Bioinformatics Lab

Ivan Gesteira Costa & Martin Manolov

Institute for Computational Genomics

Machine Learning / Classification

Gene expression data imposes challenges to classification:

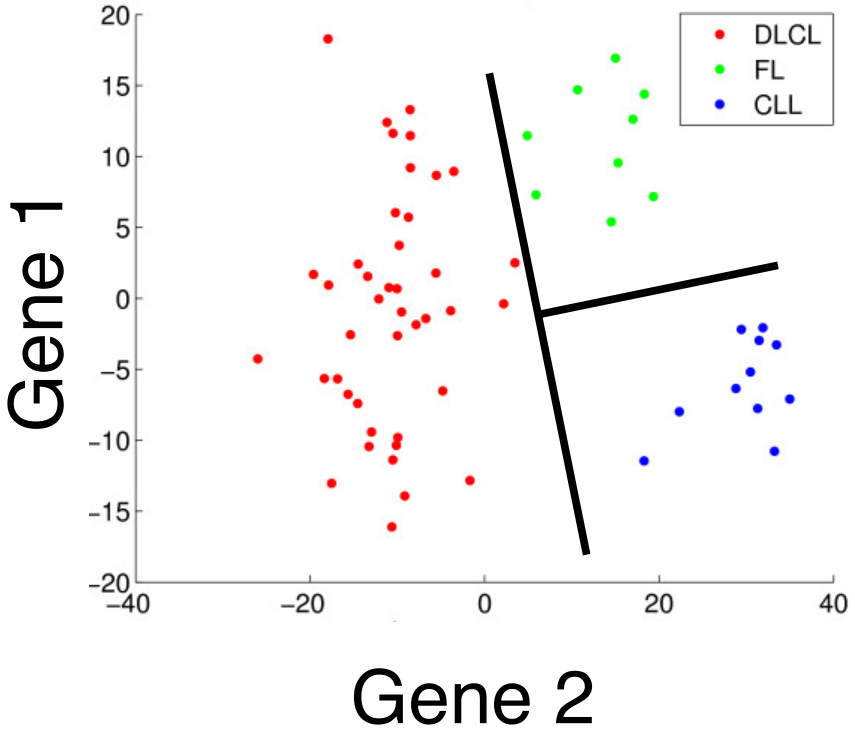
- no. of dimensions is higher (or similar) than number of samples

We need robust experimental approaches for:

- measuring the accuracy of ML methods
- finding best parameters of ML methods
- compare the performance of distinct methods.

Machine Learning - Classifier

cancer type classification



Data:

Expression matrix X (genes vs samples)

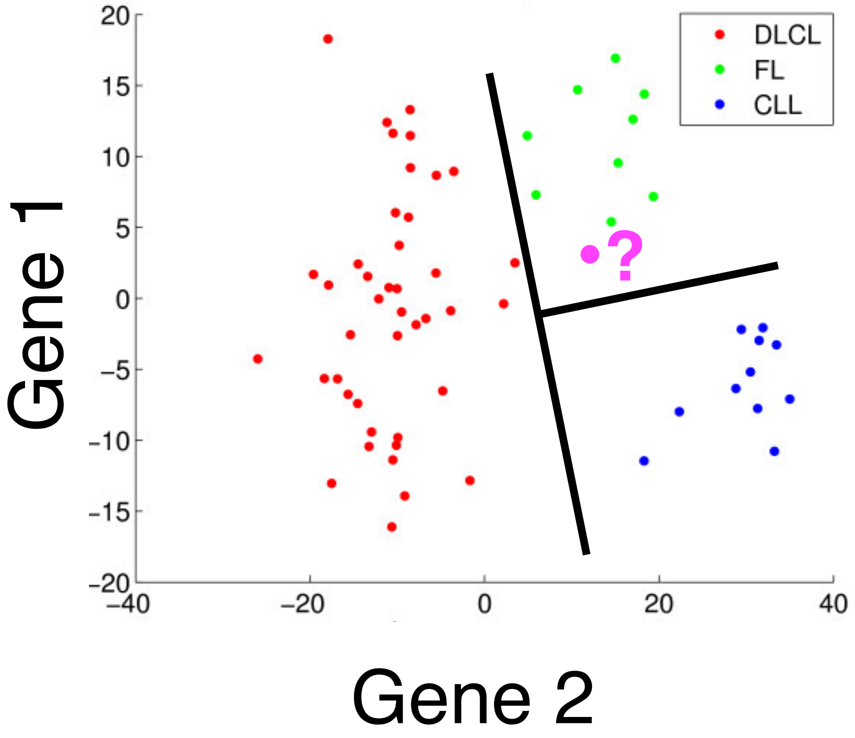
classification vector *Y* (diagnosis)

Find a function:

f(*x*) → *y*

Machine Learning - Classifier

cancer type classification



Data:

Expression matrix X (genes vs samples)

classification vector *Y* (diagnosis)

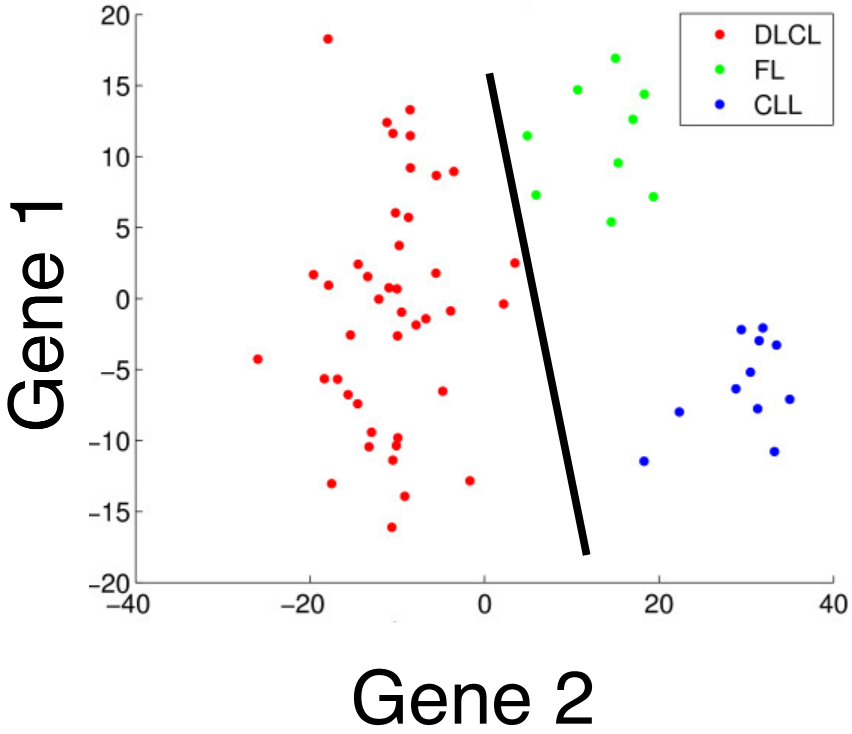
Find a function:

f(x) → y

For new samples X':

f(x') → y'

Linear Classifier



Linear Function:

$$f\,(x\,,A) = a_0 + a_1x_1 + \ldots + a_Lx_L$$

$$f\,(\,x, A) > 0 \Rightarrow \text{classe A}$$

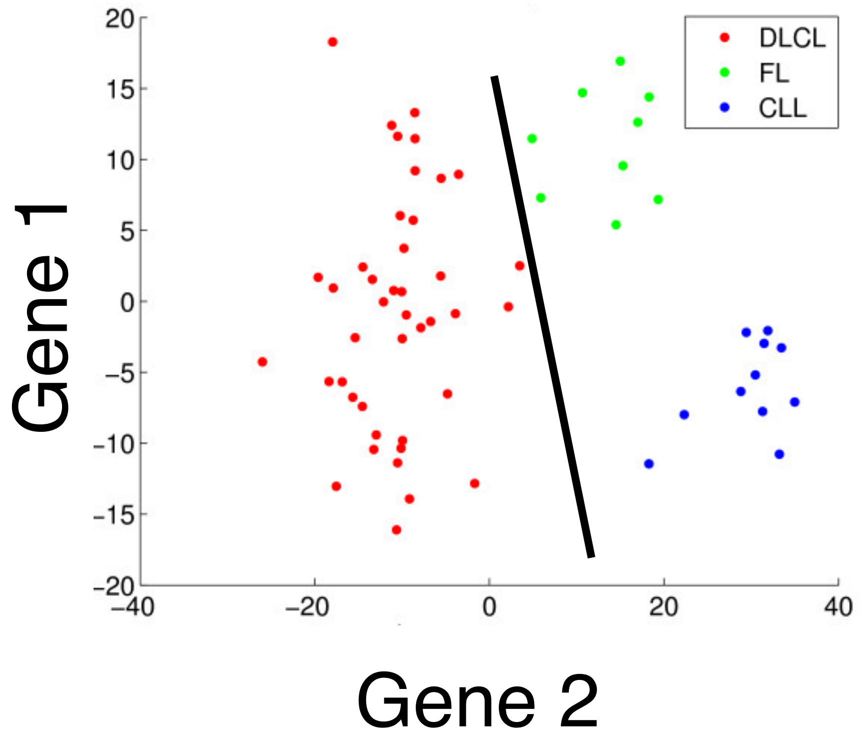
$$f\,(\,x, A) \le 0 \Rightarrow \text{classe B}$$

- **Works for 2 classes only**
  - train a function for each cancer type
- **Find coefficients**
  - with linear programming/ neural network

