Project 2022
BINF-F401

Vincent Detours

Overview of course evaluation

Written exam

- Questions assess your understanding of the theory
- It's an open book exam or 2 hours (but, don't expect to study the course during these two hours!)
- You've got a personal grade

Project

- You apply ideas and tools of the course on real data
- It's a mini research project evaluated from a written report
- It's done by groups of three students, so grade is collective

Final grade is the harmonic mean of exam and project (see https://en.wikipedia.org/wiki/Harmonic_mean). Thus, you need reasonable scores for both exam and project to pass.

Scientific background

Quantitative morphology

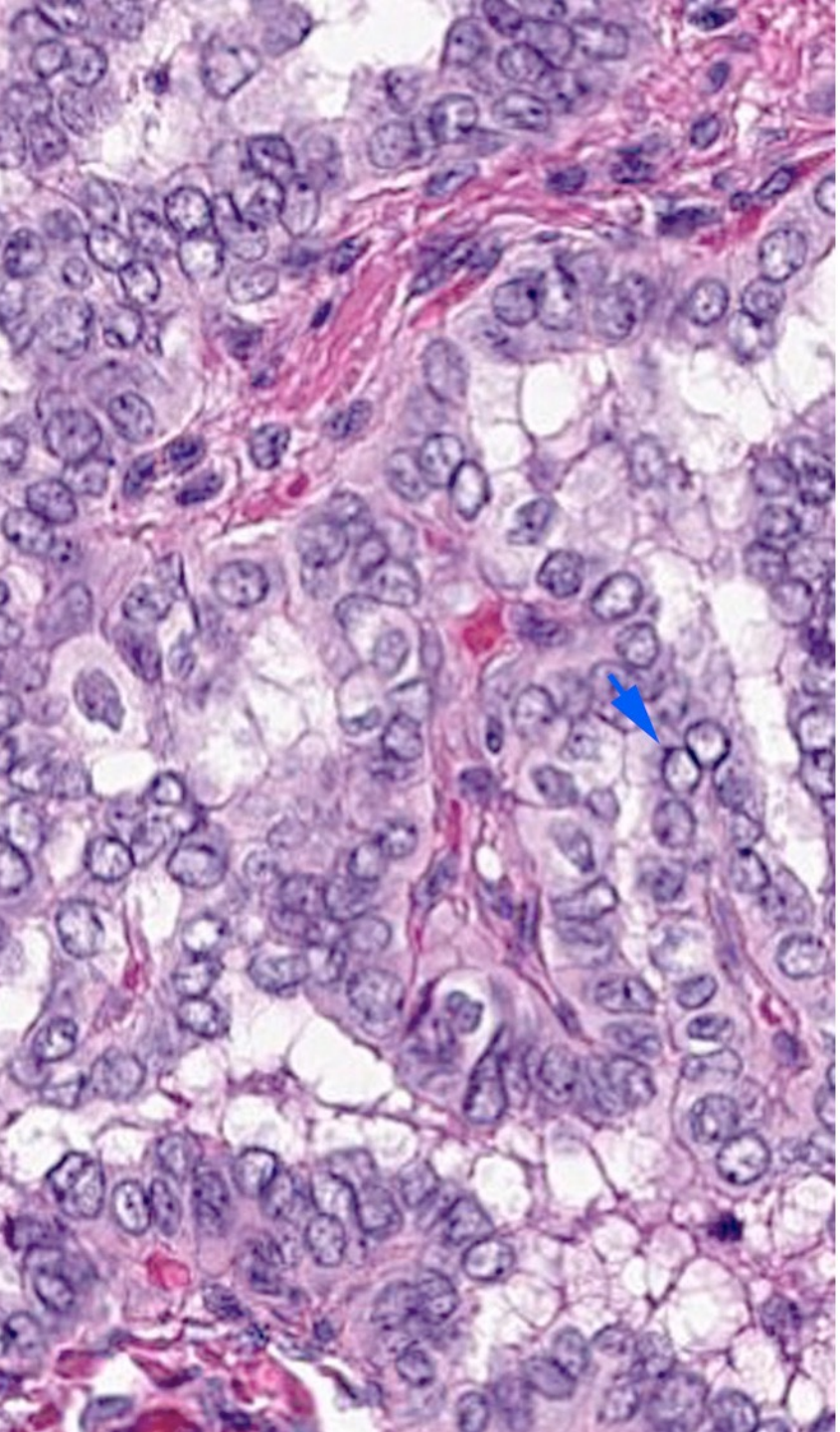
- Tissue morphology is the topic of a branch of medicine called pathology
- The diagnostic of many diseases, including cancer, rests of morphological analysis
- The main instrument of pathologists is the **microscope**

Microscope of van Leeuwenhoek (replica, 1670)

Leica confocal microscope, 2017

- The extraordinary progress of microscope technology paved the way of many conceptual advances in biology and medicine

- But the description of morphology has barely evolved over centuries: it remains qualitative and somewhat subjective



Example (thyroid cancer)

- Nuclei of cancer cells are large, irregular, light, have the aspect of ground glass. Some look like coffee beans.

This verbal description is

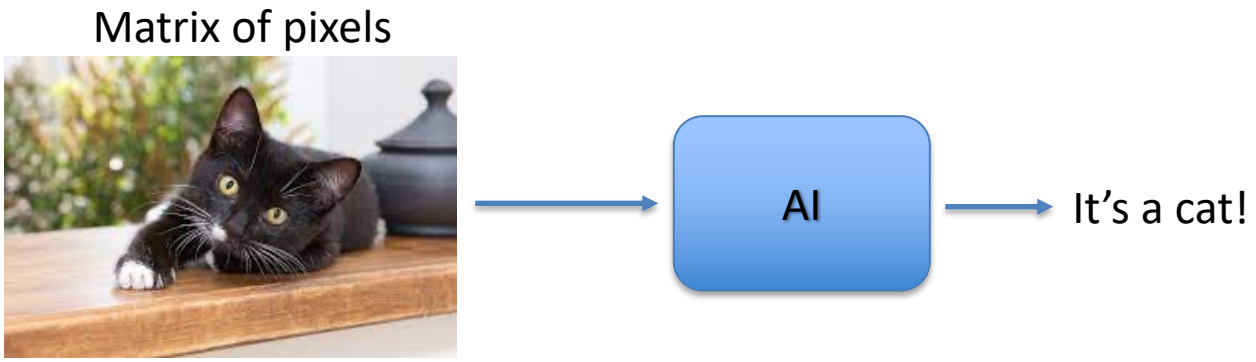
- **Qualitative**: what is the limit between normal and ‘big’ nuclei?
- **Subjective**: do two pathologists have the same idea of ‘big’, ‘irregular’ and ‘coffee bean-like’?

These shortcomings are obstacles to

- the quality of diagnostic
- our understanding of morphology

Artificial intelligence may set morphology on a quantitative basis

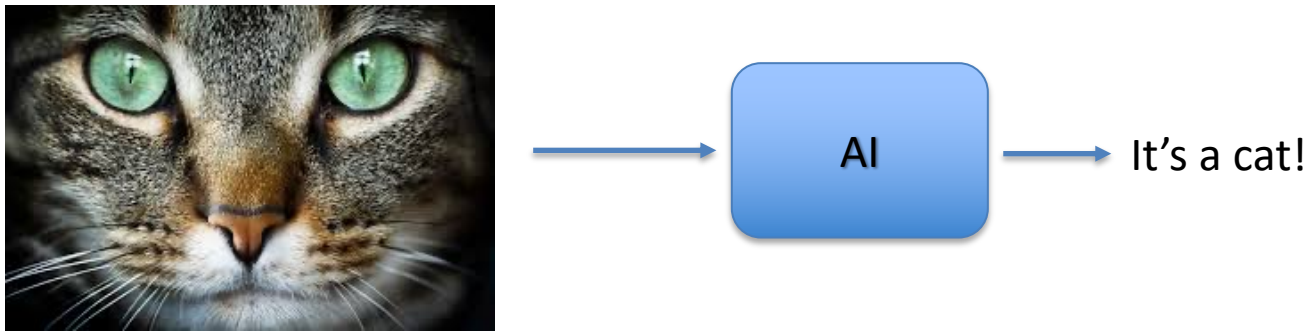
- Photos is a matrices of pixels reflecting the activation of an optical sensor
- AI extract high-level **semantic information** from raw images:



Artificial intelligence may set morphology on a quantitative basis

- Photos is a matrices of pixels reflecting the activation of an optical sensor
- AI extract high-level **semantic information** from raw images:

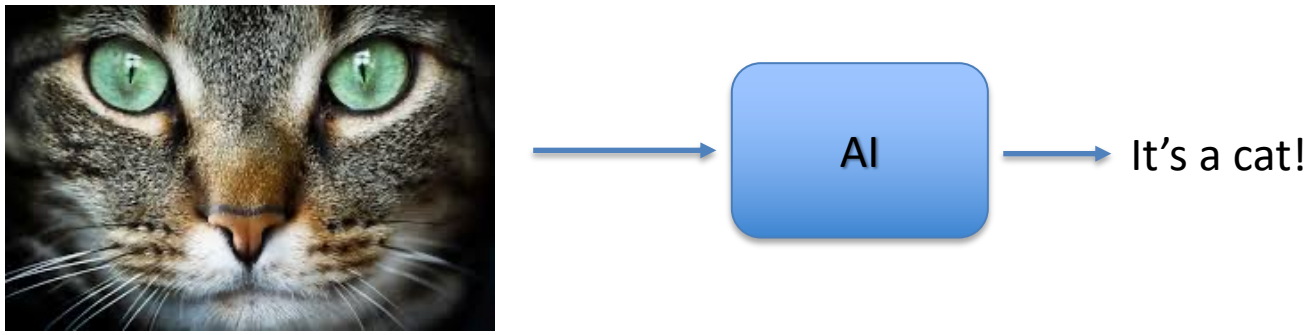
Another matrix of pixels



Artificial intelligence may set morphology on a quantitative basis

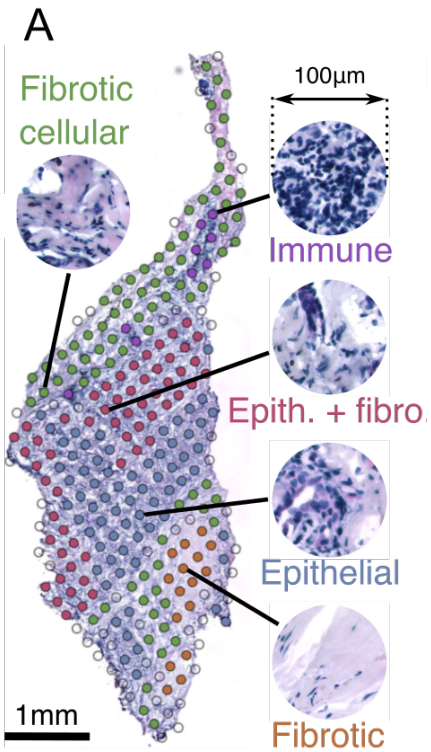
- Photos is a matrices of pixels reflecting the activation of an optical sensor
- AI extract high-level **semantic information** from raw images:

Another matrix of pixels



- Intermediate representations of images computed by inner layers of deep neural network can be exploited to turn images into numerical semantic representations and group them by (semantic) similarity.

