Lecture 3: Supervised machine learning and traditional approaches

BIODS388/BIOMED388

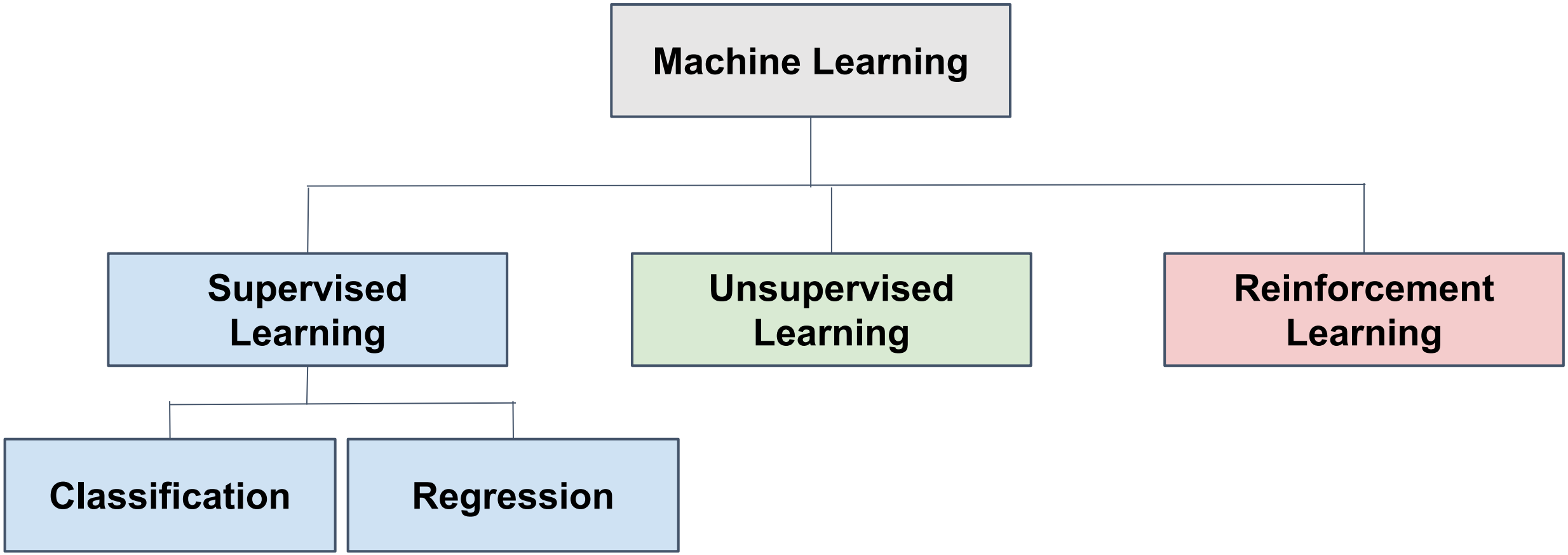
Anuj Pareek MD PhD, Mars Huang PhD Student

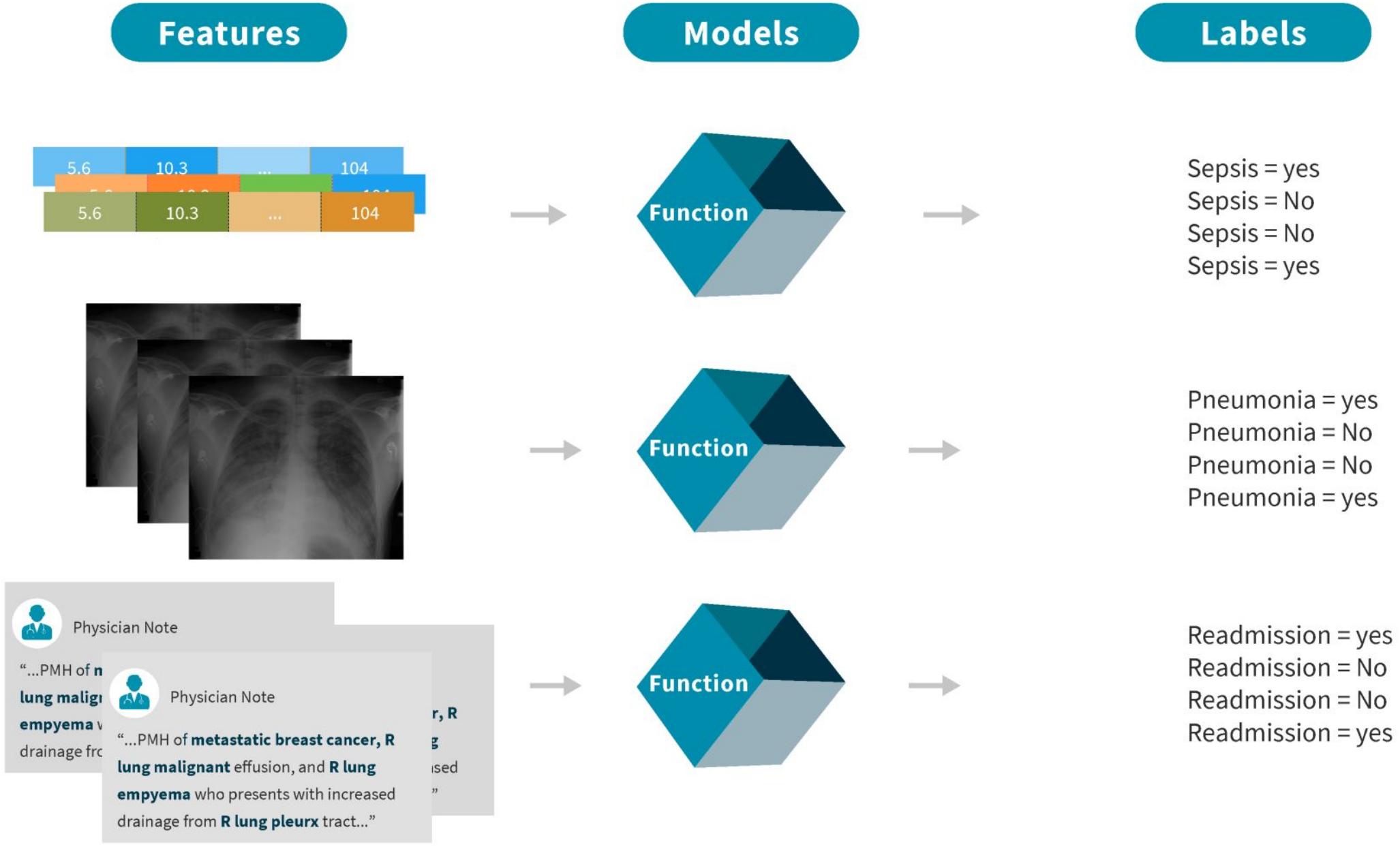
10/01/2020

Outline

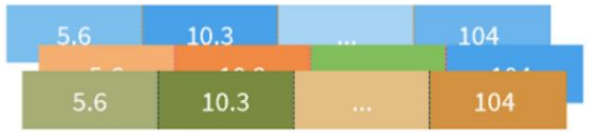
1. **Lecture 2: Recap**
2. K-nearest neighbor
3. Regression
4. Support Vector Machines (SVMs)
5. Tree-based algorithms
6. Applications

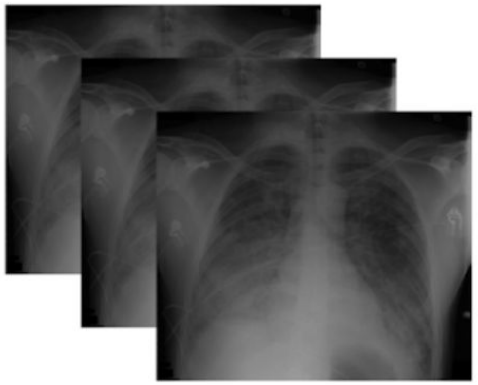
Machine Learning

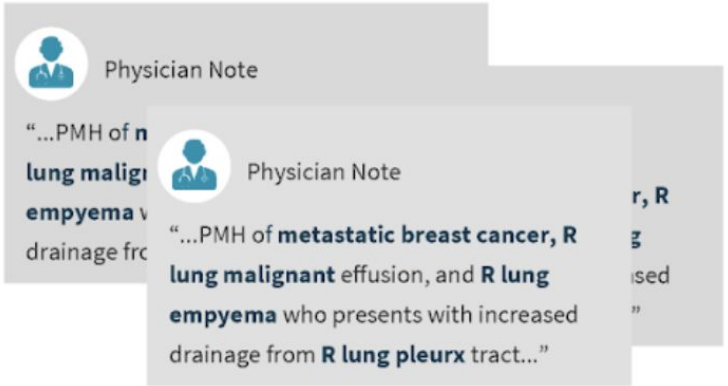



Features
Models
Labels
5.6
10.3
...
104
5.6
10.3
...
104
Function
Sepsis = yes
Sepsis = No
Sepsis = No
Sepsis = yes
Function
Pneumonia = yes
Pneumonia = No
Pneumonia = No
Pneumonia = yes
Physician Note
"...PMH of n
lung malig
empyema
drainage fr
Physician Note
"...PMH of metastatic breast cancer, R lung malignant effusion, and R lung empyema who presents with increased drainage from R lung pleurx tract..."
Function
Readmission = yes
Readmission = No
Readmission = No
Readmission = yes

Features







STRUCTURED

UNSTRUCTURED

Labels

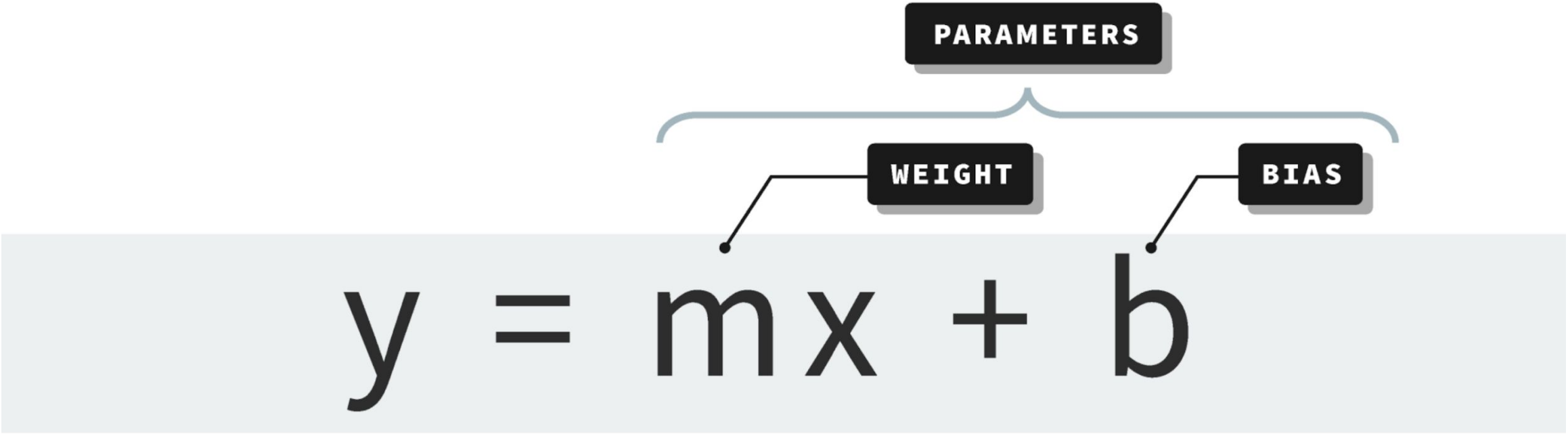
Sepsis = yes
Sepsis = No
Sepsis = No
Sepsis = yes

Pneumonia = yes
Pneumonia = No
Pneumonia = No
Pneumonia = yes

Readmission = yes
Readmission = No
Readmission = No
Readmission = yes

REGRESSION

CLASSIFICATION


PARAMETERS
WEIGHT
BIAS
y = mx + b

Outline

1. Lecture 2: Recap
2. **K-nearest neighbor**
3. Regression
4. Support Vector Machines (SVMs)
5. Tree-based algorithms
6. Applications

K-Nearest Neighbor - Classification

Fruit

Fruit

Fruit

Fruit

Animal

Animal

Animal

Animal

K-Nearest Neighbor - Classification

Fruit

Fruit

Fruit

Fruit

?

