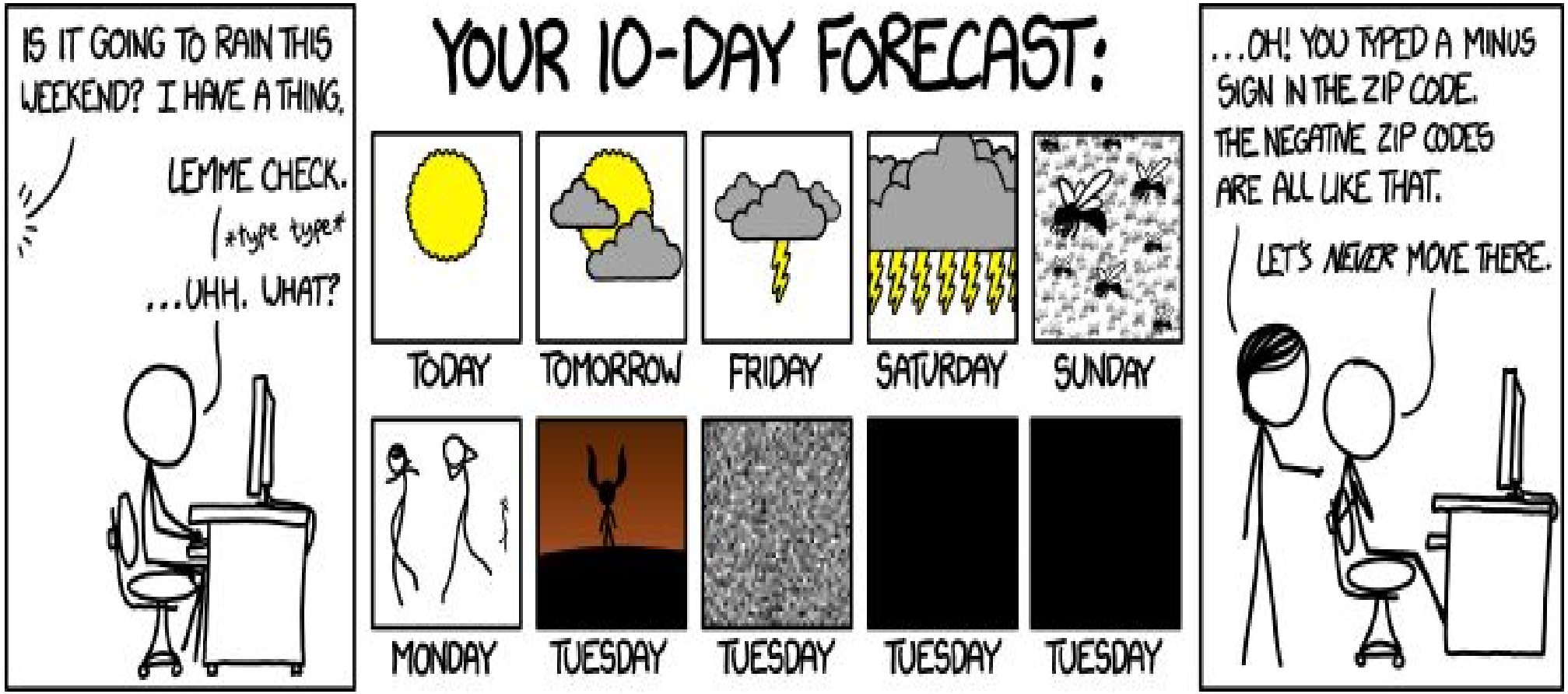
DATA**SCIENCE**

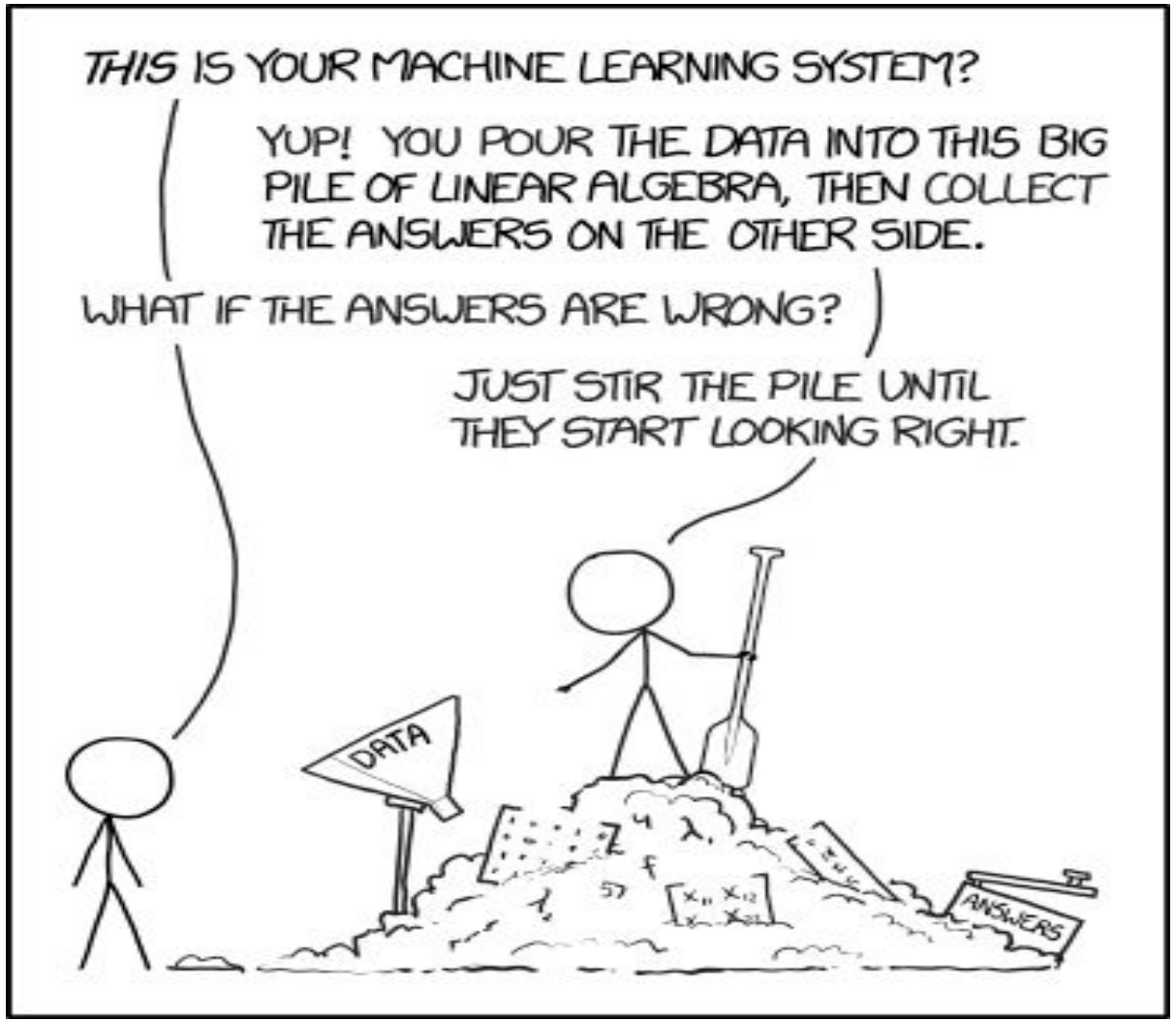
Learning to Learn Model Behavior: How to use "*human in the loop*" ?

PREDICTIVE MODELING: FUN OR MISERY?

PREDICTIONS OFTEN GO WRONG



WHEN AN ERROR OCCURS



ABOUT ME

Pramit Choudhary

@MaverickPramit

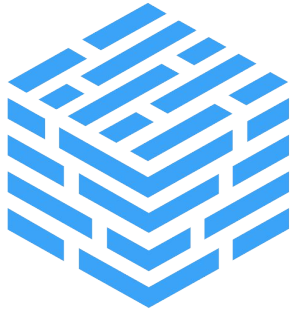
https://www.linkedin.com/in/pramitc/

https://github.com/pramitchoudhary

I am a Lead data scientist at DataScience.com. I enjoy applying and optimizing classical (Machine Learning) and Bayesian design strategy to solve real-world problems. Currently, I am exploring on better ways to evaluate and explain Model learned decision policies. I am also a member of AAAI and organizer of PyData So Cal meet-up group.

AGENDA

- DEFINE MODEL INTERPRETATION
- UNDERSTAND THE NEED FOR MODEL INTERPRETATION
- DISCUSS DICHOTOMY BETWEEN PERFORMANCE AND INTERPRETATION
- INTRODUCE SKATER
- UNDERSTANDING ANALYTICAL WORKFLOW
- DEMO
- Q&A



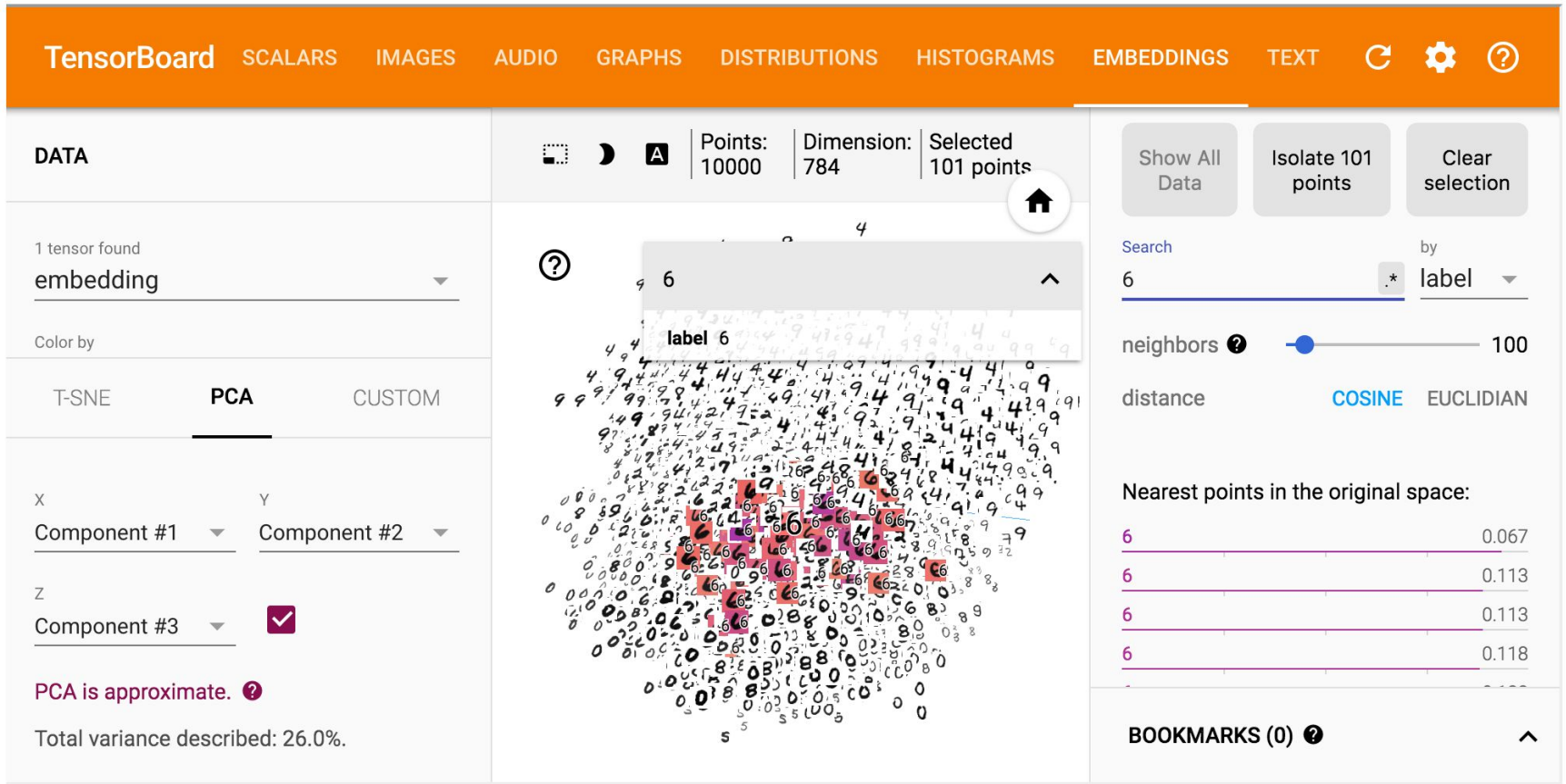
DATA**SCIENCE**.COM

DEFINE INTERPRETATION

- Definition is subjective - Data Exploration to build domain knowledge

```
In [42]: import IPython
         url = 'http://[redacted]:6006/'
         iframe = '<iframe src=' + url + ' width=1000 height=500></iframe>'
         IPython.display.HTML(iframe)
```

