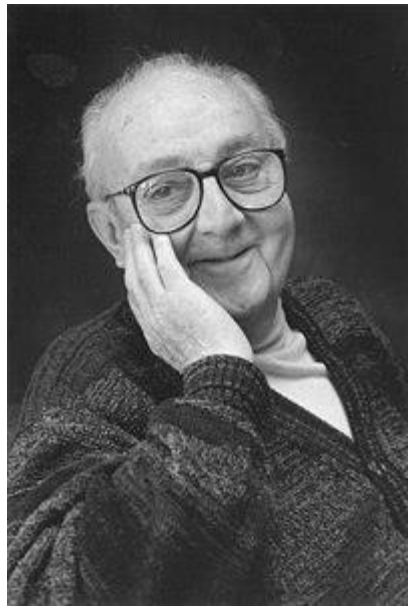
Introduction to Machine Learning

Dr. Rui Borges
Institute of Population Genetics

Introductory Course for Ph.D. students of the
Doctoral Program in Population Genetics

Population genetics: models and data



“all models are wrong, but some are useful”

George Box

“all models are wrong, but **many** are usefull”

Population genetics: mechanistic inference

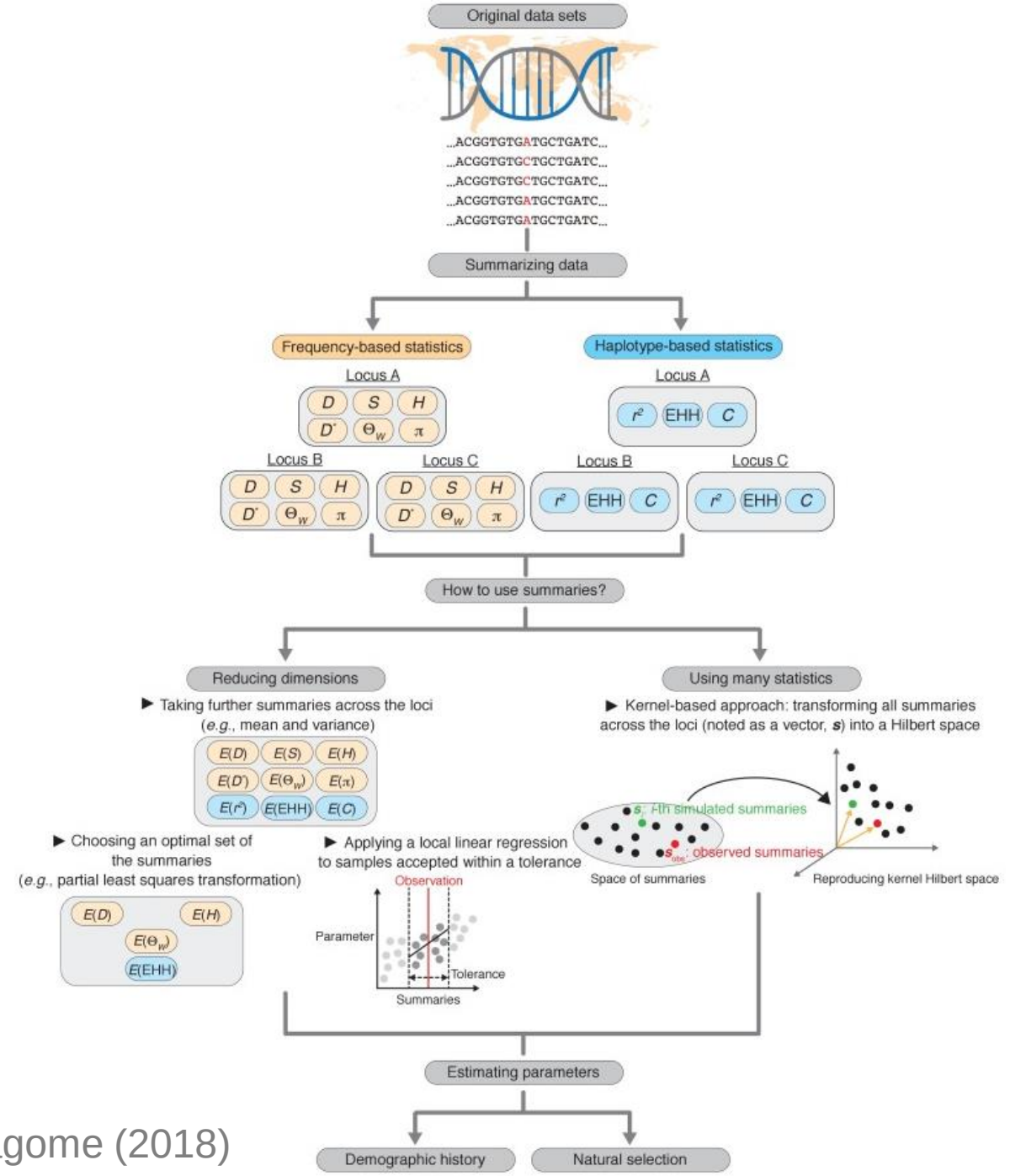
- model-based statistics has been challenged: next-generation sequencing technologies (Lavy and Meyers, 2016)
- maximum likelihood and Bayesian methods
- expectation maximization and MCMC
- cost of calculating the likelihood function

ABC: approximate Bayesian computation

- approximate the posterior distribution without the calculation of the likelihood function (Beaumont et al. 2002)
- summary statistics: capture information present in raw data
- Which summary statistics?

ABC

- widespread in population genetics Lopes and Beaumont (2010)
- curse of dimentionality: summary statistcs



Cooke and Nakagome (2018)

ABC: exercise

- Imagine that you want to fit a line $y = mx + b$ to n observed cordinates (x_i, y_i). What summary statistic(s) could you potentially use to estimate m and b using ABC?
- Write the pseudocode.

Population genetics: a data-driven field

- Population genetics has been transitioning from a theory-driven into a data-driven field
- vast amount of genomes and metadata
- Human population genomics: high-quality whole-genome sequencing from more than 150 000 individuals from the UK biobank (Halldorsson et al. 2022)

Machine learning

- General-purpose algorithms that can learn patterns present in complex and large data sets

Unsupervised learning
uncovering structure within a dataset without prior knowledge of how the data are organized

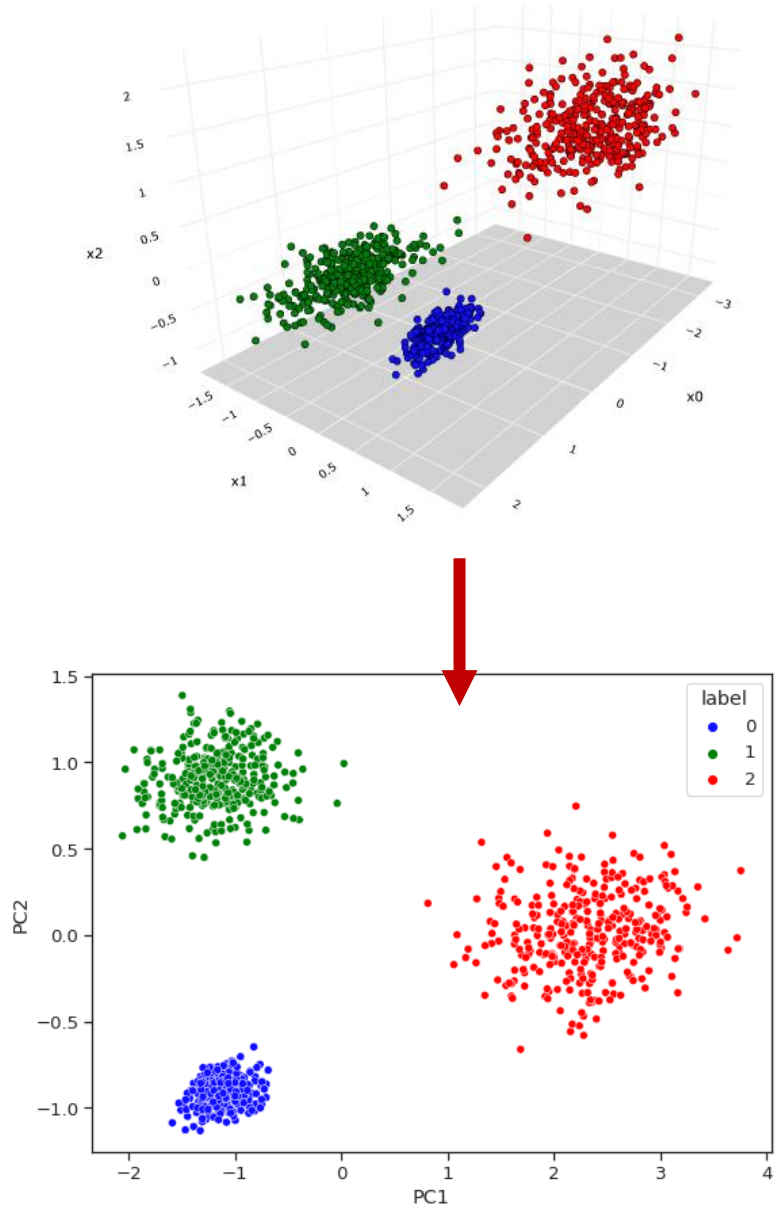
Clustering

Supervised learning
relies on prior knowledge about an example dataset to make predictions about new datapoints

Classification or Regression*

Unsupervised learning: PCA

- e.g., principal component analysis (PCA)
- PCA is a statistical technique for reducing the dimensionality of a dataset
- used to visually identify clusters of closely related data points

