# Biomarkers: Machine Learning / Statistical Learning

DAVID L. TABB, PH.D.

AUGUST 27, 2019

# Overview

- Essentials of supervised machine learning: training, validation, and testing
- Challenges to machine learning: overfitting, signal leakage
- Modes of learning: decision trees / random forests, artificial neural networks, support vector machines.

# Supervised and unsupervised learning

- Supervised learning is useful in cases where a property (label) is available for a certain dataset (training set), but is missing and needs to be predicted for other instances.
- Unsupervised learning is useful in cases where the challenge is to discover implicit relationships in a given unlabeled dataset (items are not pre-assigned).

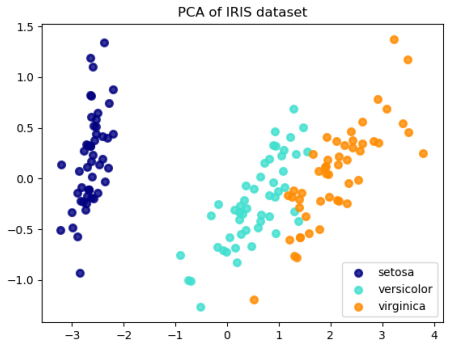
--James Le

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Unsupervised: Clustering and Principal Components Analysis

## HIERARCHICAL CLUSTERING

- Matrix of *distance metrics* shows item relationships.
- Agglomerative or divisive techniques build *dendrogram*.


## PCA

- Given feature table, PCA returns components sorted by amount of variance explained.
- Features may be correlated, but components minimize correlation.

www.informit.com/articles/article.aspx?p=357695

scikit-learn.org/stable/modules/decomposition.html

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Our conversation with a supervised learner:

- Computer, I have measured these features for all my subjects <supplies table>.
- *This* subset of subjects represents positives.
- *That* subset of subjects represents negatives.
- Find a strategy to combine information across features that can discern positives from negatives.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Competing tasks for supervised learning

Three orthogonal goals:

1. Identify a small set of features that yield the best possible classifier.
2. Use the classifier to understand the underlying biology.
3. Train the most accurate classifier.

# Feature vector

- What values might our model need to guess whether or not I have diabetes?

| Pt ID | Temp | BP | HbA1c | BMI | B Glucose |
|---|---|---|---|---|---|
| DLT001 | 37C | 120/80mm | 48 mmol/mol | 26.2 | 5.3 mmol/L |

- Each column is a feature, and the info for a particular subject is a feature vector.
- Features may be integers, floats, or values from a set. Some values may be missing!