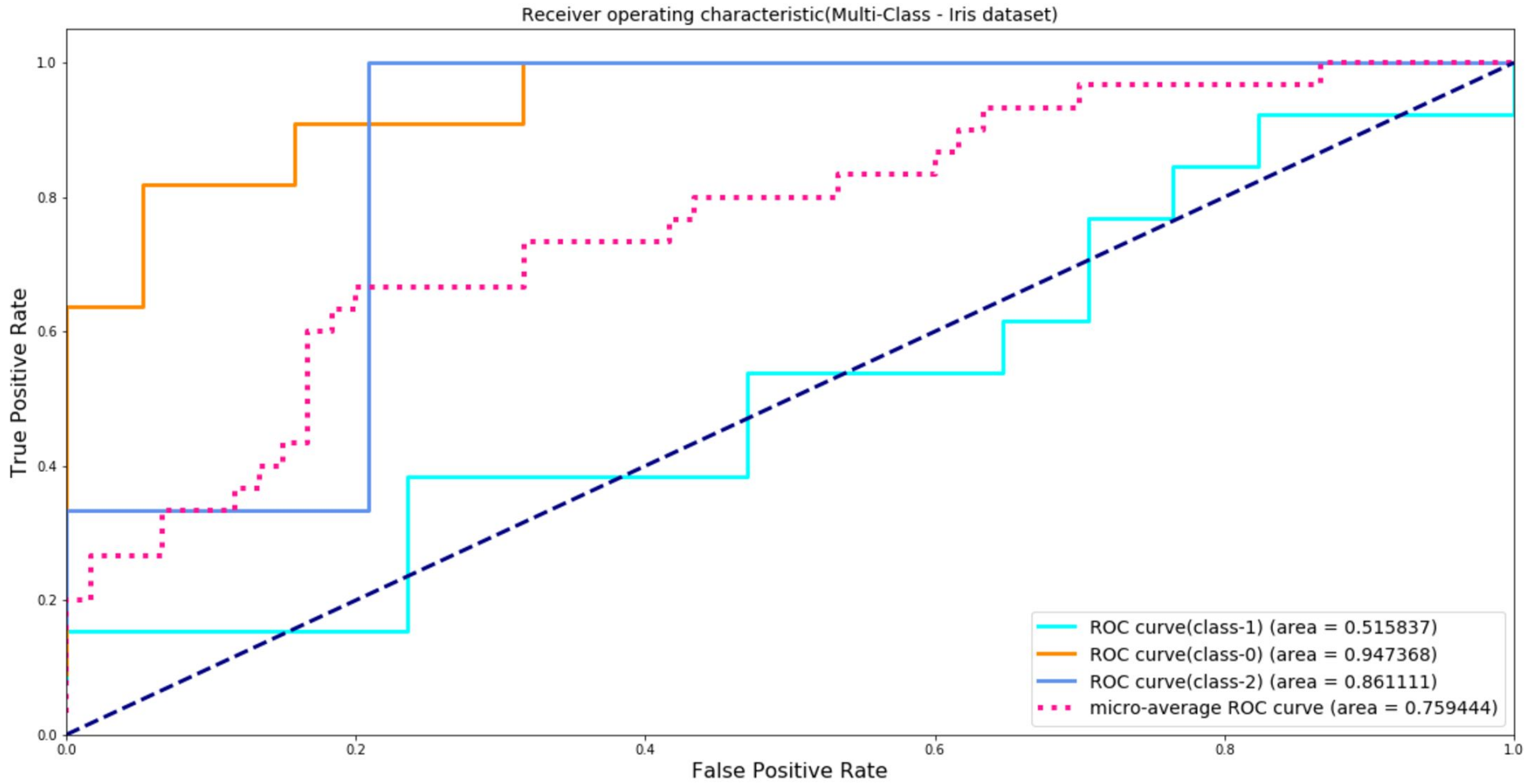
Out[42]:



DEFINE INTERPRETATION

- Definition is subjective - overlaps with Model Evaluation

```
In [77]: plot_roc_multiclass_classifier(class_label_dict=label_dict, line_width=2)
```



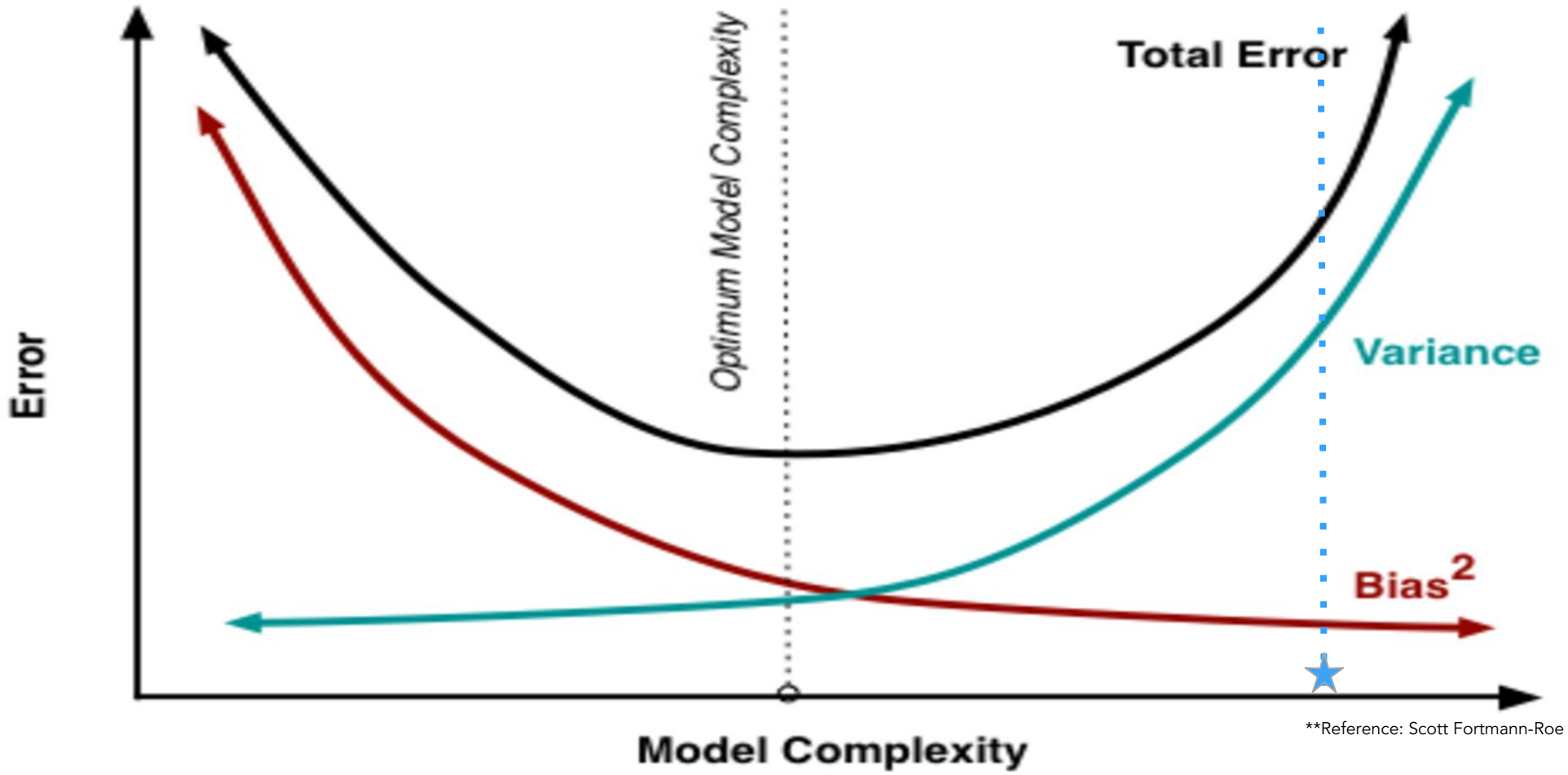
WHAT IS MODEL INTERPRETATION?

- Model interpretation is an extension of Model Evaluation to help us understand machine learning/statistical modeling behavior better if possible in a human interpretable way

- With model interpretation, one should be able to answer the following questions:
  - **Why** did the model behave in a certain way? What are the relevant variables driving a model's outcome - e.g. Customer's Lifetime Value, Fraud detection, Image Classification, Spam Detection ?
  - **What** other information can a model provide to avoid prediction errors ? What was the reason for a false positive ?
  - **How** can we trust the predictions of a "black box" model ? Is the predictive model biased ?

