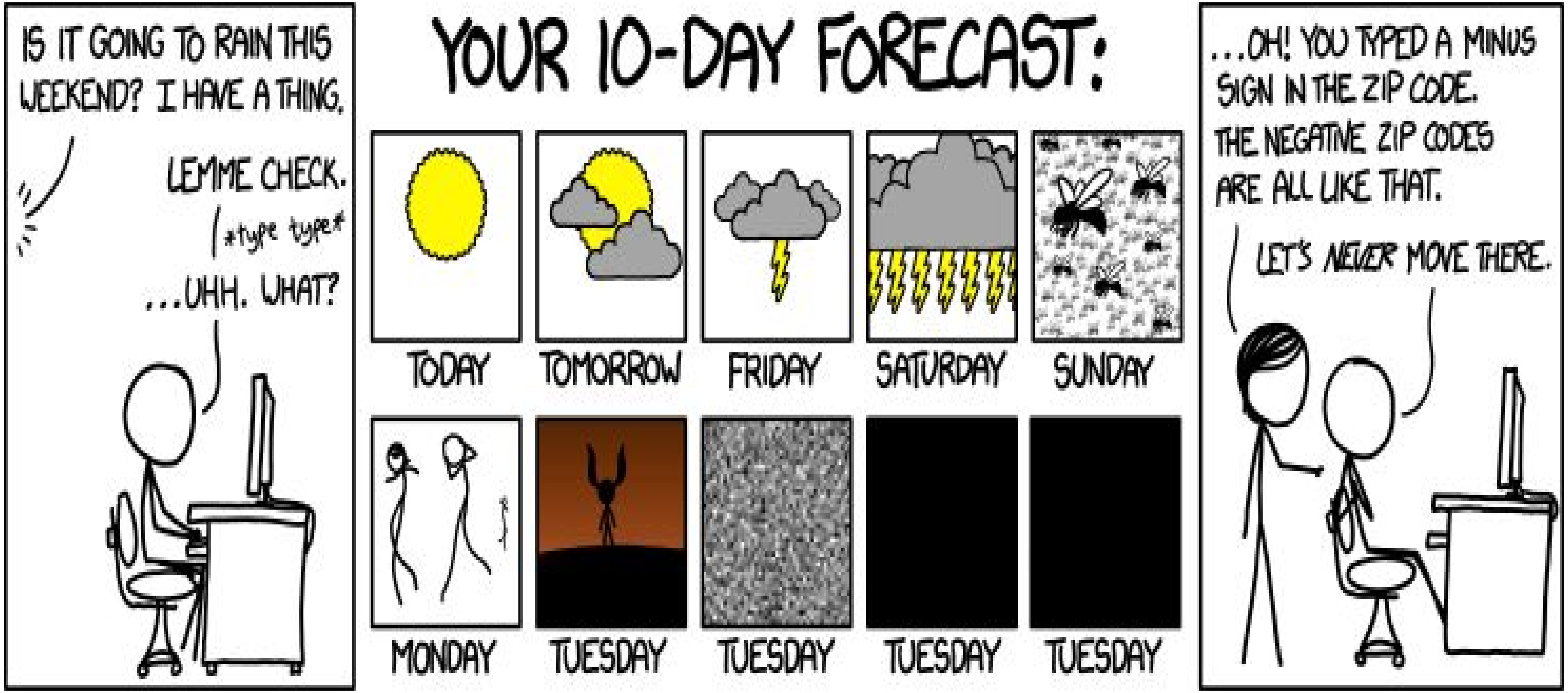
DATASCIENCE

Learning to Learn Model Behavior: How to Explain Predictions While Being Model Agnostic

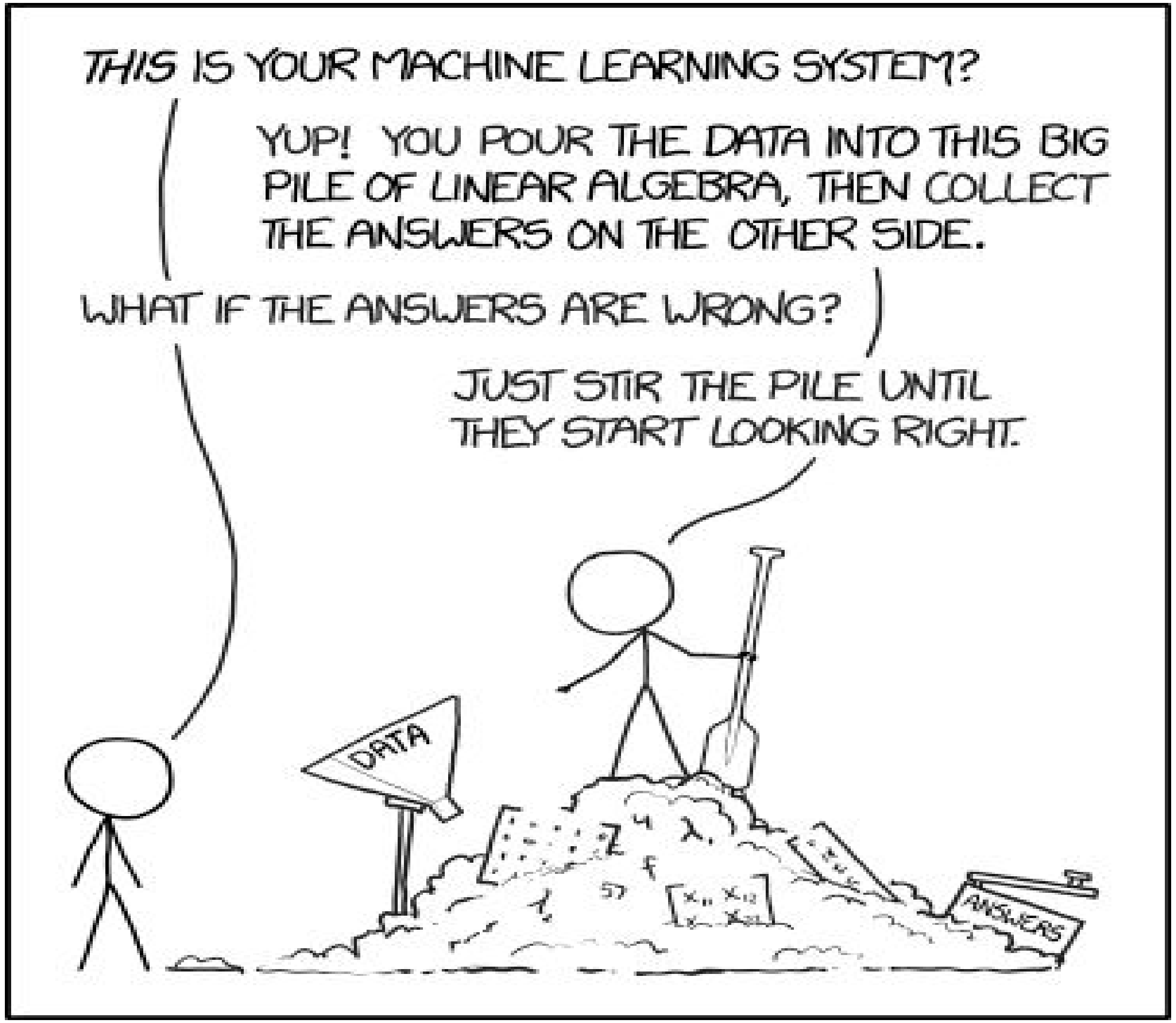
June 23, 2017

PREDICTIVE MODELING: FUN OR MISERY?

