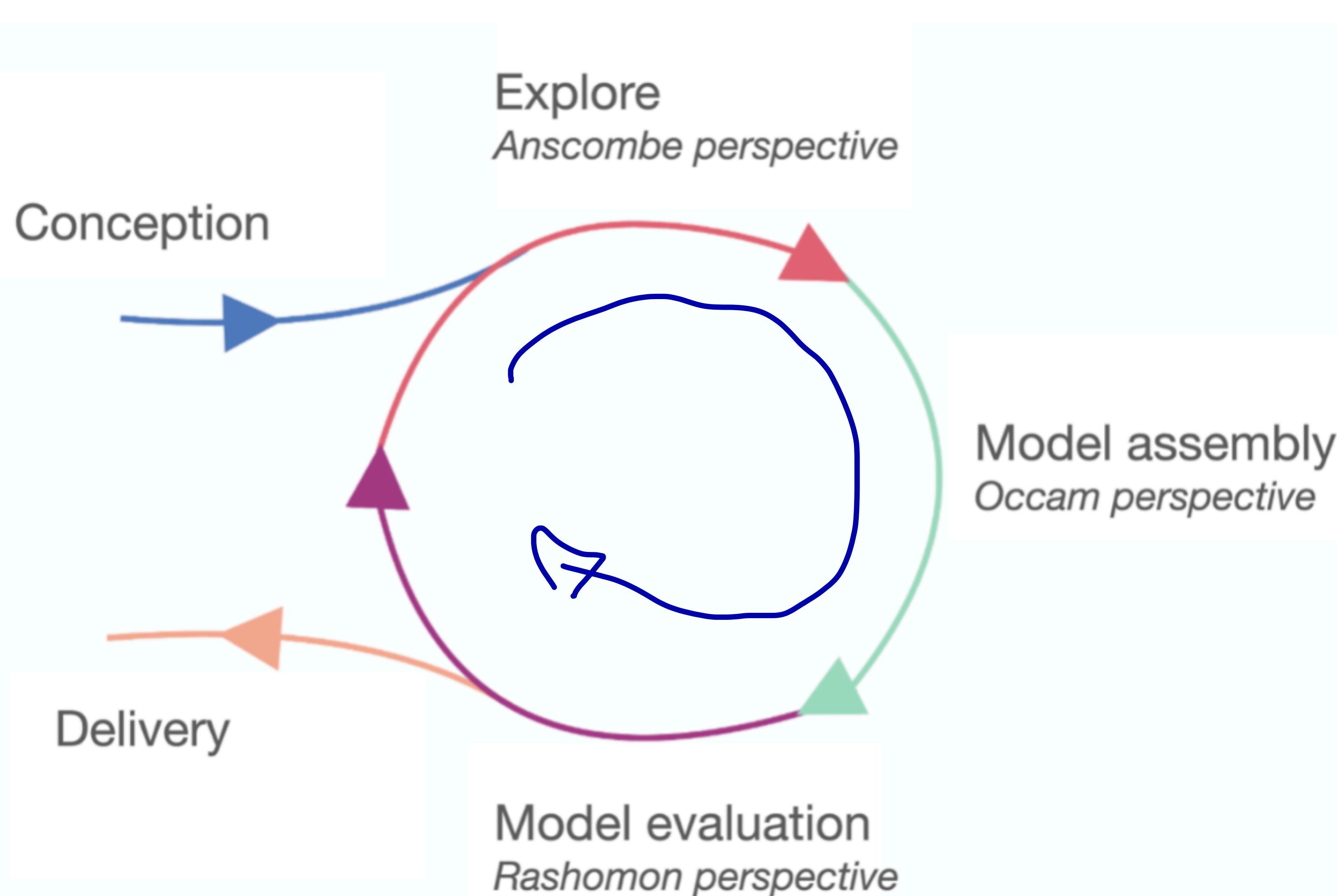


# Introduction to Responsible Machine Learning

with mlr3 and DALEX

Przemyslaw Biecek  
17-18.04.2021



## Day 1: Predictive modeling

Each block: 1.5h lecture + hands on exercises

09:00 - 10:30 Introduction to predictive modeling + EDA

10:30 - 10:45 Break

10:45 - 12:15 Hello model - first predictive model + How to measure performance

12:15 - 12:30 Break

12:30 - 14:00 Basics of random forest and boosting models

14:00 - 14:15 Break

14:15 - 15:45 Hyperparameter optimization + Wrap-up

90  
30  
30  
30

## Day 2: Model exploration

Each block: 1.5h lecture + hands on exercises

10:20 - 10:35

11:55 - 12:10

13:51 - 14:05

09:00 - 10:30 Model level analysis - variable importance

10:30 - 10:45 Break

10:45 - 12:15 Model level analysis - variable profile

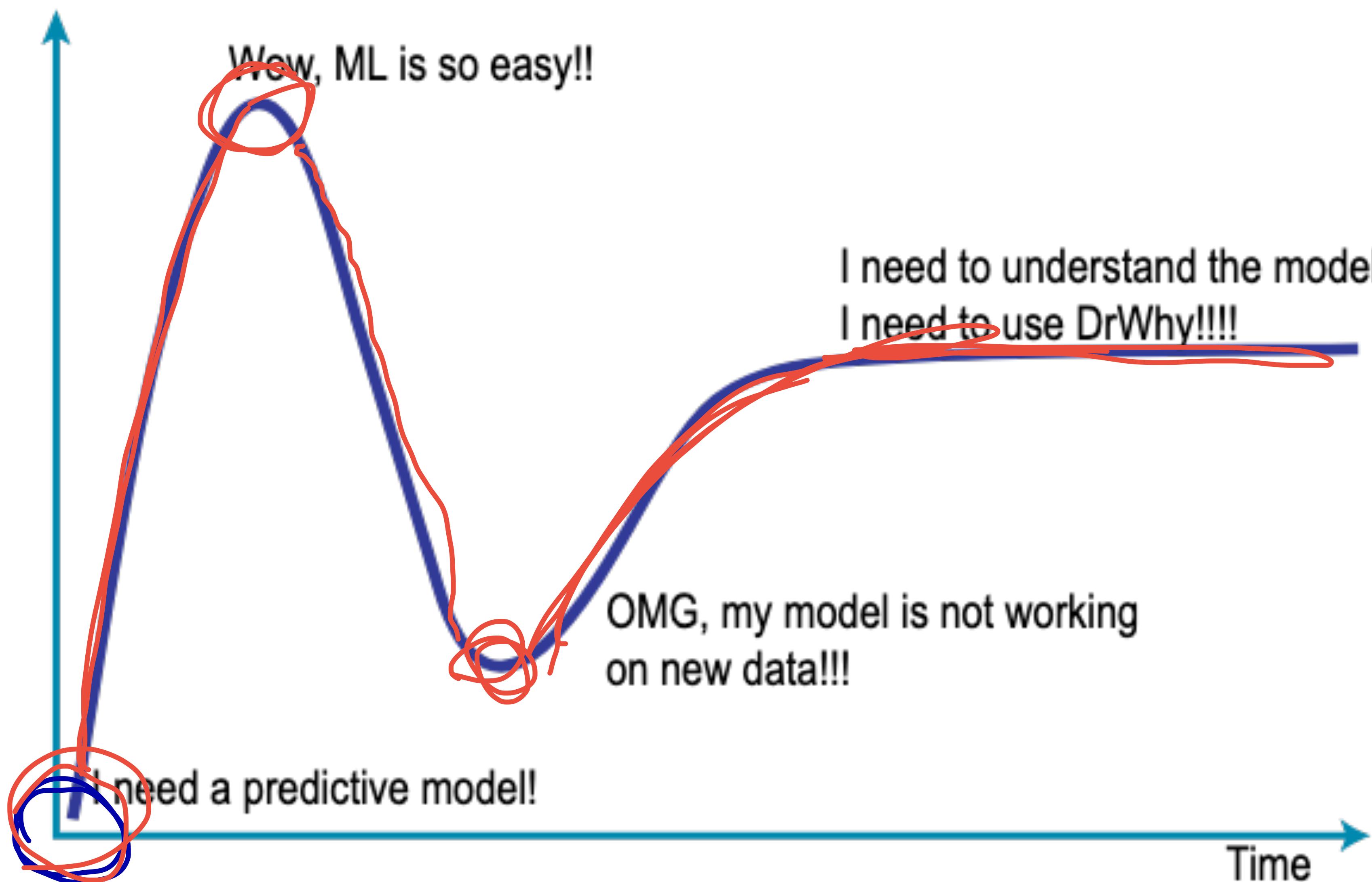
12:15 - 12:30 Break

12:30 - 14:00 Instance level analysis - variable attributions

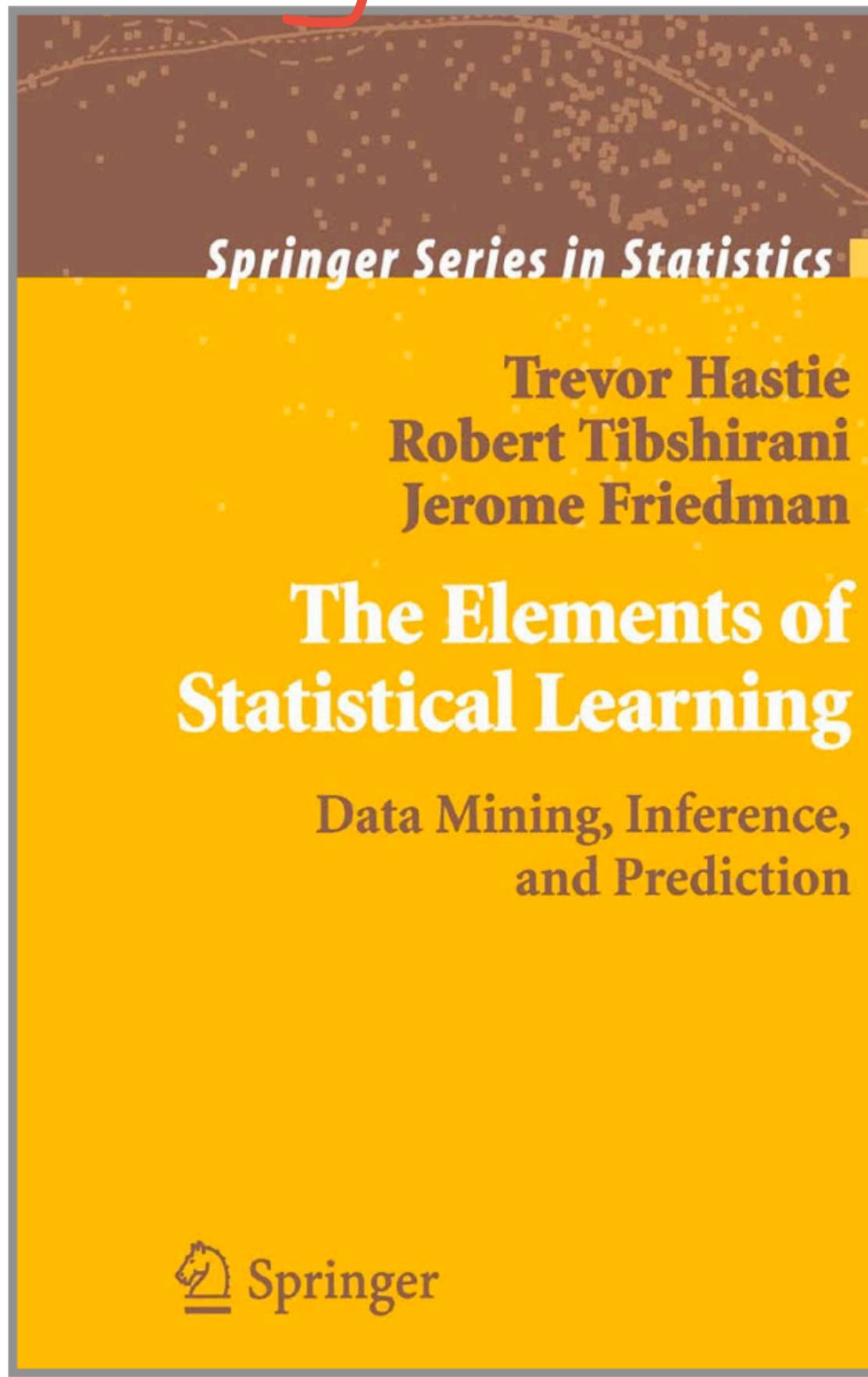
14:00 - 14:15 Break

14:15 - 15:45 Instance level analysis - variable profile + Wrap up

# Hype Cycle for Predictive Models

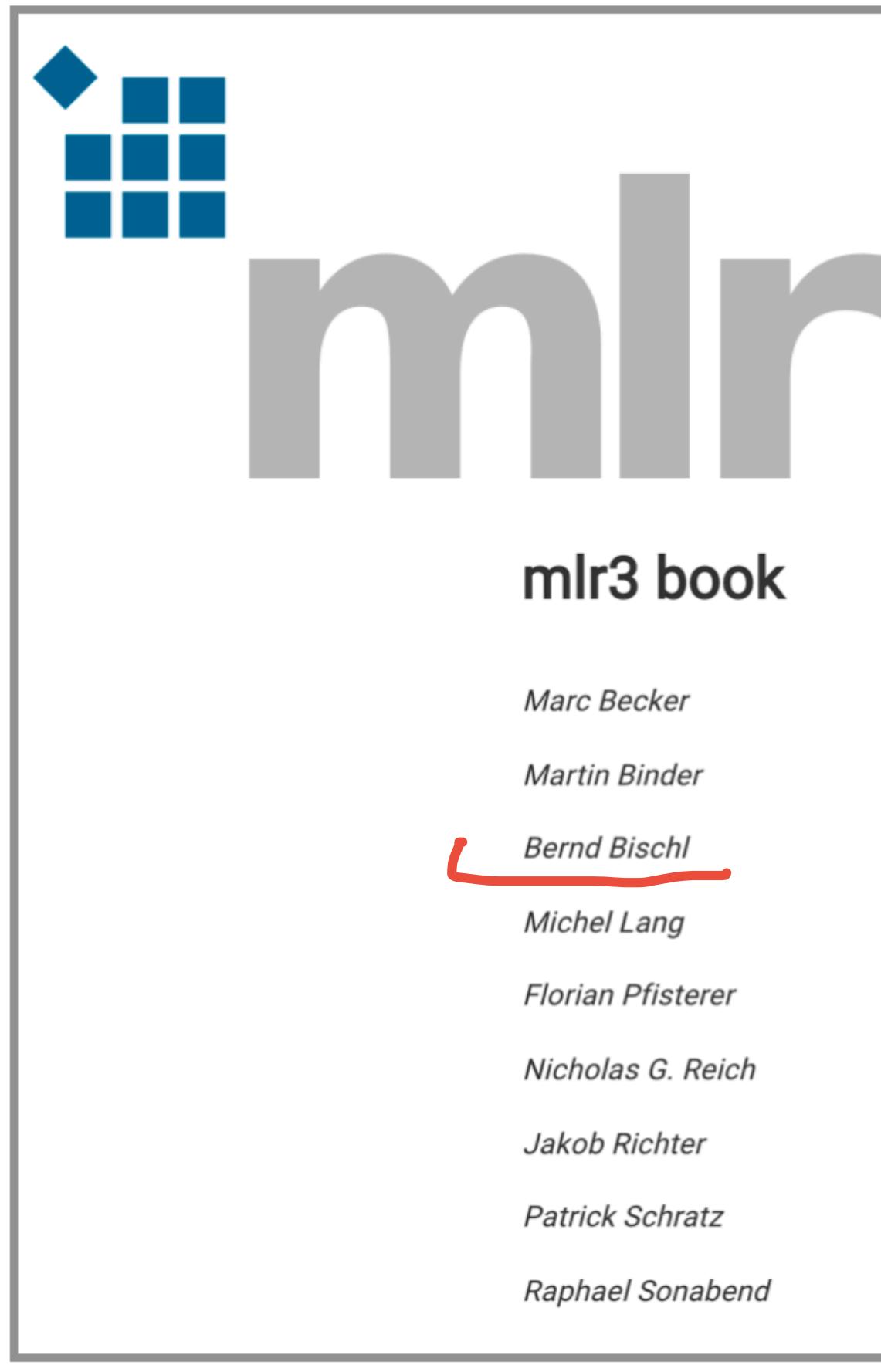


Theory



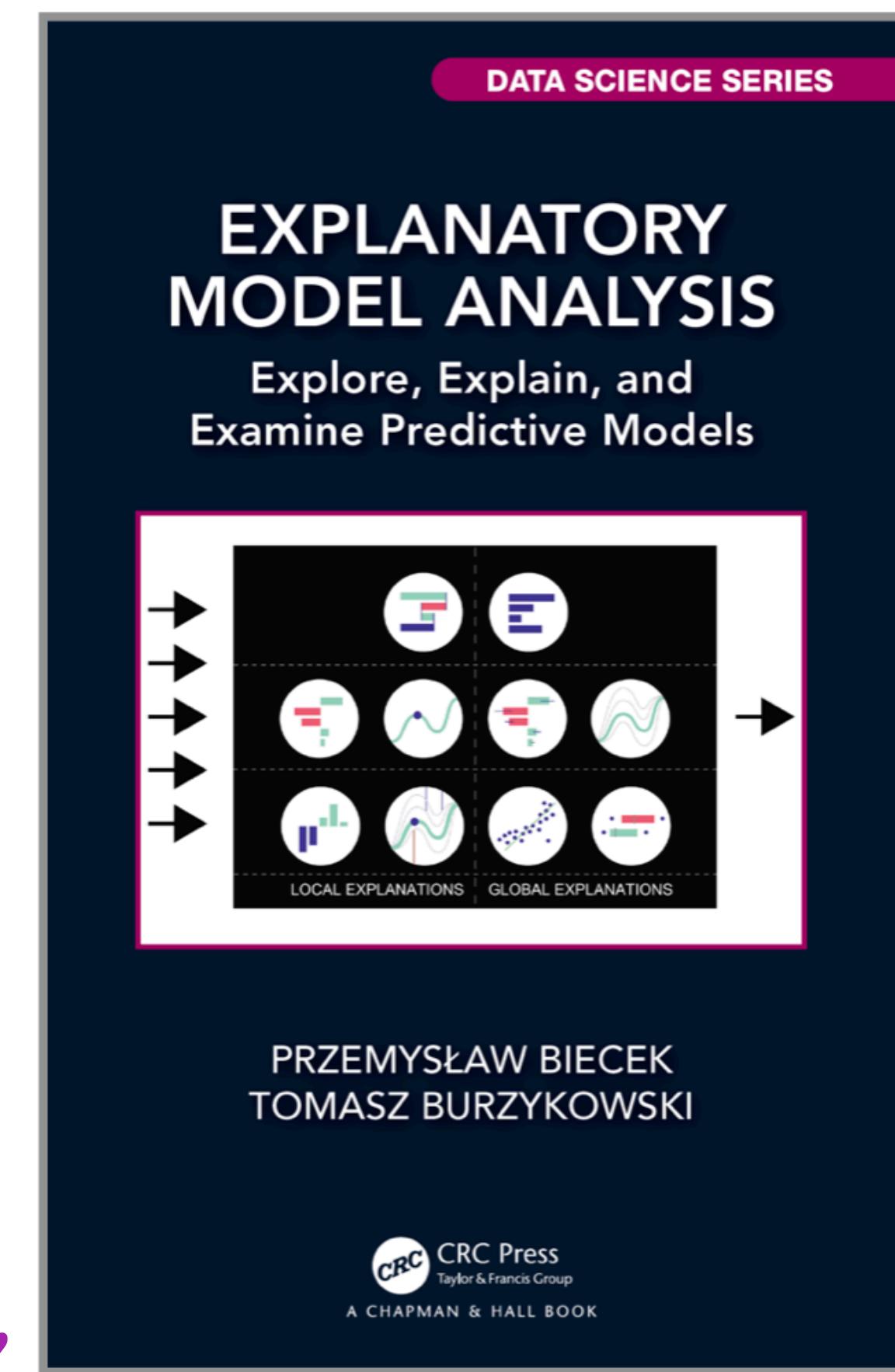
The Elements of  
Statistical Learning  
<https://www.statlearning.com/>

Tools



The mlr3 book  
<https://mlr3book.mlr-org.com/>

Day 2



Explanatory Model Analysis  
<http://ema.drwhy.ai/>

github.com/ModelOriented/DALEX/blob/master/README.md

ModelOriented / DALEX

Code Issues 12 Pull requests Actions Projects 1 Wiki Security Insights Settings

master DALEX / README.md Go to file ...

hbaniecki unify site & readme, closes #349 ✓ Latest commit ec055bf on 3 Dec 2020 History

5 contributors

107 lines (71 sloc) 9 KB Raw Blame

# moDel Agnostic Language for Exploration and eXplanation

R-CMD-check passing coverage 87% CRAN 2.2.0 downloads 103K DrWhy BackBone

Python-check passing python 3.6 | 3.7 | 3.8 pypi package 1.0.1 downloads 32k



## Overview

Unverified black box model is the path to the failure. Opaqueness leads to distrust. Distrust leads to ignoration. Ignoration leads to rejection.

The `DALEX` package xrays any model and helps to explore and explain its behaviour, helps to understand how complex models are working. The main function `explain()` creates a wrapper around a predictive model. Wrapped models may then be explored and compared with a collection of local and global explainers. Recent developments from the area of Interpretable Machine Learning/eXplainable Artificial Intelligence.

The philosophy behind `DALEX` explanations is described in the [Explanatory Model Analysis e-book](#). The `DALEX` package is a part of [DrWhy.AI](#) universe.

If you work with `scikit-learn`, `keras`, `H2O`, `tidymodels`, `xgboost`, `mlr` or `mlr3`, you may be interested in the `DALEXtra` package. It is an extension pack for `DALEX` with easy to use connectors to models created in these libraries.

<https://github.com/ModelOriented/DALEX>



## Przemysław Biecek

(53.2% of academia, 21.3% of corpo, 10.6% of startups, 14.9% of NGOs)

### Academy

6 years - Warsaw University of Technology  
13 years - University of Warsaw  
1 year - University of Wroclaw  
1 year - Polish Academy of Sciences  
4 years - Wroclaw University of Technology

### Corpo

3 years - Samsung - Principal Data Scientist  
1 year - OECD France - Research Fellow  
2 years - iQor - Head of Data Vis  
2 years - IBM Senior Data Scientist  
2 years - Netezza Senior Data Scientist