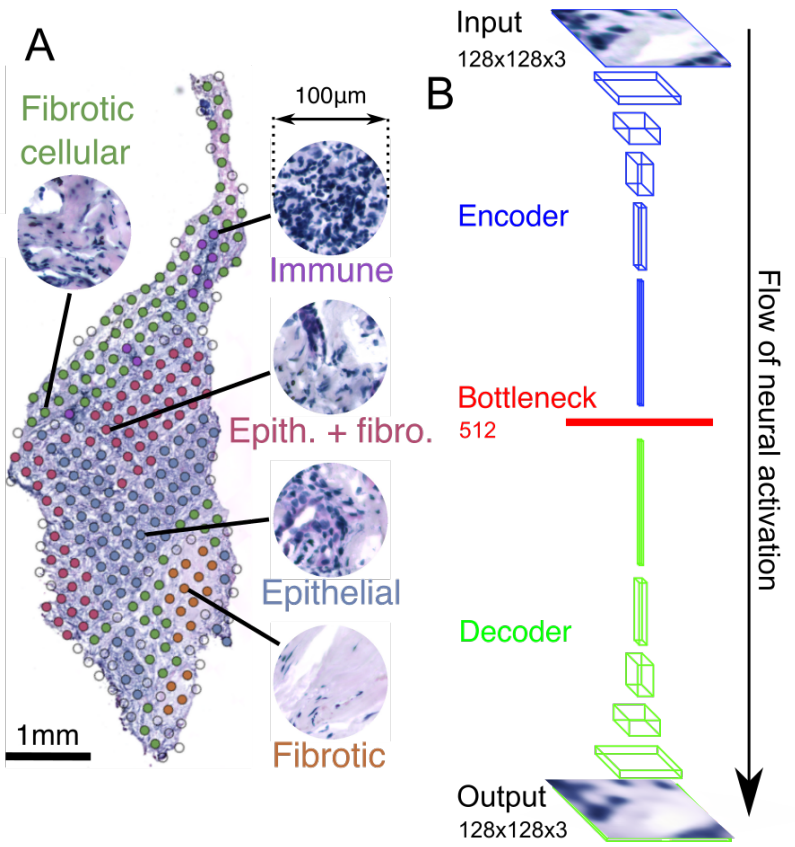
An example of unsupervised AI



A papillary thyroid cancer slice was scanned at 40X

Morphology is heterogeneous across space

An example of unsupervised AI



The slice was split into small tiles

Tiles were used to train an autoencoder

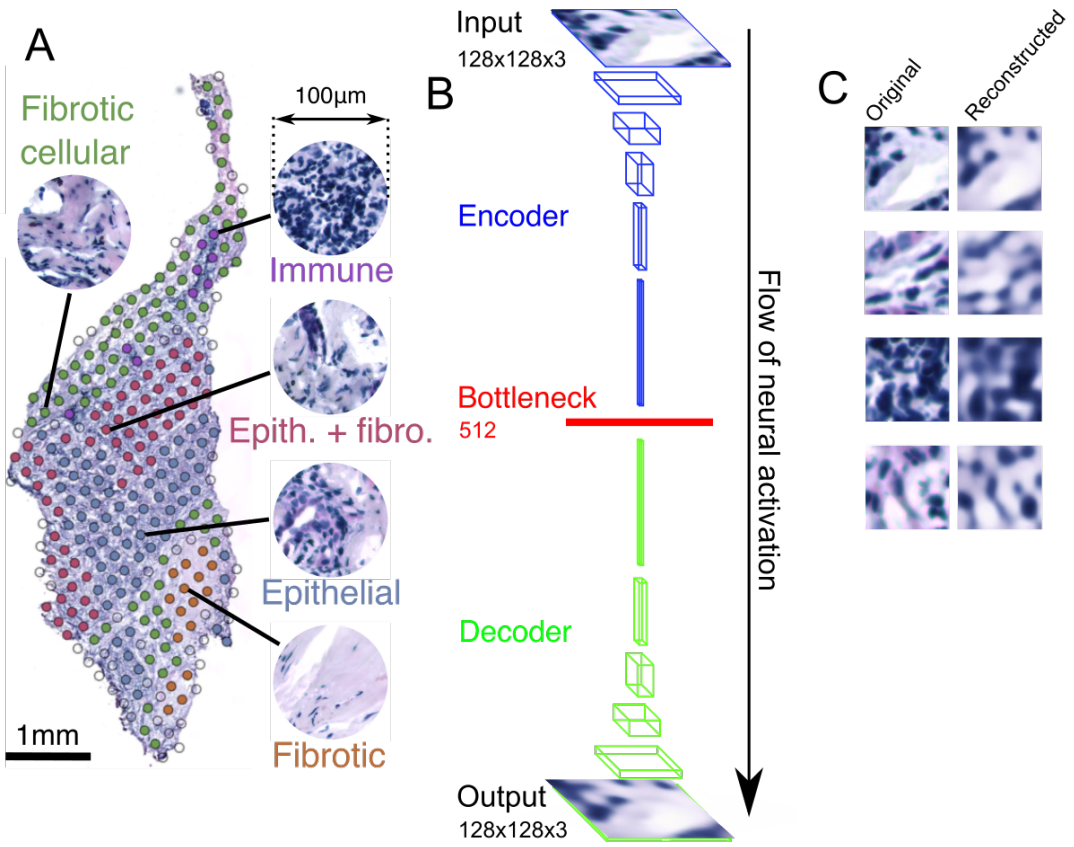
An autoencoder is a deep neural network trained to generate in its output the image it is presented as input

The trick is that the information flows through an informational bottleneck

Here the size of the bottleneck is 1% of the input image, thus the encoder must abstract out details and ignore noise

The numerical vector defining bottleneck neurons activations elicited by an image is called its ***latent representation***, it sits in a 512 dimensions space, that we also refer to as the ***morphological space***

An example of unsupervised AI



Our autoencoder does generate the input images, these are blurry versions of the original

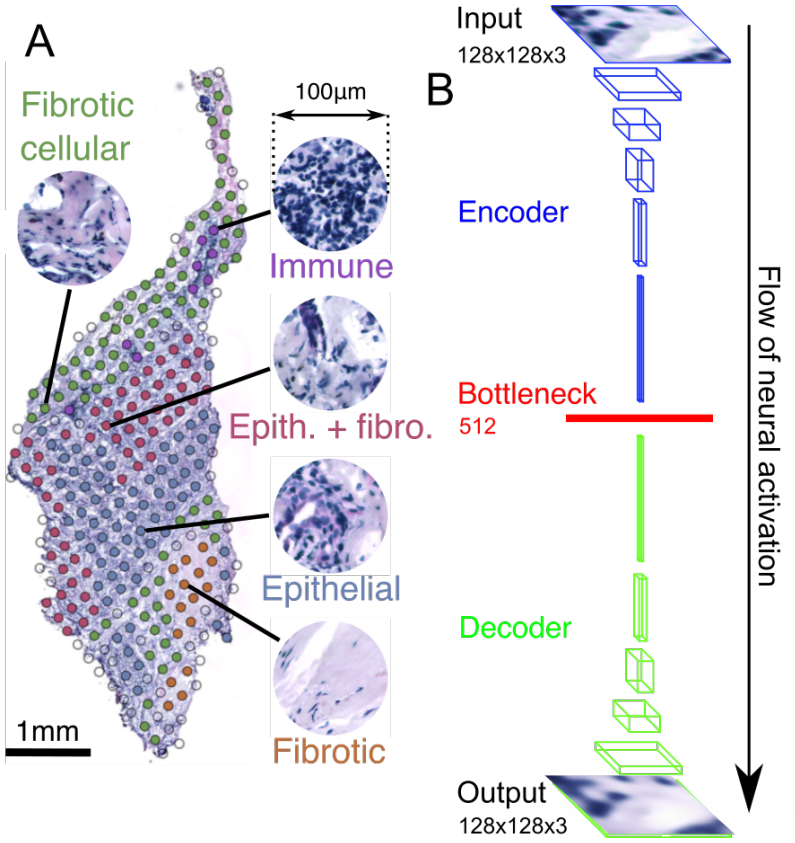
We are not interested in these reconstructed images, but in their latent representations

These are points in a 512 dimensions space, our poor brains can't cope!

So we collapse this space in 2D using the dimension reduction methods we routinely used in single cell transcriptomes analysis, e.g. t-SNE, UMAP, etc.

