PATRONAT HONOROWY
JM REKTORA

SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE

Challenges in the genetic profiling of cancer patients with applications in The Cancer Genome Atlas project

Przemysław Biecek
Warsaw University of Technology, University of Warsaw

Partner

sas THE POWER TO KNOW.

Media Partners

Medexpress.pl

ITWIZ

medycyna praktyczna

The Fourth International Workshop • Advanced Analytics and Data Science

Few words about me

- Background in Software Engineering and Mathematical Statistics, graduated in both at Wrocław University of Technology,
- Research interests in Machine Learning, Data Visualisation and Molecular Human Genetics,
- MI^2 group – the bridge between Mathematics and Computer Science at Warsaw University of Technology (MiNI PW) and University of Warsaw (MIM UW),
- Team Leader of the Genetic Mining Group

  https://github.com/geneticsMiNIng

  We support molecular biologists and physicians in their research.

Case Study:

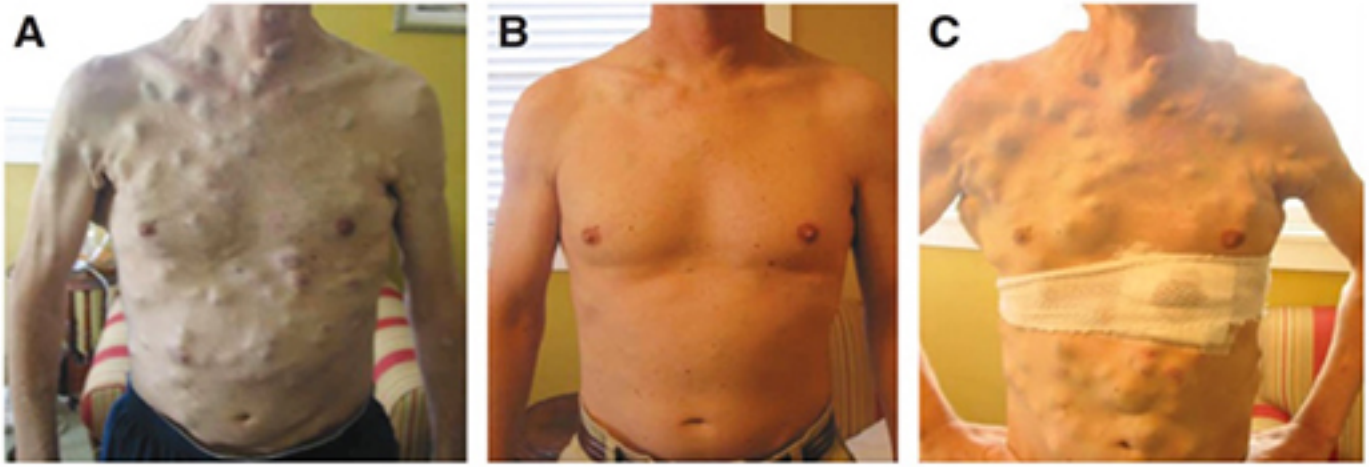
How to create a genetic signature that will score the risk of chemo-resistance for cancer patients.

Outline:

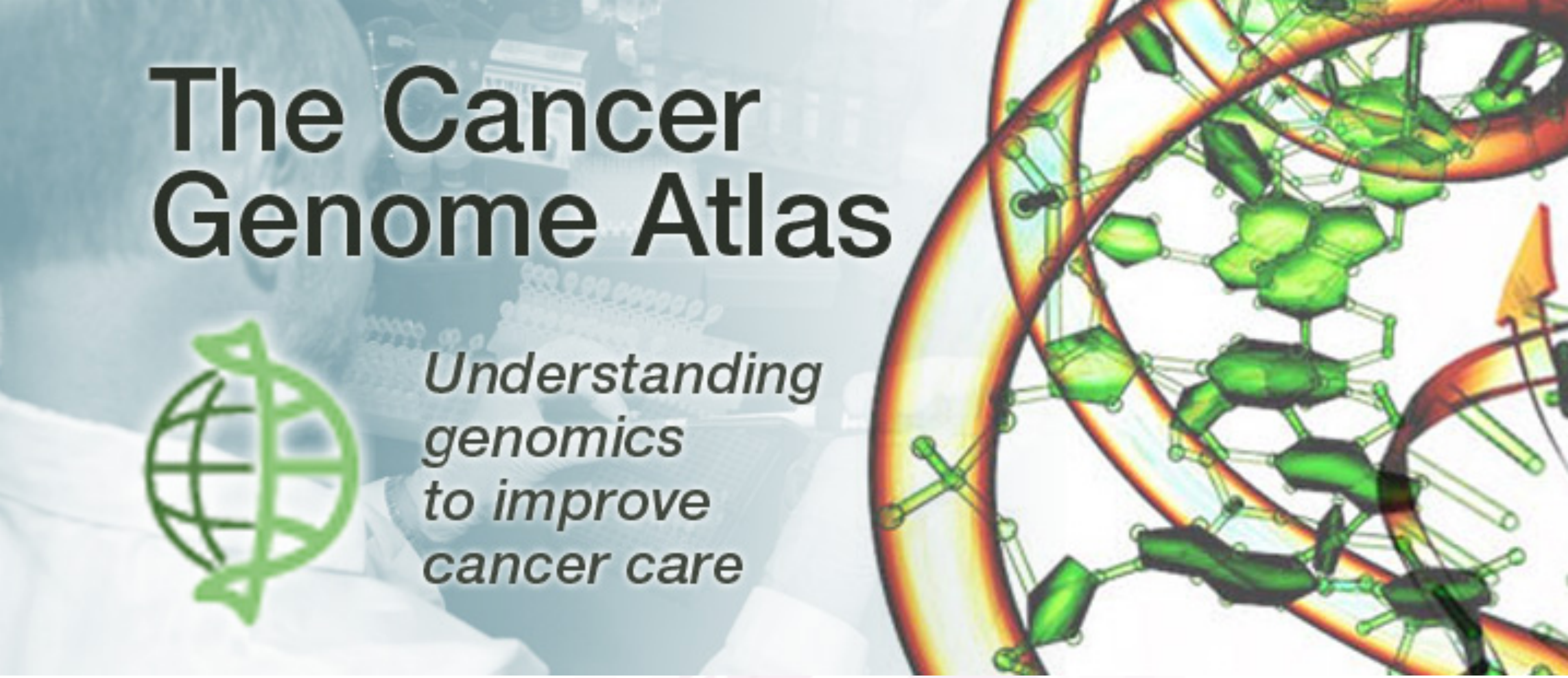
- The project overview
- *Challenge 1.* Stream of updates
- *Challenge 2.* Volume: size of the data and infrastructure
- *Challenge 3.* Modelling: training of genetic signatures
- *Challenge 4.* Integration of derived signatures
- Performance of derived signatures

How to create a genetic signature that will score the risk of chemo-resistance for cancer patients.

- More than 2000 variables in the clinical dataset. Very detailed information about treatment and outcomes.
- One can derive an index whatever the chemotherapy was successful or not (based on symptoms like short lifespan, early change in chemo treatment).



