VYSOKÁ ŠKOLA CHEMICKO-TECHNOLOGICKÁ V PRAZE

QSAR MODELLING

MARIIA MATVEIEVA

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

6th Advanced in silico Drug Design workshop/challenge 2023

Department of Physical Chemistry
Palacky University
Olomouc 2023

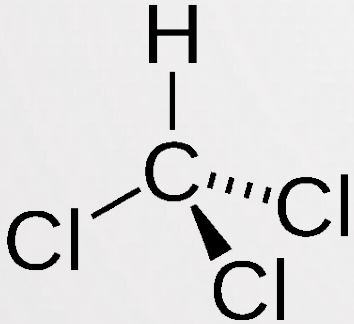
OUTLINE

1. What is **QSAR**? How did it all **begin**?
2. When/how is it applied in drug design **nowadays**?
3. **Methods**: overview
4. QSAR model **performance** and v**alidation**
5. **'Applicability domain'**
6. **Conclusions**

1. What is **QSAR**? How did it begin?

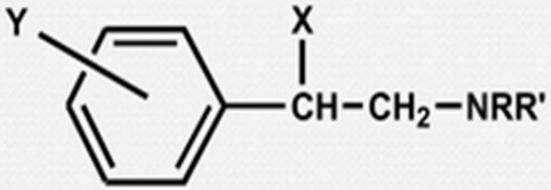
What is QSAR? How did it begin?

- 1900s lipoid theory of narcosis



- 1960s Hansch & Fujita

'Quantitative structure activity relationship'



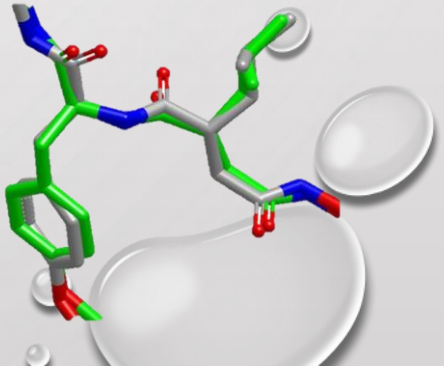
Activity = F(structure)

$$\mathrm{Log}\left(\frac{1}{C}\right) = 1.22\,\pi - 1.59\,\sigma + 7.89$$

- 1970..80s descriptors, mathematical formalism

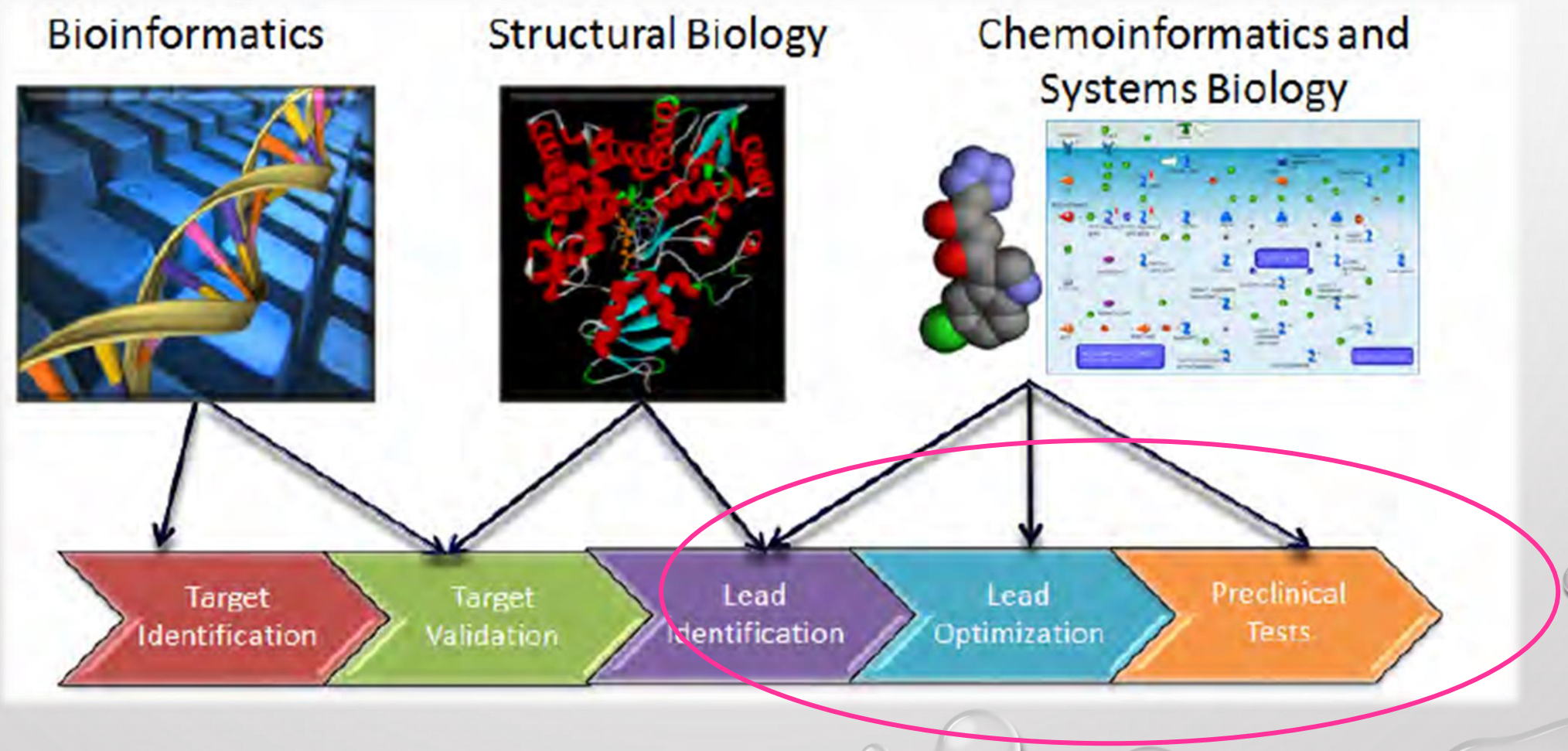
1	0	1	0	0	0	1	1

- 1980..90s 3D methods



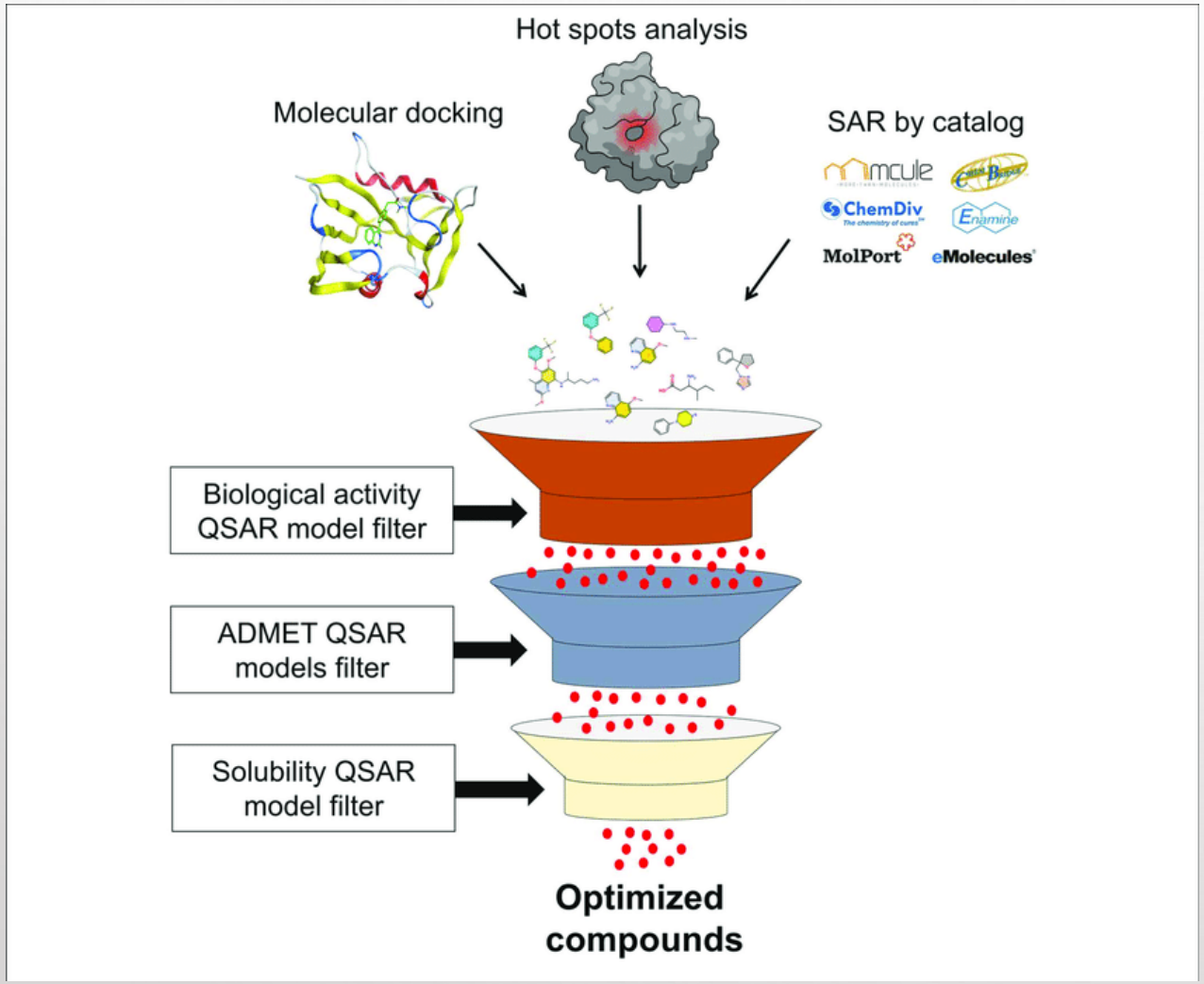
2. When/how is it applied in drug design now?

When/how is it applied in drug design now?



From book chapter: **Computational Strategies in Cancer Drug Discovery,** DOI: 10.5772/24177

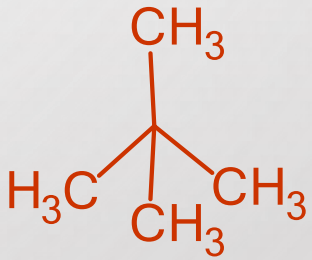
When/how is it applied in drug design now?



Activity = F(structure)

- Strength of binding to a protein (affinity)
- Inhibition of cell growth/division
- Penetration through membrane …

8.53



?

10.3389/fchem.2020.00093.

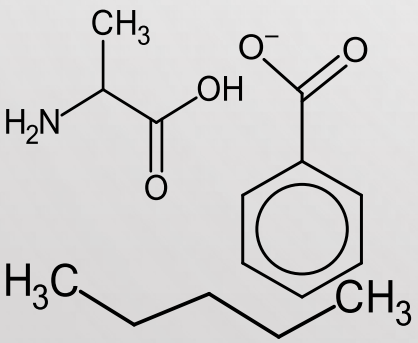
3. Methods overview

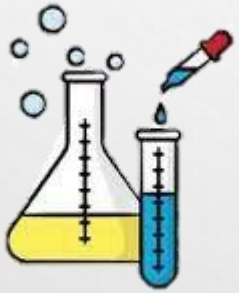
Methods overview

Data!

