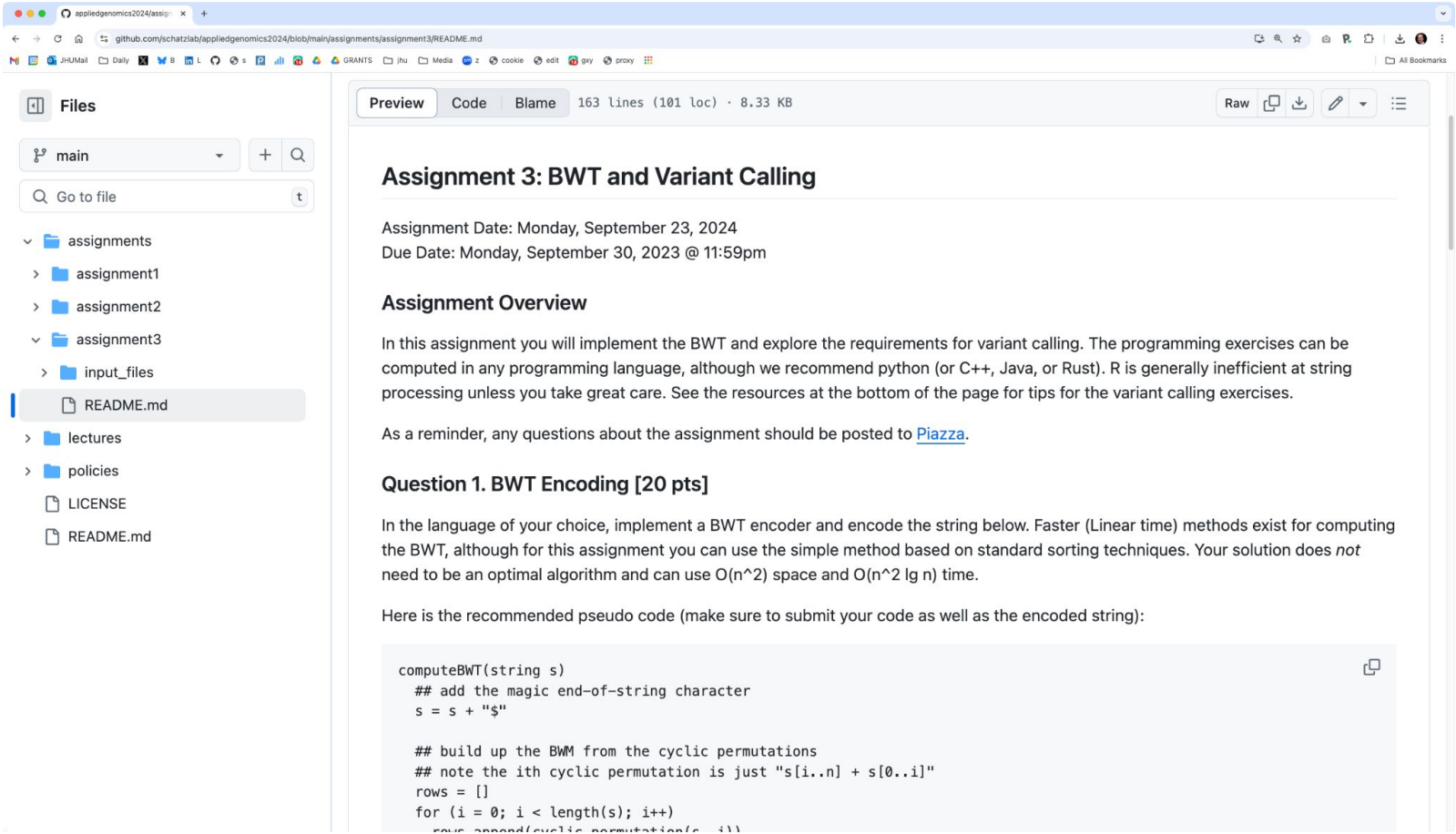
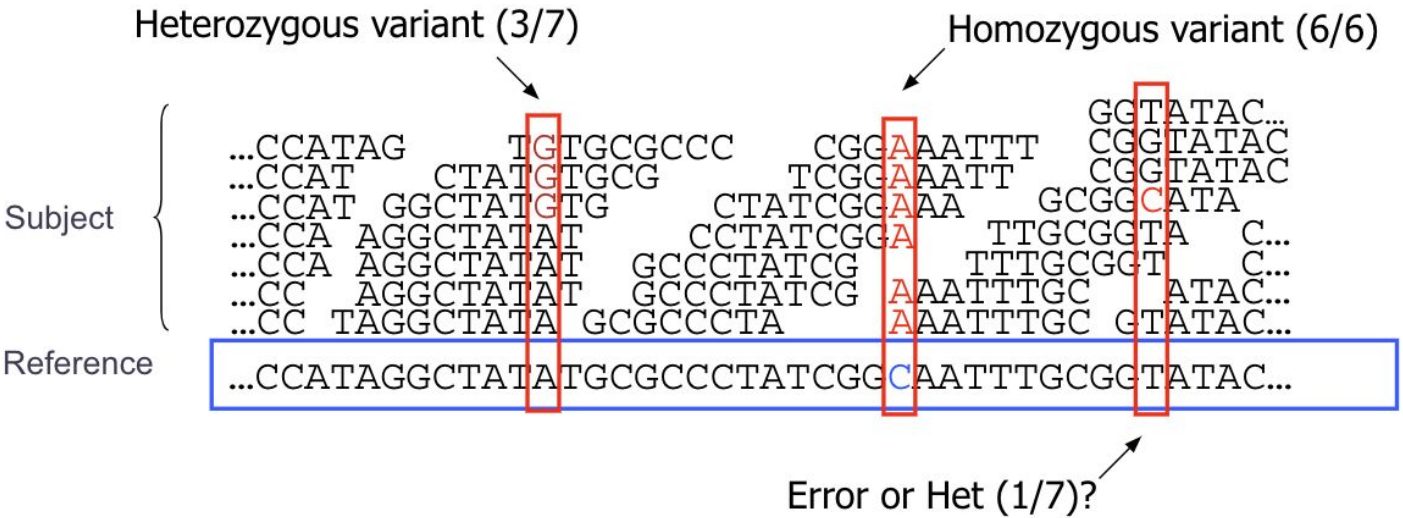
Introduction to Machine Learning

Matthew Nguyen & Michael Schatz
September 25, 2024

Assignment 3 Due Monday September 30

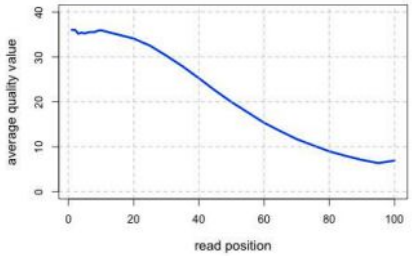


Genotyping Theory

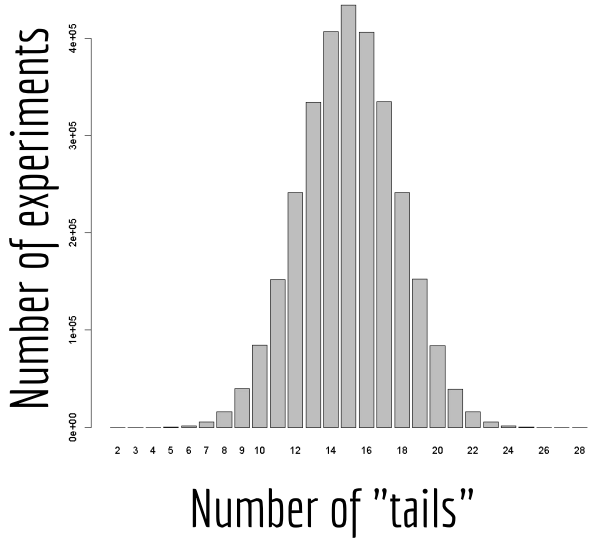


- If there were no sequencing errors, identifying SNPs would be very easy: any time a read disagrees with the reference, it must be a variant!
- Sequencing instruments make mistakes
  - Quality of read decreases over the read length
- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times

$$Q_{\text{sanger}} = -10 \log_{10} p$$



So, with 30 tosses (reads), we are much more likely to see an even mix of alternate and reference alleles at a heterozygous locus in a genome



This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to distinguish heterozygous alleles and from mere sequencing errors

P(3/30 het) <?> P(3/30 err)

PolyBayes: The first statistically rigorous variant detection tool.

letter

A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

Bayesian posterior probability

Base call + Base quality

Expected (prior) polymorphism rate

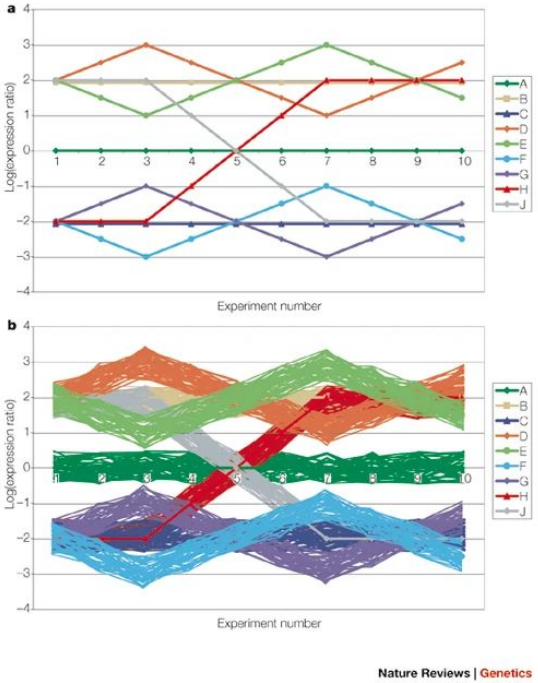
$$P(SNP) = \sum_{all\ variable\ S} \frac{\frac{P(S_1 \mid R_1)}{P_{Prior}(S_1)} \cdot \ldots \cdot \frac{P(S_N \mid R_N)}{P_{Prior}(S_N)} \cdot P_{Prior}(S_1,\ldots,S_N)}{\sum_{S_{i_1} \in [A,C,G,T]} \cdots \sum_{S_{i_N} \in [A,C,G,T]} \frac{P(S_{i_1} \mid R_1)}{P_{Prior}(S_{i_1})} \cdot \ldots \cdot \frac{P(S_{i_N} \mid R_1)}{P_{Prior}(S_{i_N})} \cdot P_{Prior}(S_{i_1},\ldots,S_{i_N})}$$

Probability of observed base composition (should model sequencing error rate)

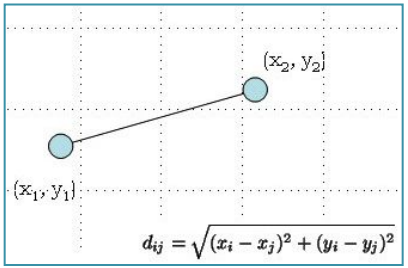
http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

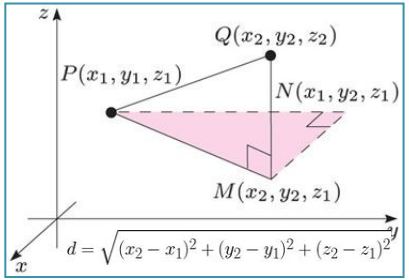
Introduction to Machine Learning: Clustering & Dimensionality Reduction

Clustering Refresher



Euclidean Distance



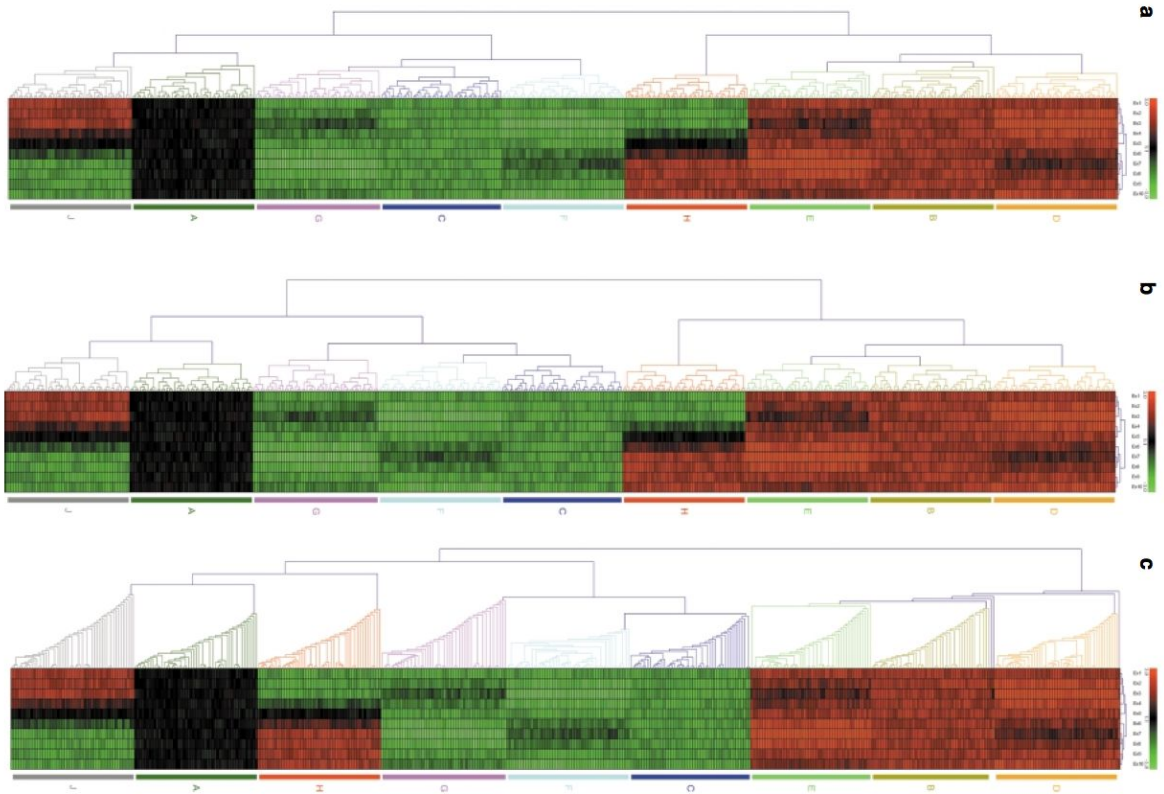


$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \mathrm{d}(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Computational genetics: Computational analysis of microarray data

