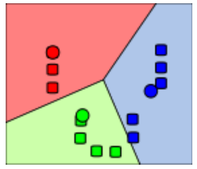
From biomolecular data to information

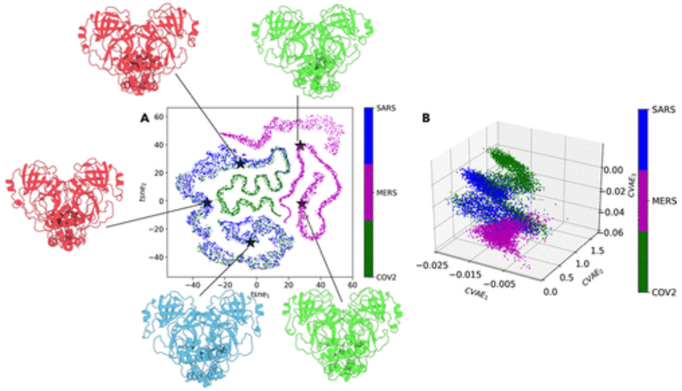
CCP5 Summer School @ University of Durham 26-27 July 2022

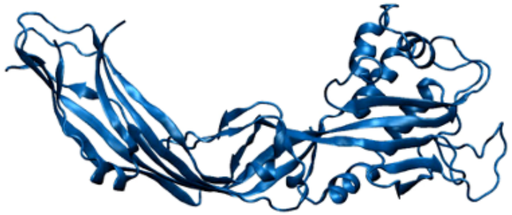
Micaela Matta
micaela.matta@kcl.ac.uk
@micaelamatta

Antonia Mey
antonia.mey@ed.ac.uk
@ppxasjsm

Matteo Degiacomi
matteo.t.degiacomi@dur.ac.uk
@MatteoDegiacomi







Schedule

Morning

09:00-11:00	Dimensionality Reduction theory and toy examples (TM)
11:00-11:30	☕ break ☕
11:30-12:30	ML Dimensionality Reduction application to protein simulations (MD)

Afternoon

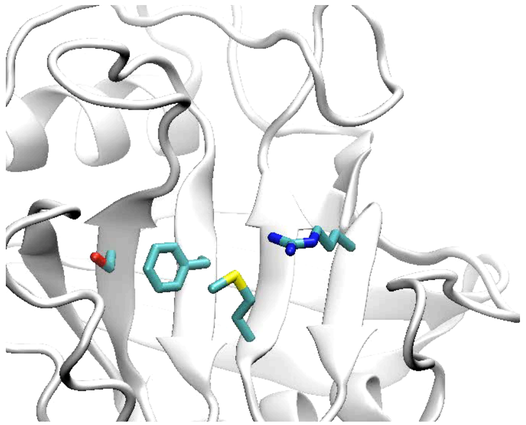
14:00-14:30	Clustering Theory (MD)
14:30 - 15:30	Clustering in practice (TM)
15:30 - 16:00	☕ break ☕
16:00 - 17:00	Classification problems (MD)

TM — Toni Mey

MD — Matteo Degiacomi

Molecular simulations of biomolecules are high dimensional

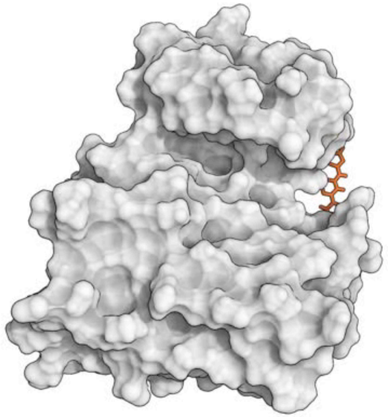
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

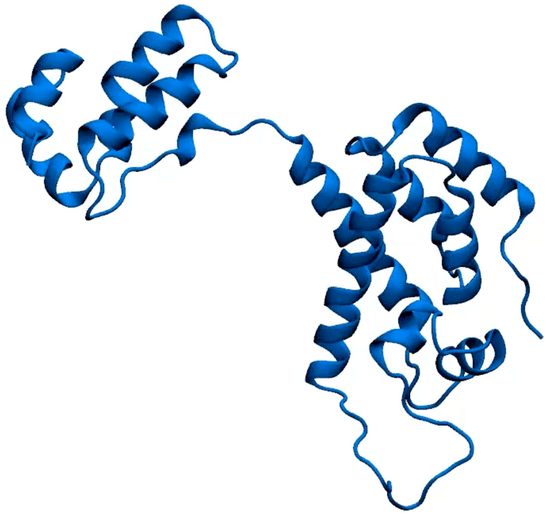
Tyrosine kinase —dasatanib



Binding affinities

Shan Y, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

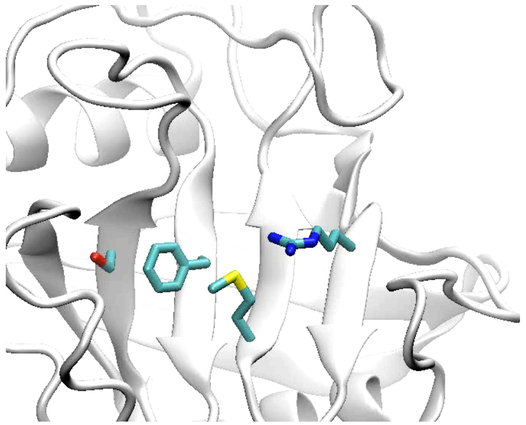


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Molecular simulations of biomolecules are high dimensional

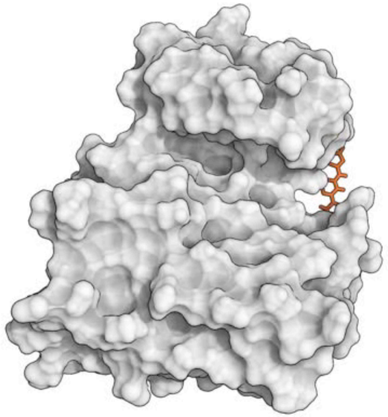
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

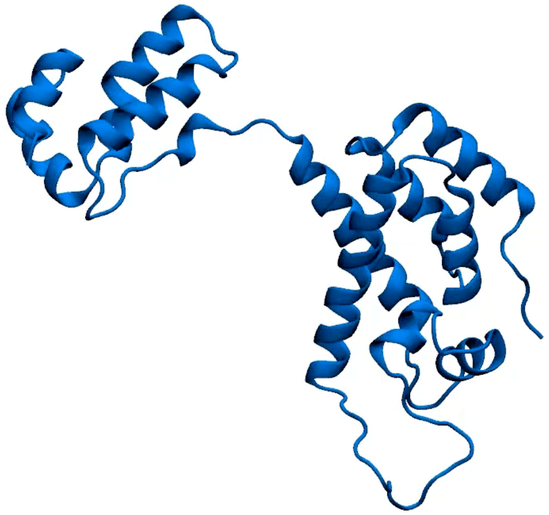
Tyrosine kinase —dasatanib



Binding affinities

Shan Y, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

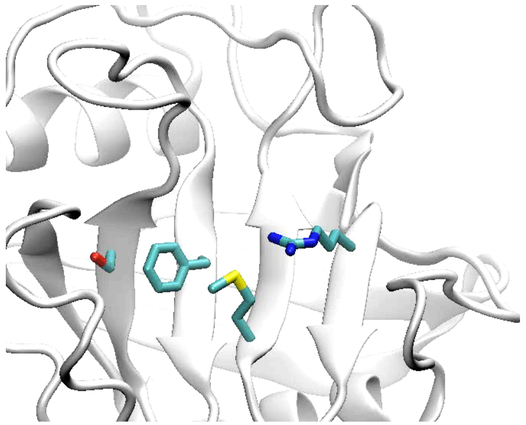


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Molecular simulations of biomolecules are high dimensional

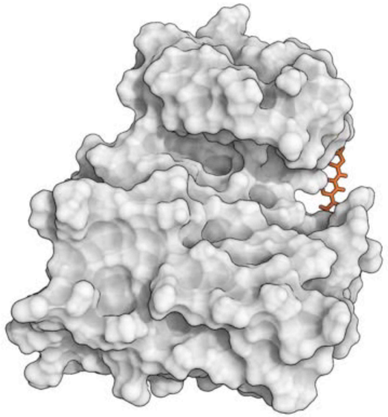
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

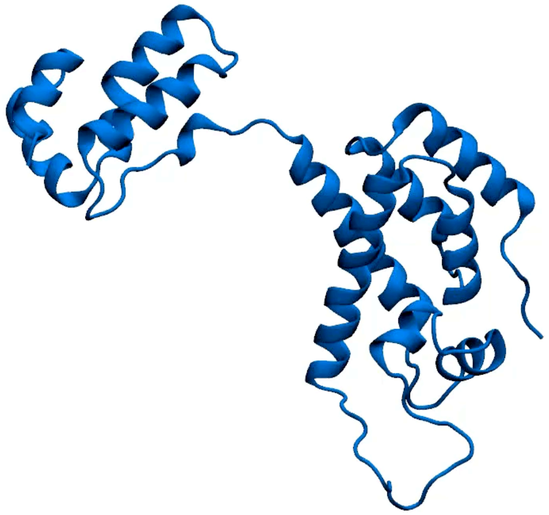
Tyrosine kinase —dasatanib



Binding affinities

Shan Y, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

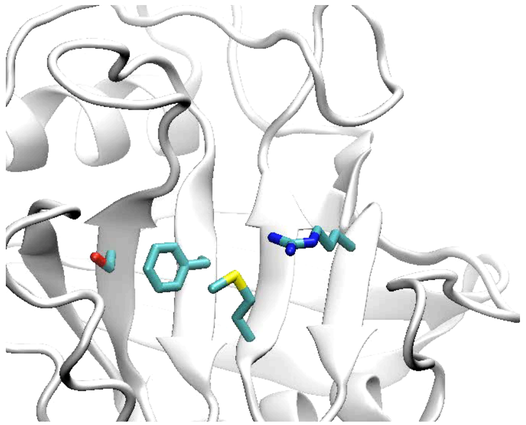


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Molecular simulations of biomolecules are high dimensional

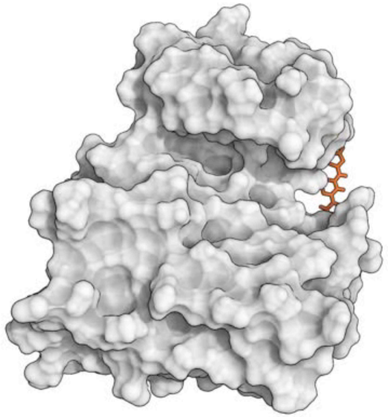
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

