Molecular networks (intro)

Wessel van Wieringen
w.n.van.wieringen@vu.nl

Department of Epidemiology and Biostatistics, VUmc
& Department of Mathematics, VU University
Amsterdam, The Netherlands

vrije Universiteit

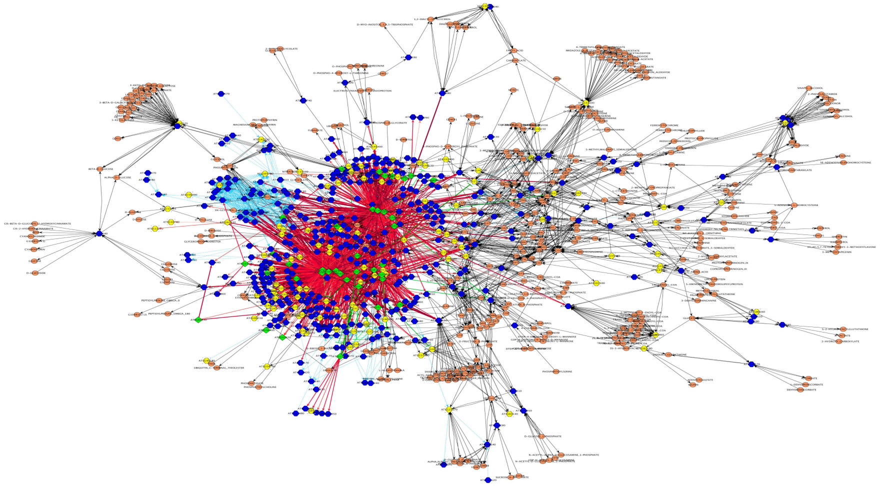
VU medisch centrum

What?

Molecular biology aims to understand the molecular processes that occur in the cell. That is, e.g.:
→ which molecules present in the cell interact?
→ how is this coordinated?

For many cellular processes, it is unknown which genes play what role.

Goal
Reconstruct the cellular regulatory network.

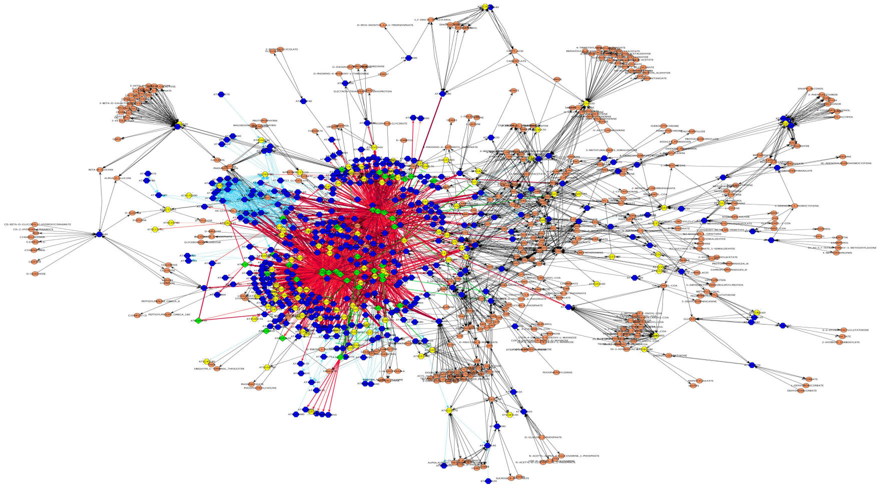


Why?

Negative motivation

→ Differentially expressed genes: boring!

→ Yet another clustering?

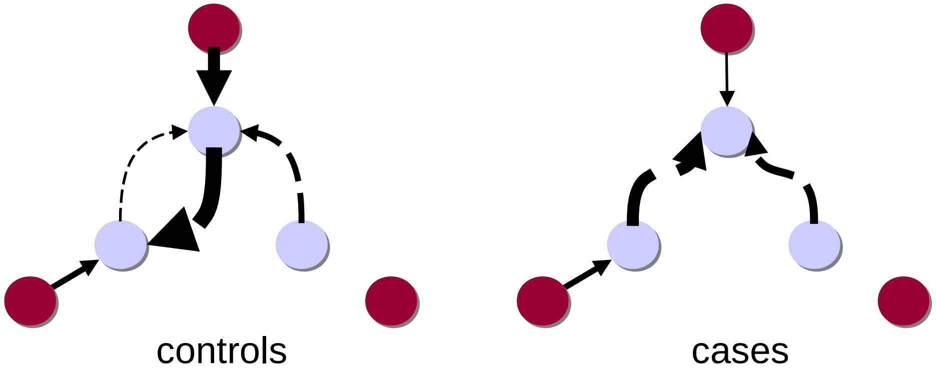


Positive motivation

→ Fancy plot.

→ Different insight.

→ Network medicine
(e.g. biomarker:
gene-gene interaction)



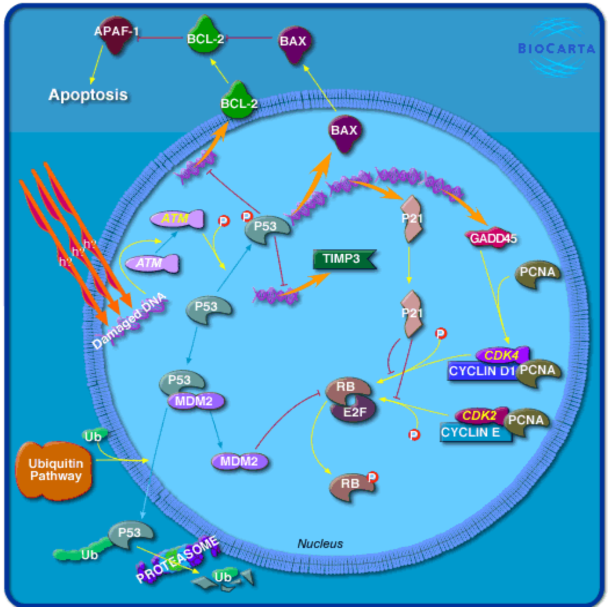
Pathway = network

Pathway = chain of chemical reactions (that processes a signal)

≈ a set of genes believed to carry one function

Pathways are loosely defined using repositories, such as:

- KEGG
- BioCarta
- GennMapp
- Reactome
- GO
- String



BioCarta: p53 signalling pathway

How?

Download from repository

→ Which? Reliable?

→ Knowledge is incomplete and biased towards a few well-studied pathways.

→ Does it apply to your situation?

Reconstruct from data

→ Data is a rare and valuable commodity!

Synthesis

Reconstruct from data with the repository as a suggestion

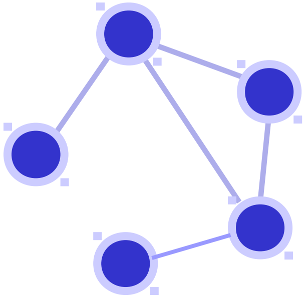
Network

Pathways are represented by a *graph* or *network*.

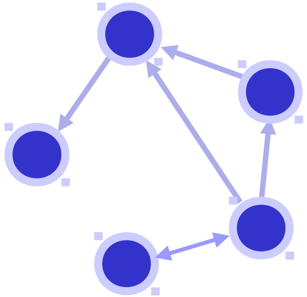
● *node* or *vertex*, representing a gene.