Institute for Computational Genomics

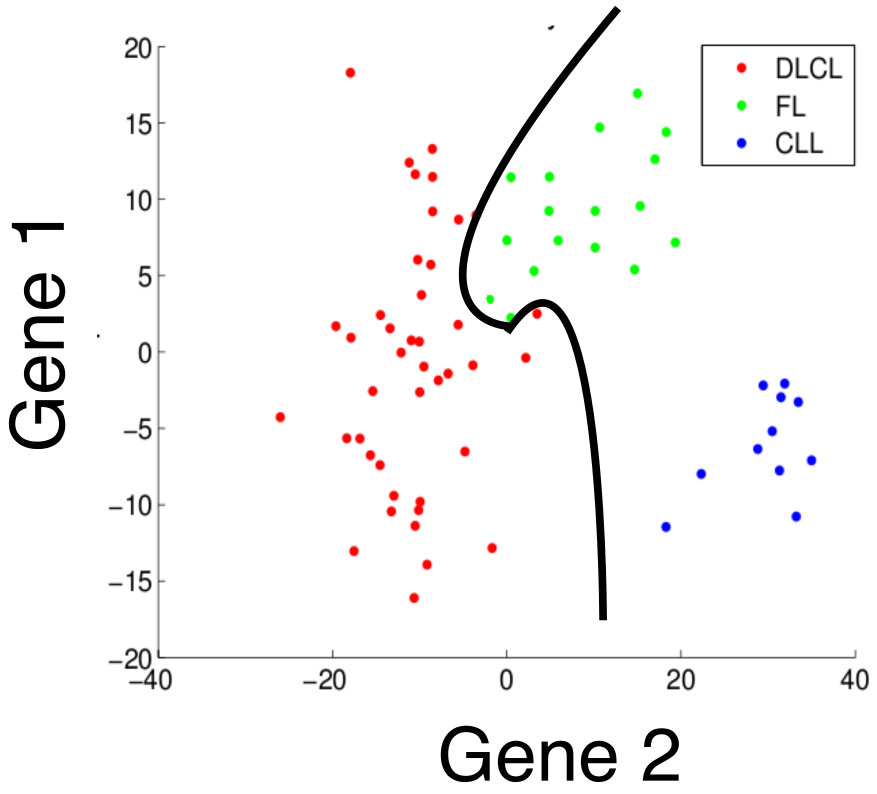
RWTH AACHEN UNIVERSITY

Linear Classifier - Problems



- Most real word problems are not linearly separable!
- There will be always some error!
- Solution: non-linear functions

Nonlinear Classifier - Problems

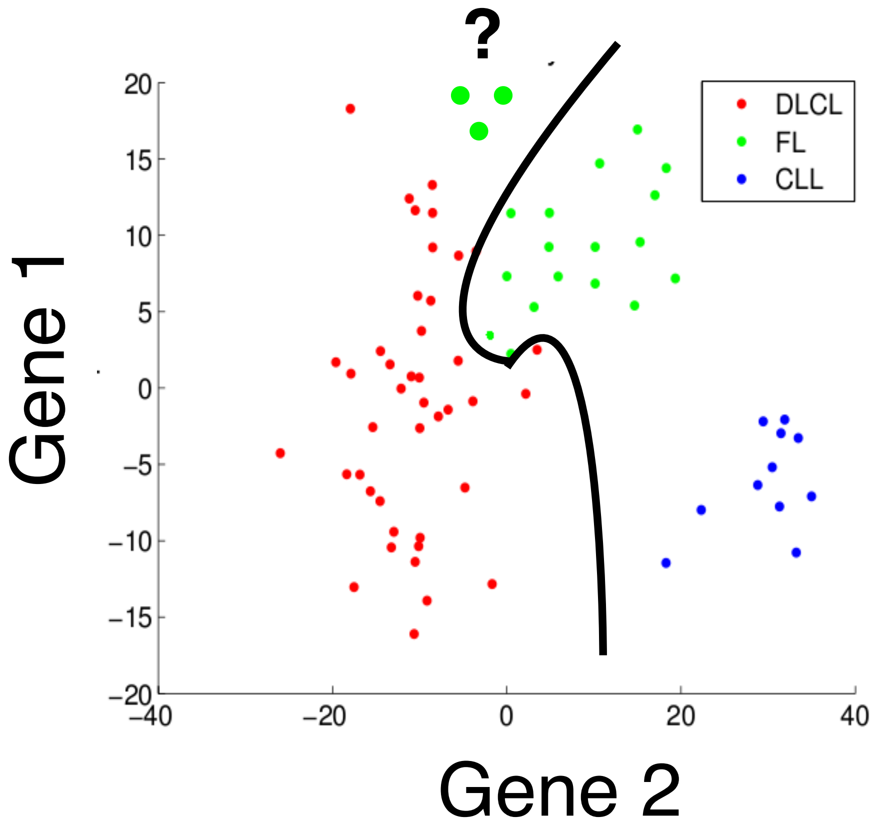


- Polinomial Function
- $f\,(x\,,A) = a_0 + a_{11}x^3{}_1 + \ldots + a_{L1}x^3{}_L$

  $a_{12}x^2{}_1 + \ldots + a_{L2}x^2{}_L$

  $a_{12}x_1 + \ldots + a_{L2}x_L$
- Third order polynomial
- Problem: overfitting

Nonlinear Classifier - Problems

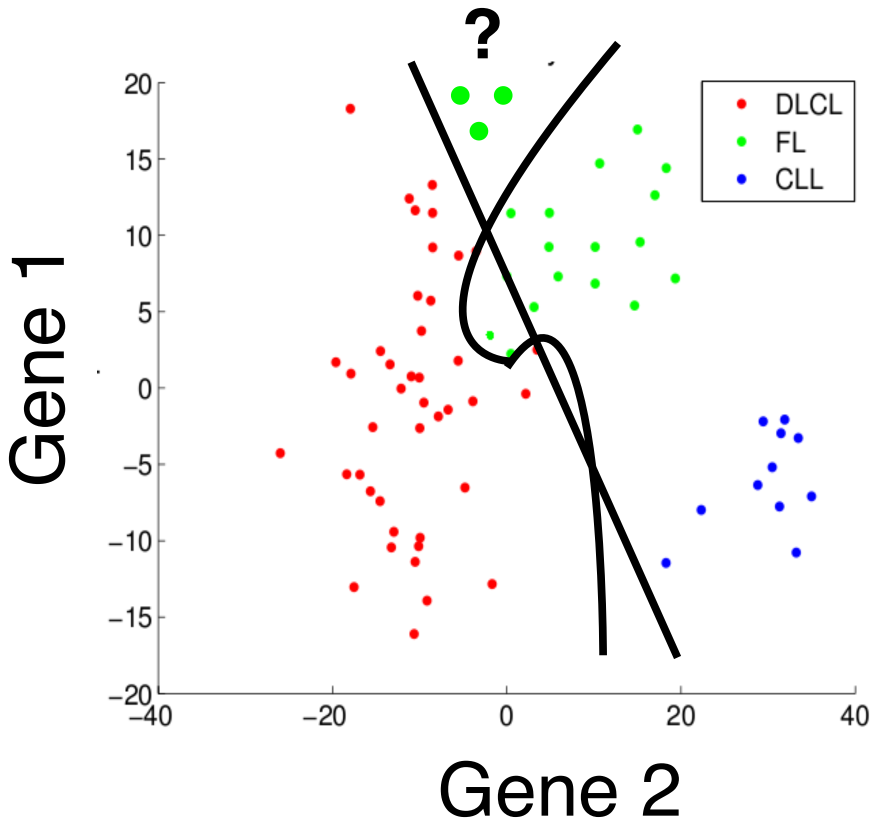


- Polinomial Function
- $f\,(x\,,\,A) = a_0 + a_{11}x^3{}_1 + \ldots + a_{L1}x^3{}_L$

  $a_{12}x^2{}_1 + \ldots + a_{L2}x^2{}_L$

  $a_{12}x_1 + \ldots + a_{L2}x_L$
- Third order polynomial
- Problem: overfitting

Nonlinear Classifier - Problems



- Polinomial Function
- $f\,(x\;,A) = a_0 + a_{11}x^3{}_1 + \ldots + a_{L1}x^3{}_L$

  $a_{12}x^2{}_1 + \ldots + a_{L2}x^2{}_L$

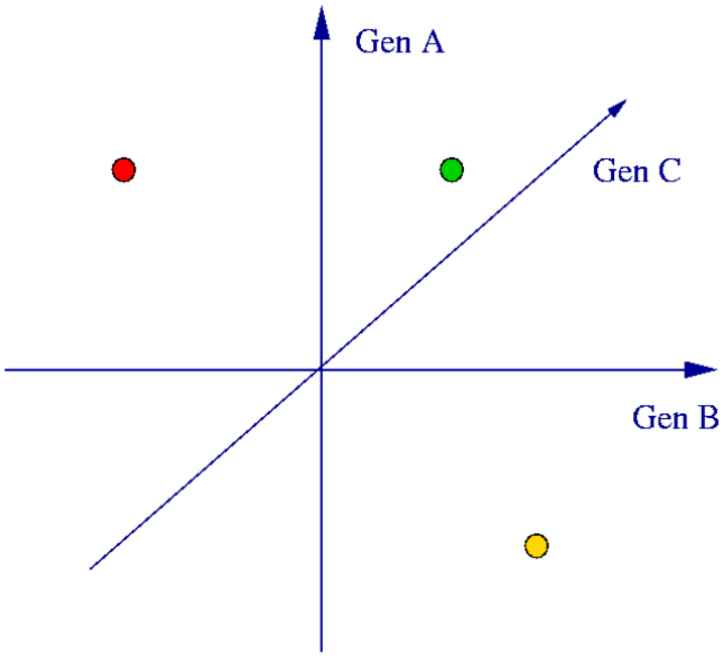
  $a_{12}x_1 + \ldots + a_{L2}x_L$
- Third order polynomial
- Problem: overfitting

Curse of Dimensionality

- Size of a Euclidean space grows exponentially with dimension
  - number of genes
- Dots (patients) are sparsely distributed in space

Curse of Dimensionality : Example

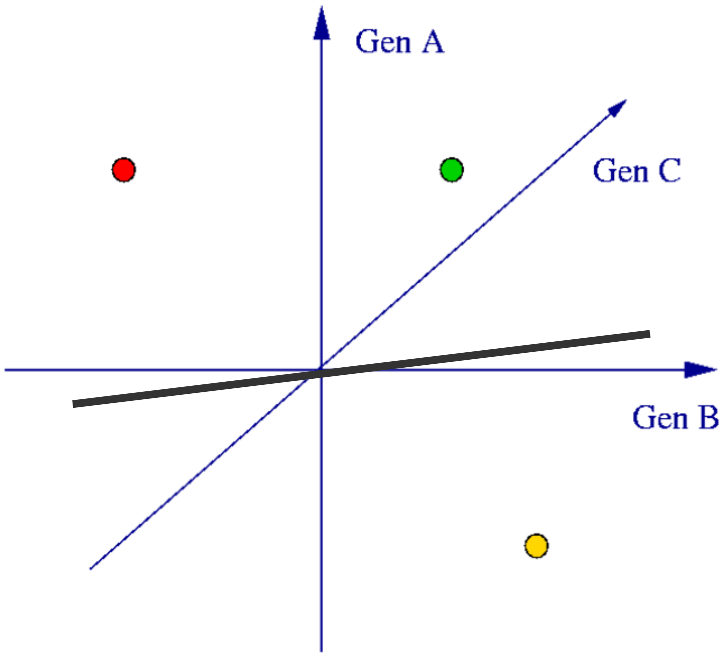
Sparse data: no of samples < no of dimensions



