Interpretable Machine Learning

Do you know what your model is doing?

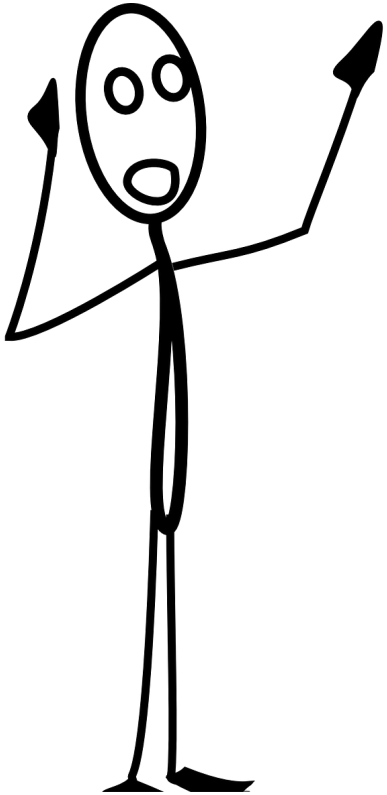
Marcel Spitzer

Offenburg, 07.03.2019

Marcel Spitzer
Big Data Scientist @ inovex

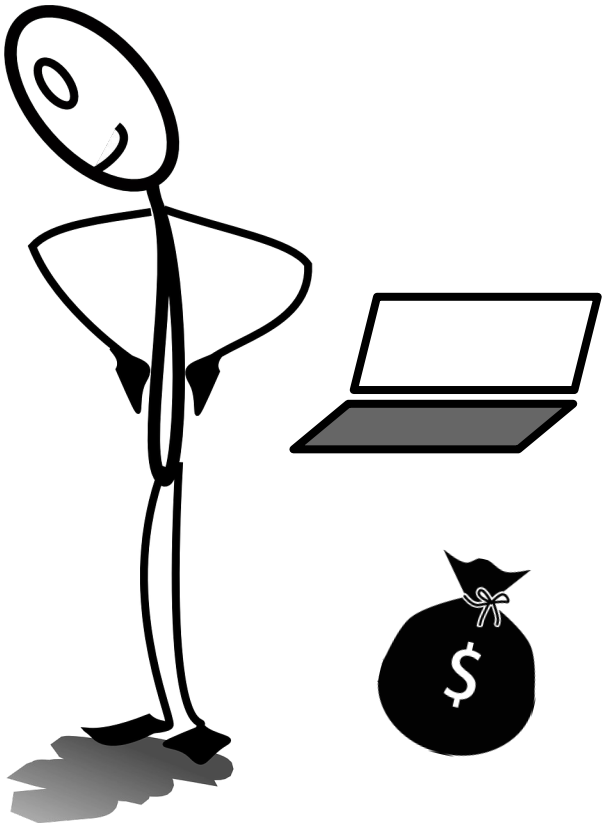
- Applied Mathematics, Data Science
- SW Engineering, Machine Learning
- Big Data, Hadoop, Spark

