30-11-2017

Monte Carlo Feature Selection and Interdependencies Discovery (MCFS-ID).

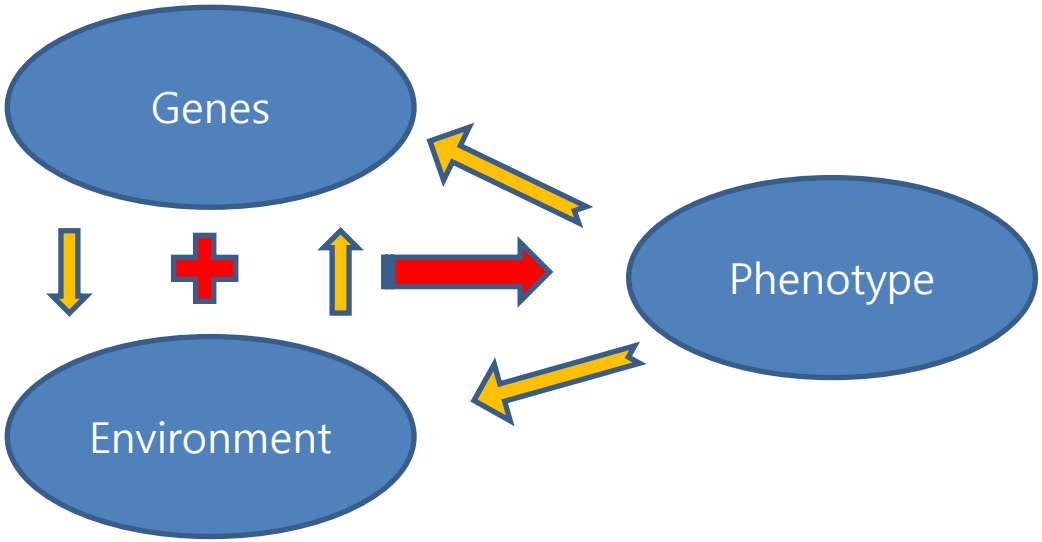
Michał Dramiński PhD

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

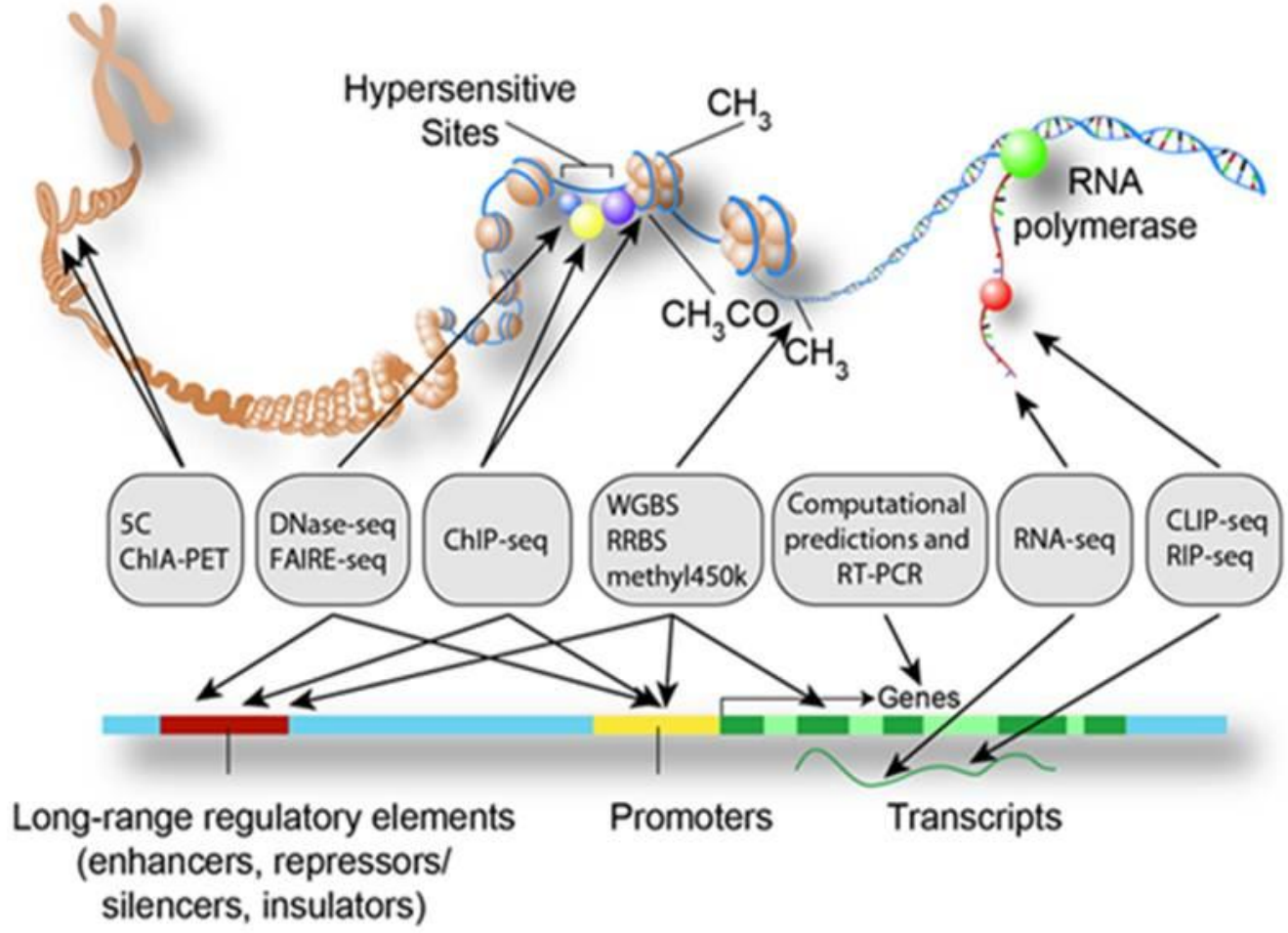
ZBO IPI PAN
http://zbo.ipipan.waw.pl

Single Gene Disorders

1. Over **10,000** human disorders are caused by a change, known as a mutation, in a **single gene**. As a whole, they affect about one percent of the population and can be easily tracked through families.
2. **Common medical problems do not have a single genetic cause!** They are associated with the effects of multiple genes in combination with lifestyle and environmental factors:
    - heart disease,
    - diabetes,
    - obesity,
    - cancer,
    - asthma/allergies,
    - autoimmune…



