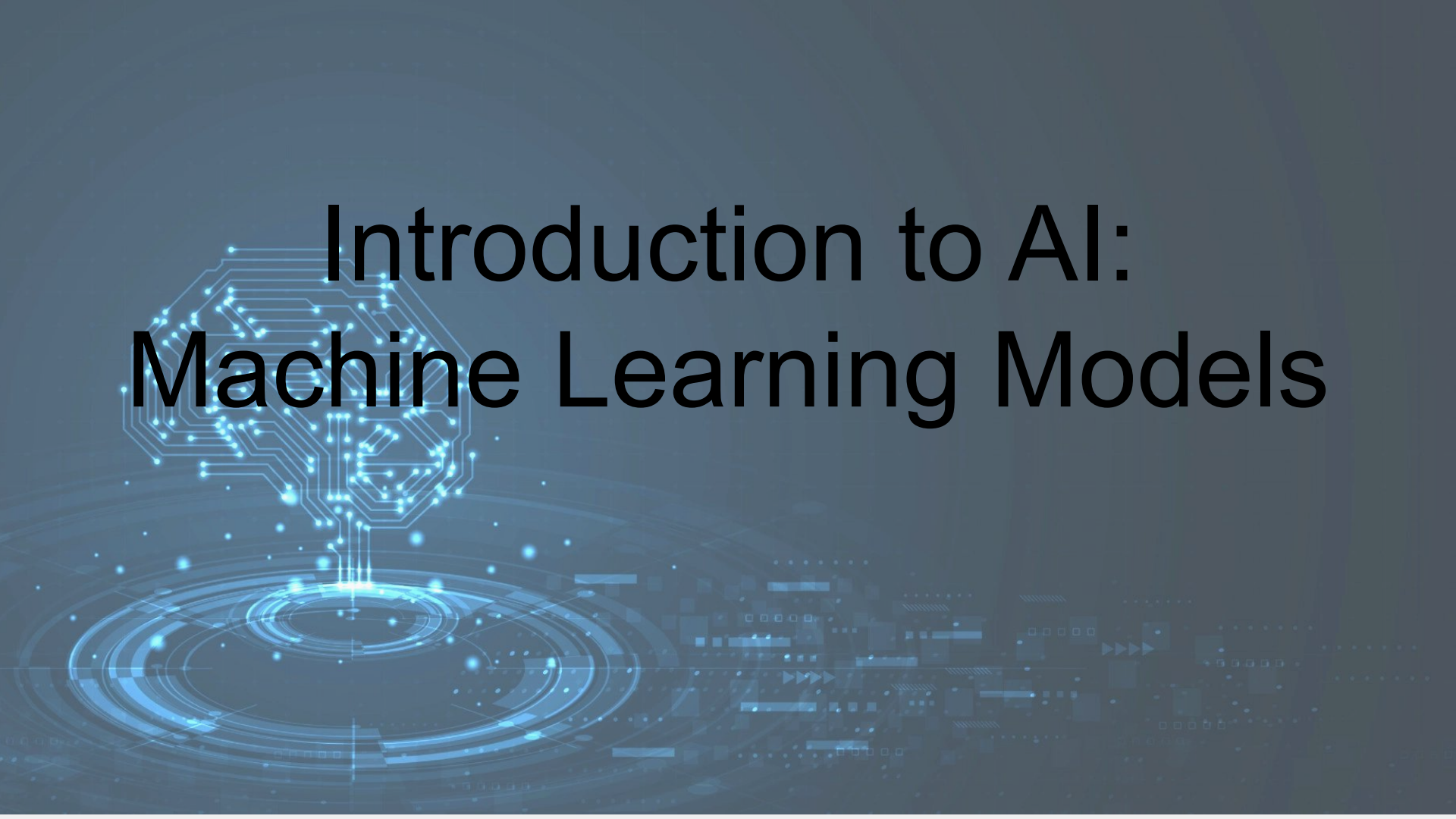
Introduction to AI: Machine Learning Models



Motivating Example

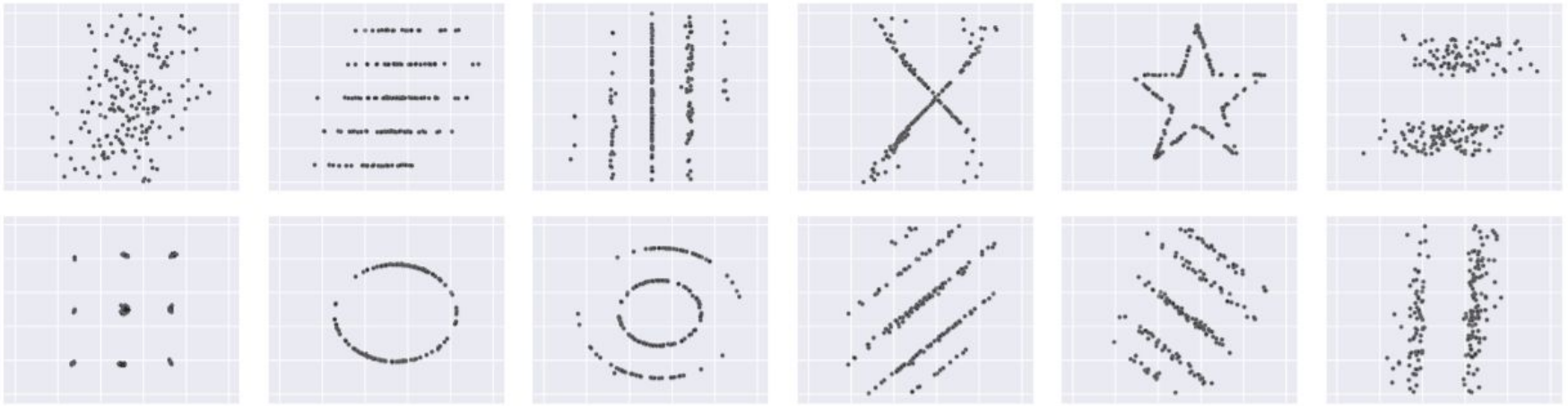
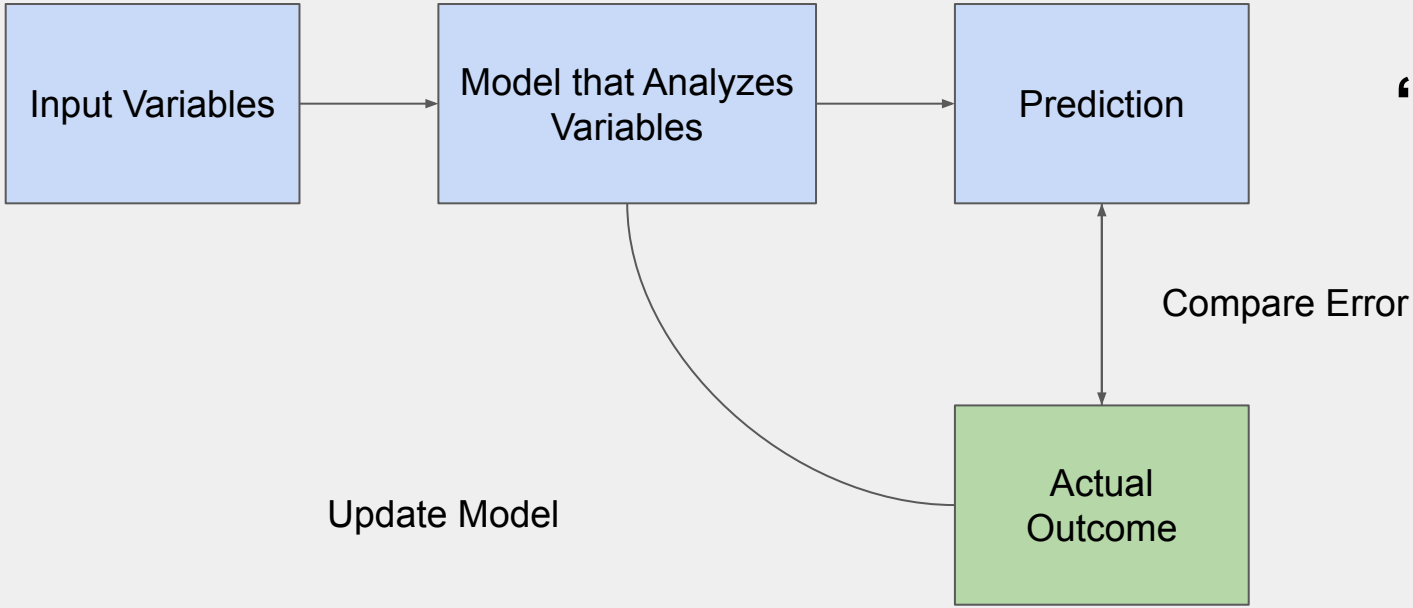


Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. ($\bar{x}$ =54.02, $\bar{y}$ = 48.09, sd_x = 14.52, sd_y = 24.79, Pearson's r = +0.32)