→ ↔ *edge* or *arrow*, representing an interaction between two genes.

— *undirected* and *directed* edges (≈ “association”)



undirected
(focus here)



directed

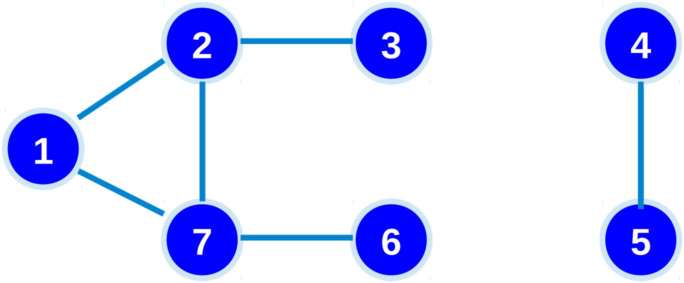
Network

Edge operationalization = direct relation
(Formally: conditional dependence)

Direct relation
Relation between two nodes without mediation of other nodes.

Indirect relation
Relation between two nodes through mediating other nodes.

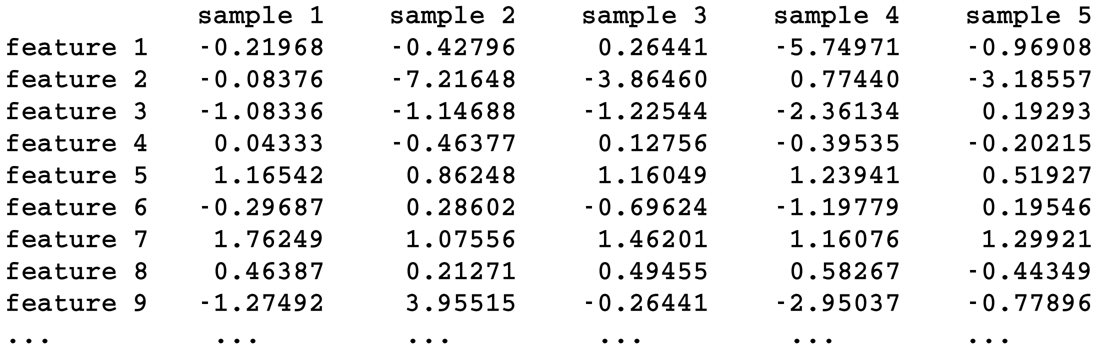
No relation
None of the above.



Relations
→ node 1 and 2: directly
→ node 3 and 6: indirectly
→ node 4 and 7: none
→ ...

With?

To reconstruct which genes interact, we have available:
→ molecular profiles of *n* samples,
→ each profile comprises *p* features.



```
             sample 1    sample 2    sample 3    sample 4    sample 5
feature 1   -0.21968    -0.42796     0.26441    -5.74971    -0.96908
feature 2   -0.08376    -7.21648    -3.86460     0.77440    -3.18557
feature 3   -1.08336    -1.14688    -1.22544    -2.36134     0.19293
feature 4    0.04333    -0.46377     0.12756    -0.39535    -0.20215
feature 5    1.16542     0.86248     1.16049     1.23941     0.51927
feature 6   -0.29687     0.28602    -0.69624    -1.19779     0.19546
feature 7    1.76249     1.07556     1.46201     1.16076     1.29921
feature 8    0.46387     0.21271     0.49455     0.58267    -0.44349
feature 9   -1.27492     3.95515    -0.26441    -2.95037    -0.77896
...          ...         ...         ...         ...         ...
```

Possibly with:
→ network suggestion,
→ group information,
→ temporal layout.

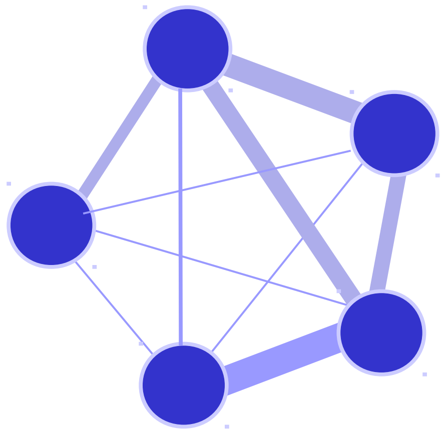
How?

Roadmap

data

```
          sample 1    sample 2     ...
gene 1    -0.21968    -0.42796     ...
gene 2    -0.08376    -7.21648     ...
gene 3    -1.08336    -1.14688     ...
gene 4     0.04333    -0.46377     ...
...      ...   ...    ...
```

