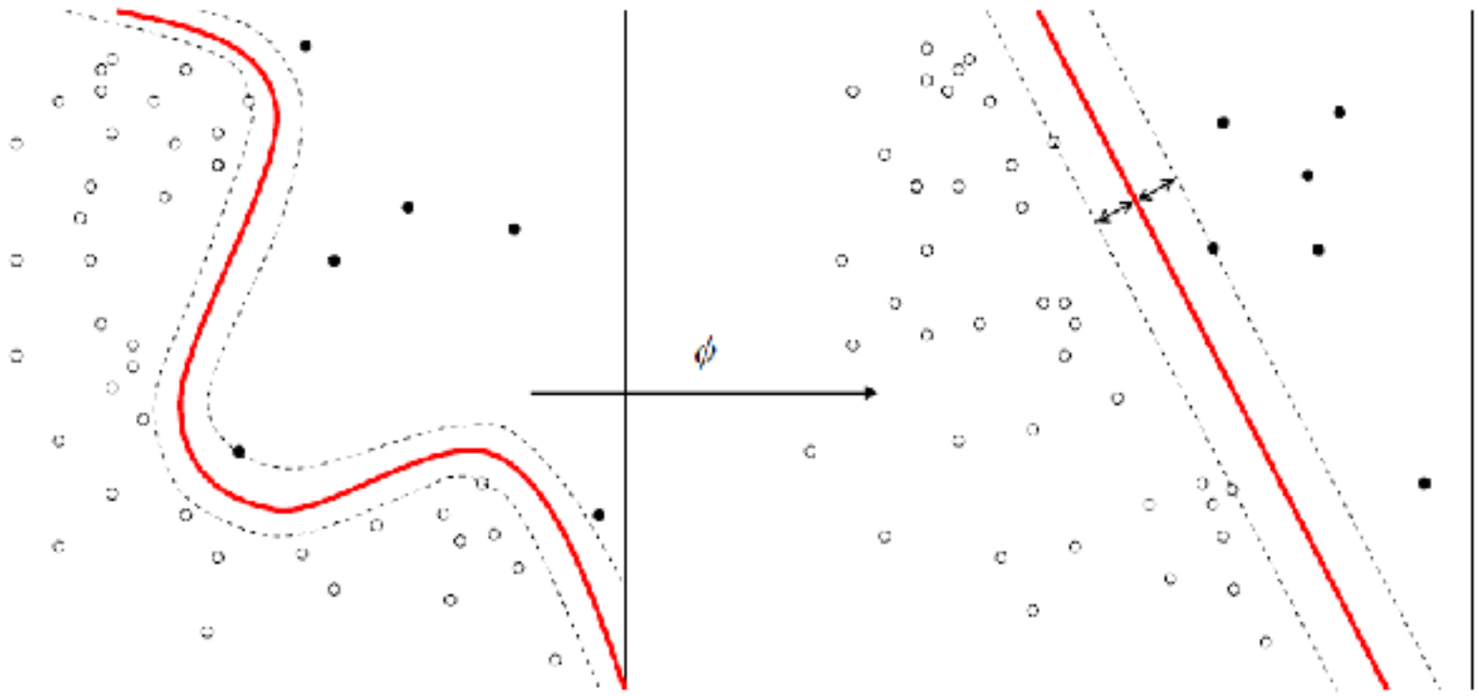


Introduction to Machine Learning

Yale University
Adam Chekroud

McCarthy Lab!
(2014)

Background

- BA Experimental Psych, MSc Neuroscience (Oxford)
- PhD Psychology (Neuroscience)

Interests

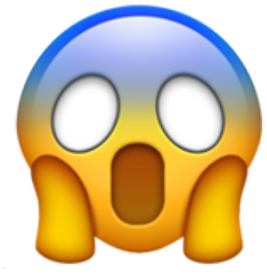
- Computational approaches to psychiatric illness
  - Optimising treatment selection in depression

Objectives

Input: training set $\{(x_i, y_i)\}_{i=1}^n$, a differentiable loss function $L($

Algorithm:

- Initialize model with a constant value:
  - $F_0(x) = \arg\min_\gamma \sum_{i=1}^n L(y_i, \gamma)$.
- For m = 1 to M:
  - Compute so-called "pseudo-residuals":
    - $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1, \ldots$
  - Fit a base learner $h_m(x)$ to pseudo-residuals, i.e. trai
  - Compute multiplier γ_m by solving the following one-
    - $\gamma_m = \arg\min_\gamma \sum_{i=1}^n L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right)$.
  - Update the model:
    - $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$.
- Output $F_M(x)$.



Objectives

- What is ML?
- Why is it useful?
- How do you do it?
  - Theory?
    - Feature selection
    - Algorithm selection
    - Model salidation
  - Practice?

What is machine learning?

- *“The goal of machine learning is to program computers to use example data or past experience to solve a given problem”*
- *Machine learning is a subfield of computer science that … explores the study and construction of algorithms that … can learn from and make predictions on data”*
- *”… algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions”*
    - Developing computer algorithms for doing — usually predicting — stuff.

What is machine learning?

Doing [stuff]:

- Systems that analyze past sales data to predict customer behavior
- Optimize robot behavior so that a task can be completed using minimum resources
- Extract knowledge from bioinformatics data

- Diverse range of methods and approaches
- Plays a central (often hidden) role in our life

Reinforcement Learning

A machine learns to interact with its (dynamic) environment (e.g. playing a computer game, driving a car)

Machine learning

"Supervised" Learning

"Unsupervised" Learning

Learn <u>without</u> structure (hard)

“Supervised” Learning

(a.k.a prediction, classification)

- We know what happened in the past (“training data”)
- Use this data to train an algorithm
- Use the algorithm to make predictions in the future, where we don’t know what will happen (“testing data”)

REVIEWS

Machine learning applications in genetics and genomics

Maxwell W. Libbrecht[1] and William Stafford Noble[1,2]

Abstract | The field of machine learning, which aims to develop computer algorithms that improve with experience, holds promise to enable computers to assist humans in the analysis of large, complex data sets. Here, we provide an overview of machine learning applications for the analysis of genome sequencing data sets, including the annotation of sequence elements and epigenetic, proteomic or metabolomic data. We

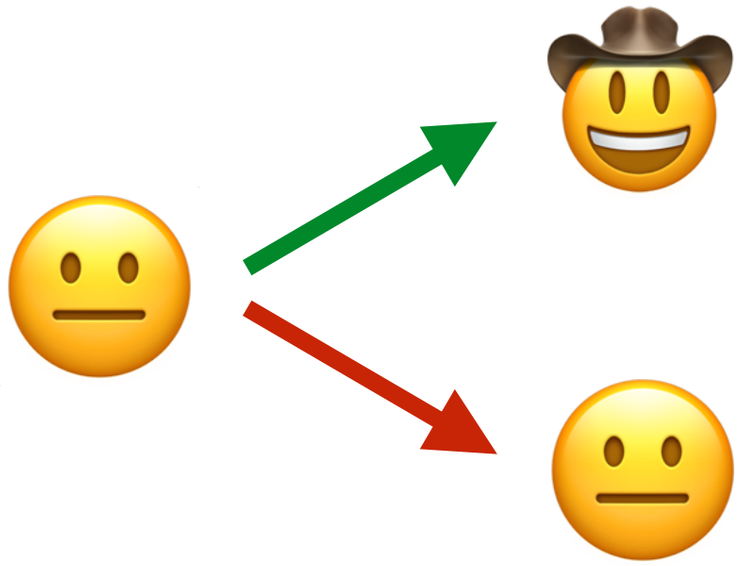
Libbrecht & Noble, 2015, *Nat. Rev. Genetics*

The Perfect Milk Machine: How Big Data Transformed the Dairy Industry

ALEXIS C. MADRIGAL | MAY 1, 2012 | TECHNOLOGY

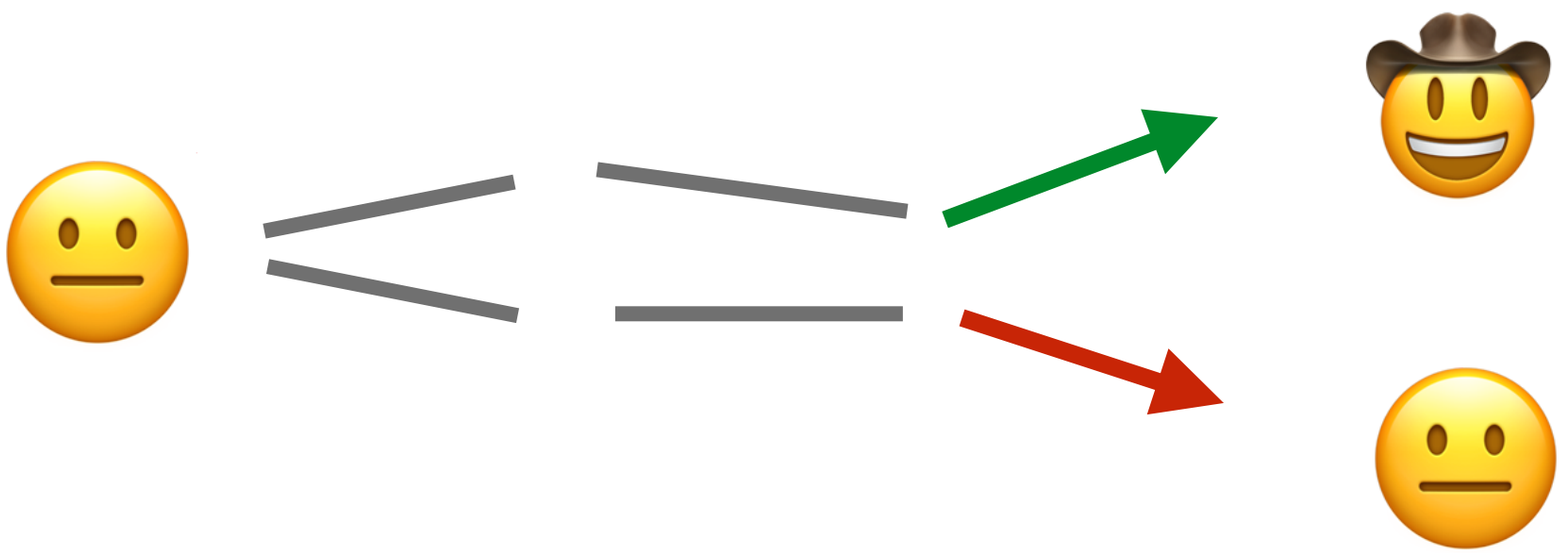
Dairy scientists are the Gregor Mendels of the genomics age, developing new methods for understanding the link between genes and living things, all while quadrupling the average cow's milk production since your parents were born.

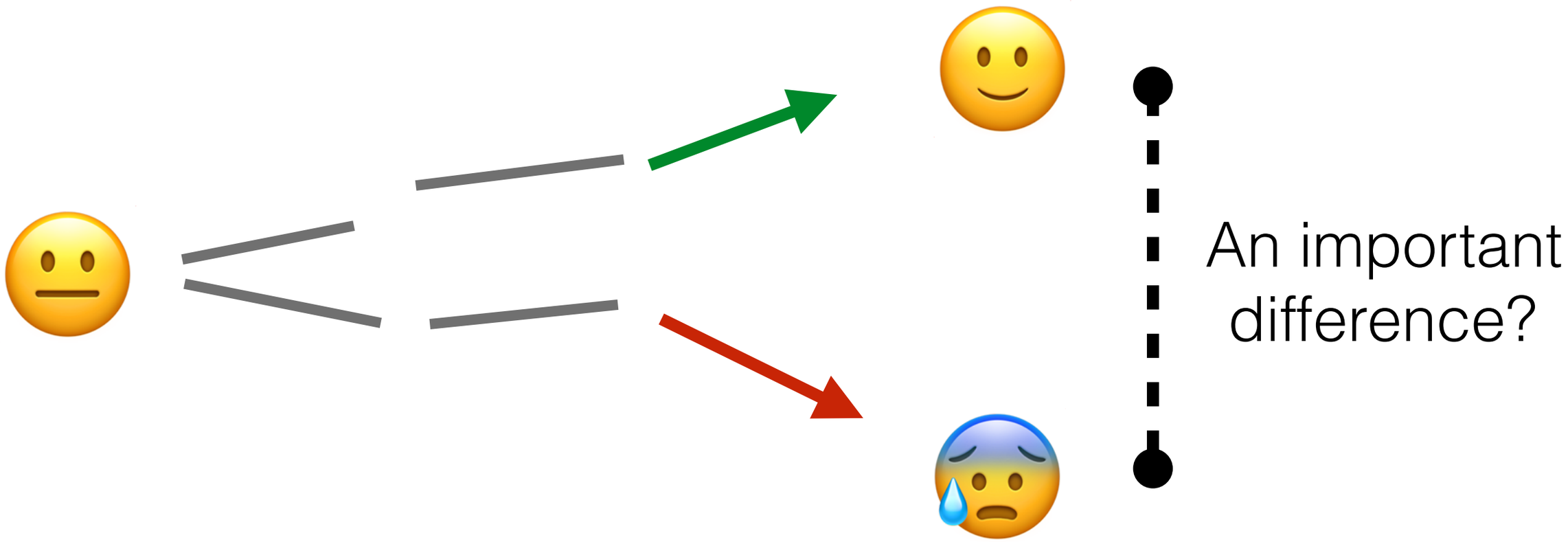
Why is it useful?



???

A long time?

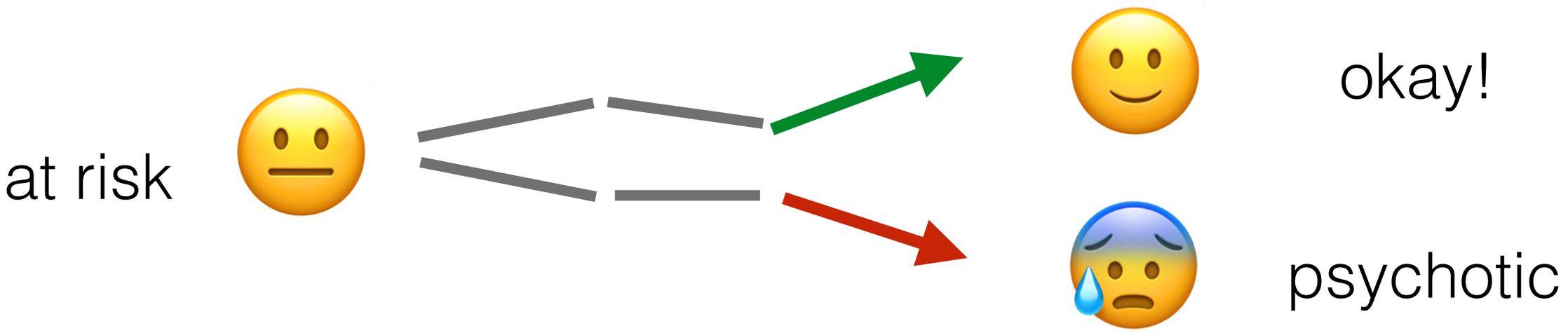




An important difference?