Quackenbush (2001) *Nature Reviews Genetics.* doi:10.1038/35076576

Hierarchical Clustering

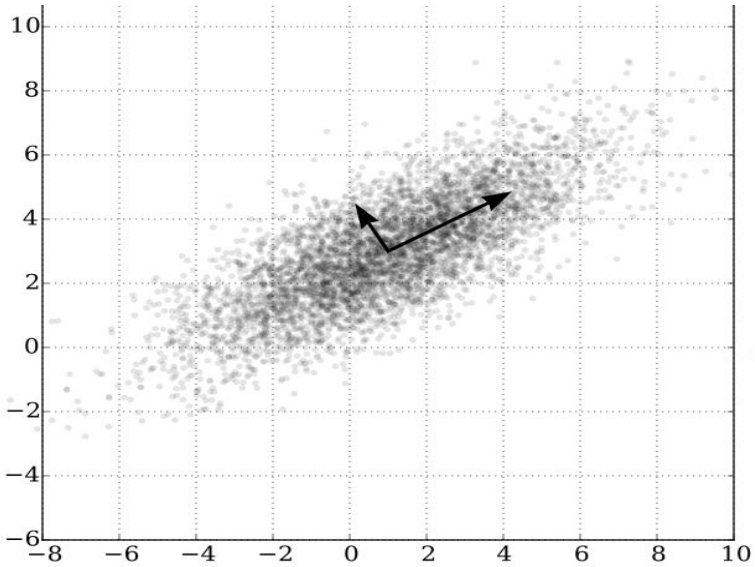


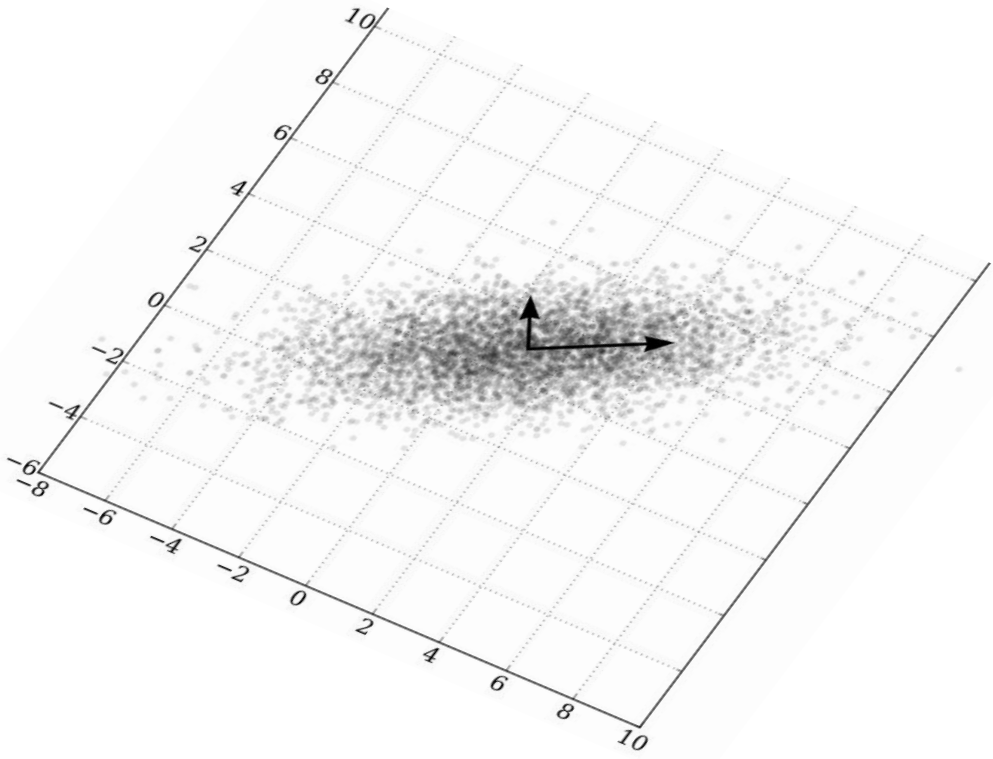
average

complete

single

Principal Components Analysis (PCA)

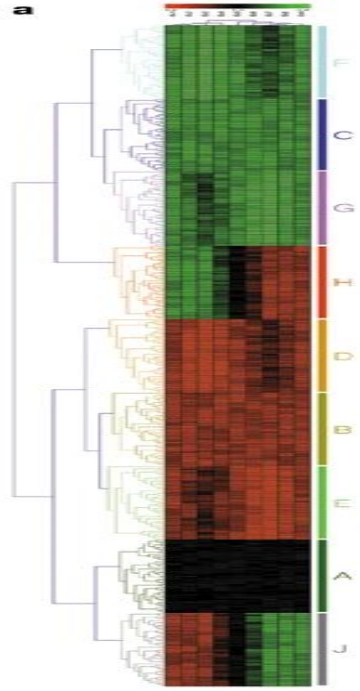


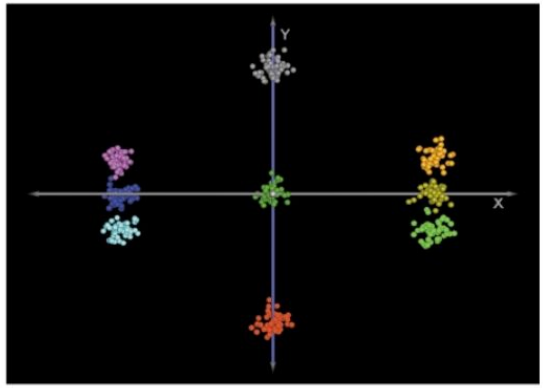


PC1: “New X”- The dimension with the most variability

PC2: “New Y”- The dimension with the second most variability

Principal Components Analysis (PCA)





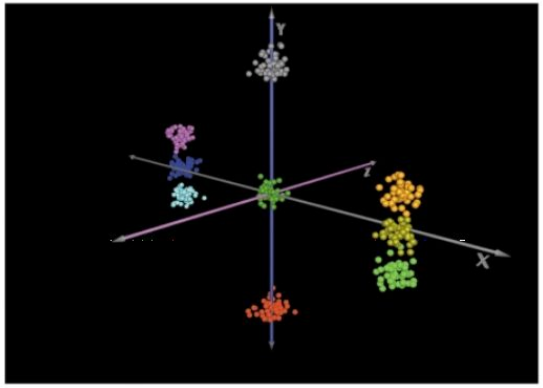
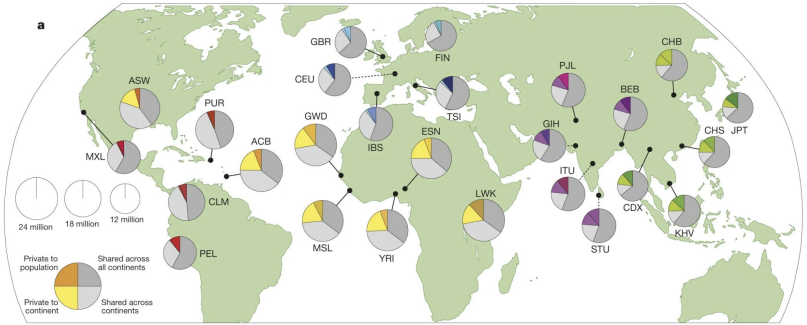
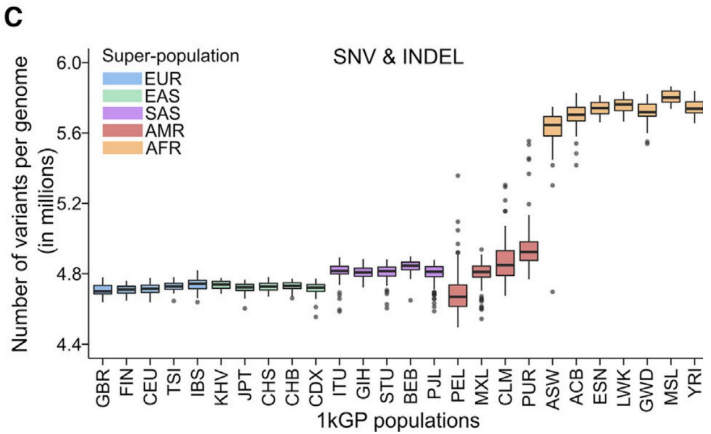


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering results for comparison.

1000 Genomes Project

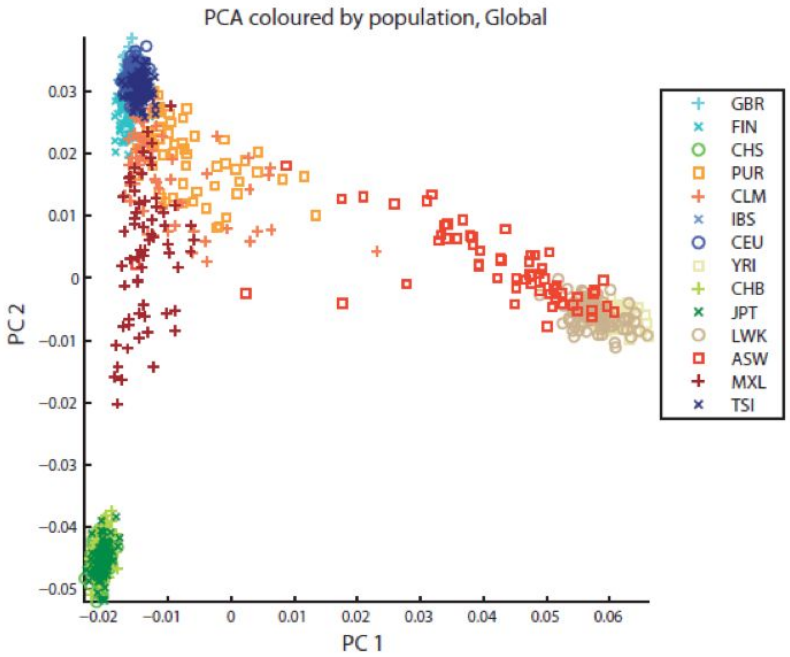


EAS	CHB	Han Chinese	Han Chinese in Beijing, China
	JPT	Japanese	Japanese in Tokyo, Japan
	CHS	Southern Han Chinese	Han Chinese South
	CDX	Dai Chinese	Chinese Dai in Xishuangbanna, China
	KHV	Kinh Vietnamese	Kinh in Ho Chi Minh City, Vietnam
	CHD	Denver Chinese	Chinese in Denver, Colorado (pilot 3 only)
EUR	CEU	CEPH	Utah residents (CEPH) with Northern and Western European ancestry
	TSI	Tuscan	Toscani in Italia
	GBR	British	British in England and Scotland
	FIN	Finnish	Finnish in Finland
	IBS	Spanish	Iberian populations in Spain
AFR	YRI	Yoruba	Yoruba in Ibadan, Nigeria
	LWK	Luhya	Luhya in Webuye, Kenya
	GWD	Gambian	Gambian in Western Division, The Gambia
	MSL	Mende	Mende in Sierra Leone
	ESN	Esan	Esan in Nigeria
AMR	ASW	African-American SW	African Ancestry in Southwest US
	ACB	African-Caribbean	African Caribbean in Barbados
	MXL	Mexican-American	Mexican Ancestry in Los Angeles, California
	PUR	Puerto Rican	Puerto Rican in Puerto Rico
	CLM	Colombian	Colombian in Medellin, Colombia
	PEL	Peruvian	Peruvian in Lima, Peru
SAS	GIH	Gujarati	Gujarati Indian in Houston, TX
	PJL	Punjabi	Punjabi in Lahore, Pakistan
	BEB	Bengali	Bengali in Bangladesh
	STU	Sri Lankan	Sri Lankan Tamil in the UK
	ITU	Indian	Indian Telugu in the UK



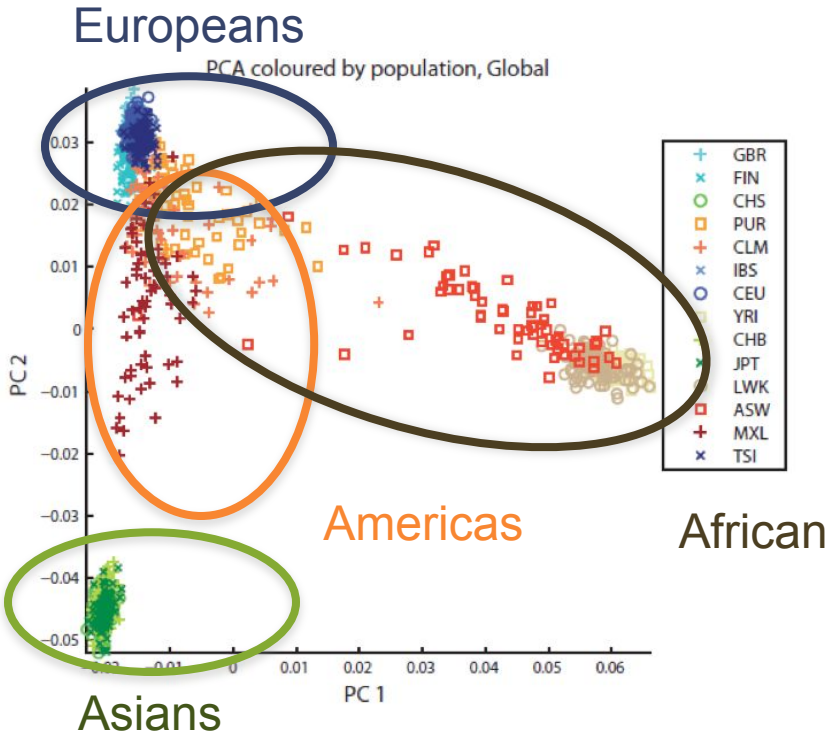
High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios
Byrska-Bishop et al. (2022) Cell. doi: 10.1016/j.cell.2022.08.004

Variation across populations



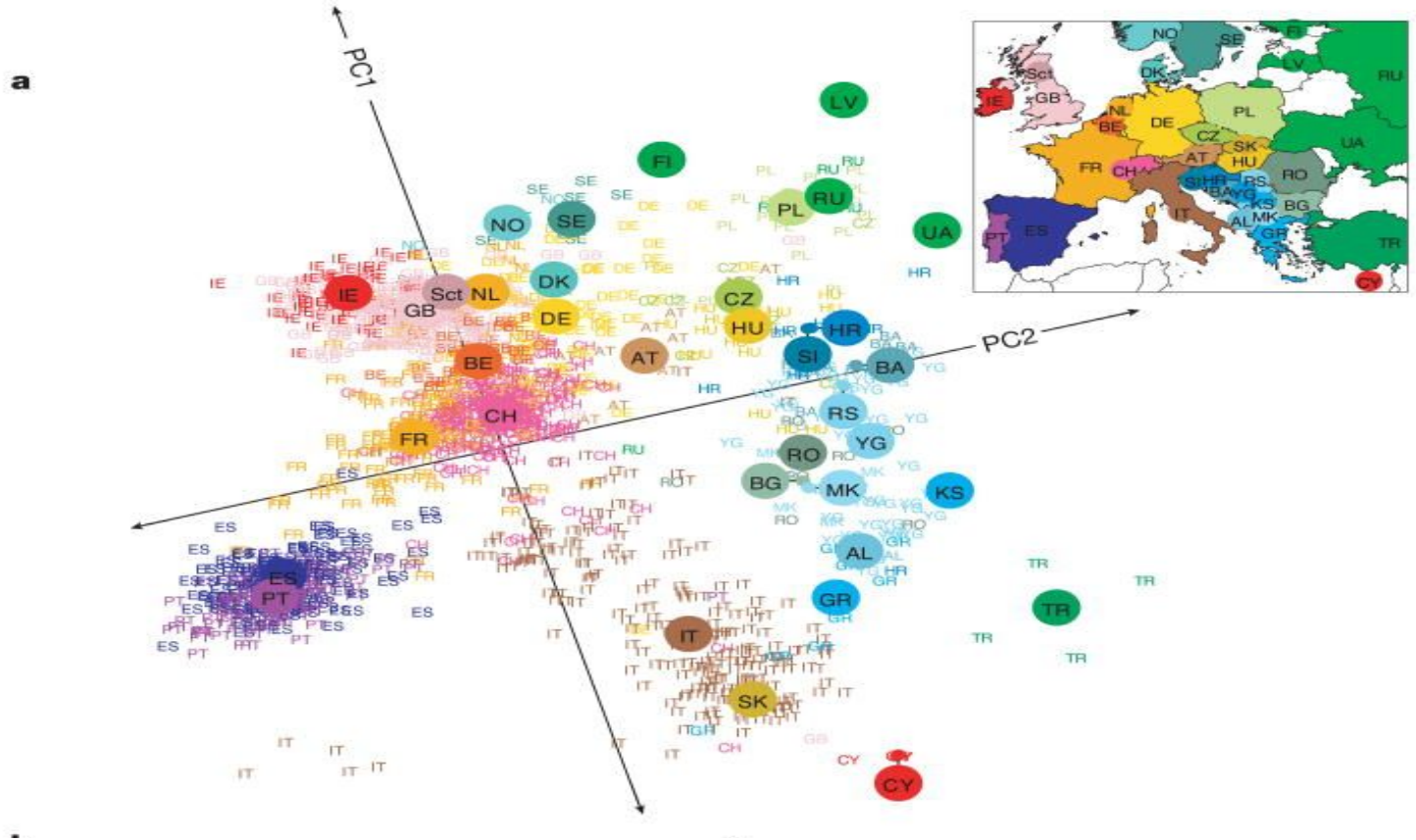
	Person1	Person2	Person3	…
SNP1	0/0	0/1	0/1	
SNP2	0/1	0/1	0/1	
SNP3	0/0	0/0	1/1	
SNP4	1/1	0/1	1/1	
SNP5	0/0	1/1	1/1	
SNP5	0/0	0/0	0/1	
…				

