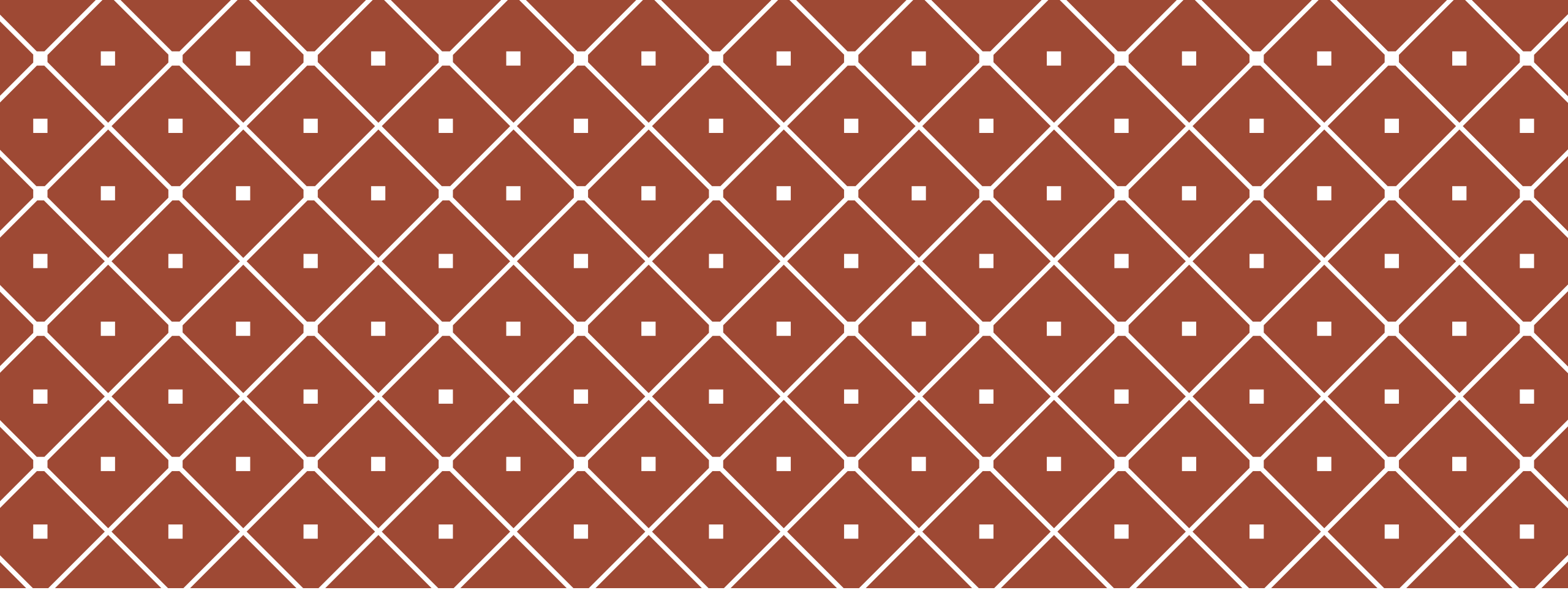


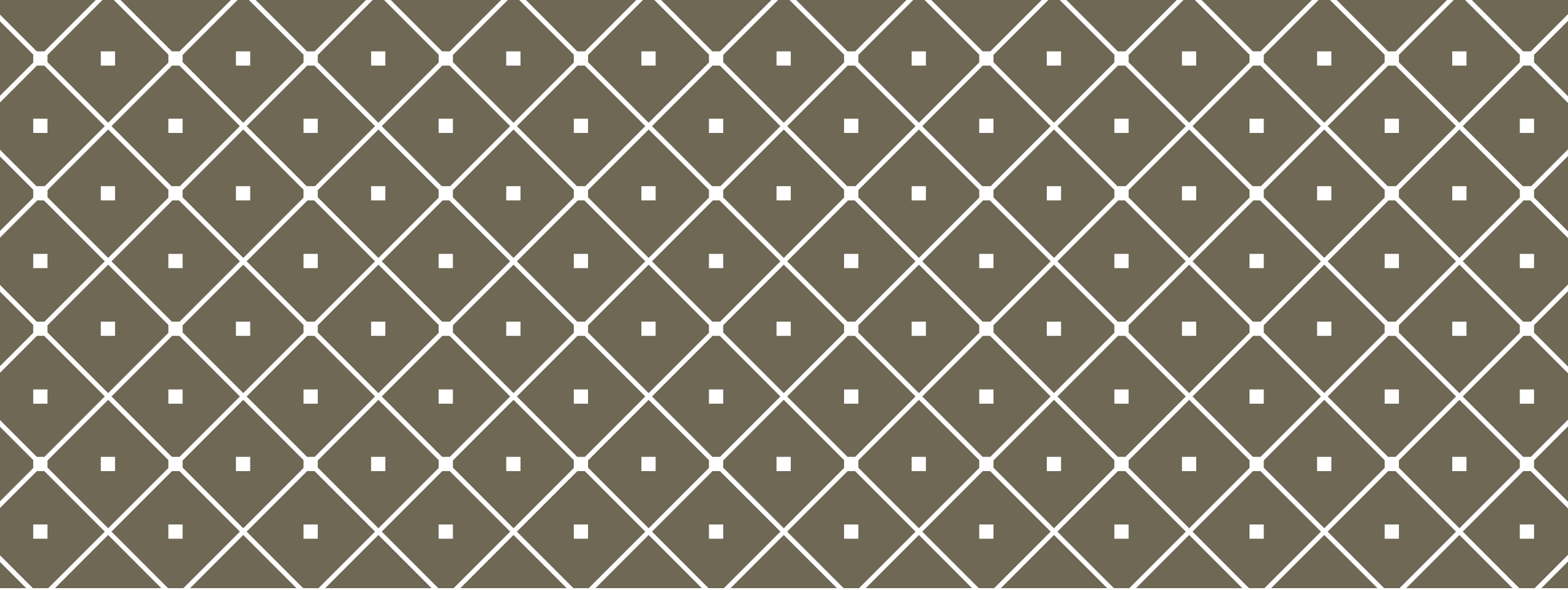
MODEL INTERPRETABILITY IN MACHINE LEARNING

Michaël Mariën, KBC Group

BEFORE WE START

1. Focus on supervised models but not on modelling
2. Focus solely on model: it is our new ground-truth
3. Focus on specific model instance: not ‘understanding architecture’
4. No focus on what “(good) explanation” is, how to measure…
5. Focus on model-agnostic methods

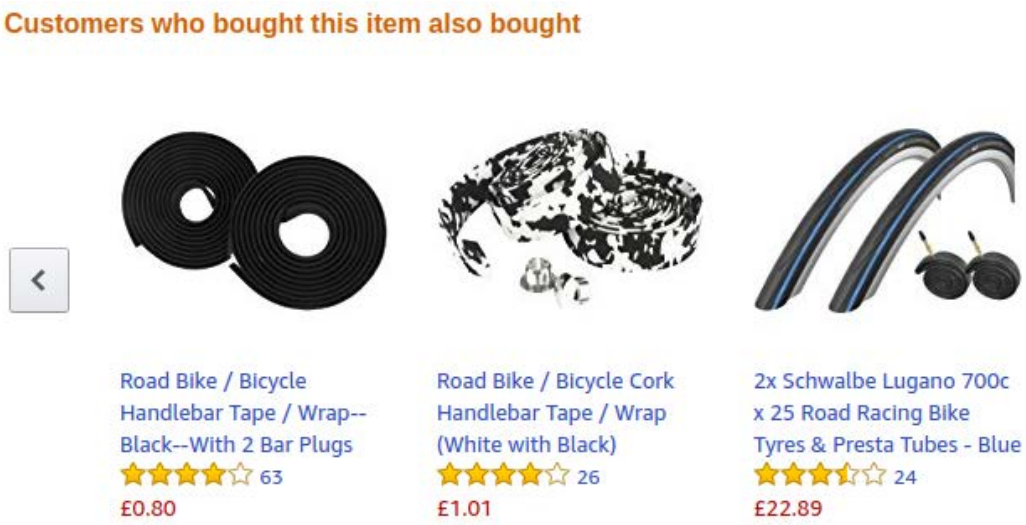
6. I’m an enthusiastic user, not a research expert



WHY?

The need for model interpretability

BUSINESS NEEDS



BUSINESS NEEDS

Marketing:

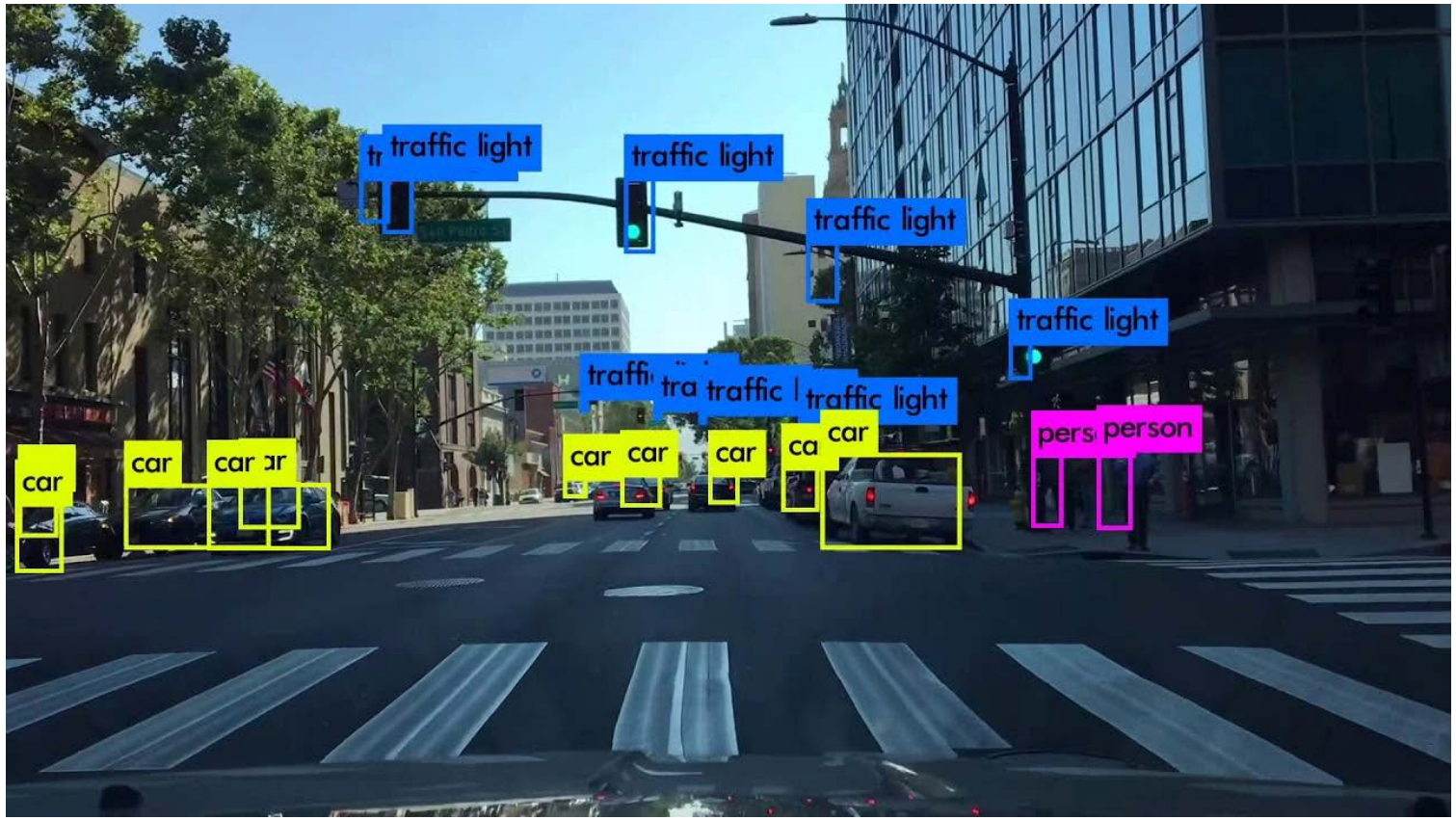
- Trust of employees to increase buy-in
- Content for employees to increase succes rate

Fraud: where to look?

Churn: how to act?

