Machine Learning:
Insights & Examples from Clinical Research

19 October 2019
Reno, NV

Petra LeBeau, Sc.D
Sr. Biostatistician III
Lead Bioinformatics Analytics Division
Rho, Inc

Outline

- Overview of Rho, Inc.
- Machine Learning Introduction
- Machine Learning Principles
- Machine Learning Pipeline
- Machine Learning Interpretation

Rho, Inc – Research Triangle Park, NC

Overview of Rho, Inc.

- Privately-held CRO
- Research Triangle Park, NC
- Founded 35 years ago
- More than 400 employees
  - 90% on-site in the RTP office
  - >50% have been at Rho for >5 years
- Federal and Commercial activities
- Strong support for Statistics and Data Management
- Provides support for all clinical research services

www.rhoworld.com

Full Service Capabilities

PIND → IND → Phase I → Phase II → Phase III → NDA → Phase IV

- INDs
- Observational
- Registries
- Mechanistic studies
- Genetics studies
- Translational research

- Concept
- Protocol Development
- Site Start Up
- Study Implementation
- Close Out
- Manuscripts & Publications
- Data Sharing

Medical and Public Health Research at Rho

Institute	Program Name	Primary Focus
NIAID / NIH	ADCT – Autoimmune Diseases Clinical Trials	Autoimmune and Stem Cell Transplantation
NIAID / NIH	ADRN – Atopic Dermatitis Research Network	Atopic Dermatitis
NIAID / NIH	CTOT – Clinical Trials of Organ Transplantation	Transplant
NIAID / NIH	ICAC – Inner City Asthma Consortium	Asthma
NIAID / NIH	ITN – Immune Tolerance Network	Autoimmune, Allergy/Asthma, and Transplant
NIDCR / NIH	CROMS – Clinical Research Operations and Management Support	Dental, Craniofacial, Pain, Oncology
NINDS / NIH	Regulatory Support Contract	Neurological Disease & Stroke

MACHINE LEARNING INTRODUCTION

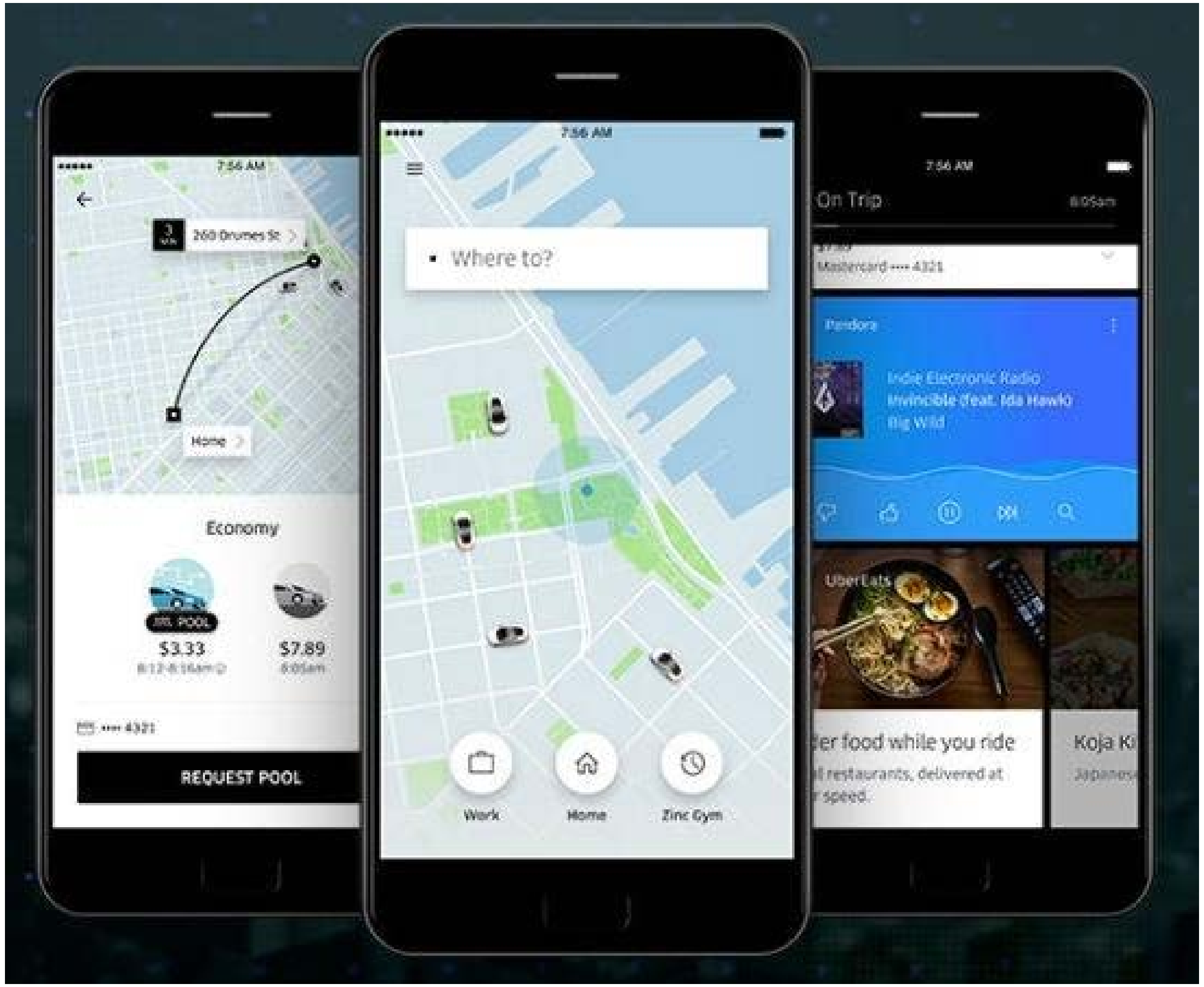
Machine learning (ML)

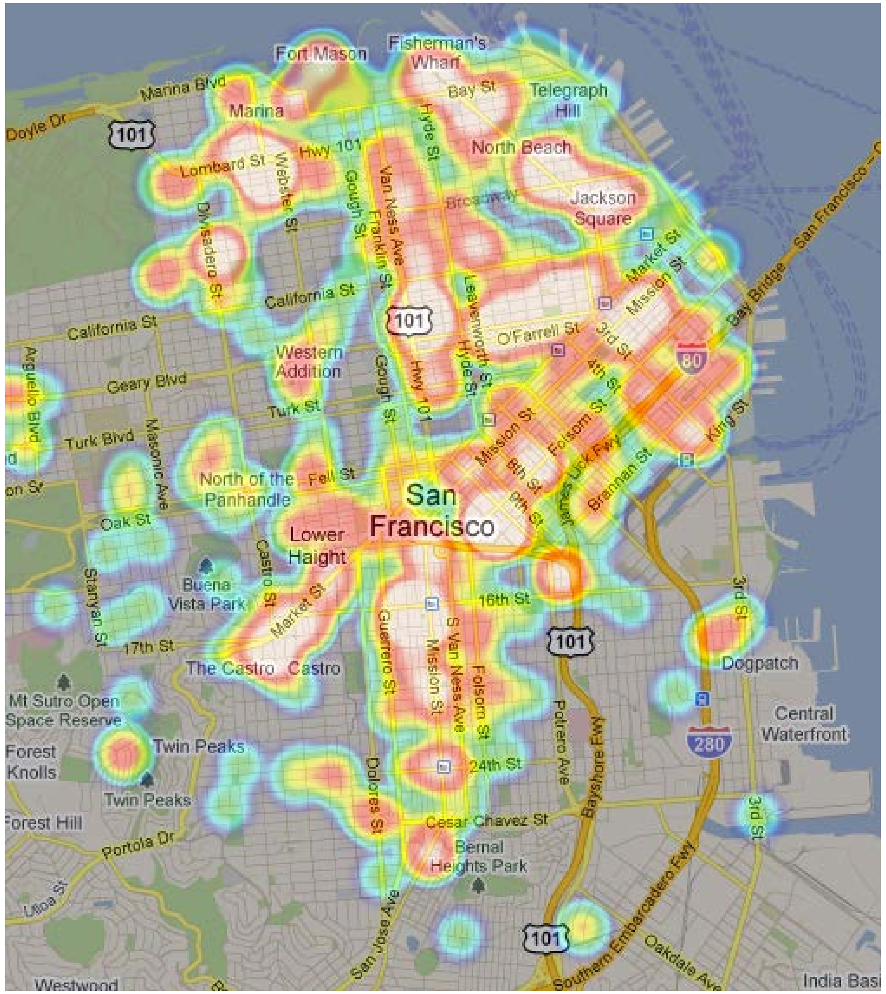
- Subset of Artificial Intelligence (AI)

"Field of study that gives computers the ability to learn without being explicitly programmed"