### Startups

1 year - Data Donuts  
2 year - Data Hero  
2 years - Positive Advisory

### NGO

7 years - SmarterPoland / BetaBit

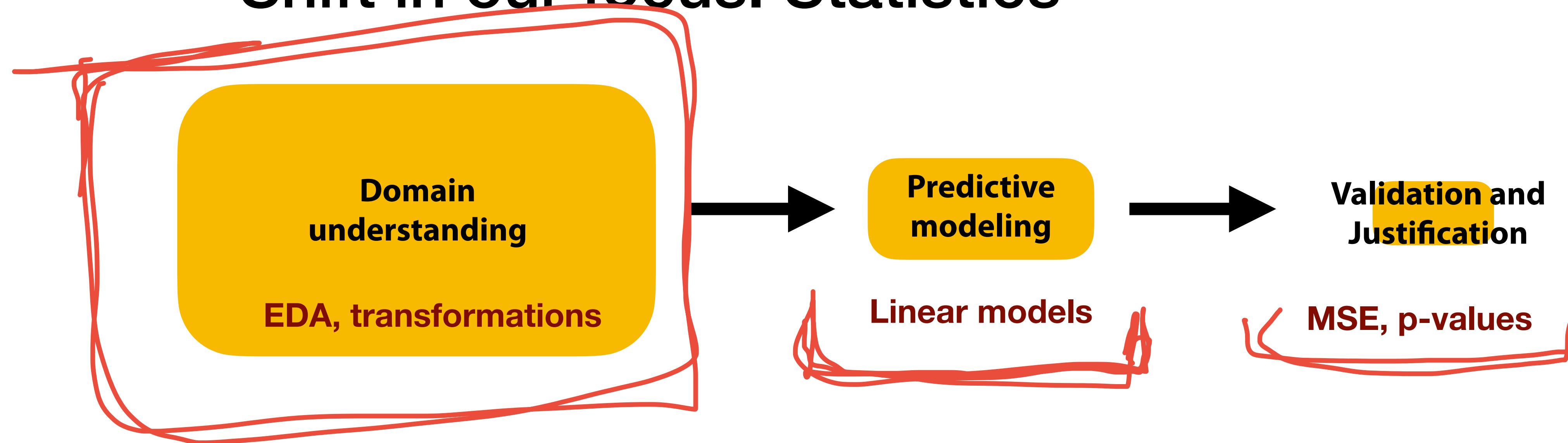
# Day 1

## Predictive modeling

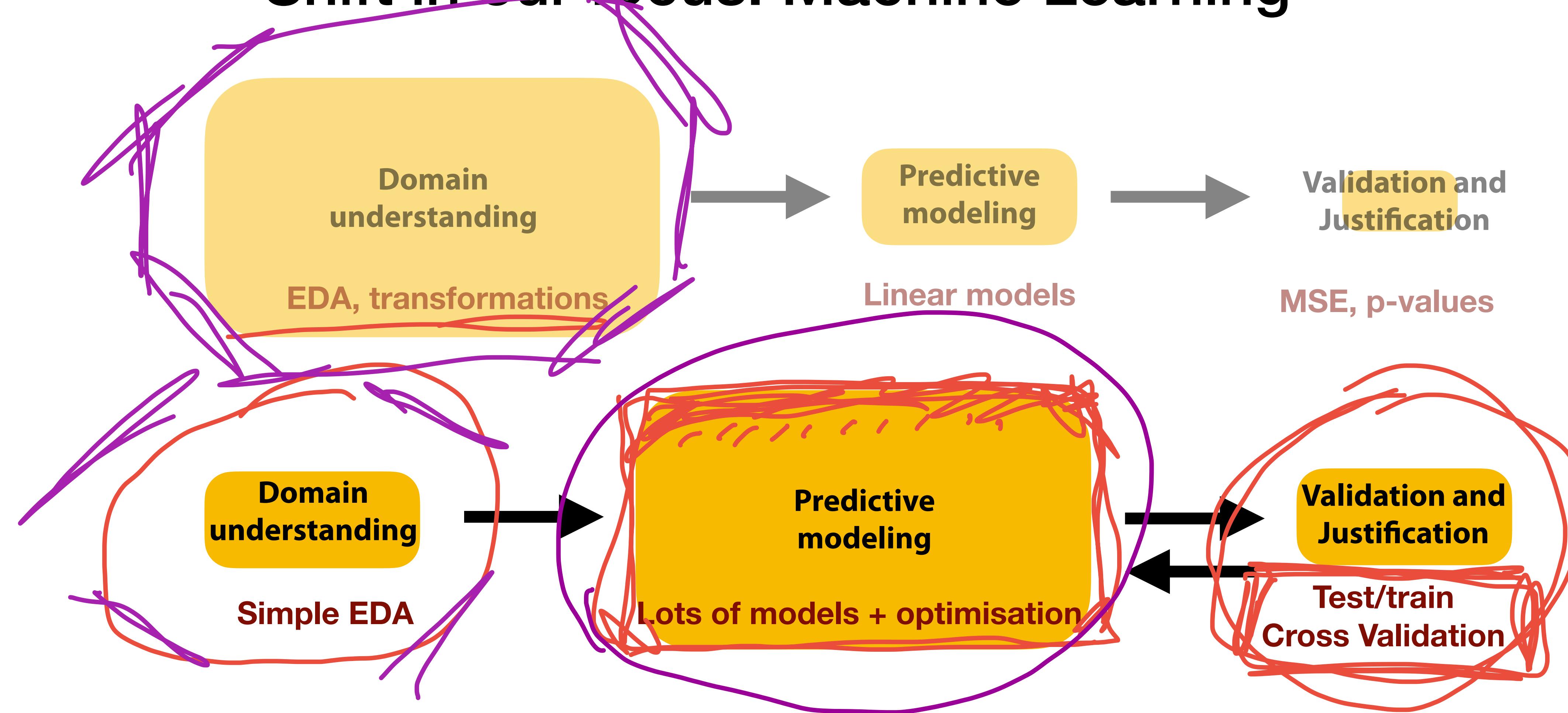
### Part 1

#### Introduction to predictive modeling + EDA

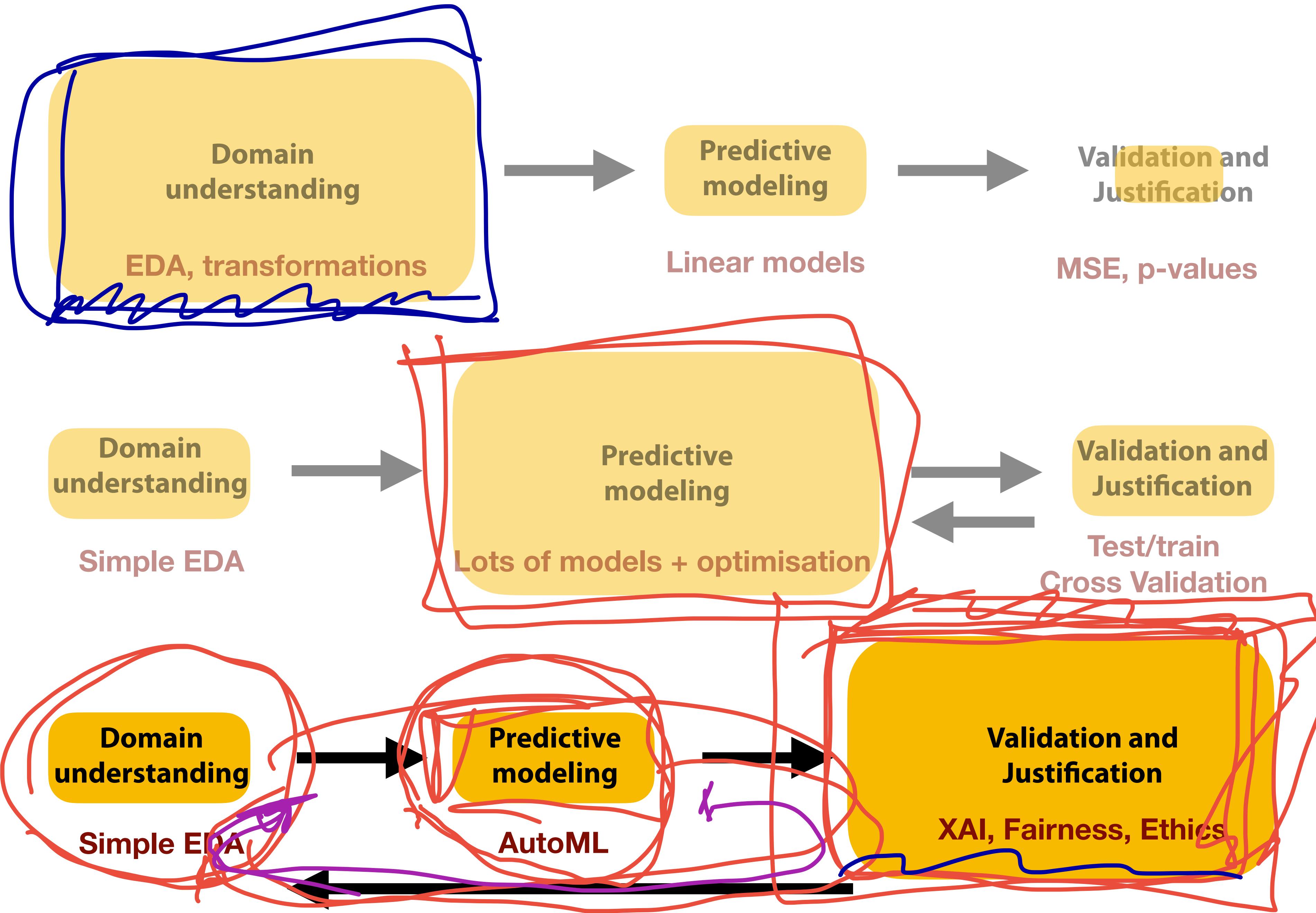
# Shift in our focus: Statistics



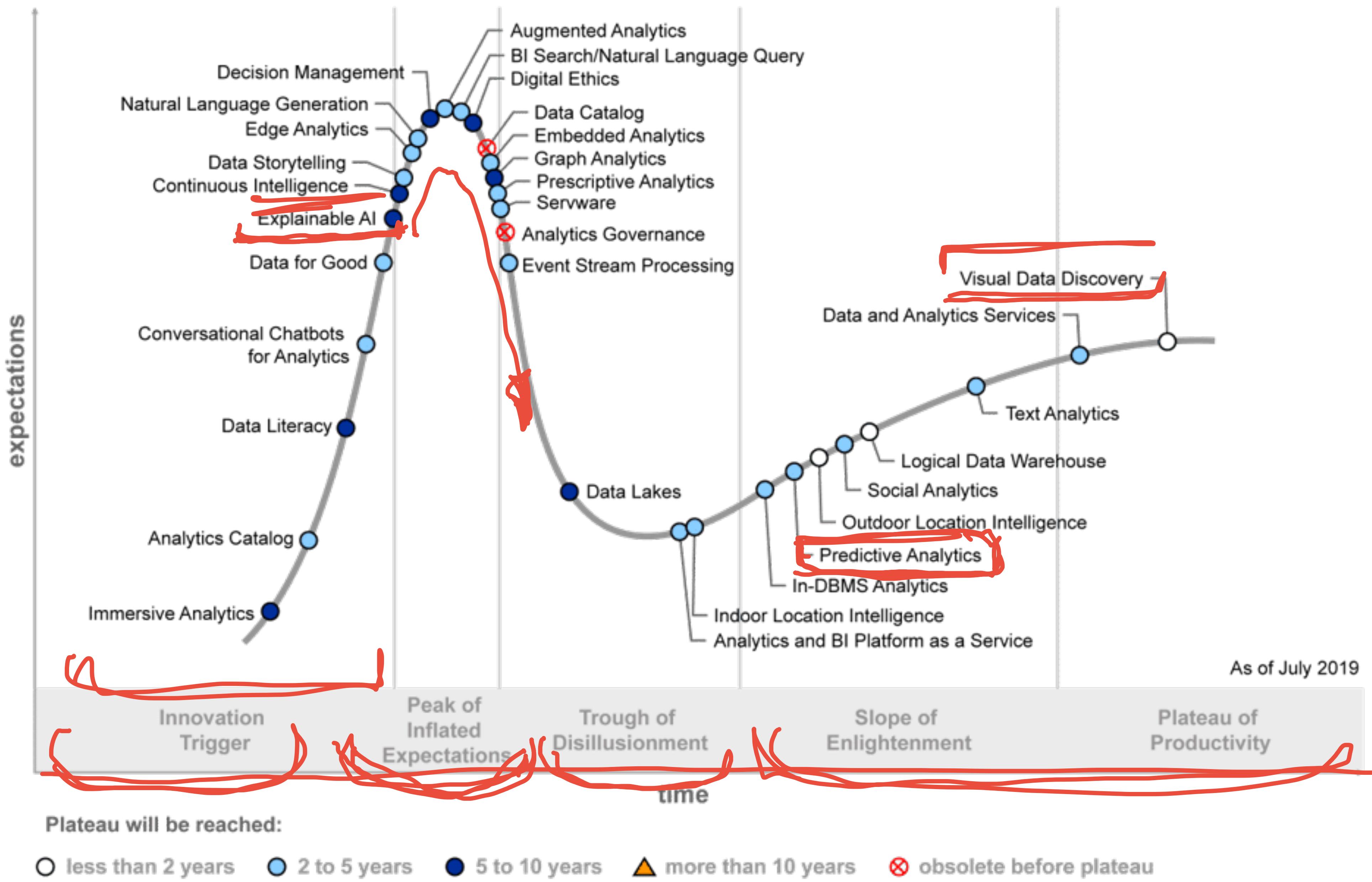
# Shift in our focus: Machine Learning

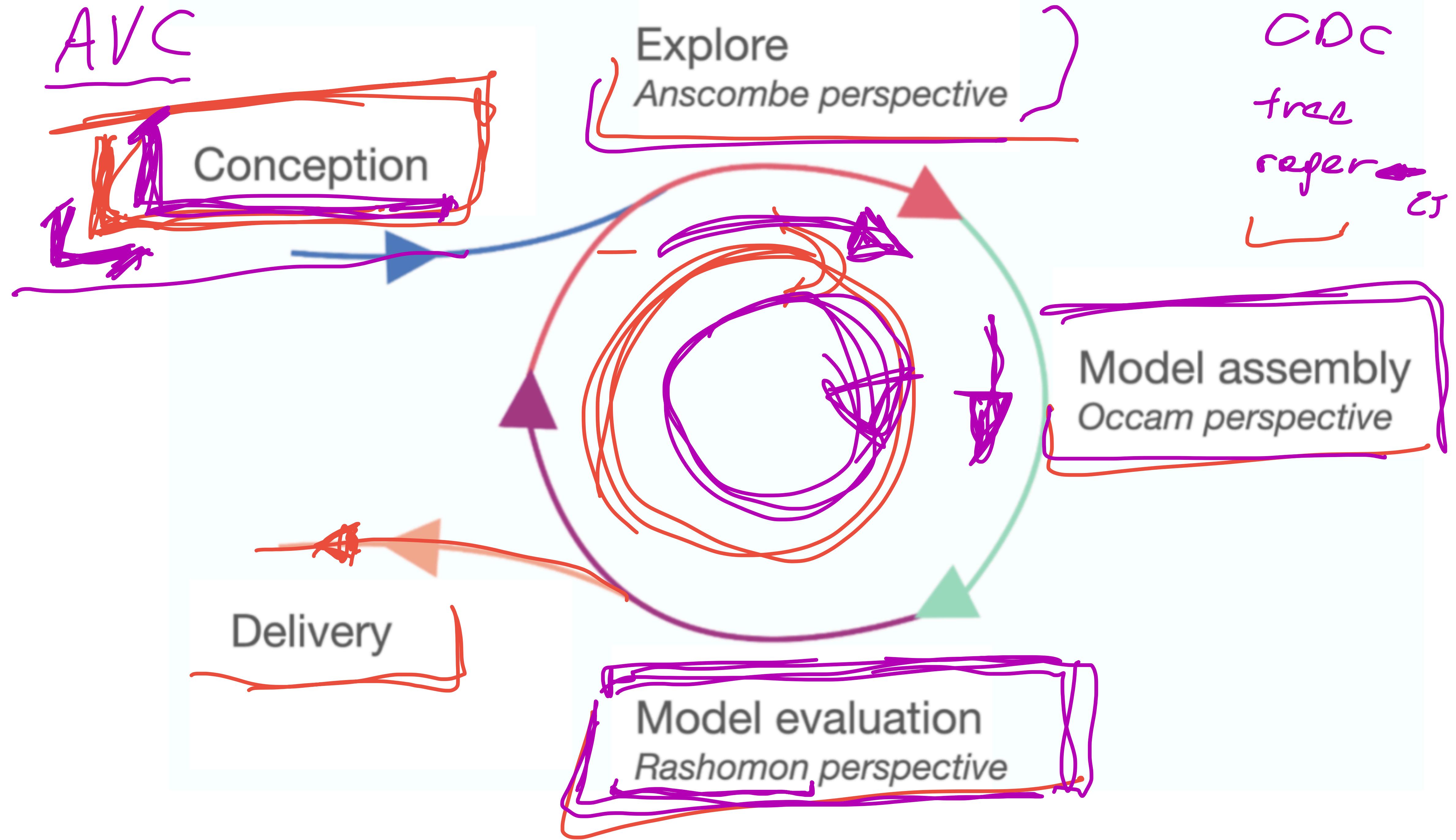


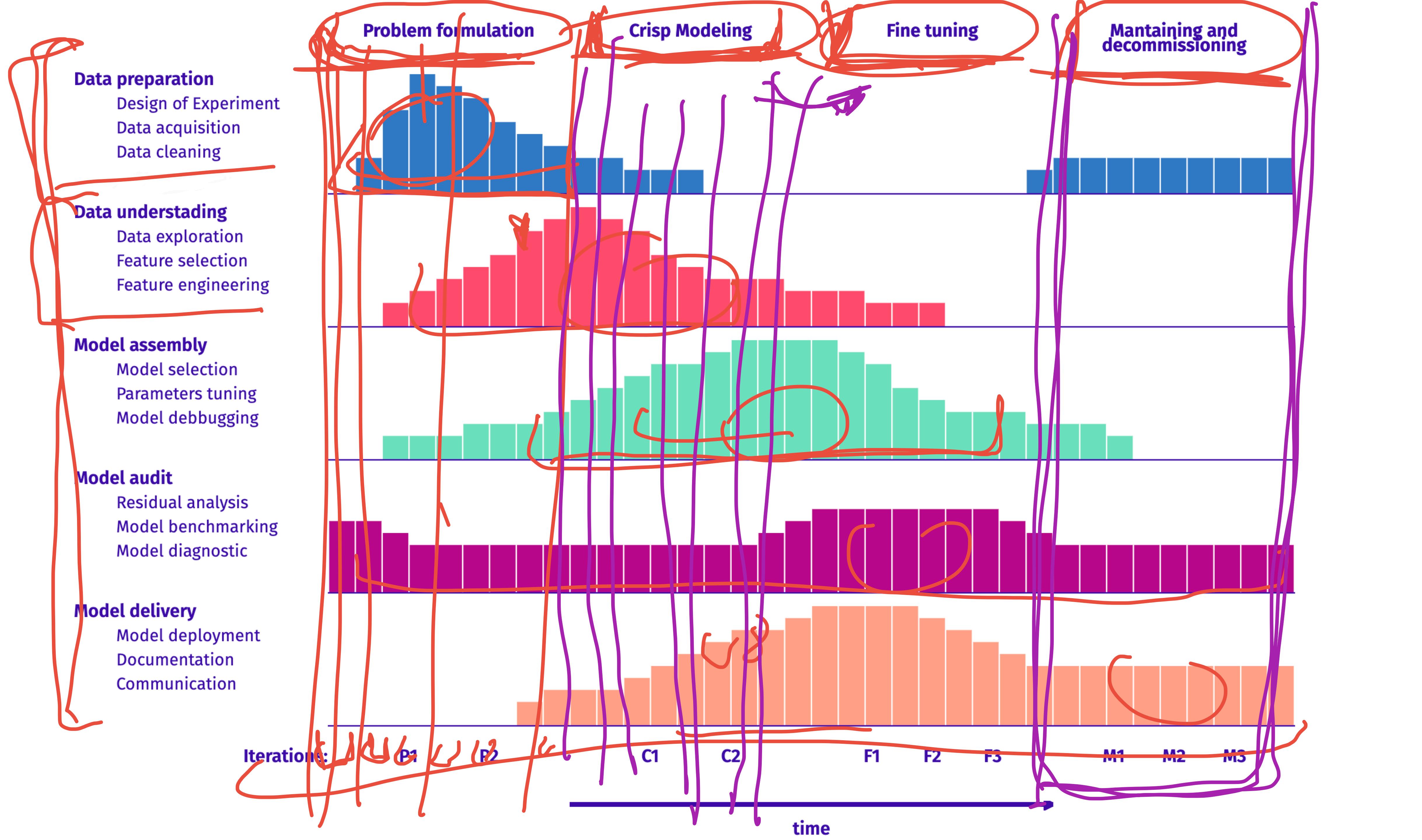
# Shift in our focus: Human Oriented ML?



# Hype Cycle for Analytics and Business Intelligence, 2019







# Day 1

## Predictive modeling

### Part 2

Hello model - first predictive model  
How to measure performance

$$MSE(f) = \frac{1}{n} \sum_i^n (f(x_i) - y_i)^2$$

$$RMSE(f) = \sqrt{MSE(f, X, y)}$$

$$ACC(f) = \frac{(TP + TN)}{n}$$

$$Prec(f) = TP / (TP + FP)$$

$$Recall(f) = TP / (TP + FN)$$

$$F_1(f) = 2 \frac{Prec(f) * Recall(f)}{Prec(f) + Recall(f)}$$

True condition		
Total population	Condition positive	Condition negative
Predicted condition		
positive	<b>True positive</b>	<b>False positive,</b> Type I error
negative	<b>False negative,</b> Type II error	<b>True negative</b>

		True condition		
Predicted Condition	Total population	Condition positive	Condition negative	Prevalence = $\frac{\sum \text{Condition positive}}{\sum \text{Total population}}$
	Predicted condition positive	True positive	False positive, Type I error	Positive predictive value (PPV), Precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$
	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$
		Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR}^+}{\text{LR}^-}$
		$F_1$ score = 2 · $\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$		
<a href="https://en.wikipedia.org/wiki/Sensitivity_and_specificity">https://en.wikipedia.org/wiki/Sensitivity_and_specificity</a>				

# Pregnancy: Sensitivity and Specificity

Introduction: None written.

[Edit Diagnosis] [Merge dx] [Add prevalence]

Tags: None. [Tag this Diagnosis](#).

The sensitivity and specificity of findings for Pregnancy are listed below. See the left navigation bar to change the display.

## Sensitive and Specific Findings

Finding	Sensitivity	Specificity	Comments, Study
Serum beta-hCG	97%	100%	as compared to urine beta hCG, for all comers to an ED for whom pregnancy tests were ordered. In this study, urine beta hCG had sens = 100%, spec = 99.2% <b>Study:</b> Am J Emerg Med. 1993 Jul;11(4):434-6. PMID: 8216535

## Specific Findings

Finding	Sensitivity	Specificity	Comments, Study
Chadwick Sign	51%	98%	Only 1 study from Dr Chadwick. Sign is "characteristic" deep violet-blue color of anterior vaginal wall. Rarely seen before 7 weeks. <b>Study:</b> JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281
Palpable Fundus	9%	97%	<b>Study:</b> JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281
Vaginal Exam Signs	18%	94%	not explicitly defined <b>Study:</b> JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281
Uterine Artery Pulsation	76%	93%	only 1 study <b>Study:</b> JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281
Morning Sickness	39%	85%	<b>Study:</b> JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281

## Sensitive Findings

Finding	Sensitivity	Specificity	Comments, Study
Urine beta-hCG Duplicate	Sensitivity = 97%		7 days after missed period. Most urine b-hCG tests are sensitive to 20 mIU/ml. If sensitivity is at 0.5 mIU/ml then should be sensitive by 7 days after fertilization. <b>Study:</b> JAMA 2001 Oct 10;286(14):1759-61. PMID 11594902
Urine beta-hCG Duplicate	Sensitivity = 90%		on 1st day of missed period. Most urine b-hCG tests are sensitive to 20 mIU/ml. If sensitivity is at 0.5 mIU/ml then should be sensitive by 7 days after fertilization. <b>Study:</b> JAMA 2001 Oct 10;286(14):1759-61. PMID 11594902
Did Not Use Birth Control	88%	42%	<b>Study:</b> JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281

Morning sickness	Pregnant	Not Pregnant
Has sickness	39	150
Has not	61	850
Sensitivity: 39.00%		Specificity: 85.00%

100

1000

Positive predictive value: 20.63% 125

Negative predictive value: 93.30% 911

Accuracy: 80.82%

**miss rate or false negative rate (FNR)**

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} = 1 - TPR$$

**false-out or false positive rate (FPR)**

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR$$

**false discovery rate (FDR)**

$$\text{FDR} = \frac{FP}{FP + TP} = 1 - PPV$$

**false omission rate (FOR)**

$$FOR = \frac{FN}{FN + TN} = 1 - NPV$$

**accuracy (ACC)**

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

**F1 score**

is the harmonic mean of precision and sensitivity

$$F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

**Matthews correlation coefficient (MCC)**

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

**Informedness or Bookmaker Informedness (BM)**

$$BM = TPR + TNR - 1$$

**Markedness (MK)**

$$MK = PPV + NPV - 1$$

**condition positive (P)**

the number of real positive cases in the data

**condition negative (N)**

the number of real negative cases in the data

**true positive (TP)**

eqv. with hit

**true negative (TN)**

eqv. with correct rejection

**false positive (FP)**

eqv. with **false alarm**, **Type I error**

**false negative (FN)**

eqv. with miss, **Type II error**

**sensitivity, recall, hit rate, or true positive rate (TPR)**

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

**specificity, selectivity or true negative rate (TNR)**

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

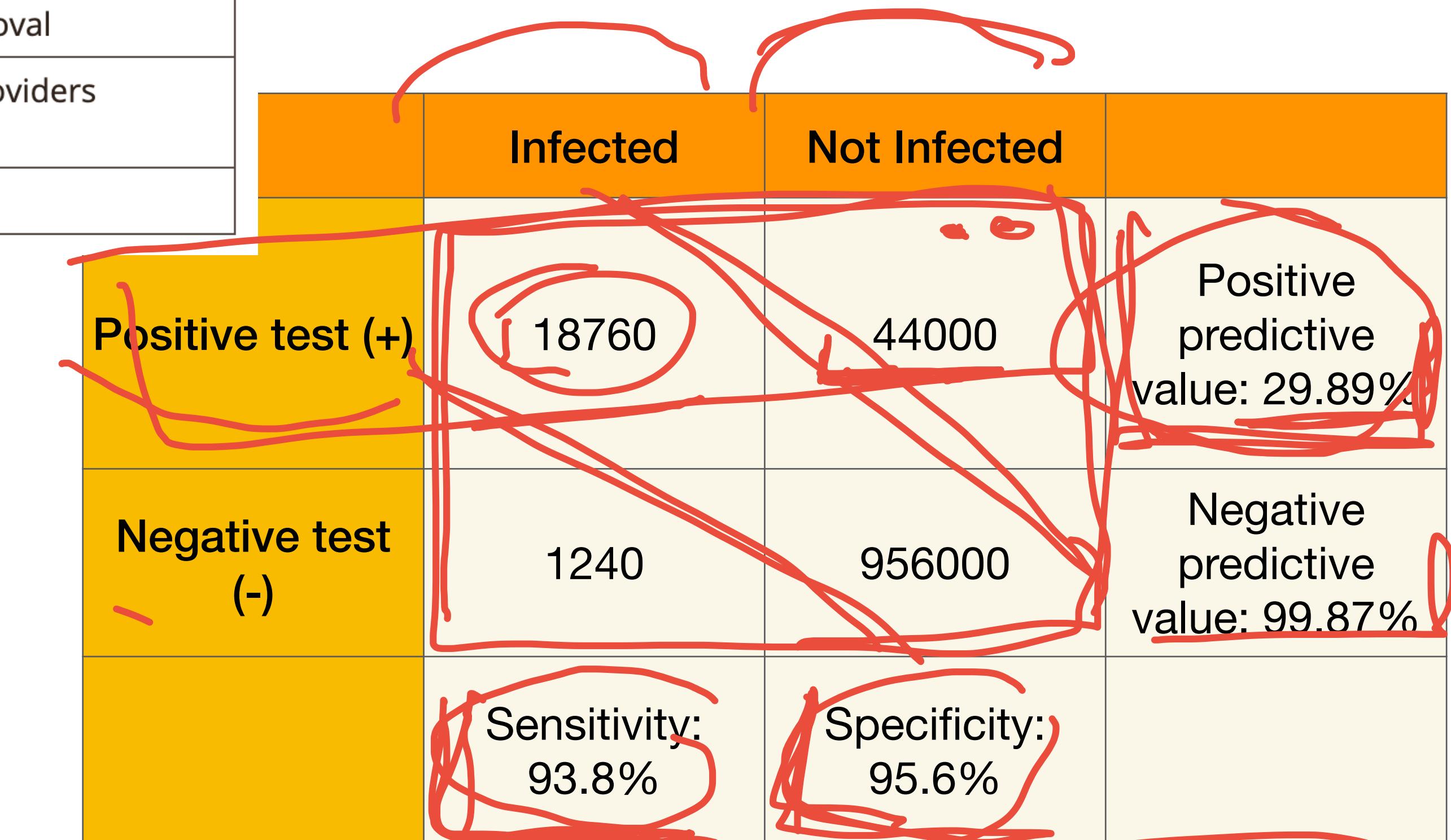
**precision or positive predictive value (PPV)**

$$PPV = \frac{TP}{TP + FP} = 1 - FDR$$

**negative predictive value (NPV)**

$$NPV = \frac{TN}{TN + FN} = 1 - FOR$$

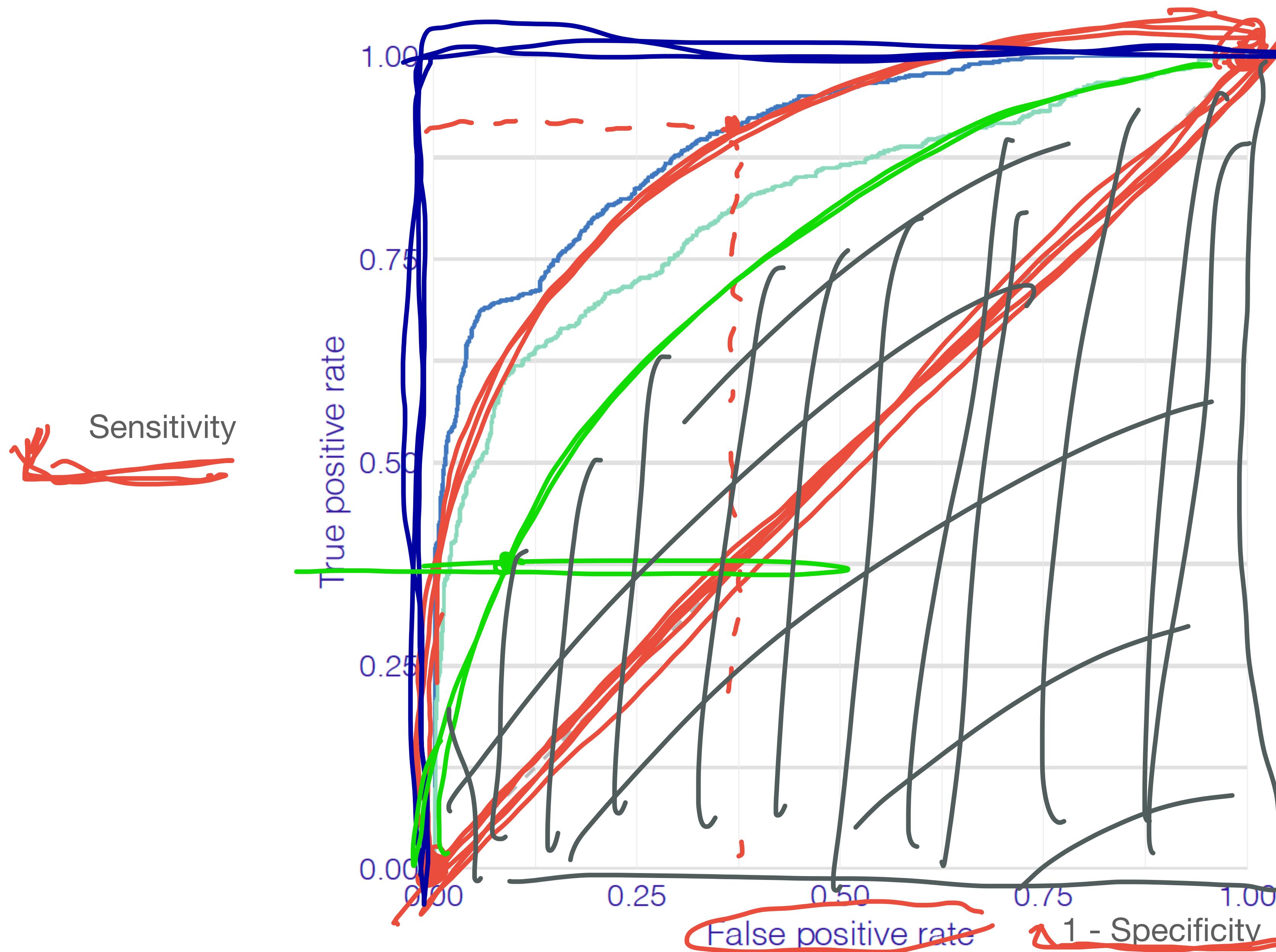
<b>Country of development</b>	US/China
<b>Type of Serological Test</b>	RDT
<b>Authors/Company</b>	<a href="#">Cellex Inc.</a>
<b>Description</b>	RDT, lateral flow assay, which detects IgM and IgG to the nucleocapsid protein of SARS-CoV-2. The sensitivity is 93.8% and specificity is 95.6%, when tested at 2 Chinese hospitals in a total of 128 COVID19 positive patients, and 250 COVID19 negative patients (as detected by RT-qPCR).
<b>Sensitivity</b>	93.8%
<b>Specificity</b>	95.6%
<b>Phase of development</b>	Approved by FDA for EUA on diagnostics, has CE approval
<b>Proposed release</b>	available for purchase by research labs/healthcare providers (product number 5513)
<b>Date</b>	April 1, 2020



# Receiver Operator Characteristic

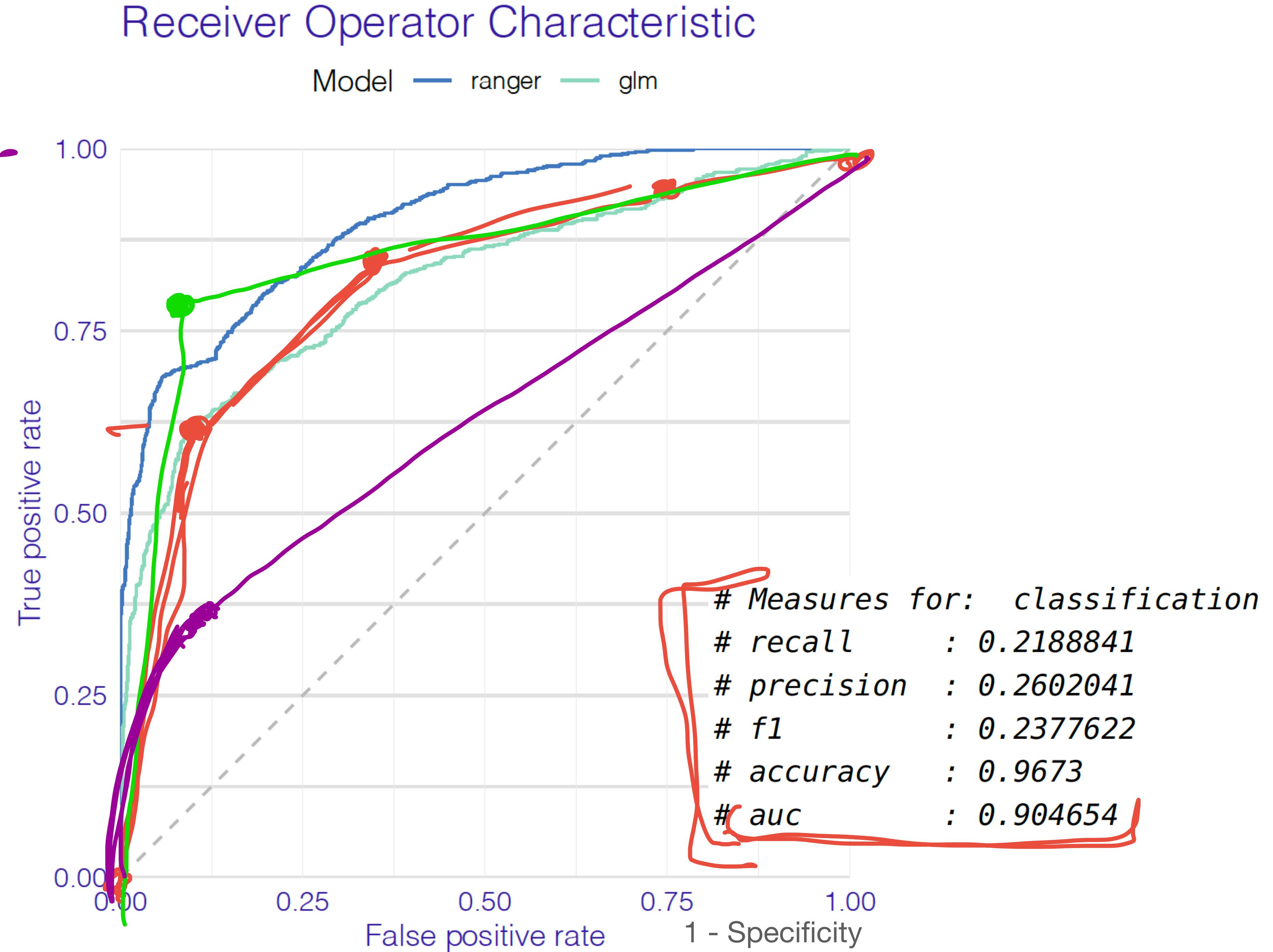
ROC

Model — ranger — glm



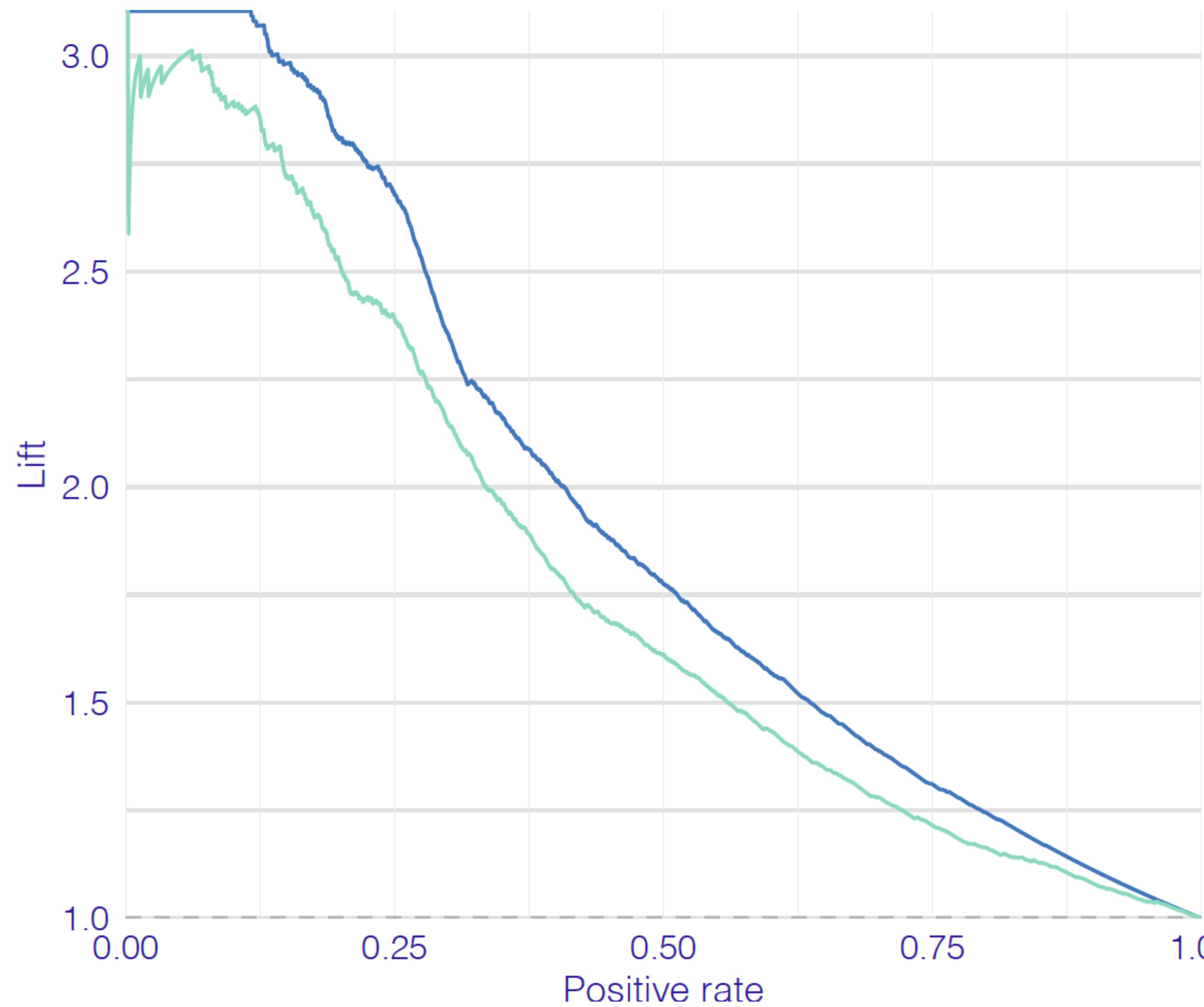
$f(x) \rightarrow P$

Sensitivity



# Lift chart

Model — ranger — glm



ranger(y ~ ., data = df)

%>%

explain()

%>%

model\_parts()

%>%

plot()