PREDICTIONS OFTEN GO WRONG



WHEN AN ERROR OCCURS


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

ABOUT US

Pramit Choudhary

Pramit Choudhary is a lead data scientist at DataScience.com. His focus is on effective ways of optimizing and applying classical (Machine Learning) and Bayesian design strategy to solve real-world problems.

Aaron Kramer

Aaron is senior data scientist at DataScience.com. With experience leading a wide variety of corporate research problems, he develops tools, algorithms, and educational materials for modern applied data science.

TEAMMATES @DATASCIENCE.COM

Ruslana

Dave

Harry

John

JR

Guo

Andrea

Ben

Tuck

AGENDA

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

DATASCIENCE.COM

WHAT IS MODEL INTERPRETATION?

- Definition is subjective

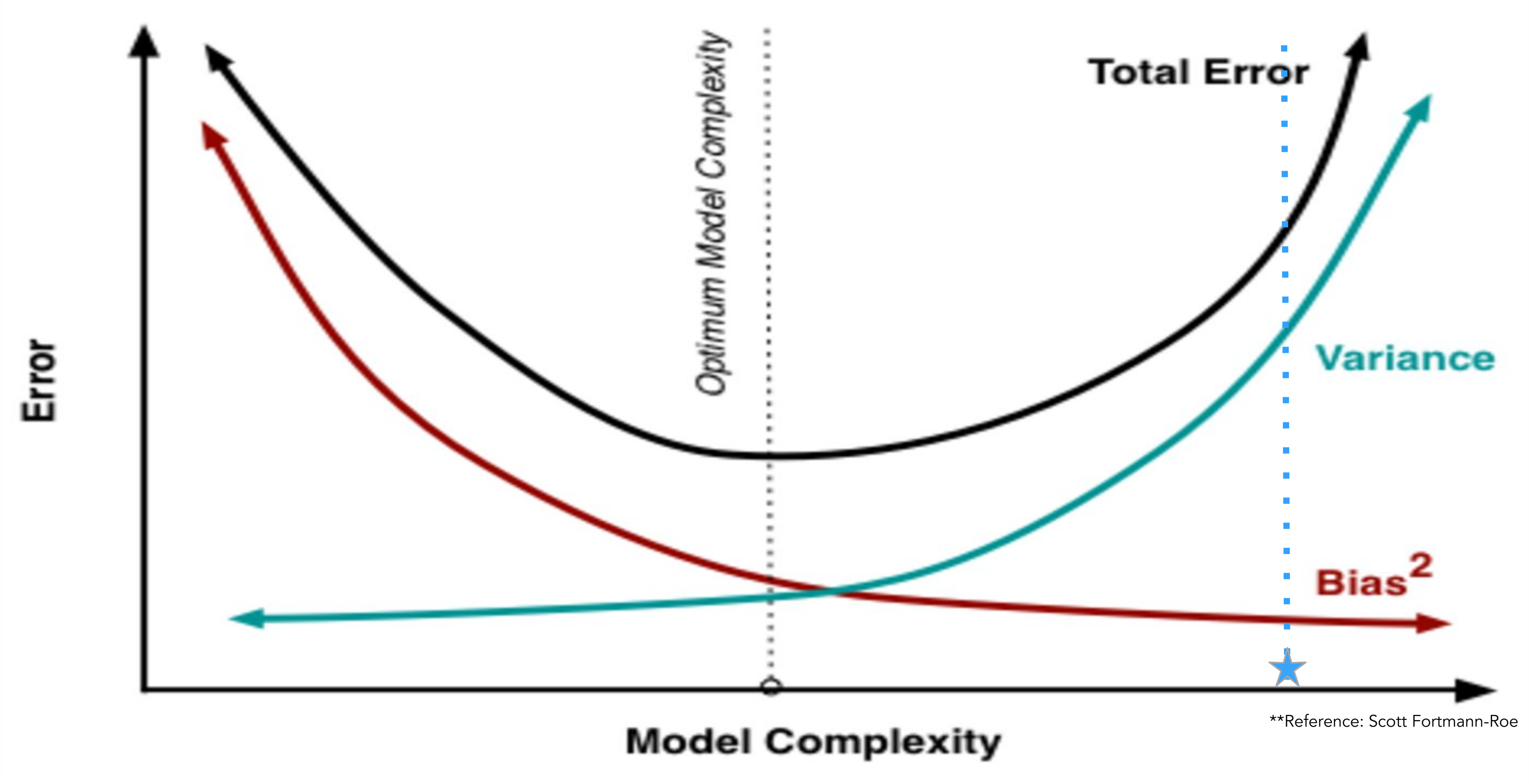
- High-level definition: Model interpretation is about understanding machine learning/statistical modeling behavior

- 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 the model outcome?

  - **What** other information can a model provide to avoid prediction errors?

  - **How** can we trust the predictions of a “black box” model?

ERROR VS MODEL COMPLEXITY



**Reference: Scott Fortmann-Roe

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

WHY MODEL INTERPRETATION?

- Helps in **exploring and discovering latent or hidden feature interactions** (useful for feature engineering/selection)

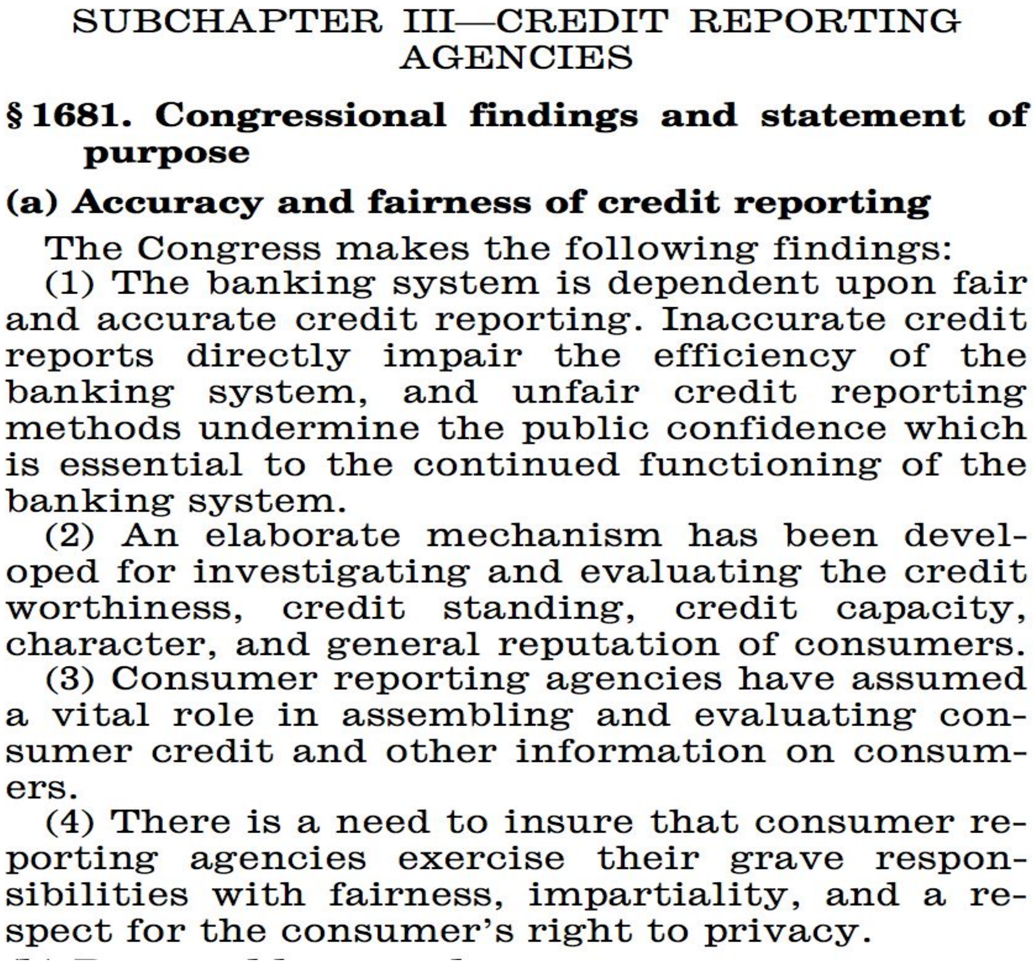
- Helps in understanding **model variability** as the environment changes (once the model is operationalized and is functional in a non-stationary environment)

- 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.