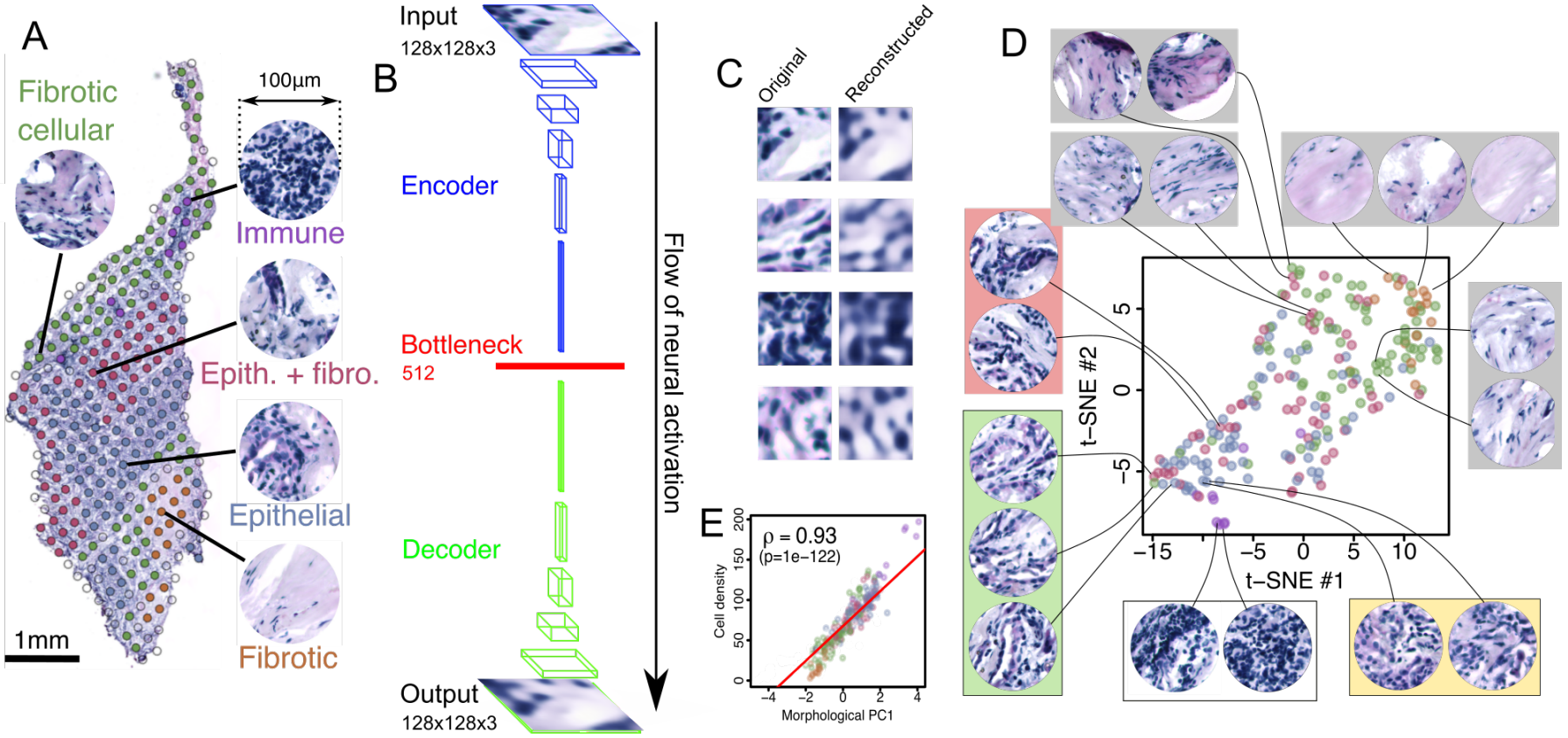
Now we can see the latent space…

An example of unsupervised AI



… and it turns out to make histological sense, grouping together images that are morphologically similar.

An example of unsupervised AI



In this example, the autoencoder has ‘discovered’ that cell density is an important organizing quantity of tissue analysis

An example of unsupervised AI

Overall,

- No human supervision/annotation and no a priori knowledge was needed to train the AI. It is truly free of cognitive biases and scientific a priori
- The AI turns the images into numerical vectors that represent tissue shape that are suitable to define similarity metrics
- The grouping of tissues in the morphological space makes histological sense
- We now have in our hands quantitative and reproducible computations in place of loose verbal statements like “these two morphologies are [similar/akin/related/share features]”
- The AI has an unlimited attention span. Large collections of high-resolution slides can be exhaustively examined

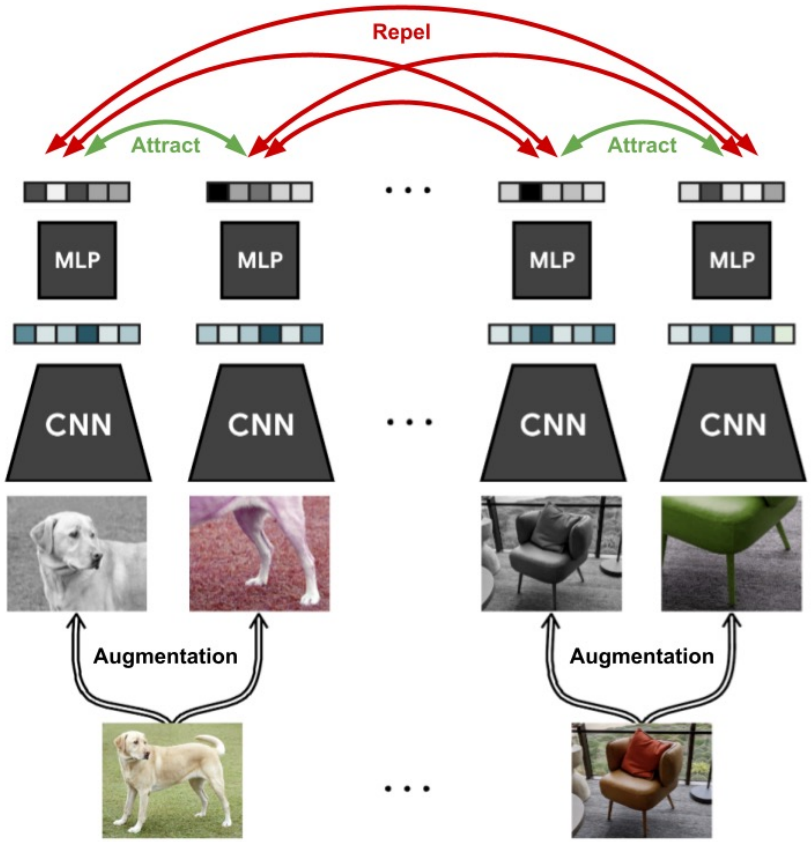
Contrastive learning

- The reconstruction objective of the autoencoder is an overkill: the exact positions of cells is morphologically irrelevant

- It is even detrimental: we need latent representations that are insensitive to technical color variations, to the orientation of the slide under the microscope, etc.

- Contrastive learning does not use a reconstruction objective and it makes it possible to train AIs invariant to specific image features

Contrastive learning



From https://simclr.github.io/

- We used the smCLR framework (Chen *et al.*, arxiv 2020)
- The AI is presented pairs of transformed images
- It is trained to guess if the two images are transformations of the same images or not
- The transformations define the invariance
- It this example, zoom/resize and color shift transformation makes the latent representations invariant to zoom level and color
- Learning is completely self-supervised because the transformations can be computed automatically

Pretext tasks

- Autoencoders and contrastive learning are instances of self-supervised AI, i.e. AI that are supervised, but with a supervision not requiring human intervention

- Another way to pretrain an AI is to supervise it with pretext task, i.e. a tasks loosely related to the final application

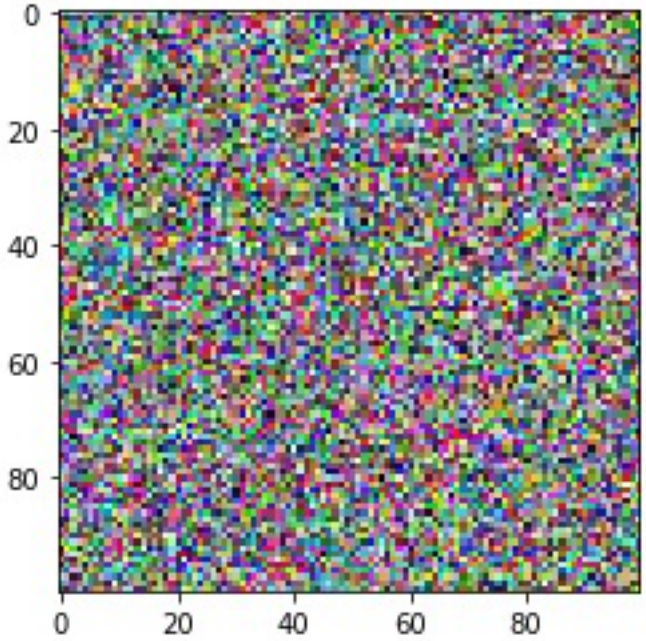
- For example, the AI underlying your project was trained to classify images from the Internet into 1000 categories of everyday life (e.g. cat, banana, train, table, ball, etc.).

- *This AI, has not seen any histological images during training*, yet this is the best we got so far for our exploration of histology!

- This is not completely unexpected after all: the fundamentals of pathologists' vision were not learned in medical school, but in their cradle while they were babies randomly gazing at their environment

The core idea of AI

Most random pixel configurations looks like this



By comparison, real world images are very special. For example, nearby pixels are typically correlated.

