Acknowledgement of Country

- The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.
- We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.
- We recognise their valuable contributions to Australian and global society.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

General Information:

- We are currently located in Building 69

Emergency evacuation point

- Food court and bathrooms are located in Building 63

- If you are experiencing cold/flu symptoms or have had COVID in the last 7 days please ensure you are wearing a mask for the duration of the module

Data Agreement

To maximize your learning experience, we will be working with genuine human genetic data, during this module.

Access to this data requires agreement to the following in to comply with human genetic data ethics regulations

If you haven't done so, please email <ctr-pdg-admin@imb.uq.edu.au> with your name and the below statement to confirm that you agree with the following:

"I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts."

Learning materials

Instructions to access WiFi/desktop/server:

https://suave-pillow-de4.notion.site/Instruction-to-Computing-Resources-dcba658c9a584e6d80a443c5d64042d8?pvs=4

Slides and practical notes:

https://cnsgenomics.com/data/teaching/GNGWS24/module[1-6]/

Module 2 - running the learning materials

https://github.com/GenomicsMachineLearning/qimr-teaching-2024/tree/main

---

Copy and paste each of the following lines into your terminal once you have logged into the workshop server:

- `/software/bin/micromamba shell init`
- `source ~/.bashrc`
- `micromamba activate /software/conda-envs/winter_school_2024`
- `git clone https://github.com/GenomicsMachineLearning/qimr-teaching-2024`
- `~/qimr-teaching-2024/runme.sh`

The output will look something like:

```
Port 3502 is available

Command to create ssh tunnel:
ssh -N -L 3502:10.10.10.10:3502 foo@10.10.10.10
Use a Browser on your local machine to go to:
localhost:3502  (prefix w/ https:// if using password)

[I 2024-06-20 05:57:41.633 ServerApp] Extension package jupyter_lsp took 0.1372s to import
[I 2024-06-20 05:57:44.647 ServerApp]     http://127.0.0.1:3502/tree?token=abc123
```

- Copy the line beginning with "ssh" into a new terminal, on your local computer, and hit [Enter].
- Copy the text beginning with "http://127.0.0.1" into a new tab in your browser, and hit [Enter].

Module 2 Cellular Omics

Room 314/315, Building 69

Aiming at interactive session, we provide the presence of a large teaching team for more one-to-one discussion.

Lecturers/Instructors: Quan Nguyen, Andrew Causer, Levi Hocki, Onkar Mulay, Prakrithi Pavithra, Andrew Newman, Xiao Tan, Feng Zhang

Module 2 Cellular Omics – Leaning Objectives

- Technologies for generating single-cell and spatial transcriptomics data
- Technologies for other spatial omics, focusing on proteomics
- Exploratory visualisation to understand the data
- Statistical analyses to discover new biological processes and biomarkers associated with disease, including cells, genes and groups of cells within the tissue. This includes:
  - Identifying cell types
  - Finding gene markers
  - Mapping cell neighbourhoods (cell communities)
  - Analysing cell-cell interactions
- Analysing spatial proteomics data and integration with spatial transcriptomics through imaging analysis techniques
- Machine learning analysis of sequencing and imaging data

Lecture 7: Tissue Segmentation and Spatial Statistics

The picture can’t be displayed.

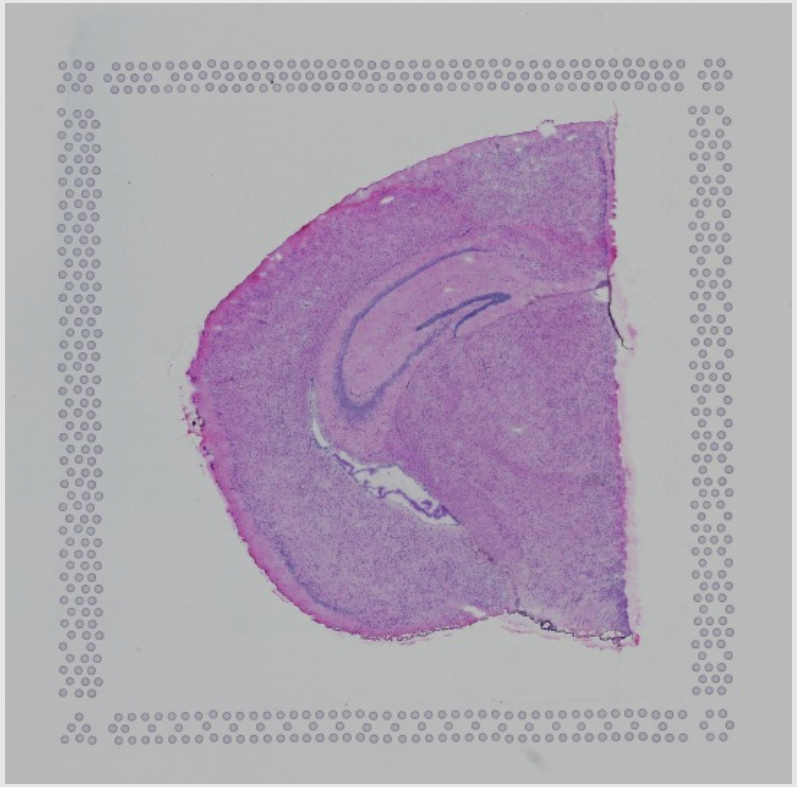
Tissue Segmentation and Spatial Statistics

Andrew Newman and Dr Quan Nguyen

The picture can't be displayed.

Visium Output

- raw_feature_bc_matrix/
  - barcodes.tsv.gz
    - List of labels used for spots (rows)
  - features.tsv.gz
    - Names of genes (columns)
  - matrix.mtx.gz
    - Counts (expression count by row, column)
- spatial/
  - scalefactors_json.json
    - Convert values (pixels, spots) from original to supplied
  - tissue_positions.csv
    - Labels (barcodes) positions in image and if considered in tissue



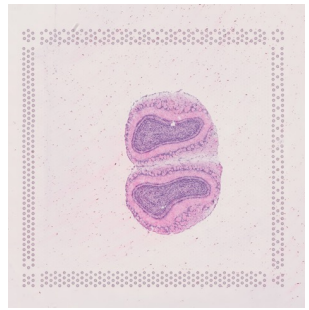
Properly Aligned Image
(triangle bottom-left)

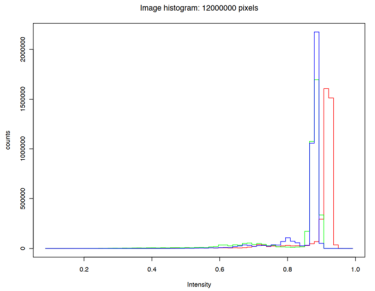
https://support.10xgenomics.com/spatial-gene-expression/software/pipelines/latest/algorithms/imaging
https://www.10xgenomics.com/support/software/space-ranger/latest/analysis/outputs/spatial-outputs

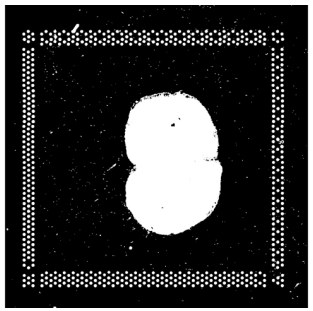
The picture can't be displayed.

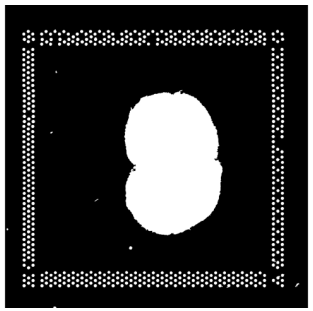
Tissue Segmentation

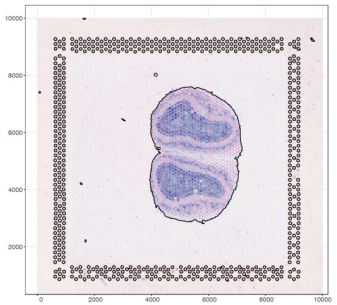
- Use the H&E image to create a boundary between the tissue under examination.
  - Extract R, G, B values
  - Pick the most discriminatory colour (blue or red)
  - Create, cleanup and visualise the mask
  - Align to fiducials
  - Compare with SpaceRanger

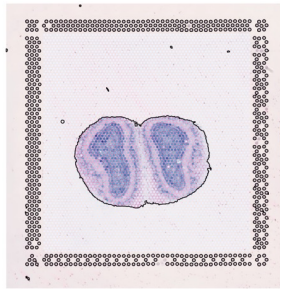


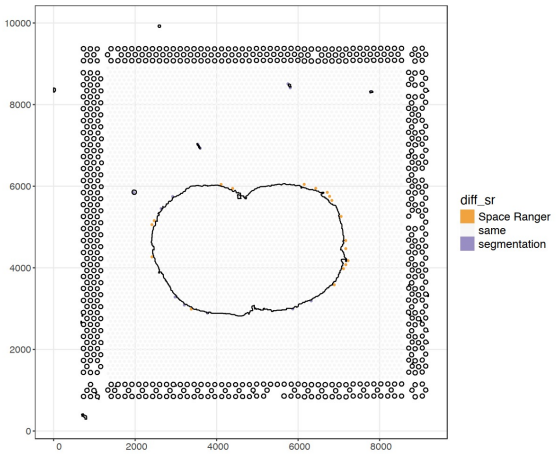






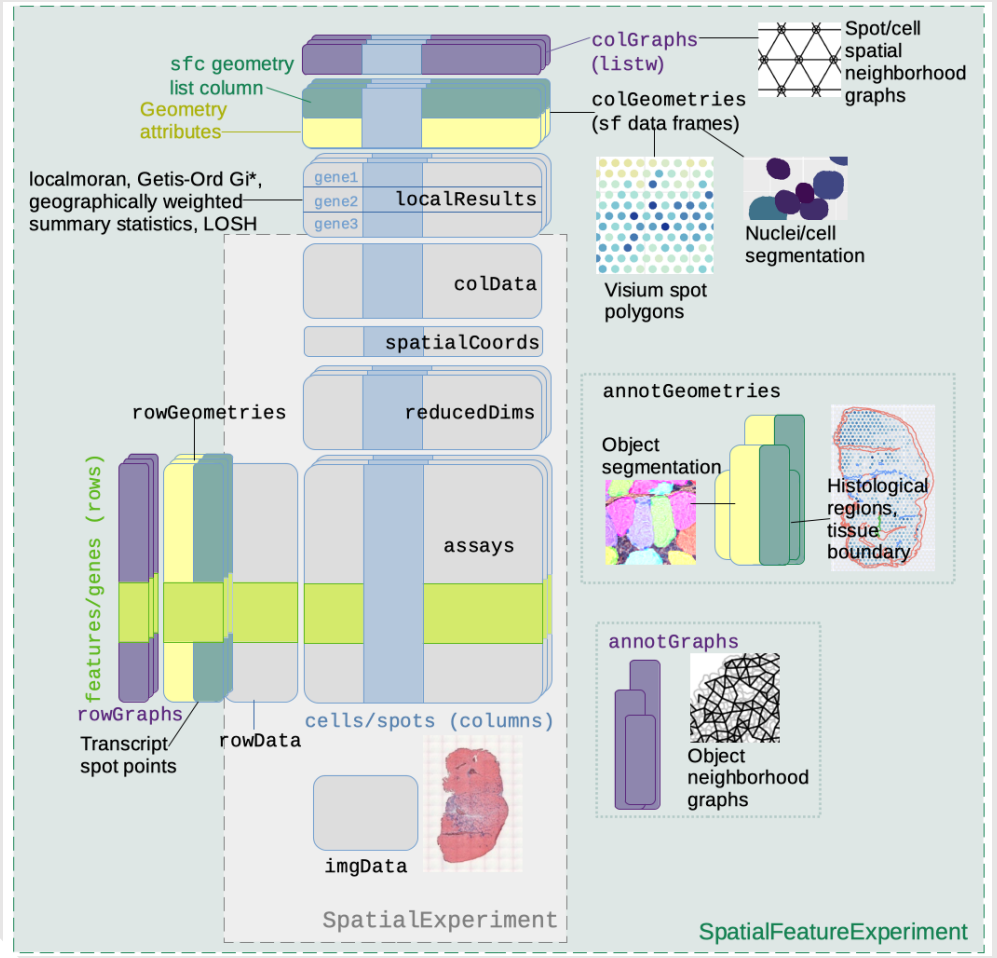






Spatial Feature Experiment

- Heavily influenced by GIS (geographic information systems) high performance geometrical operations operations
- SFE (Spatial Feature Experiment)
  - SPE (SPatial Experiment)
    - SCE (Single Cell Experiment)
- colGeometries/annotGeometries
  - Cells associated with multiple geometries (cell, tissue, etc)
  - Polygons associate with spots, tissue boundary, pathology annotations
- colGraphs/annotGraph
  - Neighbourhood graphs of geometries and their annotations.



The picture can't be displayed.

Exploratory and Confirmatory Data Analysis

Exploratory Data Analysis

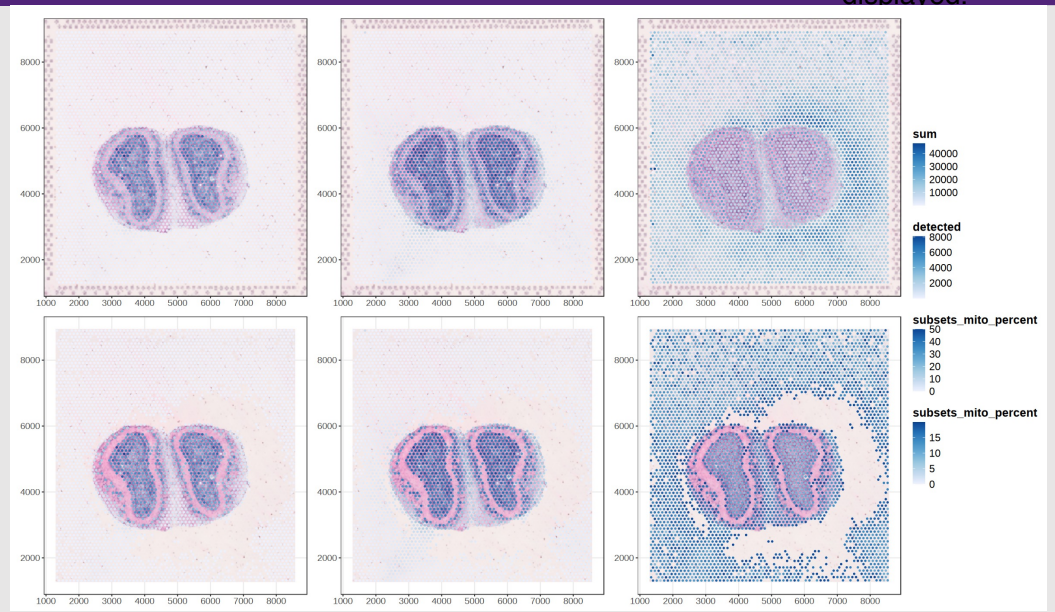
- Visualise, summarise, and analyse data sets without making assumptions about their distribution or requiring a predetermined hypothesis.
- Tasks:
  - Dimension reduction,
  - Unsupervised clustering,
  - Visualisation of data distribution, variance and skew.

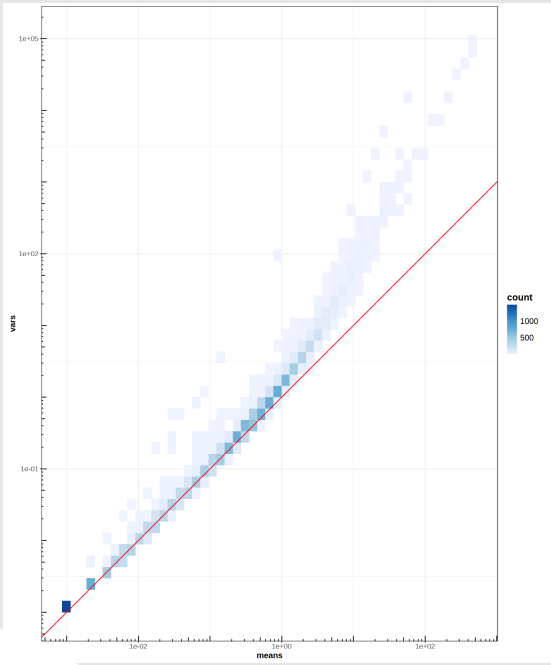
Confirmatory Data Analysis

- Test and refine hypotheses generated by EDA or based on prior knowledge
- Using data with known biological meaning such as gene markers, protein-protein interactions, ligand-receptor interactions, pathway analysis and cell atlases.
- Tasks:
  - Cell type annotation and deconvolution,
  - Spatial clustering,
  - Feature alignment,
  - Network analysis and
  - Trajectory inference.

Exploratory Data Analysis

- Observe the effect of removing mitochondrial expression on the tissue and surrounding areas.
  - Is mitochondrial important or not?
    - Parkinson's, Alzheimer's, some cancers
- The mean and variance are generally correlated
  - Genes with high mean expression and low variance maybe housekeeping genes
  - Genes with high variance related to mean expression may indicate cell type, a condition or response to stimulii

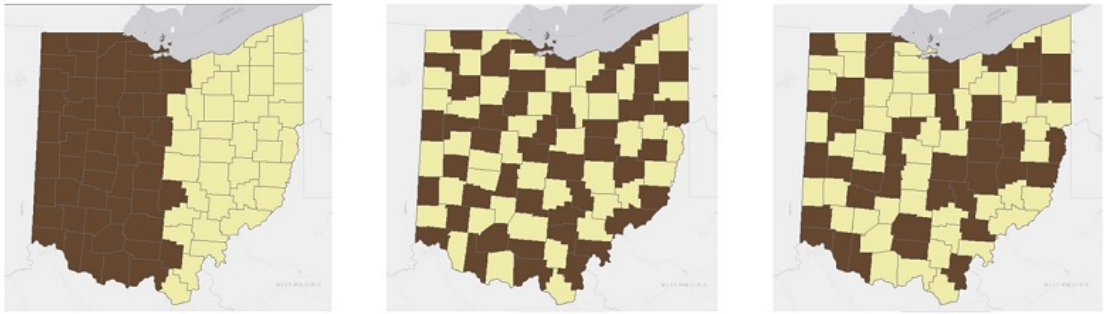




The picture can't be displayed.

Spatial Statistics

- "Everything is related to everything else, but near things are more related than distant things" Tobler's First Law of Geography
- Things change over gradients: house prices, weather, language, etc.
- Spatial Correlation
  - Positive – similar are together,
  - Negative – different are together,
  - None (zero) – random.



Positive autocorrelation | Negative autocorrelation | No spatial autocorrelation

The picture can't be displayed.

Spatial Autocorrelation

- Types of autocorrelation:
  - Variables: univariate, bivariate and multivariate
  - Locality: local and global
  - Algorithms:
    - Moran's I (local or global, univariate)
    - Geary's C (local or globa, univariate or multivariate)
    - Getis-Ord Gi (local, univariate)
    - Lee's L (bivariate)
    - MULTISPATI PCA (multivariate)

Moran's I

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where,

- n is the number of observations,
- x_i is the value at a location,
- x_j is the value at a subsequent location,
- w_{ij} is a weight indexing location of i relative to j (an edge as defined by a weighted neighbourhood graph).

The picture can't be displayed.

