

Sex Classification and Identity Verification from GC×GC of Human Scent Samples



Oleksii Kaminskyi
Jana Cechova
Petra Pojmanova
prof. RNDr. Stepan Urban, CSc.

Jan Hlavsa
Radim Spetlik
prof. Ing. Jiri Matas, PhD.

Acknowledgments

L U M I

Acknowledgments

L U M I

- computations performed on LUMI supercomputer - approx. 120k GPU hours

Acknowledgments

L U M I

- computations performed on LUMI supercomputer - approx. 120k GPU hours

∞ \$250K when computed on
rented MI250

Overview

1. We utilize raw outputs of GCxGC MS-ToF.

Overview

1. We utilize raw outputs of GCxGC MS-ToF.
2. We interpret GCxGC MS-ToF data as image.

Overview

1. We utilize raw outputs of GCxGC MS-ToF.
2. We interpret GCxGC MS-ToF data as image.
3. We align the data **WITHOUT** any human interaction.

Overview

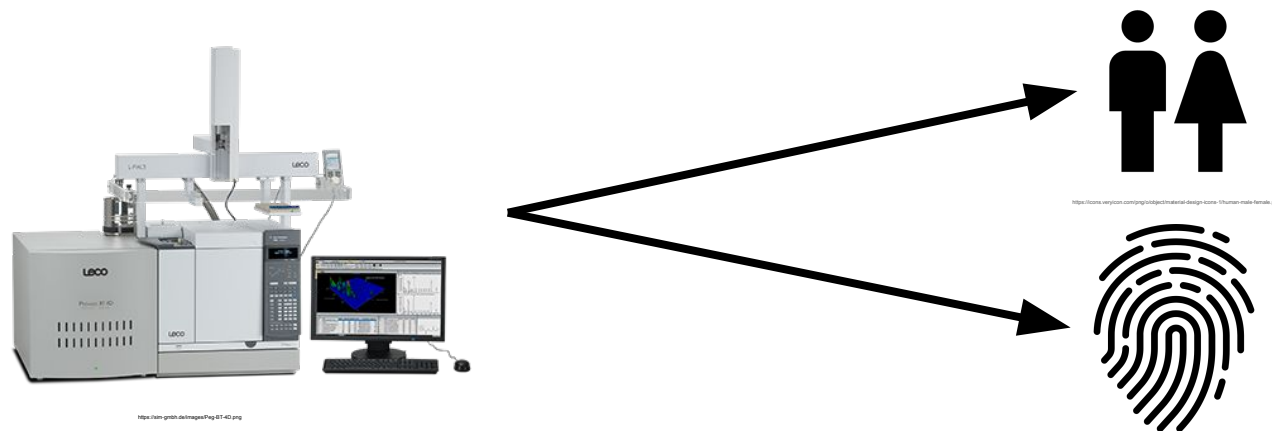
1. We utilize raw outputs of GCxGC MS-ToF.
2. We interpret GCxGC MS-ToF data as image.
3. We align the data WITHOUT any human interaction.
4. We use gradient descent to minimize error on target task.

Overview

1. We utilize raw outputs of GCxGC MS-ToF.
2. We interpret GCxGC MS-ToF data as image.
3. We align the data WITHOUT any human interaction.
4. We use gradient descent to minimize error on target task.
5. It is open-source python, available soon.

