Analiza danych o wielkim wymiarze: eksploracja czy wnioskowanie, słowo o iluzjach neodarwinizmu

Jacek Koronacki

Computational Biology Lab,
Institute of Computer Science,
Polish Academy of Sciences, Poland

http://zbo.ipipan.waw.pl

Wprowadzenie …
za pomocą przykładu

Jest rok 1854. W Soho wybucha epidemia cholery. I oto jak John Snow, 120 lat przed ukazaniem się pomnikowego dzieła Johna Tukeya *Exploratory Data Analysis*, zastosował tę analizę (odtąd w skrócie EDA) do zahamowania epidemii:

(z Jamesa R. Thompsona wspomnienia o J. W. Tukeyu)

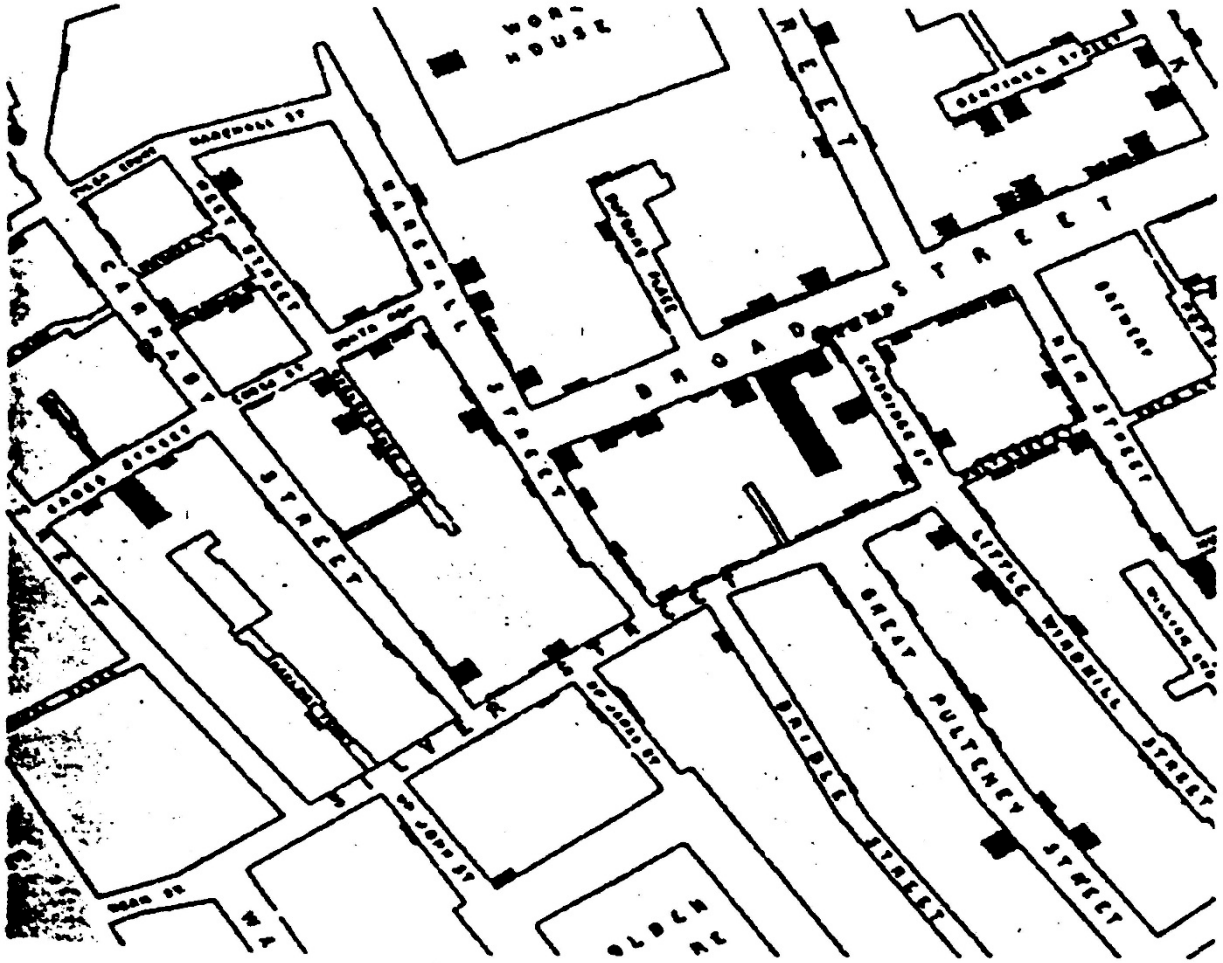


Figure 1. Vicinity of the "Broad Street Pump."

Wszakże zapobieżenie przyszłym epidemiom wymagało wytłumaczenia ich przyczyny! I około 1860 roku Ludwik Pasteur zrozumiał, iż przyczyna tkwi w zarazkach, na przykład przenoszonych przez wodę wypompowywaną spod ziemi w Soho …

Nauka stoi szukaniem związków przyczynowo skutkowych.

Ale szalenie to ciekawe – a może typowe (we wczesnej Grecji i począwszy od Średniowiecza) – że owych niespełna 160 lat temu odnalezienie przyczyny zapadalności na cholerę zostało poprzedzone eksploracyjną analizą danych.

Nominalizm czy realizm w nauce

Około 1975 roku dobiegł końca wiek sir Ronalda Fishera w analizie danych i statystyce:

Rozumowanie biegło dotąd od szczegółu do ogółu przez analizowanie danych **w świetle założeń o rozkładzie prawdopodobieństwa**, z którego dane pochodzą (np. w świetle założenia o normalności rozkładu).

Fundamentem było testowanie hipotezy o modelu w świetle obserwowanych danych i ewentualne modyfikowanie modelu.

Było do podejście, którego punktem wyjścia była hipoteza (**a hypothesis-driven approach**).

Celem było zrozumienie i wyjaśnienie badanego zjawiska.

Było to podejście realistyczne w rozumieniu arystotelesowskim i tomistycznym.

Ostatnie ćwierćwiecze XX wieku przynosi zachwyt statystyków (analityków danych) nad metodami eksploracyjnej analizy danych (EDA).

Coraz większe ilości zbieranych danych skłaniają do odejścia od arystotelosowsko-tomistycznego realizmu na rzecz **ockhamowskiego nominalizmu** radykalnych pragmatystów, którzy całą wiedzę widzą w danych.

Wybitny pionier EDA, John Tukey napisał: ***Pozwólmy przemówić danym – niech same mówią za siebie***.

W 1979 roku jego nie tak słynny, za to bardziej radykalny i też znakomity statystyk, William Eddy napisał:

The data analytic method denies the existence of "truth"; the only knowledge is empirical.

[...] The only purpose of models is to make formal implications. For far too long statisticians have concentrated on fitting models to data.

[...] If we can make without models, I think we should.

Radykalni pragmatyści byli chyba po części – zresztą słusznie – sfrustrowani modą na budowanie zbyt uproszczonych modeli.

I jak mieliśmy się wszyscy (prawie wszyscy?) analitycy danych nie stać się radykalnymi pragmatystami wobec obecnego zalewu danych (przekształcając EDA w DM, czyli *data mining*)?!

Tym bardziej, że nie chodzi tylko o rozwój technologiczny, gospodarczy, ale o nasze bezpieczeństwo (terroryzm, cyberwojny!) i zdrowie (genomika, proteomika, …)

Ale przecież dzisiejsze moce obliczeniowe pozwalają tych zbytnich uproszczeń uniknąć, przynajmniej często, czy choćby niekiedy.

No i jeśli np. rzecz nie w zapobieżeniu atakowi terrorystycznemu, lecz dokonaniu odkrycia naukowego, to jak nie wrócić do paradygmatów filozoficznego realizmu?!

Nie zgadzając się na bycie inteligentnymi zwierzętami, które nie potrzebują wyjaśnień, lecz tylko umiejętności dobrego przewidywania –

jak to postulują np. V. Mayer-Sch□onberger i K. Cukier w skądinąd interesującej książce pt. *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (John Murray (Publishers), 2013; tłum polskie 2014)

To, czego potrzebujemy i co realizujemy, to **powrót do naukowego realizmu via … matematyka poglądowa i**:

podejście dwu- lub wieloetapowe,
wychodzące od analizy dostępnych danych

(a nie od hipotezy, której prawdziwość chcemy potwierdzić),

by w kolejnych etapach
przeprowadzić wnioskowanie przyczynowo-skutkowe

i tą drogą dokonać naukowego odkrycia

(two- or multistage data-driven causal reasoning)

choć może przed nami „one-stage data-driven causal reasoning”; p. uwagi w zakończeniu i tam GPT-2.