10^3-10^6

- Comes from experiment

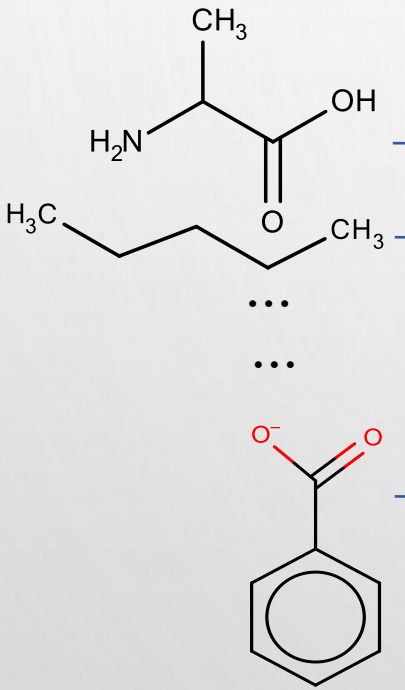




Activity
1.2
1.1
0.4
…
…
…

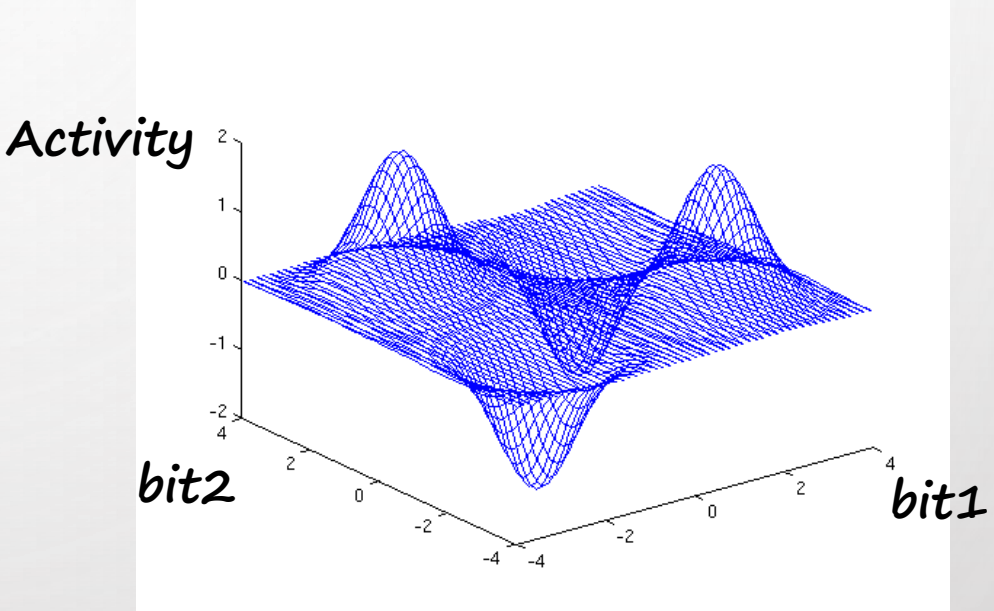
Machine Learning
a.k.a. artificial intelligence

Methods overview



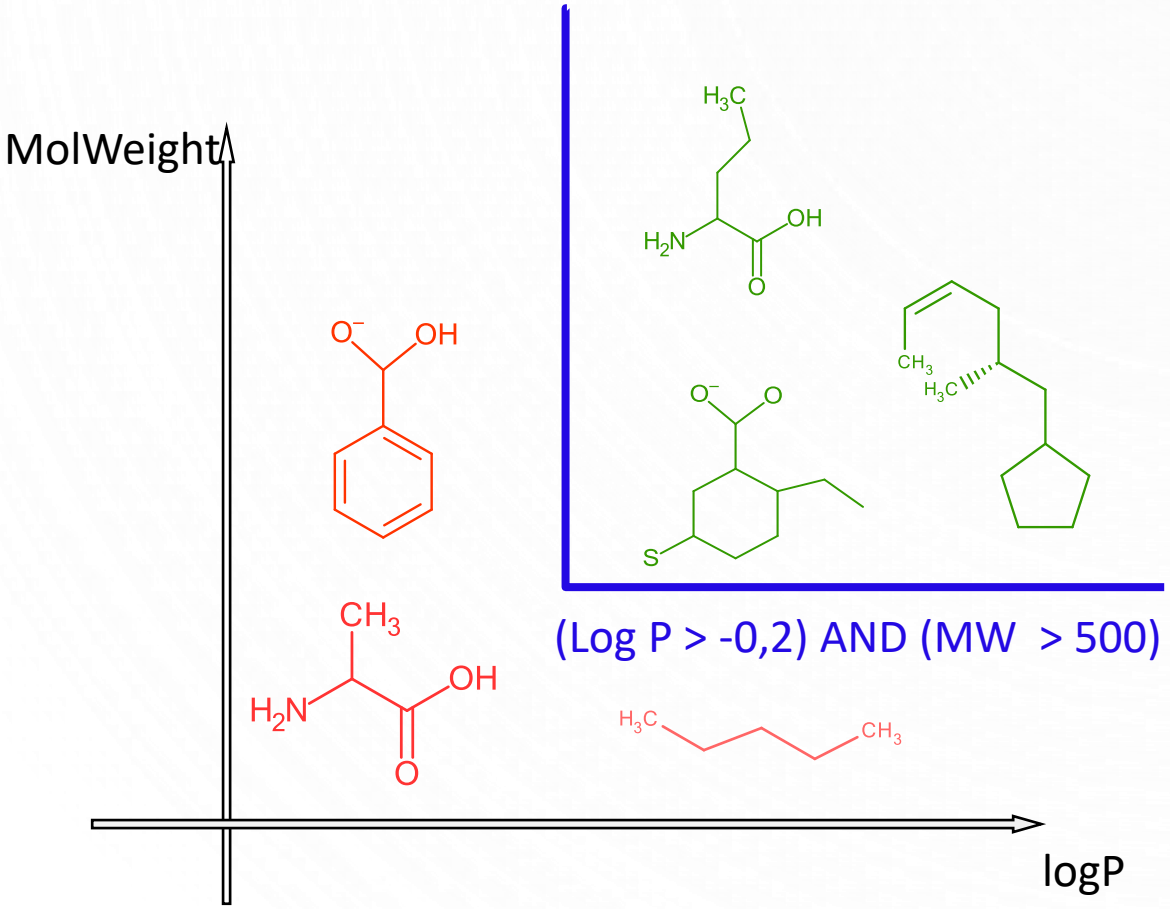
fingerprint

D_1	D_2	D_3	D_4	D_5	D_6	...	D_N
1	0	1	0	1	1	...	1
1	0	1	0	0	0	...	1
0	0	0	0	1	1	...	0
0	1	1	1	0	0	...	0
...	...	...	...	...	...	...	...
1	0	0	0	1	0	...	1

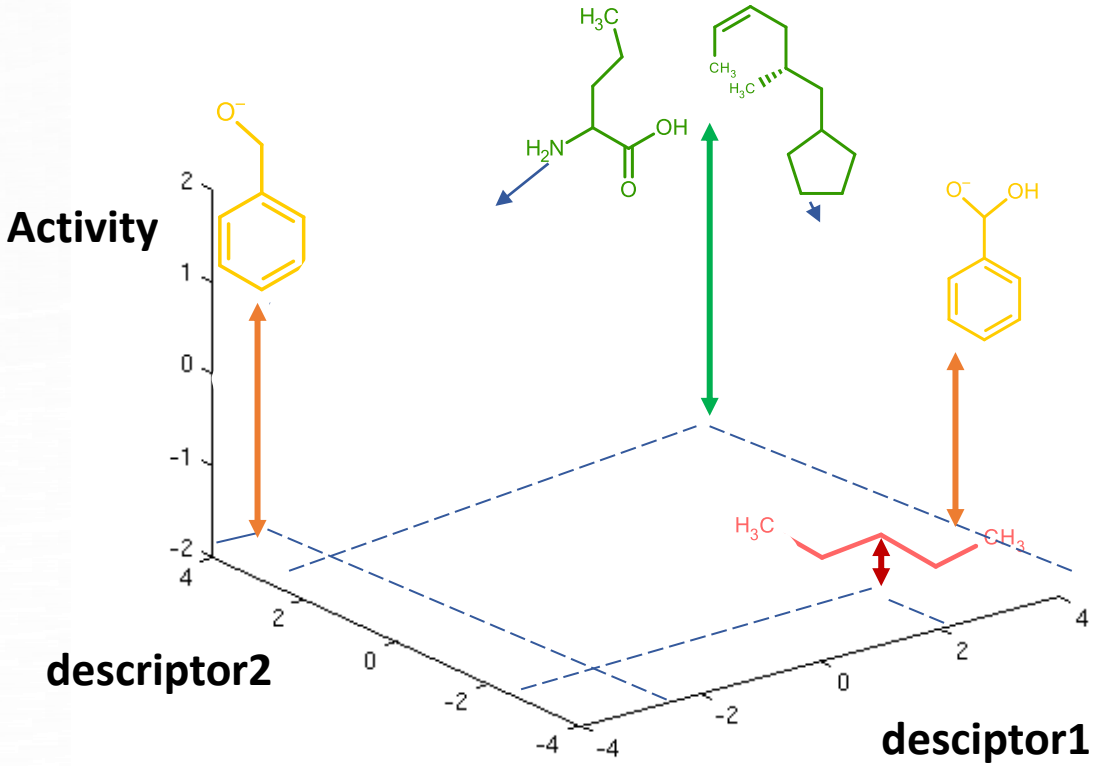


Regression & classification QSAR

Classification (Clusterization..)

