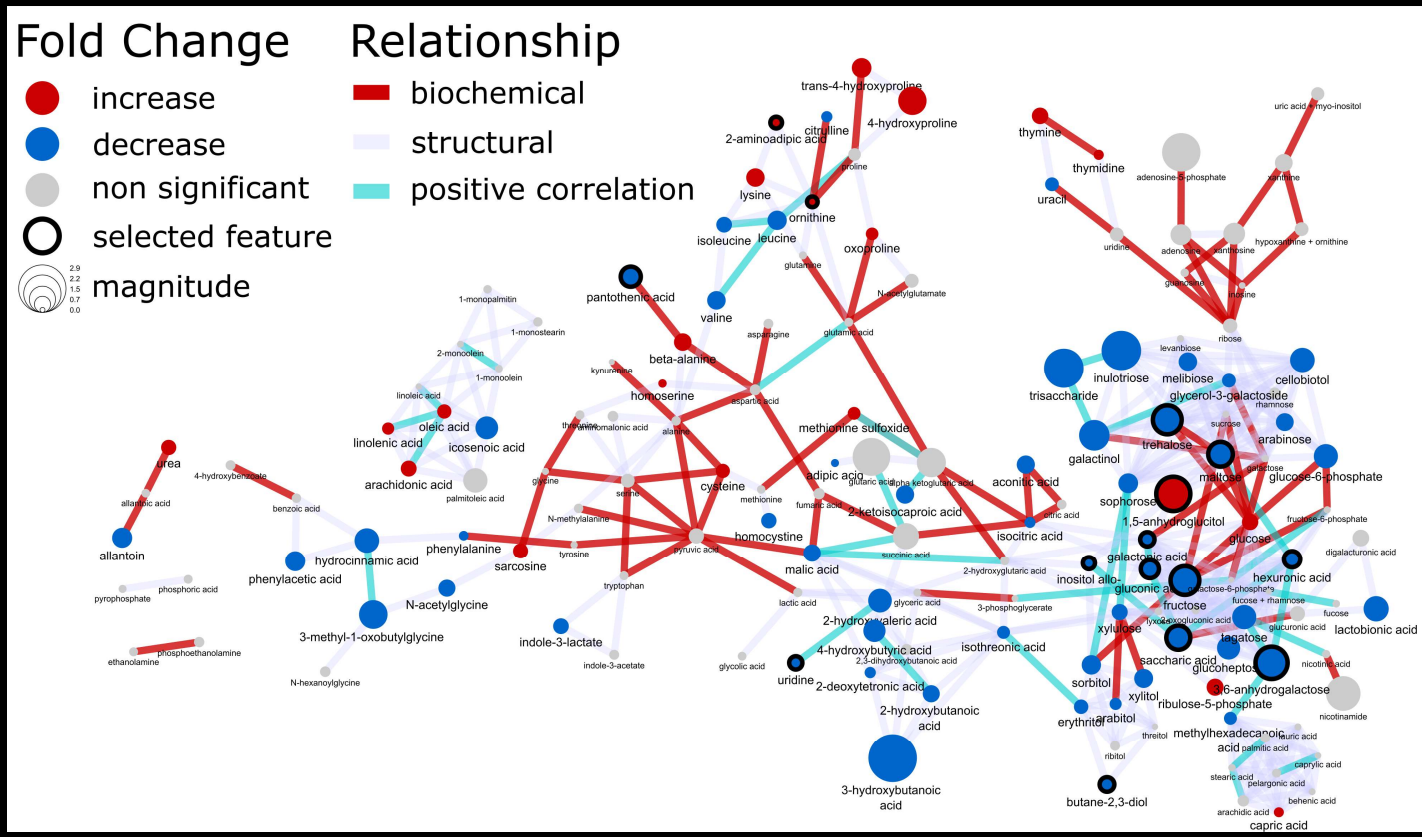
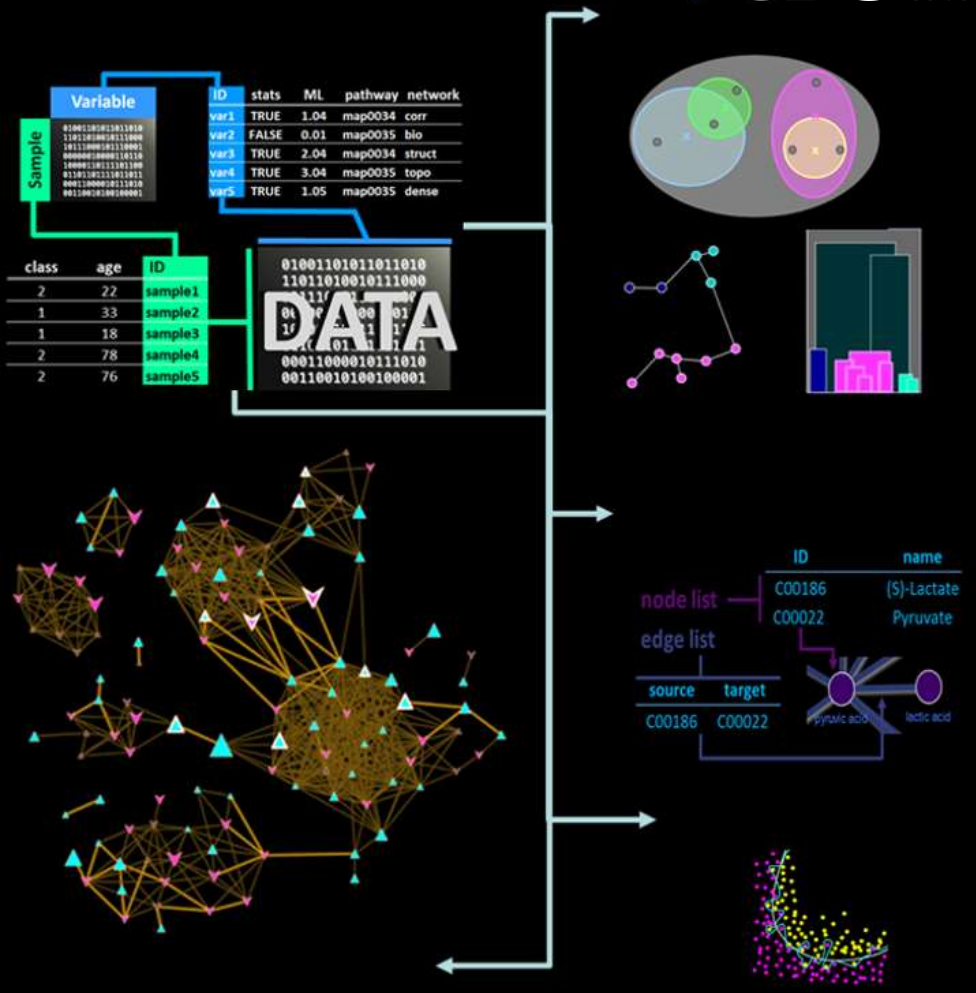
Welcome to network mapping 101

In the following course you will learn how to integrate statistical, multivariate and machine learning results within a publication quality biochemical network.



Tutorials

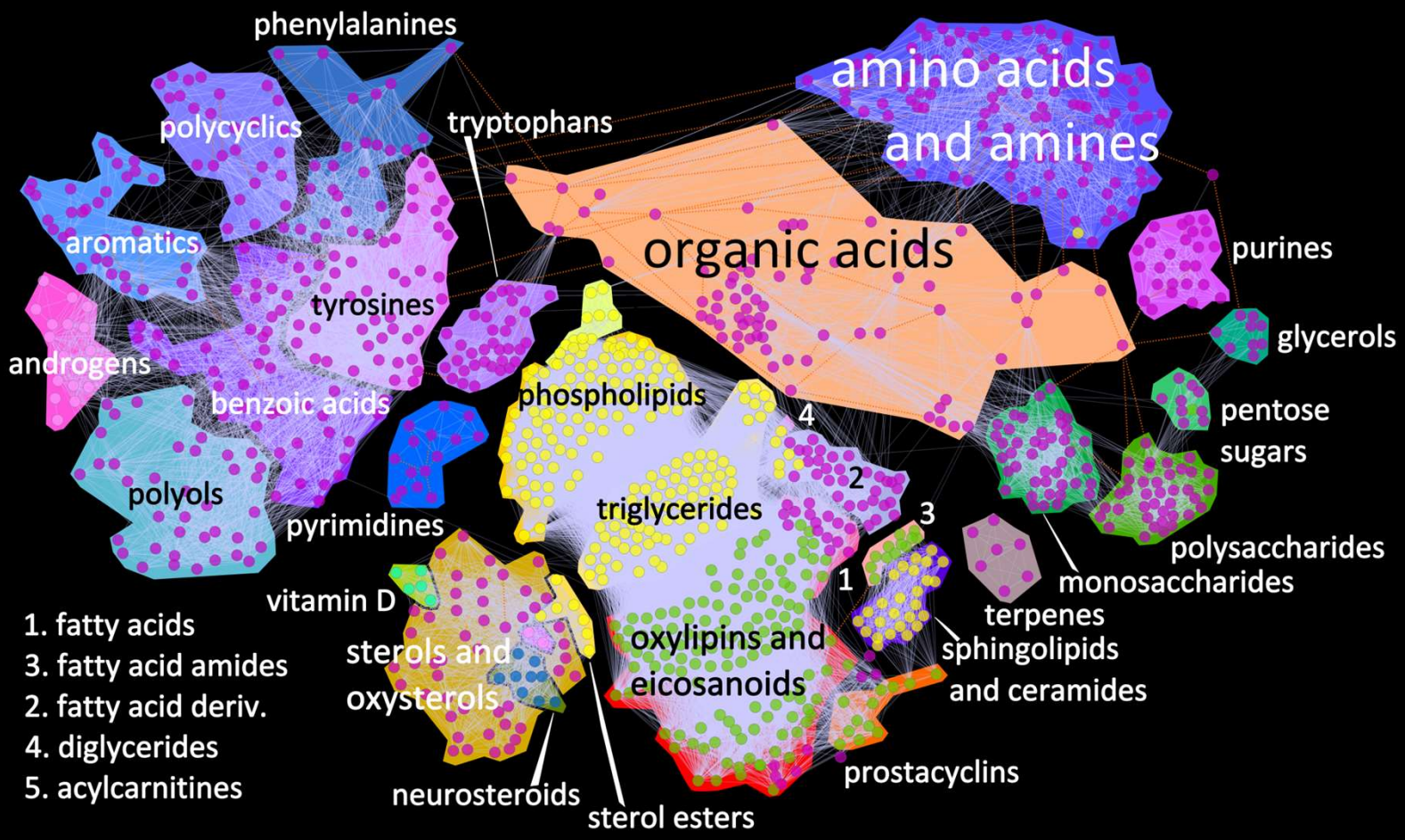
- Preparing raw data for analysis
- Statistical analysis
- Multivariate data exploration
- Supervised clustering
- Machine learning
  - classification
  - model validation
  - feature selection
- Network analysis
  - biochemical
  - structural similarity
  - correlation
- Network mapping – putting it all together



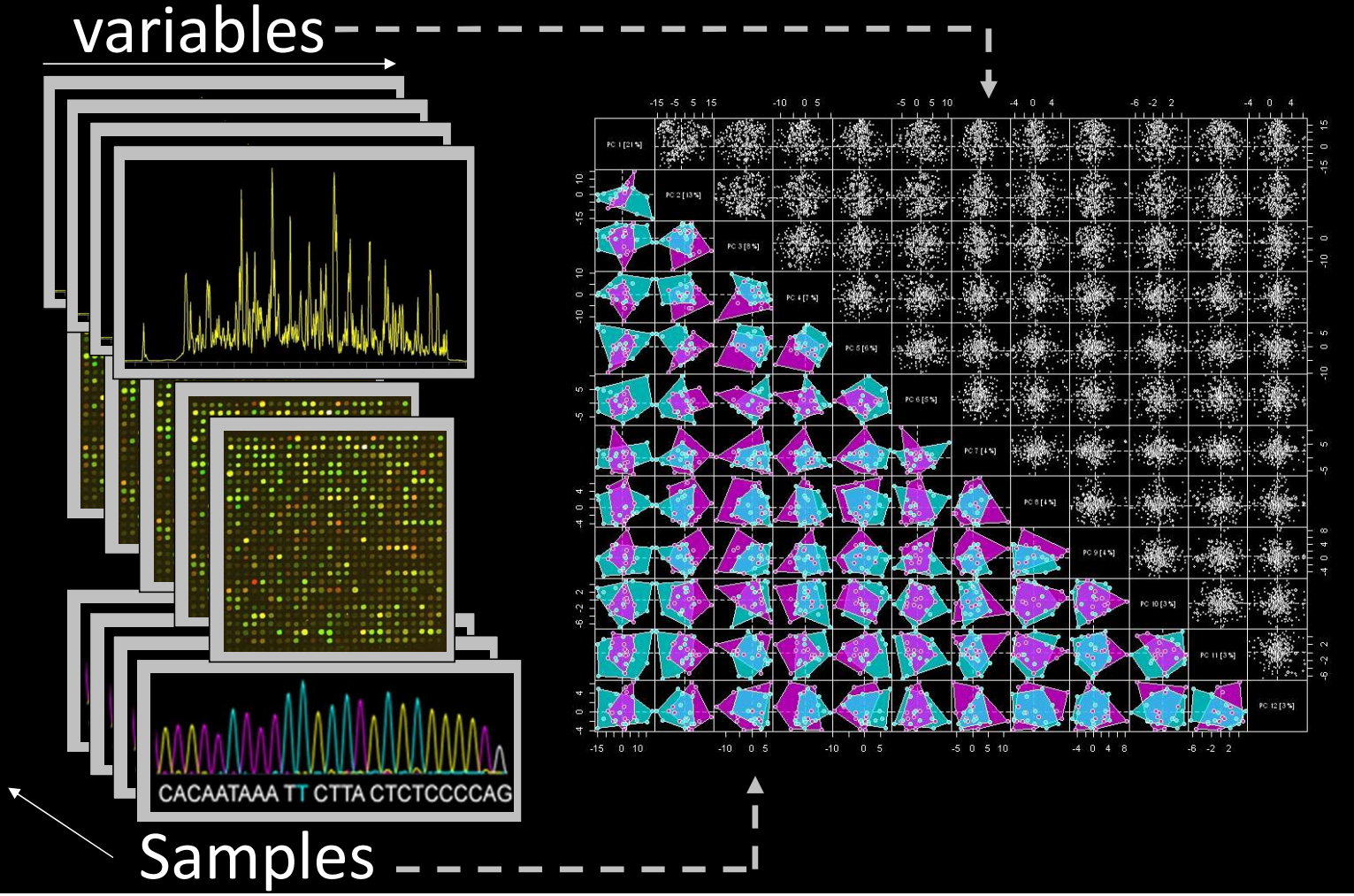
https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/#topics

Analysis at the metabolomic scale

CDS Creative Data Solutions

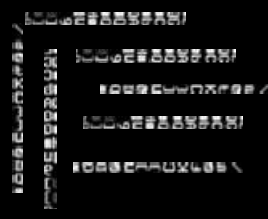


Integrate high-dimensional data



Identify what matters

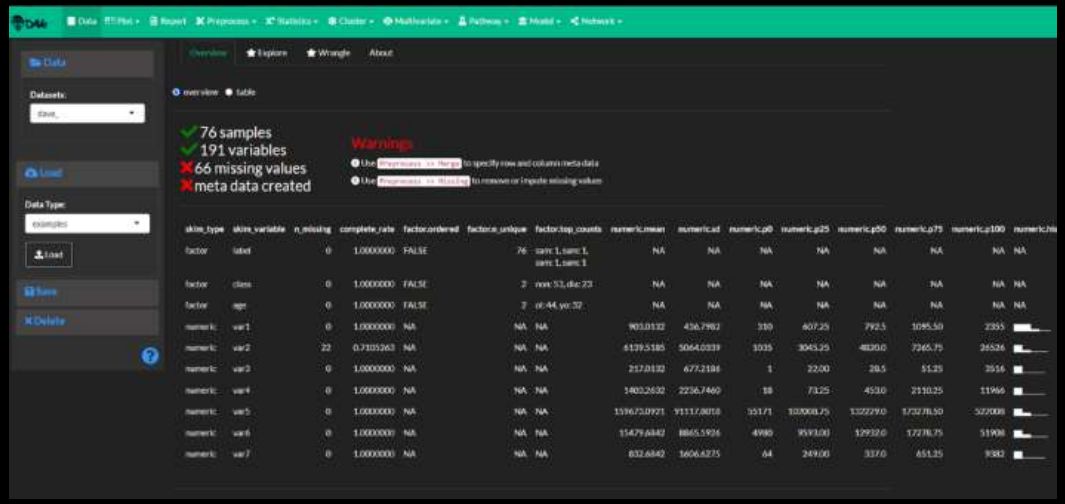
Creative Data Solutions

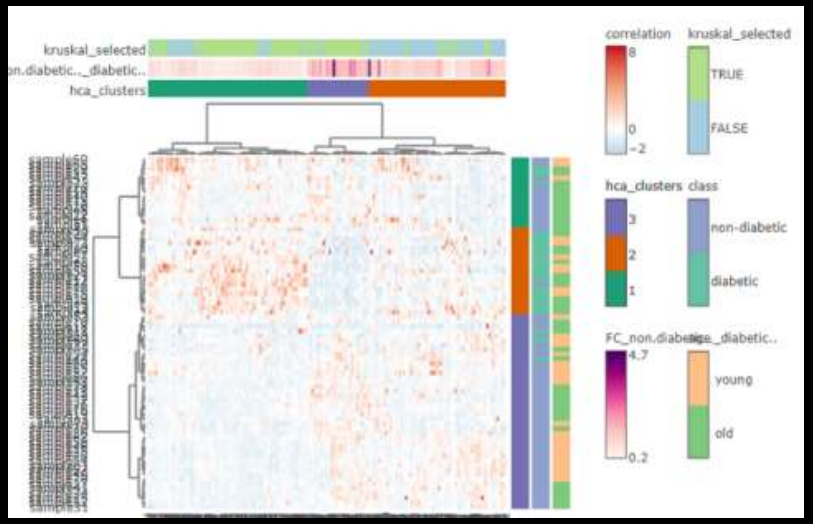
	Univariate	Multivariate	Predictive Modeling
Group 1			
Group 2			

