Challenges of Machine Learning for Transcriptomics

AI for genomics Bootcamp

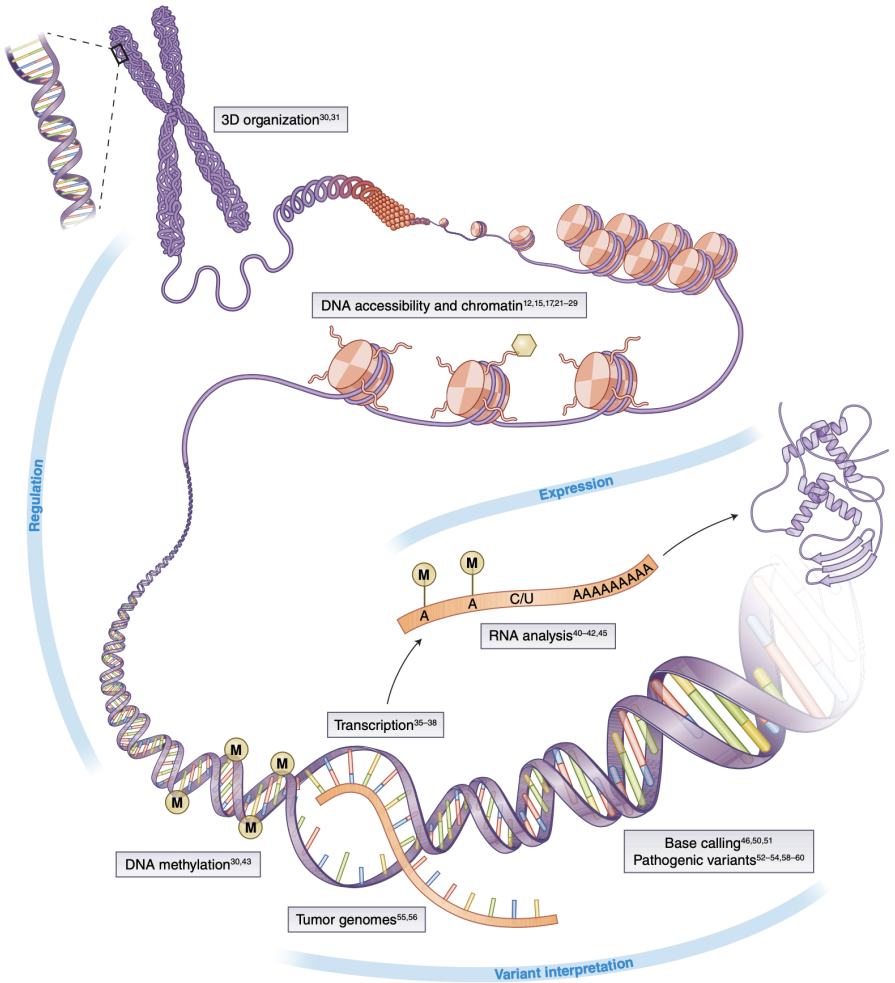
Paul Bertin

Mila, Université de Montréal

February 7, 2020

Outline

Applications of deep learning in genomics.



- Lots of applications of deep learning in genomics
- Today focus on applications to **transcriptomics** and its challenges

Figure taken from *A primer on deep learning in genomics*

What are transcriptomics?

- The study of an organism's transcriptome, the sum of all of its RNA transcripts
- We will focus on RNA-seq and single cell RNA-seq

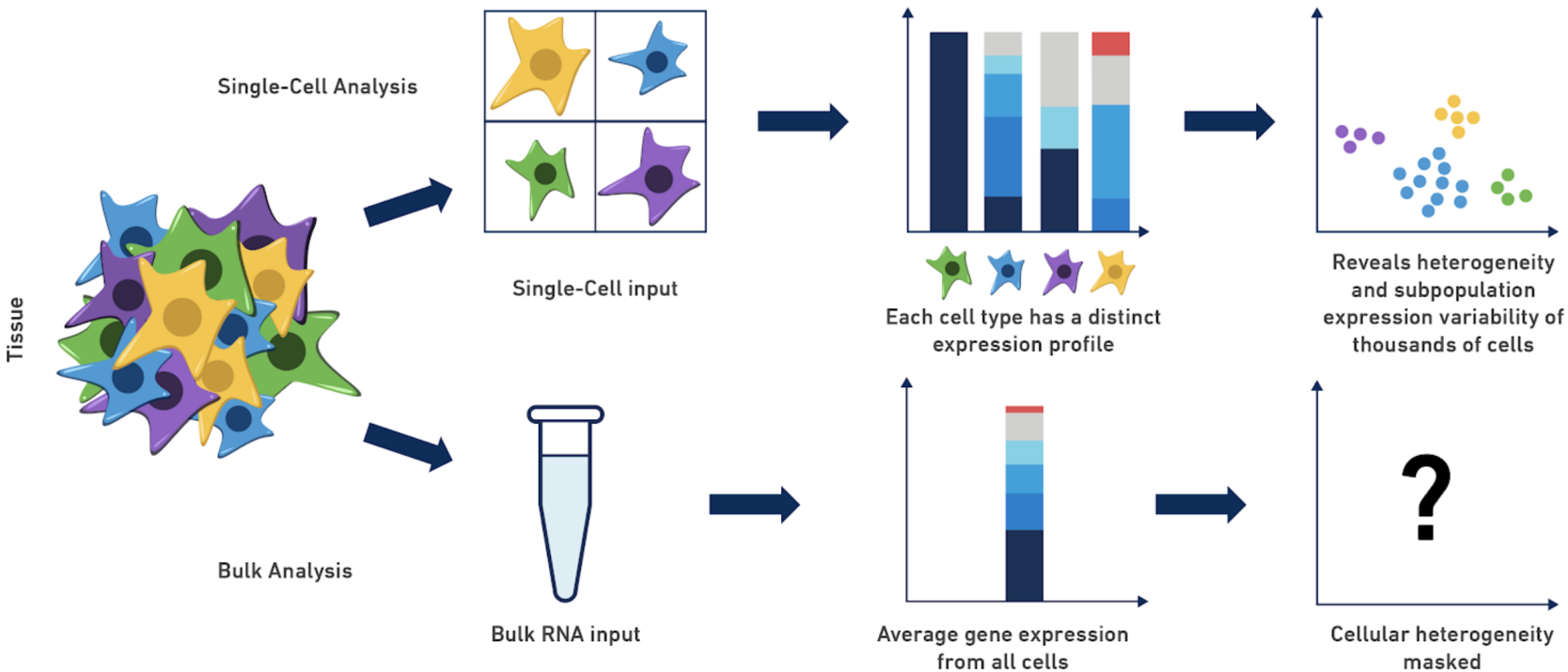
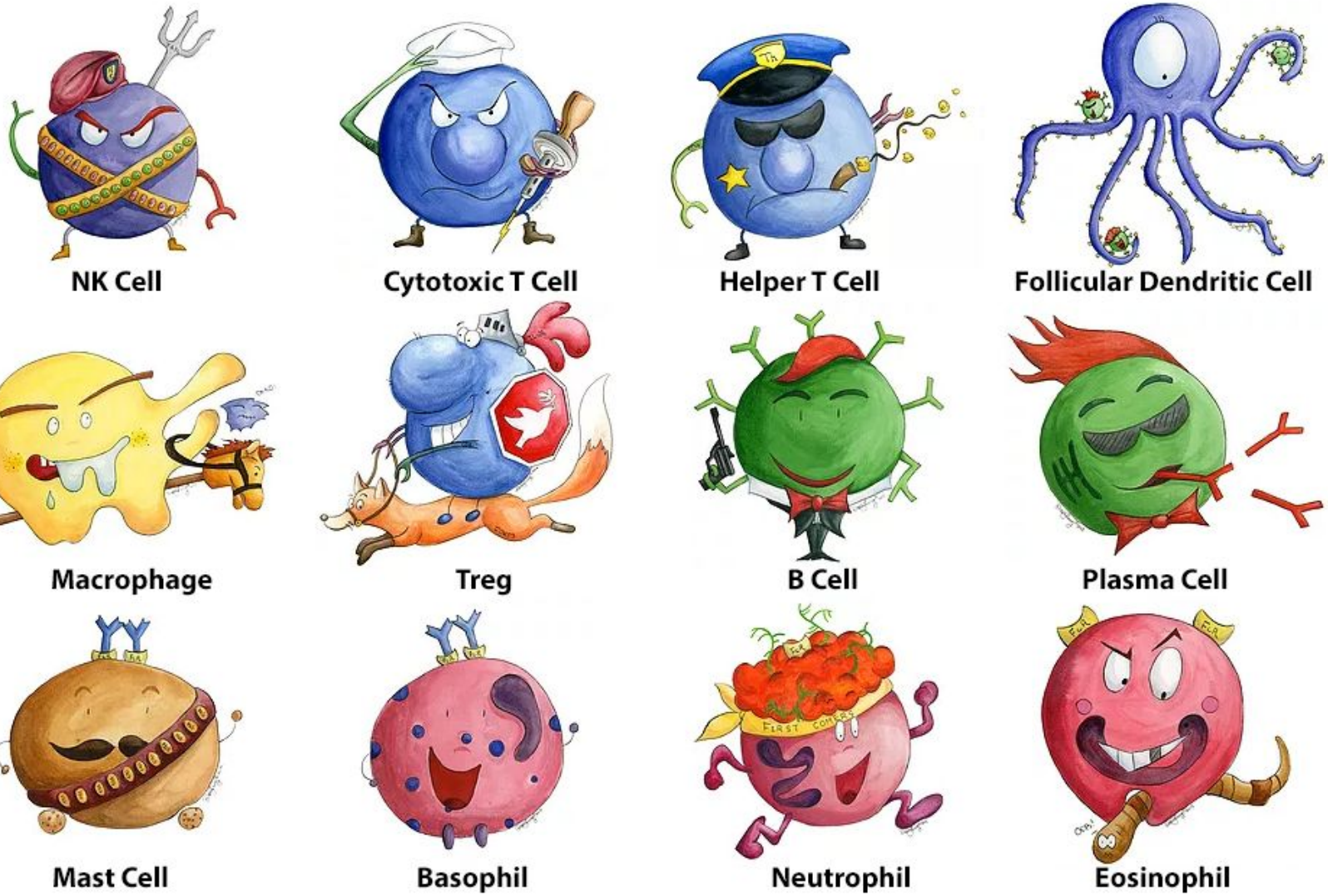


Figure taken from *https://www.biocompare.com/Bench-Tips/345311-Single-Cell-Set-Up-Sample-Preparation-Tips/*

Lots of different cell types



More and more data

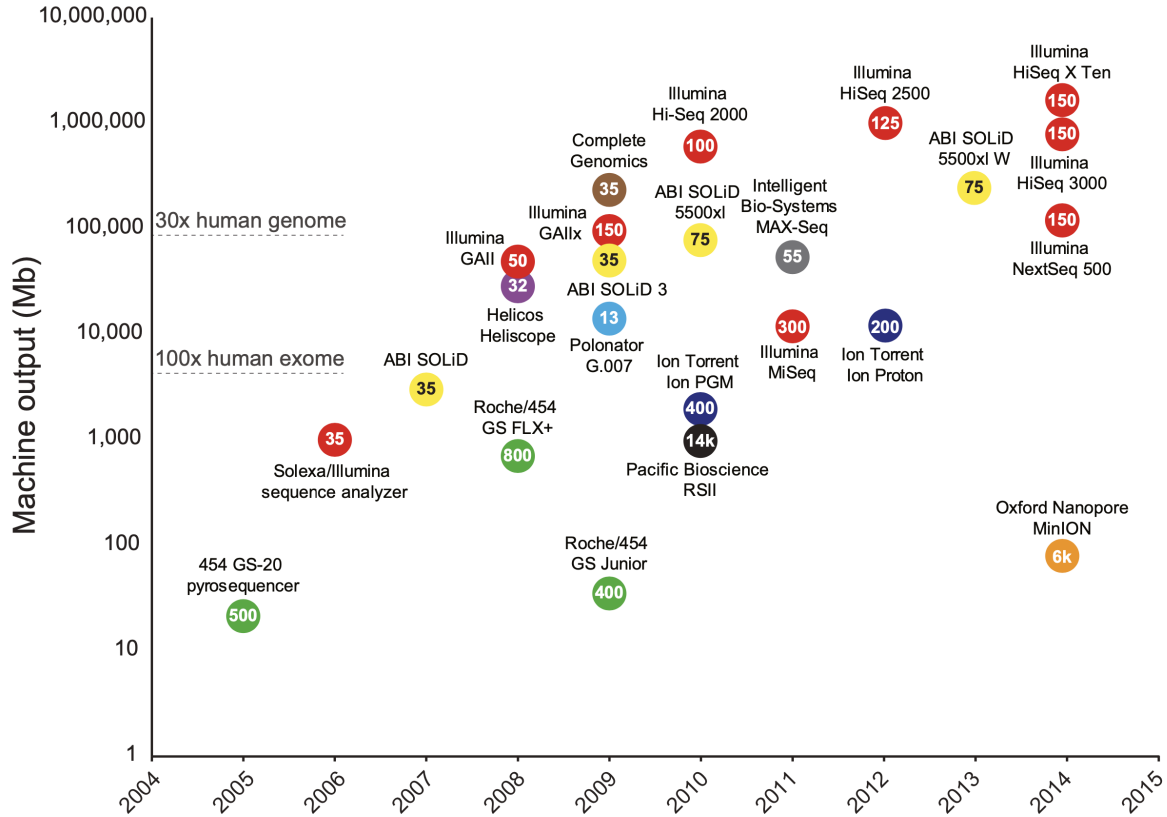


Figure: Plot of commercial release dates versus machine outputs per run are shown. Numbers inside data points denote current read lengths. Sequencing platforms are color coded.

Figure taken from *High-Throughput Sequencing Technologies*

Apply modern machine learning techniques?

Long term goals:

- Individualized medicine
- Better understanding the biology

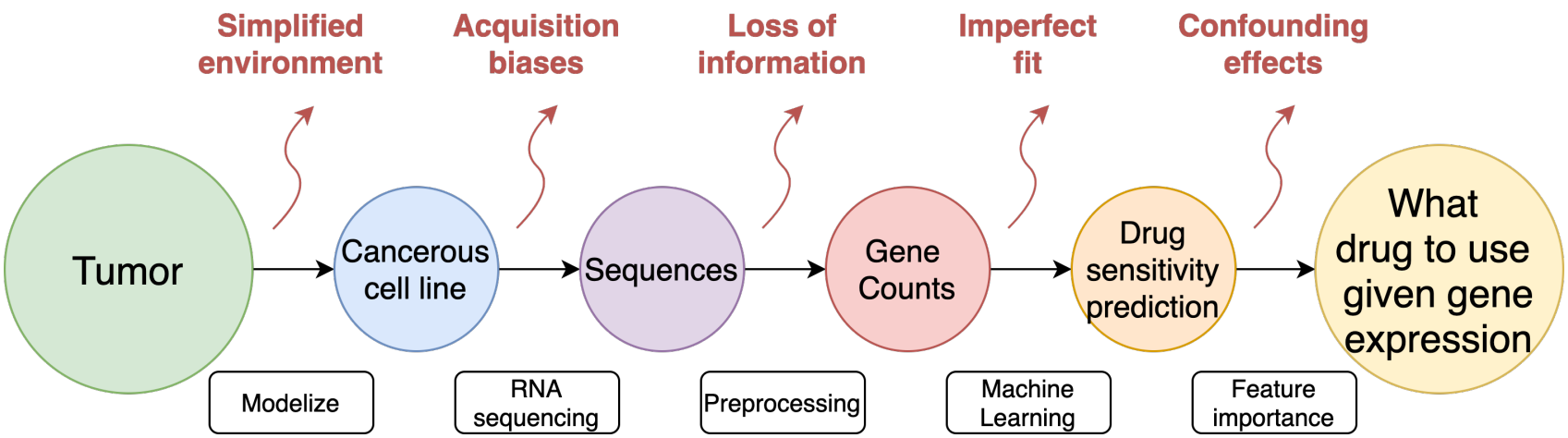
Today's objective

Identify and understand the challenges facing machine learning (ML) and deep learning (DL) techniques when applied to transcriptomics

Why should you care?

- Be aware of the limitations of usual ML
- Take those limitations into account when you use ML
- Discover fields of research in ML

How can machine learning help?

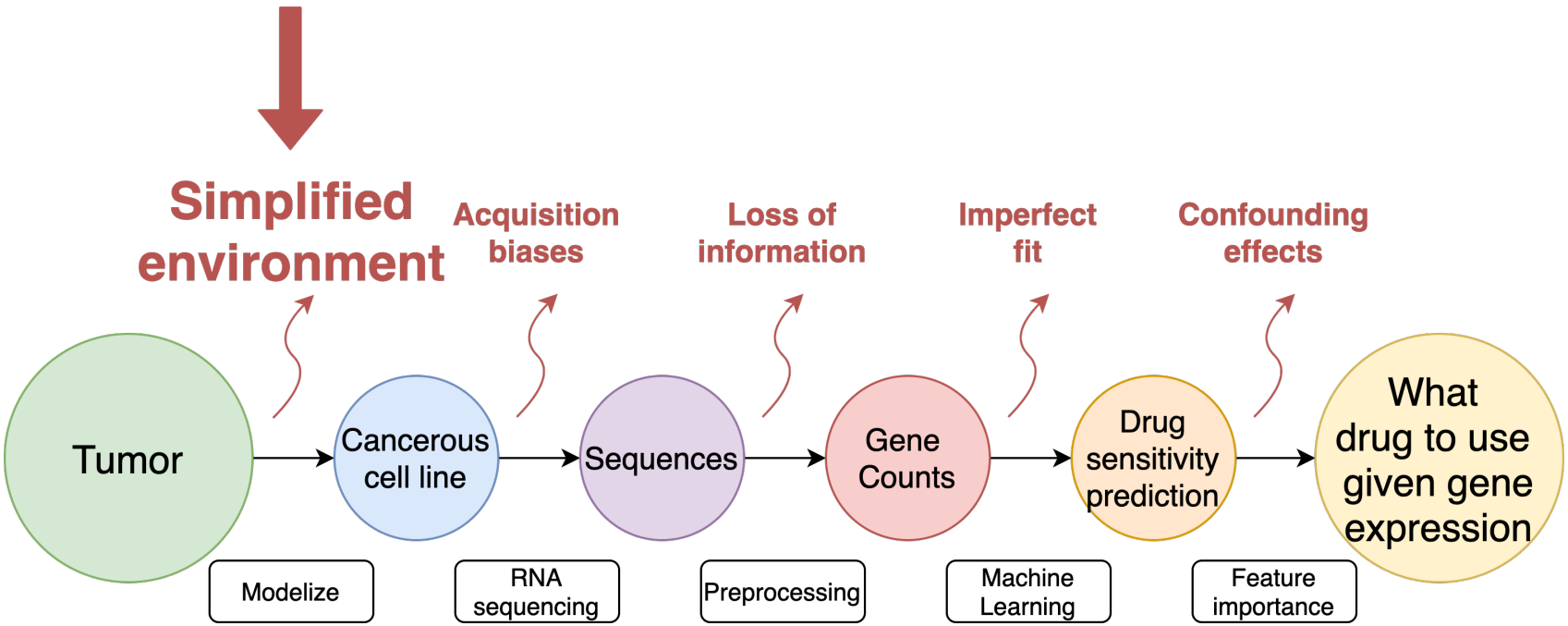


- Example of a pipeline to find better cancer treatment using Machine Learning
- Let us study the limitations associated with each step

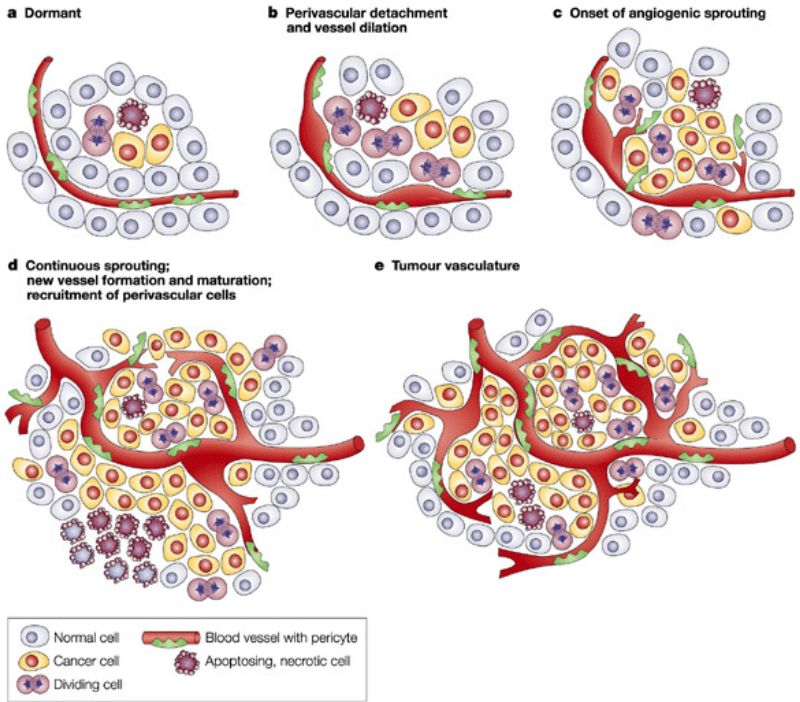
Outline

Outline

Overview


Simplified environment
Acquisition biases
Loss of information
Imperfect fit
Confounding effects
Tumor
Cancerous cell line
Sequences
Gene Counts
Drug sensitivity prediction
What drug to use given gene expression
Modelize
RNA sequencing
Preprocessing
Machine Learning
Feature importance

Cell lines, a model for tumors



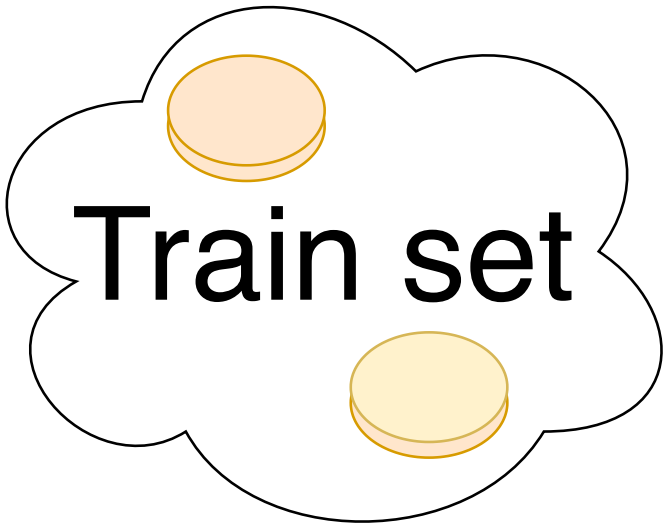
Cell lines as a model

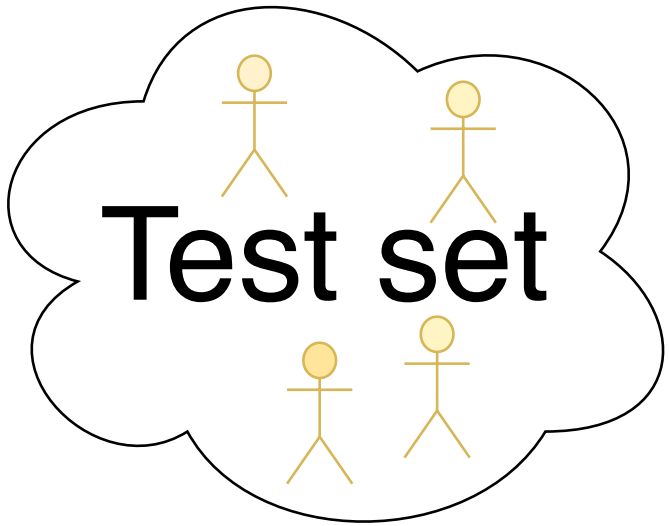
Cell lines are a simplified model of tumors

- Tumor are complex tissues
- Composed of different cell types
- Evolving in a living organism

Figure taken from *Biological Pathways Involved in Tumor Angiogenesis and Bevacizumab Based Anti-Angiogenic Therapy with Special References to Ovarian Cancer*

Why is this an issue for ML?





