Uczenie maszynowe w analizie dużych danych medycznych z przykładami w onkologii

Przemysław Biecek
Grupa MI^2

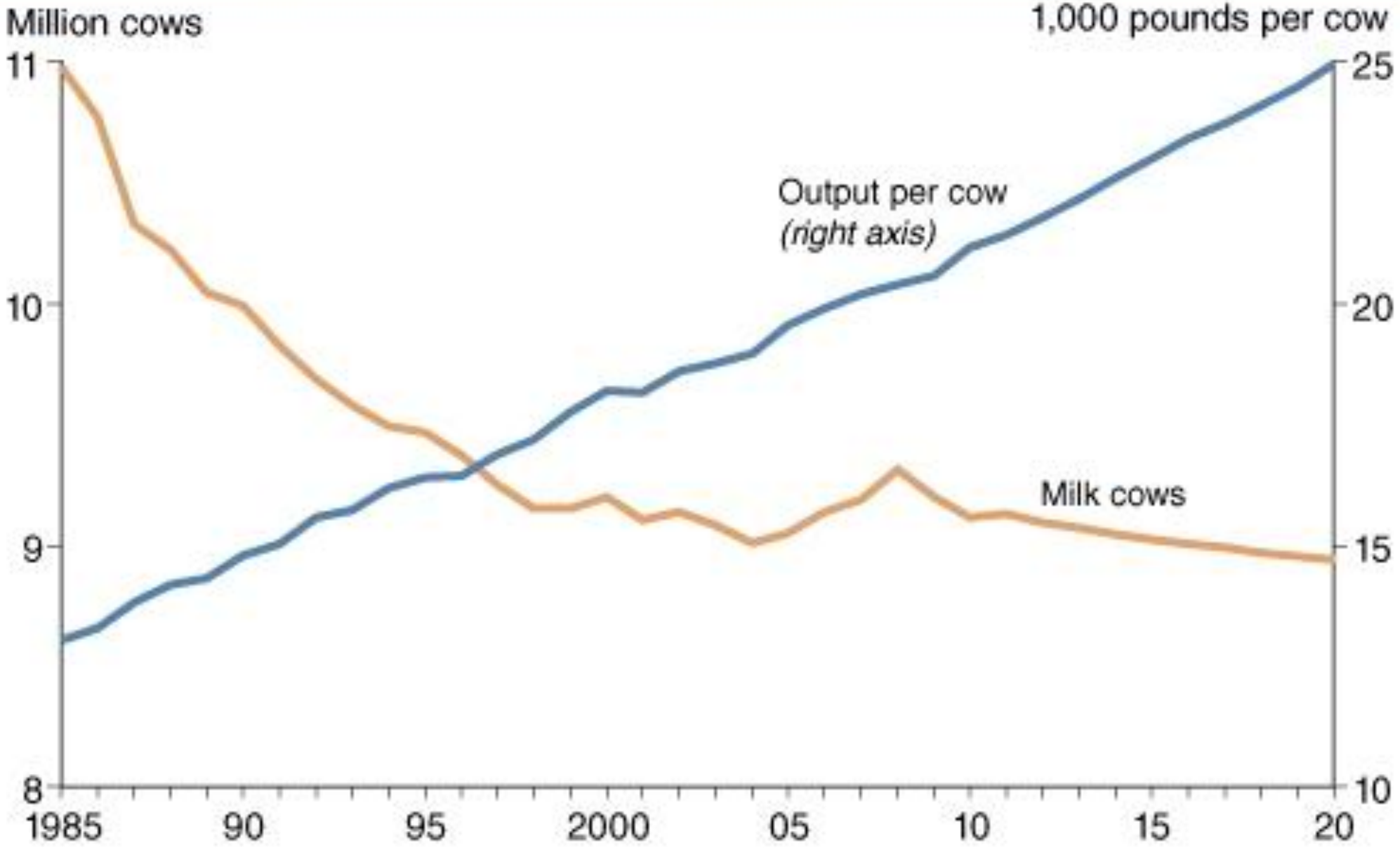
Plan

- modelowanie predykcyjne
- ,,małe dane'' w onkologii
- ,,duże dane'' w onkologii

Pytanie:

O ile wzrosła produkcja mleka na krowę w ostatnich 40 latach?

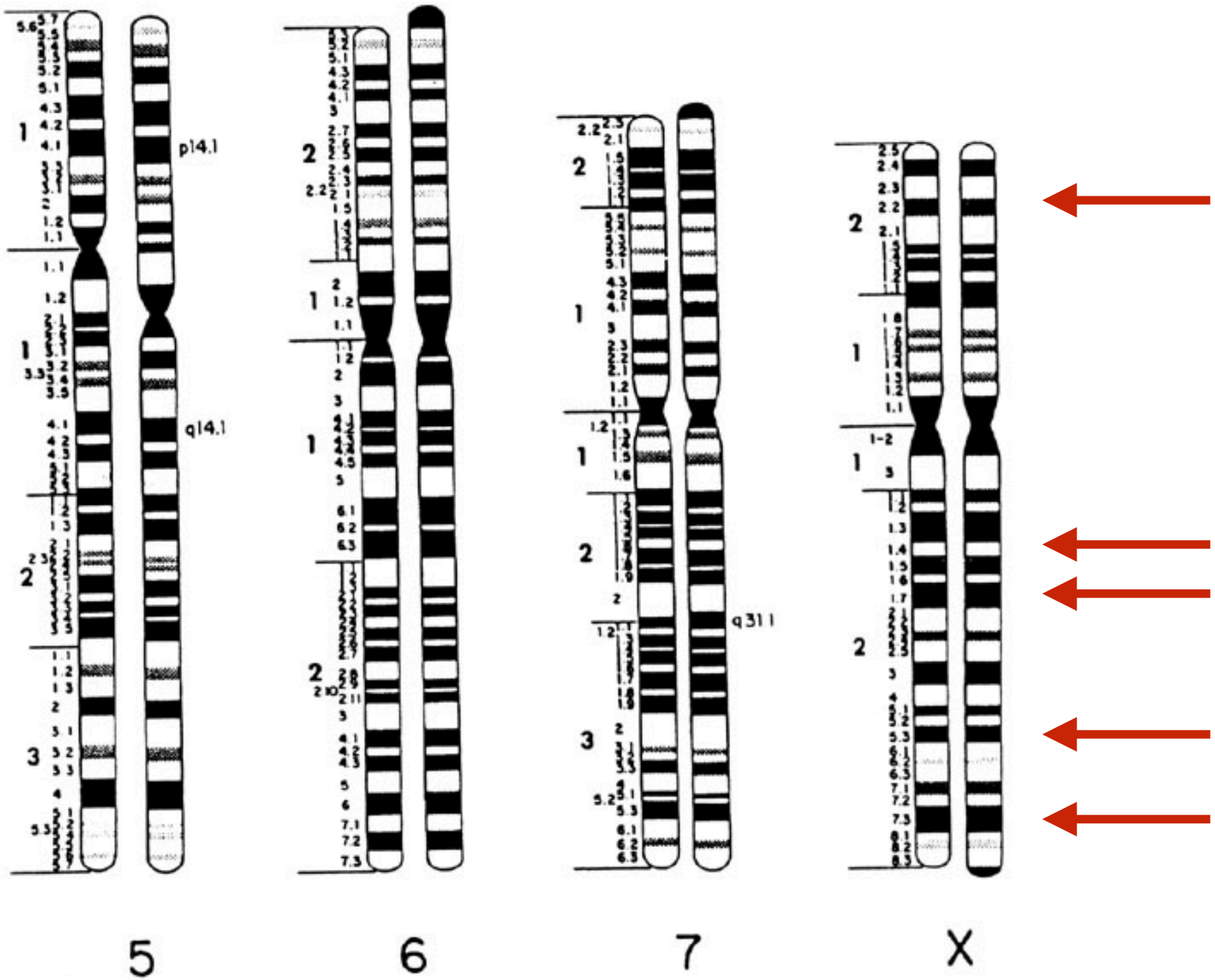
U.S. dairy herd and milk production per cow

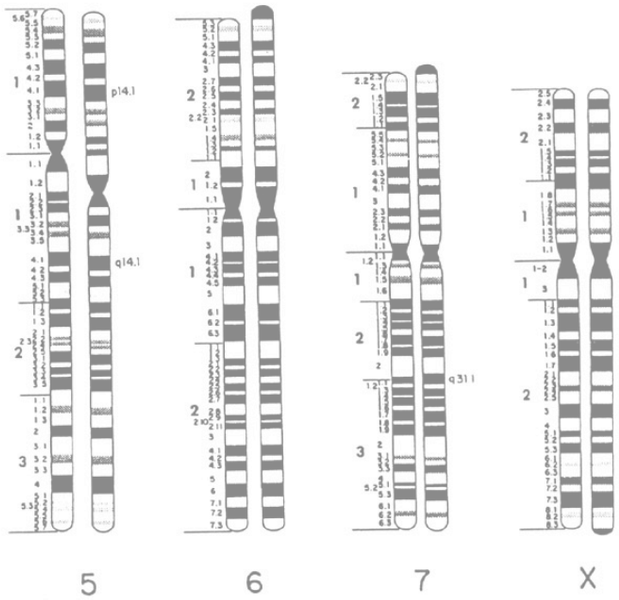


Source: USDA, Economic Research Service using USDA Agricultural Projections to 2020.

Ile mleka (średnio) wyprodukują jego córki?

QTL mapping - markery genetyczne
QTL = Quantitative Trait Loci



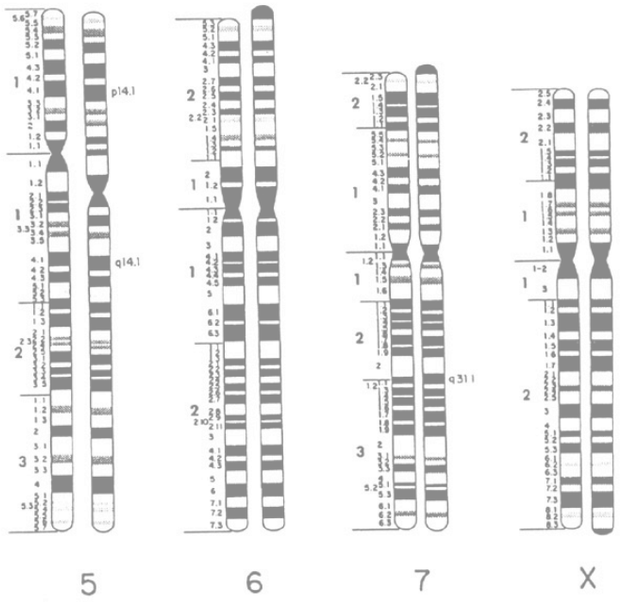


markery

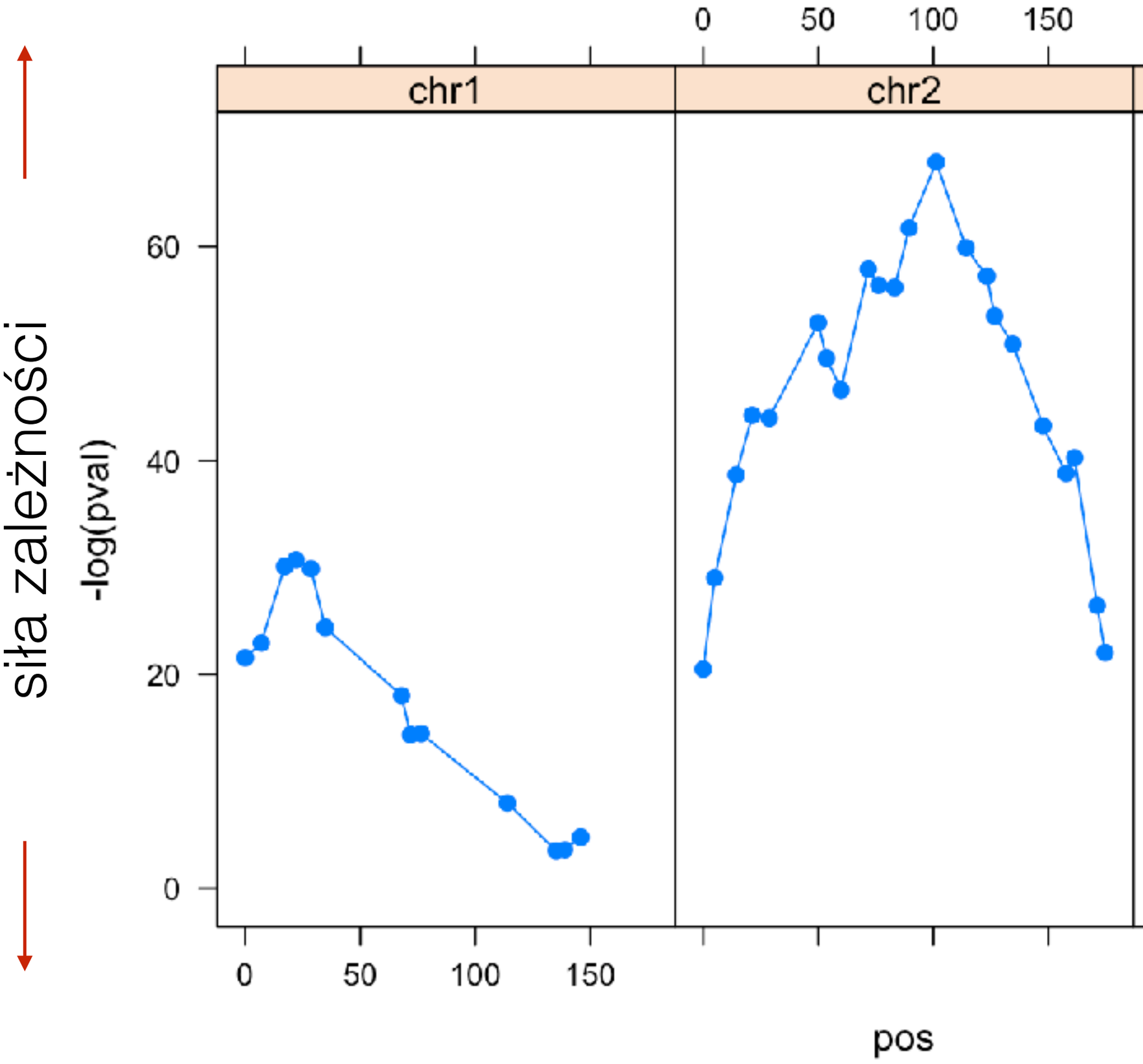
osobniki

```
> head(bs,30)
        ewg w rp v sd run gl pgk cg gpdh ninaC glt mhc ddc duc sli
BS01-1    0 0  0 0  0   0  1   0  0    0     0   0   1   1   1   1
BS01-2    2 2  2 2  2   2  1   1  1    1     1   1   1   1   0   1
BS01-3    0 0  0 0  0   0  1   1  1    1     1   1   1   1   1   1
BS01-4    2 2  2 2  0   0  0   0  0    0     0   0   1   1   1   1
BS01-5    0 0  0 0  0   0  0   0  0    0     0   0   0   0   0   0
BS02-1    0 2  2 2  2   2  1   1  1    1     1   1   1   1   1   0
BS02-2    0 0  0 0  0   0  1   1  1    1     0   0   0   0   0   0
BS02-4    0 0  0 0  0   0  1   1  1    1     1   1   0   0   0   1
BS02-5    2 2  2 2  2   2  1   1  1    1     1   1   1   1   1   1
BS03-1    0 0  0 0  0   0  0   0  0    0     0   0   0   1   1   0
BS03-2    2 2  0 0  0   0  1   1  1    1     1   1   0   0   0   0
BS03-4    2 2  2 2  0   0  1   0  1    1     1   1   1   1   1   1
BS03-5    0 0  0 0  0   0  0   0  0    0     0   0   1   1   1   0
BS04-2    2 2  2 2  2   2  0   0  0    0     0   1   0   0   0   0
BS04-3    2 0  0 0  0   0  1   1  1    1     1   1   0   0   0   1
BS04-4    2 2  2 0  0   0  0   0  0    0     0   0   1   1   1   1
BS04-5    0 0  0 0  0   0  0   0  0    0     0   0   0   0   0   0
BS05-1    2 2  2 2  2   2  1   1  1    1     1   1   1   1   1   1
BS05-2    0 0  2 2  2   2  0   0  0    0     0   0   0   0   0   0
BS05-3    2 2  2 2  2   0  0   0  0    0     0   0   1   1   1   1
BS05-5    2 2  2 2  2   2  1   0  0    0     0   0   1   1   1   0
```

