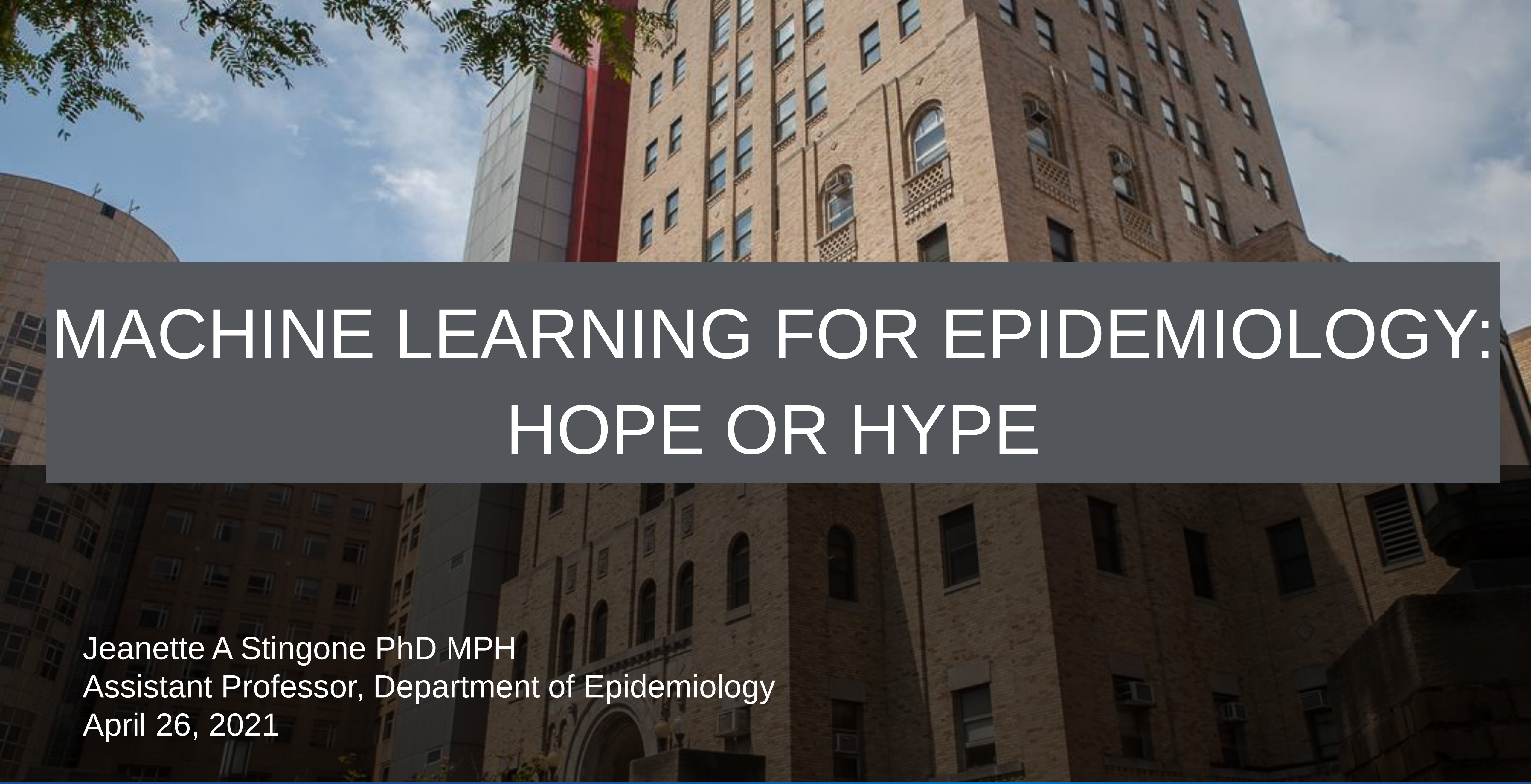


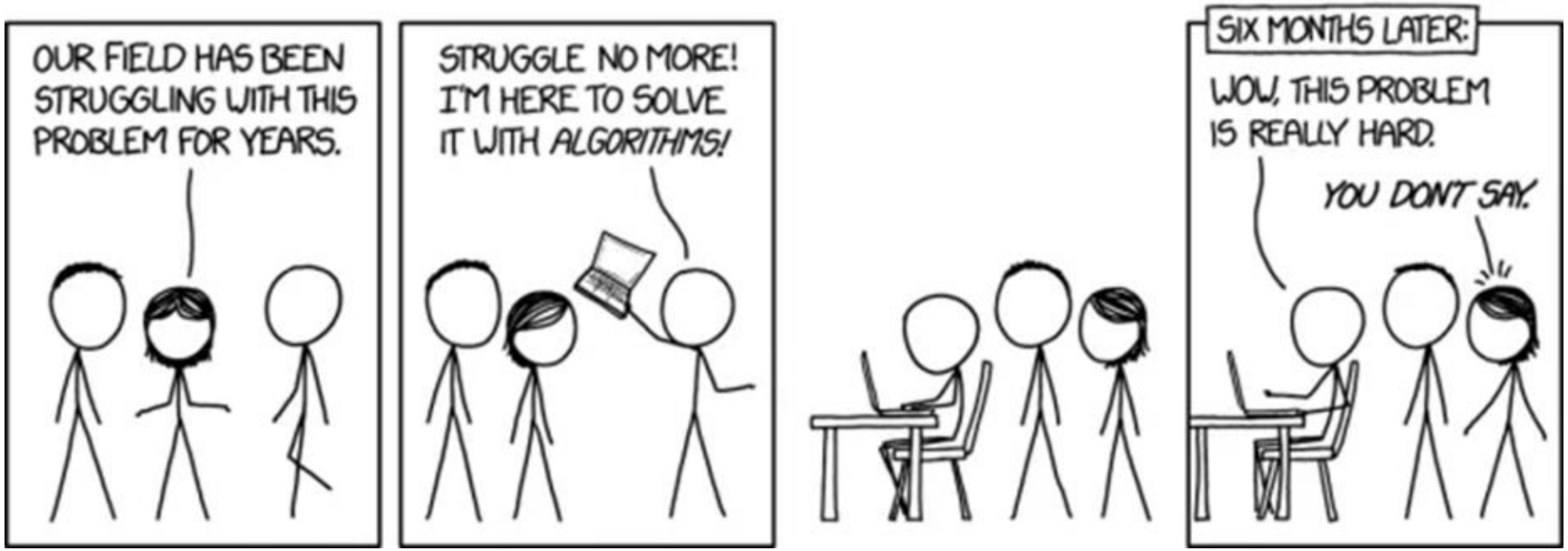
MACHINE LEARNING FOR EPIDEMIOLOGY: HOPE OR HYPE

Jeanette A Stingone PhD MPH
Assistant Professor, Department of Epidemiology

April 26, 2021

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

Machine Learning is not Magic


OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.
STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH *ALGORITHMS!*
SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.

Here to Help: https://xkcd.com/1831

The problem with the “magical” metaphor

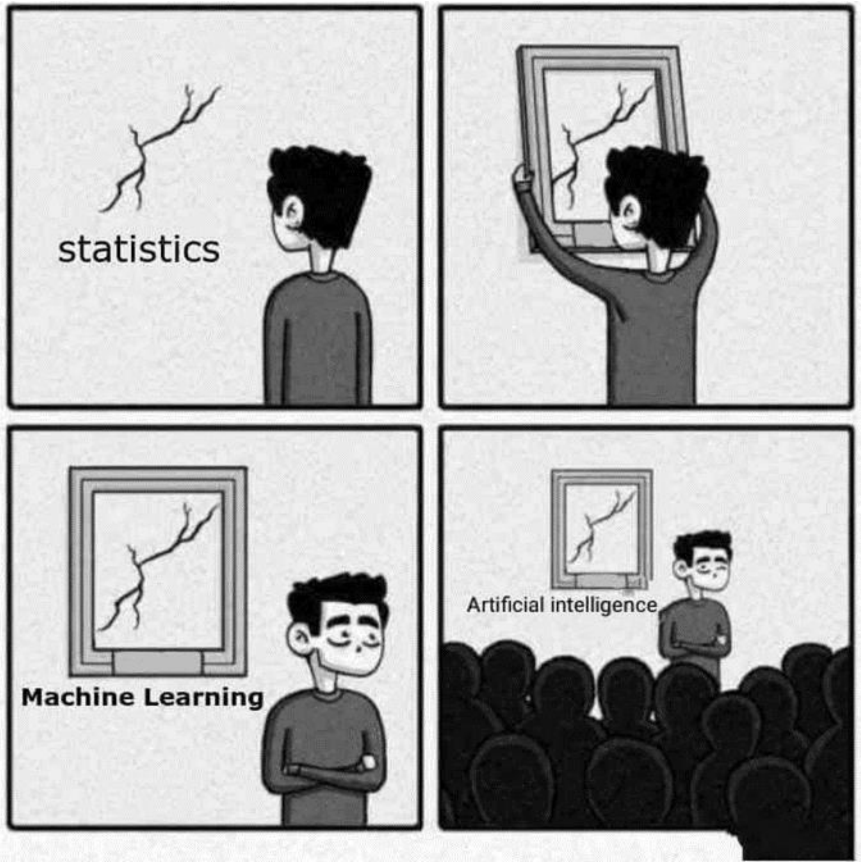
Why is Machine Learning So Magical?

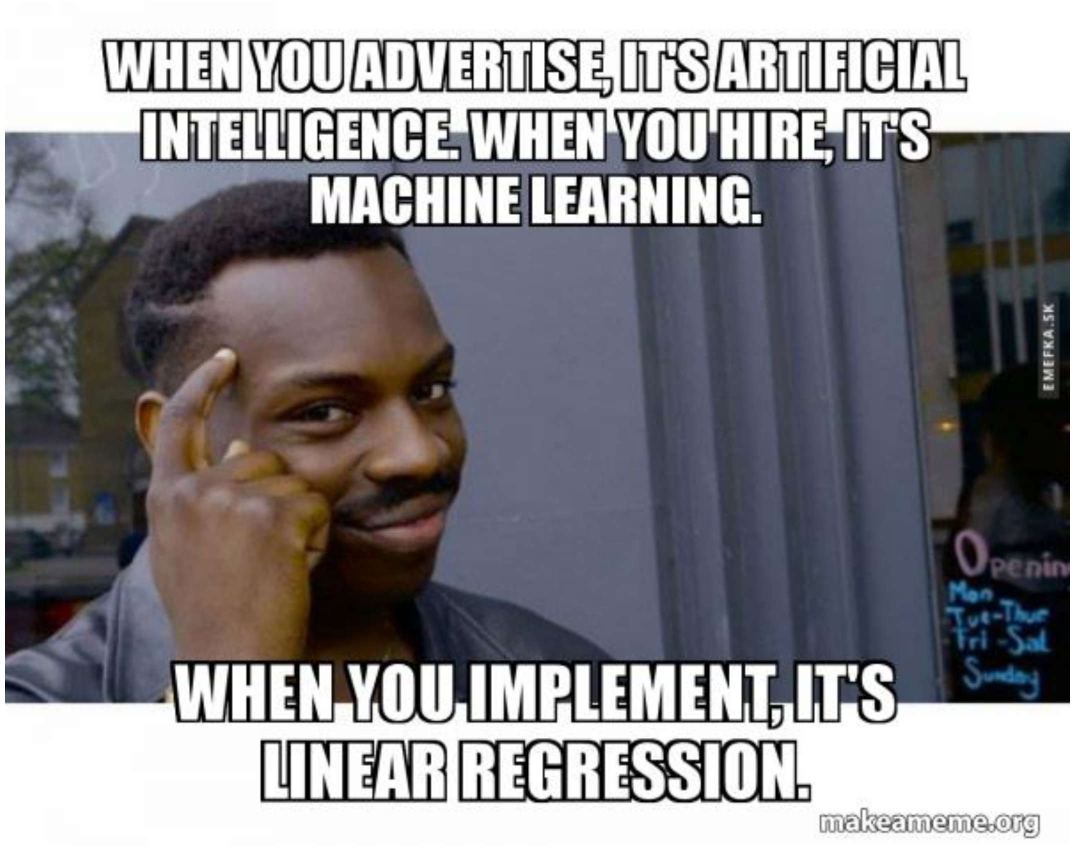
And the struggles of practising the tricks

Rick Vink Follow

Sep 6 · 4 min read ★

On the flip side, are some too cynical?




WHEN YOU ADVERTISE, IT'S ARTIFICIAL INTELLIGENCE. WHEN YOU HIRE, IT'S MACHINE LEARNING.
WHEN YOU IMPLEMENT, IT'S LINEAR REGRESSION.
makeameme.org

Epidemiologists use tools for different purposes

Questionnaire Development

CLINICAL TEST PROTOCOLS

Biological Assays

Propensity Scores

EXPOSURE MODELING

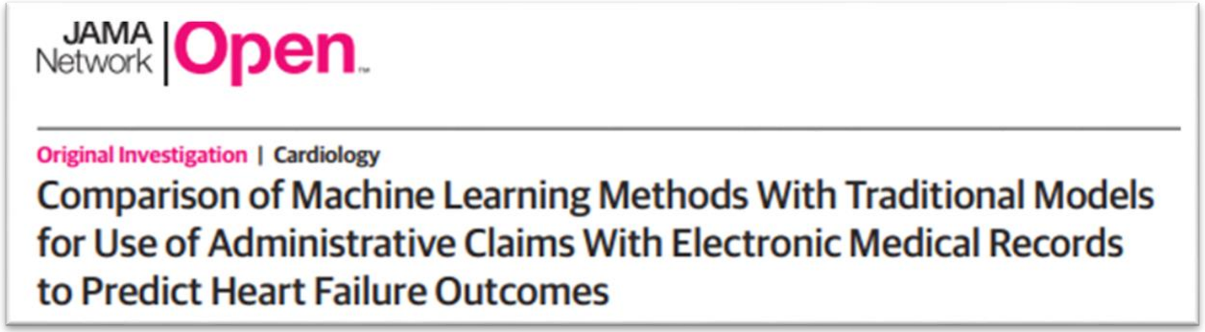
Community Engagement

Regression

AGENT-BASED MODELS

Machine Learning??

Expect Magic and You Will Be Disappointed….



JAMA Network | Open

Original Investigation | Cardiology

Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes

6113 obs in training
3389 obs for testing

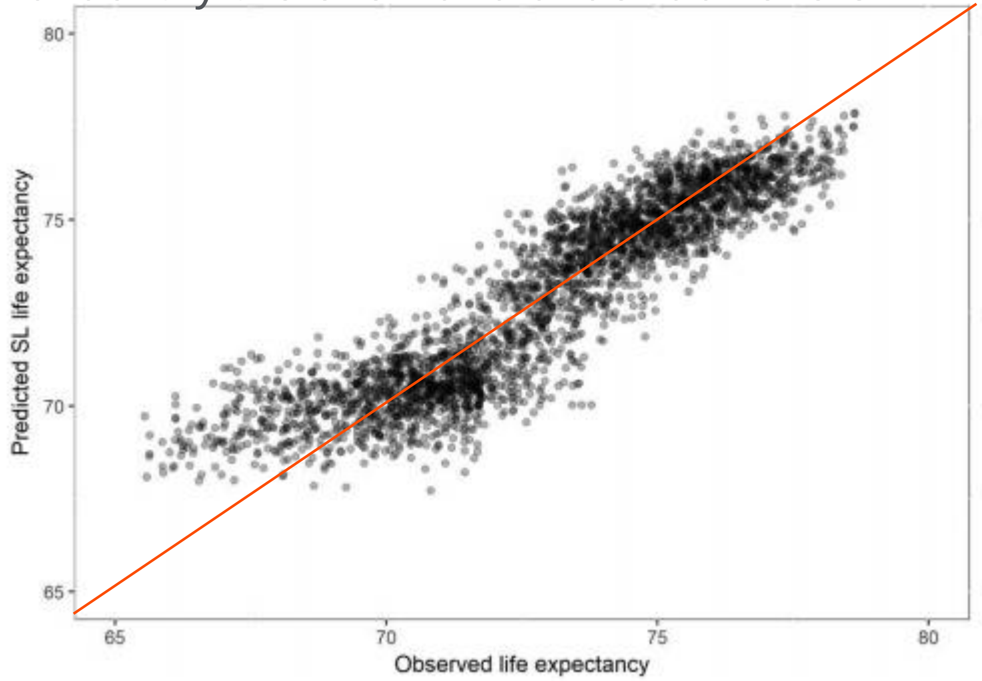
54 variables from Medicare claims
8 variables from EHR

> “In our study, we observed that when using only claims-based predictors, many of which are binary variables indicating presence or absence of medical conditions or use of specific medications, the performance improvement with machine learning approaches was minimal for prediction of most outcomes. However, when the predictor set was expanded to include EMR-based information, which included numerous laboratory test results as continuous variables, we noted that machine learning approaches generally fared better than logistic regression. This observation follows the intuition that, because tree-based machine learning approaches, such as GBM or random forests, are nonparametric and do not assume linearity for a predictor-outcome association, they are usually more adept at generating predictions based on continuous variables.”