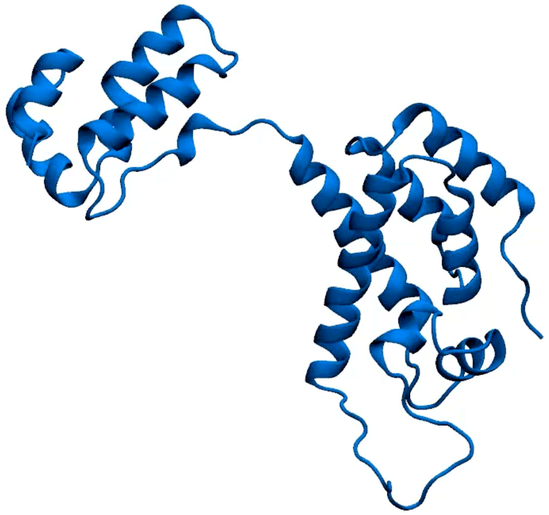
Tyrosine kinase —dasatanib



Binding affinities

Shan Y, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

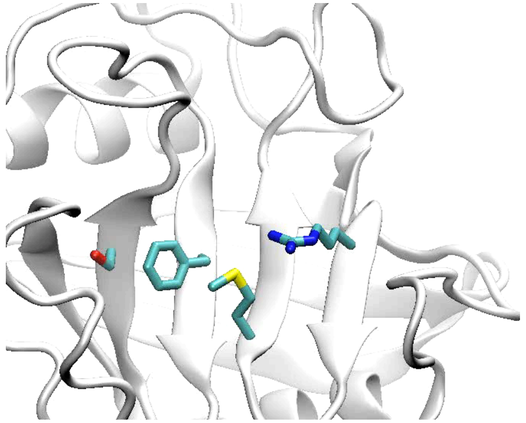


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Molecular simulations of biomolecules are high dimensional

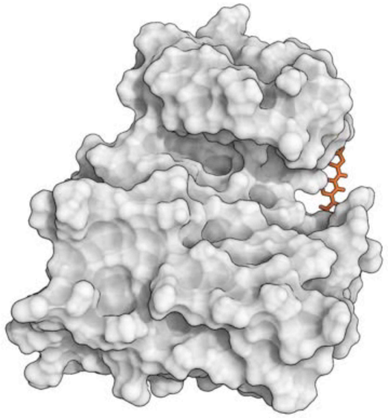
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

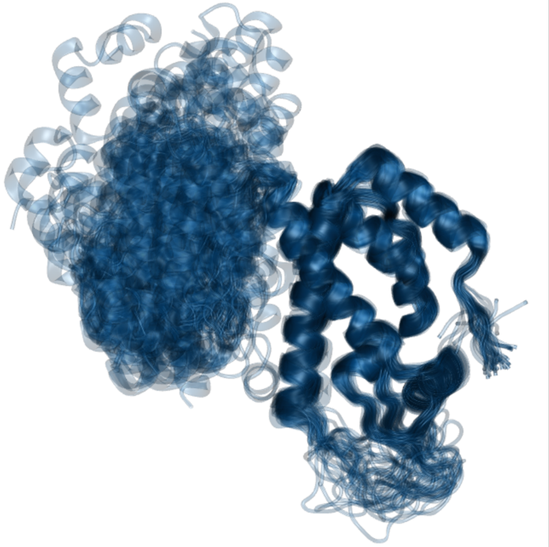
Tyrosine kinase —dasatanib



Binding affinities

Shan Y, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

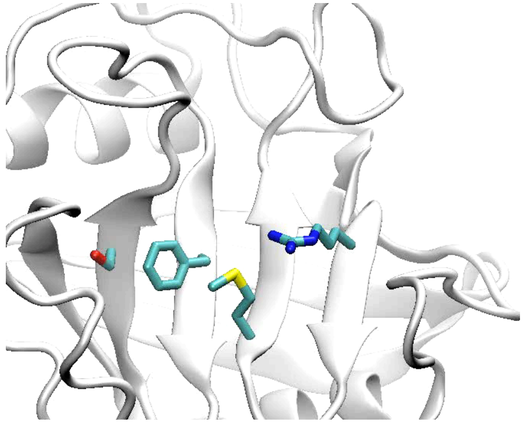


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Molecular simulations of biomolecules are high dimensional

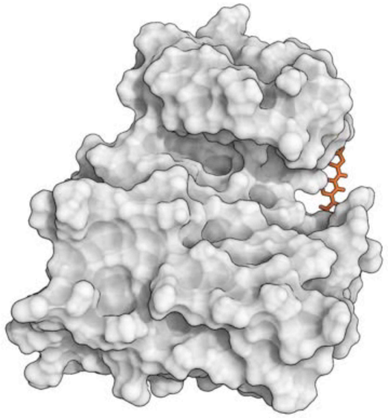
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

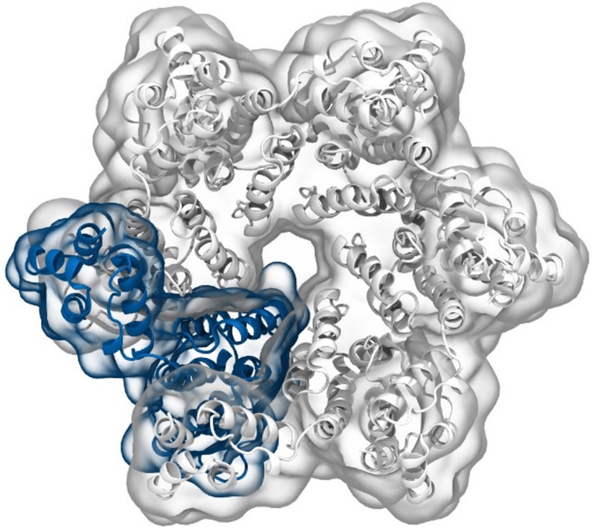
Tyrosine kinase —dasatanib



Binding affinities

Shan Y, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

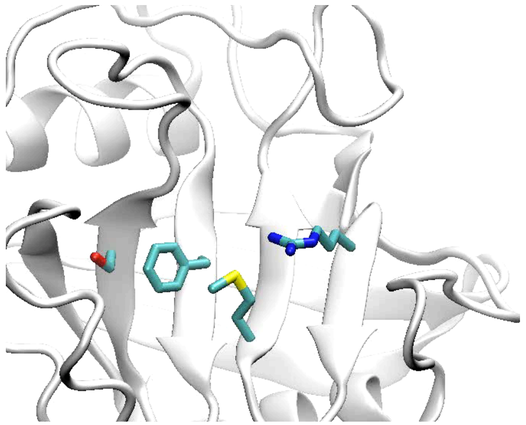


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Molecular simulations of biomolecules are high dimensional

Cyclophilin



Catalysis

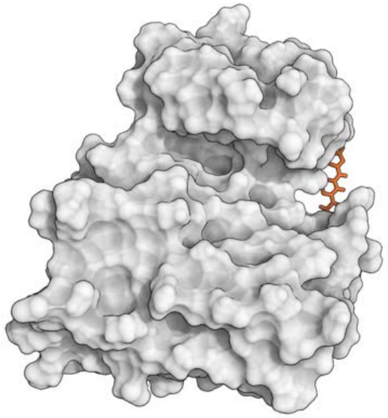
Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

45±1%
1±1%

55±1%
99±1%

Populations

Tyrosine kinase —dasatanib



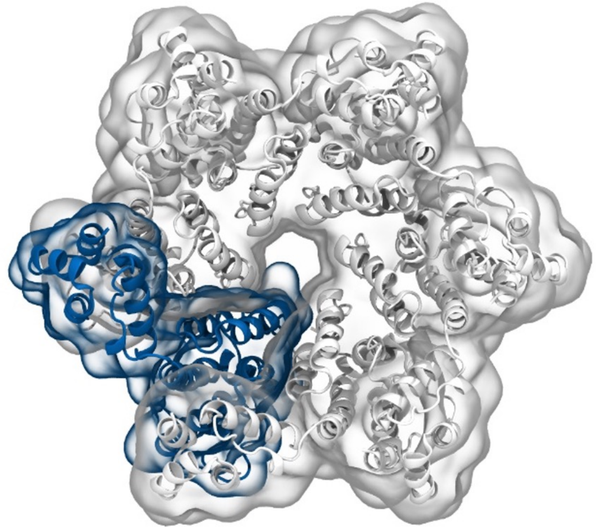
Binding affinities

Shan Y, et al. *JACS* **133,** 9181 (2011)

ΔG k_{on}

Free energies / rates

HIV Capsomer



Assembly

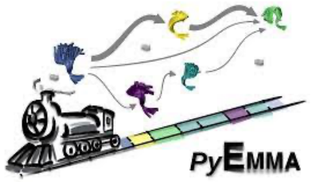
Degiacomi, *Structure* **27**, 1034 (2019)

ΔG

Free energies

Python based tools for analysis MD simulations

Ambertools



Modeller

propKa

Packmol

What is machine learning?

Artificial intelligence

Design an intelligent agent that perceives its environment and makes decisions to maximise chances of achieving its goal.

Machine learning

Gives computers the ability to learn without specifically being programmed (Arthur Samuel 1959)