Use of Neuroanatomical Pattern Classification to Identify Subjects in At-Risk Mental States of Psychosis and Predict Disease Transition

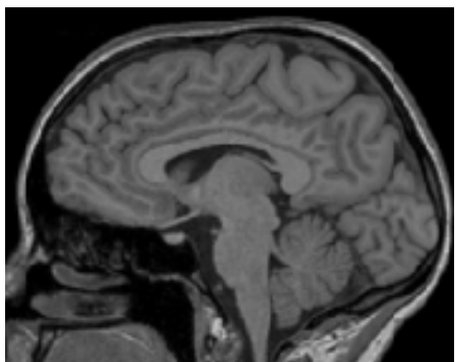
Nikolaos Koutsouleris, MD; Eva M. Meisenzahl, MD; Christos Davatzikos, PhD; Ronald Bottlender, MD; Thomas Frodl, MD; Johanna Scheuerecker, BA; Gisela Schmitt, MD; Thomas Zetzsche, MD; Petra Decker, BA; Maximilian Reiser, MD; Hans-Jürgen Möller, MD; Christian Gaser, PhD

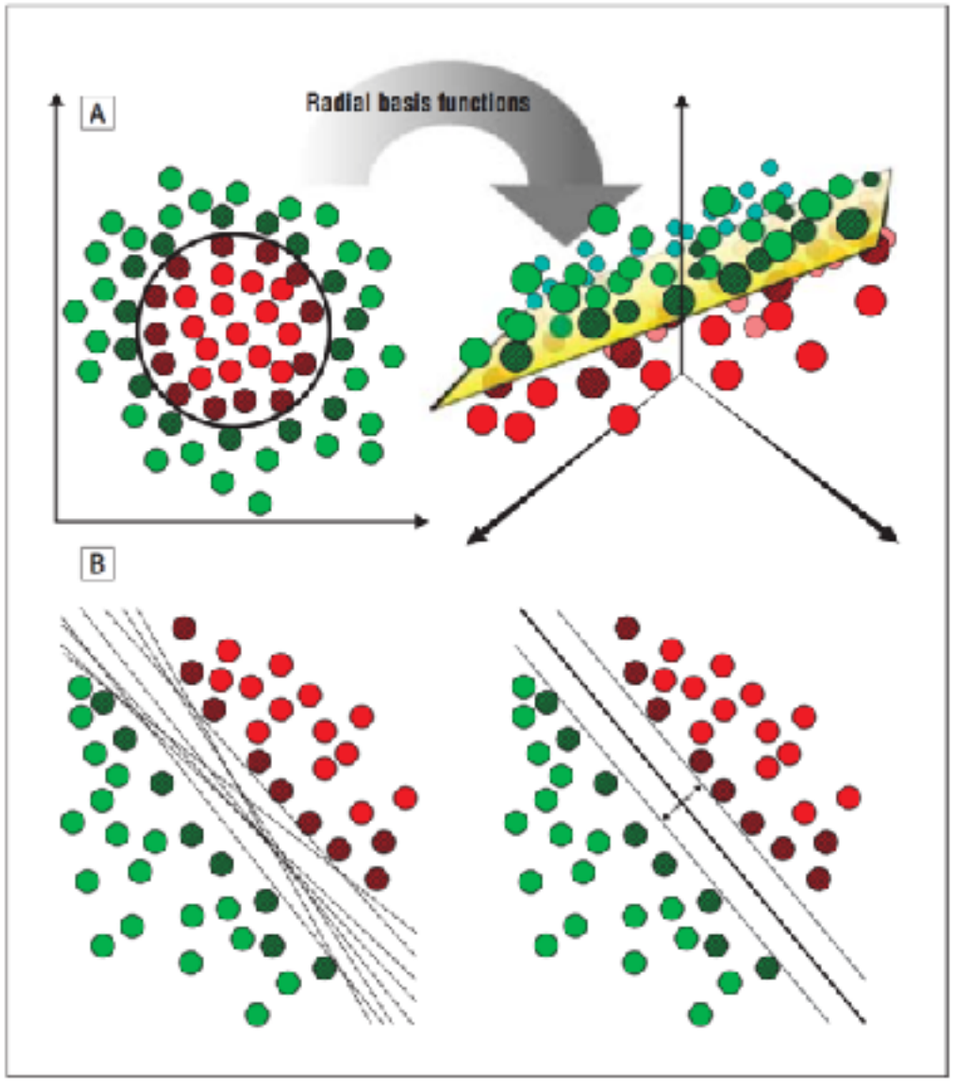


Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study

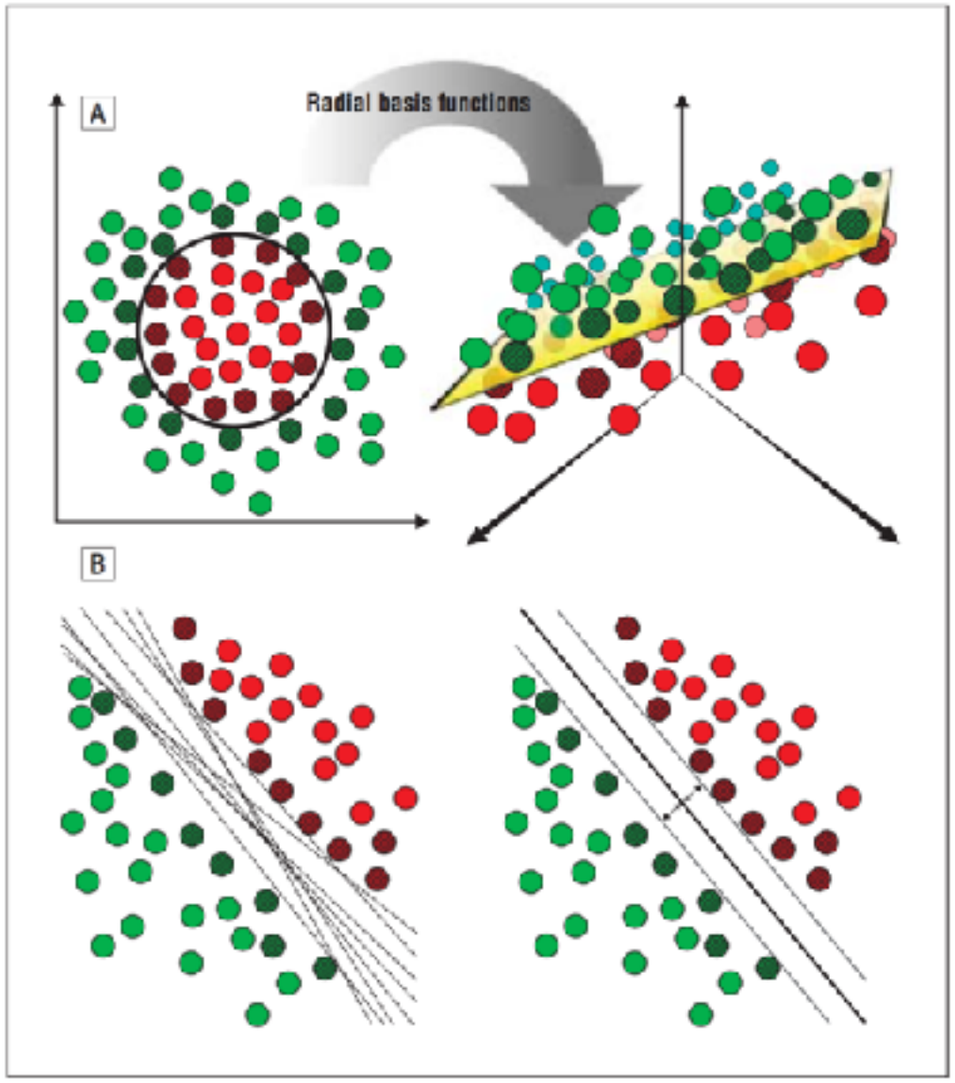
J. Mourao-Miranda[1,2], A. A. T. S. Reinders[3,4], V. Rocha-Rego[1], J. Lappin[3], J. Rondina[1], C. Morgan[3], K. D. Morgan[3], P. Fearon[3], P. B. Jones[5], G. A. Doody[6], R. M. Murray[3], S. Kapur[3] and P. Dazzan[3,7]*

(reality!)




A
Radial basis functions
B

???


A
Radial basis functions
B

The machine learning “pipeline”

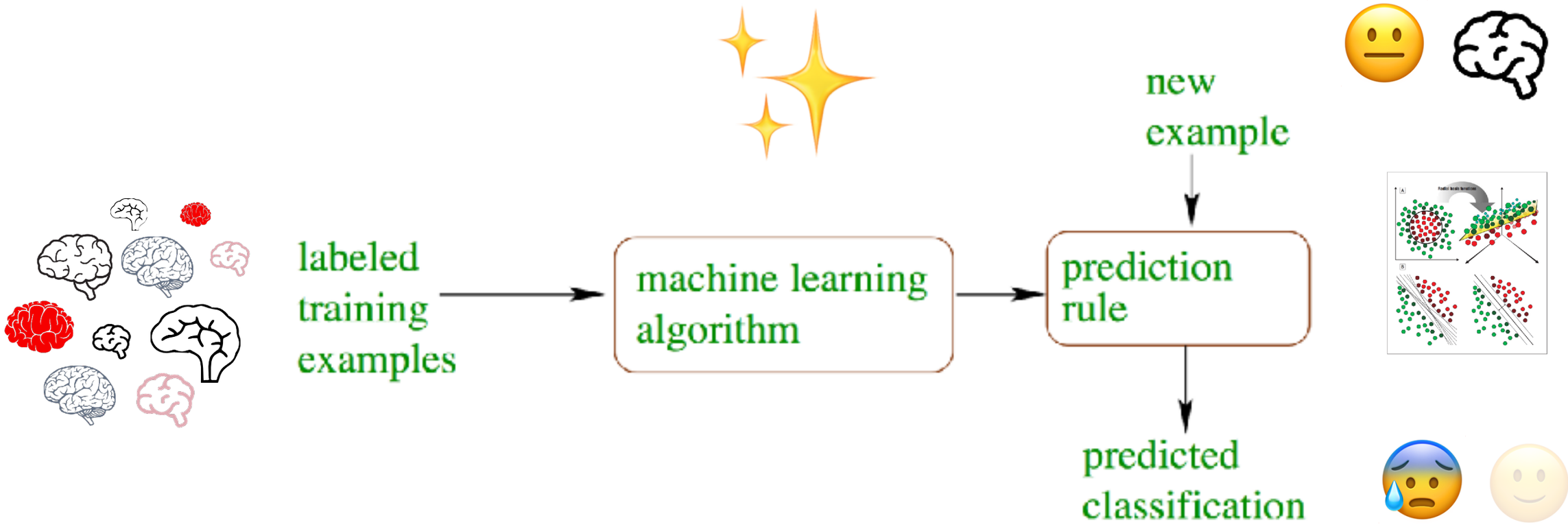
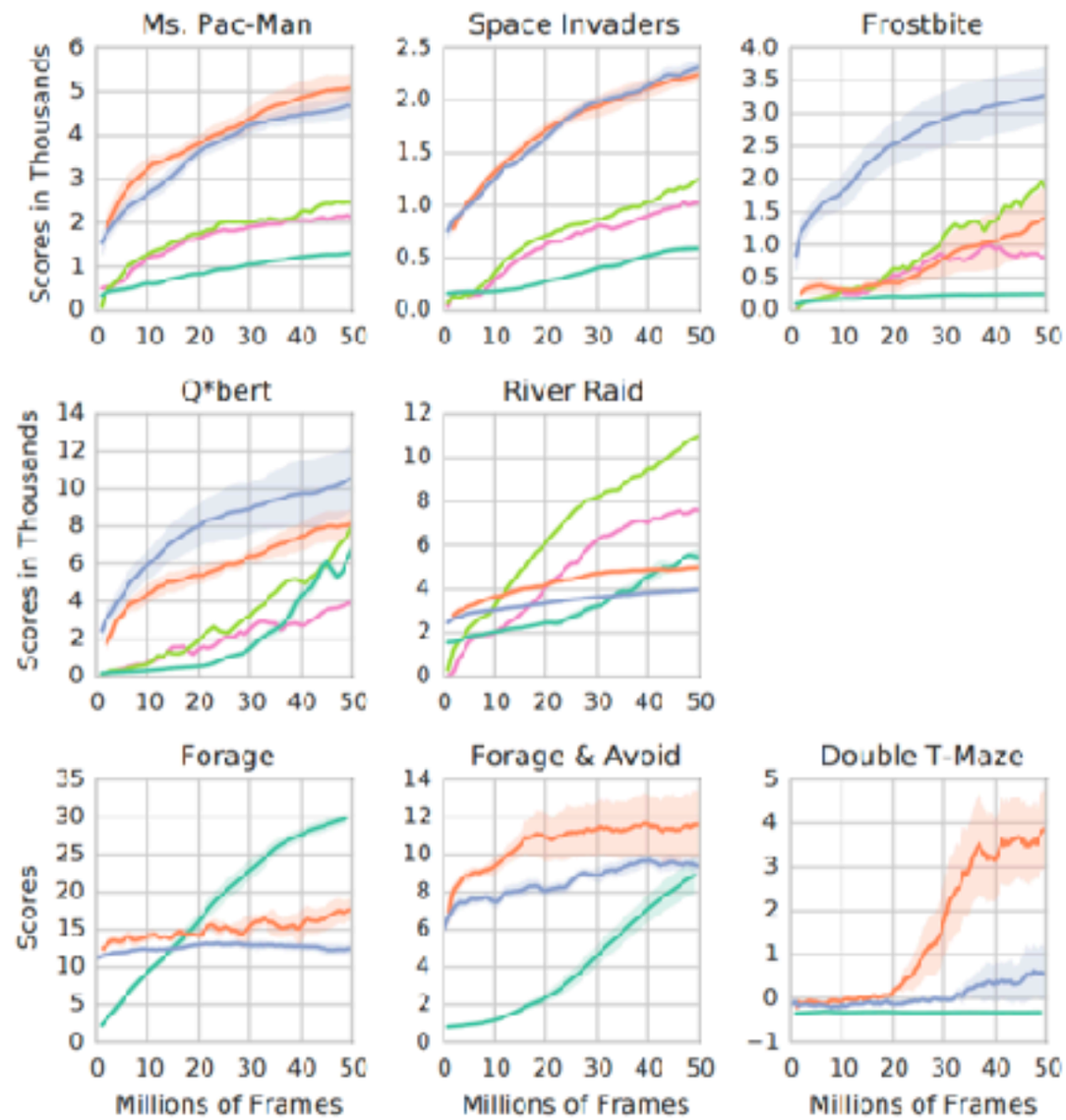


Figure 1: Diagram of a typical learning problem.

Machine learning depends on the **quantity** and **quality** of data



*Google's ATARI playing algorithm played 200 *million* games

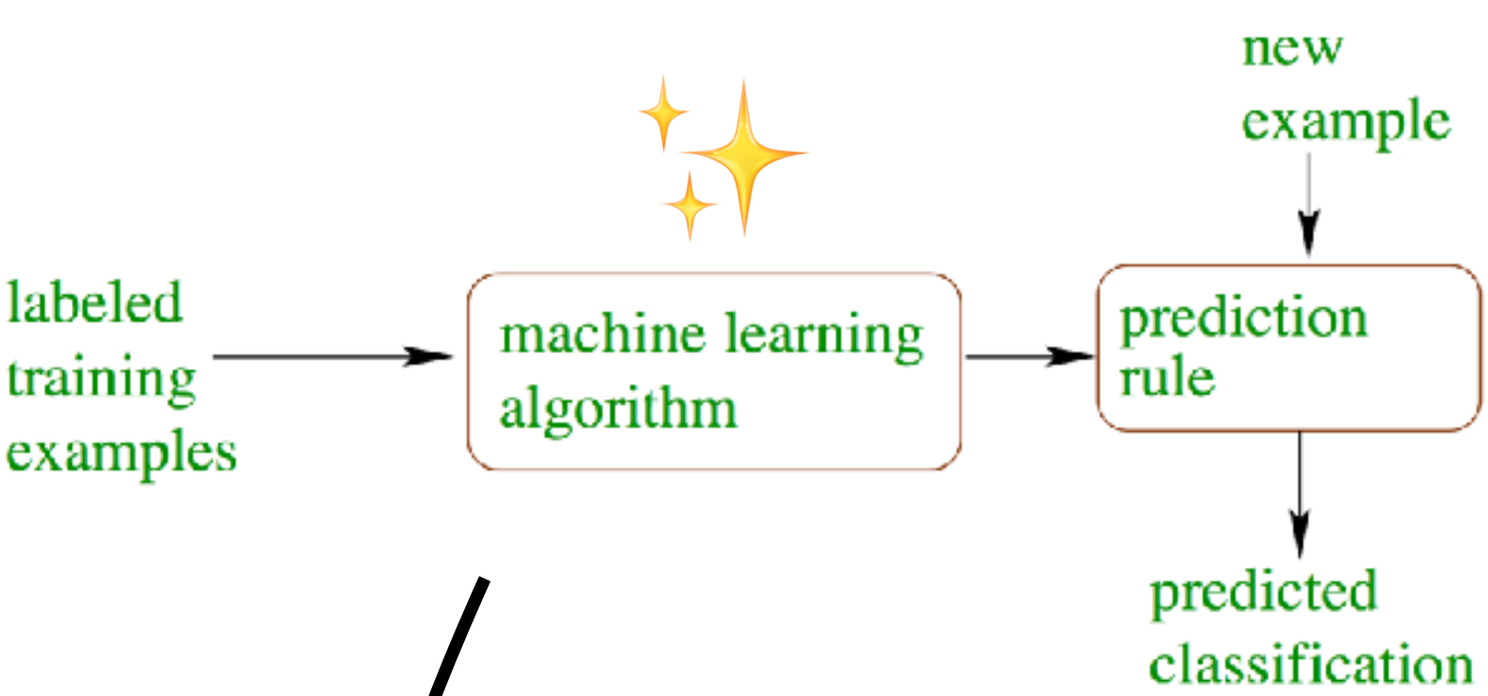
… & it "mentally replayed" each one 8 times over

Groundwork….