Fund managers

SAFETY


traffic light
traffic light
traffic light
traffic light
traffic light
car
car
car
car
car
car
car
person

REGULATION


General
Data
Protection
Regulation

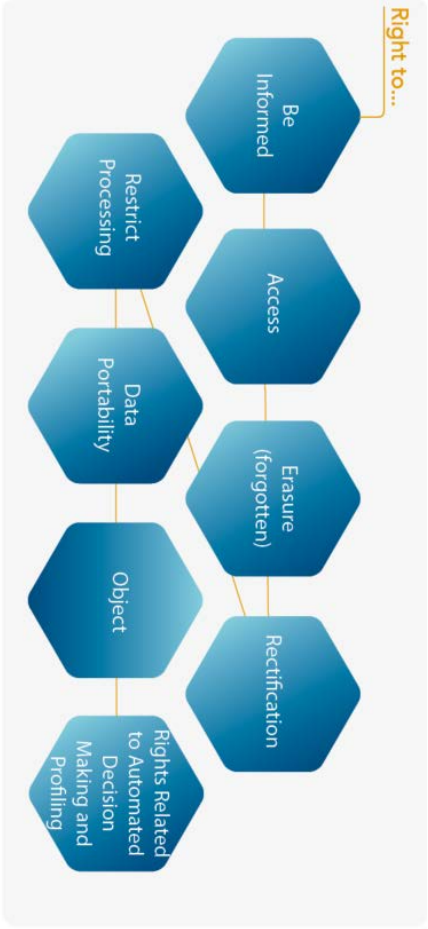

AUDIT
AUDIT
AUDIT


Discri nation

REGULATION: GDPR

The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, **access** to the personal data and the following information:

> *The existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, **meaningful information about the logic involved**, as well as the significance and the envisaged consequences of such processing for the data subject.*

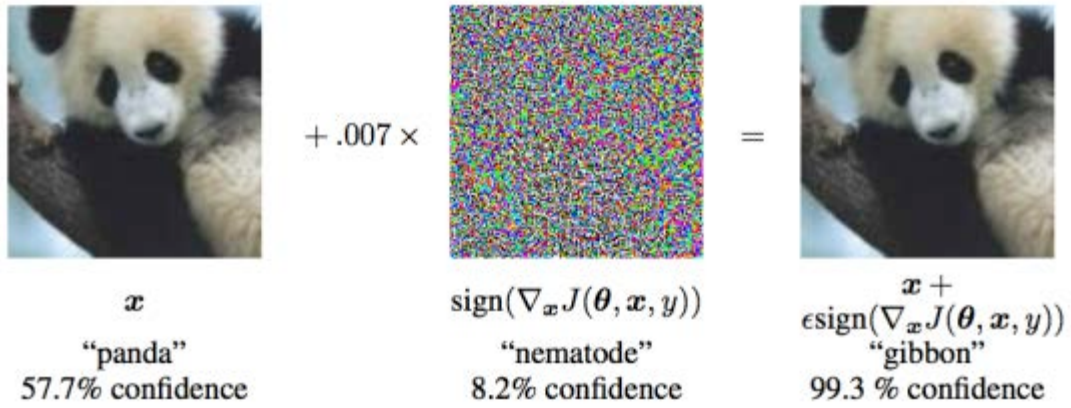


Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.

In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, **to obtain an explanation of the decision reached after such assessment** and to challenge the decision.

SCIENCE


RESEARCH
KNOWLEDGE
EXPERIMENTATION
PRACTICE
EXAMINE
PROBE
PROJECT
SCAN
SCIENCE
BOOK
FACTS
LINK
THINK
TEACHING
LEARNING
METHODOLOGY
SEARCH
CHECK
STOCK


+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
sign($\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)$)
"nematode"
8.2% confidence
$\boldsymbol{x} +$
ϵsign($\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)$)
"gibbon"
99.3 % confidence

SOCIAL ACCEPTANCE AND TRUST

POLITICS

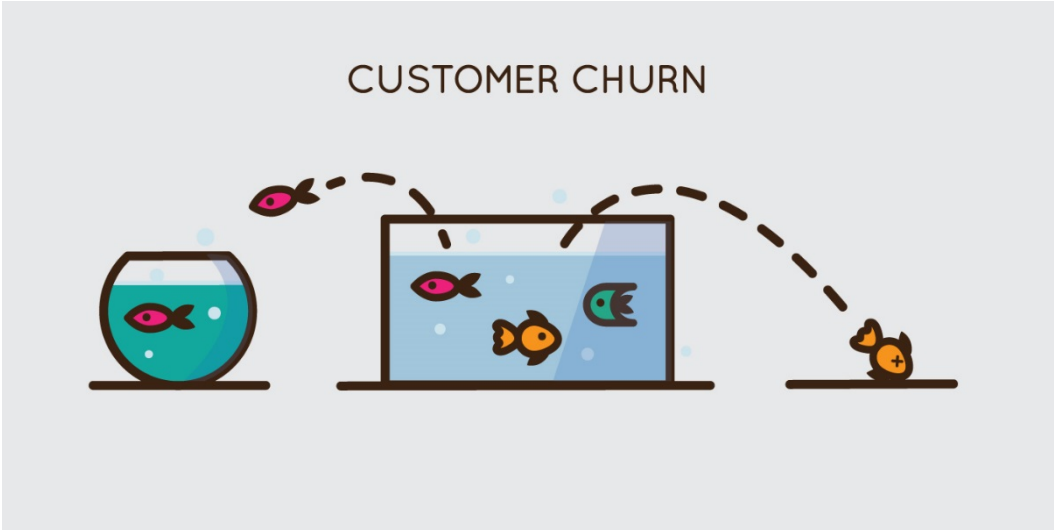
Sent to Prison by a Software Program's Secret Algorithms

Sidebar

By ADAM LIPTAK MAY 1, 2017

UNDERSPECIFICATION OF PROBLEM

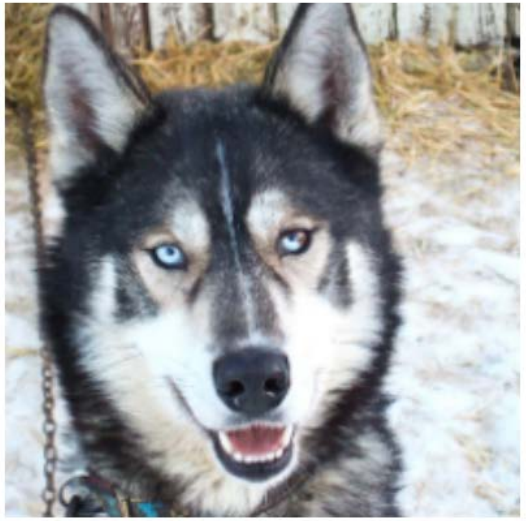
Even if you trust your AUC…



DEBUGGING / BIAS / ADVERSERIAL

CAN YOU BUILD YOUR TRUST BASED ON ACCURACY?

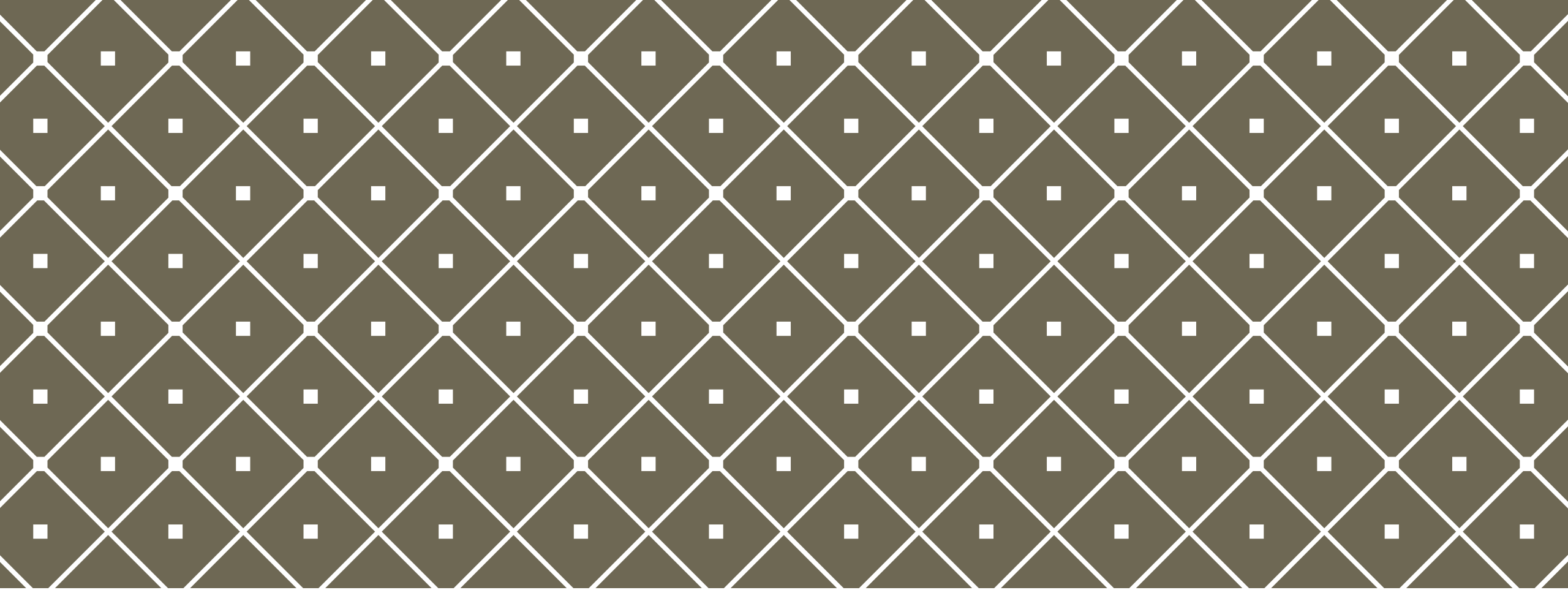
Only 1 mistake!



(a) Husky classified as wolf

(b) Explanation

Ribeiro et al: "Why Should I Trust You?" Explaining the Predictions of Any Classifier



WHAT?

Measuring and defining interpretability

WHAT IS 'INTERPRETABLE'

How to measure interpretability?

Interpretability is the degree to which a human can understand the cause of a decision. The higher the interpretability of a model, the easier it is for someone to comprehend why certain decisions (read: predictions) were made. A model has better interpretability than another model, if its decisions are easier to comprehend for a human than decisions from the second model. I will be using both the terms interpretable and explainable equally. (Tim Milner)

Can be on different levels (modular: betas, splits, convnet filters)

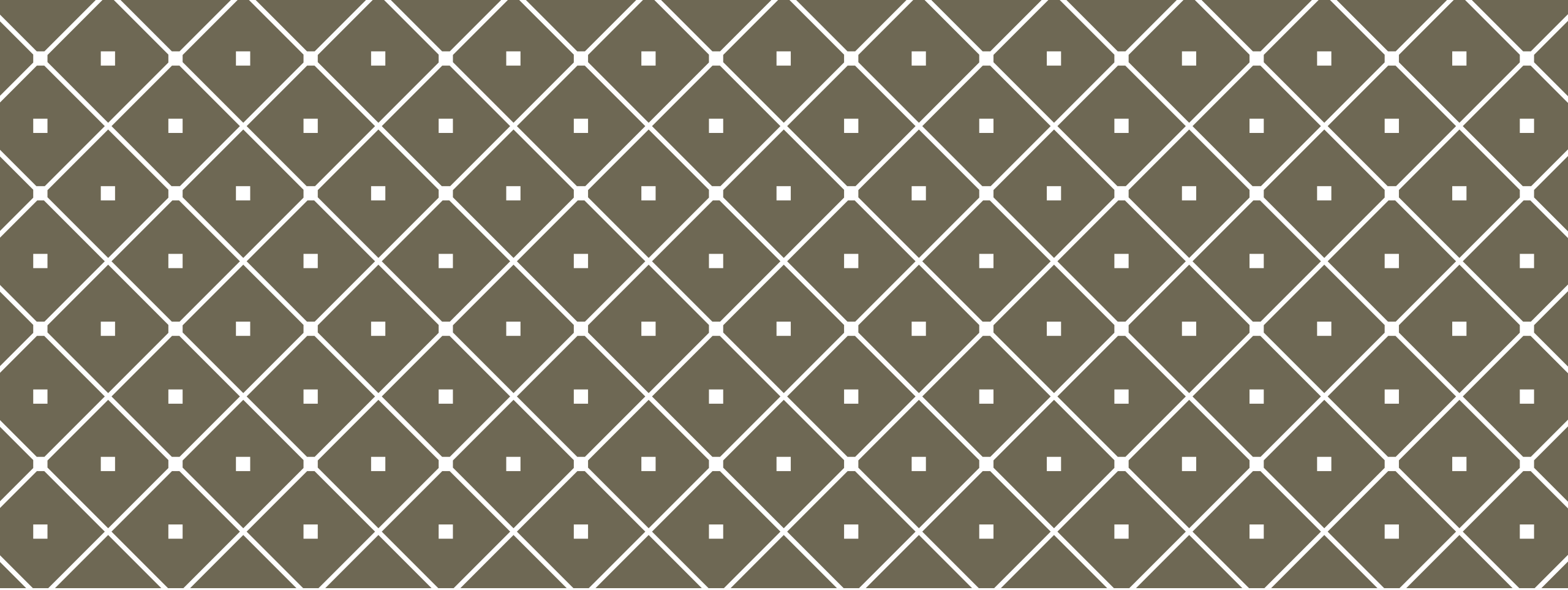
What is interpretable: depth, sparseness, complicated features, monotone?