# Curse of high dimensionality

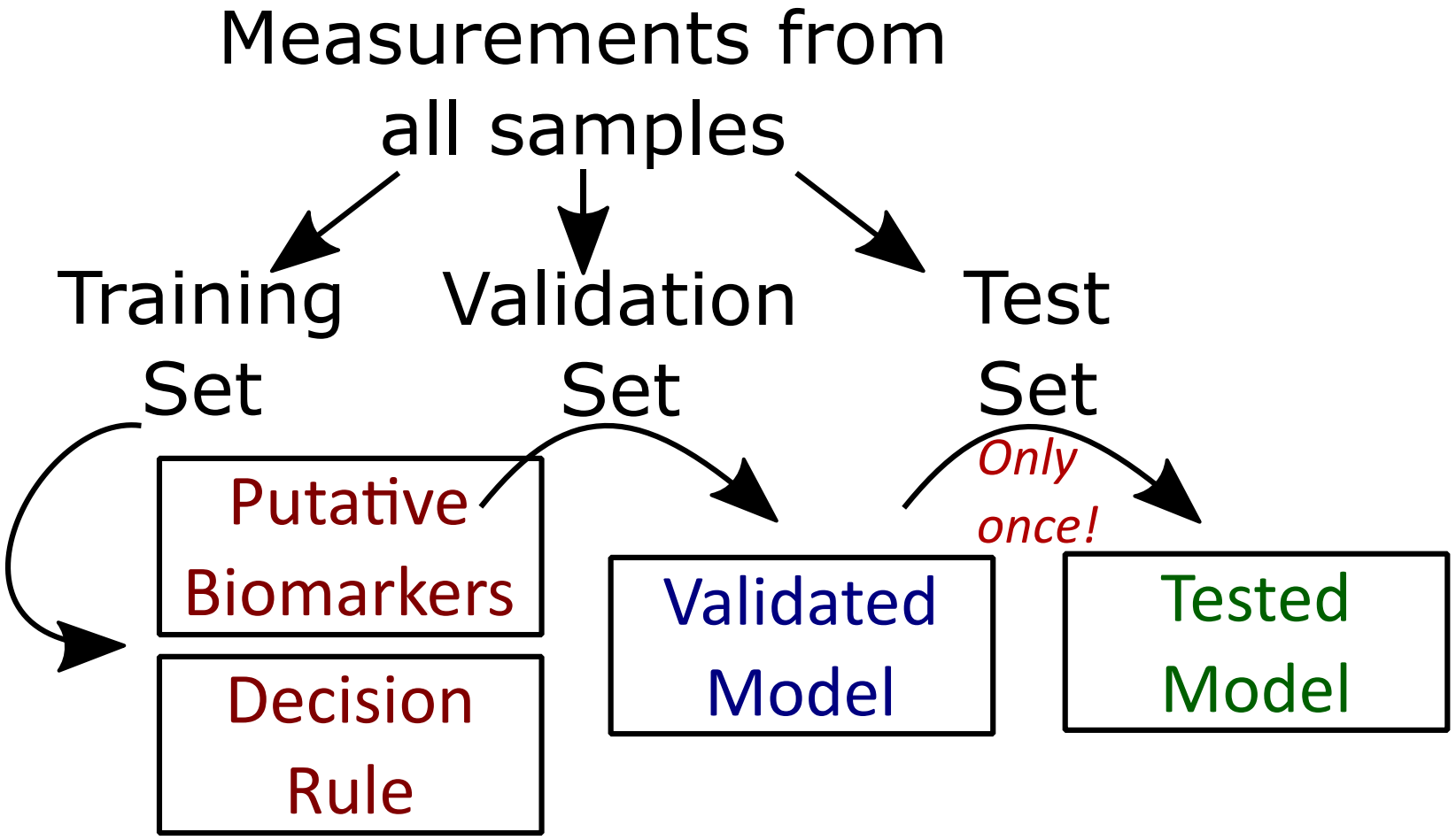
- As the number of features grows, more observations are needed to weigh features.
- Biomarker feature sets typically outnumber observations by 10 or 1000 fold.
- This is a recipe for over-fitting, giving seemingly perfect classification that will not generalize.
- Use feature selection:
  - Score each feature alone, and then use only the best.
  - Add or remove features iteratively.