Model-Free Mutual Information-Based Feature Selection and Interdependency Discovery in Classification

Niedawno temu zadanie takie zostało elegancko postawione i rozwiązane – jakkolwiek czekamy na jego dalszy rozwój –

przez

Jana Mielniczuka i Pawła Teisseyre w pracy

Stopping rules for mutual information-based feature Selection (*Neurocomputing 358, 255-274, 2019*)

W pracy oparto się na warunkowej informacji wzajemnej oraz informacji interakcyjnej (interaction information).

Informacja wzajemna (gdzie X_1, X_2 – są dyskretnymi zmiennymi losowymi, dalej zwanymi cechami, o wartościach x_i , x_j):

$I(X_1, X_2) := \sum_{i,j} p(x_i, xj) \log p(x_i, x_j) / (p(x_i) p(x_j)) =$

$KL(P_{X1,X2} || P_{X1} \times P_{X2})$

Informacja interakcyjna:

$II(X_1, X_2, Y) = I((X_1, X_2), Y) - I(X_1, Y) - I(X_2, Y)$

X_i , Y mogą być wielowymiarowe!

Cel:

Znaleźć taki podzbiór cech o ustalonej liczności $1 \leq k \leq p$, który maksymalizuje łączną informację wzajemną ze zmienną Y (klasyfikującą),

$$\arg\max_{S:|S|=k} I(X_S, Y),$$

gdzie X_S oznacza podzbiór cech $X_1, \ldots, X_p$, indeksowanych przez zbiór S z elementami ze zbioru $\{1, \ldots, p\}$.

Sekwencyjne poszukiwanie w przód: Przyjmijmy, że S jest zbiorem już wybranych cech, S^c jest jego dopełnieniem oraz X_j z S^c jest cechą kandydującą do zbioru S,

$$\arg\max_j [I(X_{S\cup\{j\}}, Y) - I(X_S, Y)] = \arg\max_j I(X_j, Y \mid X_S) =$$
$$\arg\max_j (I(X_j, Y) + II(X_j, X_S, Y))$$

Reprezentacja Möbiusa informacji interakcyjnej podpowiada łatwe do policzenia przybliżenie tej wielkości:

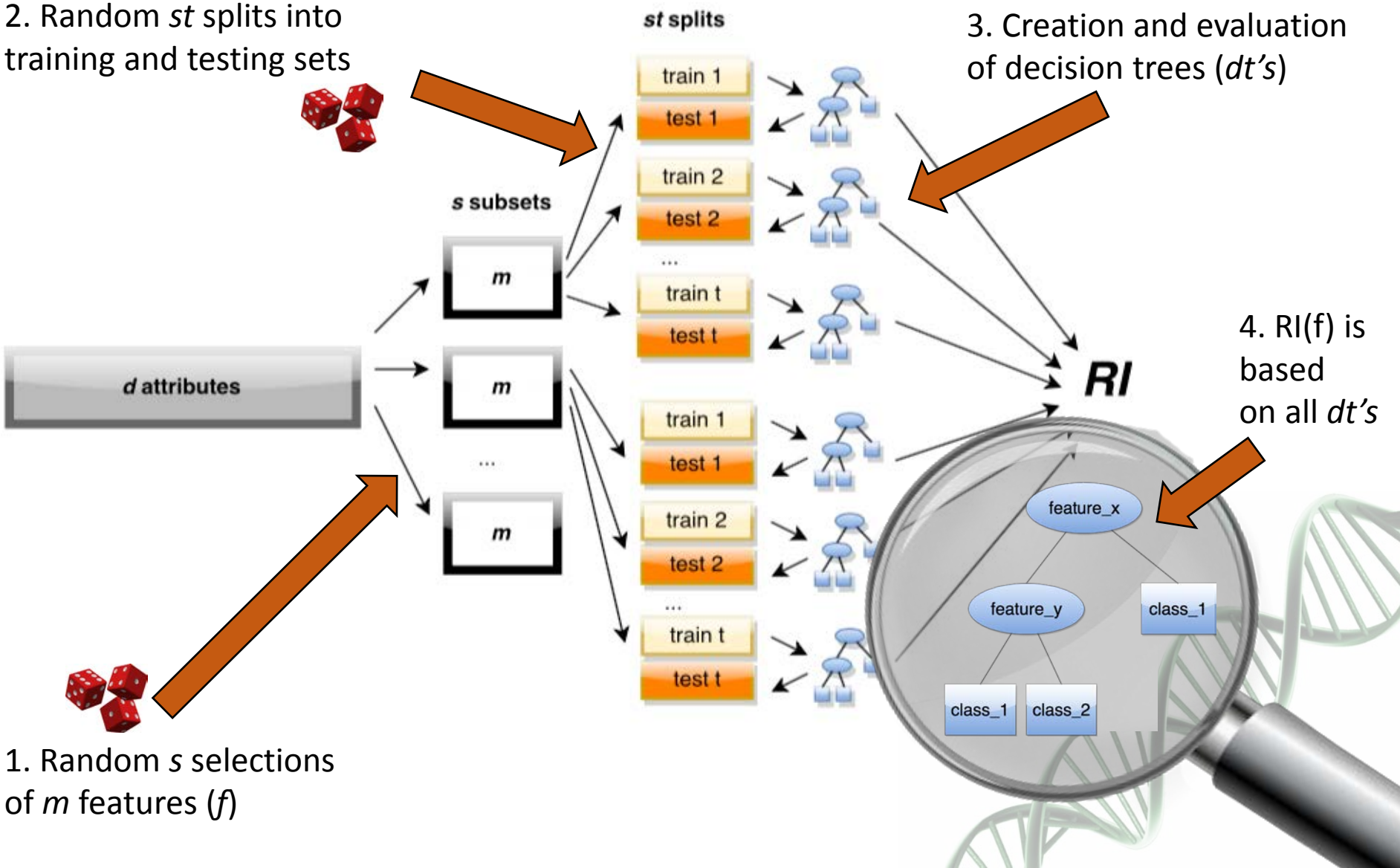
$$\arg\max_j (I(X_j, Y) + \sum_{i \varepsilon S} II(X_i, X_j, Y)).$$

Monte Carlo Feature Selection and Interdependency Discovery in Classification

(The MCFS-ID Algorithm)

Patrz *JSS*, vol. 85, issue 12, 2018

The Algorithm (MCFS-ID)



Algorytm (bez żadnych założeń o modelu):

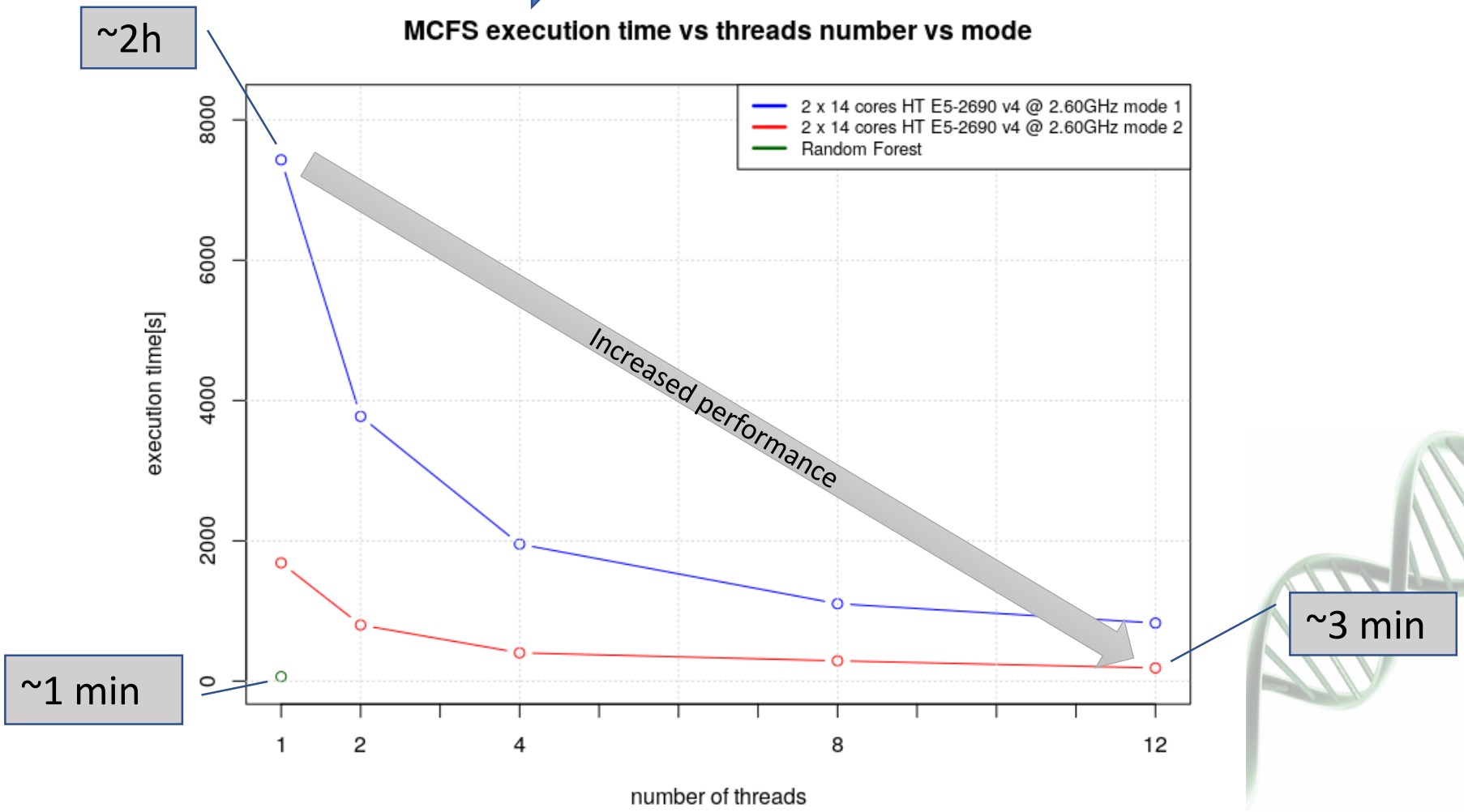
- Dokonuje rankingu cech ze względu na ich ważność dla problemu klasyfikacji (problemu orzekania o przynależności obserwacji do jednej z możliwych klas)