Variation across populations



	Person1	Person2	Person3	…
SNP1	0/0	0/1	0/1	
SNP2	0/1	0/1	0/1	
SNP3	0/0	0/0	1/1	
SNP4	1/1	0/1	1/1	
SNP5	0/0	1/1	1/1	
SNP5	0/0	0/0	0/1	
…				

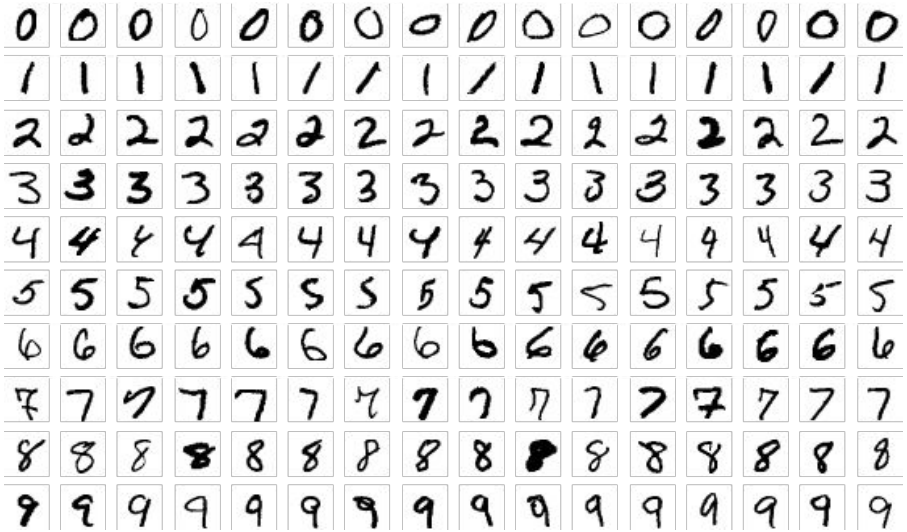
PC1 is largely correlated with African ancestry:
Human genetic diversity is dominated by African ancestry



Genes mirror geography within Europe

Novembre et al (2008) Nature. doi: 10.1038/nature07331

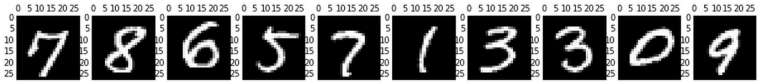
MNIST



- 70,000 images, each labeled with the correct digit (0-9)
  - 28 x 28 thumbnail images of handwritten digits
  - Grayscale image with 256 possible values (784 bytes per image)

- Widely used dataset for image classification
  - Very simple, the categories are very easy to understand

MNIST: PCA

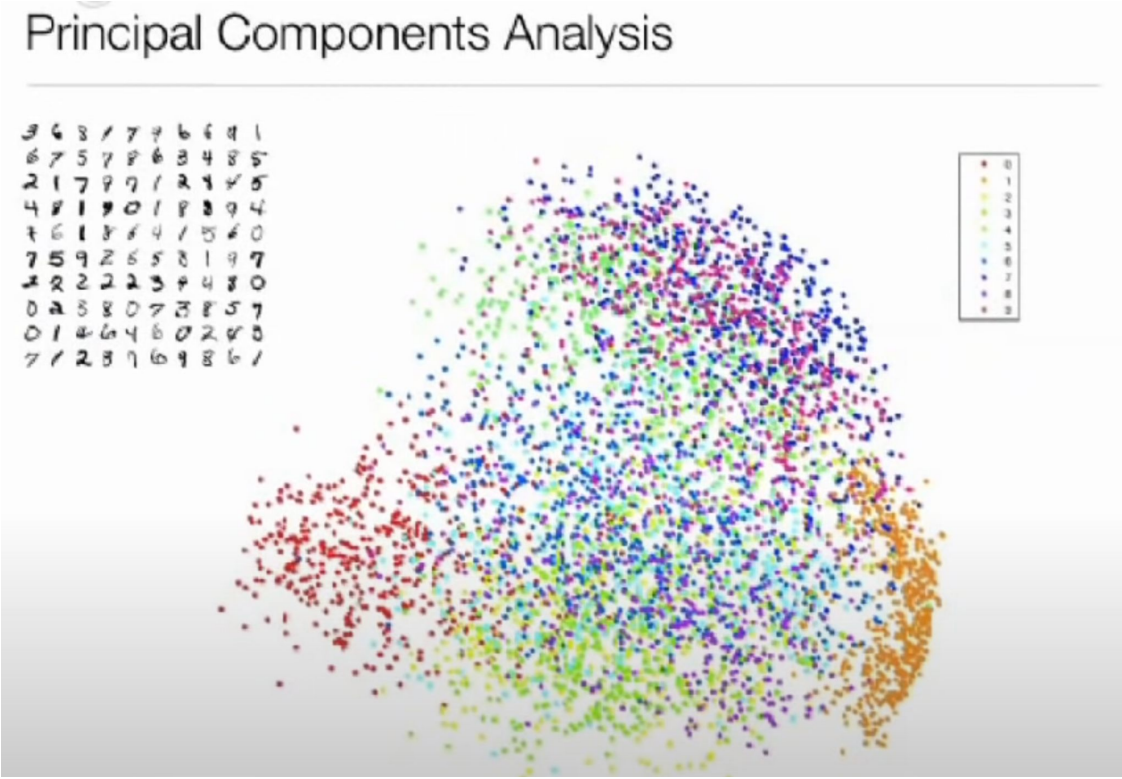


Each 28x28 image can be considered a point in 784 dimensional space!

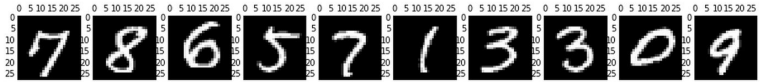
PCA roughly groups digits together

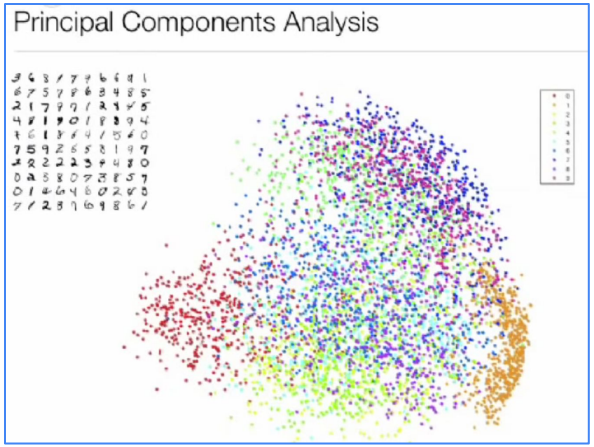
- 0 (red) in bottom left
- 1 (orange) in bottom right

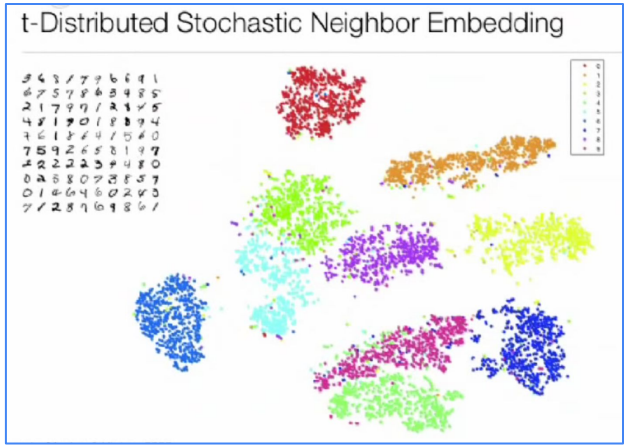
But very incomplete separation, lots of mixing of colors



MNIST: PCA and t-SNE







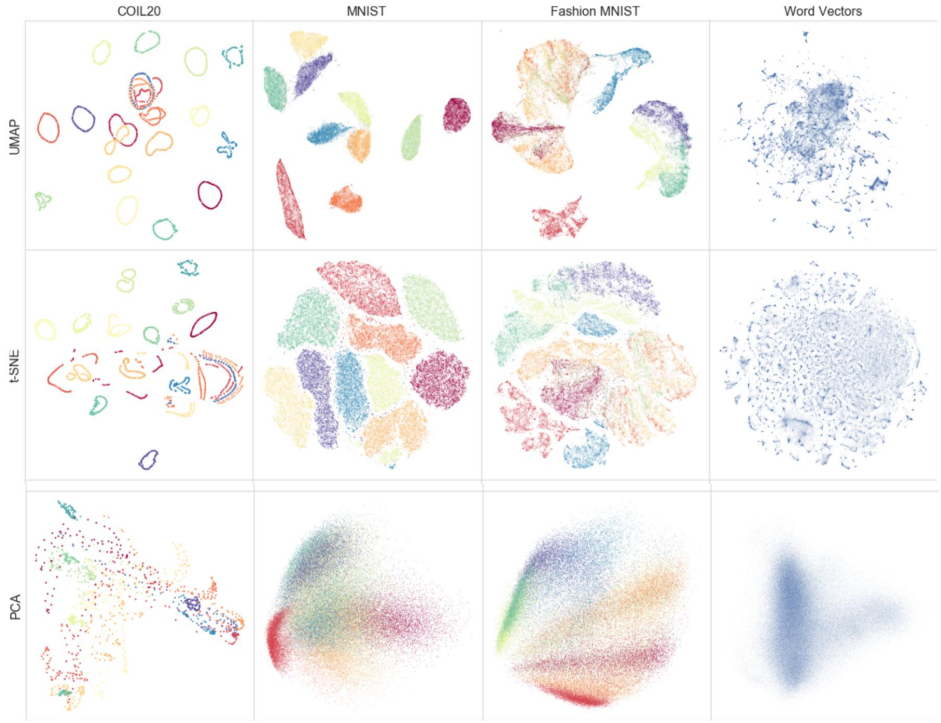
t-distributed Stochastic Neighborhood Embedding

- Non-linear dimensionality reduction technique: distances are only locally meaningful
- Rather than Euclidean distances, for each point fits a Gaussian kernel to fit the nearest N neighbors (perplexity) that define the probabilities that two points should be close together
- Using an iterative spring embedding system to place high probability points nearby

Visualizing Data Using t-SNE

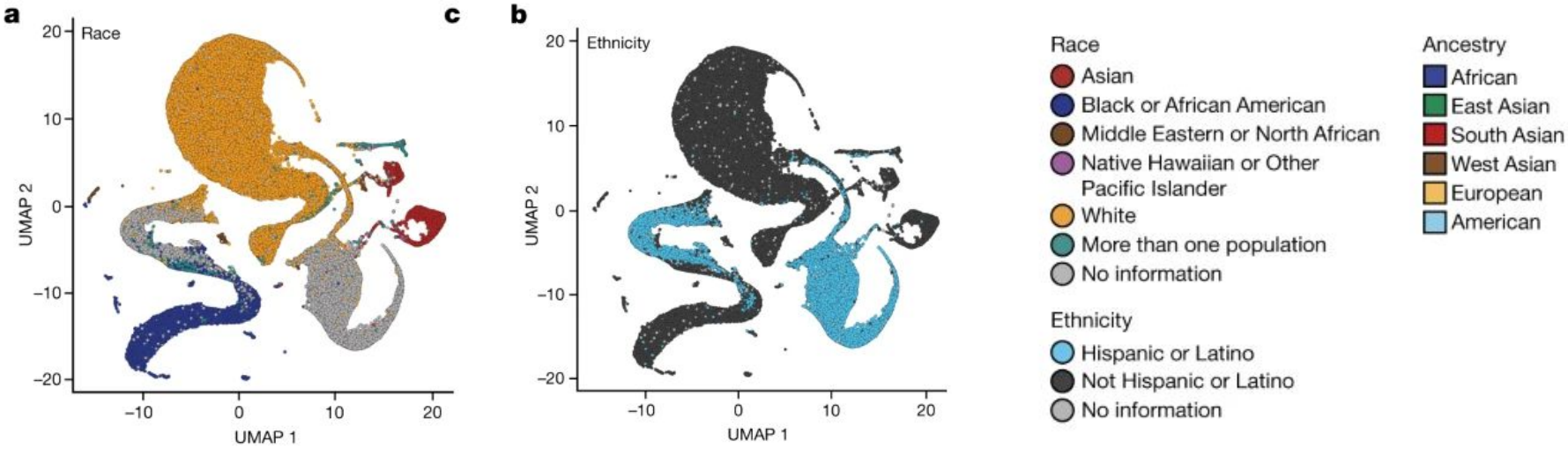
https://www.youtube.com/watch?v=RJVL80Gg3lA

UMAP



UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
McInnes et al (2018) arXiv. 1802.03426
https://www.youtube.com/watch?v=nq6iPZVUxZU
https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668

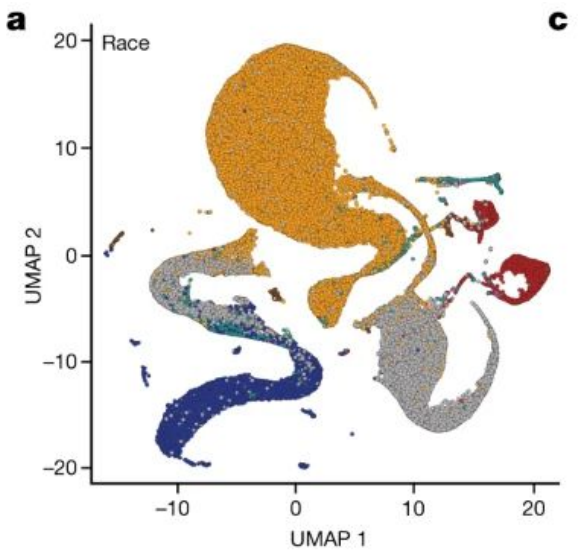
UMAP of Human Variation



Genomic data in the All of Us Research Program

The All of Us Research Program Genomics Investigators (2024) Nature. doi.org/10.1038/s41586-023-06957-x

~~UMAP of Human Variation~~



"It's a pity that All of Us used UMAP to visualize ancestry variation in their new marker paper, out today in Nature. The UMAP algorithm, by design, exaggerates the distinctiveness of the most frequent ancestries, a message that can be misinterpreted by the public. UMAP pulls unusual genotypes towards the majority clusters; in particular it fails to represent admixture in a sensible way (admixture is fundamentally additive, while UMAP is not). In this setting the messiness of Admixture or PCA plots yield a better reflection of the data"

Jonathan Pritchard
Stanford

Genomic data in the All of Us Research Program
The All of Us Research Program Genomics Investigators (2024) Nature. doi.org/10.1038/s41586-023-06957-x

Introduction to Machine Learning: Classification & Prediction

Classification Problems

Supervised learning problem:

- Given (many) labeled training examples, develop a classifier that will accurately predict an output
- Each example contains **features** used to form a prediction