mspitzer@inovex.de
@mspitzer243



Mike

Creditor

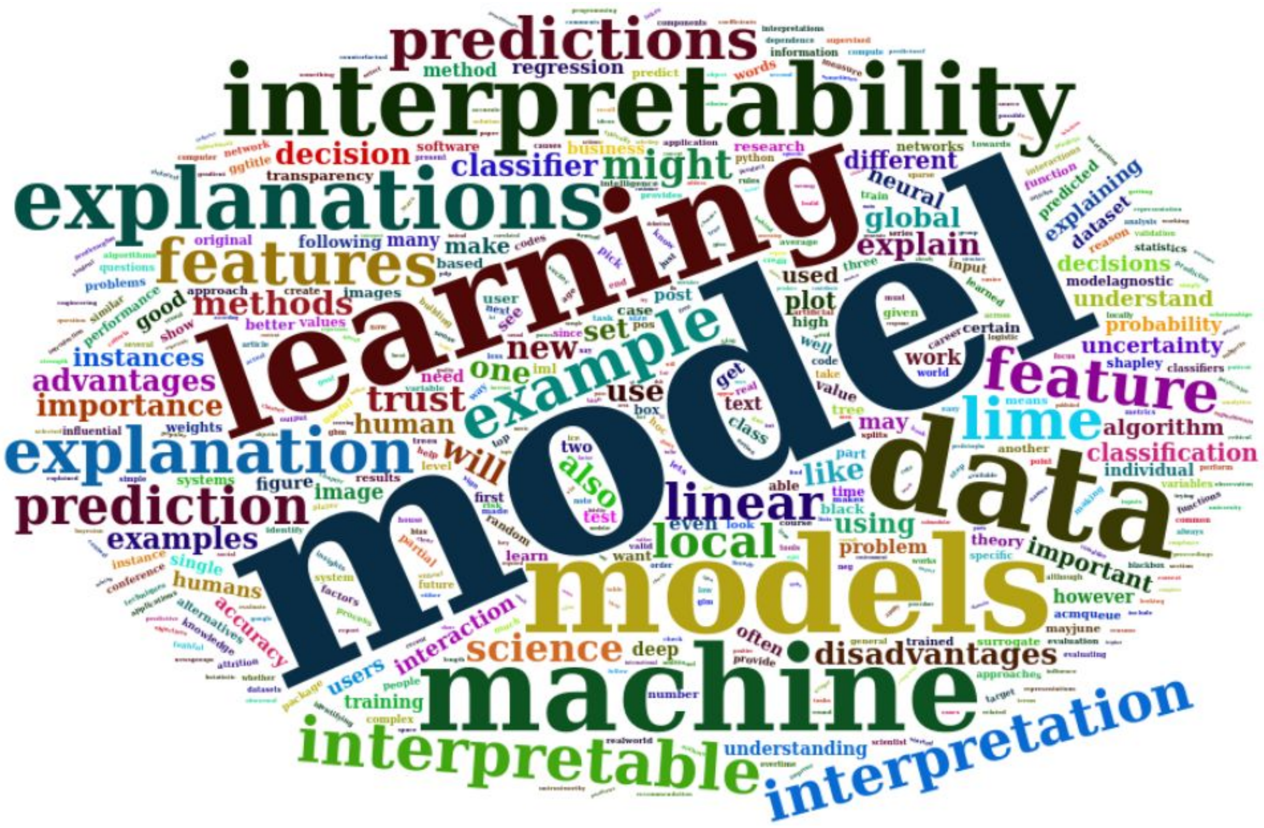



Hey! I need a loan.
Can I get one?


Hey! I need a loan.
Can I get one?


Hmmm…
No, rather not.


But...
WHY?
?


model
learning
interpretability
models
data
machine
explanations
predictions
explanation
interpretable
interpretation
feature
features
prediction
example
linear
lime
local
methods
science
disadvantages
trust
human
importance
advantages
instances
examples
decision
classifier
might
different
global
explain
decisions
understand
probability
uncertainty
algorithm
classification
important
however
new
use
one
plot
set
image
humans
accuracy
interaction
users
training
understanding
surrogate
regression
method
neural
explaining
dataset
individual
good
performance
many
make
following
original
weights
influential
figure
systems
single
instance
alternatives
knowledge
package
people
complex
business
software
transparency
network
research
networks
python
predicted
function
shapley
work
like
using
problem
theory
specific
deep
often
trained
mayjune
acmqueue
text
class
two
also
test
first
learn
random
partial
factors
will
better
values
approach
images
create
based
post
user
case
high
given
certain
used
input
three
global
statistics
modelagnostic
classifiers
world
value
tree
box
may
means
another
time
black
able
even
look
course
want
number
provide

Interpretation is the process of giving explanations to humans.

~ Kim B., Google Brain, Interpretable Machine Learning (ICML 2017)

“Interpretability is the degree to which an observer can understand the cause of a decision.”

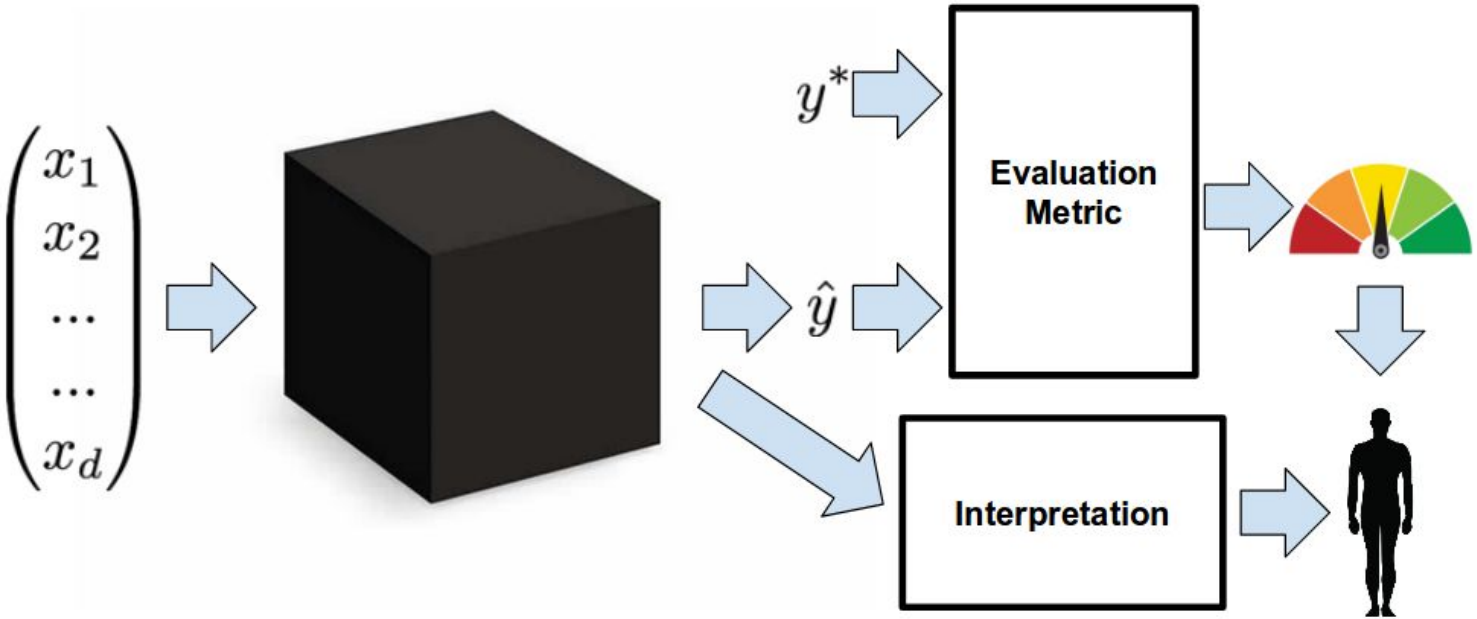
~ Miller T., 2017, Explanation in AI: Insights from the Social Sciences