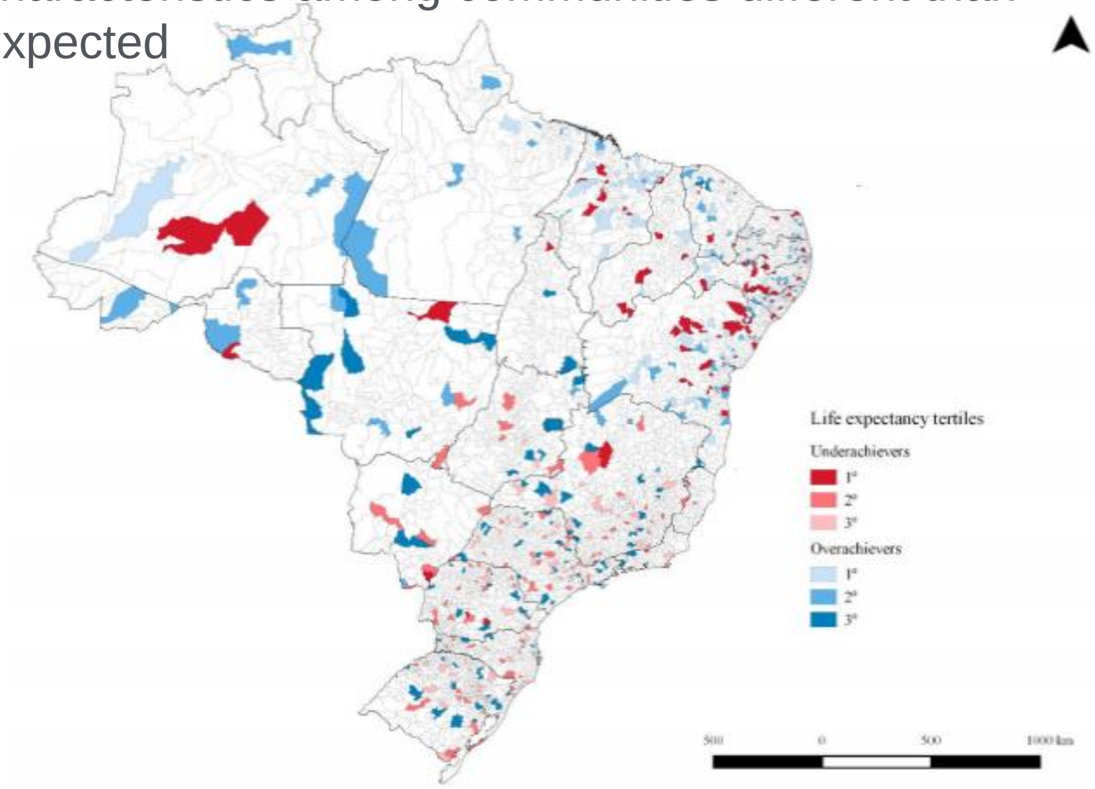
Machine Learning for Precision Public Health

“use better and more precise data to target disease prevention and control”-CDC

Step 1: Predict life expectancy based on local features and identify the over- and under-achievers



Step 2: Investigate modifiable public health characteristics among communities different than expected



Source: Chiavegatto Filho et al Epidemiology 2018; 29:836-840.

Predicting Values of Unmeasured Variables: Fetal Weight

Machine Learning for Novel Data Generation

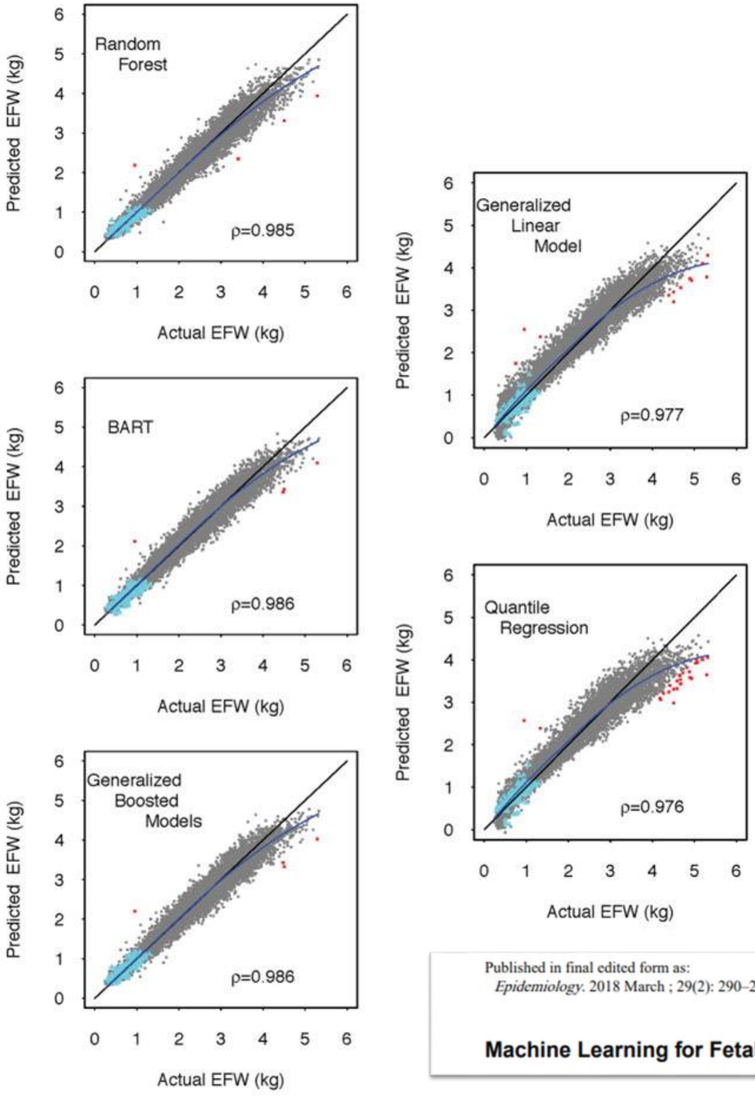


Image Analysis and Text Mining of Large Resources

Environment International
Volume 122, January 2019, Pages 3-10

ELSEVIER

A picture tells a thousand…exposures: Opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology

Scott Weichenthal [a], Marianne Hatzopoulou [b], Michael Brauer [c]

Published in final edited form as:
Epidemiology. 2018 March ; 29(2): 290–298. doi:10.1097/EDE.0000000000000788.

Machine Learning for Fetal Growth Prediction

Overview

- What is “machine learning”? Does it matter?
- How have these methods been traditionally used?
  - Prediction vs Explanation
- Types of Machine Learning and Analytic Pipelines
- Common Algorithms
- Overview of Caret Package in R

How “machine learning” is defined often depends on who you ask……

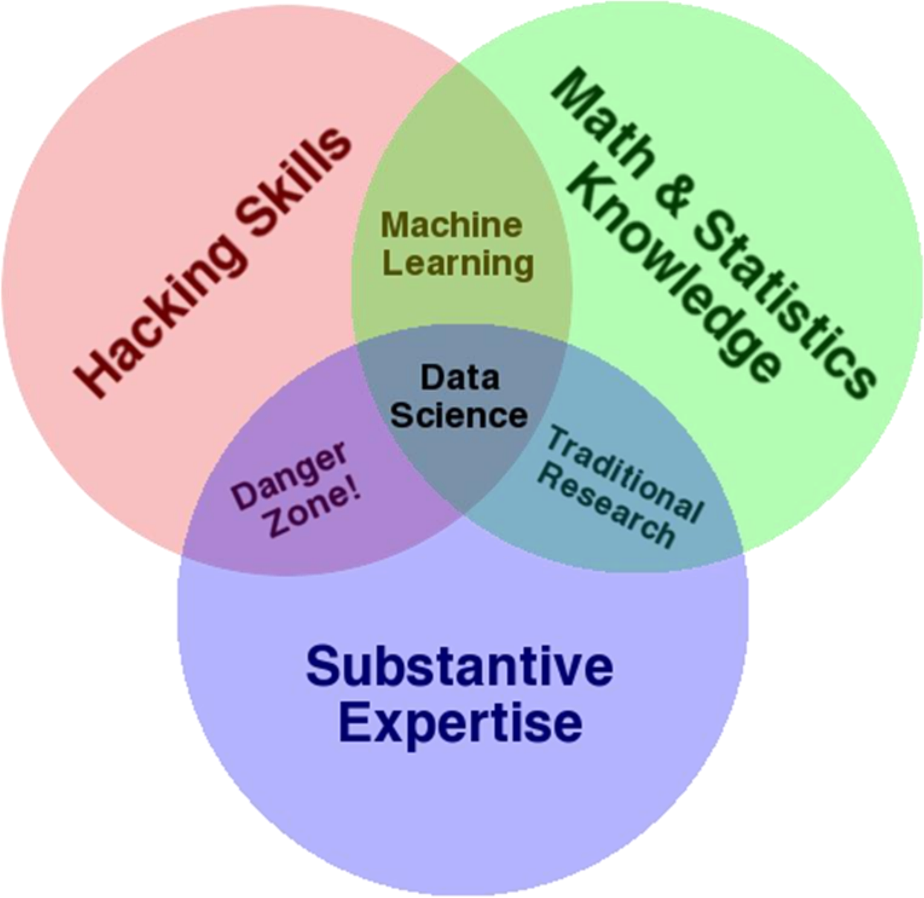
Computational methods **using experience to improve performance or to make accurate predictions**. Here, experience refers to the past information available to the learner, which typically takes the form of electronic data….In all cases, its quality and size are crucial to the success of the predictions made by the learner. -Foundations of Machine Learning Mohri, Rostamizadeh, Talwalkar, The MIT Press

A program or system that builds (trains) **a predictive model from input data**. The system uses the learned model to make useful predictions from new (never-before-seen) data drawn from the same distribution as the one used to train the model. Machine learning also refers to the field of study concerned with these programs or systems.
-Google

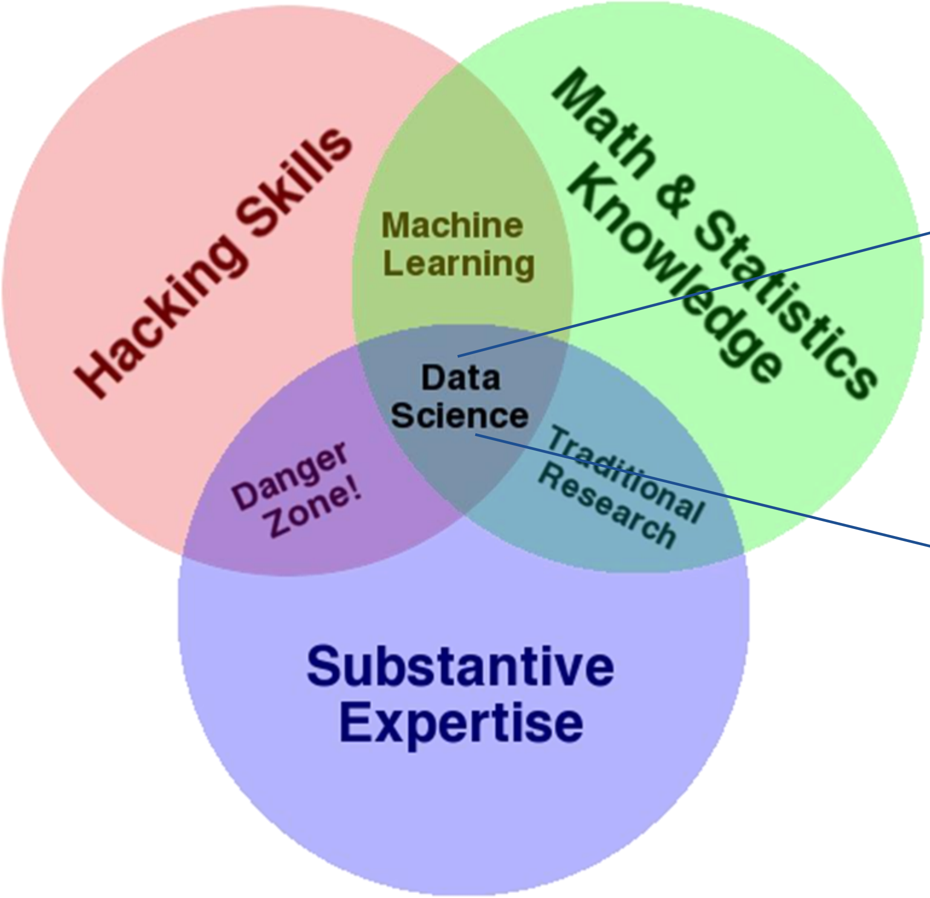
..an umbrella term for techniques that **fit models algorithmically by adapting to patterns in data**
-Mooney and Pejaver, Annual Review of Public Health

Scientific study of algorithms and statistical models that computer systems use to effectively **perform a specific task without using explicit instructions, relying on patterns and inference instead**. -Wikipedia

Machine Learning: Intersection between Computational and Mathematical/Statistical Knowledge



Machine Learning: Intersection between Computational and Mathematical/Statistical Knowledge



Machine learning combined with substantive expertise can allow us to advance our research and gain more from our scientific data

The utility of machine learning (or any complex method) will depend upon our research question and the structure/content of our data

To Explain or To Predict: What is the question…and what is the difference?

Explanatory Modeling: use of statistical models to test (or estimate) hypothesized causal associations; requires pre-existing causal model

Predictive Modeling: use of data to develop model that can predict new or future observations

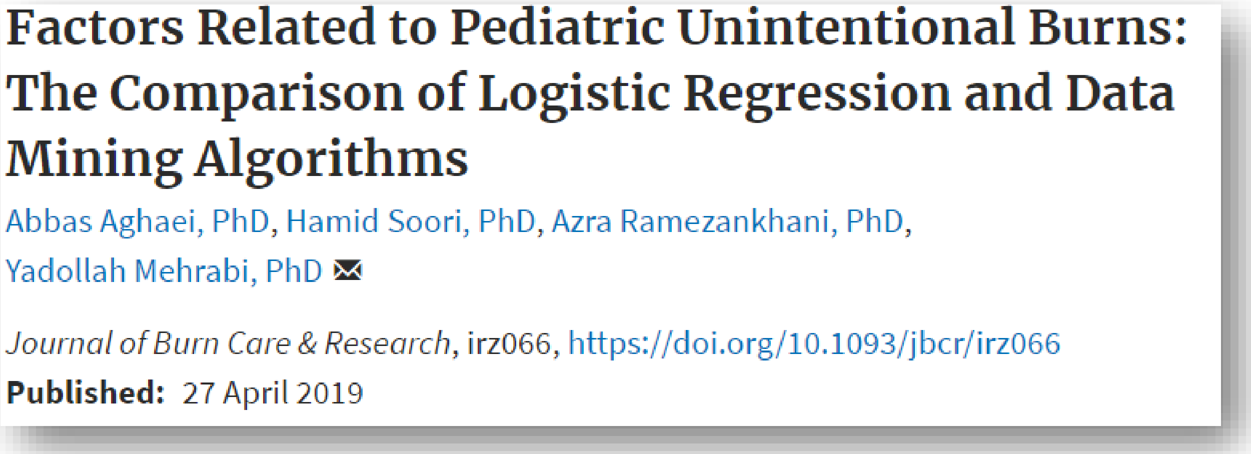
Machine learning approaches traditionally used **AND** developed for prediction goals.

- ➢ Note there are questions of prediction within explanatory modelling
  - ➢ construction of propensity scores
  - ➢ use of risk scores to account for confounding
  - ➢ predicting the counterfactual
- ➢ If goal is not prediction, do we need to adapt machine learning approaches for our goal?

But what if my goal is explanation, but I don’t have a good pre-existing causal model…..

- ➢ “By capturing underlying complex patterns and relationships, predictive modeling can suggest improvements to existing explanatory models” ---Shmueli 2010

Identifying “Predictors” using machine learning



Factors Related to Pediatric Unintentional Burns: The Comparison of Logistic Regression and Data Mining Algorithms

Abbas Aghaei, PhD, Hamid Soori, PhD, Azra Ramezankhani, PhD, Yadollah Mehrabi, PhD ✉

Journal of Burn Care & Research, irz066, https://doi.org/10.1093/jbcr/irz066

Published: 27 April 2019

Excerpt from the Abstract: The majority of the burn-related variables were related to individuals’ social welfare status and their environments. Lessening the effects of these factors could reduce the incidence of pediatric burns.

…Reliant on the assumption that a good predictor is a good explainer…..

How can Epidemiologists Benefit from Training in Machine Learning?

- Facilitate use of large and/or complex data where relationships cannot be easily visualized
  - Use of ML approaches can identify patterns in data; potentially generate hypotheses, refine metrics of exposure and/or outcome

- Make exploratory data analysis and model selection more formal
  - Similar to use of DAGs to explicitly represent assumptions of relationships between variables
  - Don’t just publish the final model, show how you arrived there.

- Greater consideration of questions of prediction and how they can benefit public health

Prediction in Epidemiology & Public Health

Interested in predicting some exposure/outcome or predictions as intermediate product toward another aim

- Disease/Outcome Forecasting
  - Predicting peak demand days for healthcare services associated with specific events

- Creation of Disease/Outcome Risk Scores
  - Triage in the ED for who needs supplemental treatments

- Construction of Propensity Scores/Inverse Probability of Treatment Weights

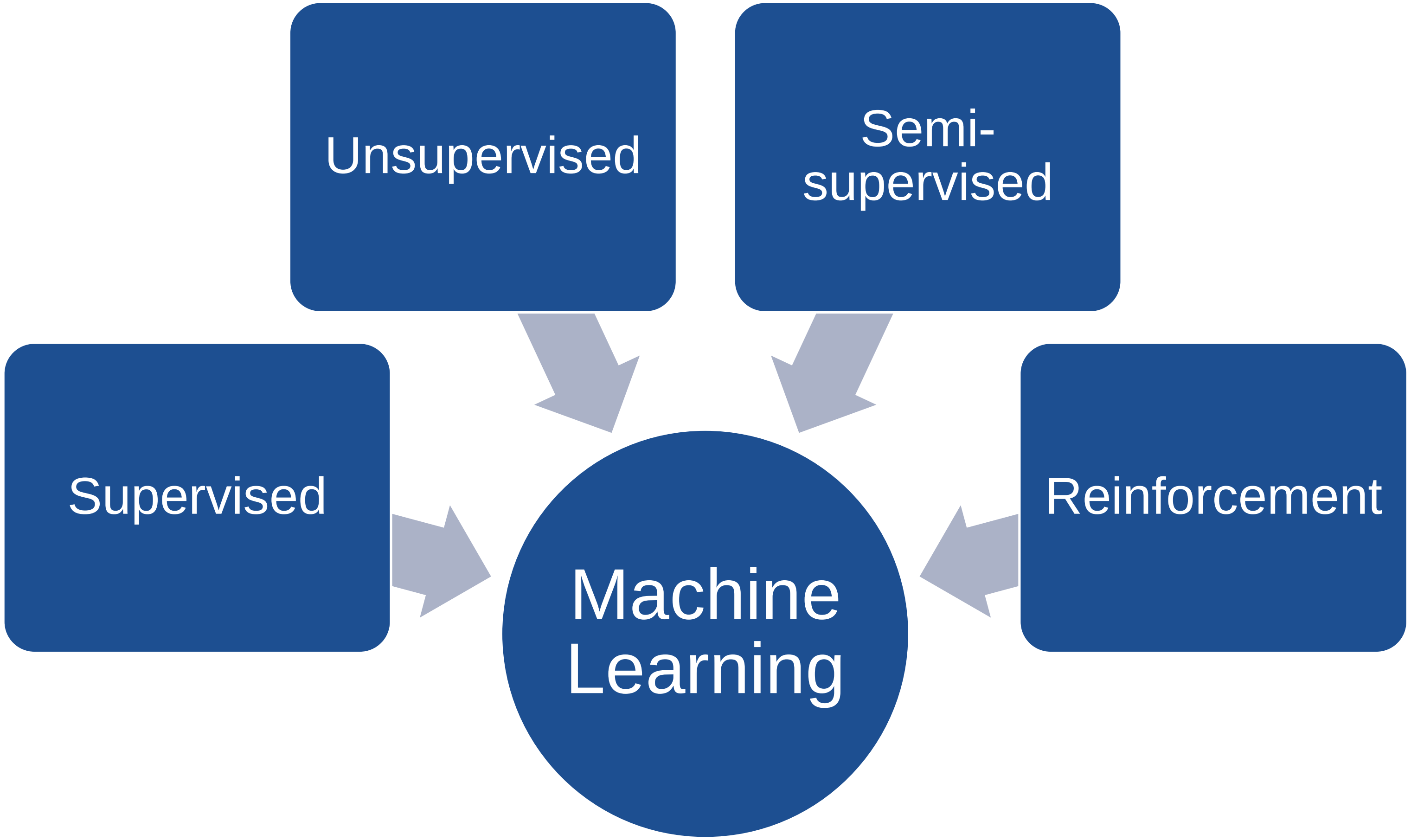
- Imputation of Missing Data; Construction of Synthetic Data

- Predicting toxicity of chemicals based on structure

- Exposure Modelling
  - Image analysis to predict exposures
    - Can calculating traffic density allow you to estimate PM emissions in areas with little air monitoring?