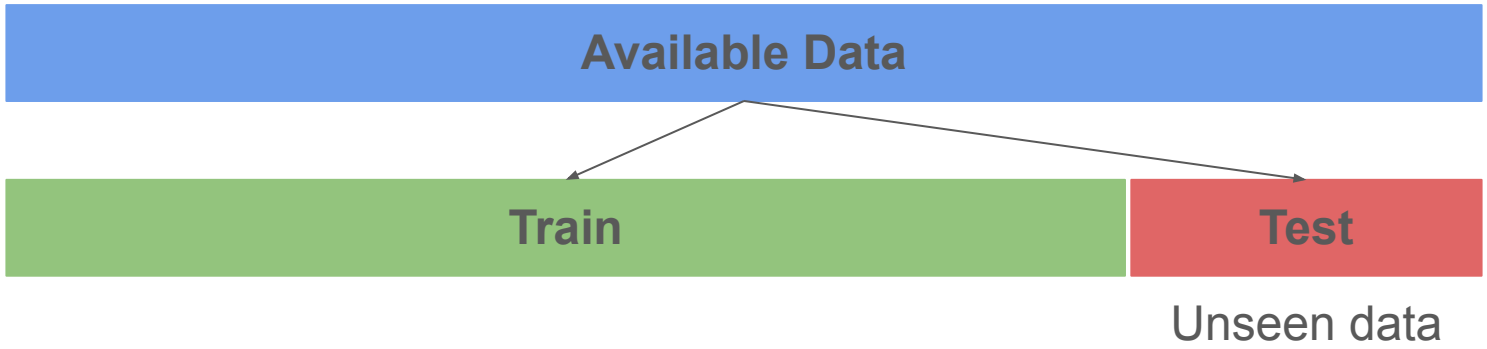
Examples:

- Given a pileup of reads, predict if there really is a variant
- Given SNVs in a person, predict ancestry (or relatedness)
- Given expression matrix, predict sample type
- Given SNVs + cell type, predict changes in expression
- Given SNVs/expression, predict disease risk
- Given SNVs/expression & disease status, predict treatment
- Given an image (photograph, xray, MRI), predict what it is (object, tumor, etc)
- Given English text, classify the sentiment (happy/sad) or predict the next work
- …

Training datasets and cross validation

We need to do two things with the data:

1. Estimate the parameters for the ML method **(Training)**
2. Evaluate how well the ML method works **(Testing)**

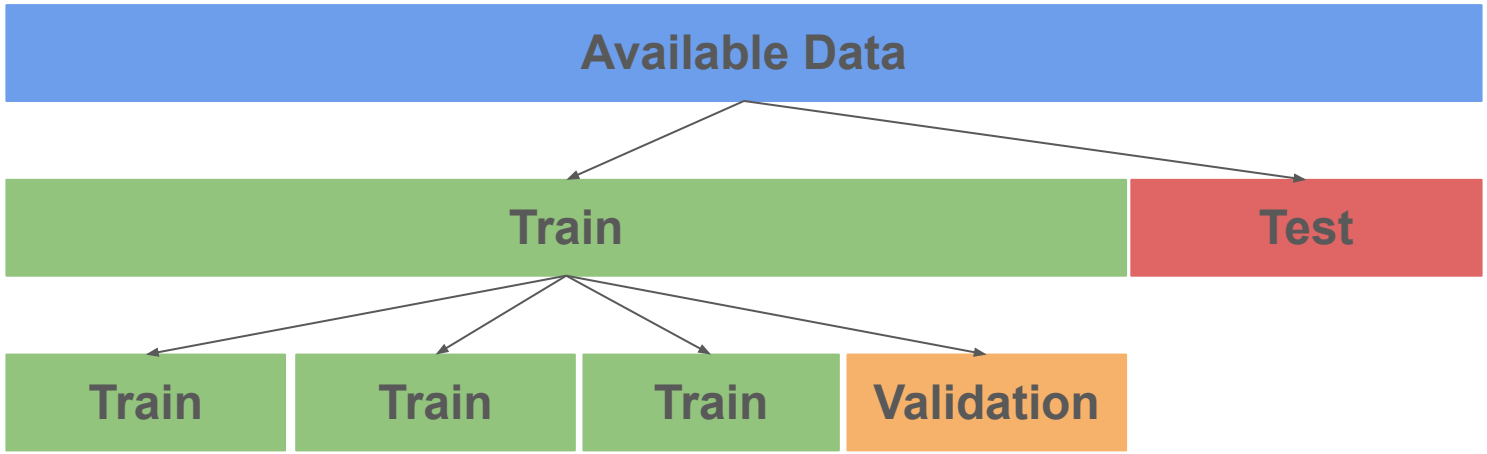


But how do we optimize the model?

Cross validation

We can split our training data into 75% train and 25% validation

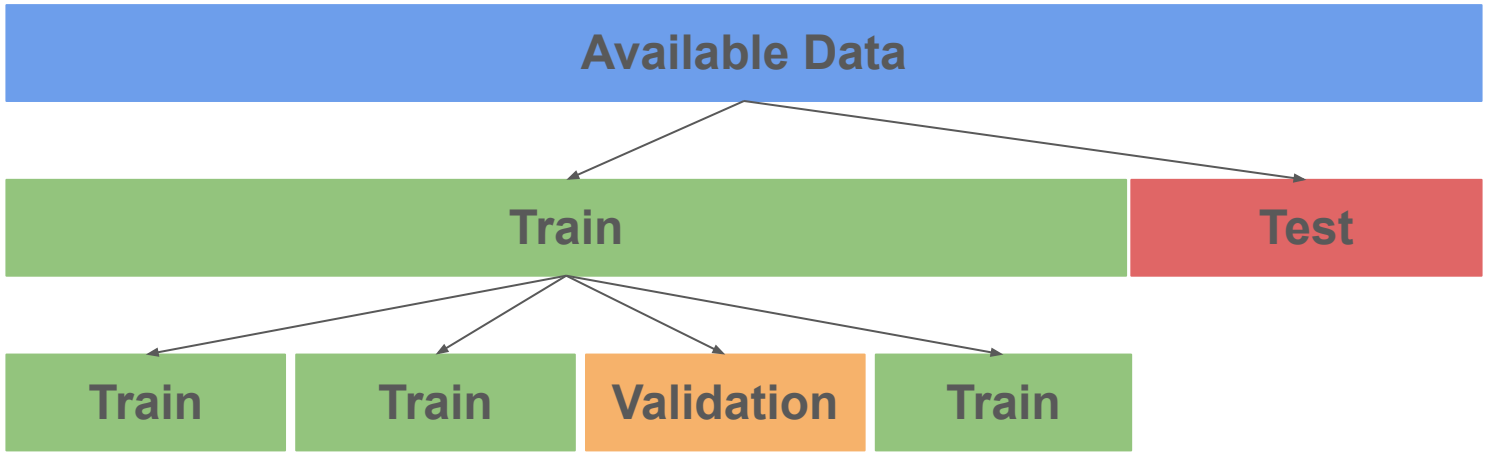
- But which block do we use for validation?



Cross validation

We can split our training data into 75% train and 25% validation

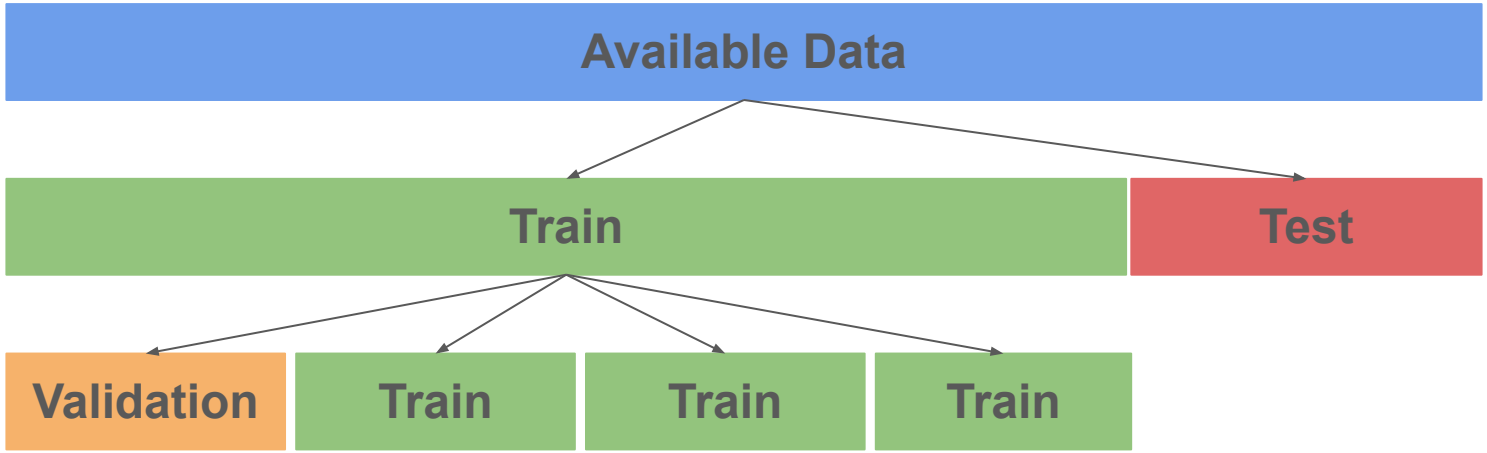
- But which block do we use for validation?



Cross validation

We can split our training data into 75% train and 25% validation

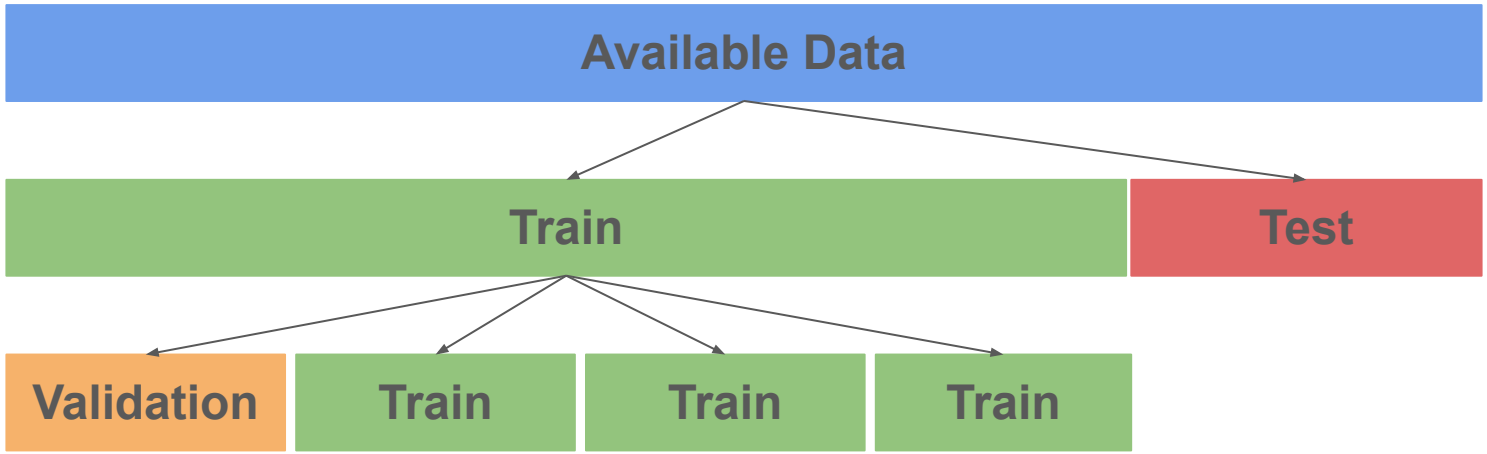
- But which block do we use for validation?



Cross validation

We can split our training data into 75% train and 25% validation

- But which block do we use for validation?
  - **Why not all of them?**



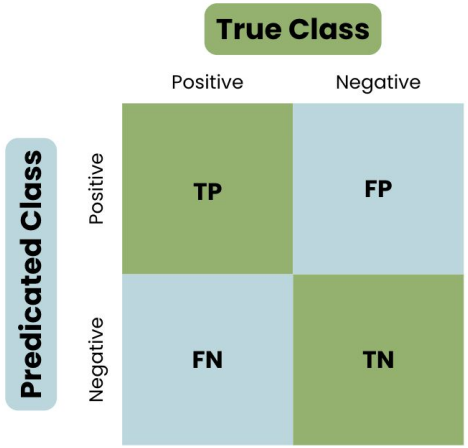
Cross validation

k-Fold cross validation

- Split training data into *k* splits

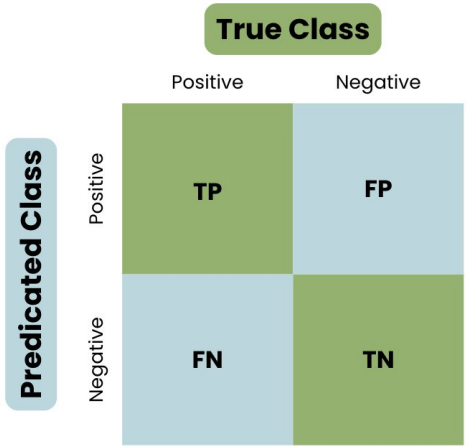
Evaluation metrics

Confusion matrix:



Evaluation metrics

Confusion matrix:

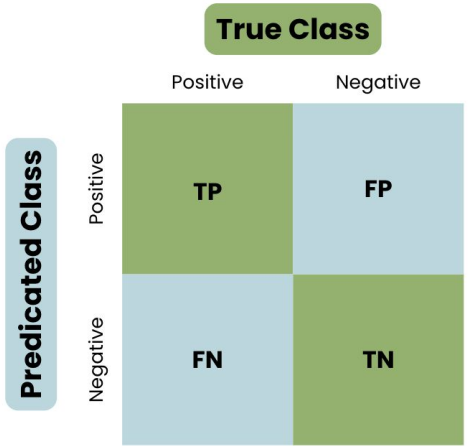


Accuracy:

$$\textit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Evaluation metrics

Confusion matrix:

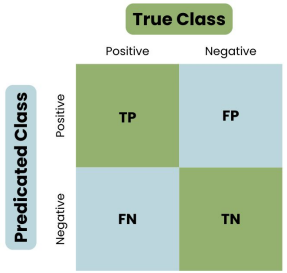


Accuracy:

- But what if the classes are unbalanced?

$$\textit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Evaluation metrics

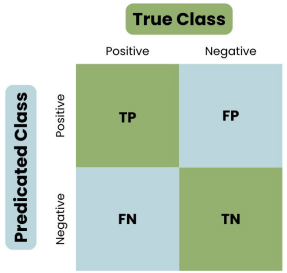


Precision:

- How many correctly predicted cases turned out to be positive
- False positive is of higher concern

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Evaluation metrics



Precision:

- How many correctly predicted cases turned out to be positive
- False positive is of higher concern

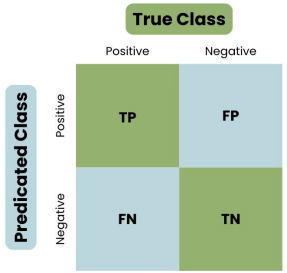
$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall (sensitivity):

- How many of the actual positive cases we were able to predict correctly
- False negative is of higher concern

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

Evaluation metrics



Precision:

- How many correctly predicted cases turned out to be positive
- False positive is of higher concern

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall (sensitivity):

- How many of the actual positive cases we were able to predict correctly
- False negative is of higher concern

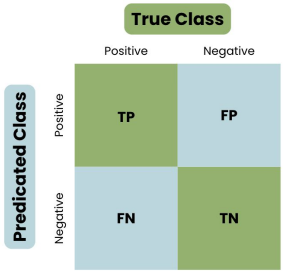
$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

F1 Score:

- Combines precision/recall (harmonic mean)

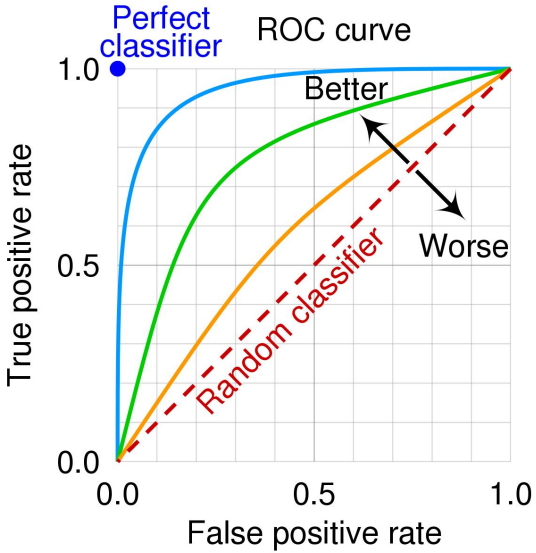
$$F1 = 2.\frac{Precision \times Recall}{Precision + Recall}$$