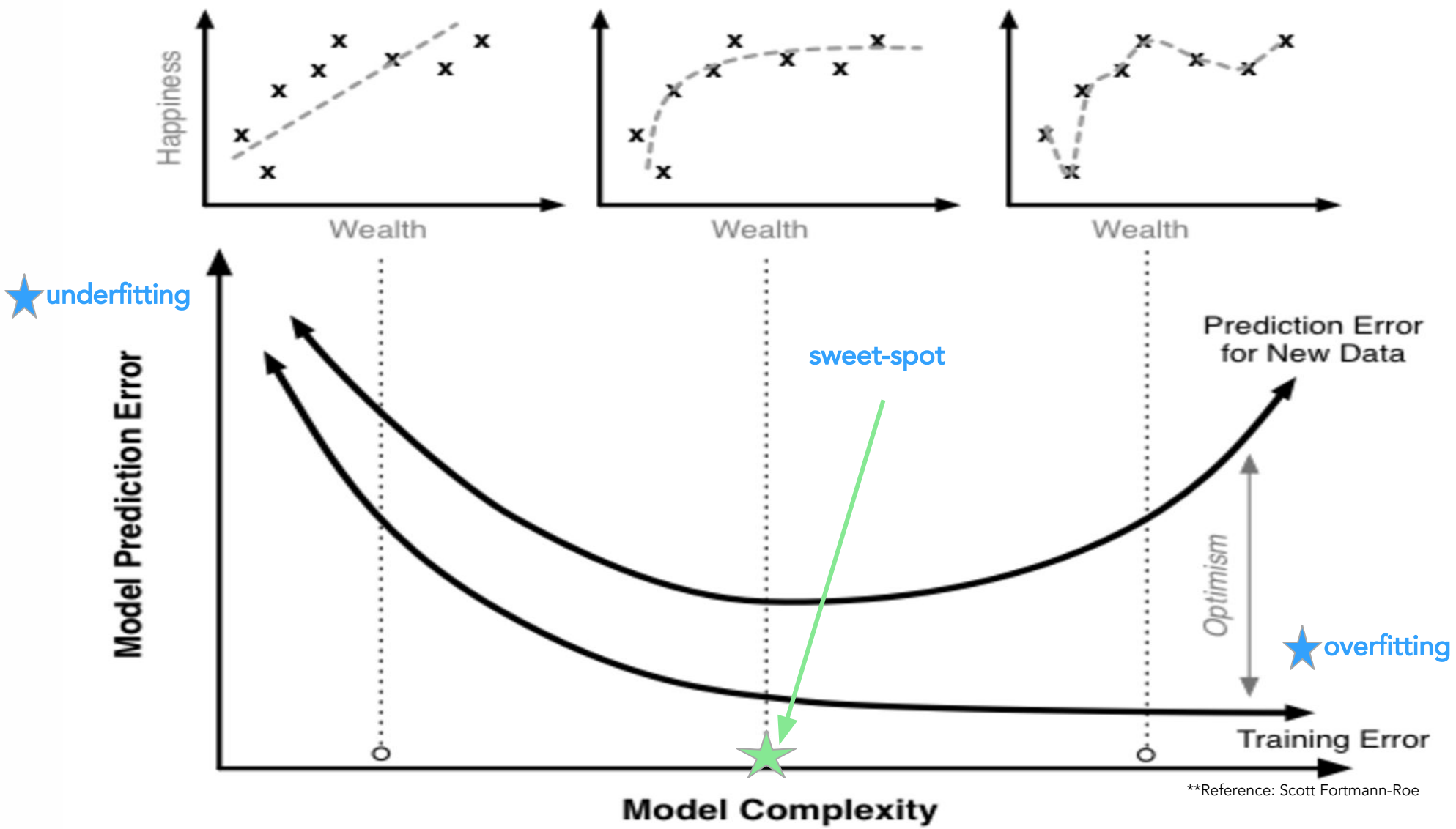
- Focus: is in-regards to Supervised learning problems

ACCURACY VS MODEL COMPLEXITY



Error(x) = **Bias**^{2} + **Variance** + Irreducible Error

Predictive Optimism



**Reference: Scott Fortmann-Roe

WHY DO WE NEED MODEL INTERPRETATION?

- Helps in **exploring and discovering latent or hidden feature interactions** (useful for feature engineering/selection)
- Helps in understanding **model variability** as the environment changes (once the model is operationalized and is functional in a non-stationary environment)
- Helps in **model comparison**
- Helps an analyst or data scientist build **domain knowledge** about a particular use case by providing an understanding of interactions

WHY MODEL INTERPRETATION?

- Brings **transparency** to decision making to enable **trust**
  - Fair Credit Reporting Act (FCRA) U.S. Code § 1681

SUBCHAPTER III—CREDIT REPORTING AGENCIES

§ 1681. Congressional findings and statement of purpose

(a) Accuracy and fairness of credit reporting

The Congress makes the following findings:

(1) The banking system is dependent upon fair and accurate credit reporting. Inaccurate credit reports directly impair the efficiency of the banking system, and unfair credit reporting methods undermine the public confidence which is essential to the continued functioning of the banking system.

(2) An elaborate mechanism has been developed for investigating and evaluating the credit worthiness, credit standing, credit capacity, character, and general reputation of consumers.

(3) Consumer reporting agencies have assumed a vital role in assembling and evaluating consumer credit and other information on consumers.

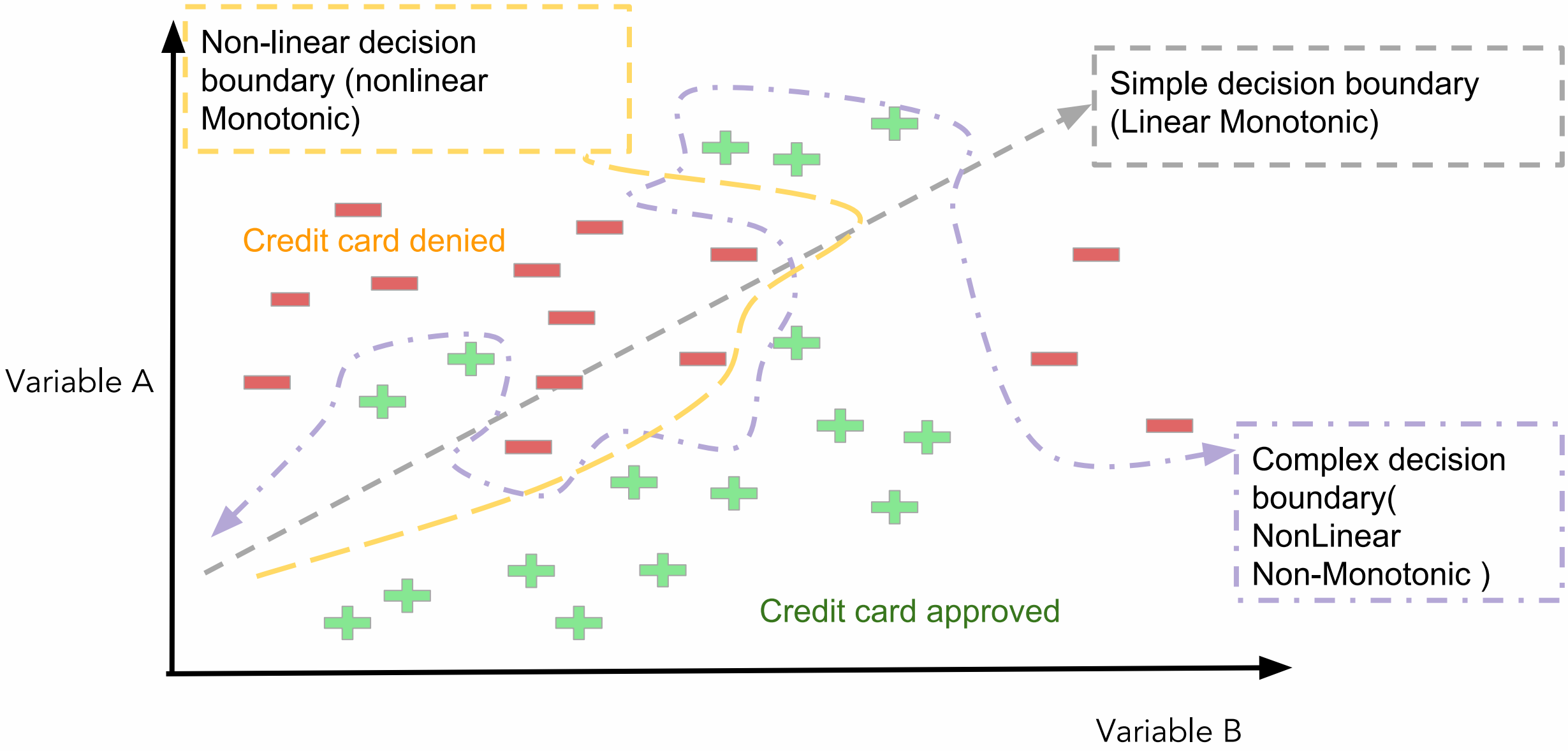
(4) There is a need to insure that consumer reporting agencies exercise their grave responsibilities with fairness, impartiality, and a respect for the consumer's right to privacy.

Mandate by U.S. government on **Fair** and **Accurate Credit** reporting. Predictive models should not be discriminative (**biased**) toward any group.

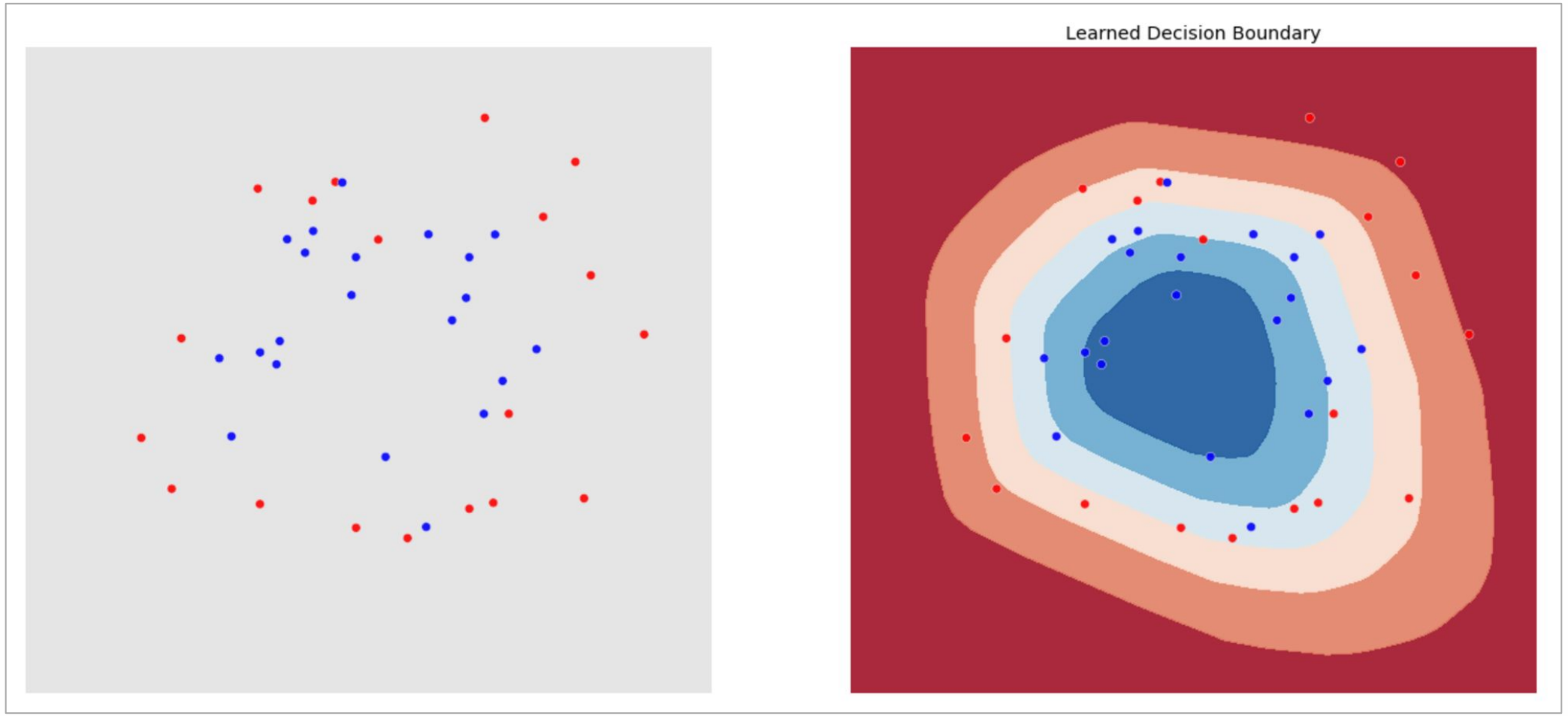
Are all predictive models interpretable?

