# Tracking information flow in biosystems from high-throughput data

Miron B. Kursa

Interdisciplinary Centre for Mathematical and Computational Modelling,
University of Warsaw

# High-throughput data

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{proteins} \rightarrow \text{lipids, sugars, small molecules}$$

- DNA and RNA are sequences of four symbols; we can read them.
- Proteins are sequences of twenty symbols; we can measure their abundance given sequences.
- We have libraries of small molecules and we can measure their abundance from their properties.

The basic functional unit at DNA level is a *gene*; we generally track products back to the gene resonsible for them.
Humans have about 20k protein-coding genes, fruit fly has 15k, bacteria have 2-5k, corn has 30k.

# Omics

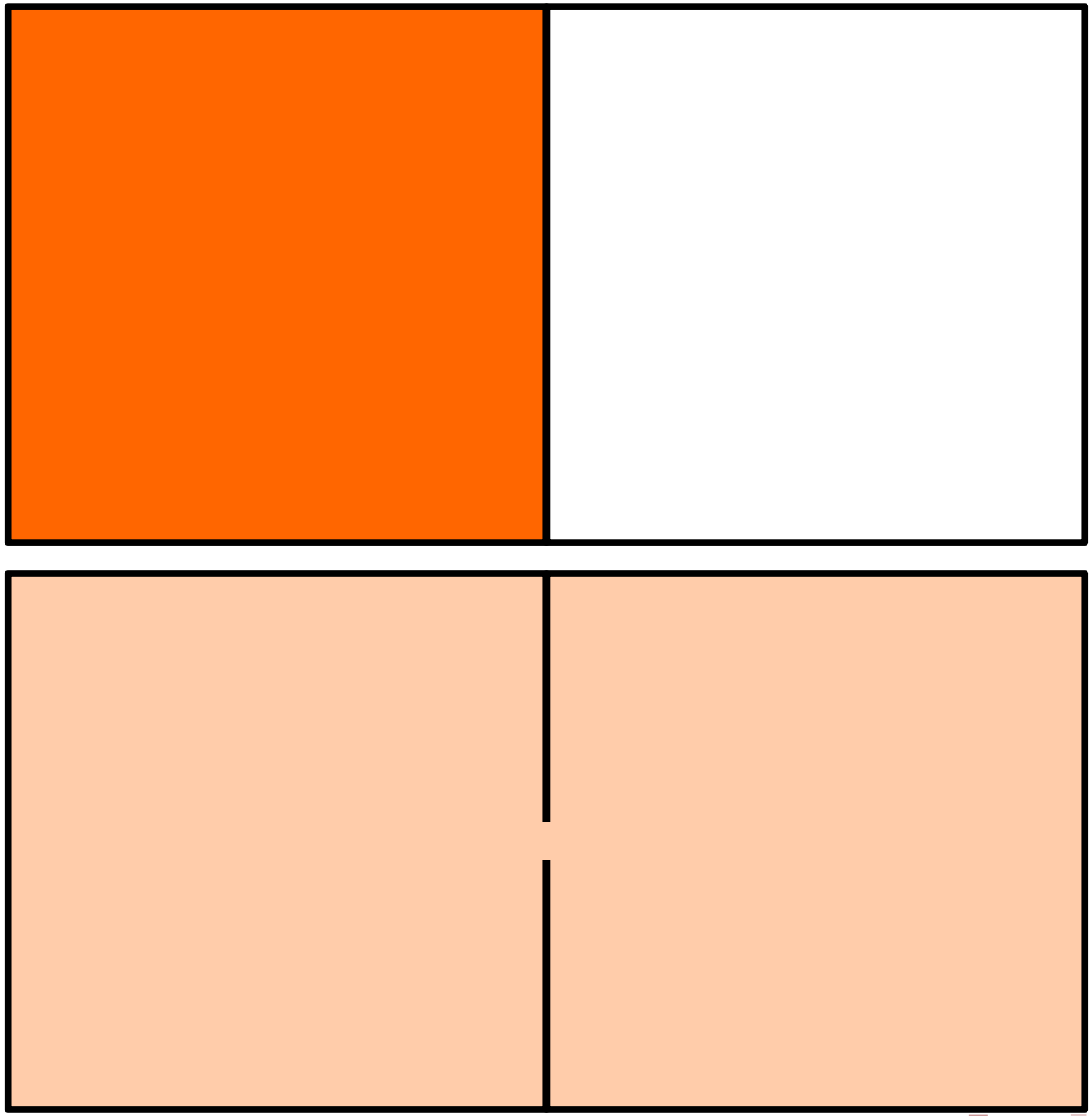
- Human has about 3G$\times$ 2 DNA symbols; genomes are *very* similar, hence genomics is generally a list of changes vs. reference; originally looked at about 1.5M differences, new data bases list hundreds of millions. This is called *genomics*. And there are modifications, DNA structure... In general, *epigenomics*.
- Most cells in human body are bacteria (35T vs. 25T RBCs and 5T other cells). Collective mix of DNA is generally demuxed into species list. This is called *metagenomics*.
- RNA can be sequenced producing a sort of gene-activity scores, this is called *transcriptomics*. The problem is that transcription varies a lot in a life of a cell, so now we try to this for every cell separately, which is called *single-cell*.
- Proteins can be cut in pieces in a smart way and then measured with mass spectroscopy; this produces a list of few thousand numbers and is called *proteomics*.
- Mass spectrometry or chromatographic techniques can measure small molecules; this is called *metabolomics* (*lipidomics*, *glycomics*, ...).

# High-throughput — statistical perspective

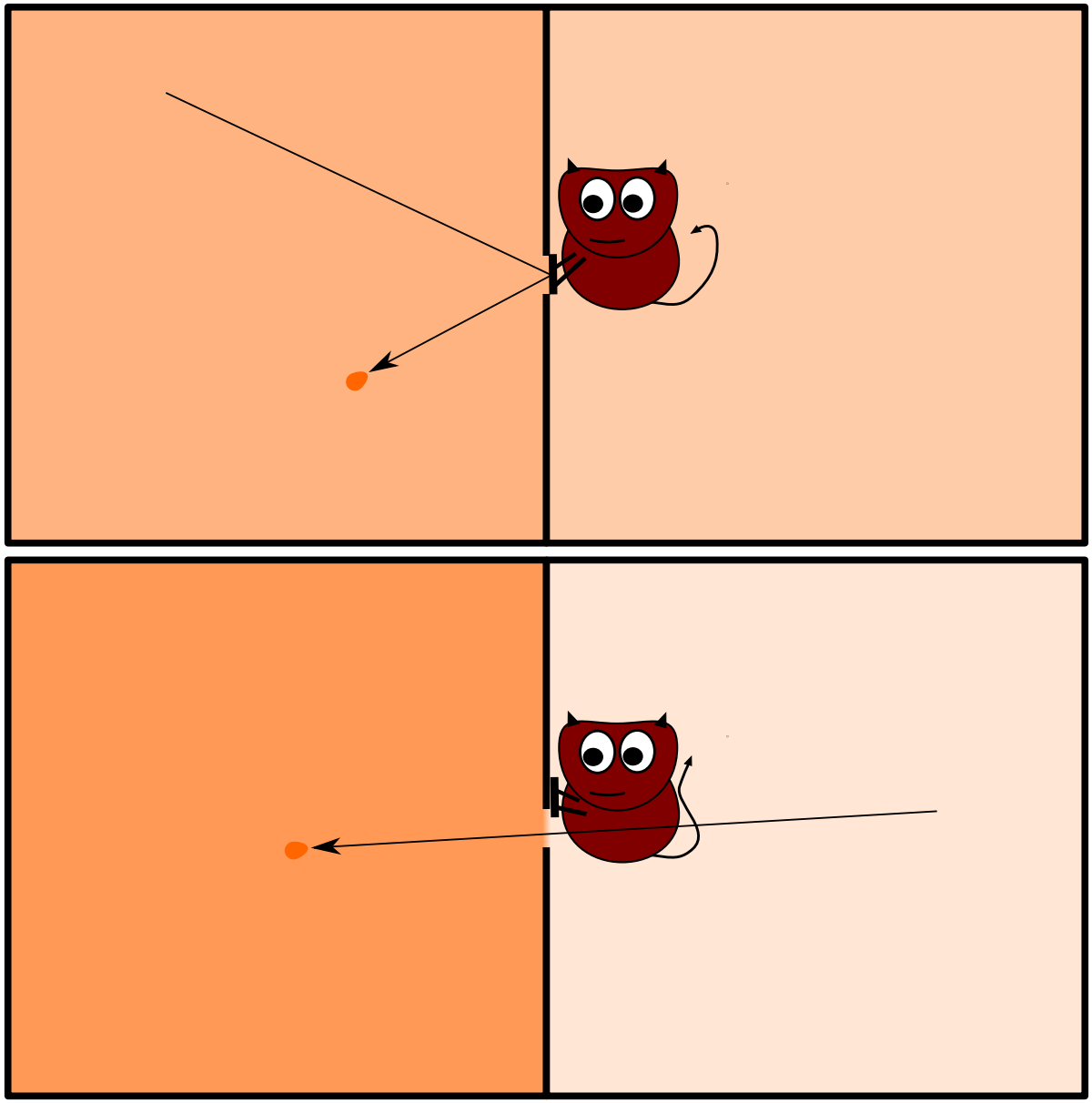
- Lots of *features*, little *observations*.
- We measure what we *can*, not what we *want* (suspect to be important).
- A vast majority of features are gonna be irrelevant.
- Feature is its context; we want to retain it to have any chance to interpret the result (not that everybody cares).
- Context information is called *ontology*; ontologies include known expression patterns, interactions (metabolic pathways in particular), taxonomic information, mentions in the literature...
- When we don't understand something, we decrease ontology resolution (*enrichment*) or look for similarities leveraging the fact that things evolve from other things via small changes rather than get clear-sheet designed (usually).

From now on, I will assume we compare two groups of observations (perturbed–control, interesting–reference); this defines feature $Y$ (*decision*, *anchor*) that maps observation to the group.

