Boas et al., 2019, Verwer & Zhang BinOCT, 2019, Aghaei et al., 2021, Gunluk et al., 2021,..)

- Dynamic Programming / Branch and Bound
  - Garofalakis et al., DTC, 2003
  - Nijssen & Fromont, DL8, 2007, 2010, Aglin et al., DL8.5, 2020, Demirovic et al., 2022
  - Angelino et al, CORELS, 2018, Hu et al., OSDT 2019, Lin et al., GOSDT, 2020, McTavish et al. 2022

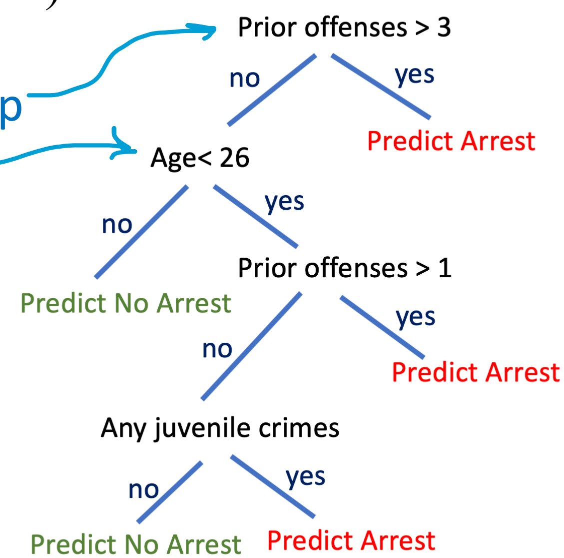
Generalized Optimal Sparse Decision Trees (GOSDT)

To figure out the optimal split at the top

Figure out the optimal split beneath it.

And the one below that.

Which eventually is a leaf.



GOSDT + Guesses (McTavish et al., AAAI 2022)

"Guessing" techniques improve speed without losing performance:

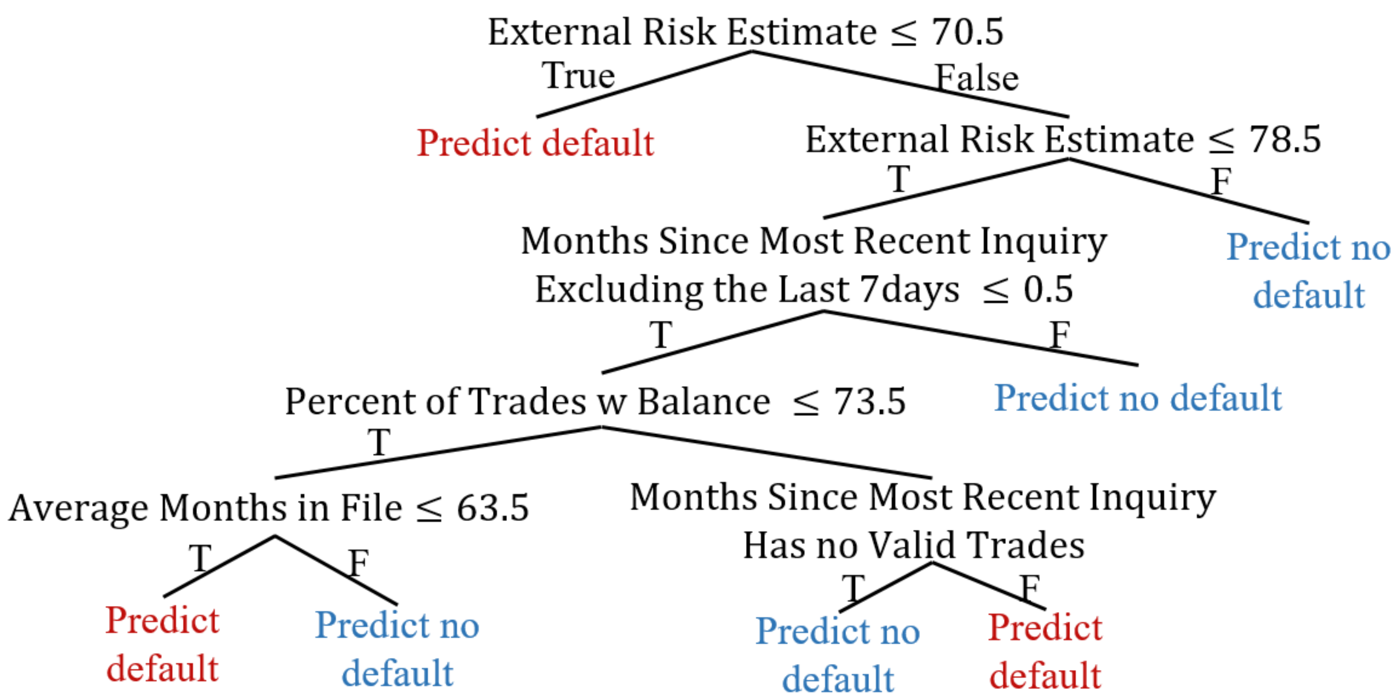
- Guess the depth. Don't search below that.

$$\mathrm{Depth}(f) \leq D$$

- Use a black box model to "guess" a lower bound on the optimal loss.
  Use it to prune parts of the search space.

- Use a random forest or boosted tree, only use its splits for the GOSDT tree.

Explainable ML Challenge (FICO dataset) tree:



- 10K data points, >1900 binary features
- training & test accuracy 72% (best black box is 73%)
- 7 leaves
- 8.1 sec

Challenges that were solved recently

Can we create trees almost as fast as CART/C4.5 create greedy trees? (Handled by GOSDT)

Can we efficiently handle continuous input variables in optimal decision trees? (Handled by "Guessing")

Can we handle constraints more gracefully? (Handled by Rashomon set work, coming up)

Can we do regression with sparse trees? (Solved by OSRT algorithm, AAAI 2023)

New Challenges: Multivariate regression. (See Jeff Simonoff). Combining trees.

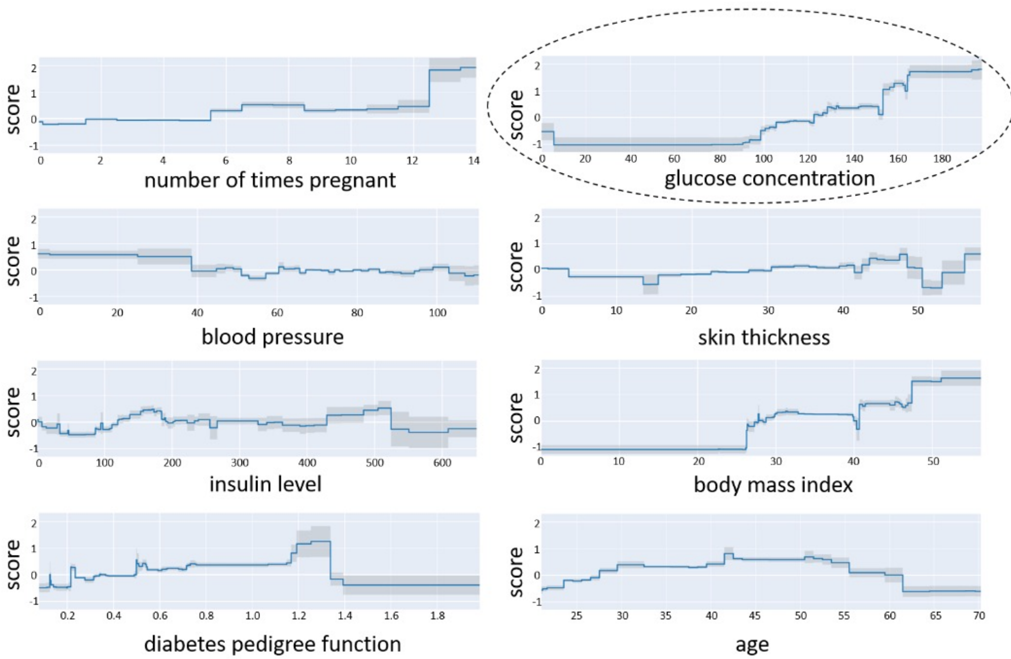
Note that code is public for GOSDT. (pip install GOSDT)

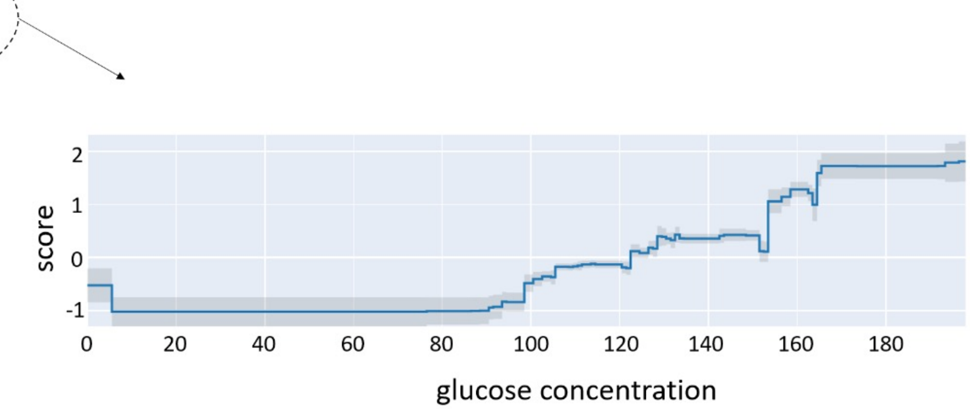
Stop here for ≤2 questions

5 principles,
Grand challenges:
- 7 DR
- 1 Logical models
- 3 GAMs

3. Generalized Additive Models

$$f(\mathbf{x}) = \sum_j f_j(x_j)$$





Credit: Slides of Rich Caruana

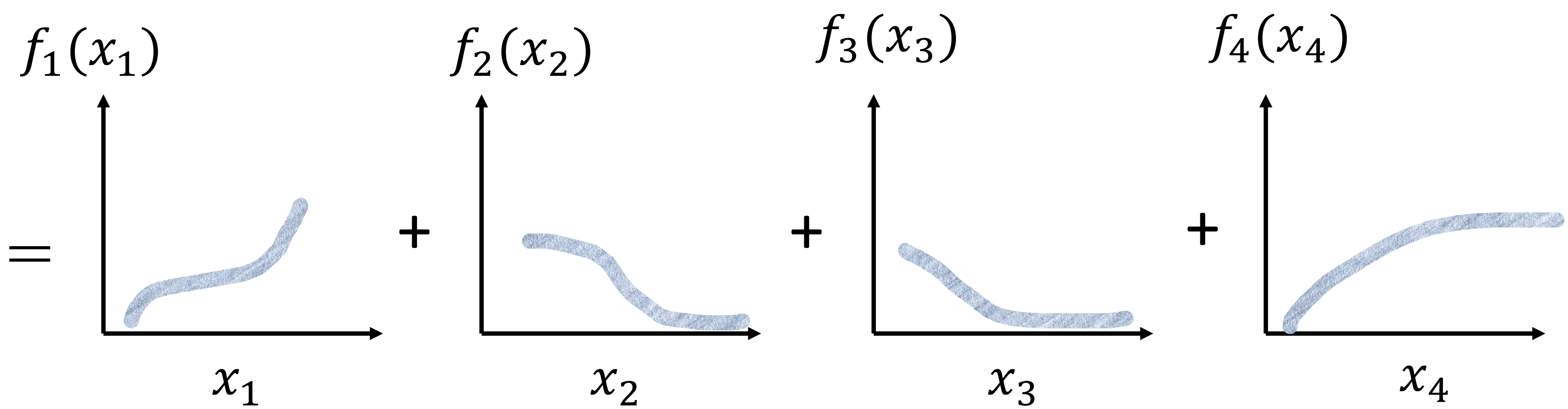
3. Generalized additive models (GAMs)

GAMs:
- very powerful, nonlinear
- uses visualization to convey contributions from each feature
- Can be trained using boosting or other ML techniques

- generally, few interaction terms
- doesn’t easily handle missing data or multiclass
- great for continuous features, not good for categorical features

