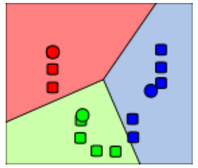
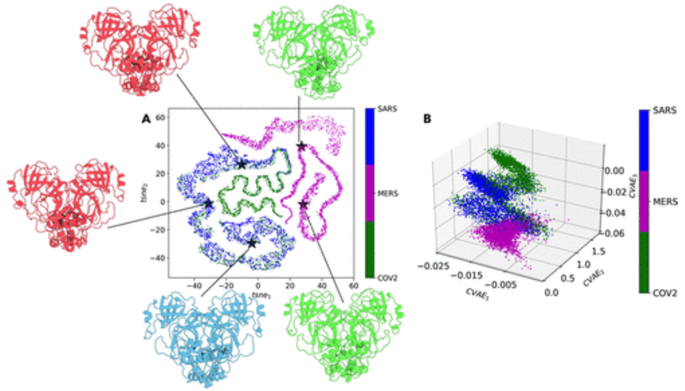
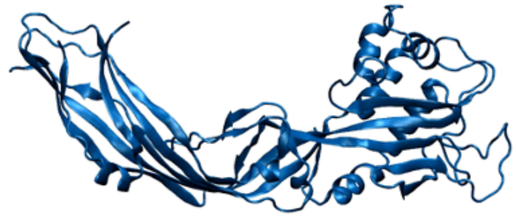
# From biomolecular data to information