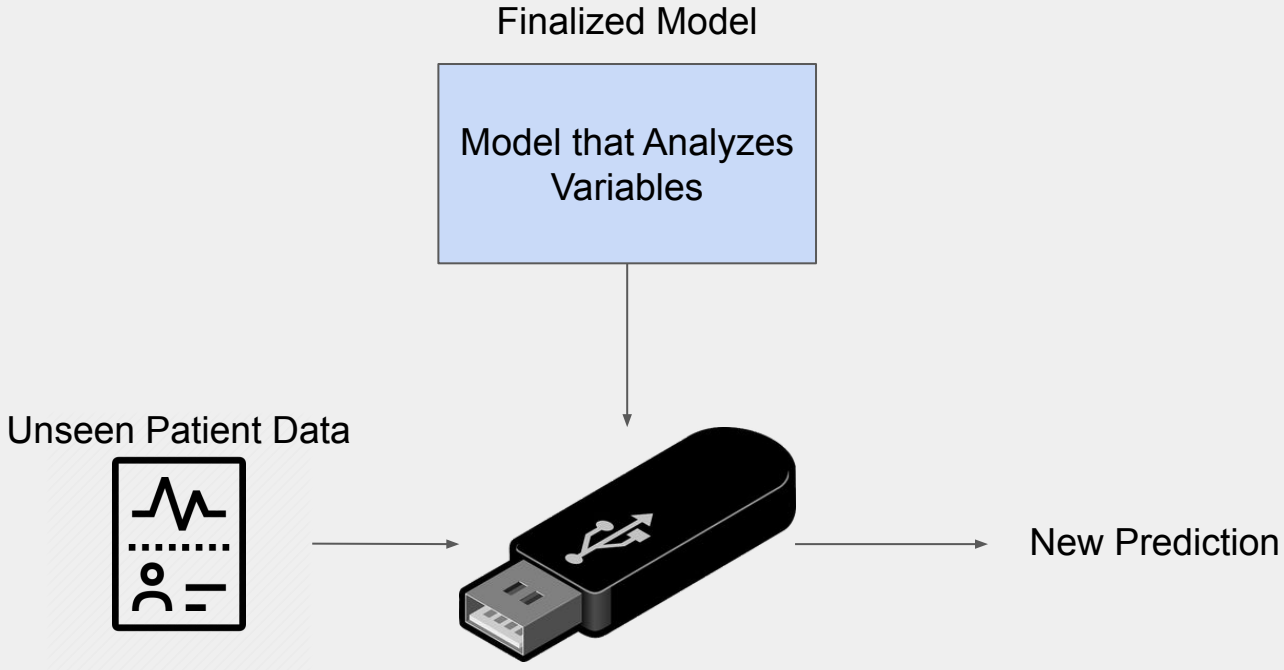
Matejka et al. (2017)

Machine Learning (supervised learning)



“Training”

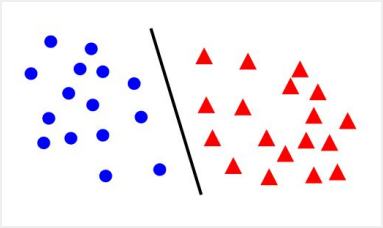
Machine Learning



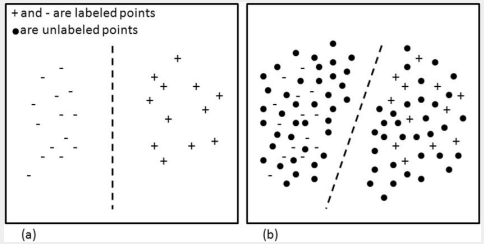
“Testing”

Learning Spectrum

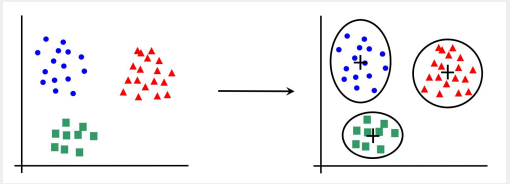
Supervised

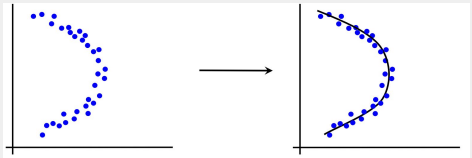


Semi-supervised



Unsupervised





Less labeled data req.

Predictive Power

Types of Supervised Learning

Regression

Input features: **Continuous**

Output prediction: **Continuous**

Classification

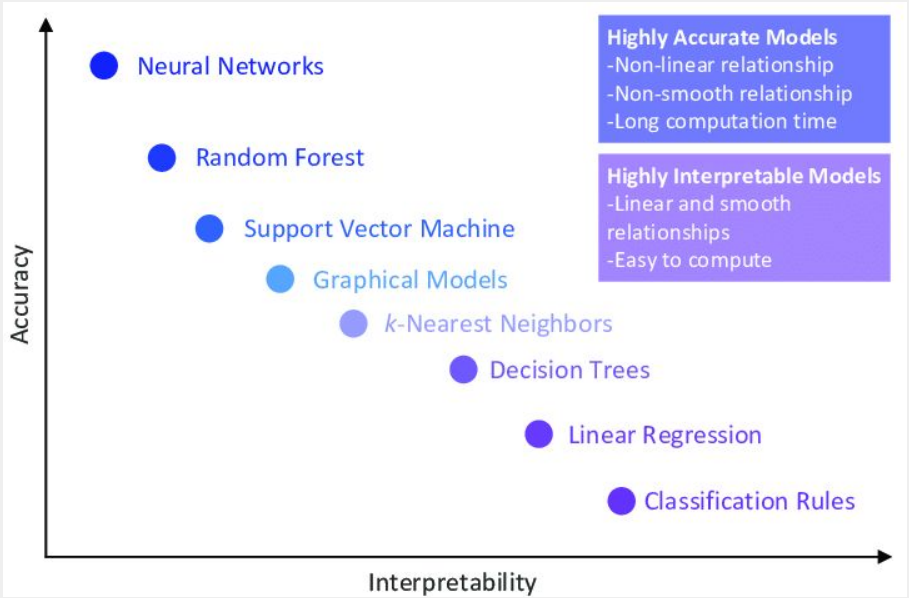
Input features: **Continuous** OR **Categorical**

Output prediction: **Categorical**

Both continuous and categorical data have to be **numeric**

Model selection - General Guidelines

- As model complexity increases:
  - Predictive accuracy increases
  - Required data increases (more parameters)
  - Potential for overfitting increases
  - Interpretability decreases



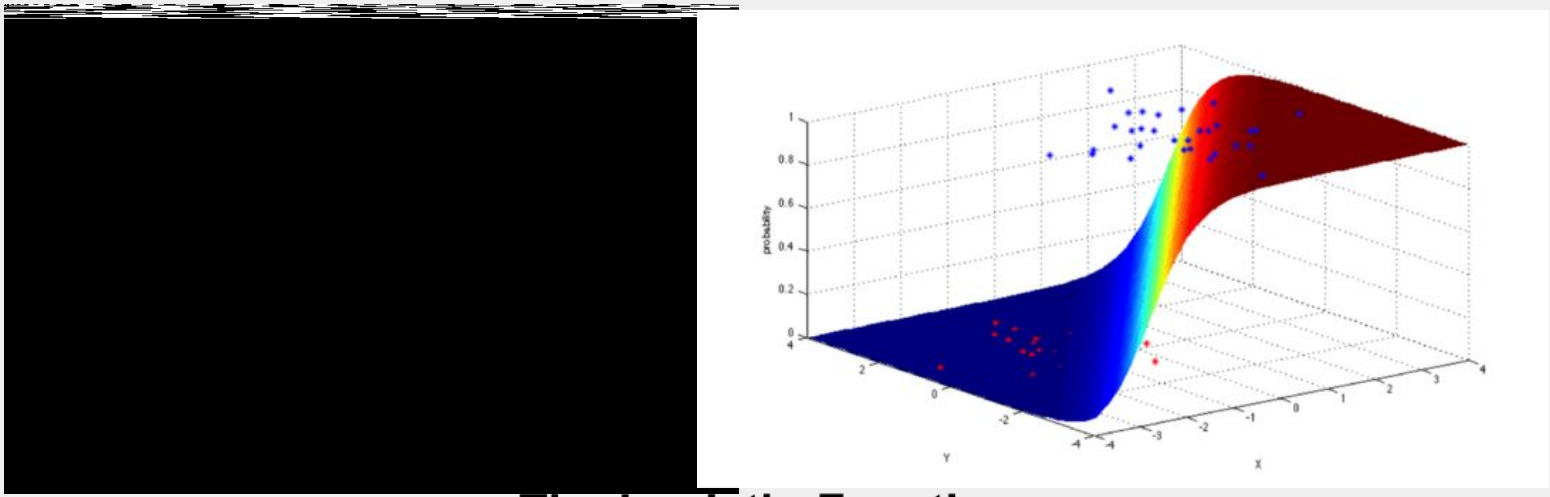
Classification Rules

CHA_2DS_2-VASc

	Condition	Points
C	Congestive heart failure (or Left ventricular systolic dysfunction)	1
H	Hypertension: blood pressure consistently above 140/90 mmHg (or treated hypertension on medication)	1
A_2	Age ≥75 years	2
D	Diabetes Mellitus	1
S_2	Prior Stroke or TIA or thromboembolism	2
V	Vascular disease (e.g. peripheral artery disease, myocardial infarction, aortic plaque)	1
A	Age 65–74 years	1
Sc	Sex category (i.e. female sex)	1

Score	Risk	Anticoagulation Therapy
0 (male) or 1 (female)	Low	No anticoagulant therapy
1 (male)	Moderate	Oral anticoagulant should be considered
2 or greater	High	Oral anticoagulant is recommended