Global Moran's I - Example

values = [1, 2, 3, 4, 5]

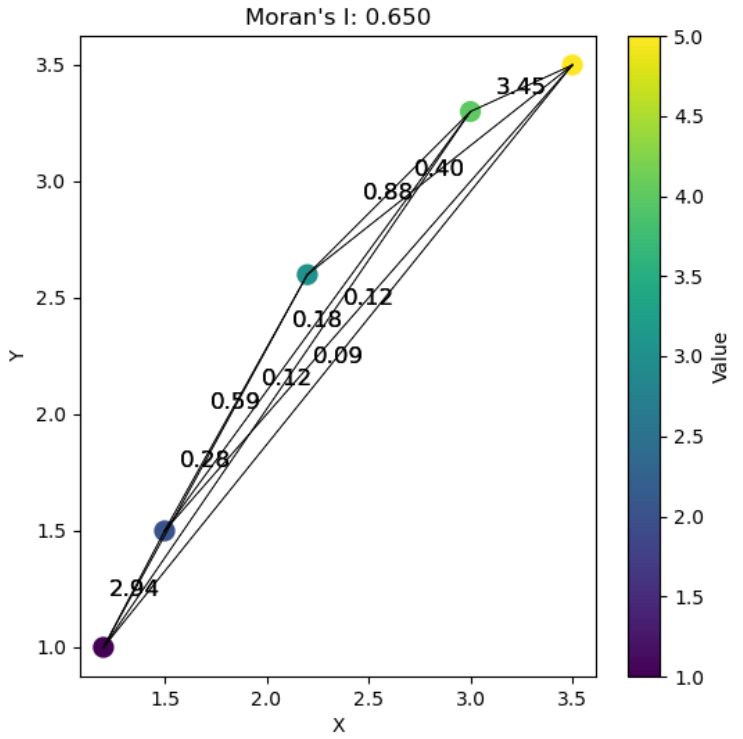
coordinates_1 = [(1.2, 1.0), (1.5, 1.5), (2.2, 2.6), (3.0, 3.3), (3.5, 3.5)]

coordinates_2 = [(1.2, 1.0), (1.5, 1.1), (2.2, 2.6), (3.0, 3.3), (3.1, 3.5)]

coordinates_3 = [(3.1, 3.5), (1.2, 1.0), (1.5, 1.1), (3.0, 3.3), (2.2, 2.6)]

Where, the weight is the squared inverse distance weights using the coordinates.

coordinates_1



The picture can't be displayed.

Global Moran's I - Example

values = [1, 2, 3, 4, 5]

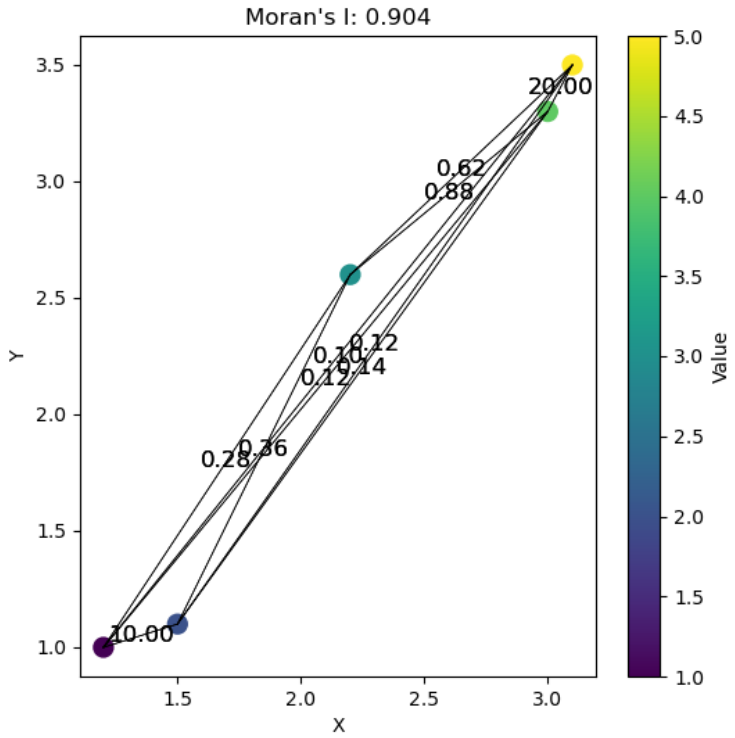
coordinates_1 = [(1.2, 1.0), (1.5, 1.5), (2.2, 2.6), (3.0, 3.3), (3.5, 3.5)]

coordinates_2 = [(1.2, 1.0), (1.5, 1.1), (2.2, 2.6), (3.0, 3.3), (3.1, 3.5)]

coordinates_3 = [(3.1, 3.5), (1.2, 1.0), (1.5, 1.1), (3.0, 3.3), (2.2, 2.6)]

Where, the weight is the squared inverse distance weights using the coordinates.

coordinates_2



The picture can't be displayed.

Global Moran's I - Example

values = [1, 2, 3, 4, 5]

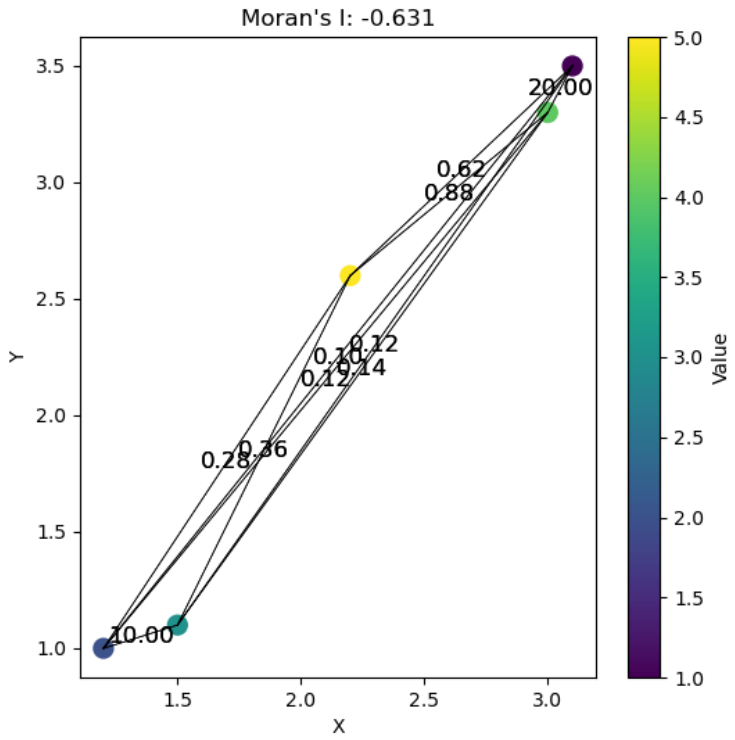
coordinates_1 = [(1.2, 1.0), (1.5, 1.5), (2.2, 2.6), (3.0, 3.3), (3.5, 3.5)]

coordinates_2 = [(1.2, 1.0), (1.5, 1.1), (2.2, 2.6), (3.0, 3.3), (3.1, 3.5)]

coordinates_3 = [(3.1, 3.5), (1.2, 1.0), (1.5, 1.1), (3.0, 3.3), (2.2, 2.6)]

Where, the weight is the squared inverse distance weights using the coordinates.

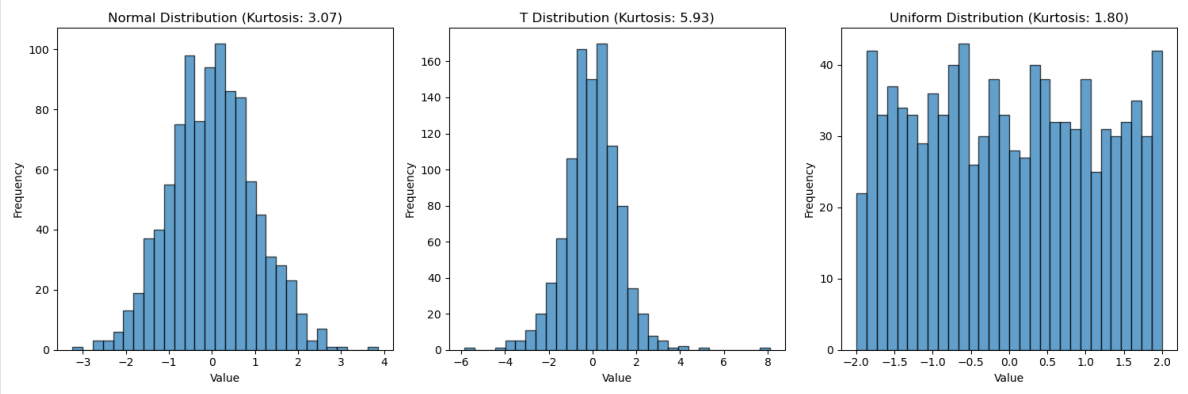
coordinates_3



The picture can't be displayed.

Kurtosis

The "K" value or Kurtosis is a measure of the "tailedness" of a distribution. Higher values indicates a distribution with a higher chance of producing outliers, conversely, lower values have a smaller chance of producing outliers and a value of 3 is the typical value for normally distributed data.



The picture can't be displayed.

Local Getis-Ord Gi*

- Compares the local sum of features and its neighbours to the global sum of features
- Detection and screening of cervical cancer in India
- Hot spots – neighbours with high prevalence of screening
- Cold spots – neighbours with low prevalence of screening

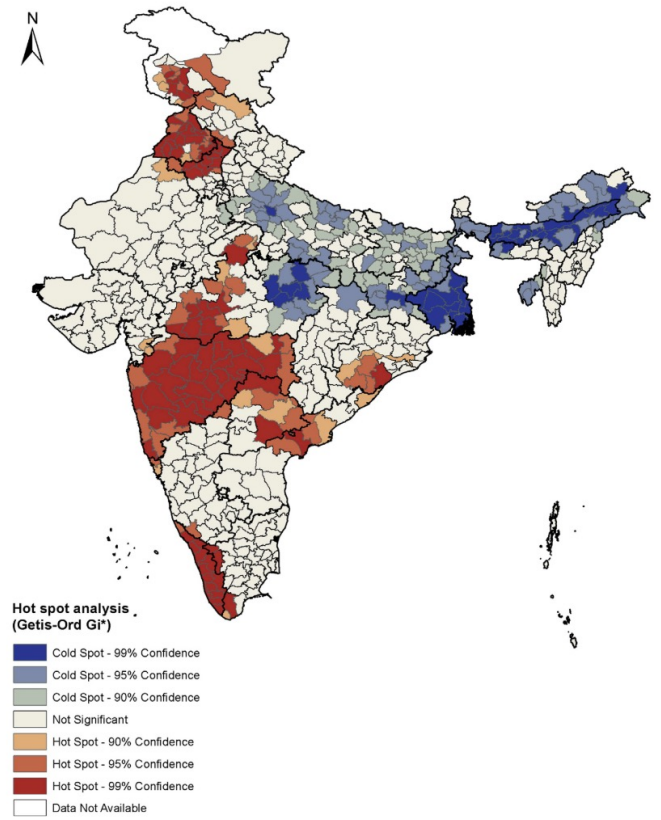


Fig. 2 – Local indicators of spatial association cluster map for the prevalence of screening across the study region.

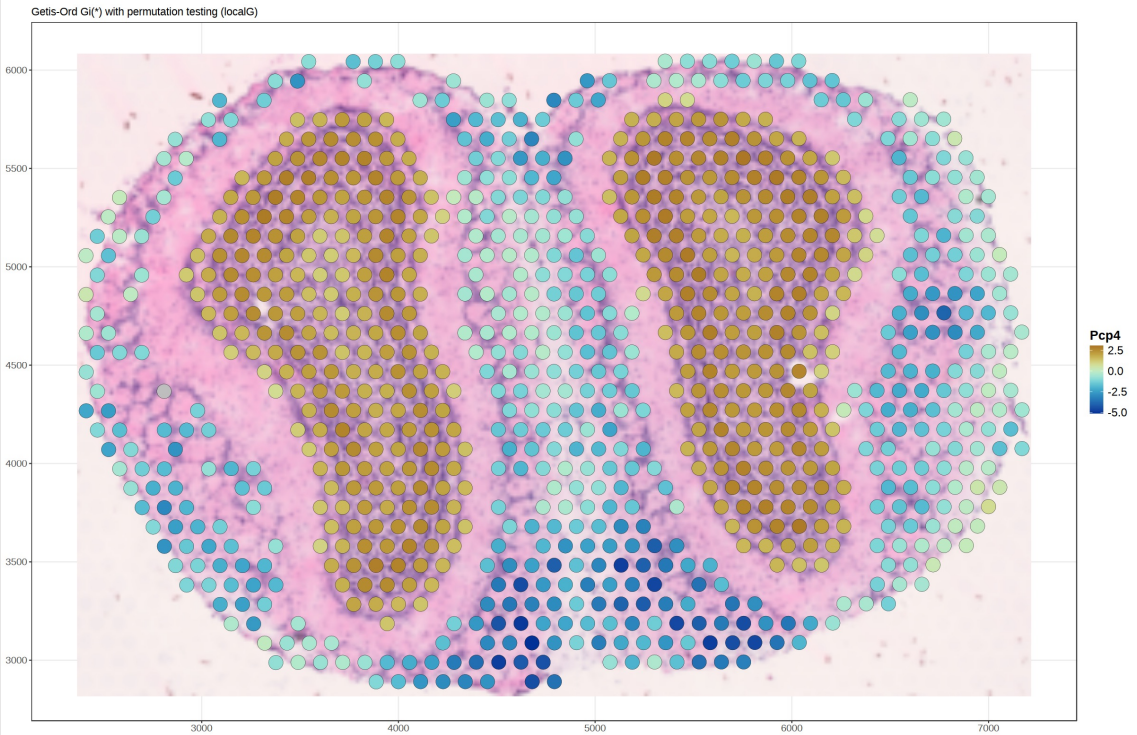
https://doi.org/10.1016/j.puhe.2019.09.008

The picture can't be displayed.

Local Getis-Ord Gi*

Applying gene expression:

- Associated with specific tissue types, developmental stages, or disease states.
- Identification of potential spatial patterns of gene regulation, such as the presence of local enhancers or suppressors that may be influencing gene expression in specific regions.

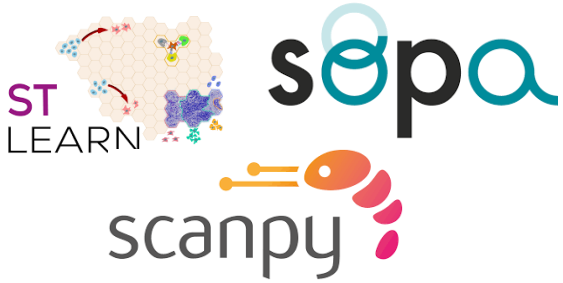


Lecture 8: Spatial Proteomics

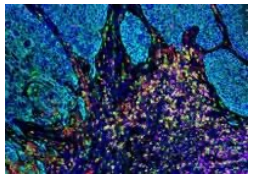
Module 2 – Part 8: Spatial Proteomics

Xiao Tan

QuPath



Python packages



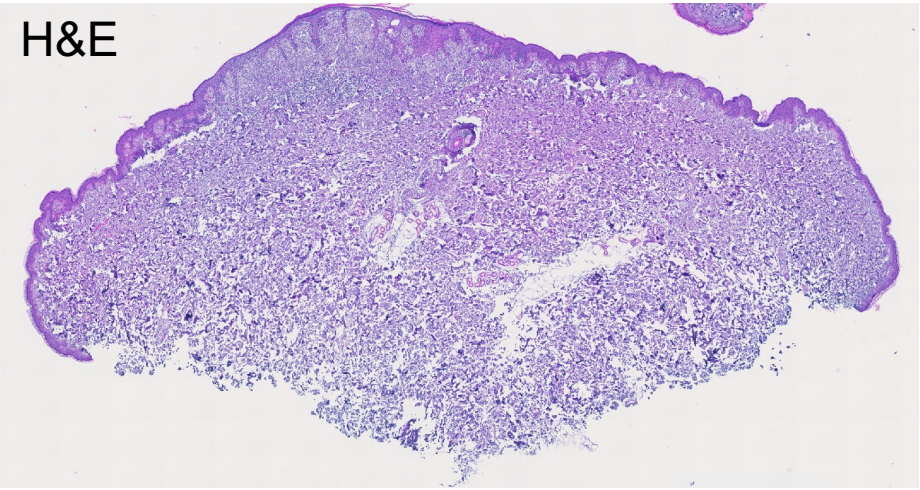
PhenoCycler (CODEX)

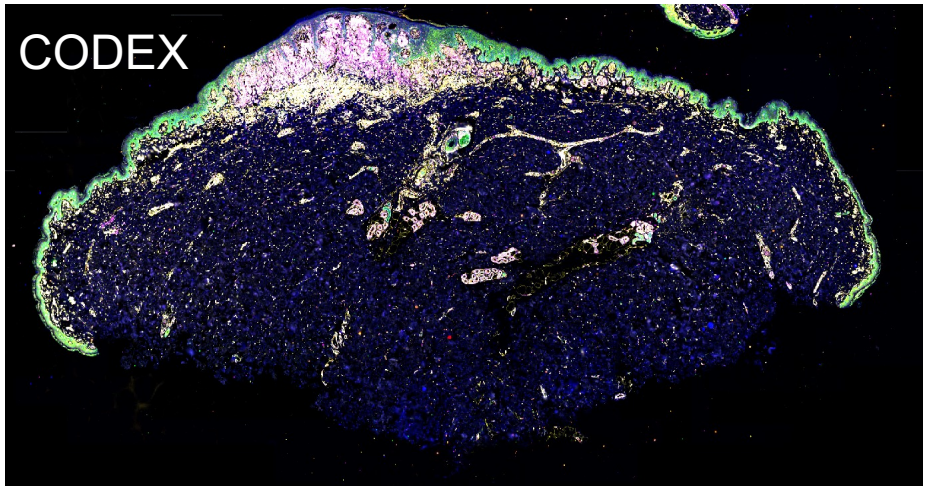
Module 2 – Part 8: Overview

Spatial Proteomics Analysis for CODEX Data

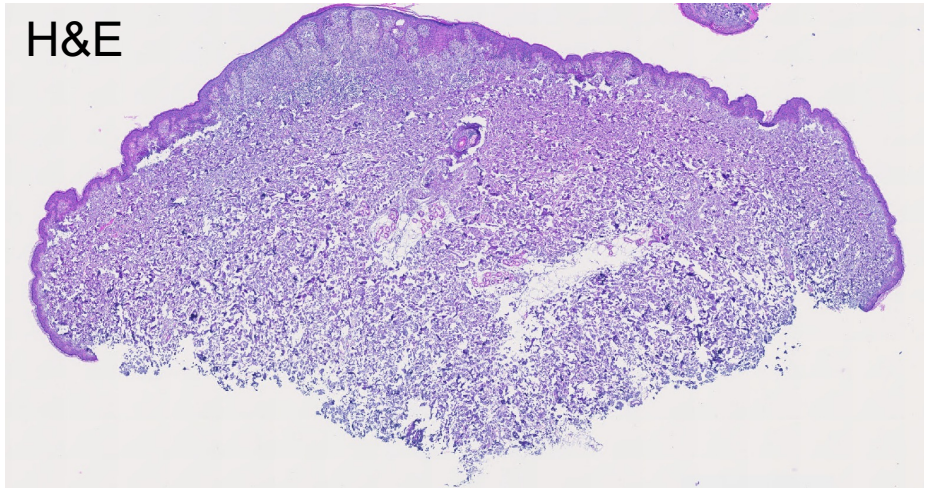
1. Overview of CODEX Data
2. Cell Segmentation
   - QuPath & Stardist
3. Downstream Analysis
   - Marker based cell type annotation
   - Cell type co-localization
   - Niche Analysis
   - Niches geometries & graph