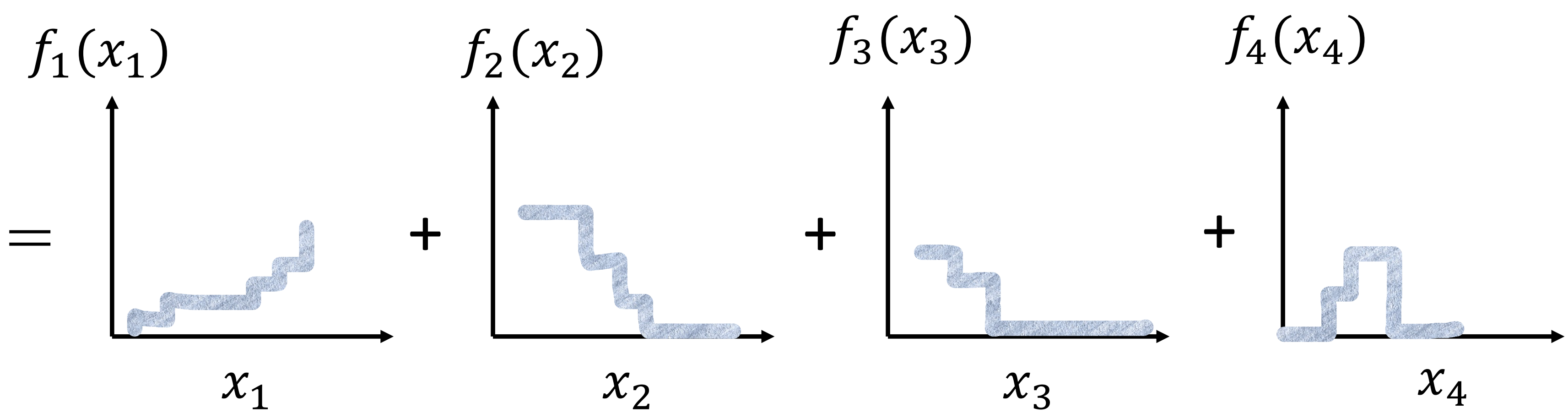
Additive Model

$$\hat{y}(x) \propto \sum_{j=1}^{p} f_j(x_j)$$



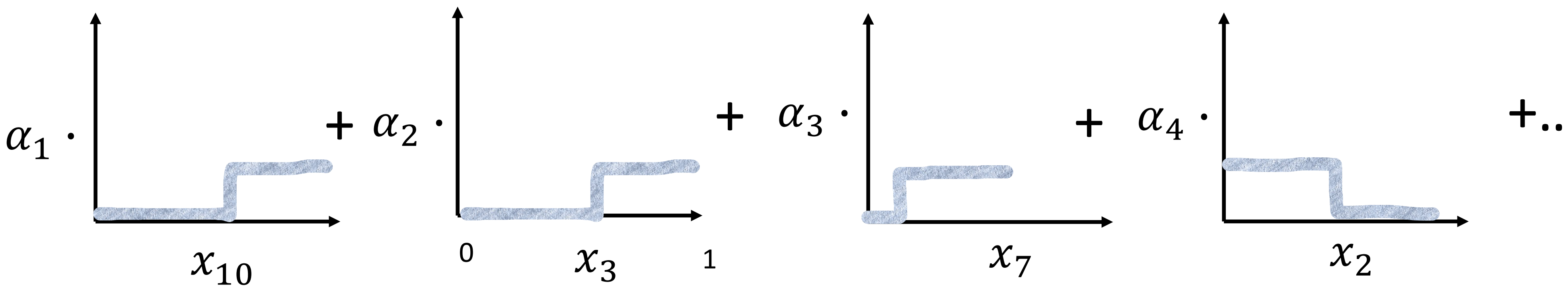
Additive Model

$$\hat{y}(x) \propto \sum_{j=1}^{p} f_j(x_j)$$



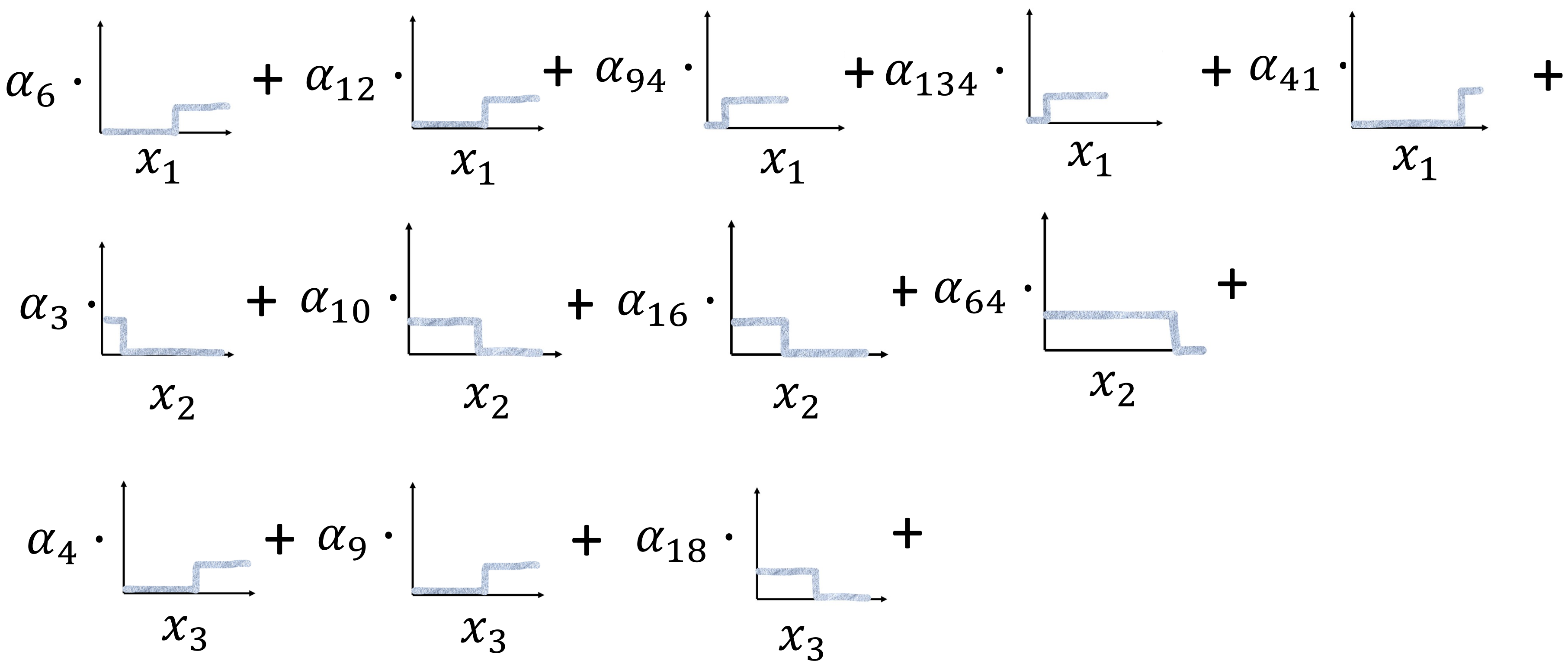
Boosted Stumps

At each iteration, the algorithm picks a feature and a threshold.



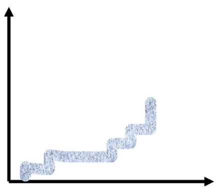
Boosted Stumps

Sort the stumps by feature

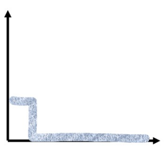
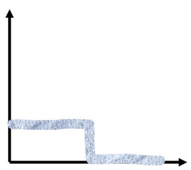
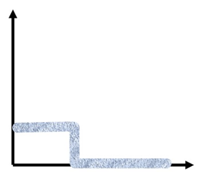
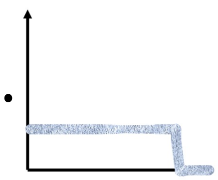


Boosted Stumps

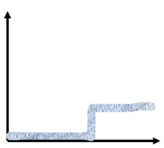
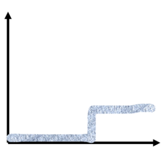
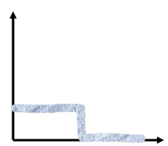
Sort the stumps by feature



x_1

$\alpha_3 \cdot$  $+ \; \alpha_{10} \cdot$  $+ \; \alpha_{16} \cdot$  $+ \; \alpha_{64} \cdot$  $+$

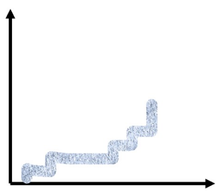
x_2 x_2 x_2 x_2

$\alpha_4 \cdot$  $+ \; \alpha_9 \cdot$  $+ \; \alpha_{18} \cdot$  $+$

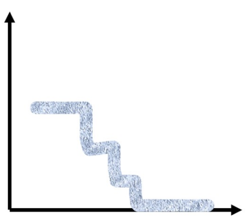
x_3 x_3 x_3

Boosted Stumps

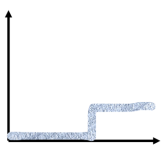
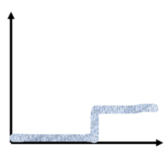
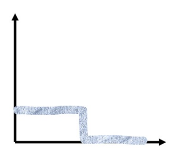
Sort the stumps by feature



x_1

$+$ 

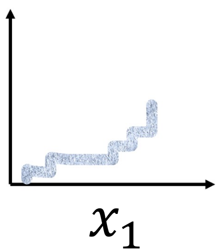
x_2

$\alpha_4 \cdot$  $+ \; \alpha_9 \cdot$  $+ \; \alpha_{18} \cdot$  $+$

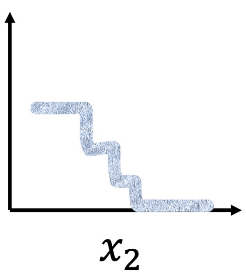
x_3 x_3 x_3

Boosted Stumps

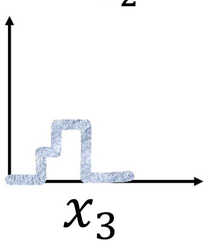
Sort the stumps by feature



$f_1(x_1)$ ← increases

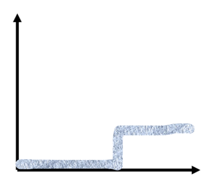
\+ 

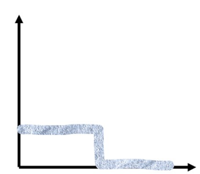
$f_2(x_2)$ ← decreases

\+ 

$f_3(x_3)$

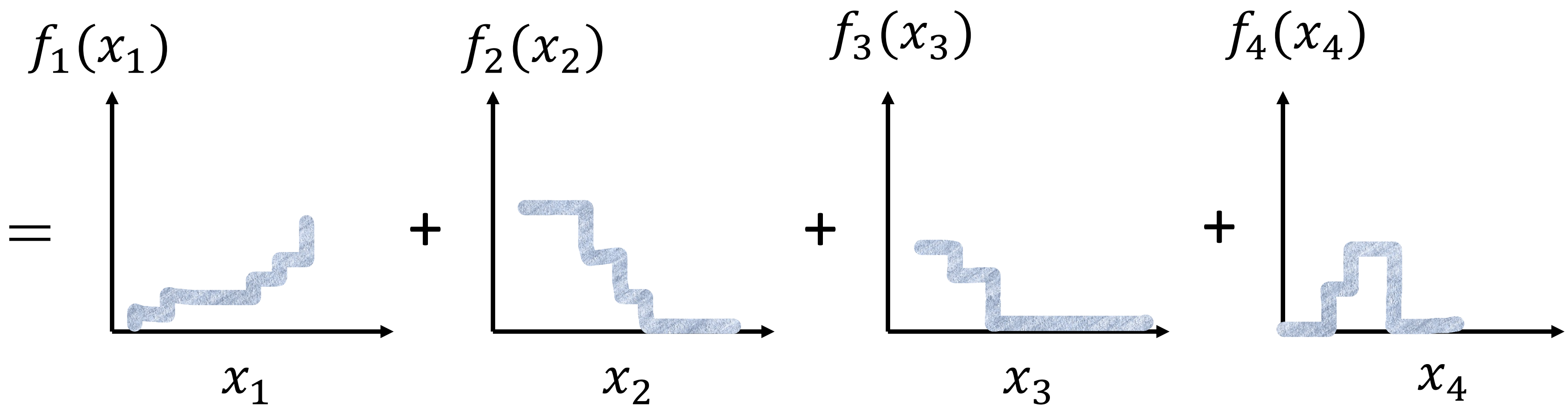
Use step functions that face different ways.