Logistic Regression



The Logistic Function

$$\text{Log}\left[\frac{Y}{(1-Y)}\right] = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n$$

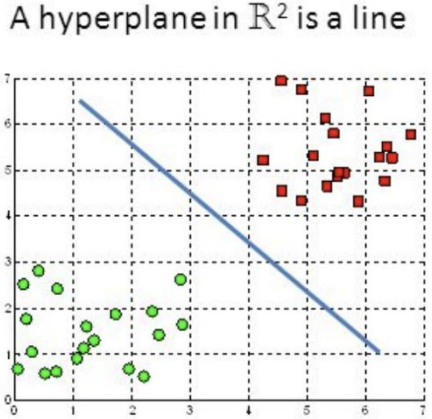
Log(Likelihood)

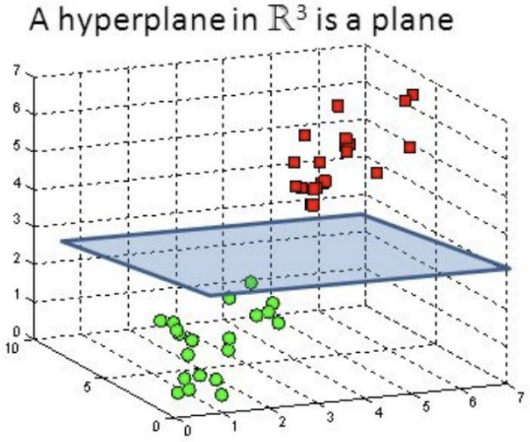
diet score (0-15)

age group (0/1)

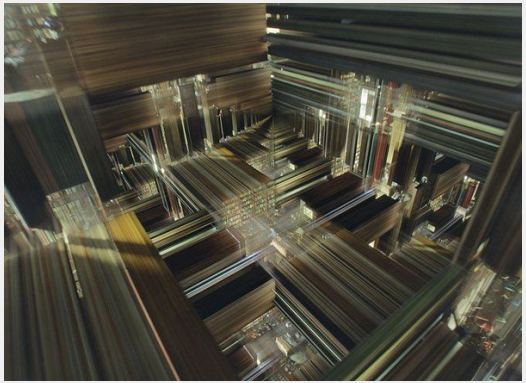
sex (0/1)

Support Vector machine

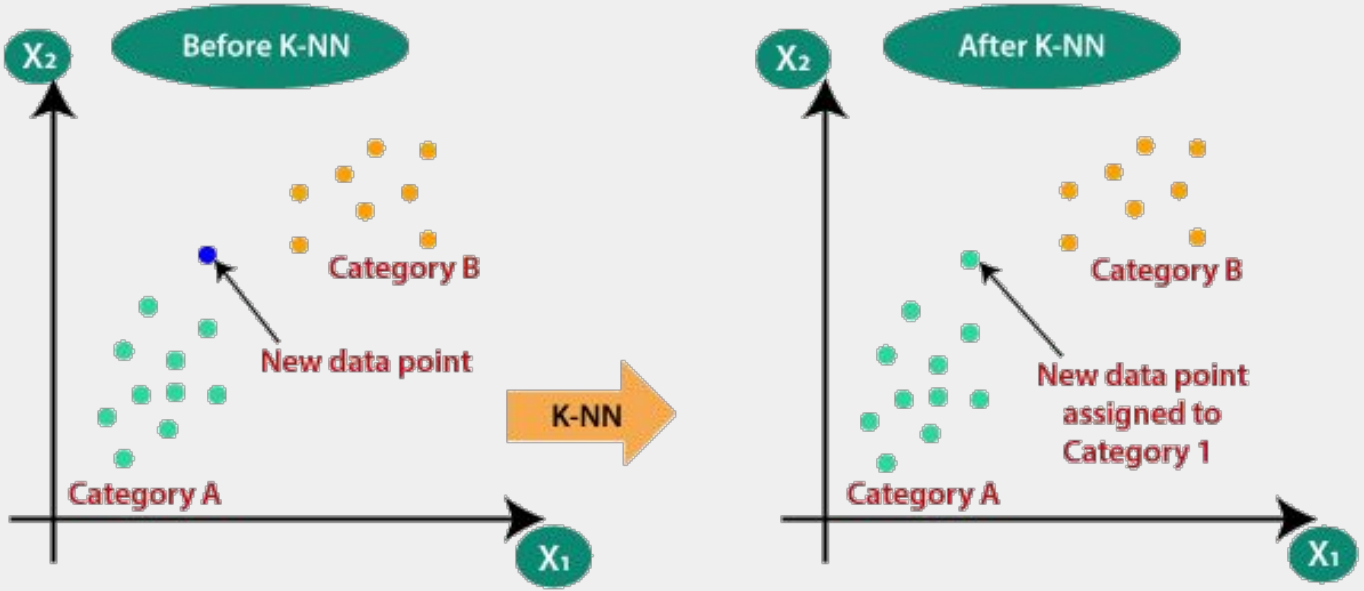




In 4D



K-nearest neighbors



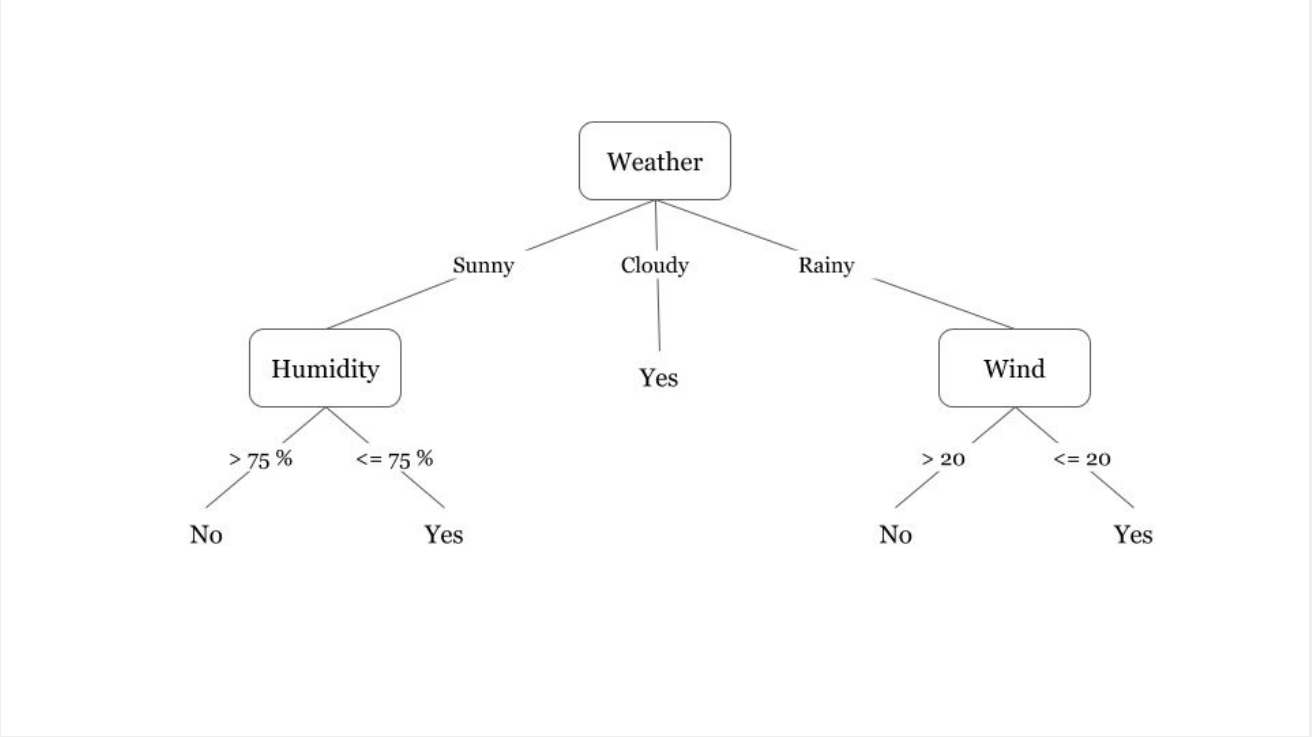
Decision Trees

- Technique for **classification or regression**
  - **Data doesn’t have to be linearly separable**
- Input and output variables can be **continuous or categorical**
- **Completely interpretable**