- Odcina cechy niosące informację o interesującym nas problemie od cech tej informacji nie niosących
  - W przypadku problemów o wielkim wymiarze wektora obserwacji algorytm działa dwuetapowo – najpierw usuwa cechy „na pewno” nie niosące żadnej informacji o problemie
  - I dopiero w drugim etapie dokonuje dokładnego rankingu

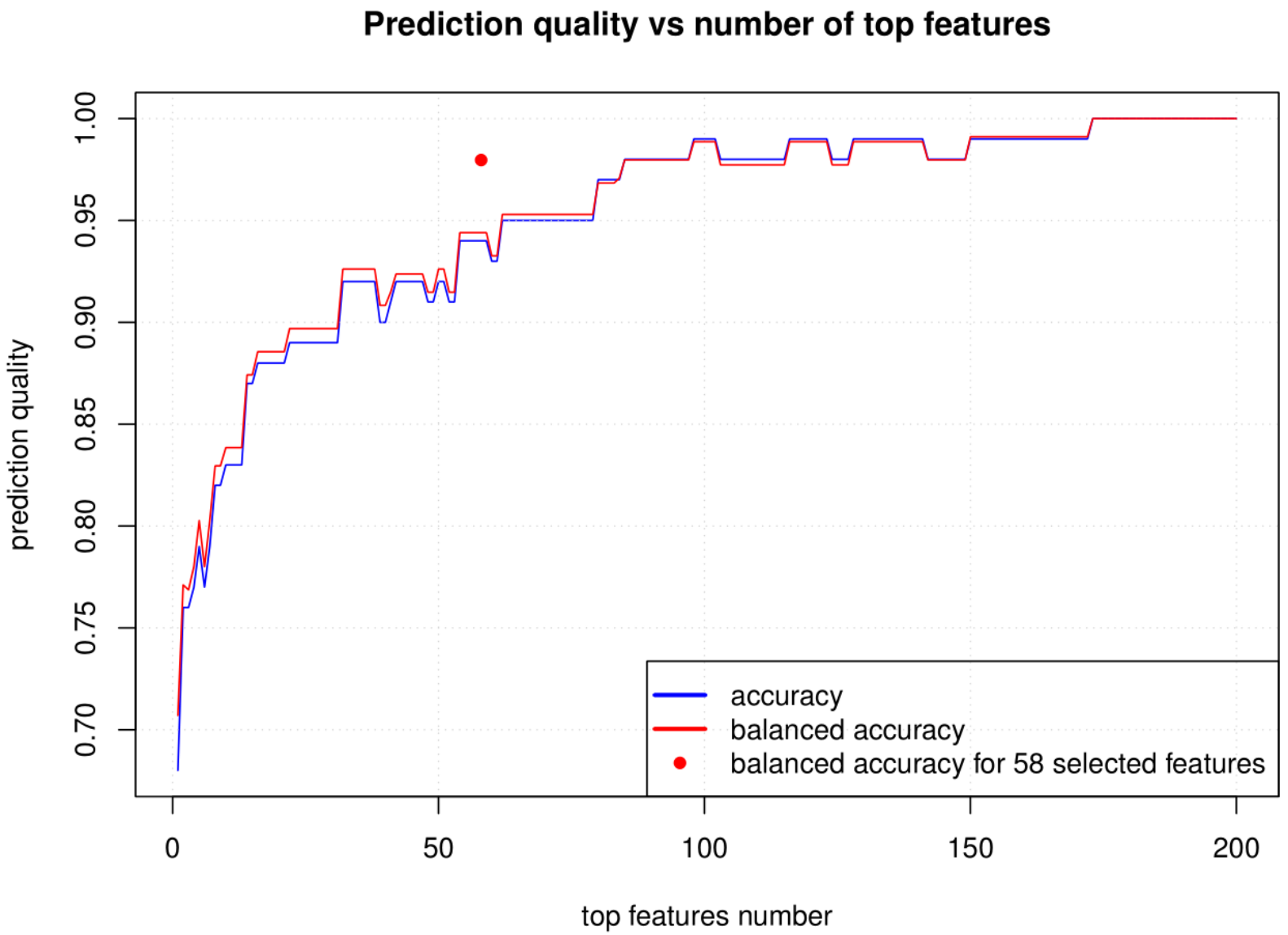
- Buduje graf współzależności między cechami (ich interakcji) w ich wspólnym orzekaniu o wartości zmiennej przypisującej obserwacje do klas, jednocześnie mierząc siłę owej współzależności między parami cech

Performance analysis

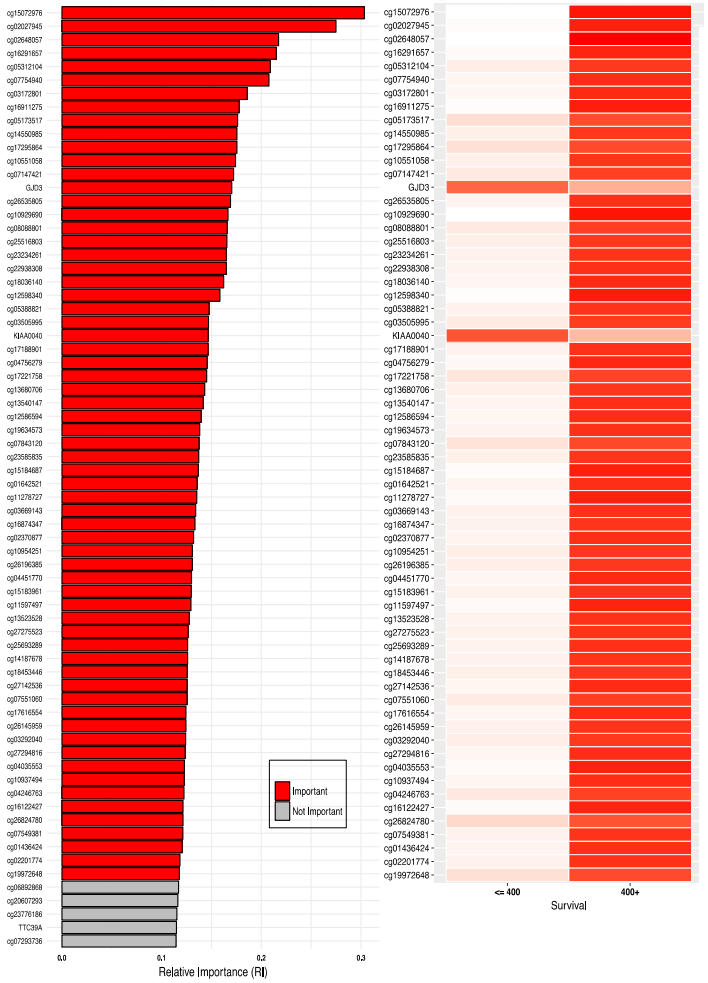
- TCGA (88 x 416 013) → (88 x 32 458) → 6.3h
- Arcene (100 x 10 000) → (100 x 2 169) → 187 sec (511 300 trees)

