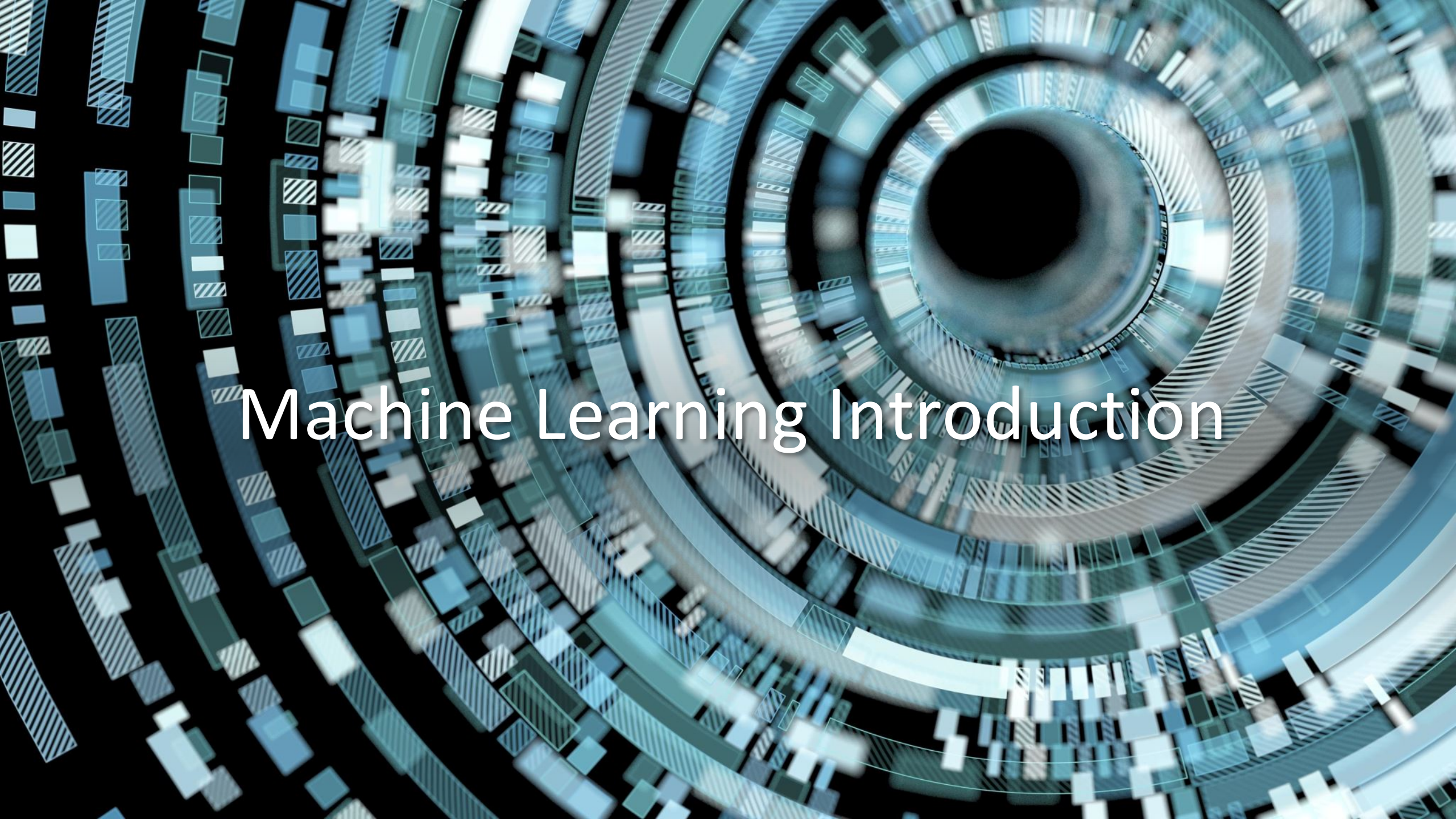
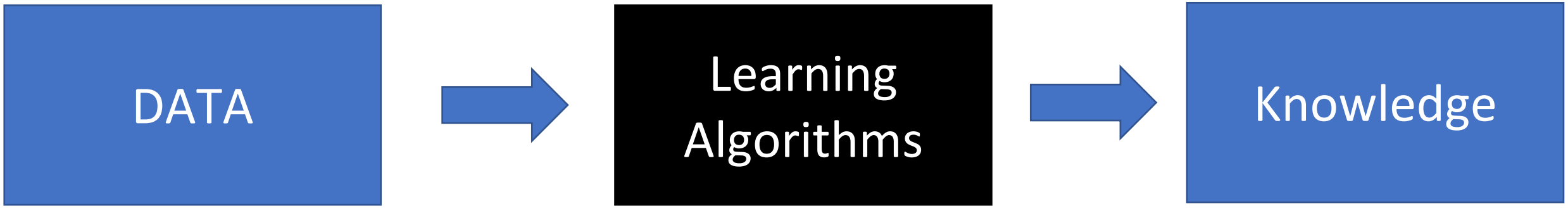

