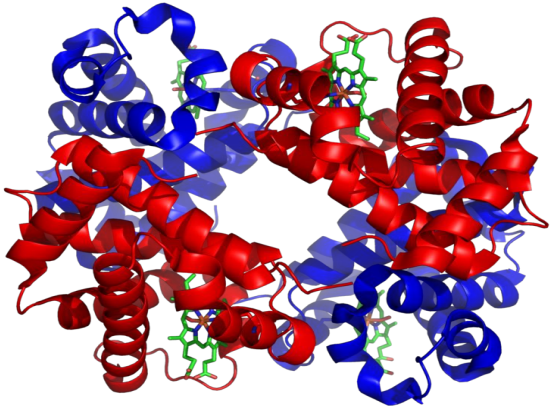
Probabilistic Network Inference

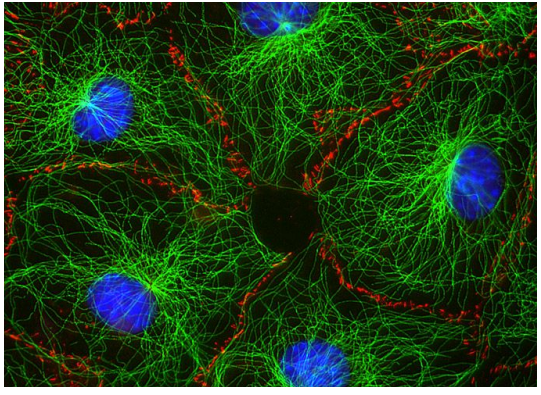
Guillermo de Anda Jáuregui
INMEGEN / CONAHCYT / C3-UNAM
@gdeandajauregui

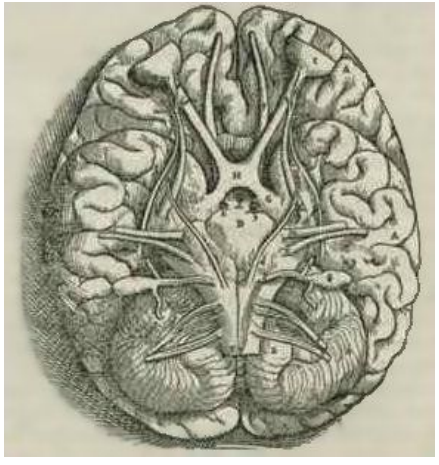
Today

- Why use networks to model correlated data
- How to generate networks from data
- Things we can learn from probabilistically inferred networks

Life is complex...







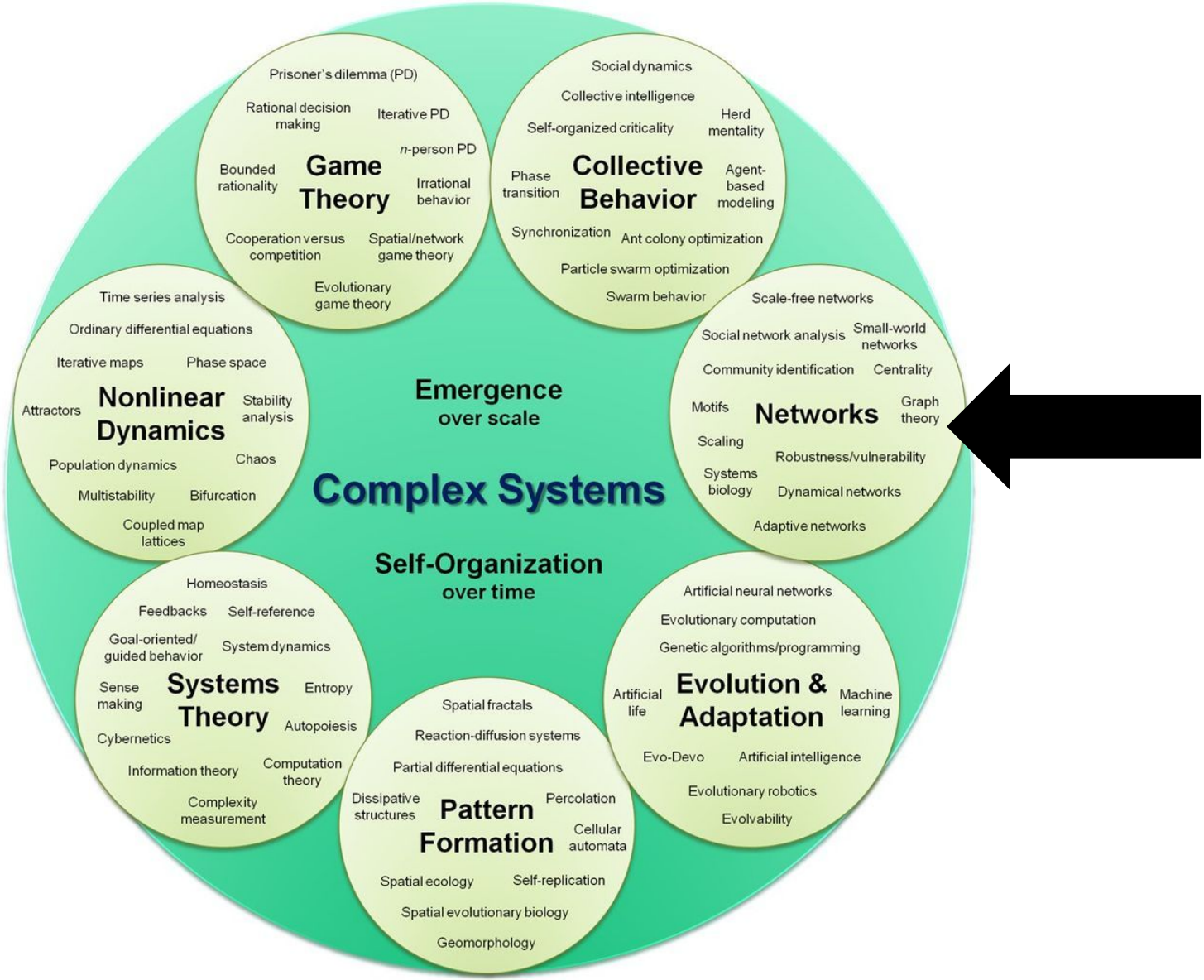
"There's no love in a carbon atom,

No hurricane in a water molecule,

No financial collapse in a dollar bill."

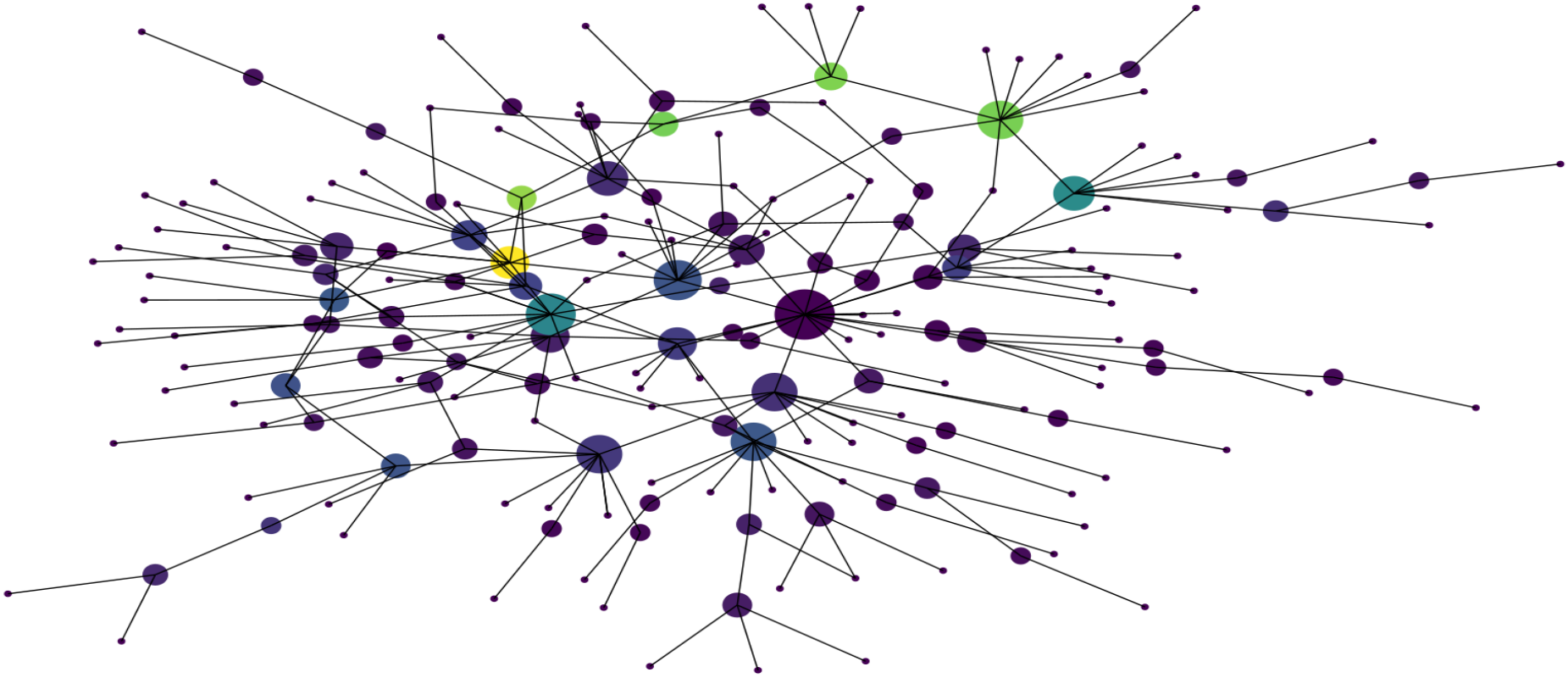
Peter Dodds


Res fuerat quondam præstans, & Gloria summa
Orbis subiectus Cæsaris Imperio,
Hic longe præstat, cuius nunc Orbis Eous,
Et Nouus, atq; alter panditur Auspitijs.
Punta de Cuba
Rio panuco
Quisſber punctus magnus continet leucas duode
eim cū dimidia, ita q duo magni puncti continent
viginti quinq leucas, cōtinet aute leucas quatuor
Italica miliaria, ita q omnes puncti qui hic cōſpi

