

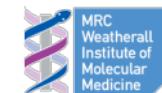


Deep Neural Networks in Genomics

Ron Schwessinger

NGS “Summer” School

27/10/2019



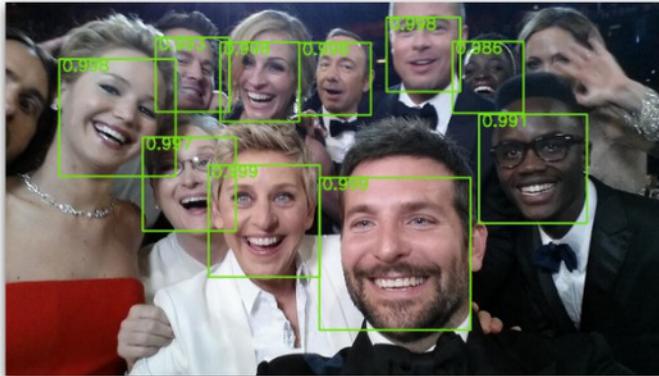
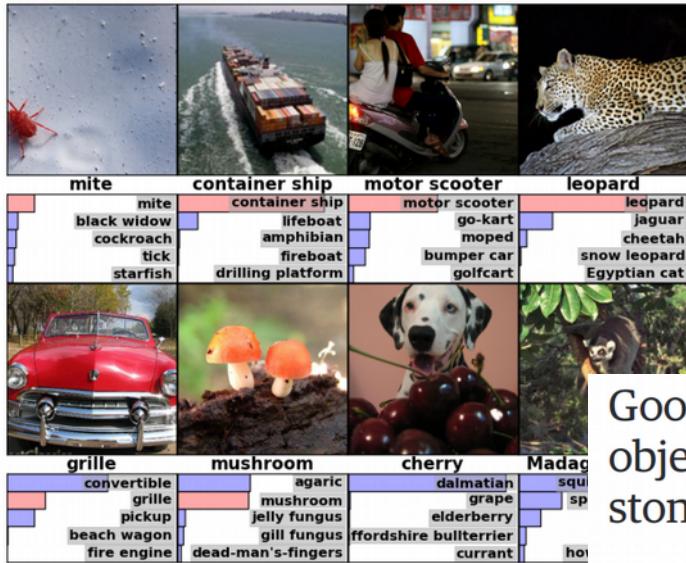
The MRC Weatherall Institute of Molecular Medicine is a strategic alliance between the Medical Research Council and the University of Oxford

Outline

- 1) Introduction to Deep Neural Networks
 - 1) Neural Networks oversimplified
 - 2) How do we learn?
 - 3) Training Process
- 2) Introduction to Convolutional Neural Networks
 - 1) Convolutions for Image Analysis
 - 2) Convolutions for DNA
- 3) Convolutional Neural Networks in Genomics
 - 1) Chromatin Feature Networks
 - 2) Utility, Strategies and Interpretation
 - 3) More Examples, More Architectures
- 4) Basic Practical Aspects
 - 1) Overfitting / Underfitting
 - 2) Train, Validation & Test Sets (& Cross-Validation)
 - 3) Learning Rate & Optimizers

1) Introduction to Deep Neural Networks

Deep Learning Out There

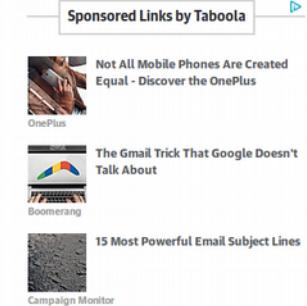


Google lands patent for automatic object recognition in videos, leaves no stone untagged

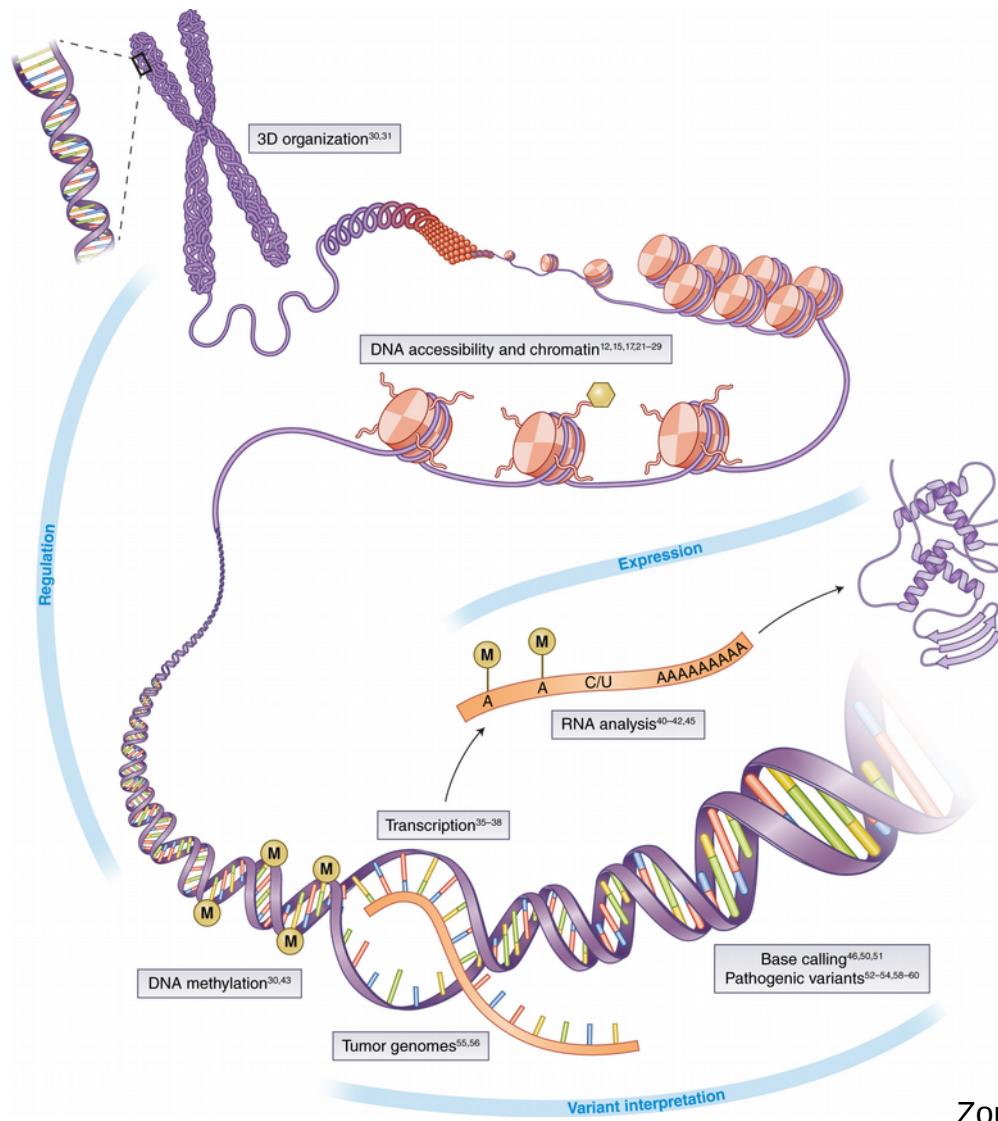
Jon Fingas , @jonfingas
08.28.12

0
Shares

f t g



Deep Learning in Genomics



Zou et al. Nature Genetics 2019

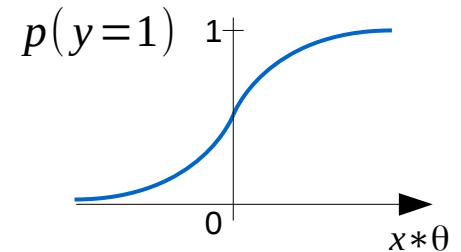
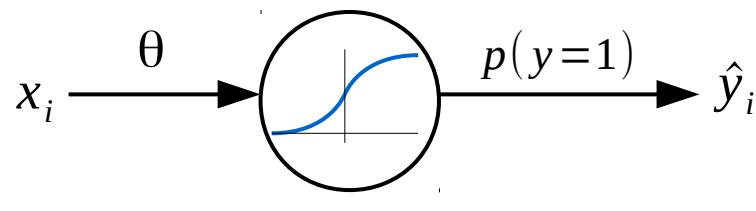
Neural Networks oversimplified

rain probability from weather forecast	tomorrow is a rainy day	y truth: is tomorrow a rainy day
x	y	\hat{y} prediction: is tomorrow a rainy day?
0.6	1	1
0.4	1	
0.2	0	
0.8	1	

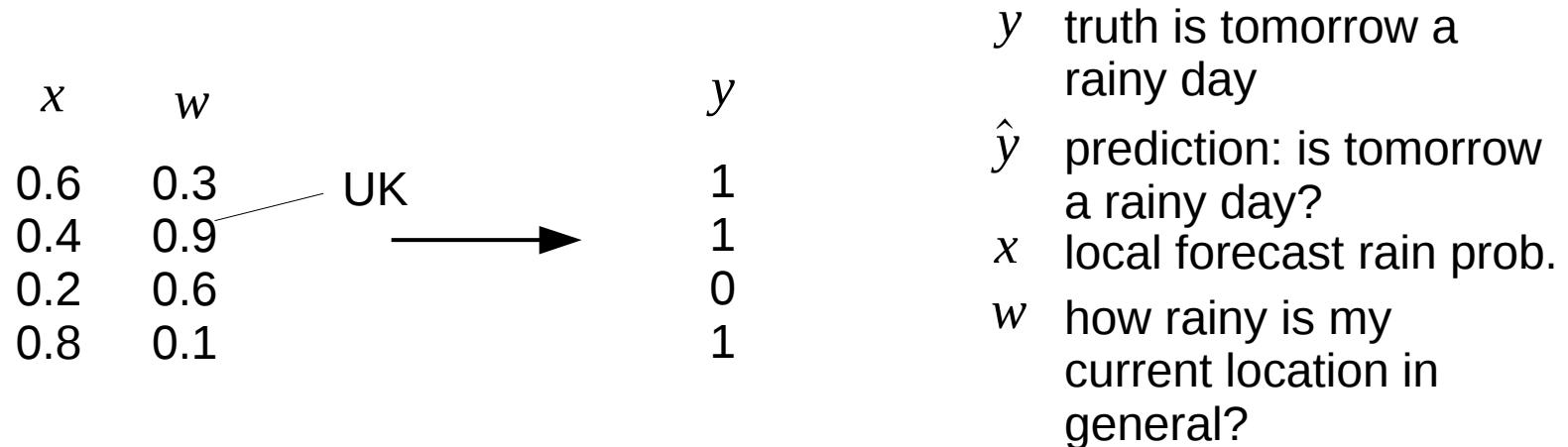
Neural Networks oversimplified

rain probability from weather forecast	tomorrow is a rainy day	y truth: is tomorrow a rainy day
x	y	\hat{y} prediction: is tomorrow a rainy day?
0.6	1	
0.4	1	
0.2	0	
0.8	1	

Single Neuron

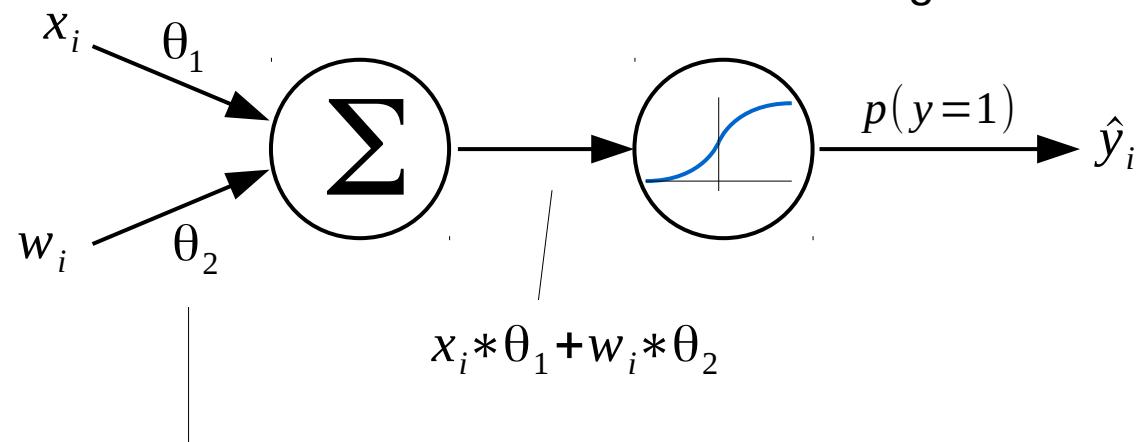


Neural Networks oversimplified



Neural Networks oversimplified

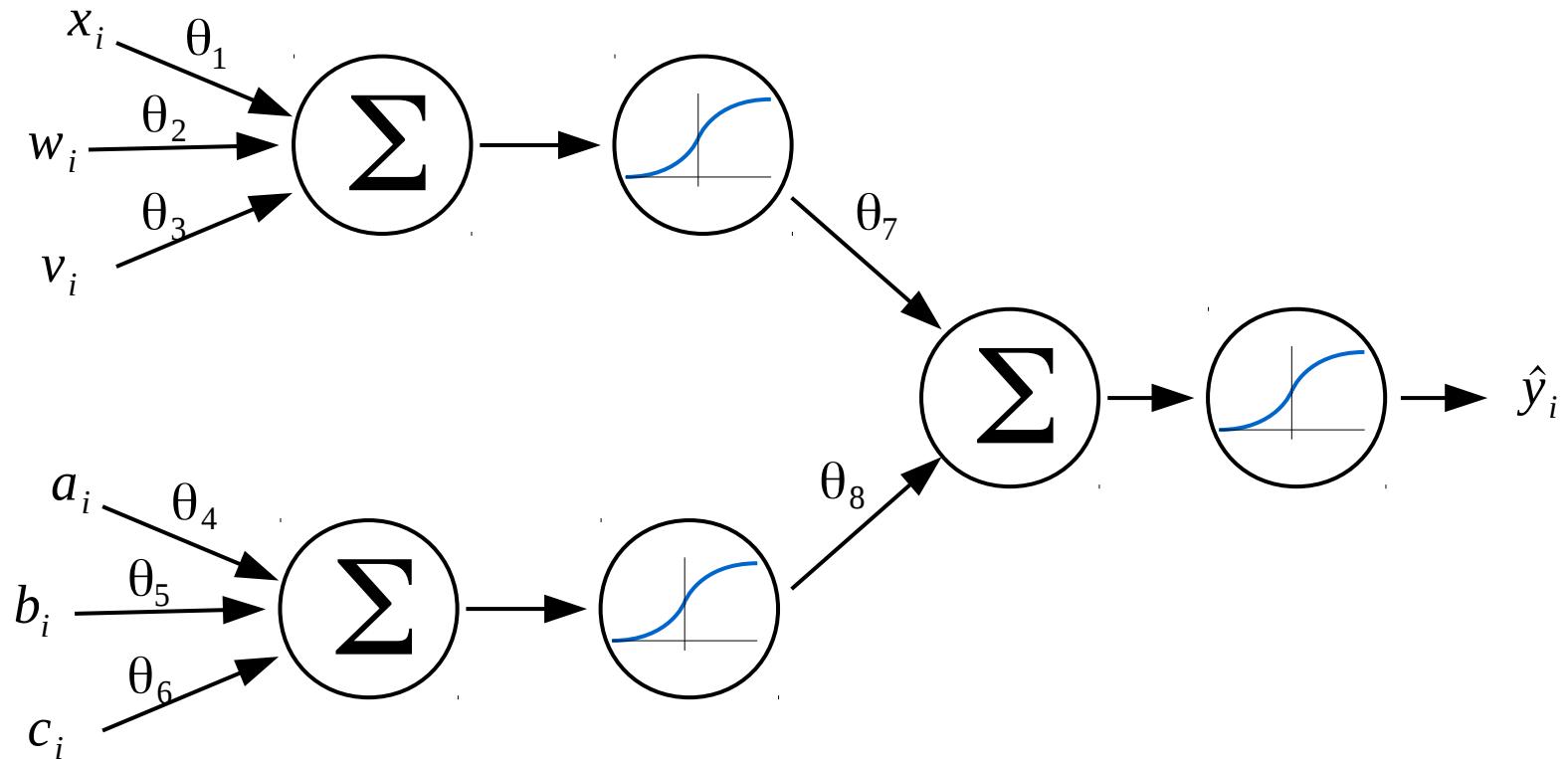
x	w		y	\hat{y}	
0.6	0.3	UK	1	1	truth is tomorrow a rainy day
0.4	0.9		1	1	prediction: is tomorrow a rainy day?
0.2	0.6		0		x local forecast rain prob.
0.8	0.1		1		w how rainy is my current location in general?