ANOVA PCA PLS

Topics

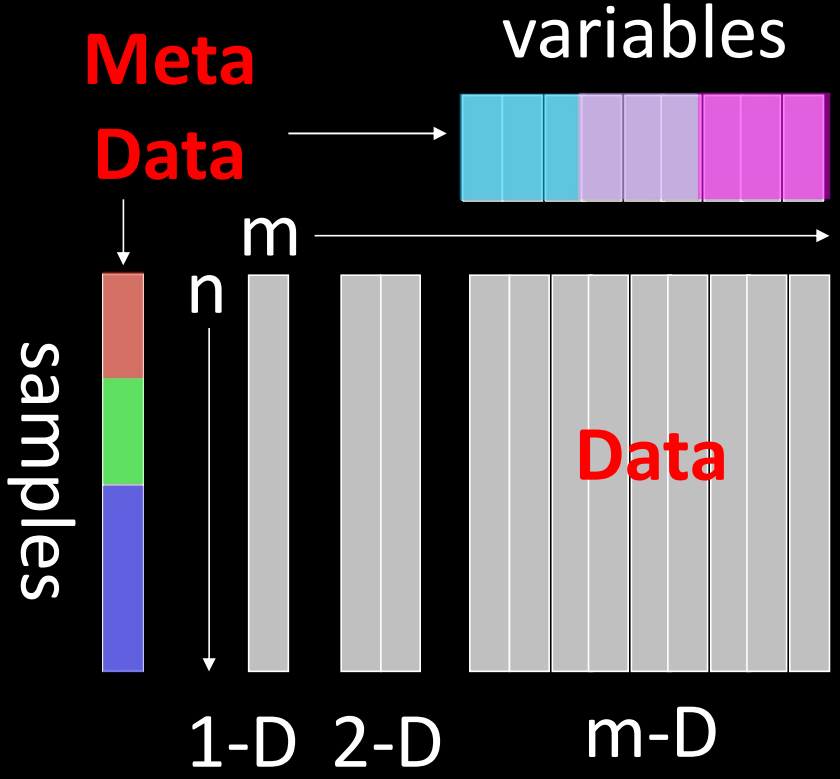
- Data preparation
- Differential expression
- Hierarchical Clustering
- Principal Components Analysis (PCA)
- Statistical analysis
- Machine learning
- Network analysis
- Network mapping



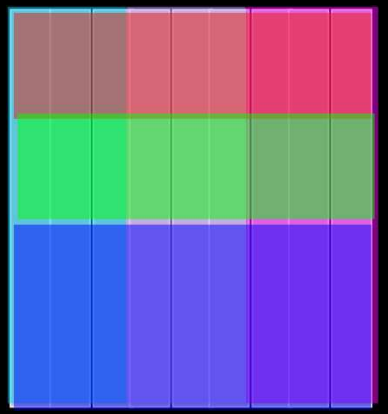


https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/#topics

How to think about data complexity



Experimental Design = complexity

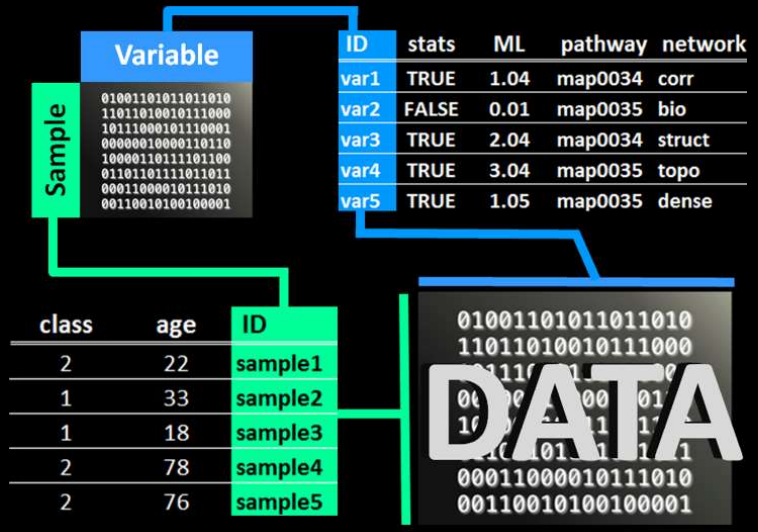


Variable # = dimensionality

Data preprocessing

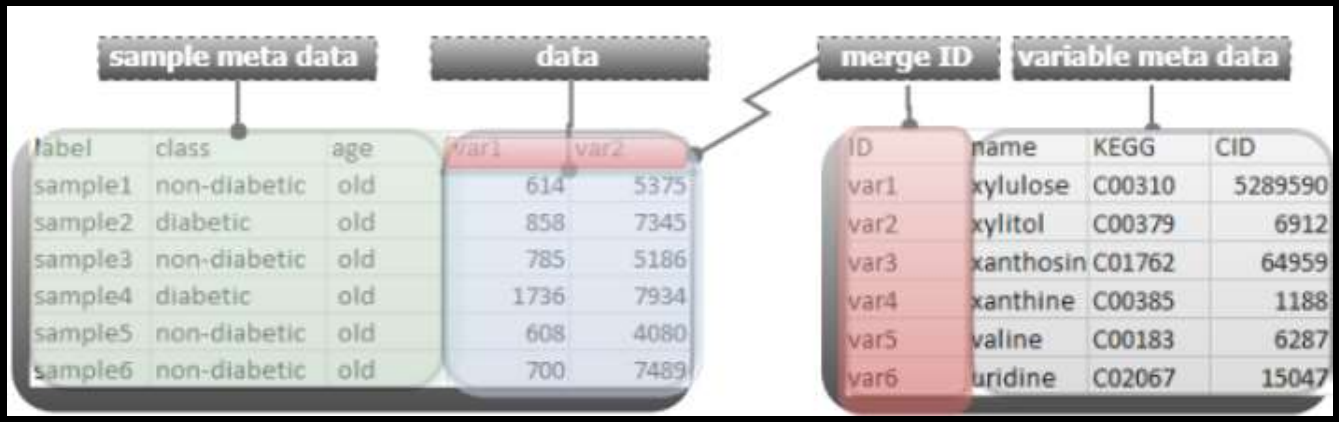
define

- data
- row meta data
- column meta data



remove and/or impute

- missing values



Your turn

Creative Data Solutions

Follow along with the following tutorial:

https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/preprocess/

Differential expression

compare

- class means

identify

- significant differences

visualize

- volcano plots
- violin plots

https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/statistics/

Your turn

Creative Data Solutions

Follow along with the following tutorial:

https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/statistics/

Hierarchical clustering (HCA)

group

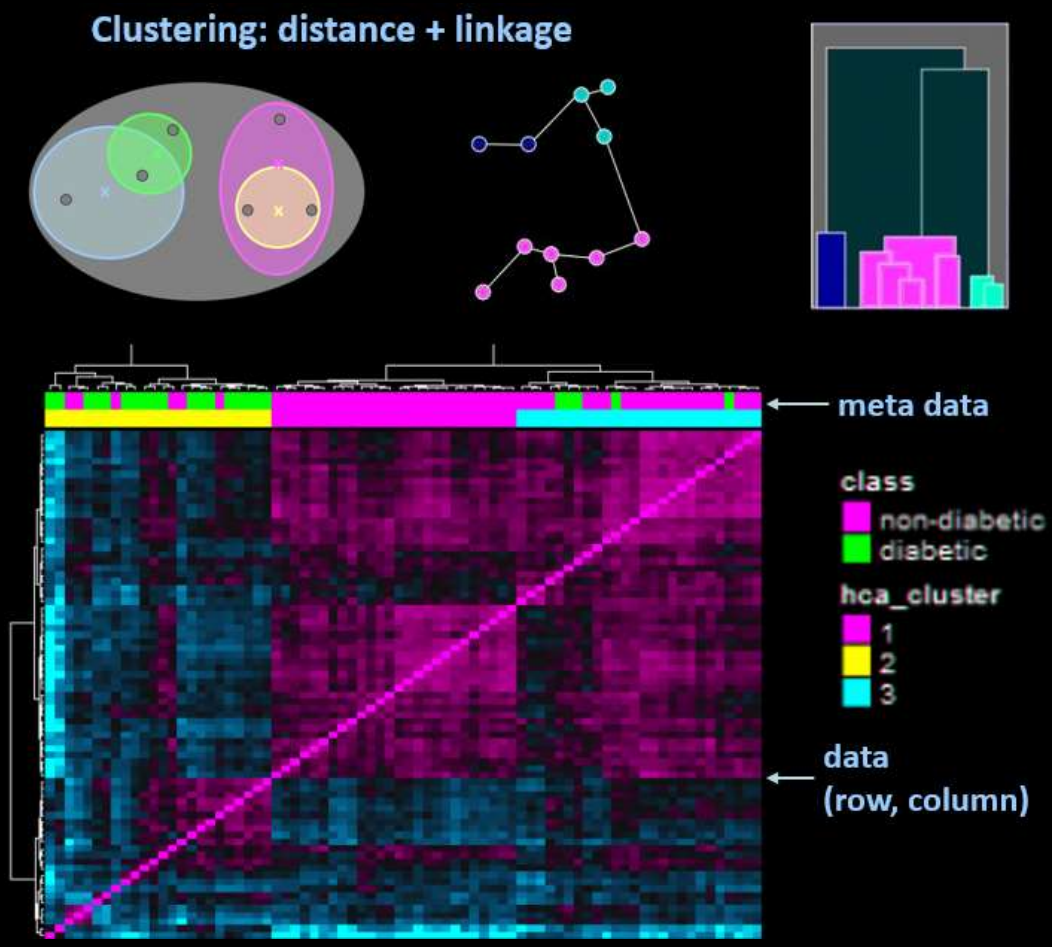
- samples and/or variables

define similarity

- correlation
- distance
- linkage

visualize

- heatmaps
- dendrograms



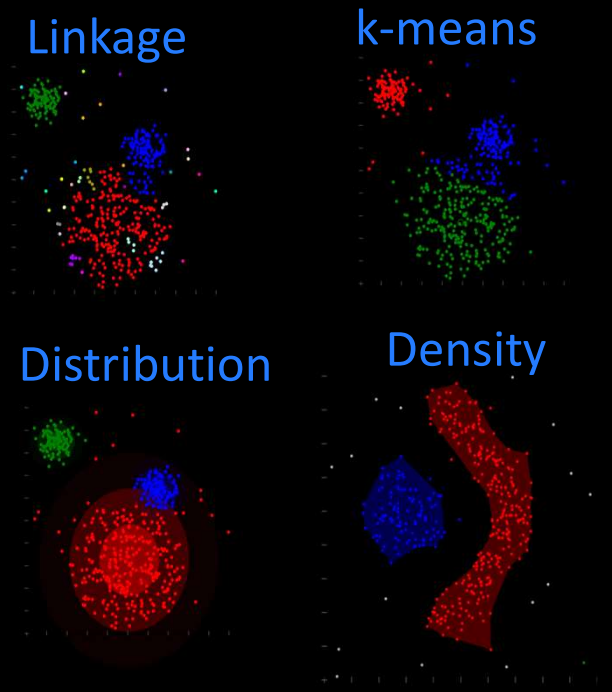
Clustering basics

Creative Data Solutions

Use the concept similarity/dissimilarity to group a collection of samples or variables

approaches

- hierarchical (linkage)
- non-hierarchical (k-NN, k-means)
- distribution (mixtures models)
- density (DBSCAN)
- self organizing maps (SOM)



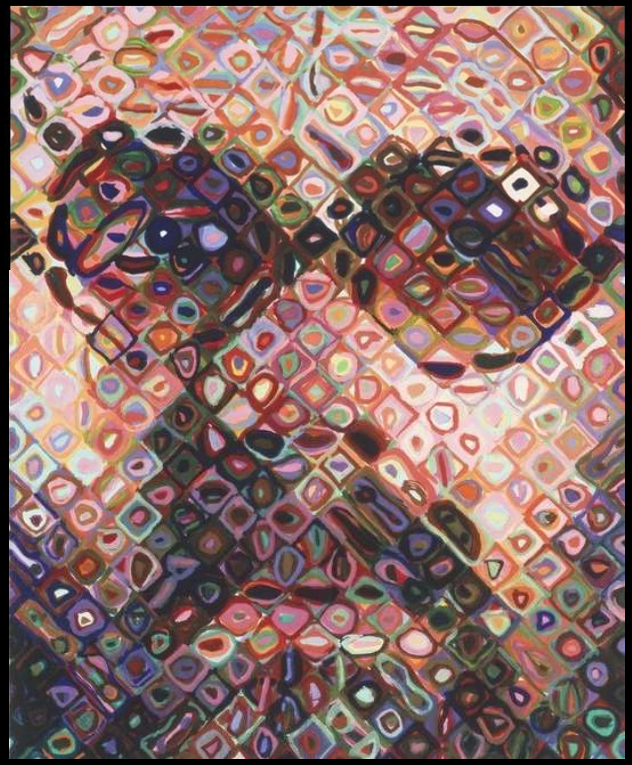
HCA goals

identify

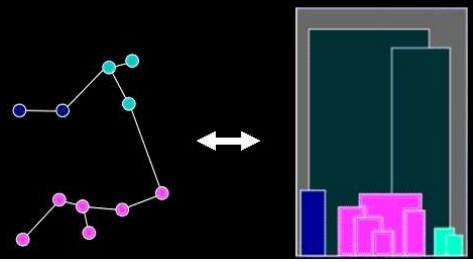
- patterns
- group structure
- relationships