Does an interpretable model always provides the best model ?

PERFORMANCE VS. INTERPRETABILITY



HOW ABOUT A MORE DIFFICULT RELATIONSHIP?



Data

Learned decision boundaries

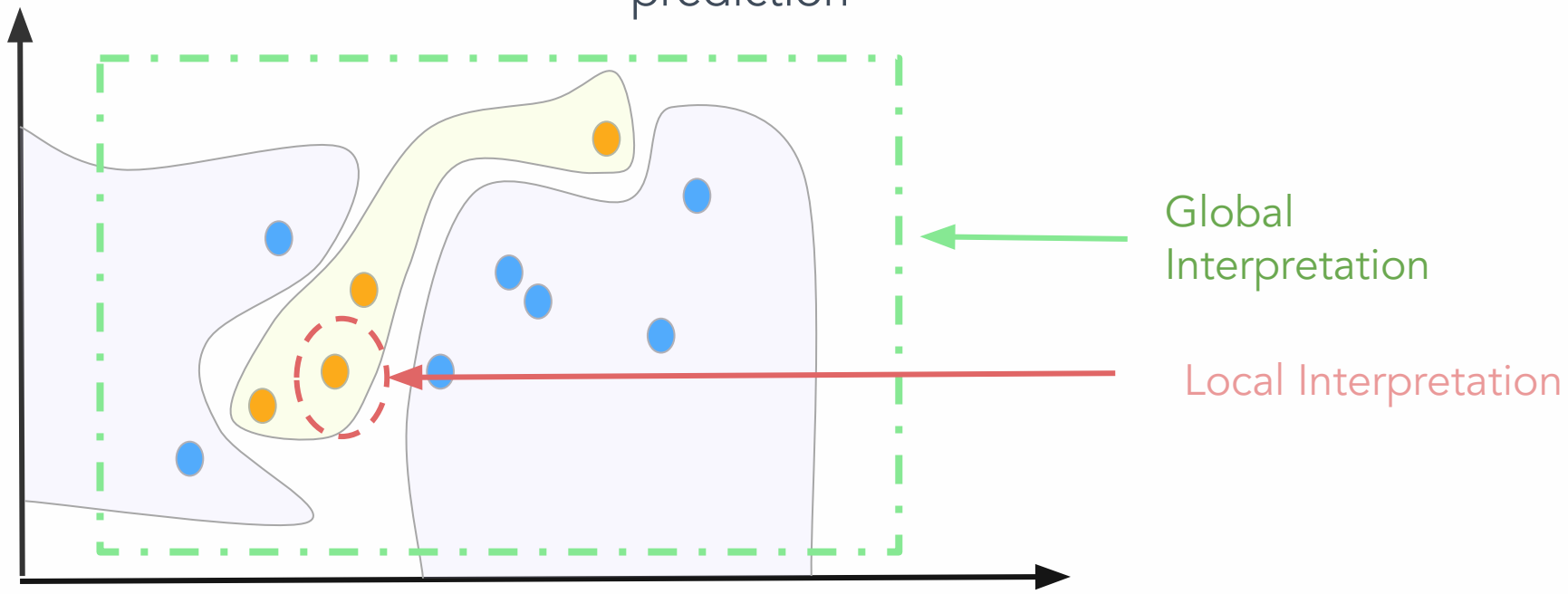
SCOPE OF INTERPRETATION

Global Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables based on the complete dataset

Local Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables wrt to a single prediction



GLOBAL INTERPRETATION

- **Relative Importance of Predictor Variable to evaluate Estimator's behavior**
  - Model-specific Feature importance - e.g.
    - Linear Model (*based on the absolute value of t-statistics*)
    - Random Forest (based *permutation importance or Gini importance*)
    - Recursive Feature Elimination(RFE) - recursively prune least important features
  - **Model Independent Feature Importance -** this will be our focus for today's discussion
    - observing entropy of predictive performance based on random perturbation of feature set
    - observing entropy of model specific scoring metric
      - Classification: f1-score, precision/recall
      - Regression: mean squared error
- Usefulness
  - Helps in identifying important covariates contributing to target prediction enabling better interpretability
  - Might help in improving accuracy and computation time by eliminating redundant or unimportant features

GLOBAL INTERPRETATION

- **Partial Dependence Plot (PDP)**
  - Helps in understanding the average partial dependence of the target function $f(Y|X_s)$ on subset of features by marginalizing over rest of the features (*complement set of features*)
  - Works well with input variable subset with low cardinality ($n \leq 2$)
  - e.g. PDPs on california housing data

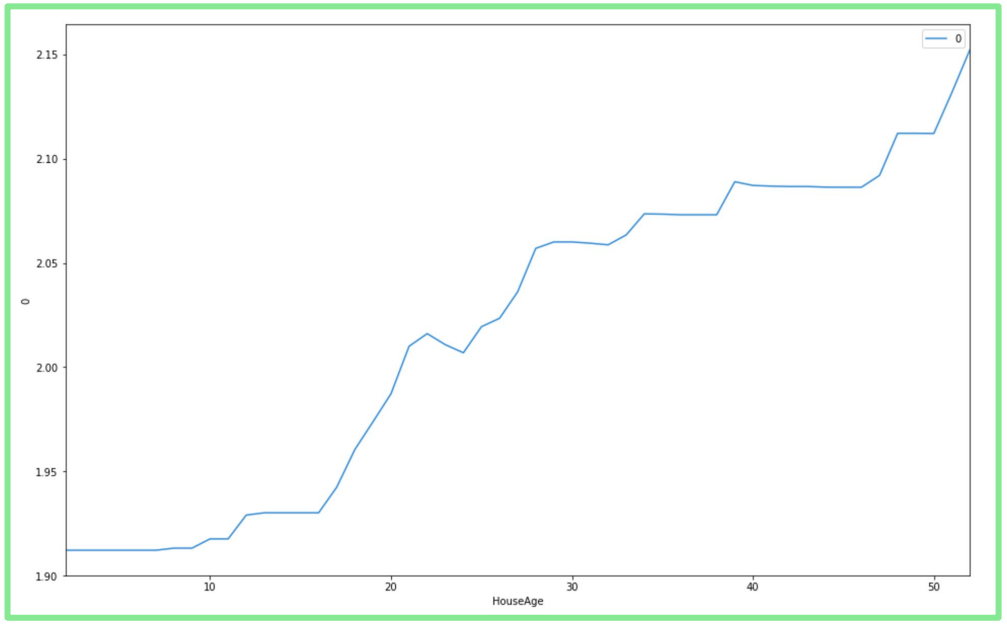


Fig A: HouseAge vs Avg. House Value

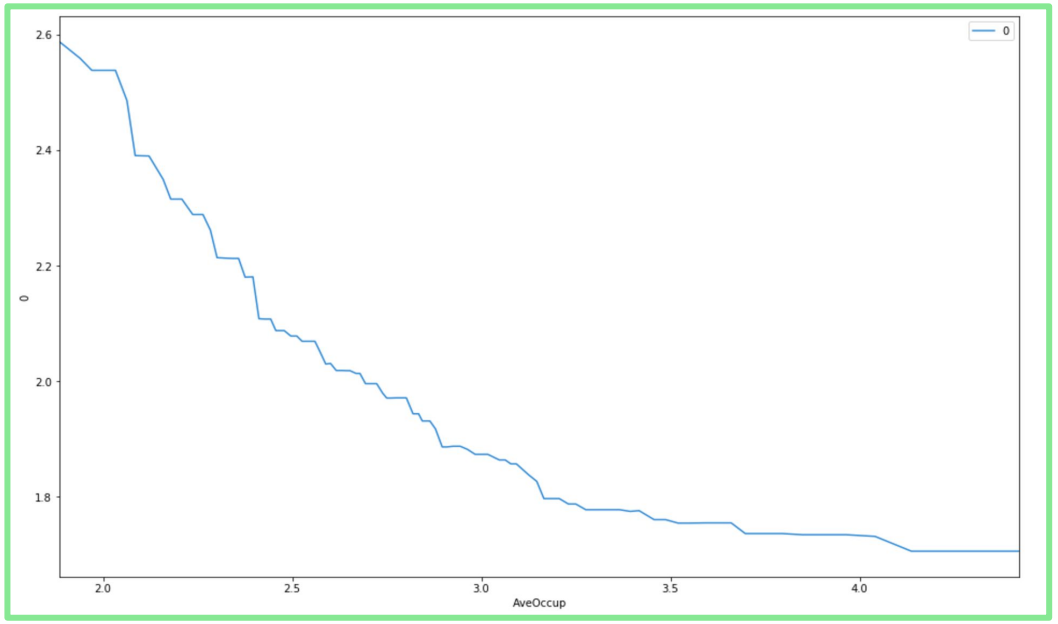


Fig B: Avg. occupants vs Avg. House Value

PDP continues ...

- Helps in understanding interaction impact of two independent features in a low dimensional space visually

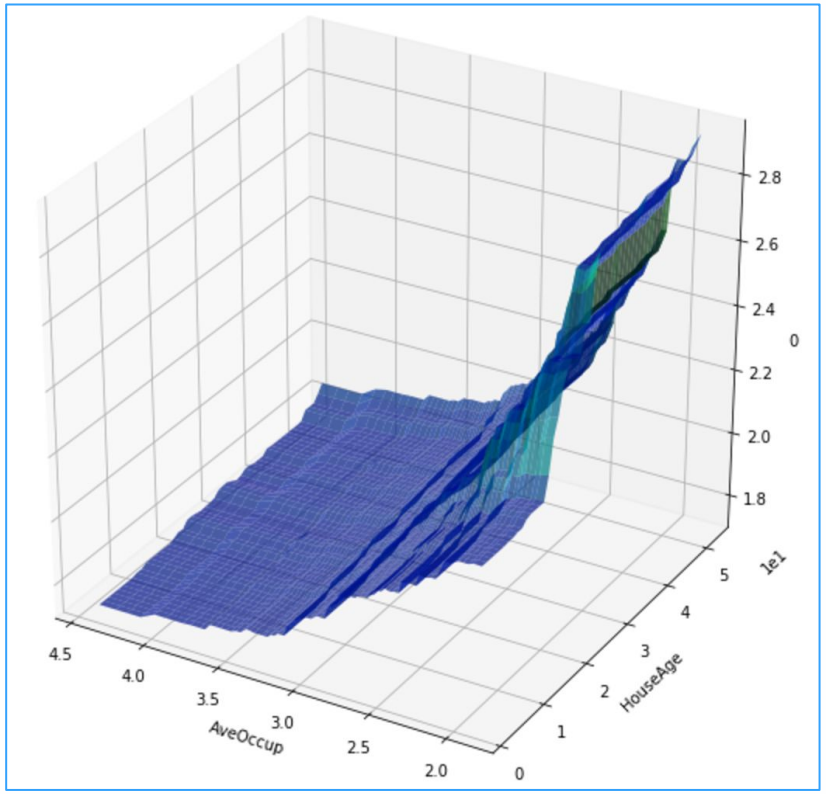
  - F on X_s where $X = X_s \cup X_c$ is

$$f_S = \mathbb{E}_{\boldsymbol{x}_C}\left[f(\boldsymbol{x}_S, \boldsymbol{x}_C)\right] = \int f(\boldsymbol{x}_S, \boldsymbol{x}_C)\mathrm{dP}\left(\boldsymbol{x}_C\right)$$

  - Average value of f() when Xs is fixed and Xc is varied over its marginal distribution

  - Integrated over values of X_c

$$\hat{f}_S = \frac{1}{N}\sum_{i=1}^{N}\hat{f}(\boldsymbol{x}_S, \boldsymbol{x}_{Ci})$$



p(HouseAge, Avg. Occupants per household) vs Avg. House Value: One can observe that once the avg. occupancy > 2, houseAge does not seem to have much of an effect on the avg. house value

PDP continues ...

