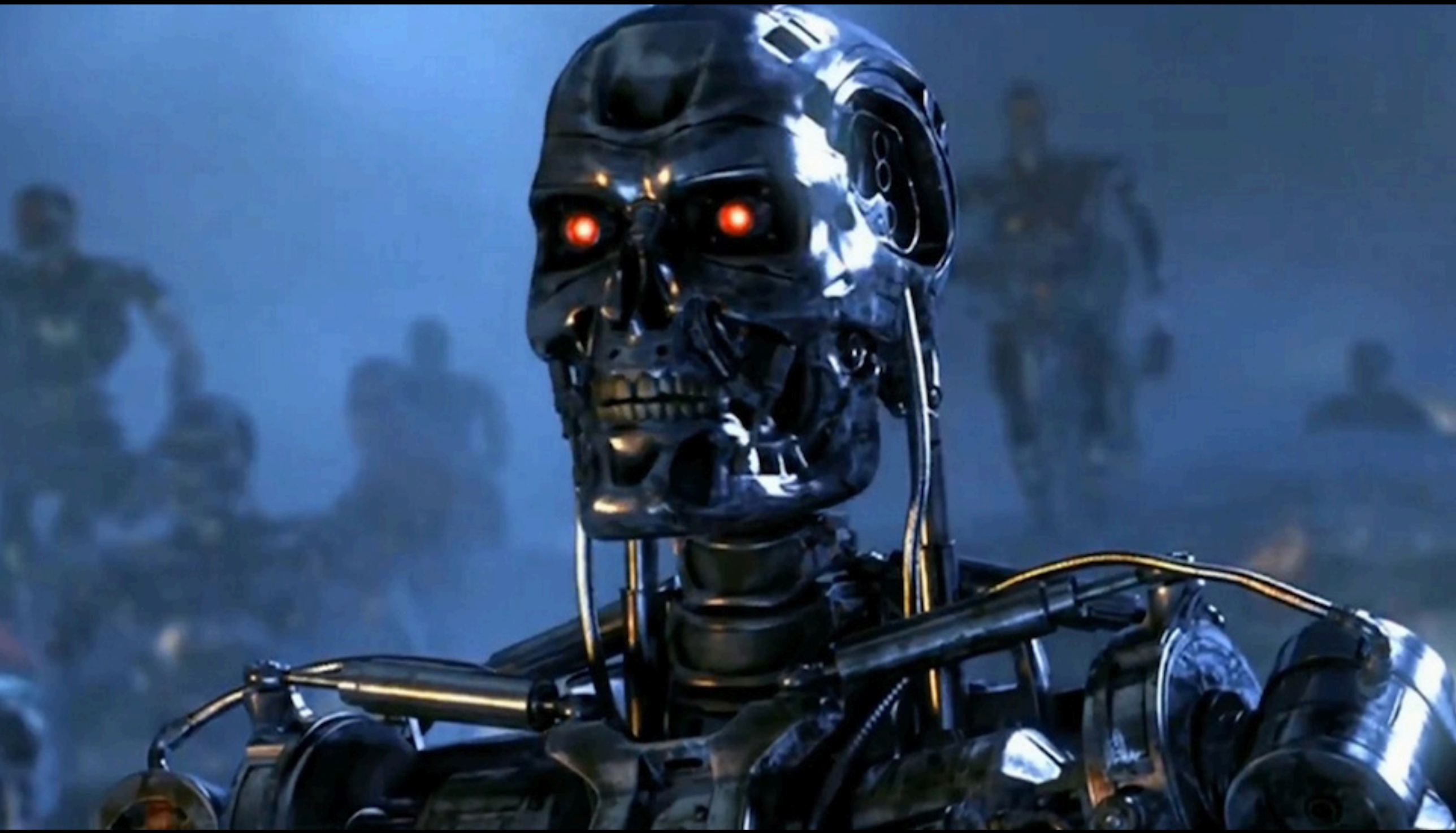
Show me your model

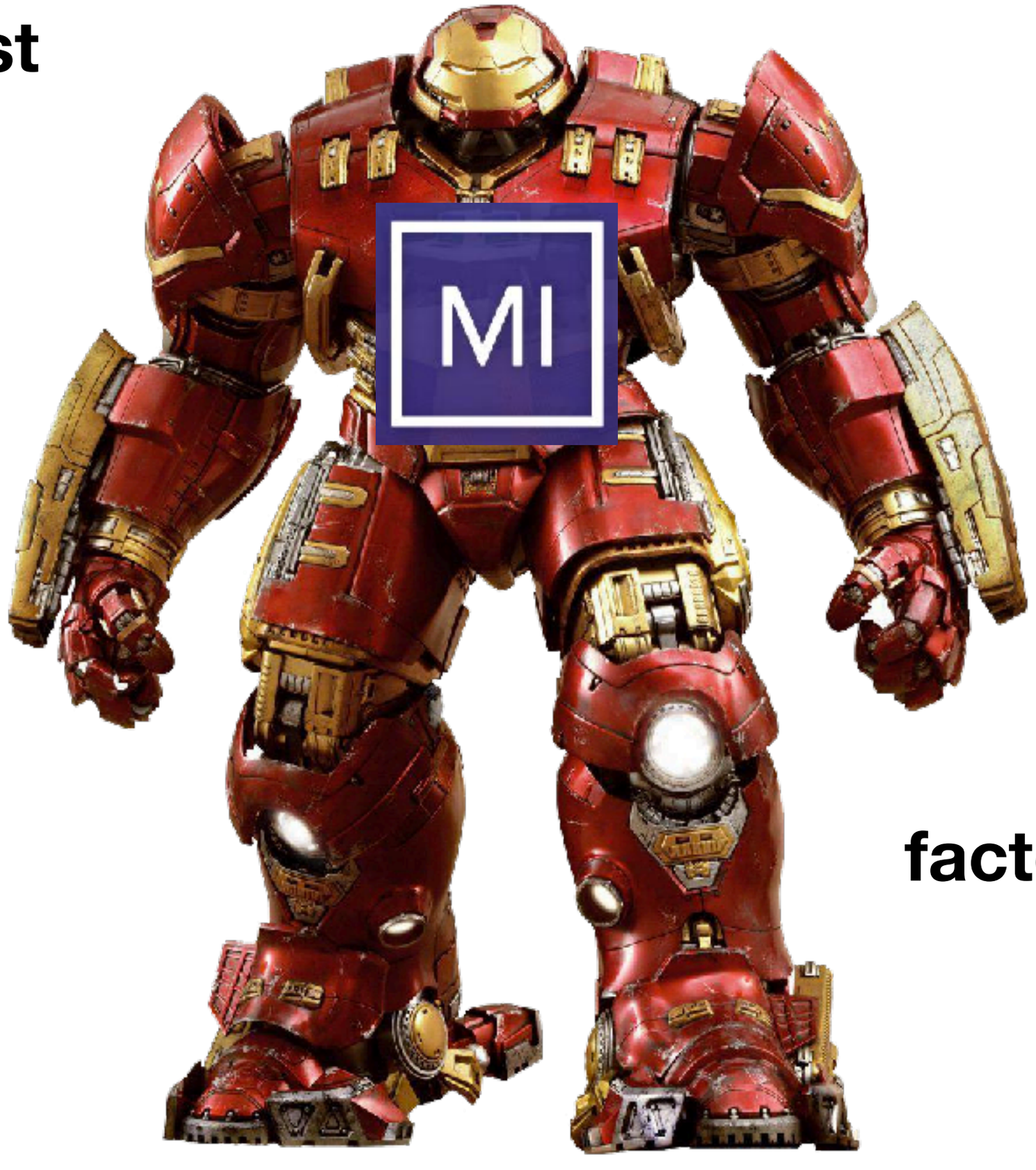
Przemysław Biecek
MI^2




Laflin
5m

randomForestExplainer

archivist



live

ggfortify

factorMerger

pdp

Czy potrzebujemy wizualizacji do zrozumienia modeli?

1) …there is **a lot of opportunity to do visualization for machine learning**. Even many of the people working in the field don't have good intuitions for how their systems work, and they need tools to inspect what they're doing, debug, etc…

https://eagereyes.org/blog/2017/eurovis-2017-conference-report-part-1
EuroVis 2017 Conference
Robert Kosara
Senior Research Scientist at Tableau