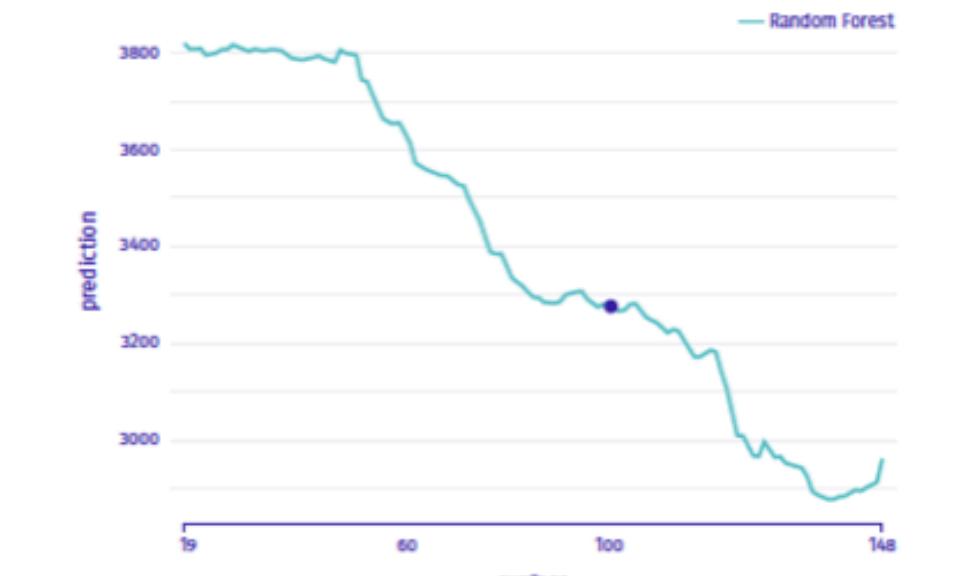
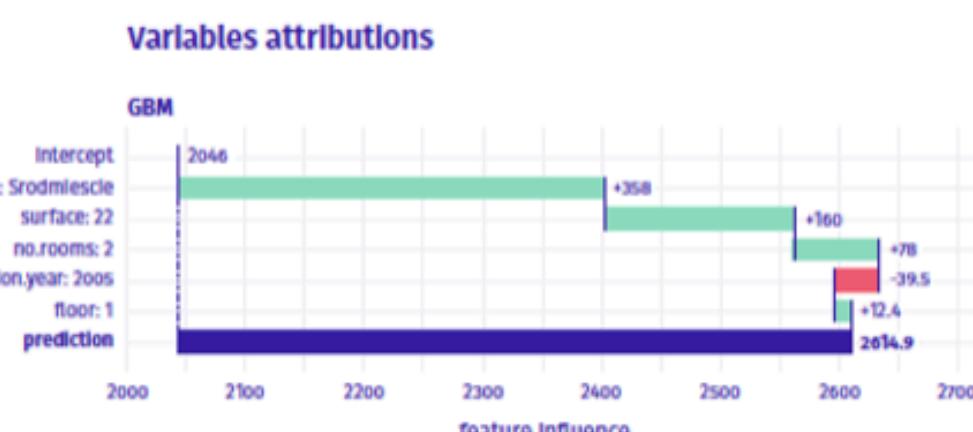
model

explainer

explanation



parsnip



G++

Java

*explainer*

model

data: data.frame

y: numeric

y\_hat: numeric

predict\_function: function (model, data)

residuals: numeric

residual\_function: function(model, data, y)

weights: numeric

model\_info: list(package, ver, type)

class: character

label: character

Day 1

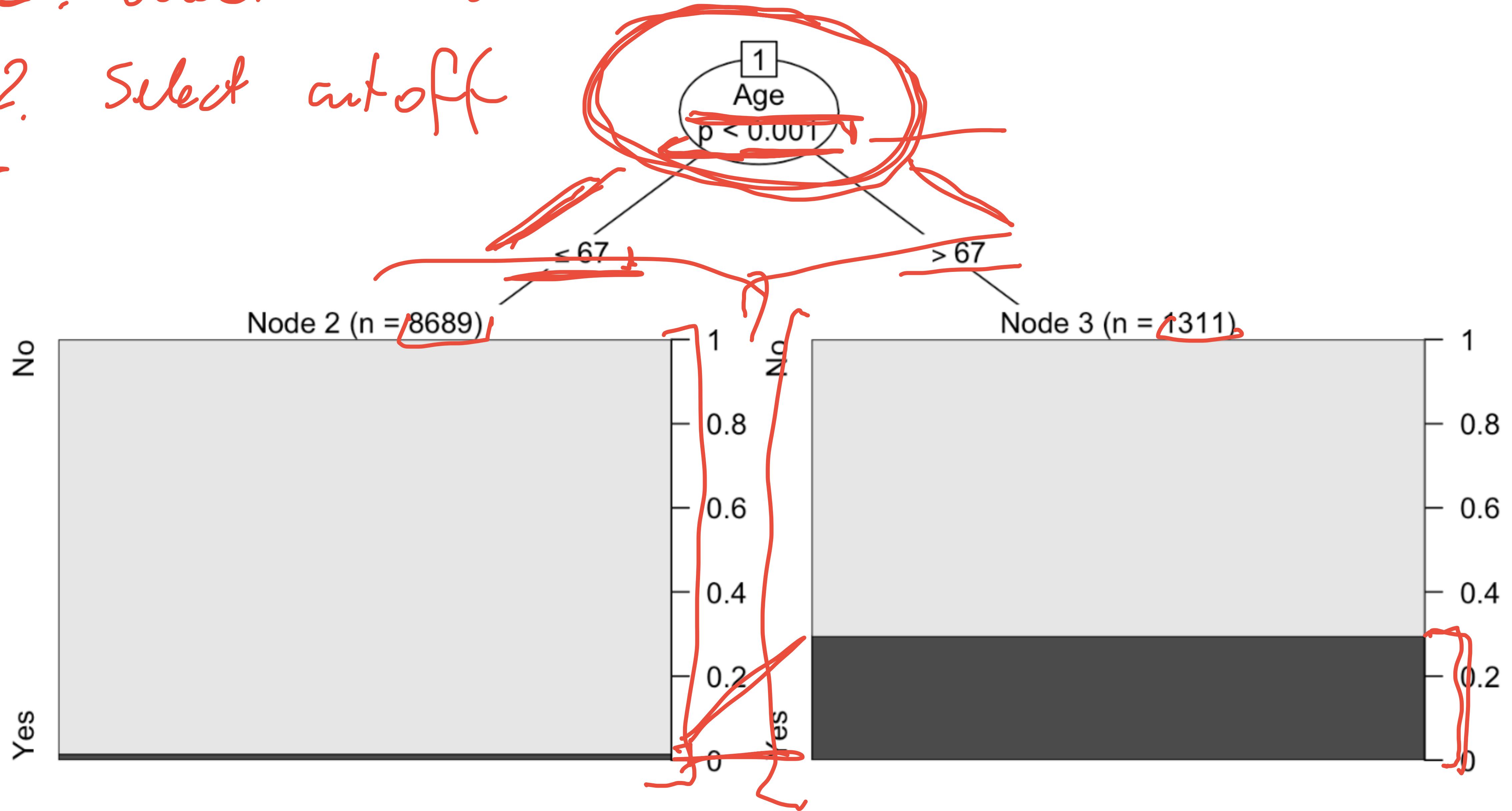
Predictive modeling

Part 3

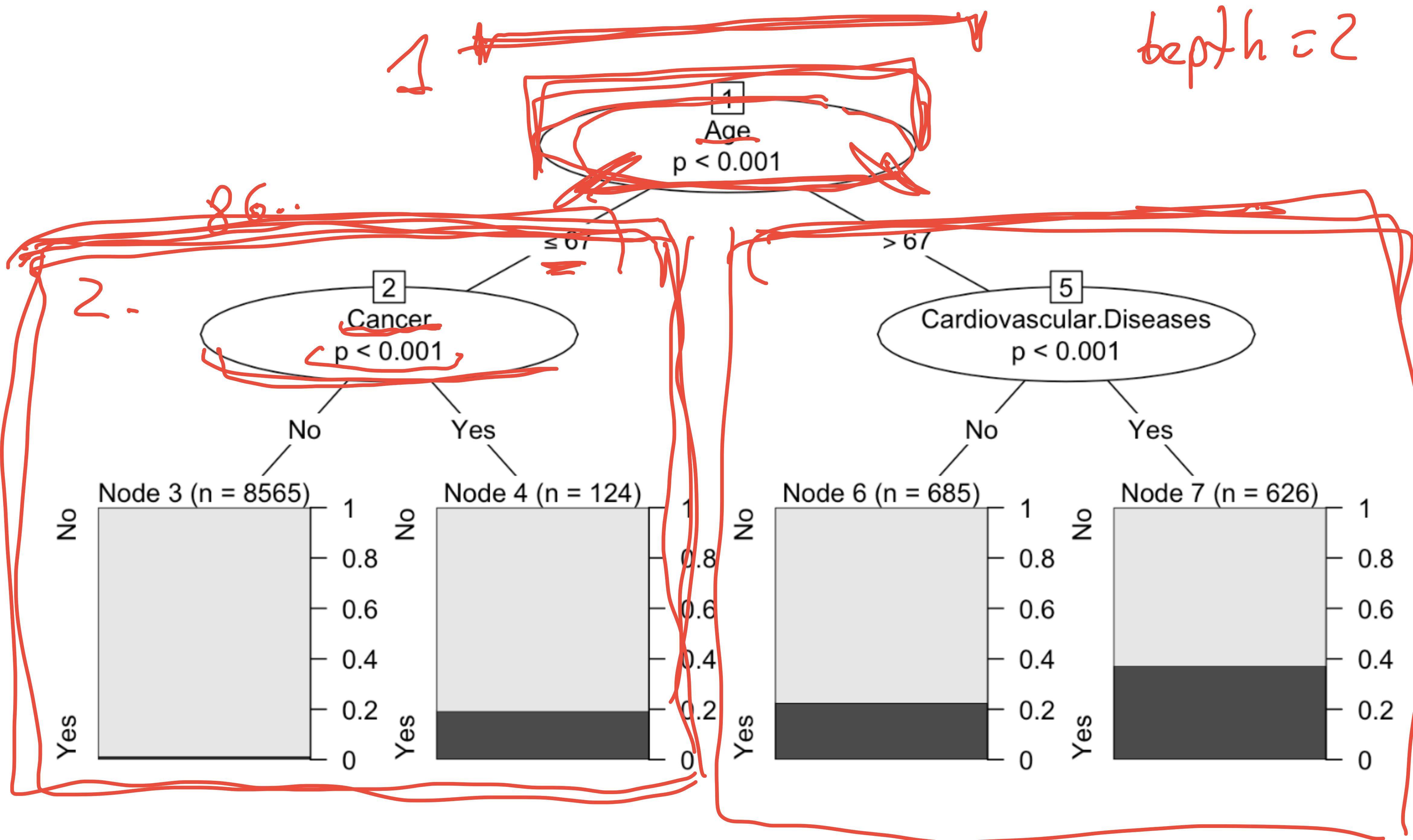
Basics of random forest and boosting models

1. Select variable

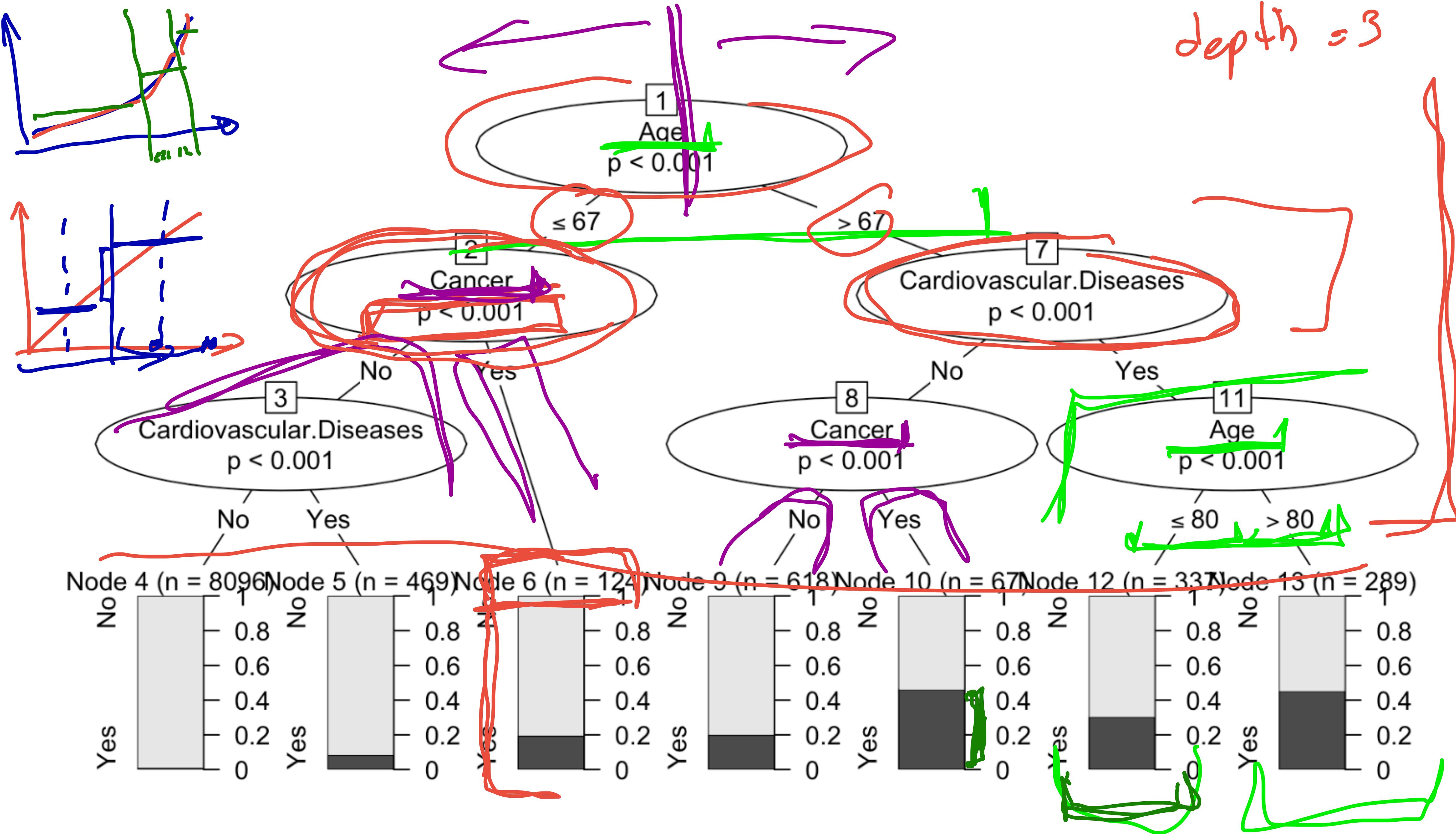
2. Select cutoff



depth = 2



dep<sup>th</sup> = 3



2001 Random Forest Breiman 2004 Boosting Friedman

Random Forests grows many classification/regression trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

PIZZOT

# im:2 Each tree

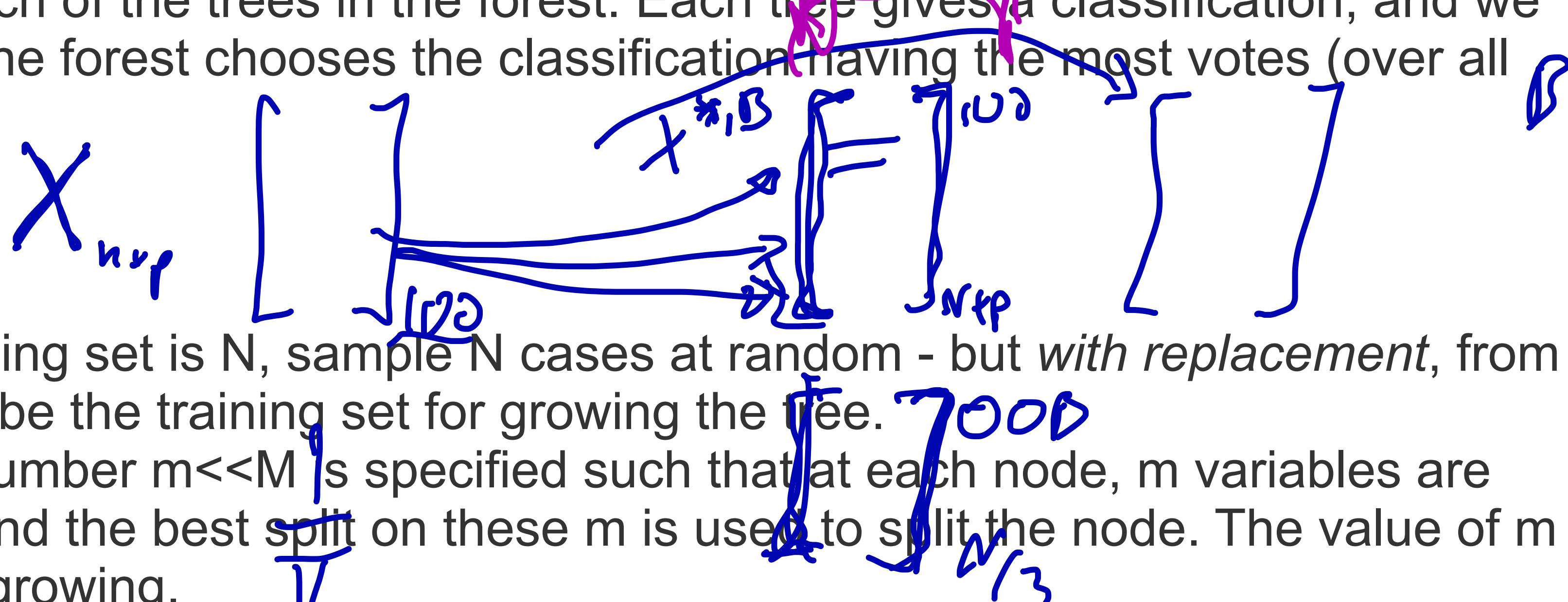
1. If the number of cases in the training set is  $N$ , sample  $N$  cases at random - but *with replacement*, from the original data. This sample will be the training set for growing the tree. **700B**

2. If there are  $M$  input variables, a number  $m < M$  is specified such that at each node,  $m$  variables are selected at random out of the  $M$  and the best ~~split~~<sup>17</sup> on these  $m$  is used to split the node. The value of  $m$  is held constant during the forest growing.

3. Each tree is grown to the largest extent possible. There is no pruning.

$$\text{MSE} = \text{Var } f + \text{bias}^2 f + \sigma^2$$

close to  
(stable)  
Sources

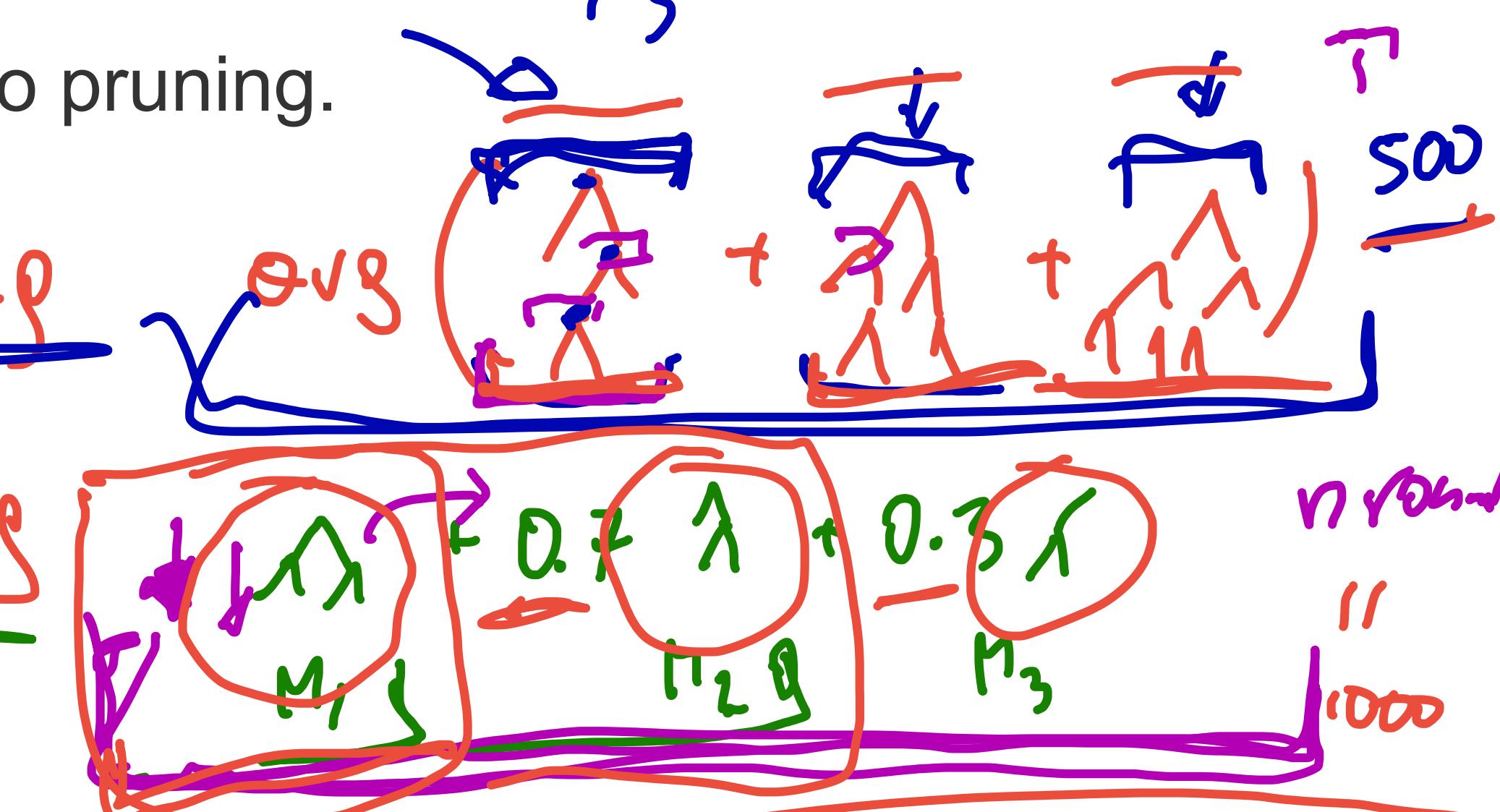


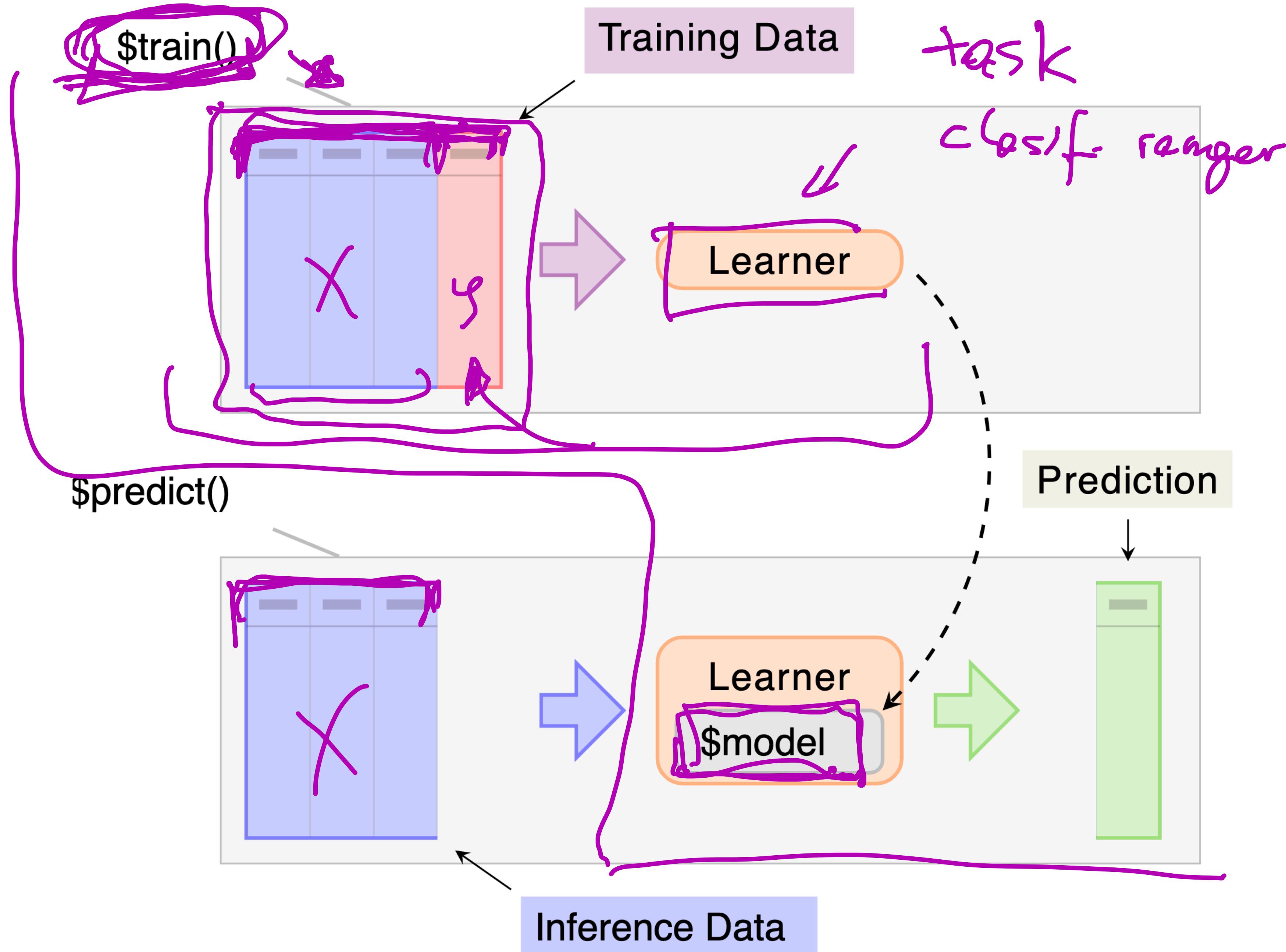
$$m = \sqrt{p}$$

$p = 10080$

# Bagging

# Boosting



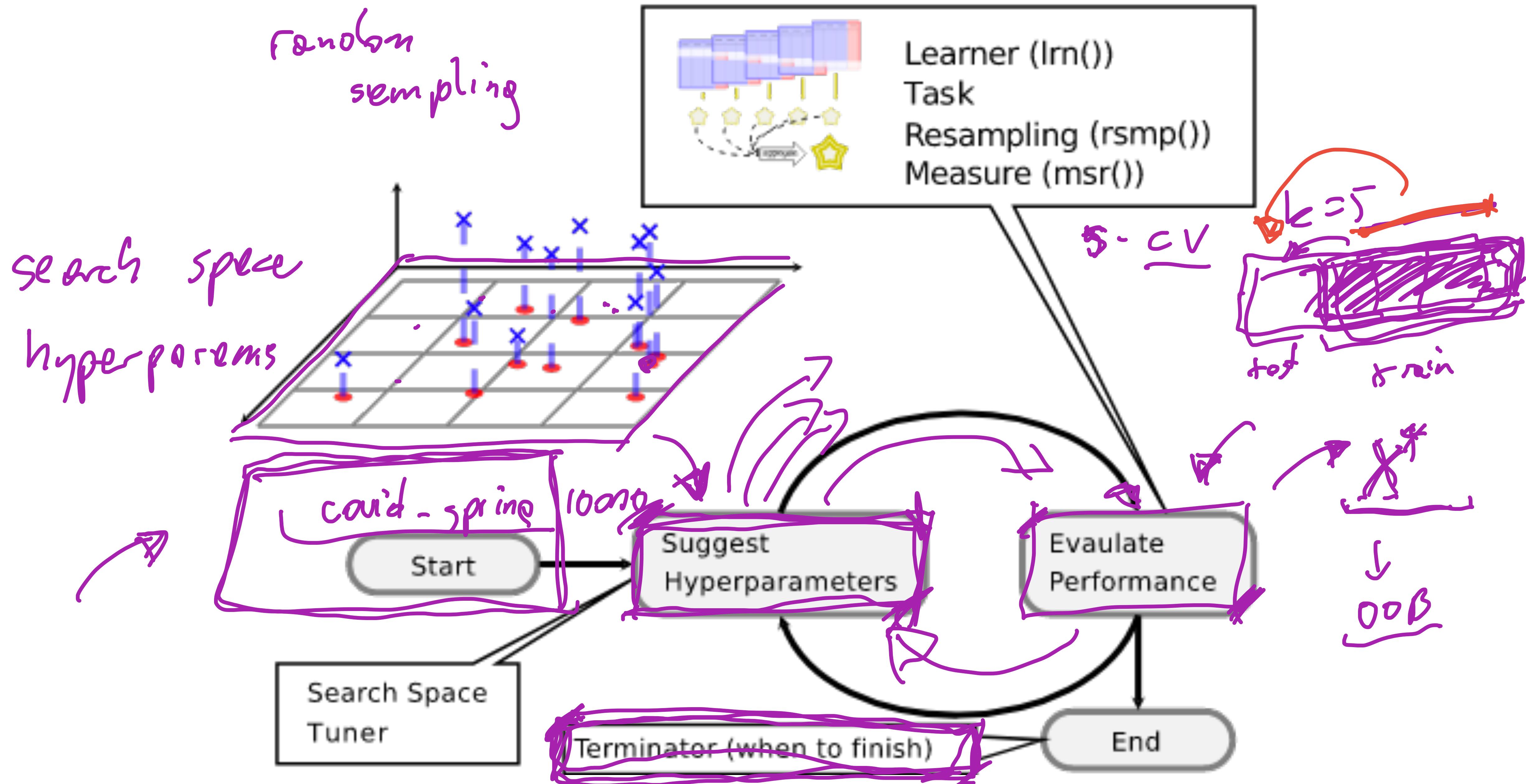


Day 1

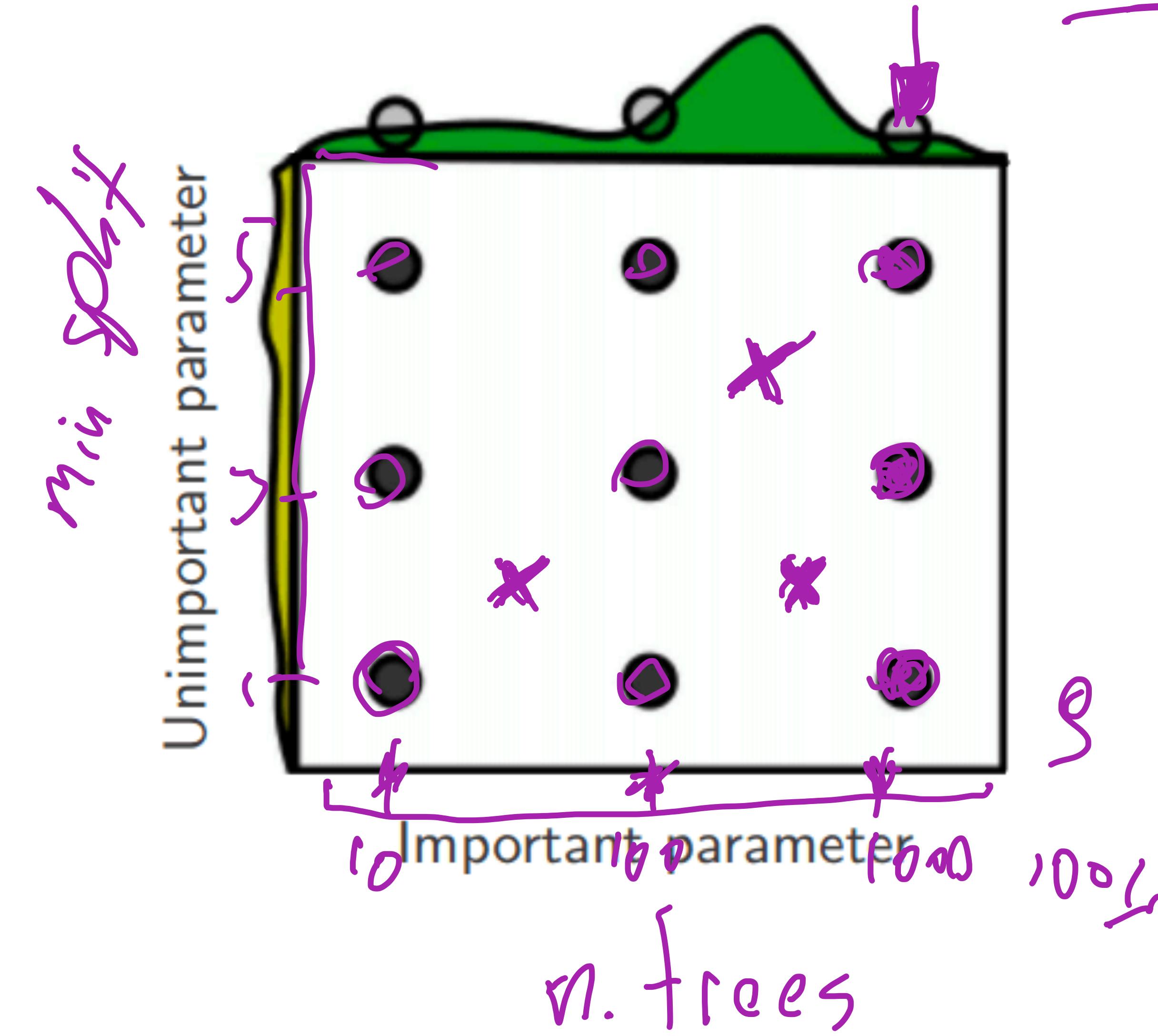
Predictive modeling

Part 4

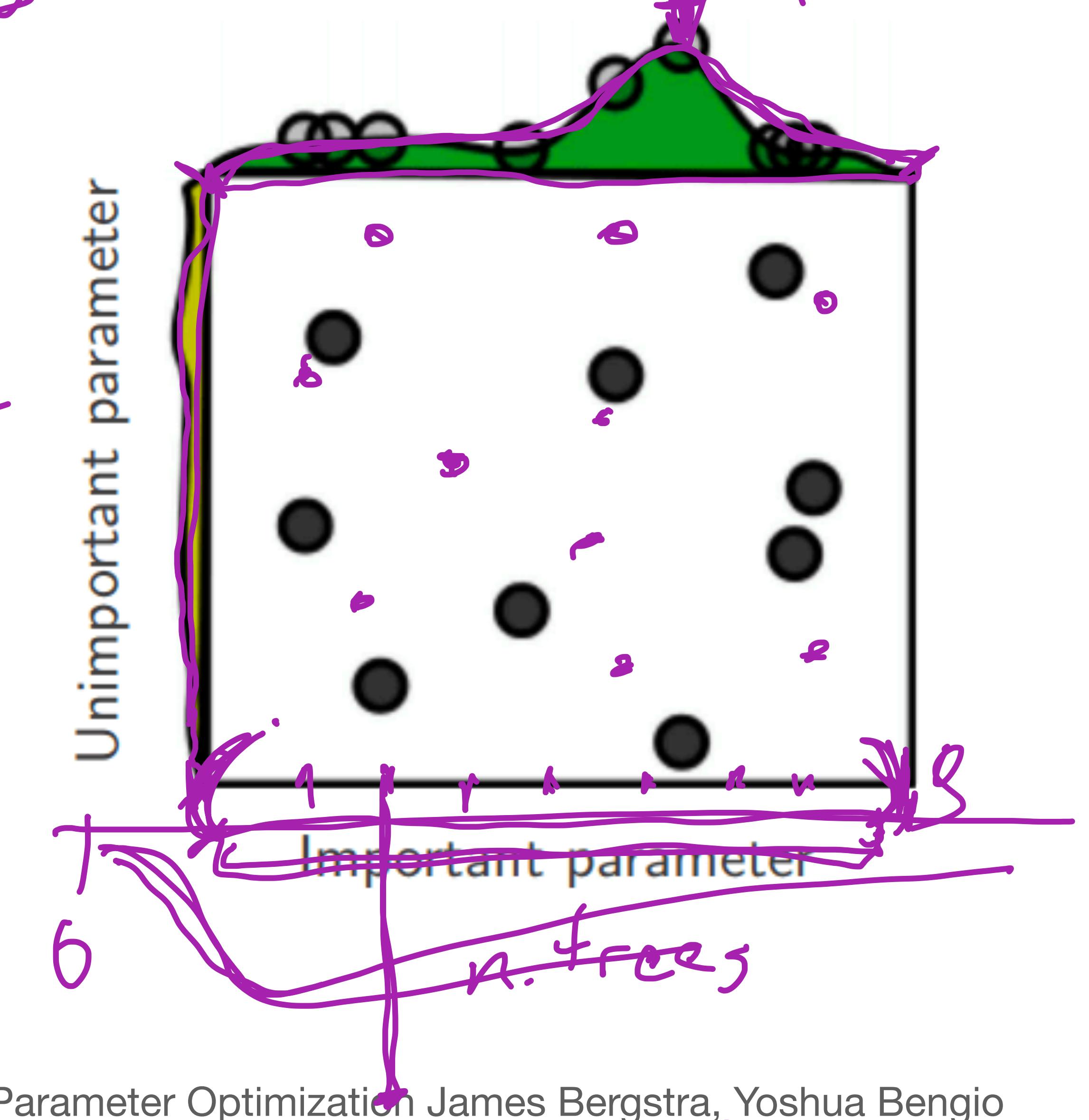
Hyperparameter optimization + Wrap-up



## Grid Layout



## Random Layout



Source: Random Search for Hyper-Parameter Optimization James Bergstra, Yoshua Bengio