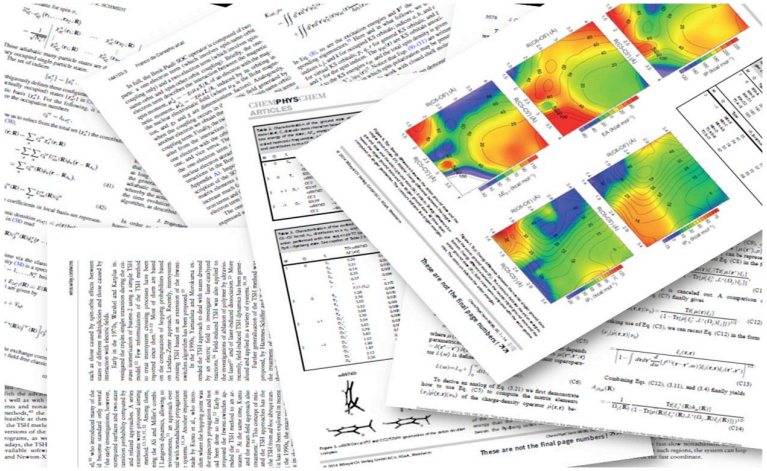
Gene expression regulation



Tissue
RNA ~ 60,000
CpG ~ 480,000
Genes ~ 23,000
SNPs ~ 10,000,000
TFs ~ 2,600
TFBS ~ 30,000
Histones:
meth/acc