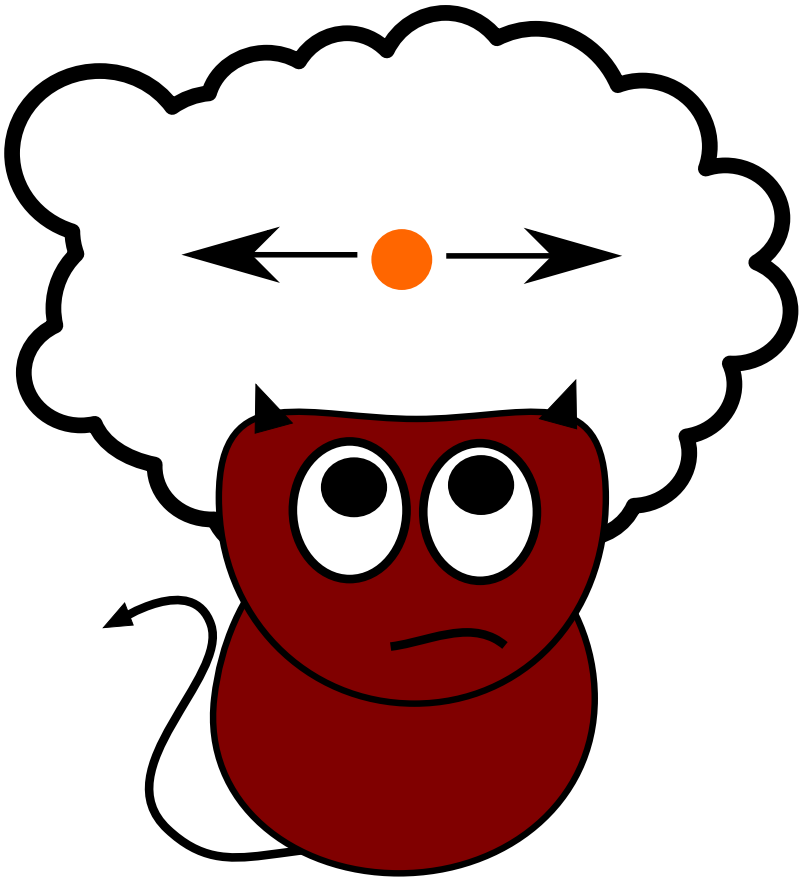
# Entropy



# Maxwell's daemon



# Maxwell's daemon



- Daemon needs *information* to properly operate the door.
- So, information is just negative entropy?
- Landauer showed that is does not break energy conservation, since requires irreversible physical process to carry computation.
- *Szilard engine* has been actually built in Japan (Toyabe 2010).
- Shannon used this to build the *information theory* as we know it today.

# Entropy

From the axioms of information theory, we get *information entropy* defined as

$$H = -\sum_{\text{states}} p_i \log p_i; \tag{1}$$

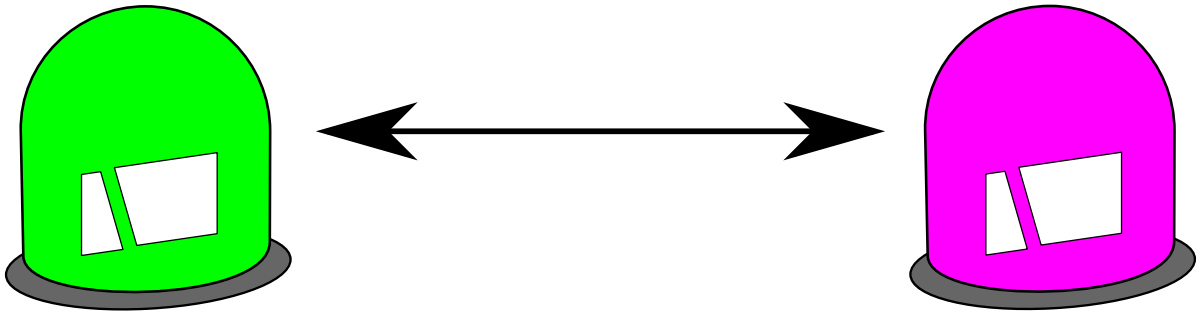
when $p_i = N_{\text{states}}^{-1}$

$$H = \log N_{\text{states}}, \tag{2}$$

which recovers Boltzmann's entropy.

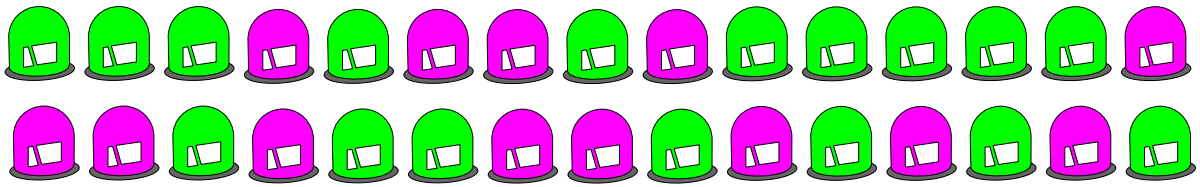
# Shannon's information entropy

$$H = -\sum_{\text{states}} p_i \log p_i \tag{3}$$

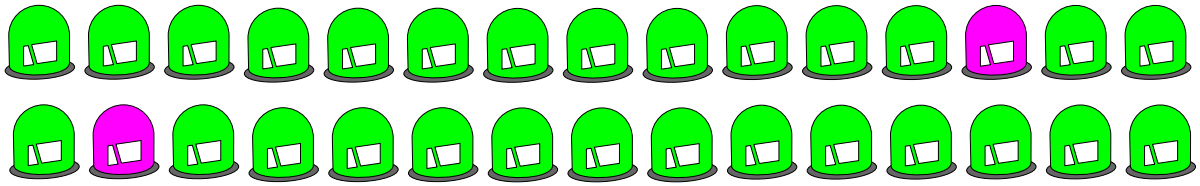


$$H = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2 = 0.693 \text{ nats} = 1 \text{ bit} \tag{4}$$

# Shannon's information entropy



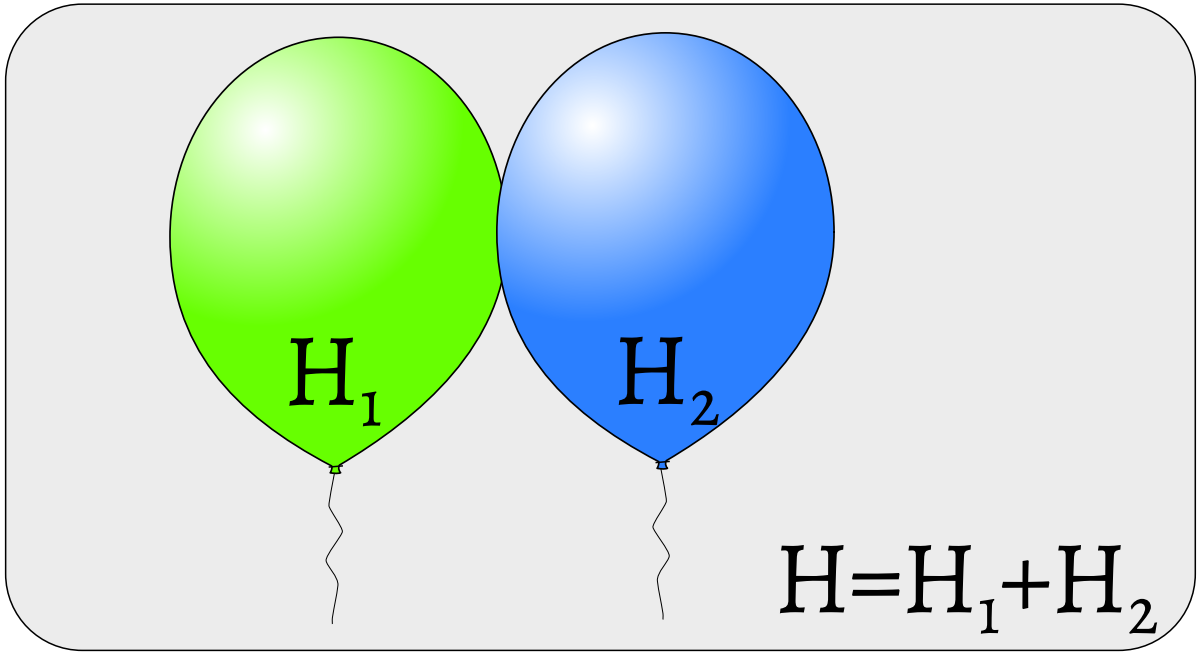
$$\hat{P}_v = \frac{13}{30} \rightarrow \hat{H} = 0.99 \text{ bits} \tag{5}$$



$$\hat{P}_v = \frac{1}{15} \rightarrow \hat{H} = 0.35 \text{ bits} \tag{6}$$

The less predictable variable is (i.e., the closest states are to be equi-probable), the higher entropy it scores. Highly-entropic variables are more informative, because it is less effective to predict their values from history.

# Extensitivity of physical entropy




States of the gas in both balloons are independent, hence

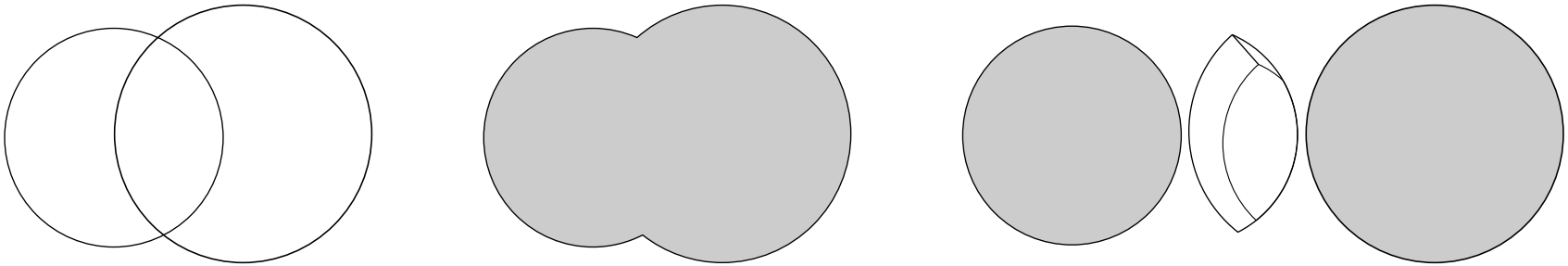
$$N_{1+2} = N_1 \times N_2, \tag{7}$$

this way

$$H_{1+2} = \log(N_1 \times N_2) = \log(N_1) + \log(N_2) = H_1 + H_2 \tag{8}$$

# Mutual information

Two random varibales are not necessary independent; then the joint probability $p(A, B) \neq p(A)p(B)$.



In the IT language, iformation can only be learned once, which means, that when we observe two variables together,

$$H(A, B) = H(A) + H(B) - \text{common information in A and B} \tag{9}$$

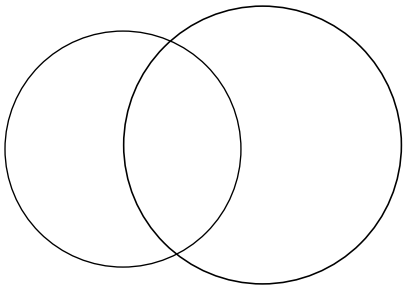
This common information is called *mutual information*

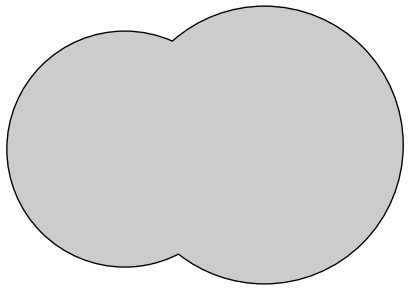
$$I(A; B) = H(A) + H(B) - H(A, B). \tag{10}$$

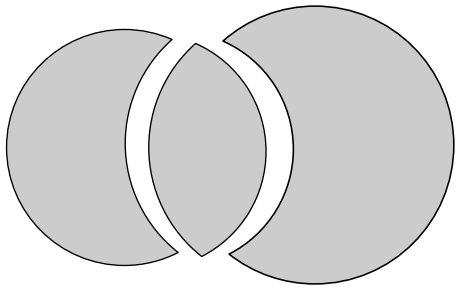
# Conditional entropy

$$I(A;B) = H(A) + H(B) - H(A,B) \tag{11}$$

This also relates to conditional entropies







$$H(A,B) = H(A|B) + I(A;B) + H(B|A); \tag{12}$$

in other words, conditioning means: *what I'm left with to learn if I already know the condition?*

# Conditional mutual information

$$I(A;B|C) = H(A|C) + H(B|C) - H(A,B|C) \tag{13}$$

Interestingly, while $H(A|C) \leqslant H(A)$,

- $I(A;B|C)$ can be smaller $I(A;B)$, indicating *redundancy*, i.e., that information in $C$ *duplicates* this shared between $A$ and $B$.
- $I(A;B|C)$ can be larger $I(A;B)$, indicating *synergy*, i.e., that information in $C$ *complements* this shared between $A$ and $B$.

