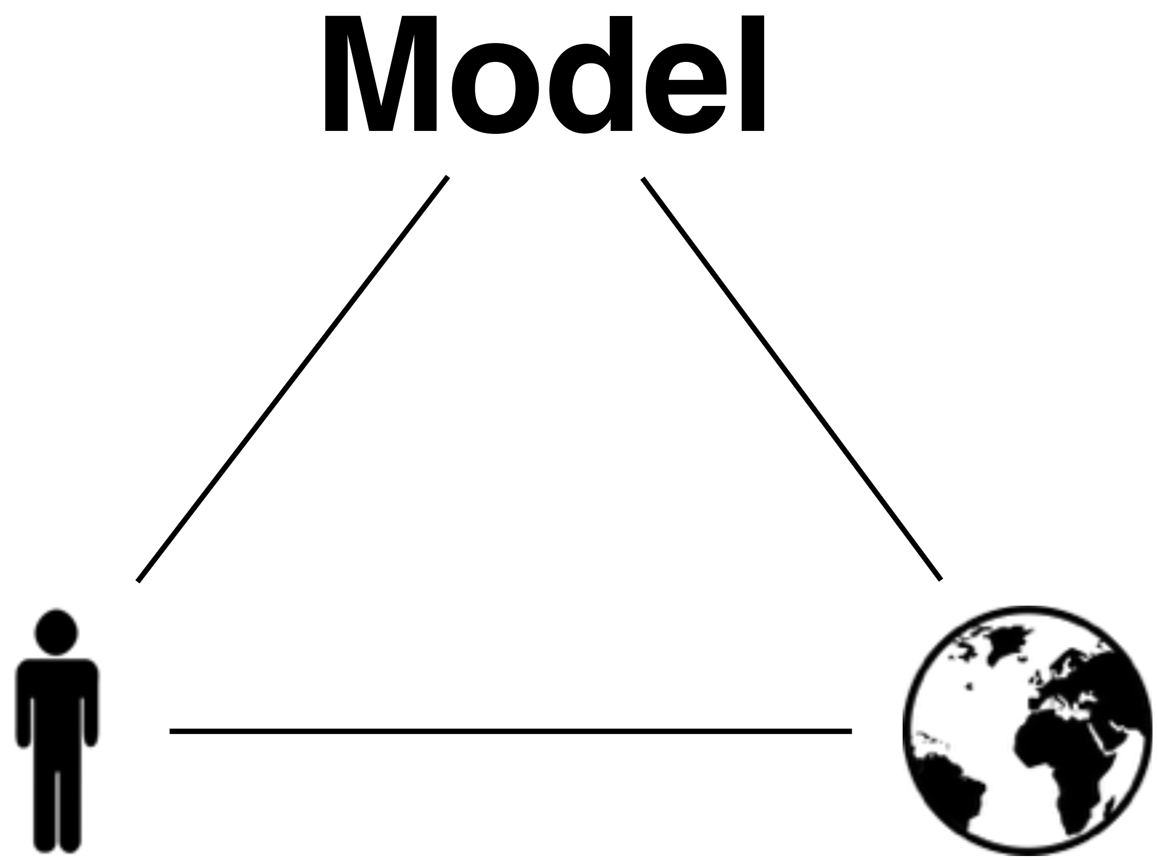
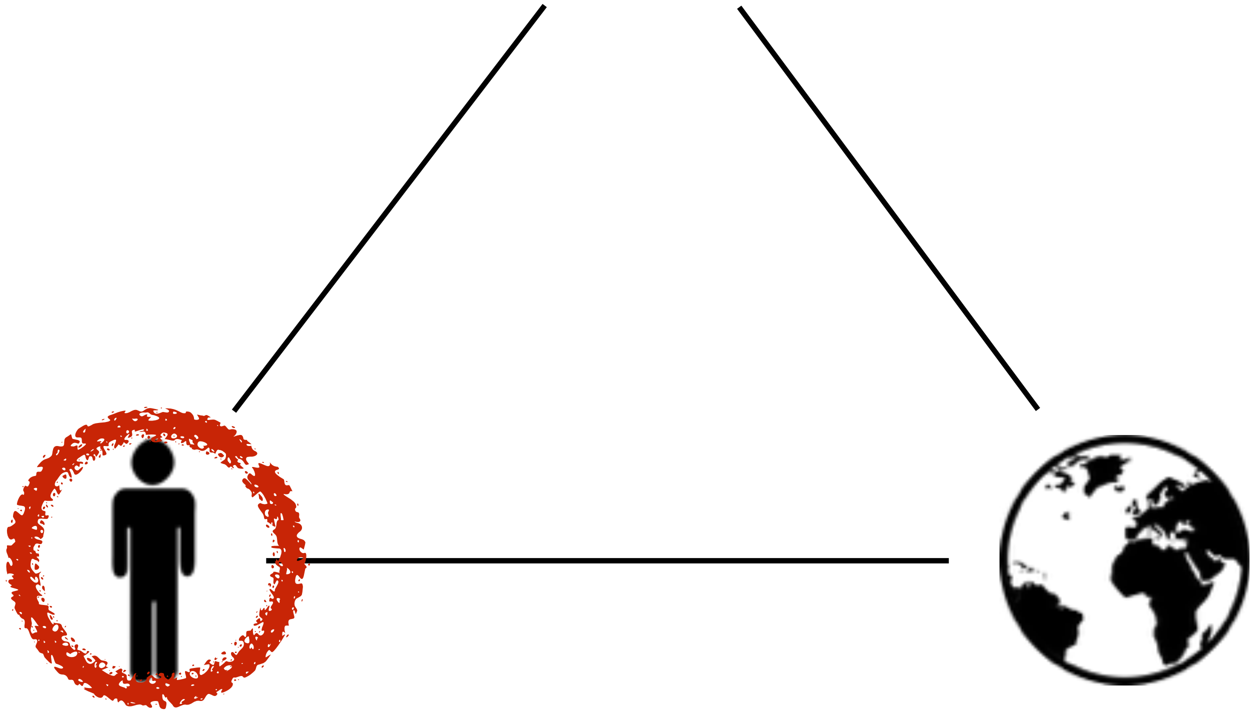
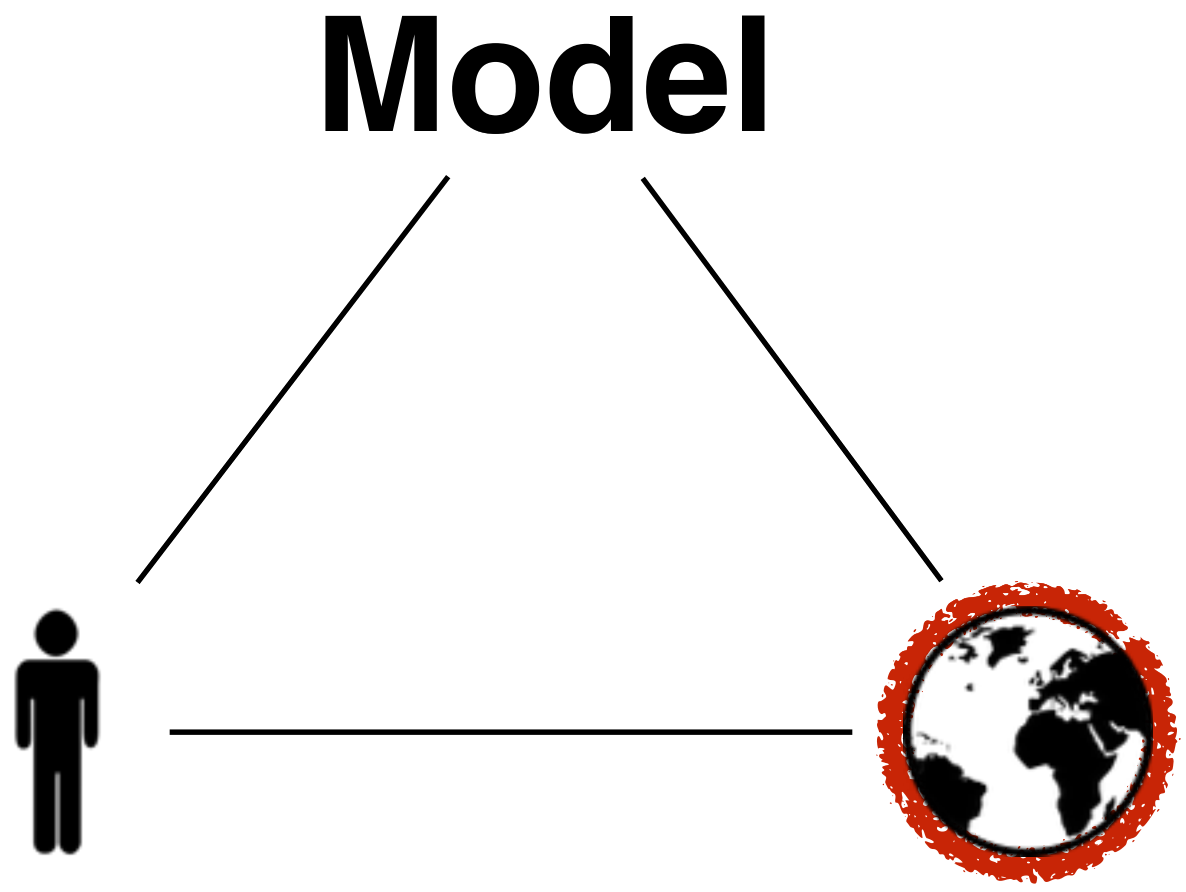
Explanatory Visual Analytics for Enhancing Human Interpretability of Machine Learning Models

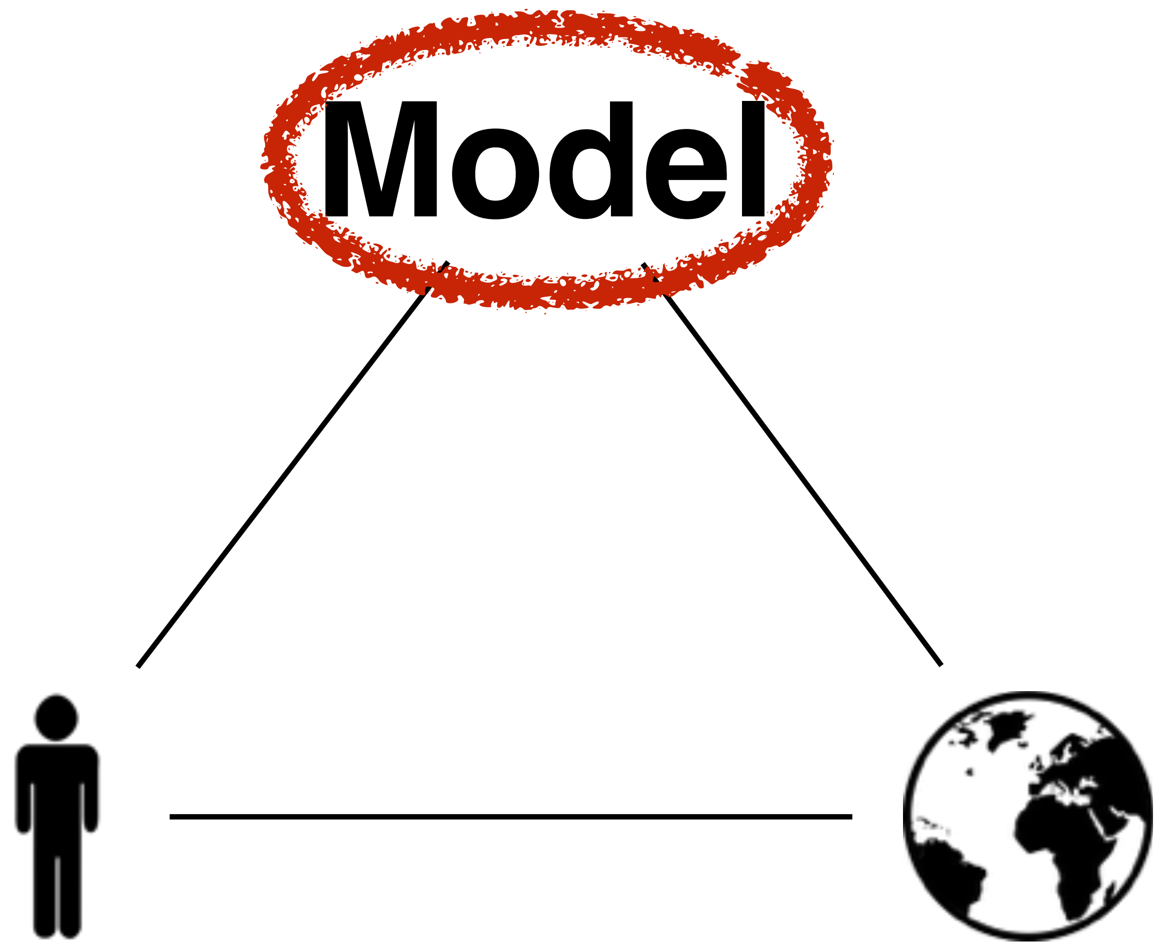
Josua Krause*, Aritra Dasgupta+, Enrico Bertini*
*NYU, +PNNL

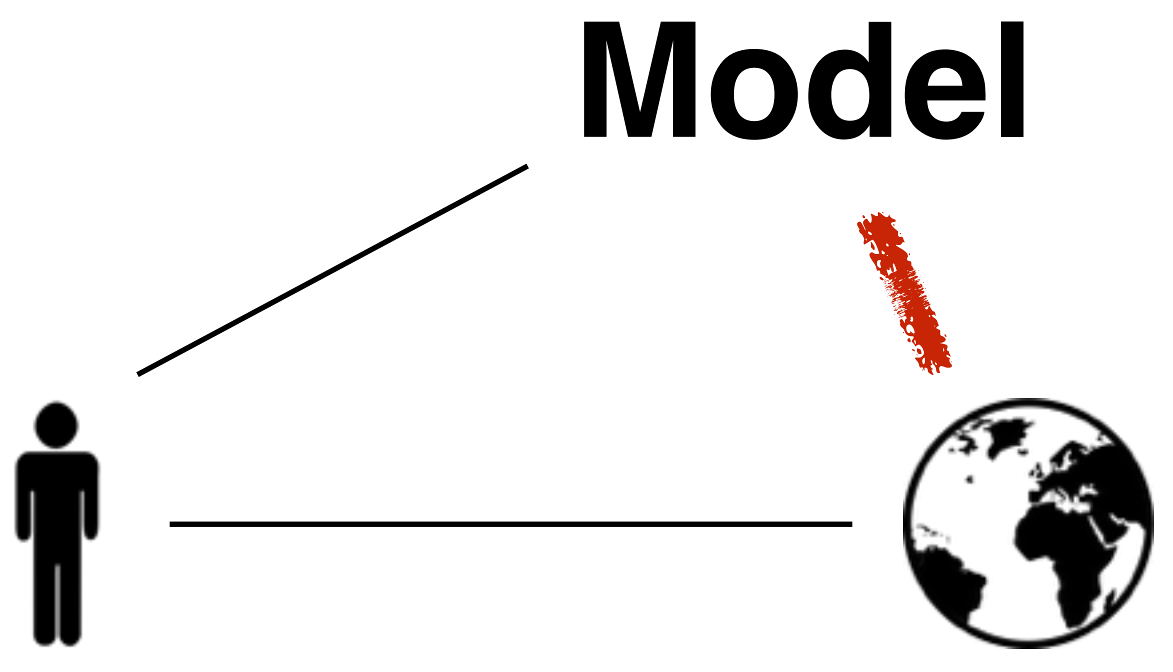

Model

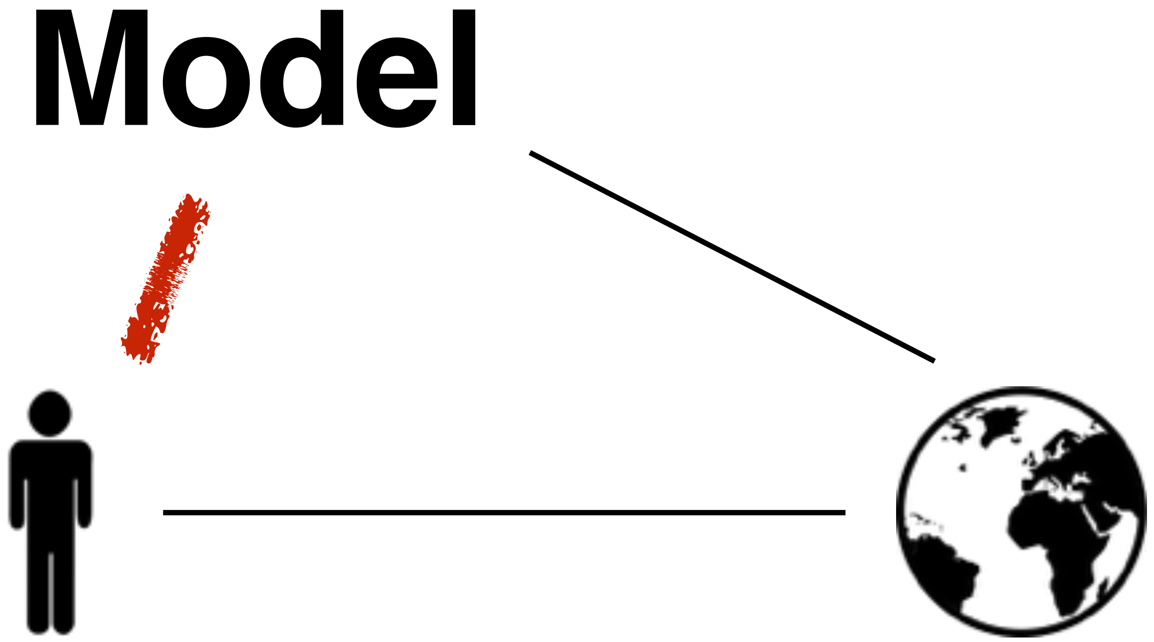
Model




Model


Model


Model


Model

Predictive Modeling

Input → **Model** → **Output**

Predictive Modeling

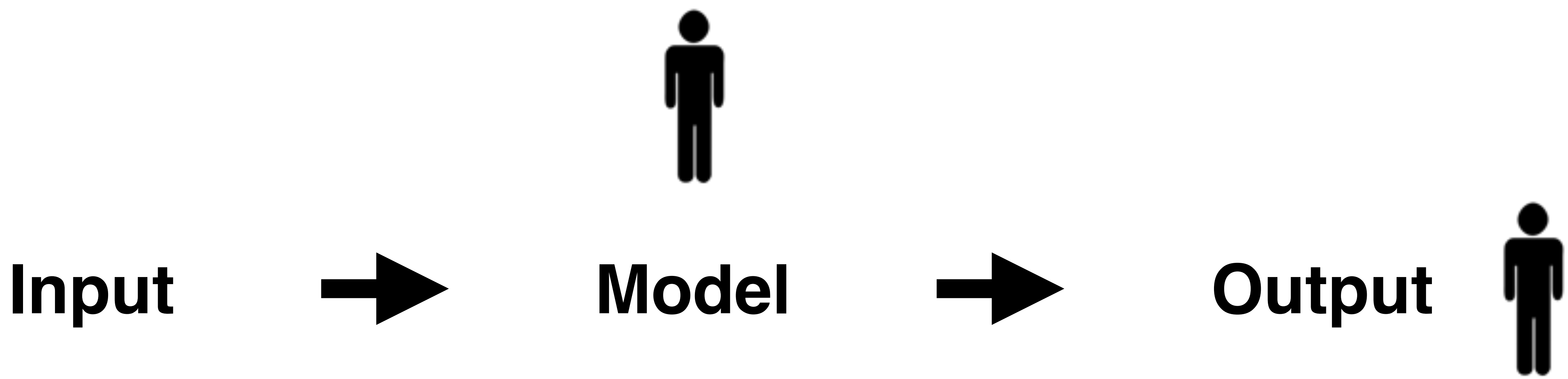
Input		**Model**		**Output**
High Dimensional Input Data	→	Trained Machine Learning Model	→	Prediction Scores