Machine Learning Introduction

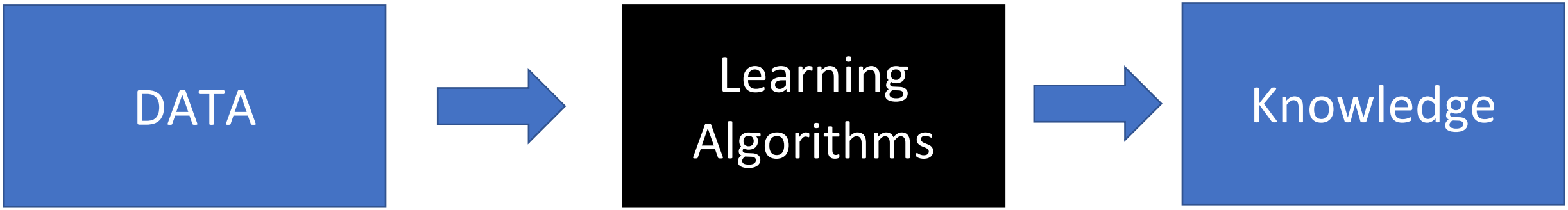
What is Machine Learning?

- Easy part: Machine
- Harder: learning
  - Short answer: Methods that can help generalize information from the observed data so that it can be used to make better decisions in the future

What is machine learning?



What is machine learning?



- Algorithms that improve their knowledge towards some task with data
- How is that different from Statistics?
- What is the relationship with AI, Data Science, Data Mining?

Can differentiate fields by their goals

- Machine Learning: Learn from one dataset and be applicable to **more in the future**
- Statistics: understanding the data **at hand**
- AI: build an intelligent agent
- Data Mining: extract patterns from large-scale data
- Data Science: the science encompassing collection, analysis and interpretation of data

Machine learning in action

- The spam filters in your email
- Self driving cars
- Google translate
- Determining which part of a genome sequence encodes a gene
- And….
  - Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Protein sequence -> protein structure

There are two very broad categories

1. Supervised learning
2. Unsupervised learning

- Both are valuable and are applied at different times to different data sets

Supervised learning

- One of the variables is labeled, so we know the “truth”
  - Class labels –categorical - (high risk/ low risk, normal/tumor
  - Continuous labels (income, age)

- Objective: predict the label from other variables – called features
  - Ex. Given a person’s age, weight, height, gender can you predict if they play football

Supervised learning

- If labels are discrete -> classification
- If labels are continuous -> regression

Unsupervised learning

- Learning without a teacher
- There is no labeled data
- The computer is supposed to learn general patterns in the data
- Typically some form of cluster analysis
- Goal: identify homogeneous subgroups

Unsupervised learning examples – Clustering

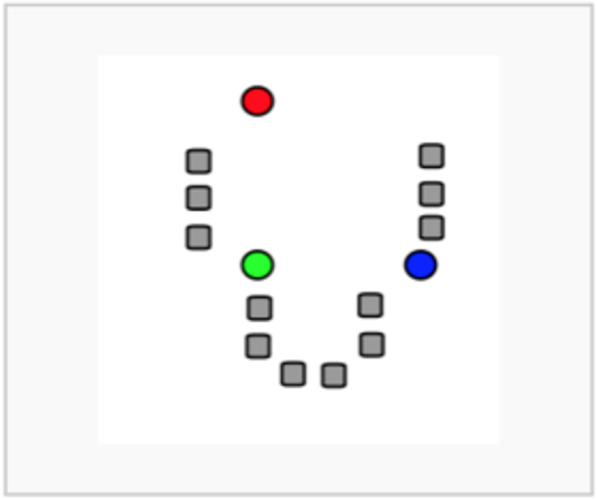
Group similar things:

Clustering methods

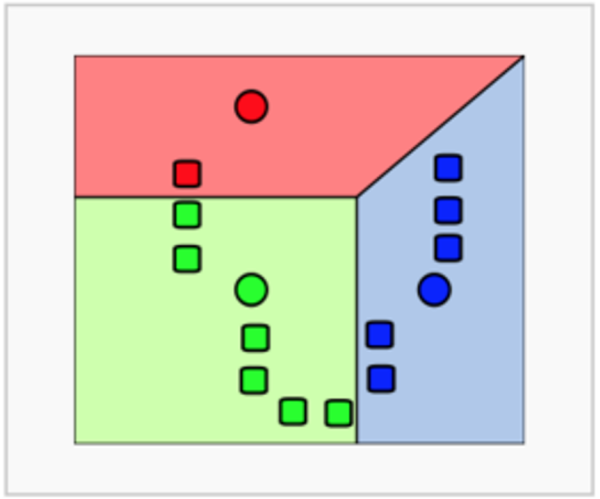
- K-means clustering
- Hierarchical clustering
- Mixture models
- There are a lot of these …