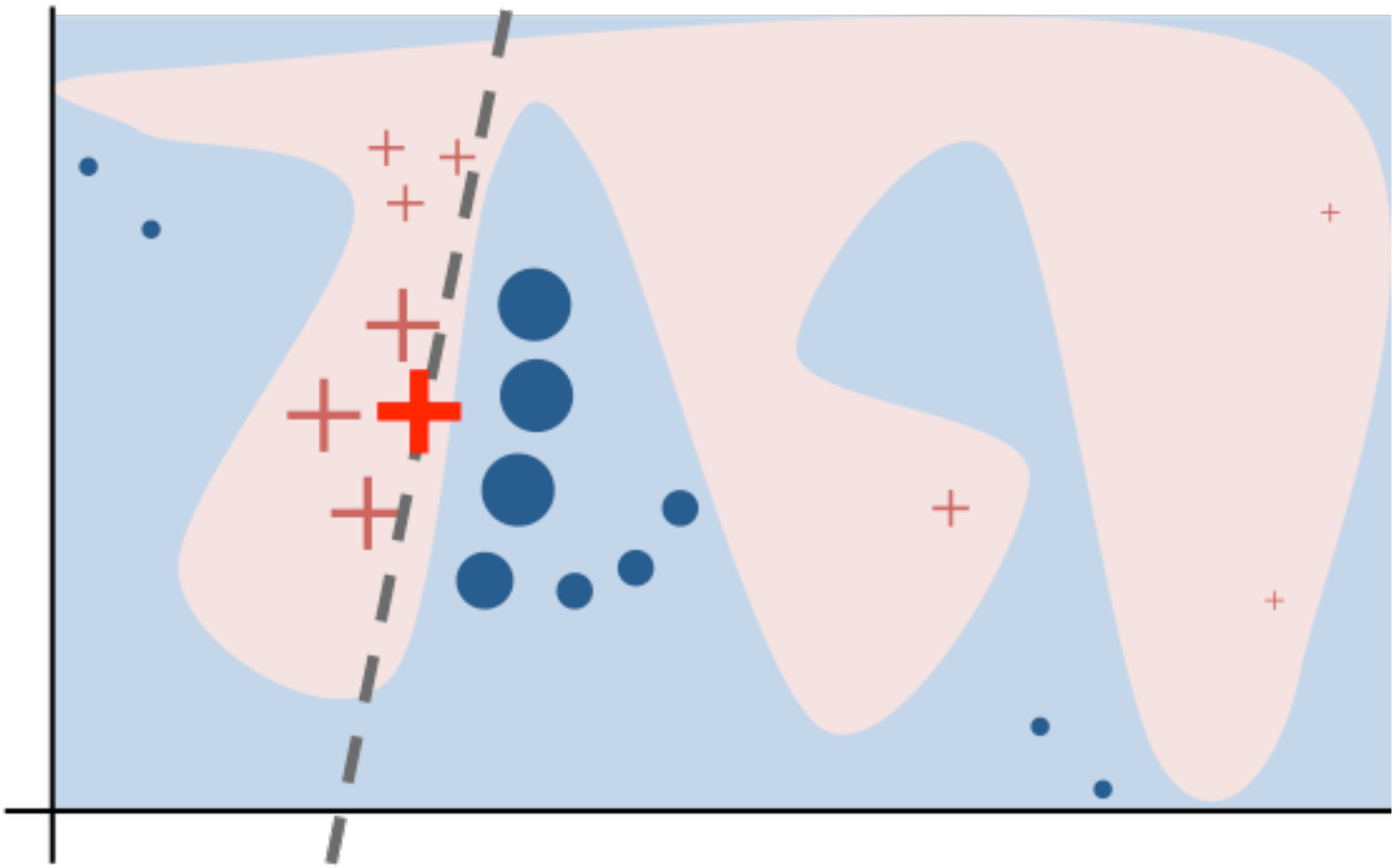
2) Understanding and trust - we need to understand models that makes important decisions for us.
3) Machine learning in regulated industry

"Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016)

Lokalne wyjaśnienia

LIME: Local Interpretable Model-agnostic Explanations

1. Generate a fake dataset around x.
2. Use black-box estimator to get target values y.
3. Train a new white-box estimator for (y,x).
4. Check prediction quality of a white-box classifier.
5. Use white-box estimator as an explanation of black-box model.



"Why Should I Trust You?" Explaining the Predictions of Any Classifier.
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016). https://arxiv.org/pdf/1602.04938.pdf

Port to R: Thomas Lin Pedersen (2017) https://github.com/thomasp85/lime

LIME: Local Interpretable Model-agnostic Explanations

(a) Original Image

(b) Explaining *Electric guitar*

Co poszło źle?

(a) Husky classified as wolf

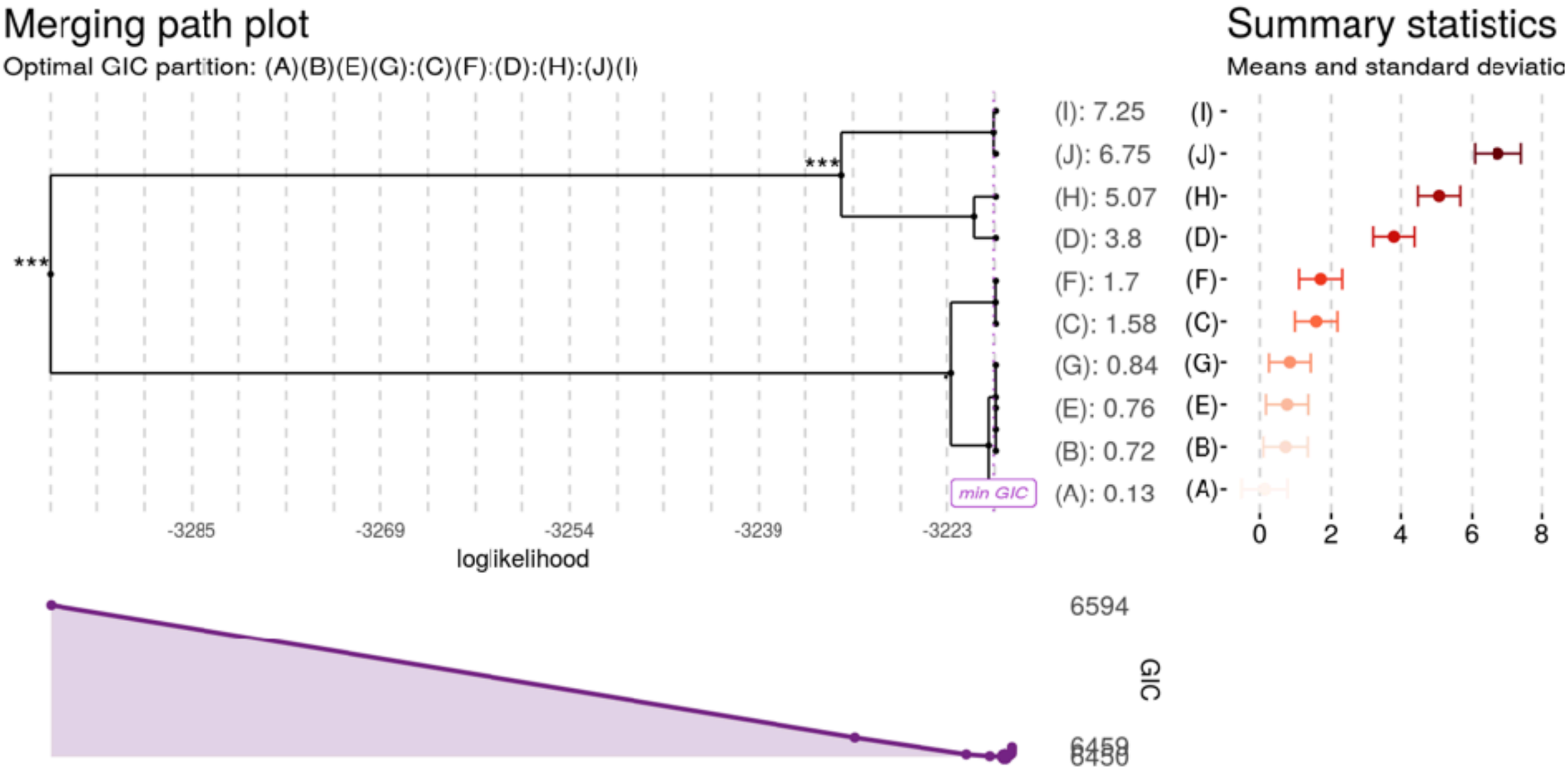
(b) Explanation

An R port https://github.com/thomasp85/lime

"Why Should I Trust You?" Explaining the Predictions of Any Classifier.
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016). https://arxiv.org/pdf/1602.04938.pdf