QTL mapping - zależności

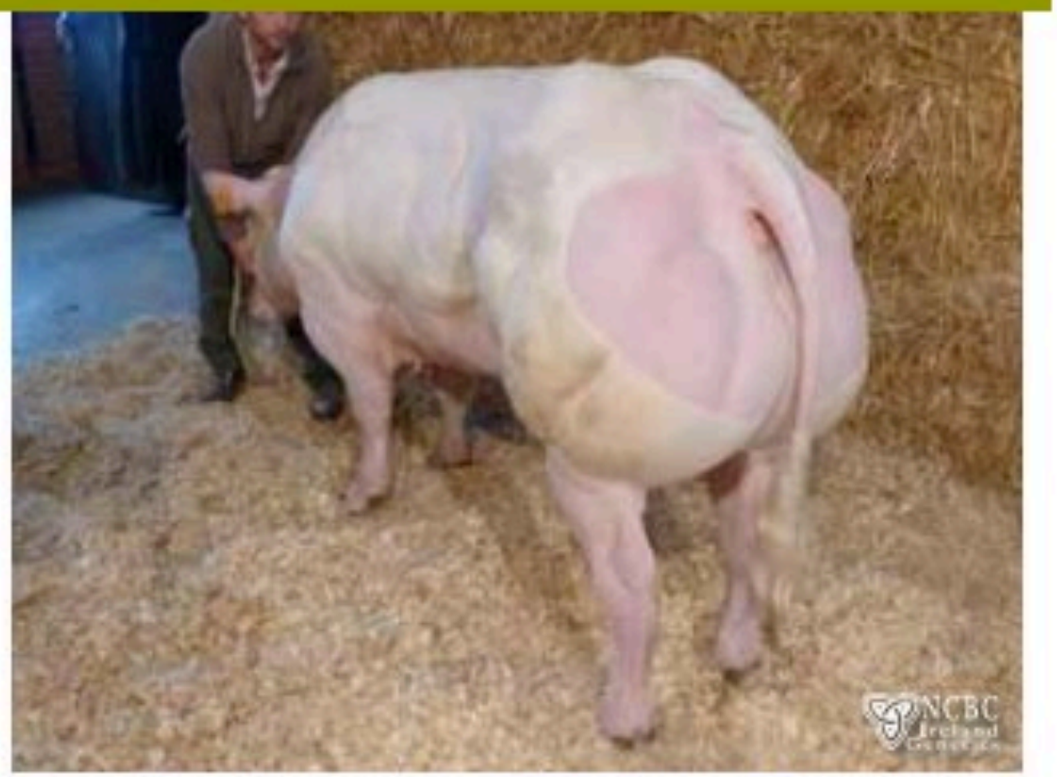



```
> head(bs,30)
       ewg w rp v sd run gl pgk cg gpdh ninaC glt m
BS01-1   0 0  0 0  0   0  1   0  0    0     0   0
BS01-2   2 2  2 2  2   2  1   1  1    1     1   1
BS01-3   0 0  0 0  0   0  1   1  1    1     1   1
BS01-4   2 2  2 2  0   0  0   0  0    0     0   0
BS01-5   0 0  0 0  0   0  0   0  0    0     0   0
BS02-1   0 2  2 2  2   2  1   1  1    1     1   1
BS02-2   0 0  0 0  0   0  1   1  1    1     0   0
BS02-4   0 0  0 0  0   0  1   1  1    1     1   1
BS02-5   2 2  2 2  2   2  1   1  1    1     1   1
BS03-1   0 0  0 0  0   0  0   0  0    0     0   0
BS03-2   2 2  0 0  0   0  1   1  1    1     1   1
BS03-4   2 2  2 2  0   0  1   0  1    1     1   1
BS03-5   0 0  0 0  0   0  0   0  0    0     0   0
BS04-2   2 2  2 2  2   2  0   0  0    0     0   1
BS04-3   2 0  0 0  0   0  1   1  1    1     1   1
BS04-4   2 2  2 0  0   0  0   0  0    0     0   0
BS04-5   0 0  0 0  0   0  0   0  0    0     0   0
BS05-1   2 2  2 2  2   2  1   1  1    1     1   1
BS05-2   0 0  2 2  2   2  0   0  0    0     0   0
BS05-3   2 2  2 2  2   0  0   0  0    0     0   0
BS05-5   2 2  2 2  2   2  1   0  0    0     0   0
```



BB2083
RACHID DE REMICHAMPAGNE

DOB: 05-APR-2013
HERD BOOK NO.: BBLBELM000857527143



CALVING DIFFICULTY		GESTATION (DAYS)	
PTA	REL.	PTA	REL.
6.3%	16%	-0.8d	42%

ICBF APR 2016

ACCORD DE WIHOGNE
JUBILAIRE DE LA CLAIE
CARESSE DE LA CLAIE

DEBORDANT ET DE BIOURGE
GOUTTE DE REMICHAMPAGNE
BOUTURE DE REMICHAMPAGNE

- First calves were born easily & good quality
- Easy calving on cows
- Widely used in Holland on dairy cows

BB4015
SIRE DE LA COUE

DOB: 01-NOV-2014
HERD BOOK NO.: BBLBELM000451780488

CALVING DIFFICULTY		GESTATION (DAYS)	
PTA	REL.	PTA	REL.
11.47%	14%	-0.88d	14%

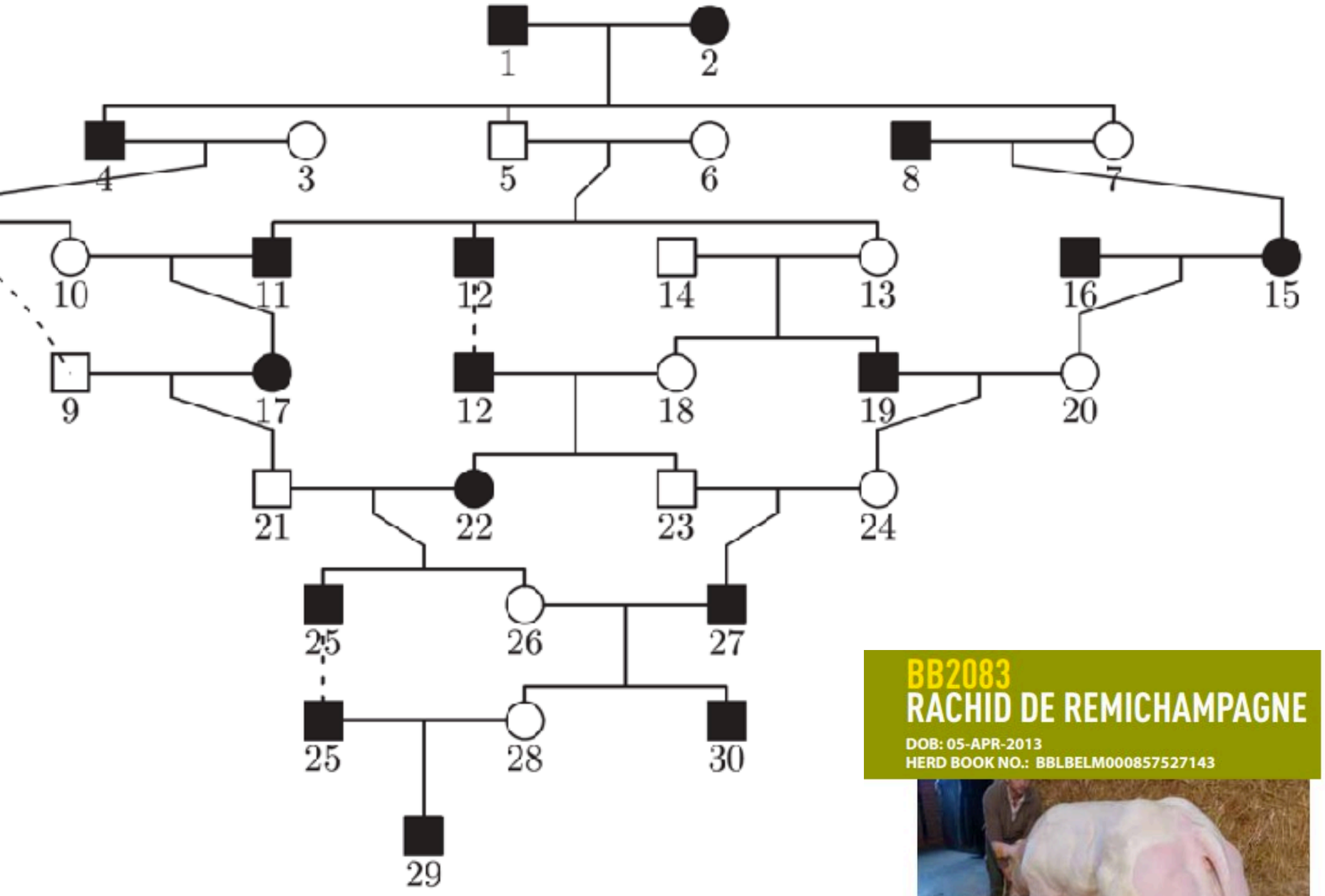
ICBF APR 2016

ORME DE SOMME
CAPPUCCINO DES AMANDIERS
LAETITIA ET DE CENTFONTAINE

HERISSON DE LA COUE
LOQUACE DE LA COUE
DELICIEUSE DE LA COUE

- Selected for crossing on dairy cows

BELGIAN BLUE



QTL mapping - modele mieszane lme4 / nlme

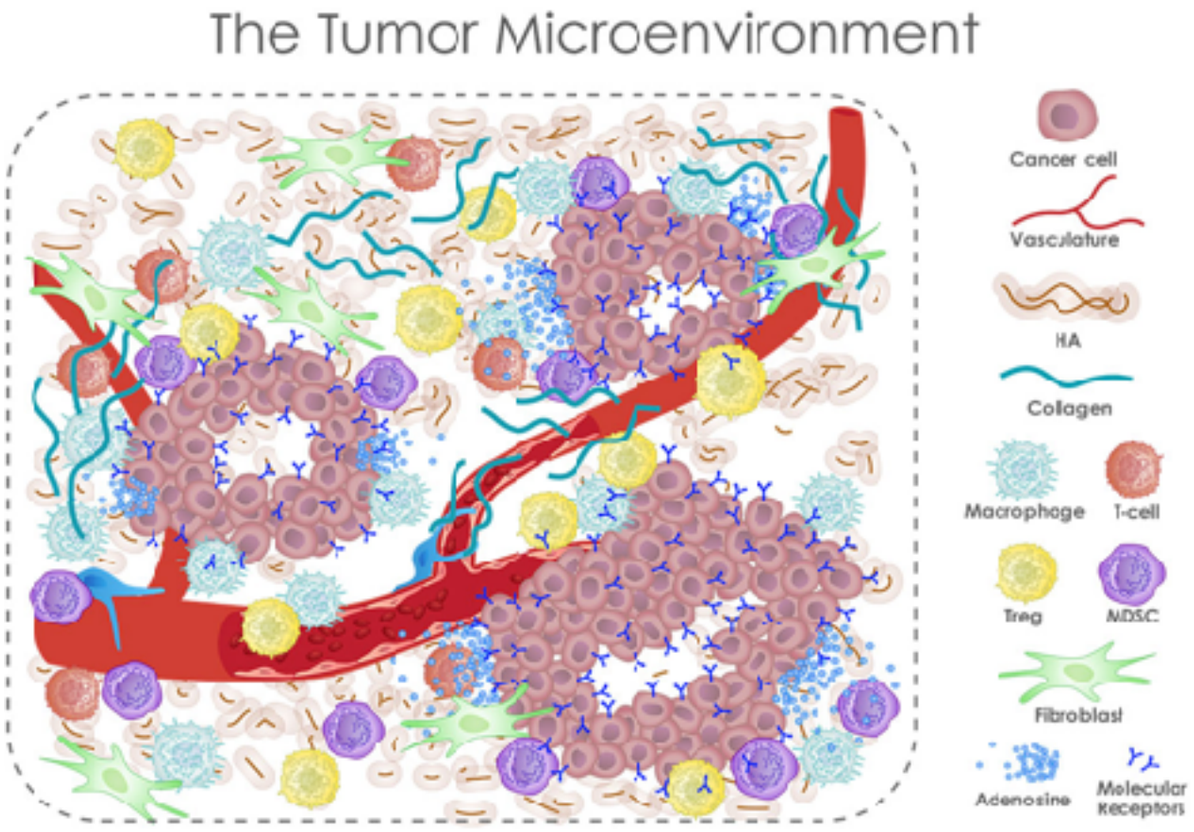
modelowanie predykcyjne

- identyfikacja tzw. actionable markers
- walidacja oparta o skuteczność

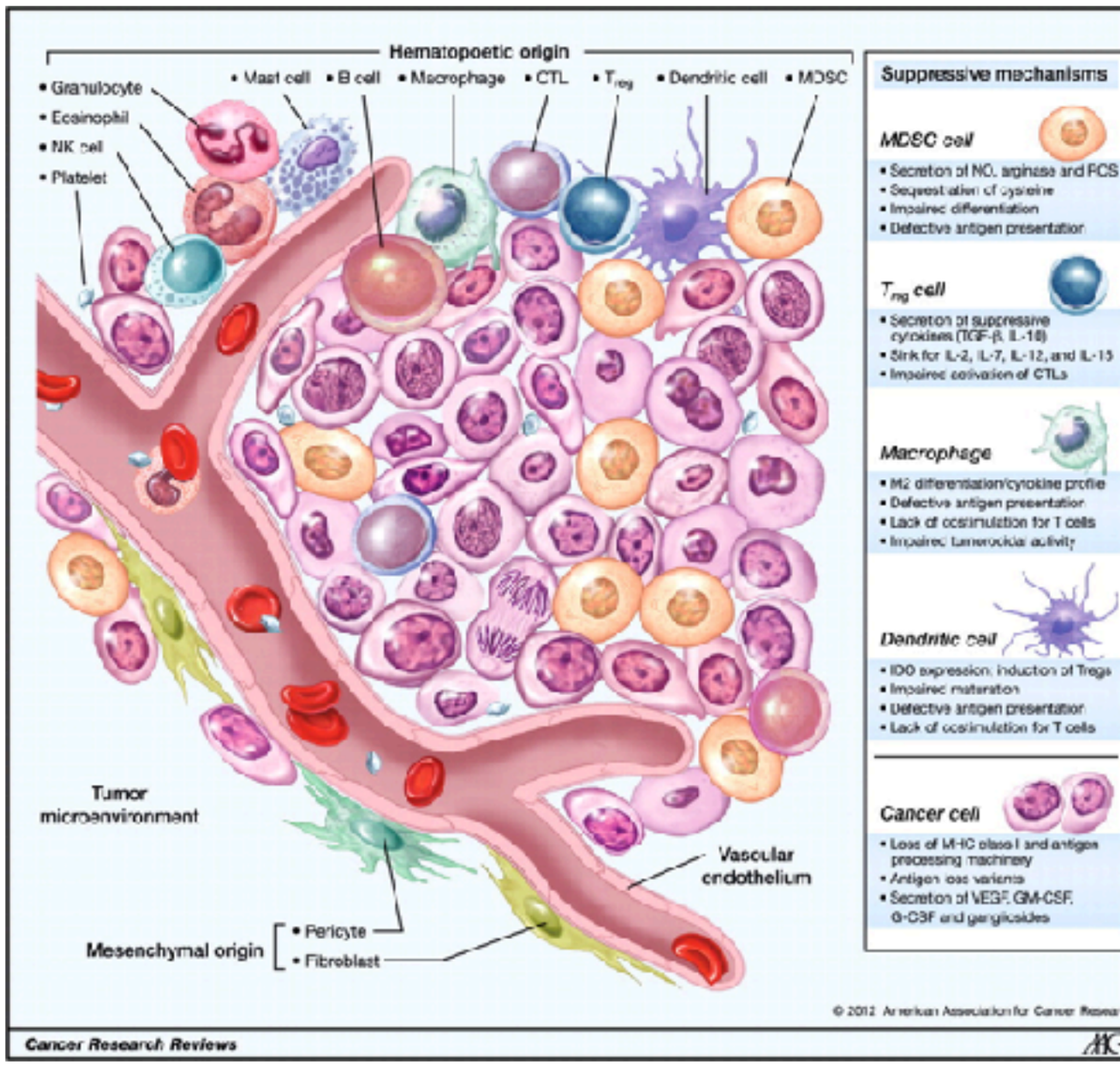
modelowanie matematyczne

- modelowanie określonego procesu
- lepsze zrozumienie opisywanego systemu

Nowotwór nie jest jednorodny,
mikro-środowisko różnorodnych komórek
bardzo szybka ewolucja - leczenie jest trudne



Whatcott et al.
Clin Cancer Res. 21:15 (2015)