Antonia Mey

Matteo Degiacomi





Durham University

matteo.t.degiacomi@dur.ac.uk

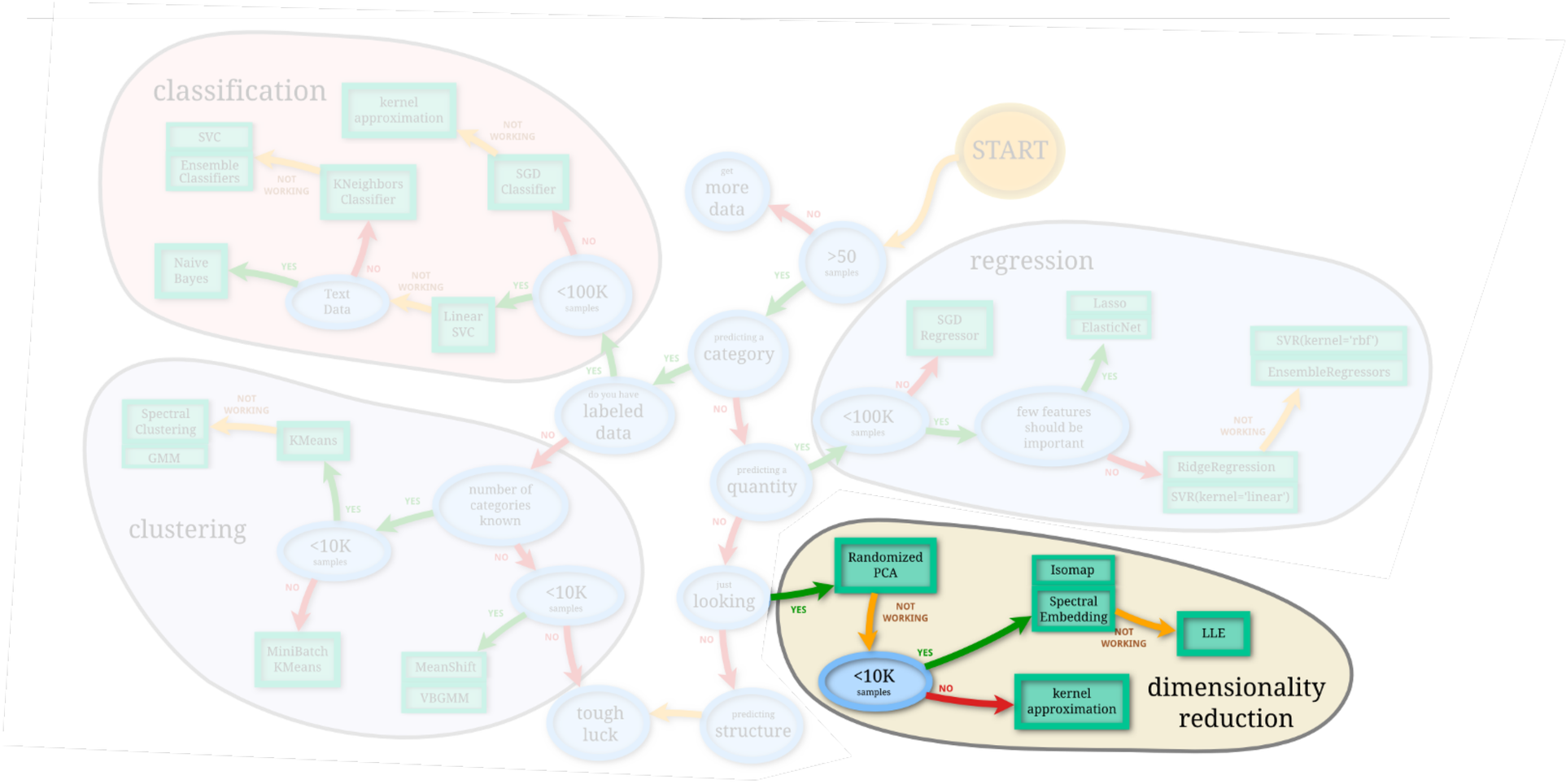
@MatteoDegiacomi

antonia.mey@ed.ac.uk

@ppxasjsm

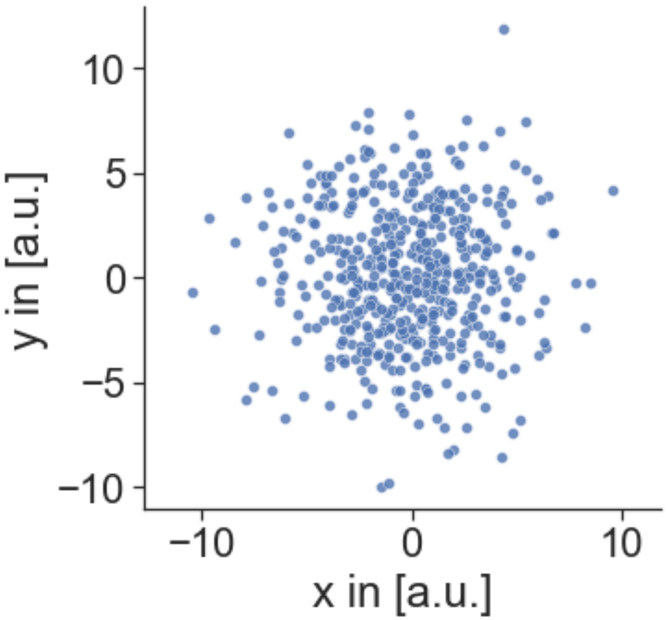