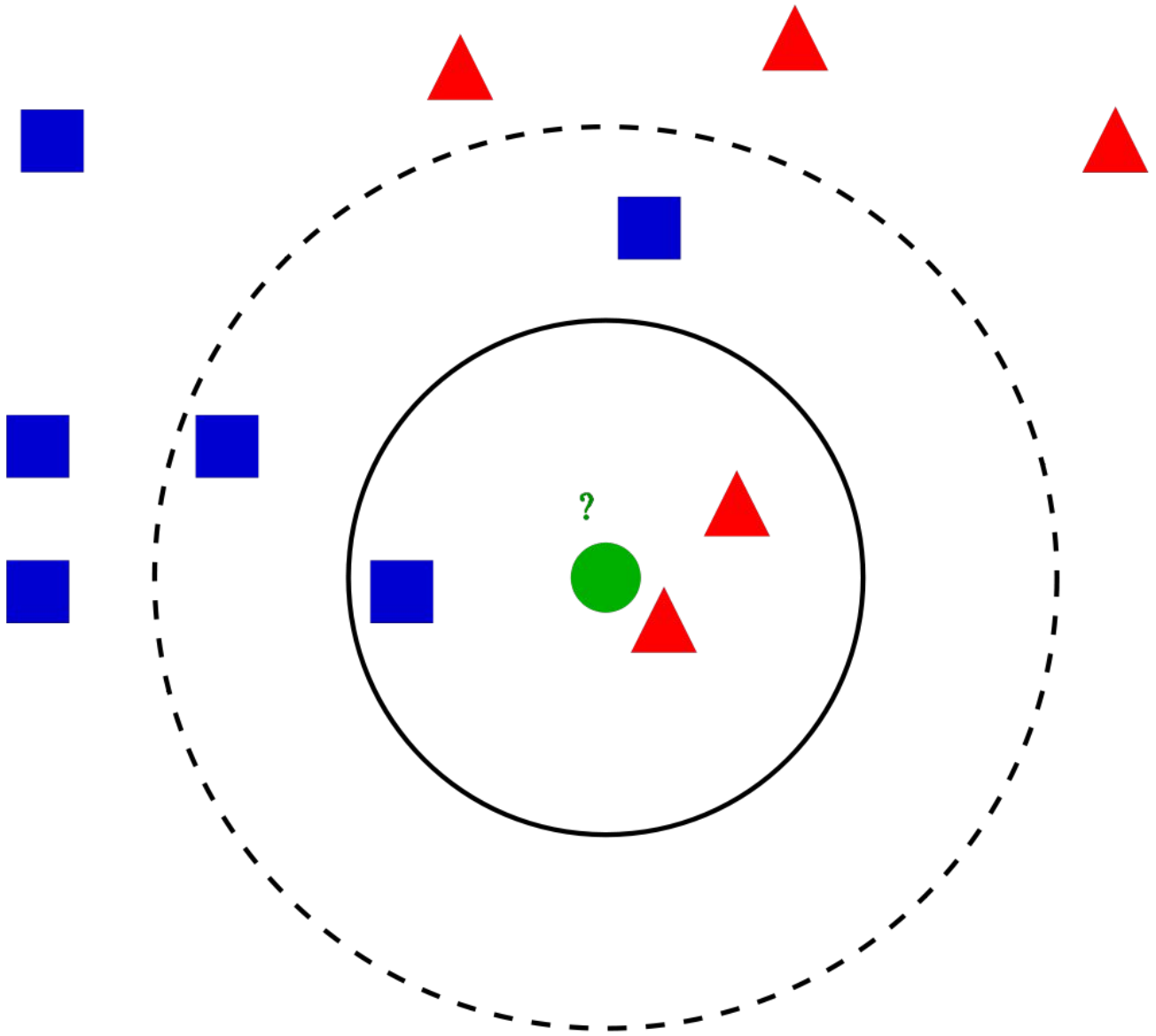
Animal

Animal

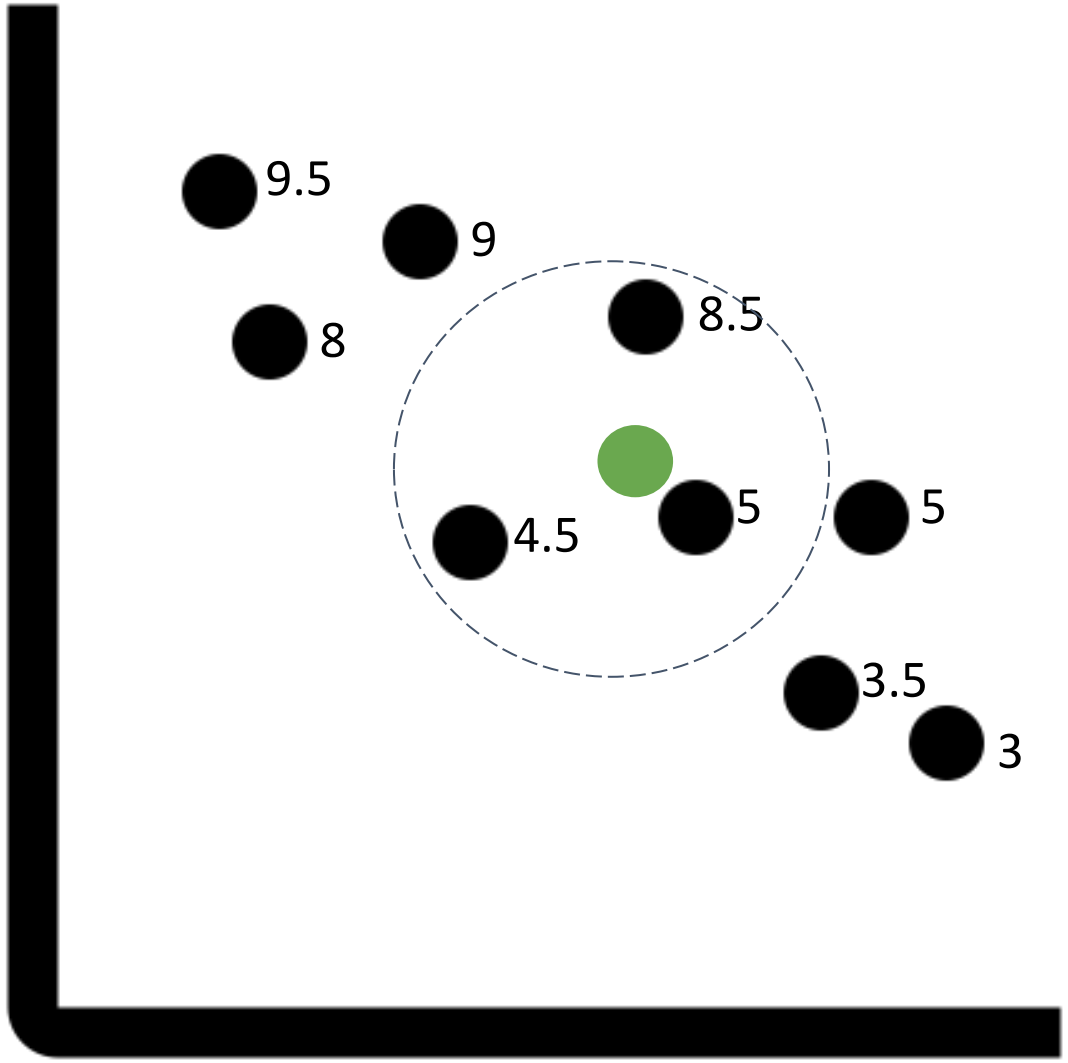
Animal

Animal

K-Nearest Neighbor - Classification



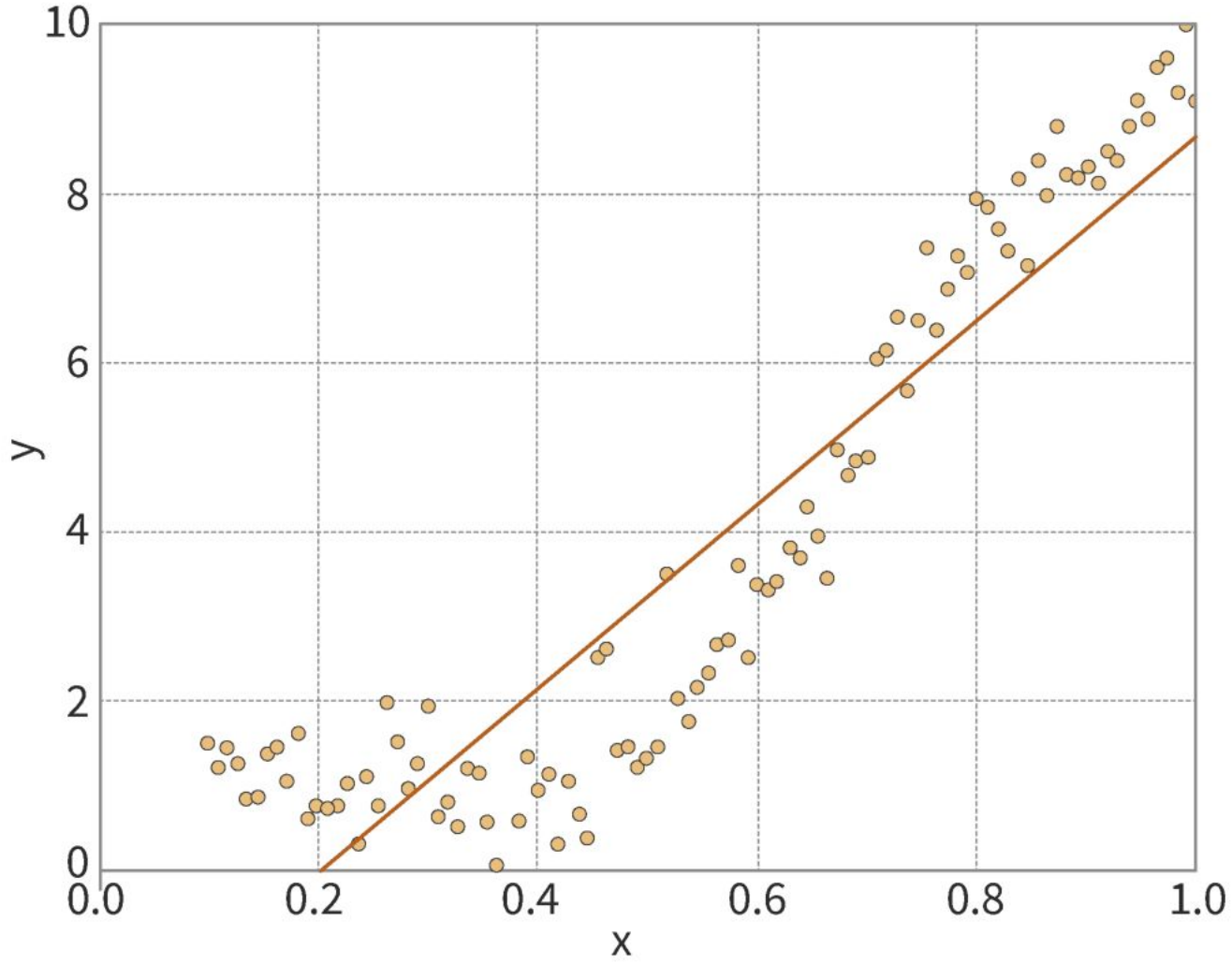
K-Nearest Neighbor - Regression



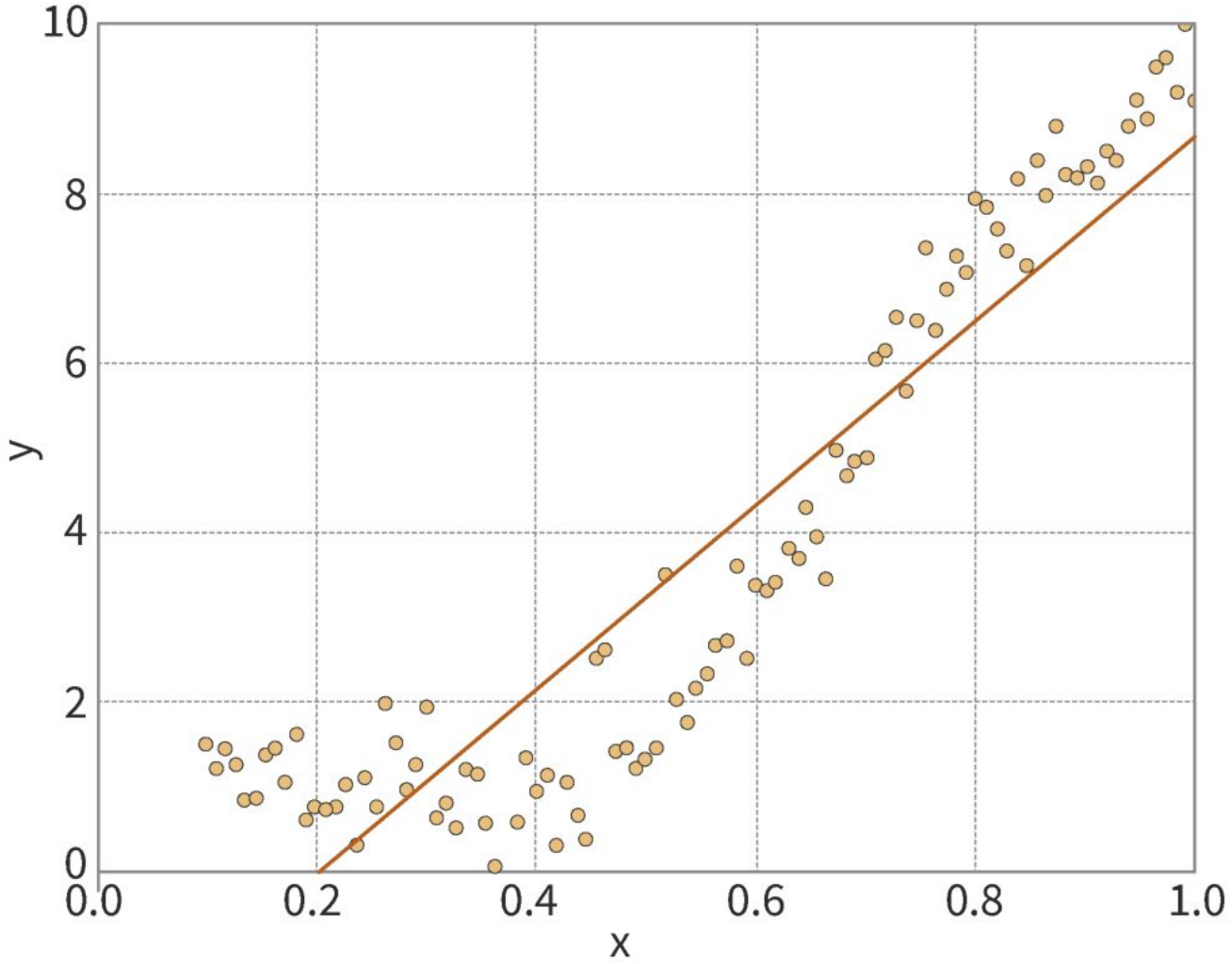
Outline

1. Lecture 2: Recap
2. K-nearest neighbor
3. **Regression**
4. Support Vector Machines (SVMs)
5. Tree-based algorithms
6. Applications

WHAT WE'VE ALREADY SEEN: LINEAR REGRESSION



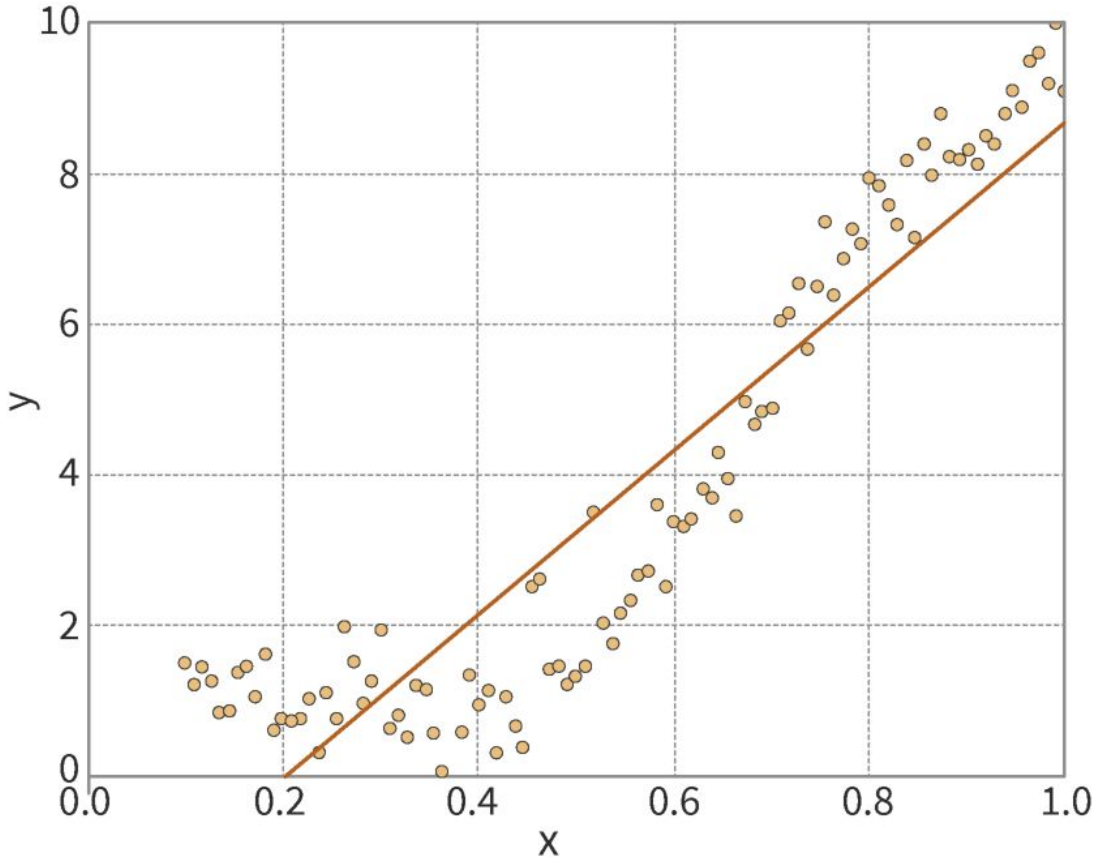
WHAT WE'VE ALREADY SEEN: LINEAR REGRESSION

