A Data Science Playbook for xAI - Navigating Interpretable and Predictive Models

Josh Poduska, Chief Data Scientist, Domino Data Lab

Why Do We Care About Model Interpretability?

Model ethics, bias, and misuse

Regulatory requirements

Trust and understanding


I think you should be more explicit here in step two
1. Data collection
& pre-processing
z = (x - u) / e
2. Model training
THEN A MIRACLE OCCURS ...
3. Evaluation
65
S. Harris


GLOBAL
LOCAL

If you had to chose only one of the following models for predicting the purchase of multi-game tickets next year, which would you choose?

1. A model that was able to find almost all the people that will buy multi-game tickets next year
2. A model that told you what combination of characteristics in a buyer are strongly associated with multi-game purchases
3. A model that told you exactly why it thinks a given person will or will not make a multi-game purchase

DOMINO

Model #1 – High Accuracy

DOMINO

Model #2 – Interpretable with Global Inference

- In general, people in cheap seats who purchased far in advance of the game do not make multi-game purchases

- In general, men who buy last-minute will make a multi-game purchase

DOMINO

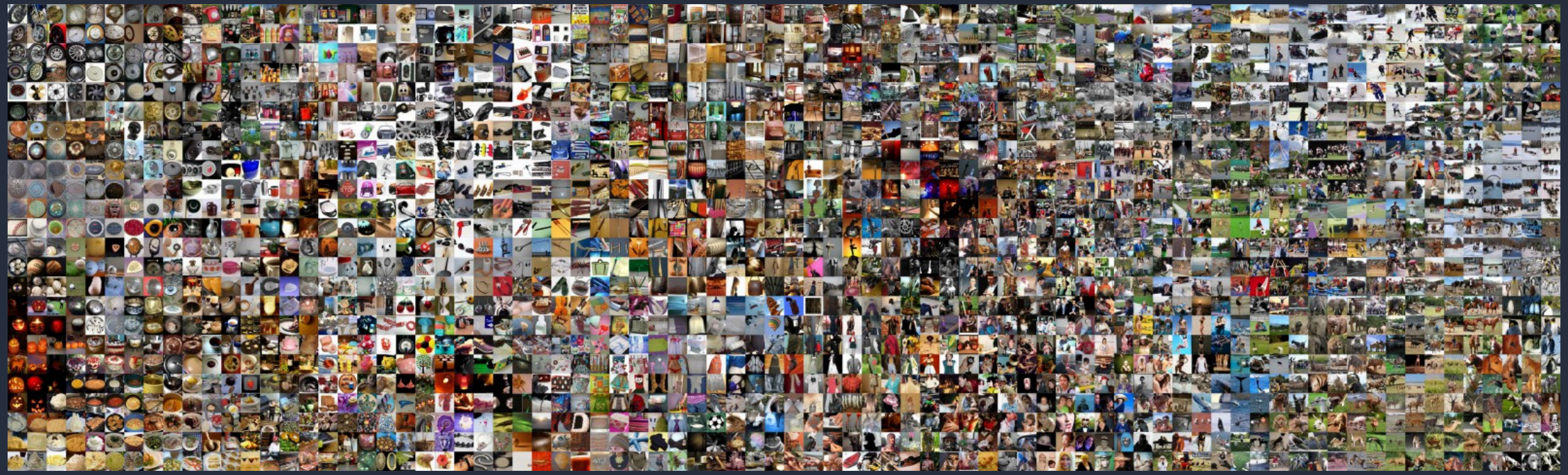
Model #3 – Interpretable with Local Inference

Yes

This person has a 0.82 probability of making a multi-game purchase this season.

0.38 due to seat location
+
0.22 due to purchased 3 days prior to game
+
0.17 due to age=45
+
0.05 due to spending more than $50 on concessions
=
0.82

DOMINO

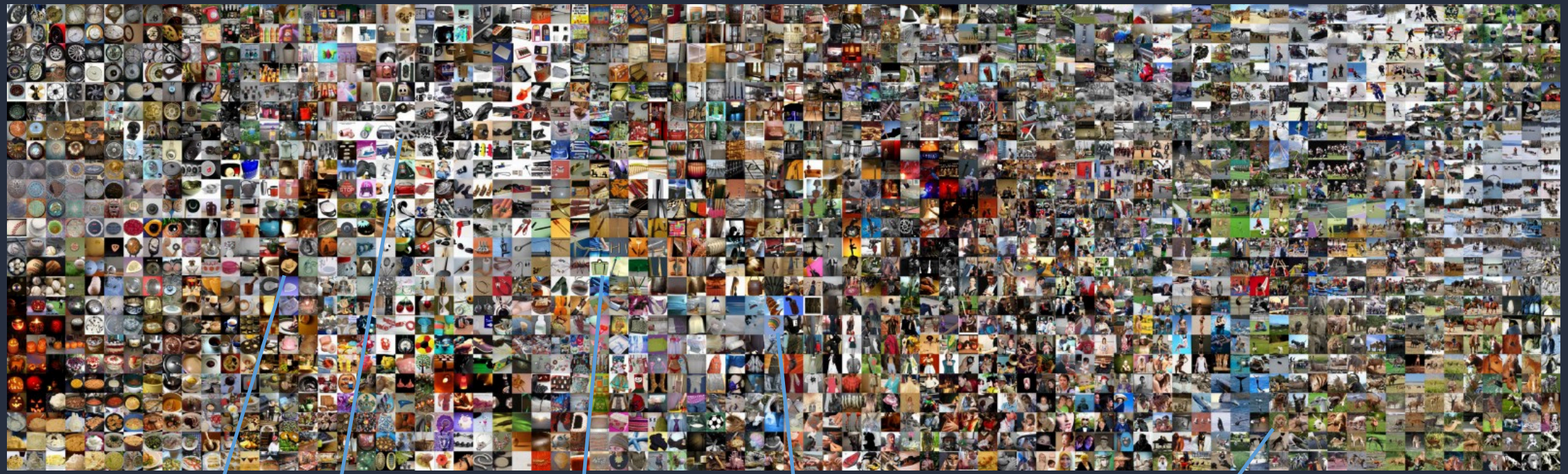


If you had to choose only one of the following three models to classify these images, which would you choose?

1. A model that was able to classify almost all these images correctly
2. A model that told you what features of images, in general, are important for classification
3. A model that told you exactly why a certain image got its classification

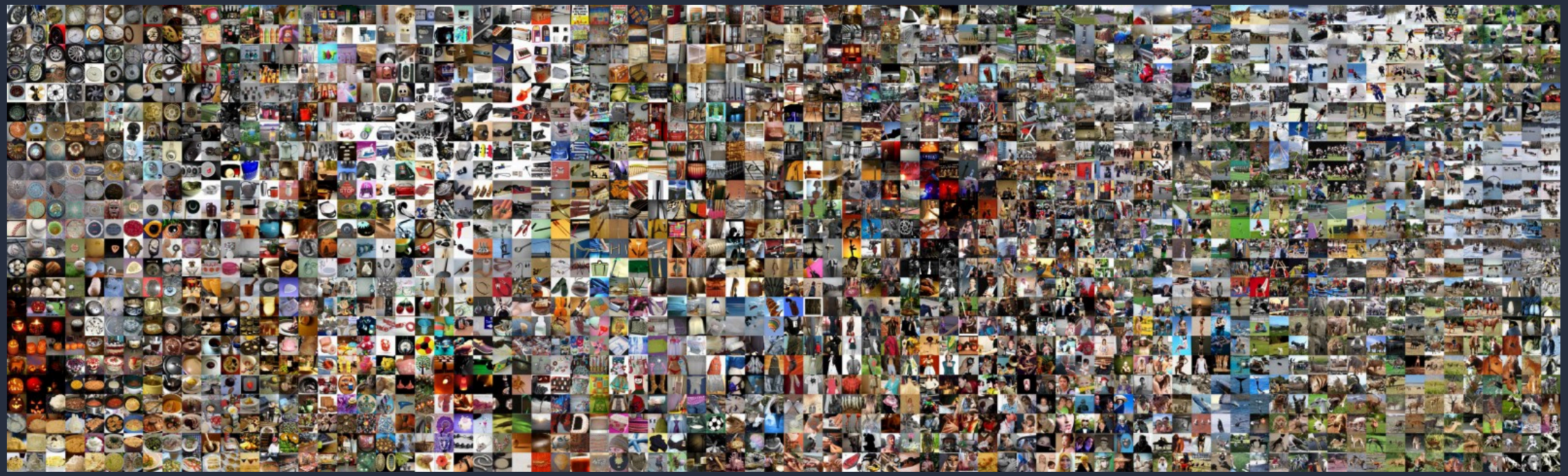
DOMINO

Model #1 – High Accuracy



DOMINO

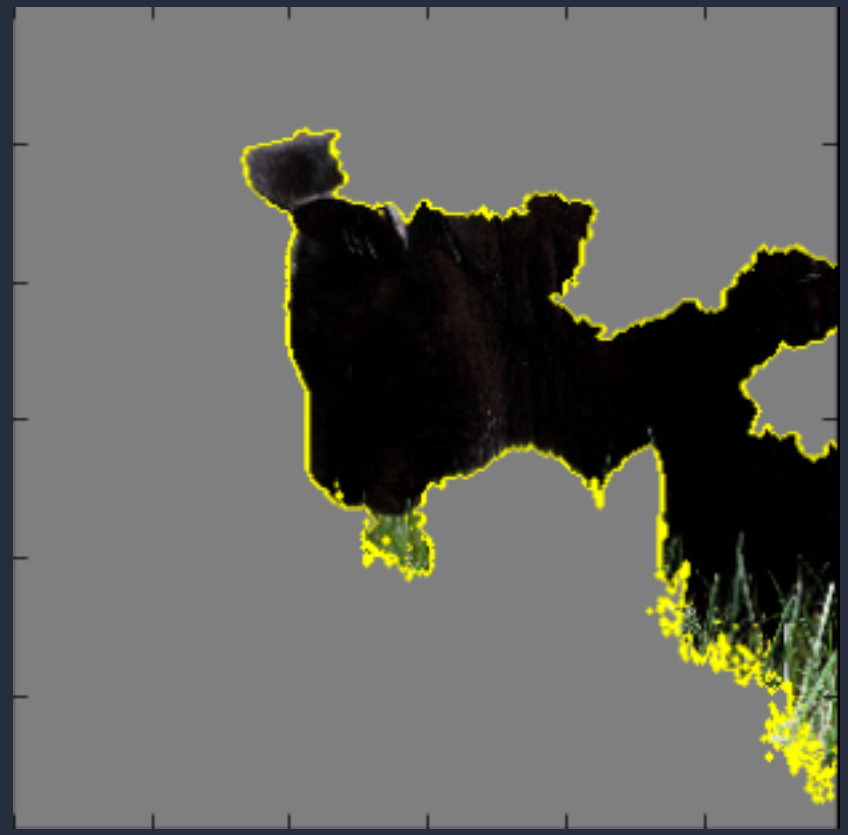
Model #2 – Interpretable with Global Inference



- In general, green objects with varying edge contrasts are plants
- In general, images with a lot of blue background are taken outside.