# Trivariate mutual information

Turns out

$$I(A;B) - I(A;B|C) = I(A;C) - I(A;C|B) = I(B;C) - I(B;C|A), \tag{14}$$

i.e., the order does not matter, and interaction is inherent to a mix.
This value is called mutual information of three variables,

$$I(A;B;C) = I(A;B) - I(A;B|C), \tag{15}$$

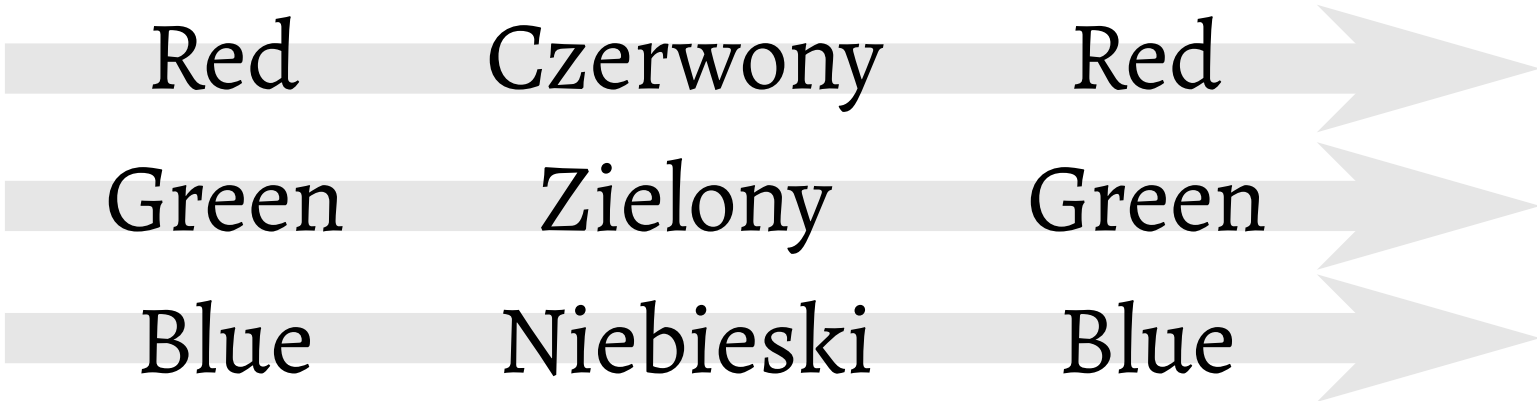
and is *positive* for redundancy, *negative* for synergy, and *zero* for independence (at the level of trivariate interactions).

# Data processing inequality

In a mediated, $A \to M \to B$ circuit, $M$ can at most destroy information, so

$$I(A;M) \geqslant I(A;B); \tag{16}$$

equality is reached only if $I(A;B|M) = 0$, that is when $M$ relies information perfectly.

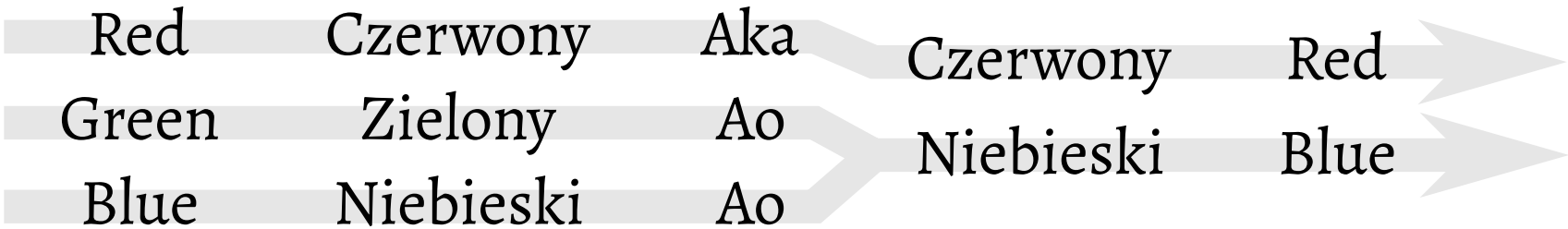


# Data processing inequality

In a mediated, $A \to M \to B$ circuit, $M$ can at most destroy information, so

$$I(A;M) \geqslant I(A;B); \tag{16}$$

equality is reached only if $I(A;B|M) = 0$, that is when $M$ relies information perfectly.



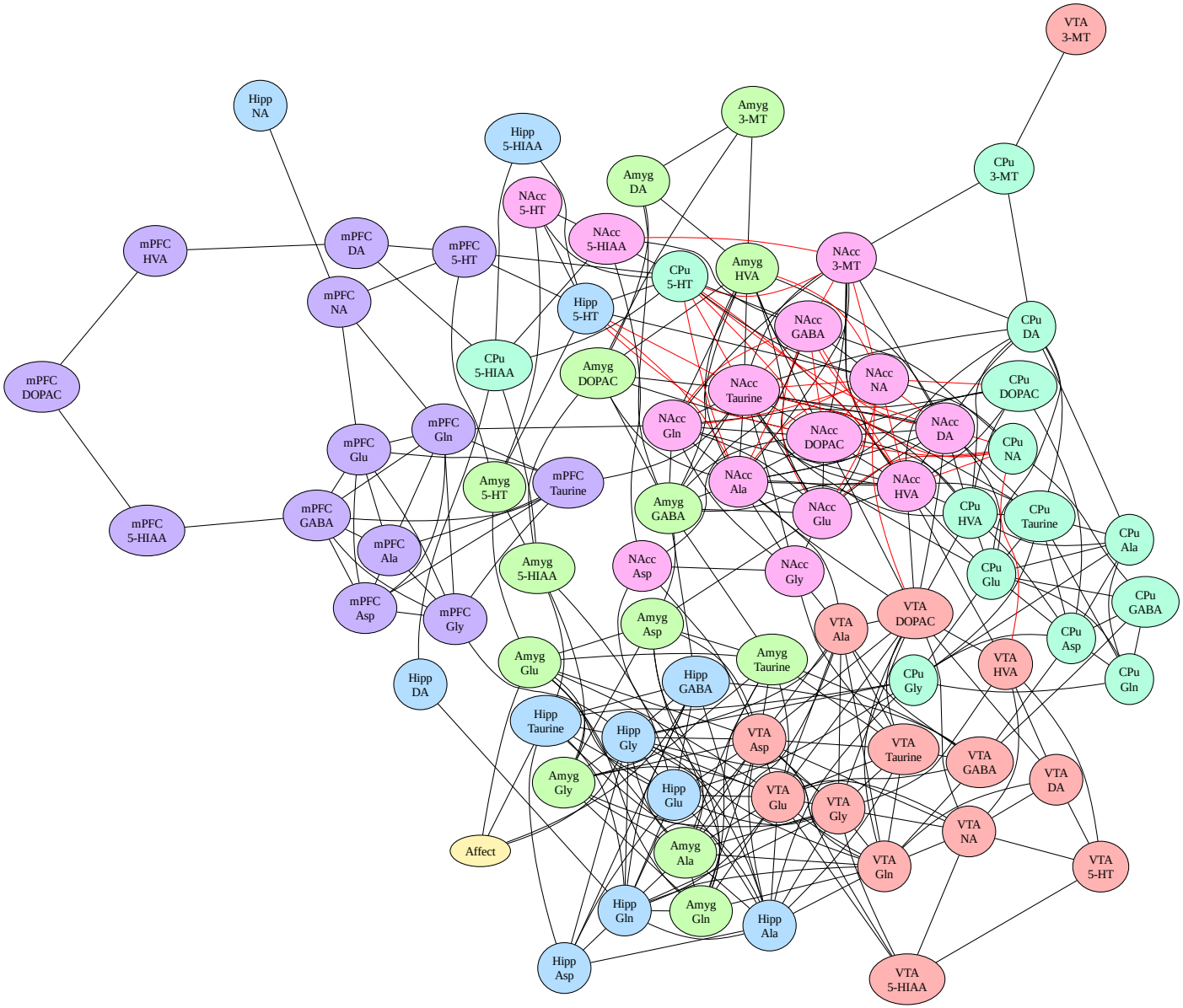
# Omic analysis

The analysis of high-throughput data has two typical aims:

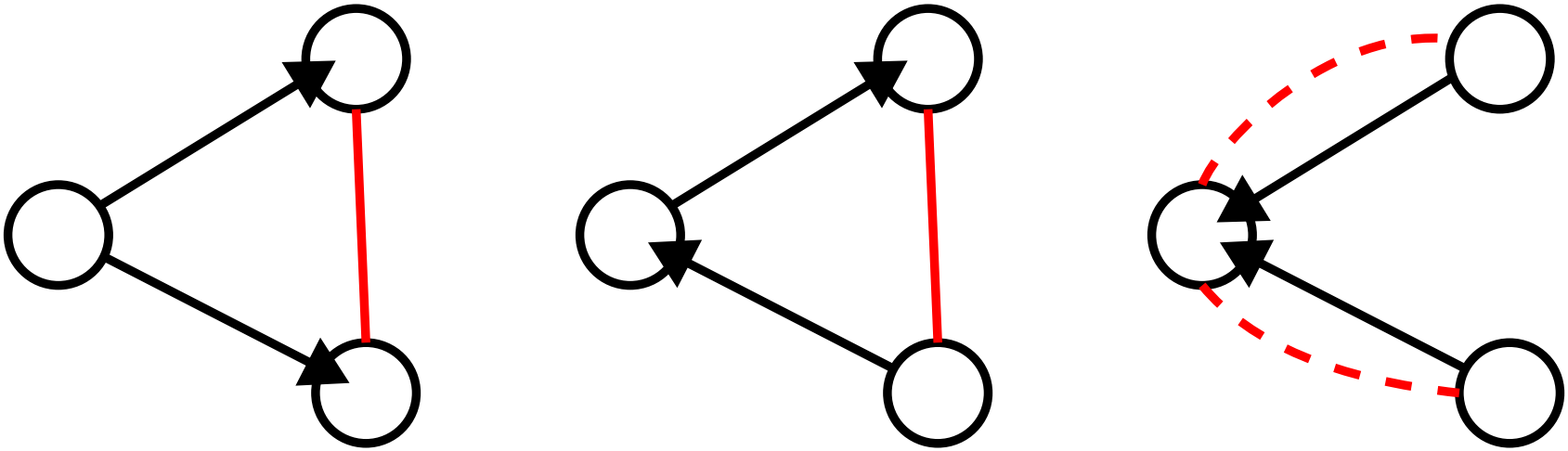
- **Differences** — checking what differs between experimental groups, like why this group is odd? Or, what happenened after we did that?
- **Interactions** — checking which agents interacted in a given context, like that this change when we fiddle with that? Or do they just move together?