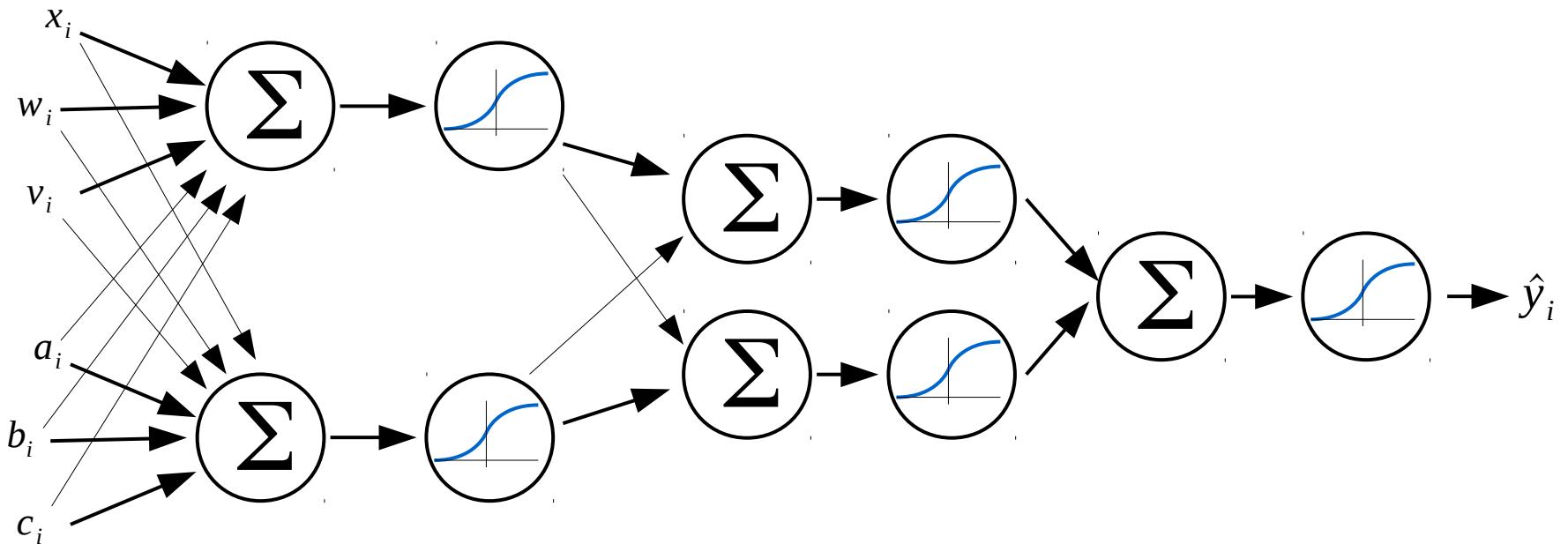
parameters

aim: tweak the parameters to get the best predictions

Neural Networks oversimplified

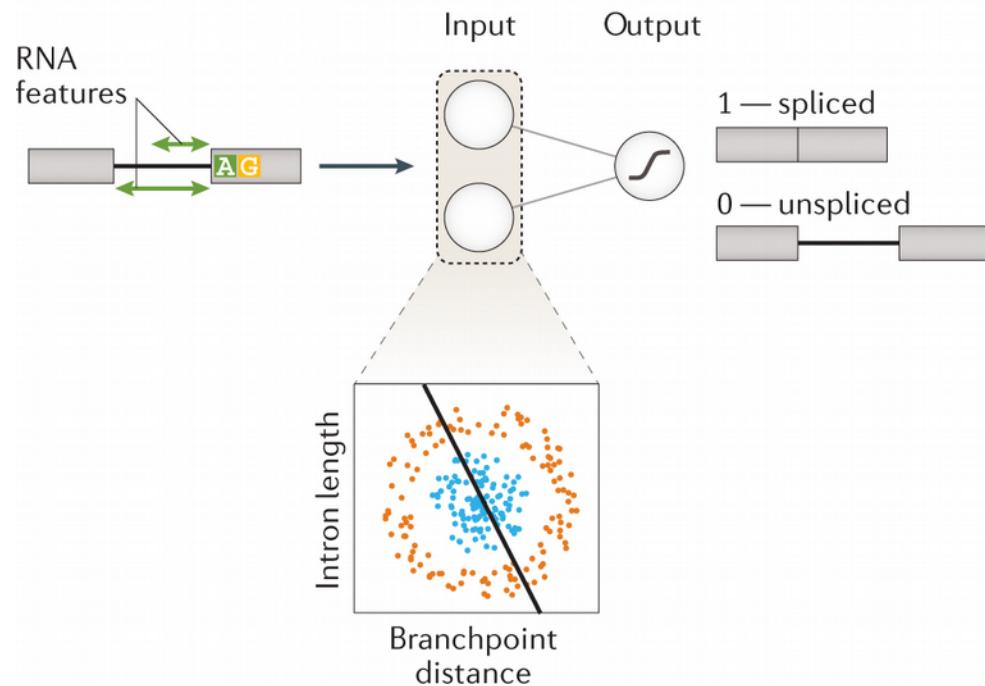


Neural Networks oversimplified

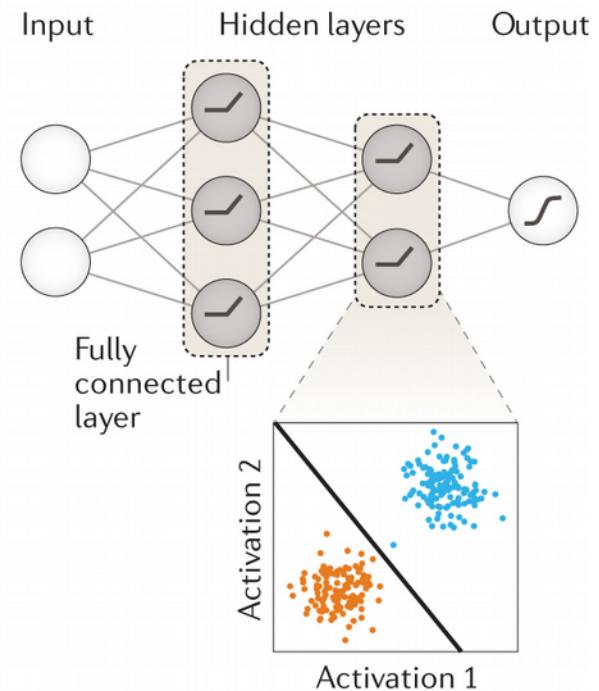


The Power of Non Linear Activations

a Single-layer neural network (logistic regression)



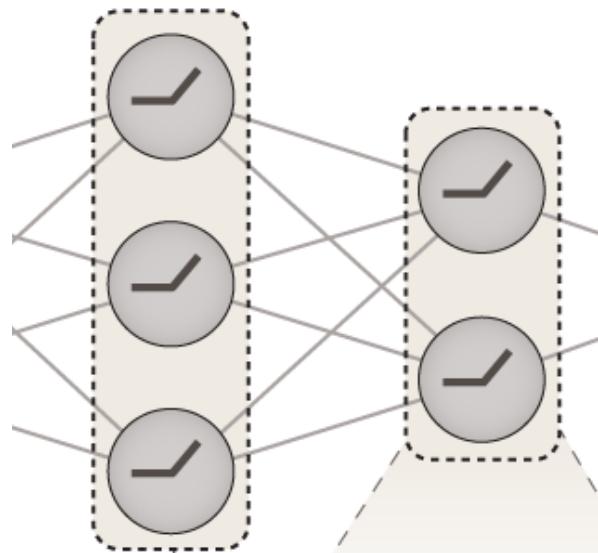
b Multilayer neural network



Gökçen et al. Nature Reviews Genetics 2019

Non Linear Activations

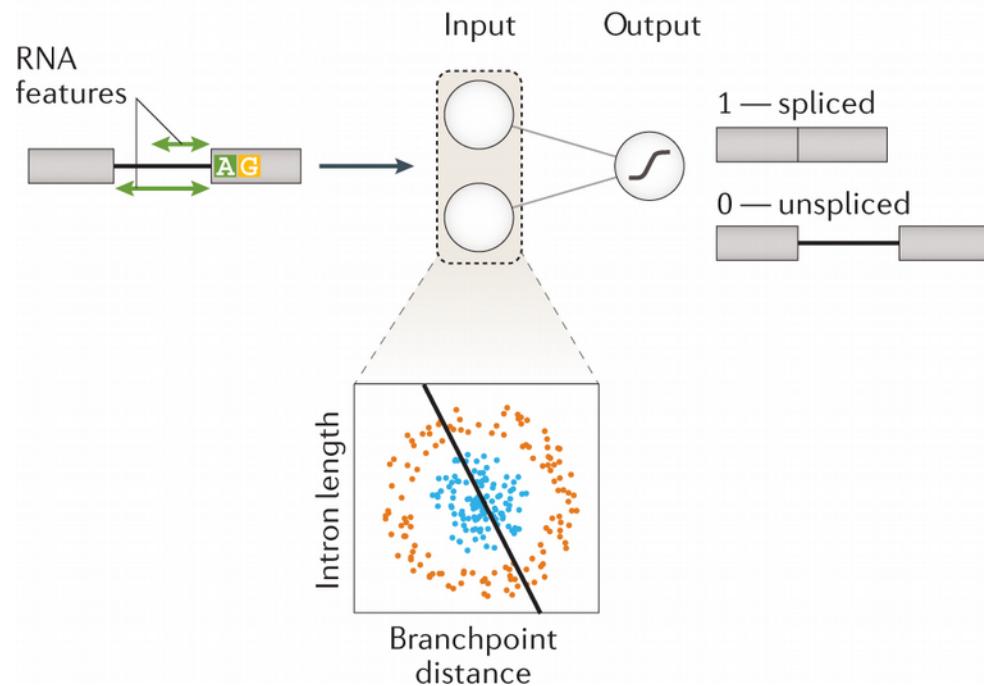
Hidden layers



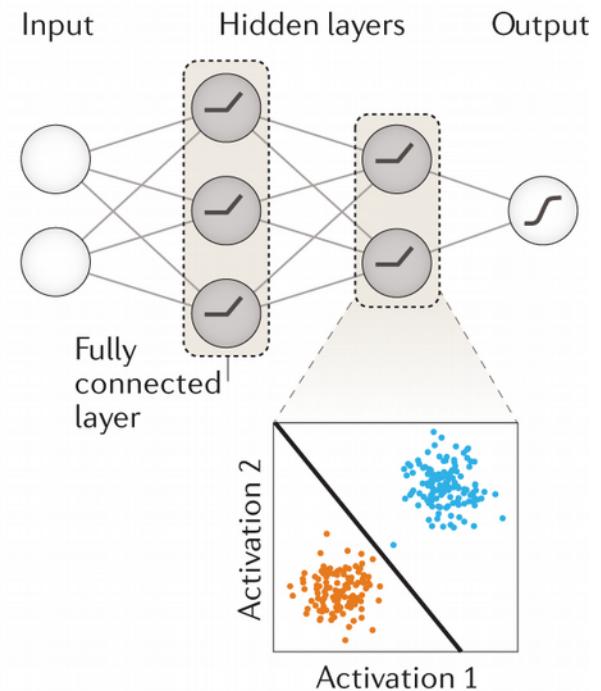
Gökcen et al. Nature Reviews Genetics 2019

The Power of Non Linear Activations

a Single-layer neural network (logistic regression)



b Multilayer neural network



Gökçen et al. Nature Reviews Genetics 2019

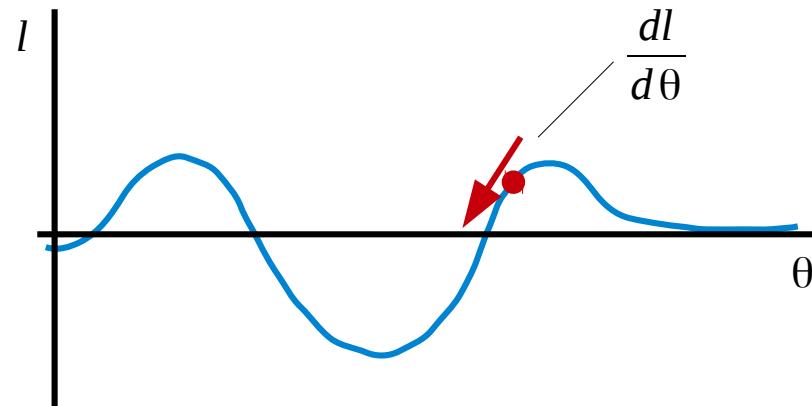
How do we learn?

Optimise **parameters** θ :

loss function $l \rightarrow$ how bad do we perform? \rightarrow minimize l

calculate **gradients!**

Optimization



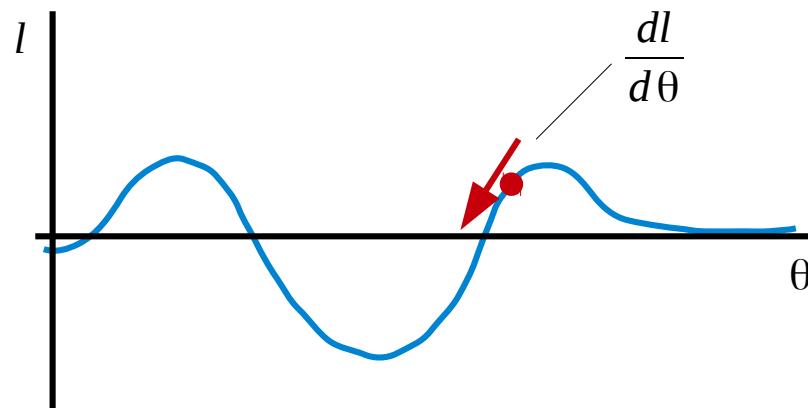
How do we learn?

Optimise **parameters** θ :

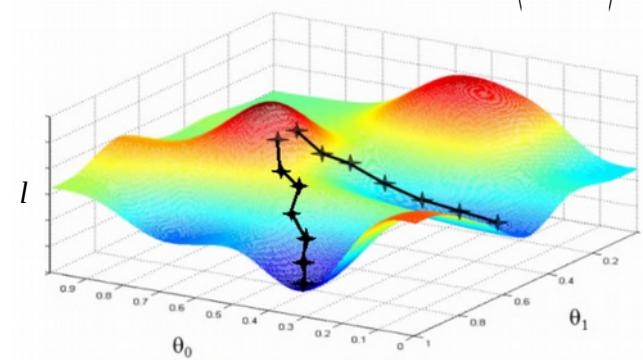
loss function $l \rightarrow$ how bad do we perform? \rightarrow minimize l

calculate **gradients!**

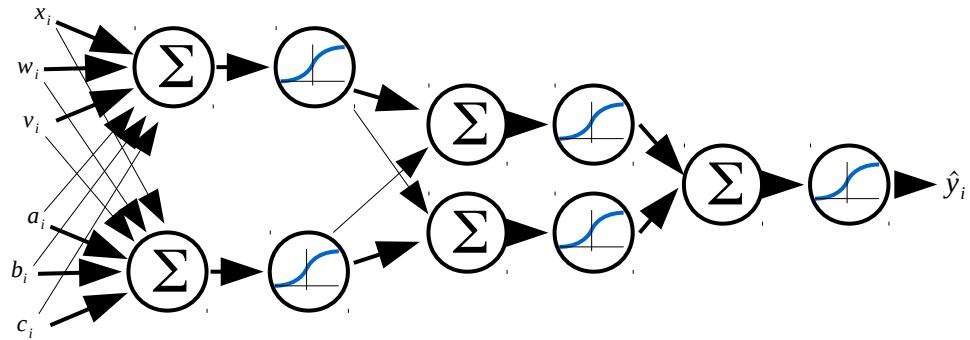
Optimization



$$\nabla J = \begin{pmatrix} \frac{\partial l}{\partial \theta_0} \\ \frac{\partial l}{\partial \theta_1} \end{pmatrix}$$



How do we learn?



- for simple NN \rightarrow gradients can be calculated by hand
- for larger architectures by computers
- but for very large/complex NNs \rightarrow normal CPU computation takes ages ...