Getting data

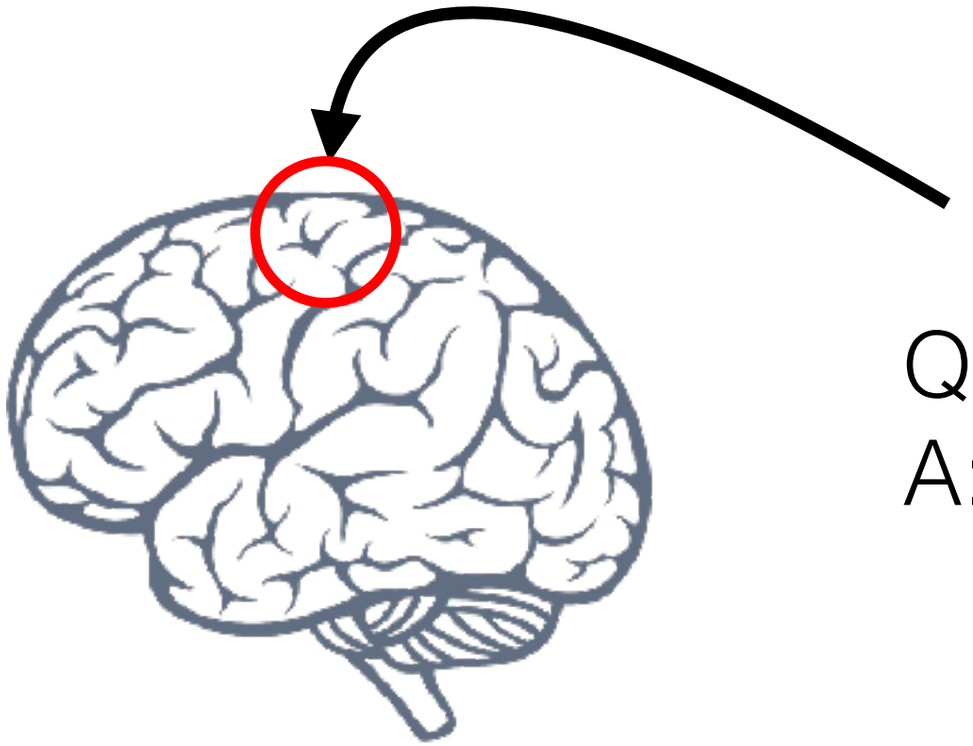
- ML is not an alternative to collecting a lot of data
    - methods were developed for problems with hundreds of thousands of data points

- ML is not a silver bullet for noisy data
    - imprecise data
    - limited signal
    - consistency in data



1. Feature selection (what info?)
2. Algorithm selection (how do we combine it?)
3. Model validation (does this “work”?)

What is a feature? (a variable)



Q: “how deep is this bit”?
A: 5

other e.g.

- time of day
- number of fb friends
- age/race/gender
- what time you go to bed
- do you have the ABCD1 gene?

Motivation for feature selection

Why?

- Models don't like too much information
- We can *create* features!

Tailoring data to our problem

- What information is predictive?
- How much do we need?
- Do we need to reduce or transform the data?
- Can we use our own knowledge to make better (custom) predictors

Feature selection

selecting a subset of (relevant) features for model construction

Motivation

Some of our data is *redundant* or *irrelevant*

- *Would your favorite color predict conversion to psychosis?*

Result?

1. Simplifies models to improve interpretability
2. Alleviates computational demand
3. Improves generalizability by avoiding *overfitting*
    - “reduction of variance”

Methods for feature selection

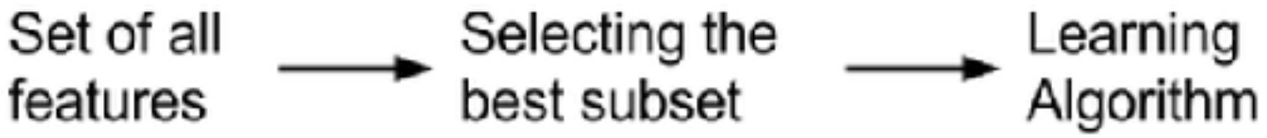
Choosing features that give us as good or better performance with less data

How?

<u>Three general classes:</u>

- Filter method
  - Minimum correlation coefficient!
- Wrapper method
  - Stepwise, or recursive feature elimination
- Embedded method
  - Least angle regression, regularization

Filter methods



Apply a statistical measure to “rate” each variable

Use these ratings to decide which variables to keep

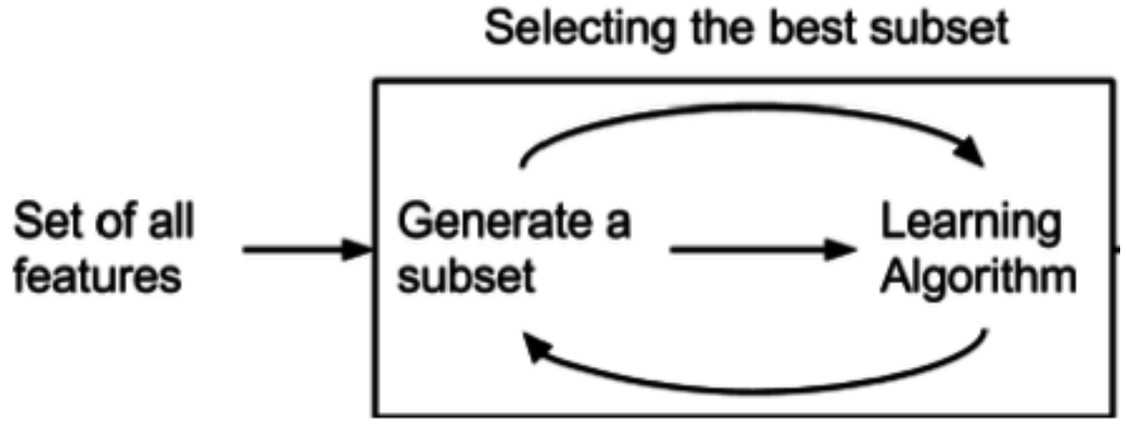
Which measures?

- Typically univariate (although decreasingly so)
    - e.g. p-val from GLM contrast in fMRI
- Can consider feature independently
    - e.g. variance of predictor
- Or with regard to the target
    - e.g. correlation coefficient, chi-squared

Pros and cons?

- Pro: *really* quick, quite effective at eliminating uninteresting variables, doesn’t usually overfit
- Con: can select redundant variables (keep two good, but correlated variables)

Wrapper methods



- Evaluate subsets of features in combination
- Use performance of model to decide which *group* of variables to keep

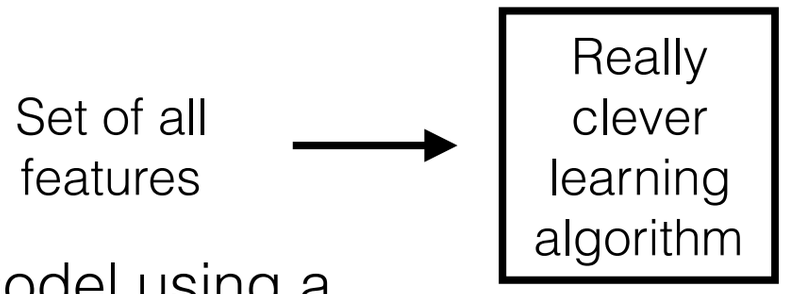
Search procedures

- Forward/Backward stepwise selection
- Recursive feature elimination
- Advanced: genetic algorithms, simulated annealing

Pros and cons?

- Pro: can detect interactions, usually give good performance, advanced methods can technically find “optimal” solution
- Con: *really* slow. Serious risk of overfitting.

Embedded methods



Definition: a machine learning algorithm that returns a model using a limited number of features

- Variable selection is built in ("embedded") to the learning algorithm
- Typically this is done using "regularization" methods

Examples:

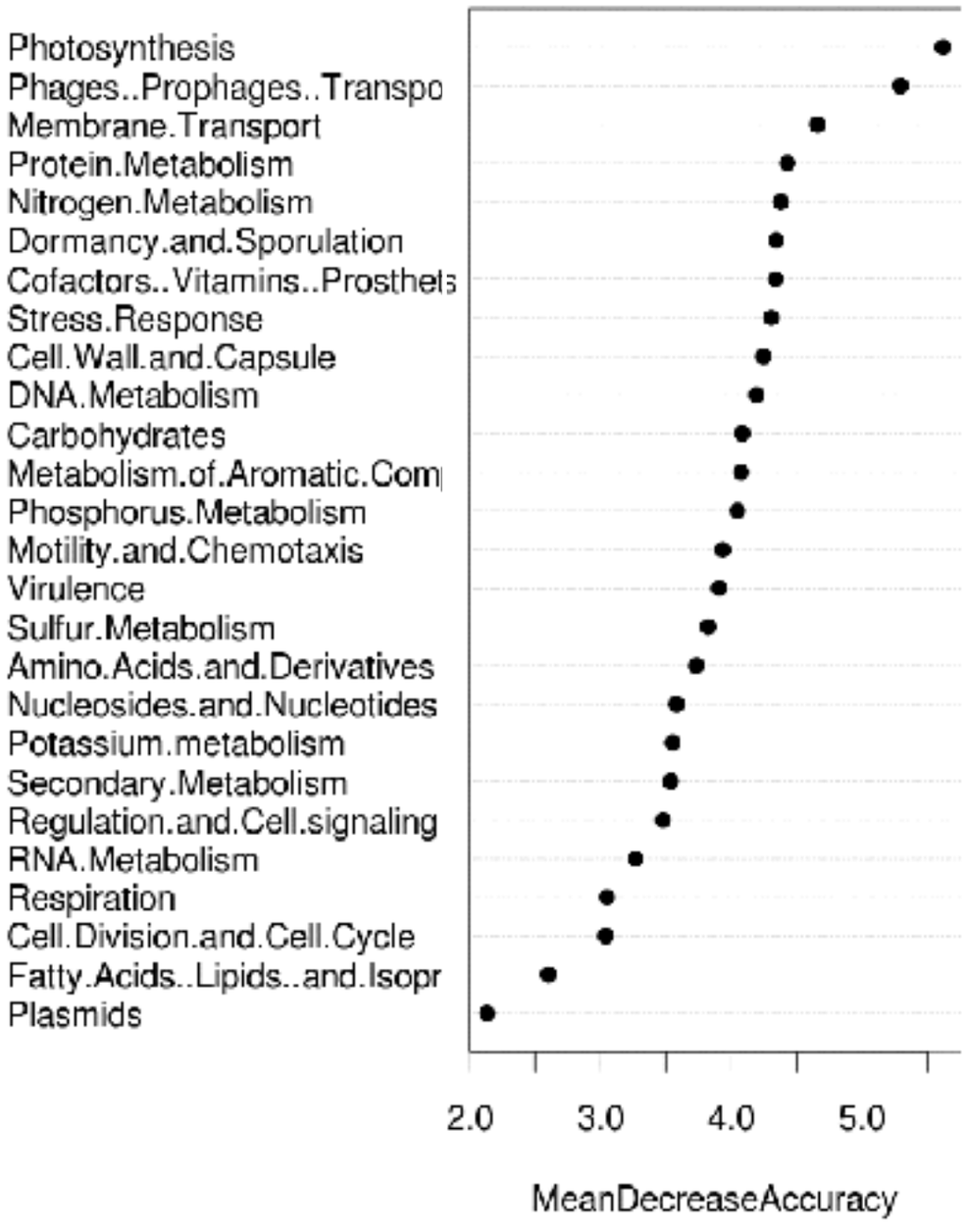
- Decision trees!
- Penalized regression
  - L1-norm: LASSO, least angle regression
  - L2-norm: Ridge regression
  - Blend: Elastic net regression
- Fancier adaptations of some models: e.g. support vector machines

Pros and cons?

- Pro: can detect interactions, don't fit many models, gives good performance
- Con: Limited selection of algorithms, technically more challenging.

Variable Importance Plot

- FS outcomes usually represented with variable importance plot

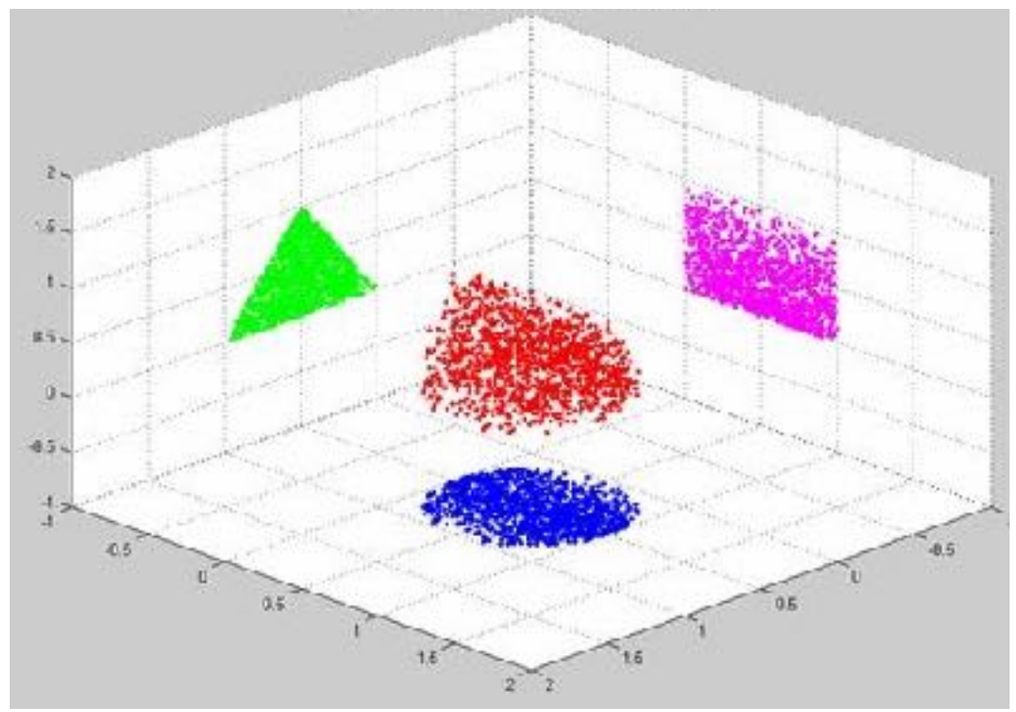


- Visual ranking of predictor utility
- Various ranking metrics available
  - r value
  - t-stat
  - reduction in RSS
  - permutation based measures

Feature Selection - Warnings

Where things can go wrong

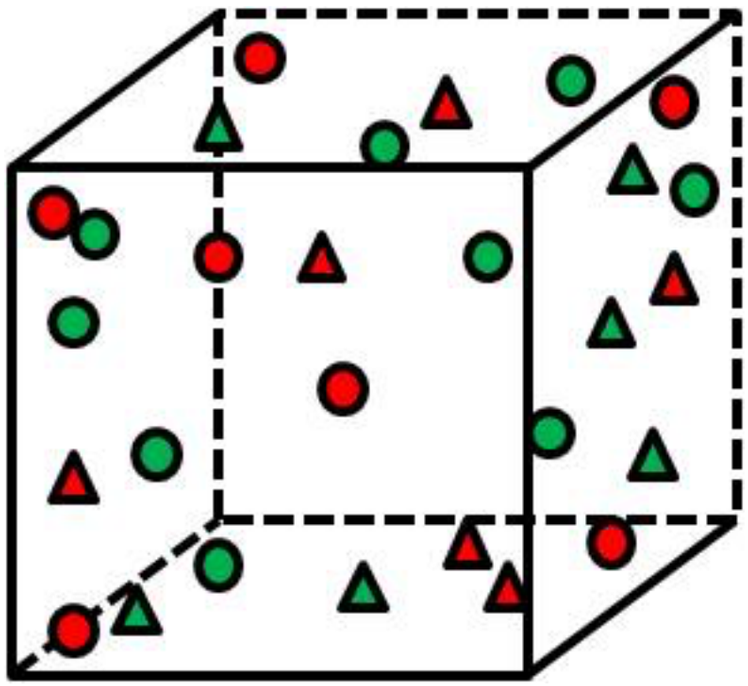
- Feature subset misrepresents reality! >
- Crash your computer

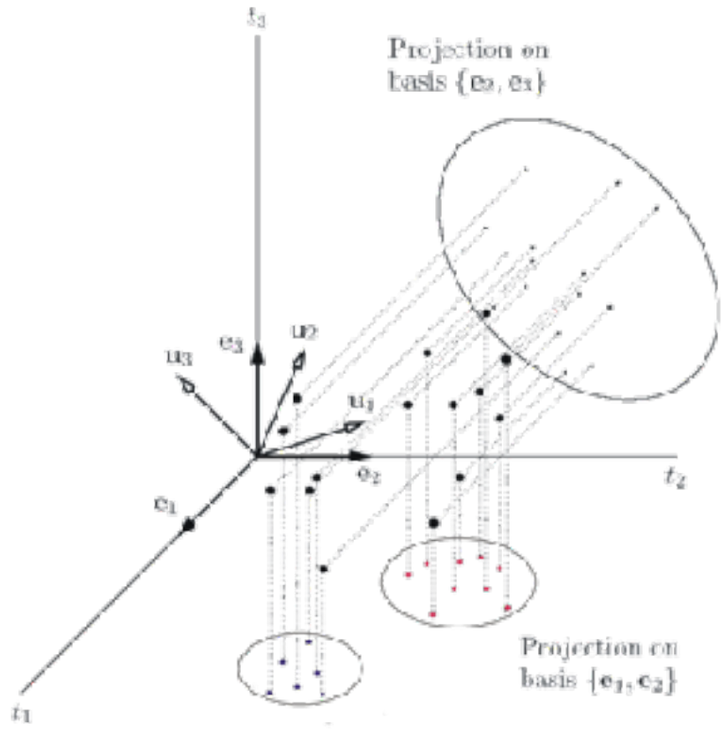


- Overfitting
- Multiple model testing

- Using same data to pick features and examine model performance
- Equivalent to test set peeking!