https://www.encodeproject.org/

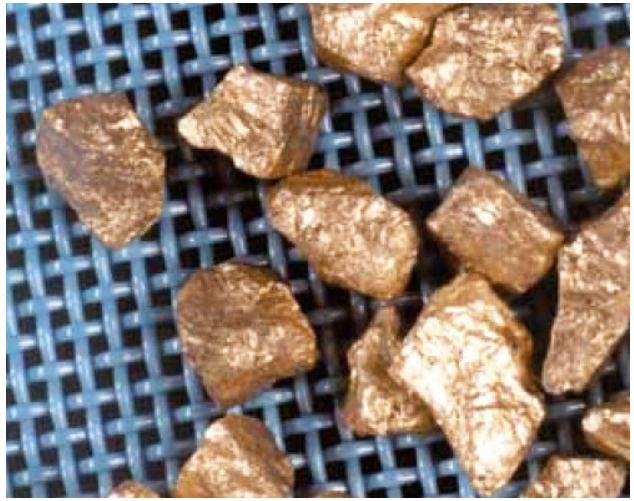
MCFS – History



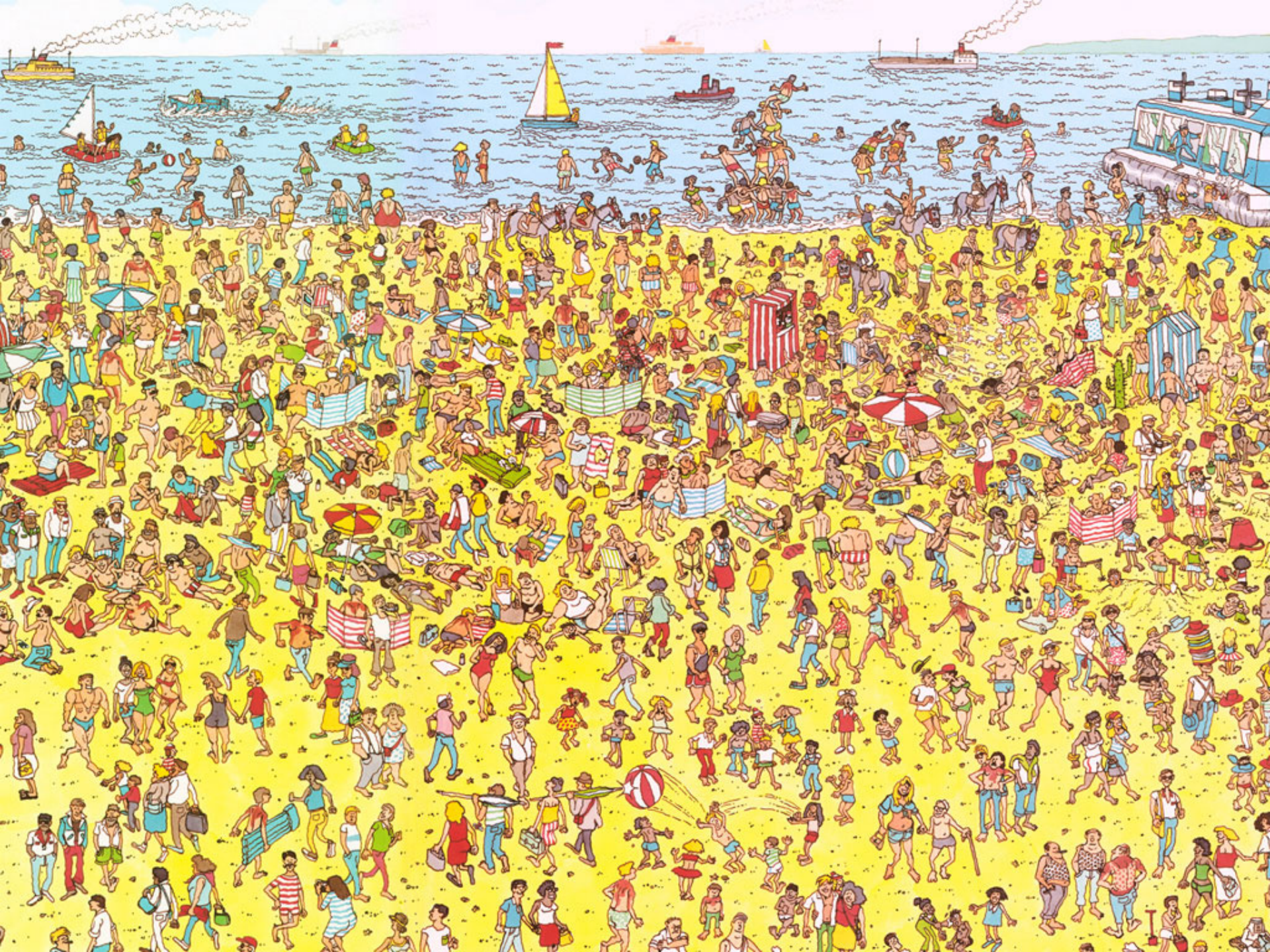
MCFS – History

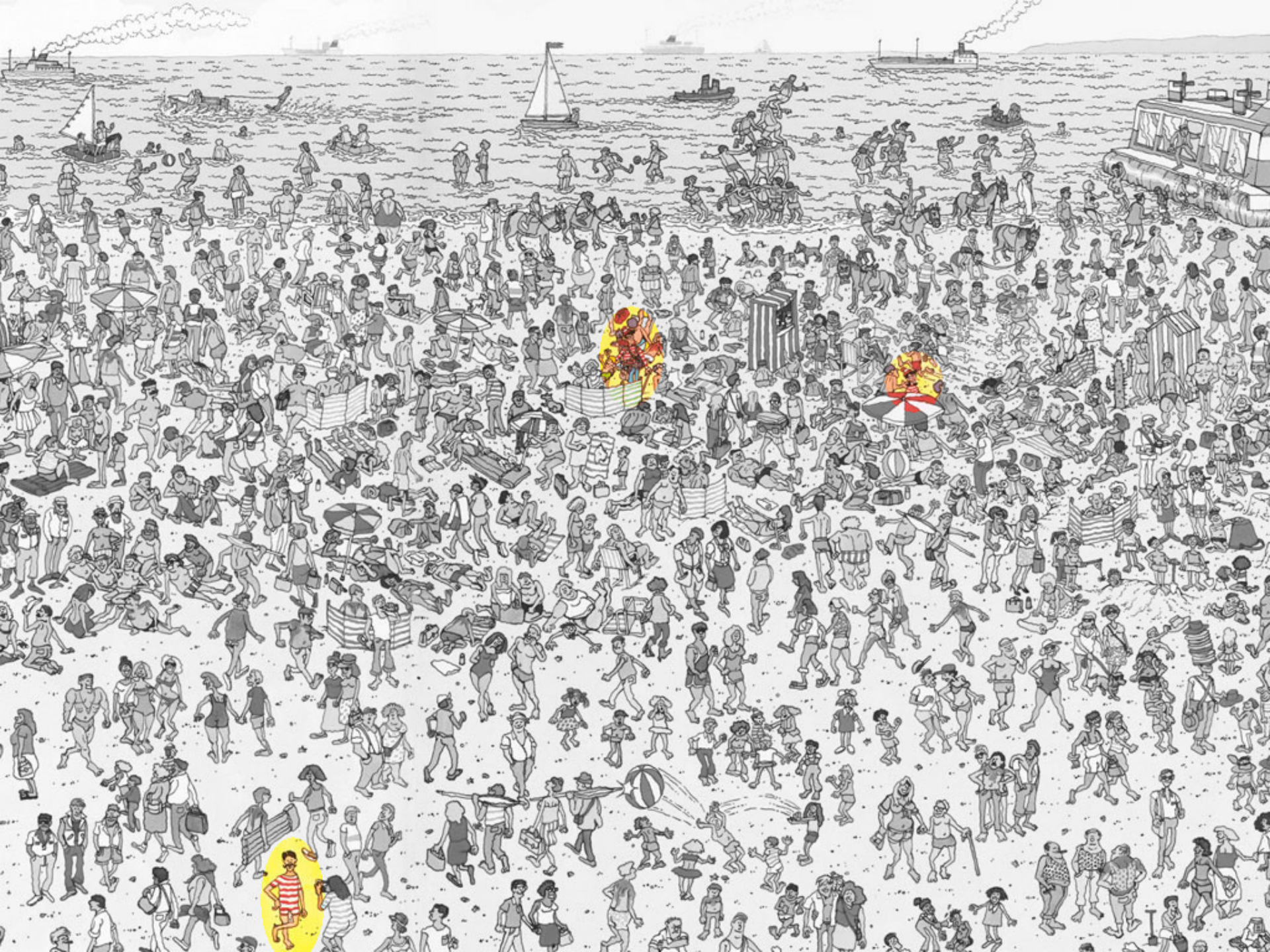
- M. Draminski, J. Koronacki. rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery. Accepted for publication in Journal of Statistical Software. Available now as vignette to rmcfs package, **2017**.
- Koronacki, J., Dramiński, M. (2017). Empirical Model Building Revisited. Models and Reality: Festschrift for James Robert Thompson, Chicago, IL: T&NO Company, **2016**.
- 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, M. Kierczak, A. Nowak-Brzezińska, J. Koronacki, J. Komorowski. "The Monte Carlo feature selection and interdependency discovery is practically unbiased." Control and Cybernetics vol 40, **2011**.
- M. Kierczak, M. Draminski, J. Koronacki, J. Komorowski. Computational analysis of molecular interaction networks underlying change of HIV-1 resistance to selected reverse transcriptase inhibitors. Libertas Academica Press. Bioinformatics and Biology Insights, **2010**.
- M. Dramiński, M. Kierczak, J. Koronacki, J. Komorowski. Monte Carlo feature selection and interdependency discovery in supervised classification. In: J Koronacki, Z Ras, S Wierzchon and J Kacprzyk, editors, Advances in Machine Learning II, Studies in Computational Intelligence. **2010.**
- M. Kierczak, K. Ginalski, M. Draminski, J. Koronacki, W. Rudnicki, J. Komorowski "A Rough Set-based model of HIV-1 RT Resistome", "Bioinformatics and Biology Insights", **2009**.
- M.Draminski, "Algorytm indukcji reguł decyzyjnych w problemach klasyfikacji i wyboru cech w zadaniach wysokowymiarowych ". Ph.D. thesis. Institute of Computer Science, Polish Academy of Sciences. **2008**.
- 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 "A Proposition of Integration of Rule Classifier Into MCFS Algorithm", Proceedings on VIII International Conference on Artificial Intelligence, AI-21, **2006**.
- 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.**

Monte Carlo Feature Selection and Interdependence Discovery



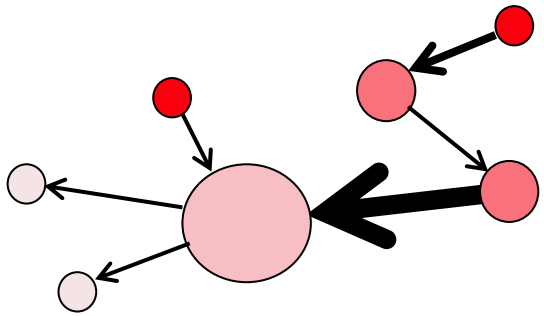
Algorithm (MCFS-ID)