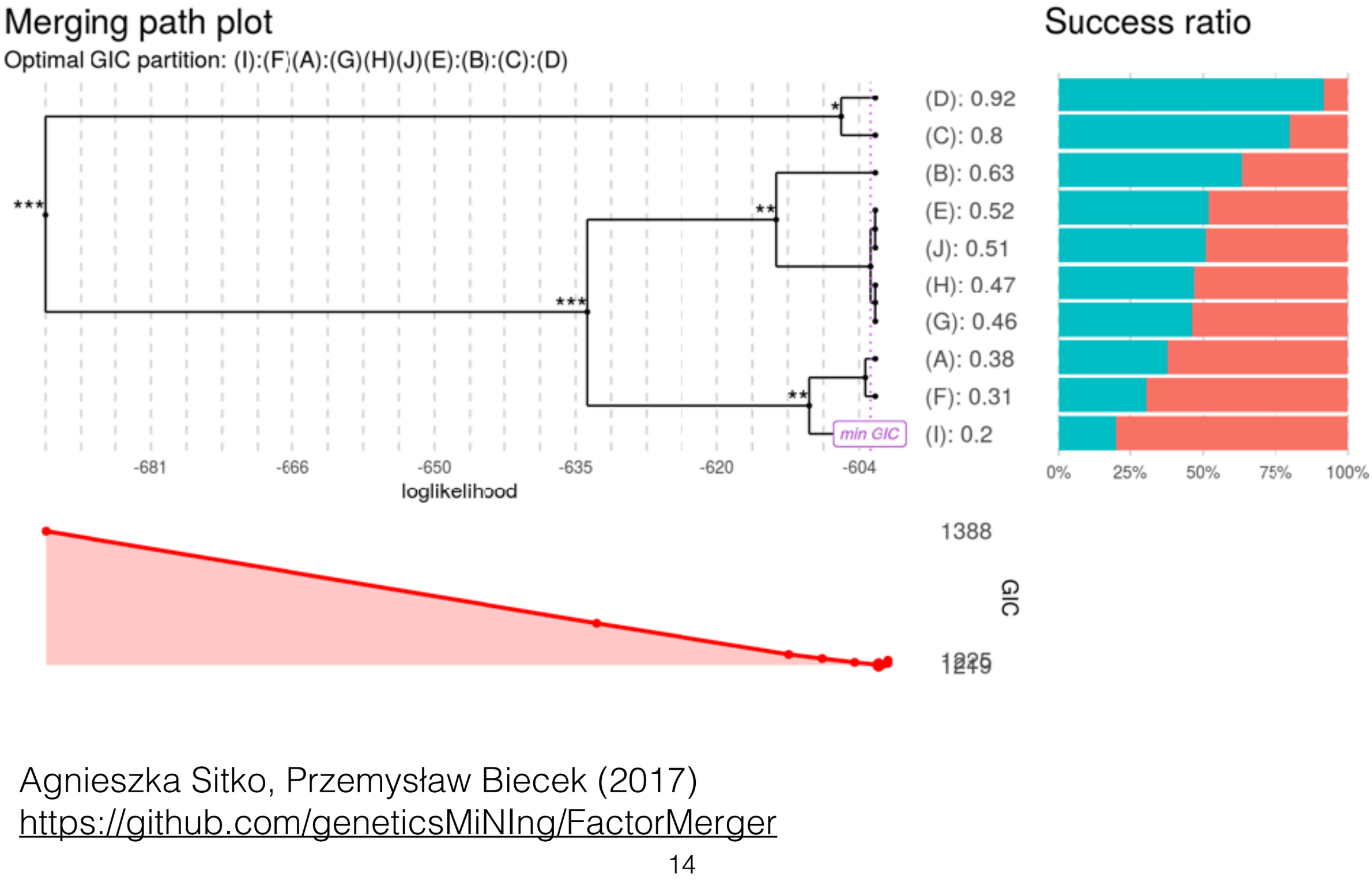
Efekt pojedynczej zmiennej

Pakiet factorMerger



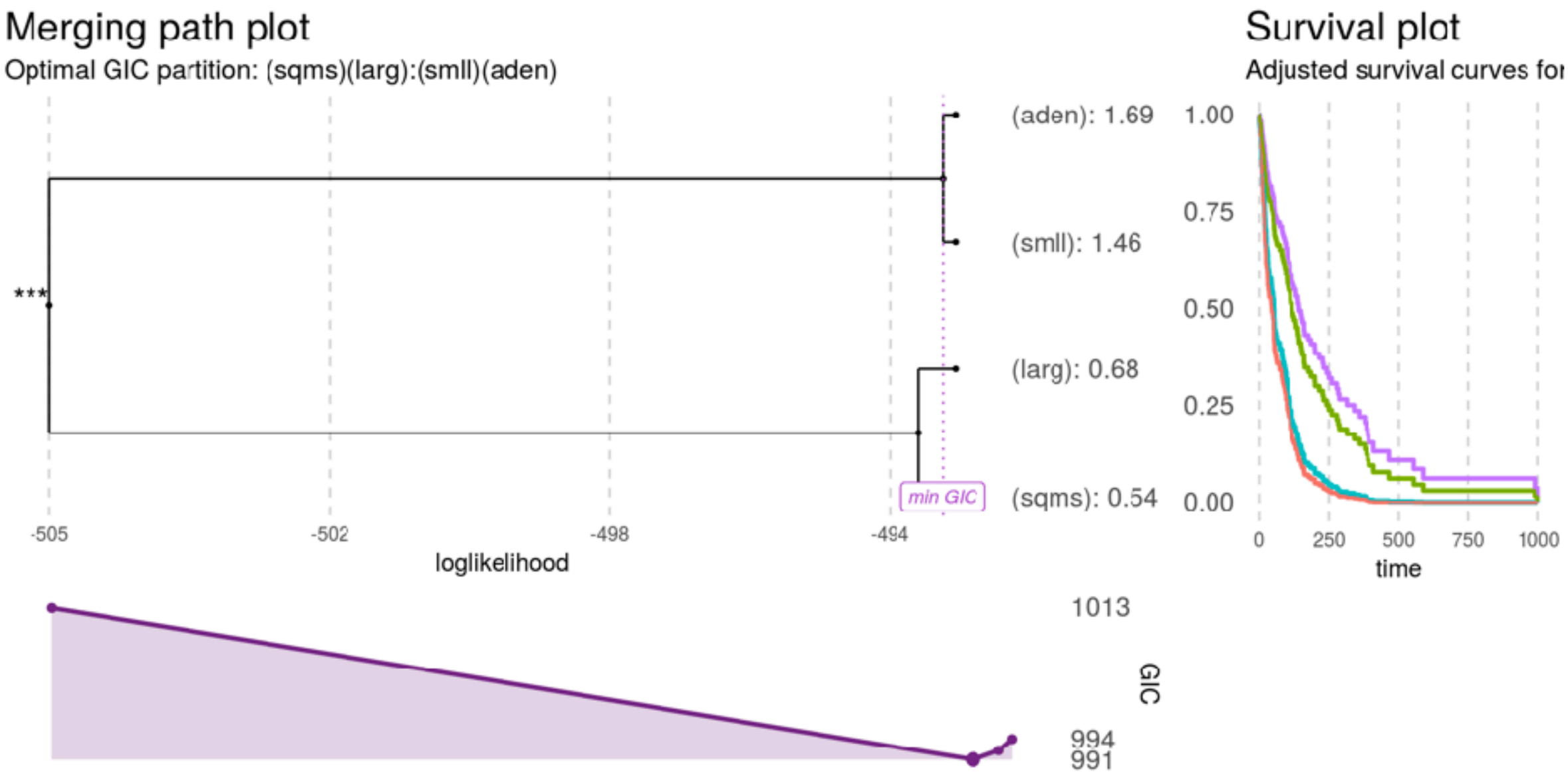
Agnieszka Sitko, Przemysław Biecek (2017)
https://github.com/geneticsMiNIng/FactorMerger

Pakiet factorMerger



Agnieszka Sitko, Przemysław Biecek (2017)
https://github.com/geneticsMiNIng/FactorMerger

Pakiet factorMerger



Agnieszka Sitko, Przemysław Biecek (2017)
https://github.com/geneticsMiNIng/FactorMerger

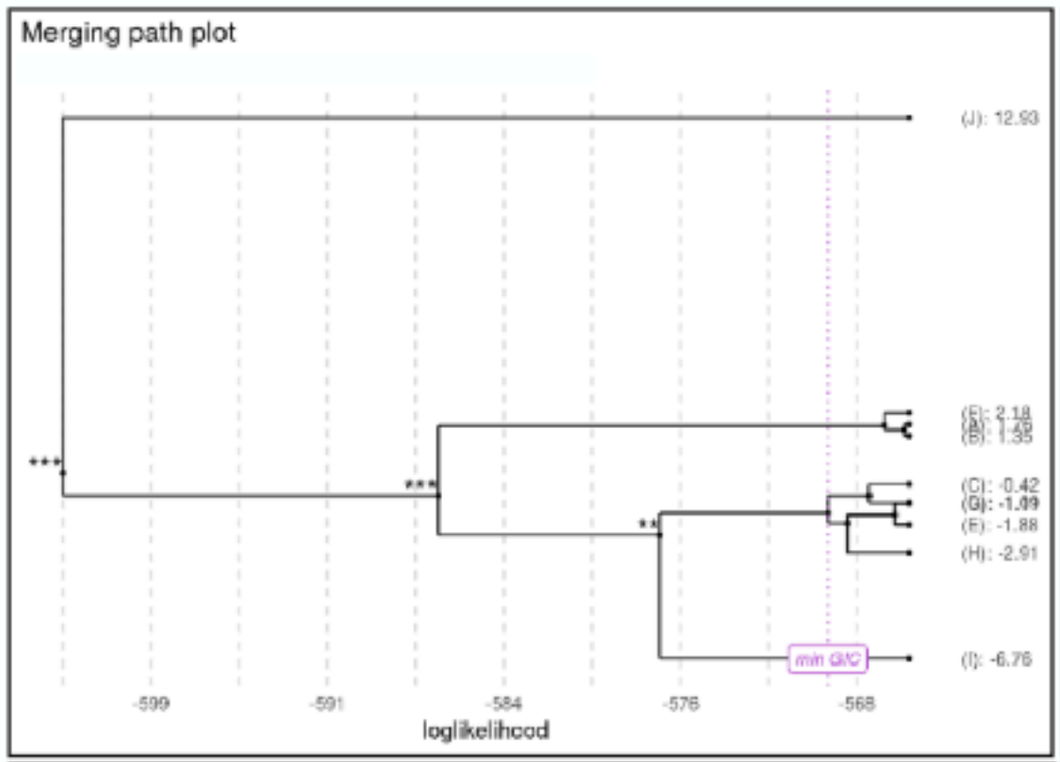
FactorMerger - set of tools to support results from post hoc testing