Predictive Modeling

Input → **Model** → **Output** 

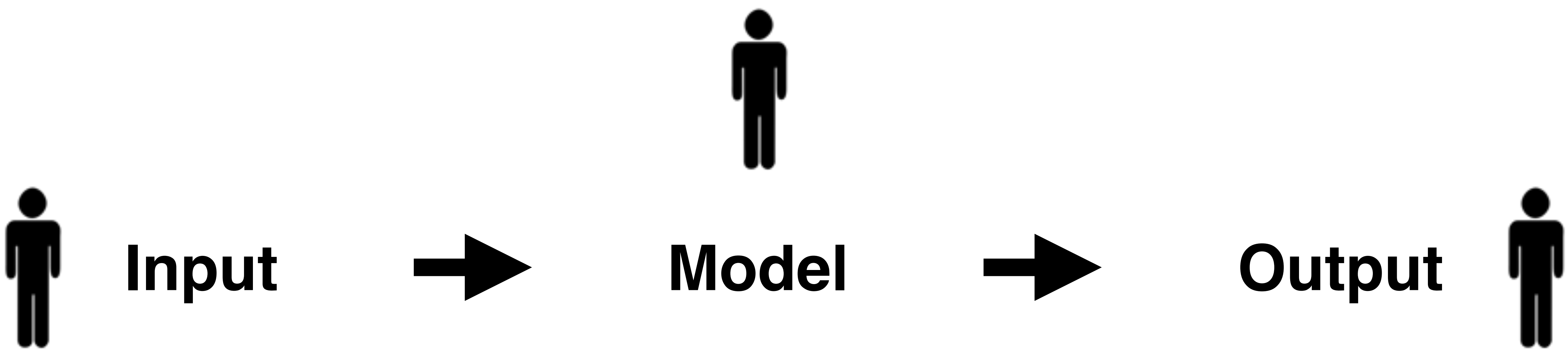
- Liability with decision
- Trust in decision

Predictive Modeling



- Model debugging
- Comparison of models

Predictive Modeling



- Find hidden associations
- Reduction of ambiguity

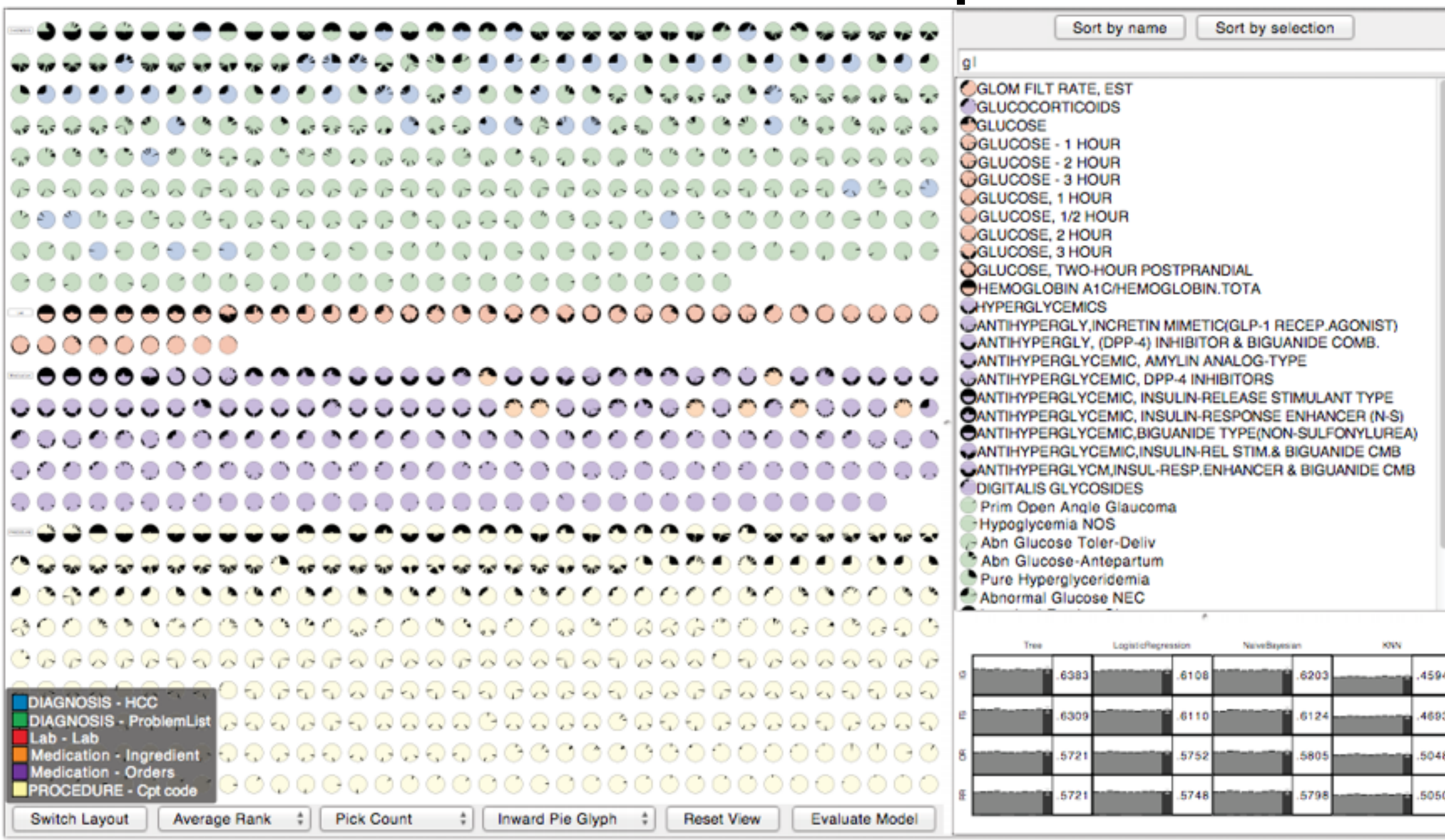
Recent Works

- INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data
  Josua Krause, Adam Perer, Enrico Bertini – *VAST 2014*
- Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models
  Josua Krause, Adam Perer, Kenney Ng – *CHI 2016*
- Using Visual Analytics to Interpret Predictive Machine Learning Models
  Josua Krause, Adam Perer, Enrico Bertini – *WHI 2016 ICML*
- Using Neural Networks for Data Mining
  Mark Craven, Jude Shavlik – *Future Generation Computer Systems 1997*
- Towards Better Analysis of Deep Convolutional Neural Networks
  Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, Shixia Liu – *VAST 2016*
- "Why Should I Trust You?" Explaining the Predictions of Any Classifier
  Marco Riberio, Sameer Singh, Carlos Guestrin – *KDD 2016*

Visual Analytics

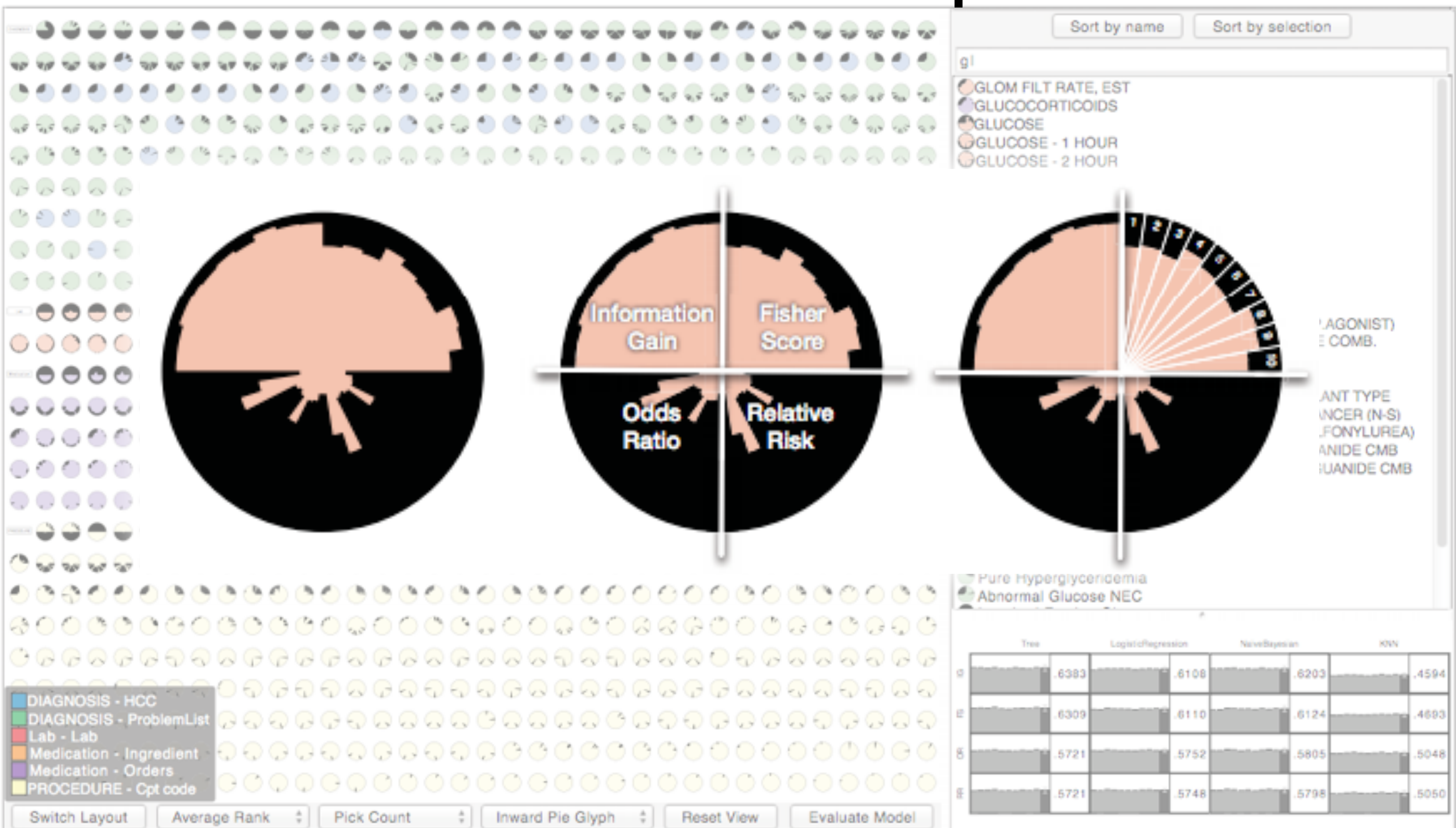
Model Output

Model Output



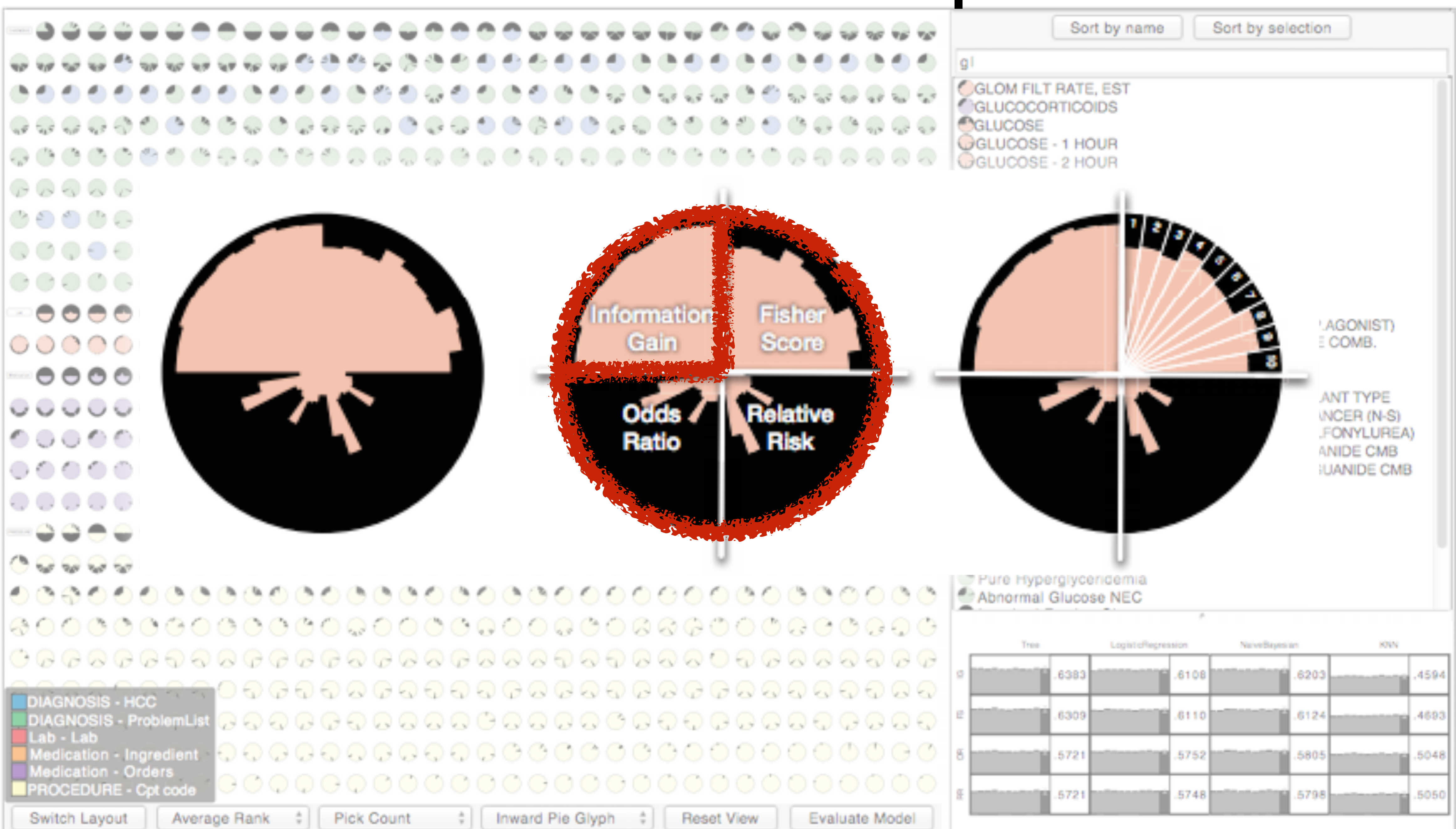
INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data
Josua Krause, Adam Perer, Enrico Bertini – *VAST 2014*

Model Output



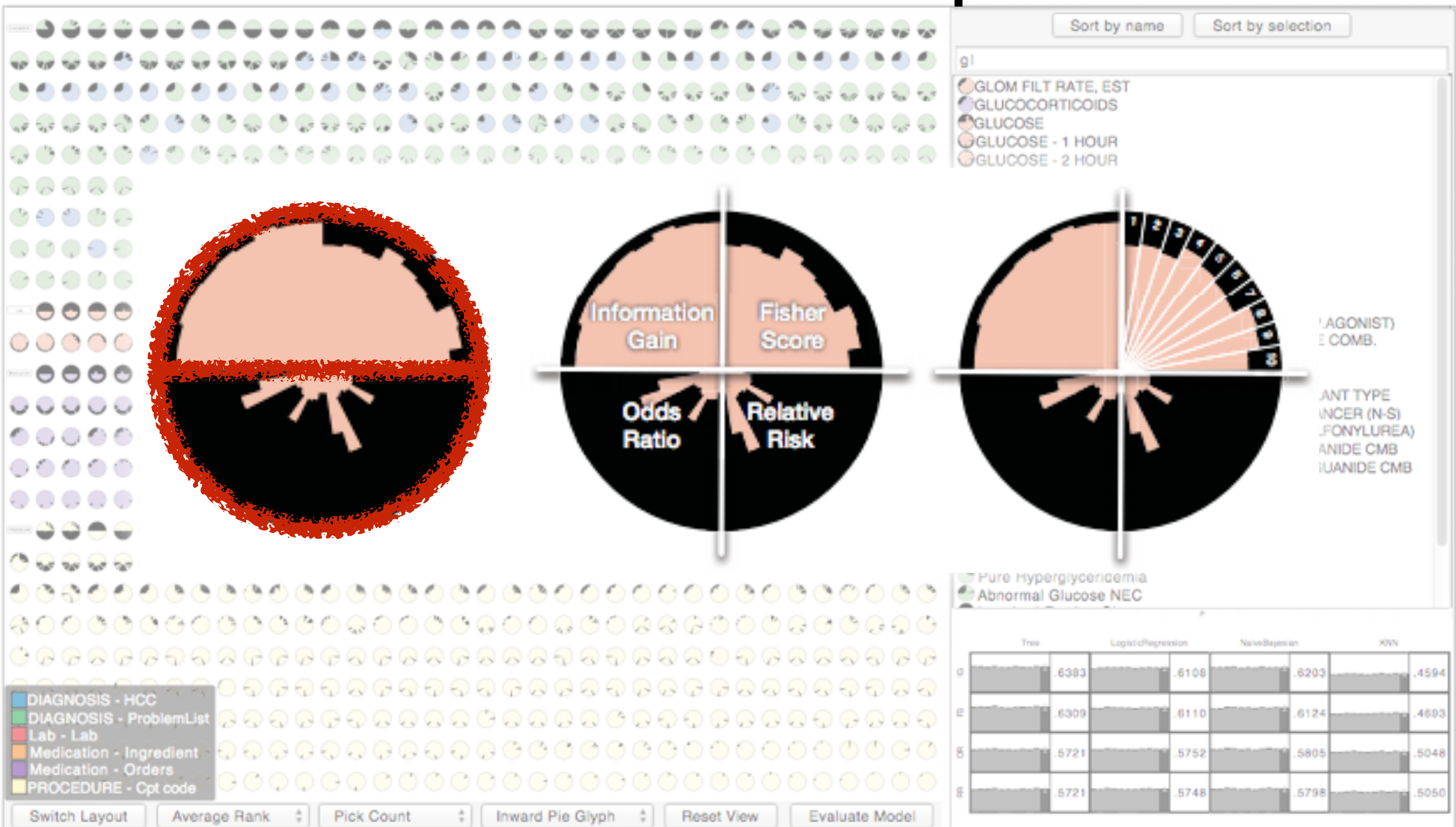
INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data
Josua Krause, Adam Perer, Enrico Bertini – *VAST 2014*

Model Output



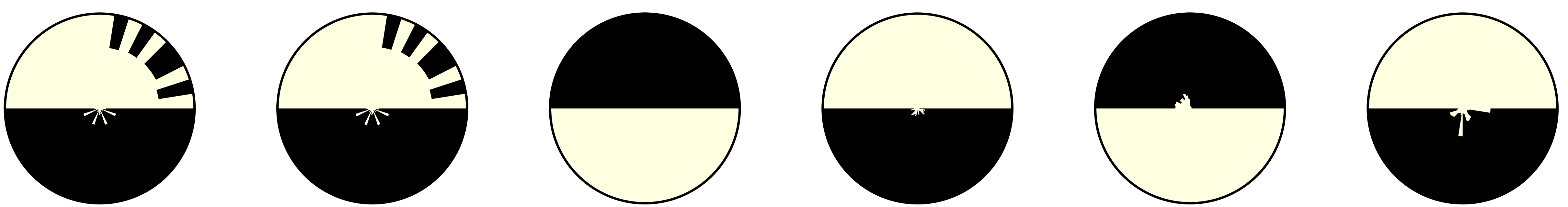
INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data
Josua Krause, Adam Perer, Enrico Bertini – *VAST 2014*

Model Output



INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data
Josua Krause, Adam Perer, Enrico Bertini – *VAST 2014*