Lab experiments

Patients

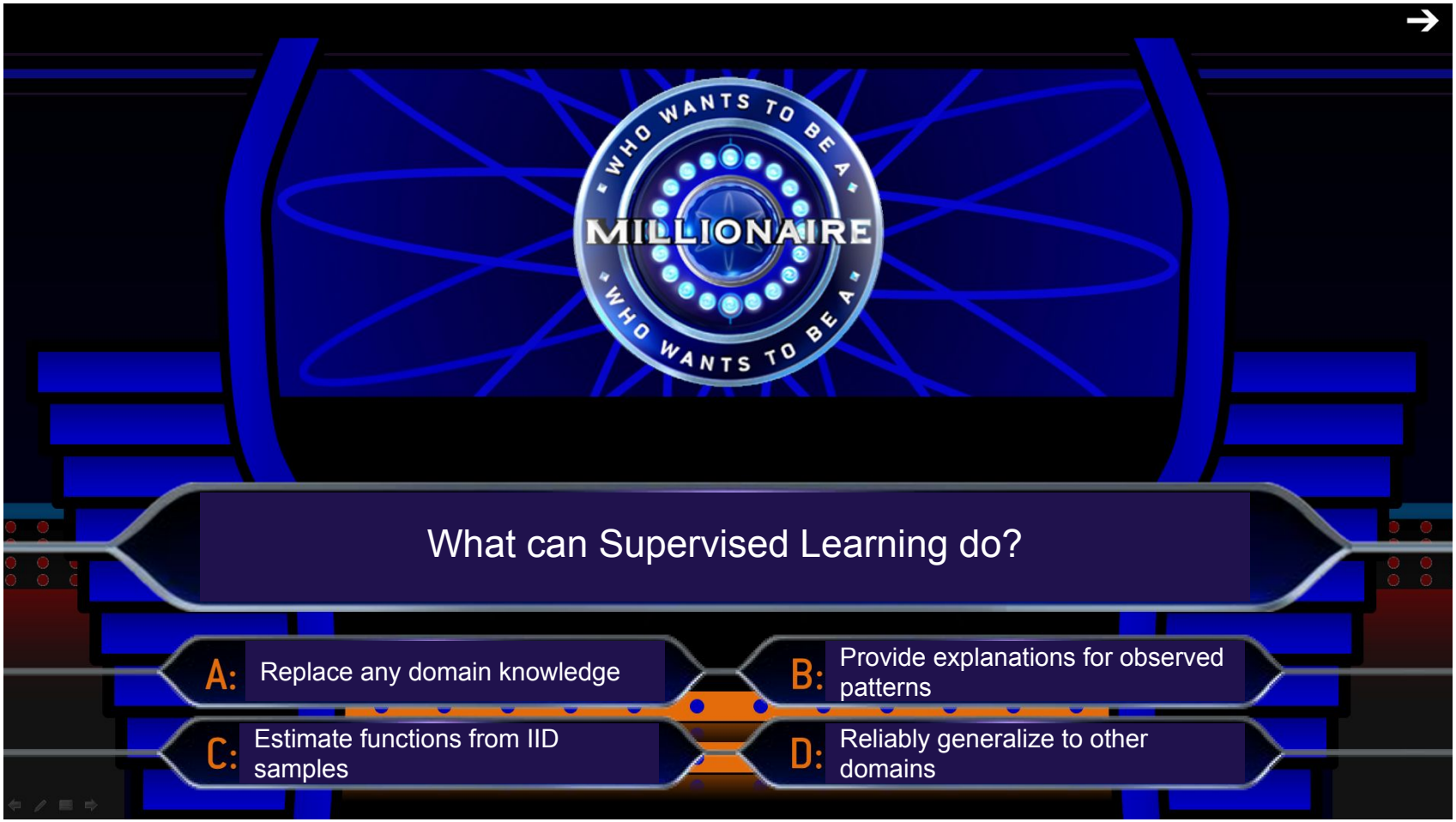
A refresher on supervised learning

Supervised learning

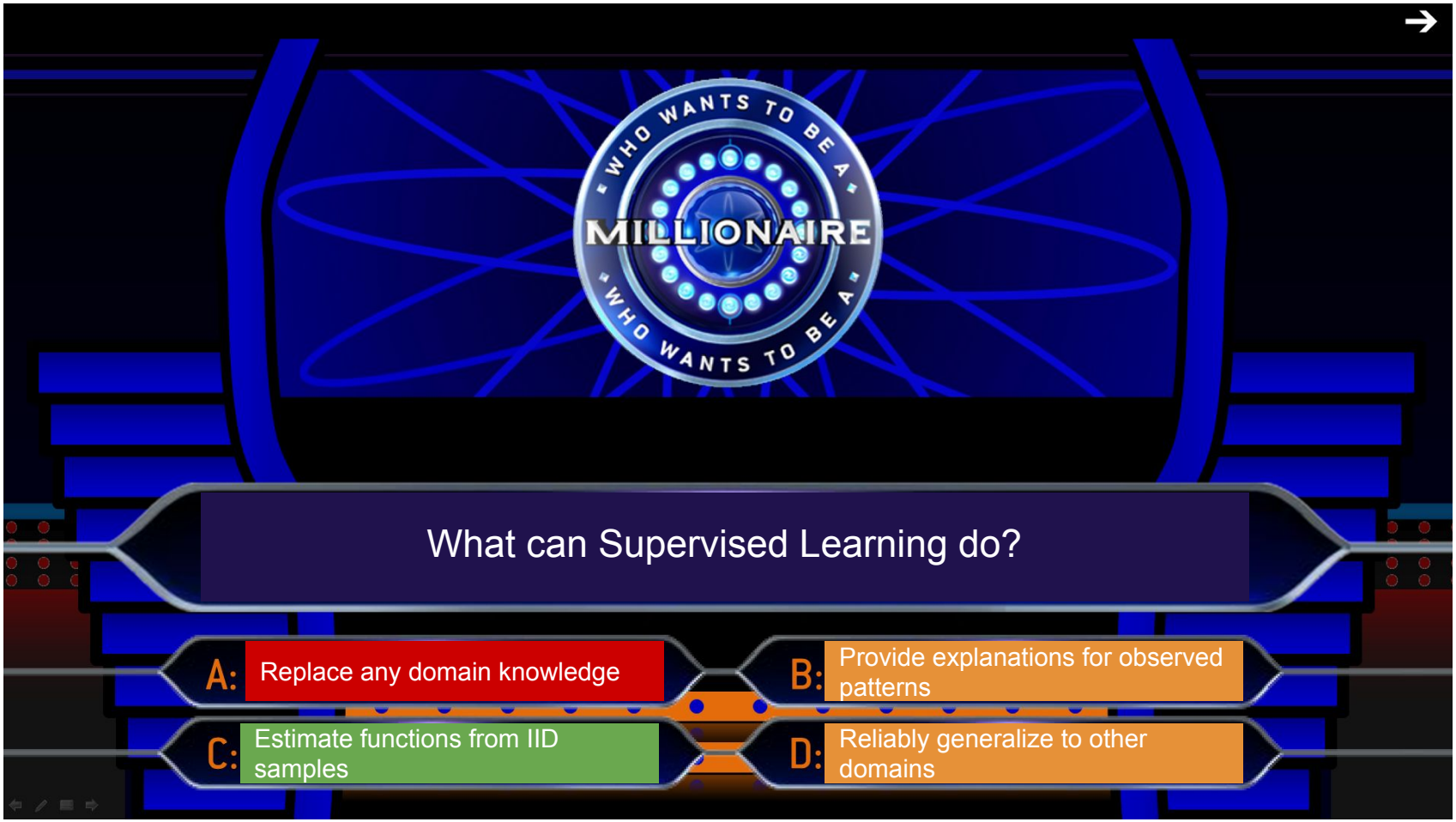
- Learn to predict target Y given input X
- We model $P_\theta(x|y)$ and learn the parameters θ based on pairs of examples

- Questions: Are there any assumptions on the dataset?

Let's have a quiz!



Machine Learning cannot do everything!



The main assumption of Supervised Learning

- **Intuition**: Pick balls at random from the same *bag* (and put the ball back before picking another one)

Independent Identically Distributed

All samples are independently drawn from a fixed probability distribution

- This assumption can be violated in several ways

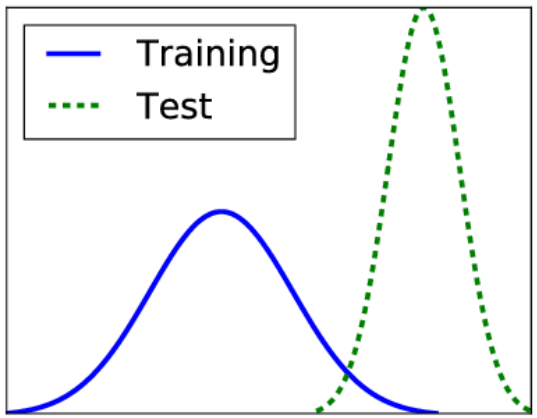


Figure: Counterexample where train and test inputs have different distributions

Covariate shift

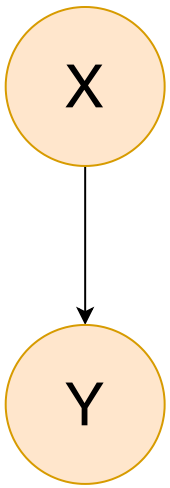
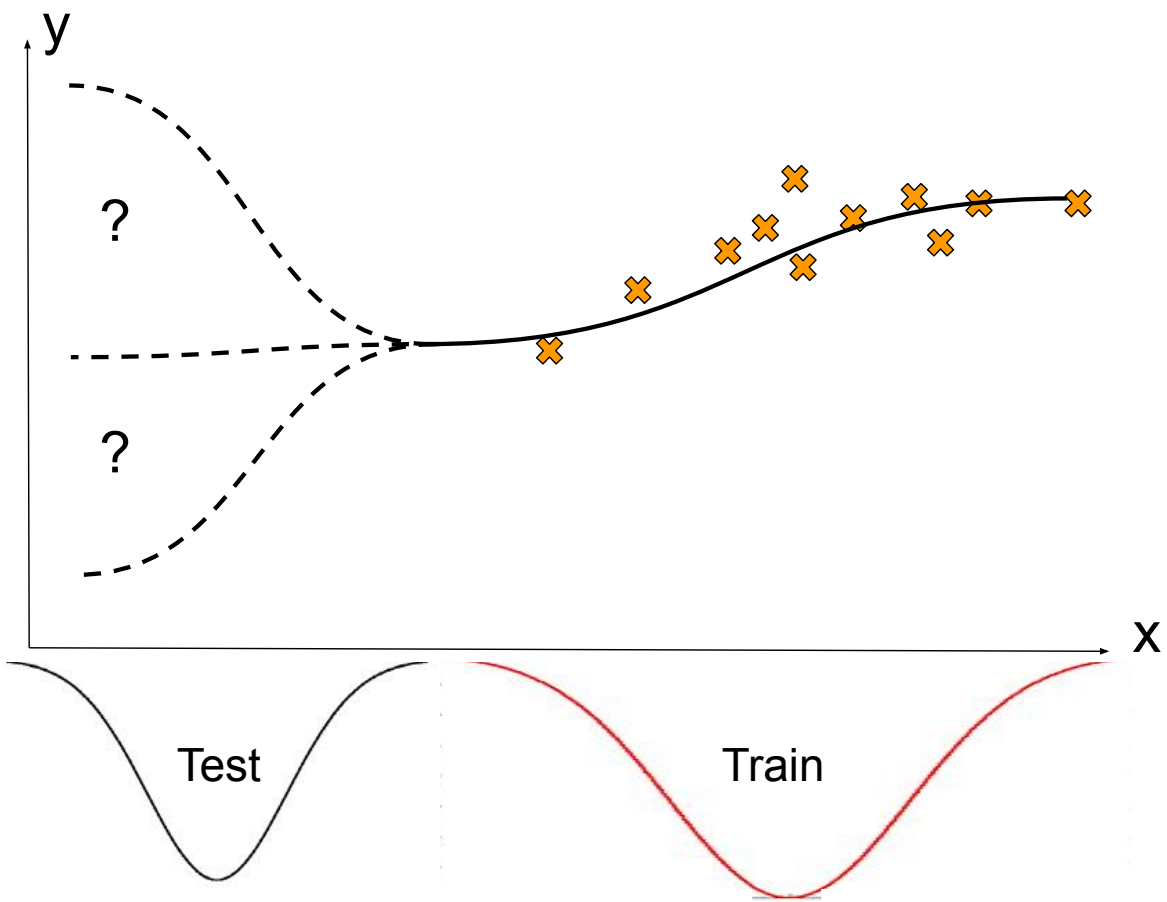


Figure: X causes Y

- Happens in X causes Y problems
- **Covariate shift**: $P(x)$ changes between train and test but $P(y|x)$ does not change
- At test time, the model will be confronted with parts of the input space that it has not seen during training

Covariate Shift



Prior probability shift

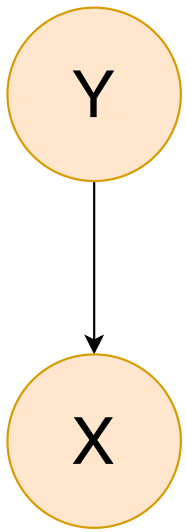


Figure: Y causes X

- Happens in Y causes X problems
- **Prior probability shift**: $P(y)$ changes between train and test but $P(x|y)$ does not change
- Difficult because both the input distribution $P(x)$ and what we model ($P(y|x)$) change

Concept shift

- **Concept shift** in X causes Y problems: $P(x)$ does not change but $P(y|x)$ changes
- **Concept shift** in Y causes X problems: $P(y)$ does not change but $P(x|y)$ changes

Different shifts together?

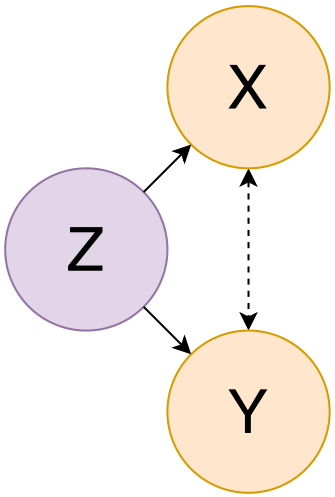


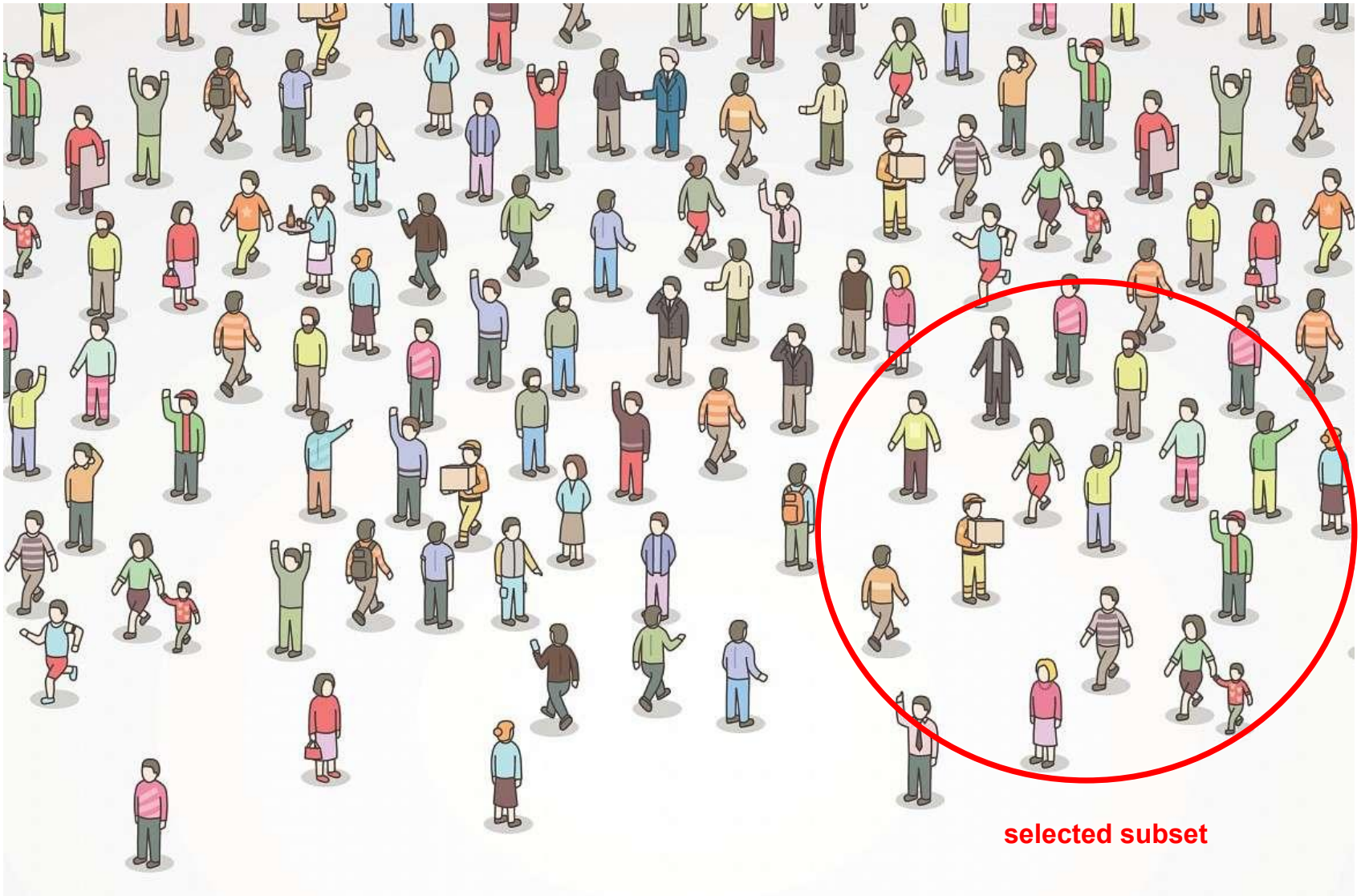
Figure: Z causes X and Y in addition to direct effects

What happens in transcriptomics?

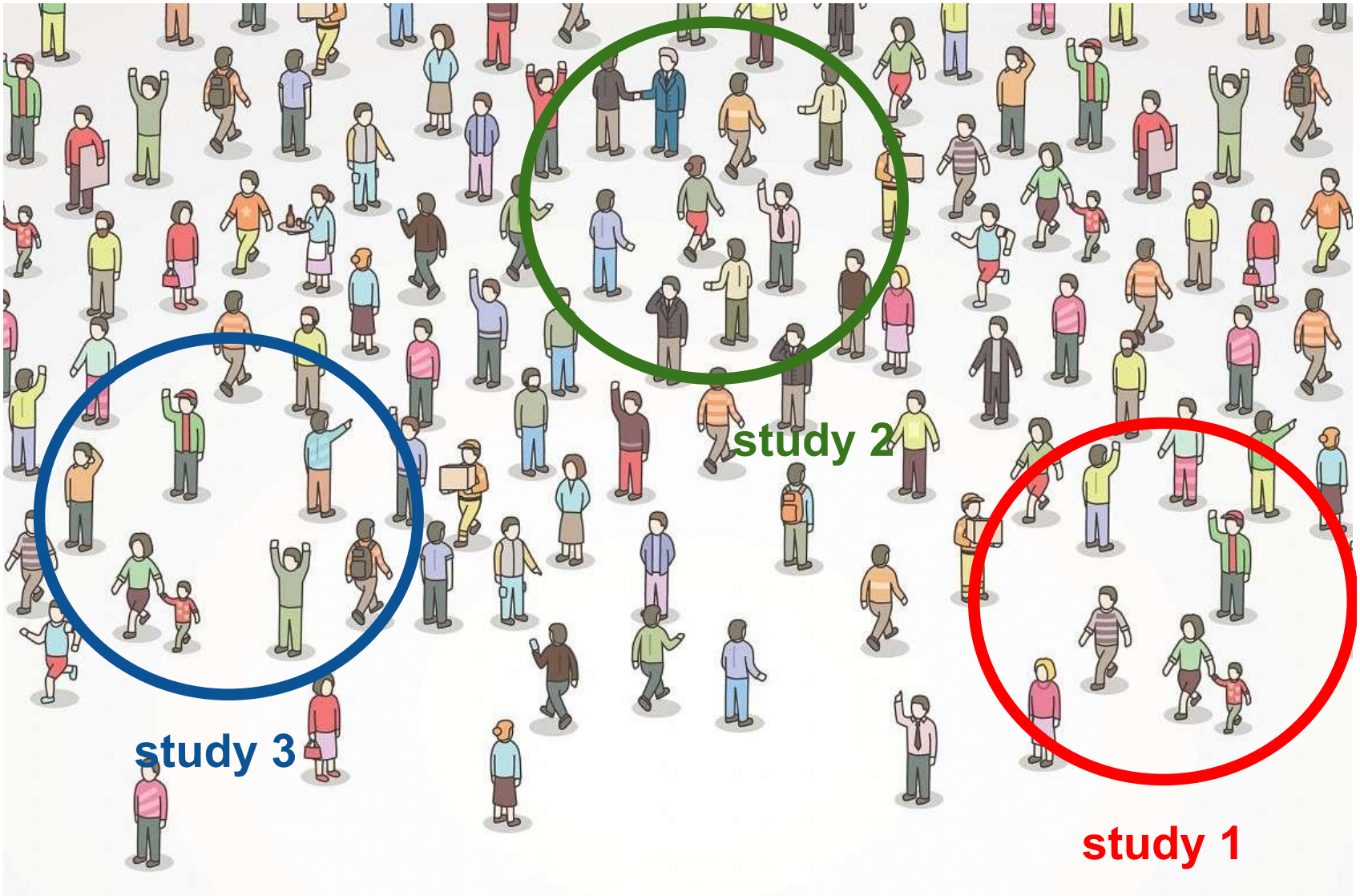
In biology, there are most certainly (very) complicated relationships between inputs and targets. Probably lots of things change together

- Examples: **covariate shift** from one individual to the other, **concept drift** from one cell type to the other.
- Quick (imperfect) fix: Normalize data
- Warning: Normalizing also means loosing information!

Selection bias



Multiple studies



Towards multi-environment learning and meta-learning?