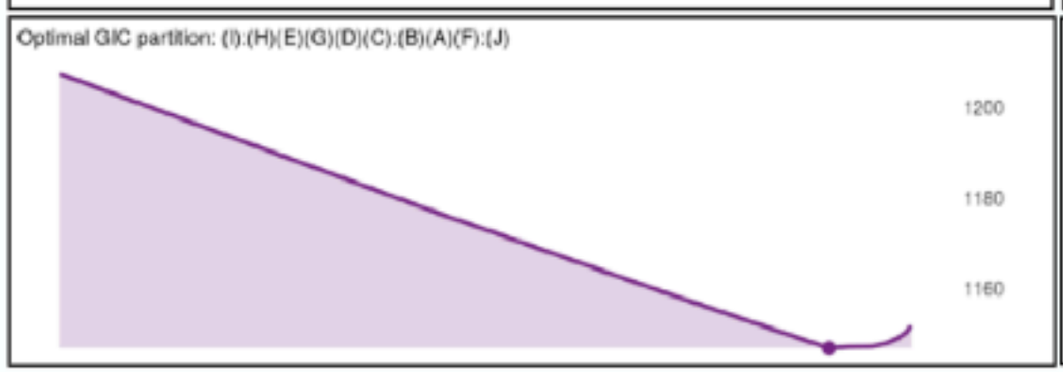
Level fusing plot

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





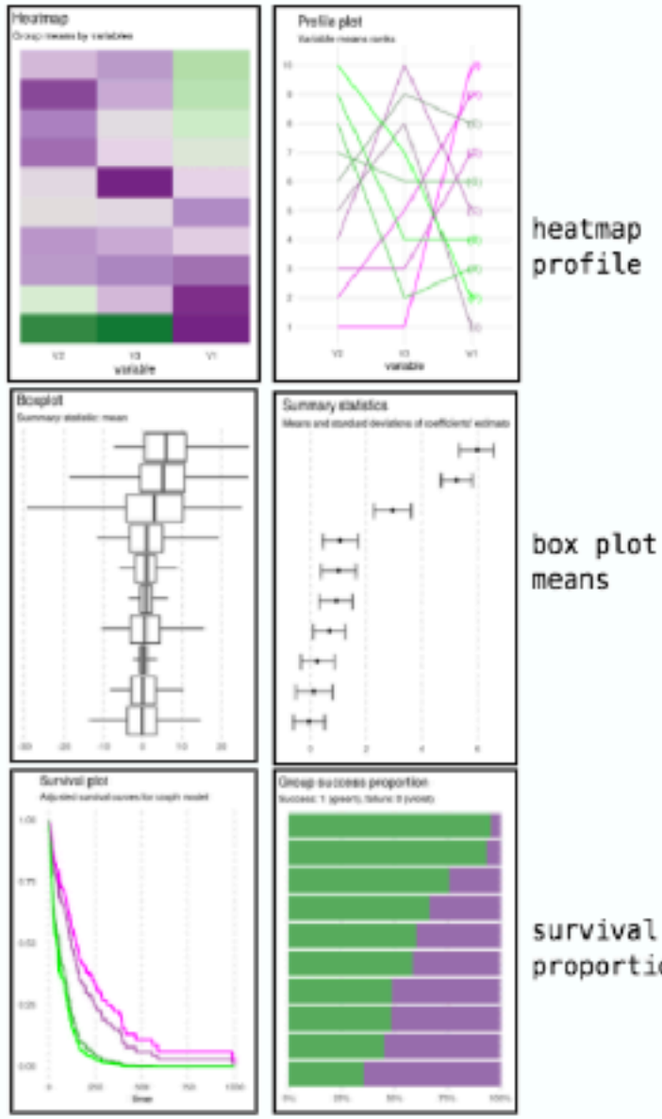
Model statistics

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



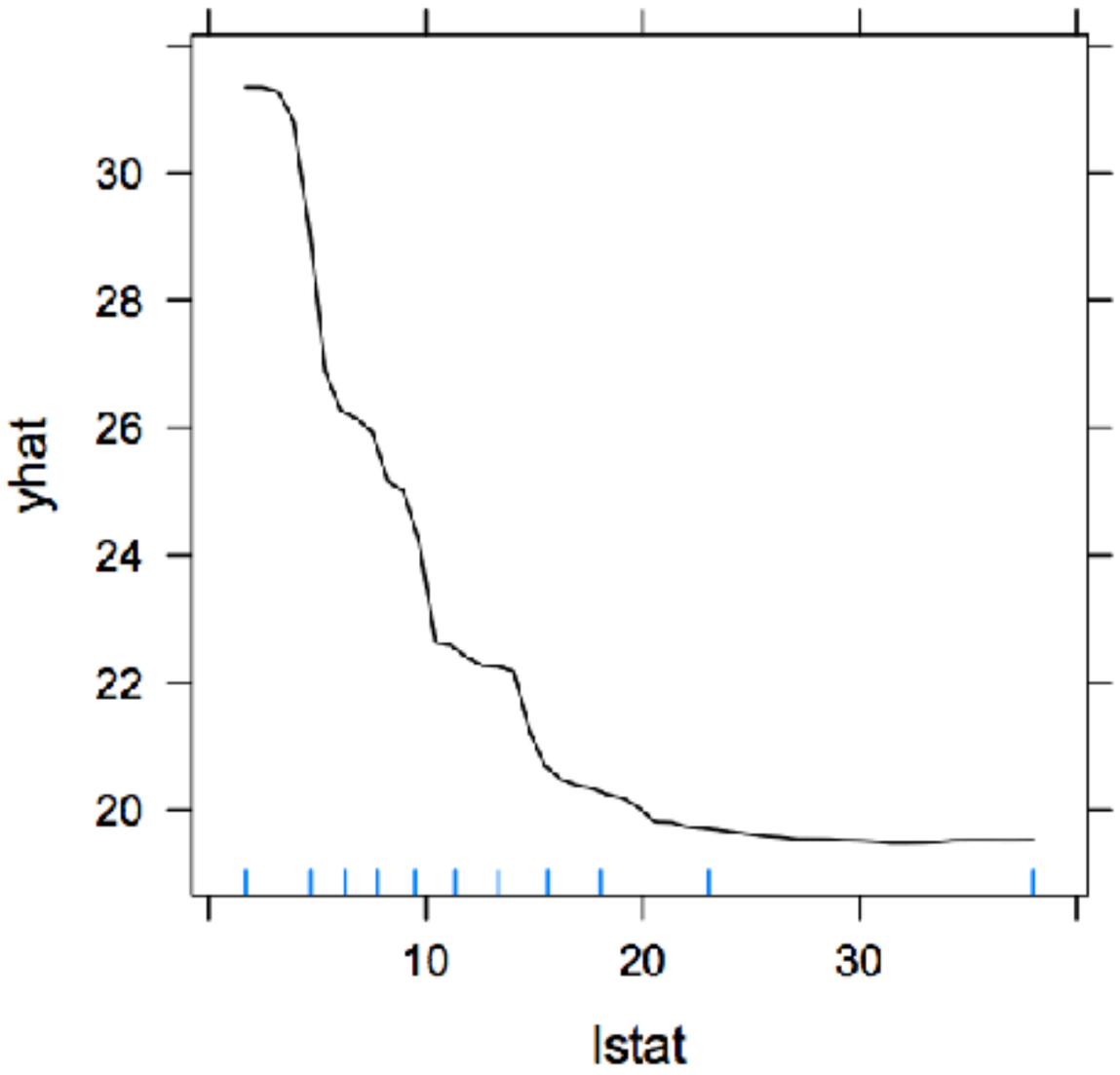
`heatmap`
`profile`

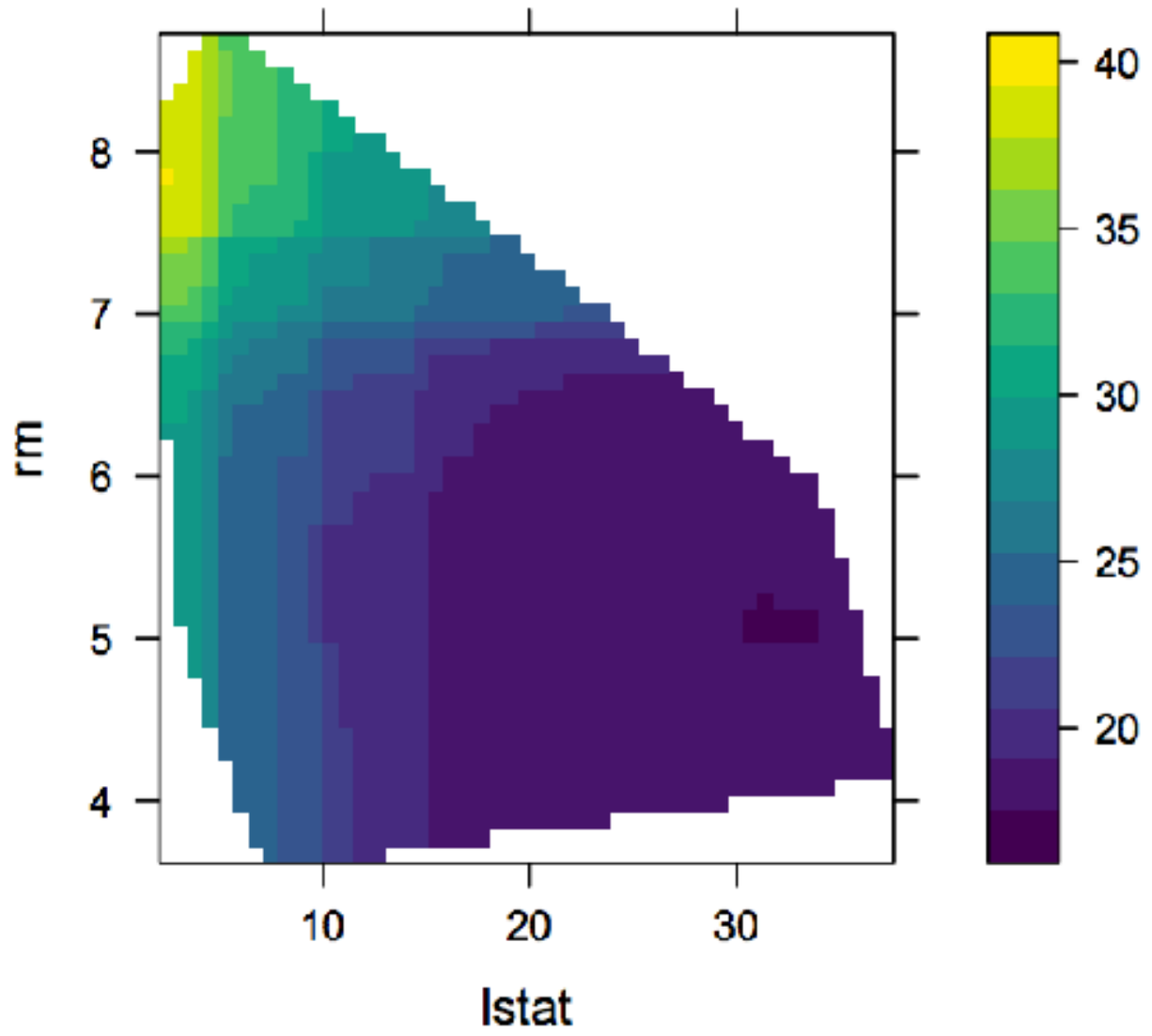
`box plot`
`means`

`survival`
`proportions`

Agnieszka Sitko, Przemysław Biecek (2017)
https://github.com/geneticsMiNIng/FactorMerger

Pakiet pdp - Partial Dependence Plots



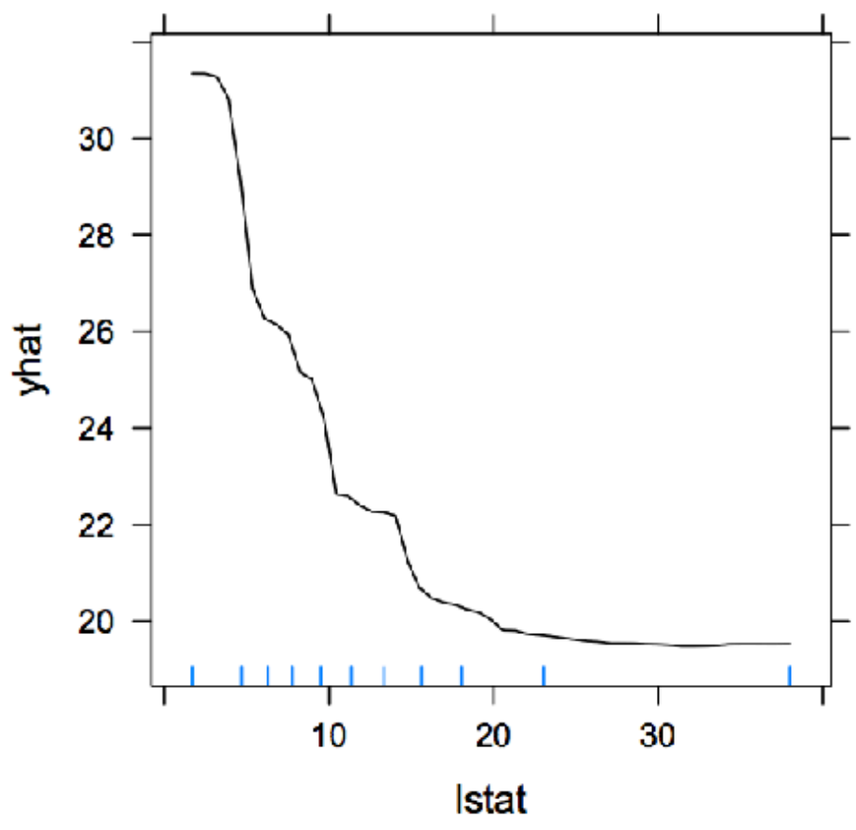


pdp: An R Package for Constructing Partial Dependence Plots. Brandon M. Greenwell (2017)
https://journal.r-project.org/archive/2017/RJ-2017-016/index.html

Pakiet pdp - Partial Dependence Plots

Type of model	R package	Object class
Decision tree	C50 (Kuhn et al., 2015)	"C5.0"
	party	"BinaryTree"
	partykit	"party"
	rpart (Therneau et al., 2015)	"rpart"
Bagged decision trees	adabag (Alfaro et al., 2013)	"bagging"
	ipred (Peters and Hothorn, 2015)	"classbagg", "regbagg"
Boosted decision trees	adabag (Alfaro et al., 2013)	"boosting"
	gbm	"gbm"
	xgboost	"xgb.Booster
Cubist	Cubist (Kuhn et al., 2014)	"cubist"
Discriminant analysis	MASS (Venables and Ripley, 2002)	"lda", "qda"
Generalized linear model	**stats**	"glm", "lm"
Linear model	**stats**	"lm"
Nonlinear least squares	**stats**	"nls"