“The curse of dimensionality”





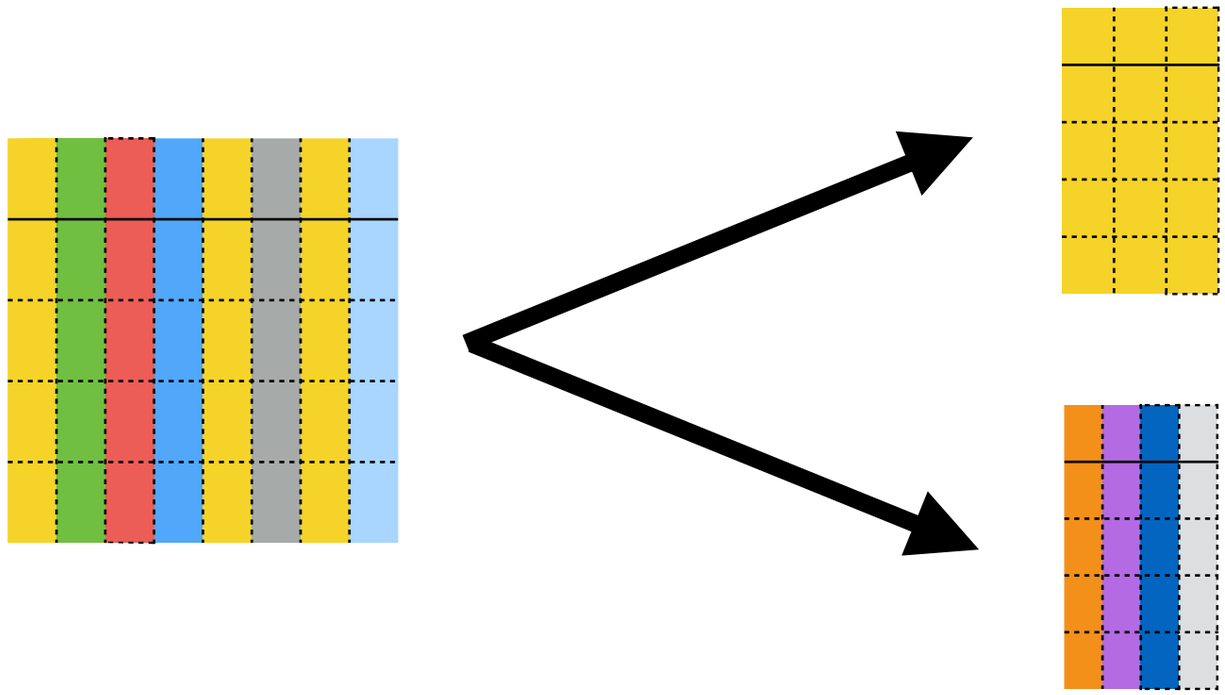
Data (usually) exist in a high-dimensional space

- Points may be close on one dimension, but very sparse in complete high-dimension space

Learning about the structure of this space is difficult!

- Requires a huge amount of data to ensure several samples with each combination of variables

Relationship with dimensionality reduction



FS: Identify three best “raw” variables

DR: Combine all variables into smaller number

- Both seek to reduce the number of attributes in the data
  - Dimensionality reduction methods create combinations of features
  - Feature selection methods include or exclude attributes that already exist, without changing them

“The curse of dimensionality”

Can we preserve informative structure in a lower-dimensional space?

We can try:

- PCA/ICA, non-Negative Matrix Factorization (nNMF), locally linear embedding

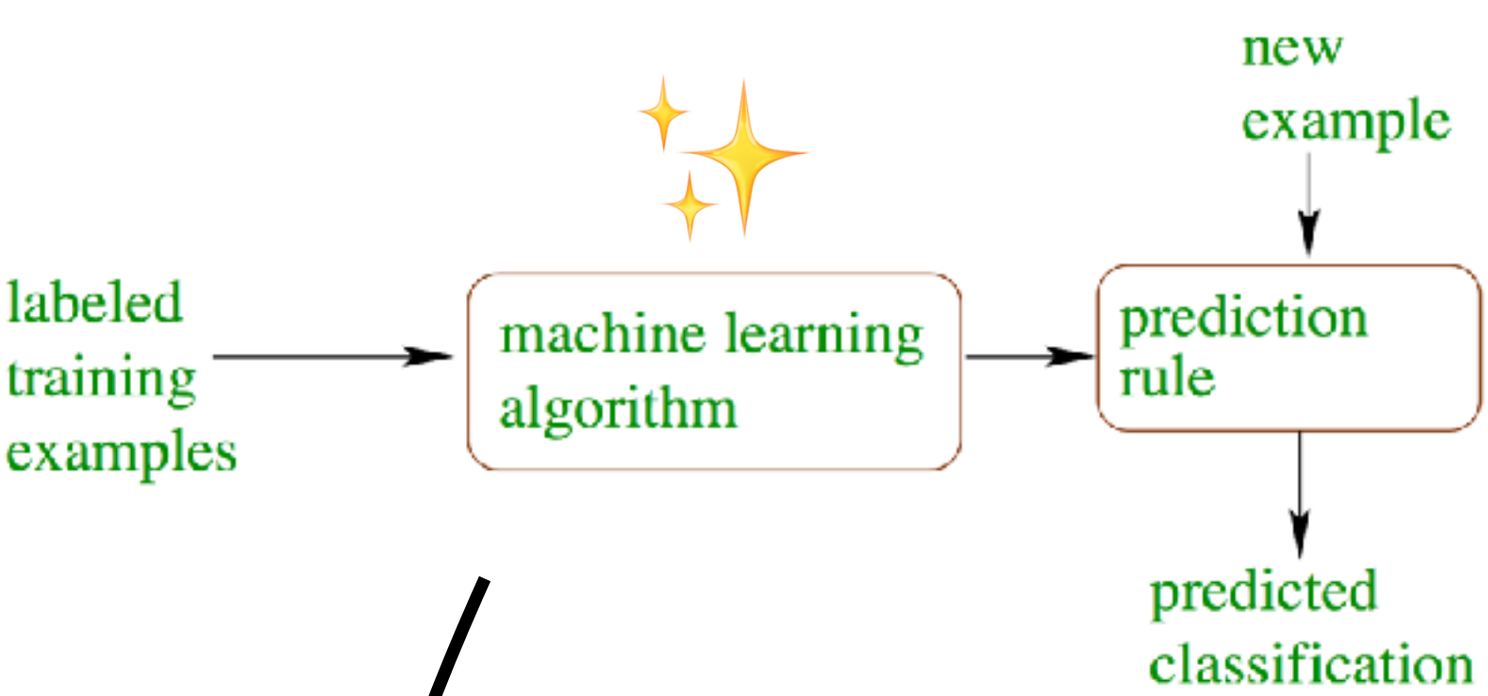
Not easy:

- Reducing dimensionality reduces information available for prediction
- Do it wrong, destroy your data

Knowledge discovery by accuracy maximization

Stefano Cacciatore[a,b,c], Claudio Luchinat[a,d,1], and Leonardo Tenori[d]

[a]Magnetic Resonance Center (CERM) and Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy; [b]Department of Medical Oncology, Dana–Farber Cancer Institute, Harvard Medical School, Boston, MA 02115; [c]Metabolomics Platform, Spanish Biomedical Research Center in Diabetes and Associated Metabolic Disorders, Rovira i Virgili University, 43007 Tarragona, Spain; and [d]FiorGen Foundation, 50019 Sesto Fiorentino, Italy

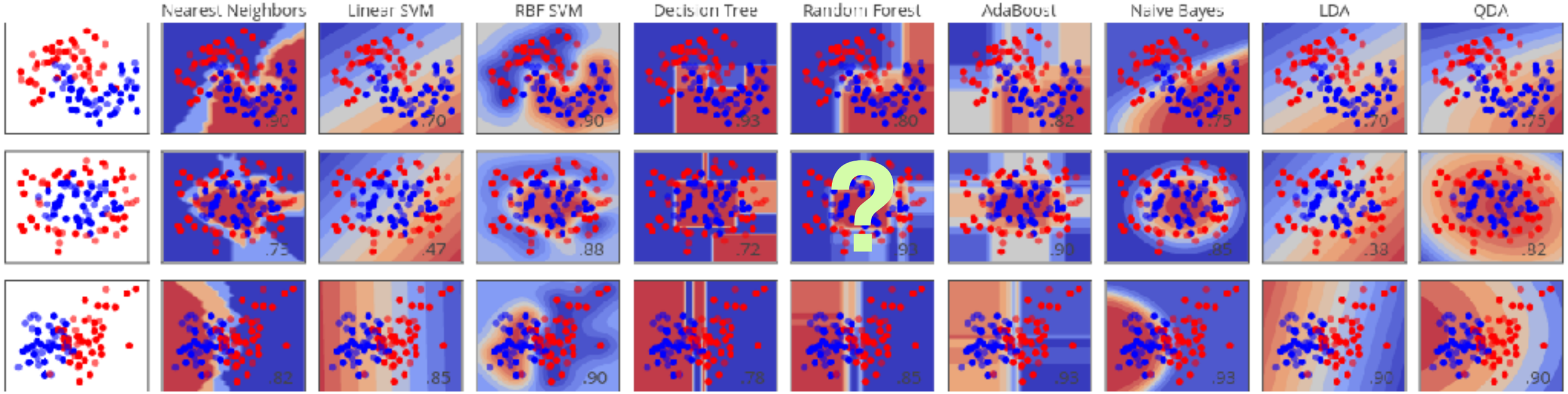


1. Feature selection (what info?)
2. Algorithm selection (how do we combine it?)
3. Model validation (does this “work”?)

Algorithm Selection

?

?



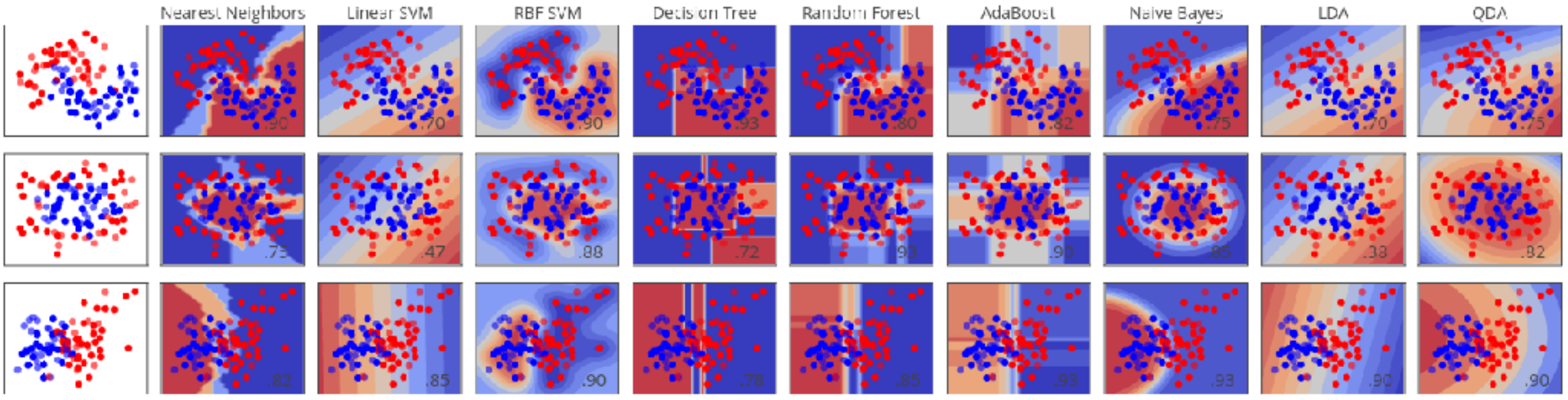
?

?

- “Algorithm Selection: Try all of them*”

* only caveat: with appropriate train-test precautions

Algorithm Selection



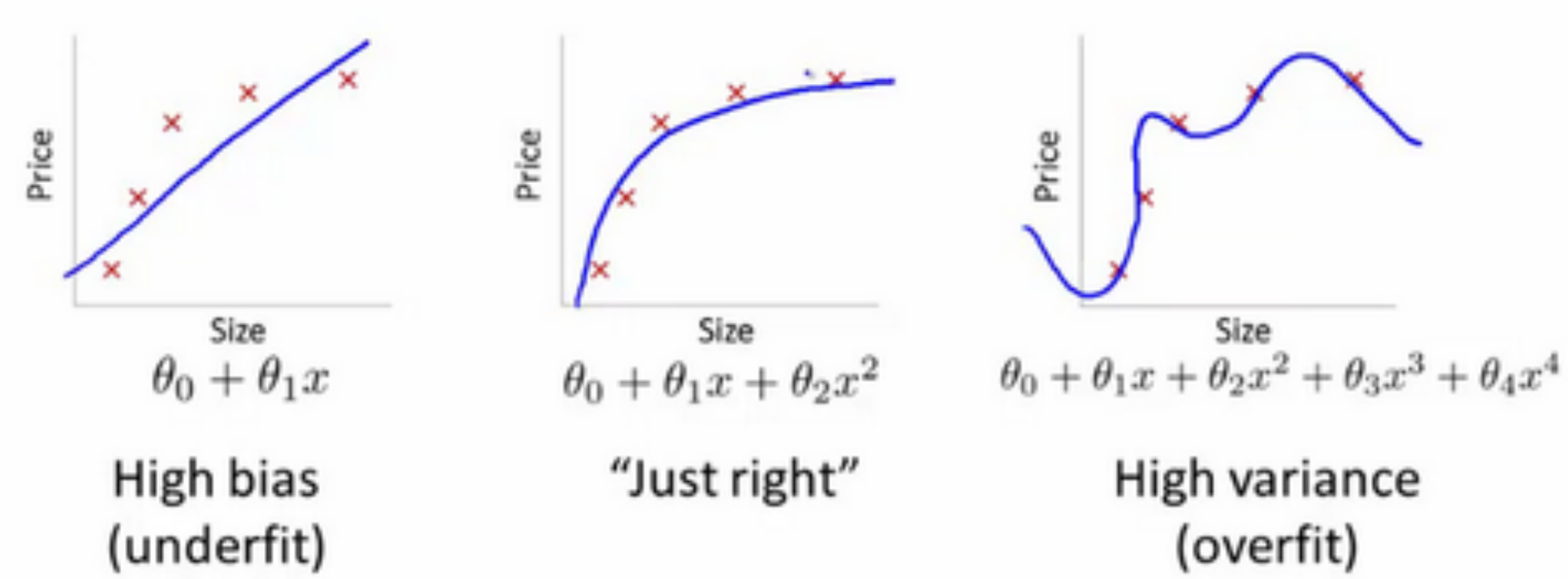
- Use predictors to “train” predictive algorithm
- Algorithm “learns” optimal mapping between your data and the outcome
    - Form and scope of mapping varies across algorithms
- Use this mapping to guess what happens for unseen data

- Many extremely powerful algorithms available “off the shelf”