THE UNIVERSITY of EDINBURGH

# The Data Mining World — Dimensionality Reduction

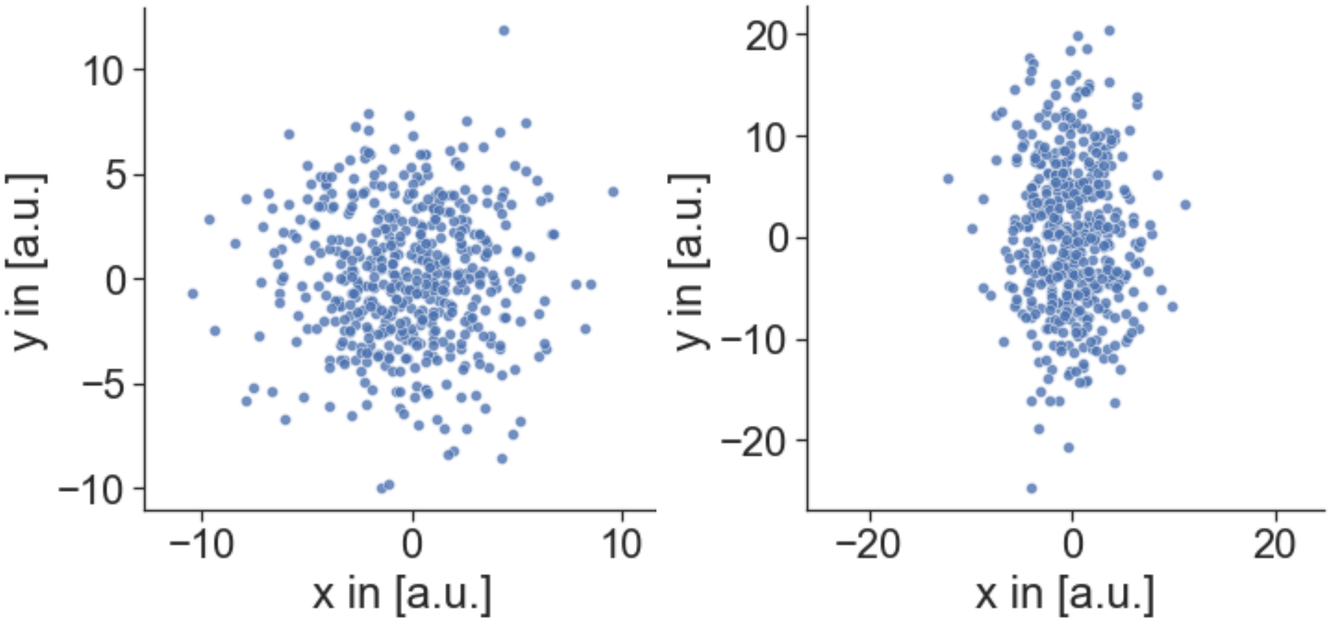



From **scikit-learn.org**

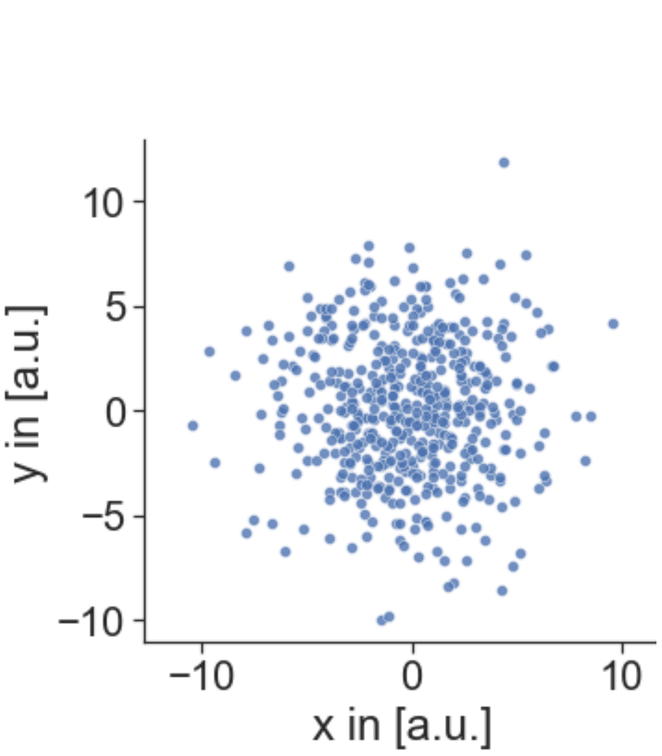
# Principal component analysis (PCA)