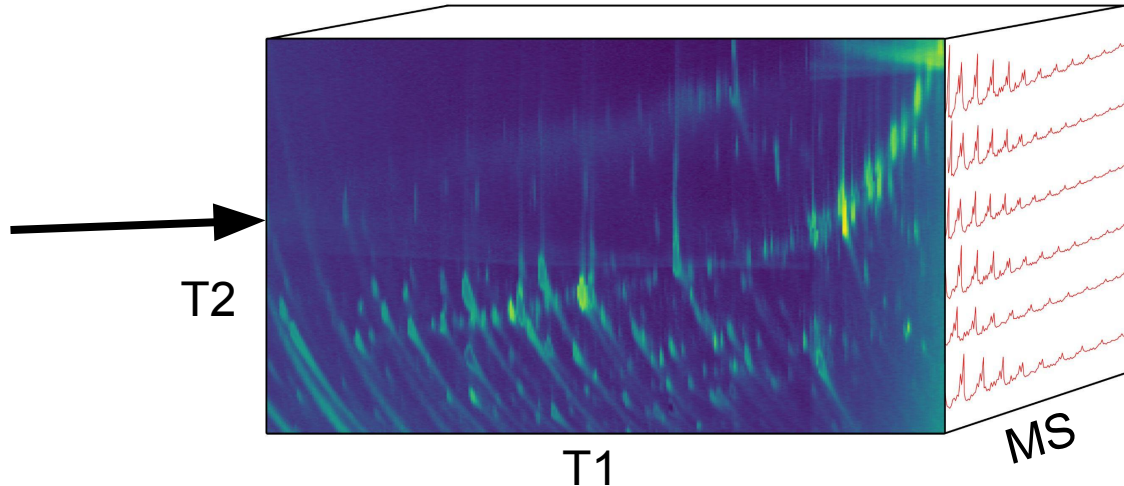
Overview

- We use raw outputs of GCxGC ToF-MS to perform
 - biological sex classification
 - given the sample of human scent identify the biological sex as male or female
 - identity verification
 - given two samples of human scent determine if they belong to the same individual



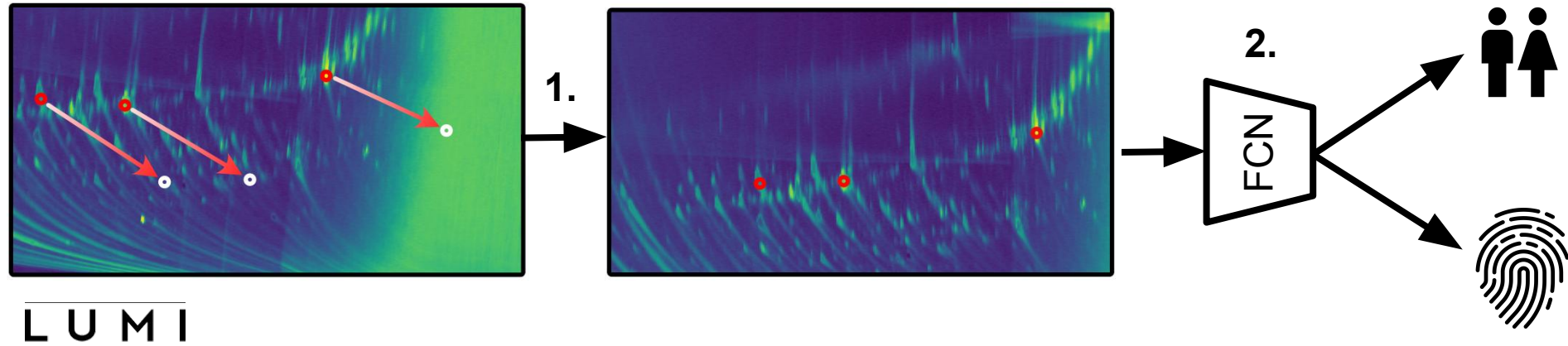
Data

- raw outputs of GCxGC ToF-MS = 3D tensors (MS x T1 x T2) approx. 4GB
- 2528 samples in total - 252 identities (130 men, 122 women)
 - split into training and validation set - with 80:20 proportions
- experiments with full data (FULL) and sum along spectral dimension (SAS)



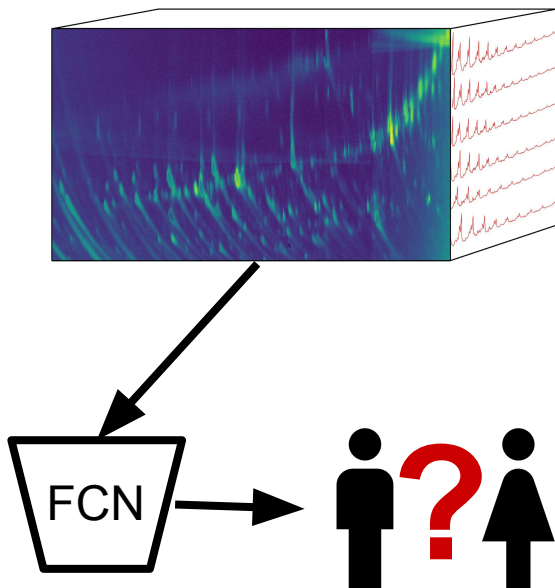
Pipeline

1. registration of the data to a canonical frame
 - detection of 24 chemical compounds followed by linear piecewise transformation, requiring ZERO hours of human work
2. sex classification/identity verification using registered data with convolutional neural networks