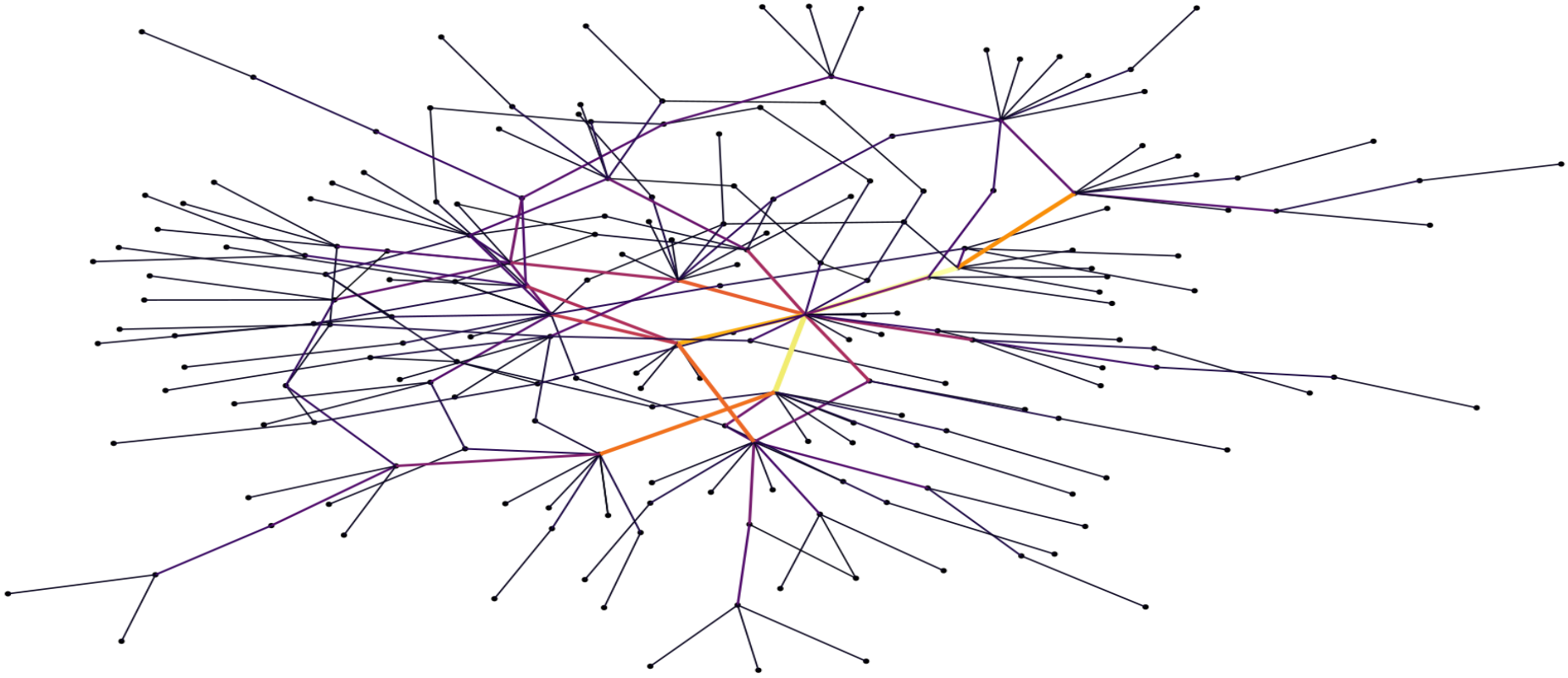

Complex Systems
Emergence
over scale
Self-Organization
over time
Game Theory
Prisoner's dilemma (PD)
Rational decision making
Iterative PD
n-person PD
Bounded rationality
Irrational behavior
Cooperation versus competition
Spatial/network game theory
Evolutionary game theory
Collective Behavior
Social dynamics
Collective intelligence
Self-organized criticality
Herd mentality
Phase transition
Agent-based modeling
Synchronization
Ant colony optimization
Particle swarm optimization
Swarm behavior
Networks
Scale-free networks
Social network analysis
Small-world networks
Community identification
Centrality
Motifs
Graph theory
Scaling
Robustness/vulnerability
Systems biology
Dynamical networks
Adaptive networks
Nonlinear Dynamics
Time series analysis
Ordinary differential equations
Iterative maps
Phase space
Attractors
Stability analysis
Population dynamics
Chaos
Multistability
Bifurcation
Coupled map lattices
Systems Theory
Homeostasis
Feedbacks
Self-reference
Goal-oriented/ guided behavior
System dynamics
Sense making
Entropy
Autopoiesis
Cybernetics
Information theory
Computation theory
Complexity measurement
Pattern Formation
Spatial fractals
Reaction-diffusion systems
Partial differential equations
Dissipative structures
Percolation
Cellular automata
Spatial ecology
Self-replication
Spatial evolutionary biology
Geomorphology
Evolution & Adaptation
Artificial neural networks
Evolutionary computation
Genetic algorithms/programming
Artificial life
Machine learning
Evo-Devo
Artificial intelligence
Evolutionary robotics
Evolvability

Networks are cool

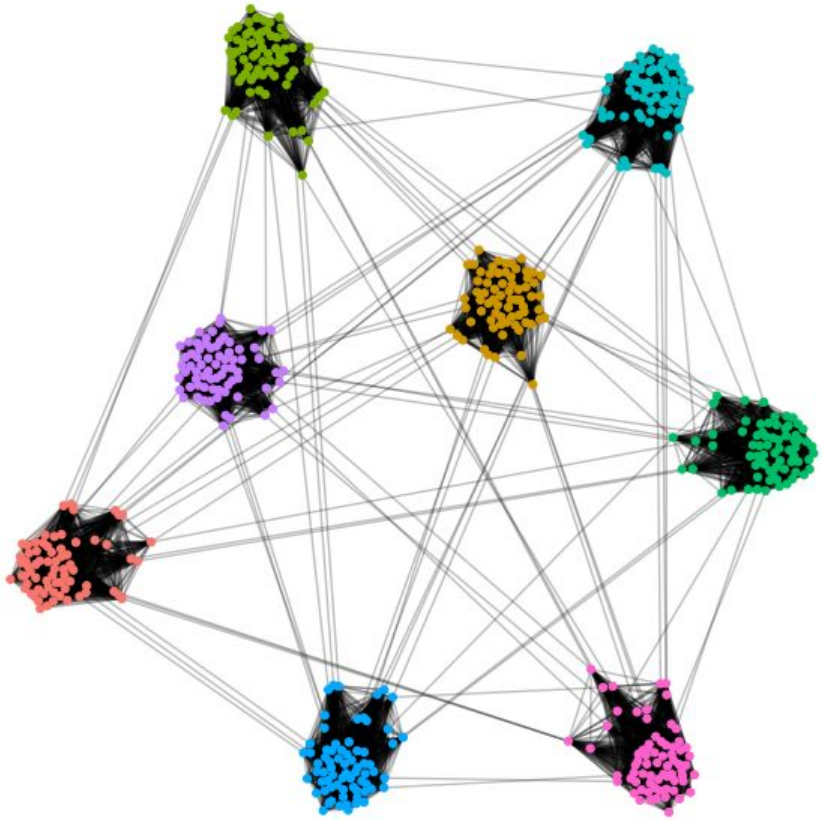
We may learn about the elements of a system...