- humans create decision systems
- humans are affected by decisions
- humans demand for explanations

(a) Husky classified as wolf (b) Explanation

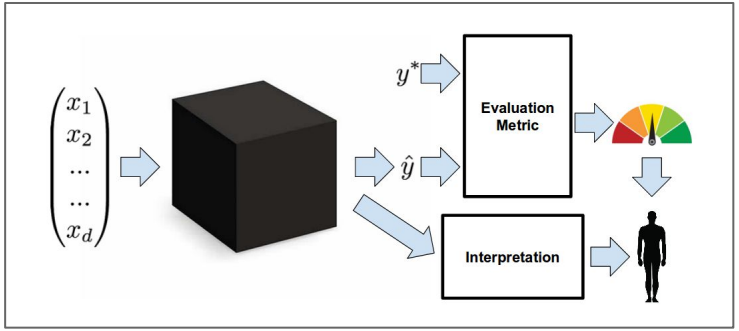
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

The additional need for interpretability



The additional need for interpretability

The decision process of a model should be **consistent to the domain knowledge** of an expert.



In particular, it ...

- should not encode bias
- should not pick up random correlation
- should not use leaked information

hard to capture in a numeric quantity

1 Use models that are **intrinsically interpretable** and known to be easy for humans to understand.

---

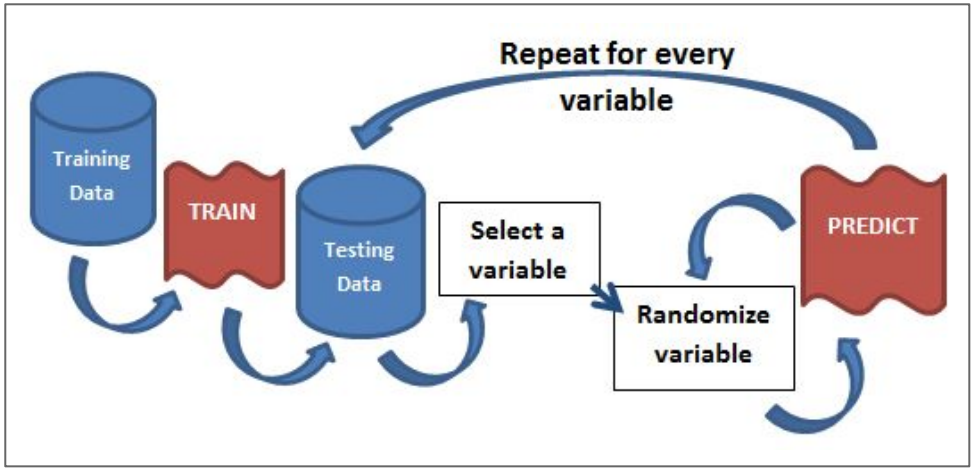
2 Train a black box model and apply **post-hoc interpretability techniques** to provide explanations.