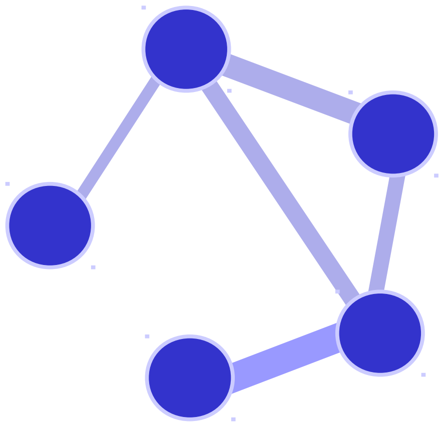
edge strength measure



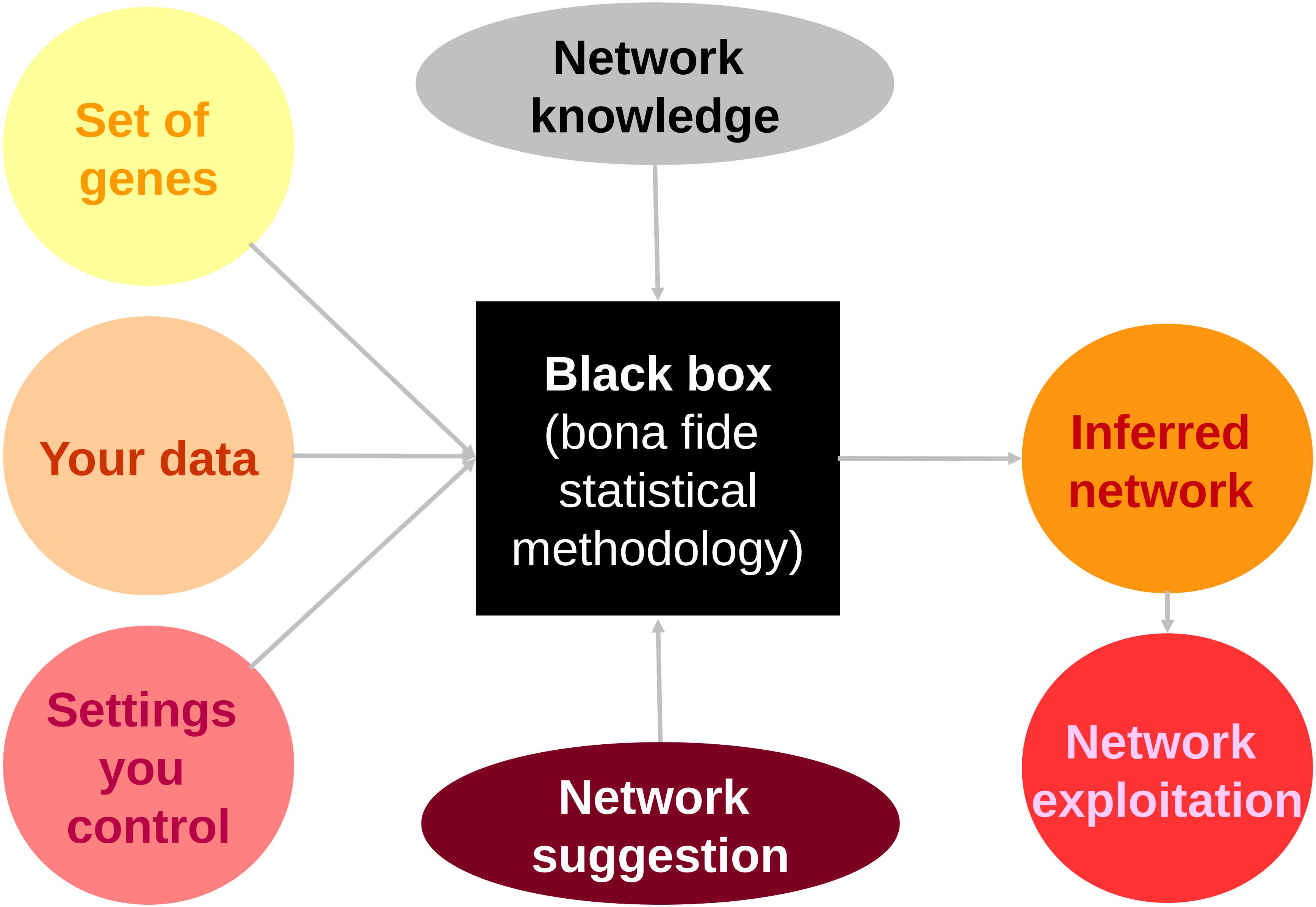
statistical test

edge strength significantly different from zero: *edge*!

inferred network



rags2ridges



Set of genes

Your data

Settings you control

Network knowledge

Black box (bona fide statistical methodology)