Multivariate adaptive regression splines (MARS)	earth (Milborrow, 2016)	"earth"
	mda (Leisch et al., 2016)	"mars"
Projection pursuit regression	**stats**	"ppr"
Random forest	**randomForest**	"randomForest"
	party	"RandomForest"
	partykit	"cforest"
	ranger (Wright, 2016)	"ranger"
Support vector machine	e1071 (Meyer et al., 2015)	"svm"
	kernlab (Karatzoglo	

Table 1: Models specifically supported by the **pdp** pacl
may still need to supply additional arguments in the cal



pdp: An R Package for Constructing Partial Dependence Plots
Brandon M. Greenwell (2017)
https://journal.r-project.org/archive/2017/RJ-2017-016/index.html

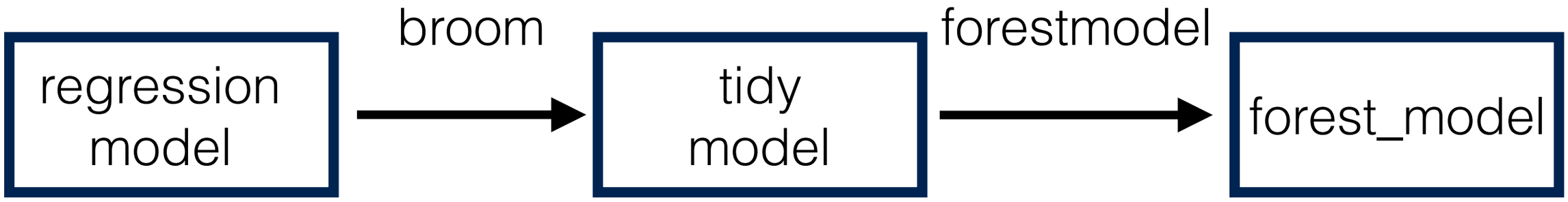
Struktura

Pakiet forestmodel

Variable		N	Hazard ratio	p
Age		180		1.01 (0.99, 1.03) 0.467
Sex	Male	113		Reference
	Female	67		0.60 (0.41, 0.88) 0.008
ECOG	0	50		Reference
	1	86		1.32 (0.85, 2.04) 0.215
	2	43		2.19 (1.33, 3.61) 0.002
	3	1		6.00 (0.79, 45.70) 0.083
Meal.Cal		180		1.00 (1.00, 1.00) 0.793

Nick Kennedy (2017) https://github.com/NikNakk/forestmodel

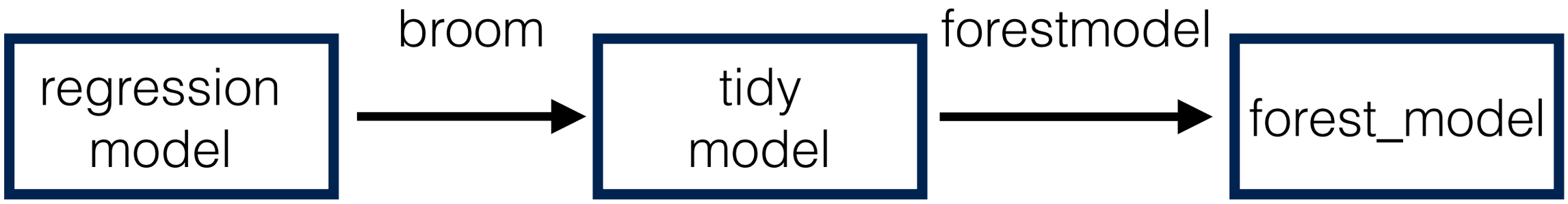
Pakiet forestmodel



Variable		N	Hazard ratio	p
Age		180		1.01 (0.99, 1.03) 0.467
Sex	Male	113		Reference
	Female	67		0.60 (0.41, 0.88) 0.008
ECOG	0	50		Reference
	1	86		1.32 (0.85, 2.04) 0.215
	2	43		2.19 (1.33, 3.61) 0.002
	3	1		6.00 (0.79, 45.70) 0.083