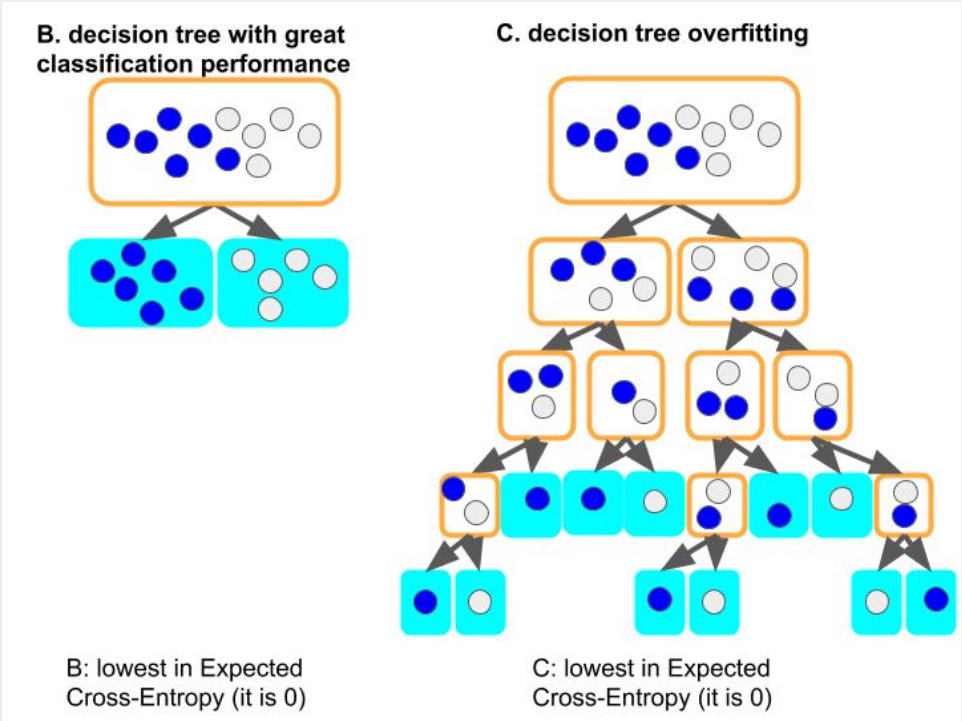
Decision Trees - Example

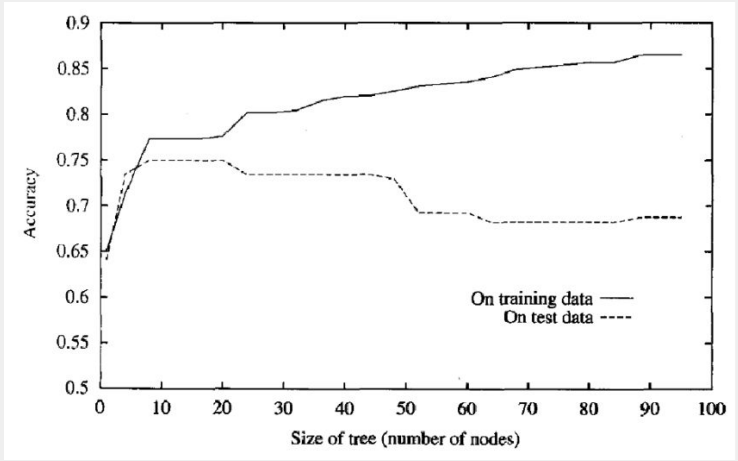
Day	Weather	Temperature	Humidity	Wind	Attack?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