http://www.erasmusmc.nl/47687/51019/4294078/chemoresistance?lang=en

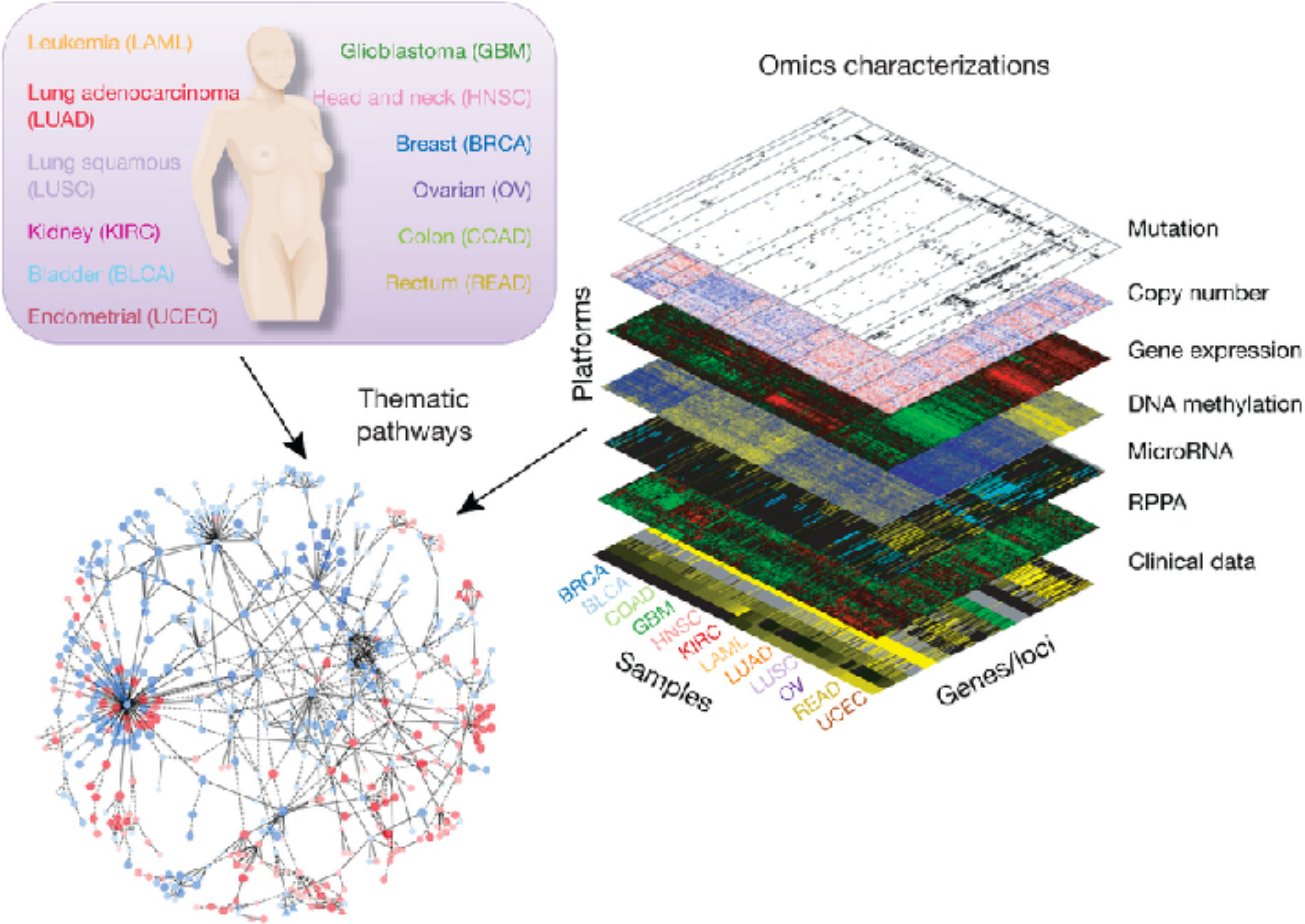


The Cancer Genome Atlas (TCGA) is a project to catalogue genetic changes responsible for cancer.

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)
Omics characterizations
Mutation
Copy number
Gene expression
DNA methylation
MicroRNA
RPPA
Clinical data
Platforms
Thematic pathways
BRCA
BLCA
COAD
GBM
HNSC
KIRC
LAML
LUAD
LUSC
OV
READ
UCEC
Samples
Genes/loci

Predictive modelling

Predictive modelling has a lot of business applications and there are good statistical tools for such modelling.

What is special in our case is the large number of predictors

- ~ 20 k – gene expression data,
- ~ 500 k – methylation data,
- ~ 1.5 M – SNP mutations

http://www.predictivemodelingnews.com/

Challenge 1:

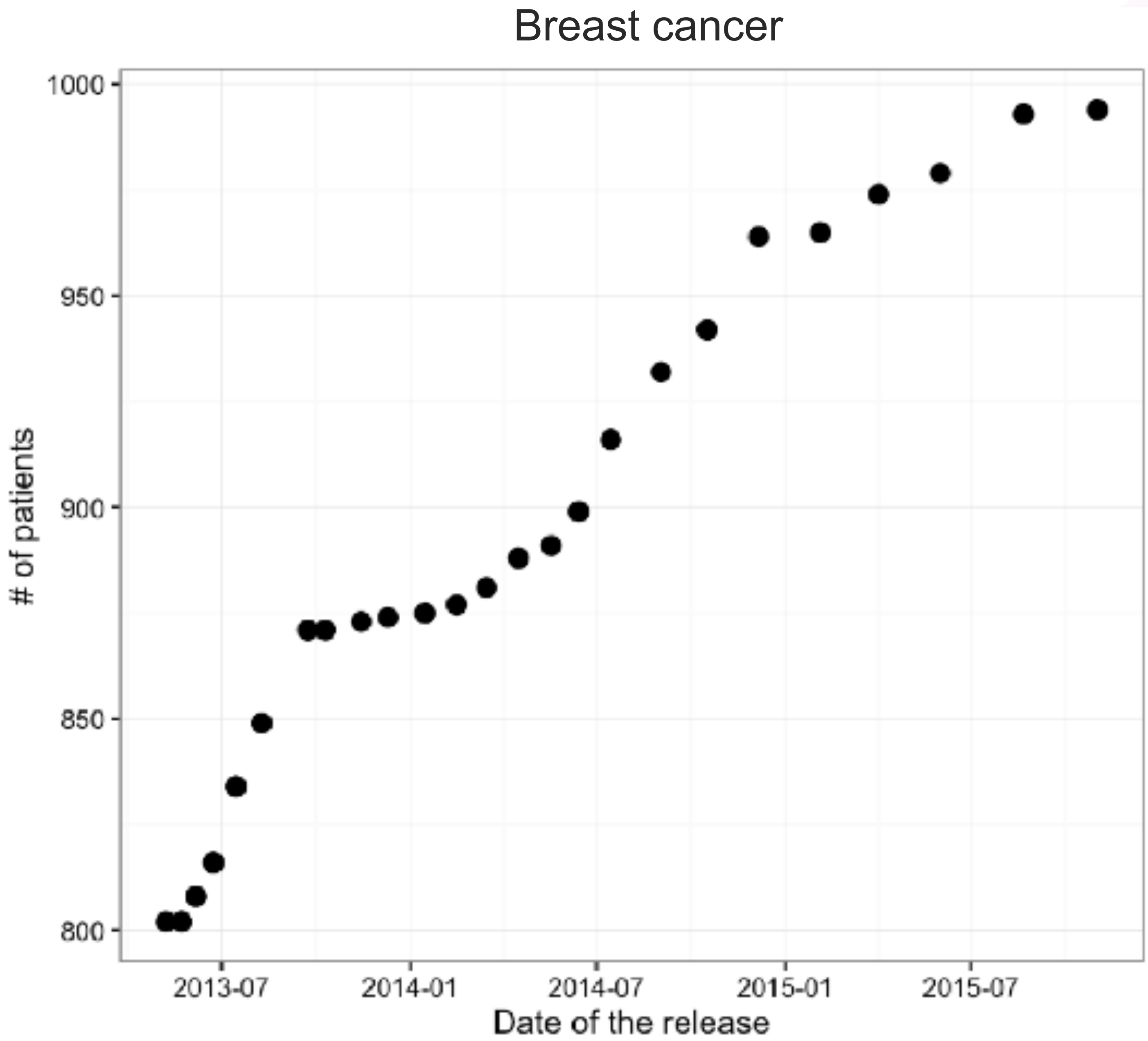
Stream of consecutive releases

In order to give researchers access to the data as soon as possible, set of consecutive releases is published. Starting from 2013 up to now there is over 30 new releases.

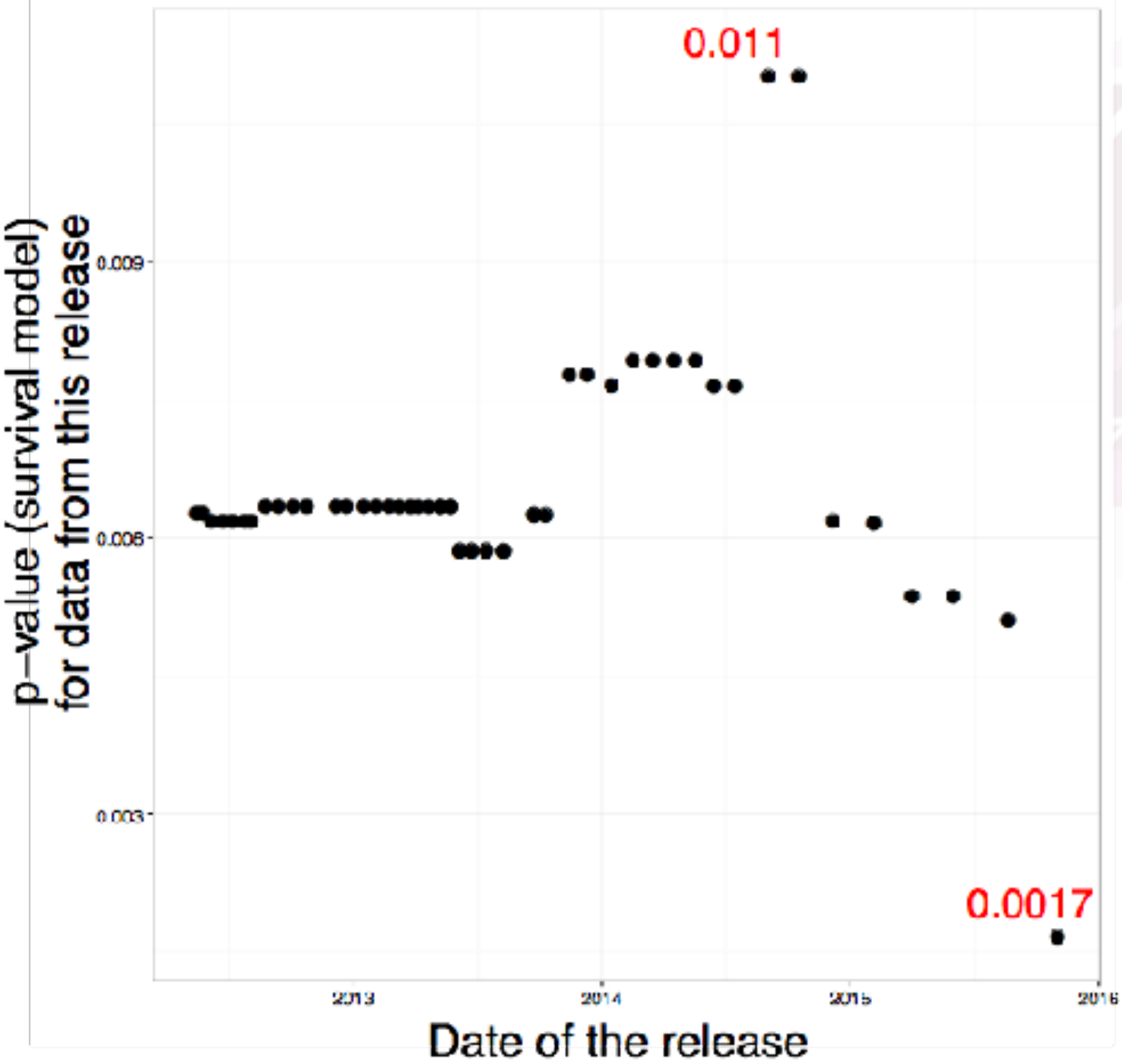
There are small differences between each release, few patients are added or removed.