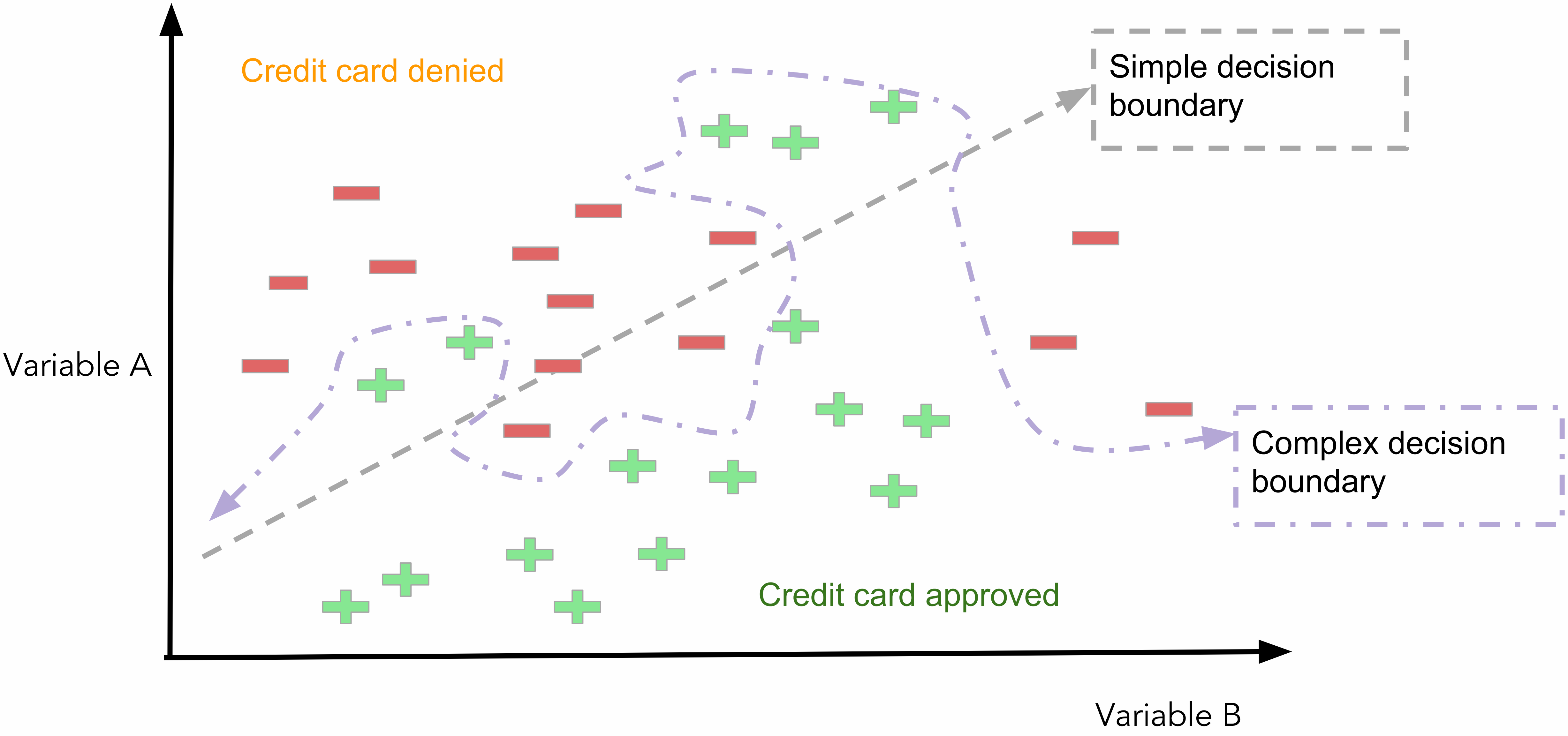
Are all predictive models interpretable?

An interpretable model may or may not predict well

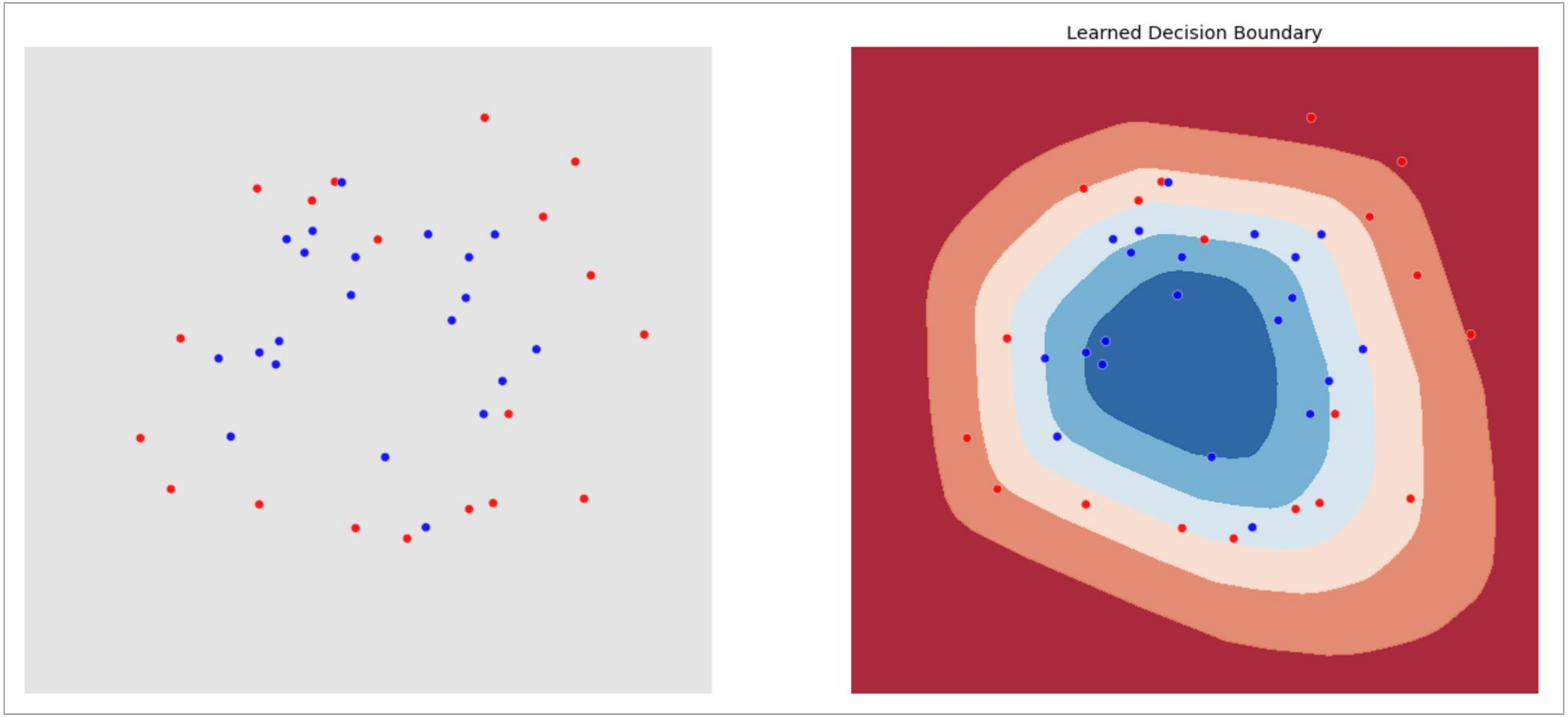
For example, we tried to benefit from an extensive set of attributes describing each of the movies in the dataset. Those attributes certainly carry a significant signal and can explain some of the user behavior. However, we concluded that they could not help at all for improving the accuracy of well tuned collaborative filtering models. Beyond selecting which features of the data to model, working with well designed models is also important ...

