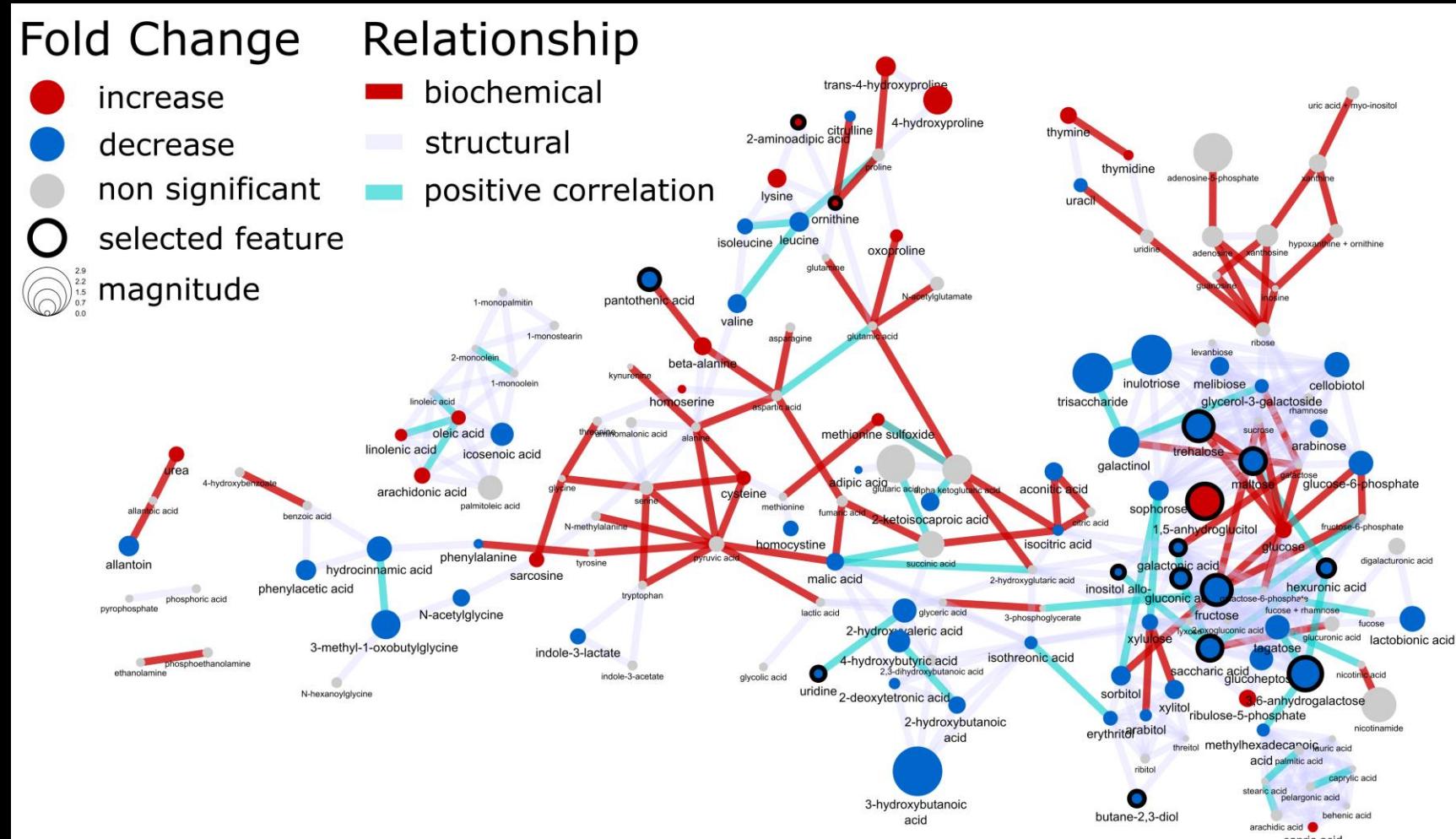


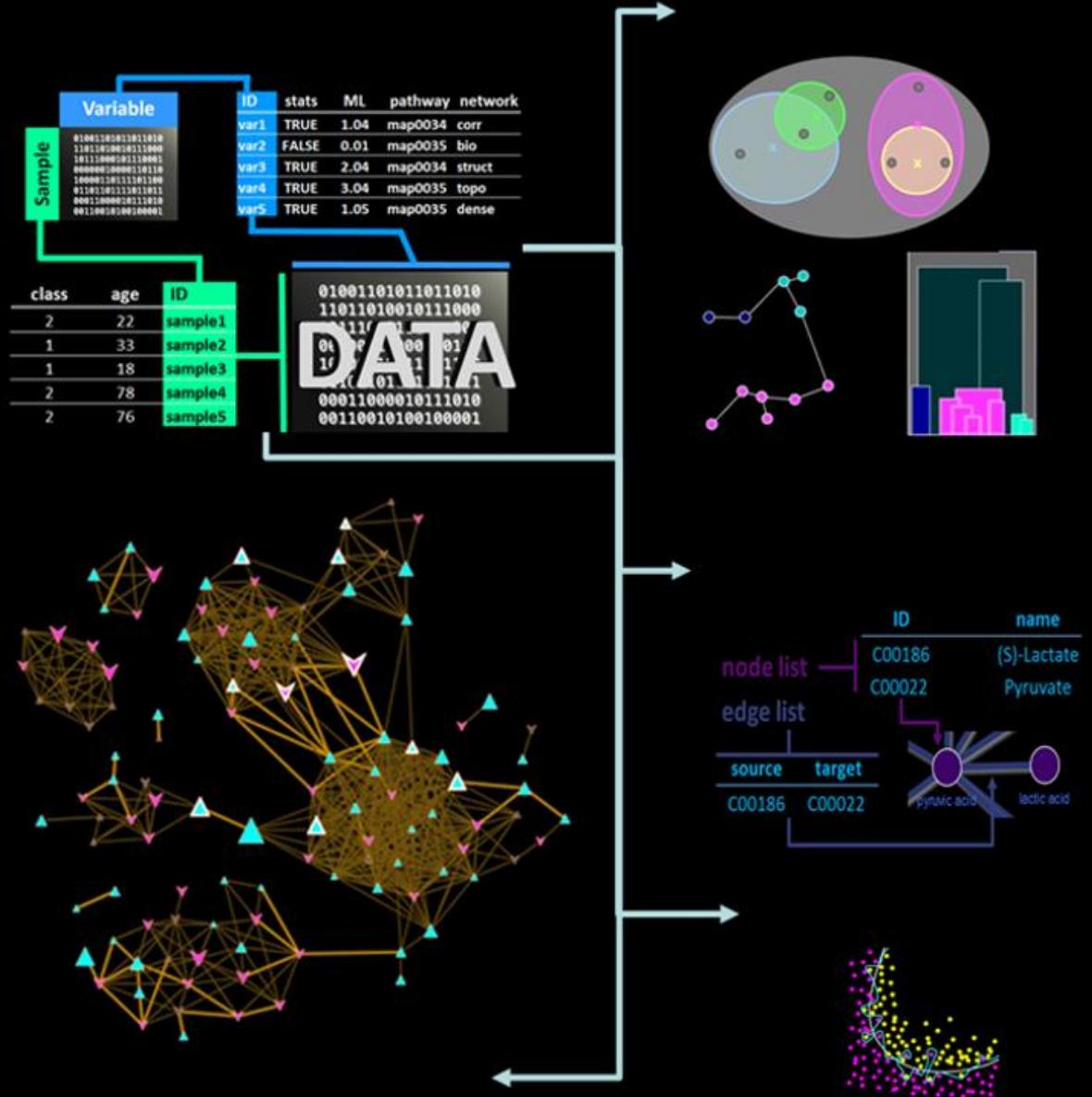
# Welcome to network mapping 101

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

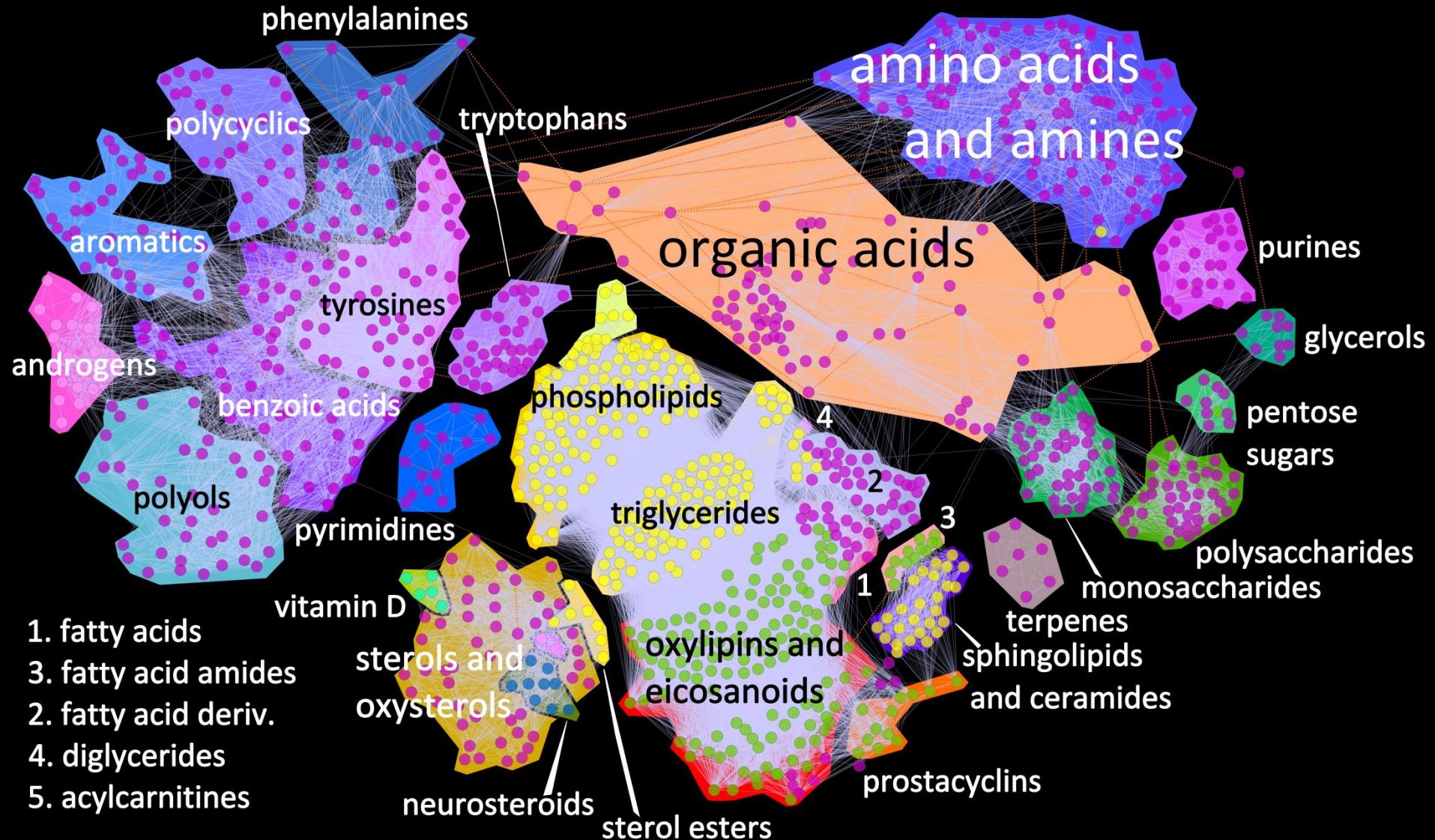


# Tutorials

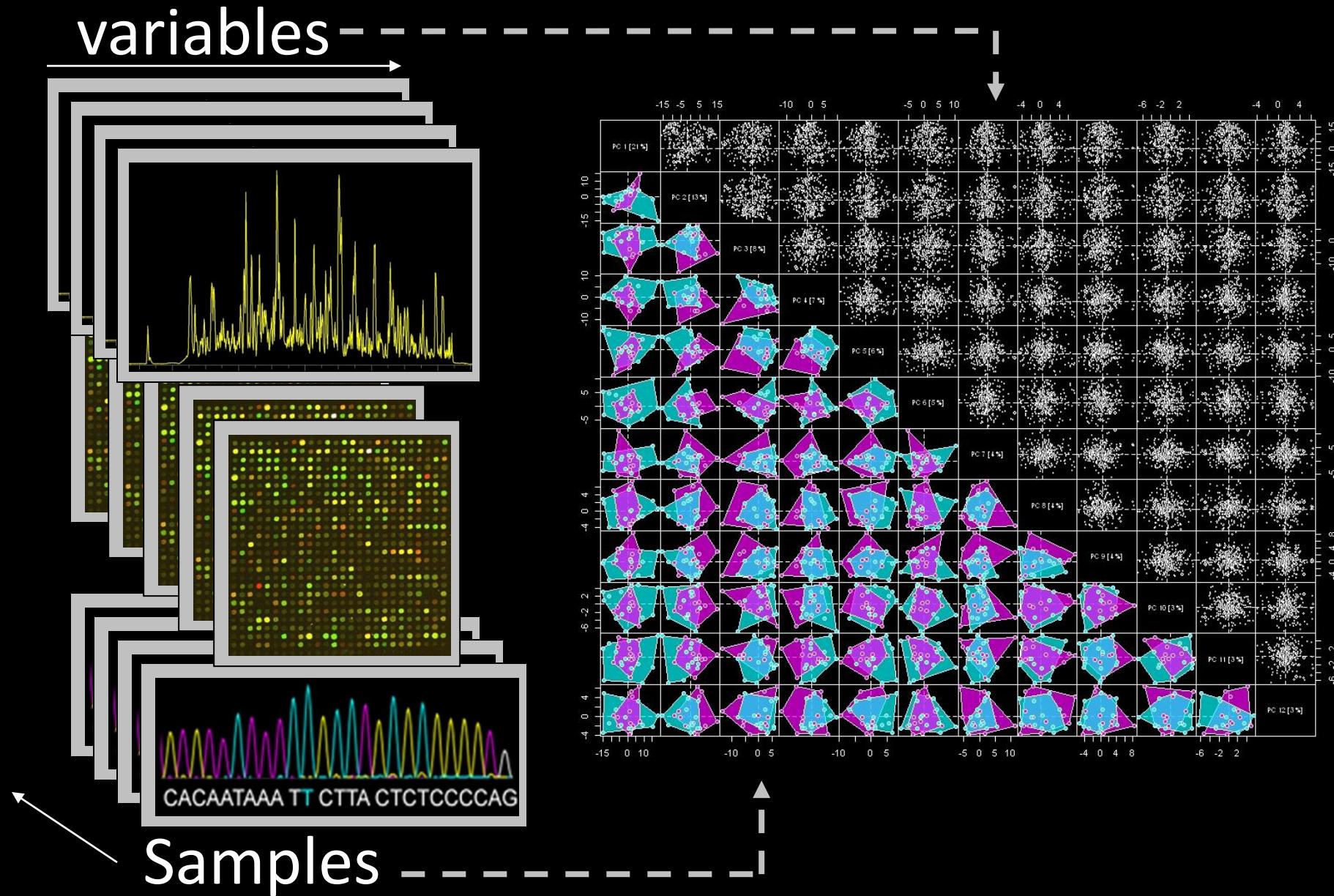
- Preparing raw data for analysis
- Statistical analysis
- Multivariate data exploration
- Supervised clustering
- Machine learning
  - classification
  - model validation
  - feature selection
- Network analysis
  - biochemical
  - structural similarity
  - correlation
- Network mapping - putting it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