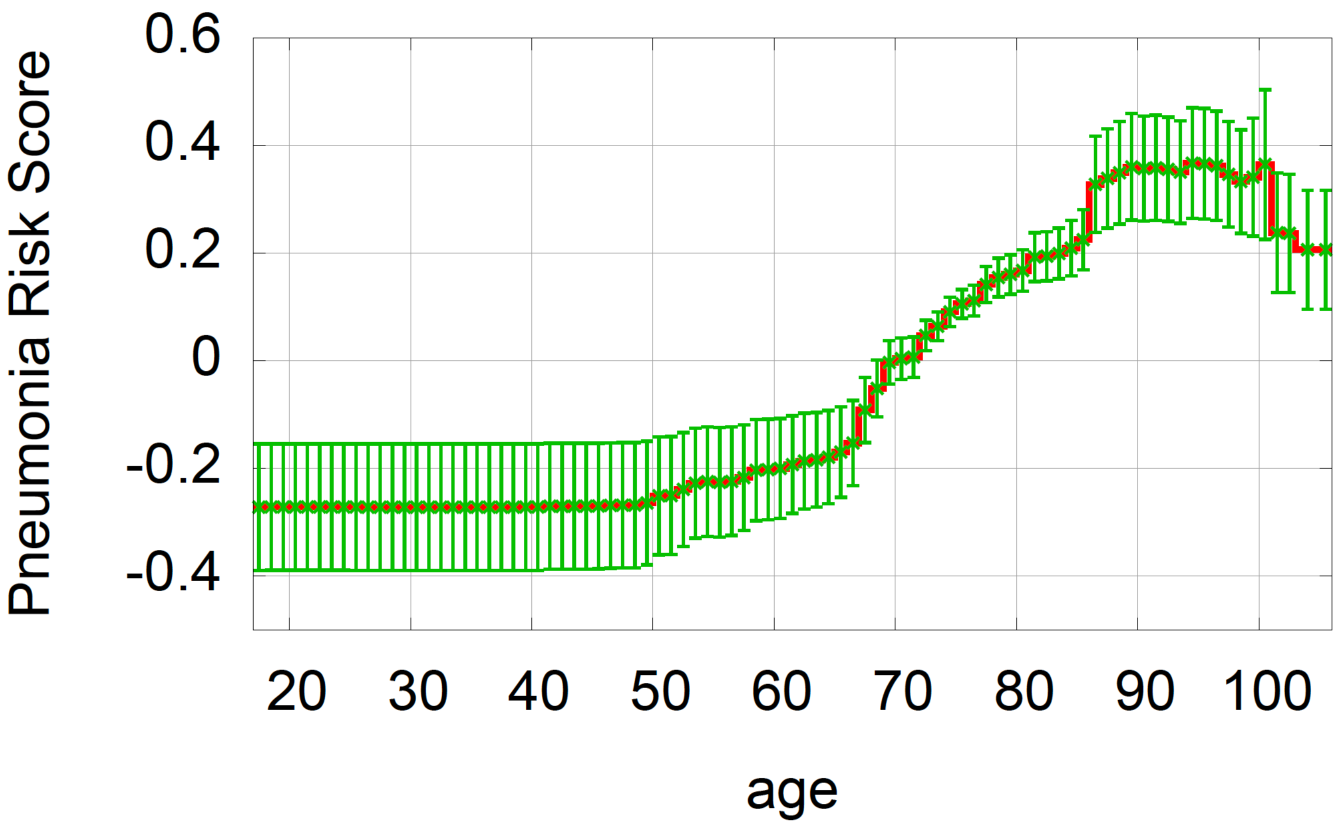




Additive Model

$$\hat{y}(x) \propto \sum_{j=1}^{p} f_j(x_j)$$





Caruana et al. KDD 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Accurate Intelligible Models with Pairwise Interactions
(Lou, Caruana, et al. 2013)

GA^{2}M – Generalized Additive Models plus Interactions

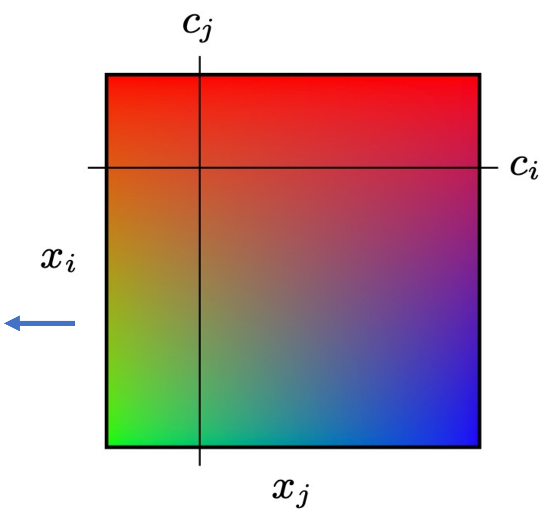
$$g(\hat{y}(x)) = \sum_{j=1}^{p} f_j(x_j) + \sum_{k \neq j} f_{kj}(x_k, x_j)$$

Algorithm for fitting GA^{2}M

Fit an additive model first (without any interactions)

Until convergence

- Add interaction term (chosen to minimize the residual)
- Refit the model $\hat{y}$ with the new interaction term.



GA^{2}M are expressive, not sparse.

The FastSparse Algorithm

Liu et al, AISTATS, 2022

Sparse Logistic Regression

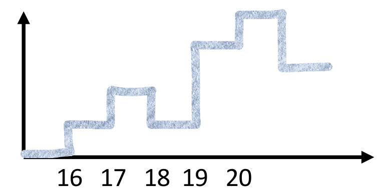
$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) + \lambda_0 \|\boldsymbol{w}\|_0$$

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



- coordinate descent (often setting coeffs to 0)

Sparse Exponential Loss Classification

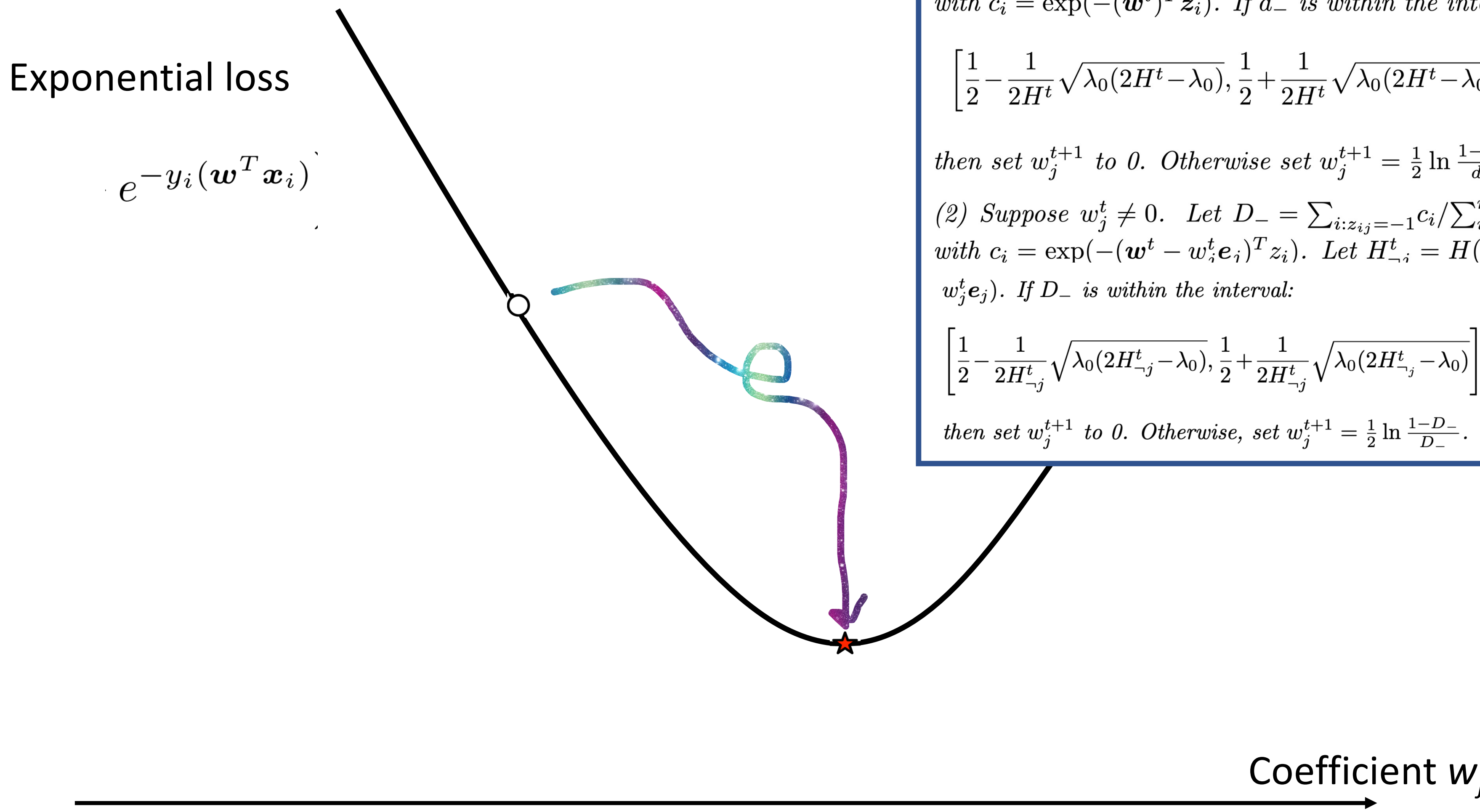
$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) + \lambda_0 \|\boldsymbol{w}\|_0$$

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

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

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

- coordinate descent (often setting coeffs to 0)



(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

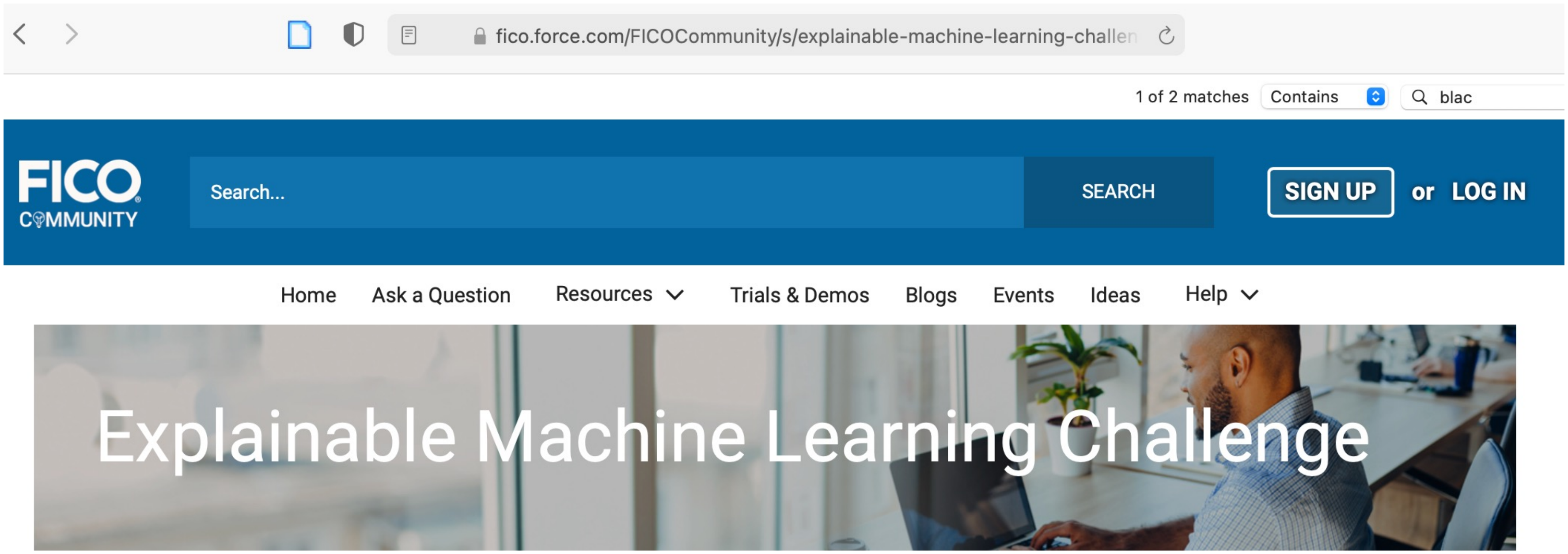
$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) + \lambda_0 \|\boldsymbol{w}\|_0$$

