# Unsupervised

Fenna Feenstra Msc
Hanze University of Applied Sciences

# Schedule

- Clustering theory
- Clustering use case (k-means)
- Distance metrics theory
- Data analysis use case (HAC)
- Dimensionality reduction theory
- Feature selection use case

# Usage

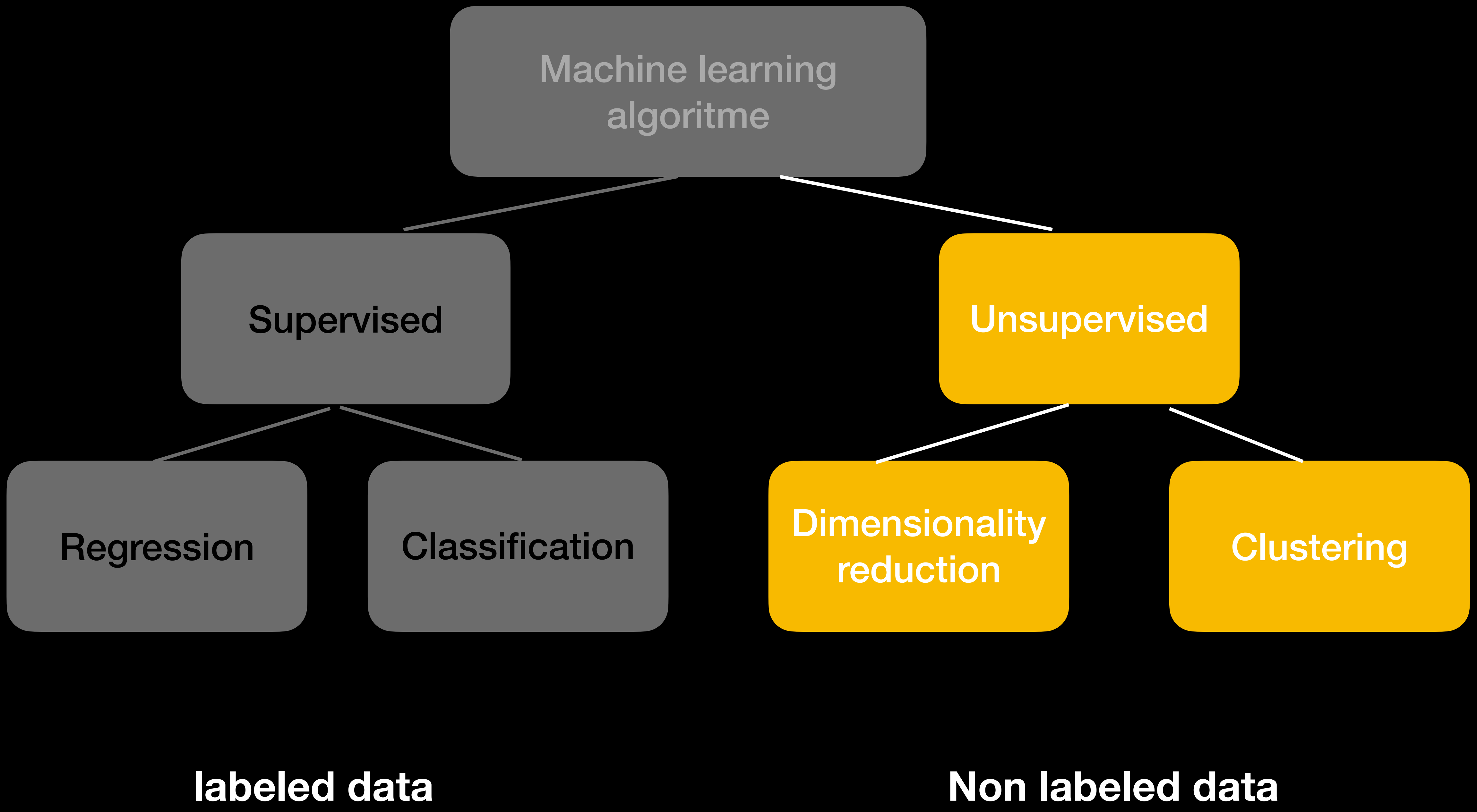
| | |
|---|---|
| **Python3** | Programming language |
| **NumPy** | Apply calculations and linear algebra |
| **Pandas** | Store and view results |
| **Matplotlib/sns** | Plot (intermediate) results |
| **Jupyter** | For demonstration purpose |
| **Pip3/conda** | Installing required packages |
| **Github** | Share code |