# Analysis at the metabolomic scale



# Integrate high-dimensional data



# Identify what matters

## Univariate

# Multivariate

# Predictive Modeling

Group 1

Group 2

ପ୍ରମାଣିତ କାନ୍ଦିତ କାନ୍ଦିତ  
କାନ୍ଦିତ କାନ୍ଦିତ କାନ୍ଦିତ  
କାନ୍ଦିତ କାନ୍ଦିତ କାନ୍ଦିତ  
କାନ୍ଦିତ କାନ୍ଦିତ କାନ୍ଦିତ  
କାନ୍ଦିତ କାନ୍ଦିତ କାନ୍ଦିତ

```

JULY91,YYY20F      +One
K-31A,4x40L      1 10000J
R033,3407F      1 10000J
R033,V14S/-      1 7/HV5711
RUYHNM0079NP?HRA-14  SUMMER
d' mddm  G-5      "n
<1>3:3 car Jh3k  h
<1>3:3 R,3:3 car D-9r
<VAR> 340#H1V2,F23-F9
<1>9-0:00#H1V2,F23-F9
[1]UC 1=2tunP P-n-9H  HYTHM0A3*
h-3:3/Vd  65/5
S-3:3/Vd  65/5
S-3:3/Vd  1 9/8
Ac5:1>13  3/3
YankeeDooDoo  3/3
J3:3/Vd  J3:3
JK/V  JK/V
OL/6  OL/6
*1/6  *1/6
S-3:3  S-3:3
sd-  sd-
?/-  ?/-
H0007C  H0007C
H0007C  H0007C
M4110  M4110

```

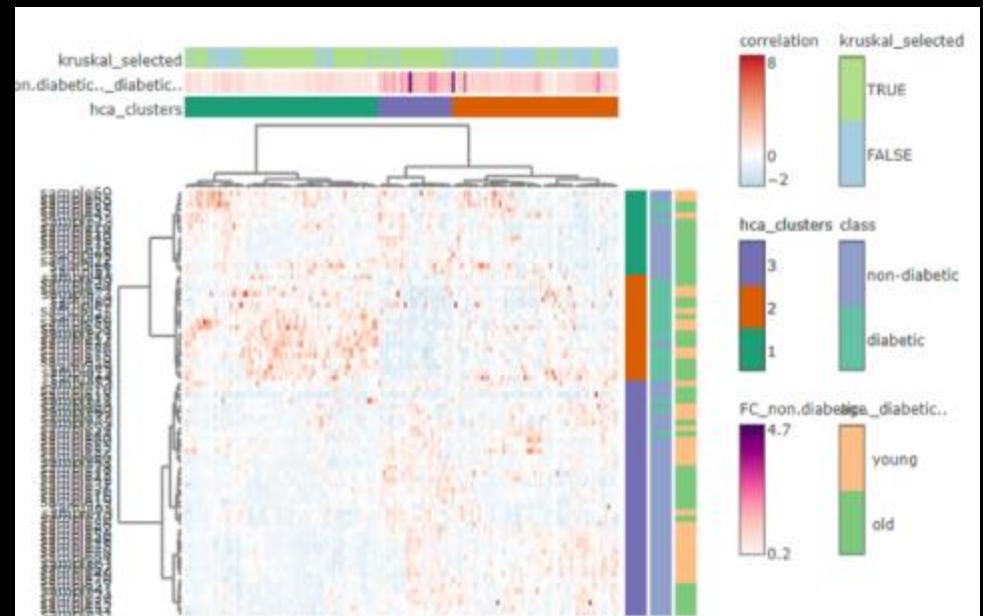
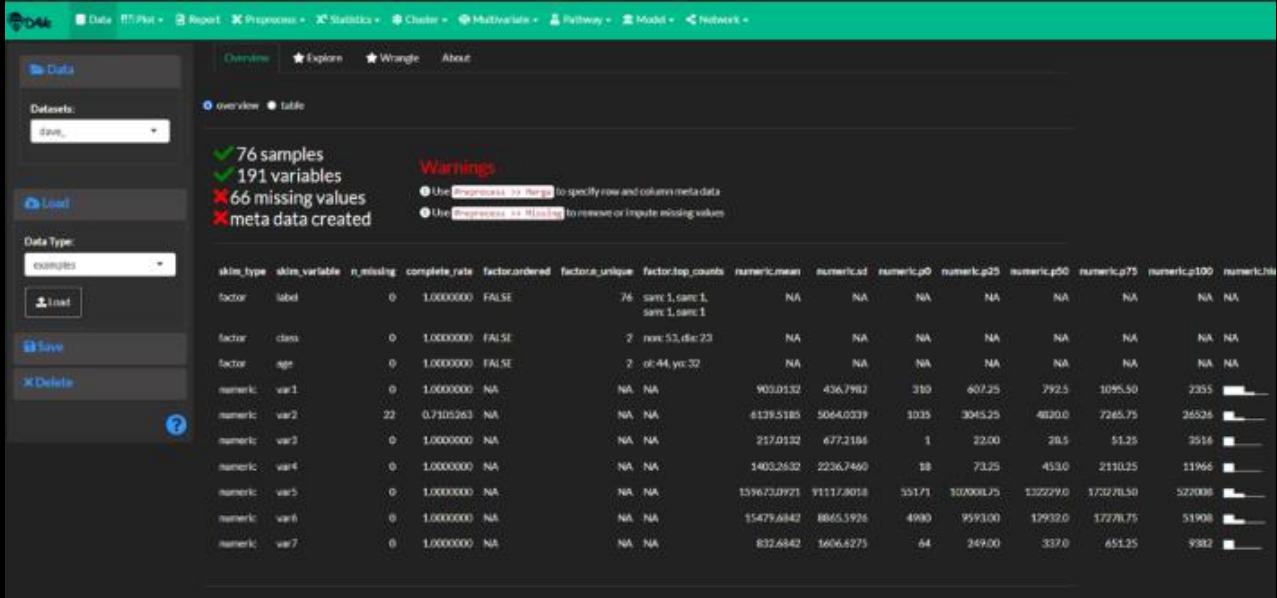
## ANOVA

PCA

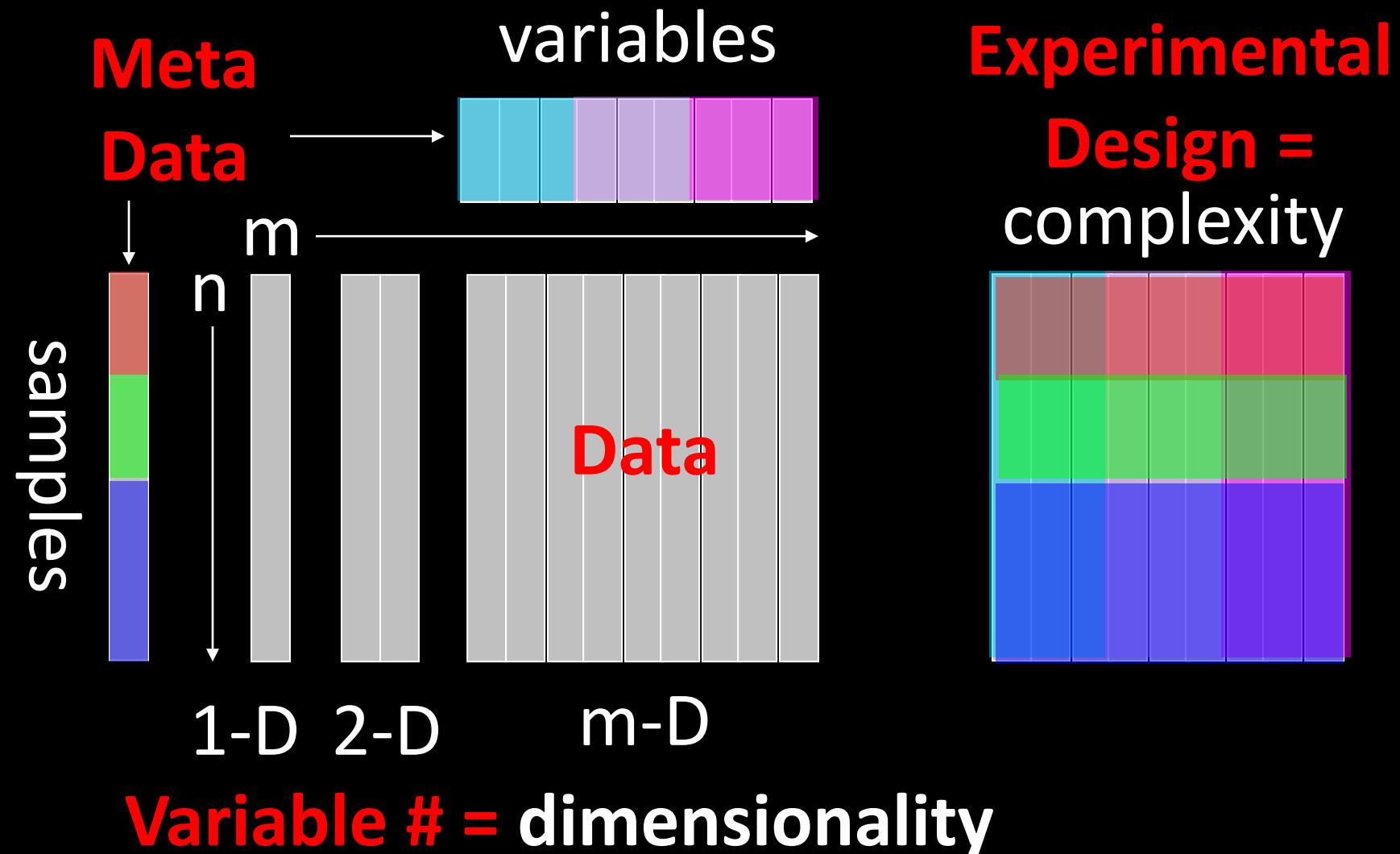
PLS

# Topics

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



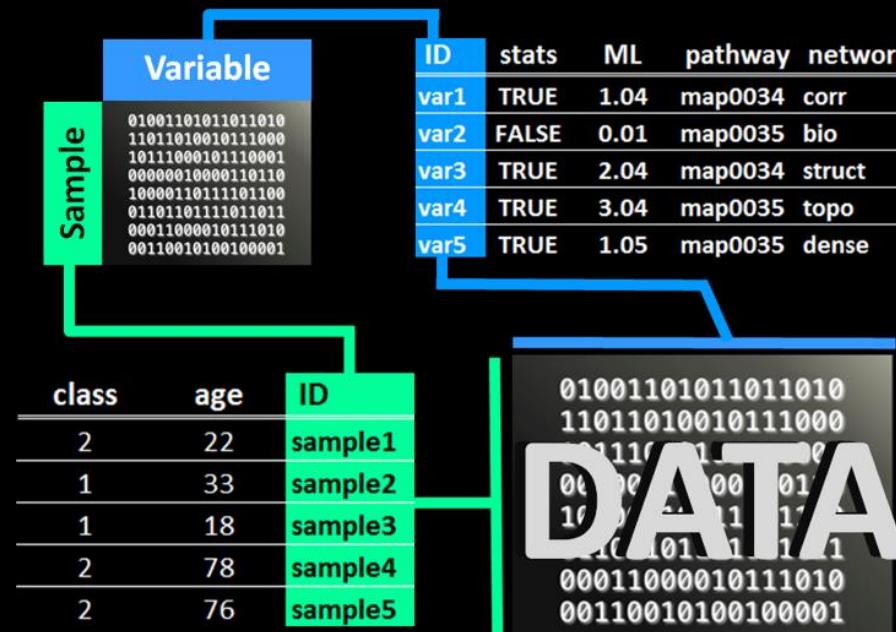
# How to think about data complexity



# Data preprocessing

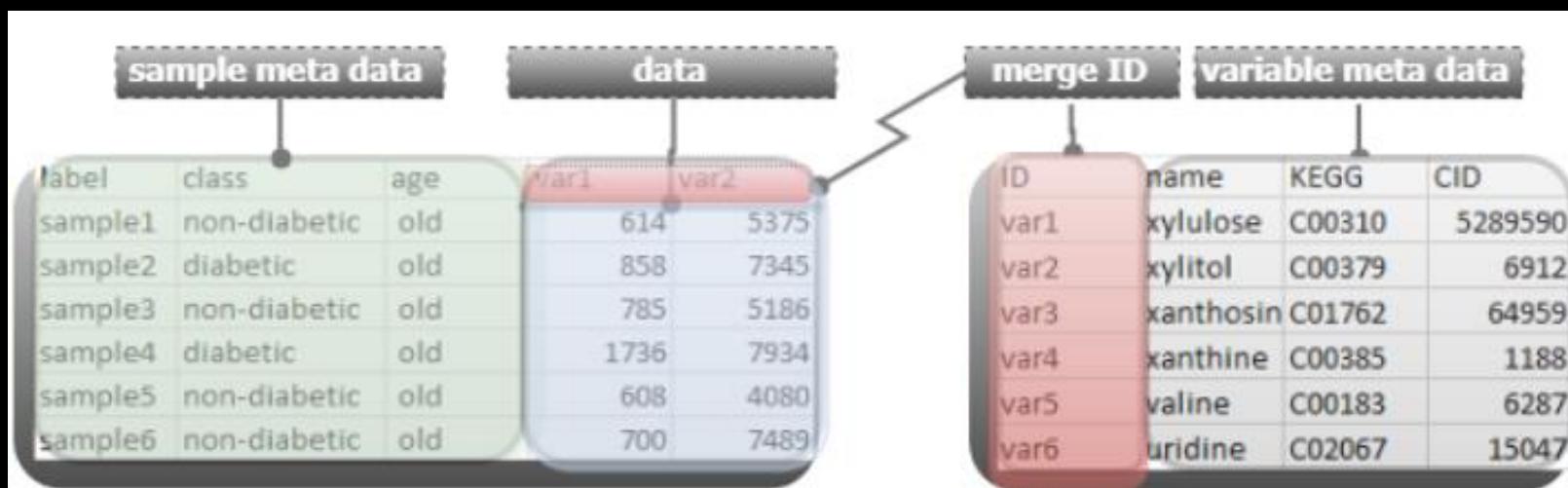
define

- data
- row meta data
- column meta data



remove and/or impute

- missing values



# Your turn

Follow along with the following tutorial:

[https://creativedatasolutions.github.io/CDS.courses/courses/network\\_mapping\\_101/docs/partial/preprocess/](https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/preprocess/)

# Differential expression

compare

- class means

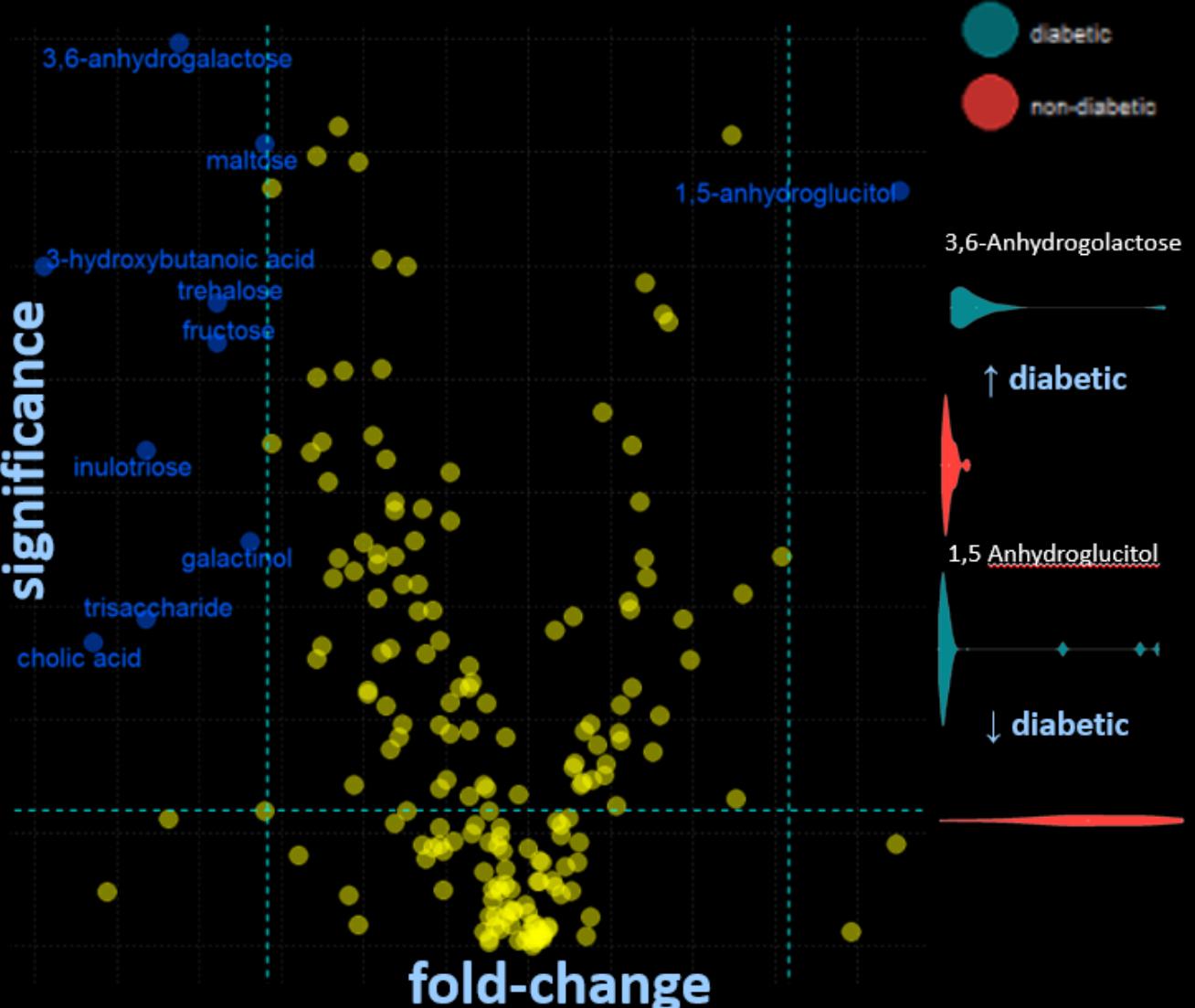
identify

- significant differences

visualize

- volcano plots
- violin plots

Simplest representation: two-class comparison



# Your turn

Follow along with the following tutorial:

[https://creativedatasolutions.github.io/CDS.courses/courses/network\\_mapping\\_101/docs/partial/statistics/](https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/statistics/)

# Hierarchical clustering (HCA)

group

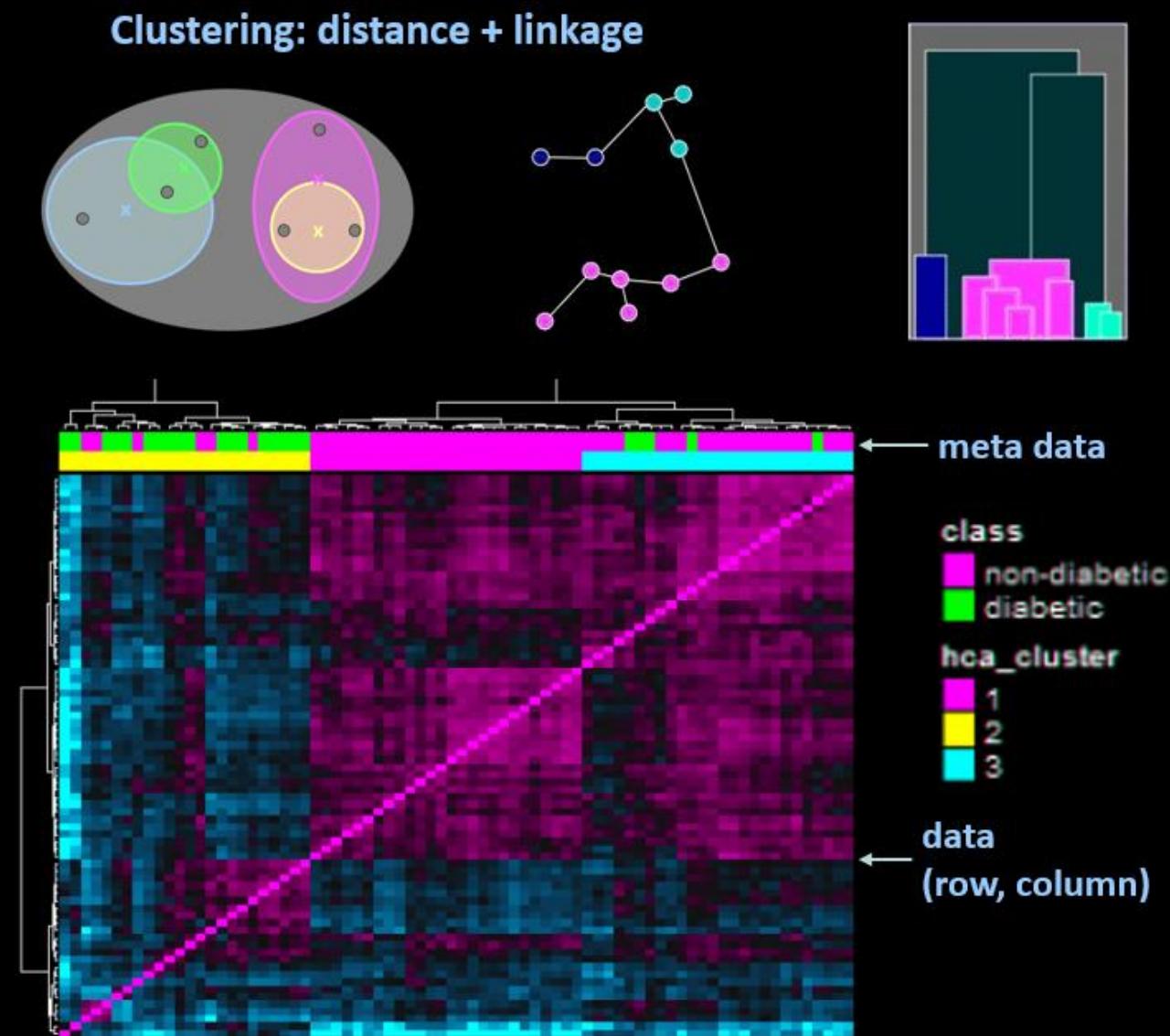
- samples and/or variables

define similarity

- correlation
- distance
- linkage

visualize

- heatmaps
- dendograms



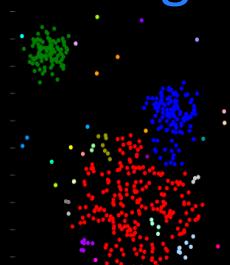
# Clustering basics

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

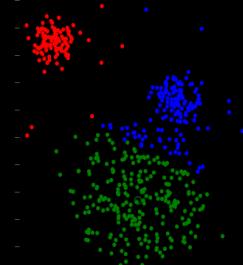
## approaches

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

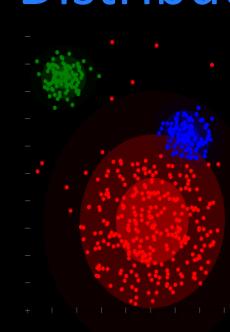
Linkage



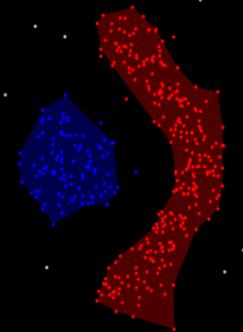
k-means



Distribution



Density



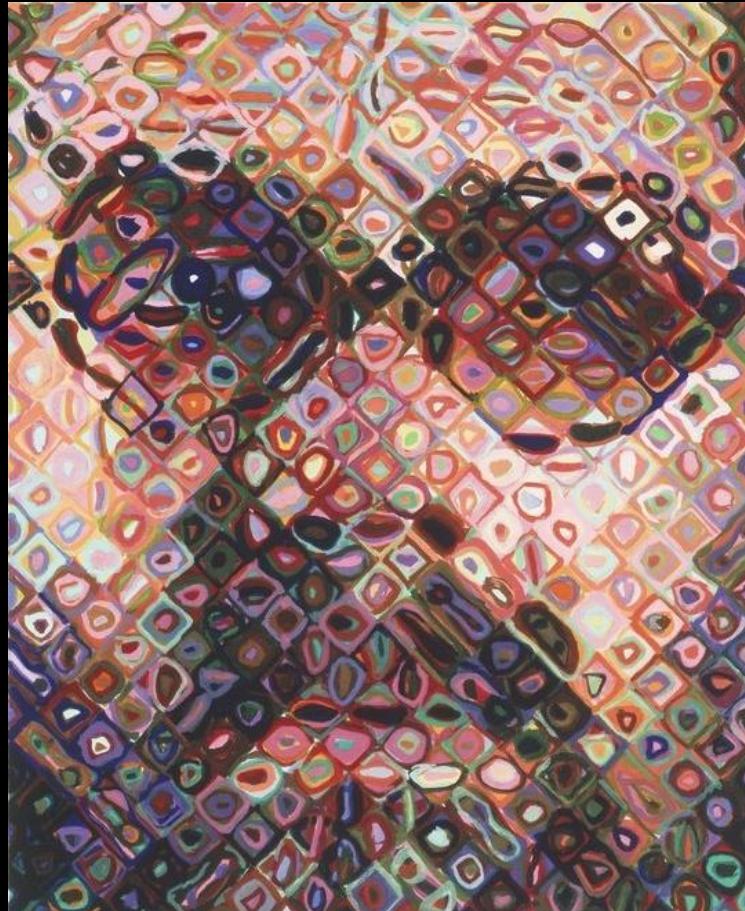
# HCA goals

identify

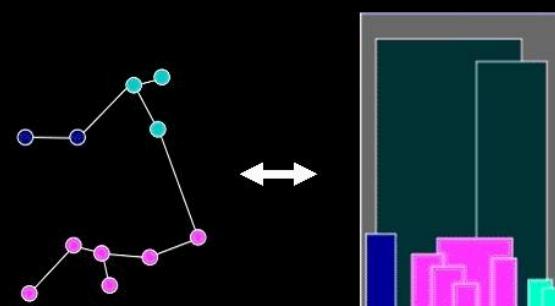
- patterns
- group structure
- relationships