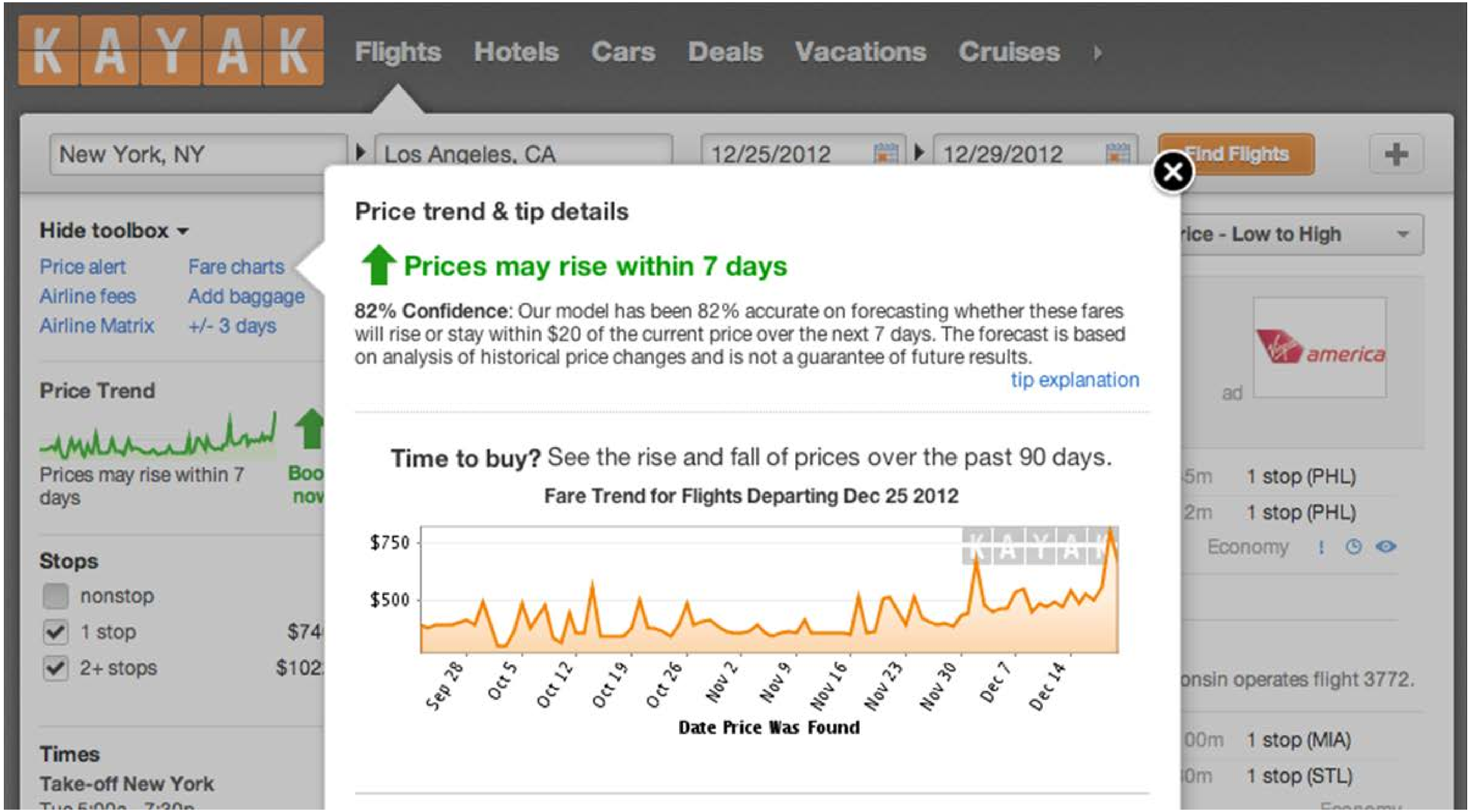
- Arthur Samuel, 1959

Ride Sharing – e.g. Uber, Lyft, Grab





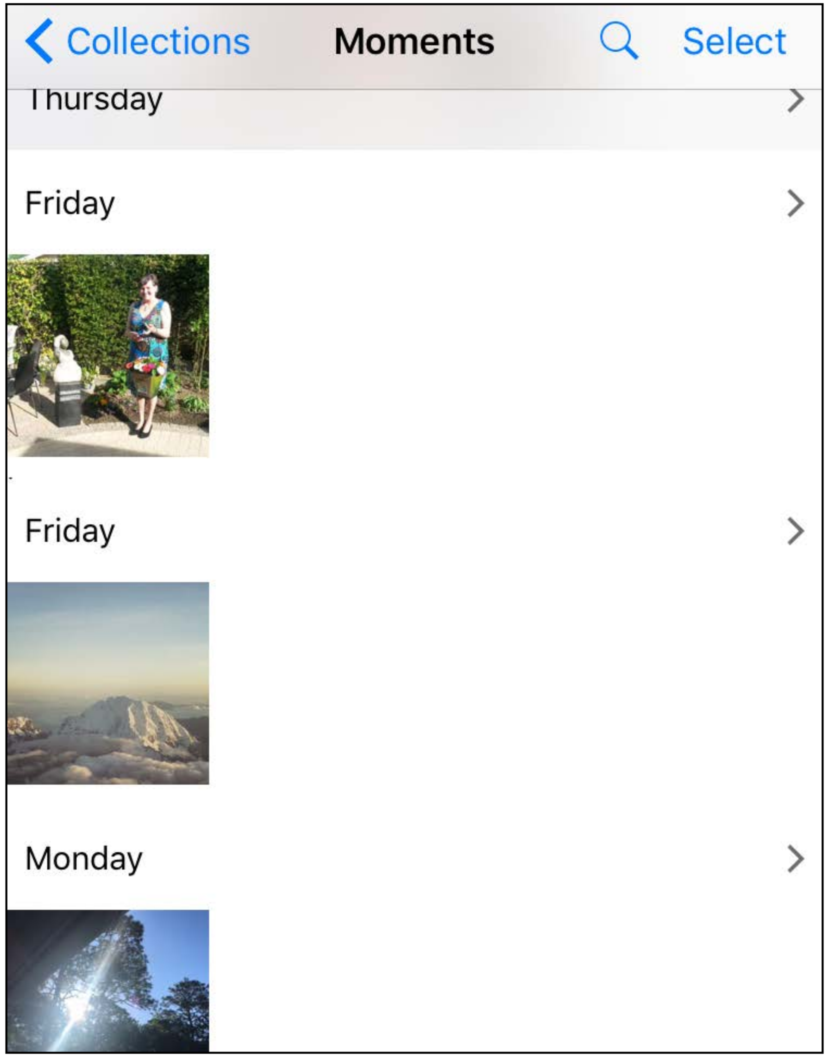
Help me Buy a (Cheaper) Ticket

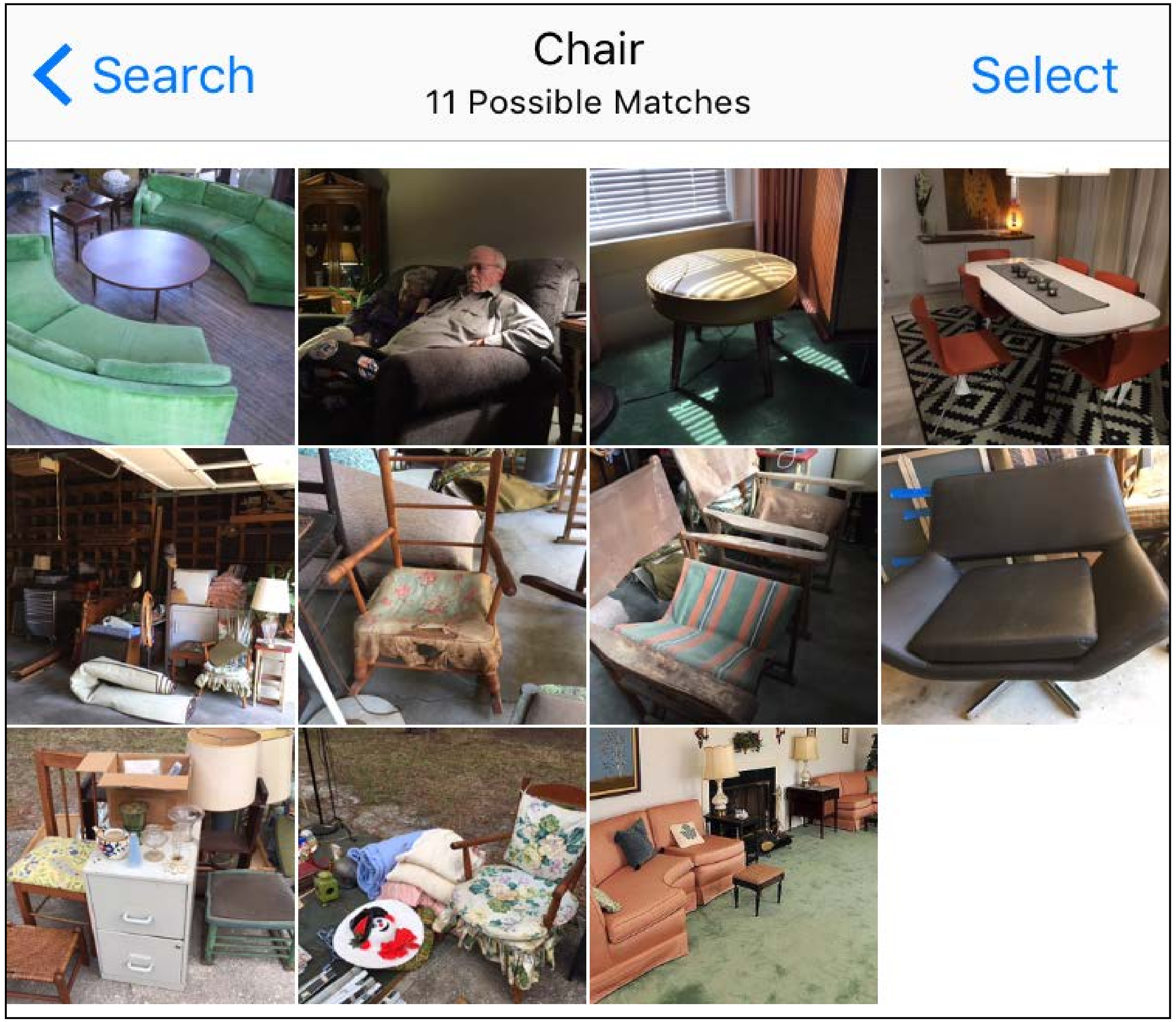


Handwriting Recognition

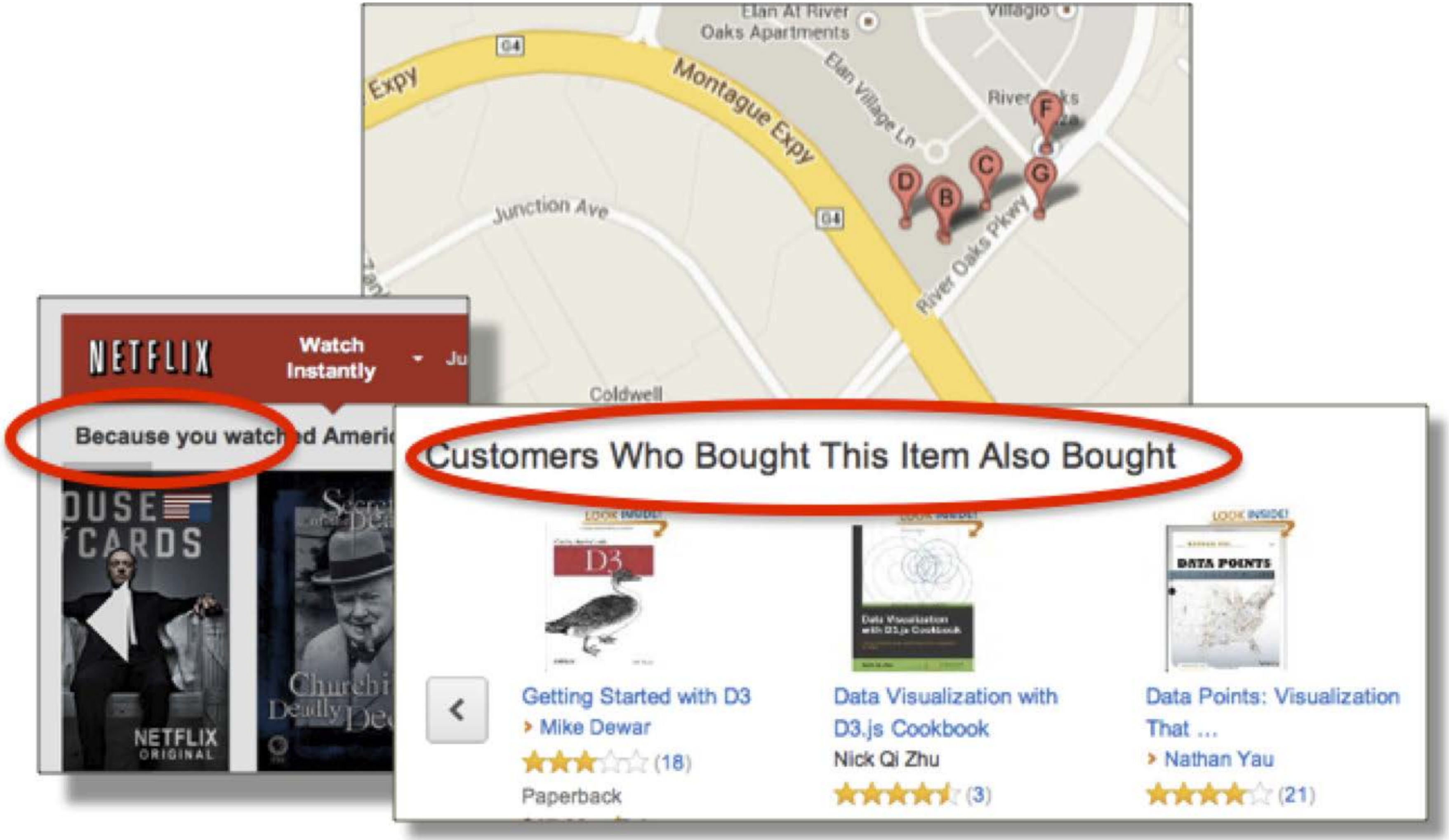
Error Rate of 1%
(99% accuracy)

Phone Photo Search





Recommendations



An Algorithmic Sense of Humor?

Context:
dealership, salesman, car, windows

Anomaly:
animal legs, mouth, teeth, furry legs

An Algorithmic Sense of Humor?

What's it going to take to get you in this car today?
Relax! It just smells the other car on you.
It runs entirely on legs.
Just don't tailgate during mating season.
It's only been driven once.
He even cleans up his road kill.
The spare leg is in the trunk.
Comfortably eats six.
She runs like a dream I once had.

What is ML?

Highly efficient algorithms
designed to “learn”
how to complete a specific task,
using past performance
to predict and improve future performance

Why do we need ML in Clinical Research?

Traditional exploratory research

- Application of statistical methods to test **causal explanations** using **a priori** theoretical constructs

Predictive research

- Application of statistical methods and/or data mining techniques, **without a priori** theoretical constructs to **predict** future outcomes
- Causality is neither a primary aim nor a requirement for variable inclusion

Common Data Sets from Clinical Trials

- Small n x p data sets (e.g. qPCR with <20 features)
- Regression techniques (e.g. logistic regression)
- Backward-elimination

STEP 1: Determine significance level to stay in model (e.g. SL = 0.05)

STEP 2: Fit full model with all possible features

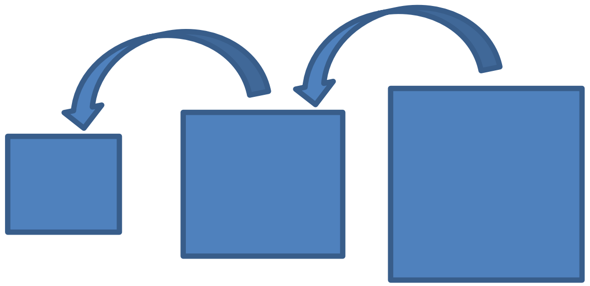
STEP 3: Consider the feature with the highest p-value.

p > SL or p <= SL

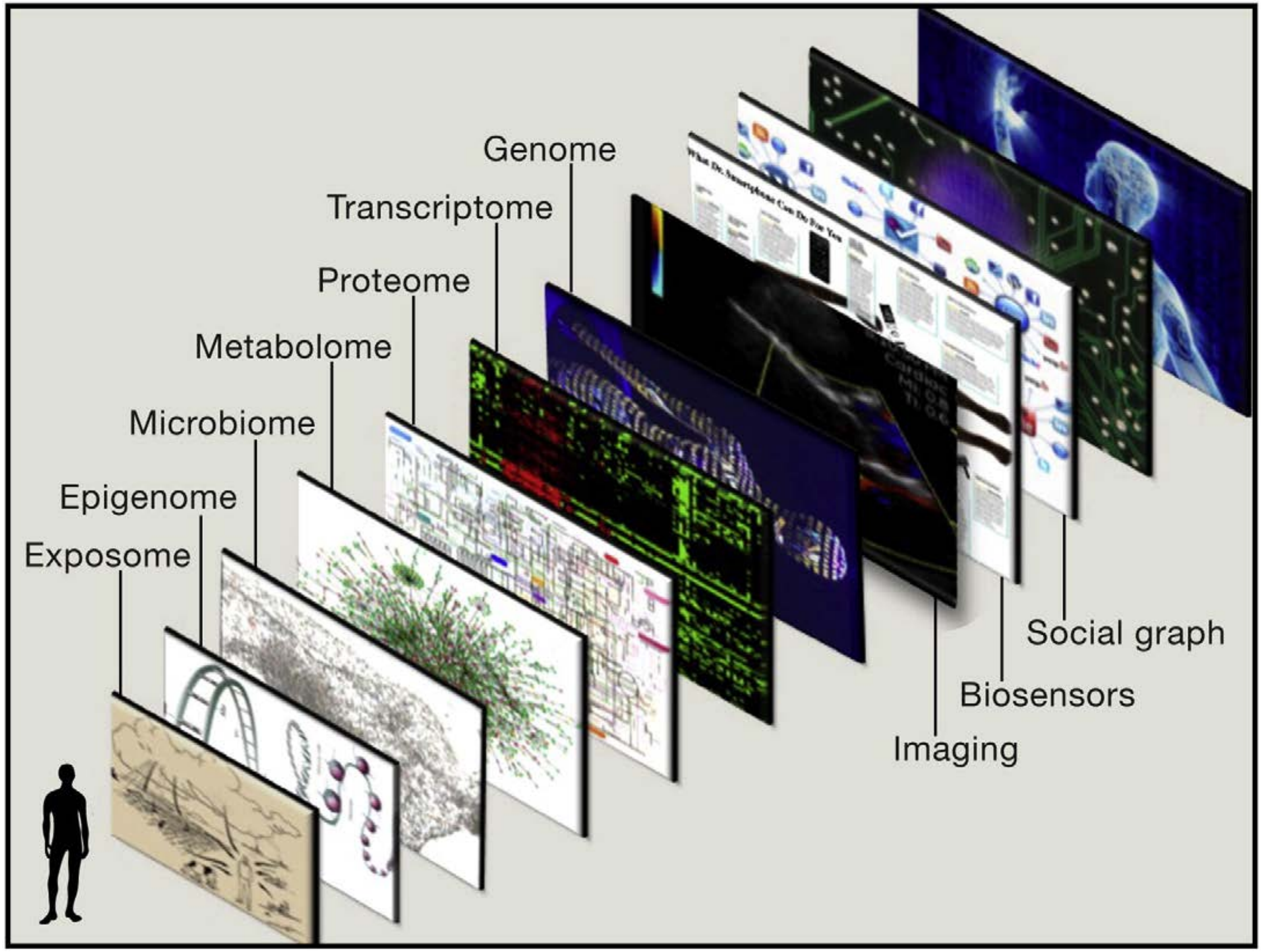
STEP 4: Remove feature

STEP 5: Fit model without this feature

Final Model



Why do we need ML?



ML is Changing Medicine

- Identification / Diagnosis
- Personalized Treatment
- Drug Discovery / Manufacturing
- Clinical Trial Research
- Radiology and Radiotherapy
- Smart Electronic Health Records
- Epidemic Outbreak Prediction

ML is Changing Medicine

Jean-Paul Pelissier / Reuters

How Artificial Intelligence Can Help Burn Victims

Machine learning allows computers to see patterns in medical images that are invisible to human doctors.

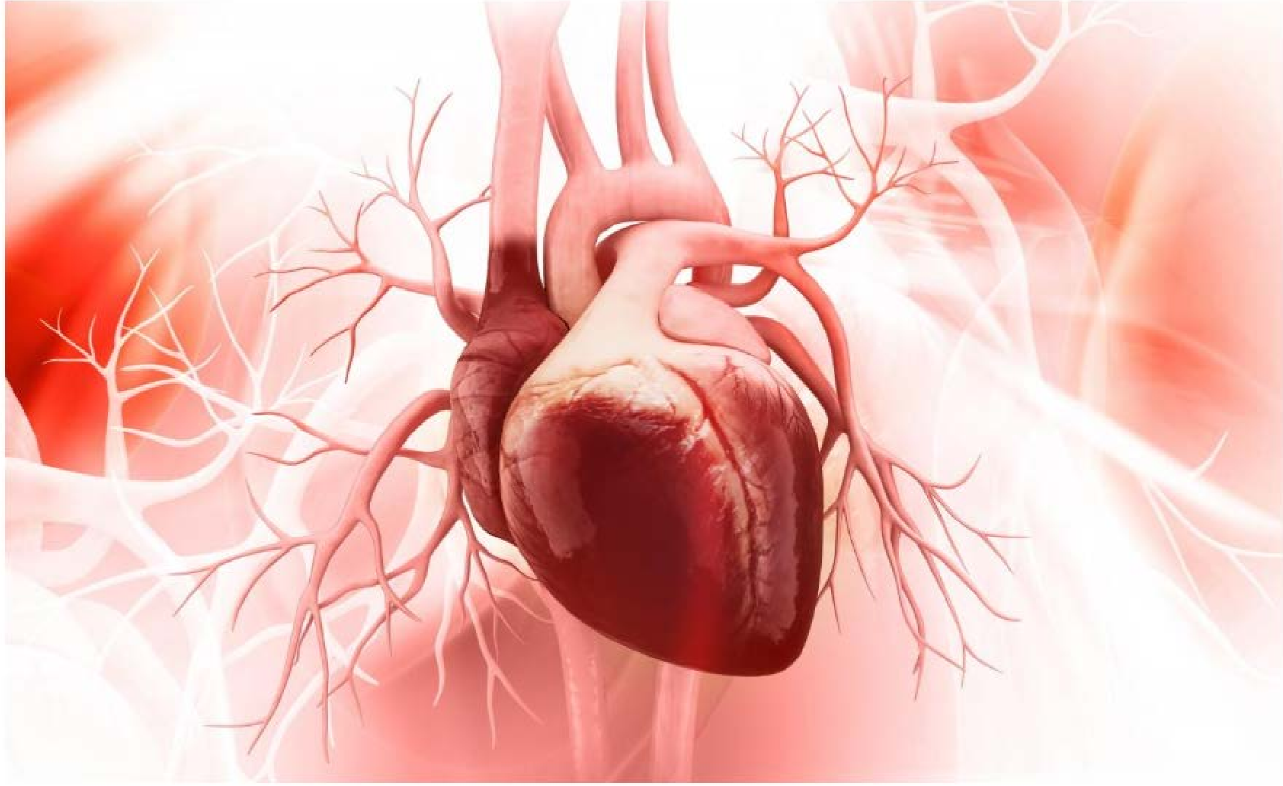
TEXT SIZE

ADRIENNE LAFRANCE | 9:52 AM ET | TECHNOLOGY

https://www.theatlantic.com/technology/archive/2016/08/how-machine-learning-could-help-burn-victims-recover-faster/495926/

Zebra Medical debuts two machine learning algorithms to predict heart disease risk

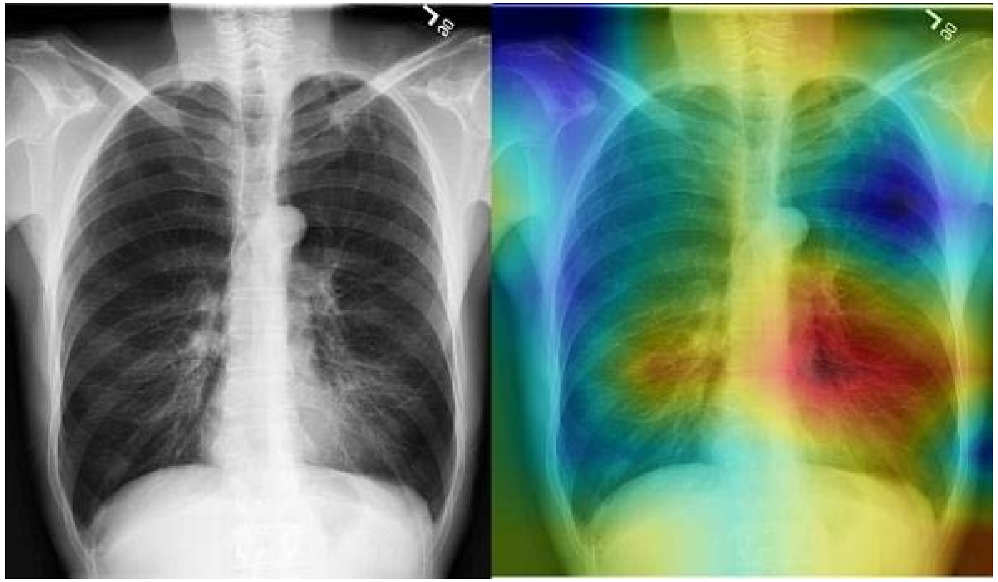
by *Amirah Al Idrus* | Aug 15, 2016 5:00am



https://www.fiercebiotech.com/medical-devices/zebra-medical-debuts-2-machine-learning-algorithms-to-predict-heart-disease-risk