- Might incorrectly articulate the interaction between predictive variable and target variable

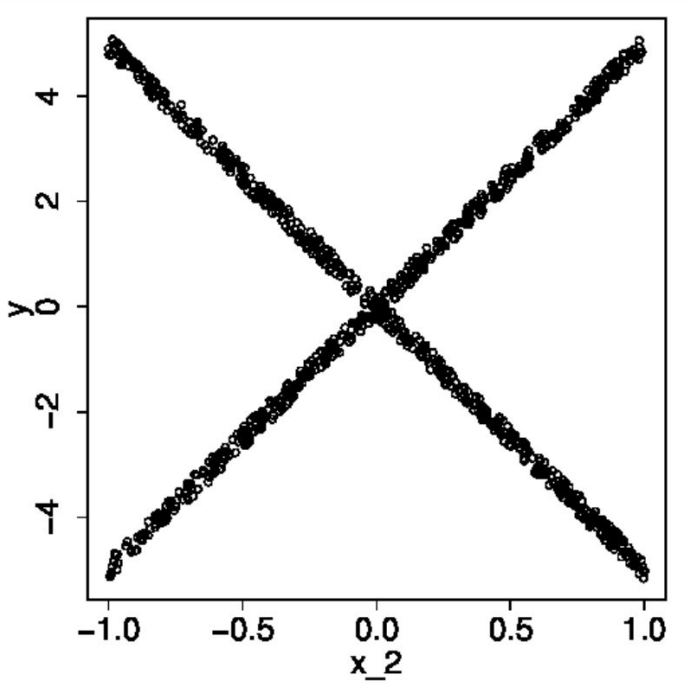


Fig A: Scatter plot

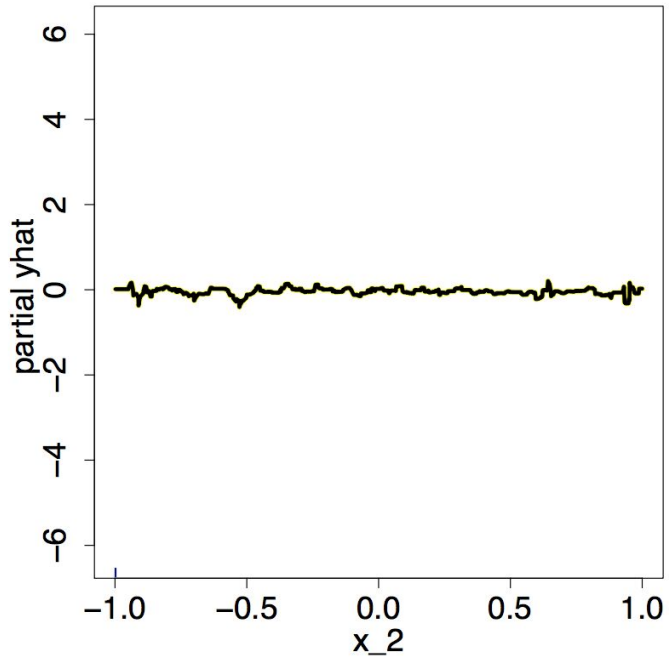


Fig B: PDP

- In *Fig A*, we plot a variable *x_2* vs Y over say a sample of 500 points
- In *Fig B*, we plot a PDP of a model for predictor variable *x_2* vs Y_hat.
- **Observation:** PDP suggests that on average x_2 has no influence on target variable

**Reference: Alex Goldstein et al.

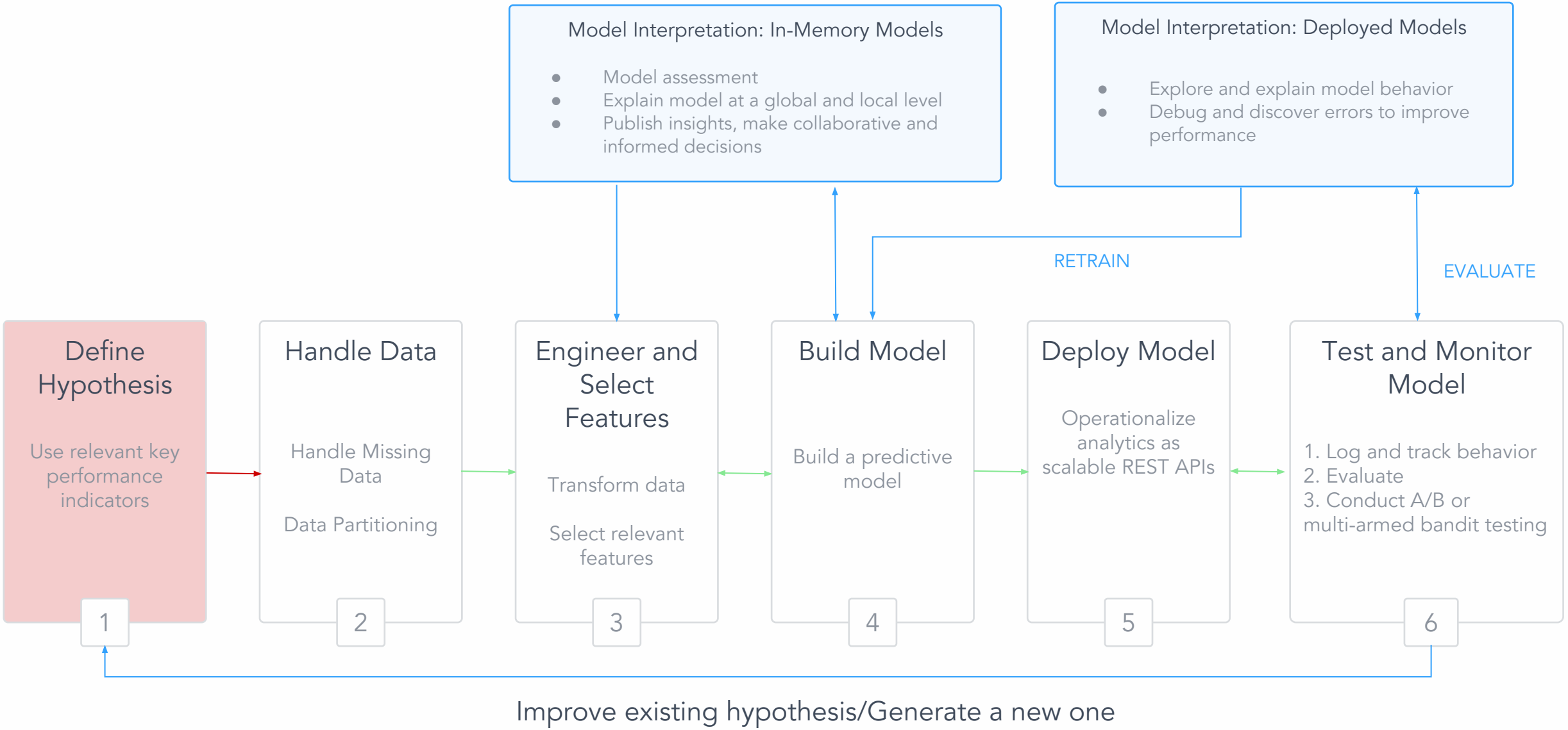
LOCAL INTERPRETATION

- Ability to **inspect and evaluate** individual prediction in human interpretable format with the help of surrogate models faithfully

$$\xi(x) = \text{argmin}_{g \in G} \; \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$

  - **ὲ :** model explanation function
  - $\mathcal{L}$: measure of fidelity
  - f : is the base model estimator
  - g ⊂ G : a set of interpretable models [*Linear Models, Decision Trees*]
  - $\prod_x$: proximity measure to define locality around an individual point
  - Ω : to regularize complexity e.g. depth of the tree, learning rate, non-zero weights for linear models

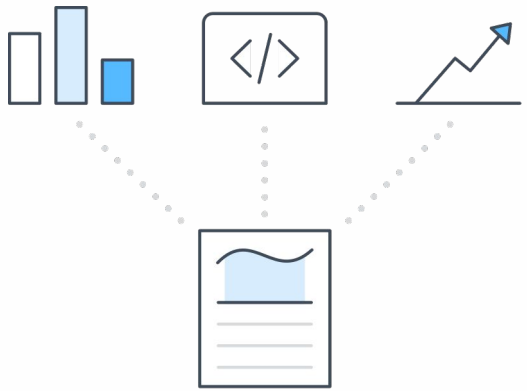
UNDERSTANDING ANALYTICAL WORKFLOW ?



HOW DO WE SOLVE THIS PROBLEM?

- Problems:
  - Data scientists are choosing easy-to-interpret models like simple linear models or decision trees over high-performing neural networks or ensembles, effectively sacrificing accuracy for interpretability
  - Community is struggling to keep pace with new algorithms and frameworks (sklearn, R packages, H20.ai)

- Possible Solution: **What if** there was an interpretation library that…
  - Is model agnostic
  - Provides human-interpretable explanation
  - Is framework agnostic (scikit-learn, H20.ai, Vowpal Wabbit)
  - Is language agnostic (R, Python)
  - Allows one to interpret third-party models (Algorithmia, indico)
  - Supports interpretation both during modeling build process and post deployment



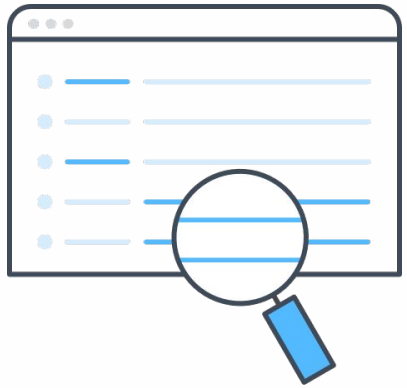
INTRODUCING ...