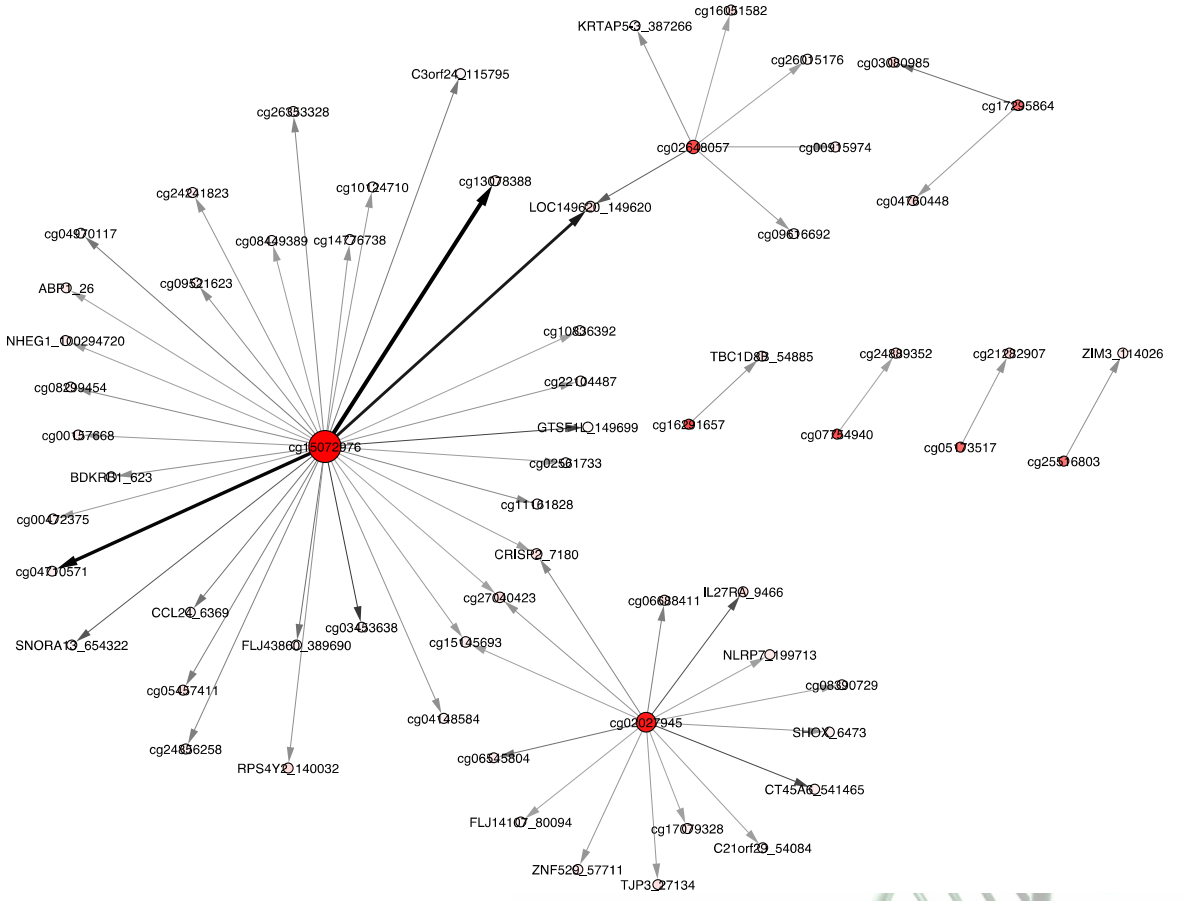


ID-Graph in prediction (Arcene dataset)



Glioma Patients Survival

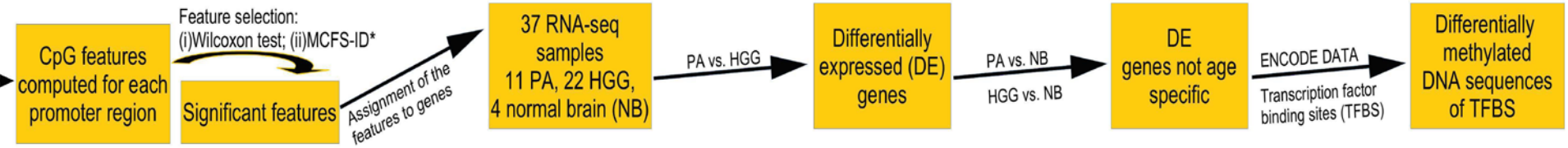




Dabrowski M.J., Draminski M., Diamanti K., …, Komorowski J., Kaminska B. & Wojtas B. **(2018)**. Unveiling new interdependencies between significant DNA methylation sites, gene expression profiles and glioma patients survival. Scientific Reports vol. 8, Article number: 4390, doi:10.1038/s41598-018-22829-1.

Multi-stage data-driven causal reasoning

By means of example*: Genome-wide mapping of DNA methylation variants affecting gene expression levels in gliomas with respect to their grade and IDH1 gene mutation status*, M. J. Dąbrowski et al., an ongoing study:



The clear functional division of the selected genes, allowed by DNA methylation analysis of TFBS, indicates their potential biological importance in pathways of gliomagenesis.

For some details, see
http://zbo.ipipan.waw.pl/files/papers/CSH_2019/Genome2019AbstractBook_p69.pdf and
http://zbo.ipipan.waw.pl/files/papers/CSH_2019/Poster2019-05-15.pdf

… Co uzasadnia kilka zdań o neodarwinizmie

(z przywołaniem dzieła Denisa Noble’a)

Kilka zdań z prac Denisa Noble'a

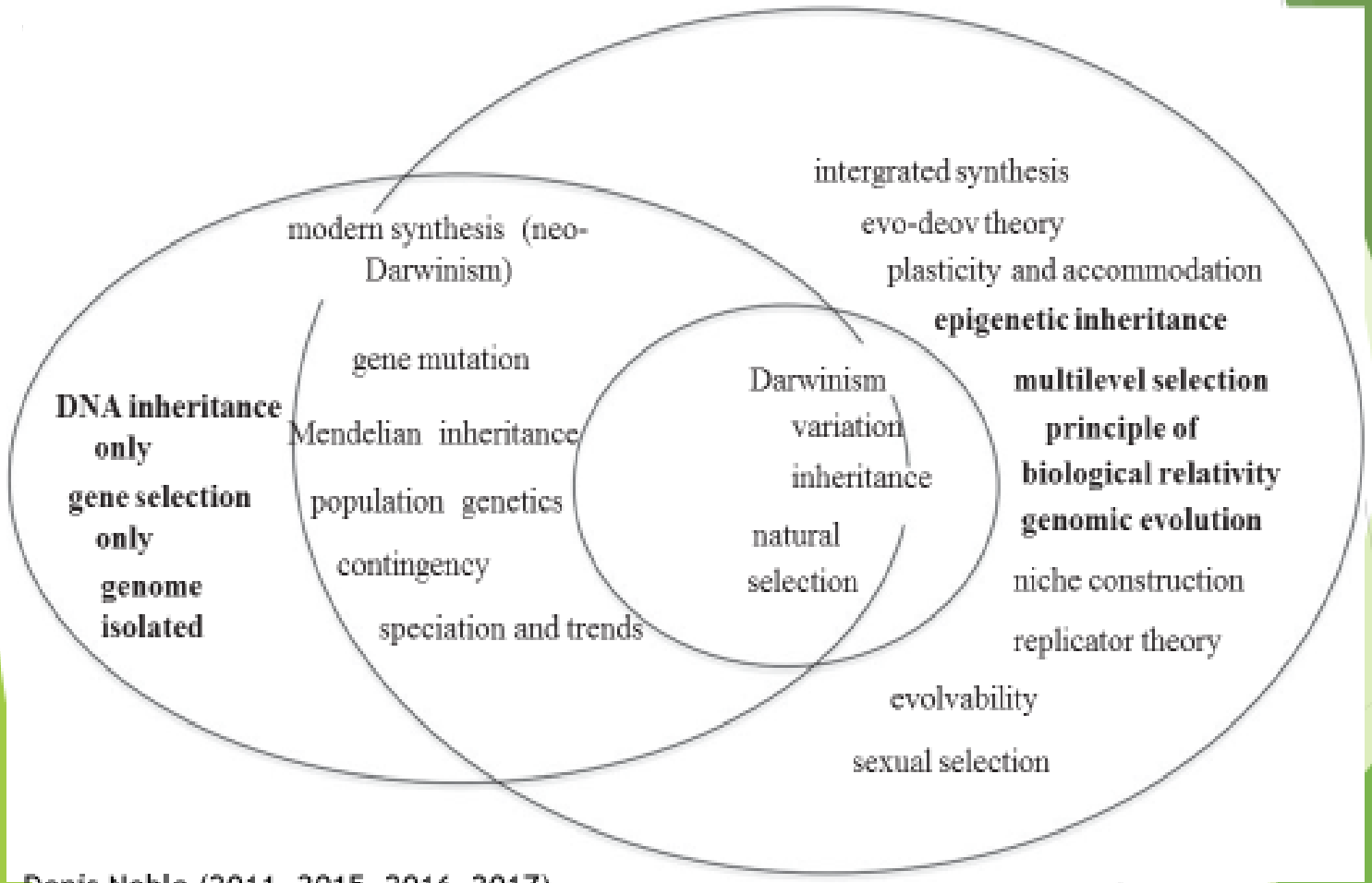
Sequence info passes one way, from DNA to protein.