AIM OF INTERPRETATION

Goal-based:
Does providing interpretability assist with down-stream tasks: fair, safety, productivity, progress of science,…

Metric-based:
is sparsity 5 a lot better than 10? Hard to judge pure on numbers without taking the task at hand into consideration

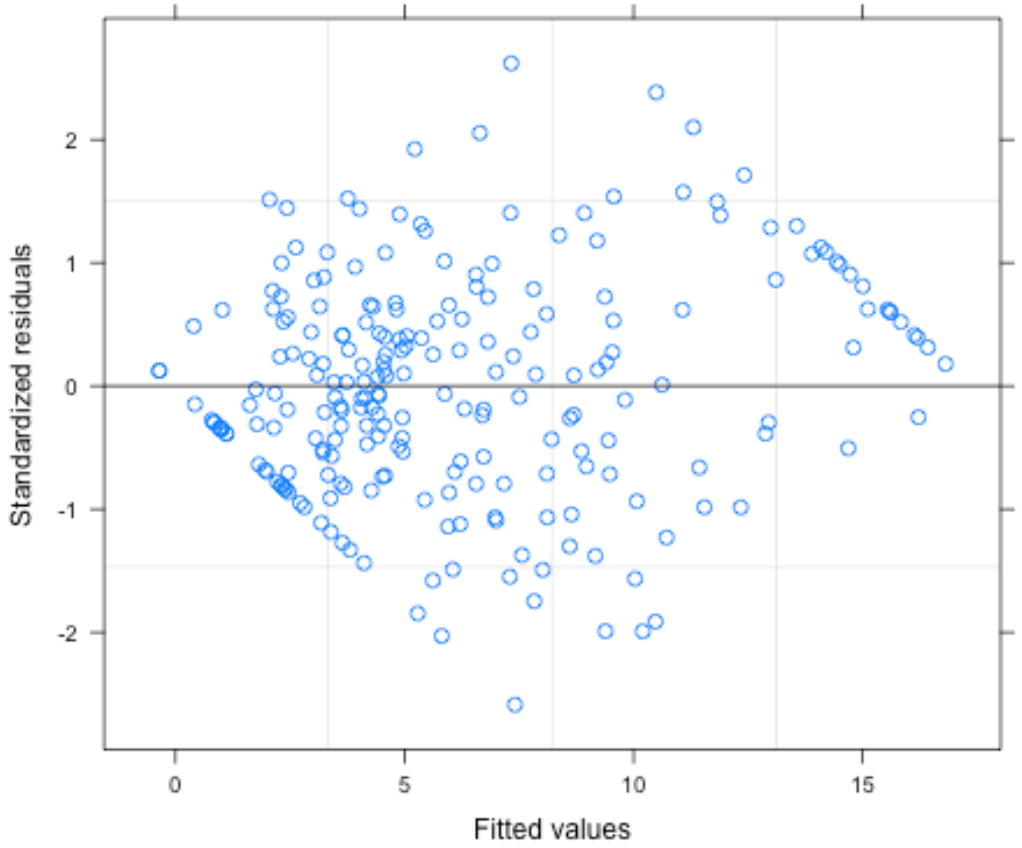
Human-based:
Humans like counterfactuals and

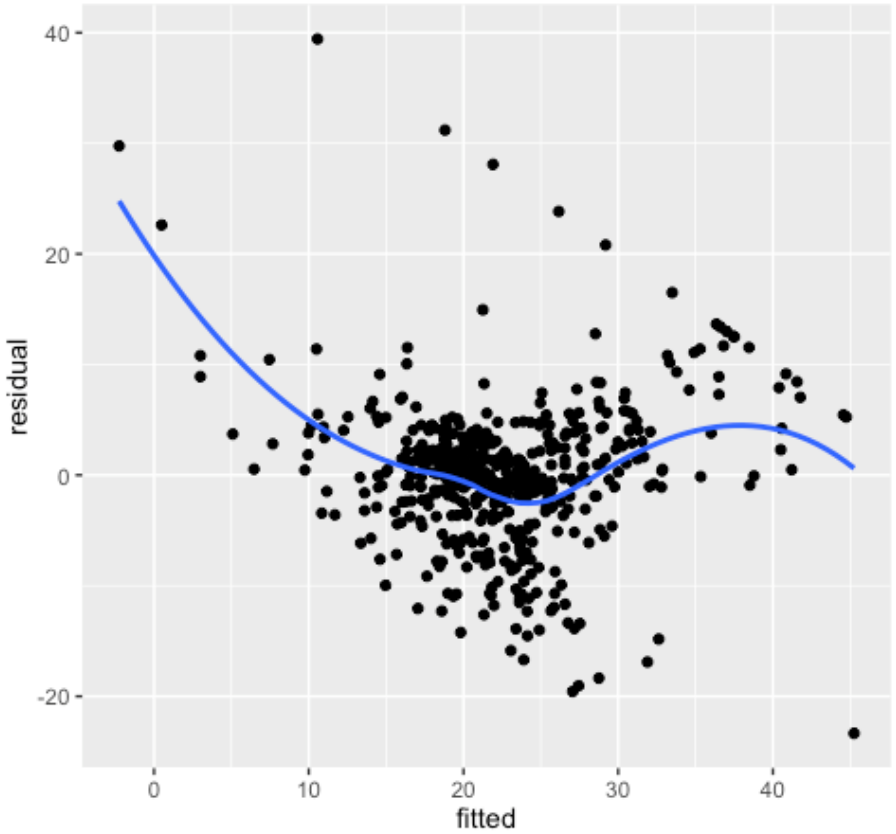


DATA

Take a look

ANALYSIS OF RESIDUALS





DIMENSIONAL REDUCTION TECHNIQUES

A variety of dimensional reduction techniques can be used to get insight into your data and visualize.

Can they help with model interpretability?

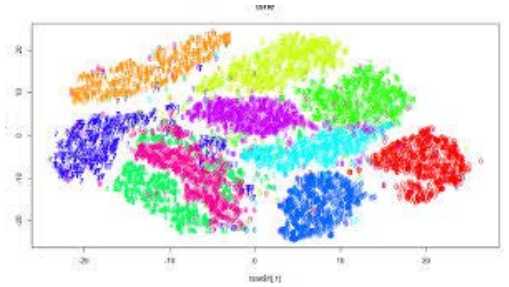
Mainly moving the complexity from the model to the features

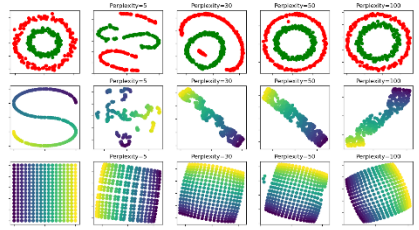
Can enhance trust if they recover known patterns

- PCA, SVD

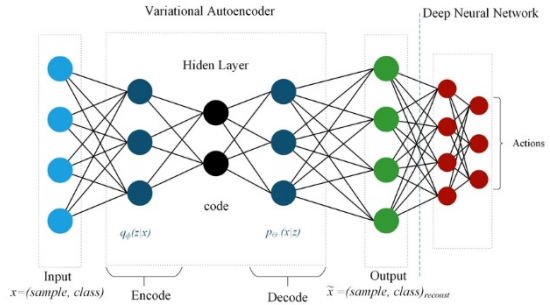
$$\underset{(I \times J)}{\mathbf{X}} = \underset{(I \times R)}{\mathbf{T}} \underset{(R \times J)}{\mathbf{P}^T} + \underset{(I \times J)}{\mathbf{E}}$$

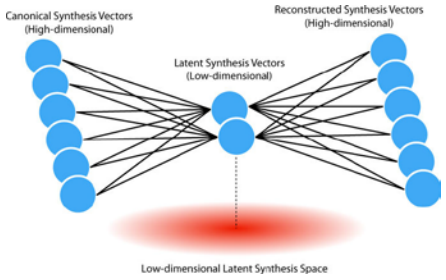
- T-SNE

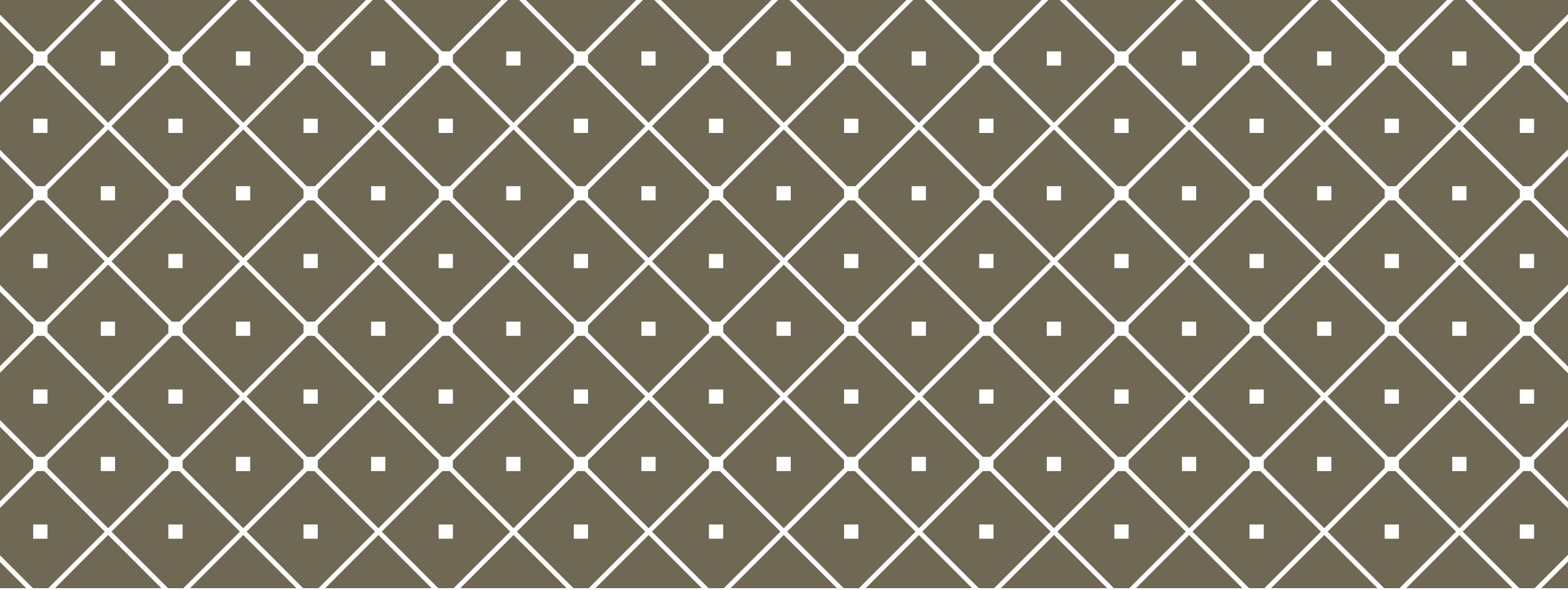




- Variational autoencoder







INTERPRETABLE MODELS

The search for white-box models

LINEAR REGRESSION

Linear models are usually considered the gold standard in interpretability (together with decision trees)

- They can provide counterfactuals only to a reference with zero features
  This can be solved with centering: the coefficients x features give the contributions to the prediction compared to reference
- Depending on the features, are they interpretable? Especially after introducing interactions, polynomials
- Some additional sparisty is required
- Is the model any good?

RULE-BASED MODELS

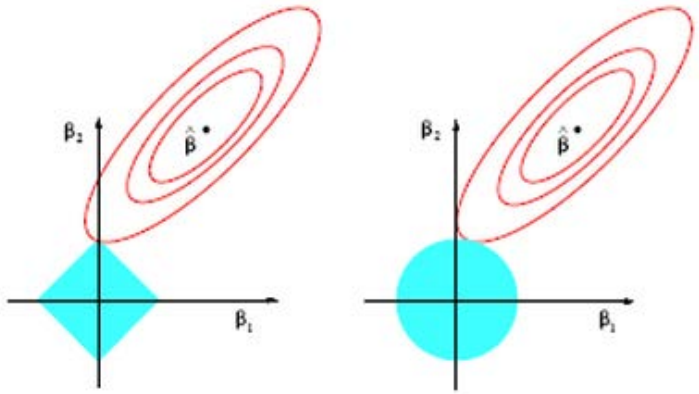
Humans really like rules, either in trees or in other approaches based on logical predicates

- To what extend is a tree interpretable? How deep can it be?
- The same for boolean, how many clauses?
- What about sparsity, can we interpret many booleans at the same time? Sometimes a local perspective fixes this.

IMPROVING LINEAR MODELS

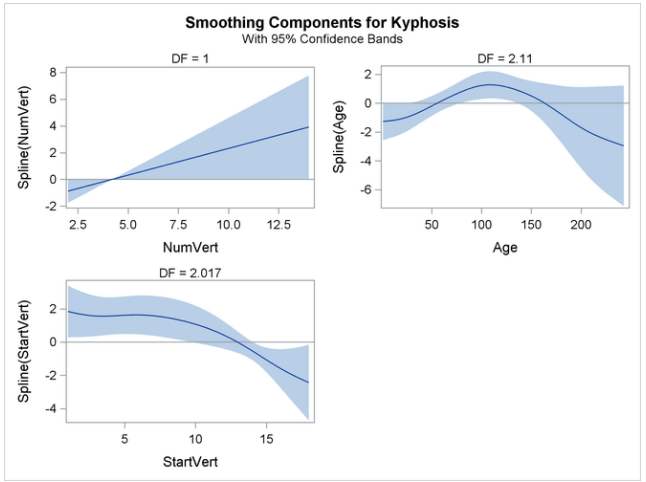
Especially in regulated industries, most machine learning techniques are still a no-go.
But progress does not halt! With small steps, such domains are catching up.
By using newer, better models, they are also changing what "interpretability" is.