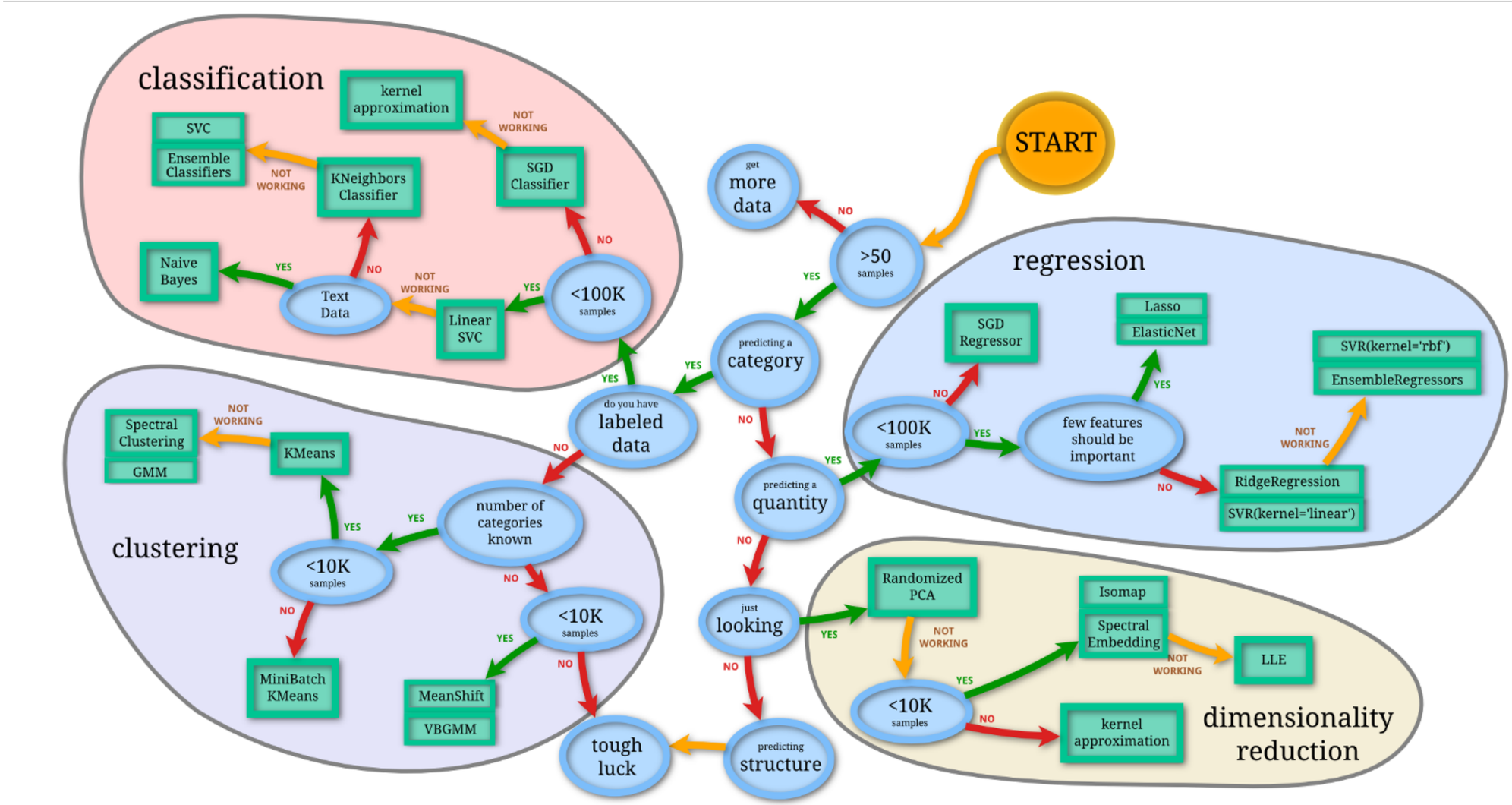
Supervised learning

Unsupervised learning

reinforcement learning

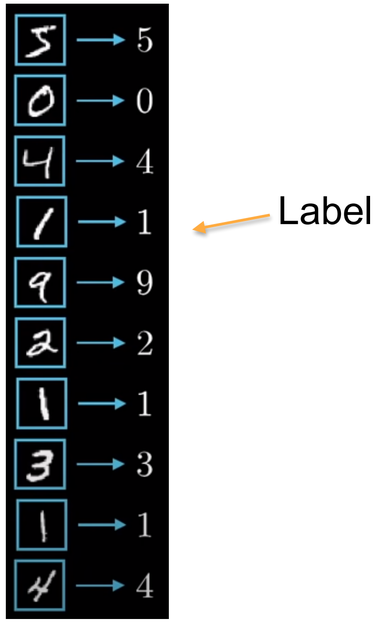
https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12

The Data Mining World

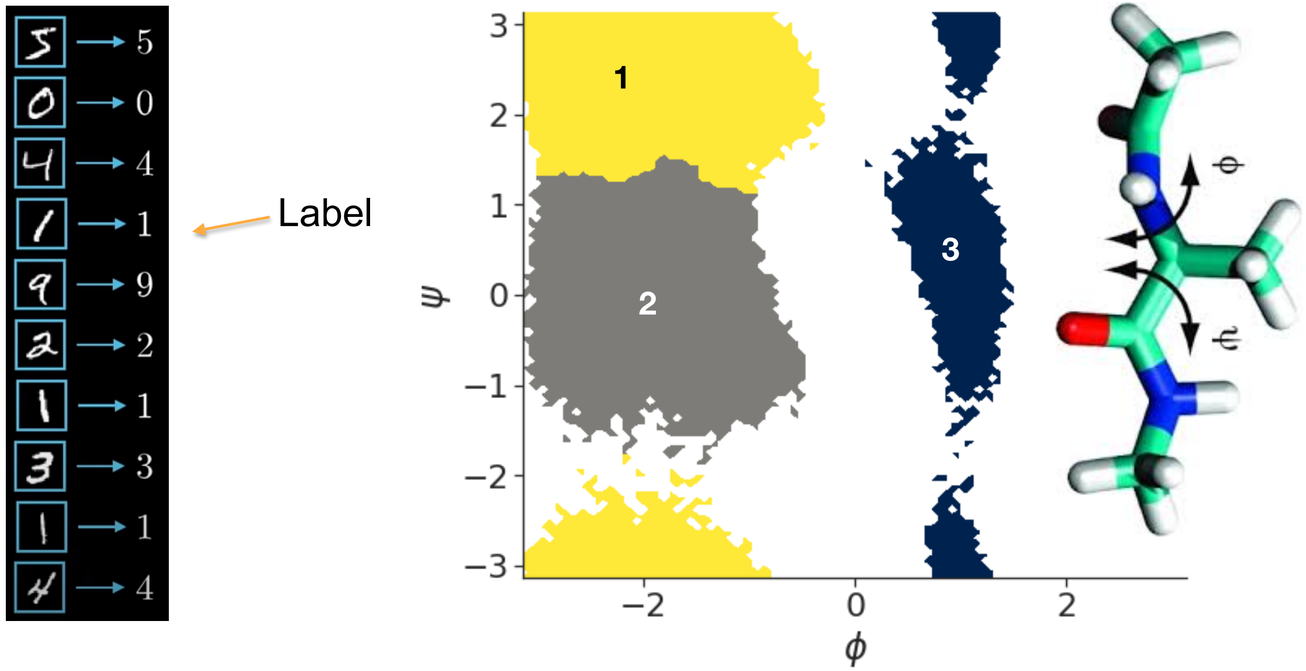


From **scikit-learn.org**

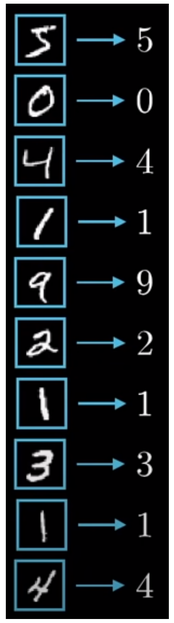
Labels are needed for supervised learning tasks



Labels are needed for supervised learning tasks

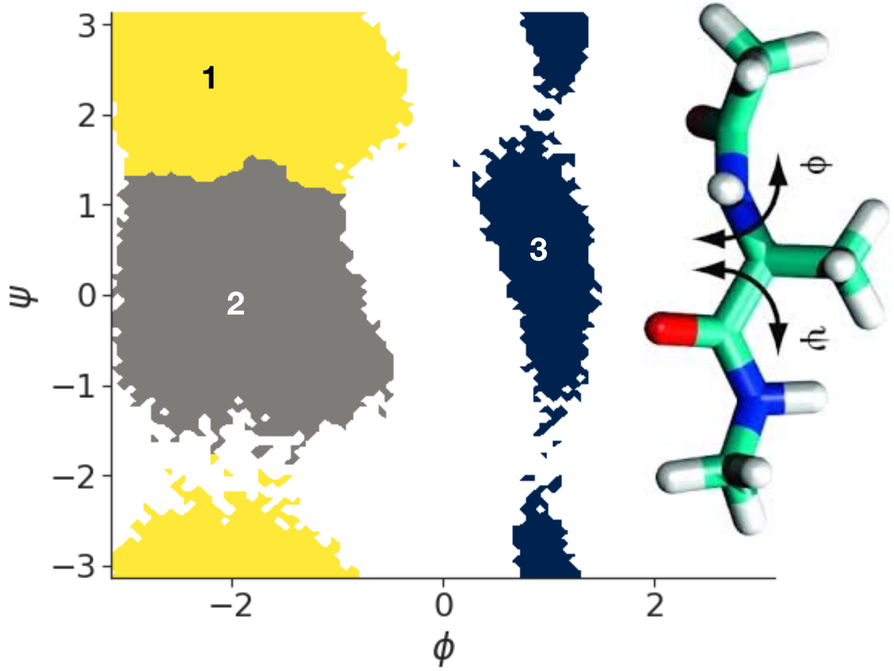


Labels are needed for supervised learning tasks

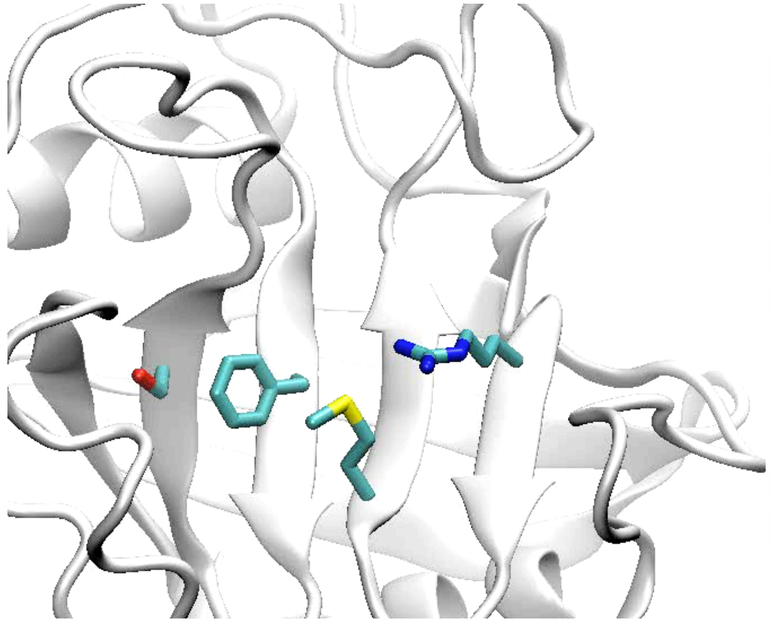


Label

Labelled states



Labelled energies



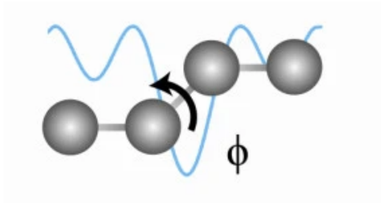
[-12008 kJ]
[-12078 kJ]
[-12045 kJ]
[-12083 kJ]
[-12062 kJ]
[-12058 kJ]
.
.
.
[-12099 kJ]
[-12093 kJ]

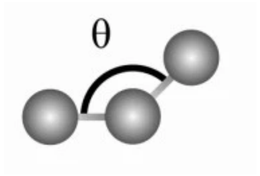
Features are possible representation of data as input