Clinical data has longer follow-ups since there is longer observation period. And there is some additional cleaning.

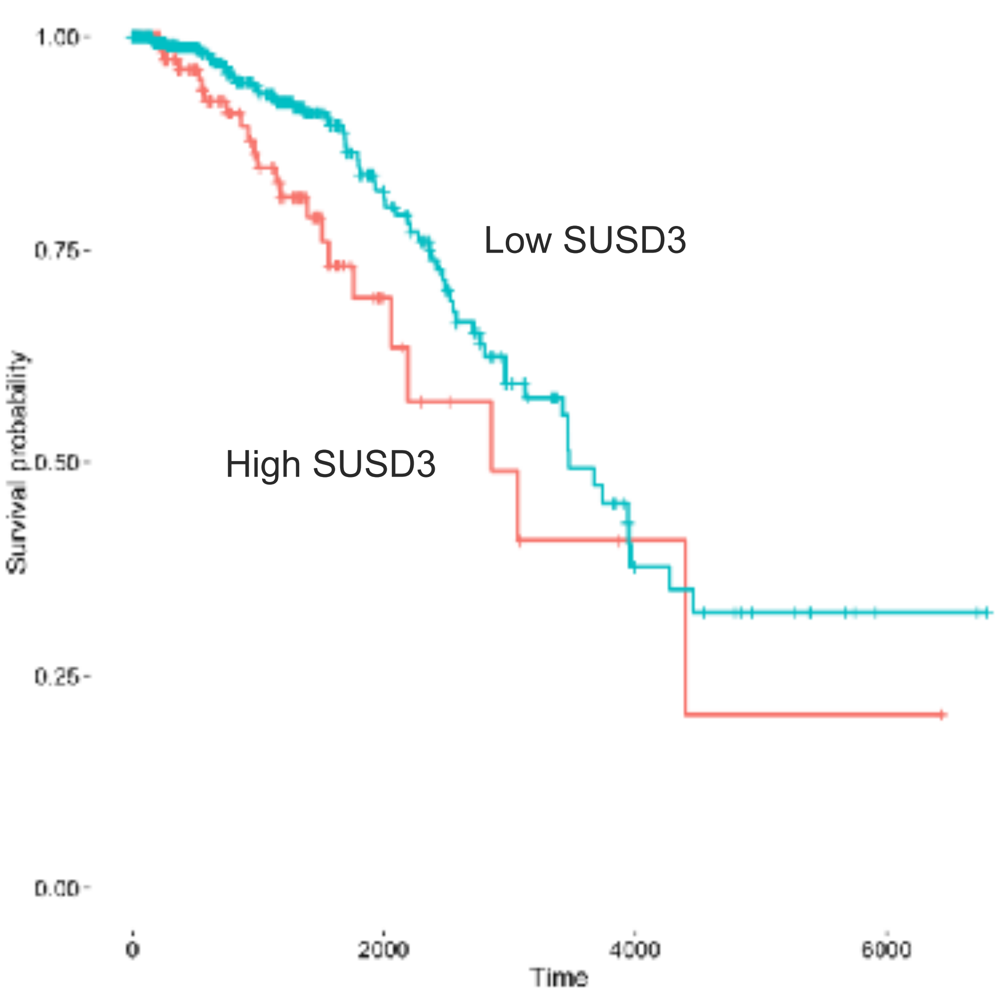
Yet, every new release always creates a risk that results will change slightly. How slightly?


Breast cancer
1000
950
900
850
800
# of patients
2013-07
2014-01
2014-07
2015-01
2015-07
Date of the release

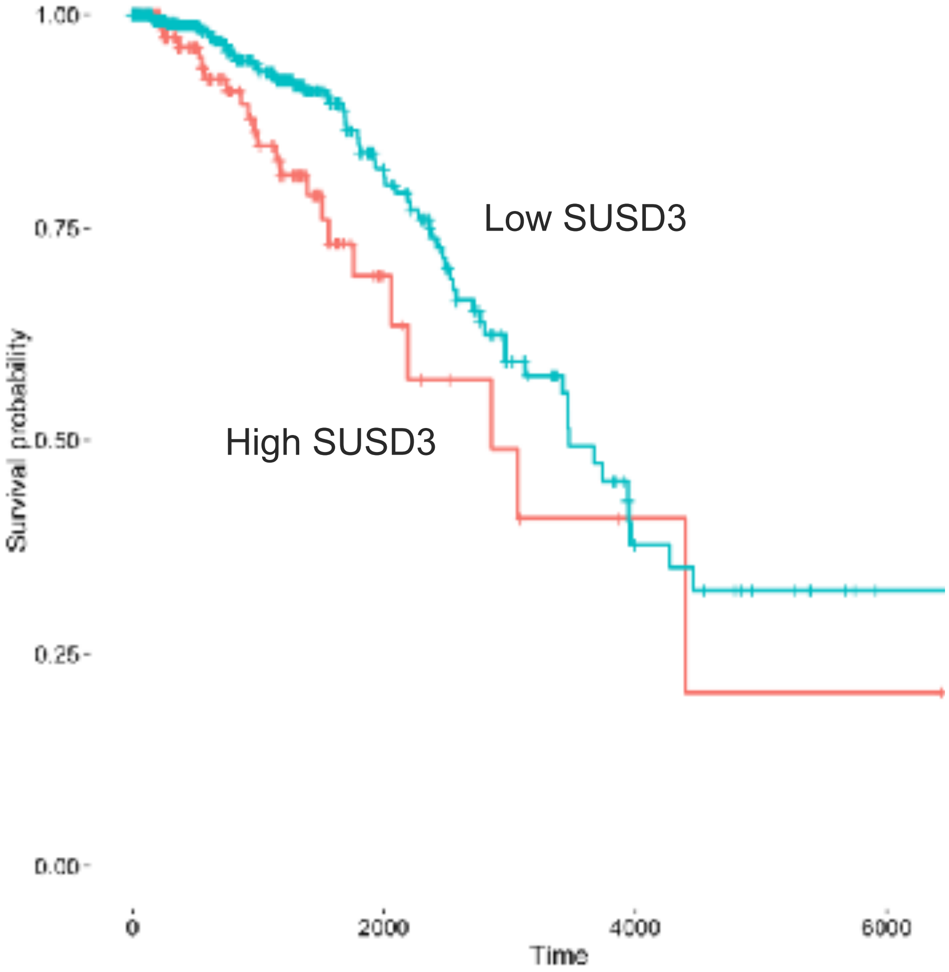
GeneSUSD3


0.011
0.0017
p−value (survival model)
for data from this release
0.009
0.006
0.003
2013
2014
2015
2016
Date of the release

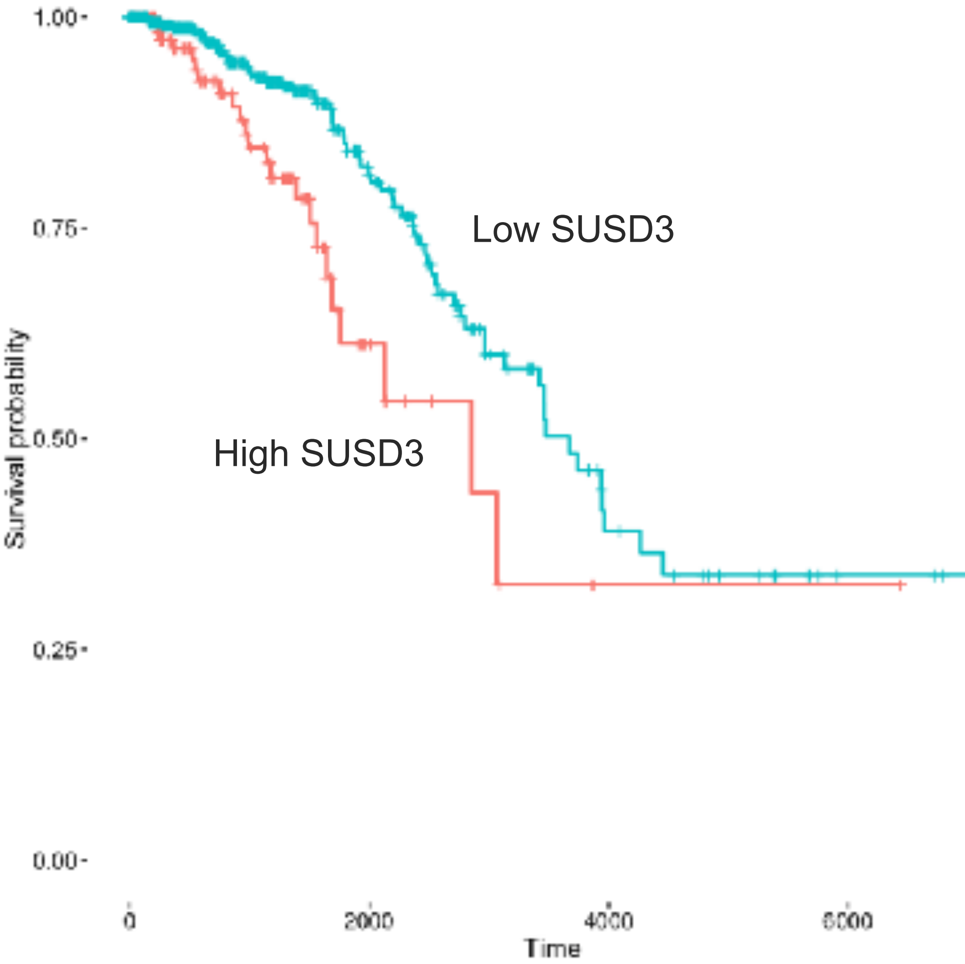
gene: SUSD3
cohort: Breast Cancer
Release: 2014 09 02
p-value 0.01