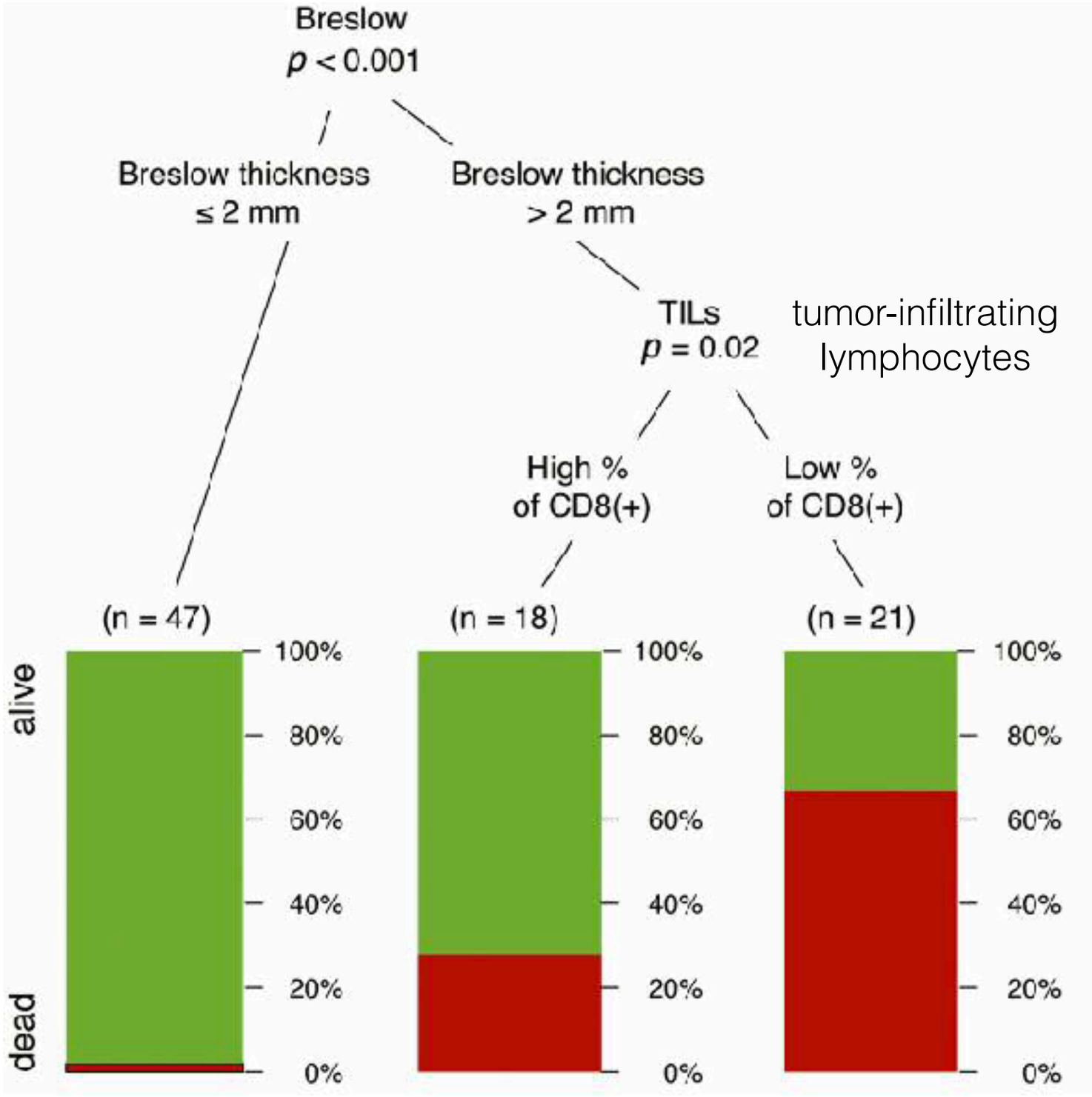
Cellular Constituents of Immune Escape
within the Tumor Microenvironment,
Cancer Research

Jako zbudować ,,mały” model predykcyjny?

Przykład:
Współpraca z Dolnośląskim Centrum Onkologii

Małe dane := dane mieszczą się w jednym pliku Excela

CSOS tygodnie 5 years	DFS tygodnie last_rec 5y (bez nawrctu)	nawrót lub czerniakowy zgon	nawrot	pM (obecność przerzutów odległych)	pN (przerzuty w węzłach chłonnych)	AJCC staging	SNLB_plus (z przerzutami)	regression	owrzodzenie (ULCER)	wiek w momencie diagnozy	płeć	breslow w mm	Breslow (pogrupowany w 4 grupy)	clark	hist_type	mitotic rate
224	224	0	0	0	0	2	0	0	0	58	0	1.1	2	3	1	0
48	48	0	0	0	0	4		0	1	79	1	5	4	4	2	2
186	186	0	0	0	0	1	0	0	0	70	1	0.5	1	2	1	0
158	155	1	1	0	1	3	1	1	1	55	1	3	3	4	3	1
173	89	1	1	0	0	3		0	0	26	0	4	3	3	2	2
226	226	0	0	0	0	1		0	0	61	0	0.4	1	2	1	0
260	260	0	0	0	0	1	0	0	0	30	1	0.75	1	3	1	0
260	260	0	0	0	0	4		0	1	49	1	9	4	4	1	2
258	258	0	0	0	0	3	0	0	1	45	0	4	3	4	1	1
166	166	1	0	0	0	4	0	0	1	61	0	9	4	4	2	1
260	260	0	0	0	0	4	0	0	1	39	0	5	4	3	3	1
204	204	0	0	0	0	1		0	0	71	0	1	1	2	1	0
247	247	0	0	0	0	3		0	1	77	1	4	3	5	2	1
138	138	0	0	0	0	3		0	1	50	1	2.2	3	3	2	2
260	260	0	0	0	1	2	1	0	1	57	1	2	2	4	1	0
170	170	0	0	0	0	1	0	0	0	69	0	0.6	1	3	1	0
257	257	0	0	0	0	2	0	0	1	40	1	2	2	3	2	1
231	231	0	0	0	0	3	0	0	1	57	1	3.5	3	4	1	2
118	118	0	0	0	0	3		0	1	75	1	3	3	4	1	0
99	99	0	0	0	0	2	0	0	0	73	0	1.8	2	3	1	0
202	97	1	1	1	0	4		0	1	76	1	8	4	5	2	2
260	260	0	0	0	1	3	1	0	0	72	1	4	3	4	1	1
200	180	1	1	0	0	3	0	0	0	71	0	2.3	3	3	2	0
231	231	0	0	0	0	2	0	0	1	73	1	1.9	2	3	1	1
190	190	0	0	0	0	2	0	0	0	53	1	1.2	2	3	1	1
35	35	1	0	1	1	3		1	1	76	1	2.5	3	4	2	2
186	186	0	0	0	0	1		0	0	71	1	0.5	1	2	1	0
251	251	0	0	0	0	2	0	0	0	33	1	1.2	2	2	1	0
243	243	0	0	0	0	2	0	0	0	33	1	1.1	2	3	1	0
5	5	1	0	0	1	4		1	1	37	0	6	4	4	1	2
167	167	0	0	0	0	1	0	0	0	55	1	0.6	1	3	1	0



BILLCD8 - A Multivariable Survival Model as a Simple and Clinically Useful Prognostic Tool to Identify High-risk Cutaneous Melanoma Patients. Donizy P, Biecek P, Halon A, Matkowski R. ***Anticancer Res. 2016 Sep;36(9):4739-47.***

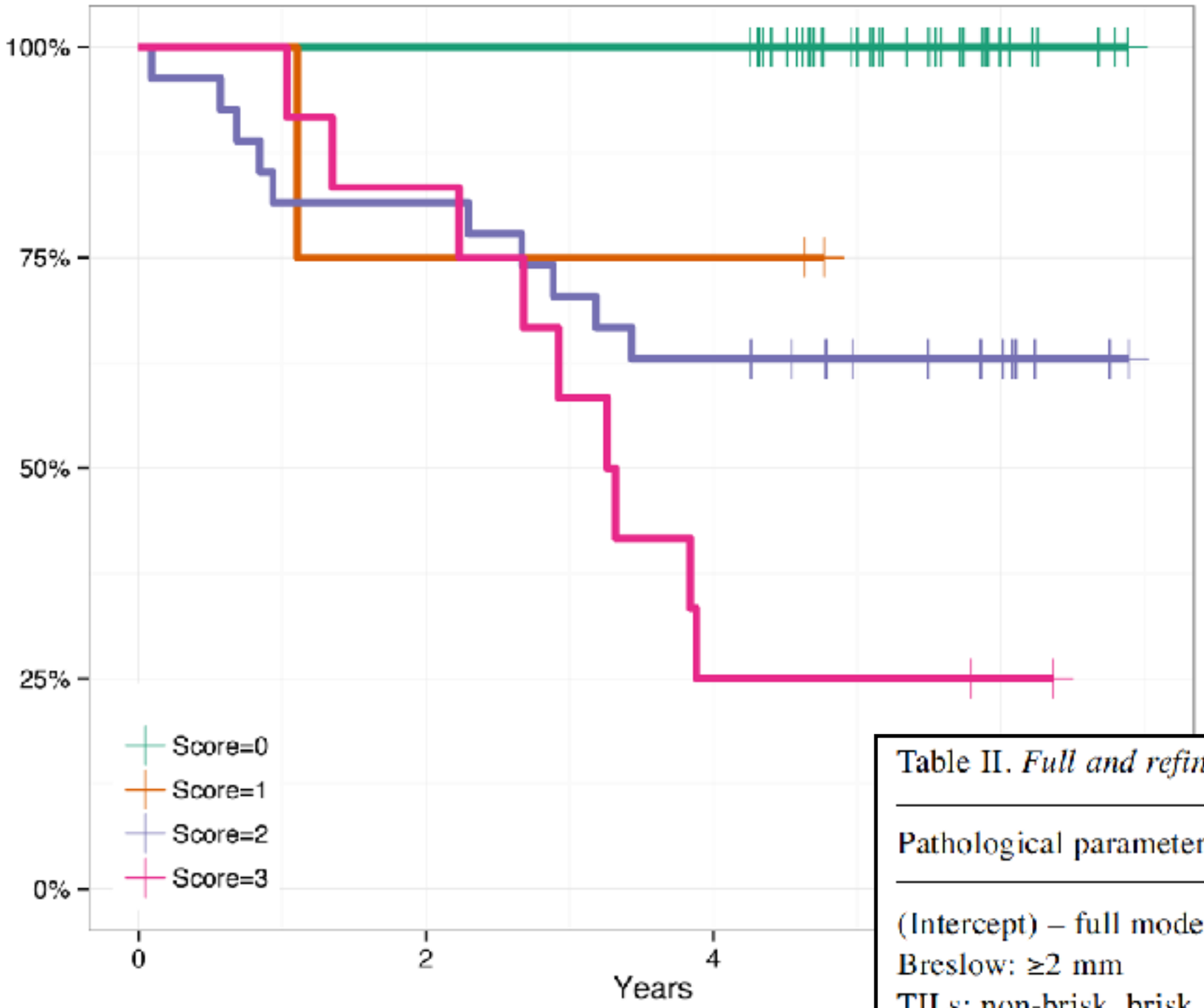


Table II. *Full and refined survival model.*

Pathological parameters	Estimate effect	*p*-Value
(Intercept) – full model	–2.27	0.12
Breslow: ≥2 mm	2.7	0.043*
TILs: non-brisk, brisk	–3.07	0.0041**
Mitotic rate: ≥1/mm^2	–1.72	0.24
LYI: present	1.38	0.074
Ulceration: present	1.85	0.11
Clark: IV, V	1.45	0.11
(Intercept) – refined model	–2.22	0.05*
Breslow: ≥2 mm	3.61	0.00089***
TILs: non-brisk, brisk	–2.02	0.0088**

BILLCD8 - A Multivariable Survival Model as a Simple and High-risk Cutaneous Melanoma Patients. Donizy P, Biecek **2016 Sep;36(9):4739-47.**

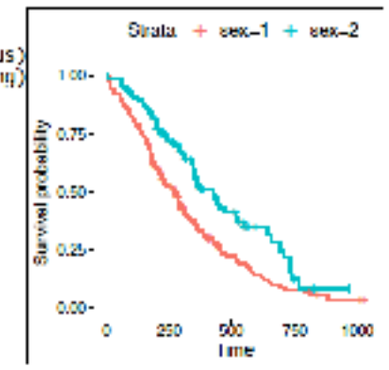
Creating Survival Plots

Informative and Elegant with *survminer*

Survival Curves

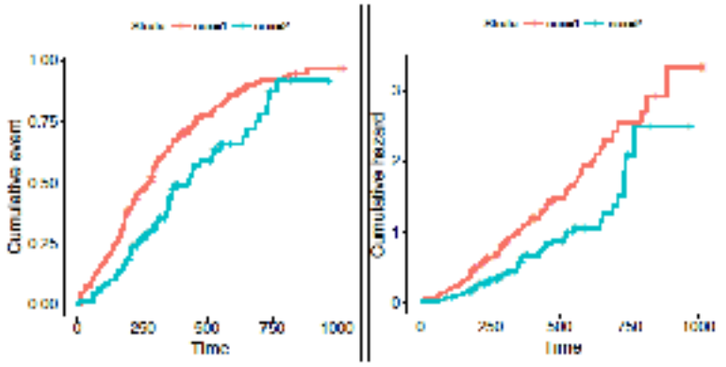
The **ggsurvplot()** function creates **ggplot2** plots from **survfit** objects.

```
library("survival")
fit <- survfit(Surv(time,status)
               ~ sex, data = lung)

class(fit)
## [1] "survfit"

library("survminer")
ggsurvplot(fit, data = lung)
```



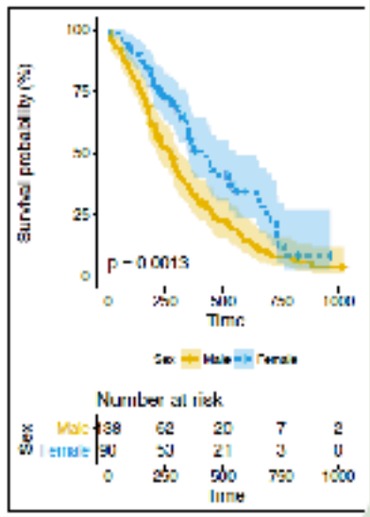
Use the **fun** argument to set the transformation of the survival curve. E.g. **"event"** for cumulative events, **"cumhaz"** for the cumulative hazard function or **"pct"** for survival probability in percentage.

```
ggsurvplot(fit, data = lung, fun = "event")
ggsurvplot(fit, data = lung, fun = "cumhaz")
```



With lots of graphical parameters you have full control over look and feel of the survival plots; position and content of the legend; additional annotations like p-value, title, subtitle.

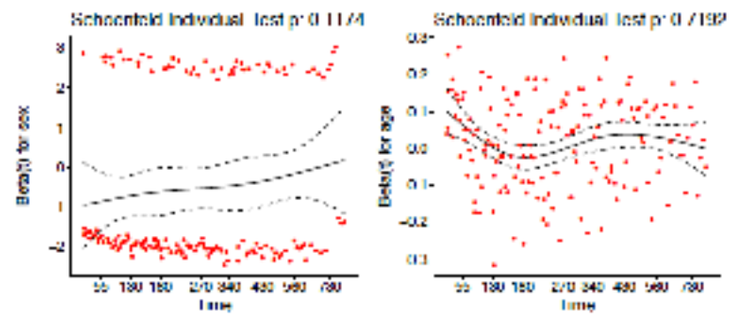
```
ggsurvplot(fit, data = lung,
    conf.int = TRUE,
    pval = TRUE,
    fun = "pct",
    risk.table = TRUE,
    size = 1,
    linetype = "strata",
    palette = c("#E7B800",
                "#2E9FDF"),
    legend = "bottom",
    legend.title = "Sex",
    legend.labs = c("Male",
                    "Female"))
```



Diagnostics of Cox Model

The function **cox.zph()** from **survival** package may be used to test the proportional hazards assumption for a Cox regression model fit. The graphical verification of this assumption may be performed with the function **ggcoxzph()** from the **survminer** package. For each covariate it produces plots with scaled Schoenfeld residuals against the time.

```
library("survival")
fit <- coxph(Surv(time, status) ~ sex + age, data = lung)
ftest <- cox.zph(fit)
ftest
##          rho chisq     p
## sex    0.1736 2.457 0.117
## age    0.0775 0.179 0.719
## GLOBAL     NA 2.651 0.266
library("survminer")
ggcoxzph(ftest)
```

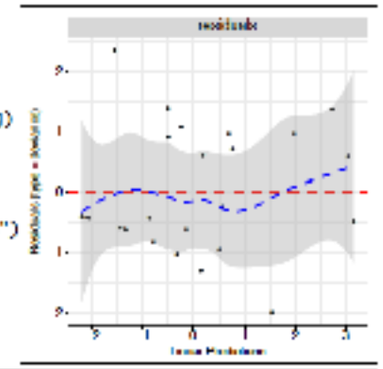


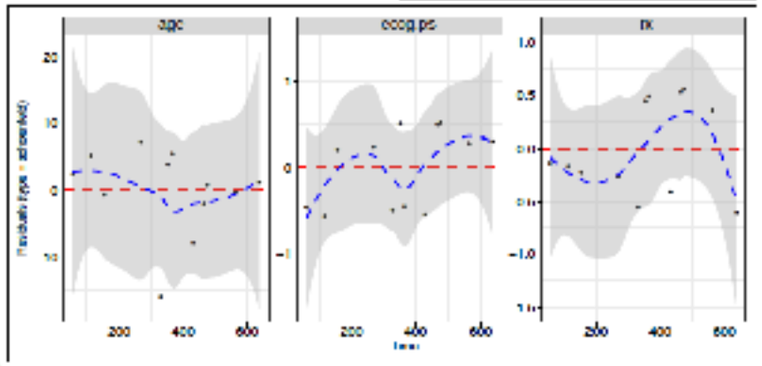
The function **ggcoxdiagnostics()** plots different types of residuals as a function of time, linear predictor or observation id. The type of residual is selected with **type** argument. Possible values are "martingale", "deviance", "score", "schoenfeld", "dfbeta", "dfbetas", and "scaledsch".