DATASCIENCE.COM

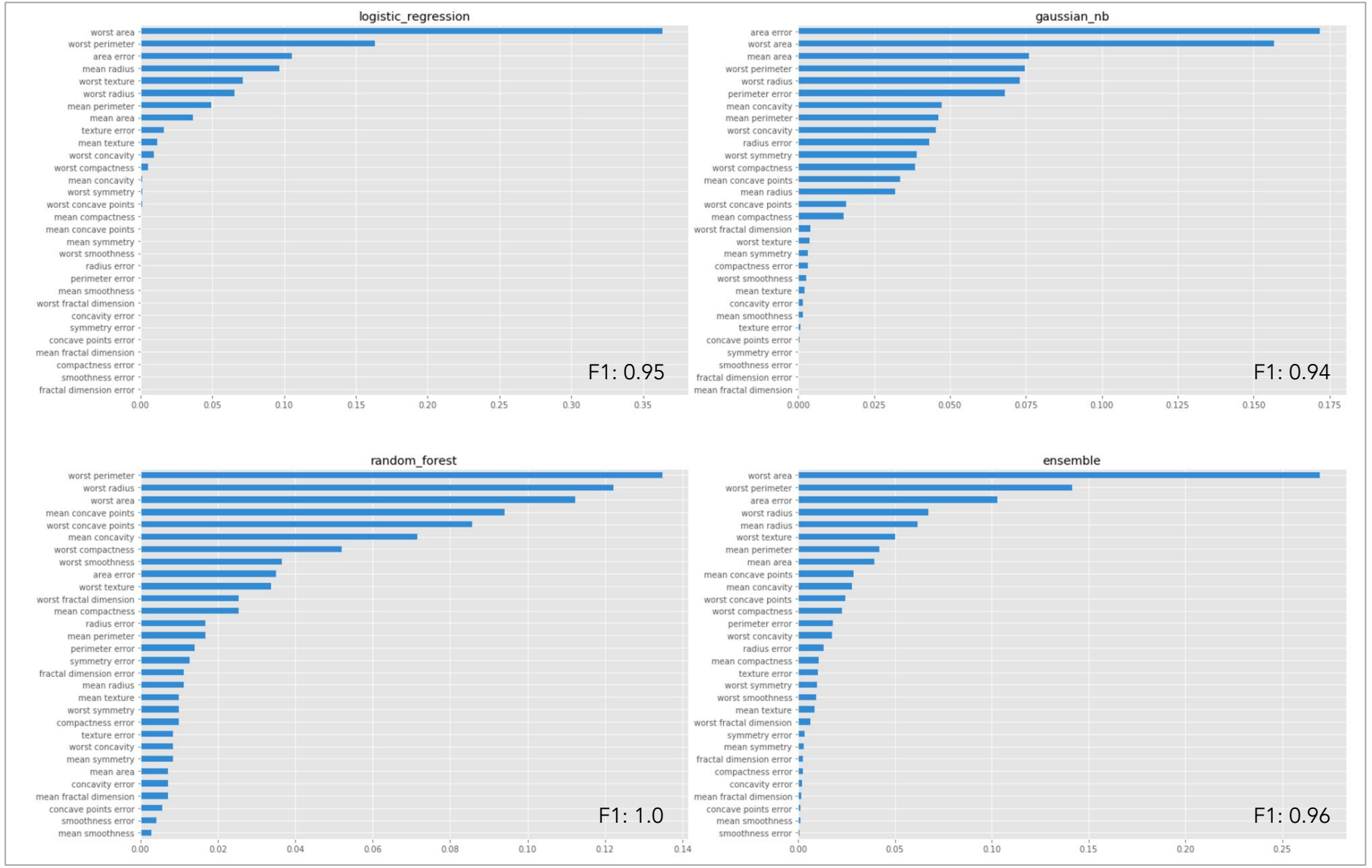
SKATER

WHAT IS SKATER?

- Python library designed to demystify the inner workings of black-box models
- Uses a number of techniques for model interpretation to explain the relationships between input data and desired output, both globally and locally
- One can interpret models both before and after they are operationalized

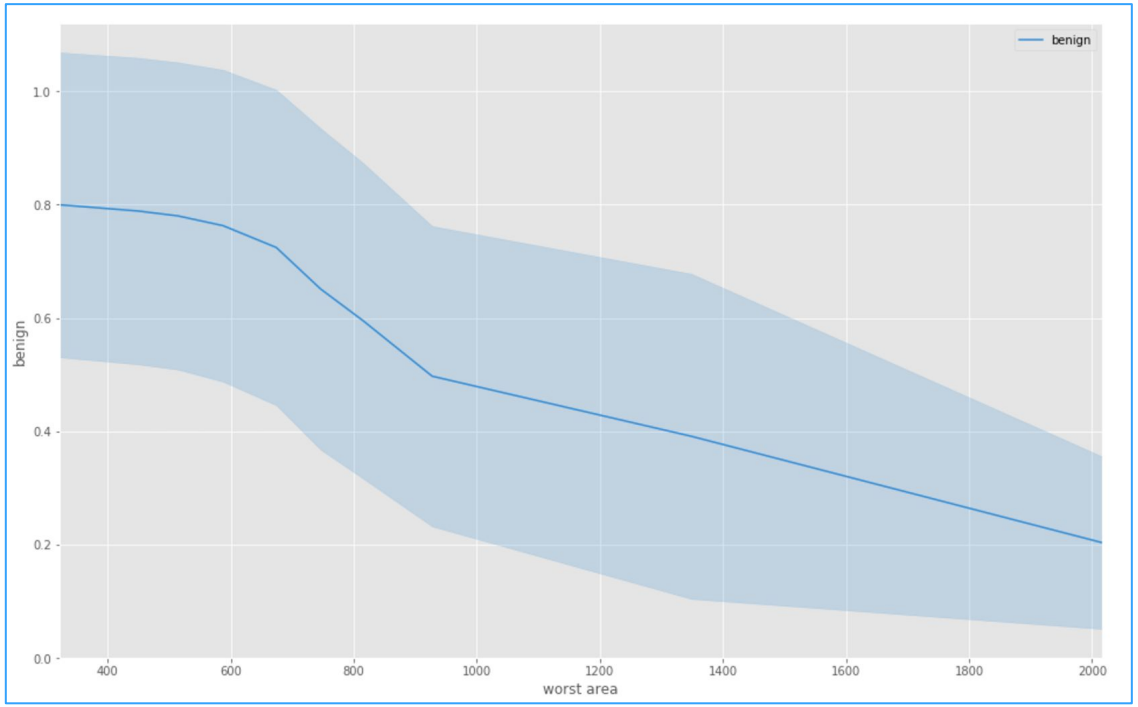


SKATER USES - Model-agnostic Variable Importance for global interpretation

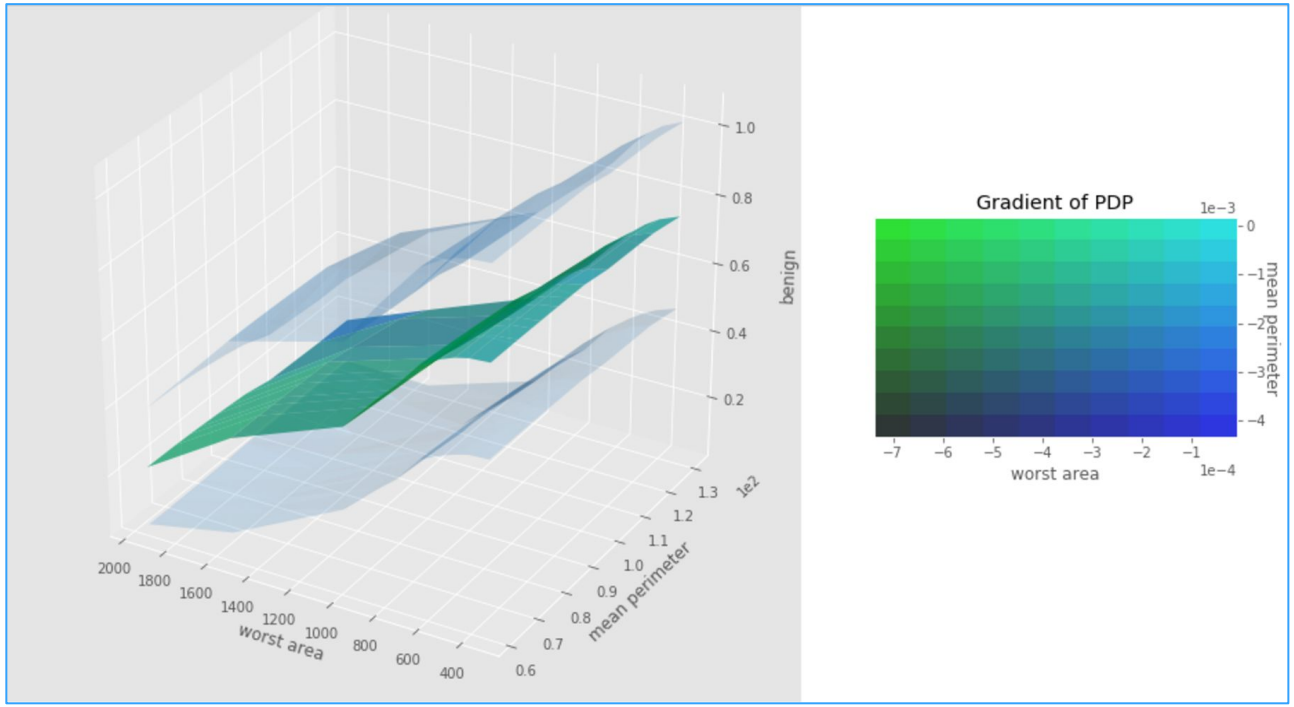


SKATER USES - Partial dependence plots for global interpretation

- A visualization technique that can be used to understand and estimate the dependence of the joint interaction of the subset of input variables to the model's response function