Two major developments:

- better algorithms to calculate partial derivatives along large networks
- using graphics cards (GPU) computation

GPUs are great for massive parallelization of simply computational tasks

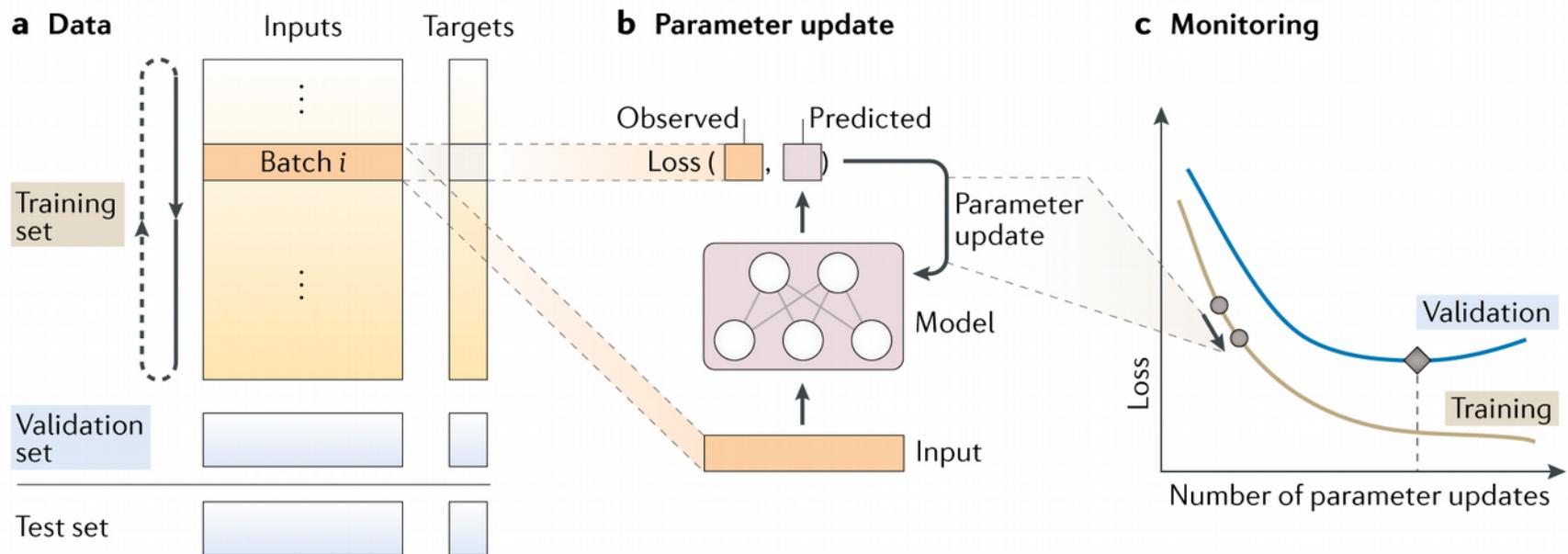
Training Process

Training:

- 1) Start with random parameters
- 2) Take a batch of inputs
- 3) Calculate predictions
- 4) Calculate your performance (loss)
- 5) Calculate gradient for every parameter
- 6) Adjust parameters along the gradient

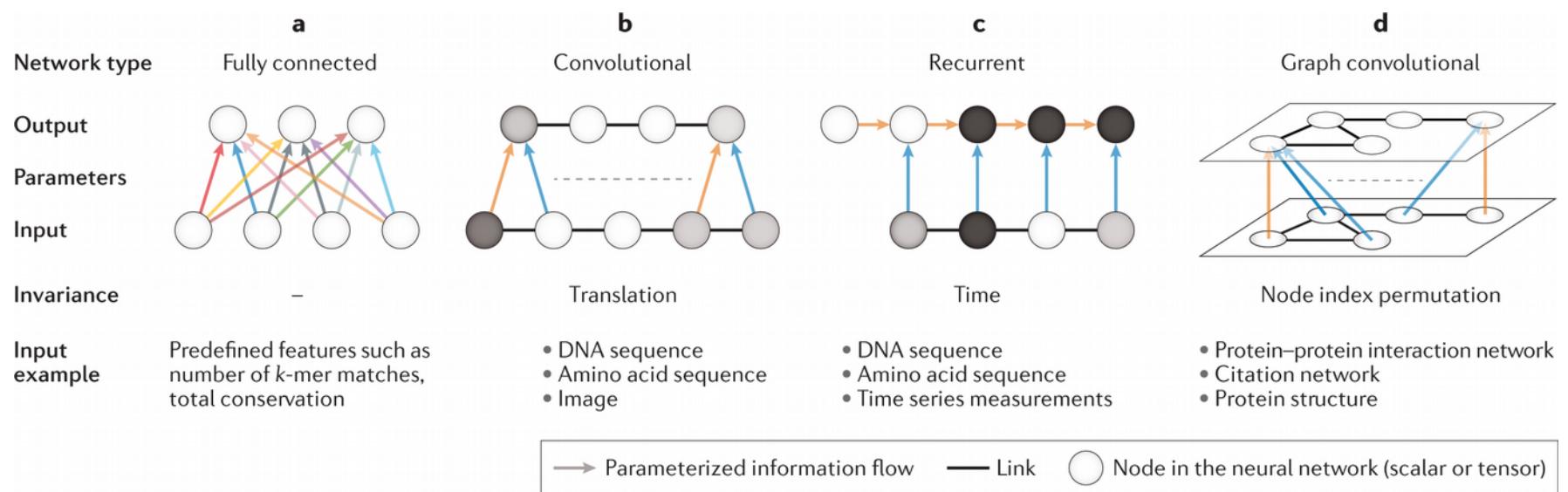
Until no improvement in performance noticeable...

Training Process



Gökçen et al. Nature Reviews Genetics 2019

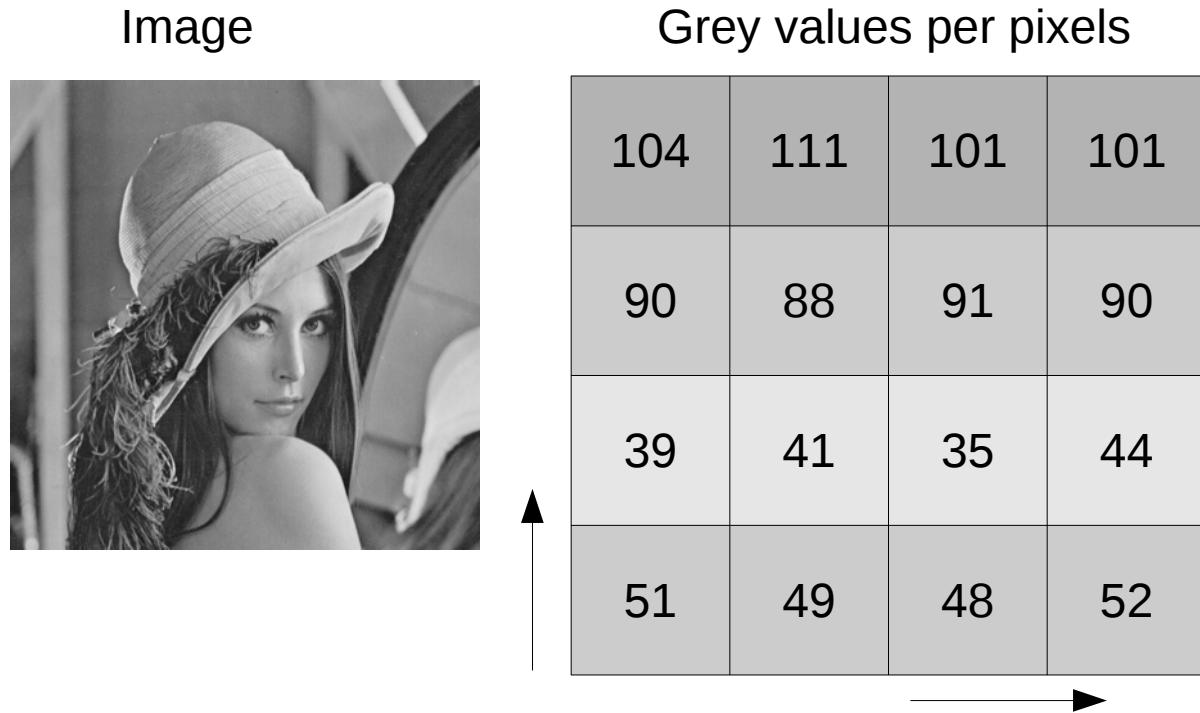
Overview of Network Types



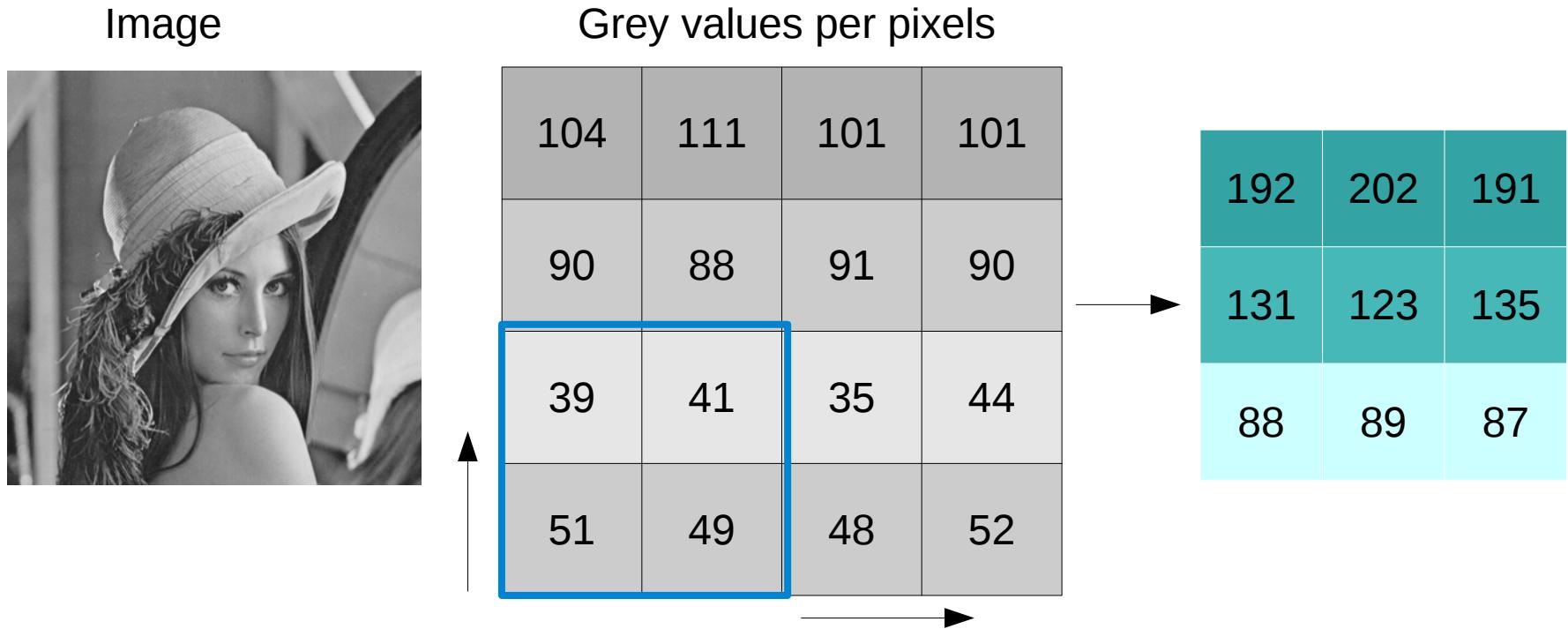
Gökçen et al. Nature Reviews Genetics 2019

2) Introduction to Convolutional Neural Networks

Convolutions for Image Analysis



Convolutions for Image Analysis



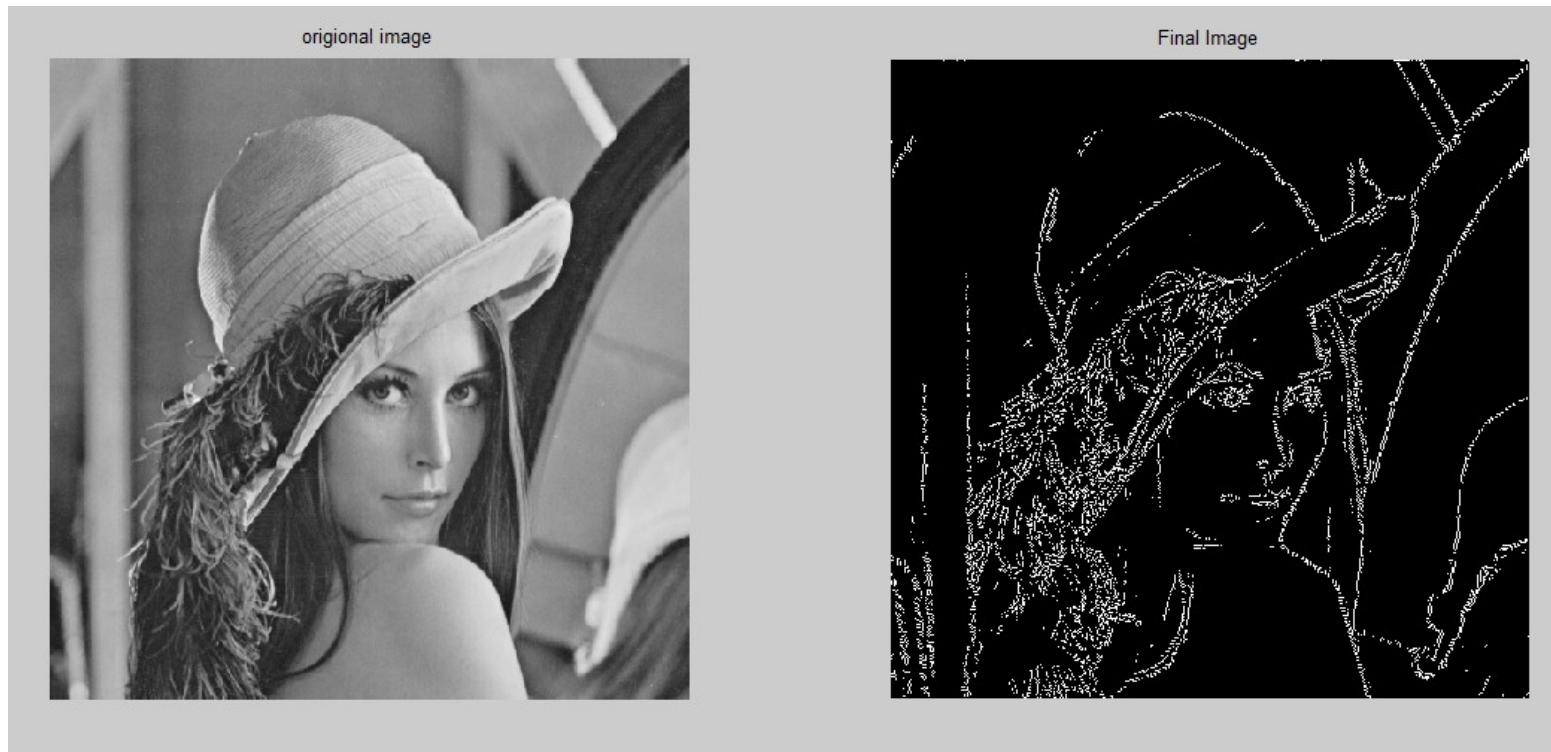
$$\sum \left(\begin{array}{|c|c|} \hline 1 & 0 \\ \hline 0 & 1 \\ \hline \end{array} * \begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array} \right) = x$$

$$1*a + 0*b + 0*c + 1*d = x$$

Filter = $\begin{matrix} \theta_1 & \theta_2 \\ \theta_3 & \theta_4 \end{matrix}$ → parameters to optimize