...and how they are connected



Plus, about the whole system, itself

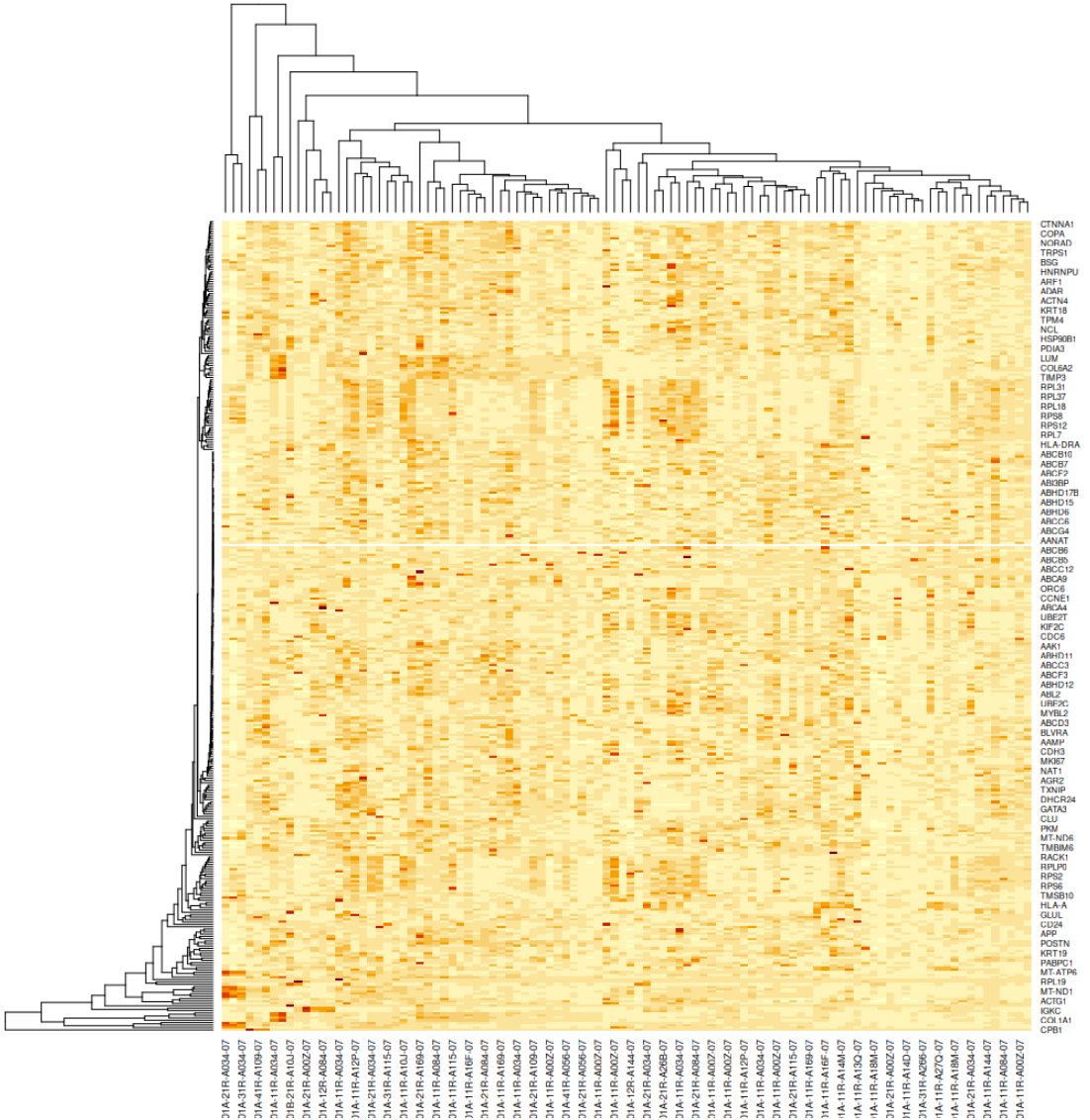


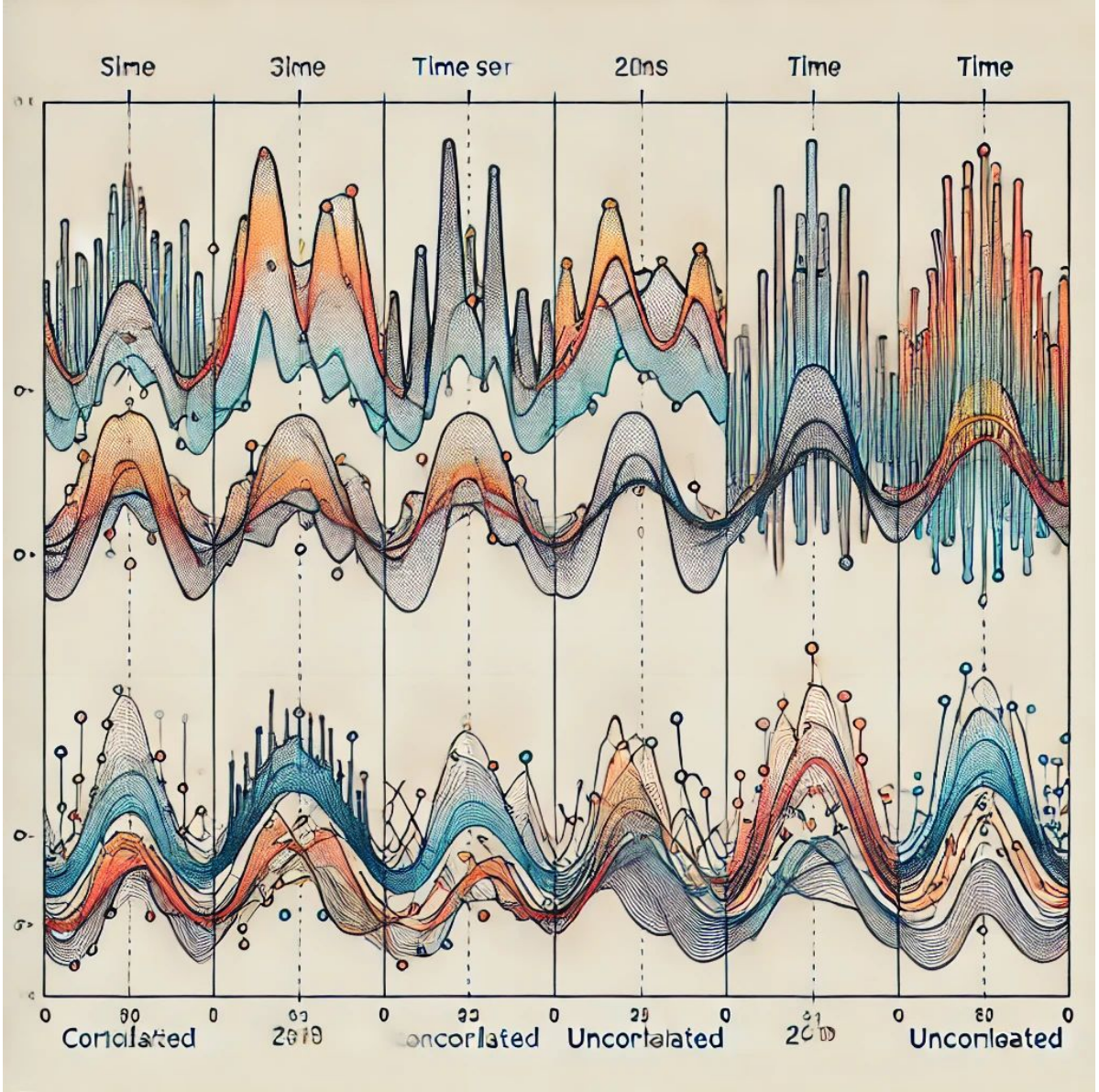
Networks in my data?


B
Ciudad Azteca
Plaza Aragon
Olimpica
Ecatepec
M. Muzquiz
Rio de los Remedios
Impulsora
Nezahualcoyotl
Villa de Aragon
Bosque de Aragon
Deportivo Oceania
Oceania
Terminal Aerea
Hangares
1 5 9 A
Pantitlan
Agricola Oriental
Canal de San Juan
Tepalcates
Guelatao
Peñon Viejo
Acatitla
Santa Marta
Los Reyes
La Paz
A
8
Constitucion de 1917
C de la UAM Estrella
Iztapalapa
Escuadron 201
Aculco
Apatlaco
Iztacalco
Coyuya
4
Santa Anita
La Viga
Viaducta
Chabacano
Mixiuhca
Velodromo
Ciudad Deportiva
Puebla
Zaragoza
Gomez Farias
Blvd. Pto. Aereo
Balbuena
Moctezuma
R. Flores Magon
R. Rubio
San Lazaro
Candelaria
Merced
Pino Suarez
Fray Servando
San Antonio Abad
Zocalo
Isabel la Catolica
San Juan de Letran
Allende
Bellas Artes
Morelos
Tepito
Lagunilla
Garibaldi
Guerrero
Buenavista
B
Canal del Norte
Eduardo Molina
Aragon
Consulado
Talisman
Bondojito
Martin Carrera
4 6
Deportivo 18 de Marzo
Indios Verdes
3
Instituto del Petroleo
Politecnico
5
La Villa - Basilica
Potrero
La Raza
Misterios
Valle Gomez
Autobuses del Norte
Lindavista
Vallejo
Norte 45
Azcapotzalco
Ferreria
Tezozomoc
El Rosario
6 7
Tlatelolco
Juarez
Hidalgo
Revolucion
Popotla
Cuitlahuac
Panteones
Cuatro Caminos
2
Aquiles Serdan
Camarones
San Joaquin
Polanco
Auditorio
Constituyentes
Tacubaya
9
Observatorio
1
Colegio Militar
Normal
San Cosme
Balderas
Cuauhtemoc
Insurgentes
Salto del Agua
Centro Medico
Chilpancingo
Juanacatlan
Chapultepec
Sevilla
Patriotismo
Niños Heroes
Hospital General
Doctores
Obrera
Lazaro Cardenas
Etiopia
Eugenia
Division del Norte
Parque de los Venados
Xola
Villa de Cortes
Nativitas
Portales
Ermita
Eje Central
Viaducta
Zapata
Coyoacan
Viveros
Copilco
Universidad
3
M. A. De Quevedo
Hospital 20 de Noviembre
Insurgentes Sur
San Antonio
San Pedro de los Pinos
Barranca del Muerto
7
Mixcoac
12
Mexicaltzingo
General Anaya
Tasqueña
2
Atlalilco
Culhuacan
San Andres Tomatlan
Lomas Estrella
Calle 11
Periferico Oriente
Tezonco
Olivos
Nopalera
Zapotitlan
Tlaltenco
Tlahuac
12


facebook

But...


CTNNA1
COPA
NCBAD
TRPS1
BSG
HNRNPU
ARF1
ADAR
ACTN4
KRT18
TPM4
NCL
HSP90B1
PDIA3
LUM
COL6A2
TIMP3
RPL31
RPL37
RPL18
RPS8
RPS12
RPL7
HLA-DRA
ABCB10
ABCE7
ABCF2
AB03BP
ABHD17B
ABHD15
ABHD6
ARCIC6
ABCG4
AANAT
ABCB6
ABCB5
ABCC12
ABCA9
ORC6
CCNE1
ABCA4
UBE2T
KIF2C
CDC6
AAK1
ARHD11
ABCC3
ABCF3
ABHD12
A8L2
UBF2C
MYBL2
ABCD3
BLVRA
AMH
CDH3
MKI67
NAT1
AGR2
TXNIP
DHCR24
GATA3
CLU
PKM
MT-ND6
TMBIM6
RACK1
RPLP0
RPS2
RPS6
TMSB10
HLA-A
GLUL
CD24
APP
POSTN
KRT19
PABPC1
MT-ATP6
RPL19
MT-ND1
ACTG1
IGKC
COL1A1
CPB1


Time
Time
Time ser
20ns
Time
Time
Correlated
Uncorlated
Uncorlated
Uncorlated

Identify statistical dependencies…
And model as networks

What’s the best measure of statistical dependency?

Pearson correlation

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Spearman correlation

$$r_s = \rho\left[\,\mathrm{R}[X], \mathrm{R}[Y]\,\right] = \frac{\operatorname{cov}\left[\,\mathrm{R}[X], \mathrm{R}[Y]\,\right]}{\sigma_{\mathrm{R}[X]}\ \sigma_{\mathrm{R}[Y]}},$$

Mutual information

$$I(x_i^h, x_j^k) = \sum_{\Omega} \sum_{\Omega'} p(x_i^h, x_j^k) \log \frac{p(x_i^h, x_j^k)}{p(x_i^h)\, p(x_j^k)}$$

Markov Random Fields

Unknown real probability density functions?

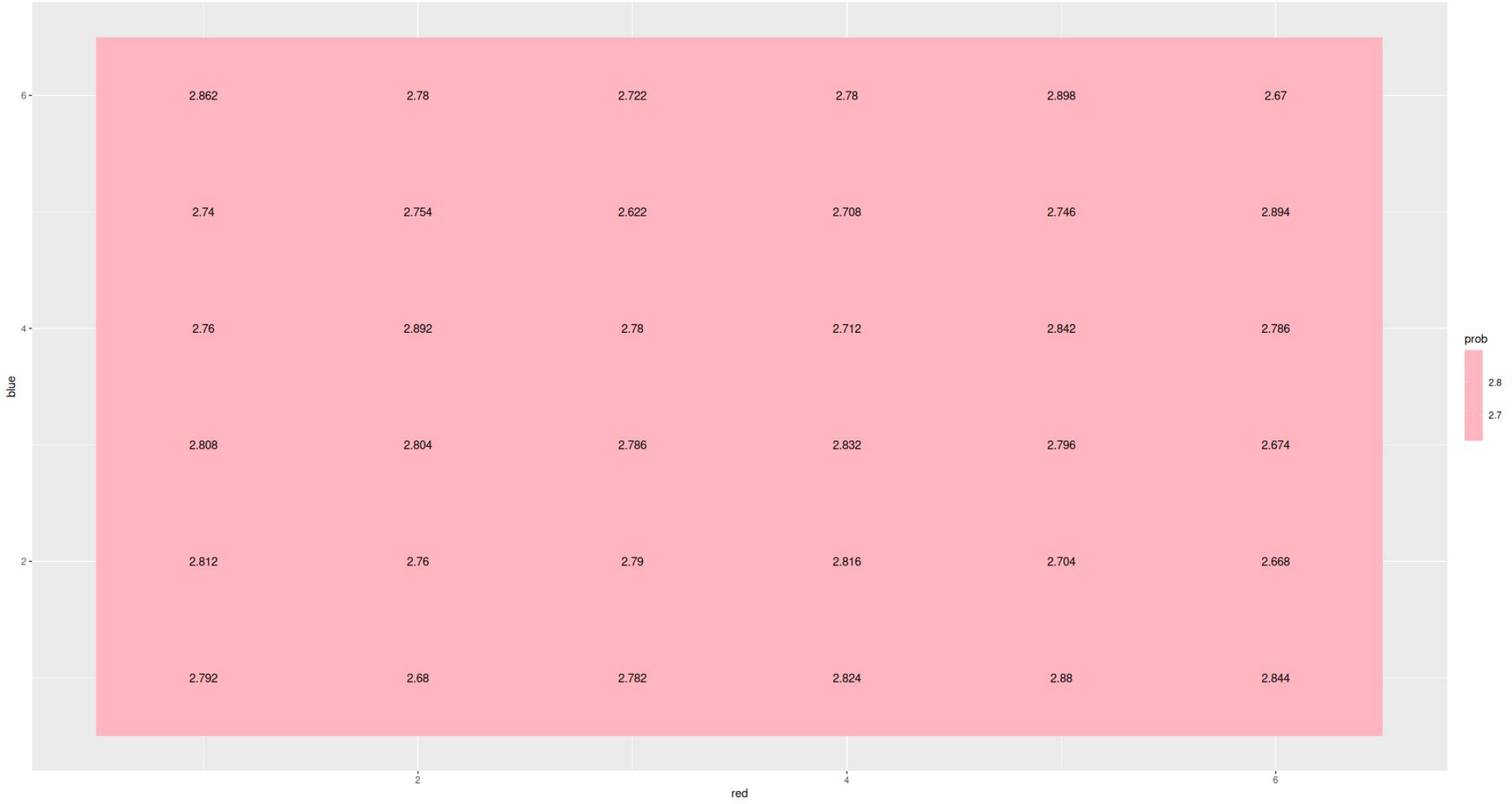
Use Empirical Probability Distributions (Big Data)