- Other learning procedures exist and can be adapted to transcriptomics
- **Multi-environment training**: assume that data comes from different environments
- **Meta-learning**: Learn to adapt fast to a new environment

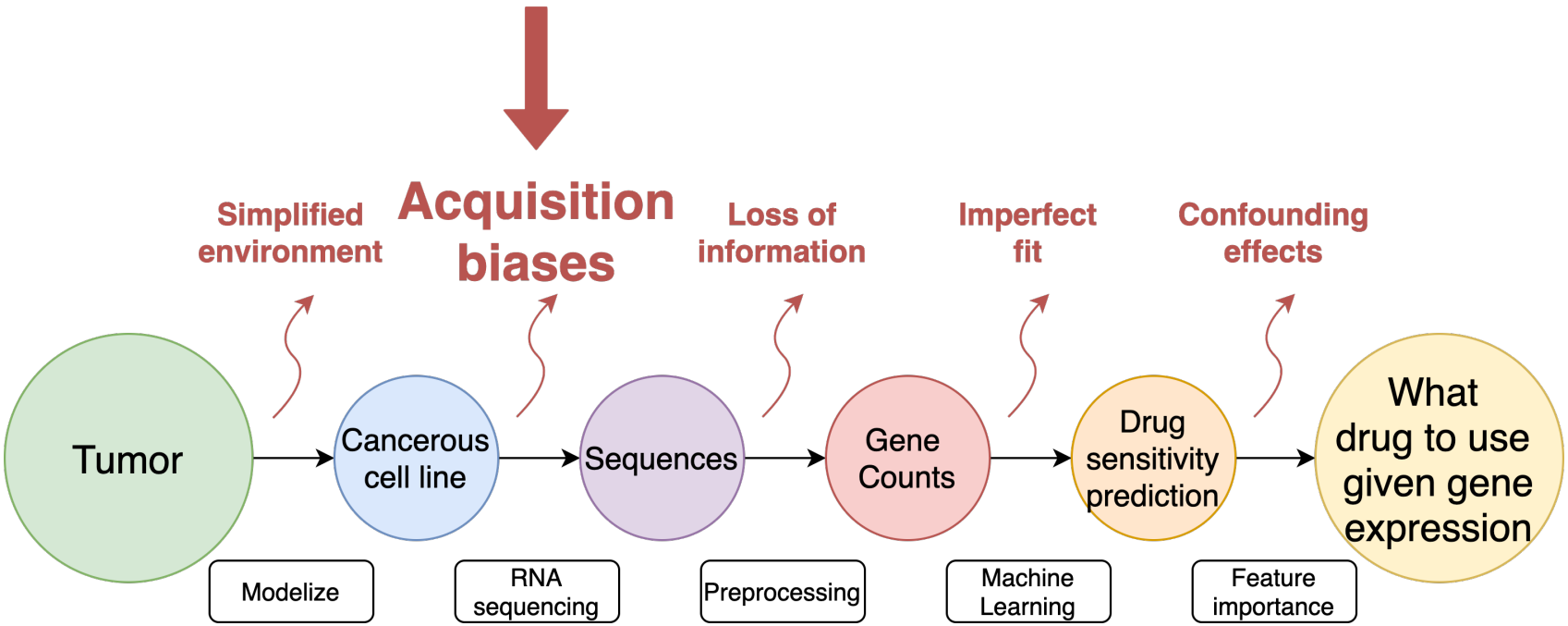
---

The TCGA Meta-Dataset Clinical Benchmark

Mandana Samiei[1] **Tobias Würfl**[2] **Tristan Deleu**[3] **Martin Weiss** [3]
Francis Dutil[5] **Thomas Fevens**[1] **Geneviève Boucher**[3,4] **Sebastien Lemieux**[3,4]
Joseph Paul Cohen[3]

Outline

Overview


Acquisition biases
Simplified environment
Loss of information
Imperfect fit
Confounding effects
Tumor
Cancerous cell line
Sequences
Gene Counts
Drug sensitivity prediction
What drug to use given gene expression
Modelize
RNA sequencing
Preprocessing
Machine Learning
Feature importance

The acquisition process

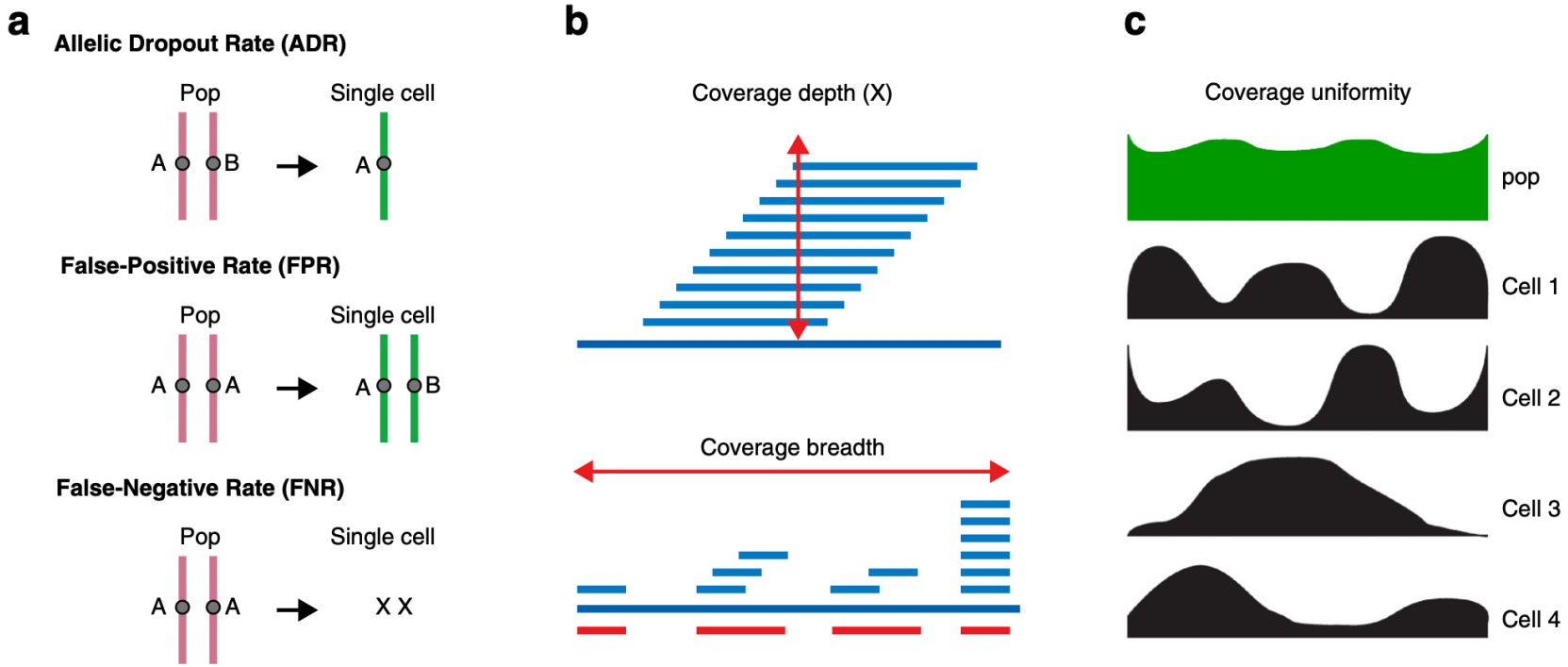
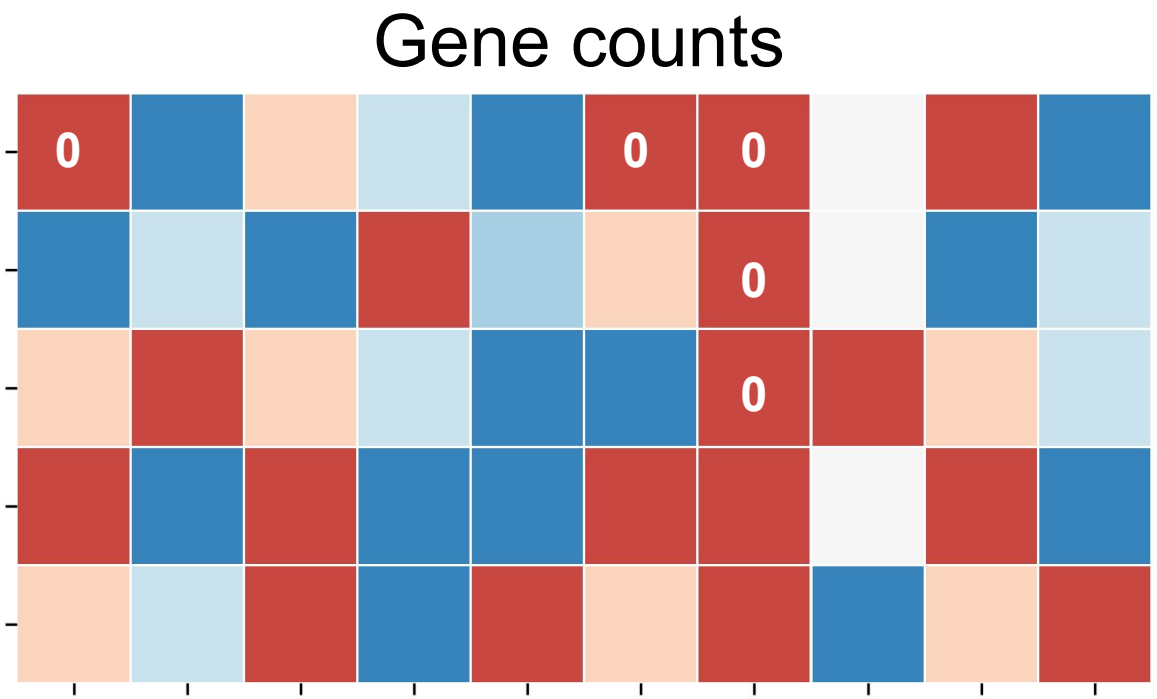


Figure 3 Technical errors and coverage in single-cell sequencing data. (a) Technical errors that occur in single-cell sequencing (SCS) data include: false-positive errors, allelic dropout events and false-negative errors due to insufficient coverage. 'Pop' indicates a population of cells. **(b)** Coverage metrics in SCS data include coverage depth and total physical coverage, or breadth. **(c)** Coverage uniformity, or 'eveness' in SCS data can vary from cell to cell, but is often more uniform in standard genomic DNA sequencing experiments using populations of cells.

Figure taken from *Cancer genomics: one cell at a time*

Dropout in single cell



Denoising Autoencoder

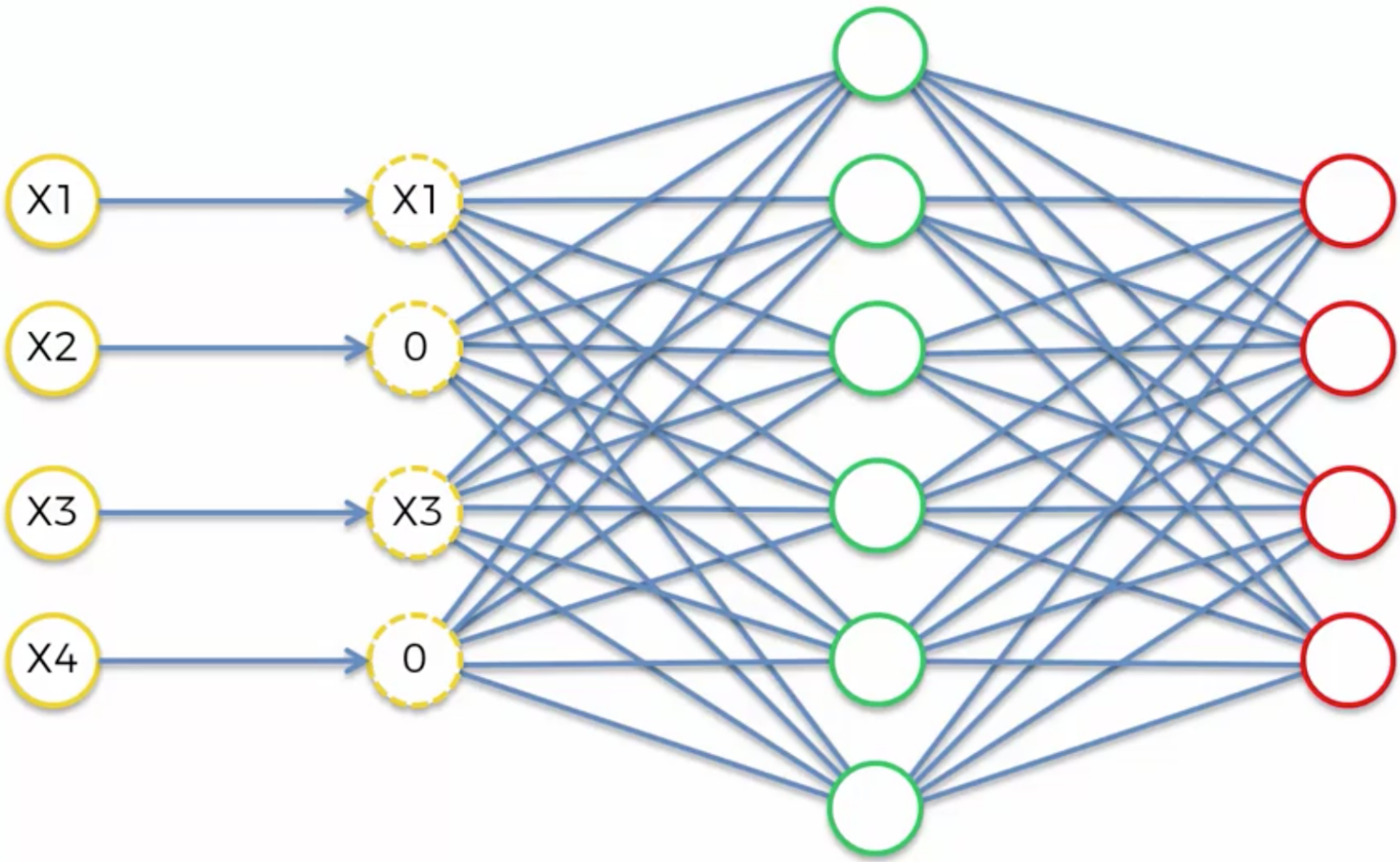
- How to denoise the data?

Extracting and Composing Robust Features with Denoising Autoencoders

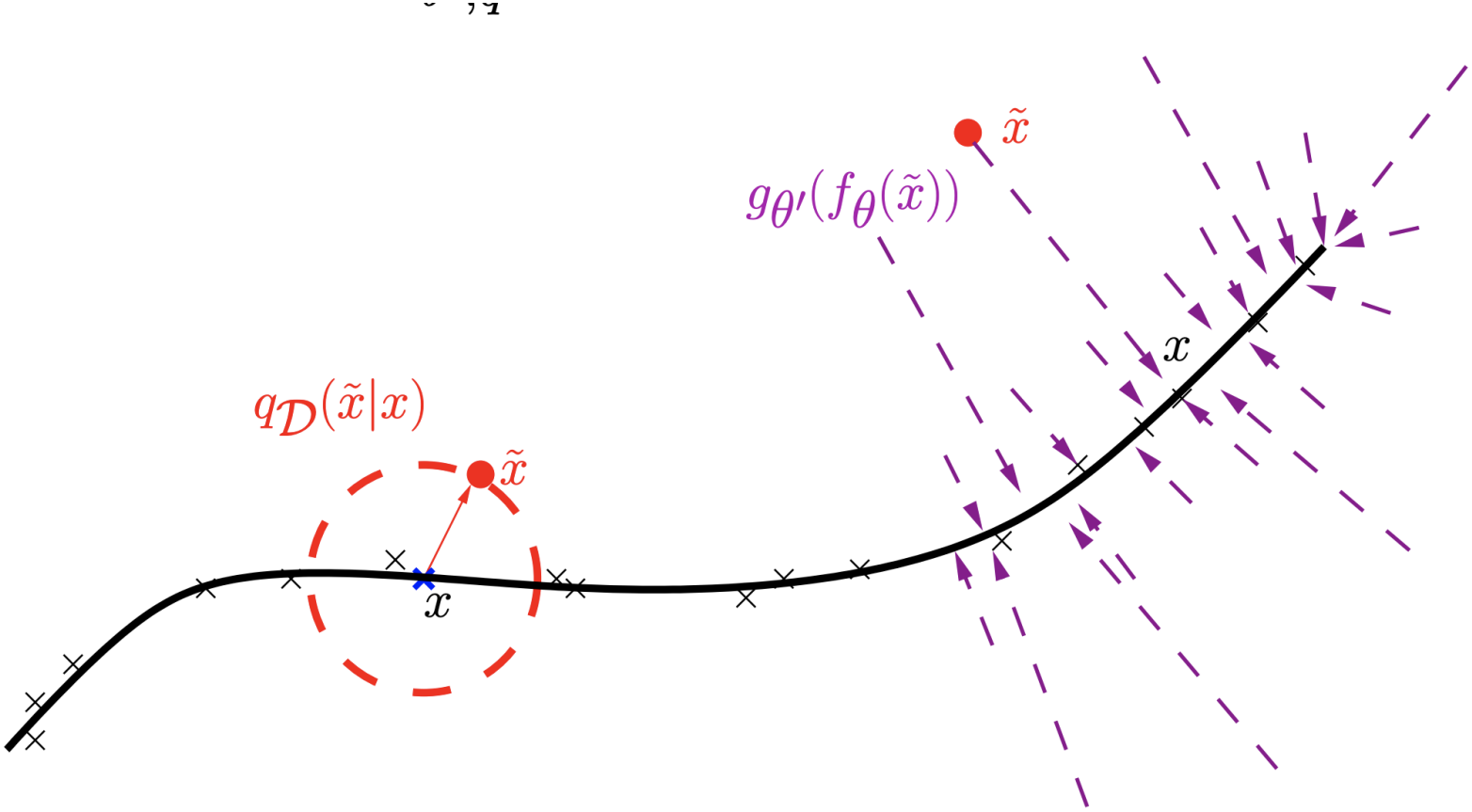
---

Pascal Vincent VINCENTP@IRO.UMONTREAL.CA
Hugo Larochelle LAROCHEH@IRO.UMONTREAL.CA
Yoshua Bengio BENGIOY@IRO.UMONTREAL.CA
Pierre-Antoine Manzagol MANZAGOP@IRO.UMONTREAL.CA
Université de Montréal, Dept. IRO, CP 6128, Succ. Centre-Ville, Montral, Qubec, H3C 3J7, Canada

Denoising Autoencoder



Denoising Autoencoder



Tan *et al. BMC Bioinformatics* (2017) 18:512
DOI 10.1186/s12859-017-1905-4

BMC Bioinformatics

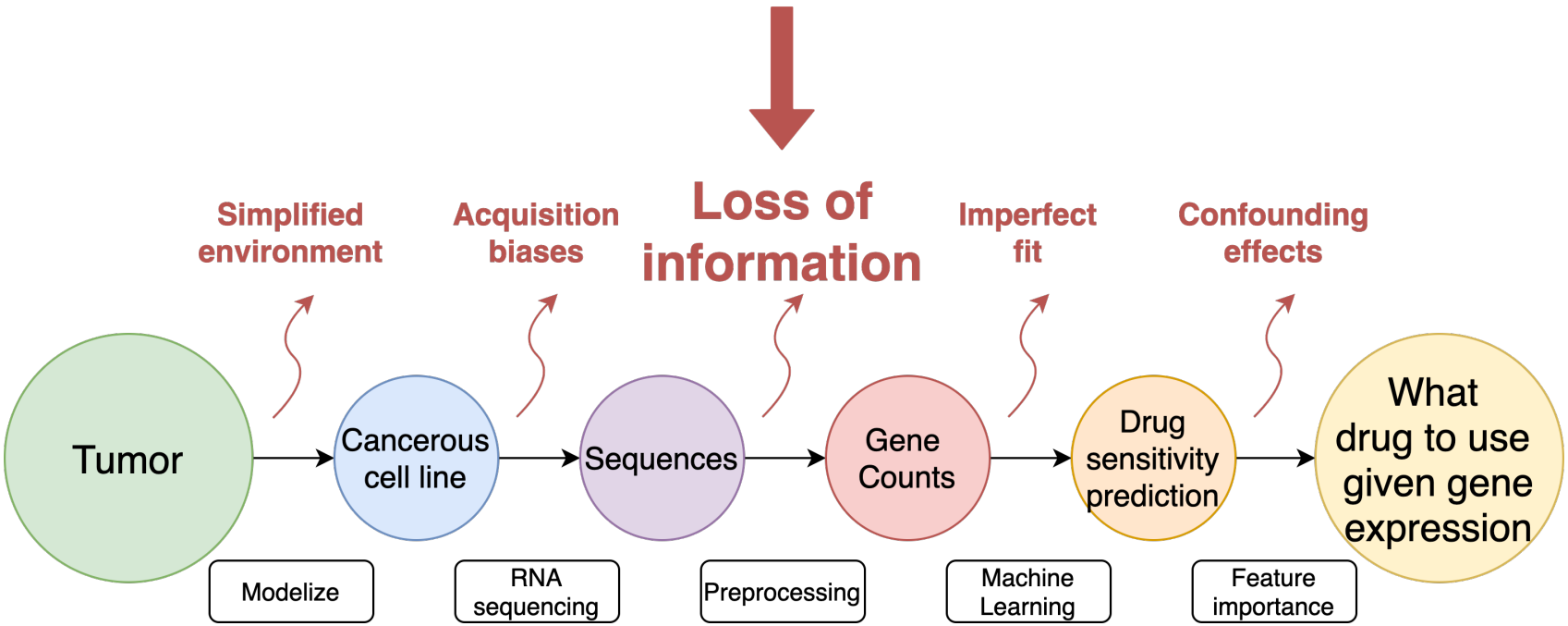
METHODOLOGY ARTICLE Open Access

ADAGE signature analysis: differential expression analysis with data-defined gene sets

Jie Tan[1], Matthew Huyck[2,3], Dongbo Hu[2], René A. Zelaya[2], Deborah A. Hogan[3] and Casey S. Greene[2*]

Outline

Overview


Loss of information
Simplified environment
Acquisition biases
Imperfect fit
Confounding effects
Tumor
Cancerous cell line
Sequences
Gene Counts
Drug sensitivity prediction
What drug to use given gene expression
Modelize
RNA sequencing
Preprocessing
Machine Learning
Feature importance

The acquisition process

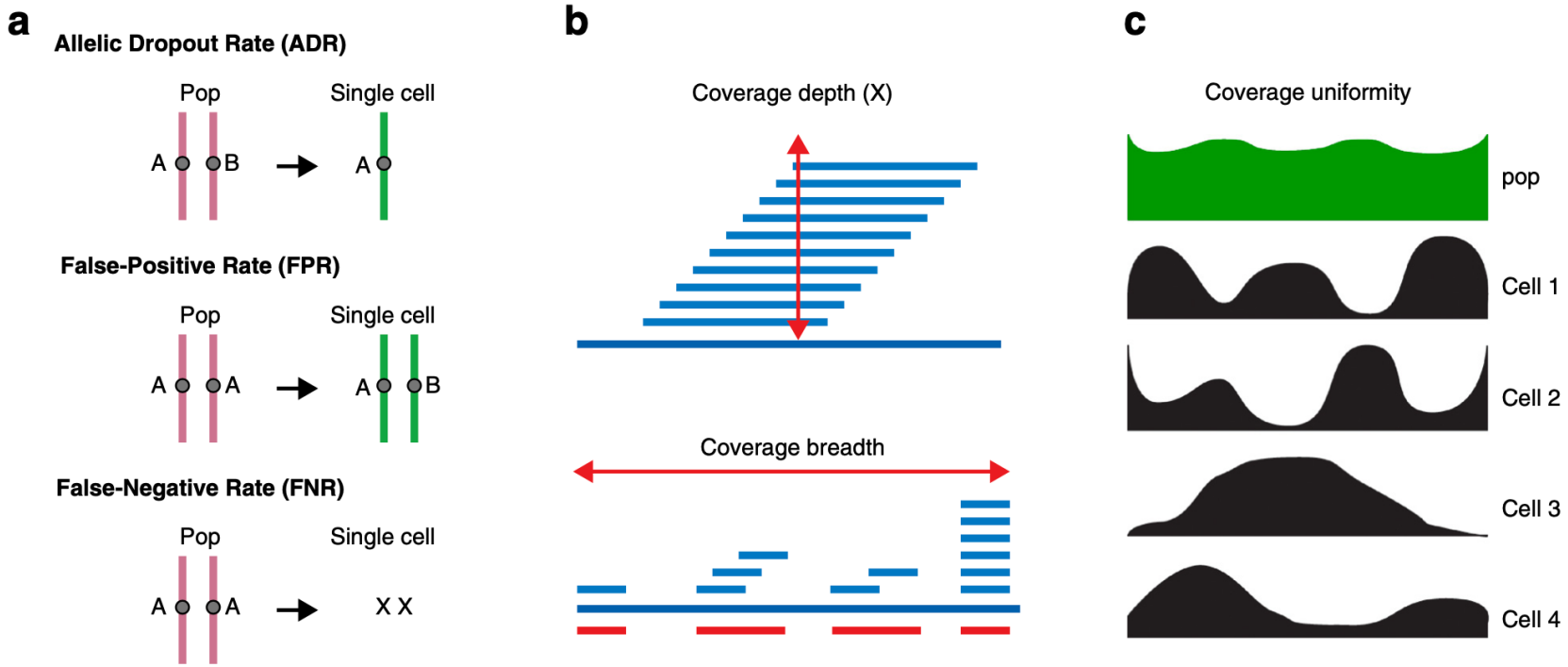


Figure 3 Technical errors and coverage in single-cell sequencing data. (a) Technical errors that occur in single-cell sequencing (SCS) data include: false-positive errors, allelic dropout events and false-negative errors due to insufficient coverage. 'Pop' indicates a population of cells. **(b)** Coverage metrics in SCS data include coverage depth and total physical coverage, or breadth. **(c)** Coverage uniformity, or 'eveness' in SCS data can vary from cell to cell, but is often more uniform in standard genomic DNA sequencing experiments using populations of cells.

Figure taken from *Cancer genomics: one cell at a time*

Different preprocessings

A key step

Preprocessing is a key step that determines what data will be fed to our machine learning model. Each technique comes with limitation and drawbacks

- Need to account for different **total amounts of reads** in the different samples
- Need to account for the **lengths of the genes**
- Classic normalization methods: RPKM (Read Per Kilobase Million), FPKM (Fragment Per Kilobase Million), TPM (Transcripts Per Kilobase Million)
- **Alignment**: different reference genomes can be used from one dataset to the other!