The **ox.scale** argument defines what shall be plotted on the OX axis. Possible values are "linear.predictions", "observation.id", "time".

Logical arguments **hline** and **sline** may be used to add horizontal line or smooth line to the plot.

```
library("survival")
library("survminer")
fit <- coxph(Surv(time, status) ~
             sex + age, data = lung)

ggcoxdiagnostics(fit,
    type = "deviance",
    ox.scale = "linear.predictions")

ggcoxdiagnostics(fit,
    type = "schoenfeld",
    ox.scale = "time")
```

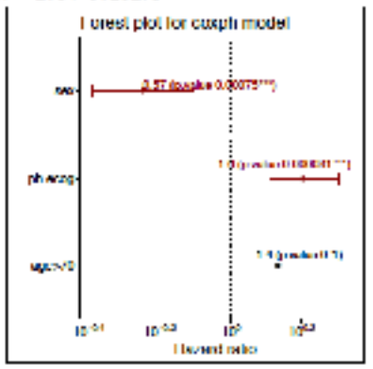




Summary of Cox Model

The function **ggforest()** from the **survminer** package creates a forest plot for a Cox regression model fit. Hazard ratio estimates along with confidence intervals and p values are plotted for each variable.

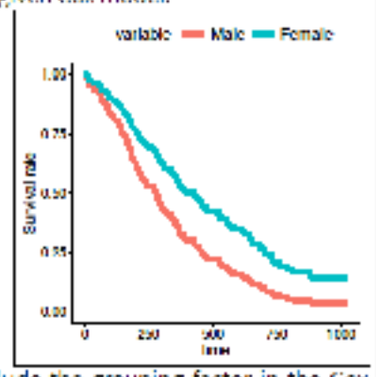
```
library("survival")
library("survminer")
lung$age <- ifelse(lung$age > 70, ">70","<= 70")
fit <- coxph( Surv(time, status) ~ sex + ph.ecog + age, data = lung)
fit

## Call:
## coxph(formula = Surv(time, status) ~ sex + ph.ecog + age, data = lung)
##
##              coef exp(coef) se(coef)     z        p
## sex     -0.5670    0.5672   0.1678 -3.37 0.000753
## ph.ecog  0.4700    1.6000   0.1130  4.16 3.1e-05
## age>70   0.3070    1.3594   0.1817  1.69 0.091075
##
## Likelihood ratio test=31.6  on
## n= 227, number of events= 164

ggforest(fit)
```



The function **ggcoxadjustedcurves()** from the **survminer** package plots Adjusted Survival Curves for Cox Proportional Hazards Model. Adjusted Survival Curves show how a selected factor influences survival estimated from a Cox model.

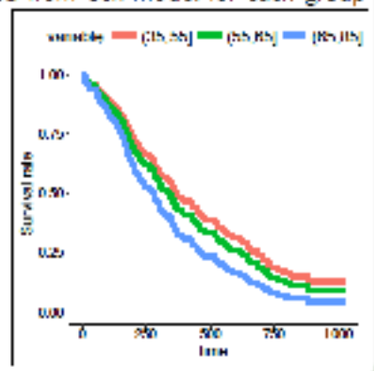
Note that these curves differ from Kaplan Meier estimates since they present expected ssurvival based on given Cox model.

```
library("survival")
library("survminer")

lung$sex <- ifelse(lung$sex == 1,
                   "Male", "Female")

fit <- coxph(Surv(time, status) ~
             sex + ph.ecog + age,
             data = lung)

ggcoxadjustedcurves(fit, data = lung,
             variable=lung$sex)
```



Note that it is not necessary to include the grouping factor in the Cox model. Survival curves are estimated from Cox model for each group defined by the factor independently.

```
lung$age3 <- cut(lung$age,
                 c(35,55,65,85))

ggcoxadjustedcurves(fit, data = lung,
             variable = lung$age3)
```



This onepager presents the **survminer** package [Alboukadel Kassambara, Marcin Kosinski 2017] in version 0.3.
See https://github.com/kassambara/survminer/ for more details.

BILLCD8 - A Multivariable Survival Model as a Simple and Clinically Useful Prognostic Tool to Identify High-risk Cutaneous Melanoma Patients. Donizy P, Biecek P, Halon A, Matkowski R. ***Anticancer Res. 2016 Sep; 36(9):4739-47.***

Ductal carcinoma in situ on stereotactic biopsy of suspicious breast microcalcifications: Expression of SPARC (Secreted Protein, Acidic and Rich in Cysteine) can predict postoperative invasion. Szynglarewicz B, Kasprzak P, Donizy P, Biecek P, Halon A, Matkowski R. ***J Surg Oncol. 2016 Oct;114(5):548-556.***

Golgi-Related Proteins GOLPH2 (GP73/GOLM1) and GOLPH3 (GOPP1/MIDAS) in Cutaneous Melanoma: Patterns of Expression and Prognostic Significance. Donizy P, Kaczorowski M, Biecek P, Halon A, Szkudlarek T, Matkowski R. ***Int J Mol Sci. 2016 Oct 1;17(10).***

Screen-detected ductal carcinoma in situ found on stereotactic vacuum-assisted biopsy of suspicious microcalcifications without mass: radiological-histological correlation. Szynglarewicz B, Kasprzak P, Biecek P, Halon A, Matkowski R. ***Radiol Oncol. 2016 Apr 23;50(2):145-52.***

HER-2 expression in immunohistochemistry has no prognostic significance in gastric cancer patients. Halon A, Donizy P, Biecek P, Rudno-Rudzinska J, Kielan W, Matkowski R. ***ScientificWorldJournal. 2012;2012:941259.***

Prognostic role of c-met expression in breast cancer patients. Gisterek I, Lata E, Halon A, Matkowski R, Szelachowska J, Biecek P, Kornafel J. ***Rep Pract Oncol Radiother. 2011 Jun 8;16(5):173-7.***

Correlation between hepatocyte growth factor receptor and vascular endothelial growth factor-A in breast carcinoma. Gisterek I, Matkowski R, Suder E, Hałoń A, Szelachowska J, Łata E, Łacko A, Biecek P, Kornafel J. ***Folia Histochem Cytobiol. 2010 Jan 1;48(1):78-83.***

Serum vascular endothelial growth factors a, C and d in human breast tumors. Gisterek I, Matkowski R, Lacko A, Sedlaczek P, Szewczyk K, Biecek P, Halon A, Staszek U, Szelachowska J, Pudelko M, Bebenek M, Harlozinska-Szmyrka A, Kornafel J. ***Pathol Oncol Res. 2010 Sep;16(3):337-44.***

BILLCD8 - A Multivariable Survival Model as a Simple and Clinically Useful Prognostic Tool to Identify High-risk Cutaneous Melanoma Patients. Donizy P, Biecek P, Halon A, Matkowski R. ***Anticancer Res. 2016 Sep; 36(9):4739-47.***

Ductal carcinoma in situ on stereotactic biopsy of suspicious breast microcalcifications: Expression of SPARC (Secreted Protein, Acidic and Rich in Cysteine) can predict postoperative invasion. Szynglarewicz B, Kasprzak P, Donizy P, Biecek P, Halon A, Matkowski R. ***J Surg Oncol. 2016 Oct;114(5):548-556.***

Golgi-Related Proteins GOLPH2 (GP73/GOLM1) and GOLPH3 (GOPP1/MIDAS) in Cutaneous Melanoma: Patterns of Expression and Prognostic Significance. Donizy P, Kaczorowski M, Biecek P, Halon A, Szkudlarek T, Matkowski R. ***Int J Mol Sci. 2016 Oct 1;17(10).***

Screen-detected ductal carcinoma in situ found on stereotactic vacuum-assisted biopsy of suspicious microcalcifications without mass: radiological-histological correlation. Szynglarewicz B, Kasprzak P, Biecek P, Halon A, Matkowski R. ***Radiol Oncol. 2016 Apr 23;50(2):145-52.***

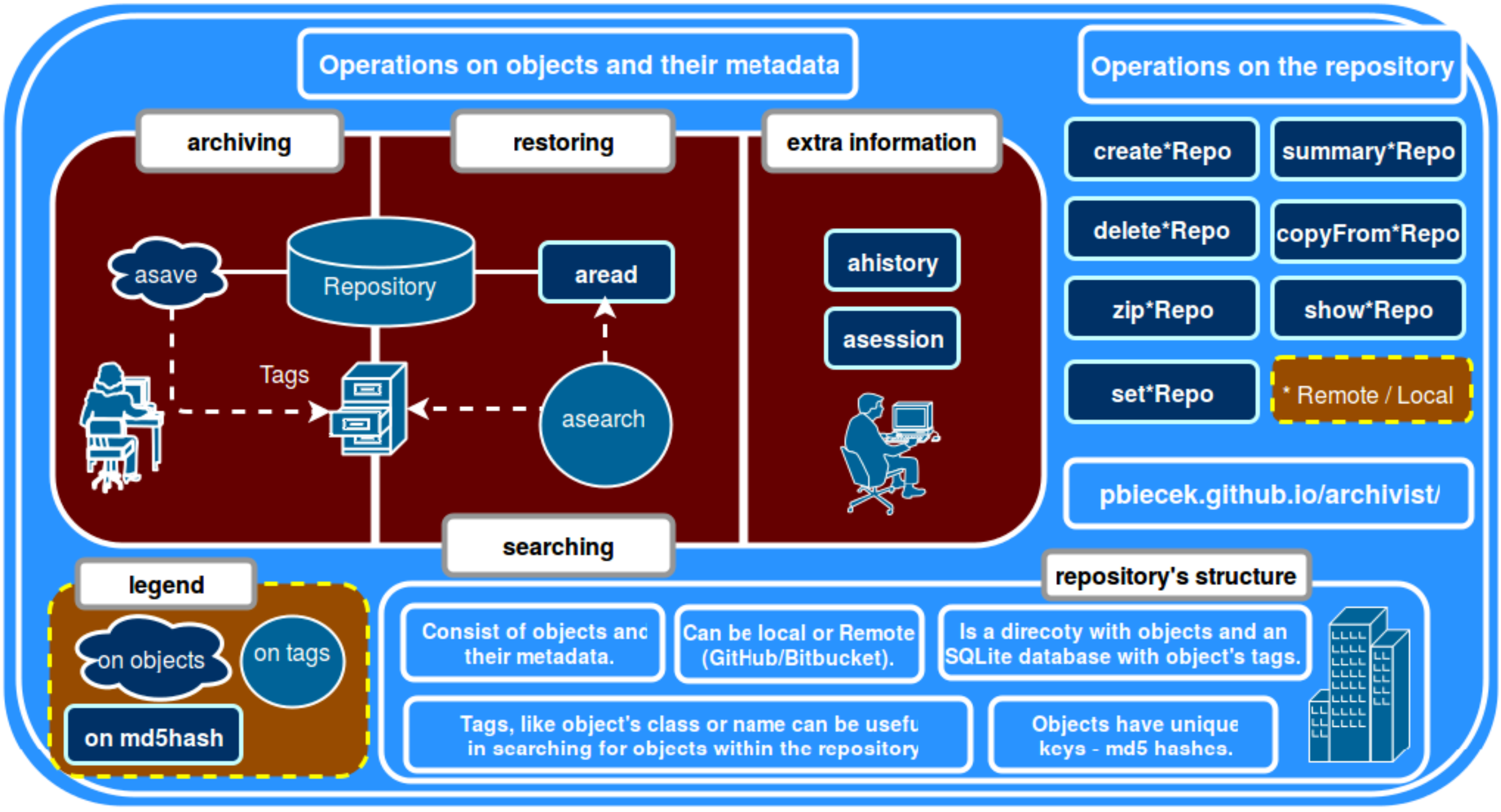
HER-2 expression in immunohistochemistry has no prognostic significance in gastric cancer patients. Halon A, Donizy P, Biecek P, Rudno-Rudzinska J, Kielan W, Matkowski R. ***ScientificWorldJournal. 2012;2012:941259.***

Prognostic role of c-met expression in breast cancer patients. Gisterek I, Lata E, Halon A, Matkowski R, Szelachowska J, Biecek P, Kornafel J. ***Rep Pract Oncol Radiother. 2011 Jun 8;16(5):173-7.***

Correlation between hepatocyte growth factor receptor and vascular endothelial growth factor-A in breast carcinoma. Gisterek I, Matkowski R, Suder E, Hałoń A, Szelachowska J, Łata E, Łacko A, Biecek P, Kornafel J. ***Folia Histochem Cytobiol. 2010 Jan 1;48(1):78-83.***

Serum vascular endothelial growth factors a, C and d in human breast tumors. Gisterek I, Matkowski R, Lacko A, Sedlaczek P, Szewczyk K, Biecek P, Halon A, Staszek U, Szelachowska J, Pudelko M, Bebenek M, Harlozinska-Szmyrka A, Kornafel J. ***Pathol Oncol Res. 2010 Sep;16(3):337-44.***

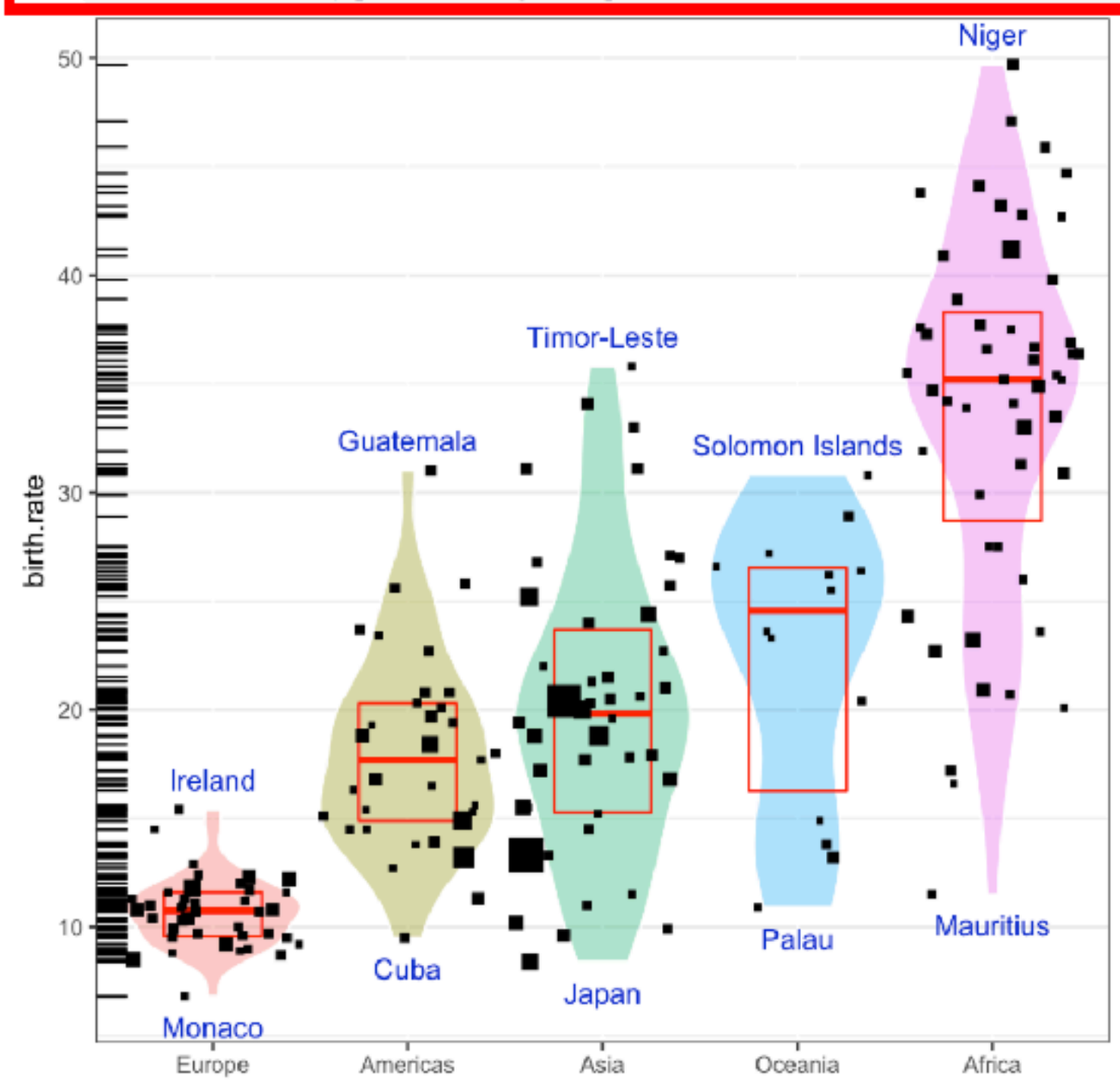
archivist - zarządzanie obiektami



https://github.com/pbiecek/archivist

```
ggplot(countries, aes(x=continent, y=birth.rate, label=country)) +
  geom_violin(scale="width", aes(fill=continent), color="white", alpha=0.4) +
  stat_summary(fun.data = "q3", geom = "crossbar",
               colour = "red", width = 0.5) +
  geom_jitter(aes(size=(population)^0.9),position=position_jitter(width = .45, height = 0),
              shape=15) +
  geom_rug(sides = "l") +
  geom_text(data=countriesMin, vjust=2, color="blue3") +
  geom_text(data=countriesMax, vjust=-1, color="blue3") +
  theme_bw() + xlab("") + theme(legend.position="none", panel.grid.major.x = element_line(color="whi
te"))
```

```
Load: archivist::aread('pbiecek/Eseje/arepo/ba7f58fafe7373420e3ddce039558140')
```