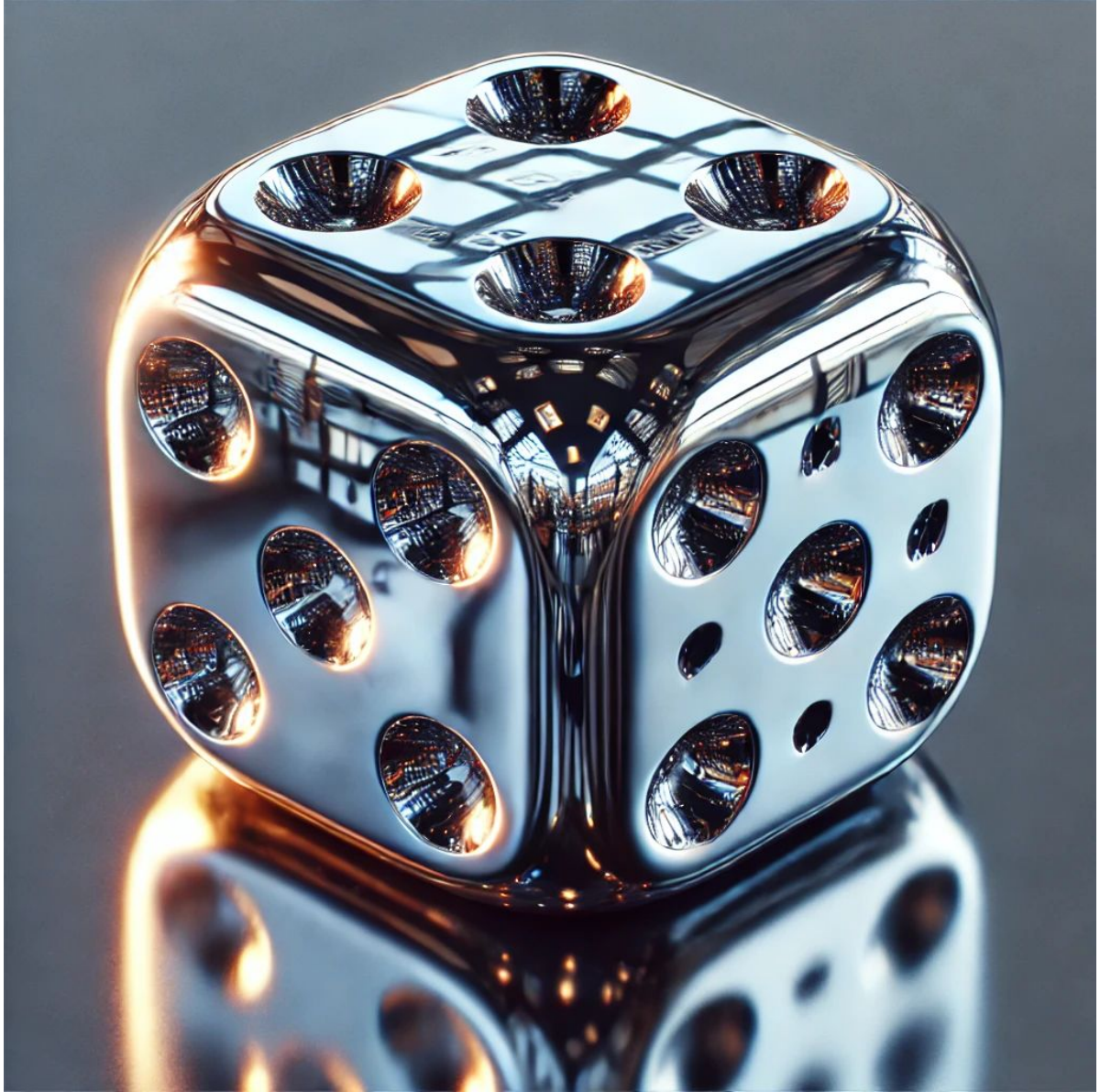
Glivenko-Cantelli theorem (uniform convergence)

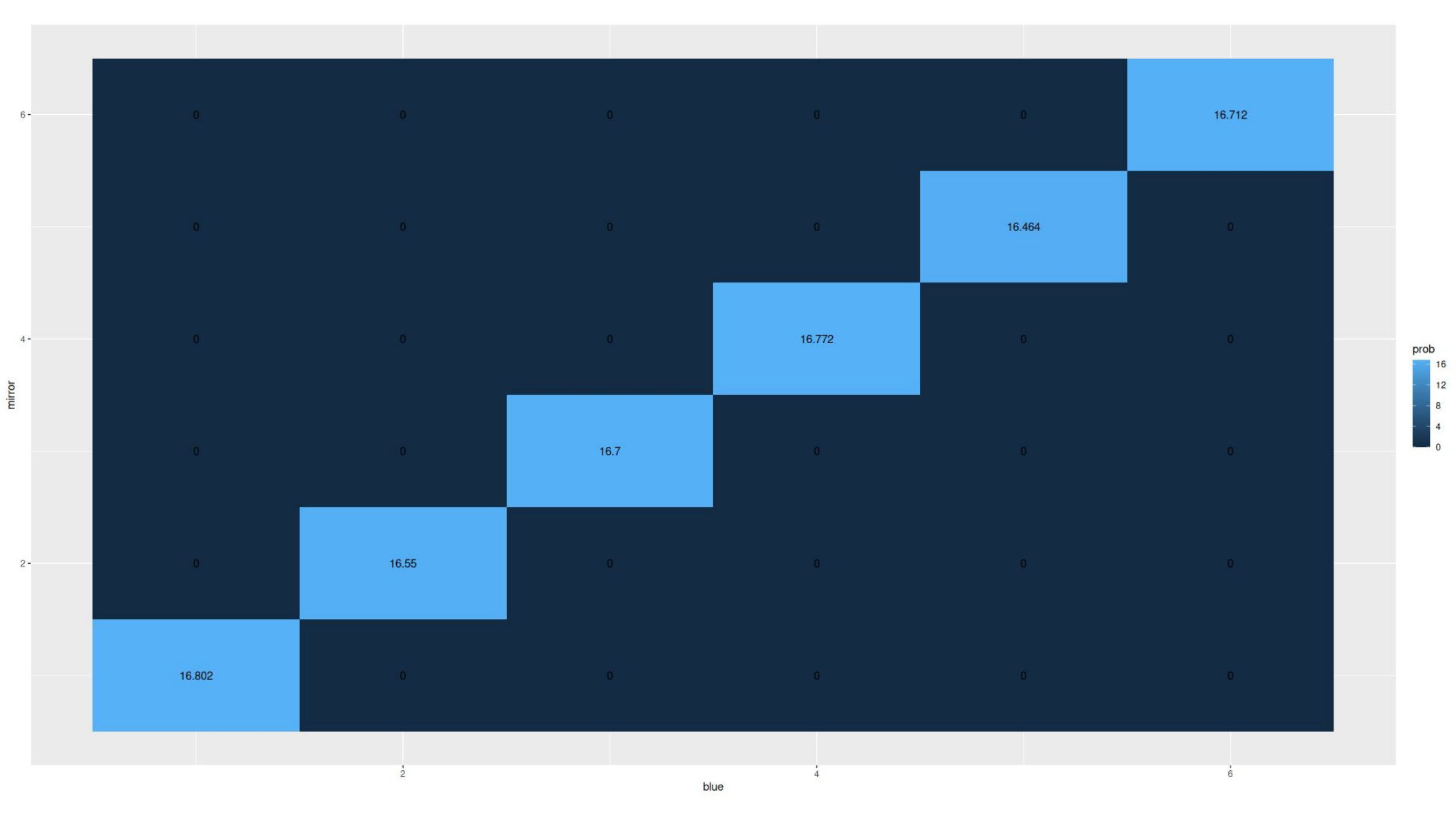
$$||F_n - F||_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0 \qquad a.s.$$

Glivenko 1933, Cantelli 1933

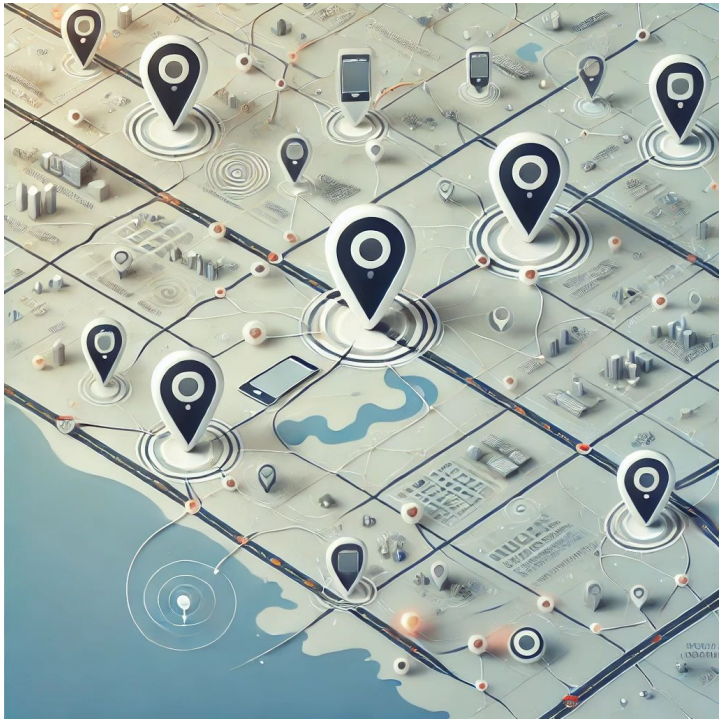
```r
# excercise 1: random variables (dice) ----

library(tidyverse)
library(infotheo)

# make two fair dice (red and blue) ----
# Each die has outcomes from 1 to 6

set.seed(725)
{red_dice <- sample(1:6, size = 50000, replace = TRUE)
blue_dice <- sample(1:6, size = 50000, replace = TRUE)
}

dice_data <- data.frame(red = red_dice, blue = blue_dice) |> as_tibble()
# plot probability density function (uniform) ----
```