- three genes
- 2 patients with known cancer type(red/yellow)
- 1 unknown cancer type(green)

Curse of Dimensionality : Example

Sparse data: no of samples < no of dimensions

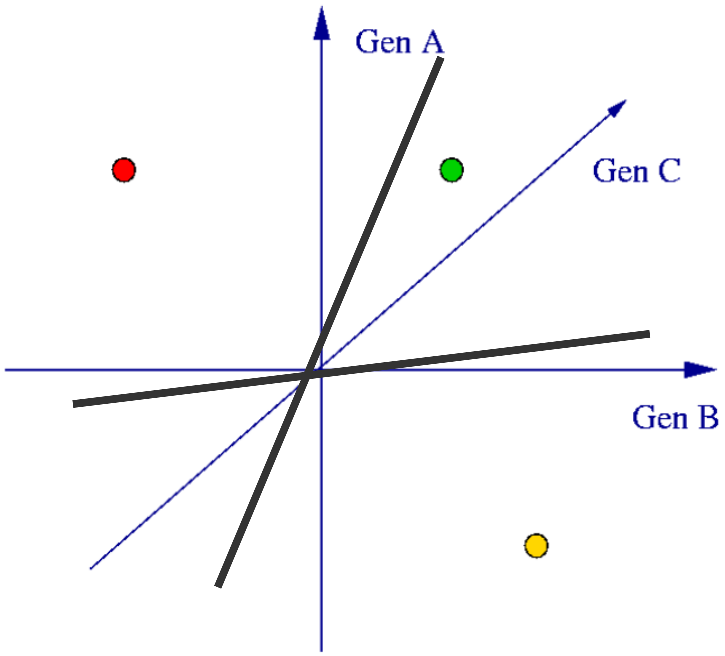


Perfect classifier (on training!)

- three genes
- 2 patients with known cancer type(red/yellow)
- 1 unknown cancer type(green)

Curse of Dimensionality : Example

Sparse data: no of samples < no of dimensions

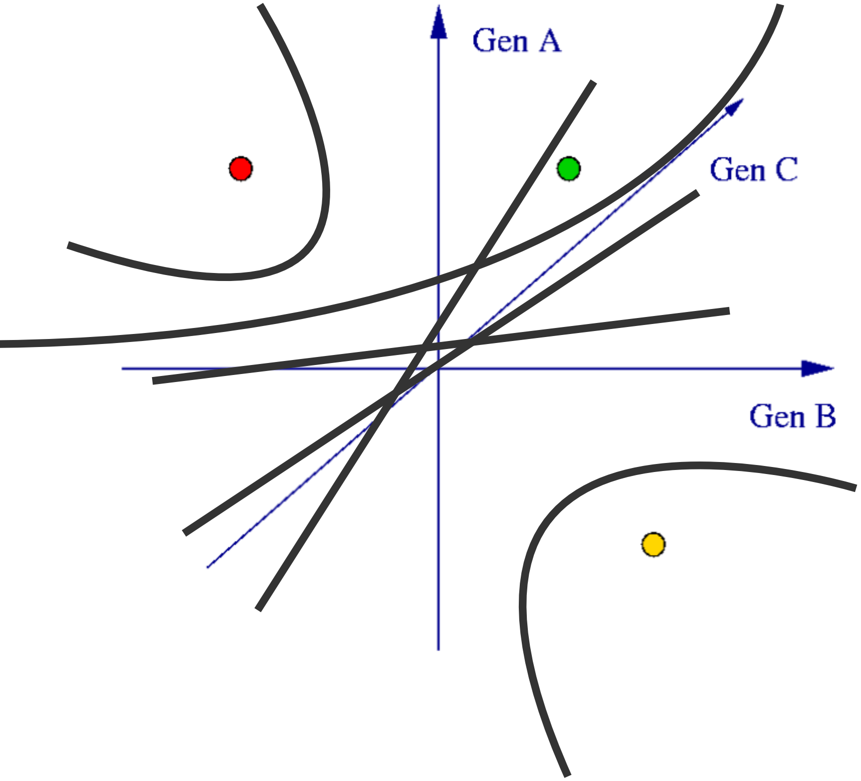


- three genes
- 2 patients with known cancer type(red/yellow)
- 1 unknown cancer type(green)

More perfect classifiers (on training!)
Hard to generalize 1

Curse of Dimensionality : Example

Sparse data: no of samples < no of dimensions



- There are millions of perfect linear classifiers
- And even if non-linear classifiers are considered!

Dealing with Curse of Dimensionality

- Have a proper training / test evaluation procedure
- Use simple classifiers
- Reduce the dimension of your data:
    - feature selection
    - PCA or tSNE (black box!)

Classifier Evaluation

1. **Statistics to measure the classification performance**

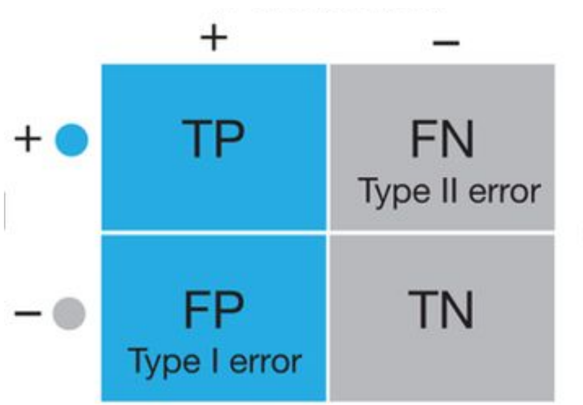
2. Data splitting strategies to avoid overfitting
- ML learns training data but do not generalize to unseen data

Classification Metrics

Measures for two class problem

Predicted Class

Actual Class



	+	–
+ ●	TP	FN Type II error
– ●	FP Type I error	TN

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

$$\text{F1 Score} = \frac{2{*}TP}{2{*}TP+FP+FN}$$

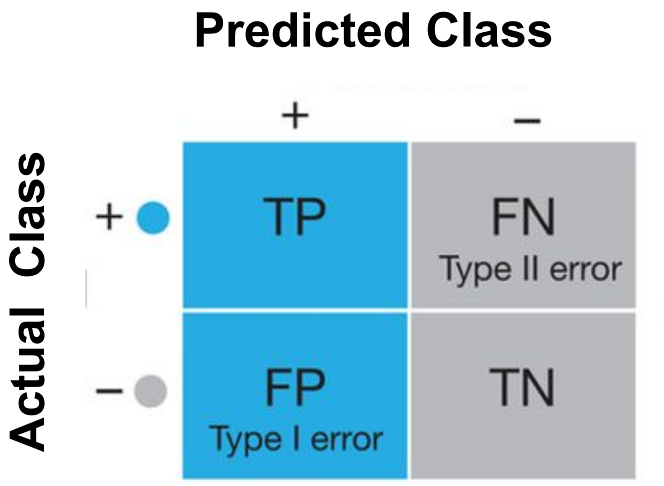
$$\text{Precision} = TP / TP + FP$$

$$\text{Sensitivity/Recall} = TP / TP + FN$$

Source: Lever et al., Nat. Methods (2016)

Classification Metrics

Measures for two class problem



Actual Class \ Predicted Class	+	–
+ ●	TP	FN Type II error
– ●	FP Type I error	TN

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

$$\textbf{F1 Score} = \frac{2*TP}{2*TP+FP+FN}$$

Precision = TP / TP + FP

Sensitivity/Recall = TP / TP + FN

Extension for multi class:

- evaluate class vs. others / use average accuracy / F1.

Class imbalance:

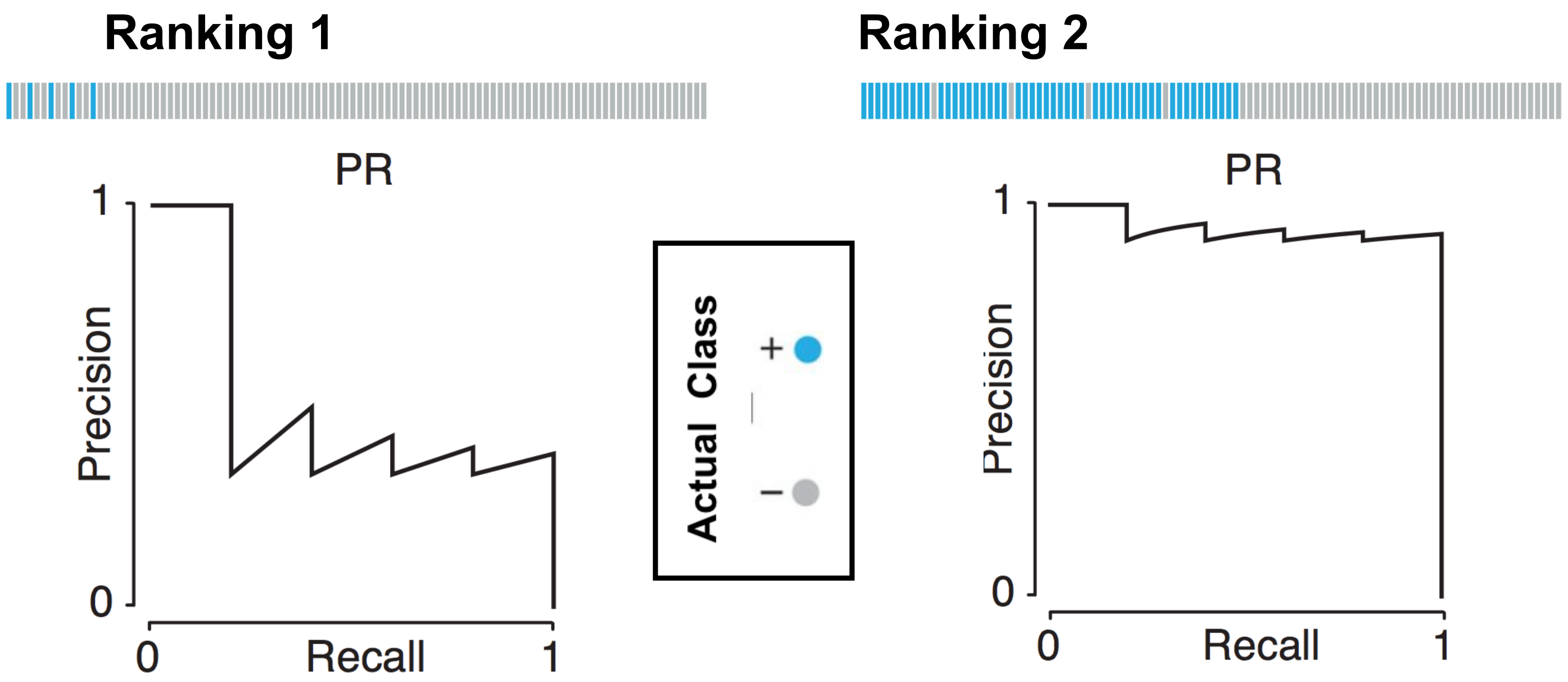
- usually number of negatives is larger / classifiers with low Precision might still have high Acc/Sensitivity

Source: Lever et al., Nat. Methods (2016)

Classification Metrics / Class Imbalance

Precision - Recall (PR) curves

- requires ranking of classification, i.e. class probability

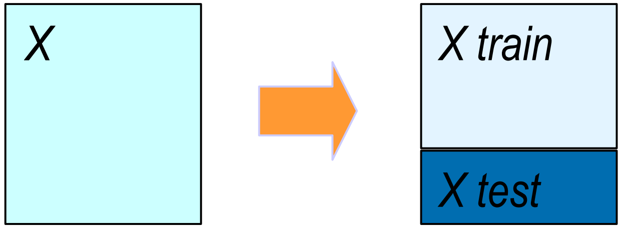


- area under the PR curve -> higher area indicates best classifiers!