K-means very briefly

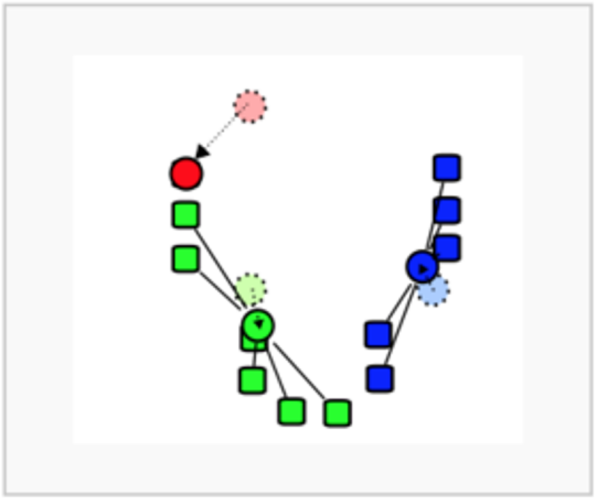
Demonstration of the standard algorithm



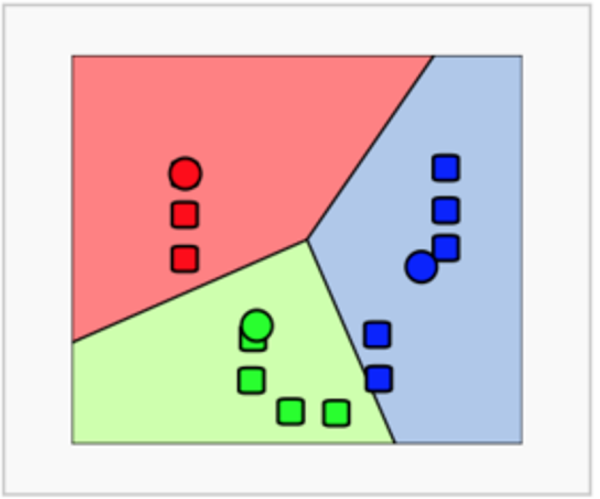
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

brilliant.org

In Statistics

- Fit a linear regression vis the sum of squared deviations
- Goal: the line that fits best on average

- THIS IS NOT THE GOAL IN MACHINE LEARNING
- Why?

Statistics vs Machine Learning

- Stats model that fits the data really well (large R^2)
  - May be overfitting the data
  - Won't work well on new data

- Need to do something different.

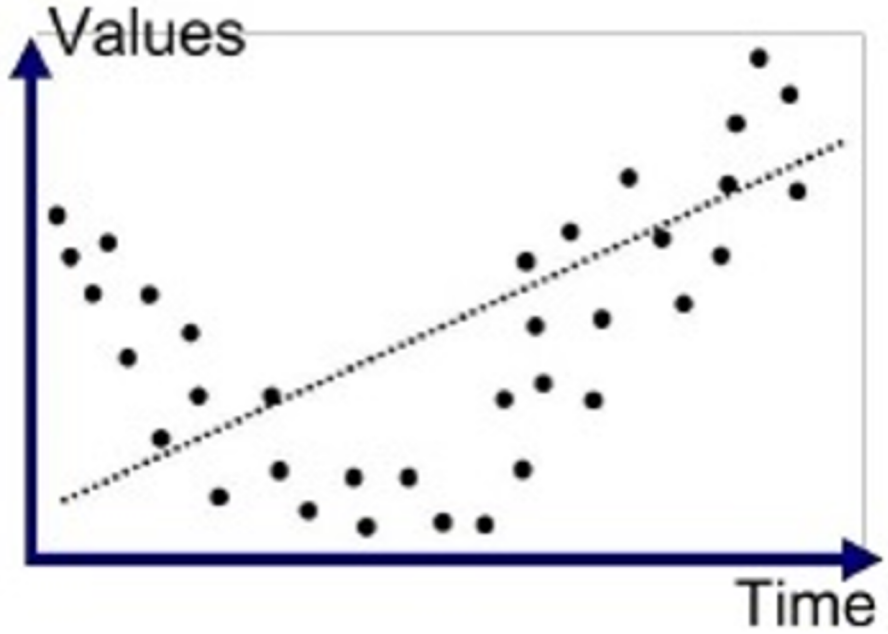
Something different: Reserve a test set

- Standard method:

1. Split your data set into training set and test set
2. Train your data on the training set
3. See how well it works on the test set
4. Take the model that does the best on the test set

The Variance-Bias Trade-off

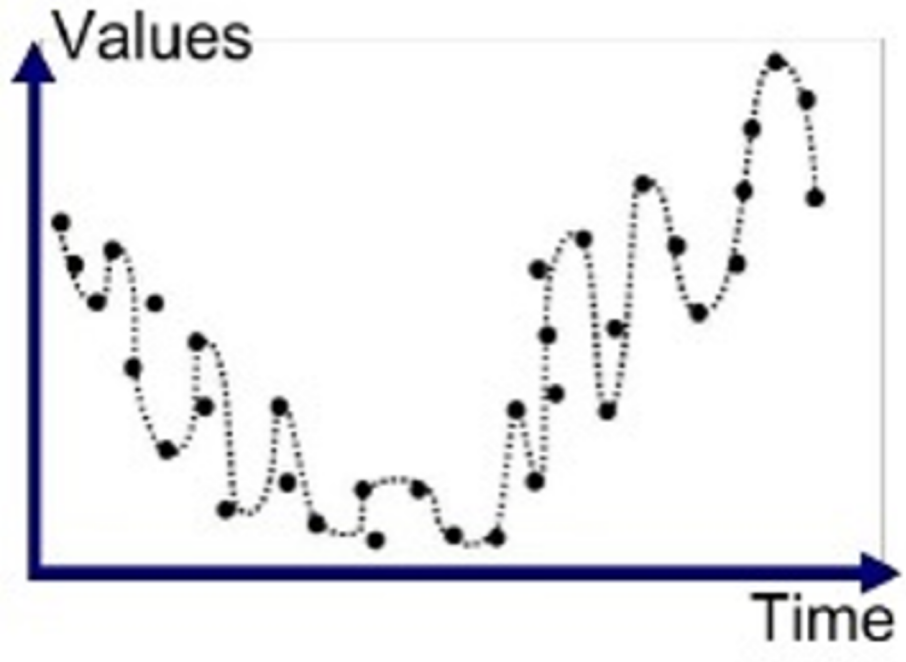
- Some models are too simple
- Under-fit the data, are not capturing all the available information
- Do equally well (or bad) on training and test data
- These are *high-bias* models