ML is Changing Medicine

https://arxiv.org/abs/1606.05718



Deep Learning Algorithm beats radiologists in diagnosing x-rays

https://arxiv.org/abs/1711.05225

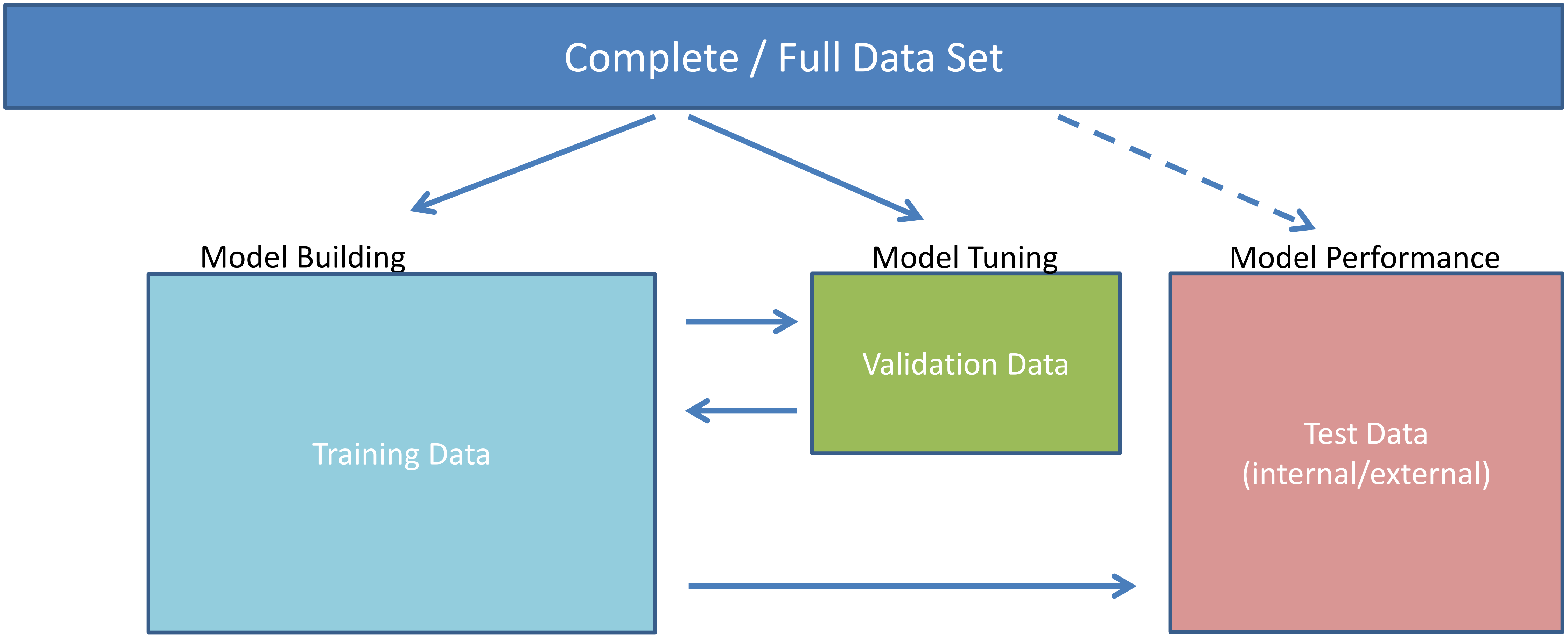
MACHINE LEARNING PRINCIPLES

ML Algorithms

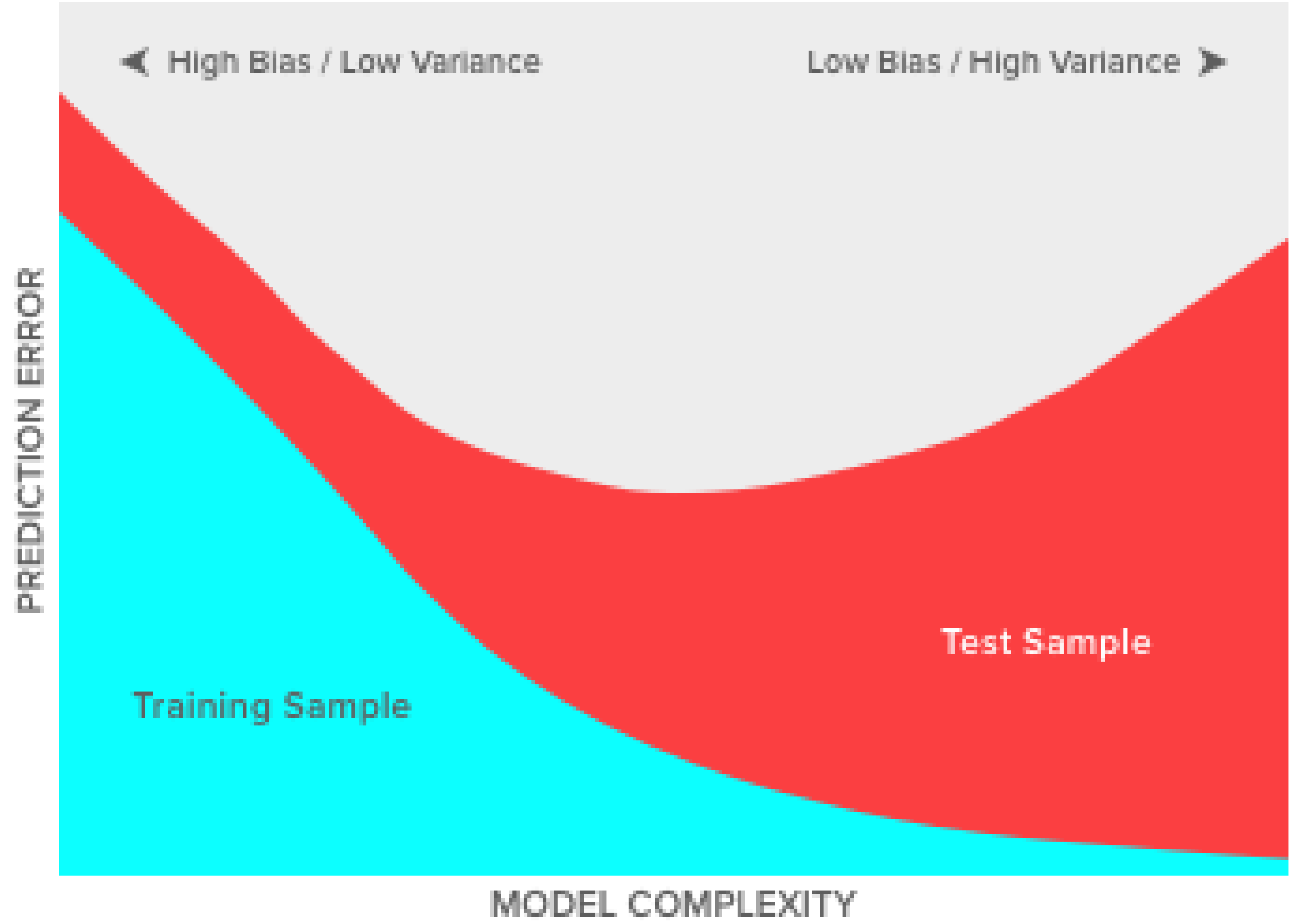
Unsupervised *Not Labeled Data*	**Supervised** *Labeled Data*	
Clustering	**Classification** *Non-ordered or Discrete Data*	***Regression*** *Ordered or Continuous Data*
• Hierarchical • K-means • Gaussian Mixture Model	• Random Forest • Neural Network • Gradient Boosting Trees • Decision Tree • Logistic Regression • Naïve Bayes • SVM	• Decision Tree • Linear Regression • Random Forest • Neural Network • Gradient Boosting Tree

Training ~ Validation ~ Test

LOCK-BOX APPROACH



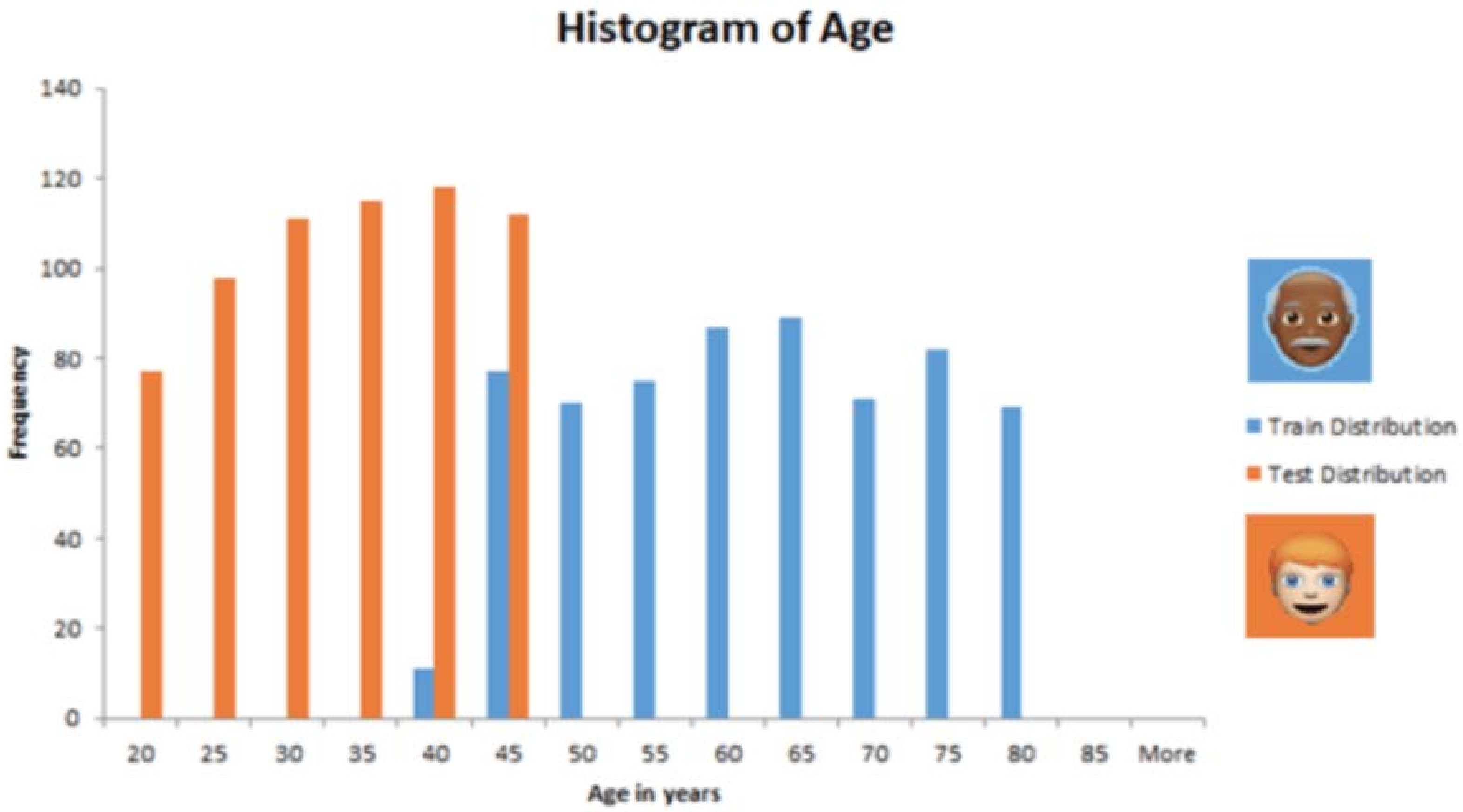
Bias-Variance Trade-off



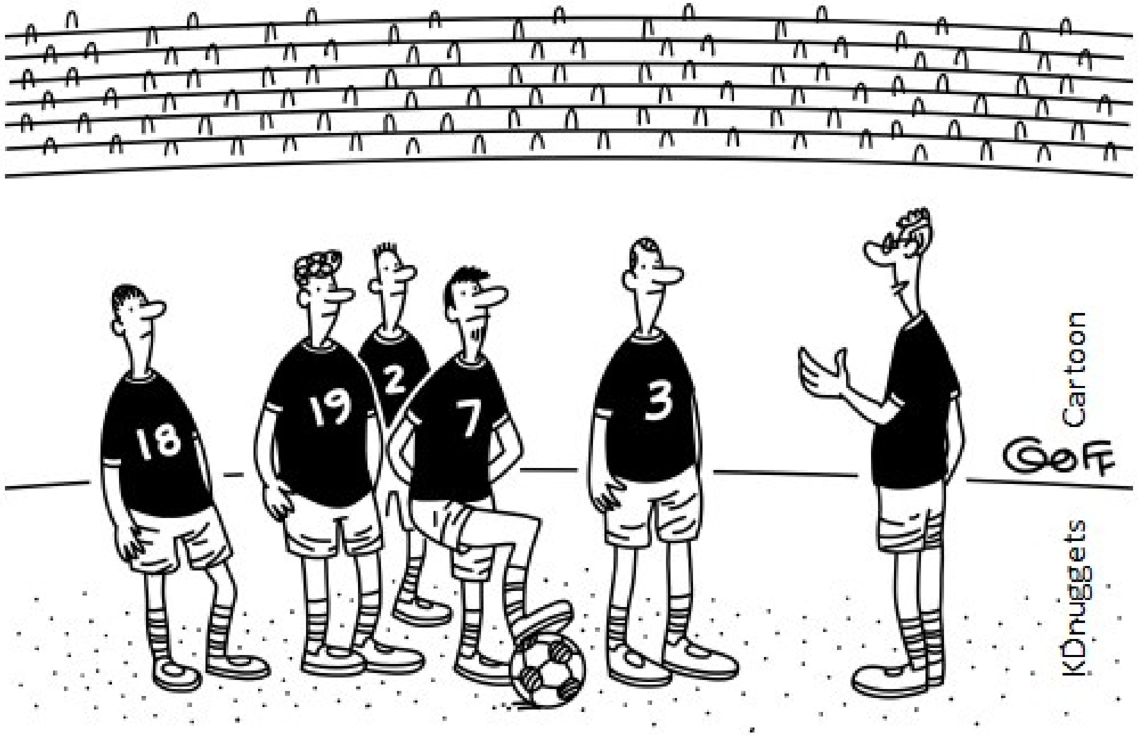
Overfitting:
An analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably.
~ Oxford Dictionaries

Control over- & underfitting

- Check similarity of train and test datasets



Control over- & underfitting



"Remember, the other team

is using Machine Learning on your

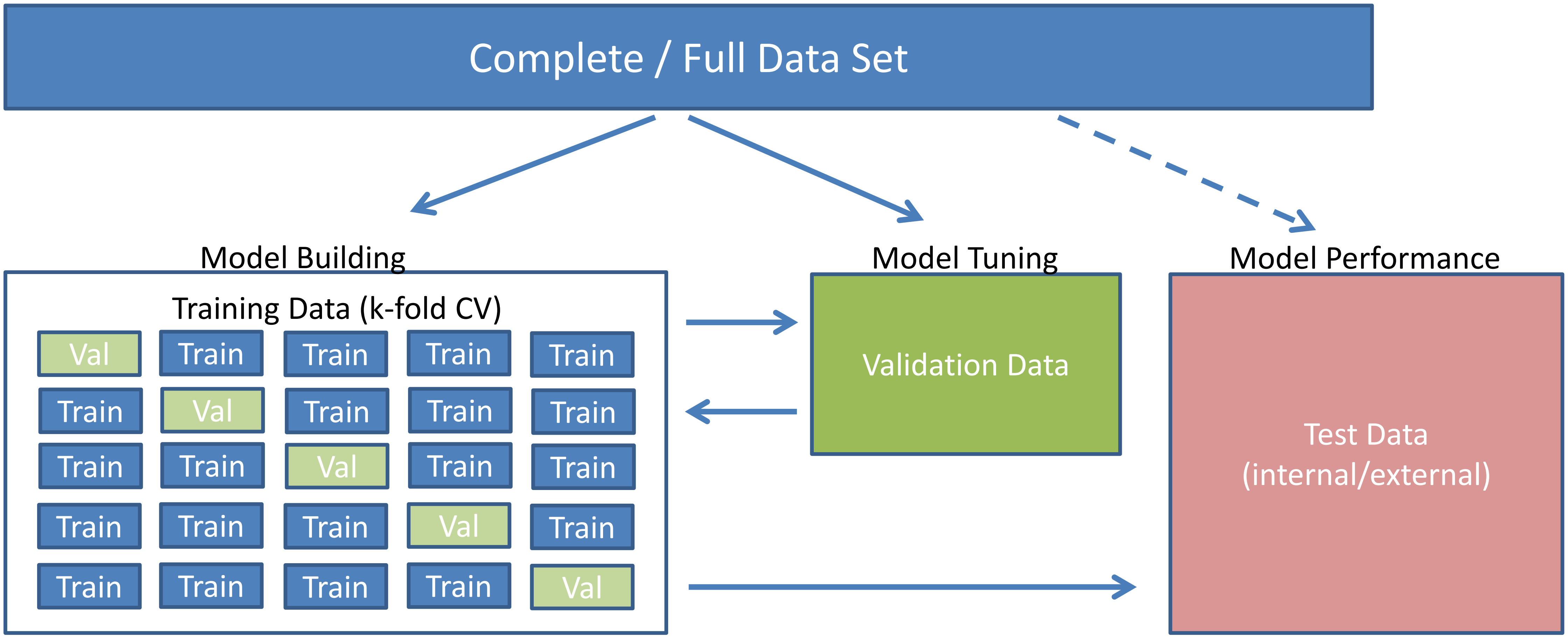
games to predict your play.