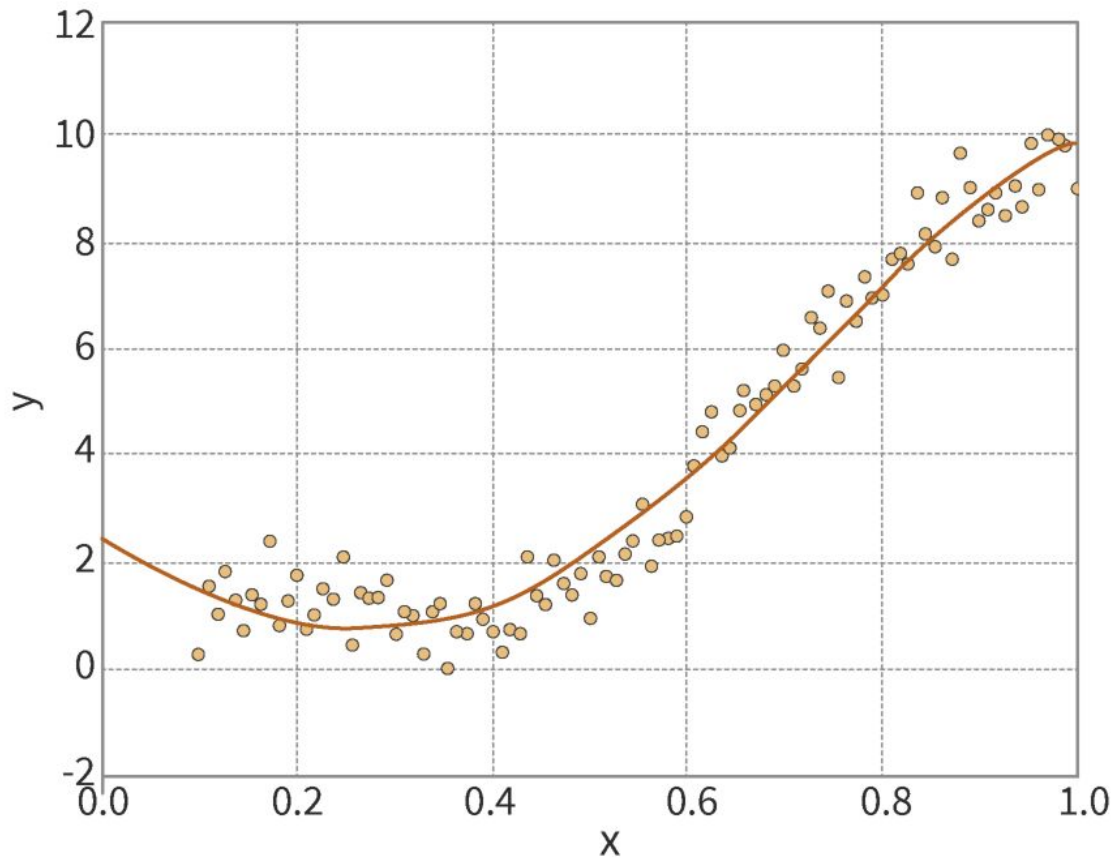


What if a curve would be a better fit?

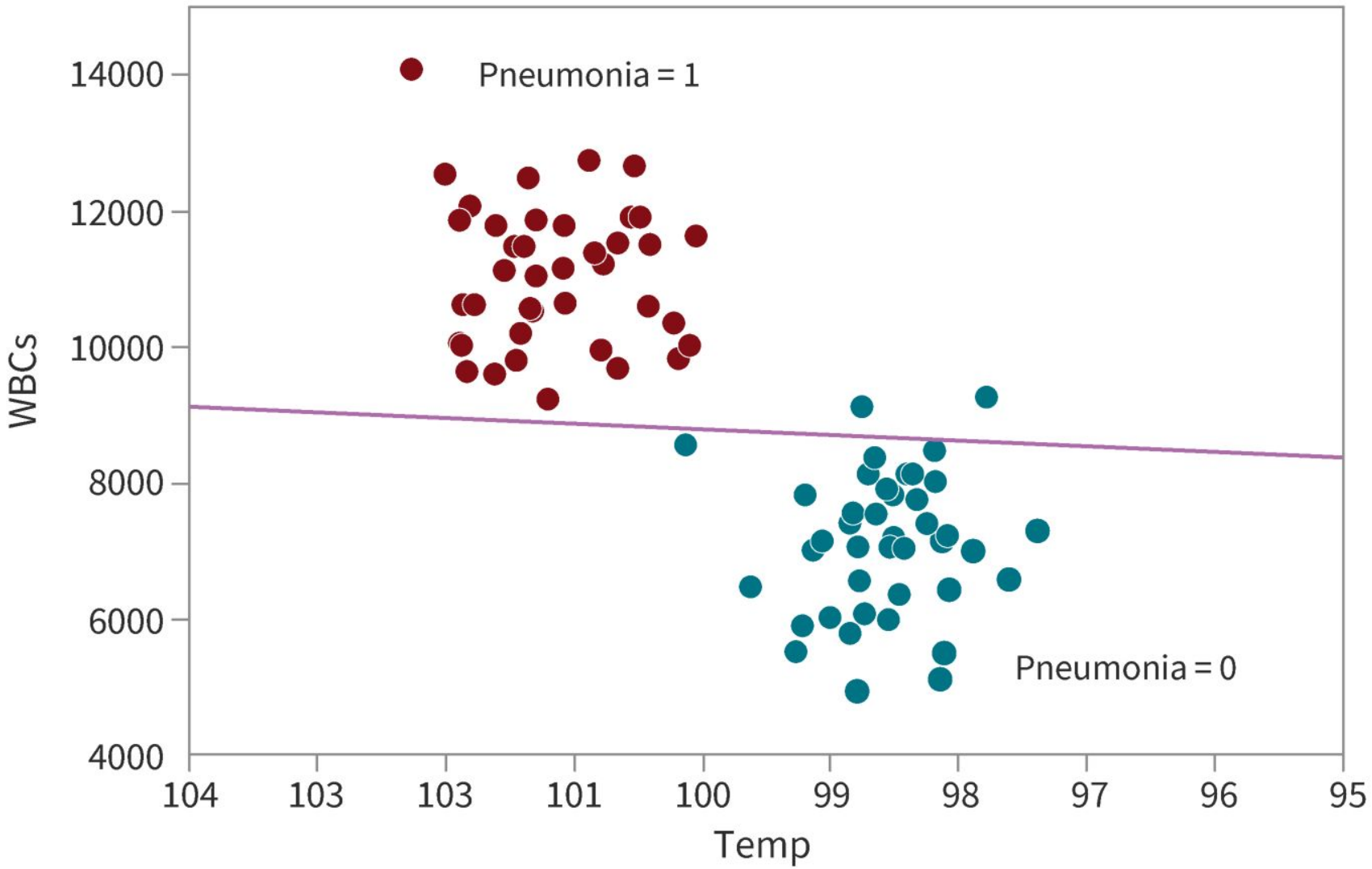
ONE VARIANT: POLYNOMIAL REGRESSION

Choosing the right model for your data is important!