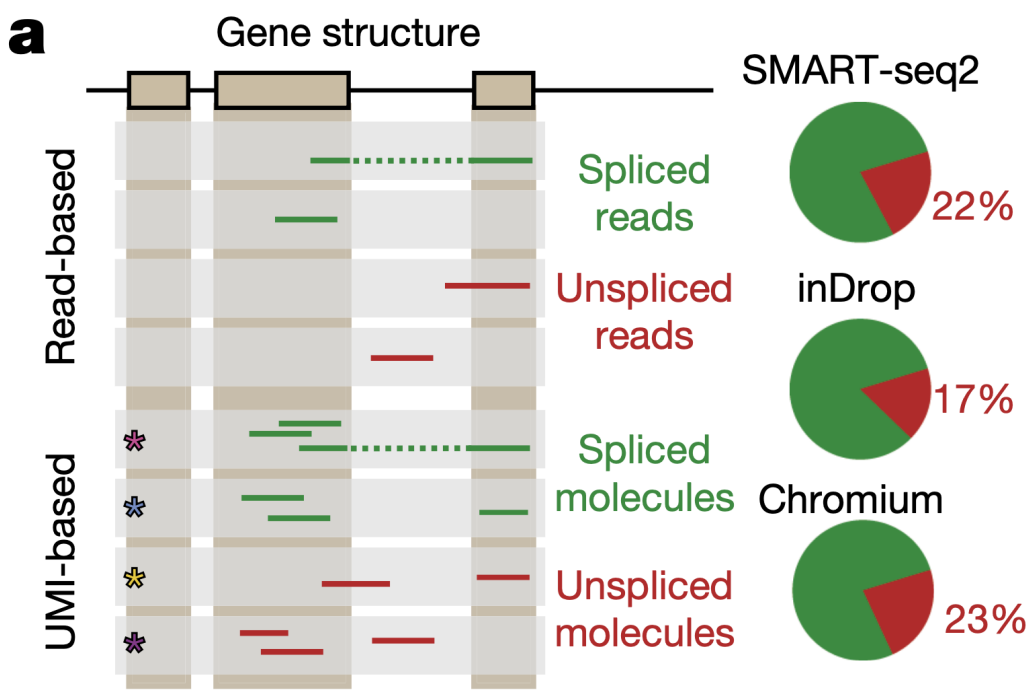
Tremendous information loss!

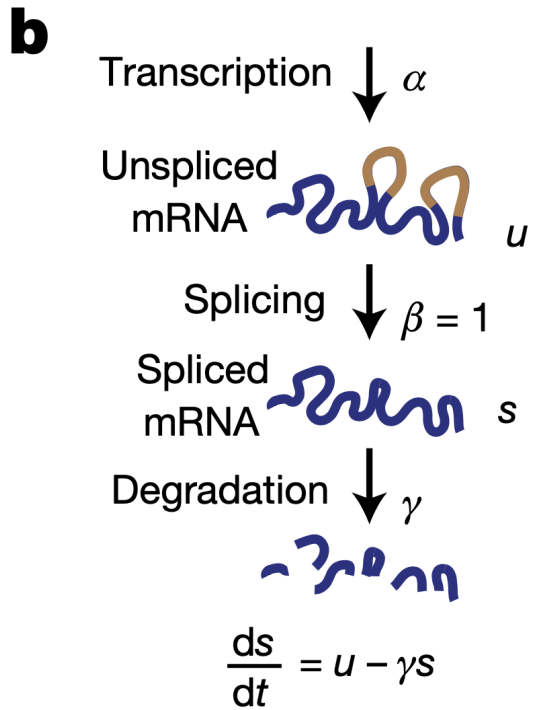
Towards a more clever preprocessing?

A lot of **information is lost** during those preprocessing steps, **limiting what can be achieved downstream**. Moreover, the non standardized normalization and alignment limit our ability to transfer knowledge from one dataset to the other.

- Could we do better?
- Example: RNA velocity inference using splicing information.

RNA velocity





you can visit `https://scvelo.org`

Figure taken from *RNA velocity of single cells*

RNA velocity

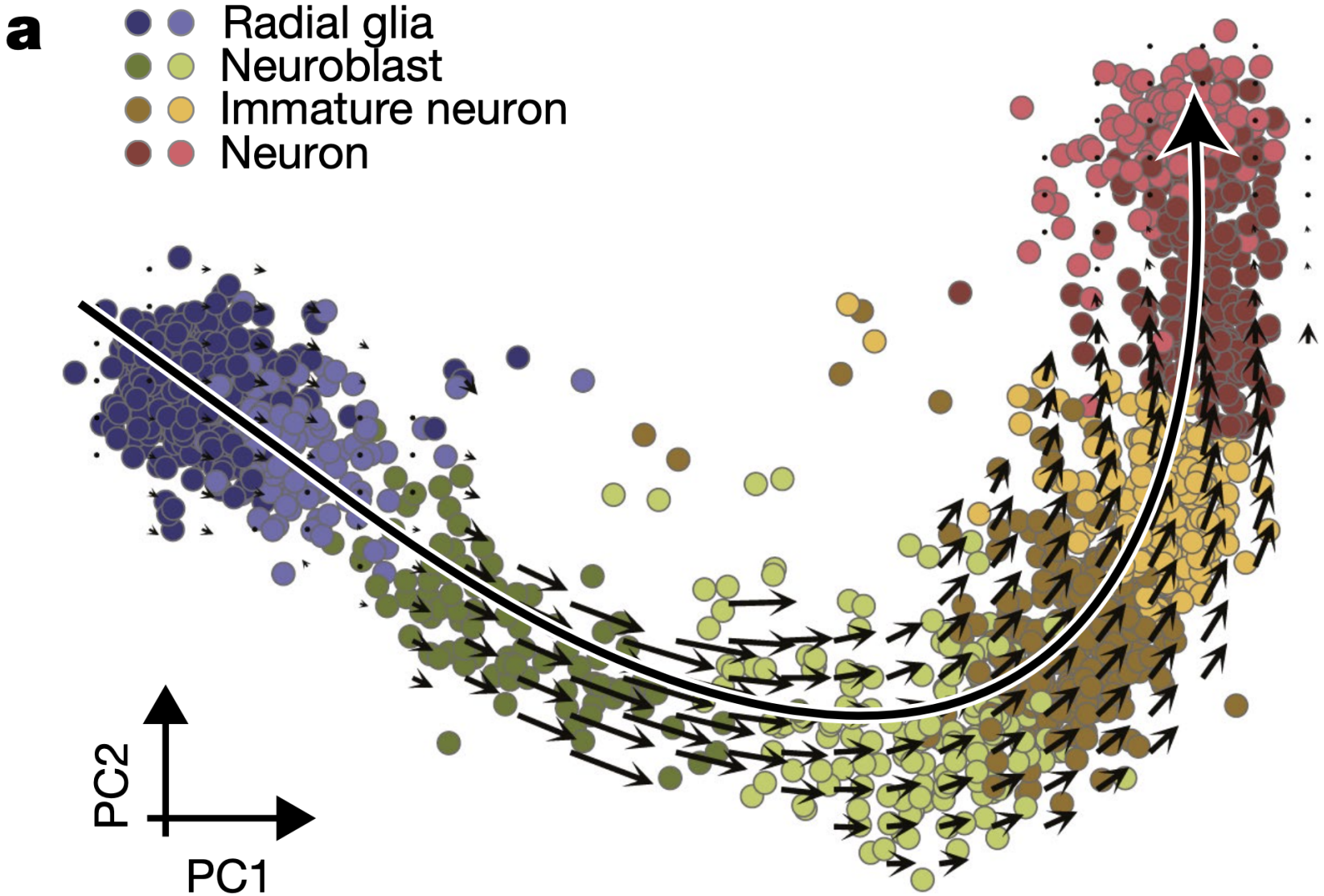
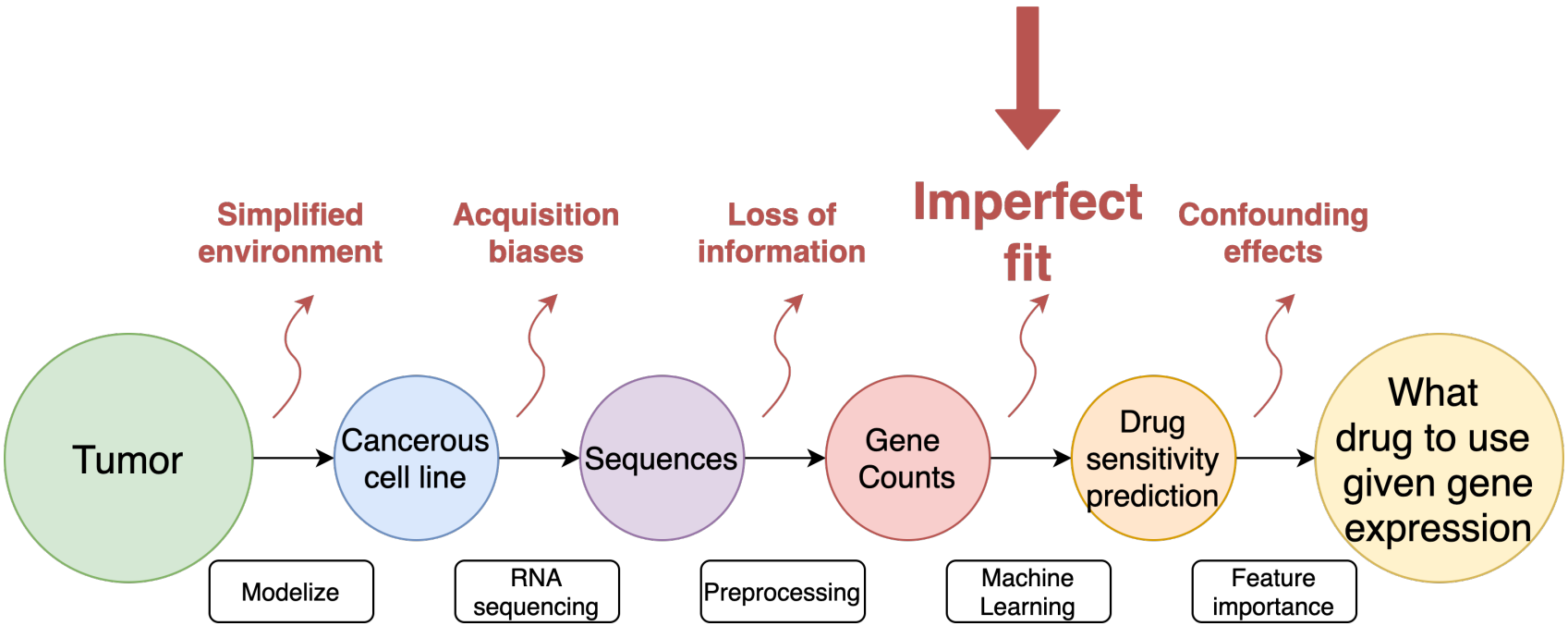


Figure taken from *RNA velocity of single cells*

Outline

Overview


Simplified environment
Acquisition biases
Loss of information
Imperfect fit
Confounding effects
Tumor
Cancerous cell line
Sequences
Gene Counts
Drug sensitivity prediction
What drug to use given gene expression
Modelize
RNA sequencing
Preprocessing
Machine Learning
Feature importance

Outline

Fat data

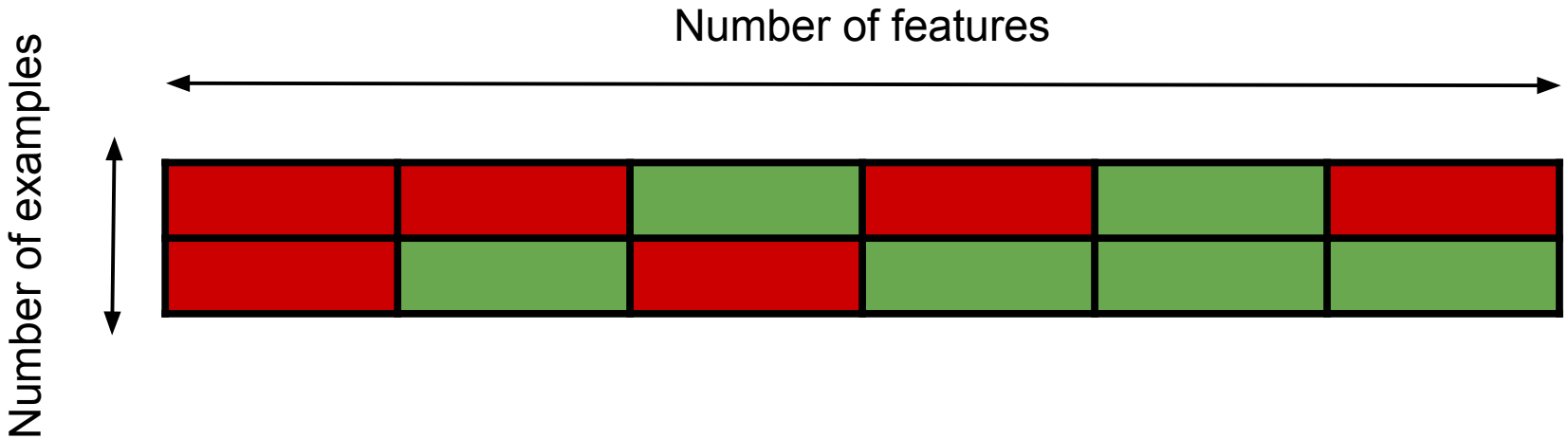
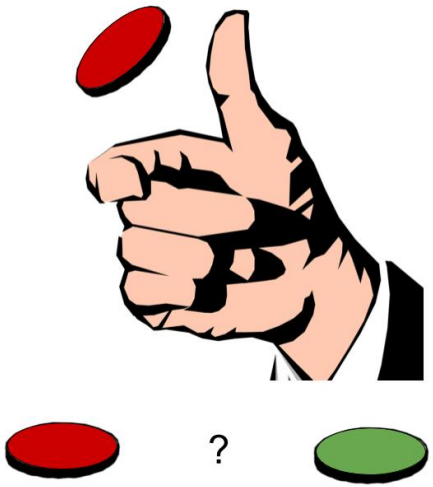
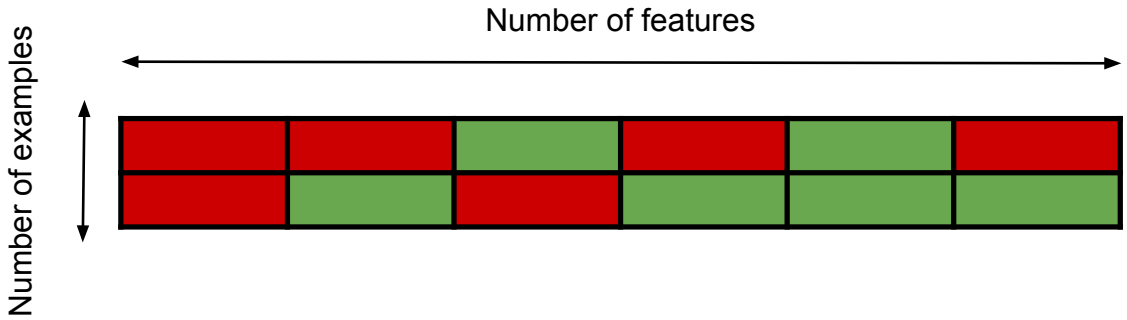


Figure: When data has lots of feature but few examples, data is called fat

Fat data: Beware of spurious correlations!



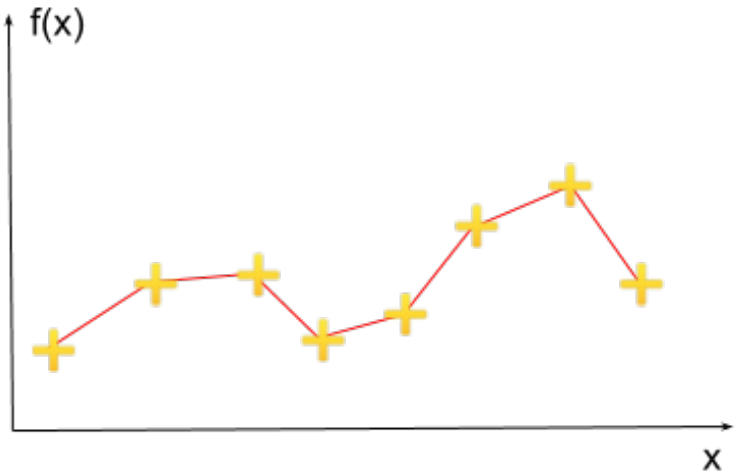


- **Spurious correlations**: with fat data, features can be highly correlated together out of chance!
- Example: binary, independent features and 2 samples. Some features will have a correlation of 1 out of chance!

Conclusion

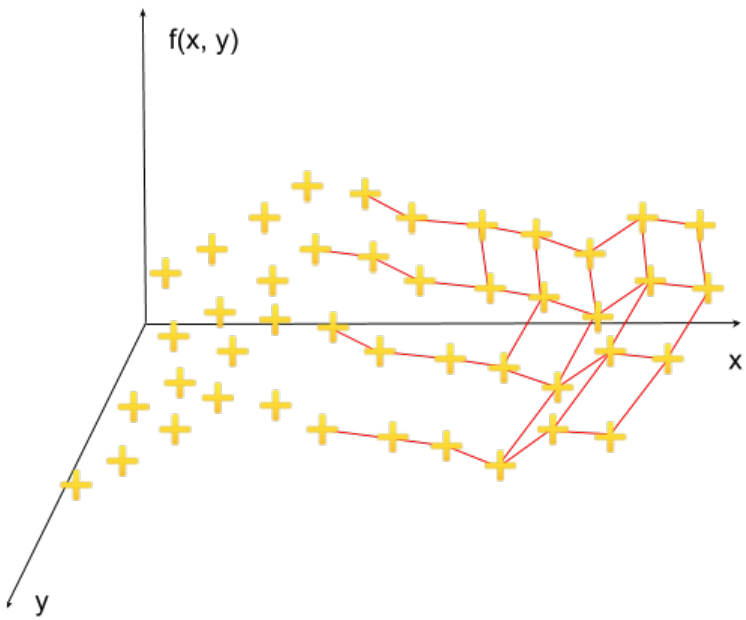
In high dimension, you need lots of examples!

Estimate a function of 1 variable



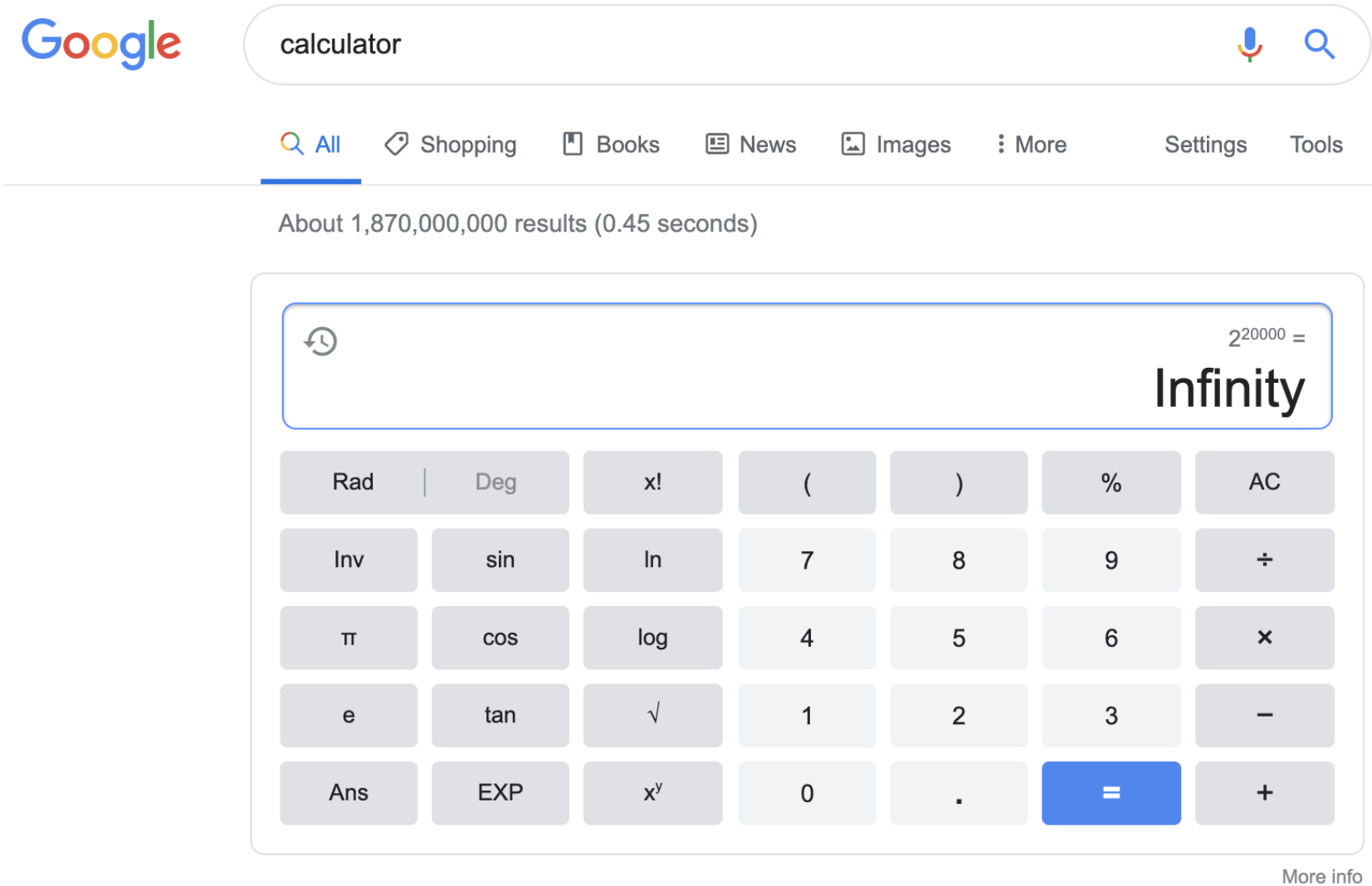
- We want a *high* density of samples (distance between two sample points $< \frac{1}{N}$) in order to estimate the function reliably
- N samples to estimate the function on $[0, 1]$

Estimate a function of 2 variables



- To have the same density of sample in 2 dimensions we need N^2 samples
- In dimension 3, we need N^3 samples...

And with 20,000 dimensions?



Volumes in high dimension

Cover the space in d dimensions

The volume of a hypercube $[0, a] \times [0, a] \times ... \times [0, a]$ of dimension d is :

$$\mathbf{a^d}$$

In high dimension, volumes are very big!

- This is why **kNN does not work in high dimension**
- How to estimate a function in dimension 20k?
- **Machine learning is about making the right assumptions** to overcome the need for many samples.