So, kick the ball with your other foot!"

Control over- & underfitting

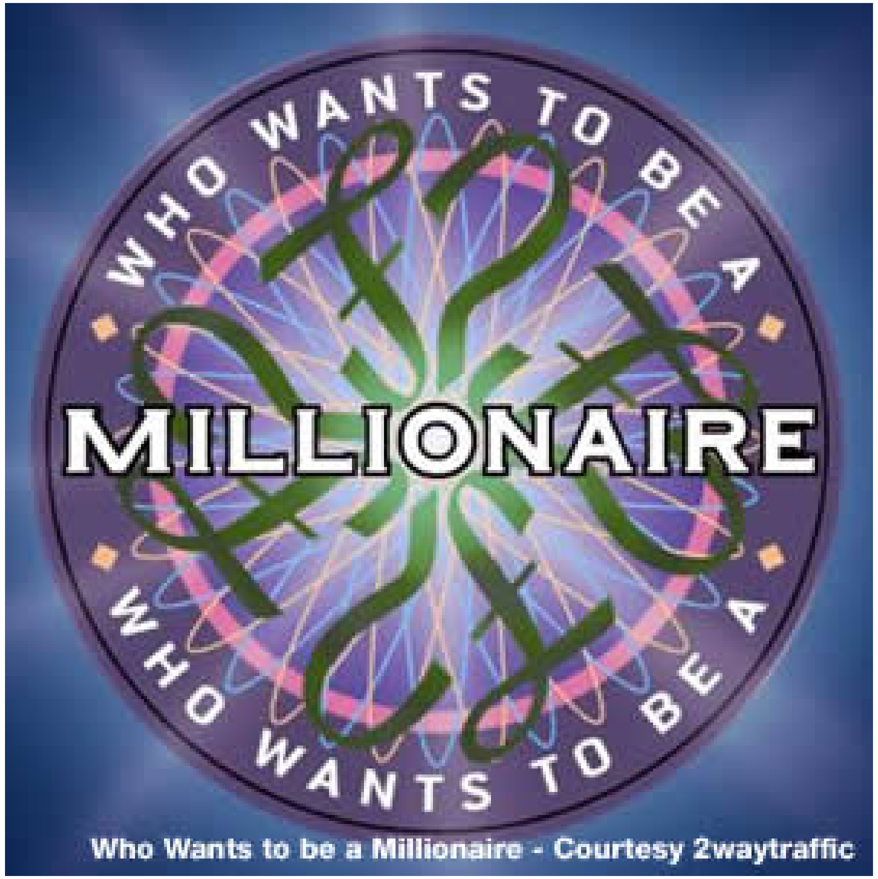
- Check similarity of train and test datasets
- Dimension reduction
- Check for Independent and Identically Distributed (IID) observations
- K-fold cross-validation
- Ensembles/Stacking/Bagging/Randomized Averaging

Training ~ Validation ~ Test



Wisdom of Crowds

Francis Galton



Wisdom of Crowds in ML

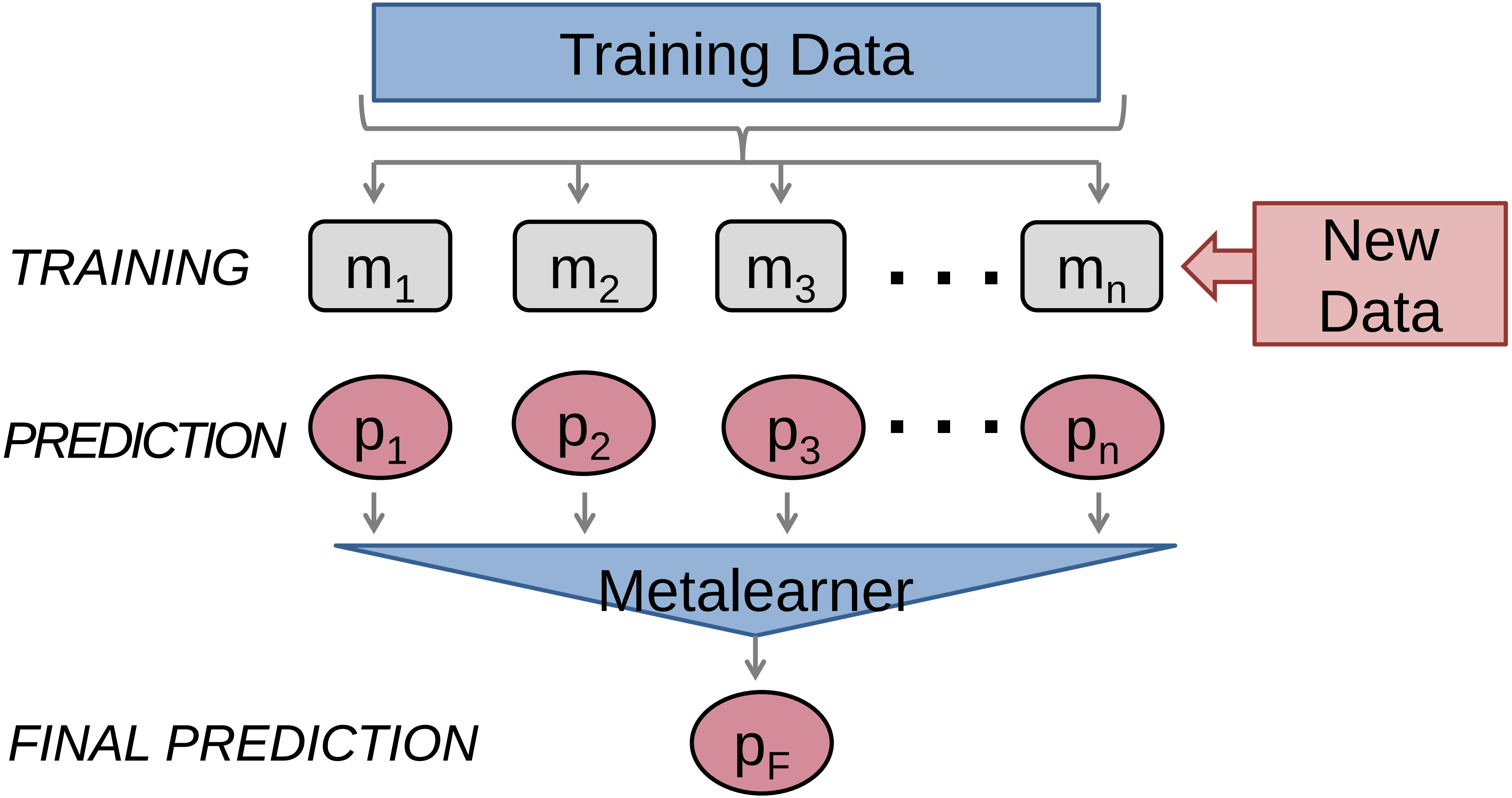
Ensembles:

Combine a diverse set of models

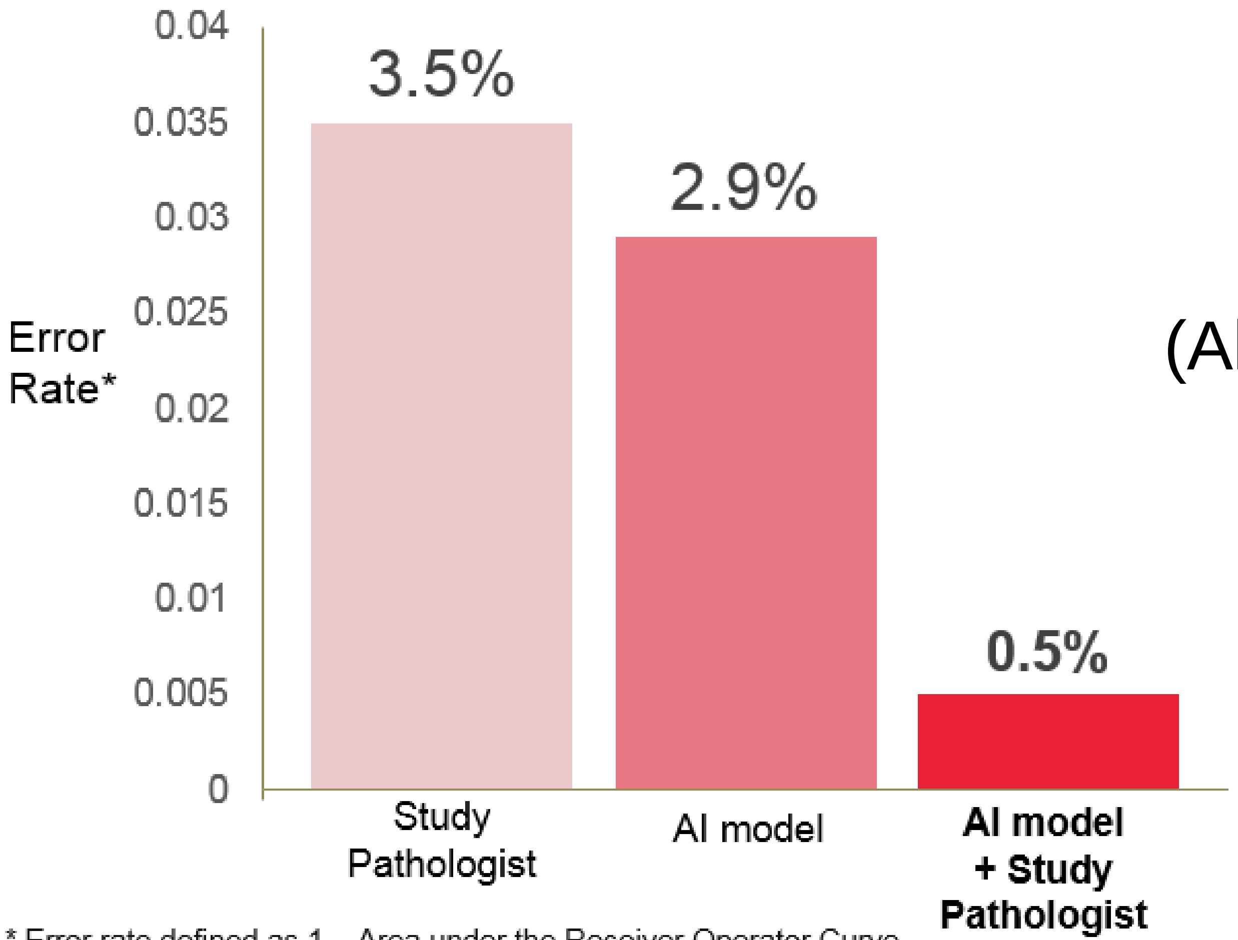
into a stronger,

high-performing model

Ensemble Learning / Stacked Models



Ensemble of Machine & Human



(Algorithm + Pathologist) > Pathologist

* Error rate defined as 1 – Area under the Receiver Operator Curve
** A study pathologist, blinded to the ground truth diagnoses, independently scored all evaluation slides.

https://blogs.nvidia.com/blog/2016/09/19/deep-learning-breast-cancer-diagnosis/

Clinical Trial Challenges

- Small sample size
  - – Resampling / Cross-validation
  - – External test set

- Imbalanced classifier
  - – Can get very high prediction accuracy with low event rates
  - – Solutions:
    - Simple models
    - Data class weighting
    - Oversample / Undersample
    - Search over a variety of models & perform hyper-parameter search

!!! Remain critical !!!

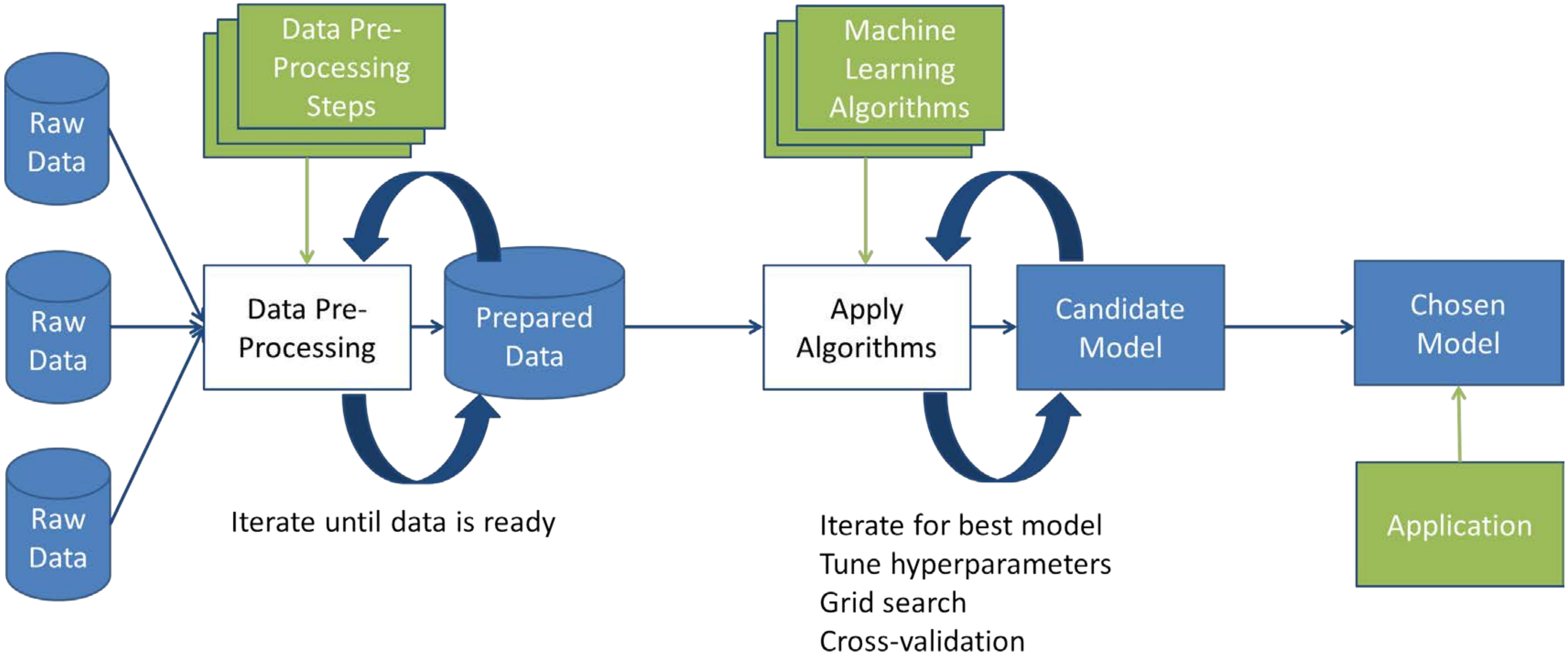
Traditional / Predictive Modeling

Traditional statistical model may be better	ML may be better
Uncertainty is inherent and signal-noise ratio is not large	Signal-noise ratio is large and outcome being predicted doesn't have a strong component of randomness
Isolate effects of small number of variables	Overall prediction is the goal, without the need to succinctly describe the impact of any one variable
Uncertainty in overall prediction or the effect of a predictor is sought	Not interested in estimating uncertainty in forecasts or in effects of selected predictors
Small sample size	Sample size is HUGE
Isolate effects of "special" variables such as treatment as a risk factor	No need to isolate effect of a special variable such as treatment
Entire model needs to be interpretable	Not concerned that the model is a 'black box'

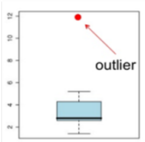
Source: Frank Harrell, Statistical Thinking Blog https://www.fharrell.com/post/stat-ml/

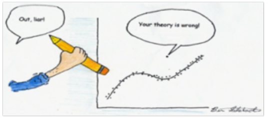
MACHINE LEARNING PIPELINE

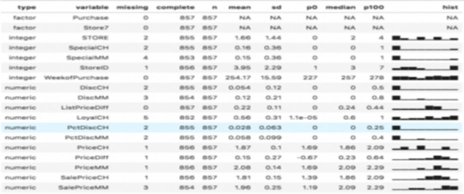
ML Pipeline


Raw Data
Raw Data
Raw Data
Data Pre-Processing Steps
Data Pre-Processing
Prepared Data
Iterate until data is ready
Machine Learning Algorithms
Apply Algorithms
Candidate Model
Iterate for best model
Tune hyperparameters
Grid search
Cross-validation
Chosen Model
Application