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

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

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

- coordinate descent (often setting coeffs to 0)

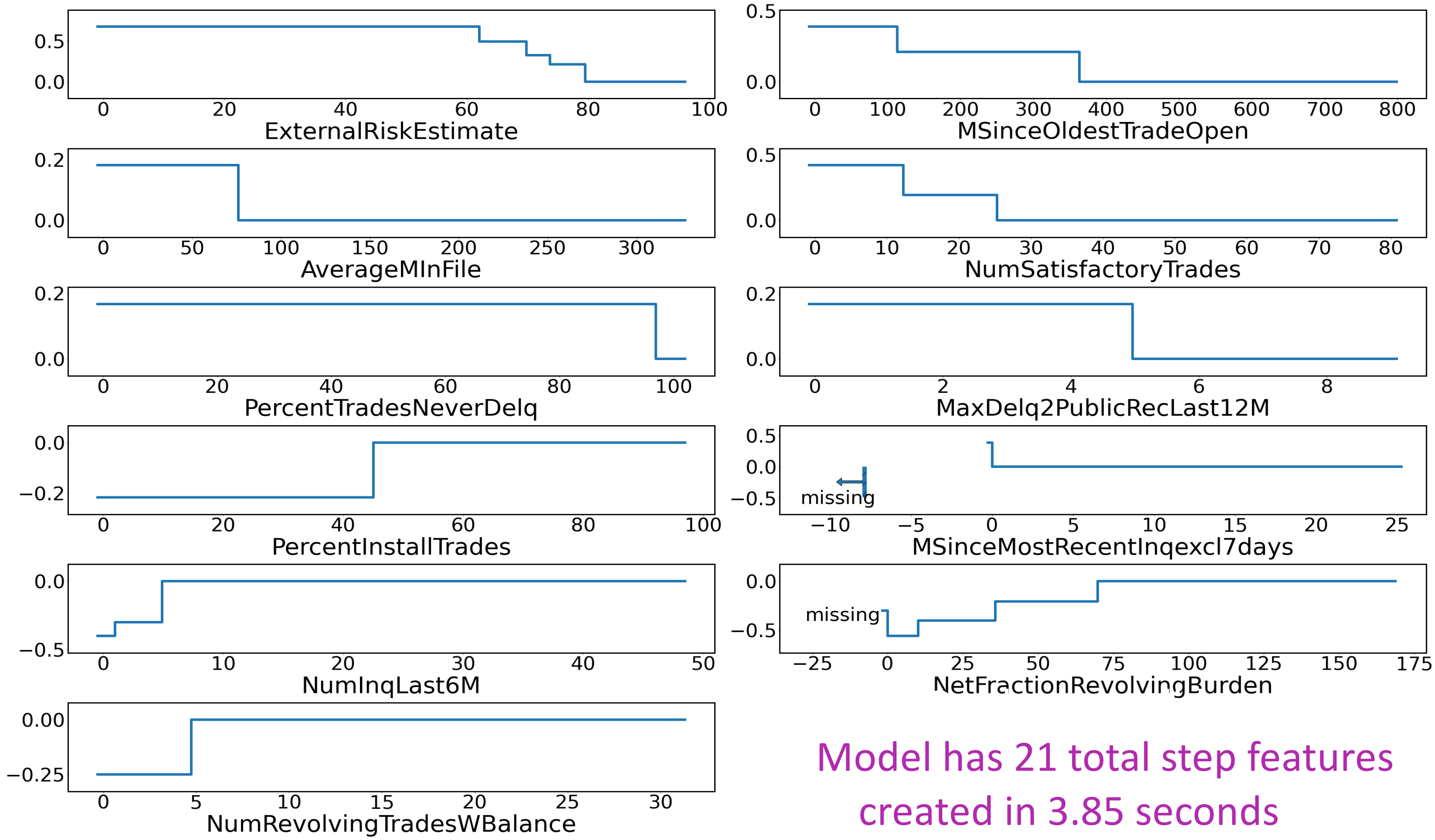


Home Equity Line of Credit (HELOC) Dataset

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.

This dataset → 1917 binary features

Generalized Additive Model on the FICO Dataset



Model has 21 total step features created in 3.85 seconds

Challenges that were tackled recently

If a GAM shows counterintuitive relationships between features and outcomes, can we use this to troubleshoot?

Chen et al., Missing Values and Imputation in Healthcare Data: Can Interpretable Machine Learning Help? CHIL, 2023

User interaction with GAMs (more later).

Still a Challenge: user-specified shape functions

Note that GA2M is available in the interpML package, and FastSparse is also public.

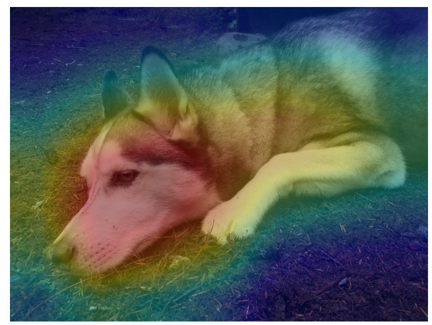
Stop here for ≤2 questions and a quick break

5 principles,
Grand challenges:
- 7 DR
- 1 Logical models
- 3 GAMs

Warning (before I move into neural networks)

- One does not need neural networks for tabular data.
  - There are lots of papers on neural networks for tabular data
- The meaning of interpretability needs to be defined for non-tabular data.
- The neural networks people use the words "interpretable" and "explainable" interchangeably.
  - There are a lot of papers and websites claiming "interpretability" when they are explaining neural networks.
  - Even papers called "Interpretable CNNs" are not necessarily interpretable.
- In my view, "saliency" is not sufficient for interpretability.

”Explaining” deep NN’s with saliency maps doesn’t work

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps		 “Explanation”	

Do you trust the network now?

Lots of work in radiology on attention maps now…

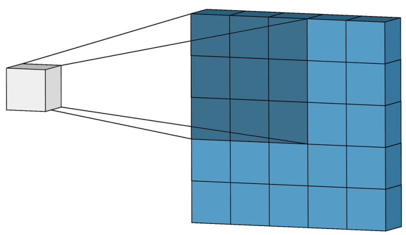
Switch speakers

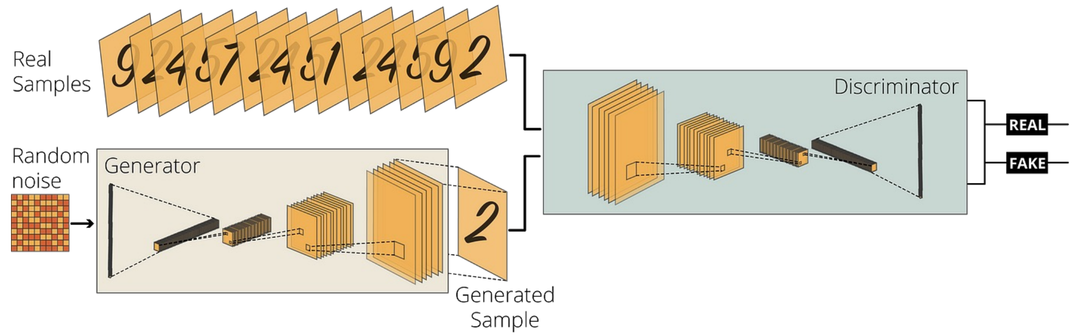
5 principles,

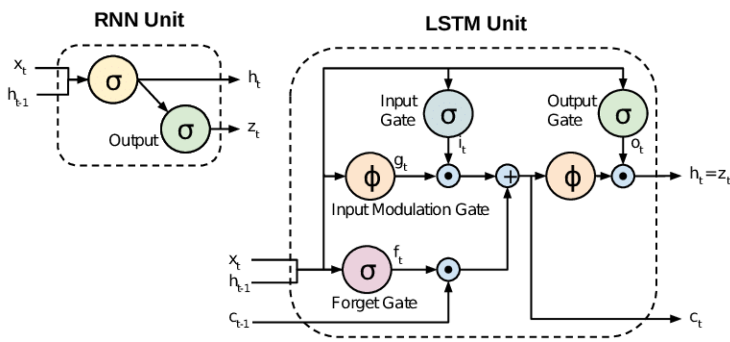
Grand challenges:

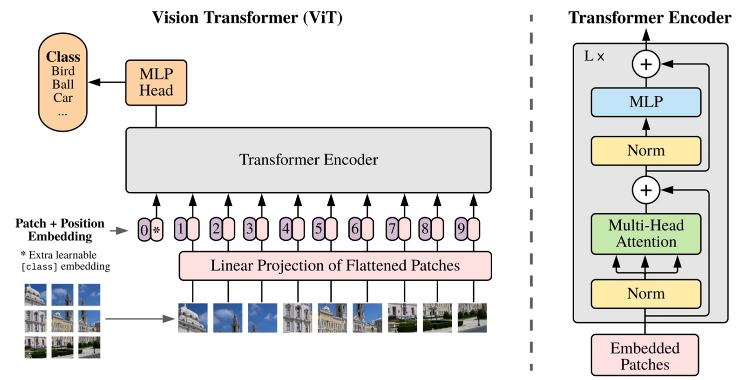
- 7 DR
- 1 Logical models
- 3 GAMs
- 11 Generalizable NN methods
- 4 Case-based reasoning

Provide methods that generalize to new architectures









Open questions

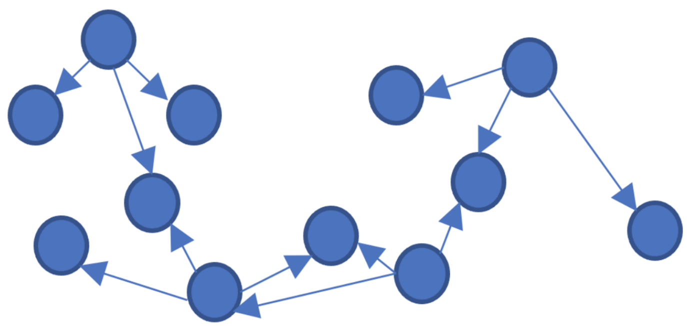
11.1 Can we produce interpretable methods that can apply to both old and new architectures?