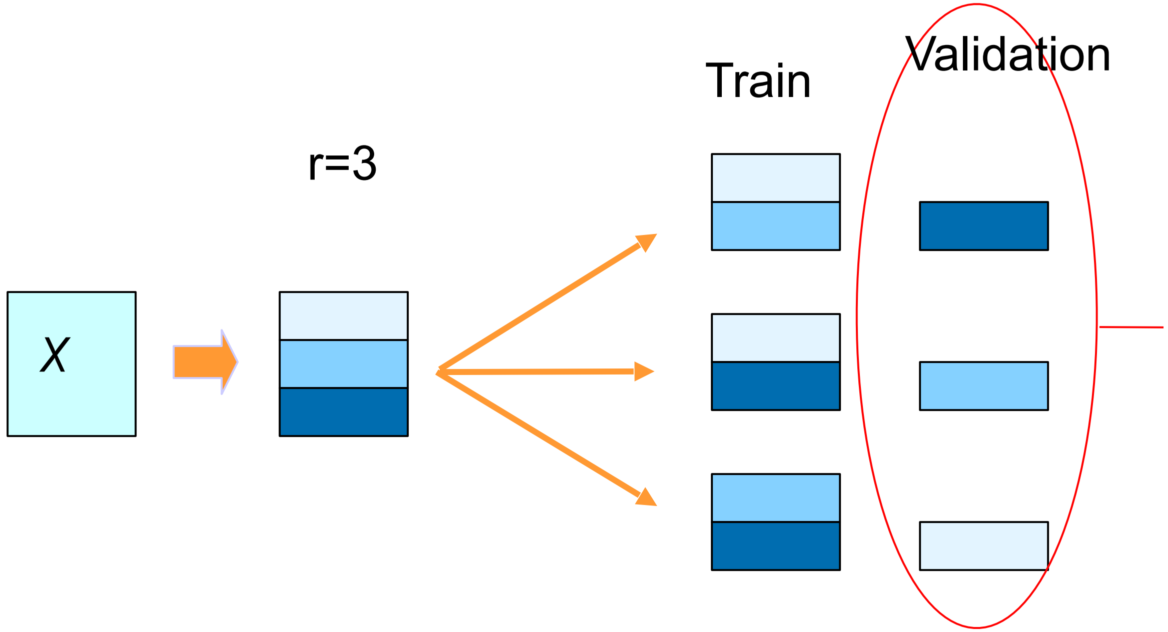
Source: Lever et al., Nat. Methods (2016)

Classifier Evaluation

- The performance of your classifier needs to be evaluated at test data:
  - an independent "test data set"
  - cross-validation



Cross-validation



X

r=3

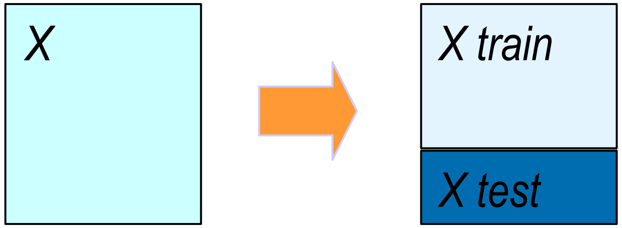
Train

Validation

mean and std of accuracy, precision and sensitivity

Classifier Evaluation

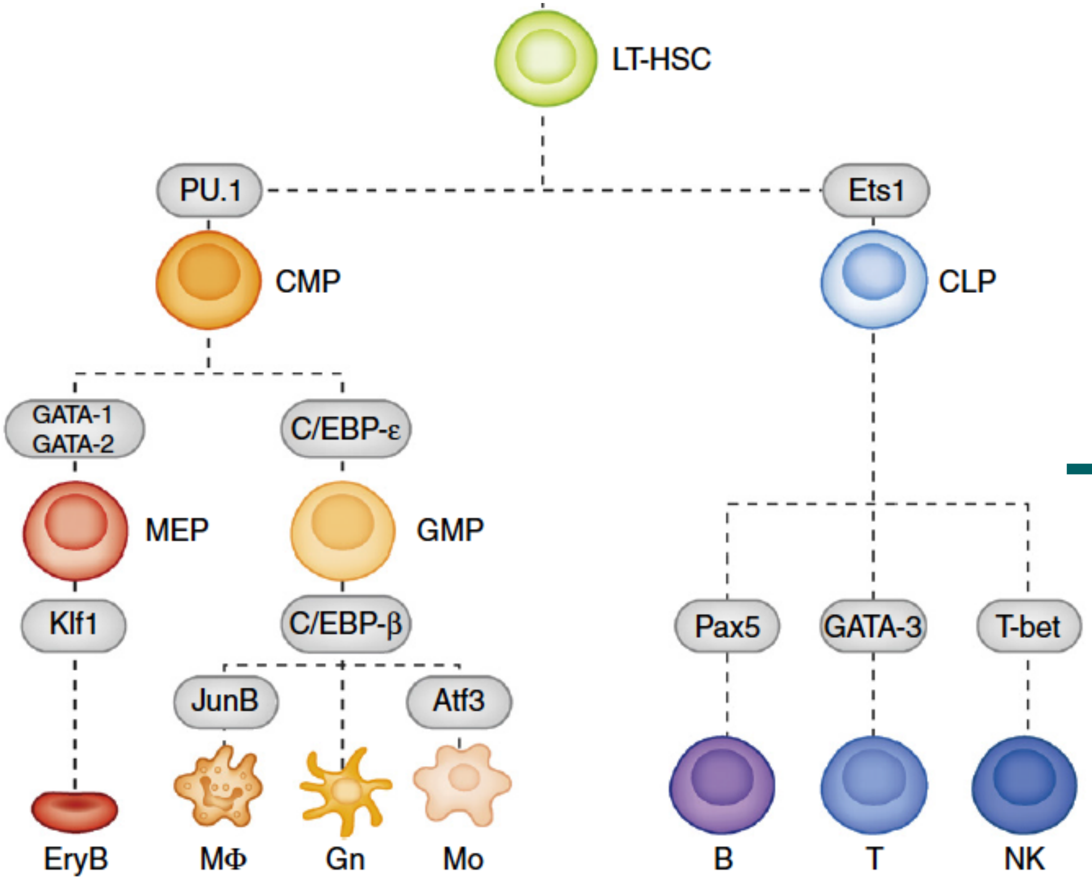
- The performance of your classifier needs to be evaluated at test data:
  - an independent "test data set"
  - cross-validation



- Never use test data to improve classification (choose a better classifier or marker gene)
  - For this you need to establish validation data (or nested cross- validation approach)