Data Pre-Processing

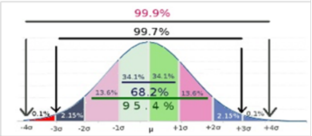


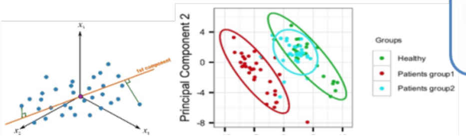




group	group_A	group_B	group_C
A	1	0	0
B	0	1	0
C	0	0	1

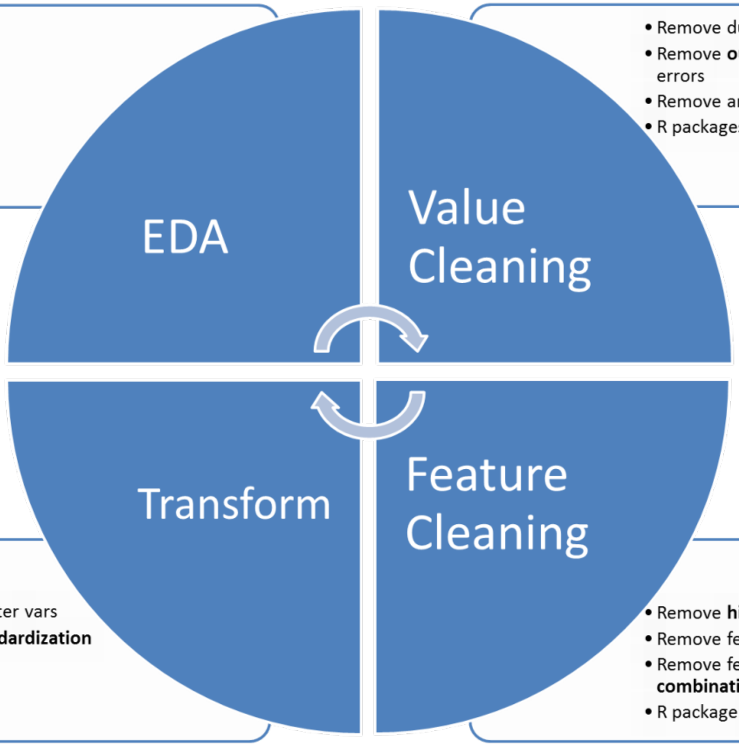
$$z = \frac{x - \mu}{\sigma}$$





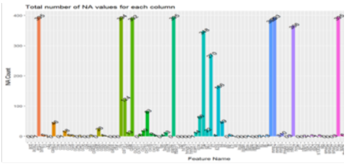
- **Summarize, Visualize**
- R Packages: *skimr, ggplot2*

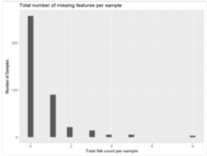
- Remove duplicate records
- Remove **outliers** if true measurement errors
- Remove and/or impute **missing values**
- R packages: *mice, caret*

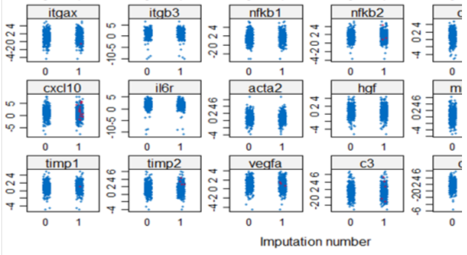


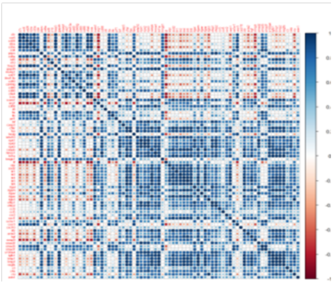
- **One-Hot Encoding** of character vars
- **Normalization/Scaling/Standardization**
- **Dimension Reduction**: PCA
- R packages: *caret*

- Remove **highly correlated** features / PCA
- Remove features with **near-zero variance**
- Remove features that are **linear combinations** of other features
- R package: *caret*



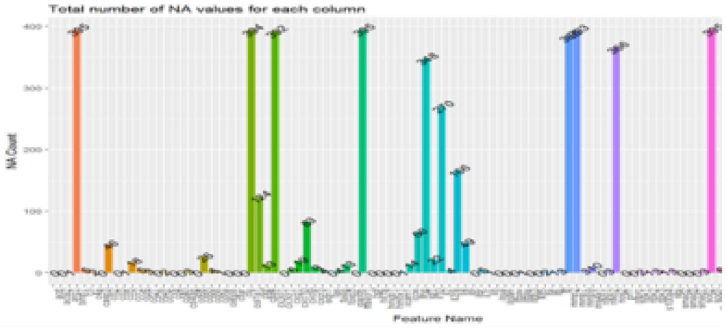


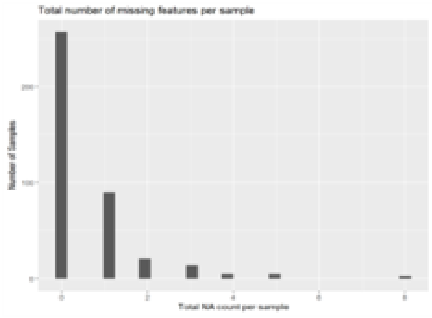


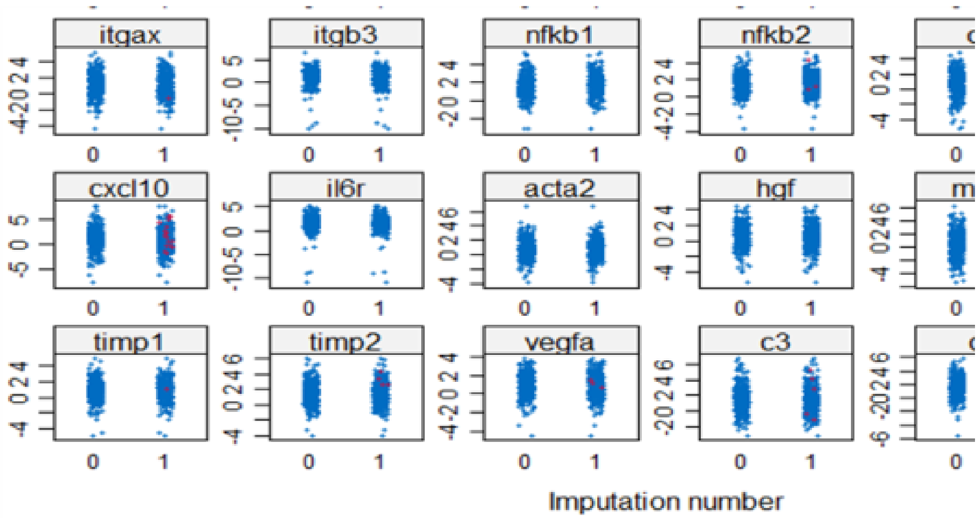


Value Cleaning

- Remove duplicate records
- Remove **outliers** if true measurement errors
- Remove and/or impute **missing values**
- R packages: *mice*, *caret*

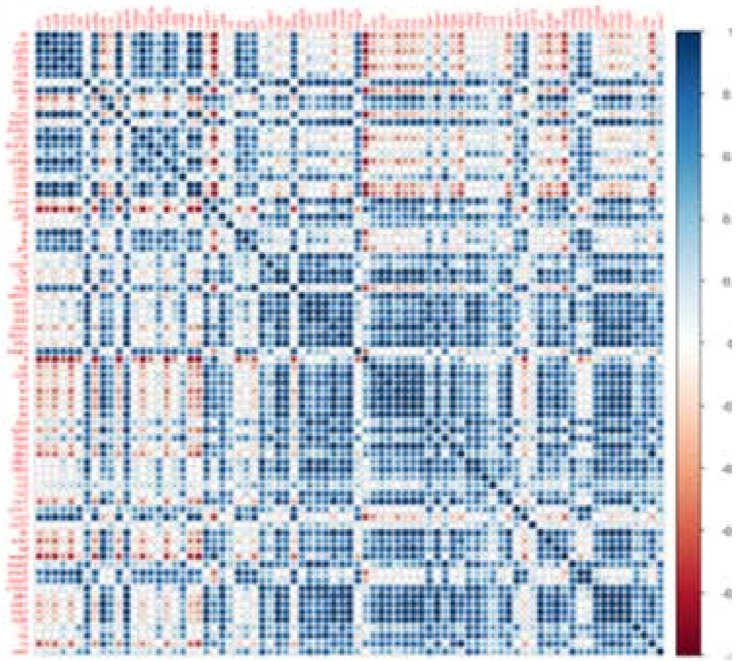






Feature Cleaning

- Remove **highly correlated** features / PCA
- Remove features with **near-zero variance**
- Remove features that are **linear combinations** of other features
- R package: *caret*

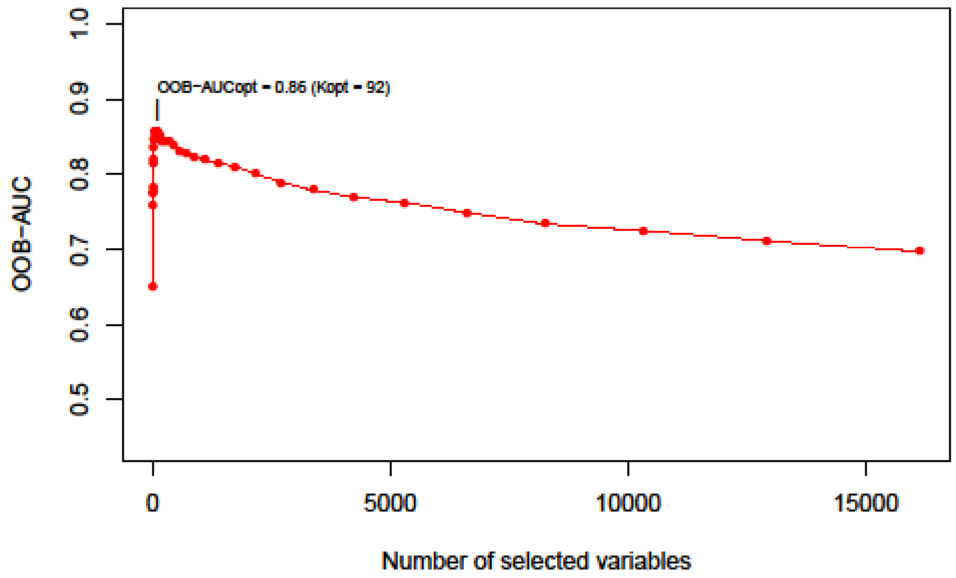


Feature Selection

- Select discriminative features

  E.g. differentially expressed between groups
- Recursive Feature Elimination

  E.g. through optimization of the AUC
    - R::AUCRF



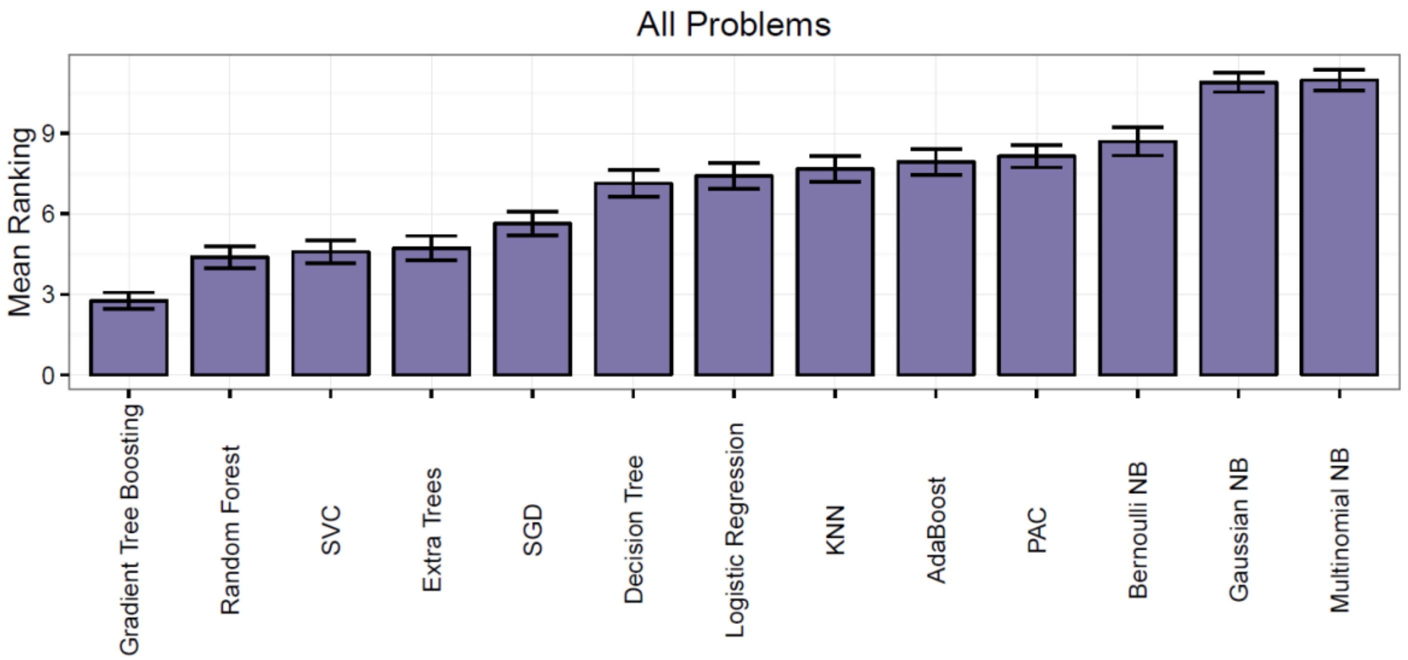
- Algorithm decides
  - Feature importance
  - Least assumptions
  - Allows for possible interactions/mediations of features

Algorithm Selection

Familiar	versus	Unfamiliar – black box
Simple		Complex
Easy to interpret / understand clinically		Harder to interpret / understand
Reasonable prediction		Excellent prediction with high-accuracy

Algorithm Selection

Paper by Olsen et al 2018 “Data-driven advice for applying machine learning to bioinformatics problems”:



- Use Ensemble Trees when in time crunch
- No silver bullet algorithm
  - Test a suite of algorithms AND
  - Test a suite of parameters for each algorithm (tuning)

!!! REMAIN CRITICAL !!!

Model Tuning / Evaluation

Validate on:

- Train / validation split (hold-out set)
- K-fold cross-validation
- Bootstrap resampling
- Visualization

Performance metrics:

Supervised

- AUC / logloss
- MSE / RMSE

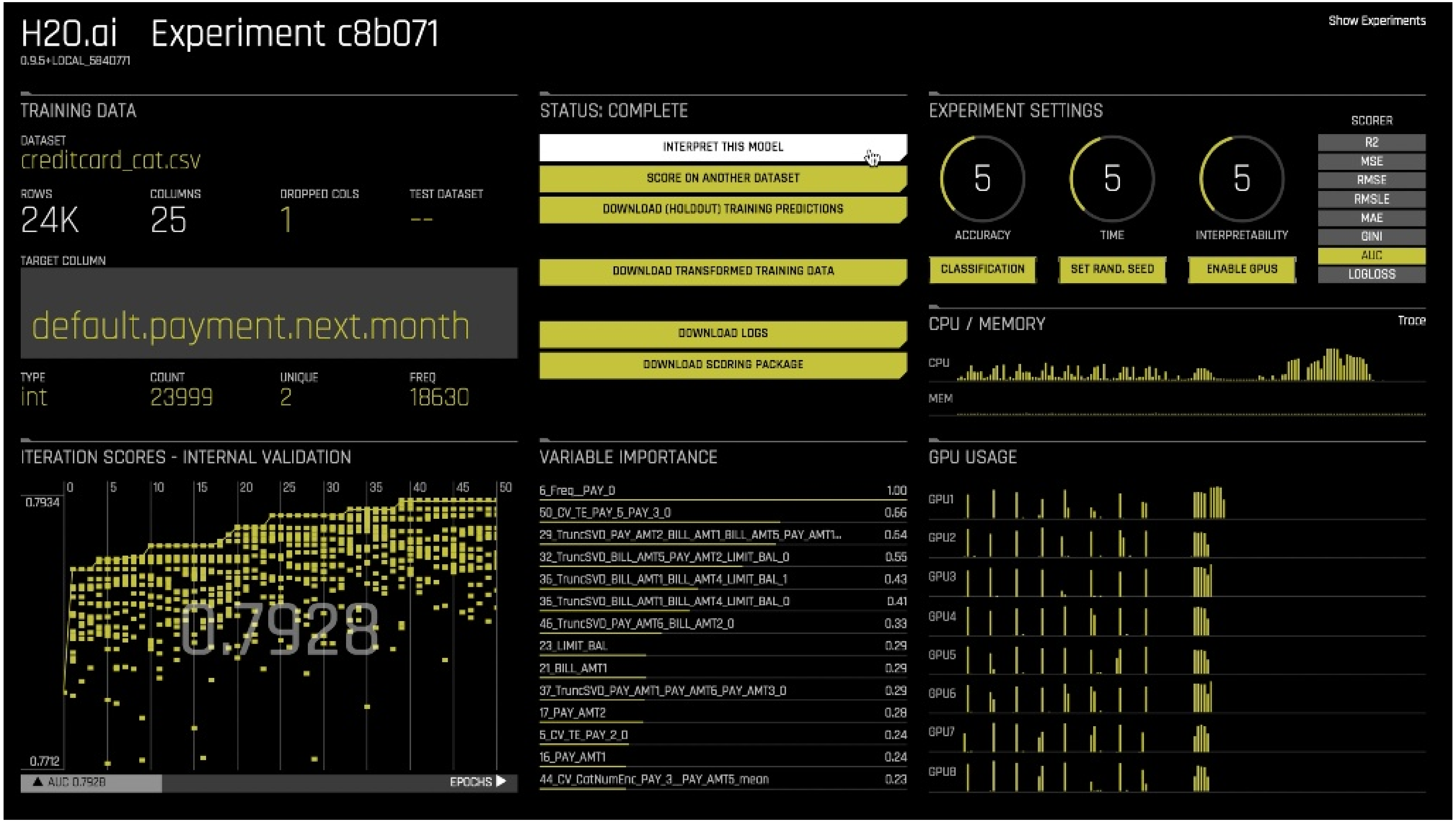
Unsupervised

- AIC, BIC, TSS

In Practice

- Demand for ML experts has outpaced the supply, despite the surge of people entering the field
- ML software
  - Easy to use interfaces
  - Non-expert use
- Automate process of training a large selection of candidate models