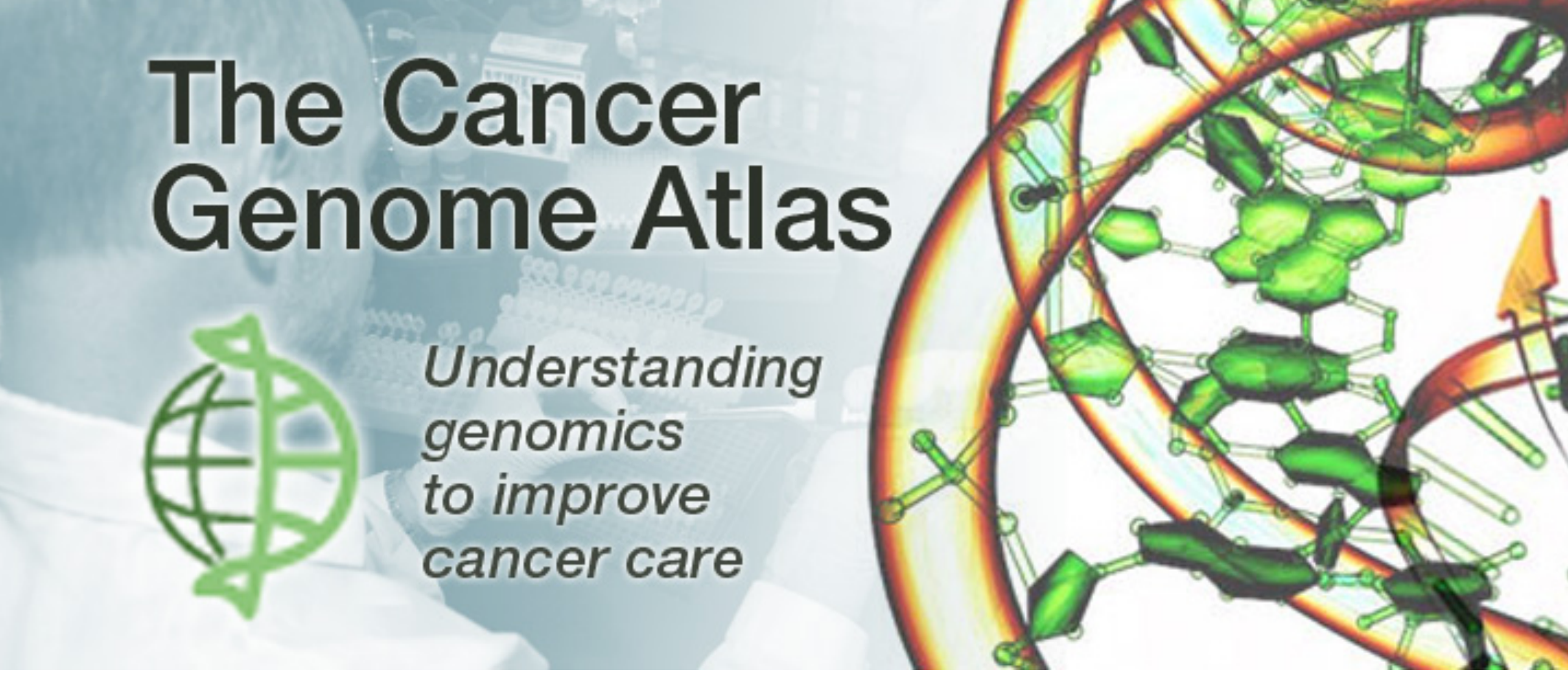


Jako zbudować ,,duży” model predykcyjny?

Przykład:

Współpraca z Wielkopolskim Centrum Onkologii

The Cancer Genome Atlas



The Cancer Genome Atlas (TCGA)

Multi-dimensional maps of the key genomic changes in **33 types of cancer**. **2.5 petabytes of data** describing tumour tissue and matched normal tissues from more than **11,000 patients** is **publicly available**. The data have contributed to more than a thousand studies of cancer by independent researchers.

Source: http://cancergenome.nih.gov/abouttcga/overview

Leukemia (LAML)

Lung adenocarcinoma (LUAD)

Lung squamous (LUSC)

Kidney (KIRC)

Bladder (BLCA)

Endometrial (UCEC)

Glioblastoma (GBM)

Head and neck (HNSC)

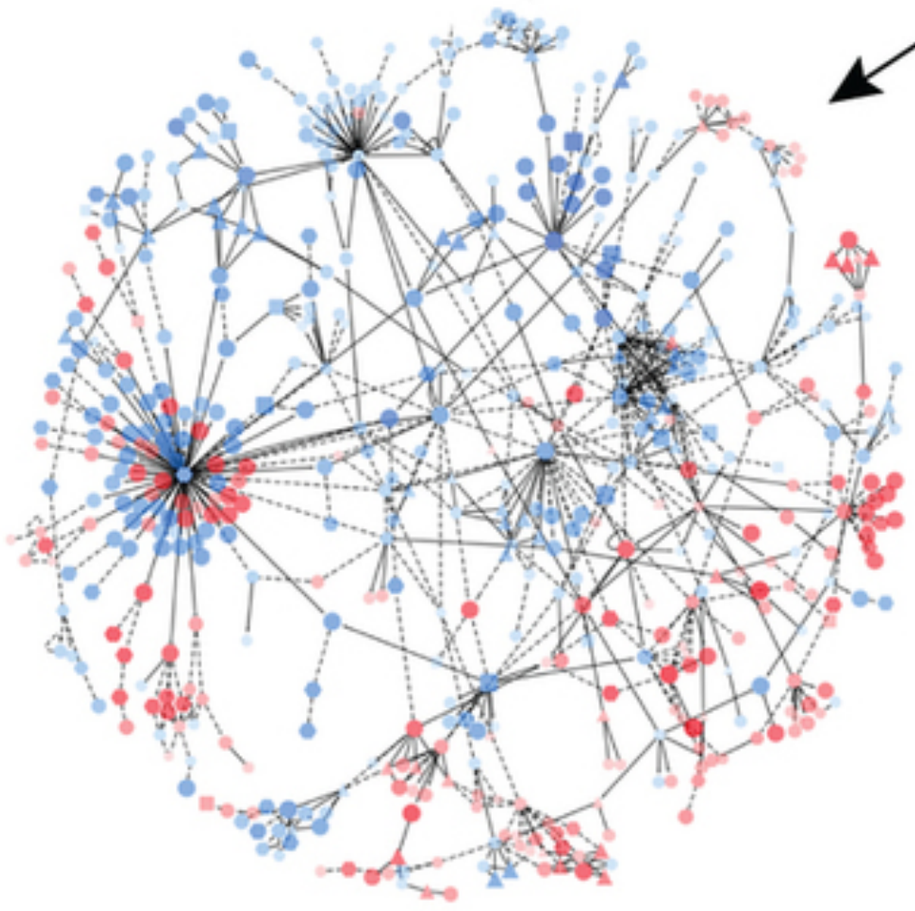
Breast (BRCA)

Ovarian (OV)

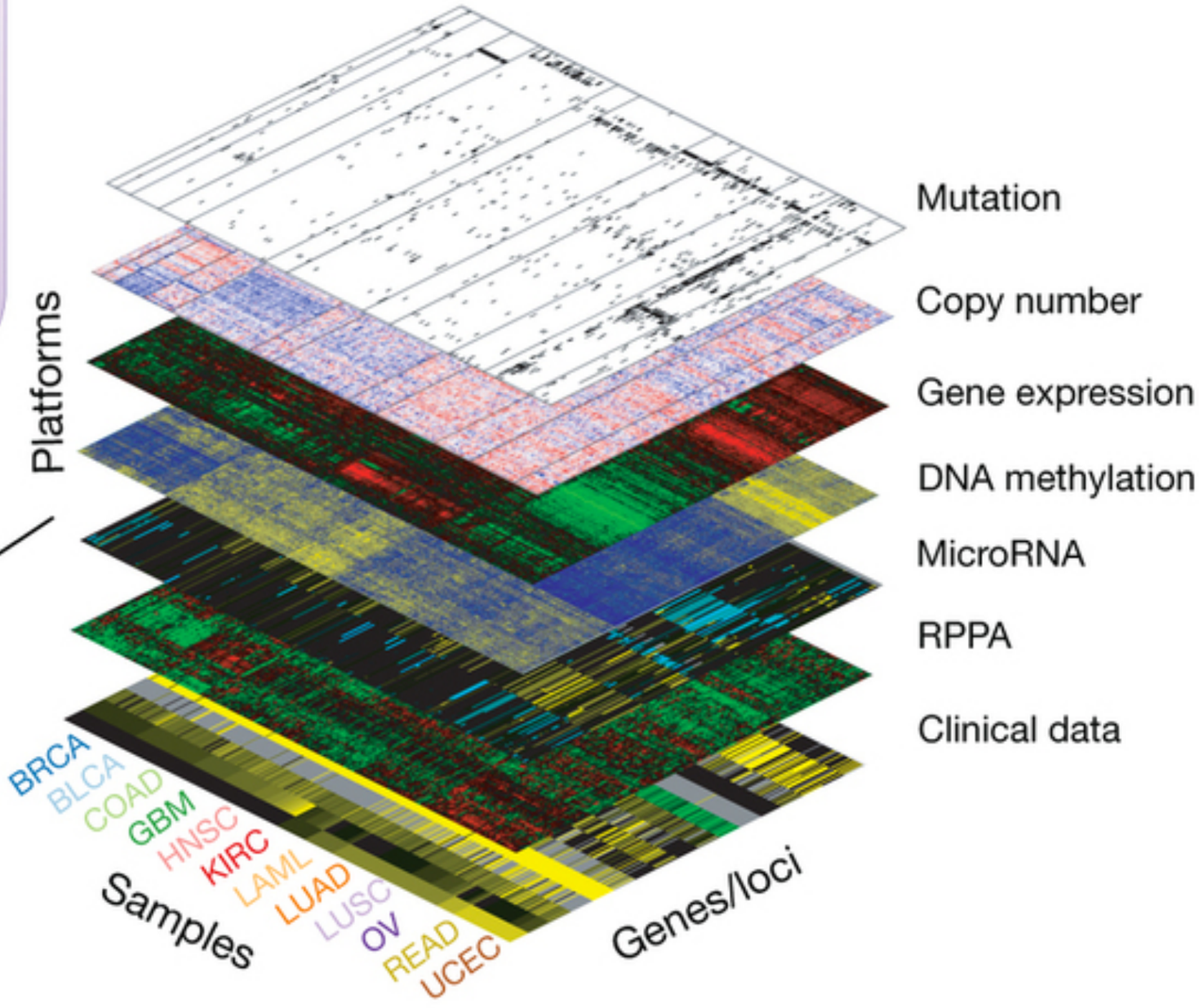
Colon (COAD)

Rectum (READ)

Thematic pathways



Omics characterizations



Kyle Chang et al, The Cancer Genome Atlas Pan-Cancer analysis project, Article in Nature Genetics 45(10):1113-20

Czy możemy wykorzystać ,,małe modele'' na dużych komputerach?

Nie! - *the curse of big data*.

- Modele typu drzewa decyzyjne nie wykorzystują dużej liczby cech. Aby z tym sobie poradzić można stosować bagging, boosting lub inne komitety klasyfikatorów.

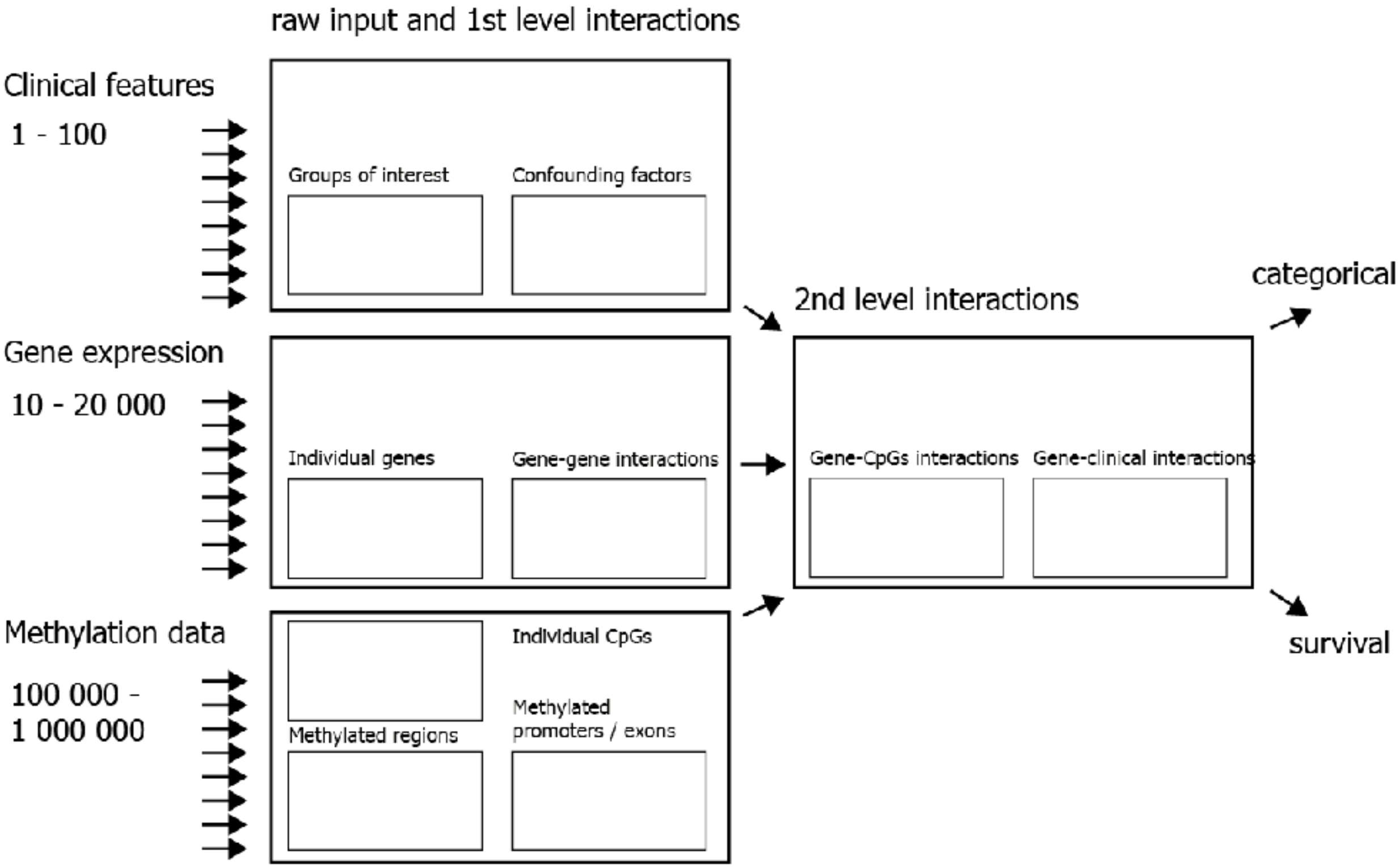
- Modele regresyjne maja dużą wariancję. Potrzebujemy bardziej stabilnych ocen parametrów, takich jak LASSO, regresja grzbietowa lub ,,sieci elastyczne''.

- Różne źródła danych różnią się bardzo ze względu na liczbę zmiennych - brak prostego rozwiązania.