Decision Trees - Example

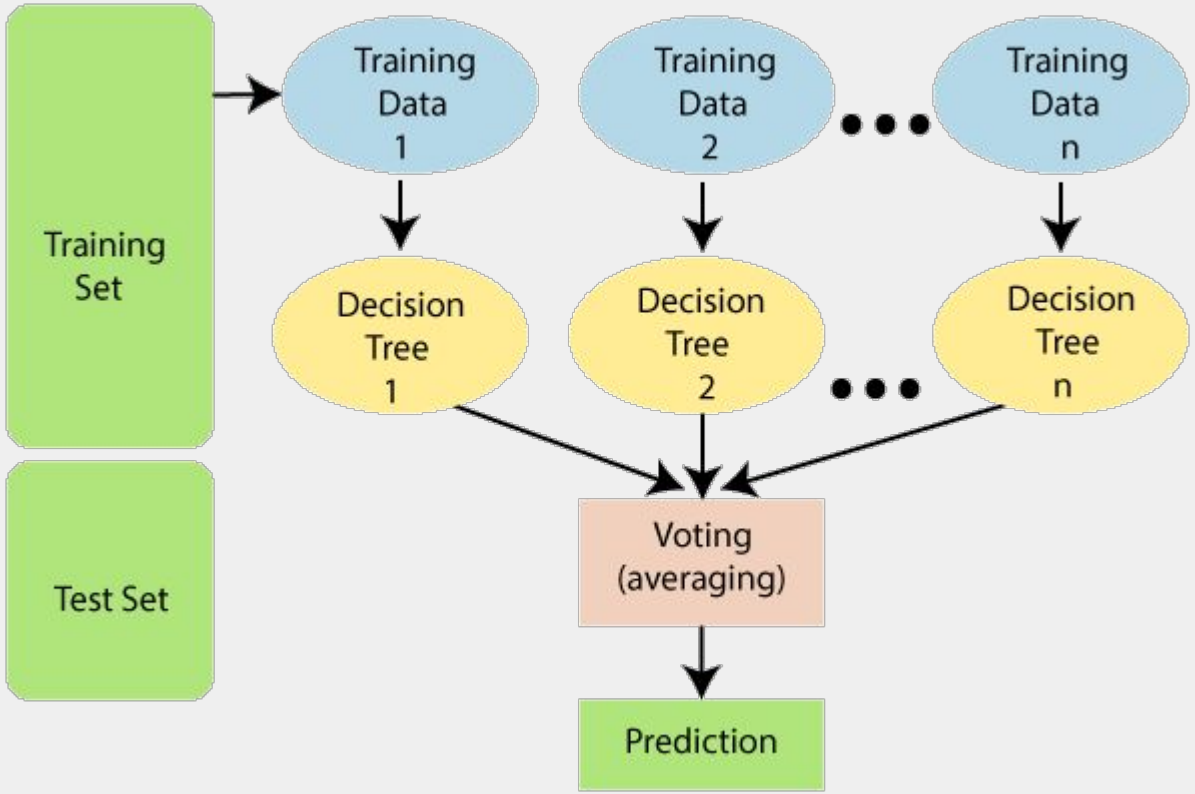


Decision Tree - Overfitting

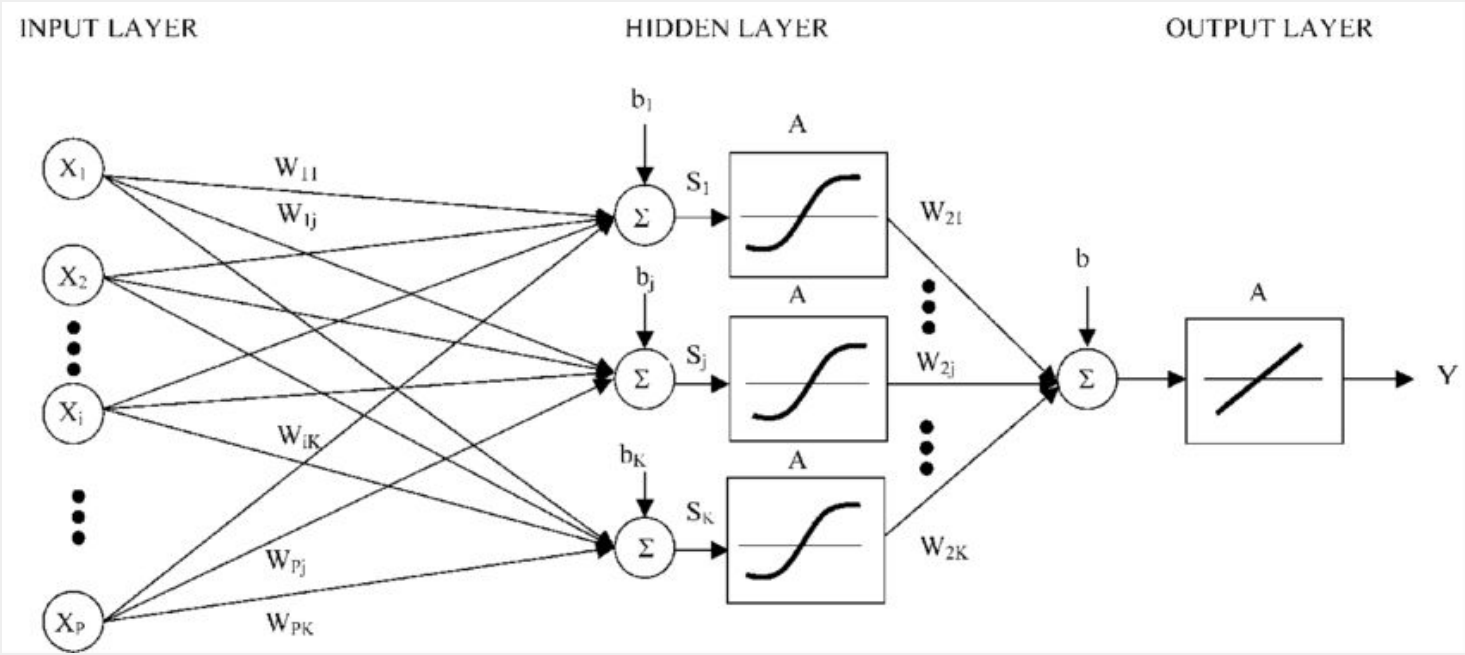




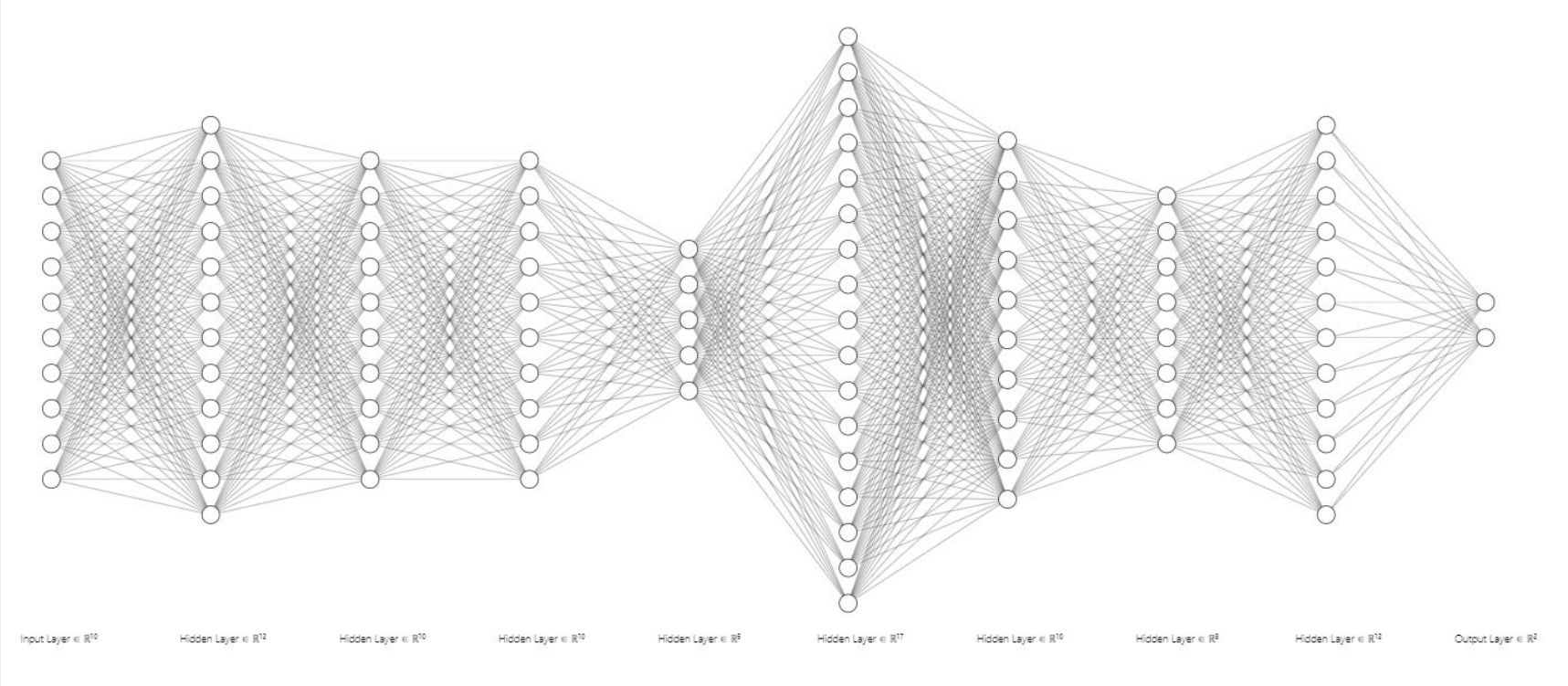
Random Forest



Neural Network

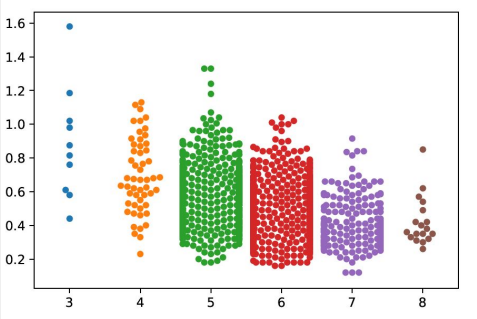


Scaling Up


Input Layer ∈ ℝ¹⁰
Hidden Layer ∈ ℝ¹²
Hidden Layer ∈ ℝ¹⁰
Hidden Layer ∈ ℝ¹⁰
Hidden Layer ∈ ℝ⁶
Hidden Layer ∈ ℝ¹⁷
Hidden Layer ∈ ℝ¹⁰
Hidden Layer ∈ ℝ⁸
Hidden Layer ∈ ℝ¹²
Output Layer ∈ ℝ²

Training Process

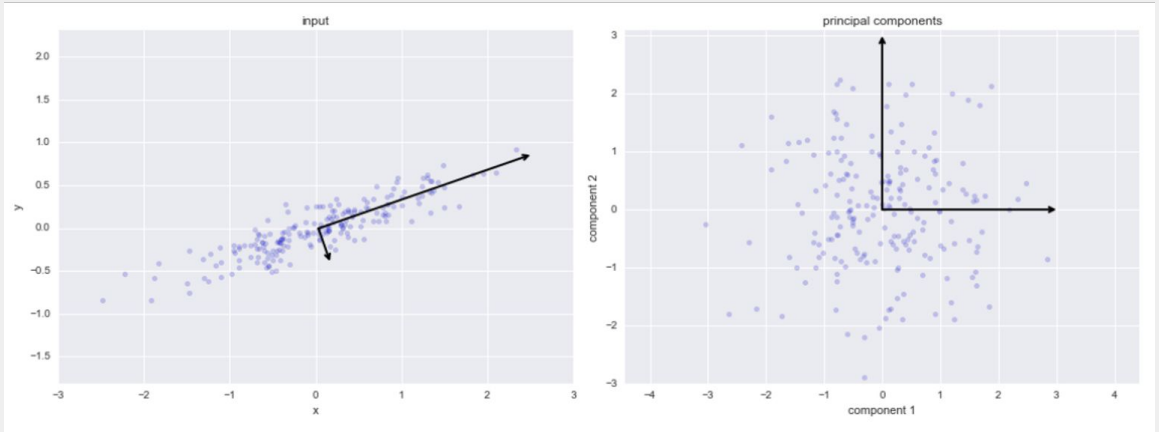
- You've picked an algorithm, now what?

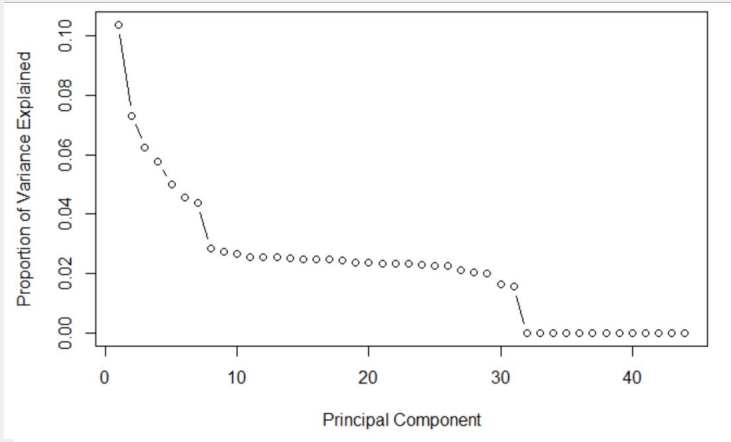


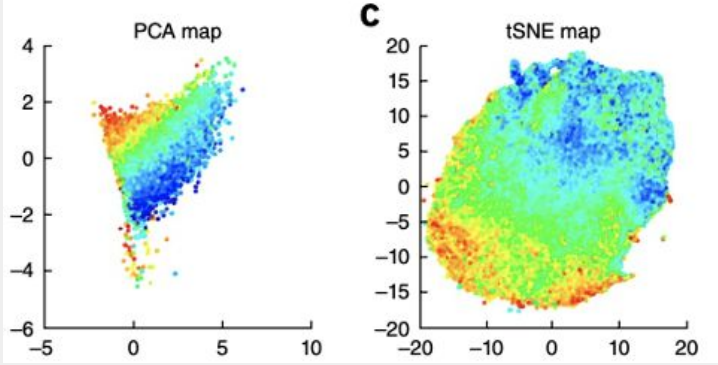
1. Explore the data
   a. Visualize (plot) it
   b. Inspect the values
2. Prepare your data:
   a. 'Clean' it - remove outlier samples, remove/fill missing values, etc
   b. Decide which features you want to use. Potential for dimensionality reduction/clustering
   c. Know which variables are continuous or categorical
   d. Know what you are predicting

Dimensionality Reduction

- Excessive features make models more difficult to train
- Techniques to combat this problem:
  - Principal Component Analysis (PCA)
  - t-Stochastic Neighbor Embedding (tSNE)
  - Initial univariate stats, rank by odds ratios

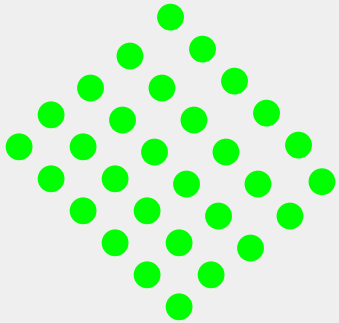


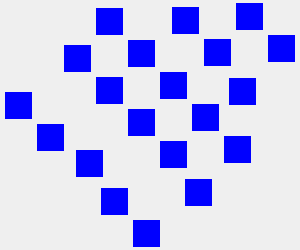


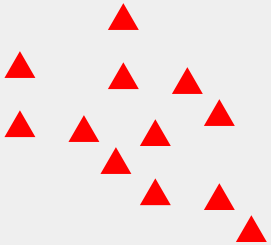


Training Process

3. **Split your data** into training, validation and test sets
    a. The test set should only be used once at the very end to check your model performance