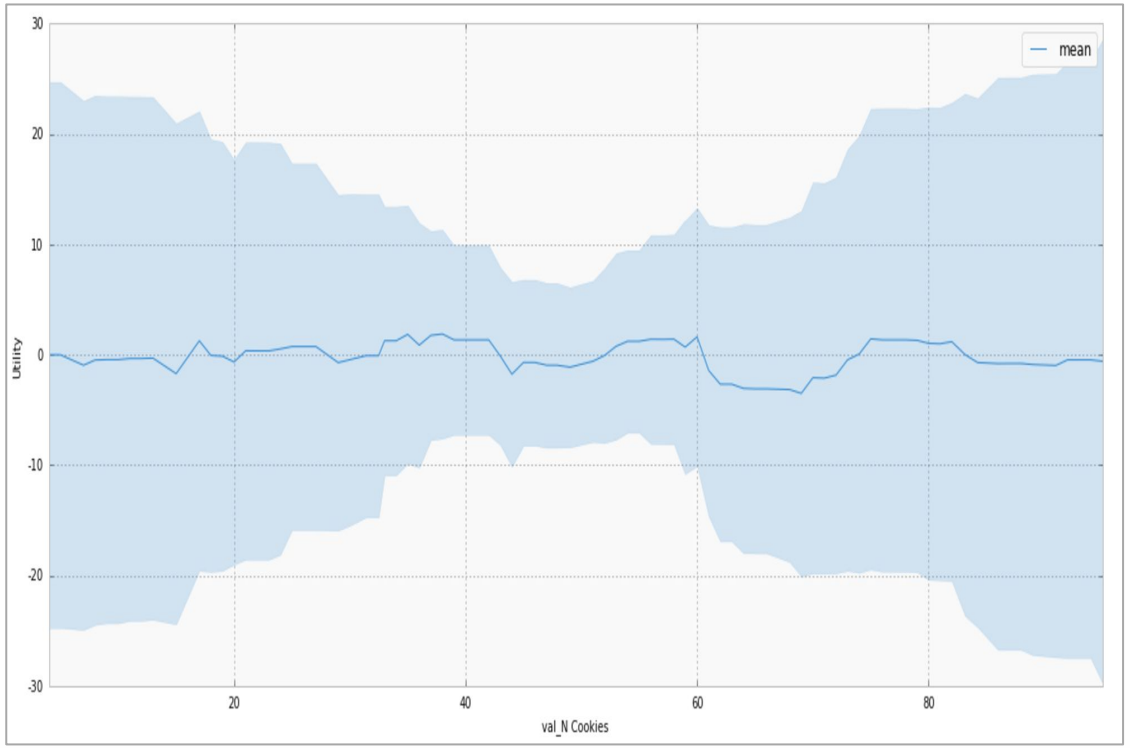
a. One-way interaction



b. Two-way interaction

PDPs continued

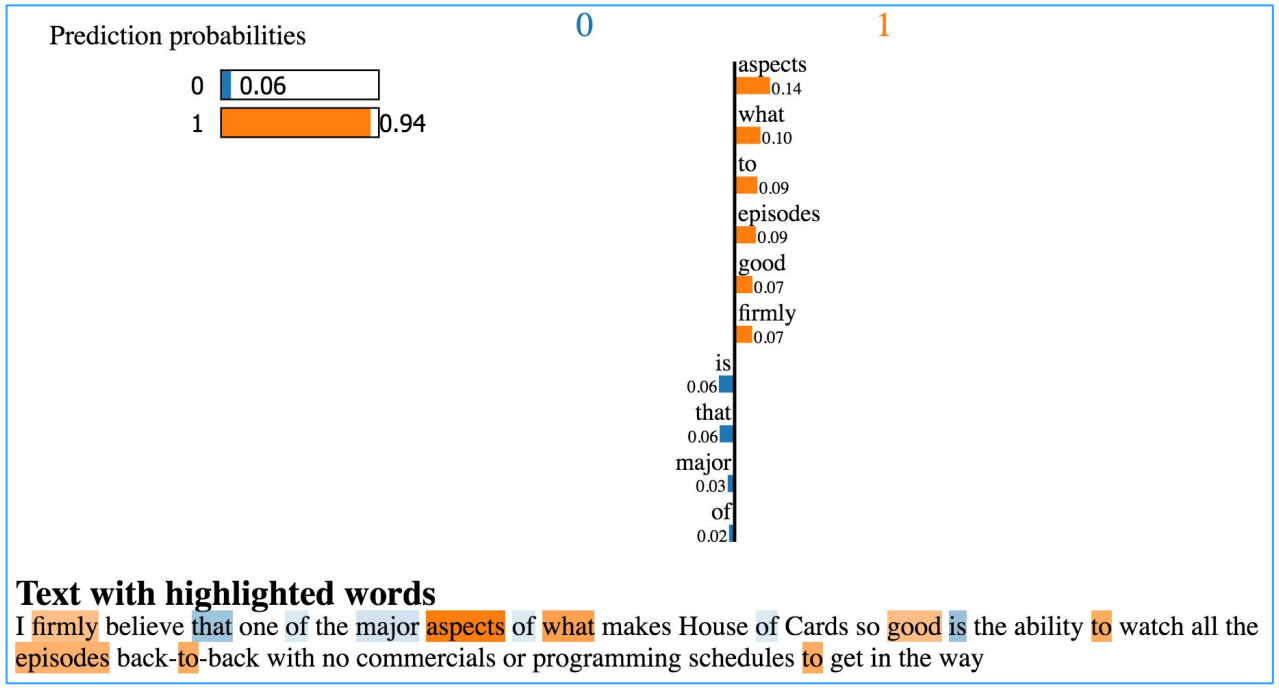
- PDPs suffers from cancellation effect because of averaging
- Variance effect helps in highlighting this cancellation



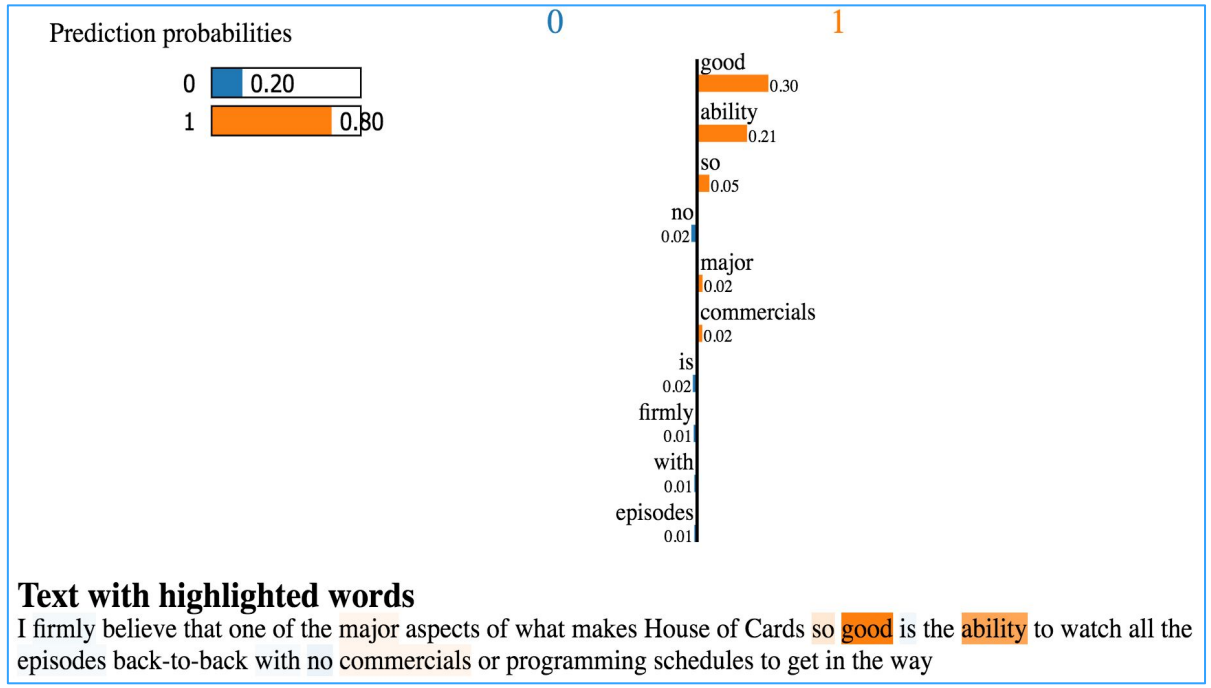
One-way interaction with variance

SKATER USES - Local Interpretable Model-Agnostic Explanations (LIME) for local interpretation

- A novel technique developed by Marco, Sameer and Carlos to explain the behavior of any classifier or regressor in an human interpretable way using linear surrogate models to approximate around the vicinity of a single prediction



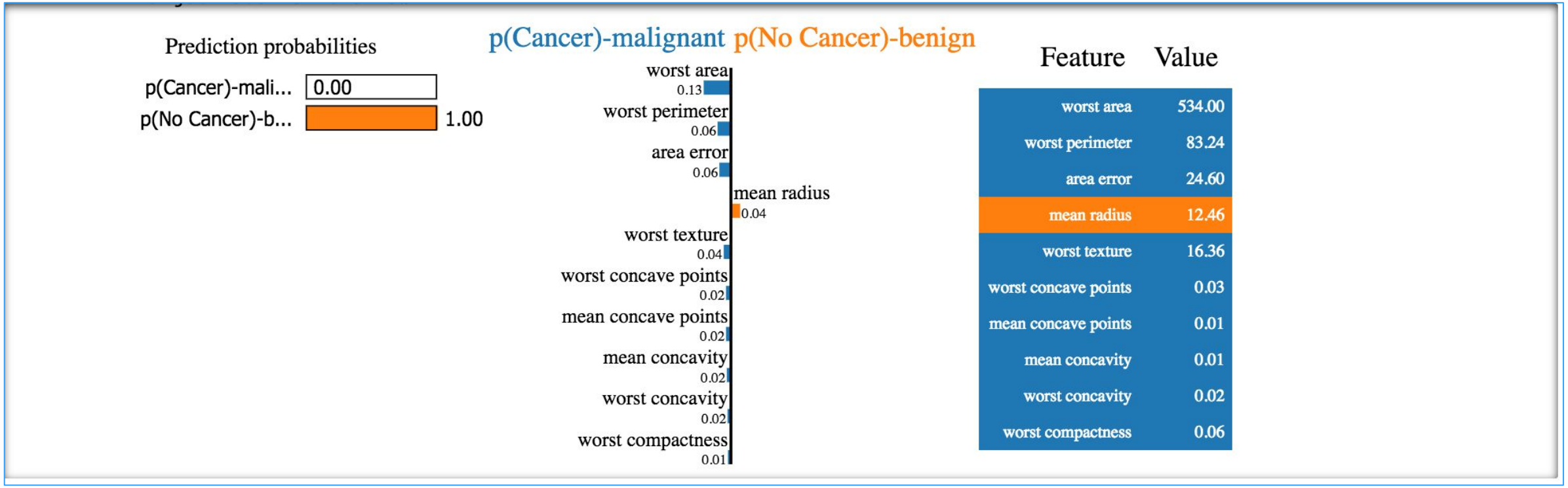
Deployed Model - indico.io



Deployed Model - algorithmia

LIME continues ...

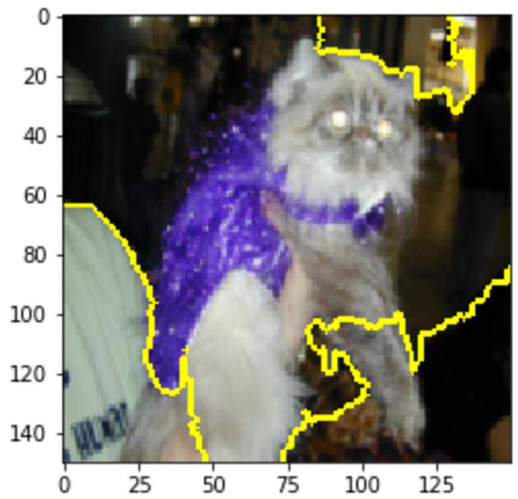
- Regression
  - Gold Label : No Cancer
  - Predicted(*y_hat*): No Cancer



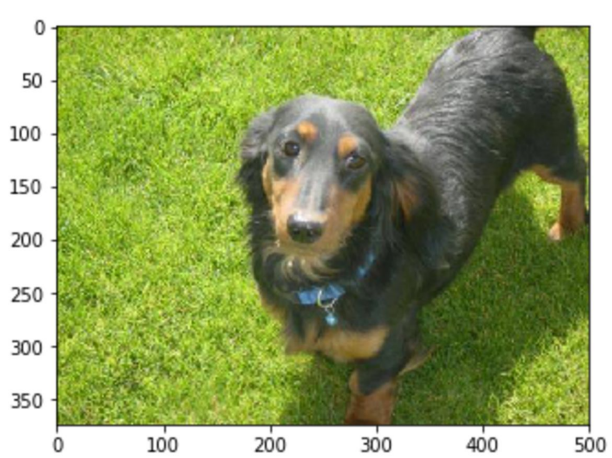
SKATER USES

- LIME for image interpretability (*experimental*)

highlight the feature boundaries

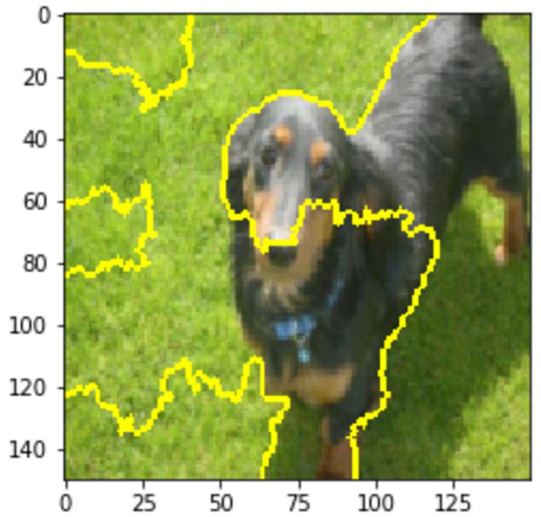


Will this be classified correctly ?



highlight the feature boundaries

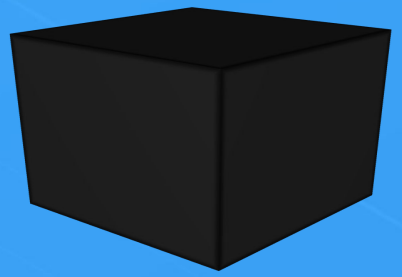
Got classified as a “dog” but doesn’t seem convincing



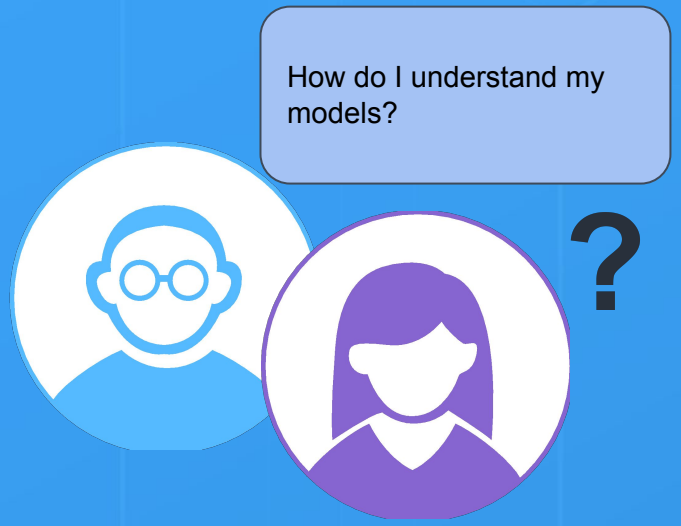
- Which features ?
- Was it the green background ?