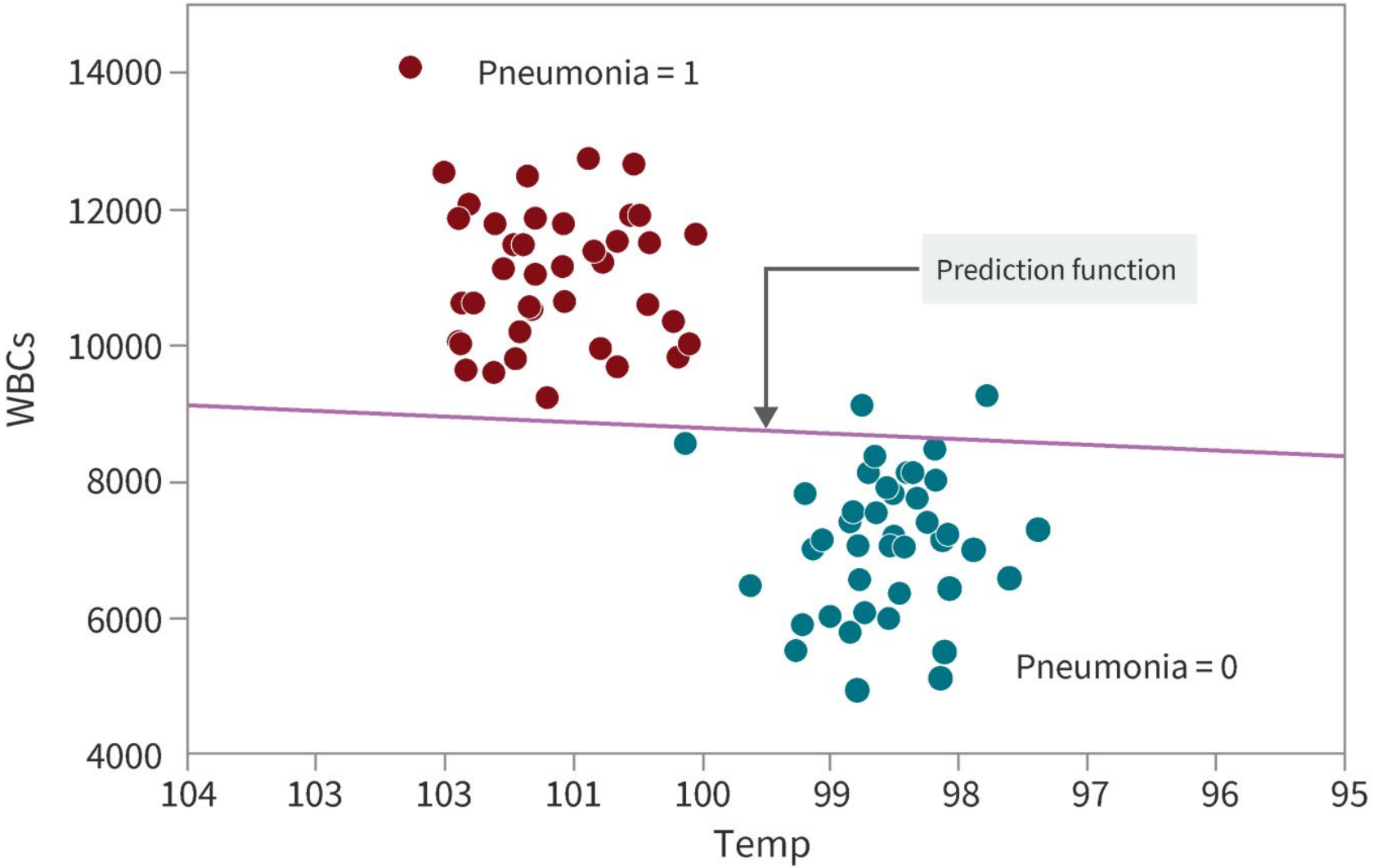


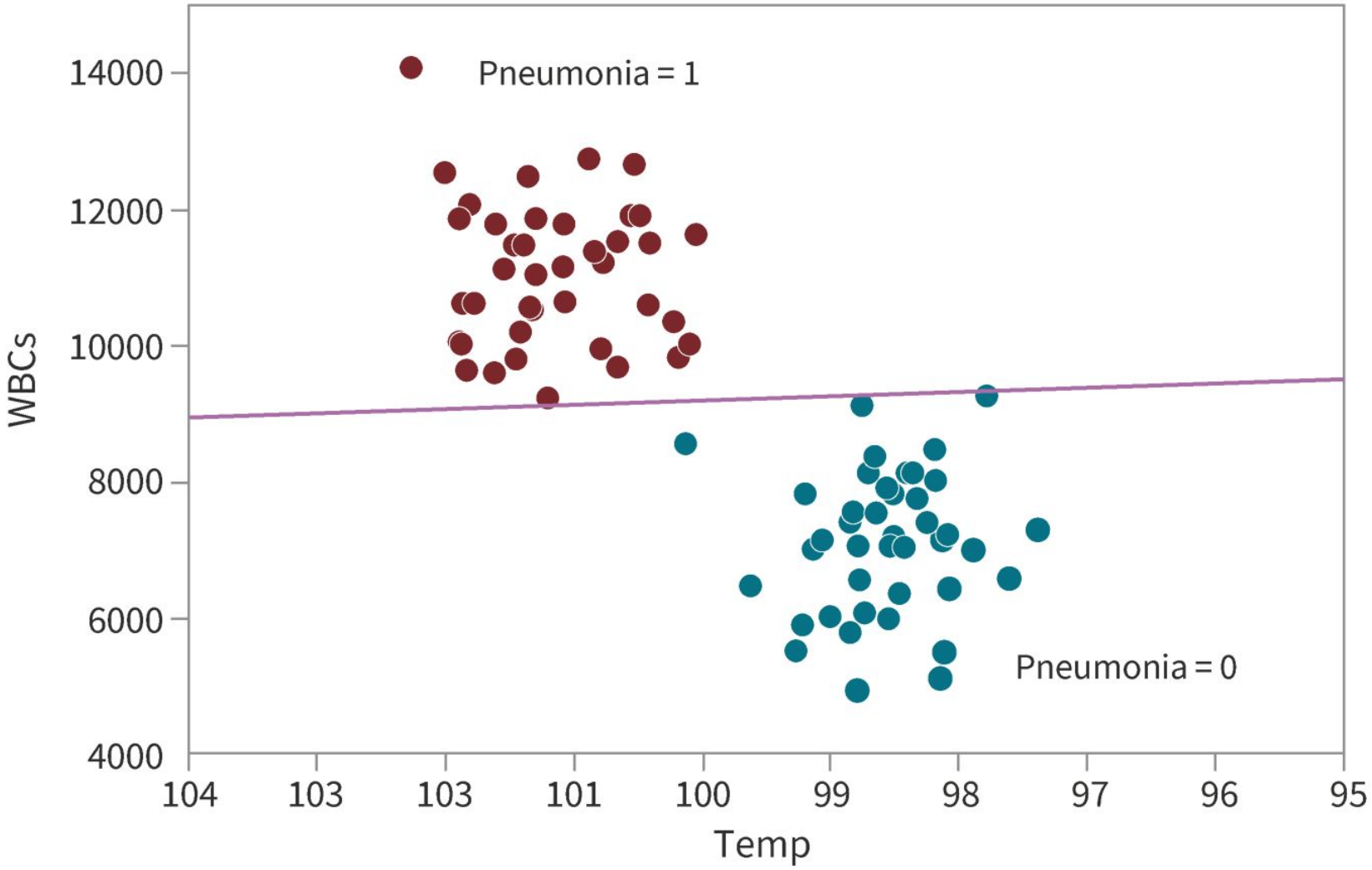


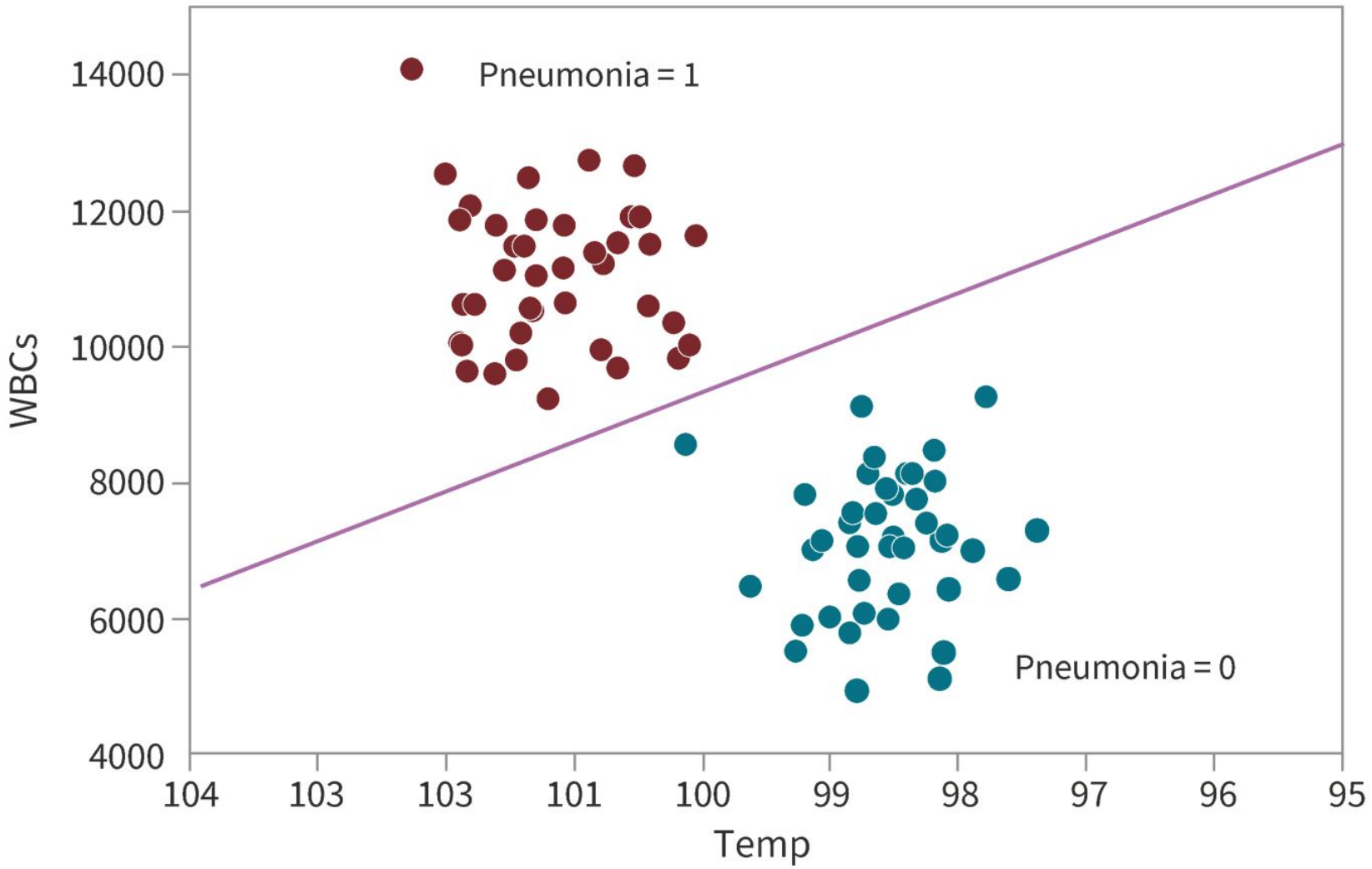
REMEMBER: LOGISTIC REGRESSION



REMEMBER: LOGISTIC REGRESSION



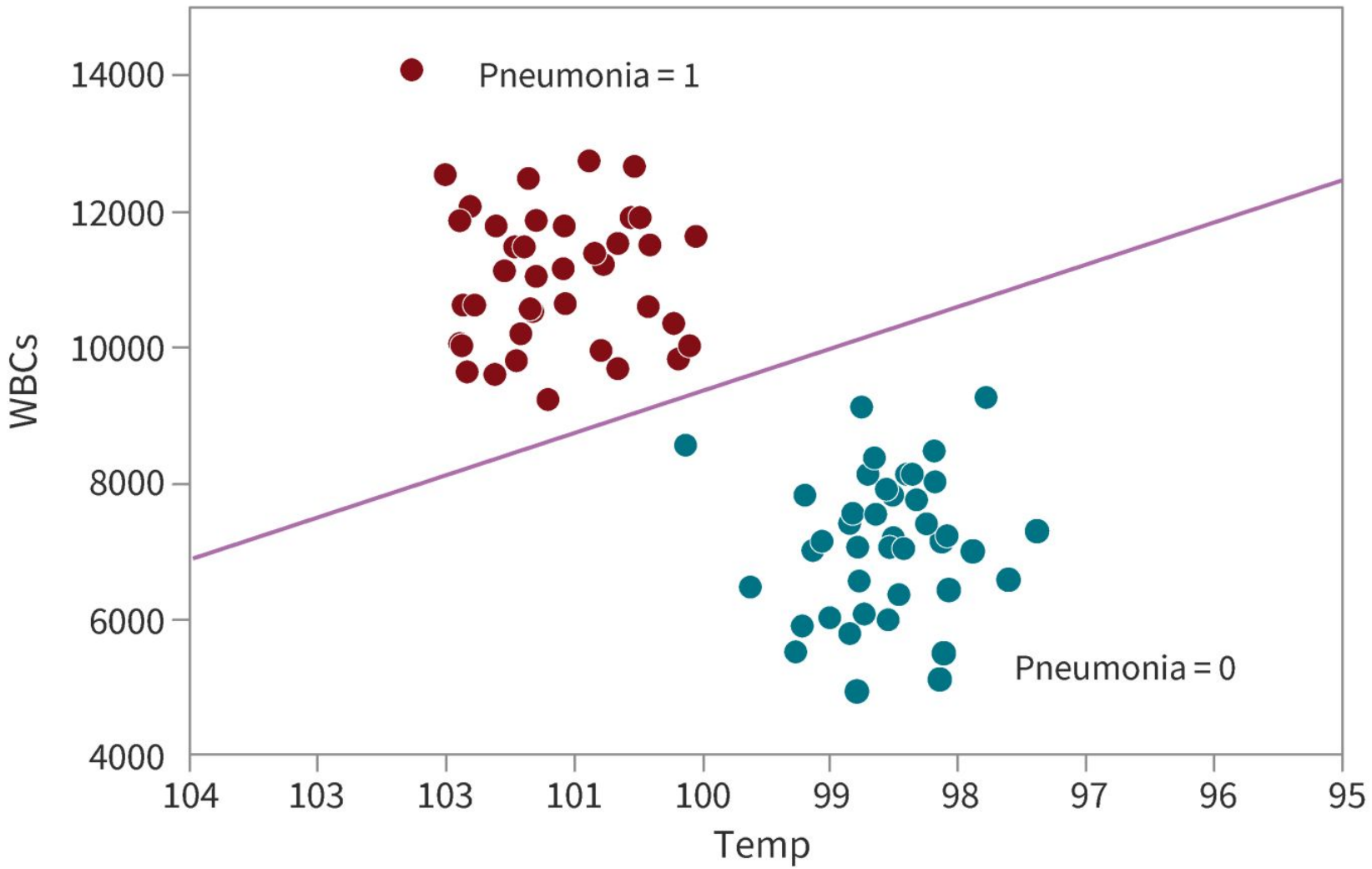

Pneumonia = 1
Pneumonia = 0
WBCs
14000
12000
10000
8000
6000
4000
104
103
103
101
100
99
98
97
96
95
Temp


Pneumonia = 1
Pneumonia = 0
WBCs
14000
12000
10000
8000
6000
4000
104
103
103
101
100
99
98
97
96
95
Temp

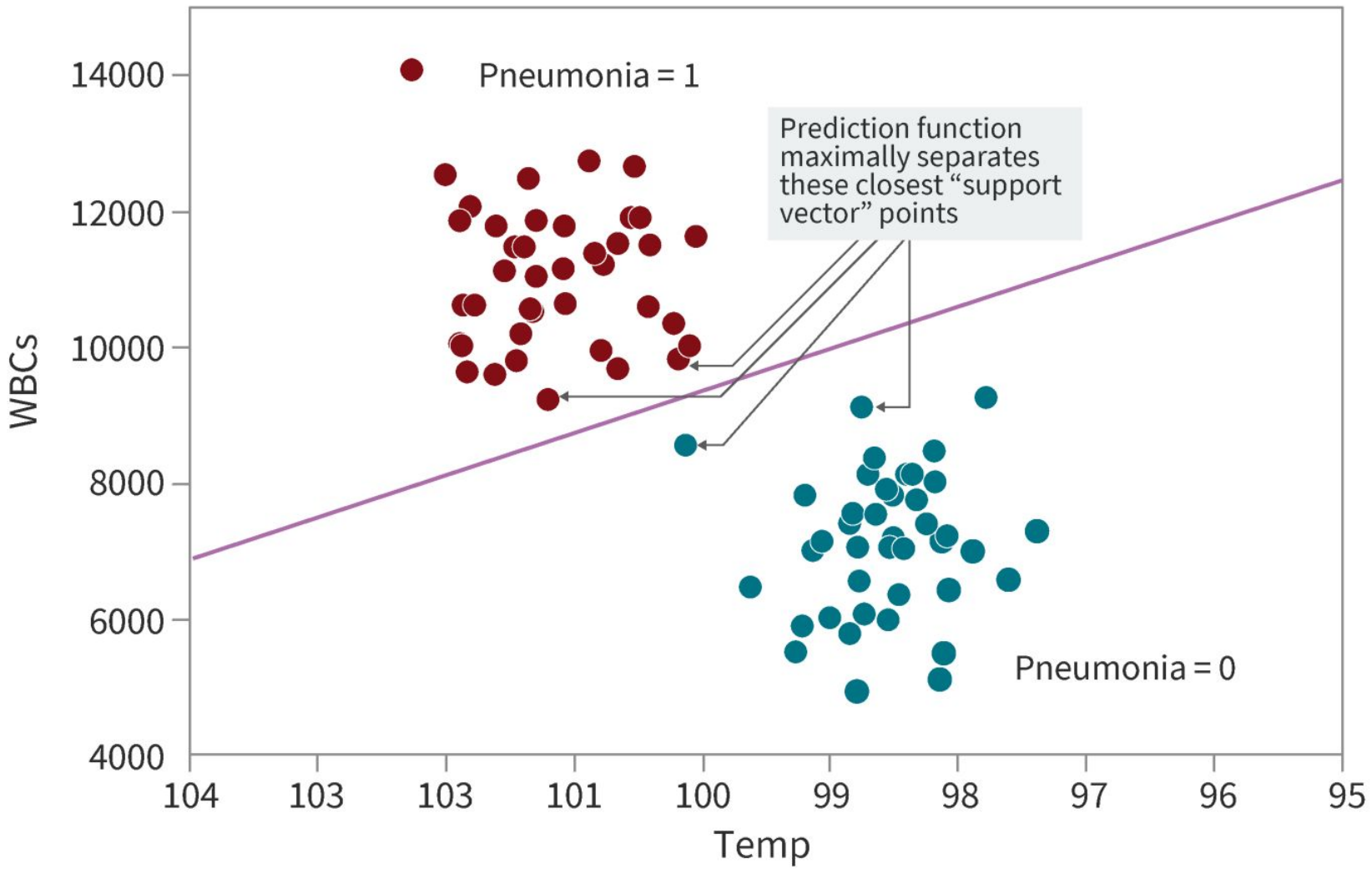
Outline

1. Lecture 2: Recap
2. K-nearest neighbor
3. Regression
4. **Support Vector Machines (SVMs)**
5. Tree-based algorithms
6. Applications

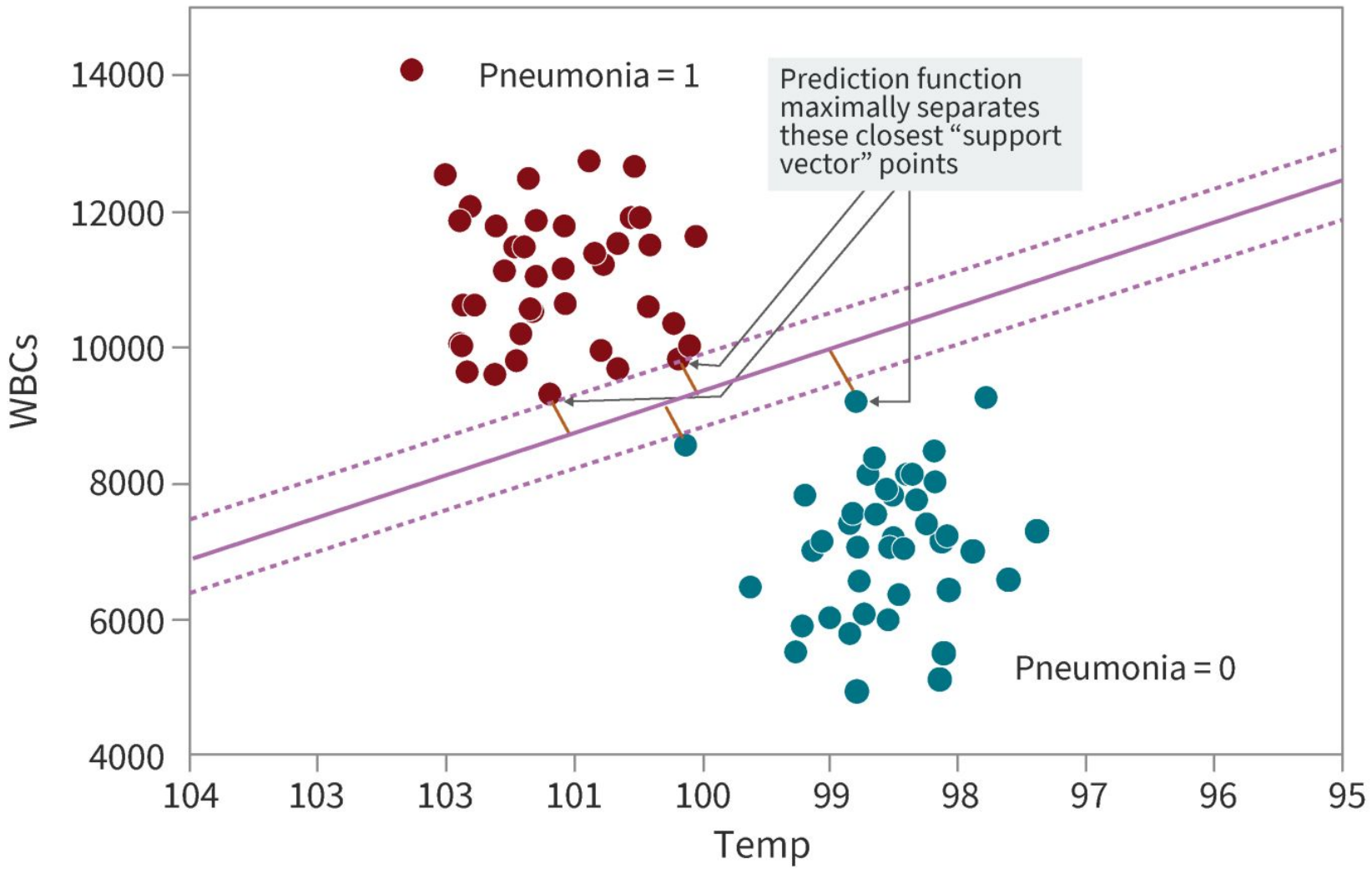
SUPPORT VECTOR MACHINES (SVMs)