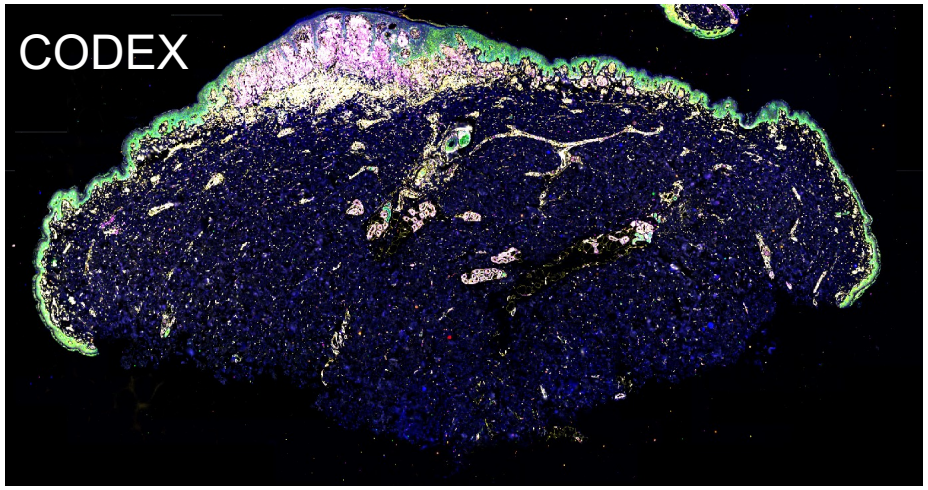
Datasets – Melanoma (Skin)

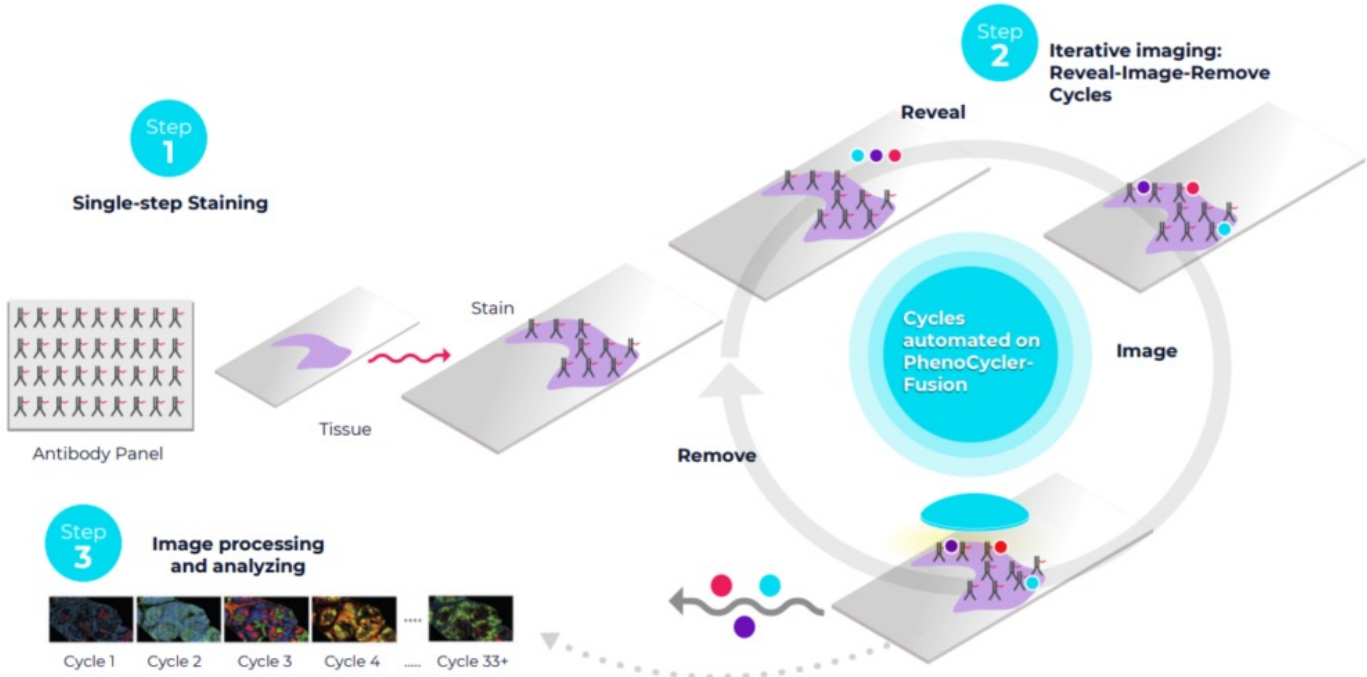




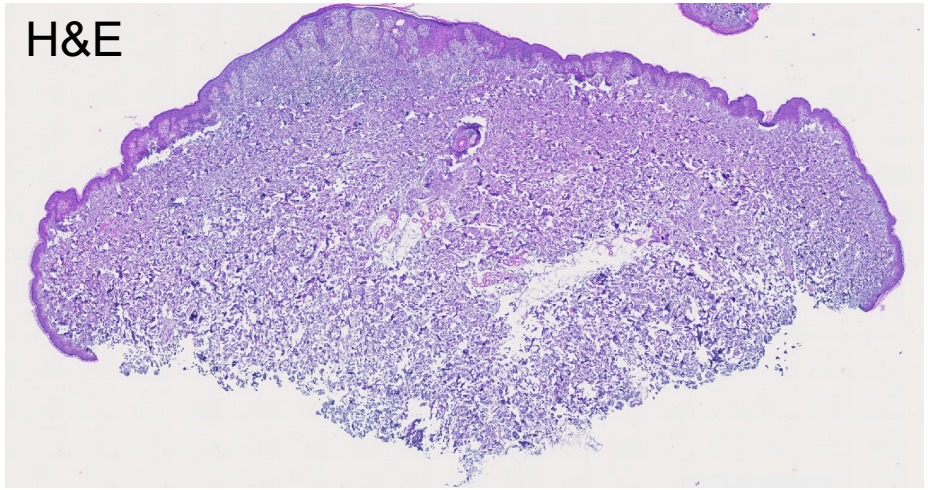
Datasets – Melanoma (Skin)



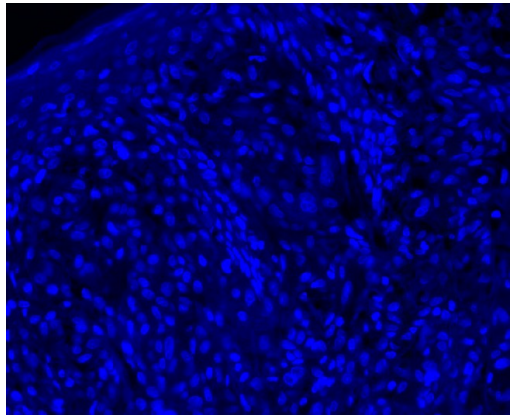




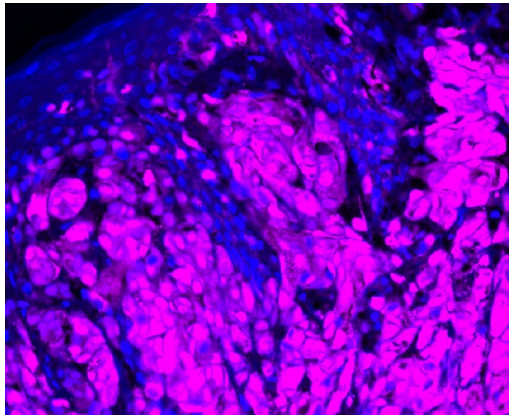
Datasets – Melanoma (Skin)

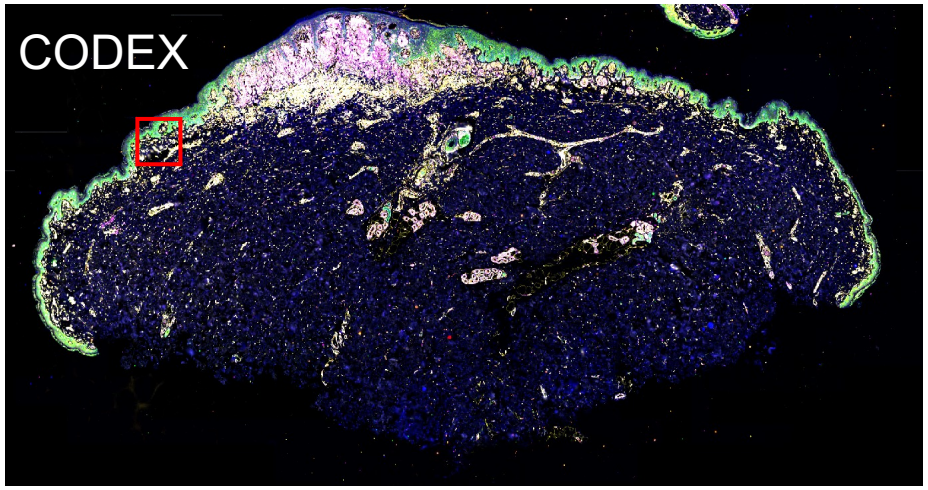


DAPI

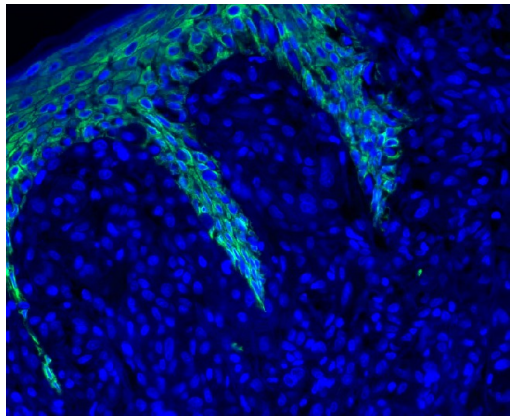


S100B

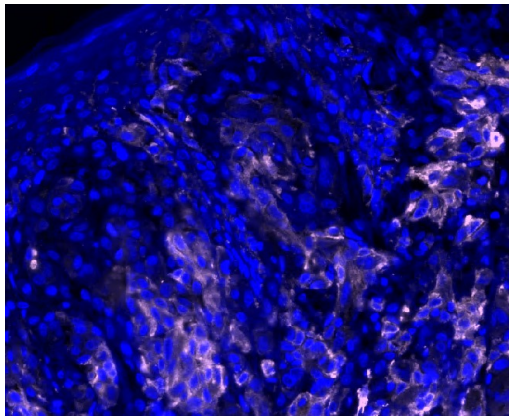




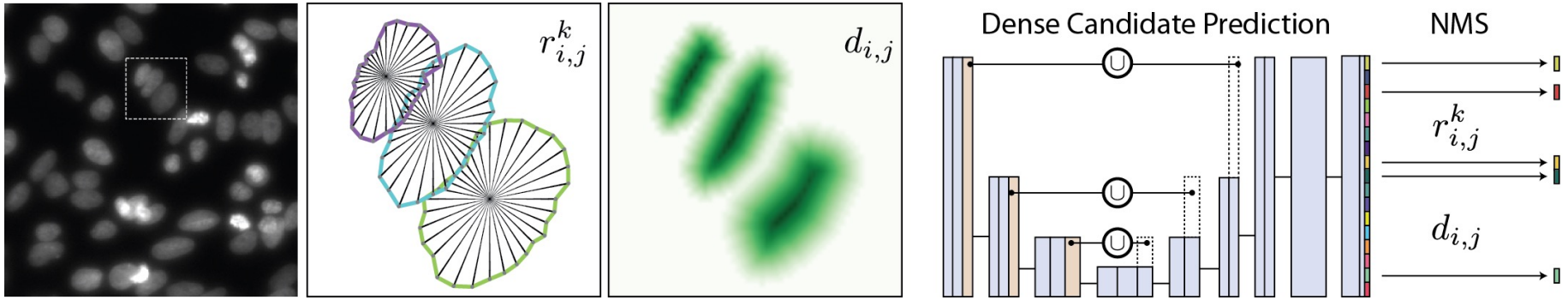
Keratin 14

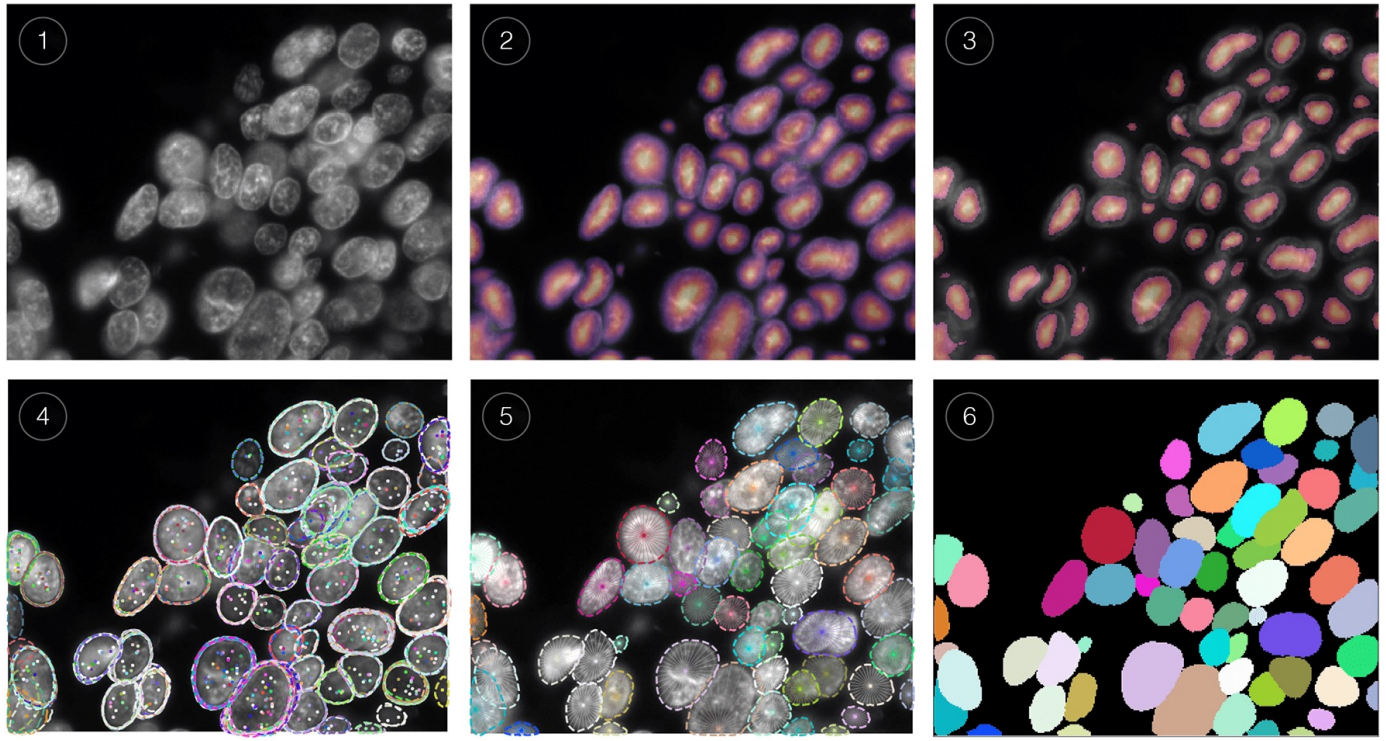


PMEL

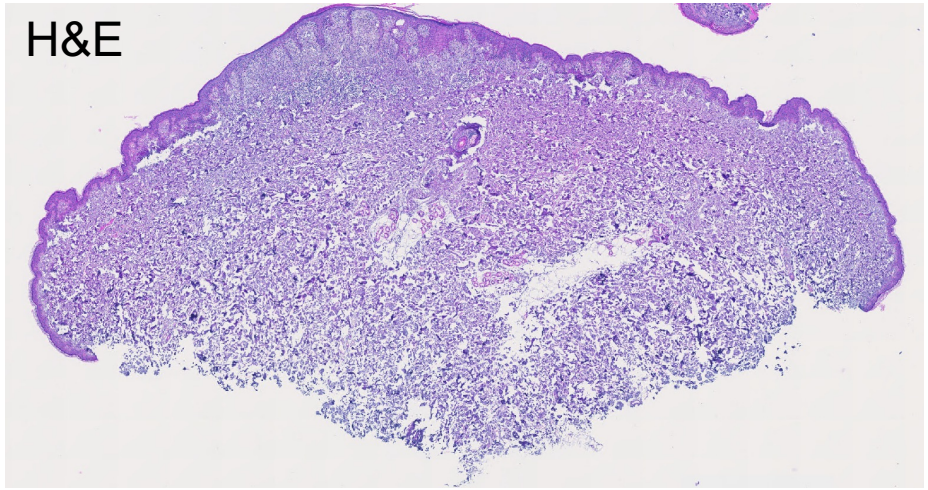


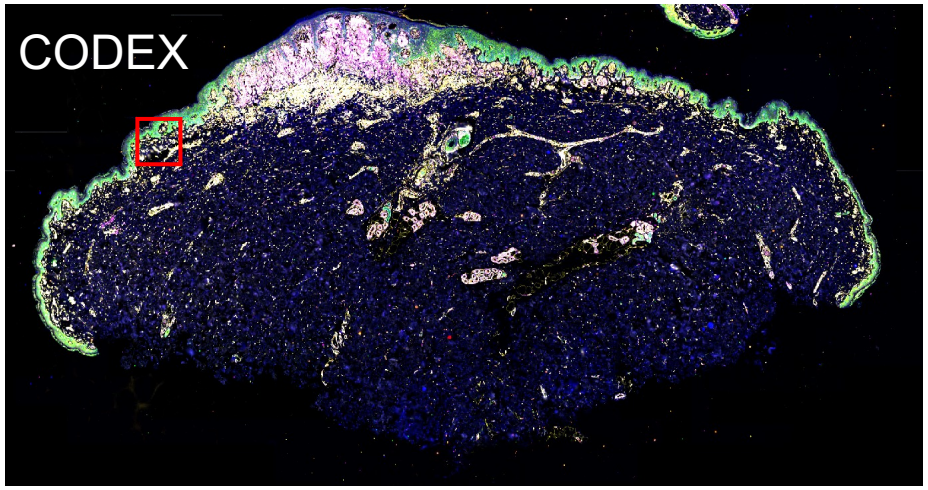
1. *StarDist* Cell Segmentation



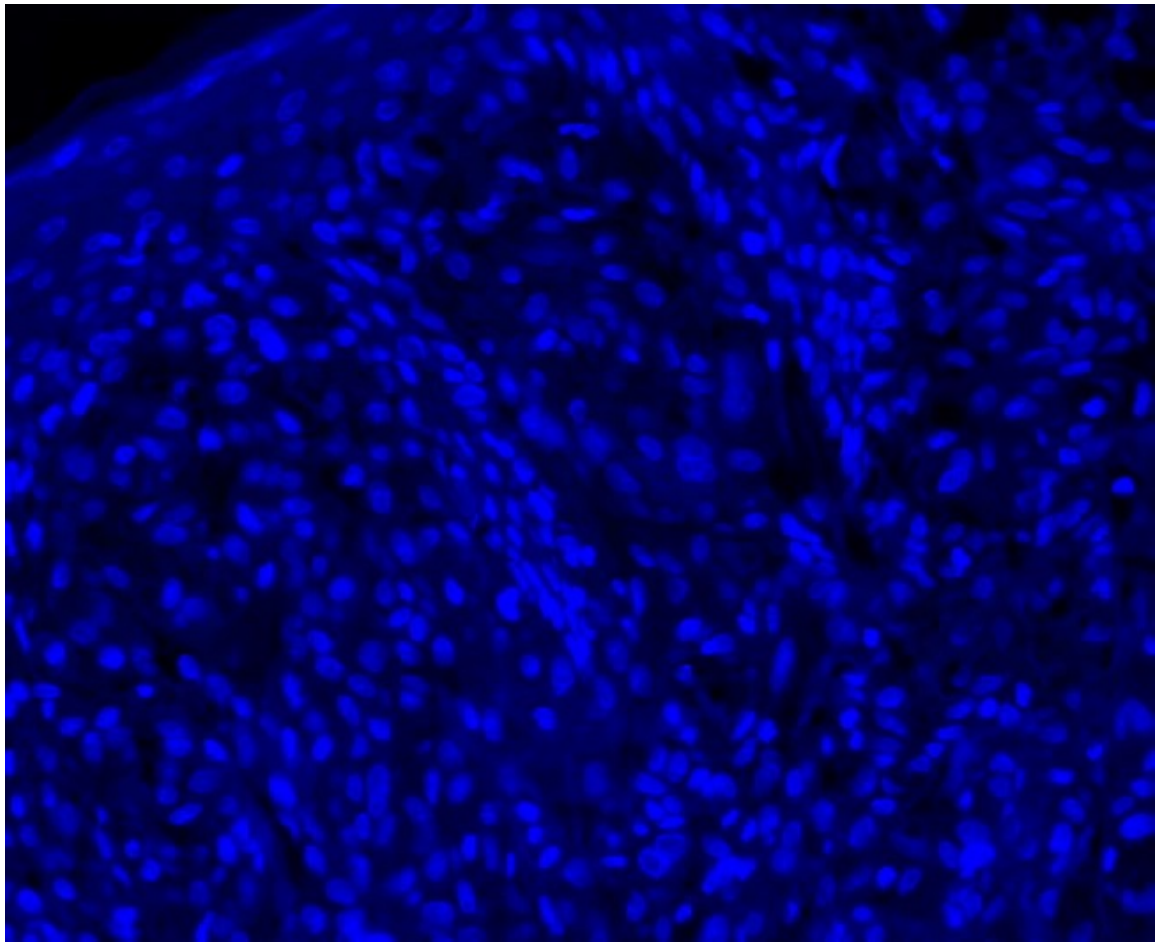


Datasets – Melanoma (Skin)