Yet, much another info passes from the organism to the genome. It must be so to produce many different patterns of gene expression, which enable many different phenotypes (e.g. many different cell types in the same body) to be generated from the same genome. In addition to controlling relative expression levels, the organism also makes use of protein-mediated protein processing to add yet another layer of control following transcription.

This info is conveyed by patterns of transcription factors, genome marking, histone marking and many RNAs, which in turn control the patterns of gene expression. These controls are exerted through preferential targeted binding to the genome or histone proteins.

Moreover, non-DNA information can be transmitted across generations.

Podsumowanie Denisa Noble'a



Denis Noble (2011, 2015, 2016, 2017)

Zamiast zakończenia jeszcze słowo o starym i nowym w modelowaniu, czyli …

… jeszcze o współczesnych drogach do modelu

- Stare nie umiera nigdy, choć przyjmuje nową postać. Jak w LTCM, czy lepszym Reneaissance Technologies i Medallion Fund Jamesa Simonsa.

[Tytułem wtrętu kilka ciekawych nazwisk i nazw: Robert i Rebekah Mercerowie, Peter Thiel, Palantir Technologies, Steve Bannon, Breitbart, Cambridge Analytica.]

- Nowe zapowiada generyczne modelowanie nierozłącznie splecione z uczeniem się z danych. Jak GPT-2 (z udziałem tzw. głębokiego uczenia się).

GPT-2

Human-written context:
In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

I GPT-2 na to:
Dr. Jorge P´erez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans.

[…] P´erez and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said P´erez. (To be continued)

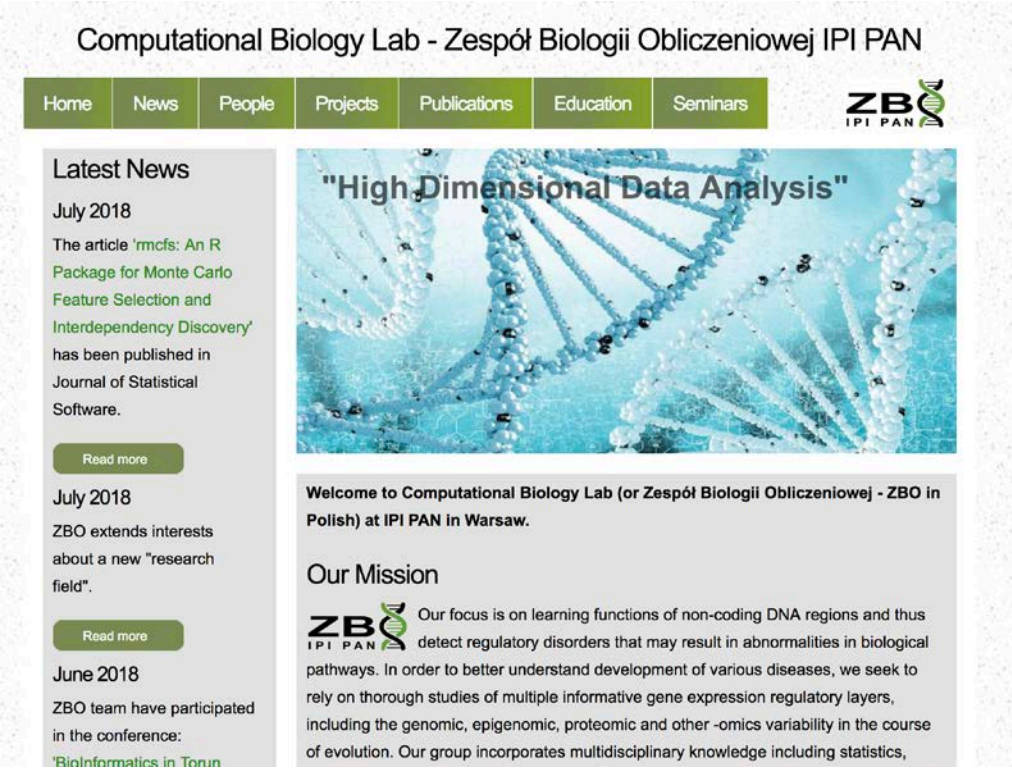
GPT-2, contd:

P´erez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. P´erez stated, “We can see, for example, that they have a common ‘language,’ something like a dialect or dialectic.”

Dr. P´erez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America. […]

Bardzo dziękuję za cierpliwość



http://zbo.ipipan.waw.pl

I jeszcze aneks,
ale już nie podczas wykładu

MCFS-ID – Past and Present

- Dramiński M., Koronacki J. rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery. Journal of Statistical Software vol. 85(12), doi:10.18637/jss.v085.i12, **2018.**
- Koronacki, J., Dramiński, M. Empirical Model Building Revisited. Models and Reality: Festschrift for James Robert Thompson, Chicago, IL: T&NO Company, **2017**.
- M.Dramiński, M.J.Dabrowski, K.Diamanti, J.Koronacki, J. Komorowski, "Discovering networks of interdependent features in high-dimensional problems" in "Big Data Analysis: New Algorithms for a New Society", eds. Nathalie Japkowicz and Jerzy Stefanowski, Studies in Big Data, ISSN 2197-6503, **2015**.
- [···]
- M.Draminski, A.Rada-Iglesias, S.Enroth, C.Wadelius, J. Koronacki, J.Komorowski "Monte Carlo feature selection for supervised classification", BIOINFORMATICS 24(1): 110-117, **2008**.
- [...]
- M.Draminski, J. Koronacki, J.Komorowski "A study on Monte Carlo Gene Screening", Proceedings of the New Trends in Intelligent Information Processing and Web Mining IIS'2005 Symposium, Gdansk, Poland, Springer-Verlag, **2005.**

Relative Importance (MCFS-ID)

Important attribute:

Occurs in many decision trees (DT)

Is located nearby the root & separates many objects

The relative importance of feature g_k, RI_{g_k}, is defined as

$$\mathrm{RI}_{g_k} = \sum_{\tau=1}^{s \cdot t} \mathrm{wAcc}_{\tau}^{u} \sum_{n_{g_k}(\tau)} \mathrm{IG}(n_{g_k}(\tau)) \left(\frac{\text{no. in } n_{g_k}(\tau)}{\text{no. in } \tau} \right)^{v},$$