11.2 Can we produce explanations that are useful in multiple domains?

Case-based reasoning

Provide explanations that generalize across domains

Case-based reasoning

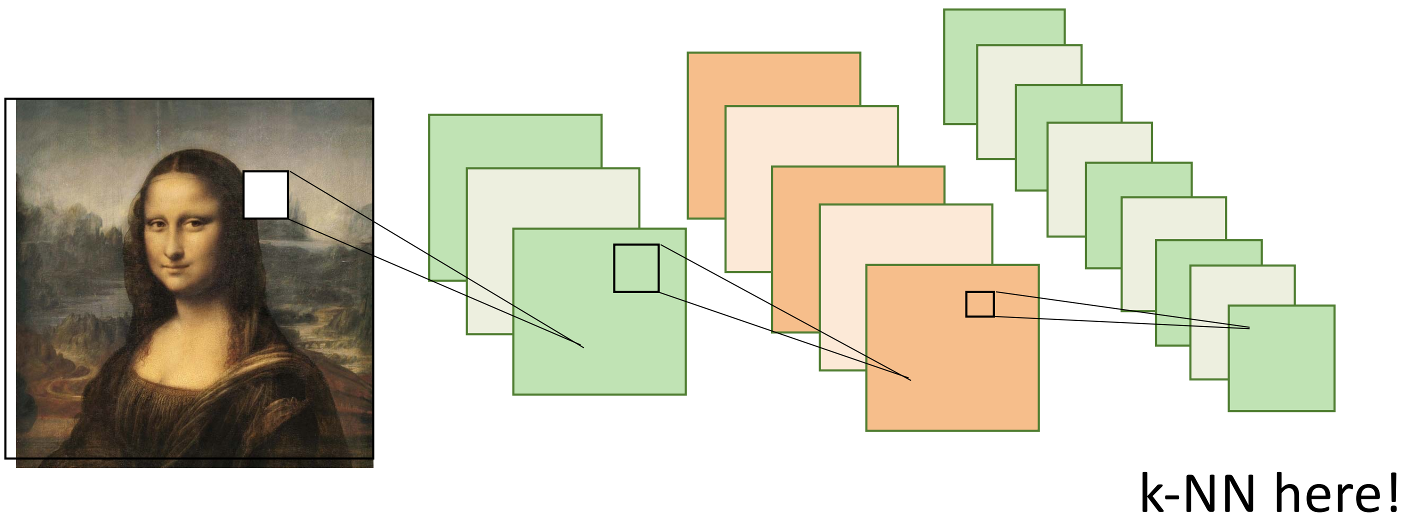


Nearest neighbour

- can be used for any data type
- extremely powerful, can even be used for images
- dates to the beginning of AI, K-nearest neighbors

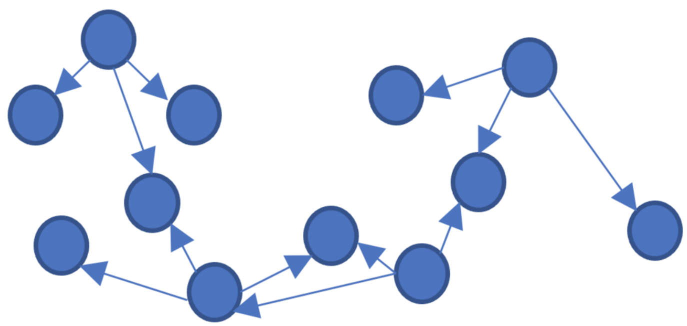
- Case-based reasoning is a paradigm that involves solving a new problem using known solutions to similar past problems (Aamodt and Plaza, 1994)

- k-nearest neighbors (kNN) (Fix and Hodges, 1951; Cover and Hart, 1967). No training required
  - Weinberger and Saul (2009) adaptive k-NN
  - Salakhudinov & Hinton (2017) Deep k-NN
  - Papernot & McDaniel (2018) Deep k-NN where neighbors from every layer in the network are used.
  - Card et al. (2019) Deep weighted averaging classifier – classification based on latent space distances



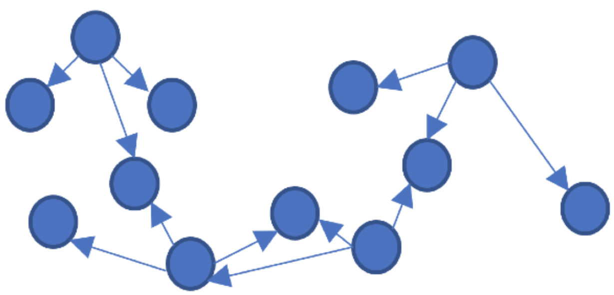
Introducing prototype models

Case-based reasoning

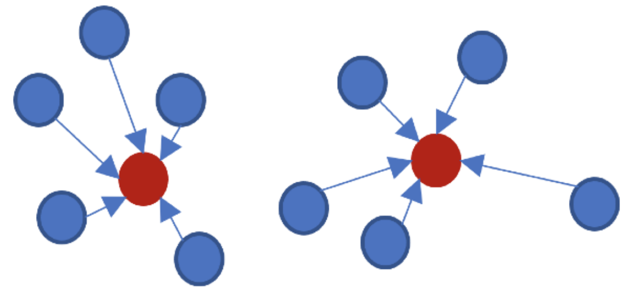


Nearest neighbour

n comparisons



Nearest neighbors



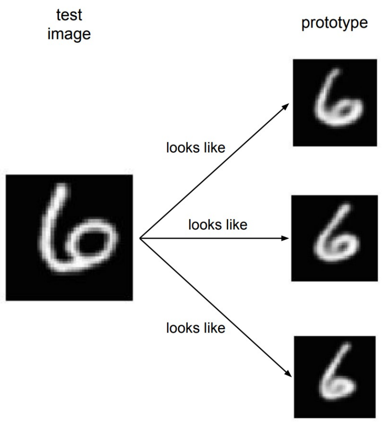
Prototype models

- Computationally expensive
- Can show a bad neighbour (misclassified, not representative)

- Need to learn the prototypes as well as distance metric
- Global interpretability
- Prototype editing

Case granularity

ProtoPNet (Chen et al. 2019)



Whole image

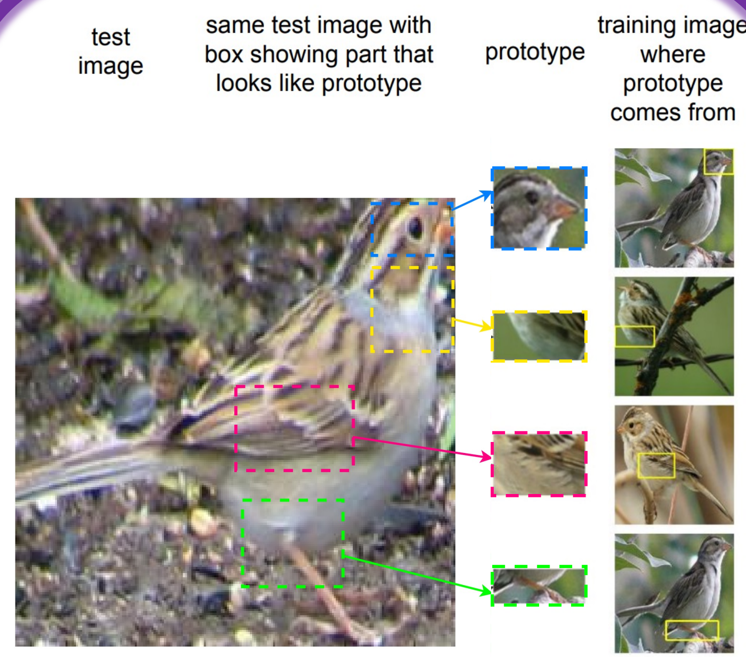
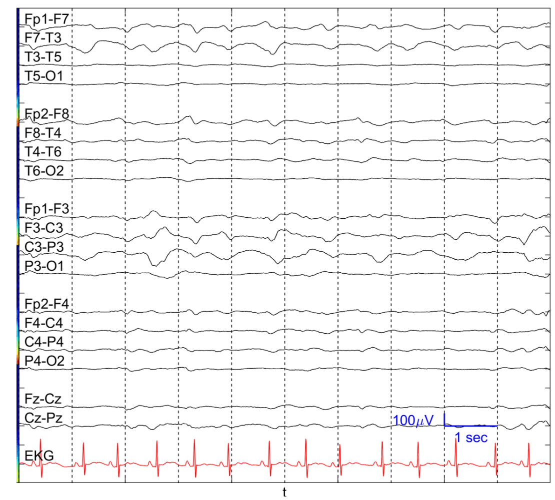


Image part



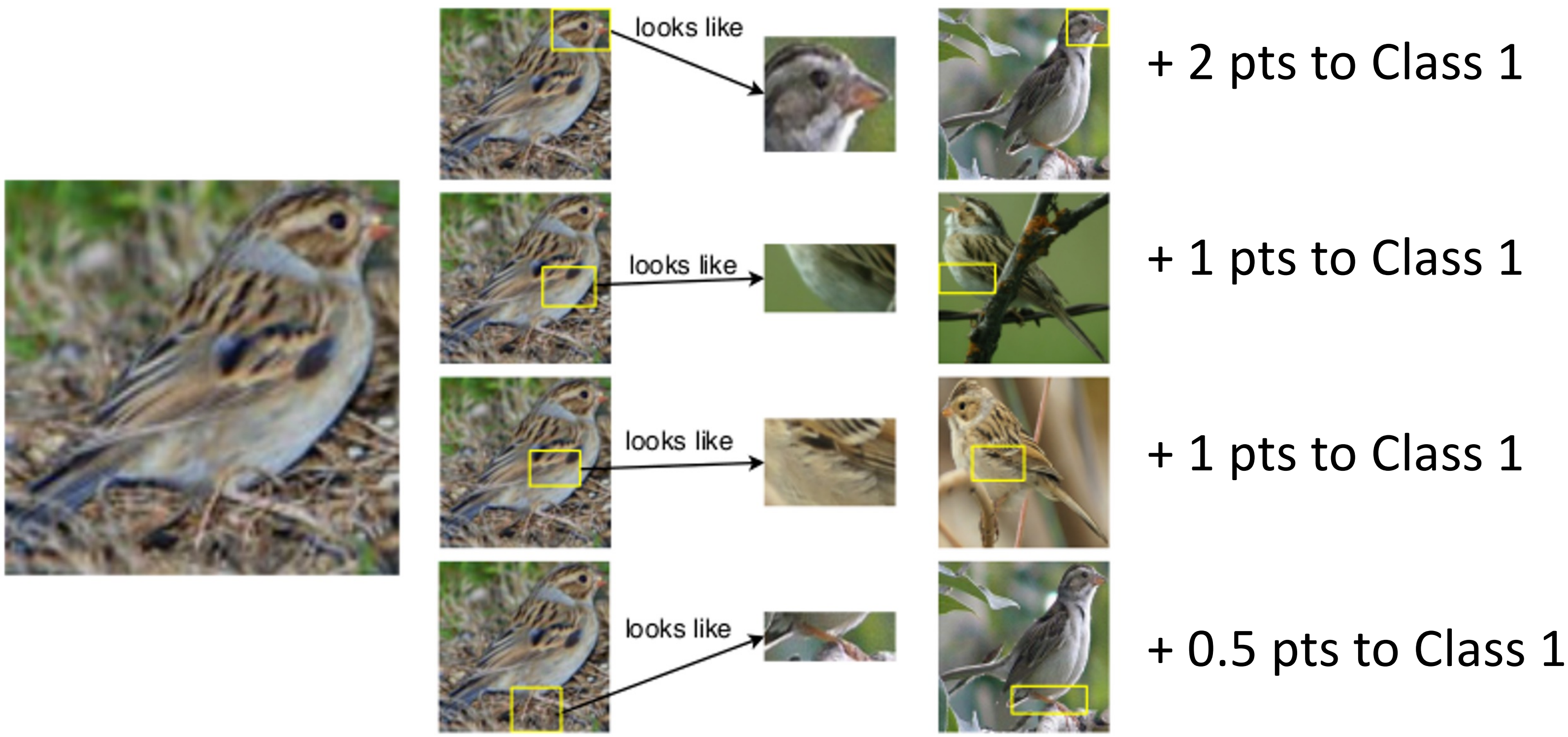
Time segment

How would you describe why this bird is a clay-colored sparrow?