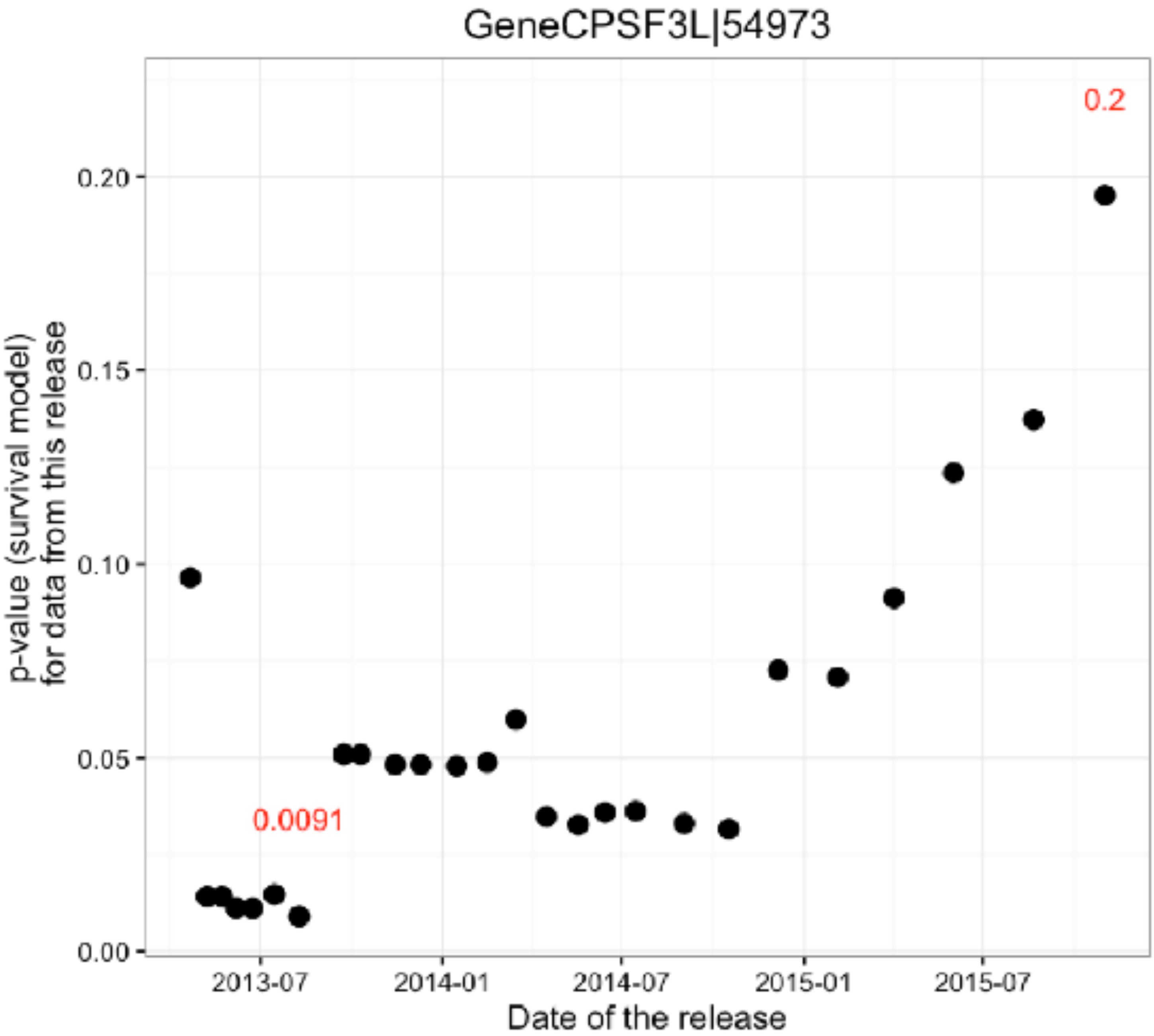


gene: SUSD3
cohort: Breast Cancer
Release: 2014 09 02
p-value 0.01

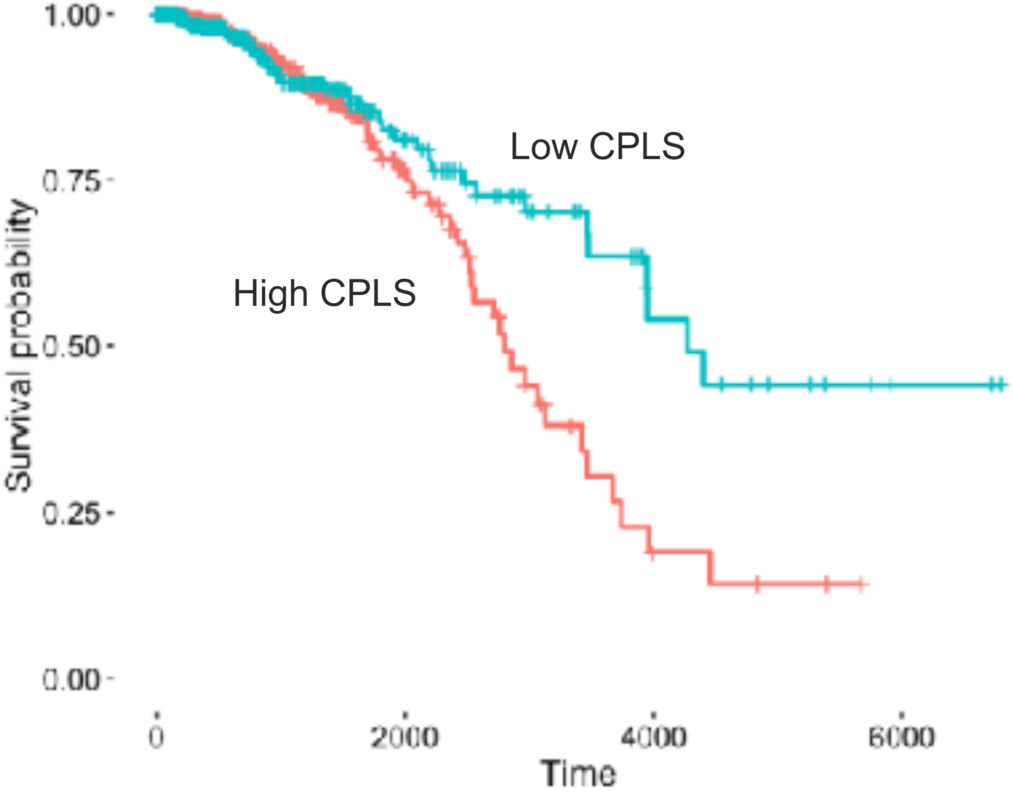


gene: SUSD3
cohort: Breast Cancer
Release: 2015 08 21
p-value 0.001

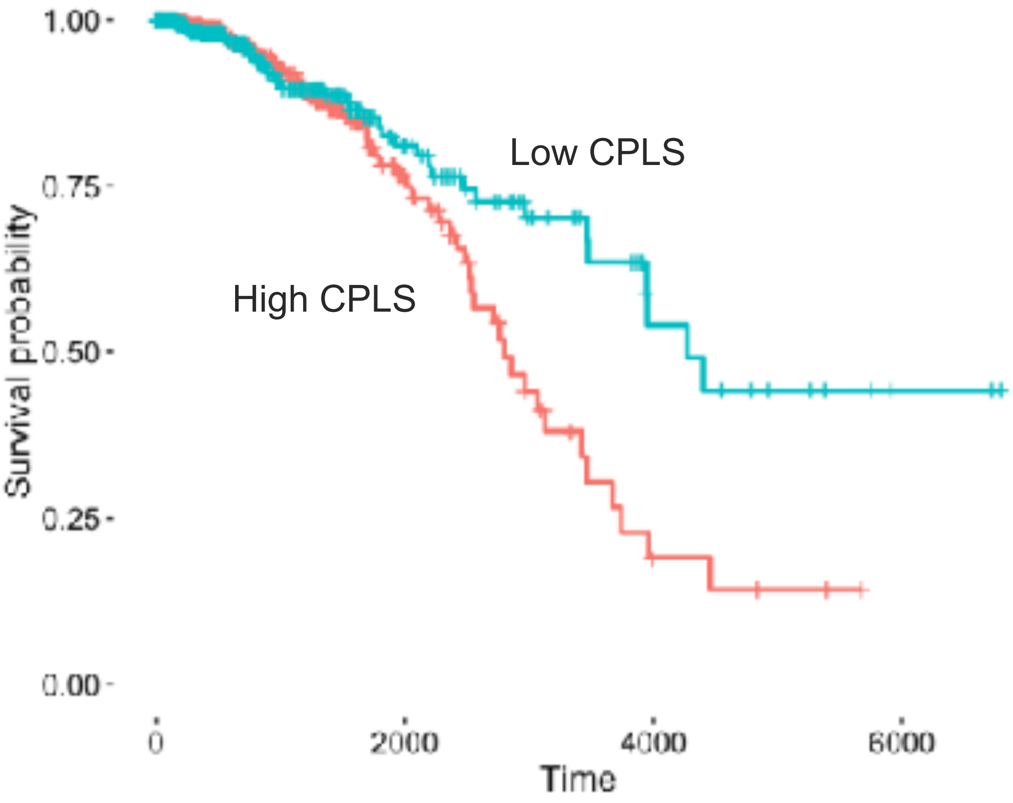



GeneCPSF3L|54973
0.2
0.0091
0.20
0.15
0.10
0.05
0.00
p-value (survival model)
for data from this release
2013-07
2014-01
2014-07
2015-01
2015-07
Date of the release