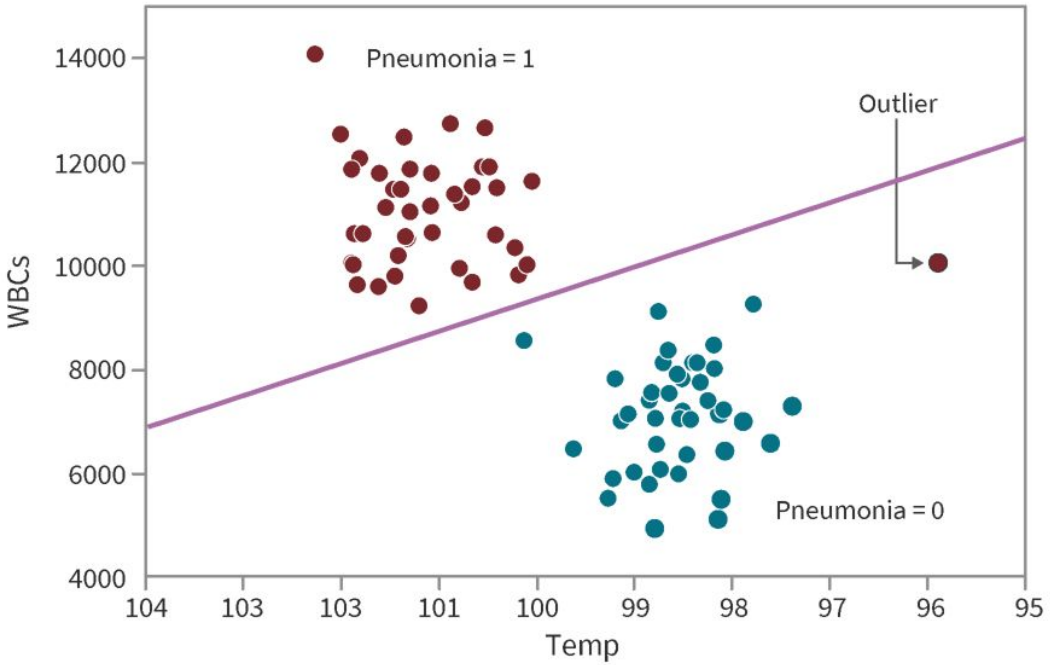


SUPPORT VECTOR MACHINES (SVMs)

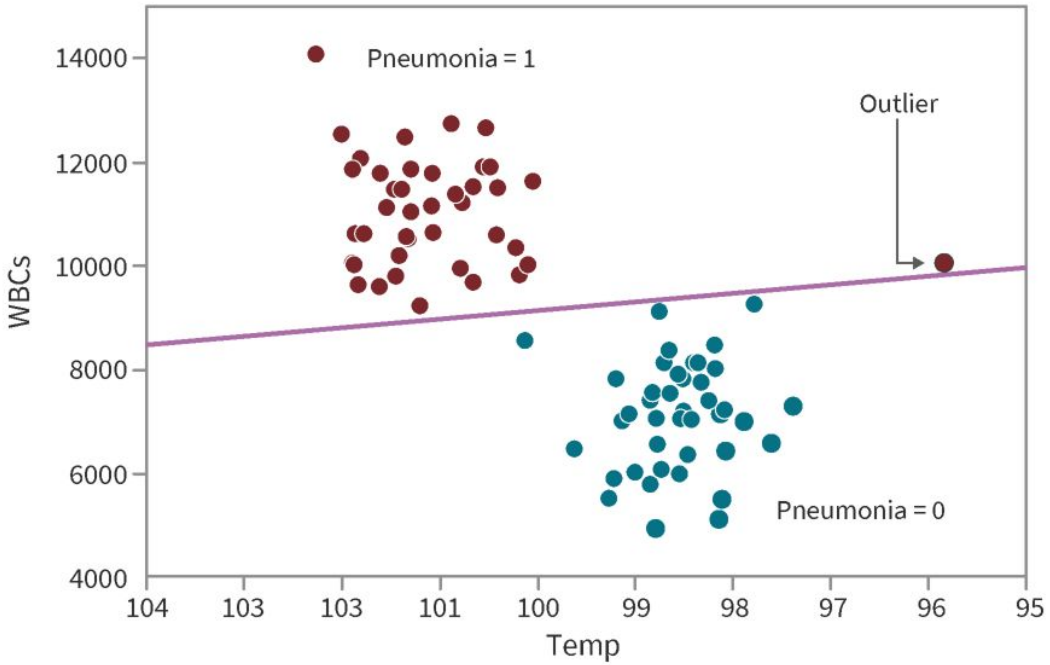


SUPPORT VECTOR MACHINES (SVMs)