evaluate and refine hypothesis

reduce complexity



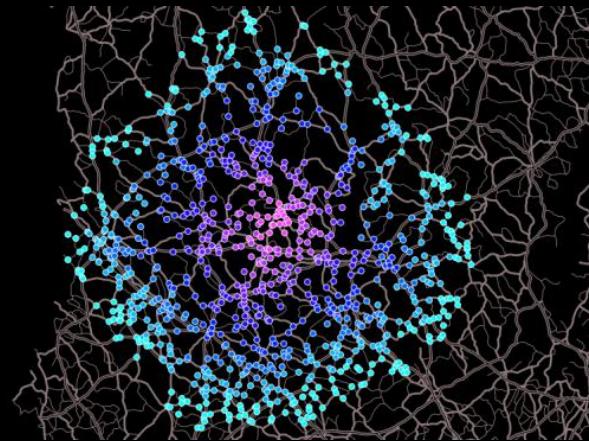
Artist: Chuck Close



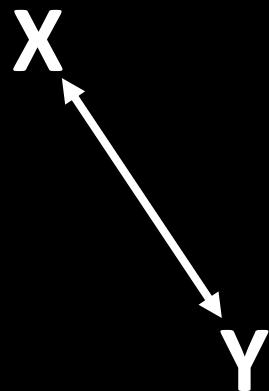
# HCA methods

## distance

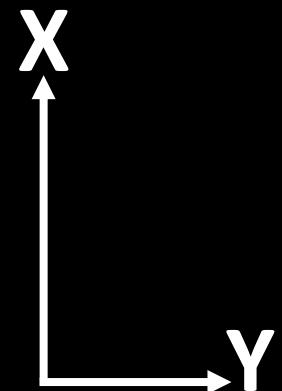
- defines “nearness” or similarity



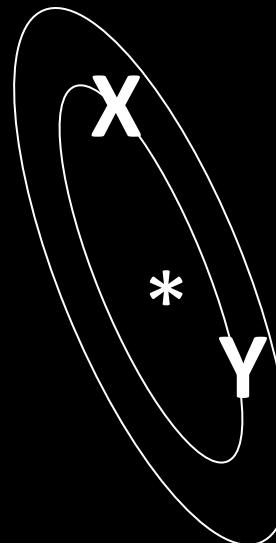
Euclidean



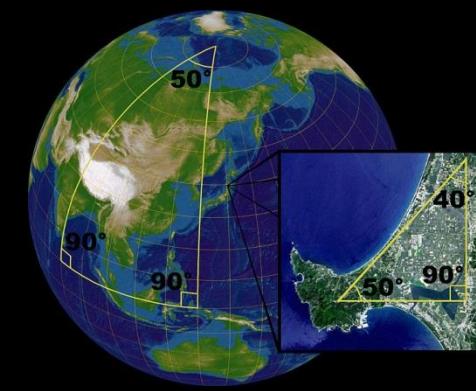
Manhattan



Mahalanobis



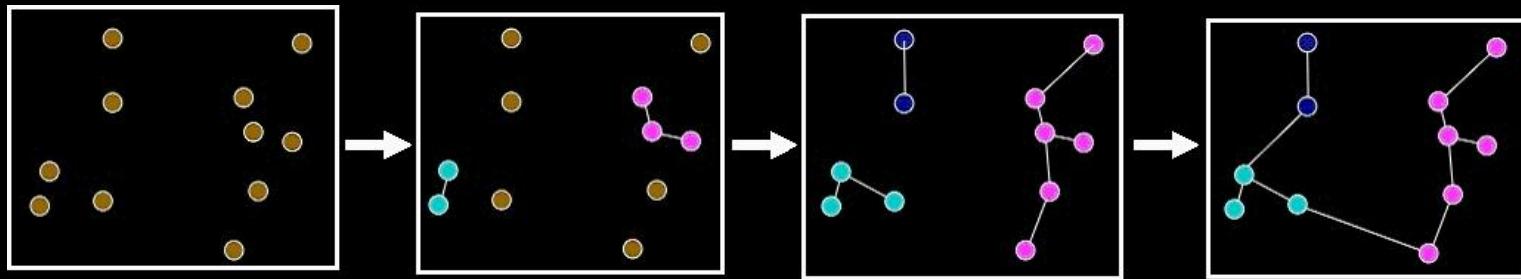
non-Euclidean



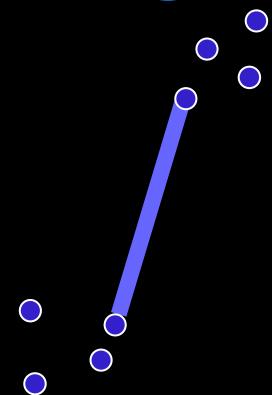
# HCA methods

## linkage or agglomeration

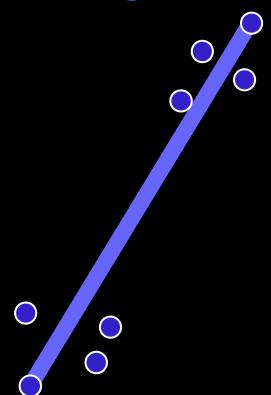
- how samples or variables are connected or grouped



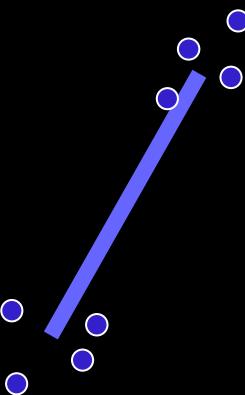
**single**



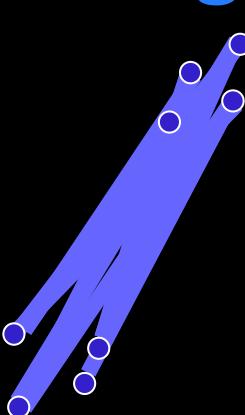
**complete**



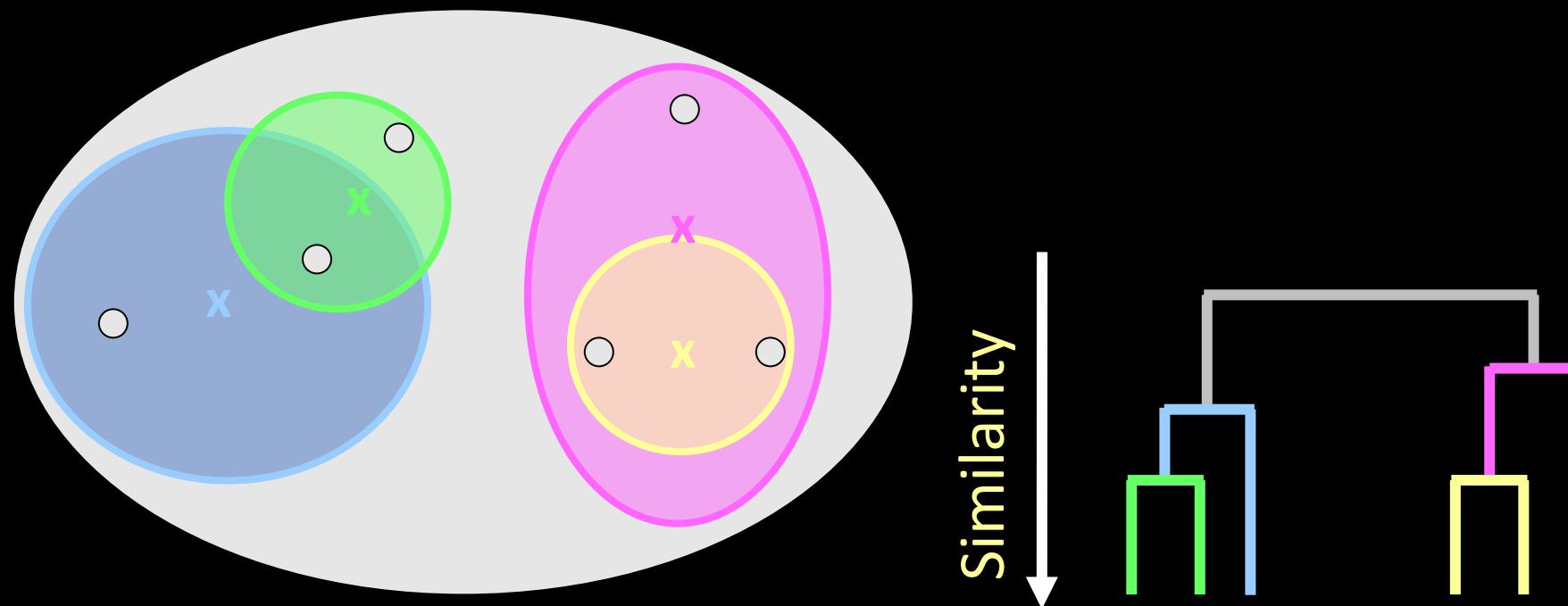
**centroid**



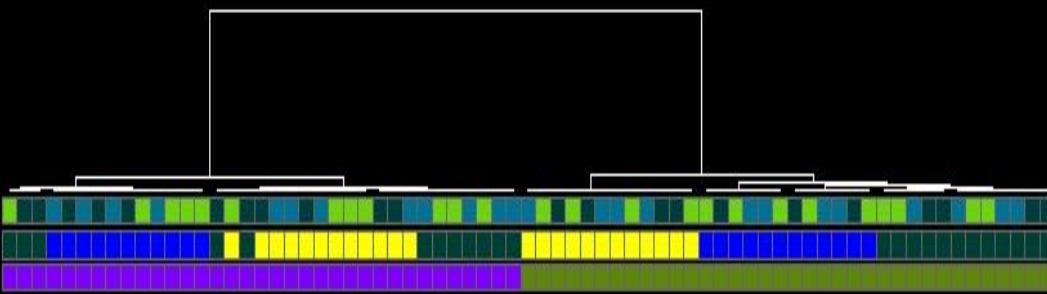
**average**



# HCA process

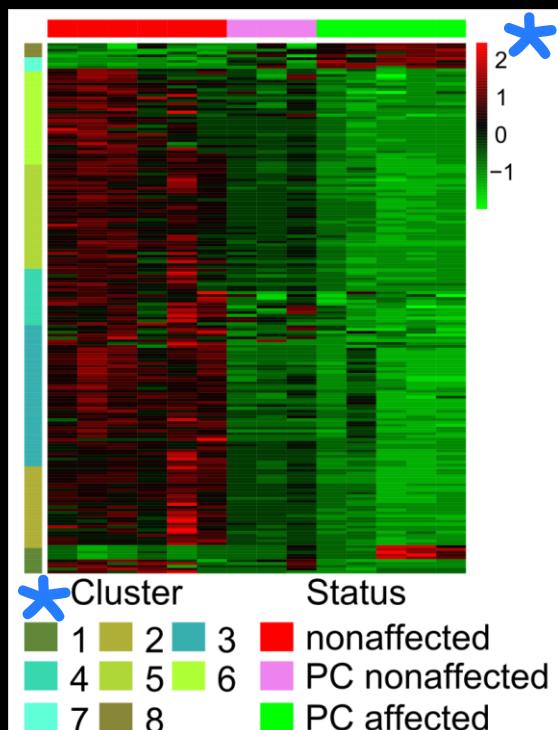
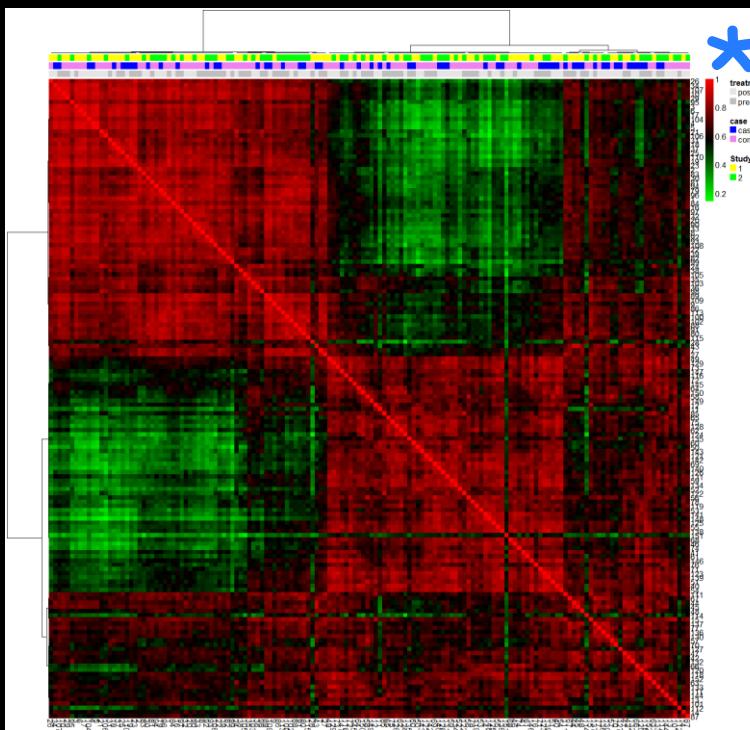


# HCA interpretation



overview

confirmation



How does my metadata  match my data structure?

# Your turn

Follow along with the following tutorial:

[https://creativedatasolutions.github.io/CDS.courses/courses/network\\_mapping\\_101/docs/partial/clustering/#heirarchical-clustering](https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/clustering/#heirarchical-clustering)