Differences are generally digged though some rudimentary testing; my brick in this garden is *Boruta* (a relevant feature selection method based on RF), which does have some traction.
Interactions are modelled with, obviously, graphs; more effort is spent on analysing graphs than making them, correlation coefficients still rule here. To be fair, proper network inference is borderline impossible.

# Correlation network



# Correlation network problems



- **They're undirected.** Causality aside, we don't see inconsistencies.
- **They have non-causal edges.** The worse part is the tangle of numerous indirect links, e.g. $A \to C$ from $A \to B \to C$.
- **They lack causal edges.** We don't see subtle effects or any non-marginal parts of multivariate ones.
- **They are a bad model.** In real life, we have loops (a lot of them), and multi-modalities, and extra few magnitudes of complexity.

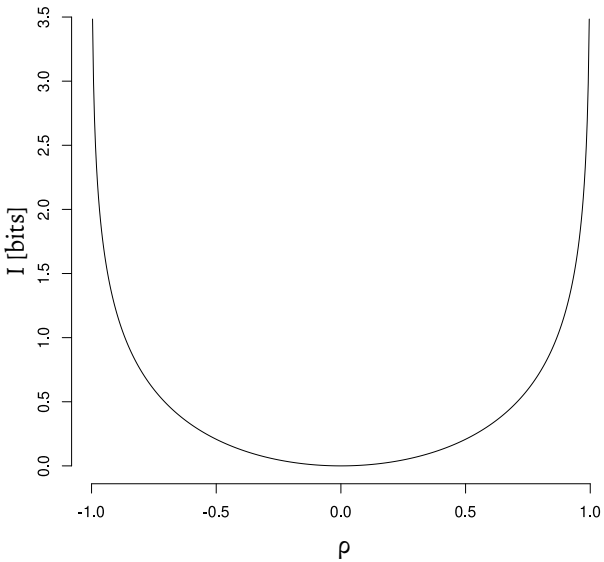
Causality is a lost cause; we can have a latent confounder that have set-up everything to deceive us.

# Mutual information & correlation

x As it turns up, for $X, Y \sim \mathcal{N}$, in a differential entropy framework,

$$I(X;Y) = -\log \sqrt{1-\rho(X,Y)^2}, \tag{17}$$

where $\rho(X,Y)$ is the Pearson correlation.




Correlation of 0 has 0 bits, of 50% (or -50%) has 0.21 bits, 90% 1.2 bits, and 99% 2.8 bits. 100% is oops...

Can we omit the singularity, drop normality assumptions and have nice things like trivariate MI?

# Kendall transformation

Instead of messing with differential entropy, we may just want to preserve ranking, like in non-parametric statistics.

| $A$ |
|---|
| 7.7 |
| 8.2 |
| 3.2 |
| 5.3 |
| 2.8 |

| | 7.7 | 8.2 | 3.2 | 5.3 | 2.8 |
|---|---|---|---|---|---|
| 7.7 | – | △ | ▽ | ▽ | ▽ |
| 8.2 | ▽ | – | ▽ | ▽ | ▽ |
| 3.2 | △ | △ | – | △ | ▽ |
| 5.3 | △ | △ | ▽ | – | ▽ |
| 2.8 | △ | △ | △ | △ | – |

$$A^{\mathcal{K}} = [\triangledown, \triangle, \triangle, \triangle, \triangle, \triangle, \triangle, \triangle, \triangledown, \triangledown, \triangledown, \triangle, \triangledown, \triangledown, \triangle, \triangle, \triangledown, \triangledown, \triangledown, \triangledown]$$

- Equal values are encoded by the third state, □.
- Hence, $n$ numbers get encoded into a $m := n(n-1)$ ternary values.

# KT & correlation

- Kendall correlation coefficient

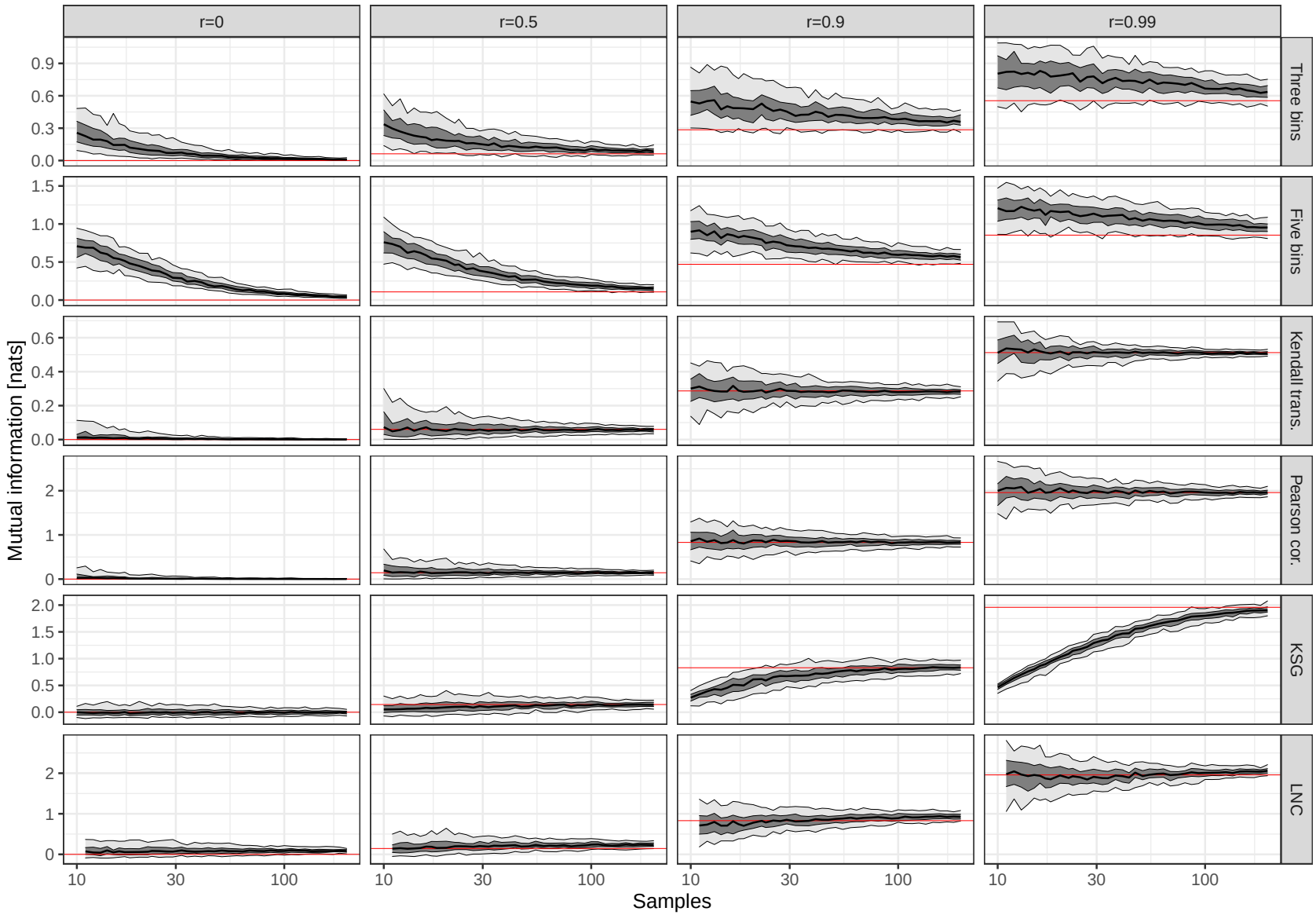
$$\tau(A, B) = \frac{2}{m} \sum_i \mathbf{1}(A_i^{\mathcal{K}} = B_i^{\mathcal{K}}) - 1.$$

- Mutual information is a function of $\tau$

$$I(A^{\mathcal{K}}, B^{\mathcal{K}}) = \tau \log \sqrt{\frac{1+\tau}{1-\tau}} + \log \sqrt{1-\tau^2}.$$

Doesn't explode for $\tau = \pm 1$!

# Bivariate normal benchmark



# KT & binary variables

| $X$ |
|---|
| a |
| b |
| a |
| a |
| b |

| | a | b | a | a | b |
|---|---|---|---|---|---|
| a | – | ▽ | □ | □ | ▽ |
| b | △ | – | △ | △ | □ |
| a | □ | ▽ | – | □ | ▽ |
| a | □ | ▽ | □ | – | ▽ |
| b | △ | □ | △ | △ | – |

$$X^{\mathcal{K}} = [\triangle, \square, \square, \triangle, \triangledown, \triangledown, \triangledown, \square, \square, \triangle, \square, \triangle, \square, \triangle, \square, \triangle, \triangledown, \square, \triangledown, \triangledown]$$

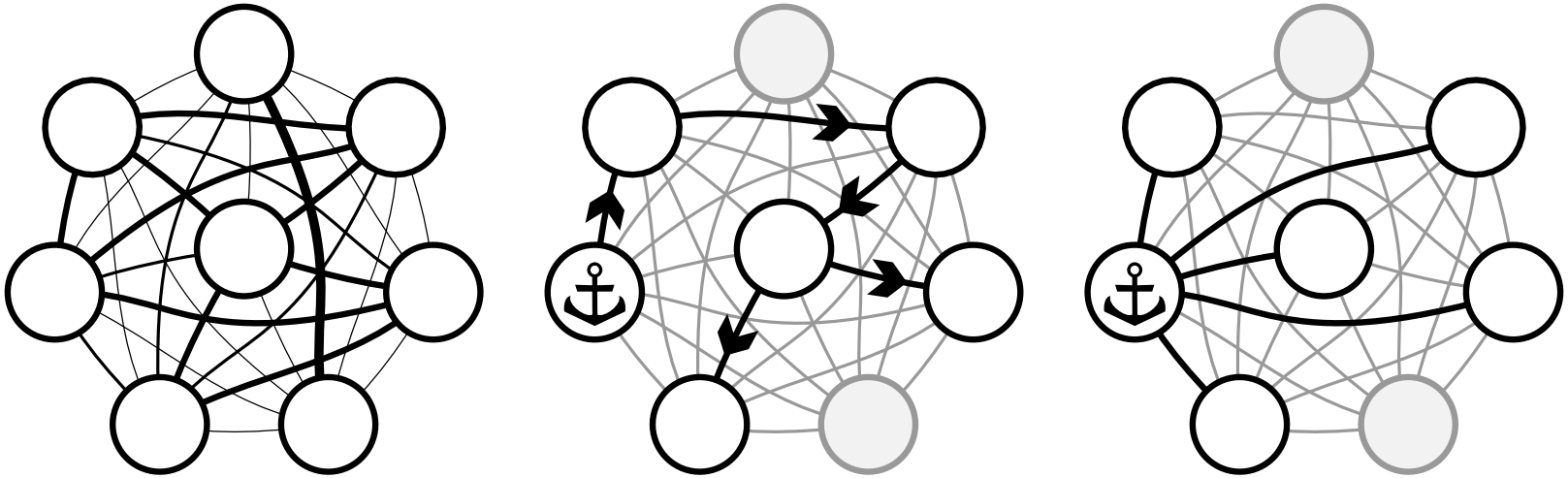
MI turns out to be a function of the area under ROC curve, $A$:

$$I^{\mathcal{K}}(A; a; b) = \frac{2n_a n_b}{n(n-1)} \left( A \log \frac{A}{1-A} + \log(2-2A) \right), \tag{18}$$

Also Mann-Whitney-Wilcoxon test, as its statistic $U = n_a n_b (1 - A)$.