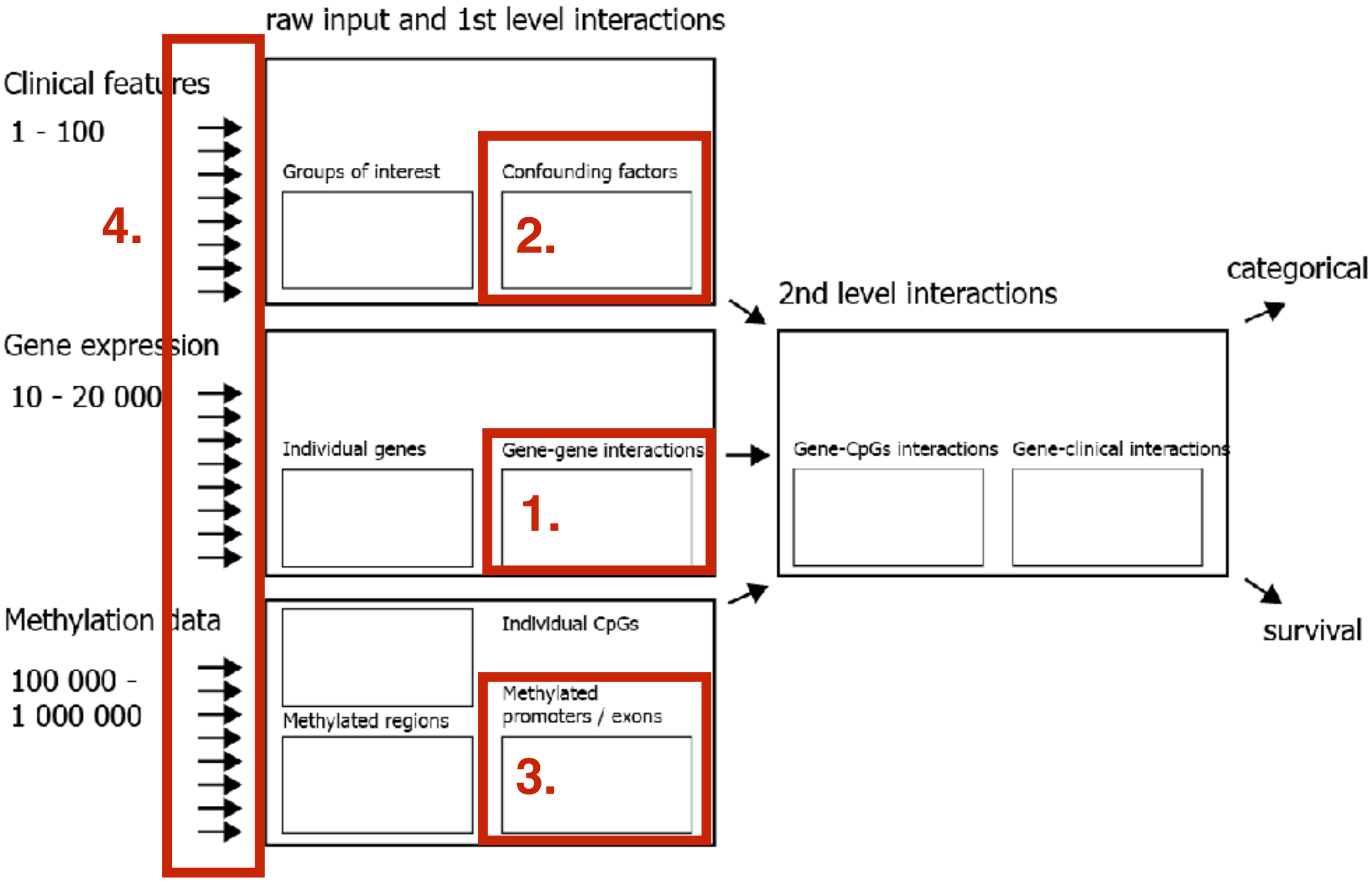
- Etykiety klas są silnie niezbalansowane - brak prostego rozwiązania, undersampling / oversampling czasem pomaga.

- …

Learning with Structure: MLGenSig (Machine Learning Genetic Signatures)



Learning with Structure: MLGenSig (Machine Learning Genetic Signatures)



Problem 1

Gene2Vec = word2vec for genes

- 20 000 genów -> 200 000 000 par genów (interakcje)
- Nie możemy testować 200 milionów hipotez !!!
- Jako wybrać kandydatury do testowania interakcji?

Problem 1

Gene2Vec = word2vec for genes

- 20 000 genów -> 200 000 000 par genów (interakcje)
- Nie możemy testować 200 milionów hipotez !!!
- Jako wybrać kandydatury do testowania interakcji?

- Algorytm Word2vec produkuje d-wymiarowe reprezentacje słów (d ~ 100).
- *Words that occur in similar contexts tend to have similar meanings (Harris, 1954; Firth, 1957; Deerwester et al., 1990). – If words have similar row vectors in a word–context matrix, then they tend to have similar meanings.*

You shall know a word by the company it keeps

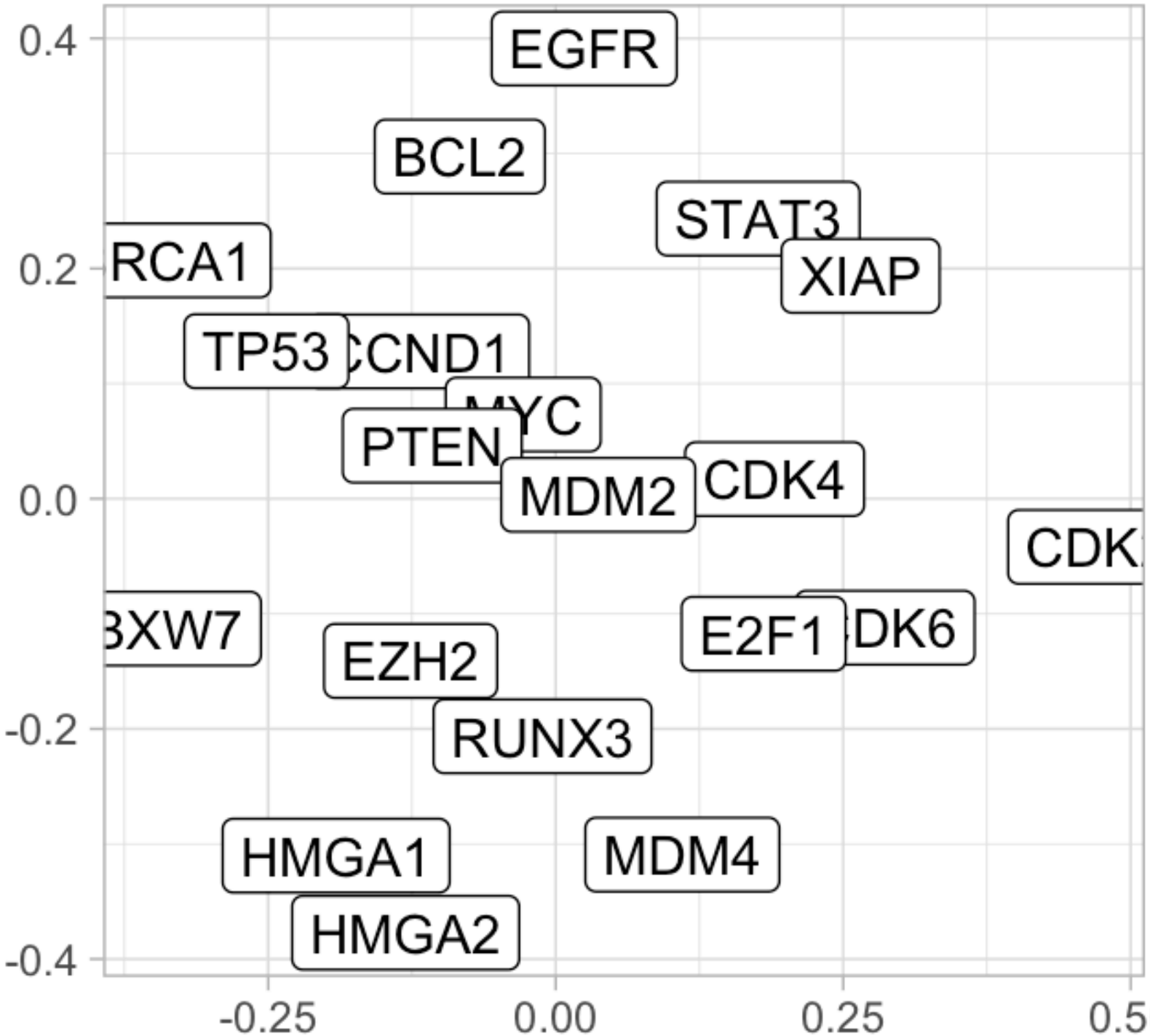
Problem 1

1. Pobierz streszczenia z bazy PUBMED [~25GB skompresowanych danych]

2. Wyznacz kontekst występowania słów i termów

3. Użyj GloVe lub innego algorytmu word2vec aby znaleźć 50-, 200, 600-wymiarową reprezentacje słów

4. Znajdź geny bliskie sob - kandydatury do badań interakcji

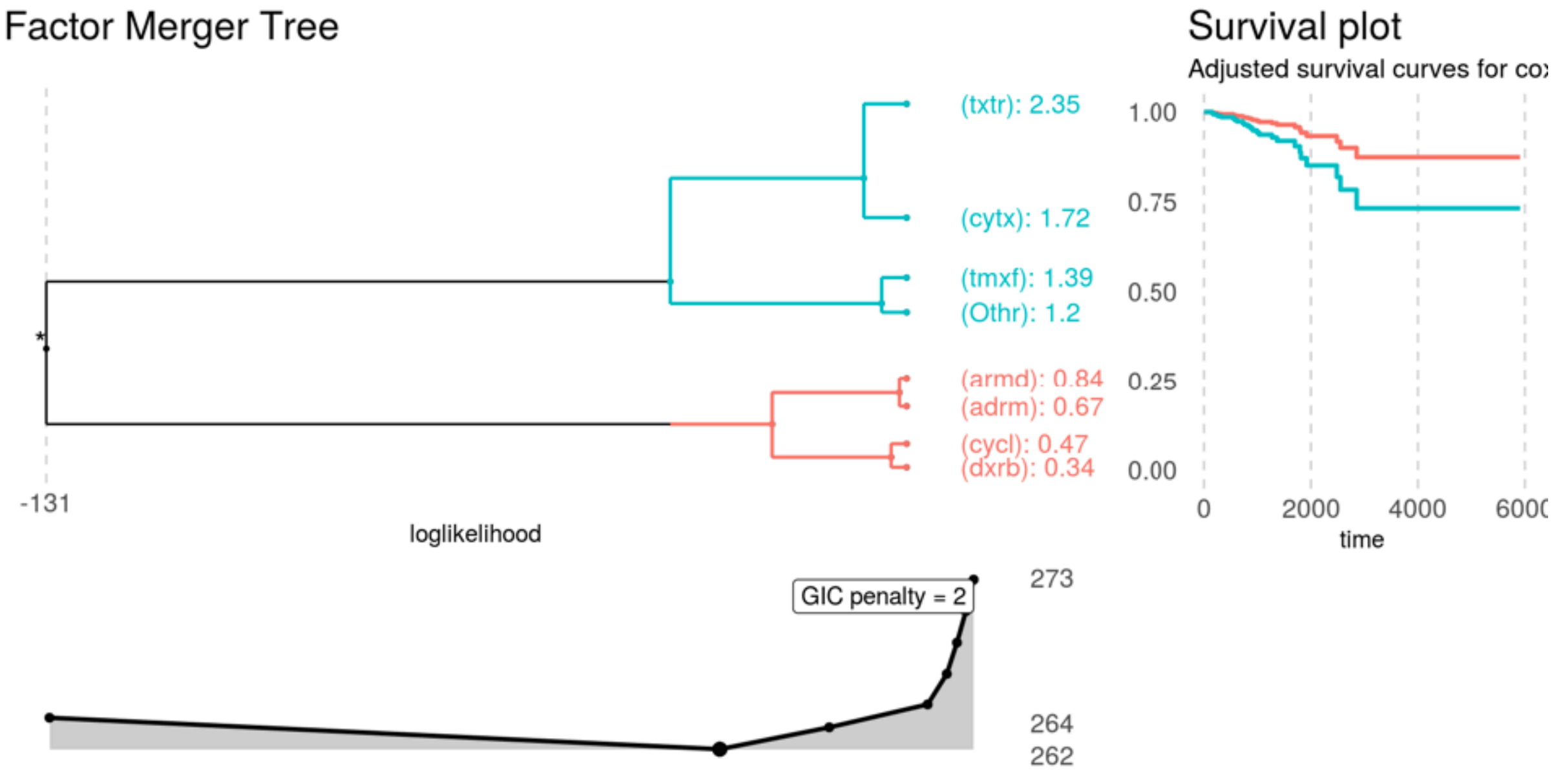


Project Lead: Kornel Kiełczewski

Problem 2

Project Lead:
Agnieszka Sitko

FactorMerger:
an extended hierarchical post-hoc testing



Problem 2

Project Lead:
Agnieszka Sitko

factorMerger:
an extended hierarchical post-hoc testing

FactorMerger - set of tools to support results from post hoc testing

Level fusing plot

The top-left plot shows level fusing paths (merging paths). With arguments **family=**, **show=**, **fuse=**, **spacing=**, one can select how to merge factors and what shall be presented on OX/OY axes.

Argument	Summary
panel = "all"	All panels
panel = "left"	Only left two panels
panel = "top"	Only top two panels
panel = "merging"	Merging path plot

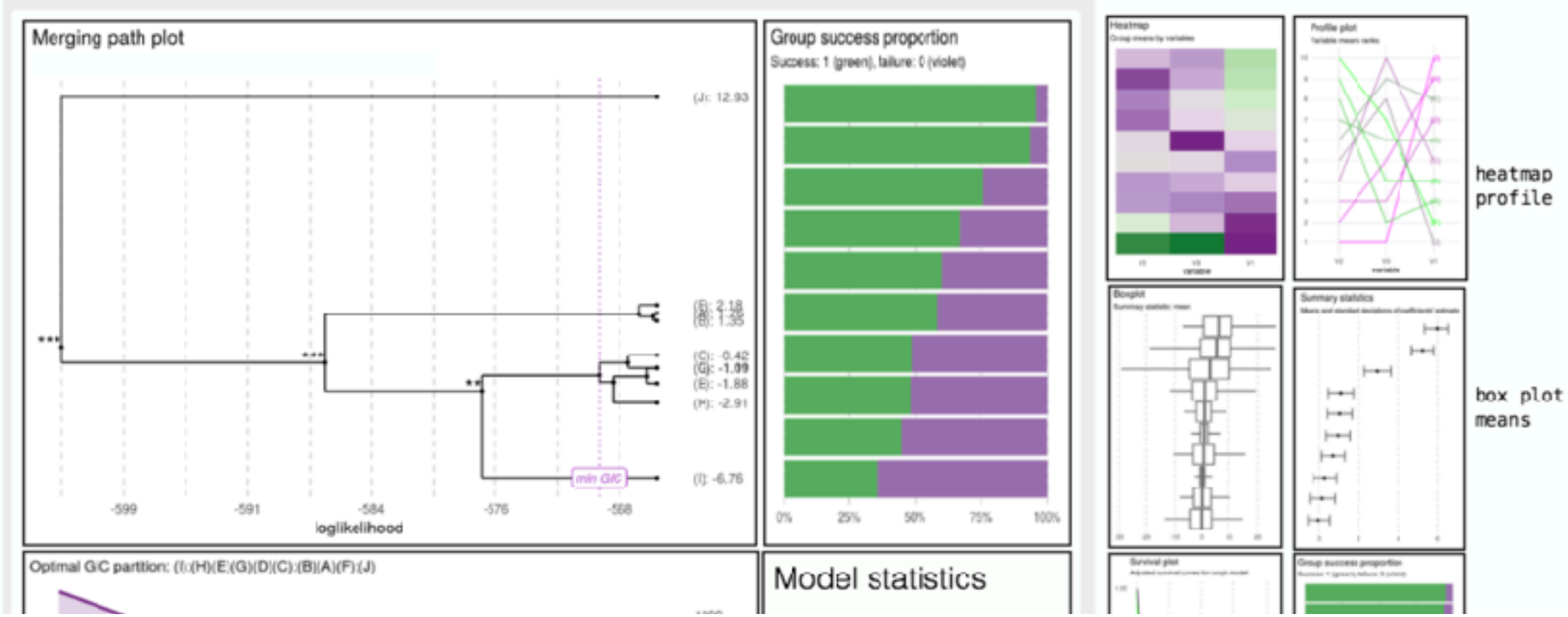
Argument	Summary
show = "likelihood"	Plot likelihood on OX axis
show = "p-value"	Plot p-values on OX axis
fuse = "all2all"	Compare all pairs of groups
fuse = "nearby"	Compare nearby groups
fuse = "cluster"	DMR4glm algorithm
spacing = "equidistant"	Levels equidistant on OY scale
spacing = "effects"	Levels according to their effects
family = "gaussian"	For one-dimensional Gaussian
family = "mgaussian"	For multi dimensional Gaussian
family = "binomial"	For binomial regression
family = "survival"	For Cox regression

Group summaries

The top-right panel shows group characteristics.

Use the parameter **summary=** to select the most suitable presentation.