Problem Definition

Cell Differentiation & Gene Expression

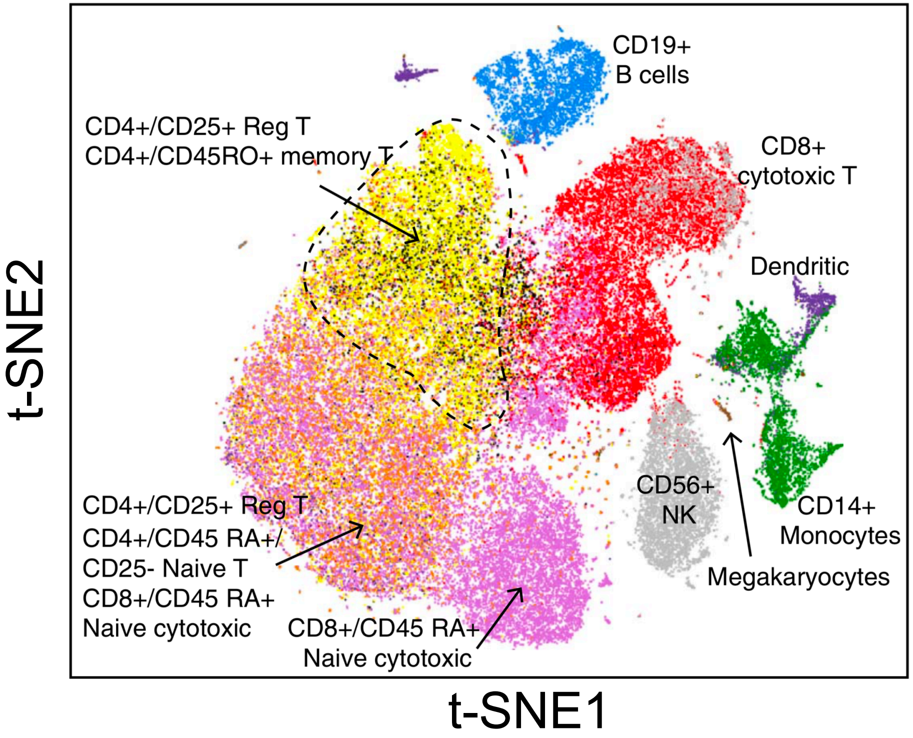


	Cell 1	Cell 2	…
Gene 1	25	918	
Gene 2	0	456	
Gene 3	20	342	
Gene 4	0	214	
…			

Source: Amit (2016), *Nature Immunoloy.*

Gene Expression of Lymphoid Cells

PBMCs from Humans

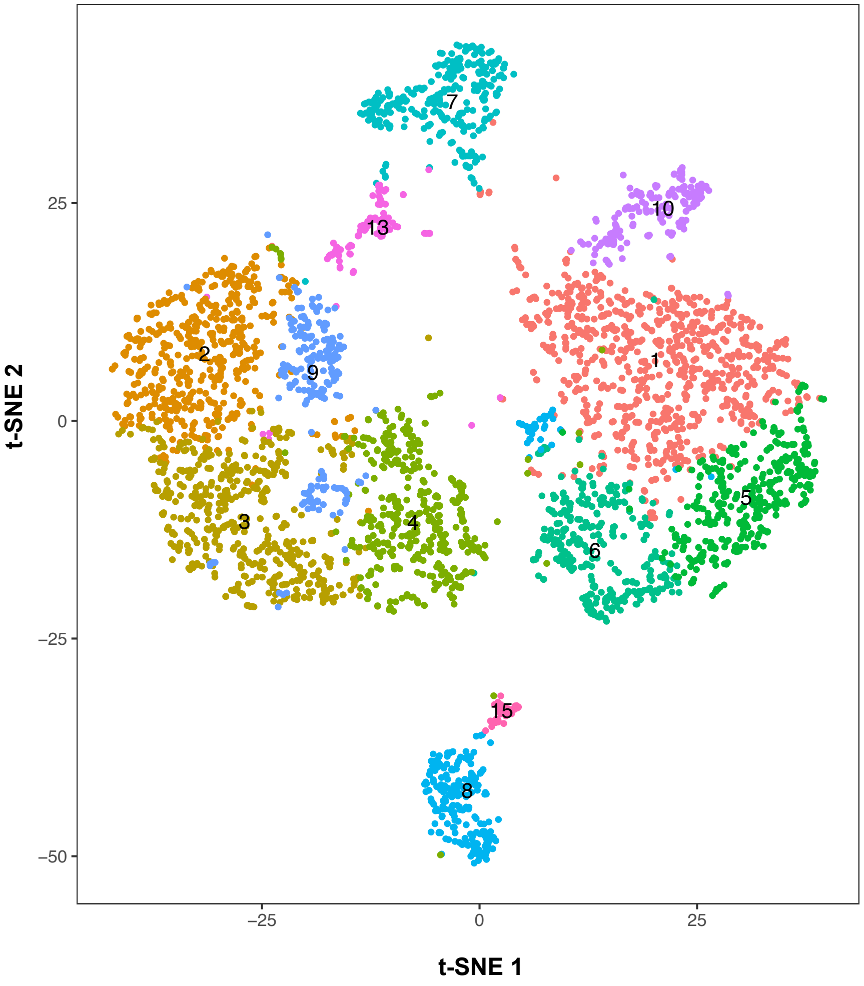


Single cell RNA-seq from 68k cells

Source: Zheng et al. 2017 & Buenrostro et al. 2018

Basics Bioinformatics - Clustering

Gut Immune Cells - 12 groups

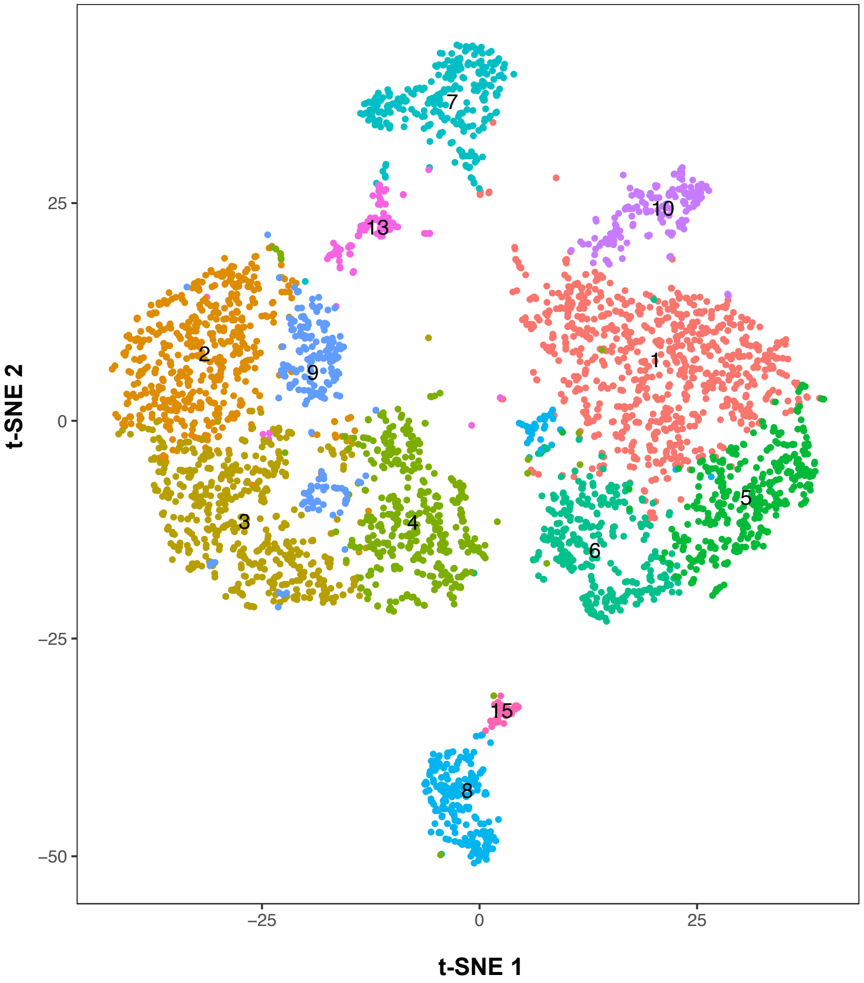


Clustering - identify cells with similar expression patterns
- based on PCA (20 dimension)

How to identify cell types?

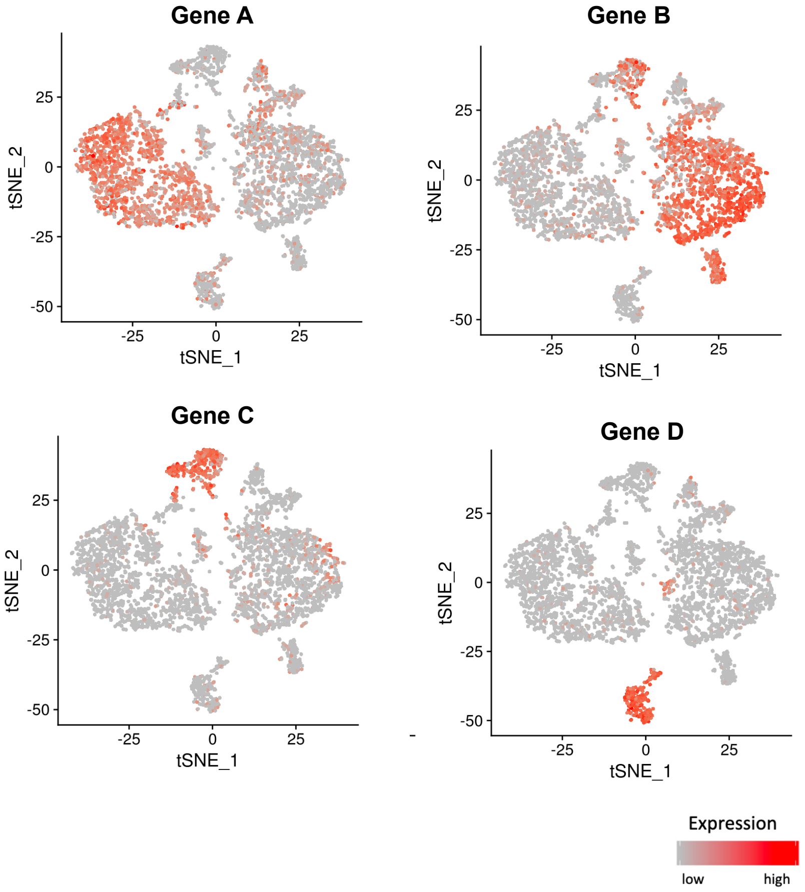
Cell Identity with an Expert

Gut Immune Cells - 12 groups



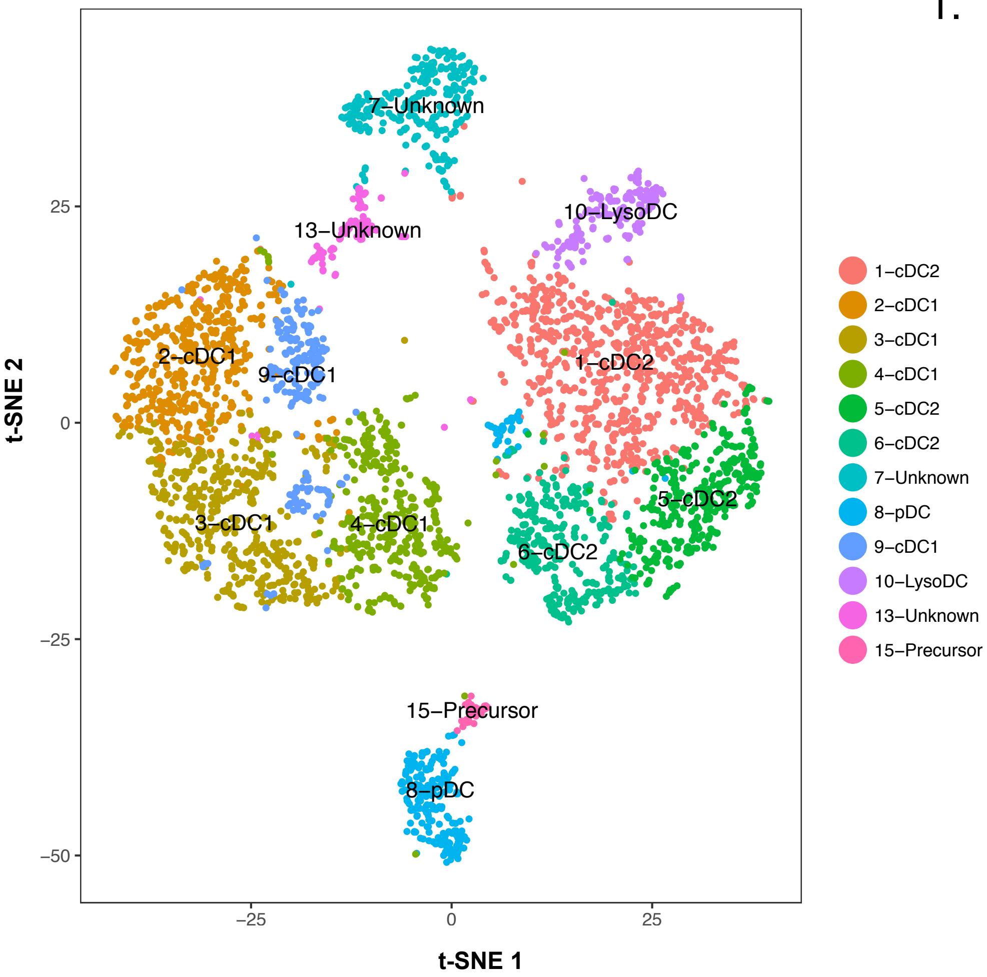
Check expression of:

1. known genes



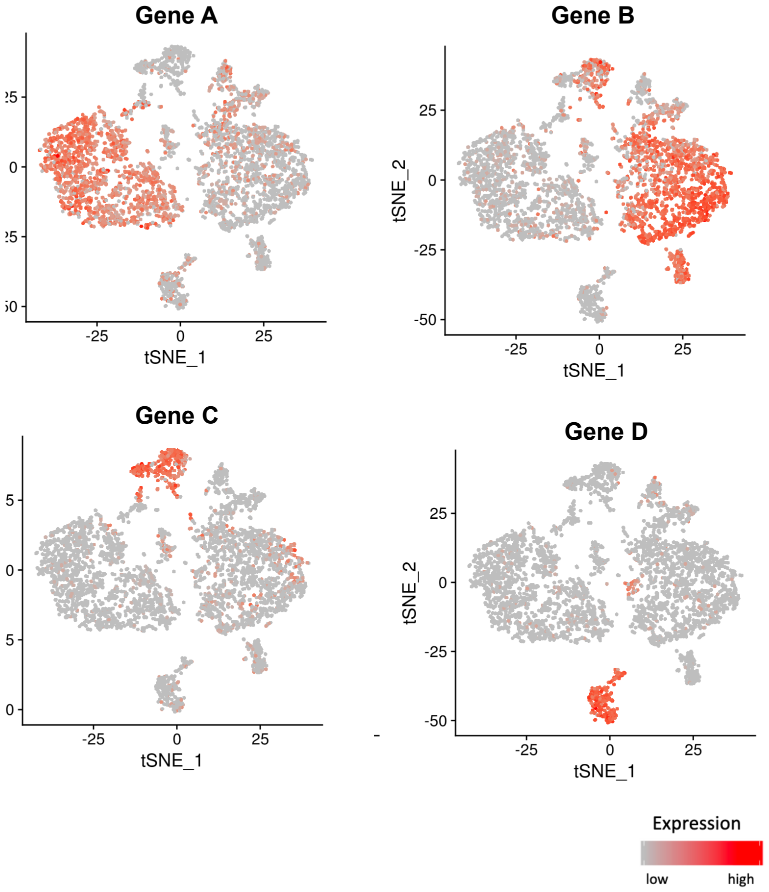
Cell Identity with an Expert

Gut Immune Cells - 12 groups



Check expression of:

1. known genes

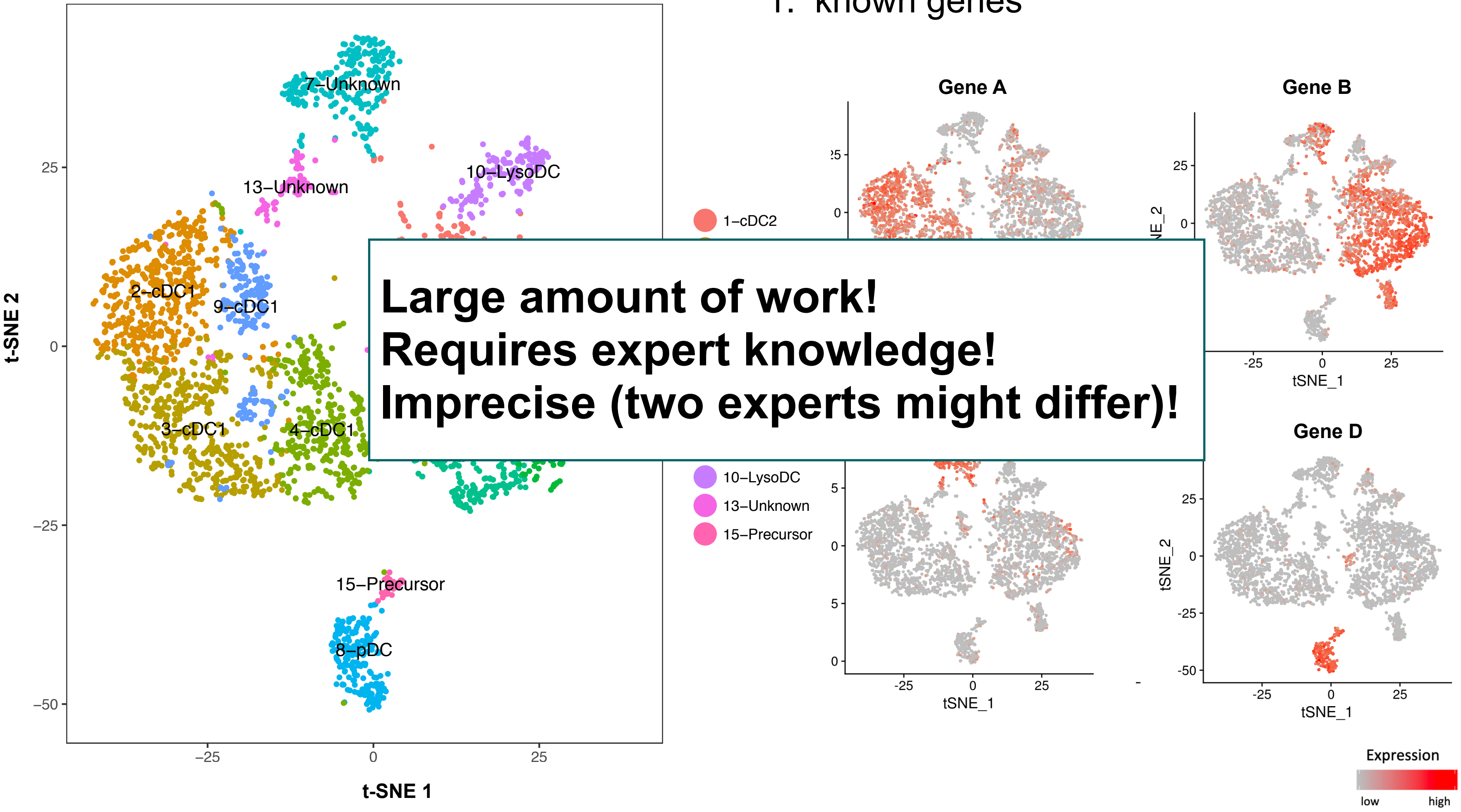


Cell Identity with an Expert

Gut Immune Cells - 12 groups

Check expression of:

1. known genes



Large amount of work!
Requires expert knowledge!
Imprecise (two experts might differ)!

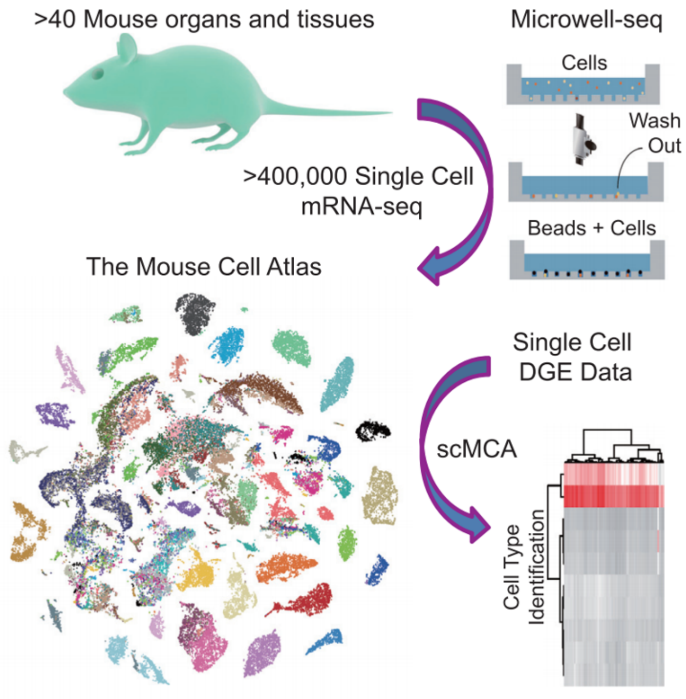
Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

Automatic Cell Identification

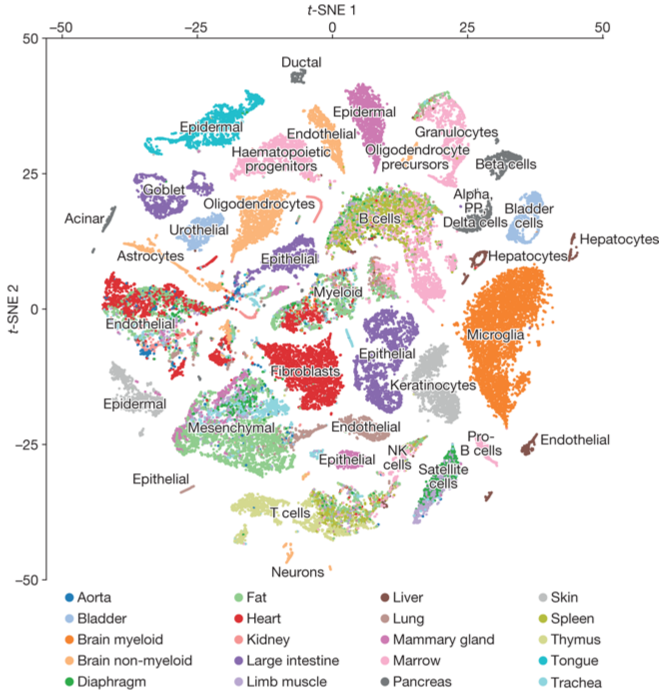
Large consortia provide gene expression and annotation of cells
- annotation is based on *cell ontology*

Mouse cell atlas (MCA)