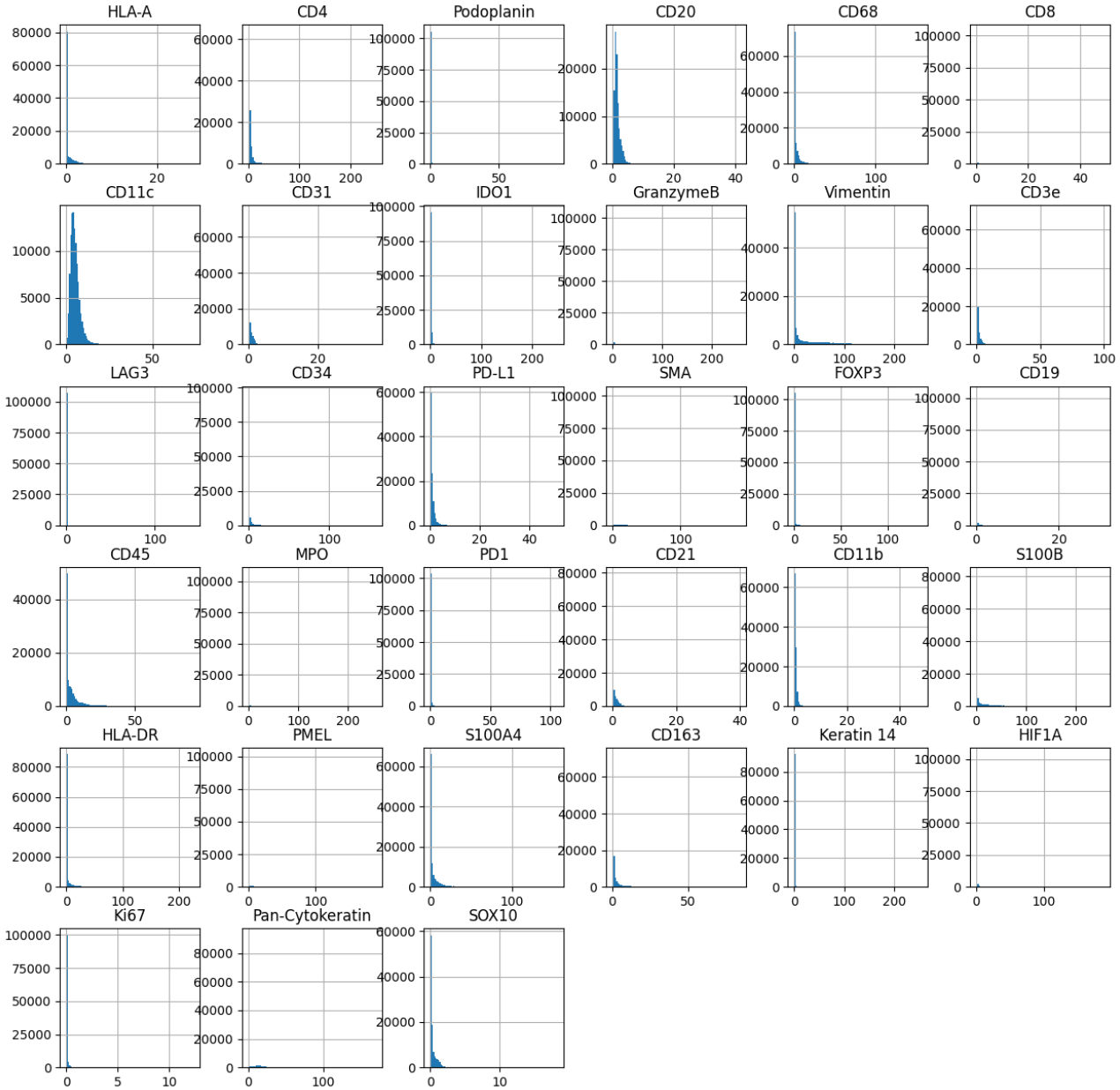




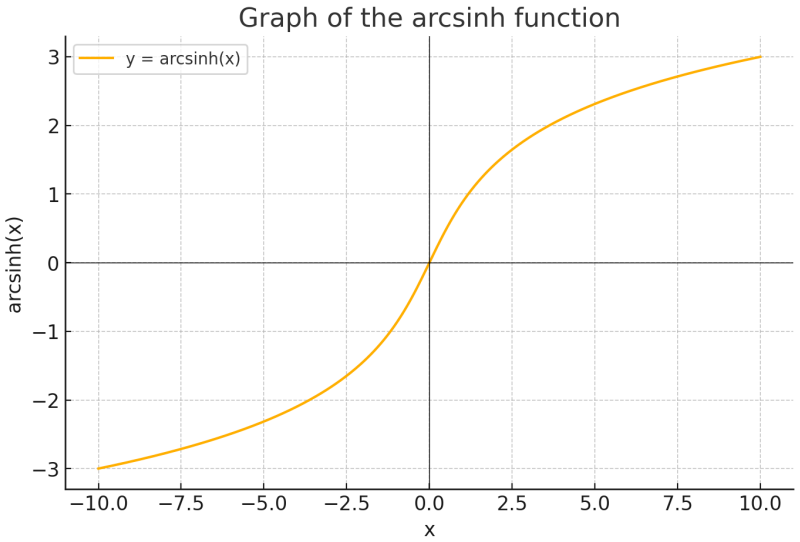
DAPI Cell segmentation



1. Protein Measurements

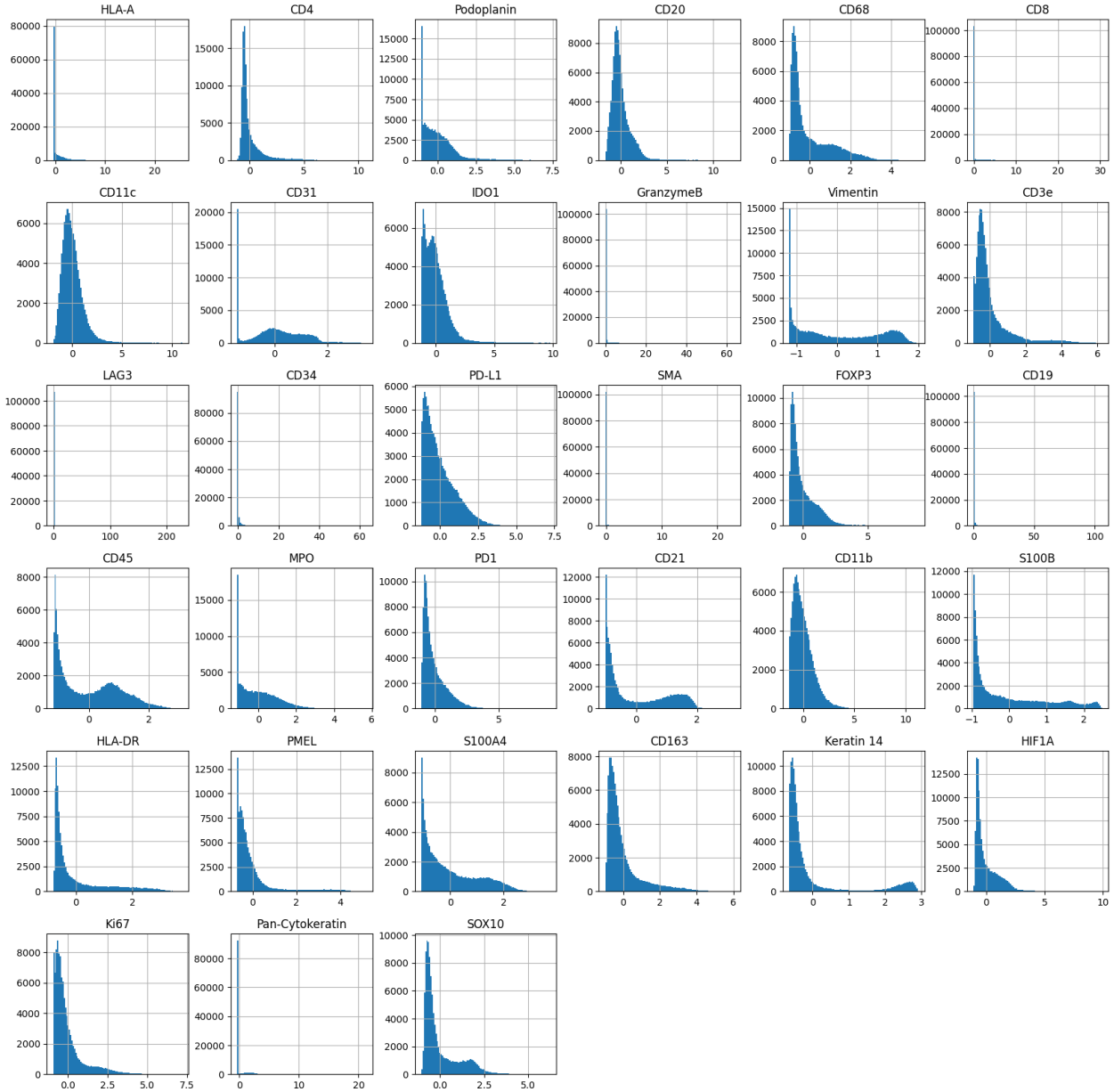


Normalisation: $$\operatorname{asinh}\left(\frac{X}{5Q(0.2, X)}\right)$$



Standardisation: $$z = \frac{(X - \mu)}{\sigma}$$

1. Protein Measurements



2. Cell type annotation

Marker list

…

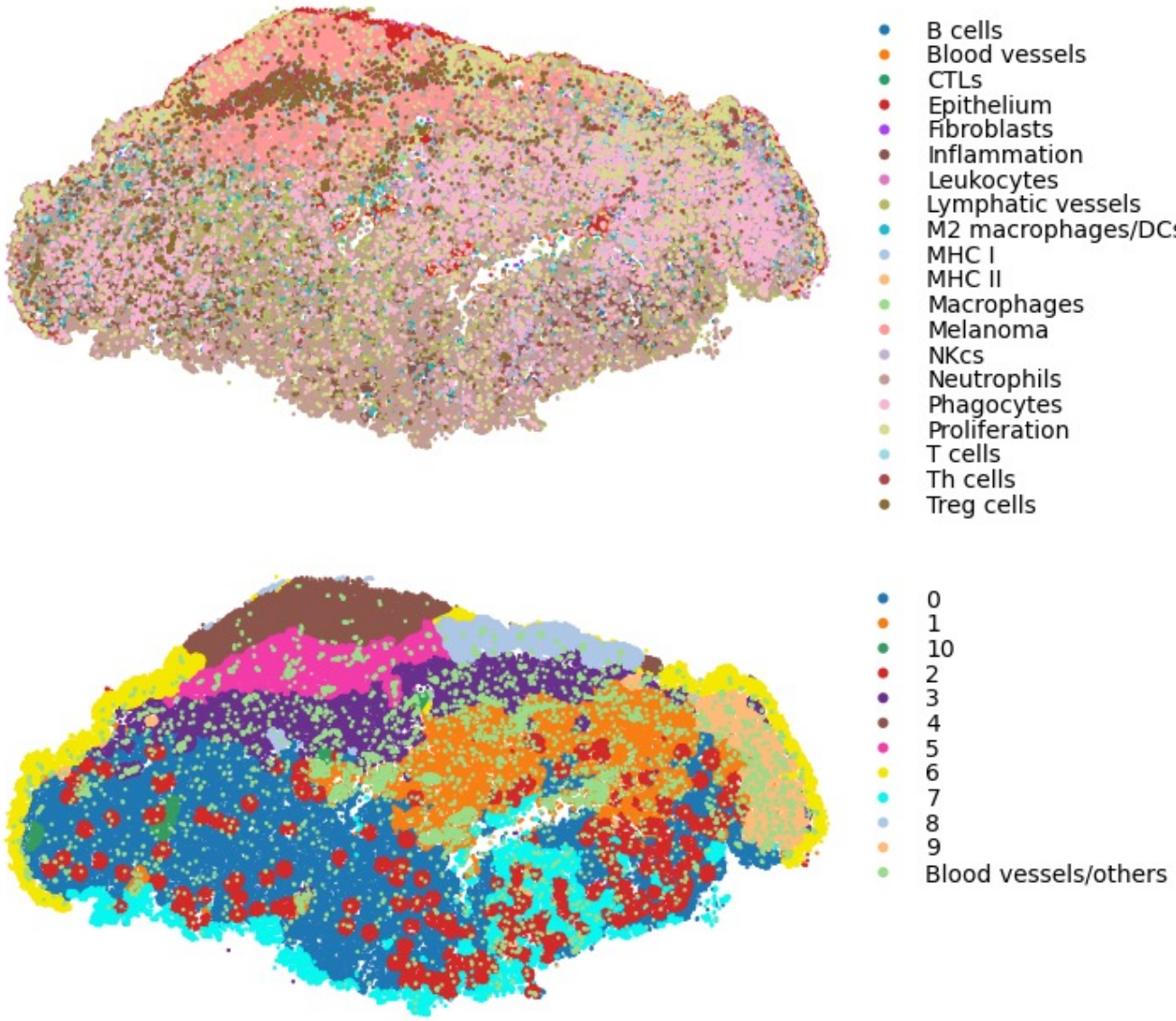
'Keratin 14': 'Epithelium',

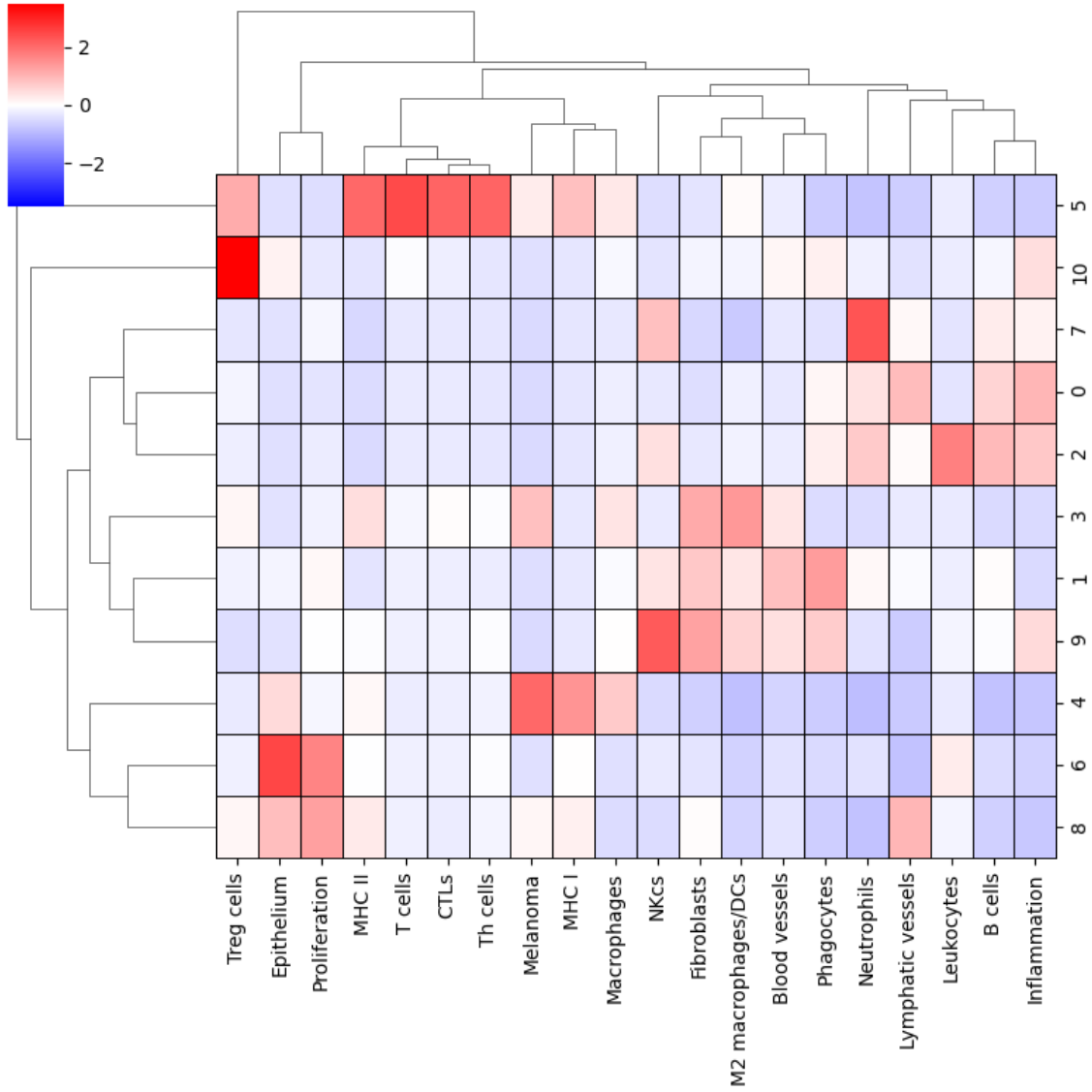
'Ki67': 'Proliferation',

'S100B': 'Melanoma',

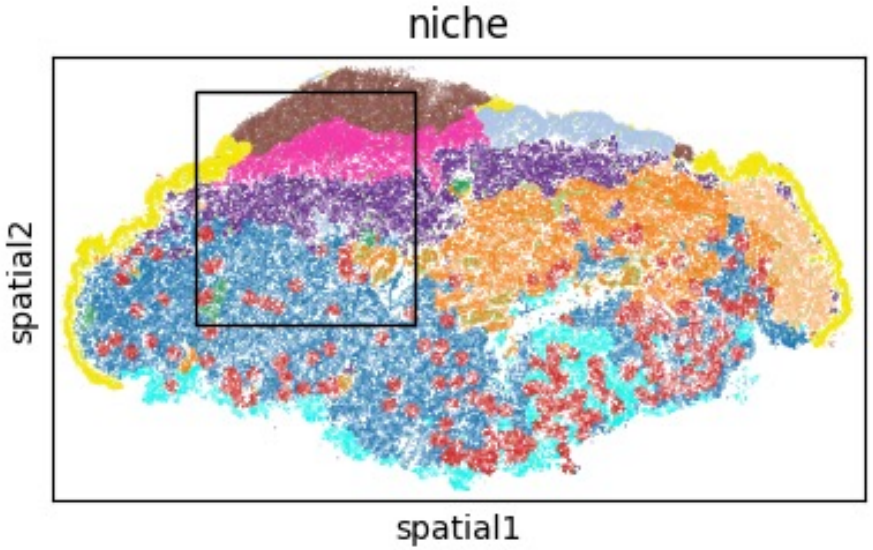
…

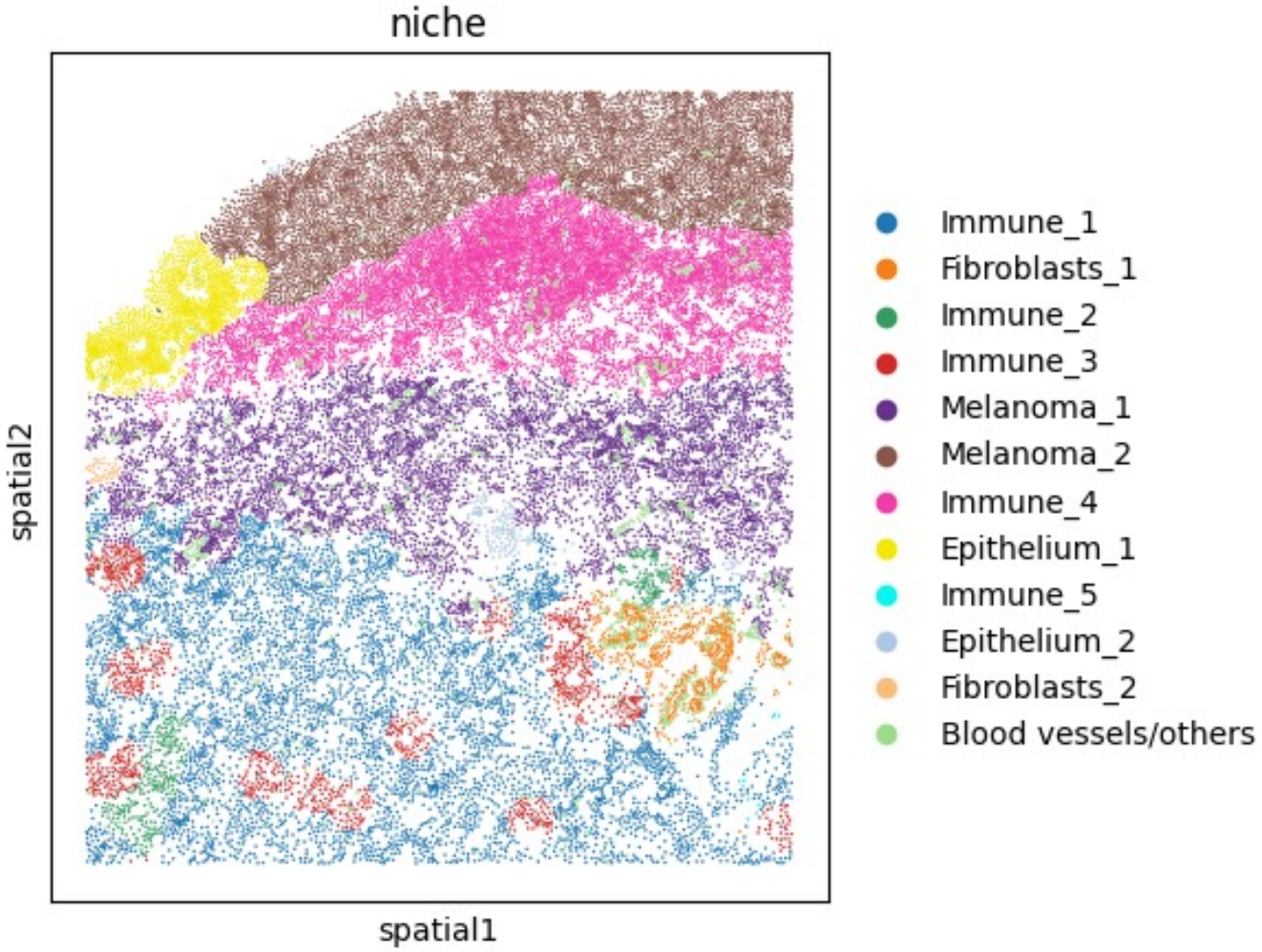
2. Cell type niche



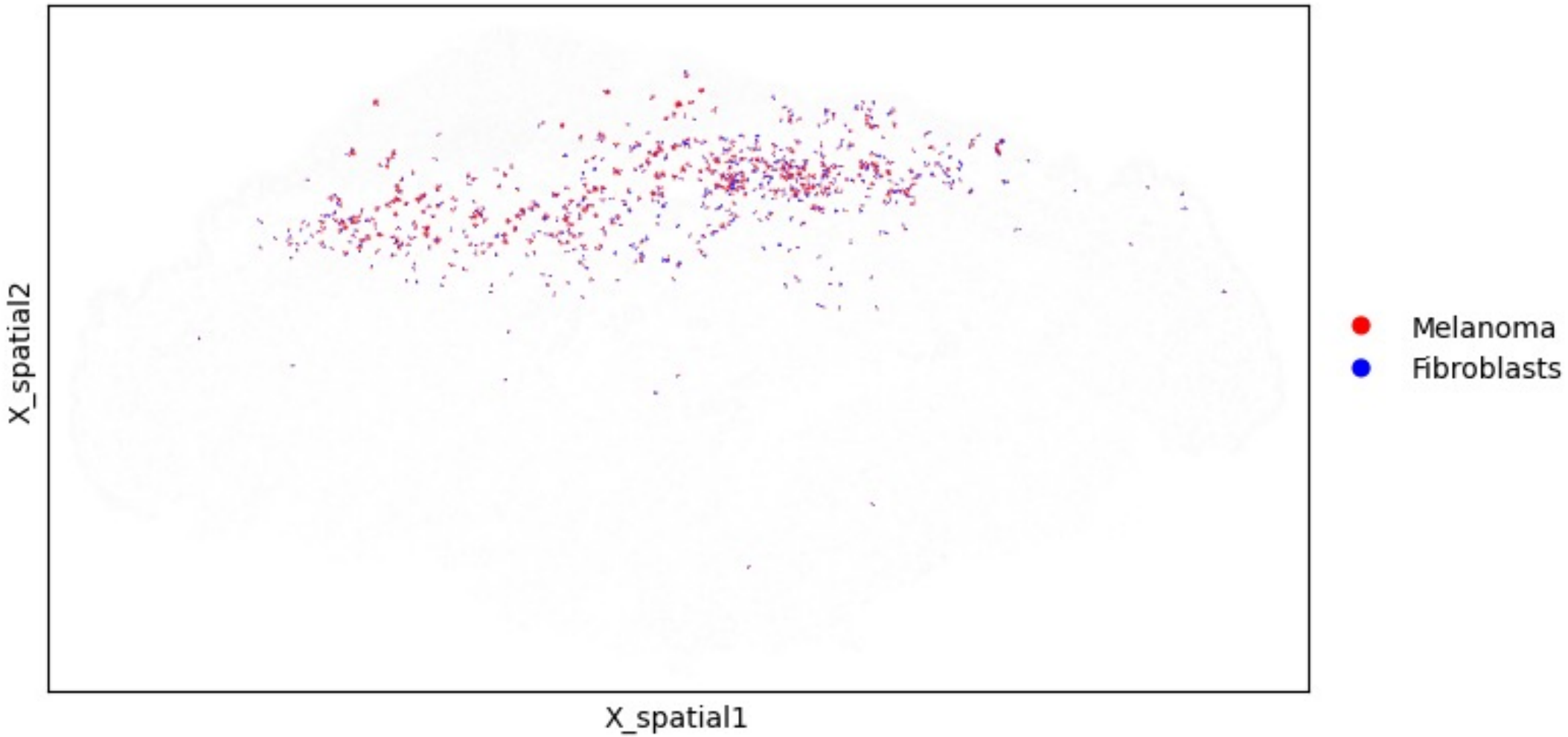


3. Cell type niche components

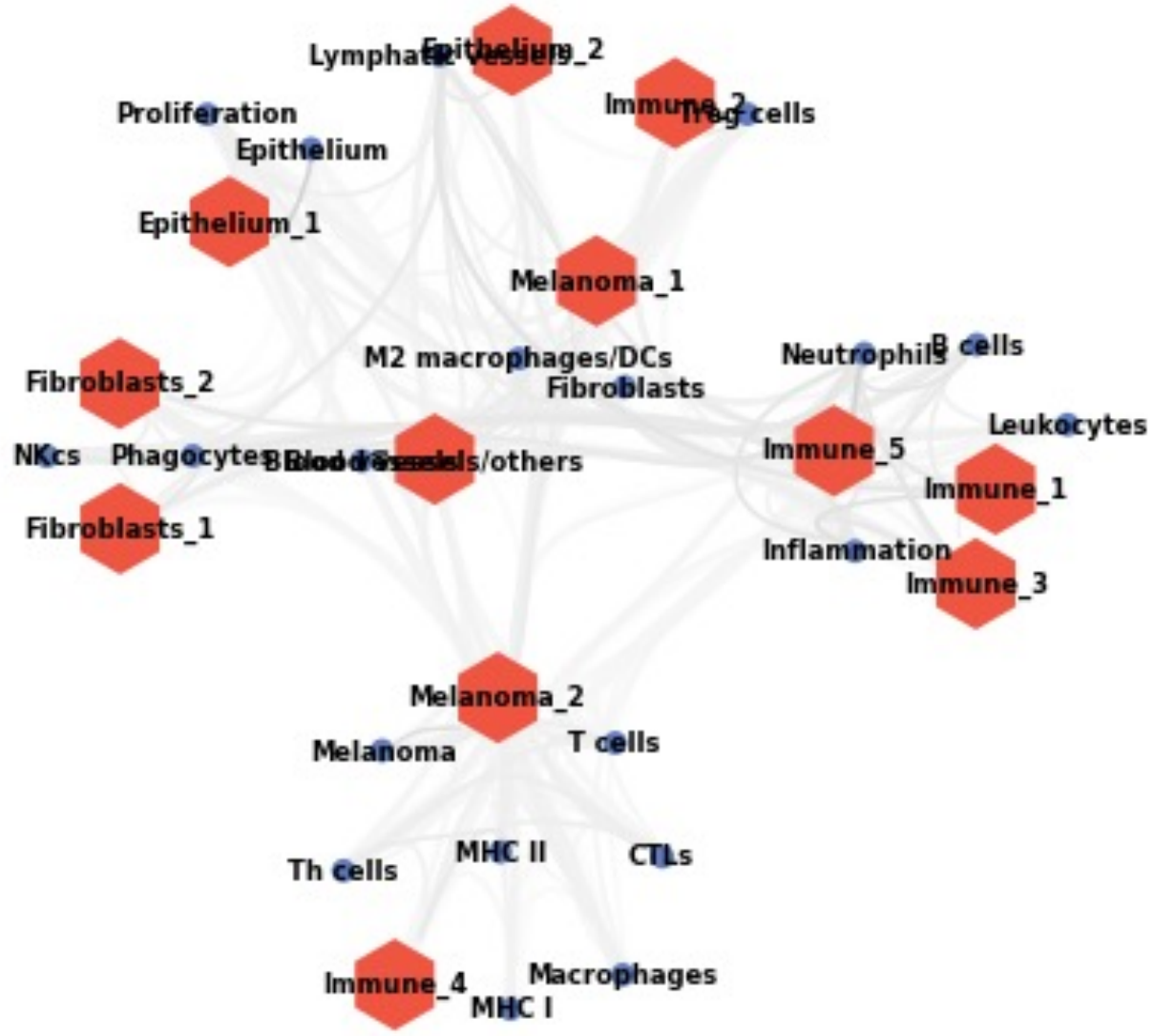




3. Co-localization between cell types

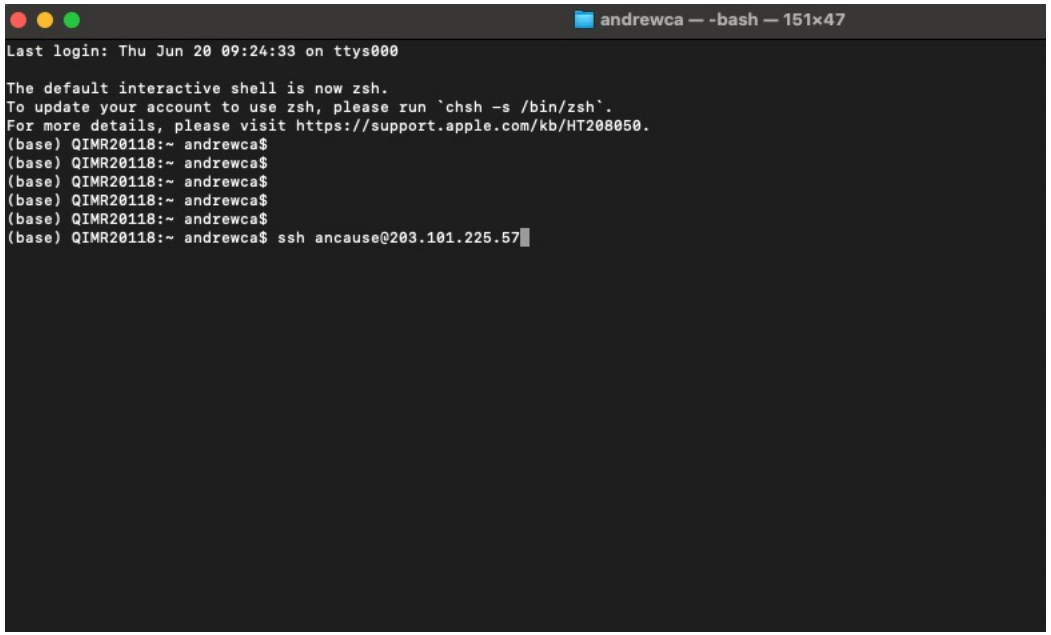


3. Cell-type / Niche network



Running the Practical

Terminal PowerShell



1. Log into your account:

   ssh {username}@203.101.225.57
   username & password from winter school email

2. Follow these commands:

   - /software/bin/micromamba shell init
   - source ~/.bashrc
   - micromamba activate /software/conda-envs/winter_school_2024
   - git clone https://github.com/GenomicsMachineLearning/qimr-teaching-2024 /scratch/$USER/qimr-teaching-2024
   - /scratch/$USER/qimr-teaching-2024/runme.sh

3. Open JuperterNotebook:

   /software/005-spatial-proteomics/CODEX_analysis.ipynb

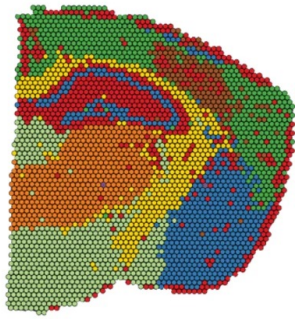
Lecture 9: Machine Learning

Module 2 – Part 9: Machine Learning

Xiao Tan & Quan Nguyen

TensorFlow PyTorch

Python packages



10X Visium

Module 2 – Part 9: Deep Learning

Deep learning for spatial transcriptomics data

1. Basic Concept

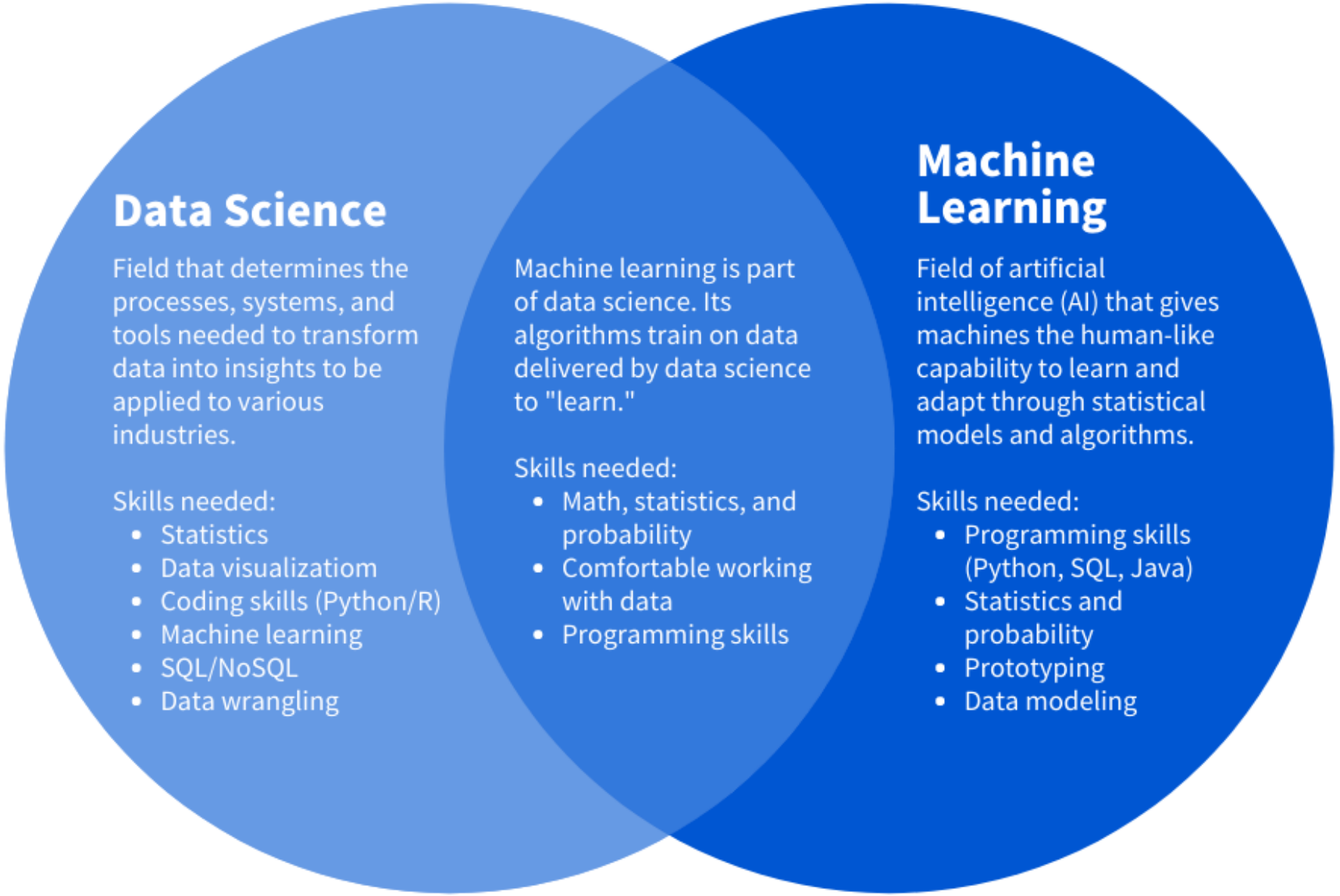
2. Neural Network model (NN)

- Key elements
- Model architectures

3. NN in spatial omics

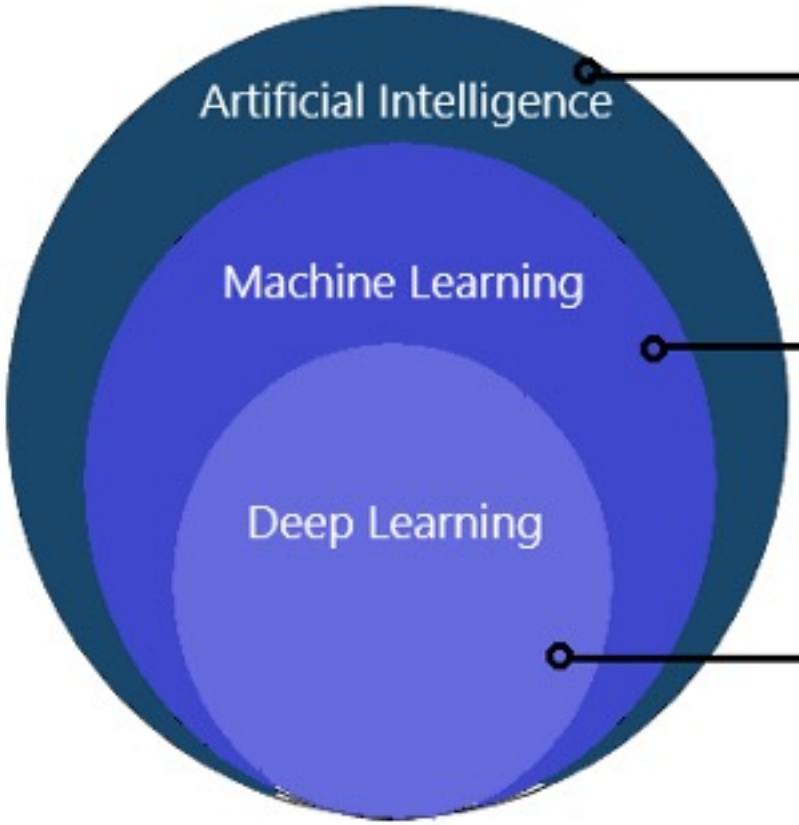
- Convolutional neural network (CNN)
- Autoencoder

Data science and machine learning



(Coursera, 2022)

Machine learning, statistical learning, deep learning



Artificial Intelligence

AI enables machines to think without any human intervention.