Post-hoc interpretability techniques

	Global	Local
Model-specific	Model Internals, Intrinsic Feature Importance	Rule Sets (Tree Structure)
Model-agnostic	Partial Dependence Plots, Feature Importance (perm-based), Global Surrogate Models	Individual Conditional Expectation, Local Surrogate Models

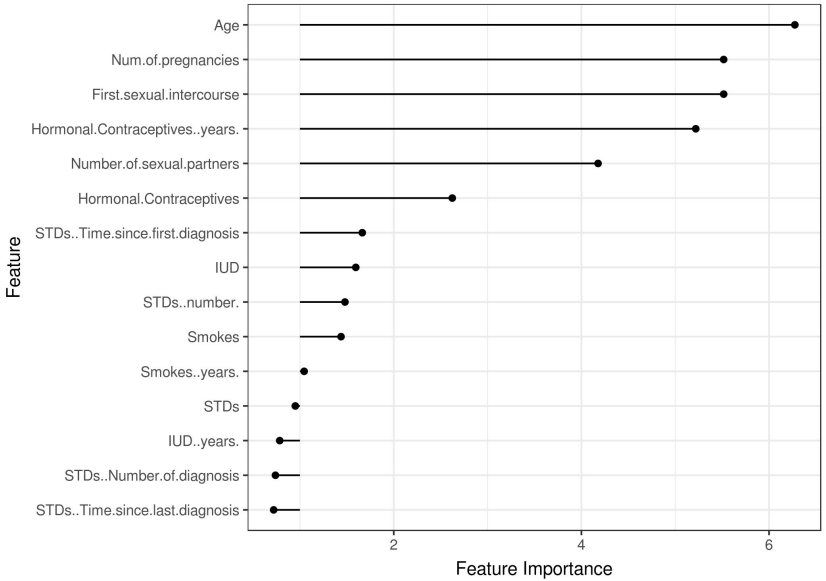
Feature Shuffling

- averages degradation measured by a certain loss function after repeatedly permuting single features
- feature is important if the error significantly increases after a shuffle



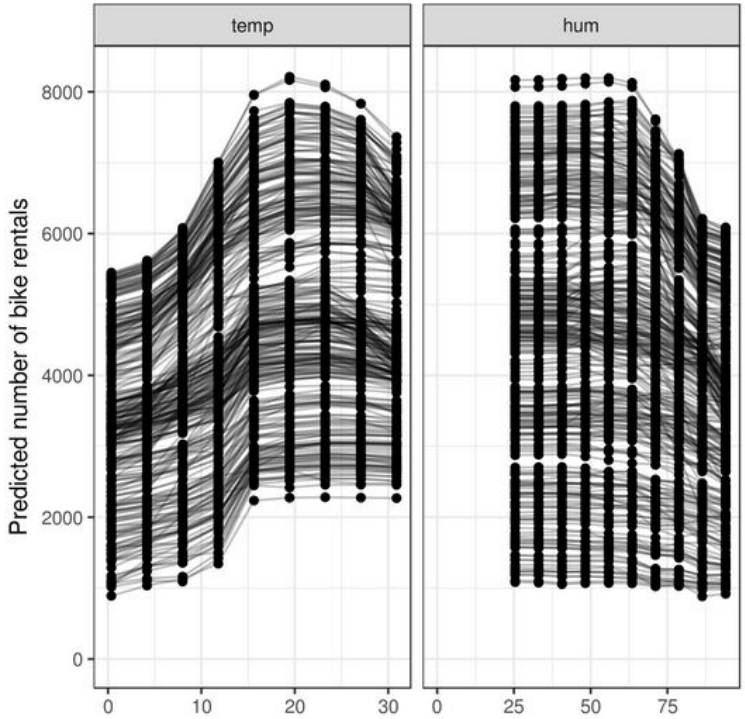
Feature Shuffling

- estimates feature importance
- highly compressed, global insight
- tied to some loss function
- not applicable in high dimensional domains (e.g. image/text classification)



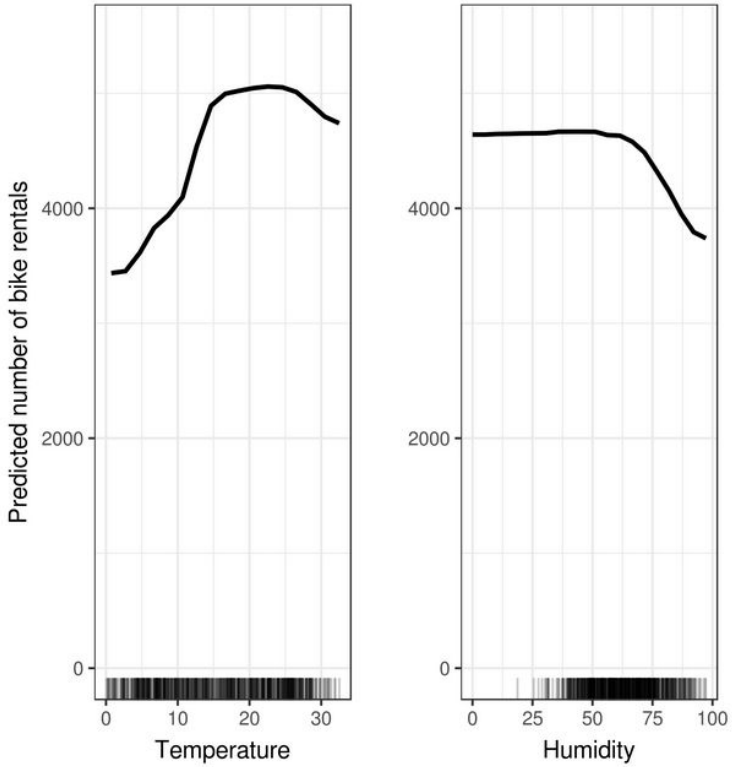
Individual Conditional Expectation (ICE)