rmcfs: a R implementation of MCFS-ID

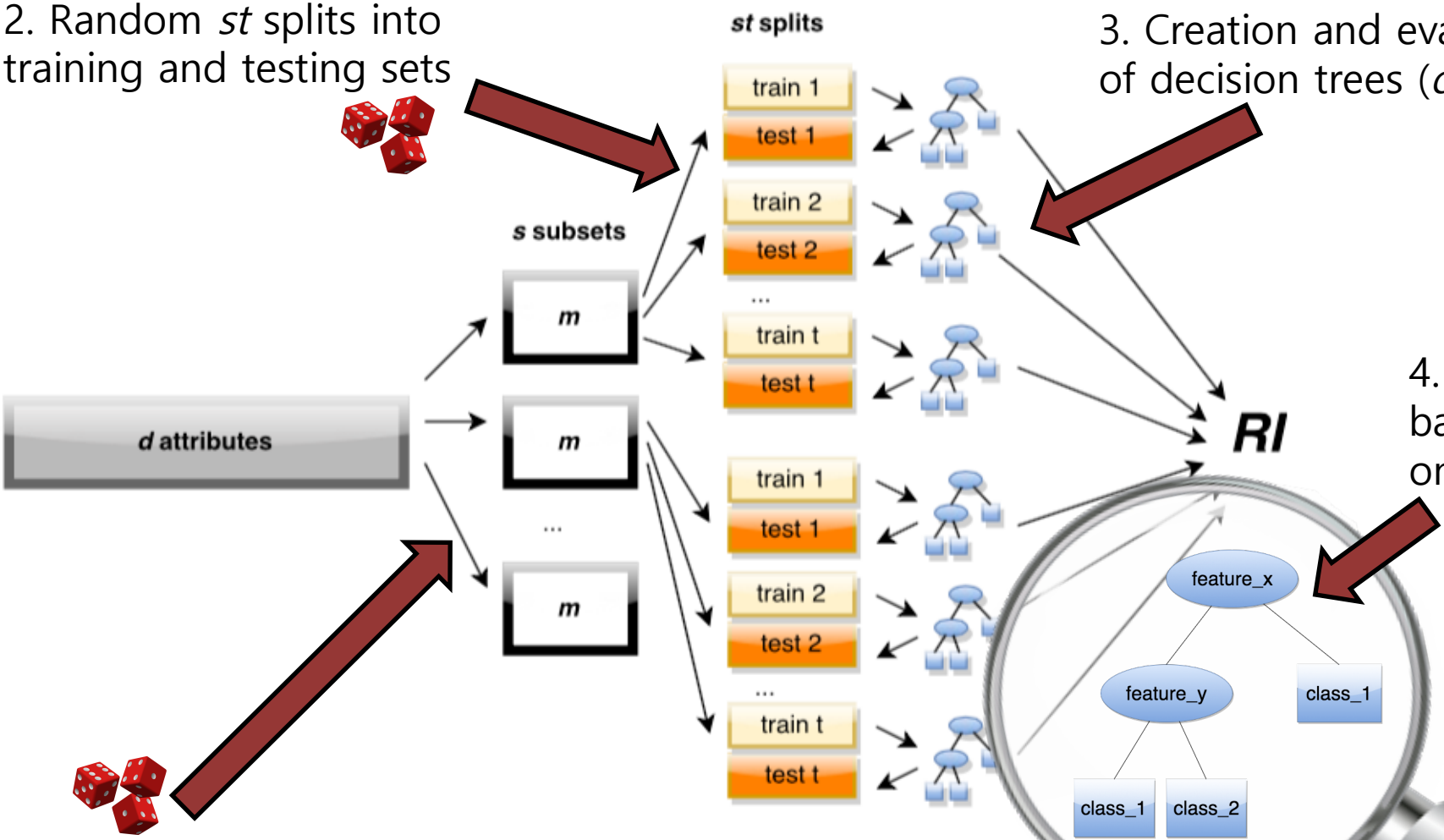
- rmcfs is a R package - publicly available on **CRAN**.
- **Multithreaded parallel** implementation in Java.
- Provides **ranking of features** with cutoff point of the most significant features.
- Provides **interdependency directed graph** (ID-Graph) that shows non linear relations between features.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



Algorithm (MCFS-ID)

2. Random *st* splits into training and testing sets

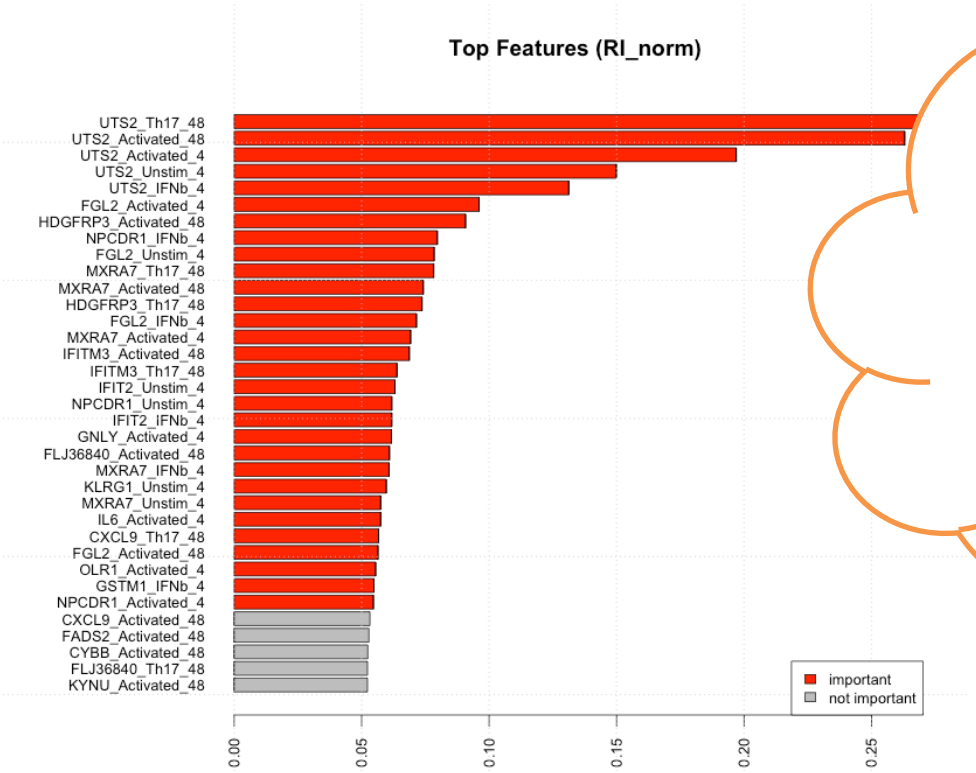
3. Creation and evaluation of decision trees (*dt*)



4. RI(f) is based on all *dt*

1. Random *s* selections of *m* features (*f*)

Relative Importance (MCFS-ID)



Important attribute:

1. occurs in many decision trees (DT),
2. is located nearby the root (separates many objects),
3. separates classes within the node with high quality,
4. DTs based on it perform well on unseen data.