Solution to the Netflix Prize Bell et al., '08

PERFORMANCE VS. INTERPRETABILITY



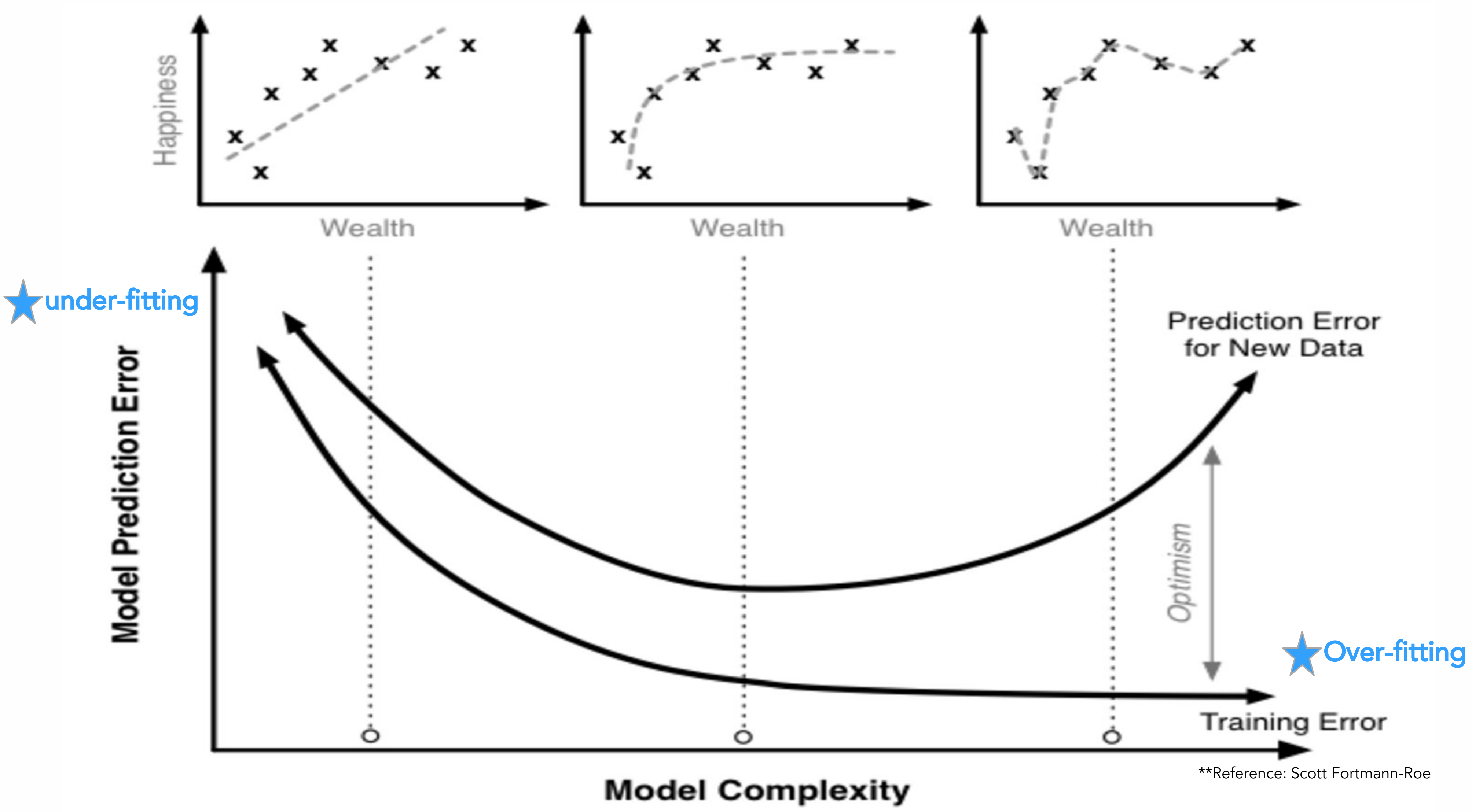
HOW ABOUT A MORE DIFFICULT RELATIONSHIP?



Data

Learned decision boundaries

Predictive Optimism

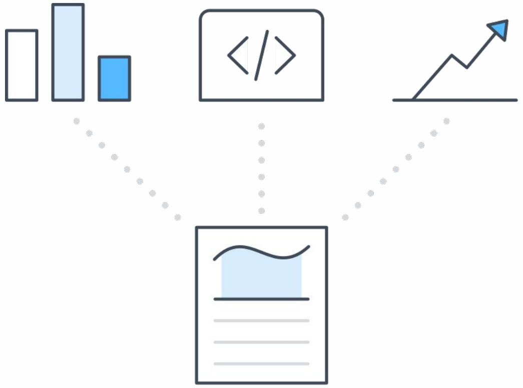


**Reference: Scott Fortmann-Roe

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 (H20.ai, sklearn, R packages)

- 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 analytical workflows during modeling process and post deployment



INTRODUCING ...

DATASCIENCE.COM

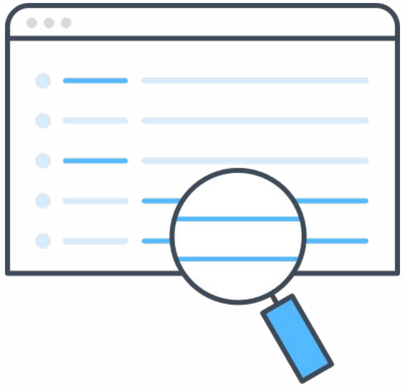
SKATER

WHAT IS SKATER?