Evaluation metrics



AUC-ROC

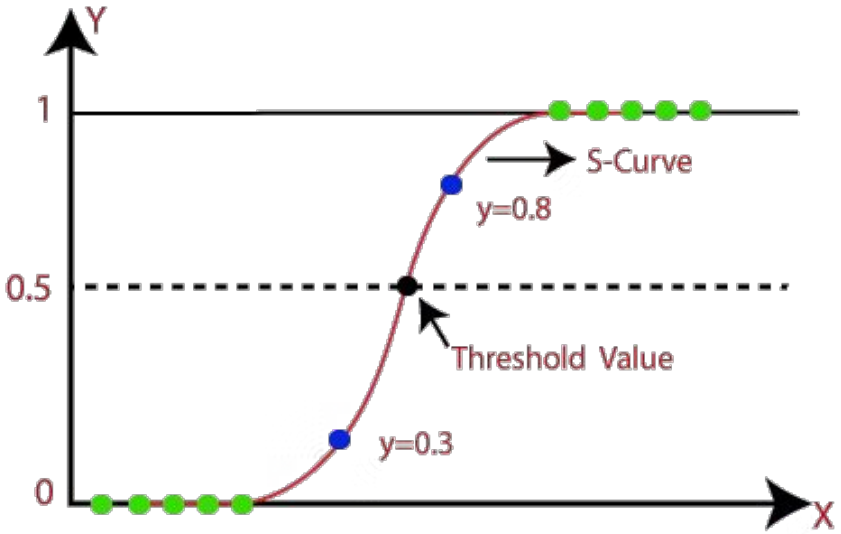
- Receiver Operator Characteristic (ROC) plots the true positive rate against the false positive rate
- Area Under the Curve (AUC) measure how the model distinguishes classes



Logistic Regression

Logistic regression assigns a weight to each feature to form a prediction

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$



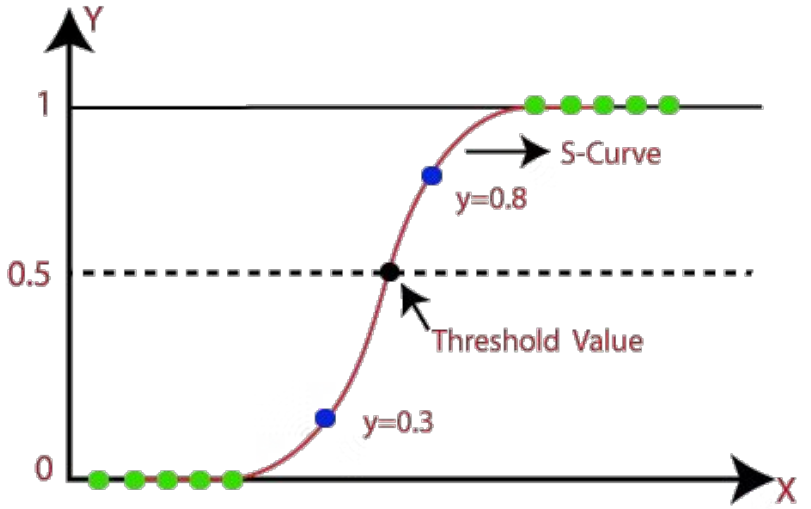
https://encord.com/blog/what-is-logistic-regression/

Logistic Regression

Logistic regression assigns a weight to each feature to form a prediction

Pros	Cons
• Easy to understand • Good accuracy when data is linearly separable • Less inclined to **overfit**	• Assumes linearity between output and features • No multicollinearity between features

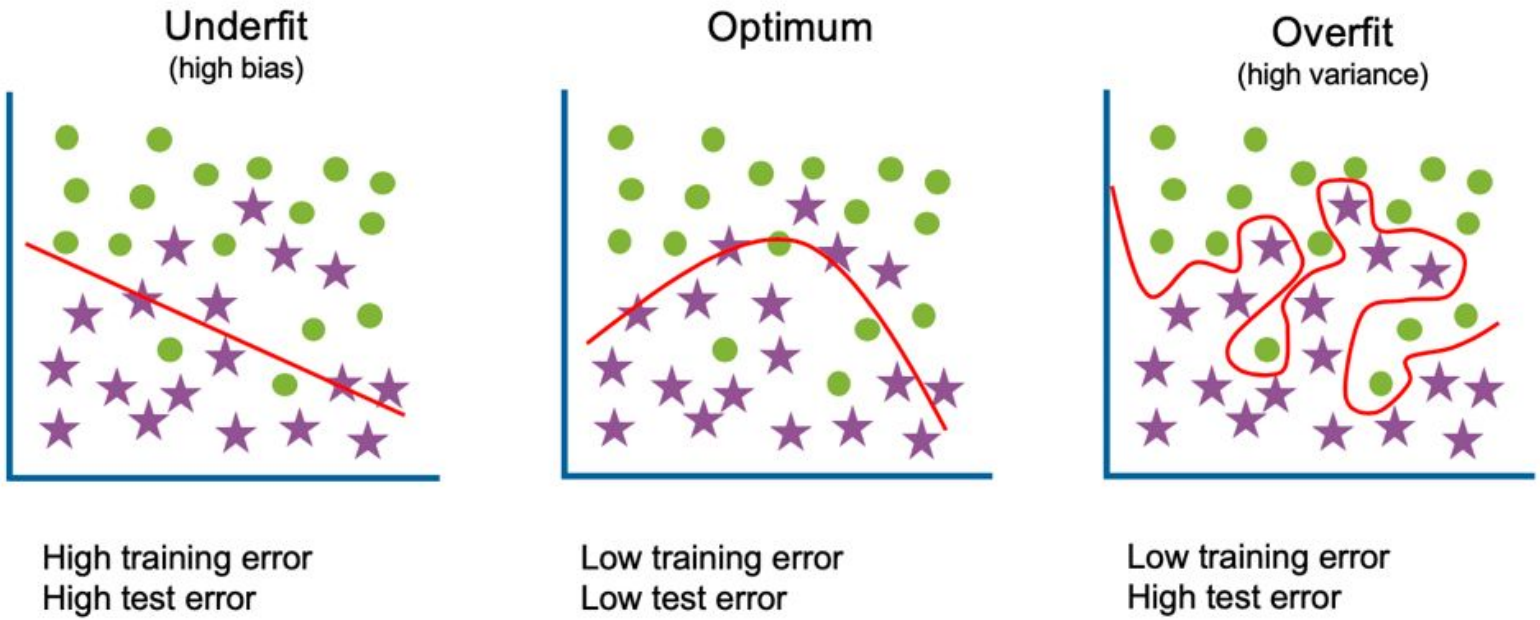
$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$



https://encord.com/blog/what-is-logistic-regression/

Overfitting

Model is very accurate on training data but can't generalize to new data



https://skyengine.ai/se/skyengine-blog/129-overfitting-in-machine-learning-and-computer-vision

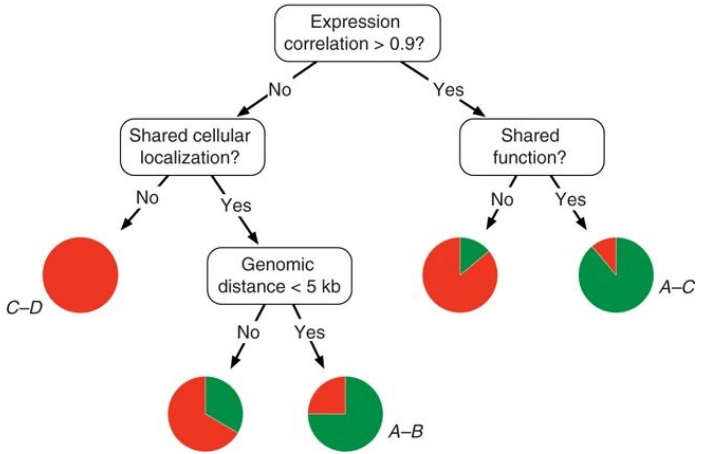
Decision Trees

Decision tree learns simple decision rules inferred from features

a

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
⋮					

b

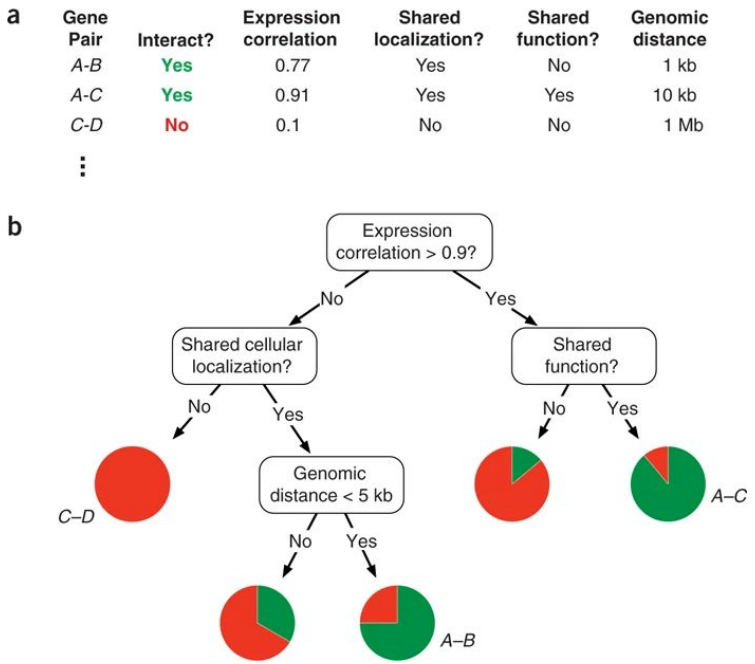


Kingsford, C., Salzberg, S. What are decision trees?. *Nat Biotechnol* 26, 1011–1013 (2008). https://doi.org/10.1038/nbt0908-1011

Decision Trees

Decision tree learns simple decision rules inferred from features

Pros	Cons
• Easy to understand • Can be visualized • Can handle multiple outputs	• Can create overcomplicated trees that overfit • Small variations in data can lead to completely different tree

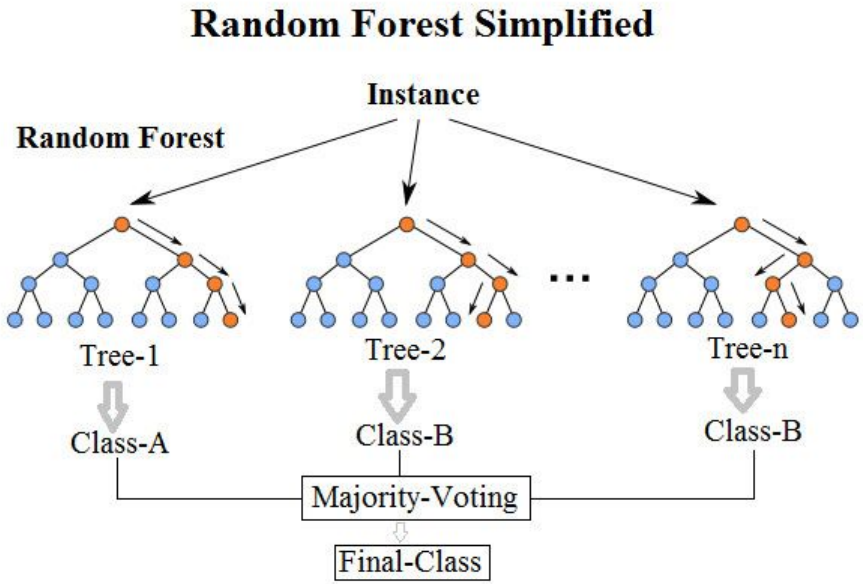


Kingsford, C., Salzberg, S. What are decision trees?. *Nat Biotechnol* 26, 1011–1013 (2008). https://doi.org/10.1038/nbt0908-1011

Random Forests

An ensemble of decision trees built from sample of training data

- Prediction is the averaged output of all decision trees



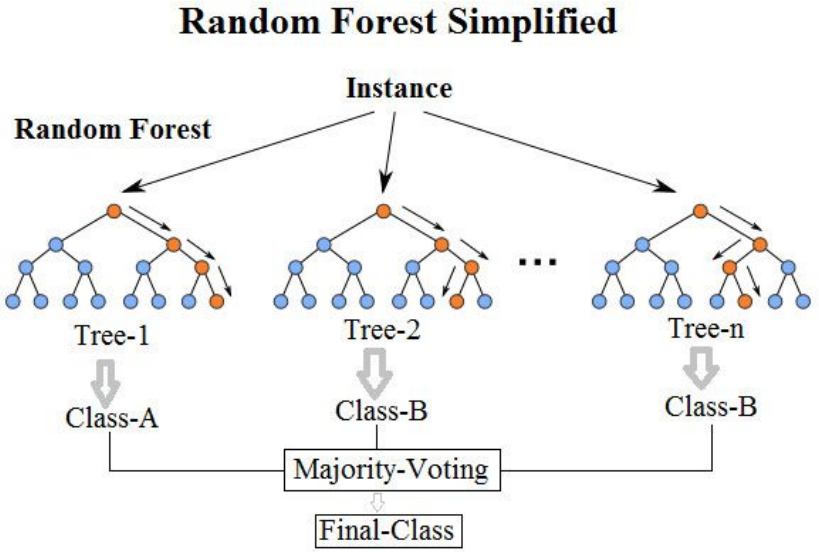
https://www.linkedin.com/pulse/random-forest-algorithm-interactive-discussion-niraj-kumar/

Random Forests

An ensemble of decision trees built from sample of training data

- Prediction is the averaged output of all decision trees

Pros	Cons
• Highly accurate • Intrinsically interpretable • Can learn non-linear decision boundaries • Parallel processing	• High computational complexity • Can overfit because of noise

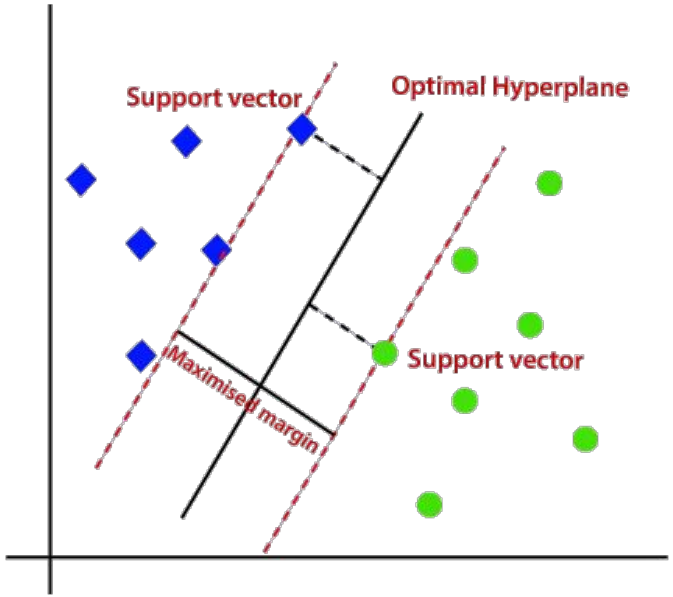


https://www.linkedin.com/pulse/random-forest-algorithm-interactive-discussion-niraj-kumar/

Support Vector Machine (SVM)