What are the different types of machine learning?

Types of Machine Learning



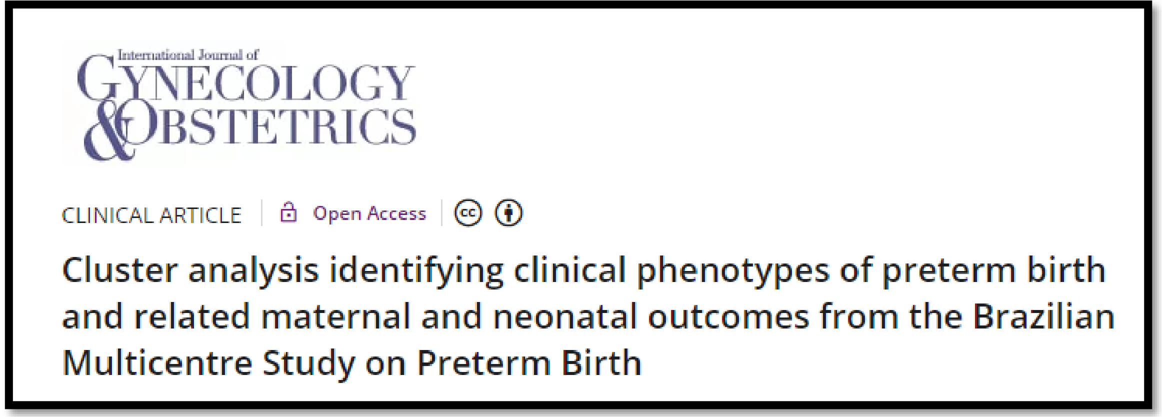
Unsupervised

Context: for each observation of the inputs (predictor/exposure/independent variables), there is no associated output (response measurement); also described as data are “unlabeled”

Algorithm identifies patterns within the vector of inputs and generates an output that seeks to understand or represent the relationships between variables and/or observations.

Addresses: Clustering and Dimension Reduction Problems

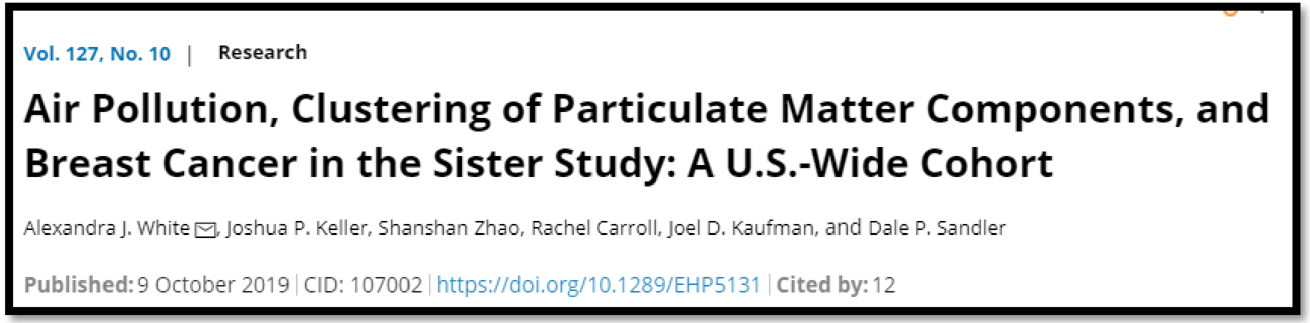
Clustering to refine the outcome classification


International Journal of GYNECOLOGY & OBSTETRICS

CLINICAL ARTICLE | Open Access

Cluster analysis identifying clinical phenotypes of preterm birth and related maternal and neonatal outcomes from the Brazilian Multicentre Study on Preterm Birth

Clustering for Exposure Assessment


Vol. 127, No. 10 | Research

Air Pollution, Clustering of Particulate Matter Components, and Breast Cancer in the Sister Study: A U.S.-Wide Cohort

Alexandra J. White, Joshua P. Keller, Shanshan Zhao, Rachel Carroll, Joel D. Kaufman, and Dale P. Sandler

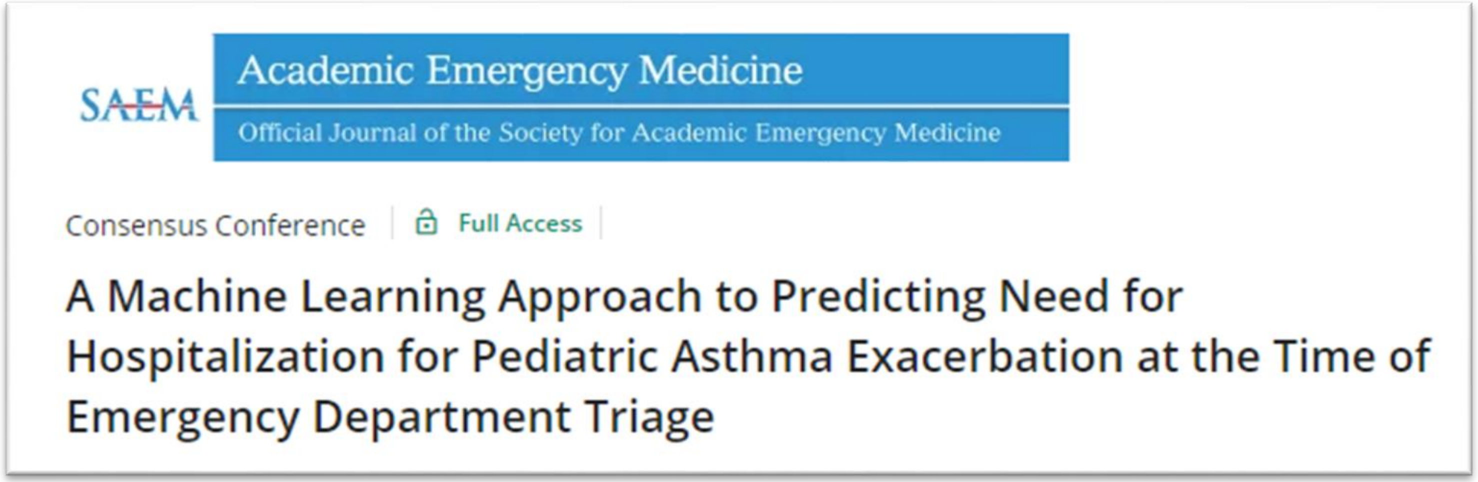
Published: 9 October 2019 | CID: 107002 | https://doi.org/10.1289/EHP5131 | Cited by: 12

Supervised

Context: for each observation of the inputs (predictor/exposure/independent variables), there is an associated output (response measurement); also described as data are “labeled”

Algorithm learns how to use inputs to generate outputs through training and receives feedback by looking at actual outcomes; process is “supervised”

Addresses: Regression, Classification and Estimation Problems



SAEM

Academic Emergency Medicine

Official Journal of the Society for Academic Emergency Medicine

Consensus Conference | Full Access

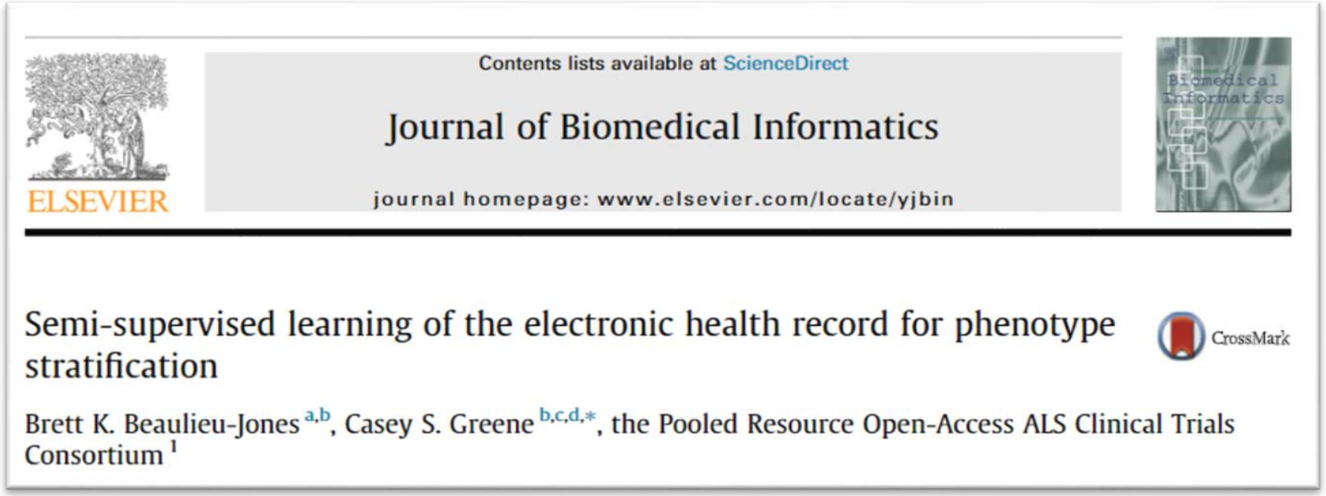
A Machine Learning Approach to Predicting Need for Hospitalization for Pediatric Asthma Exacerbation at the Time of Emergency Department Triage

Semi-supervised

Context: for each observation of the inputs (predictor/exposure/independent variables), only a (typically) small subset has an associated output (response variable); combination of labeled and unlabeled data

Algorithm learns from subset of labeled data and then applies what it has learned to the unlabeled data.

Addresses: Useful when it is overly human or resource intensive to obtain sufficient amounts of labeled data to train a supervised algorithm



Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

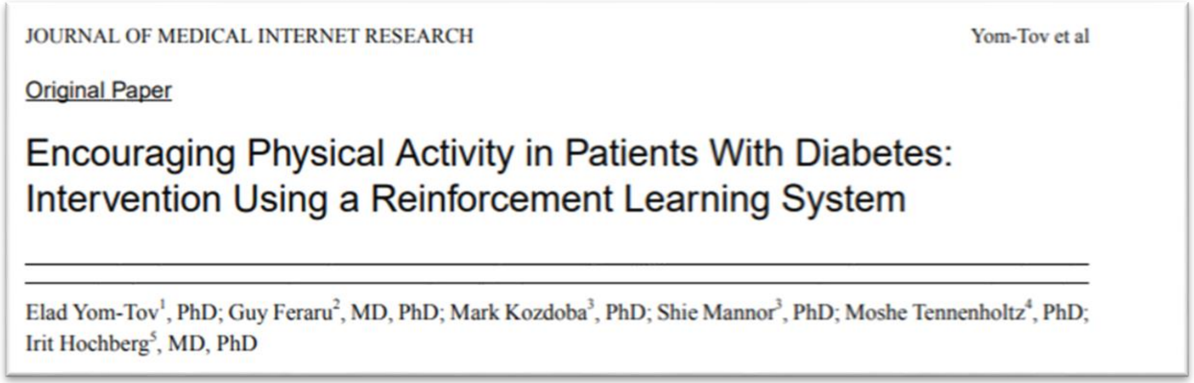
ELSEVIER

Semi-supervised learning of the electronic health record for phenotype stratification

Brett K. Beaulieu-Jones [a,b], Casey S. Greene [b,c,d,*], the Pooled Resource Open-Access ALS Clinical Trials Consortium [1]

Reinforcement

Algorithm learns how to act in a given environment through maximization of reward; needs to anticipate future rewards from short-term actions

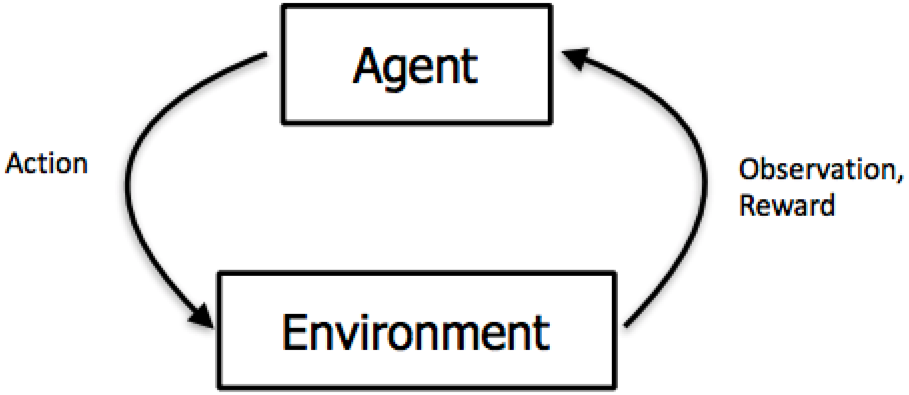


JOURNAL OF MEDICAL INTERNET RESEARCH Yom-Tov et al

Original Paper

Encouraging Physical Activity in Patients With Diabetes: Intervention Using a Reinforcement Learning System

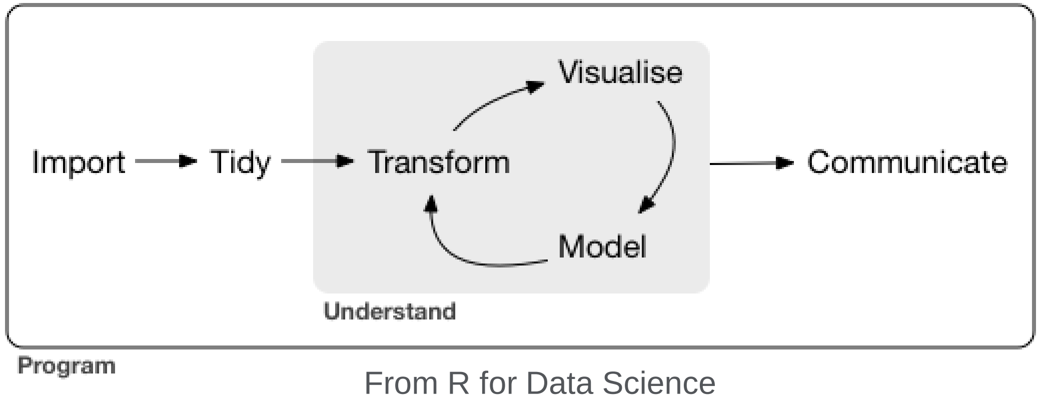
Elad Yom-Tov[1], PhD; Guy Feraru[2], MD, PhD; Mark Kozdoba[3], PhD; Shie Mannor[3], PhD; Moshe Tennenholtz[4], PhD; Irit Hochberg[5], MD, PhD



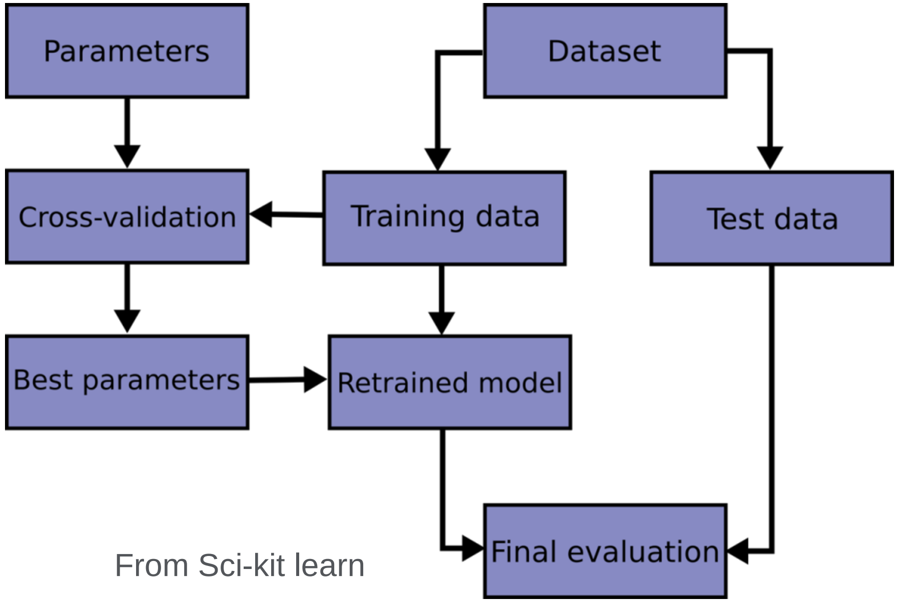
Pipelines

Ordered set of tasks to accomplish a specific task or goal; Visual study protocol