Penalized Regression



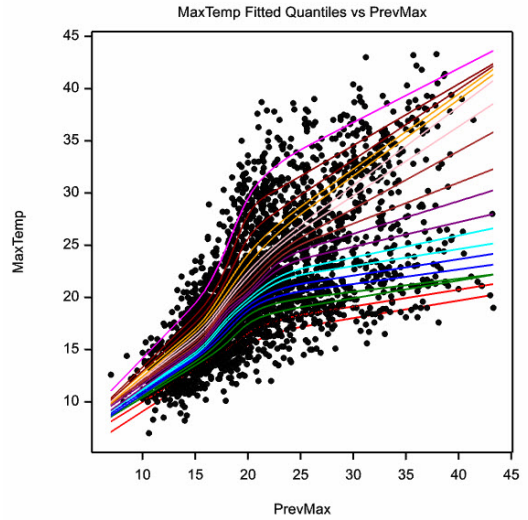
No real p-values, statistics

Generalized Additive Models



Splines, poly, logs and all that

Quantile Regression

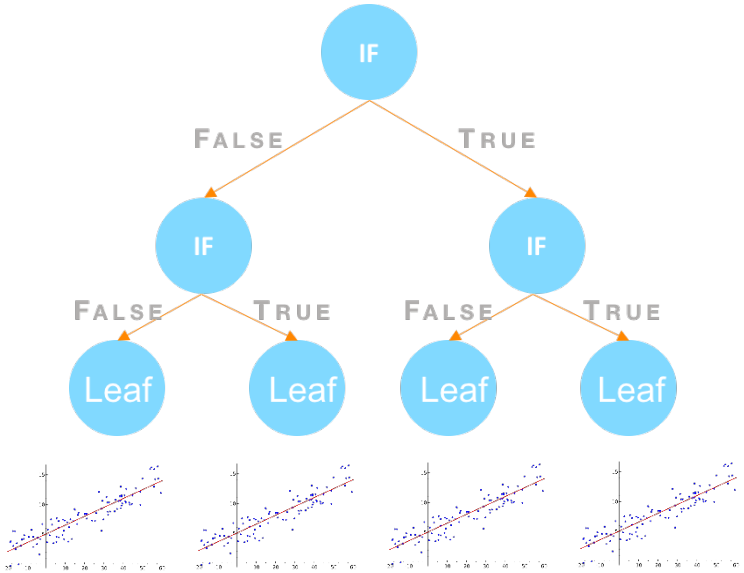


Different loss functions

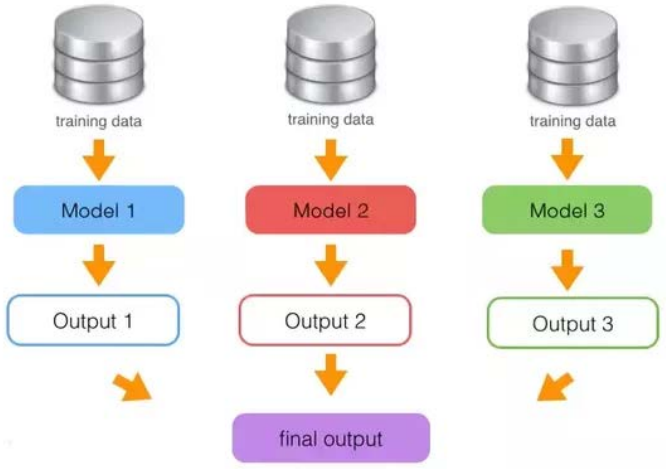
EMBRYO STATE OF THE ART

Try to use simpler versions of complex models, perhaps shift complexity to features.
Or use their insights, features, relations as input for simpler models, thus maintaining the required well-behaved properties. Distillation is another promising method.

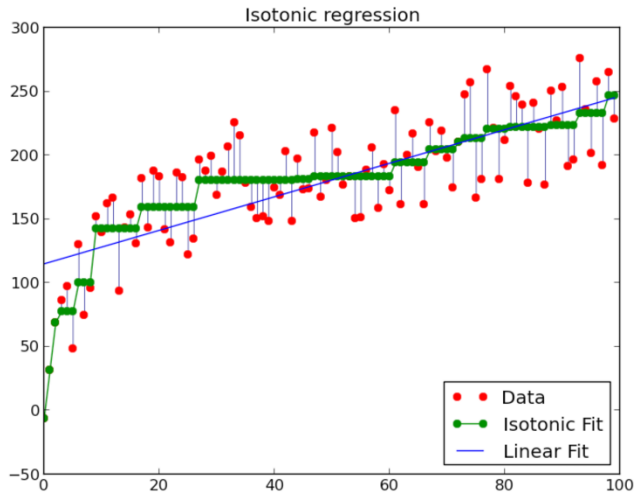
Segment + model



Stacking ensembles



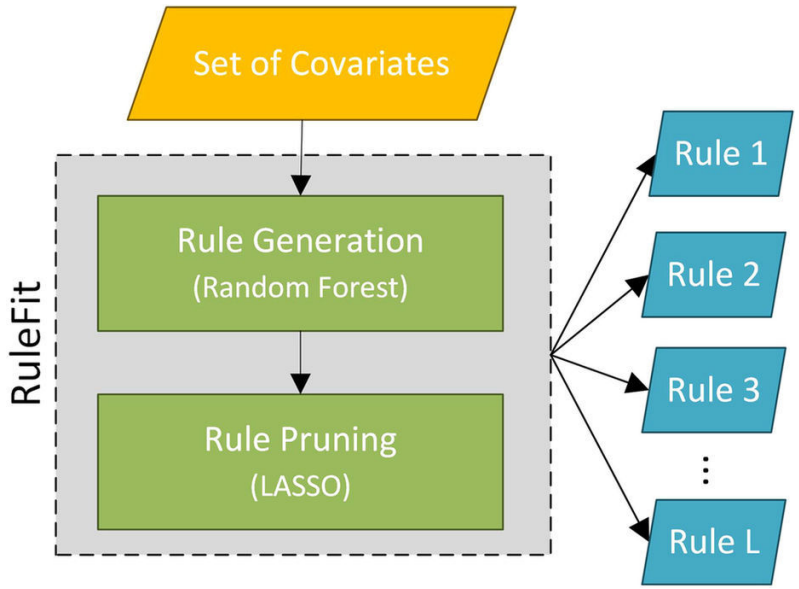
Monotonicity constraints



KIND OF STATE OF THE ART: NEW METHODS

New methods, designed to be more accurate than traditional models, but interpretable.

RuleFit



Decision tree + Lasso

Antecedent Mining

Consider *minSup*= 0.5 and *minConf*= 0.5:

ID	Sequences
seq1	{a, b},{c},{f},{g},{e}
seq2	{a, d},{c},{b},{a, b, e, f}
seq3	{a},{b},{f},{e}
seq4	{b},{f, g}

A sequence database

→

ID	Rule	Support	Confidence
r1	{a, b, c}⇒{e}	0.5	1.0
r2	{a}→{c, e, f}	0.5	0.66
r3	{a, b} →{e, f}	0.5	1.0
r4	{b} →{e, f}	0.75	0.75
r5	{a} →{e, f}	0.75	1.0
r6	{c} →{f}	0.5	1.0
r7	{a}→{b}	0.5	0.66
…	…	…	…

Some rules found

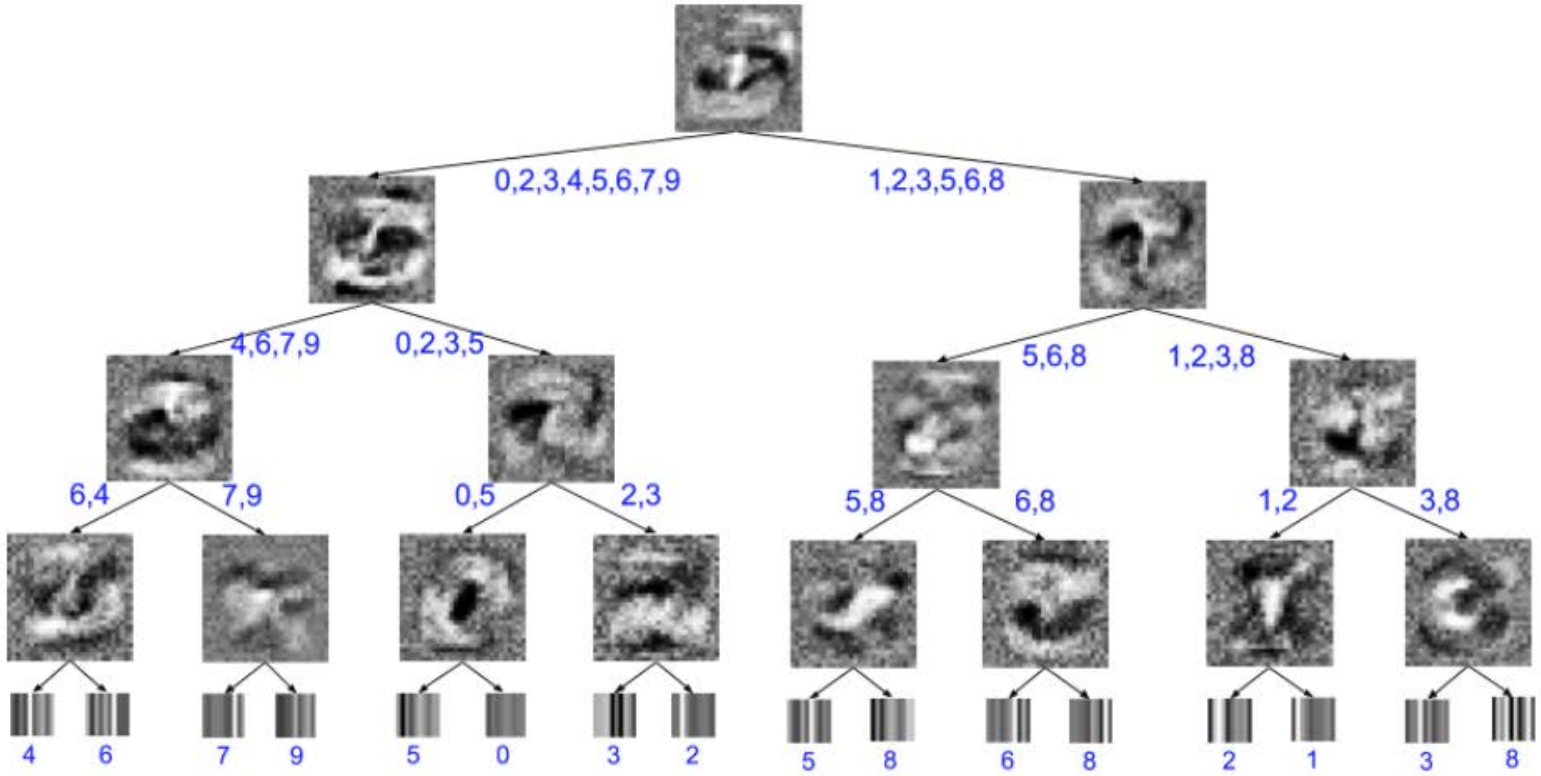
FP growth, Eclat, Claudien, Bayesian Rule List

Friedman et al: Ribeiro et al: Predictive Learning via Rule Ensembles

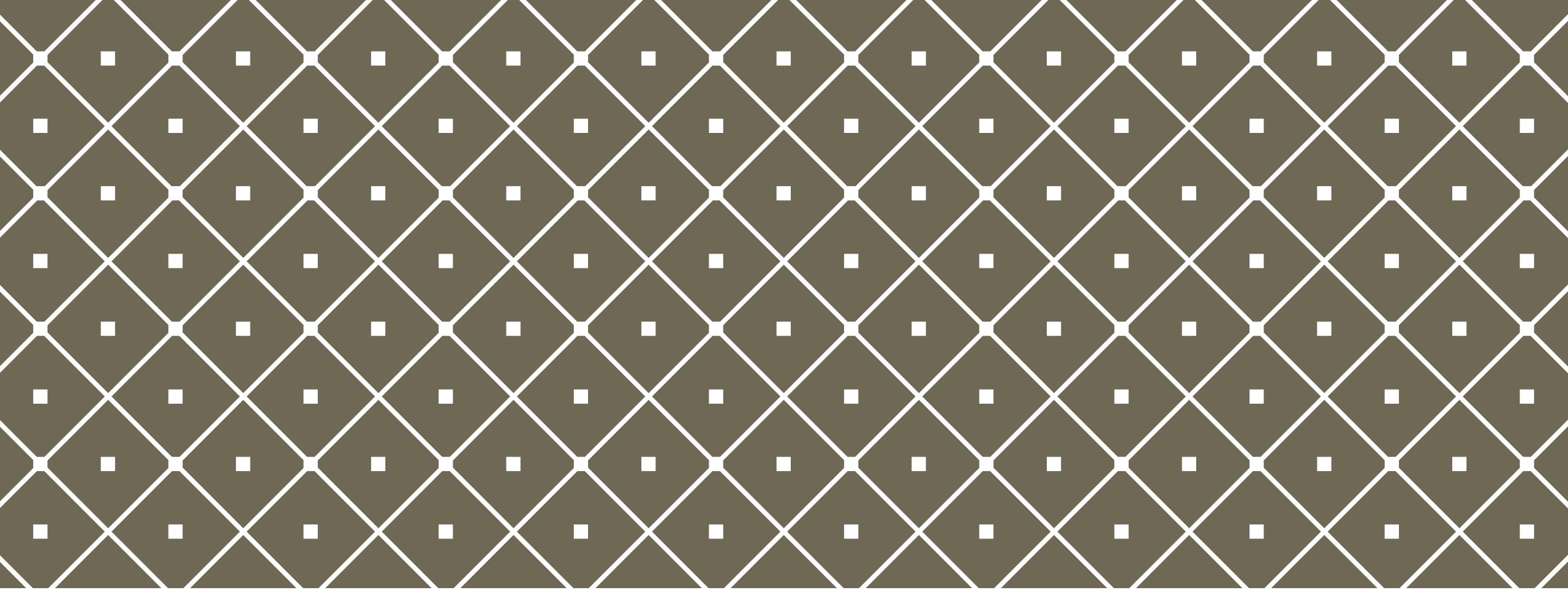
KIND OF STATE OF THE ART: NEW METHODS

New methods, designed to be more accurate than traditional models, but interpretable.

Distilling neural network in soft decision tree



Frosst et al: Distilling a Neural Network Into a Soft Decision Tree



INTERPRETING COMPLEX MODELS

Dealing with black box models

INTERPRETING COMPLEX MODELS

The object of study is the model / response function and it's complexity

In the previous chapter, we looked at linear or monotonic function

For general models: global feature importance methods for selection

INTERPRETATION METHOD REQUIREMENTS

Model agnostic: a method that can be applied to as large a class of models as possible is preferable. Note that often model specific method can be turned into model agnostic ones via a surrogate model (but this can be suboptimal)

Explanation flexibility: there are different forms of explanation. In some cases it might be useful to have a linear coefficients, in other cases a decision tree, reason codes, or visualization of feature importance

Representation flexibility: The method needs to be applied to a interpretable features, not necessarily the same as those of the model. Surrogate models can port specific to agnostic (perhaps exclude bias of method for certain models)

APPROXIMATE VS EXACT

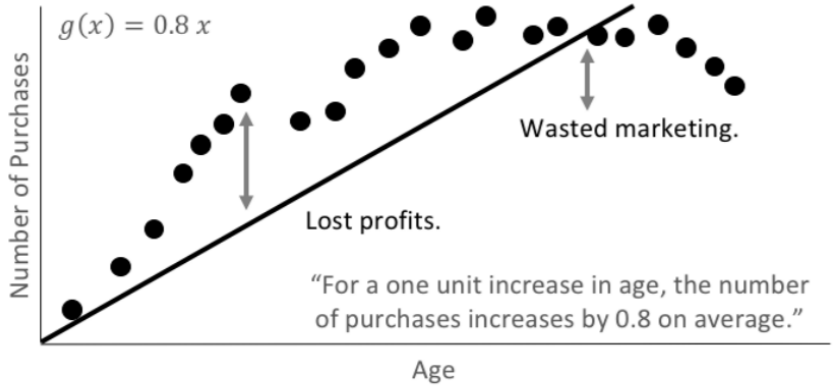


Figure 1: An illustration of *approximate* model with *exact* explanations.