Model #3 – Interpretable with Local Inference

The model said this was a black bear

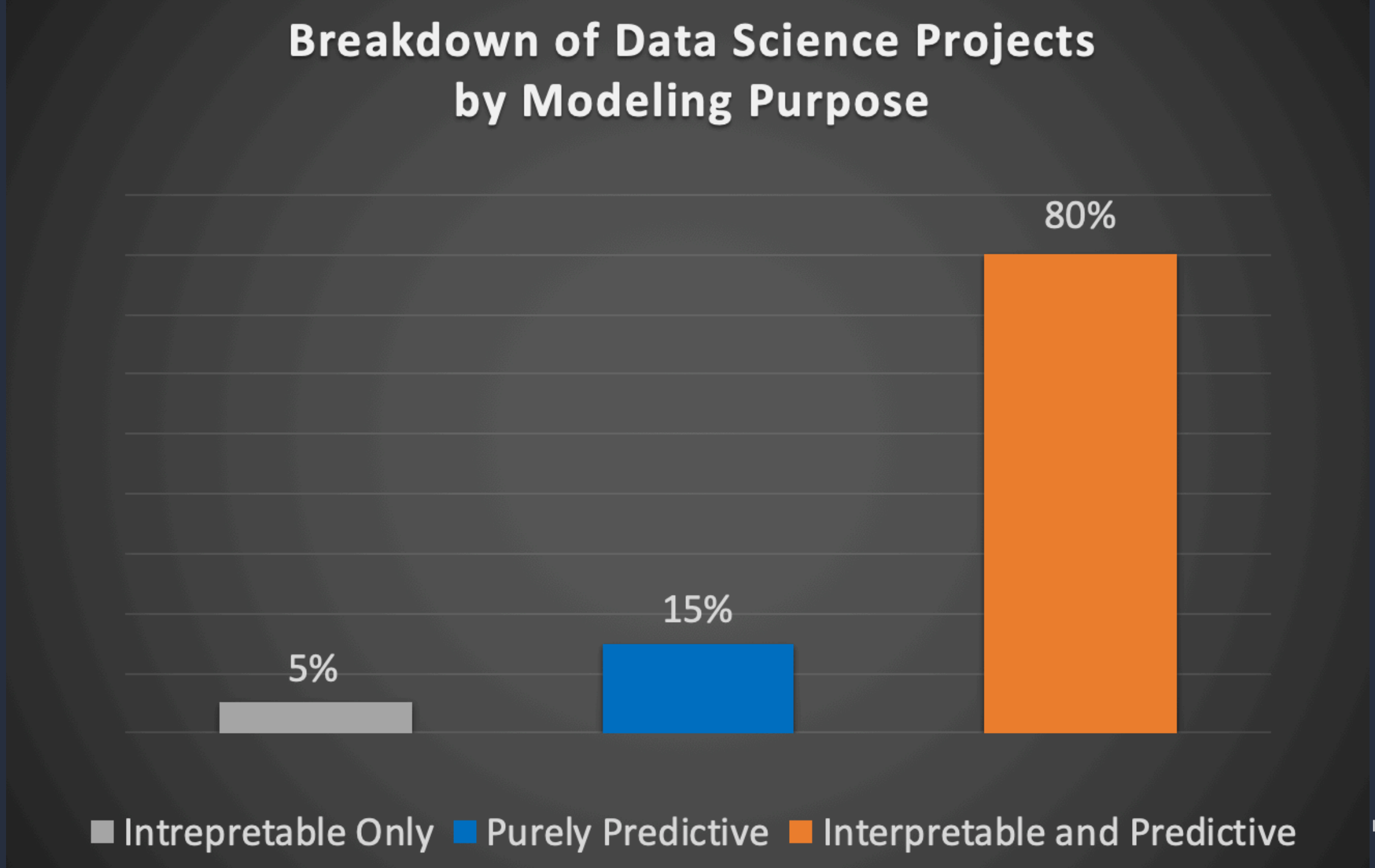


because of this super-pixel object

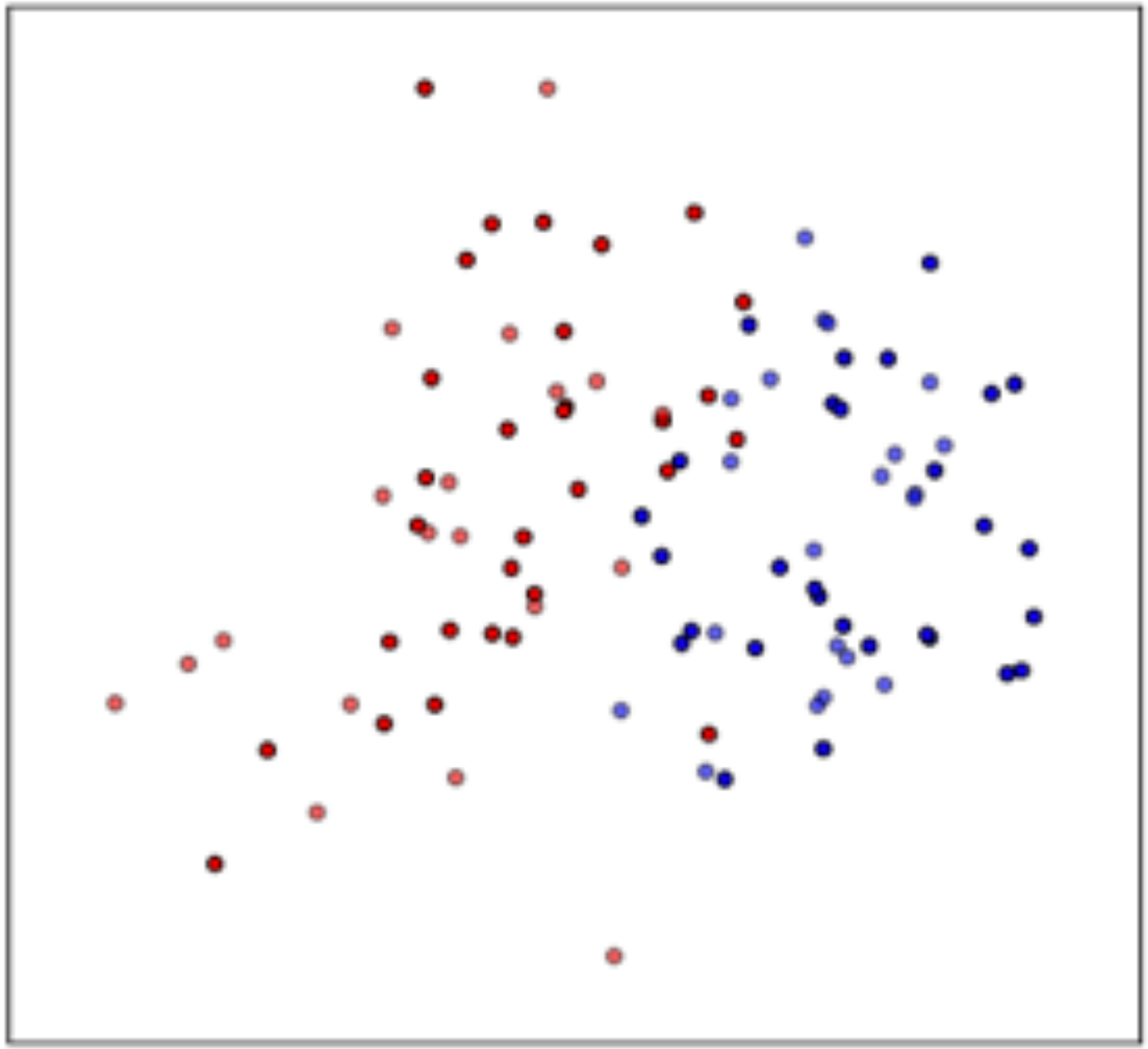
DOMINO

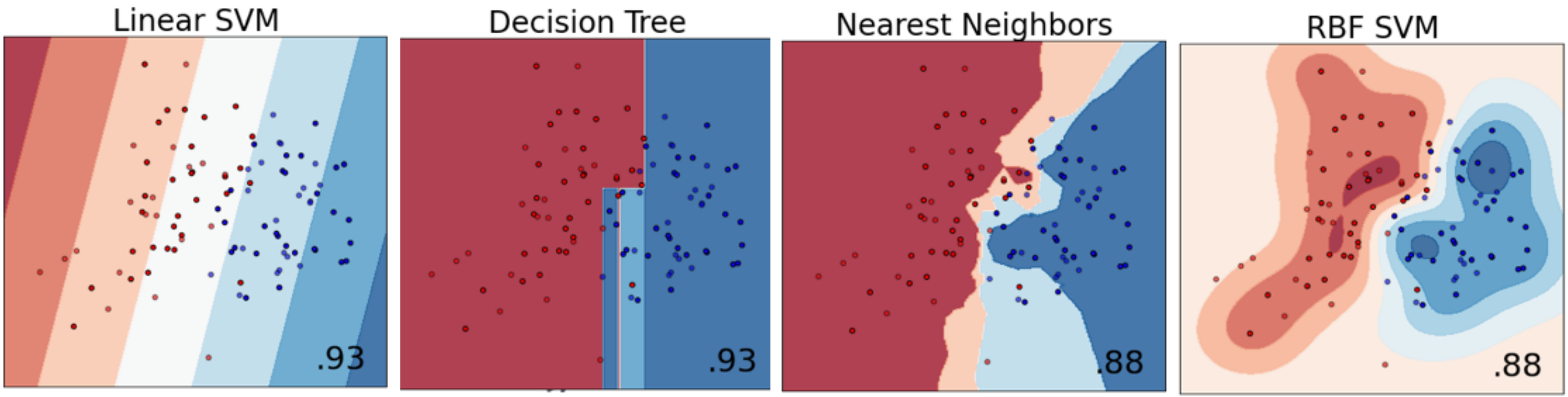

Interpret, Predict, or a Little of Both?
DOMINO

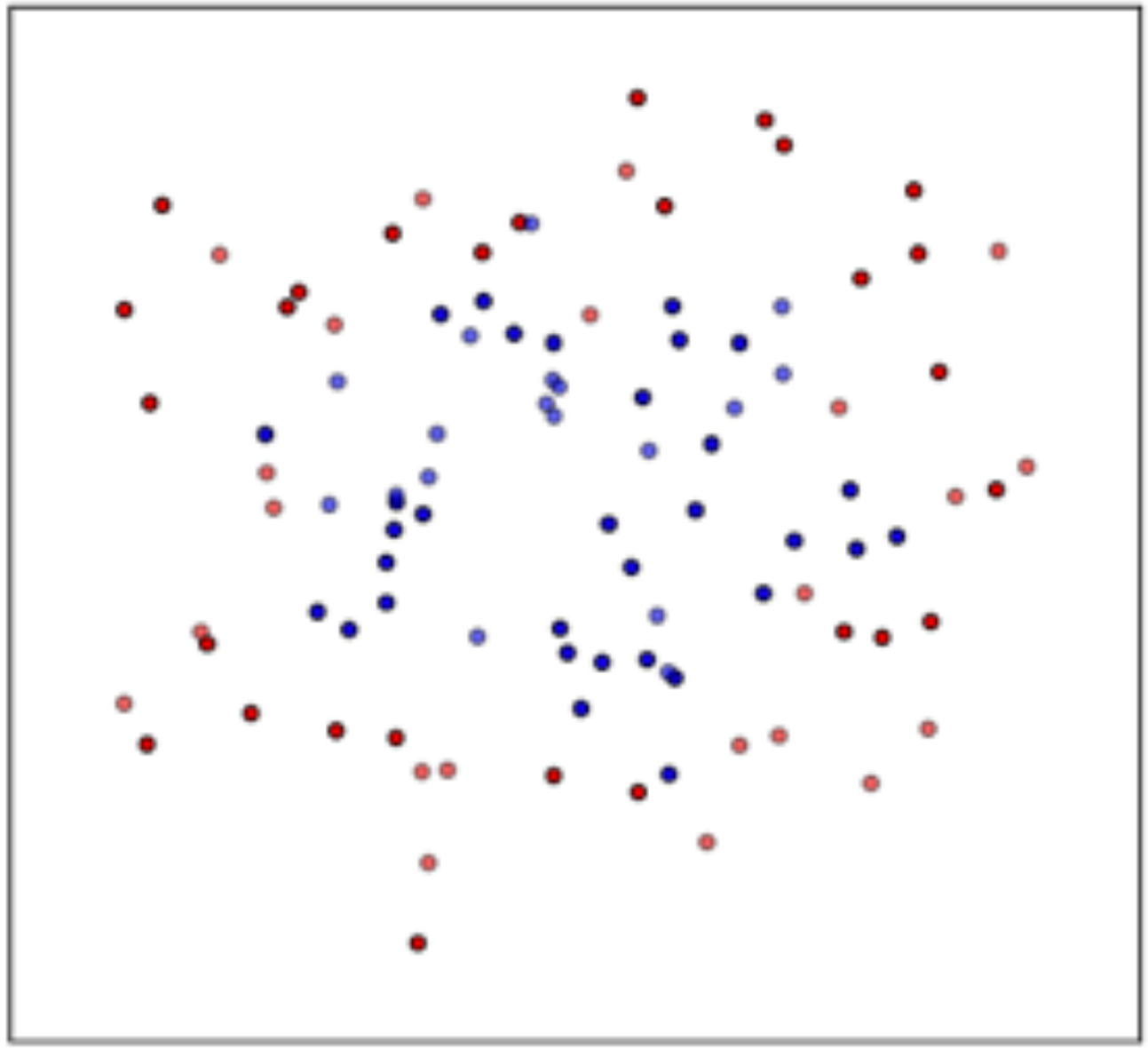

Interpret, Predict, or a Little of Both?
DOMINO

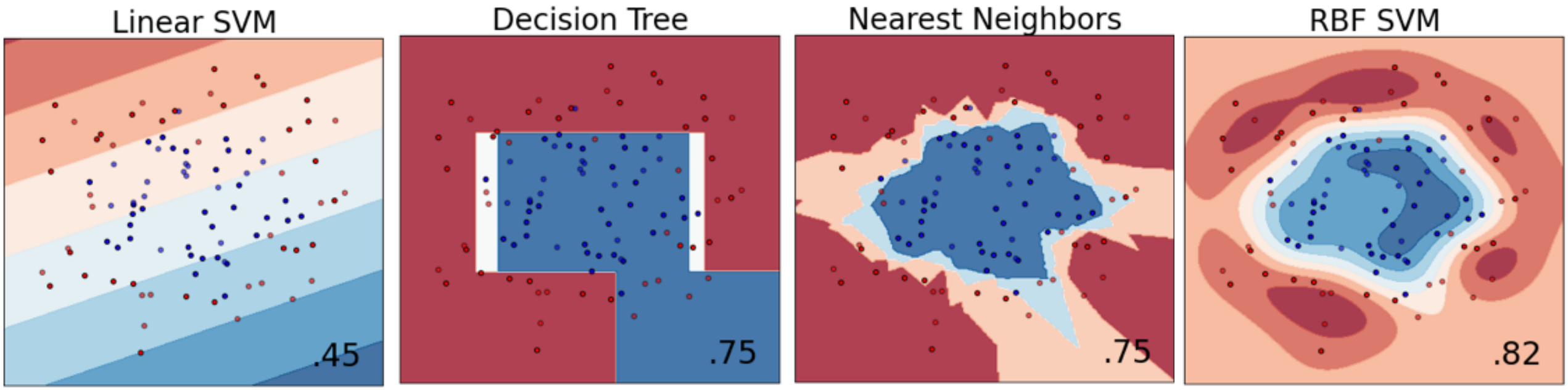

Breakdown of Data Science Projects
by Modeling Purpose
80%
15%
5%
Intrepretable Only
Purely Predictive
Interpretable and Predictive


POLICE
"Does your car have any idea why my car pulled it over?"
PAUL NOTH




Linear SVM
.93
Decision Tree
.93
Nearest Neighbors
.88
RBF SVM
.88




Linear SVM
.45
Decision Tree
.75
Nearest Neighbors
.75
RBF SVM
.82


HIGH DIVE
20 FEET

Define Interpretable and Predictive Models

Suppose there is a theory stating that inputs X cause response Y via the function, F.

$$Y=F(X)$$

Let x and y be measurable variables corresponding to samples taken from X and Y respectively, and let f be a model that approximates F.

- Interpretable modeling seeks f that best explains F

- Predictive modeling seeks the f that best predicts future y

Define Interpretable and Predictive Models

The holy grail is to find a model f that theoretically and structurally matches F and also perfectly predicts future y observations.

You won't find such an f

Trade-offs must be made

The Elements of Statistical Learning, 1st Ed. p196
Hastie, Tibshirani, Friedman


I'M STILL WORKING HERE
CAN'T YOU SEE HOW EXCITED I AM??
memegen.com

Fully Specified Models - Local Interpretability

Estimating house price based on square footage (X1) and bedrooms (X2)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

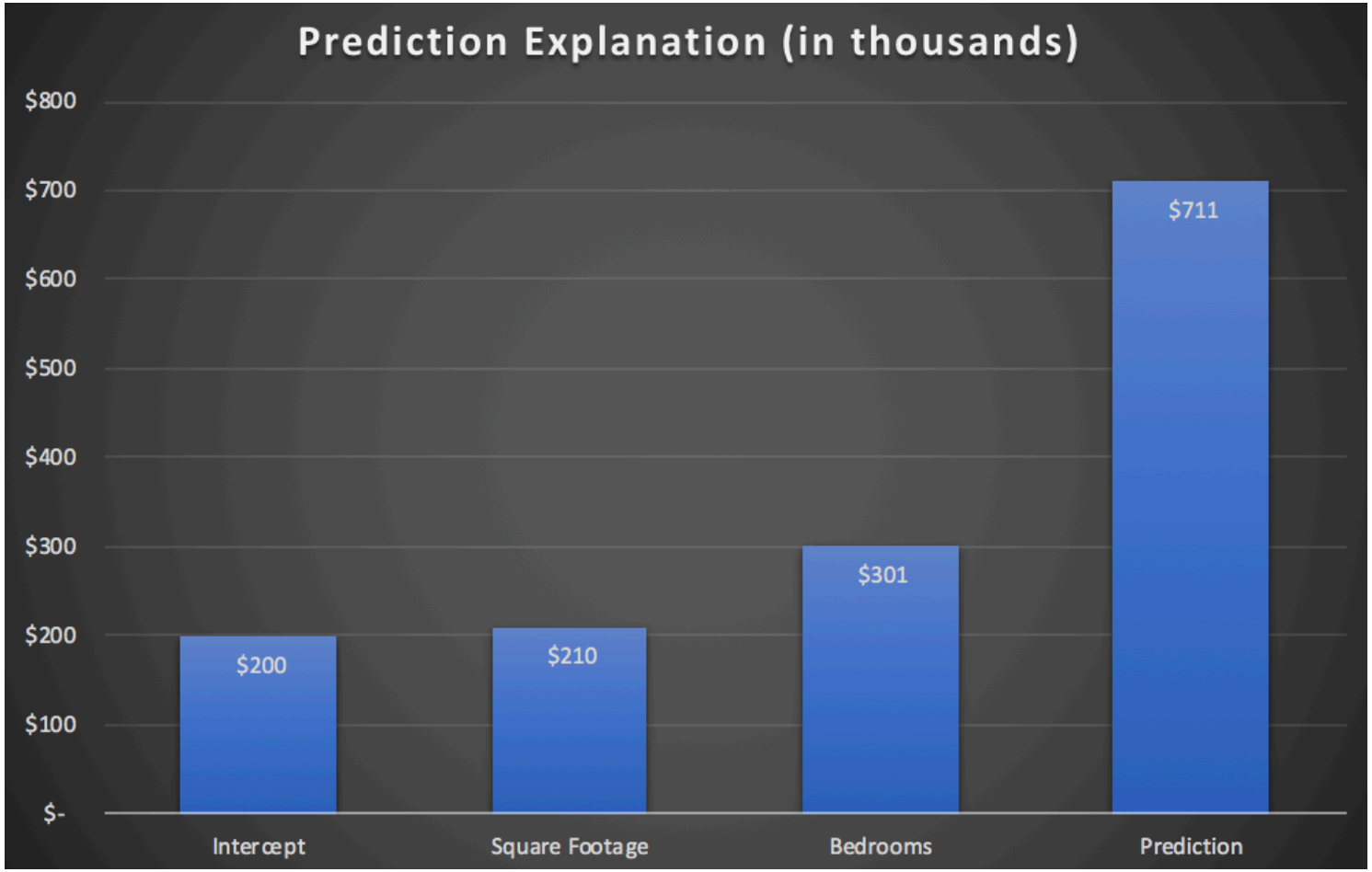
Fit to the data

$$y = 200 + 0.07 x_1 + 75.3 x_2$$

Make a prediction with 3000 sqft and 4 bedrooms.