Driverless AI by H2O



AutoML by H2O

- Simple wrapper function, interface with R and Python
- Performs a large number of modeling-related tasks
  - no need for lots of lines of code
- Freeing up time to focus on other aspects of the data science pipeline
  - E.g. data pre-processing, feature engineering, model deployment
- Automatic training and tuning of many models with user-specified stopping criteria
  - Time-limit
  - Performance metric
- Automatically trains Ensembles on the collection of individual models
  - Produces highly predictive ensemble model which in most cases will be the top preforming model in the AutoML Leaderboard

RStudio in the Cloud

- Free-up local resources
- Work from any location
- Lots of computing power at your fingertips

RStudio image

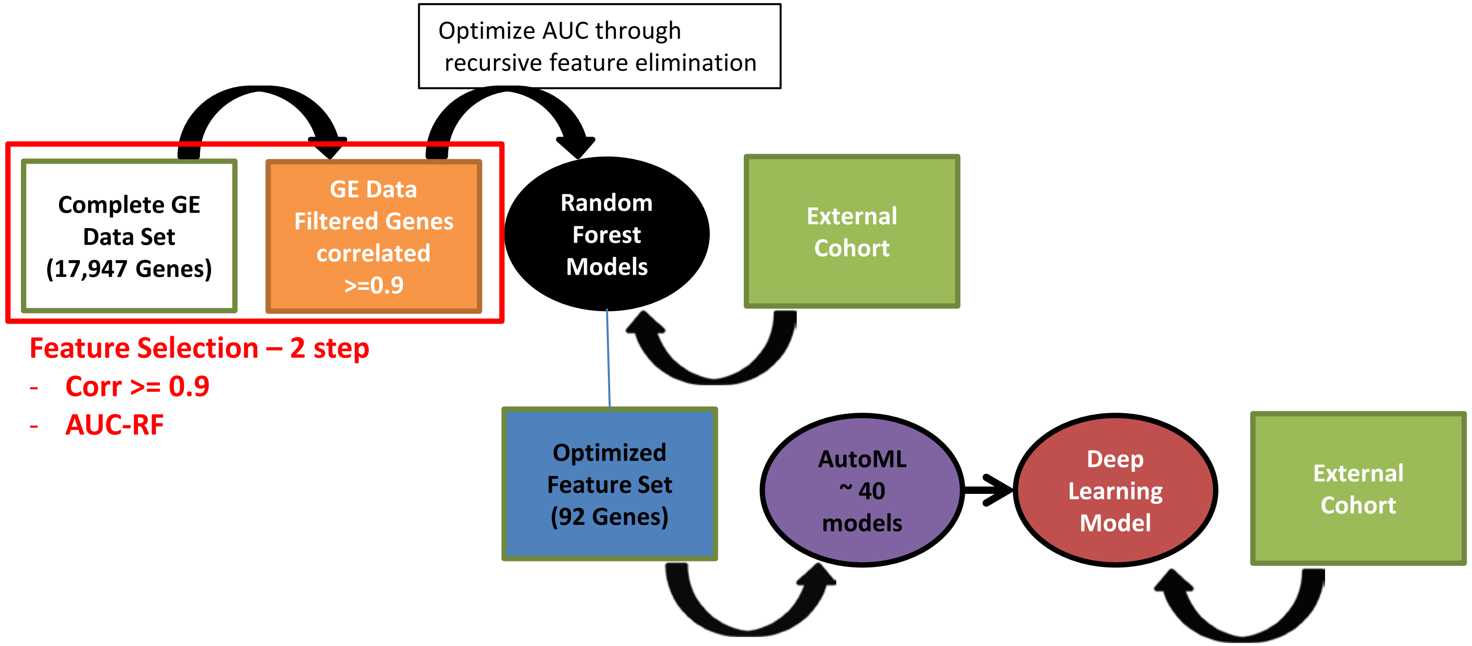
Scalable

Parallel computing

Minimal cost

Example: Classification Pipeline

- Subacute rejection versus Transplant Excellent
- 536 samples
- Gene expression of 17,947 genes



Features	AUC (test)
92	0.67

Features	AUC (test)
92	0.73

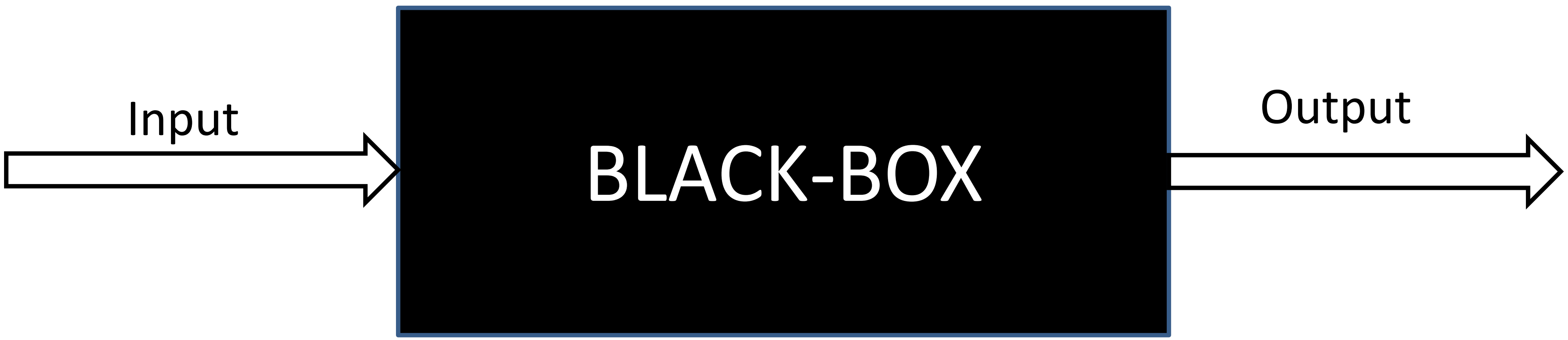
Example: H2O AutoML R Code

```
library(h2o)
h2o.init(nthreads = -1)
train <- h2o.importFile(data_path)
test <- h2o.importFile(data_path)

y     <- 'pheno'
x     <- setdiff(names(train), y)

aml <- h2o.automl(x = x,
                  y = y,
                  training_frame = train,
                  max_runtime_secs = 3600/2,
                  stopping_metric = "AUC",
                  seed = 12345)

## Save model
h2o.saveModel(aml@leader, path="/project/automl_results", force=TRUE)

## print leaderboard
 print(aml@leaderboard)

## Predictions & Performance
 h2o.predict(aml, test)
 h2o.performance(aml@leader, test)
```

MACHINE LEARNING INTERPRETATION

ML Challenges

- ML models are often hard to explain – lots of high-degree interactions and non-linear model behavior.
- Some algorithms learn how to weigh complex combinations of input variables



What is ML Interpretability?

“The ability to ***explain*** or to present

in ***understandable*** terms

to a ***human***”

-- Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning.” In: arXiv preprint 2017

URL: https://arxiv.org/pdf/1702.08608.pdf

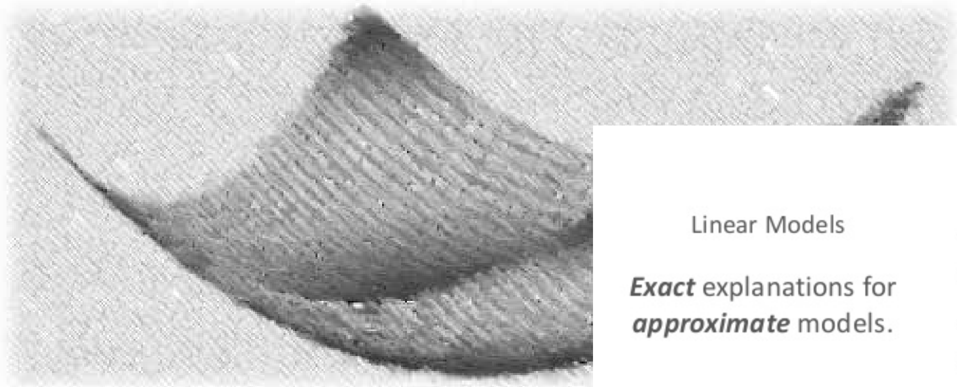
What is a Good Explanation?

“When you can no longer keep asking why”

-- Gilpin, Leilani H et al. (2018). “Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning.” In: arXiv preprint arXiv:1806.00069. URL: https://arxiv.org/pdf/1806.00069.pdf.

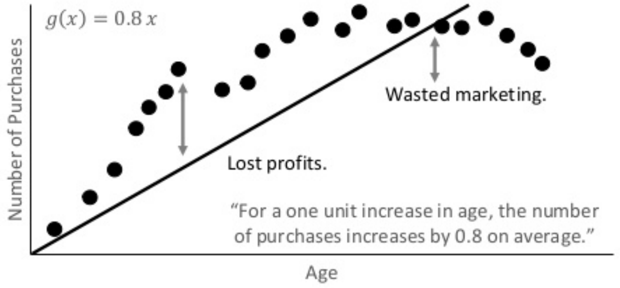
Interpretable Model vs a ML model

Linear Models
Strong model locality
Usually stable models and explanations



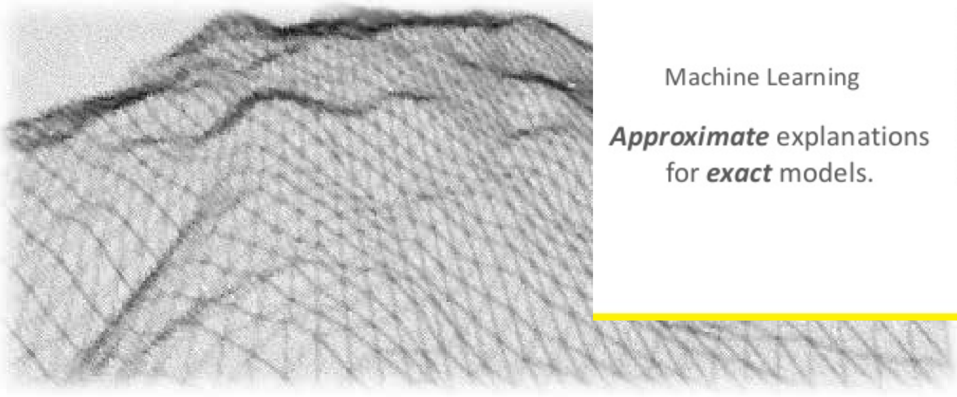
Linear Models

Exact explanations for ***approximate*** models.



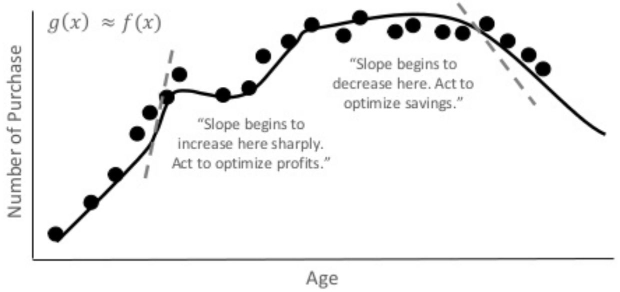
E.g. Linear Regression $y = w_0 + w_1x_1 + w_2x_2 + \ ... + wnxn$

Machine Learning
Weak model locality
Sometimes unstable models and explanations
(a.k.a. The Multiplicity of Good Models)



Machine Learning

Approximate explanations for ***exact*** models.



Why should we Care about Interpretability?

ML models have entered **critical areas** like health care, justice systems and financial industry.

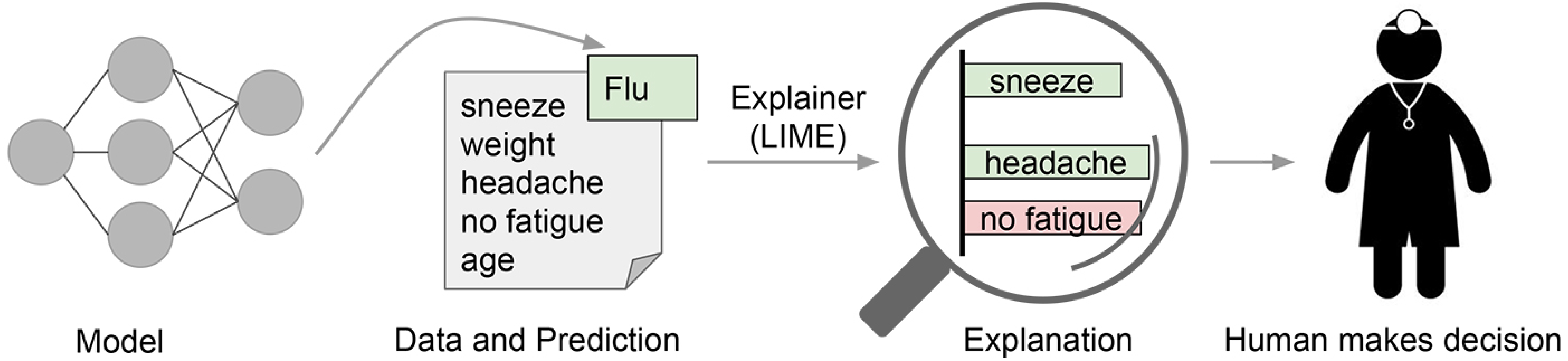
- **Financial industry**: highly regulated and loan issuers are required by law to make fair decisions and explain their credit models to provide reasons whenever they decide to decline a loan application
- A **medical diagnosis** model is responsible for human life. How can we be confident enough to treat a patient as instructed by a black-box model?
- When using a **criminal decision** model to predict the risk of recidivism at court, we have to make sure the model behaves in an equitable, honest and nondiscriminatory manner → prevent sociological biases.
- If a **self-driving car** suddenly acts abnormally and we cannot explain why, are we going to be comfortable enough to use the technique in real traffic in larger scale?

What do we want to Achieve with Interpretability?

Answer questions like

- *Why did the algorithm make certain decisions*?
- *What variables were the most important in predictions?*
- *Is the model trustworthy*?