Specific definition varies by field and their perspective on data analytics

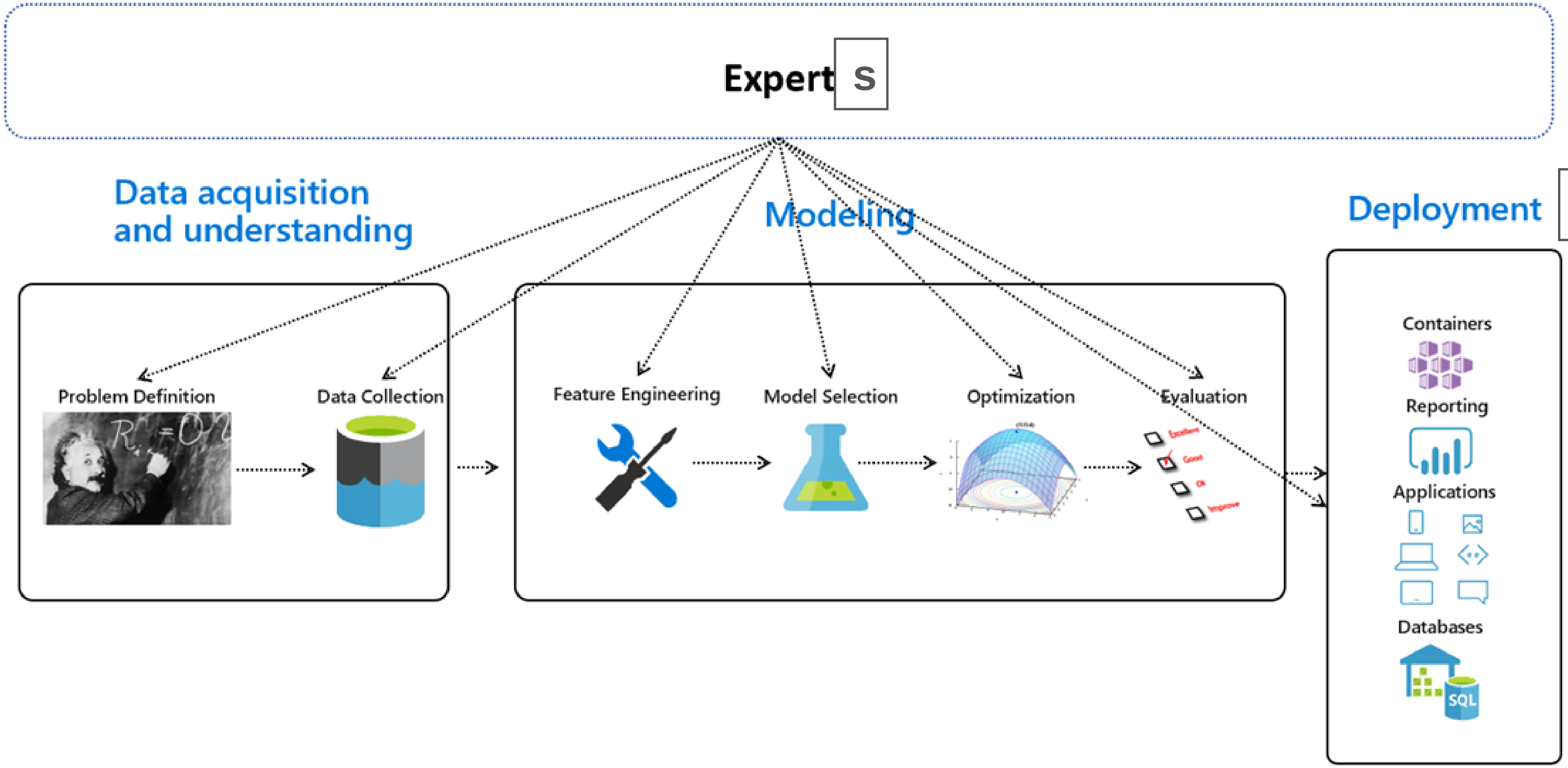


From R for Data Science
https://r4ds.had.co.nz/



From Sci-kit learn

Pipeline of entire Data LifeCycle



s

Or Dissemination

Source: Medium.com

Commonly used Terms in Machine Learning pipelines

Small n, large-p vs Small p, large-n problem

- n-number of individuals in dataset, p-number of features for each individual
- Refers to shape of dataset (wide vs long) with each having specific set of challenges

Parameters

- a variable, internal to the model, and derived from the data; often saved as part of final model
- Example: β in a regression model

Hyperparameters

- a variable, external to the model and often set by the programmer/analyst; used to estimate model parameters or to optimize the algorithm; can also be called tuning parameter
- Example: number of trees in a random forest

Tuning

- Customization of a model by varying the hyperparameters to determine the values that provide the optimal performance

Evaluation: Focus on Prediction

Classification

- Accuracy
- Confusion Matrix
- ROC Curves and Area Under the Curve
- Calibration

Regression

- Mean Square Error
- R-squared

How we Evaluate (again focus on prediction): Differentiating Training, Validation and Testing

Data Partitioning

- ○ Splitting a dataset into random subsets for use in either training, validating or testing the machine learning model
- ○ Use of different subsets
  - ○ Training: used by algorithm to learn the resulting model
  - ○ Validation: used to compare performance of models produced by different algorithms, hyperparameters, …
  - ○ Test/Hold Out: used to obtain final metrics of performance and results of the model

Sample size typically dictates how data are partitioned.
More data used for training than testing in the context of prediction.
Also includes creation of K-folds for cross-validation. Folds are equal sized.

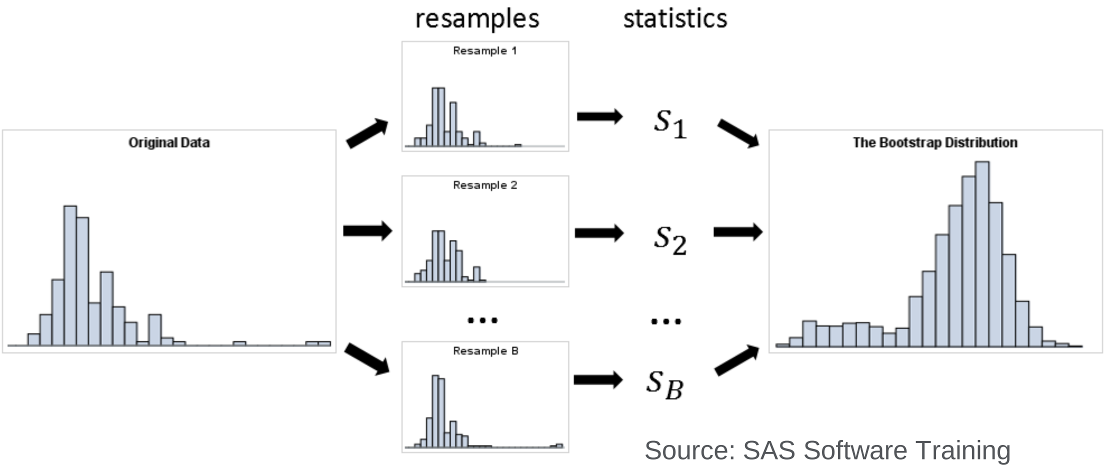
Resampling Methods

Bootstrapping

- Iteratively sampling with replacement
- Used to estimate parameters and draw inferences on a population
- Used in ensemble methods e.g. bagging

Cross-validation

- Validation technique
- Partition data into k non-overlapping subsets
- Estimate model parameters on k-1 subsets (training) then apply model in the held-out subset for evaluation metrics
- Repeat k times
- Similar Approach for Leave-one-out Cross-Validation



Source: SAS Software Training

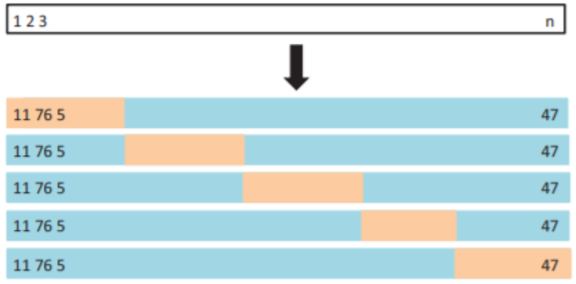
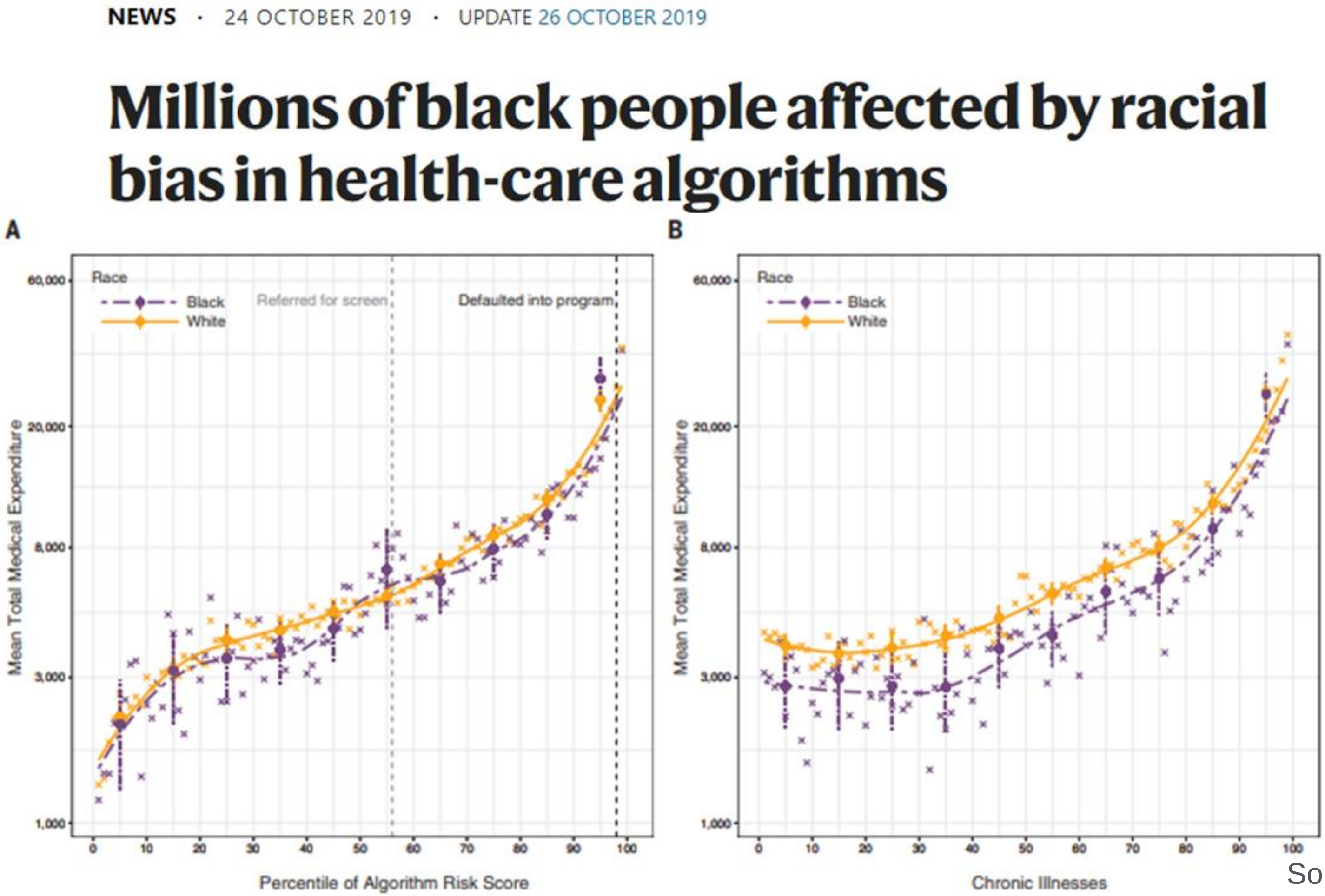


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

Source: Introduction to Statistical Learning in R

Evaluating in the Real World: Do models behave as expected?



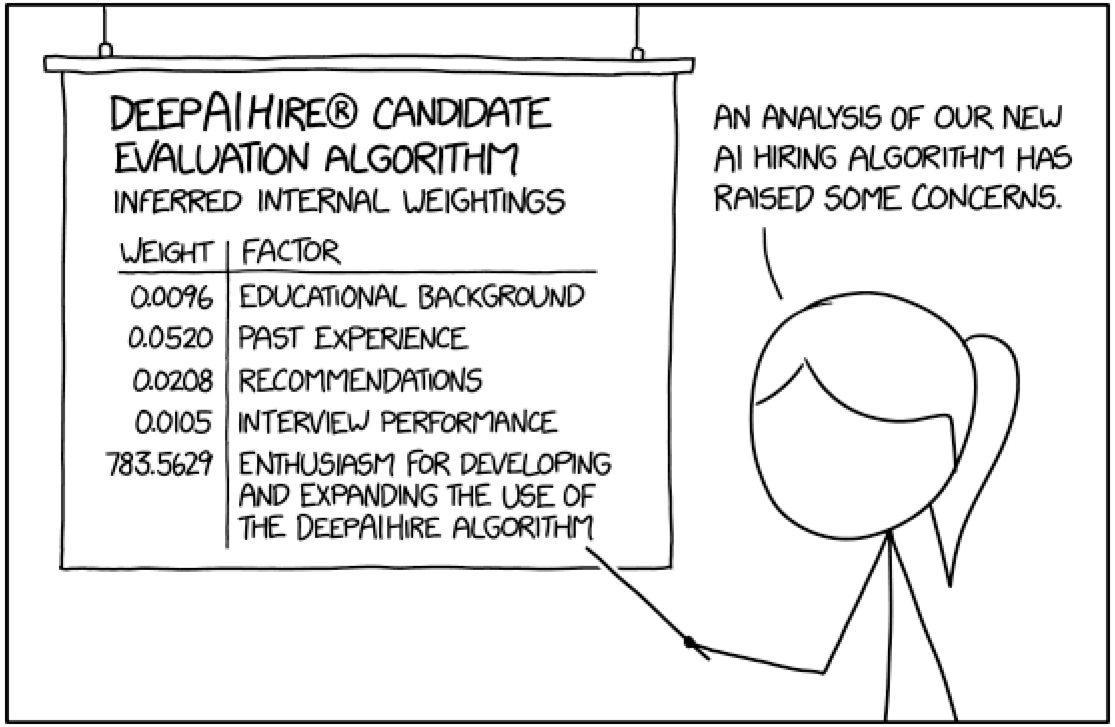
NEWS · 24 OCTOBER 2019 · UPDATE 26 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

Source: Obermeyer et al Science 366:447-453.

Know the Data that Drives Your Predictions

Are there biases in our past actions?



DEEPAIHIRE® CANDIDATE EVALUATION ALGORITHM

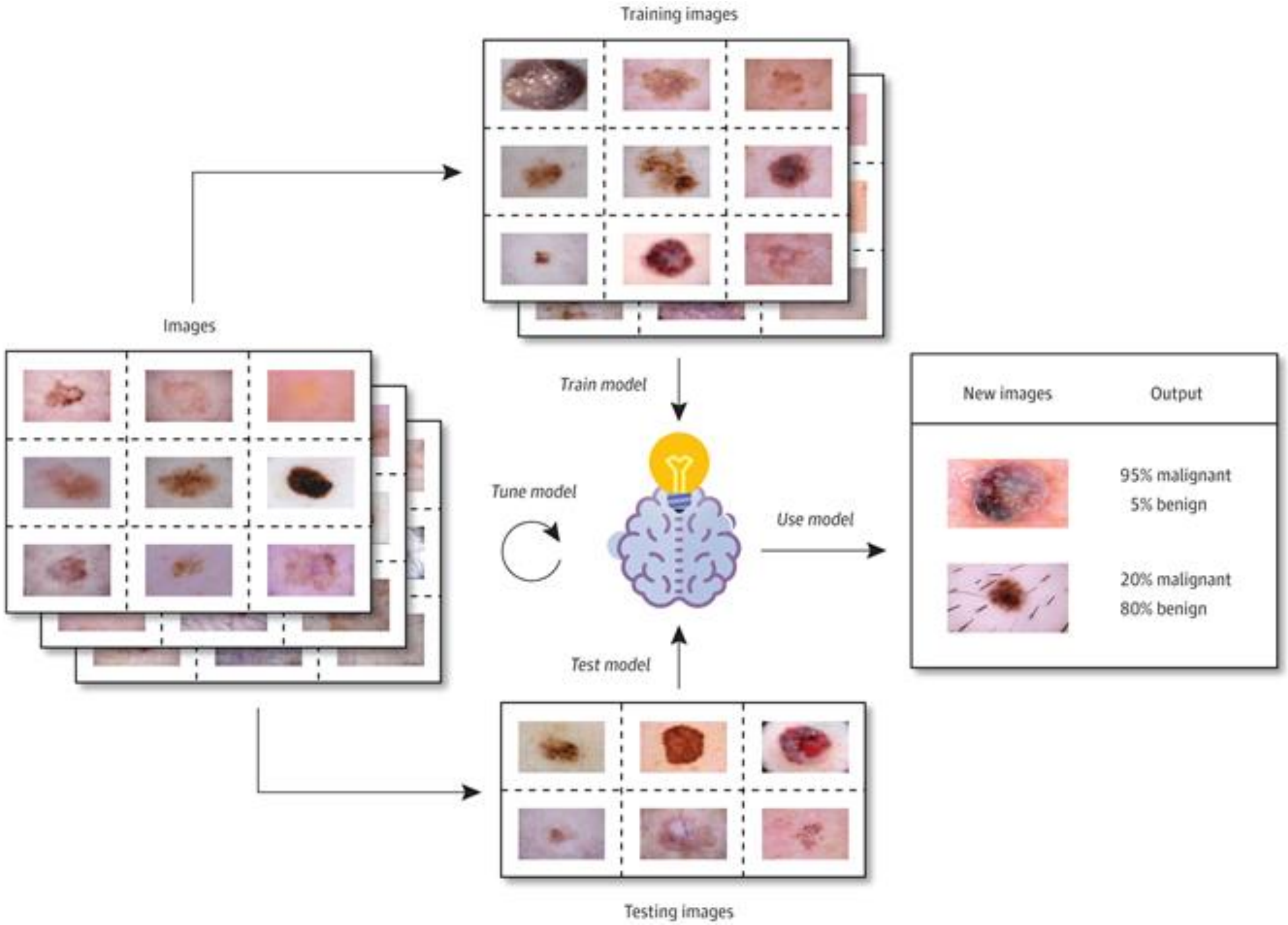
INFERRED INTERNAL WEIGHTINGS

WEIGHT	FACTOR
0.0096	EDUCATIONAL BACKGROUND
0.0520	PAST EXPERIENCE
0.0208	RECOMMENDATIONS
0.0105	INTERVIEW PERFORMANCE
783.5629	ENTHUSIASM FOR DEVELOPING AND EXPANDING THE USE OF THE DEEPAIHIRE ALGORITHM

AN ANALYSIS OF OUR NEW AI HIRING ALGORITHM HAS RAISED SOME CONCERNS.