Nick Kennedy (2017) https://github.com/NikNakk/forestmodel

Pakiet forestmodel

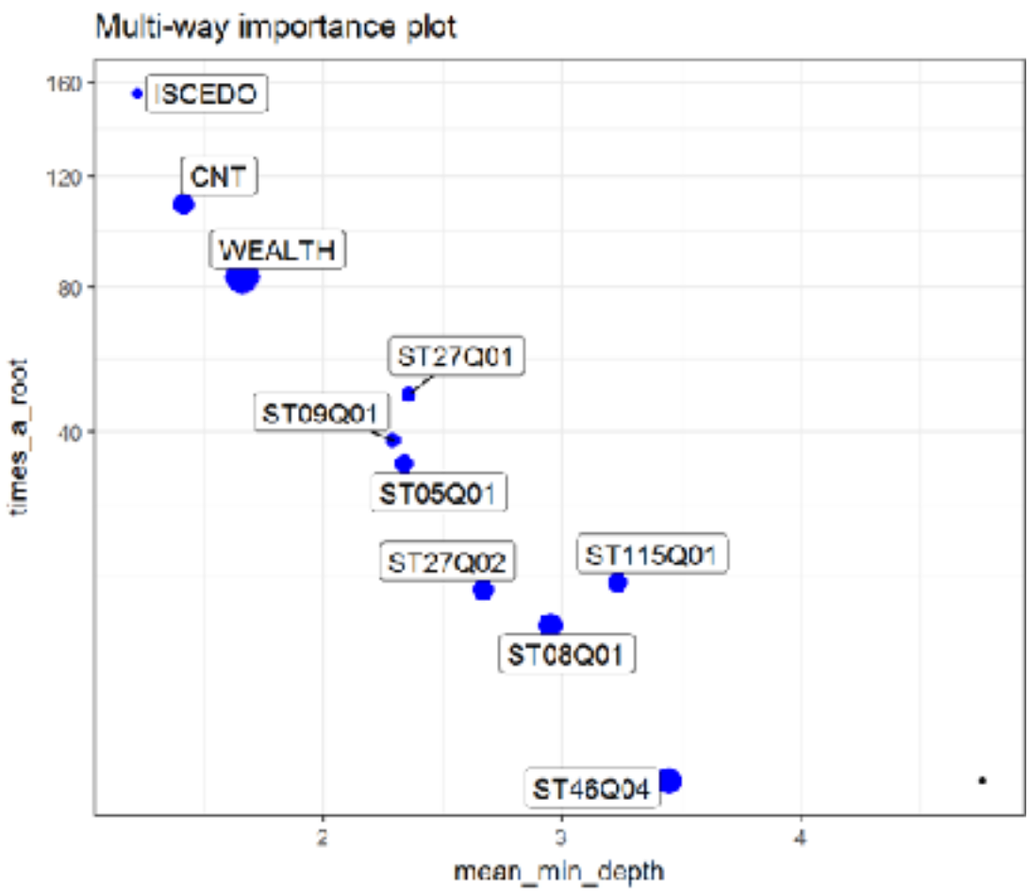


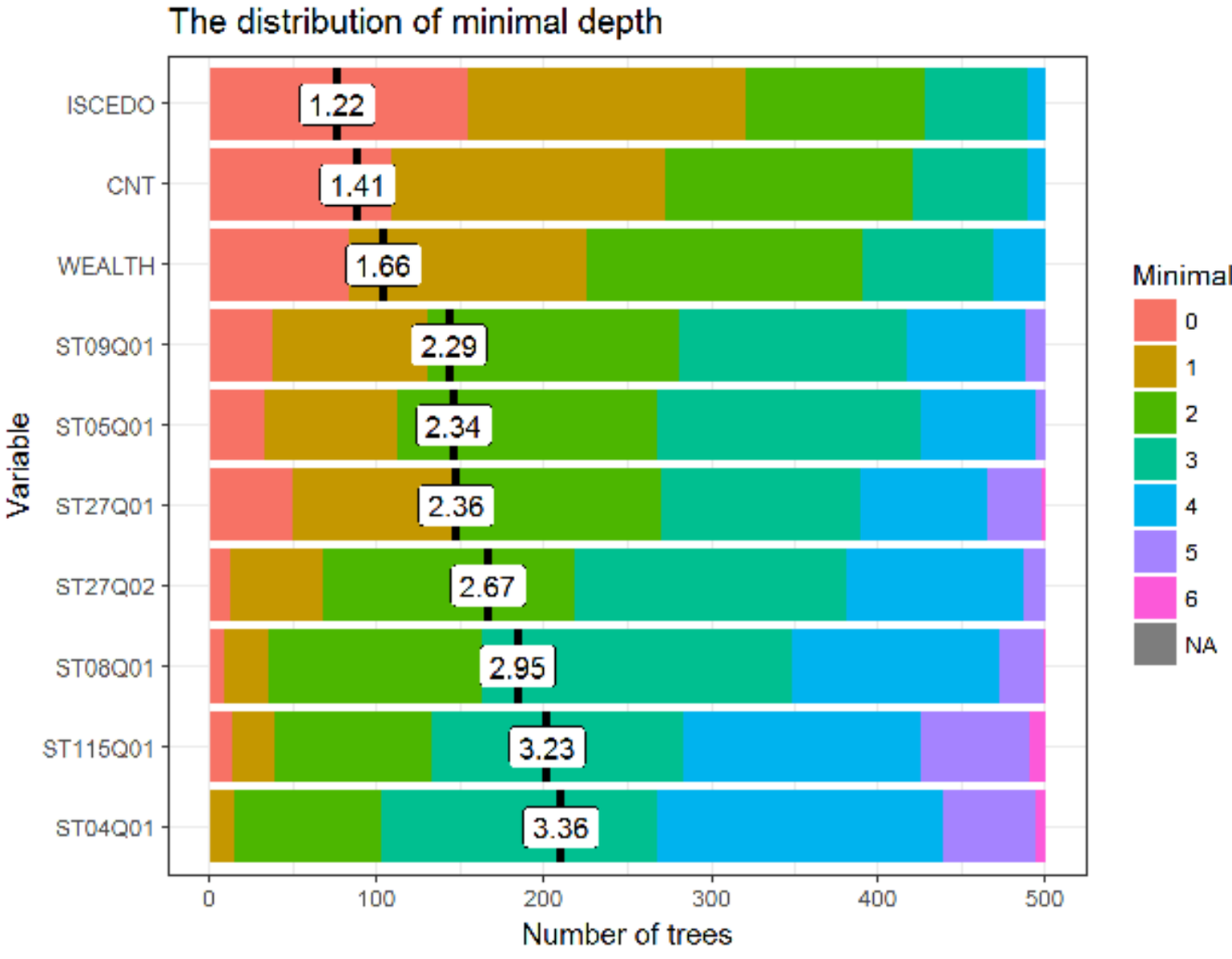
Class	tidy	glance
aareg	x	x
acf	x	
anova	x	
aov	x	
aovlist	x	
Arima	x	x
betareg	x	x
biglm	x	x

spec	x		
stanfit	x		
stanreg	x	x	
summary.glht	x		
summary.lm	x	x	
summaryDefault	x	x	
survexp	x	x	
survfit	x	x	
survreg	x	x	x

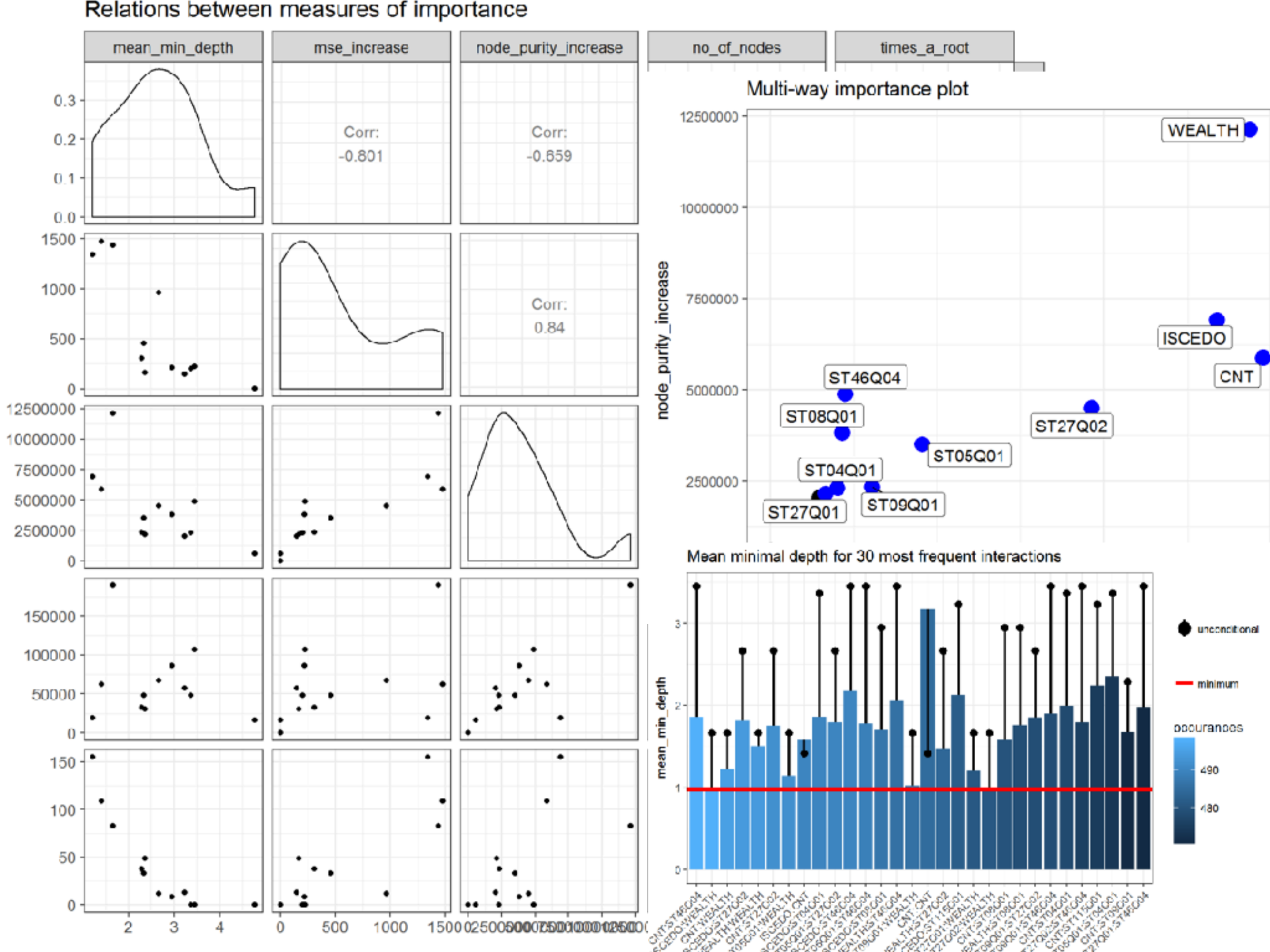
broom: An R Package for Converting Statistical Analysis Objects Into Tidy Data Frames, David Robinson (2014) arXiv:1412.3565v2 [stat.CO]