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

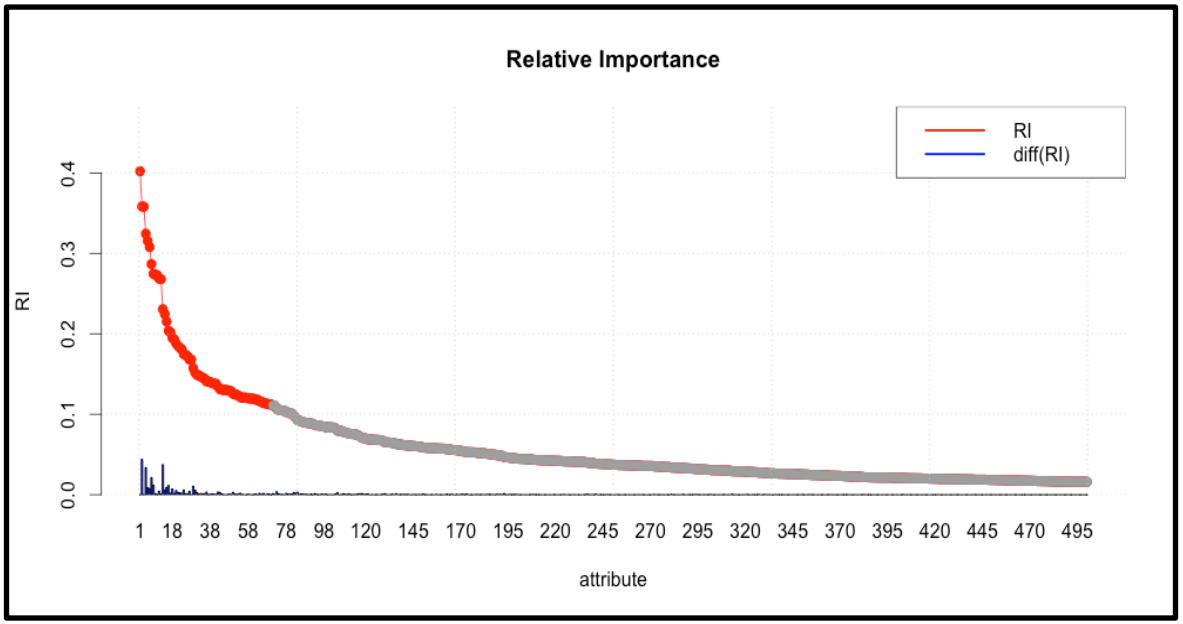
$$\mathrm{RI}_{g_k} = \sum_{\tau=1}^{s \cdot t} \mathrm{wAcc}_{\tau}^{u} \sum_{n_{g_k}(\tau)} \mathrm{IG}(n_{g_k}(\tau)) \left(\frac{\text{no. in } n_{g_k}(\tau)}{\text{no. in } \tau} \right)^{v},$$

1 4 3 2

Cut-off point in MCFS-ID

Cut Off (MCFS-ID)

Cuf off value can be determined by 3 different methods.

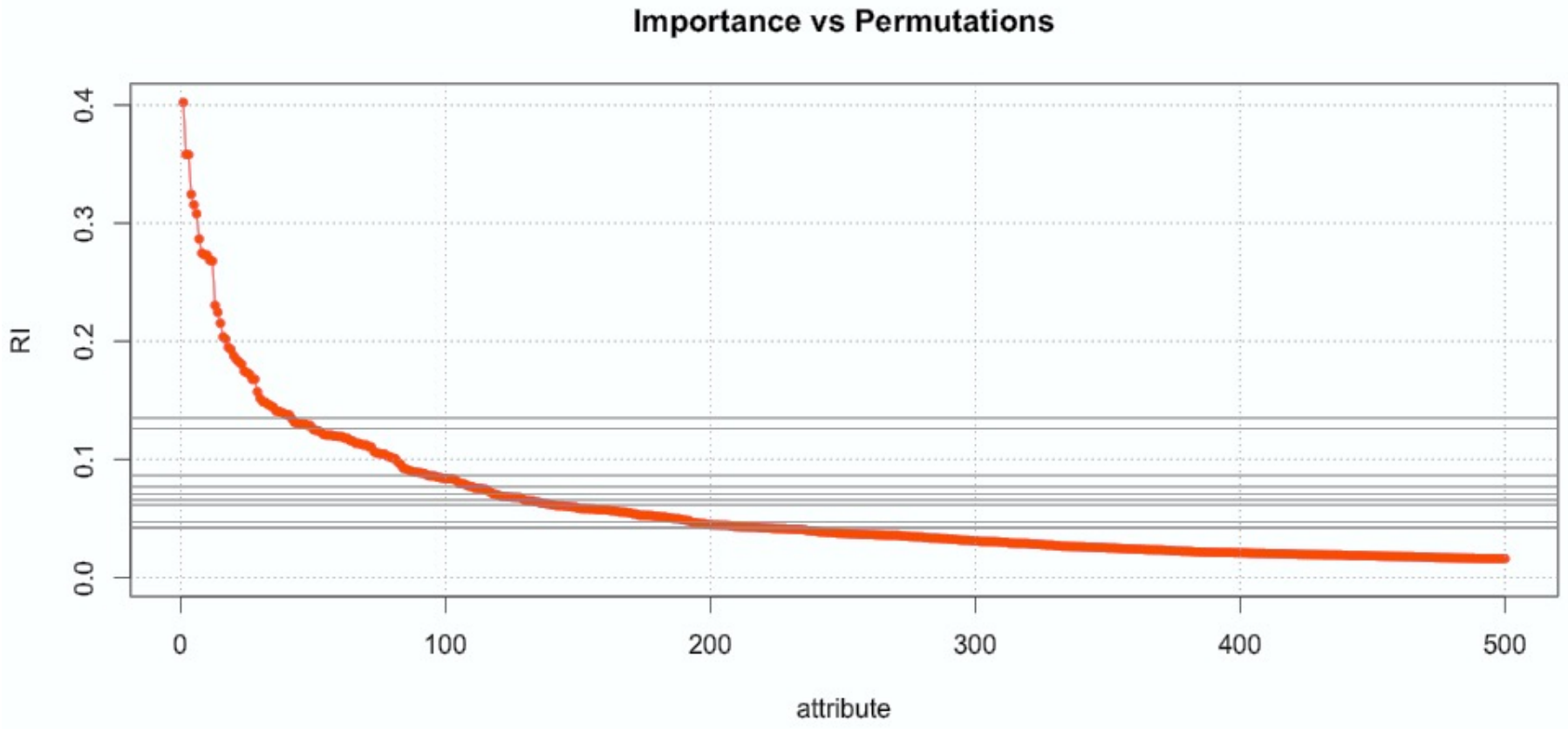


Methods:

1. Critical angle
2. Kmeans
3. **Permutations (max RI)**

Cut Off – permutations (MCFS-ID)

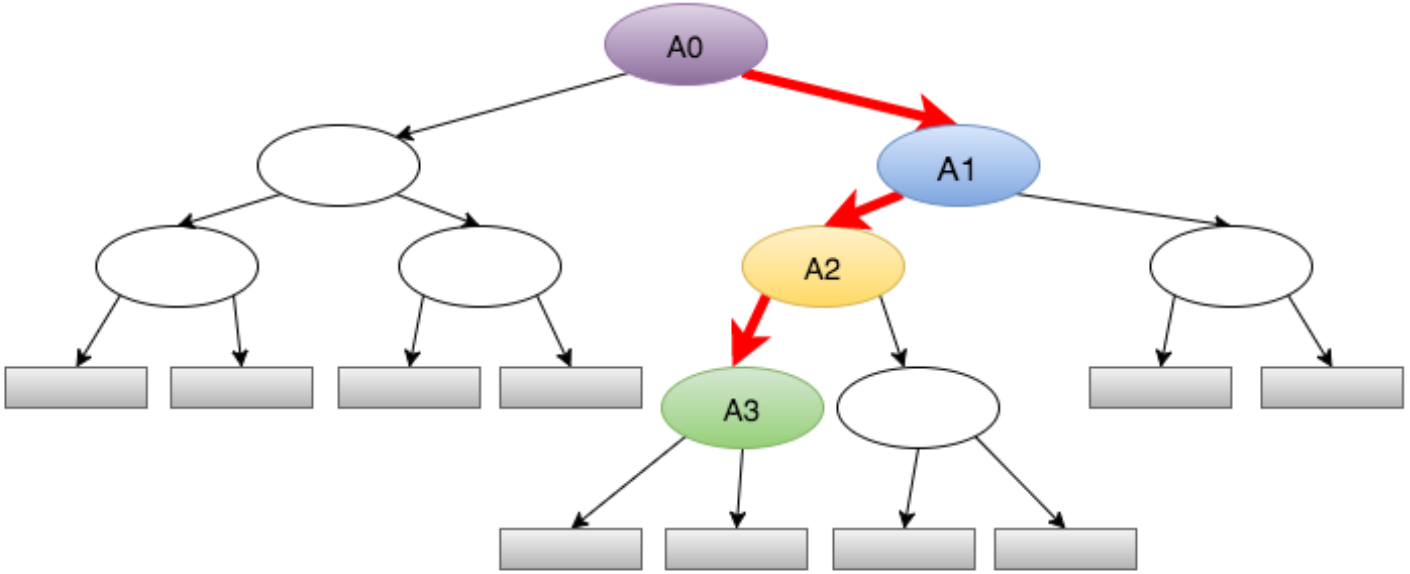
Permute x times decision attribute and run mcfs. Compare all maximal RI obtained from such experiments to the reference experiment (with original decision attribute).

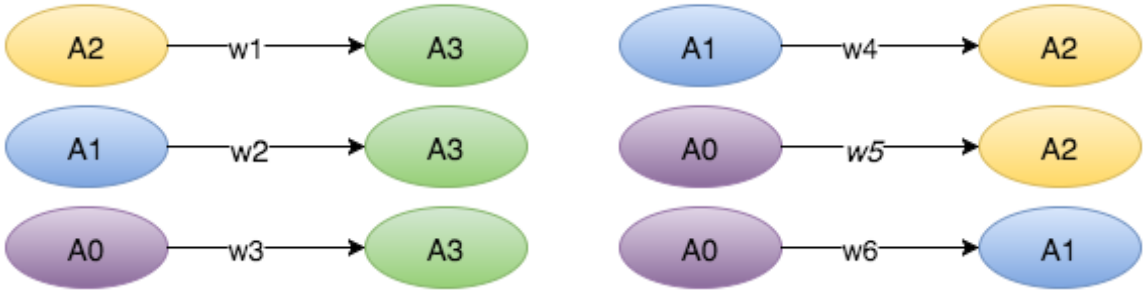


ID-Graphs

ID Graph (MCFS-ID)

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

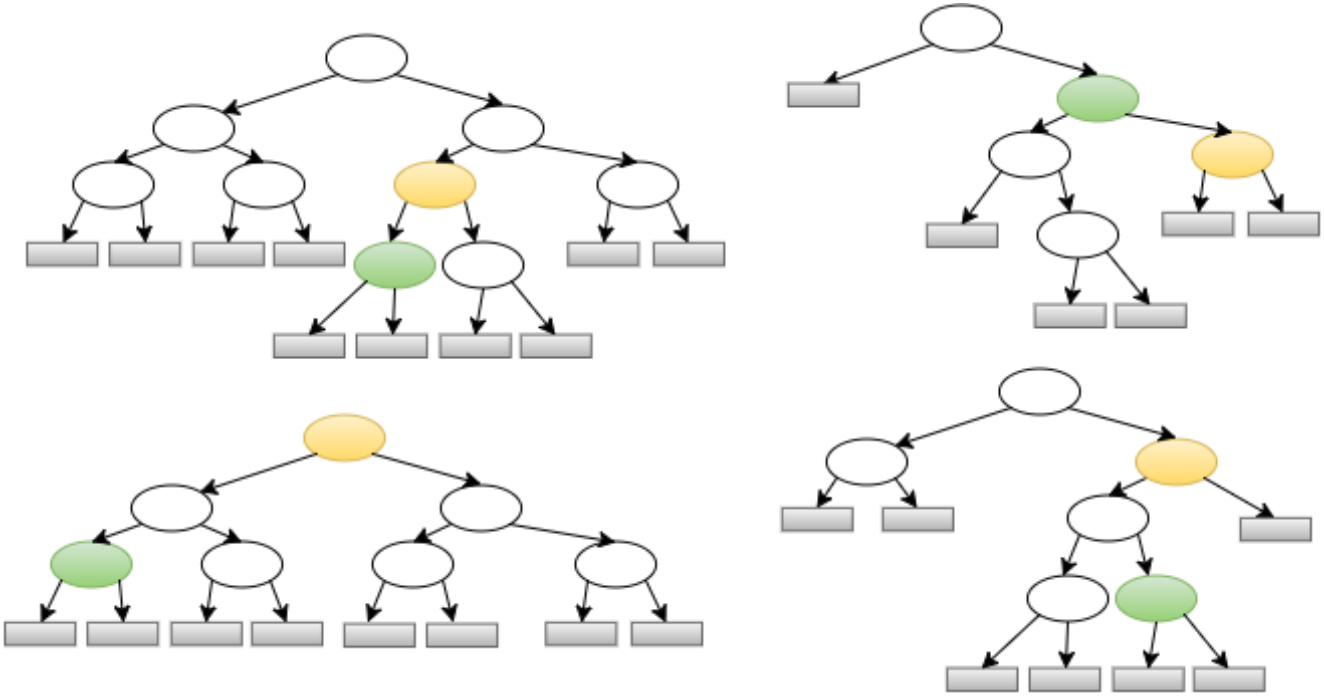




ID Graph (MCFS-ID)