Source: XKCD https://xkcd.com/2237/

Bias doesn't have to be explicitly programmed



From: **Machine Learning and Health Care Disparities in Dermatology**
JAMA Dermatol. 2018;154(11):1247-1248. doi:10.1001/jamadermatol.2018.2348

Commonly-used Algorithms

Unsupervised

K-Means*

Hierarchical clustering*

PCA

Self Organizing Maps

Gaussian Mixture Models

Supervised

Support Vector Machines

Naïve Bayes

K-Nearest Neighbors*

Regularized Regression*

Decision Trees*

Neural Networks

Ensemble*

Bagging

- Random Forest

Boosting

- XGBoost

Stacking

- SuperLearner

*To be reviewed in more detail in later slides

Clustering

Broad set of techniques for finding subgroups or clusters in a dataset-ISLR

Difference between clusters>>> difference within clusters

All observations are forced into a cluster

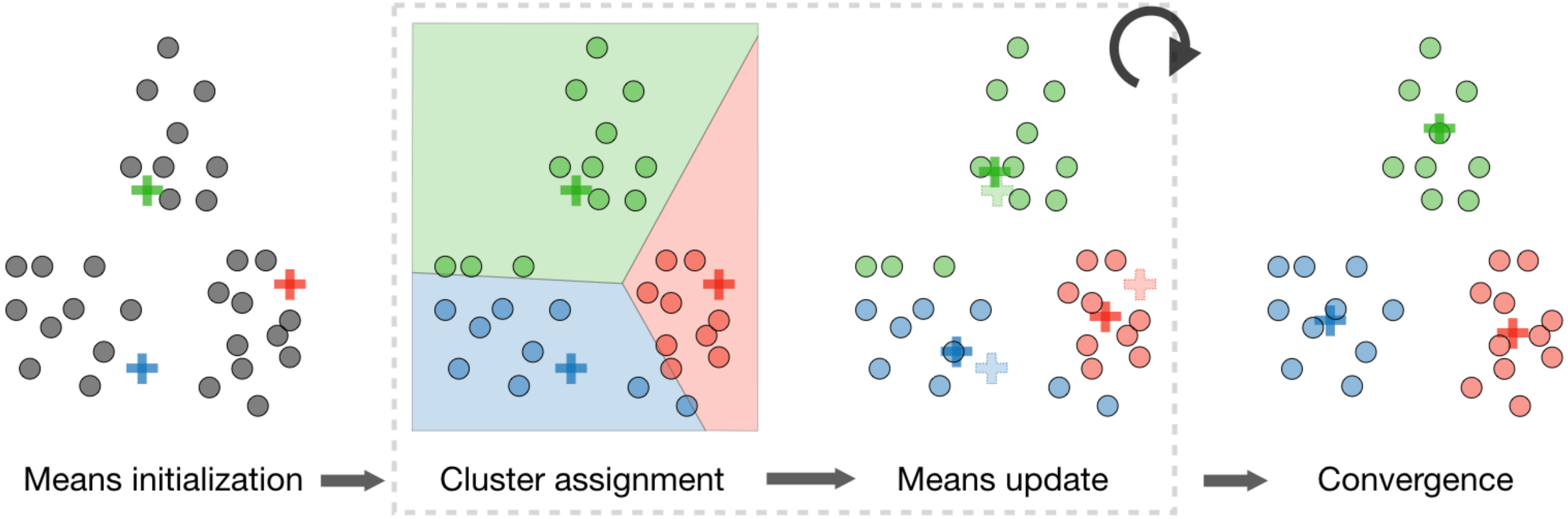
Two common algorithms: K-means and Hierarchical, but also k-prototypes, PAM, etc.

Applications

Identify clusters of observations based on features: phenotypic subgroup identification

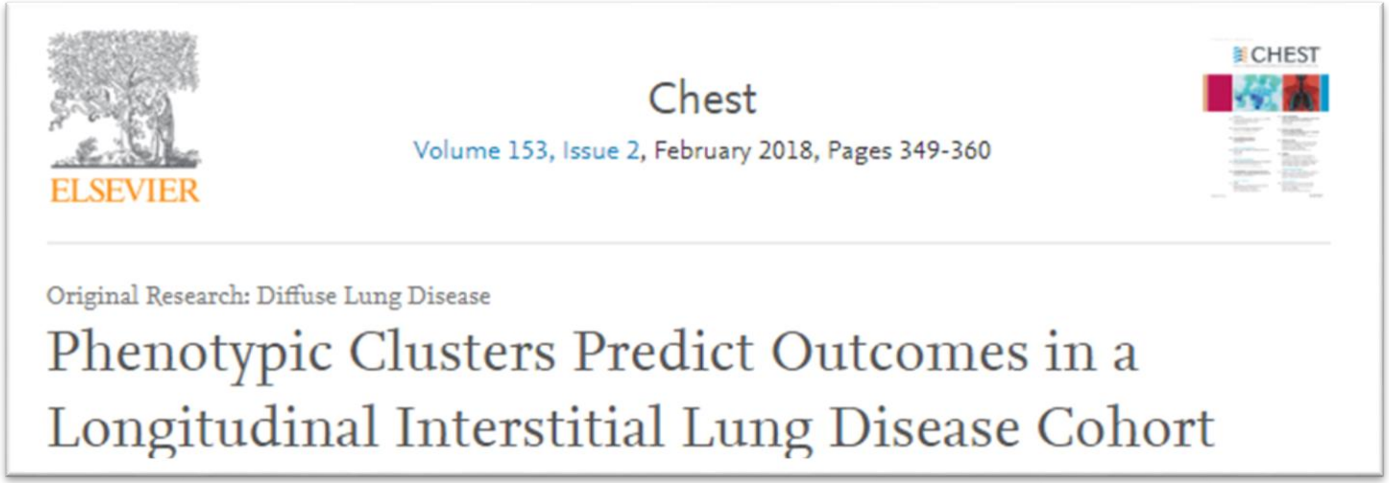
Identify clusters of features based on observations: identifying genetic expression patterns

K-means clustering: user must specify k



Source: CS-229 Course Material

Application: Identify Phenotypic Subtypes



ELSEVIER

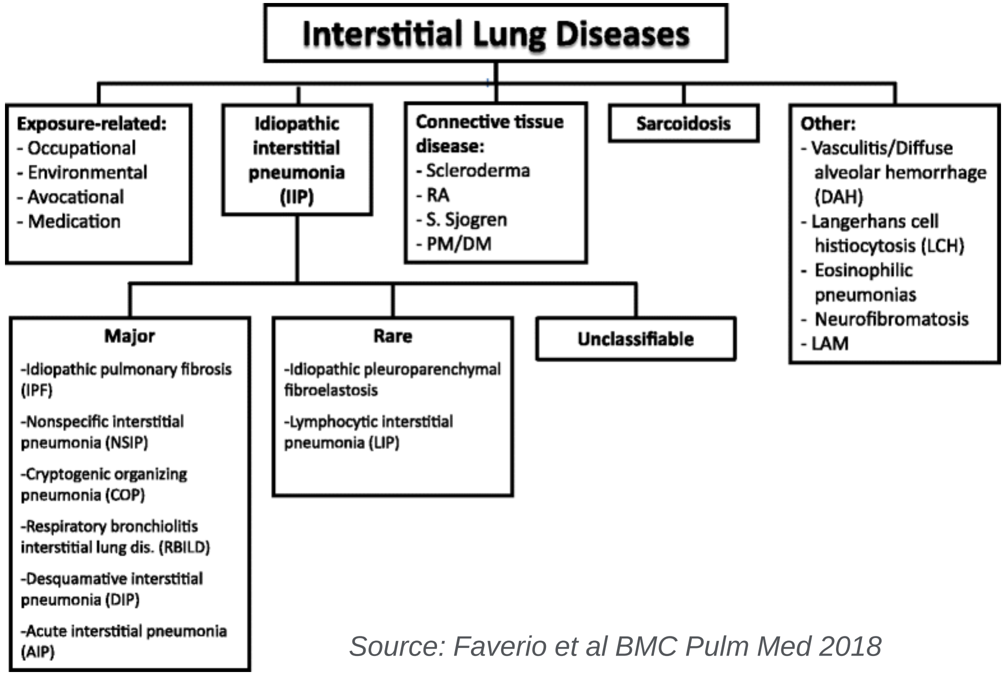
Chest

Volume 153, Issue 2, February 2018, Pages 349-360

Original Research: Diffuse Lung Disease

Phenotypic Clusters Predict Outcomes in a Longitudinal Interstitial Lung Disease Cohort

Goal: “Identify distinct clinical phenotypes in heterogeneous diseases”



Interstitial Lung Diseases

Exposure-related:
- Occupational
- Environmental
- Avocational
- Medication

Idiopathic interstitial pneumonia (IIP)

Major

-Idiopathic pulmonary fibrosis (IPF)

-Nonspecific interstitial pneumonia (NSIP)

-Cryptogenic organizing pneumonia (COP)

-Respiratory bronchiolitis interstitial lung dis. (RBILD)

-Desquamative interstitial pneumonia (DIP)

-Acute interstitial pneumonia (AIP)

Rare

-Idiopathic pleuroparenchymal fibroelastosis

-Lymphocytic interstitial pneumonia (LIP)

Unclassifiable

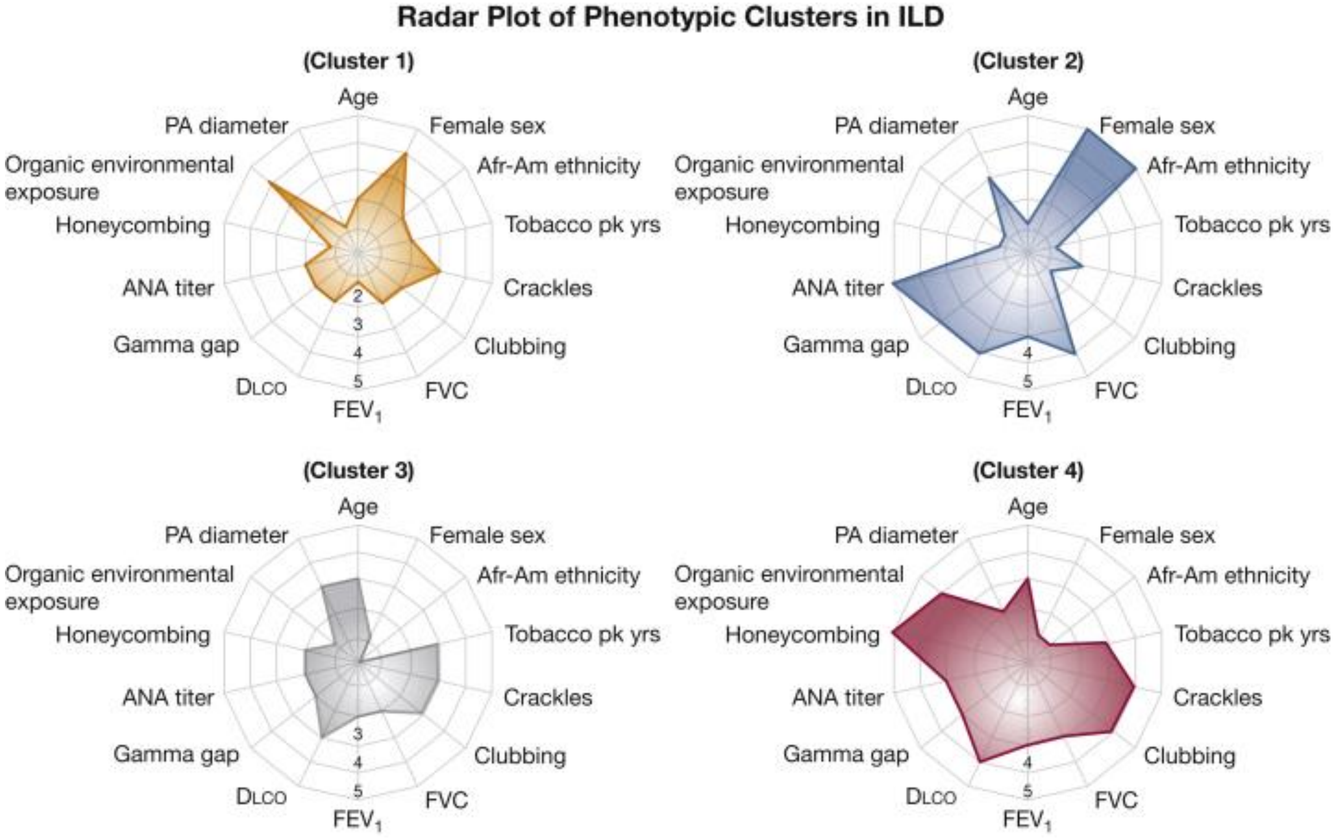
Connective tissue disease:
- Scleroderma
- RA
- S. Sjogren
- PM/DM

Sarcoidosis

Other:
- Vasculitis/Diffuse alveolar hemorrhage (DAH)
- Langerhans cell histiocytosis (LCH)
- Eosinophilic pneumonias
- Neurofibromatosis
- LAM

Source: Faverio et al BMC Pulm Med 2018

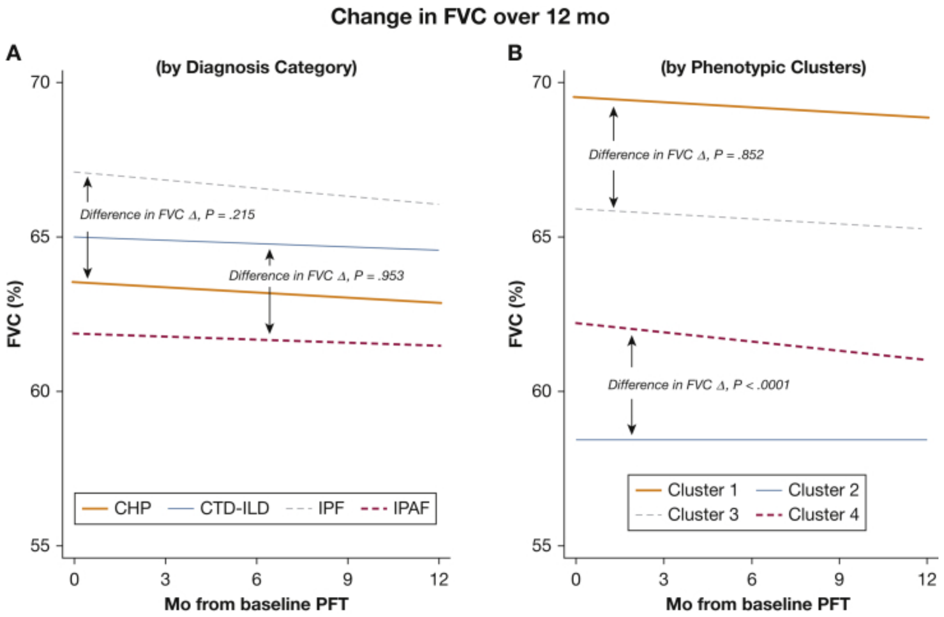
Method: Partitioning around Medoids (PAM)



Why PAM?

Similar to K-means, but more robust to outliers

Relies on median distances rather than means



Source: Adegunsoye et al Chest 2018

Hierarchical Clustering

Builds nested clusters in a successive manner

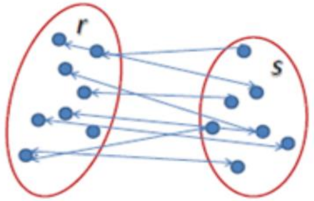
Agglomerative: each observation starts in its own cluster, and pairs of clusters are merged successively; better in discovering small clusters

Divisive: all observations start in the same cluster and splits are performed recursively; better in discovering large clusters

Merges and splits decided by cluster **dissimilarity.** Dissimilarity computed by **distance** and **linkage**.