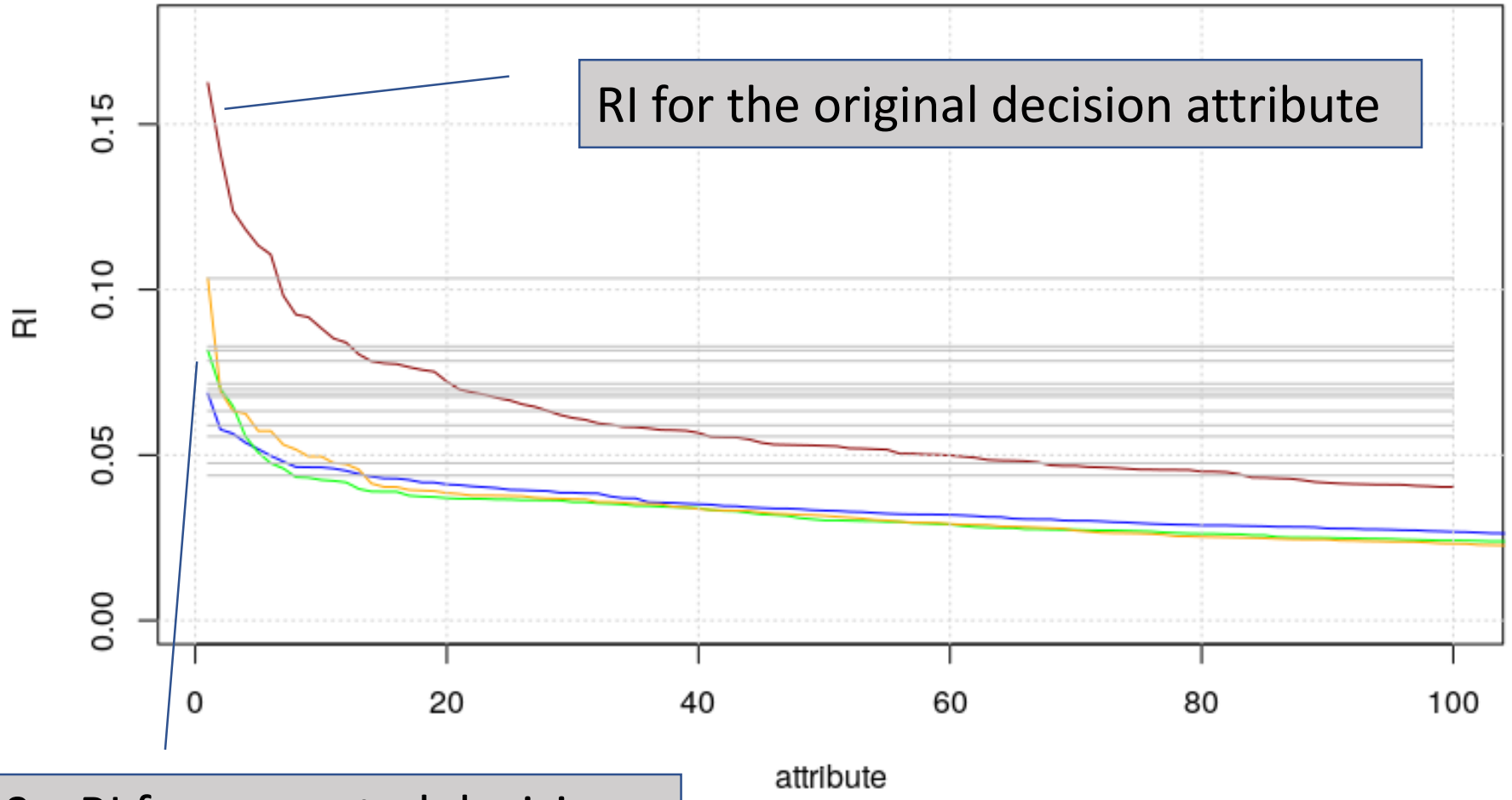
DTs based on it perform well on unseen data

Separates classes within the node with high quality

Cut-off point in MCFS-ID

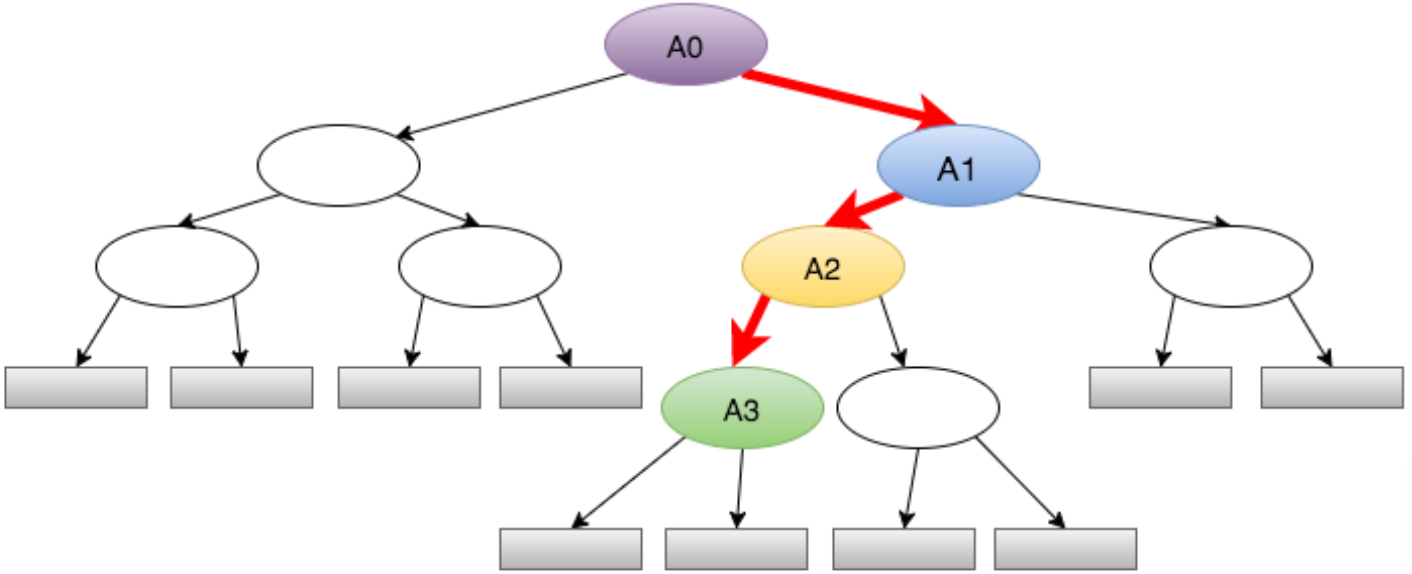
Cut Off – Permutations (MCFS-ID)

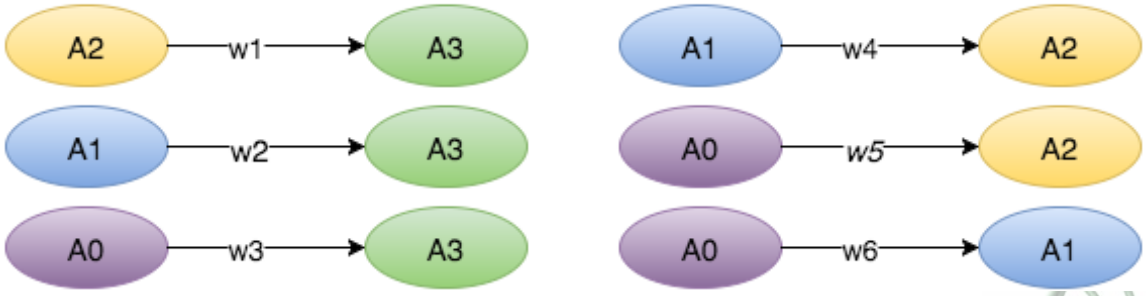
Permute x times the decision attribute and run mcfs. Compare all maximal RIs obtained from such experiments with the reference experiment (with original decision attribute).



ID Graphs (MCFS-ID)

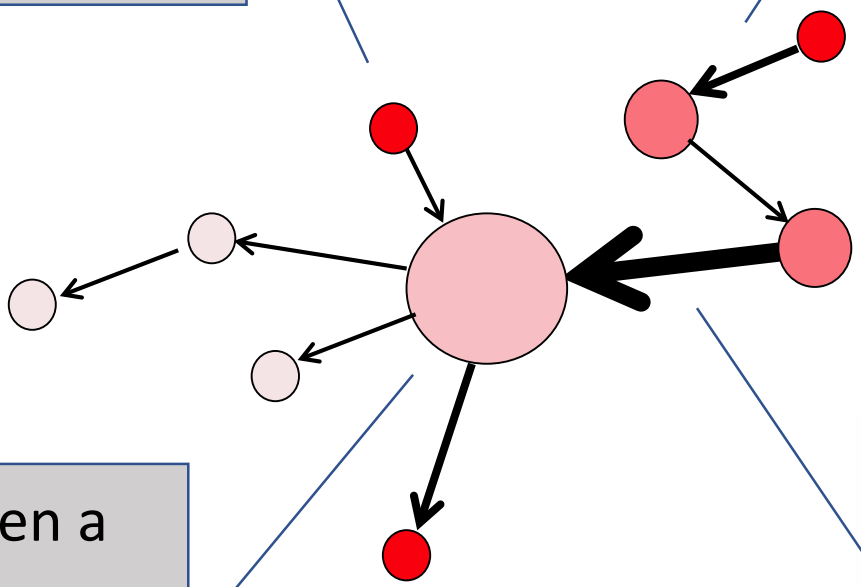
We assume that when two features co-occur along the same path in a decision tree, they are interdependent