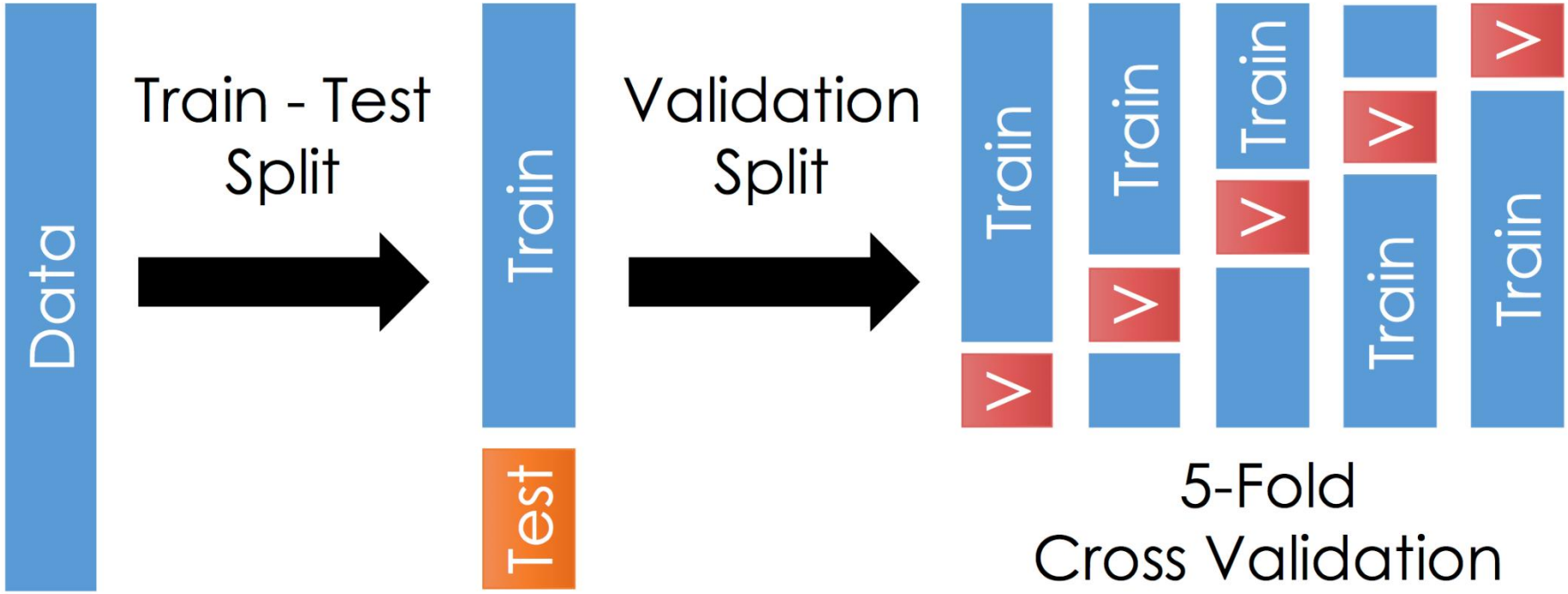
Slide from Nathan Edwards, Georgetown University

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Machine and statistical learning for predictive models



# Cross validation useful when data are limiting



UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Commandments of Machine Learning

1. Thou shalt not represent thy model's performance using the validation set.
2. Thou shalt not overfit thy model.
3. Thou shalt not generalize outside thy scope.
4. Thou shalt scrub thy input of signal leakage.
5. Thou shalt expect model instability.
6. Thou shalt not use ML uncritically.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Overfitting creates a model that fails to generalize.



http://dominicwilcox.com/portfolio/bed/

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Signal leakage gives the model info that lets it cheat.

- Return to our model to detect people with diabetes. What if we add a feature that reports units of insulin injected each day?

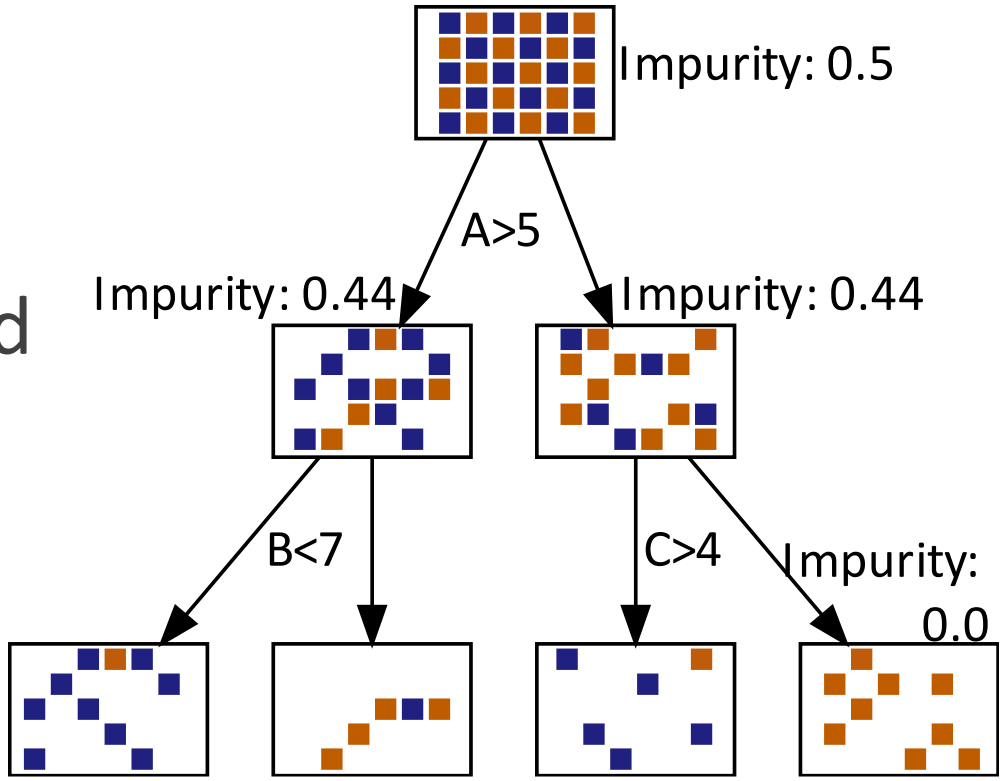
| Pt ID | Temp | BP | HbA1c | BMI | B Glucose | Insulin/day |
|---|---|---|---|---|---|---|
| DLT001 | 37C | 120/80mm | 48 mmol/mol | 26.2 | 5.3 mmol/L | 40 units |