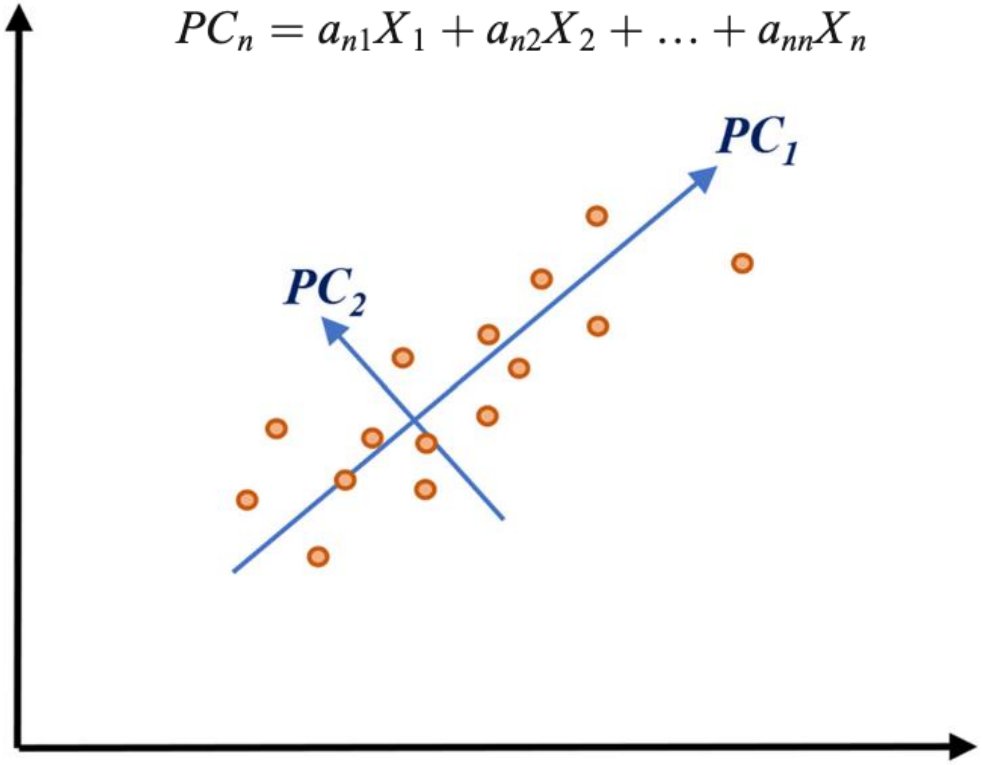


PCA

- linearly transforms the data into a new coordinate system where most of the variation in the data can be described with fewer dimensions
- many studies use the first two principal components

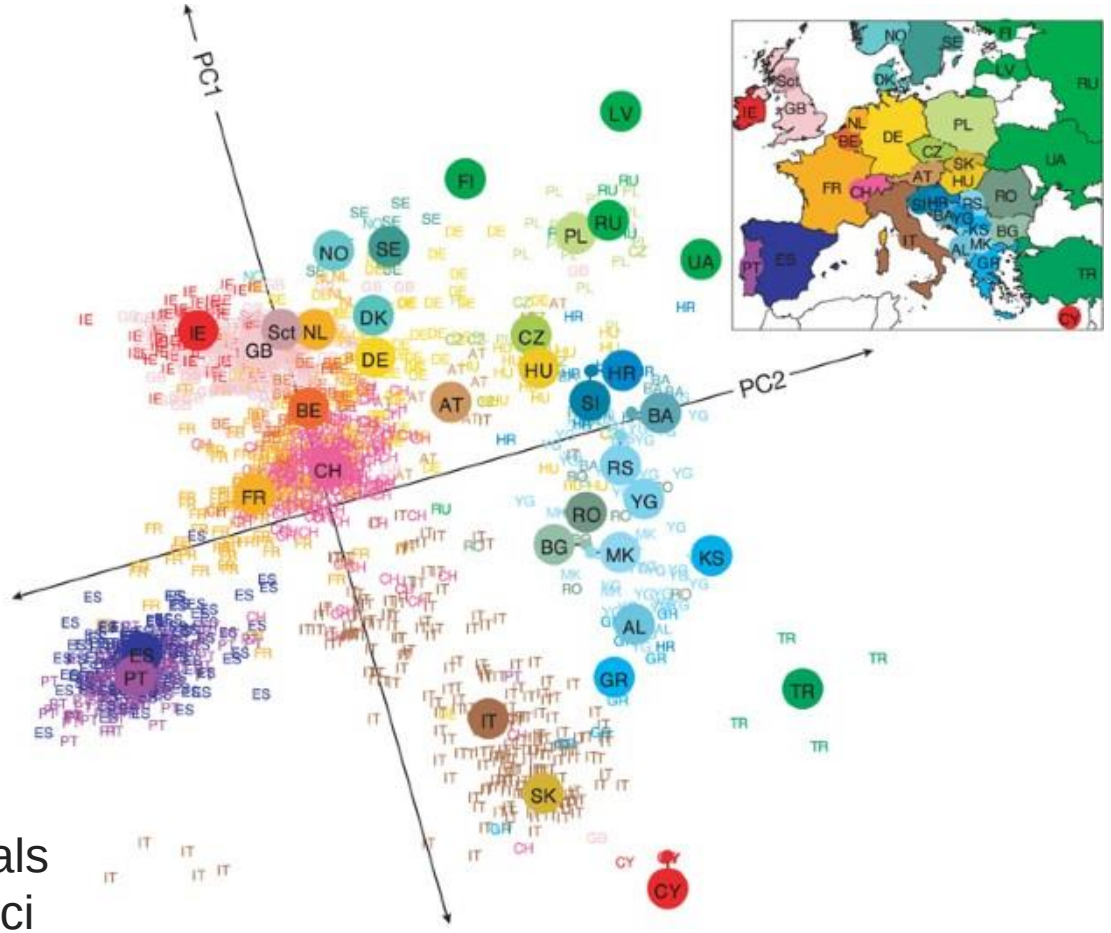
New coordinate system:

$$PC_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1n}X_n$$
$$PC_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2n}X_n$$
$$\vdots$$
$$PC_n = a_{n1}X_1 + a_{n2}X_2 + \ldots + a_{nn}X_n$$



PCA in population genetics

- PCA can be used for discovering unknown relatedness relationships among individuals



3 192 European individuals genotyped at 500 568 loci

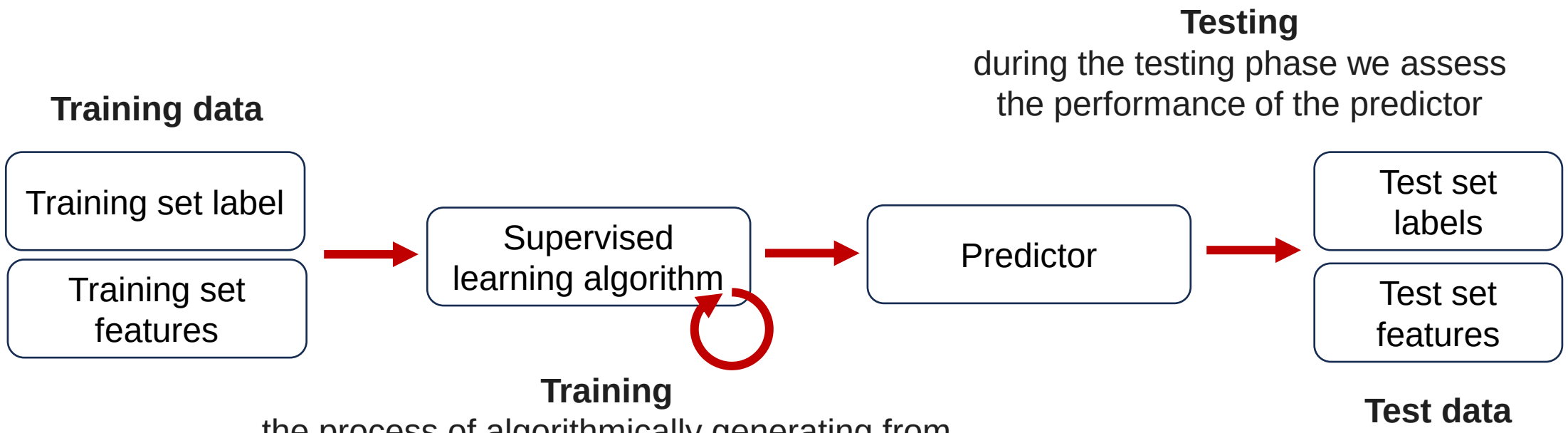
Novembre et al. 2008

PCA: exercise

- Learning lower dimensional representation can save memory usage
- Learning lower dimensional representation can remove redundancies and noises in data
- When we use PCA, we need data to be labelled
- PCA extracts the variance structure from high dimensional data such that the variance of projected data is minimized
- Different individual principal components are linearly uncorrelated
- The dimension of original data representation is always higher than the dimension of transformed representation of PCA

Supervised learning

The general framework:



Supervised learning

How to build a good predictor?

- **loss function:** a measure of how correctly the response variable was predicted
- minimize the value of the risk function during training

Task	**Error type**	**Function**
Regression	Mean-squared error	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
	Mean absolute error	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Classification	Cross entropy	$-\frac{1}{n}\sum_{i=1}^{n}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)]$

Some important concepts

How do we test the predictor?

- Confusion matrix and accuracy for classification tasks
- Mean-squared error or the coefficient of determination for regression tasks

Accuracy:

$$\frac{\sum_{i\in[n]} c_{ii}}{\sum_{i,j\in[n]} c_{ij}}$$

Confusion matrix →

Predicted \ Expected	1	2	3	4
1	52	3	7	2
2	2	28	2	0
3	5	2	25	12
4	1	1	9	40

Supervised learning: exercise

- Two predictors were trained for a classification problem. The accuracy of these predictors on the training and testing set are shown in the following tables.

Predictor A	Accuracy
Training data	0.89
Test data	0.82

Predictor B	Accuracy
Training data	0.99
Test data	0.78

- Which predictor would you choose? Why?