evaluate and refine hypothesis

reduce complexity



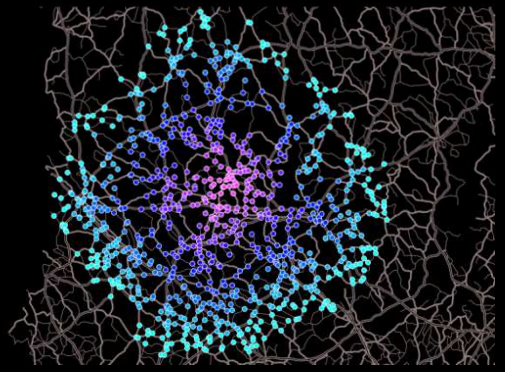
Artist: Chuck Close



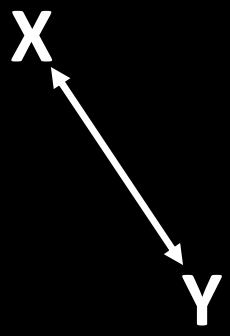
HCA methods

distance

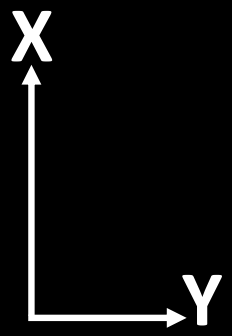
- defines "nearness" or similarity



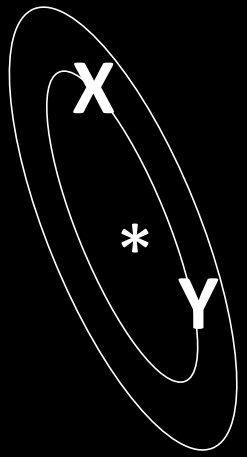
Euclidean



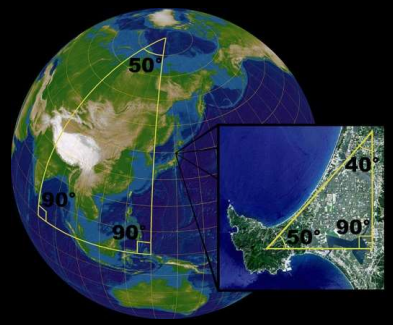
Manhattan



Mahalanobis



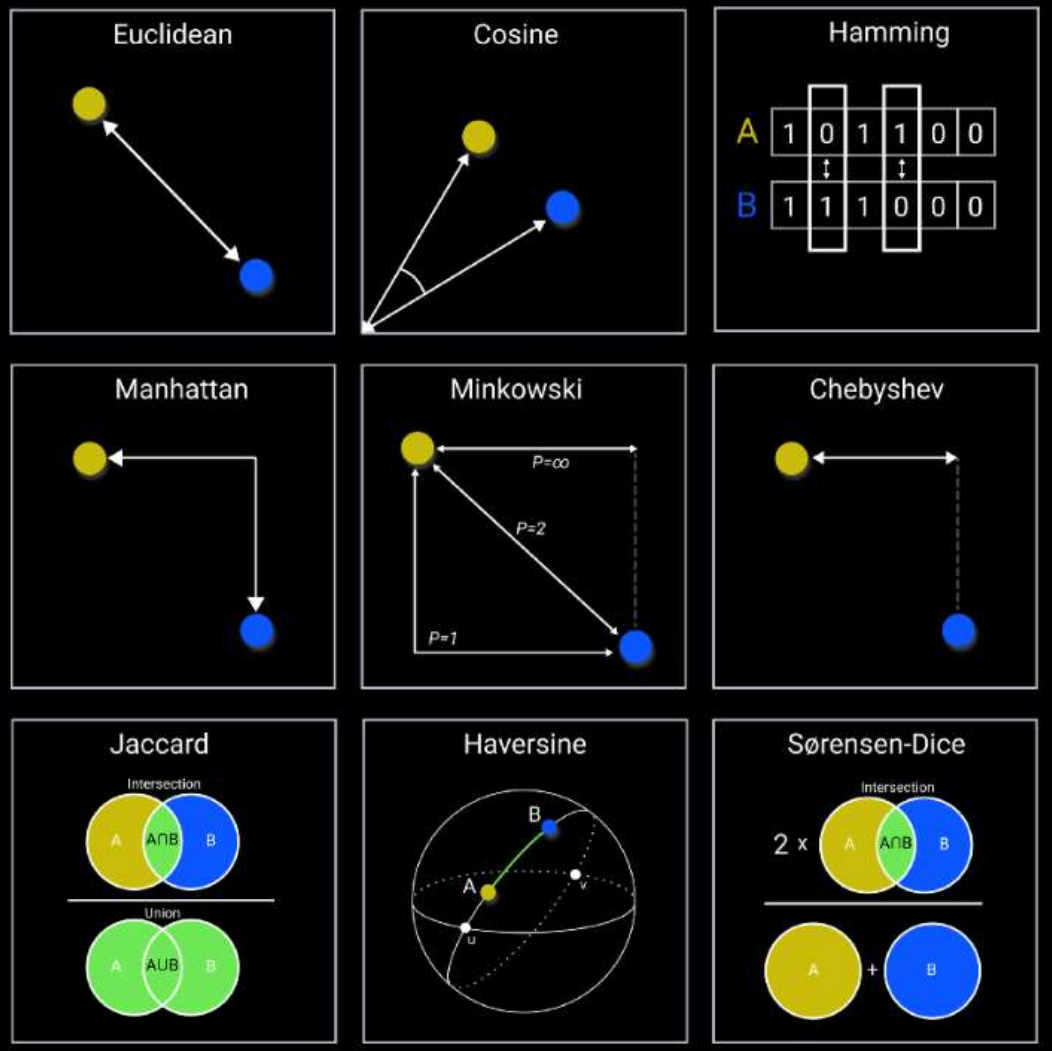
non-Euclidean



HCA methods

distance

- defines "nearness" or similarity

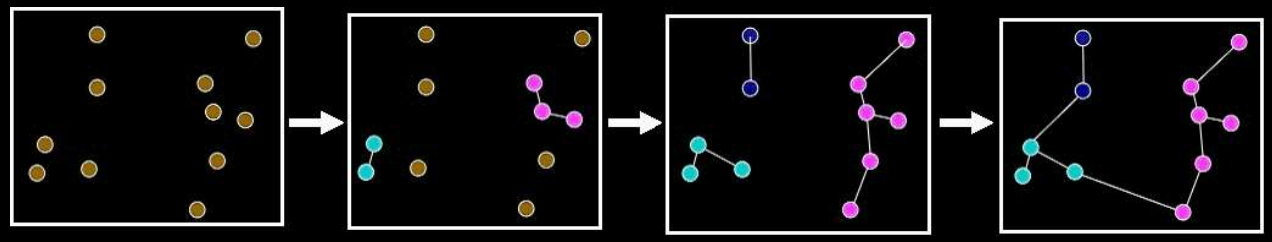


https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

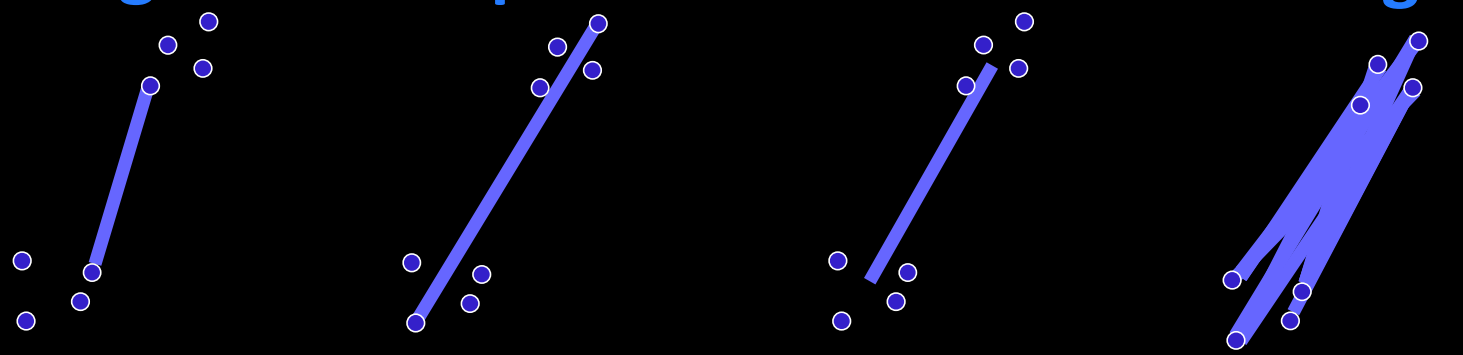
HCA methods

linkage or agglomeration

- how samples or variables are connected or grouped

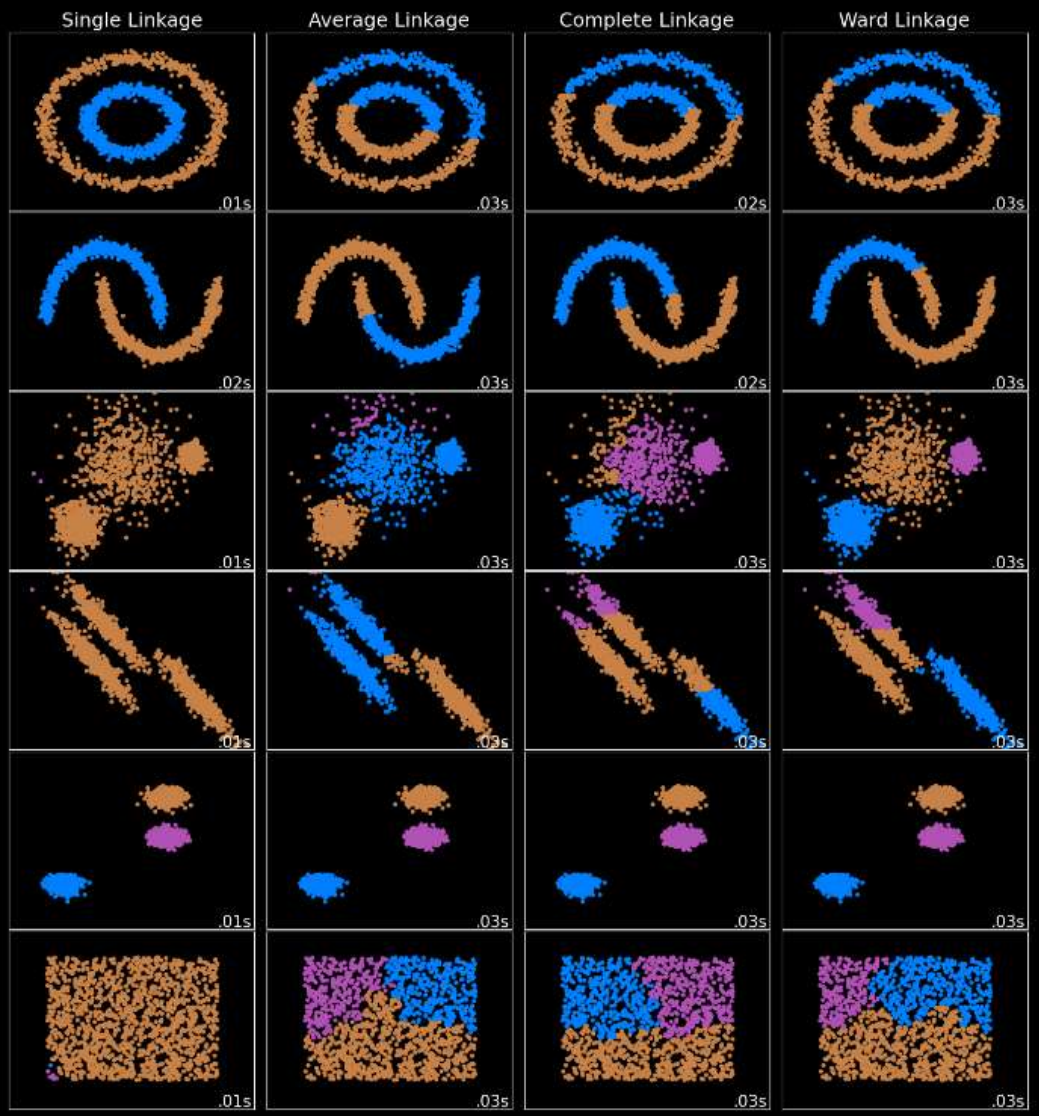


single complete centroid average



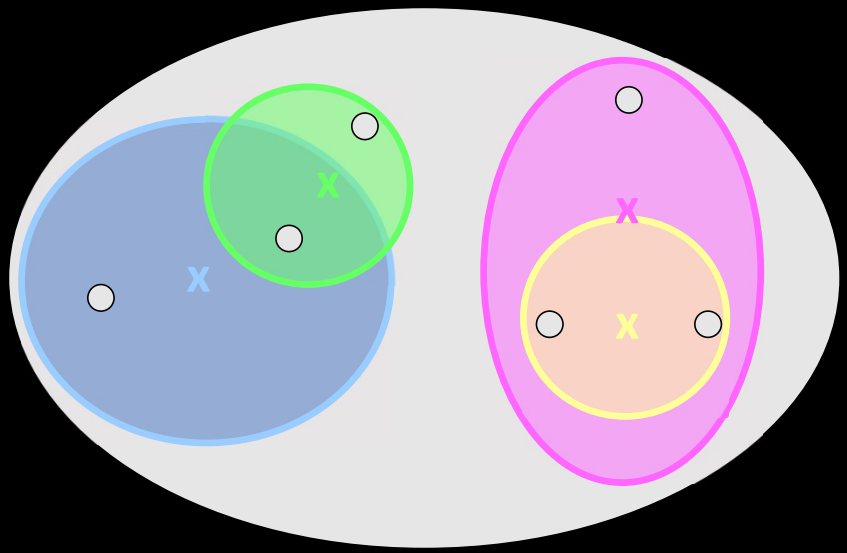
HCA methods

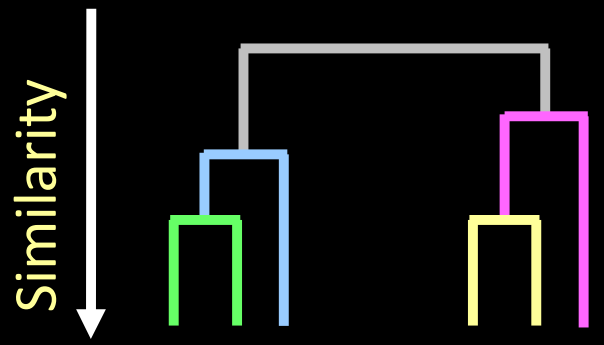
linkage or agglomeration



https://scikit-learn.org/stable/auto_examples/cluster/plot_linkage_comparison.html

HCA process

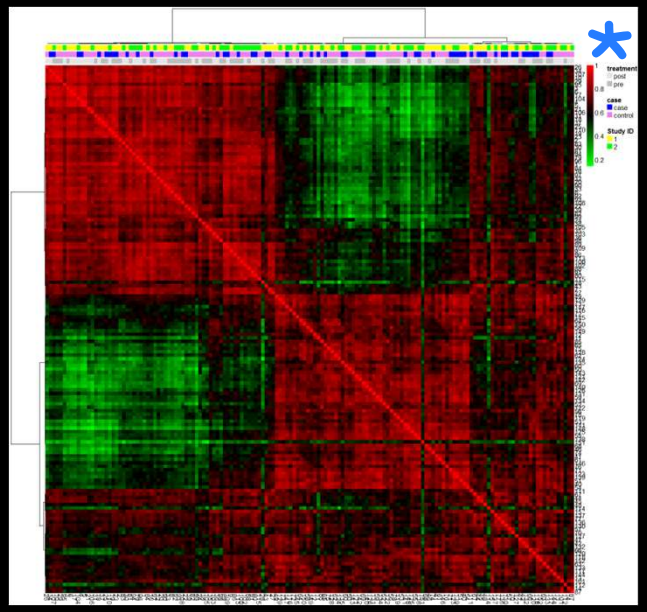




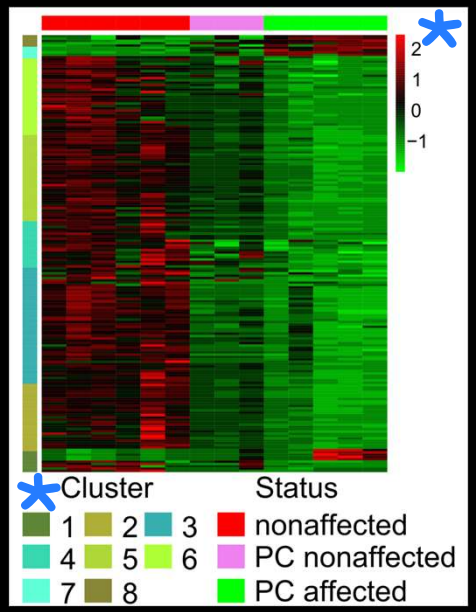
HCA interpretation

How does my metadata match my data structure? *

overview



confirmation



Your turn

Creative Data Solutions

Follow along with the following tutorial:

https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/clustering/#heirarchical-clustering

Principal Components Analysis (PCA)

CDS Creative Data Solutions

reduce

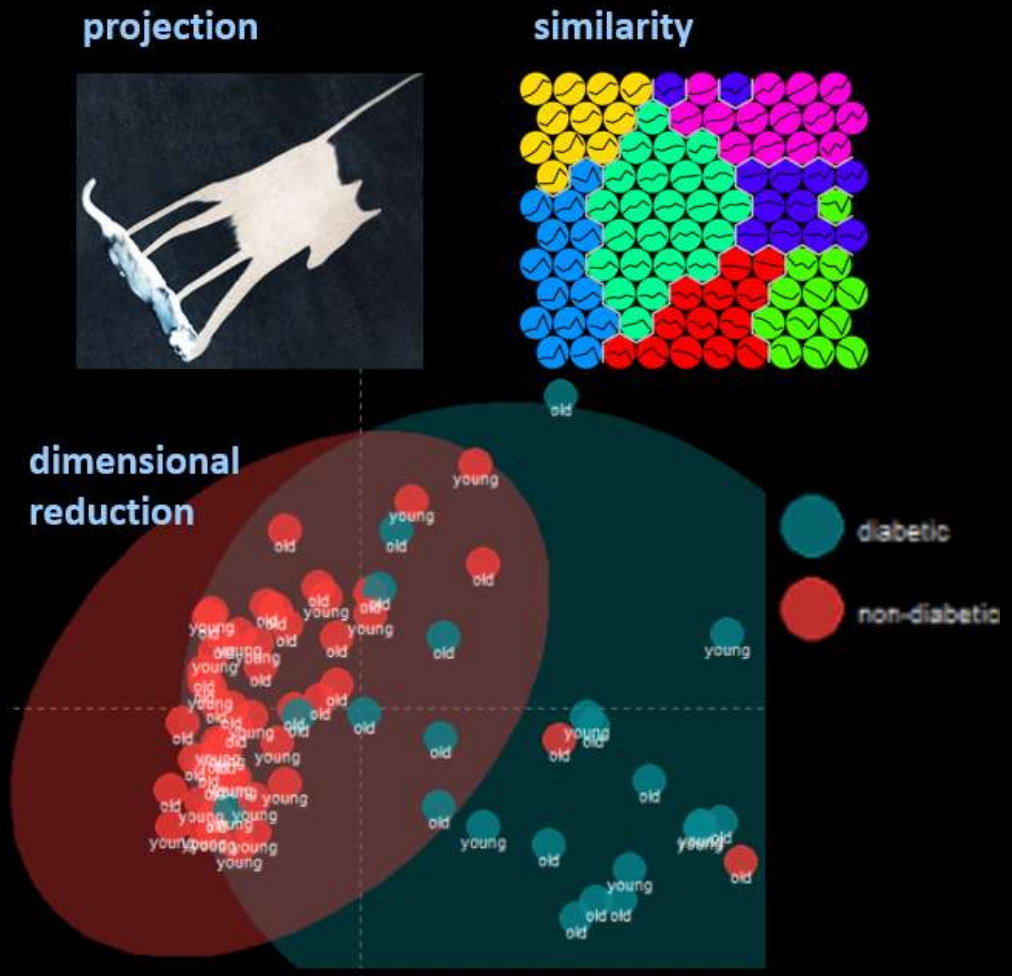
- dimensionality

maximize

- variance explained

visualize

- variance explained
- outliers
- sample scores
- variable loadings



https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/multivariate/#pca

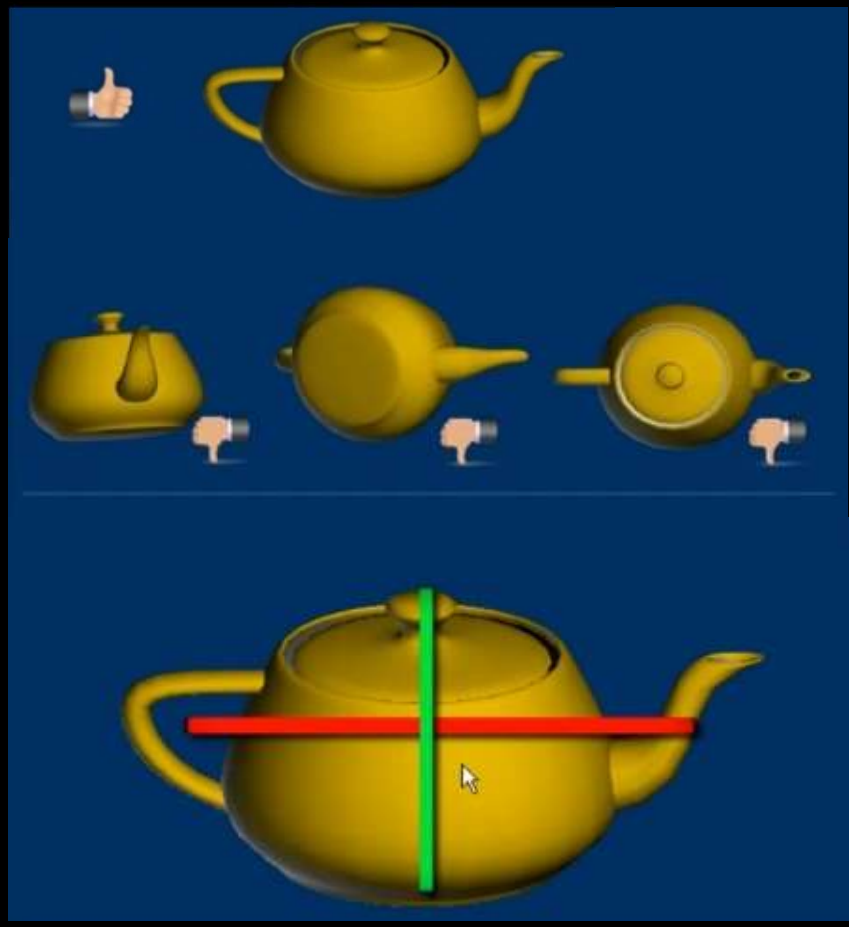
PCA goals

Principal Components (PCs)