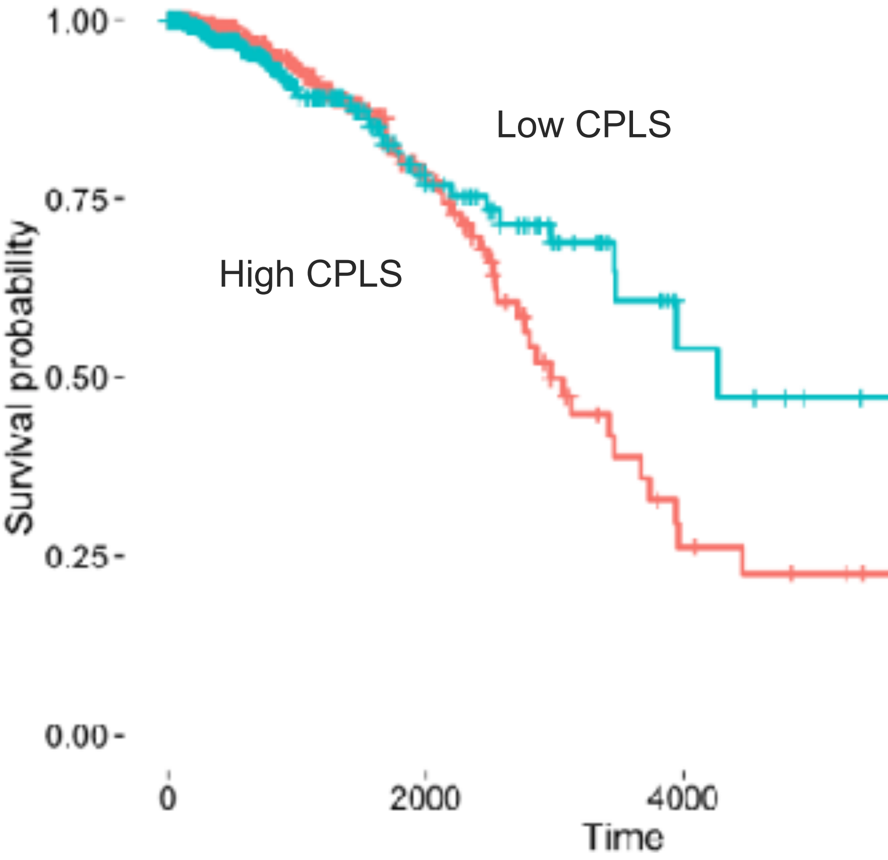
gene: CPLS
cohort: Breast Cancer
Release: 2013 07
p-value 0.0091



gene: CPLS
cohort: Breast Cancer
Release: 2013 07
p-value 0.0091



gene: CPLS
cohort: Breast Cancer
Release: 2015 08 21
p-value 0.2

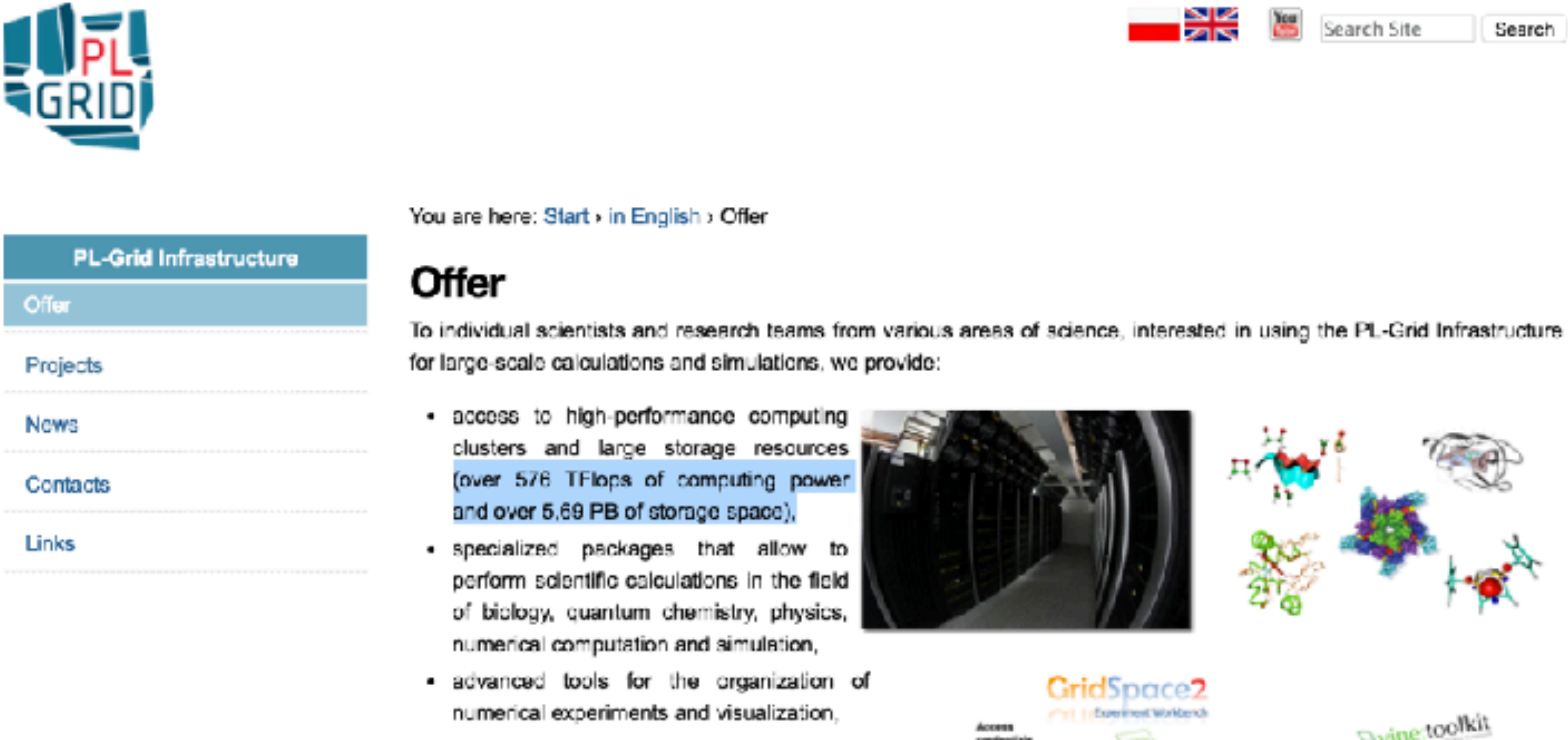


Challenge 2.

Volume: size of the data and infrastructure

To store and process the data of this size we used PL-Grid infrastructure.

Sometimes the ‘fat’ nodes with 512GB of RAM are useful.