2.862
2.78
2.722
2.78
2.898
2.67
2.74
2.754
2.622
2.708
2.746
2.894
2.76
2.892
2.78
2.712
2.842
2.786
2.808
2.804
2.786
2.832
2.796
2.674
2.812
2.76
2.79
2.816
2.704
2.668
2.792
2.68
2.782
2.824
2.88
2.844
6
4
2
blue
2
4
6
red
prob
2.8
2.7

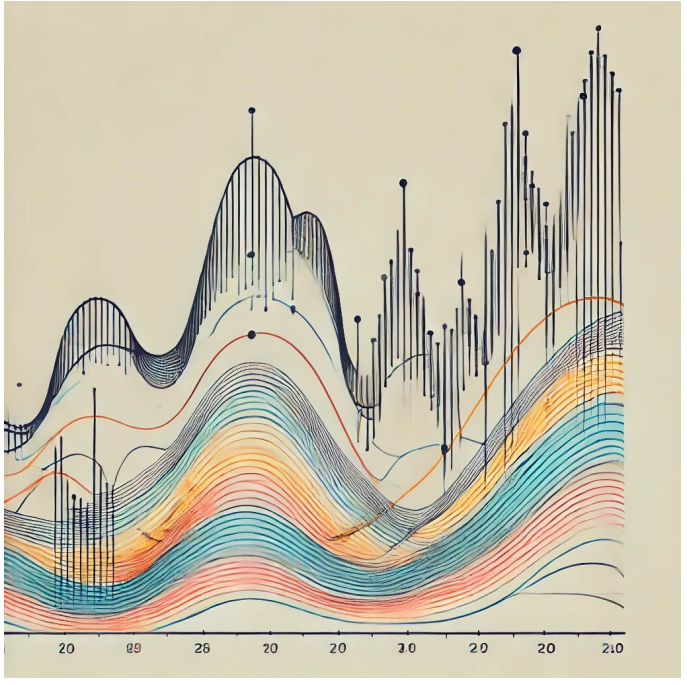



0
0
0
0
0
16.712
0
0
0
0
16.464
0
0
0
0
16.772
0
0
0
0
16.7
0
0
0
16.55
0
0
0
0
16.802
0
0
0
0
6
4
2
mirror
2
4
6
blue
prob
16
12
8
4
0

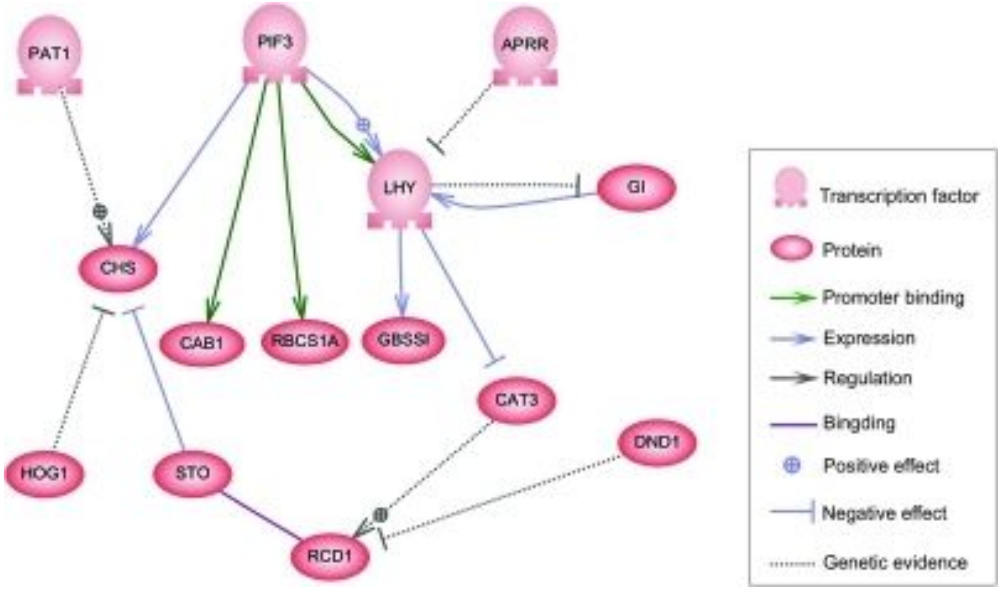
Some use cases



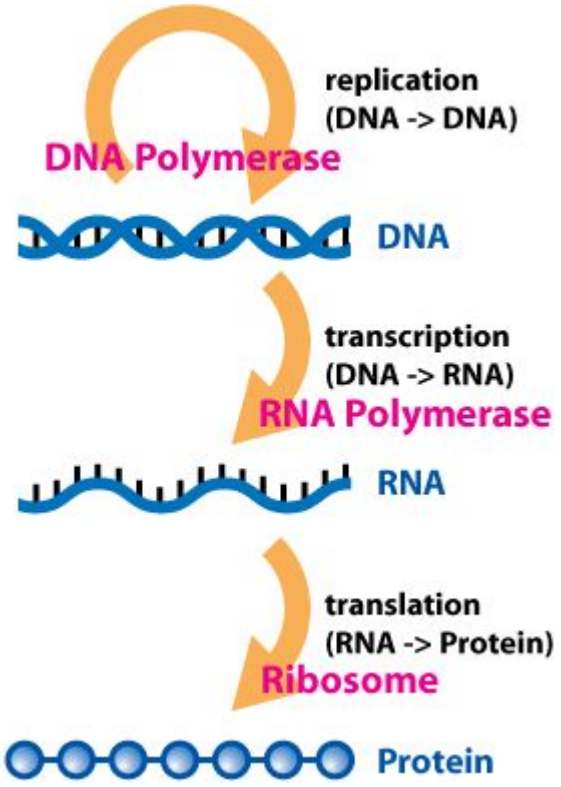
Some use cases

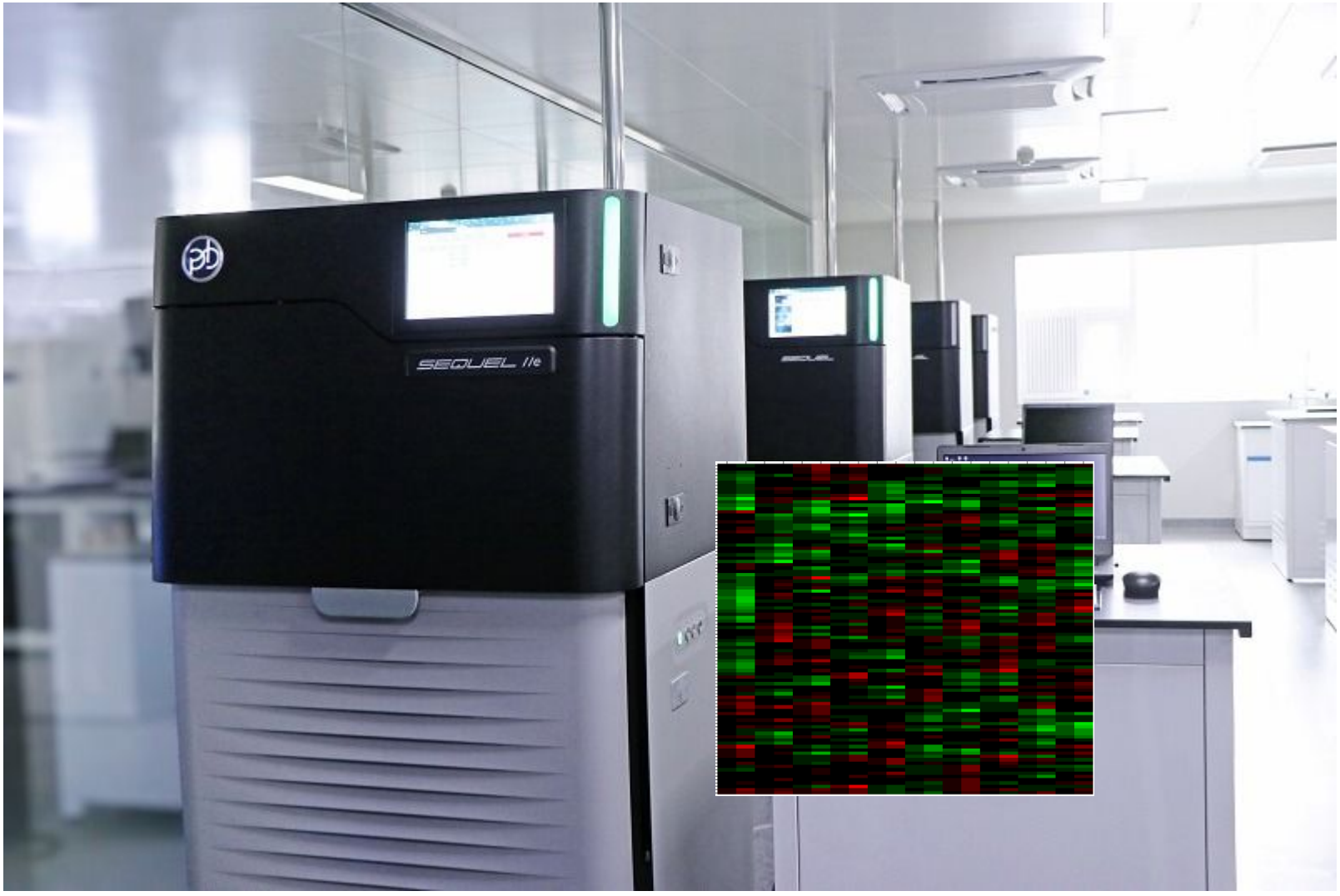


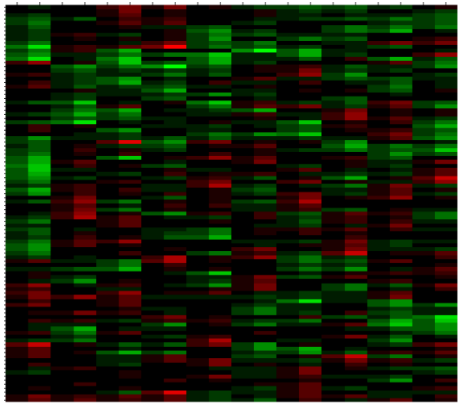
Some use cases



My (our) use case: genomics




SEQUEL IIe



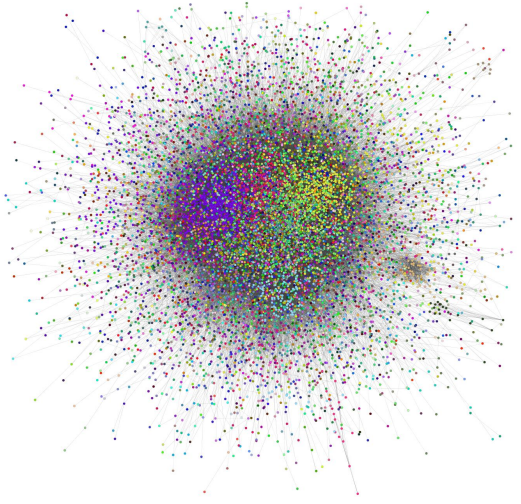
similarity

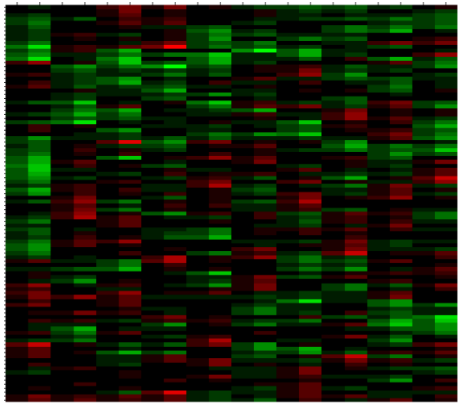
filtering

	g_1	g_2	...	g_n
g_1	1	0.7	...	0.9
g_2	0.7	1	...	0.2
...	...	...	1	...
g_n	0.9	0.2	...	1