- Explain hypothesis
- Explain why phenomena are happening
- Complete transparency & accountability
- Ensure ML models are unbiased, fair and trustworthy



Flu
sneeze
weight
headache
no fatigue
age
Explainer
(LIME)
sneeze
headache
no fatigue
Model
Data and Prediction
Explanation
Human makes decision

SAMSI – IMSM Workshop

samsi
NSF Duke NCSU UNC

- Statistical and Mathematical Sciences Institute
  https://www.samsi.info/
- Workshops, Visiting scholars, Research fellows, Outreach

- Industrial Mathematical and Statistical Modeling (IMSM) workshop
- 1995 – present
- Held in July in SAS Hall at NCSU
  - ~ 6 industrial problem presenters
  - ~ 6 faculty mentors
  - 30 – 45 math/stat/engineering graduate students
  - 9 days to complete the project
- Rho was problem presenter for 9th consecutive year this year

SAMSI Project: Predicting Liver Disease

Open Source Data set:

- Indian Liver Patient data set -- From North East India
- 583 subjects: 416 with liver disease & 167 without
- For model building: 467 Training & 116 Test

Outcome: Liver Disease (Yes/No)

Predictors (10): Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, Albumin and Globulin Ratio

GBM model built with H2O

→ Develop/explore methodologies to explain and visualize how this model made its predictions

ML Interpretability Methodologies

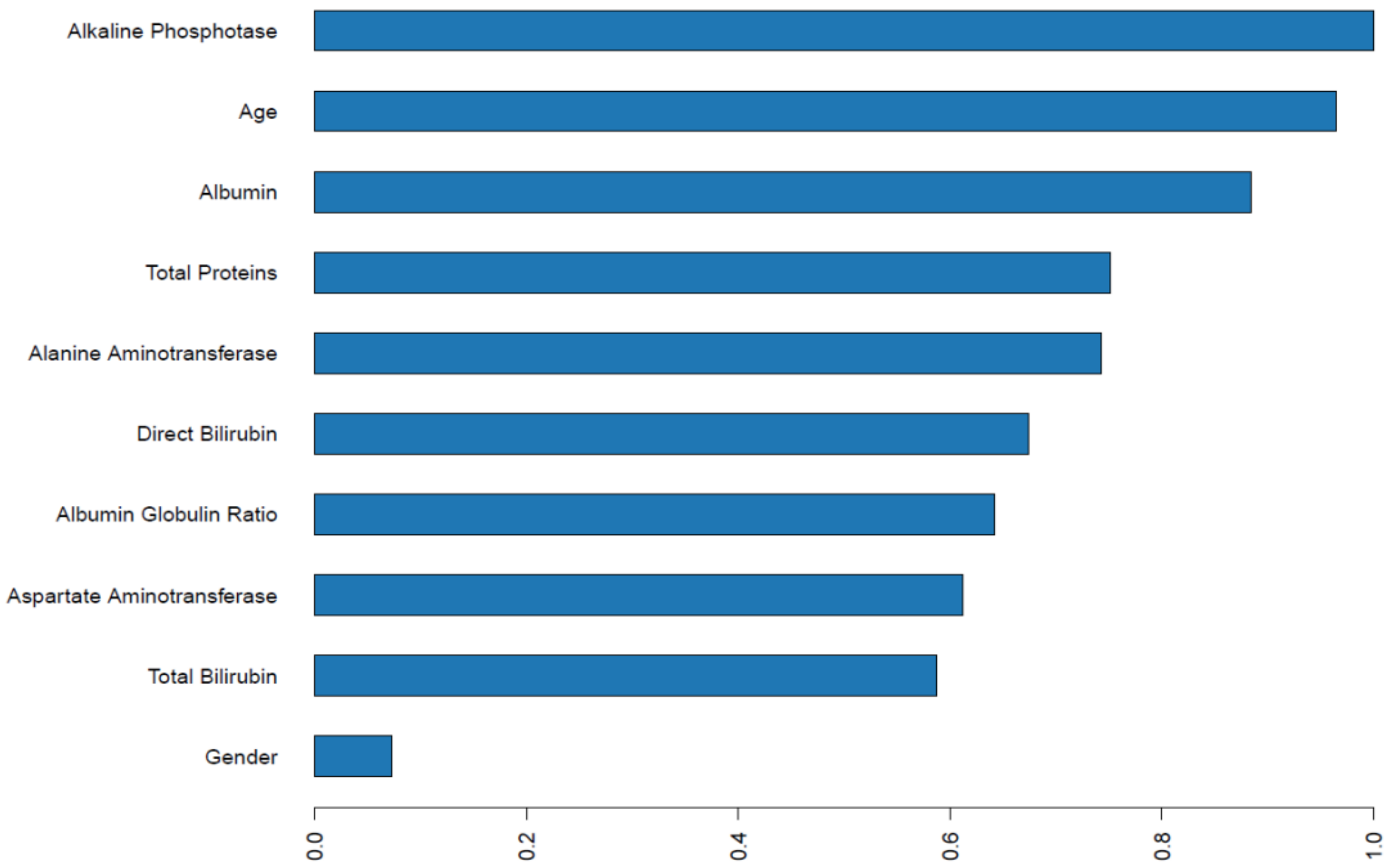
Global Interpretation:

- Variable Importance Plots (VIP)
- Surrogate Models
- Partial Dependency Plots (PDP)

Local Interpretation:

- Individual Conditional Expectation (ICE) plots
- Local Interpretable Model-agnostic Explanations (LIME) plots

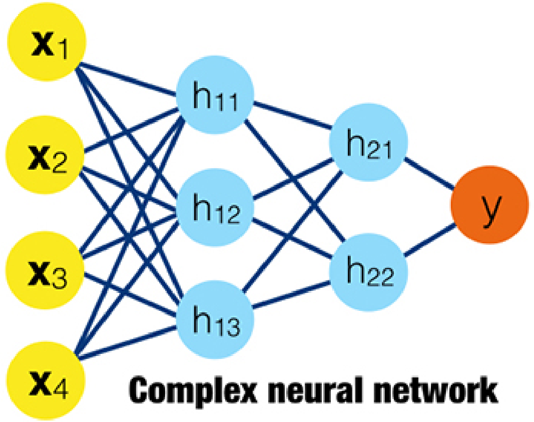
Global - Variable Importance Plots



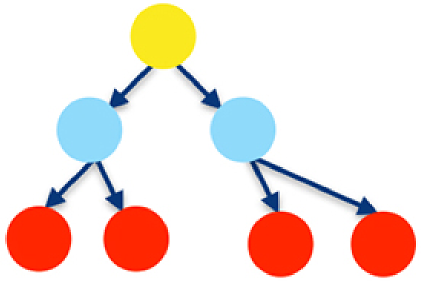
Global – Surrogate Model

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

1. Train a complex machine learning model

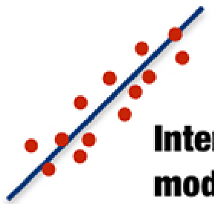


Complex neural network



Interpretable decision tree

Or



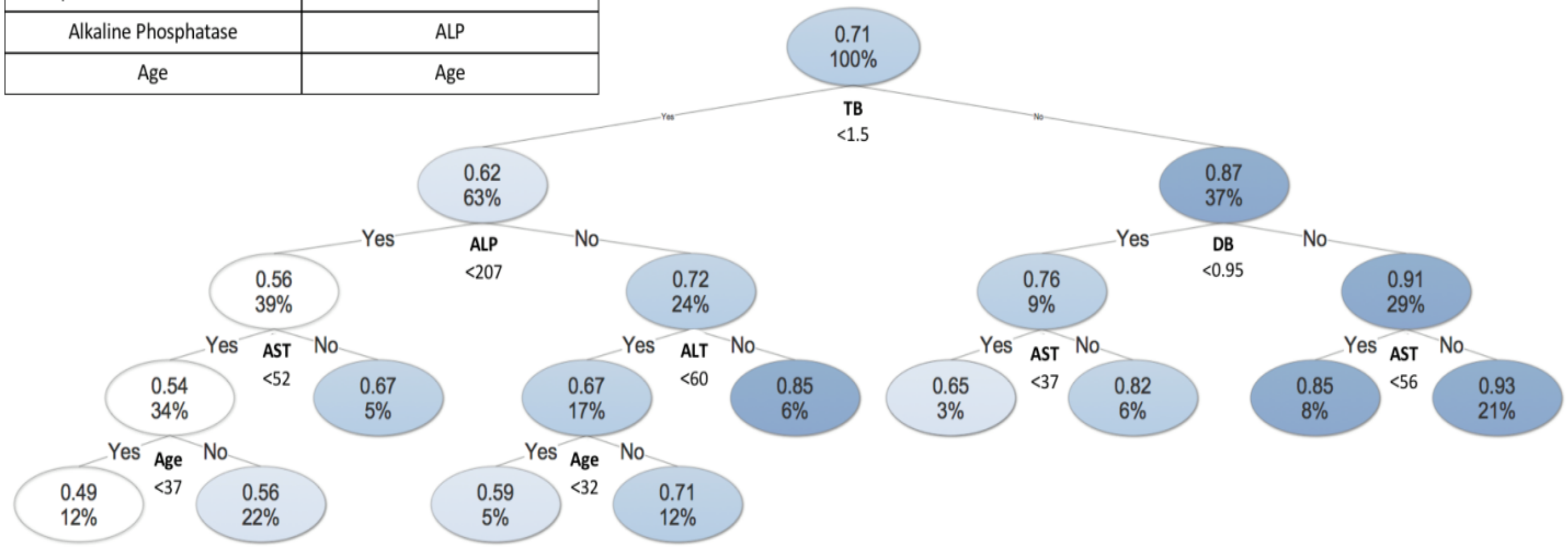
Interpretable linear model

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

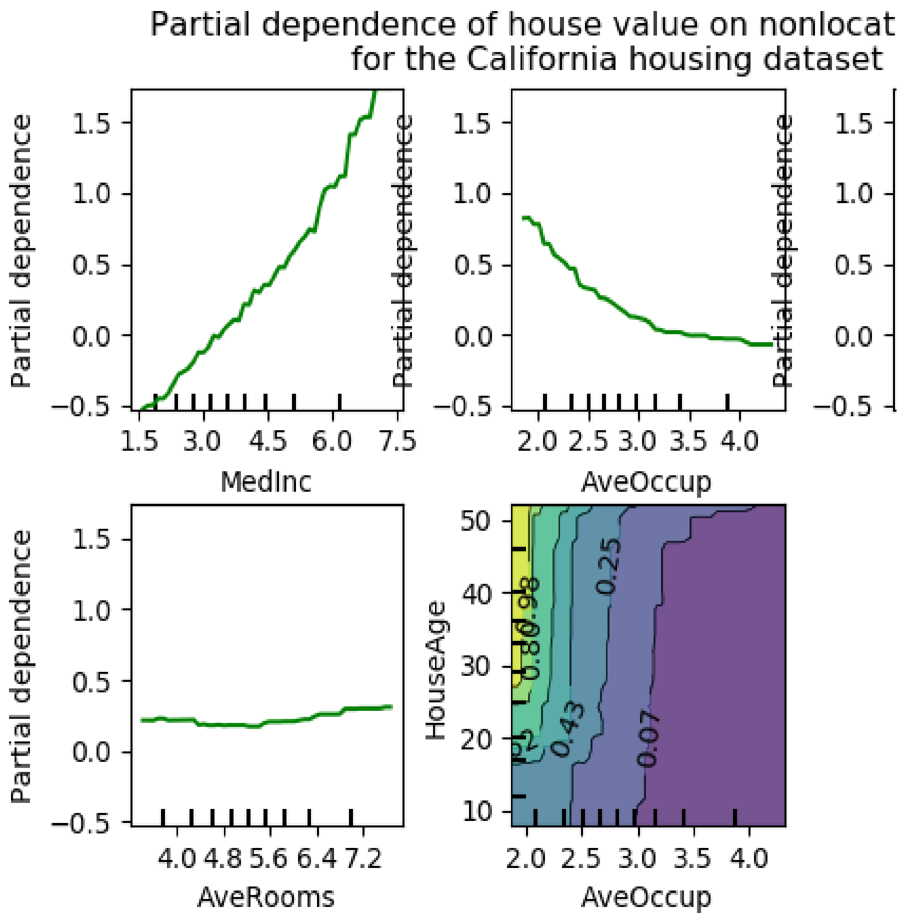
2. Train an interpretable model on the original inputs and the predicted target values of the complex model

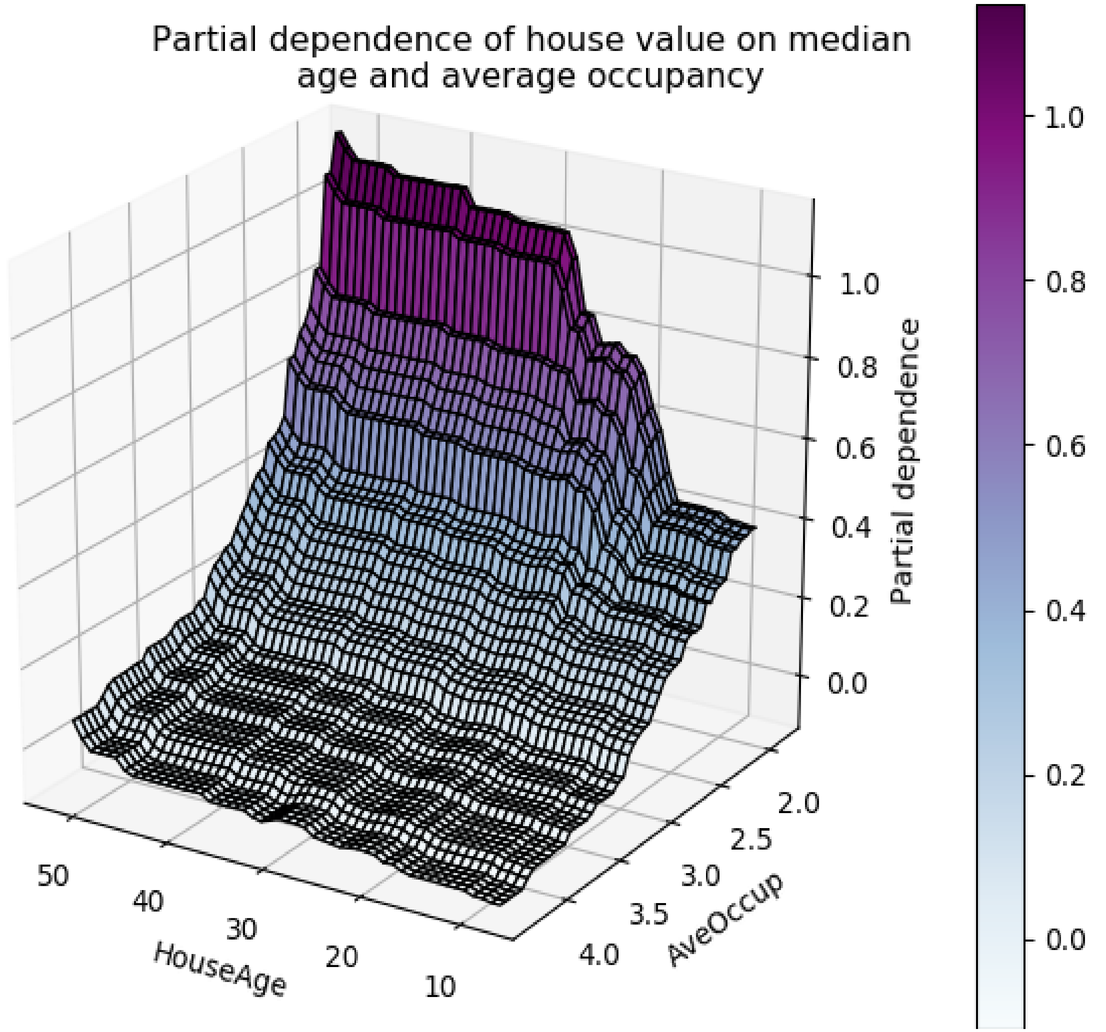
Global – Surrogate Model

Key	
Variable	**Abbreviation**
Total Bilirubin	TB
Direct Bilirubin	DB
Alanine Aminotransferase	ALT
Aspartate Aminotransferase	AST
Alkaline Phosphatase	ALP
Age	Age



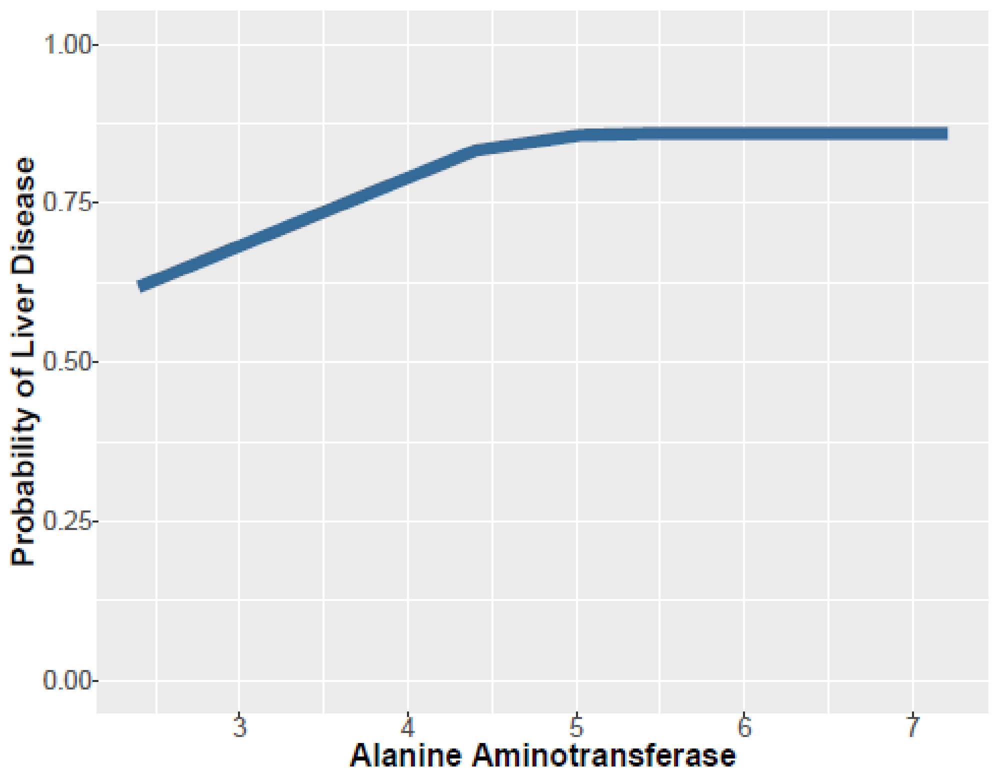
Global – Partial Dependency Plots (PDP)