Weight of the edge in ID-graph is large if:

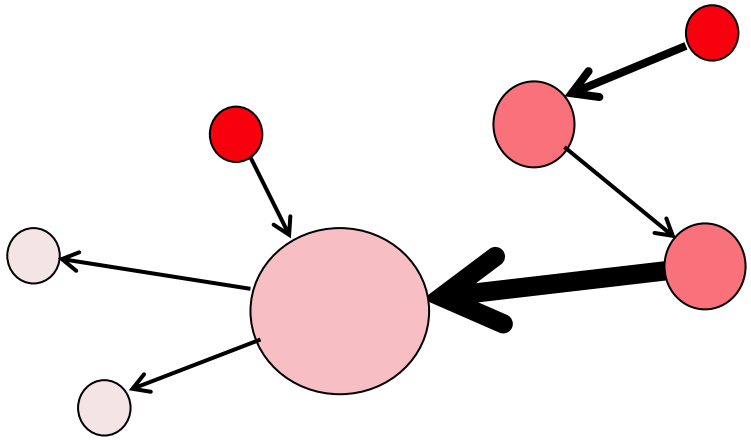
- Both nodes frequently co-occur on the same path in many trees.
- Child node has high (IG) and contains relatively large portion of ancestor objects.



ID Graph (MCFS-ID)

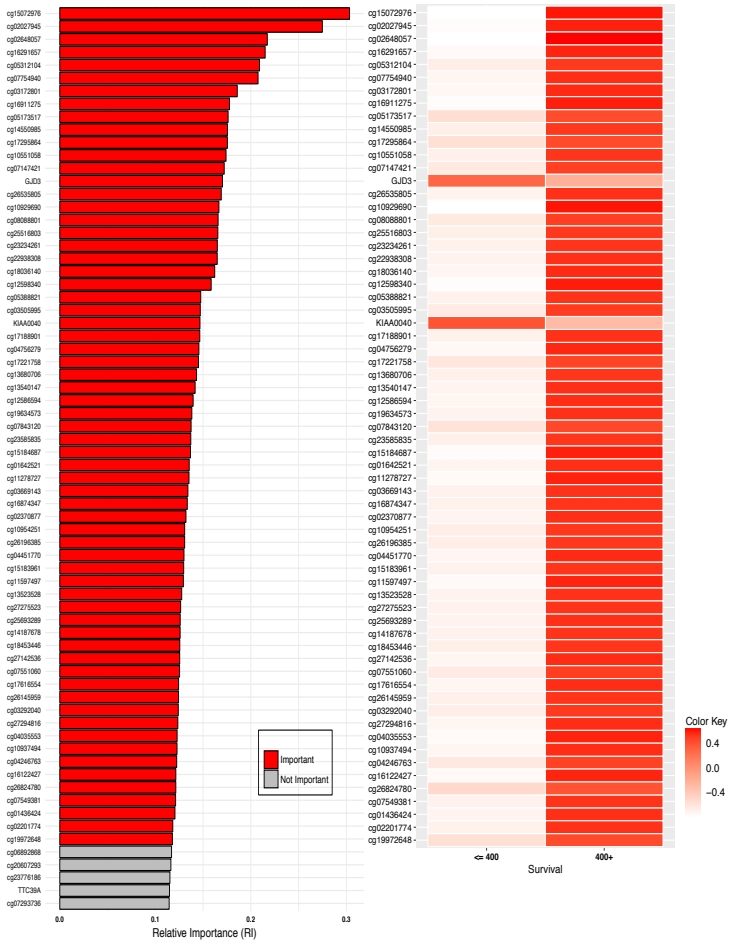
ID-graph Plot rules:

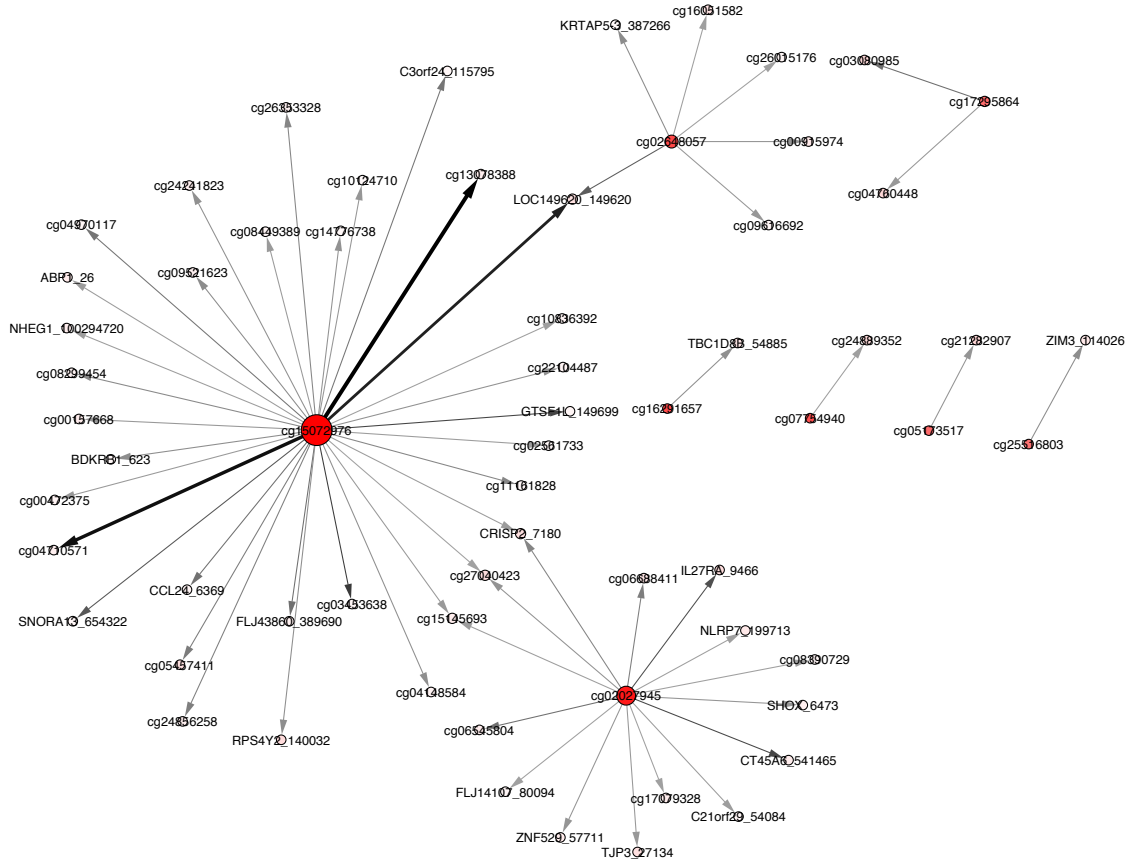
- The higher RI of the feature, the stronger color intensity;
- The more often a feature co-occurs with others the larger the size of the node;
- The larger the weight of the edge, the thicker the connection is;
- The direction of the links follows the path in the tree: from ancestors to their children.