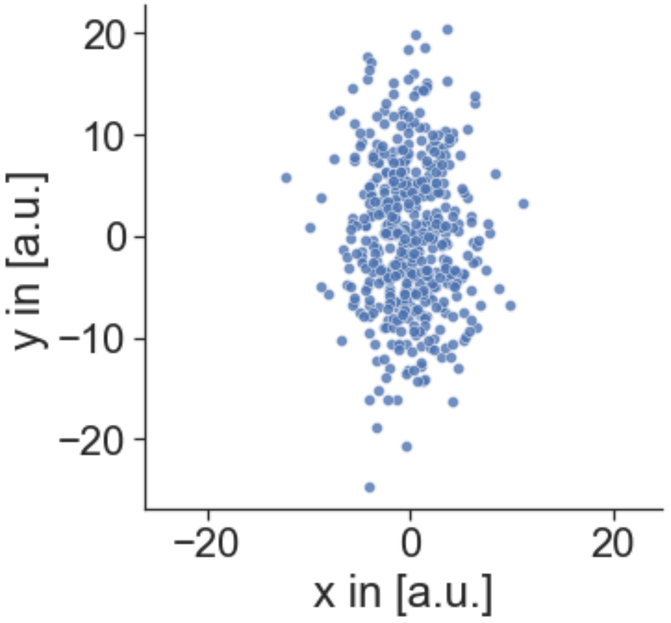


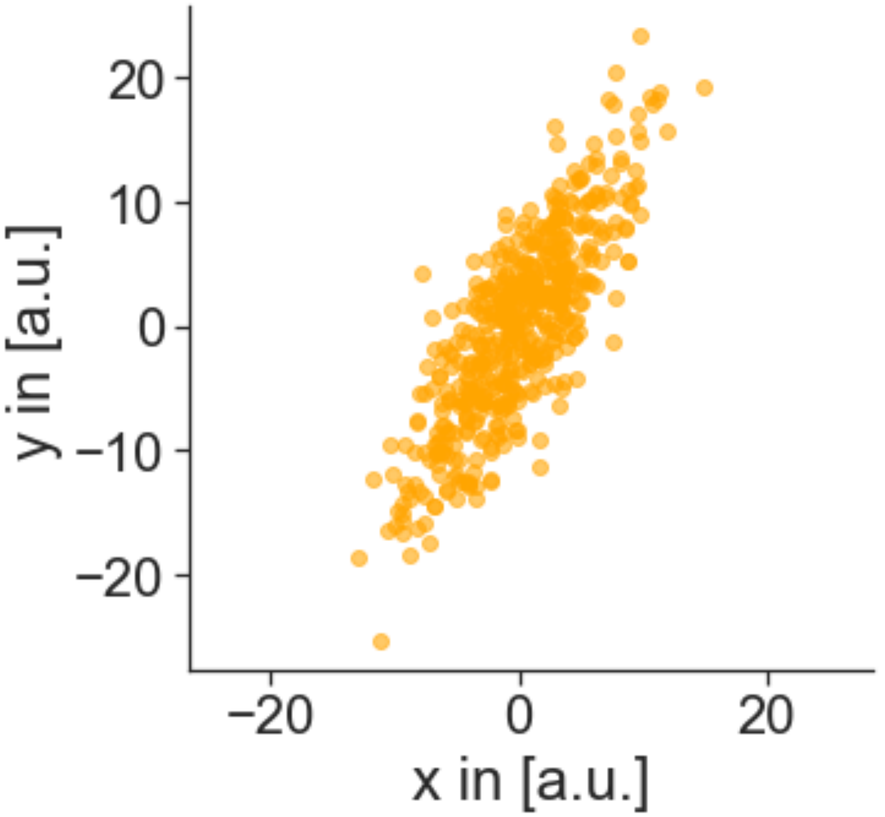
# Principal component analysis (PCA)



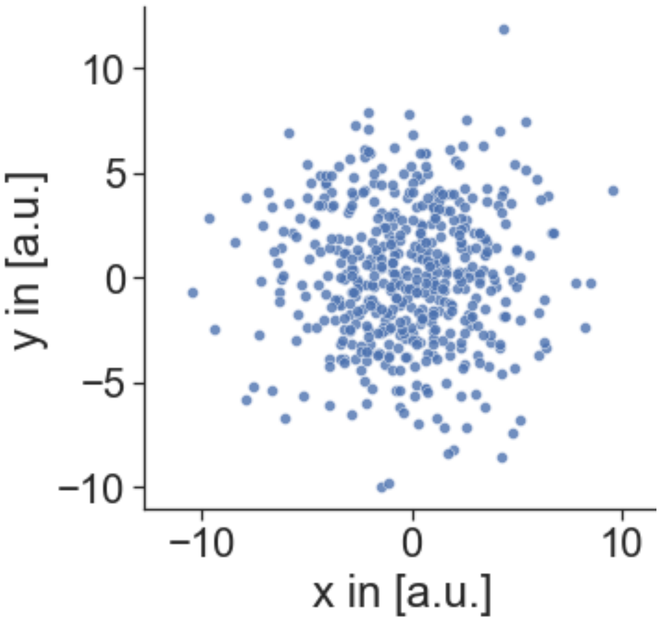
# Principal component analysis (PCA)