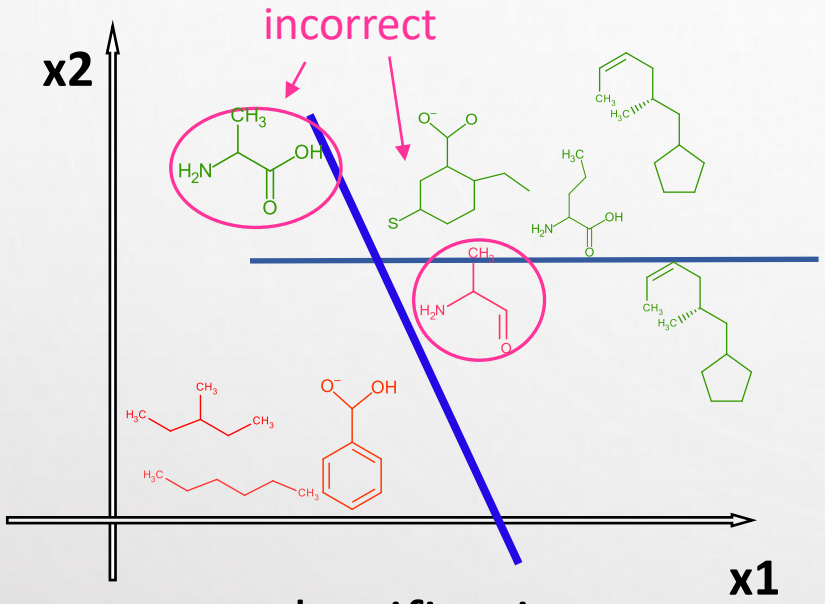


Regression

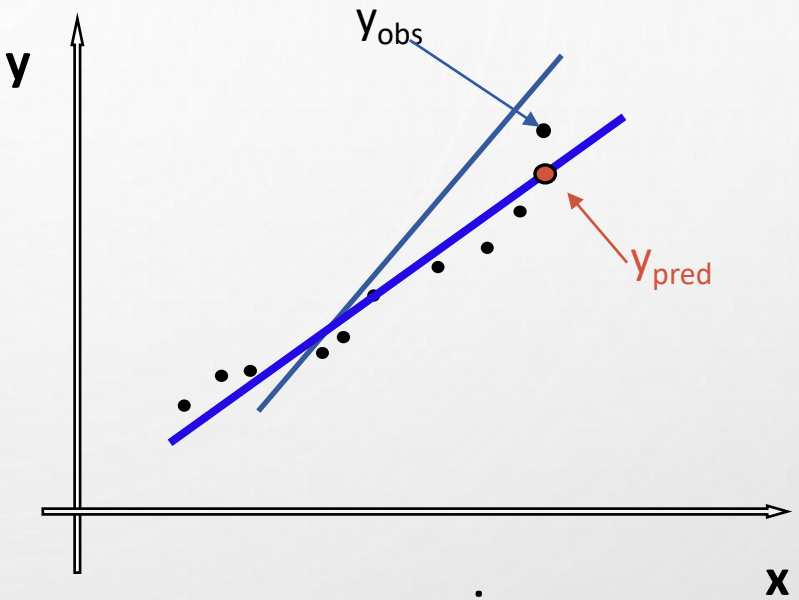


Optimization problem

Minimize (maximize) objective function



classification



regression

Accuracy:

$[_{N\ correctly\ classified}]/N$

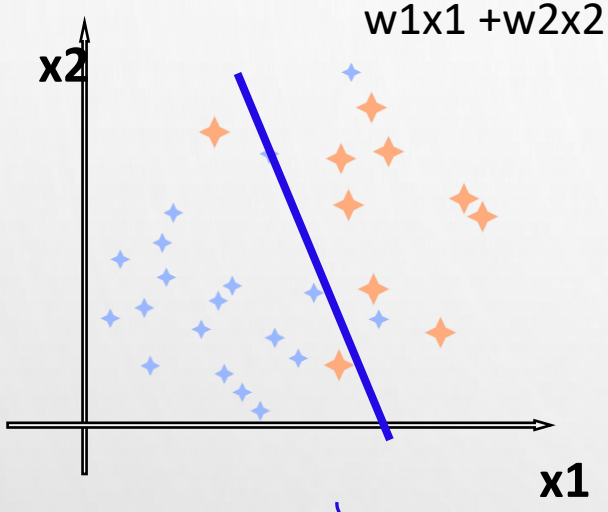
Sum of squared errors:

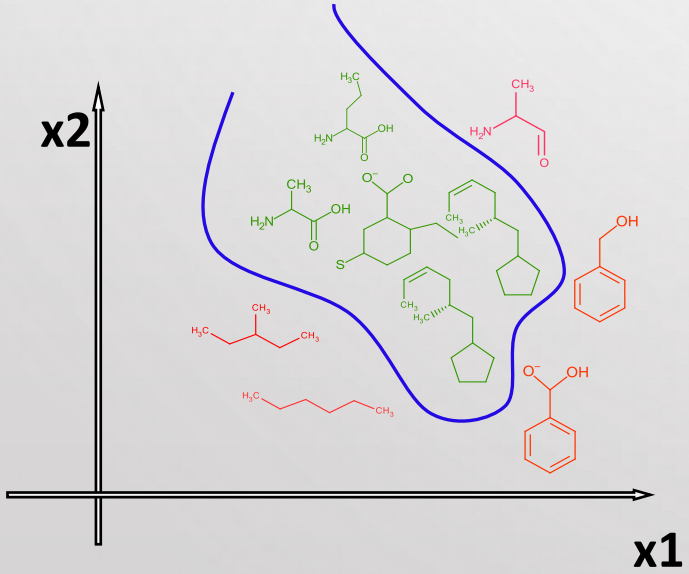
$L = \sum(y_{obs} - y_{pred})^2$

Linear & nonlinear models

classification

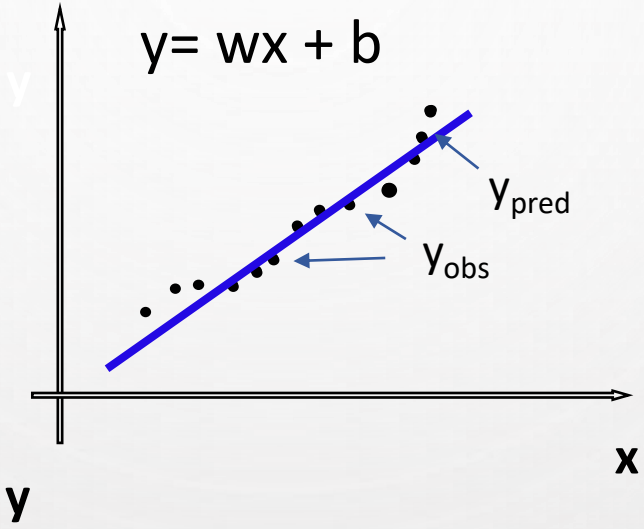
$w1x1 + w2x2 + b = 0$



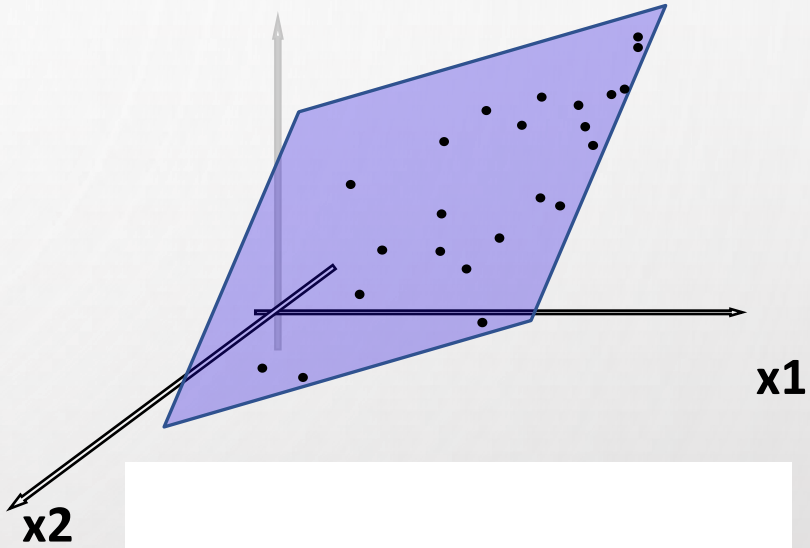


regression

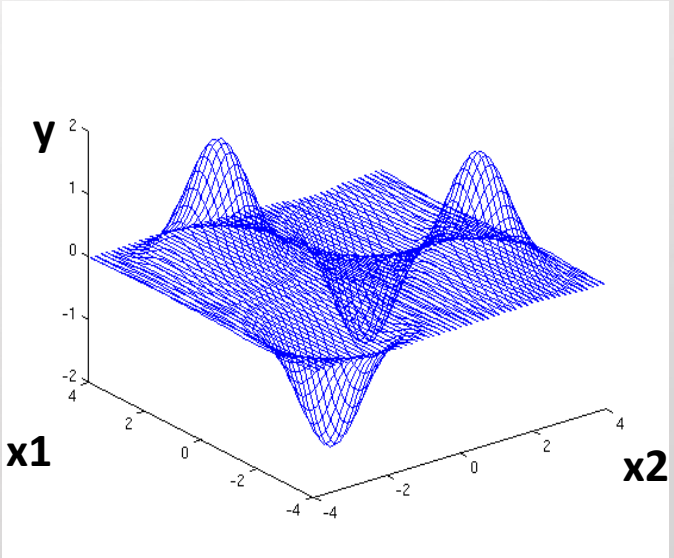
$y = wx + b$



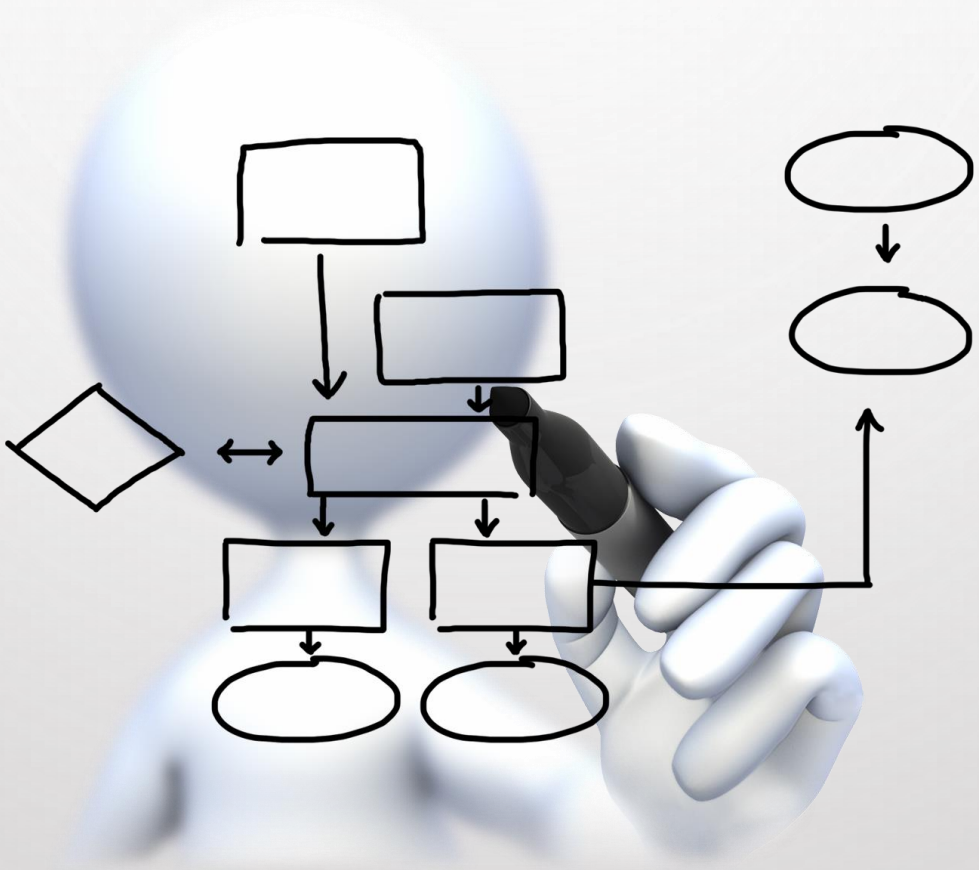
$Y = w_1x_1 + w_2x_2 + b$



Non-linear



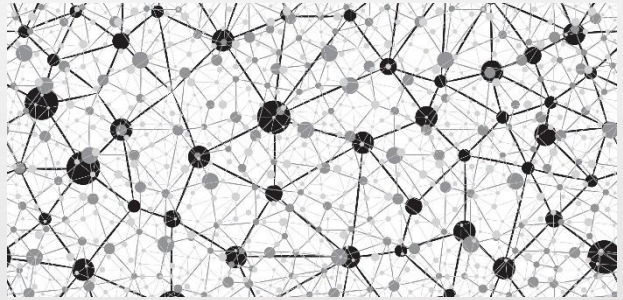
ALGORITHMS…



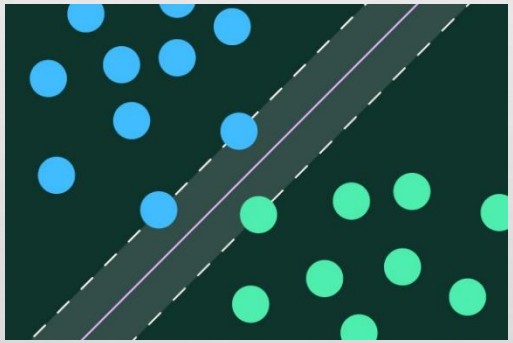
Algorithms

- RANDOM FOREST
- GRADIENT BOOSTING

- NEURAL NETWORKS, DEEP LEARNING

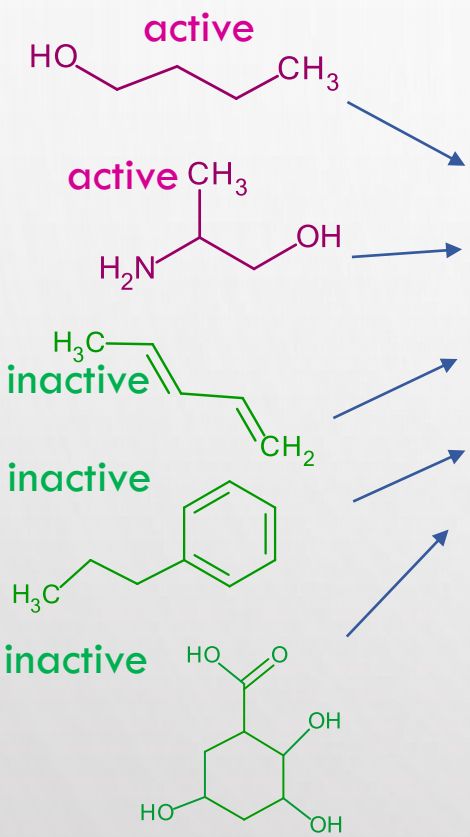


- SUPPORT VECTOR MACHINES



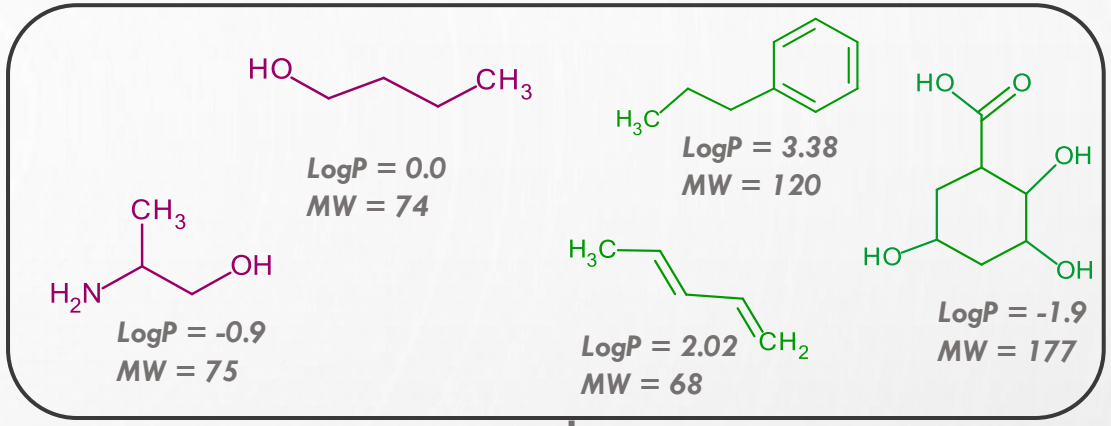
Random Forest

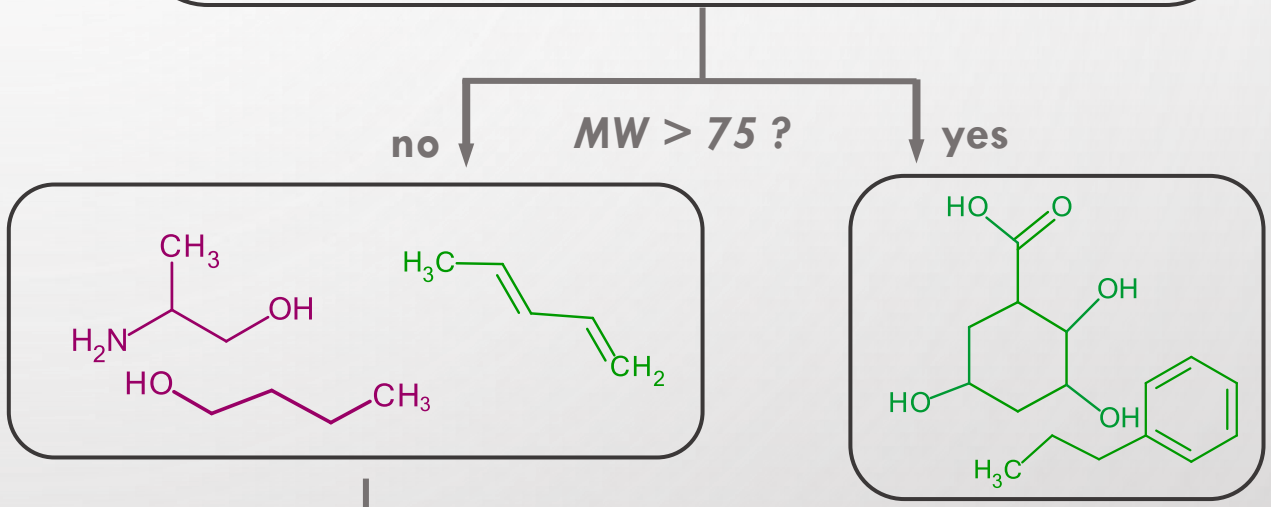
Decision Tree

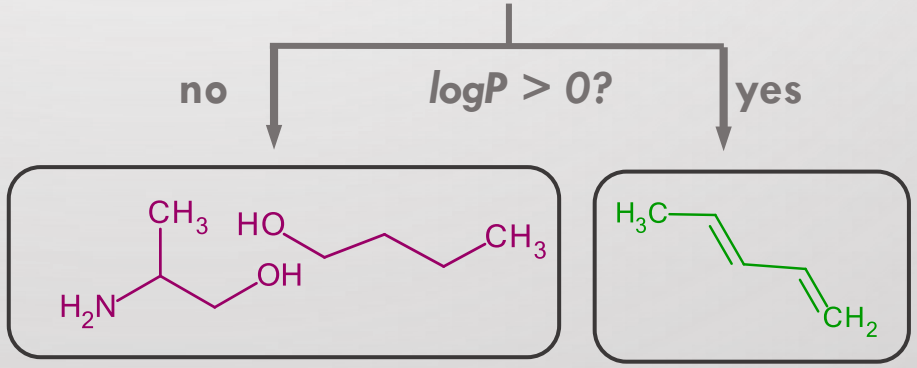


descriptors

LogP	Molecular Weight
0	74
-0.9	75
2.02	68
3.38	120
-1.9	177

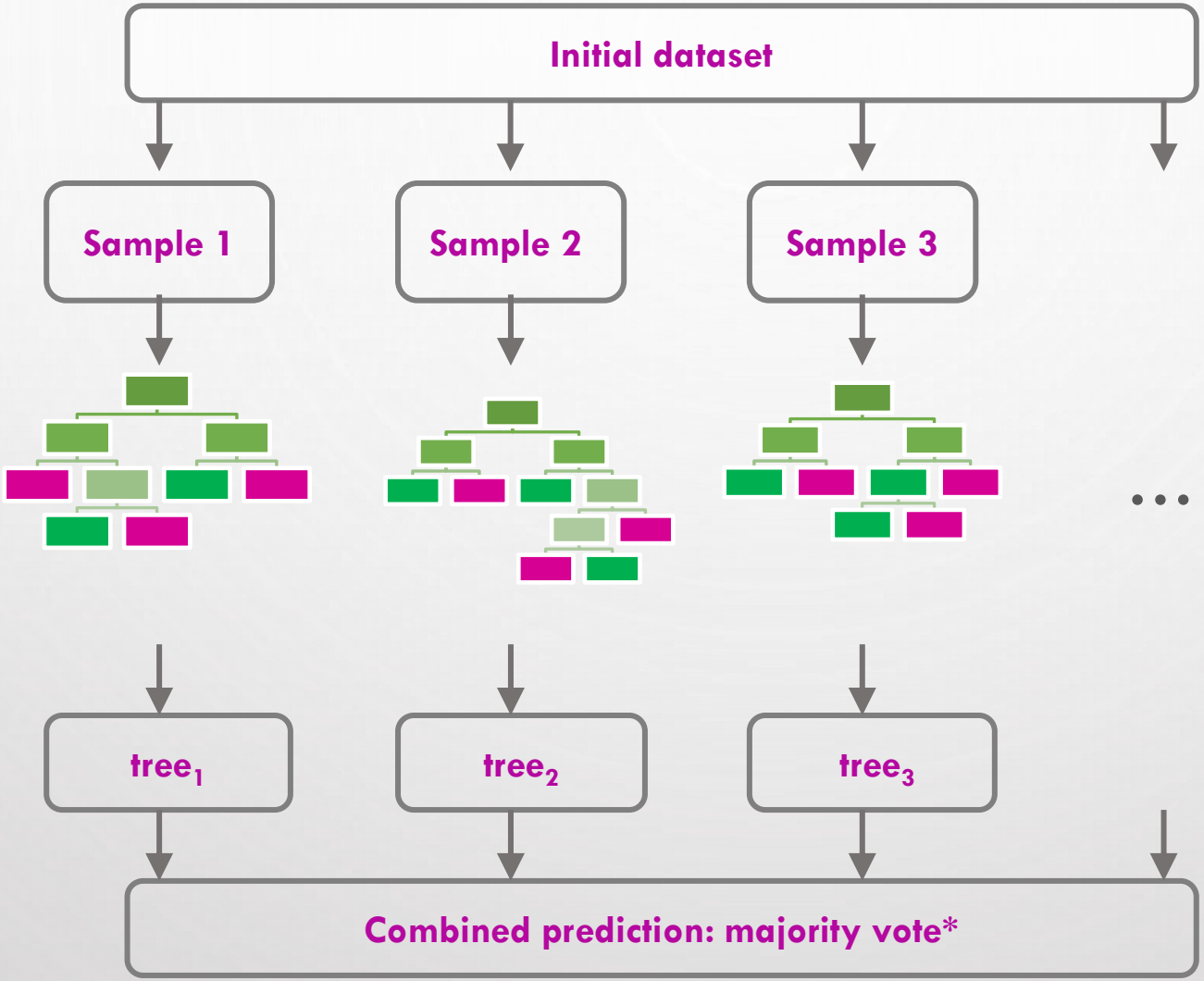




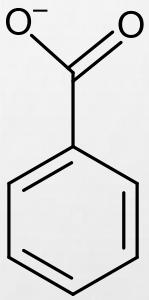


MODEL: "MW >75 OR (MW<=75 & LOG P >0) → ACTIVE "