# Material

| Library | Docs | cheatsheet |
|---|---|---|
| python | https://www.python.org/doc/ | https://perso.limsi.fr/pointal/_media/python:cours:mementopython3-english.pdf |
| pandas | pandas.pydata.org/pandas-docs/stable/ | https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf |
| numpy | https://docs.scipy.org/doc/numpy | https://s3.amazonaws.com/dq-blog-files/numpy-cheat-sheet.pdf |
| matplotlib | https://matplotlib.org/contents.html | https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Scikit_Learn_Cheat_Sheet_Python.pdf |
| sklearn | scikit-learn.org | https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Scikit_Learn_Cheat_Sheet_Python.pdf |

# Book

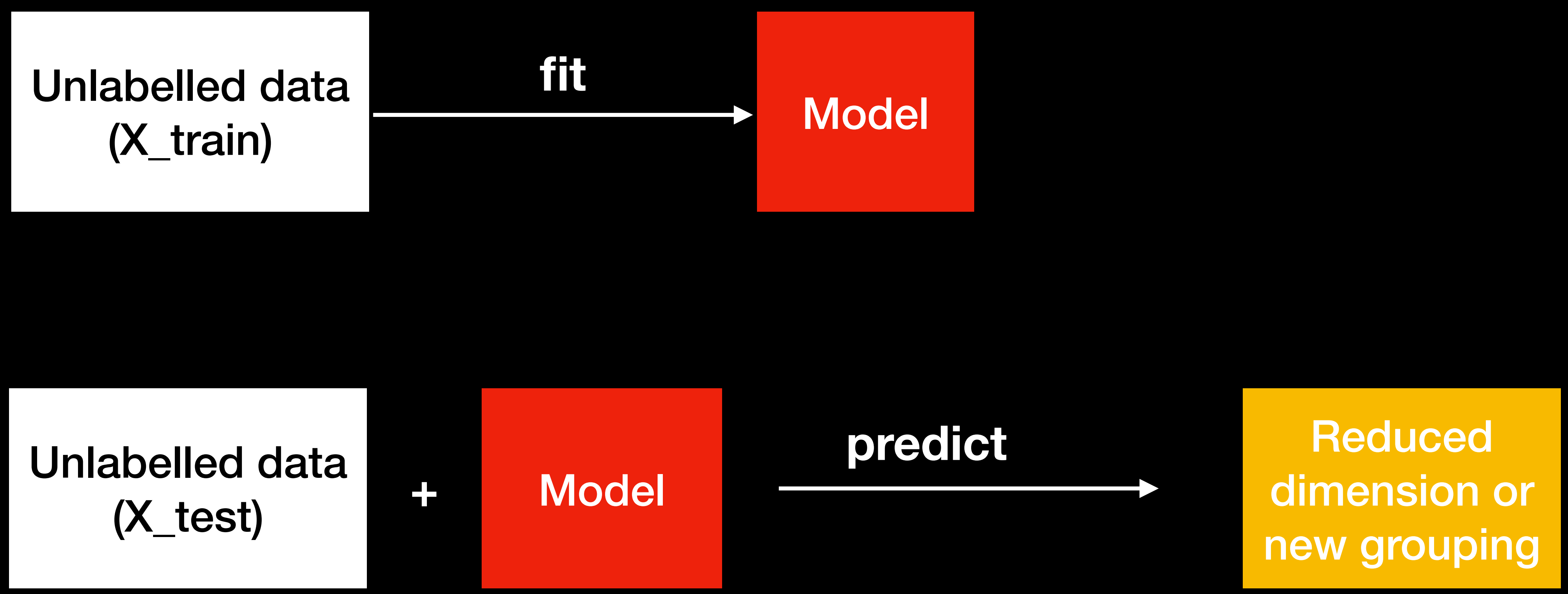
# Machine learning



# Use cases

- Try to find structures in our dataset
  - Clustering: identify unknown structures (group data, segmentation analysis, anomaly detection, classification, improve supervised learning by modelling per cluster)
  - Dimensionality reduction: use structural characteristics to simplify data (image compression, text analysing, feature selection, improve plotting)