Features from coordinates

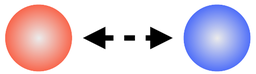
Charge

Dihedral





Angle



Distances

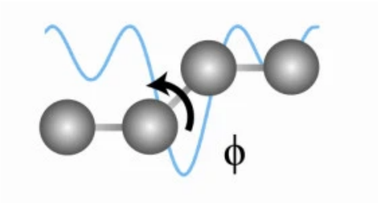
Bonds

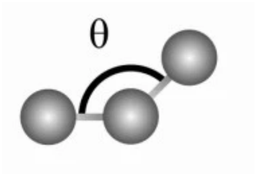
Features are possible representation of data as input

Features from coordinates

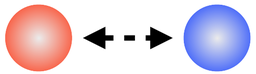
Charge

Dihedral





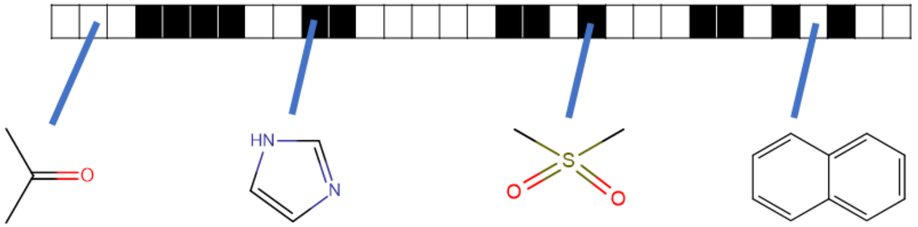
Angle

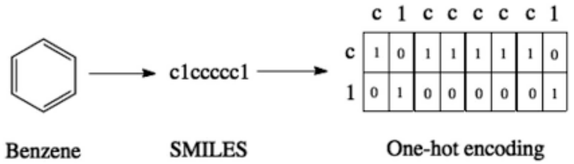


Distances

Bonds

Other features



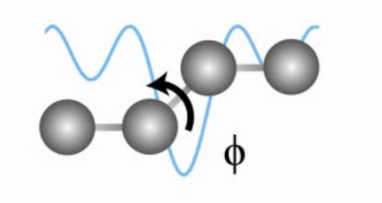


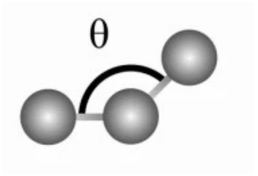
Smiles string

Features are possible representation of data as input

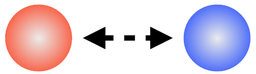
Features from coordinates

Charge





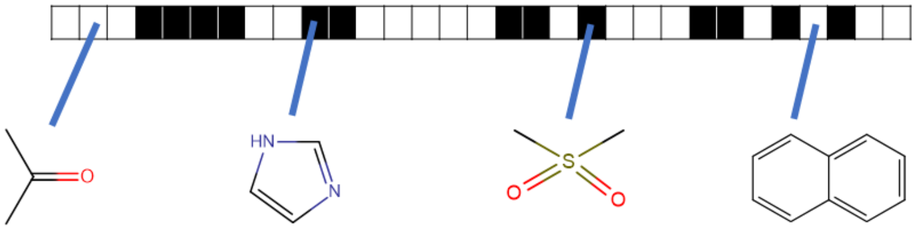
Angle

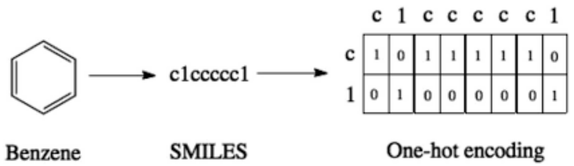


Distances

Bonds

Other features





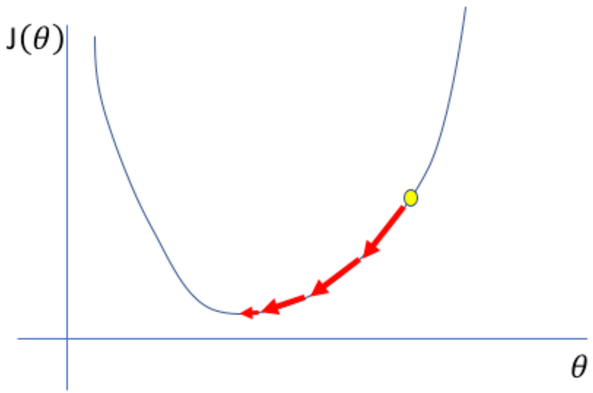
Smiles string

Feature vectors

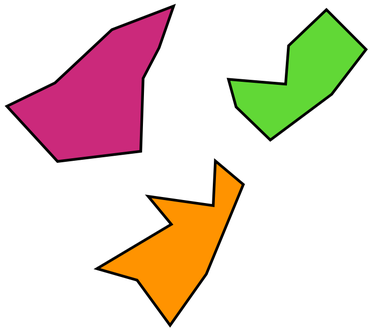
```
['ATOM:ACE 1 CH3 1 x',
 'ATOM:ACE 1 CH3 1 y',
 'ATOM:ACE 1 CH3 1 z',
 'ATOM:ACE 1 C 4 x',
 'ATOM:ACE 1 C 4 y',
 'ATOM:ACE 1 C 4 z',
 'ATOM:ACE 1 O 5 x',
 'ATOM:ACE 1 O 5 y',
 'ATOM:ACE 1 O 5 z',
 'ATOM:ALA 2 N 6 x',
 'ATOM:ALA 2 N 6 y',
 'ATOM:ALA 2 N 6 z',
 'ATOM:ALA 2 CA 8 x',
 'ATOM:ALA 2 CA 8 y',
 'ATOM:ALA 2 CA 8 z',
 'ATOM:ALA 2 CB 10 x',
 'ATOM:ALA 2 CB 10 y',
 'ATOM:ALA 2 CB 10 z',
 'ATOM:ALA 2 C 14 x',
 'ATOM:ALA 2 C 14 y',
 'ATOM:ALA 2 C 14 z',
 'ATOM:ALA 2 O 15 x',
 'ATOM:ALA 2 O 15 y',
 'ATOM:ALA 2 O 15 z',
 'ATOM:NME 3 N 16 x',
 'ATOM:NME 3 N 16 y',
 'ATOM:NME 3 N 16 z',
 'ATOM:NME 3 C 18 x',
 'ATOM:NME 3 C 18 y',
 'ATOM:NME 3 C 18 z']
```

Hyper parameters open up an array of modelling choices

Learning rate

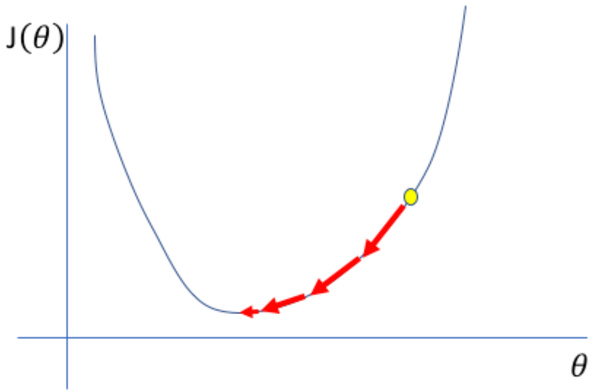


Hyper parameters open up an array of modelling choices

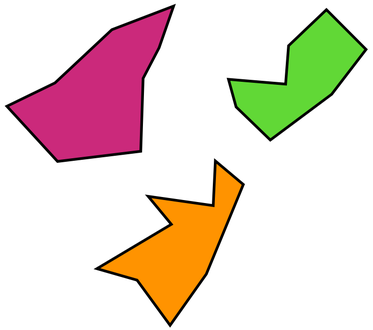


N-clusters

Learning rate



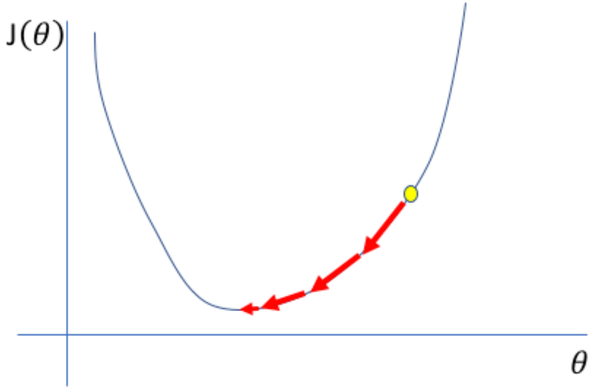
Hyper parameters open up an array of modelling choices