Forest



*majority vote:



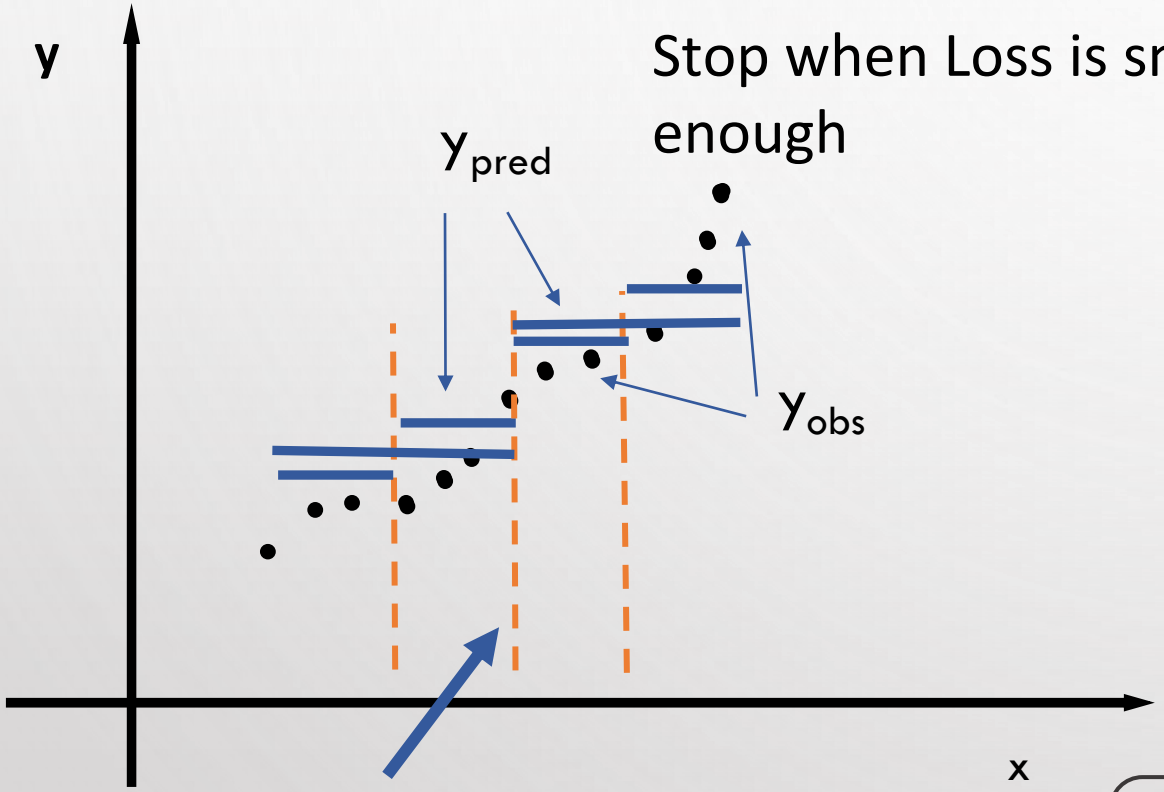
Active?

5 trees: Yes

2 trees: No

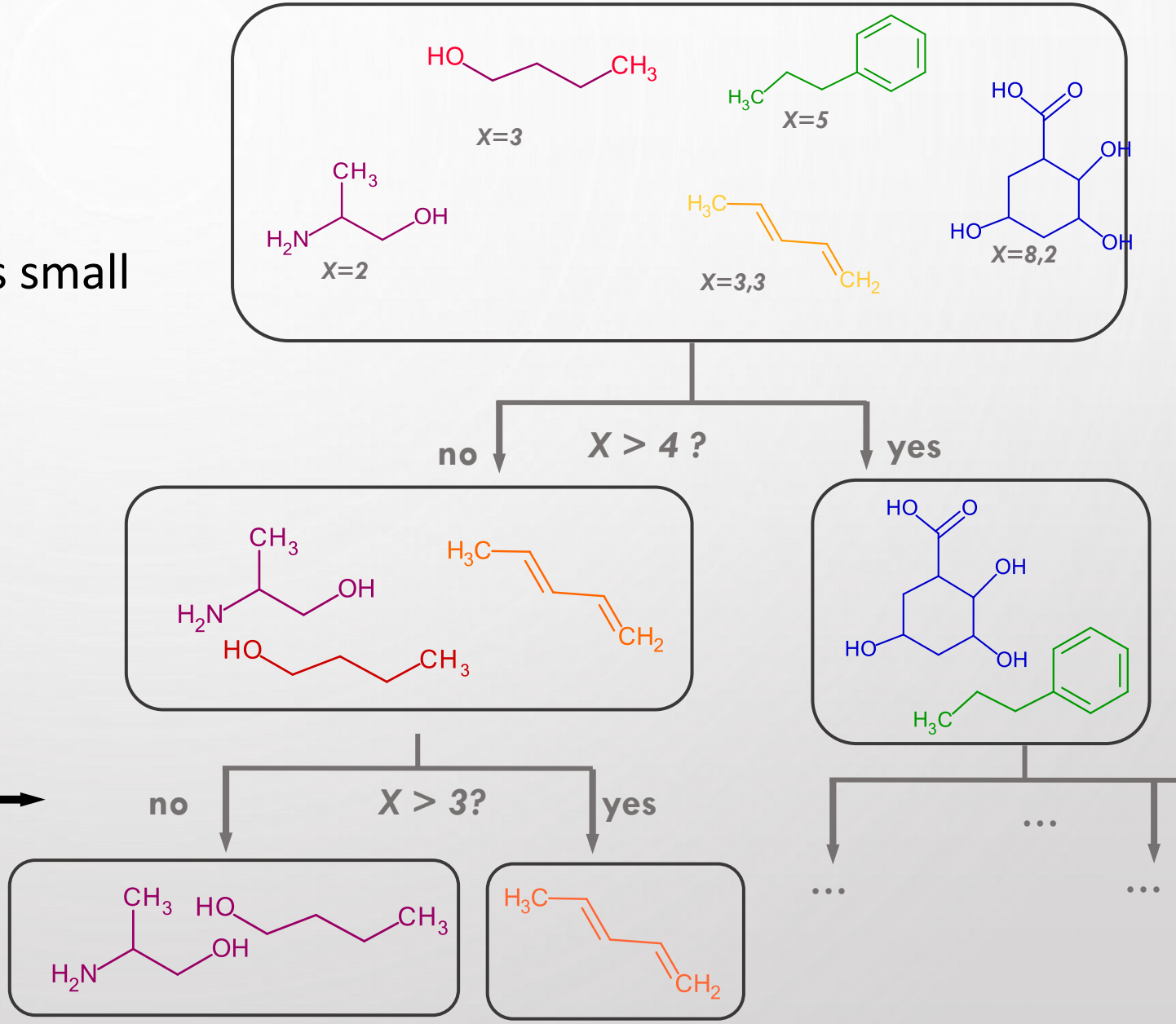
Random Forest: Yes

Decision tree (Regression)