WITHOUT INTERPRETATION ...

$f(Y|X)$

Data →  → Evaluate

Black box model



WITH SKATER ...

Data →  → Evaluate

Unboxed model

- Partial dependence plot
- Relative variable importance
- Local Interpretable Model Explanation (LIME)
- More coming soon ...

R or Python model (linear, nonlinear, ensemble, neural networks)

Scikit-learn, caret and rpart packages for CRAN

H20.ai, Algorithmia, etc.

COMING SOON ...

- Predictions as conditional statements: An interpretable model, with series of decision rules
  - Given a dataset, mine a set of antecedents
  - Possible to observe and learn a manageable set of rules and their orders

```
The rules list is :
If      {Pclass=3,Sex_Encoded=0} (rule[67]) then positive probability = 0.43023256
else if {Sex_Encoded=0} (rule[81]) then positive probability = 0.95081967
else if {Pclass=3} (rule[78]) then positive probability = 0.16746411
else if {Pclass=2,Parch=0} (rule[44]) then positive probability = 0.07246377
else  (default rule)  then positive probability = 0.40000000
```

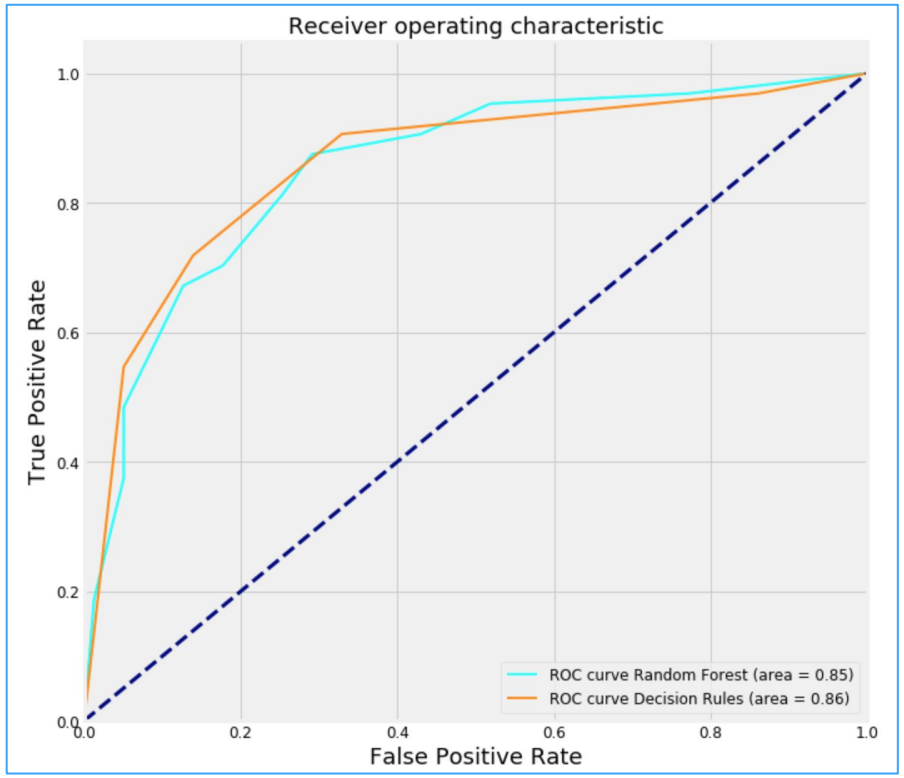


Fig: Series of rules capturing the p(Survival) on titanic dataset

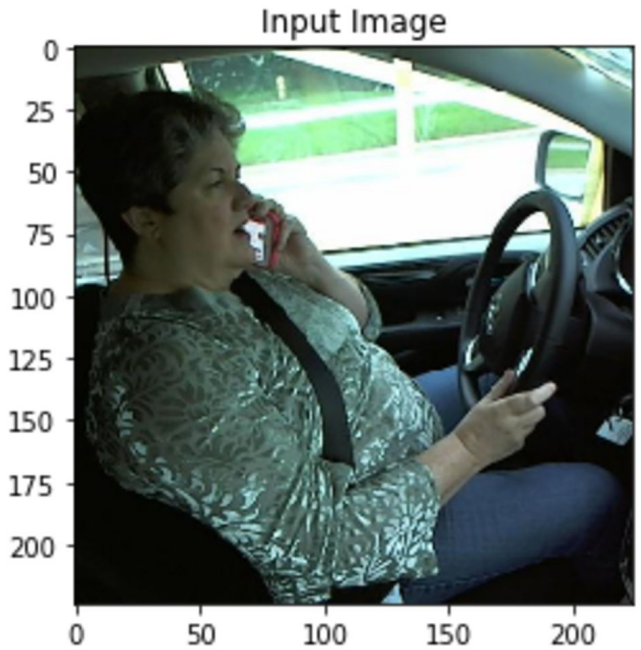
JUPYTER'S INTERACTIVENESS

- Human in the loop is very useful for Model Evaluation
- Being able to do it in a convenient way, increases efficiency
- Interactiveness,
  - Jupyter Widgets: - UI controls to inspect code and data interactively
  - Enables collaboration and sharing:
    - Widgets can be serialized and embedded in
      - html web pages,
      - Sphinx style documents
      - html-converted notebooks on nbviewer

  - Jupyter dashboards
    - is a dashboard layout extension
    - helpful in organizing notebook outputs - text, images, plots, animations in report like layout

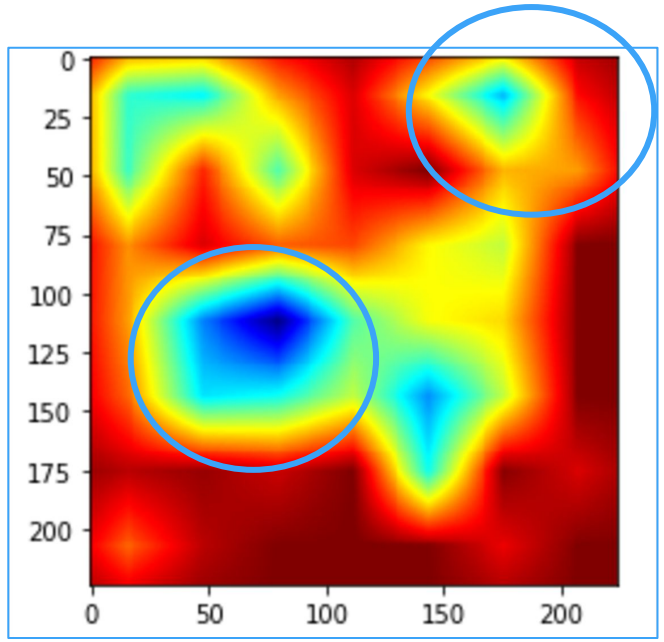
A QUICK GLIMPSE INTO THE FUTURE

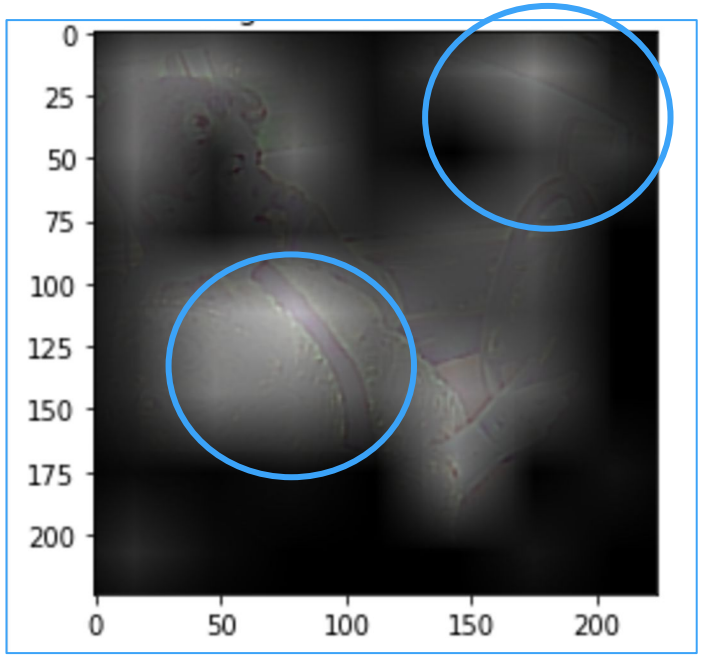
Visual QnA: Is the person driving the car safely ?



Top 5 predictions:

1. seat belt = 0.75
2. limousine = 0.051
3. golf cart = 0.017
4. minivan = 0.015
5. car mirror = 0.015





SPECIAL THANKS

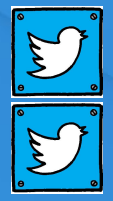
- Special thanks to Aaron Kramer(*one of the original authors of Skater*), Ben Van Dyke and rest of the datascience.com teammates for helping out with Skater

- Thank you to IDEAS for providing us the opportunity to share our thoughts with a wider community

Q&A

info@datascience.com

pramit@datascience.com



@DataScienceInc

@MaverickPramit

Help wanted(Skater): https://tinyurl.com/yd6tnc7l

Appendix

References:

- A. Weller, "Challenges for Transparency": https://arxiv.org/abs/1708.01870
- Max Kuhn, Variable Importance Using The caret pkg: http://ftp.uni-bayreuth.de/math/statlib/R/CRAN/doc/vignettes/caret/caretVarImp.pdf
- Friedman' 01, Greedy Function Approximation: A gradient boosting machine: https://statweb.stanford.edu/~jhf/ftp/trebst.pdf
- Recursive Feature Elimination: https://arxiv.org/pdf/1310.5726.pdf
- LIME: https://arxiv.org/pdf/1602.04938v1.pdf
- Nothing Else Matters: https://arxiv.org/pdf/1611.05817v1.pdf
- Peeking Inside the Black Box: https://arxiv.org/abs/1309.6392