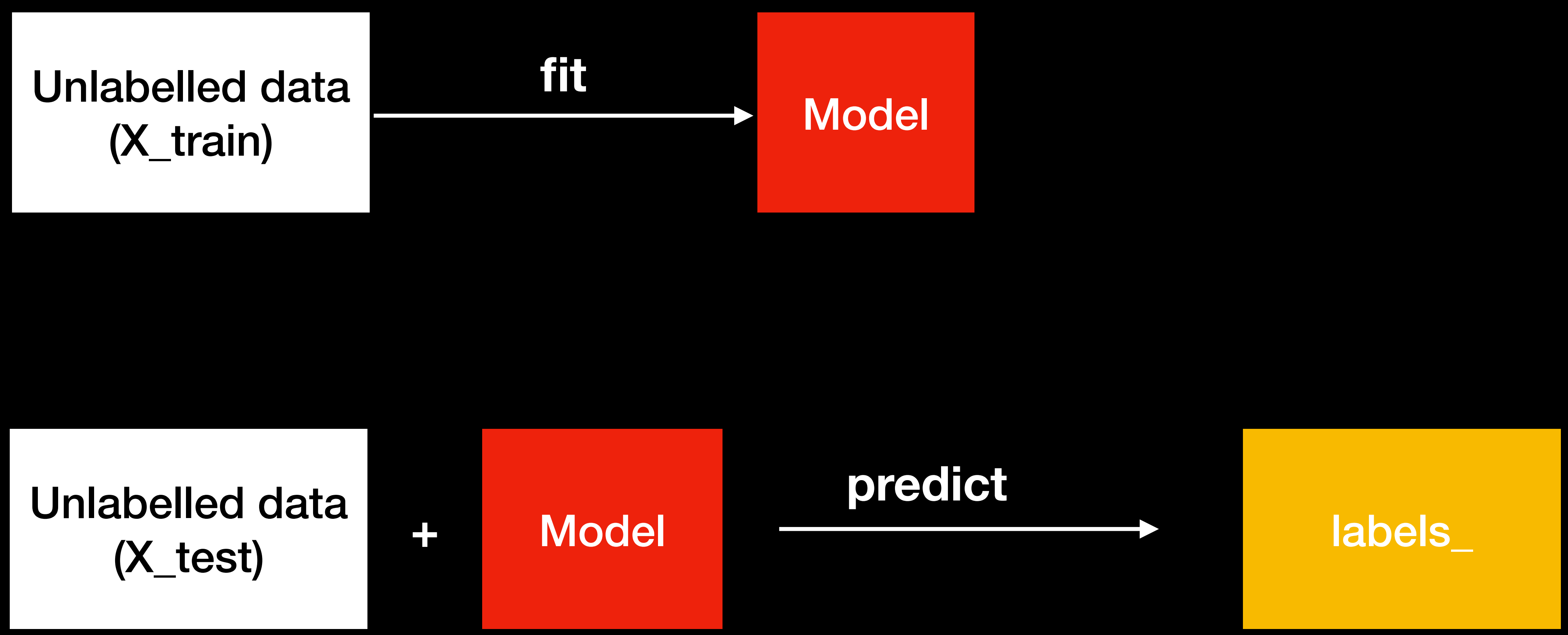
# General method



# Clustering

- Group data points based on certain similarities. Data does not have labels, best algorithm depends on use case. Most common:
    - K-means
    - Meanshift
    - HAC
    - DBSCAN

# Cluster method



# Examples

- Group documents by topic
- Group organism by genetic information
- Group colours into cluster-IDs to compress image