Pakiet randomForestExplainer
Co się kryje w lesie?





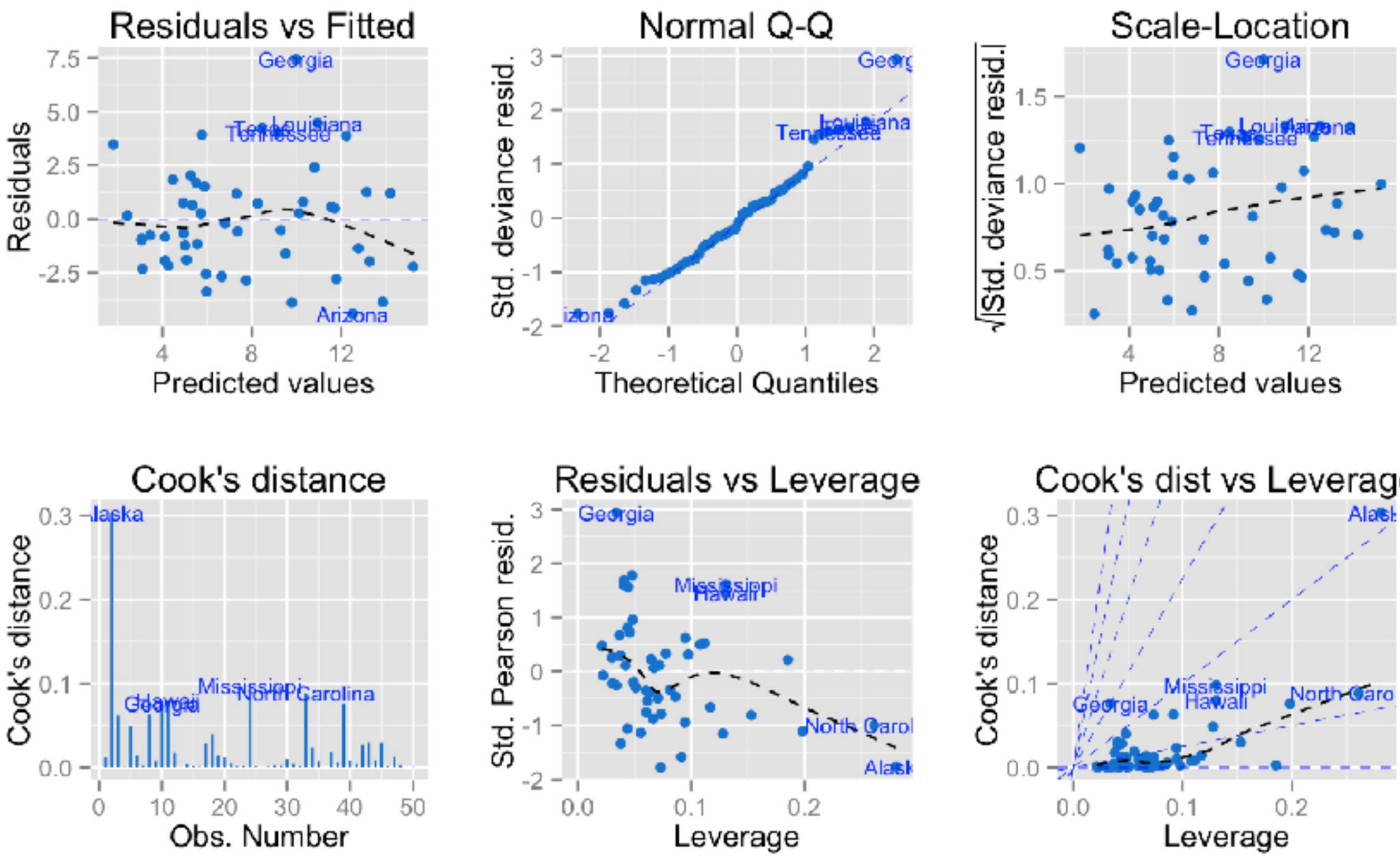
Aleksandra Paluszynska, Przemyslaw Biecek (2017)
https://github.com/geneticsMiNIng/BlackBoxOpener
John Ehrlinger (2015) ggRandomForests: Random Forests for Regression


Relations between measures of importance
mean_min_depth
mse_increase
node_purity_increase
no_of_nodes
times_a_root
Corr:
-0.801
Corr:
-0.659
Corr:
0.84
Multi-way importance plot
node_purity_increase
WEALTH
ISCEDO
CNT
ST46Q04
ST08Q01
ST27Q02
ST05Q01
ST04Q01
ST09Q01
ST27Q01
Mean minimal depth for 30 most frequent interactions
mean_min_depth
unconditional
minimum
occurances

Diagnostyka

Pakiet ggfortify

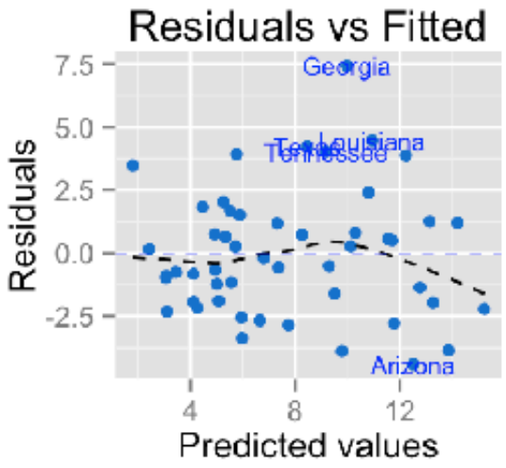
lm

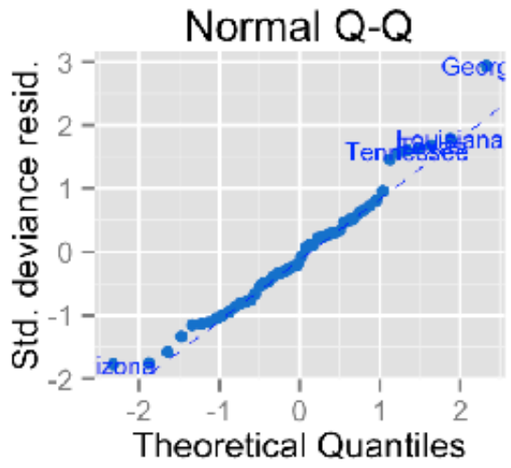


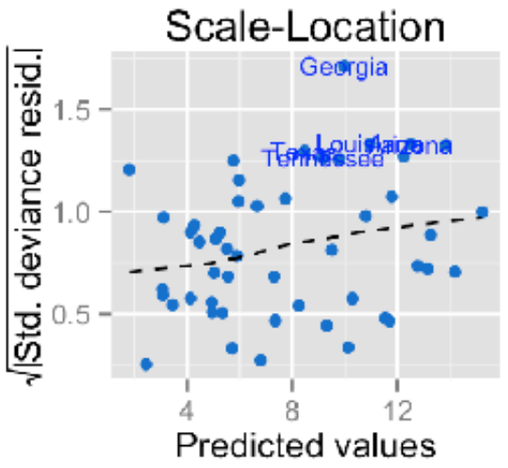
Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.