- shows dependence of the response on a feature per instance
- single curve results from varying a certain feature for a given instance
- inconsistent pattern indicates multicollinearity



Partial Dependence Plots (PDP)

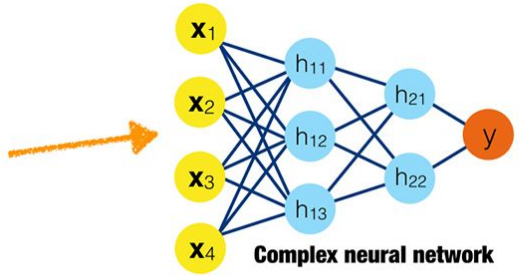
- PDP curve is the result of averaging ICE curves
- very intuitive, easy to understand
- assumption of independence is a strong drawback



Global Surrogate Models

BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.18	MORT	7
1	0.42	HELOC	10
0	0.11	MORT	10
0	0.21	MORT	1

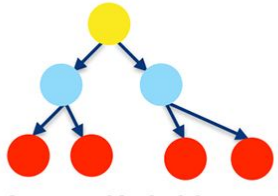
1. Train a complex machine learning model



Complex neural network

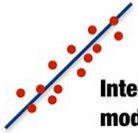
BAD	PREDICTED_BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.47	0.18	MORT	7
1	0.82	0.42	HELOC	10
0	0.18	0.11	MORT	10
0	0.12	0.21	MORT	1