**Explore**  
*Anscombe perspective*

Conception

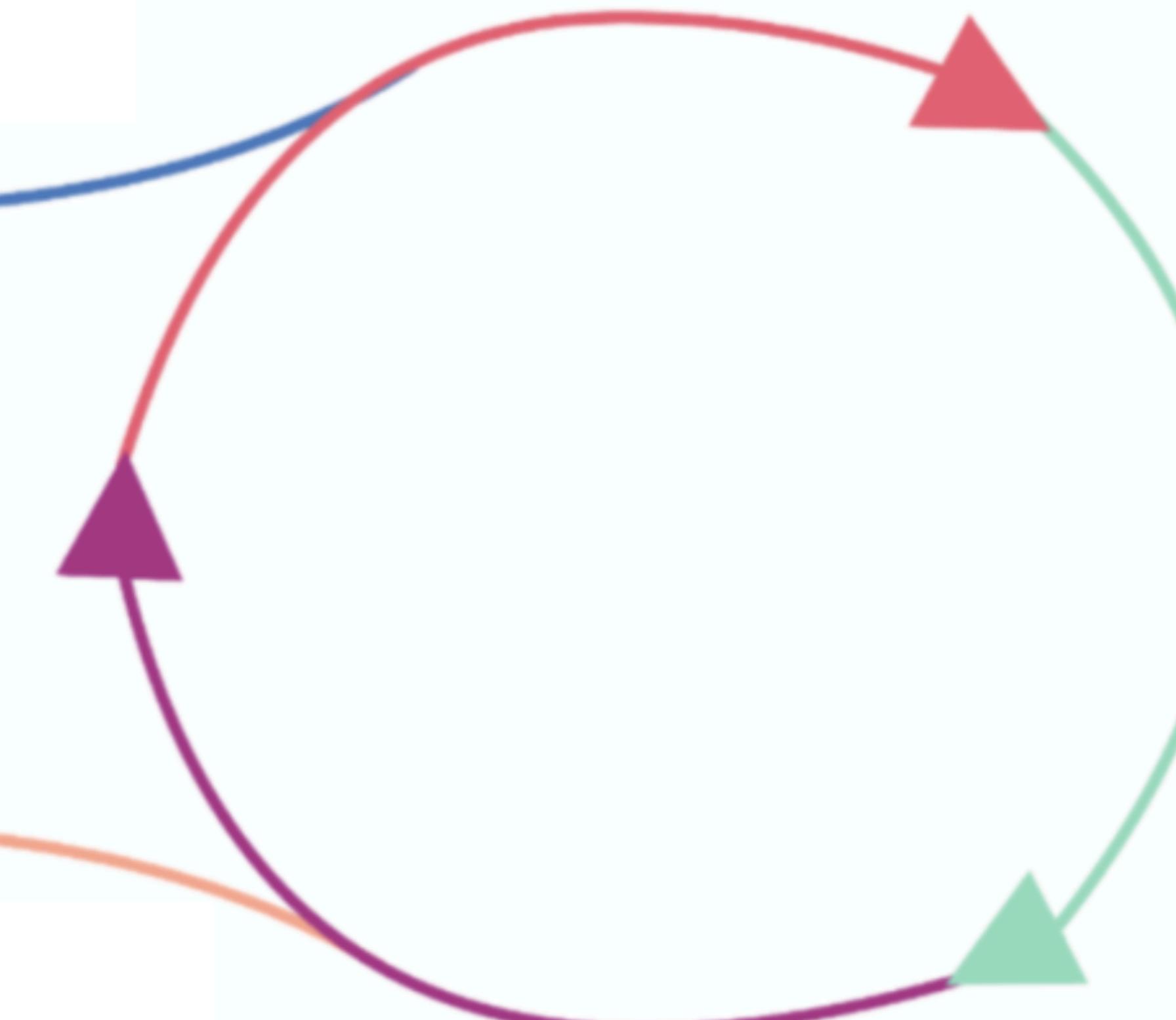


Model assembly  
*Occam perspective*

Delivery



Model evaluation  
*Rashomon perspective*



# Day 2

## Model exploration

## Machine Learning is Creating a Crisis in Science

The adoption of machine-learning techniques is contributing to a worrying number of research findings that cannot be repeated by other researchers.

**Kevin McCaney**

Wed, 02/27/2019 - 11:28



Photo credit: metamorworks/iStock

# Google Flu Trends

From Wikipedia, the free encyclopedia

**Google Flu Trends** was a [web service](#) operated by [Google](#). It provided estimates of [influenza](#) activity for more than 25 countries. By aggregating [Google Search](#) queries, it attempted to make accurate predictions about flu activity. This project was first launched in [2008](#) by Google.org to help predict outbreaks of flu.<sup>[1]</sup>

Google Flu Trends is now no longer publishing current estimates. Historical estimates are still available for download, and current data are offered

Forbes

Billionaires    Innovation    Leadership    Money    Con

61,215 views | Mar 23, 2014, 09:00am

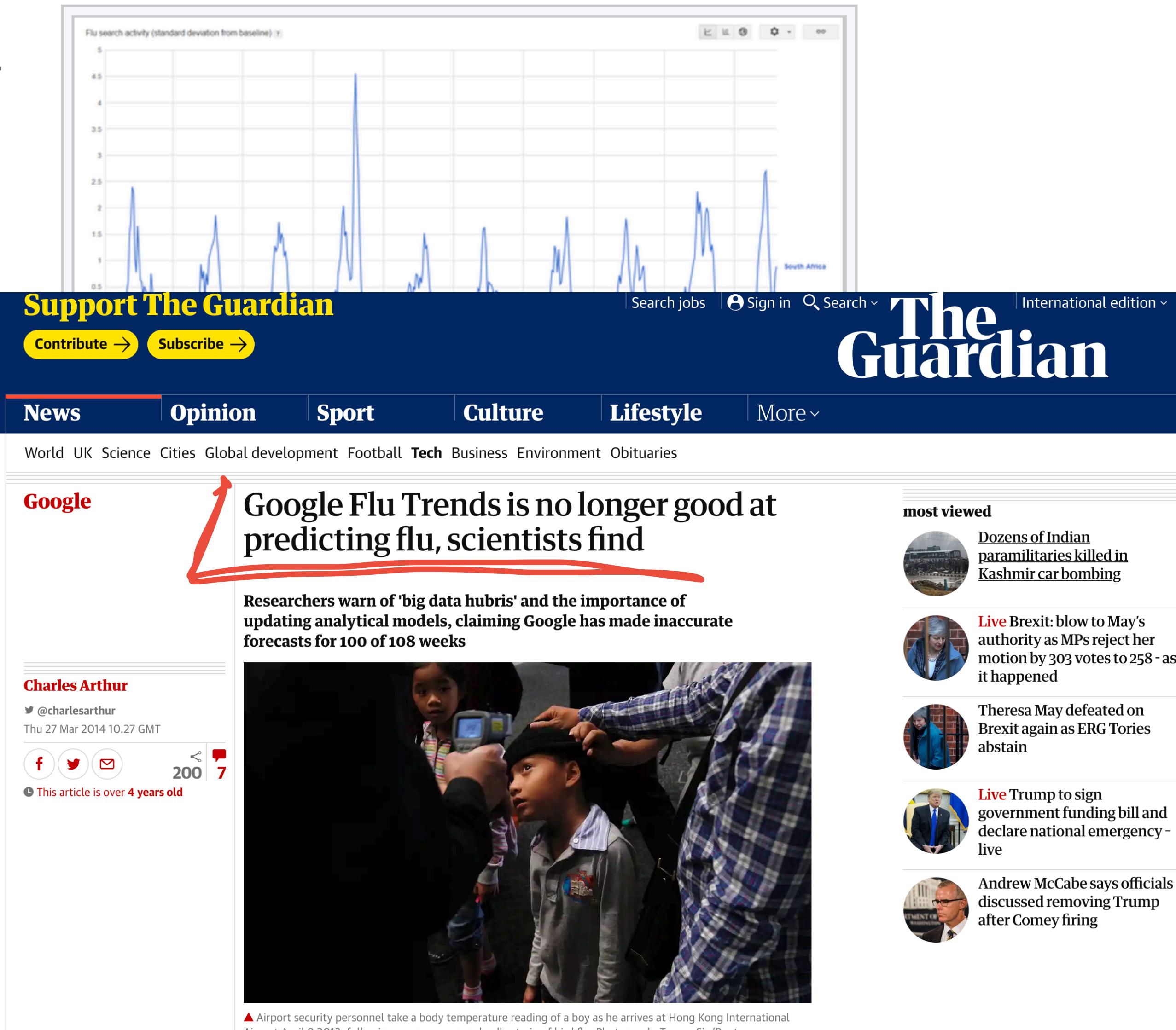
## Why Google Flu Is A Failure

Steven Salzberg Contributor 

Pharma & Healthcare

f It seemed like such a good idea at the time.

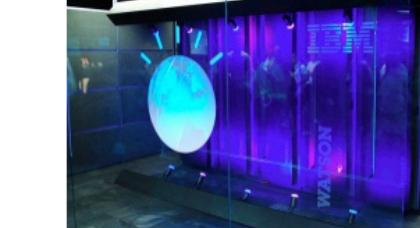
<https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/>





## **Report: IBM Watson delivered ‘unsafe and inaccurate’ cancer recommendations**

JULY 25, 2018 BY FINK DENSFORD — LEAVE A COMMENT



Internal documents from **IBM Watson Health** (NYSE:IBM) indicate that the company’s **Watson for Oncology** product often returns “multiple examples of unsafe and incorrect treatment recommendations,” according to a new report from **STAT News**.

The documents come from slides presented last year by IBM Watson Health’s deputy chief health officer, according to the report, and include feedback from customers that indicated the product is “often inaccurate” and that its recommendations bring to light “serious questions about the process for building content and the underlying technology.”

The issues were blamed on training the Watson product received by IBM engineers and physicians at the Memorial Sloan Kettering Cancer Center, which included “synthetic,” or hypothetical patients and cases, instead of real patient data, **STAT reports**.

BUSINESS NEWS

OCTOBER 10, 2018 / 5:12 AM / A MONTH AGO

## Amazon scraps secret AI showed bias against women

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's specialists uncovered a big problem: their

The group created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes. The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes, the people said.

Instead, the technology favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured," one person said.

### Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist

THE VERGE

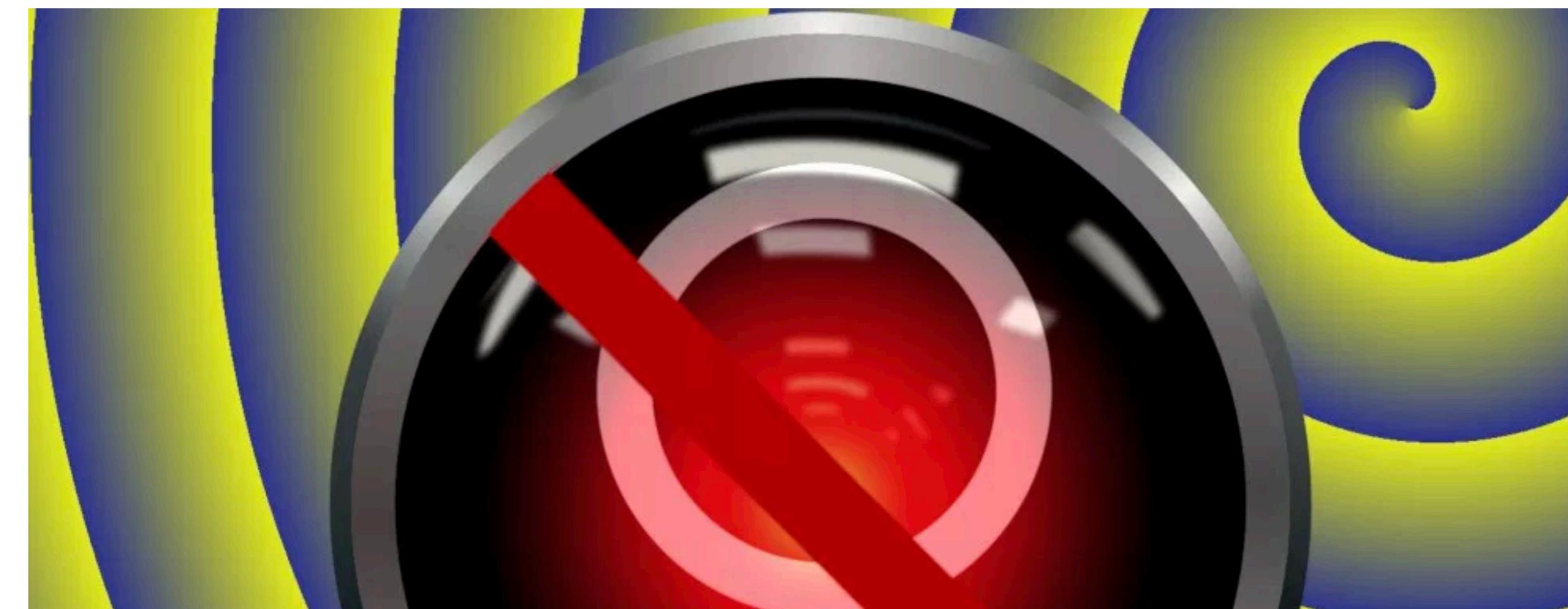
TECH ▾ SCIENCE ▾ C

TECH ▾ AMAZON ▾ ARTIFICIAL INTELLIGENCE

## Amazon reportedly scraps internal AI recruiting tool that was biased against women

*The secret program penalized applications that contained the word "women's"*

21



# COMPAS (software)

From Wikipedia, the free encyclopedia

**COMPAS**, an acronym for Correctional Offender Management Profiling for Alternative Sanctions, is a case management and decision support tool developed by Northpointe (now equivant<sup>↗</sup>) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.<sup>[1][2]</sup>

## Contents [hide]

### 1 Risk Assessment

- 1.1 Pretrial Release Risk scale
- 1.2 General Recidivism scale
- 1.3 Violent Recidivism scale

### 2 References

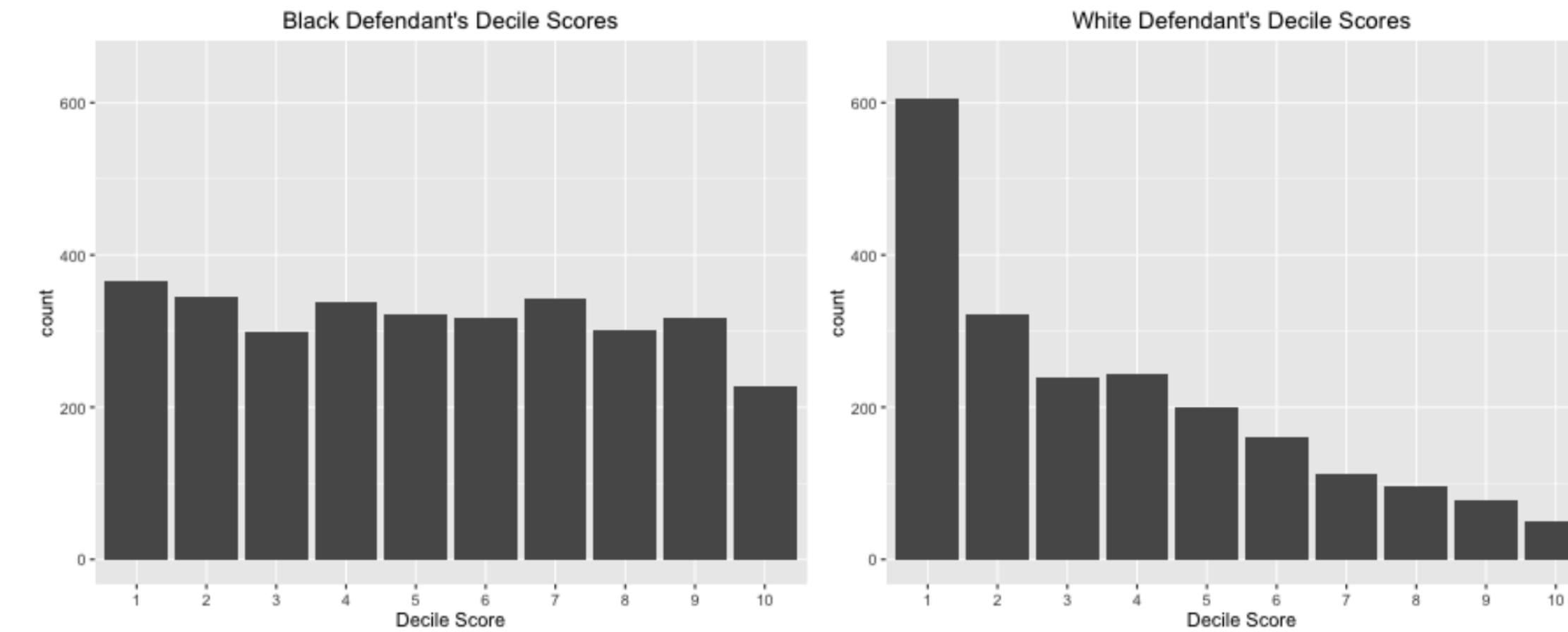
## Analysis

We analyzed the COMPAS scores for "Risk of Recidivism" and "Risk of Violent Recidivism." We did not analyze the COMPAS score for "Risk of Failure to Appear."

We began by looking at the risk of recidivism score. Our initial analysis looked at the simple distribution of the COMPAS decile scores among whites and blacks. We plotted the distribution of these scores for 6,172 defendants who had not been arrested for a new offense or who had recidivated within two years.

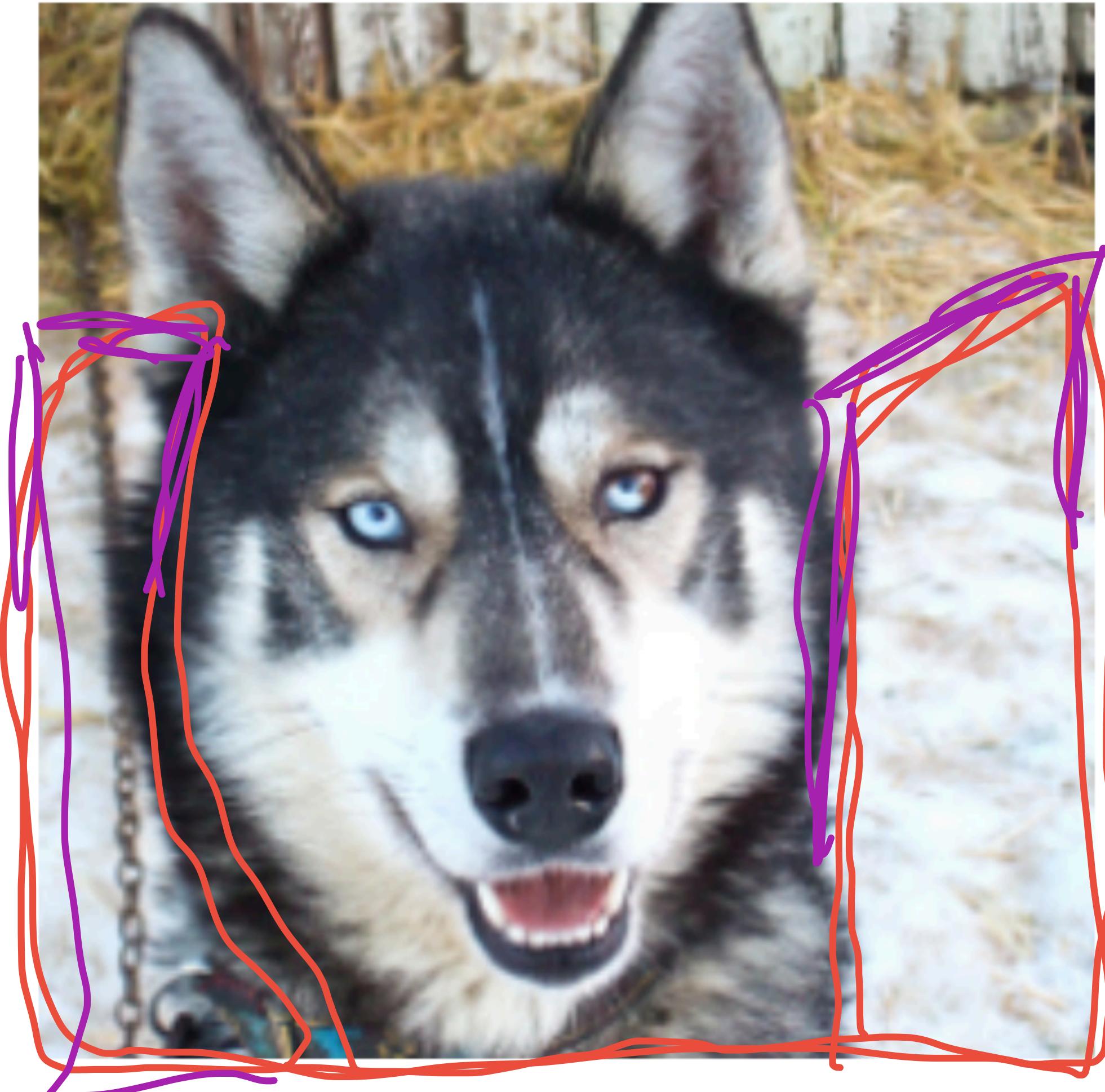
## Risk Assessment [ edit ]

The COMPAS software uses an algorithm to assess potential scales for general and violent recidivism, and for pretrial risk. In the Practitioner's Guide, the scales were designed using behavioral relevance to recidivism and criminal careers."<sup>[3]</sup>



These histograms show that scores for white defendants were skewed toward lower-risk categories, while black defendants were evenly distributed across scores. In our two-year sample, there were 3,175 black defendants and 2,103 white defendants, with 1,175 female defendants and 4,997 male defendants. There were 2,809 defendants who recidivated within two years in this sample.

# Model debugging



(a) Husky classified as wolf

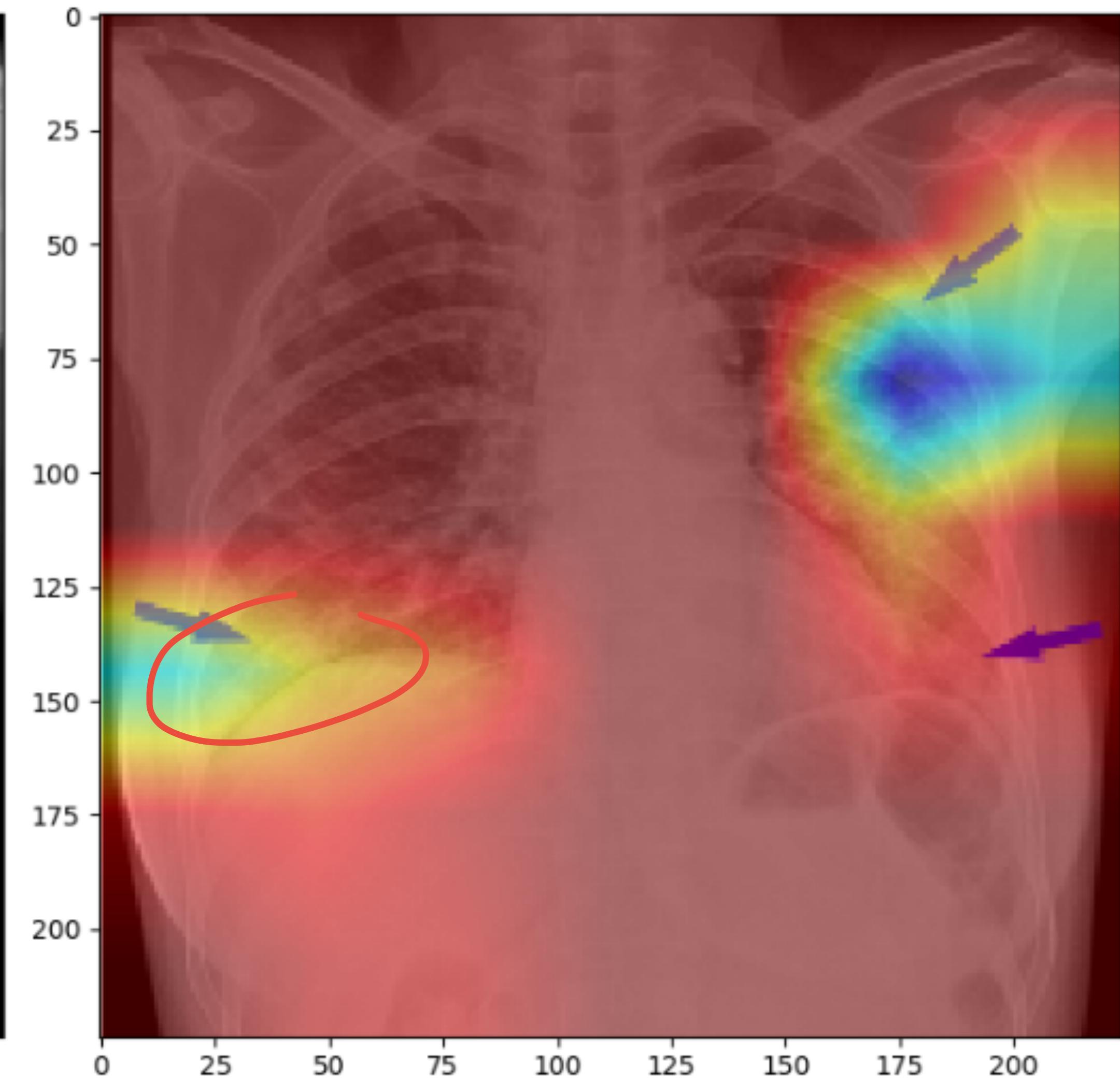
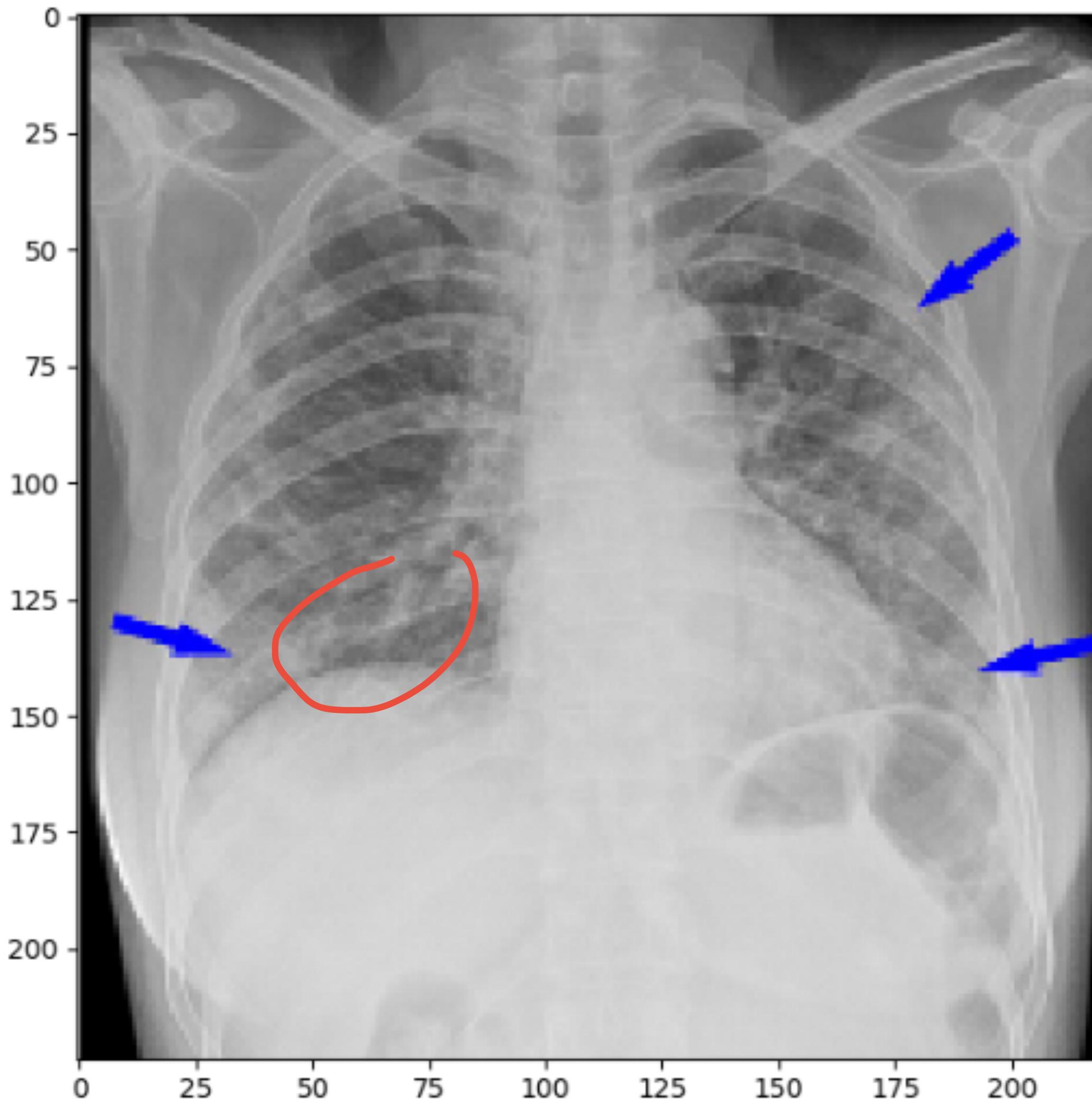


(b) Explanation

"Why Should I Trust You?" Explaining the Predictions of Any Classifier.