Network suggestion

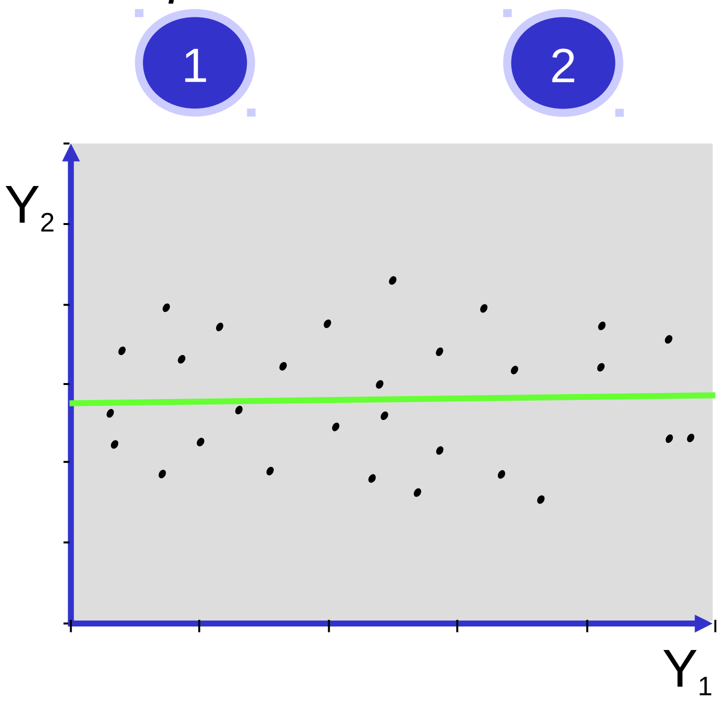
Inferred network

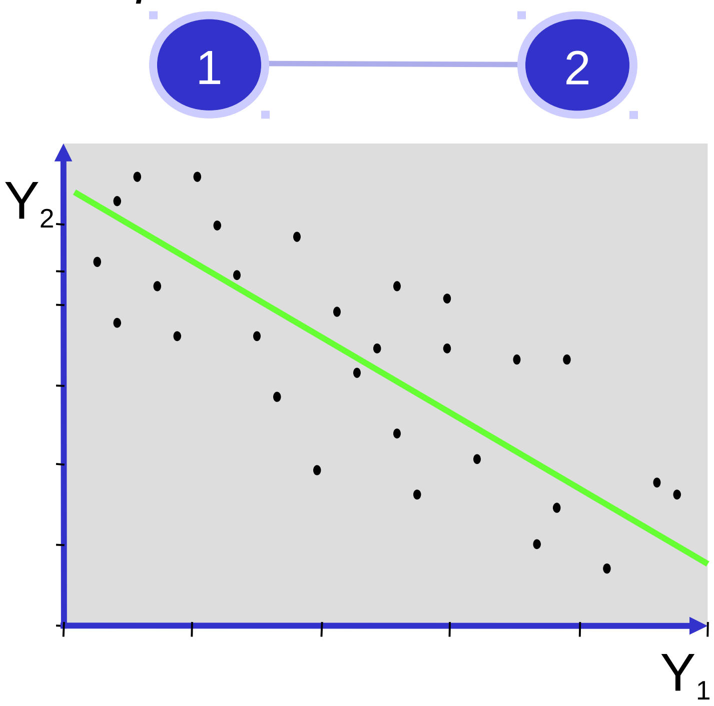
Network exploitation

Two-gene pathway

Two-gene pathway

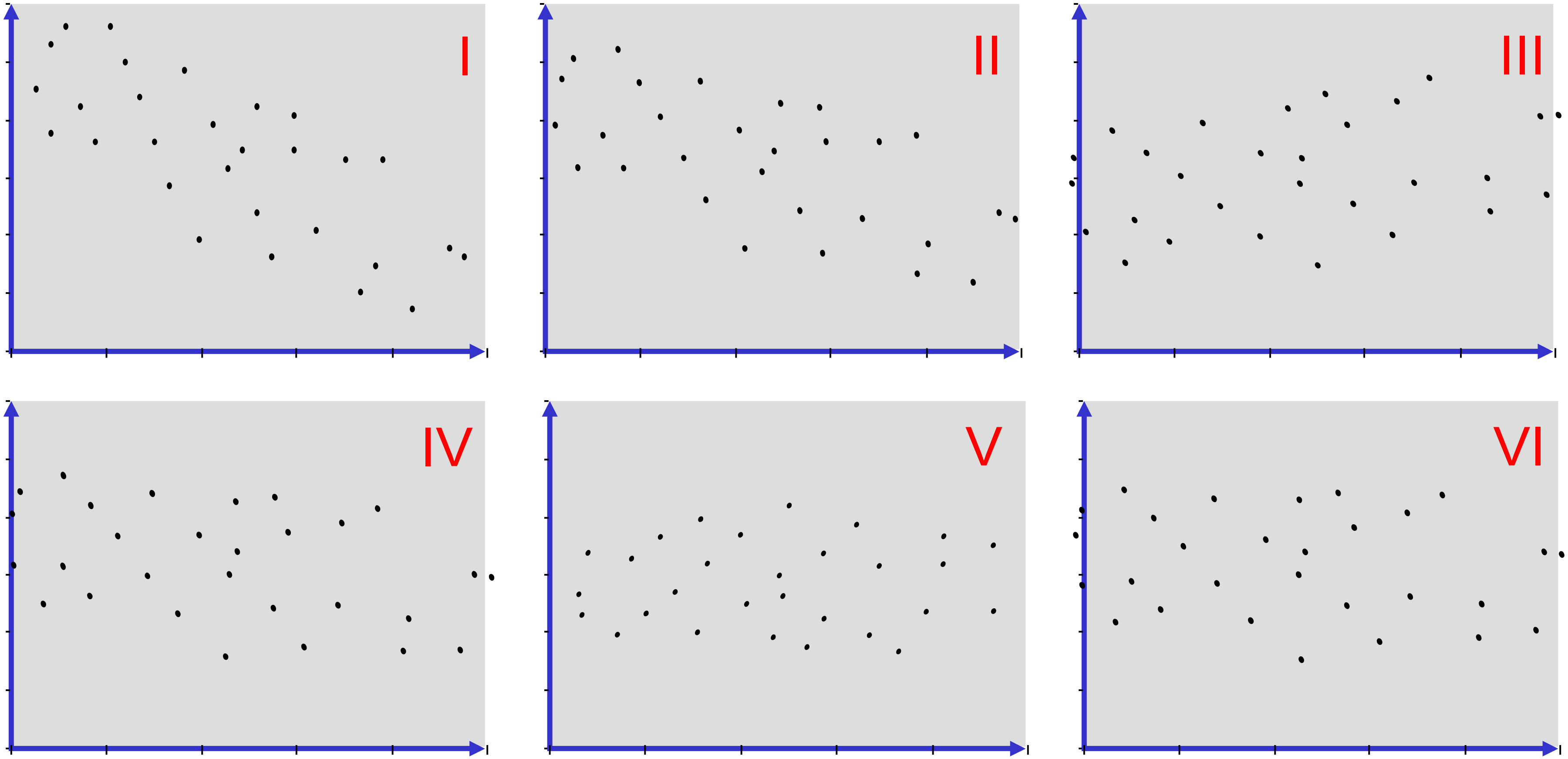
Two-gene pathways comprise two genes, and ignore the possibility there may be more.


Independence


Dependence

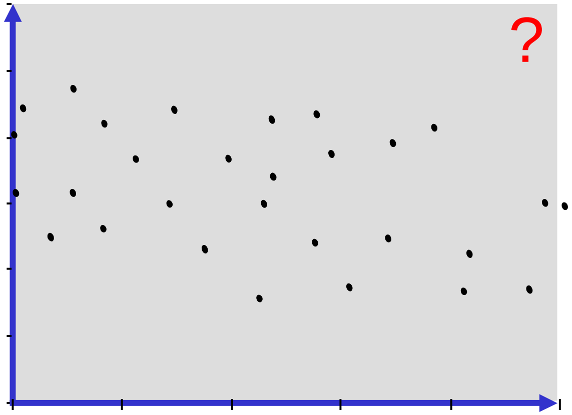
Two-gene pathway

Scatterplots of data on two random variables.
Which show association?



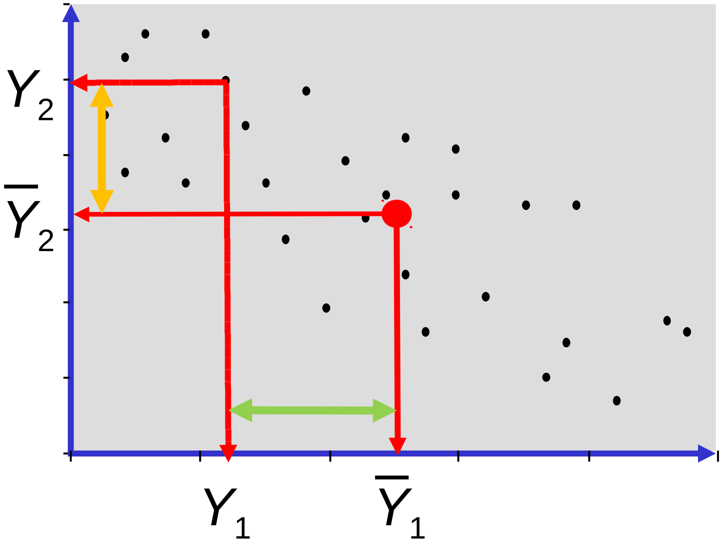
Two-gene pathway

→ Assess association between two random variables graphically.
→ Not very exact and in boundary cases no consensus.



Ideally, a measure of interrelatedness of the two variables.

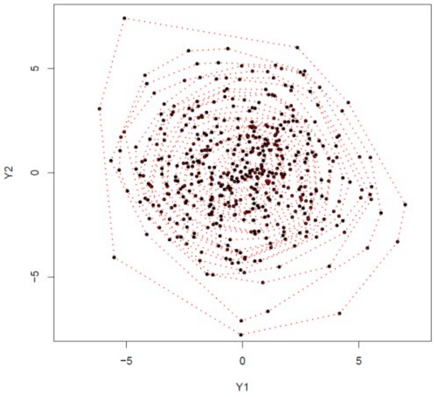
Correlation measures whether a change in one variable systematically coincides with a change in another variable.



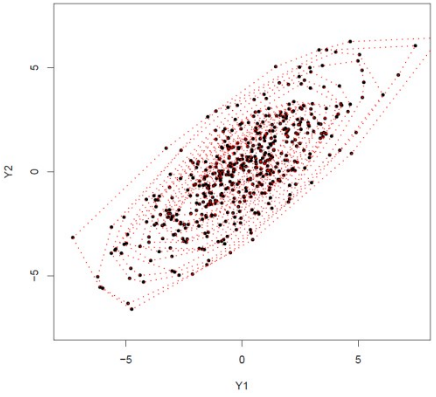
Two-gene pathway

Two-gene system

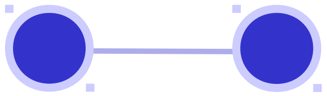
Calculate correlation between any two genes. If the correlation is large (in some sense), the two genes interact.



r = 0.027



r = 0.693



Interpretation pitfall

Eat chocolate, win the Nobel!