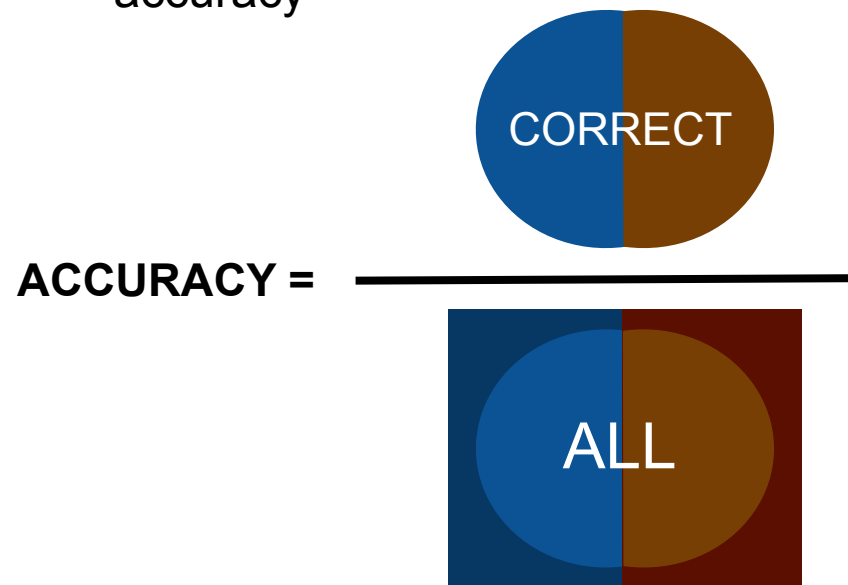


Sex classification

- given SAS or FULL data, predict sex



- error measure - classification accuracy

$$\text{ACCURACY} = \frac{\text{CORRECT}}{\text{ALL}}$$


The equation shows that accuracy is the ratio of correct classifications to the total number of classifications. The 'CORRECT' part is visualized as a circle split into two halves (blue and brown), and the 'ALL' part is visualized as a square containing a larger circle, also split into two halves (blue and brown).

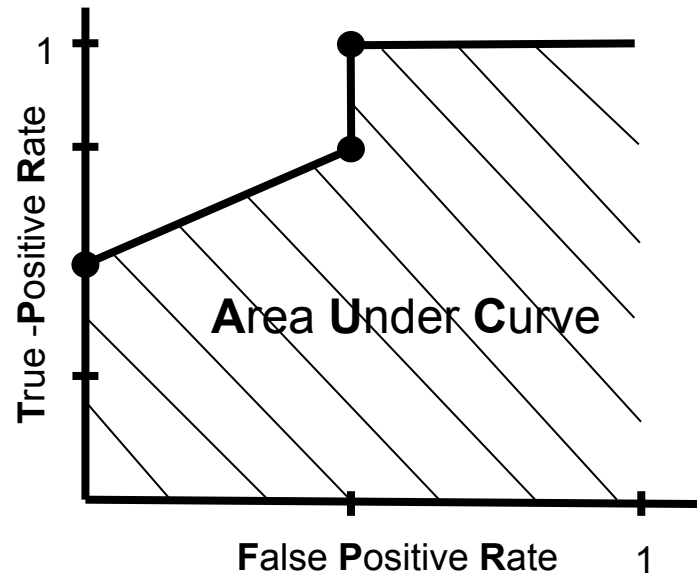
Sex classification results on validation dataset, 10 fold X-val

	Sex Classification Model	Accuracy (%)
SAS	LinearCNN(X)	78.50 ± 9.76
	CNN(X)	75.00 ± 5.48
	LeNet5(X)	84.50 ± 5.22
FULL	MaxPoolNN(X)	91.00 ± 5.83
	PatchNN(X)	94.50 ± 6.10