- Because people without diabetes do not inject insulin, insulin injections are specific predictors and yet are useless to prediction.

https://insidebigdata.com/2014/11/26/ask-data-scientist-data-leakage/

# Decision Trees (Classification and Regression Trees)

- Each branch is decided by value on one feature.
- Gini impurity should decrease by branch.
- Stop splitting when impurity is zero or too few points remain.



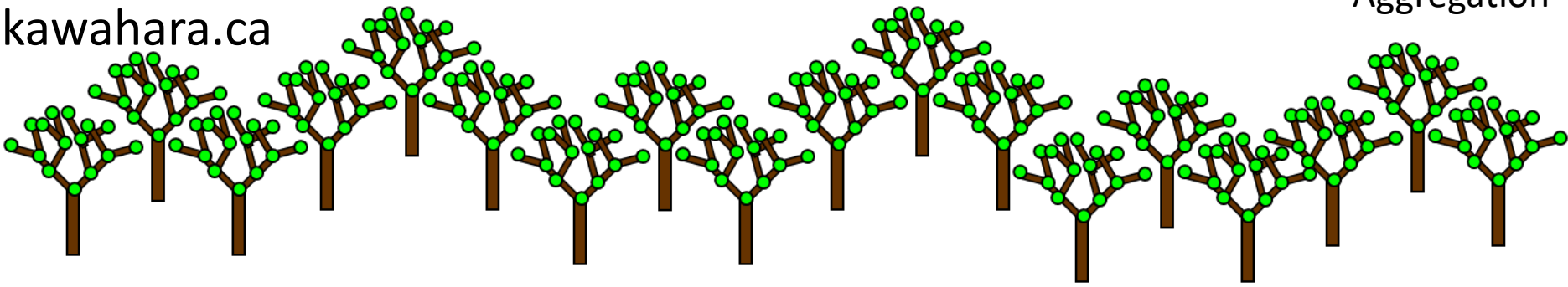
UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Random Forests are *ensembles* of Decision Trees

- An *ensemble* yields a *strong learner* by combining results from many *weak learners*.
- Many decision trees are built from samples of the subjects and samples of features.
- RF are relatively robust against odd data.