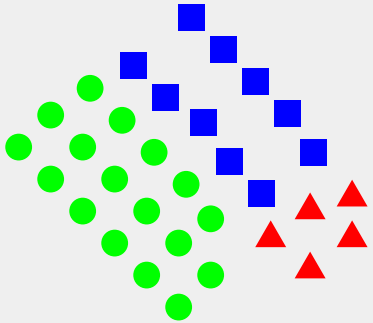




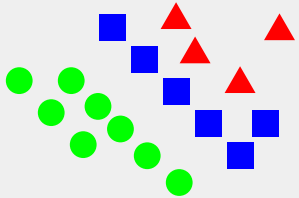


Training Process

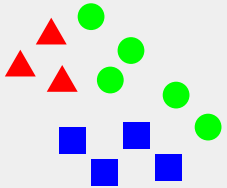
3. **Split your data** into training, validation and test sets
    a. The test set should only be used once at the very end to check your model performance



Training



Validation



Testing

Stratified - equal proportion of events in each set

Training Process

4. **Train** the ML algorithm with the training set

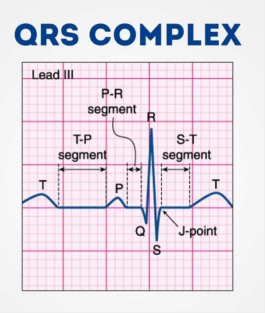
5. **Validate** results on the validation set
   a. Cross validation: Multiple different validation sets can be created from the original data

6. Change ‘hyperparameters’ and repeat training and validation if necessary

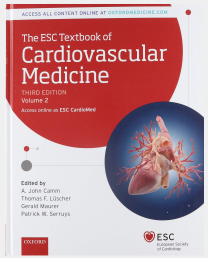
7. Good validation performance → Check results on the **test set**!

Parameters vs. Hyperparameters

- Parameters are the **weights** / **biases** that the ML model is fitting to the data

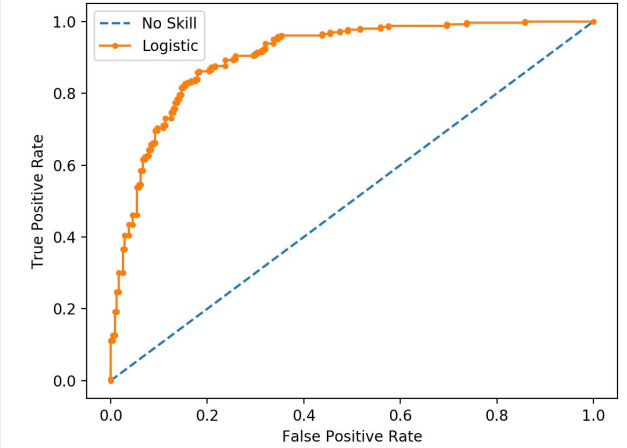


- Hyperparameters are parameters that govern **how** the ML model learns
  - The maximum depth of a decision tree
  - Number of iterations
  - Optimization technique
  - Batch size
  - Learning rate



Evaluating Performance

- Classification:
  - Accuracy, recall, precision,
  - Receiver operating characteristic (ROC) curve
- Regression:
  - Mean ‘distance’ error



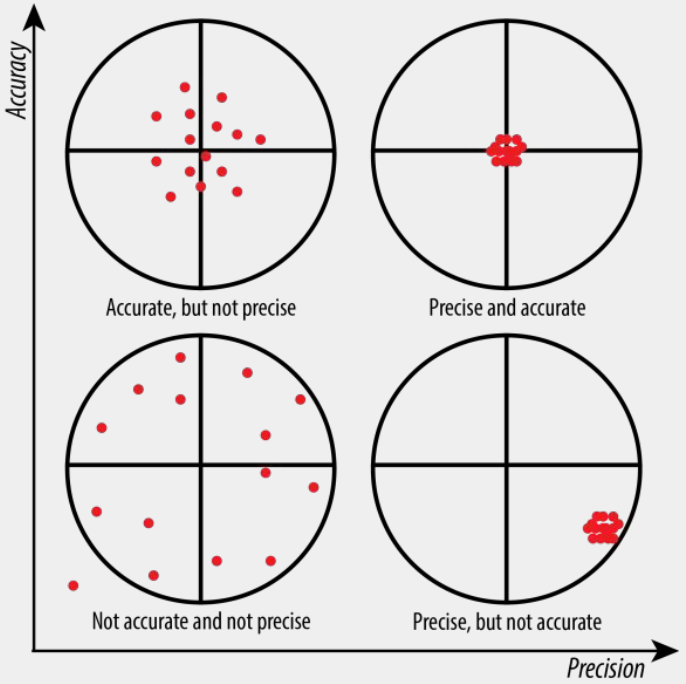
- Task dependent:
  - Imaging - pixel-based classification errors (e.g. Dice Coefficient)

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Model selection - General Guidelines

- Pick the model with the ‘best’ cross-validation performance
  - Is a 0.1% increase in performance meaningful?
  - Always use statistical tests to identify *real* differences

- Other considerations: Complexity, explainability, training time (cost), amount of labeled data



SCIENTIFIC REPORTS
nature research

OPEN

Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer

Marta Bogowicz[1,2,16*], Arthur Jochems[2,16], Timo M. Deist[2], Stephanie Tanadini-Lang[1], Shao Hui Huang[3], Biu Chan[3], John N. Waldron[3], Scott Bratman[3], Brian O'Sullivan[3], Oliver Riesterer[1,4], Gabriela Studer[1,5], Jan Unkelbach[1], Samir Barakat[2], Ruud H. Brakenhoff[6], Irene Nauta[6], Silvia E. Gazzani[7], Giuseppina Calareso[8], Kathrin Scheckenbach[9], Frank Hoebers[10], Frederik W. R. Wesseling[10], Simon Keek[2], Sebastian Sanduleanu[2], Ralph T. H. Leijenaar[2], Marije R. Vergeer[11], C. René Leemans[6], Chris H. J. Terhaard[12], Michiel W. M. van den Brekel[13], Olga Hamming-Vrieze[14], Martijn A. van der Heijden[13], Hesham M. Elhalawani[15], Clifton D. Fuller[15], Matthias Guckenberger[1] & Philippe Lambin[2]

PICO

P: Patients at any of 6 cohort hospitals with head and neck tumors, CECT, and HPV histology

I: Prediction of HPV status and overall survival with AI algorithm

C: No comparator

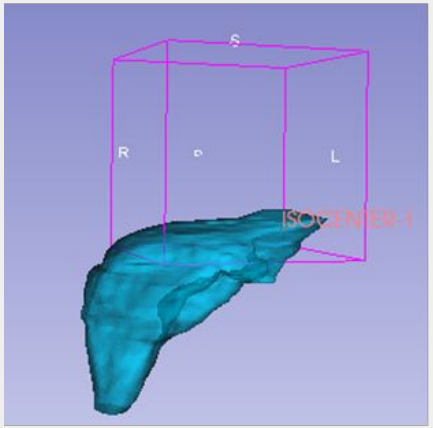
O: Prediction accuracy/sens/spec/auc

What methods did they use?

Radiomics analysis. Radiomic features were extracted from the primary tumor region. The treatment defined gross tumor volume (GTV) was visually assessed for the presence of artifacts and slices with artifacts were manually removed from the contour. Images were resampled to 3.3 mm cubic voxels using linear interpolation. The Hounsfield unit range was set to (−20, 180) to limit the analysis to soft tissue. In total, 981 features were extracted with the Z-Rad radiomics software implementation[17]:

- shape (n = 18).
- intensity distribution (n = 17).
- texture (n = 90): the Gray Level Co-occurrence Matrix (n = 26), the Neighborhood Gray Tone Difference Matrix (n = 4), the Gray Level Run Length Matrix (n = 14), the Gray Level Size Zone Matrix (n = 14), the Gray Level Distance Zone Matrix (n = 16) and the Neighboring Gray Level Dependence Matrix (n = 16).
- wavelet transform (n = 856).

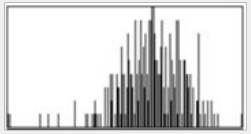
Radiomics



1.

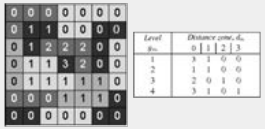
$$skewness = \frac{\mu_3}{\sigma^3} = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^3}{\left(\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2}\right)^3}$$

2.



$$entropy = -\sum_{i=1}^{N_g} p(i)\log_2\left(p(i)+\epsilon\right)$$

108.



$$GLV = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j|\theta)(i-\mu)^2$$

Feature selection. First, data quality check was performed. Missing values were assessed and features with more than 20% missing values were excluded. Similarly, to avoid outliers, features with skewed distribution (skewness > 5) were excluded. The exclusion criteria were evaluated in the entire dataset for the centralized learning and per cohort for distributed learning. In the distributed learning, the union of features excluded per cohort was considered as the excluded subset.