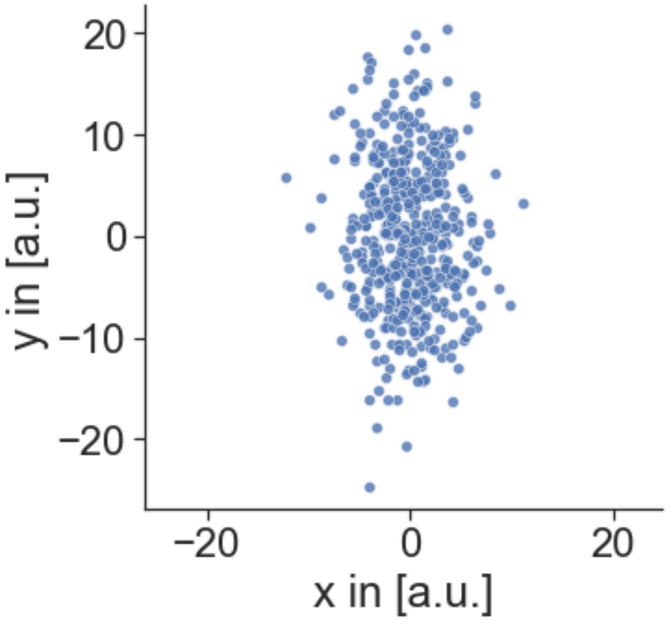


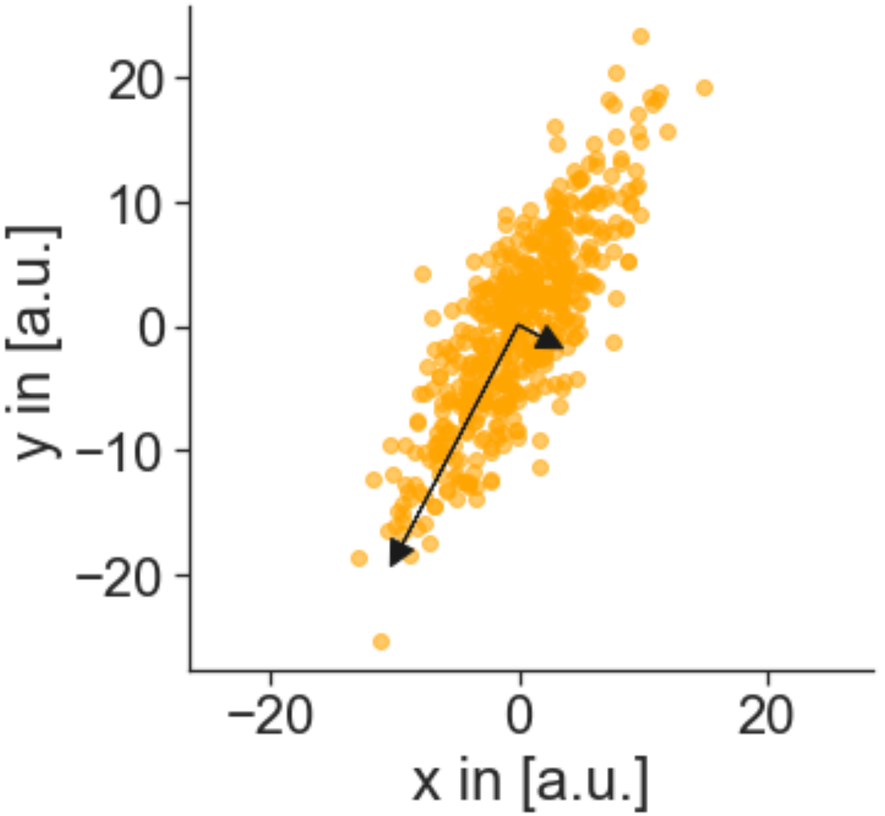



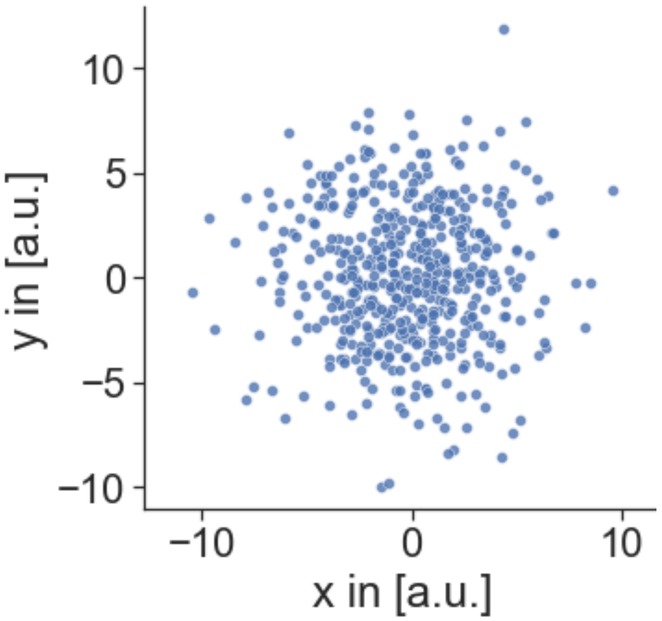
# Principal component analysis (PCA)