Underfitted

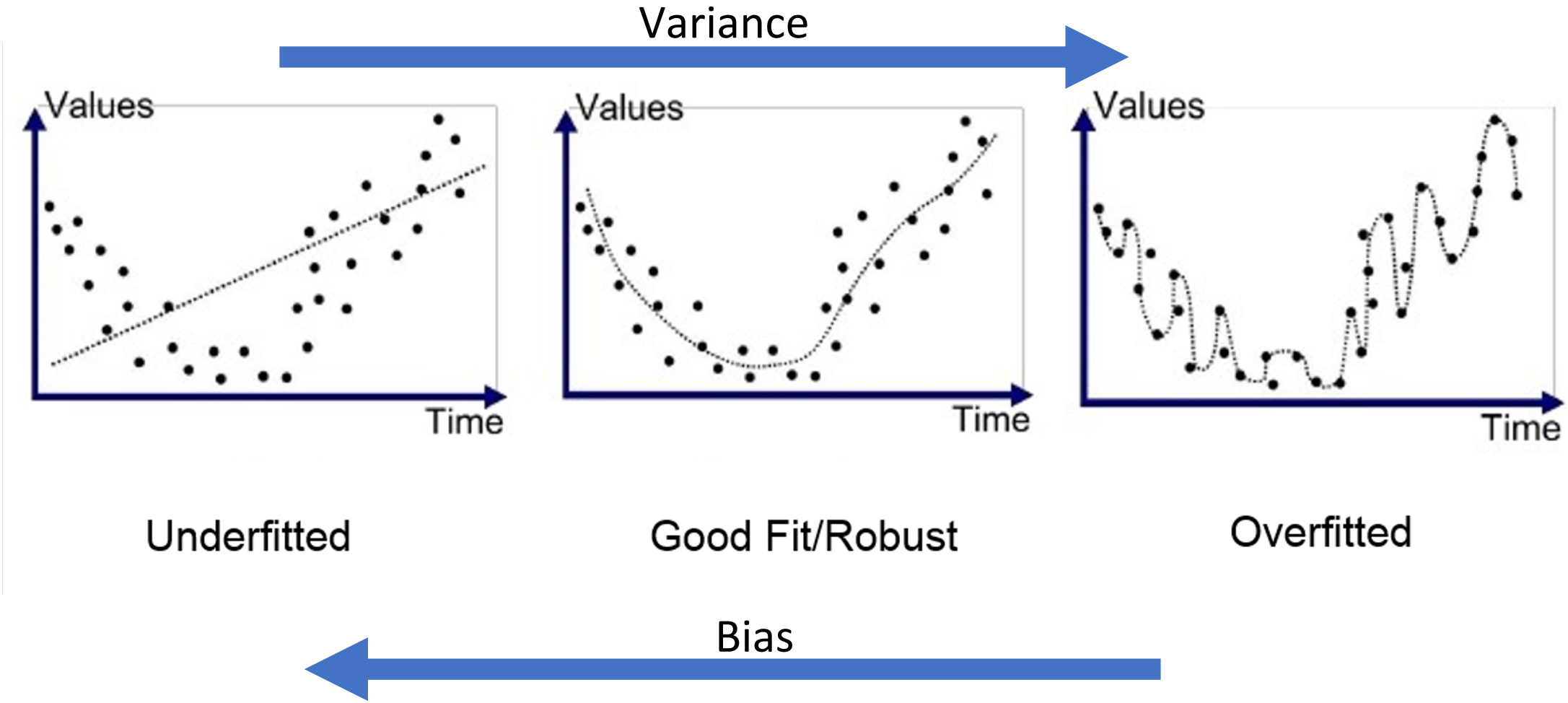
The Variance-Bias Trade-off

- Some models are too complex
- Overfit the training data
- Do badly on test/new data
- These are *high-variance* models



Overfitted

Goal: Land in the middle (low-variance, low-bias)



Model Performance

Metric to see how well a model works

Can’t just rely on the one test set

Metric of Performance depends on model

- Classification:
  - Accuracy (misclassification)