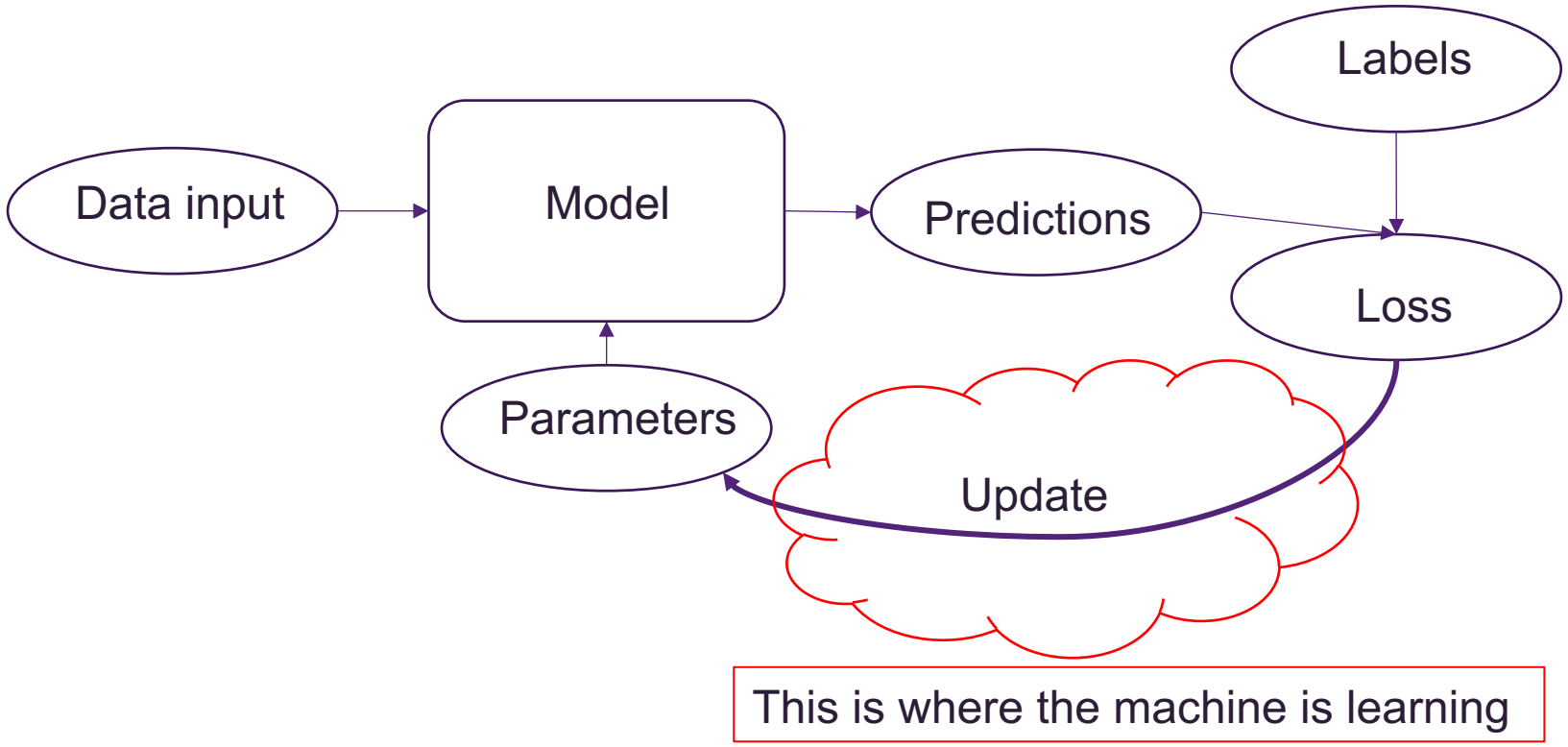
Machine Learning

Subset of AI that uses statistical learning algorithms that learn pattern in data over time

Deep Learning

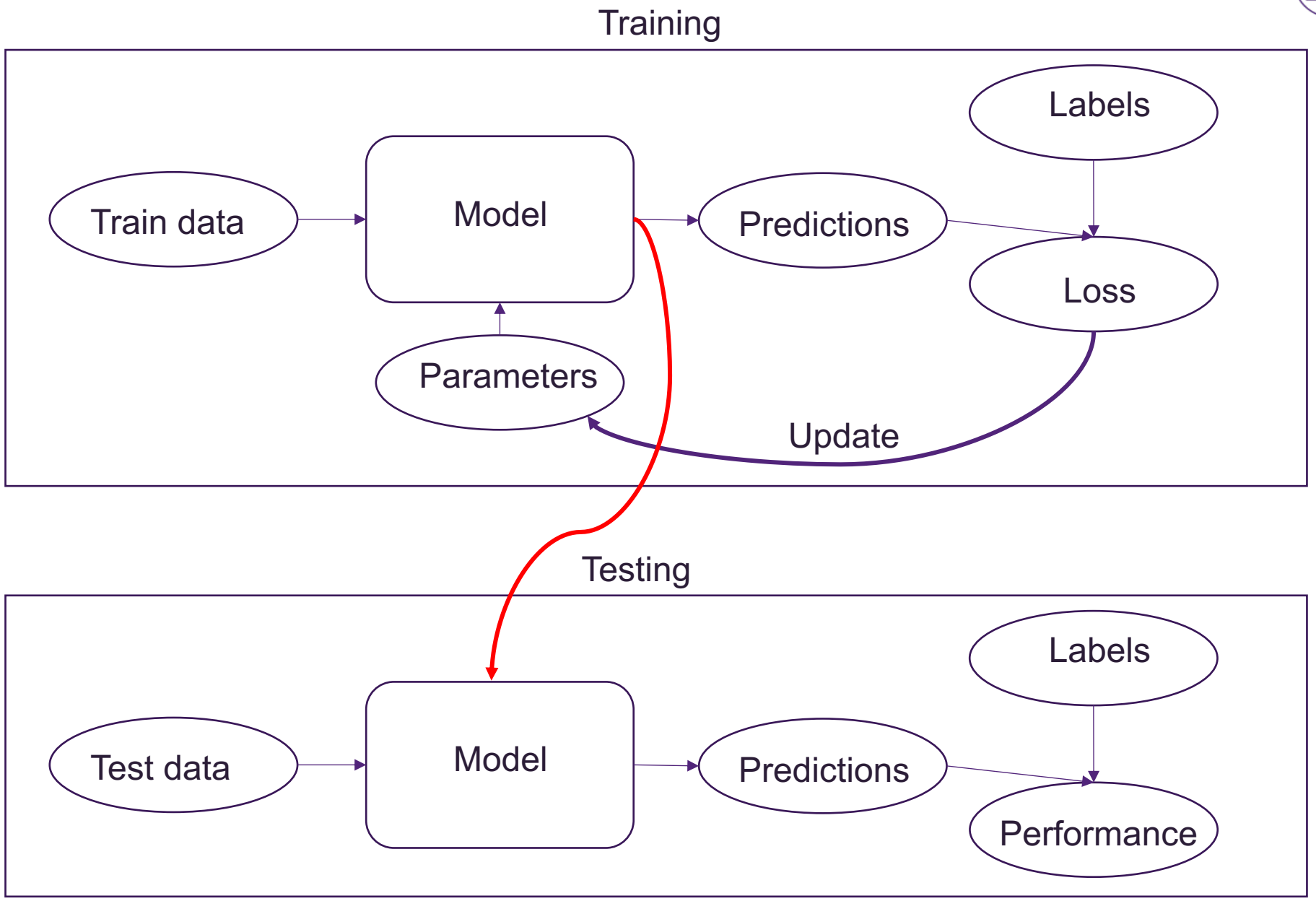
Subset of ML that filters the data through multiple layers

Machine learning – Loss function

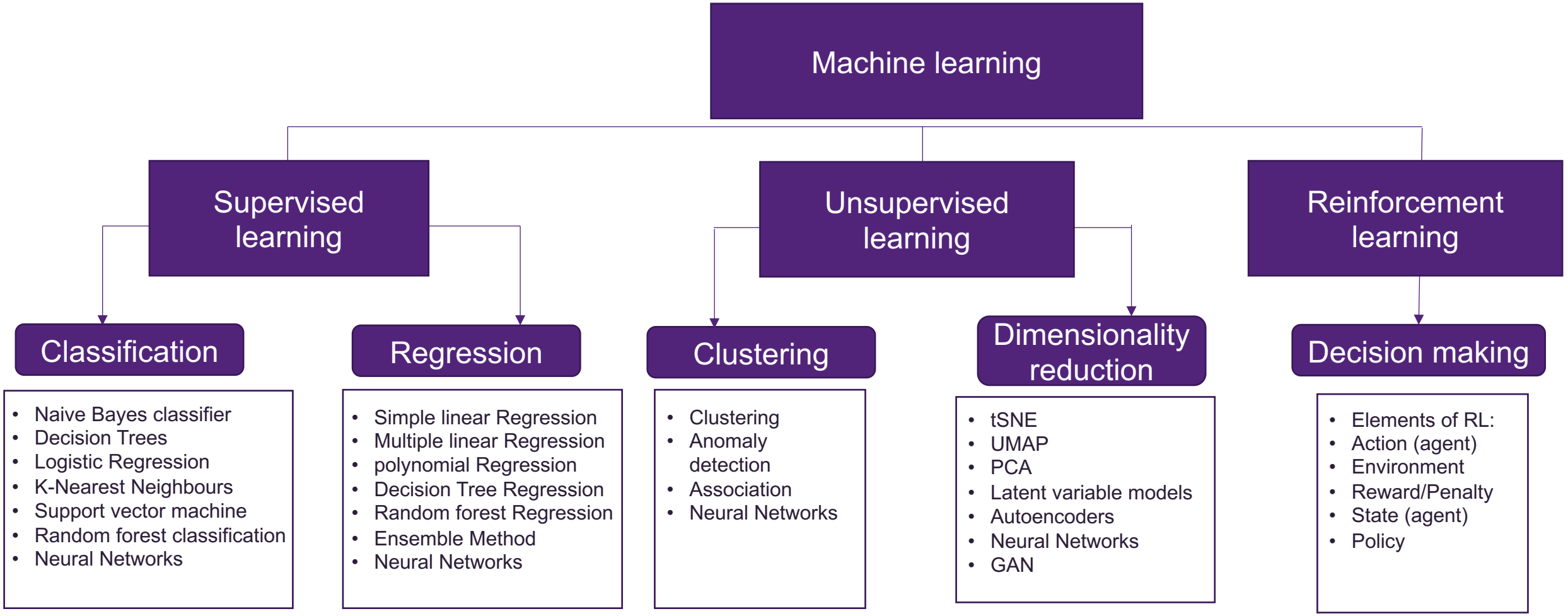


- ML: The training of programs developed by allowing a computer to learn from its experience (rather than through manually coding the individual steps)
- Loss function is where ML meets statistical models
- (hyper)Parameters are where machine learning deviate from statistical models

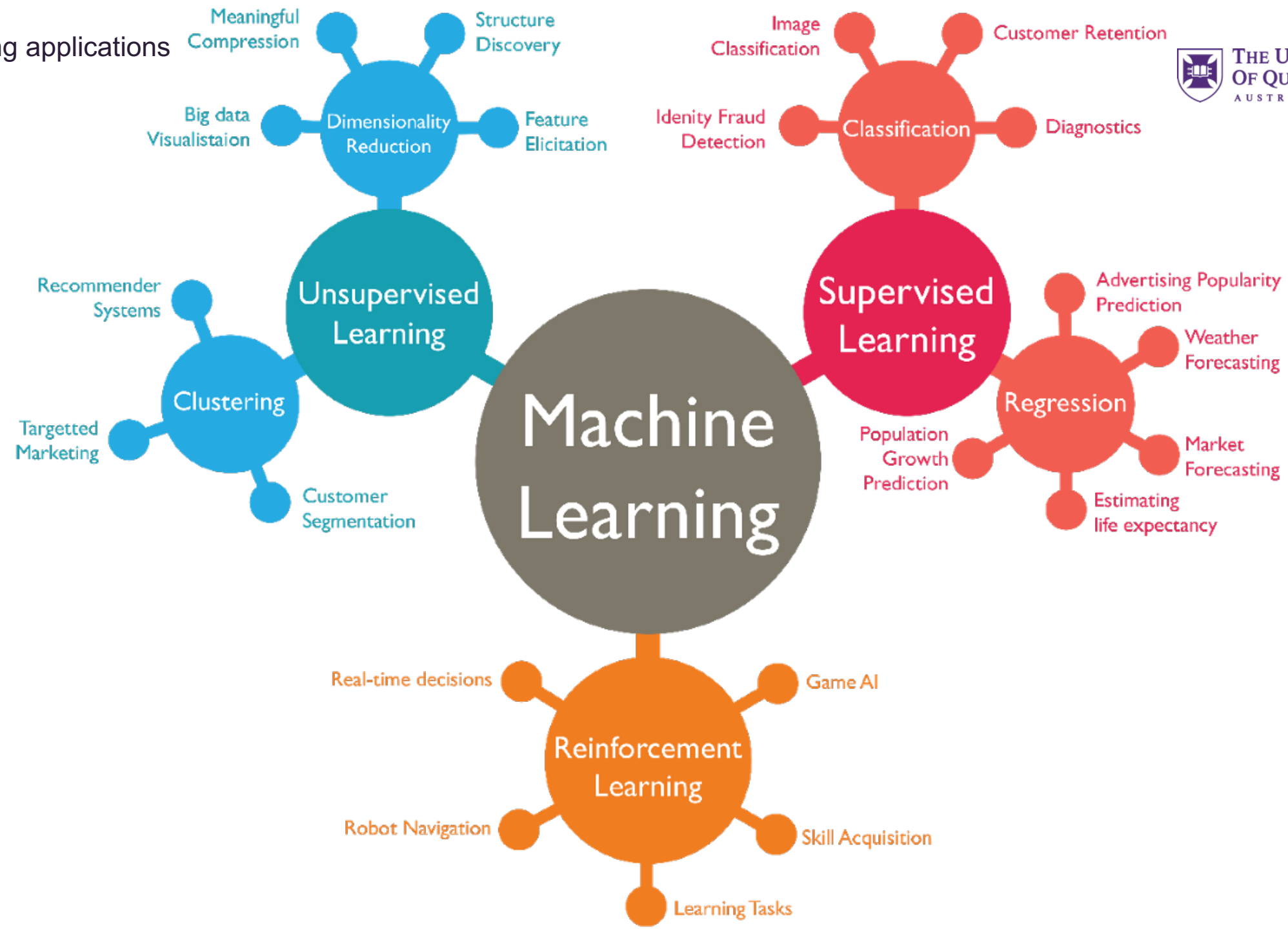
Machine learning – Training and testing datasets