Supervises machine learning algorithms

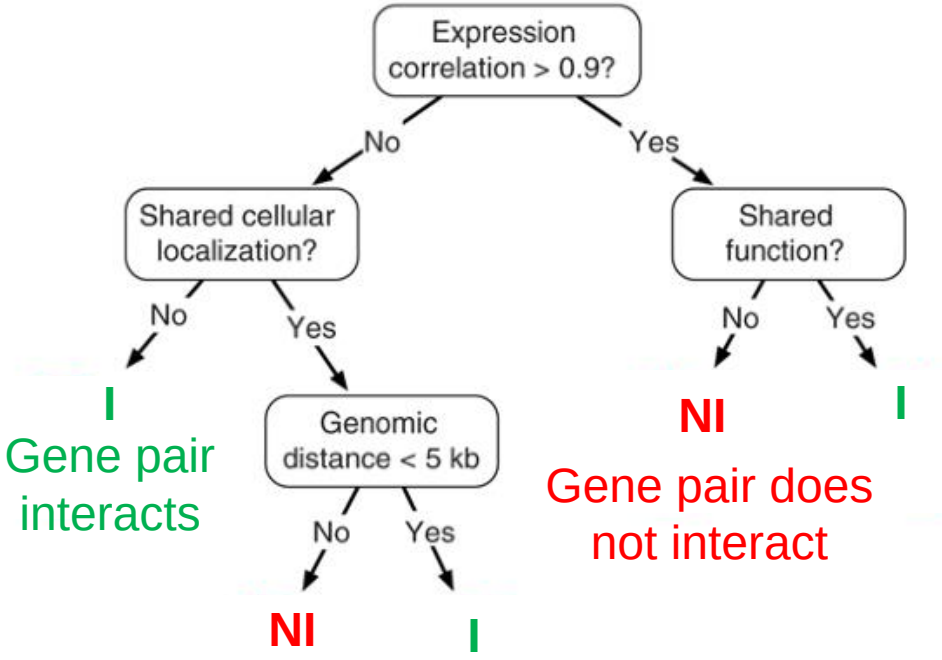
- Rich history in population genetics

1. Decision trees
2. Random forests
3. Boosting
4. Support vector machines
5. Deep learning

Decision trees (DT)

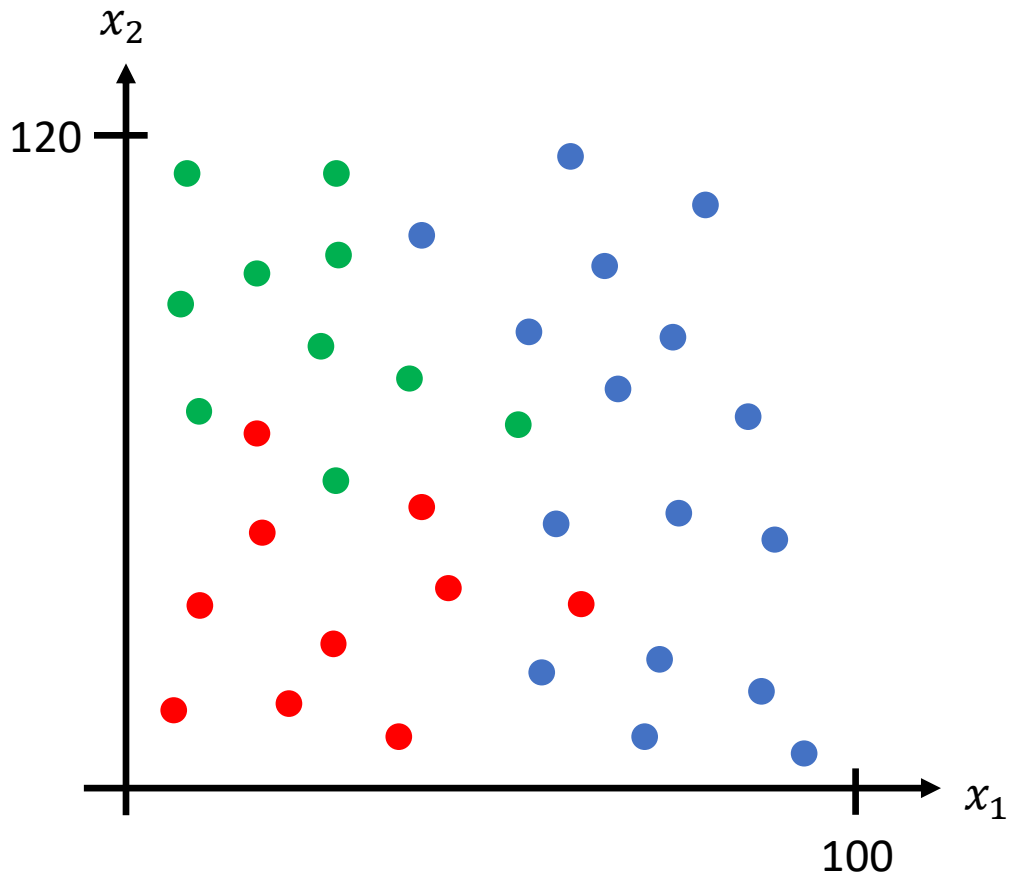
- a hierarchical structure that predicts the response variable of an example by if-else statements on features
- at the next level of the tree another feature is examined
- the predicted value is determined by which leaf of the tree is reached at the end of this process

Gene Pair	Interact?	Expression correlation	Shared localization?	Shared function?	Genomic distance
A-B	Yes	0.77	Yes	No	1 kb
A-C	Yes	0.91	Yes	Yes	10 kb
C-D	No	0.1	No	No	1 Mb
⋮					



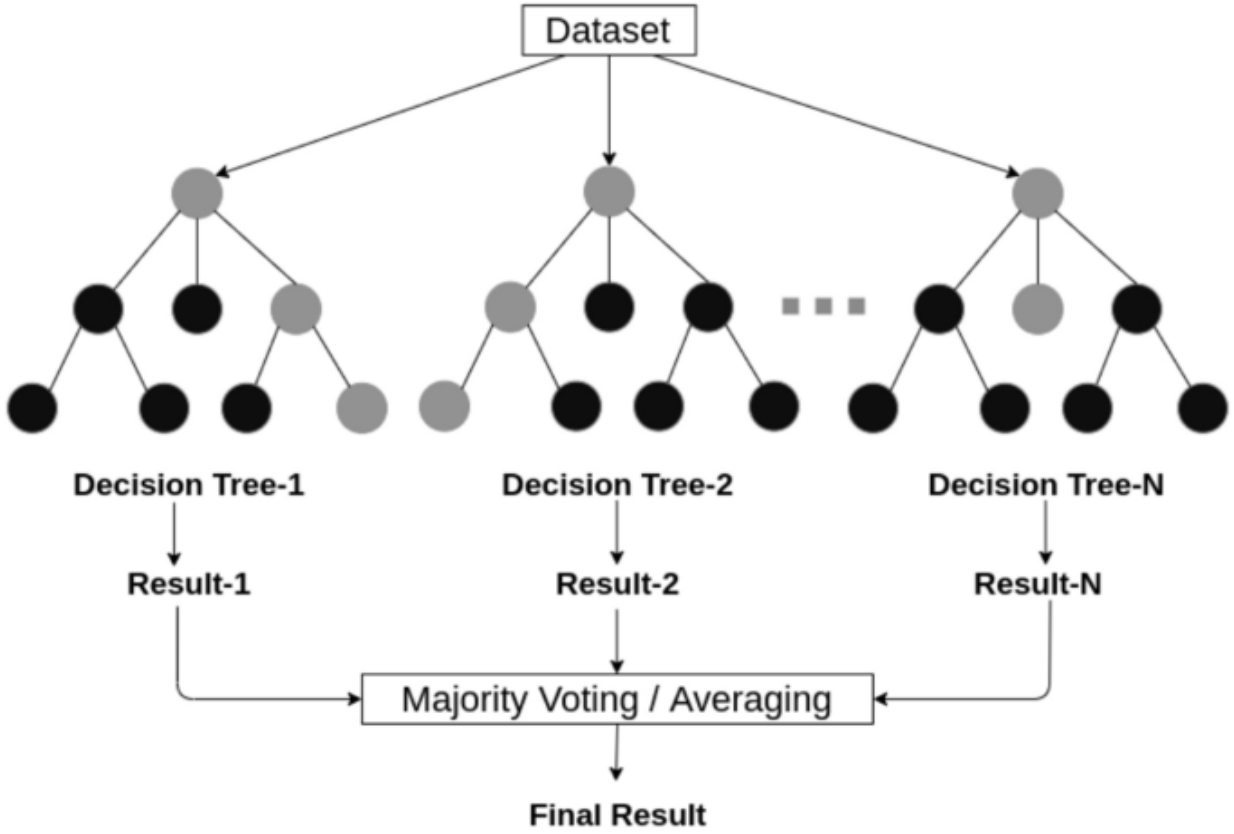
DT: exercise

Build a decision tree for this data:



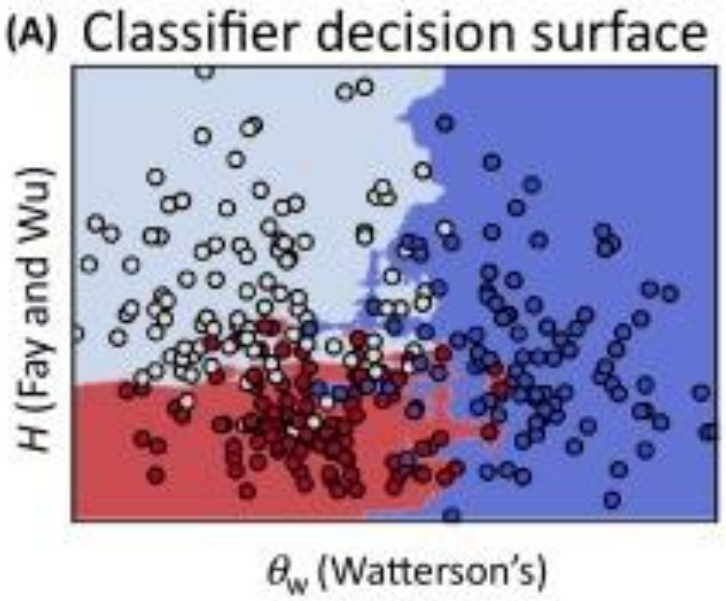
Random forests (RF)