- Python library designed to demystify the inner workings of black-box models

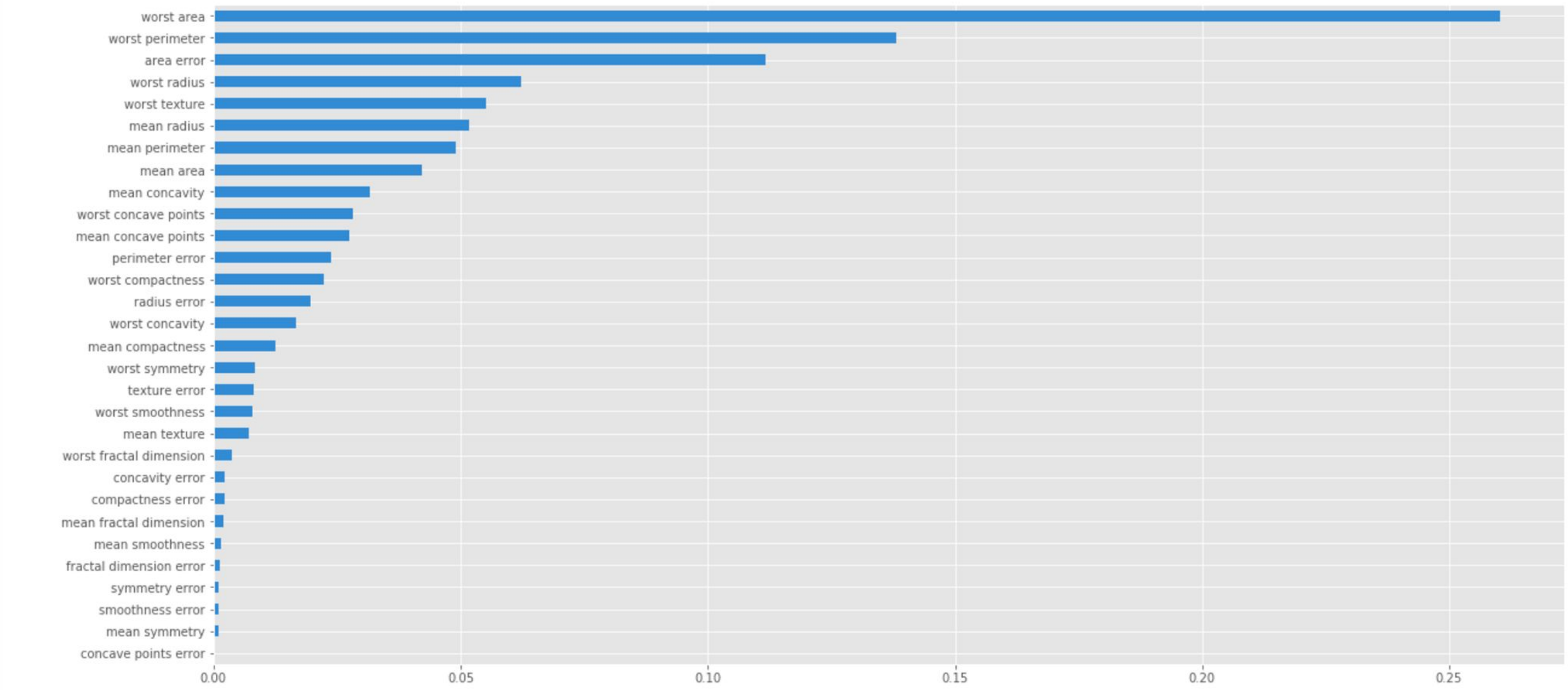
- Uses a number of techniques for model interpretation to explain the relationships between input data and desired output, both globally and locally

- One can interpret models both before and after they are operationalized



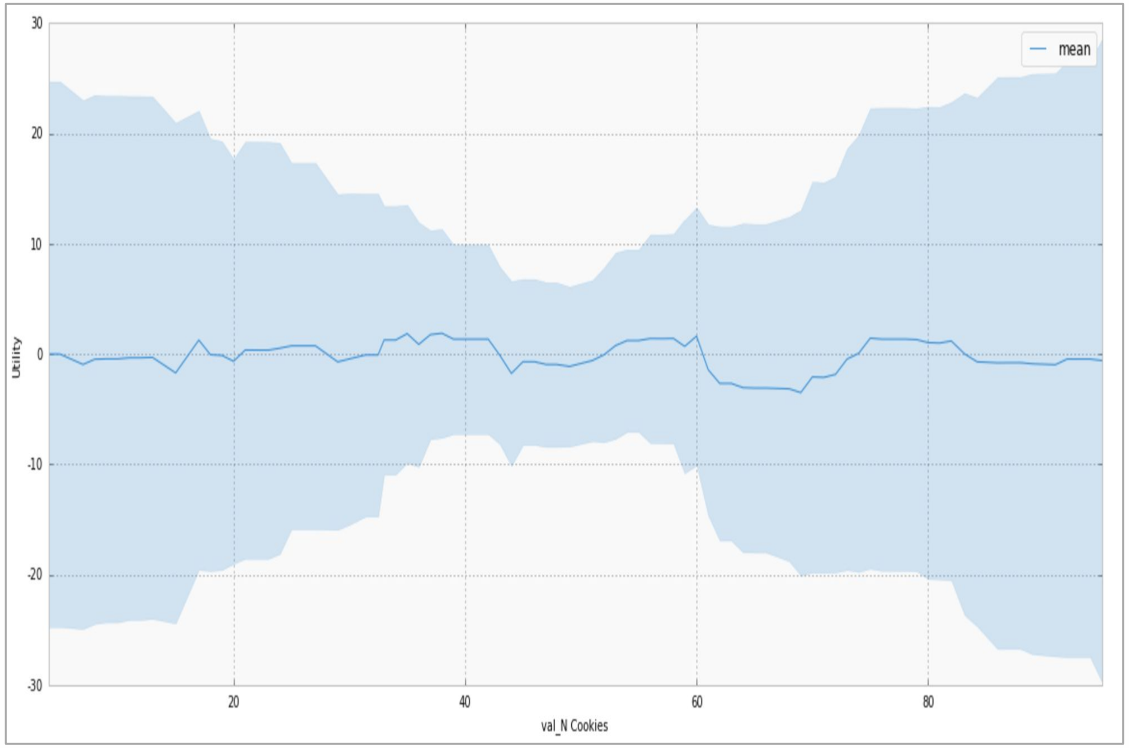
SKATER USES

- Model-agnostic variable importance for global interpretation
  - Helps in evaluating the importance of each independent input variable using variable perturbation