200 + 0.07*3000 + 75.3*4 = $711.2K

Features may need to be normalized and model assumptions validated



Fully Specified Model Coefficients – Global Interpretability

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	**87.830**	6.385		13.756	.000	75.155	100.506
	age	**-.165**	.063	-.176	-2.633	.010	-.290	-.041
	weight	**-.385**	.043	-.677	-8.877	.000	-.471	-.299
	heart_rate	**-.118**	.032	-.252	-3.667	.000	-.182	-.054
	gender	**13.208**	1.344	.748	9.824	.000	10.539	15.877

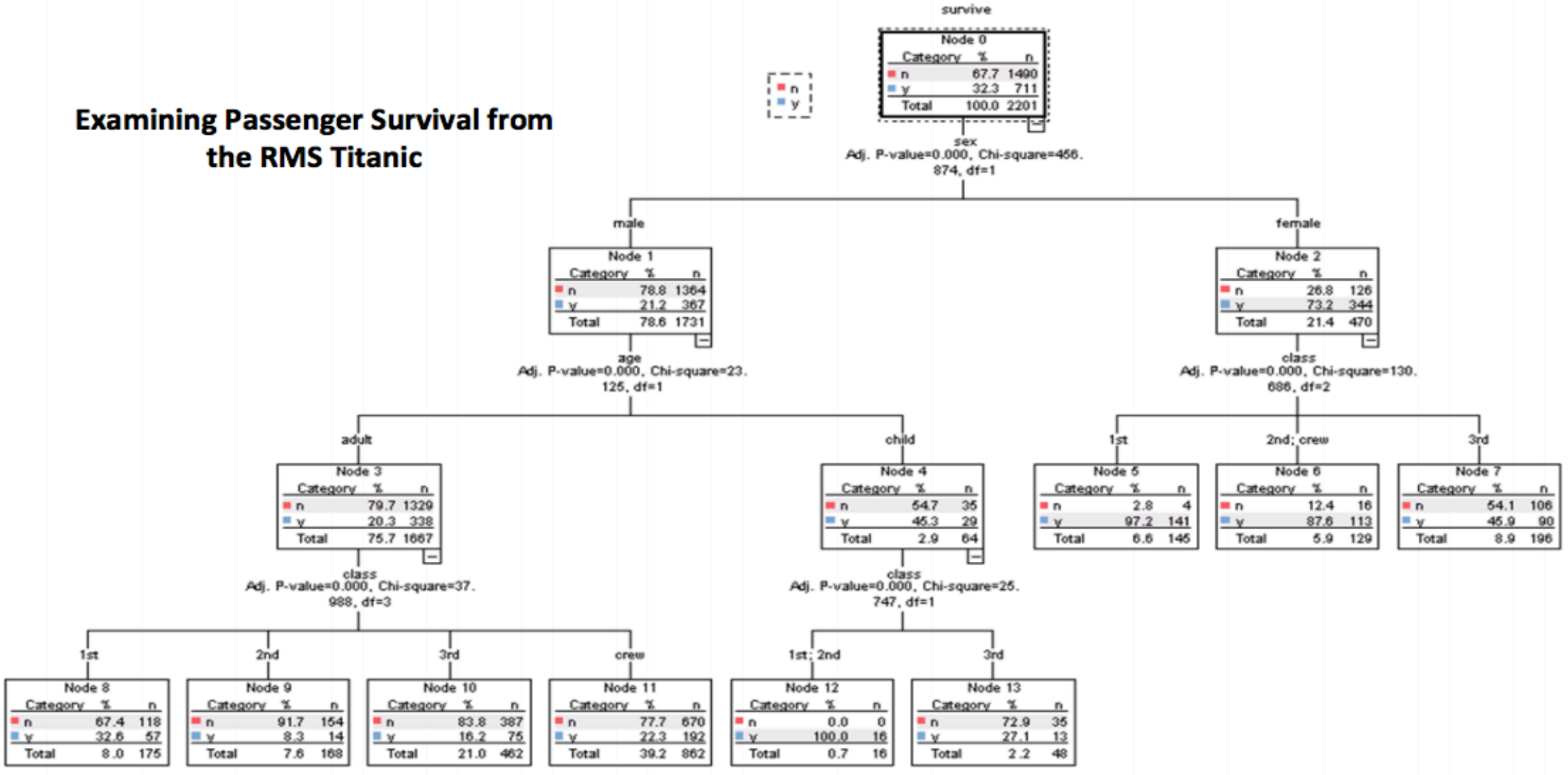
a. Dependent Variable: VO2max

Features may need to be normalized

and model assumptions validated

DOMINO

Decision Tree Nodes - Local and Global



Variable Importance - Global Interpretability

Model-specific

Accuracy

Gini

Variance

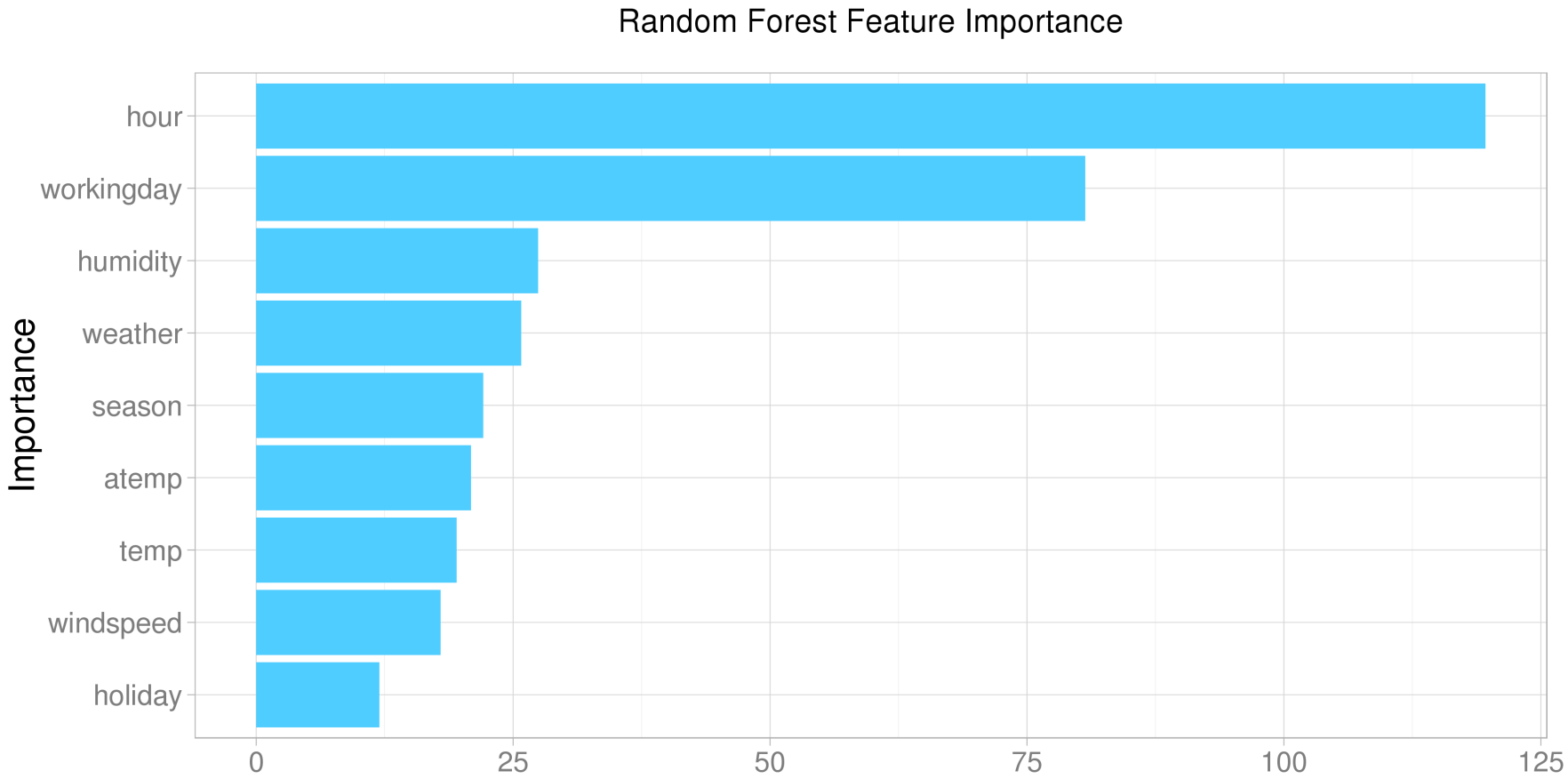
Gain

Feature value Permutations

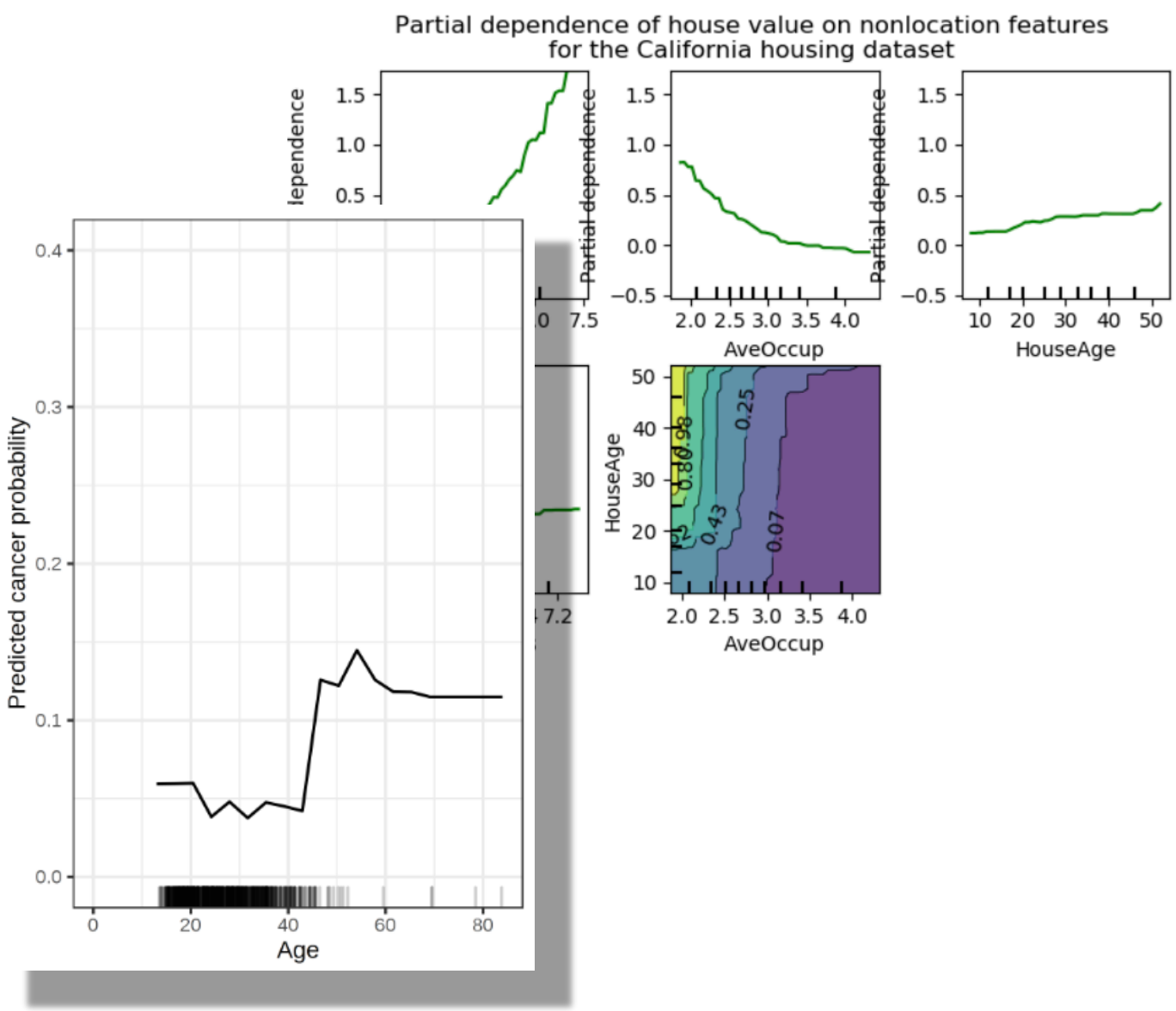
Spearman's correlation flavors

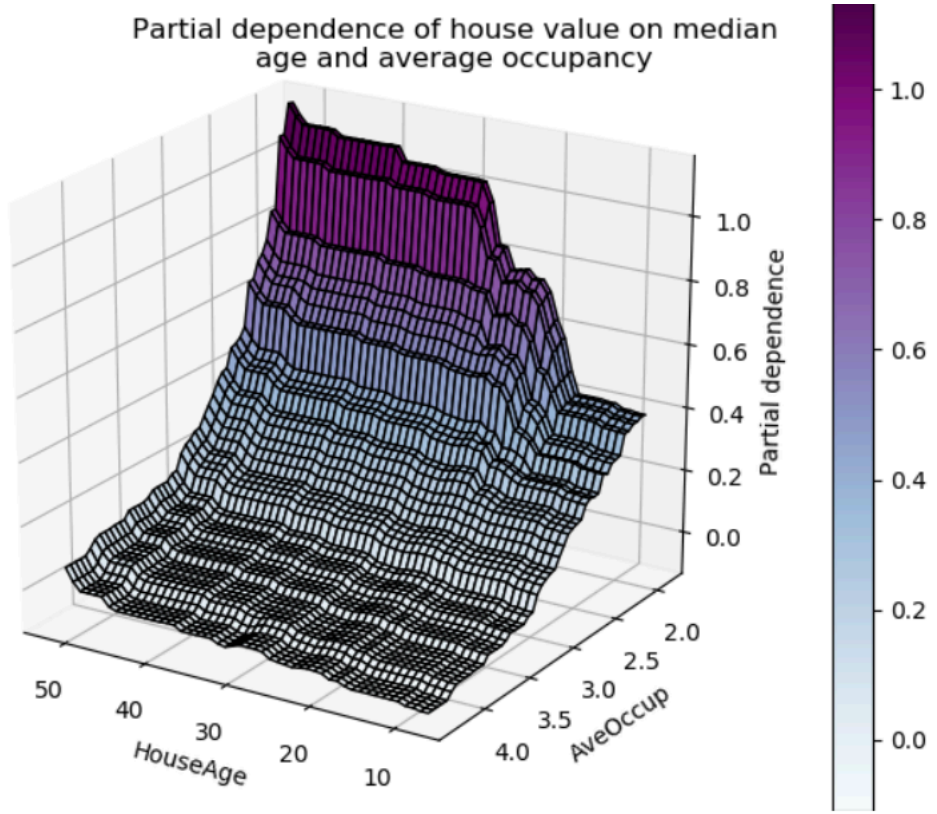
Model agnostic

permutations



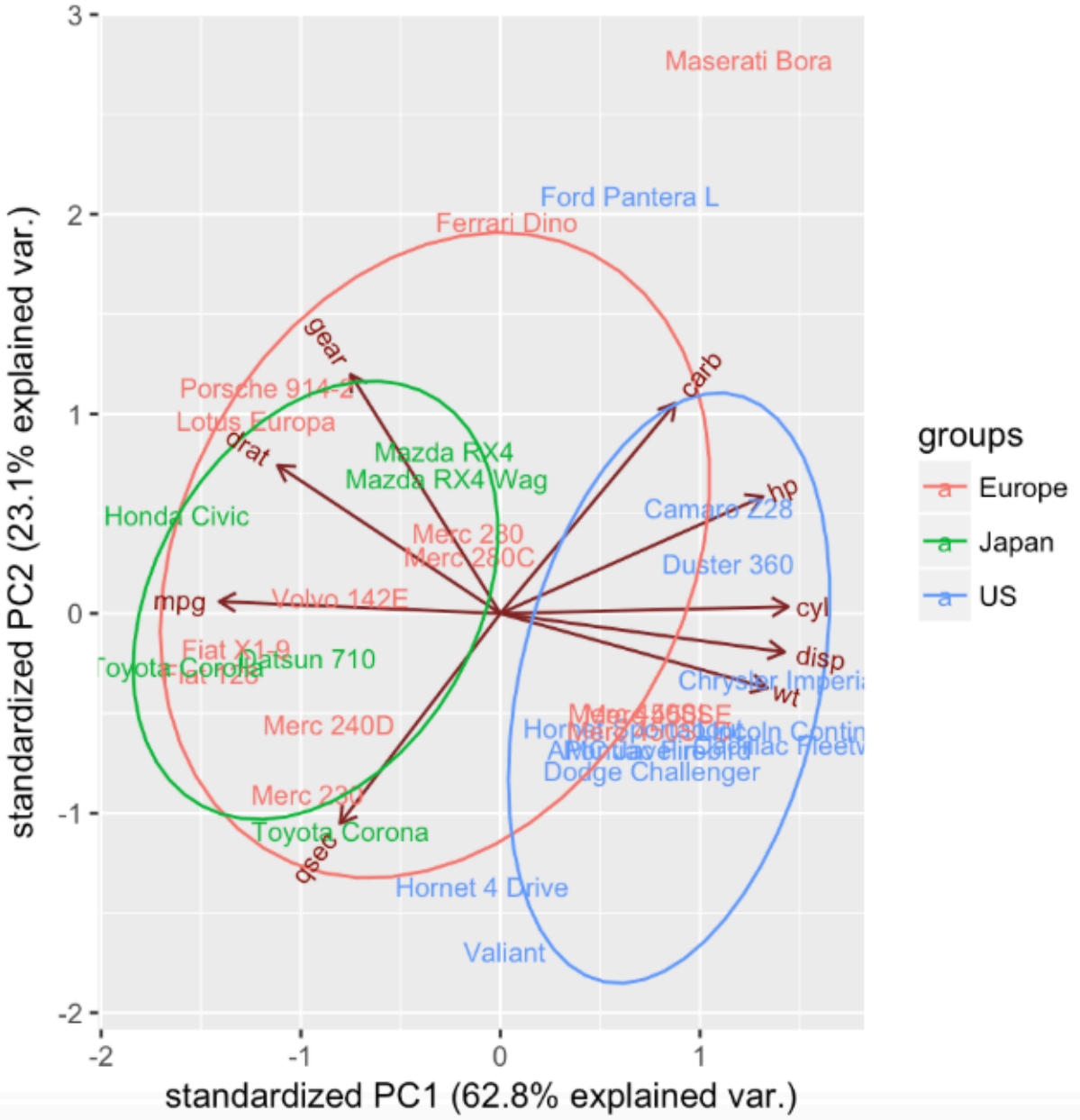
Partial Dependency Plots - Global Interpretability





DOMINO

PCA – Global Interpretability



DOMINO

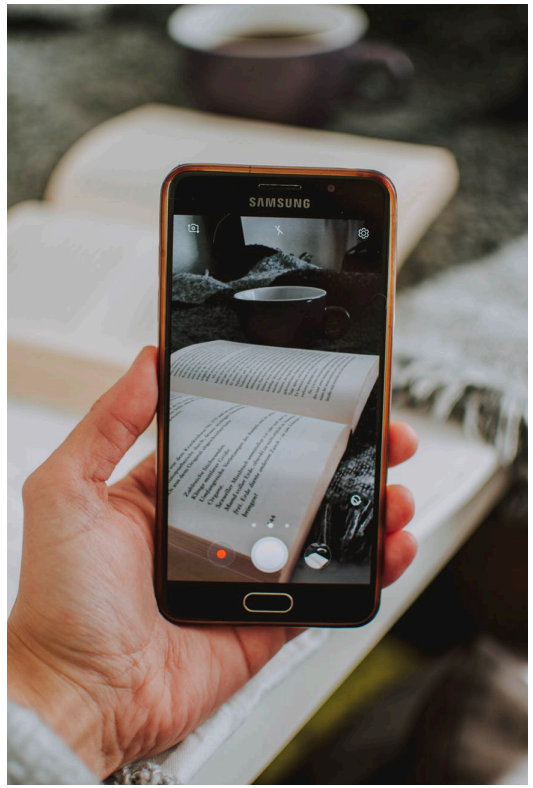
Tabular vs Image & Text

ML Models

Tabular Data

DNNs

Image & Text Data



DOMINO

A Note on Deep Neural Nets

Pixel Influence

- Saliency maps (reconstruction option)
- LIME & Anchor
- Deep SHAP
- Saliency Perturbations

Concept Extraction

- TCAV & ACE
- AILens
- CapsNet (reconstruction option)
- Interpretable CNNs (interpretability built-in)

Limitations of Saliency/Pixel Influence Approaches

1) Local exaplainability only
2) No control over concepts these maps pick
3) Saliency maps produced by randomized networks are similar to that of the trained network (Adebayo et al., 2018)
4) Simple meaningless data processing steps, may cause saliency methods to result in significant changes (Kindermans et al., 2017).