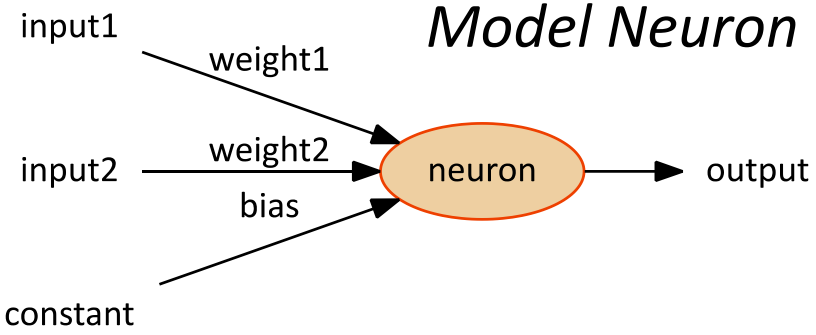
"bagging"
Bootstrap
Aggregation

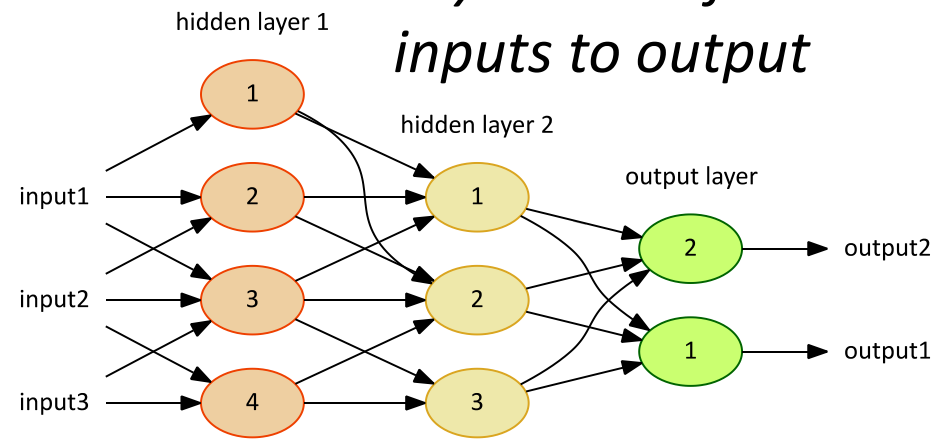
kawahara.ca



https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

# Artificial Neural Networks and Deep Learning




- *Back Propagation* updates hidden layers after each example to favor weights that minimize error in supervised learning.
- *Deep Feedforward Networks* employ many hidden layers to represent abstraction better and allow for Big Data input sets in *unsupervised* learning

GOOGLE'S ARTIFICIAL BRAIN LEARNS TO FIND CAT VIDEOS

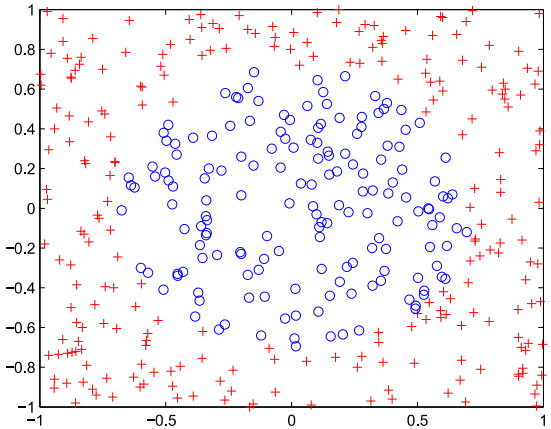
*Wired* (2012)

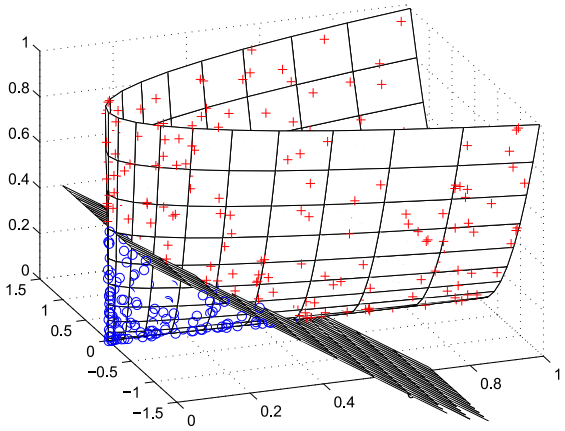
A. Krogh. *Nature Biotech*. (2008) 26: 195-197.

Y. LeCun et al. *Nature* (2015) 521: 436-444.

# Support Vector Machines

- *Kernel function* computes sample distances.
- *Hyperplane* is divider in high-dimensional space to separate cohorts best.
- *Support vectors* are tangents from hyperplane to closest training points.




Wikimedia Commons: Machine Learner



W.S. Noble. *Nature Biotech*. (2006) 24: 1565-1567.

# Takeaway Messages

- Including labels in training constitutes *supervised* learning.
- Many predictive models could be produced from a given training set. The best models excel in a never-before-seen test set, not in a reused validation set.