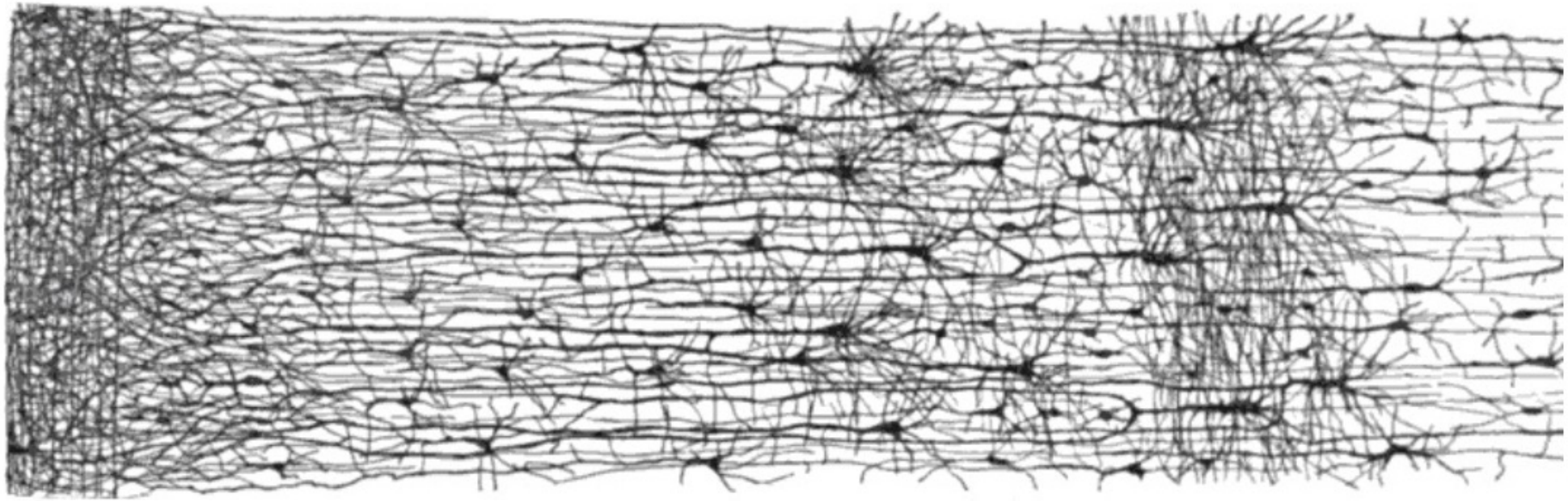
Machine learning categories and tasks

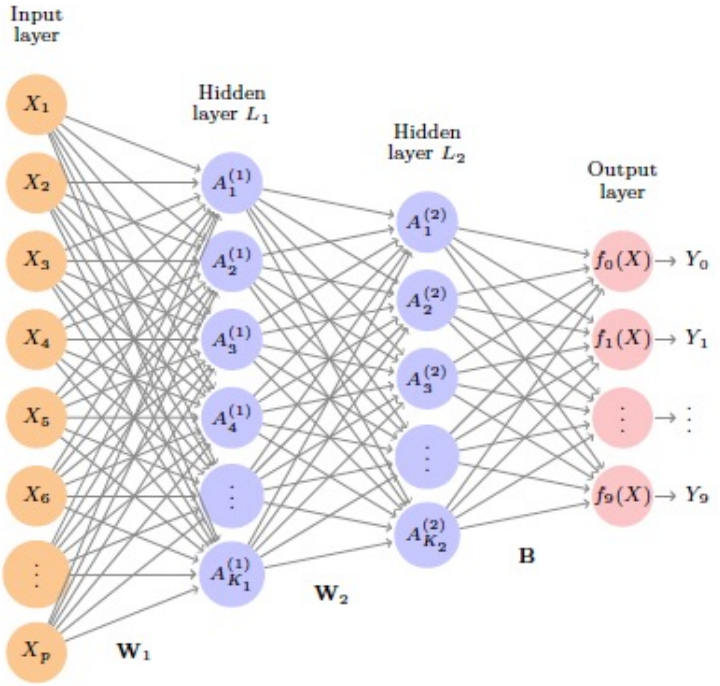


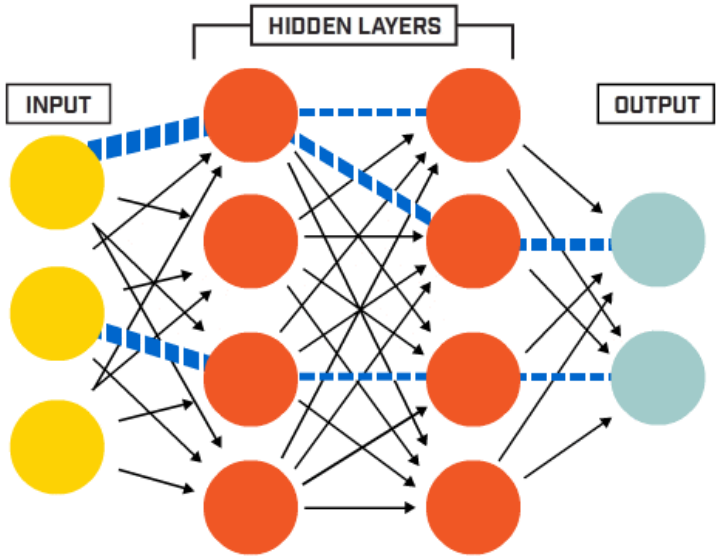
Machine learning applications



Neural network and multilayer perceptrons

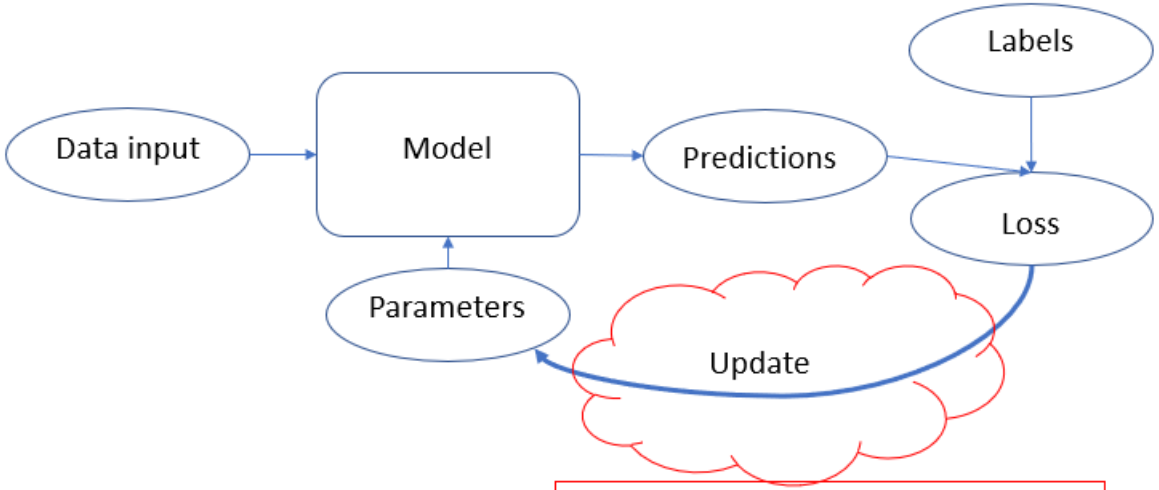


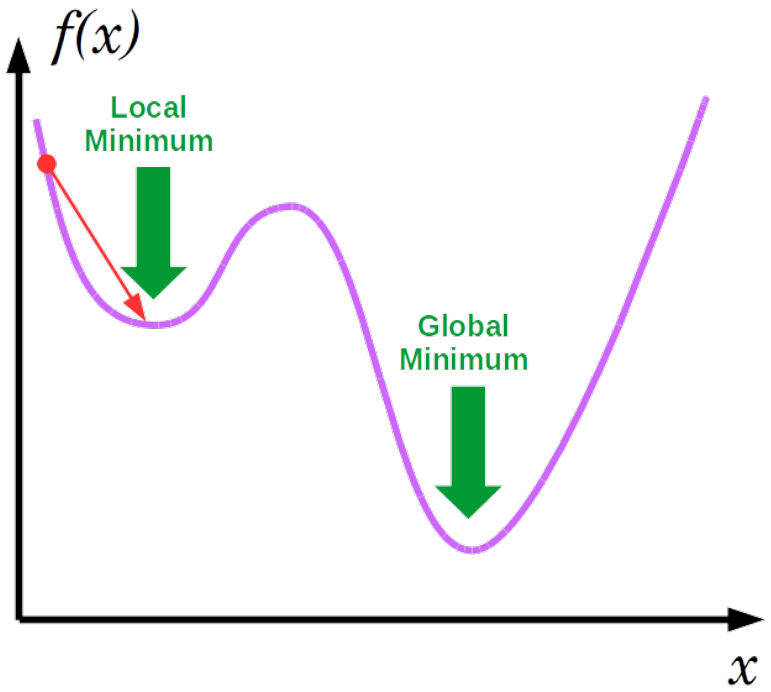


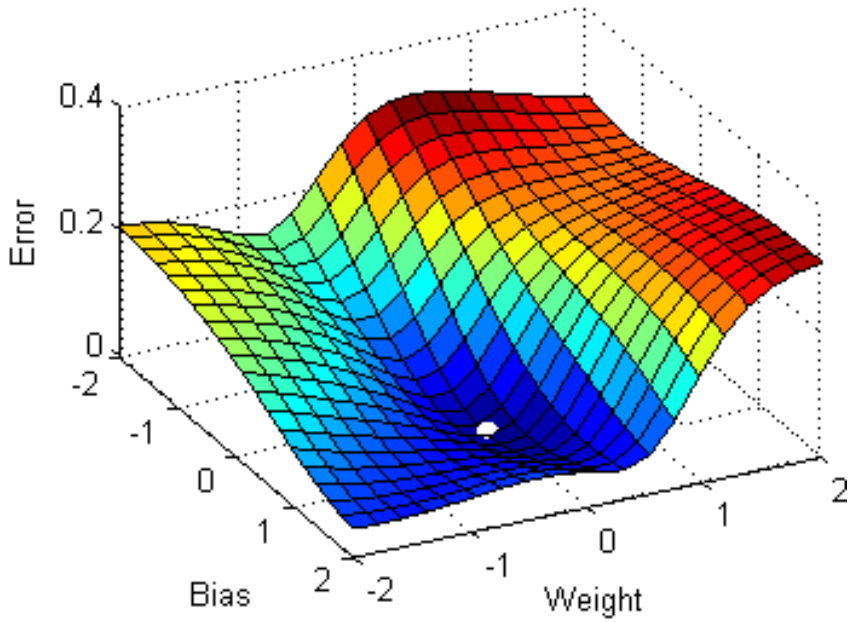


(Towards AI, 2019)

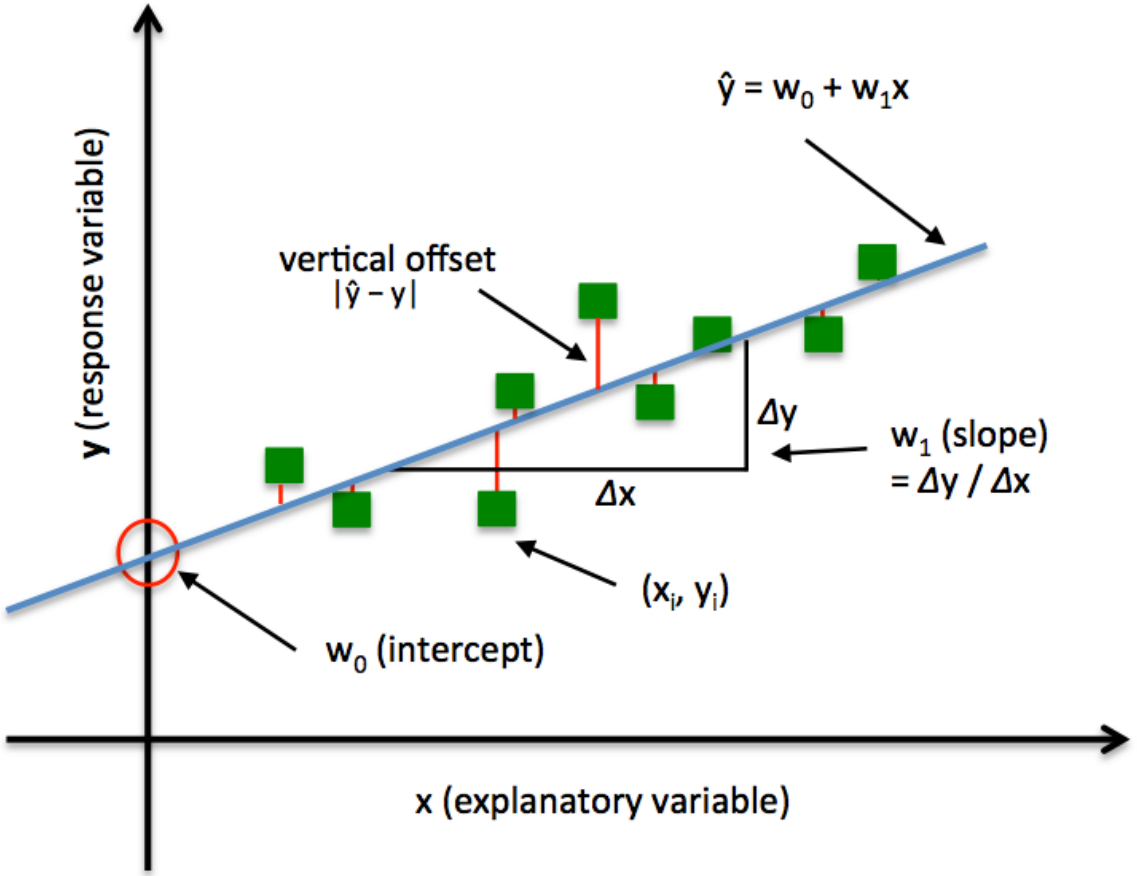
Machine learning – Loss function







From linear regression to machine learning



$$Error = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

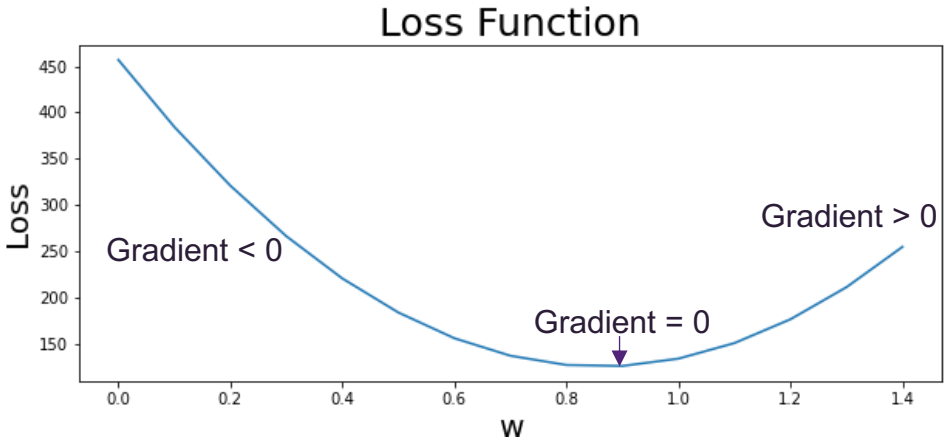
$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - w_0 - w_1 X_i)^2$$

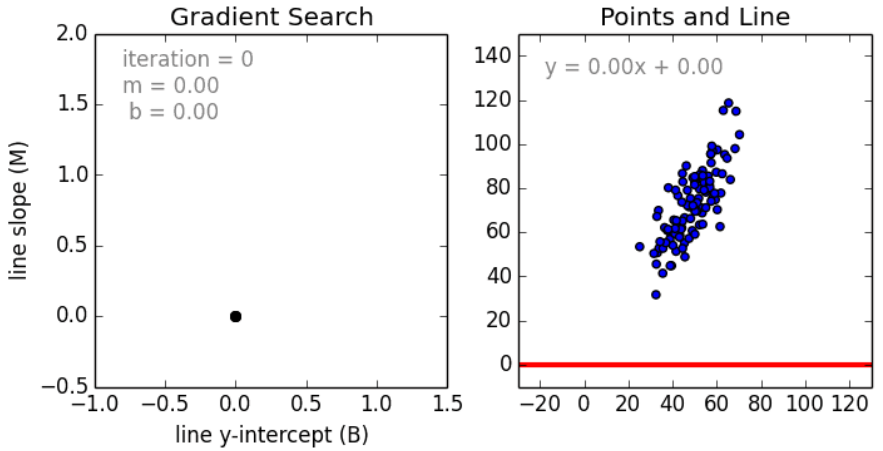
= Objective function
= Loss function
= J(w_0, w_1)

Minimize wrt w_0 and w_1 by gradient descent

http://rasbt.github.io/mlxtend/user_guide/regressor/LinearRegression/

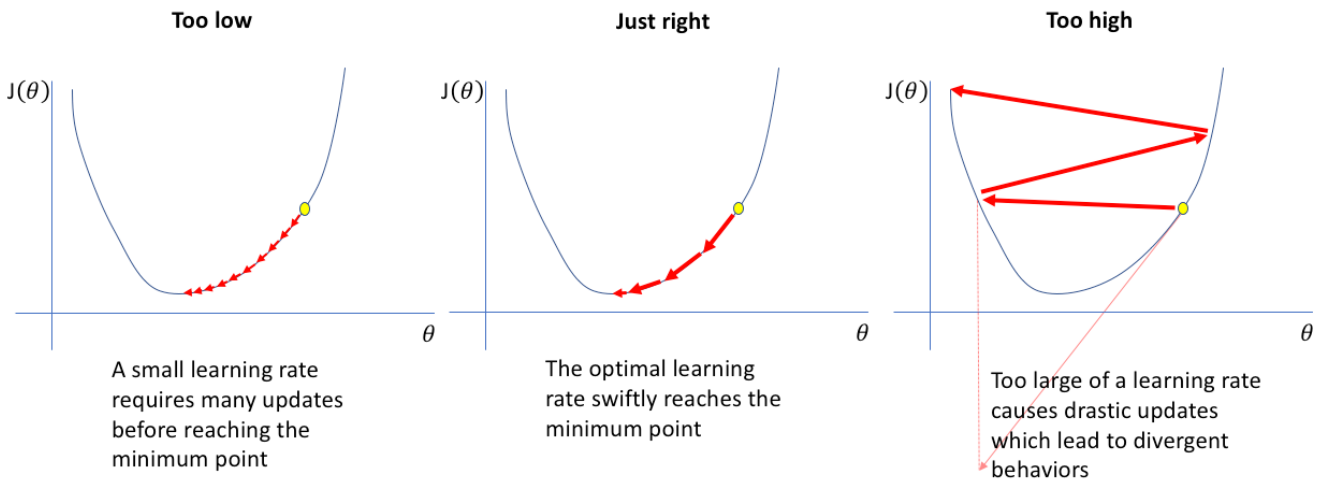
General terms: Gradient Descent Example for Linear Regression





$$w^t = w^{t-1} - \alpha \nabla J(w^{t-1})$$

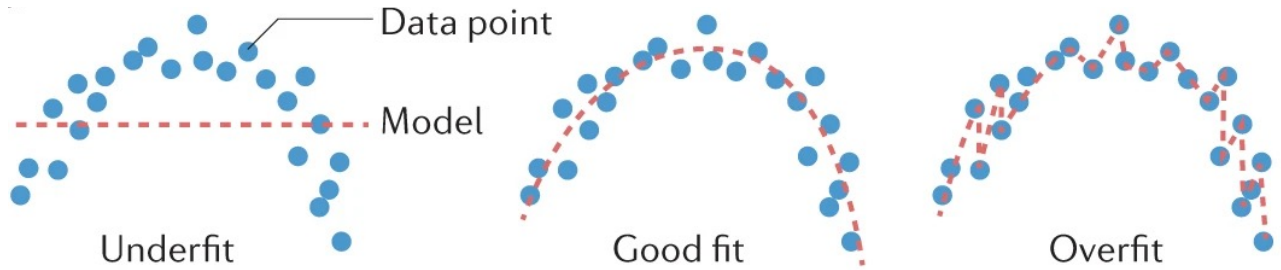
α is the learning rate (step length)
Effect of learning rate →

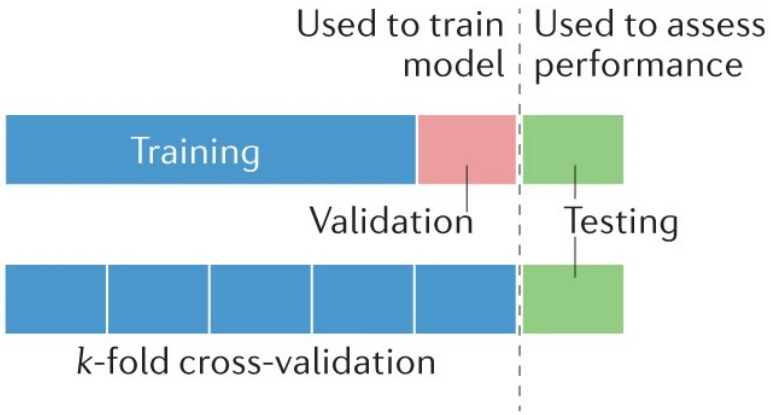


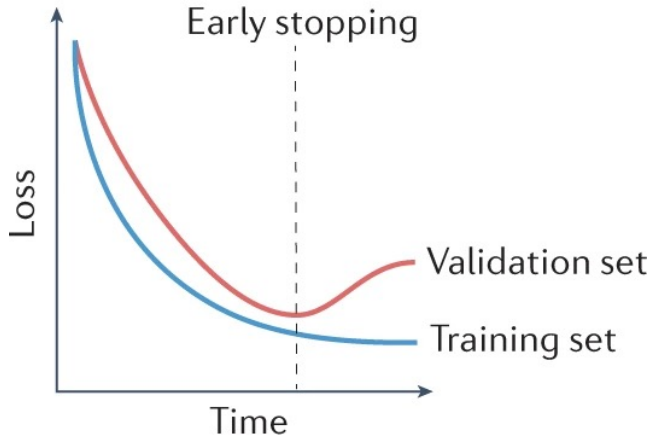
https://www.jeremyjordan.me/nn-learning-rate/

https://github.com/mattnedrich/GradientDescentExample

General terms : Overfitting and how to reduce

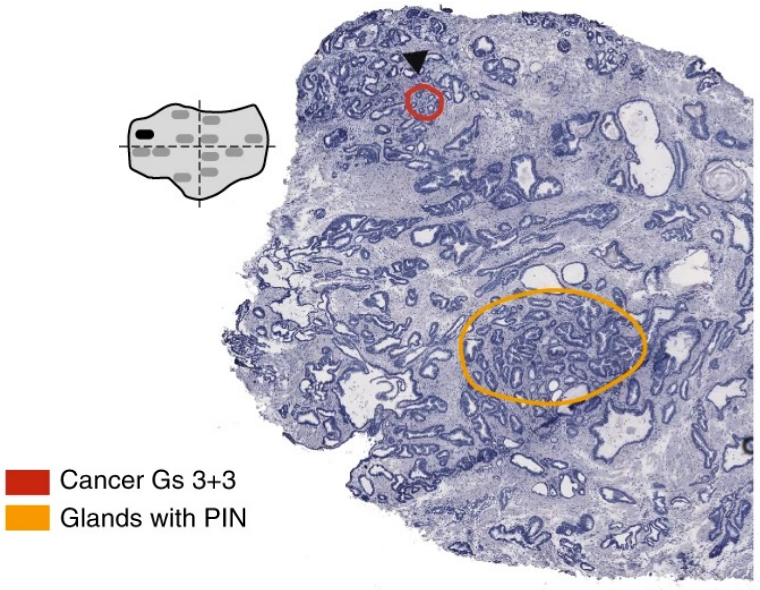




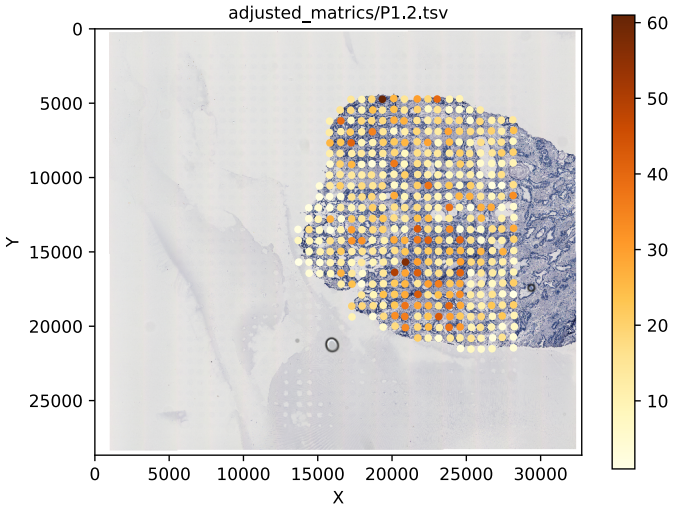


Spatial Transcriptomics: expression + image + location

	14.96x10.06	15.92x10.05	17x10.06	17.89x10.06
STARD7 ENSG00000084090	0	0	1	0
WDR1 ENSG00000071127	1	1	1	1
NDUFB2 ENSG00000090266	2	4	2	1
BAIAP2L1 ENSG00000006453	2	1	8	1