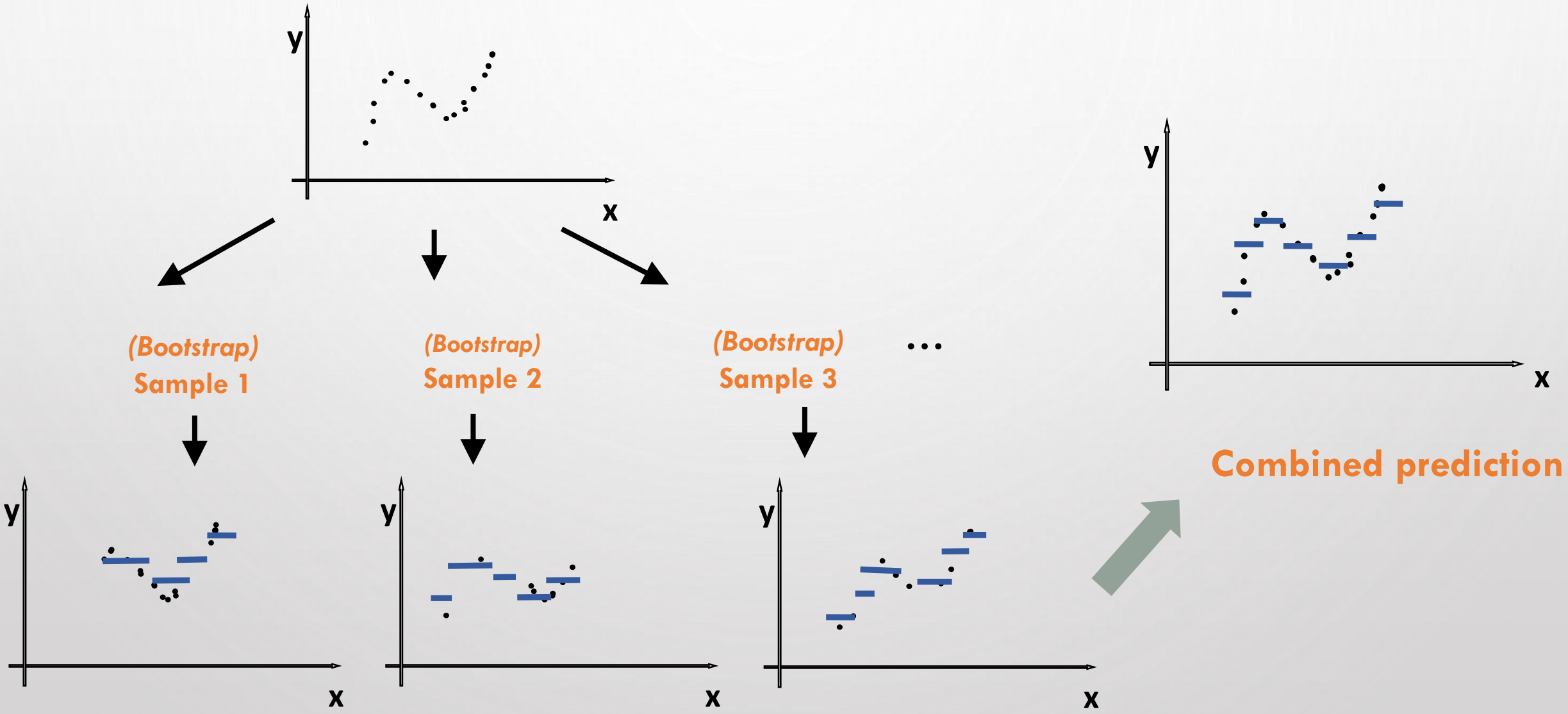


Stop when Loss is small enough

Choose split and y_{pred} that minimizes error

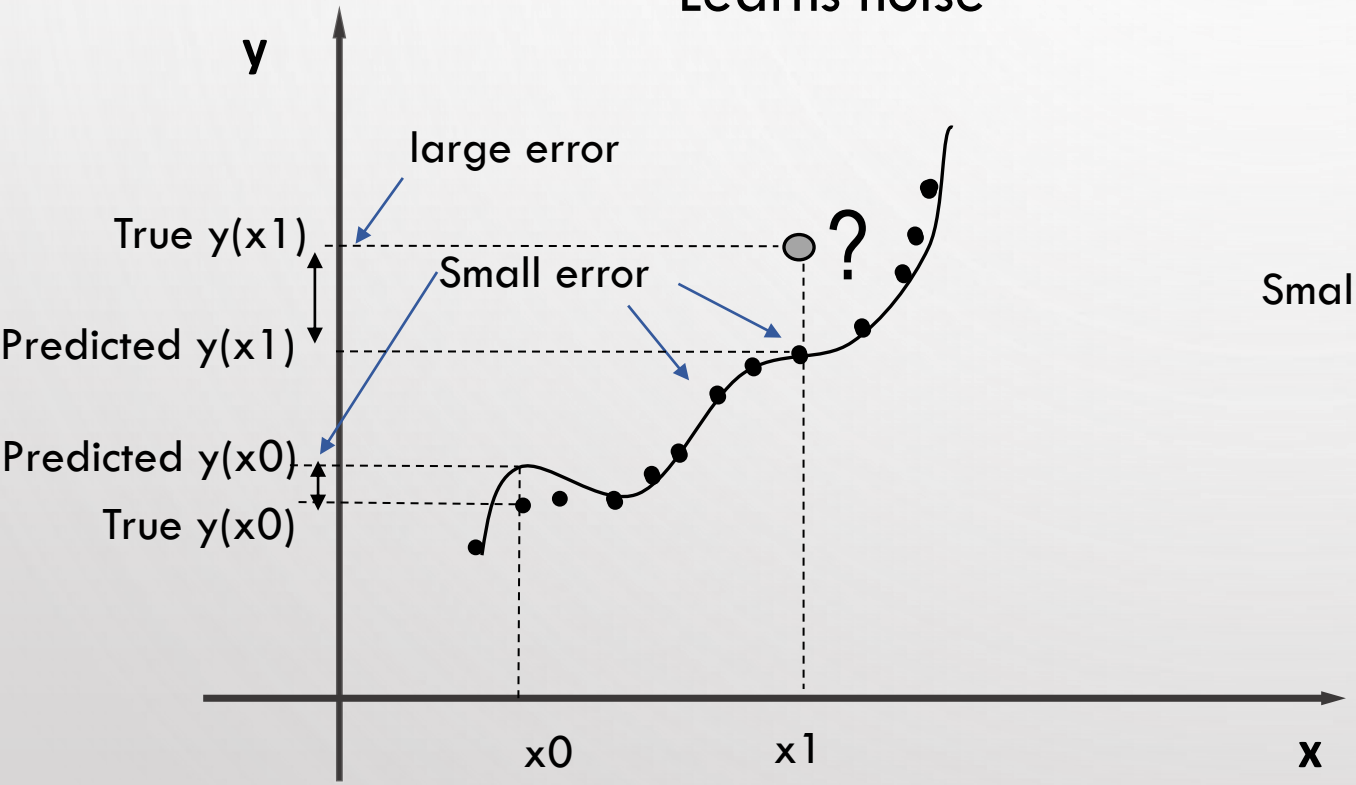


Random Forest (regression)