You are here: Start › in English › Offer

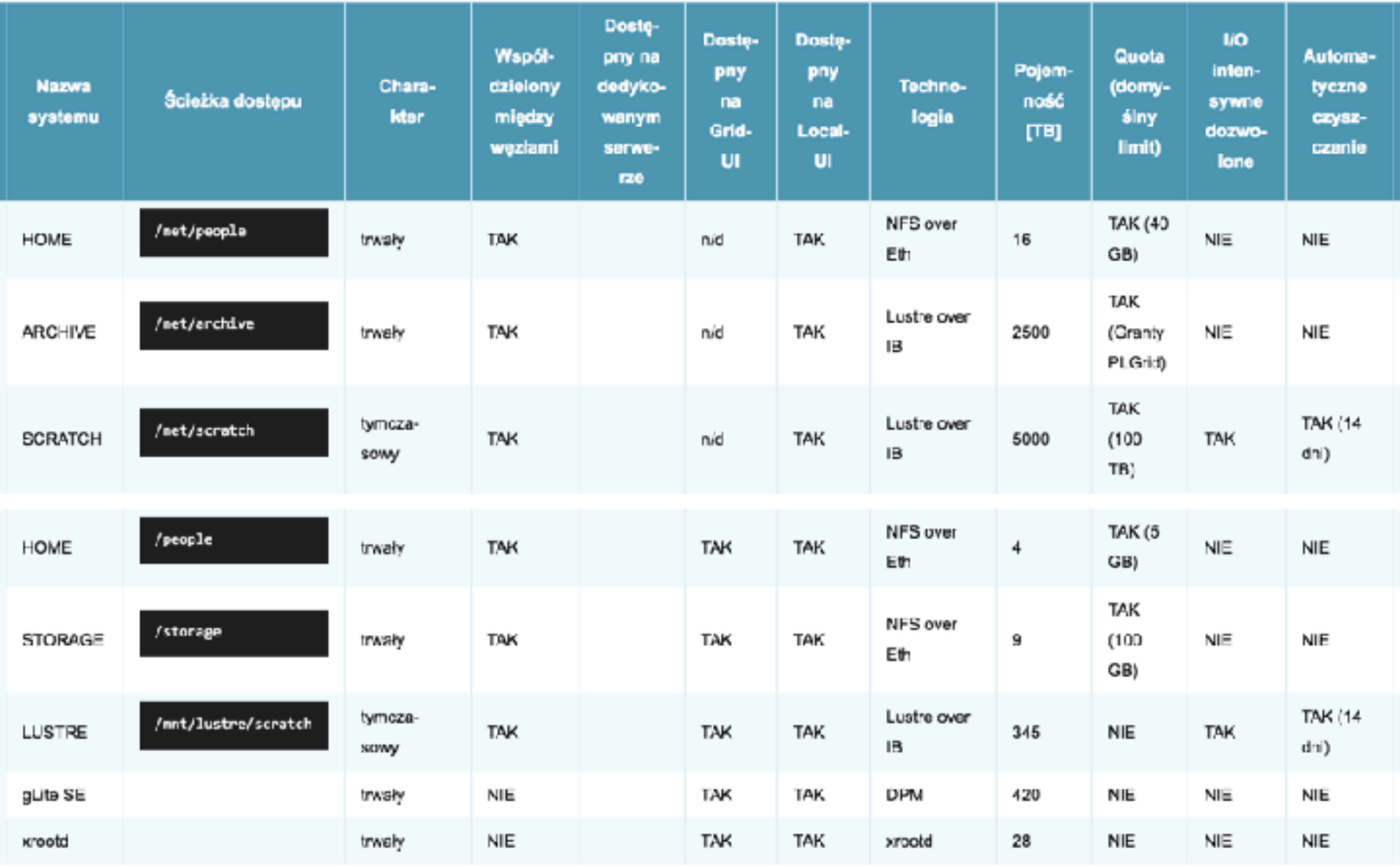
Offer

To individual scientists and research teams from various areas of science, interested in using the PL-Grid Infrastructure for large-scale calculations and simulations, we provide:

- access to high-performance computing clusters and large storage resources (over 576 TFlops of computing power and over 5.69 PB of storage space),
- specialized packages that allow to perform scientific calculations in the field of biology, quantum chemistry, physics, numerical computation and simulation,
- advanced tools for the organization of numerical experiments and visualization,

Challenge 2.

Volume: size of the data and infrastructure



Nazwa systemu	Ścieżka dostępu	Charakter	Współdzielony między węzłami	Dostępny na dedykowanym serwerze	Dostępny na Grid-UI	Dostępny na Local-UI	Technologia	Pojemność [TB]	Quota (domyślny limit)	I/O intensywne dozwolone	Automatyczne czyszczenie
HOME	/net/people	trwały	TAK		n/d	TAK	NFS over Eth	16	TAK (40 GB)	NIE	NIE
ARCHIVE	/net/archive	trwały	TAK		n/d	TAK	Lustre over IB	2500	TAK (Granty PLGrid)	NIE	NIE
SCRATCH	/net/scratch	tymczasowy	TAK		n/d	TAK	Lustre over IB	5000	TAK (100 TB)	TAK	TAK (14 dni)
HOME	/people	trwały	TAK		TAK	TAK	NFS over Eth	4	TAK (5 GB)	NIE	NIE
STORAGE	/storage	trwały	TAK		TAK	TAK	NFS over Eth	9	TAK (100 GB)	NIE	NIE
LUSTRE	/mnt/lustre/scratch	tymczasowy	TAK		TAK	TAK	Lustre over IB	345	NIE	TAK	TAK (14 dni)
gLite SE		trwały	NIE		TAK	TAK	DPM	420	NIE	NIE	NIE
xrootd		trwały	NIE		TAK	TAK	xrootd	28	NIE	NIE	NIE

Challenge 3.

Modelling: training of genetic signatures

We are going to test three sate of the art classifiers:

1. **Logistic regression** with L1 regularisation (LASSO, additive model).
2. **Gradient Boosting** with decision trees (model that allows for low order interactions),
3. **Random Forest** (model that allows for deep interactions),

Two approaches to variable preselection (one dimensional screening vs. vimp).

Here we will show only results for mRNA samples:

TCGA dataset: https://bioconductor.org/packages/release/data/experiment/html/RTCGA.rnaseq.html

Challenge 3. Modelling:

Logistic regression with regularisation

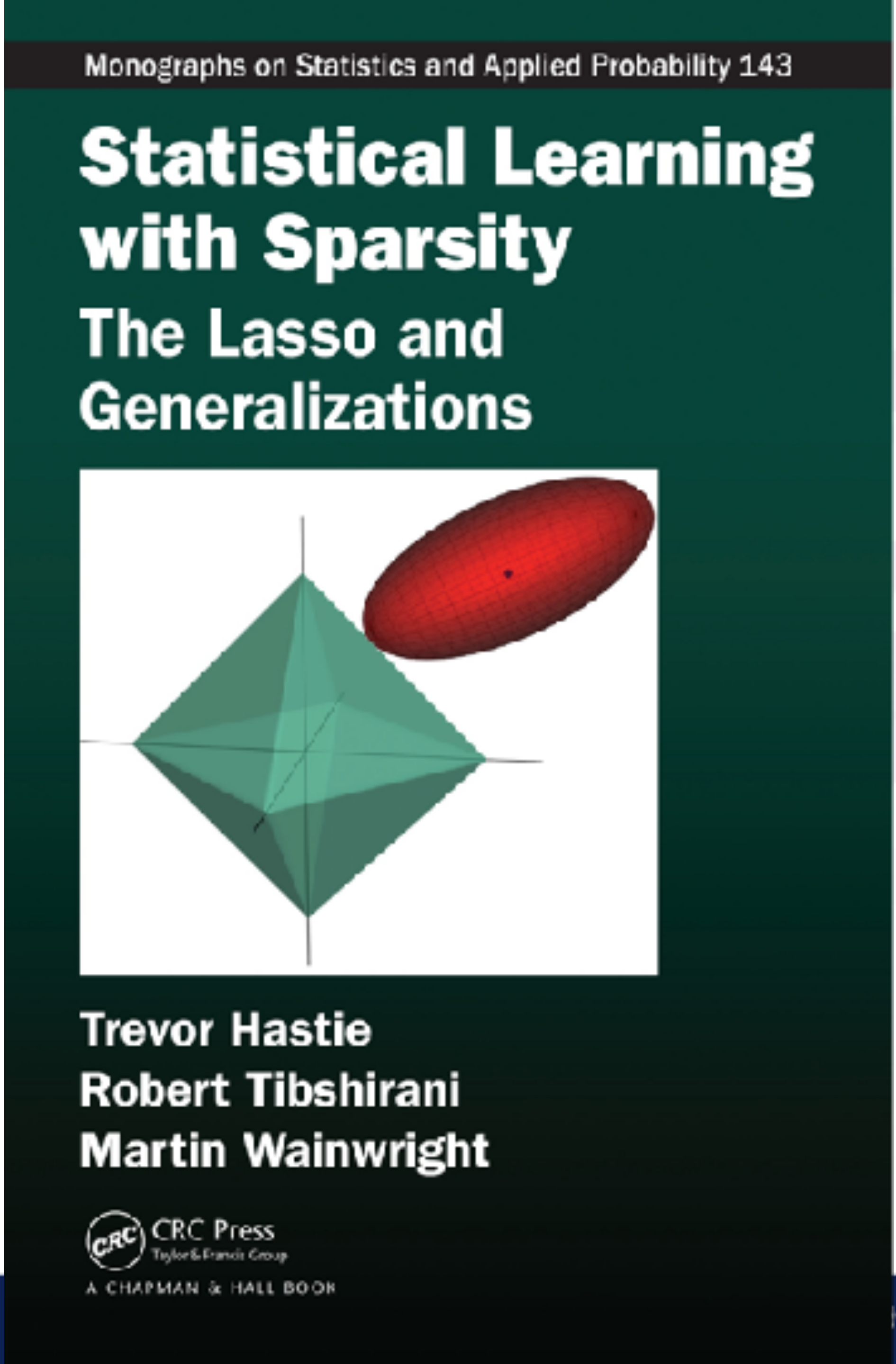
- A model based approach + LASSO regularisation [least absolute shrinkage and selection operator].

- The transformation of gene expression is very important. We are using log (1+ scaled counts).

logit(score) = X_1 b_1 + … + X_p b_p

- **Linear method.** Final score is a monotonic transformation of linear combination of features/genes. **Easier interpretation.**

- When working with high dimensional data the regularisation is important to improve the predictive properties. Here we are using the lasso regularisation L1 penalty.



Monographs on Statistics and Applied Probability 143

Statistical Learning with Sparsity

The Lasso and Generalizations

Trevor Hastie
Robert Tibshirani
Martin Wainwright

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Trevor Hastie

Robert Tibshirani

Martin Wainwright

Logistic regression with regularisation

- [removed]

Challenge 3. Modelling:

Random Forest

- Random forest is an ensemble of decision trees. Trees (that are grown very deep) are known to have low bias, but very high variance.