- Images sit in a very high dimensional space (here, 100x100x3 variables)
- The vast majority of pixels configurations do not exist in the real world
- Images of the real world are embedded in a small part of the image space
- This is even more true of image in sub-domains, like cats, dogs, human faces, etc.
- AI algorithms are statistical machines that learn these embeddings from data

Scientific background

The Canguilhem study

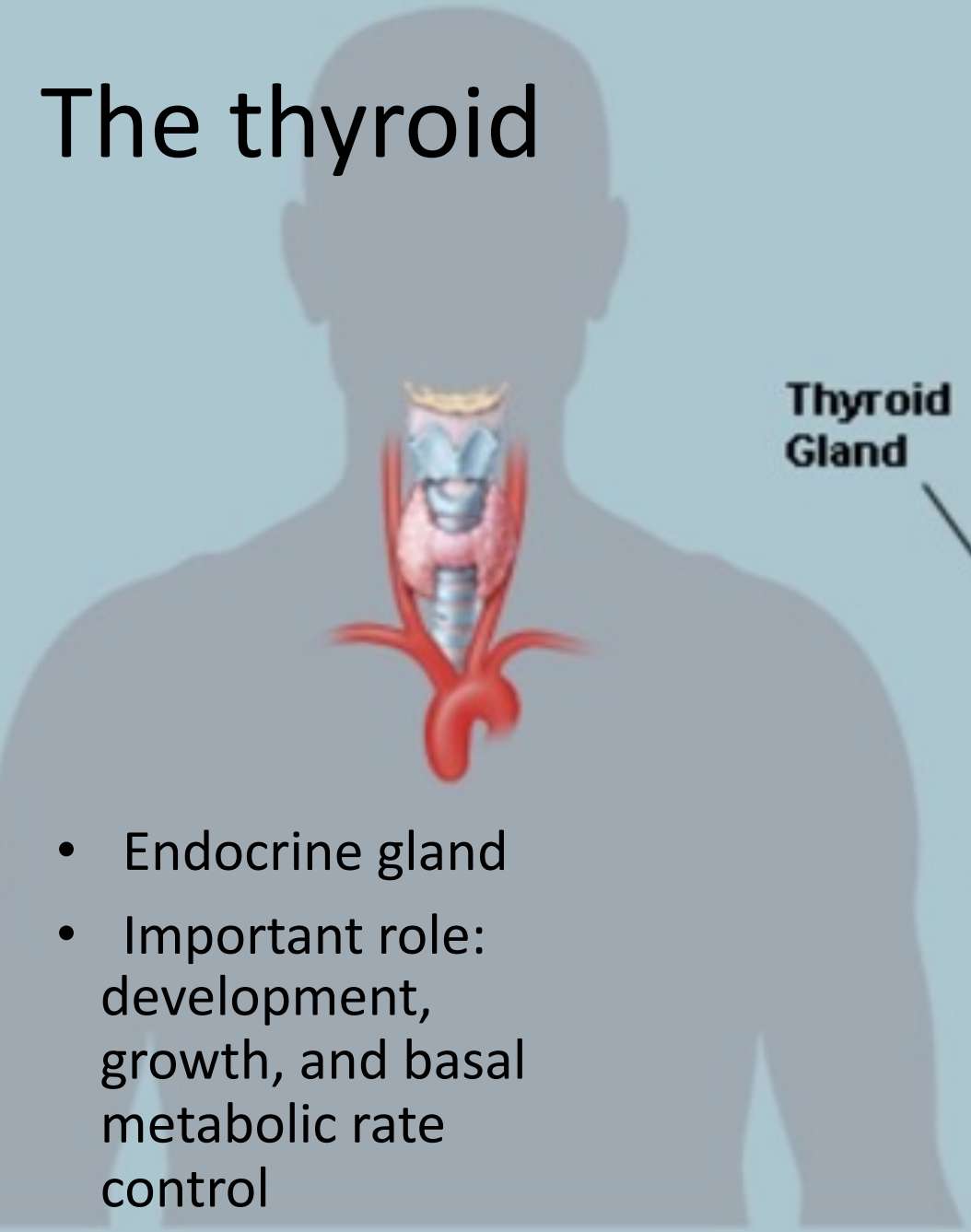
Canguilhem and the normal thyroid

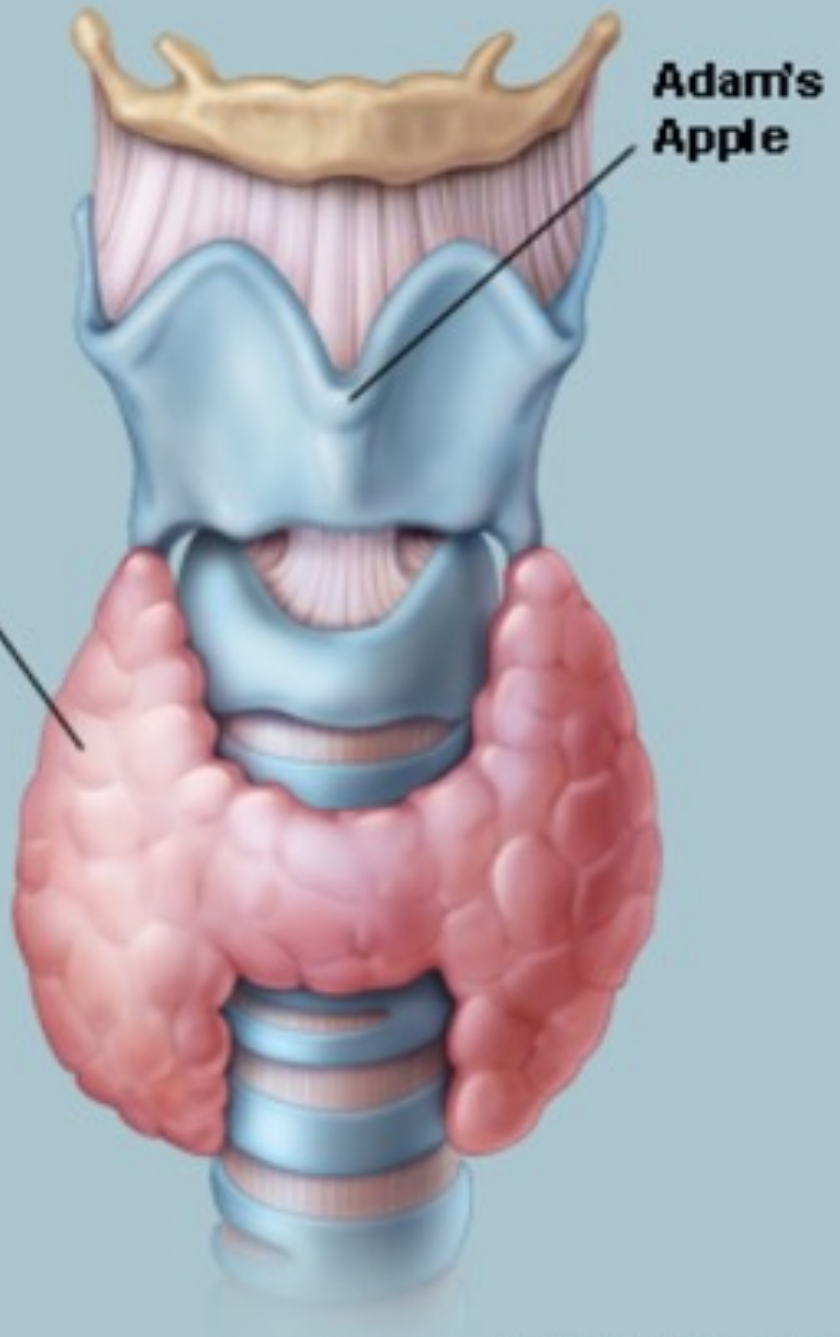
- Essay "*Le normal et le pathologique* "
  by Georges Canguilhem, 1943
  - criticized the definition of disease states as mere deviations from a norm

- This is very relevant to the thyroid cancer overtreatment problem

- My lab propose:
  - to survey **the diversity of normal thyroid morphologies**
  - to achieve this we will develop a novel **unbiased** and **quantitative** approach to morphology