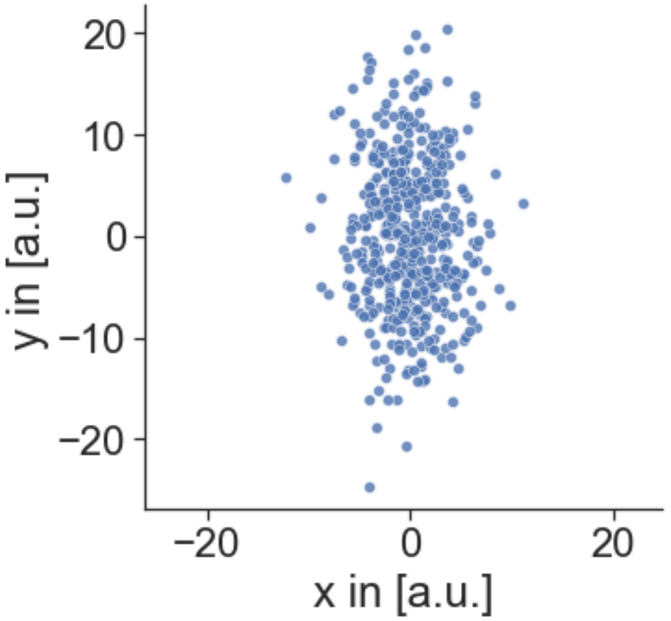


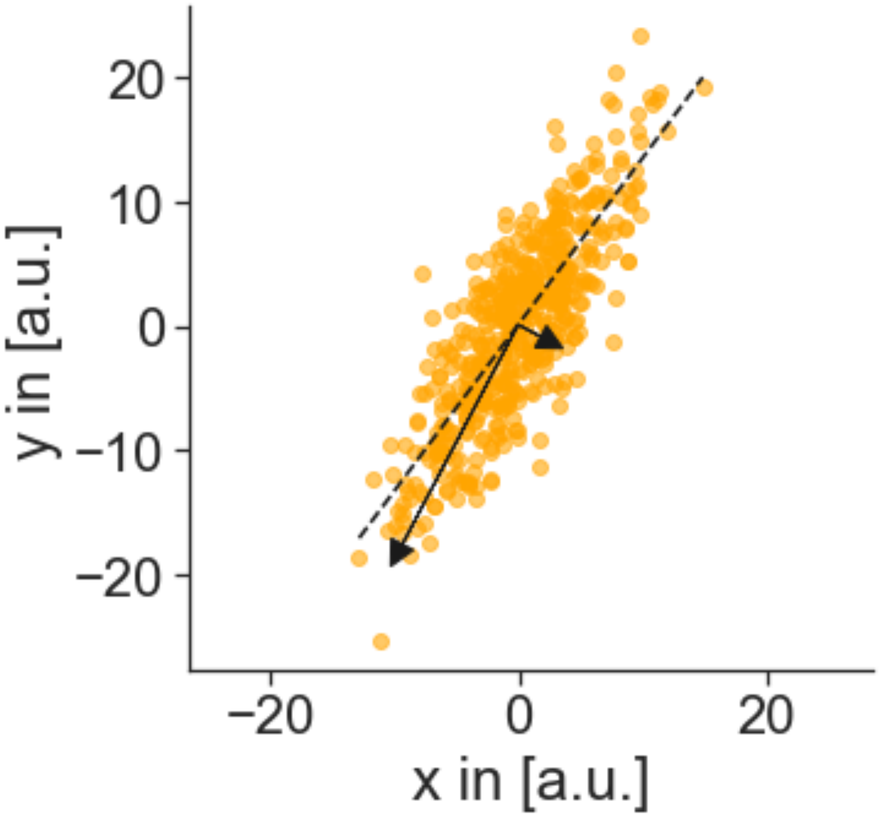