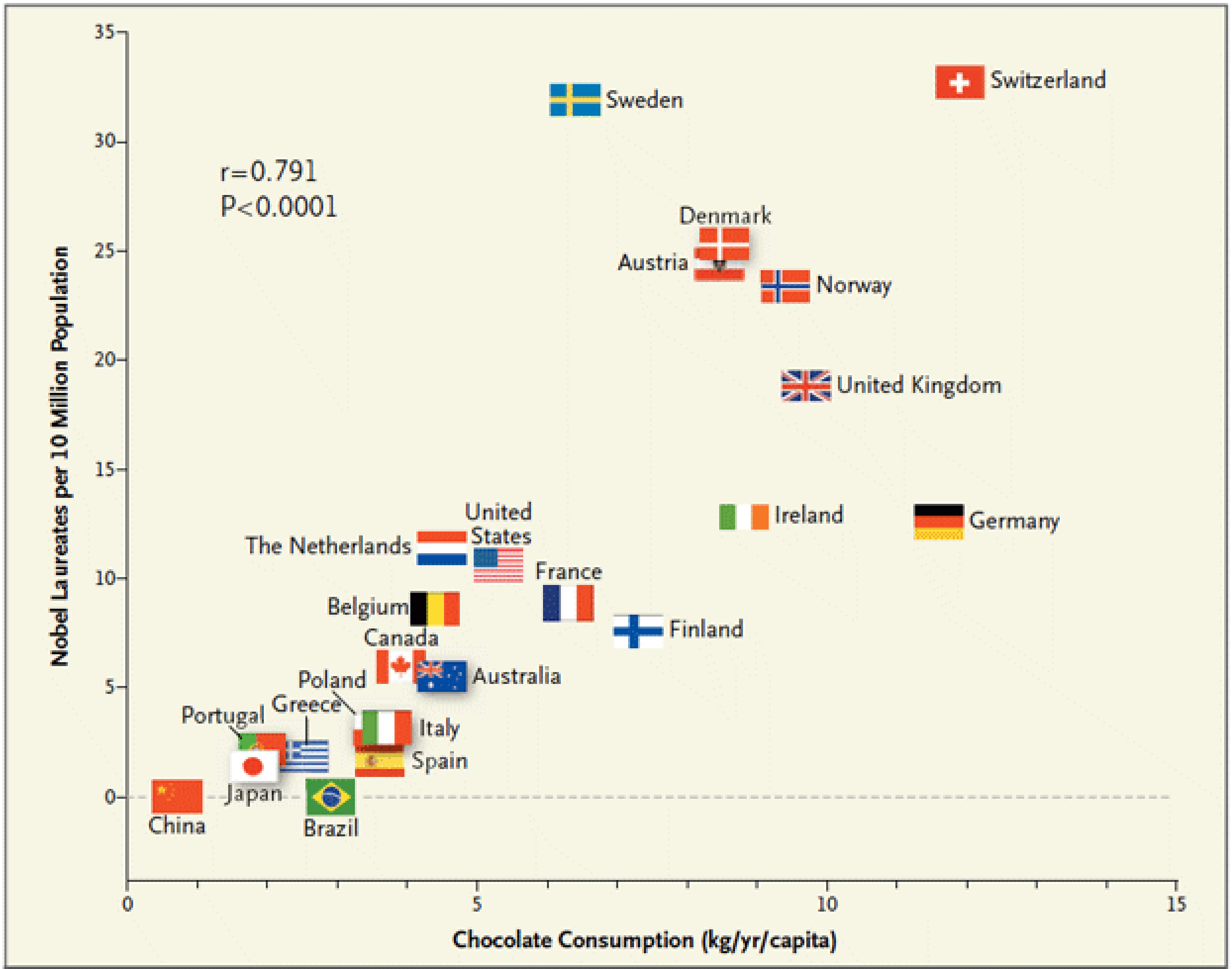


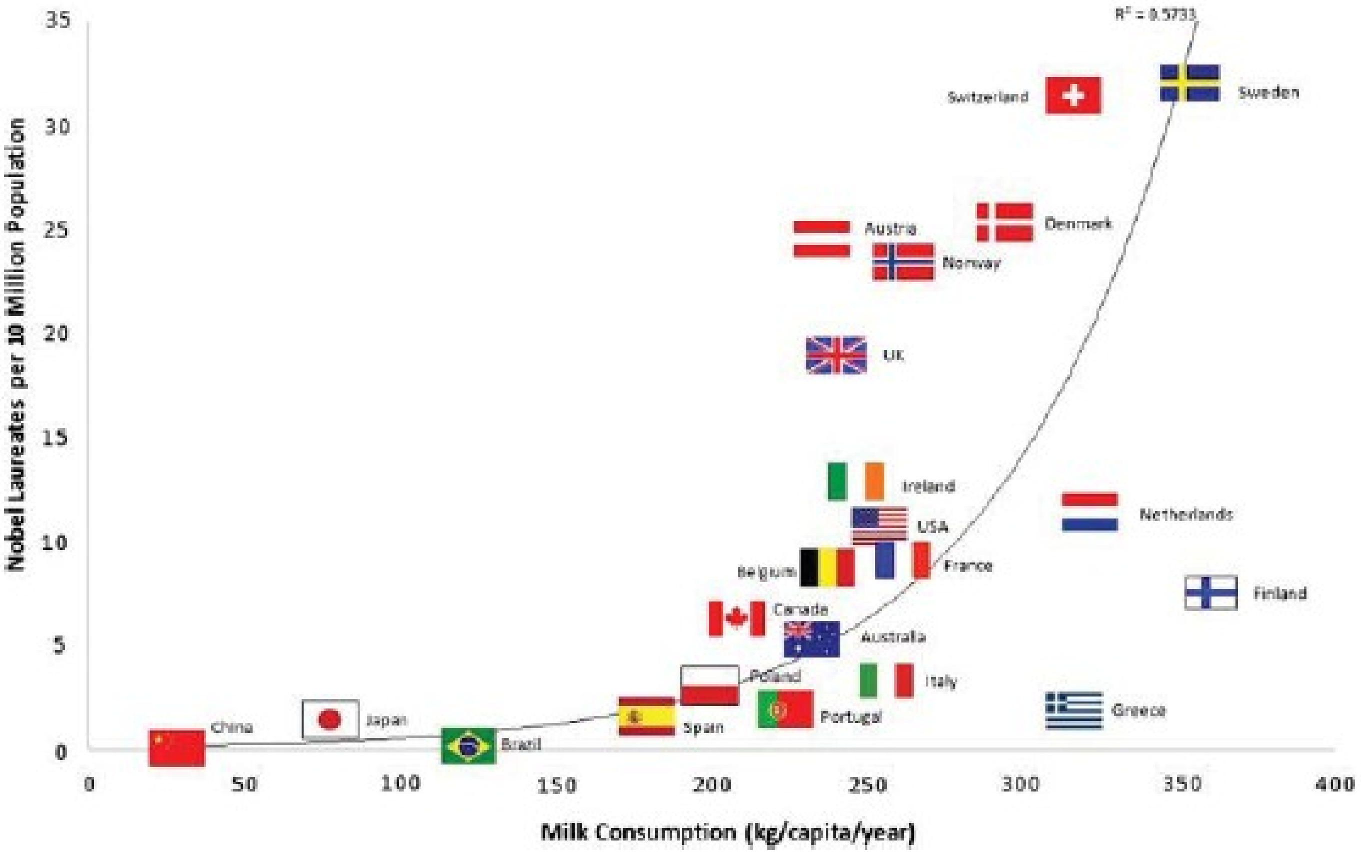
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

Messerli, 2012.

Interpretation pitfall

Even better: drink milk, win the Nobel!