- non-supervised
- projection of the data which maximize variance explained

results

1. eigenvalues = variance explained
2. scores = new coordinates for samples (rows)
3. loadings = linear combination of original variables

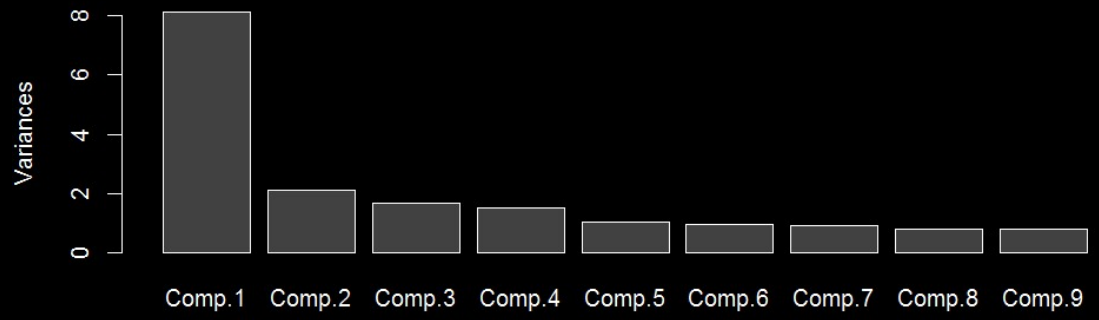


James X. Li, 2009, VisuMap Tech.

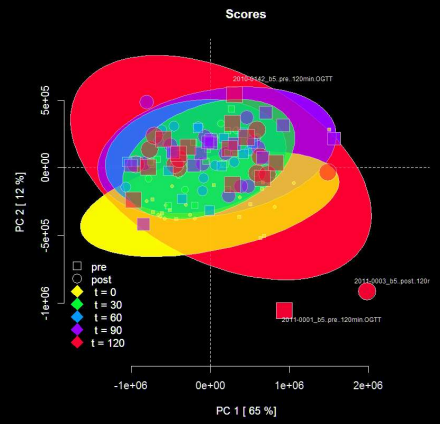
PCA interpretation

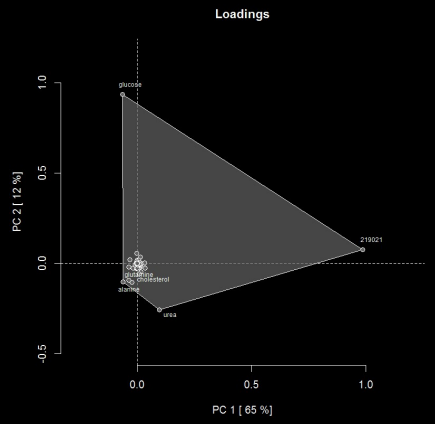
Creative Data Solutions

Variance explained (eigenvalues)

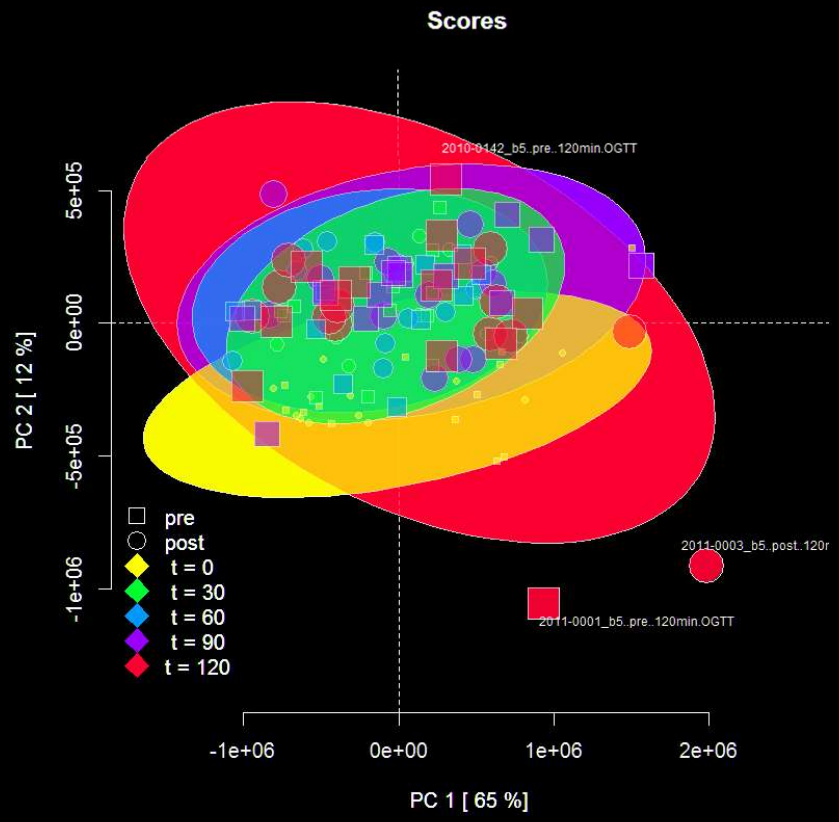


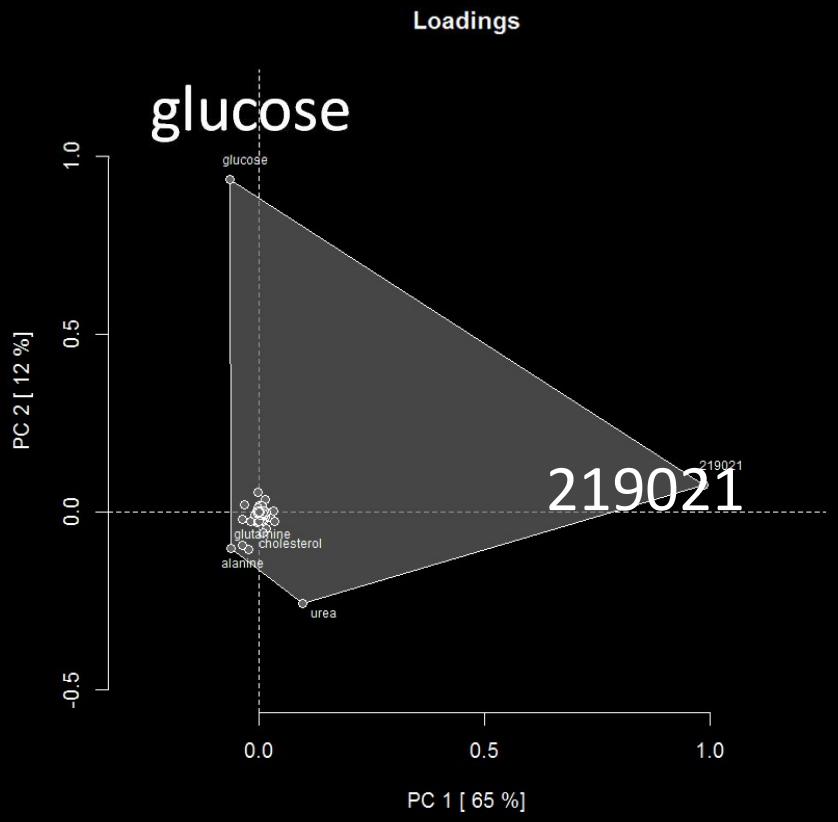
Row (sample) scores and column (variable) loadings





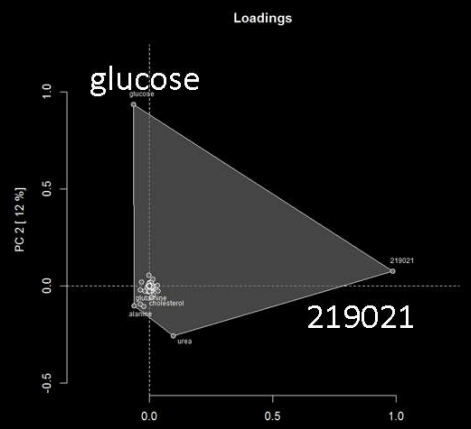
PCA example





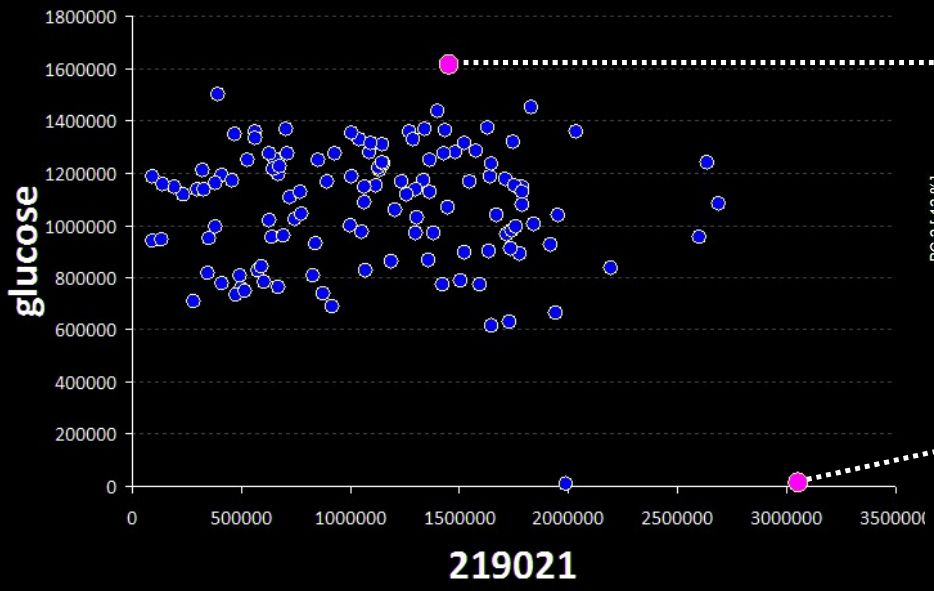
*no scaling or centering

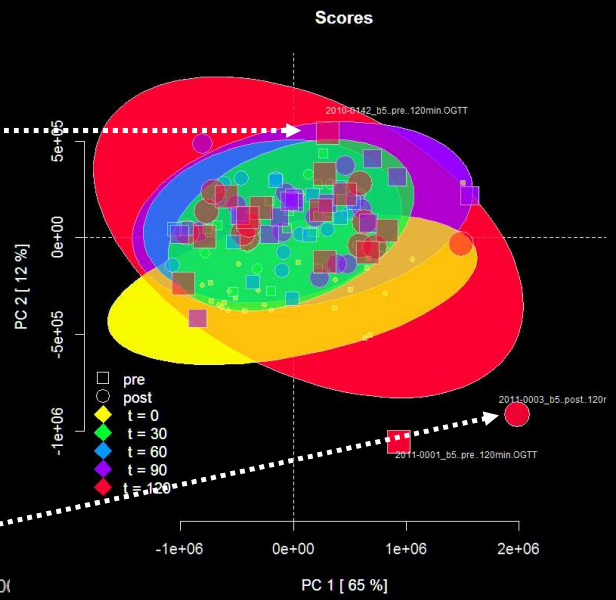
Relationship between scores and loadings



loadings

top loading variable's scatterplot





scores

Your turn

CDS Creative Data Solutions

Follow along with the following tutorial:

https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/multivariate/#pca

Machine learning

predict

- sample classification

optimize

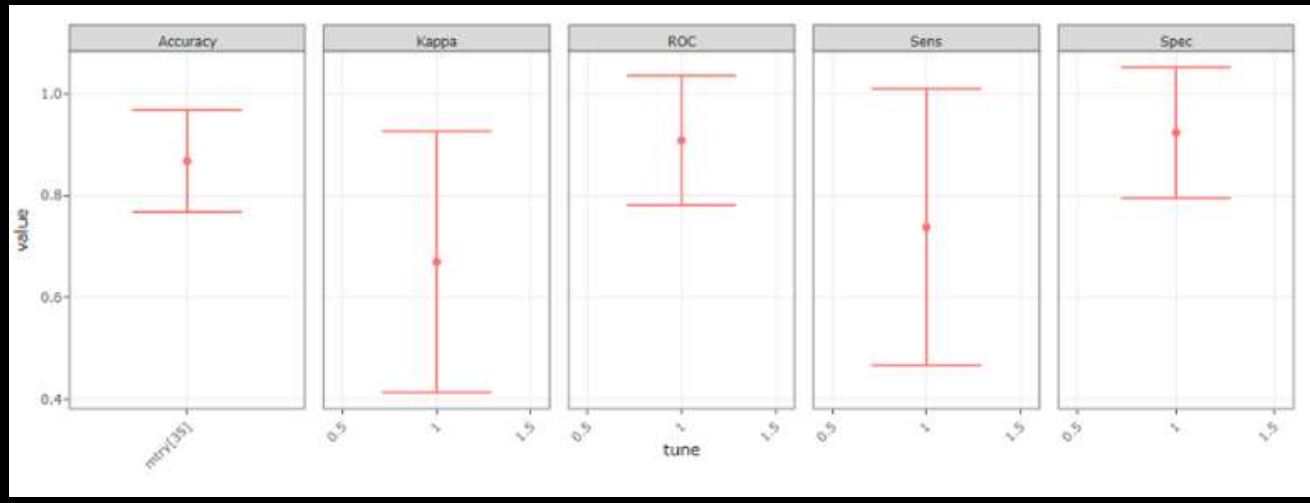
- model performance

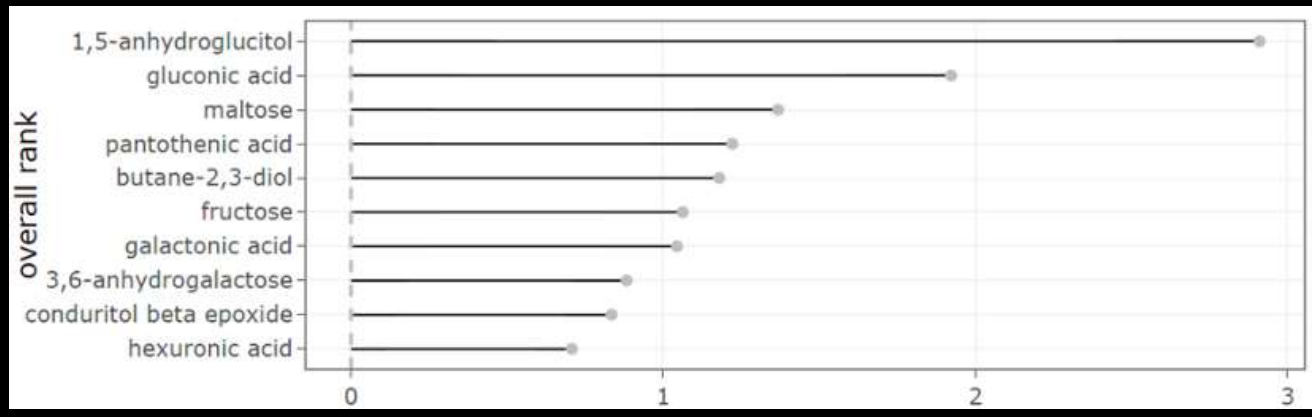
select

- important features

visualize

- model performance
- feature importance



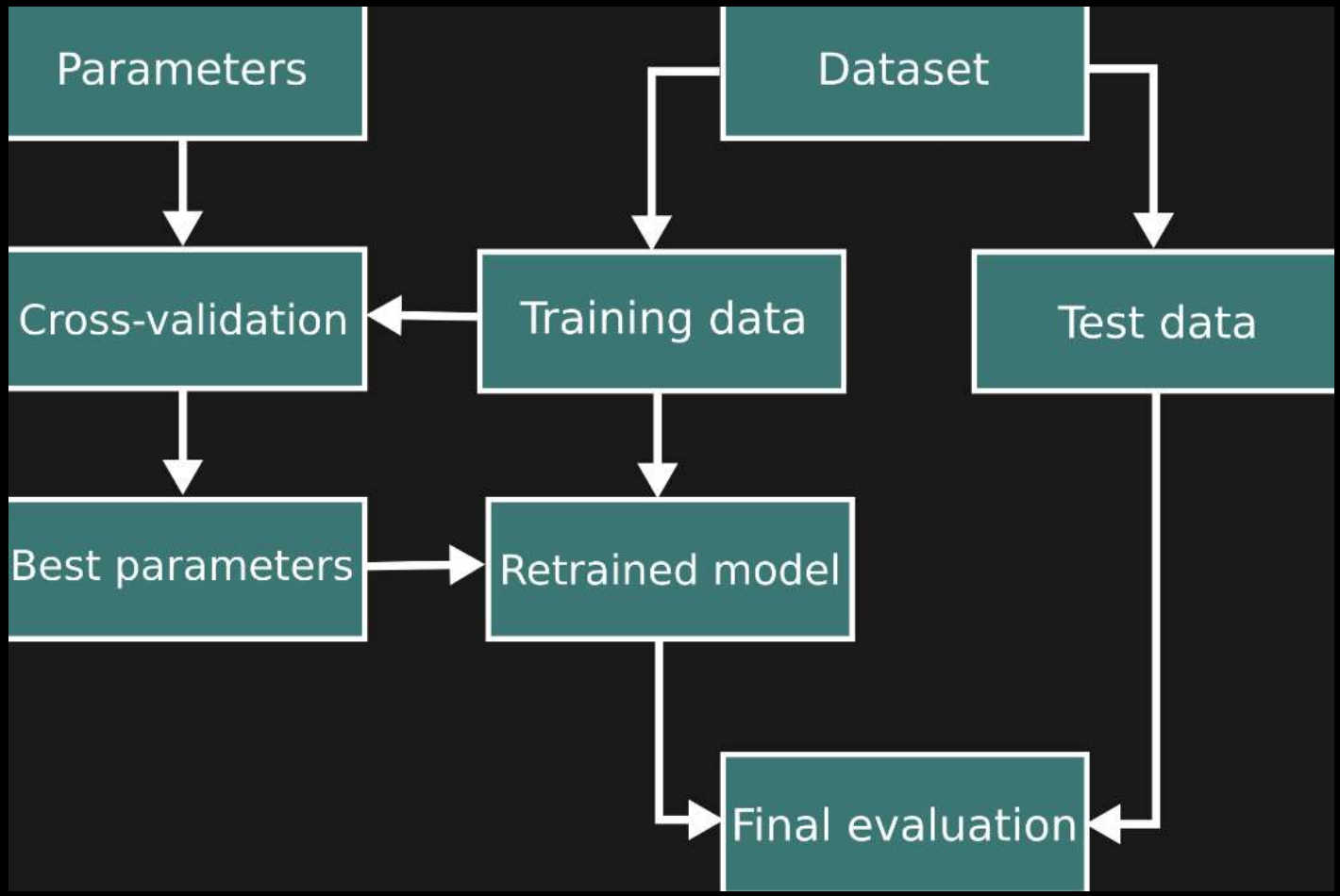


https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/model/

Machine learning

CDS Creative Data Solutions

model validation

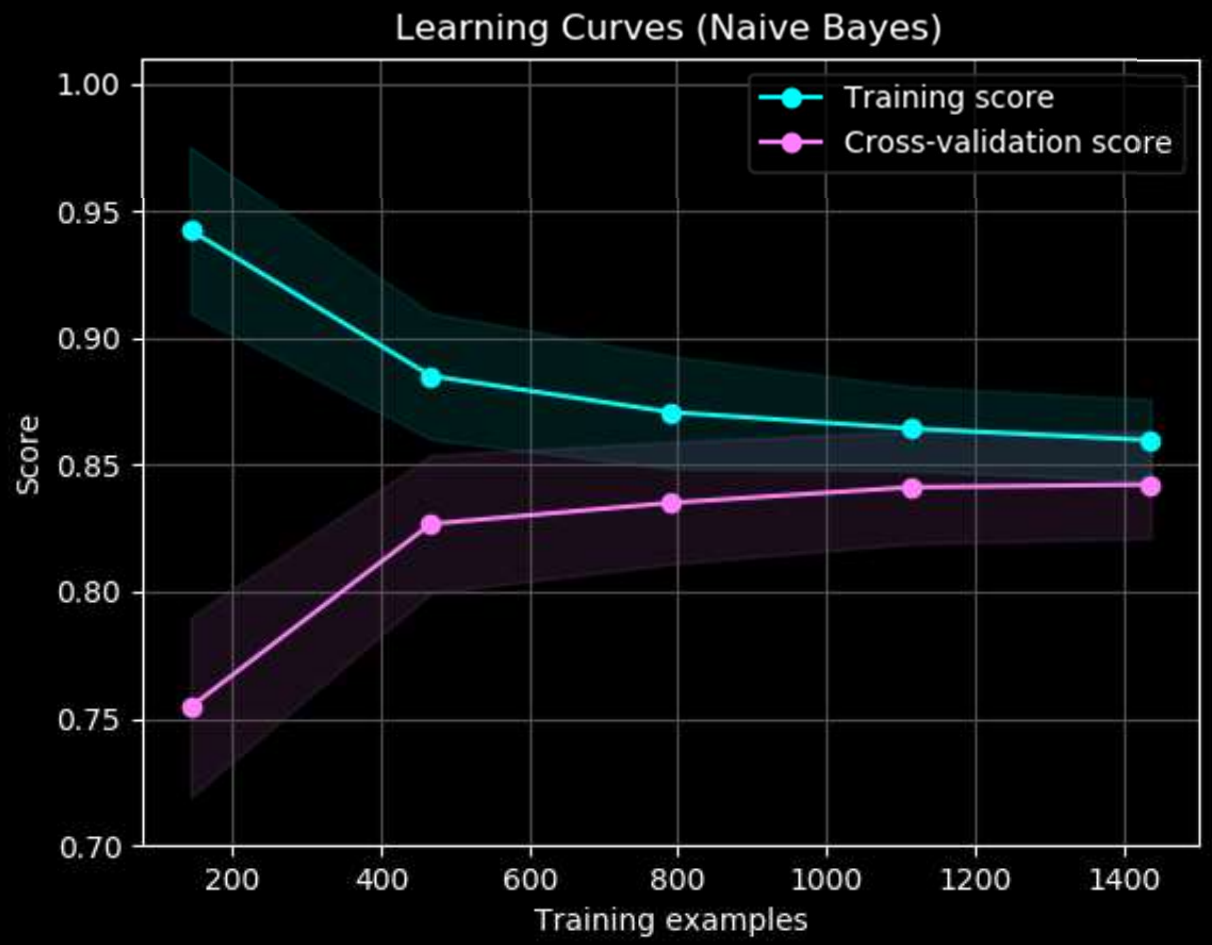


https://scikit-learn.org/stable/modules/cross_validation.html

Machine learning

model training

and validation

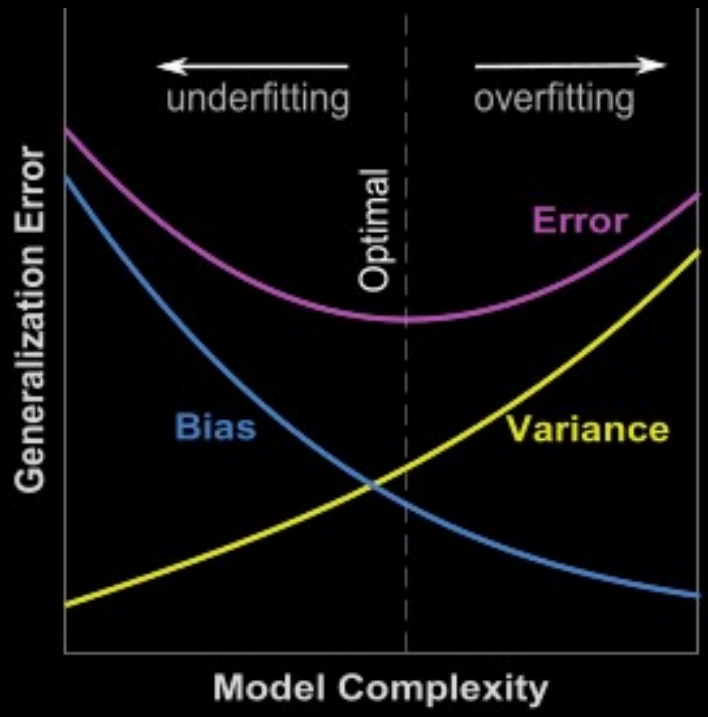


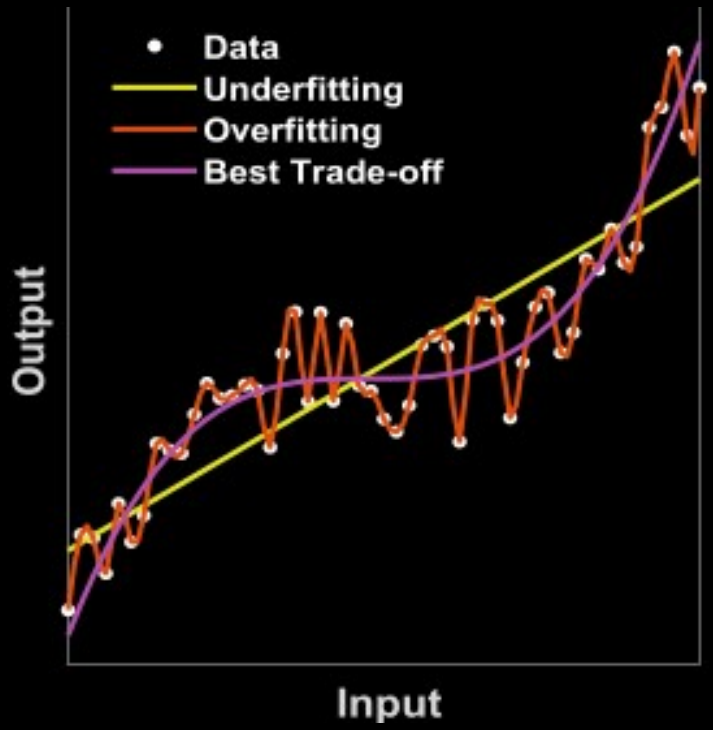
https://en.wikipedia.org/wiki/Learning_curve_(machine_learning)

Machine learning

CDS Creative Data Solutions

bias vs. variance





https://www.researchgate.net/publication/359896672_A_Review_of_Machine_Learning_Methods_Applied_to_Structural_Dynamics_and_Vibroacoustic/figures?lo=1

Machine learning

classification performance

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers

Machine learning

CDS Creative Data Solutions

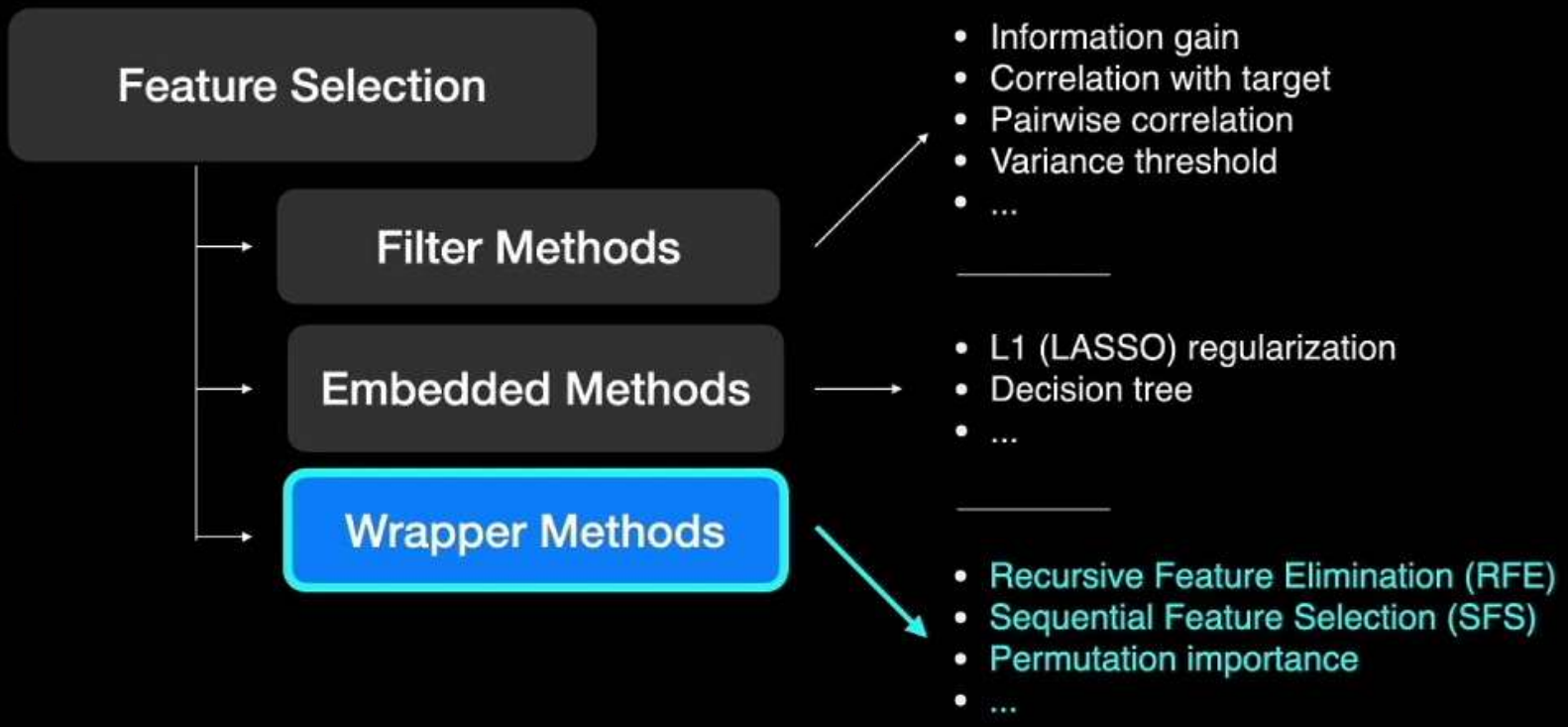
cross-validation

https://scikit-learn.org/stable/modules/cross_validation.html

Machine learning

CDS Creative Data Solutions

feature selection



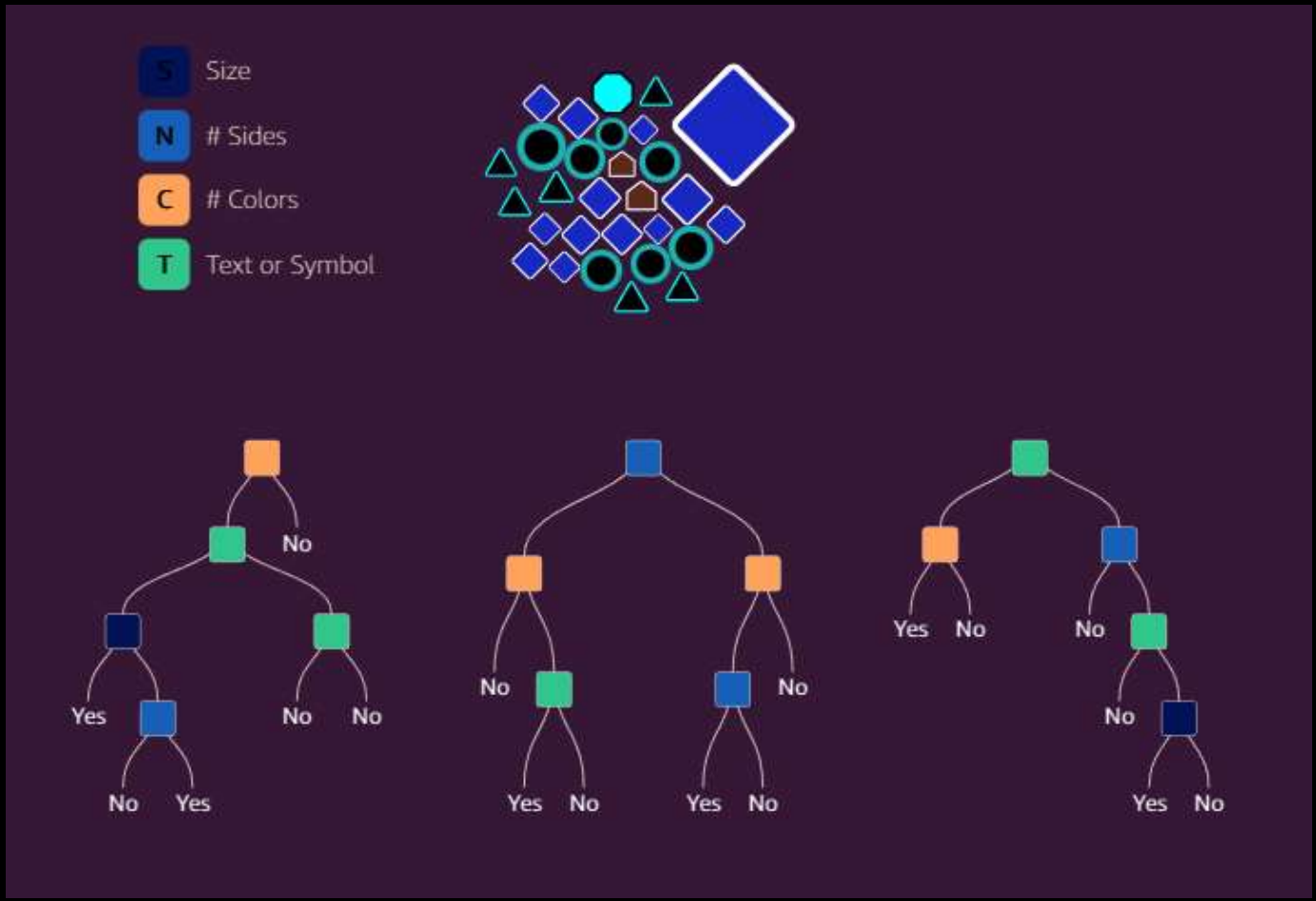
https://scikit-learn.org/stable/modules/feature_selection.html#feature-selection