Outline

Automatic feature extraction

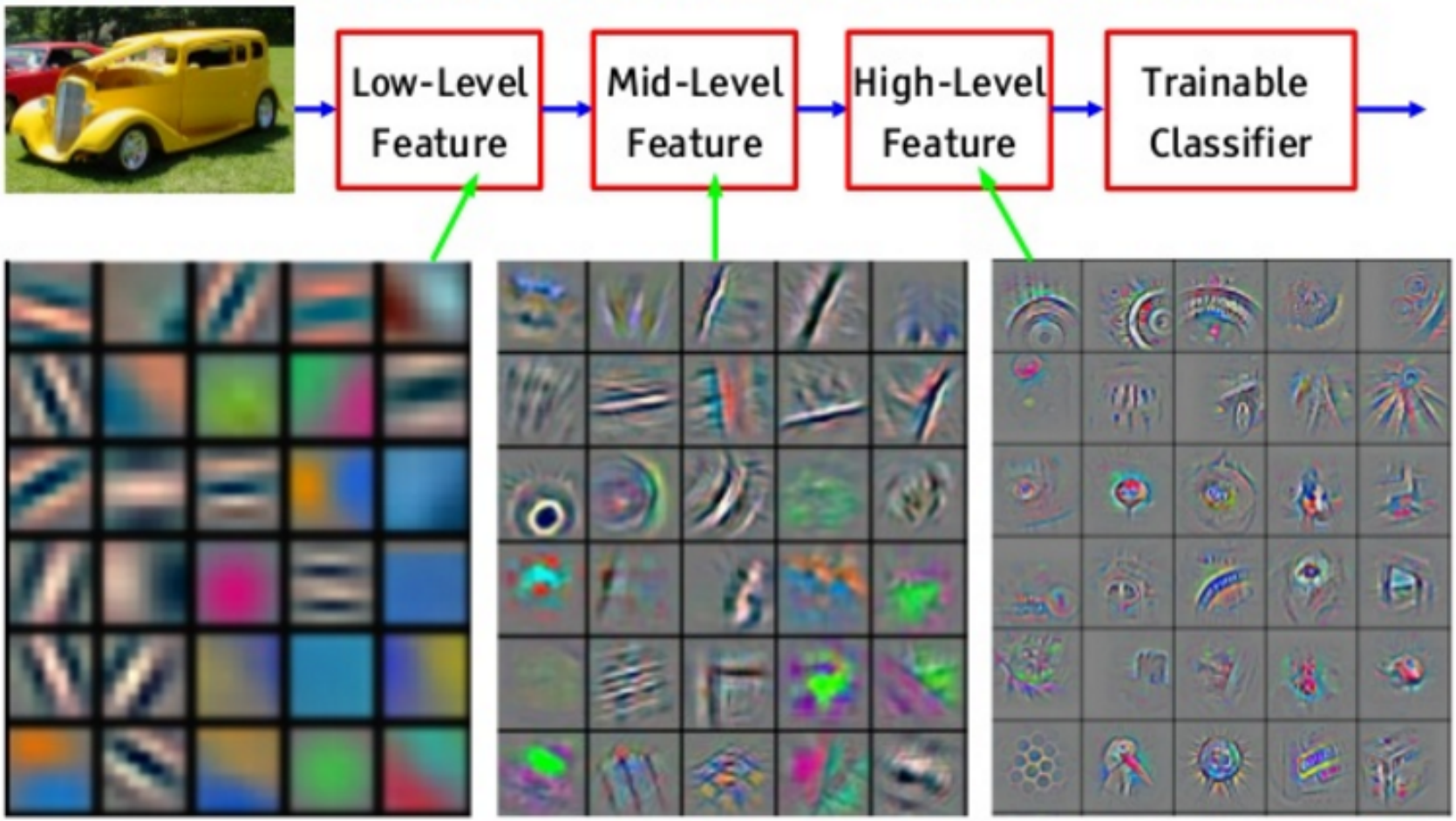
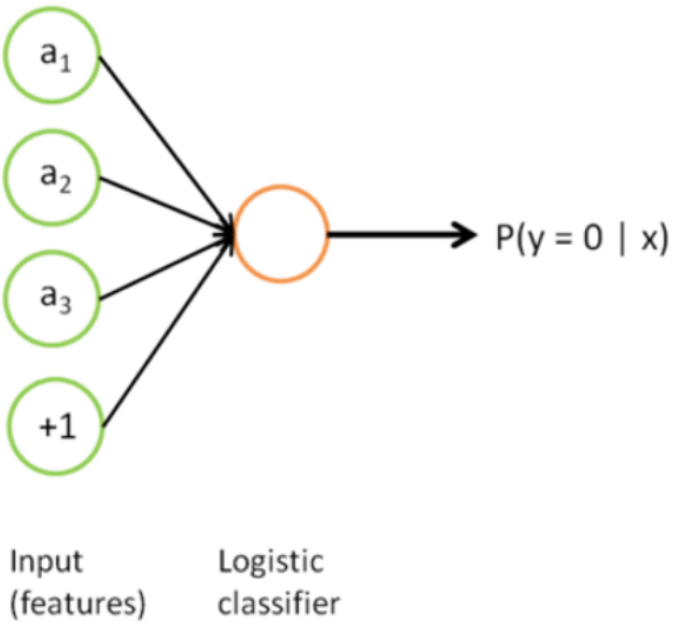
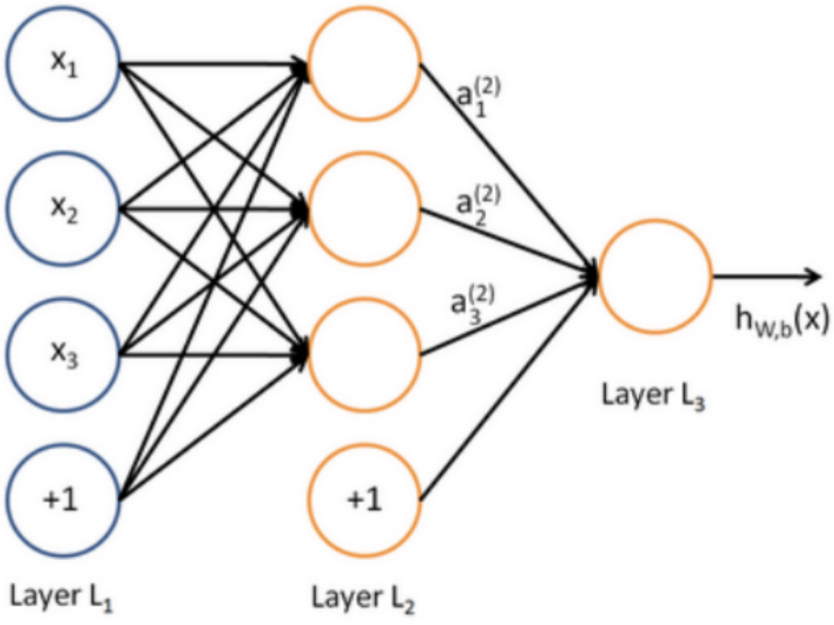


Figure by Y. LeCun

Comparison with logisitic regression



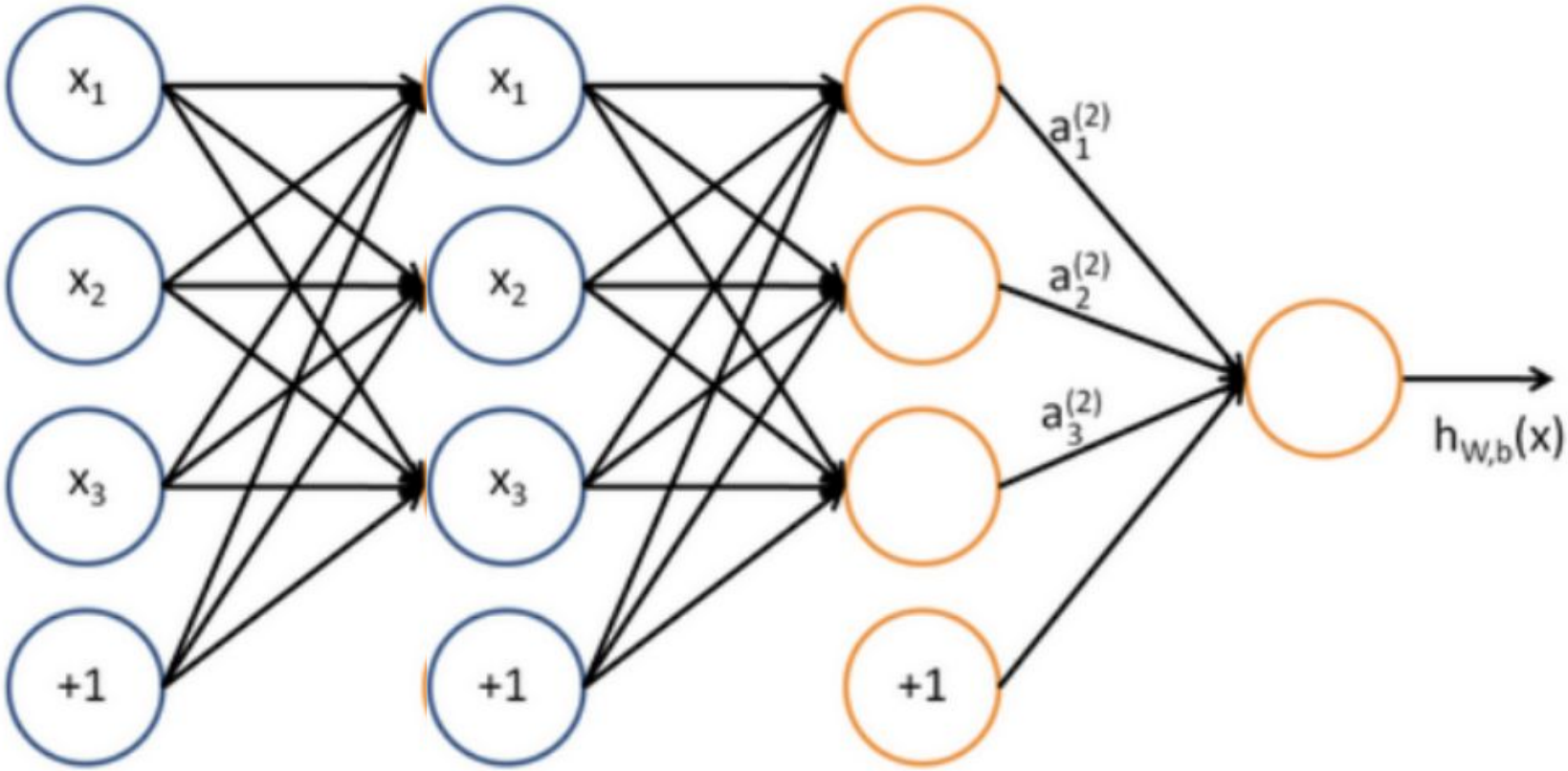
Logistic Regression



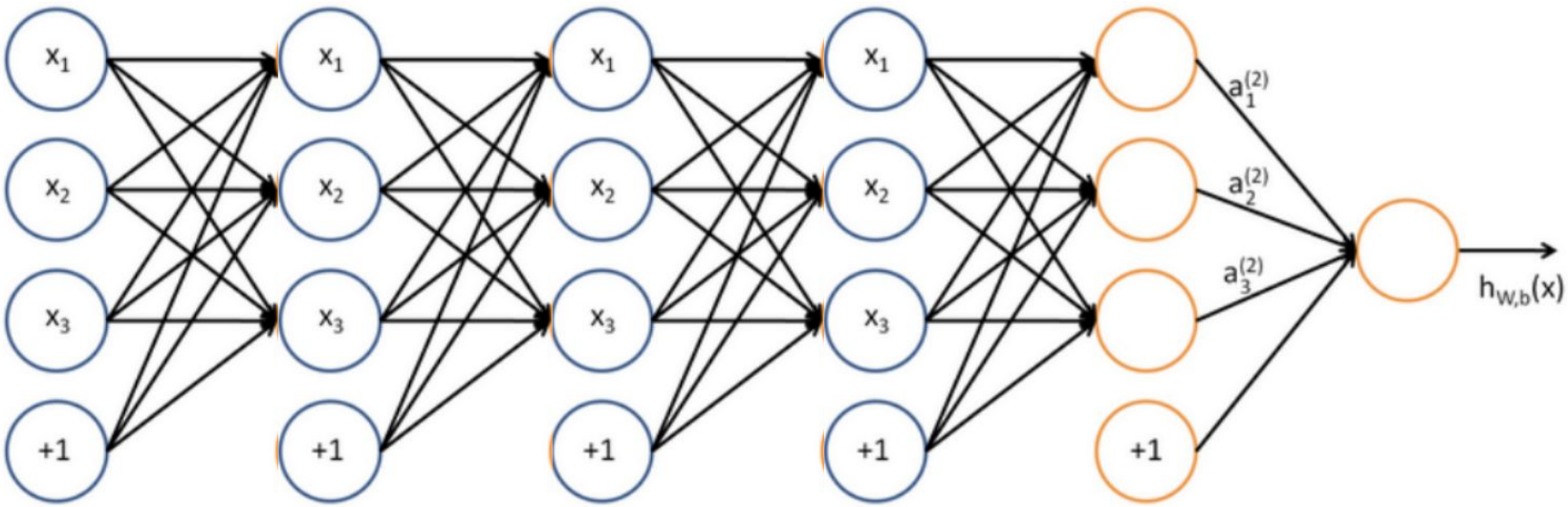
Neural Network

- 1 hidden layer NN: the features fed to the logistic regression are learnt

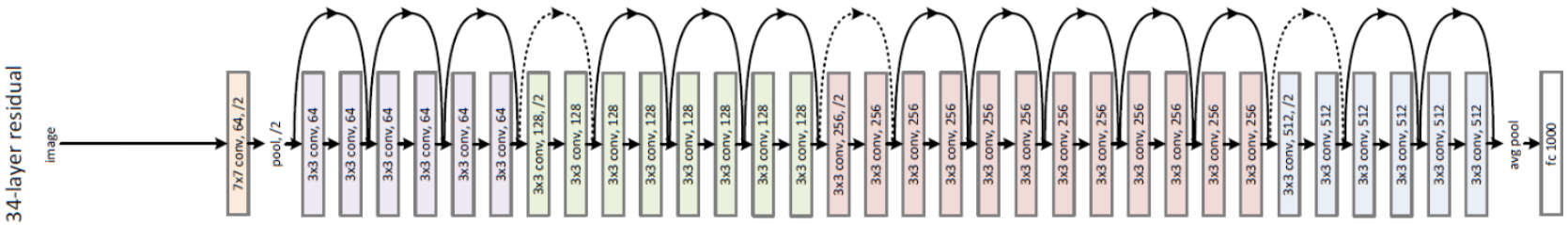
Going deeper!



Going deeper!

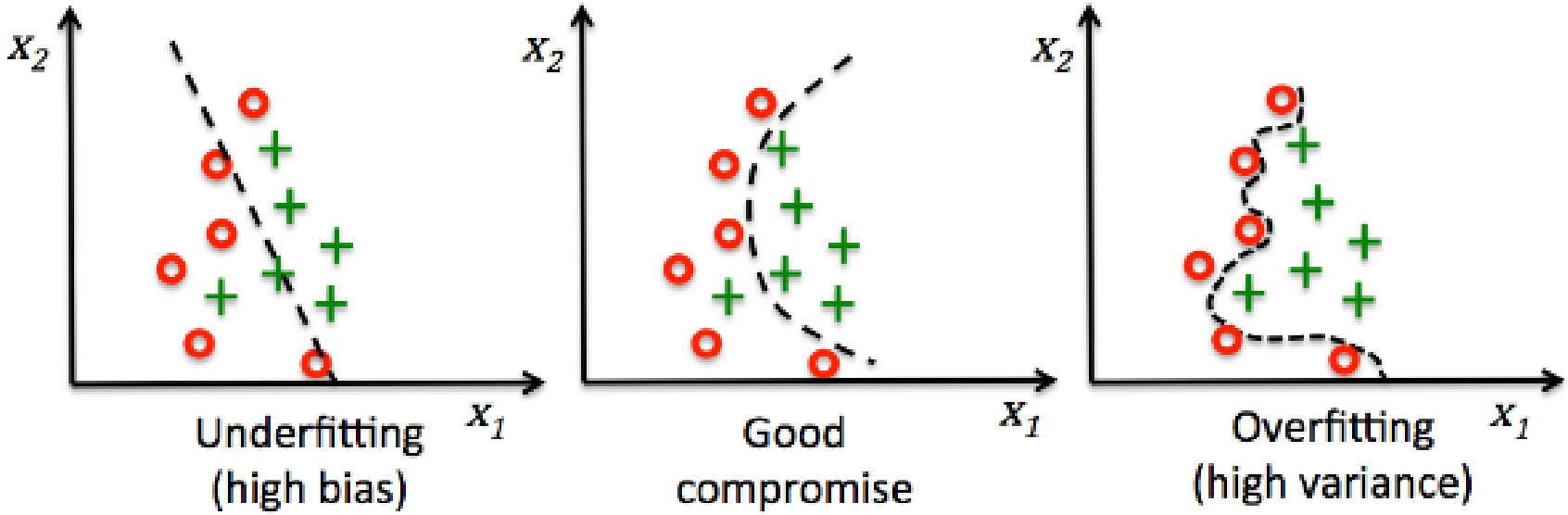


Going deeper!



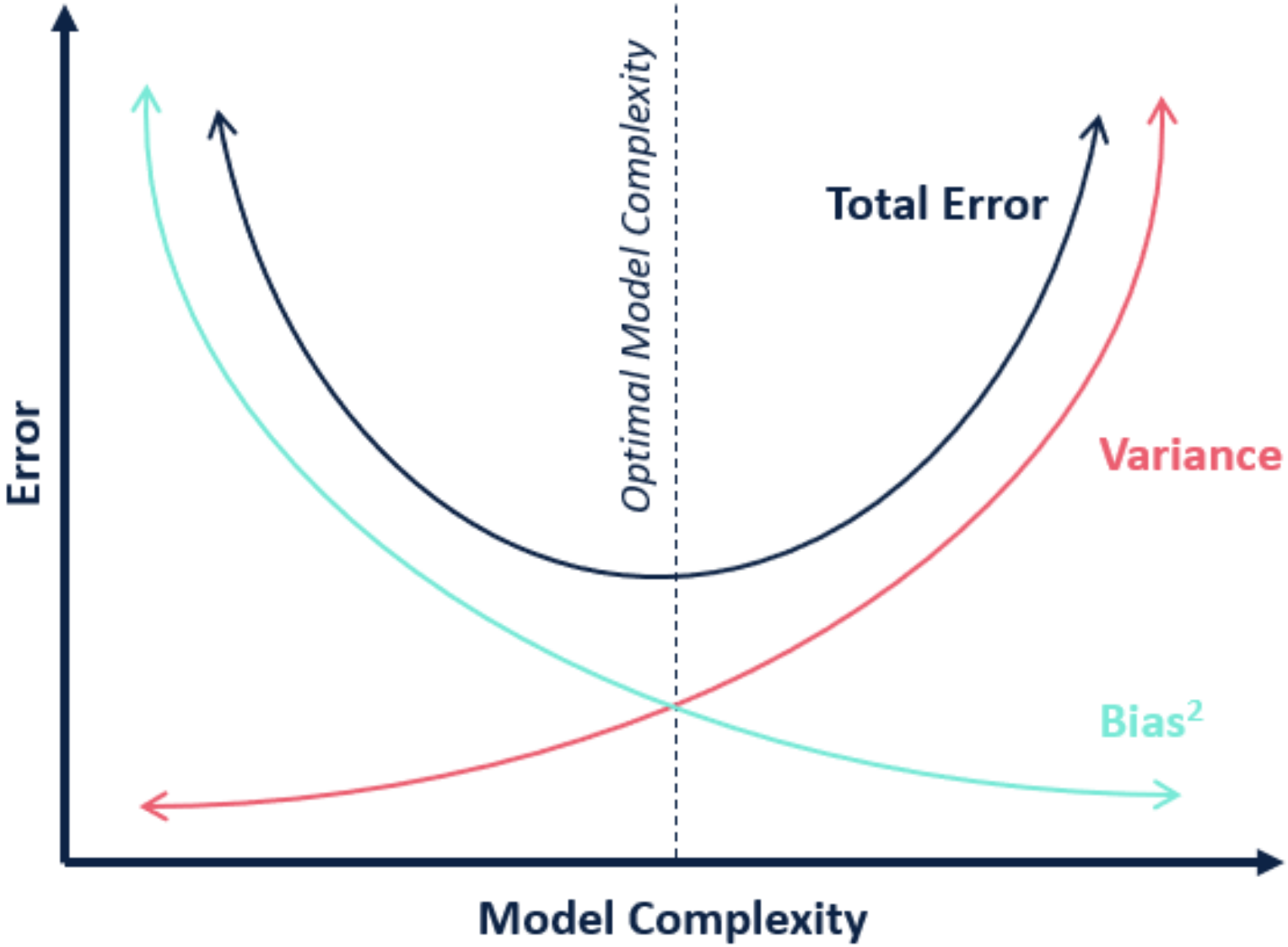
- Example: **Resnet**, a 34 layer network!
- Needs additional trick (residuals) for forward and backward signals to pass through

The bias variance trade-off



- High model complexity: high variance low bias
- Low model complexity: low variance high bias
- **You have to choose the right model complexity!** (regularization, model depth,...)

The bias variance trade-off



Why does ML work so well in Computer Vision?

Why does ML work so well in Computer Vision?

People have made made simplifying assumptions that hold well in Computer Vision

- Let us dive into the details of CNNs

Fully connected

- X_L and X_{L+1} activations vectors in layers L and $L+1$
- $X_{L+1} = \sigma(W_L X_L + B)$

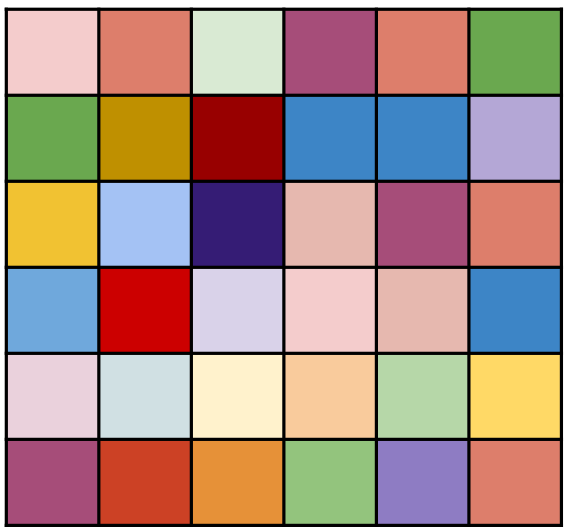
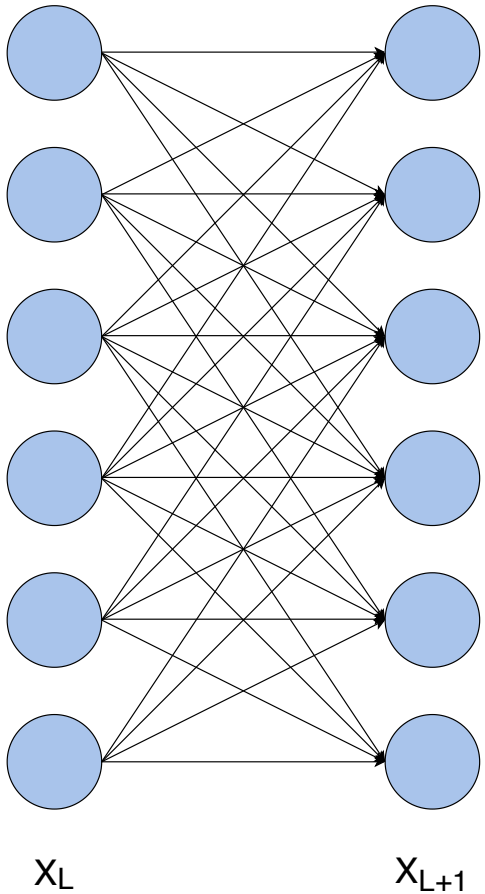
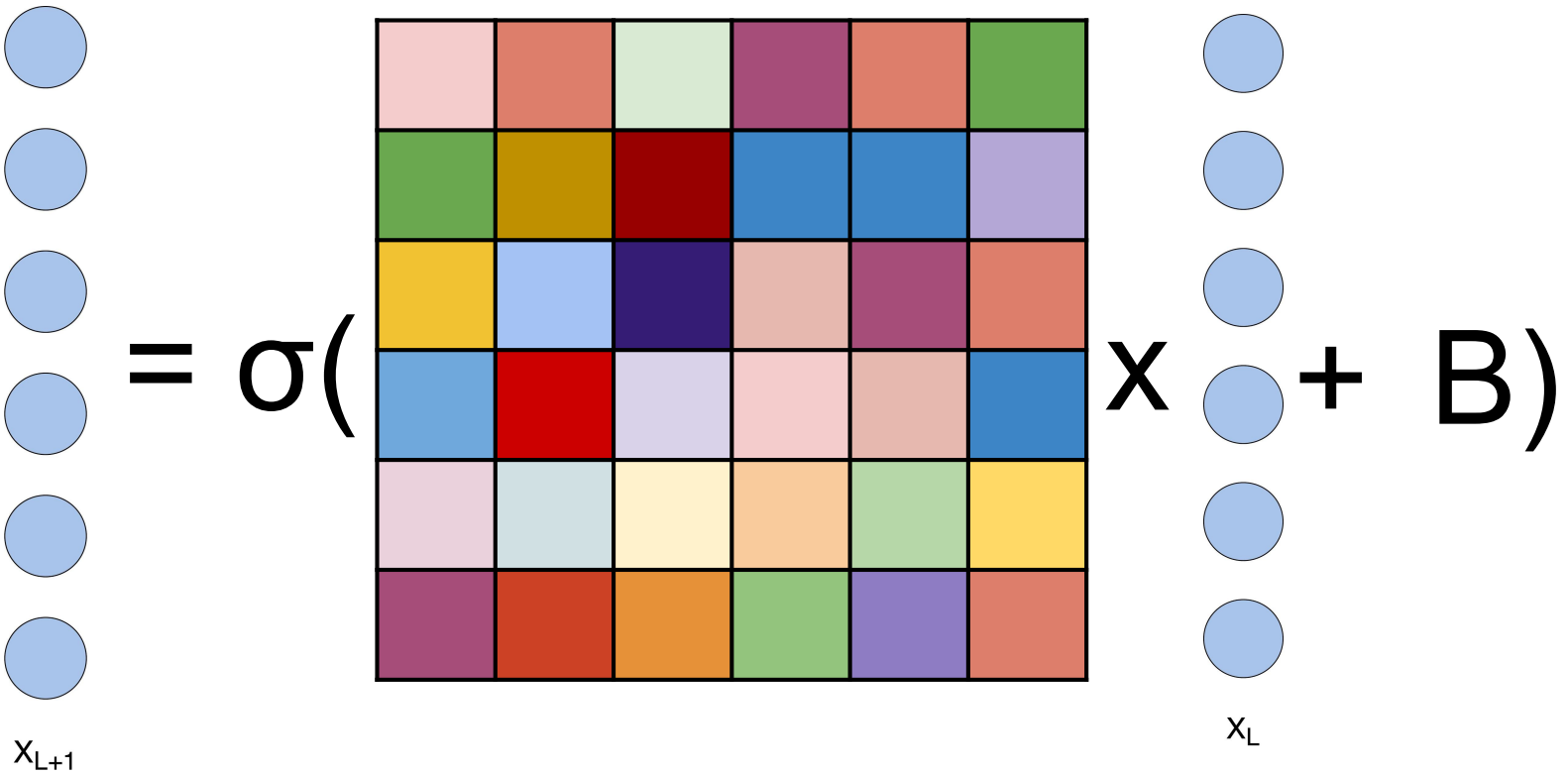