ID Graph - Plot Rules

The higher RI of a feature, the stronger its color intensity

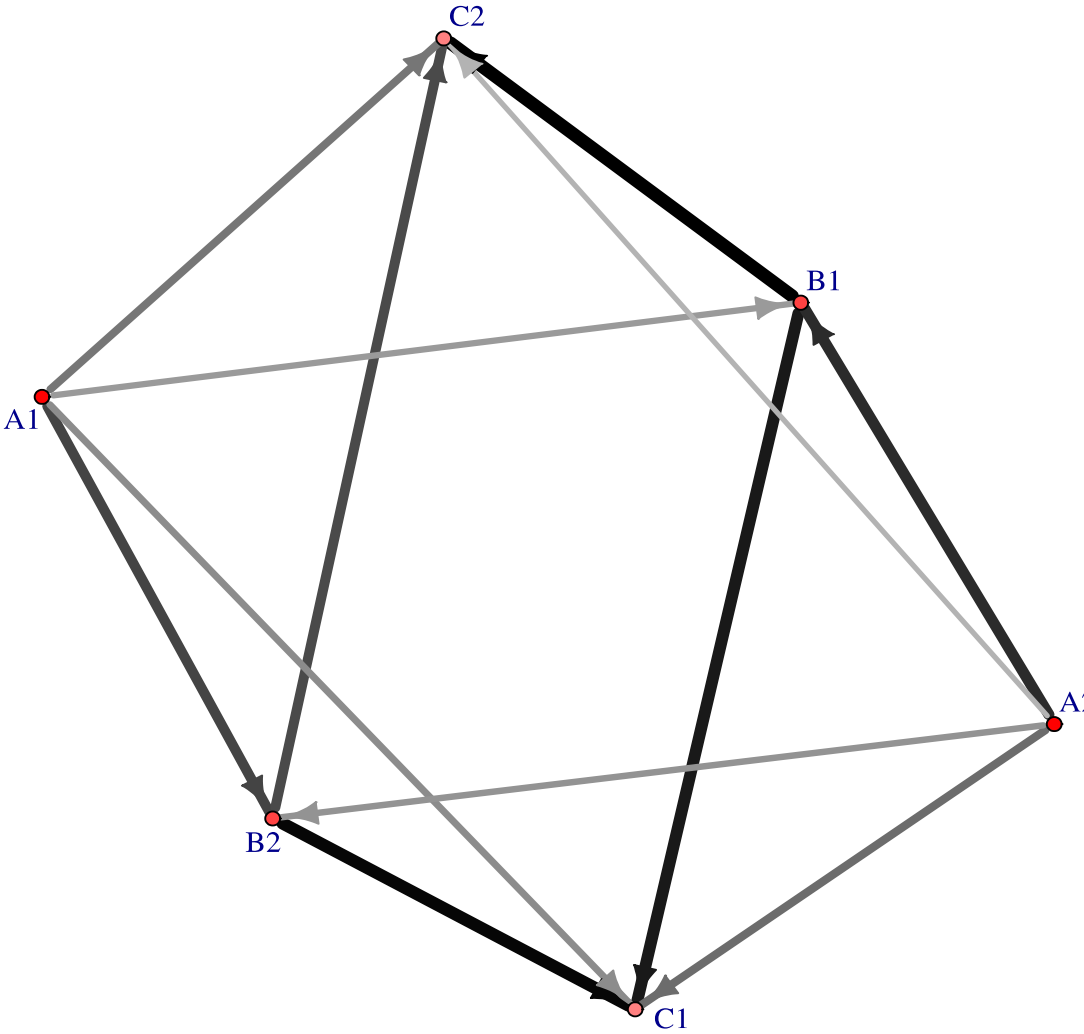


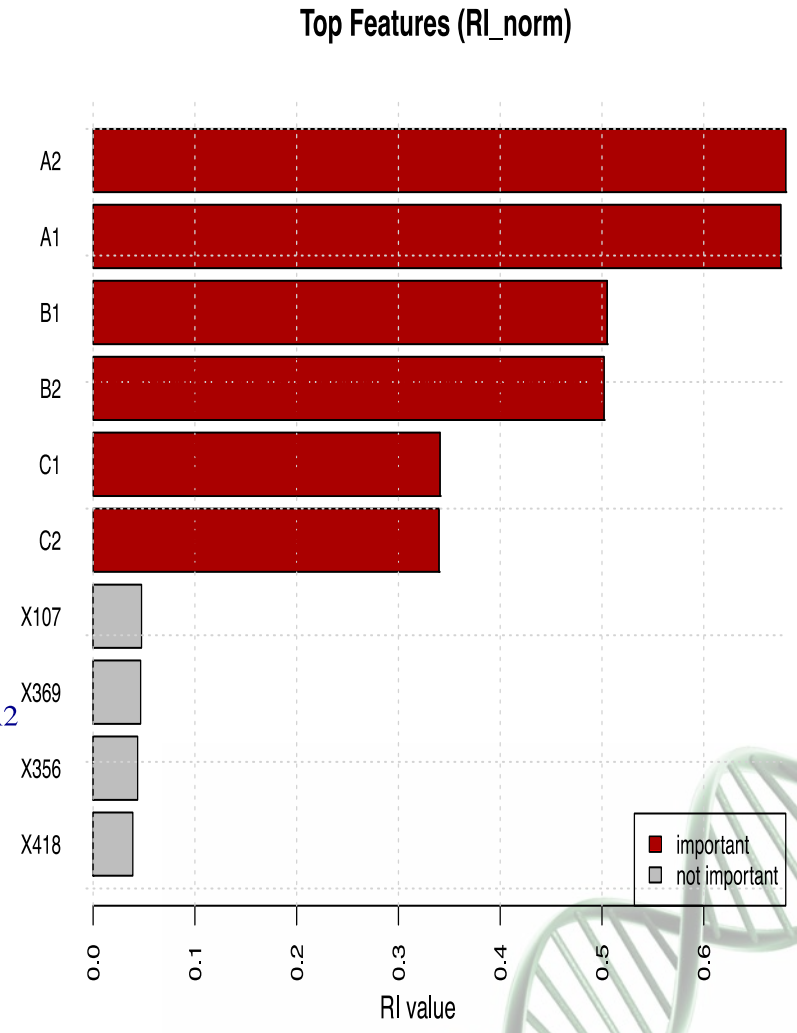
Direction of the links follows the path in the tree: from ancestors to their children