400.000 cells on 40 tissues

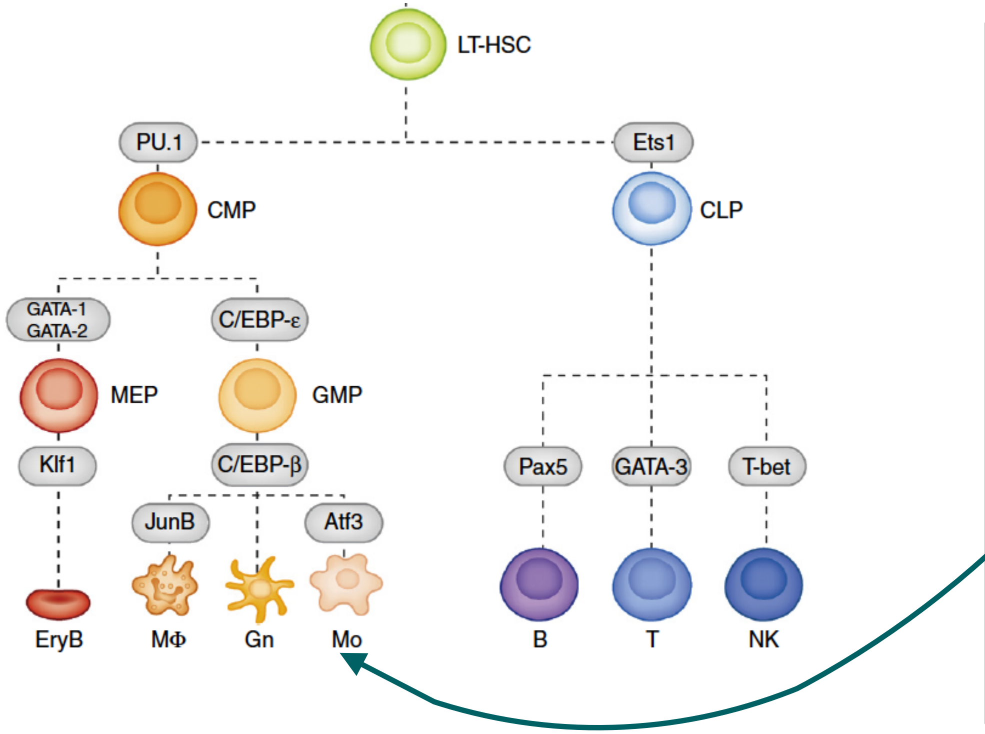
Tabula Muris (TM)

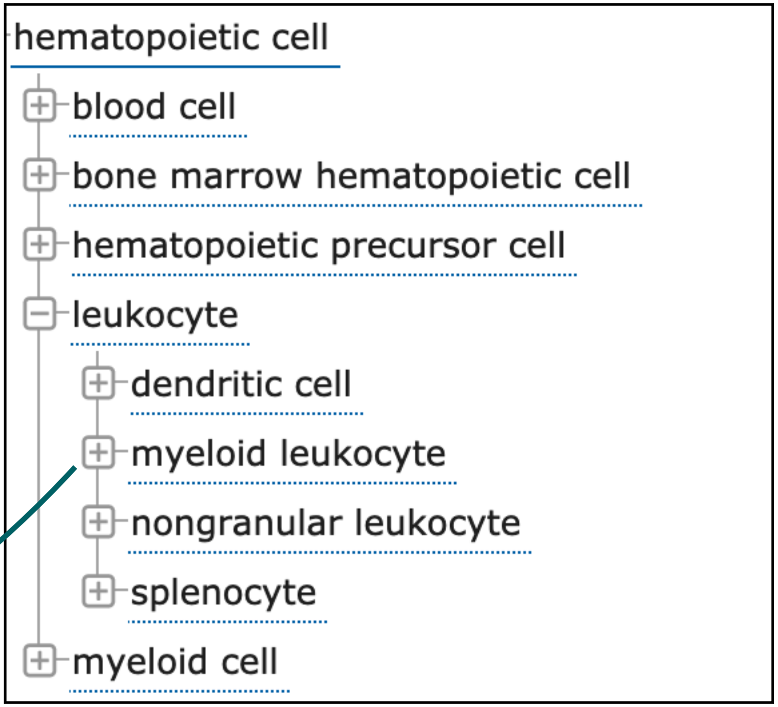


100.000 cells on 20 tissues

Cell Ontology

Controlled vocabulary for cell types in animals




hematopoietic cell
- ⊞ blood cell
- ⊞ bone marrow hematopoietic cell
- ⊞ hematopoietic precursor cell
- ⊟ leukocyte
  - ⊞ dendritic cell
  - ⊞ myeloid leukocyte
  - ⊞ nongranular leukocyte
  - ⊞ splenocyte
- ⊞ myeloid cell

https://www.ebi.ac.uk/ols/ontologies/cl

Available as Json format at:
https://github.com/obophenotype/cell-ontology

Overall Design / Basic Approach

Use machine learning for cell type classifiers:

- elastic net, Neural Networks, Random Forests

For each organ from MCA build a classifier:

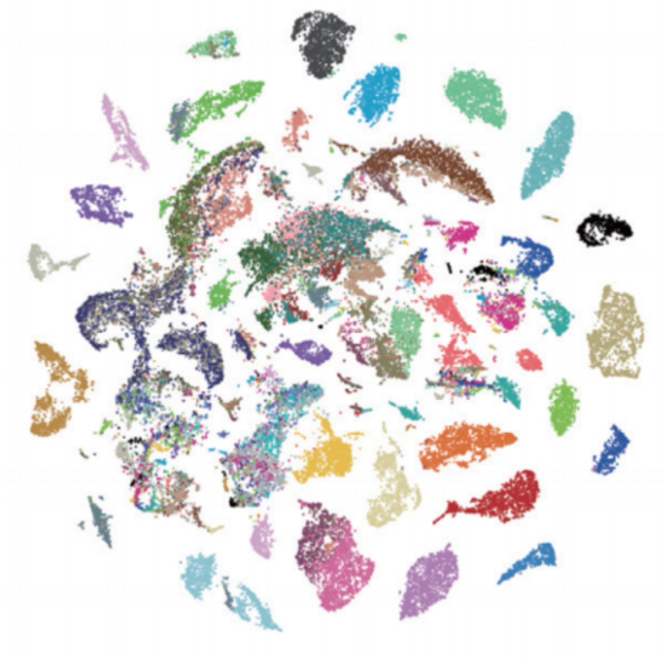
- i.e. Peripheral-Blood from MCA
- check/revise cell annotation (using cell ontology)
- use this data for classifier training/parameter selection with **cross-validation**
- use **area under PR curve** for selection

Test data:

- Find respective organ in TM (i.e. bone marrow)
- Revise cell annotation
- Measure cell type accuracy (PR curve) of MCA model in TM data

Automatic Cell Identification

Mouse cell atlas & Tabula Muris



400.000 cells on 40 tissues

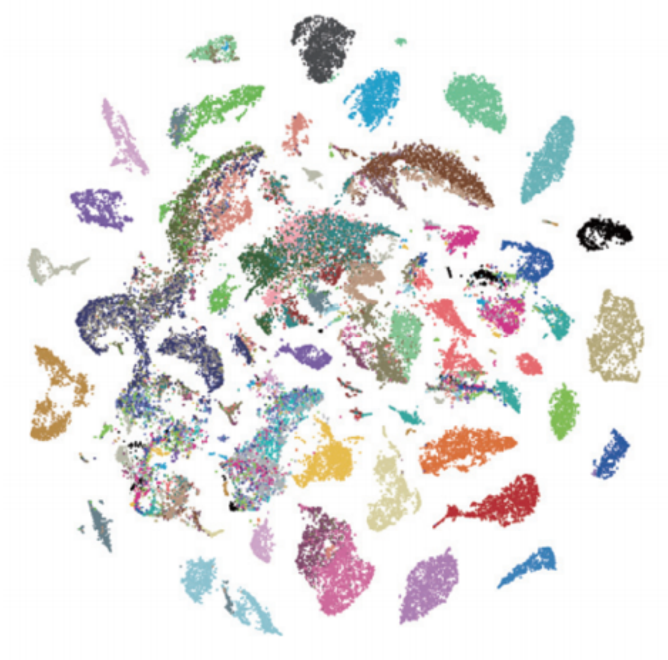
Use pre-annotated cells to build classifiers to annotate novel single cell data (diseases)

Methodological questions:

1. Which machine learning methods to use?
   - Neural networks, statistical methods, ….
2. Feature selection (vs. Blackbox)
   - Find reliable markers from classifiers?
3. Are classifiers robust on sparse data?
   - Evaluate performance when reducing number of reads

Automatic Cell Identification

Mouse cell atlas
& Tabula Muris



400.000 cells on 40 tissues

Challenges:

1. Detect unknown/unseen cells?
   - Detect progenitor cells?

2. Build classifiers across tissues/ whole body?

3. Annotate human samples with mouse trained classifiers?

Challenges: Unseen cells

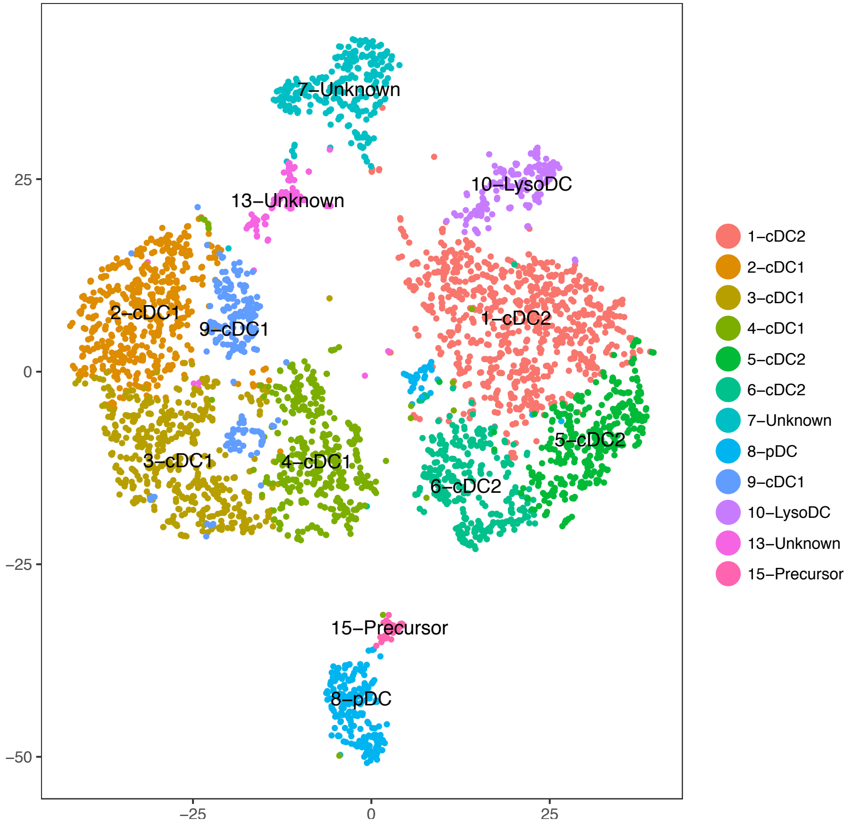
Test data has cell types, which are not included in your classifier.