- Regression
  - Mean Square error

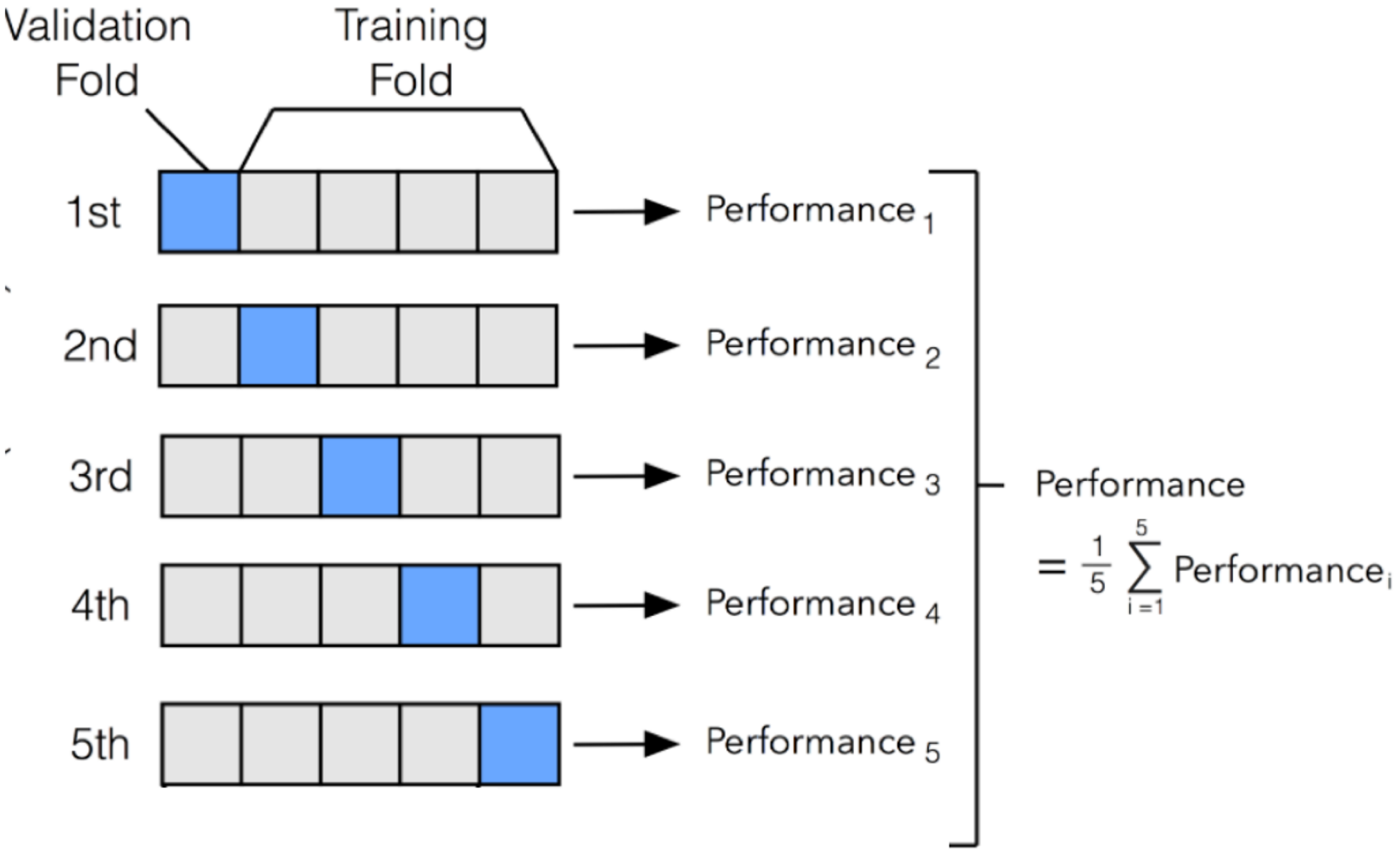
How can we make more test data?