SVMs find a hyperplane that best separates the data into classes

- Maximize margin between different classes



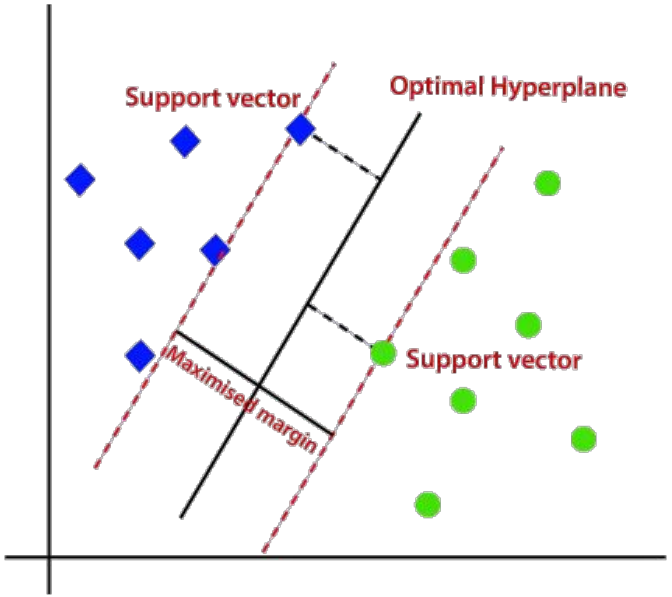
https://medium.com/@ambika199820/support-vector-machine-svm-algorithm-explained-c8facc10bcfa

SVM

Support Vector Machines find a hyperplane that best separates the data into classes

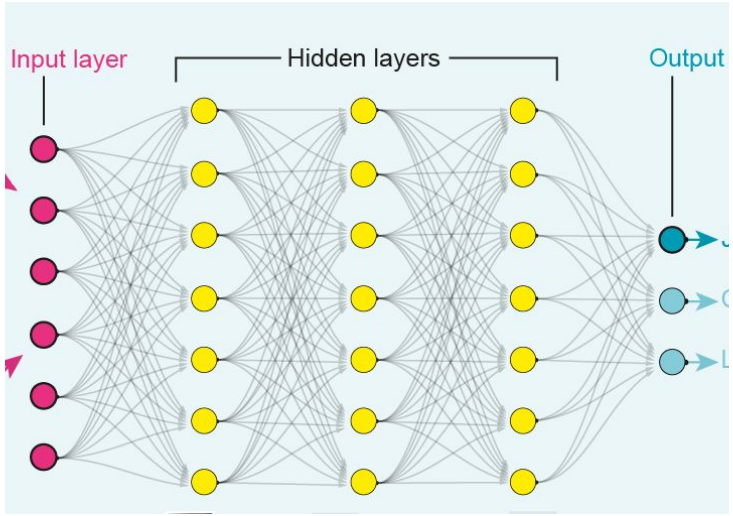
- Maximize margin between different classes

Pros	Cons
• Effectively handles high dimensional data • Can work well with small datasets • Can model non-linear decision boundaries	• Does not work for large data sets • Need to choose kernel • Limited to two-class problems • No probabilistic interpretation



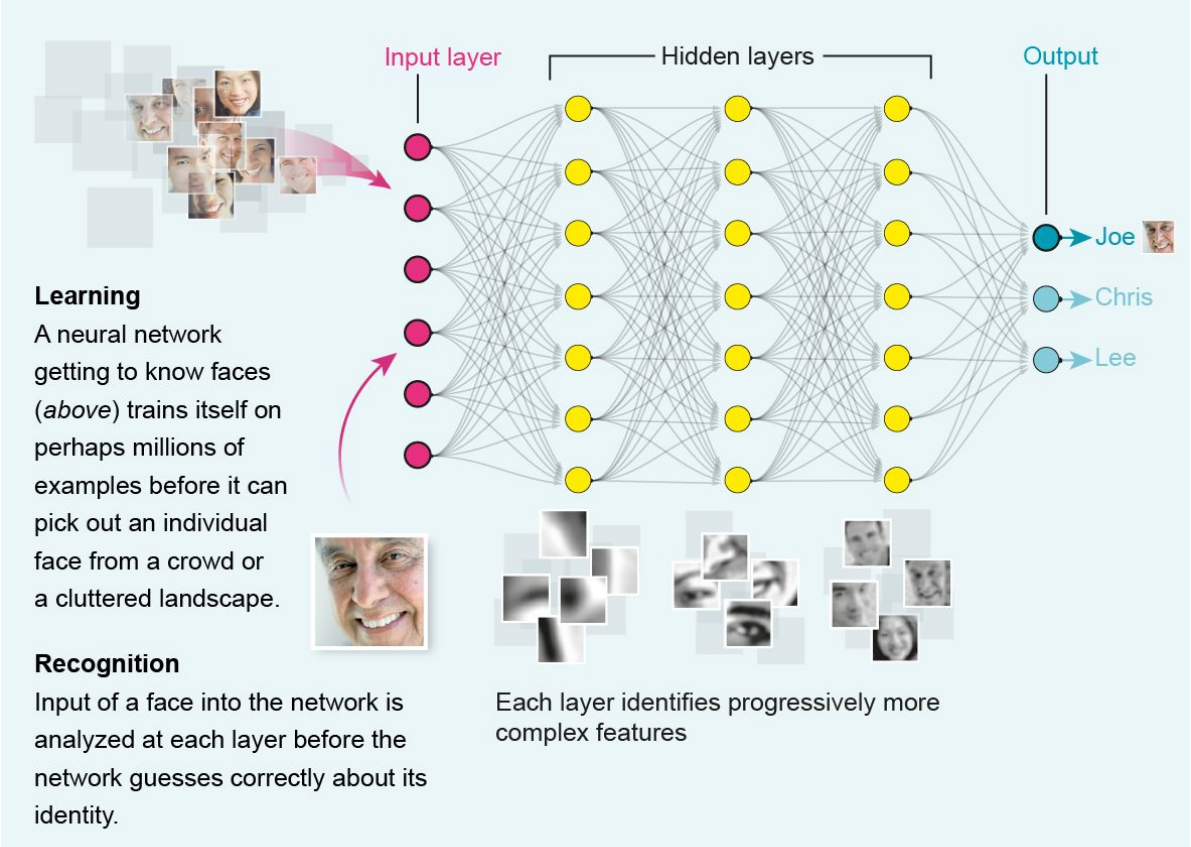
https://medium.com/@ambika199820/support-vector-machine-svm-algorithm-explained-c8facc10bcfa

Artificial Neural Networks



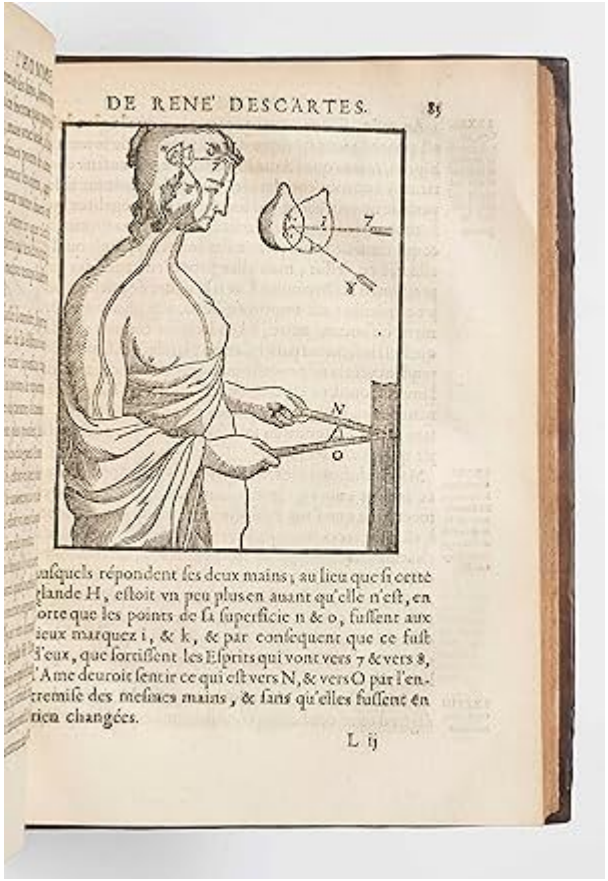
https://www.codingdojo.com/blog/artificial-neural-network-brain-behind-todays-smart-technology

Artificial Neural Networks

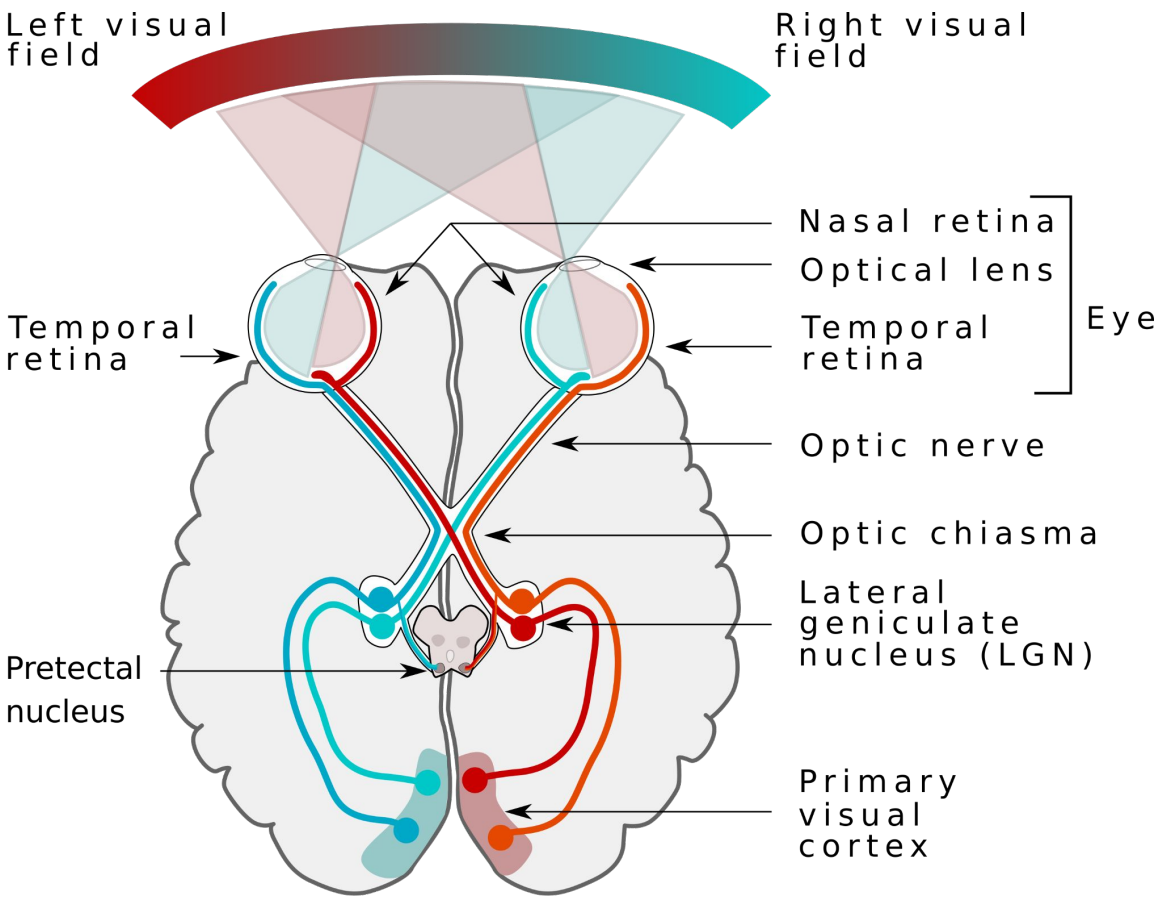


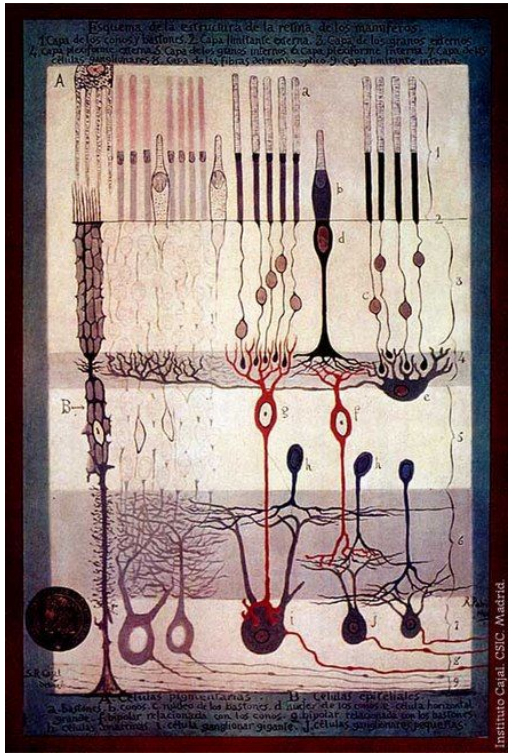
https://www.codingdojo.com/blog/artificial-neural-network-brain-behind-todays-smart-technology

Neural Circuitry



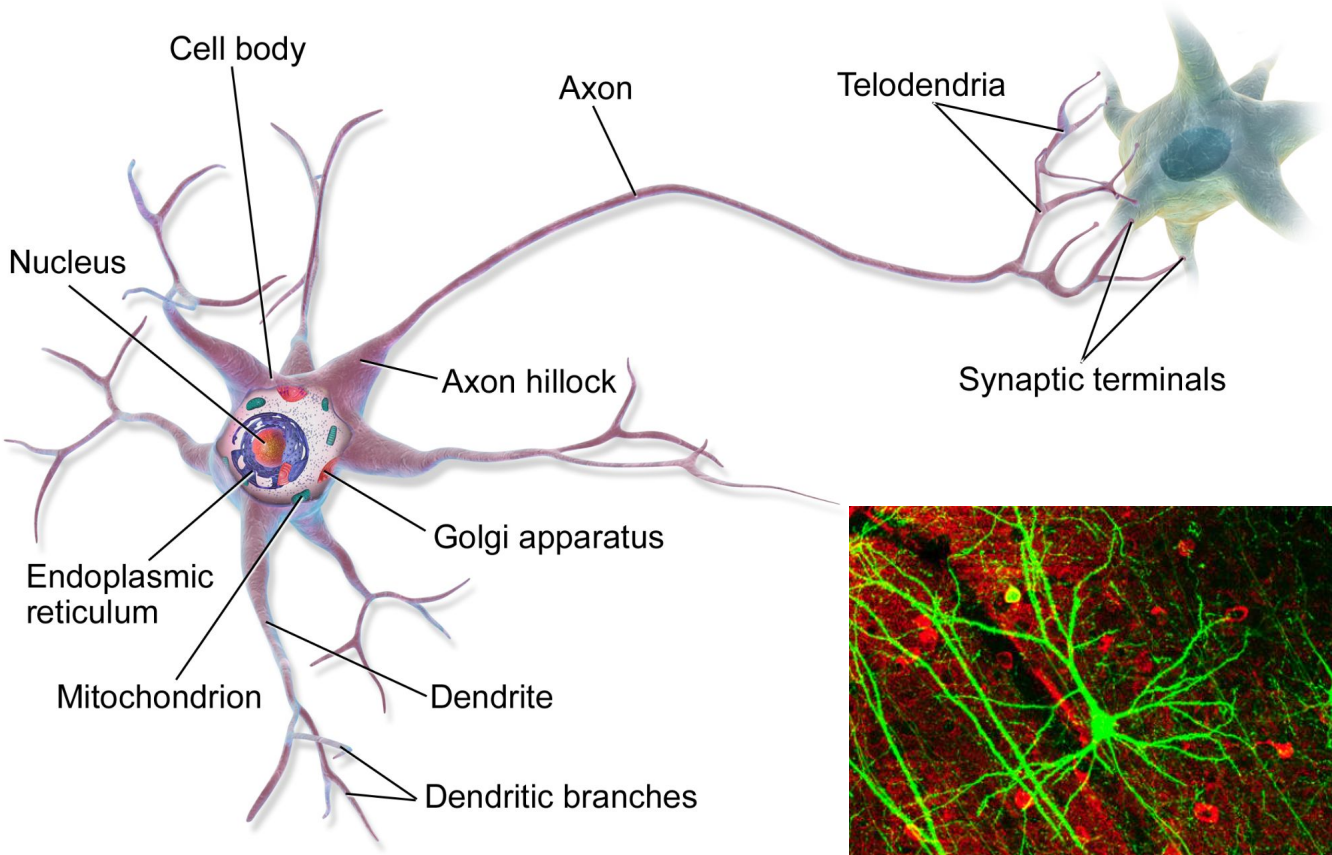
Human Visual Cortex

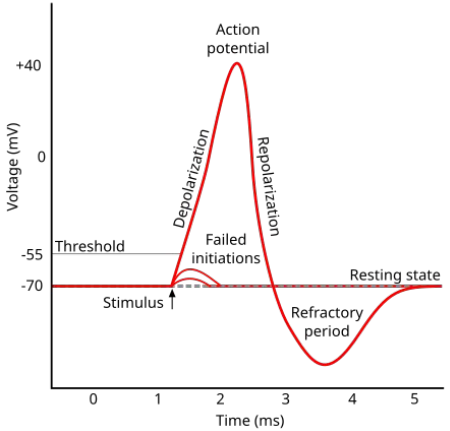


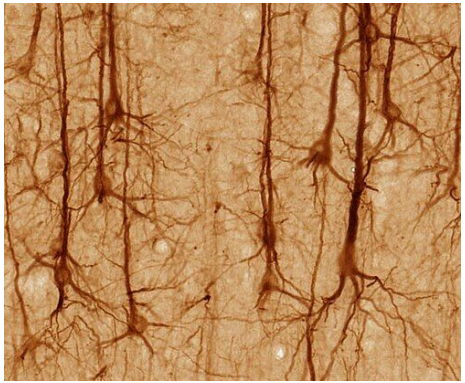


Structure of the Mammalian Retina
S. Ramón y Cajal (1900)