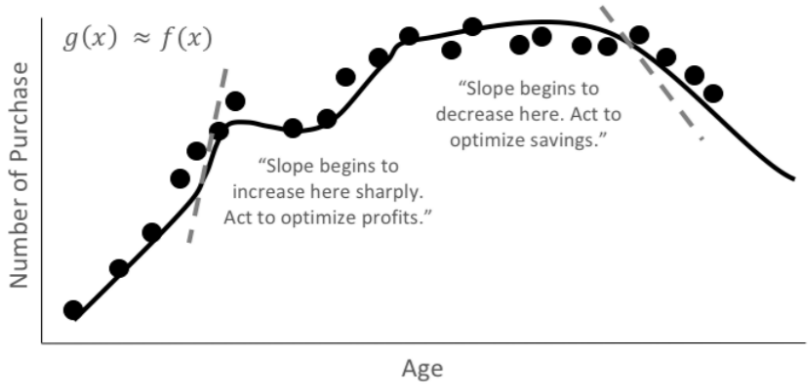


Figure 2: An illustration of an *exact* model with *approximate* explanations. Here $f(x)$ represents the true, unknown target function, which is approximated by training a machine learning algorithm on the pictured data points.

Hall et al: Machine Learning Interpretability with H20 Driverless AI

GLOBAL OR LOCAL

Global: The more classic approach.

Approximately give insight in the total mapping from input to output (modelled conditional distribution)

Often based on averages over all data points.

Attempt to port the safe world of interpretable models as a whole to black box models

Local: Newer approach, often sensitivity based

Understand the mapping locally, on small patches of input (or even output)

Idea is that well-chosen, small local sections are more well-behaved

Can usually be made global

Attempt to port the safe world of interpretable models to local parts of black box models

SURROGATE MODELS

Global technique, model compression, distillation or retrain of an interpretable model (typically on outcomes of complex model) , typically regression or decision tree.

- Very often used: random forest to single decision tree
- Fast debugging, confirmation of domain knowledge from interactions, features
- Some measure of similarity between models
- Is it possible? We choose complex model for a reason.
- How important is that difference in performance?

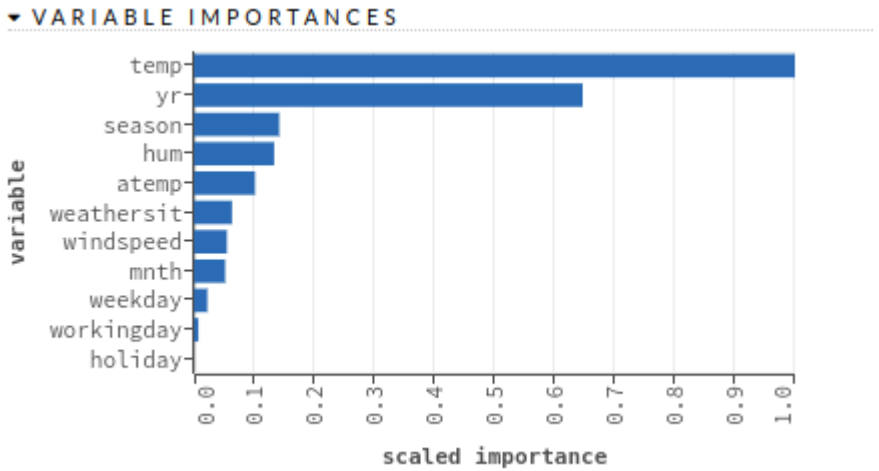
GLOBAL VARIABLE IMPORTANCES

Global technique to determine which variables are important. A column is completely permuted and the change in error (or outcomes) determines how important this column is.

Some model-specific methods exists (greedy methods for trees, betas for regression)

- Extremely compressed view, can be judged in seconds

- Need access to labels in test set
- The model agnostic method is unsigned



Fisher et al: Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the 'Rashomon' Perspective.
Breiman: Random Forests

PARTIAL DEPENDENCE PLOT

The goal of the partial dependence plot is to focus on one (or a few) feature x_{pdp} and study its marginal effect on the model f's predictions.

Partial dependence is a function of x_{pdp}:

$$f_{x_{pdp}}(x_{pdp}) = E_{x_{pdp}}[f(x_{pdp}, x_{\setminus pdp})] = \int f(x_{pdp}, x_{\setminus pdp}) \mathrm{dP}(x_{\setminus pdp})$$

In practice, we can use a samping approach for huge data sets:

$$f_{x_{pdp}}(x_{pdp}) \approx \frac{1}{n} \sum_{i=1}^{n} f(x_{pdp}, x_{\setminus pdp})$$

Friedman: Greedy Function Approximation: A Gradient Boosting Machine

PARTIAL DEPENDENCE PLOT

Global technique, model-agnostic to study a single feature by averaging [2]out the others.

- Intuitive for layman, causal, easy to explain: give all data the same value, visual
- Capture non-monotonic behaviour easily
- Sensitivity, stability testing easily possible
- Varying PD plot should correspond to important features
- Only 1 or 2 features in same plot
- **If features are correlated we average over very unlikely data points** (local averaging)
- **Averaging can hide very different individual reponses:** strong interactions

INDIVIDUAL CONDITONAL EXPECTATION PLOTS

The goal of ICE is to see for each row how the prediction changes when we change one fixed feature (and keep all others fixed). Non-linear sensitivity analysis for one feature

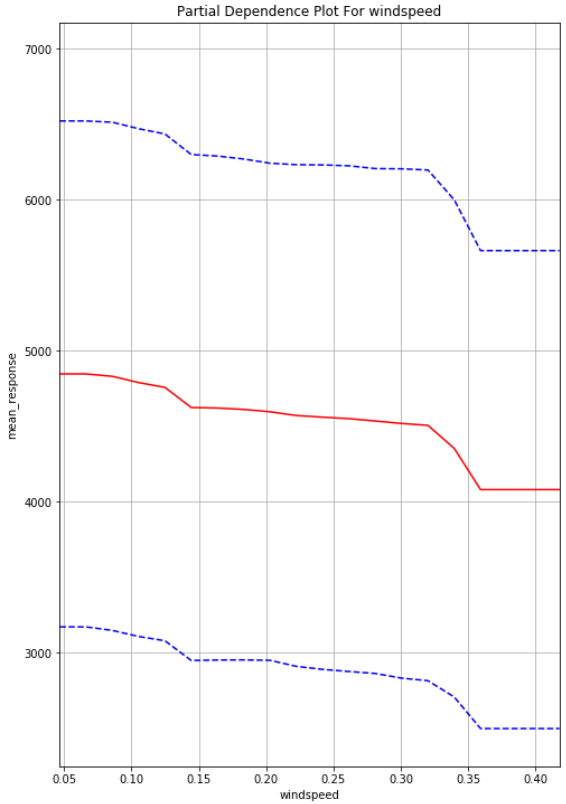
Un-averaged version of PDP, can capture interactions between features and unmask heterogenous effects

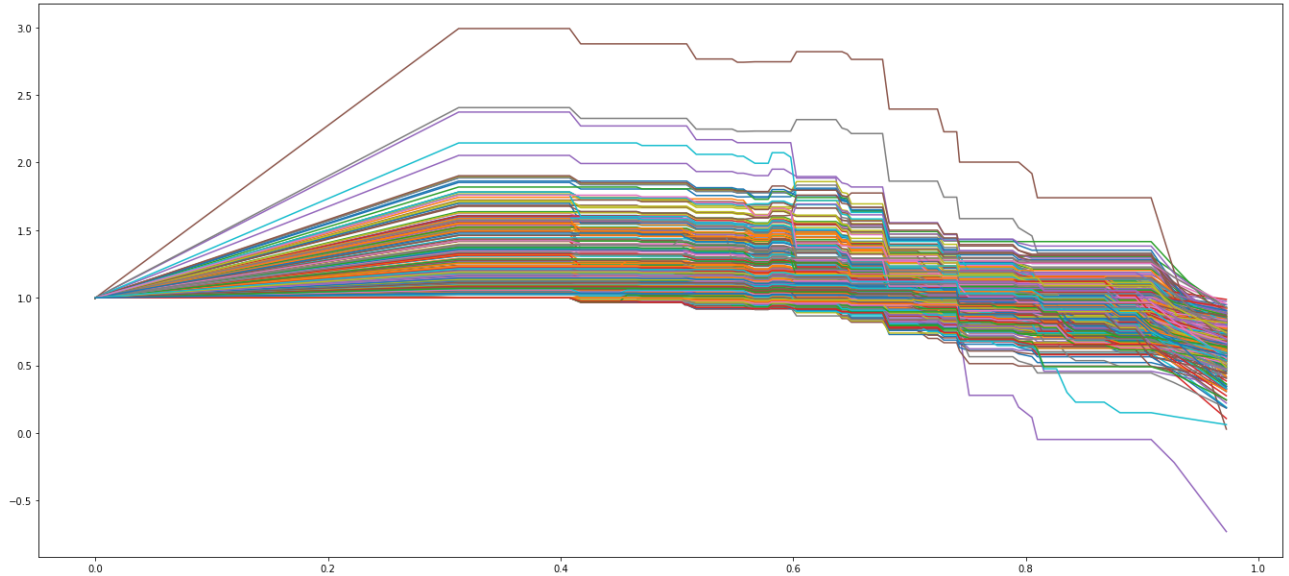
Often a centered version is used for user friendliness

Theory: derivative ICE version

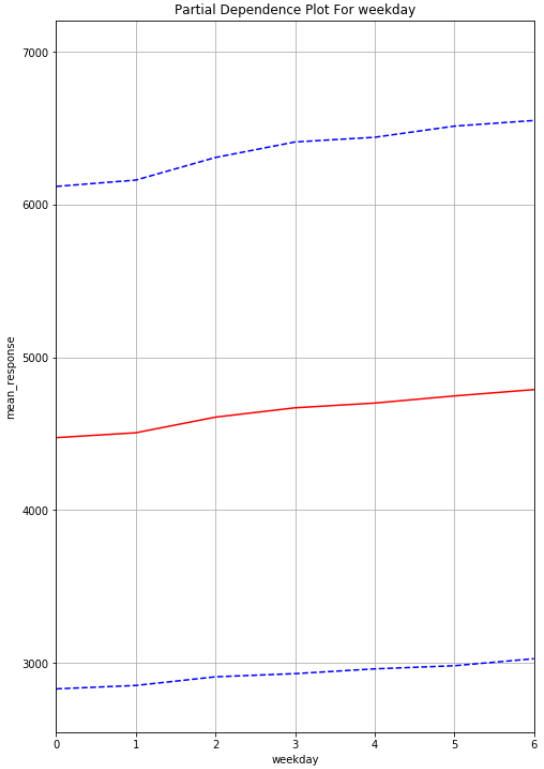
Goldstein et al: Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation

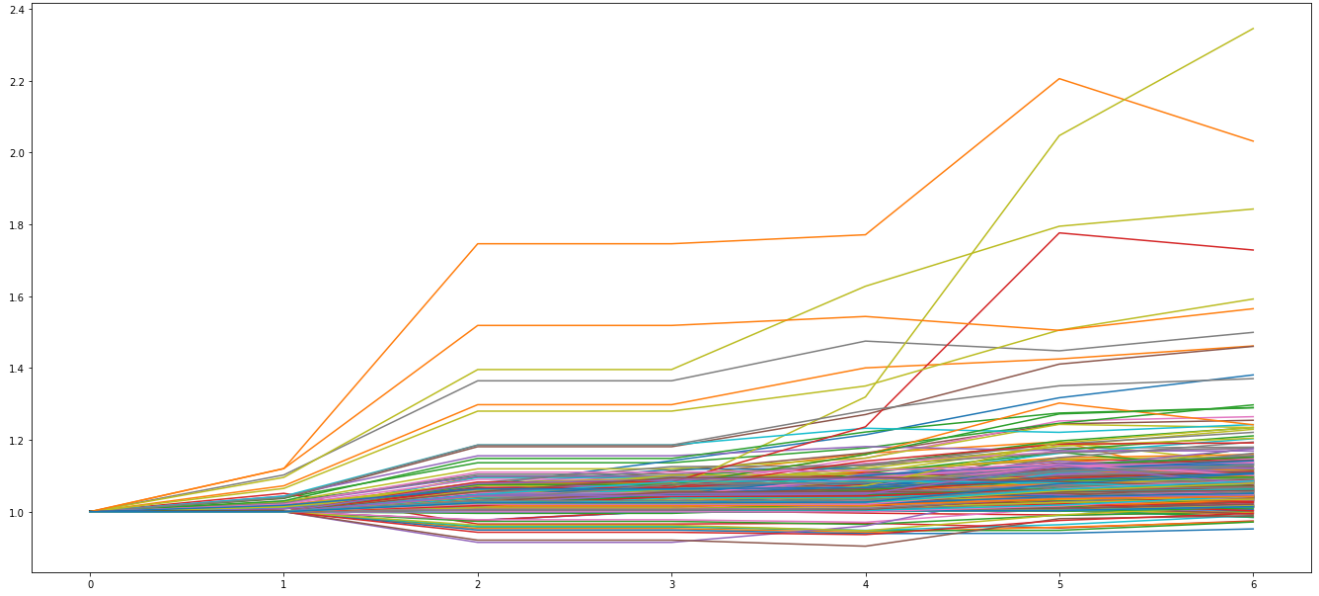
INDIVIDUAL CONDITONAL EXPECTATION PLOTS





INDIVIDUAL CONDITONAL EXPECTATION PLOTS





INDIVIDUAL CONDITONAL EXPECTATION PLOTS

Local technique, model-agnostic to relation between feature and response

- Very intuitive for layman, causal, easy to explain, visual
- Interactions taken into account (some what)
- Identifying outliers and unexpected behavior:
- Varying PD plot should correspond to important features
- Only 1 in same plot
- Cluttered plots
- **If features are correlated we plot lines of very unlikely data points**

LEAVE ONE COVARIATE OUT

The idea behind LOCO is simple. For a row, replace one feature with missing, mean,… and compare the prediction before and after this replacement. Repeat over all features and use differences as feature importance.