Choosing an algorithm

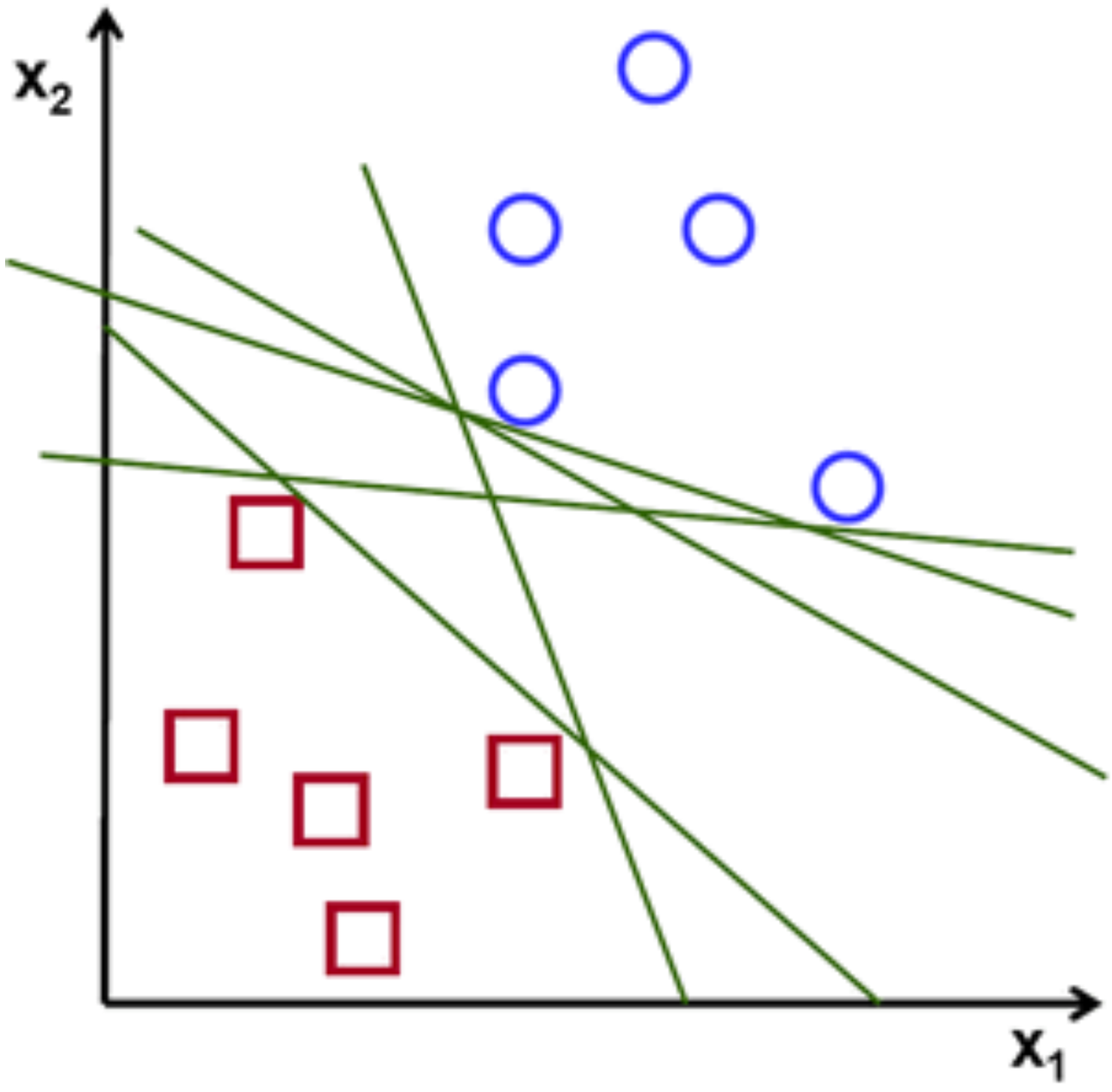
Accuracy

- More accuracy is not always better (!!)
- Is an approximation sufficient?
- Is an approximation better?!
    - Simpler models less likely to over fit



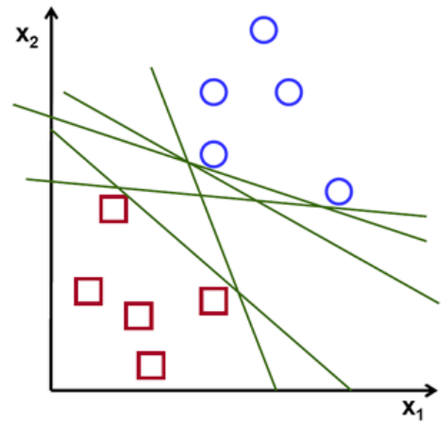
c.f. Bias-variance tradeoff

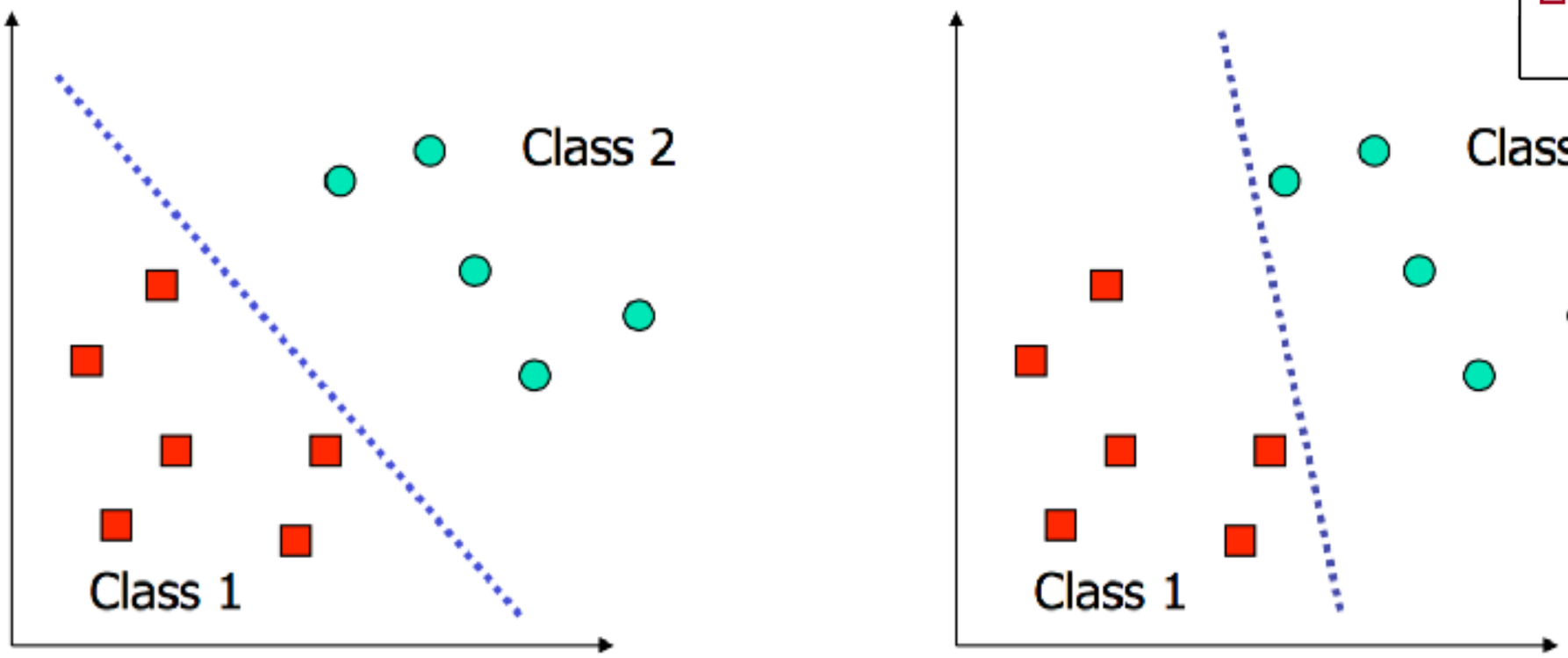
Support Vector Machine



Aim: find a boundary that (optimally) separates our classes

Support Vector Machine

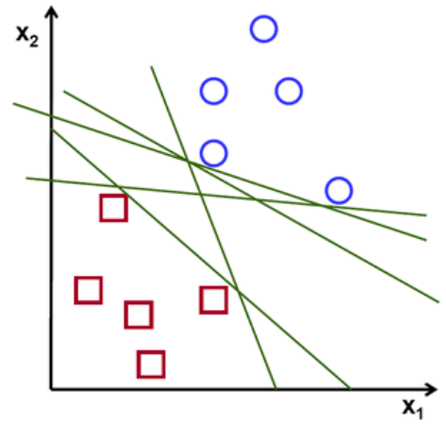


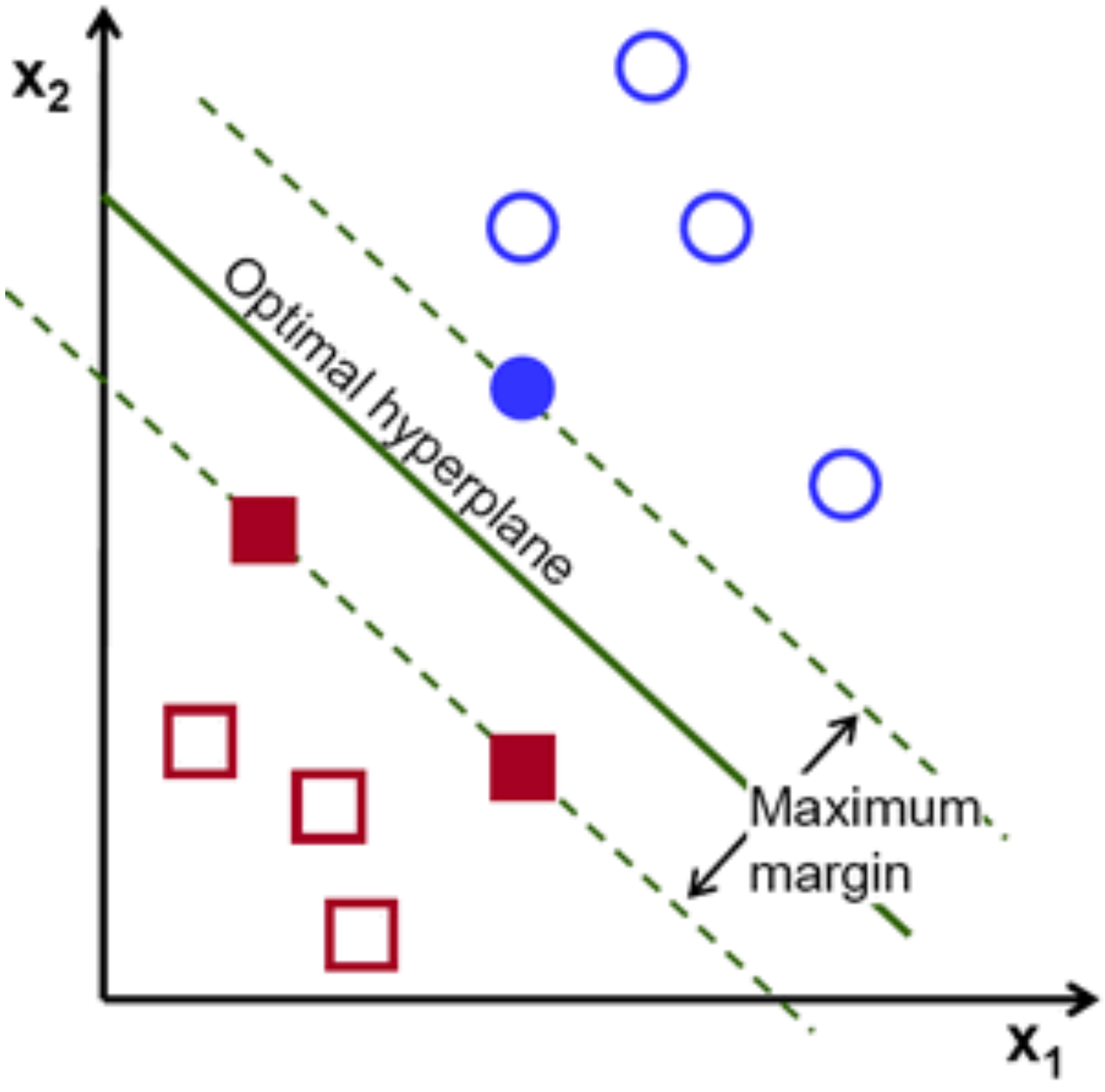


(Bad boundaries)

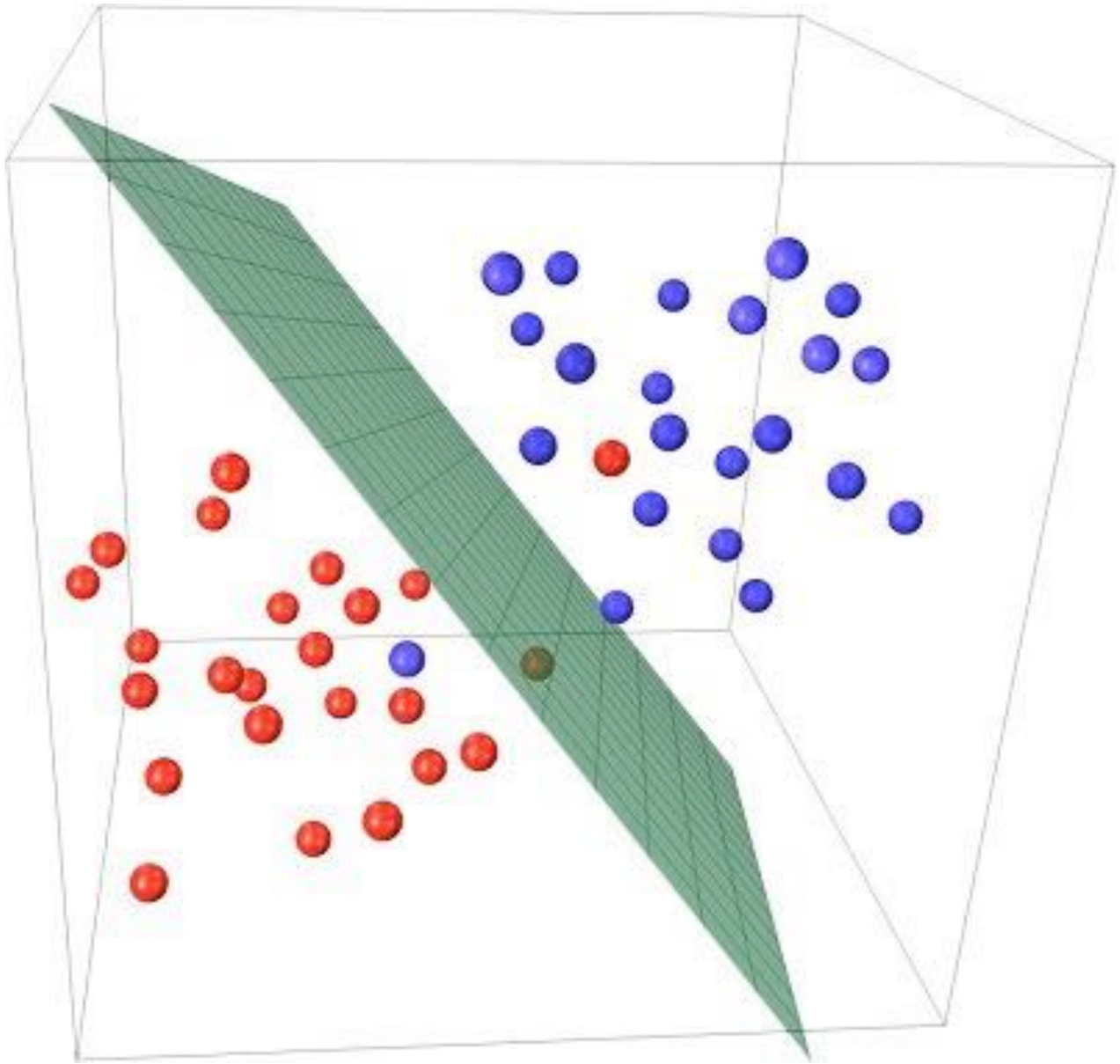
- Max-margin theory: boundary should be as far away from the data as possible