Figure: matrix W_L



Fully connected



Convolutional layer

- $X_{L+1} = \sigma(W_L X_L + B)$
- parameter sharing : constraints on W_L

0.2	-0.7				
1.4	0.2	-0.7			
	1.4	0.2	-0.7		
		1.4	0.2	-0.7	
			1.4	0.2	-0.7
				1.4	0.2

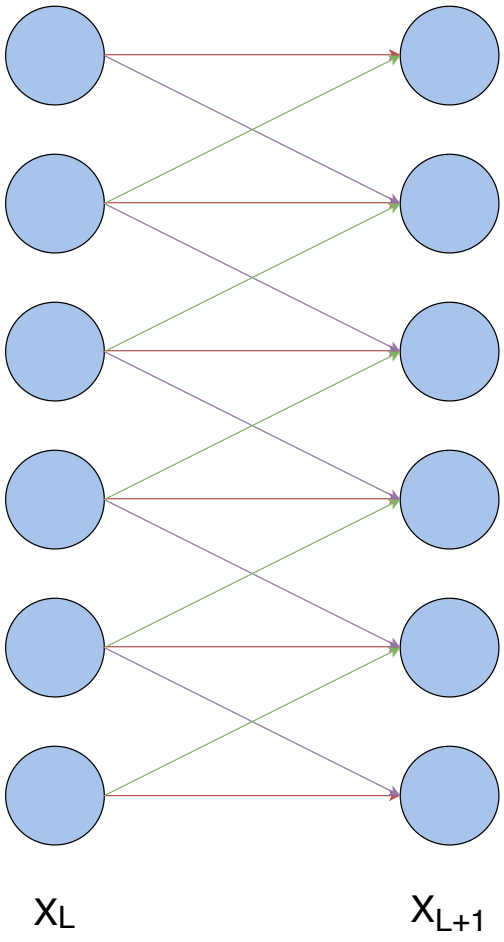
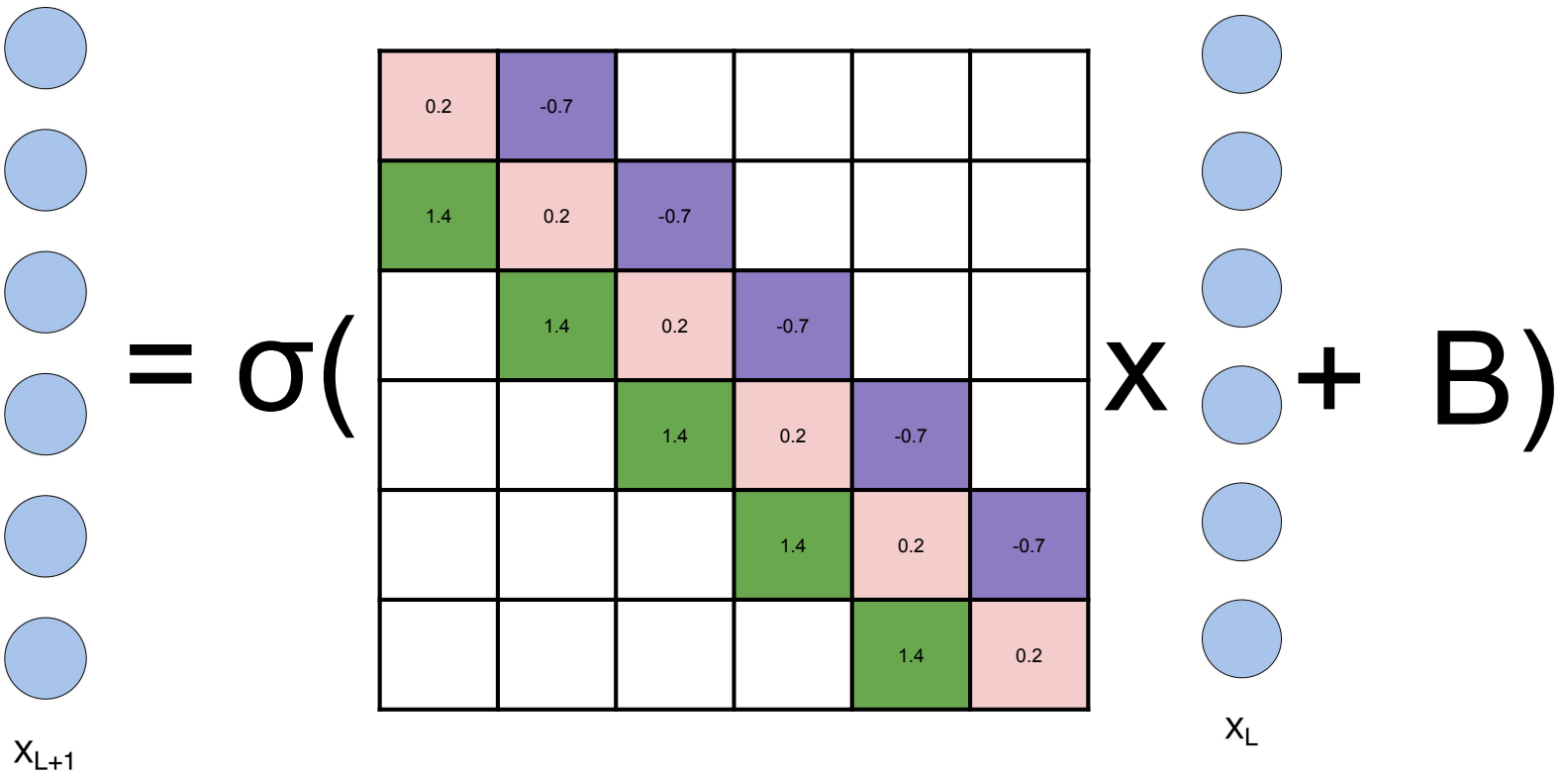
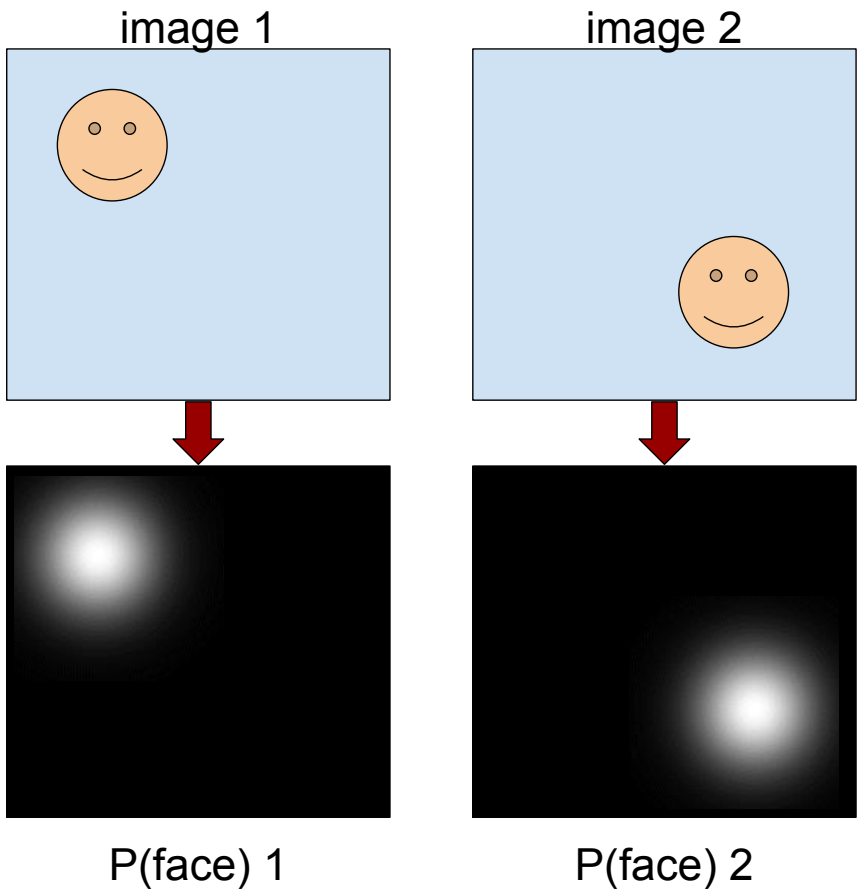


Figure: matrix W_L with constraints

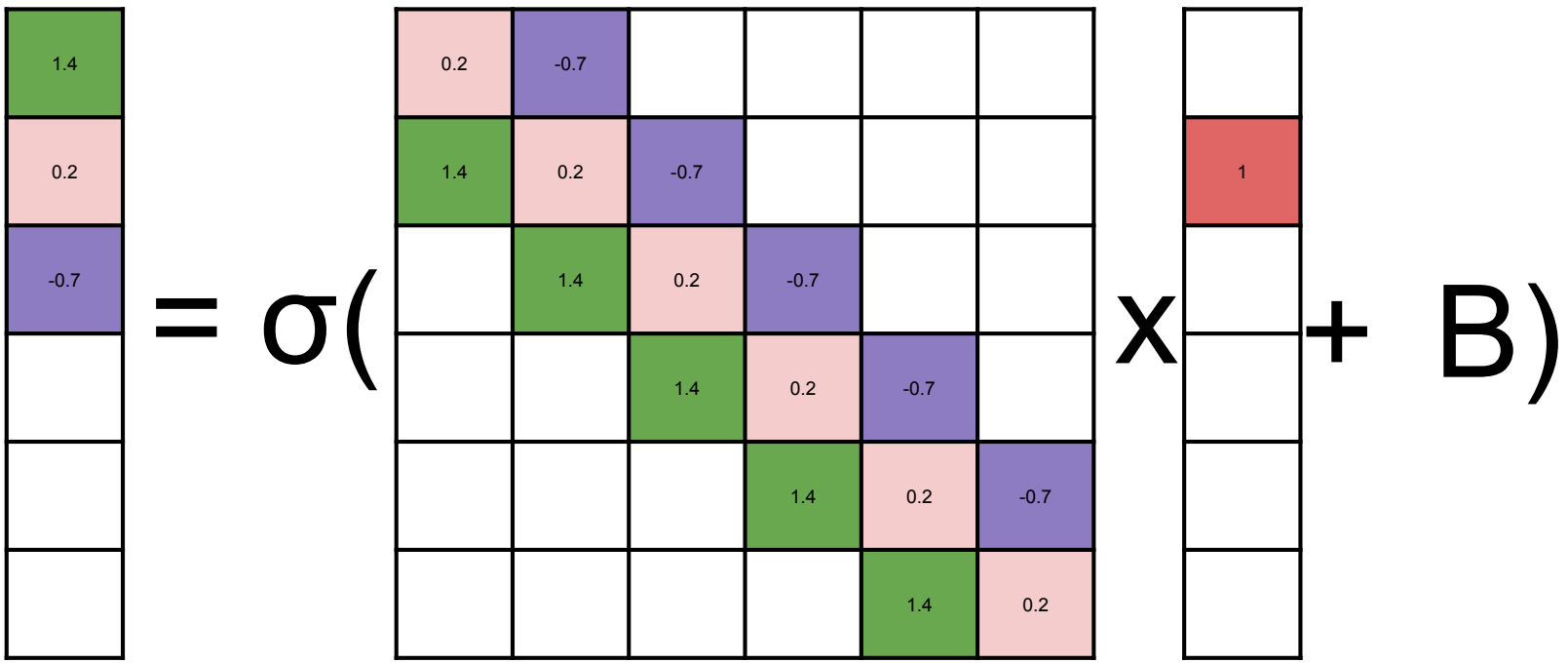
Convolutional layer



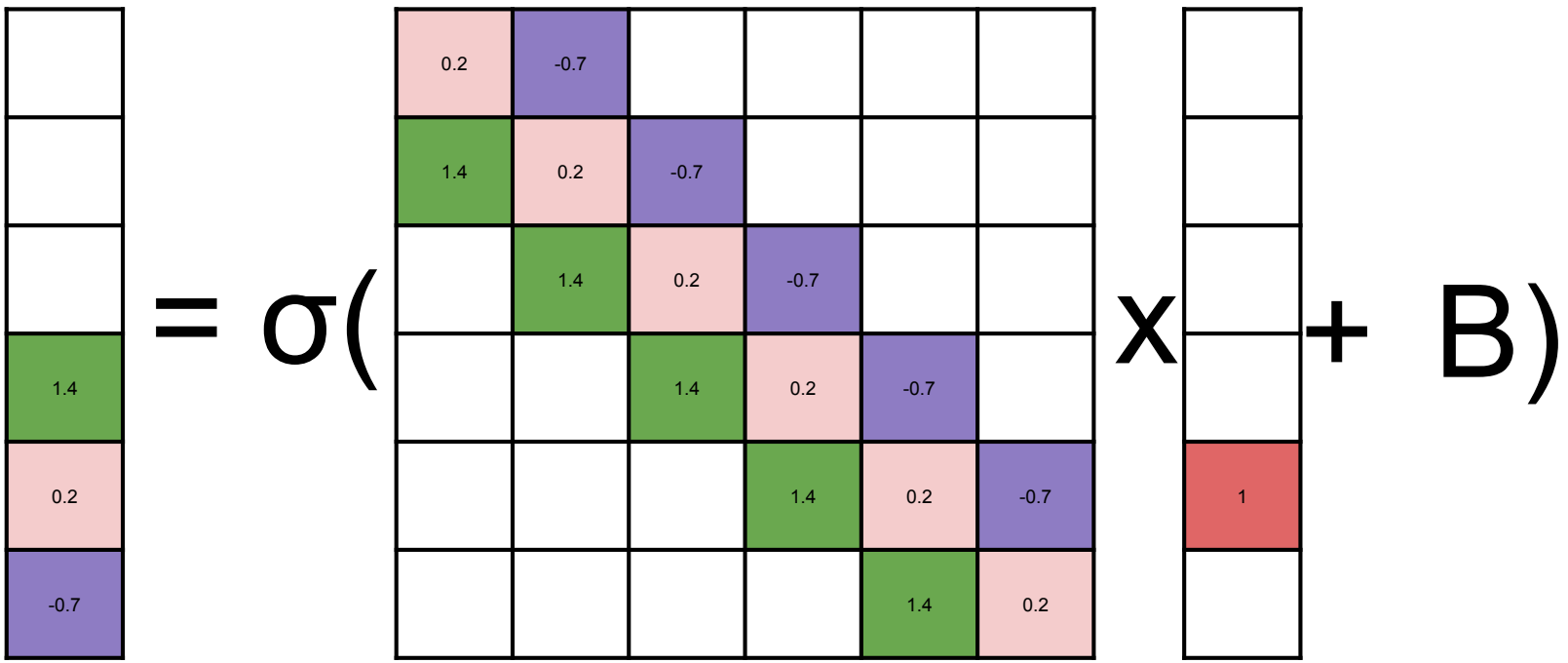
Equivariance



Equivariance



Equivariance



Outline

What are the right assumptions for gene expression data?

- We would like to add **prior knowledge** and/or make **biologically grounded assumptions**

What are the right assumptions for gene expression data?

- Use gene interaction graphs?
- Assume similarity of processes between genes?
- Assume similar perturbation response between individuals/species?

Outline

Incorporate Graph Prior Knowledge?

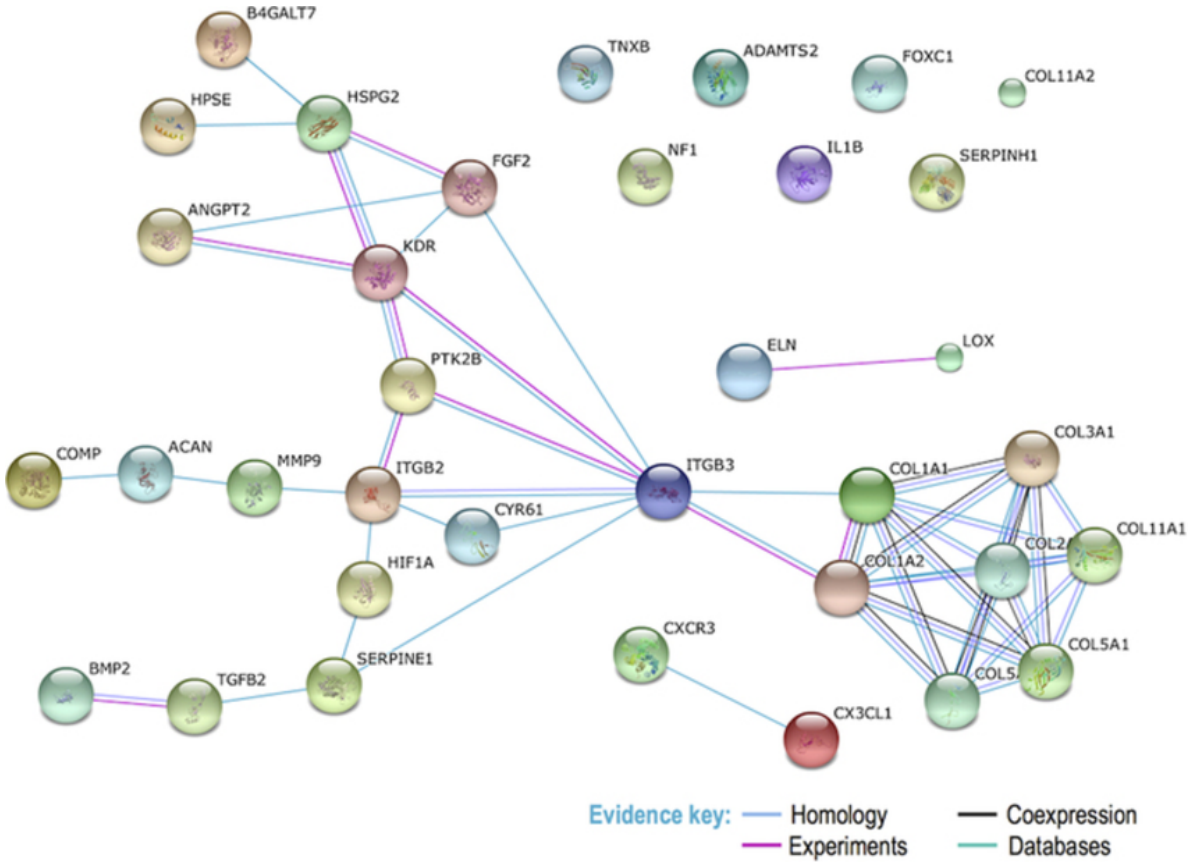
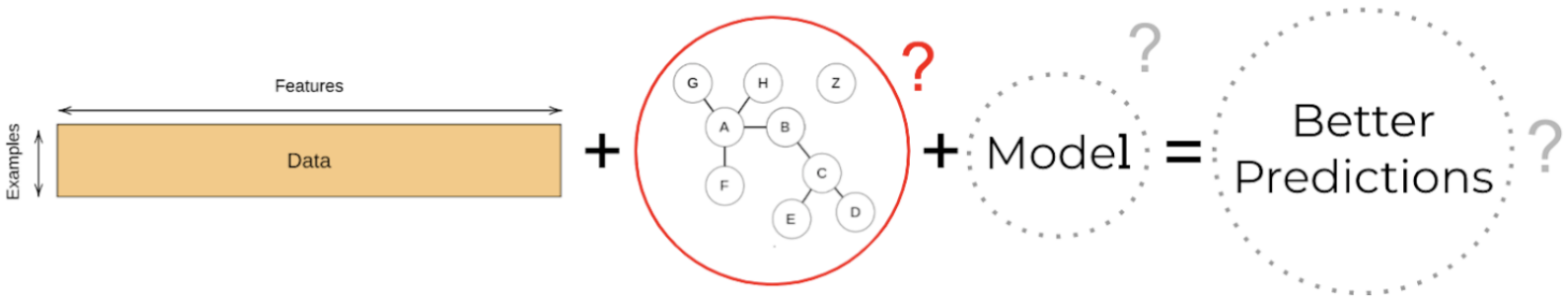


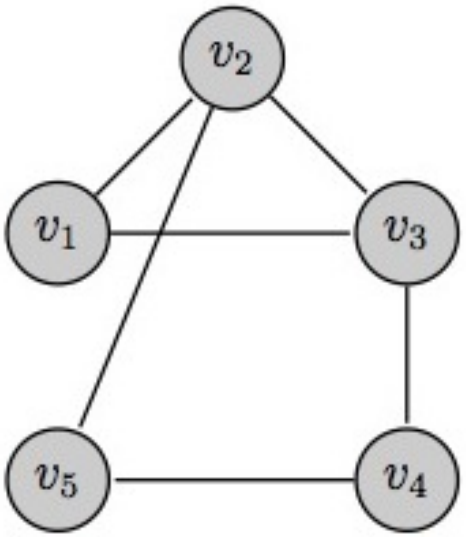
Figure: Example of a curated graph: StringDB

Incorporate Graph Prior Knowledge?



- **Idea**: Use gene interaction graphs to **constrain a ML model**
- 2 questions:
  - How to use the graph in a machine Learning model?
  - Are curated graphs well suited for gene expression data?

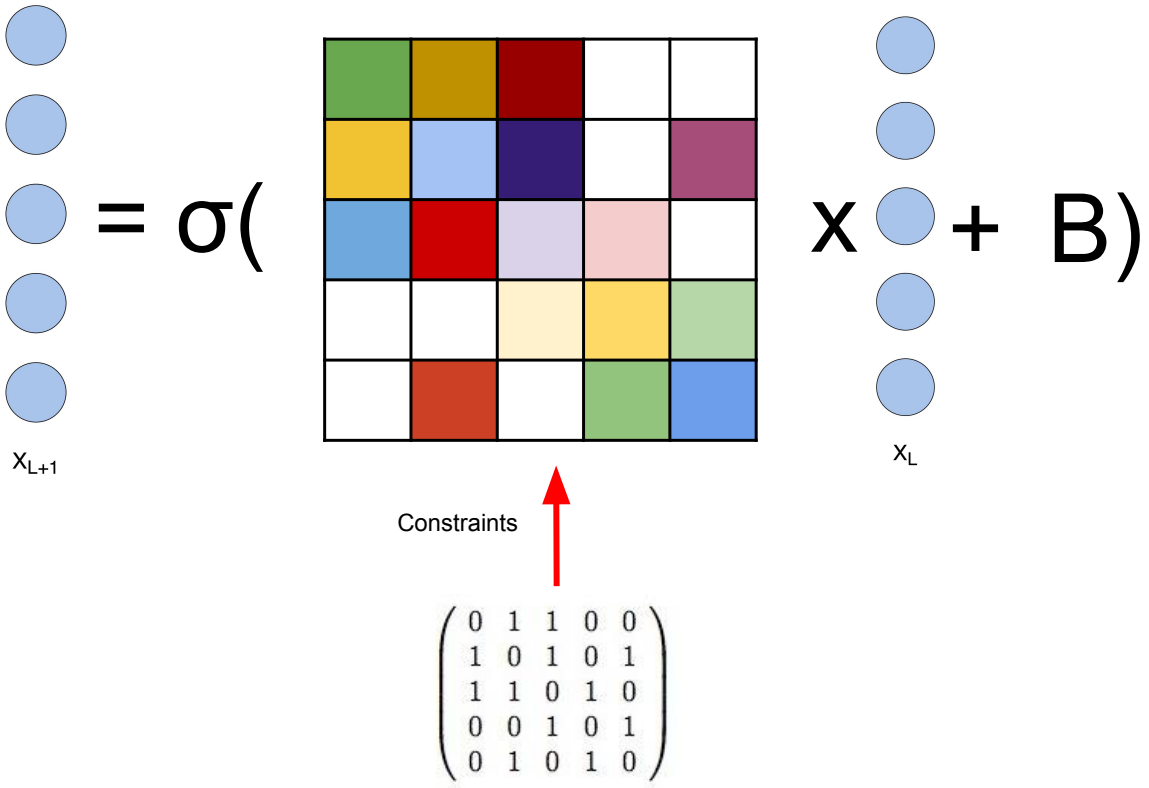
Adjacency matrix of a graph



$$A_G = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

- We can represent an undirected graph by its adjacency matrix
- **Adjacency matrix**: the value at coordinates (i, j) is 1 if nodes i and j are connected, 0 otherwise

Constraining the model: a simple example



- This is one simple example
- Deep learning with graphs is a dynamic field of research!