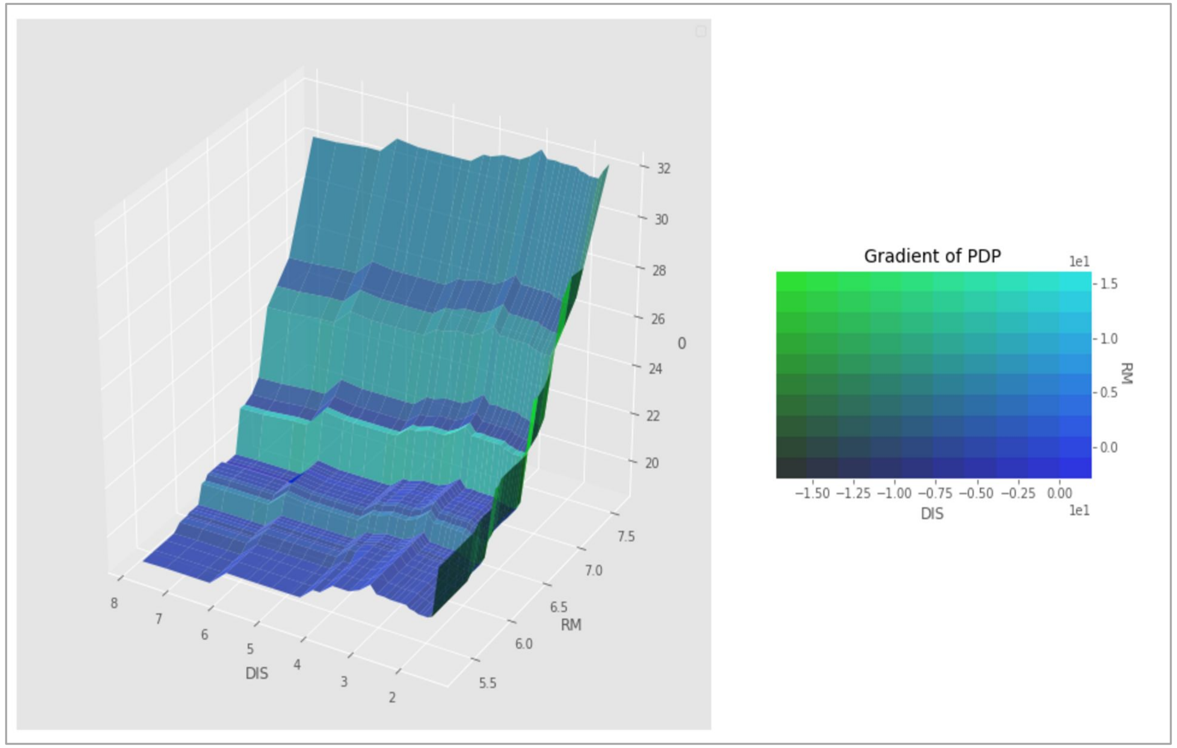


SKATER USES

- 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



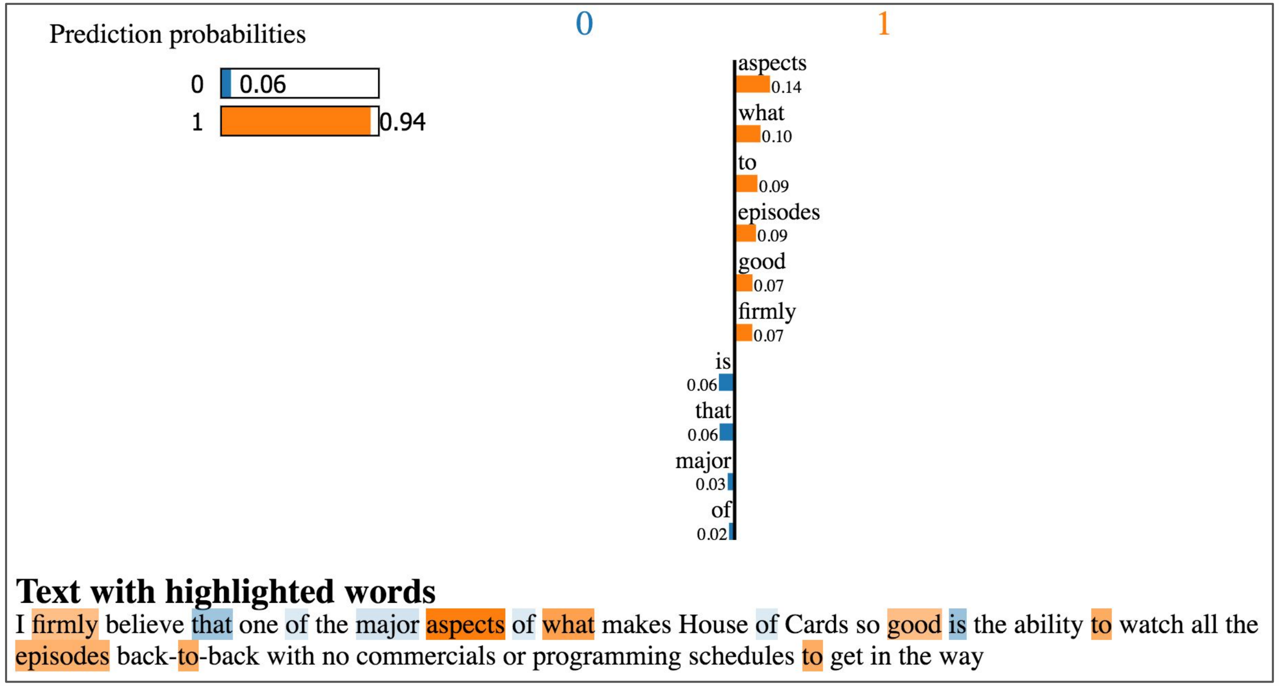
One-way interaction with variance



Two-way interaction

SKATER USES

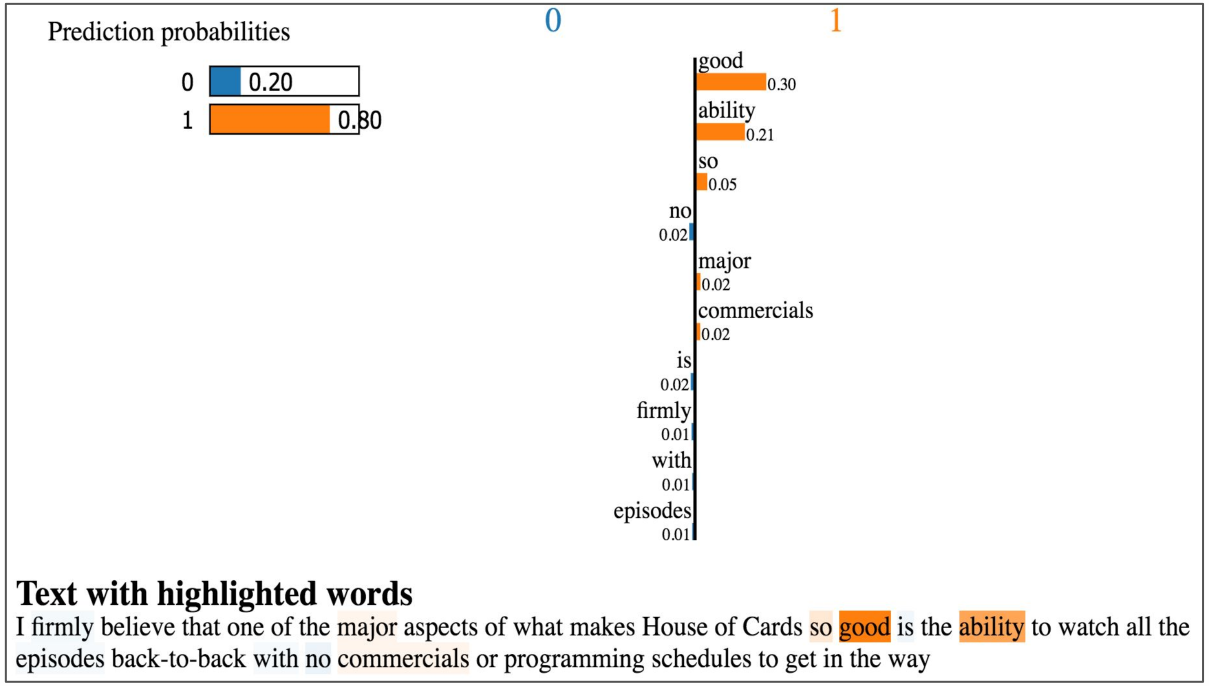
- Improved local interpretable model-agnostic explanations (LIME) for local interpretation
  - A novel technique developed by Marco, Sameer and Carlos to explain the behavior of any classifier or regressor in an human interpretable and faithful manner using surrogate models



Text with highlighted words

I firmly believe that one of the major aspects of what makes House of Cards so good is the ability to watch all the episodes back-to-back with no commercials or programming schedules to get in the way