5) Saliency maps may also be vulnerable to adversarial attacks (Ghorbani et al., 2017).

ML Models DNNs

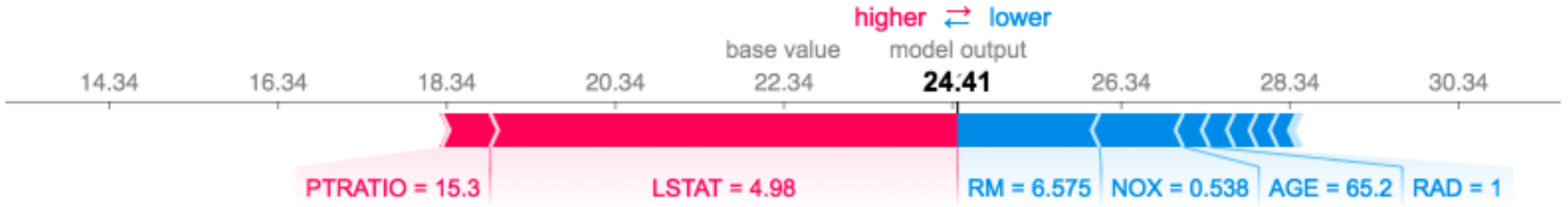
SHAP (SHapley Additive exPlanations)

Scott Lundberg, PhD candidate in the Paul G. Allen School of Computer Science & Engineering at the University of Washington

- Built on the idea of Shapley values
  - the average marginal contribution of a feature value over all possible coalitions
- SHAP is an approximation that assigns each feature an importance value for a prediction
- Only approach to guarantee the properties of consistency and accuracy
- Works on any black box model
- Very computationally expensive for some models

ML Models DNNs

SHAP - Local Interpretability

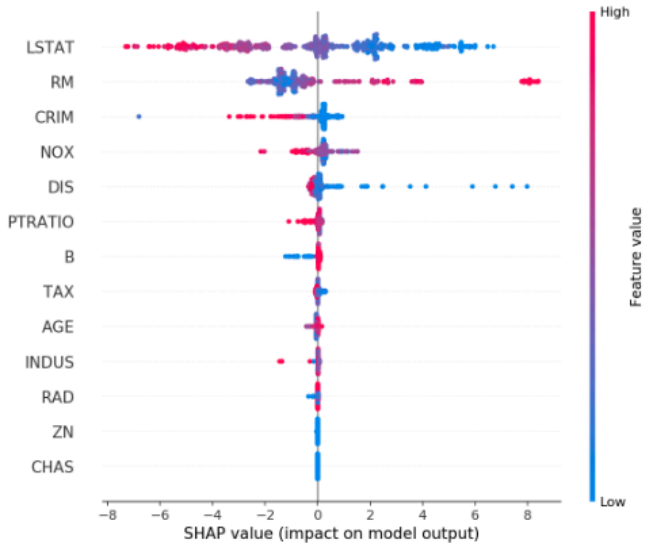


DOMINO

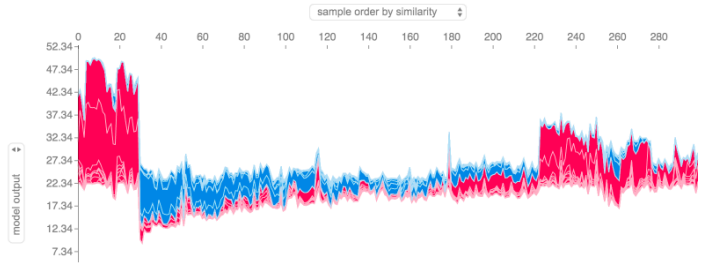
ML Models DNNs

SHAP - Global Interpretability

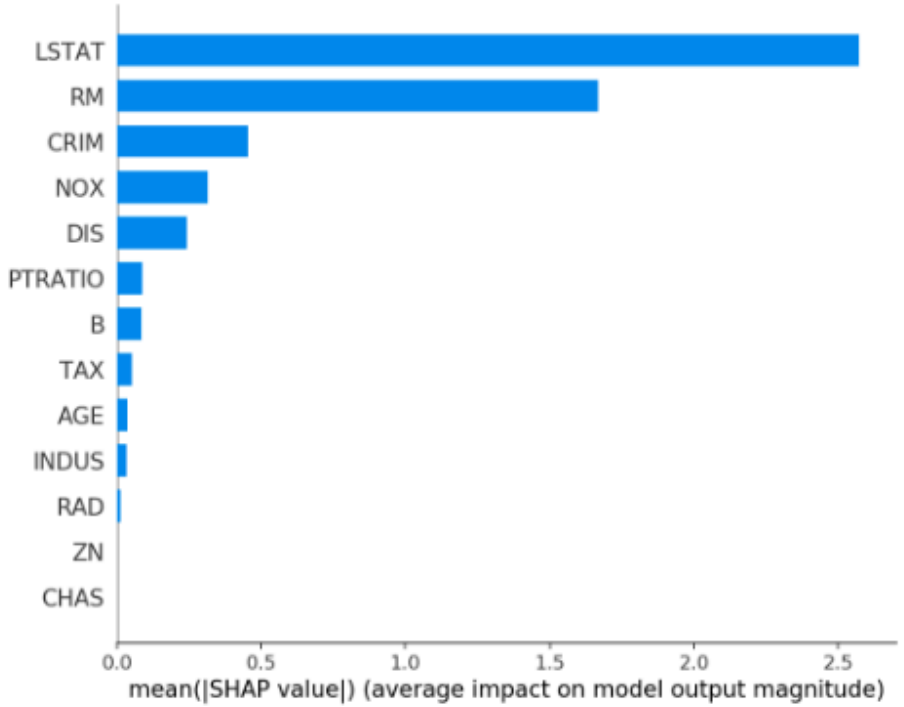
```
# summarize the effects of all the features
shap.summary_plot(shap_values, X)
```



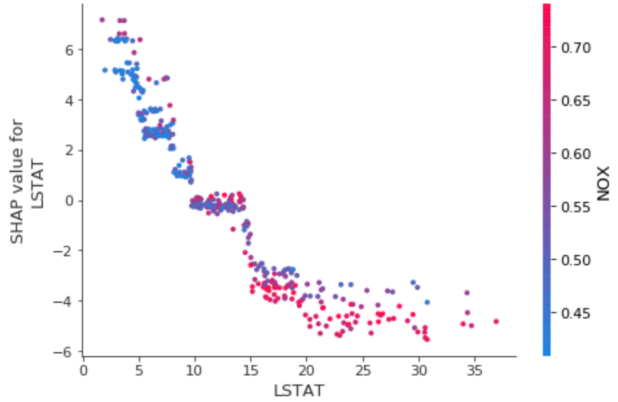

```
# visualize the training set predictions
shap.force_plot(explainer.expected_value, shap_values, X)
```




```
shap.summary_plot(shap_values, X, plot_type="bar")
```




```
shp_plt = shap.dependence_plot("LSTAT", shap_values_train,
```



ML Models DNNs

SHAP

- http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions
- https://christophm.github.io/interpretable-ml-book/shapley.html
- https://github.com/slundberg/shap
- https://blog.dominodatalab.com/shap-lime-python-libraries-part-1-great-explainers-pros-cons/
- https://blog.dominodatalab.com/shap-lime-python-libraries-part-2-using-shap-lime/
- https://arxiv.org/pdf/1602.04938.pdf

ML Models DNNs

LIME (Local Interpretable Model-agnostic Explanations)

Marco Tulio Ribeiro Ph.D., Researcher at Microsoft Research, in the Adaptive Systems and Interaction group and Affiliate Assistant Professor at the University of Washington