2. Train an interpretable model on the original inputs and the predicted target values of the complex model



Interpretable decision tree

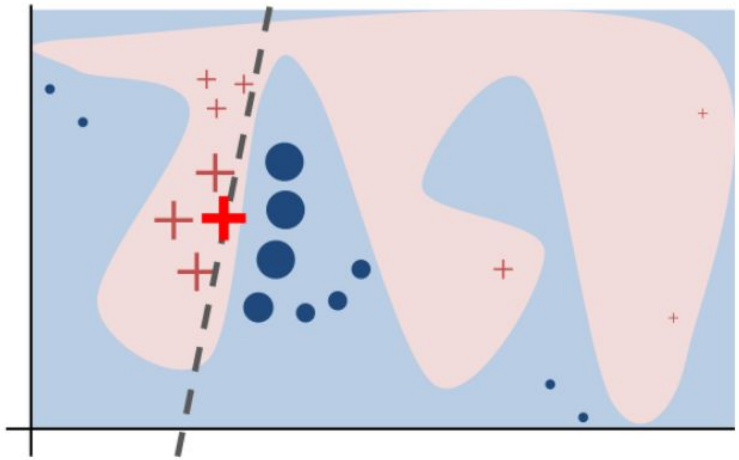
Or



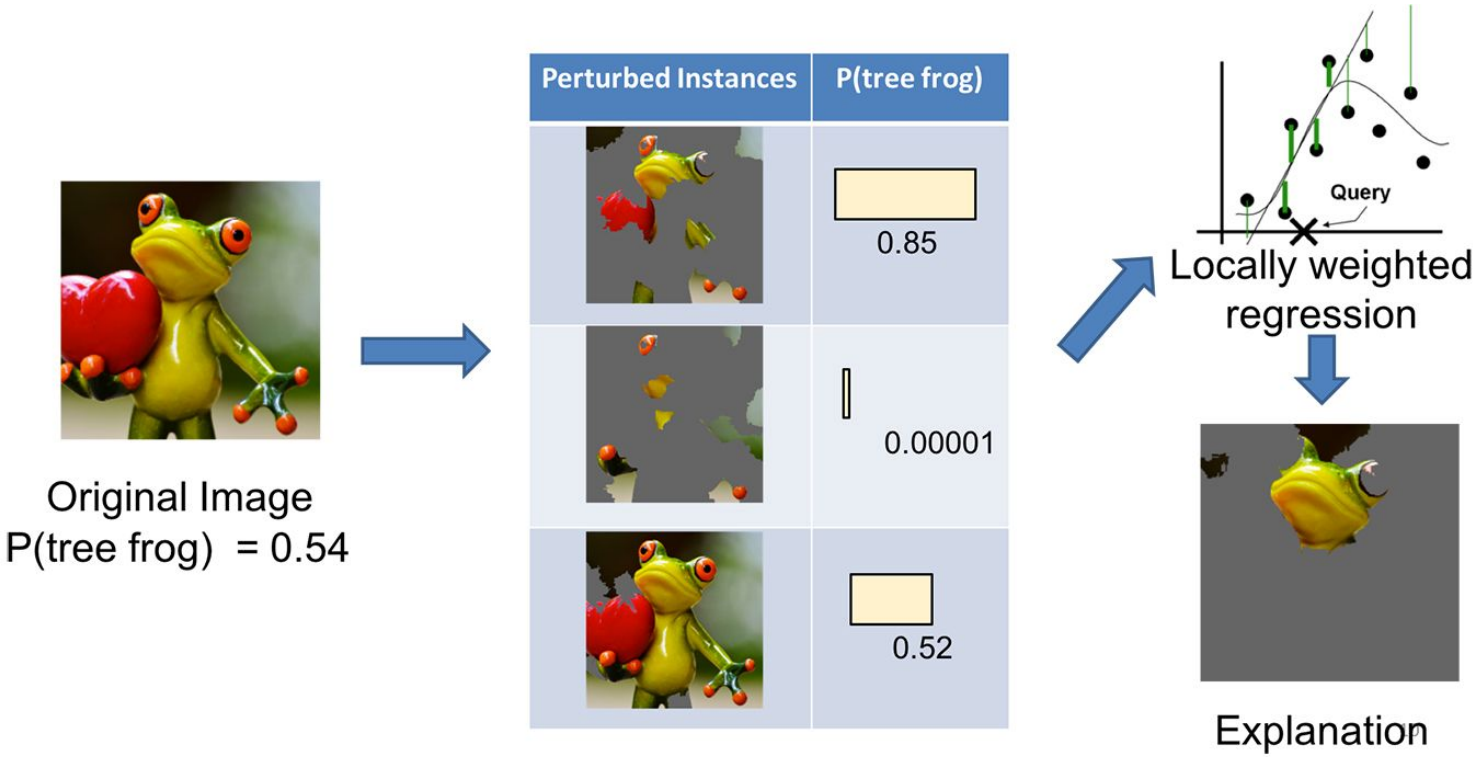
Interpretable linear model

Local Surrogate Models: LIME

- feeds original model with small variations of instance to be explained
- sampled instances are weighted by proximity to the instance of interest
- interpretable models are fit locally on observed outcome