- You train data did not contained enough cells
- new cell types only found in a disease condition (test data).
- …

Build classifiers that recognise unknown cells

- classifiers have a confidence level
- Indicate that cells with low confidence are unknown

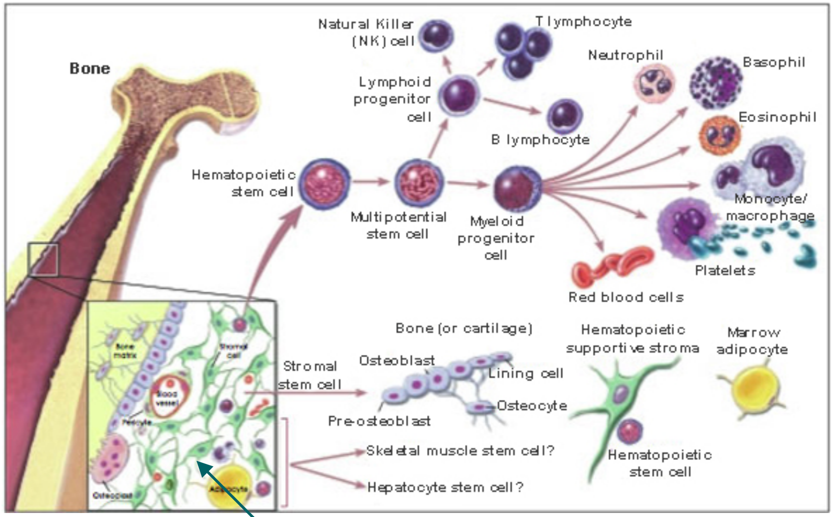
Example: gut immune cells



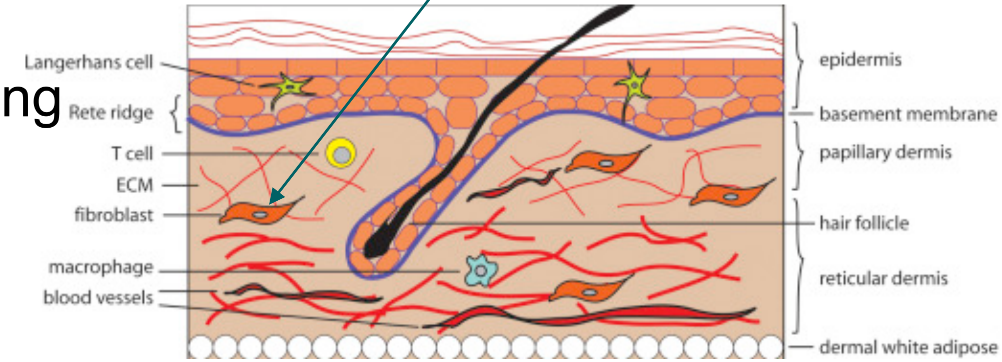
Challenges: Cross organs classification

Most cells are tissue specific
- parenchyma cells
    - aveoli in lungs
    - hepatocytes in liver
    - …

Some cells are in several orgasms:
- stromal cells -> adipose cells, bone cells, fibroblast
- immune cells
- these cells might differ depending of the tissue.



stromal cells

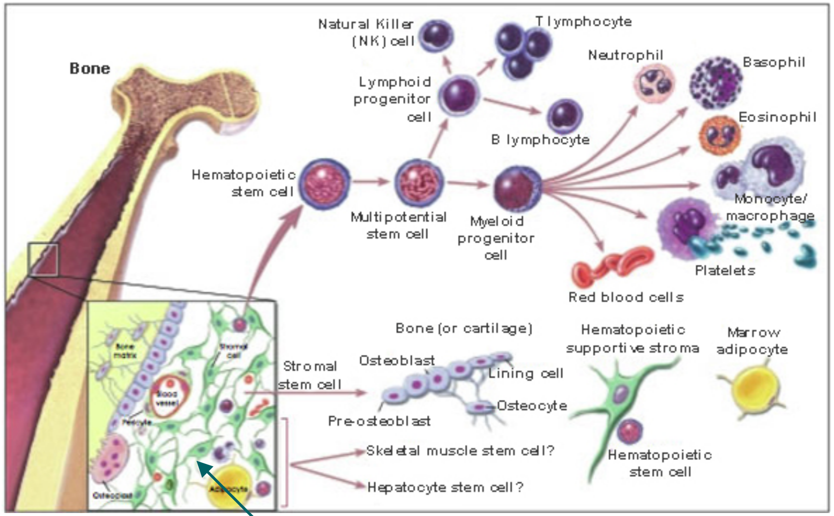


Challenges: Cross organs classification

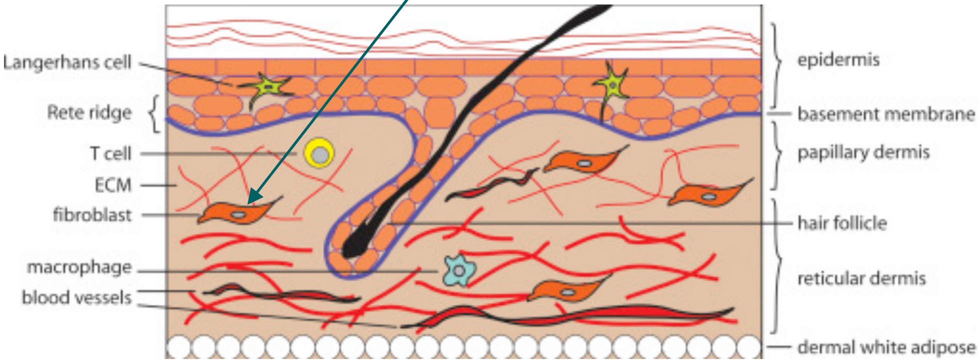
We known the origin/organ of a data.

What is the best strategy to build classifiers?

- a classifier per tissue?
- whole body classifiers?
- combination: per tissue for parenchyma cells and whole body for others?



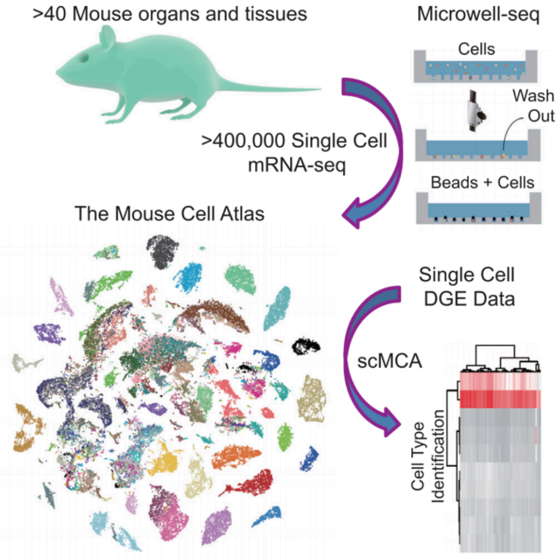
stromal cells



Challenges: Cross organism classification

Use mouse data to classify human samples
- gene names can be mapped but gene function might differ.

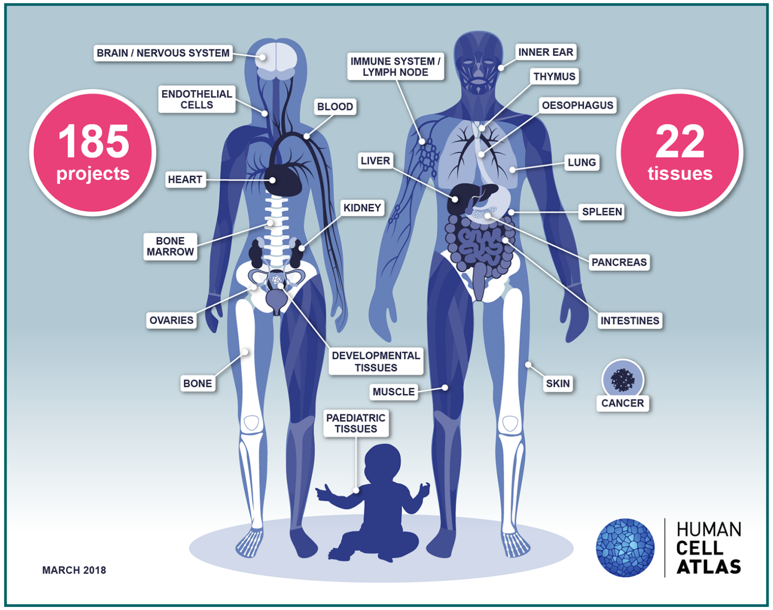
Mouse cell atlas (MCA)



400.000 cells on 40 tissues

Classify

Human Cells



Still being built

Project Proposal

- **Groups: 3-4 participants each**
- **Each group addresses a method problem and challenge**

Method problem

1. Which machine learning methods?
2. Feature selection?
3. Are classifiers robust on sparse data?

Challenges

1. Detect unknown cells
2. Cross tissues/whole body classifiers?
3. Cross organism classifier?

- **Build classifiers and evaluate then on all MCA/TM data**

- additional tasks and data might be defined during the course.

- **Projects code should be deposited in gitlab (git.rwth-aachen.de)**

Calendar

27.05.2019 to 8.07.2019 – Project Development

15.07.2019 – Project Presentation

Links

- Machine learning libraries:
  - python - scikits - https://scikit-learn.org/stable/
  - python & gpu - https://keras.io/
  - R - several individual packages
    - i.e. http://topepo.github.io/caret/index.html
    - seurat / low level single cell and cluster analysis
    - - https://satijalab.org/seurat/

- Cell Ontology:
  - https://github.com/obophenotype/cell-ontology

- Single cell data repositories:
  - Tabua Muris (TM)
    https://figshare.com/articles/MCA_DGE_Data/5435866
  - Mouse cell atlas (MCA)
    https://figshare.com/articles/MCA_DGE_Data/5435866

Relevant data is already at the RWTH Cluster
/hpcwork/nova0028/BioinfoLab/data

Thank you!