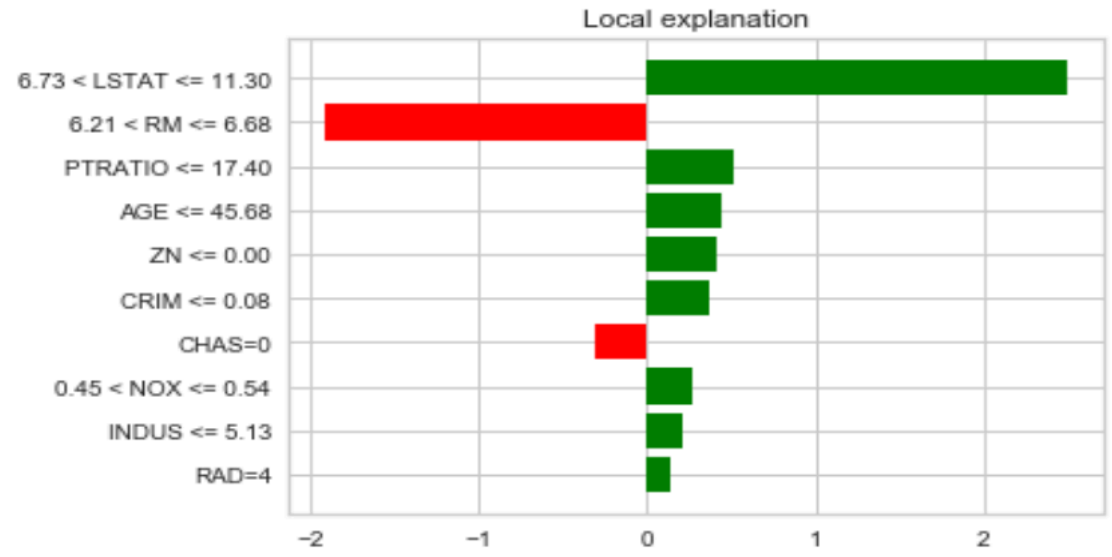
- Builds sparse linear models around each prediction to explain how the black box model works in that local vicinity
- v 0.1.1.34 of the python package doesn't work out-of-the-box on all models
- Can be thought of as an approximation to Shapley values
- Much faster execution times
- Output is easier for most people to understand

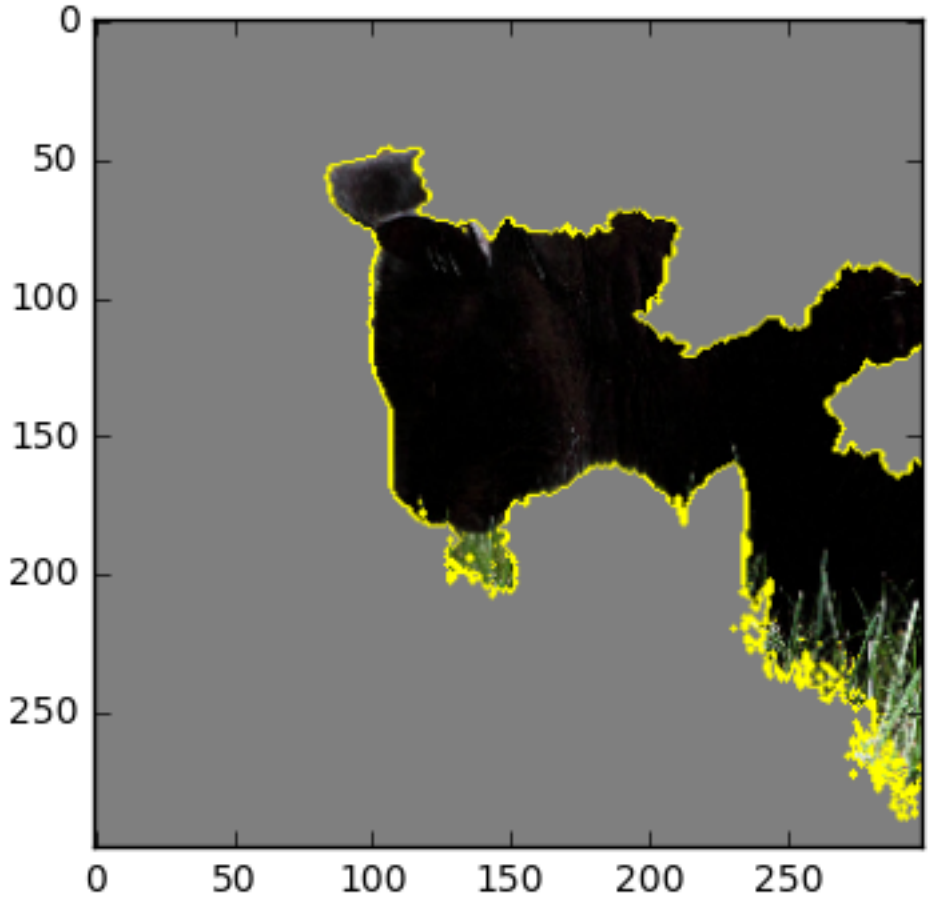
ML Models DNNs

LIME - Local Interpretability

```
In [59]: exp.as_pyplot_figure()
```

Out[59]:





DOMINO

ML Models DNNs

LIME

- https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf
- https://cran.r-project.org/web/packages/lime/vignettes/Understanding_lime.html
- https://github.com/marcotcr/lime
- https://blog.dominodatalab.com/shap-lime-python-libraries-part-1-great-explainers-pros-cons/
- https://blog.dominodatalab.com/shap-lime-python-libraries-part-2-using-shap-lime/
- https://arxiv.org/pdf/1602.04938.pdf

ML Models DNNs

Anchor

Marco Tulio Ribeiro Ph.D., Researcher at Microsoft Research, in the Adaptive Systems and Interaction group and Affiliate Assistant Professor at the University of Washington

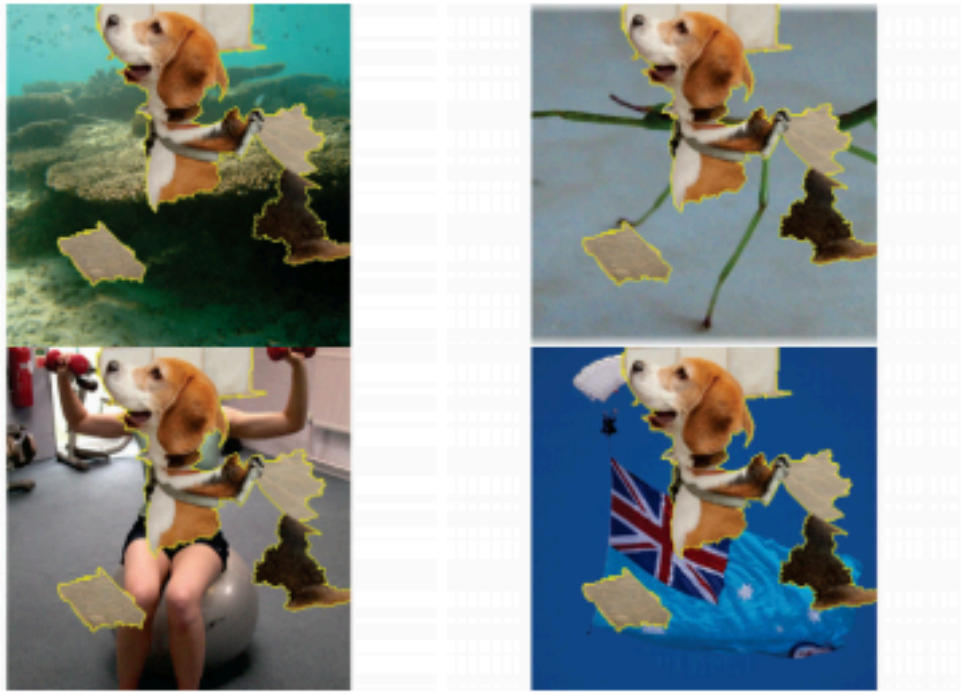
- Model-agnostic local explanations based on if-then rules
- Highlights the part of the input that is sufficient for the classifier to make the prediction, making them intuitive and easy to understand.
- For images, start with LIME pixel influence, then test for image anchoring super-pixels

ML Models DNNs

Anchor

(a) Original image

(b) Anchor for "beagle"



(c) Images where Inception predicts $P(\text{beagle}) > 90\%$

ML Models DNNs

Anchor

	If	Predict
adult	No capital gain or loss, never married	$\leq$ 50K
	Country is US, married, work hours $>$ 45	$>$ 50K
rcdv	No priors, no prison violations and crime not against property	Not rearrested
	Male, black, 1 to 5 priors, not married, and crime not against property	Re-arrested
lending	FICO score $\leq$ 649	Bad Loan
	649 $\leq$ FICO score $\leq$ 699 and \$5, 400 $\leq$ loan amount $\leq$ \$10, 000	Good Loan

Table 3: Generated anchors for Tabular datasets

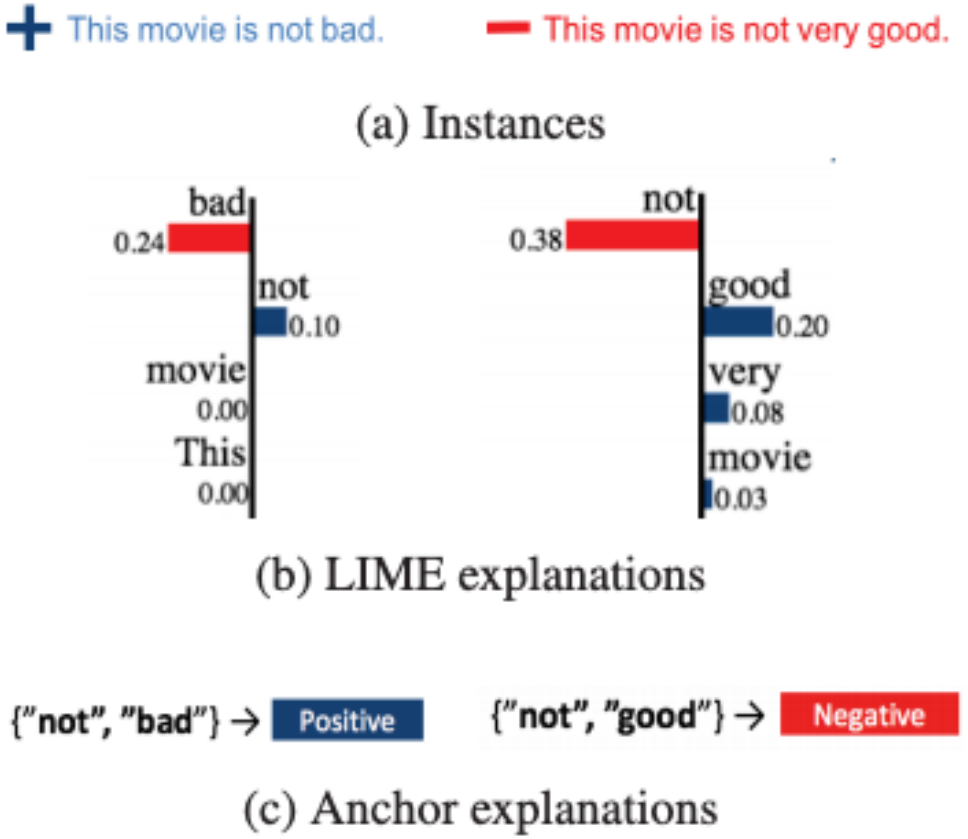


Figure 1: Sentiment predictions, LSTM

DOMINO

ML Models DNNs

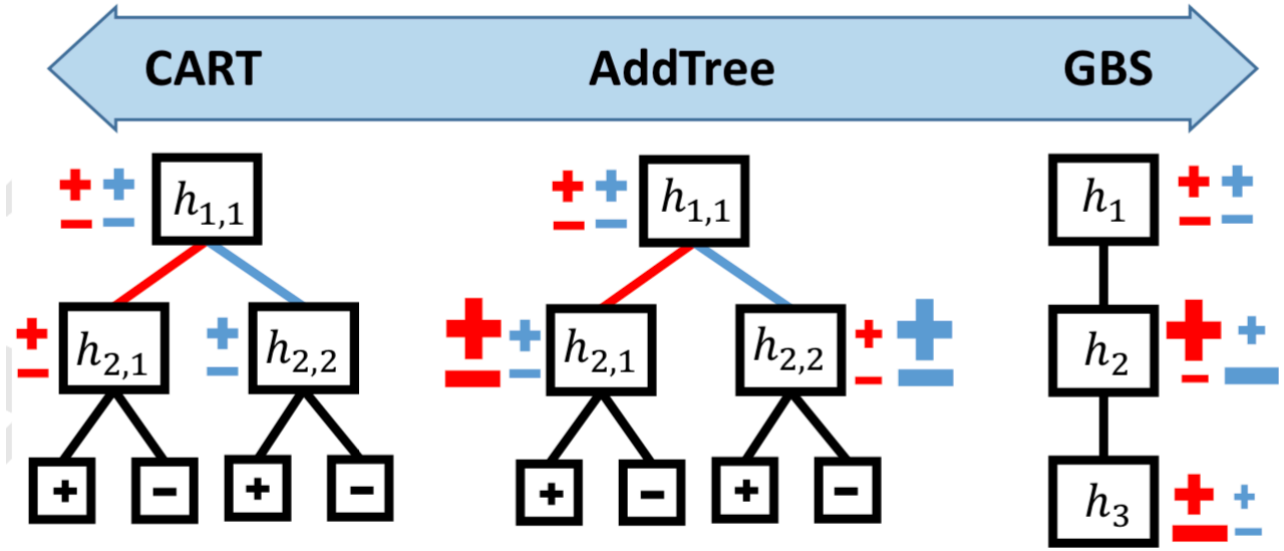
Anchor

https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982/15850
https://github.com/marcotcr/anchor (python - code is still raw as of Jan 2020)

AddTree

Gilmer Valdes, Ph.D., Assistant Professor Physics Department of Radiation Oncology UCSF Medical Center