Machine learning

Creative Data Solutions

Random Forest

- cross-validation
- feature selection

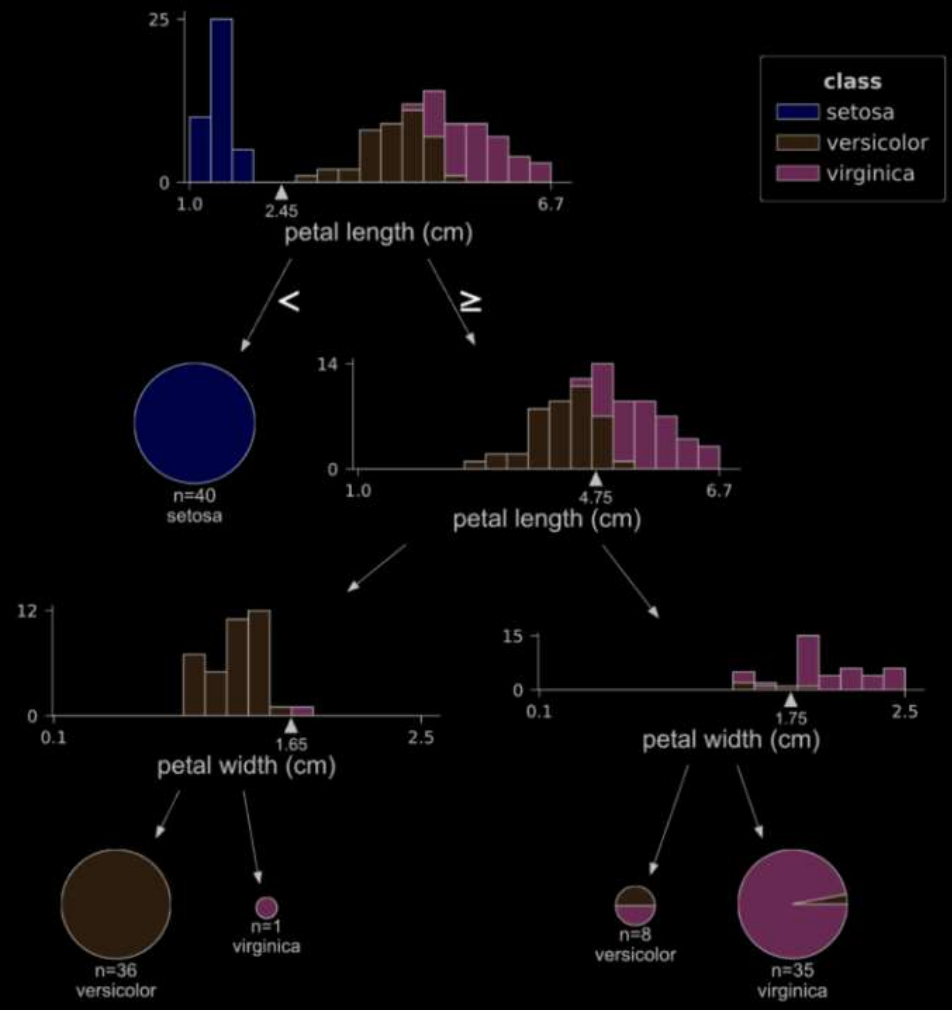


https://mlu-explain.github.io/random-forest/

Machine learning

Random Forest

- decision path

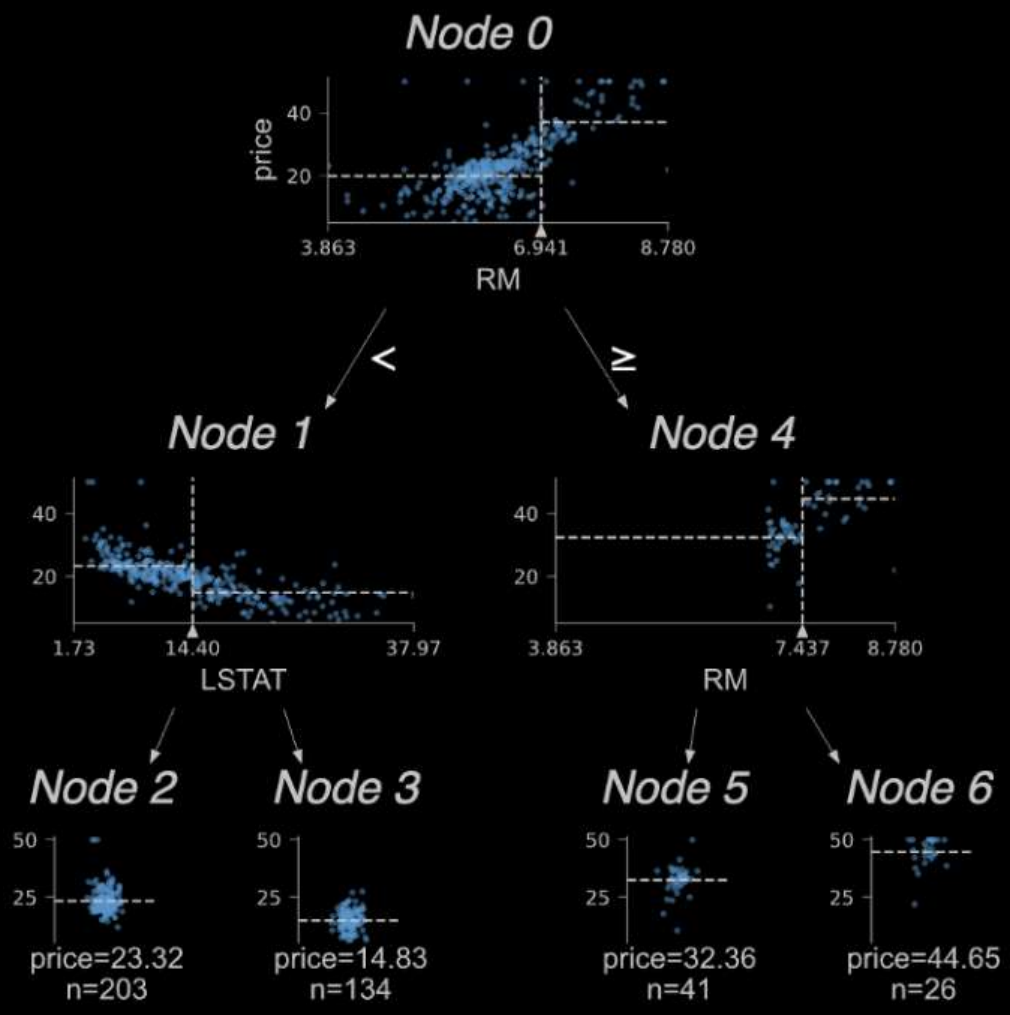


https://github.com/parrt/dtreeviz

Machine learning

Random Forest

- decision path



https://github.com/parrt/dtreeviz

Machine learning

CDS Creative Data Solutions

autoML

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.7767	0.8309	0.6056	0.7114	0.6413	0.4823	0.4950	1.3870
lr	Logistic Regression	0.7564	0.8043	0.5056	0.6941	0.5786	0.4145	0.4285	1.1230
gbc	Gradient Boosting Classifier	0.7562	0.8239	0.5667	0.6731	0.6031	0.4314	0.4431	0.0900
ada	Ada Boost Classifier	0.7526	0.8016	0.5889	0.6524	0.6091	0.4310	0.4394	0.0800
lightgbm	Light Gradient Boosting Machine	0.7524	0.8028	0.5778	0.6614	0.6086	0.4299	0.4381	0.1430
rf	Random Forest Classifier	0.7488	0.8035	0.5111	0.6849	0.5740	0.4023	0.4182	0.2350
ridge	Ridge Classifier	0.7452	0.0000	0.4722	0.6844	0.5492	0.3816	0.3997	0.0150
lda	Linear Discriminant Analysis	0.7452	0.7912	0.4833	0.6783	0.5563	0.3859	0.4017	0.0130
xgboost	Extreme Gradient Boosting	0.7449	0.7896	0.5722	0.6442	0.5984	0.4140	0.4207	0.2640
knn	K Neighbors Classifier	0.7153	0.7261	0.5111	0.5962	0.5405	0.3379	0.3467	0.0220
et	Extra Trees Classifier	0.7134	0.7573	0.4333	0.6079	0.4968	0.3072	0.3204	0.1810
dt	Decision Tree Classifier	0.7075	0.6741	0.5722	0.5635	0.5630	0.3445	0.3481	0.0130
nb	Naive Bayes	0.6817	0.7064	0.2389	0.5527	0.3288	0.1657	0.1905	0.0110
svm	SVM - Linear Kernel	0.6015	0.0000	0.3611	0.3419	0.3251	0.0851	0.0924	0.0170
qda	Quadratic Discriminant Analysis	0.5759	0.5889	0.4833	0.4062	0.3705	0.1011	0.1281	0.0180

https://pycaret.gitbook.io/docs/get-started/quickstart?q=leaderboard

Your turn

Creative Data Solutions

Follow along with the following tutorial:

https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/model/

Network analysis

network mapping

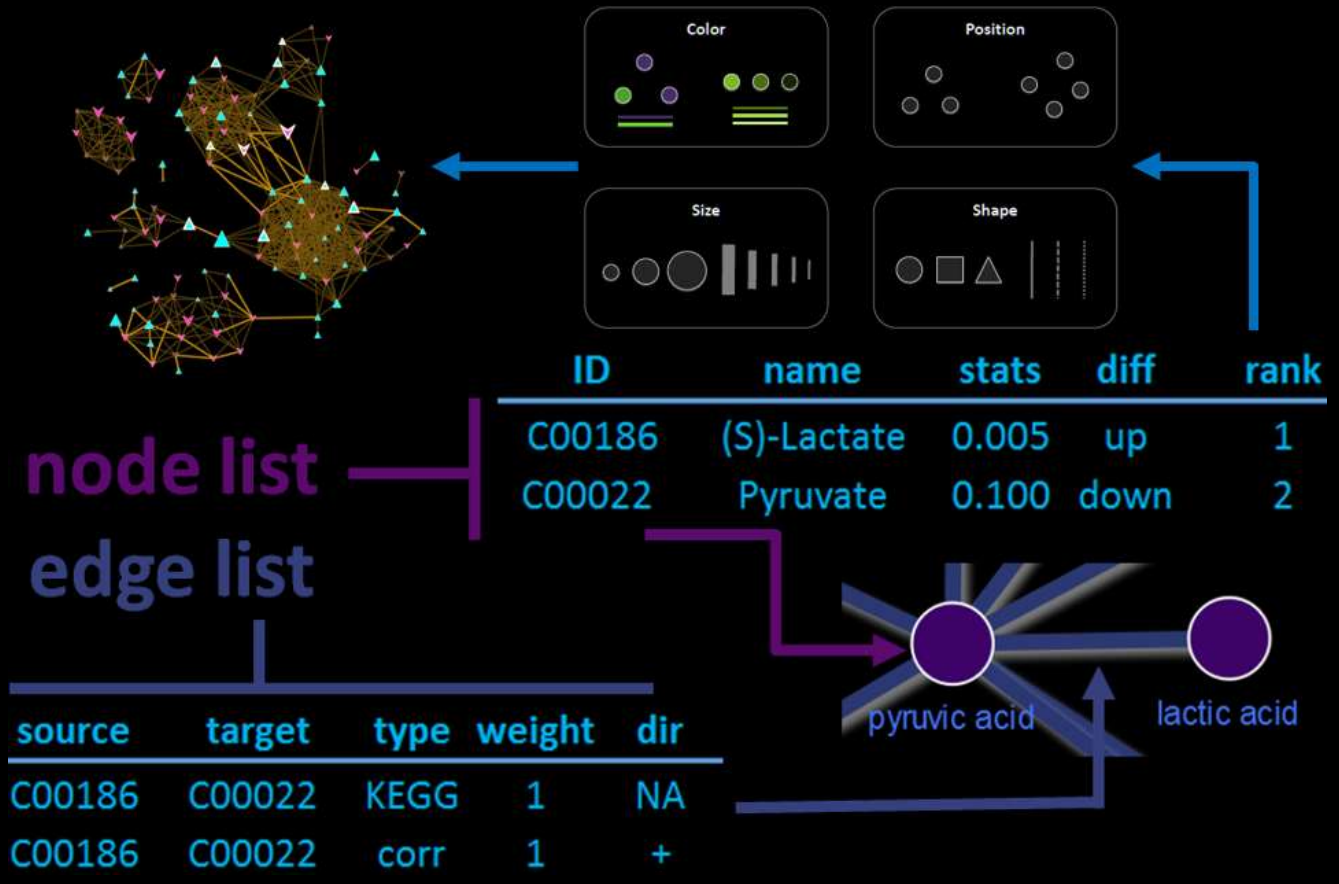
- transform variables

network calculation

- regularized correlation
- biochemical
- structural similarity
- model performance

visualize

- interactive networks



https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/network/

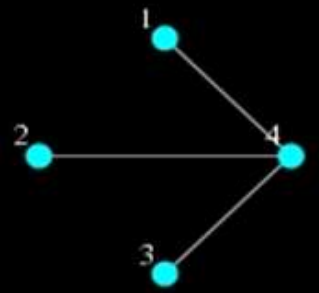
Components for network mapping

connections (edges)

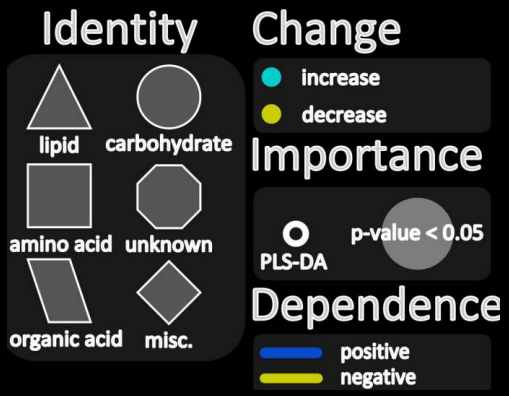
- empirical dependency (correlation)
- biochemical (substrate/product)
- chemical similarity
- …

nodes (vertices)

- magnitude
- importance
- direction
- relationships
- …



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$



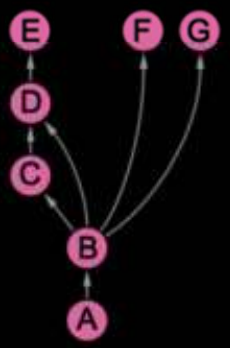
Network data structures

adjacency matrix

Undirected

	A	B	C	D	E	F	G	Degree
A	0	1	1	1	1	1	0	5
B	1	0	0	0	0	1	0	2
C	1	0	0	0	0	0	0	1
D	1	0	0	0	0	0	0	1
E	1	0	0	0	0	0	0	1
F	1	1	0	0	0	0	1	3
G	0	0	0	0	0	1	0	1

Directed



	A	B	C	D	E	F	G	Out-degree
A	0	1	0	0	0	0	0	1
B	0	0	1	1	0	1	1	4
C	0	0	0	1	0	0	0	1
D	0	0	0	0	1	0	0	1
E	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0

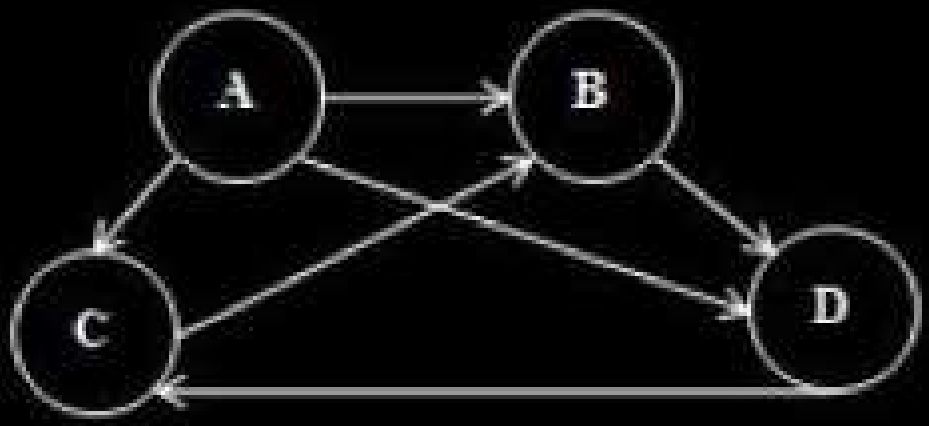
Weighted

	A	B	C	D	E	F	G	Degree
A	0	8	12	12	12	16	12	72
B	8	0	0	0	0	4	0	12
C	12	0	0	0	0	0	0	12
D	12	0	0	0	0	0	0	12
E	12	0	0	0	0	0	0	12
F	16	4	0	0	0	0	12	32
G	12	0	0	0	0	12	0	24

https://www.steveclarkapps.com/graphs/

Network data structures

adjacency list



Node	Adjacent Node(s)
A	B C D
B	D
C	B
D	C

https://www.steveclarkapps.com/graphs/

Network data structures

edge list and
vertices list
(node attributes)

edgesList

source	target
1	2
1	4
1	5
2	3
2	5
2	6
3	6
4	5
5	6
...	

verticesList

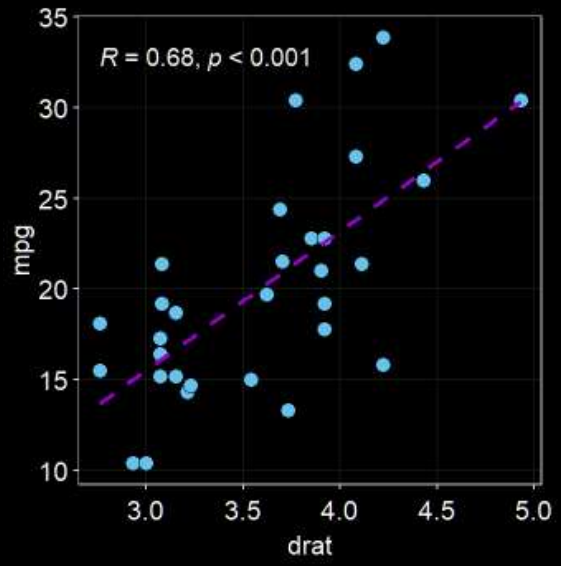
index
1
2
3
4
5
6
...