# Principal Components Analysis (PCA)

reduce

- dimensionality

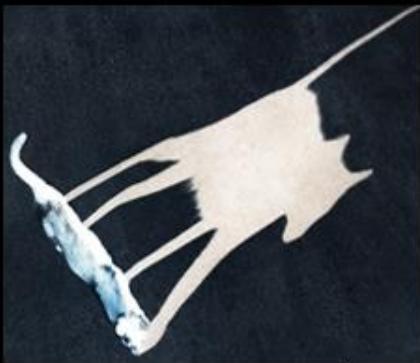
maximize

- variance explained

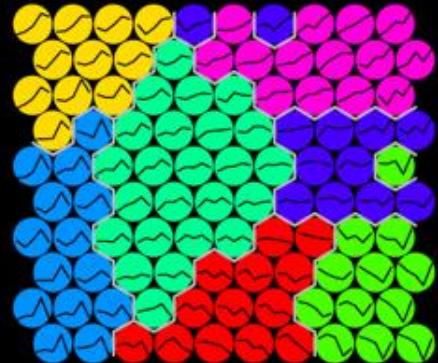
visualize

- variance explained
- outliers
- sample scores
- variable loadings

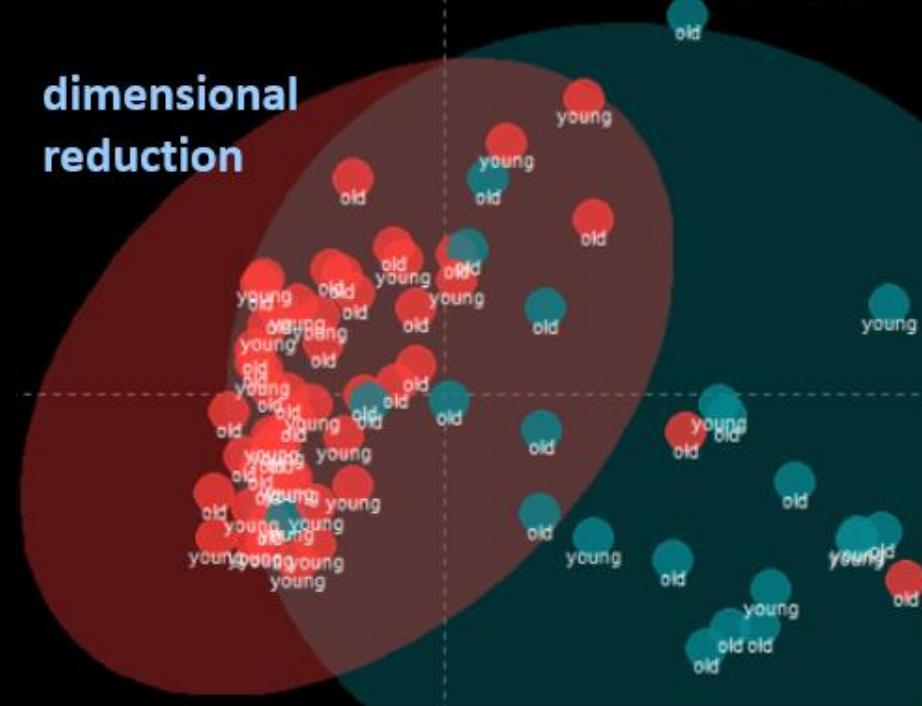
projection



similarity



dimensional reduction



diabetic  
non-diabetic

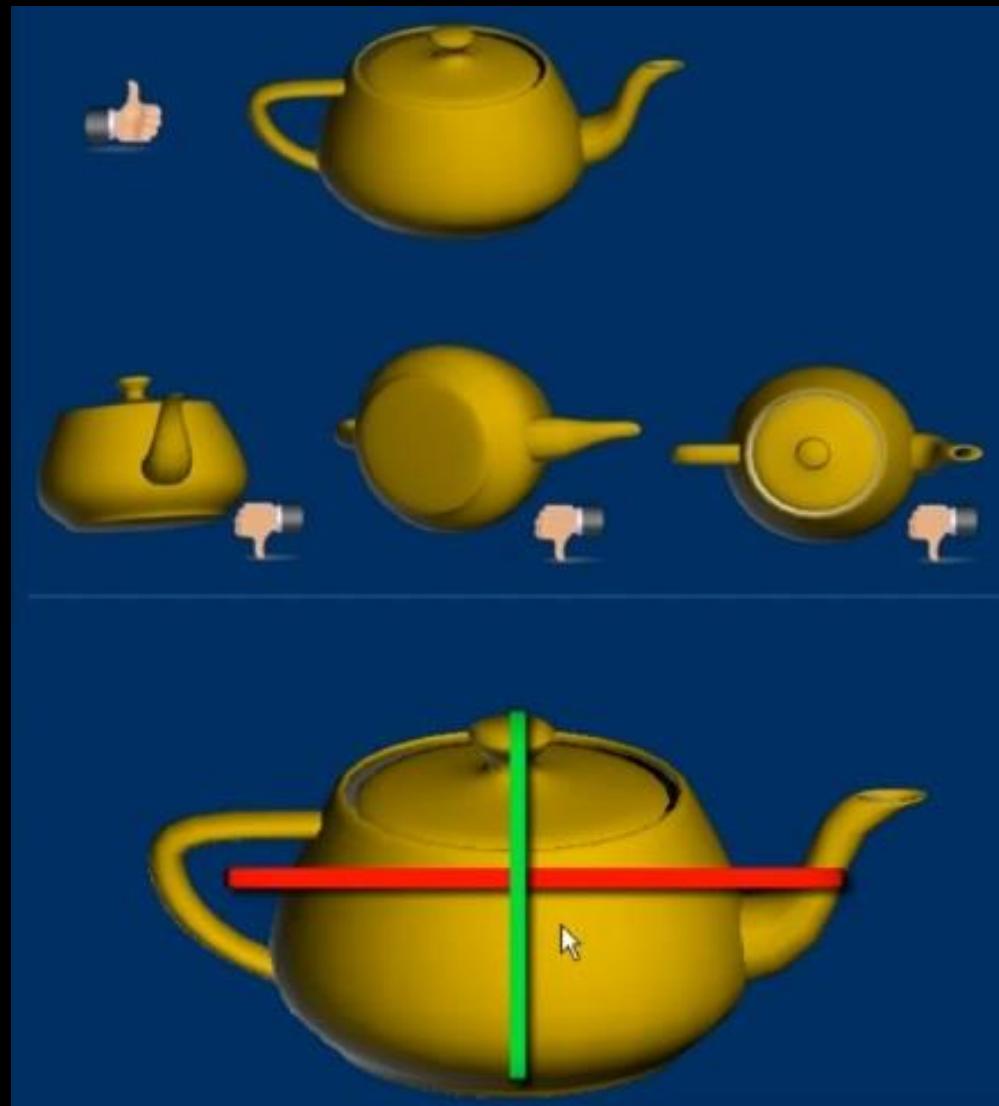
# PCA goals

Principal Components (PCs)

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

results

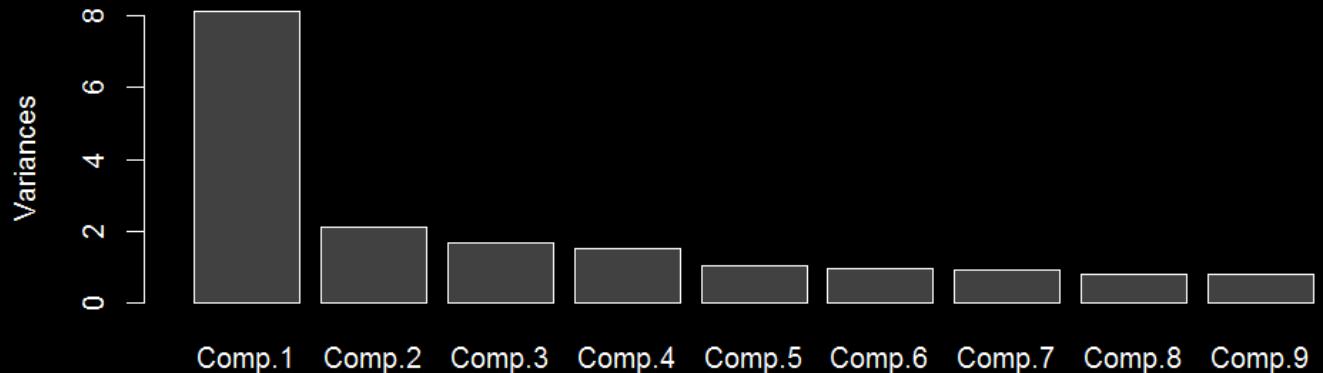
1. eigenvalues = variance explained
2. scores = new coordinates for samples (rows)
3. loadings = linear combination of original variables



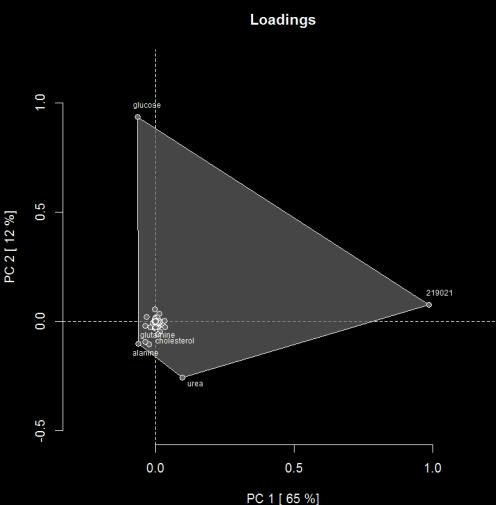
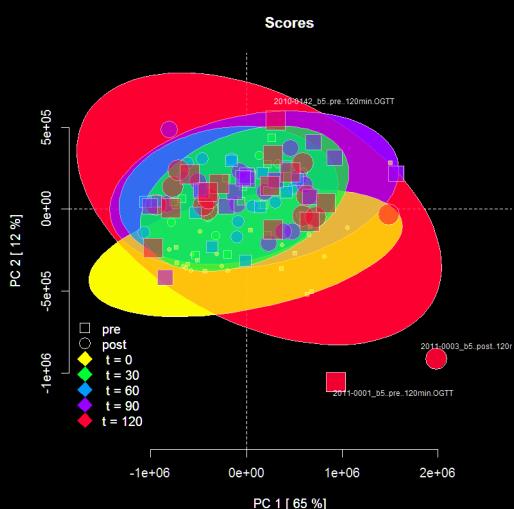
James X. Li, 2009, VisuMap Tech.

# PCA interpretation

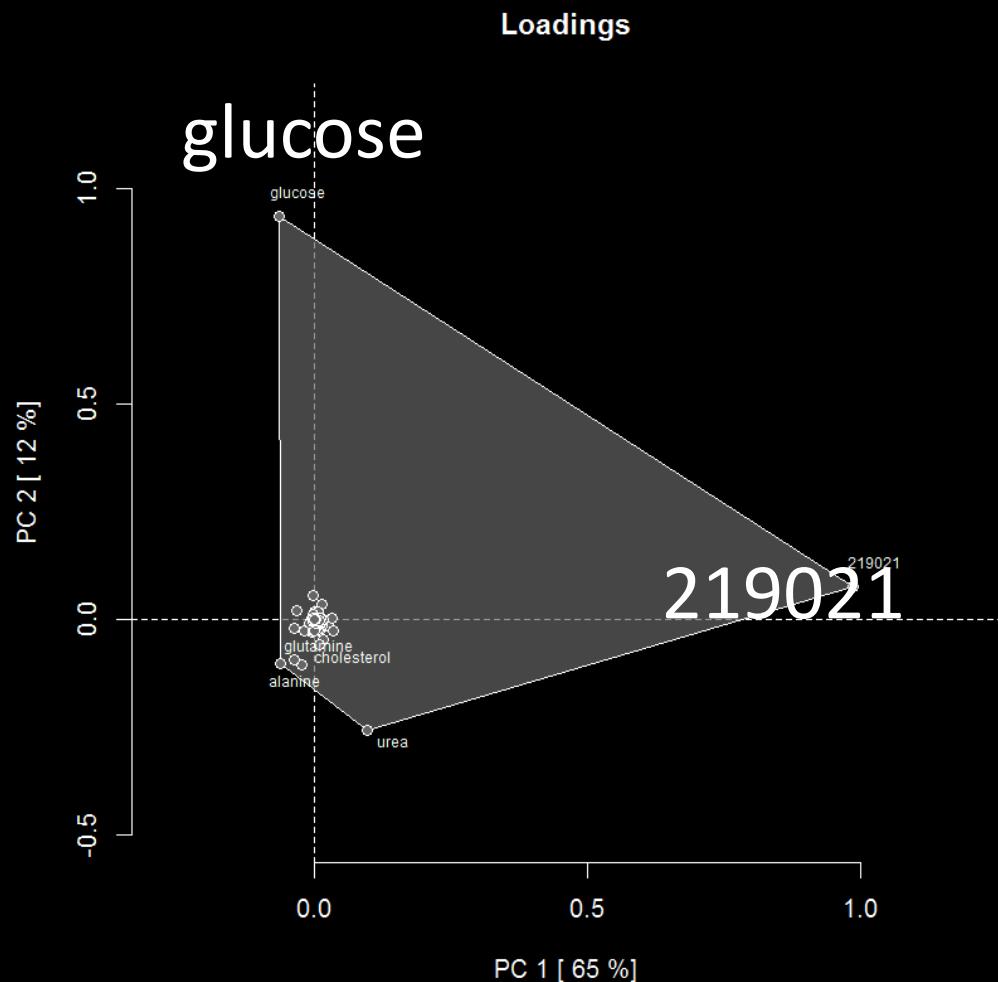
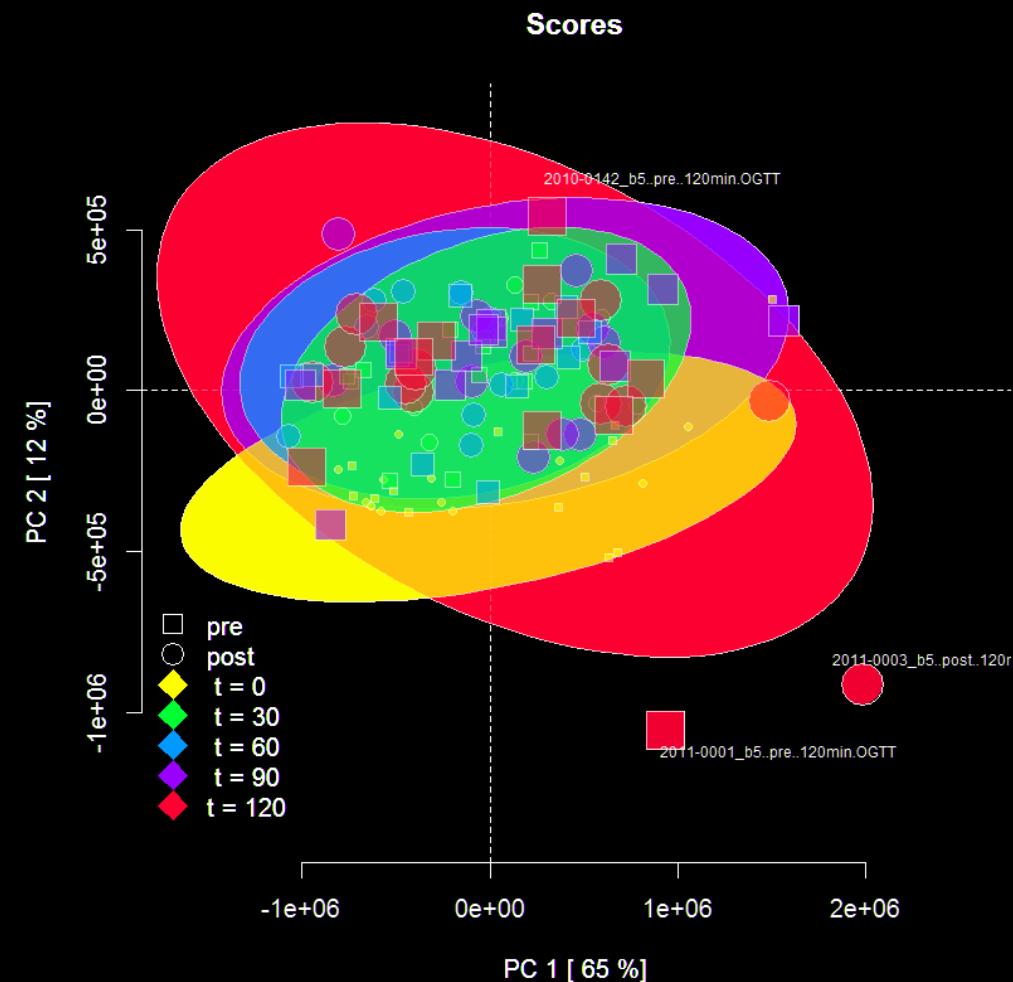
Variance explained (eigenvalues)