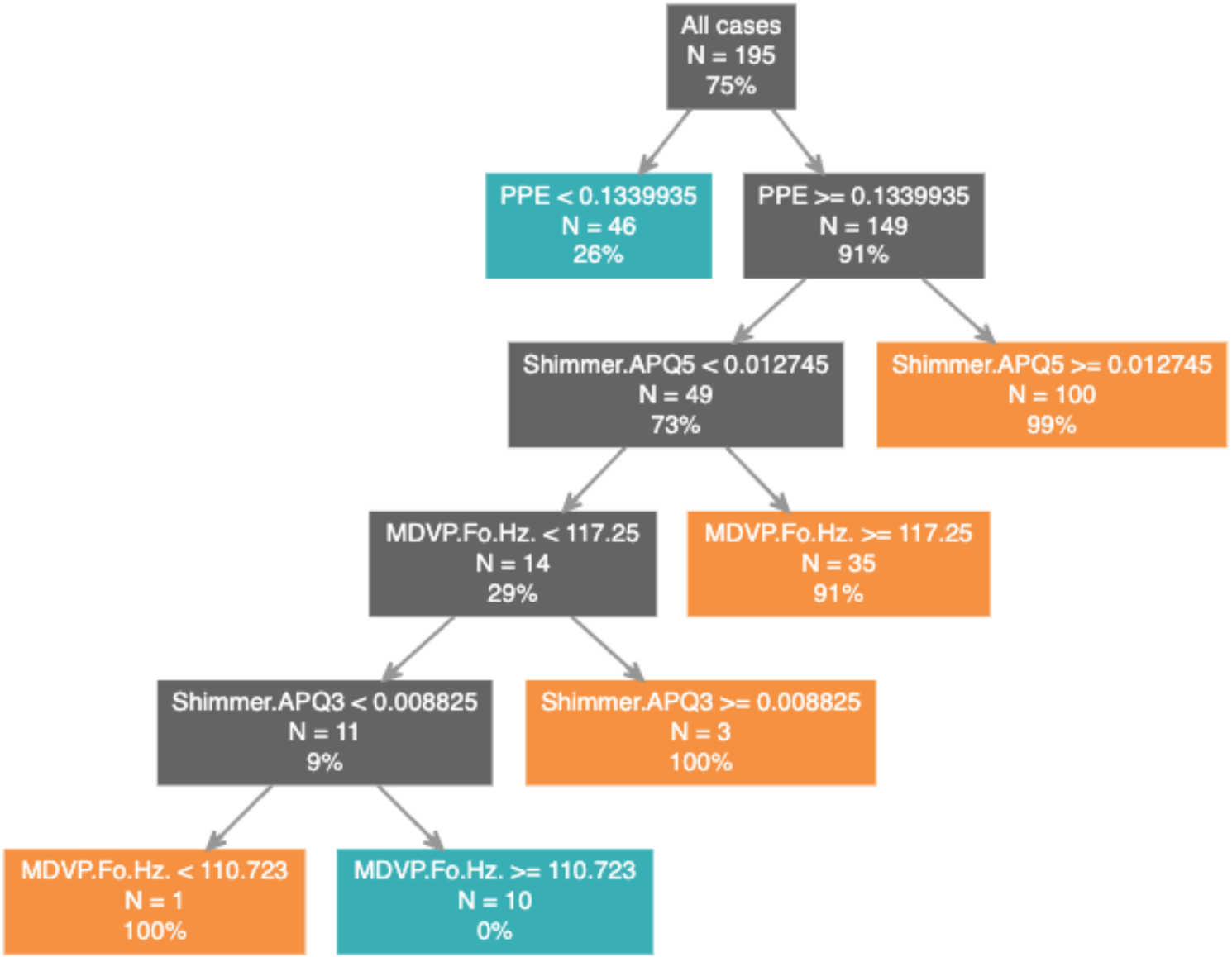
- Not an interpretability tool, but an algorithm with interpretability as a hyperparameter
- Grows a single, highly interpretable tree with accuracy close to ensemble methods
- GBT is additive. CART is full interactive. Show that these models exist along a spectrum, revealing previously unseen connections between these approaches
- MATLAB library for MediBoost &
  R library for AddTree



AddTree - Local Interpretability

```
                                                levelName Estimate
1  All cases                                                       1
2   ¦--PPE < 0.1339935                                             0
3   °--PPE >= 0.1339935                                            1
4       ¦--Shimmer.APQ5 < 0.012745                                 1
5       ¦   ¦--MDVP.Fo.Hz. < 117.25                               0
6       ¦   ¦   ¦--Shimmer.APQ3 < 0.008825                        0
7       ¦   ¦   ¦   ¦--MDVP.Fo.Hz. < 110.723                      1
8       ¦   ¦   ¦   °--MDVP.Fo.Hz. >= 110.723                     0
9       ¦   ¦   °--Shimmer.APQ3 >= 0.008825                      1
10      ¦   °--MDVP.Fo.Hz. >= 117.25                             1
11      °--Shimmer.APQ5 >= 0.012745                               1
```

AddTree - Global Interpretability



AddTree

https://www.nature.com/articles/srep37854
https://egenn.github.io/rtemis/rtemis_mediboost_vignette.html
https://egenn.github.io/rtemisdocs/index.html
www.mediboostml.com
https://www.pnas.org/content/116/40/19887
https://rtemis.netlify.com/addtree
https://rtemis.netlify.com/setup.html
https://arxiv.org/abs/1903.09731

- Last link is to new research from this team – a paper on Expert-Augmented Machine Learning (EAML), an automated method that guides the extraction of expert knowledge and its integration into machine-learned models.

- For a somewhat similar approach that uses DNNs to build a single tree see GENESIM
- https://arxiv.org/pdf/1611.05722.pdf

InterpretML

Equal contributions: Samuel Jenkins & Harsha Nori & Paul Koch & Rich Caruana. Open sourced May 15, 2019.

- GA²M, like AddTree, is not an interpretability technique but a modeling approach aimed at simultaneous accuracy and interpretability

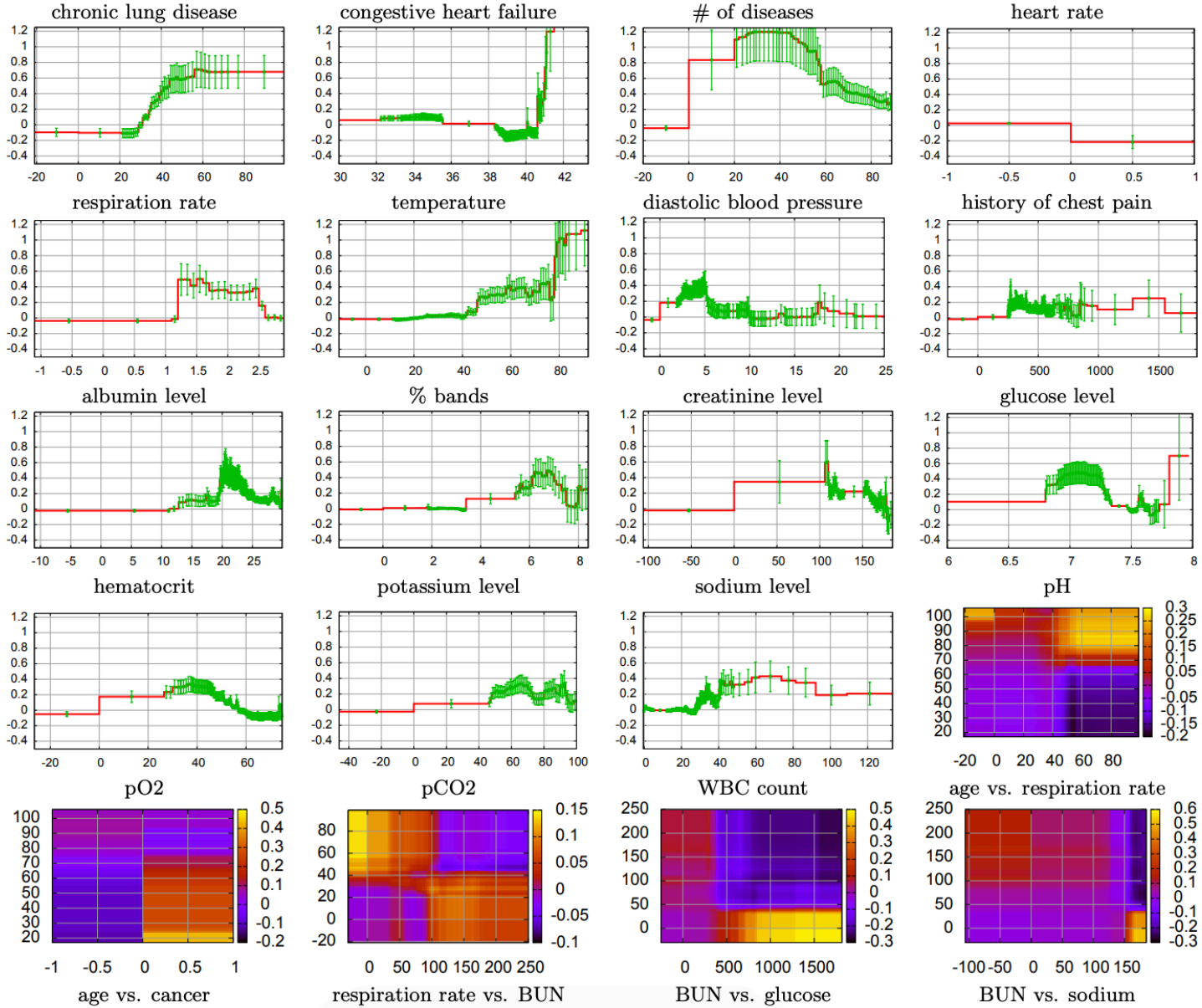
- GAMs are arguably less interpretable than GLMs due to the parameter function, but can provide a better fit

$$g(\mathbf{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

- Each f_i is often a spline but can also be a regression tree or a boosted tree

- GA²M adds pairwise interactions to GAM by ranking all possible pairs of interactions in the residuals (top k selected via cross validation)

GA²M - Global Interpretability



InterpretML

- https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/KDD2015FinalDraftIntelligibleModels4HealthCare_igt143e-caruanaA.pdf
- https://en.wikipedia.org/wiki/Generalized_additive_model
- https://github.com/microsoft/interpret
- https://drive.google.com/drive/u/2/folders/1PF6GEWF9KK_VPMwAYG0pylpmKIJVIKSL

Testing with Concept Activation Vectors (TCAV)

Been Kim, Ph.D., Research Scientist at Brain. Debuted at ICML 2018. Follow-up work (ACE) published at NeurIPS 2019.

- Global explainability. Typically on images. Assessing the importance of high-level, human interpretable concepts.
  - e.g. wheels and decals to identify a police van or strips to identify a zebra
- Post training, so not limited to feature inputs and not limited to learning from training examples (you provide your own concept examples)
- Derive CAVs by training a linear classifier between a concept's examples and random counter examples and then taking the vector orthogonal to the decision boundary.
  - If I make this picture more like the concept will the probability of the label change?

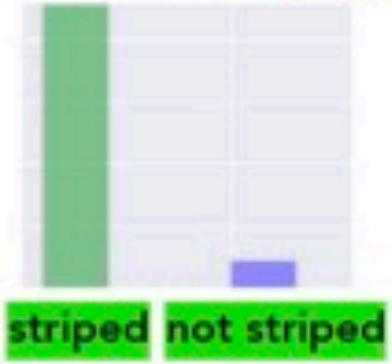
A trained machine learning model (e.g., neural network)

$p(z)$

zebra-ness

Was striped concept important to this zebra image classifier?

TCAV score for **Zebra**



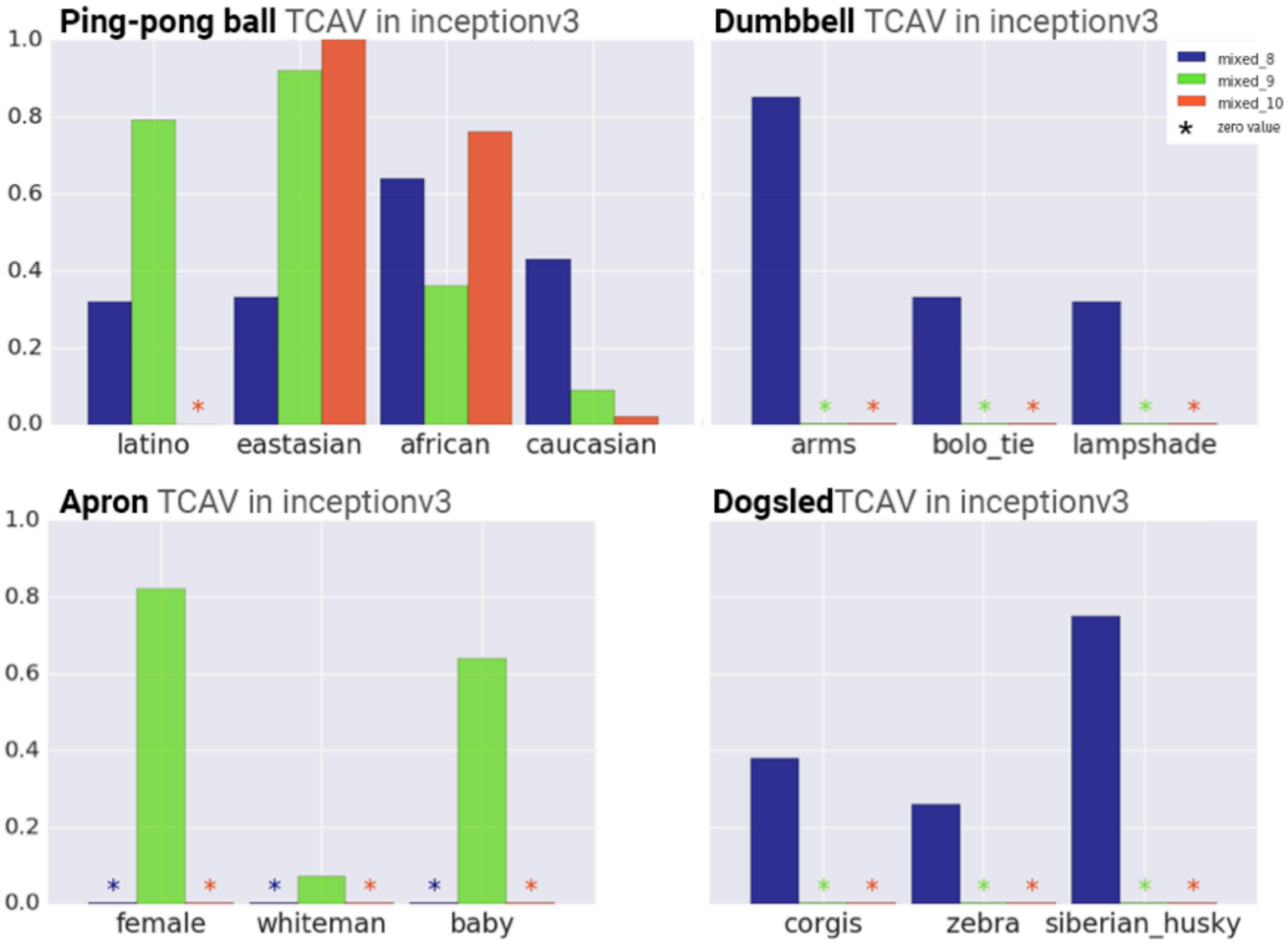
TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

Testing with Concept Activation Vectors (TCAV)

- Not a push-button solution. User must:
  - Hand-design the concept and provide 50 – 200 examples that represent that concept and 50 – 200 random examples that do not represent that concept
  - Pick intermediate model layers to use for TCAV
  - Write your own model wrapper with loss function and data handling
  - Write a class that returns your activations

- Other drawbacks:
  - TCAV works if and only if your network has learned about the concept
  - Might introduce bias in choice of concept examples

TCAV – Global Interpretability



Testing with Concept Activation Vectors (TCAV)

- https://youtu.be/Ff-Dx79QEEY
- https://www.youtube.com/watch?v=DNk-hcSV1pY
- http://proceedings.mlr.press/v80/kim18d/kim18d.pdf
- https://arxiv.org/abs/1711.11279
- https://www.slideshare.net/SessionsEvents/interpretability-beyond-feature-attribution-quantitative-testing-with-concept-activation-vectors-tcav-123005780
- https://github.com/tensorflow/tcav
- https://github.com/tensorflow/tcav/blob/master/Run%20TCAV.ipynb

ACE

Amirata Ghorbani, PhD student at Stanford under Dr. James Zou

- Automated Concept-based Explanation (ACE) seeks to eliminate the need for hand-designed concepts and human bias in concept selection such as used in TCAV.
- Takes a trained classifier and a set of images of a class as input. Extracts concepts (usually in the form of pixel groups) in that class and returns each concept's importance.
- Current steps to implement are manual and not trivial. TCAV is the last step and can be replaced by another method.
- Showed that concepts are sufficient for prediction across tested models (similar result as Anchors)