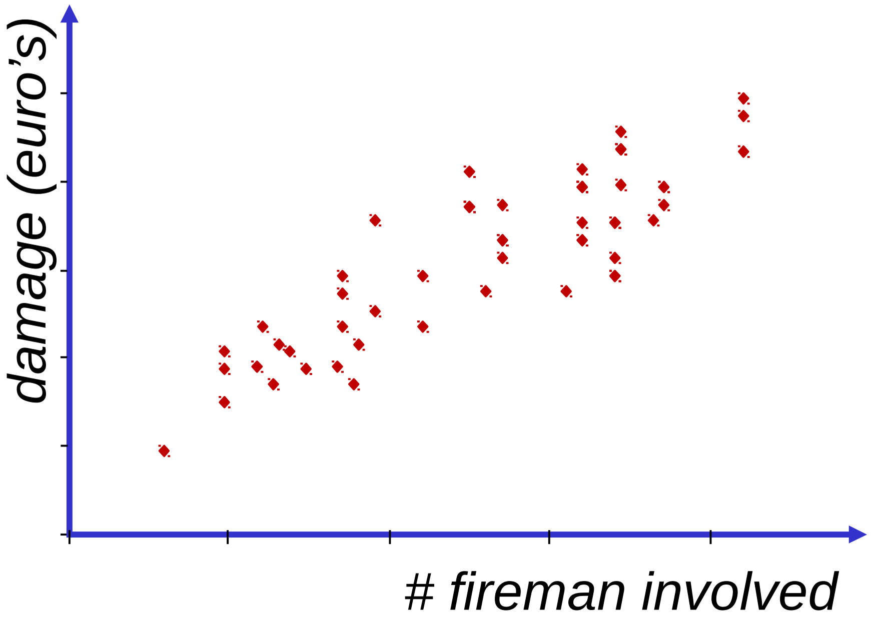


Best: drink chocolate-milk, win the Nobel?

Linthwaite, Fuller, 2013.

Interpretation pitfall

Does the involvement of more fireman result in more damage?



Possible interpretations of these data:

$$\begin{tikzcd} X \arrow[r] & Y \end{tikzcd}$$

More firemen result in more damage.

$$\begin{tikzcd} X & Y \arrow[l] \end{tikzcd}$$

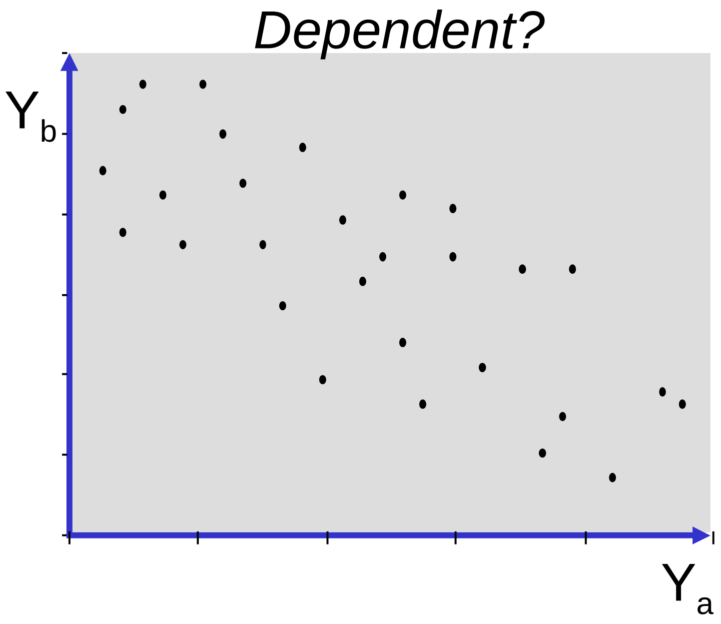
More damage results in more firemen.

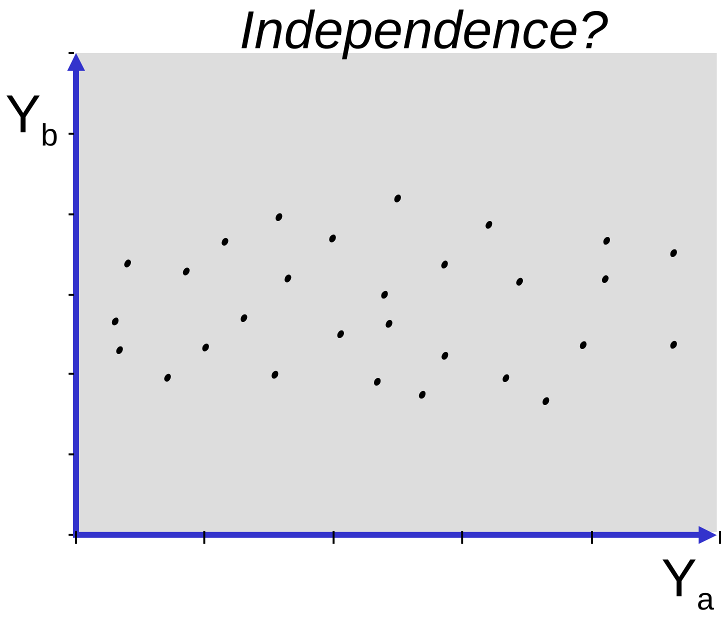
$$\begin{tikzcd} X & & Y \\ & Z \arrow[ul] \arrow[ur] & \end{tikzcd}$$

A bigger fire (Z) results in more firemen and more damage.

Interpretation pitfall

What to conclude about the relation between the expression levels of gene A and B?





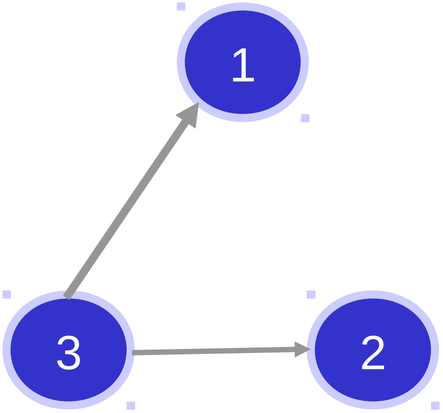
Question

Could other genes be responsible for observed (in)dependence?

Multi-gene pathways

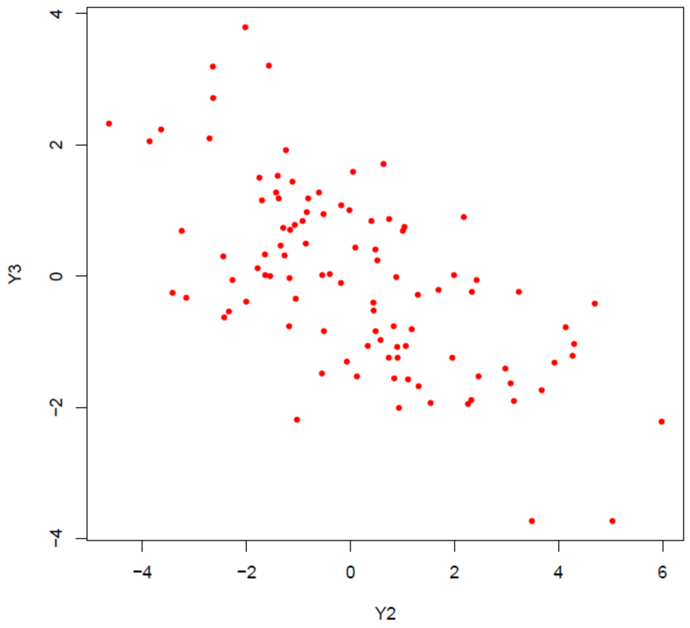
A possible model:

$$Y_1 = b_1 * Y_3 + \text{error}$$
$$Y_2 = b_2 * Y_3 + \text{error}$$



Correlation between nodes 1 and 2 may be nonzero!