Different algorithms prefer different features but yield similar results

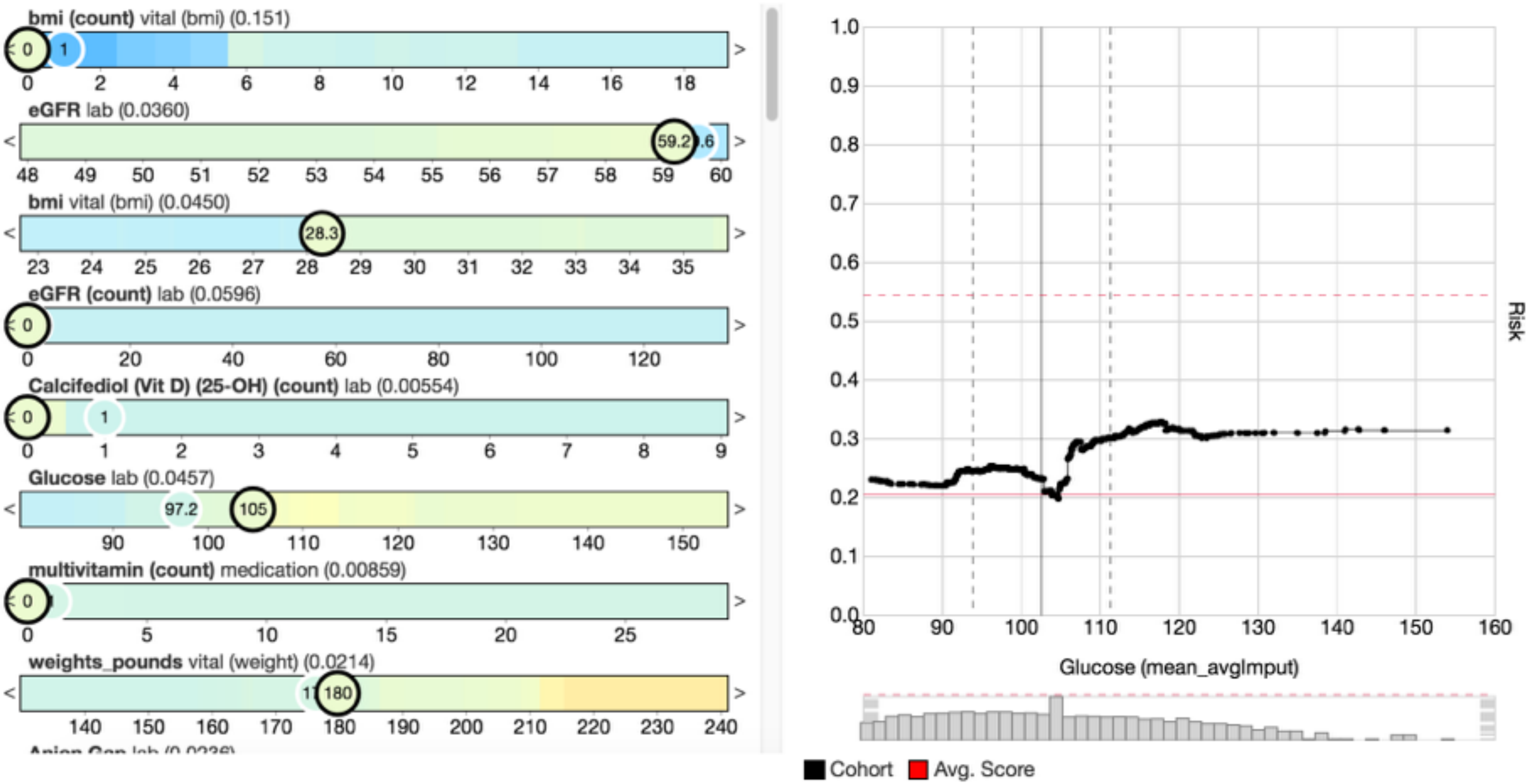


Visual Analytics

Model Output

Model Interaction

Model Interaction

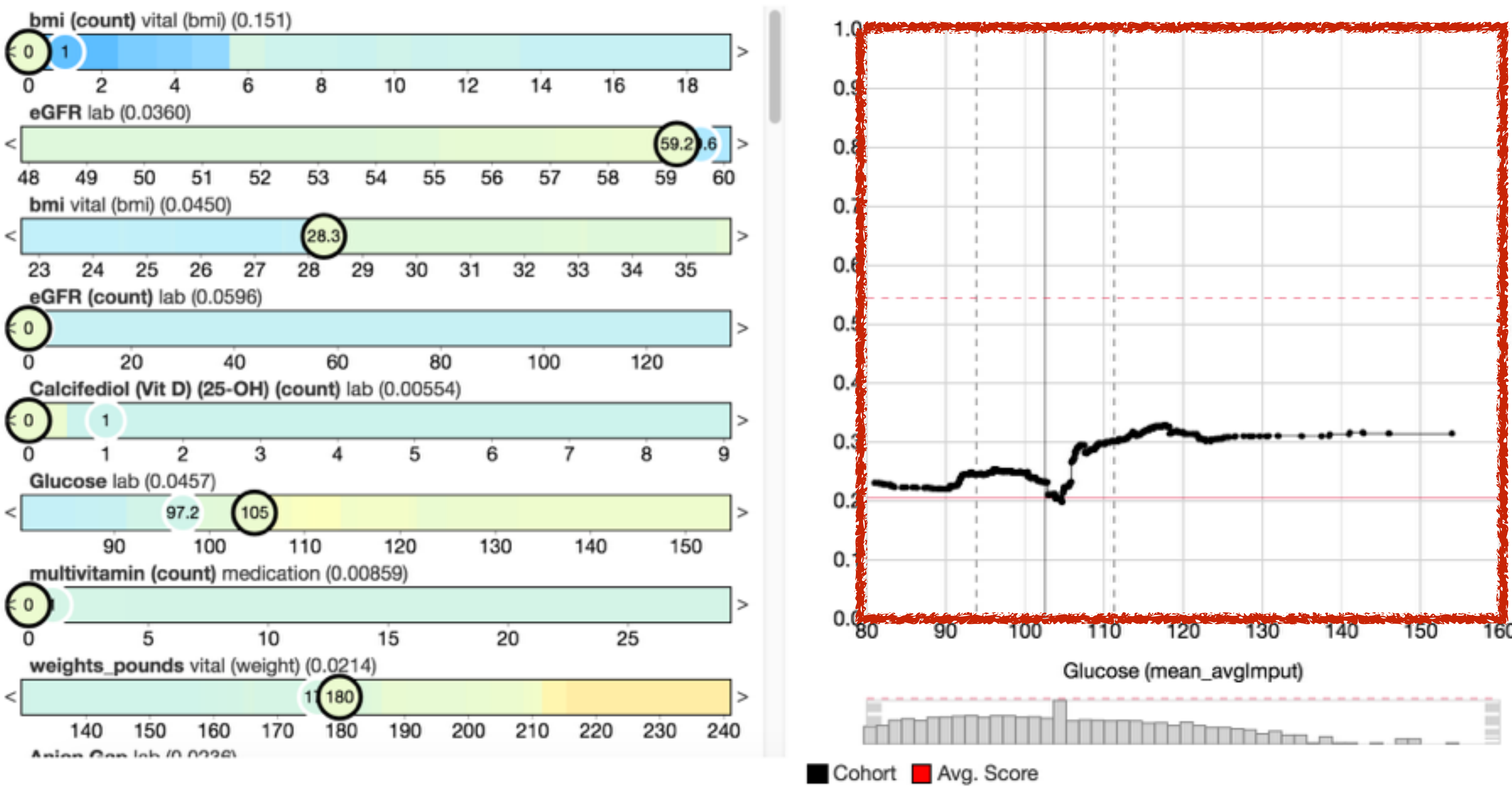


Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models
Josua Krause, Adam Perer, Kenney Ng – *CHI 2016*

Model Interaction

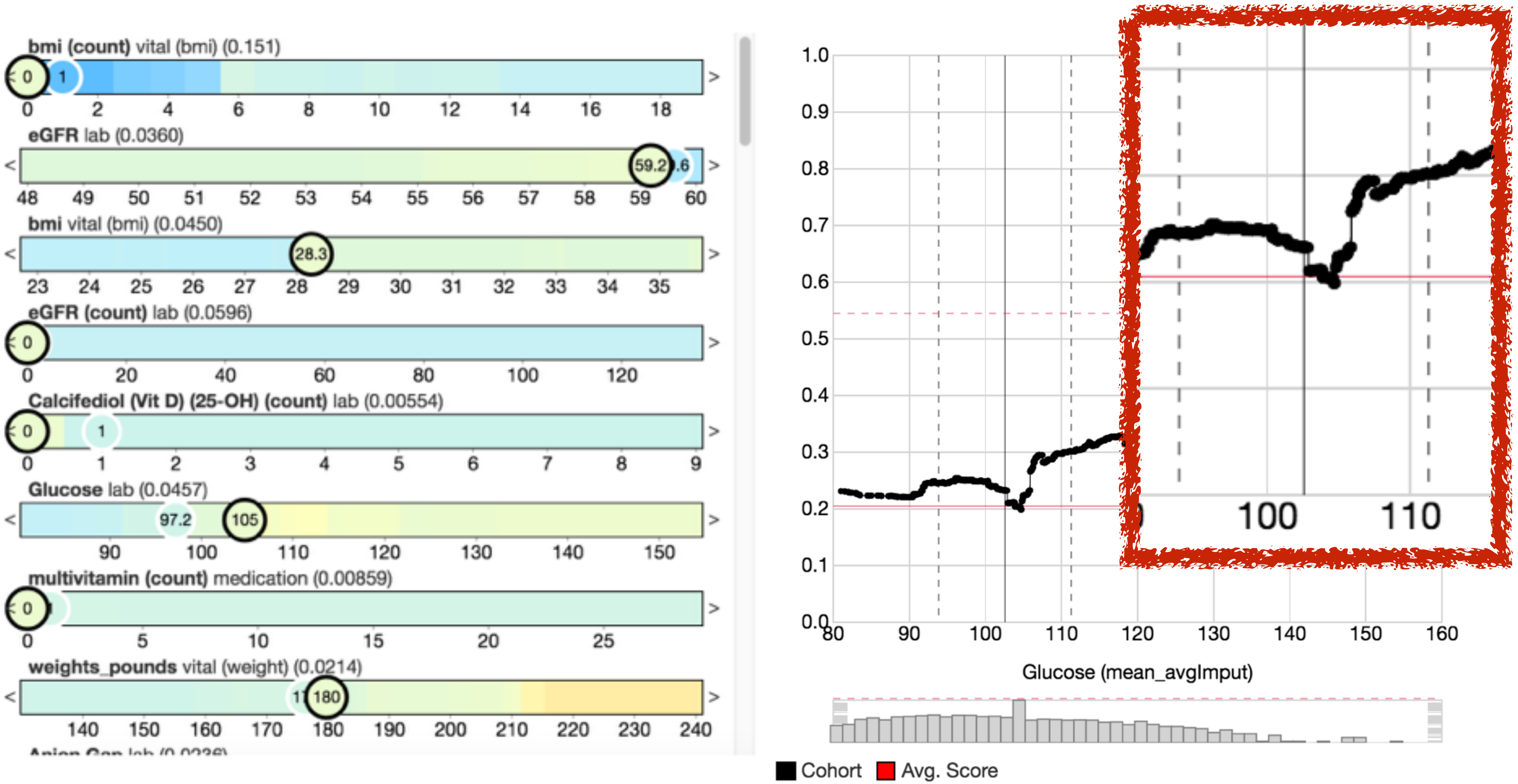
- What are expected values?
- What needs to change for flipping the label?

Model Interaction



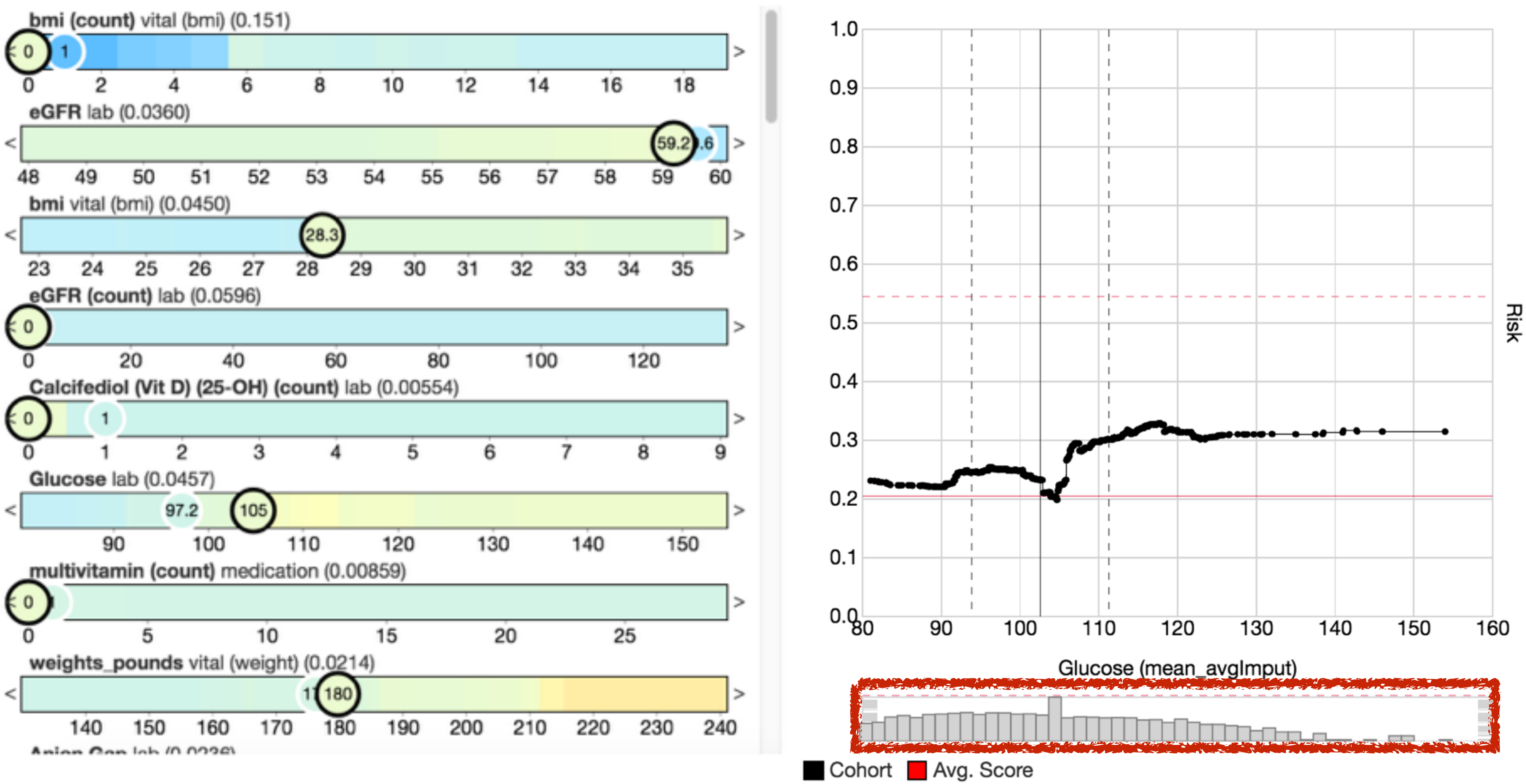
Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models
Josua Krause, Adam Perer, Kenney Ng – *CHI 2016*

Model Interaction



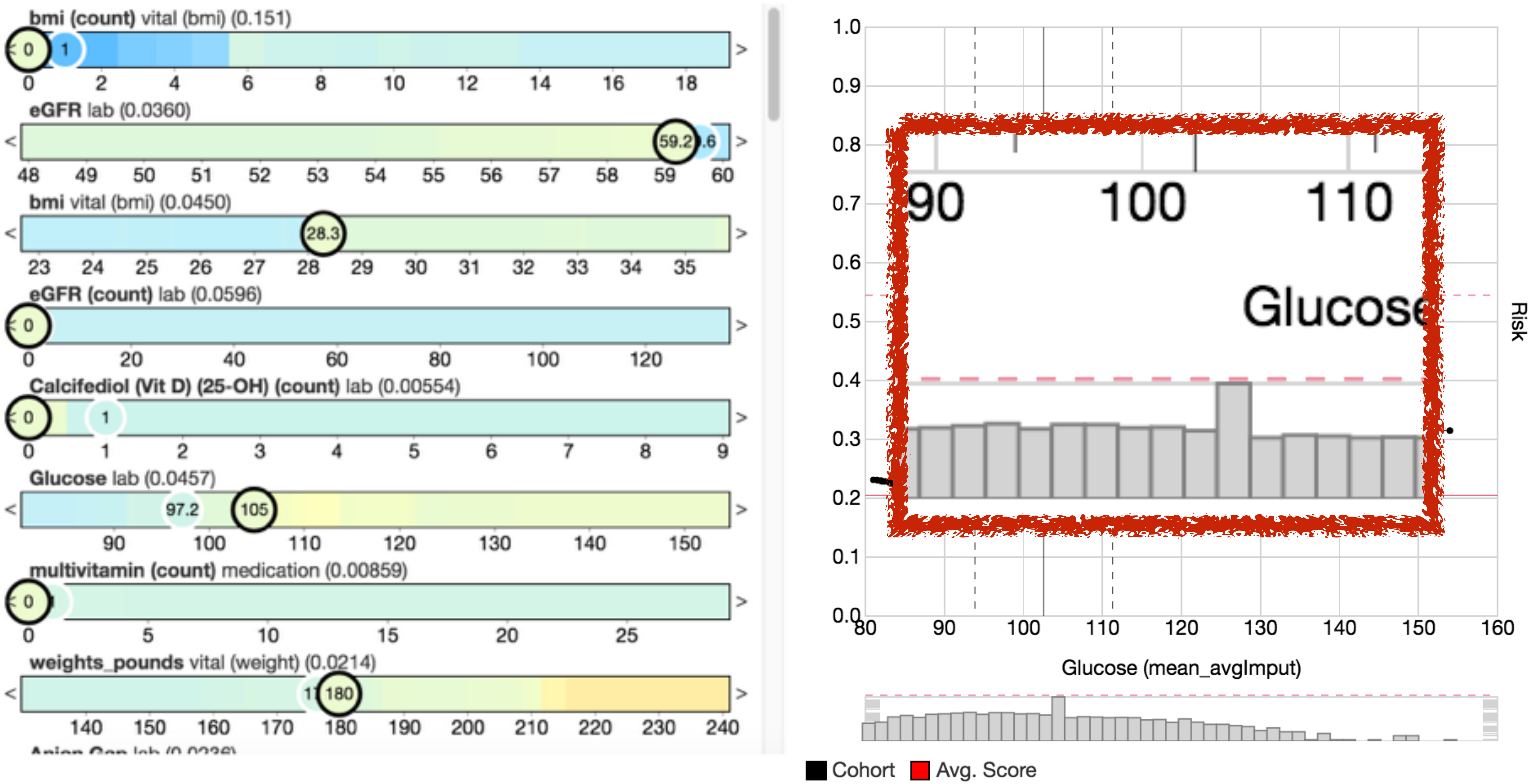
Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models
Josua Krause, Adam Perer, Kenney Ng – *CHI 2016*

Model Interaction



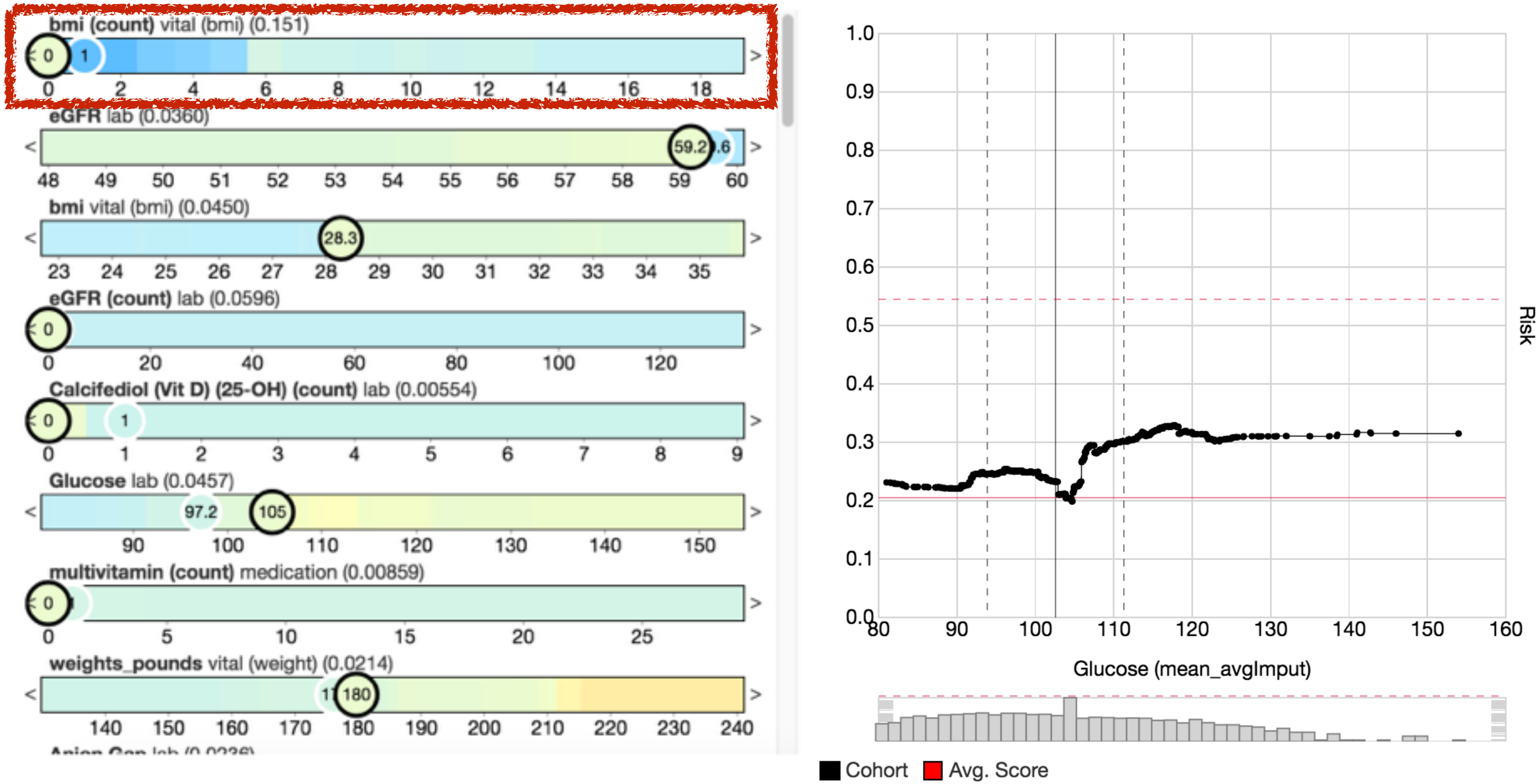
Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models
Josua Krause, Adam Perer, Kenney Ng – *CHI 2016*