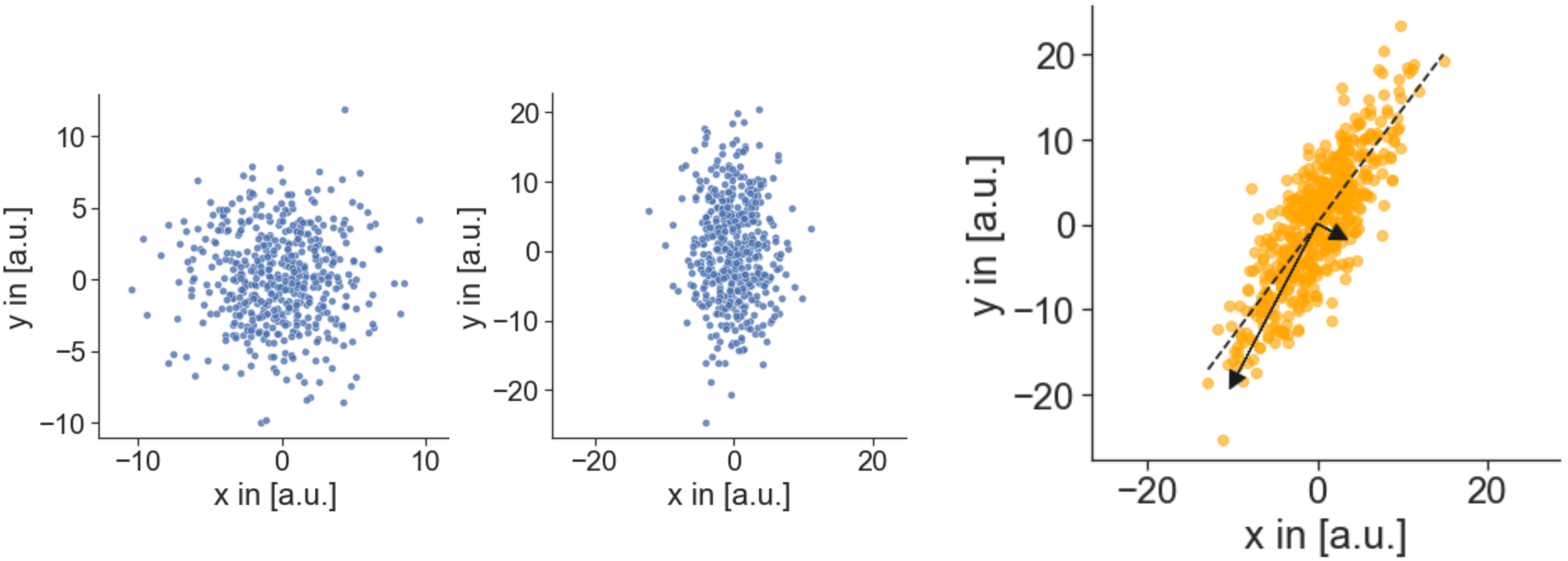
# Principal component analysis (PCA)