It does not seem to work!

- **Curated graphs do not seem to be well suited for gene expression data** when using all genes

---

Analysis of Gene Interaction Graphs as Prior Knowledge for Machine Learning Models

---

Paul Bertin
Mila, Université de Montréal
Montréal, Canada

Mohammad Hashir
Mila, Université de Montréal
Montréal, Canada

Martin Weiss
Mila, Université de Montréal
Montréal, Canada

Vincent Frappier
Mila, Université de Montréal
Montréal, Canada

Theodore J. Perkins
Ottawa Hospital Research Institute
University of Ottawa
Ottawa, Canada

Geneviève Boucher
Institute for Research in Immunology and Cancer
Université de Montréal
Montréal, Canada

Joseph Paul Cohen
Mila, Université de Montréal
Montréal, Canada

A current debate

Graph biased feature selection of genes is better than random for many genes

Jake Crawford *† Casey S. Greene †‡

- A current debate: What if you choose the right genes?

What's next?

Could there be an interplay between graph curation and ML model performance?

- Identify genes that hurt performance

Outline

Parameter sharing among genes?

Low dimensional state of the cell

Gene expressions are far from being independent, the **data has a lot of structure**.

- Idea: Use **Representation Learning** with low dimensional latent spaces (*e.g.* dim ~ 500)
  - We can perform analysis in the lower dimensional latent space (*e.g.* fit a prediction model)
  - But we still need a matrix of shape (20k, 500): **lots of parameters**!
- Lots of things (regulatory processes, effects) might be similar among genes
- We can **share parameters among genes** $\rightarrow$ Diet Networks

Published as a conference paper at ICLR 2017

DIET NETWORKS: THIN PARAMETERS FOR FAT GENOMICS

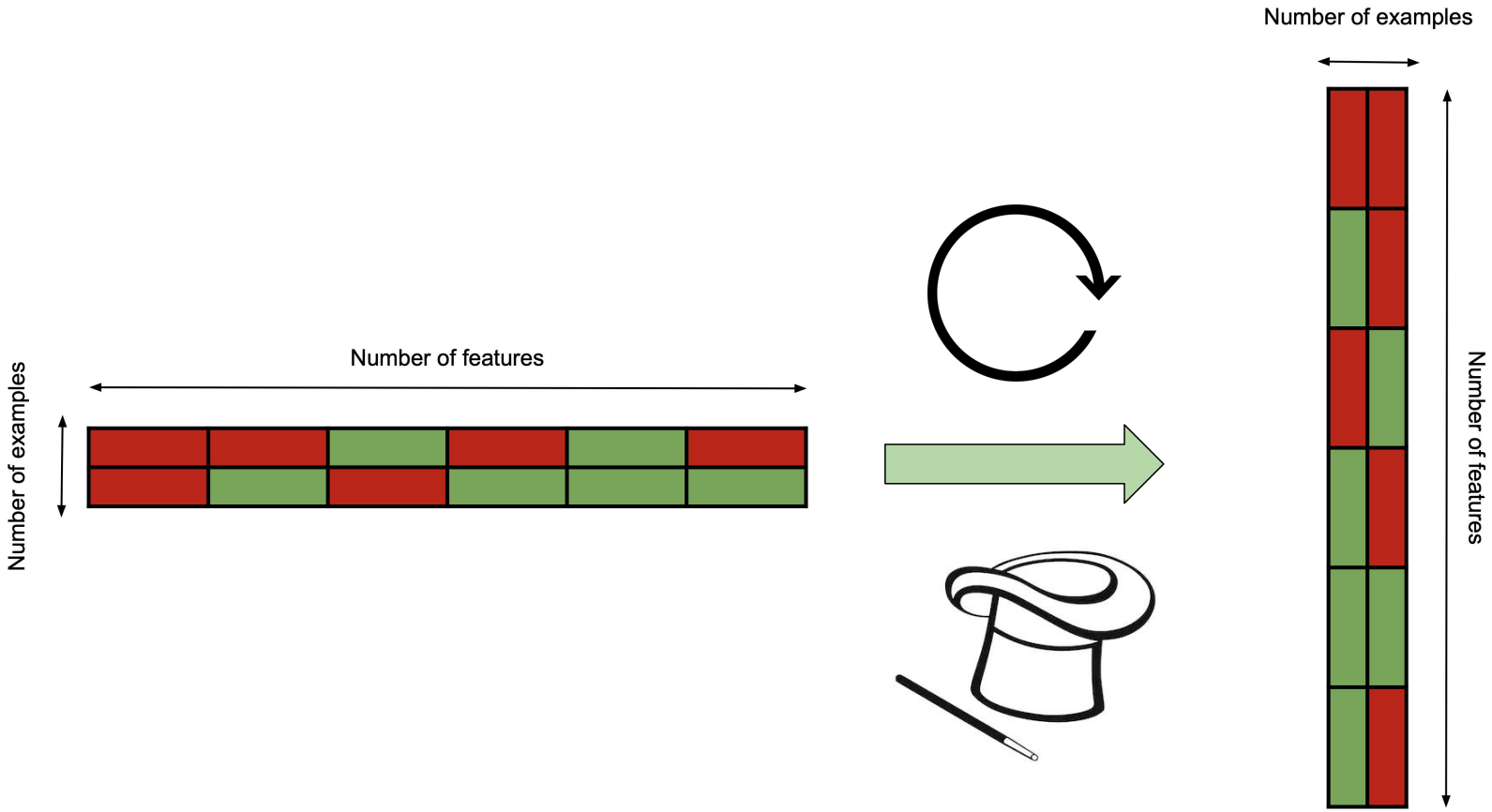
Adriana Romero[*], **Pierre Luc Carrier**[*],
Akram Erraqabi, Tristan Sylvain,
Alex Auvolat, Etienne Dejoie
Montreal Institute for Learning Algorithms
Montreal, Quebec, Canada
firstName.lastName@umontreal.ca, except
adriana.romero.soriano@umontreal.ca
and pierre-luc.carrier@umontreal.ca

Marc-André Legault[1], **Marie-Pierre Dubé**[1,2,3]
[1]University of Montreal, Faculty of Medicine
[2]Montreal Heart Institute,
[3]Beaulieu-Saucier Pharmacogenomics Centre
Montreal, Quebec, Canada
marc-andre.legault.1@umontreal.ca
marie-pierre.dube@umontreal.ca

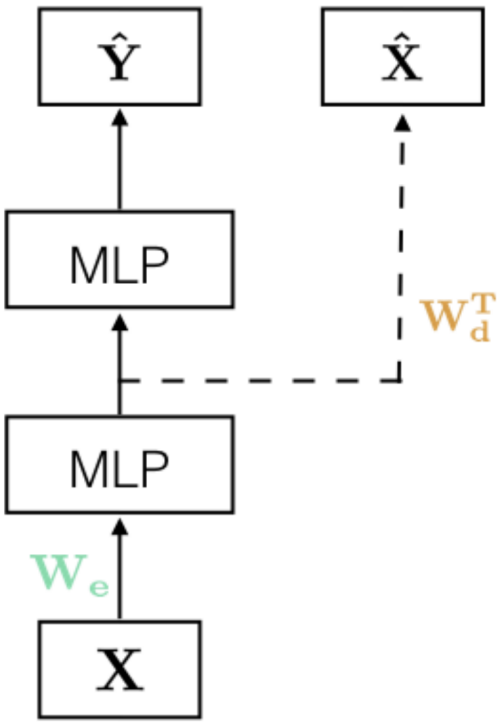
Julie G. Hussin
Wellcome Trust Centre for Human Genetics
University of Oxford
Oxford, UK
julieh@well.ox.ac.uk

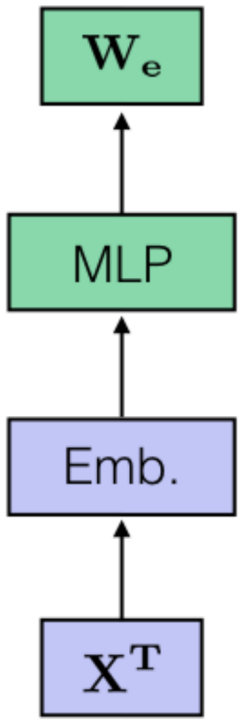
Yoshua Bengio
Montreal Institute for Learning Algorithms
Montreal, Quebec, Canada
yoshua.umontreal@gmail.com

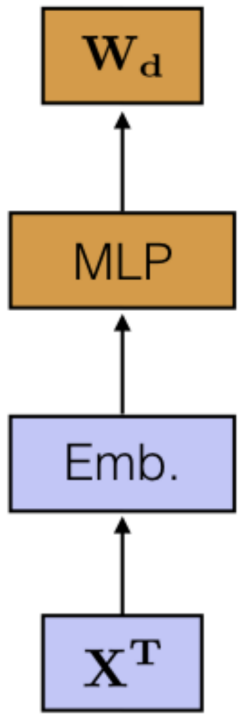
Diet Networks: Magic trick!



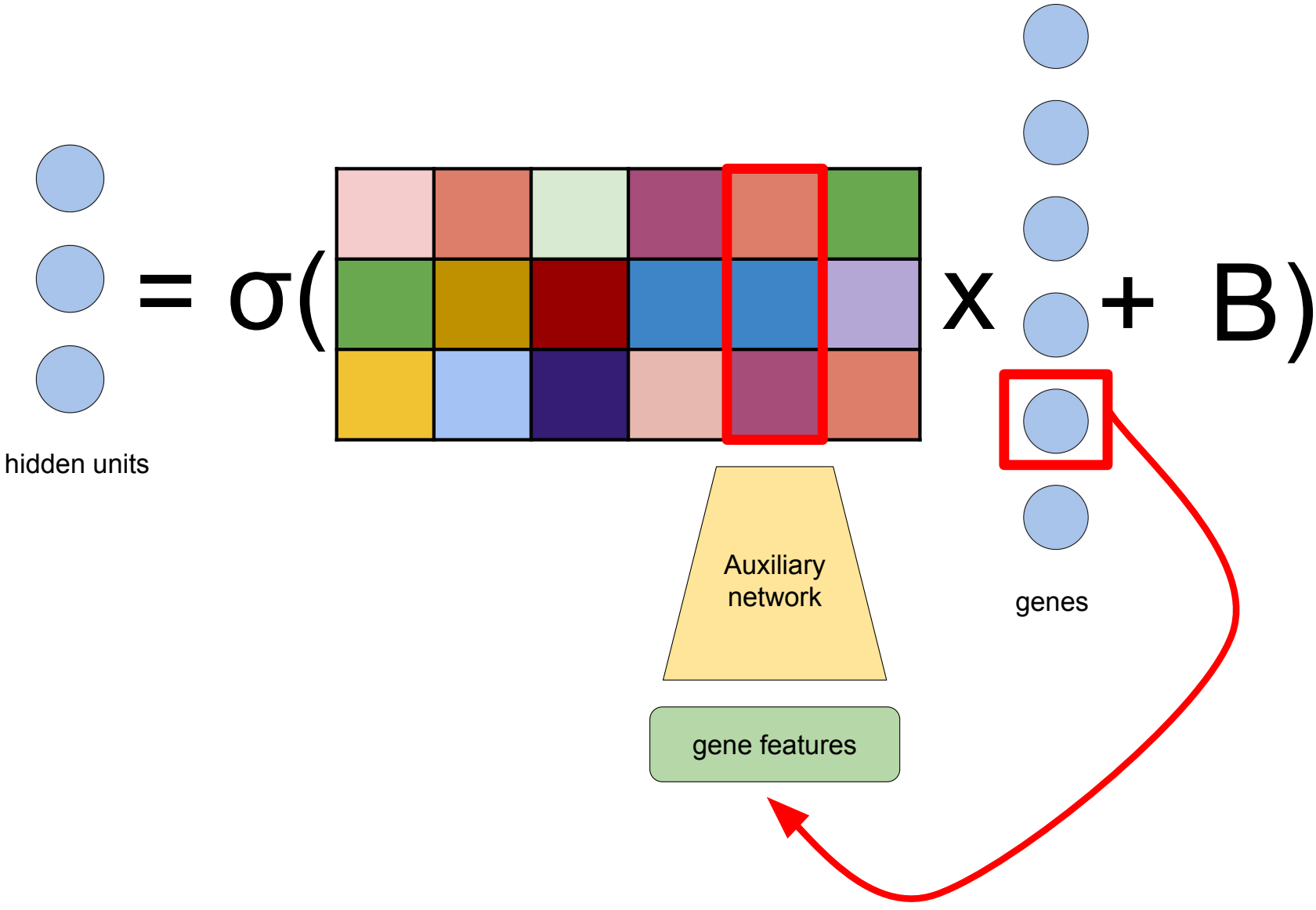
Diet Networks







Diet Networks



Diet Networks

- Not yet applied successfully to gene expression data!
- The features that are fed to the auxiliary networks have to contain **the relevant information about the task you want to solve**!

Open question

What features to use for gene expression data?

Outline

nature methods

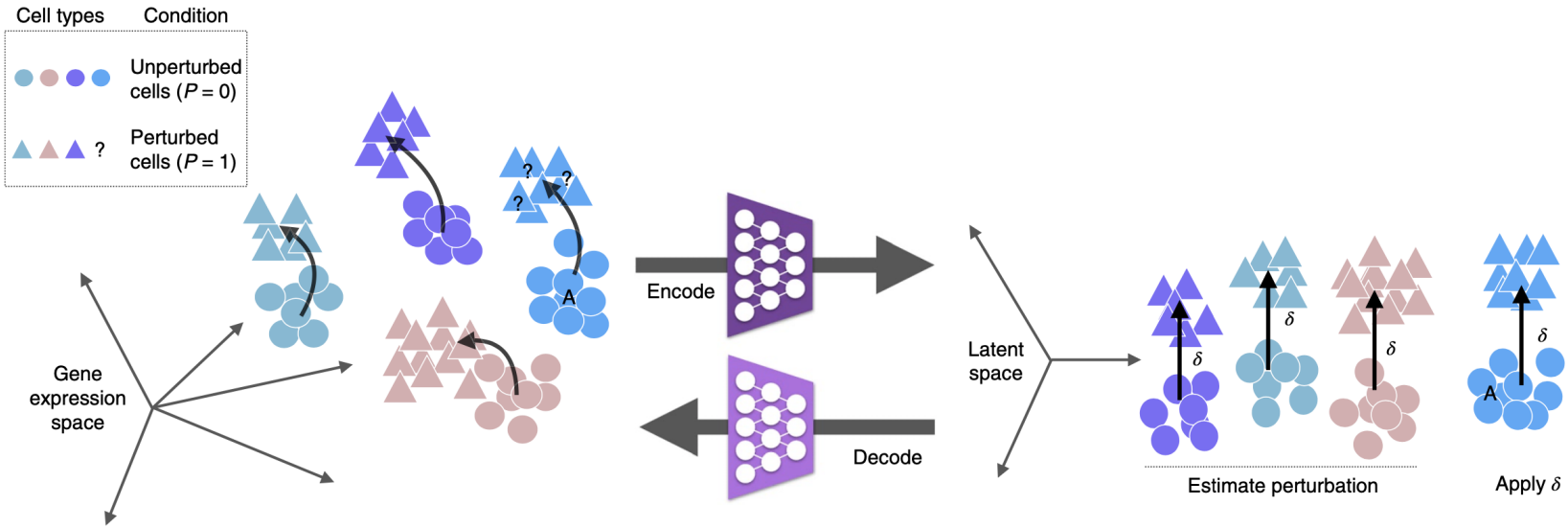
ARTICLES

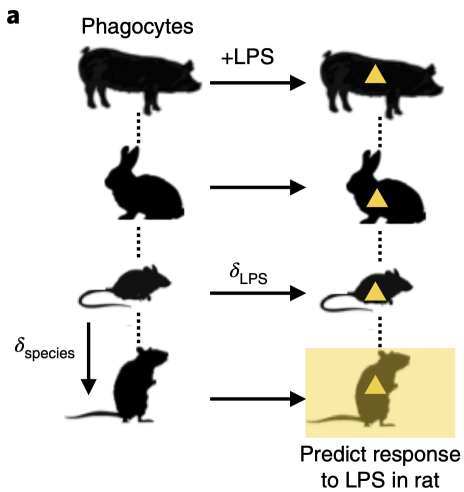
https://doi.org/10.1038/s41592-019-0494-8

scGen predicts single-cell perturbation responses

Mohammad Lotfollahi[1,2], F. Alexander Wolf[1*] and Fabian J. Theis[1,2,3*]

Response to perturbation

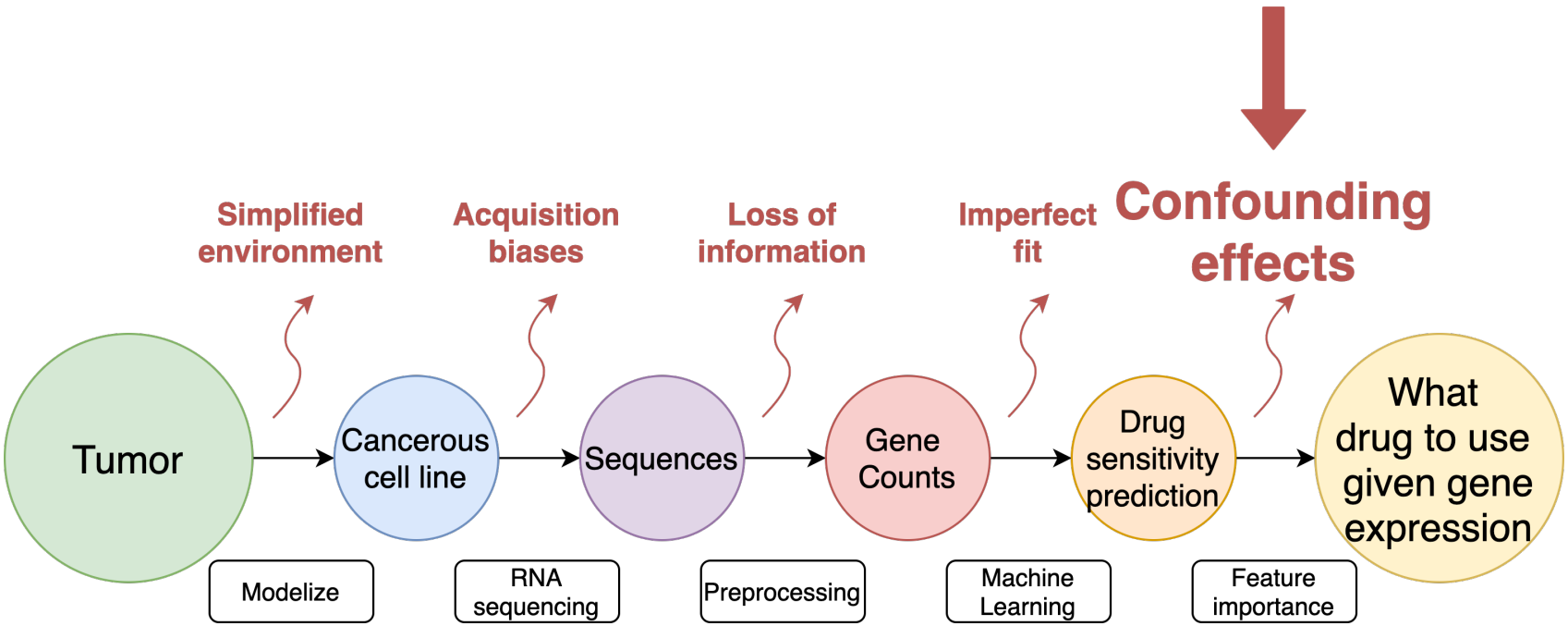




- Observations: **arithmetic in latent space seem to make sense** (*e.g.* Word2Vec)
- Assumption: **response to perturbation is the same in latent space** across species/cell types

Outline

Overview


Simplified environment
Acquisition biases
Loss of information
Imperfect fit
Confounding effects
Tumor
Cancerous cell line
Sequences
Gene Counts
Drug sensitivity prediction
What drug to use given gene expression
Modelize
RNA sequencing
Preprocessing
Machine Learning
Feature importance

Want to get some explanations from the model?

How to better understand what is happening?