Convolutions for Image Analysis

Edge detection – Sobel operator



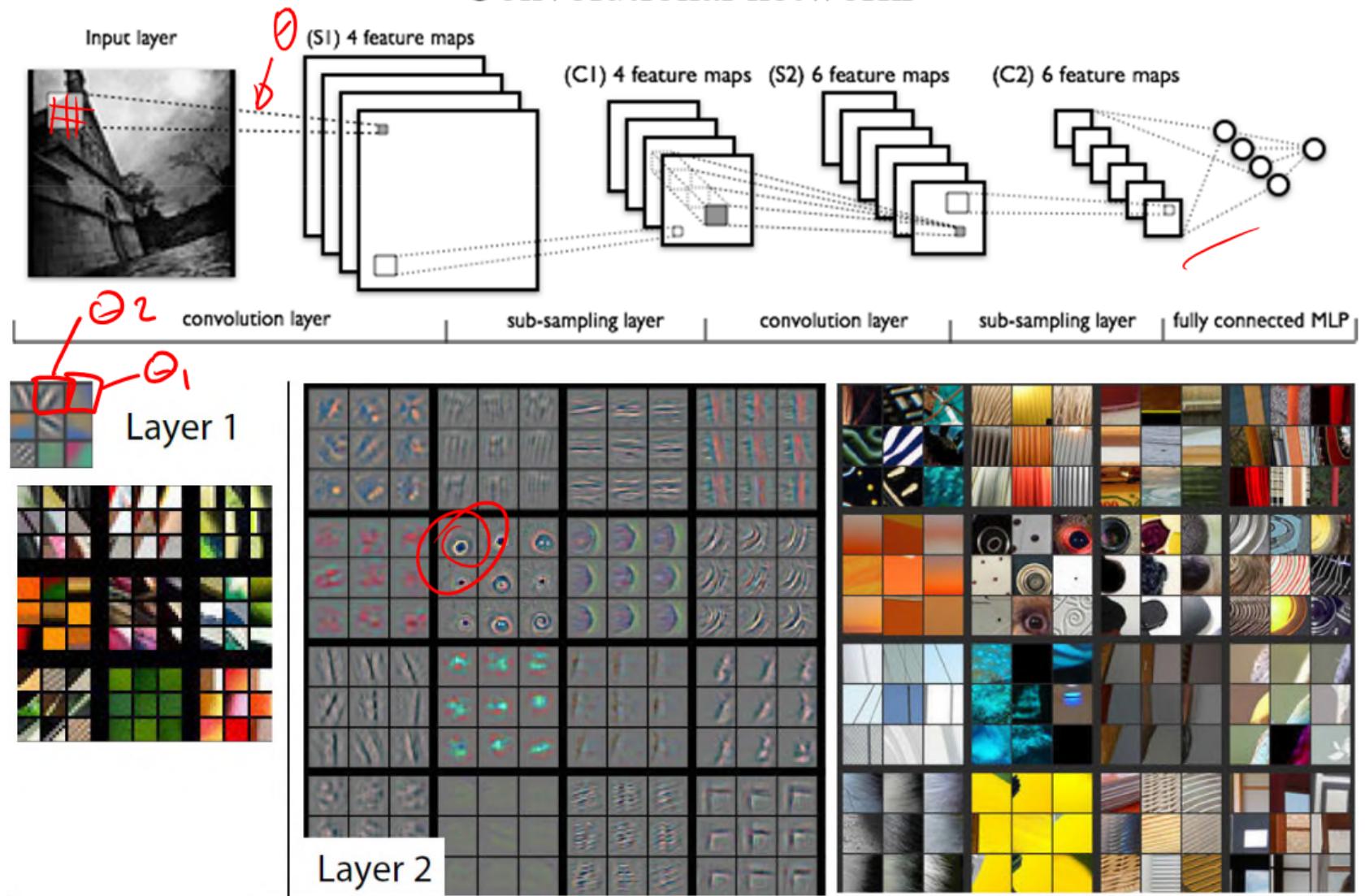
-1	0	1
-2	0	2
-1	0	1



1	2	1
0	0	0
-1	-2	-1

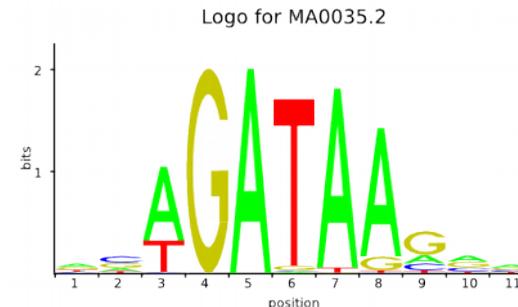


Convolutional networks



[Matthew Zeiler & Rob Fergus]

Convolutions for DNA



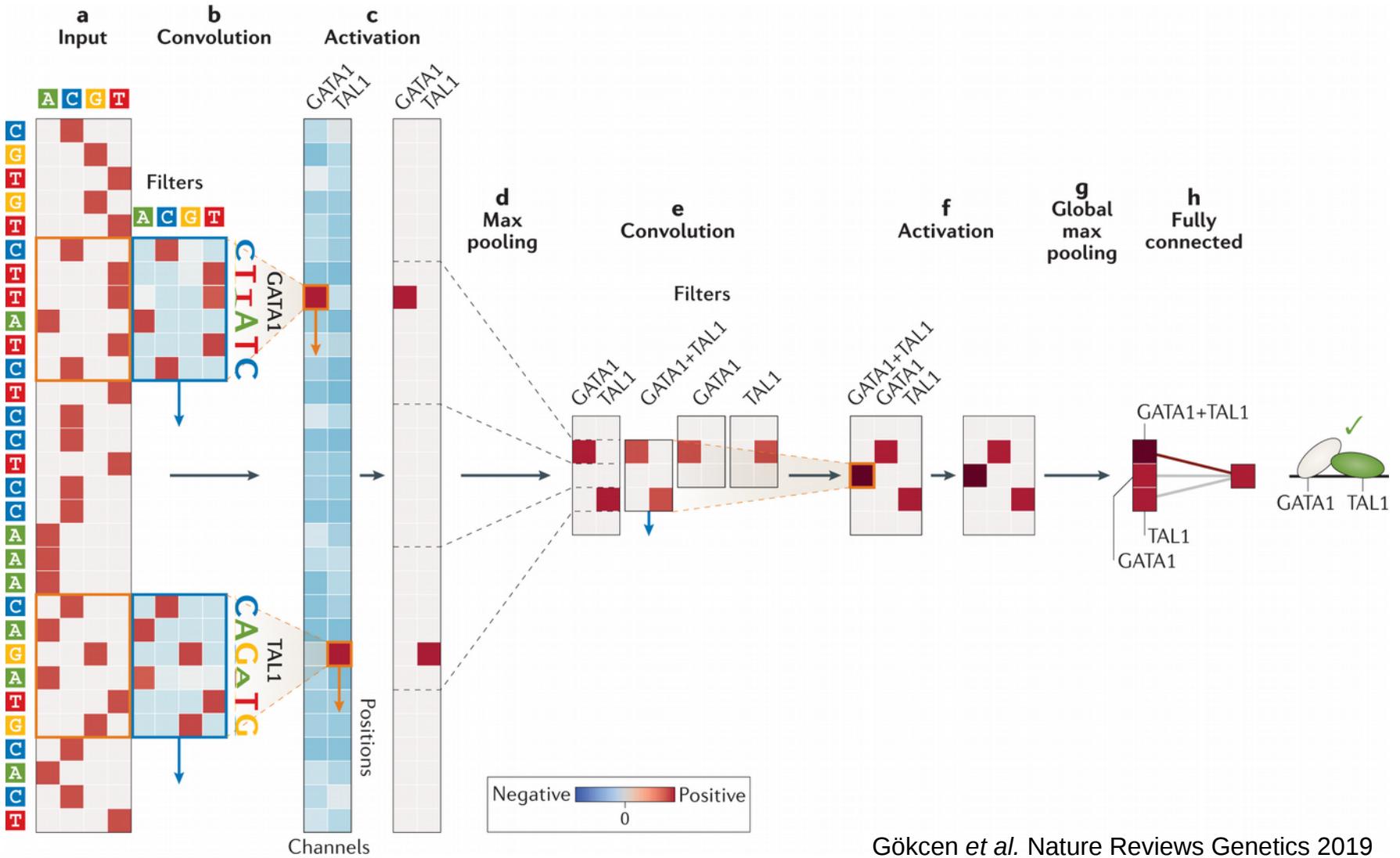
A	[1423	708	2782	0	4000	27	3887	3550	799	1432	1487]
C	[560	1633	31	0	0	29	0	4	681	897	829]
G	[1242	1235	10	4000	0	109	6	383	2296	1360	1099]
T	[775	424	1177	0	0	3835	107	63	224	311	585]

hot coded sequence

	A	C	G	A	T	A	A	G	T	A	G	G	G	C	T	A	T	T	C	C	...		
A	[1	0	1	0	1	0	1	1	0	0	1	0	1	0	0	0	0	1	0	0	0	...
C	[0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	...
G	[0	0	0	1	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	...
T	[0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	1	1	0	0	...

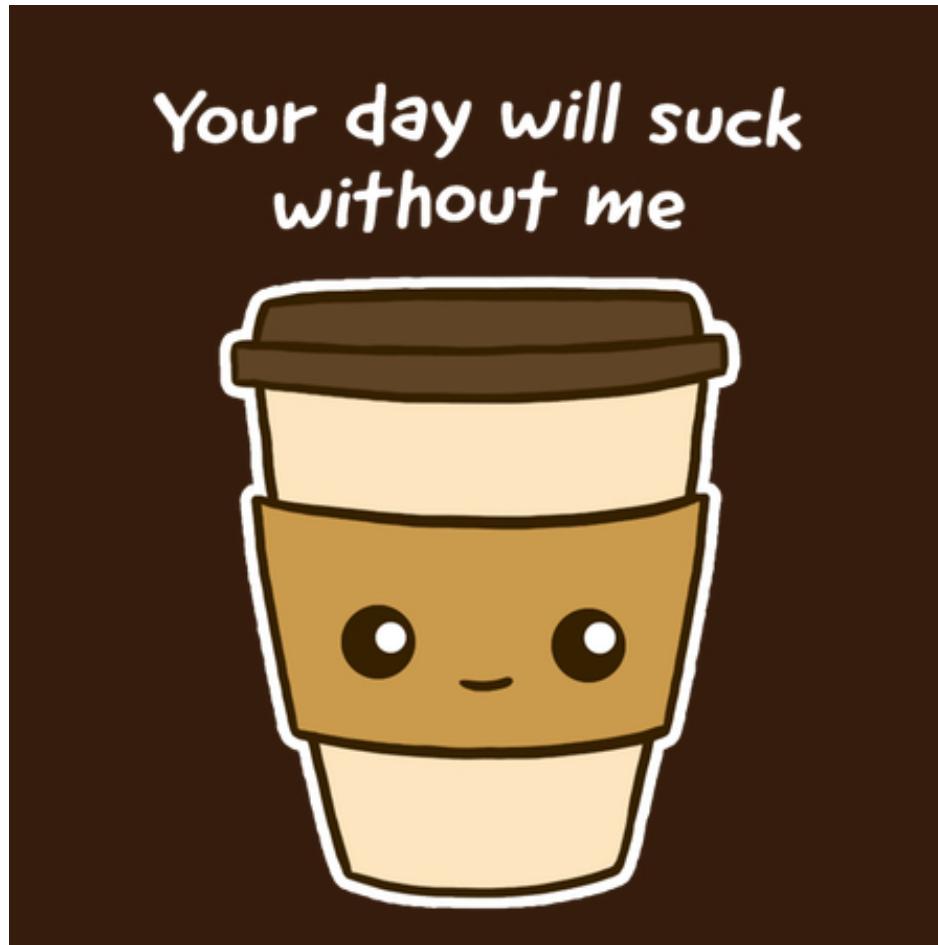
1005 506 501 501 40 55 501 ...

Convolutions of Convolutions for DNA



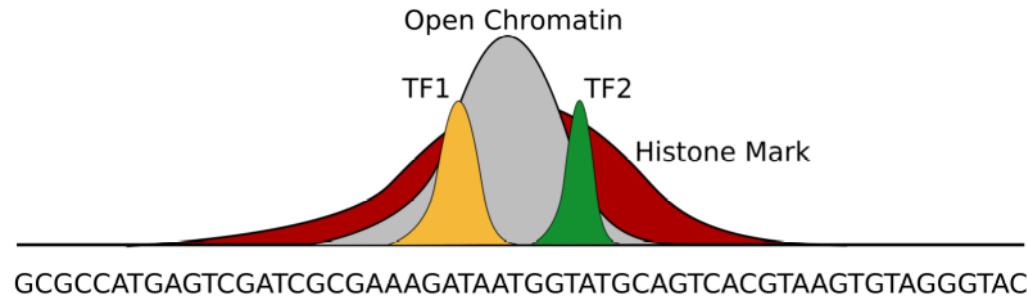
Gökçen et al. Nature Reviews Genetics 2019

Coffee Break!

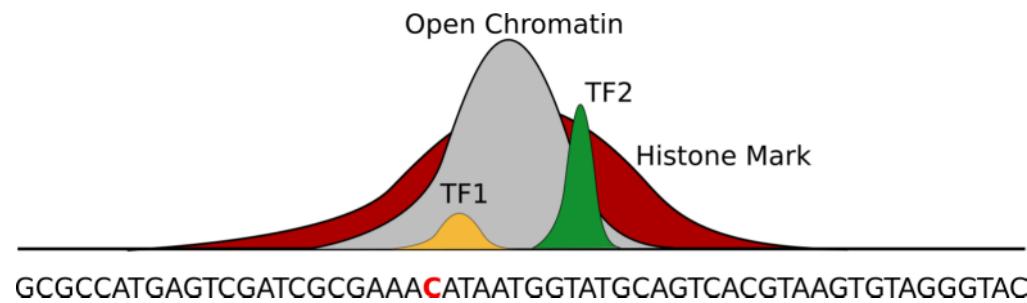
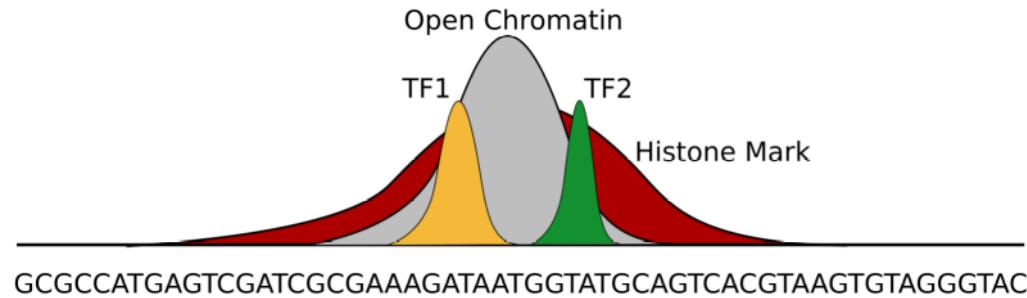


3) Convolutional Neural Networks in Genomics

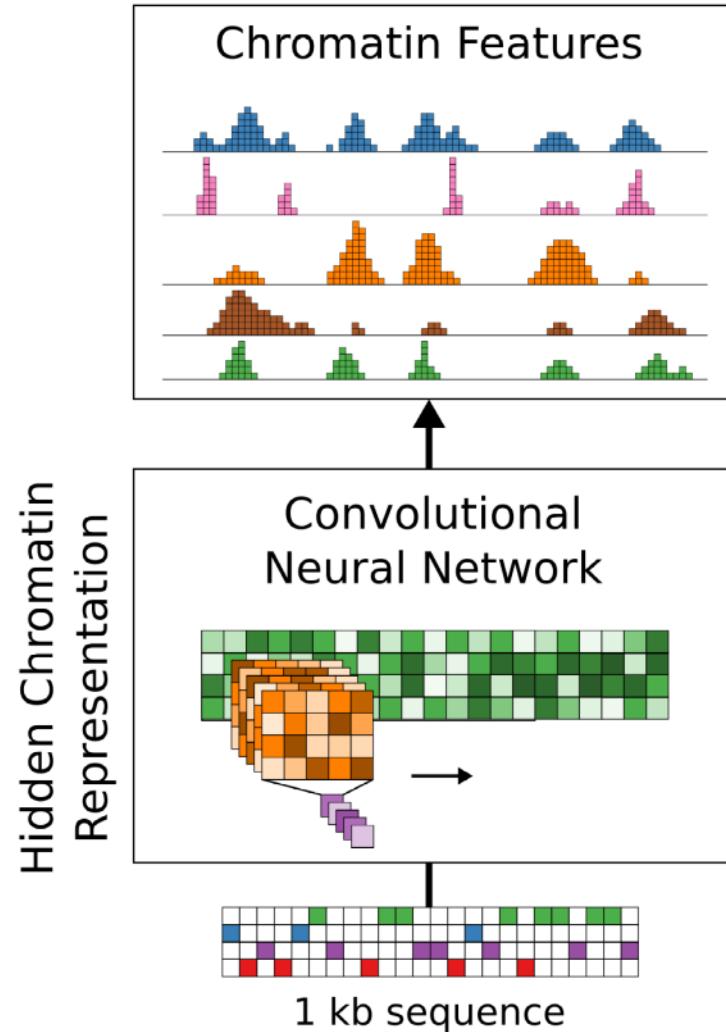
Chromatin Feature Networks



Chromatin Feature Networks

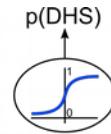


Chromatin Feature Networks

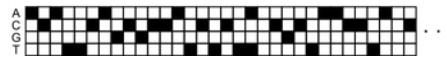


Chromatin Feature Networks

Output

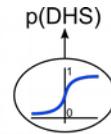


Sequence



Chromatin Feature Networks

Output



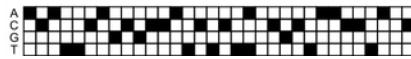
Layer 1:

Max pooling

Convolution

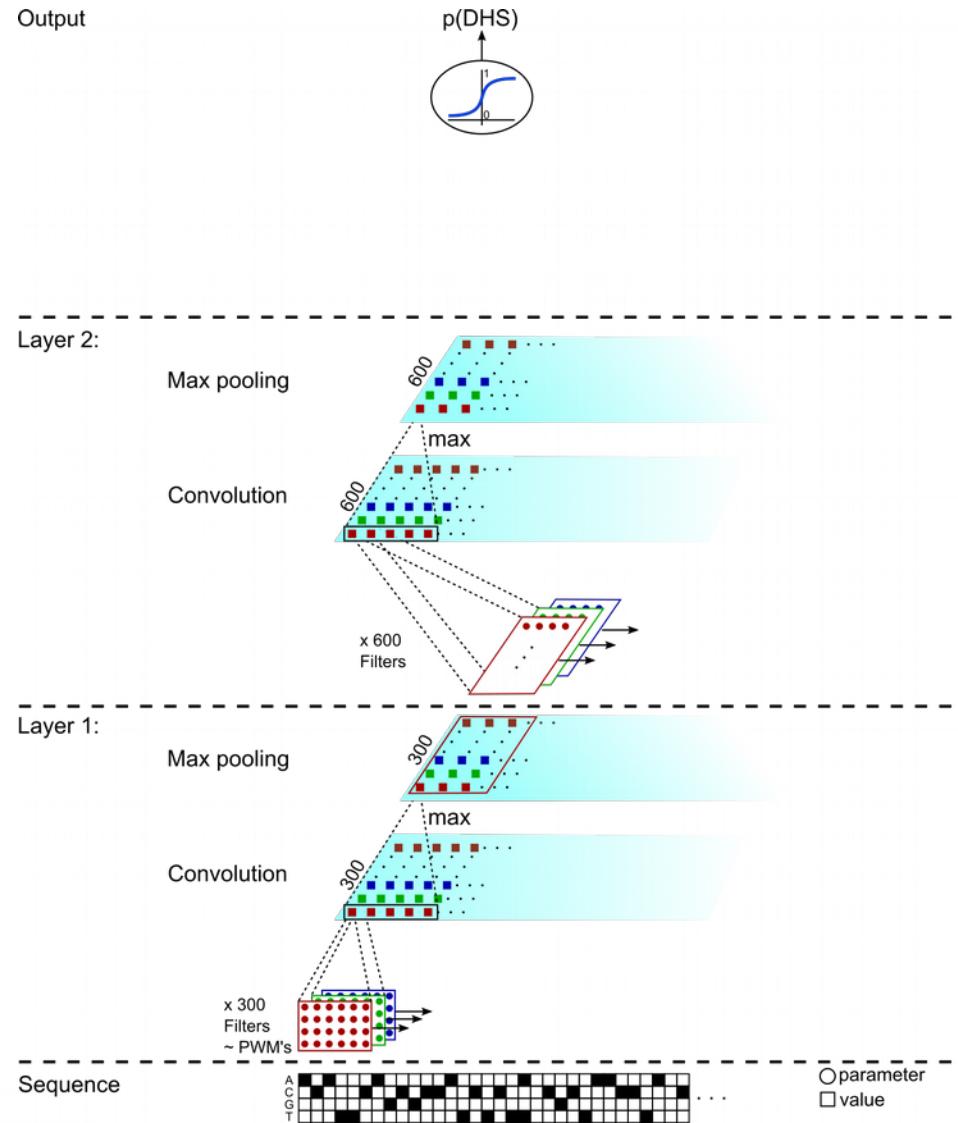
x 300
Filters
~ PWM's

Sequence

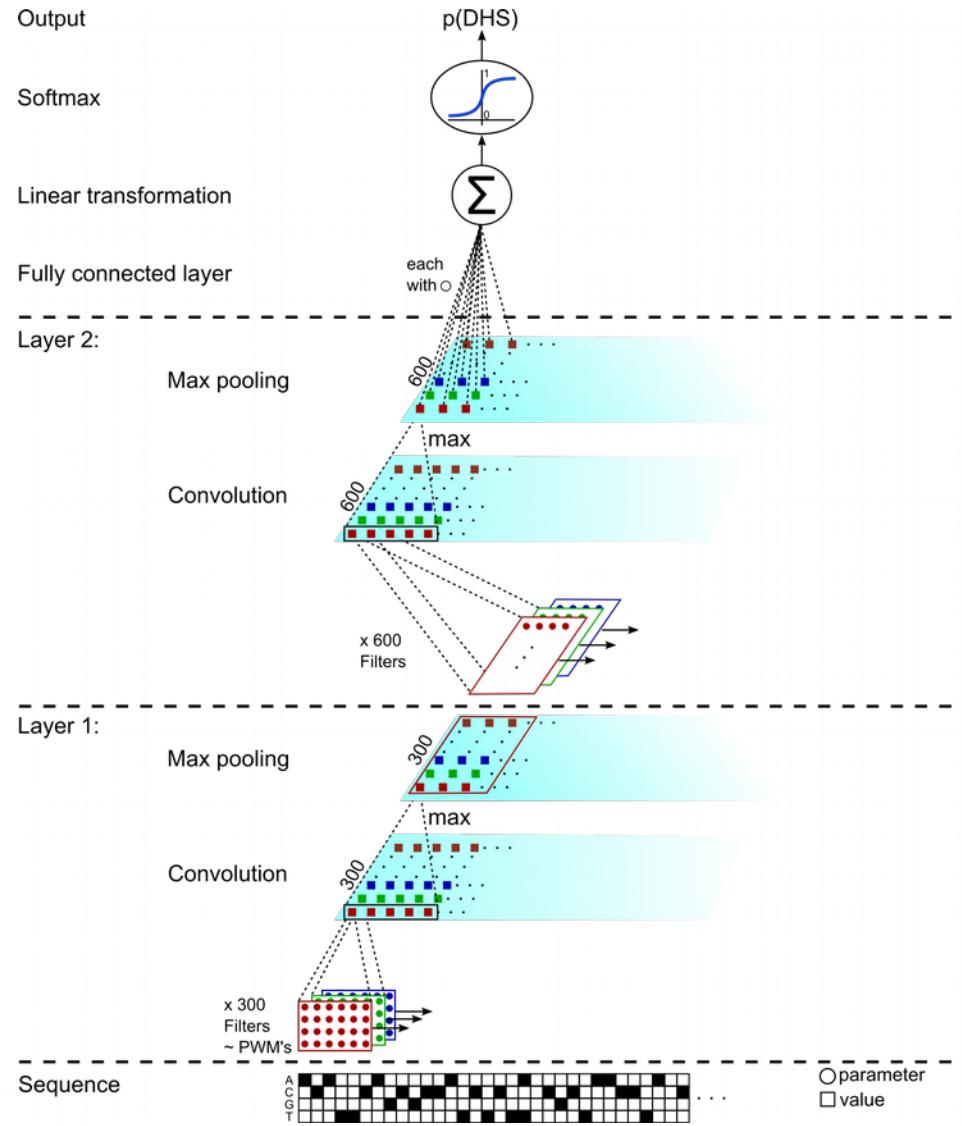


○ parameter
□ value

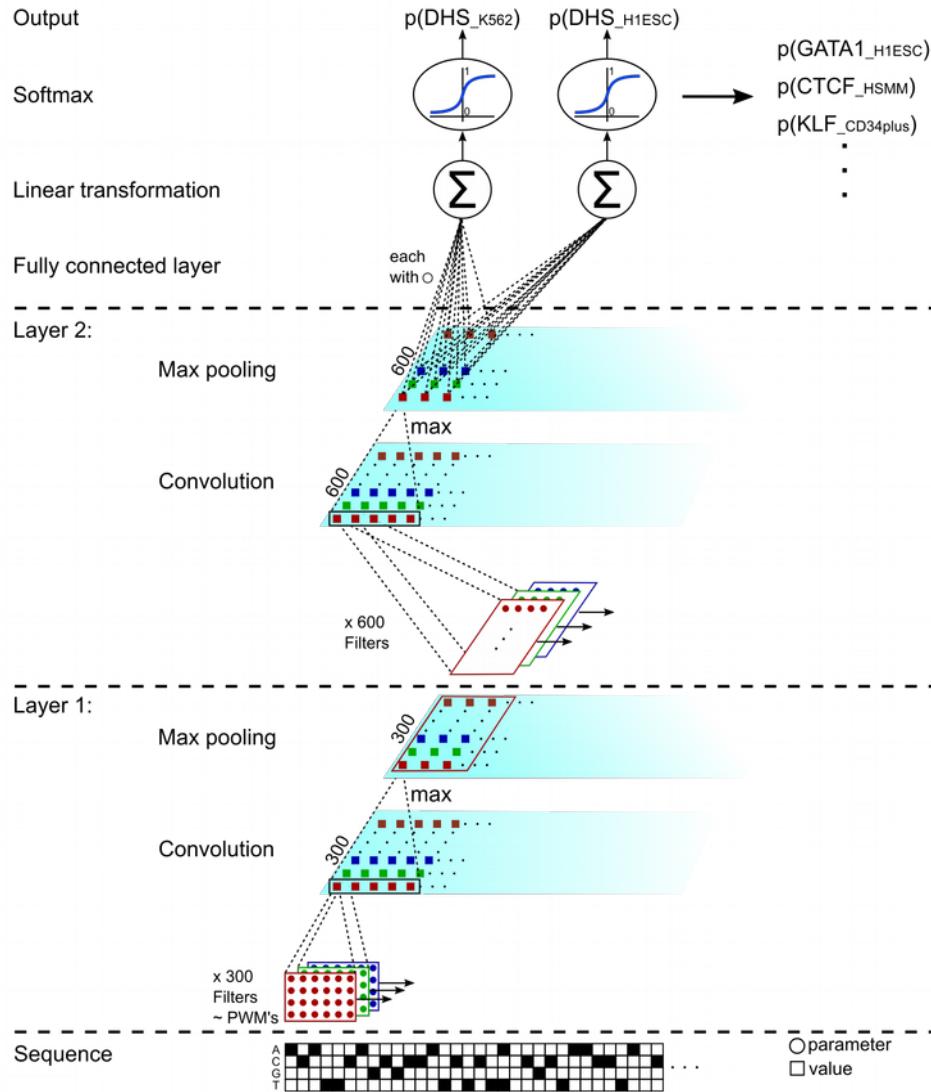
Chromatin Feature Networks



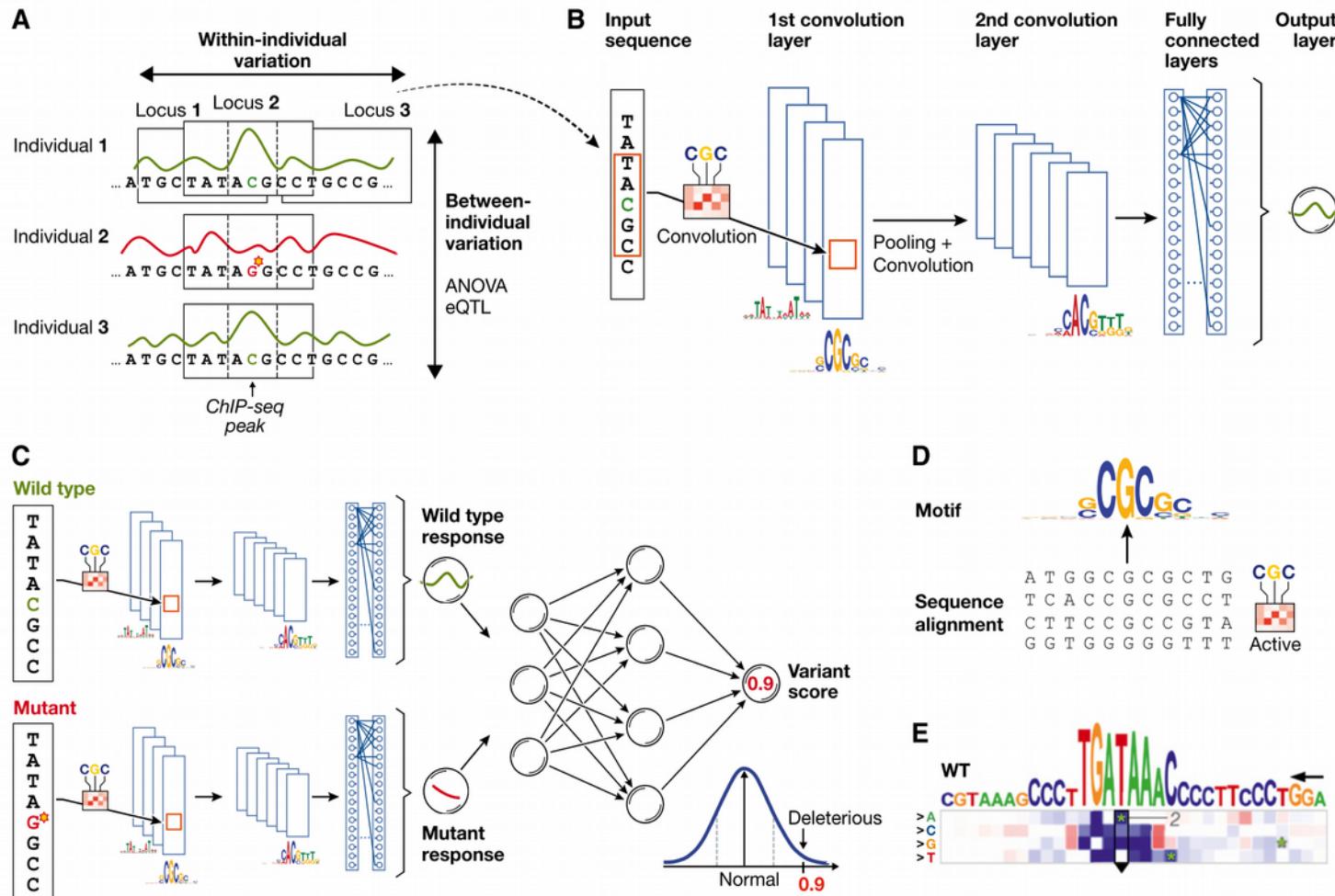
Chromatin Feature Networks



Chromatin Feature Networks

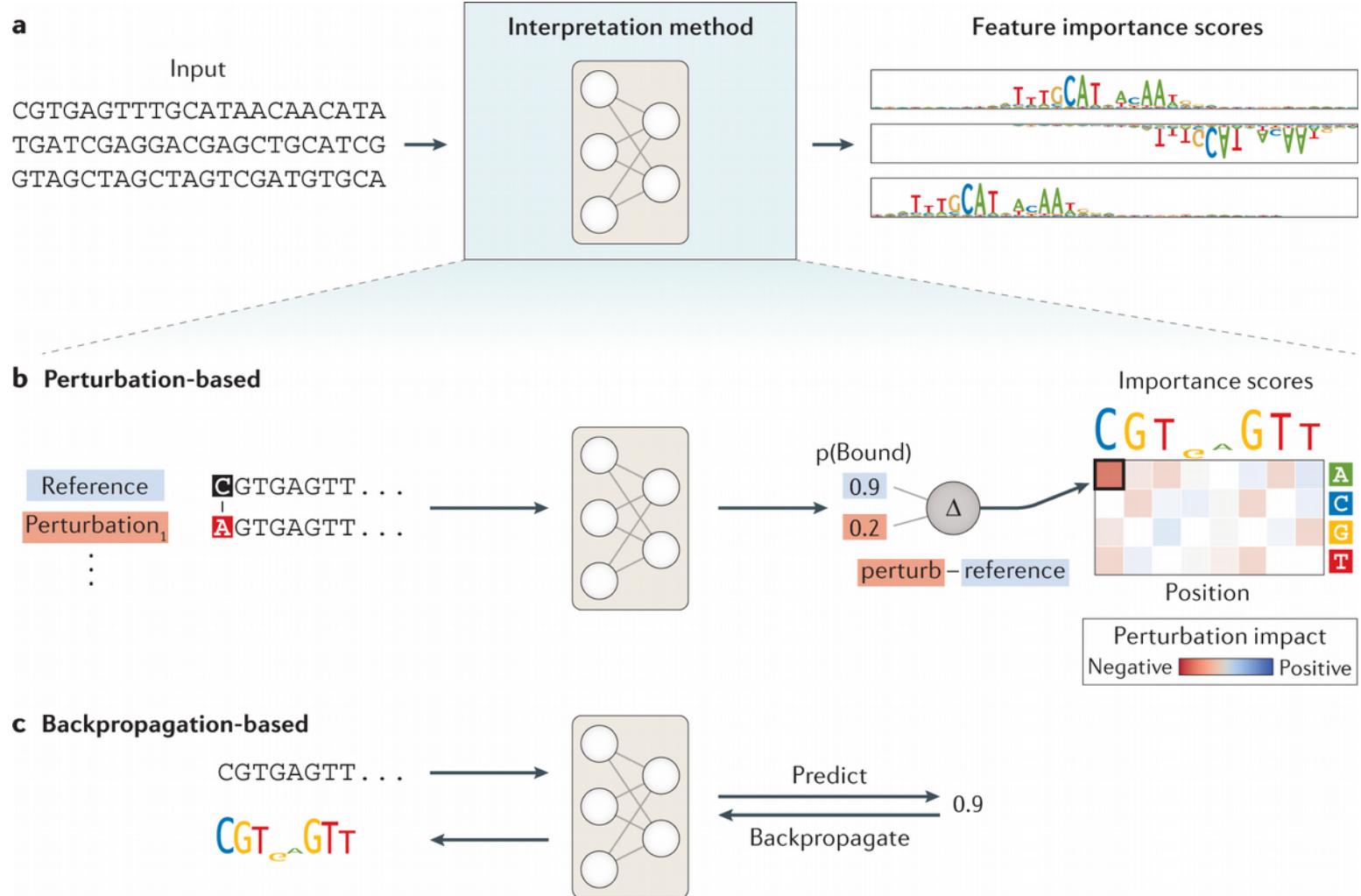


Utility



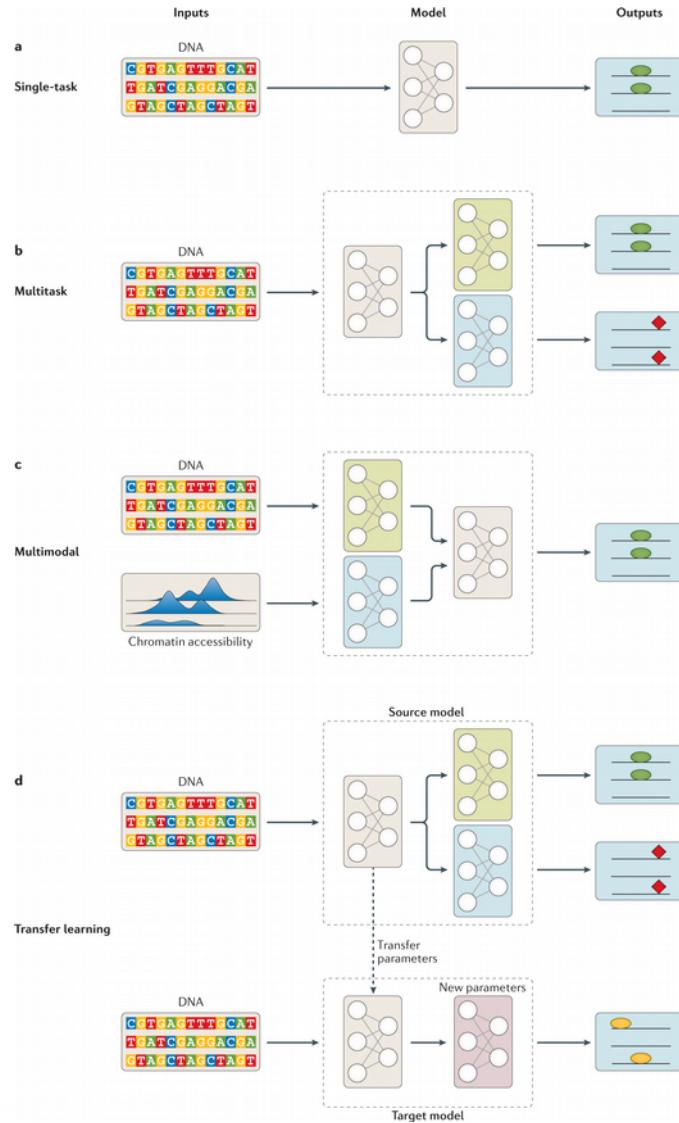