Correlation networks

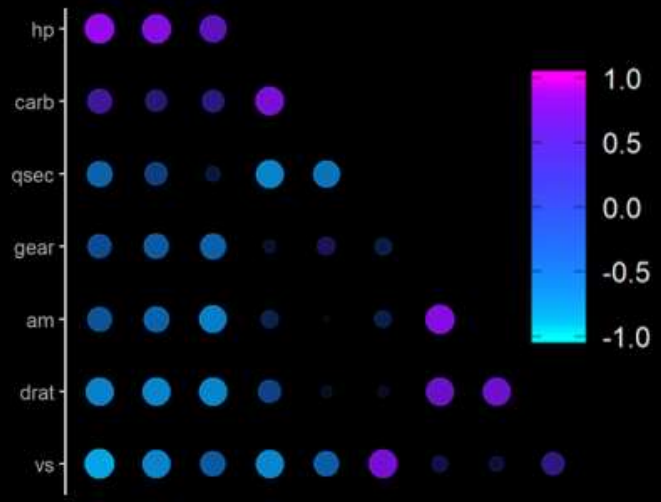
Creative Data Solutions

Connect molecules based on strength of their correlation or partial-correlation

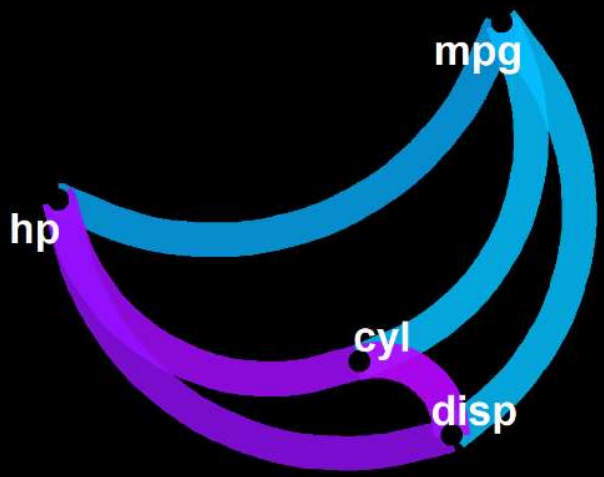
bivariate



multivariate

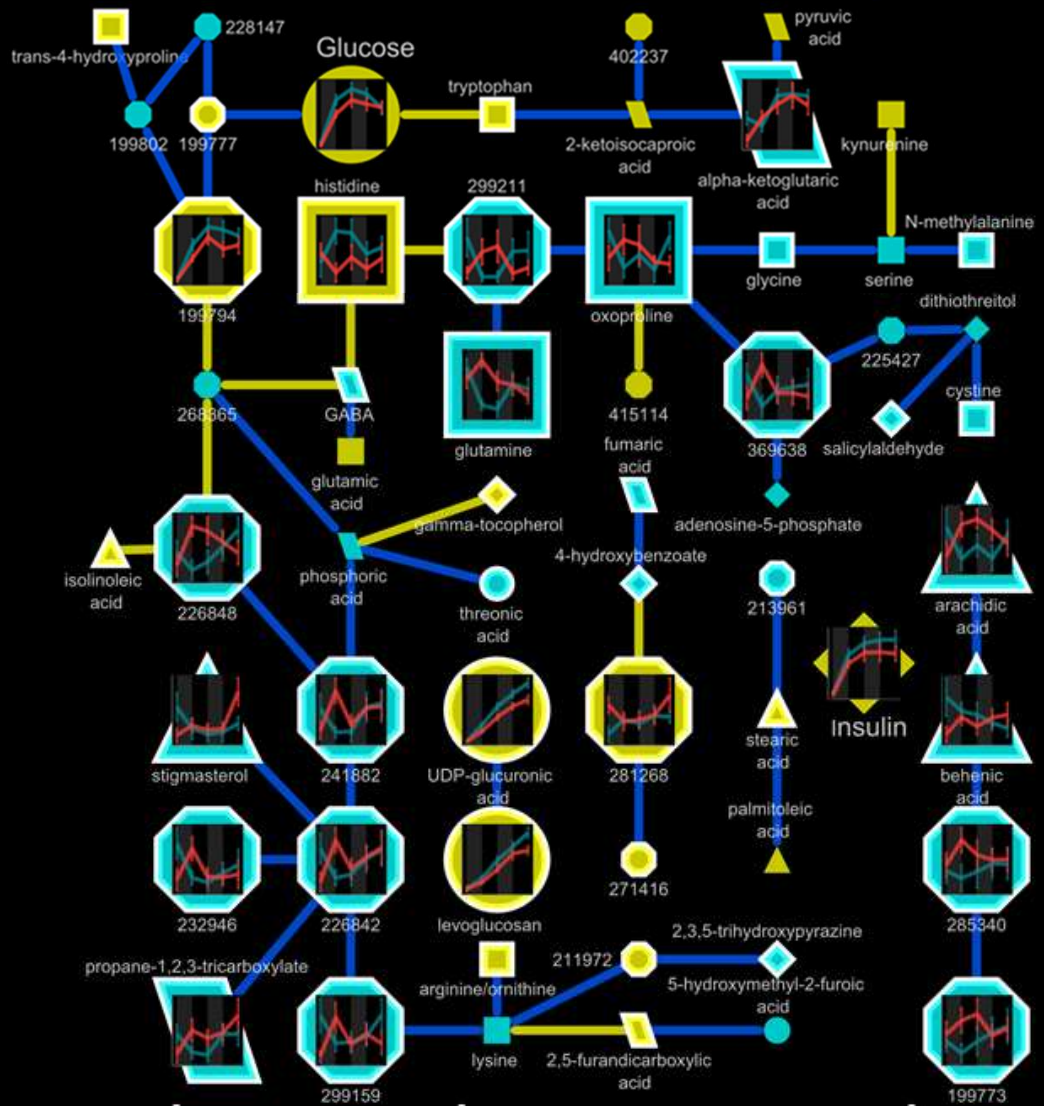


network

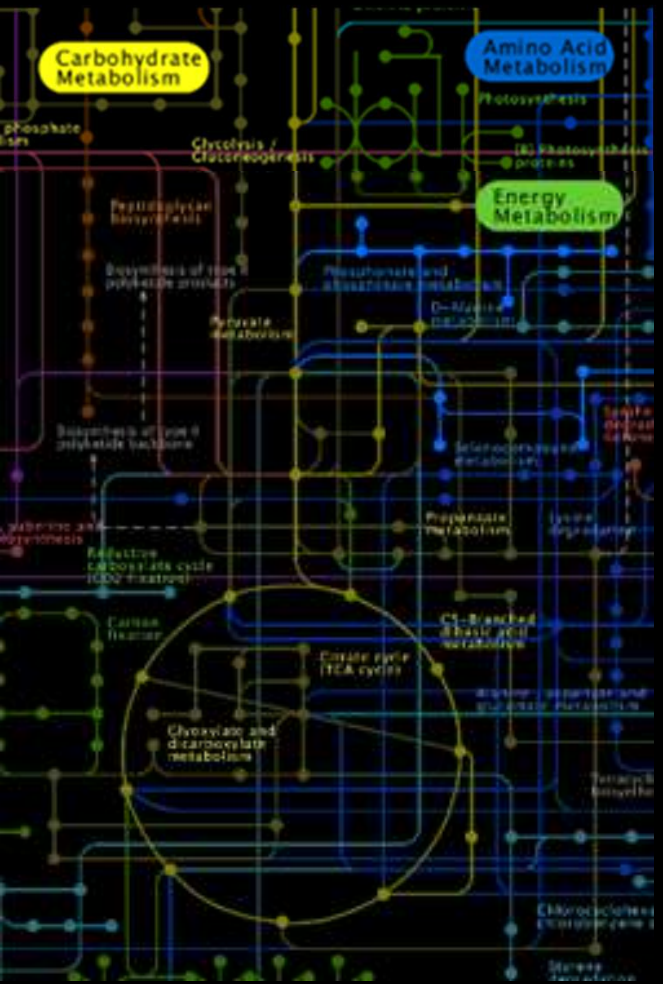


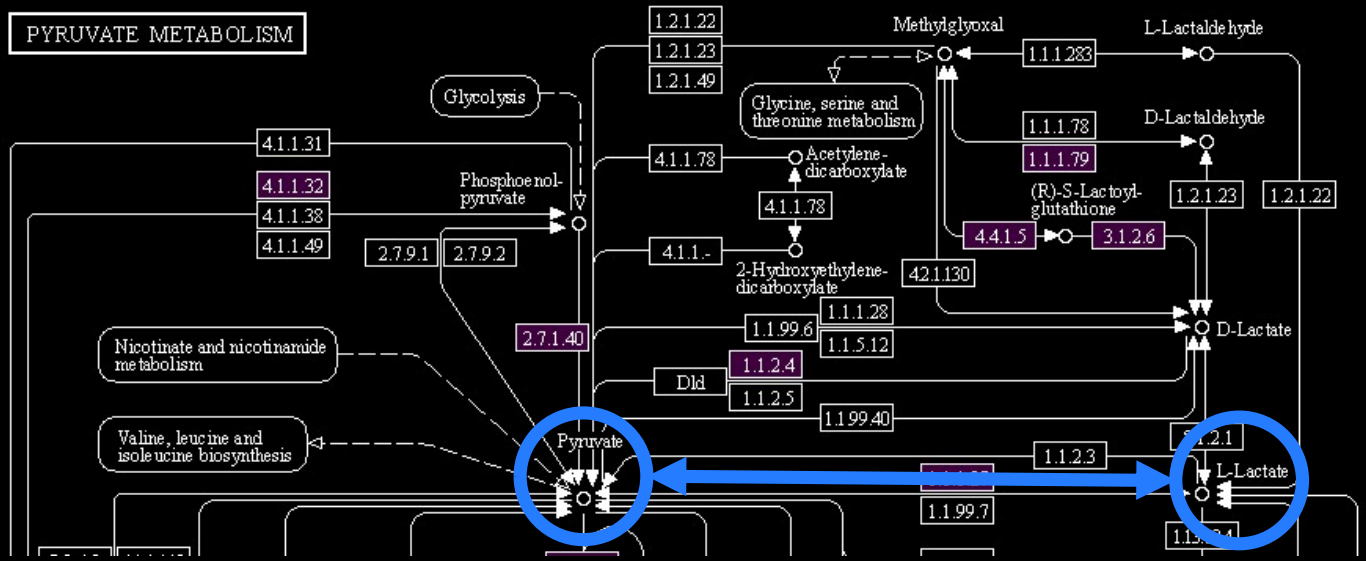
Correlation example

regularized correlation network showing relationships in metabolic timeseries measurements for two classes of samples



Biochemical networks





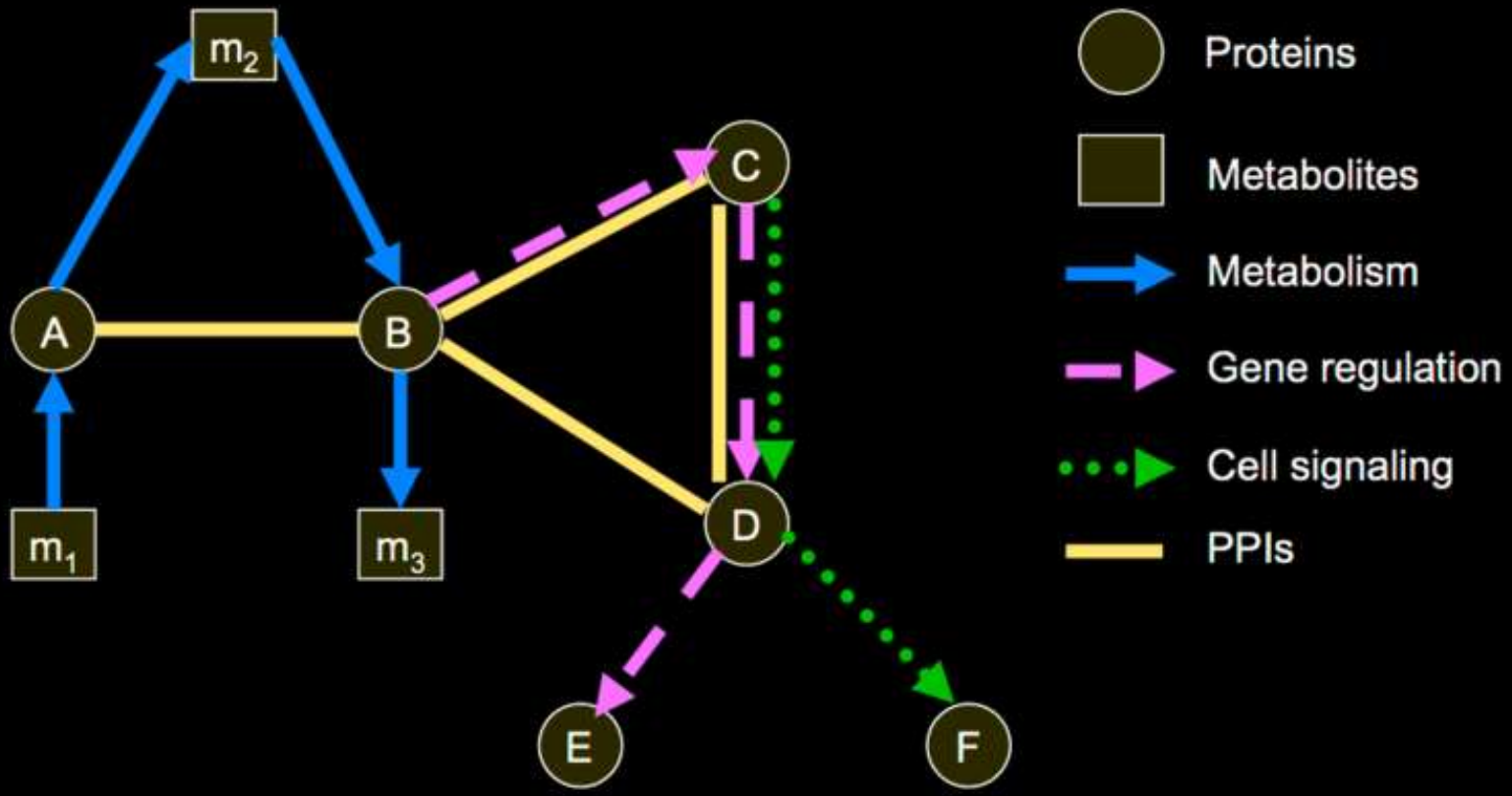
nodes

ID	name	stats	diff	rank
C00186	(S)-Lactate	0.005	up	1
C00022	Pyruvate	0.100	down	2

edges

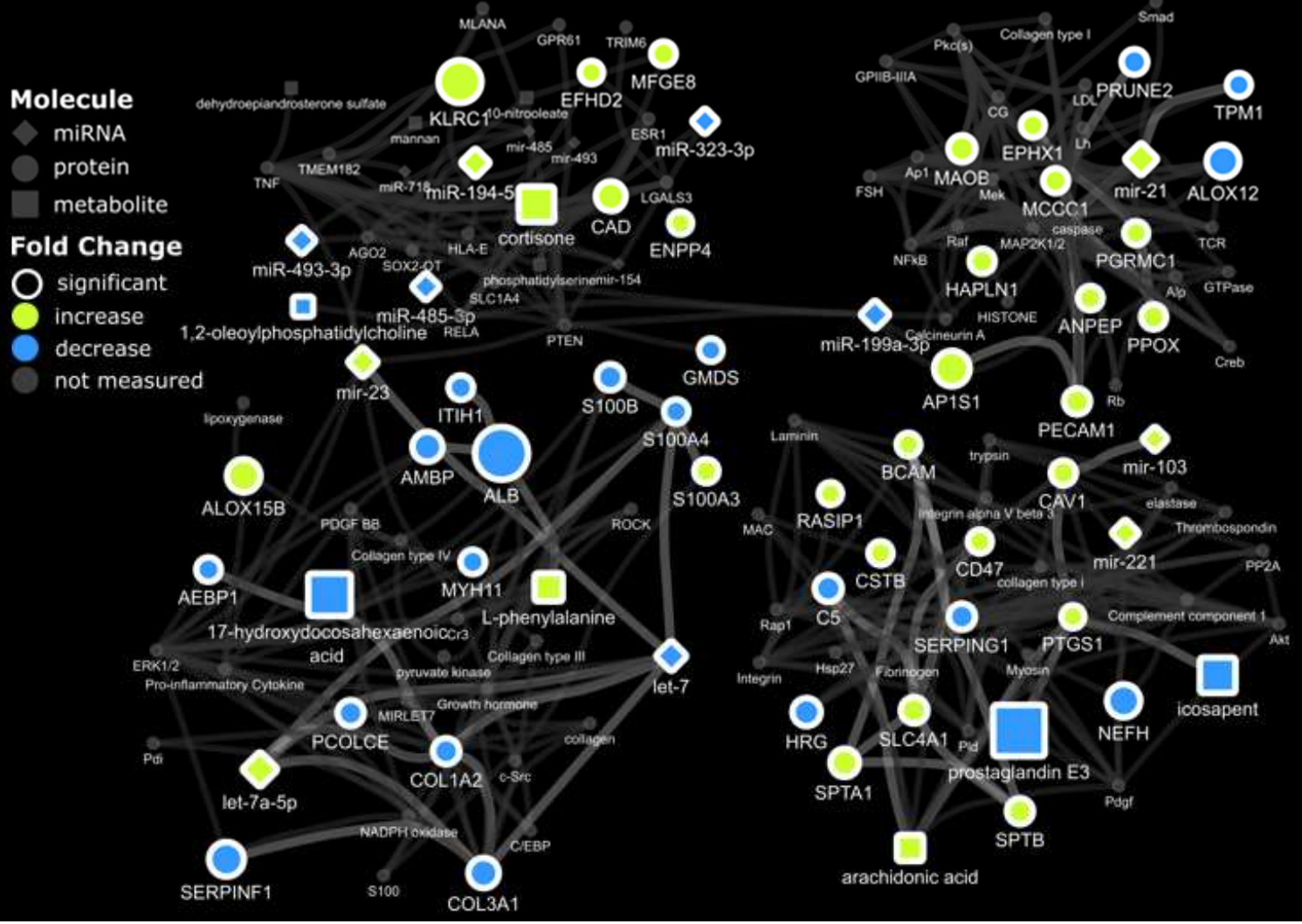
source	target	type	weight	dir
C00186	C00022	KEGG	1	NA