Model Interaction



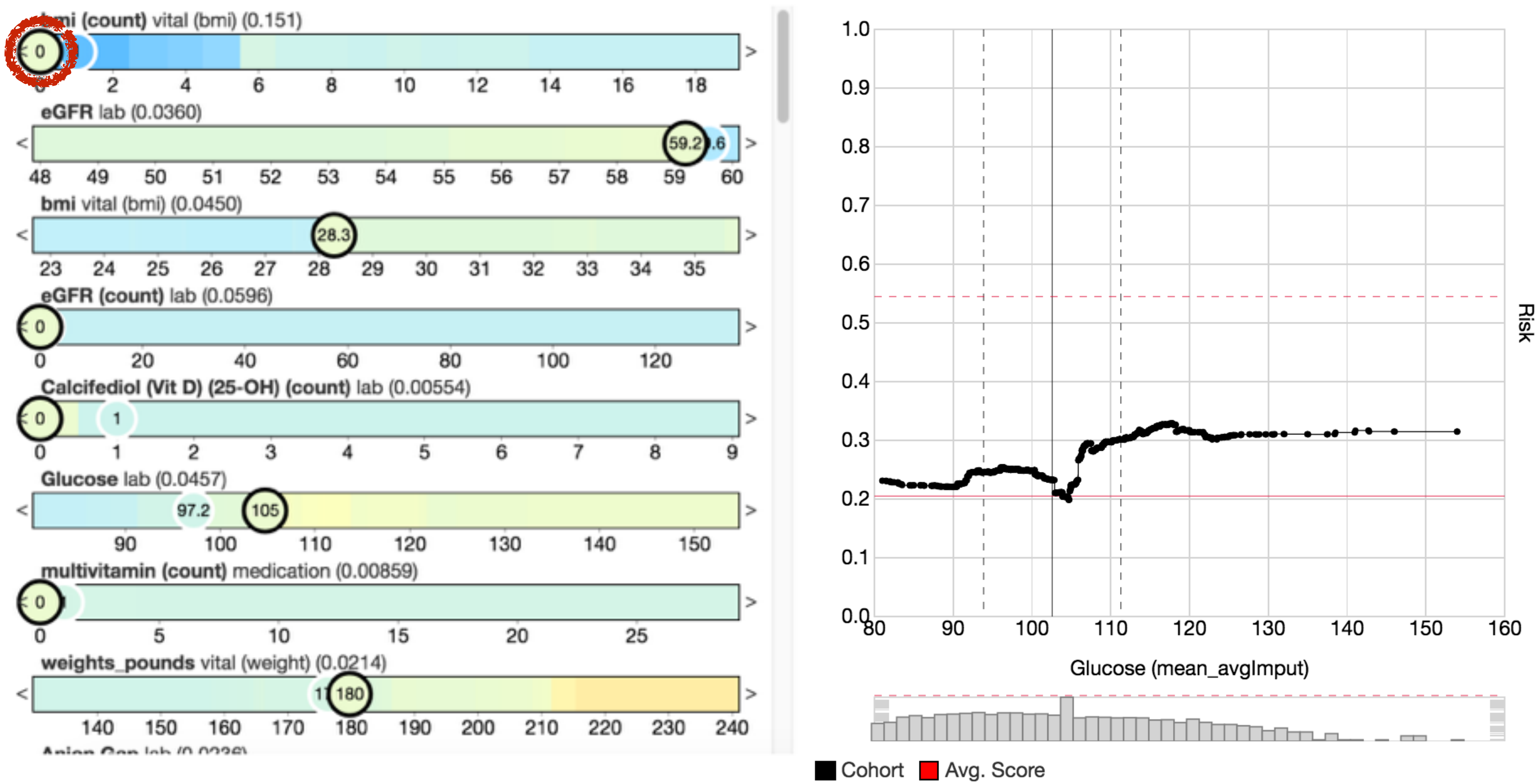
Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models
Josua Krause, Adam Perer, Kenney Ng – *CHI 2016*

Model Interaction



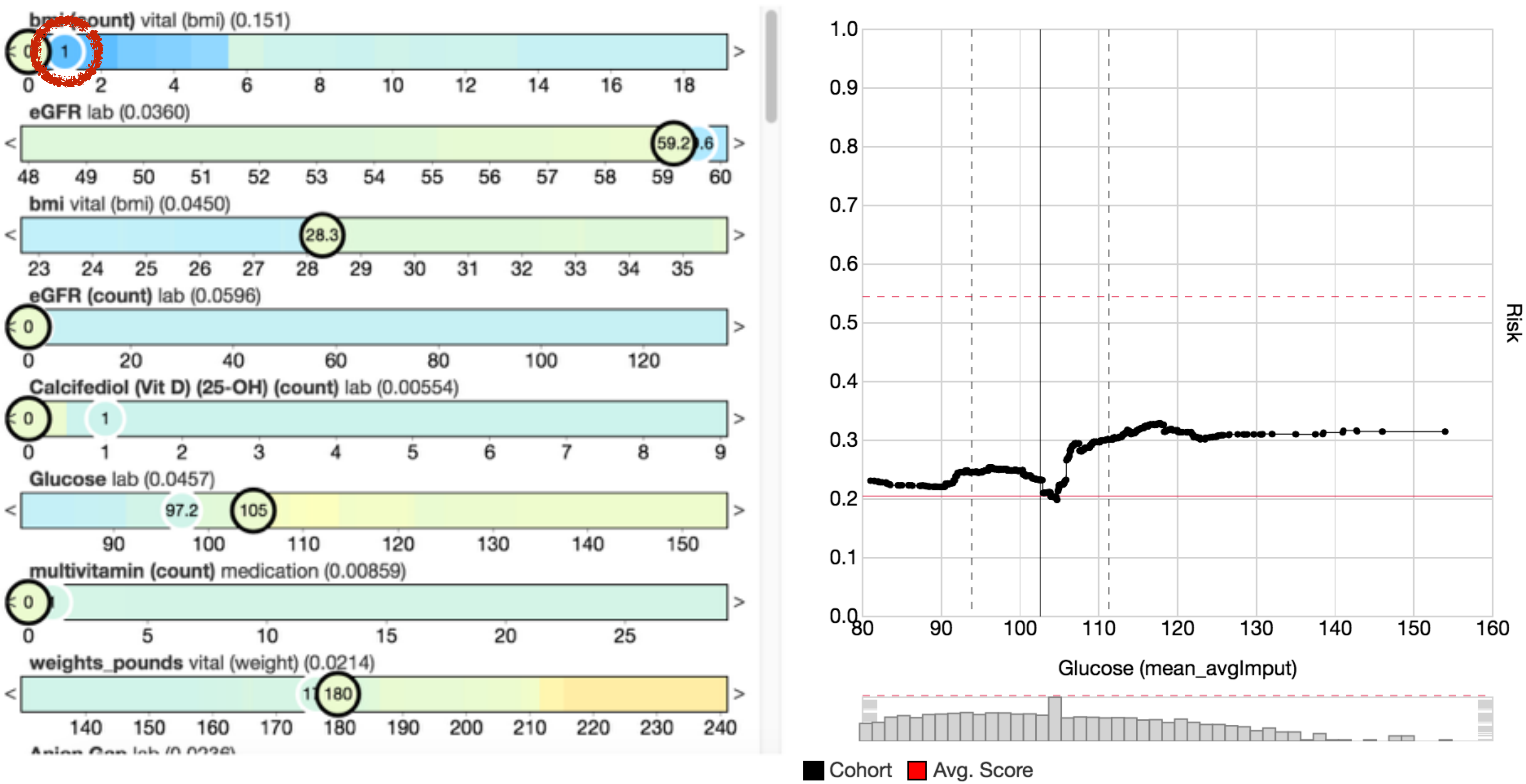
Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models
Josua Krause, Adam Perer, Kenney Ng – *CHI 2016*

Model Interaction



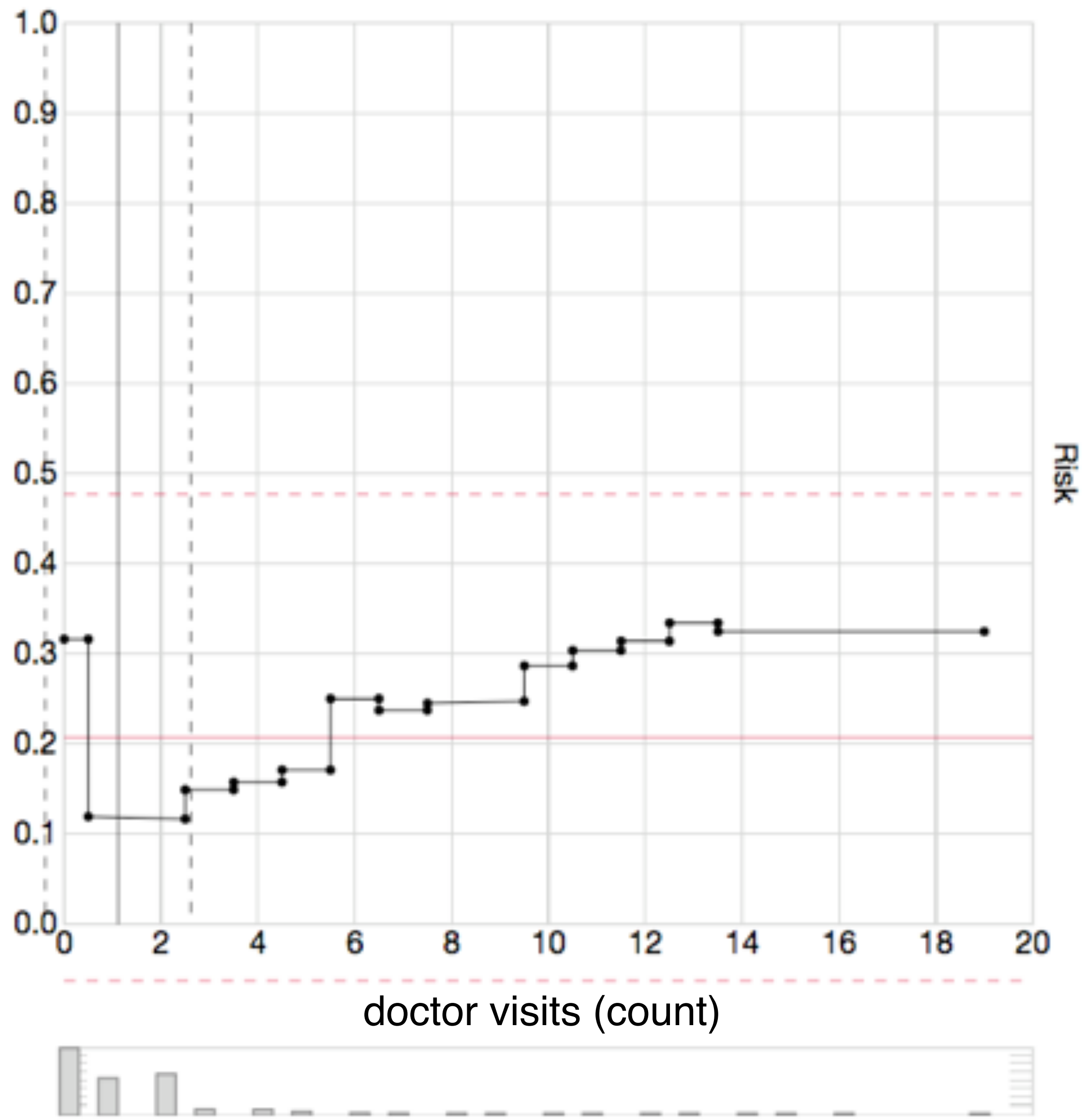
Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models
Josua Krause, Adam Perer, Kenney Ng – *CHI 2016*

Model Interaction



Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models
Josua Krause, Adam Perer, Kenney Ng – *CHI 2016*

doctor visits



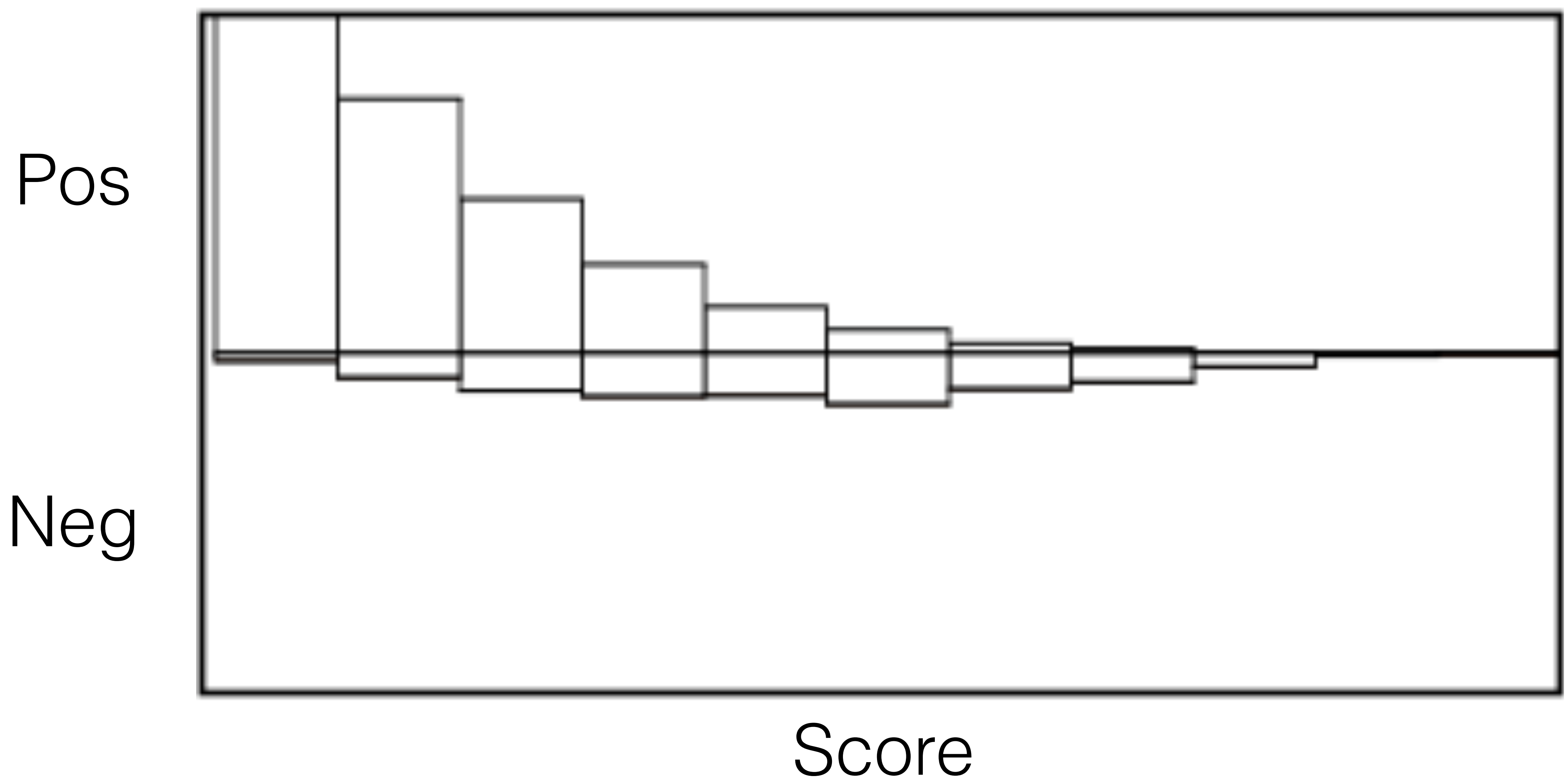
Visual Analytics

Model Output

Model Interaction

Model Output

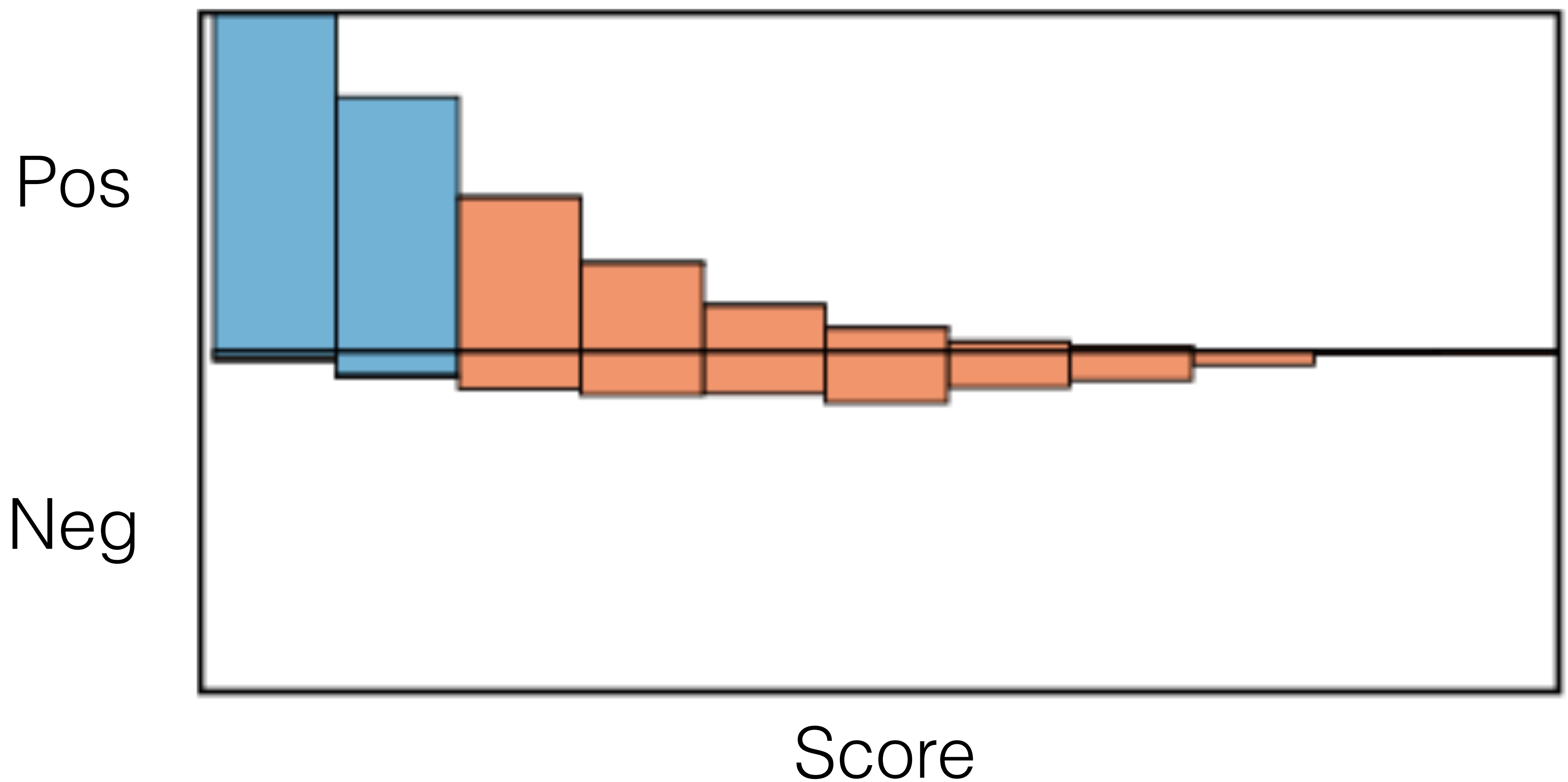
Distribution of Items wrt. Prediction Score