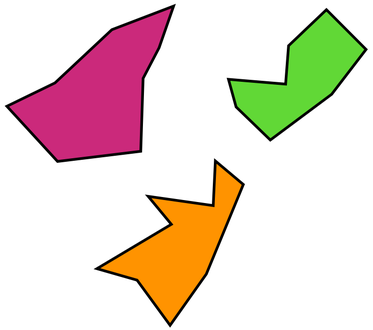
N-clusters

How many dimensions for reduction

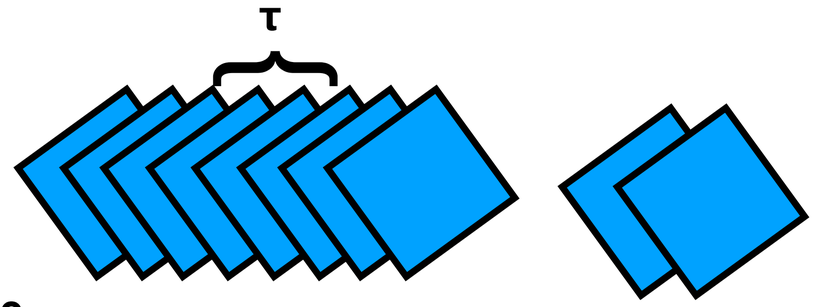
Learning rate



Hyper parameters open up an array of modelling choices

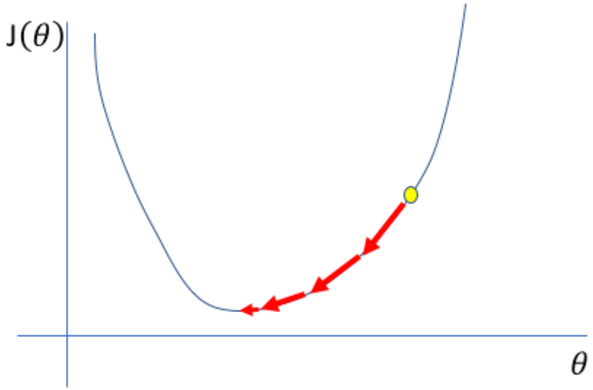


N-clusters



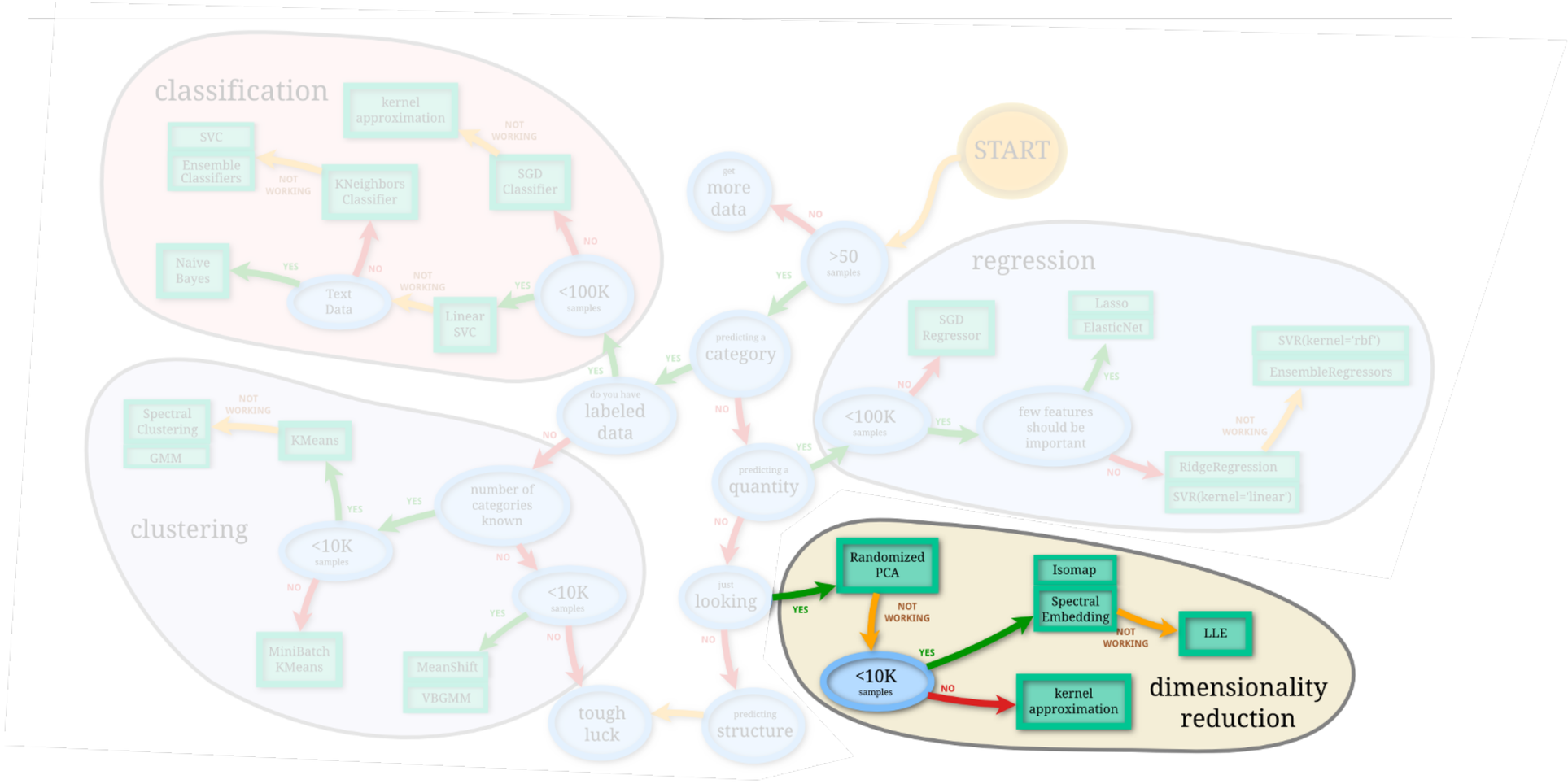
Striding through your trajectory

Learning rate



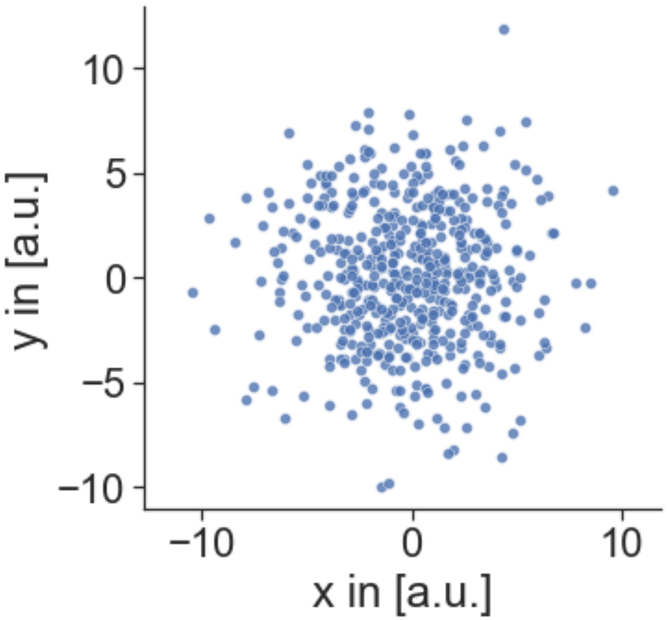
How many dimensions for reduction

The Data Mining World — Dimensionality Reduction

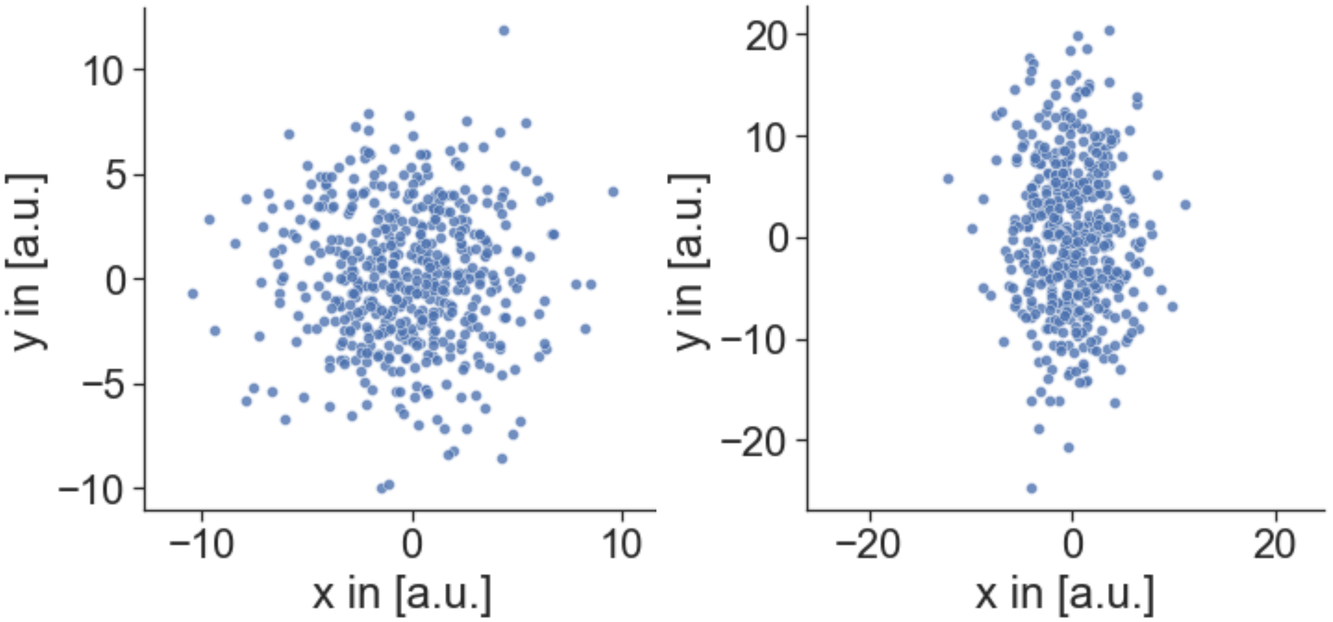


From **scikit-learn.org**

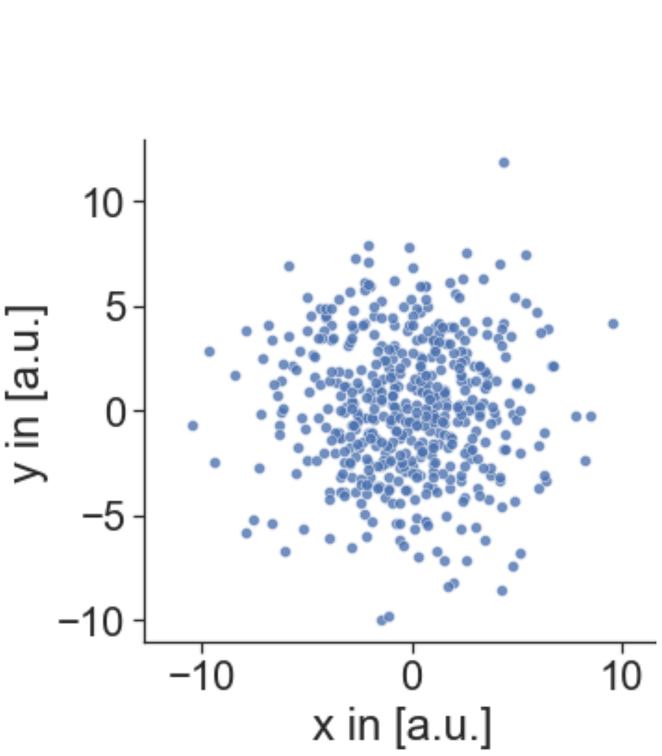
Principal component analysis (PCA)