Multi-Omic networks



https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/types-of-biological-networks/

Multi-Omic networks



Multi-Omic networks

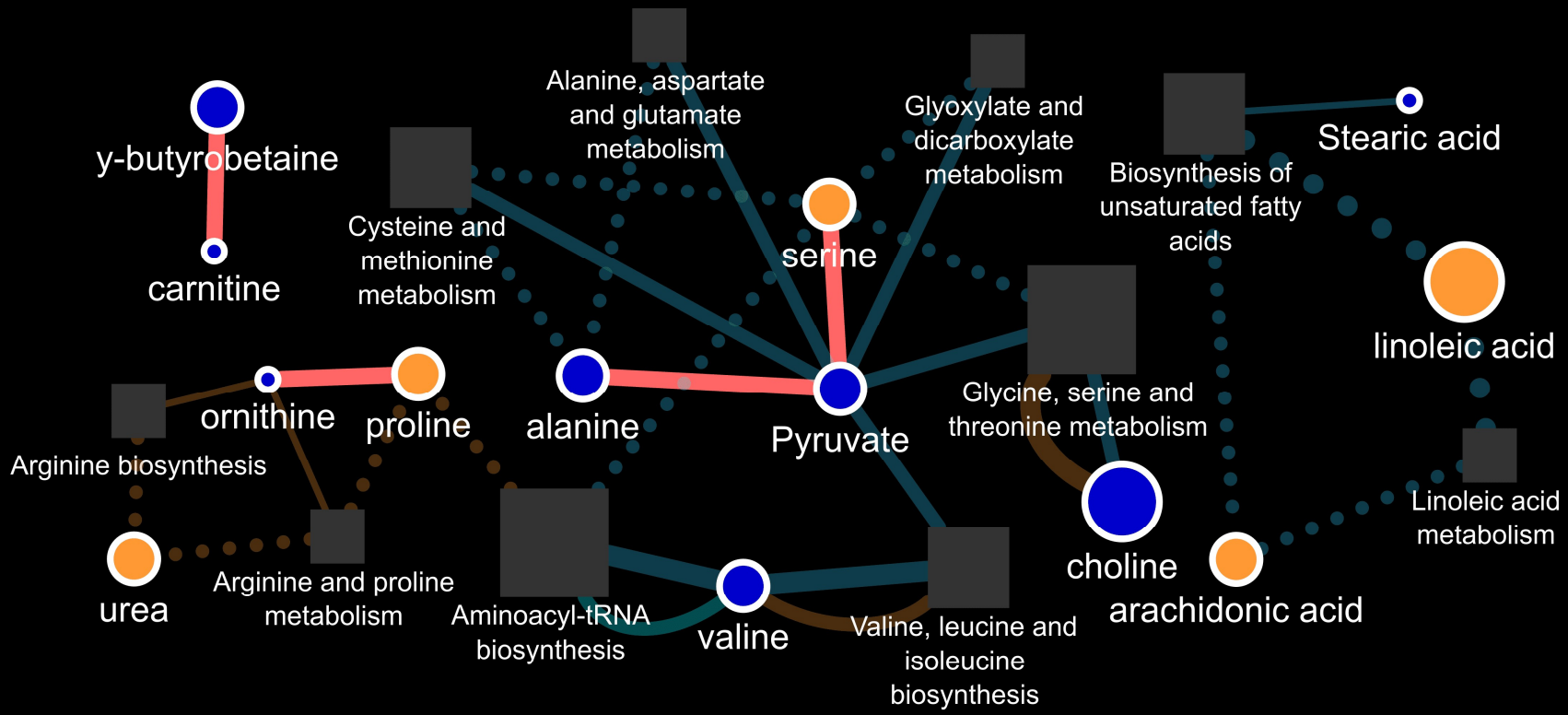
CDS Creative Data Solutions

Type

- metabolite
- pathway
- increase
- decrease

Relationship

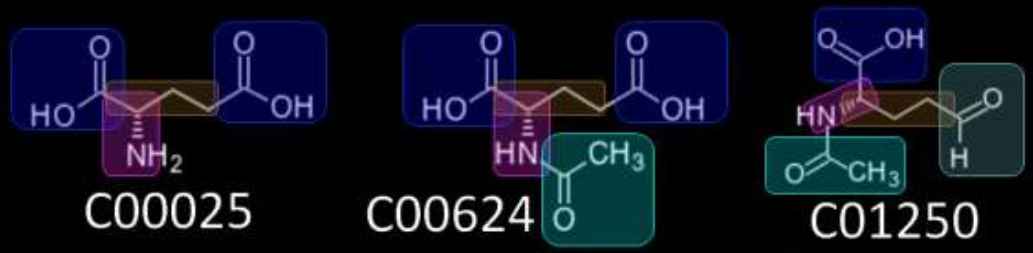
- biochemical
- blood
- saliva
- decrease
- increase



Structural similarity networks

- Use structure to generate molecular fingerprint
- Calculate similarities between metabolites based on fingerprint

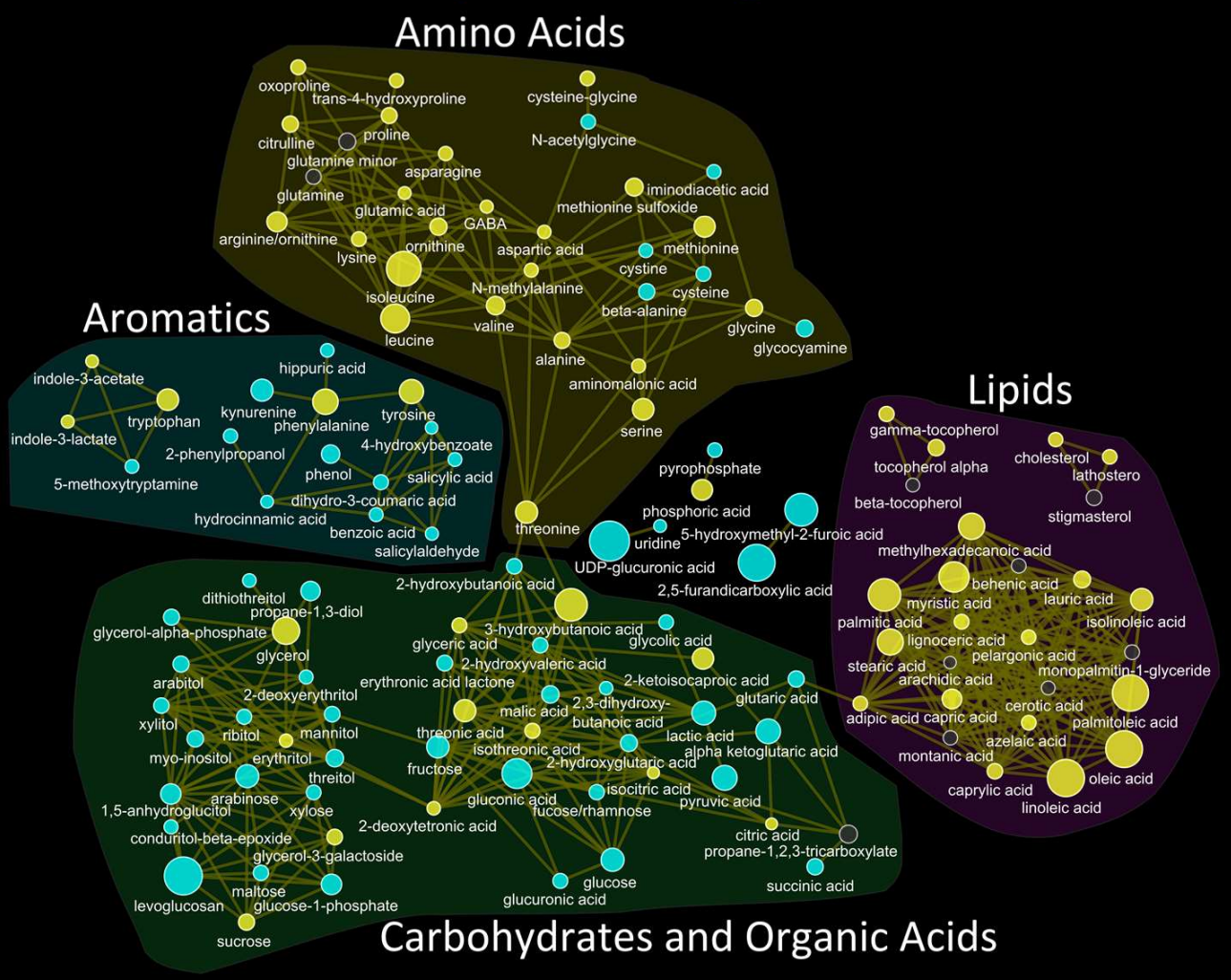
Chemical mapping
of substructure comparison
using PubChem



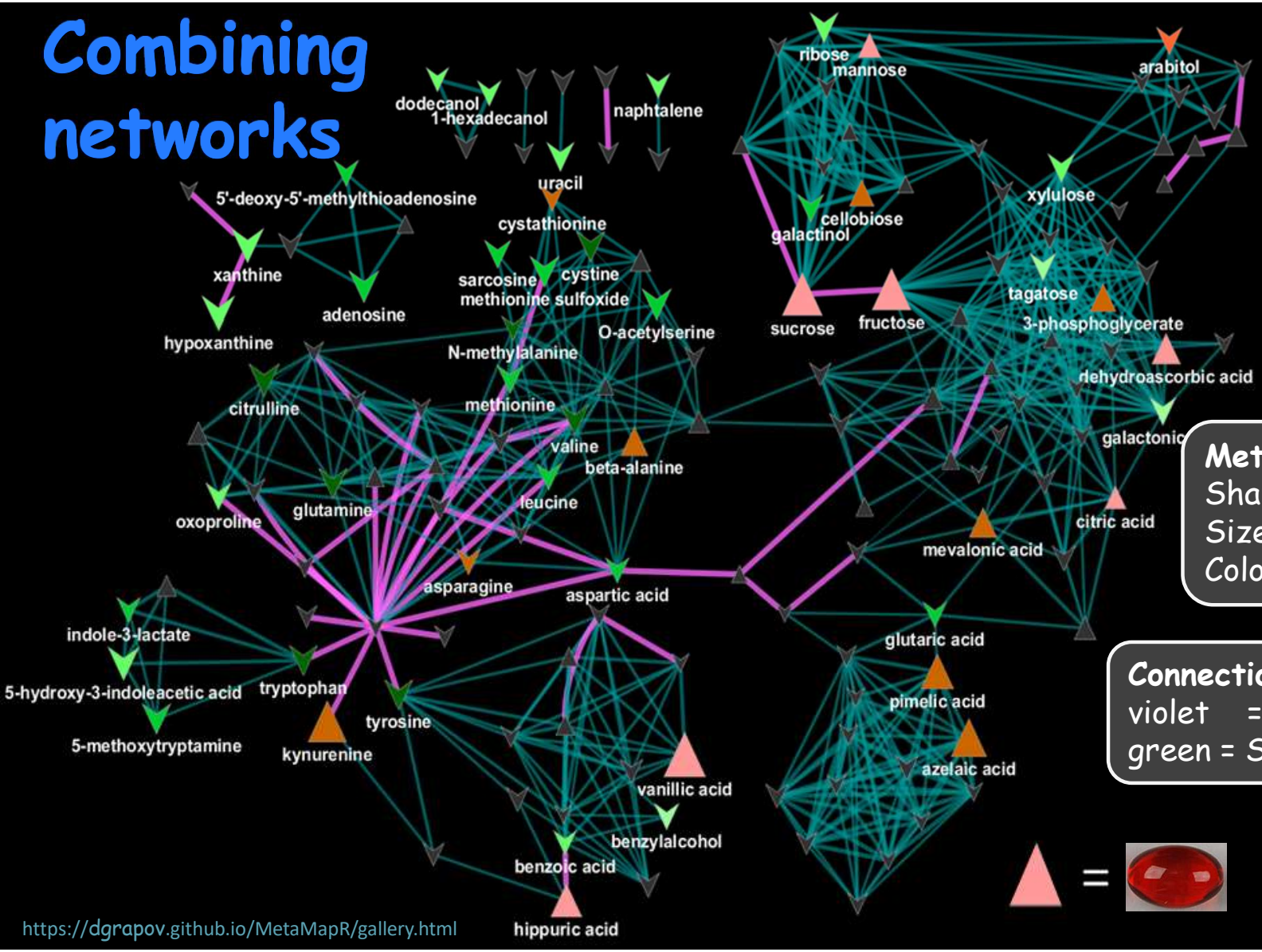
substructure matrix decomposition and
Tanimoto chemical similarity calculations

BMC Bioinformatics 2012, **13**:99 doi:10.1186/1471-2105-13-99

Structural similarity example



Combining networks



Metabolites
Shape = increase/decrease
Size = importance (loading)
Color = correlation

Connections
violet = Biochemical relationships
green = Structural similarity

https://dgrapov.github.io/MetaMapR/gallery.html

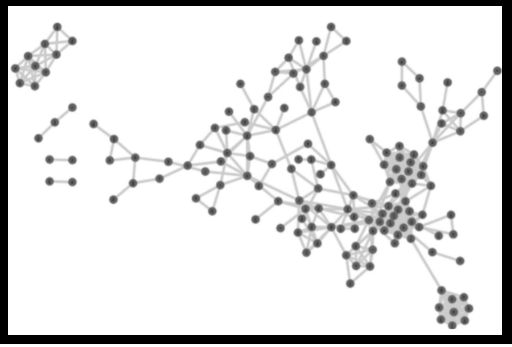
Your turn

Creative Data Solutions

Follow along with the following tutorial:

https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/network/

Network refinement and visualization



learn

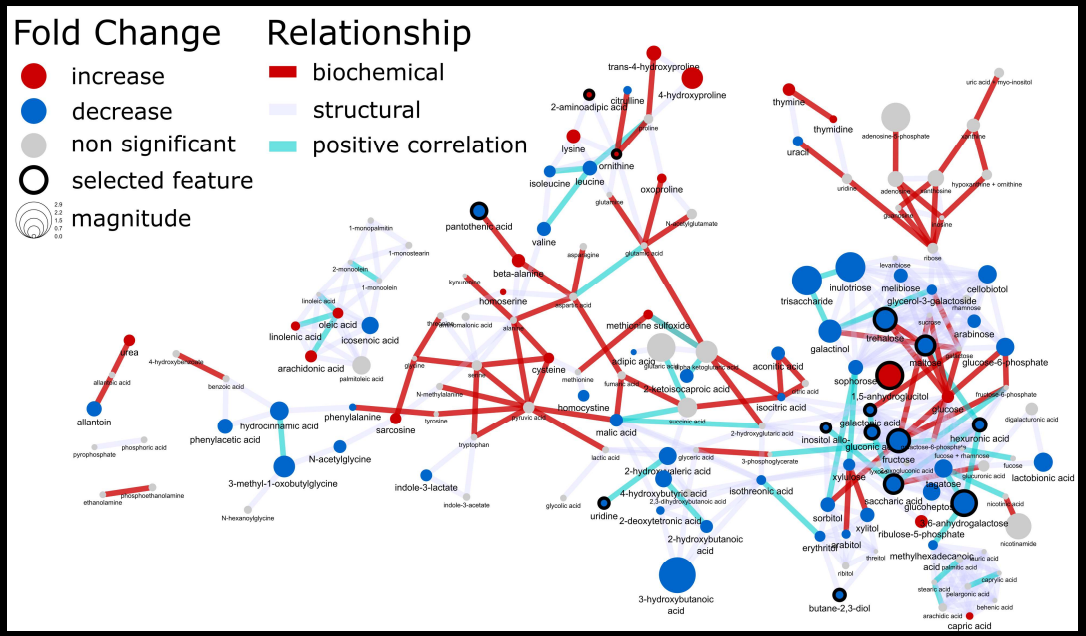
- Cytoscape basics

map variables to

- node attributes
- edge attributes

optimize

- layout
- legend
- publication quality figure



https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/cytoscape/

Your turn

Creative Data Solutions

Follow along with the following tutorial:

https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/cytoscape/