# Principal component analysis (PCA)

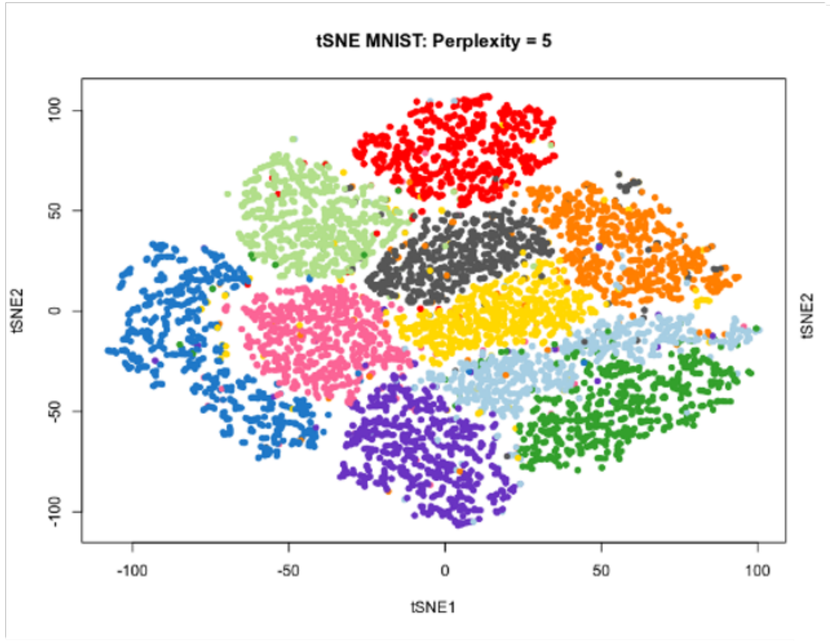


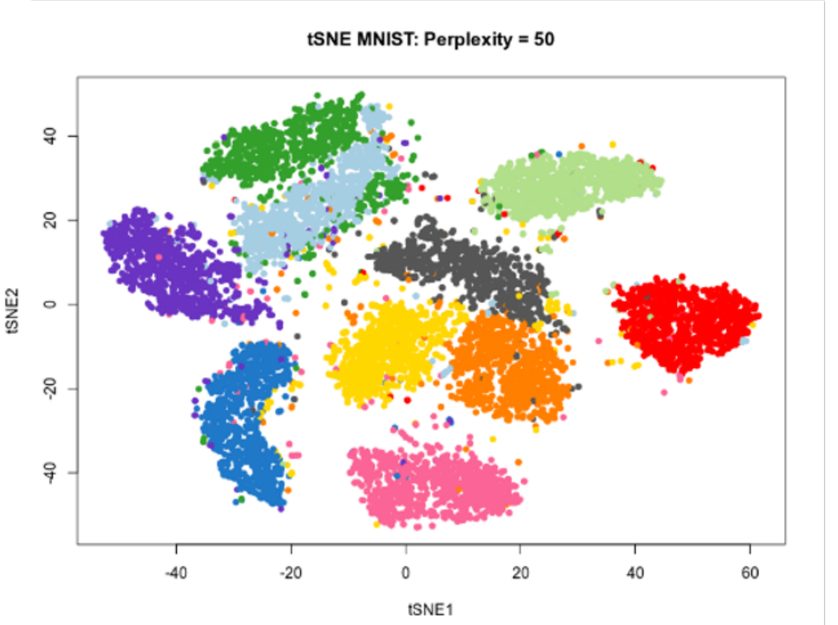

- PCA is an **orthogonal linear transformation** that **maximises the variance** across the first component
- A linear regression fit **minimises the error** with regard to all data points.
- PCA can be used as a tool for **dimensionality reduction**

# T-distributed Stochastic Neighbour Embedding (t-SNE)

- Useful for visualisation, project high-dimensional data in 2 or 3 dimensions.
- Controlled by one main parameters: “perplexity”
- Relative distance between points not quantitatively meaningful

MNIST: database of written digits



# Post-its


# Schedule

## Thursday — Antonia Mey

| | |
|---|---|
| 13:30-14:45 | ML Clustering |
| 14:45-15:00 | ☕ break ☕ |
| 15:00-16:20 | ML Dimensionality Reduction |
| 16:20-16:30 | Closing remarks |
| 18:00-onwards | Informal social event |

## Friday — Matteo Degiacomi

| | |
|---|---|
| 09:30-10:45 | ML Classification |
| 10:45-11:00 | ☕ break ☕ |
| 11:00-12:30 | ML Regressions and Neural Networks |
| 12:30-13:30 | Lunch |
| 13:30-onwards | Bring your own problem |