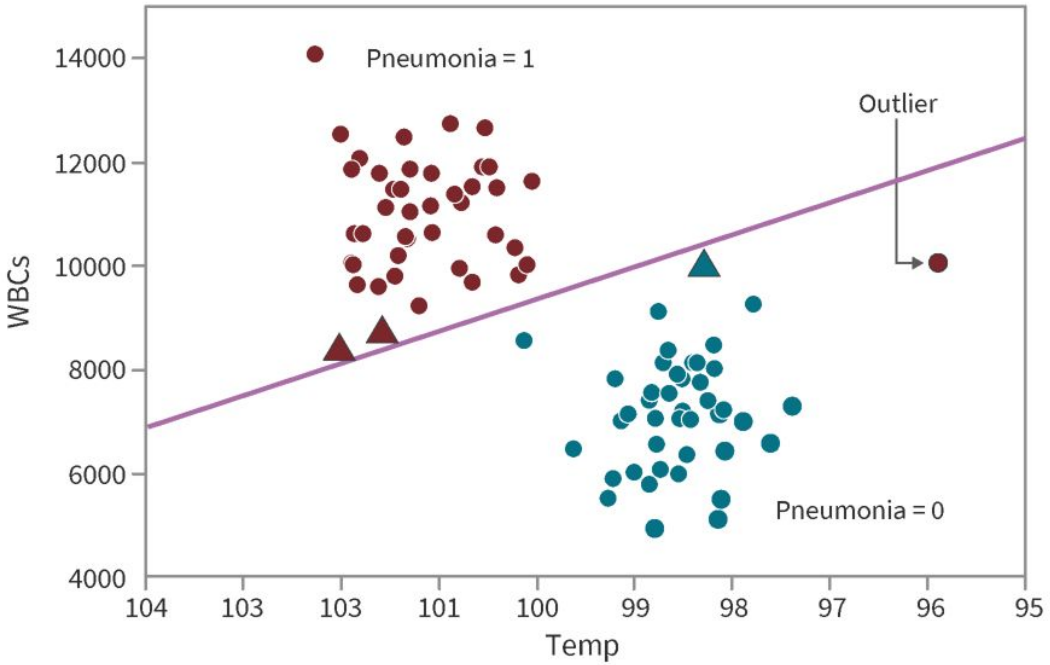




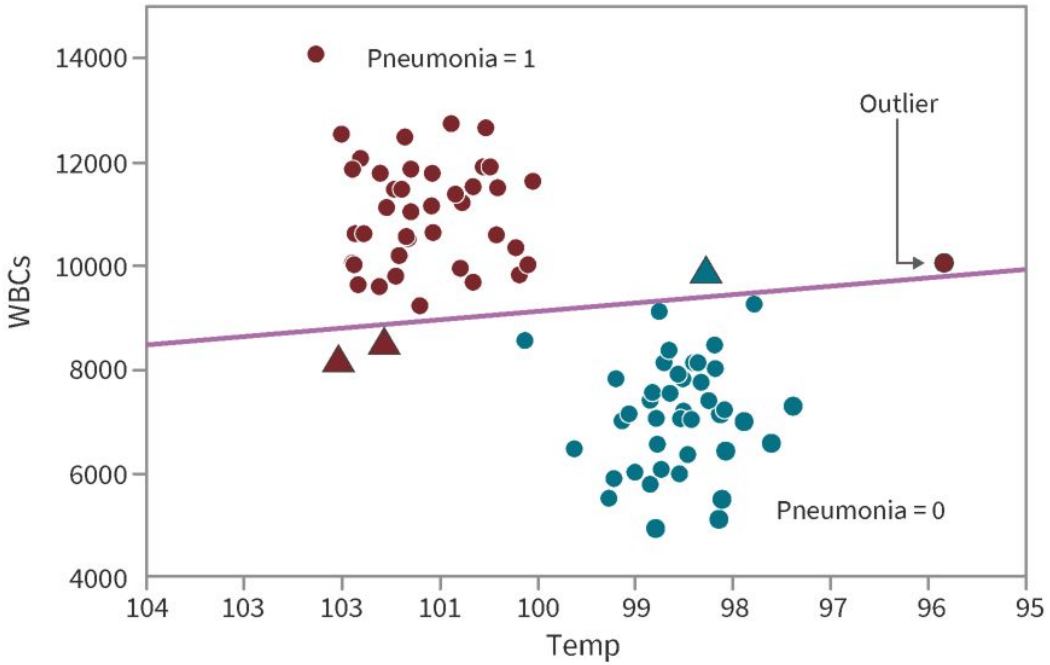
SVM



Logistic Regression

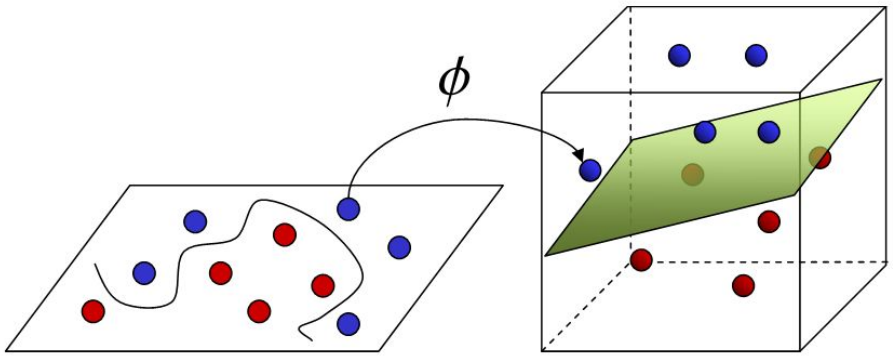


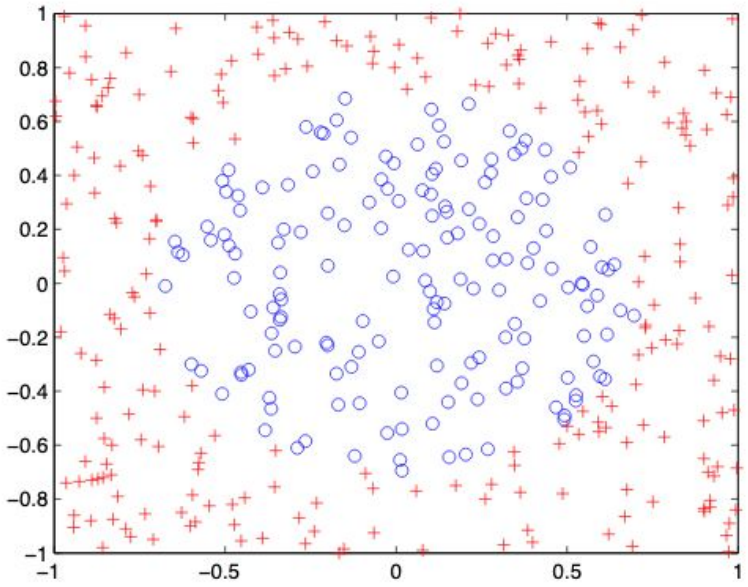
SVM

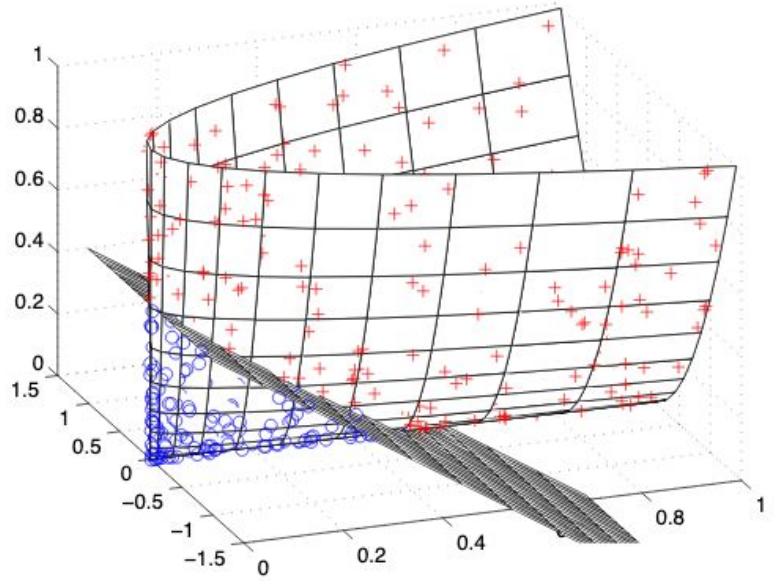


Logistic Regression

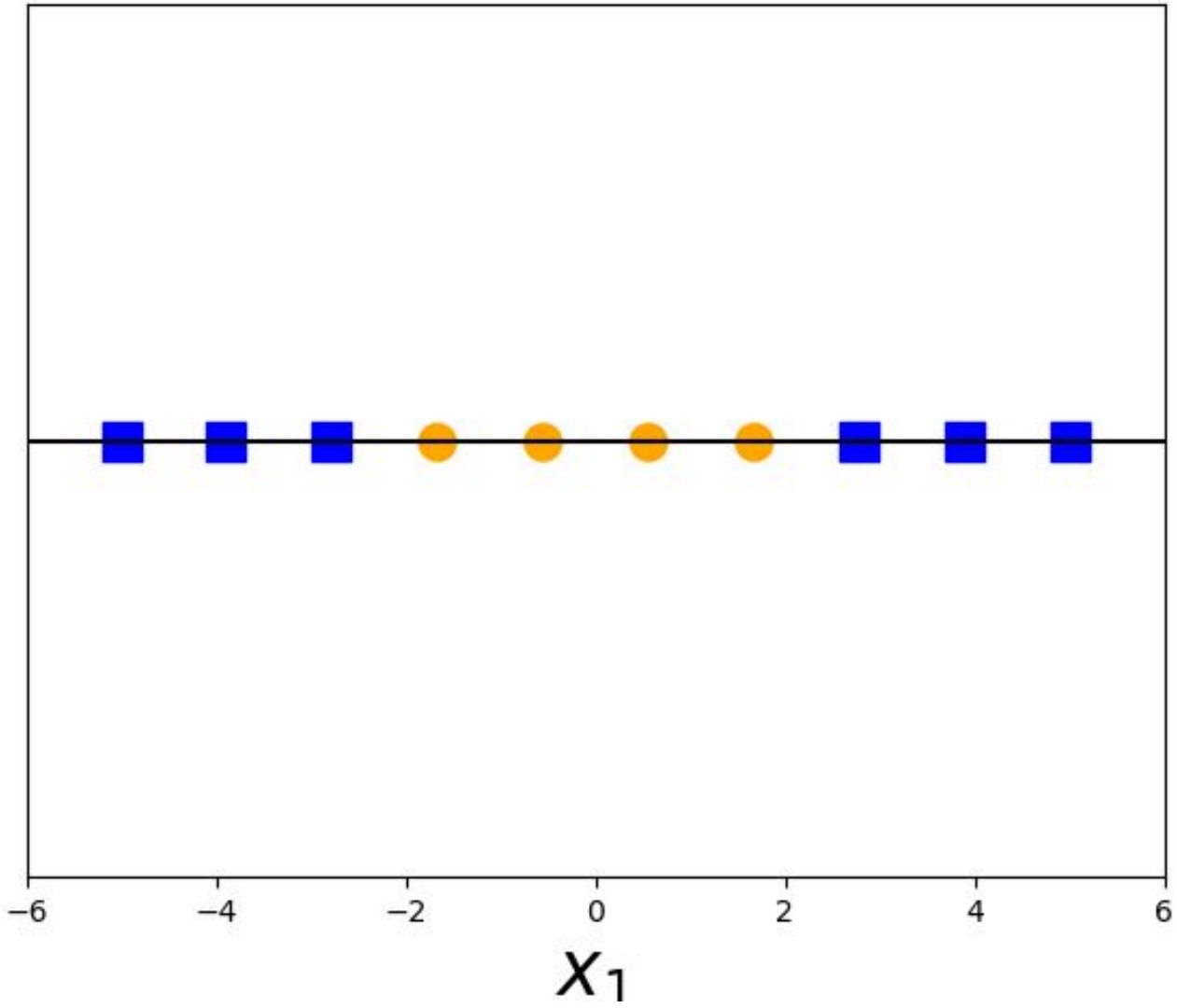
The Kernel Trick



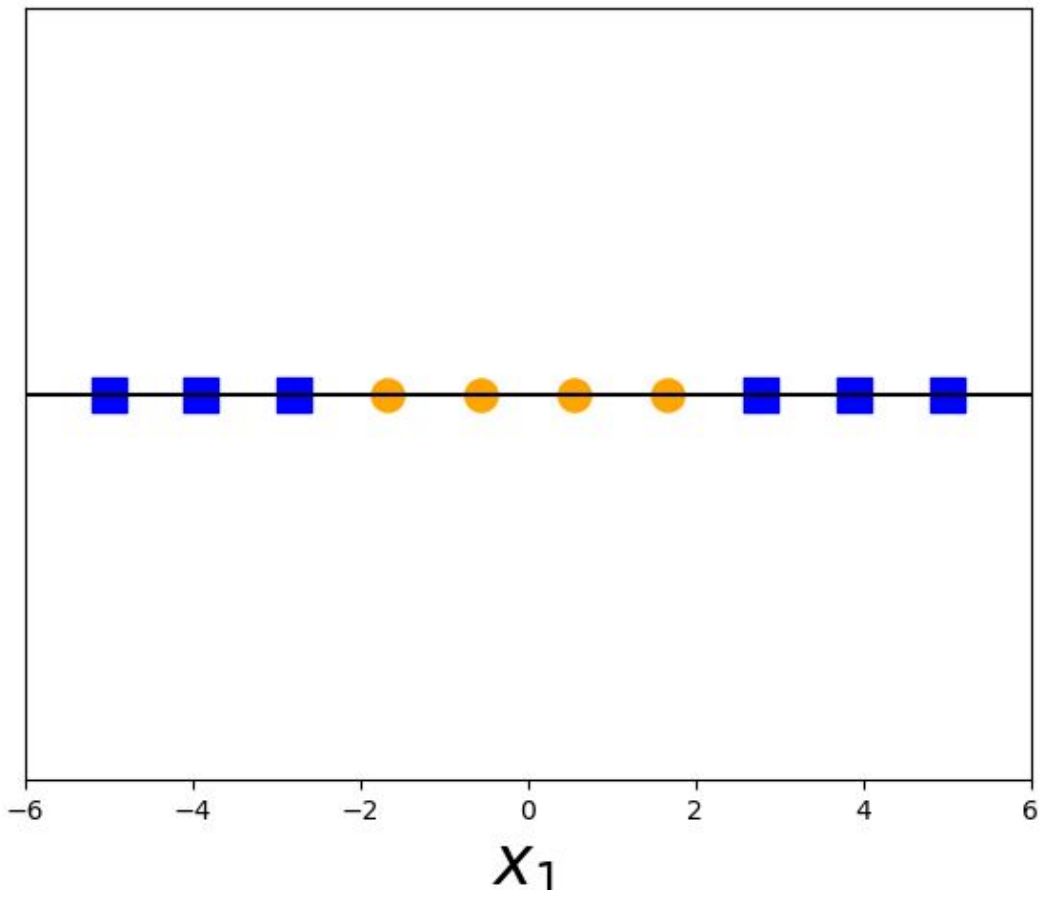


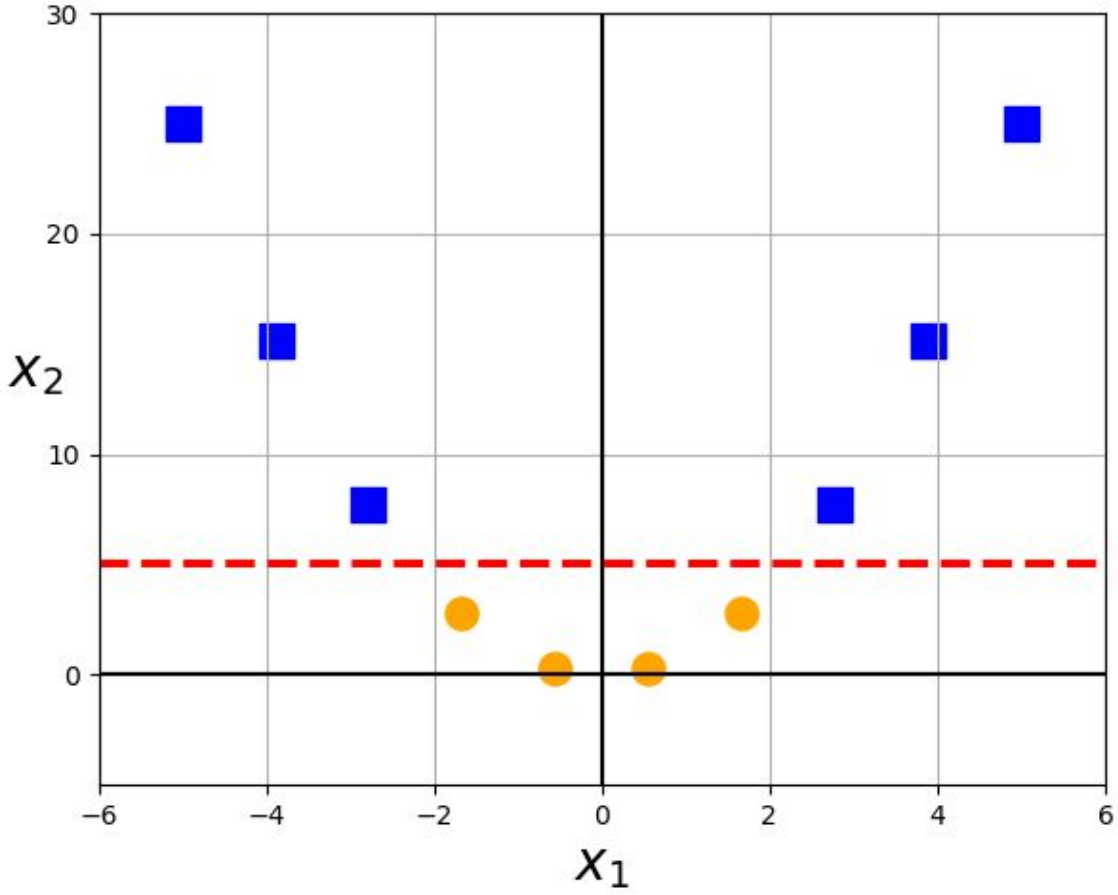


Simple Kernel Trick Example



Simple Kernel Trick Example

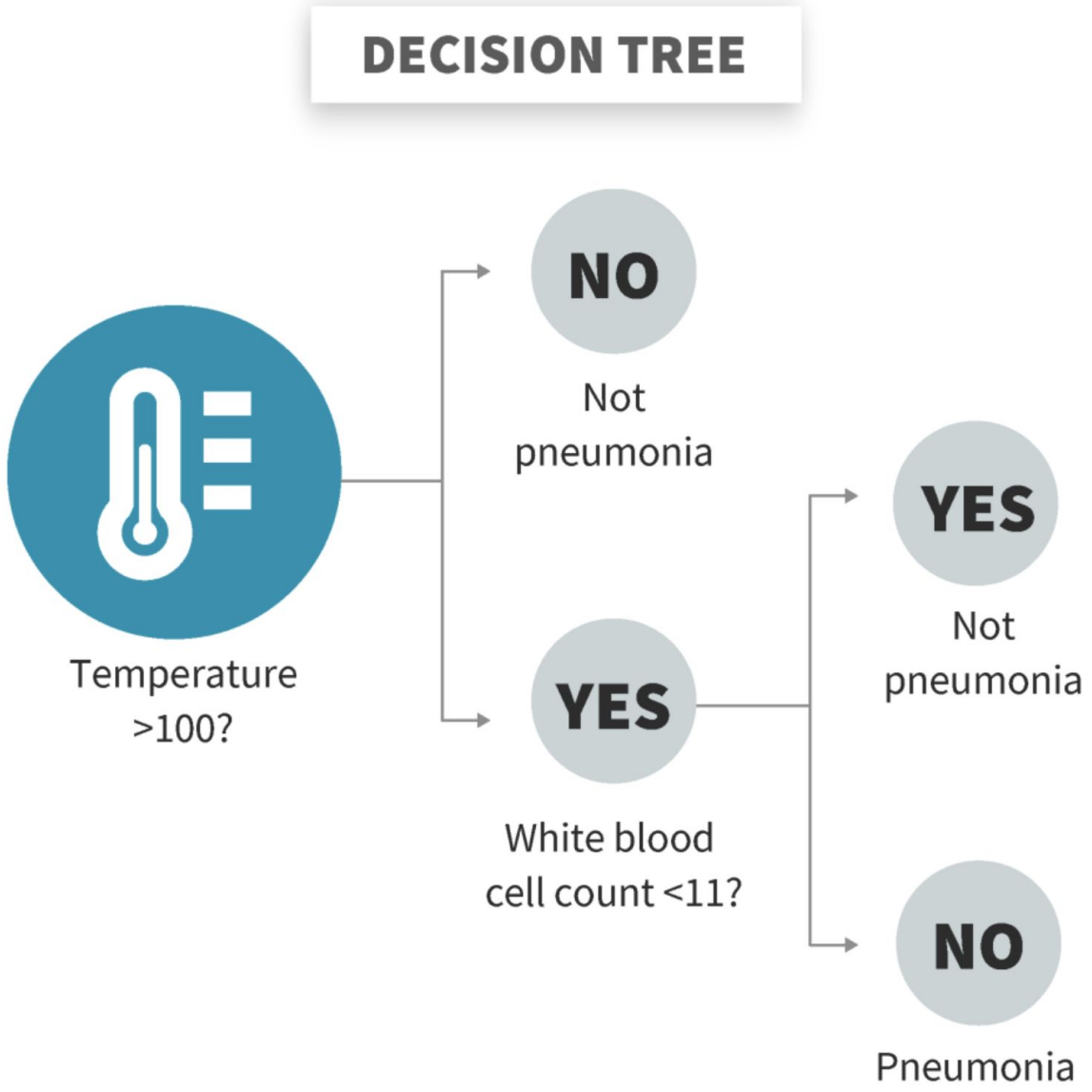




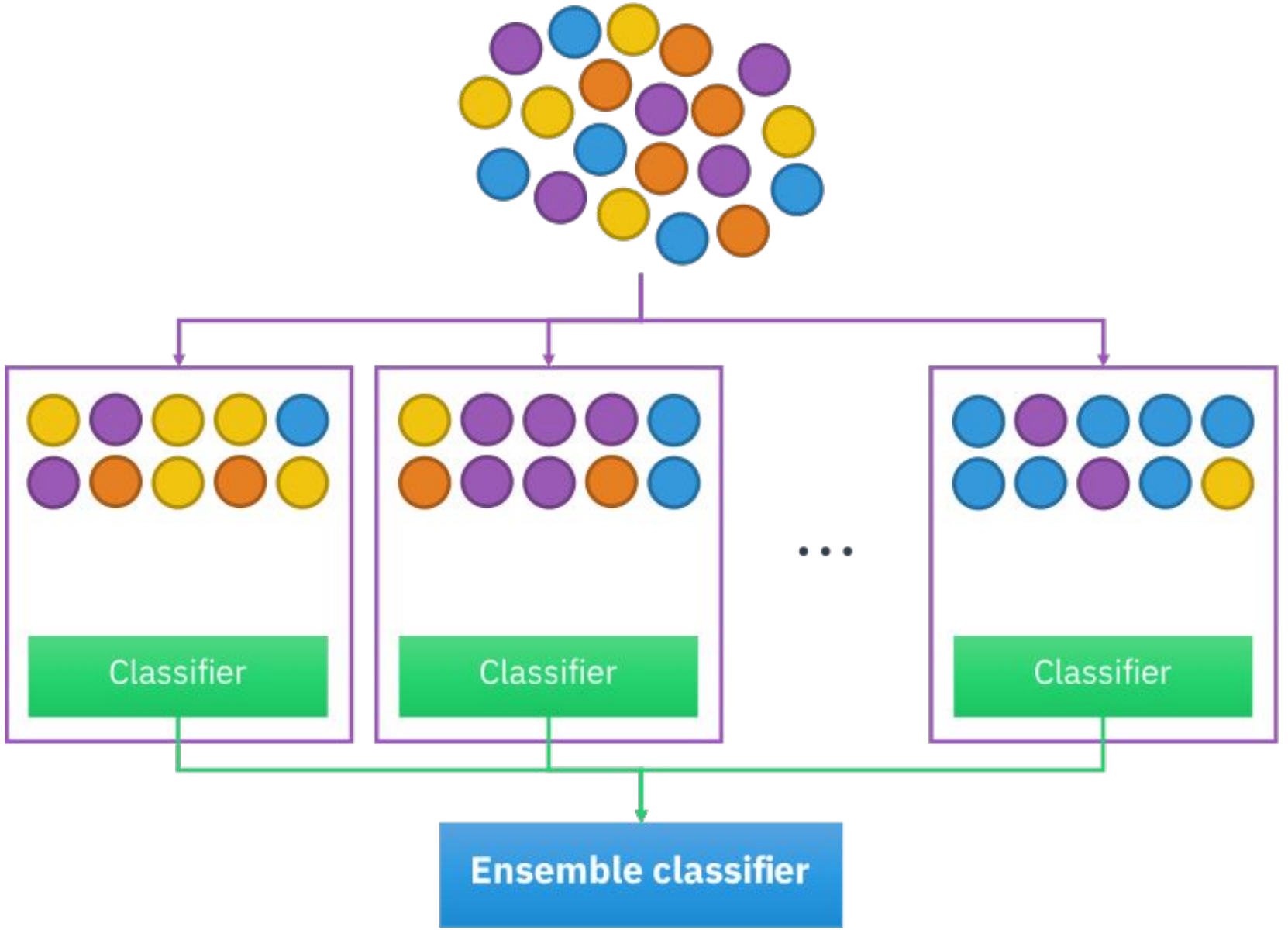
Outline

1. Lecture 2: Recap
2. K-nearest neighbor
3. Regression
4. Support Vector Machines (SVMs)
5. **Tree-based algorithms**
6. Applications