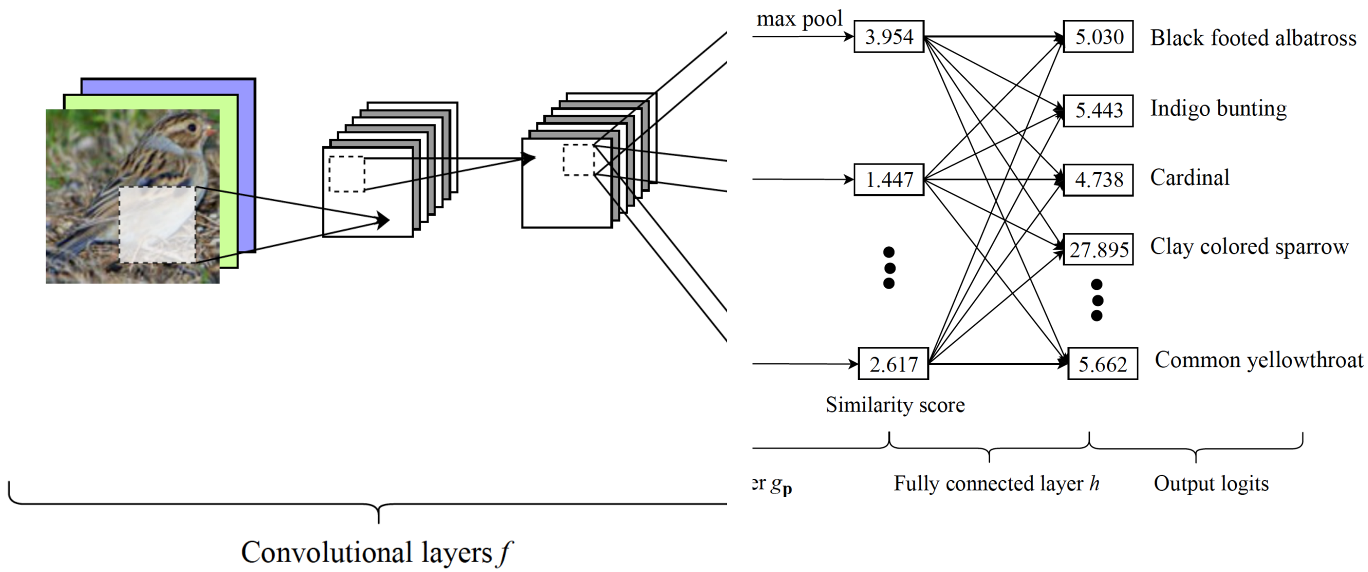
Compare parts of the bird to typical parts from the class

+ 2 pts to Class 1

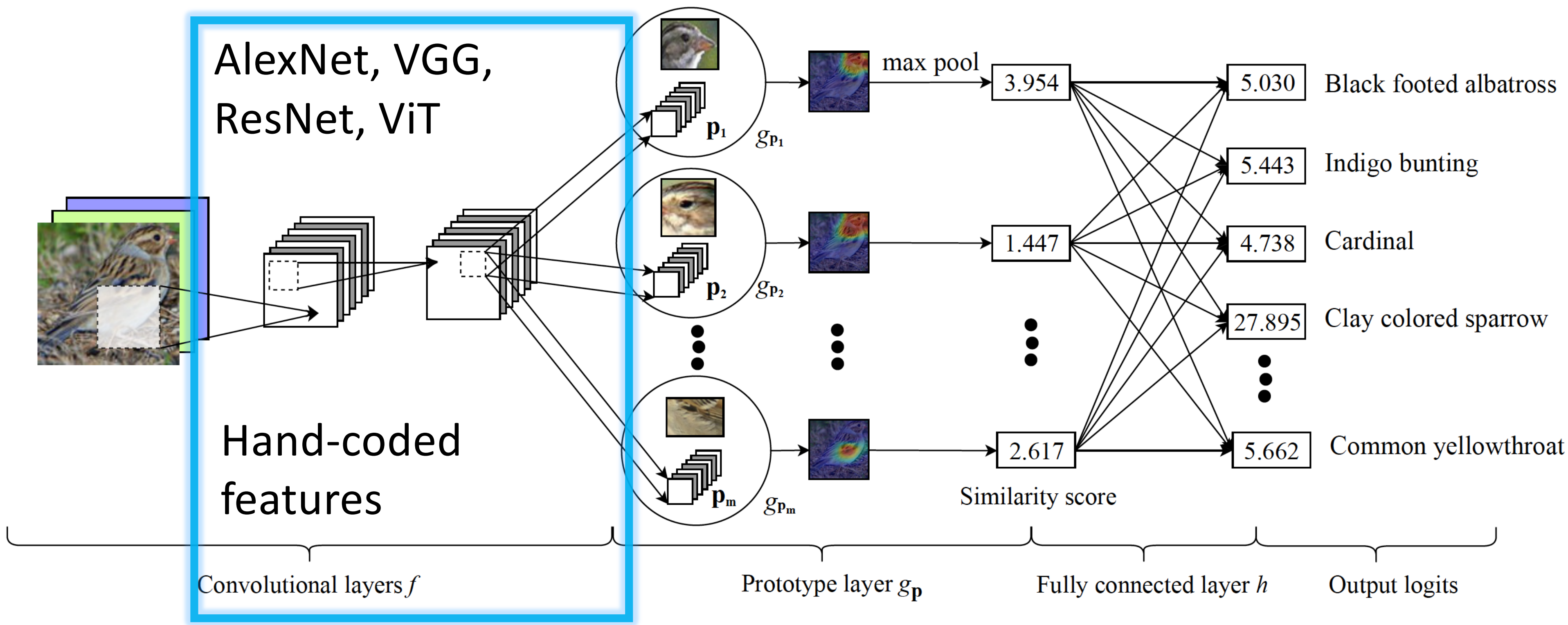
Add evidence from many parts to make a prediction



Take any “standard” black box CNN…



And transform it to be interpretable



Semi-ProtoPNet Deep Neural Network for the Classification of Defective Power Grid Distribution Structures

by Stefano Frizzo Stefenon [1,2,*], Gurmail Singh [3], Kin-Choong Yow [3] and Alessandro Cimatti [1]

1 Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy

2 Department of Mathematics, Informatics and Physical Sciences, University of Udine, Vi
33100 Udine, Italy

3 Faculty of Engineering and Applied Science, University of Regina, Wascana Parkway 3
0A2, Canada

XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations

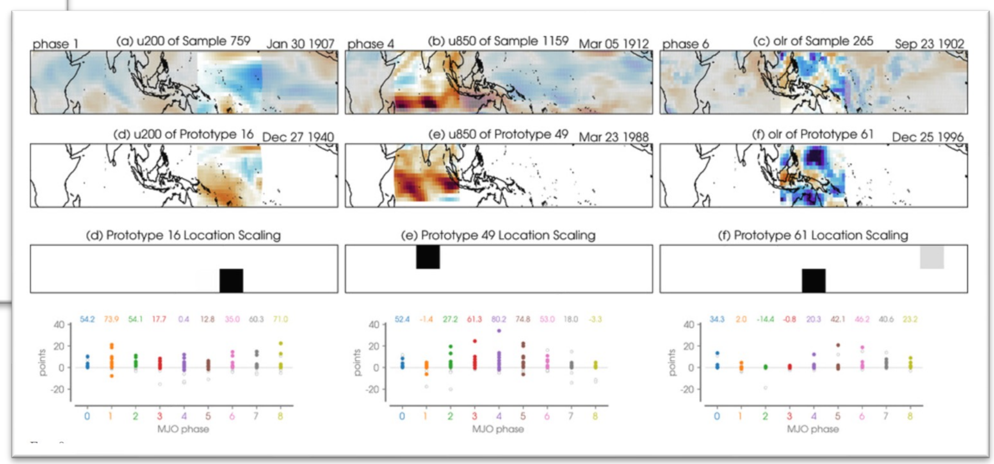
Eunji Kim, Siwon Kim, Minji Seo, Sungroh Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15719-15728

Reevaluating the Safety Impact of Inherent Interpretability on Deep Neural Networks for Pedestrian Detection

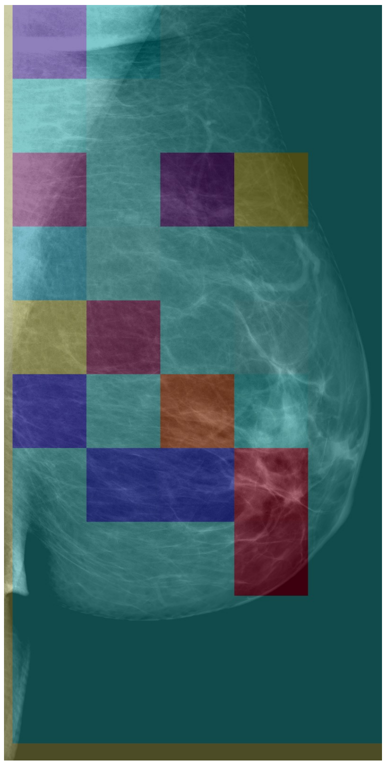
Patrick Feifel[1,2] Frank Bonarens[1] Frank Köster[2,3]

frank.bonarens@stellantis.com

[1] Stellantis, Opel Automobile GmbH
[2] Carl von Ossietzky Universität Oldenburg
[3] Deutsches Zentrum für Luft- und Raumfahrt



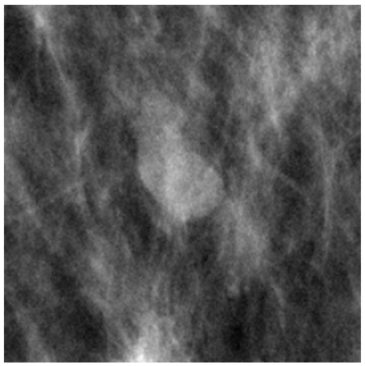
An Example Application of ProtoPNet: Computer-Aided Mammography



- For domains where deep learning dominates
- High-stakes decisions

- Constrain logic
  - Model decision is based on similarity to “prototypical” cases
  - Prototypes relate to known medical feature