Using Visual Analytics to Interpret Predictive Machine Learning Models
Josua Krause, Adam Perer, Enrico Bertini – *2016 ICML Workshop on Human Interpretability in Machine Learning*

Model Output

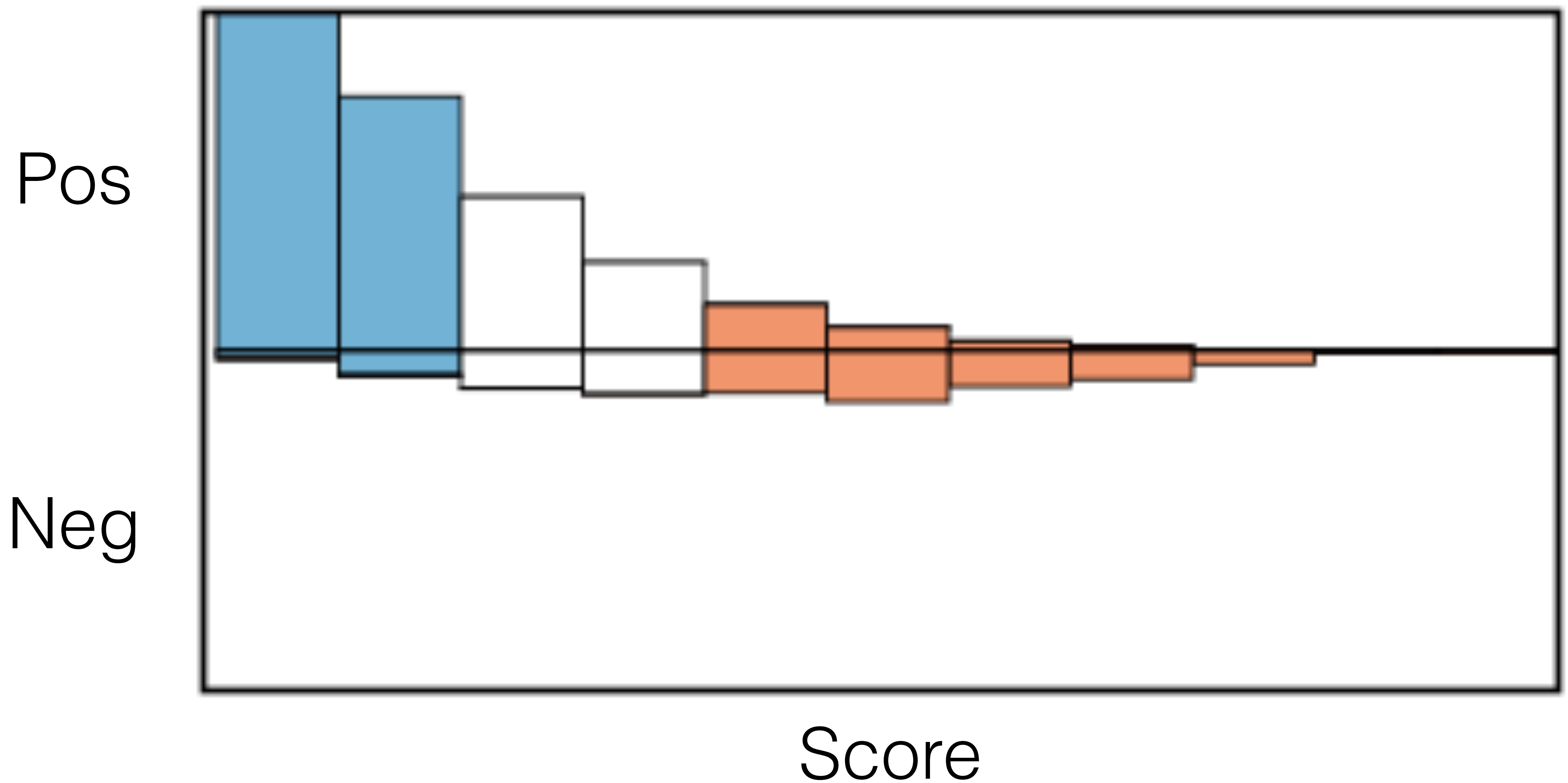
Distribution of Items wrt. Prediction Score



Using Visual Analytics to Interpret Predictive Machine Learning Models
Josua Krause, Adam Perer, Enrico Bertini – *2016 ICML Workshop on Human Interpretability in Machine Learning*

Model Output

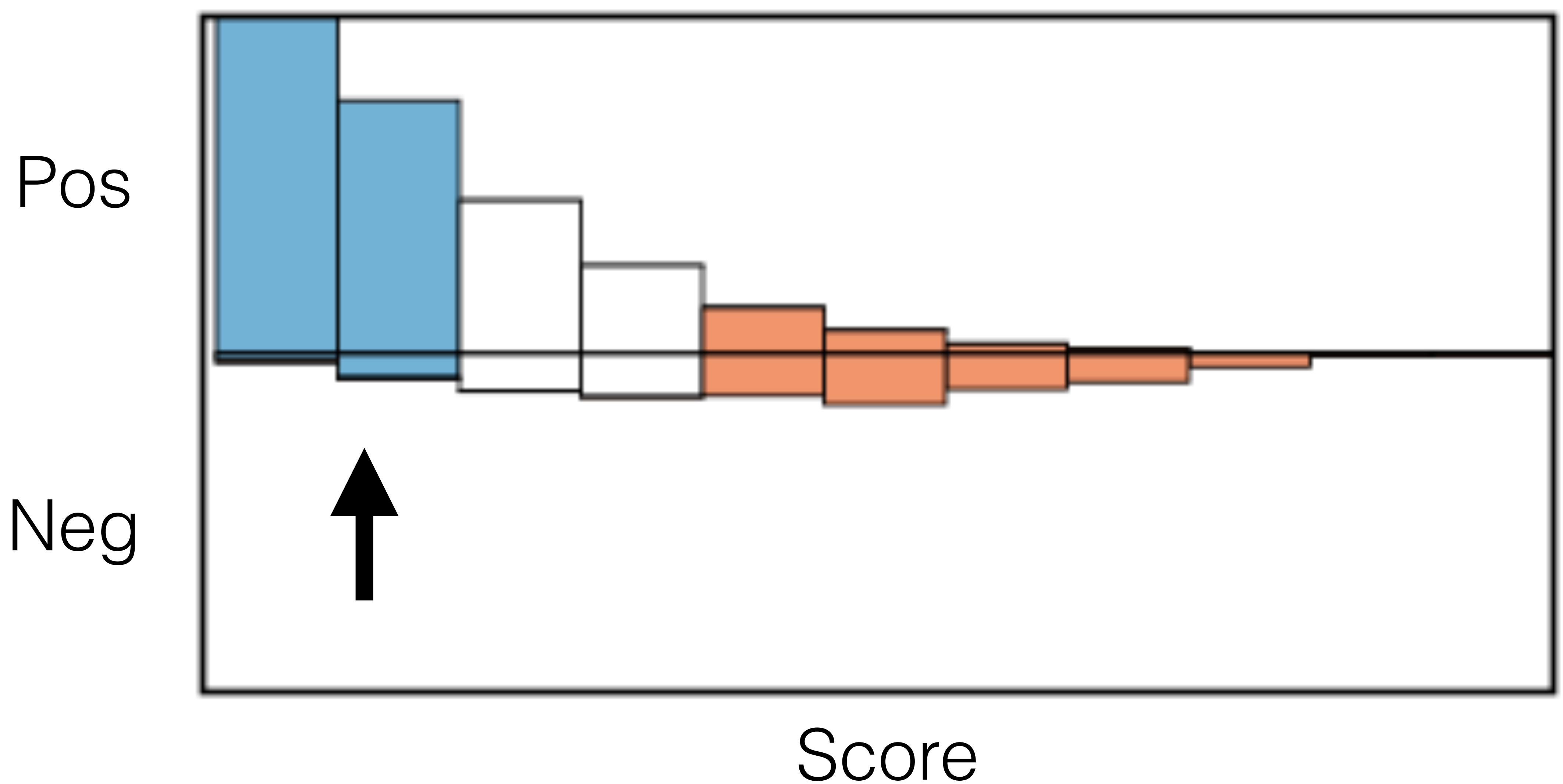
Distribution of Items wrt. Prediction Score



Using Visual Analytics to Interpret Predictive Machine Learning Models
Josua Krause, Adam Perer, Enrico Bertini – *2016 ICML Workshop on Human Interpretability in Machine Learning*

Model Output

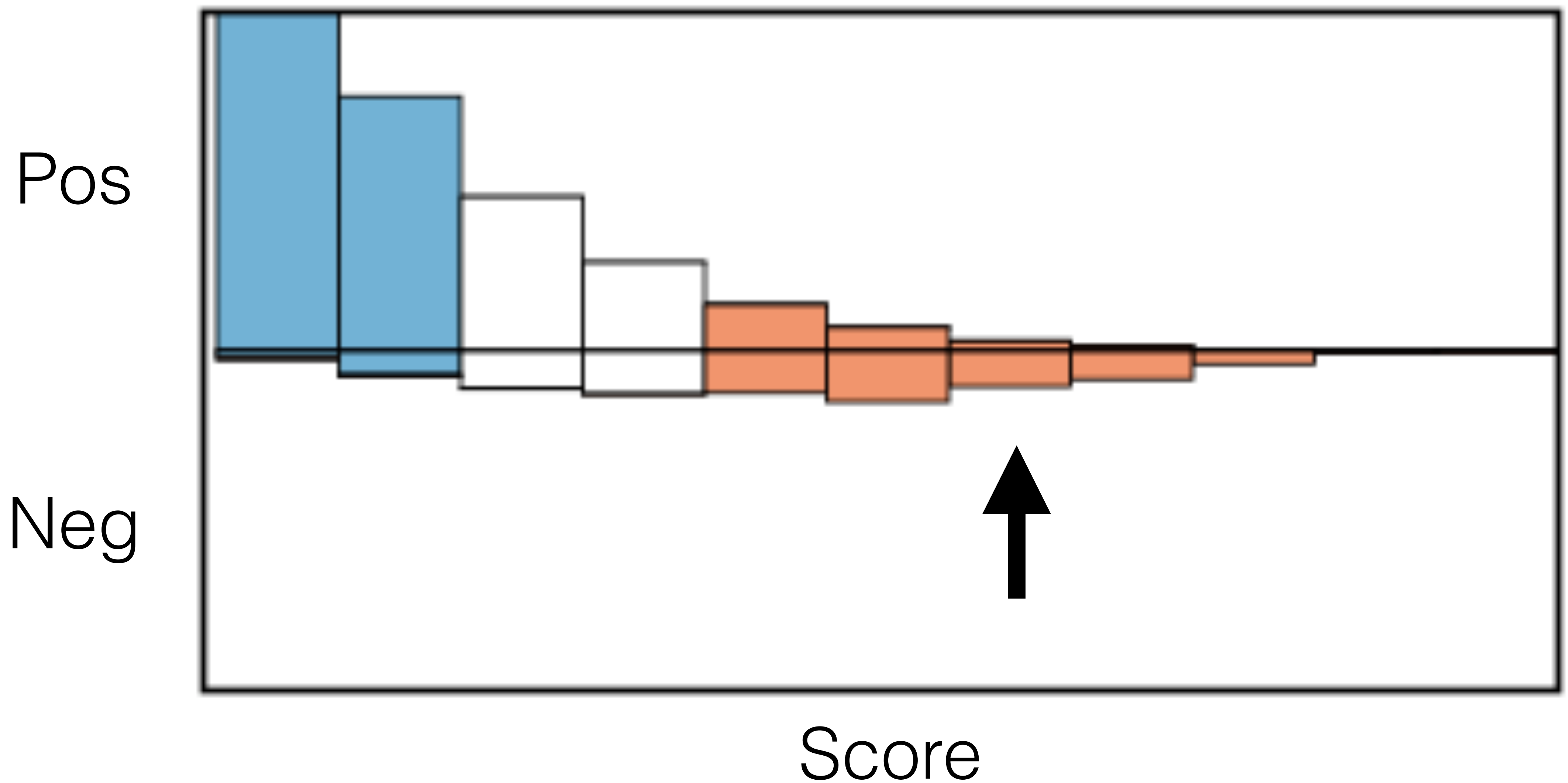
Distribution of Items wrt. Prediction Score



Using Visual Analytics to Interpret Predictive Machine Learning Models
Josua Krause, Adam Perer, Enrico Bertini – *2016 ICML Workshop on Human Interpretability in Machine Learning*

Model Output

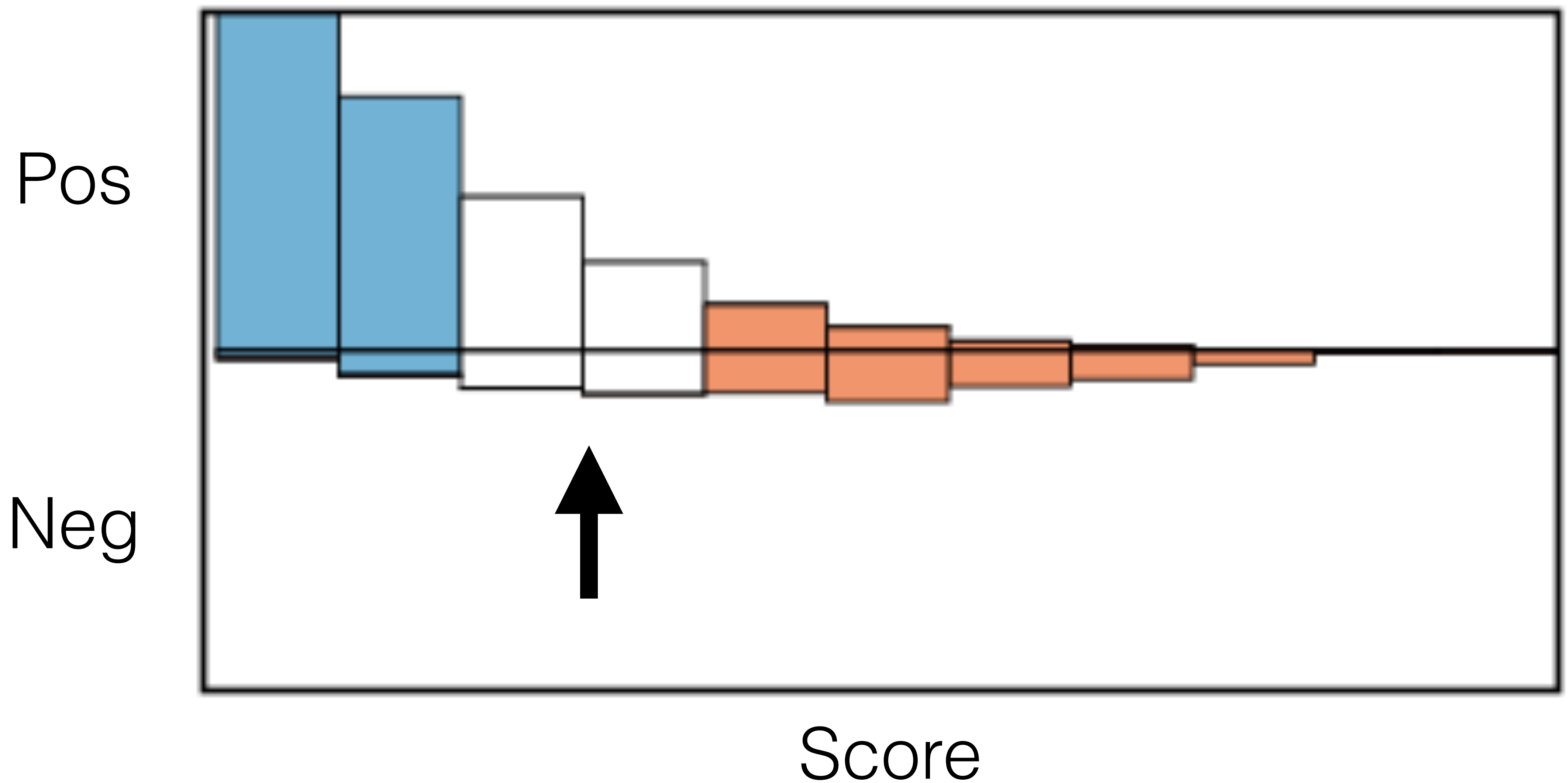
Distribution of Items wrt. Prediction Score



Using Visual Analytics to Interpret Predictive Machine Learning Models
Josua Krause, Adam Perer, Enrico Bertini – *2016 ICML Workshop on Human Interpretability in Machine Learning*