The thyroid

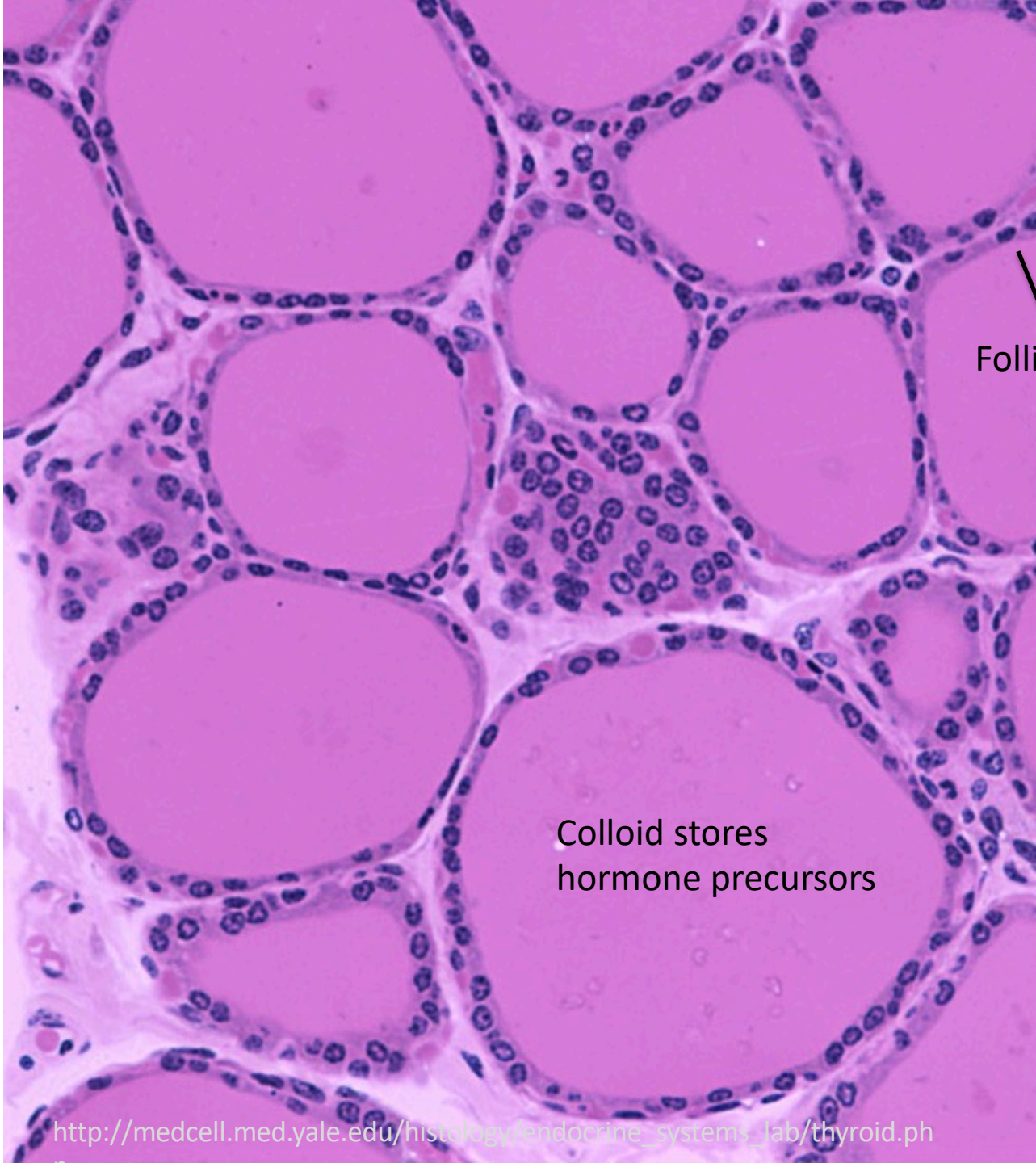




- Endocrine gland
- Important role: development, growth, and basal metabolic rate control

Thyroid follicles

- The main functional units of the thyroid are follicles containing colloid
- Colloid is a viscous material composed predominantly of the thyroid hormone precursor protein
- This image is the textbook ‘cliché’ of the perfectly healthy thyroid of a young subject



Normal thyroid: asymptomatic population vs. textbook

Thyroid nodules are found in examinations *not* related to thyroid disease:

- neck palpation: 21% of patients
- ultrasonography: 67% of patients

Autopsy series suggest that:

- undiagnosed microcarcinomas in 22-36% of cases
- autoimmune thyroiditis in
  - 27% of women
  - 7% of men

Overall goal of the study (**your project** and beyond)

A. Survey the morphologies present in a large collection of asymptomatic thyroid histological slices

B. Survey the correlation of morphological variations with
   1. genotypes
   2. **Gene expression**
   3. **Clinical data**

- In your project you will explore B2 and B3!
- Don't worry, we have computed morphological variations across 893 thyroid slices from GTEx (see 1st lecture)
- I'll provide nice pre-formated matrices for morphological variations, clinical data and gene expression

Technical aspects

GTEx thyroid samples

Overview of GTEx

- 54 organs from 948 donors, most with
  - High-throughput genotypes
  - Genome-wide gene expression (RNA-seq)
  - Histology images (20X scans)
  - donor level clinical annotations

- GTEx samples were collected postmortem

- For an overview of the GTEx v8 release, see https://www.gtexportal.org/home/tissueSummaryPage

Overview of GTEx

For the thyroid there are

– 893 histology images

– 893 clinical records

– 465 RNA-seq profiles

Each project group will be provided a subset of 100-200 samples, each with a specific clinical bias. So, don’t copy/paste results of your classmate. And don’t worry if they get result different from yours.

Clinical data: demographics/health

The clinical data matrix (clinical-data.tsv) includes,

AGE: age at death

SEX: 1-male, 2=female

HGHT: height of donor

WGHT: weight of donor

BMI: general indicator of the body fat an individual is carrying based upon the ratio of weight to height

Clinical data: technical

The clinical data matrix (clinical-data.tsv) also includes,

COHORT: two types of donors, ‘Organ donor’ or ‘Postmortem’

TRISCHD: Ischemic Time (minutes). It’s the time elapsed between the presumed donor death and tissue collection.

DTHHRDY: Hardy scale. A number from 0 to 4 summarizing the circumstances of death

0=Ventilator Case
1=Violent and fast death
2=Fast death of natural causes
3=Intermediate death
4=Slow death

Details about clinical annotations here:

https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs000424/phs000424.v8.p2/pheno_variable_summaries/phs000424.v8.pht002742.v8.GTEx_Subject_Phenotypes.data_dict.xml

Clinical data: miscellaneous

The clinical data matrix (clinical-data.tsv) also includes,

SUBJID: the GTEx ID of the subject, e.g. GTEX-111CU

SMPLID: the GTEx ID of the organ, e.g. GTEX-111CU-0226

SMPTHNTS: the sample's pathology notes taken by GTEx pathologists who examined the histological slices (possibly useful if you want to dive deeper into the project, they found quite few incidental diseases, failed dissection, etc.)

IMGURL: link to the interactively zoomable high resolution scan of the histological slice, e.g. https://brd.nci.nih.gov/brd/specimen/GTEX-111CU-0226

Gene expression matrix

I'll also provide expression data (RNA-read-counts.tsv):

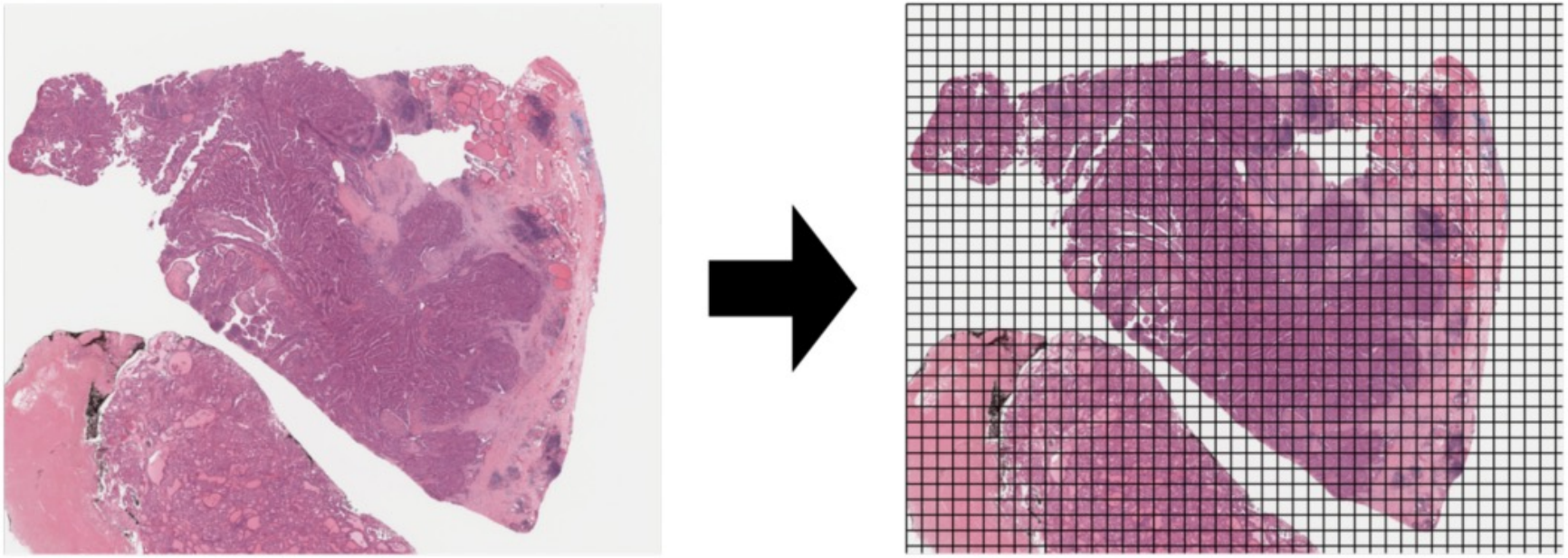
- There is one row per transcripts (56200 in total)
- Transcript ENSEMBL IDs and gene symbols are in the first and second columns
- Beside these, each column stands for a sample
- Expression was recorded as raw, unormalized, read counts, i.e. a read count of 11 for gene G means that 11 reads align on G
- Raw read counts are suitable inputs for differential gene expression software such as DESeq2 or edgeR
- Column names of the expression matrix are indexed with **SMPLID**

Morphological count matrix

I’ll provide a count matrix of the morphologies identified by AI (morphological-counts.tsv)

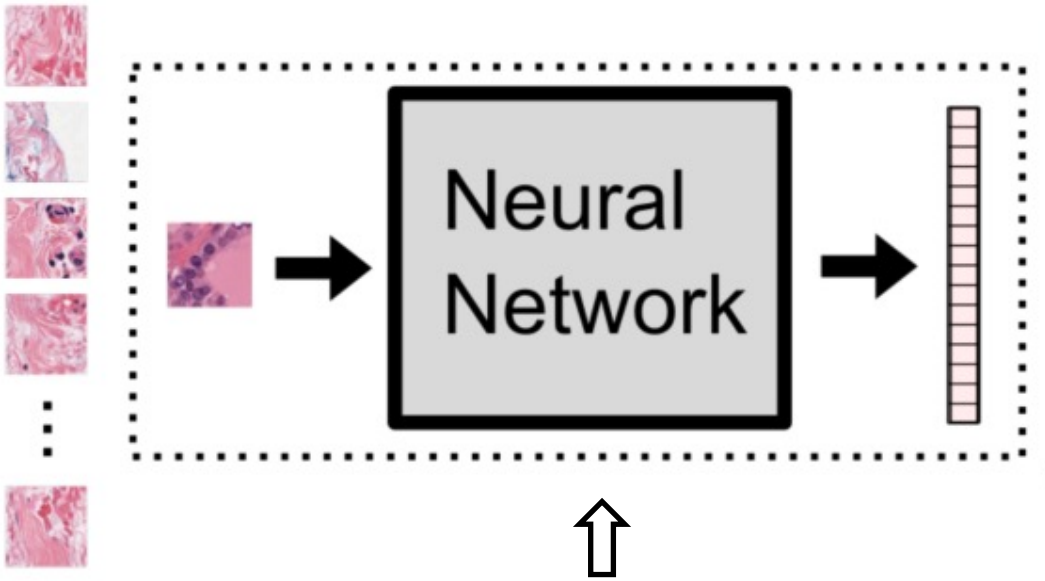
Here is how it was computed…

Step 1: tile the high-resolution image



- Images were scanned at 20X magnification
- Each tiles is 224x224 px^2 or 110x110 $micron^2$
- There are 5,000-10,000 tiles/image
- And about 7,000,000 tiles in the 893 thyroid images