# Paradigm merge

But what if we could merge network inference with feature selection?



- The basic idea is to trace effects of the perturbation ($Y$).
- By tracing information we skip loops (no re-learning); the resulting graph will be a tree.
- How to formalise this?

# Influence path

We have our MI network, a weighted undirected graph

$$G(V,E): \quad V = X, \quad E = V^2, \quad w : (v_i, v_j) \to I(v_i; v_j). \tag{19}$$

We would like to find a path $P = (p_1, \ldots, p_n) \in V^n$, so that

- $p_1$ and $p_n$ are user-defined (the path is anchored);
- $\min_i I(p_{i-1}; p_i; p_{i+1})$ is maximal and positive (the bottleneck trio is widest possible);
- $\forall_i I(p_{i-1}; p_i) \geqslant I(p_{i-1}; p_{i+1})$ (DPI holds locally).
- $\forall_i I(Y; p_i) \geqslant I(Y; p_{i+1})$ (DPI holds globally).

This seems as an optimisation problem, but it turns out we can use a trick, and use a *pair graph*, a weighed directed graph

$$G^*(E,F): \quad E = X^2, \quad F = E^2, \tag{20}$$

$$w^* : ((v_i, v_j), (v_{j'}, v_k)) \to \iota(v_i, v_j, v_k) \cdot \mathbf{1}(j = j'). \tag{21}$$

# Pair graph

$$G^*(E,F): \quad E = X^2, \quad F = E^2,$$

$$w^* : ((v_i, v_j), (v_{j'}, v_k)) \to \iota(v_i, v_j, v_k) \cdot \mathbf{1}(j = j').$$

The $\iota$ function allows us to formulate requirements in an algorithmically friendly way;

$$\iota_0(A, B, C) := \max\{I(A; B; C), 0\} \tag{22}$$

will propagate us through widest paths, while

$$\iota_*(A, B, C) := \iota_0(A, B, C) \cdot \mathbf{1}(I(A; B) \geqslant I(A; C)) \tag{23}$$

enforces local DPI, and

$$\iota_\to(A, B, C) := \iota_*(A, B, C) \cdot \mathbf{1}(I(Y; B) \geqslant I(Y; C)) \tag{24}$$

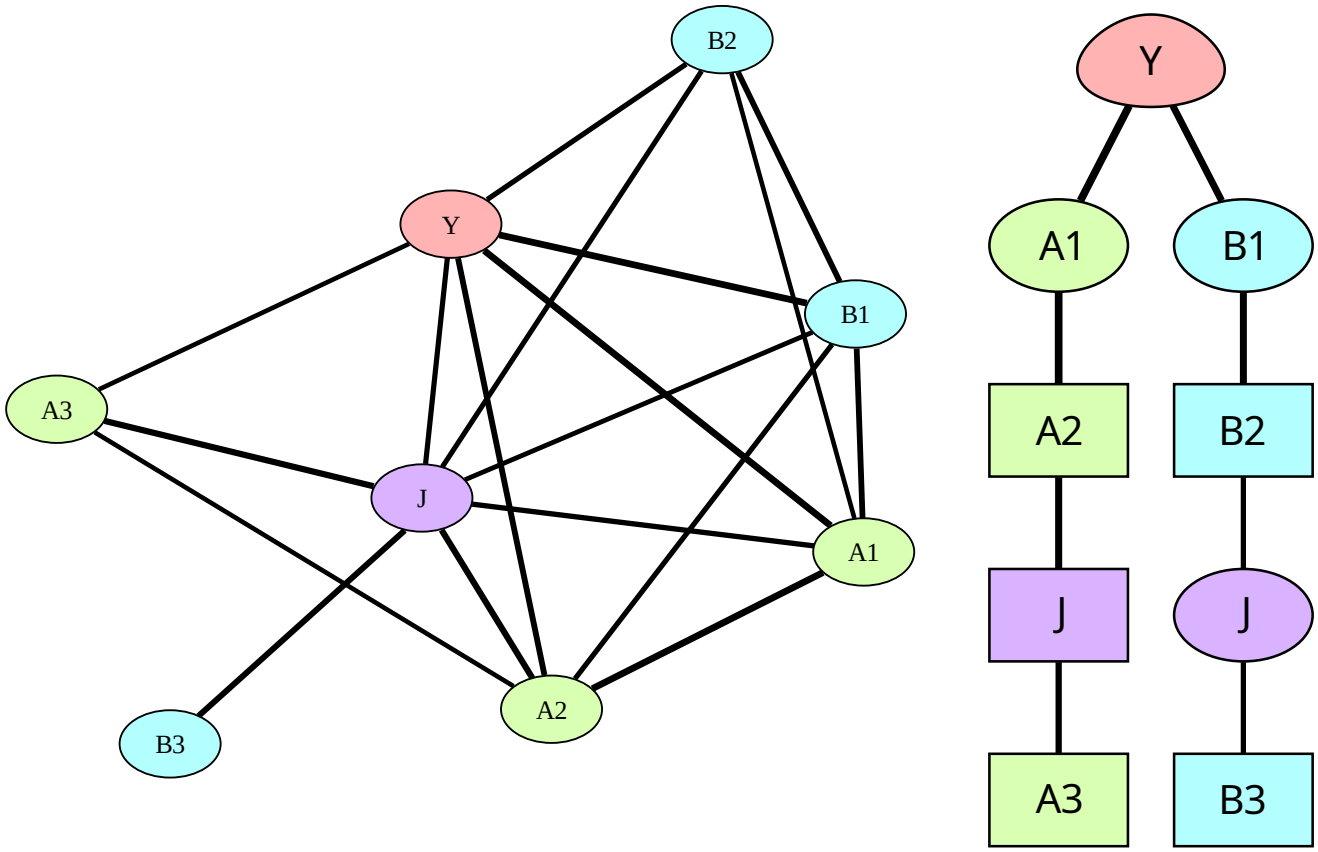
the global one.

# Widest paths on the edge graph

- Having $G^*$ and $\iota$ we can apply Dijkstra's algorithm to find a widest path tree from a designated vertex $Y$ to any other.
- We start by initiating the queue with all $(Y, x_i, x_{j\neq i})$ stubs of paths and expanding them according to maximal minimal $\iota$ value over a path.
- We can round small scores into 0 in $\iota$ to cap weak mediation; this adds a parameter, though.
- In practice, $G^*$ should not be constructed explicitly and proper $\iota$ caching is crucial for efficiency; also, queue is saturated with increase-key operations.

This is the *Vistla* method.