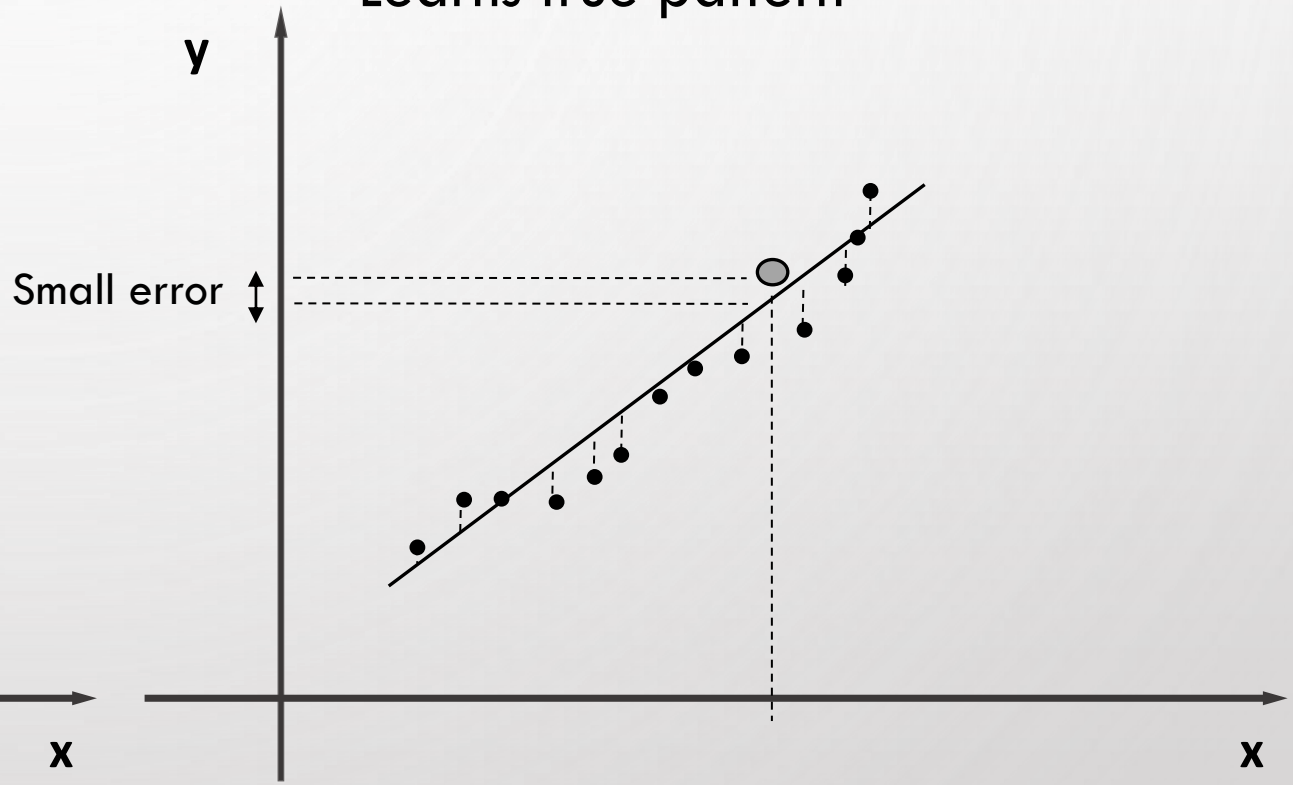


Overfitting

Overfitted model: weak predictor
Learns noise

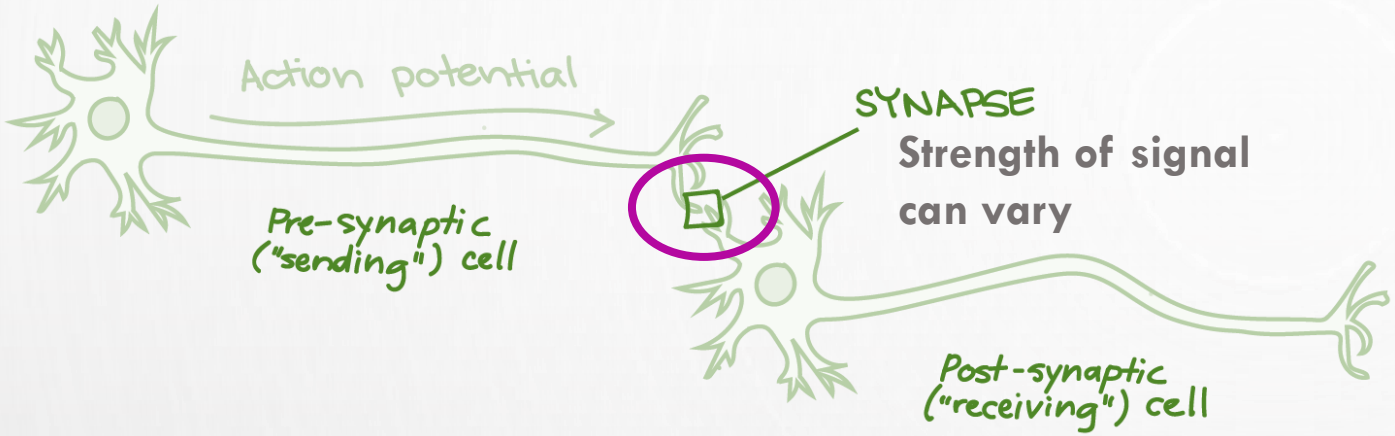


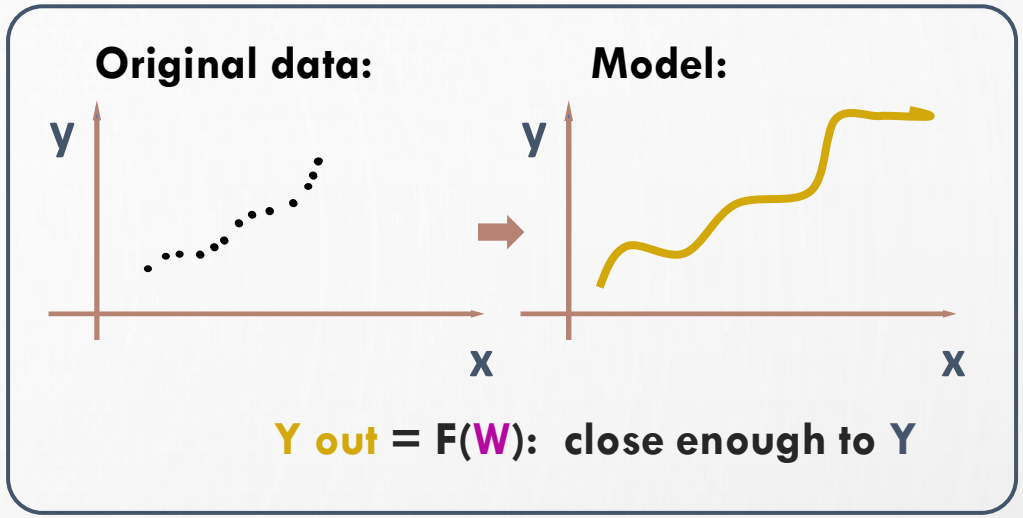
Strong predictor
Learns true pattern



NEURAL NETWORKS & DEEP LEARNING

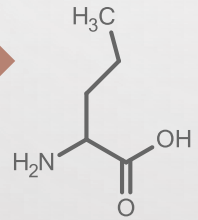
Brain: source of inspiration





Simple model

Activation inside neuron

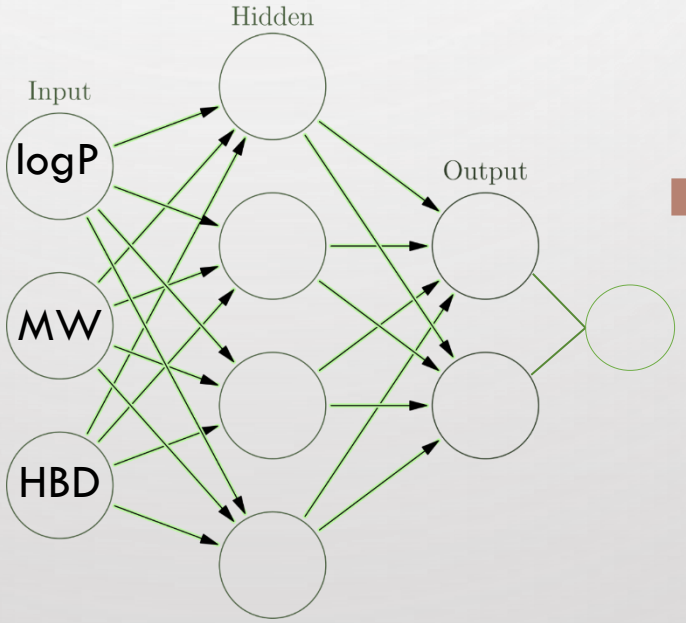


Neurons = nodes

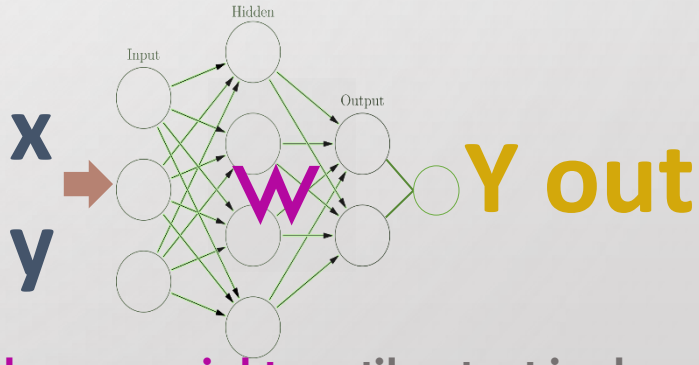
Synapses = weights

WEIGHT = value to multiply signal by

Combine neurons in network:

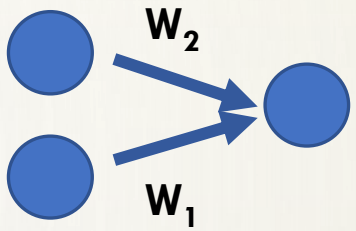


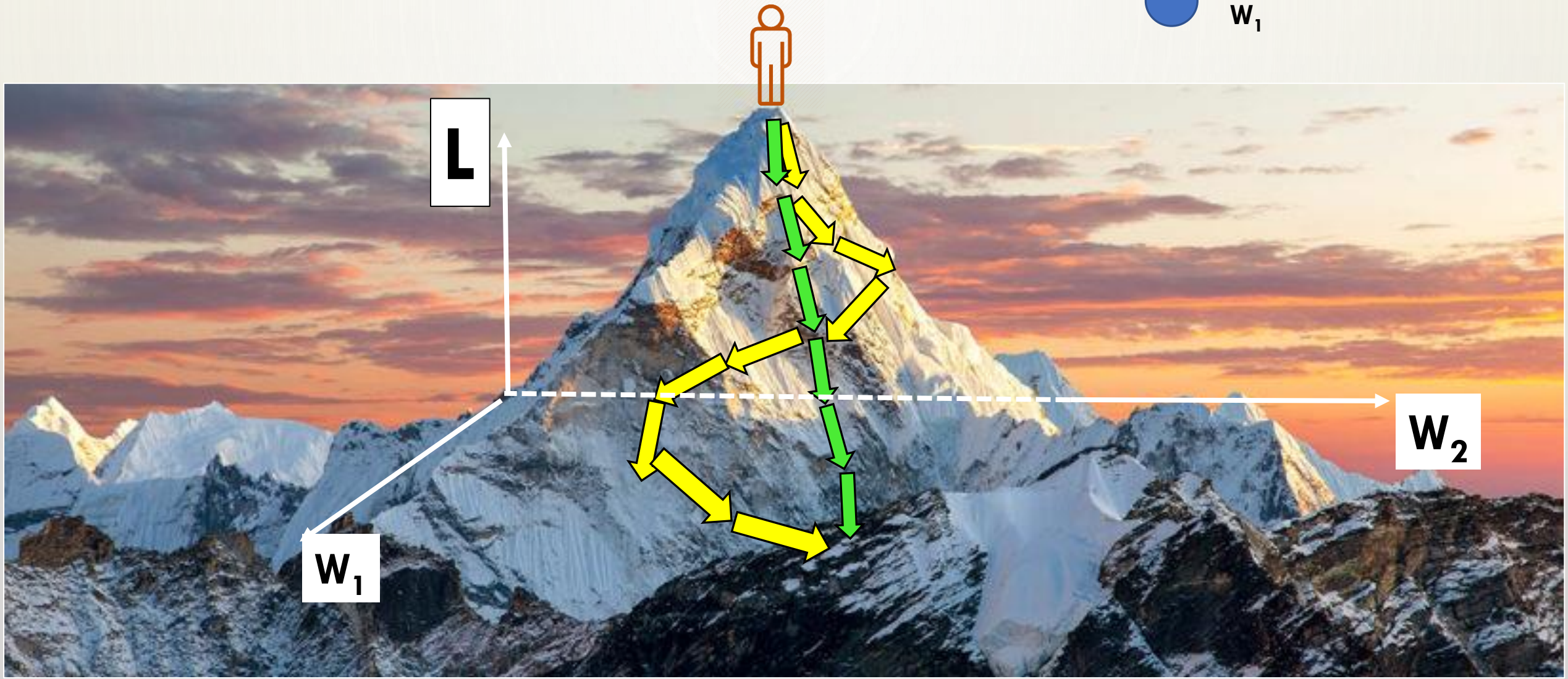
Feed the data to network:



change weights until output is close enough to expected

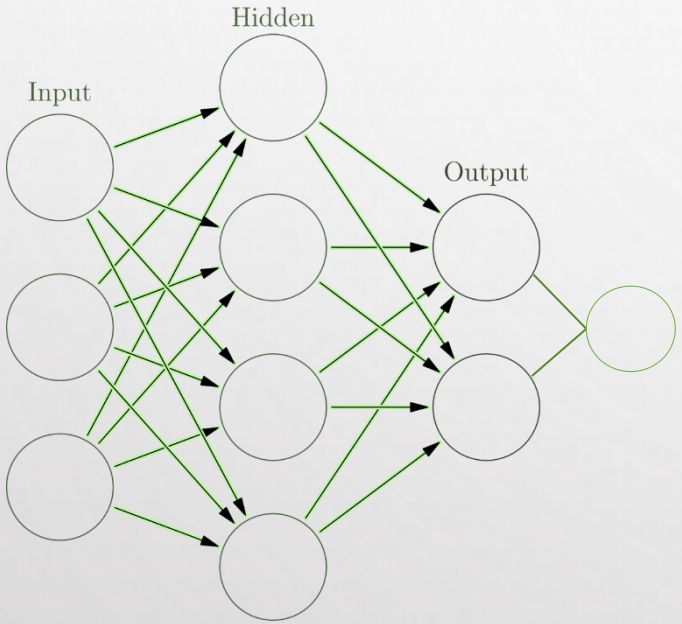
GRADIENT DESCENT





NETWORK ARCHITECTURES

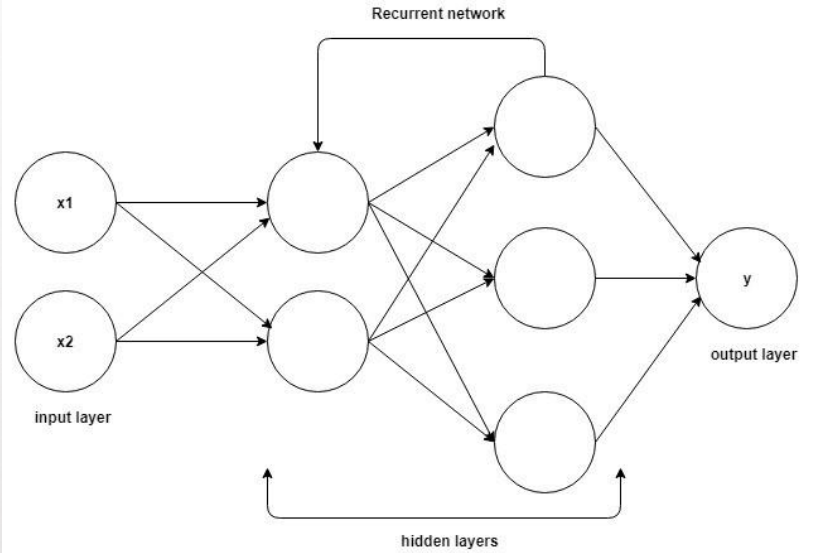
- MULTILAYER PERCEPTRON, MLP



- RECURRENT NETS: THERE ARE BACKWARD CONNECTIONS

SMILES strings as input

CCCN(cccccc)