Support Vector Machine



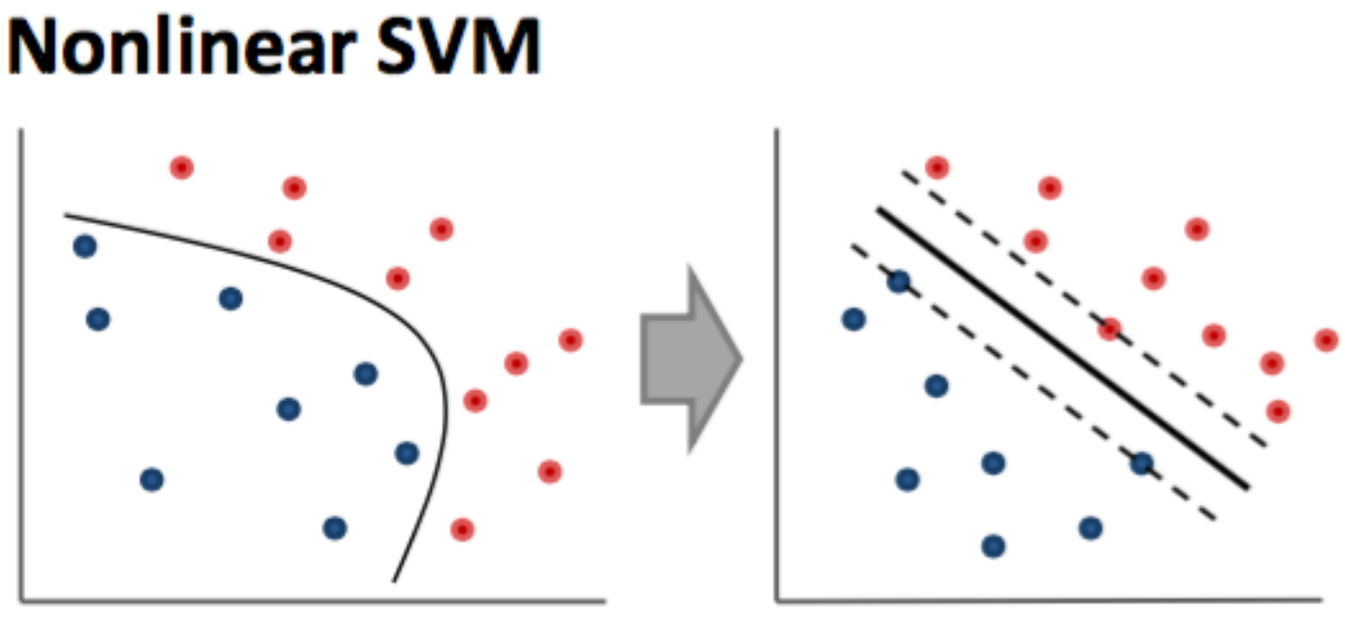


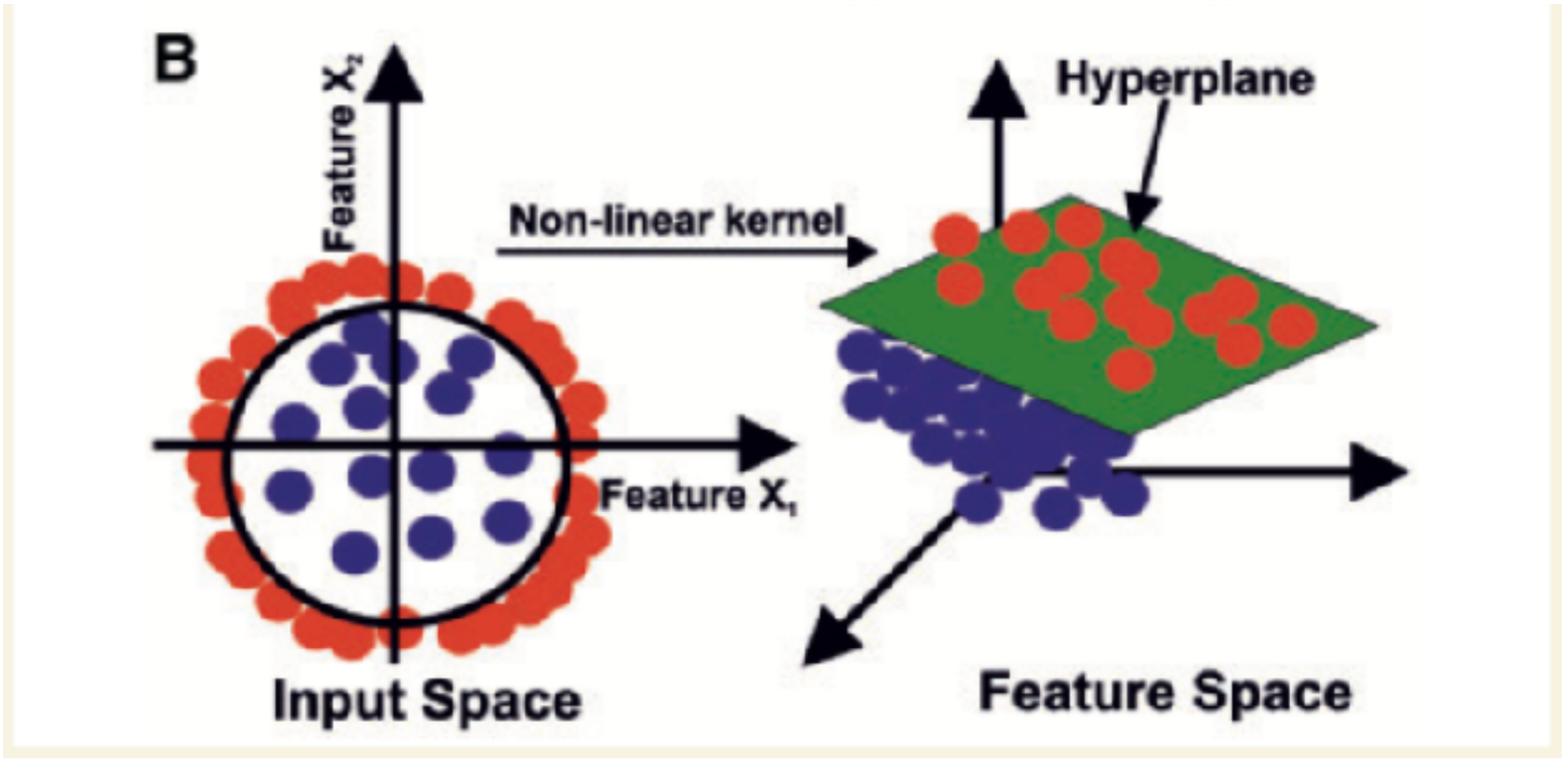
Support Vector Machine

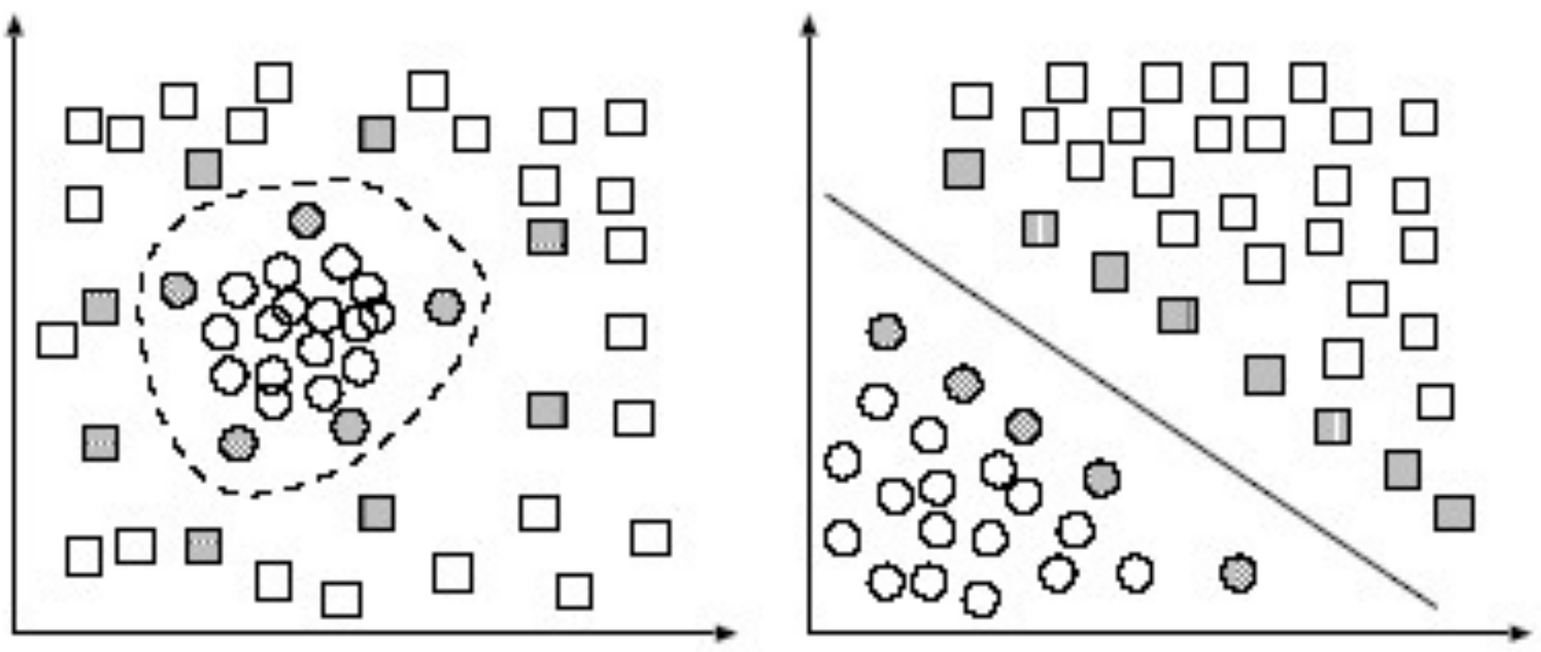


(hyperplane is not just in 2-3D)

Support Vector Machine

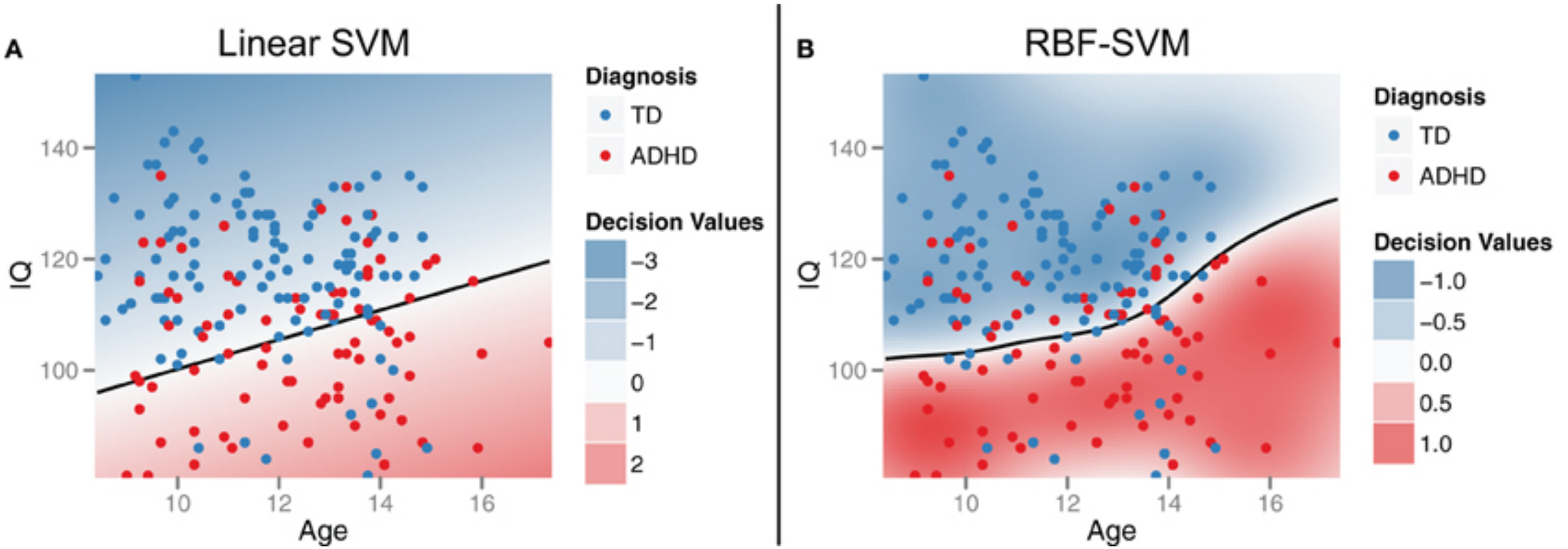




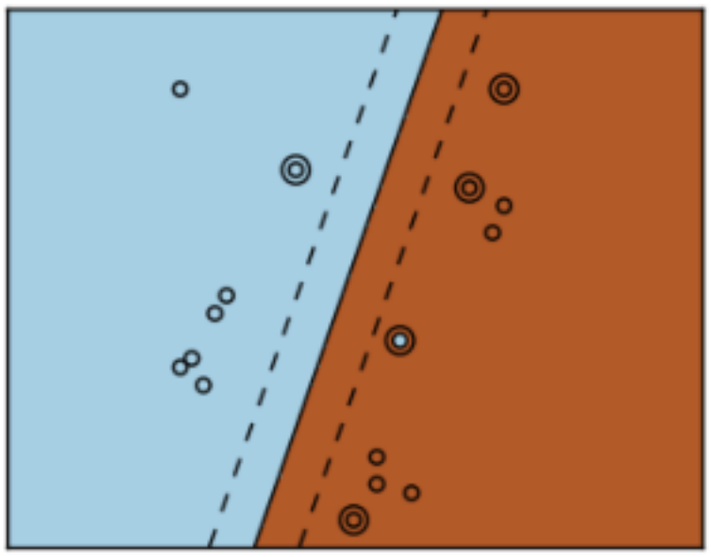


(a) Radial Basis Function (b) RBF mapping

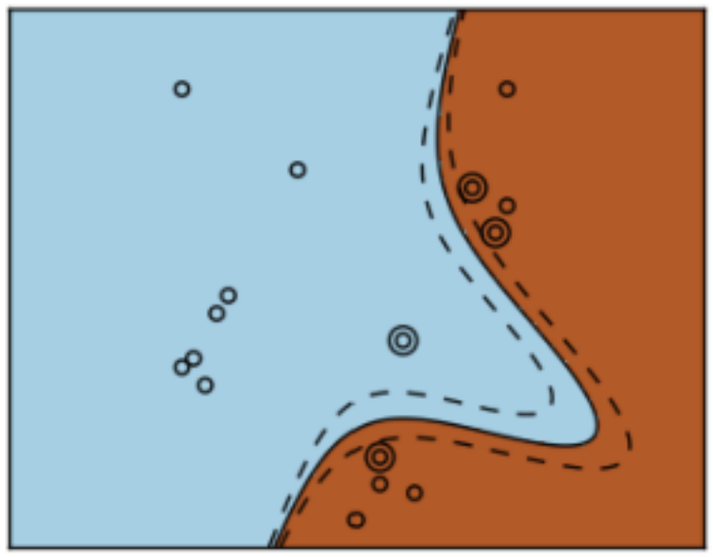
Separable classification with Radial Basis kernel functions in different space. Left: original space. Right: feature space.



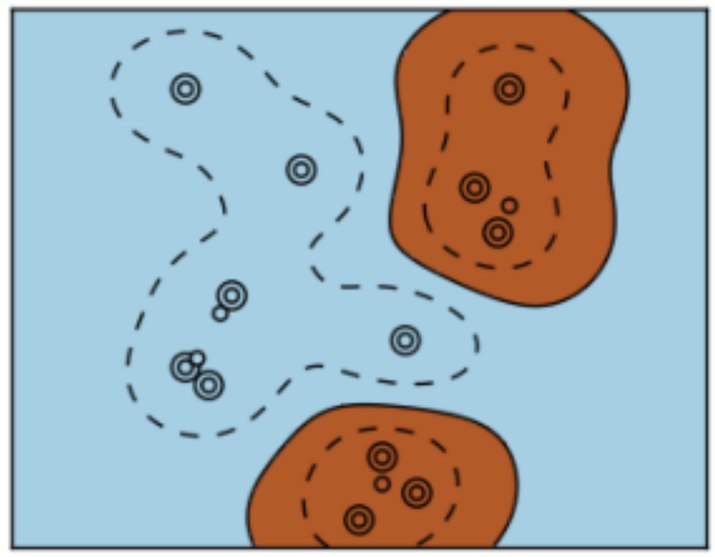
SVM - finer details (optional)