# *Junction* example

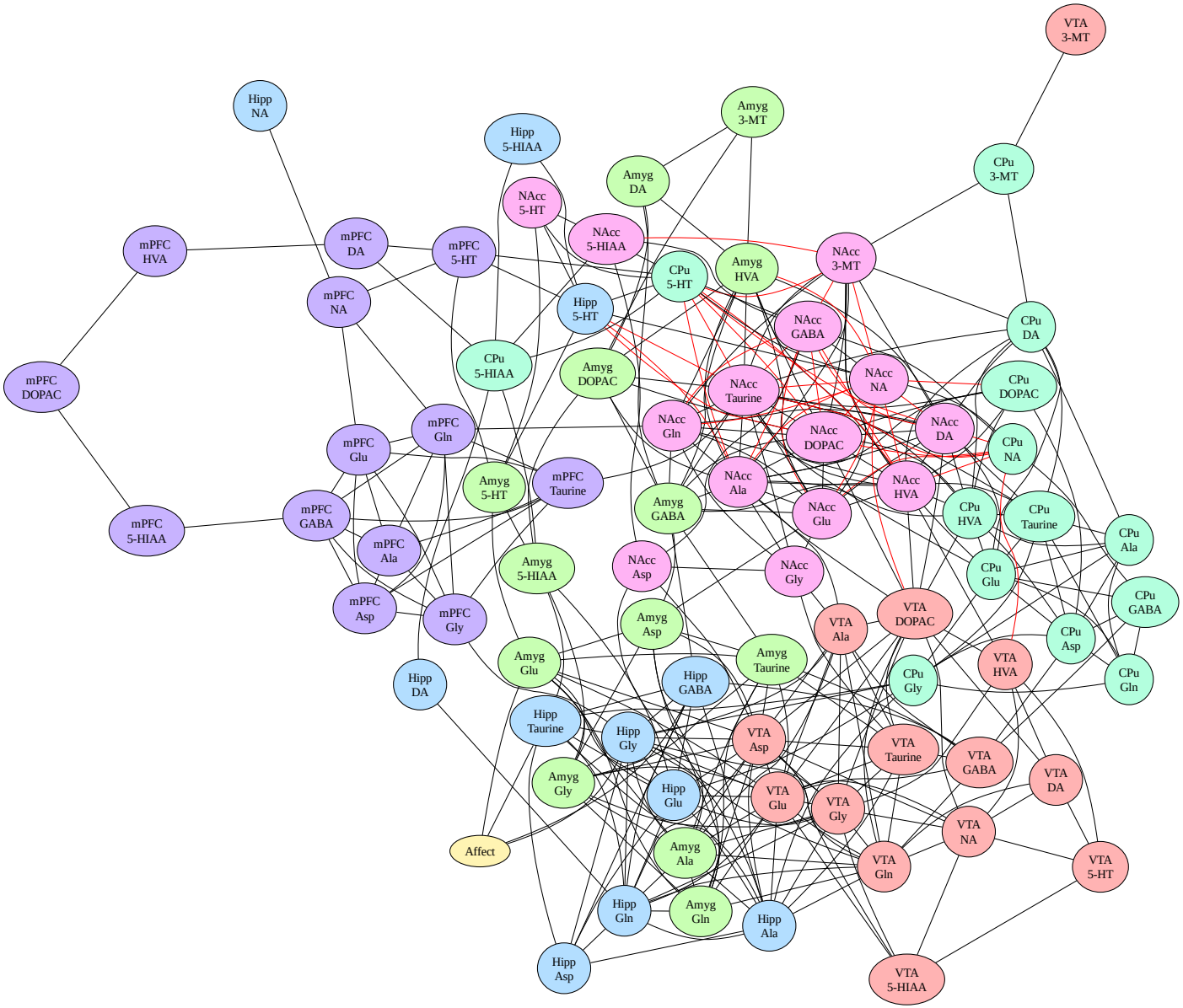


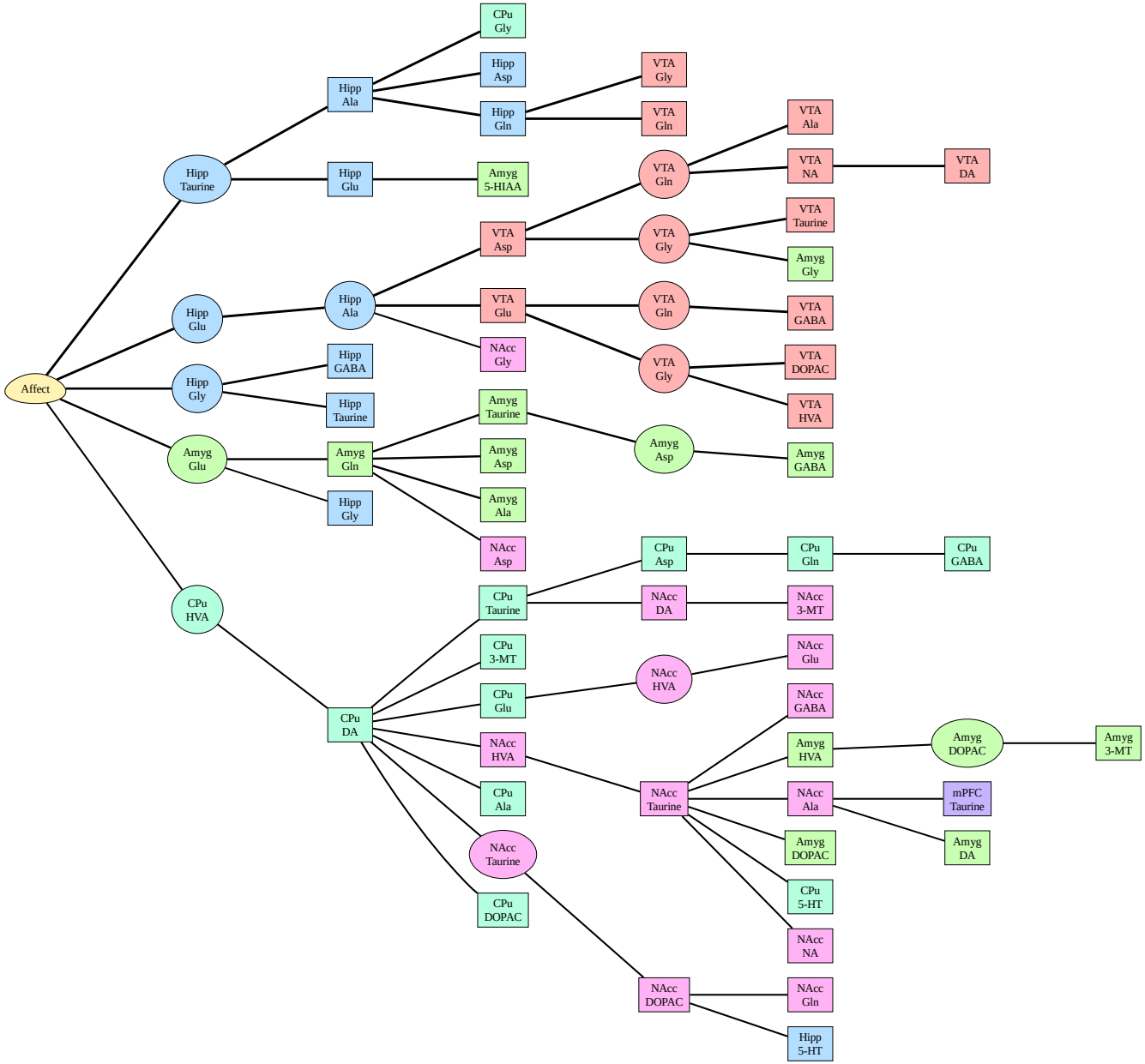

This data was generated from two Bayes networks, $Y \to A_1 \to A_2 \to J_A \to A_3$ and $Y \to B_1 \to B_2 \to J_B \to B_3$, but only $J = J_A \times J_B$ is in the data.

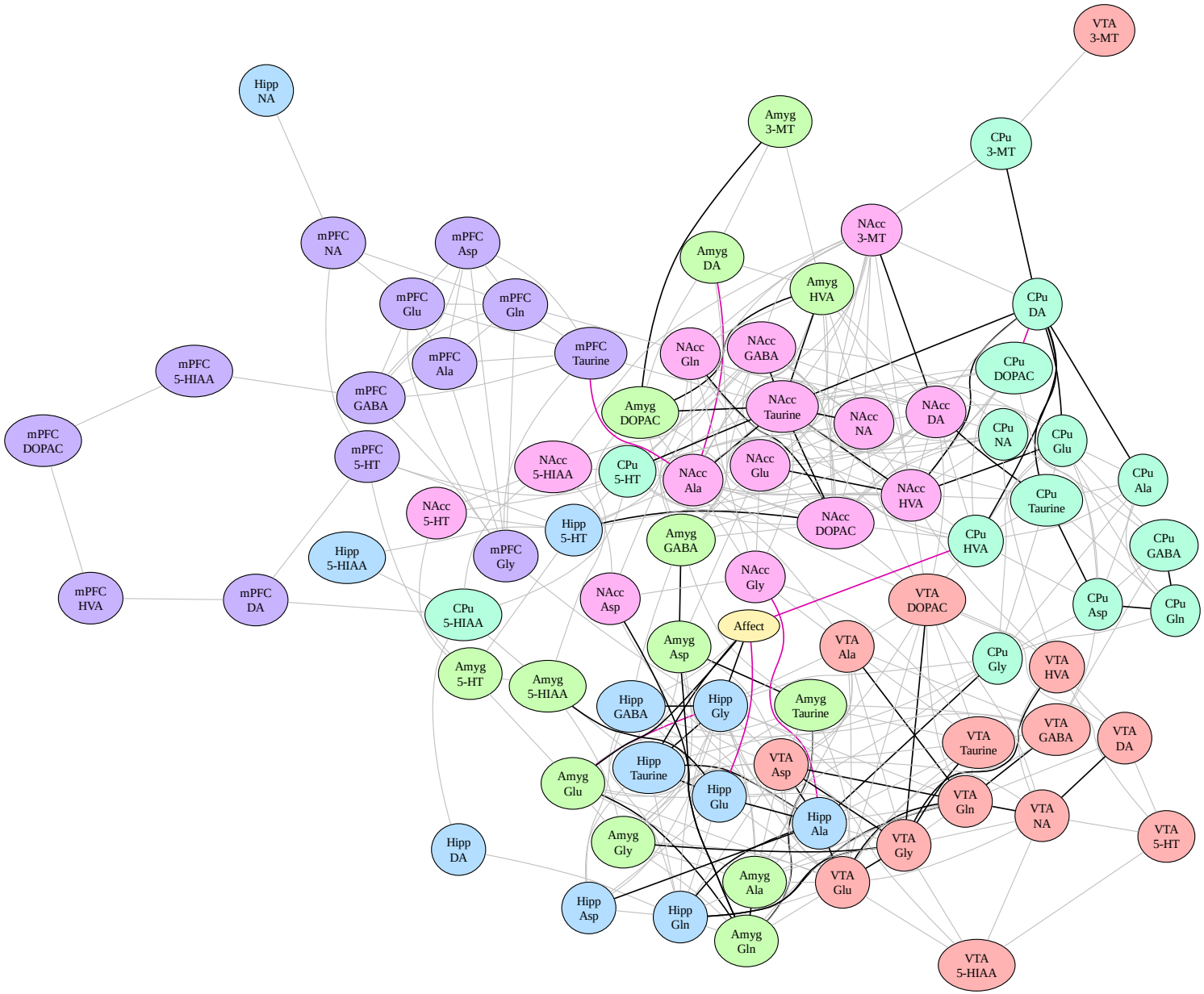
Boxes are path ends (*leaves*), while ellipses mark *relays*.

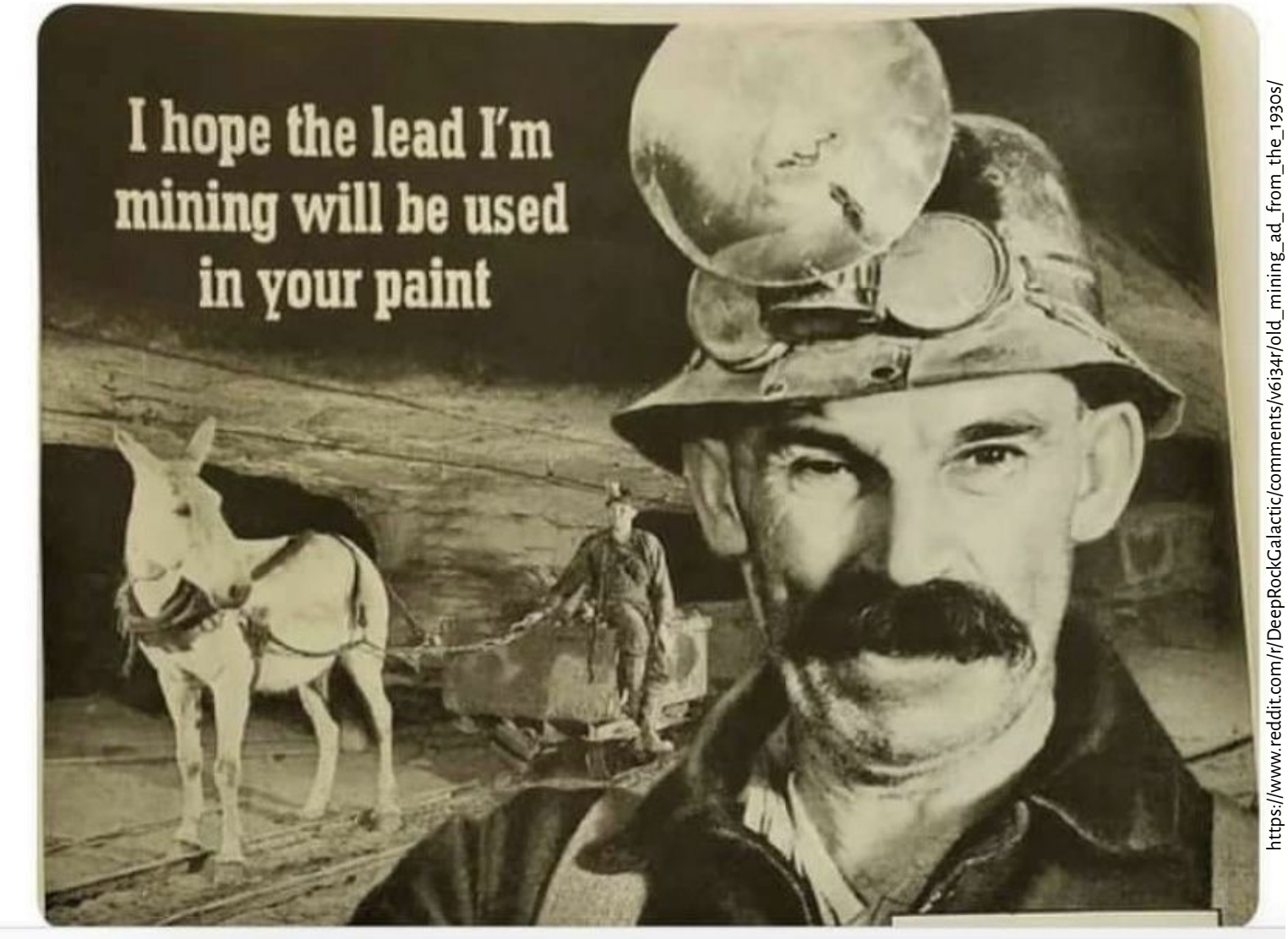
# *Affect* network

- Compilation of data from three papers.
- Rats from various experiments, 52 with induced affect, 53 not.
- Metabolomic study: 14 substances in 6 brain structures.