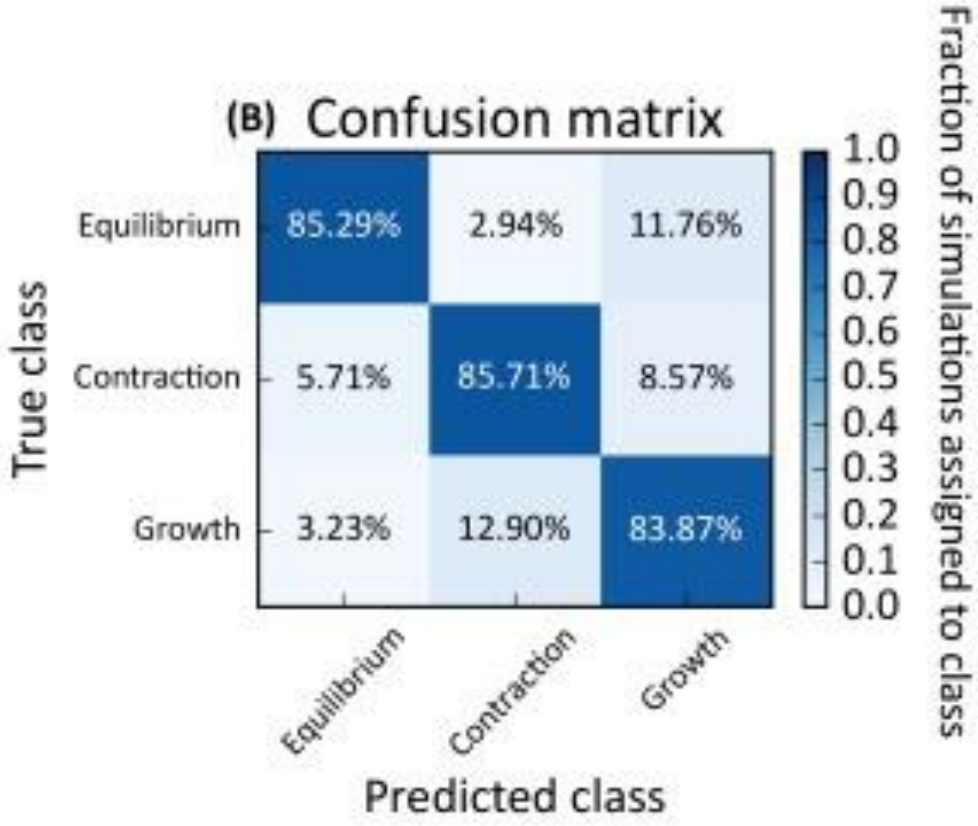
- ensemble of semi-randomly generated decision trees
- runs through each tree in the forest, and these trees then vote to determine the predicted value
- random forests can perform both classification and regression



Random forests in population genetics

- RF algorithm to classify demographic scenarios of population contraction, expansion or constant size

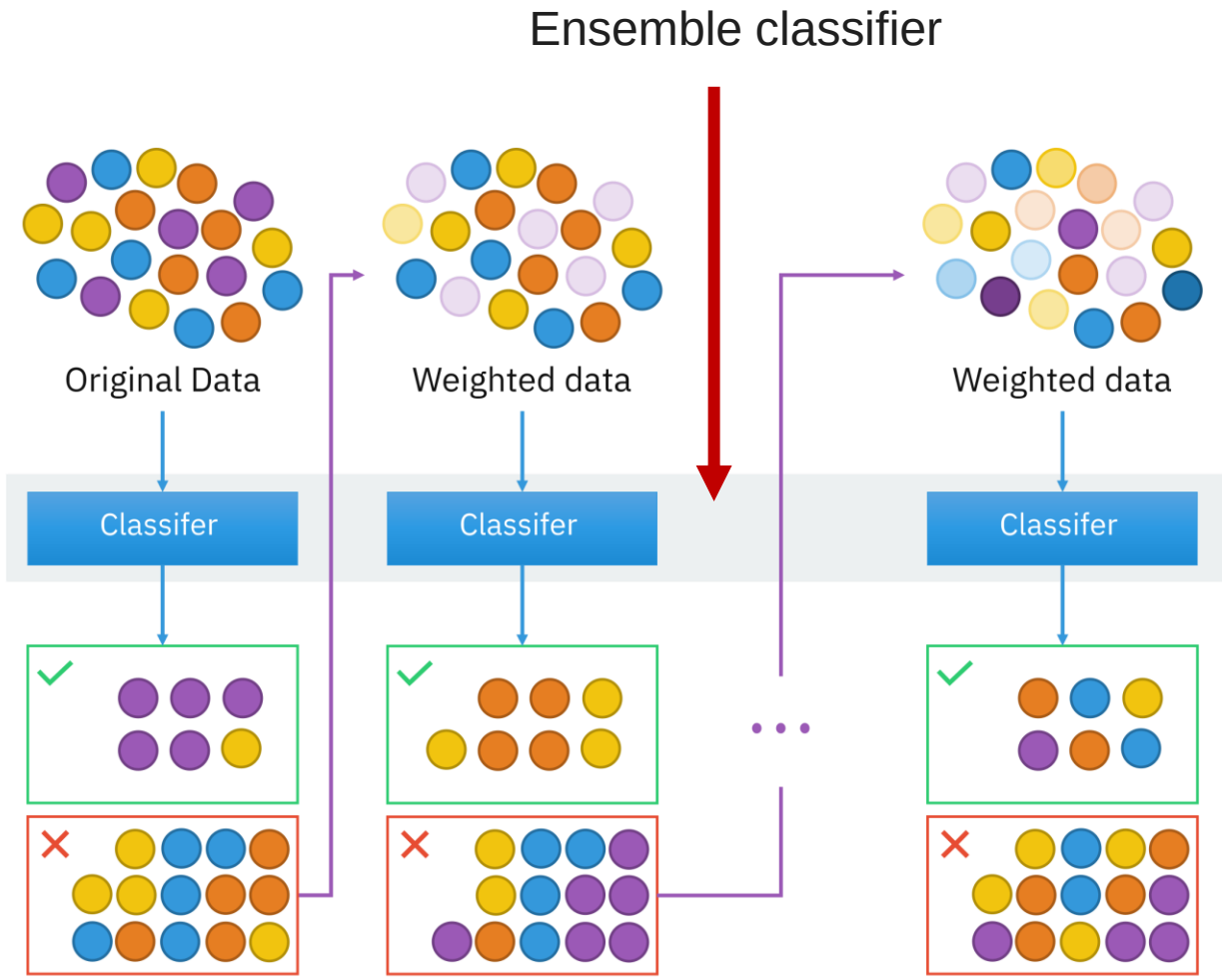




Schrider and Kern (2018)