Angermueller et al. 2016

Interpretation



Gökçen et al. Nature Reviews Genetics 2019

Learning Strategy

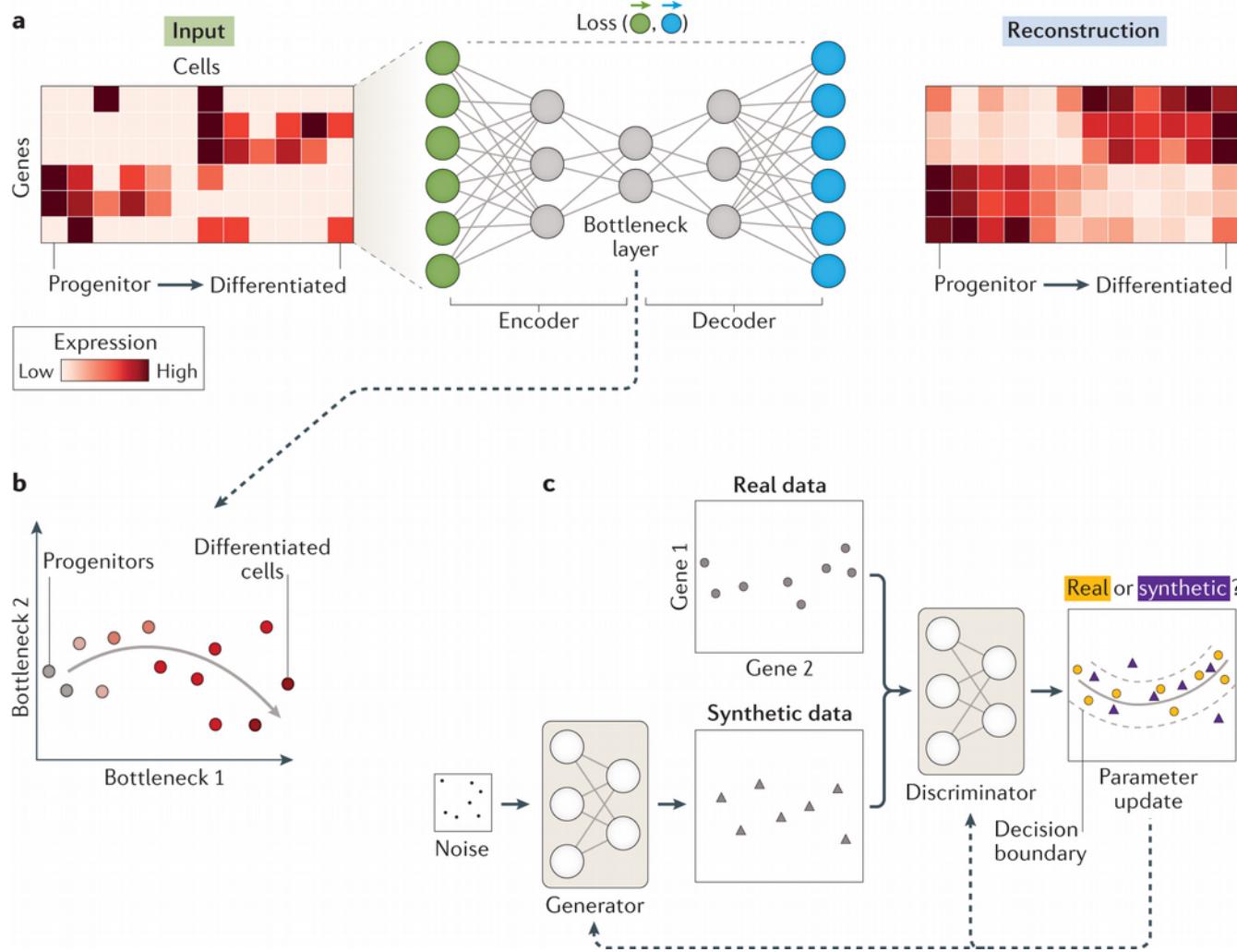


Gökçen et al. Nature Reviews Genetics 2019

More Examples, More Architectures

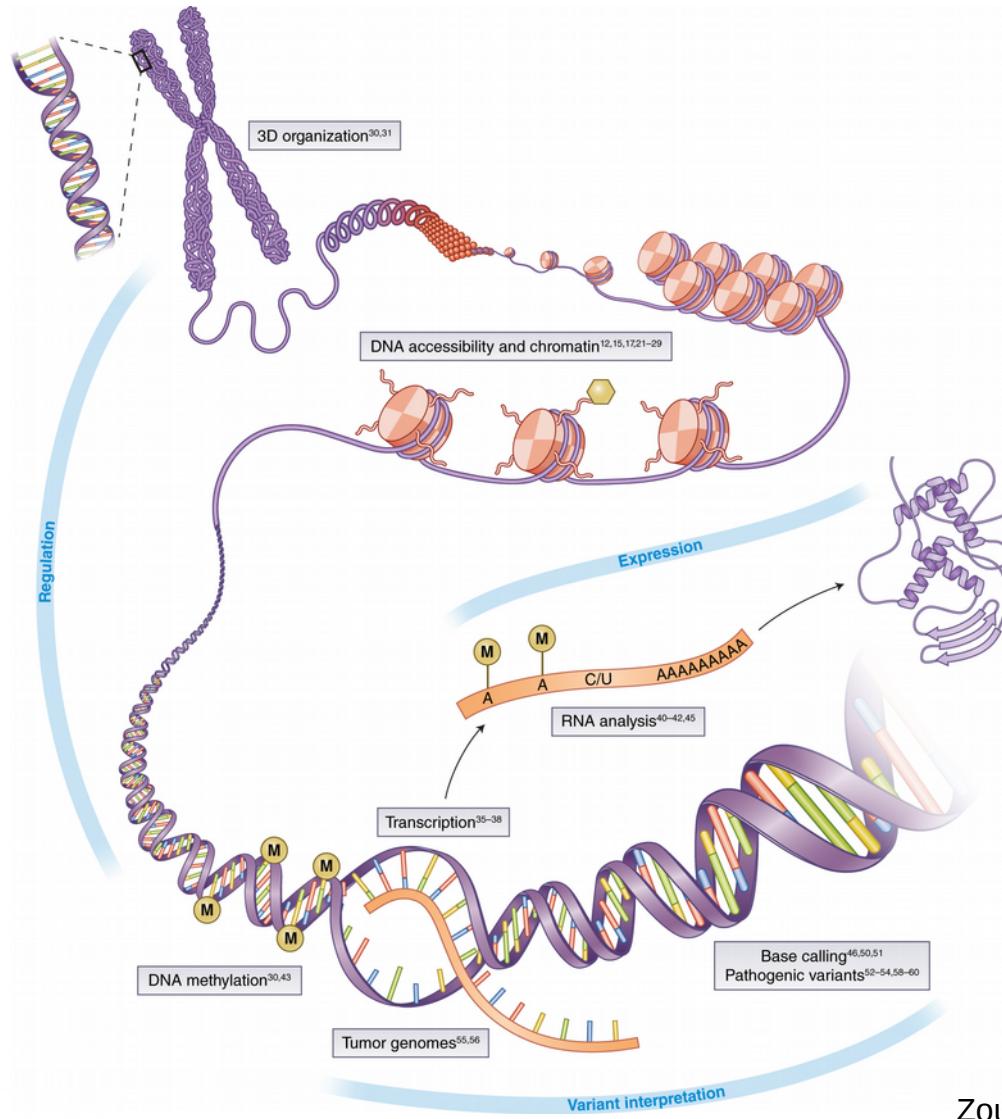
Gökcen *et al.* Nature Reviews Genetics 2019

AutoEncoders and GANs



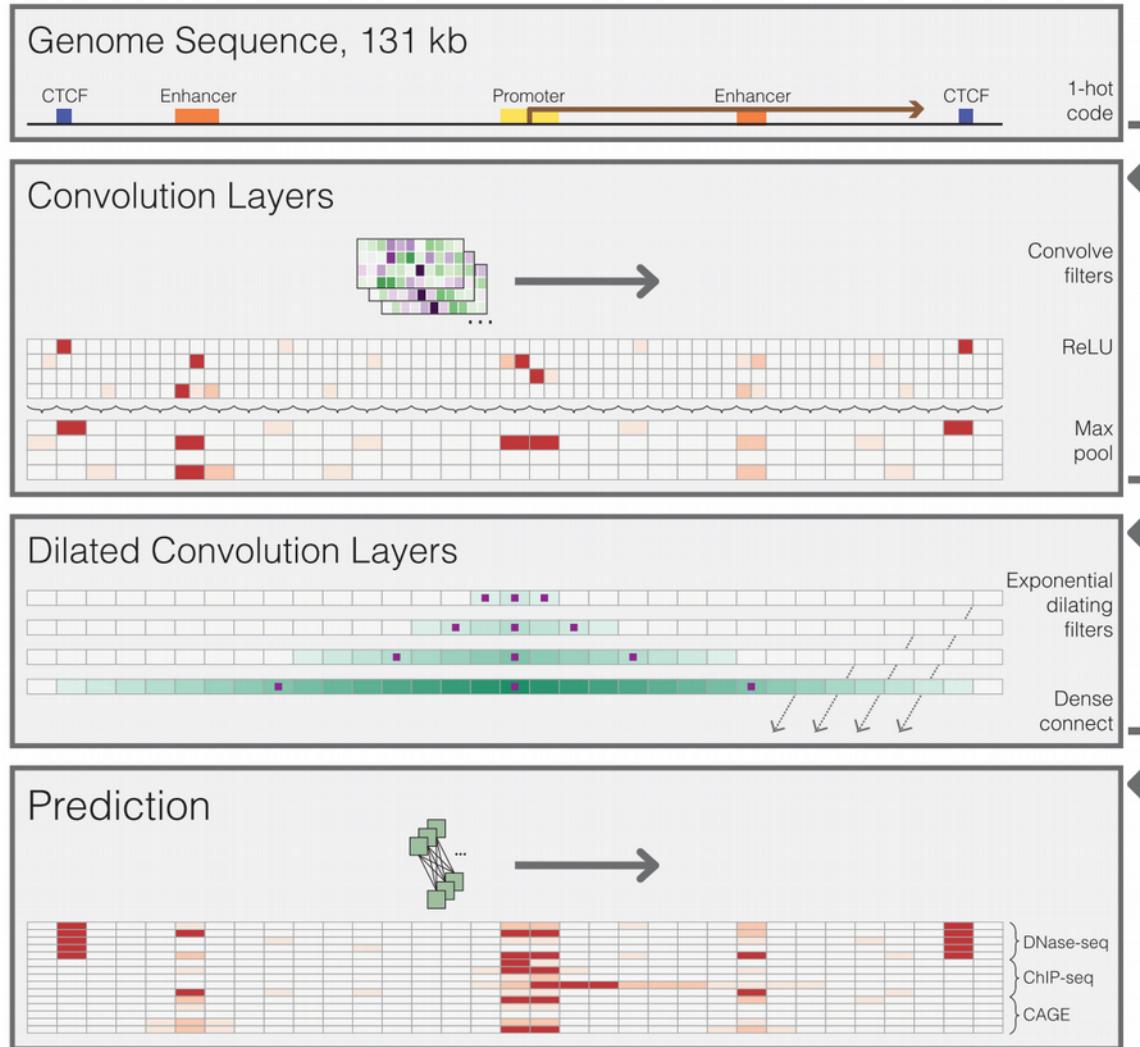
Gökçen et al. Nature Reviews Genetics 2019