Step 2: compute tiles' latent representations

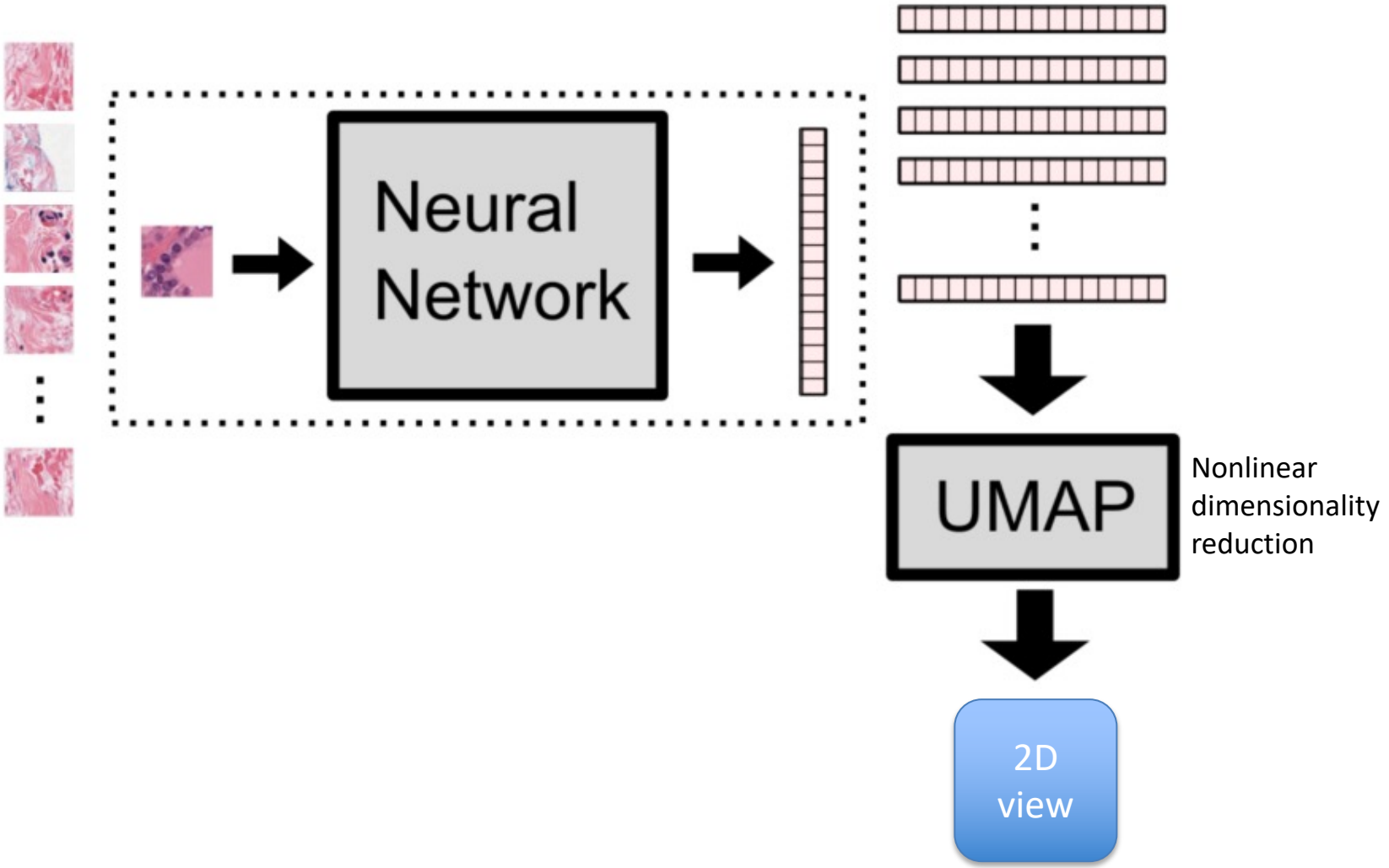


The network is trained beforehand (with a supervised pretext task *not* related to histology, see previous slides)

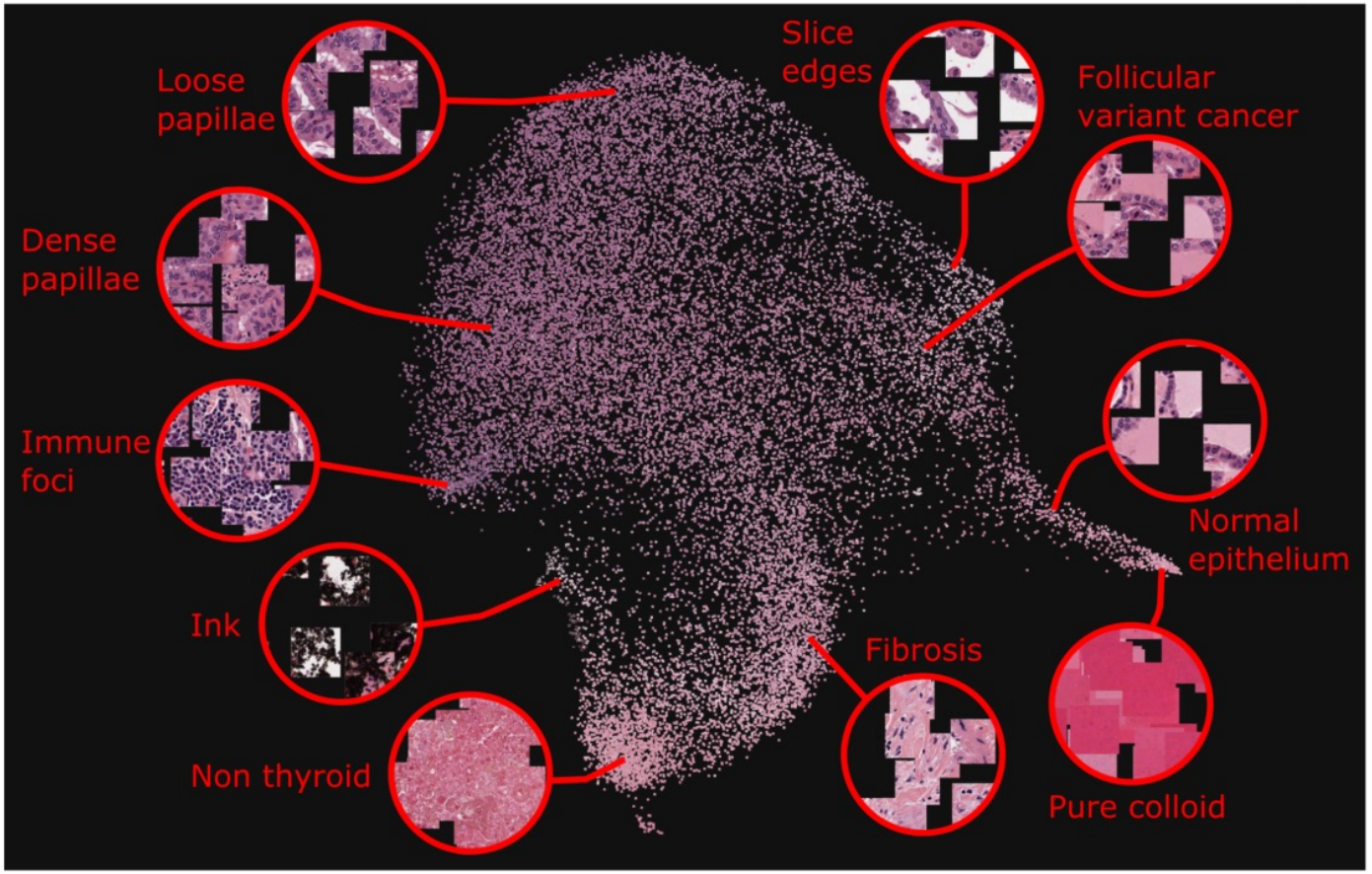
Now a tile is a point in a 384 dimensional space,

384 << 224x224x3

Step 2: display tiles' latent representations (not part of project)

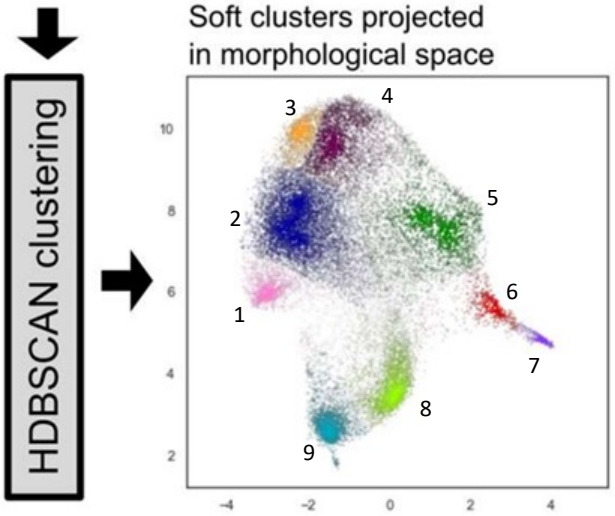


Step 2: display tiles' latent representations (not part of project)



Check out interactive version at
https://www-hpda.ulb.ac.be/iribhm/ai/morphological-space/demo/#