Row (sample) scores and column (variable) loadings



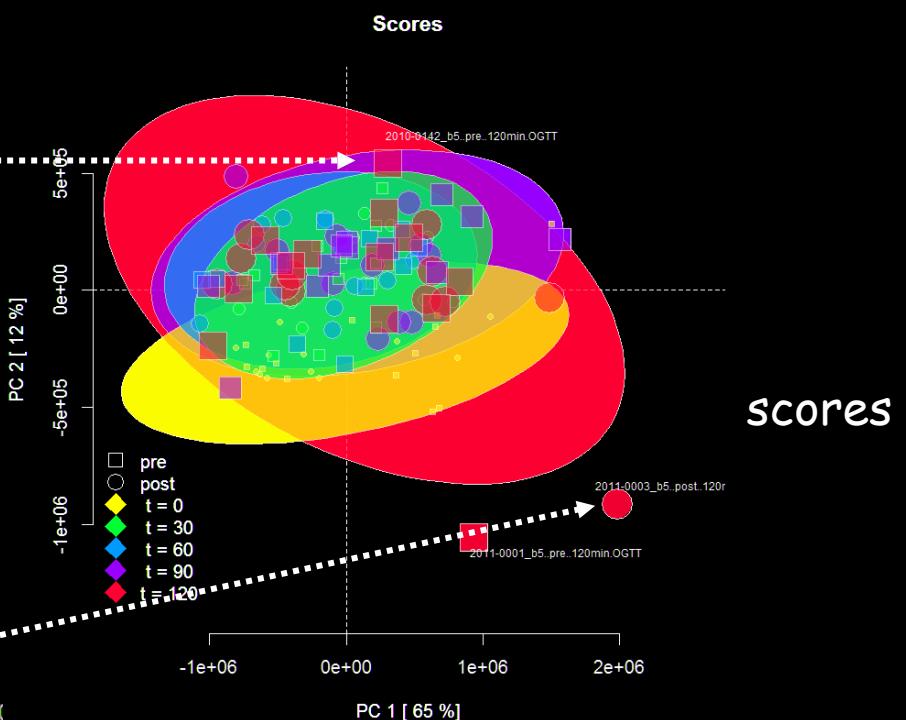
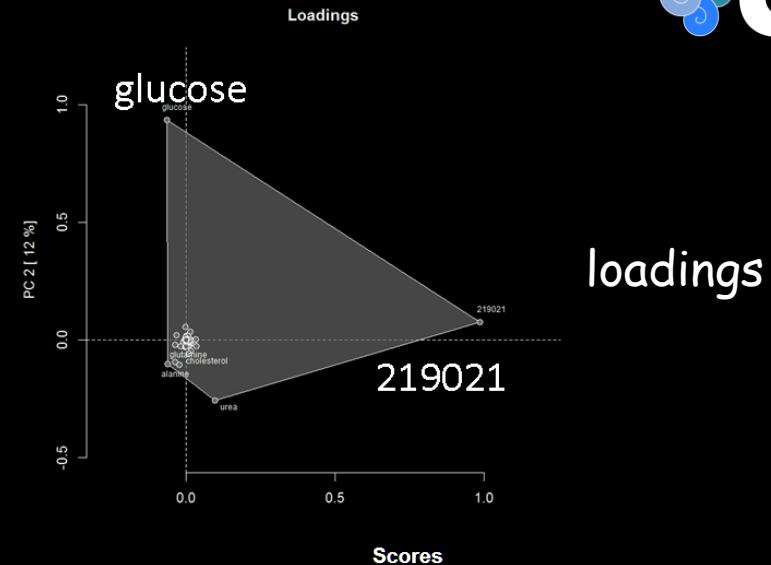
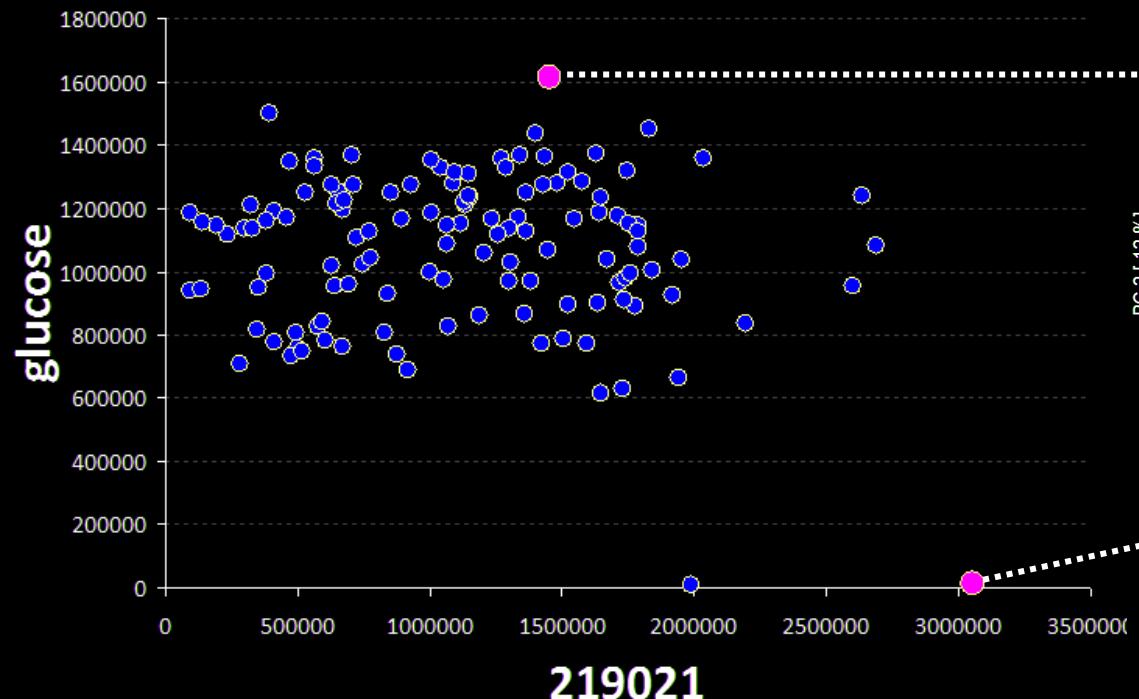
# PCA example



\*no scaling or centering

# Relationship between scores and loadings

top loading variable's scatterplot



# Your turn

Follow along with the following tutorial:

[https://creativedatasolutions.github.io/CDS.courses/courses/network\\_mapping\\_101/docs/partial/multivariate/#pca](https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/multivariate/#pca)

# Machine learning

**predict**

- sample classification

**optimize**

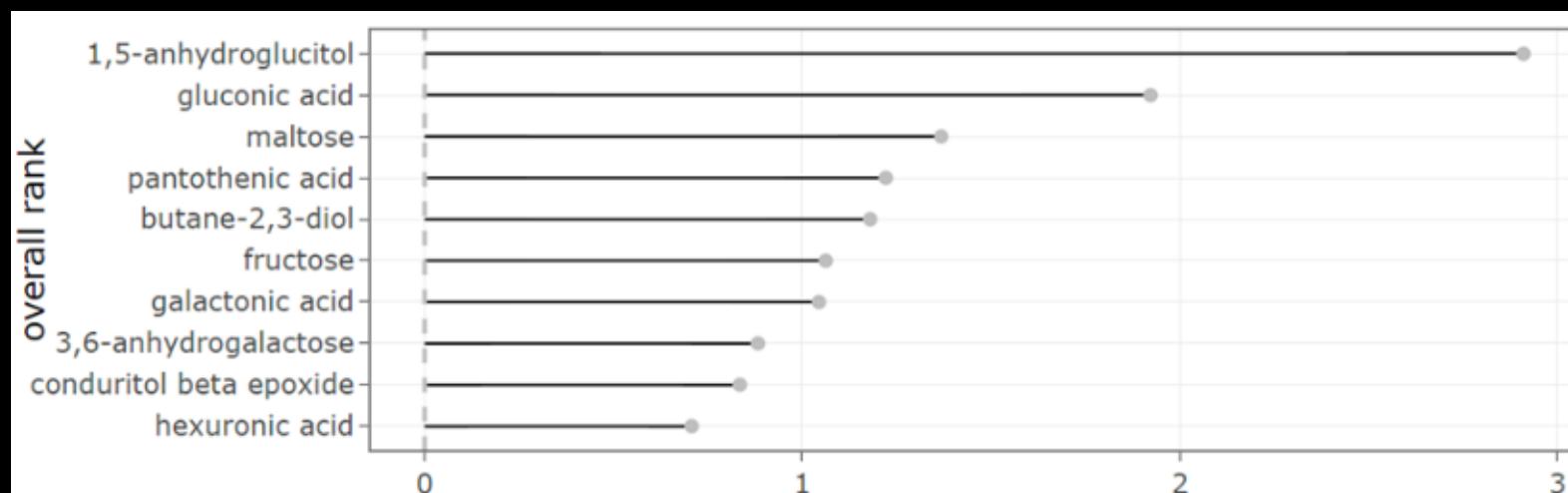
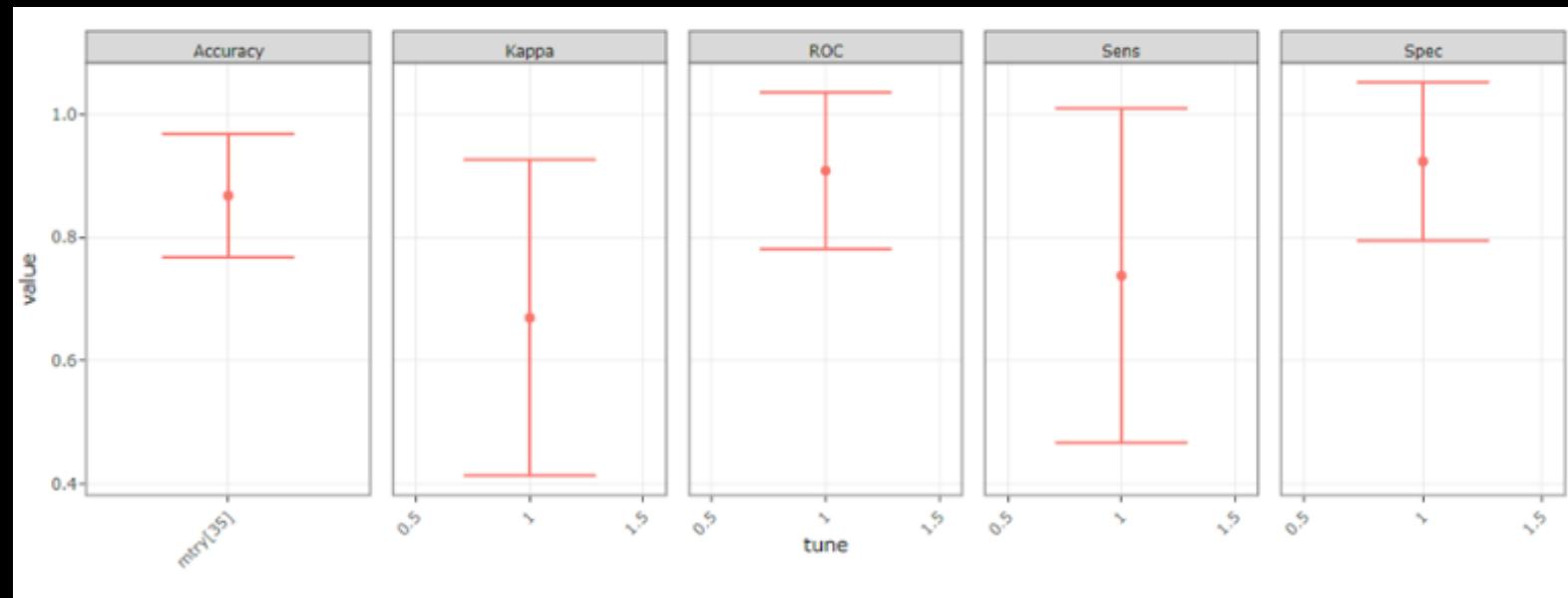
- model performance

**select**

- important features

**visualize**

- model performance
- feature importance



# Your turn

Follow along with the following tutorial:

[https://creativedatasolutions.github.io/CDS.courses/courses/network\\_mapping\\_101/docs/partial/model/](https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/model/)

# Network analysis

## network mapping

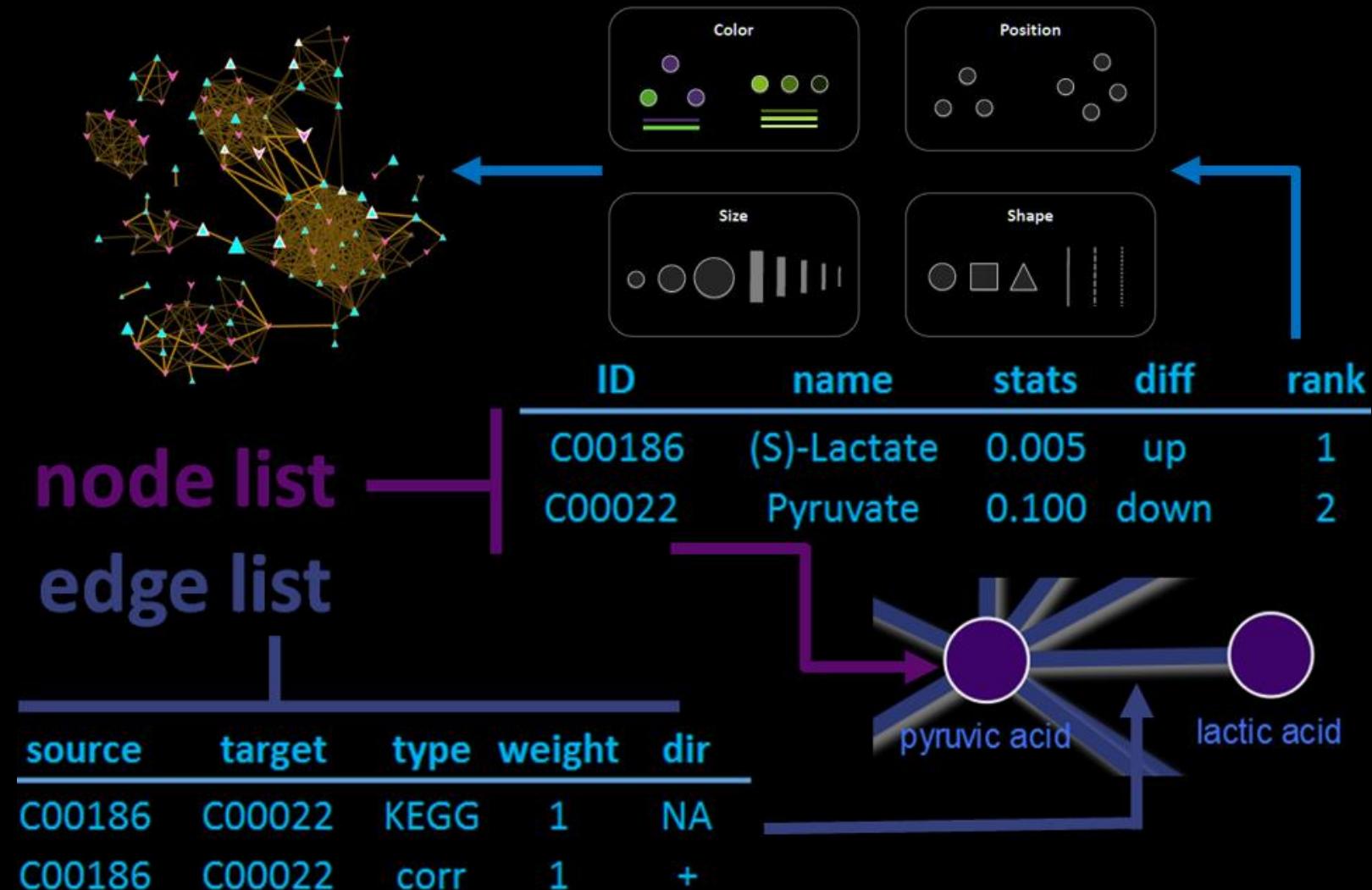
- transform variables

## network calculation

- regularized correlation
- biochemical
- structural similarity
- model performance

## visualize

- interactive networks



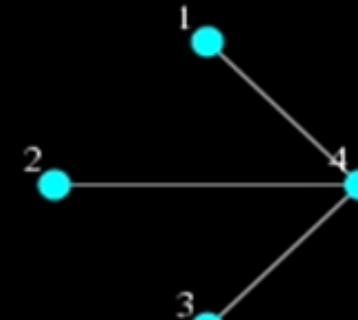
# Components for network mapping

## connections (edges)

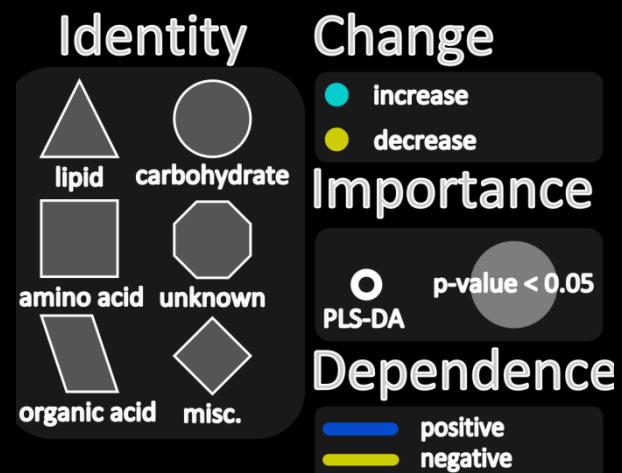
- empirical dependency (correlation)
- biochemical (substrate/product)
- chemical similarity
- ...