- Random forests are averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance (significantly) and introduce bias (slightly). *Breiman, L. (2001), Random Forests, Machine Learning 45(1), 5-32.*

- **Non-linear method.** Holds some similarities to k-nearest neighbours. Harder interpretation of the model.

Random Forest

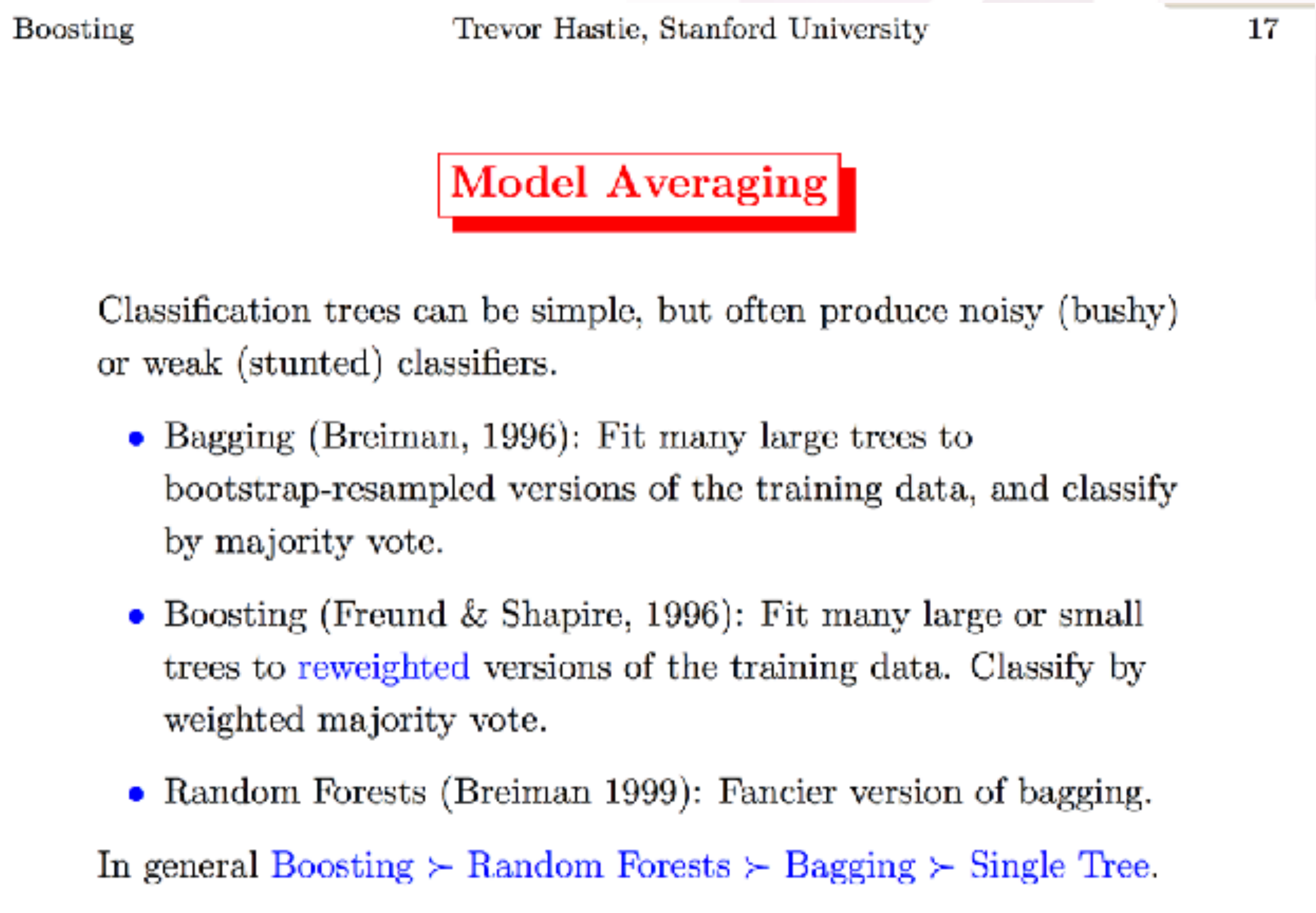
- [removed]

Challenge 3. Modelling:

Gradient Boosting Machine

- Very popular on Kaggle. Expected good predictive properties.
- GBM is also based on an ensemble of decision trees. But the trees are not trained on independent bootstrapped samples. Instead, consecutive trees are trained on residuals from previous models.
- It resembles gradient optimisation, since consecutive steps are improvements of current solution. *Friedman, J. (1999) Greedy Function Approximation: A Gradient Boosting Machine.*
- **Non-linear method.** Harder interpretation of the model.

What Trevor Hastie said...



Boosting Trevor Hastie, Stanford University 17

Model Averaging

Classification trees can be simple, but often produce noisy (bushy) or weak (stunted) classifiers.

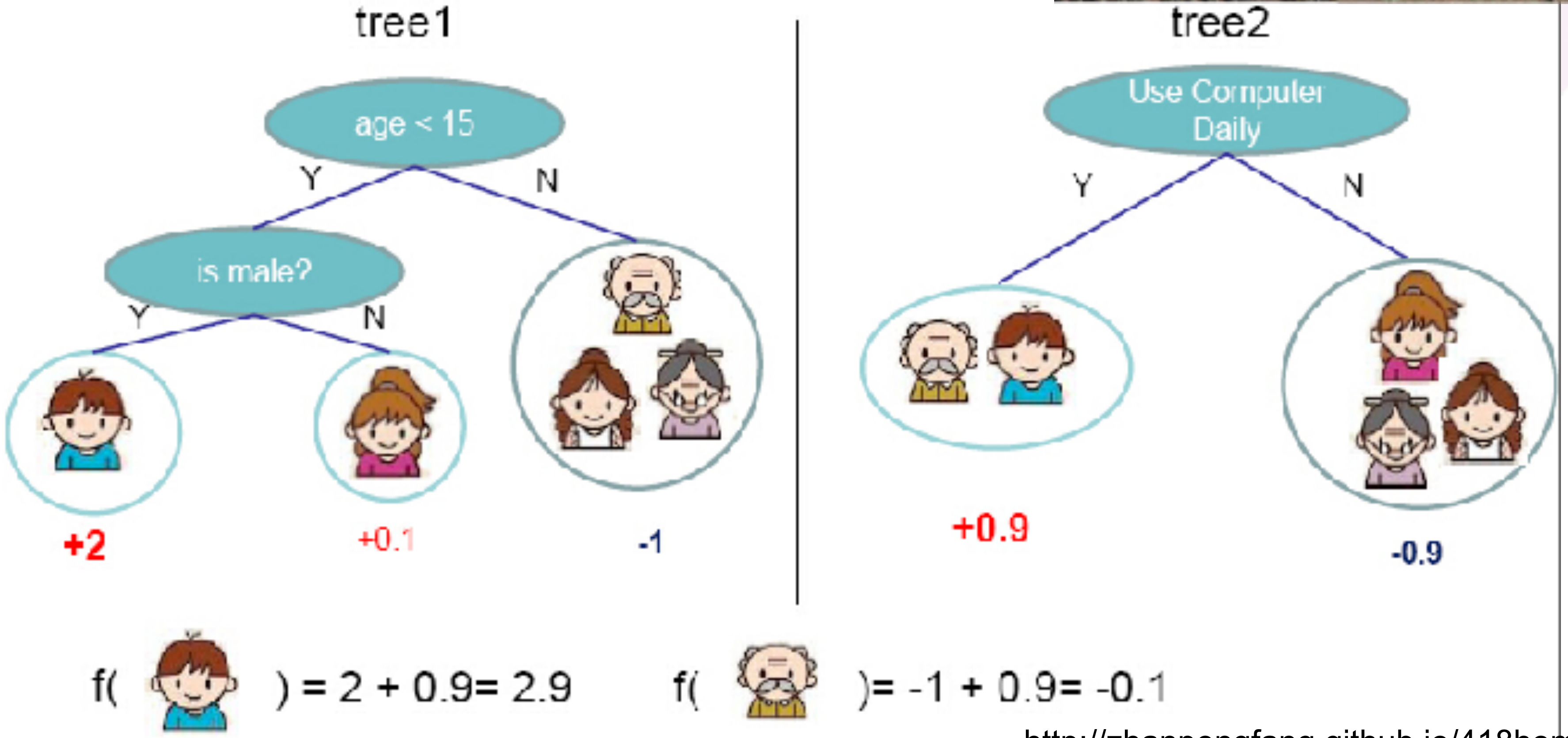
- Bagging (Breiman, 1996): Fit many large trees to bootstrap-resampled versions of the training data, and classify by majority vote.
- Boosting (Freund & Shapire, 1996): Fit many large or small trees to reweighted versions of the training data. Classify by weighted majority vote.
- Random Forests (Breiman 1999): Fancier version of bagging.

In general Boosting $\succ$ Random Forests $\succ$ Bagging $\succ$ Single Tree.