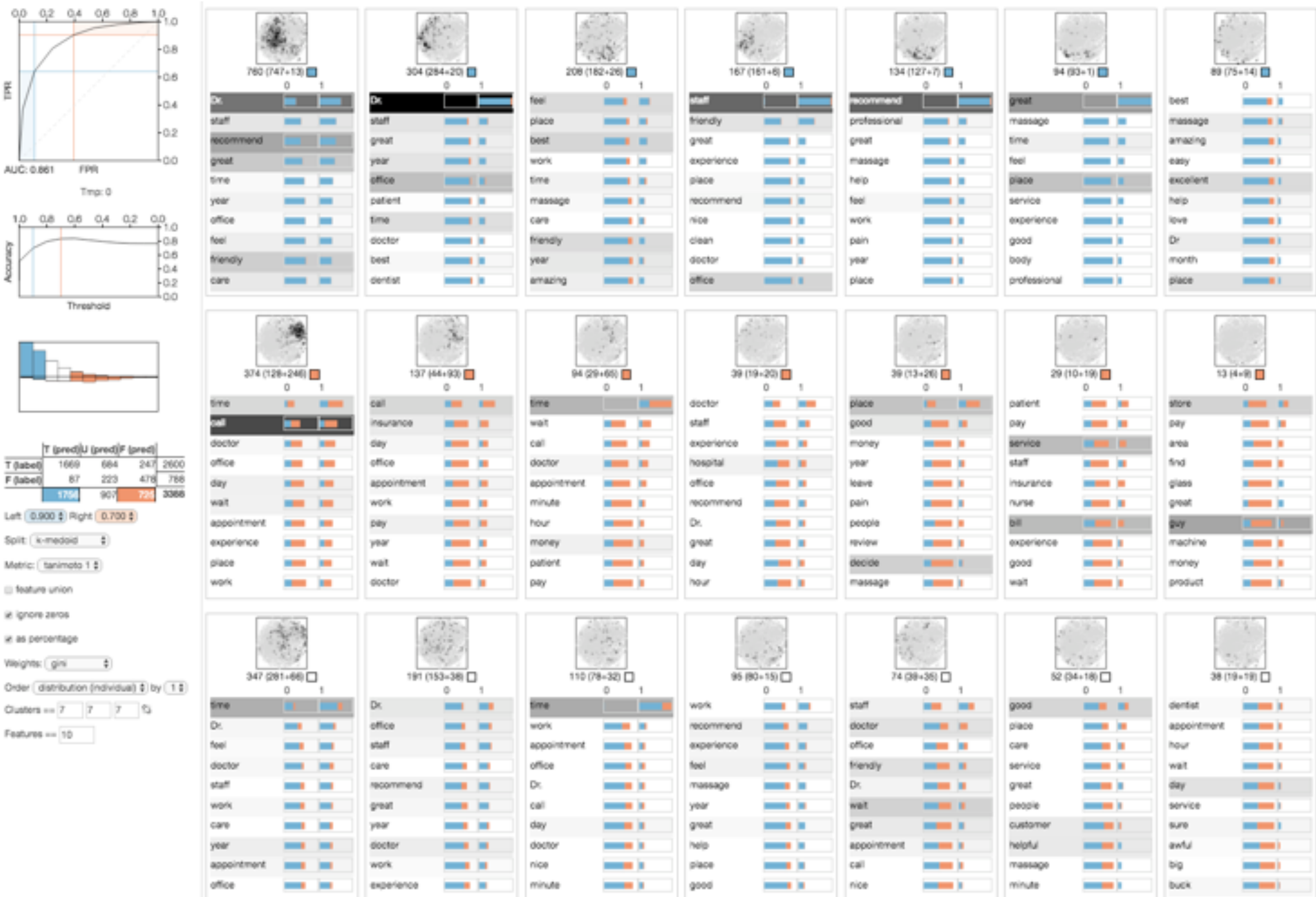
Model Output

Distribution of Items wrt. Prediction Score



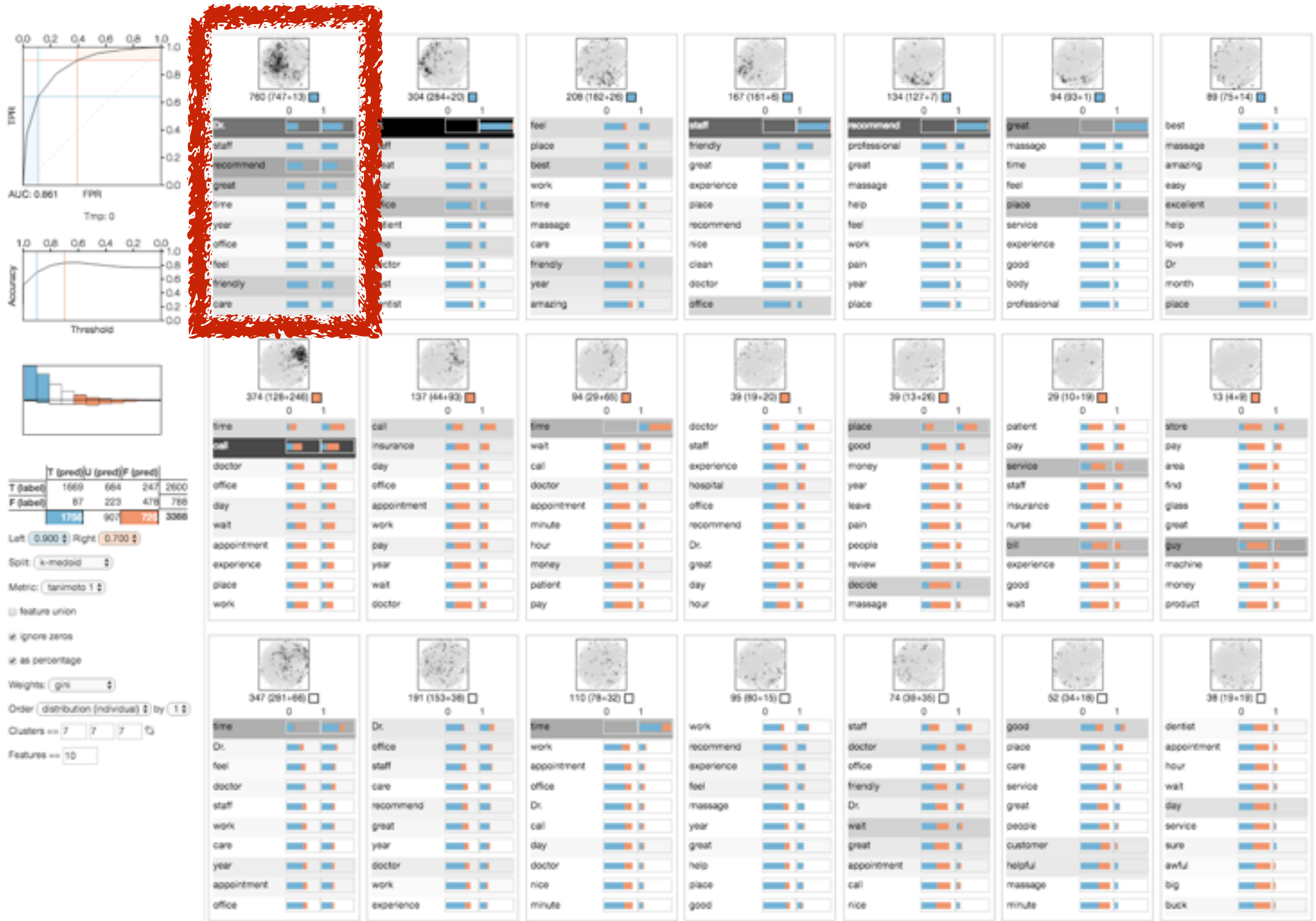
Using Visual Analytics to Interpret Predictive Machine Learning Models
Josua Krause, Adam Perer, Enrico Bertini – *2016 ICML Workshop on Human Interpretability in Machine Learning*

Item Subsets



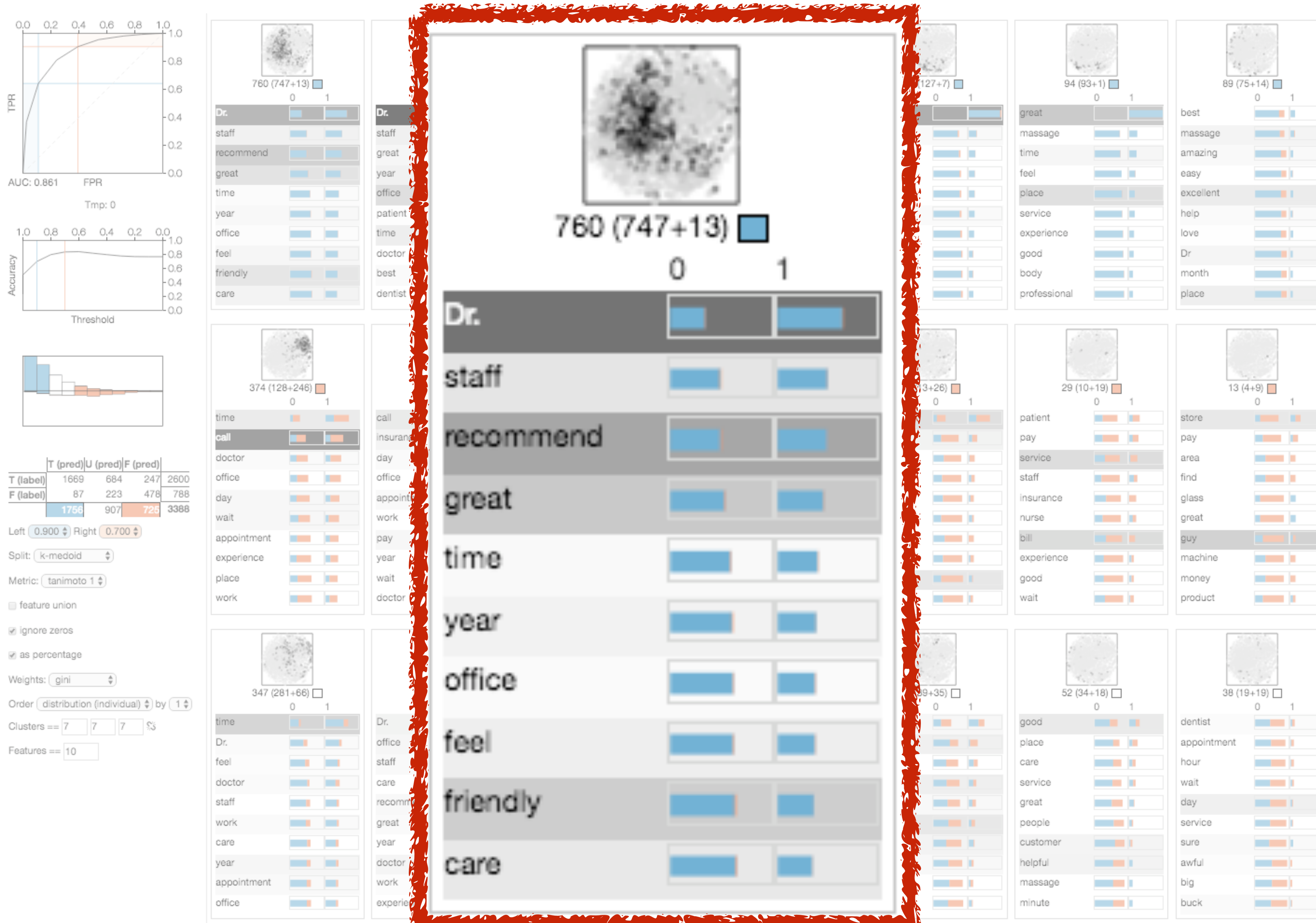
Using Visual Analytics to Interpret Predictive Machine Learning Models
Josua Krause, Adam Perer, Enrico Bertini – *2016 ICML Workshop on Human Interpretability in Machine Learning*

Item Subsets

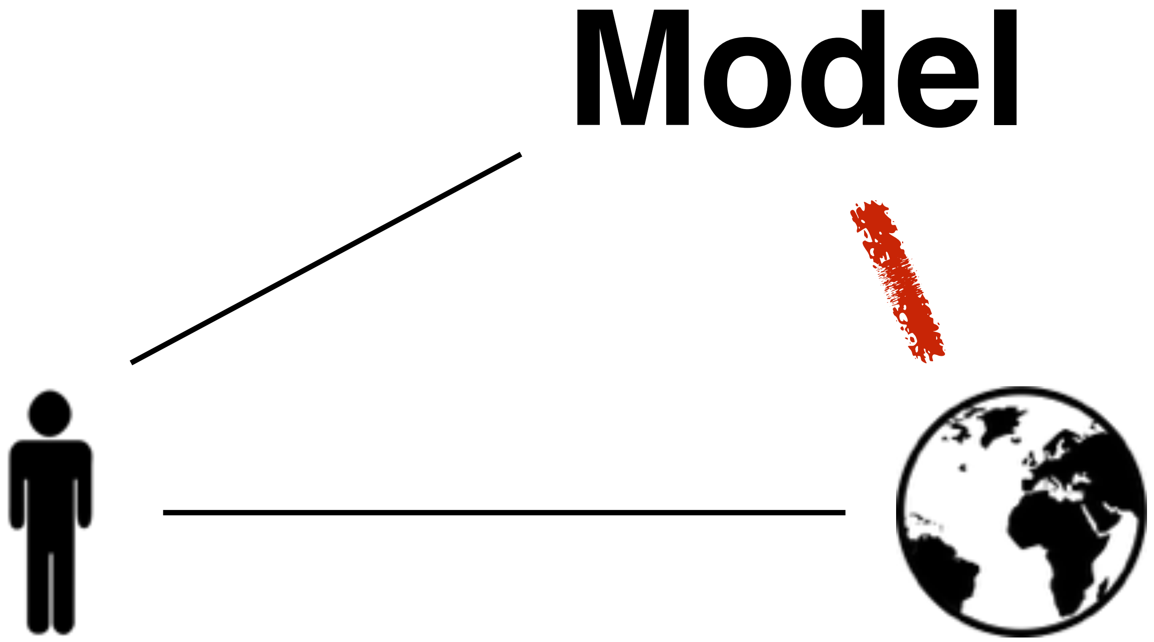


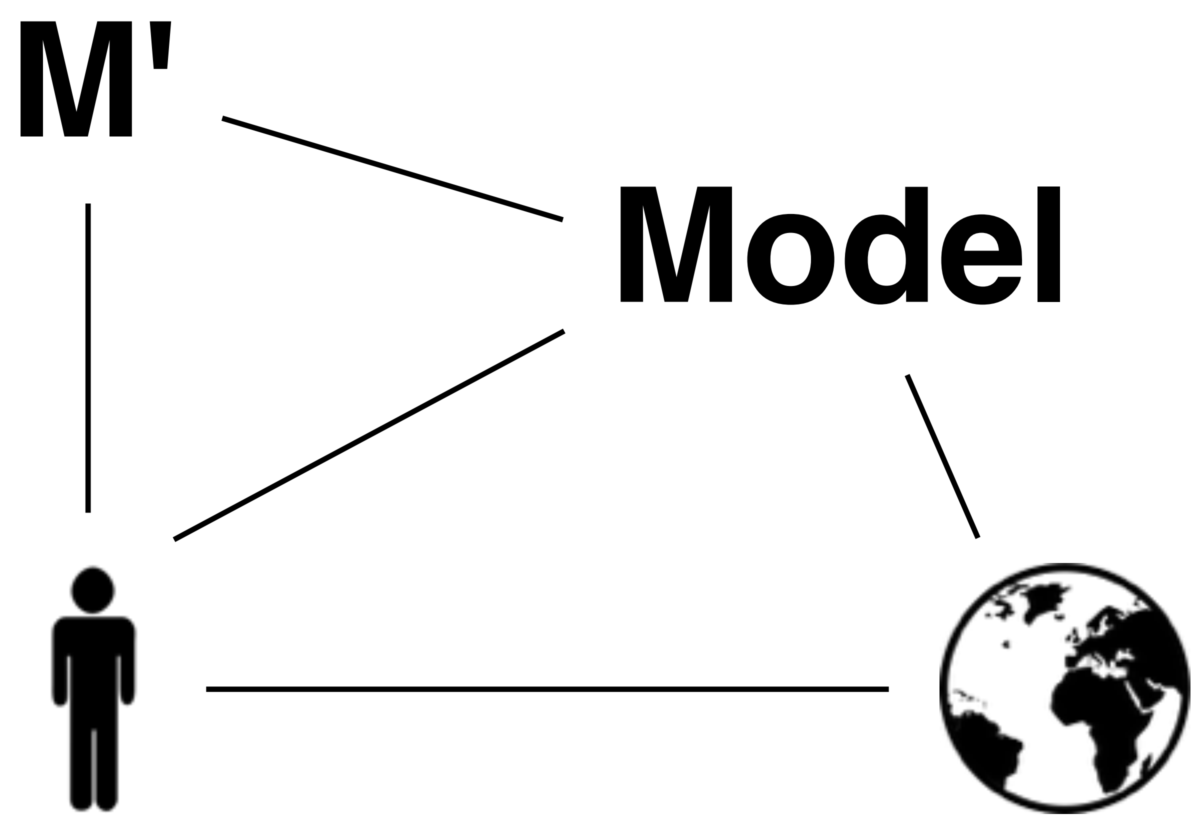
Using Visual Analytics to Interpret Predictive Machine Learning Models
Josua Krause, Adam Perer, Enrico Bertini – *2016 ICML Workshop on Human Interpretability in Machine Learning*

Item Subsets

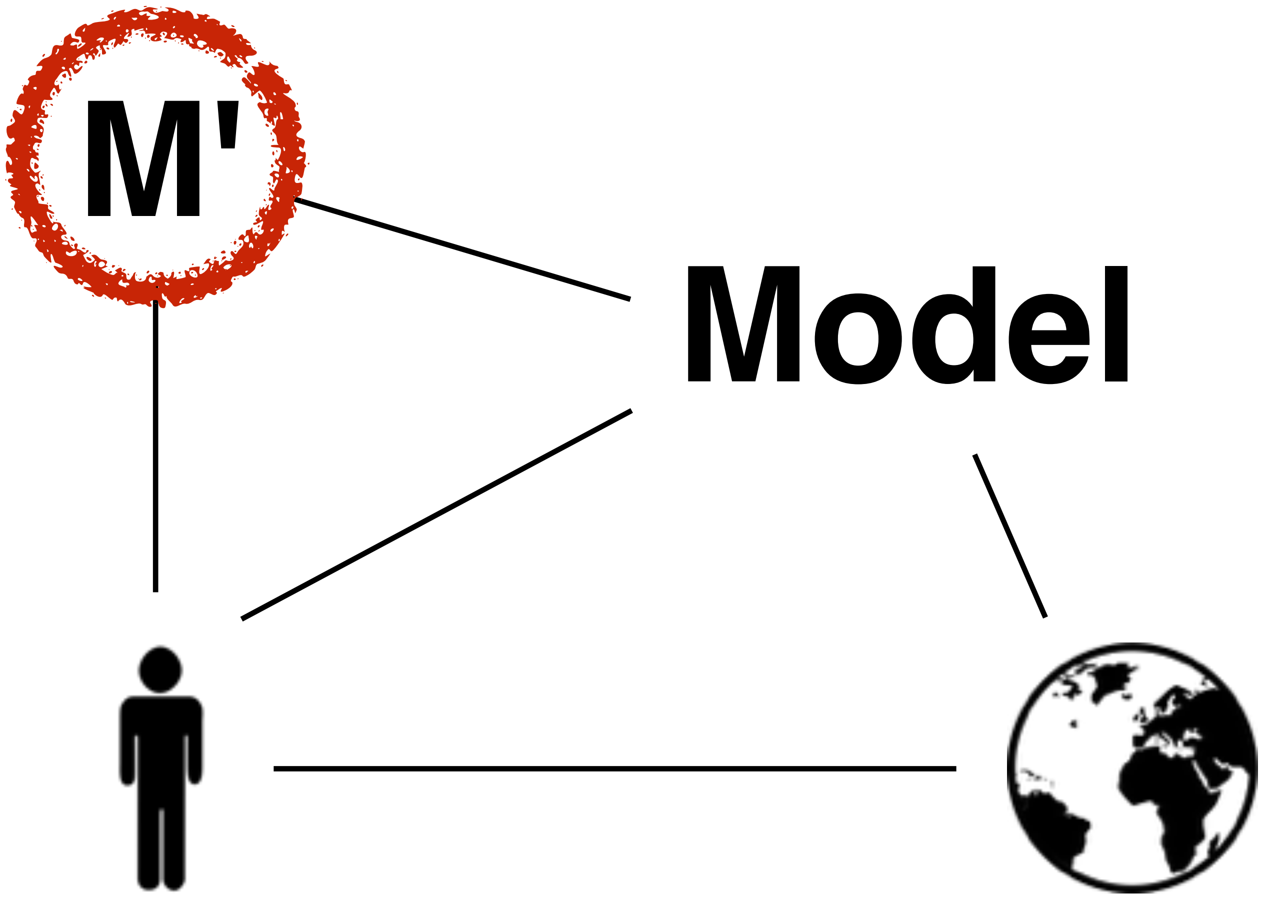


Using Visual Analytics to Interpret Predictive Machine Learning Models
Josua Krause, Adam Perer, Enrico Bertini – *2016 ICML Workshop on Human Interpretability in Machine Learning*