Step 3: cluster tiles in latent space



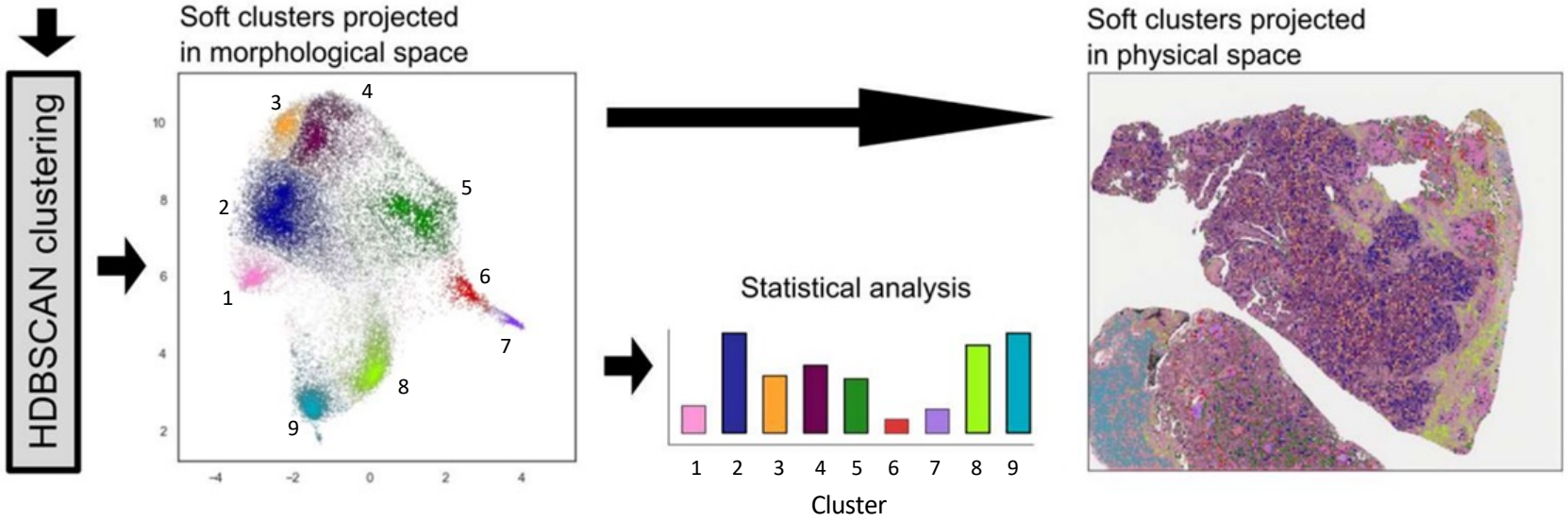
We clustered the tiles based on the latent representations from step 2

For this project we defined 64 clusters, we call these ‘morphological clusters’ (64 is somewhat arbitrary, we are still working on this)

The clustering is done on the 7,000,000 tiles at once, so morphological clusters are representative of the 893 thyroids in GTEx

An atlas of the morphological clusters is available in resource file morphological-atlas.pdf

Step 4: cluster tiles in latent space



One we have the morphological cluster identities of all tiles in a histology image, we can

- Compute the number of tiles in each cluster, that's a morphological summary, and use is in further calculation. That's what your project is about!
- Overlay morphological cluster IDs on the actual image
- Etc.

Morphological count matrix

- I'll provide a matrix of the count of tiles in the morphological clusters (morphological-counts.tsv)
- The columns stand for morphological clusters (indexed from 0 to 63)
- The line stand for samples (indexed by **SMPLID**)
- For example, if matrix entry (GTEX-111CU-0226, Mophological-cluster-23) = 286 :
  - It means thyroid GTEX-111CU-0226 has 286 tiles in Mophological-cluster-23.
  - Since tiles have a defined physical surface, 110x110 micron^2 (0.11x0.11 mm^2) it also mean that Mophological-cluster-23 spans 286x0.11^2=3.5 mm^2 in histological image of thyroid GTEX-111CU-0226

Morphological count matrix

- Thus, just as the RNA-seq read count matrix gives the expression of genes across the samples, the morphological count matrix gives the **expression of morphological clusters**

- Both are count data!

- Just as you will do with RNA-seq counts, I suggest you analyze morphological differential expression with DeSeq2 or edgeR

Your project step-by-step

Questions to be addressed

Overview of the project

In a nutshell,

The overall goal is to compare the morphological description of tissues, thyroid glands, to their transcriptome

The morphological descriptions rests on unsupervised deep learning models

Q1: explore clinical variables

1. How are they distributed?

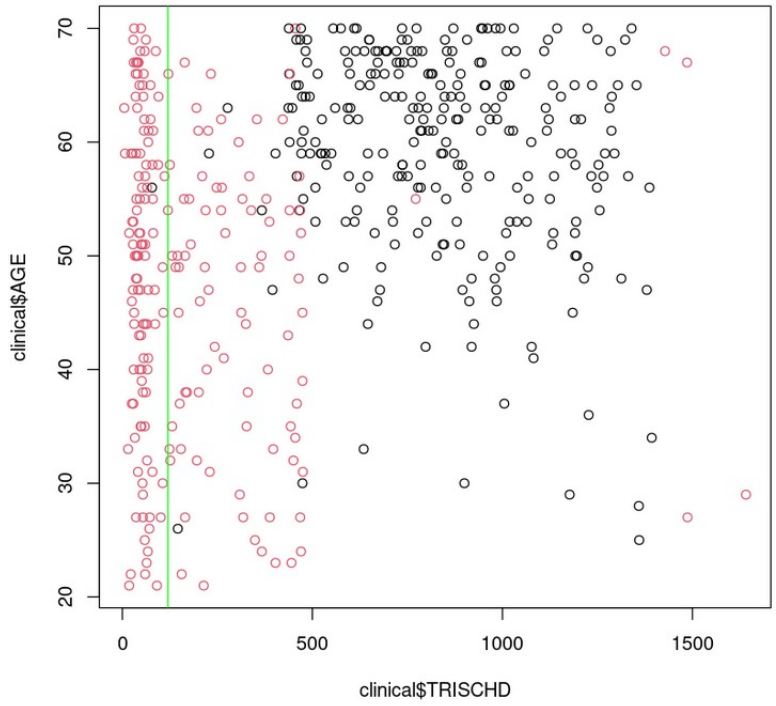
2. Are they correlated?

3. In particular, are some technical variables possibly confounding demographic/health variables?

Hint: Tools to address this may include PCA analysis, showing a variable's correlation matrix, specific plots,…

Hint: **DTHHRDY** is more likely a categorical variable.

Q1: explore clinical variables



COHORT:
Organ donors
Postmortem

Hint: Q1.3 is very important!

Here we see that AGE is confounded by TRISCHD and COHORT

Thus, any statistical analysis of AGE will need to be adjusted (i.e. normalized) for TRISCHD and COHORT

(Your data won't necessarily look like this…)

Q2: Clinical data vs. morphology

1. Compute systematically associations between clinical variables and morphological cluster counts. The purpose is to compare the magnitude of the associations of the different variables with morphology.

2. Discuss the association with technical variables.

3. For non-technical variables, redo the analysis with adjustment for the confounding technical variables, if any is reported in Q2.2. Report and discuss significant associations.

Q2: Clinical data vs. morphology

Hint: This is a differential expression analysis where we study 'differential **morphological cluster** expression' instead of differential **gene** expression. So, I suggest you use proven tool developed for transcript count data, like DESeq2, edgeR, etc.

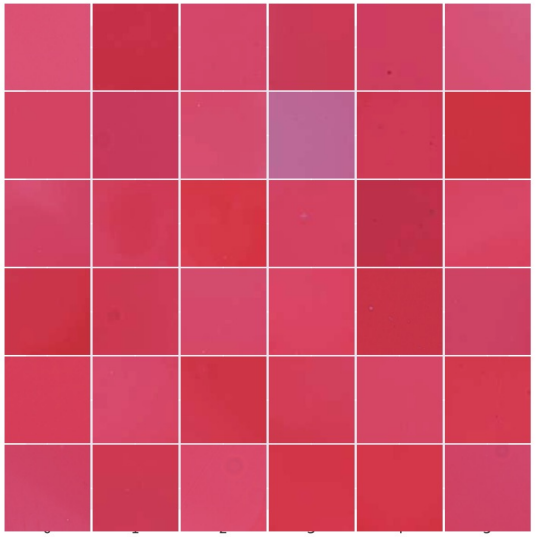
Hint: These tools implements multivariate model formulas that will enable the treatment of confounders in Q2.3. For example, formula `'~AGE+X+Y'` will compute the multivariate association of morphological clusters with variables AGE, X and Y. So, if you look at AGE from this model it will be adjusted for the variations of X and Y.

Thus, DESeq2 not only automatically handles normalization for total count and other intensity biases of count data, it also enables you to cancel out your variables of choice in the analysis.

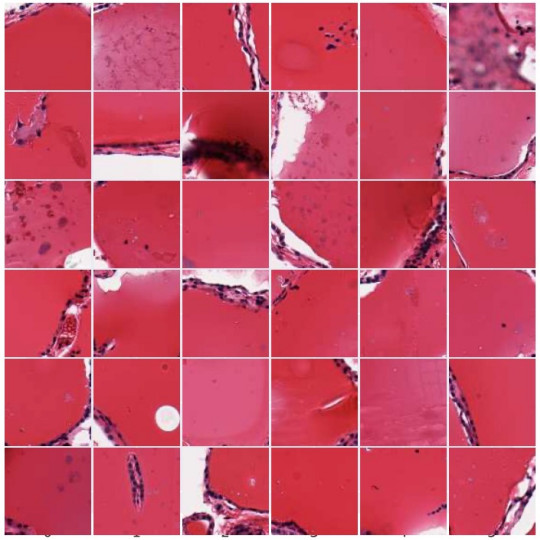
Hint: the function 'as.formula' (turns char string into a formula) could be useful to automate your analysis…

Q2: Clinical data vs. morphology

36 tiles from Morphological cluster 59 (colloid)



36 tiles from Morphological cluster 13 (edge of follicle)



Hint: Counts of different morphological clusters may be correlated with one another, as shown in the atlas. One reason is that patterns on a scale of 110x110 micron^2 may by part of the same larger-scale morphological entity.