Deployed Model - indico.io

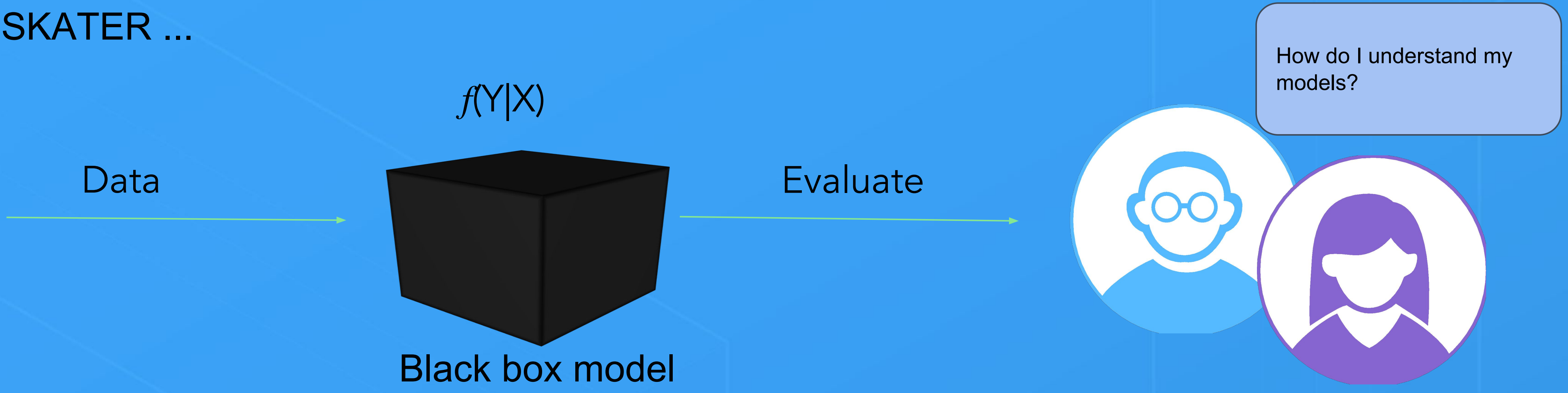


Text with highlighted words

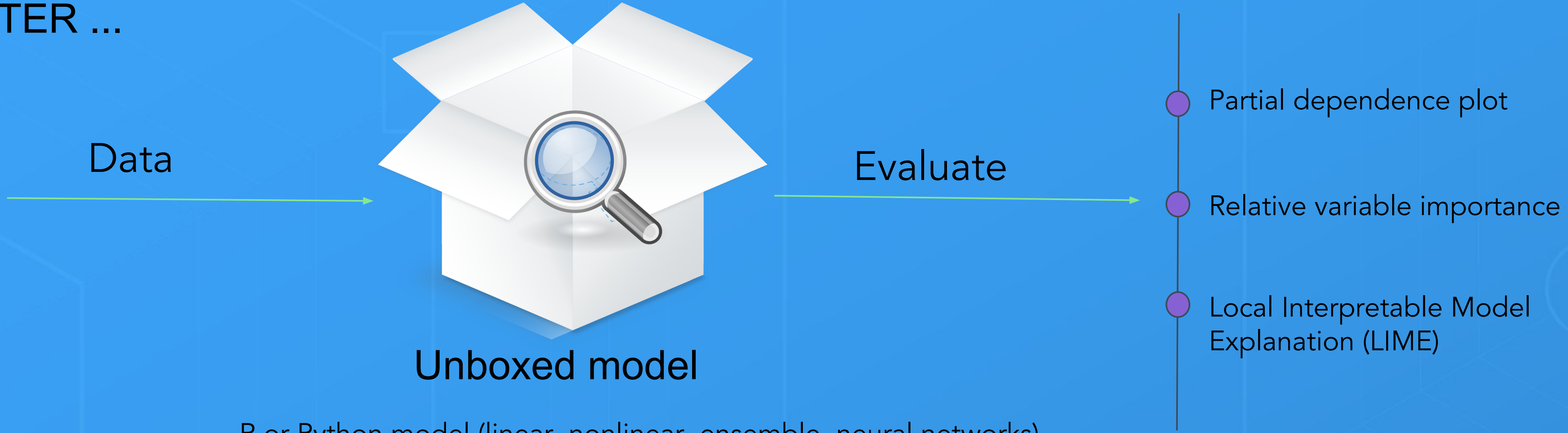
I firmly believe that one of the major aspects of what makes House of Cards so good is the ability to watch all the episodes back-to-back with no commercials or programming schedules to get in the way

Deployed Model - algorithmia

WITHOUT SKATER ...



WITH SKATER ...



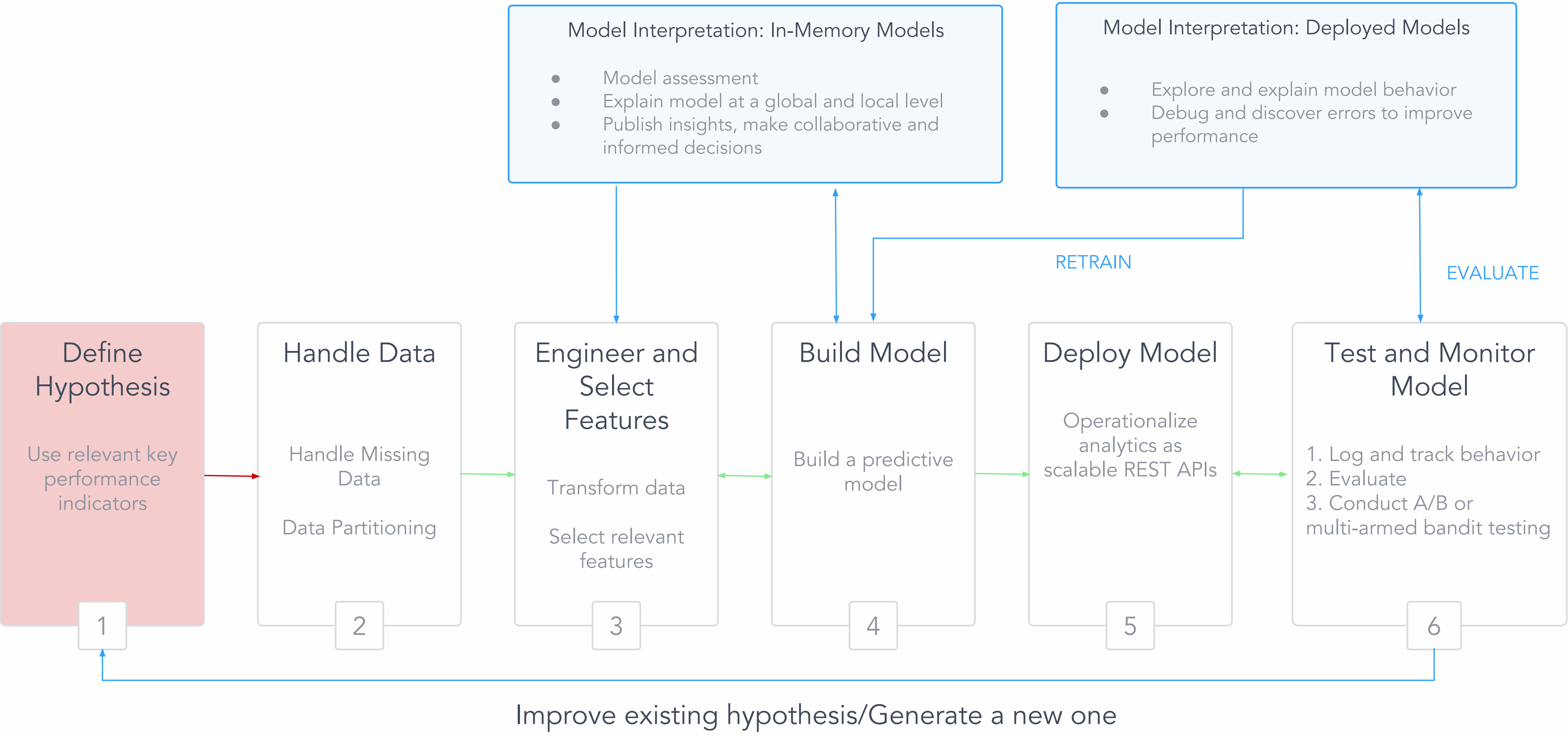
- Partial dependence plot
- Relative variable importance
- Local Interpretable Model Explanation (LIME)

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

Scikit-learn, caret and rpart packages for CRAN

H20.ai, Algorithmia, etc.