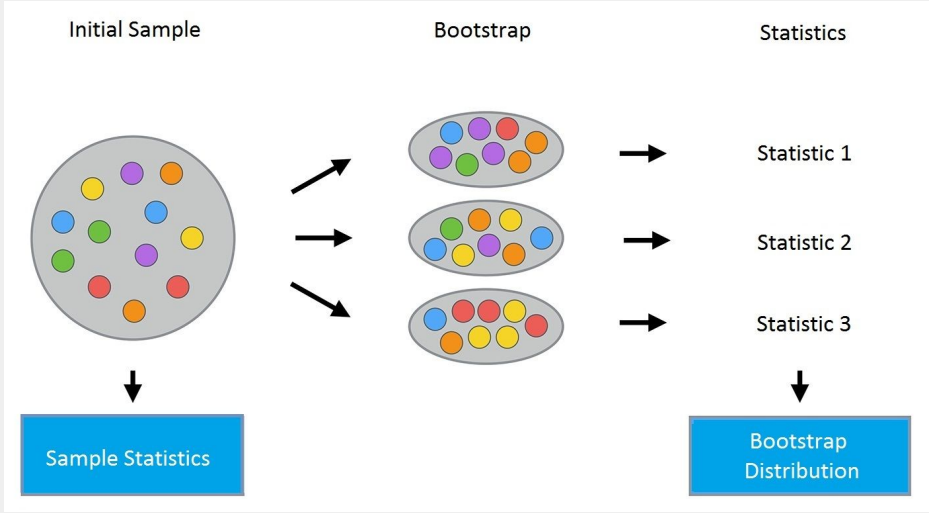
Next, inter-features correlations were assessed (Fig. 1). Features were scaled with the z-score. In distributed learning, the global mean and standard deviation per feature were obtained by sharing local statistics on mean, dispersion from mean and number of patients in the cohort. The global correlations were estimated as weighted average of fisher transformed local correlation coefficients. The average linkage hierarchical clustering (Python SciPy library v. 1.3.0) was performed on the set of inter-features correlation coefficients with a 0.6 cutoff, separately for the centralized and distributed learning.

Finally, to select a feature representative per cluster a univariate logistic regression was performed on the entire dataset (centralized learning) as well as the separate cohorts (distributed learning). In the centralized learning, per cluster, the feature with the highest area under the receiver operator characteristic curve (AUC) was chosen if the false discovery rate <0.05. In the distributed learning, per cohort and per cluster, the feature with the highest AUC was chosen to represent each cluster. In the central sever the cohort-specific sets were compared

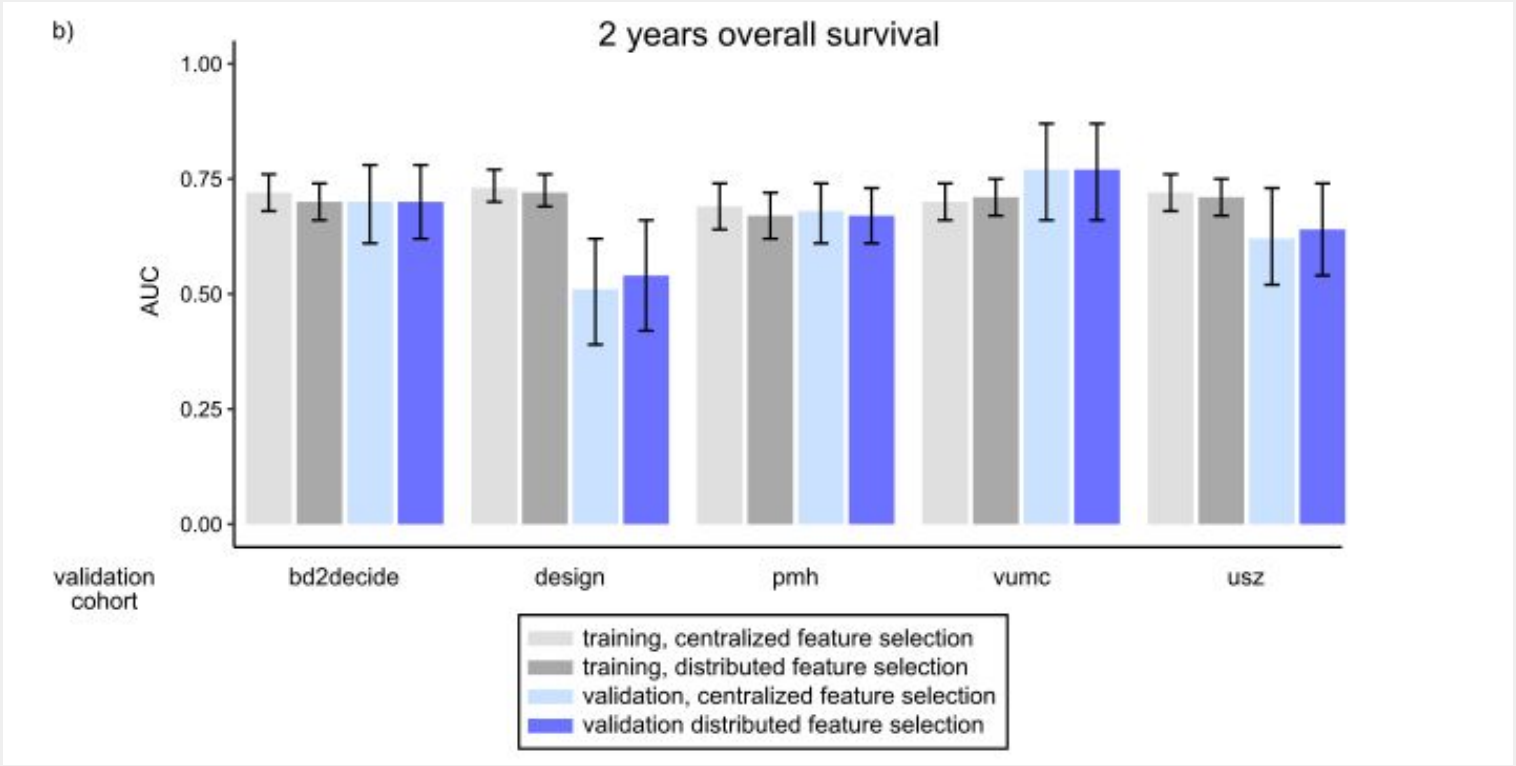
Classification. A multivariate logistic regression model was trained for both outcomes, HPV and 2 year overall survival (2yOS). In the centralized learning, the model was fitted with a GLM (generalized linear models)

Comparison of the models. Five models were created to predict HPV status and another five to predict 2yOS. For each of the models, four cohorts were used for training and one was left out for external validation (patients with unknown status were excluded from modeling of the respective outcome). The prognostic power of a model was evaluated in the validation cohort. Models were trained in a distributed and centralized manner for comparison.

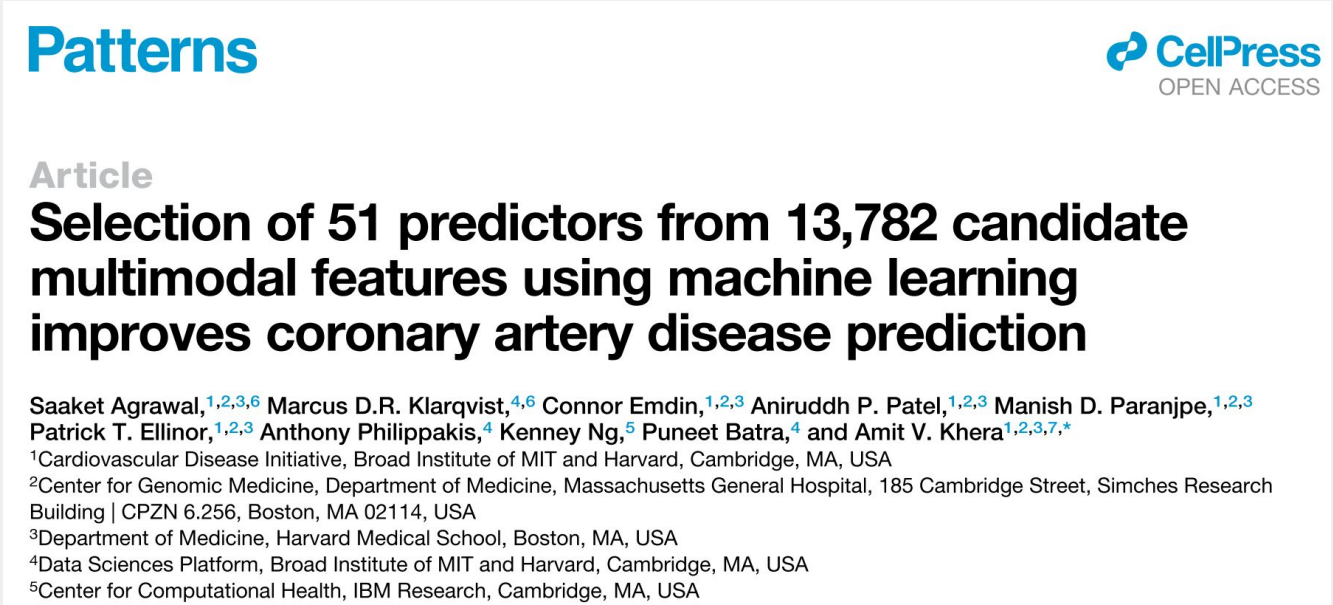
Cross-Validation or Bootstrapping!!!