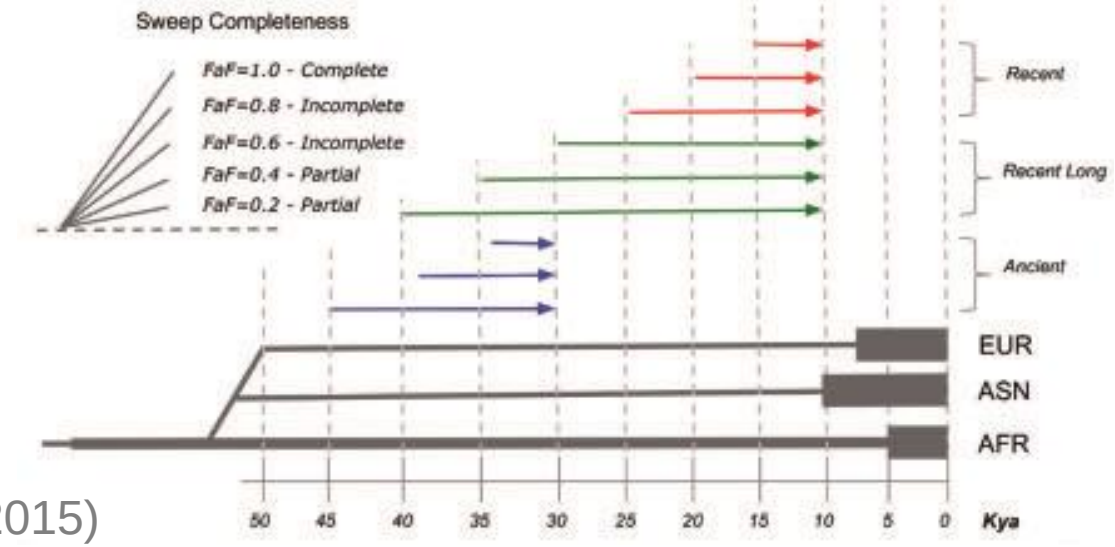
Boosting

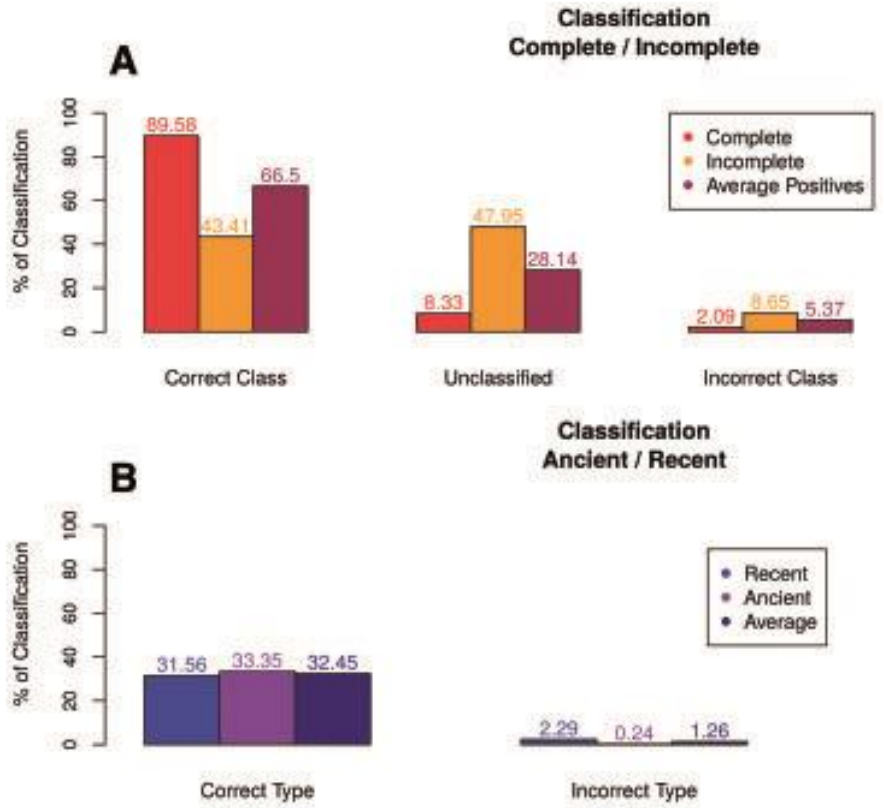
- a class of techniques that iteratively constructs a set of predictors
- the new predictor to be added focuses on samples the current set of predictors has struggled with



Boosting in population genetics

- Hierarchical boosting method to classify the time and completeness of sweeps in human populations

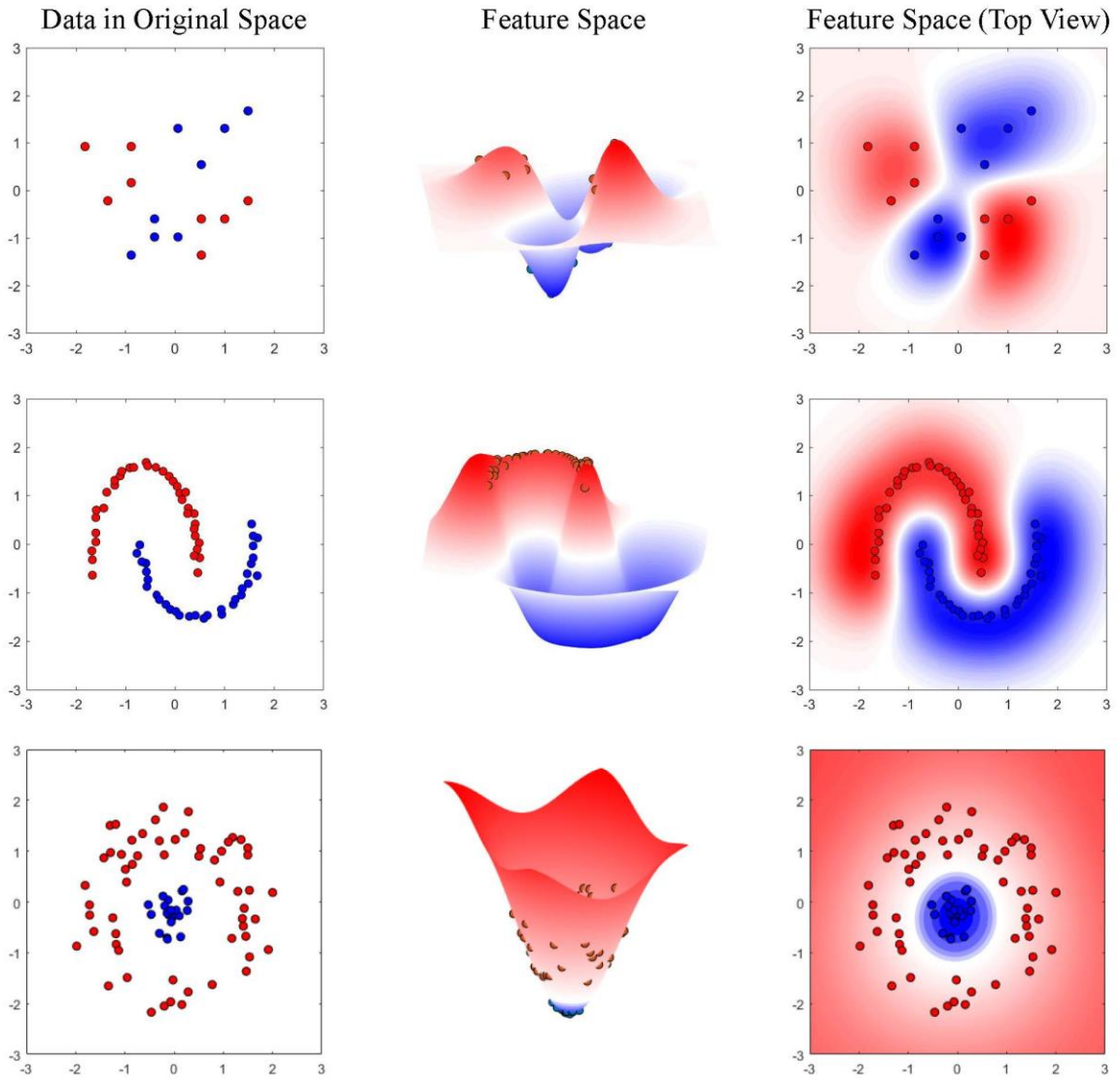




Pybus et al. (2015)

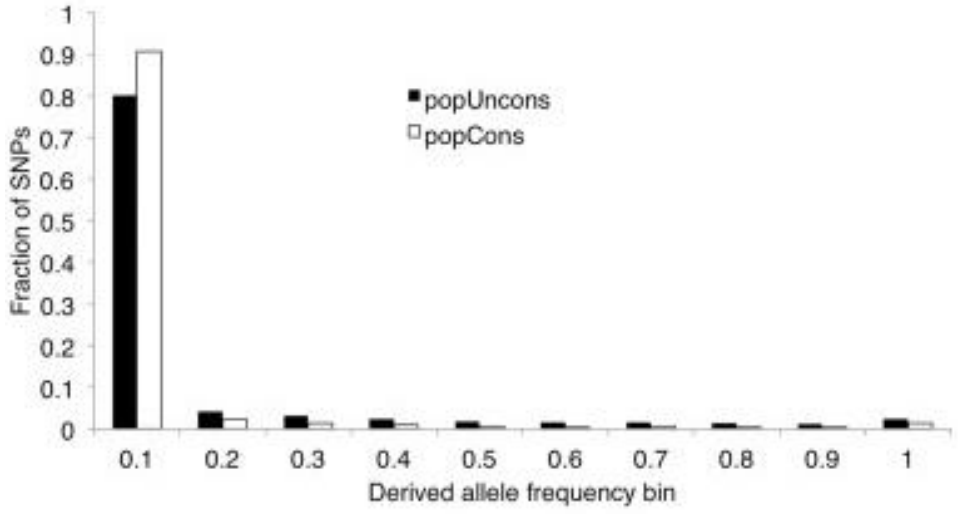
Support vector machines (SVM)

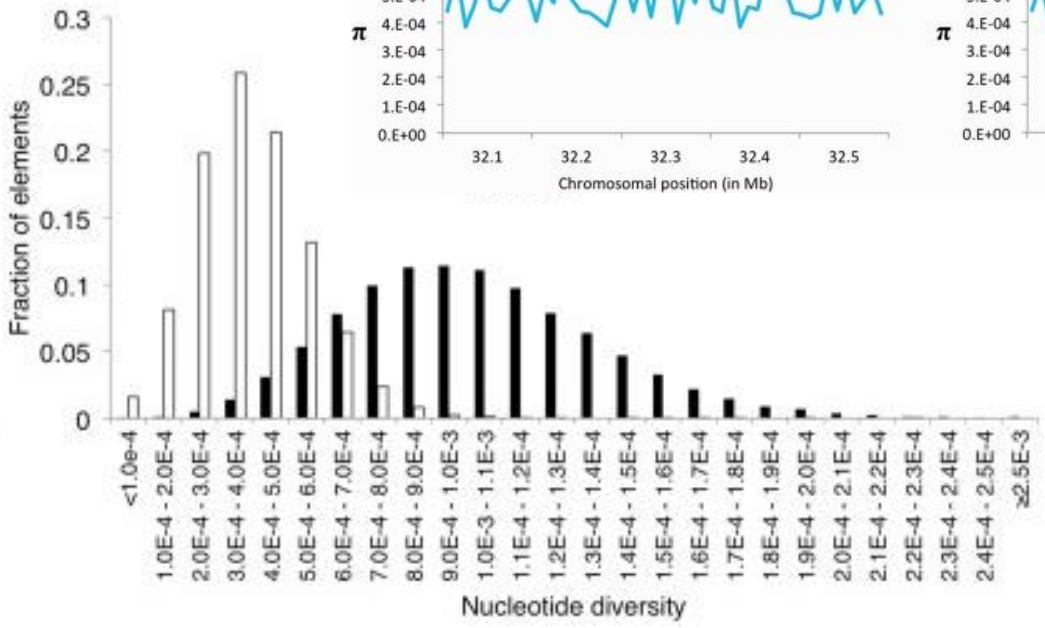
- seek to find the hyperplane that optimally separates two classes of training data
- data are often mapped to high-dimensional space using a kernel function
- accomplish multiclass classification or regression