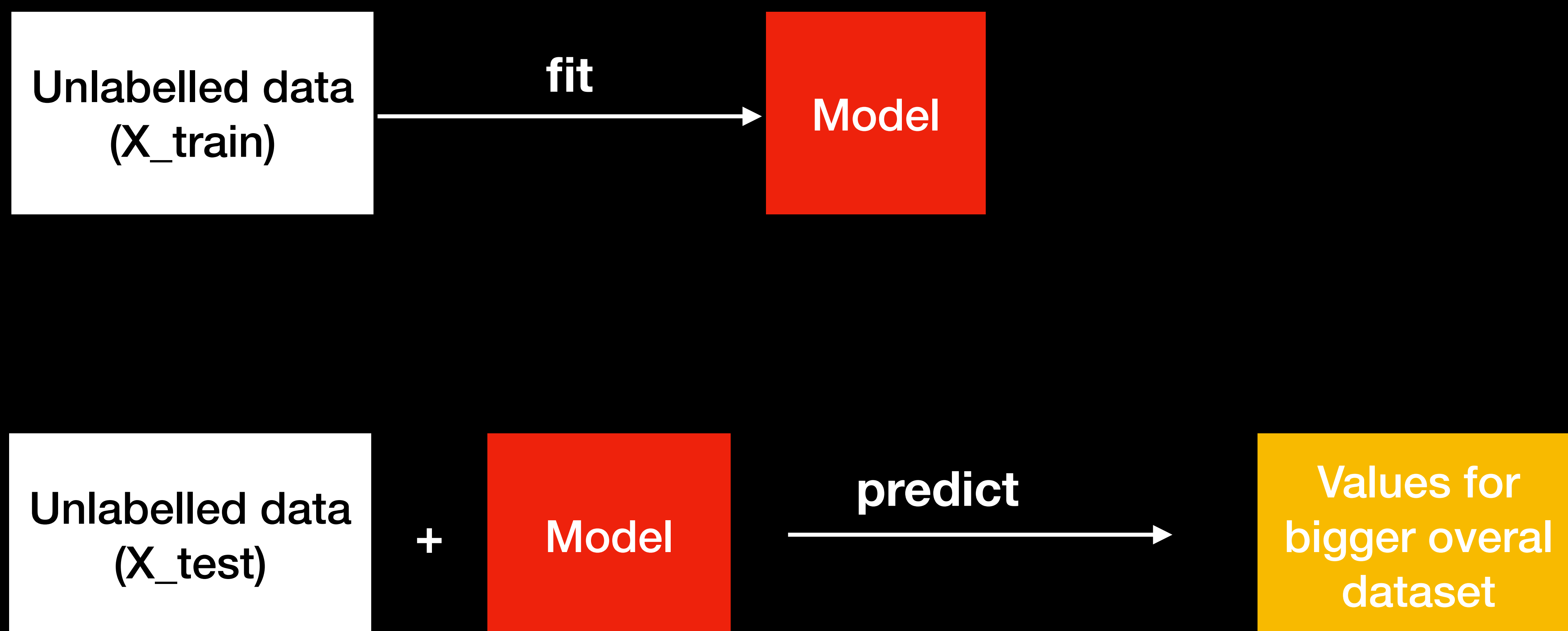
# Clustering methods

| Method | K-means | Mean Shift | HAC | DBSCAN |
|---|---|---|---|---|
| Class | KMeans | MeanShift | Agglomerative Clustering | DBSCAN |
| Essential Parameters | n_clusters | bandwidth | n_clusters, linkage, affinity | epsilon, min_samples metric |
| Distance methods | Euclidean | Euclidean | euclidean/l1/ L2/Manhattan/ cosine/ precomputed | About 23 different metrics |
| Use case | find a few clusters of ~ same size | determine the (large) amount of clusters | get a full tree (informative) | tons of data and weirds shapes |
| Remark | Fast, but bad on non spherical clusters | Slow, bad on weird shapes | Slow, good on weird shapes or uneven cluster sizes | Highly configurable ,good performance |

# Dimensionality reduction

- Coming up with a lower dimensional representation of our original data that maintains the majority of the information important for us in the original dataset (create new features as combination of original features). Most common:
    - PCA (linear principal component analysis)
    - KernelPCA (non linear principal component analysis)
    - MDS (Multidimensional Scaling)
    - t-SNE (t-distributed Stochastic Neighbor Embedding)
    - NMF (non negative matrix factorisation)