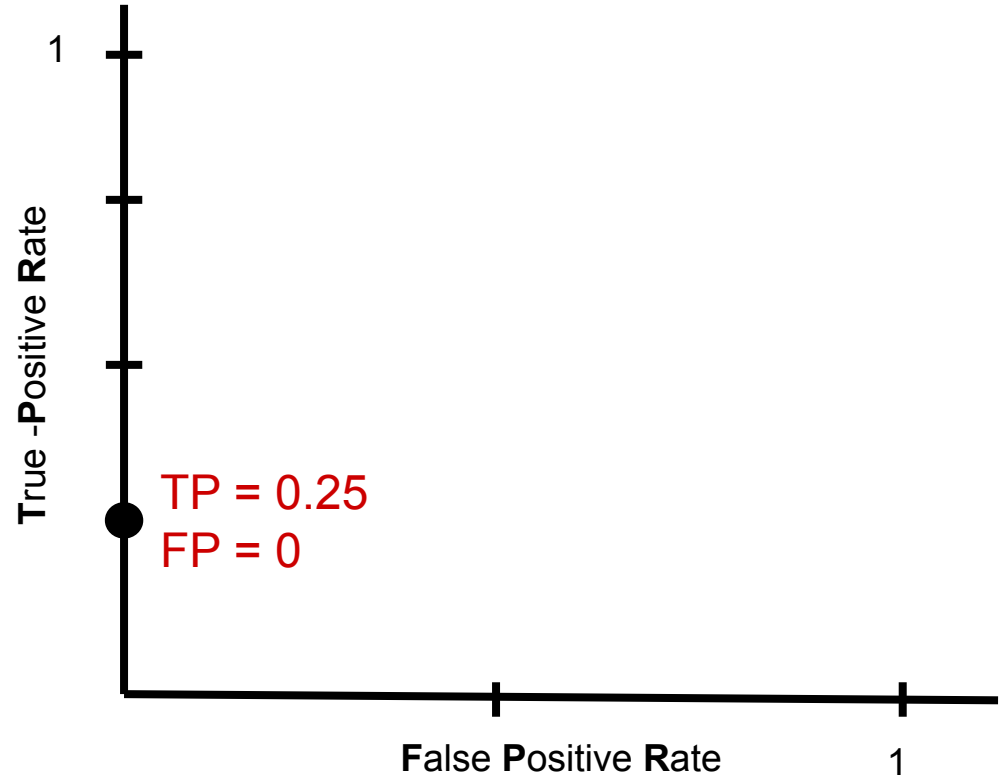
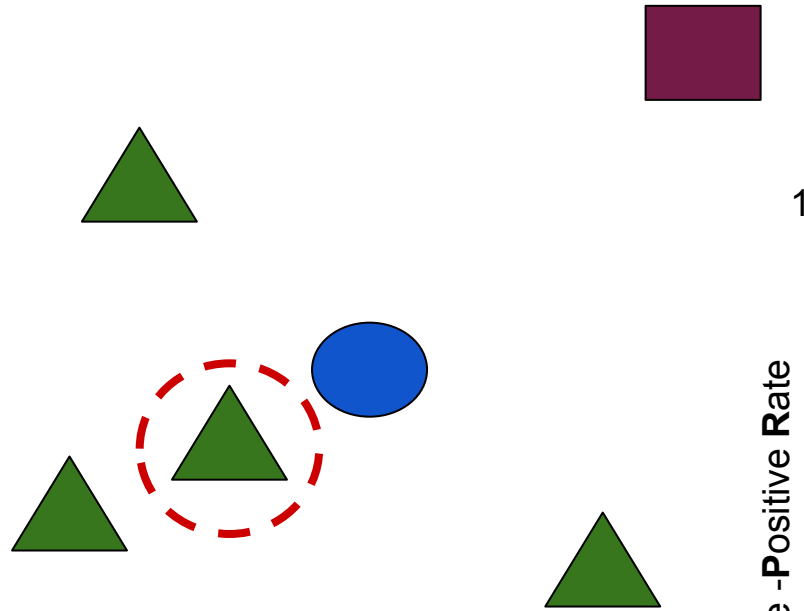