Simulation shows this: →

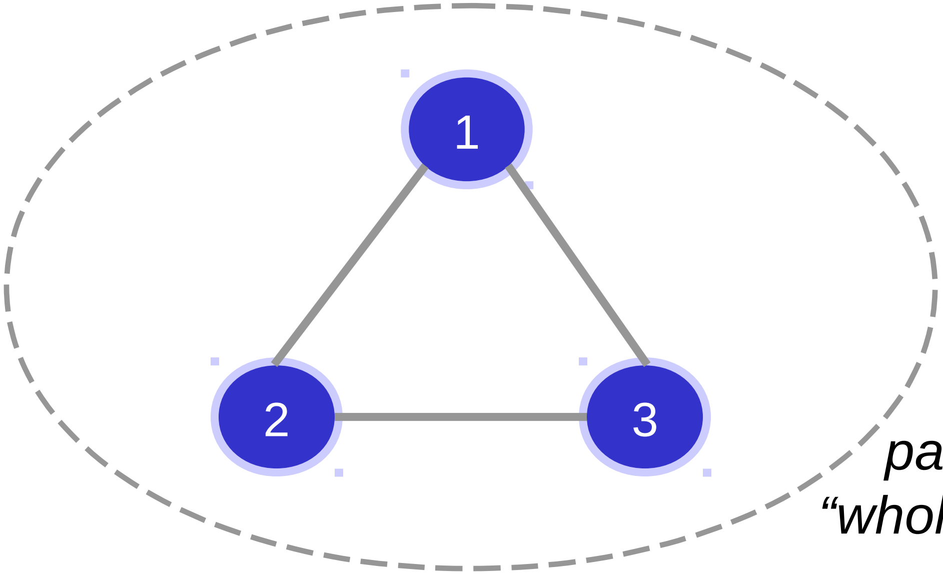


Even though there is no direct (causal) relationship between node 1 and 2 they may be correlated.

Multi-gene pathways

Multi-gene pathways

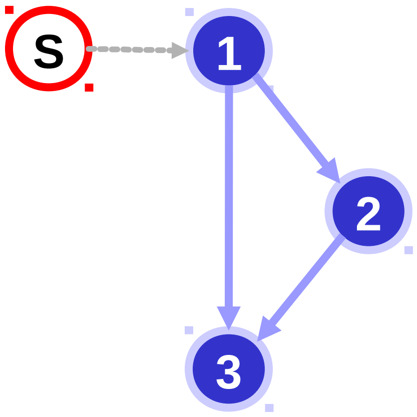
Multi-gene pathways comprise of more than two genes, and assume no gene “lives” outside the pathway.



pathway = “whole universe”

Correlate all possible gene pairs. This ignores the other genes, and only assesses (direct + indirect) association between a gene pair.

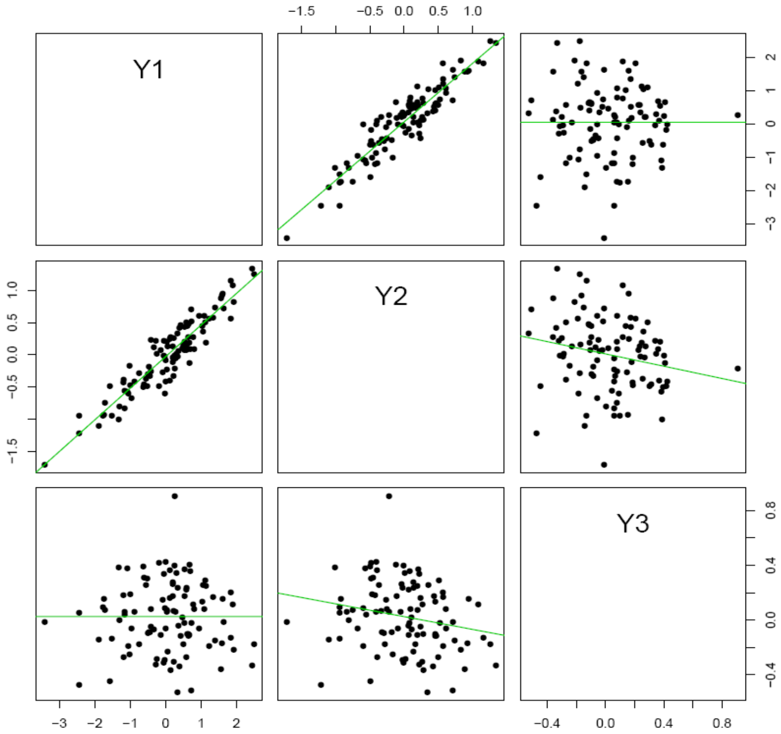
Multi-gene pathways



Example

Consider a pathway of 3 genes with underlying regulatory network:

Data



Correlation matrix:

```
      Y1     Y2     Y3
Y1 1.000  0.930  0.000
Y2 0.930  1.000 -0.211
Y3 0.000 -0.211  1.000
```

Multi-gene pathways

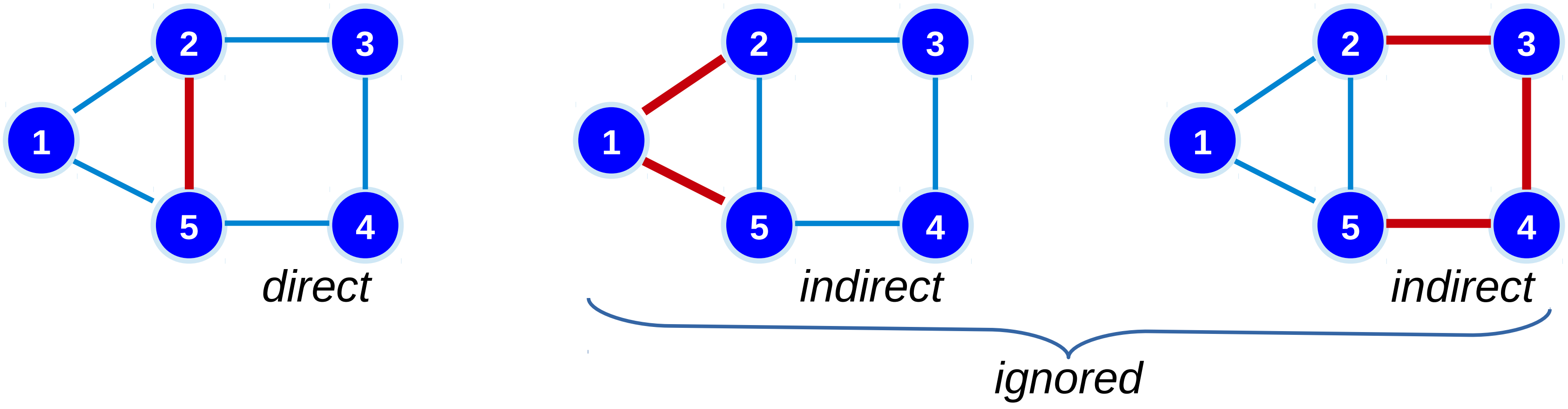
Partial correlation coefficient

Correlation between two variables when taking into account other variables.

The partial correlation measures the direct relation between node A and B while excluding effects of nodes in C.