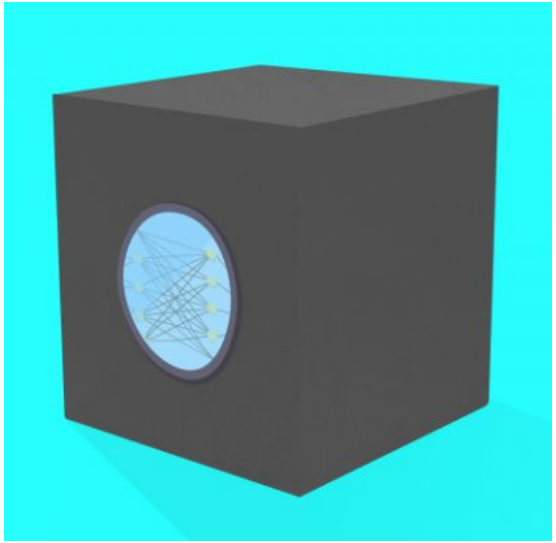


Local Surrogate Models: LIME



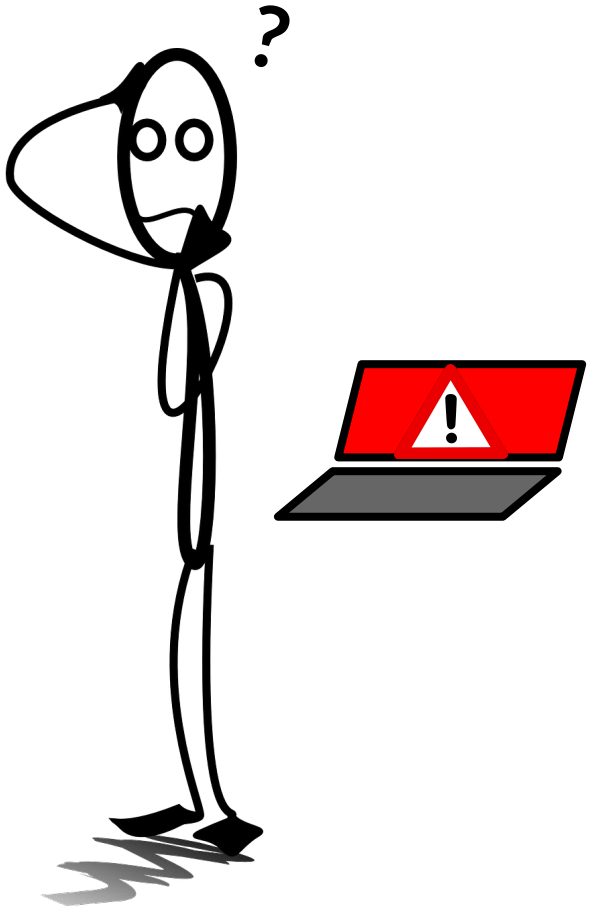
Conclusion

- performance metrics are crucial for evaluation, but they **lack explanations**
- criteria like fairness and consistency are much harder if not **impossible to quantify**
- the problem with blackboxes is the **lack of trust** caused by their opaque nature
- **transparency is key** to achieving trust and acceptance in the mainstream



Conclusion

don’t end up like this!



Conclusion

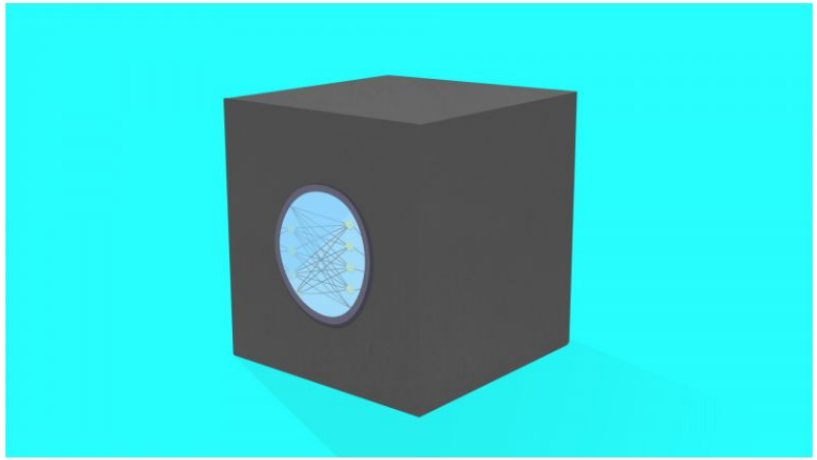
don't end up like this!

Resources

- Molnar C., 2018, Interpretable Machine Learning - A Guide for Making Black Box Models Explainable
- Gill N., Hall P., 2018, An Introduction to Machine Learning Interpretability
- Zhao Q., Hastie T., 2017, Causal Interpretations of Black-Box Models
- Kim B., Doshi-Velez F., 2017, Interpretable Machine Learning: The fuss, the concrete and the questions
- Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. Why should i trust you? Explaining the predictions of any classifier

inovex BLOG

Strategies Applications Infrastructure Analytics Methods



Machine Learning Interpretability: Do You Know What Your Model Is Doing?

Gepostet am: 13. Februar 2019

Marcel Spitzer

Vielen Dank

Marcel Spitzer
Big Data Scientist
mspitzer@inovex.de

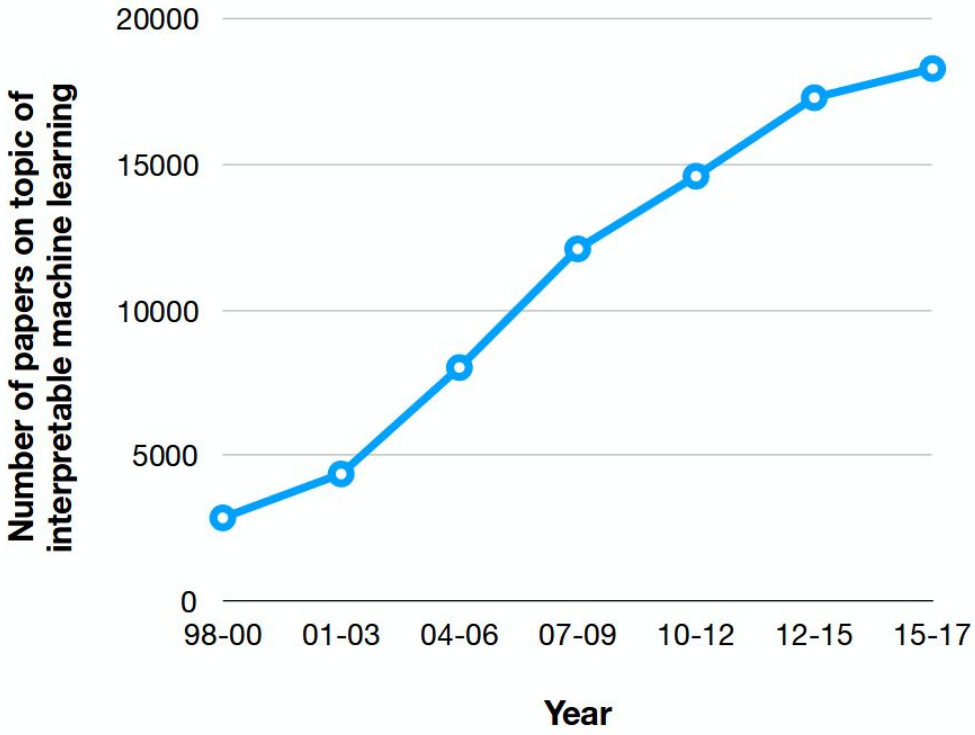
inovex GmbH
Schanzenstraße 6-20
Kupferhütte 1.13
51063 Köln