Identity verification

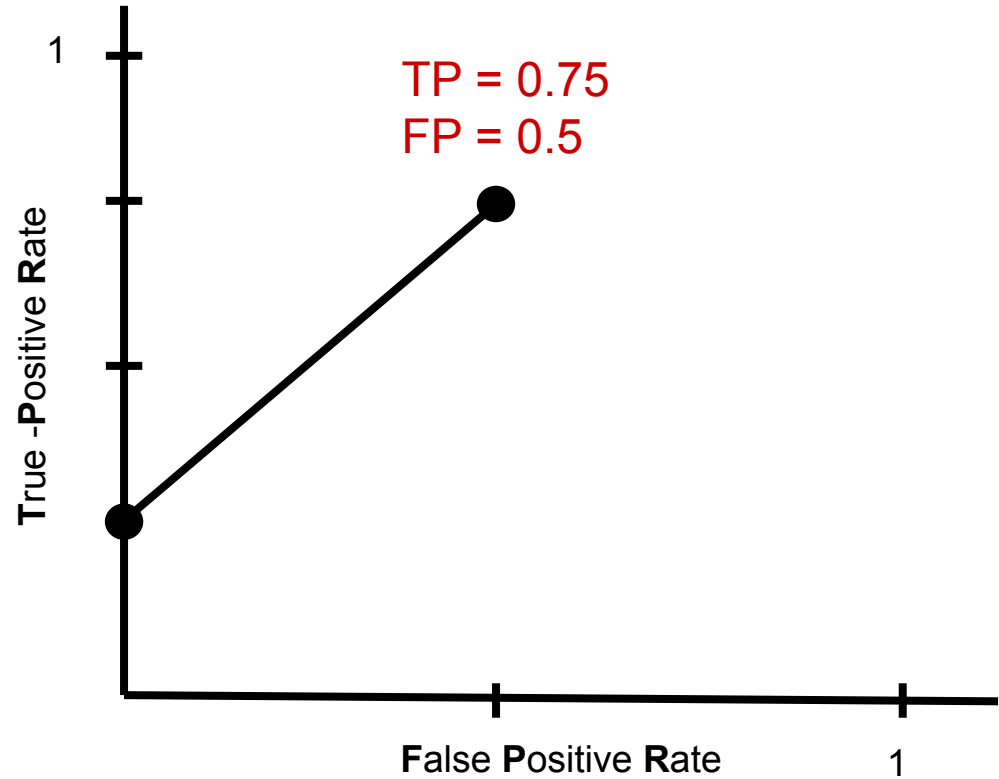
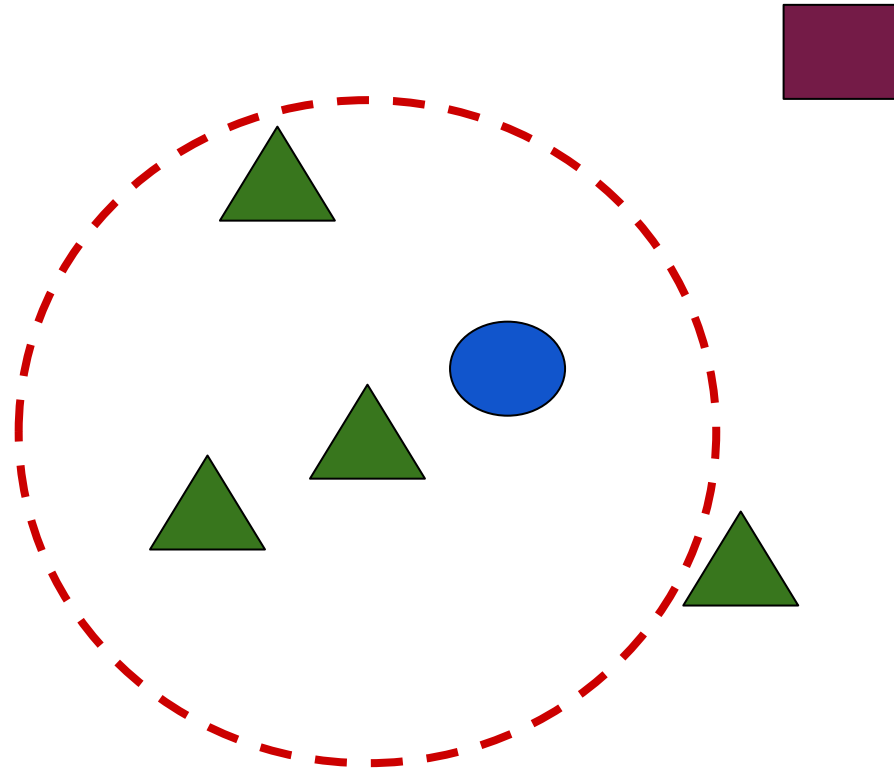
- verify whether two SAS (or two FULL) samples belong to the same identity
- error measure - **Area Under Receiver Operating Characteristic**