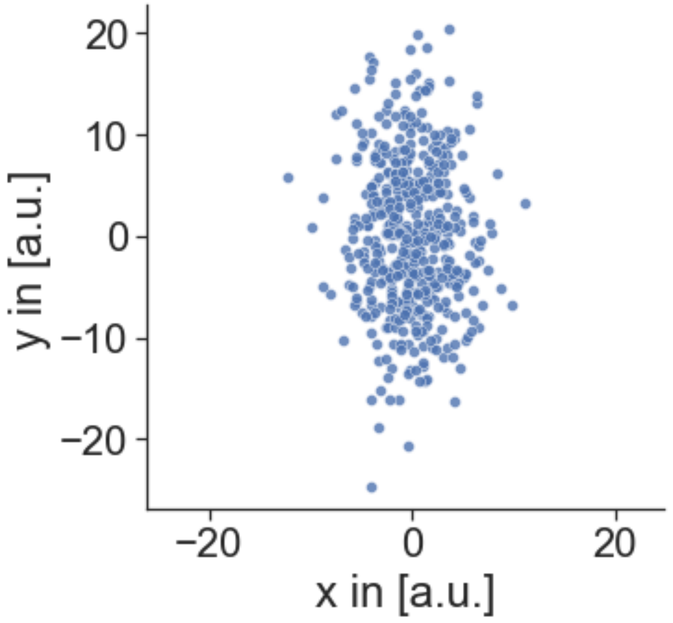


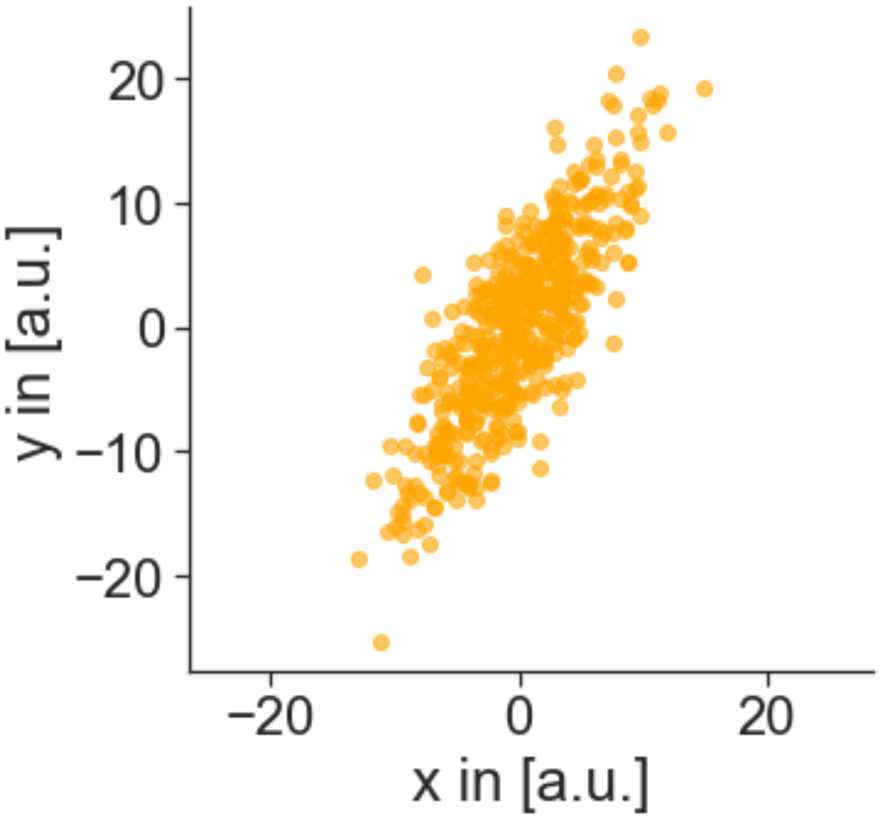
Principal component analysis (PCA)



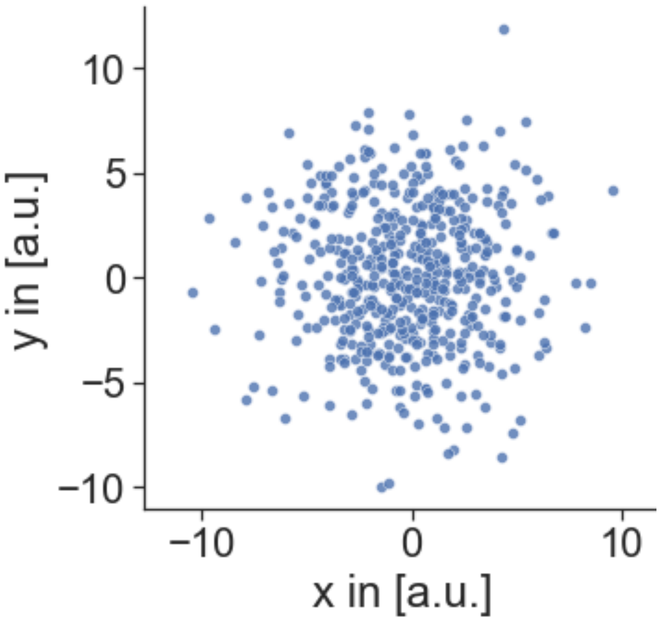
Principal component analysis (PCA)

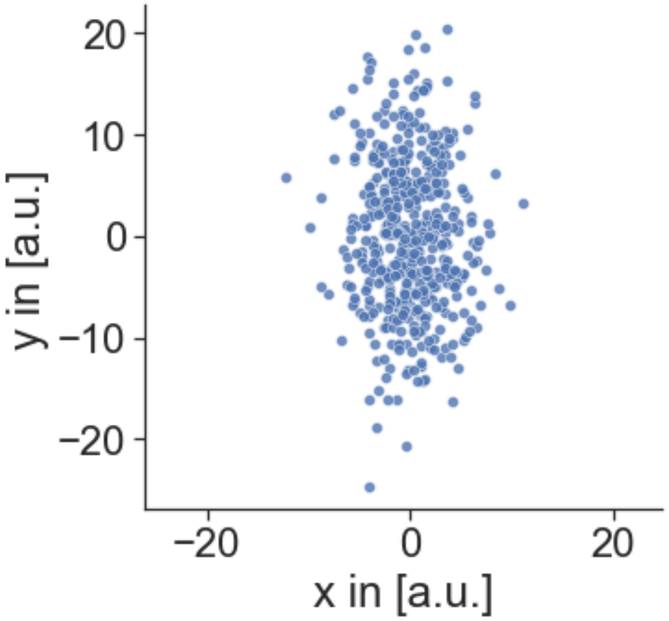


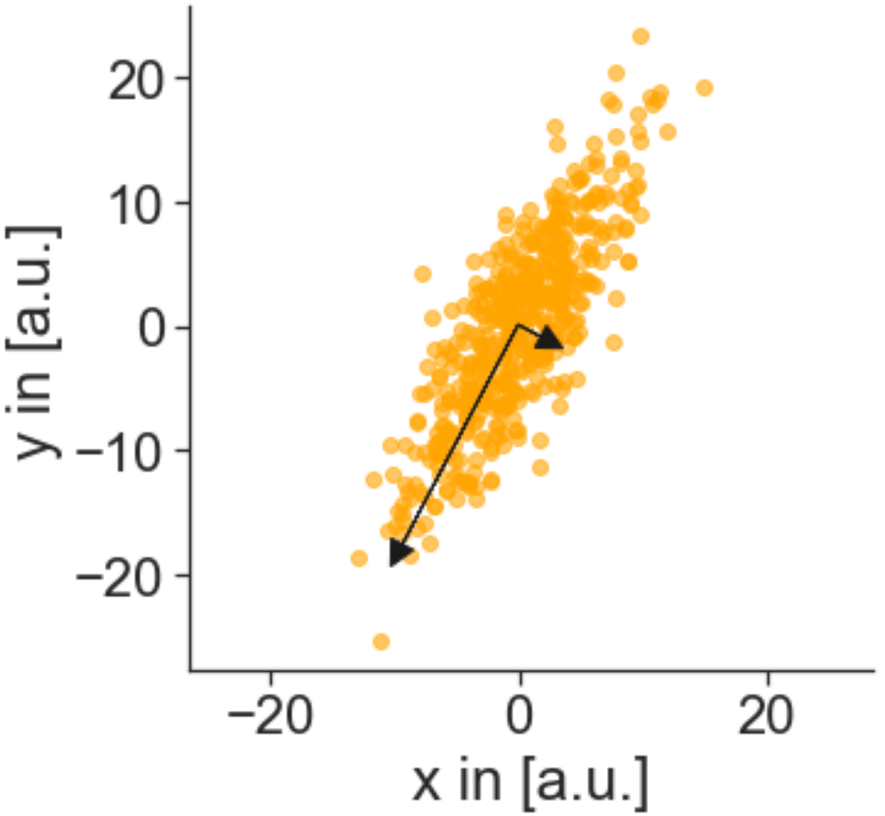




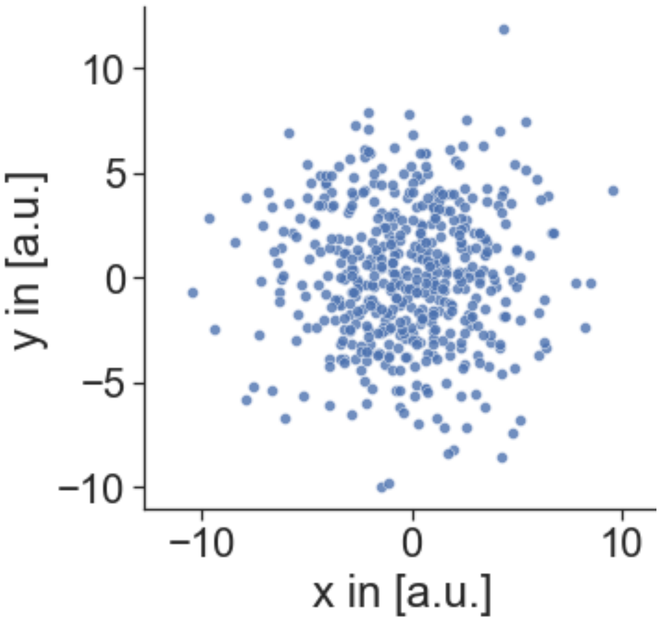
Principal component analysis (PCA)