The more often a feature co-occurs with others, the larger the size of the node

The larger the weight of an edge, the thicker the connection is

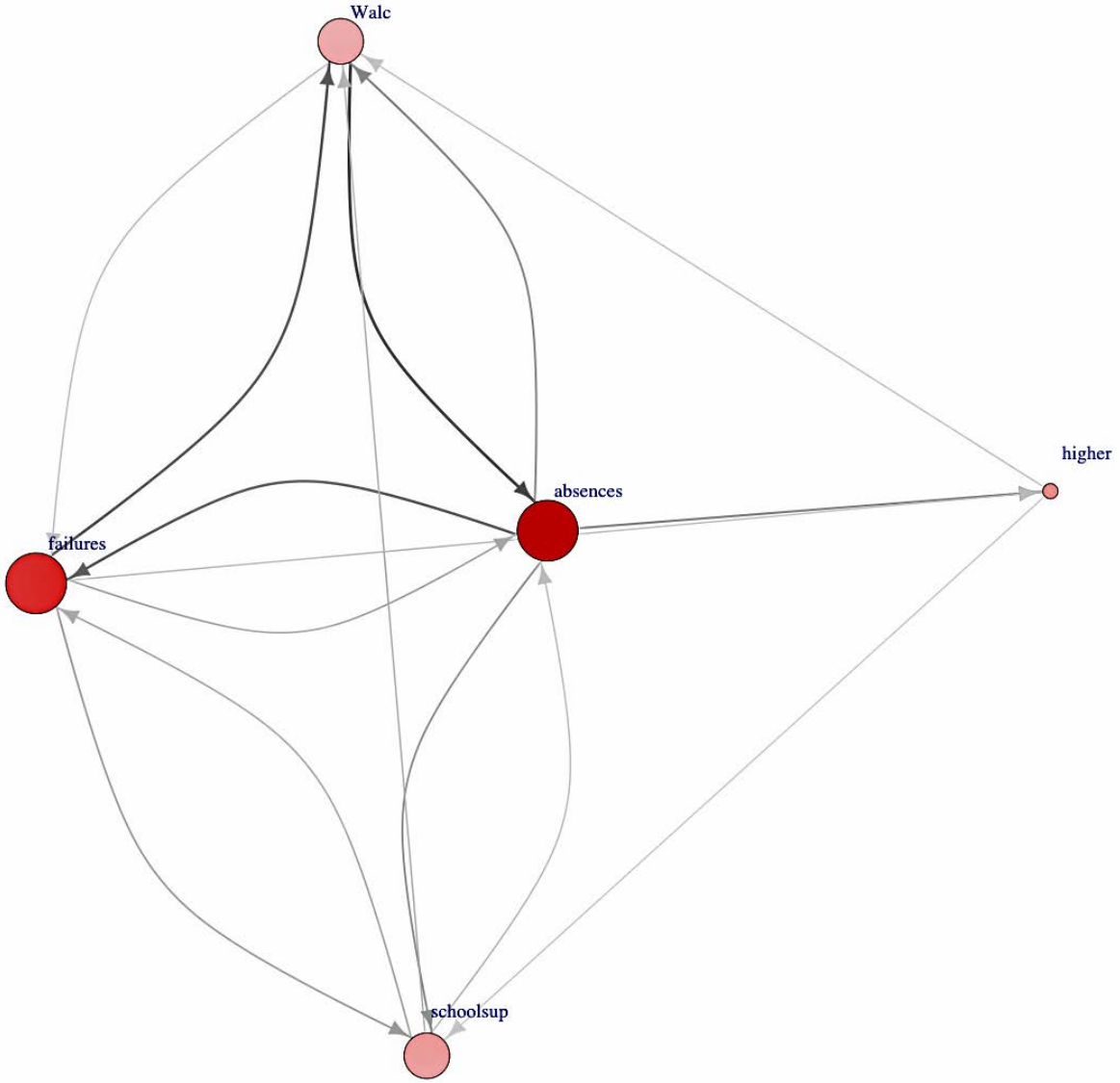
Artificial Data

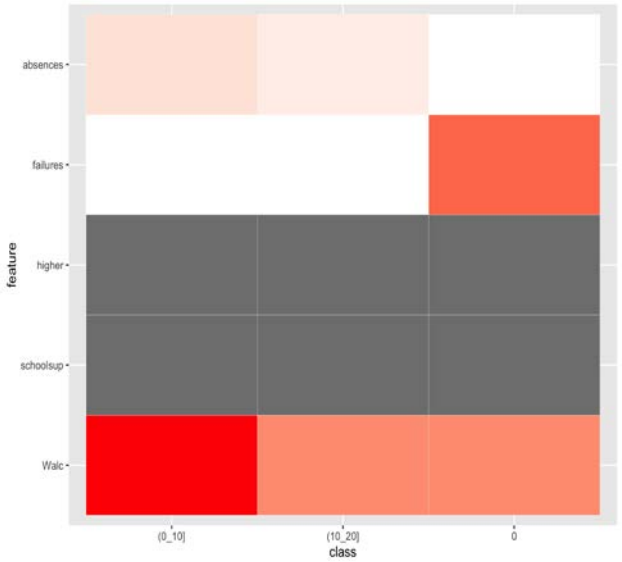




Dramiński M., Koronacki J. (**2018**) rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery. Journal of Statistical Software vol. 85(12), doi:10.18637/jss.v085.i12.

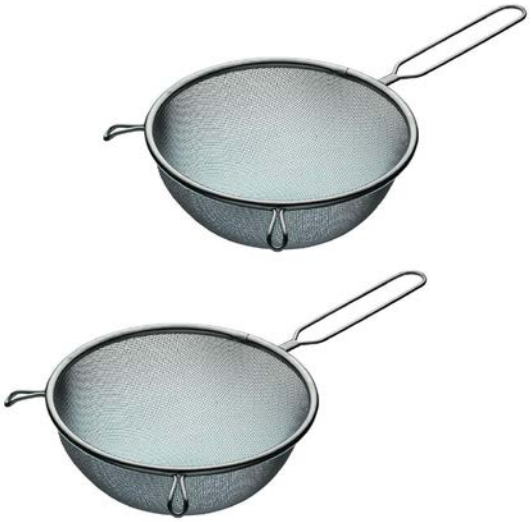
Student Performance Data Set (UCI)



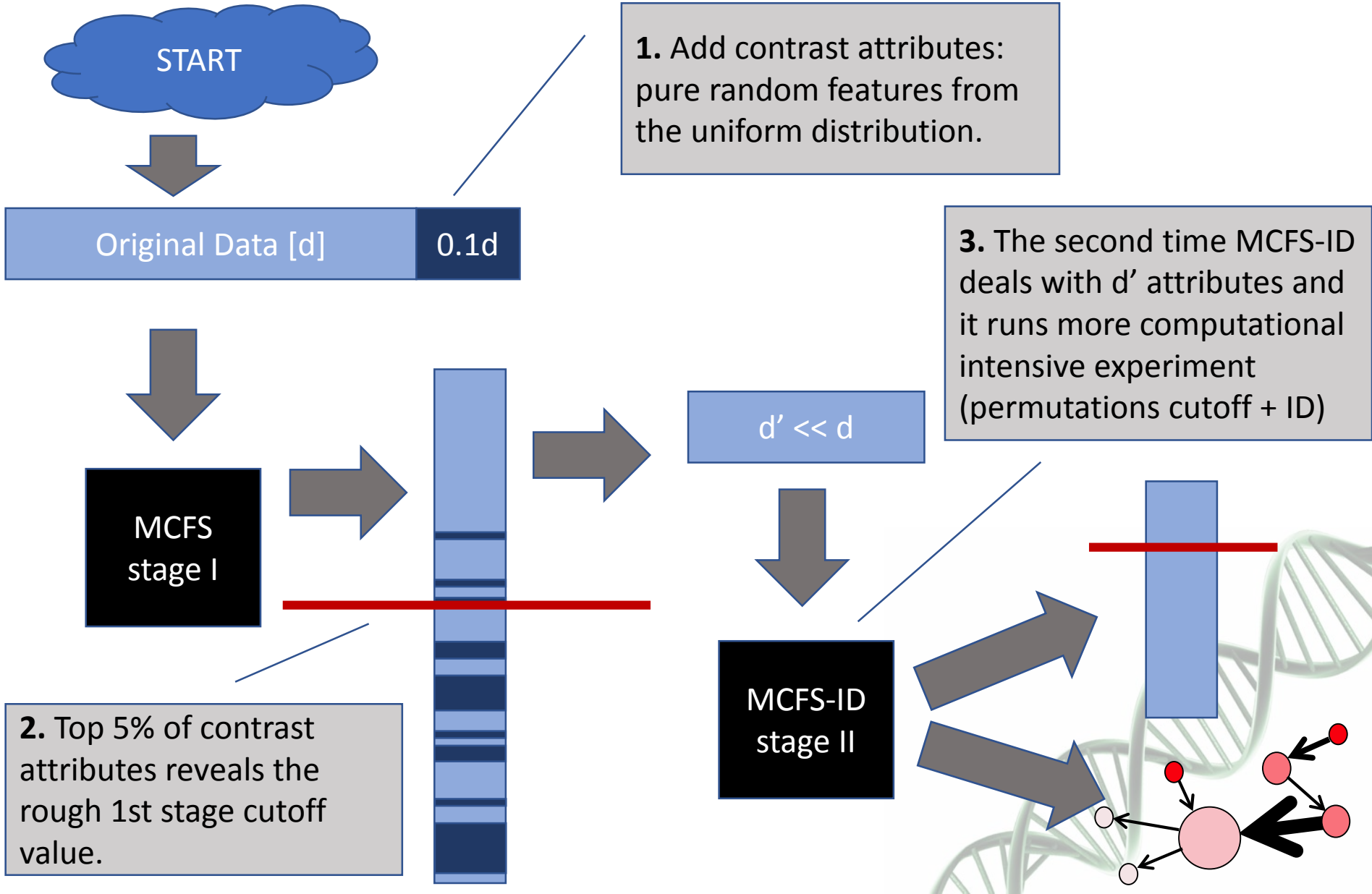


Koronacki, J., Dramiński, M. (**2017**). Empirical Model Building Revisited. Models and Reality: Festschrift for James Robert Thompson, Chicago, IL: T&NO Company.

Two stage filtering



Two stage filtering



1. Add contrast attributes: pure random features from the uniform distribution.

2. Top 5% of contrast attributes reveals the rough 1st stage cutoff value.

3. The second time MCFS-ID deals with d' attributes and it runs more computational intensive experiment (permutations cutoff + ID)

rmcfs: An R package ver. 1.3.0

rmcfs: a R implementation of MCFS-ID

- **rmcfs** is a R package - publicly available on **CRAN**.
- **Multithreaded parallel** implementation in Java.
- Provides **ranking of features** with cutoff point of the most significant features.
- Provides **interdependency directed graph** (ID-Graph) that shows non linear relations between features. These are not correlations!
- ID-Graph describes frequent interdependecies in the observed decision trees. In fact, the edges in the ID-Graph describe weighted conditional probabilities of attributes' occurrence.