It is model agnostic and can be seen as a local version of global feature importances.

Takes into account non-linear effects.

Lei et al: Distribution-free predictive inference for regression

LEAVE ONE COVARIATE OUT

Local technique, model-agnostic to determine feature importance effect on respons

- Signed
- Can be made global (confidence intervals)

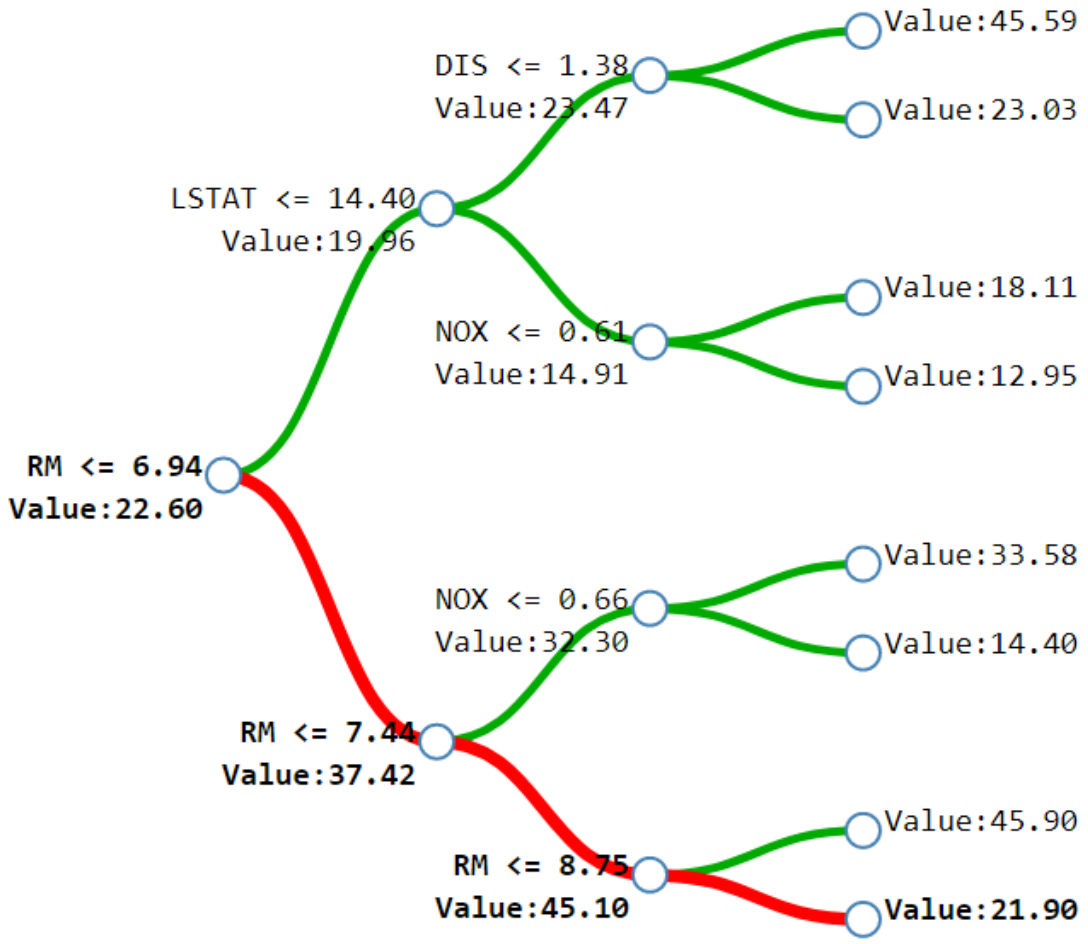
- Feature-by-feature
- Strong complex interactions interfer with the abilities of LOCO

TREEINTERPRETER

Very simple method, rapid increase in popularity

Model specific: Decision Tree and Random Forest (averaging)

Follow the path of a datapoint and keep track of how the prediction changes at each splitting node



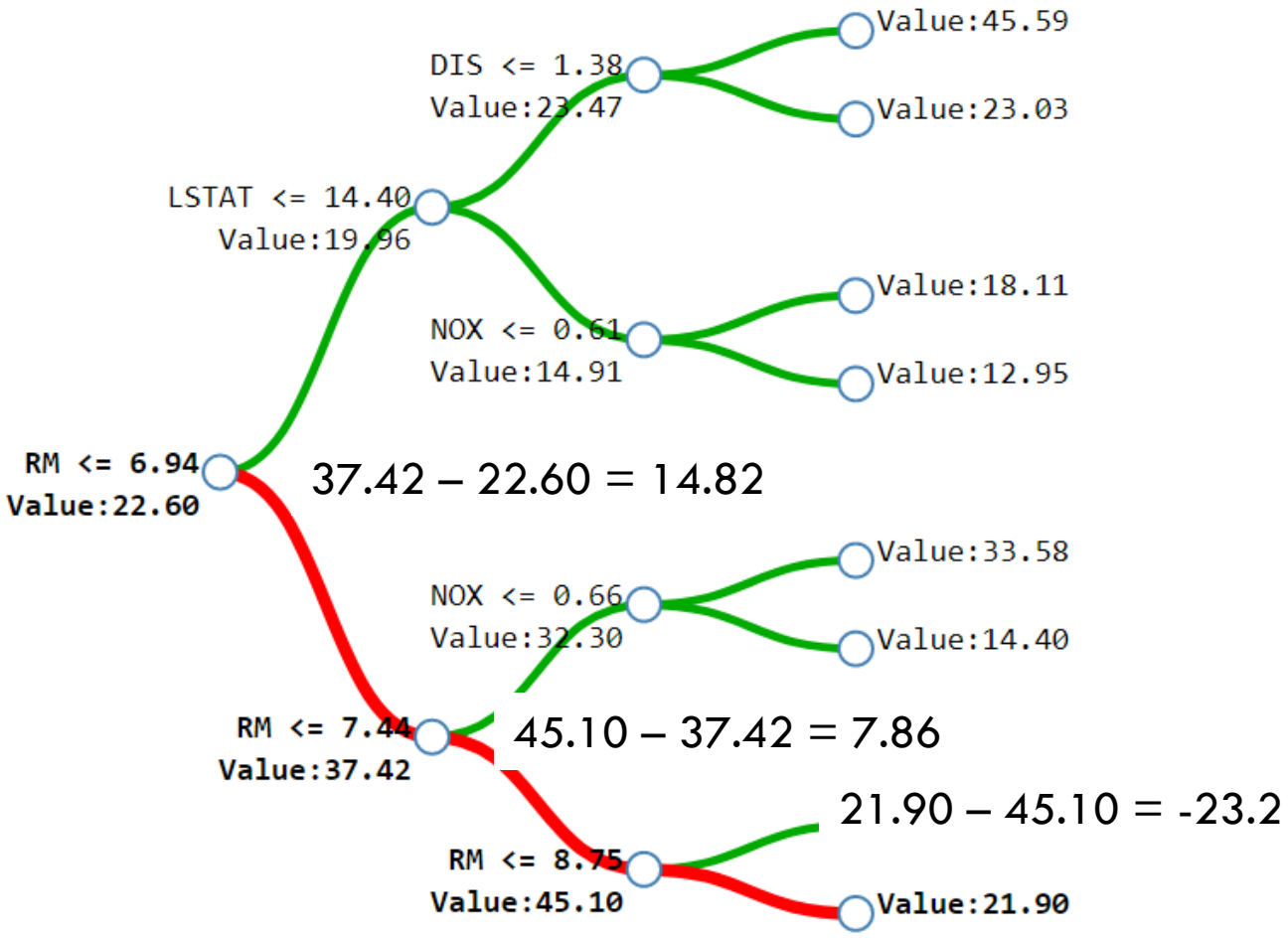
Saabas: http://cs.ioc.ee/~tarmo/tday-kao/saabas-slides.pdf

TREEINTERPRETER

Very simple method, rapid increase in popularity

Model specific: Decision Tree and Random Forest (averaging)

Follow the path of a datapoint and keep track of how the prediction changes at each splitting node



TREEINTERPRETER

Local technique, model-specific to determine feature importance on the prediction in trees for every split in the tree

- Human understandable: rules and the effects that sum to final prediction
- Signed

- Not sensitivity analysis: small change can make point follow different path
- Biased towards the leaves, not the root of the tree, can be very counterintuitive

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

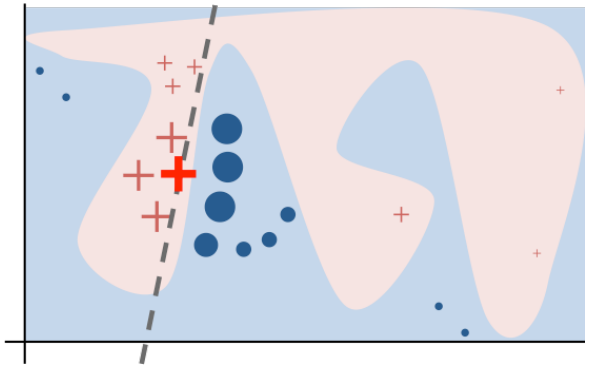
Lime is a formal approach to sensitivity analysis.

The idea is that locally everything is simpler:

- Linearity: Taylor expansion $f(x) \approx f(a) + f'(a)(x - a)$
- Sparsity: locally, a lot of features are constant and unimportant

Approximate model locally with linear model, decision tree…

What is local? What is closeness?



Ribeiro et al: "Why Should I Trust You?" Explaining the Predictions of Any Classifier

LIME

1. Sample perturbations: $z' \in \{0,1\}^{d'} \leftrightarrow \ z$ in data
   This corresponds to putting data to "missing" (words, pixels, features)
2. Locally weighted linear regression g (metric *D* to be chosen)
   $\pi_x(z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right)$ and $min \sum_{z,z'} \pi_x(z)\left(f(z) - g(z')\right)^2$
3. Select K best features with Lasso (for sparsity)
4. Refit locally weighted linear regression with K best features
5. Use local model to interpret the effects for the chosen data point and to study counterfactuals etc

LIME FOR TEXT

1. Pick data point to be explained, human interpretation of text is based on word
2. Create perturbed data by removing words
3. Distance can be cosine, TF-IDF, word embedding related?
4. Fit a weighted lasso model
5. Interpret

LIME FOR IMAGES

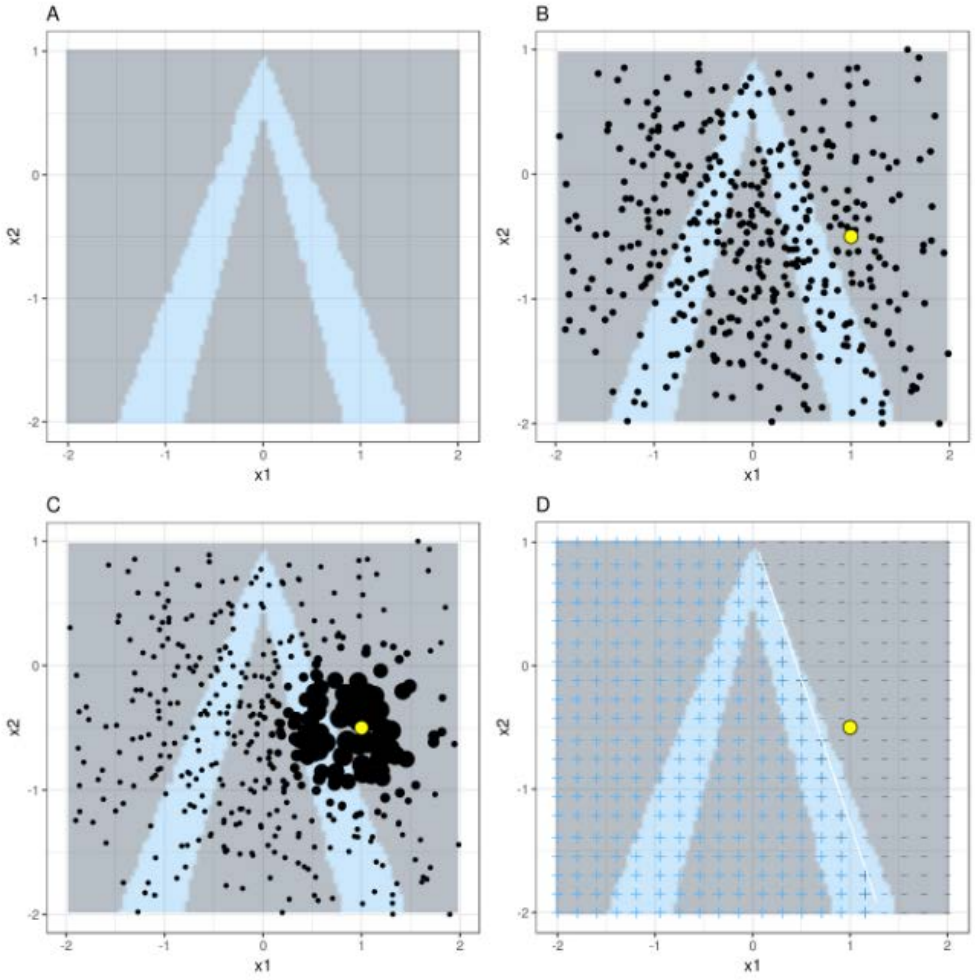
1. Pick data point to be explained, human interpretation of text is based on superpixels (Felzenszwalb)
2. Create perturbed data by removing / graying
3. Distance can L2,…?
4. Fit a weighted lasso model
5. Interpret