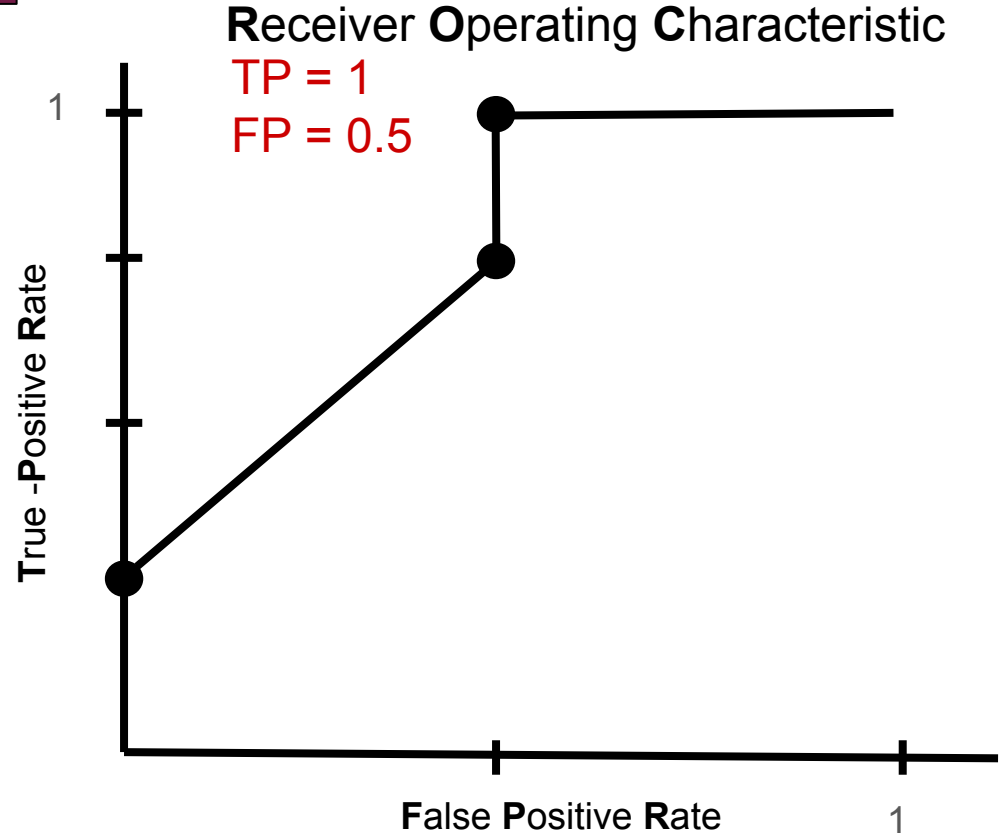
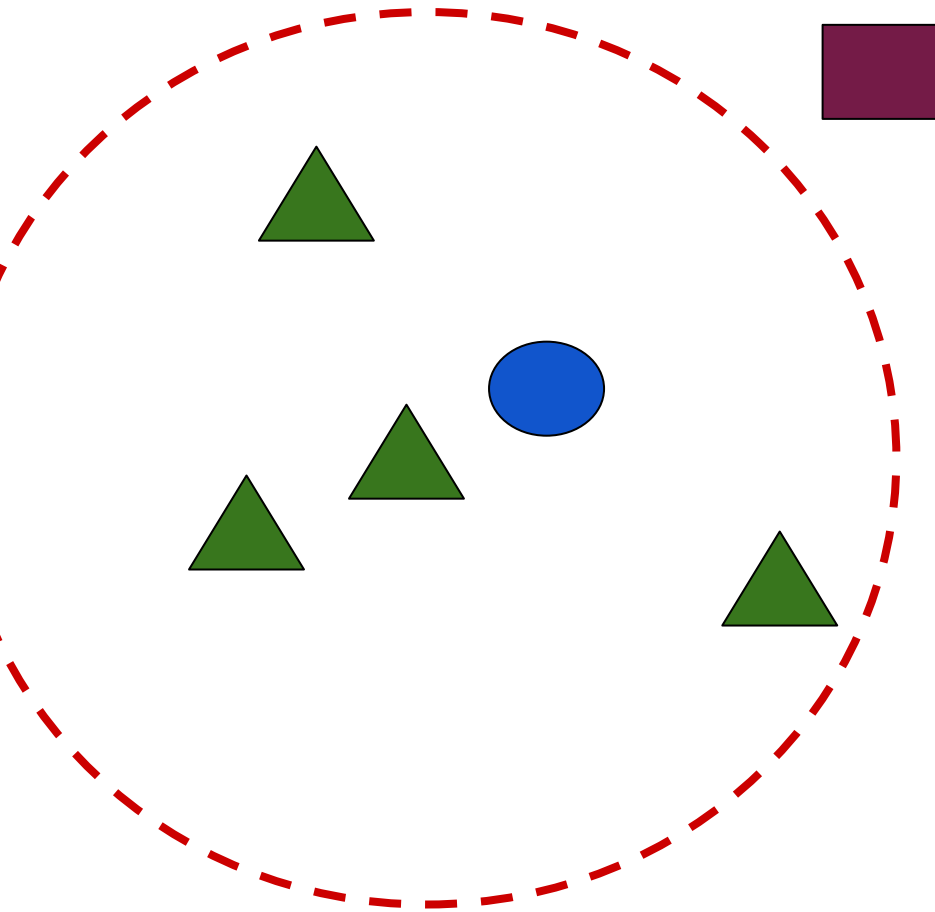