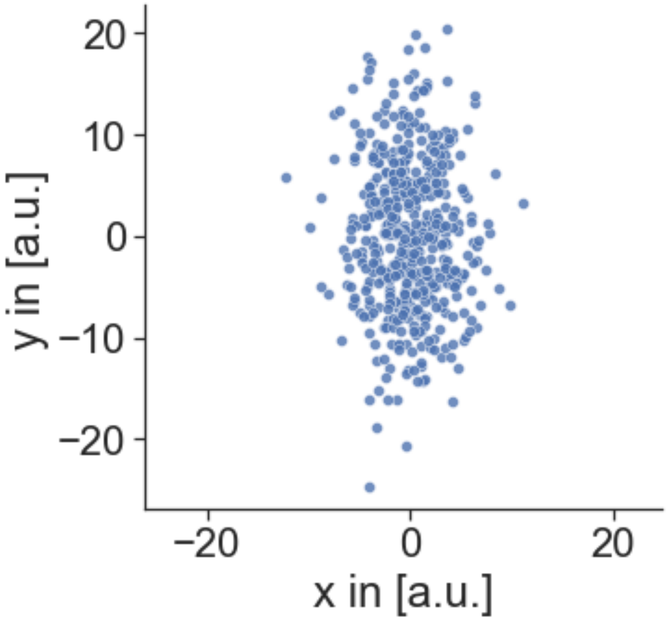


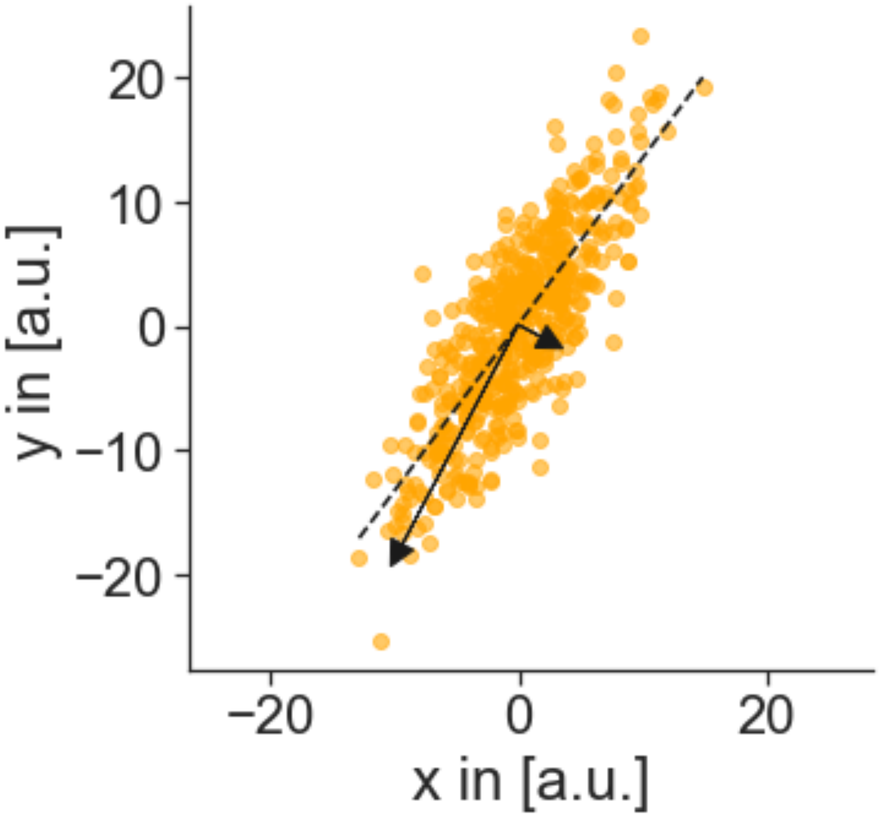




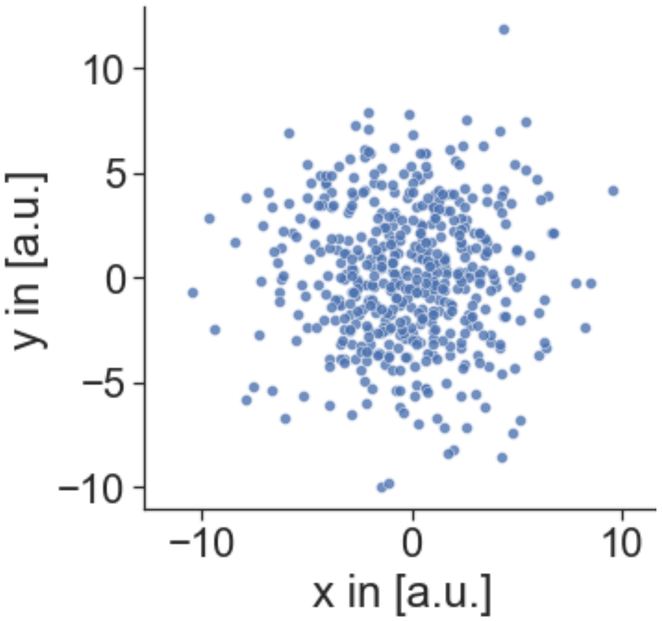
Principal component analysis (PCA)

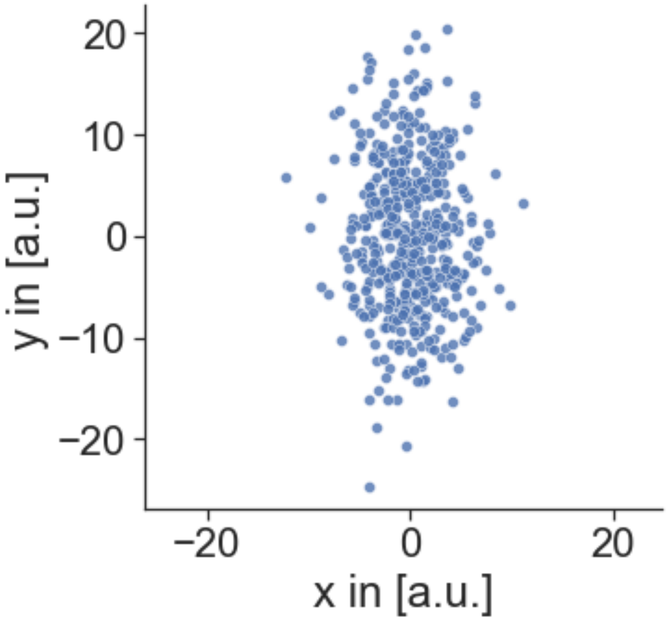


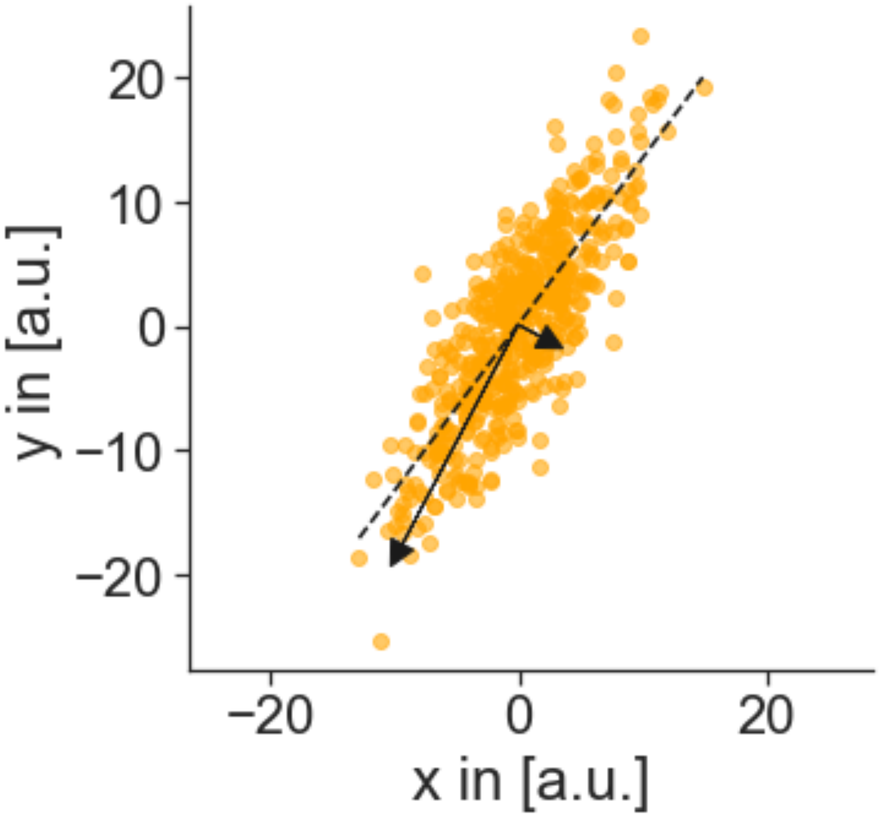




Principal component analysis (PCA)







- 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**

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**

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**

$\mathbf{C}(0)$ Is the covariance matrix of the data **X**

Solving the generalised eigenvalue problem

$$\mathbf{C}(0)\mathbf{W} = \mathbf{W}\boldsymbol{\Sigma}$$

Gives an eigenvector matrix **W** that will allow the transform of the original data **X** onto a new basis **T** that maximises the variance.

$$\mathbf{T} = \mathbf{X}\mathbf{W}$$

It is possible to choose m eigenvectors to project onto by only using the first m-columns of **W**.

Time-lagged independent component analysis

- tICA is a **linear transform** similar to PCA
- The transform is chosen such that amongst all linear transforms, tICA **maximizes the autocorrelation** of transformed coordinates.

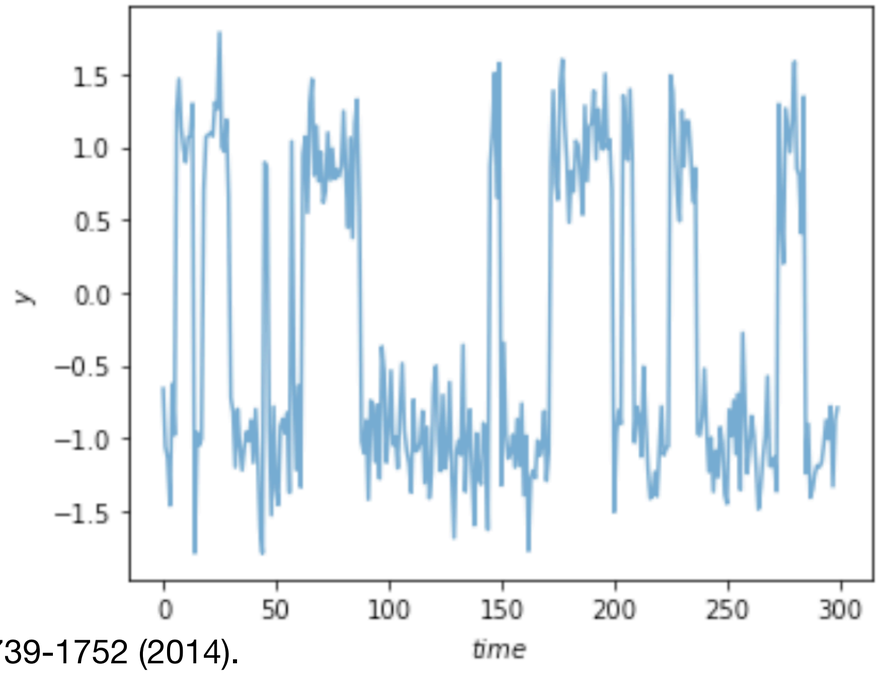
[1]: F. Nüske, B. Keller, G. Pérez-Hernández, **A. S. J. S. Me**y and F. Noé: *J. Chem. Theory Comput.* **10**, 1739-1752 (2014).

[2]: G. Perez-Hernandez, et al. *J. Chem. Phys.* **139**, 015102 (2013).

[3]: C. R. Schwantes et al. *J. Chem. Theory Comput.* **9**, 2000-2009 (2013).

Time-lagged independent component analysis

- tICA is a **linear transform** similar to PCA
- The transform is chosen such that amongst all linear transforms, tICA **maximizes the autocorrelation** of transformed coordinates.



[1]: F. Nüske, B. Keller, G. Pérez-Hernández, **A. S. J. S. Me**y and F. Noé: *J. Chem. Theory Comput.* **10**, 1739-1752 (2014).