Results


b)
2 years overall survival
1.00
0.75
0.50
0.25
0.00
AUC
validation cohort
bd2decide
design
pmh
vumc
usz
training, centralized feature selection
training, distributed feature selection
validation, centralized feature selection
validation distributed feature selection

Motivating Example



Patterns

CellPress
OPEN ACCESS

Article

Selection of 51 predictors from 13,782 candidate multimodal features using machine learning improves coronary artery disease prediction

Saaket Agrawal,[1,2,3,6] Marcus D.R. Klarqvist,[4,6] Connor Emdin,[1,2,3] Aniruddh P. Patel,[1,2,3] Manish D. Paranjpe,[1,2,3] Patrick T. Ellinor,[1,2,3] Anthony Philippakis,[4] Kenney Ng,[5] Puneet Batra,[4] and Amit V. Khera[1,2,3,7,*]

[1]Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[2]Center for Genomic Medicine, Department of Medicine, Massachusetts General Hospital, 185 Cambridge Street, Simches Research Building | CPZN 6.256, Boston, MA 02114, USA

[3]Department of Medicine, Harvard Medical School, Boston, MA, USA

[4]Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA

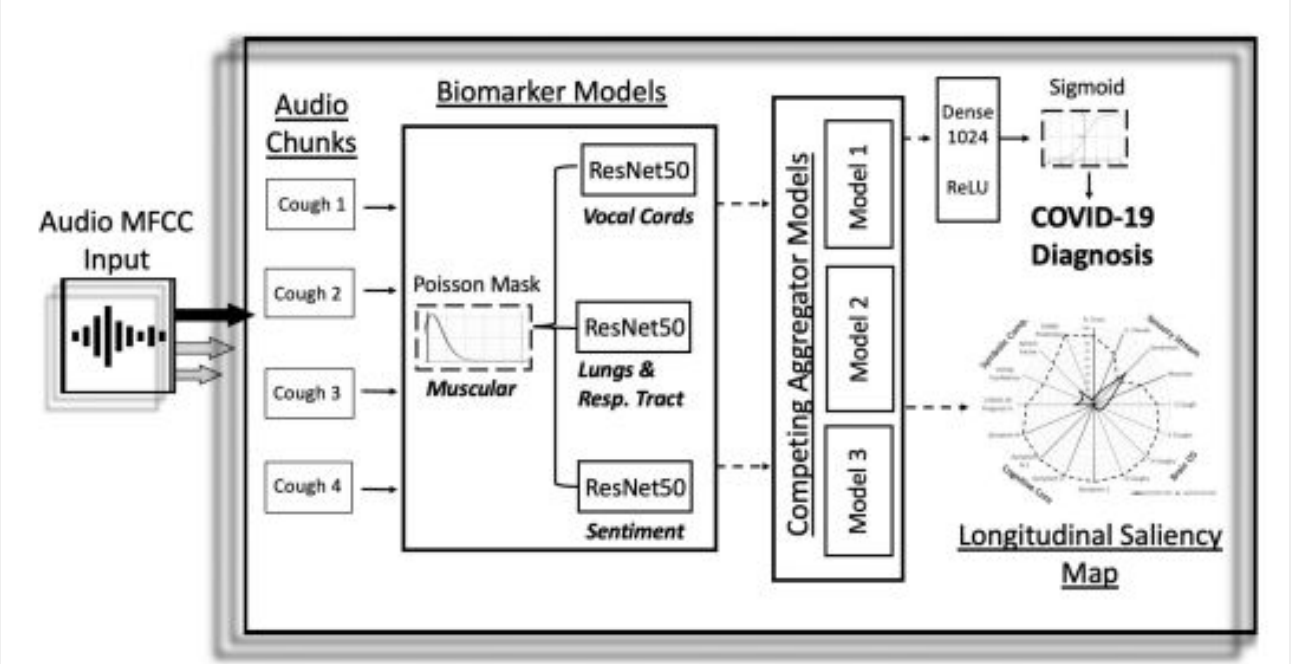
[5]Center for Computational Health, IBM Research, Cambridge, MA, USA

COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings

Jordi Laguarta, Ferran Hueto, and Brian Subirana

Classification problem: classify COVID from cough

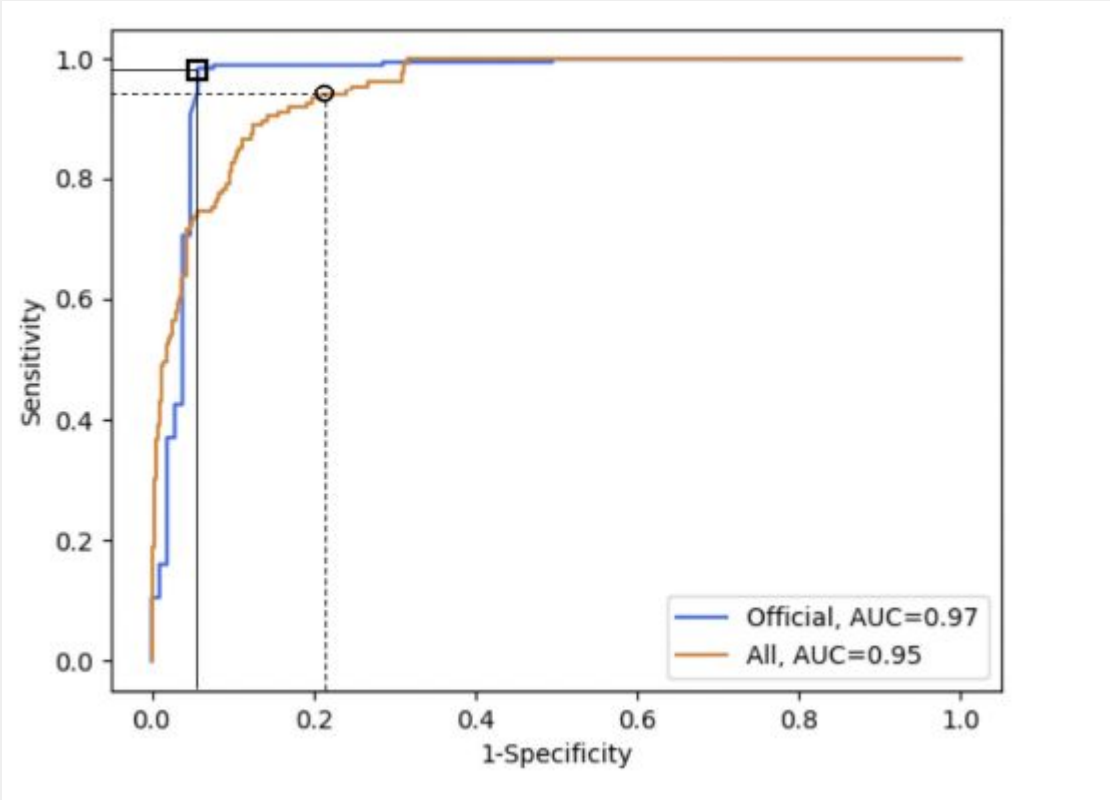
on Alzheimer's, which significantly improves the COVID-19 discrimination accuracy of our architecture. *Results:* When validated with subjects diagnosed using an official test, the model achieves COVID-19 sensitivity of 98.5% with a specificity of 94.2% (AUC: 0.97). For asymptomatic subjects it achieves sensitivity of 100% with a specificity of 83.2%. *Conclusions:* AI techniques can produce a free, non-invasive, real-time, any-time, instantly distributable, large-



Biomarker	Model Name	COVID(%)	Alzheimer(%)
R. Tract	Cough	23	9
Sentiment	Intonation	8	19
Vocal cords	WW 'THEM'	19	16
R.Tract&Sent.	Cough&Tone.	0	0
R.Tract&V.cords	Cough&WW	1	6
Sent.&V.cords	Tone.&WW	0	3
In All 3		34	41
In Neither 3		15	6

train and validate the COVID-19 discriminator. 4256 subjects (80%) were used for training and 1064 (20%) for validation. Table I provides more details on the patient distribution for the randomly sampled patients selected from the dataset.

No cross-validation, but sample size is large


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1-Specificity
Official, AUC=0.97
All, AUC=0.95

but…

	Positives			Negatives			Total	
	#	%	Hit(%)	#	%	Hit(%)	%	Hit(%)
Number of Patients	2660	50.0	94.0	2660	50.0	78.4	100.0	86.4
COVID-19 Diagnostic								
Official Test	475	17.9	98.5	224	8.4	94.2	13.1	97.1
Doctor Assessment	962	36.2	98.8	523	19.7	92.8	27.9	96.7
Personal Assessment	1223	46.0	89.5	1913	71.9	72.6	58.9	79.2
Symptoms								
No Symptoms 'Official'	102	3.8	100.0	114	4.3	83.2	4.1	91.1
No Symptoms	196	7.4	100.0	2029	76.3	78.3	41.8	80.2
Fever	656	24.7	98.1	34	1.3	88	13.0	97.6
Tiredness	1428	53.7	93.2	210	7.9	81.2	30.8	91.7
Sore Throat	1064	40.0	99.9	205	7.7	84.8	23.9	97.5
Diff. Breathing	680	25.6	99.3	49	1.8	77.1	13.7	97.8
Chest Pain	680	25.6	99.4	58	2.2	87.3	13.9	98.4
Diarrhea	652	24.5	94.2	100	3.8	71.8	14.1	91.2
Cough	1724	64.8	99.8	262	9.8	91.1	37.3	98.7

but…

	Positives			Negatives			Total	
	#	%	Hit(%)	#	%	Hit(%)	%	Hit(%)
No Symptoms 'Official'	102	3.8	100.0	114	4.3	83.2	4.1	91.1
Cough	1724	64.8	99.8	262	9.8	91.1	37.3	98.7

What is this study missing?

- Diverse dataset
- Resampling, either cross-validation or bootstrapping
- Comparison with status quo methods (RAT, or a classification rule of flu-like symptoms = COVID)

Is the study valuable?

- “Proof of concept” - **is there a signal?**