HOW DOES IT FIT INTO AN ANALYTICAL WORKFLOW?



Model Interpretation: In-Memory Models

- Model assessment
- Explain model at a global and local level
- Publish insights, make collaborative and informed decisions

Model Interpretation: Deployed Models

- Explore and explain model behavior
- Debug and discover errors to improve performance

Define Hypothesis

Use relevant key performance indicators

1

Handle Data

Handle Missing Data

Data Partitioning

2

Engineer and Select Features

Transform data

Select relevant features

3

Build Model

Build a predictive model

4

Deploy Model

Operationalize analytics as scalable REST APIs

5

Test and Monitor Model

1. Log and track behavior
2. Evaluate
3. Conduct A/B or multi-armed bandit testing

6

Improve existing hypothesis/Generate a new one

SKATER DEMO

INTERPRETATION ROADMAP

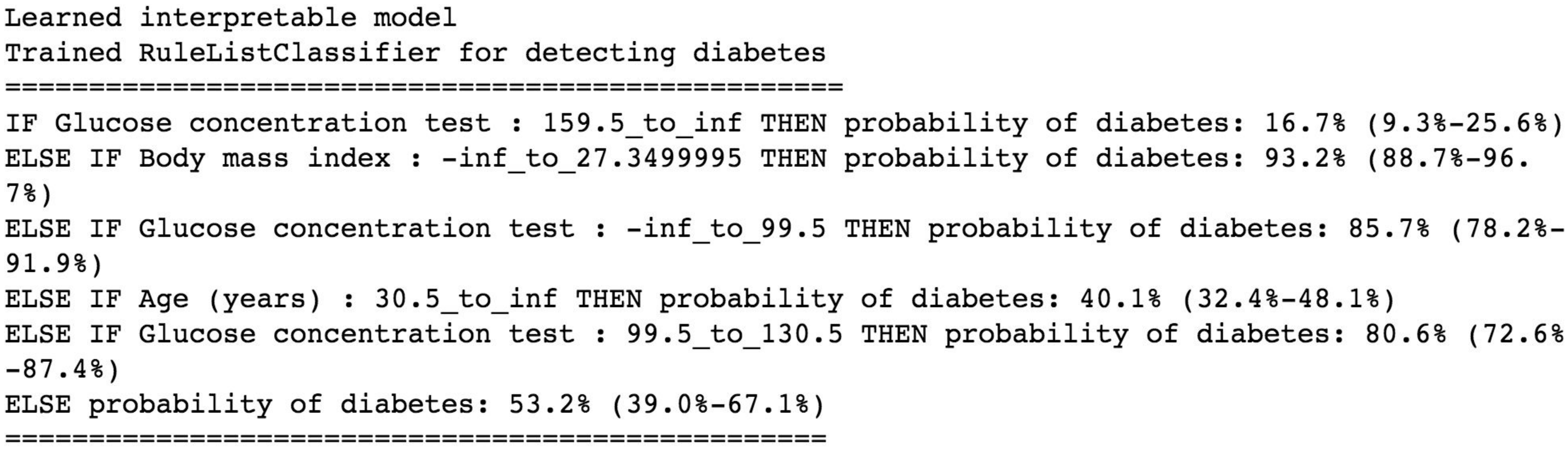
- More improvements to Deployed Model - H20, VW, Spark-MLLib

- Possible Future Work
    - Individual Conditional Expectation
    - Local Interpretation: e.g. Anchors for Local Interpretation
    - Global Interpretation: e.g. Probabilistic Rule based Models
    - Better support for Image Interpretability
        - extension of LIME
        - Class Activation Maps
    - Better ways to detect Interaction effect among model variables

- Help wanted: https://goo.gl/W17q4i

A QUICK GLIMPSE INTO THE FUTURE

- Bayesian Rule Lists: An interpretable model, with series of decision statement



```
Learned interpretable model
Trained RuleListClassifier for detecting diabetes
==================================================
IF Glucose concentration test : 159.5_to_inf THEN probability of diabetes: 16.7% (9.3%-25.6%)
ELSE IF Body mass index : -inf_to_27.3499995 THEN probability of diabetes: 93.2% (88.7%-96.
7%)
ELSE IF Glucose concentration test : -inf_to_99.5 THEN probability of diabetes: 85.7% (78.2%-
91.9%)
ELSE IF Age (years) : 30.5_to_inf THEN probability of diabetes: 40.1% (32.4%-48.1%)
ELSE IF Glucose concentration test : 99.5_to_130.5 THEN probability of diabetes: 80.6% (72.6%
-87.4%)
ELSE probability of diabetes: 53.2% (39.0%-67.1%)
=================================================
```

Q&A

dstm@datascience.com

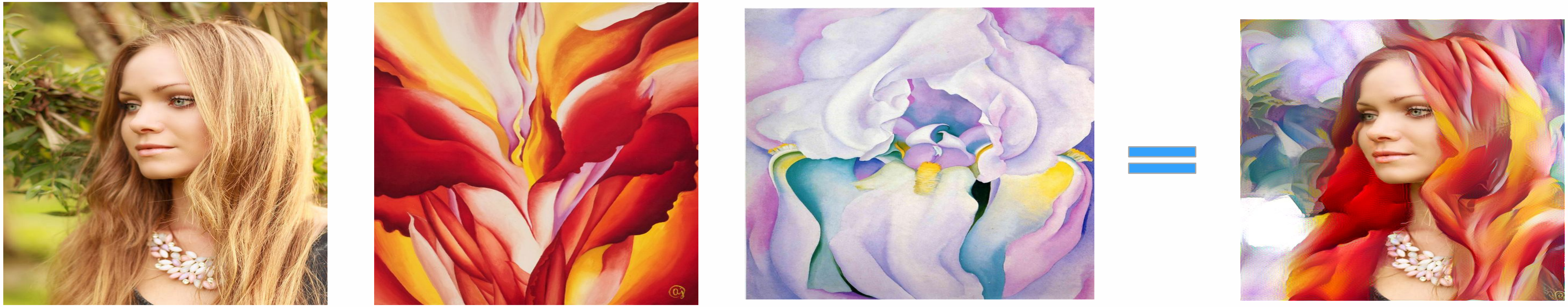
pramit@datascience.com

Appendix

UPCOMING TALKS AND EVENTS

PyData
Seattle 2017

○ "Learn to be a painter using Neural Style Painting"



UPCOMING TALKS AND EVENTS

- PyData Seattle 2017
  - ○ Jean-rene.gauthier and Ben Van Dyke: **“Implementing and Training Predictive Customer Lifetime Value Models in Python”**

- JUPYTERCON
  - ○ Aaron Kramer: **“Interactive natural language processing with SpaCy and Jupyter”**

-

  Andrea Trevino: will be a panelist discussing data science best practices with industry leaders from Google, Netflix, eHarmony, Live Nation and others