Receiver Operating Characteristic

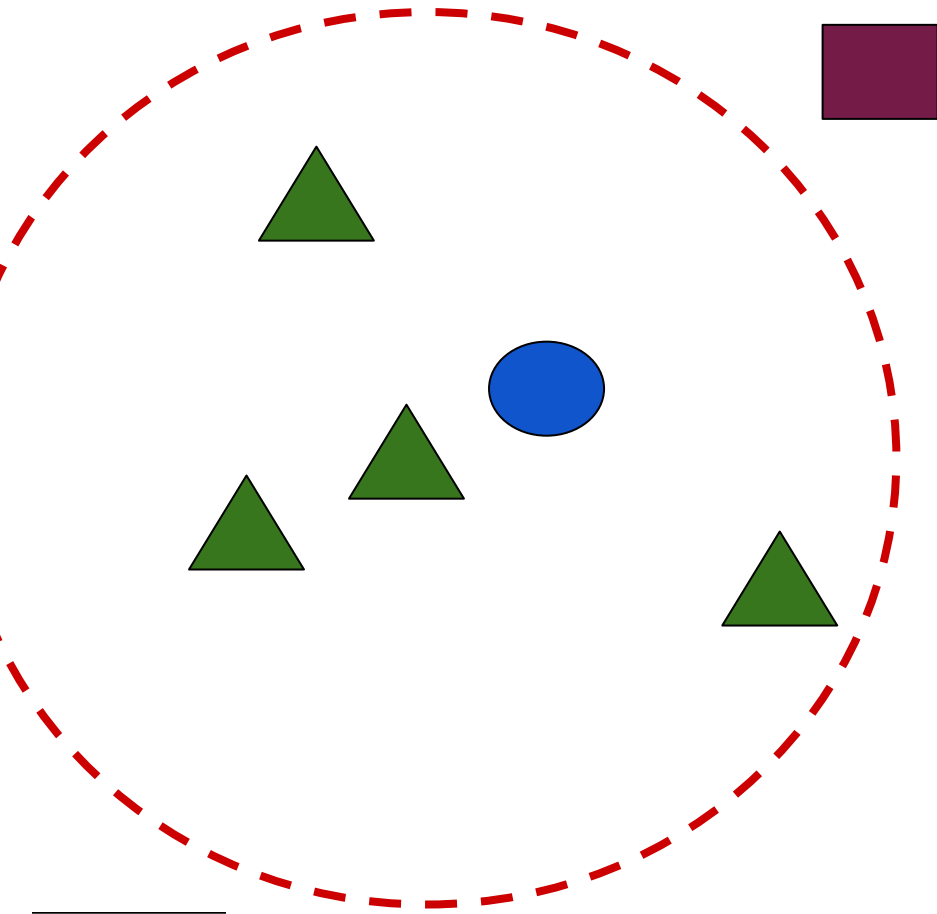


Receiver Operating Characteristic

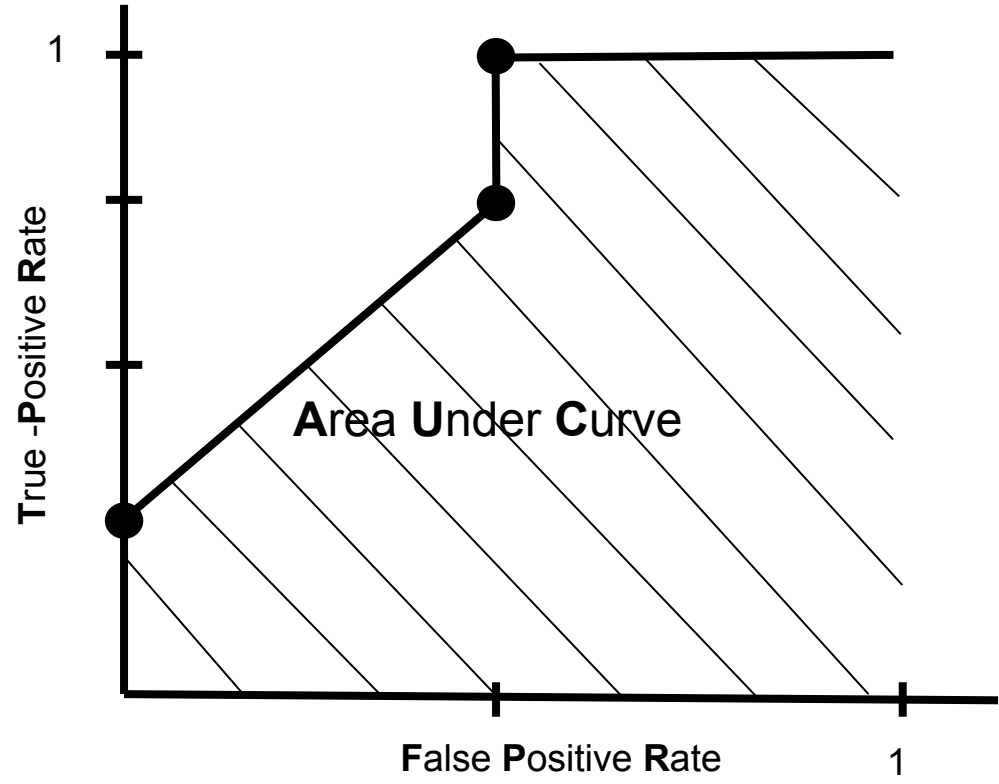




LUMI

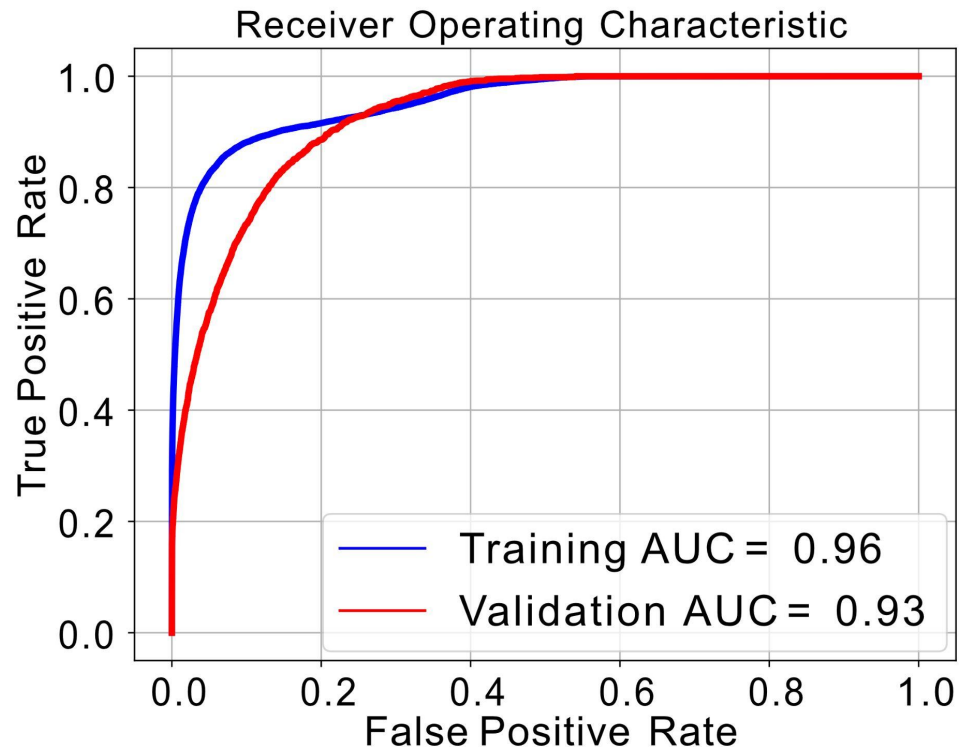


Receiver Operating Characteristic



Identity verification results

- although conceptually simple, computationally very intensive
- The main advantages:
 - zero human work
 - reproducibility
 - open-source



Thank you for your attention

Preprint



Python codes