Take multiple random splits to generate training and test data

See overall performance across different splits

This is called: cross-validation

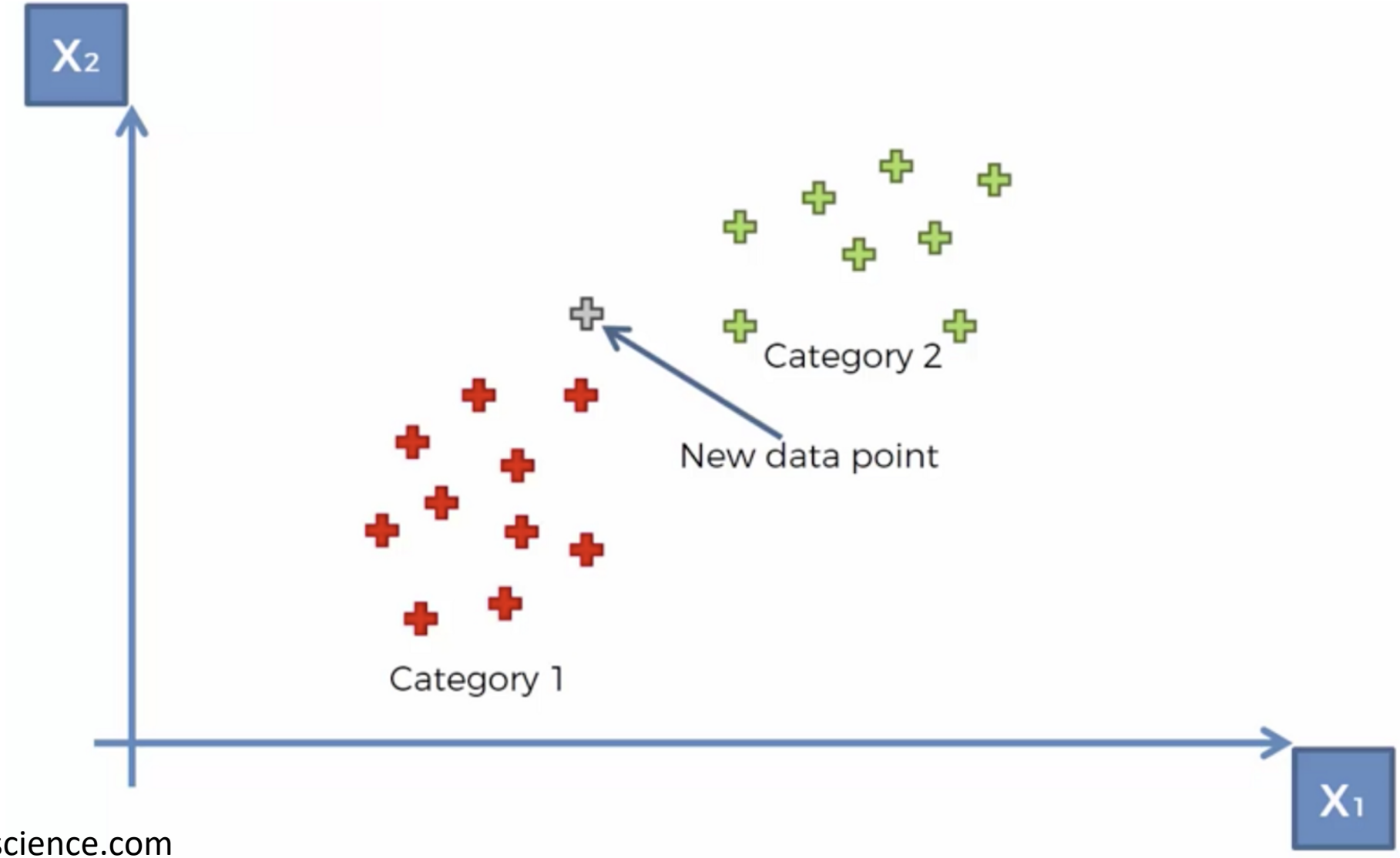
Cross validation



Supervised learning methods

- K- nearest Neighbors
- Decision trees
- Ensemble methods
  - Random forests
  - Boosted trees
- Support Vector Machines