Gradient Boosting Machine

- [removed]

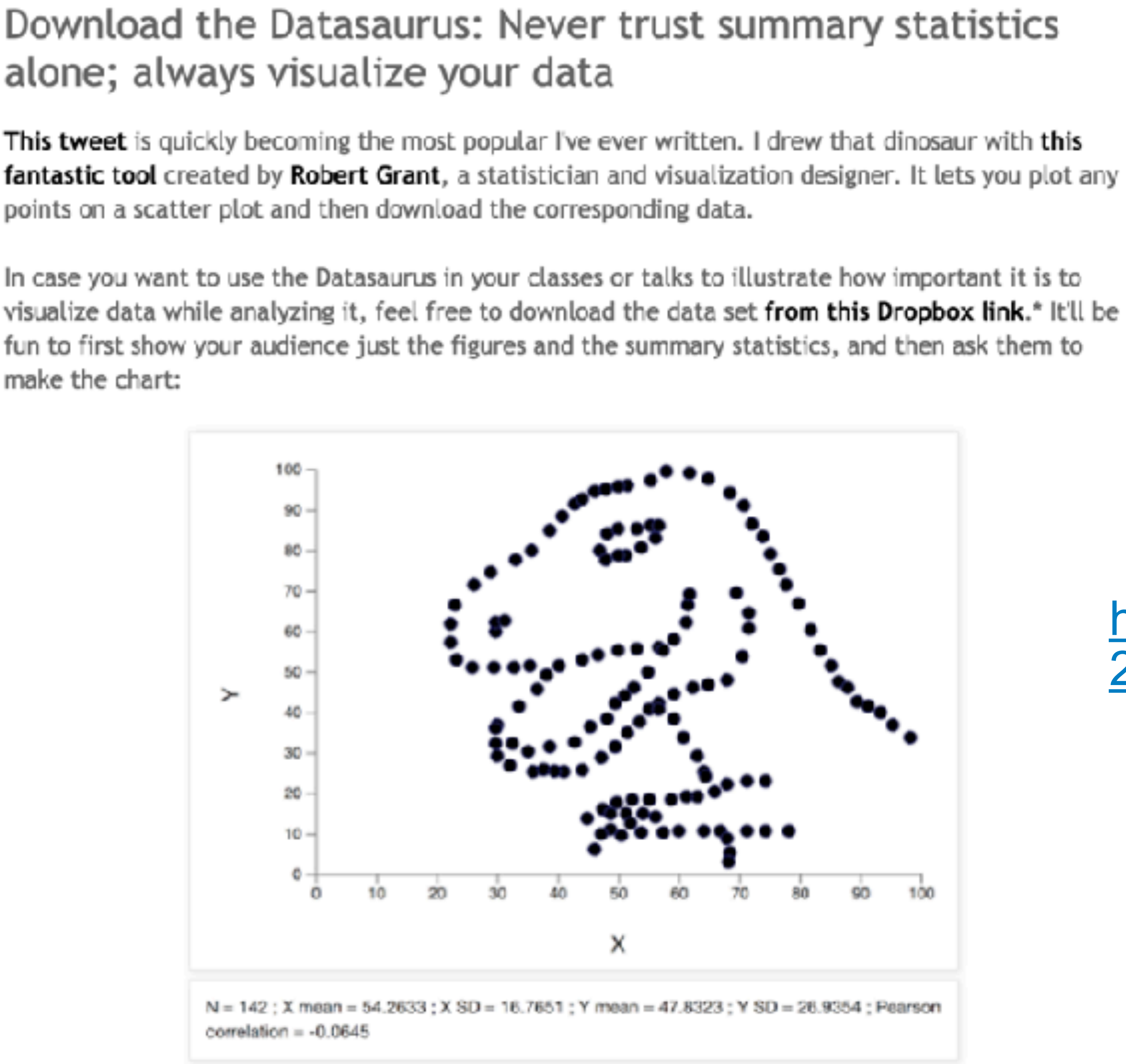
Random Forest vs Gradient Boosted Trees



http://zhanpengfang.github.io/418home.html

Top 150 most important biomarkers (for domain validation)

[removed]



Download the Datasaurus: Never trust summary statistics alone; always visualize your data

This tweet is quickly becoming the most popular I've ever written. I drew that dinosaur with **this fantastic tool** created by **Robert Grant**, a statistician and visualization designer. It lets you plot any points on a scatter plot and then download the corresponding data.

In case you want to use the Datasaurus in your classes or talks to illustrate how important it is to visualize data while analyzing it, feel free to download the data set **from this Dropbox link.*** It'll be fun to first show your audience just the figures and the summary statistics, and then ask them to make the chart:

N = 142 ; X mean = 54.2633 ; X SD = 16.7651 ; Y mean = 47.8323 ; Y SD = 26.9354 ; Pearson correlation = -0.0645

http://bit.ly/2e3JiAQ

Challenge 4.

Multi platform integration of signatures

Model stacking:

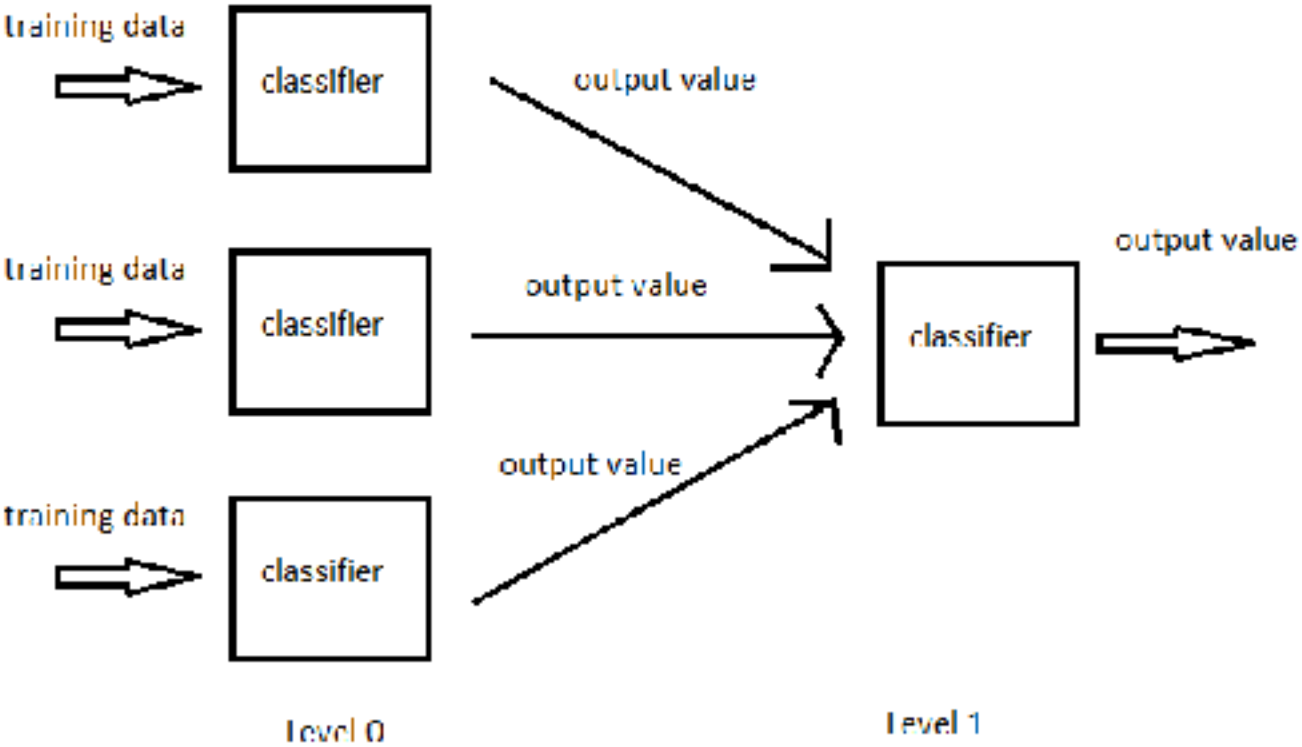
Merging of Single Platform Classifiers into a Cross Platform Classifier

There are different ways to combine classifiers for different platforms.

Here we are using model stacking.

http://www.chioka.in/stacking-blending-and-stacked-generalization/

Concept Diagram of Stacking



Survival curves for selected cancers in TCGA

- [removed]

Final thoughts

1. Data visualisation is very important. It helps to validate and communicate findings.
2. The reproducibility of findings is important. For large, complex and live data this may be a real issue.
3. For large datasets the infrastructure is very important. You cannot generate a sample if you cannot access the data.
4. In the genetic profiling the scoring time is not crucial, but the modelling and validation are crucial. One model is enough, but it’s pretty big.
5. Black boxes may be effective, but since these are life and death decisions the model transparency is also very important.
6. Data comes from various platforms. Integration into a single model may be an issue.

Przemysław Biecek

Odkrywać! Ujawniać! Objaśniać!

Zbiór esejów o sztuce prezentowania danych

Acknowledgements

Warsaw University of Technology

Marcin Kosiński (archivist, RTCGA)
Kornel Kiełczewski (mRNA/miRNA)
Katarzyna Fąk (mRNA-seq)
Witold Chodor (RTCGA)
Paulina Auguścik (splicing)
Barbara Sozańska (mRNA)

Greater Poland Cancer Centre

Maciej Wiznerowicz
Urszula Oleksiewicz
Marta Gładych
and rest of the team

US Embassy Science Fellow

Arcady Mushegian

MD Anderson Cancer Center

Parantu Shah

International Institute of Molecular and Cell Biology

Maciej Żylicz
Alicja Żylicz
Bartosz Wawrzynów
Maciej Olszewski
Marcin Herok

PATRONAT HONOROWY
JM REKTORA

SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE

The Fourth International Workshop

Advanced Analytics and Data Science

Partner

sas THE POWER TO KNOW.

Media Partners

Med express.pl

ITWIZ

medycyna praktyczna