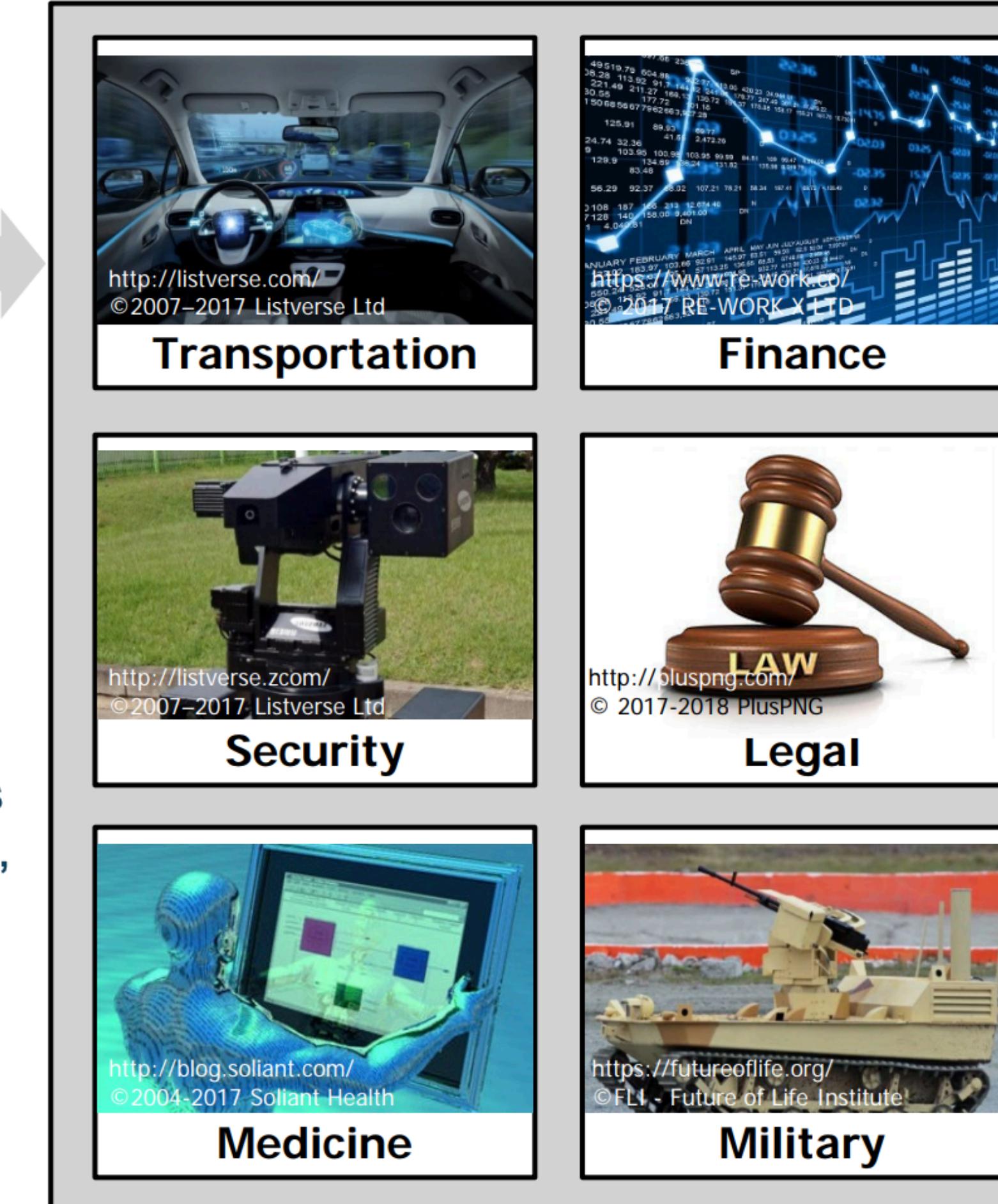
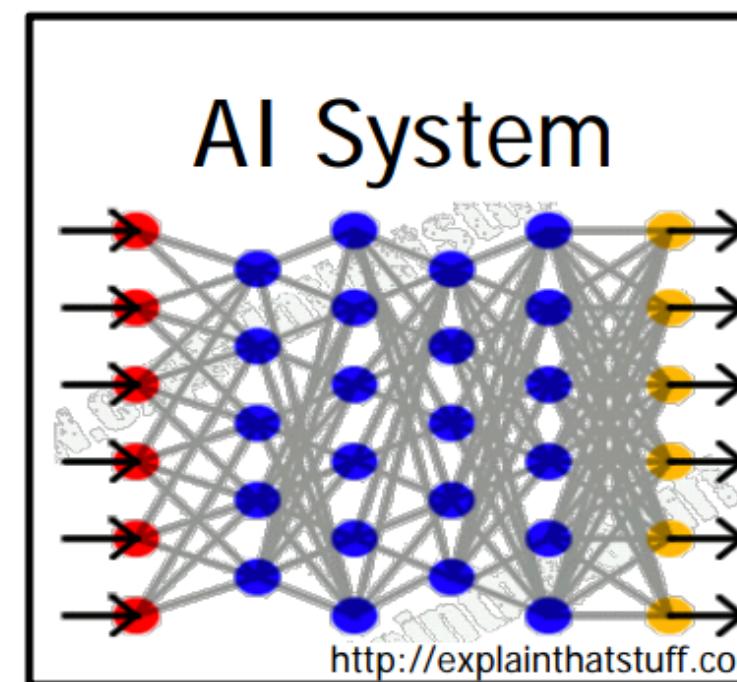
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016). <https://arxiv.org/pdf/1602.04938.pdf>

Ground Truth Class: 1 (COVID-19)  
Predicted Class: 1 (COVID-19)  
Prediction probabilities: ['0.00', '1.00']





# The Need for Explainable AI



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- The current generation of AI systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to users
- Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificially intelligent partners

# Right to explanation

From Wikipedia, the free encyclopedia

In the [regulation of algorithms](#), particularly [artificial intelligence](#) and its subfield of [machine learning](#), a **right to explanation** (or **right to an explanation**) is a [right](#) to be given an [explanation](#) for an output of the algorithm. Such rights primarily refer to [individual rights](#) to be given an explanation for decisions that significantly affect an individual, particularly legally or financially. For example, a person who applies for a loan and is denied may ask for an explanation, which could be "Credit bureau X reports that you declared bankruptcy last year; this is the main factor in considering you too likely to default, and thus we will not give you the loan you applied for."

Some such [legal rights](#) already exist, while the scope of a general "right to explanation" is a matter of ongoing debate.

## Contents [hide]

- 1 Examples
  - 1.1 Credit score in the United States
  - 1.2 European Union
  - 1.3 France
- 2 Criticism
- 3 See also
- 4 References
- 5 External links

# Prawo do wyjaśnienia decyzji kredytowej dla każdego! Sukces Panoptykonu!

13.02.2019



Po dzisiejszym posiedzeniu połączonych komisji sejmowych mamy dobre wiadomości: udało się przekonać rząd i posłów do wprowadzenia zmian w prawie bankowym, które zmienią proces przyznawania kredytów. W miejsce „czarnej skrzynki”, która wypluwa niezrozumiałe decyzje bez uzasadnienia, ma się pojawić przejrzysta procedura, wzmacniająca prawa osób ubiegających się o kredyt. To duży przełom w relacjach bank–klient i odpowiedź ustawodawcy na postulaty, które zgłaszamy od dawna.

## Krótką historią długiej walki

Od momentu, w którym pojawił się rządowy projekt tzw. [ustawy sektorowej](#) (wdrażającej RODO w 160 ustawach) i propozycje zmian w prawie bankowym, walczyliśmy o to, żeby każdy konsument starający się o kredyt miał prawo zażądać od banku wyjaśnienia oceny zdolności kredytowej i decyzji, którą w jego sprawie podjął bank. W pierwszej wersji projektu było przewidziane jedynie szczegółowe i nieprecyzyjne sformułowane prawo do wyjaśnienia, dotyczące wyłącznie decyzji podejmowanych w sposób automatyczny, czyli bez udziału człowieka. Uznaliśmy, że to za mało, i walczyliśmy o więcej. Tym bardziej, że od kilku lat prawo do wyjaśnienia oceny zdolności kredytowej (w każdej sytuacji) mają przedsiębiorcy.

Na pierwszym posiedzeniu specjalnej podkomisji powołanej do prac nad tym pakietem zaapelowaliśmy do rządu i posłów o wprowadzanie dalej idących gwarancji dla konsumentów. Ten pomysł poparli posłowie wszystkich partii, a przewodniczący podkomisji Edward Siarka wezwał Ministerstwo Cyfryzacji do wypracowania kompromisu ze „stroną społeczną” i przedstawicielami banków. Negocjacje i praca nad brzmieniem przepisów zajęły trzy tygodnie, ale warto było poczekać.

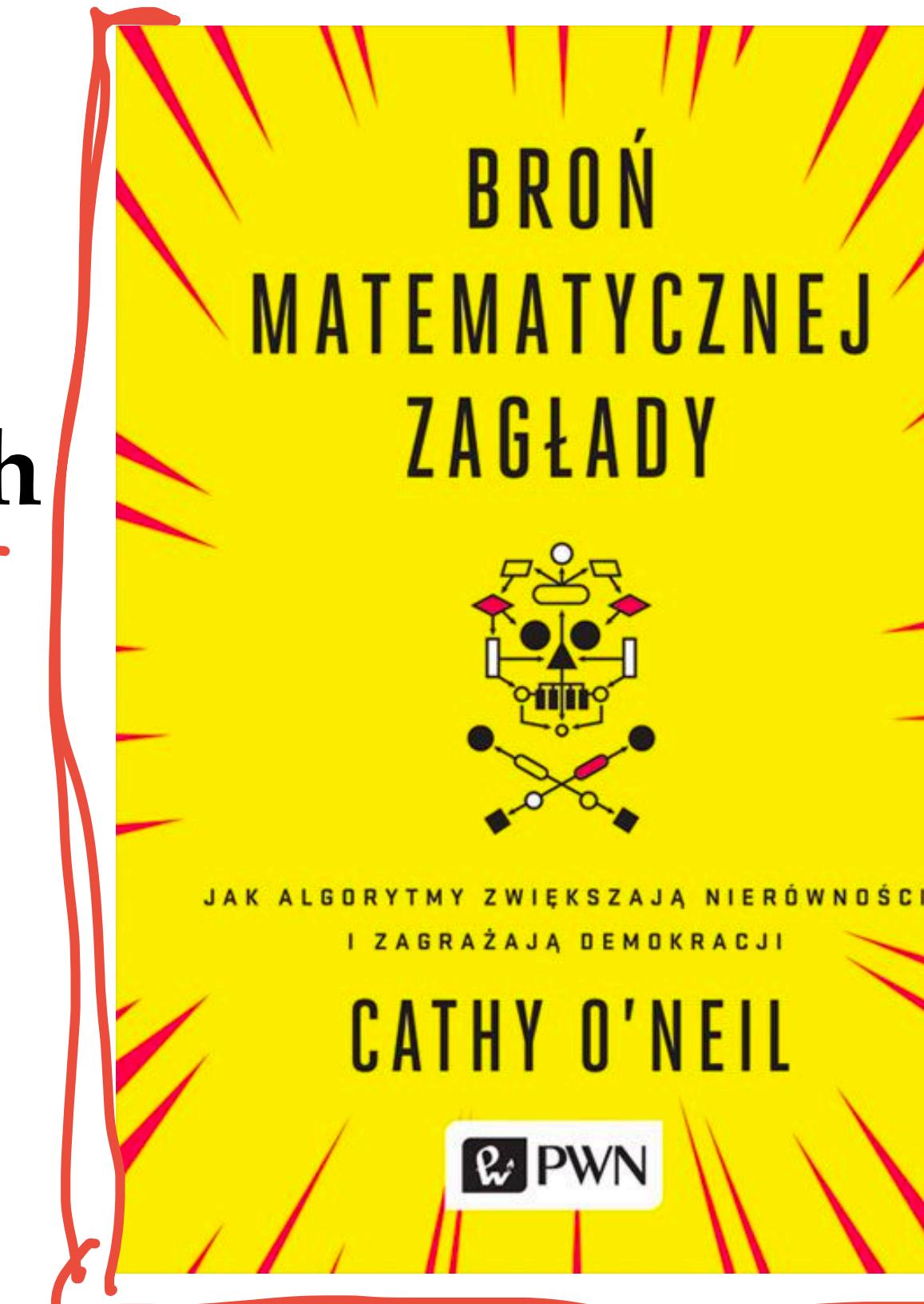
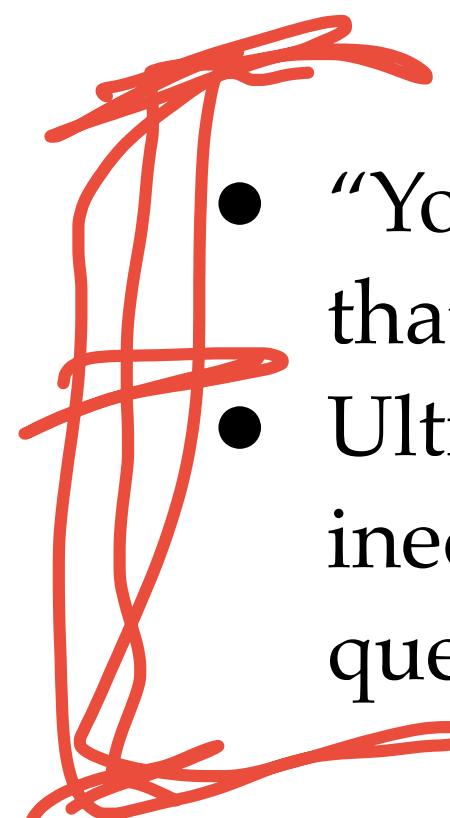
## Koniec z „czarną skrzynką” przy udzielaniu kredytów

**Na wniosek ubiegającego się o kredyt konsumenta bank przedstawi mu czynniki, w tym dane osobowe,**

Na wniosek ubiegającego się o kredyt konsumenta bank przedstawi mu czynniki, w tym dane osobowe, które miały wpływ na ocenę zdolności kredytowej. To kwintesencja zgłoszonej dzisiaj przez rząd i popartej przez posłów ze wszystkich ugrupowań zmiany w prawie bankowym.

## Cathy O'Neil: The era of blind faith

~~black boxes~~  
~~in big data must end~~



- “You don’t see a lot of skepticism,” she says. “The algorithms are like shiny new toys that we can’t resist using. We trust them so much that we project meaning on to them.”
- Ultimately algorithms, according to O’Neil, reinforce discrimination and widen inequality, “using people’s fear and trust of mathematics to prevent them from asking questions”.

# Day 2

## Model exploration

### Part 1

#### Model level analysis - variable importance



The central assumption of the  
machine learning models:

The future will be  
similar to the past

The central assumption of the

machine learning models:

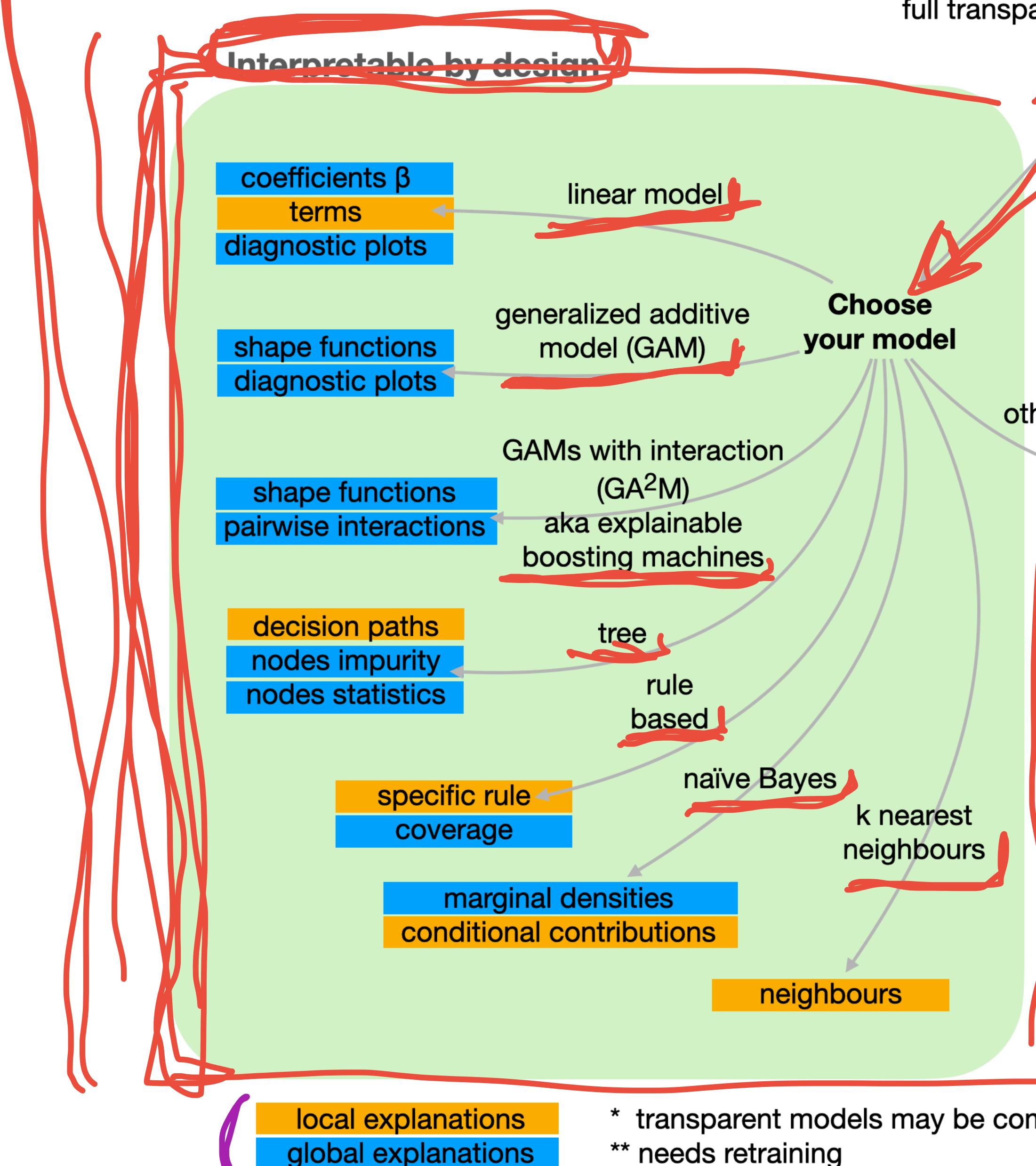
# COVID19

The future will be

similar to the past

# START HERE

Transparent model or  
rough explanations?



Choose  
your model

other

sorry  
let's try with  
rough explanation

full transparency \*

rough are ok

**Model specific**

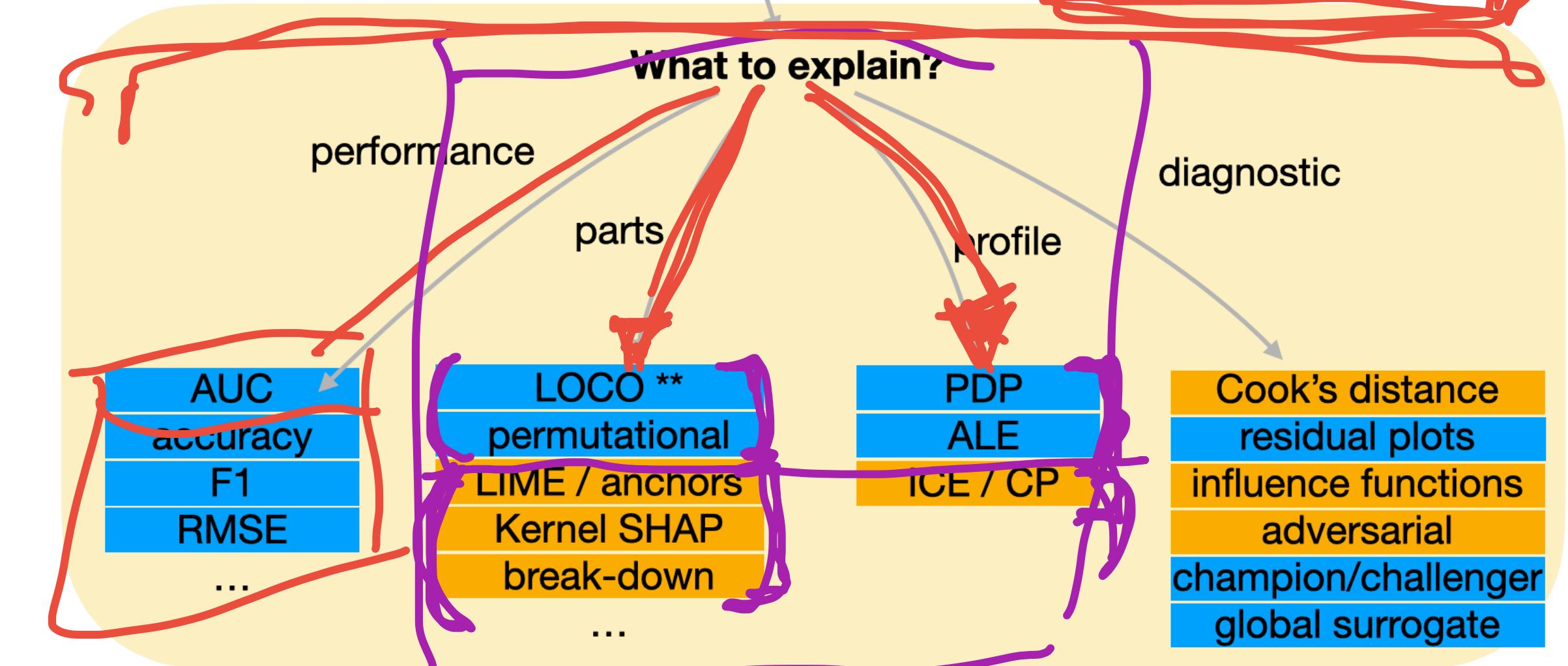
tree ensemble

neural network

**Model agnostic**

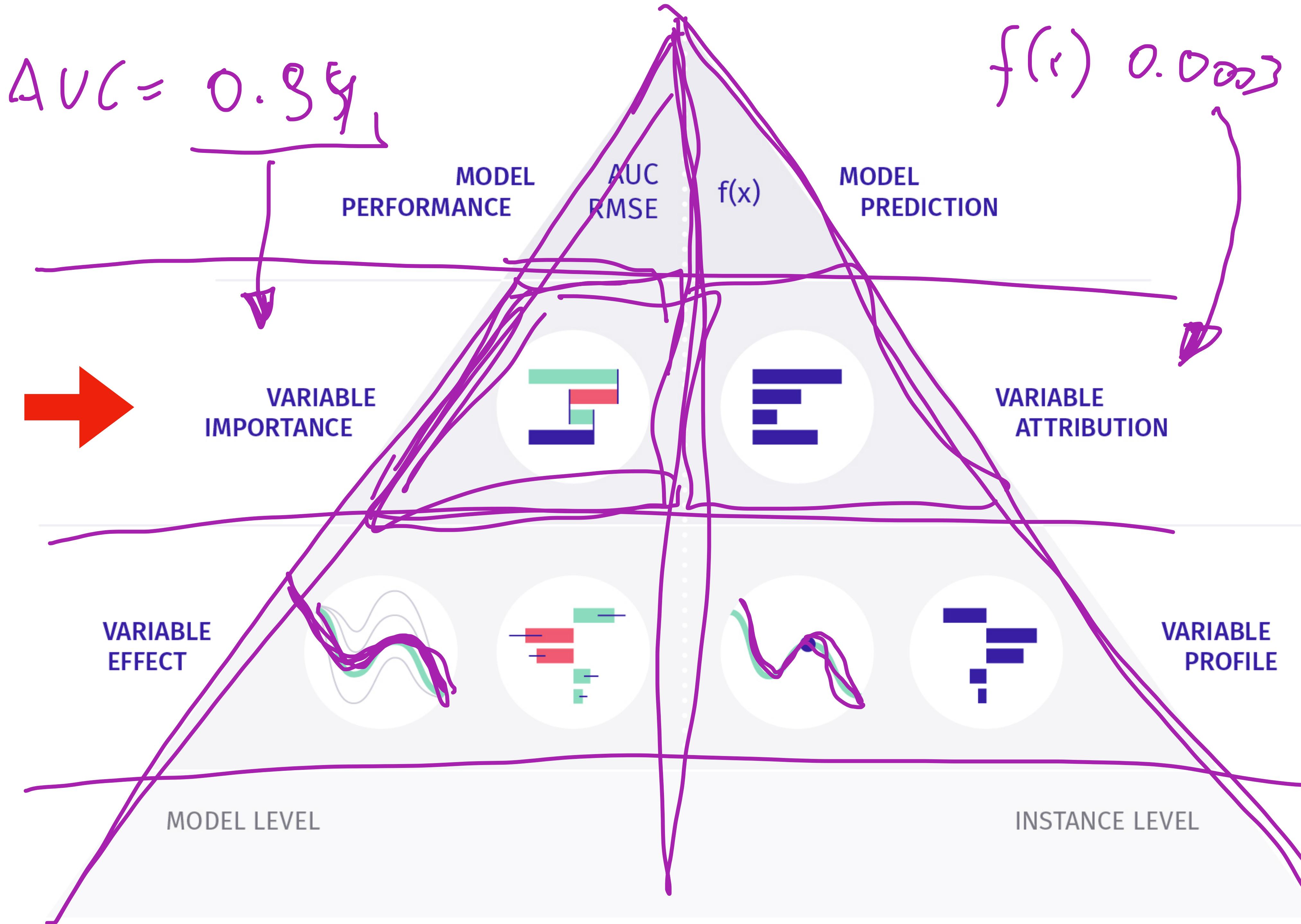
other  
(e.g. model stacking)

Choose  
your model



\* transparent models may be complex

\*\* needs retraining



# Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon" Perspective

arXiv.org > stat > arXiv:1801.01489

Search...

Help | Advance

Statistics > Methodology

## All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously

Aaron Fisher, Cynthia Rudin, Francesca Dominici

(Submitted on 4 Jan 2018 (v1), last revised 23 Dec 2019 (this version, v5))

Variable importance (VI) tools describe how much covariates contribute to a prediction model's accuracy. However, important variables for one well-performing model (for example, a linear model  $f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$  with a fixed coefficient vector  $\boldsymbol{\beta}$ ) may be unimportant for another model. In this paper, we propose model class reliance (MCR) as the range of VI values across all well-performing model in a prespecified class. Thus, MCR gives a more comprehensive description of importance by accounting for the fact that many prediction models, possibly of different parametric forms, may fit the data well. In the process of deriving MCR, we show several informative results for permutation-based VI estimates, based on the VI measures used in Random Forests. Specifically, we derive connections between permutation importance estimates for a single prediction model, U-statistics, conditional variable importance, conditional causal effects, and linear model coefficients. We then give probabilistic bounds for MCR, using a novel, generalizable technique. We apply MCR to a public data set of Broward County criminal records to study the reliance of recidivism prediction models on sex and race. In this application, MCR can be used to help inform VI for unknown, proprietary models.

Comments: The final, published article is now available at [this http URL](#). This version contains minor changes to the introductory sections, and some typo corrections. The title of this article changed twice during the revision process (see v1 and v3 on arxiv)

Subjects: **Methodology (stat.ME)**

Journal reference: Journal of Machine Learning Research 20 (177), 1–81, 2019

Cite as: [arXiv:1801.01489](#) [stat.ME]

(or [arXiv:1801.01489v5](#) [stat.ME] for this version)

### Bibliographic data

<https://arxiv.org/abs/1801.01489>



# Cynthia Rudin

Professor of Computer Science, ECE, and Statistics, [Duke University](#)  
Zweryfikowany adres z cs.duke.edu - [Strona główna](#)

OBSERWUJ

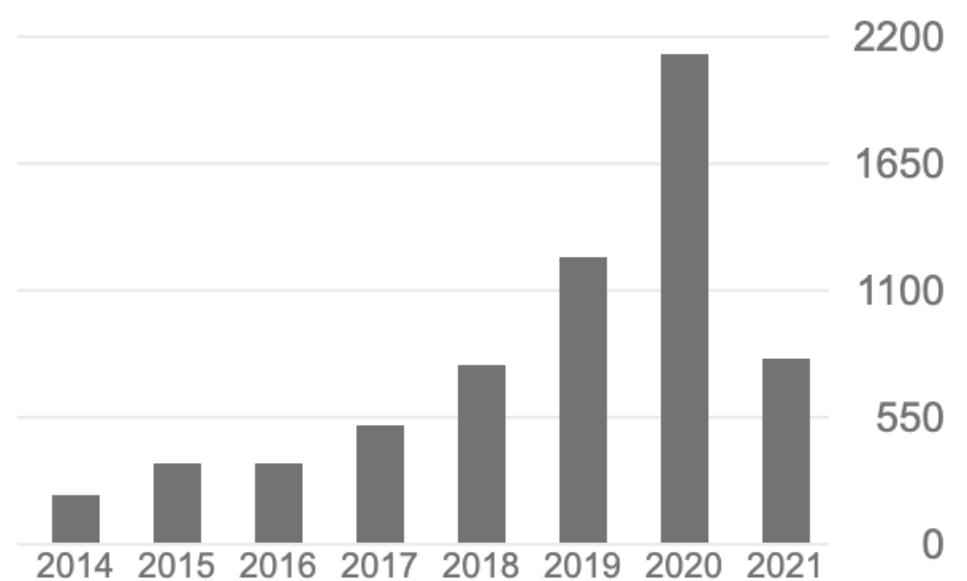
machine learning interpretability data science

TYTUŁ	CYTOWANE PRZEZ	ROK
<b>Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead</b> <small>C Rudin Nature Machine Intelligence 1 (5), 206-215</small>	976 *	2019
Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model <small>B Letham, C Rudin, TH McCormick, D Madigan Annals of Applied Statistics 9 (3), 1350-1371</small>	544 *	2015
All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. <small>A Fisher, C Rudin, F Dominici Journal of Machine Learning Research 20 (177), 1-81</small>	341 *	2019
The P-Norm Push: A simple convex ranking algorithm that concentrates at the top of the list <small>C Rudin The Journal of Machine Learning Research 10, 2233-2271</small>	235 *	2009
Supersparse linear integer models for optimized medical scoring systems <small>B Ustun, C Rudin Machine Learning 102 (3), 349-391</small>	230 *	2016
Machine learning for the New York City power grid <small>C Rudin, D Waltz, RN Anderson, A Boulanger, A Salleb-Aouissi, M Chow, ... IEEE transactions on pattern analysis and machine intelligence 34 (2), 328-345</small>	229	2011
Margin-based ranking and an equivalence between AdaBoost and RankBoost <small>C Rudin, RE Schapire The Journal of Machine Learning Research 10, 2193-2232</small>	214 *	2009
The bayesian case model: A generative approach for case-based reasoning and prototype classification <small>B Kim, C Rudin, JA Shah</small>	199	2014

Cytowane przez

[WYSWIETL WSZYSTKO](#)

	Wszystkie	Od 2016
Cytowania	7304	5854
h-indeks	41	35
i10-indeks	89	73



Dostęp publiczny

[WYSWIETL WSZYSTKO](#)

0 artykułów

23 artykuły

niedostępne

dostępne

Objęte finansowaniem

Współautorzy

[WYSWIETL WSZYSTKICH](#)



Berk Ustun  
Google AI / UCSD

>



Tyler H. McCormick  
University of Washington

>



David Madigan  
Professor of Statistics, Northeast...