Multiple distance metrics for calculating dissimilarity

- Euclidian
- Squared Euclidian
- Manhattan
- Maximum
- Mahalanobis

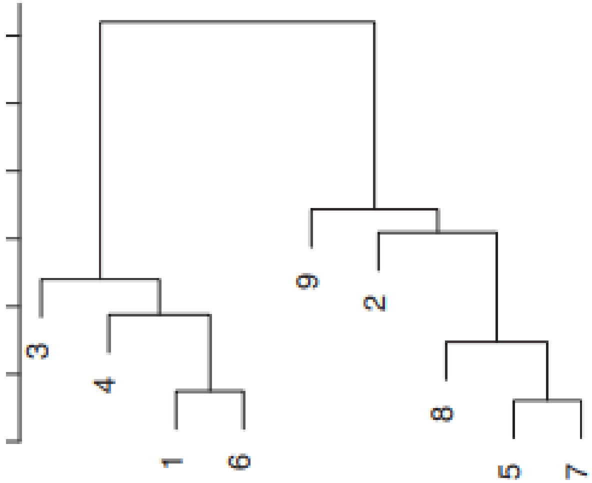


Multiple criterion for linkage

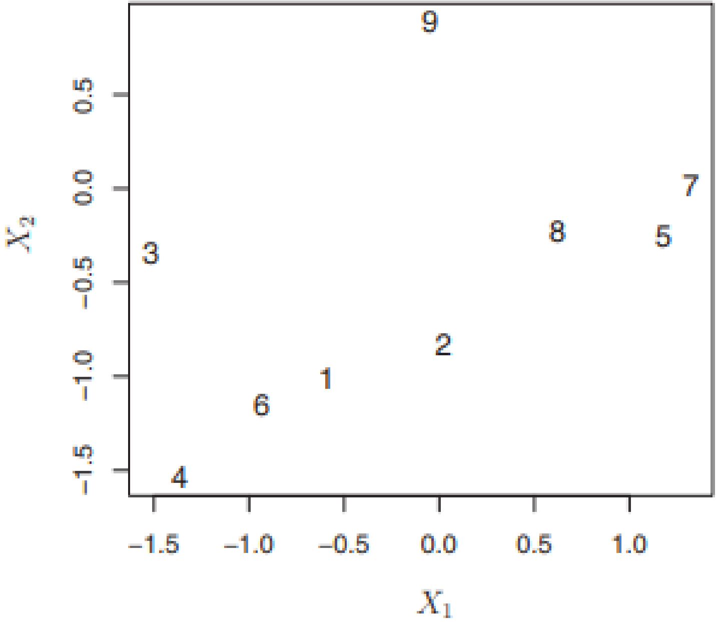
- Complete
- Single
- Average

Dendrogram: visualization of hierarchical clustering

Dendrogram after clustering



Raw Data



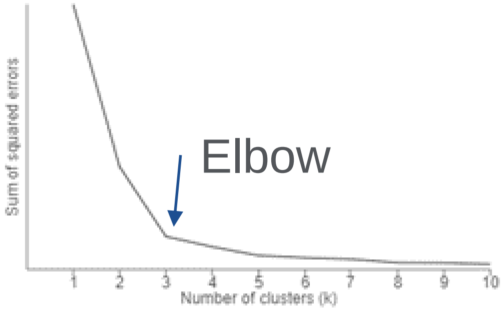
Source ISLR

Again, need to determine optimal number of clusters

Evaluating Number of Clusters

Goal: minimize the intra-cluster variation (i.e. homogeneous clusters)

Elbow: Plot the within-cluster sum of squares. Optimal clusters is identified at the bend in the plot



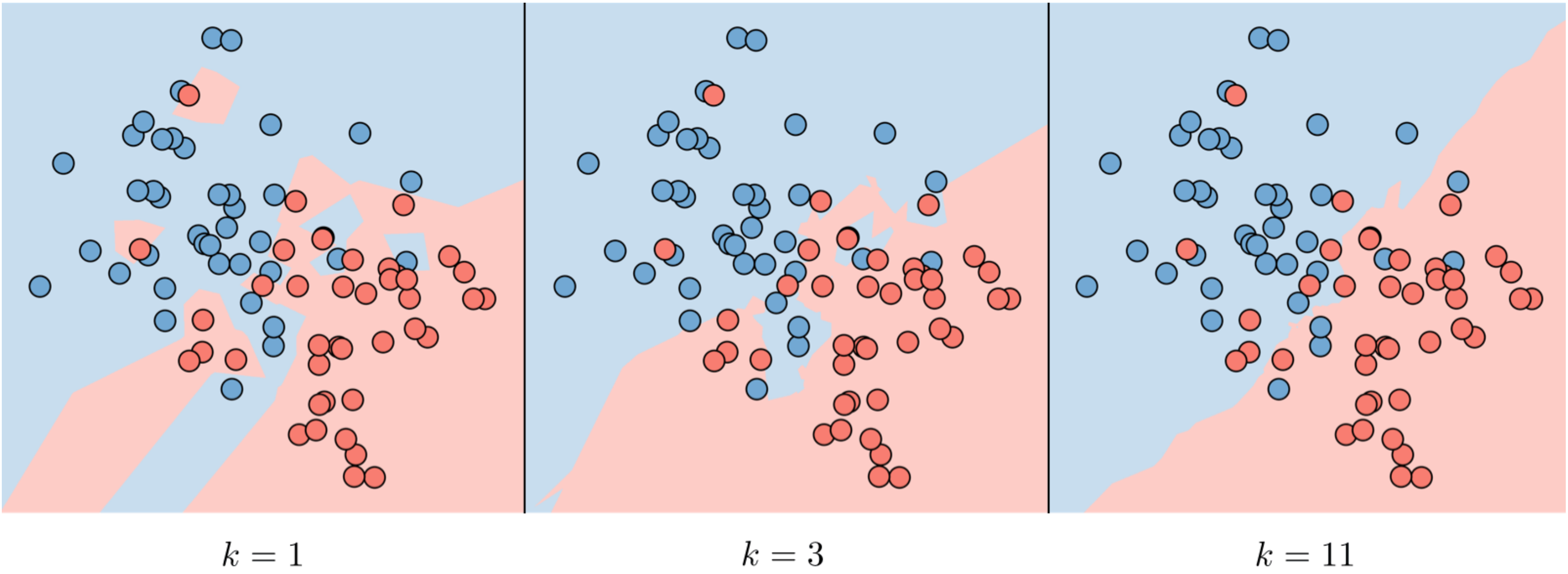
Silhouette: measure of the quality of the clustering, maximize s

$s = \frac{b-a}{\max(a,b)}$, where a is mean distance between a sample and other samples within the same class and b is the mean distance between a sample and other points in the next nearest cluster

Gap-statistic: compares total within cluster variation with their expected values under the null reference distribution. Requires boostrapping to generate reference distribution. Maximize the gap-statistic to identify optimal number of clusters.

K-nearest neighbors

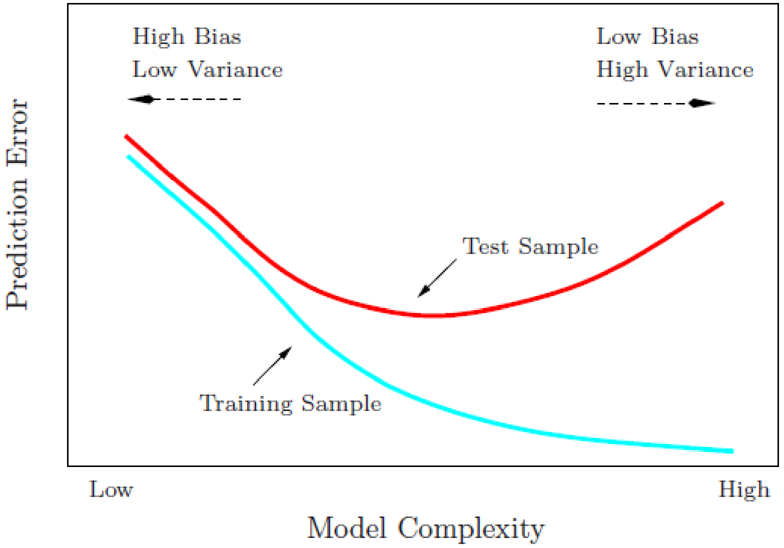
Non-parametric approach where the predicted value of an observation is determined by the nature of its k-neighbors from the training set.



Source: CS229 Course Material

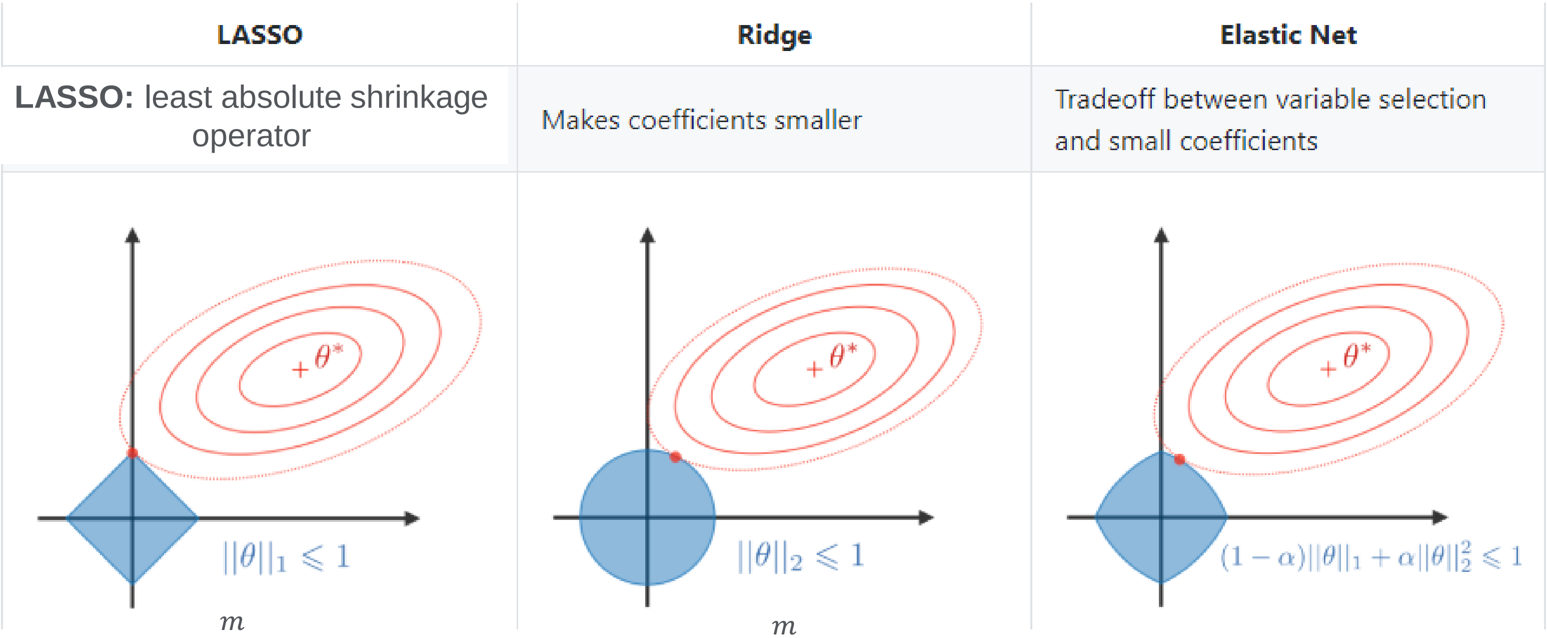
Regularized Regression (shrinkage)

Addresses limitation of regression: Increase flexibility/complexity by including power terms, interactions, etc. decreases bias but can increase variance (overfitting)



$$\underbrace{\sum_{i=1}^{N}\left(y_i-\theta_0-\sum_{j=1}^{p}\theta_j\cdot x_{ij}\right)^2}_{\substack{\text{Residual sum of squares}\\ \text{Typical least squares loss function}}} + \left.\lambda\|\theta\|\right\}\text{Penalty function}$$

Difference in the algorithms: penalty



LASSO	Ridge	Elastic Net
LASSO: least absolute shrinkage operator	Makes coefficients smaller	Tradeoff between variable selection and small coefficients

θ^*
$||\theta||_1 \leqslant 1$
θ^*
$||\theta||_2 \leqslant 1$
θ^*
$(1-\alpha)||\theta||_1 + \alpha||\theta||_2^2 \leqslant 1$

$$\lambda \sum_{j=1}^{m} |\beta|$$

L1 Regularization

$$\lambda \sum_{j=1}^{m} \beta^2$$

L2 Regularization

Combines L1 and L2 Regularization

Elastic Net:

$$\frac{\sum_{i=1}^{n} (y_i - x_i^J \hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{\beta}_j^2 + \alpha \sum_{j=1}^{m} |\hat{\beta}_j| \right)$$

Alpha is the hyperparameter that controls the loss function that differentiates LASSO, Ridge and Elastic Net

Varying Lambda (λ)

Controls the penalty function

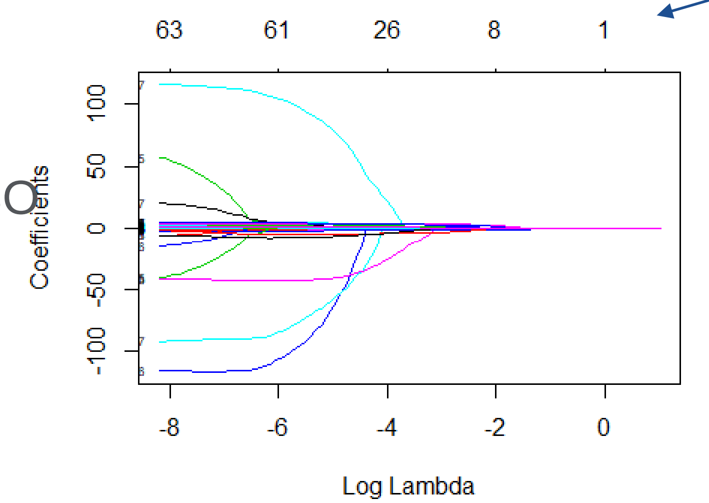
λ = 0: penalty=0 and produces standard regression coefficients

λ -> ∞: penalty=large and regression coefficients shrunk toward 0

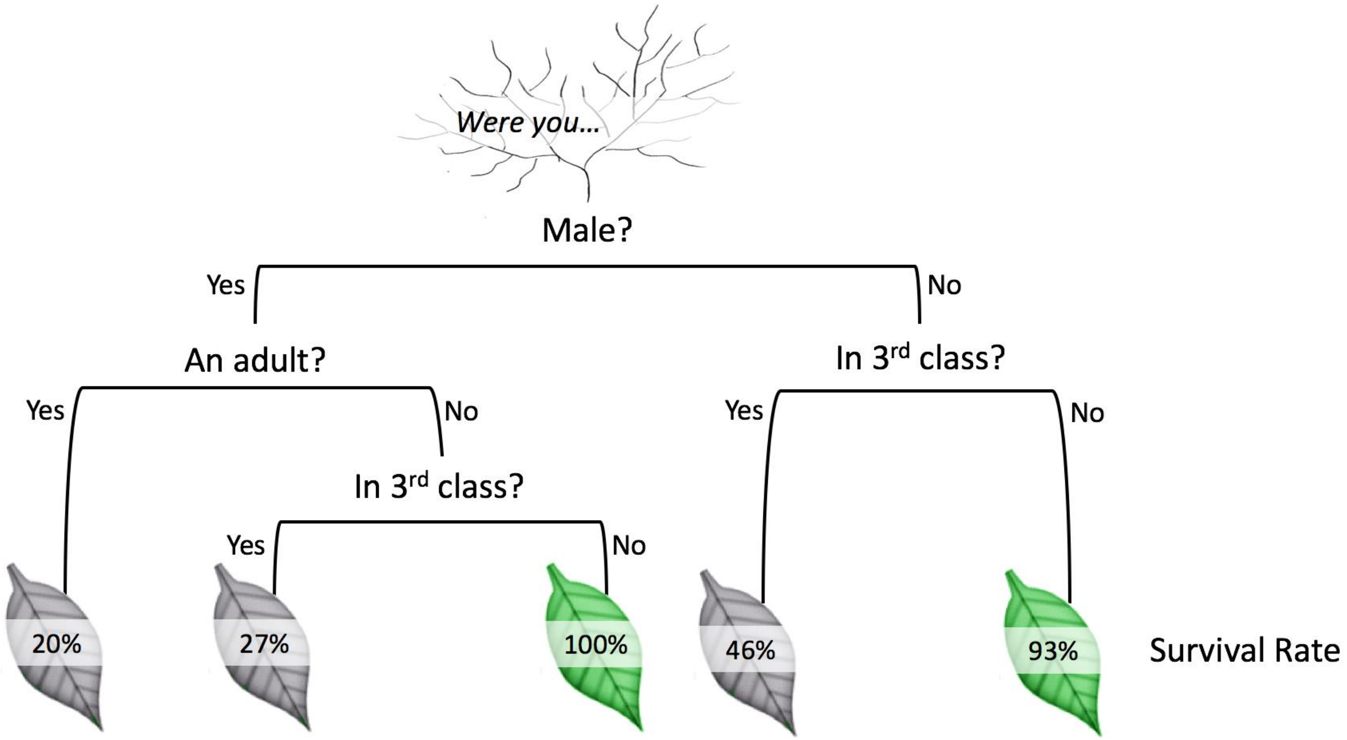
λ : chosen through **cross-validation**

Number of features

Plot of coefficients from LASSO when varying λ



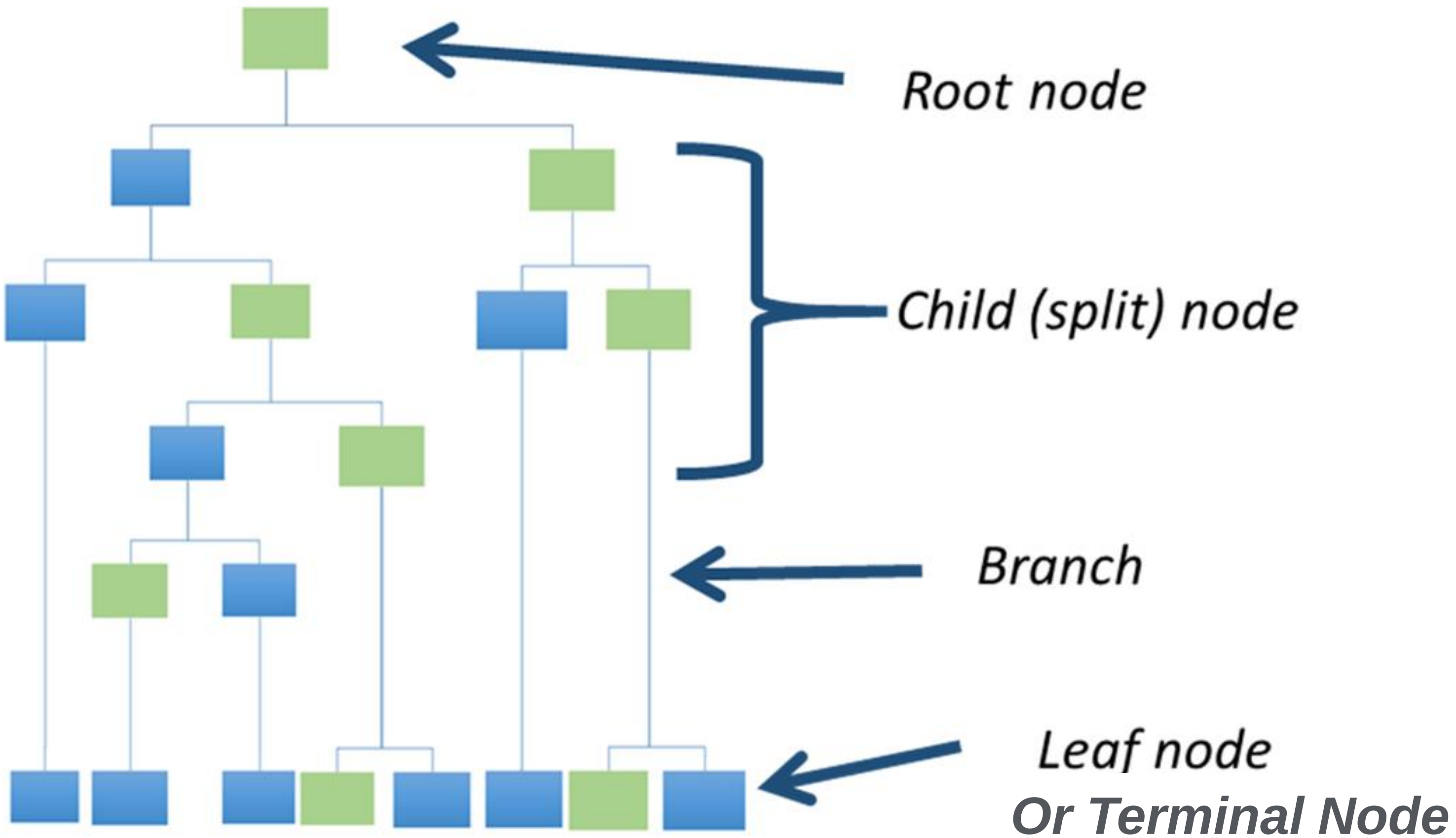
Intro to Tree-Based Methods: Titanic Example