(Berglund et al, 2018)



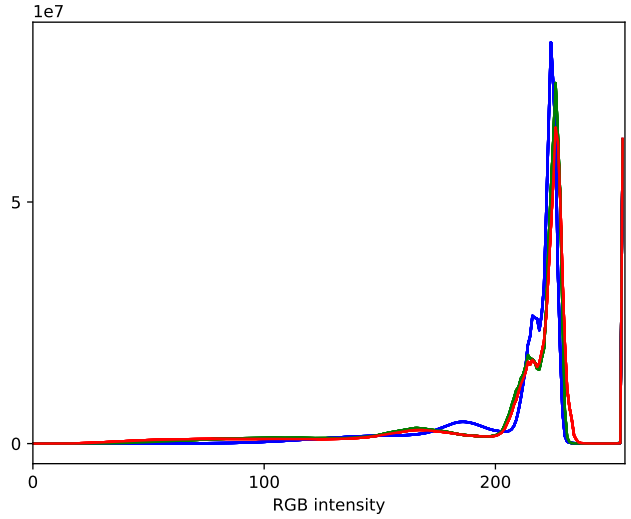
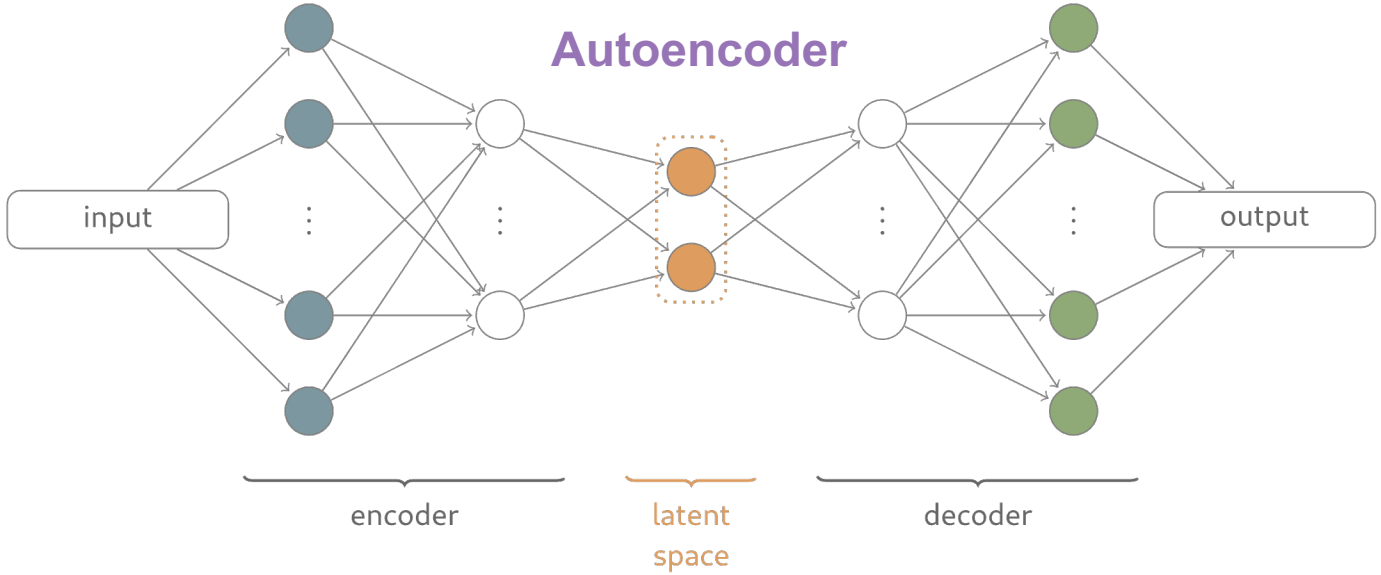


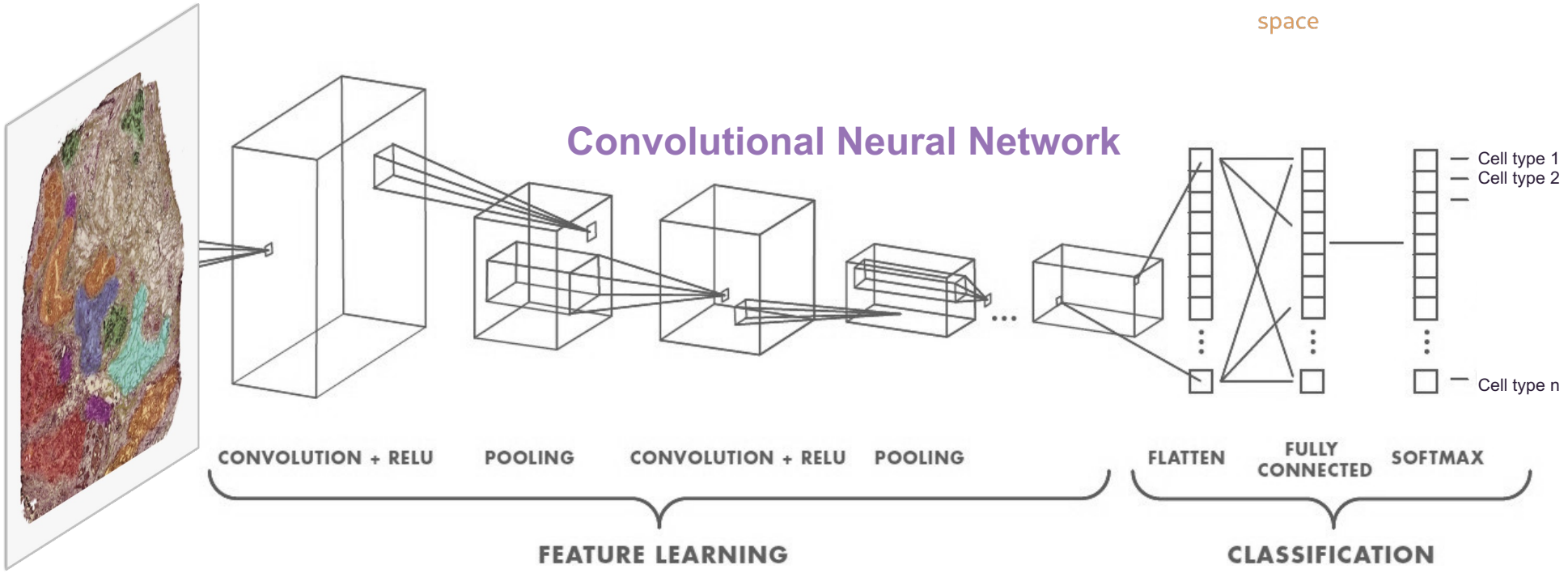
Image mode=RGB, size=32768x28672, (28672, 32768, 3)

Neural Network for Spatial Transcriptomics

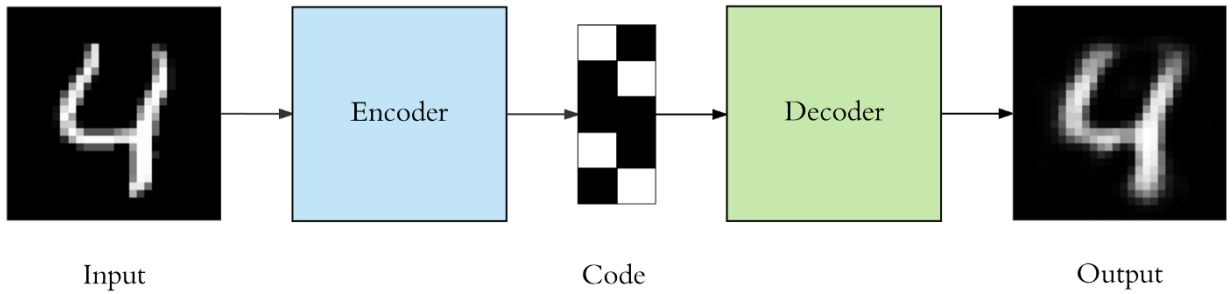
Two neural network (NN) architectures

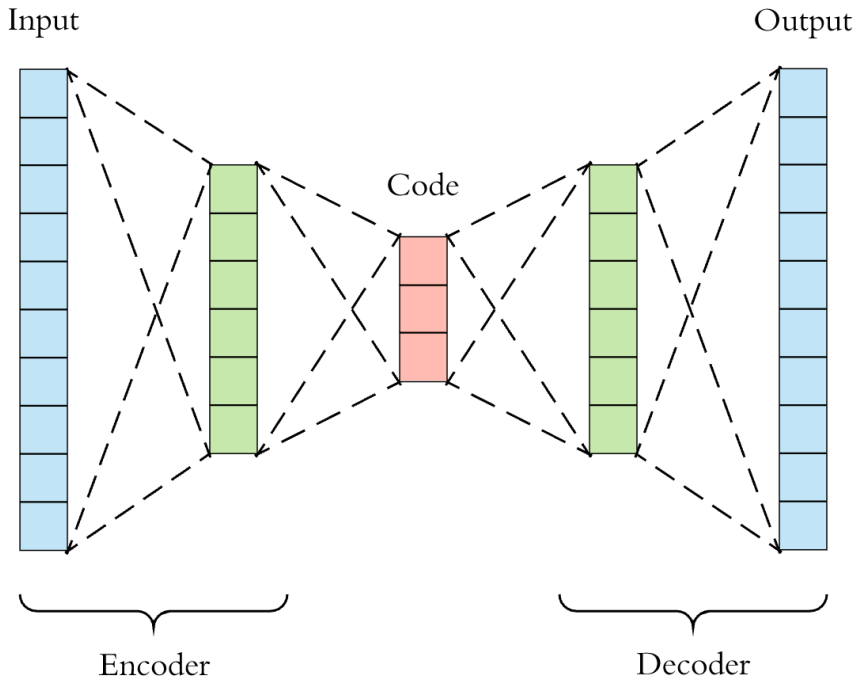
- Convolutional Neural Network (CNN) for feature extraction
  - Designed for spatial imaging data
- Autoencoder (AE) for combining data
  - Find informative shared latent space





Autoencoder





Input is the same shape as the output - compress the input into a lower-dimensional *code (latent-space representation)*

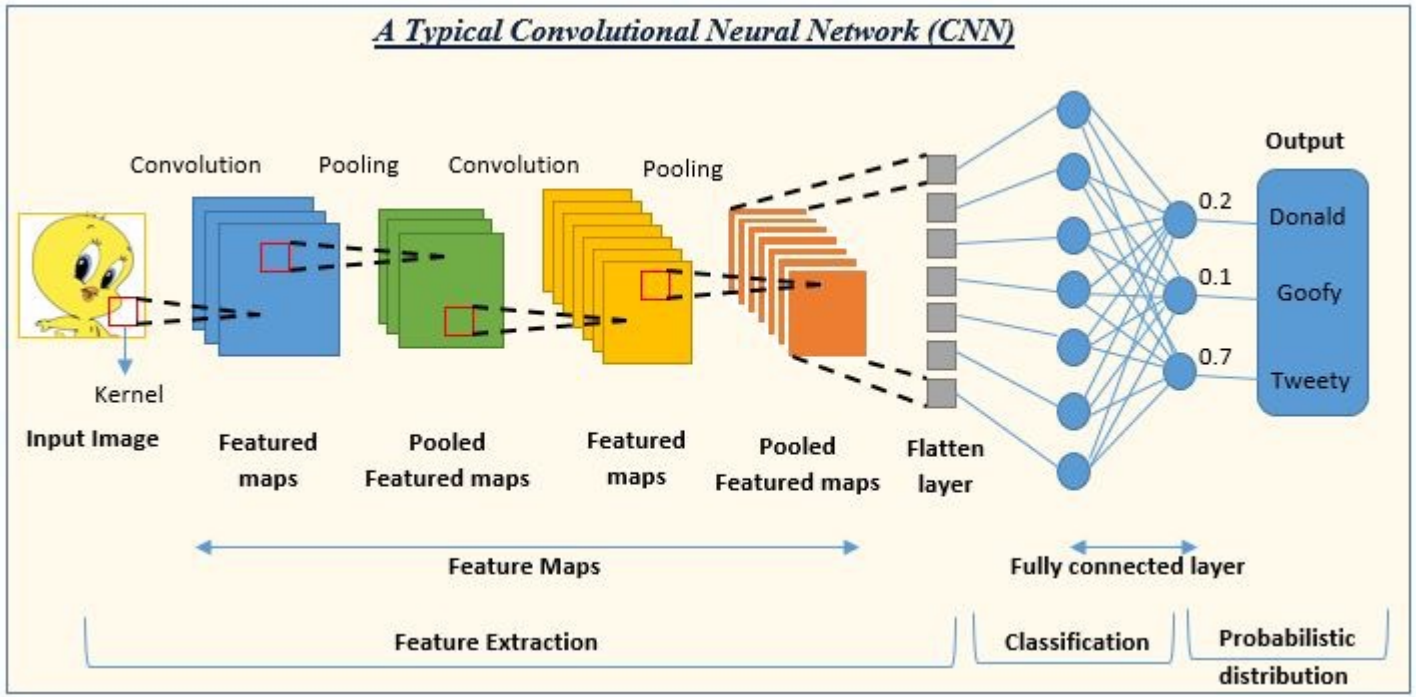
The latent space is determinate

Loss function: KL divergence

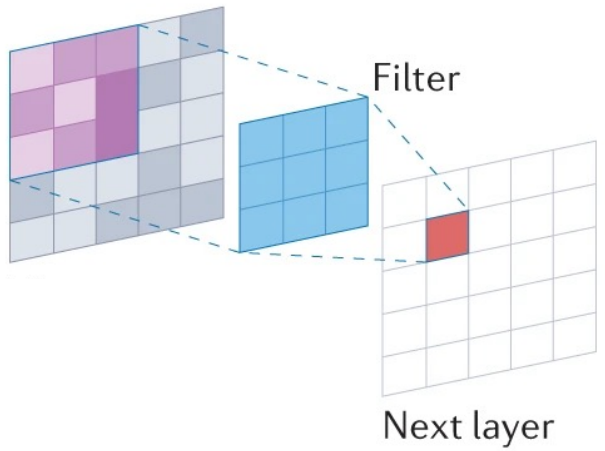
Practical 9.1

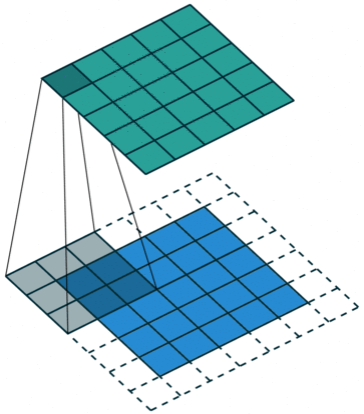
https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

CNN: convolutional neural network



Convolution



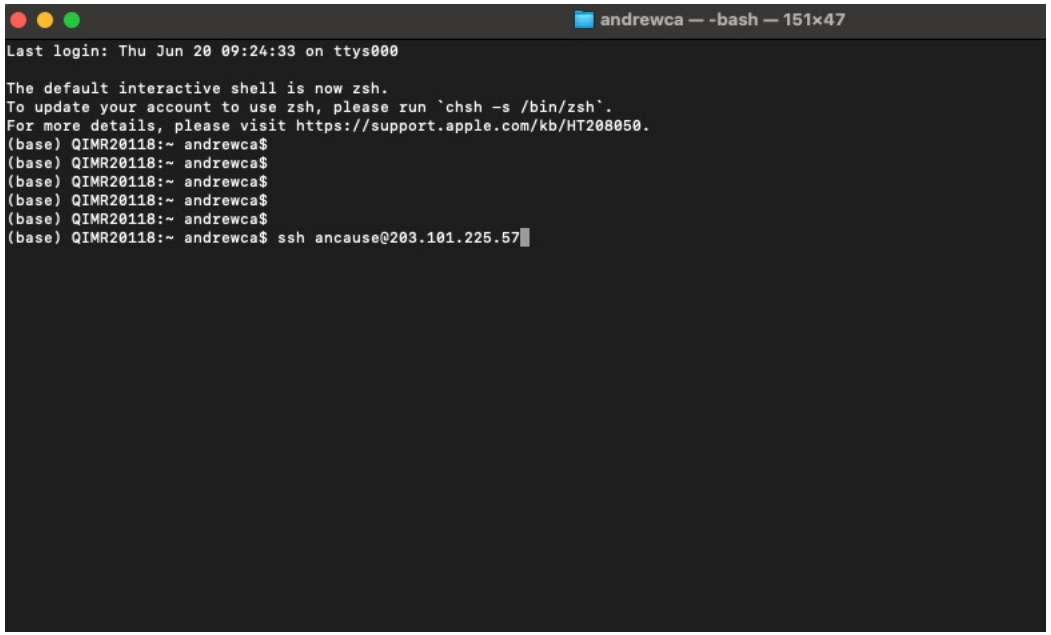


Practical 9.2

https://www.analyticsvidhya.com/blog/2022/01/convolutional-neural-network-an-overview/
https://www.freecodecamp.org/news/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050/

Running the Practical

Terminal PowerShell



1. Log into your account:

   ssh {username}@203.101.225.57
   username & password from winter school email

2. Follow these commands:

   - /software/bin/micromamba shell init
   - source ~/.bashrc
   - micromamba activate /software/conda-envs/winter_school_2024
   - git clone https://github.com/GenomicsMachineLearning/qimr-teaching-2024 /scratch/$USER/qimr-teaching-2024
   - /scratch/$USER/qimr-teaching-2024/runme.sh

3. Open JuperterNotebook:

   /software/006-deep-learning/Deep_learning_tutorial.ipynb