Deep Learning in Genomics



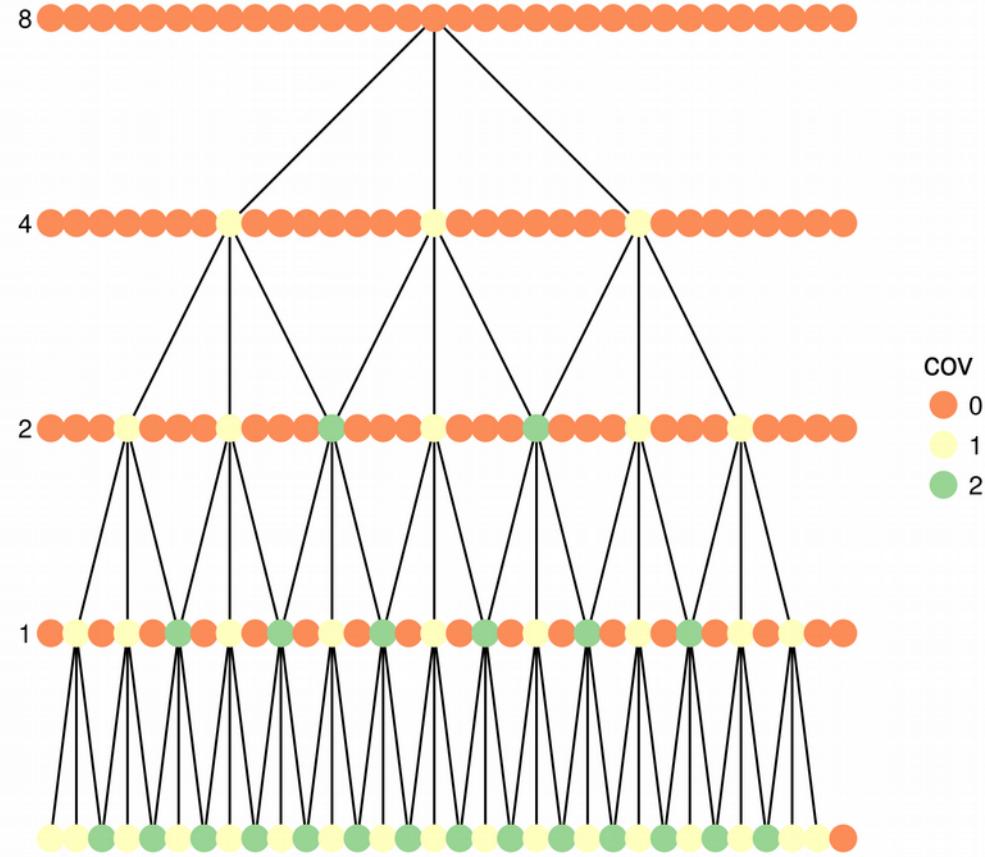
Zou et al. Nature Genetics 2019

Examples – Dilated Networks



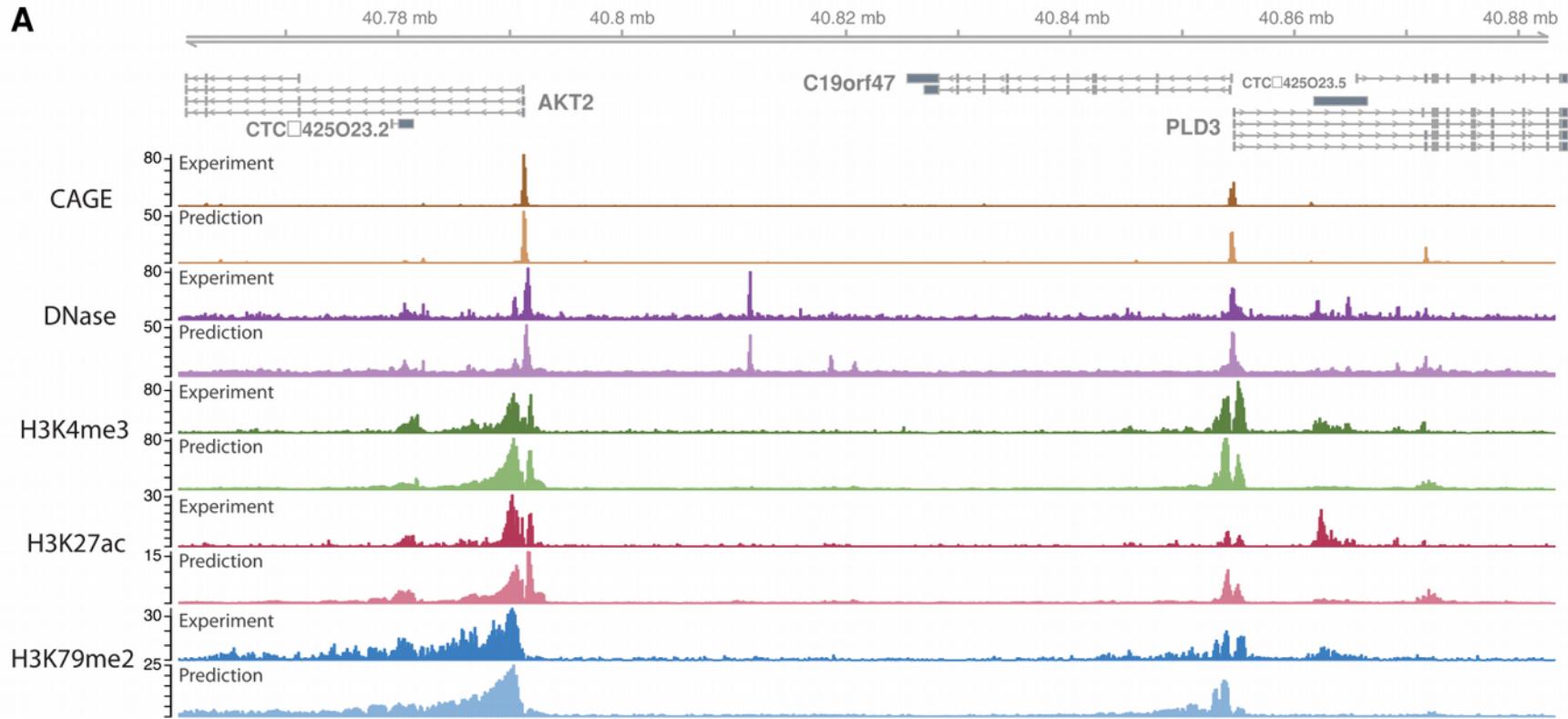
Kelley et al. Genome Research 2018

Examples – Dilated Networks

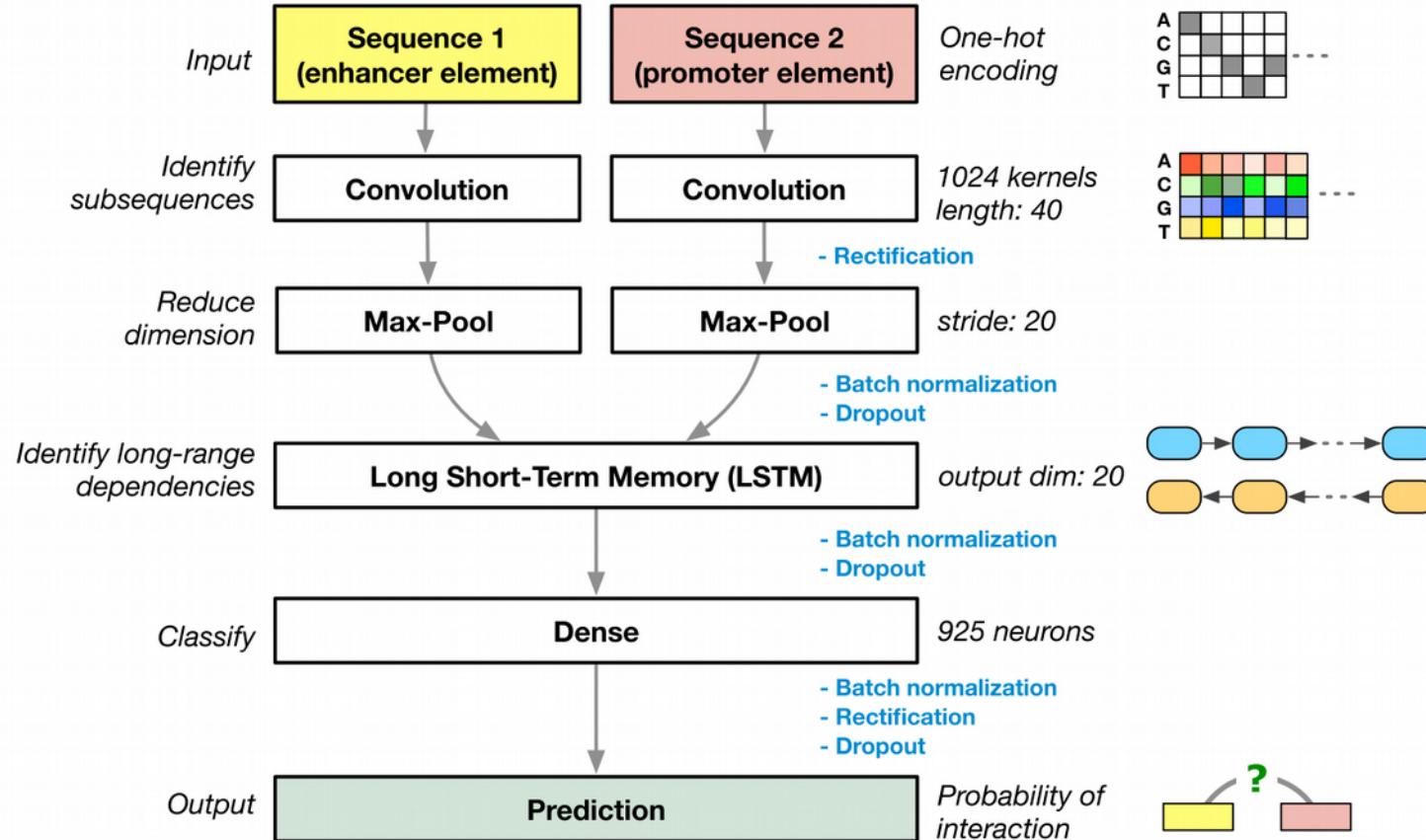


Examples – Dilated Networks

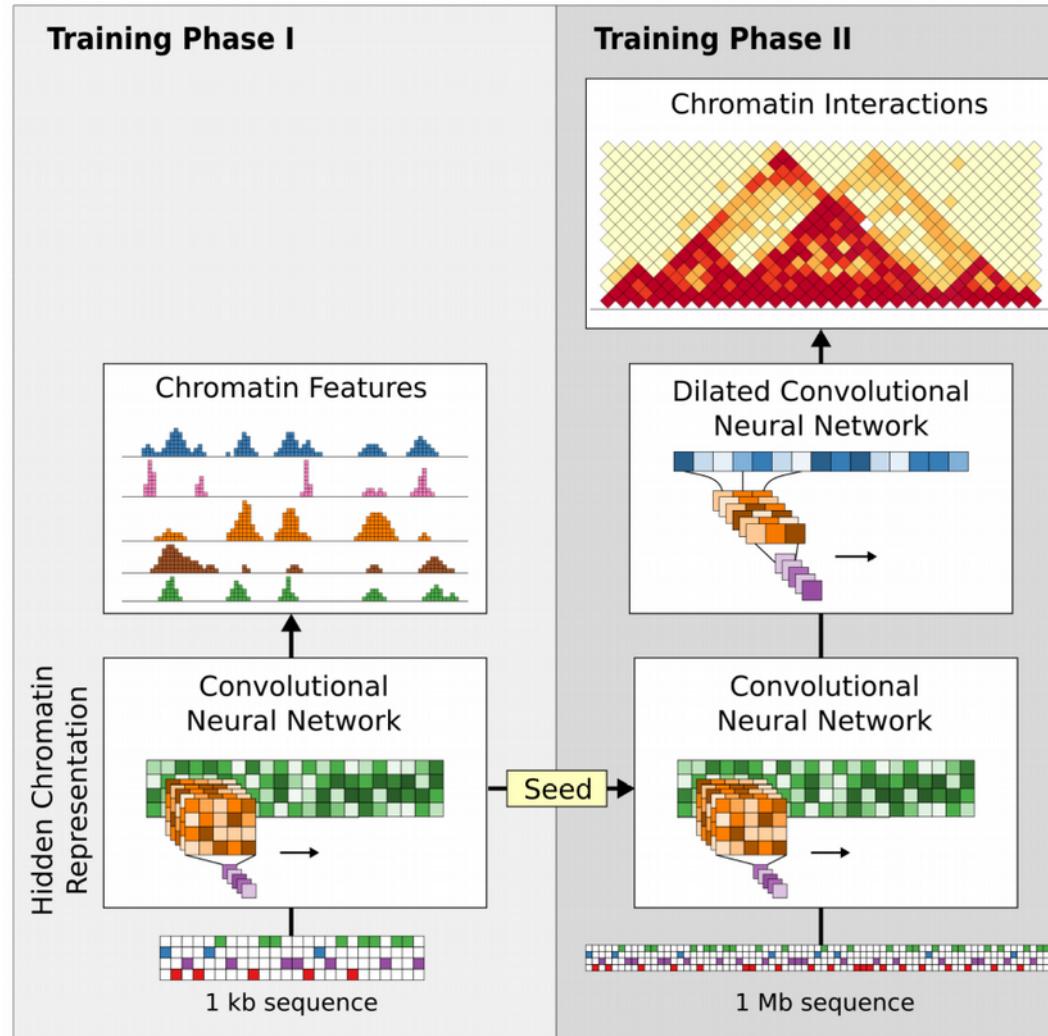
A



Examples – Enhancer – Promoter - Contacts



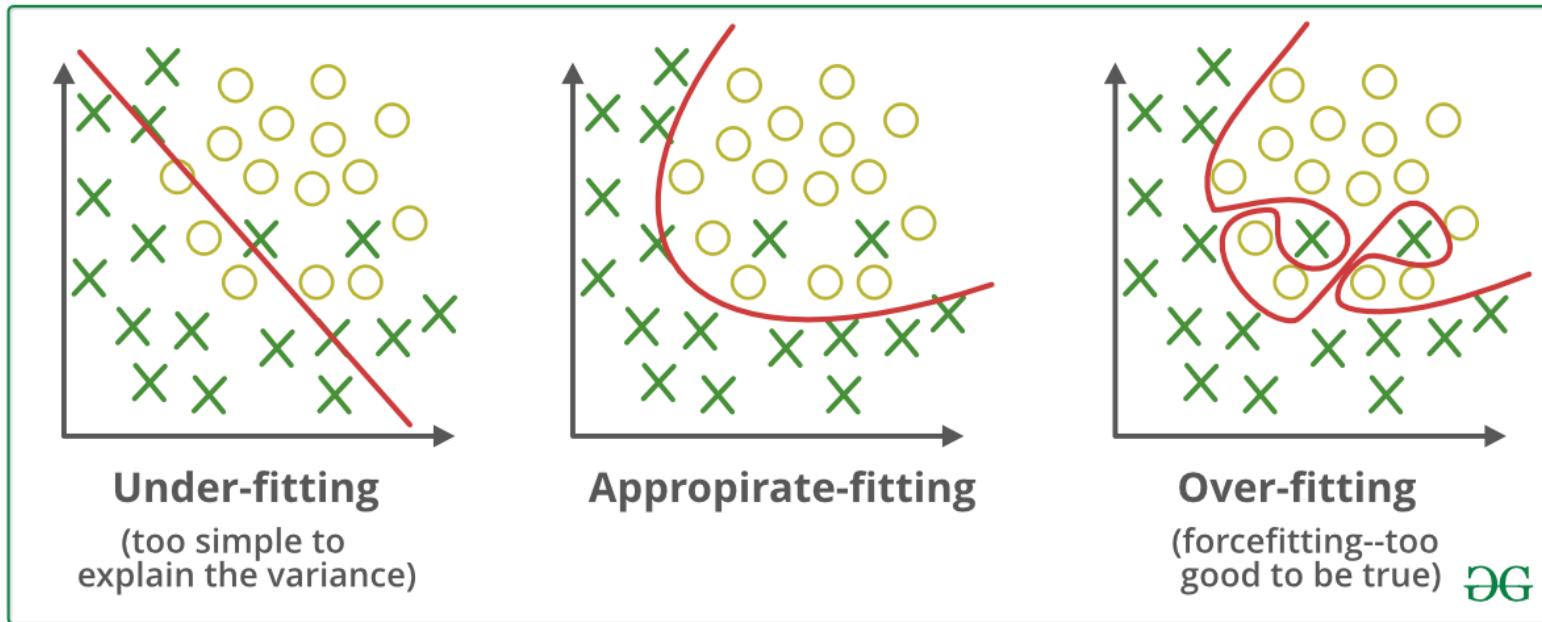
Examples – 3D Genome



Schwessinger et al. bioRxiv 2019

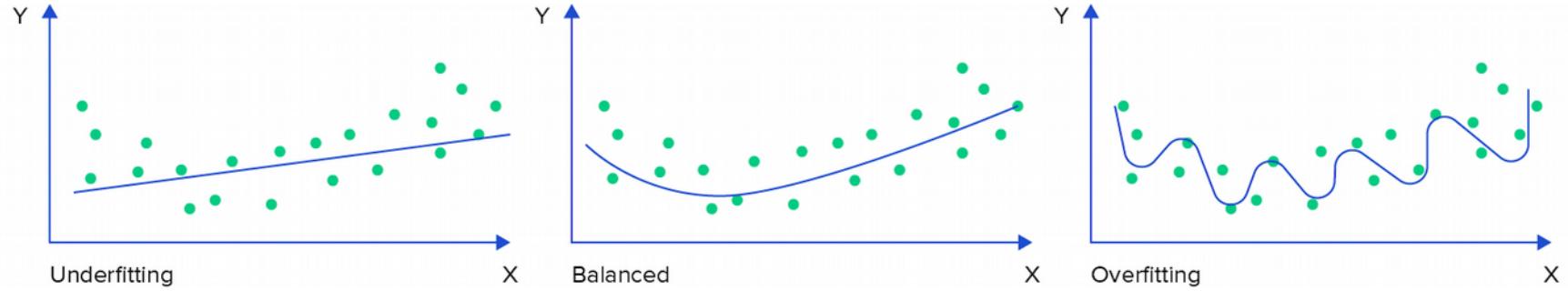
4) Basic Practical Aspects

Overfitting / Underfitting



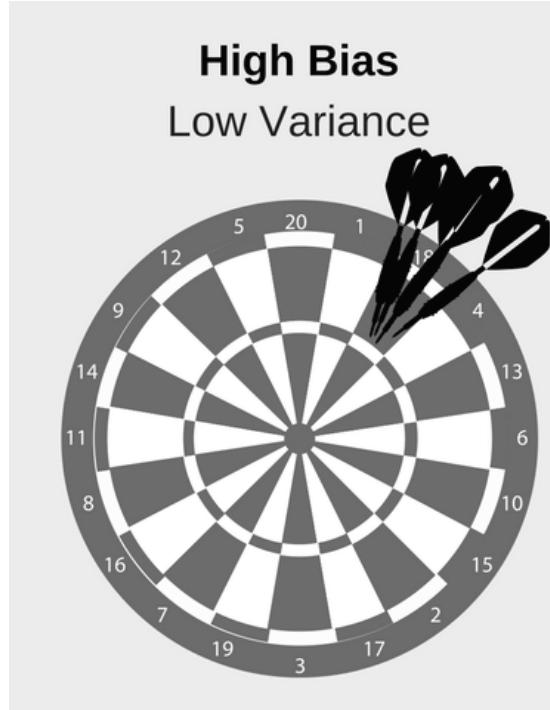
Source towardsdatascience.com

Overfitting / Underfitting

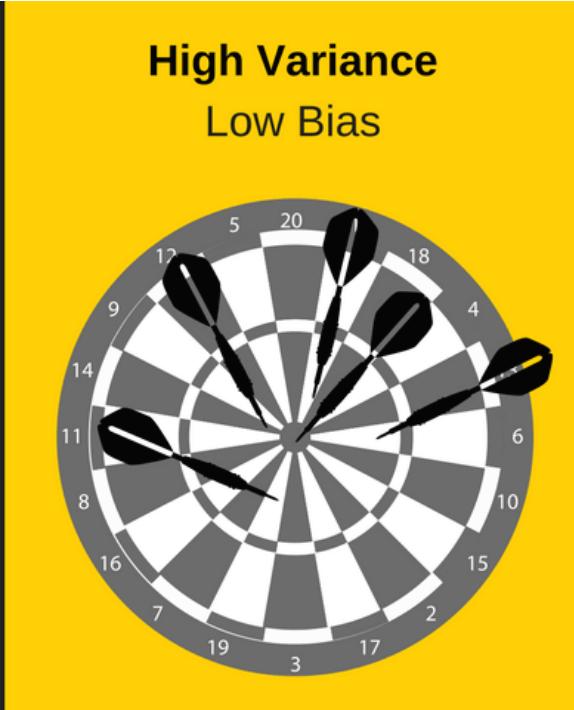


Source towardsdatascience.com

Overfitting / Underfitting



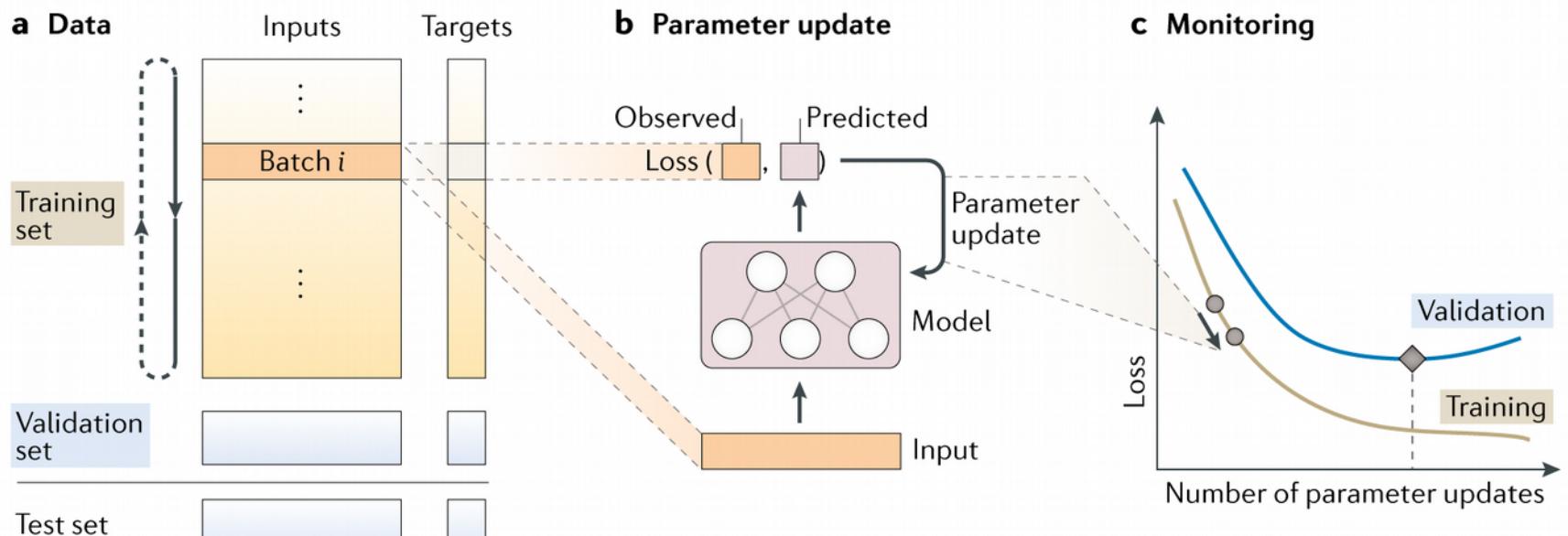
Model to simple
underfitting



Model to complex
overfitting

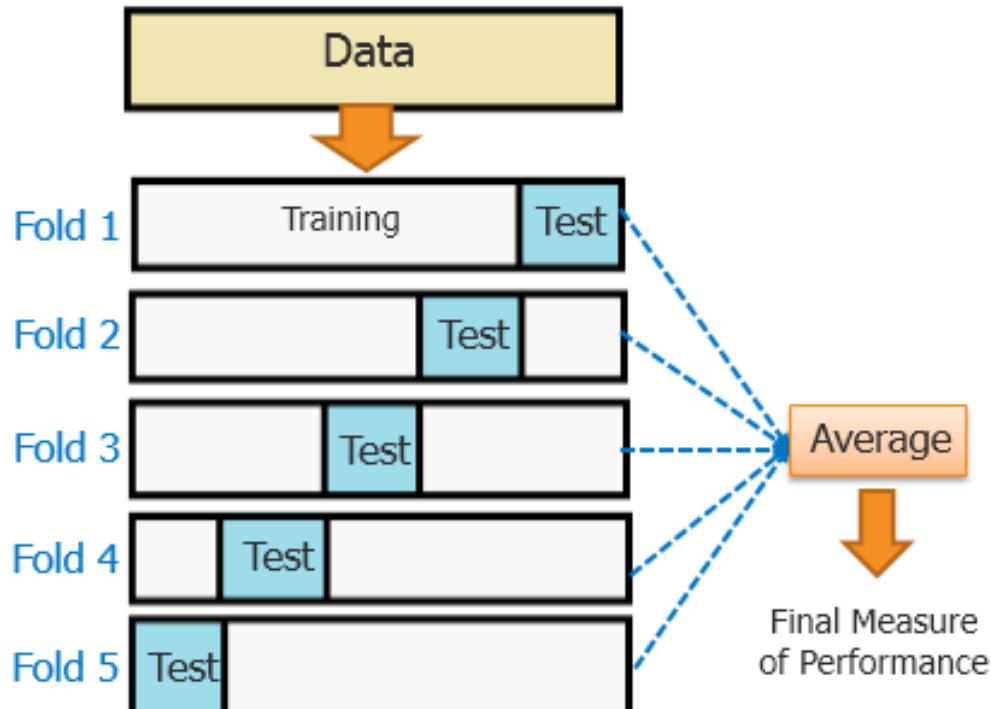
Source [EDS](#)

Train, Validation & Test Sets (& Cross-Validation)



In practice: Split data into stratified train, valid. and test sets. Use cross-validation if low on data points.

Train, Validation & Test Sets (& Cross-Validation)



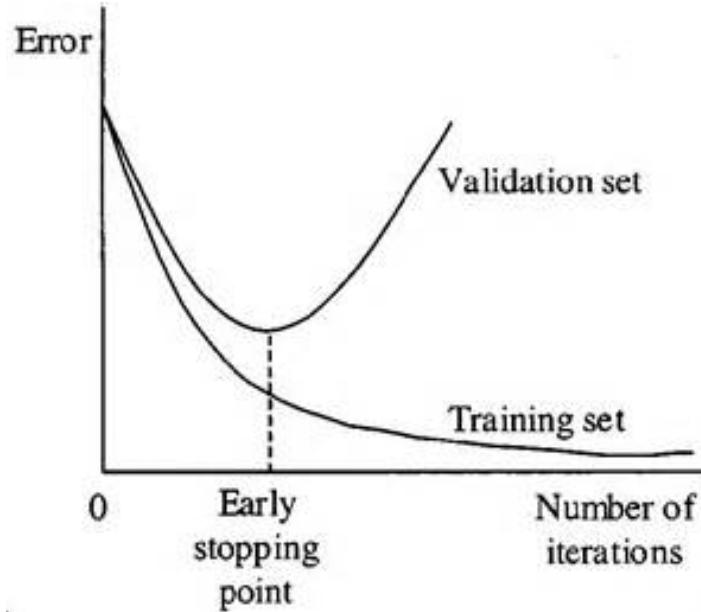
Source blog.contact sunny.com