How to know what the model is *looking at*? Let us investigate feature importance techniques and their limitations

Outline

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

Karen Simonyan **Andrea Vedaldi** **Andrew Zisserman**
Visual Geometry Group, University of Oxford
{karen,vedaldi,az}@robots.ox.ac.uk

Saliency Maps

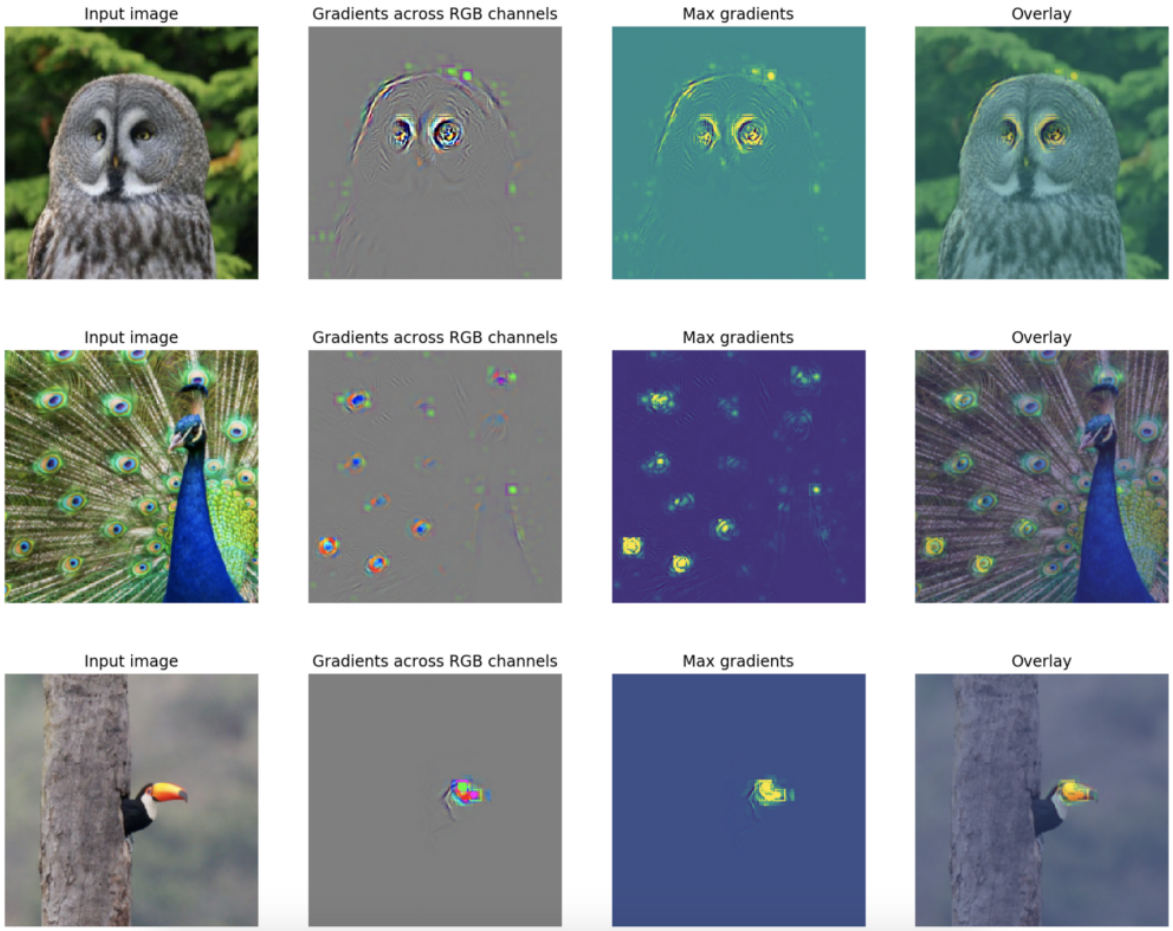
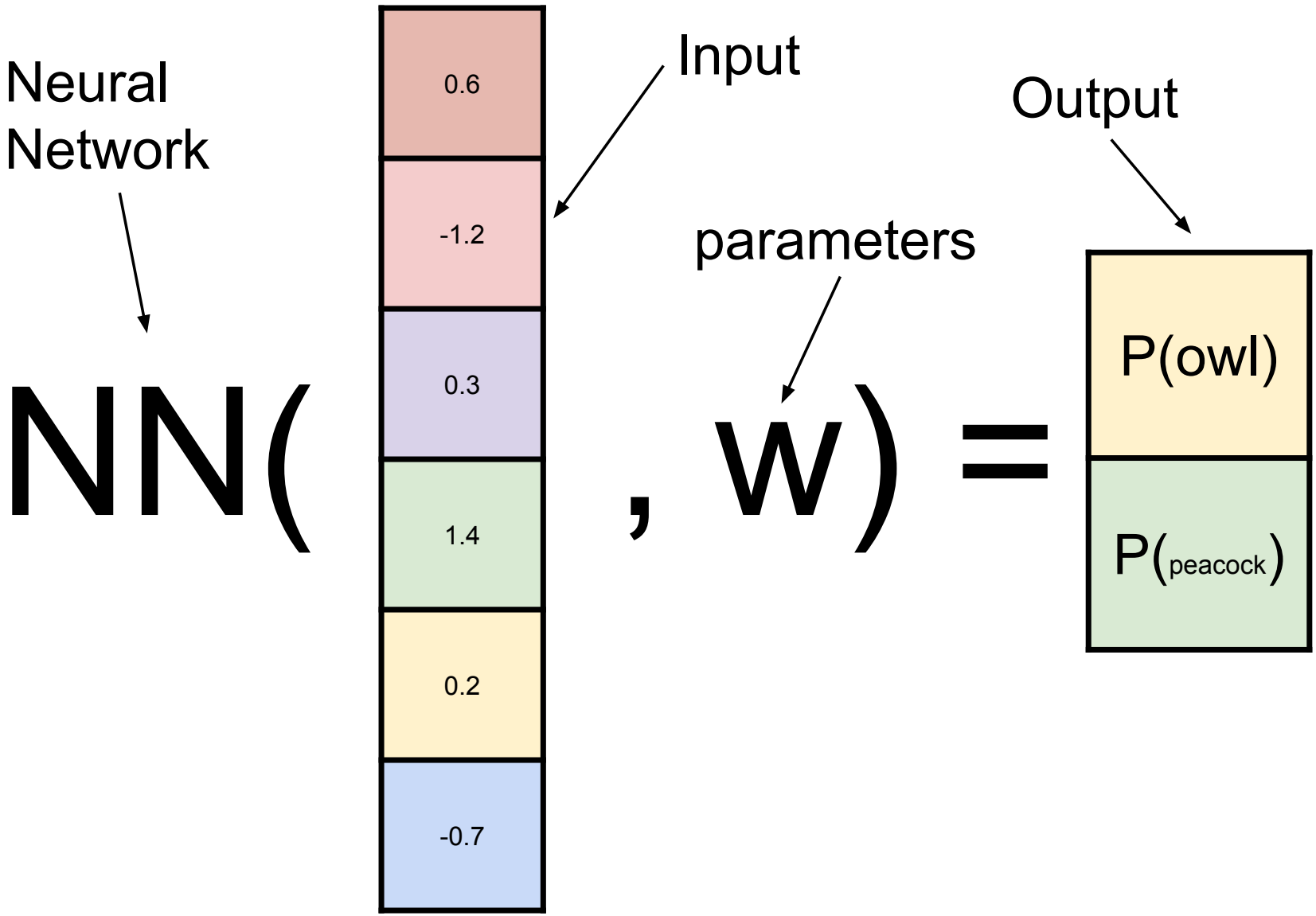


Figure taken from
https://mc.ai/feature-visualisation-in-pytorch%E2%80%8A-%E2%80%8Asaliency-maps/

Parametric models

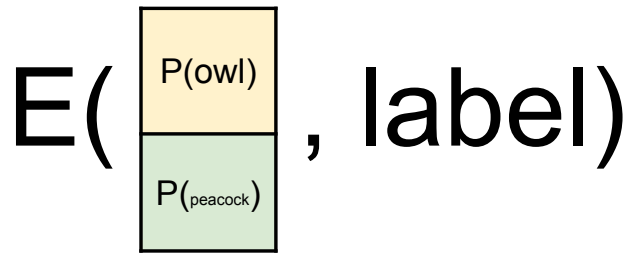


Usual training

How is a model usually trained?

- Iterate:
  - Compute the error for a given input
  - Compute the gradient of the error w.r.t each parameter $\frac{\partial E}{\partial w_i}$ using backpropagation
  - Update the parameters in order to lower the error:

$$w_i^{t+1} = w_i^t - \lambda \frac{\partial E}{\partial w_i}$$



- λ is called the learning rate
- In practice we use several inputs at once (in a **batch**)
- Other *gradient descent* algorithms exist (*e.g. Adam*)

Note

This is called **Stochastic Gradient Descent**

Back to the computation graph

$$\frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$$

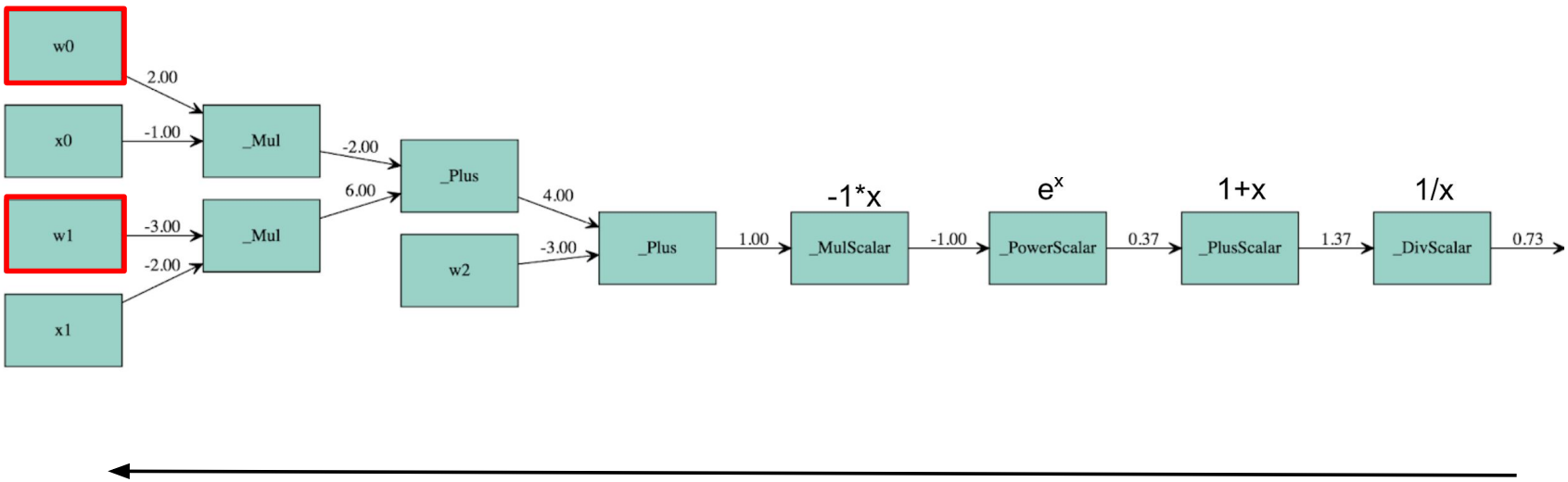


Figure by J. P. Cohen

Back to the computation graph

$$\frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$$

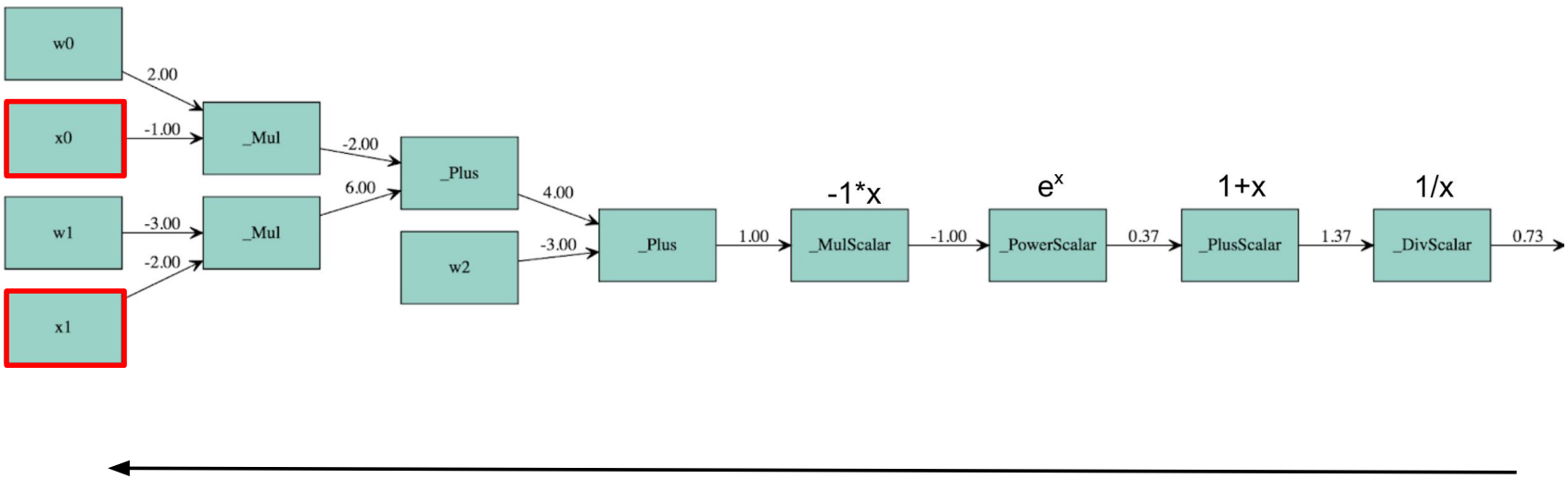
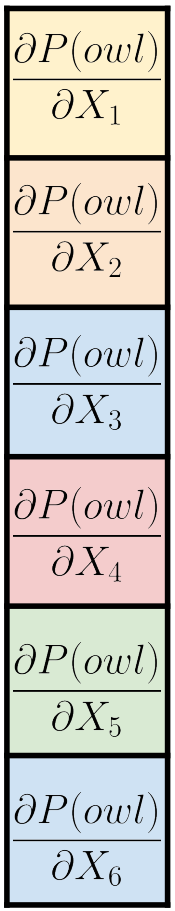


Figure by J. P. Cohen

Saliency Maps



- For a given class probability, *e.g.* $P(owl)$, compute the **gradient with respect to the input** $\frac{\partial P(owl)}{\partial x_i}$
- We get a real number for each input feature

Interpretation

$\frac{\partial P(owl)}{\partial x_i}$: how much the class probability $P(owl)$ depends on feature x_i

Deep Dream

Inceptionism: Going Deeper into Neural Networks

Wednesday, June 17, 2015

Posted by Alexander Mordvintsev, Software Engineer, Christopher Olah, Software Engineering Intern and Mike Tyka, Software Engineer

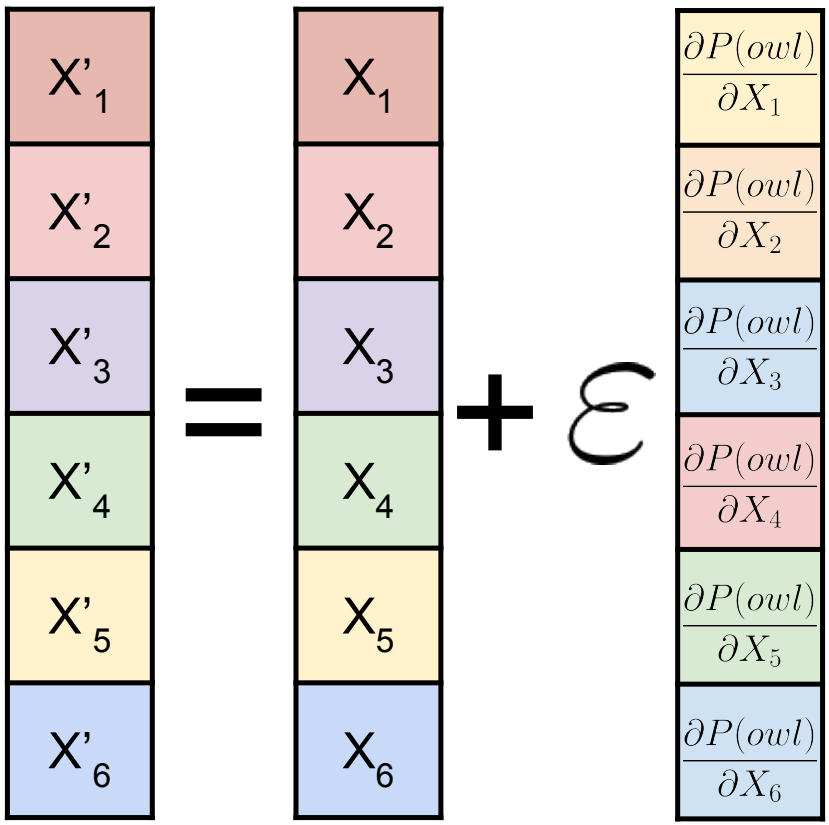
Intuition

Make the model dream the input that would maximize a given class probability

- Gradient ascent in the input space to maximize a given class probability

Deep Dream: Gradient Ascent in the input space

- Update the input by iterating:



Deep Dream

- The input image has been updated in order to maximize the probability of the *dog* class

At the end of the lecture

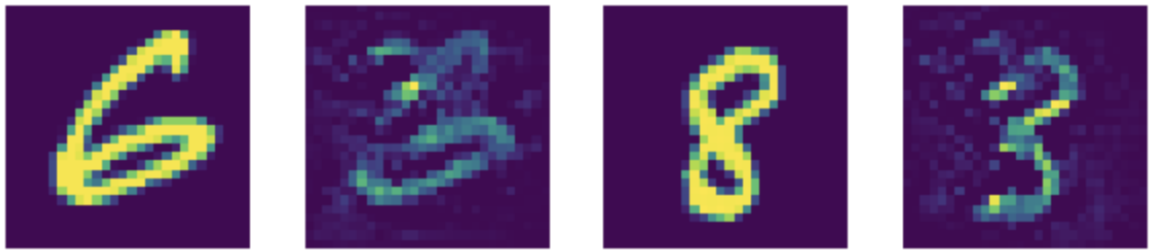


Figure: You will learn how to dream 3s from other numbers!

Visit the following notebook: **Google colab**

Limitations of feature importance methods

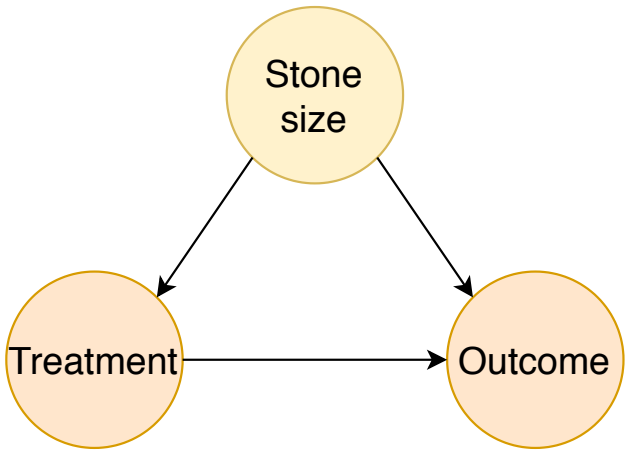
Limitations of feature importance methods

- Feature importance methods can be very noisy and difficult to interpret for gene expression data.
- **Feature importance does not provide a causal explanation as the prediction can be confounded**.

Outline

Simpson's paradox

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



- Two treatments for kidney stones
- **Which one is better?**

Figure: The size of kidney stones has an effect on both treatment assignment and outcome

Figure taken from `https://en.wikipedia.org/wiki/Simpson%27s_paradox`

Simpson's paradox: an example from regression

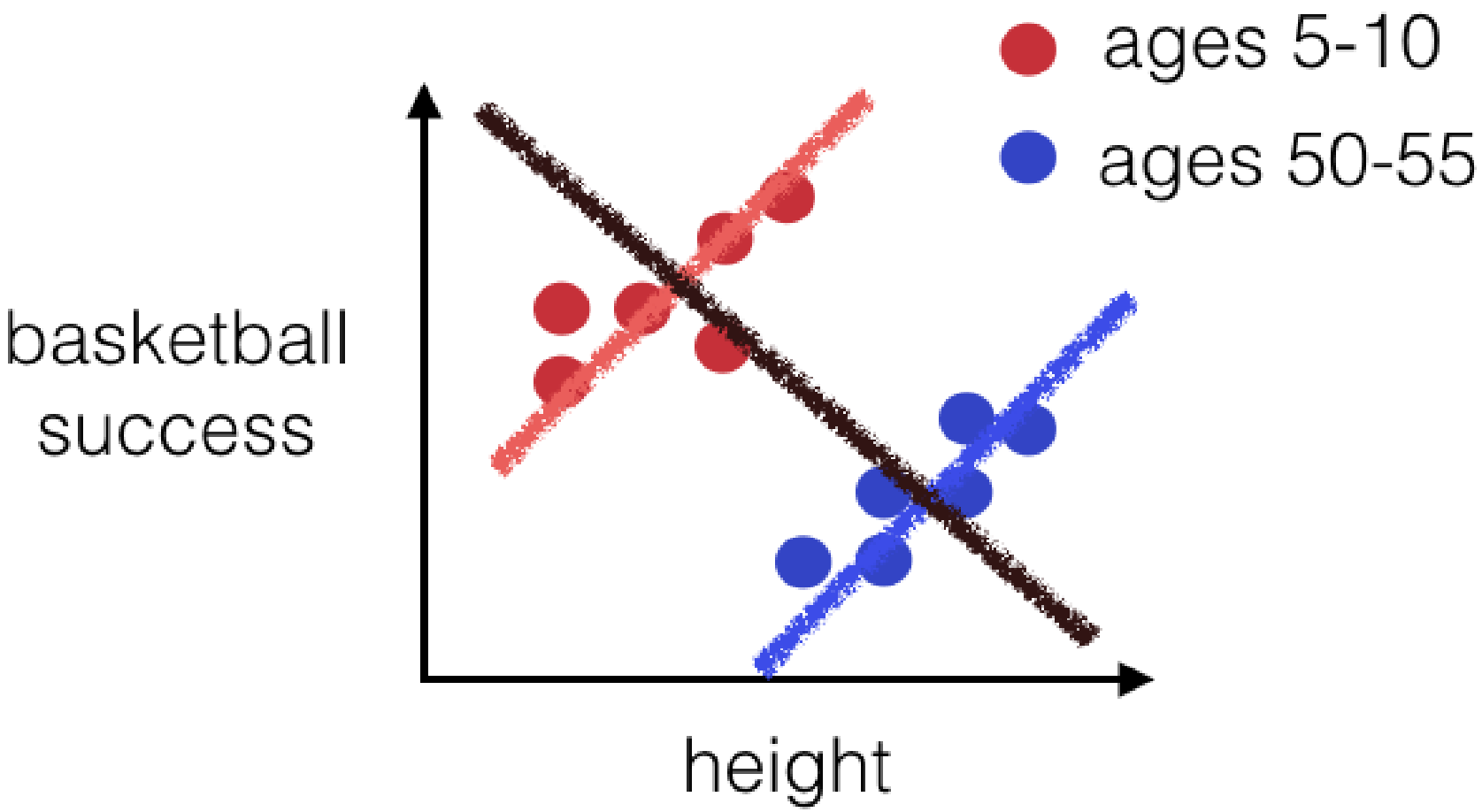
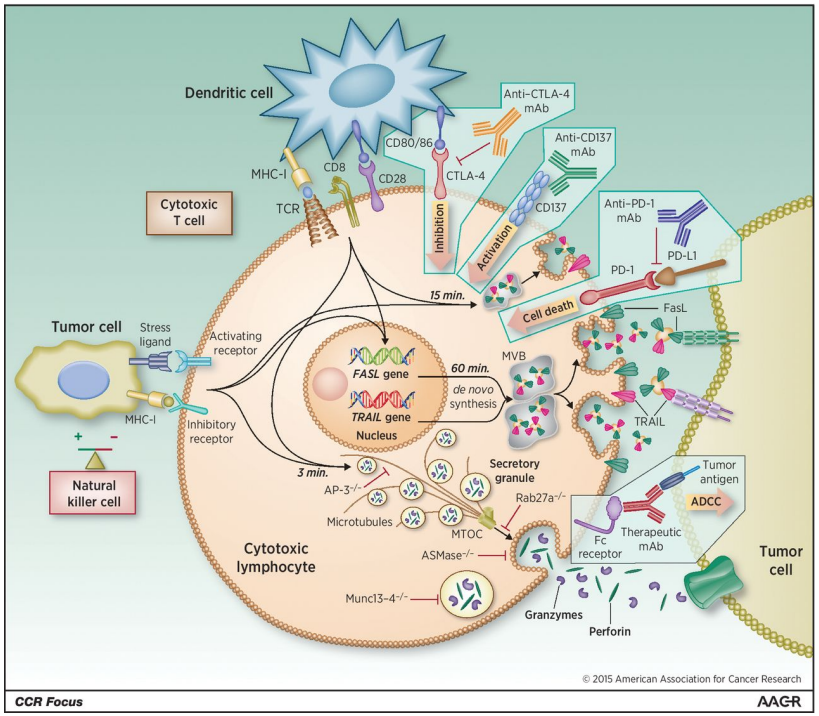
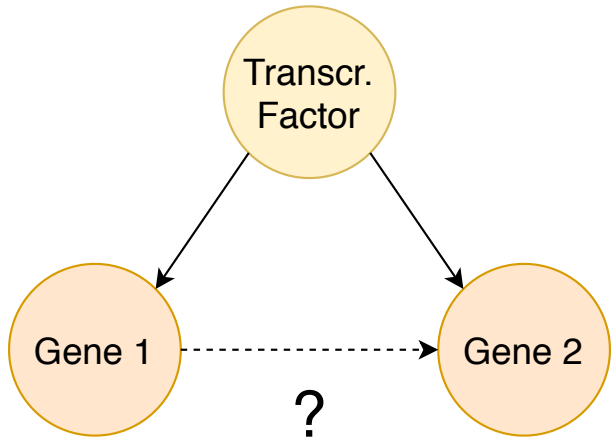


Figure: If we do not take age into account, we may conclude that height has a negative influence on basketball performance!

How to understand the mechanisms of the cell?





- Would you like to identify the effect of a gene on another gene?
- Lots of confounders
- Current area of research (module networks...)

Figure taken from *https://clincancerres.aacrjournals.org/content/21/22/5047*

Simpson's paradox: Take-away

Take-away

If you are provided with data that contains several partitions[1], you may want to **fit a model on a whole data** as well as **separate models for each partition**.

- You can analyse the *story* told by feature importance techniques applied to the different models.
- If all models agree, you have an interpretation that is robust across partitions (but no guarantee that the story is true...)
- If not, you may want to investigate further (lab experiments?)

In transcriptomics, there are lot of unobserved confounders!
(*e.g.* non coding parts of the genome)

[1] *e.g.* cell lines, cell types, expression level of important Transcr. Factors

Outline

Conclusion

Conclusion

- We investigated several challenges of Machine Learning when it is applied to transcriptomic data.
- **We need to design models making the right assumptions for gene expression data!**

Practice time!

Visit the following notebook: **Google colab**