## nodes (vertices)

- magnitude
- importance
- direction
- relationships
- ...



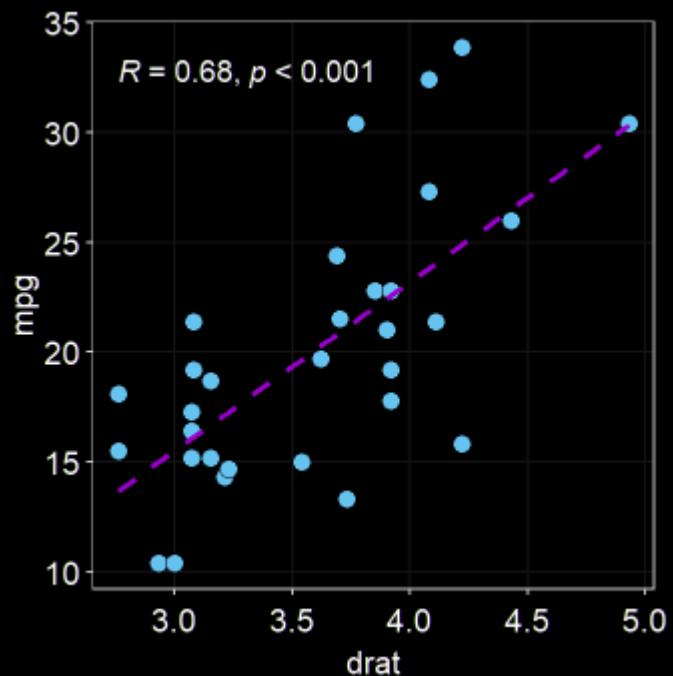
	1	2	3	4
1	0	0	0	1
2	0	0	0	1
3	0	0	0	1
4	1	1	1	0



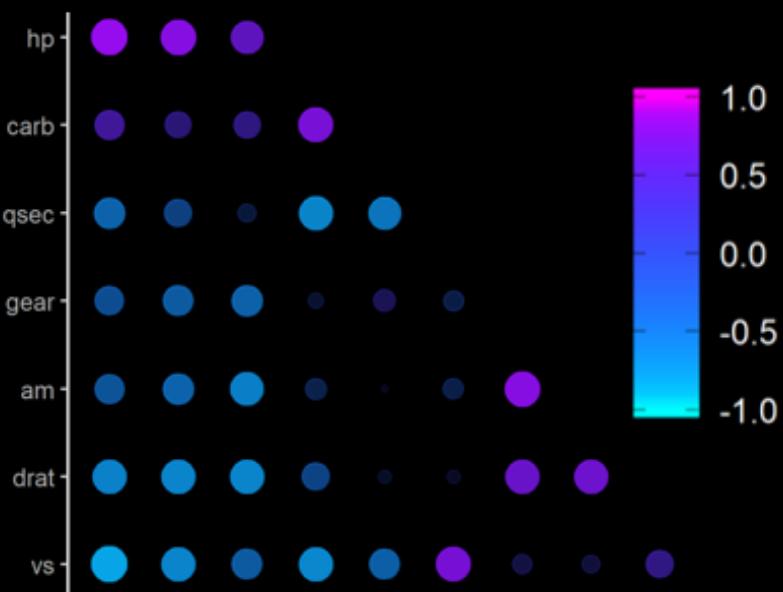
# Correlation networks

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

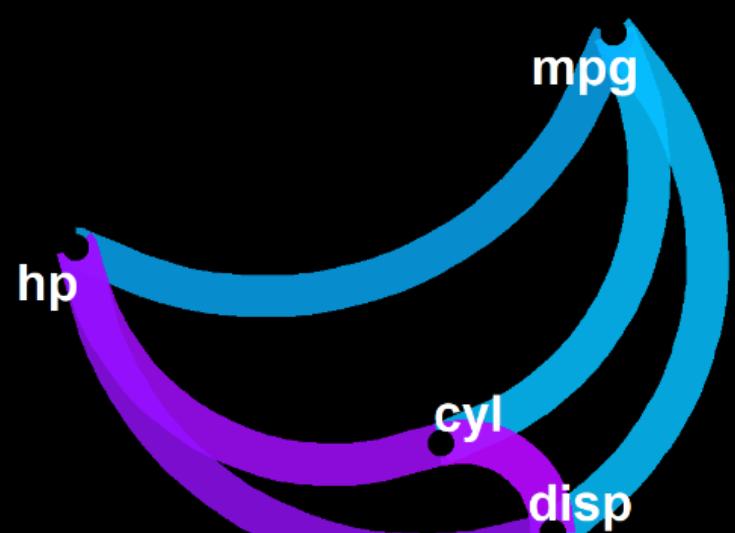
bivariate



multivariate

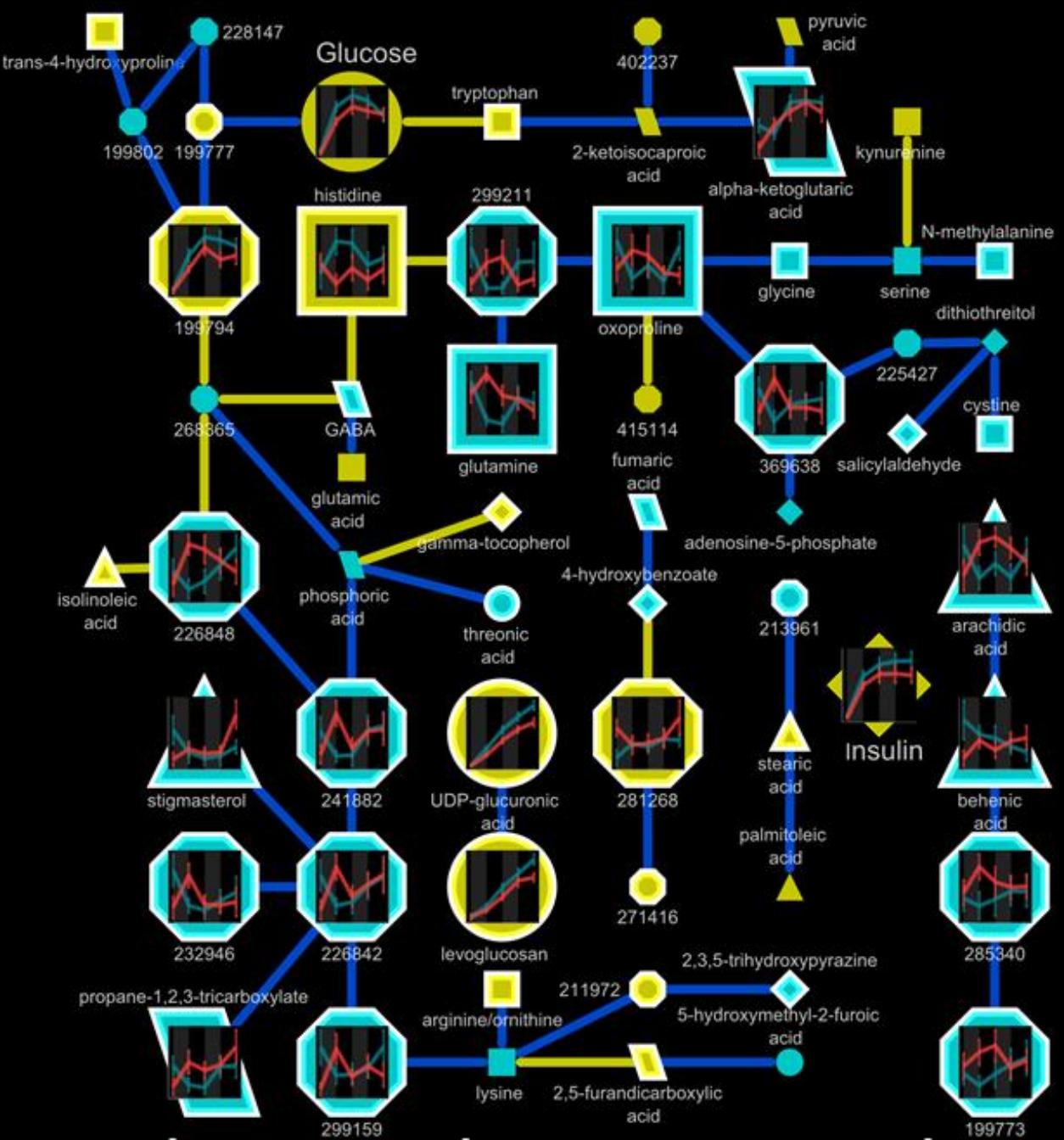


network

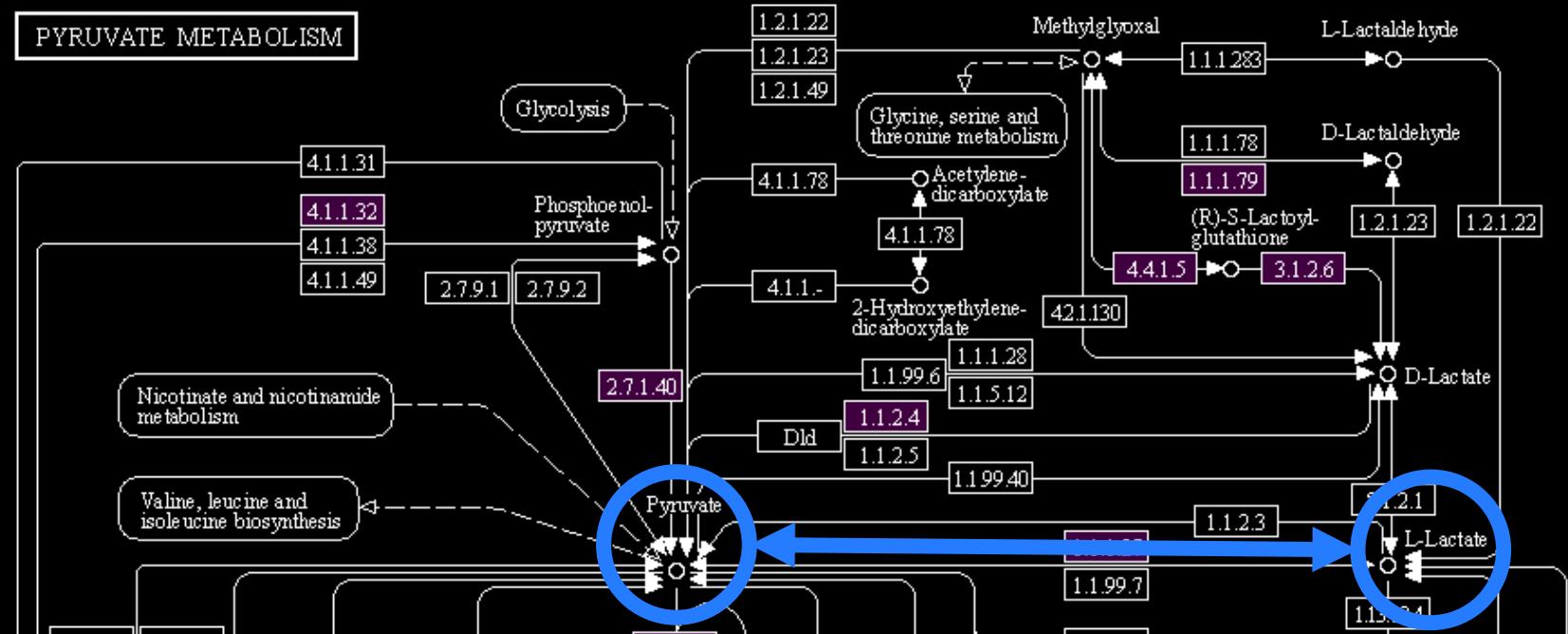
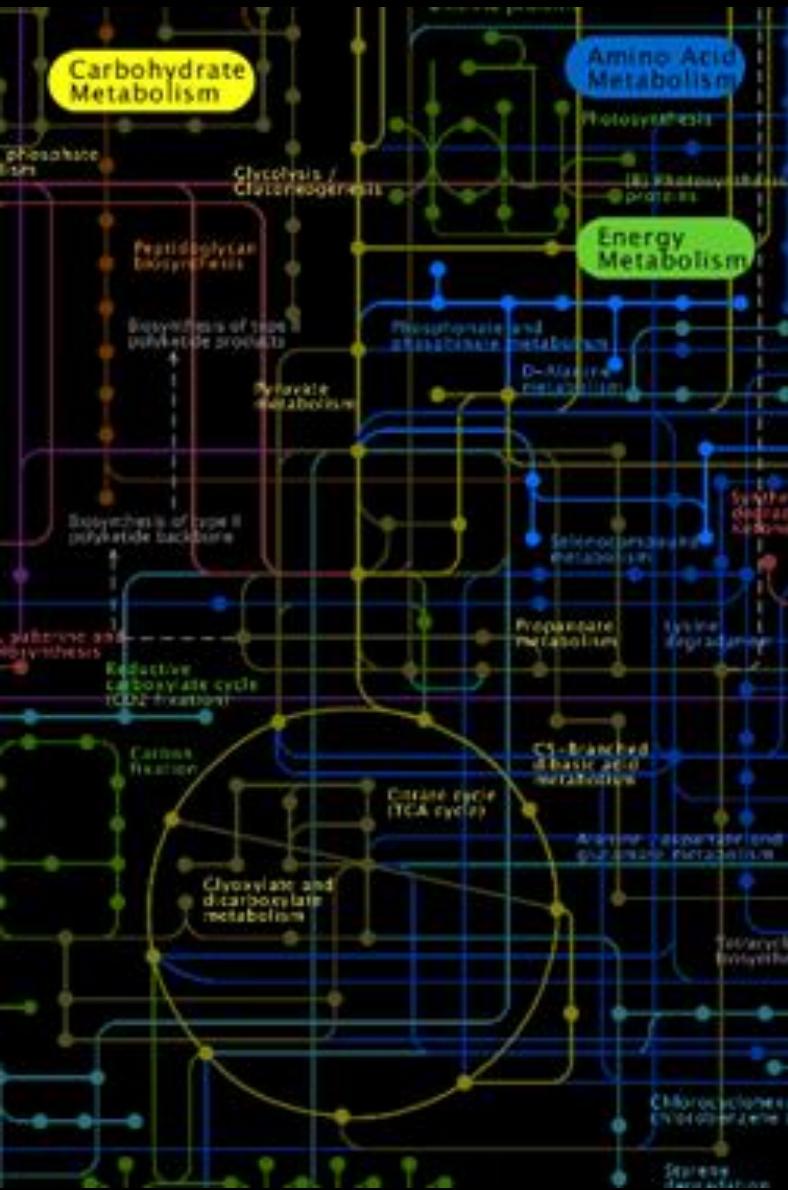