>

# Variable importance for random forest

2021

← → ⌂ stat.berkeley.edu/~breiman/RandomForests/cc\_home.htm#varimp



## Variable importance

In every tree grown in the forest, put down the oob cases and count the number of votes cast for the correct class. Now randomly permute the values of variable m in the oob cases and put these cases down the tree.

Subtract the number of votes for the correct class in the variable-m-permuted oob data from the number of votes for the correct class in the untouched oob data. The average of this number over all trees in the forest is the raw importance score for variable m.

If the values of this score from tree to tree are independent, then the standard error can be computed by a standard computation. The correlations of these scores between trees have been computed for a number of data sets and proved to be quite low, therefore we compute standard errors in the classical way, divide the raw score by its standard error to get a z-score, and assign a significance level to the z-score assuming normality.

If the number of variables is very large, forests can be run once with all the variables, then run again using only the most important variables from the first run.

For each case, consider all the trees for which it is oob. Subtract the percentage of votes for the correct class in the variable-m-permuted oob data from the percentage of votes for the correct class in the untouched oob data. This is the local importance score for variable m for this case, and is used in the graphics program **RAFT**.

[https://www.stat.berkeley.edu/~breiman/RandomForests/cc\\_home.htm#varimp](https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#varimp)

# Do we need a new measure for Variable Importance?

Several common approaches for variable selection, or for describing relationships between variables, do not necessarily capture a variable's importance. Null hypothesis testing methods may identify a relationship, but do not describe the relationship's strength. Similarly, checking whether a variable is included by a sparse model-fitting algorithm, such as the Lasso (Hastie et al., 2009), does not describe the extent to which the variable is relied on. Partial dependence plots (Breiman et al., 2001; Hastie et al., 2009) can be difficult to interpret if multiple variables are of interest, or if the prediction model contains interaction effects.

Another common VI procedure is to run a model-fitting algorithm twice, first on all of the data, and then again after removing  $X_1$  from the data set. The losses for the two resulting models are then compared to determine the importance, or “necessity,” of  $X_1$  (Gevrey et al., 2003). Because this measure is a function of two prediction models rather than one, it does not measure how much either individual model relies on  $X_1$ . We refer

## Variable importance as the drop in the loss after variable permutation

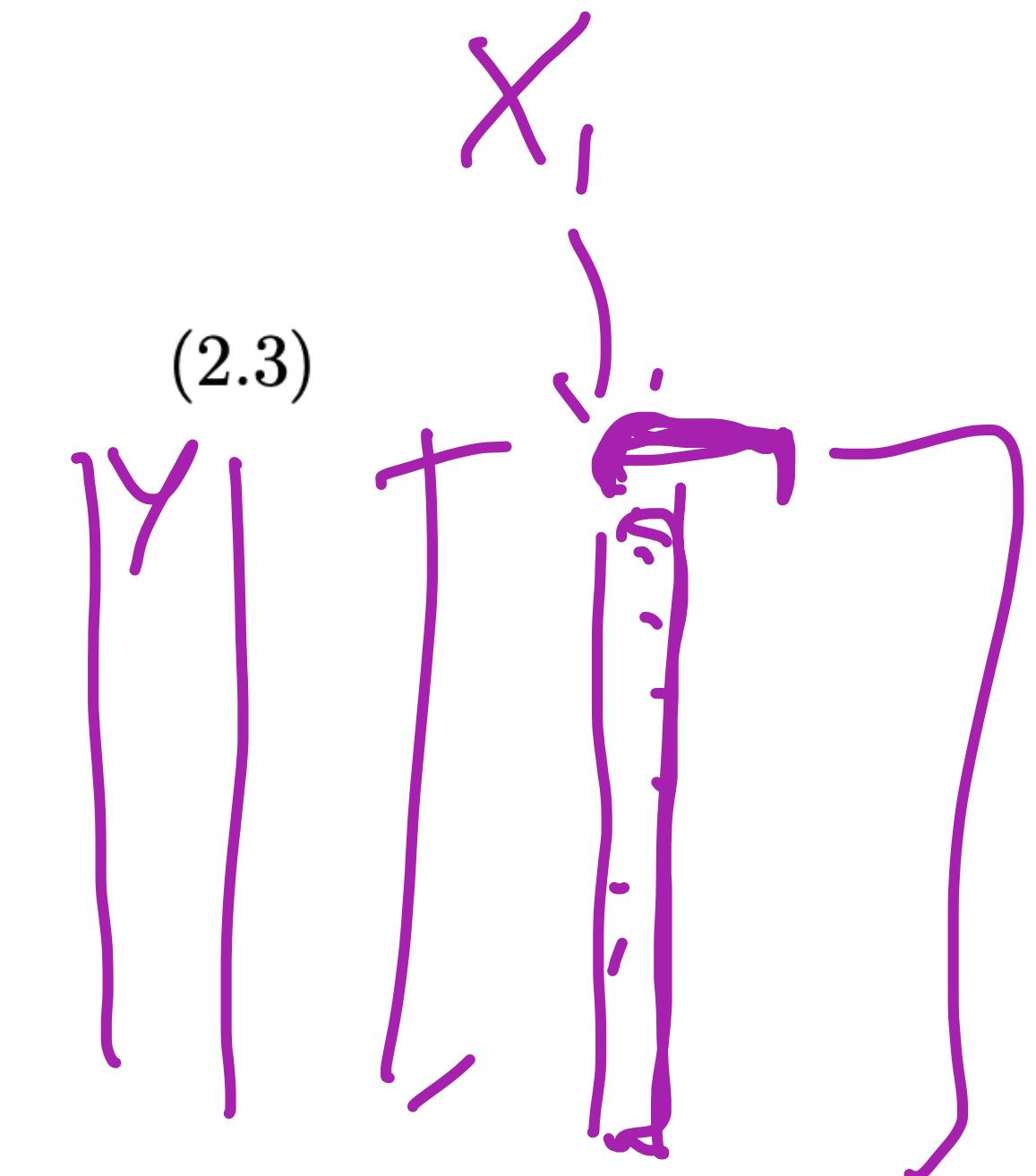
$$MR(f) := \frac{\text{Expected loss of } f \text{ under noise}}{\text{Expected loss of } f \text{ without noise}}$$

$$\frac{AVC(f, X^{[e_j]})}{AVC^{(2.1)}(f, X)}$$

The added noise must satisfy certain properties, namely, it must render  $X_1$  completely uninformative of the outcome  $Y$ , without altering the marginal distribution of  $X_1$  (for details, see Section 3, as well as Breiman, 2001; Breiman et al., 2001).

$$\widehat{MR}(f) := \frac{\text{In-sample loss of } f \text{ under noise}}{\text{In-sample loss of } f \text{ without noise}},$$

(2.3)



## One permutation is not enough. Check all of them

As a reference point, we compare  $e_{\text{switch}}(f)$  against the standard expected loss when none of the variables are switched,  $e_{\text{orig}}(f) := \mathbb{E}L(f, (Y, X_1, X_2))$ . From these two quantities, we formally define *model reliance* (MR) as the ratio,

$$MR(f) := \frac{e_{\text{switch}}(f)}{e_{\text{orig}}(f)}, \quad (3.1)$$

$$e_{\text{switch}}(f) := \mathbb{E}L\{f, (Y^{(b)}, X_1^{(a)}, X_2^{(b)})\}$$

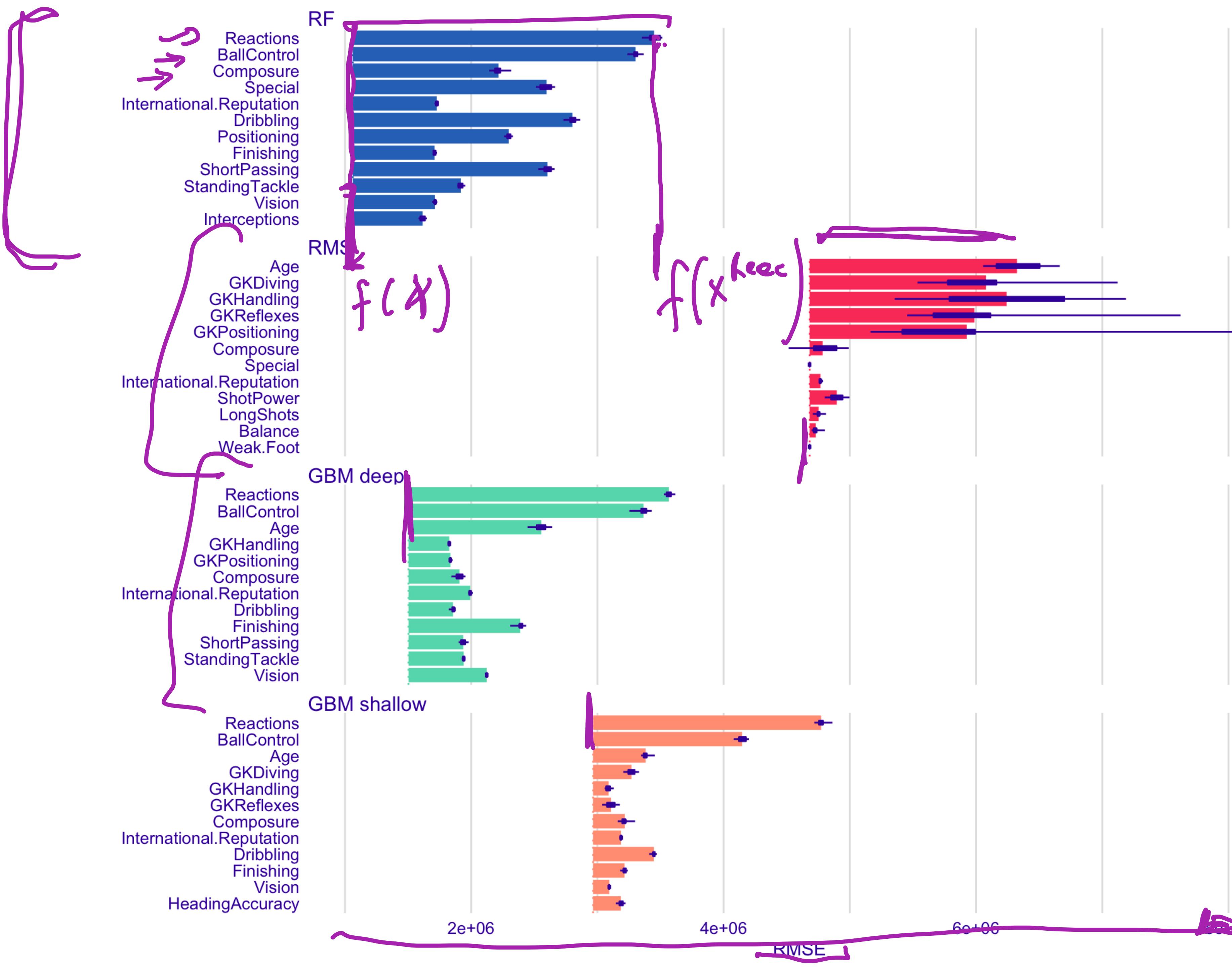
Given a model  $f$  and data set  $\mathbf{Z} = [\mathbf{y} \ \ \mathbf{X}]$ , we estimate  $MR(f)$  by separately estimating the numerator and denominator of Eq 3.1. We estimate  $e_{\text{orig}}(f)$  with the standard empirical loss,

$$\hat{e}_{\text{orig}}(f) := \frac{1}{n} \sum_{i=1}^n L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[i,\cdot]})\}. \quad (3.2)$$

We estimate  $e_{\text{switch}}(f)$  by performing a “switch” operation across all observed pairs, as in

$$\hat{e}_{\text{switch}}(f) := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} L\{f, (\mathbf{y}_{[j]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]})\}. \quad (3.3)$$

$B = 100$



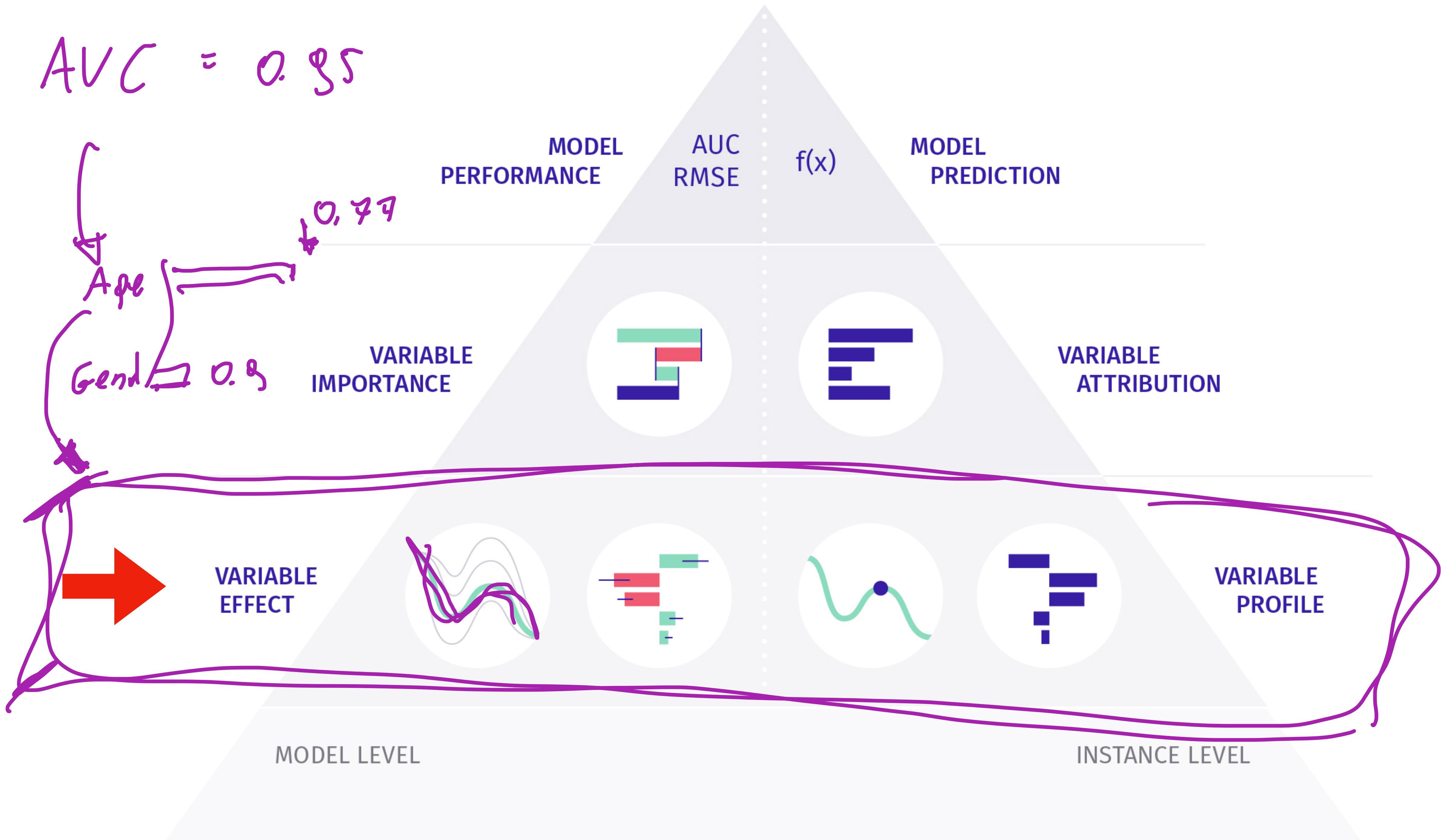
Day 2

Model exploration

Part 2

Model level analysis - variable profile

$$AVC = 0.95$$



# Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models

arXiv.org &gt; stat &gt; arXiv:1612.08468

Search...

All fields

Help | Advanced Search

Statistics &gt; Methodology

## Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models

Daniel W. Apley, Jingyu Zhu

(Submitted on 27 Dec 2016 (v1), last revised 19 Aug 2019 (this version, v2))

When fitting black box supervised learning models (e.g., complex trees, neural networks, boosted trees, random forests, nearest neighbors, local kernel-weighted methods, etc.), visualizing the main effects of the individual predictor variables and their low-order interaction effects is often important, and partial dependence (PD) plots are the most popular approach for accomplishing this. However, PD plots involve a serious pitfall if the predictor variables are far from independent, which is quite common with large observational data sets. Namely, PD plots require extrapolation of the response at predictor values that are far outside the multivariate envelope of the training data, which can render the PD plots unreliable. Although marginal plots (M plots) do not require such extrapolation, they produce substantially biased and misleading results when the predictors are dependent, analogous to the omitted variable bias in regression. We present a new visualization approach that we term accumulated local effects (ALE) plots, which inherits the desirable characteristics of PD and M plots, without inheriting their preceding shortcomings. Like M plots, ALE plots do not require extrapolation; and like PD plots, they are not biased by the omitted variable phenomenon. Moreover, ALE plots are far less computationally expensive than PD plots.

Comments: The R package ALEPlot is available on CRAN. The new version contains refined definitions of ALE effects, a new illustrative example, theorems and proofs of asymptotic properties of ALE effects and estimators, and extra implementation details

Subjects: **Methodology (stat.ME)**

Cite as: [arXiv:1612.08468](#) [stat.ME]

(or [arXiv:1612.08468v2](#) [stat.ME] for this version)

### Bibliographic data

Select data provider: [Semantic Scholar](#) | [Prophy](#) [Disable Bibex(What is Bibex?)]

- No references available from data provider.

Articles recently added or updated may not have propagated to data providers yet. If you believe there is an error, contact [Semantic Scholar](#).

### Submission history

<https://arxiv.org/abs/1612.08468>

### Download:

- [PDF](#)
- [Other formats](#)  
(license)

Current browse corner

stat.ME

< prev | next >  
new | recent | 1612

Change to browse mode

stat

References & Citations

- [NASA ADS](#)

Export citation  
[Google Scholar](#)

### Bookmark



Science  
WISE



# Daniel Apley

[Northwestern University](#)

Zweryfikowany adres z u.northwestern.edu

statistics machine learning manufacturing quality control

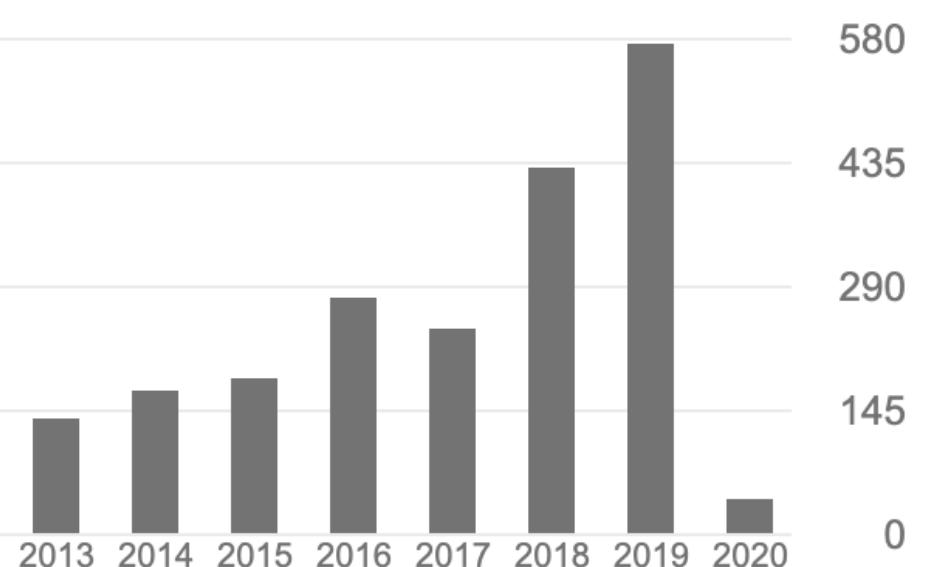
OBSERWUJ

Cytowane przez

[WYSWIETL WSZYSTKO](#)

[Wszystkie](#) [Od 2015](#)

Cytowania	3202	1754
h-indeks	26	22
i10-indeks	61	40



TYTUŁ	CYTOWANE PRZEZ	ROK
<a href="#">Quantification of model uncertainty: Calibration, model discrepancy, and identifiability</a> PD Arendt, DW Apley, W Chen Journal of Mechanical Design 134 (10), 100908	198	2012
<a href="#">Understanding the effects of model uncertainty in robust design with computer experiments</a> DW Apley, J Liu, W Chen Journal of Mechanical Design 128 (4), 945-958	183	2006
<a href="#">Diagnosis of multiple fixture faults in panel assembly</a> DW Apley, J Shi	172	1998
<a href="#">Local Gaussian process approximation for large computer experiments</a> RB Gramacy, DW Apley Journal of Computational and Graphical Statistics 24 (2), 561-578	158	2015
<a href="#">The GLRT for statistical process control of autocorrelated processes</a> DW Apley, J Shi IIE transactions 31 (12), 1123-1134	155	1999
<a href="#">A non-stationary covariance-based Kriging method for metamodeling in engineering design</a> Y Xiong, W Chen, D Apley, X Ding International Journal for Numerical Methods in Engineering 71 (6), 733-756	148	2007
<a href="#">A better understanding of model updating strategies in validating engineering models</a> Y Xiong, W Chen, KL Tsui, DW Apley Computer methods in applied mechanics and engineering 198 (15-16), 1327-1337	139	2009
<a href="#">A factor-analysis method for diagnosing variability in multivariate manufacturing processes</a> DW Apley, J Shi Technometrics 43 (1), 84-95	126	2001
<a href="#">A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality</a> MA Bessa, R Bostanabad, Z Liu, A Hu, DW Apley, C Brinson, W Chen, ... Computer Methods in Applied Mechanics and Engineering 320, 633-667	125	2017
<a href="#">The Autoregressive T2 Chart for Monitoring Univariate Autocorrelated Processes</a> DW Apley, F Tsung Journal of Quality Technology 34 (1), 80-96	116	2002

## Dataset level

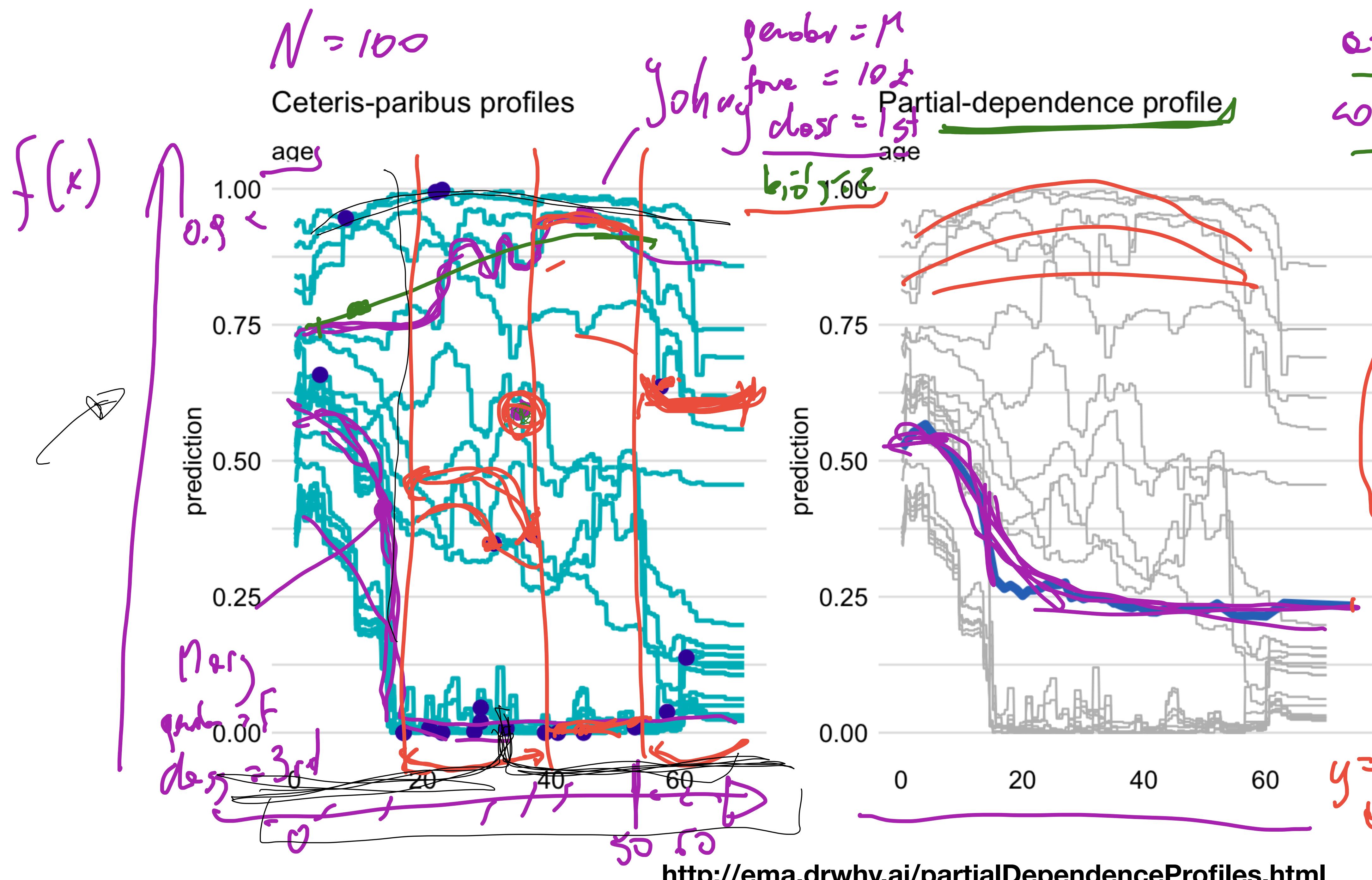
The value of a PD profile for model  $f()$  and explanatory variable  $X^j$  at  $z$  is defined as follows:

\*

$$g_{PD}^{f,j}(z) = E_{X^{-j}}[f(X^j|z)]. \quad (18.1)$$

Usually, we do not know the true distribution of  $X^{-j}$ . We can estimate it, however, by the empirical distribution of  $N$ , say, observations available in a training dataset. This leads to the use of the average of CP profiles for  $X^j$  as an estimator of the PD profile:

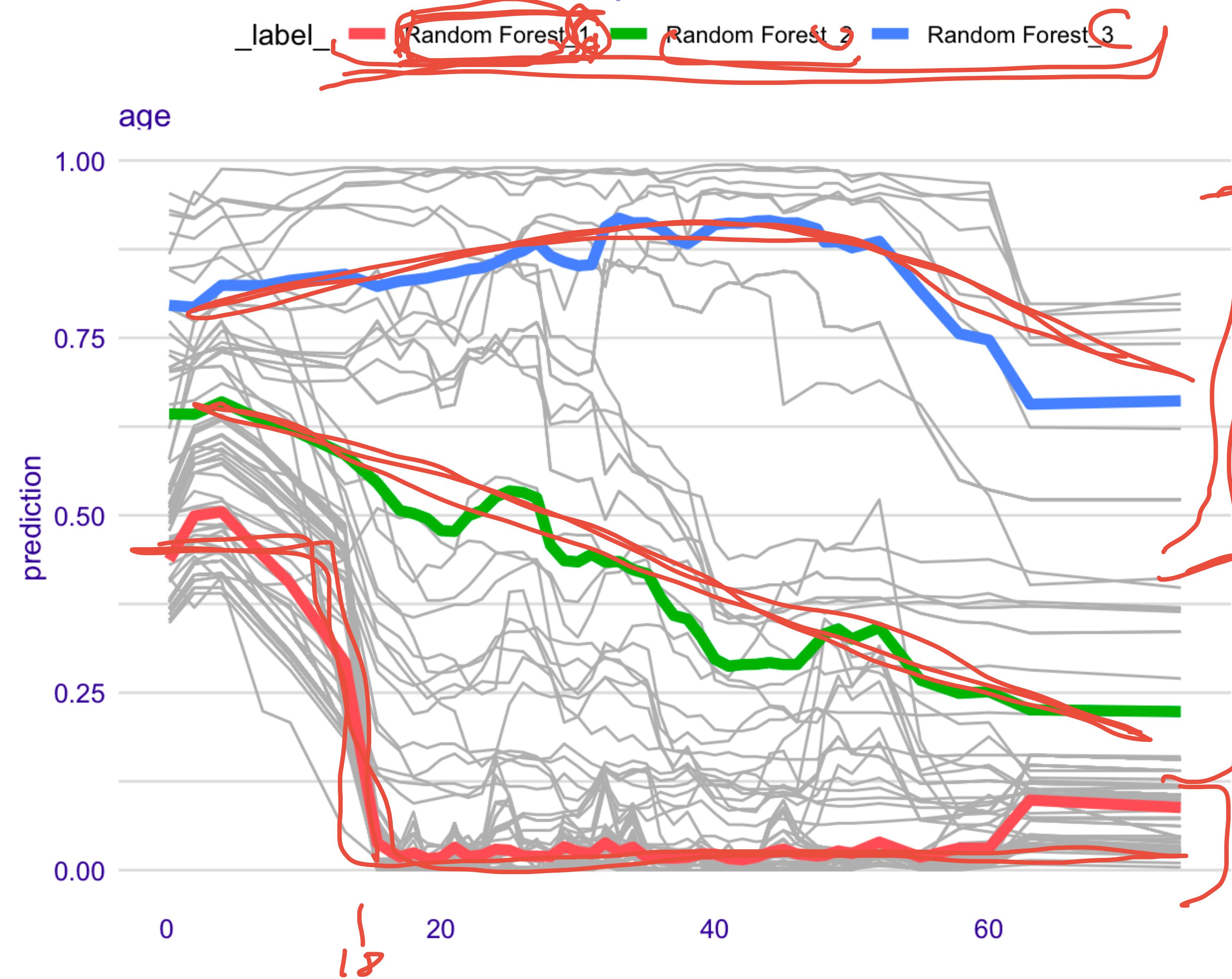
$$\hat{g}_{PD}^{f,j}(z) = \frac{1}{N} \sum_{i=1}^N f(x_i^j|z) \quad (18.2)$$



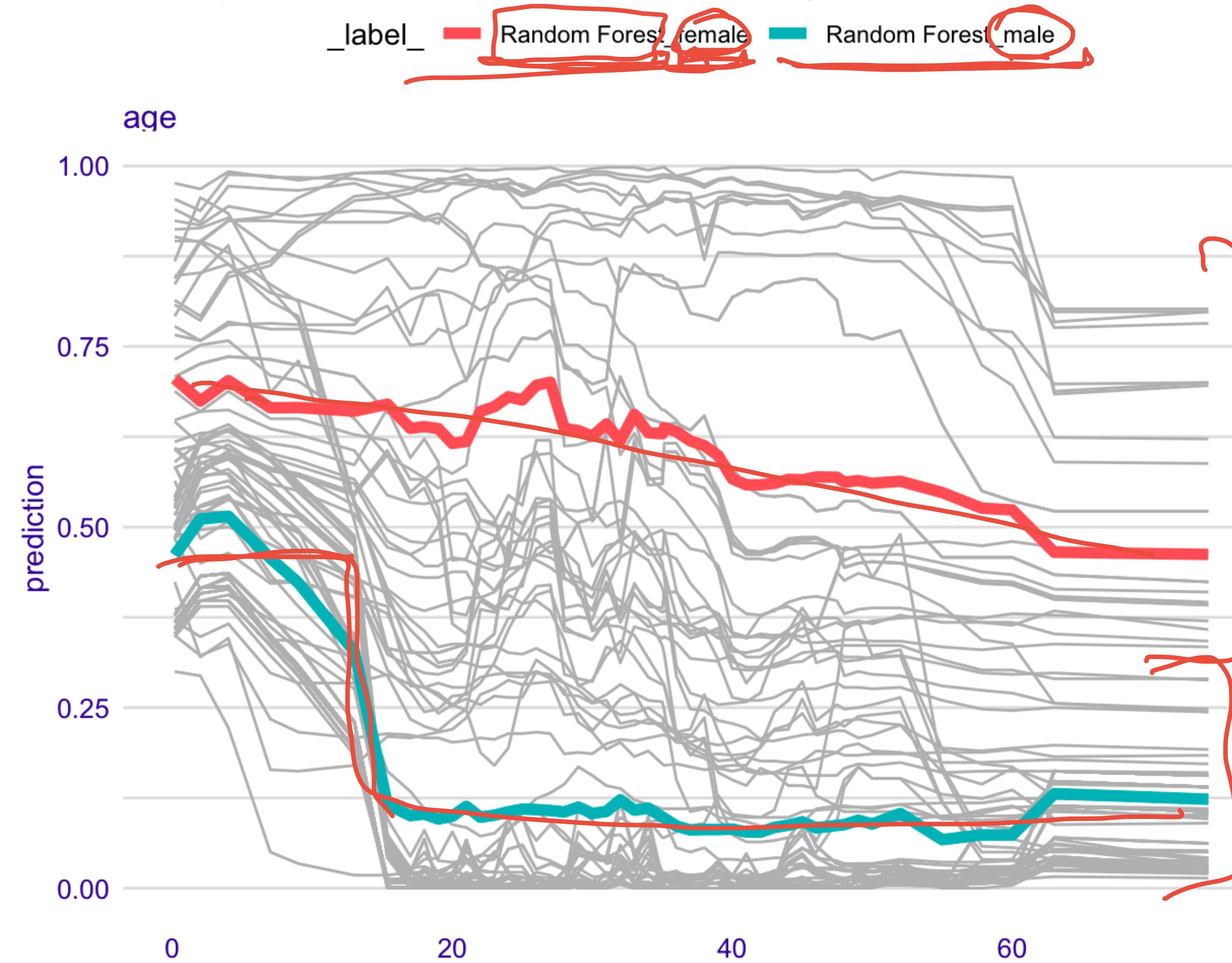
<http://ema.drwhy.ai/partialDependenceProfiles.html>

additiveness  
correlation  
 Accumulated  
 Local  
 Effects  
 ALE  
 $y = 0.6 \cdot x_1 + 0.3 \cdot x_2$