Argument	Summary
summary = "heatmap"	For mgaussian
summary = "profile"	For mgaussian
summary = "boxplot"	For gaussian
summary = "means"	For gaussian
summary = "survival"	For Cox regression
summary = "proportions"	For binomial regression

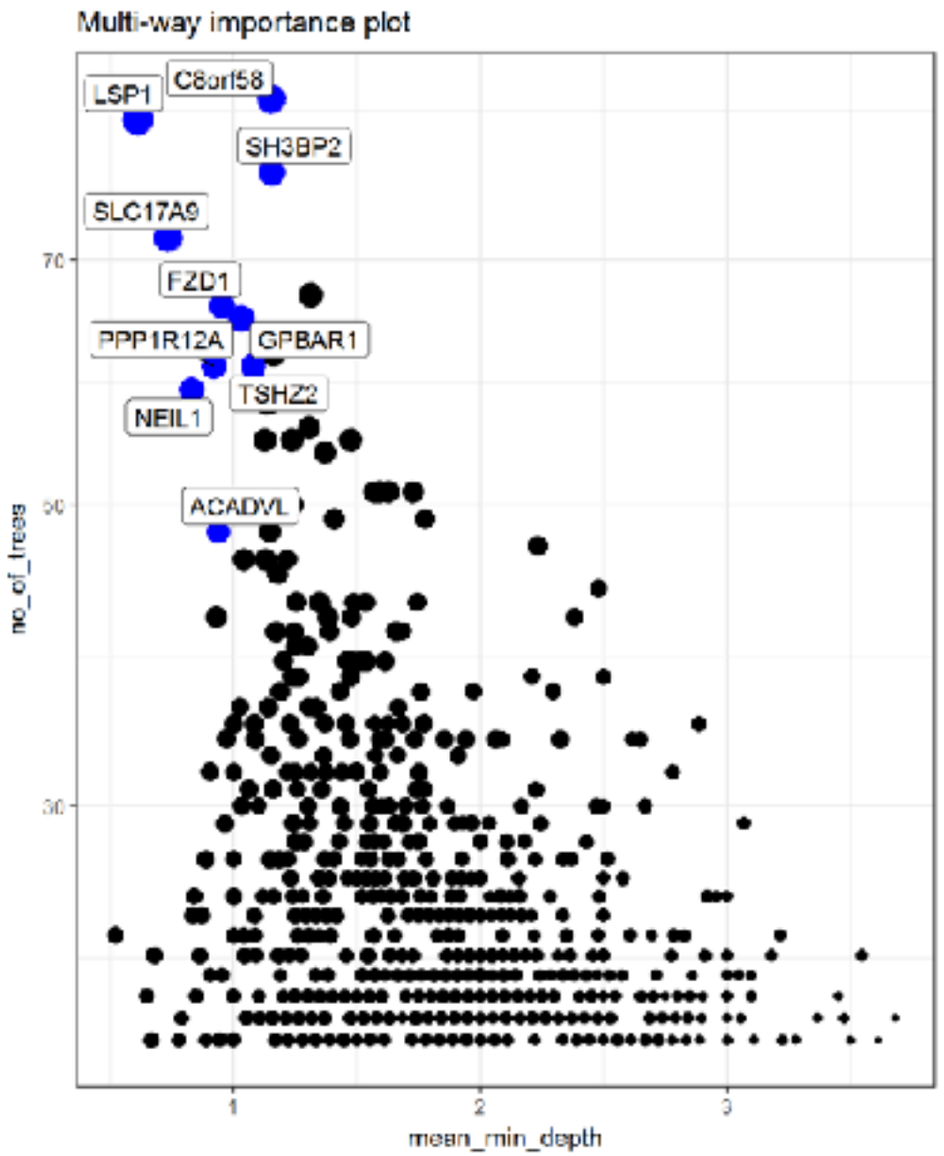


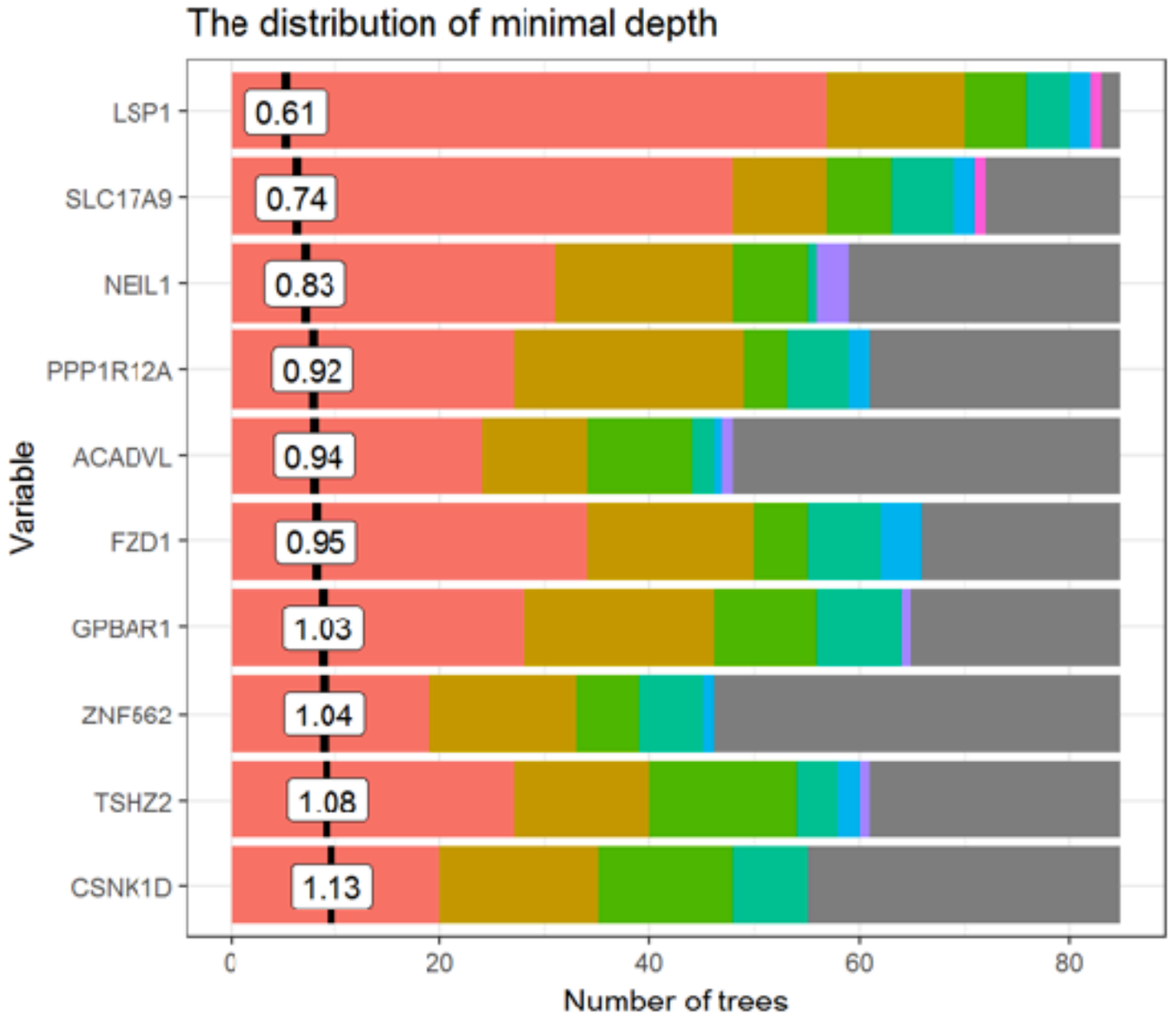
https://github.com/geneticsMiNIng/FactorMerger

Problem 3

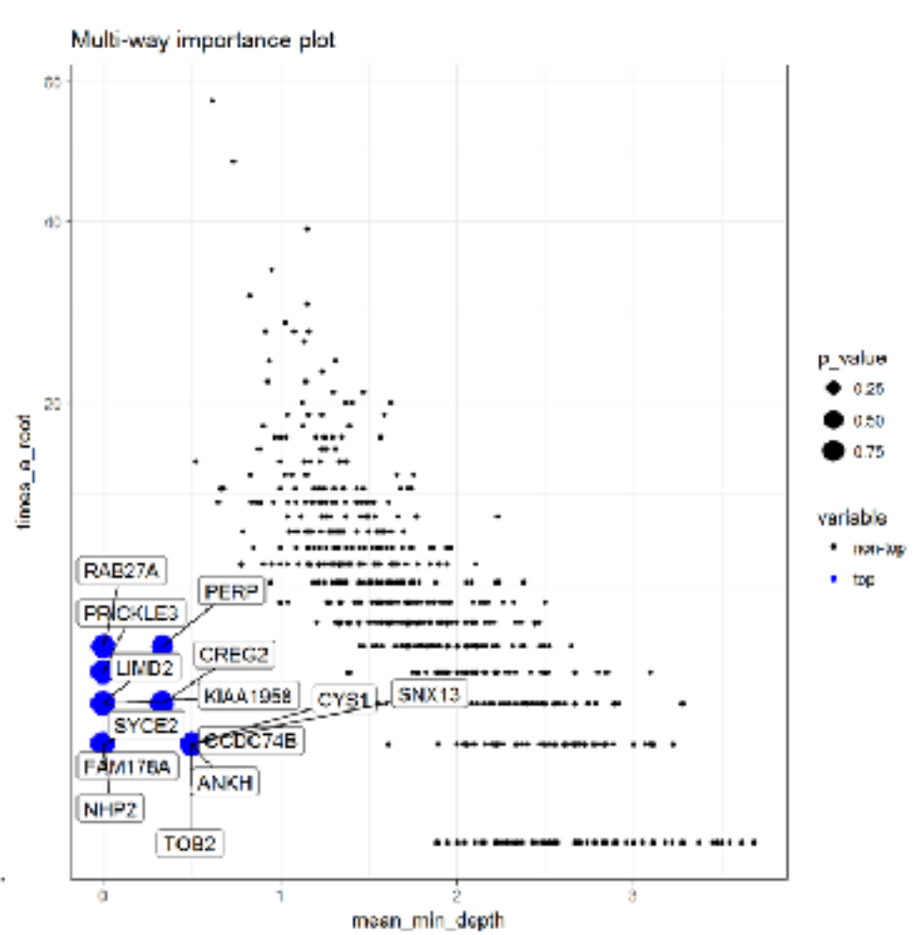
randomForestExplainer:
What is in the (random) forest?