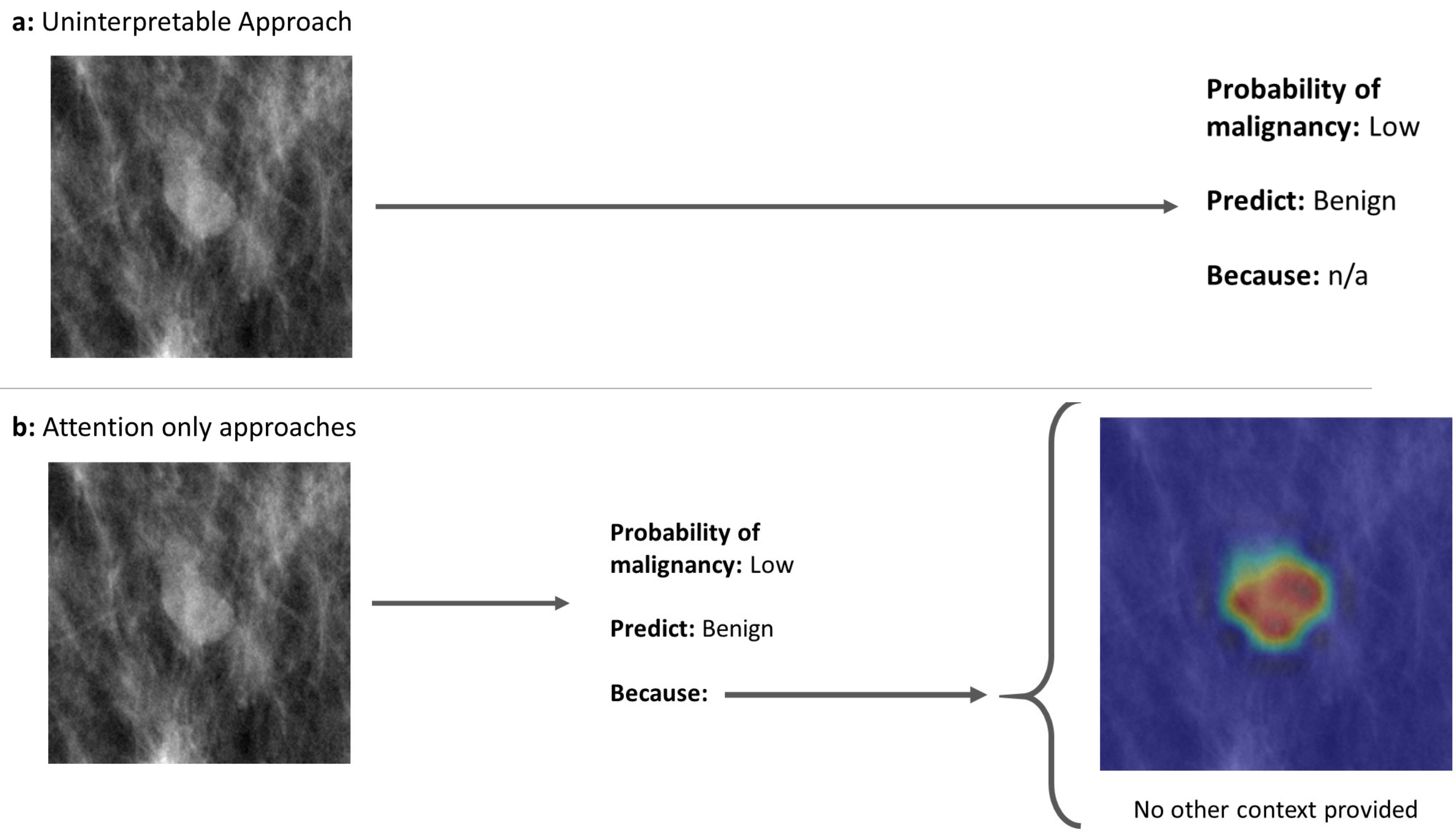
a: Uninterpretable Approach



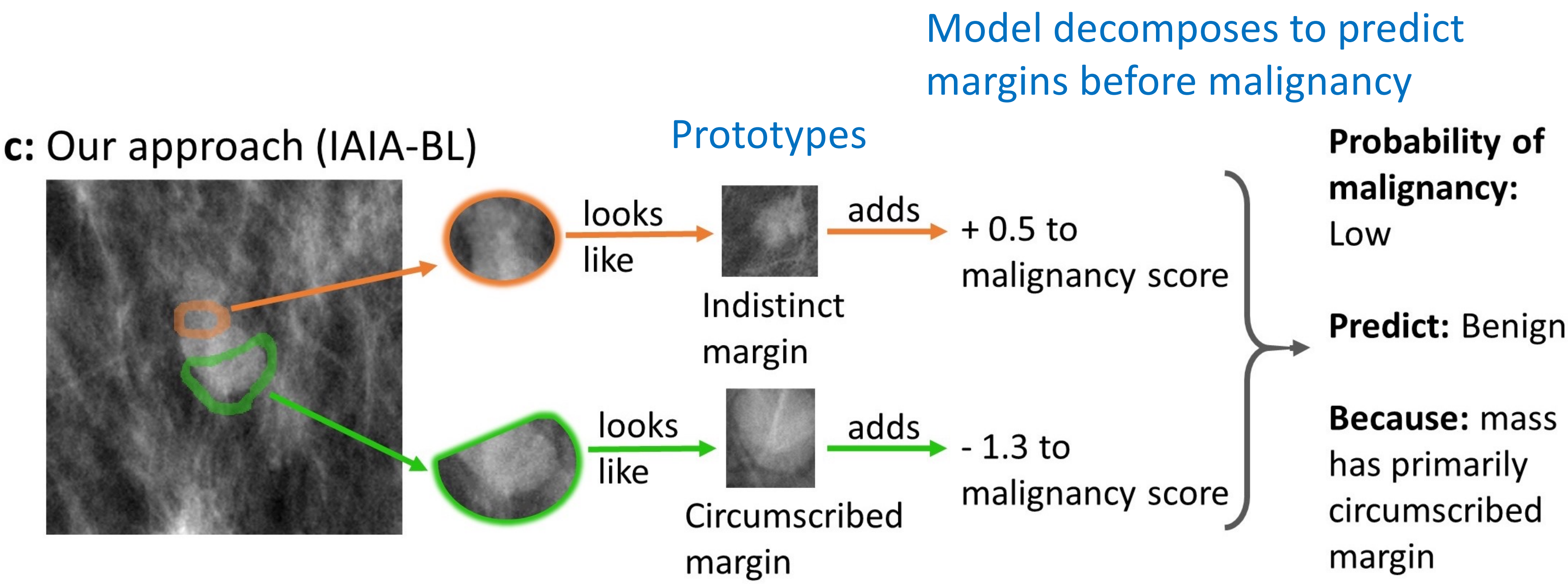
Probability of malignancy: Low

Predict: Benign

Because: n/a


a: Uninterpretable Approach
Probability of malignancy: Low
Predict: Benign
Because: n/a
b: Attention only approaches
Probability of malignancy: Low
Predict: Benign
Because:
No other context provided

Interpretable AI algorithm for Breast Lesions (IAIA-BL)



Open questions

- 4.1: How to integrate prior knowledge or human supervision into prototype learning?
  - humans may want to prune prototypes, design them, or specify a region/feature of interest where the prototypes should focus.
  - How to make this generalizable to many domains? (Challenge 11.2)

- 4.2: How to troubleshoot a trained prototype-based model to improve the quality of prototypes?
  - How can we replace a “bad” prototype?
  - Posthoc pooling of prototypes (Rymarczyk et al., 2022) (Rymarczyk et al., 2021)

Open questions

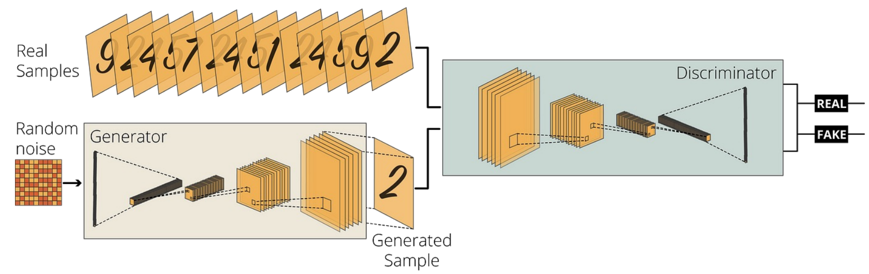
- 4.3: Representations are linked to context, how to represent that context?
  - Prototype shows a part of an image, but the area around that affect the representation as well
  - Donnelly et al., 2022

Stop here for ≤2 questions

- 5 principles,
- Grand challenges:
  - 7 DR
  - 1 Logical models
  - 3 GAMs
  - 11 Generalizable NN methods
  - 4 Case-based reasoning
  - 12 Generative models
  - 13 NLP

State of generative models

- Generative models
  - GANs (Goodfellow et al., 2014)
    - Generative Adversarial Nets
  - Stable diffusion
    - DALL-E 2
    - Midjourney
- Interpretability in this space
  - Ross et al., 2021 Interactive Reconstruction
  - Sahiner et al., 2021 Replace the deep NNs

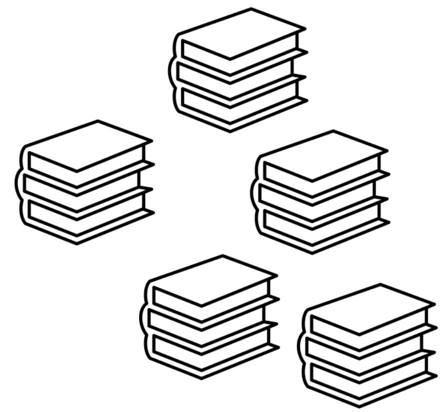


Open problems

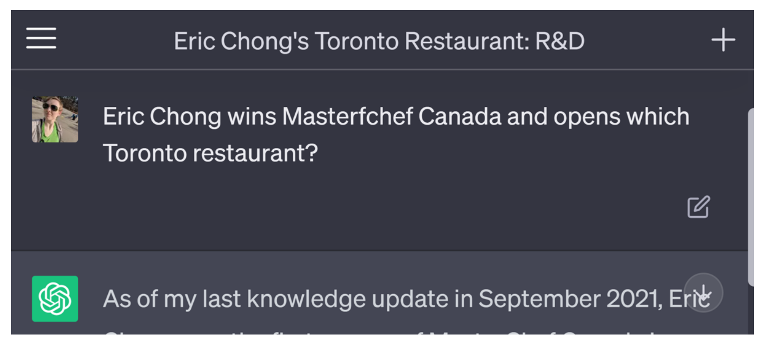
- 12.1 What constitutes an explanation for this type of task?
  - An exhaustive list of source material is impractical
  - Unified rubric for explanations of this task
- 12.2 Can you have one true explanation when you have multiple outputs?
  - How can you quantify variability in the output?

Interpretability in natural language processing (NLP)

- Classification of text segments
- Extraction of key information / summarization
- Information retrieval



- Generating text from prompts



Is self-explanation interpretable enough?

User Can you explain this meme?

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.

GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.

The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world.

The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

Some existing approaches for interpretability in NLP

- Keyphrase extraction / keyword extraction
  - Hasan et al., 2014 survey paper
- Rationale extraction
  - EMNLP (Lei, Barzilay, Jaakkola, 2016)
  - SPECTRA (Guerreiro and Martins, 2021)
- Prototype-based methods for classification
  - ProSeNet (Ming et al., 2019)
- Ask the AI to generate its own explanation of itself

- Pre-prompt the LLM with correct/verified information
  - Retrieval-Augmented Generation (Lewis et al., 2020) LLaVA-Med (Li et al., 2023)
  - Esteva et al., 2021 retrieves specific paragraphs of source database
  - Qiao et al., 2023 have a survey paper of 100s of these

Open Questions

- 13.1 What quality of explanation is good enough?
  - I would argue that the generation of plausible looking explanations with no guarantee of their truthfulness will be insufficient.
  - Is looking at most recent sources and prompts enough?
  - Concerns about explanation faithfulness (Lyu et al., 2023)

- 13.2 Generalizing methods that worked on RNN+LSTM architecture to a transformer-based model
  - Some flavour of 11.1, the development of methods that generalize to new architectures

Stop here for ≤2 questions

5 principles,

Grand challenges:

- 7 DR
- 1 Logical models
- 3 GAMs
- 11 Generalizable NN methods
- 4 Case-based reasoning
- 12 Generative models
- 13 NLP
- 9 Rashomon sets

9. Characterization of the “Rashomon” set of good models

Many realities, no one truth.

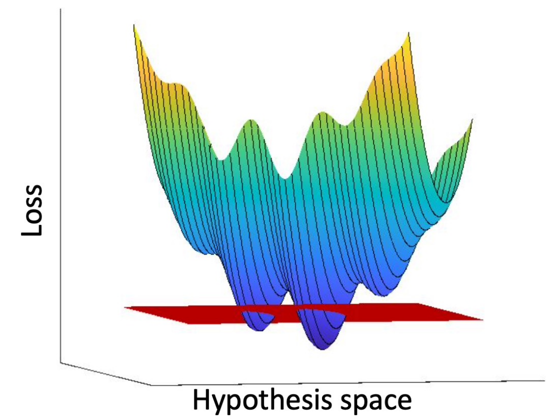
9. Characterization of the "Rashomon" set of good models

- The Rashomon set has models with low loss:

$$R(\mathcal{F}, f^*, \epsilon) = \{f \in \mathcal{F} \text{ such that } Loss(f) \leq Loss(f^*) + \epsilon\}$$

Could the Rashomon set be the key to everything?

- Can the Rashomon set explain why simple-yet-accurate models exist for tabular data?
- Can the Rashomon set help us with a key challenge in ML, namely interacting with users?
- Can the Rashomon set help us understand variable importance?