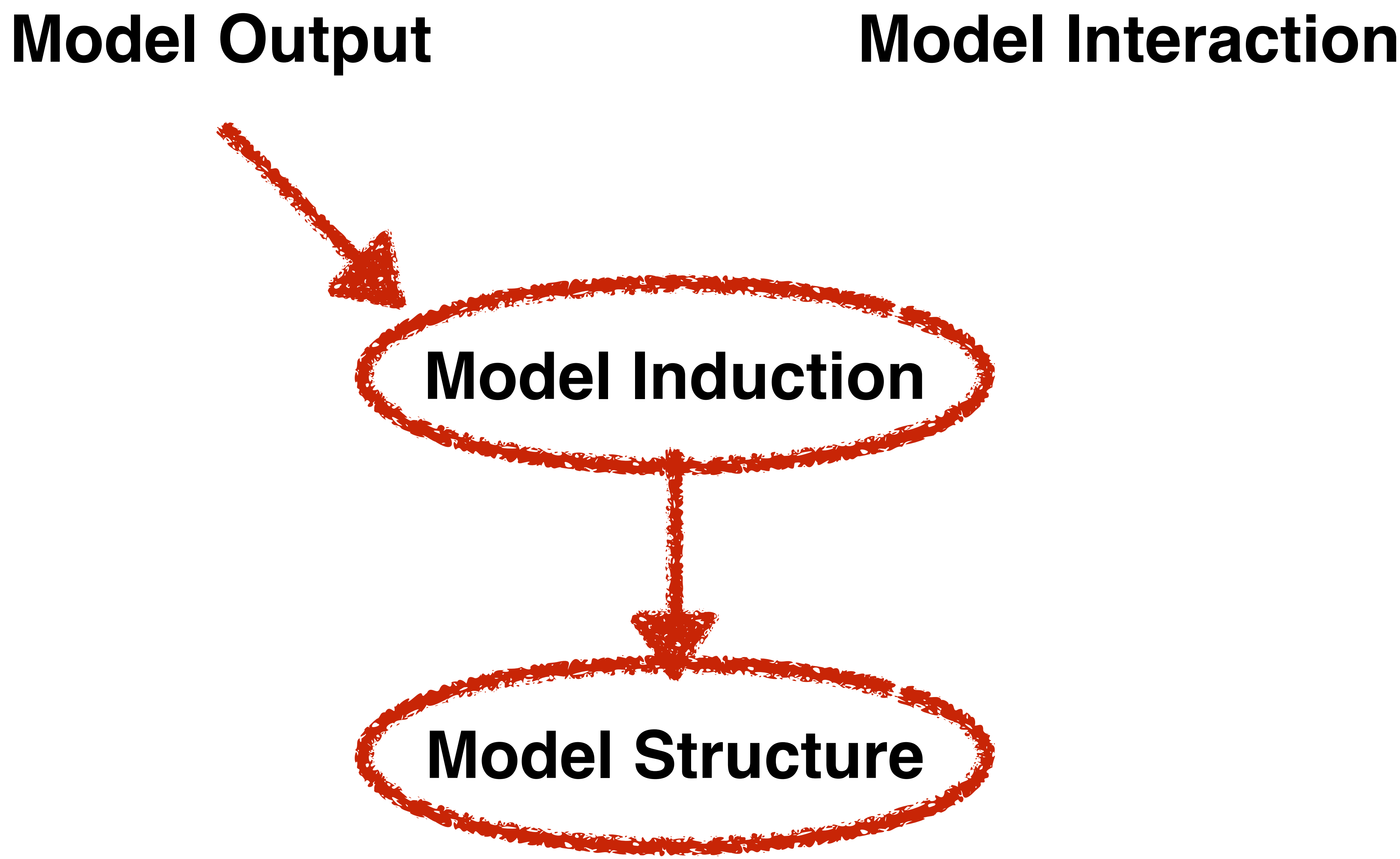

Model


M'
Model

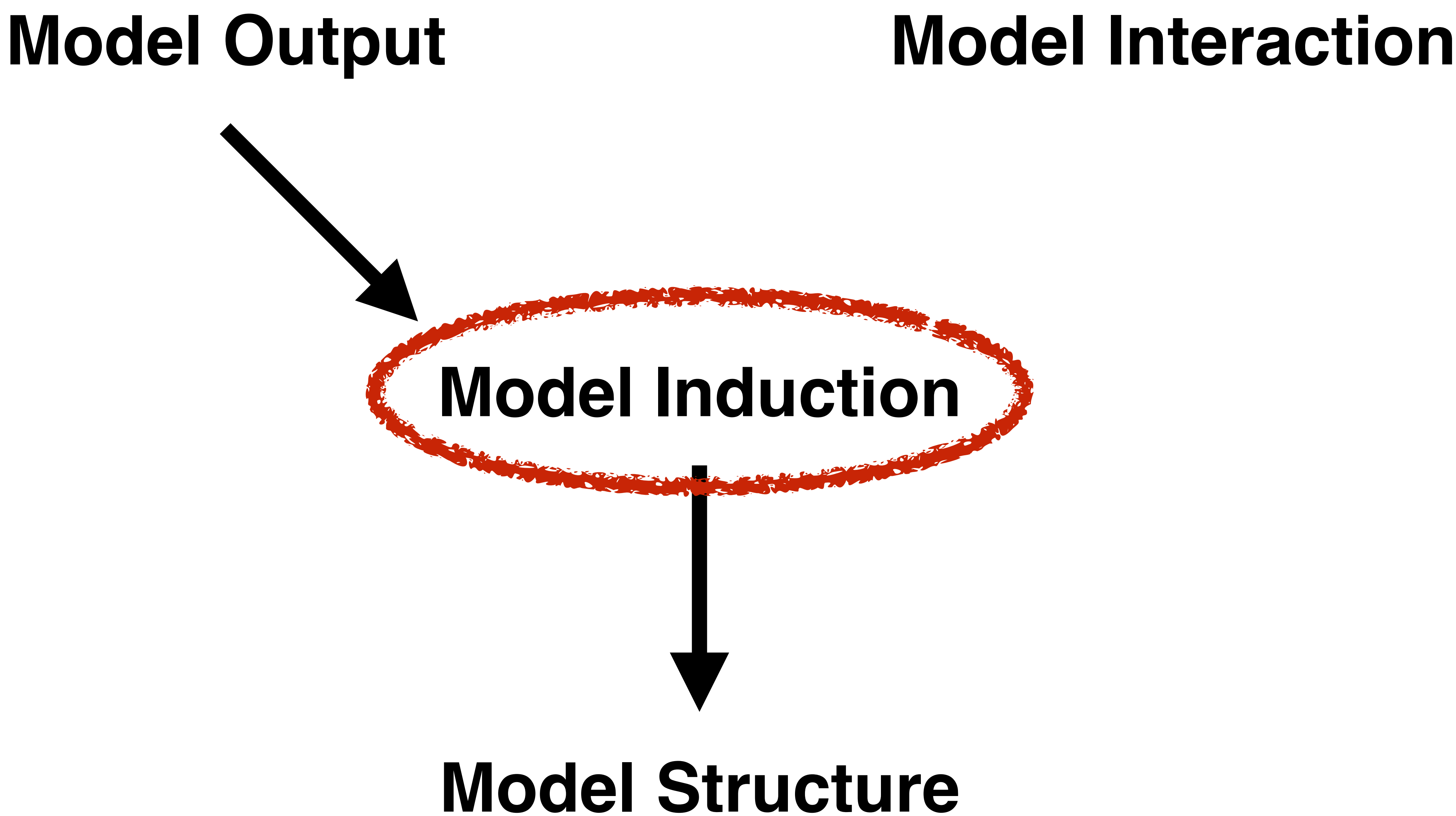
Induced Model



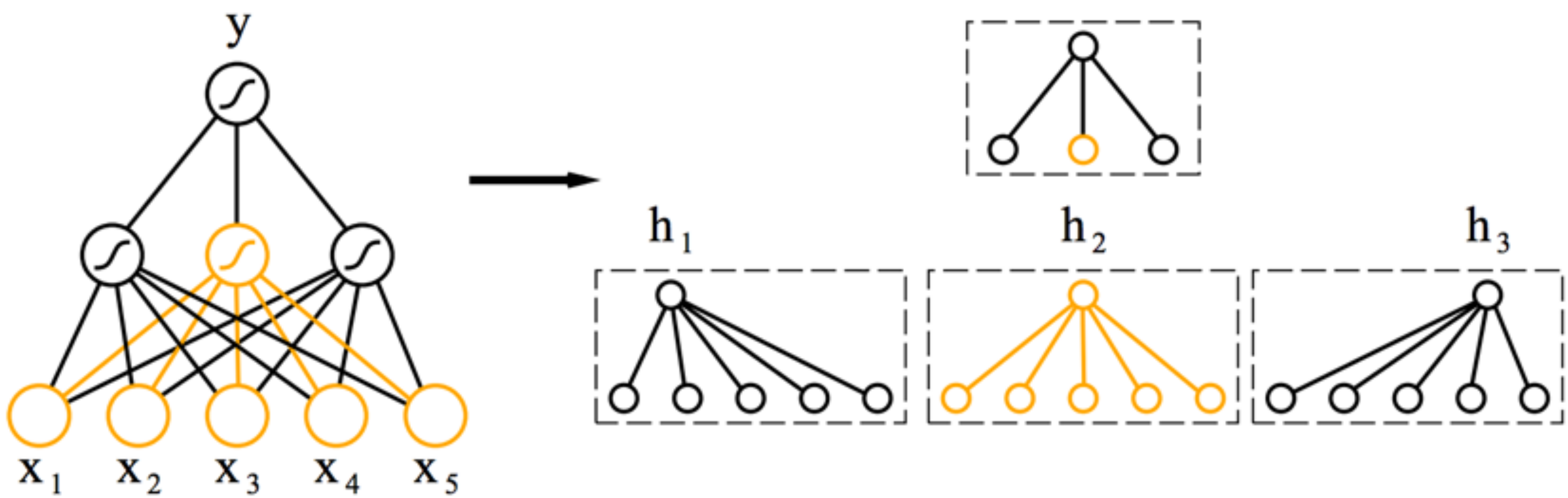
Visual Analytics



Visual Analytics



Model Induction



extracted rules:

$$
\begin{aligned}
y &\leftarrow h_1 \vee h_2 \vee h_3 \\
h_1 &\leftarrow x_1 \wedge x_2 \\
h_2 &\leftarrow x_2 \wedge x_3 \wedge x_4 \\
h_3 &\leftarrow x_5
\end{aligned}
$$

Using Neural Networks for Data Mining
Mark Craven, Jude Shavlik – *Future Generation Computer Systems 1997*

Visual Analytics

Model Output

Model Interaction

Model Induction

Model Structure

Code and Internal State

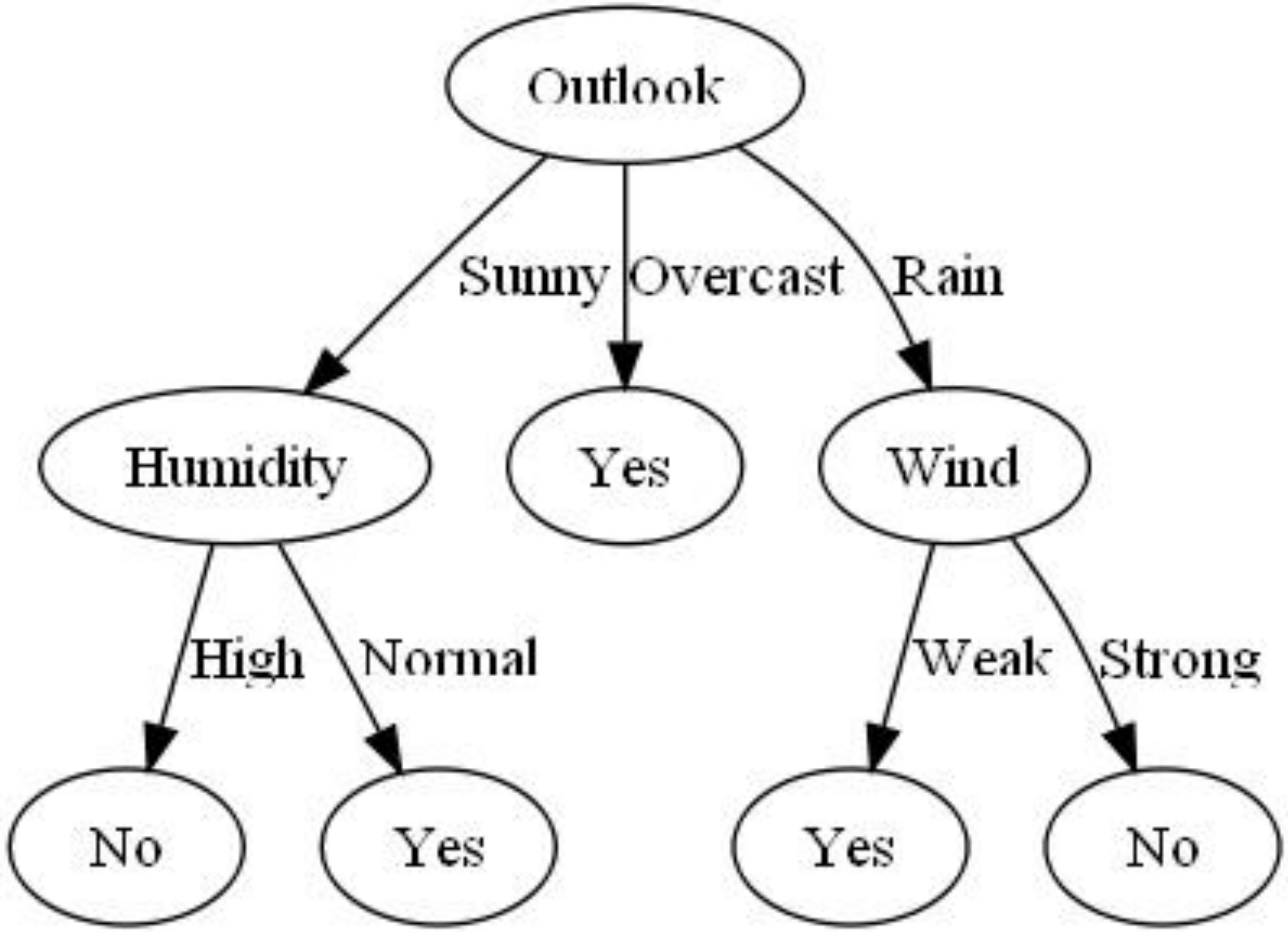
```
with nogil:
    for i in range(n_samples):
        node = self.nodes
        # While node not a leaf
        while node.left_child != _TRE
            # ... and node.right_chi
            if X_ptr[X_sample_stride
                    X_fx_stride * no
                node = &self.nodes[no
            else:
                node = &self.nodes[no

        out_ptr[i] = <SIZE_t>(node -
```

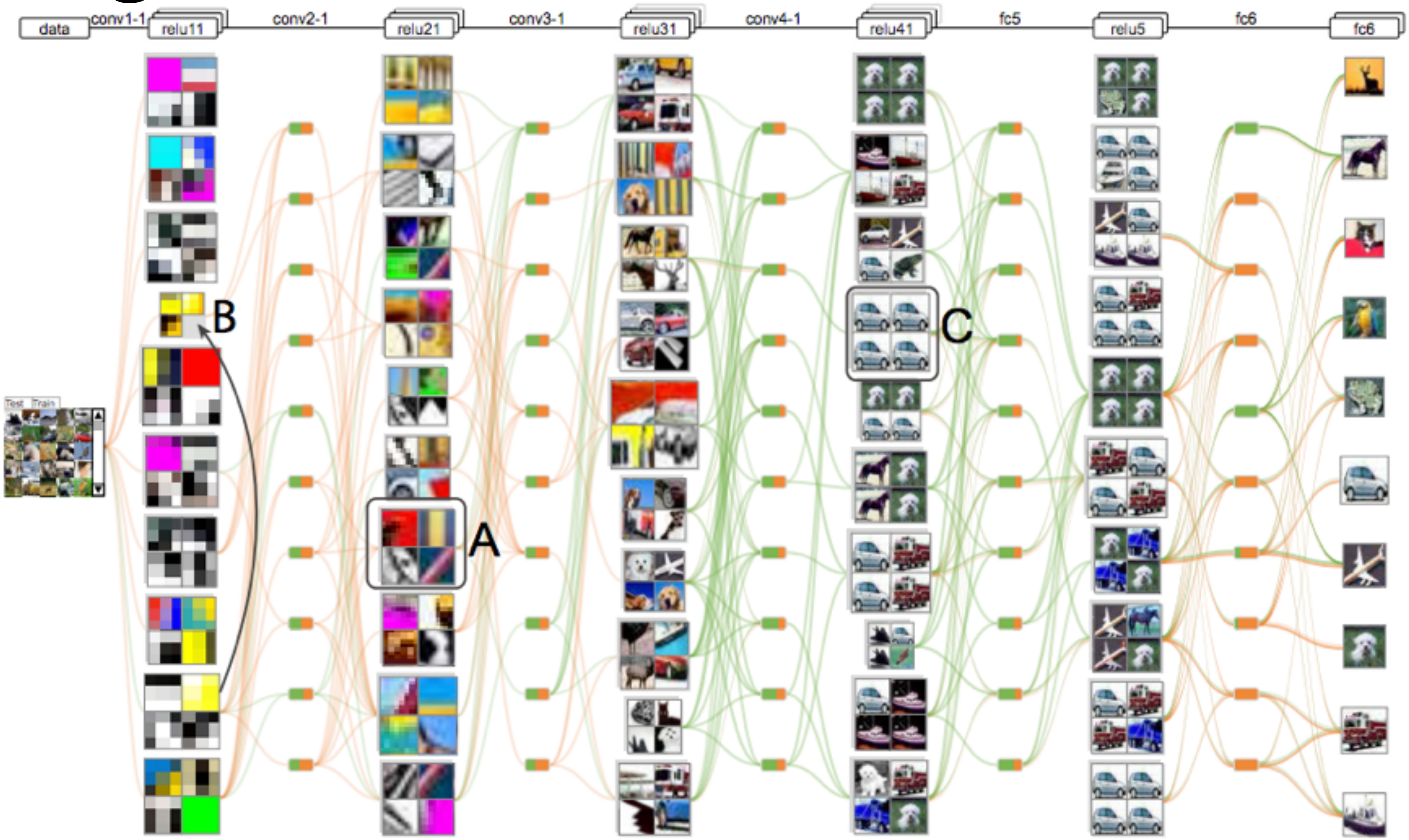
Code from scikit-learn tree implementation

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 2
39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 5
-1, -1, -1, -1, -1, -1, -1, -1, 83, 84, 85, 86, -1, -1, -1, -1, -1, 92, -1, -1
113, 114, 115, 116, 117, 118, -1, -1, -1, -1, 123, 124, -1, -1, -1, 12
-1, 146, -1, -1, -1, -1, -1, -1, -1, -1, 155, 156, -1, -1, 159, -1, -1, 16
179, 180, 181, 182, 183, -1, -1, 186, -1, -1, 189, -1, -1, -1, -1, -1, 19
211, 212, 213, 214, 215, 216, 217, 218, -1, -1, -1, -1, -1, -1, -1, -1
-1, -1, 246, -1, -1, 249, 250, 251, 252, 253, 254, 255, 256, -1, 258,
-1, 278, 279, -1, -1, -1, 283, 284, 285, 286, -1, 288, -1, -1, -1, -1, -
-1, 310, -1, -1, 313, 314, 315, -1, -1, -1, -1, 320, -1, -1, -1, 324, 325
-1, 344, 345, -1, -1, -1, 349, 350, -1, -1, -1, 354, 355, -1, -1, 358, -1
-1, -1, -1, -1, 380, 381, -1, -1, -1, 385, 386, 387, 388, 389, 390, 391
-1, 409, -1, -1, -1, 413, -1, 415, -1, -1, 418, -1, -1, 421, 422, 423,
441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 4
-1, 473, -1, -1, 476, 477, -1, -1, -1, 481, 482, 483, -1, -1, -1, 487, -1
-1, -1, -1, 508, 509, -1, -1, -1, -1, 514, -1, -1, 517, 518, -1, -1, -1, 52
539, 540, 541, 542, 543, 544, -1, -1, -1, 548, -1, -1, -1, -1, 553, -1,
572, 573, 574, -1, -1, -1, -1, 579, 580, 581, 582, 583, 584, 585, 58
-1, -1, -1, -1, -1, 605, -1, -1, 608, -1, -1, -1, 612, -1, -1, 615, 616, 6
-1, -1, 636, 637, 638, 639, 640, -1, -1, -1, -1, -1, -1, -1, 648, 649, 6
-1, -1, 669, 670, 671, 672, -1, 674, -1, -1, -1, -1, -1, -1, 681, -1, -1, -
699, 700, 701, 702, 703, 704, 705, 706, 707, 708, 709, -1, -1, -1, -
-1, -1, 733, -1, -1, 736, -1, -1, -1, 740, -1, -1, 743, -1, -1, 746, 747,
-1, -1, -1, 767, 768, 769, 770, 771, -1, -1, -1, 775, -1, -1, 778, -1,
795, 796, 797, 798, 799, 800, 801, 802, 803, 804, 805, 806, 807, 8
-1, -1, -1, -1, -1, -1, -1, -1, 830, 831, 832, 833, -1, 835, -1, -1, -1, -
-1, -1, -1, -1, 861, -1, -1, 864, -1, -1, 867, 868, -1, -1, -1, 872, 873,

Algorithm and Internal State



Algorithm and Internal State



Towards Better Analysis of Deep Convolutional Neural Networks
Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, Shixia Liu – *VAST 2016*