	g_1	g_2	...	g_n
g_1	0	1	...	1
g_2	1	0	...	0
...	...	...	0	...
g_n	1	0	...	0

Network





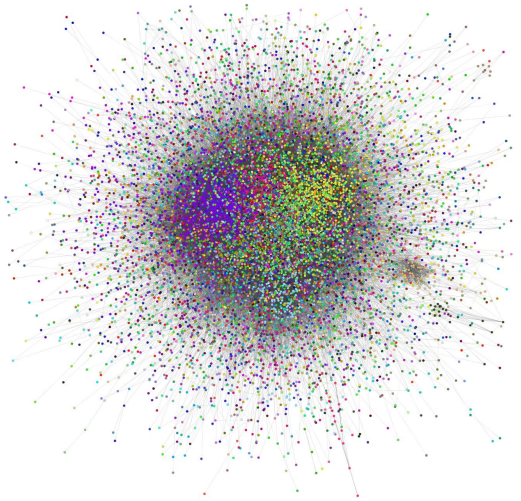
similarity

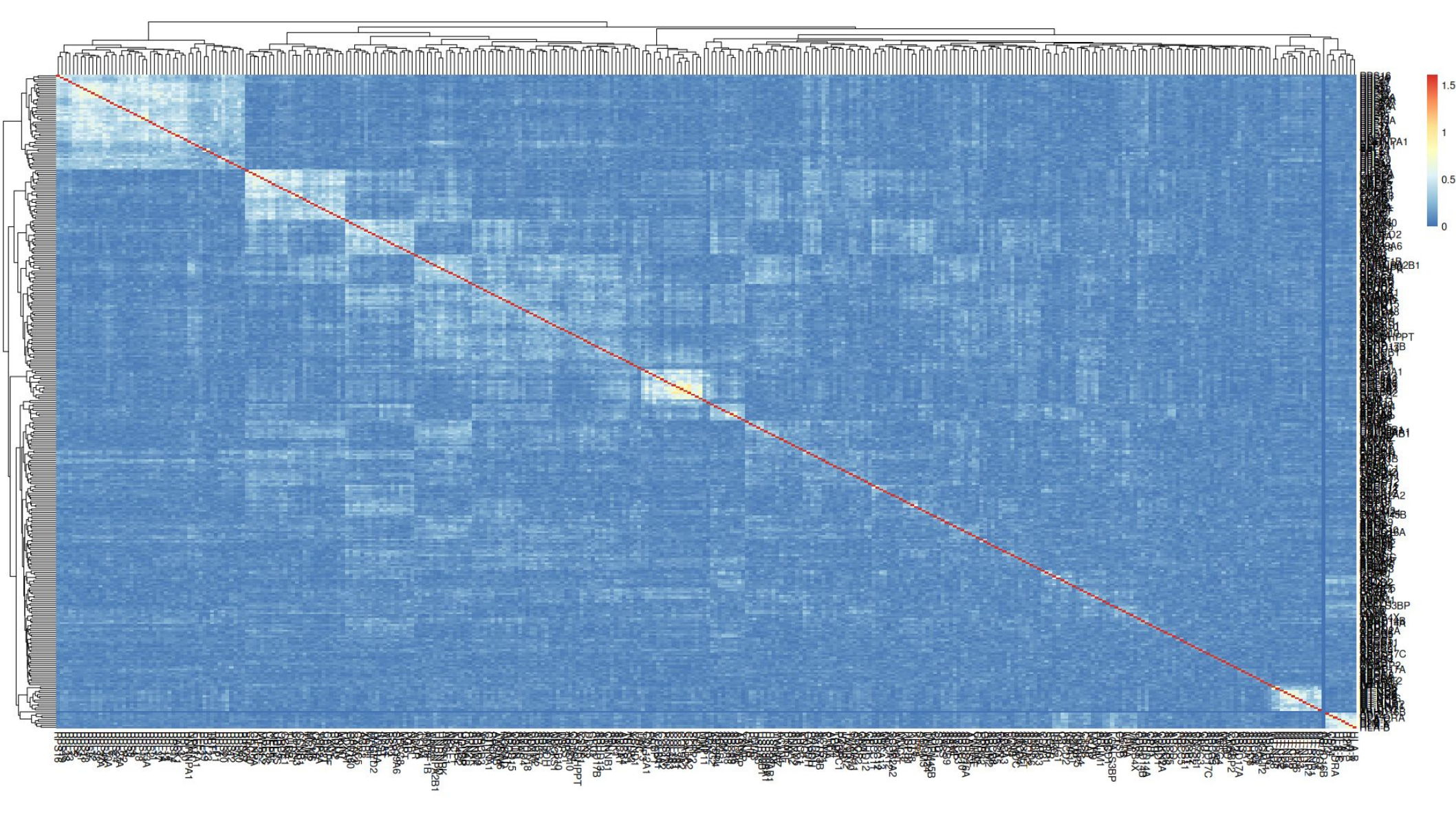
	g_1	g_2	...	g_n
g_1	1	0.7	...	0.9
g_2	0.7	1	...	0.2
...	...	...	1	...
g_n	0.9	0.2	...	1

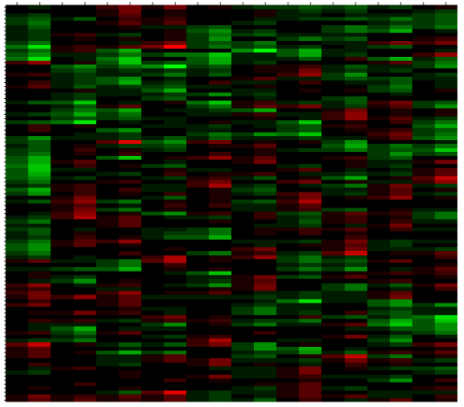
filtering

	g_1	g_2	...	g_n
g_1	0	1	...	1
g_2	1	0	...	0
...	...	...	0	...
g_n	1	0	...	0

Network




1.5
1
0.5
0



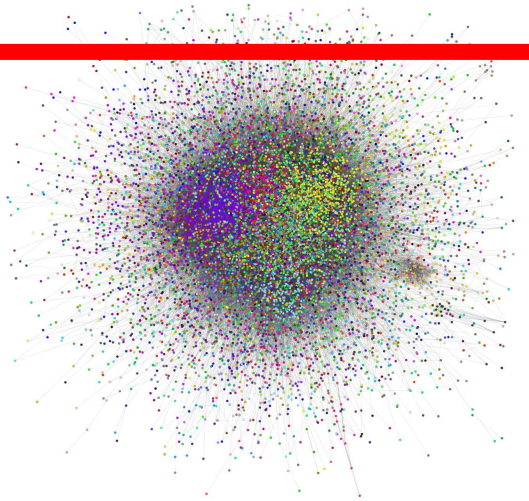
similarity

	g_1	g_2	...	g_n
g_1	1	0.7	...	0.9
g_2	0.7	1	...	0.2
...	...	...	1	...
g_n	0.9	0.2	...	1