(a)

9. Characterization of the “Rashomon” set of good models

- The Rashomon set has models with low loss:

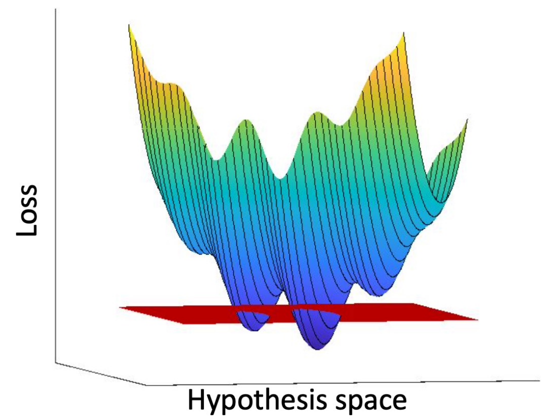
$$R(\mathcal{F}, f^*, \epsilon) = \{f \in \mathcal{F} \text{ such that } Loss(f) \leq Loss(f^*) + \epsilon\}$$

“Rashomon Set” Theory

On the Existence of Simpler Machine Learning Models

Lesia Semenova, Cynthia Rudin, and Ronald Parr

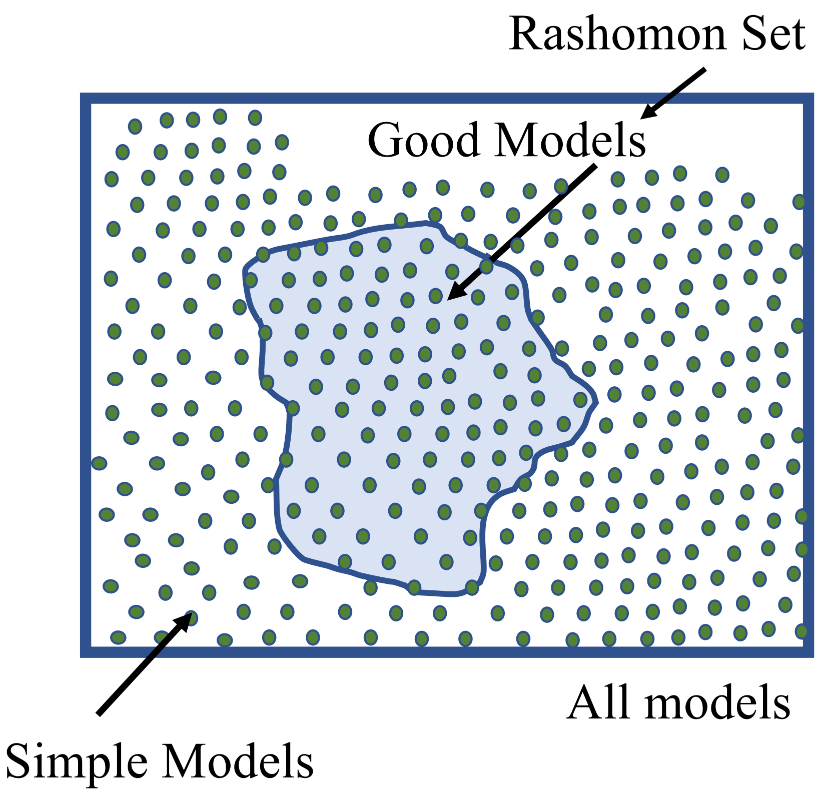
ACM Conference on Fairness, Accountability, and Transparency, 2022



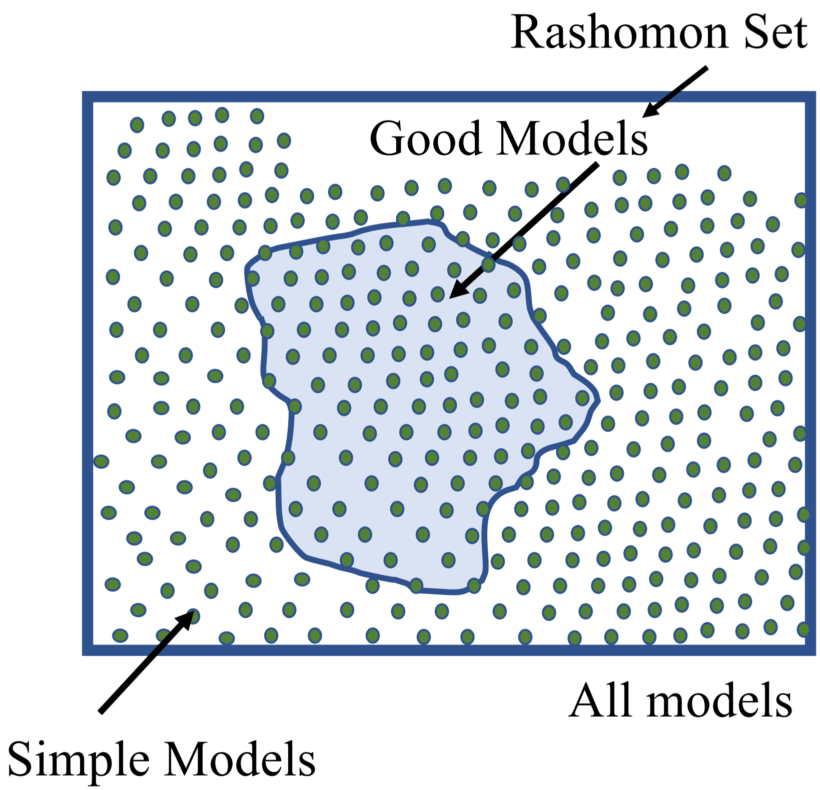
(a)

- Many datasets have large Rashomon sets
- If the Rashomon set is large, it is likely to contain interpretable yet accurate models.
- Thus, many datasets yield interpretable models.

The “Rashomon Set” Theory



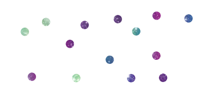
The “Rashomon Set” Theory



Large Rashomon sets are correlated with:

The existence of simpler models.

More label/feature “noise”.



Implication:

Optimizing for simplicity
won’t sacrifice accuracy.

For details, see “On the existence of Simpler Machine Learning Models,” Semenova, Rudin & Parr, ACM FAccT, 2022

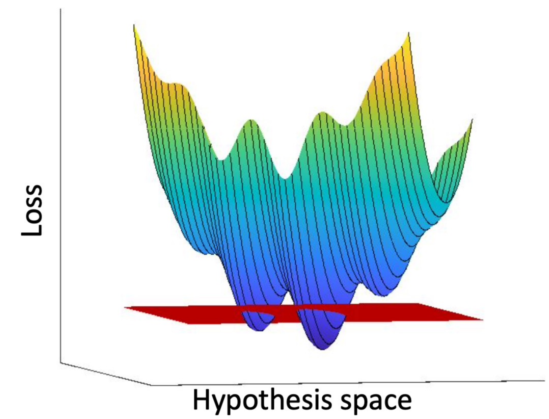
9. Characterization of the "Rashomon" set of good models

- The Rashomon set has models with low loss:

$$R(\mathcal{F}, f^*, \epsilon) = \{f \in \mathcal{F} \text{ such that } Loss(f) \leq Loss(f^*) + \epsilon\}$$

Could the Rashomon set be the key to everything?

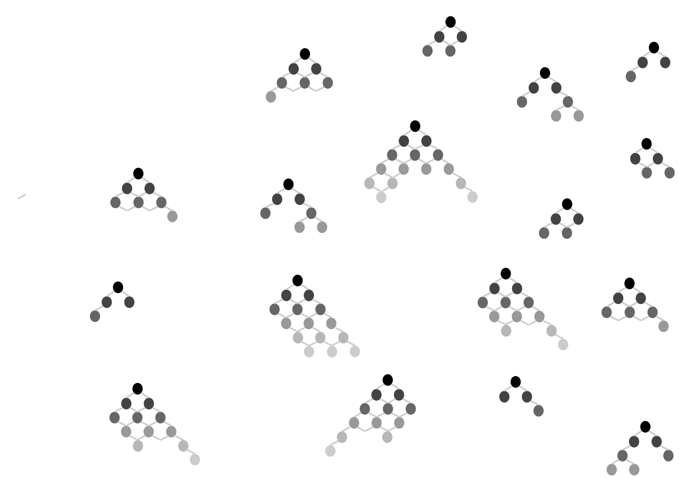
- Can the Rashomon set explain why simple-yet-accurate models exist for tabular data?
- Can the Rashomon set help us with a key challenge in ML, namely interacting with users?
- Can the Rashomon set help us understand variable importance?



(a)

A New Paradigm of Machine Learning

Training Set ⟶ Algorithm ⟶ Many Predictive Models



Exploring the Whole Rashomon Set of Sparse Decision Trees

Rui Xin*
Duke University
rui.xin926@duke.edu

Chudi Zhong*
Duke University
chudi.zhong@duke.edu

Zhi Chen*
Duke University
zhi.chen1@duke.edu

Takuya Takagi
Fujitsu Laboratories Ltd.
takagi.takuya@fujitsu.com

Margo Seltzer
The University of British Columbia
mseltzer@cs.ubc.ca

Cynthia Rudin
Duke University
cynthia@cs.duke.edu

NeurIPS 2022

TreeFARMS = Trees FAst RashoMon Sets

TreeFARMS returns *all* almost-optimal trees

TIMBERTREK: Exploring and Curating Sparse Decision Trees with Interactive Visualization

Zijie J. Wang[1] Chudi Zhong[2] Rui Xin[2] Takuya Takagi[3] Zhi Chen[2]
Duen Horng Chau[1] Cynthia Rudin[2] Margo Seltzer[4]

bit.ly/timbertrek

IEEE Vis 2022

TreeFARMS returns *all* almost-optimal trees

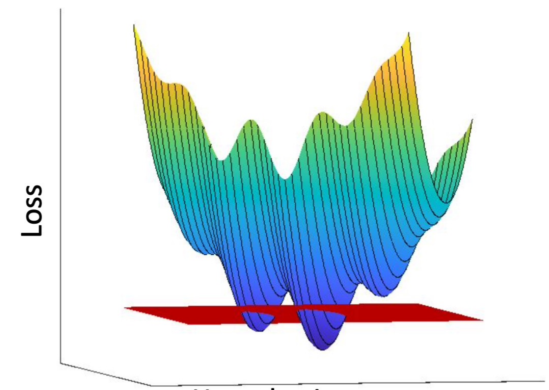
9. Characterization of the "Rashomon" set of good models

- The Rashomon set has models with low loss:

$$R(\mathcal{F}, f^*, \epsilon) = \{f \in \mathcal{F} \text{ such that } Loss(f) \leq Loss(f^*) + \epsilon\}$$

Could the Rashomon set be the key to everything?

- Can the Rashomon set explain why simple-yet-accurate models exist for tabular data?
- Can the Rashomon set help us with a key challenge in ML, namely interacting with users?
- Can the Rashomon set help us understand variable importance?



(a)

Challenges that were tackled recently

Can we handle constraints on models?

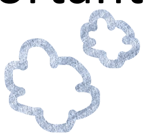
Just filter the Rashomon set!

- fairness
- monotonicity
- multiple performance objectives

Challenges that were tackled recently

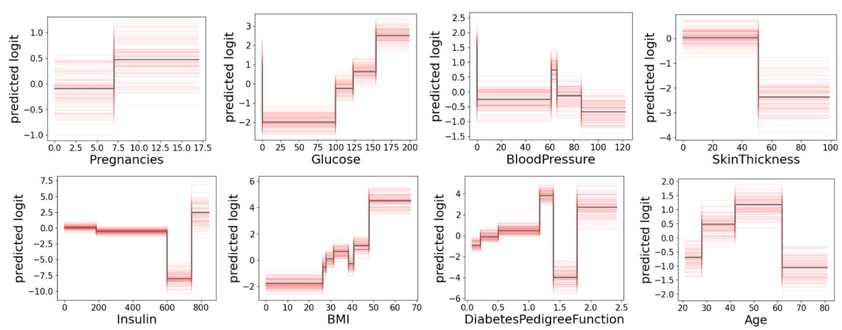
Can we project the Rashomon set onto variable importance axes to see how often variables are important within the Rashomon set?

“Variable Importance Clouds”



Dong and Rudin. Exploring the Cloud of Variable Importance for the Set of All Good Models, Nature Machine Intelligence, 2020.

Can we get Rashomon sets for other model classes?



Chen et al., 2023. Understanding and Exploring the Whole Set of Good Sparse Generalized Additive Models.

5 principles,

Grand challenges:

- 7 DR
- 1 Logical models
- 3 GAMs
- 11 Generalizable NN methods
- 4 Case-based reasoning
- 12 Generative models
- 13 NLP
- 9 Rashomon sets

Done!

arXiv.org > cs > arXiv:2103.11251

Search...

Help | Advanced S

Computer Science > Machine Learning

[Submitted on 20 Mar 2021]

Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, Chudi Zhong

Interpretability in machine learning (ML) is crucial for high stakes decisions and troubleshooting. In this work, we provide fundamental principles for interpretable ML, and dispel common misunderstandings that dilute the importance of this crucial topic. We also identify 10 technical challenge areas in interpretable machine learning and provide history and background on each problem. Some of these problems are classically important, and some are recent problems that have arisen in the last few years. These problems are: (1) Optimizing sparse logical models such as decision trees; (2) Optimization of scoring systems; (3) Placing constraints into generalized additive models to encourage sparsity and better interpretability; (4) Modern case-based reasoning, including neural networks and matching for causal inference; (5) Complete supervised disentanglement of neural networks; (6) Complete or even partial unsupervised disentanglement of neural networks; (7) Dimensionality reduction for data visualization; (8) Machine learning models that can incorporate physics and other generative or causal constraints; (9) Characterization of the "Rashomon set" of good models; and (10) Interpretable reinforcement learning. This survey is suitable as a starting point for statisticians and computer scientists interested in working in interpretable machine learning.

Thanks