Denoted: $\rho(Y_a, Y_b \mid \mathbf{Y}_c)$

Partial correlation between nodes 2 and 5:



direct

indirect

indirect

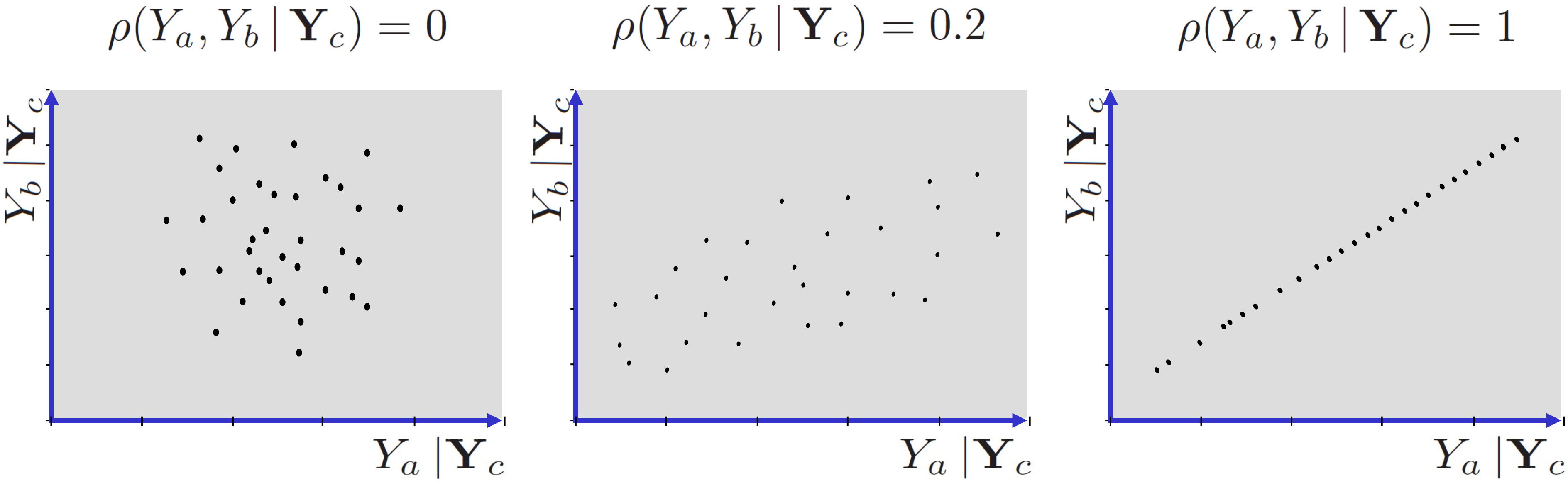
ignored

Multi-gene pathways

The partial correlation is a correction, thus:

$$\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) \in [-1, 1]$$

with:



$$\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) = 0$$

$$\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) = 0.2$$

$$\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) = 1$$

Multi-gene pathways

Interpretation

Let Y_1, Y_2, Y_3 be random variables. Then, $\rho(Y_1, Y_2 \,|\, Y_3) \approx$ amount of information in Y_1 on Y_2 after removal of all information on either of them contained in Y_3.

$$\rho(Y_1, Y_2 \,|\, Y_3) = 0$$

```
Call:
lm(formula = Y1 ~ 0 + Y2 + Y3)

Coefficients:
   Estimate Pr(>|t|)
Y2 -0.01444     0.638
Y3  1.01584    <2e-16 ***
```

Y_2 adds nothing to Y_3 in explaining variation in Y_1.

$$\rho(Y_1, Y_2 \,|\, Y_3) \neq 0$$

```
Call:
lm(formula = Y1 ~ 0 + Y2 + Y3)

Coefficients:
   Estimate Pr(>|t|)
Y2  0.24869 2.95e-15 ***
Y3  0.96542  < 2e-16 ***
```

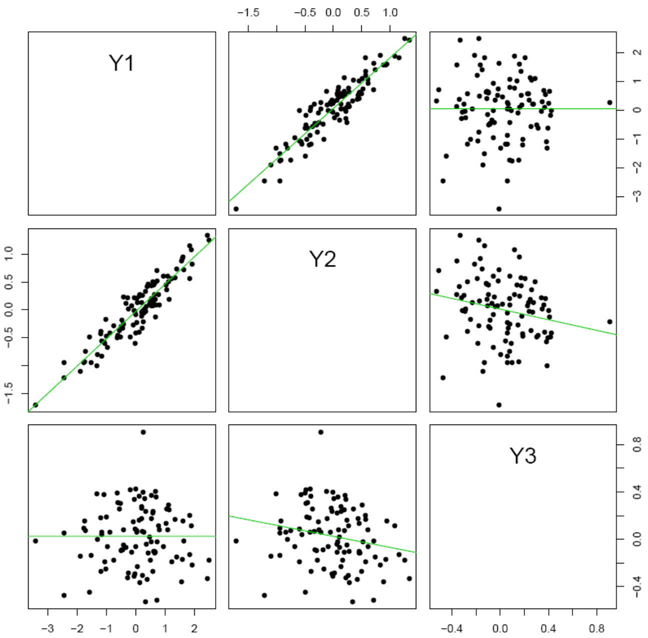
Y_2 does add to Y_3 in explaining variation in Y_1.

Multi-gene pathways

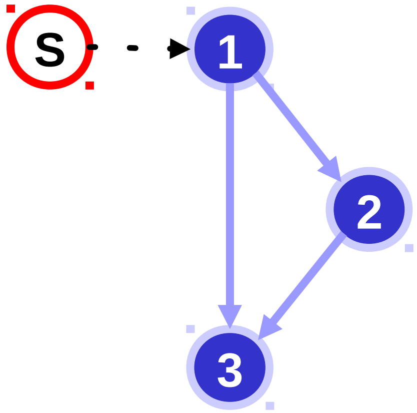
Example (continued)

Partial correlation matrix:

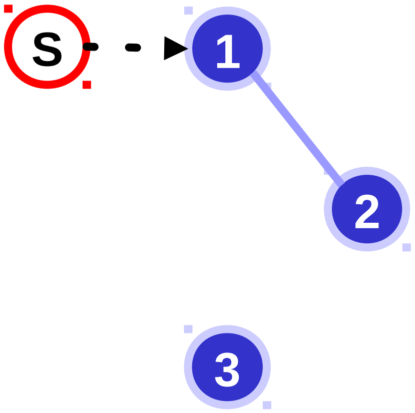
```
          Y1      Y2      Y3
Y1     1.000   0.952   0.549
Y2     0.952   1.000  -0.576
Y3     0.549  -0.576   1.000
```



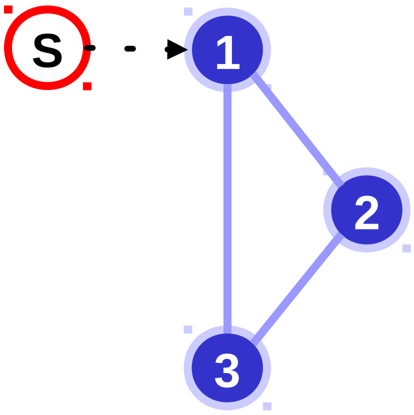
underlying regulatory network



reconstructed (correlation)



reconstructed (partial correlation)



Multi-gene pathways

rags2ridges

→ calculates partial correlations ...

→ … from high-dimensional data (in ridge fashion).

→ identifies the network from partial correlations.

→ allows for incorporation of network suggestion.

→ visualizes network.

→ exploits network (model) for down-stream purposes.

→ deals with group information: differential networks.

glasso

→ … from high-dimensional data (in lasso fashion).

ragt2ridges

→ sibling of **rags2ridges** for time course data.