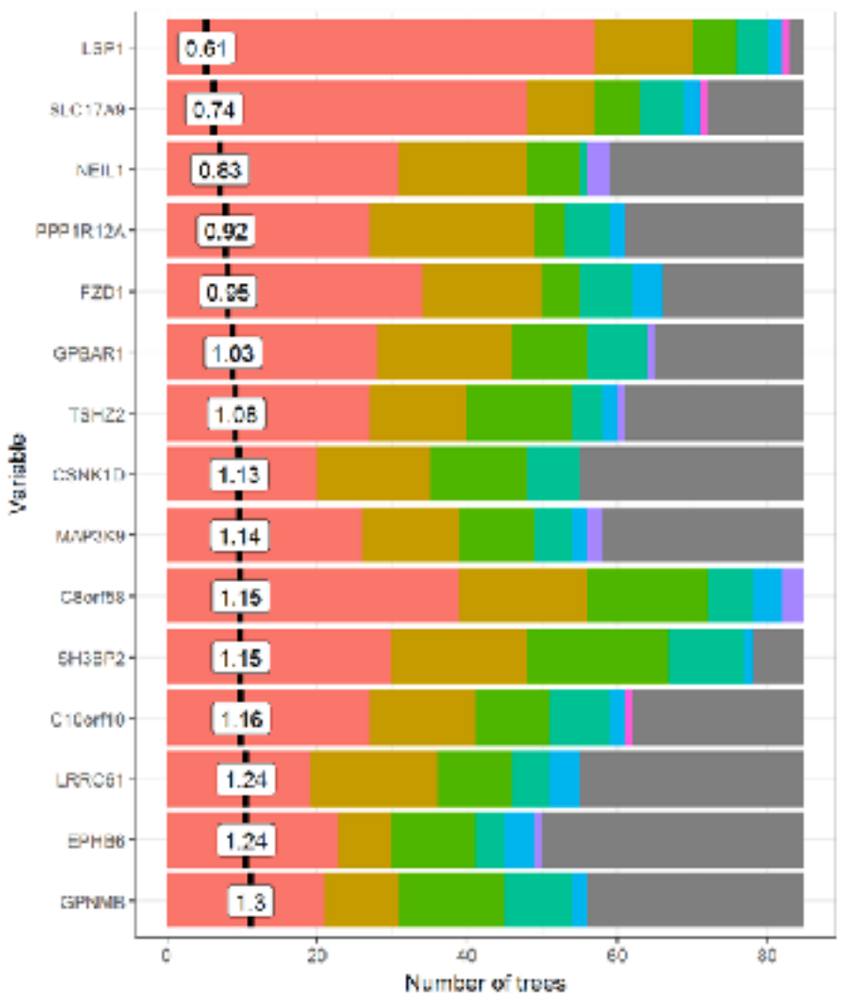
Project Lead:
Ola Paluszynska



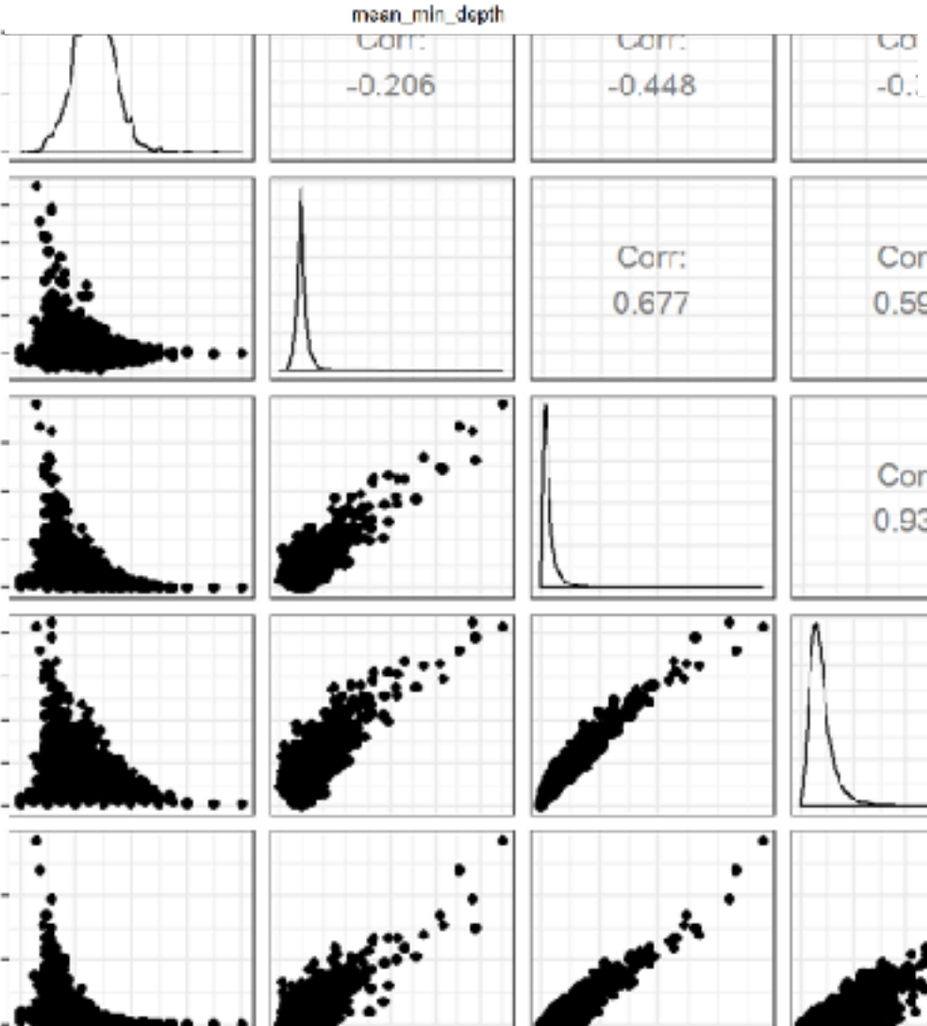


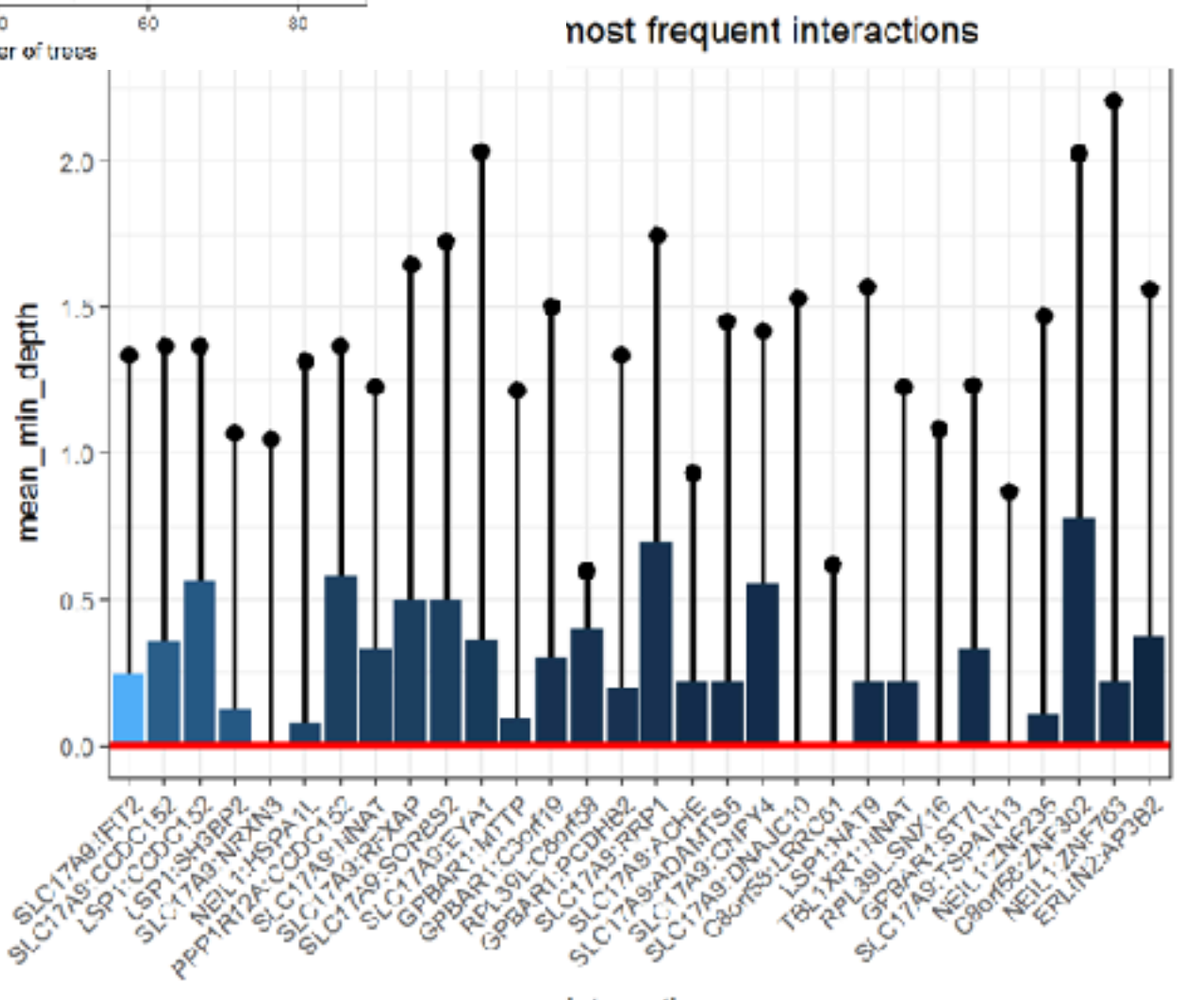
https://github.com/geneticsMiNIng/BlackBoxOpener





Project Lead: Ola Paluszynska





Problem 4

Ważność cech

z Random Forest

Cechy kliniczne

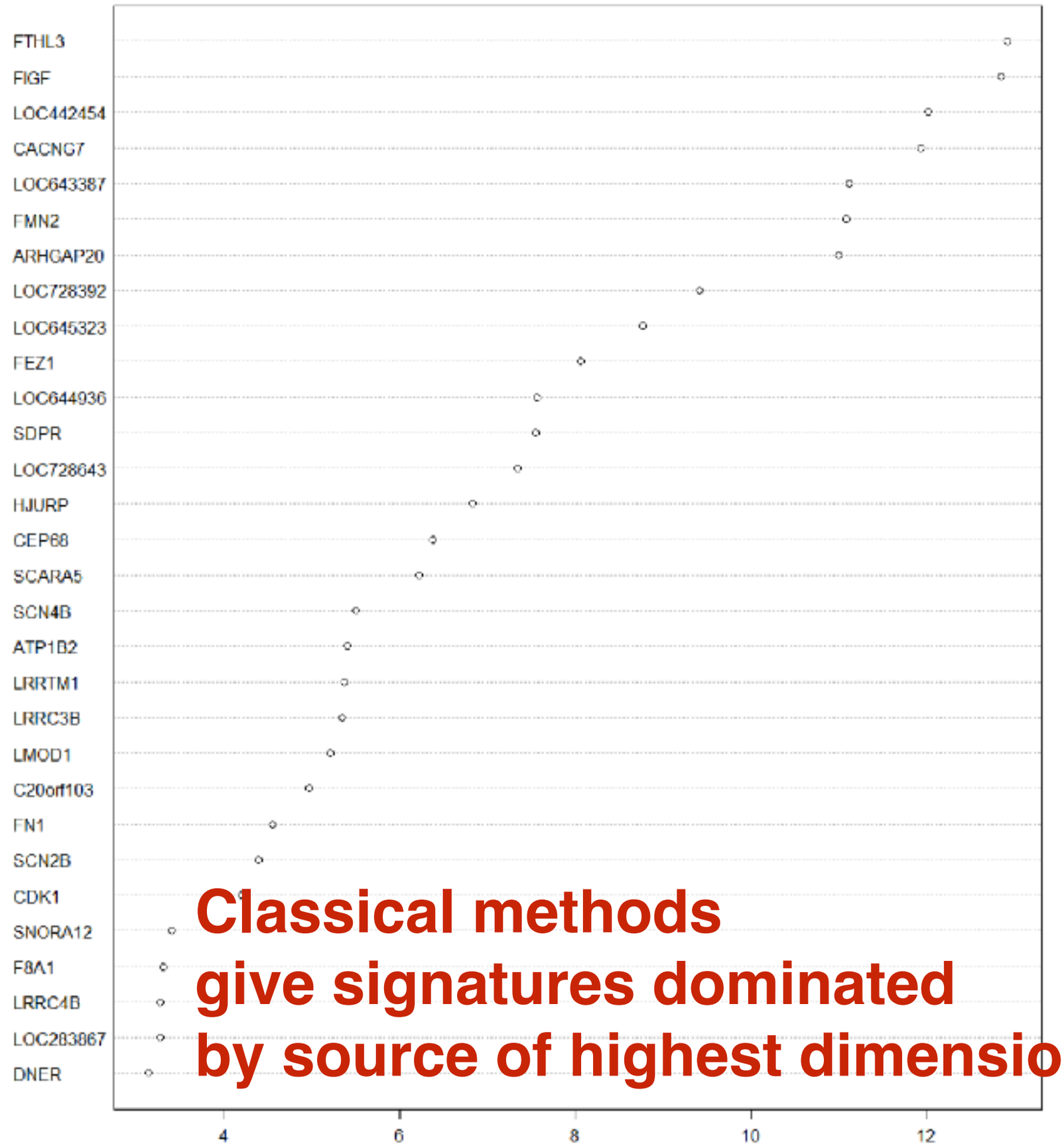
rozmiar: 1-100

Cechy mRNA

rozmiar: 100 - 20 000

Cechy metylacji

rozmiar: 10 000 - 1 000 000



Classical methods give signatures dominated by source of highest dimension

Problem 4 Modified Bayesian Information Criteria

$$BIC = -2 \log L(M_i, \hat{\theta}_i) + k_i \log(n)$$

$$mBIC = -2 \log L(M_i, \hat{\theta}_i) + k_i^{cli} p_{cli} + k_i^{gen} p_{gen} + k_i^{met} p_{met}$$

Dopasowanie modelu

kara za cechy kliniczne

kara za cechy mRNA

kara za cechy metylacyjne

W podobny sposób możemy dodawać więcej źródeł danych

Problem 4 Modified Bayesian Information Criteria

$$mBIC = -2 \log L(M_i, \hat{\theta}_i) + k_i^{cli} p_{cli} + k_i^{gen} p_{gen} + k_i^{met} p_{met}$$

Locating multiple interacting quantitative trait loci using robust model selection. Andreas Baierl, Andreas Futschik, Małgorzata Bogdan, Przemysław Biecek. ***Computational Statistics & Data Analysis, 51 (12) p. 6423–6434***

Extending the Modified Bayesian Information Criterion (mBIC) to Dense Markers and Multiple Interval Mapping. Małgorzata Bogdan, Florian Frommlet, Przemysław Biecek, Riyan Cheng, Jayanta K. Ghosh, R.W. Doerge. ***Biometrics 64 (4) p. 1162–1169 (2008)***

The R Package bgmm: Mixture Modeling with Uncertain Knowledge. Przemysław Biecek, Ewa Szczurek, Martin Vingron and Jerzy Tiuryn ***Journal of Statistical Software, 47(3), pp. 1-31 (2012)***

Problem 5

RTCGA

Factory of R packages with TCGA data

Home

Web Applicatio

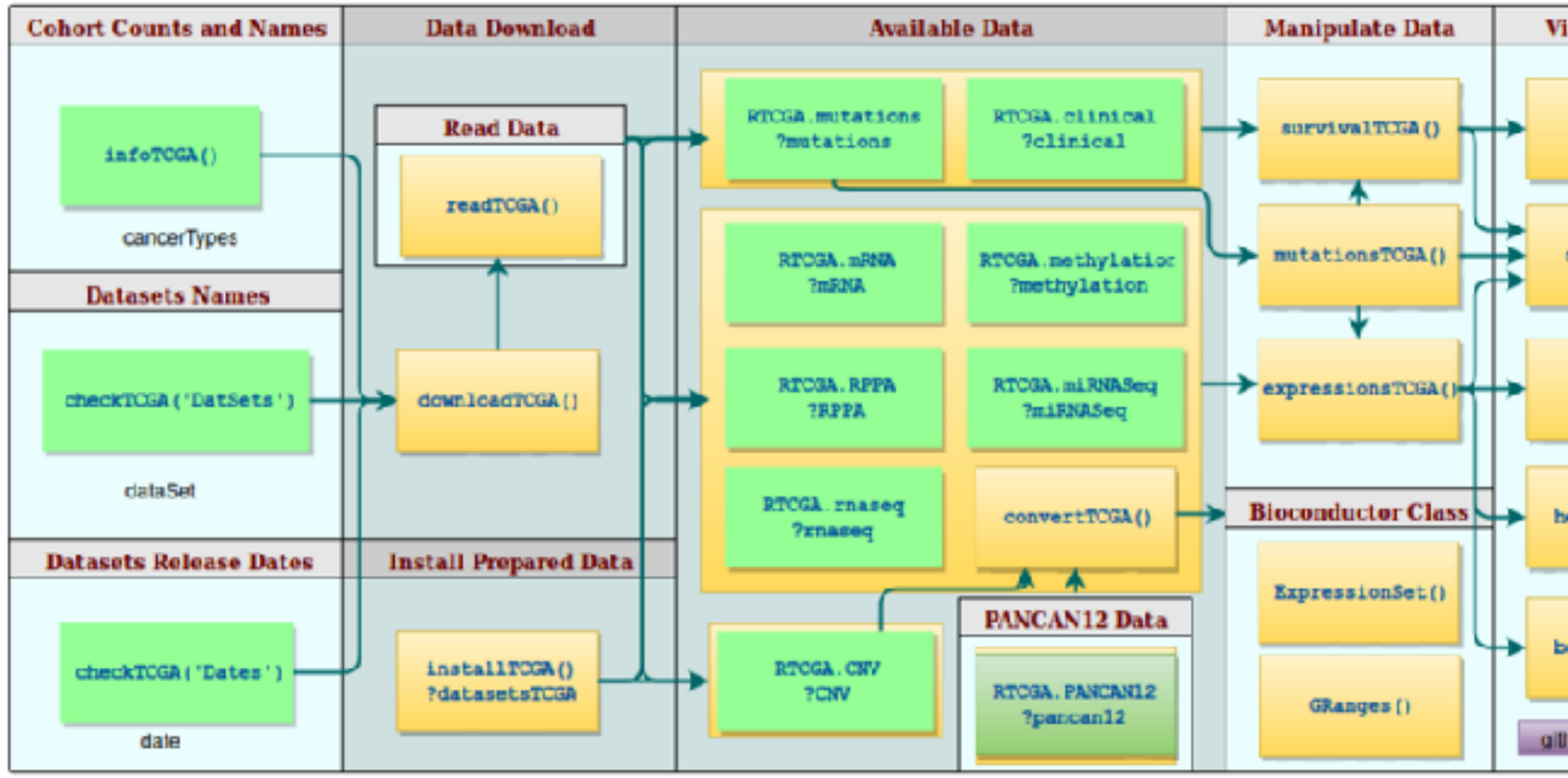
Cox Survival A

TCGA Data Dov

Visualizations

GitHub project

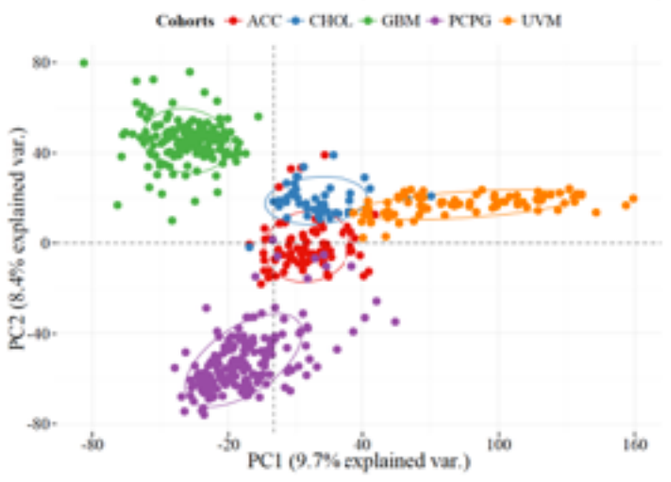
Workflow of RTCGA package



This can be reproduced in R with

```
library(RTCGA.rnaseq)
rnaseqBiplot(c("ACC", "CHOL", "GBM", "PCPG", "UVM"))
```

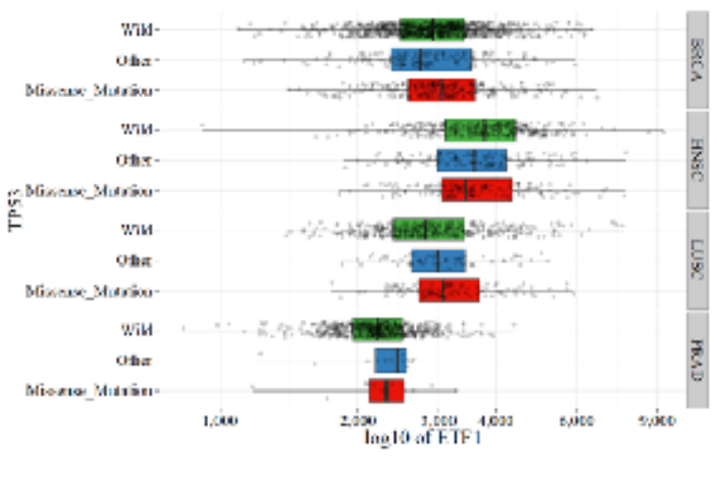
load to R with archivist::aread('MarcinKosinski/ISOBM/56')



This can be reproduced in R with

```
library(RTCGA.mutations);library(RTCGA.rnaseq)
mutationsBox(c('BRCA','HNSC','LUSC','PRAD'),'TP53','ETF1')
```

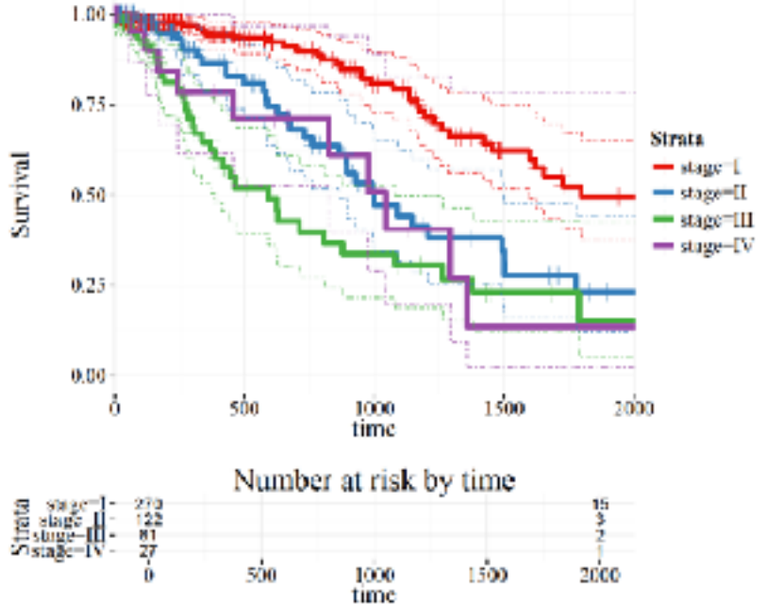
load to R with archivist::aread('MarcinKosinski/ISOBM/3b')



This can be reproduced in R with

```
library(RTCGA.clinical)
clinicalStageSurvival(LUAD.clinical, xlims = c(0,2000))
```

load to R with archivist::aread('MarcinKosinski/ISOBM/75')



Maintainer: Marcin Kosiński

Dziękuję za uwagę

1 czerwca o 12:00 otwarcie

MI DATA LAB