Neurons


Cell body
Axon
Telodendria
Nucleus
Axon hillock
Synaptic terminals
Golgi apparatus
Endoplasmic reticulum
Mitochondrion
Dendrite
Dendritic branches


Action potential
+40
0
Voltage (mV)
Depolarization
Repolarization
Threshold
-55
Failed initiations
Resting state
-70
Stimulus
Refractory period
0
1
2
3
4
5
Time (ms)



Vision & Pattern Recognition

Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex

D. H. HUBEL AND T. N. WIESEL • *Neurophysiology Laboratory, Department of Pharmacology Harvard Medical School, Boston, Massachusetts*

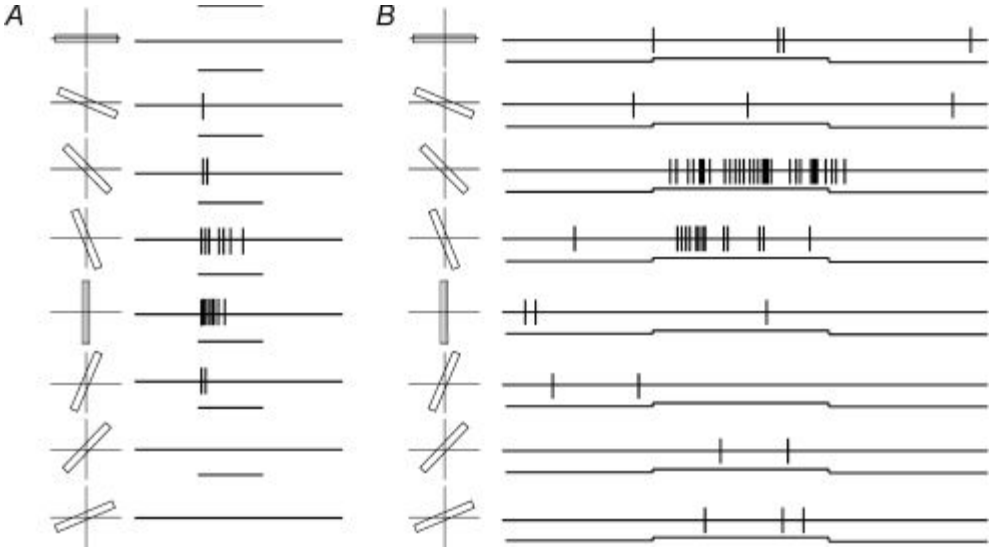
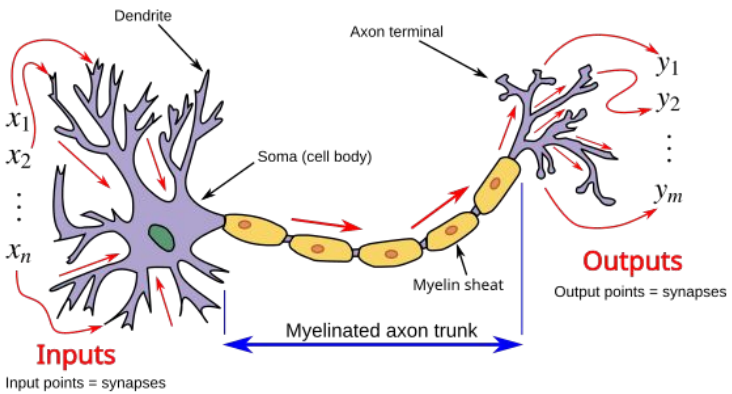


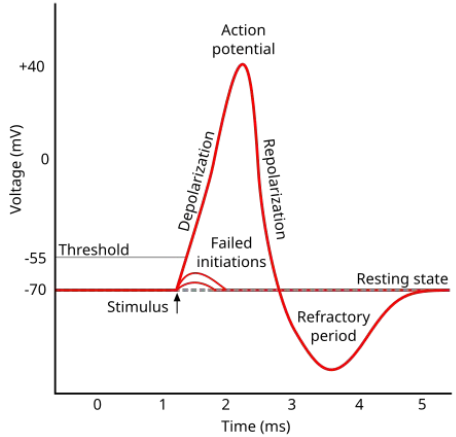
Table 1. Simple Cortical Fields

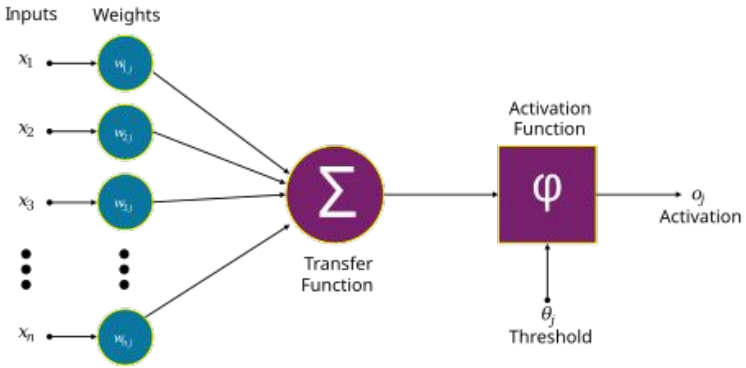
	Text-fig.	*No. of cells*
(*a*) Narrow concentrated centres		
(i) Symmetrical flanks		
Excitatory centres	*2C*	23
Inhibitory centres	*2D*	17
(ii) Asymmetrical flanks		
Excitatory centres	—	28
Inhibitory centres	*2E*	10
(*b*) Large centres; concentrated flanks	*2F*	21
(*c*) One excitatory region and one inhibitory	*2G*	17
(*d*) Uncategorized	—	117
Total number of simple fields		233

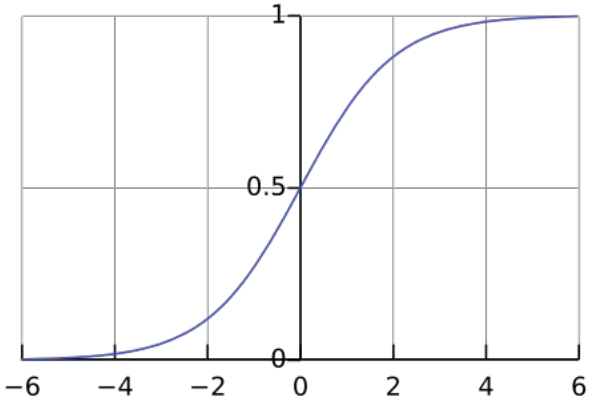
Example simple cells in the striate cortex of anaesthetized paralysed cat (A) and the awake fixating monkey (B). The comparison illustrates the similar response to oriented slits of light in the anaesthetized, paralysed cats and awake behaving monkeys. The example from the awake monkey shows the same qualitative orientation tuning but a slightly higher background rate that from the anaesthetized cat. Traced from Fig. 3A of Hubel & Wiesel (1959) and from Fig. 5 of (Wurtz, 1969c).

Biological vs Artificial Neurons



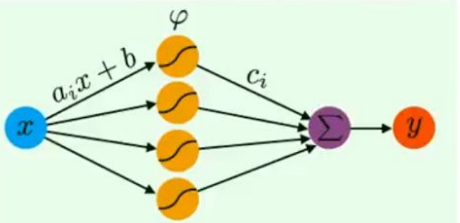






The non-linear activation function (sigmoid) is needed to model non-linear relationships

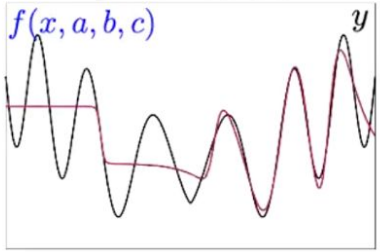
ANNs are “Universal Approximators”



1 hidden layer perceptron:

$$y \approx f(x,a,b,c) \stackrel{\text{def.}}{=} \sum_{i=1}^{p} c_i \varphi(a_i x + b_i)$$

$p = 6$ neurons



$p = 20$ neurons

Approximation by Superpositions of a Sigmoidal Function*

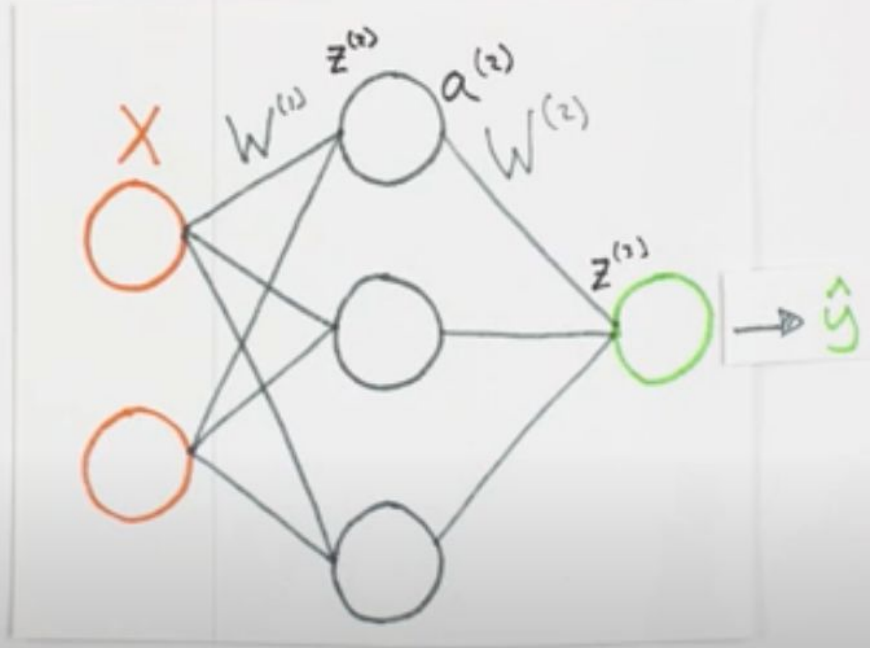
G. Cybenko†

Abstract. In this paper we demonstrate that finite linear combinations of compositions of a fixed, univariate function and a set of affine functionals can uniformly approximate any continuous function of n real variables with support in the unit hypercube; only mild conditions are imposed on the univariate function. Our results settle an open question about representability in the class of single hidden layer neural networks. In particular, we show that arbitrary decision regions can be arbitrarily well approximated by continuous feedforward neural networks with only a single internal, hidden layer and any continuous sigmoidal nonlinearity. The paper discusses approximation properties of other possible types of nonlinearities that might be implemented by artificial neural networks.

Key words. Neural networks, Approximation, Completeness.

Training with Backpropagation: Random initialization

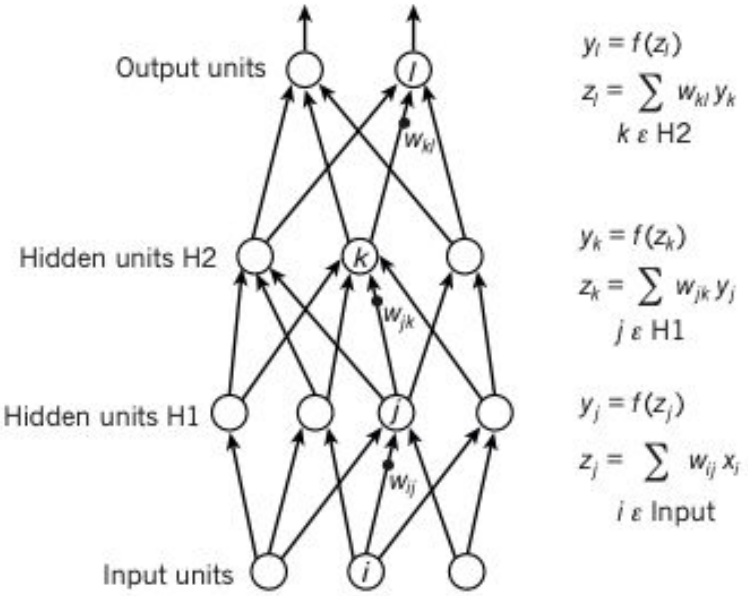
```python
class Neural_Network(object):
    def __init__(self):
        #Define Hyperparameters
        self.inputLayerSize = 2
        self.outputLayerSize = 1
        self.hiddenLayerSize = 3

        #Weights (Parameters)
        self.W1 = np.random.randn(self.inputLayerSize, \
                                  self.hiddenLayerSize)
        self.W2 = np.random.randn(self.hiddenLayerSize, \
                                  self.outputLayerSize)

    def forward(self, X):
        #Propagate inputs though network
        self.z2 = np.dot(X, self.W1)
        self.a2 = self.sigmoid(self.z2)
        self.z3 = np.dot(self.a2, self.W2)
        yHat = self.sigmoid(self.z3)
        return yHat

    def sigmoid(self, z):
        #Apply sigmoid activation function to scalar, vector, or matrix
        return 1/(1+np.exp(-z))
```



https://youtu.be/bxe2T-V8XRs?si=LwLxcyLgS4WFIyml

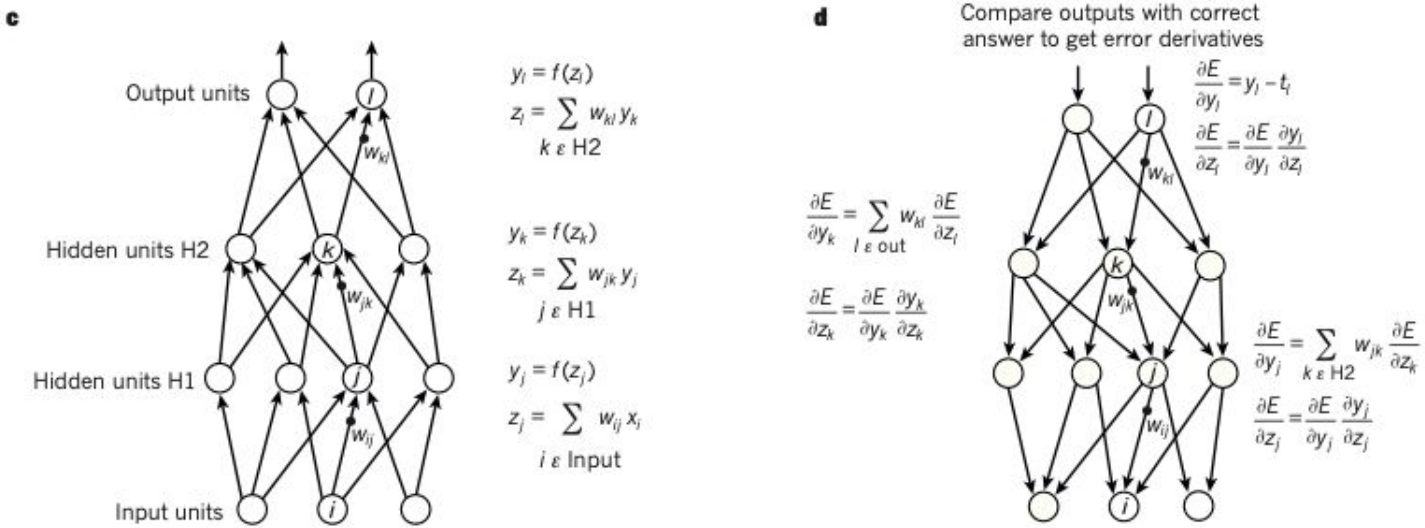
Training with Backpropagation: Forward evaluation

c



Deep Learning
LeCun, Bengio, Hinton (2020). Nature. doi:10.1038/nature14539

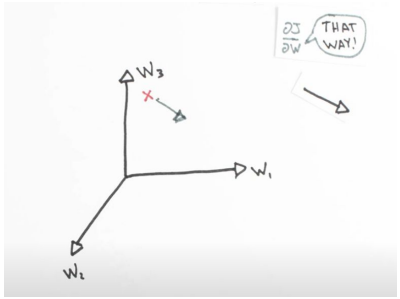
Training with Backpropagation: Backpropagation