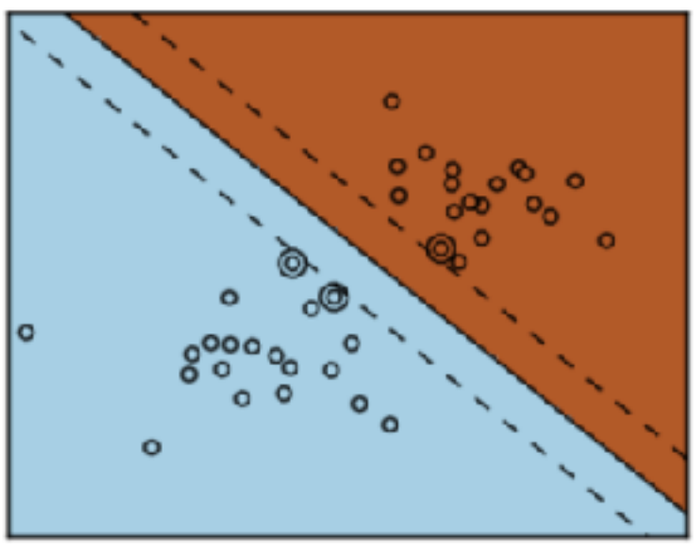
Linear

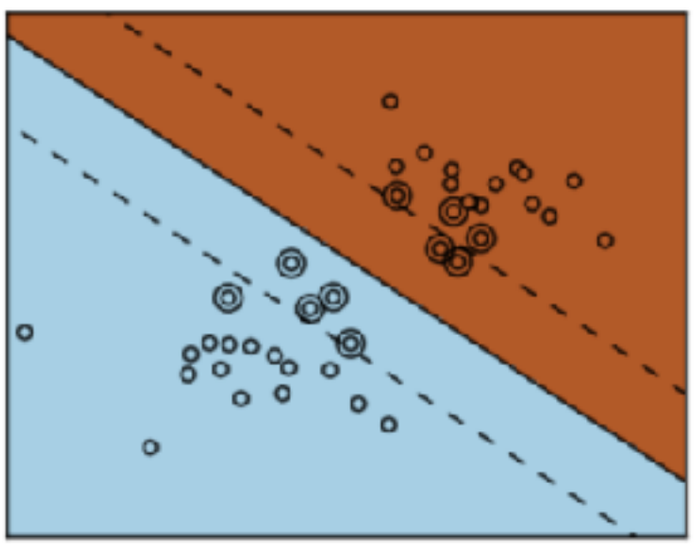


Polynomial



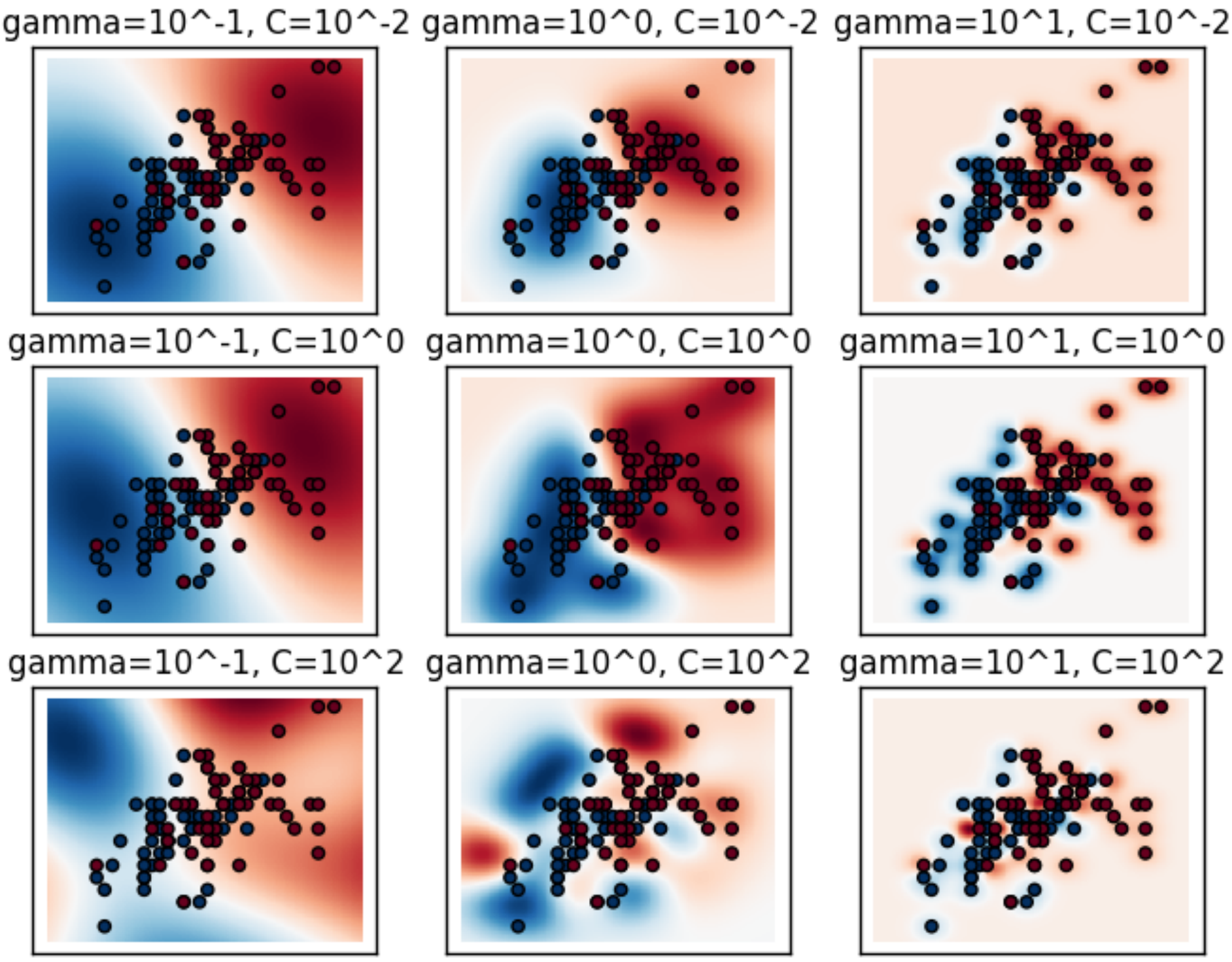
Gaussian/radial basis





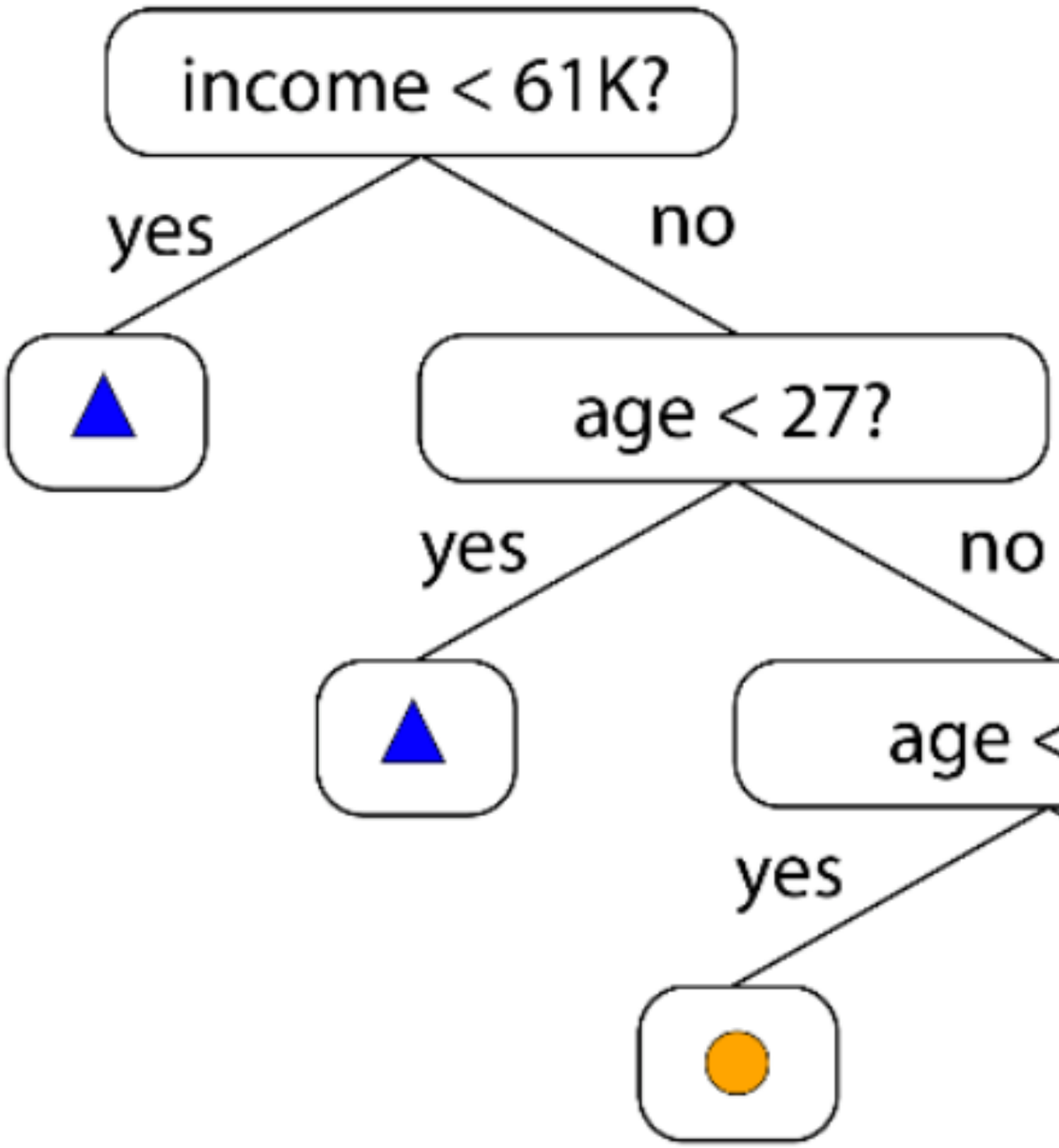
Varying “cost”, *C*

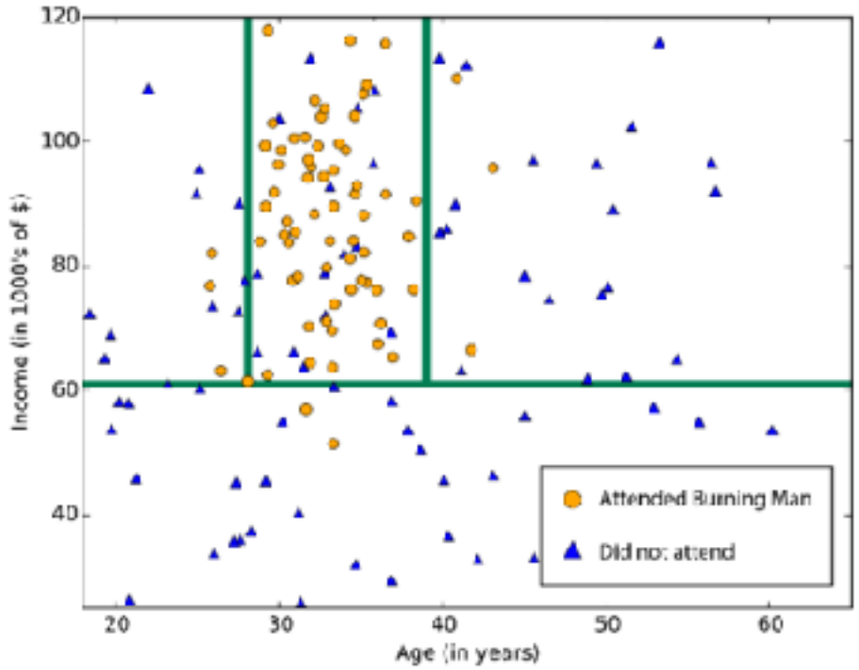
SVM - finer details (optional)

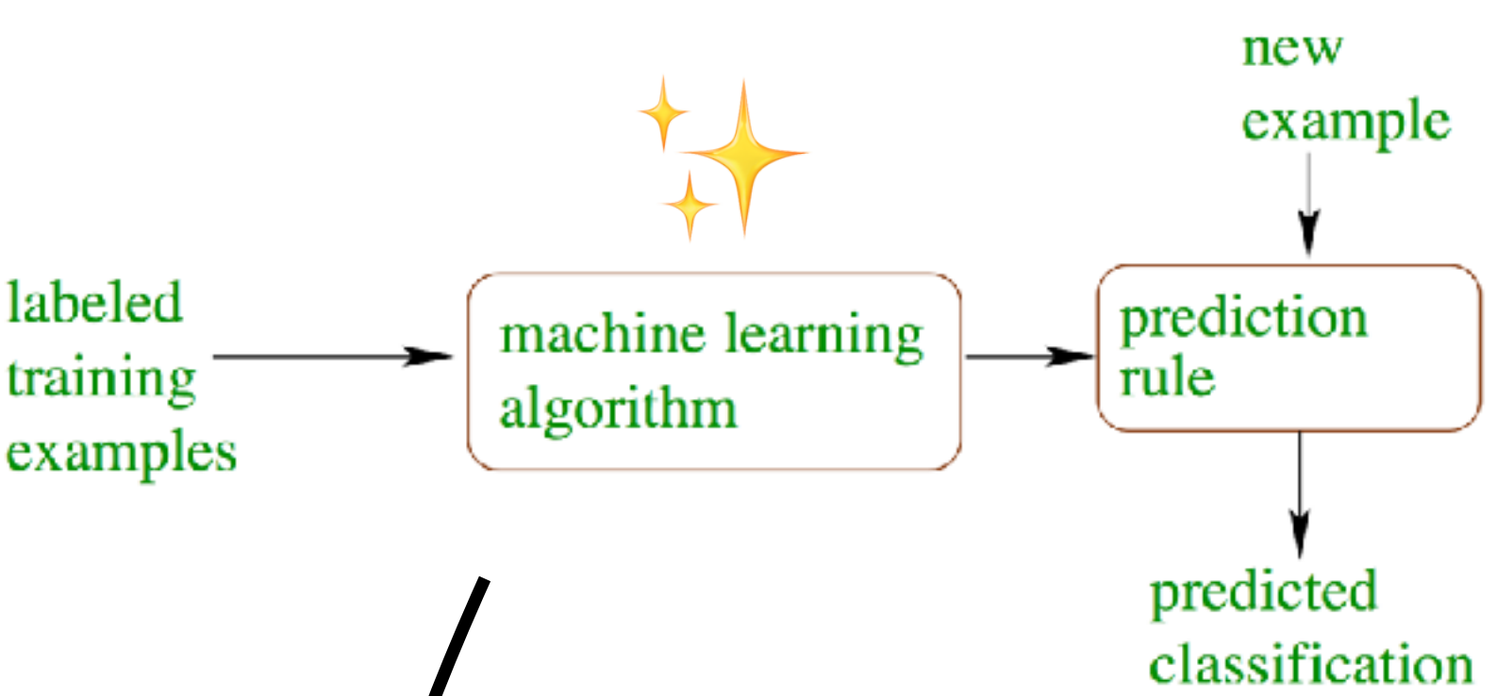


—-> See scikit learn SVM documentation for more information

Decision trees







1. Feature selection (what info?)
2. Algorithm selection (how do we combine it?)
3. Model validation (does this “work”?)

Testing, testing, testing!

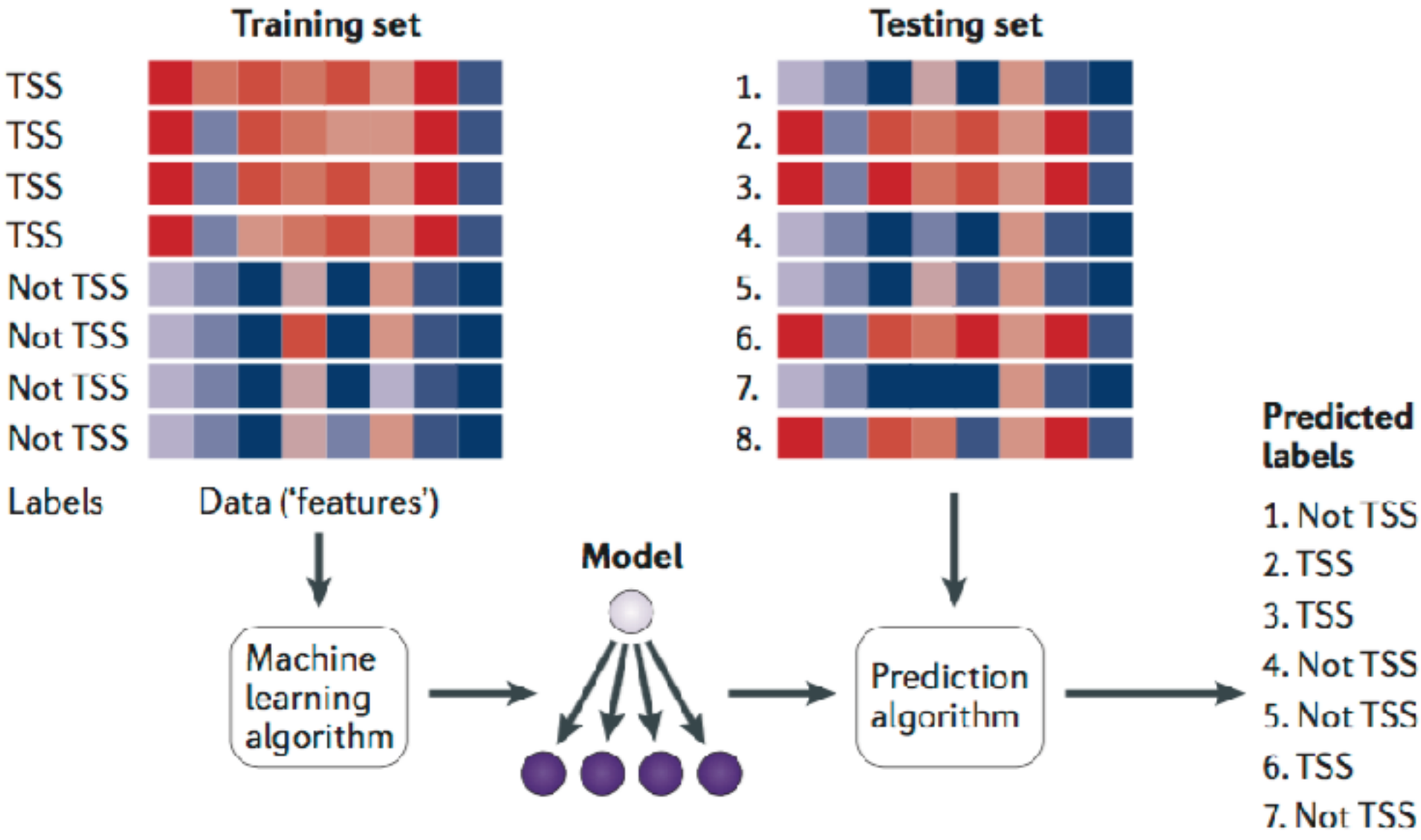
Model Performance?