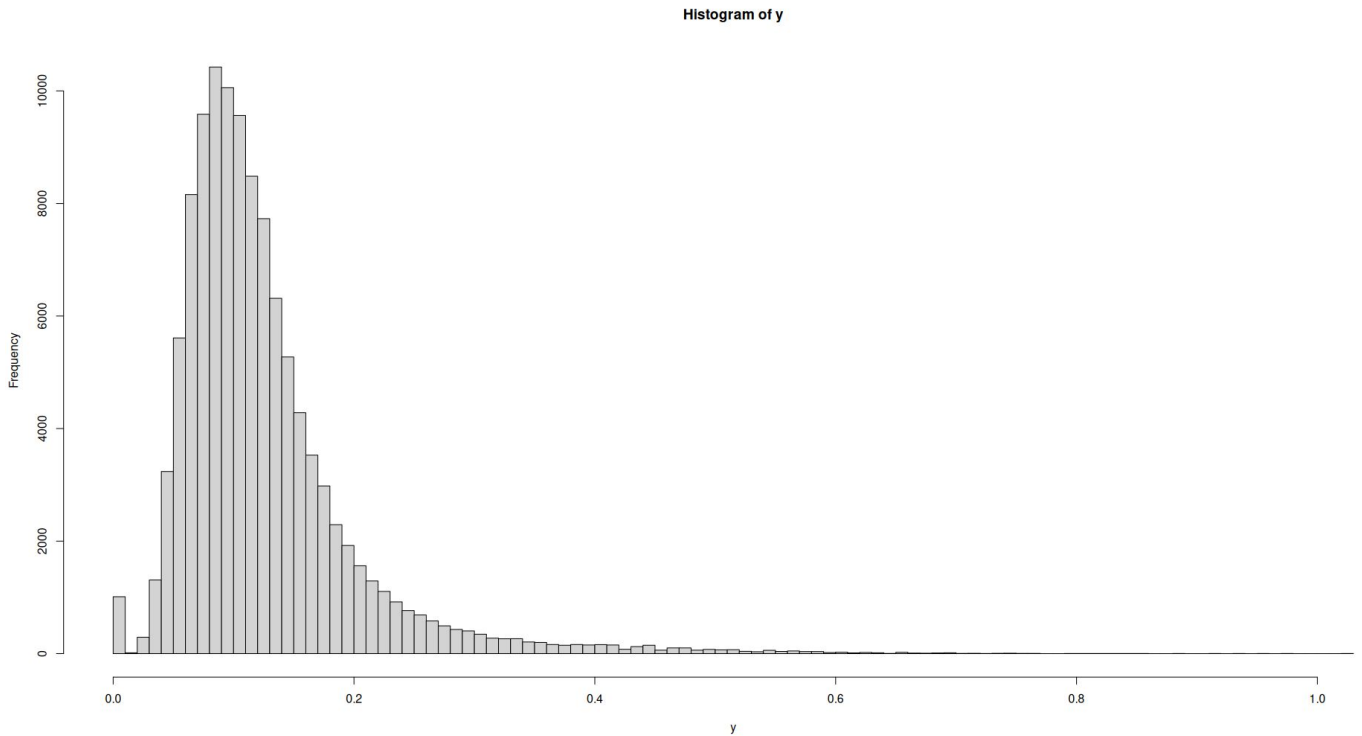
filtering

	g_1	g_2	...	g_n
g_1	0	1	...	1
g_2	1	0	...	0
...	...	...	0	...
g_n	1	0	...	0

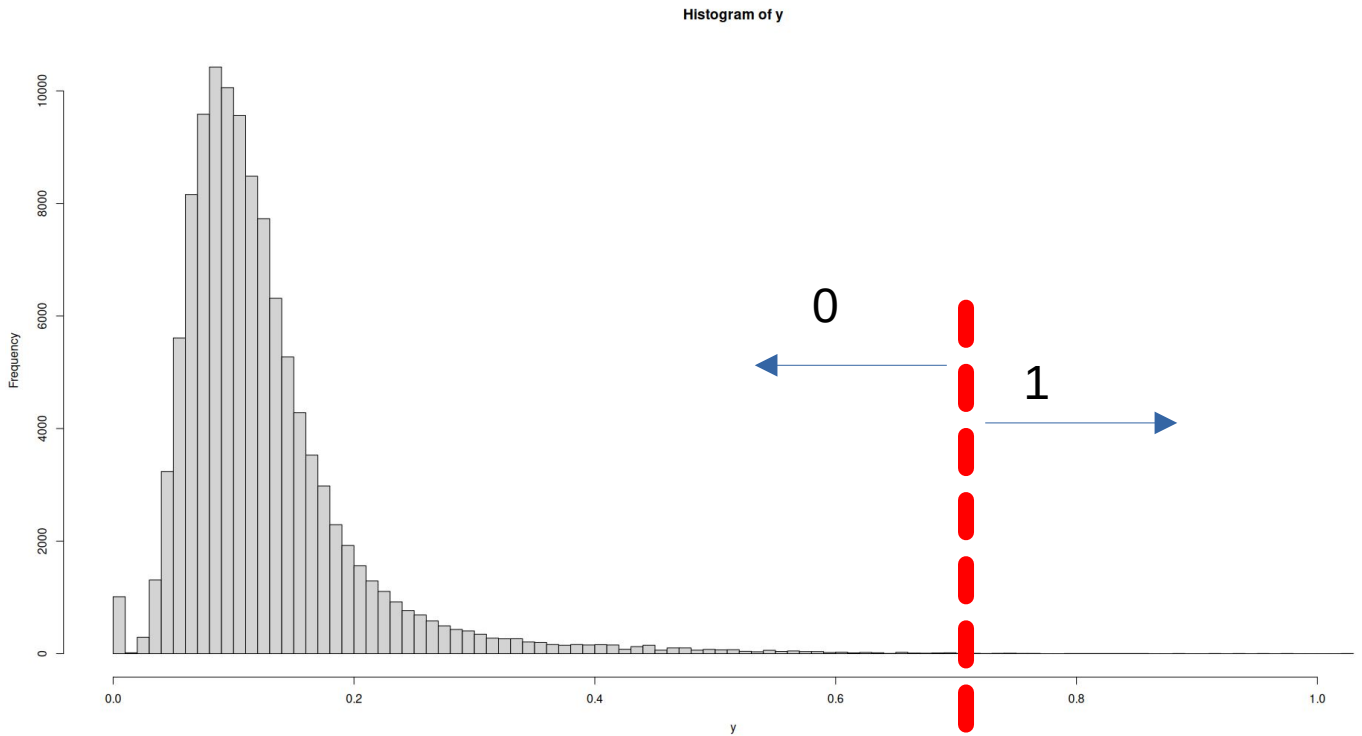
Network



Basic: thresholding (aka: “relevance networks”)

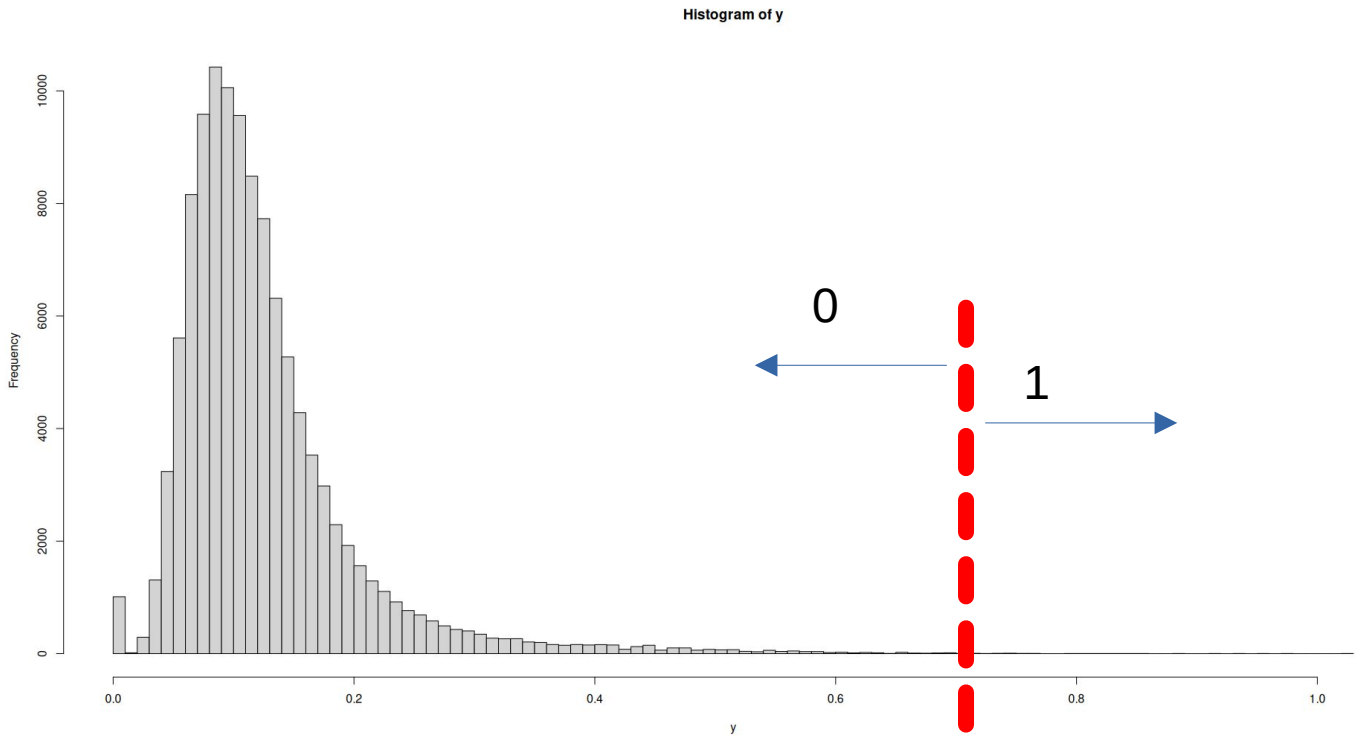


Basic: thresholding (aka: “relevance networks”)



https://doi.org/10.1142/9789814447331_0040

Basic: thresholding (aka: “relevance networks”)



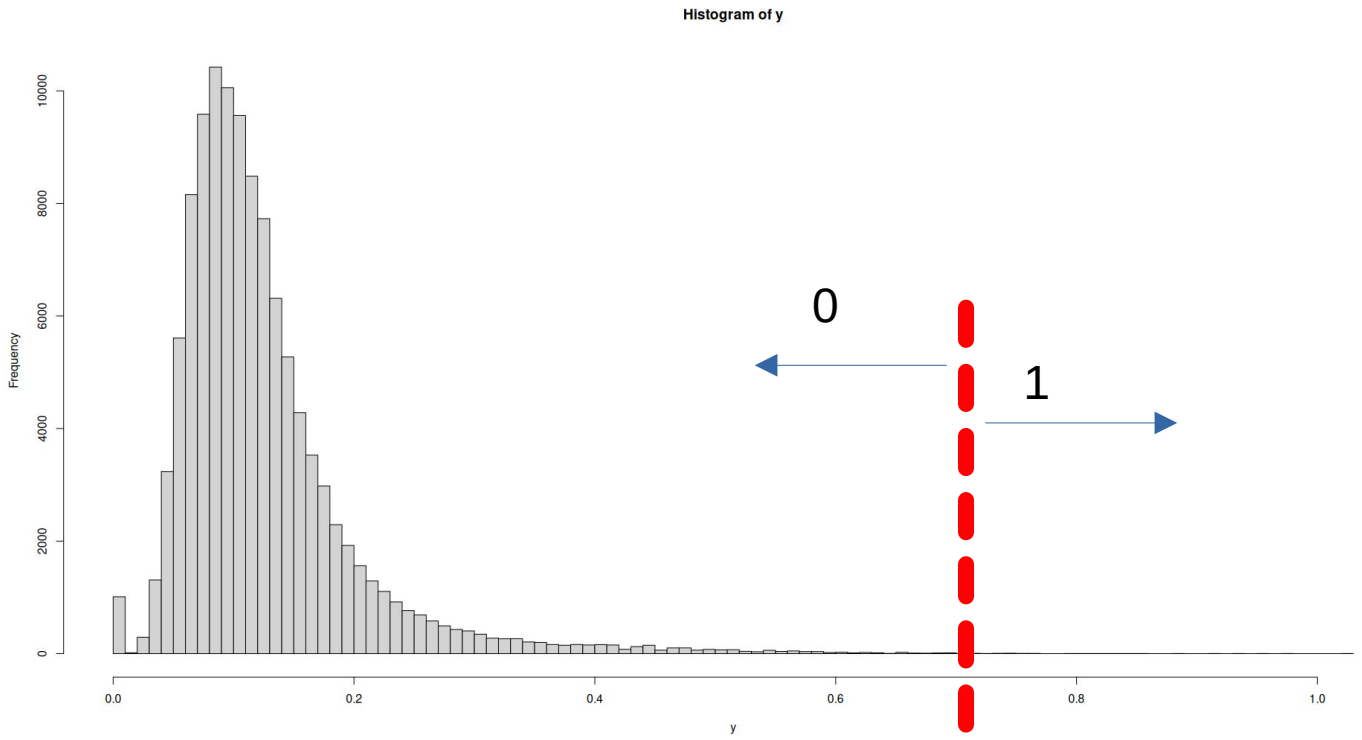
Rationale:

MI ~ permutation p.value

$$p(I \geq I_0 | \bar{I} = 0) \propto e^{-\alpha M I_0}$$

See: https://doi.org/10.1186/1471-2105-7-s1-s7

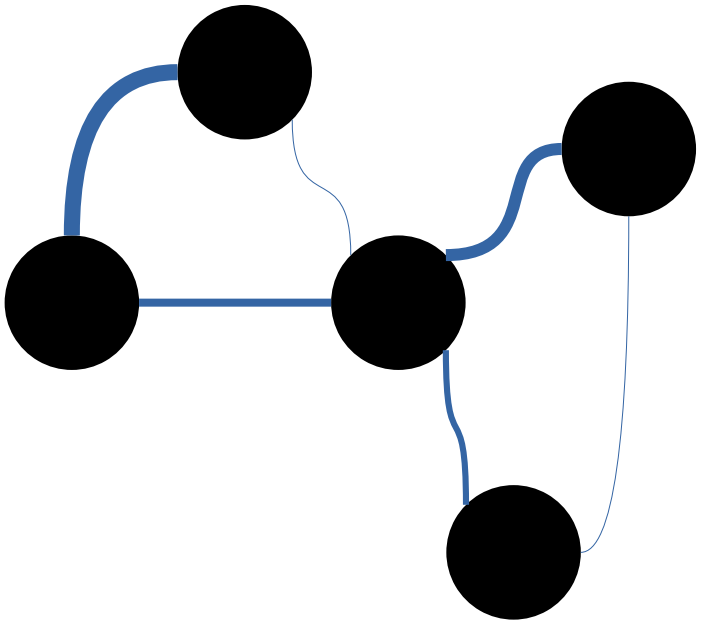
Basic: thresholding (aka: “relevance networks”)



Cool… so where do I set the threshold?