Why do we need interpretability?

safety	system should provide sound decisions
curiosity	understand something unexpected
debugging	behaviour should be predictable
optimality	optimize for true objectives

When we may not need interpretability

low risk	no significant consequences
awareness	problem is well-studied
vulnerability	prevent people from gaming the system



NIPS 2016 workshop on Interpretable Machine Learning for Complex Systems

ICML 2016 Workshop on Human Interpretability in Machine Learning

NIPS 2017 Workshop on Interpreting, Explaining and Visualizing Deep Learning

NIPS 2017 symposium and workshop: interpretable and Bayesian machine learning

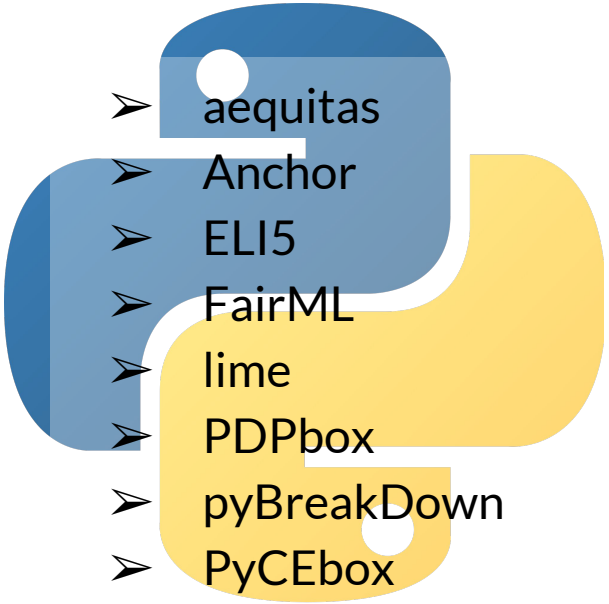
ICML 2017 Workshop on Human Interpretability in Machine Learning

ICML 2018 Workshop on Human Interpretability in Machine Learning

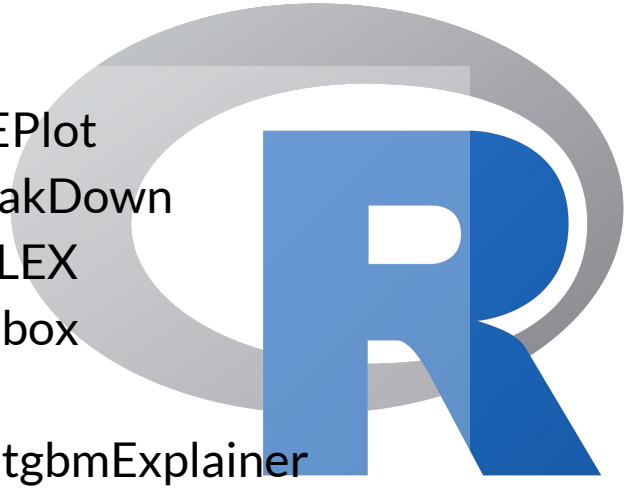
Recommendations for interpretability techniques

- **Who is the recipient?**
  - Lay-Men → rather intuitive, example-based local explanations
  - Analysts → global surrogates, perm-based feature importance
  - Authorities → intrinsically interpretable models

- **What are the explanations used for?**
  - Debug/Improve → PDP & ICE curves
  - Decision support → rule-based explanations
  - Auditing/Legal → intrinsically interpretable models



- aequitas
- Anchor
- ELI5
- FairML
- lime
- PDPbox
- pyBreakDown
- PyCEbox
- SHAP
- Skater
- tensorflow/model-analysis
- TreeInterpreter



- ALEPlot
- breakDown
- DALEX
- ICEbox
- iml
- lightgbmExplainer
- lime
- live
- pdp
- vip
- xgboostExplainer