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Ribeiro et al: "Why Should I Trust You?" Explaining the Predictions of Any Classifier

LIME FOR STRUCTURED DATA

1. Pick data point to be explained
2. Create perturbed data by independently sampling data based on statistics
3. Distance can L2,…?
4. Fit a weighted lasso model
5. Interpret



Molnar: Interpretable Machine Learning

LIME

Lime gives is a local model-agnostic method via sensitivity analysis. It gives a local model, the coefficients can be used for trends or specific explanations of data points

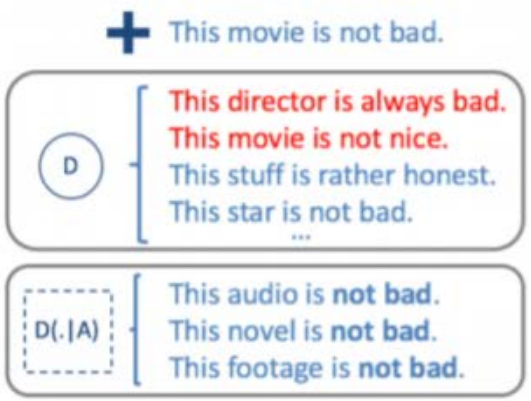
Several modifications are possible. K-lime is K-means followed by regression in each cluster. Cluster can be human determined to incorporate domain knowledge

- Local Model, versatile
- Sparse
- So many hyperparameters: Lasso, distance, kernel, model choice,…
- Conflicting local predictions
- Sum of importances might not sum to predicition
- Computationally expensive

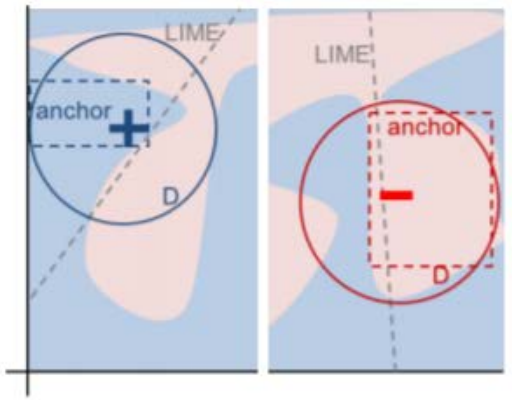
ANCHOR

Anchor is a recent development by the authors of LIME. It looks for rules (predicates) that the selected data point satisfies, as well as other points in the same predicted class (and no point in the other predicted class). I.e for instance with the anchor true, they all have the same prediciton.

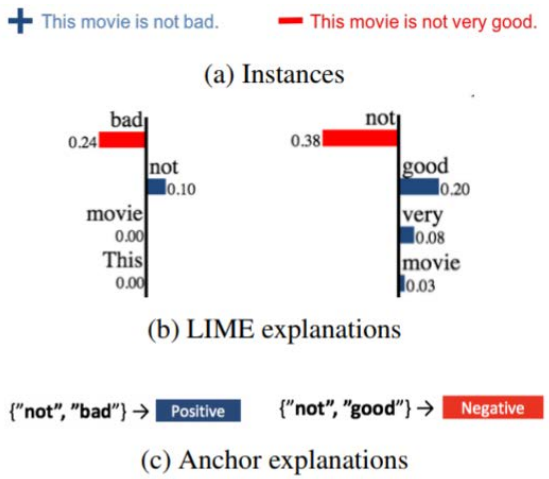
$$E_{D(Z|A)}[1_{f(x)=f(z)}] > \tau, A(x) = 1$$



(a) $\mathcal{D}$ and $\mathcal{D}(.|A)$



(b) Two toy visualizations



Ribeiro et al: Anchors: High Precision Model-Agnostic Explanations

WHICH METHOD TO USE?

We introduced several methods. What methods should be used in which situation? And what if they don't agree?

Moreover, for a certain data set several good models can be made (non convexity). Behavior and outcome of interpretation techniques can change across models, especially if different techniques are used. This unstability of explanations is very unsettling.

Therefore, it is good to consider several methods and distill those explanations that reoccur in most.

Such common insights are not always there, as the methods offer a different perspective on a very complicated function.

Focus more on stability of predictions, less on that of the model?

SHAPLEY VALUES

We have tons of possible "feature importances" that explain a prediction.

Just for Random Forest, we have lime, gain, split count, treeinterpreter, loco, permutation…

Can we get to a prefered method by posing extra conditions?

Consistency. If a model relies more on a feature, then the attributed *importance* for that feature should not decrease (comparison)

1. effect on expected *accuracy* when feature is removed (global)
2. change in the model's *output* when feature is removed (local)

Accuracy. The sum of all the feature importances gives the total importance (meaning)

Strumbelj et al: Explaining prediction models and individual predictions with feature contributions
Lundberg et al: A Unified Approach to Interpreting Model Predictions
Lundberg et al: Consistent feature attribution for tree ensembles

SHAPLEY VALUES

Consistency: only permutation

Trees are greedy: features near the root are most important. The tree methods are biased towards lower splits. When feature is more important, it's reported variable importance drops (inconsistent)

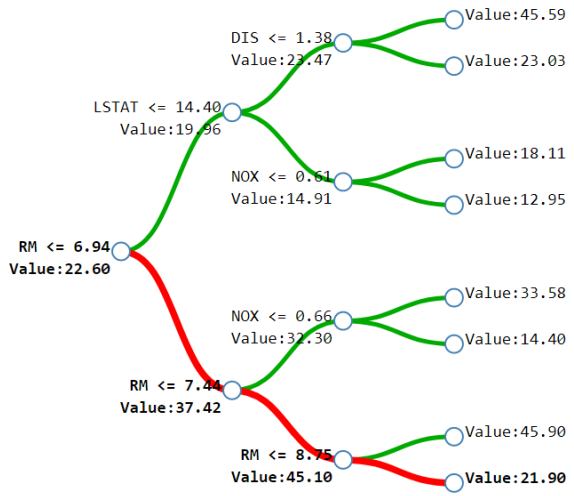
Very clear with TreeInterpreter

SHAPLEY VALUES

Consistency: only permutation

Trees are greedy: features near the root are most important. The tree methods are biased towards lower splits. When feature is more important, it's reported variable importance drops (inconsistent)

Very clear with TreeInterpreter



SHAPLEY VALUES

Accuracy: TreeInterpreter and gain

They work their way towards the final prediction / importance.

Permutation, LOCO not

They are computed separately for each feature, no reason for them to have this property.

Lime is heuristic

no reason for exact accurcy or consistency relation to hold

SHAPLEY VALUES

Combine the good properties of the permutation method and the stepwise importance methods

Focus on row. We want to assign a value to all features how much they contribute to overall prediction. This is a common task in game theory.

The Shapley value fairly assigns a value to each individual player (feature) depending on their contribution (to the prediction) in all possible coalitions.

The payout of a coalition of features S is

$$v(x_S) = E_{X_C}\big(f(X_S, X_C)\big) - E_X\big(f(X)\big) = f_S(x_S) - f_\emptyset(x_\emptyset)$$

Which is the PD of X_S minus the mean respons.

SHAPLEY VALUES

The payout of a coalition of features S is

$$v(x_S) = E_{X_C}\big(f(X_S, X_C)\big) - E_X\big(f(X)\big) = f_S(x_S) - f_\emptyset(x_\emptyset)$$

Which is the PD of X_S minus the mean respons.

Now suppose a new feature *j* joins the already present features S.

It's contribution is

$$\Delta_j(x_S) = \ v\big(x_{S\cup j}\big) - v(x_S) = f_{S\cup j}\big(x_{s\cup j}\big) - f_S(x_S)$$

This is already built up in such a way that it satisfies the accuracy criterium.

Intuitively, we build this up stepwise as TreeInterpreter

SHAPLEY VALUES

To make the method consistent, we look at the permutation idea, for a permutation π , let $B_j(\pi)$ be the indices before j in the permutation

The Shapley value is defined as

$$\phi_j(x) = \frac{1}{p!}\sum_{\pi} \Delta_{\mathrm{j}}\left(x_{B_j(\pi)}\right)$$

We are averaging differences in predictions over all possible orderings of the features.

The feature values enter a room in random order. All feature values in the room participate in the game (contribute to the prediction). The Shapley value is the average marginal contribution of feature value by joining whatever features already entered the room before.

https://christophm.github.io/interpretable-ml-book/

SHAPLEY ADDITIVE EXPLANATIONS

More formal

we want an additive feature attribution method: a linear function of binary variables (indicating feature presence)

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$$

Local accuracy: The explanations are truthful: $g(z') = f(z)$

Missingness: Missing features have no attributed impact to the model predictions

C*onsistency*: given two models f_1, f_2 and some input i always has more impact on the outcome of f_1, regardless of the other inputs, than i has a higher attribution in model f_1

SHAPLEY ADDITIVE EXPLANATIONS

Only one option that satisfies these 3 constraints and uses conditional expectation as function of simplified input

$$E[f(z)|\, z_S] \approx g(z')$$

For linear models $f(x) = \sum_{j=1}^{M} w_j x_j + b$, we get

$$\phi_0(f, x) = b \text{ and } \phi_i(f, x) = w_j(x_j - E[x_j])$$

SHAPLEY ADDITIVE EXPLANATIONS

Only one option that satisfies these 3 constraints and uses conditional expectation as function of simplified input

$$E[f(z)|\, z_S] \approx g(z')$$

Surprisingly, this method encompasses several others:

LIME (choice of kernel), DeepLift, Layer-Wise Relevance Propagation, classic Shapley values, LOCO, TreeInterpreter,…

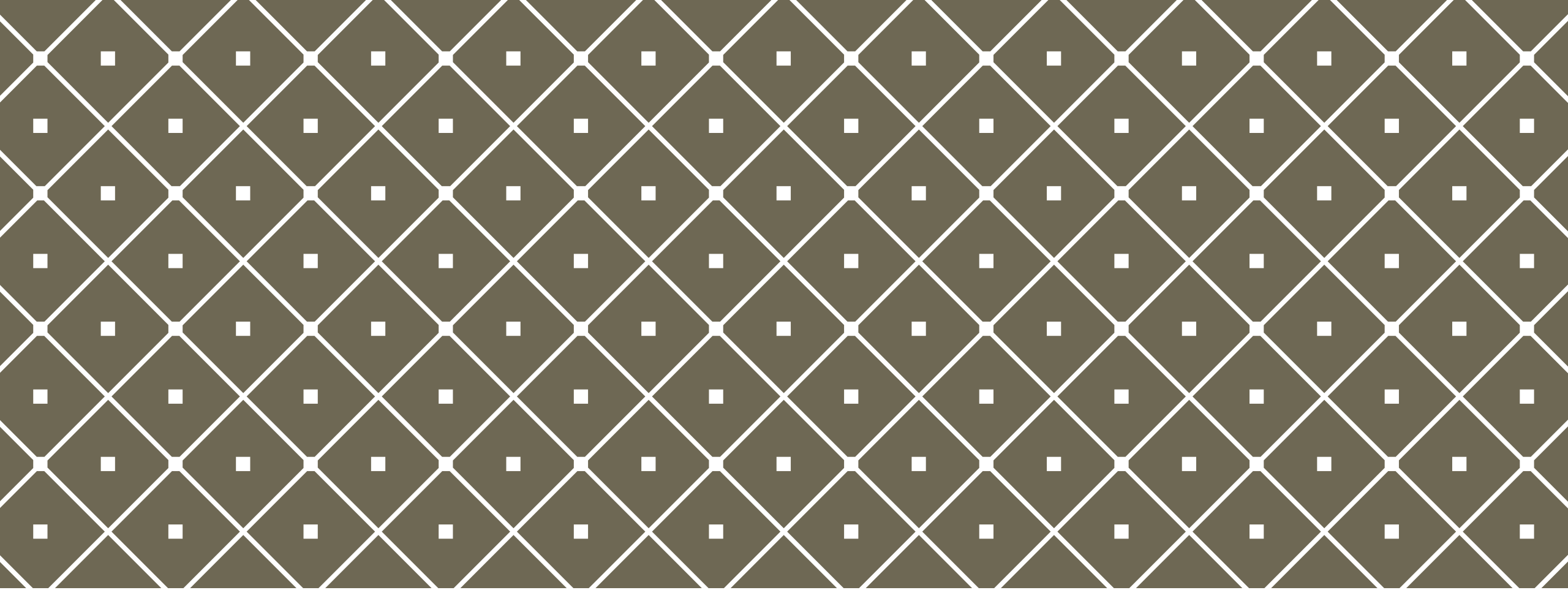
Several methods for calculating:
Model Agnostic: Shapley sampling values, Kernel SHAP
Model Specific: TreeSHAP, Max SHAP, Deep SHAP

Strumbelj et al: Explaining prediction models and individual predictions with feature contributions
Lundberg et al: A Unified Approach to Interpreting Model Predictions
Lundberg et al: Consistent feature attribution for tree ensembles

SHAPLEY ADDITIVE EXPLANATIONS

SHAP gives a theoretical footing, from game theory, to feature importances for a specific local instance.