Glioma Patients Survival

Features reduced from 416 013 to 65!

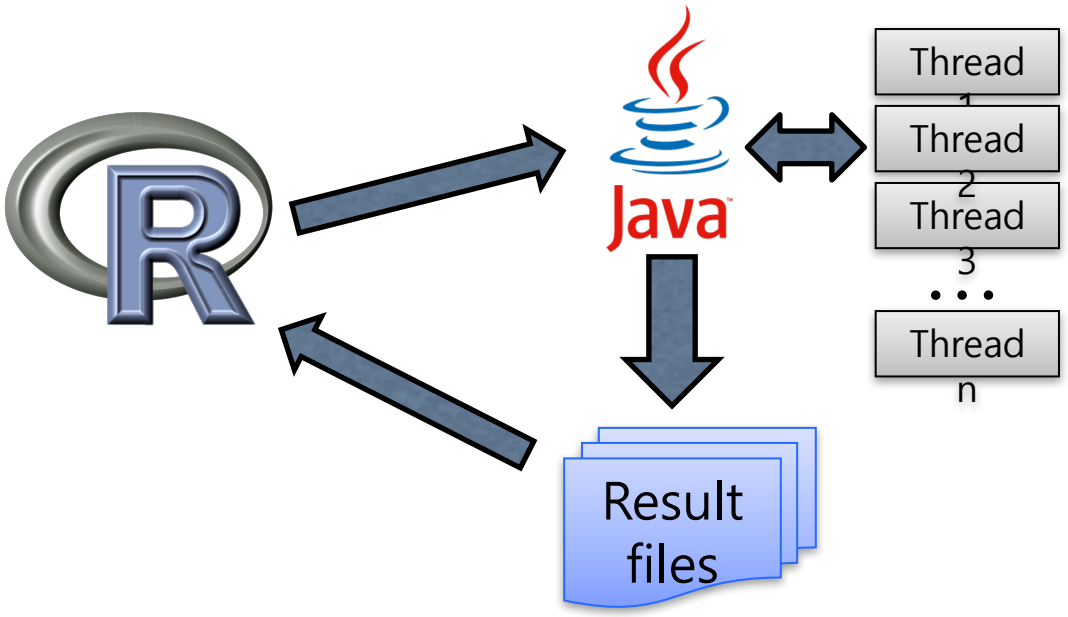




rmcfs: An R package

R package – Implementation overview

1. Function `mcfs` is a wrapper on Java mcfs 2.0.x implementation.
   - Check in bash if Java (1.6 or higher) is installed in your system: `java –version`
2. Problems with R library rJava?
   - Run in bash: `sudo R CMD javareconf`



rmcfs package – functions

rmcfs library implements – 16 functions available to the user

```
showme()
artificial.data()
fix.data()
prune.data()

mcfs()
build.idgraph()
```

```
read.adx()
read.adh()
write.adx()
write.adh()
write.arff()

export.result()
import.result()
```

```
plot.mcfs()
print.mcfs()
plot.idgraph()

S3 plot()
S3 print()
S3 plot()
```

https://cran.r-project.org/web/packages/rmcfs/index.html
http://www.ipipan.eu/staff/m.draminski/mcfs.html

Thank You!

Computational Biology Lab:

Professors

Jacek Koronacki
Jan Komorowski

Assistant Professors

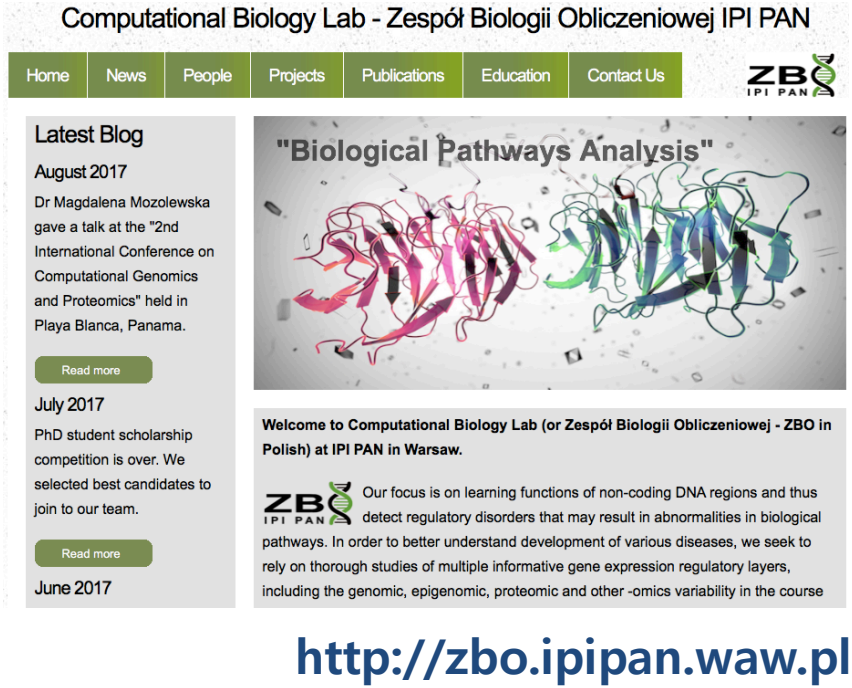
Michal J. Dabrowski
Michal Draminski
Magdalena A. Mozolewska

PhD Students

Agata Dziedzic
Ilona E. Grabowicz
Marta Jardanowska

Undergraduate Students

Aleksandra Fronc
Mateusz Kacprowski
Marta Lukasik
Aleksandra Mikulska
Michal Nowakowski
Adam Przybyłek



http://zbo.ipipan.waw.pl