# dimensionality reduction



# Dimensions reduction example

fit →

Compression Model

+

Compression Model

predict (reconstruct picture) →

# Curse of dimensionality

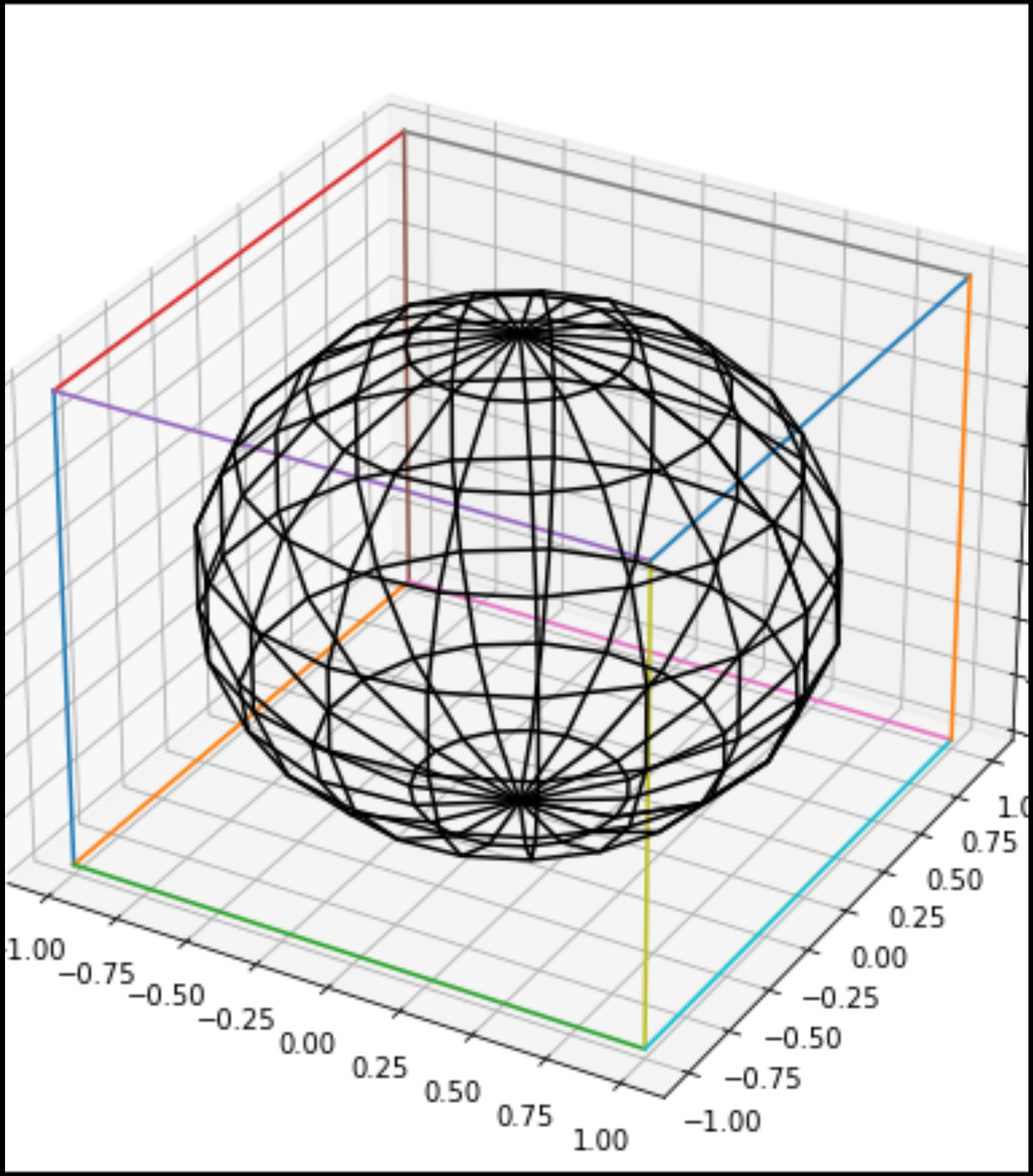
Curse of Dimensionality describes the explosive nature of increasing data dimensions and its resulting exponential increase in computational efforts required for its processing and/ or analysis

2 dim -> 79% inside sphere

3 dim -> 52% inside sphere

10 dim -> 0.25% inside sphere

high-dimensional space leads to sparse data; clustering is difficult when points are far away from each other



# unsupervised+ supervised

Use dimensionality reduction to lower number of features

- PCA
- Matrix factorisation
- t-SNE
- MDS

https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/

# methods

| Method | PCA | KernelPCA | MDS | NMF |
|---|---|---|---|---|
| Class | PCA | KernelPCA | MDS | NMF |
| Essential Parameters | n_components | n_components, kernel, gamma | n_components | n_components, init |
| Use case | In case of linear combination of features | Non linear relationships | Non linear relationships | Only for positive values (words, images) |
| Remark | Preserves variance as much as possible | Requires more computation | Preserving distance between points rather than variance | |