ACE with TCAV

ACE

- https://arxiv.org/abs/1902.03129
- https://github.com/amiratag/ACE

AILens

Anupam Datta, Ph.D., Professor, Electrical and Computer Engineering Department, CMU Silicon Valley

- *Influence-directed explanations* for deep networks (CNNs)
- Combine traditional input influence for a single instance and activation neuron identification for higher-level concepts
- Identify neurons with high influence and visualization techniques to provide an interpretation for the concept they represent
- Isolate regions across multiple images that share a common concept and influence the prediction

AILense - Global and Local Interpretability

AILens

- https://www.ailens.io/
- http://www.andrew.cmu.edu/user/danupam/datta-sen-zick-oakland16.pdf
- https://arxiv.org/pdf/1802.03788.pdf
- https://www.andrew.cmu.edu/user/danupam/

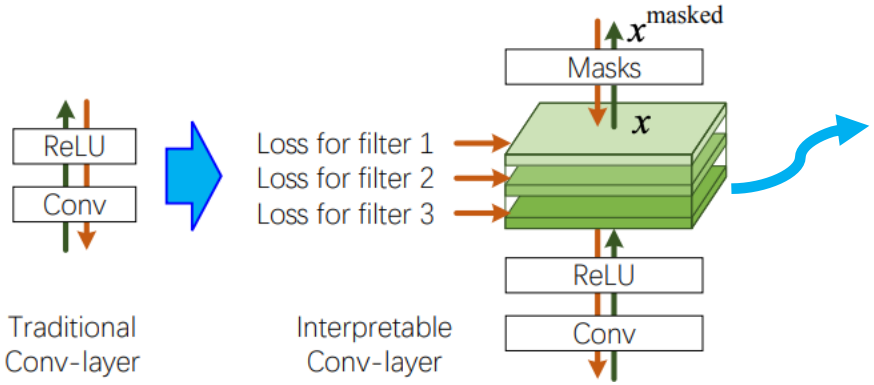
Interpretable CNNs

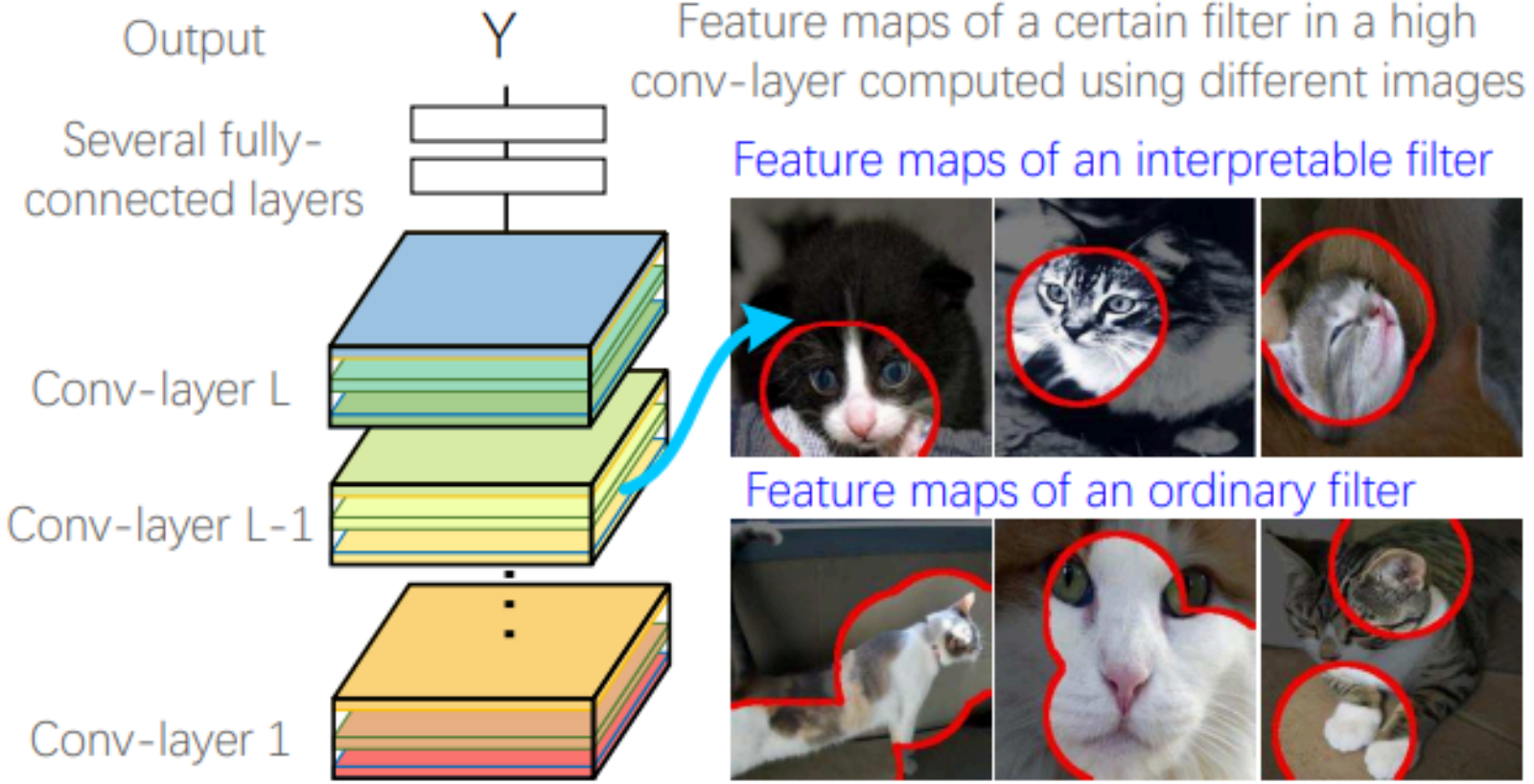
Quanshi Zhang, postdoctoral researcher at UCLA

- Without any additional human supervision, uses filters to automatically push high conv-layer representations in a CNN to represent an understandable object part.
- The interpretable CNN does not change the loss function on the top layer and uses the same training samples as the original CNN.
- Predictive loss occurs but is often minimal. Predictive improvement due to regularization can be realized.

Interpretable CNNs

- Each filter must encode a distinct object part that is exclusive to a single category
- The filter must be activated by a single part of the object, rather than repetitively appear on different object regions.
  - Assumes that high-level parts don't contain textual details
  - User must find the right layer





Interpretable CNNs

http://openaccess.thecvf.com/content_cvpr_2018/papers/Zhang_Interpretable_Convolutional_Neural_CVPR_2018_paper.pdf
https://github.com/zqs1022/interpretableCNN (MATLAB only as of Jan 2020)

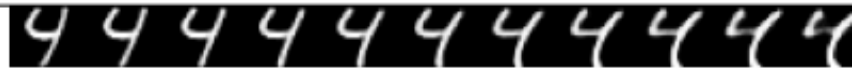
CapsNet

Geoffrey Hinton, Google Brain

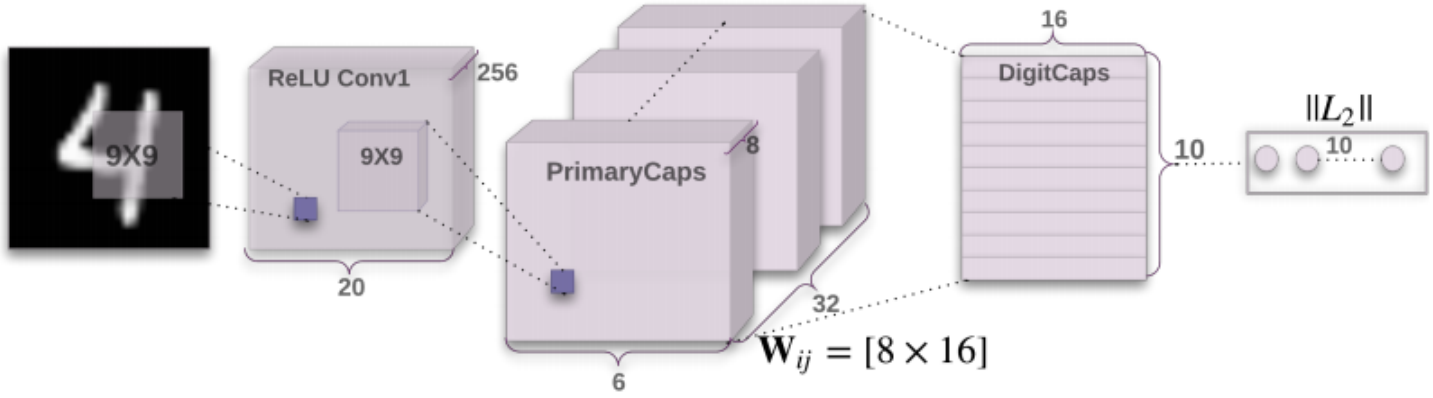
- New research. Only works well on MNIST-like images currently. Shows promise.
- Capsules represent properties of a particular entity in an image such as position, size, orientation, velocity, albedo, hue, texture, etc.
- The model learns to encode these relationship between a part and whole which leads to viewpoint invariant knowledge that generalizes to new viewpoints.
- Can add a reconstruction loss routine so that use capsules can regenerate an image.

CapsNet

Captured Properties

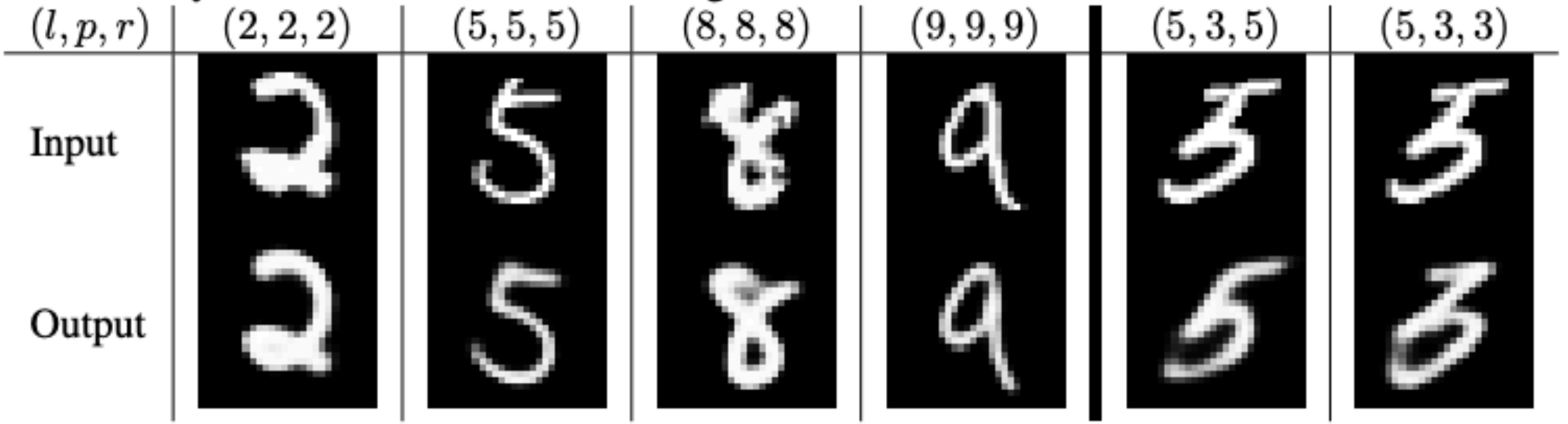
Scale and thickness	
Localized part	
Stroke thickness	
Localized skew	
Width and translation	
Localized part	

Capsule Generation



CapsNet

Figure 3: Sample MNIST test reconstructions of a CapsNet with 3 routing iterations. (l, p, r) represents the label, the prediction and the reconstruction target respectively. The two rightmost columns show two reconstructions of a failure example and it explains how the model confuses a 5 and a 3 in this image. The other columns are from correct classifications and shows that model preserves many of the details while smoothing the noise.