VTA 3-MT
Hipp NA
Amyg 3-MT
Hipp 5-HIAA
CPu 3-MT
Amyg DA
NAcc 5-HT
mPFC HVA
mPFC DA
mPFC 5-HT
NAcc 5-HIAA
NAcc 3-MT
Amyg HVA
CPu 5-HT
mPFC NA
Hipp 5-HT
NAcc GABA
CPu DA
CPu 5-HIAA
Amyg DOPAC
NAcc NA
mPFC DOPAC
NAcc Taurine
CPu DOPAC
mPFC Gln
NAcc Gln
NAcc DOPAC
NAcc DA
mPFC Glu
CPu NA
Amyg 5-HT
mPFC Taurine
NAcc Ala
NAcc HVA
Amyg GABA
mPFC GABA
NAcc Glu
CPu HVA
CPu Taurine
CPu Ala
mPFC 5-HIAA
mPFC Ala
Amyg 5-HIAA
NAcc Asp
NAcc Gly
CPu Glu
CPu GABA
mPFC Asp
mPFC Gly
VTA DOPAC
Amyg Asp
VTA Ala
CPu Asp
Amyg Glu
Amyg Taurine
VTA HVA
Hipp DA
Hipp GABA
CPu Gly
CPu Gln
Hipp Taurine
Hipp Gly
VTA Asp
VTA Taurine
VTA GABA
Amyg Gly
VTA DA
Hipp Glu
VTA Glu
VTA Gly
VTA NA
Affect
Amyg Ala
VTA Gln
VTA 5-HT
Hipp Gln
Amyg Gln
Hipp Ala
Hipp Asp
VTA 5-HIAA


Affect
Hipp Taurine
Hipp Glu
Hipp Gly
Amyg Glu
CPu HVA
Hipp Ala
Hipp Glu
Hipp GABA
Hipp Taurine
Amyg Gln
Hipp Gly
CPu DA
CPu Gly
Hipp Asp
Hipp Gln
Amyg 5-HIAA
VTA Asp
VTA Glu
NAcc Gly
Amyg Taurine
Amyg Asp
Amyg Ala
NAcc Asp
CPu Taurine
CPu 3-MT
CPu Glu
NAcc HVA
CPu Ala
NAcc Taurine
CPu DOPAC
VTA Gly
VTA Gln
VTA Gly
VTA Gln
VTA Gly
Amyg Asp
CPu Asp
NAcc DA
NAcc HVA
NAcc Taurine
NAcc DOPAC
VTA Ala
VTA NA
VTA Taurine
Amyg Gly
VTA GABA
VTA DOPAC
VTA HVA
Amyg GABA
CPu Gln
NAcc 3-MT
NAcc Glu
NAcc GABA
Amyg HVA
NAcc Ala
Amyg DOPAC
CPu 5-HT
NAcc NA
NAcc Gln
Hipp 5-HT
VTA DA
CPu GABA
Amyg DOPAC
Amyg 3-MT
mPFC Taurine
Amyg DA


VTA 3-MT
Hipp NA
Amyg 3-MT
CPu 3-MT
NAcc 3-MT
mPFC NA
mPFC Asp
Amyg DA
Amyg HVA
CPu DA
mPFC Glu
mPFC Gln
NAcc Gln
NAcc GABA
mPFC 5-HIAA
mPFC Taurine
CPu DOPAC
mPFC Ala
mPFC GABA
Amyg DOPAC
NAcc Taurine
NAcc NA
NAcc DA
CPu NA
CPu Glu
mPFC DOPAC
mPFC 5-HT
NAcc 5-HIAA
CPu 5-HT
NAcc Ala
NAcc Glu
CPu Ala
NAcc HVA
CPu Taurine
NAcc 5-HT
Hipp 5-HT
NAcc DOPAC
CPu HVA
CPu GABA
Hipp 5-HIAA
Amyg GABA
mPFC Gly
NAcc Gly
mPFC HVA
mPFC DA
NAcc Asp
VTA DOPAC
CPu Asp
CPu Gln
CPu 5-HIAA
Affect
Amyg Asp
VTA Ala
Amyg 5-HT
Amyg 5-HIAA
CPu Gly
VTA HVA
Hipp GABA
Hipp Gly
Amyg Taurine
VTA GABA
VTA Taurine
VTA DA
Hipp Taurine
VTA Asp
Amyg Glu
Hipp Glu
VTA Gln
VTA NA
Hipp Ala
Hipp DA
Amyg Gly
VTA Gly
VTA 5-HT
Amyg Ala
VTA Glu
Hipp Asp
Hipp Gln
Amyg Gln
VTA 5-HIAA




https://www.reddit.com/r/DeepRockGalactic/comments/v6i34r/old_mining_ad_from_the_1930s/

`gitlab.com/mbq/vistla` or `./praznik` for more information theory fun; both are also on CRAN.

# Supplement: Boruta

# Relevance

To recall, we have one distinguished feature $Y$ (decision) which anchors our interest in the system; the set of other features will be referred to as $X$.

The feature $X_i$ is *relevant* iff

$$\exists_{Q \subset X} I(X_i; Y|Q) > 0. \tag{25}$$

In particular, $X_i$ is *strongly* relevant iff additionally $I(X_i; Y|X/X_i) > 0$, and *weakly* relevant otherwise. Weak relevance means that there is some subset of $X$, called *spoiler*, that holds the same information as $X_i$.

Similarly, we can define *base*, a minimal $B \subset X$ for which $I(X_i; Y|B) > 0$. Features relevant for $B = \emptyset$ can be called *simple*, as they are relevant merely due to their *marginal* relevance $I(X_i; Y)$.

The (all) *relevant feature selection* is a selection which retains only the relevant features.

# Small redundant XOR toy data

| $A$ | $B$ | $N_1$ | $N_2$ | $N_3$ | $A \vee B$ | $A \wedge B$ | $\neg A$ | $Y = A \otimes B$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| | | | | | ... | | | |
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

It has an ambiguous solution, so *everything is redundant*, consequently *nothing is strongly relevant*.

# Relevant feature selection — problems

$$X_i \text{ relevant} \leftrightarrow \exists_{Q \subset X} I(X_i; Y|Q) > 0. \tag{26}$$

- The search space of $2^{|X|}$ is huge.
- The estimation of $I$ is hard, especially when joint states of $Q$ are numerous.
- Small MI estimates are overshot (*dark MI*), hence $> 0$ is too naïve.
- Relevance is not causal (as neither mathematical model) — this only means $X_i$ is useful for reconstructing $Y$.

We approached that with the *Boruta* algorithm.

# Boruta vs. huge search space

The search space of $2^{|X|}$ is huge.

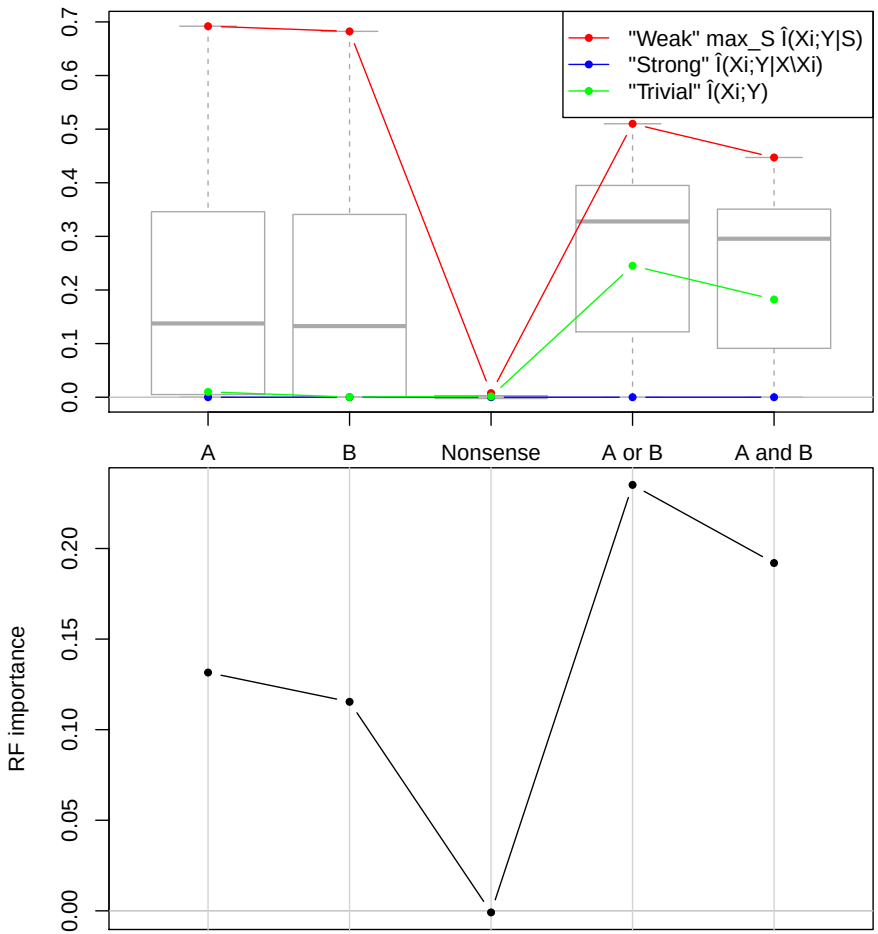
- It is not if we limit the interaction depth somehow.
- *Margin leaks* cause higher-order interactions to show up in lower-order scans, allowing for targeted optimisation: find strong margins first, and use them as bases for later search. In fact 'margin-tight' interactions are incredibly hard to engineer, and so are unlikely in practice.
- It may just make zero sense to look beyond certain depth, especially in $p \gg n$ cases — spurious interactions are just too likely.

# Boruta vs. CMI estimation

Estimation of $I$ is hard, especially when joint states of $Q$ are numerous.

- This is essentially the problem of what we can model — like it is nearly impossible to guess NC-PRNG form from its output, but while we know its form, it is usually trivial to fit original parameters. We must rely on heuristics, and there is no universally optimal one.
- Limiting considered interaction depth limits $Q$'s worst-case dimensionality.
- Estimation for continuous data is basically impossible.