In practice: Split data into stratified train, valid. and test sets. Use cross-validation if low on data points.

Overfitting / Underfitting

Tackle overfitting by in practice by:

- Optimizing regularization (L1 penalty, L2 penalty, Dropout)
- Use more data
- Early stopping
- Ensemble
- ...

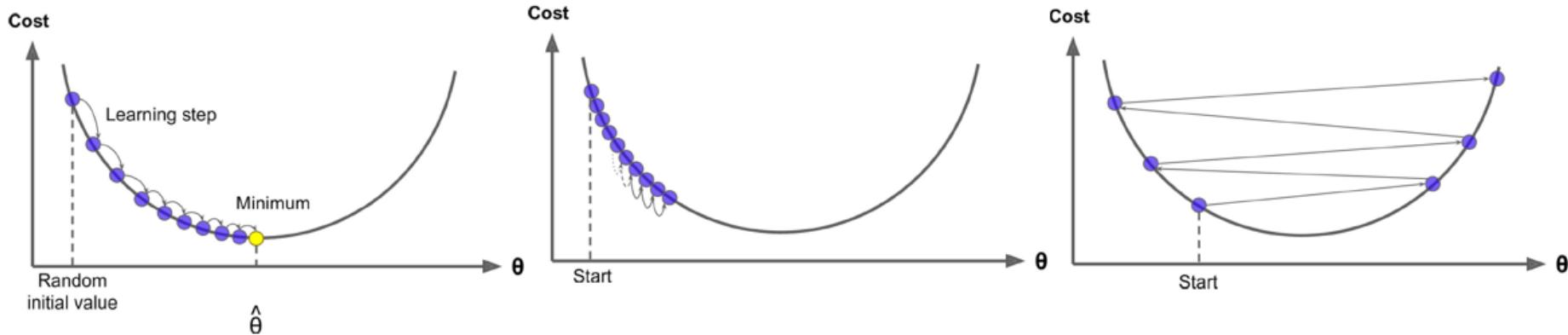


Overfitting / Underfitting

Tackle underfitting by in practice by:

- Increase the complexity / capacity of your model
- Use more data

Learning Rate & Optimizer



Optimum learning rate : The model adjusts weights (θ) in subsequent training loops to arrive at cost minima.

Slow learning rate : Converges to cost minima but very slowly.

Fast learning rate : may not converge to cost minima and the cost might keep increasing with further training loops.

Image Credit : "Hands-on Machine Learning with Scikit-Learn and TensorFlow" by Aurelien Geron

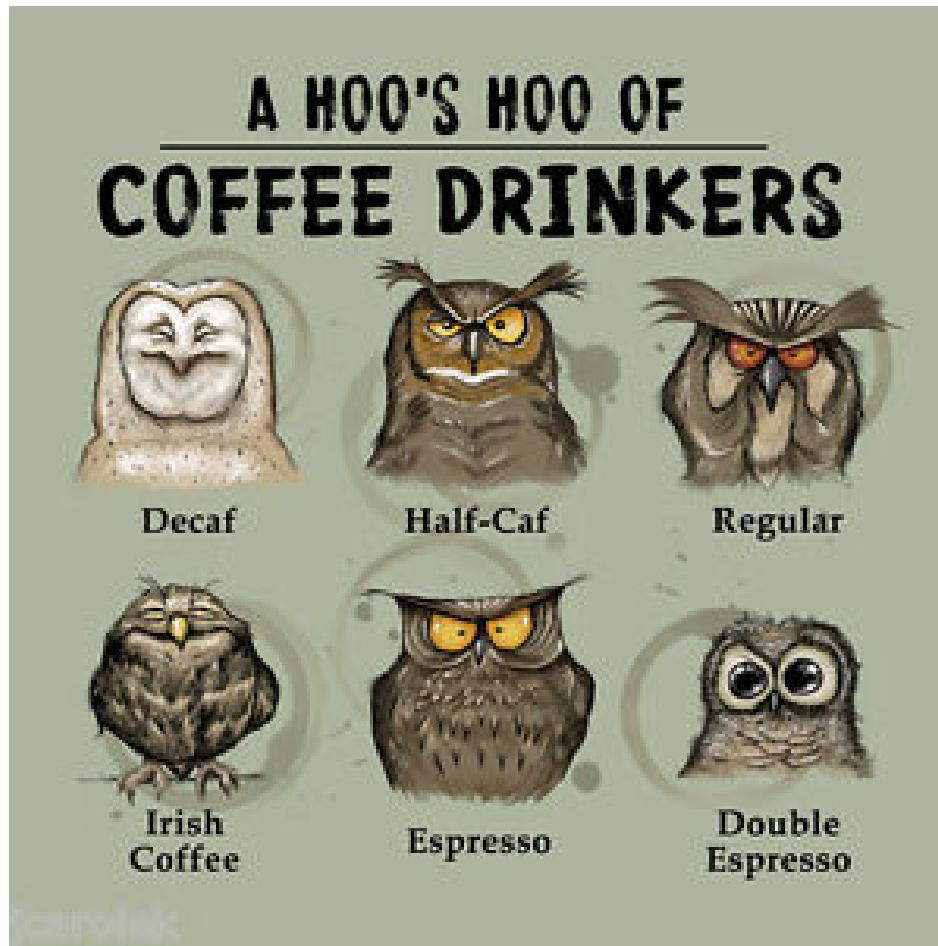
Disrupt4.0



In practice: optimize the learning and pot. try different optimizers. But ADAM is very good default. Also use batches where you can.

Questions?

Coffee Break!



Practical

Head over to <https://github.com/NGSchoolEU/ngs19>

Scroll down to **Deep learning methods for genomics** and use the **Colab** link.

https://colab.research.google.com/drive/1SRHe_SXmKeXImNBR6tnhFQ3eThM4-iZu