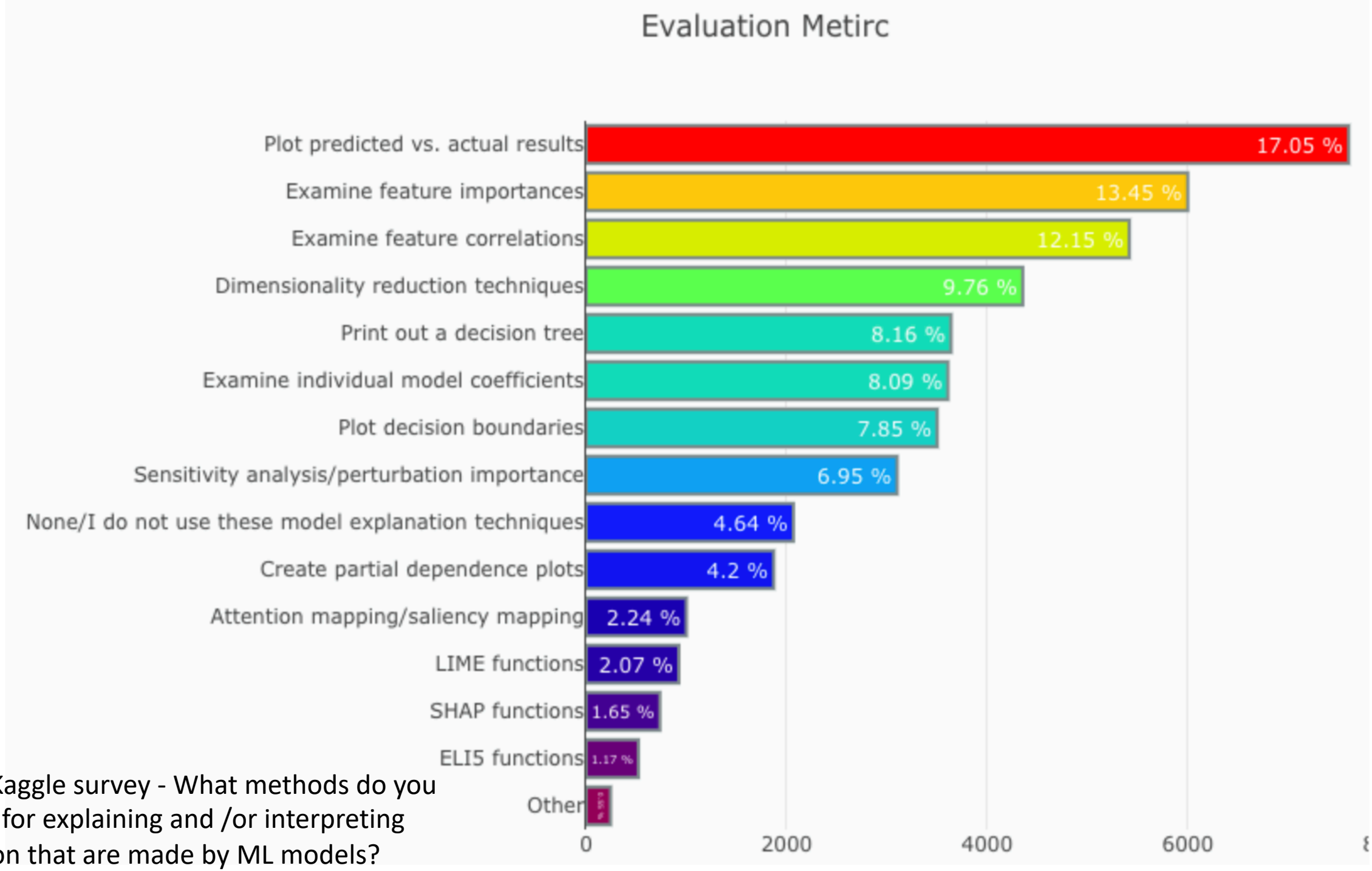


CapsNet

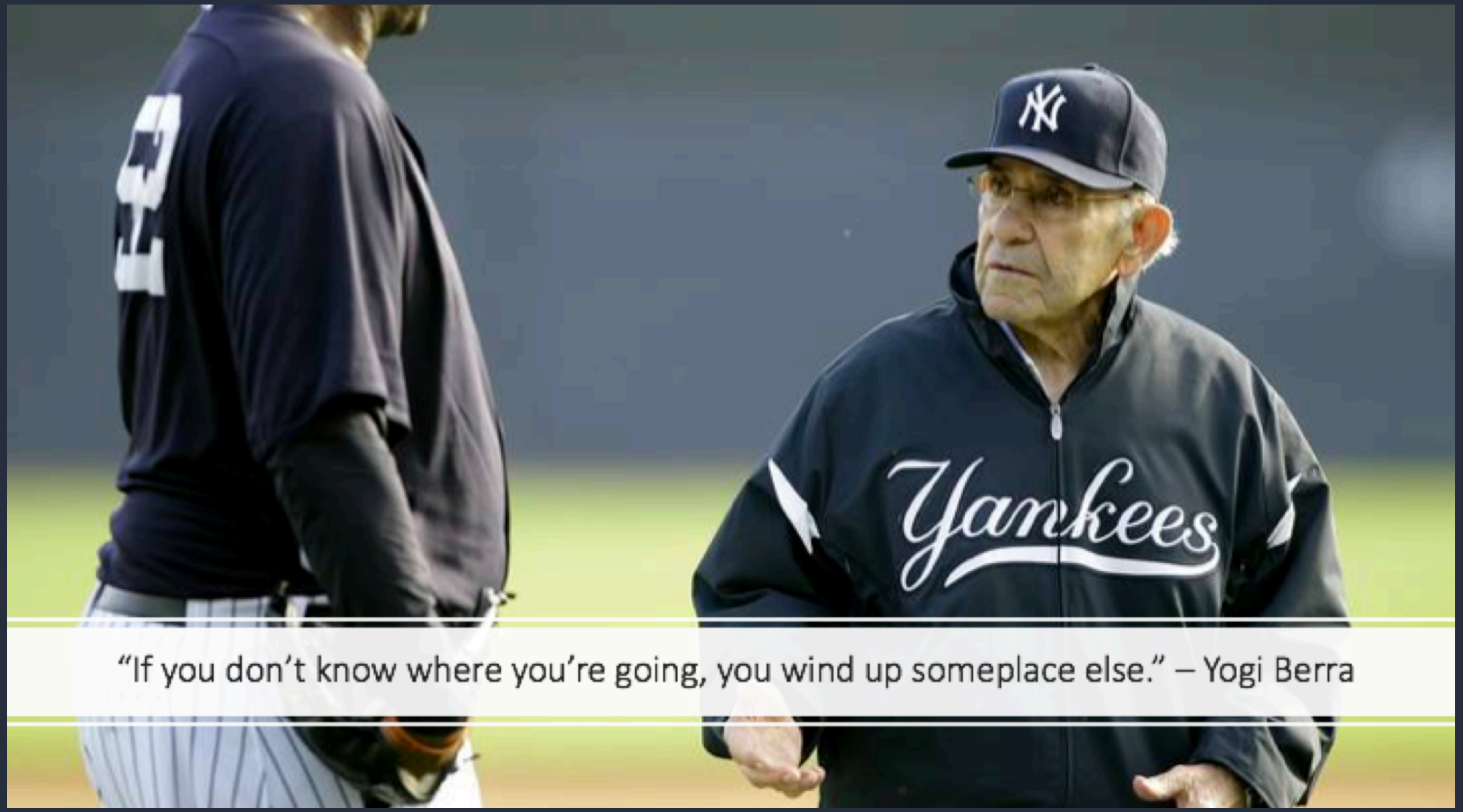
https://papers.nips.cc/paper/6975-dynamic-routing-between-capsules.pdf
https://github.com/loretoparisi/CapsNet (and many others on github making their own implementation of CapsNet)

A Few Examples

- Llyods
  - SHAP, LIME, and sticking to GLMMs
- FICO
  - Recently had an xAI challenge
  - Doing a bunch of traditional approaches
  - Not using LIME but is aware of it
- Bloomberg
  - Sticking to GLMMs
  - Avoiding RNNs for NLP
- Capitol One
  - Documented benefits from DL but not putting much into production due to regulations and potential bias



2018 Kaggle survey - What methods do you prefer for explaining and /or interpreting decision that are made by ML models?



"If you don't know where you're going, you wind up someplace else." – Yogi Berra

Global Interpretability With Predictive Power Is Rare

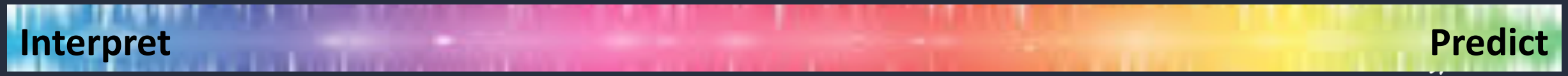
- For DNNs…
- Interpretable CNNs and AILense (Influence Directed Explanations) are promising options
- TCAV and ACE show promise but are still to hard to implement for many projects

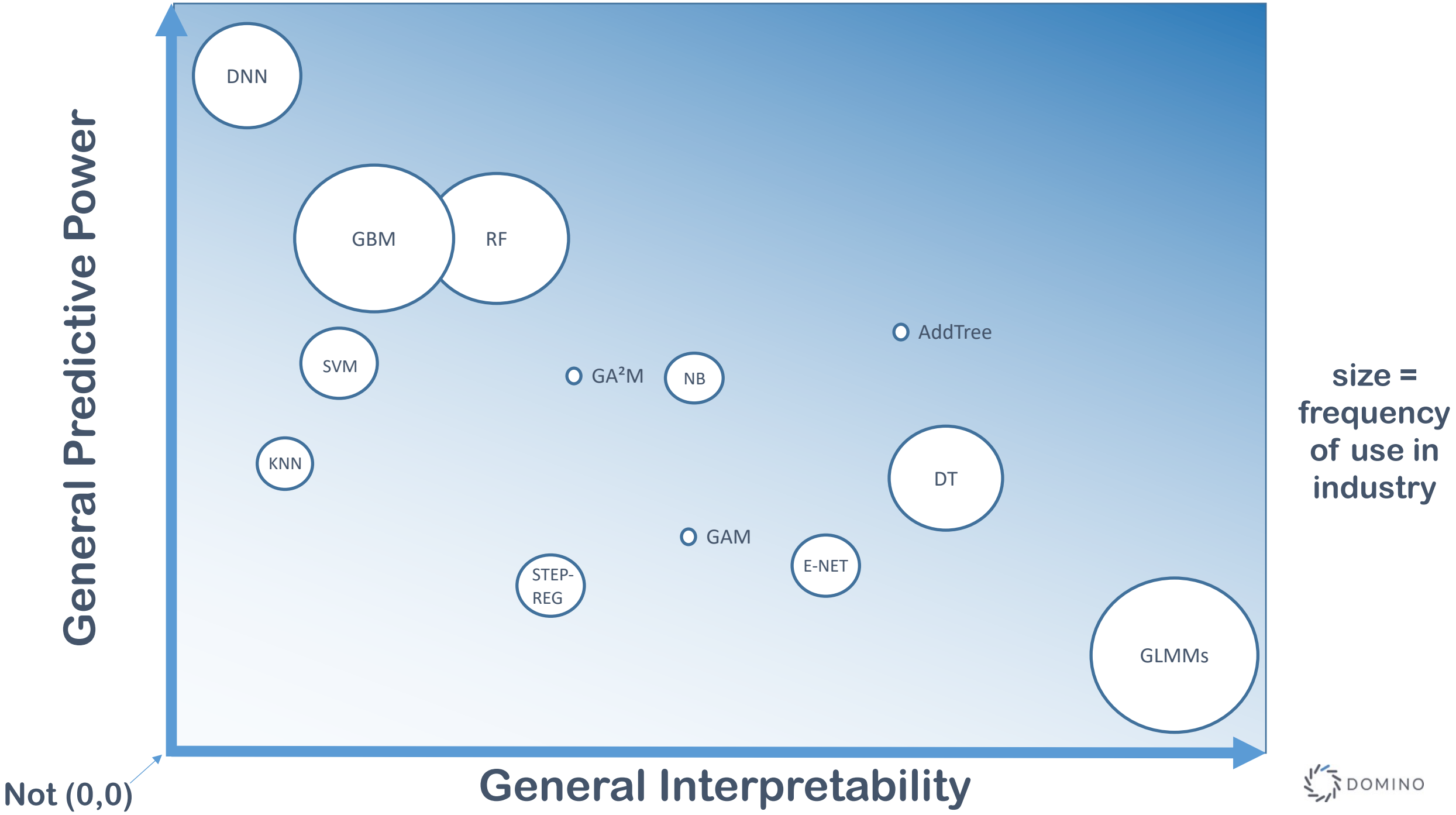
Global Interpretability With Predictive Power Is Rare

- For tradition ML models…
- All interpretability tools leave a lot to be desired. Must sacrifice predictive power.
- Compare approaches
  - DNN, GBM, or RF combined with a SHAP
  - Trees, E-Net, and NB are a good compromise
  - AddTree is promising if your customers are good with trees and/or the tree is not too big
  - GLMM might predict well and will at least give a baseline
- Don’t rely on global inference with ML Models

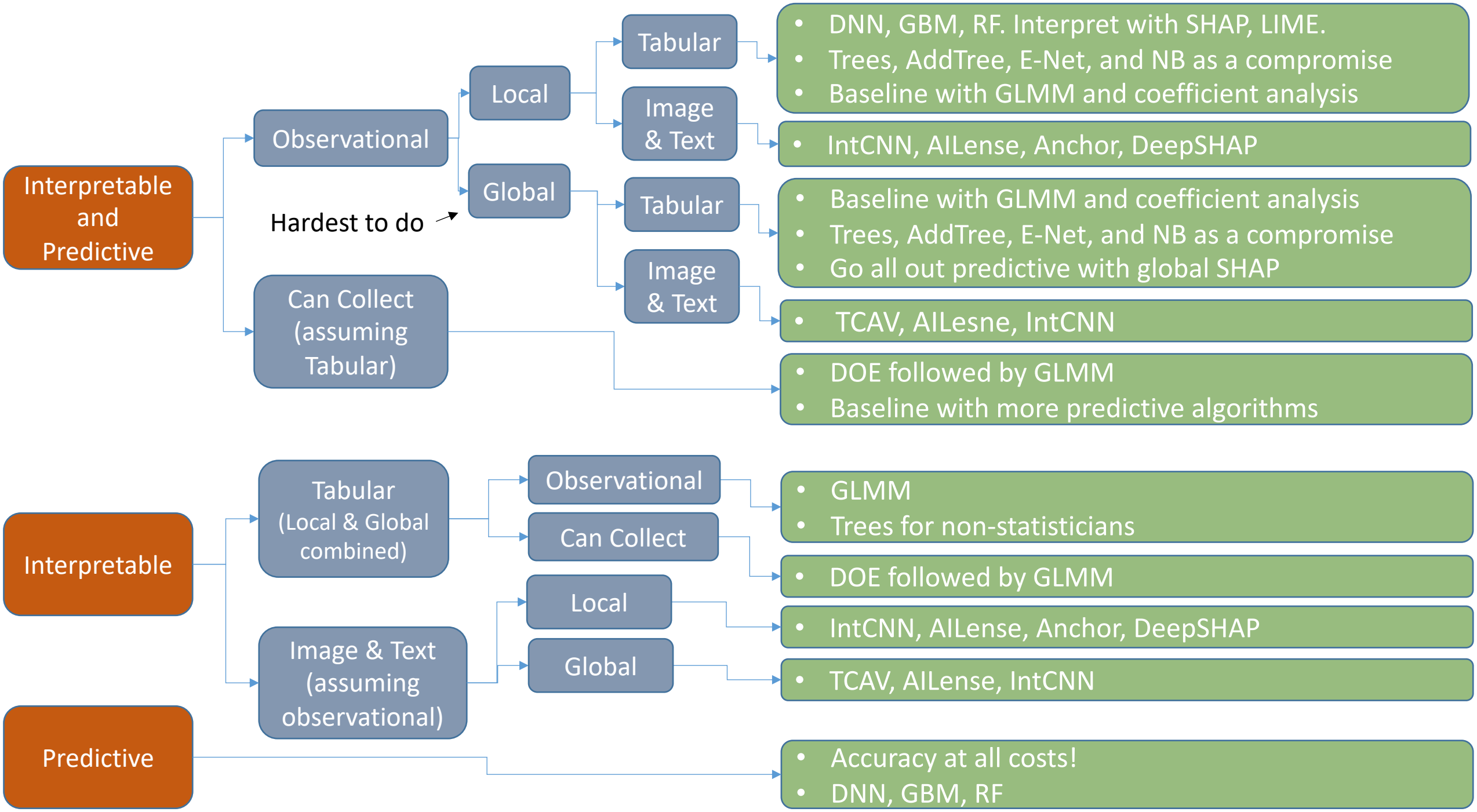
Figure Out Where You Are On The Spectrum

- What is the goal of the project?
- What decisions will be made or influenced as a direct result of the project?
- Who will make those decisions (person, machine)?
- How will they consume the output of the model?
- What kind of model output is needed (score/probability, binary prediction, quantify effect of factors, etc.)?
- How frequently will they have to make the decisions?
- What is the value of a correct decision and the cost of a wrong decision?
- How important is it to explain each individual prediction and why?
- How important is it to explain the way the model works in general (or to know the structure of relationships in the system/data) and why?




General Predictive Power
DNN
GBM
RF
SVM
KNN
GA²M
NB
AddTree
DT
GAM
E-NET
STEP-
REG
GLMMs
size =
frequency
of use in
industry
Not (0,0)
General Interpretability

DOMINO


Interpretable and Predictive
Observational
Local
Tabular
DNN, GBM, RF. Interpret with SHAP, LIME.
Trees, AddTree, E-Net, and NB as a compromise
Baseline with GLMM and coefficient analysis
Image & Text
IntCNN, AILense, Anchor, DeepSHAP
Global
Tabular
Baseline with GLMM and coefficient analysis
Trees, AddTree, E-Net, and NB as a compromise
Go all out predictive with global SHAP
Hardest to do
Image & Text
TCAV, AILesne, IntCNN
Can Collect (assuming Tabular)
DOE followed by GLMM
Baseline with more predictive algorithms
Interpretable
Tabular (Local & Global combined)
Observational
GLMM
Trees for non-statisticians
Can Collect
DOE followed by GLMM
Image & Text (assuming observational)
Local
IntCNN, AILense, Anchor, DeepSHAP
Global
TCAV, AILense, IntCNN
Predictive
Accuracy at all costs!
DNN, GBM, RF