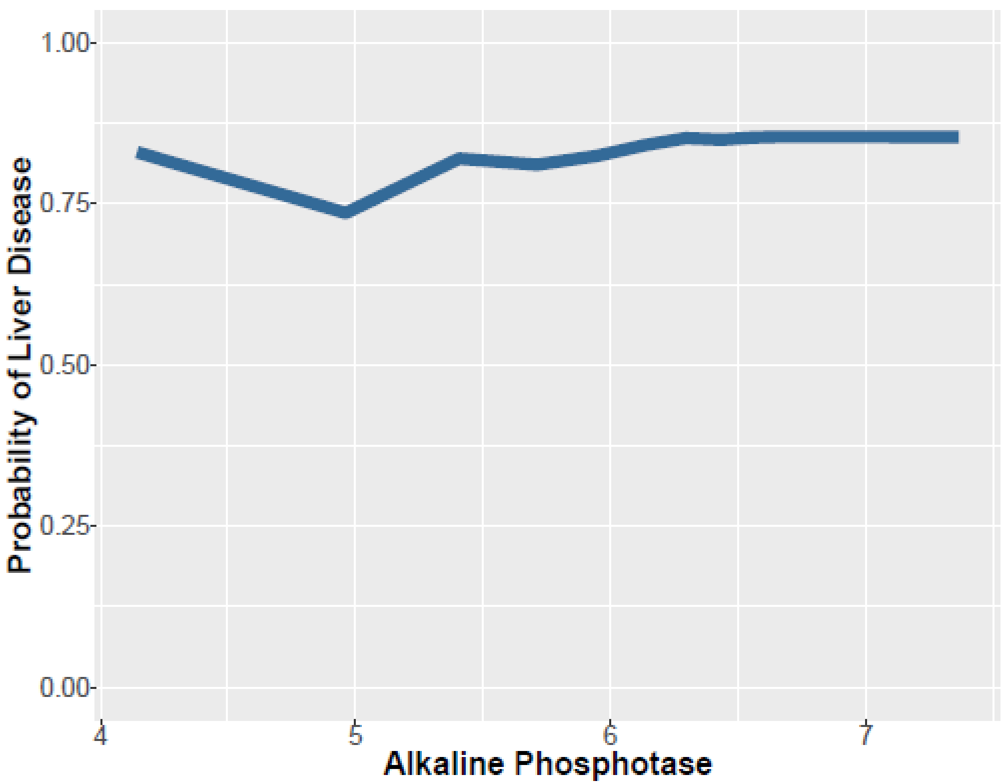




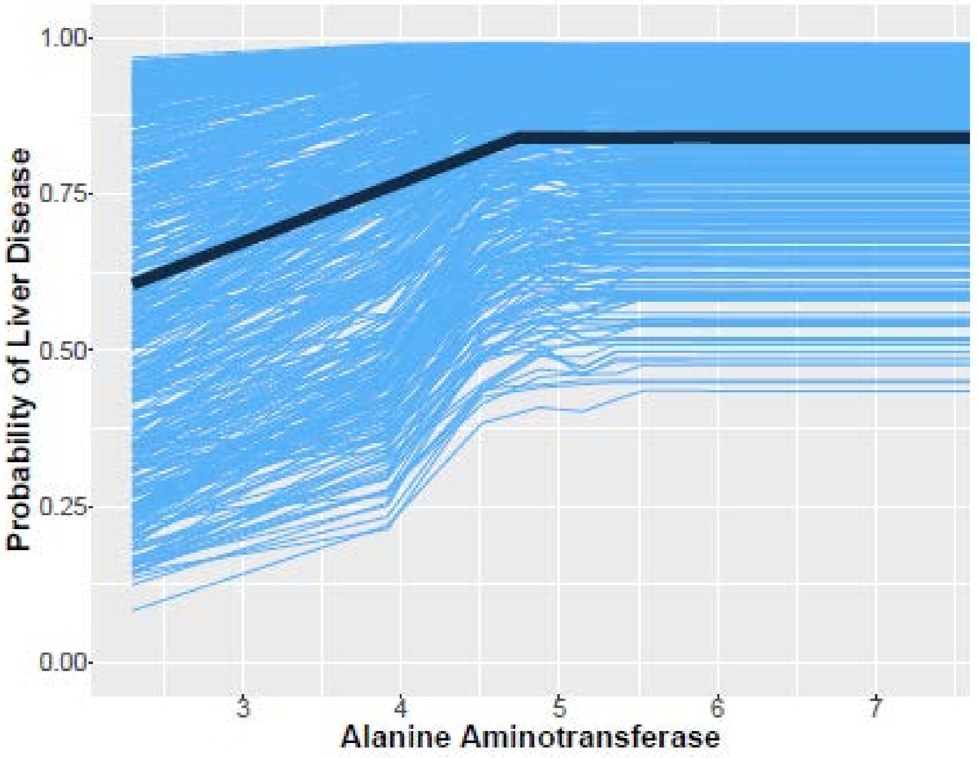
$$HomeValue = MedInc + AveOccup + HouseAge + AveRooms$$

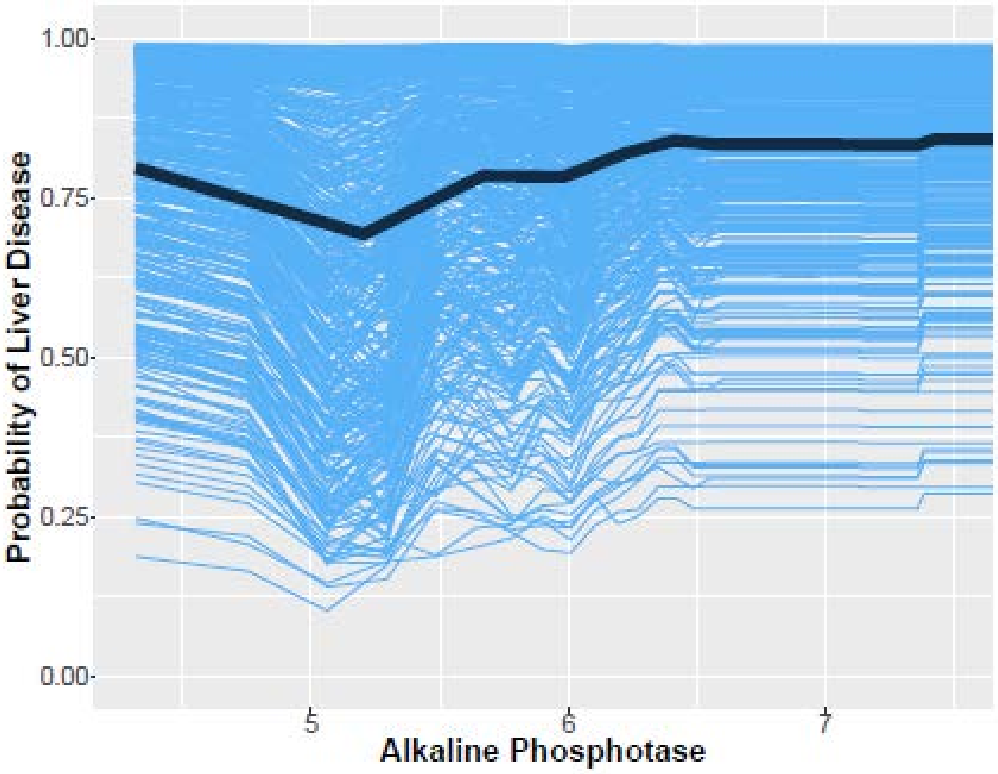
Global – Partial Dependency Plots (PDP)





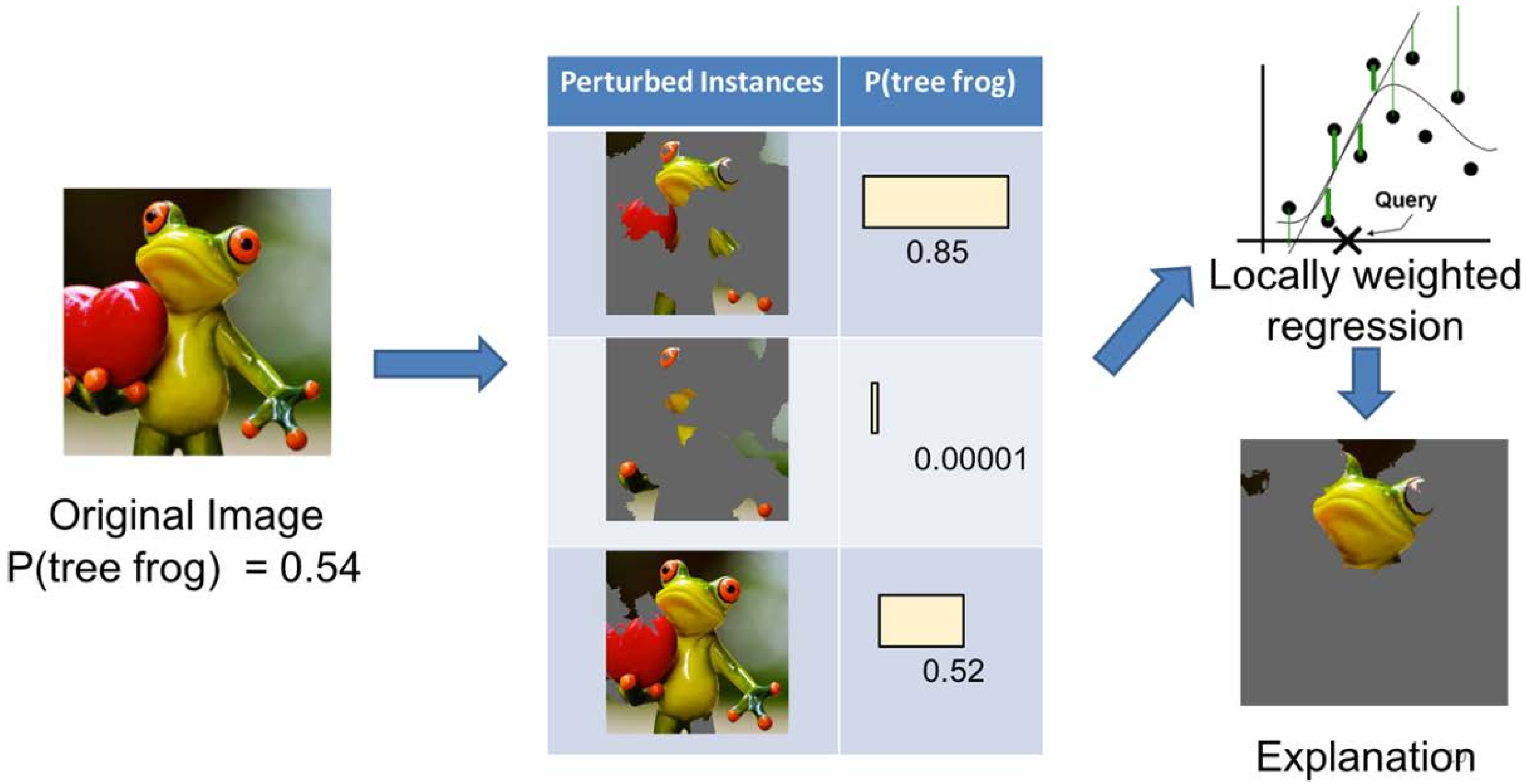
Local – Individual Conditional Expectation (ICE) plots



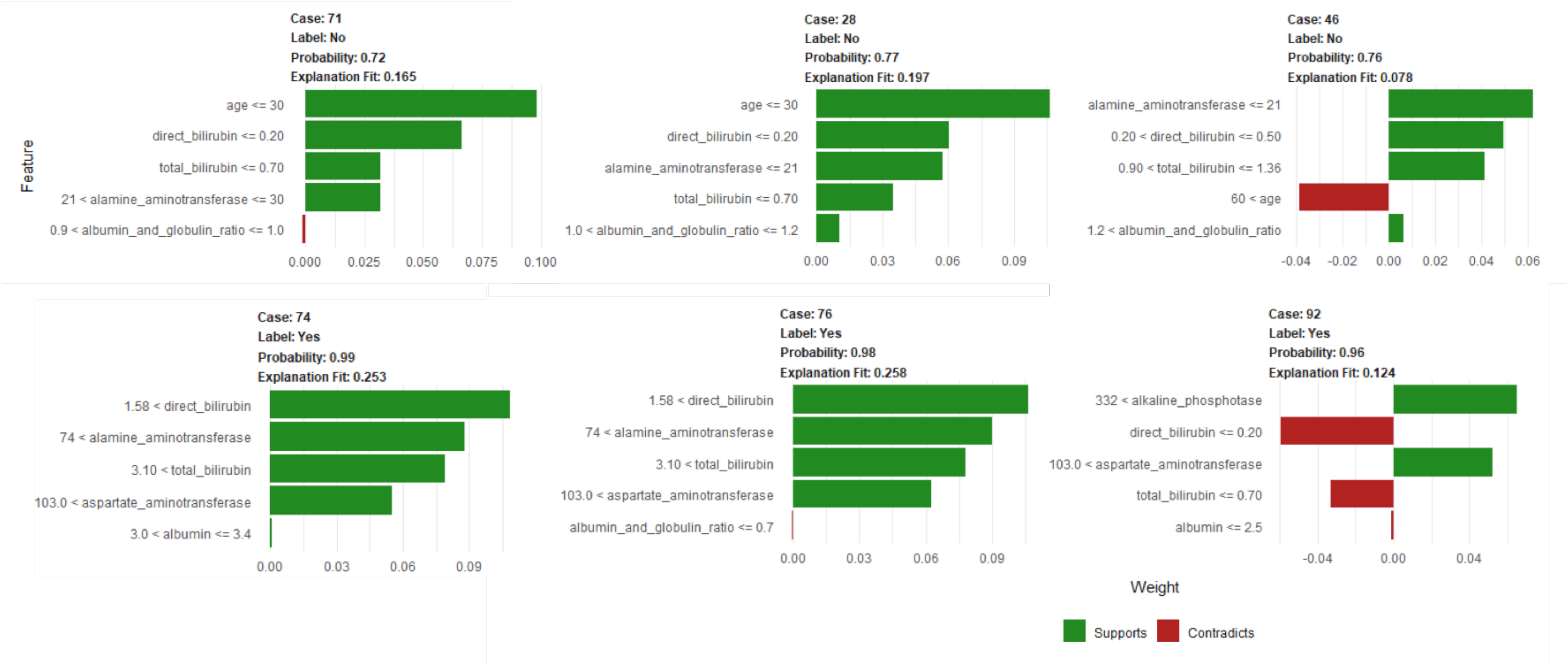


Visualizes the dependency of the predicted response on a feature for EACH instance separately, resulting in multiple lines

Local Interpretable Model-agnostic Explanations (LIME)



Local Interpretable Model-agnostic Explanations (LIME)



Conclusions on Liver Disease Prediction

- Global-Variable Importance Plot (VIP)
    1. Alkaline Phosphatase
    2. Age
    3. Albumin
- Global/Local – PDP & Independent Conditional Expectation (ICE)
    - Alkaline phosphatase (extreme lower and higher levels)
    - Age and total bilirubin (direct positive relation)
- Local-Locally Interpretable Model Explanations (LIME)
    - Age (elderly), alanine phosphatase, and bilirubin (higher levels)

Interpretation Summary

- Use simpler low-fidelity or sparse explanations to understand more complex high-fidelity explanations
- Global and local explanatory techniques are often needed to explain a model
- Seek consistent results across multiple explanatory techniques
- To increase adoption, production deployment of explanatory methods must be straightforward → work in progress

Resources - R packages

- Caret http://topepo.github.io/caret/index.html
- mlr https://cran.r-project.org/web/packages/mlr/index.html
- h2o https://cran.r-project.org/web/packages/h2o/index.html
  https://www.h2o.ai/
- DALEX https://cran.r-project.org/web/packages/DALEX/index.html
- Lime https://t.co/Ztn5YgfVvH
  https://cran.r-project.org/web/packages/lime/index.html
- ShapleyR https://t.co/pZLhbVVs5a
- Iml https://cran.r-project.org/web/packages/iml/vignettes/intro.html
- ICEbox https://github.com/kapelner/ICEbox
- live https://t.co/zaQnLBtbfO
- xgboostExplainer https://t.co/1wgpqD8HL4
- breakDown https://t.co/FmvLqJXsFO and https://pbiecek.github.io/breakDown/

Resources

- Book on ML interpretation
  https://christophm.github.io/interpretable-ml-book/
- Book: An Introduction to Machine Learning Interpretability
  https://www.safaribooksonline.com/library/view/an-introduction-to/9781492033158/
- Blogs
  https://ssearch.oreilly.com/?q=+interpretability
  https://lilianweng.github.io/lil-log/2017/08/01/how-to-explain-the-prediction-of-a-machine-learning-model.html
- Implementations (examples) with LIME:
  http://projects.rajivshah.com/inter/ReasonCode_NFL.html
  http://www.business-science.io/business/2017/09/18/hr_employee_attrition.html (LIME and H2O)

Acknowledgement

Agustin Calatroni,
Principal Statistical Scientist
Rho, Inc