# CMI & Random Forest




Random Forest importance

$$J_{\mathrm{rf}}(X_i) \approx \langle I(X_i; Y|Q) \rangle_{Q \in 2^{X-i}} , \qquad (27)$$

which can be shown asymptotically (Louppe 2014).

Tree is an implementation of *find strong margin, use it as a base for next scan* idea.

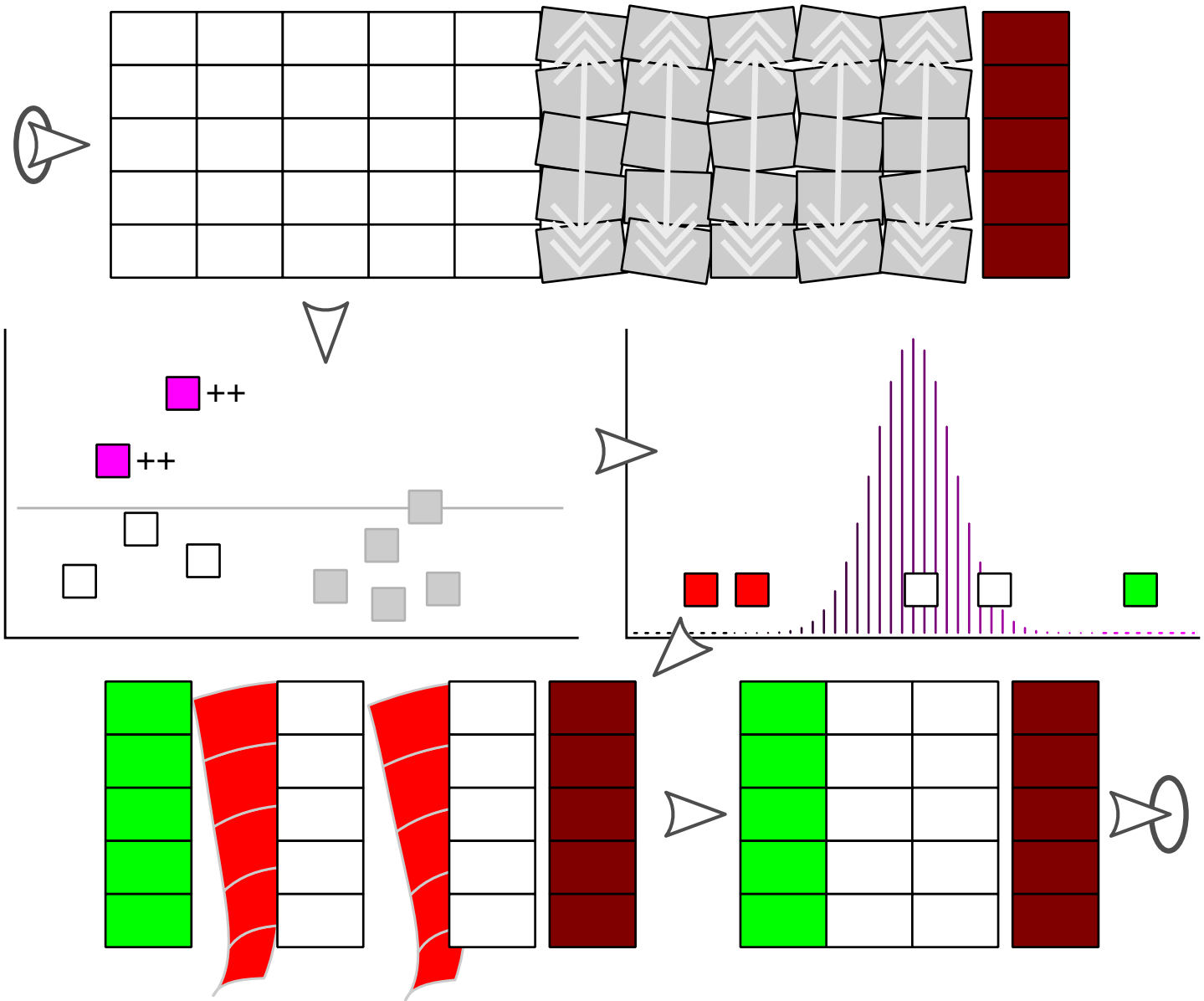
RF also handles 'reasonable' mapping of continuous features (slight biases my apply).

# Boruta vs. dark MI

MI estimates are overshot, hence $> 0$ is too naïve.

- We can estimate the baseline MI by bootstrap.
- They have to be mixed with true features to properly cover interaction.
- They have to preserve actual distribution within the information system.
  - . . . but this chicken & egg situation, as this is a solution to what-is-relevant problem.
  - We can somewhat settle on marginal distributions (which can be achieved by mixing).

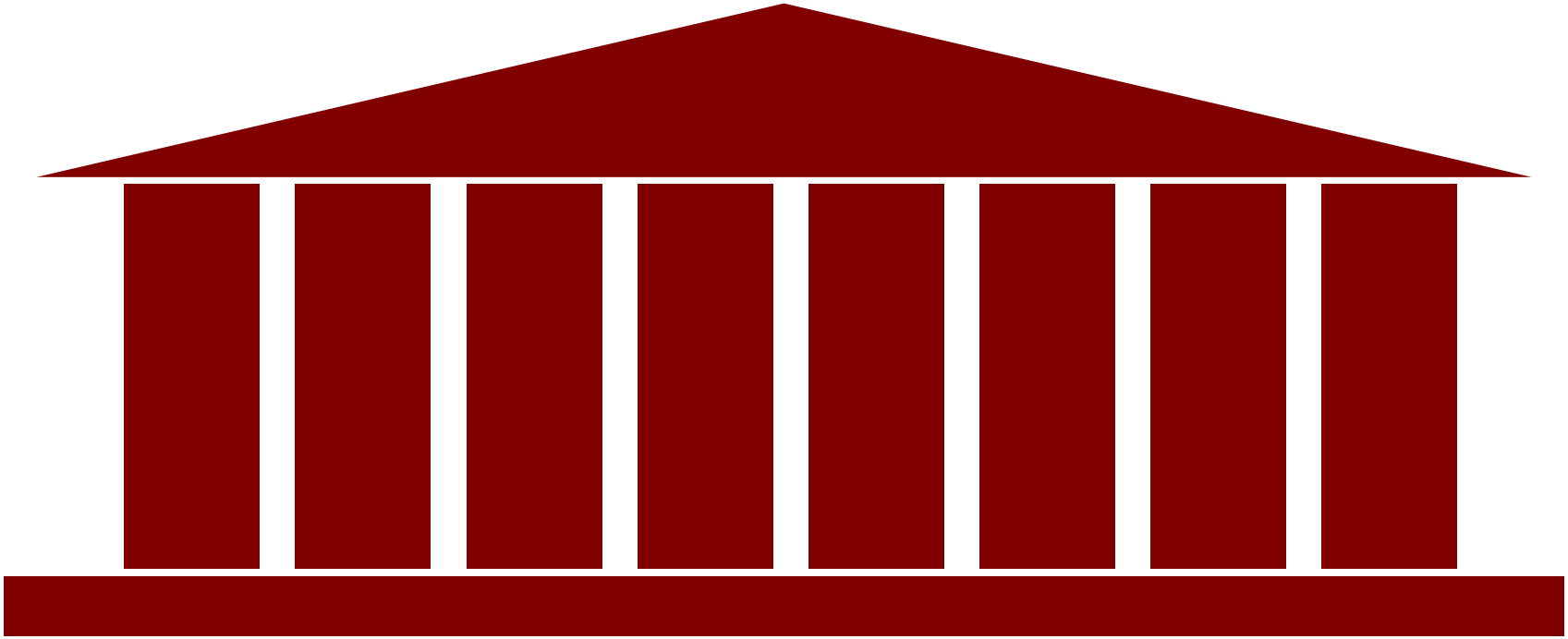
# Boruta



# Optimal feature selection

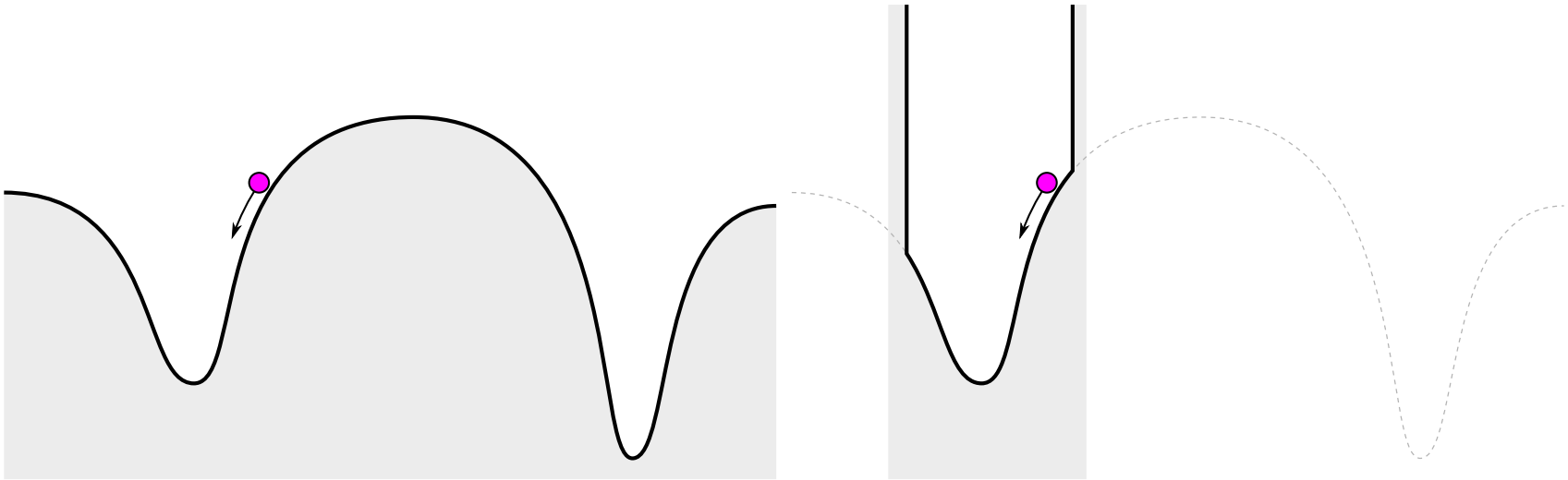
- The conventional knowledge is that feature selection should be done to optimise model accuracy.
- In fact, the ideal Bayes classifier requires only strongly relevant features (Nilsson 2007).
  . . . yet only when it exists, i.e., is unique.
- In $p \gg n$ setup, typical to omics, the prediction is usually overdetermined — there are several possible options to build an equally good classifier. Consequently, the problem of optimal selection is underdetermined, like in the SRX toy problem.
- What's worse, random noise may be sometimes assembled in a model even better than real signals.

# Ambivalent redundancy dilemma



Redundant elements can be safely removed, right?

# Constrained error optimisation



After wrong selection, local or spurious optimum behaves like a global one.