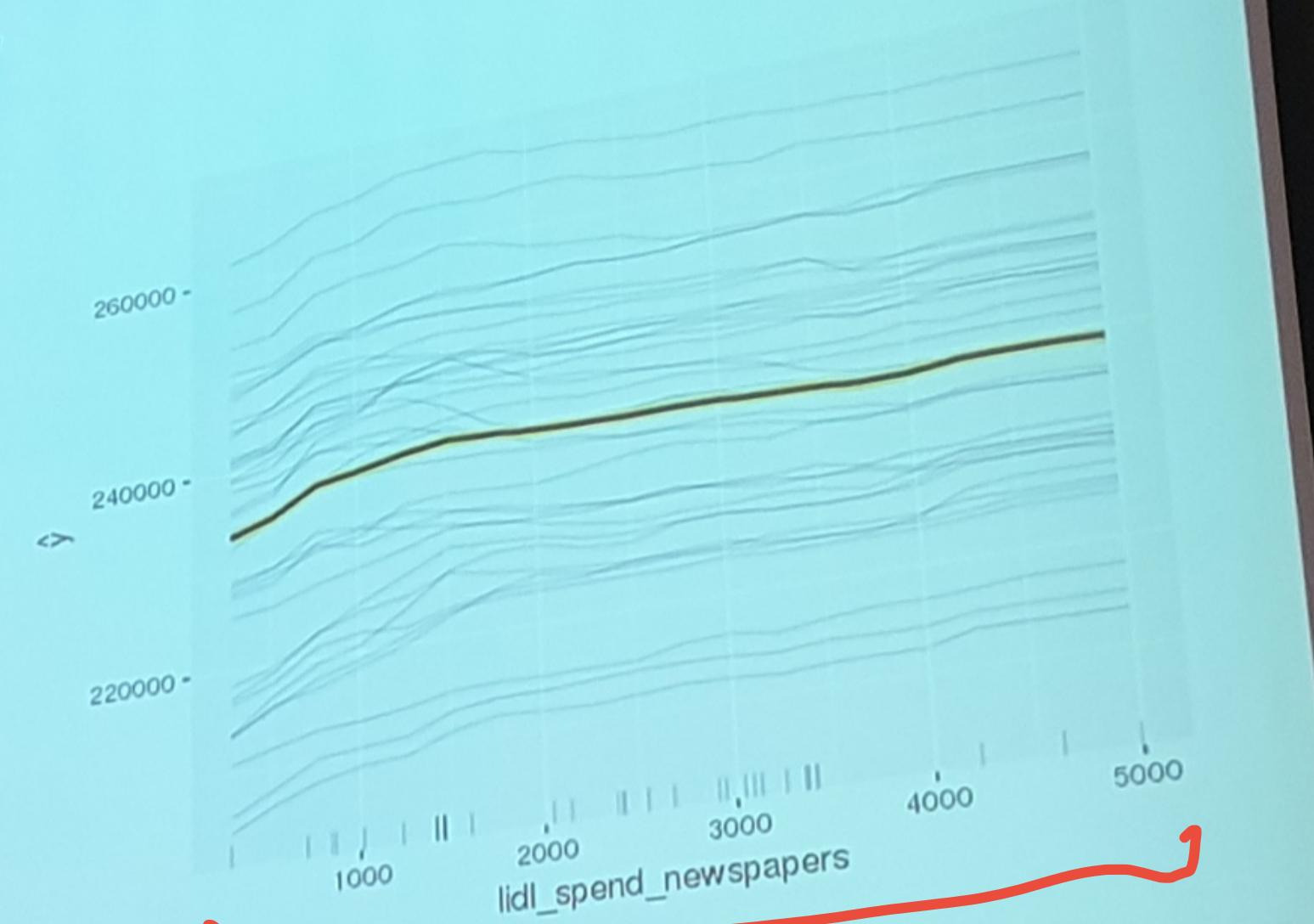
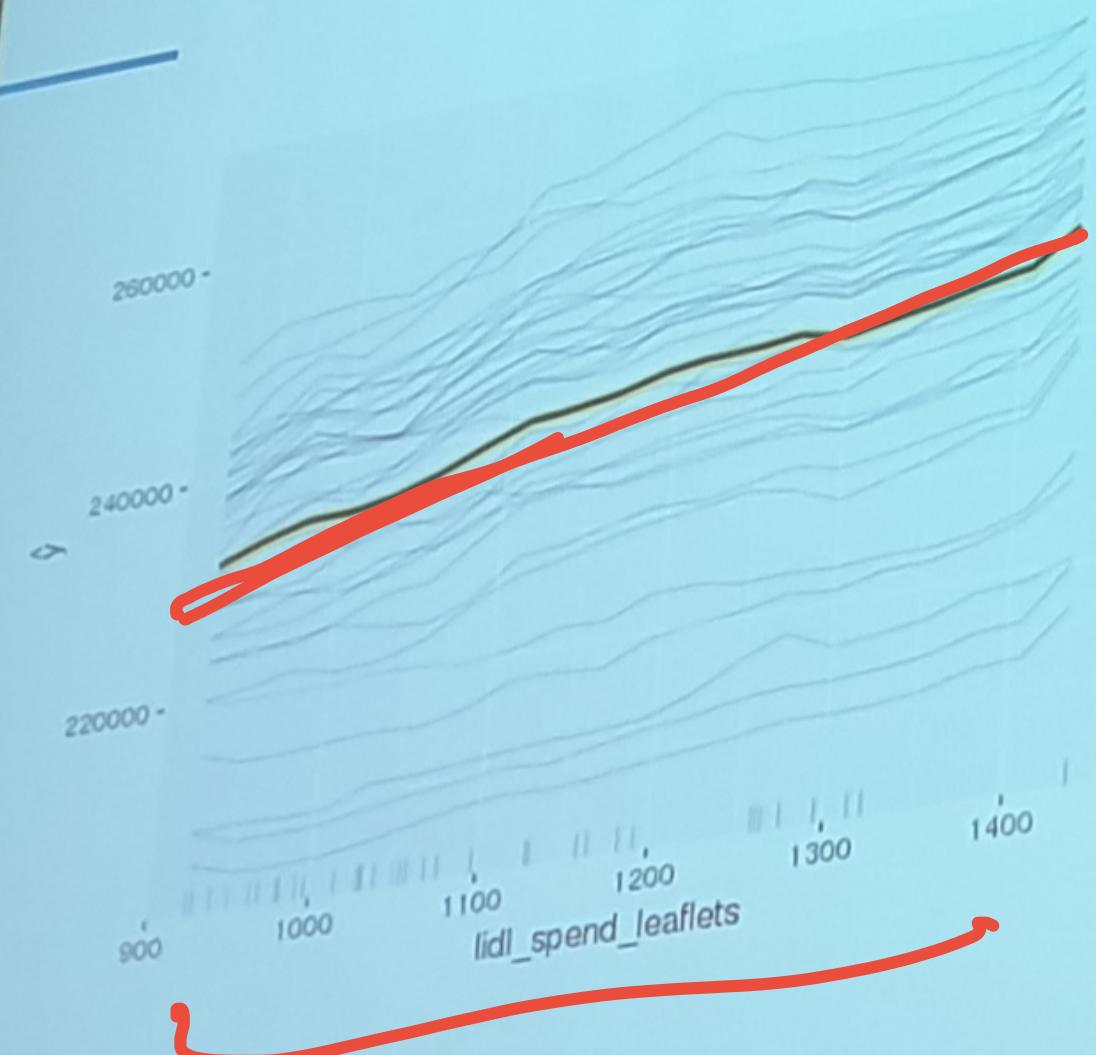
# Three clusters for 100 CP profiles



# Groups of Ceteris paribus profiles by Sex



## Situation 2. Channel contribution.

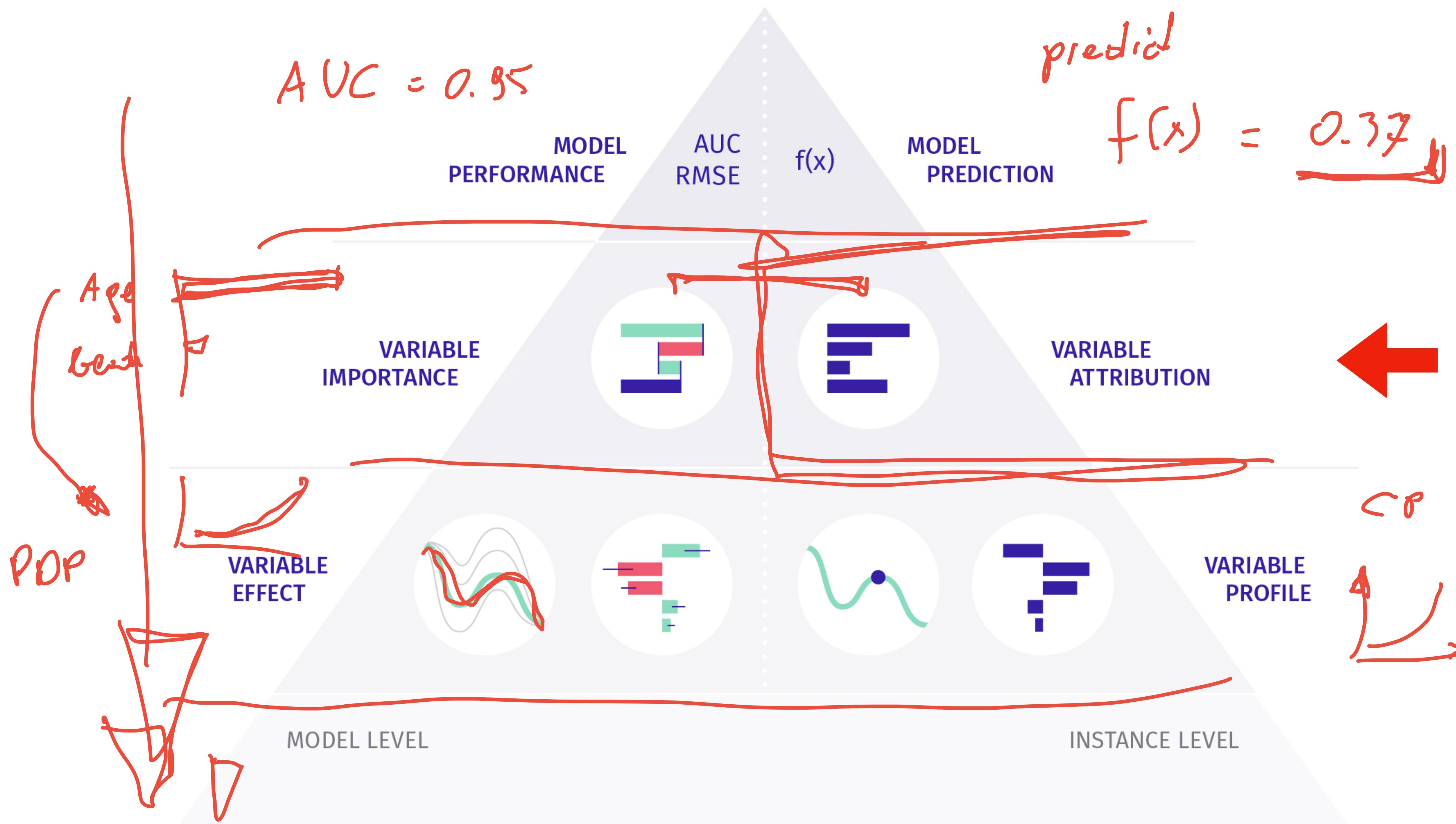


# Day 2

## Model exploration

### Part 3

#### Instance level analysis - variable attributions



SHAP

---

# A Unified Approach to Interpreting Model Predictions

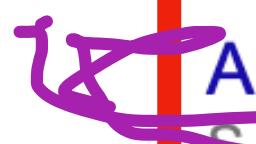
---

**Scott M. Lundberg**  
Paul G. Allen School of Computer Science  
University of Washington  
Seattle, WA 98105  
slund1@cs.washington.edu

**Su-In Lee**  
Paul G. Allen School of Computer Science  
Department of Genome Sciences  
University of Washington  
Seattle, WA 98105  
suinlee@cs.washington.edu

## Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

TYTUŁ	CYTOWANE PRZEZ	ROK
 <a href="#">A unified approach to interpreting model predictions</a> S Lundberg, SI Lee arXiv preprint arXiv:1705.07874	2717	2017
<a href="#">Consistent individualized feature attribution for tree ensembles</a> SM Lundberg, GG Erion, SI Lee arXiv preprint arXiv:1802.03888	350	2018
<a href="#">From local explanations to global understanding with explainable AI for trees</a> SM Lundberg, G Erion, H Chen, A DeGrave, JM Prutkin, B Nair, R Katz, ... Nature machine intelligence 2 (1), 56-67	339	2020
<a href="#">Explainable machine-learning predictions for the prevention of hypoxaemia during surgery</a> SM Lundberg, B Nair, MS Vavilala, M Horibe, MJ Eisses, T Adams, ... Nature biomedical engineering 2 (10), 749-760	240	2018
<a href="#">A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia</a> SI Lee, S Celik, BA Logsdon, SM Lundberg, TJ Martins, VG Oehler, ... Nature communications 9 (1), 1-13	91	2018
<a href="#">Advances in neural information processing systems</a> SM Lundberg, SI Lee, I Guyon New York: Curran Associates	91	2017
<a href="#">An unexpected unity among methods for interpreting model predictions</a> S Lundberg, SI Lee arXiv preprint arXiv:1611.07478	71	2016
<a href="#">Explainable AI for trees: From local explanations to global understanding</a> SM Lundberg, G Erion, H Chen, A DeGrave, JM Prutkin, B Nair, R Katz, ... arXiv preprint arXiv:1905.04610	70	2019

**Definition 1** Additive feature attribution methods have an explanation model that is a linear function of binary variables:

## Properties

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

where  $z' \in \{0, 1\}^M$ ,  $M$  is the number of simplified input features, and  $\phi_i \in \mathbb{R}$ .

### Property 1 (Local accuracy)

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^M \phi_i x'_i \quad (5)$$

The explanation model  $g(x')$  matches the original model  $f(x)$  when  $x = h_x(x')$ , where  $\phi_0 = f(h_x(\mathbf{0}))$  represents the model output with all simplified inputs toggled off (i.e. missing).

### Property 2 (Missingness)

$$x'_i = 0 \implies \phi_i = 0 \quad (6)$$

Missingness constrains features where  $x'_i = 0$  to have no attributed impact.

The third property is *consistency*. Consistency states that if a model changes so that some simplified input's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease.

**Property 3 (Consistency)** Let  $f_x(z') = f(h_x(z'))$  and  $z' \setminus i$  denote setting  $z'_i = 0$ . For any two models  $f$  and  $f'$ , if

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \quad (7)$$

for all inputs  $z' \in \{0, 1\}^M$ , then  $\phi_i(f', x) \geq \phi_i(f, x)$ .

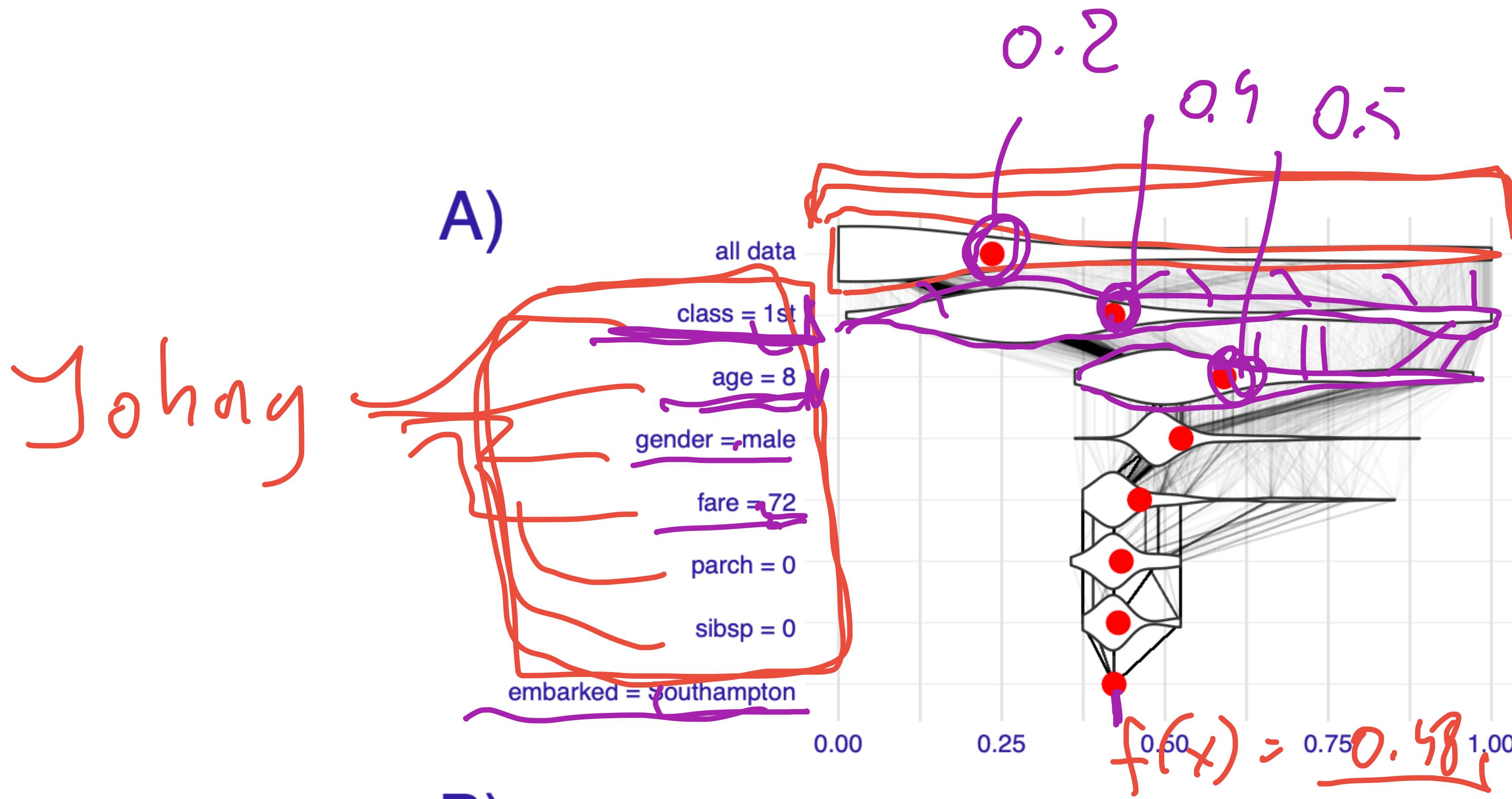
**Theorem 1** Only one possible explanation model  $g$  follows Definition 1 and satisfies Properties 1, 2, and 3:

$$\phi_i(f, x) = \sum_{z' \subset x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \quad (8)$$

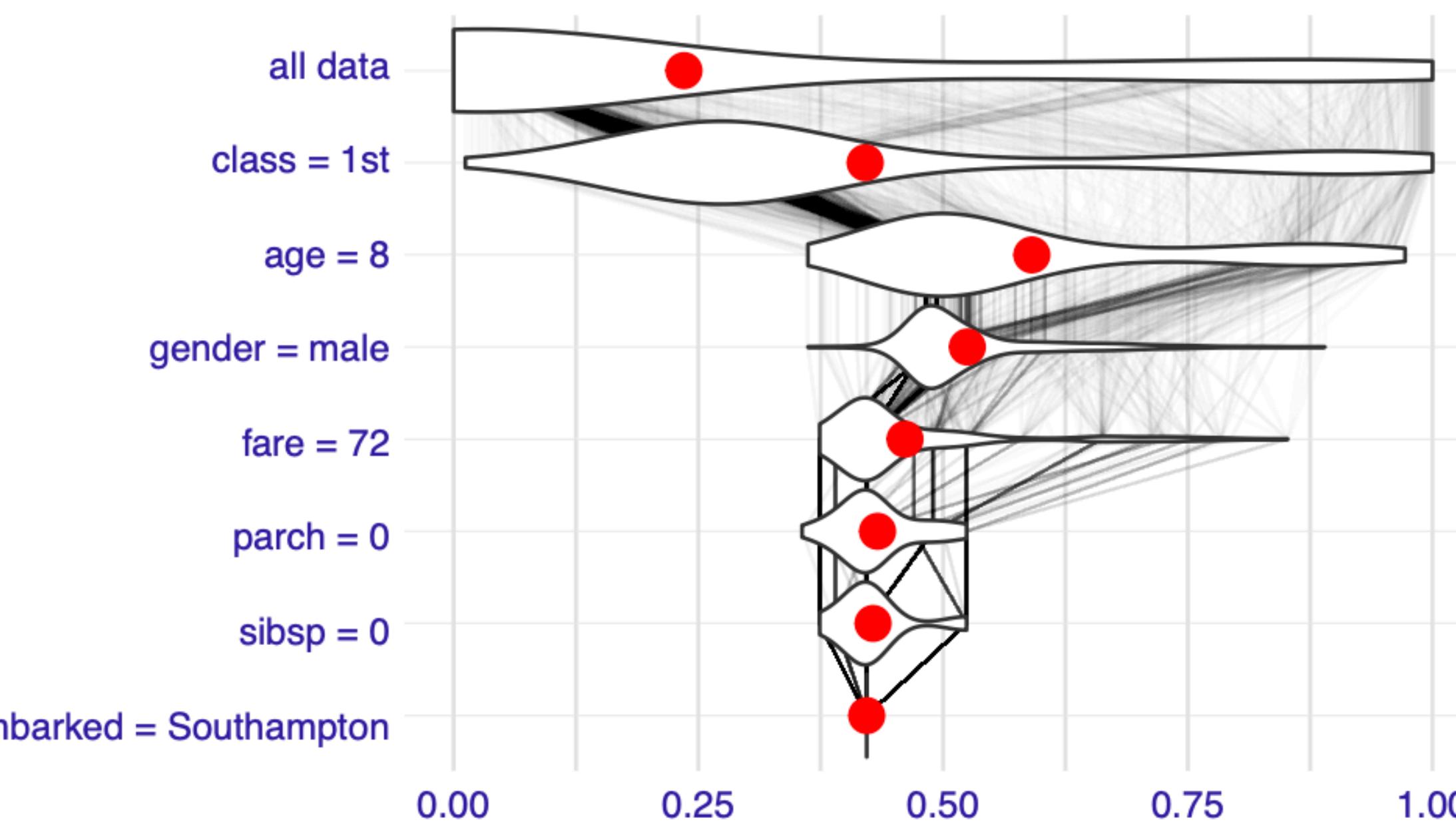
# Additive shifts for non additive models

class = 1st

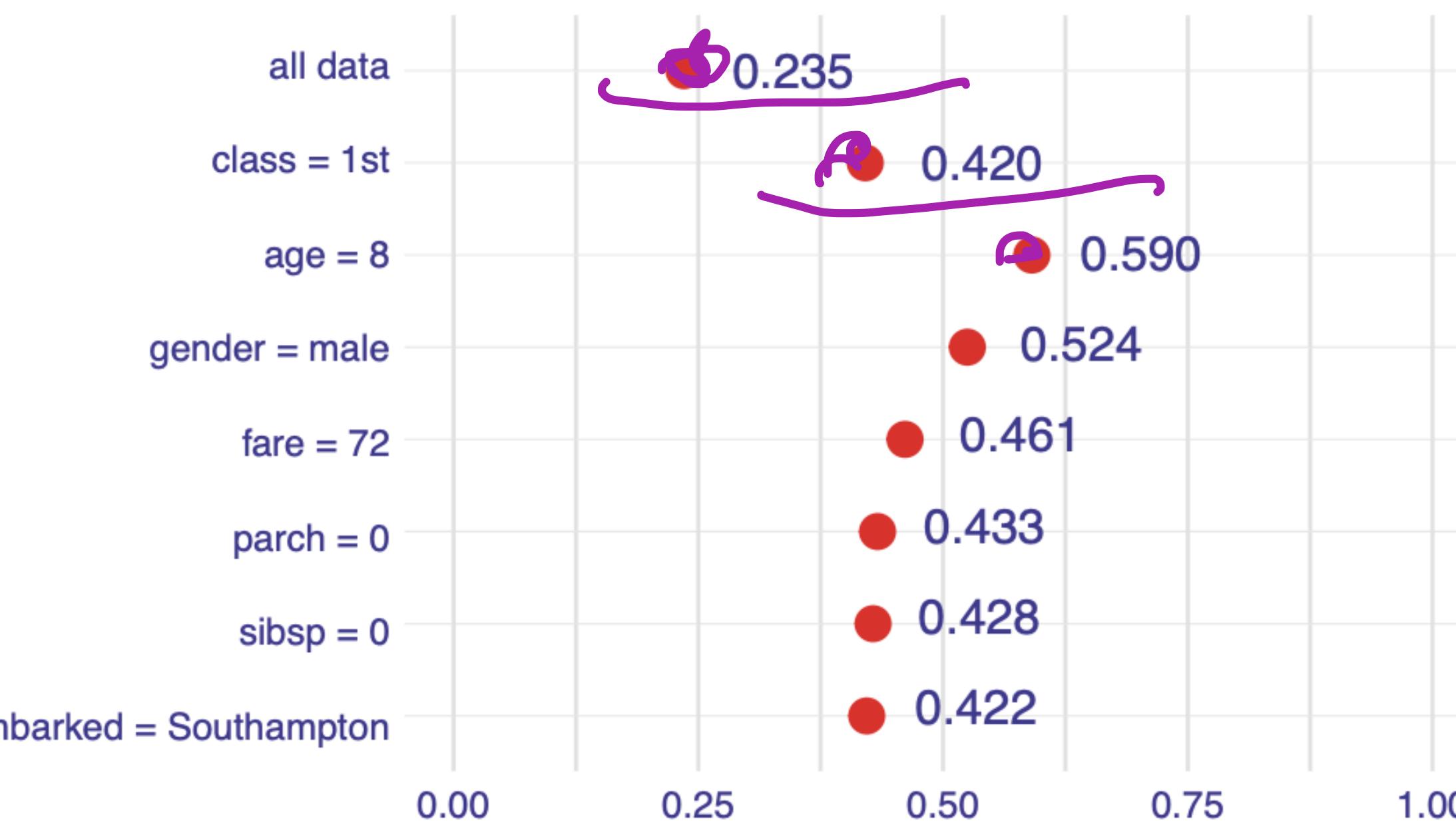
## Model agnostic approach



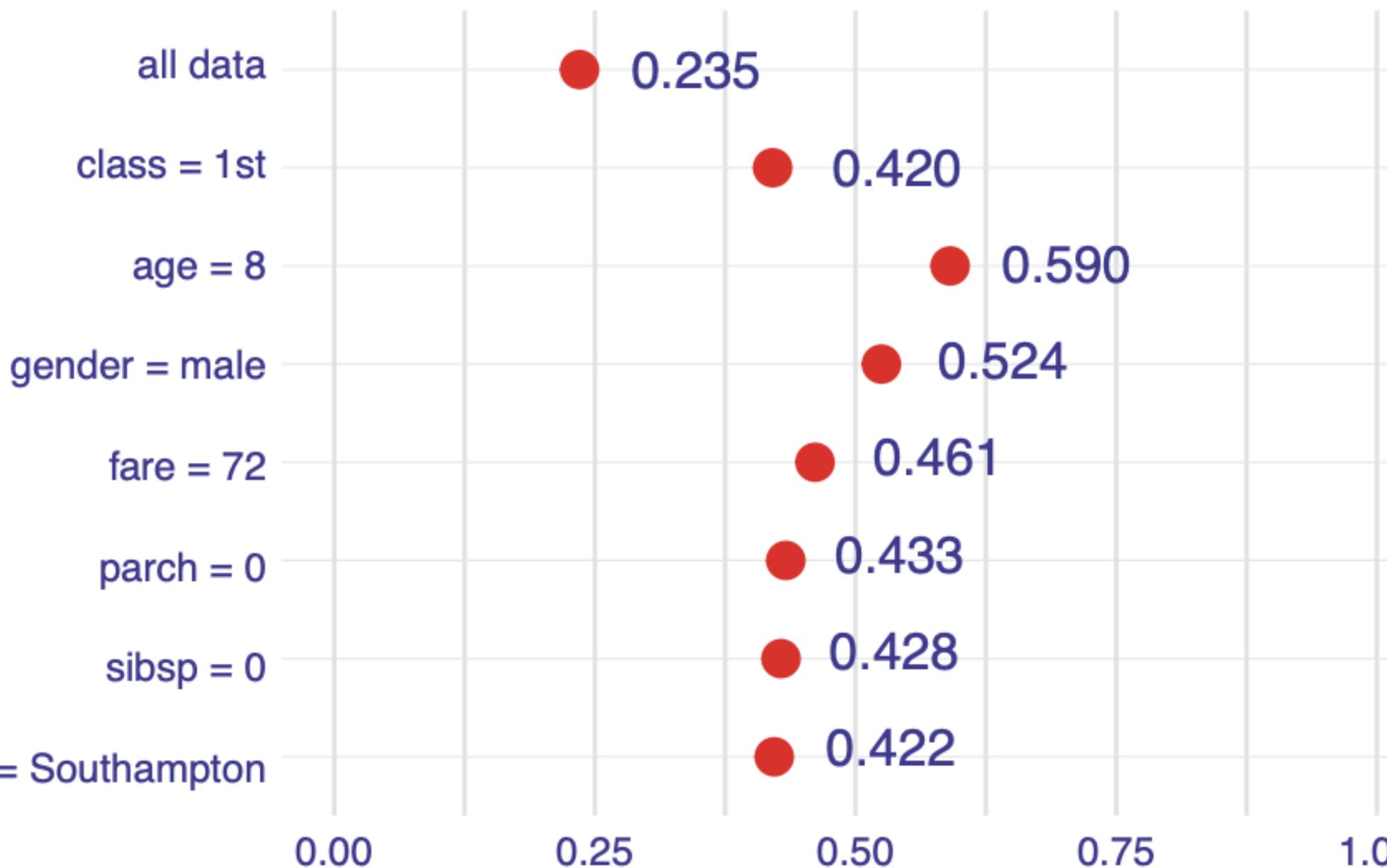
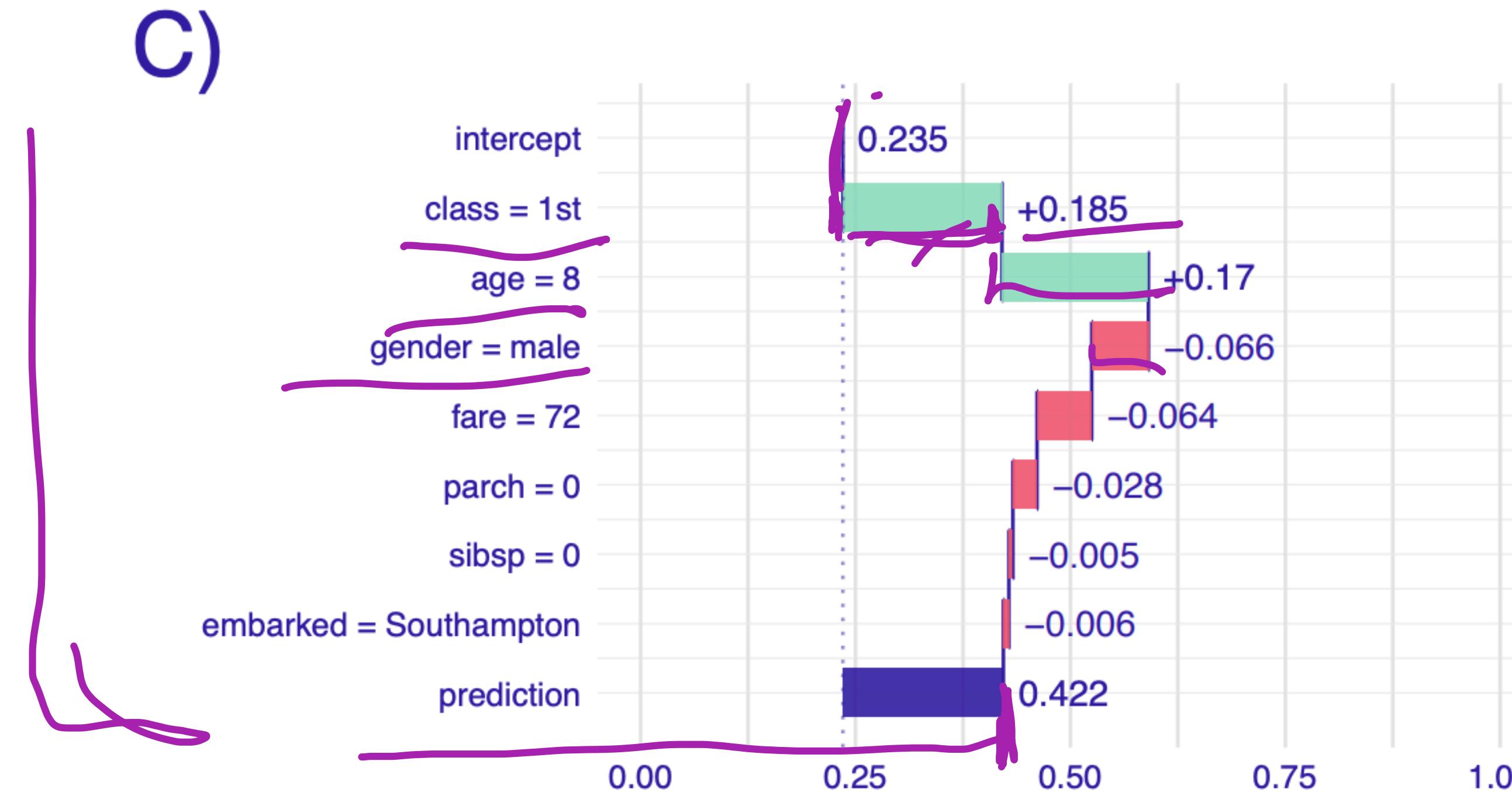
A)