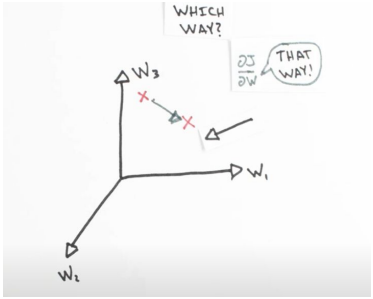


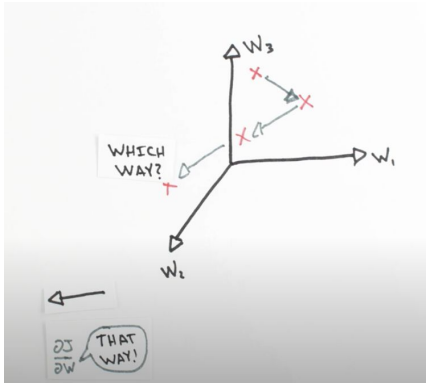
- Calculations along every edge, but the math is generally very basic
  - Addition, multiplication, exponentiation, etc
- Also easy to parallelize (especially with GPUs)

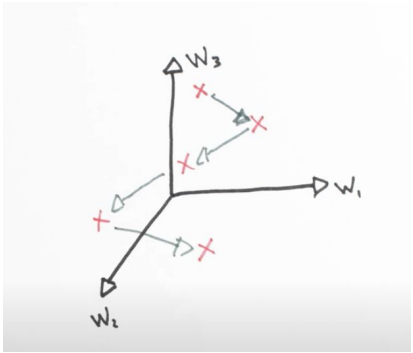
Deep Learning
LeCun, Bengio, Hinton (2020). Nature. doi:10.1038/nature14539

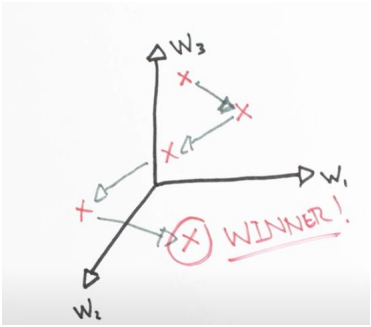
Training with Backpropagation: Gradient Descent
Minimize errors between predictions and true labels







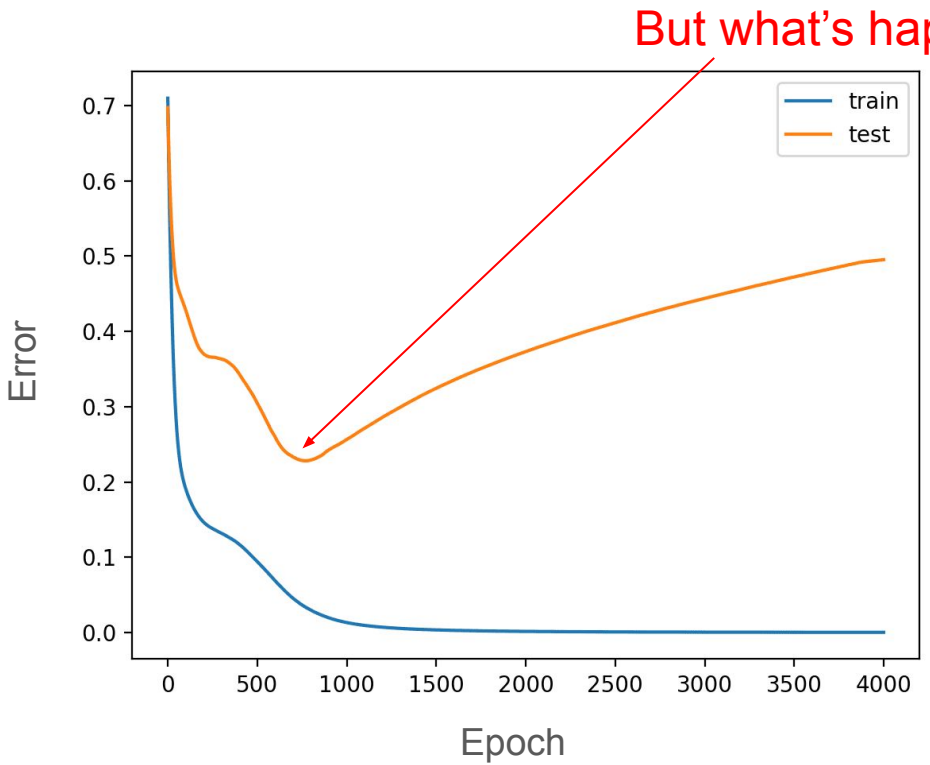




Ball rolling down a hill: start with a random initialization, then stochastic gradient descent by iteratively computing the partial derivatives of the functions

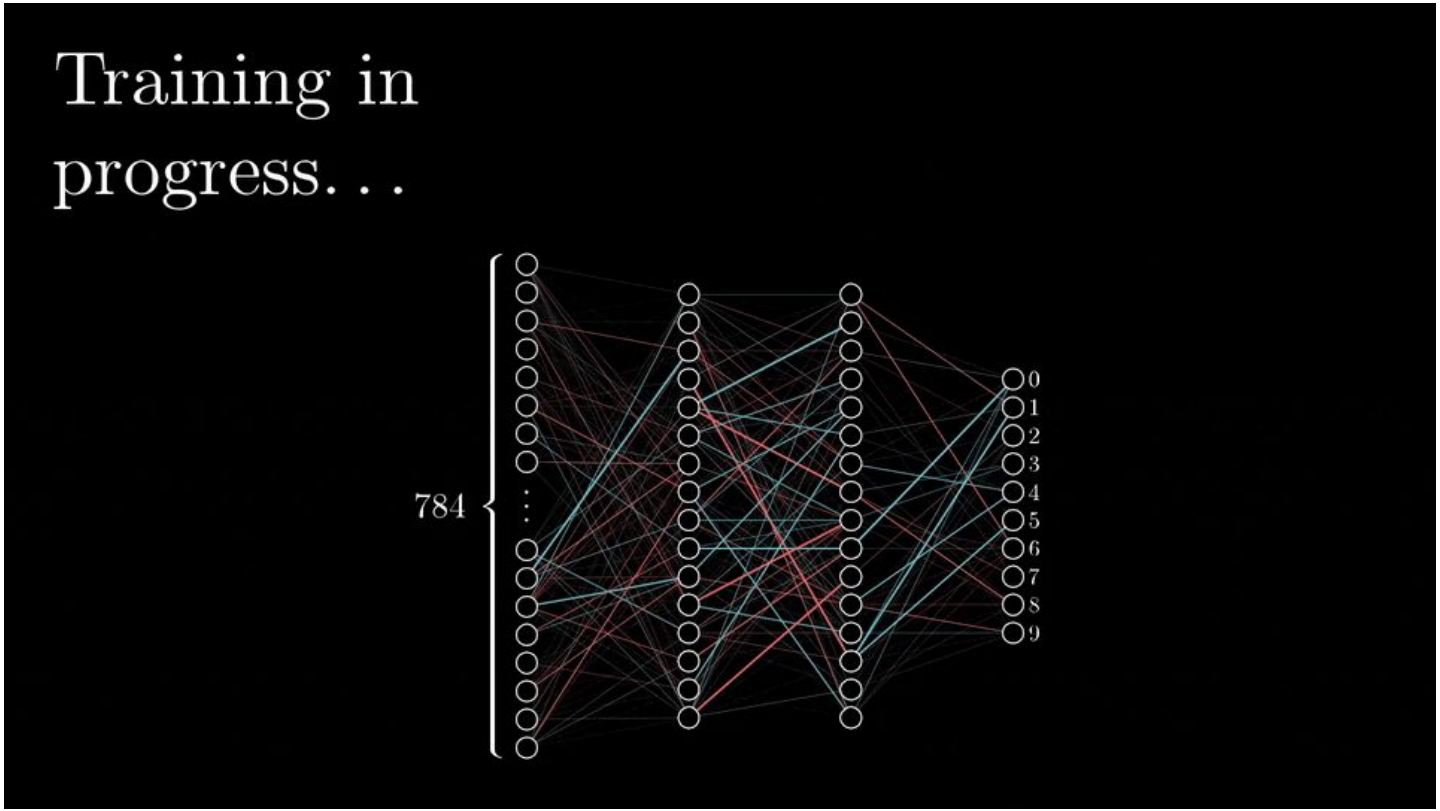
Training ANN through multiple epochs

Error goes down with more training epochs



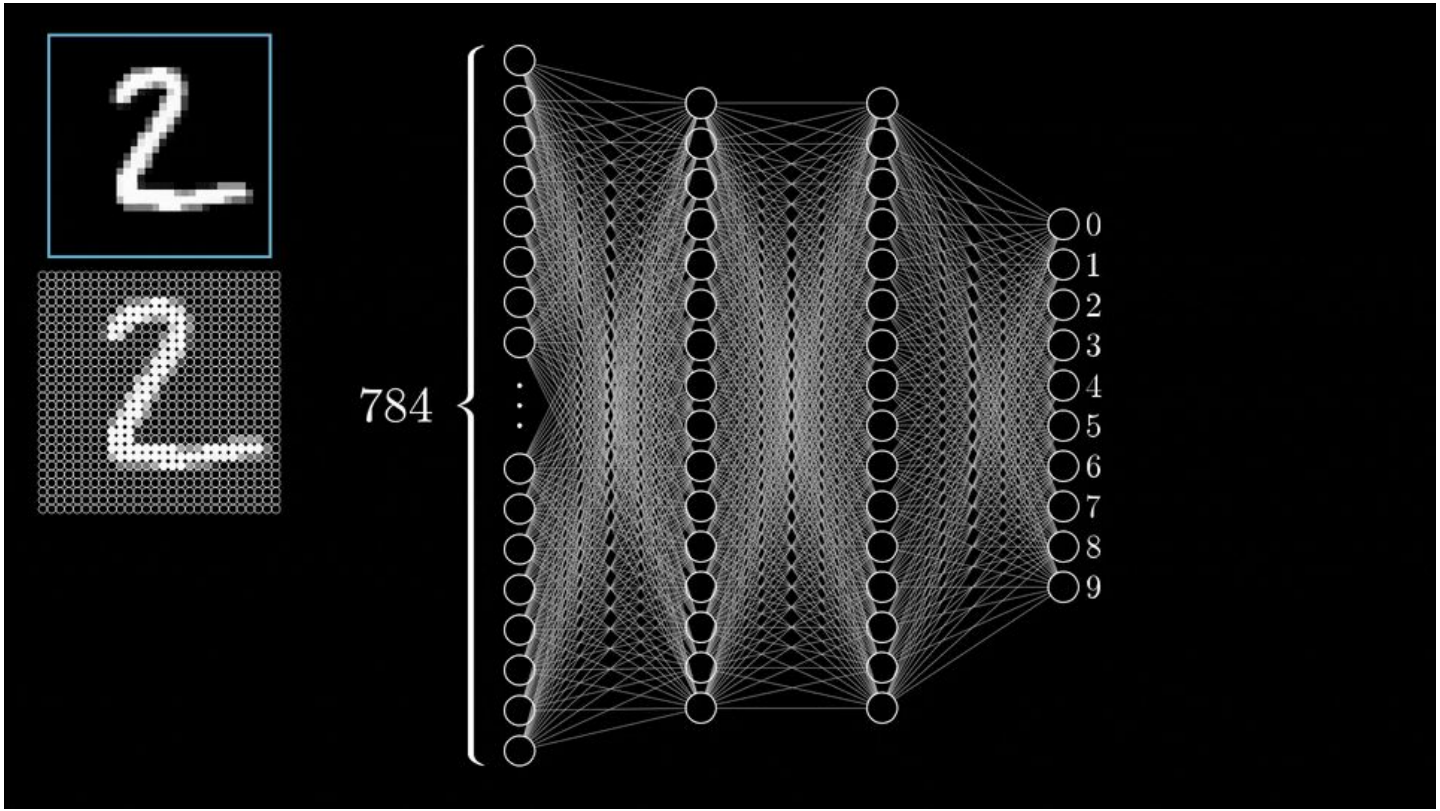
Training for too long can lead to overfitting!

Recap - Training



https://youtu.be/IHZwWFHWa-w

Recap - Evaluation

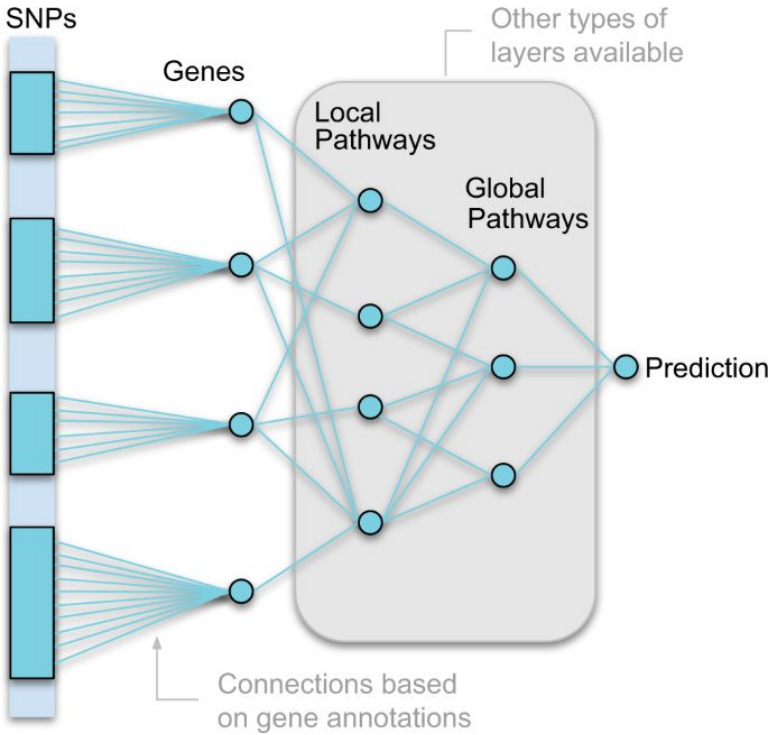


https://www.youtube.com/watch?v=Ilg3gGewQ5U

ANNs in Bioinformatics

GenNet: predicting phenotypes from genetic data

- Connections between layers defined through prior biological knowledge
- Basis for predicting many phenotypes
  - E.g. Schizophrenia, hair and eye color



van Hilten, A., Kushner, S.A., Kayser, M. *et al.* GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun Biol* 4, 1094 (2021).