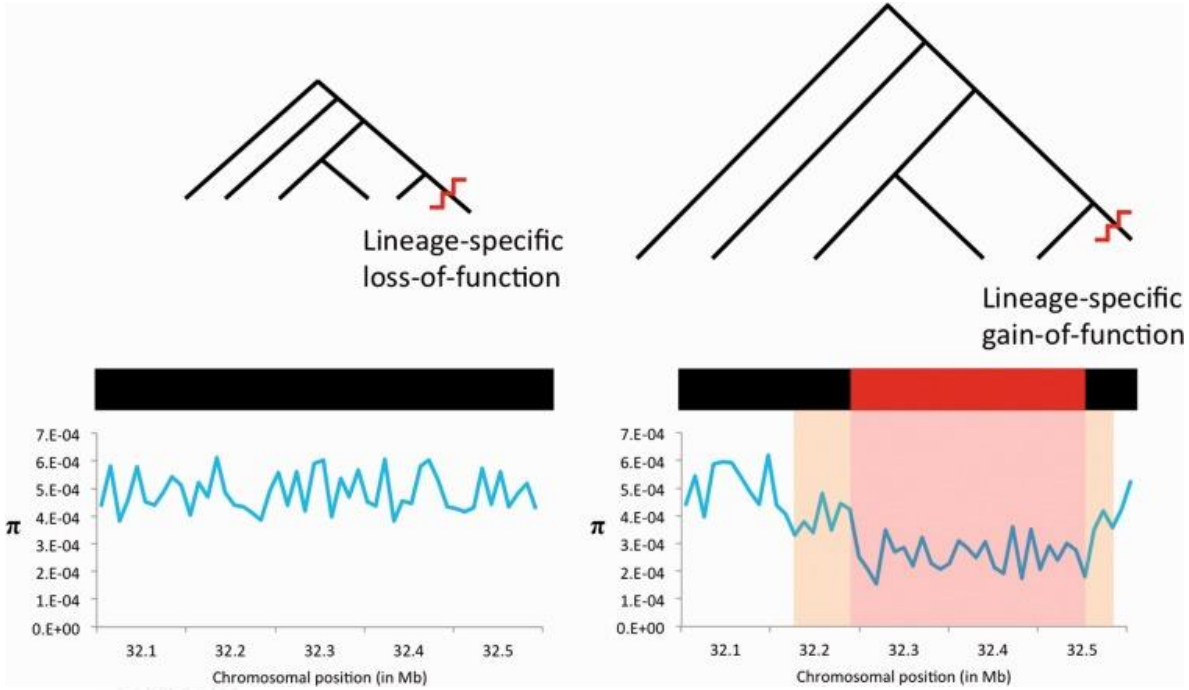


SVM in population genetics

- SVM to classify 10 kb genomic windows as either constrained or unconstrained



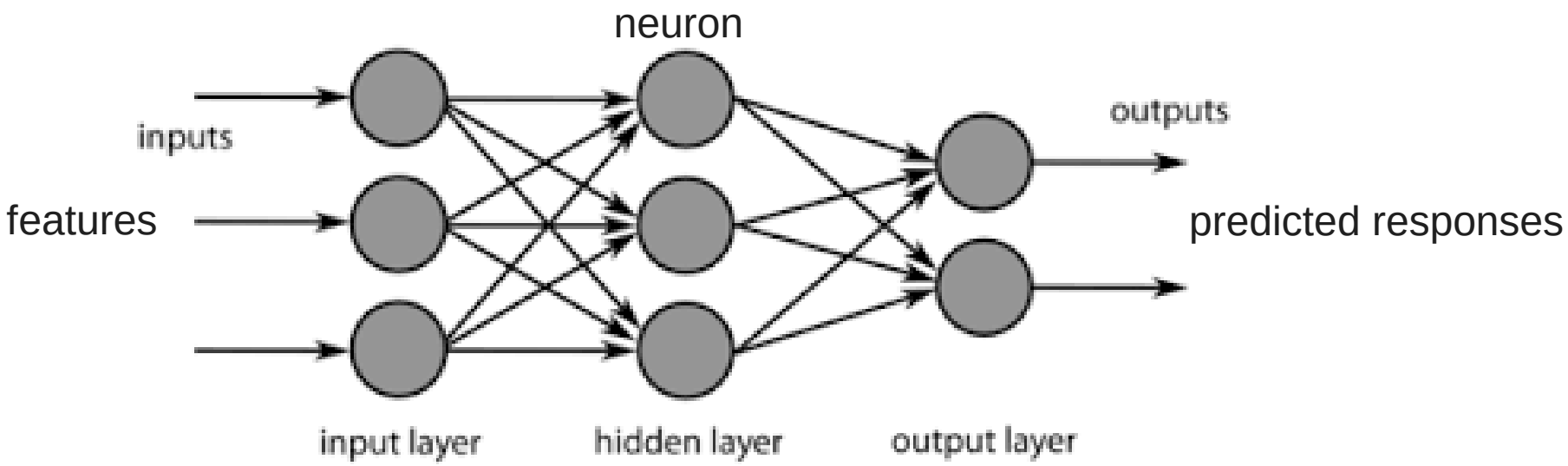




Schrider and Kern (2018)

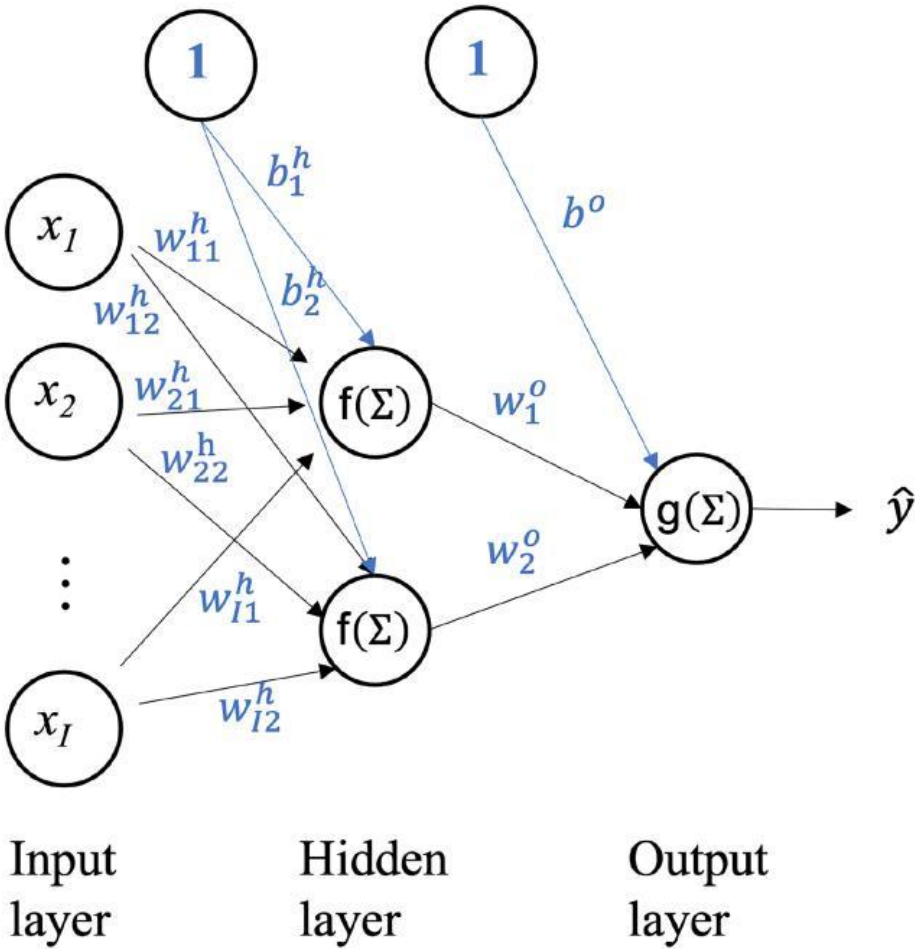
Deep learning

- learning using networked algorithmic models that contain multiple hidden layers between the input and output layers



Artificial neural networks (ANNs)

- a network of layers of one or more neurons
- receive weighted inputs from each neuron in the previous layer
- perform a linear combination on these inputs which is then passed through an activation function

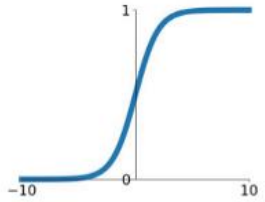


ANNs: activation functions

- **activation function**: a function that calculates the output of the node

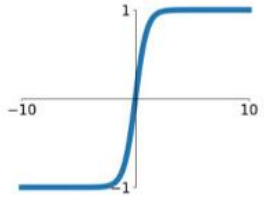
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



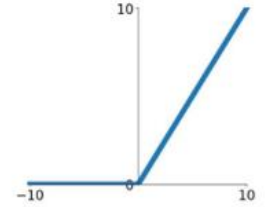
tanh

$$\tanh(x)$$



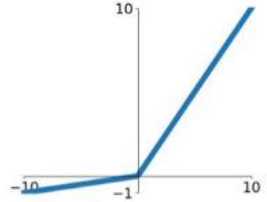
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

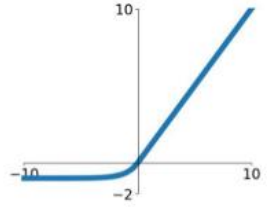


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

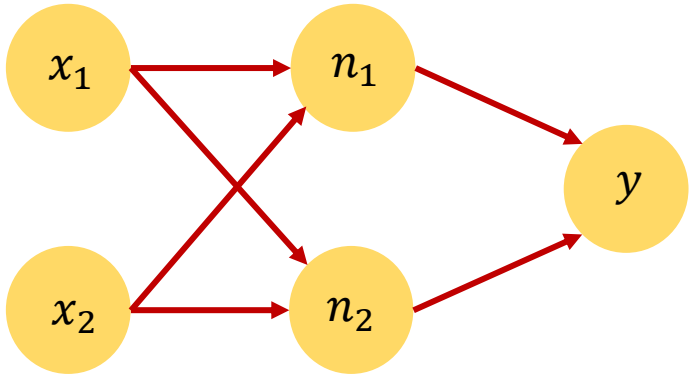
ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



ANNs: exercise

- Write down the formula of y based on this artificial neural network:



- Further assume the activation function is linear.