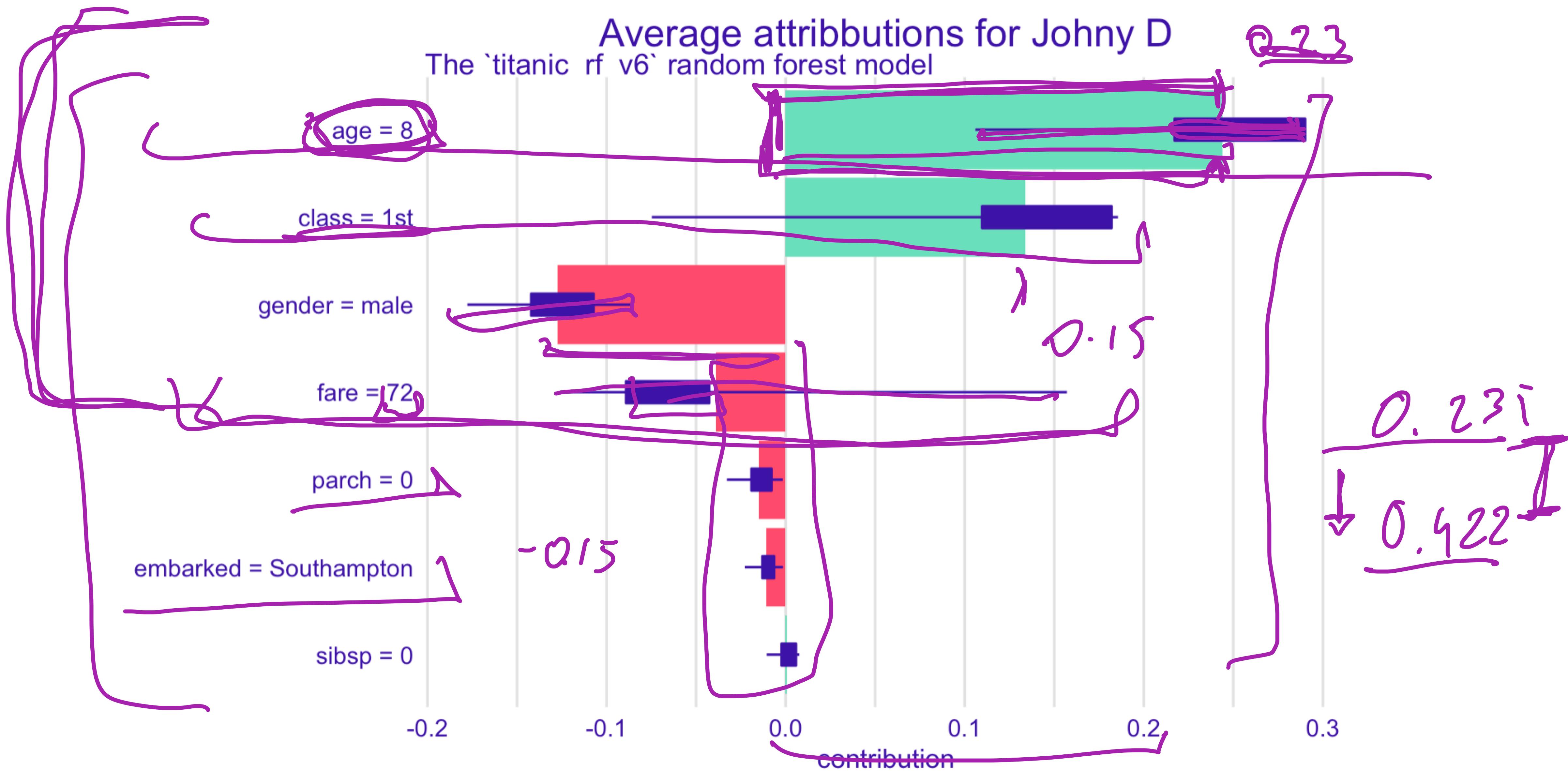
B)



C)

**B)****C)**

# Take an average effect



# Great unification!

## 2.1 LIME

To find  $\phi$ , LIME minimizes the following objective function:

$$\xi = \arg \min L(f, g, \pi_{x'}) + \Omega(g). \quad (2)$$

## 2.2 DeepLIFT

*DeepLIFT* was recently proposed as a recursive prediction explanation method for deep learning [8, 7]. It attributes to each input  $x_i$  a value  $C_{\Delta x_i \Delta o}$  that represents the effect of that input being set to a reference value as opposed to its original value. This means that for DeepLIFT, the mapping  $x = h_x(x')$  converts binary values into the original inputs, where 1 indicates that an input takes its original value, and 0 indicates that it takes the reference value. The reference value, though chosen by the user, represents a typical uninformative background value for the feature.

DeepLIFT uses a "summation-to-delta" property that states:

$$\sum_{i=1}^n C_{\Delta x_i \Delta o} = \Delta o, \quad (3)$$

## 2.3 Layer-Wise Relevance Propagation

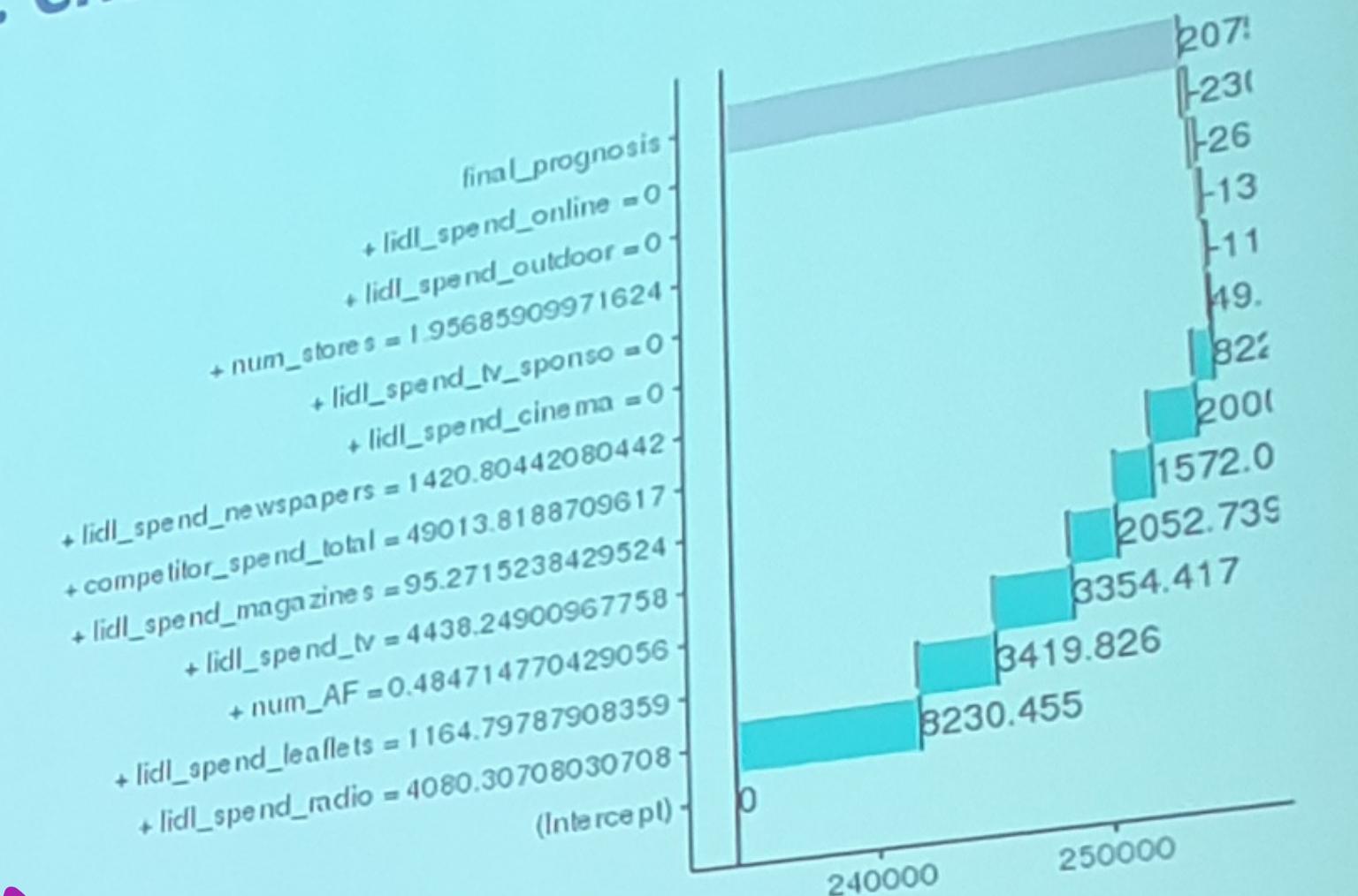
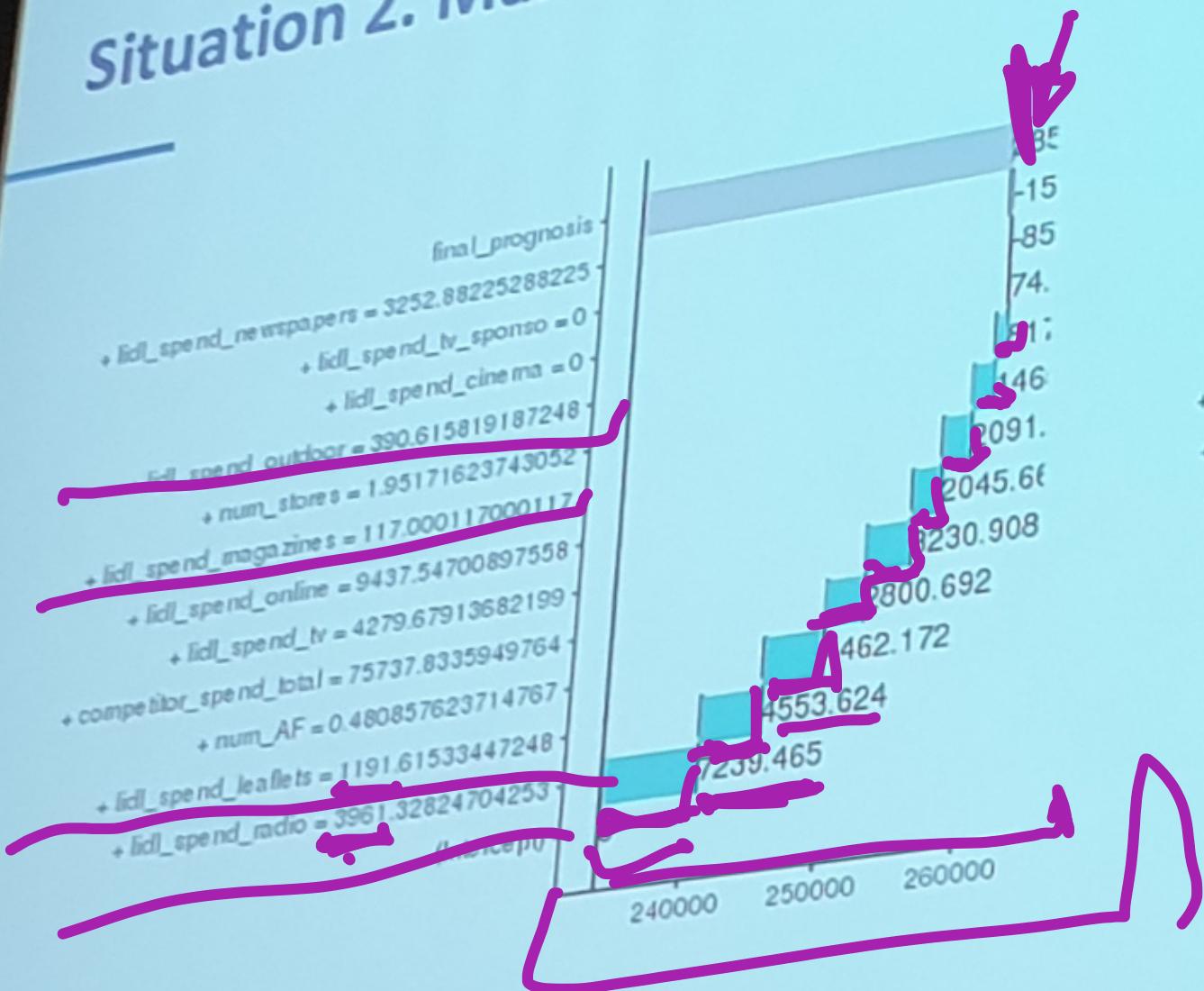
The *layer-wise relevance propagation* method interprets the predictions of deep networks [1]. As noted by Shrikumar et al., this method is equivalent to DeepLIFT with the reference activations of all neurons fixed to zero. Thus,  $x = h_x(x')$  converts binary values into the original input space, where 1 means that an input takes its original value, and 0 means an input takes the 0 value. Layer-wise relevance propagation's explanation model, like DeepLIFT's, matches Equation 1.

## 2.4 Classic Shapley Value Estimation

Three previous methods use classic equations from cooperative game theory to compute explanations of model predictions: Shapley regression values [4], Shapley sampling values [9], and Quantitative Input Influence [3].

*Shapley regression values* are feature importances for linear models in the presence of multicollinearity. This method requires retraining the model on all feature subsets  $S \subseteq F$ , where  $F$  is the set of all features. It assigns an importance value to each feature that represents the effect on the model prediction of including that feature. To compute this effect, a model  $f_{S \cup \{i\}}$  is trained with that feature

## Situation 2. Marketing Mix Model. Channel contribution.

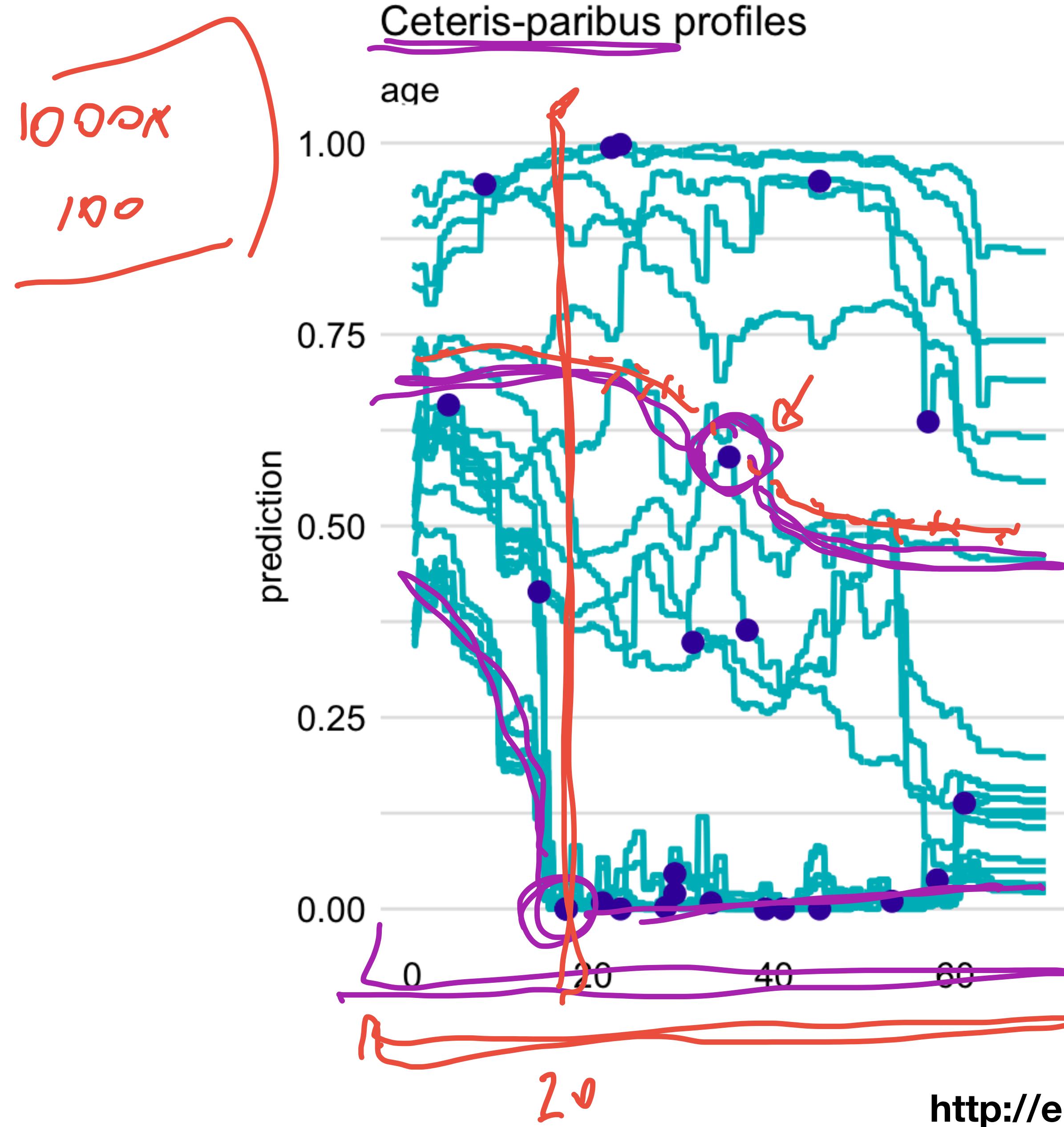


Day 2

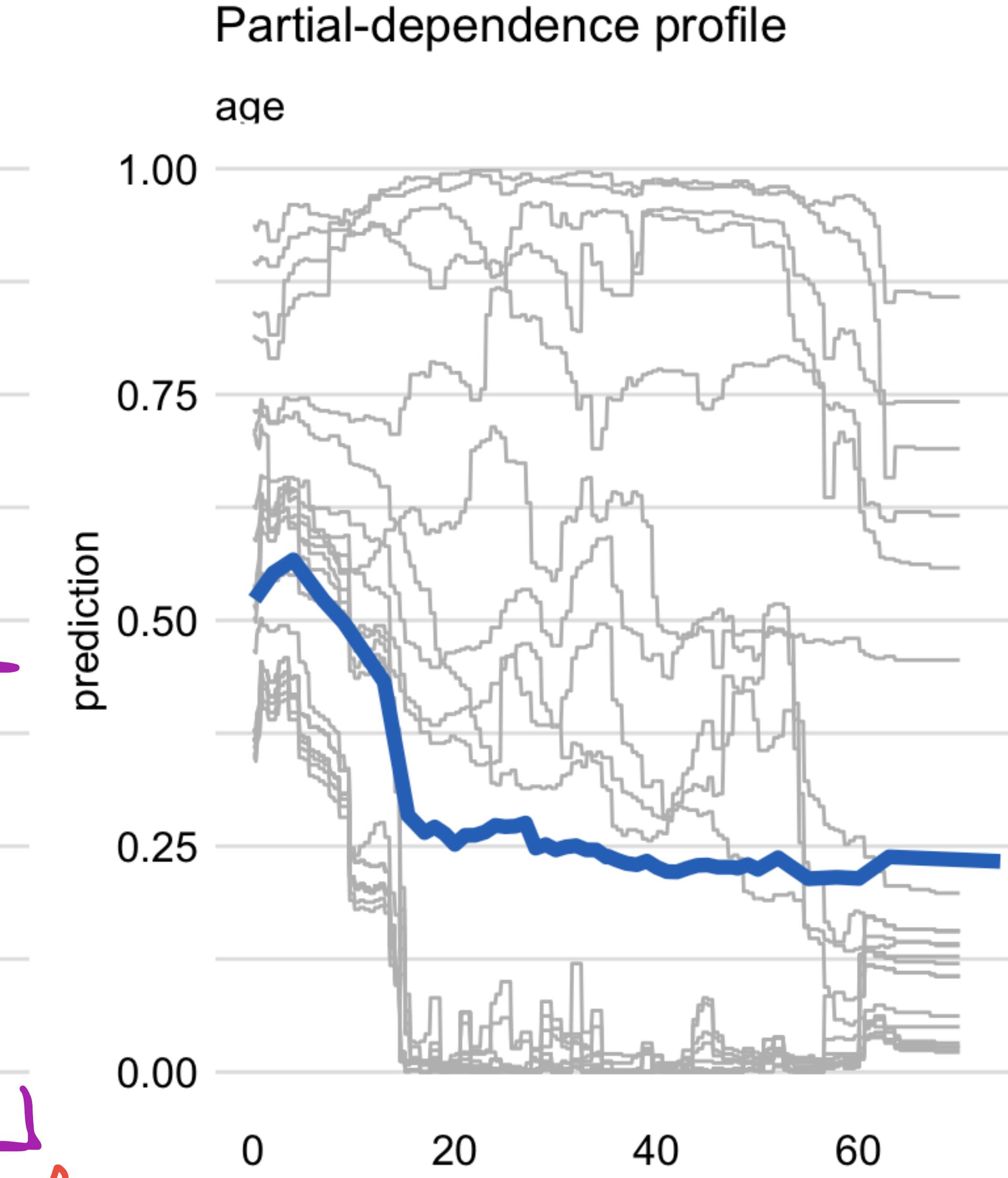
Model exploration

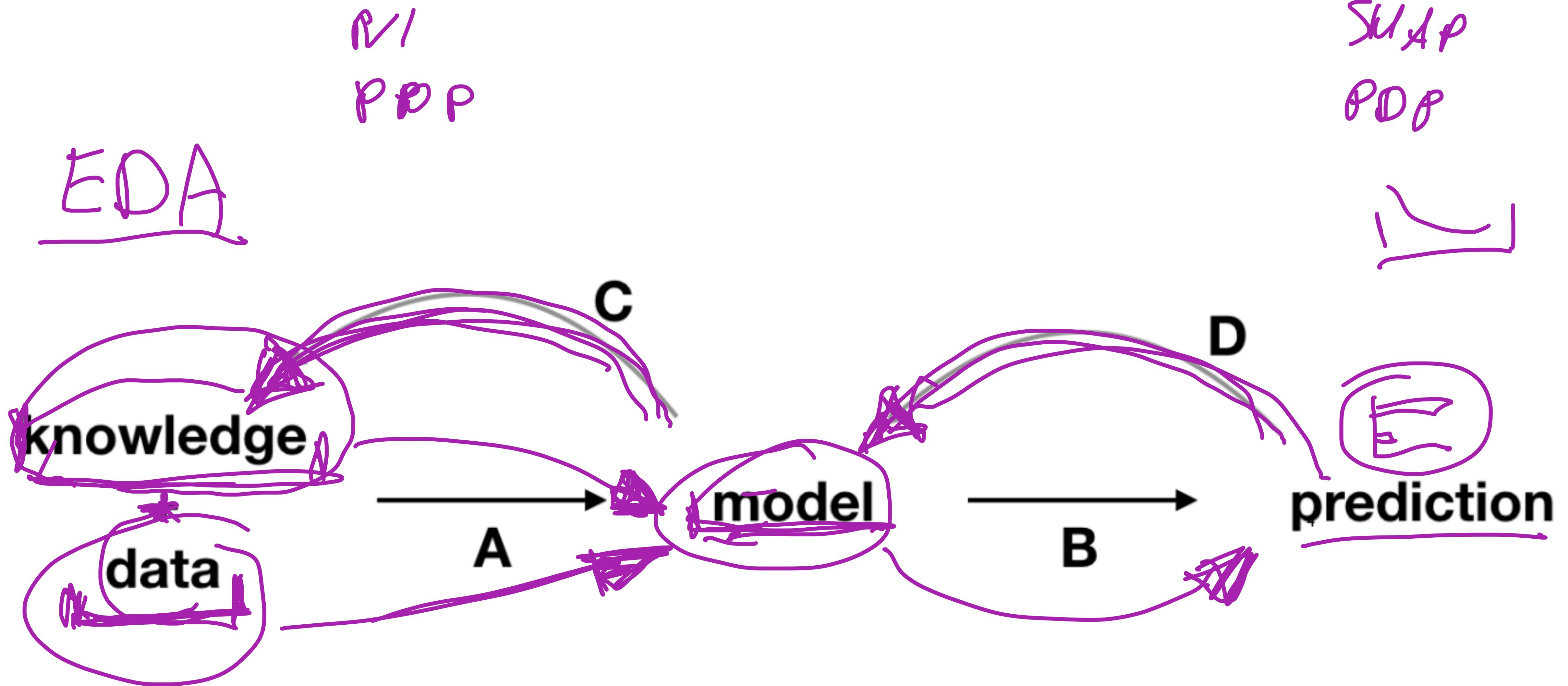
Part 4

Instance level analysis - variable profile

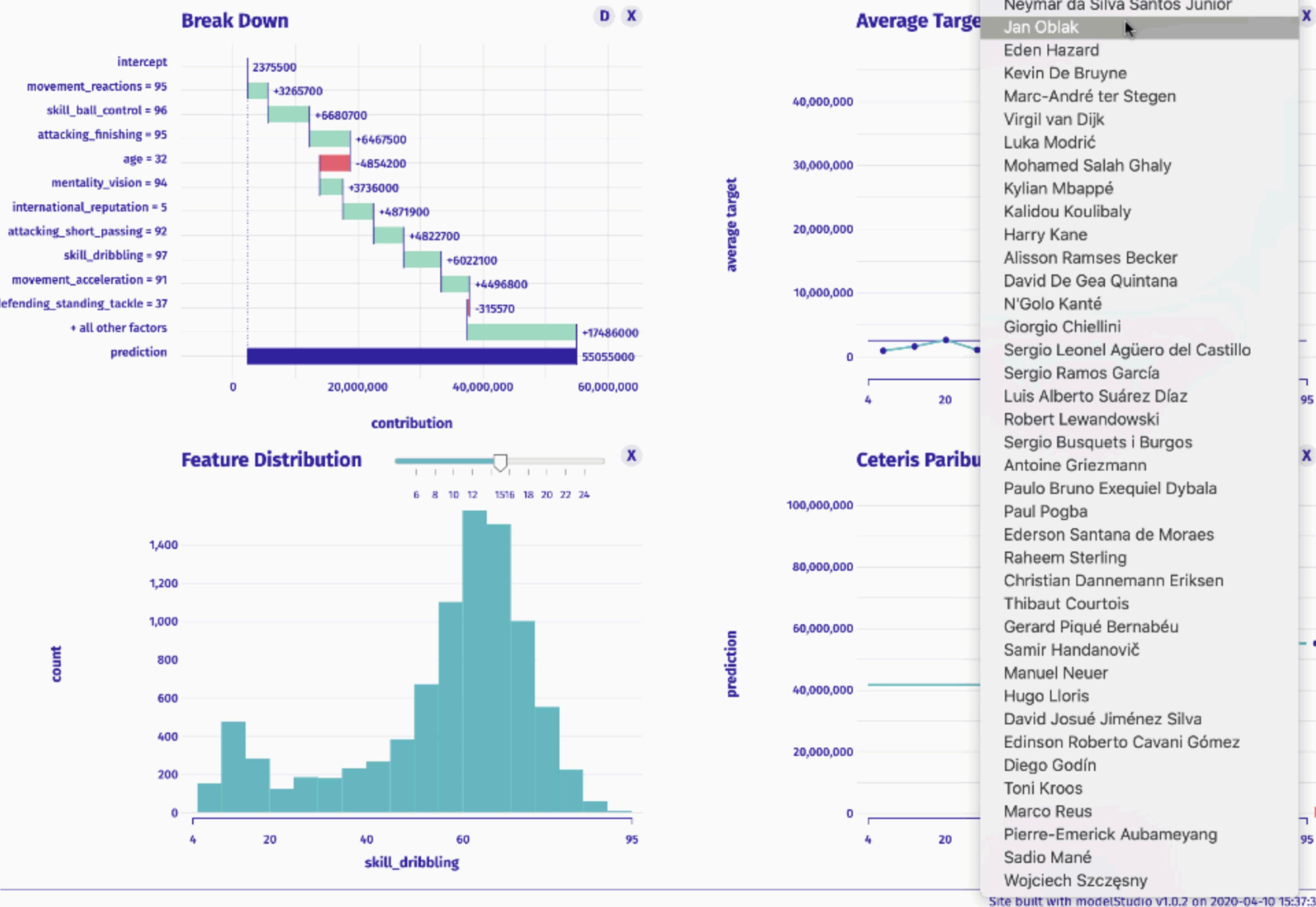


<http://ema.drwhy.ai/partialDependenceProfiles.html>





# Interactive Model Studio for FIFA 20 (GBM model)



- ✓ Lionel Andrés Messi Cuccittini
- Cristiano Ronaldo dos Santos Aveiro
- Neymar da Silva Santos Junior
- Jan Oblak**
- Eden Hazard
- Kevin De Bruyne
- Marc-André ter Stegen
- Virgil van Dijk
- Luka Modrić
- Mohamed Salah Ghaly
- Kylian Mbappé
- Kalidou Koulibaly
- Harry Kane
- Alisson Ramses Becker
- David De Gea Quintana
- N'Golo Kanté
- Giorgio Chiellini
- Sergio Leonel Agüero del Castillo
- Sergio Ramos García
- Luis Alberto Suárez Díaz
- Robert Lewandowski
- Sergio Busquets i Burgos
- Antoine Griezmann
- Paulo Bruno Exequiel Dybala
- Paul Pogba
- Ederson Santana de Moraes
- Raheem Sterling
- Christian Dannemann Eriksen
- Thibaut Courtois
- Gerard Piqué Bernabéu
- Samir Handanović
- Manuel Neuer
- Hugo Lloris
- David Josué Jiménez Silva
- Edinson Roberto Cavani Gómez
- Diego Godín
- Toni Kroos
- Marco Reus
- Pierre-Emerick Aubameyang
- Sadio Mané
- Wojciech Szczęsny

Site built with [modelStudio v1.0.2](#) on 2020-04-10 15:37:35

# Introduction to Responsible Machine Learning

with mlr3 and DALEX

Przemyslaw Biecek  
17-18.04.2021

<https://www.linkedin.com/in/pbiecek/>