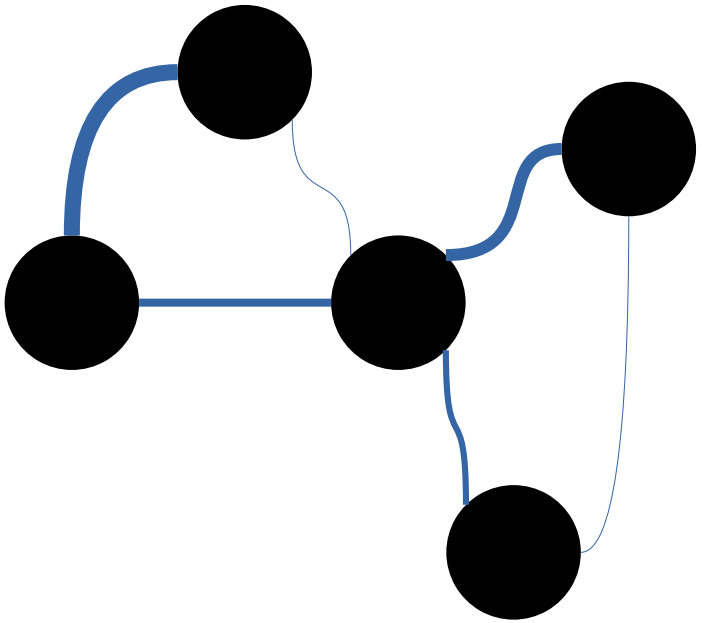
… need phenomenological input

Local thresholding (Backbone extraction)



How important is the edge in for a particular node?
- In terms of *edge weight* and *node strength*

Local thresholding (Backbone extraction)

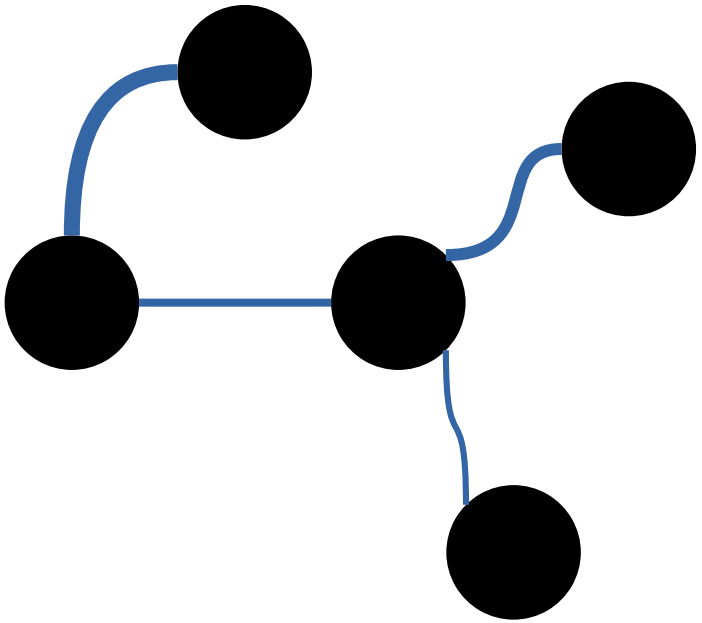


How important is the edge in for a particular node?
- In terms of *edge weight* and *node strength*

$$\Upsilon_i(k) \equiv kY_i(k) = k\sum_j p_{ij}^2.$$

Disparity measure

Local thresholding (Backbone extraction)

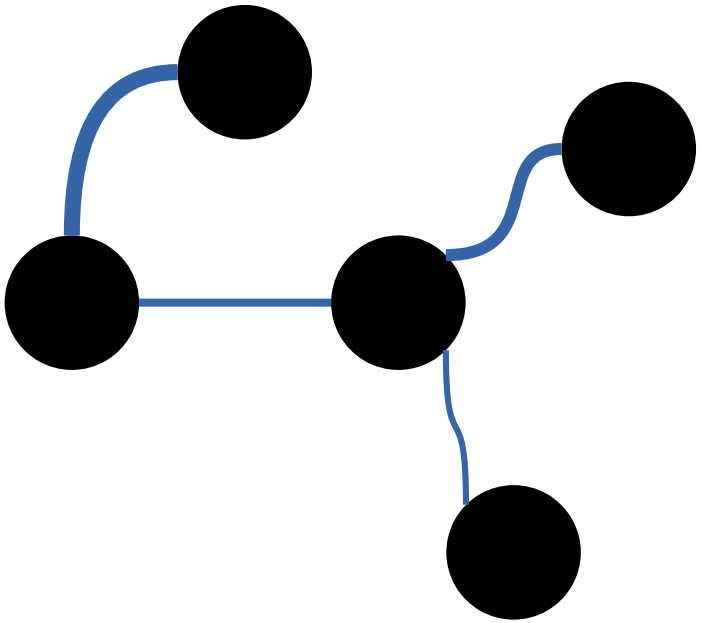


How important is the edge in for a particular node?
- In terms of *edge weight* and *node strength*

$$\Upsilon_i(k) \equiv kY_i(k) = k\sum_j p_{ij}^2.$$

Disparity measure

Local thresholding (Backbone extraction)



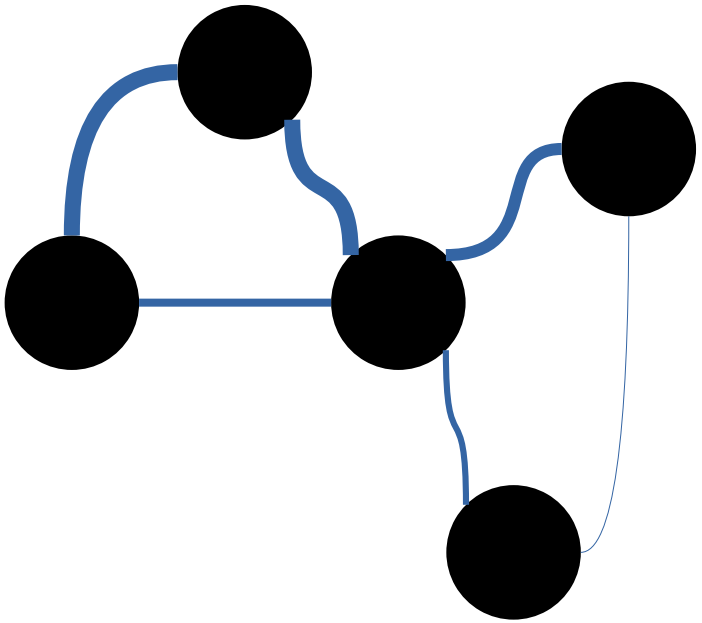
How important is the edge in for a particular node?
- In terms of *edge weight* and *node strength*

O(n³)
=(

$$\Upsilon_i(k) \equiv kY_i(k) = k\sum_j p_{ij}^2.$$

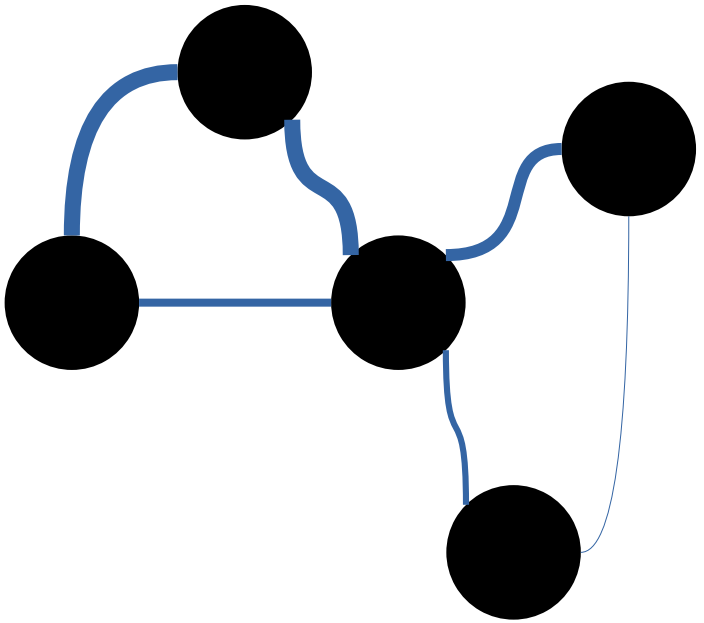
Disparity measure

Data processing inequality



Are nodes really interdependent, or do they appear to be so because they share a neighbor?

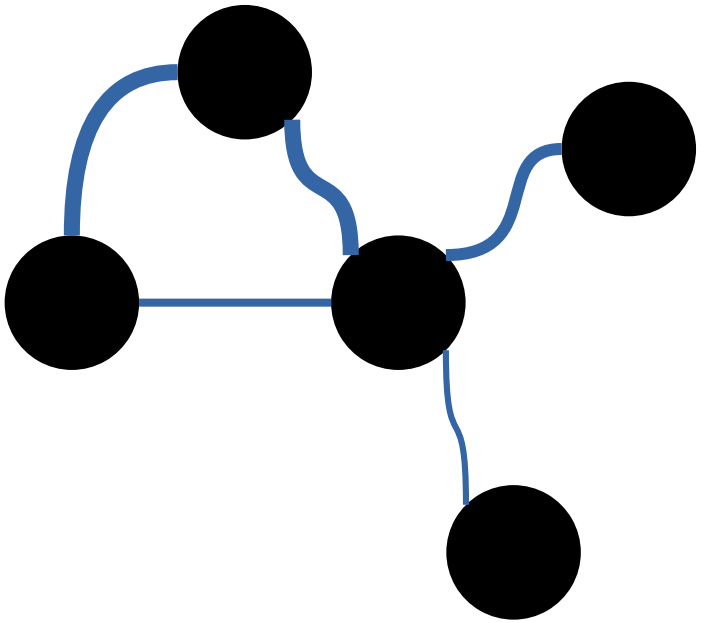
Data processing inequality



Are nodes really interdependent, or do they appear to be so because they share a neighbor?

$$I(X;Y) \geqslant I(X;Z)$$

Data processing inequality



Are nodes really interdependent, or do they appear to be so because they share a neighbor?

$$I(X;Y) \geqslant I(X;Z)$$

Now that we have a network…
It’s time to learn about it!

Microscale: how are nodes connected?
Mesoscale: what local features are there?
Macroscale: what’s the shape of the network?

Open Science

Free Software

Collaborative Culture

gdeanda@inmegen.edu.mx

@gdeandajauregui

github.com/
guillermodeandajauregui

guillermodeandajauregui.github.io