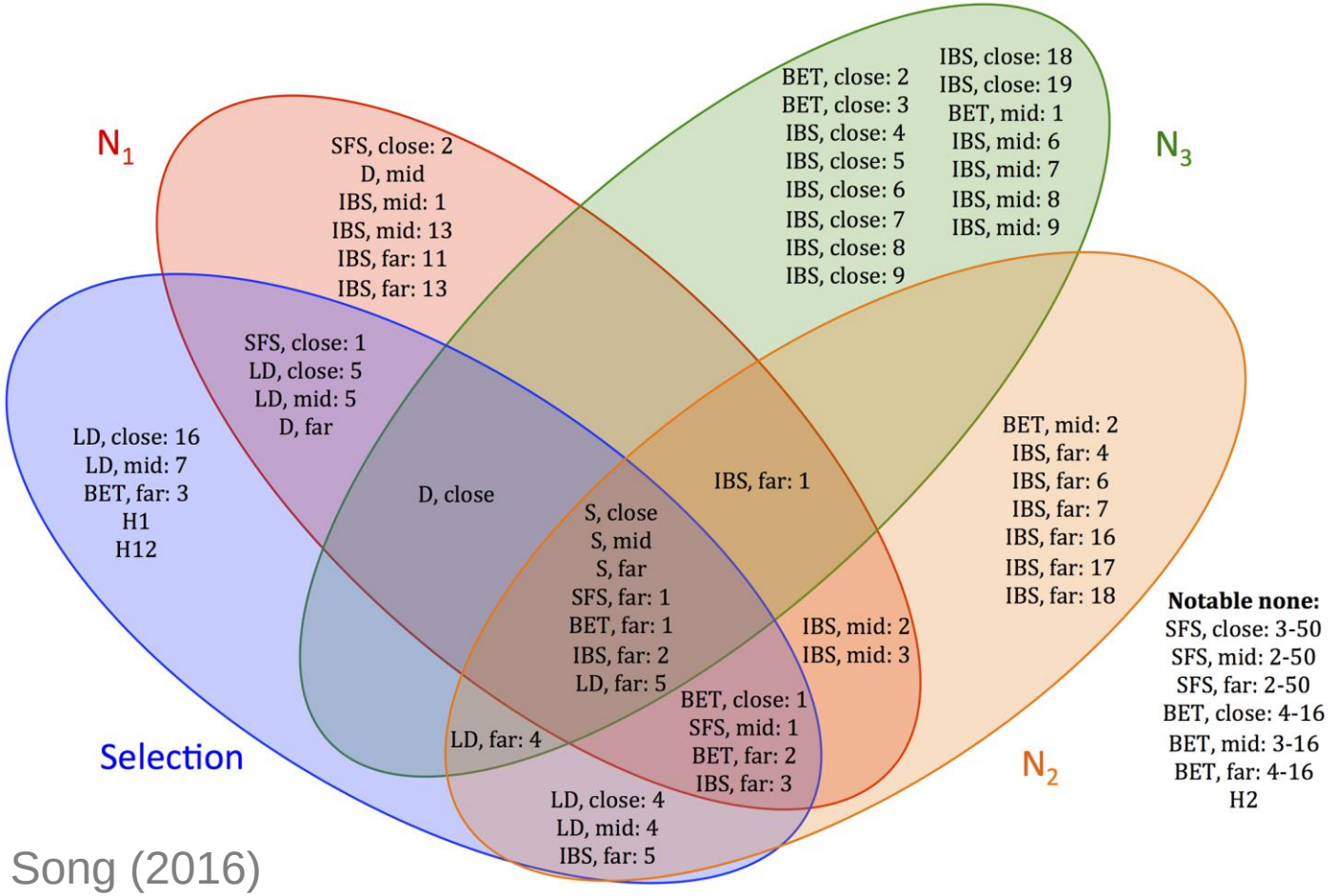
ANNs in population genetics

- African *Drosophila melanogaster* population
- windows classification: neutral, hard sweep, soft sweep, and balancing selection

Sheehan and Song (2016)

ANNs in population genetics

- most informative summary statistics
- LD statistics are most important for selection



Sheehan and Song (2016)

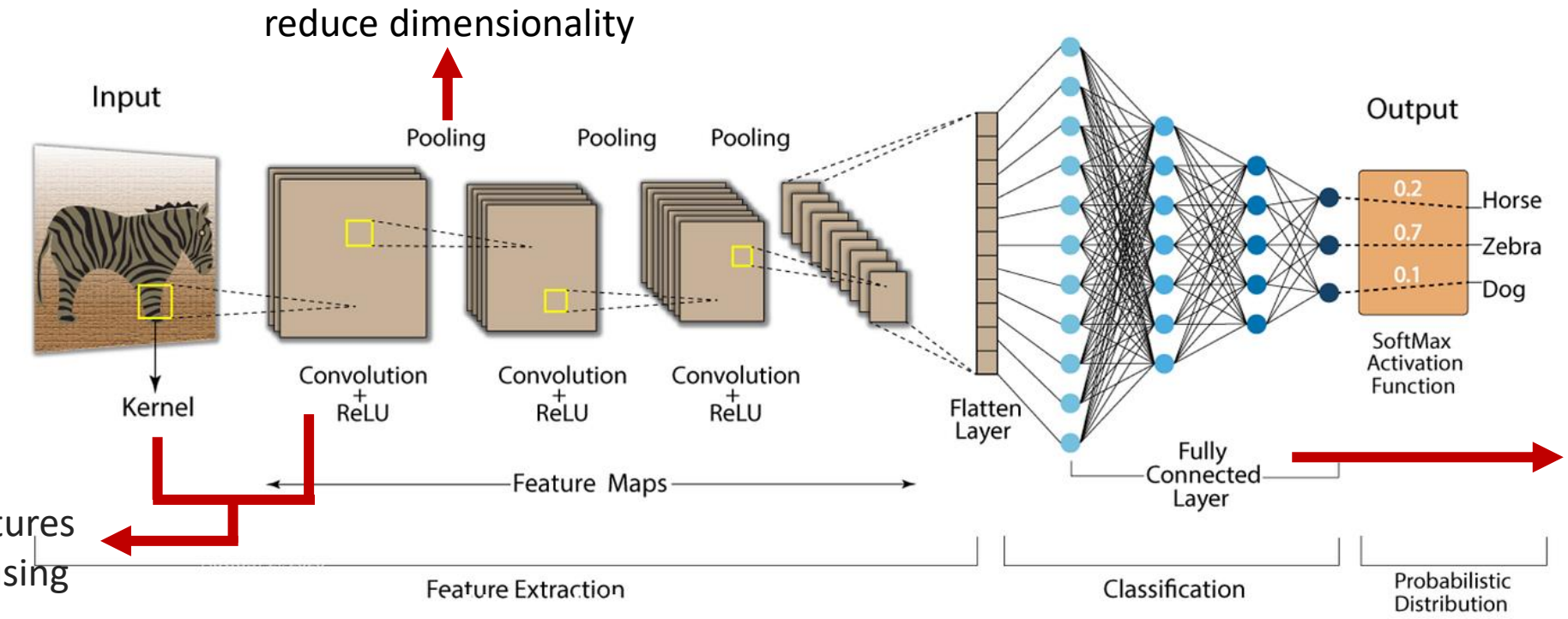
ANNs: exercise

Visit playground.tensorflow.org and assemble a neural network that can satisfactorily classify the orange and blue dots of this two datasets:

- For that, play with the activation function, the number of hidden layers and number of neurons per layer.

Convolutional neural networks (CNNs)

- designed to analyse grid-like data, such as images
- are characterized by a feature extraction phase