- How do we quantify performance?
  - Choose a key performance metric
    - Accuracy, PPV; R^2, Mean absolute error?

Model Validation?

- Is this model a good model of reality?
  - How will our model do on future data?
    - test-train split; leave-one-out; split half, k-fold

How good is my model?



1 - Get the predictions from our model

2 - Compare them to the truth ——>

$$\text{accuracy} = \frac{\#\text{ correct predictions}}{\#\text{ total data points}}$$

How good is my model?

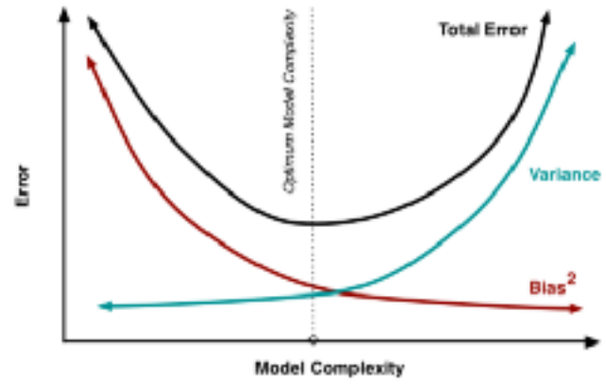
What if outcomes are not equally likely?

- If only 1% of people suicide, guessing “no suicide” for everyone will give 99% accuracy!

Confusion matrix

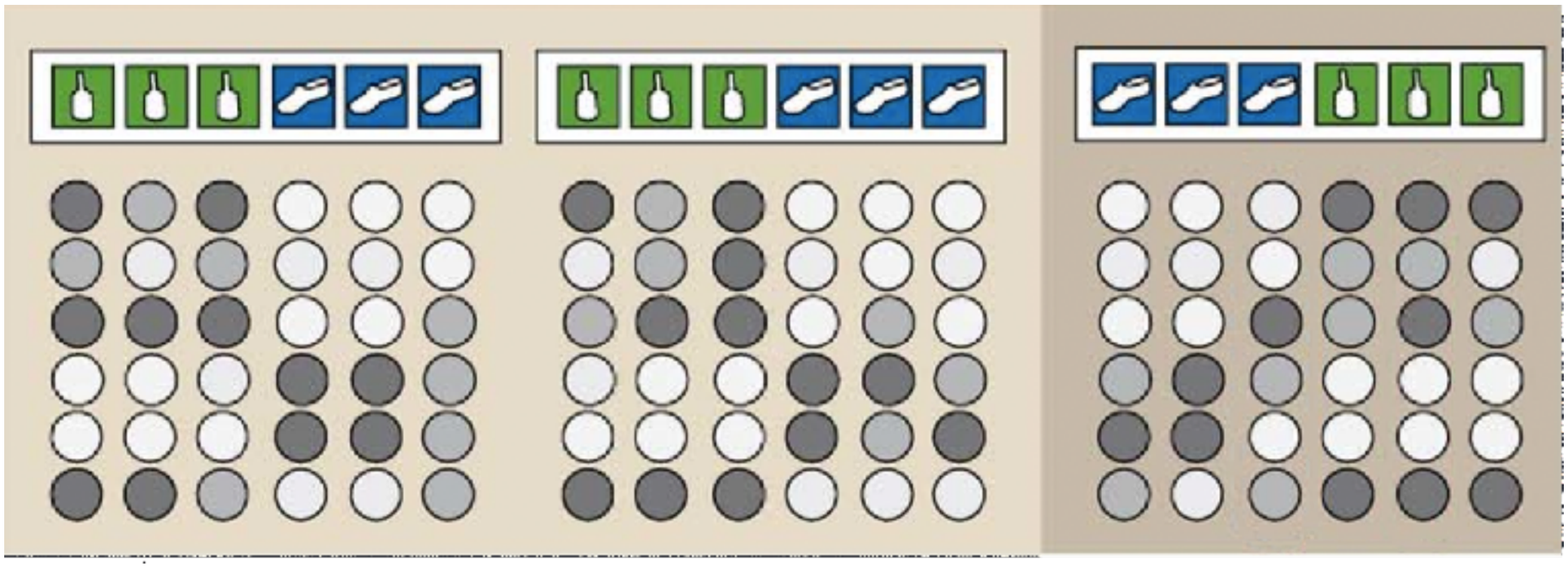
		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	**Specificity** = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

Internal Validation



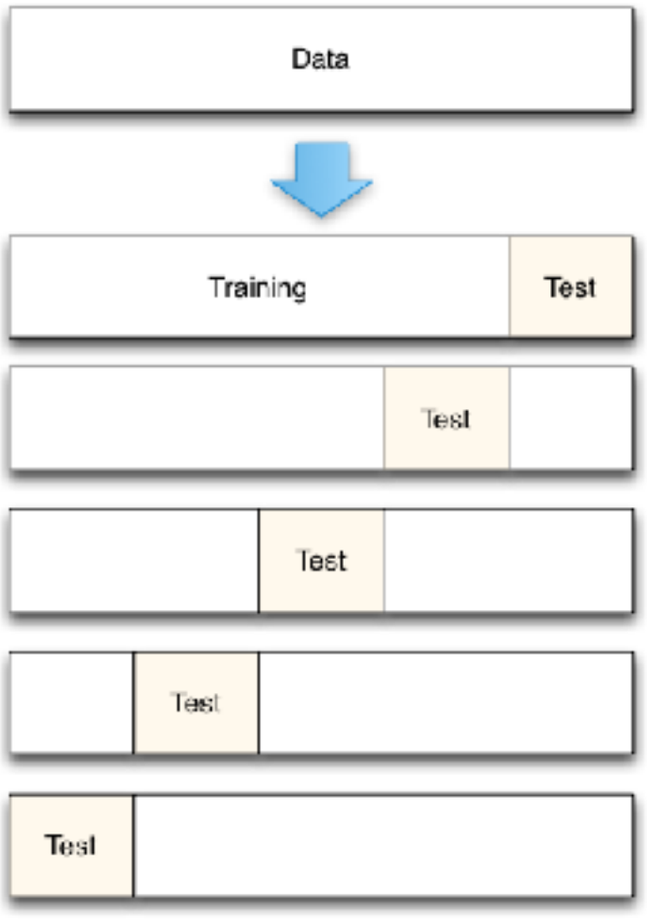
- Two motivations for internal validation:
    1. Estimate variance of modeling approach
    2. No point testing model on things it has seen before!

Easiest solution: split data into training and testing

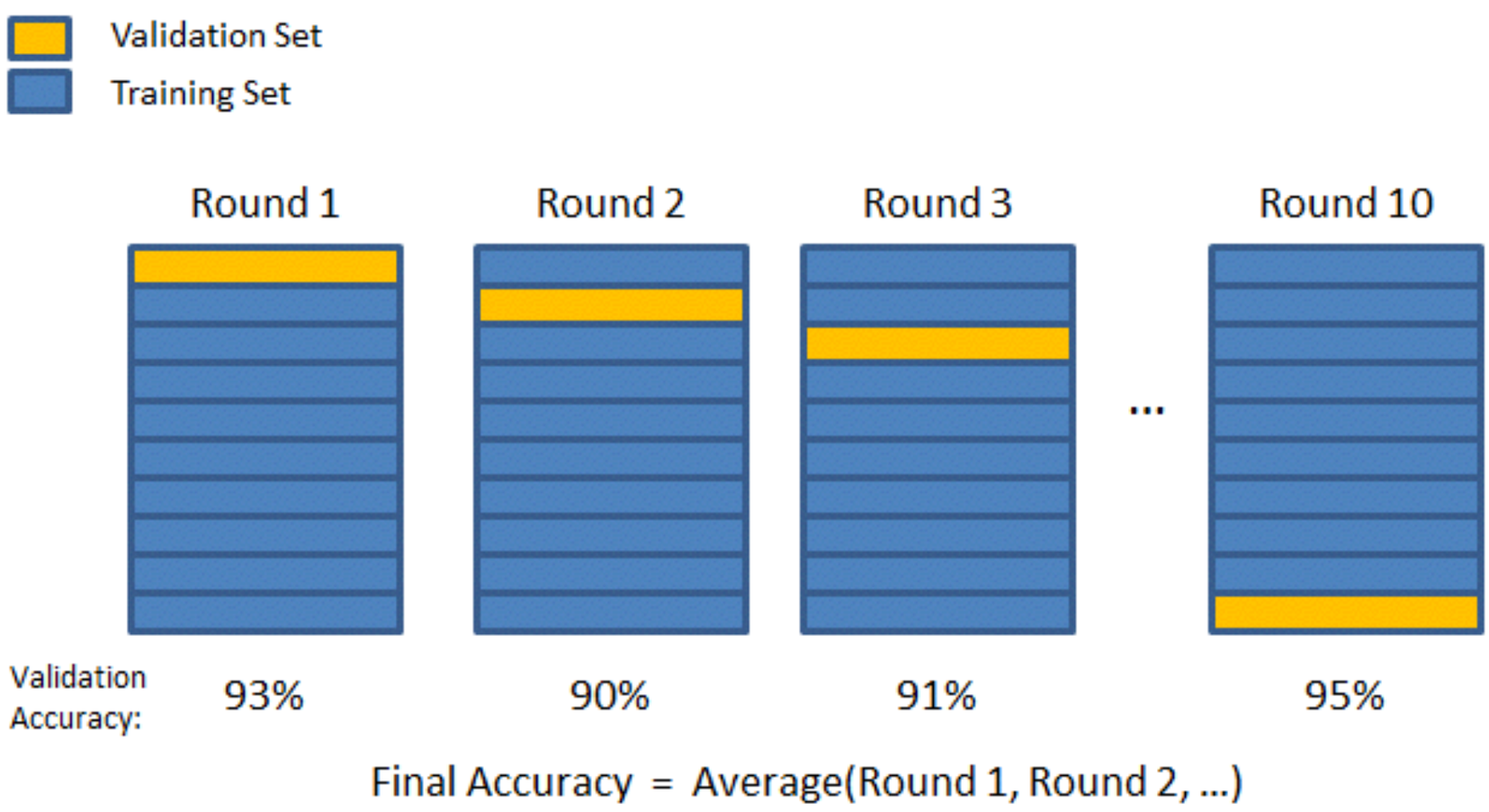


Run 1 Run 2 Run 3

Better CV



5-fold CV



10-fold CV

Is my model *real* ?

- Internal validation allows us to know how variable our model is in our current data, and try not to overfit
  - But how well will our model perform *in general*?
- Examine performance on totally new data
  - If good, gives us confidence that the model is picking up on “real” signal
    - Better candidate for the truth!

Best practice: prospective validation

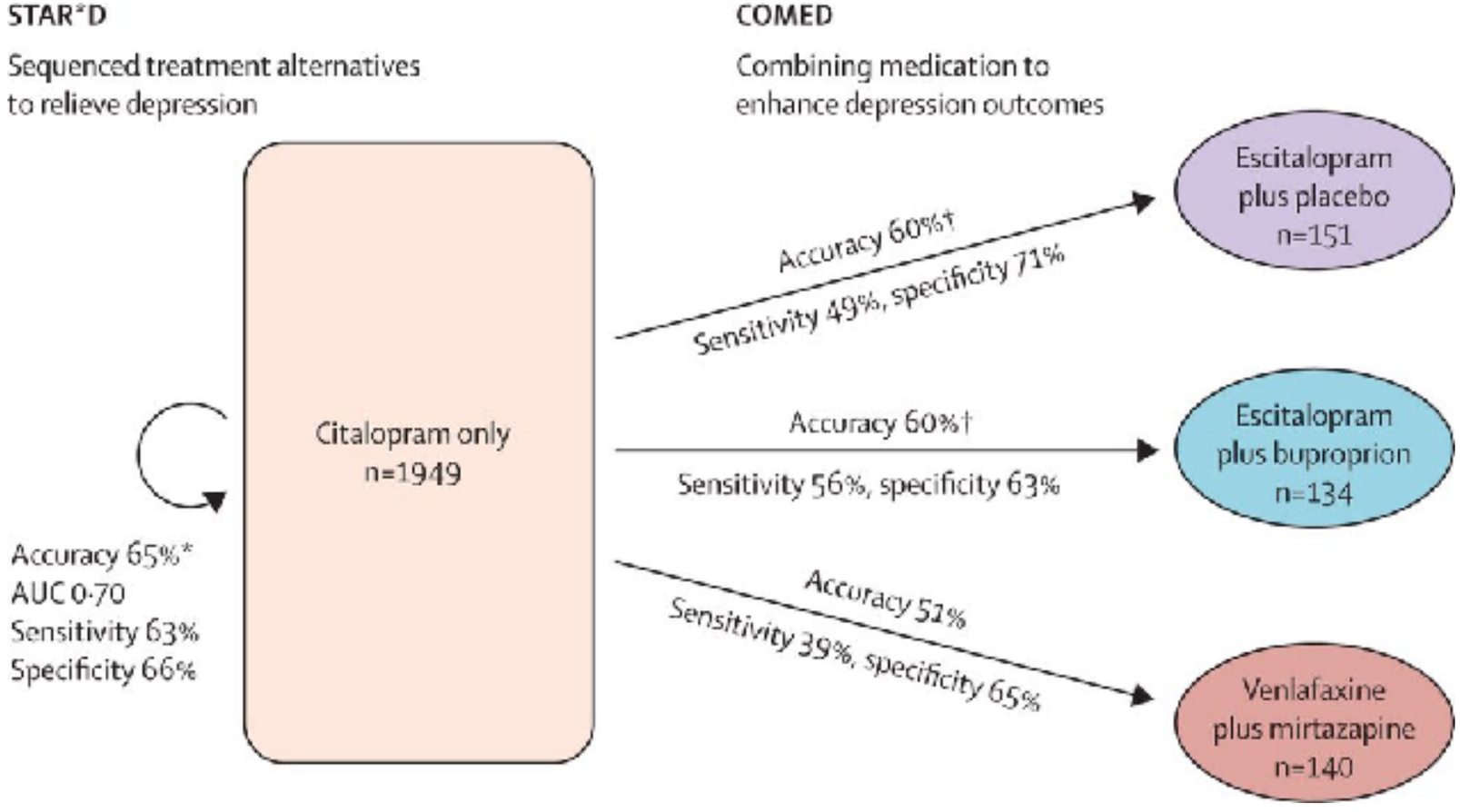
Is my model *real* ?

Articles

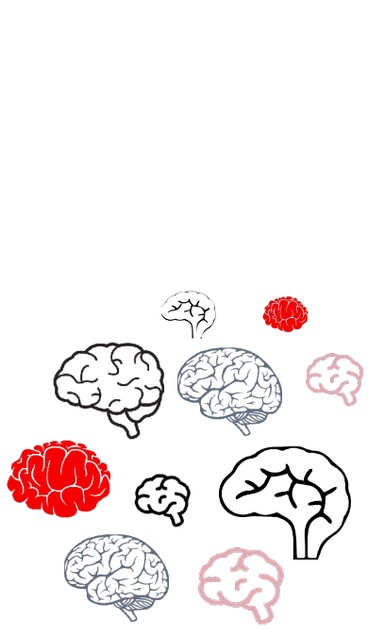
Cross-trial prediction of treatment outcome in depression: a machine learning approach

Adam Mourad Chekroud, Ryan Joseph Zotti, Zarrar Shehzad, Ralitza Gueorguieva, Marcia K Johnson, Madhukar H Trivedi, Tyrone D Cannon, John Harrison Krystal, Philip Robert Corlett

- Trained in one trial, tested prospectively in another trial
- Model performance was weak (60%)
- Some (significant) models did not generalize!



Summary



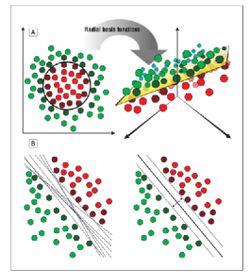
labeled
training
examples

machine learning
algorithm

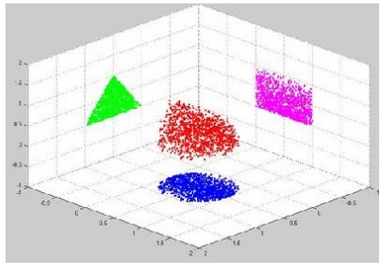
new
example

prediction
rule

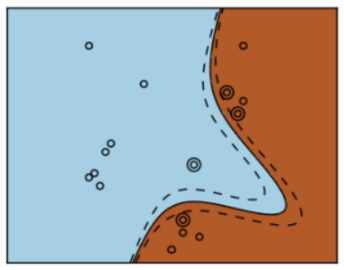
predicted
classification



1



2



3