Visual Analytics

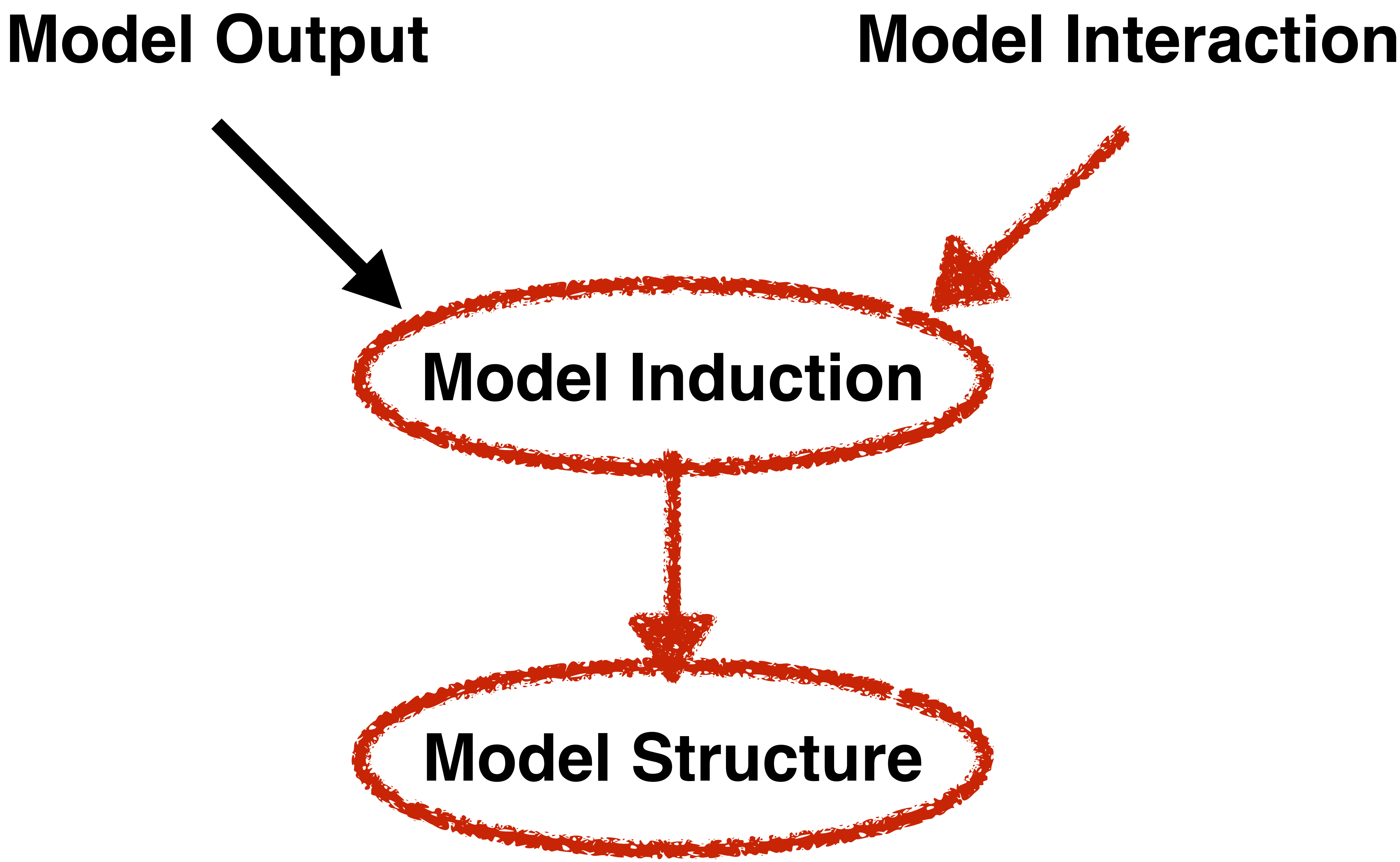
Model Output

Model Interaction

Model Induction

Model Structure

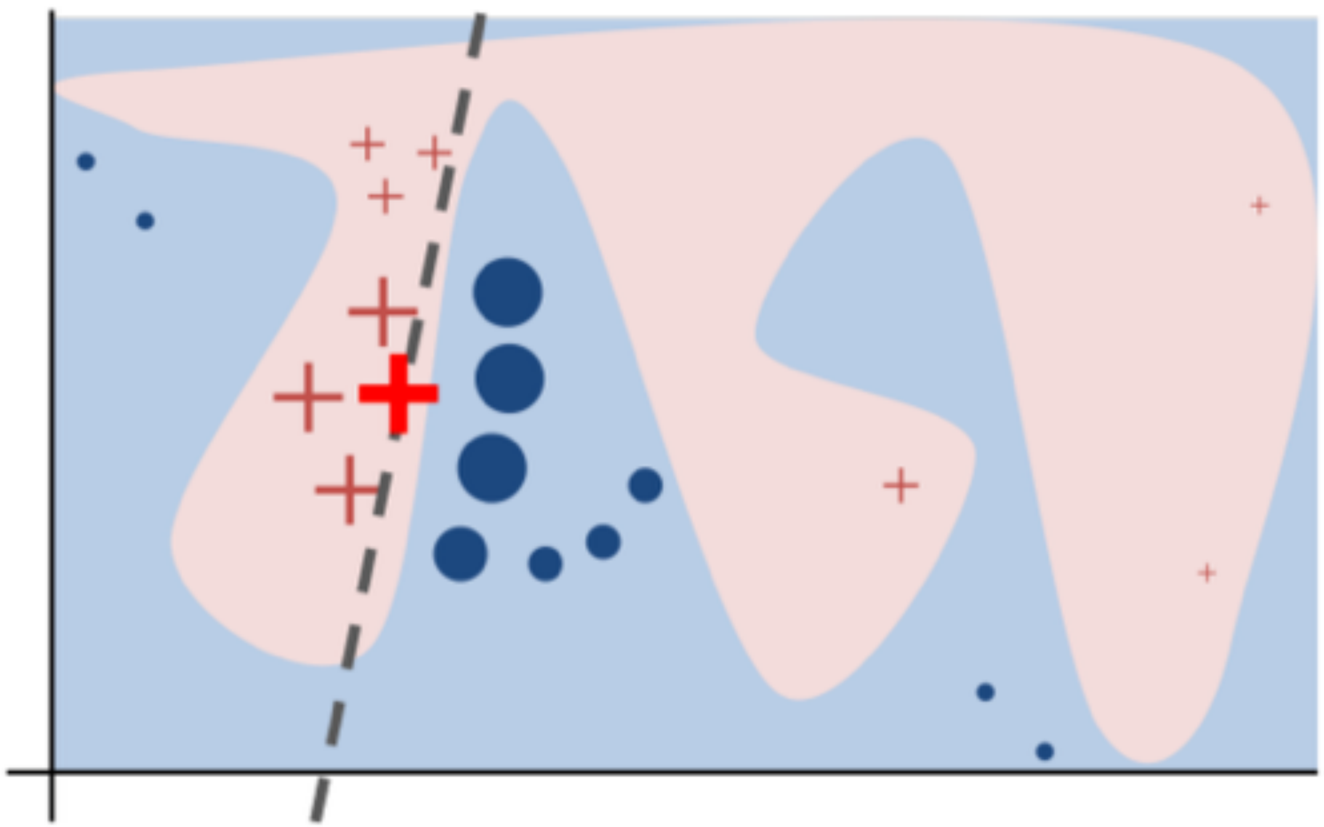
Visual Analytics

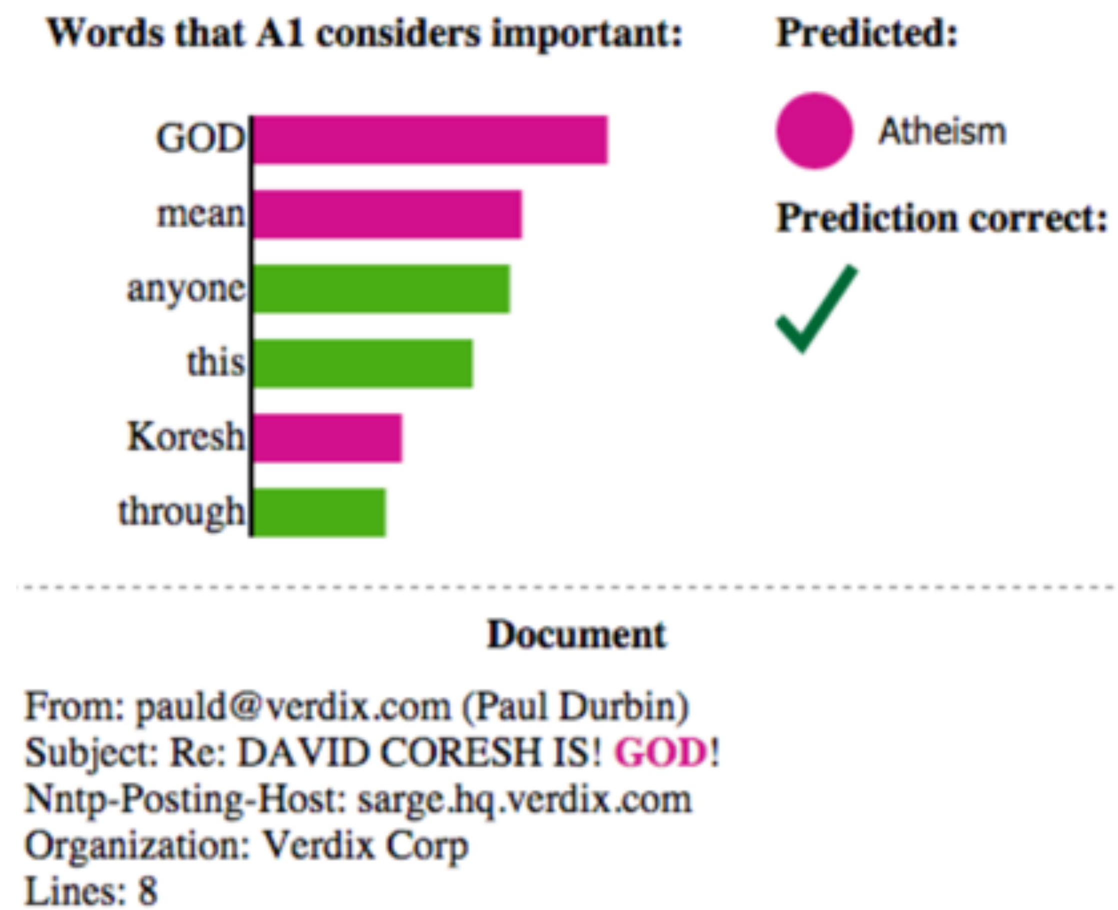


Model Interaction

LIME

Local Interpretable Model-agnostic Explanations





"Why Should I Trust You?" Explaining the Predictions of Any Classifier
Marco Riberio, Sameer Singh, Carlos Guestrin – KDD 2016

Visual Analytics

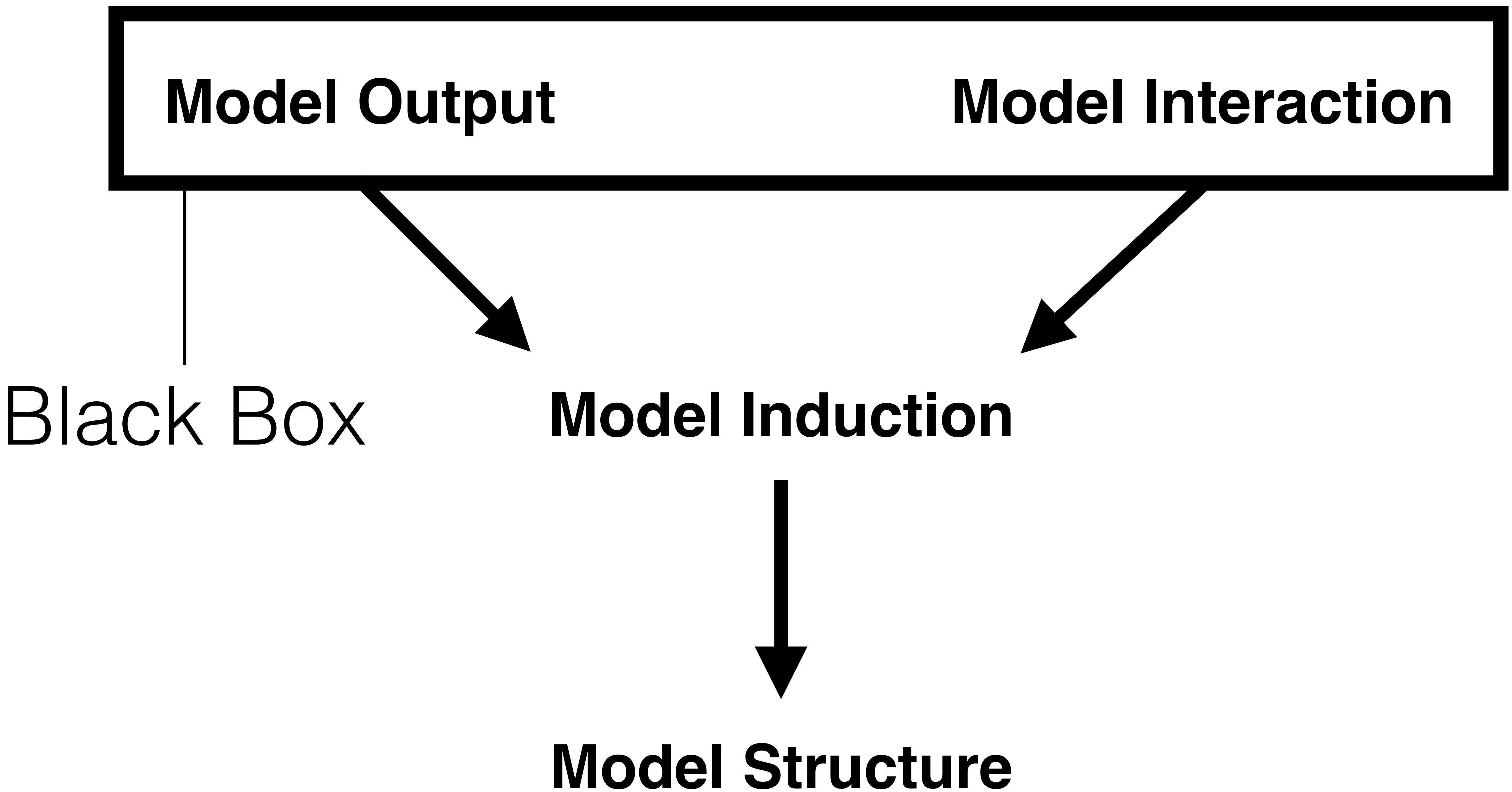
Model Output

Model Interaction

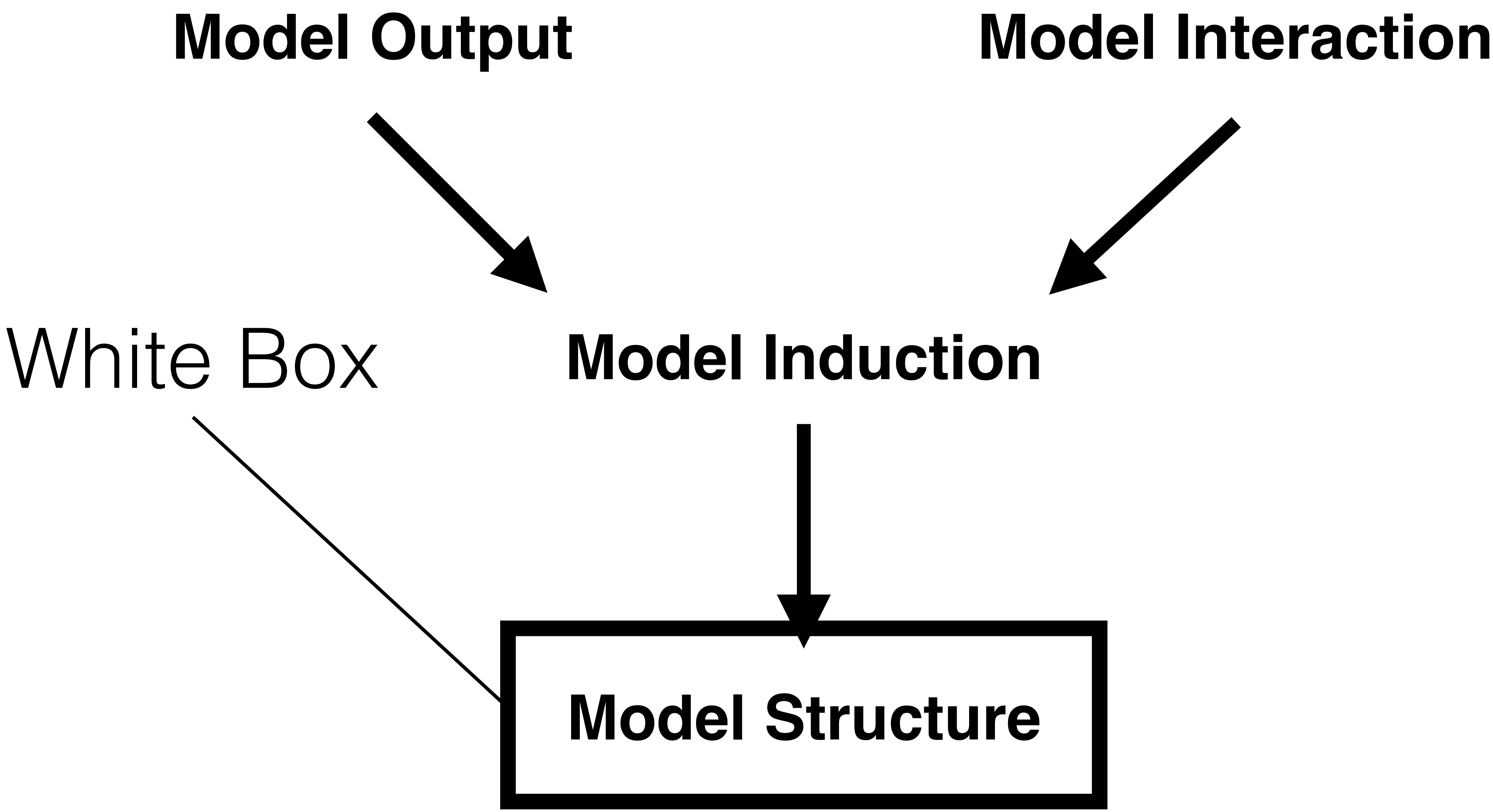
Model Induction

Model Structure

Visual Analytics



Visual Analytics



Reflections

- More black box explanations
    - Generalizable
    - Current strategies are still far from optimal
- Gray box explanations
    - Leveraging feature transformations / convolution
- Methods from ML community don't utilize Visual Analytics
- Generalize over data types
- How to deal with non-interpretable or inferred features?
- Item level explanation --> Population level explanation

The Mythos of Model Interpretability

Zachary C. Lipton ZLIPTON@CS.UCSI
University of California, San Diego 9500 Gilman Drive, La Jolla, CA 92093 USA

Abstract

Supervised machine learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? We want models to be not only good, but interpretable. And yet the task of *interpretation* appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable. Despite this ambiguity, many papers proclaim interpretability axiomatically, absent further explanation. In this paper, we seek to refine the discourse on interpretability. First, we examine the motivations underlying interest in interpretability, finding them to be diverse and occasionally discordant. Then, we address model properties and techniques thought to confer interpretability, identifying transparency to humans and post-hoc explanations as competing notions. Throughout, we discuss the feasibility and desirability of different notions, and question the oft-made assertions that linear models are interpretable and that deep neural networks are not.

the literature suggests the latter to be the case. I
motives for interpretability and the technical desc
of interpretable models are diverse and occasiona
cordant, suggesting that *interpretability* refers to m
one concept. In this paper, we seek to clarify both,
ing that *interpretability* is not a monolithic concep
fact reflects several distinct ideas. We hope, thro
critical analysis, to bring focus to the dialog.

Here, we consider supervised learning but not ot
chine learning paradigms, such as reinforcement l
and interactive learning. This scope derives from o
inal interest in the oft-made claim that linear mo
preferable to deep neural networks on account of i
terpretability (Lou et al., 2012). To gain conceptua
we ask the refining questions: *What is interpretabi
why is it important?* Broadening the scope of dis
seems counterproductive with respect to our aims.
search investigating interpretability in the context
forcement learning, we point to (Dragan et al., 2013
studies the human interpretability of robot actions.

To contextualize any definition of interpretability,
consider the motives that it addresses (expanded
Many papers motivate interpretability as a means to
der trust (Kim, 2015; Ridgeway et al., 1998). But w
cisely is trust? Some equate trust with understandin
others equate trust with confidence in a model's a

ICML 2016

Explanations Considered Harmful? User Interactions with Machine Learning Systems

puter

ht/license statement here. ACM now
cation options:
ds the copyright on the work. This is the

etain copyright, but ACM receives an
nse.
r(s) wish to pay for the work to be open
e must be paid to ACM.
to hold the appropriate release statement
n Verdana 7 point font. Please do not
x.
ned a unique DOI string to be included here.

Abstract

It has been suggested that the intelligibility of machine learning system behavior is an important factor in ensuring that users can identify that the system has erred, understand how the system operates and that thereby they are better able to provide appropriate feedback to the machine learning system to improve its accuracy. There has been increasing research into how to make machine learning intelligible to users without a background in AI, and it has been shown that providing explanations of a system's reasoning has many benefits. In this paper we review recent work in this area but also point to instances when explanations might have less desirable effects. Further work is warranted to understand how best to expose the reasoning of machine learning systems to improve their usability.

Author Keywords

Machine learning; explanations; reliability; intelligibility.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

CHI 2016

Thanks!

Model Output

Model Interaction

Model Induction

Model Structure

Slides at http://bit.ly/2eIyP8R

Josua Krause[*], Aritra Dasgupta[+], Enrico Bertini[*]
[*]NYU, [+]PNNL