Nowadays superseded by **Transformers**

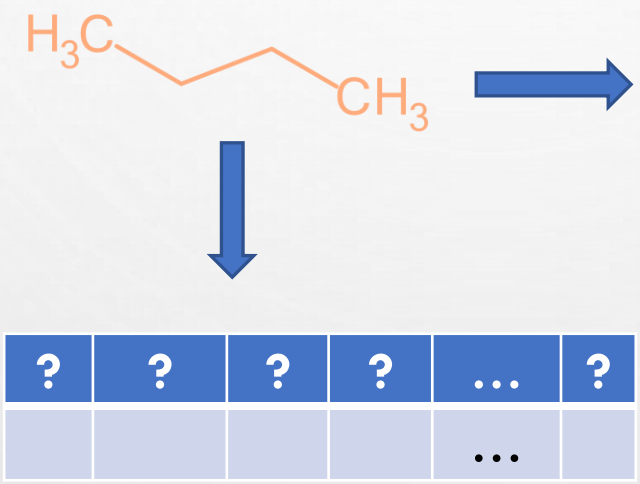
NETWORK ARCHITECTURES

traditional descriptors

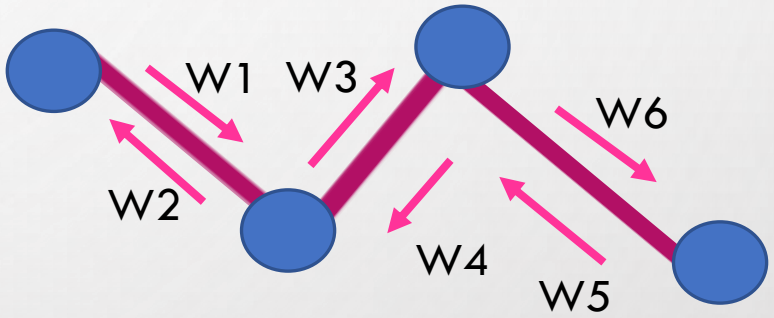
D_1	D_2	D_3	D_4	…	D_N
1	0	1	0	…	1

modelling

Graph convolutional networks



Graph

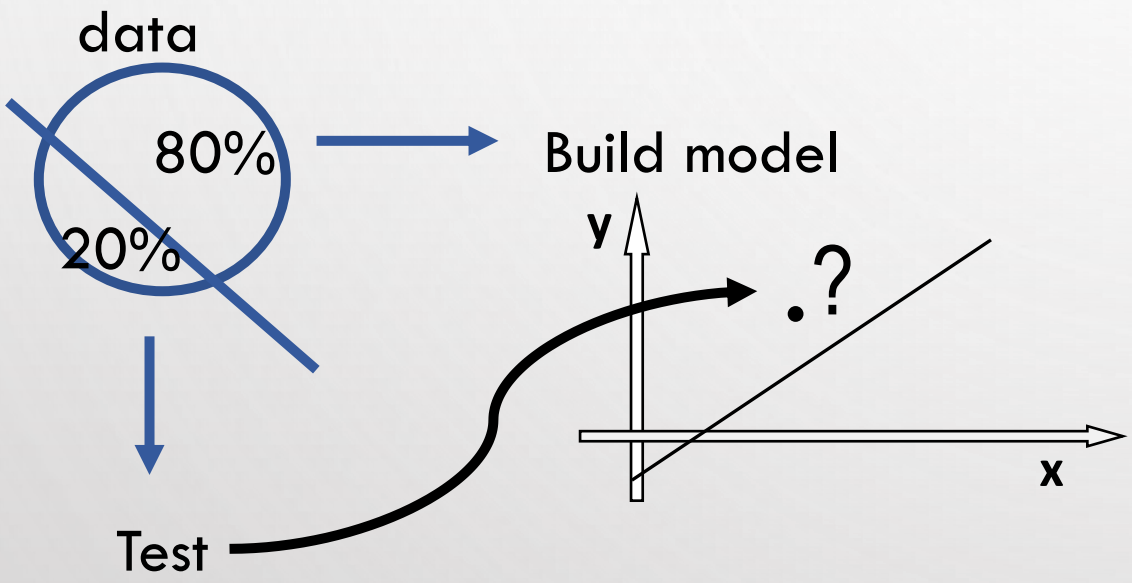


Modelling: find best W

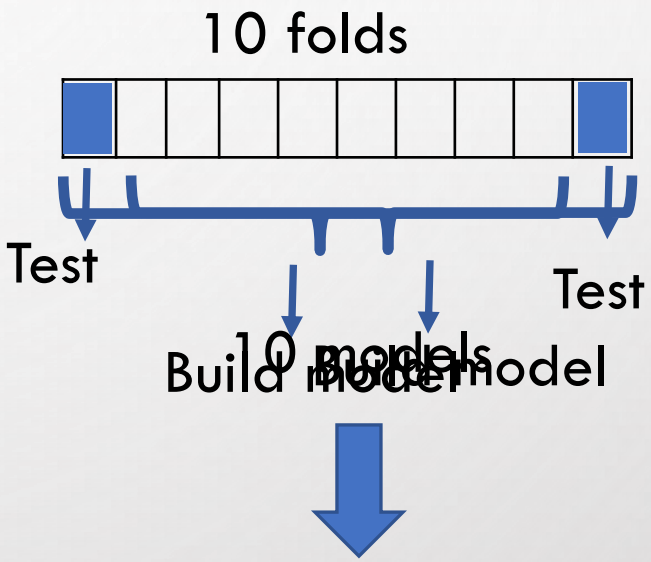
4. Model Validation, performance

Model Validation

1. EXTERNAL TEST



2. k-fold cross-validation



predictions of different folds are combined to calculate the final y_{pred}

MODEL PERFORMANCE/METRICS

MODEL PERFORMANCE: CLASSIFICATION

Confusion matrix	Predicted 1	Predicted 0
observed 1	true positive (TP)	false negative (FN)
observed 0	false positive (FP)	true negative (TN)

$$\text{Specificity} = \frac{TN}{TN + FP}$$

performance for class 0

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

performance for class 1

$$\text{accuracy} = \frac{TP + TN}{N}$$

Overall accuracy

$$\text{Balanced accuracy} = \frac{\text{Specificity} + \text{Sensitivity}}{2}$$

Used for imbalanced data

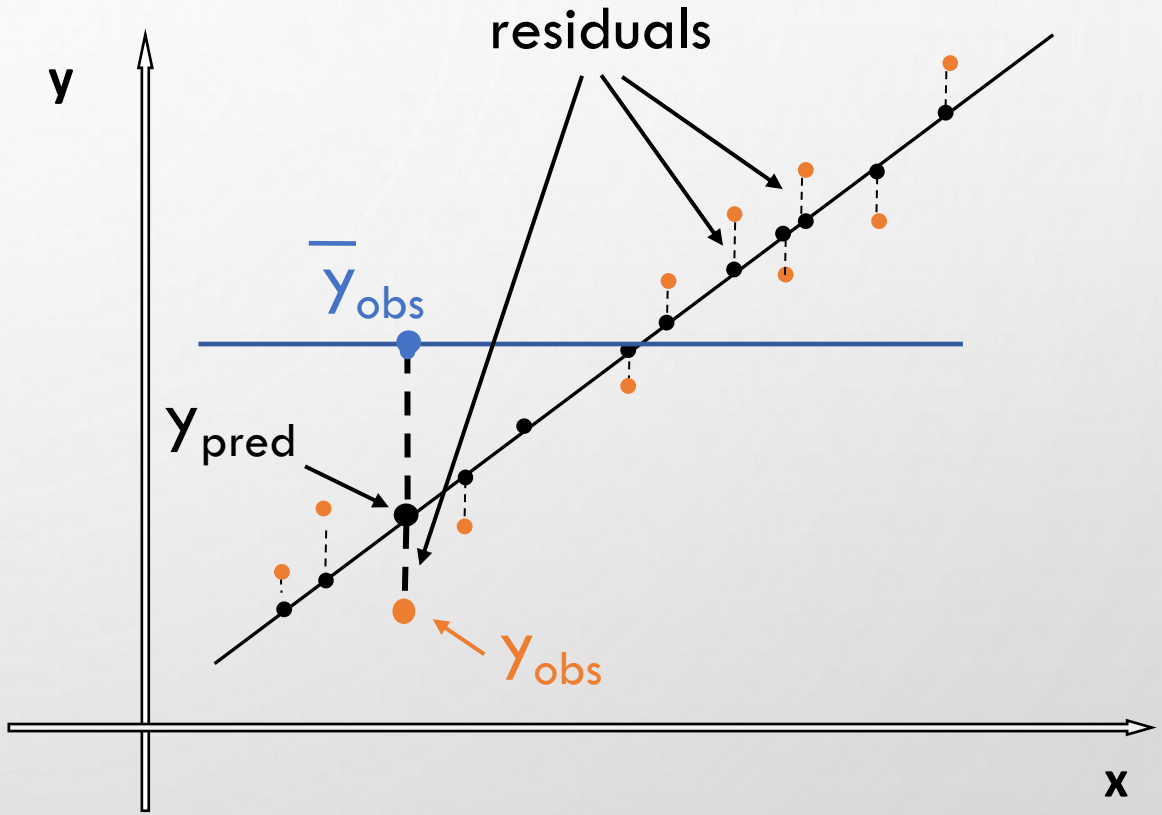
MODEL PERFORMANCE: REGRESSION

Determination coefficient (cross-validated)

$$Q^2 = 1 - \frac{\sum_i (y_{i,pred} - y_{i,obs})^2}{\sum_i (y_{i,obs} - \bar{y}_{obs})^2}$$

Root mean squared error (cross-validated)

$$RMSE = \sqrt{\frac{\sum_i (y_{i,pred} - y_{i,obs})^2}{N}}$$



5. APPLICABILITY DOMAIN

Moldel: relationship between structure and boiling point of alkanes

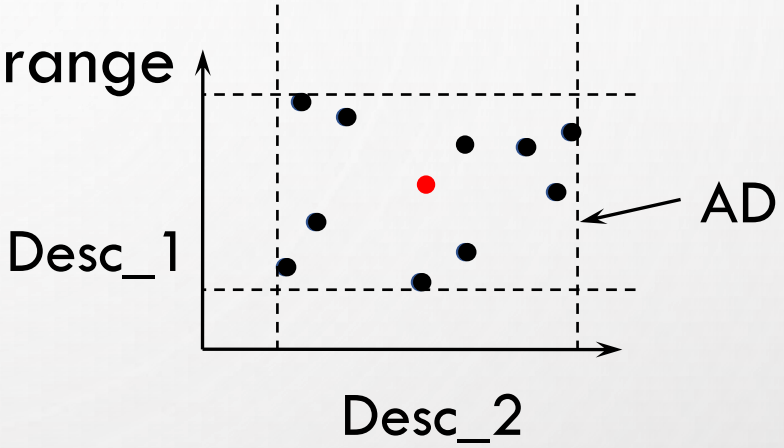
$$BP = F(structure)$$

Question: Can you predict boiling point for acetic acid?

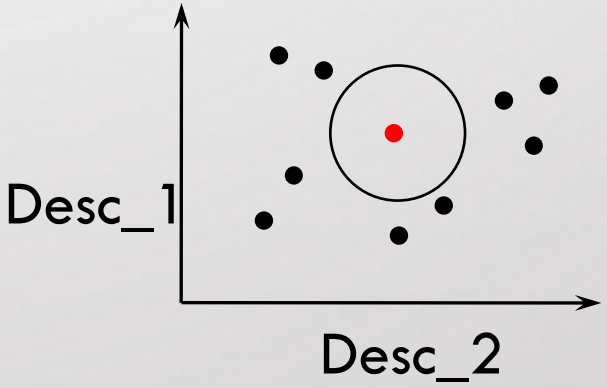
Model applicability domain

APPLICABILITY DOMAIN

- Bounding box - based on descriptor range
  - internal regions are usually empty, especially if the number of descriptors is large
  - it doesn't take into account descriptor correlation



- Distance from training set compounds in descriptor space



NOT COVERED HERE..

- METHODS: SUPPORT VECTOR MACHINES, GRADIENT BOOSTING MACHINES
- INVERSE QSAR: ACTIVITY → TO STRUCTURE = MOLECULE GENERATORS
- HYPERPARAMETERS: HOW MANY TREES ARE GOOD IN RANDOM FOREST? WHAT TYPE/SIZE OF NEURAL NETWORK IS OPTIMAL? – SEE TUTORIAL!
- MODEL EXPLAINABILITY - INTERPRETATION

6. CONCLUSIONS

CONCLUSIONS

- QSAR is a mathematical methodology to produce models relating moelcules' structure to their activity
- Fast and easy way to screen large chemical libraries
- Doesn't need structural target information, relies purely on ligand
- Requires carefull validation - models can have low error during training, but in real-world task perform poorly
- Applicability domain should be considered
- Work best when combined together with docking, pharmacophore models etc.

THANK YOU!