[2]: G. Perez-Hernandez, et al. *J. Chem. Phys.* **139**, 015102 (2013).

[3]: C. R. Schwantes et al. *J. Chem. Theory Comput.* **9**, 2000-2009 (2013).

Time-lagged independent component analysis

- tICA is a **linear transform** similar to PCA
- The transform is chosen such that amongst all linear transforms, tICA **maximizes the autocorrelation** of transformed coordinates.

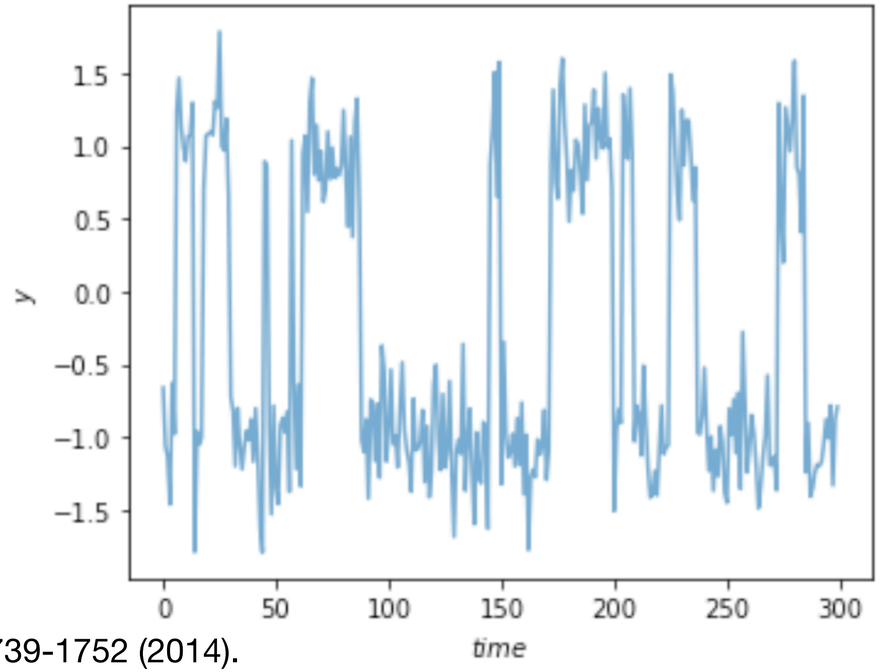
$\mathbf{r}(t) = (r_i(t))_{i=1,\ldots,D}$ D-dimensional input data vector that is mean free, i.e.

$$\mathbf{r}(t) = \mathbf{r}(t) - \langle \mathbf{r}(t) \rangle_t$$

Computing the covariance of the data at t = 0 and t=τ which is the lag-time chosen.

$$c_{ij}(\tau) = \langle r_i(t) r_j(t+\tau) \rangle_t$$

This allows the computation of two covariance matrices:
$\mathbf{C}(0)$ and $\mathbf{C}(\tau)$



[1]: F. Nüske, B. Keller, G. Pérez-Hernández, **A. S. J. S. Me**y and F. Noé: *J. Chem. Theory Comput.* **10**, 1739-1752 (2014).

[2]: G. Perez-Hernandez, et al. *J. Chem. Phys.* **139**, 015102 (2013).

[3]: C. R. Schwantes et al. *J. Chem. Theory Comput.* **9**, 2000-2009 (2013).

Time-lagged independent component analysis

- tICA is a **linear transform** similar to PCA
- The transform is chosen such that amongst all linear transforms, TICA **maximizes the autocorrelation** of transformed coordinates.

Entries of the covariance matrix can be computed as:

$$c_{ij}(\tau) = \frac{1}{N-\tau-1} \sum_{t=1}^{N-\tau} r_i(t) r_j(t+\tau)$$

[1]: F. Nüske, B. Keller, G. Pérez-Hernández, **A. S. J. S. Me**y and F. Noé: *J. Chem. Theory Comput.* **10**, 1739-1752 (2014).

[2]: G. Perez-Hernandez, et al. *J. Chem. Phys.* **139**, 015102 (2013).

[3]: C. R. Schwantes et al. *J. Chem. Theory Comput.* **9**, 2000-2009 (2013).

Time-lagged independent component analysis

- tICA is a **linear transform** similar to PCA
- The transform is chosen such that amongst all linear transforms, TICA **maximizes the autocorrelation** of transformed coordinates.

Entries of the covariance matrix can be computed as:

$$c_{ij}(\tau) = \frac{1}{N-\tau-1} \sum_{t=1}^{N-\tau} r_i(t) r_j(t+\tau)$$

$\mathbf{C}(0)$ Will be a symmetric matrix. The symmetry of $\mathbf{C}(\tau)$ will need to be enforced with:

$$\mathbf{C}(\tau) = \frac{1}{2}(\mathbf{C}_d(\tau) + \mathbf{C}_d^{\mathsf{T}}(\tau))$$

We can now solve the generalised eigenvalue problem:

$$\mathbf{C}(\tau)\mathbf{U} = \mathbf{C}(0)\mathbf{U}\mathbf{\Lambda}$$

Eigenvector matrix containing ICs **Diagonal matrix with eigenvalues**

[1]: F. Nüske, B. Keller, G. Pérez-Hernández, **A. S. J. S. Me**y and F. Noé: *J. Chem. Theory Comput.* **10**, 1739-1752 (2014).

[2]: G. Perez-Hernandez, et al. *J. Chem. Phys.* **139**, 015102 (2013).

[3]: C. R. Schwantes et al. *J. Chem. Theory Comput.* **9**, 2000-2009 (2013).

Time-lagged independent component analysis

- tICA is a **linear transform** similar to PCA
- The transform is chosen such that amongst all linear transforms, TICA **maximizes the autocorrelation** of transformed coordinates.

Entries of the covariance matrix can be computed as:

$$c_{ij}(\tau) = \frac{1}{N-\tau-1} \sum_{t=1}^{N-\tau} r_i(t) r_j(t+\tau)$$

$\mathbf{C}(0)$ Will be a symmetric matrix. The symmetry of $\mathbf{C}(\tau)$ will need to be enforced with:

$$\mathbf{C}(\tau) = \frac{1}{2}(\mathbf{C}_d(\tau) + \mathbf{C}_d^{\mathsf{T}}(\tau))$$

We can now solve the generalised eigenvalue problem:

$$\mathbf{C}(\tau)\mathbf{U} = \mathbf{C}(0)\mathbf{U}\mathbf{\Lambda}$$

Eigenvector matrix containing ICs

Diagonal matrix with eigenvalues

$$\mathbf{z}^{\mathsf{T}}(t) = \mathbf{r}^{\mathsf{T}}(t)\mathbf{U}$$

M columns of full rank U for DR

[1]: F. Nüske, B. Keller, G. Pérez-Hernández, **A. S. J. S. Me**y and F. Noé: *J. Chem. Theory Comput.* **10**, 1739-1752 (2014).

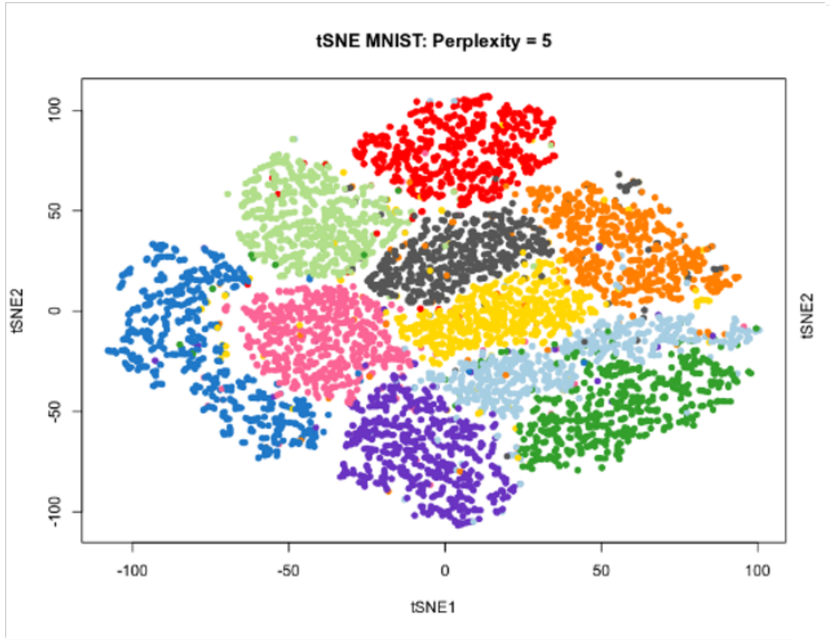
[2]: G. Perez-Hernandez, et al. *J. Chem. Phys.* **139**, 015102 (2013).

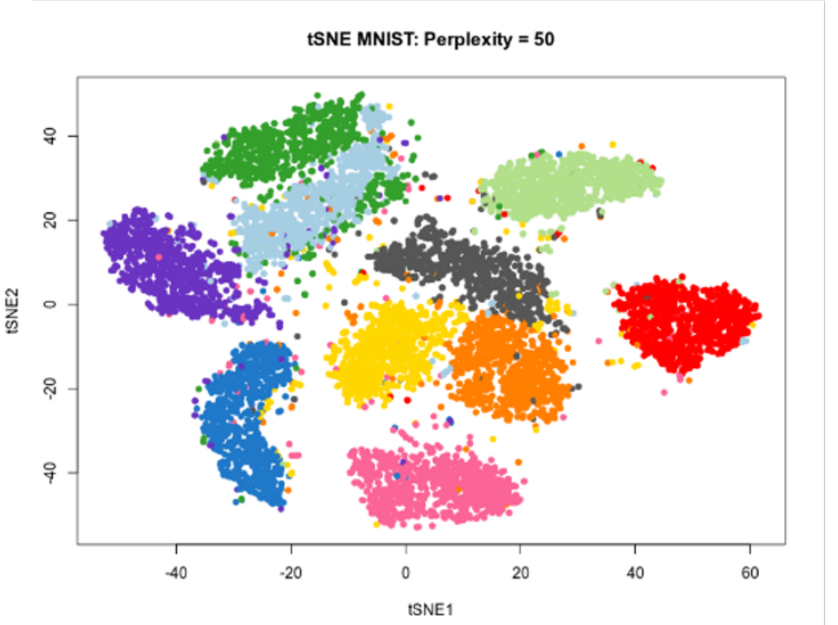
[3]: C. R. Schwantes et al. *J. Chem. Theory Comput.* **9**, 2000-2009 (2013).

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





Let’s head to GitHub and try some of these concepts

https://github.com/CCPBioSim/MDAnalysis_ML_workshop

You will need to update your repository if you cloned it before with:

```
$: git pull origin main
```

Or using your favourite way of updating a git repository

Post-its