Source: algobeans.com

Structure of a Tree

Trees generated through recursive partitioning



Root node

Child (split) node

Branch

Leaf node

Or Terminal Node

Key Terms

“Greedy” Algorithm: makes the optimal choice at each step

- Makes “best” first split without consideration of subsequent splits

Node purity: homogeneity of a node in relation to the labels of the observations contained

- Goal is often to maximize node purity to obtain optimal classification or prediction

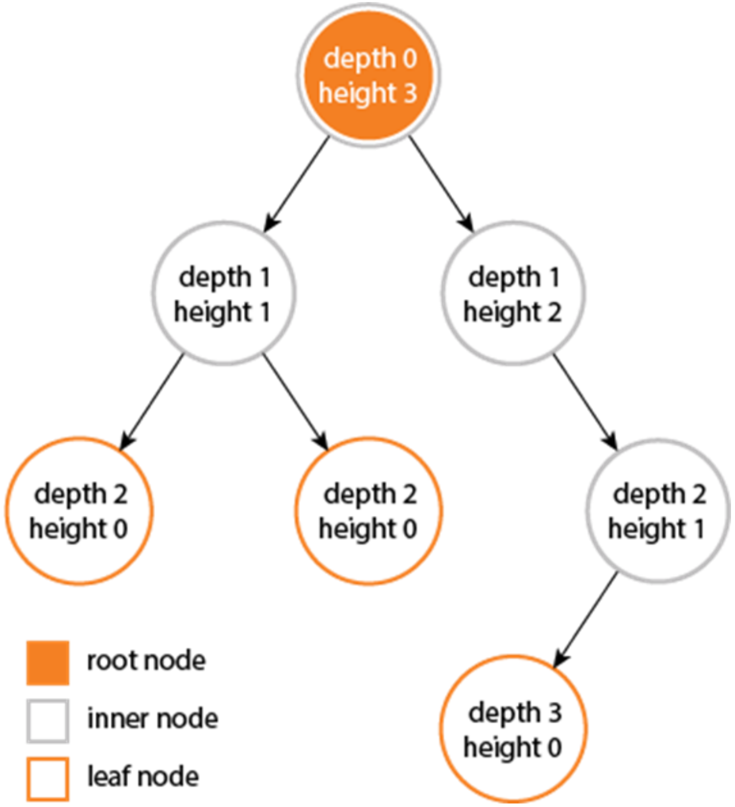
Measures of purity

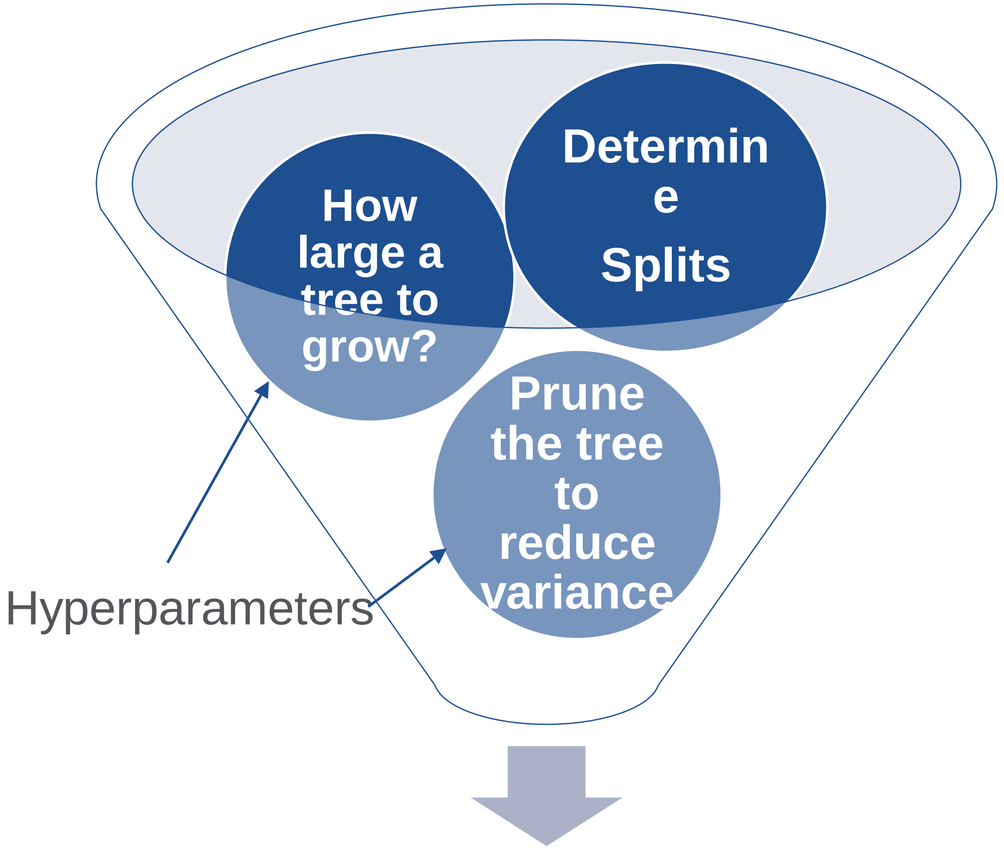
- Gini coefficient, entropy, variance, mean square error…. and others

Surrogate split: split using another feature that most closely resembles the consequences of the original split

- Often how tree-based methods handle missing data

Growing a Tree: Analytic Considerations




How large a tree to grow?

Determin e

Splits

Prune the tree to reduce variance

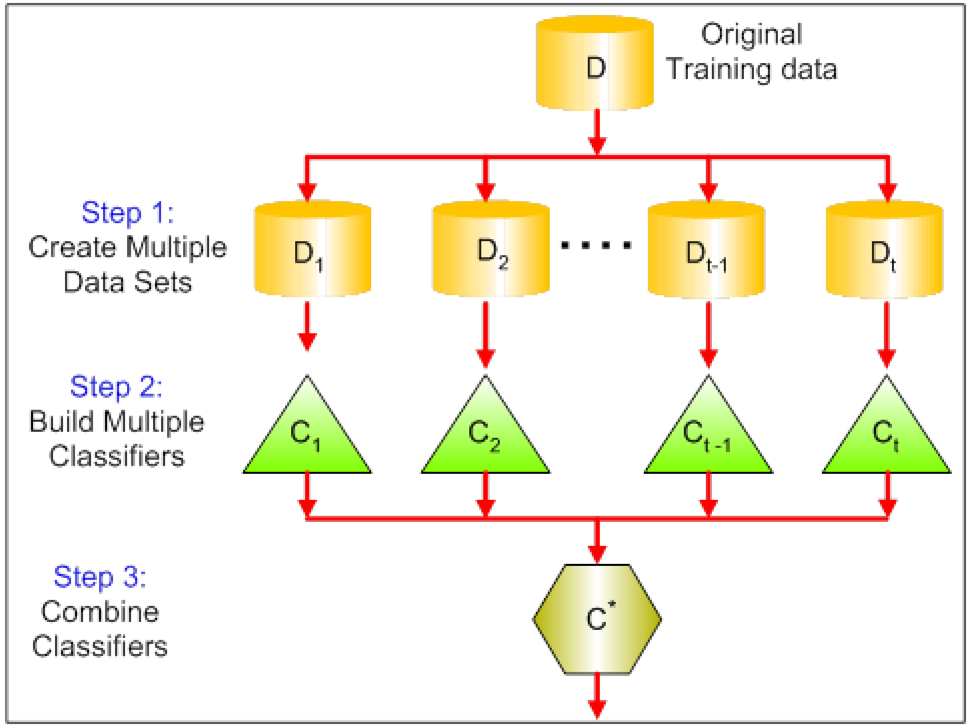
Hyperparameters

Constructing a tree-based model

Ensemble Methods

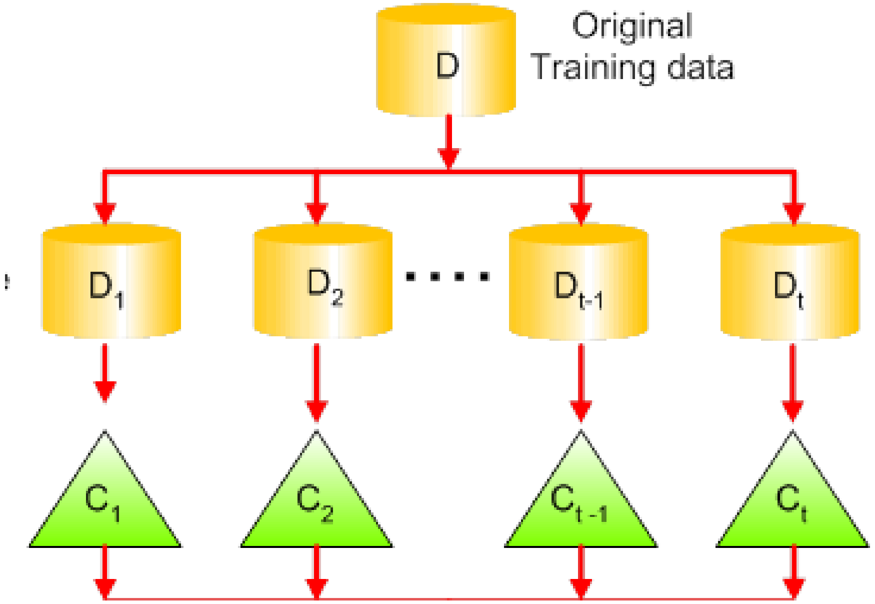
IMPROVE PREDICTION

Ensembles **combine** several base models in order to produce an optimal prediction



Simple Techniques to Combine: Classification

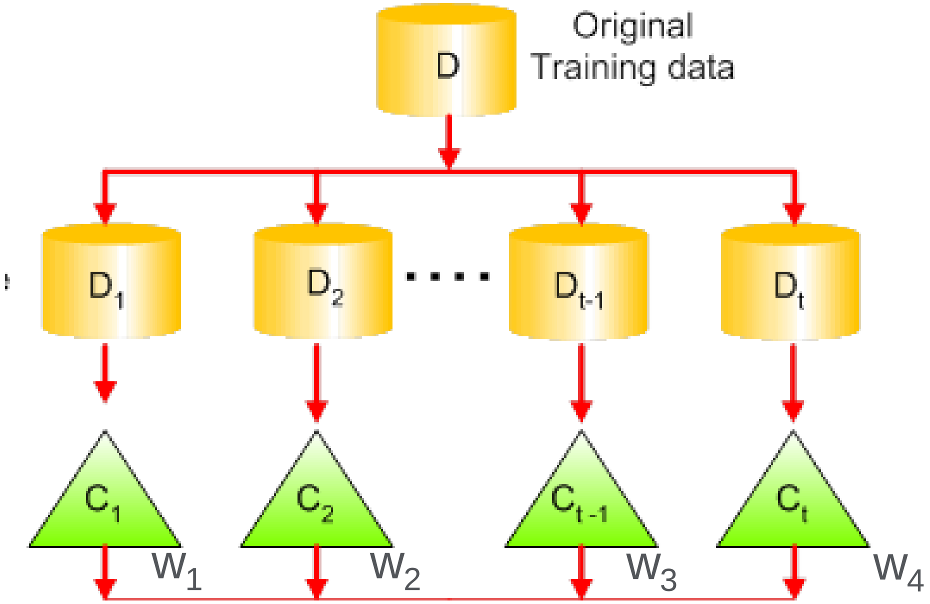
Majority Voting



C_1=1 C_2=1 C_3=0 C_4=1

C=1

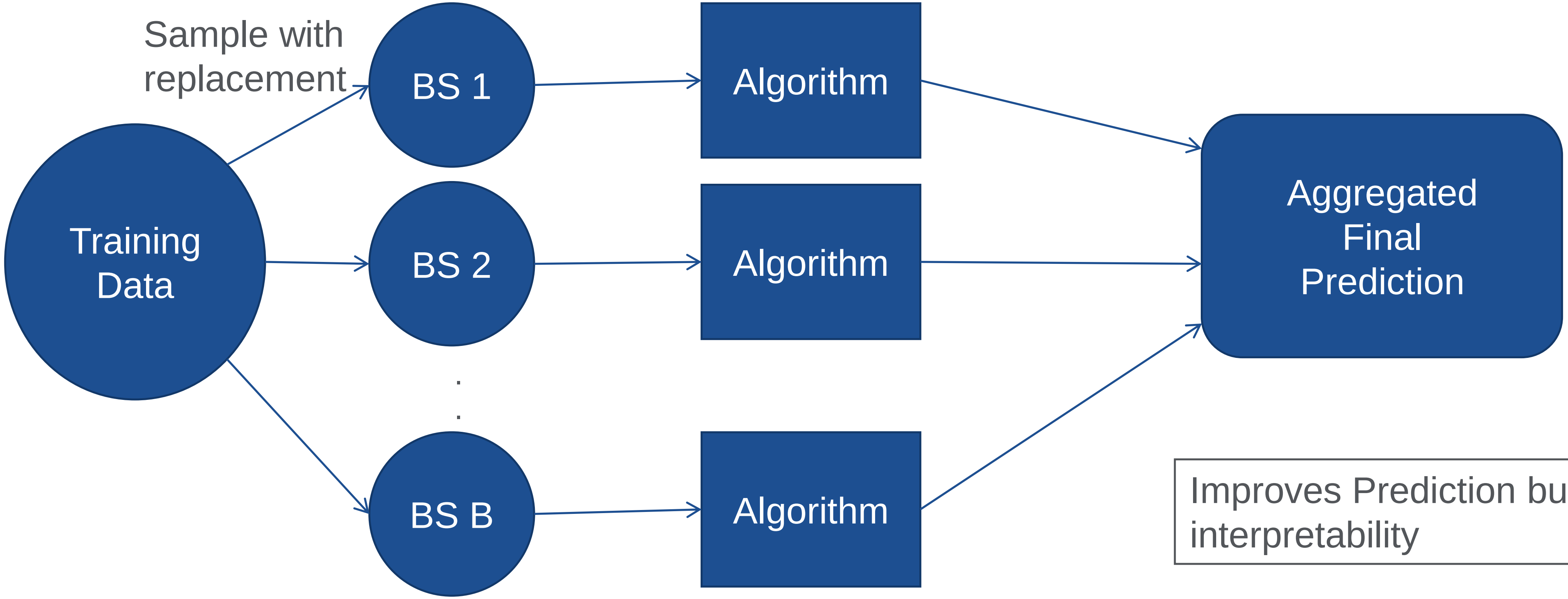
Weighted Voting



Final Prediction (C) a weighted combination with weights calculated based on predictive ability

Advanced Techniques: Bagging

Bagging: Bootstrap Aggregation; average results (or voting) across bootstrapped samples of data



Improves Prediction but may reduce interpretability

Random Forest: Extension of Bagging

At each split, we choose a random sample of *m* predictors from the full set of *p* features

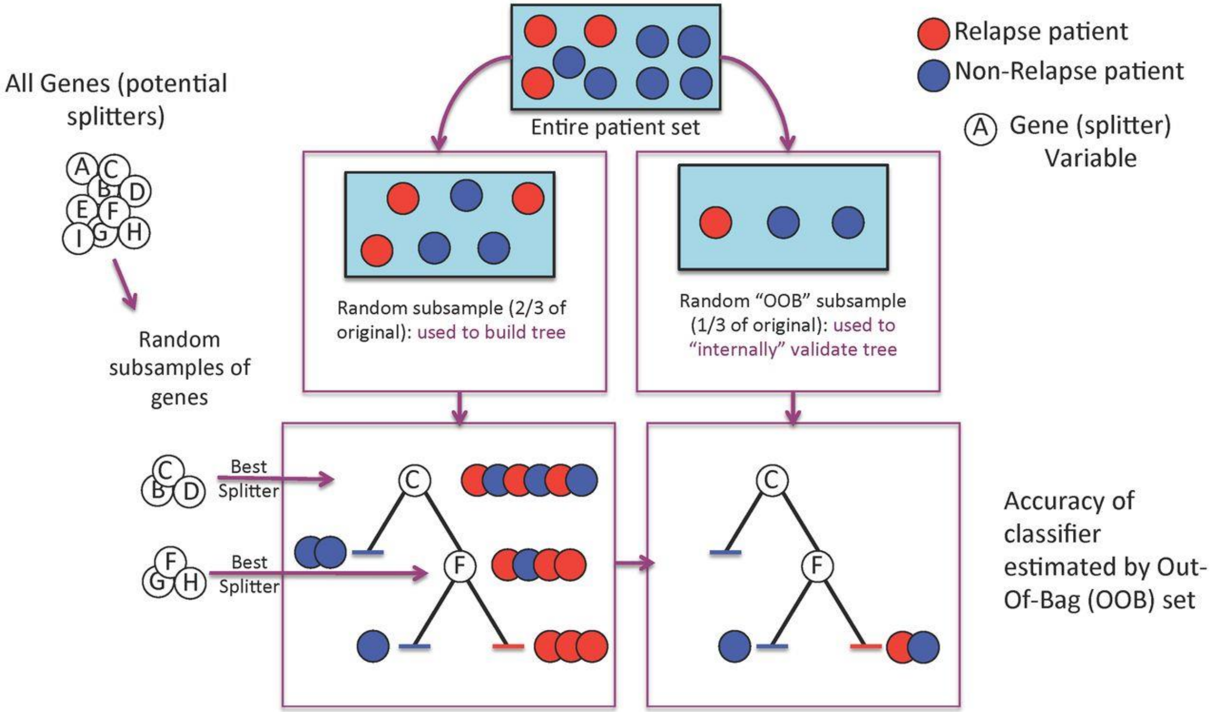
When m=p, then bagging and RF are equivalent

Randomly sampling from features de-correlates the trees

Provides opportunity for all features to contribute to prediction

Improves prediction overall, even though individual trees may be weaker

For prediction, want m to be small, typically m=√p


All Genes (potential splitters)
A C B D E F I G H
Random subsamples of genes
Entire patient set
Relapse patient
Non-Relapse patient
A Gene (splitter) Variable
Random subsample (2/3 of original): used to build tree
Random "OOB" subsample (1/3 of original): used to "internally" validate tree
C B D Best Splitter
F G H Best Splitter
C
F
Accuracy of classifier estimated by Out-Of-Bag (OOB) set

Source: BioInformatics Handbook

Variable Importance Factors

Measure of Individual Variable Contribution to the Overall Prediction

Accuracy-based Importance

Within OOB, record the prediction/classification error.