K-nearest neighbor



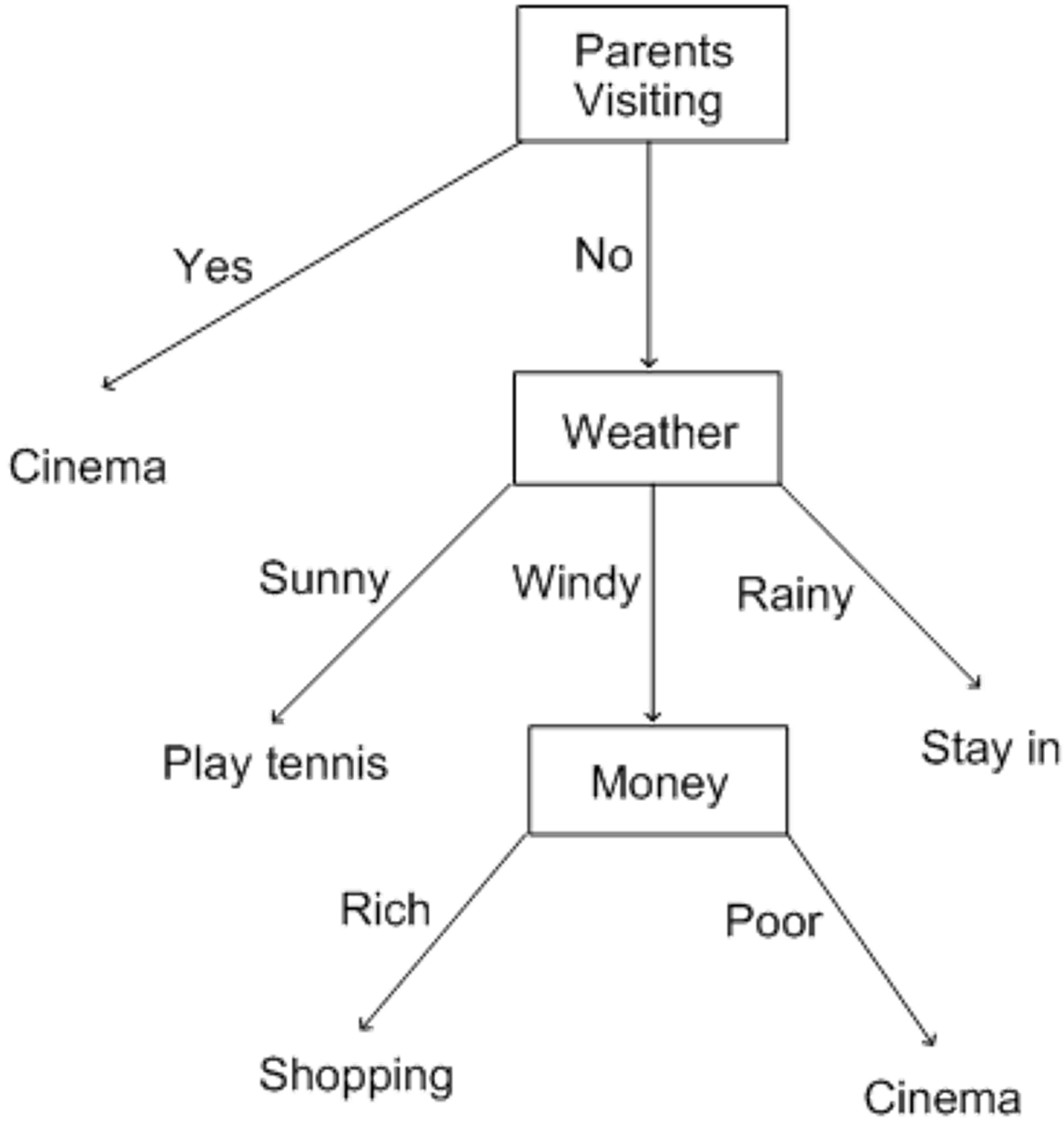
towardsdatascience.com

K-nearest neighbor

- For each data point, find its k nearest neighbors in the predictor space
  - Decide on a distance metric

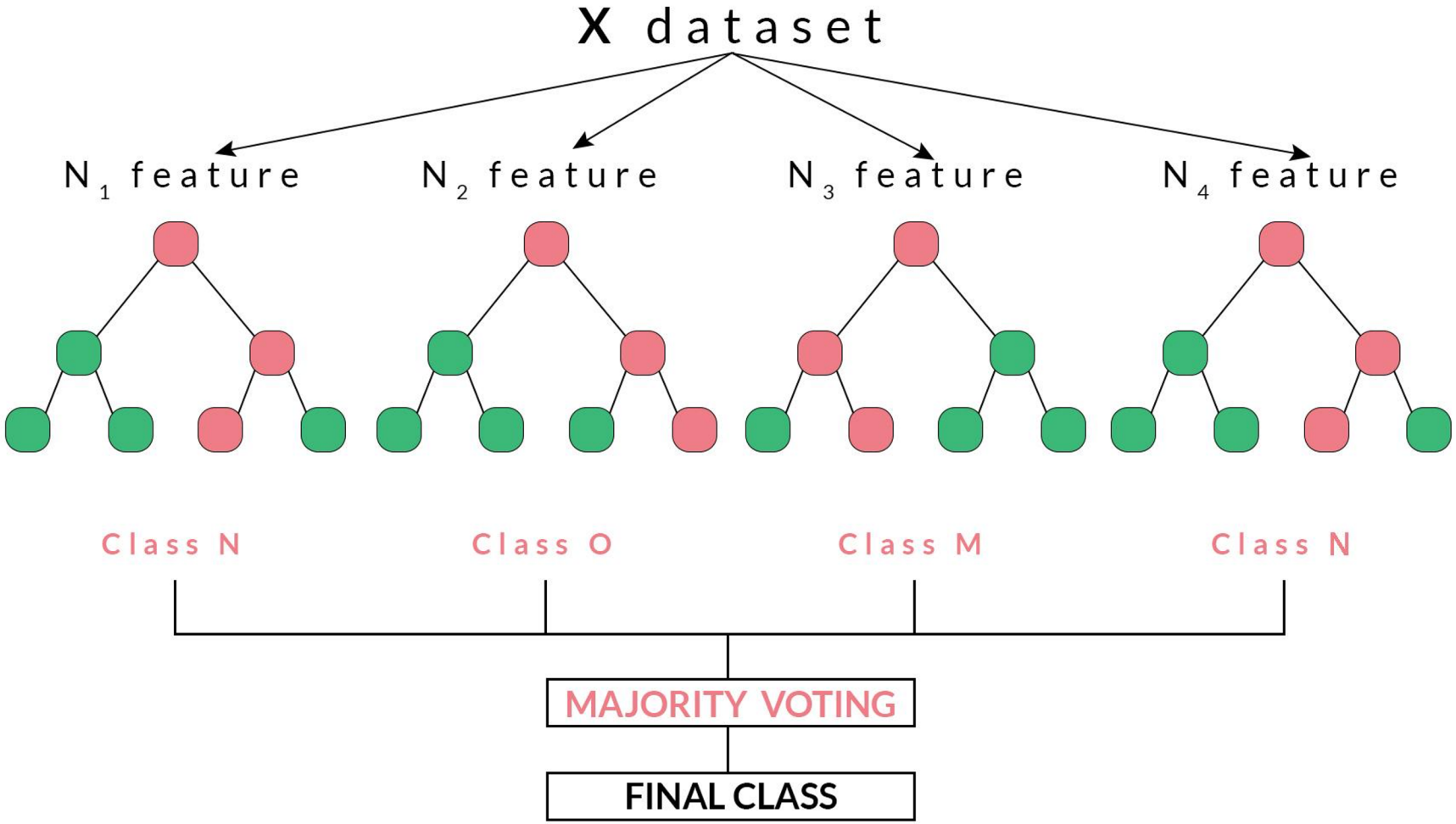
- Prediction of that data point is the
  - Average of the labels (regression)
  - The most prevalent of the labels (classification)
  - Observed in those neighbors