Decision Tree



Bagging



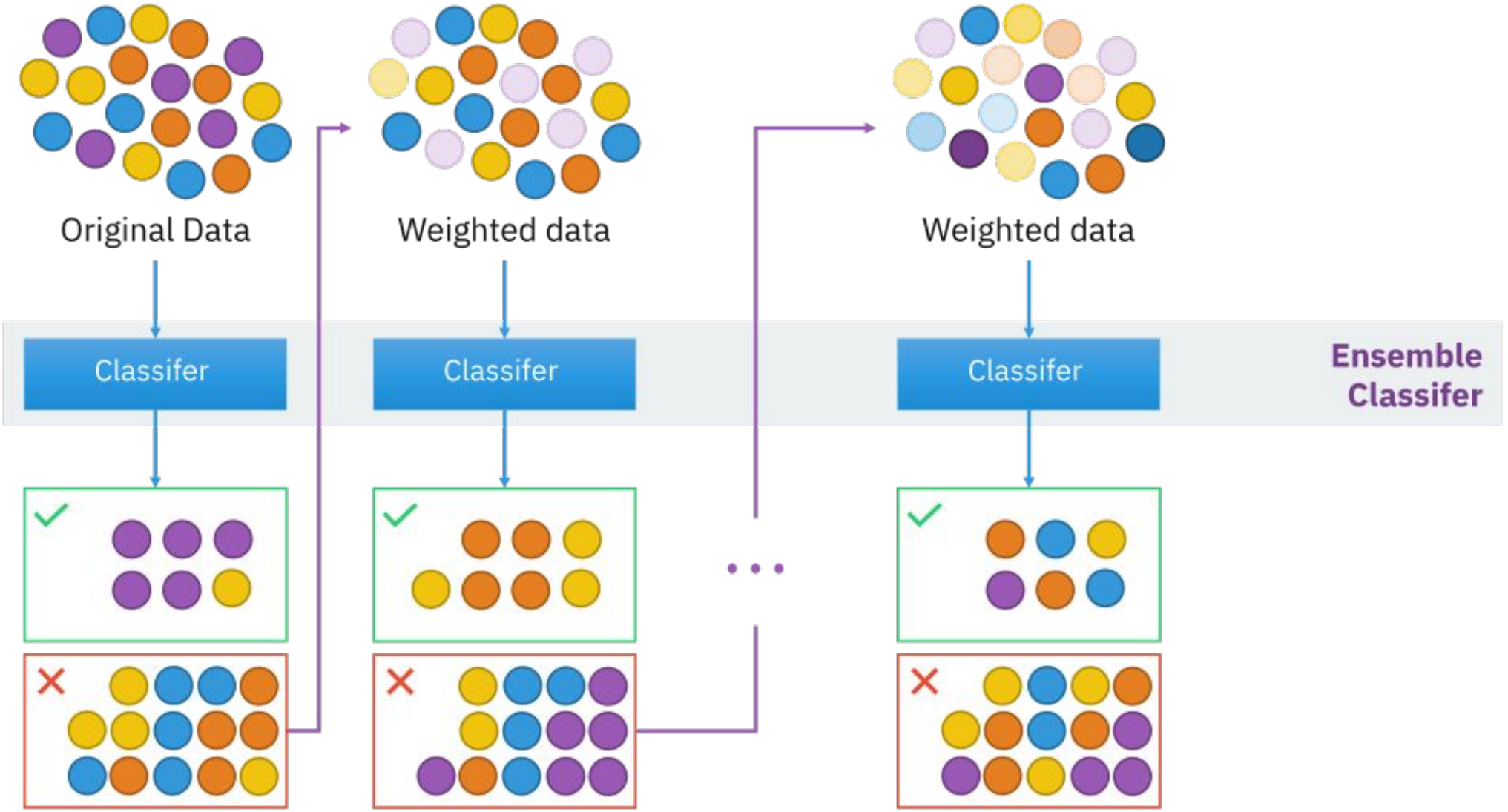
Original Data

Bootstrapping

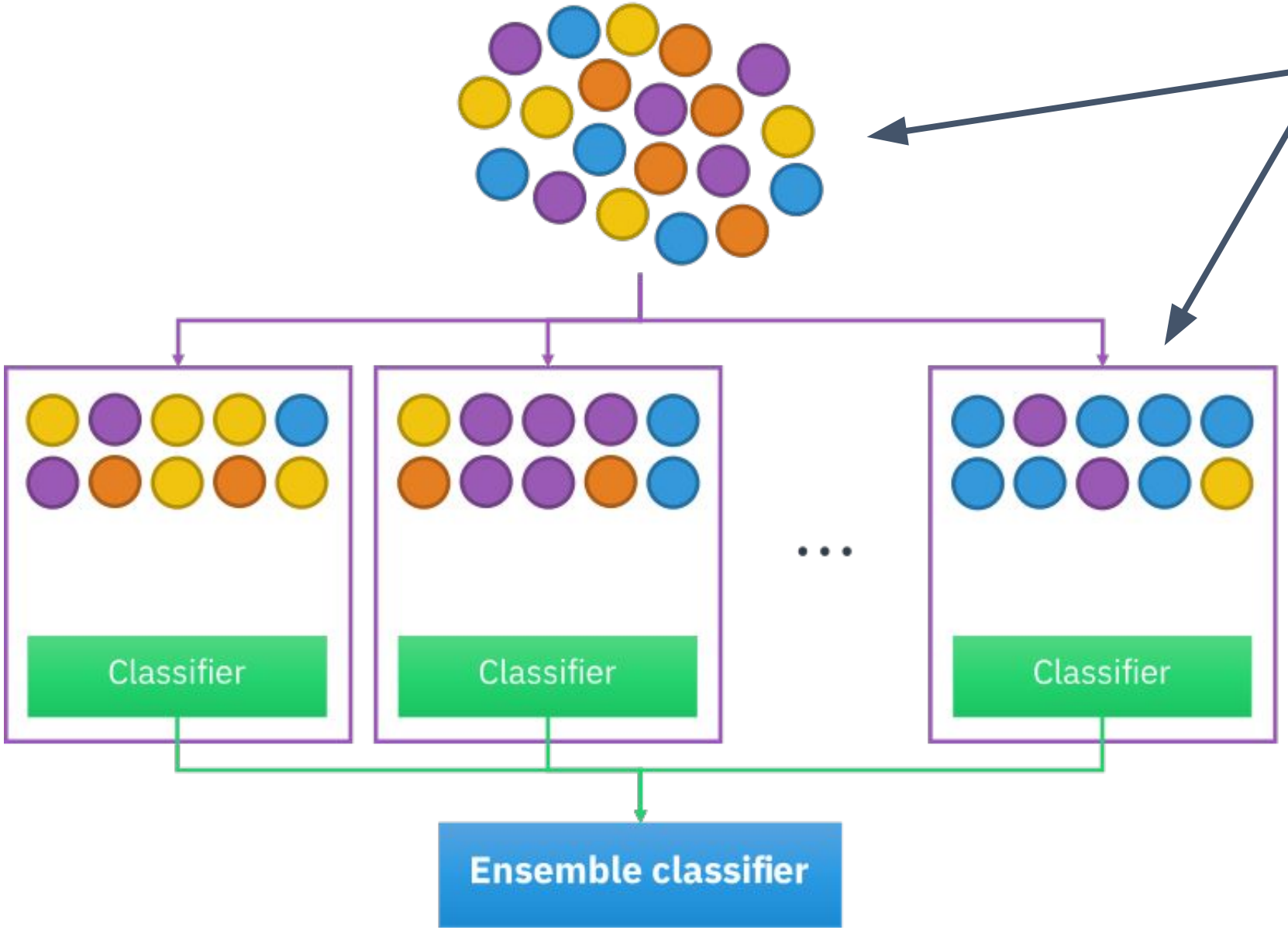
Aggregating

Bagging

Boosting


Original Data
Weighted data
Weighted data
Classifer
Classifer
Classifer
Ensemble Classifer

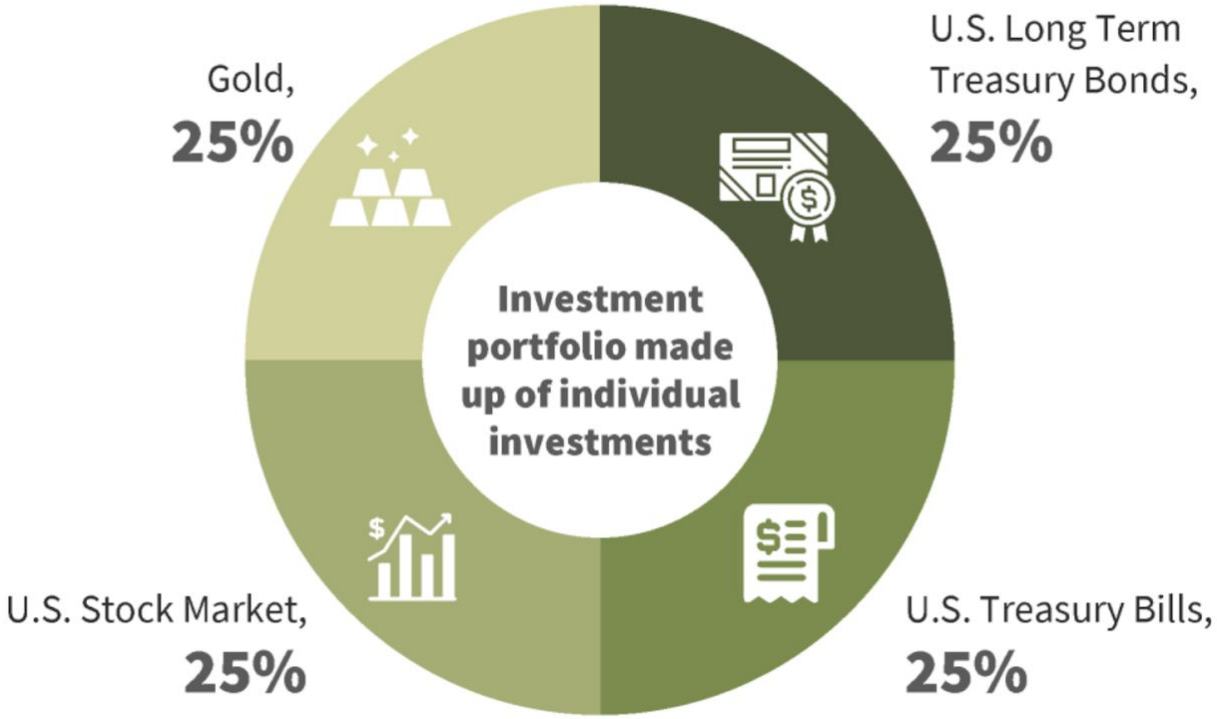
Random Forest

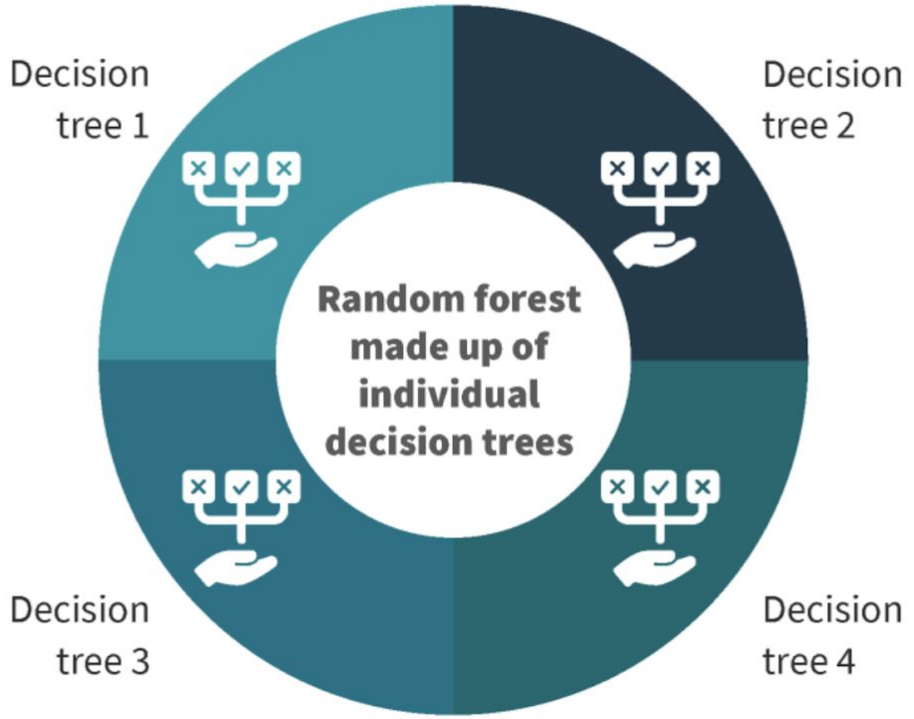


Unlike bagging, in Random Forest different **input features** are selected for each classifier rather than data points

Random Forest

Random Forest





Outline

1. Lecture 2: Recap
2. K-nearest neighbor
3. Regression
4. Support Vector Machines (SVMs)
5. Tree-based algorithms
6. **Applications**

Scenario 1

Data: Levels of 20 different blood transcriptomic markers *(TLR7, PSME1 etc.)* proposed to be associated with Major Depressive Disorder (MDD) from persons *with* and *without* MDD.

For each person it is known whether that person had MDD at the time of blood test.

Objective: Build a blood-based diagnostic test for Major Depressive Disorder

***Discuss**: How will you solve this using Machine Learning?*

OPEN

Citation: Transl Psychiatry (2016) 6, e931; doi:10.1038/tp.2016.198

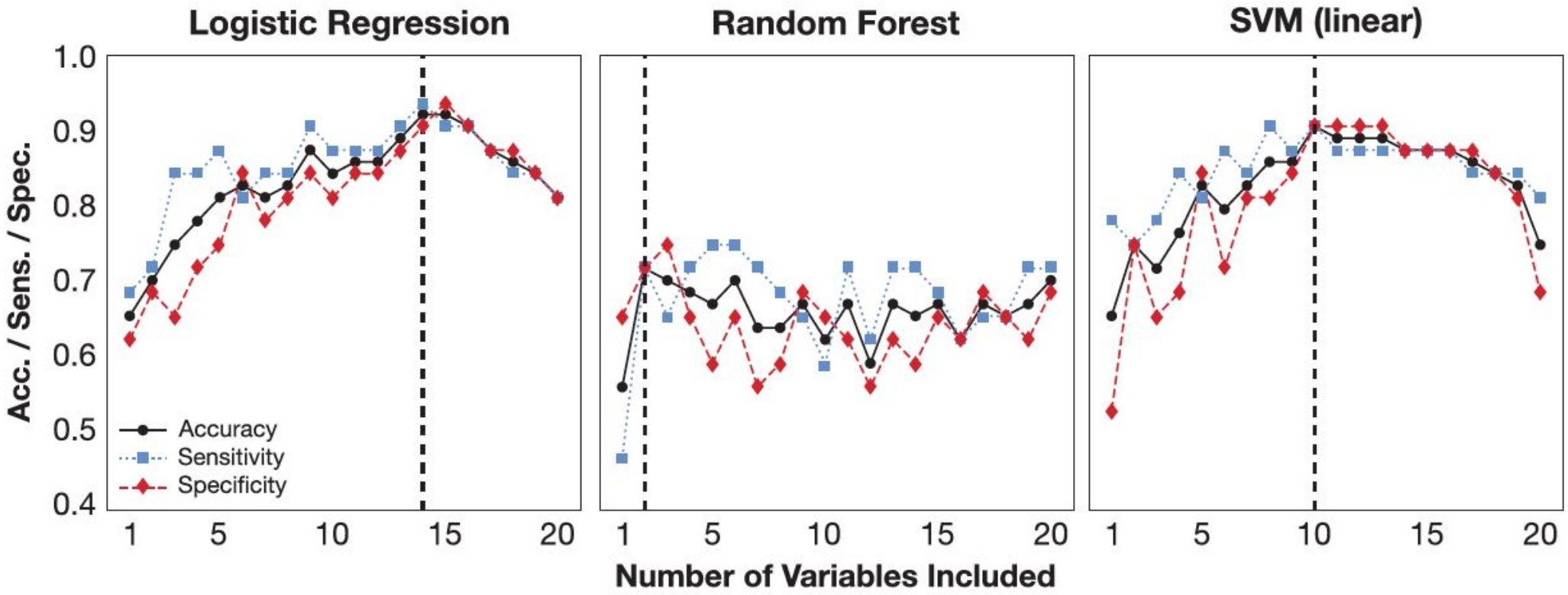
www.nature.com/tp

Paper 1

ORIGINAL ARTICLE

A support vector machine model provides an accurate transcript-level-based diagnostic for major depressive disorder

JS Yu[1], AY Xue[1], EE Redei[2] and N Bagheri[1]



Logistic Regression: Classification accuracy of 92.2% with 14 transcript variables. Sensitivity and specificity of 93.8% and 90.6%, respectively.