Permute the feature and recalculate prediction/classification

Difference in two errors are averaged over trees and normalized.

Node purity based Importance (Gini importance for classification)

Within OOB, total decrease in node impurities from splitting on variable averaged over all trees.

Advanced Techniques: Boosting

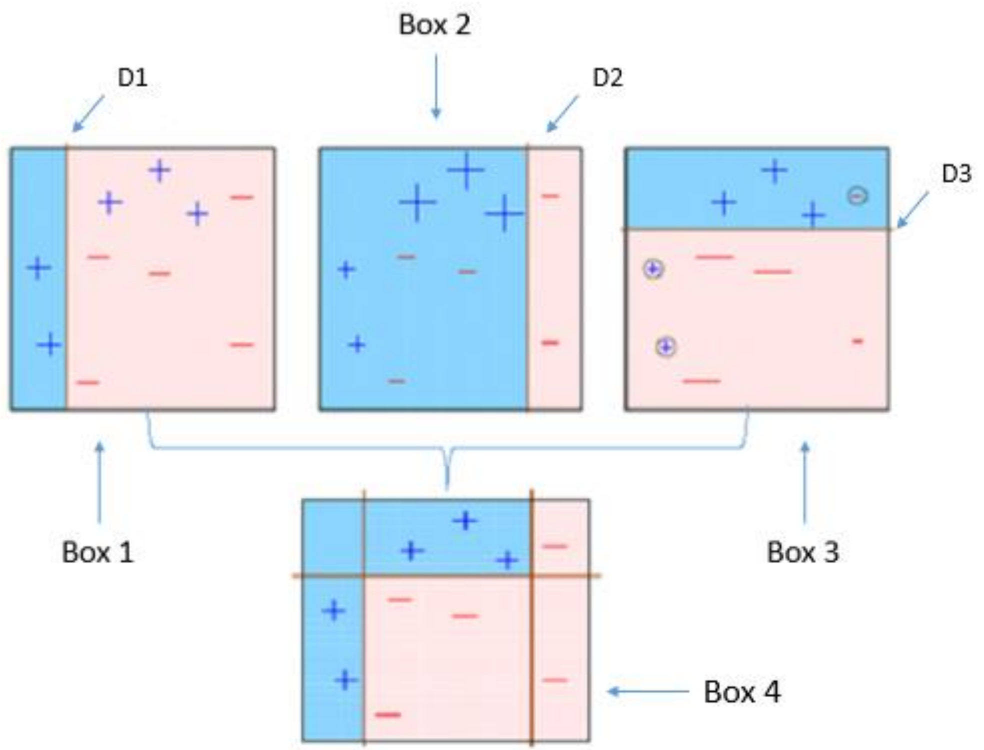
General technique that can be applied to diverse algorithms; popular with trees as base

Goal: Convert series of weak learners to a strong learner by learning algorithms sequentially

Example: Results from first algorithm provide information to second, etc.

Both provide an initial strong learner and then add to the initial learner with weaker learners

In adaptive boosting, information is weights of data points. If classified correctly, gets smaller weight so algorithm focuses on misclassified datapoints.



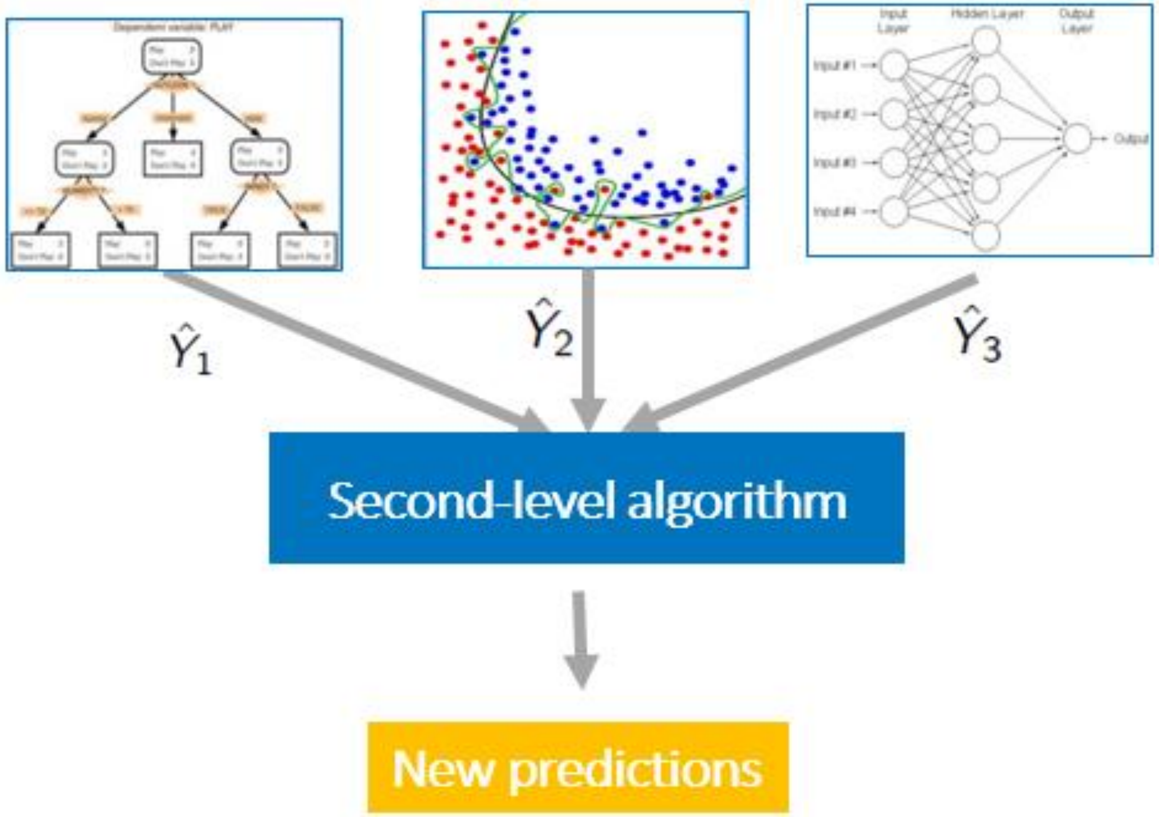
Source: Desarda Understanding AdaBoost

Advanced Techniques: Stacking

Rather than combining results of multiple learners, stacking uses predictions from learners as inputs to additional learners.

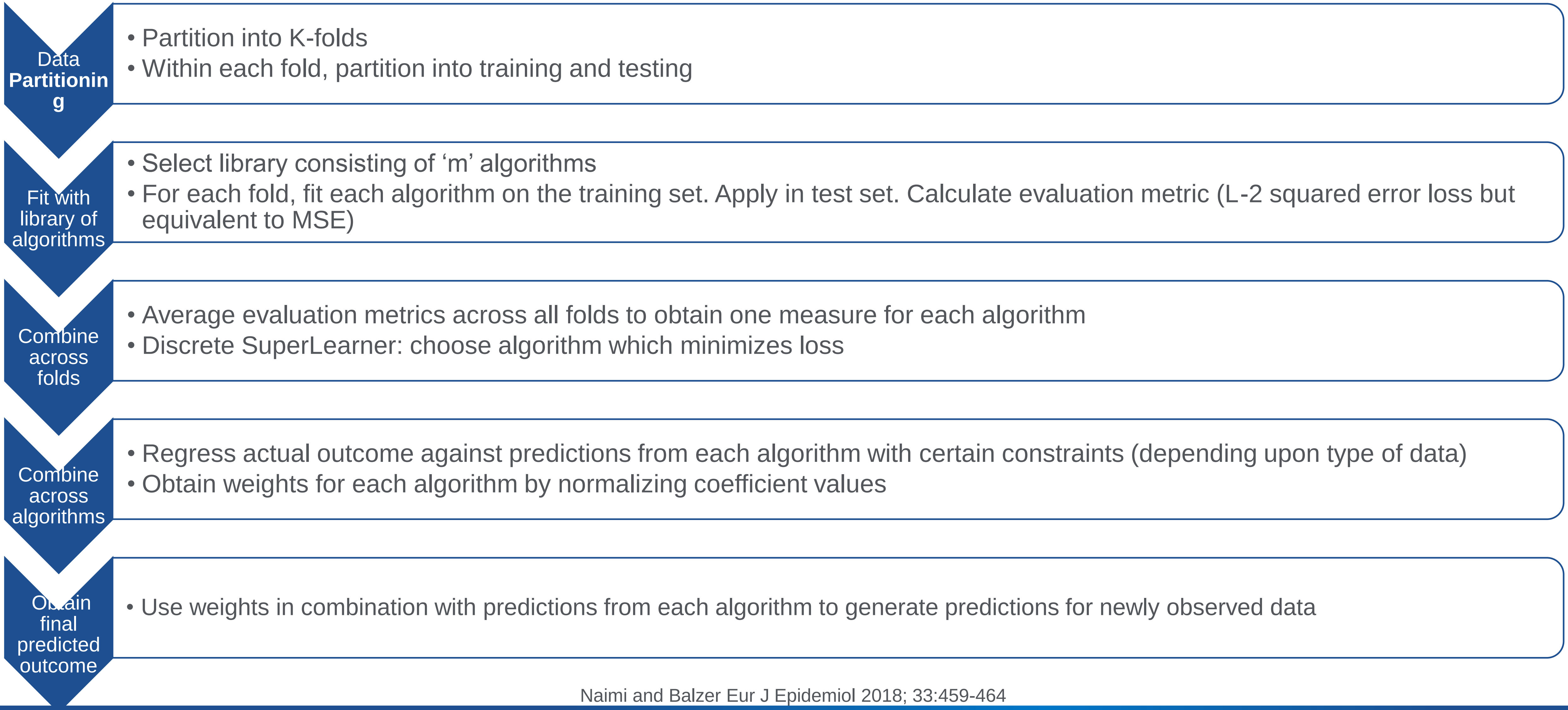
Optimal performance when diverse learners are used in the initial predictions

Common Example in Epidemiology: SuperLearner



Source: SAS Blogs

Overview of Super Learner



Data **Partitioning**

- Partition into K-folds
- Within each fold, partition into training and testing

Fit with library of algorithms

- Select library consisting of ‘m’ algorithms
- For each fold, fit each algorithm on the training set. Apply in test set. Calculate evaluation metric (L-2 squared error loss but equivalent to MSE)

Combine across folds

- Average evaluation metrics across all folds to obtain one measure for each algorithm
- Discrete SuperLearner: choose algorithm which minimizes loss

Combine across algorithms

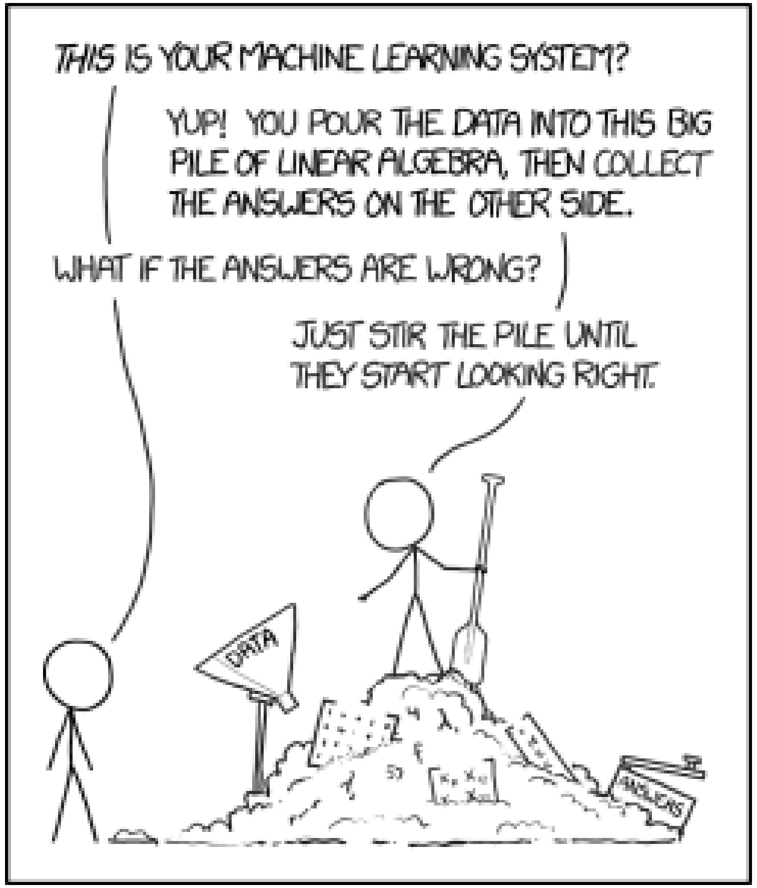
- Regress actual outcome against predictions from each algorithm with certain constraints (depending upon type of data)
- Obtain weights for each algorithm by normalizing coefficient values

Obtain final predicted outcome

- Use weights in combination with predictions from each algorithm to generate predictions for newly observed data

Naimi and Balzer Eur J Epidemiol 2018; 33:459-464

Critical thinking is not optional….


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

Credit: XKCD

Consider needs of research question….

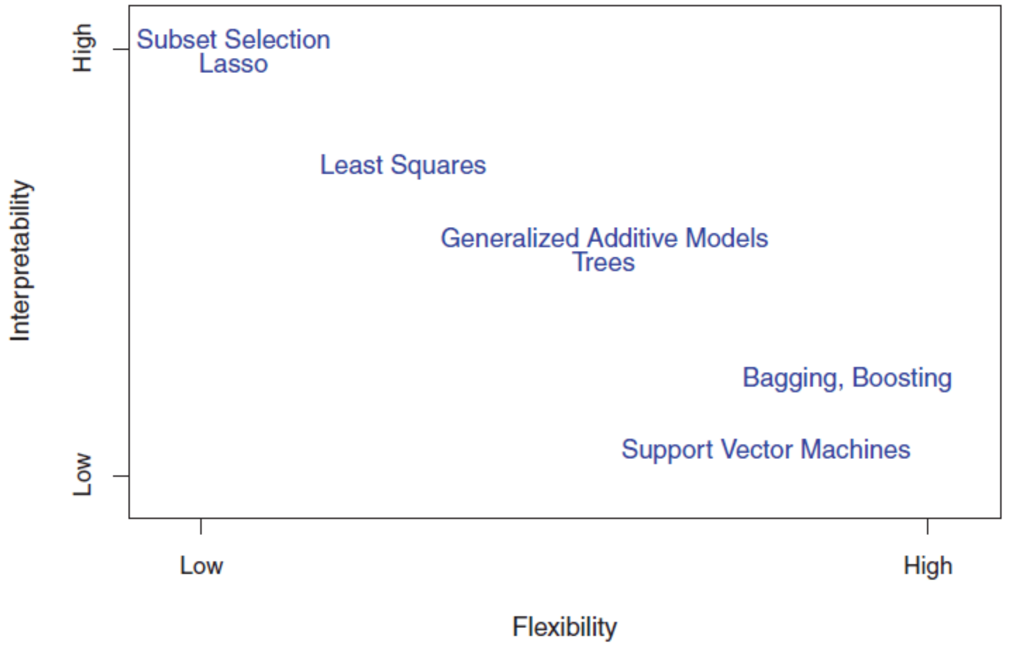


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

Source: ISLR

What types of epidemiologic questions/tasks benefit from high flexibility?

What types require more interpretability?

What does interpretability mean in the context of machine learning?

Consider particulars of the data…

Are data highly correlated?

Do you anticipate non-linear effects?

Are you interested in interactions between features/exposures?

Multiple avenues to integrate machine learning into epidemiologic research and practice..

Complex algorithms won’t solve all our problems, but they can enhance our research if applied *thoughtfully*

Largely underutilized potential of using machine learning for other tasks like data harmonization, data-linkage and knowledge modelling

Funding and Institutional Support:
NIEHS Career Development Award (R00ES027022); HHEAR Data Science Center-Subcontract (U2CES026555)
Columbia Calderone Early Faculty Prize
Columbia DSI Scholars Program