Decision Trees



quora.com

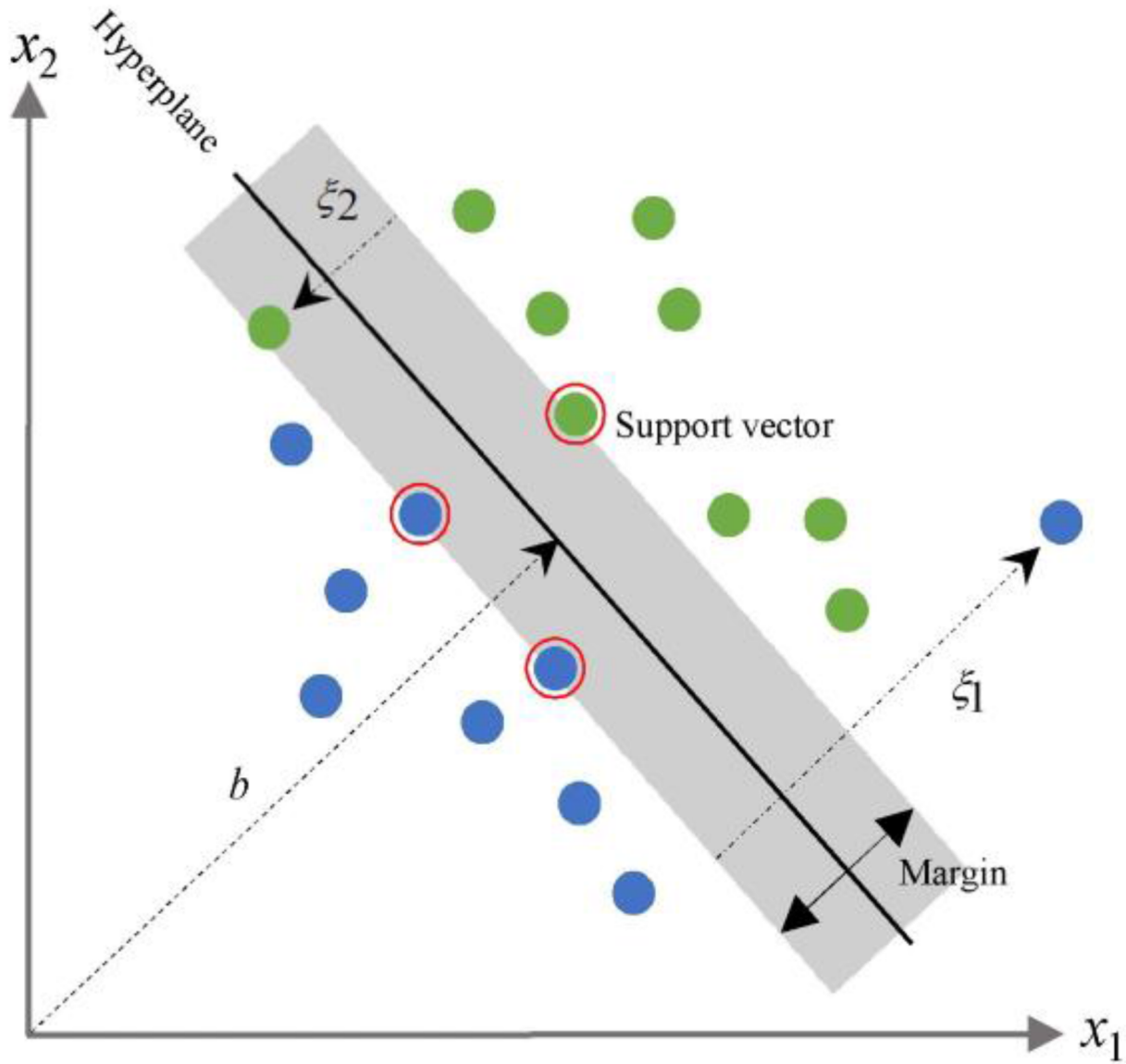
Random Forest



blog.quantinsti.com

Support Vector Machines

- Give higher importance to the points near the boundary



mdpi.com

Supervised Learning Methods

- Tend to focus on local solution
- We will be focusing on decision trees

The Scikit-learn Package

- Does lots of the python machine learning
- We will use this extensively in fitting machine learning models to data
- **General method:**

1. Import an ML algorithm from sklearn
2. Define the characteristics of the model
3. Fit training data to the model
4. Predict on test fata
5. Compute performance metrics and cross-validation