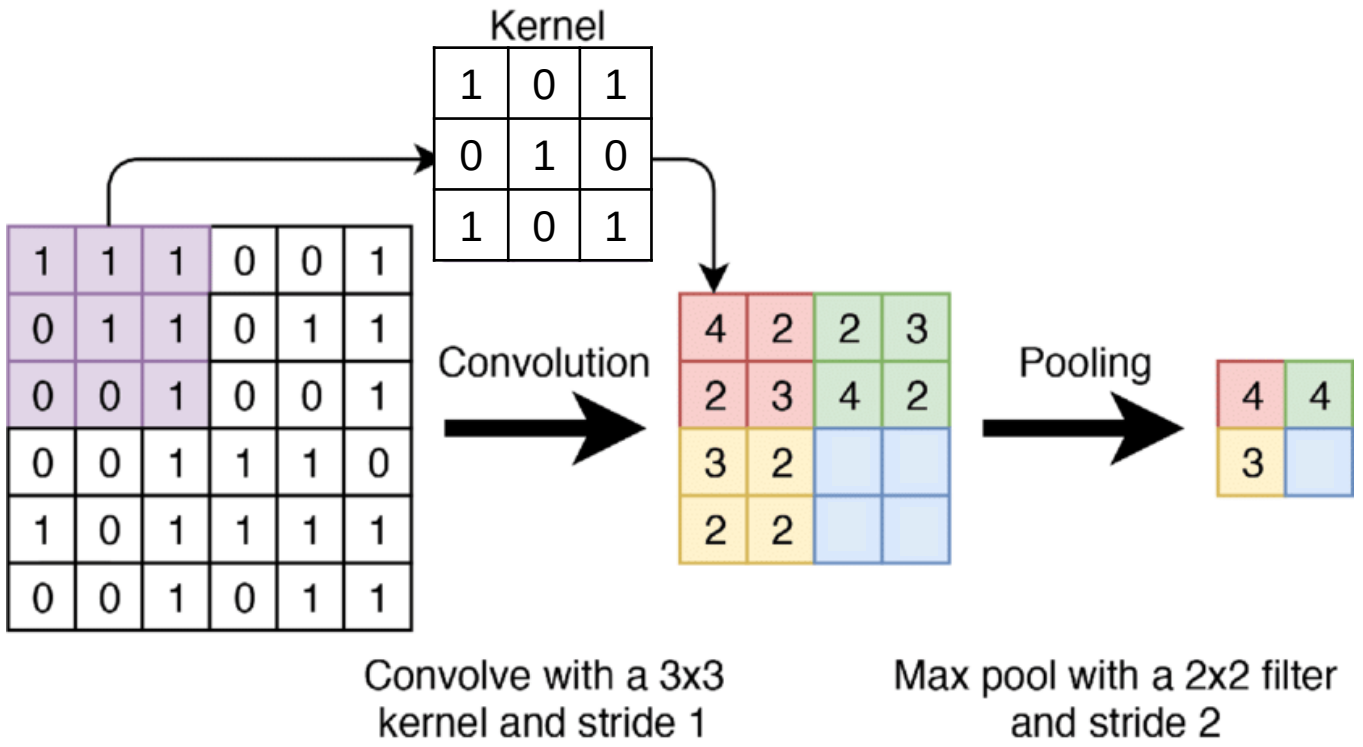


extract the features in the images using kernels

match the produced feature maps with the labels of the original image

CNNs: convolution and pooling

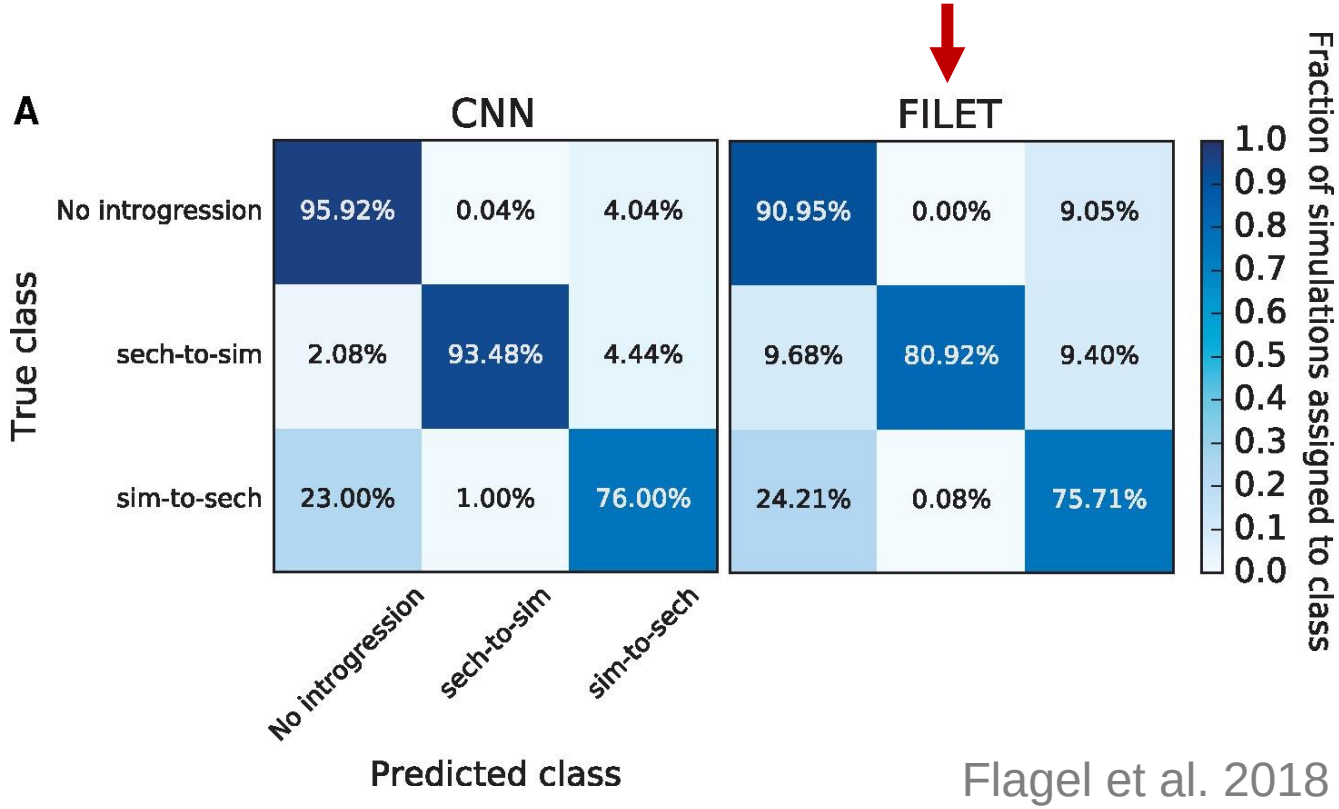
- What is convolution and pooling?



CNNs in population genetics

- characterize introgression between *Drosophila simulans* and *D. sechellia* species pair
- there is evidence for recent gene flow

a machine learning approach that leverages a vector of 31 summary statistics

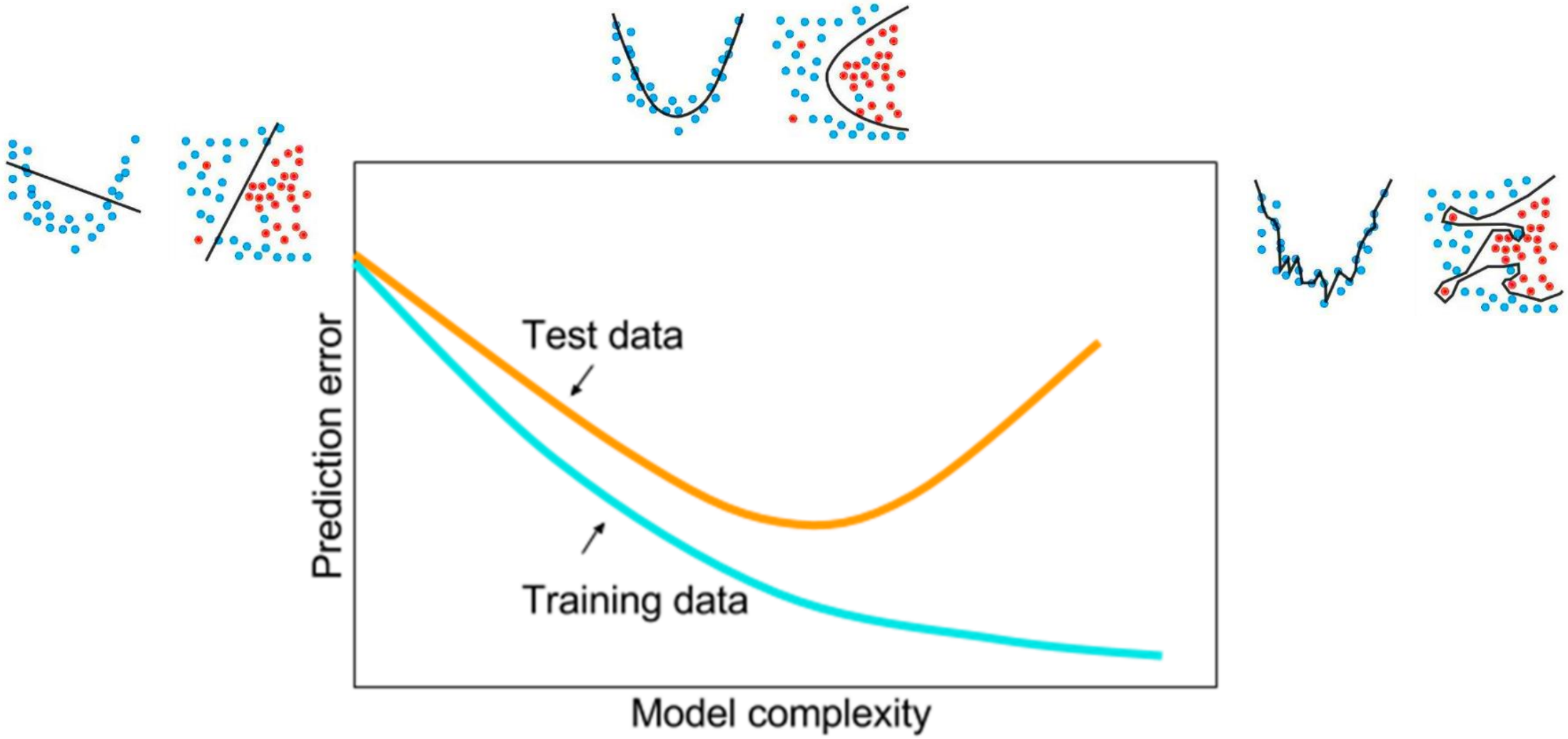


Flagel et al. 2018

Deep learning: an important math result

- the universal approximation theorem
- states that every continuous function can be approximated arbitrarily closely by a neural network with just one hidden layer.
- holds only for restricted classes of activation functions

Deep learning: model complexity & error



Concluding remarks

- the future of population genomic analysis rests in our ability to make sense of large and ever-growing datasets
- supervised ML techniques represent a new paradigm for analysis in the context of high-dimensional data produced by an unknown or imprecisely parameterized model
- ML provides robust, computationally efficient inference for several problems that are difficult to gain traction on via classical statistical approaches

Tutorial

Detecting selective sweeps from Evolve and Resequence experiments using deep learning and decision trees

Group discussion

Future challenges in machine learning

Group discussion

- ML applications relying on simulated training data must make modelling assumptions. Are we limited by the current simulators?
- Can ML substitute population genetic simulation?
- How feasible will parameter estimation be in more complex evolutionary models?
- Can ML do better than standard population genetic statistics?
- How can we use ML for assessing uncertainty on parameter estimates?