# Correlation example

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



# Biochemical networks



# Structural similarity networks

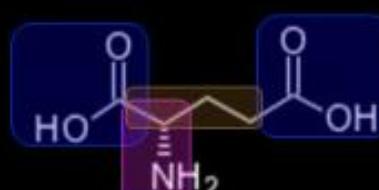
- Use structure to generate molecular fingerprint
- Calculate similarities between metabolites based on fingerprint
- PubChem service for similarity calculations

[http://pubchem.ncbi.nlm.nih.gov//score\\_matrix/score\\_matrix.cgi](http://pubchem.ncbi.nlm.nih.gov//score_matrix/score_matrix.cgi)

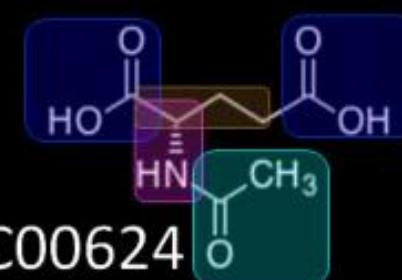
- online tools

<http://uranus.fiehnlab.ucdavis.edu:8080/MetaMapp/homePage>

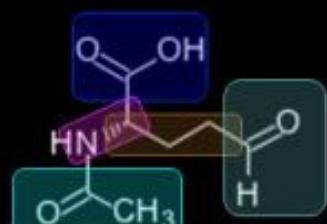
**Chemical mapping**  
of substructure comparison  
using PubChem



C00025



C00624

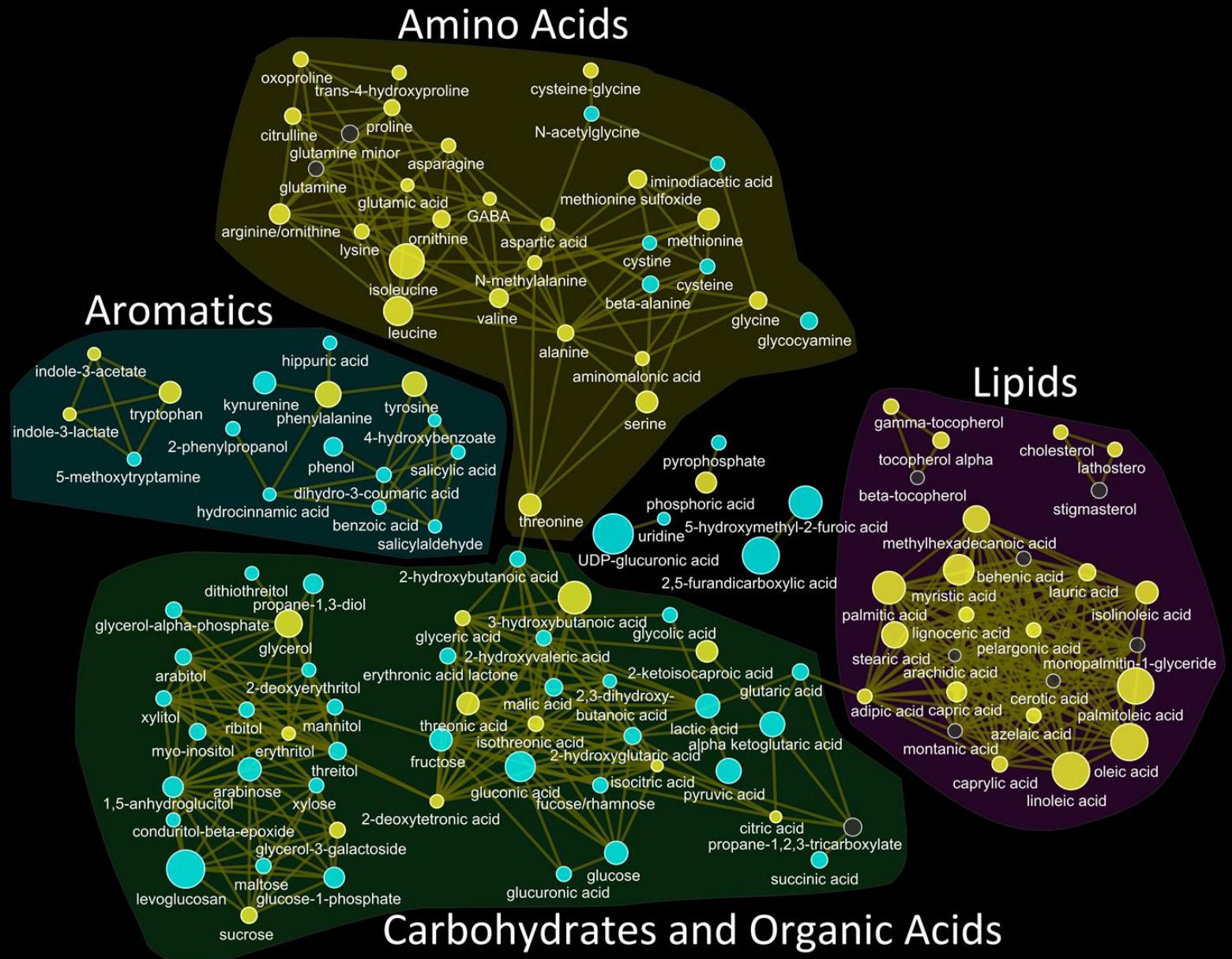


C01250

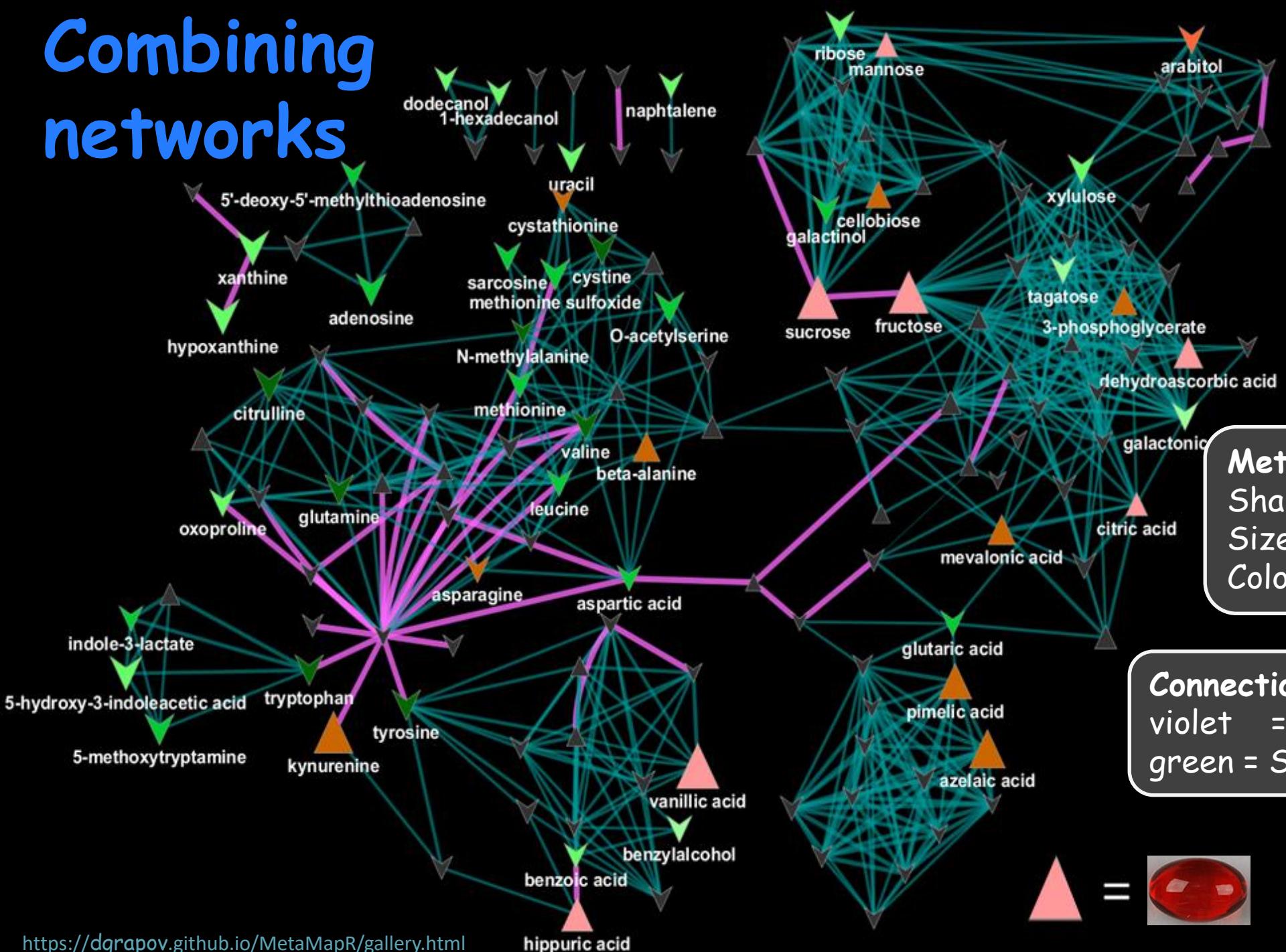
substructure matrix decomposition and  
Tanimoto chemical similarity calculations

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



# Your turn

Follow along with the following tutorial:

[https://creativedatasolutions.github.io/CDS.courses/courses/network\\_mapping\\_101/docs/partial/network/](https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/network/)

# Network refinement and visualization

learn

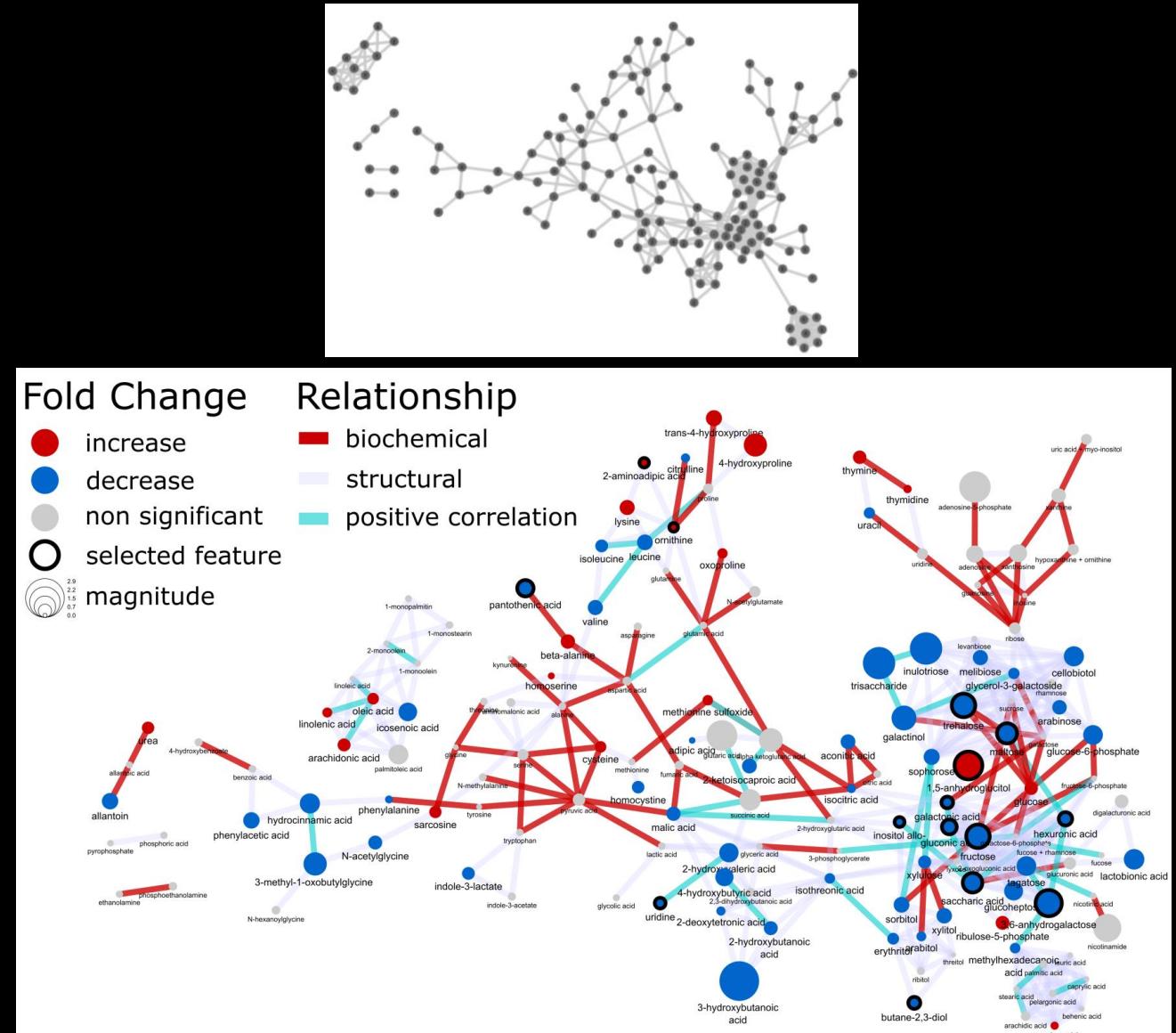
- Cytoscape basics

map variables to

- node attributes
- edge attributes

optimize

- layout
- legend
- publication quality figure



# Your turn

Follow along with the following tutorial:

[https://creativedatasolutions.github.io/CDS.courses/courses/network\\_mapping\\_101/docs/partial/cytoscape/](https://creativedatasolutions.github.io/CDS.courses/courses/network_mapping_101/docs/partial/cytoscape/)