A trivial example: these two clusters are both part of follicles (center and edge), so they tend to co-occur in the same samples

Less trivial sources of correlation include disease states, e.g. autoimmunity

By the way, the order of the cluster in the atlas is based on a (meta-) clustering of the morphological cluster counts.

Q3: Morphology vs. gene expression

1. Report:
   - The number of significant down-regulated and up-regulated genes associated with each morphological cluster
   - Report the the 10 most significant up-regulated genes

2. Same as Q3.1 with REACTOME gene sets

3. Discuss the results, technically and biologically

Q3: Morphology vs. gene expression

Hint: Use dedicated RNA-seq tools that take count data as input (e.g. DESeq2, edgeR, etc.) for Q3.1

Hint: Morphological cluster counts are directly related to image size, and image size depends on how the pathologist cut the organ as much as on its actual volume. It's more relevant to look for the genes associated with *morphological cluster proportions* rather than raw counts.

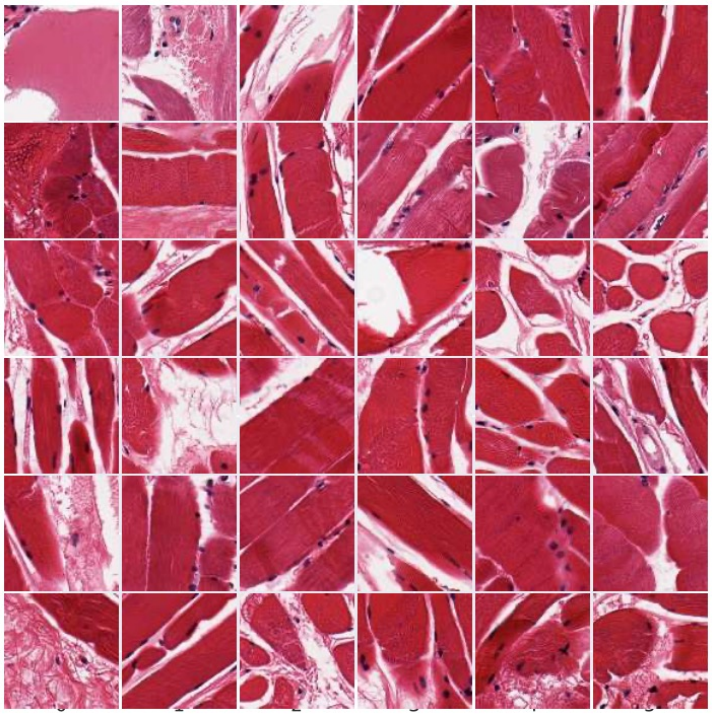
Hint: I highly recommend you filter transcripts beforehand. You don't need transcripts that are not/little expressed in the thyroid. It's also a good idea to focus on transcripts showing high variability (use the median average deviation, a.k.a. mad() in R). Don't hesitate to be drastic in your filtering (e.g. it's OK to remove >50% of all transcripts)

The filtering reduces the multiple testing problem, of course. But but it also reduce the computational burden. After appropriate filtering, I could run DESeq2 analyses for all 64 morphological clusters on 465 samples in ~1 hour and a peak RAM usage < 4GB

Hint: Just as in Q2, you may need to adjust for technical confounders.

Hint: DESeq2, edgeR handle multiple testing. But you'll have another layer of multiple testing since you will run these algorithms multiple times.

Q3: Morphology vs. gene expression



36 tiles from morphological cluster 14

Hint: Surgeons sometime take muscle adjacent with the thyroid

That's another technical confounder

Our AI identified a 'muscle cluster' (Morphological cluster 14)

You may use it as an adjustment covariate in your analysis, rather than as a target variable

Q3: Morphology vs. gene expression

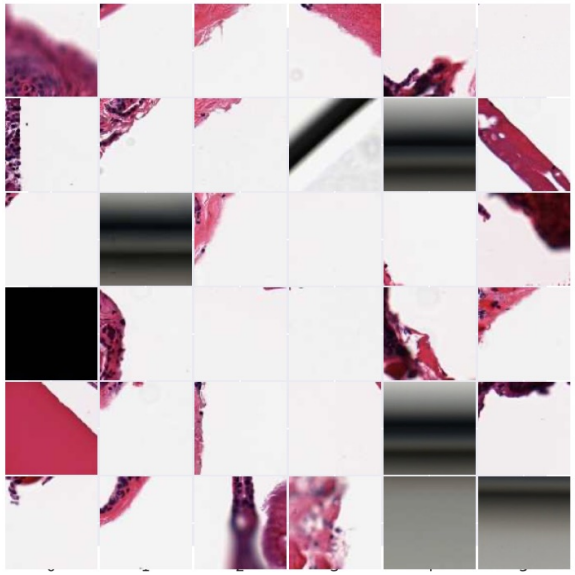
Hint: In Q3.2, running 64 times GSEA may not be computationally tractable. I suggest you use the fgsea package, which takes as input a gene ranking (i.e. provided by the differential analyses of Q3.1 output)

(Yes, I know, it's not using the the gold standard sample permutation of GSEA. But I tried it for you and it's reasonable in this context where the transcriptional signal is quite massive)

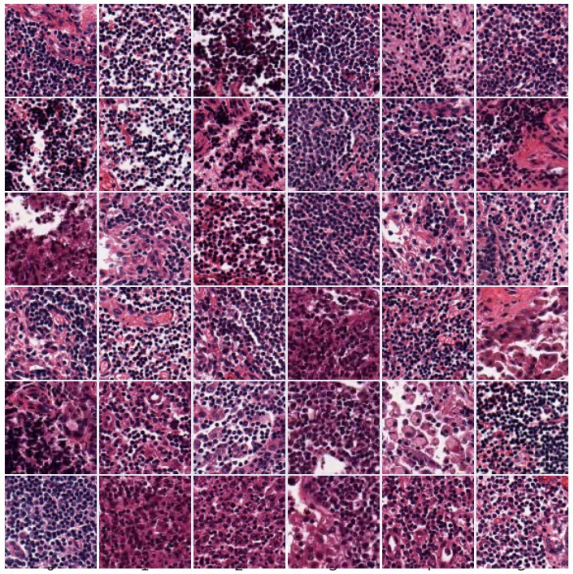
Hint: The REACTOME gene sets are provided in resource file c2.cp.reactome.v7.5.1.symbols.gmt. Fgsea provides a function to read *.gmt files

Q3: Morphology vs. gene expression

Morphological cluster 58 (tissue edges and various artifacts)



Morphological cluster 45 (dense lymphocytes aggregates)



Here are two remarkable clusters to guide Q3.3.

Hint: Think of them as negative and positive controls.

Hint: Can you estimate the prevalence of inflammation in your collection of asymptomatic thyroids?

Your project step-by-step

Handing out your report

Project report

Your are asked to describe your work in a written report

It includes :

- An short introduction to the topic
- A careful description of methods and results
- A discussion of your results for each questions

Dumping code and results is *not* good enough, you must show me you understand what you are doing

Project report

- The report is handed out as a **PDF** file or a **Jupyter notebook**

- Since I am asking quite a few results (e.g. series of genes differentially expressed in 64 analysis), you may put the larger outputs as appendices to the report, or hide them in your notebook, and discuss the most salient findings in the main text

- I am not defining a specific report length. Good reports are neither too short, nor too long. 15-20 pages (appendix not included) seems a reasonable balance.

Project report

- This report is a good exercise for your master thesis and for scientific writing in general

- Try to be specific, precise and quantitatively accurate. For example, don't write '*many genes are associated with cluster Z*', but '*N genes are associated with cluster Z*', where N is the number you have computed

Grading of your projects

- Questions Q1-Q3: **15pt**
- Quality of your PDF report (it needs to be concise, clear, precise, well presented), **3pt**
- Reproducibility: all your analyses must be reproducible, i.e. all the steps are carefully documented, including parameter settings and software versions, **2pt**
- …and **2pt** to be gained beyond 20 for the motivated ones who go off the beaten track with original analyses, etc.

Handing out your project report

- Deadline is June 9th 23:59:59.
- Send me the report by e-mail (Vincent.Detours@ulb.be), including 'BINF-F401' in the subject line
- Get back to me if I don't acknowledge receipe of your report within 3 days.

Your project step-by-step

Rules of the game

Groups

- There will be 9 groups of 3 and one group of 4 students
- I'll form the group at random to avoid social biases
- Each group is assigned a specific subset of the data, some with a biased selection of donors
- You will be informed by mail about which group you've been assigned to

Rules of the game

Choose your own weapons. I do advise R, but accept alternatives.

You are encouraged to communicate with one another, but

- Remember, each group has its own unique dataset
- Beware of herd effects, the majority and/or the leading figures can be dead wrong
- Plagiarism is not acceptable

Rules of the game

You will be able to ask questions during the remaining two course sessions

I'll also address your questions by e-mail, if

- You use your brain and do your home work before asking, e.g. I won't reply if the answer is in the R documentation or can be obtained from a basic web search. (I basically request the same etiquette as in any technical forum on the Internet.)

- The subject line of all your correspondance with me must start with 'BINF-F401: '

Some tips from past experience…

- This science project your technical choices must be argumented and the calculations must be reproducible, i.e. all the information needed to rerun the analyses must be provided. *In science details do matter*.

- Dumping graphics and stats in a document is not enough. The students who got the best grades examined critically their results and demonstrated that they understood the limits and biological significance of what they did.

- Technical problems are hard to anticipate, so don’t wait for the last minute to discover them, and then be left with no time to overcome them.