- Theoretical footing, fair distribution among players / variables
- Allows for fair comparison between row, or models
- Completely model agnostic in theory
- Misinterpretation, it is not gain, or TreeInterpreter
- Conflicting local predictions
- Not sparse,
- No model so no counterfactual, causal inductions
- Computationally extremely expensive



DEEP LEARNING

Let the model speak

INTERPRETING DEEP NETWORKS

When a network is given more resources than the minimum needed to solve a given task , the symmetric, low-order, local solutions that humans seem to prefer are not the ones that the network chooses from the vast number of solutions available; indeed , the generalized delta method and similar learning procedures do not usually hold the "human " solutions stable against perturbations.

Denker et al: Large Automatic Learning, Rule Extraction, and Generalization

INTERPRETING DEEP NETWORKS

Use forward propagation to perturb or gray out the input, this is slow

- Maximum activation analysis
- Deconvolution

Use Backprop to compute attributions for all input features in a single

- Layer-wise Relevance Propagation: assign relevance starting at activation of final output
- Deeplift: similar as LRP but compare relative to the activations with a reference input

The last two methods, are explanations: $f(x) = \sum_{input\ i} h_i$

Alternatively, we can look at sensitivity or saliency methods

- Absolute value of partial derivatie of output to all inputs
- Partial derivatives x feature value: like in linear models (also explanation, not just saliency)

An attribution method with uniqueness properties as SHAP is integrated gradients

INTEGRATED GRADIENTS

We want to distribute the prediction score over the inputs in proportion to their contribution relative to a baseline

We do not evaluate partial derivatives at just at input x, but average the derivatives when input varies linearly from x to x_{ref}

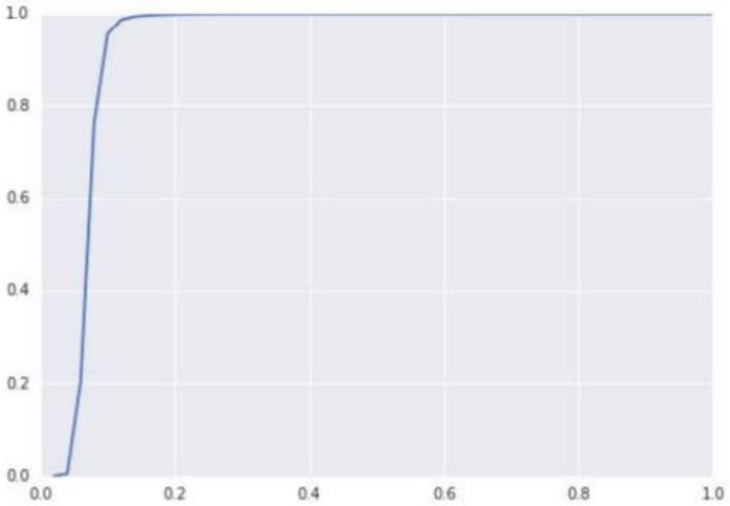
Idea is to find the 'sweet spot' when we can make a decision

Imagine turning on the lights in a room, when can you decide who's in the room? This is the crucial moments to look which inputs are crucial

Sundararajan et al: Axiomatic Attribution for Deep Networks

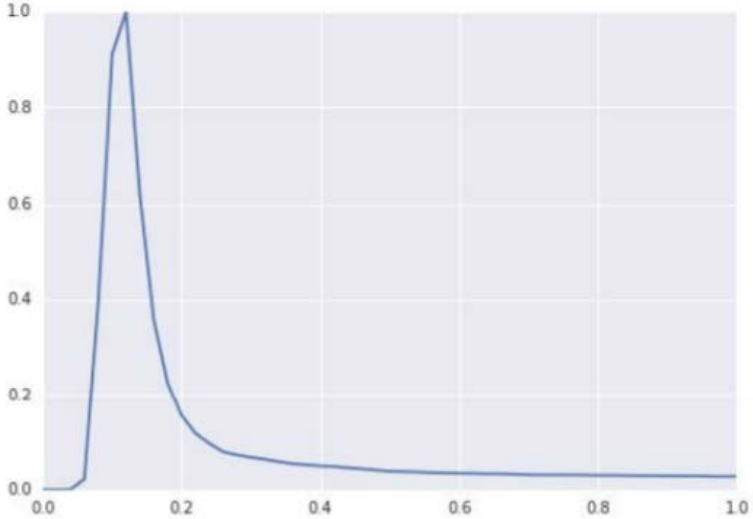
INTERPRETING DEEP NETWORKS

Predictions



Intensity α

Gradient (average)



Intensity α

http://www.unofficialgoogledatascience.com/2017/03/attributing-deep-networks-prediction-to.html

Taly et al: Why did the network make this prediction? (http://theory.stanford.edu/~ataly/Talks/sri_attribution_talk_jun_2017.pdf)

INTEGRATED GRADIENTS

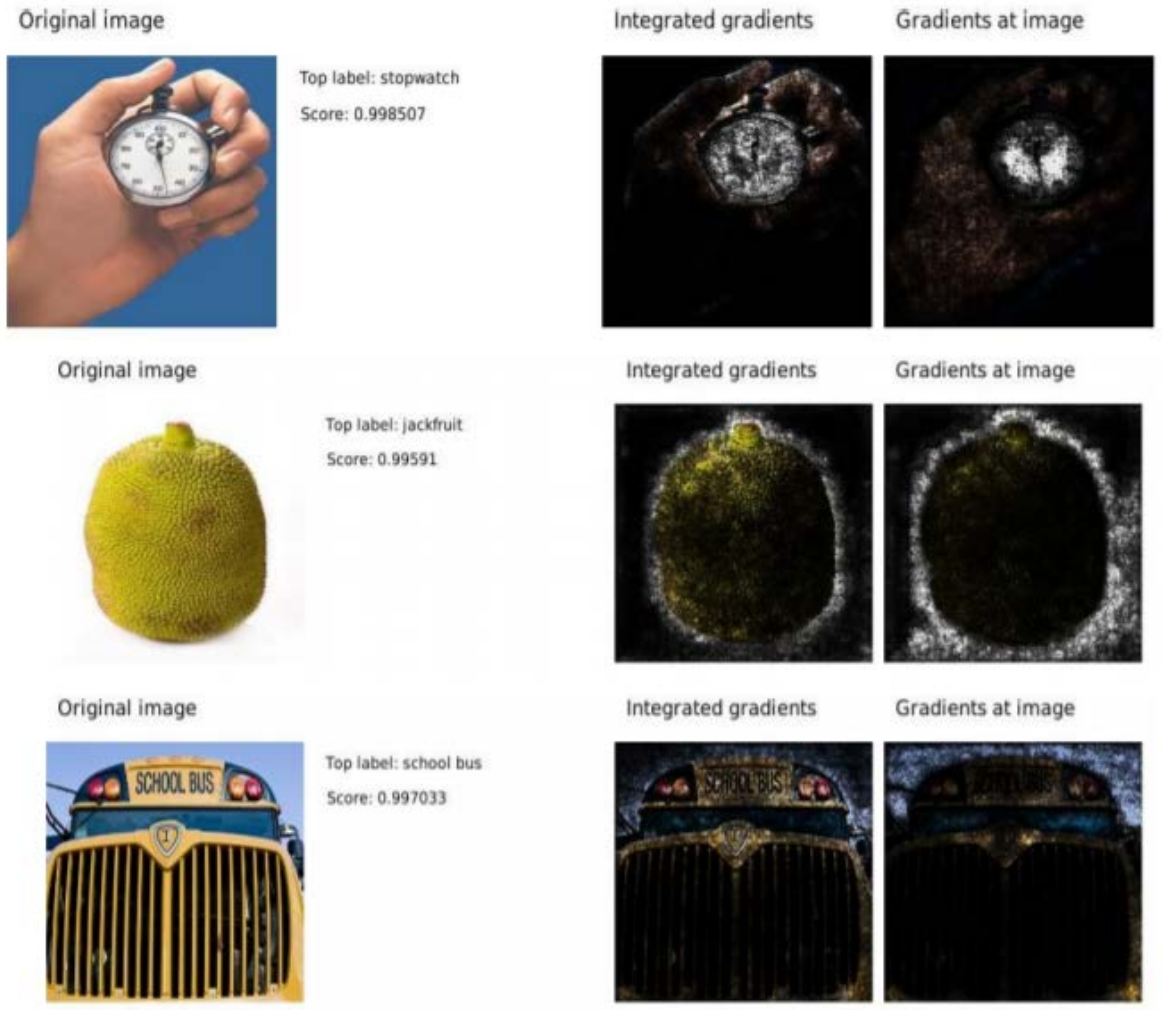
We want to distribute the prediction score over the inputs in proportion to their contribution relative to a baseline

$$G_i(x) = x_i \int_0^1 \frac{\delta f_i}{\delta x_i}\left(\alpha x + (1-\alpha)x_{ref}\right)d\alpha$$

Top label: mosque
Score: 0.999127

http://www.unofficialgoogledatascience.com/2017/03/attributing-deep-networks-prediction-to.html
Taly et al: Why did the network make this prediction? (http://theory.stanford.edu/~ataly/Talks/sri_attribution_talk_jun_2017.pdf)

INTEGRATED GRADIENTS



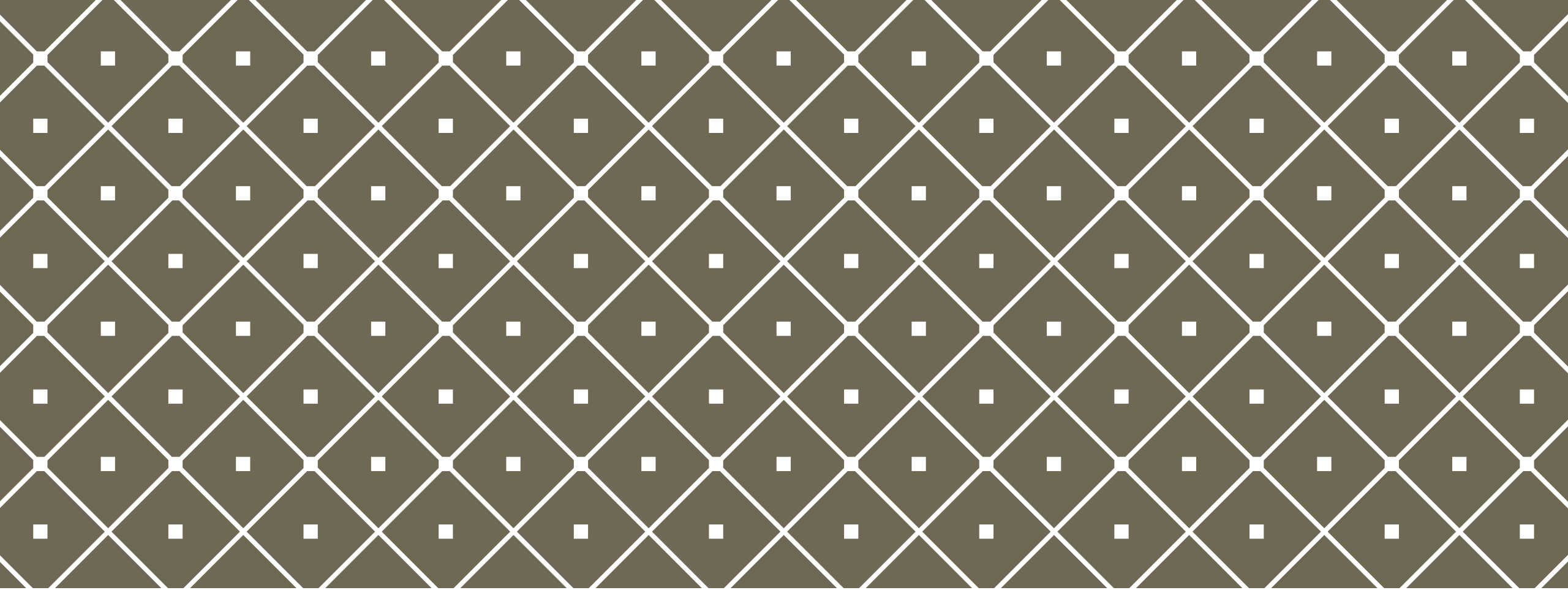
INTEGRATED GRADIENTS

Integrated gradients is the unique method that satisfies:

- **Sensitivity:** If changing a variable, changes the output, it gets a nonzero attribution (and vice versa)
- **Implementation Invariance:** models that have the same outputs give the same attributions
- **Linearity:** attributions for sum of models is sum of attributions
- **Symmetry:** if model is symmetric in some input, the attributions are equal
- **Completeness:** the sum of the attributions is equalt to predicition – baseline

Unsurprisingly, there is a connection with game theory and Shapley values

Sundararajan et al: Axiomatic Attribution for Deep Networks



CONCLUSION & REFERENCES

Resources & tools

CONCLUSION

Interpretability of complex machine learning or deep learning models is a new and booming field, every day new articles and methods appear.

For now, there is no golden standard and it is advisable to combine several techniques for interpretation.

Don’t trust your models blindly, but neither your interpretation methods!

REFERENCES

General References

https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning
https://christophm.github.io/interpretable-ml-book/

Packages

LIME: https://github.com/marcotcr/lime
SHAP: https://github.com/slundberg/shap
ICE: https://github.com/AustinRochford/PyCEbox
TreeInterpreter: https://github.com/andosa/treeinterpreter
General: https://github.com/datascienceinc/Skater
General: https://github.com/TeamHG-Memex/eli5
Commercial: http://docs.h2o.ai/driverless-ai/latest-stable/docs/booklets/MLIBooklet.pdf