Linear SVM: Classification accuracy of 90.6% with 10 transcript variables. Sensitivity and specificity of 90.6% and 90.6%, respectively

Scenario 2

Data: Several physiological variables and blood-test parameters measured at time-points (assume every 30 mins) from hospitalized patients; *MAP, heart rate, age, C-reactive protein, S-glucose, SpO_2*

For each time-point it is known whether the patient has *circulatory failure* at that time-point.

Objective: Predict whether patient will go into circulatory failure within next 8 hours.

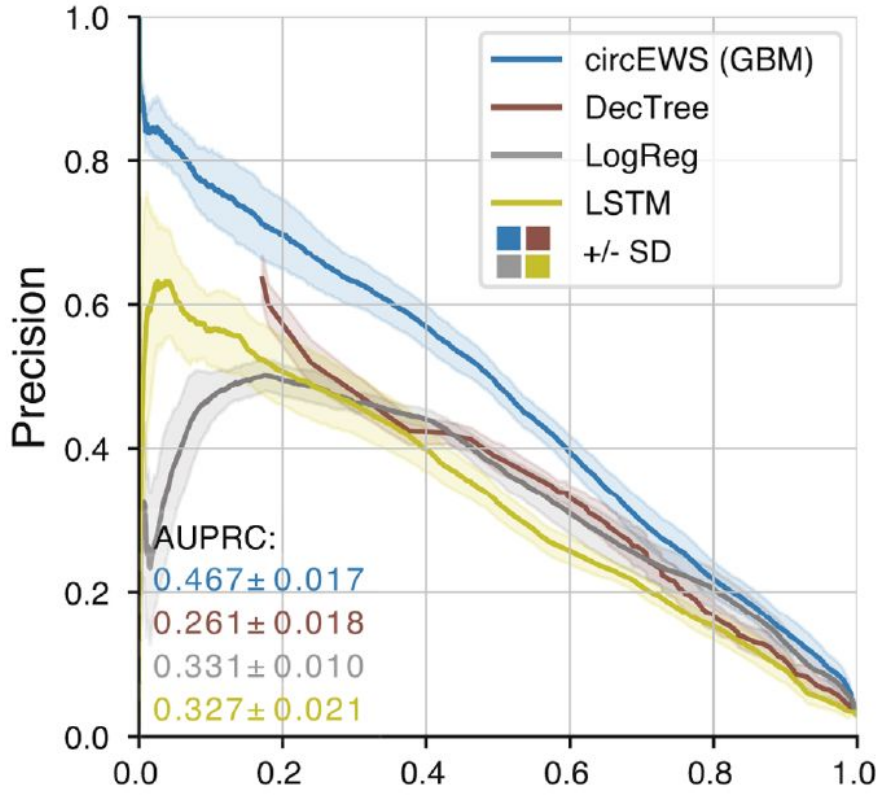
***Discuss**: How will you solve this using Machine Learning?*

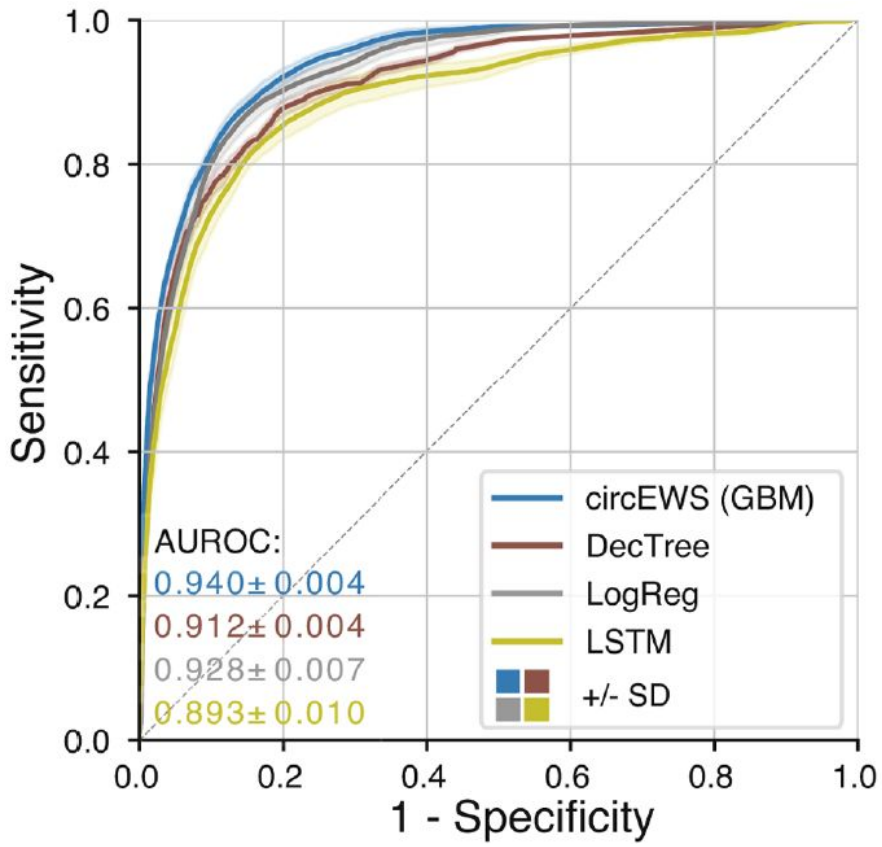
Paper 2

ARTICLES
https://doi.org/10.1038/s41591-020-0789-4
nature medicine
Check for updates

Early prediction of circulatory failure in the intensive care unit using machine learning

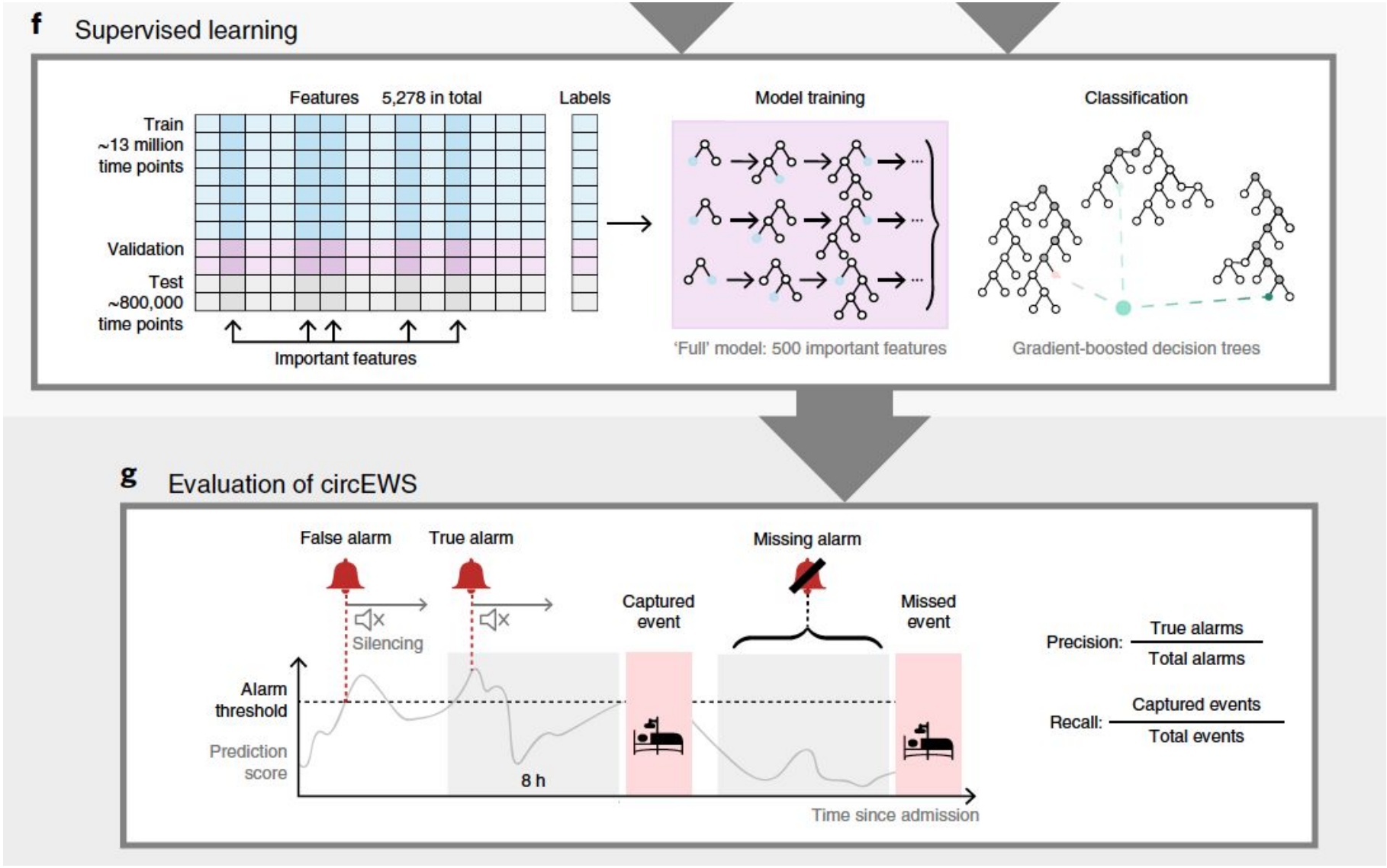
Stephanie L. Hyland[1,2,3,4,10], Martin Faltys[5,10], Matthias Hüser[1,4,10], Xinrui Lyu[1,4,10], Thomas Gumbsch[6,7,10],





Metrics: Precision (PPV) and Recall (Sensitivity). 1-Specificity is FPR.

Paper 2

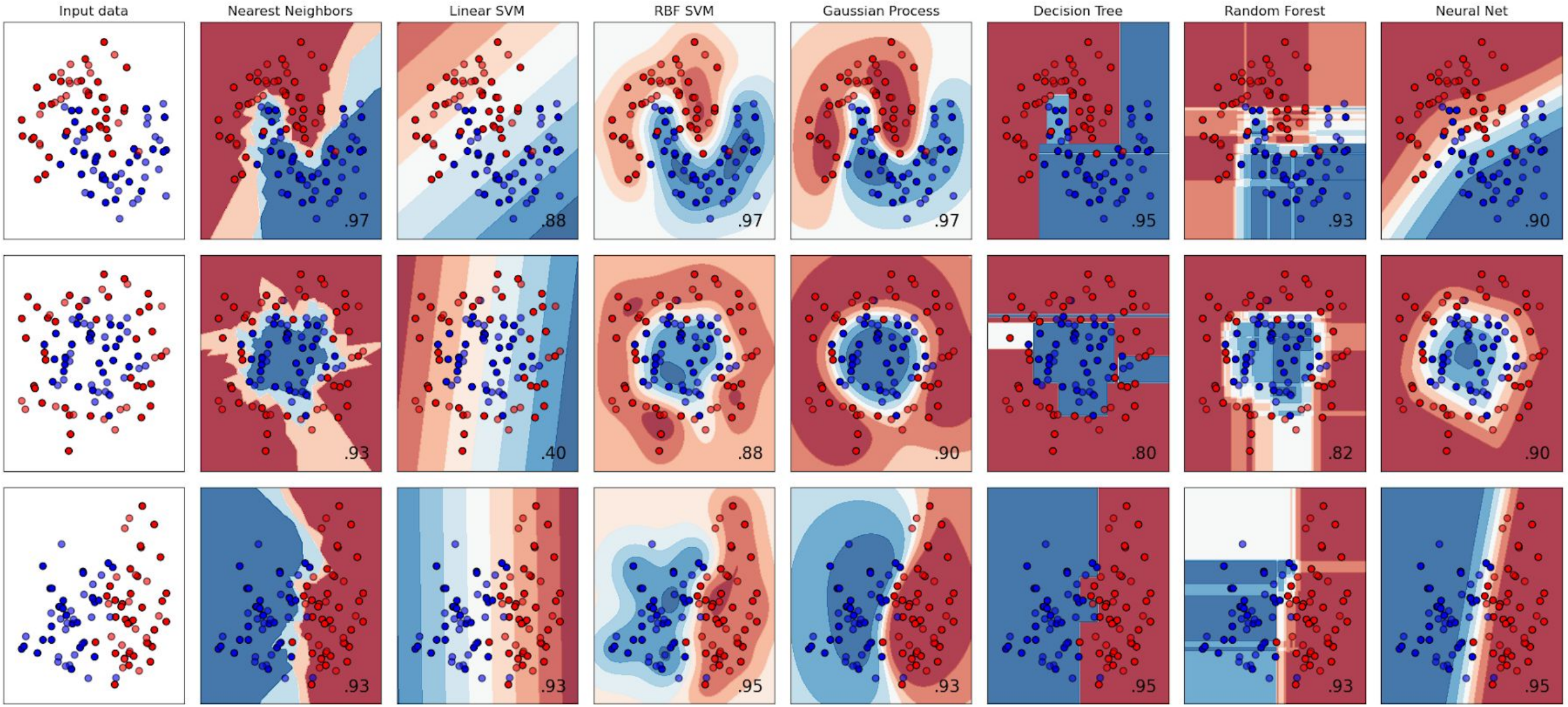


Gradient Boosted Decision Trees: predicts 90% of circulatory-failure events in the test set, with 82% identified more than 2h in advance, resulting in an area under the receiver operating characteristic curve of 0.94 and an area under the precision-recall curve of 0.63

When to use what model

1. Categorize the problem you are trying to solve
   - Supervised, Unsupervised, Reinforcement
2. Figure out the subgroup
   - i.e. for supervised, recognize if it is classification or regression
3. Understand the type of data you are dealing with
   - Size
   - Type
   - Structure
4. Other considerations
   - Speed
   - Computational resources

Comparison of Approaches for Classification



source: https://scikit-learn.org

Summary

Today we covered

- K-nearest neighbor
- Regression
- Support Vector Machines (SVMs)
- Tree-based algorithms
- Applications

Coming up: Fundamentals of deep learning and neural networks