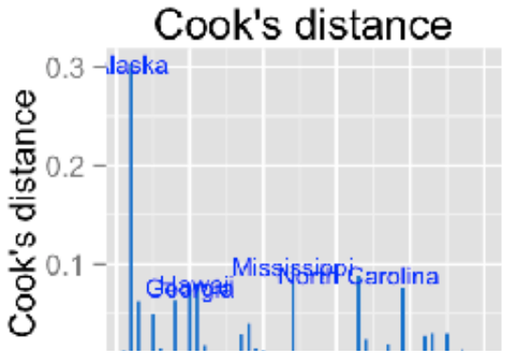
Pakiet ggfortify

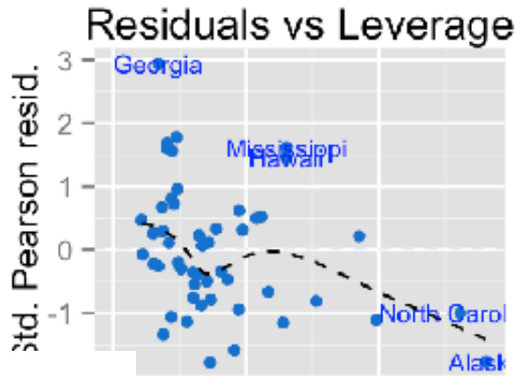
lm

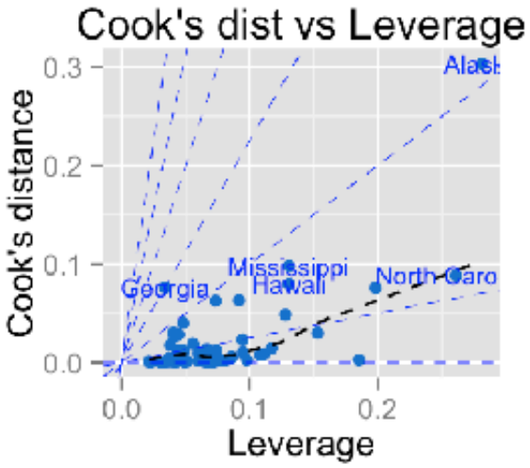




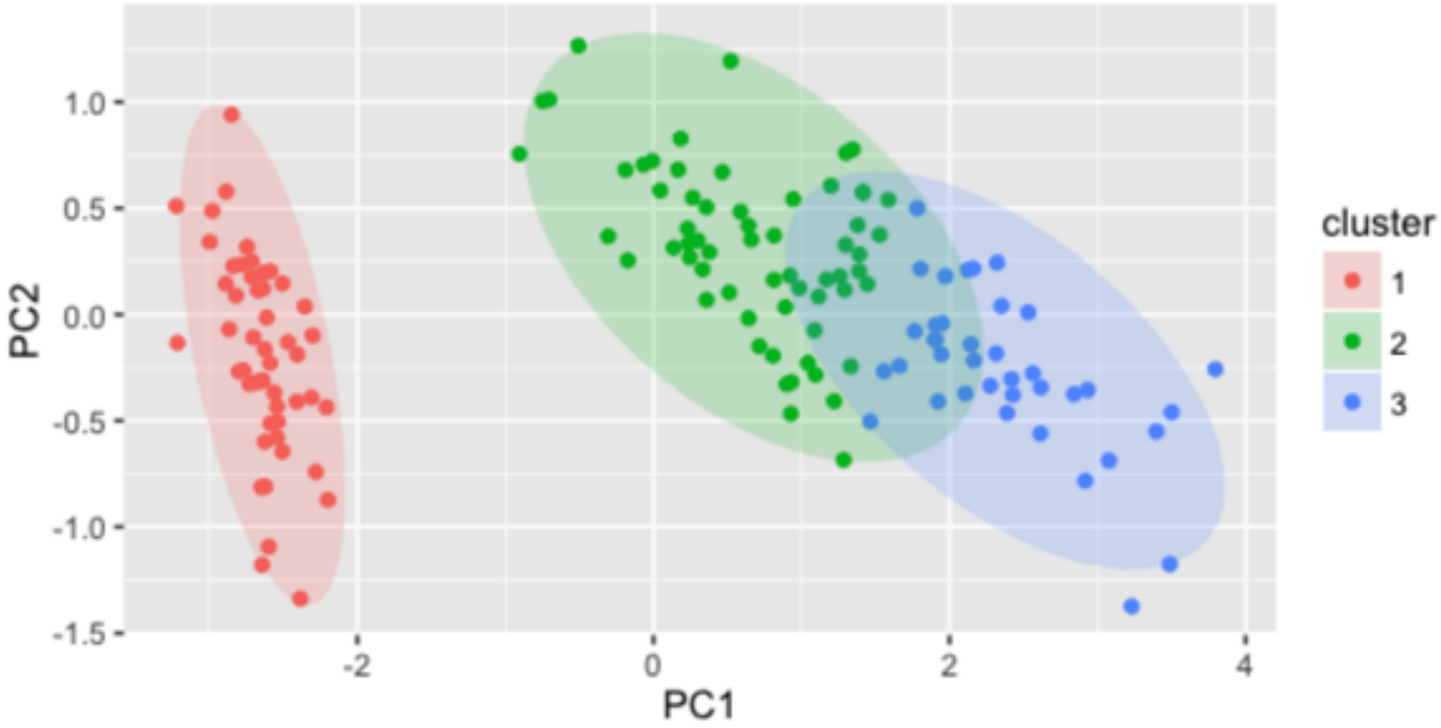






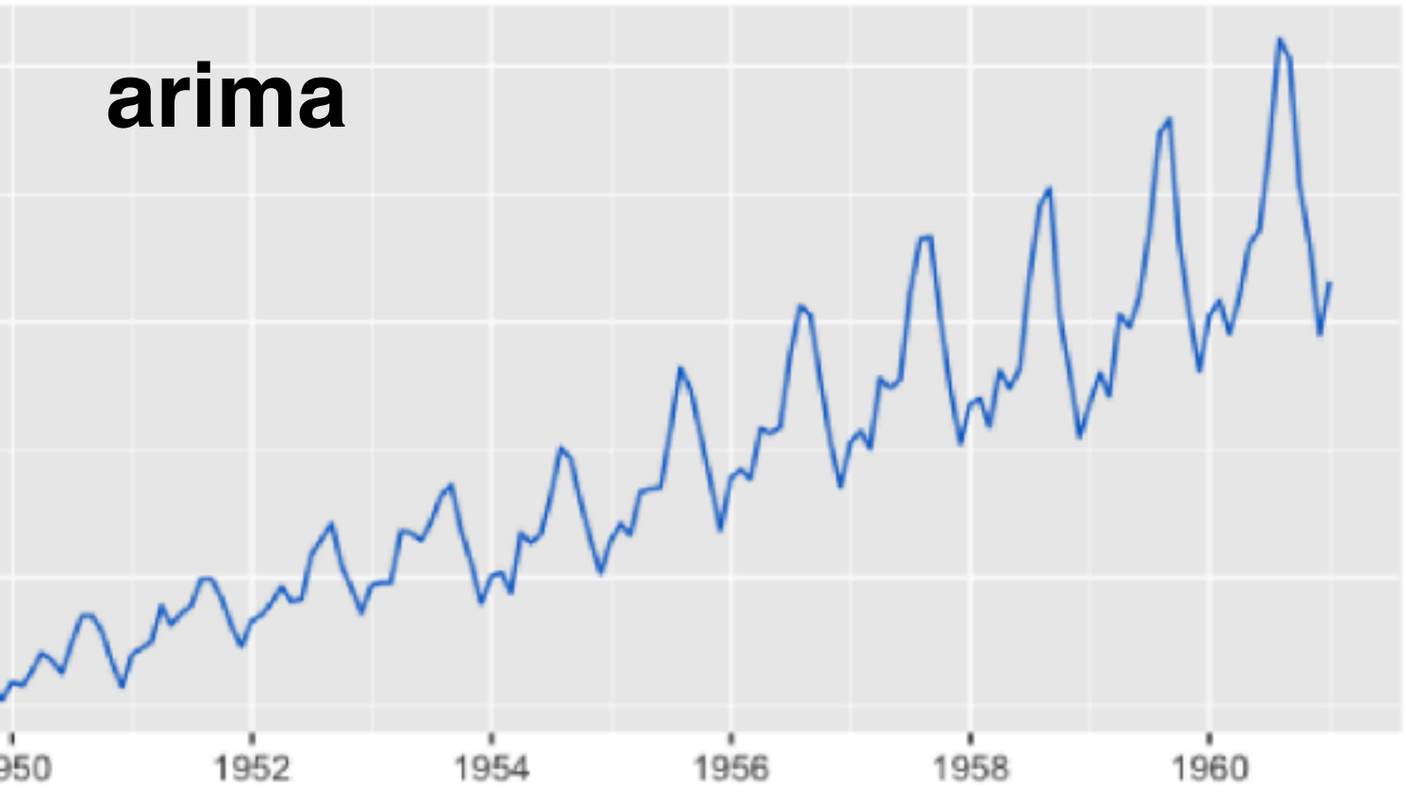


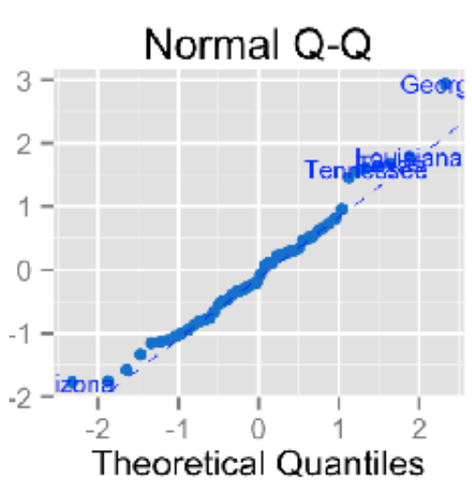
PAM

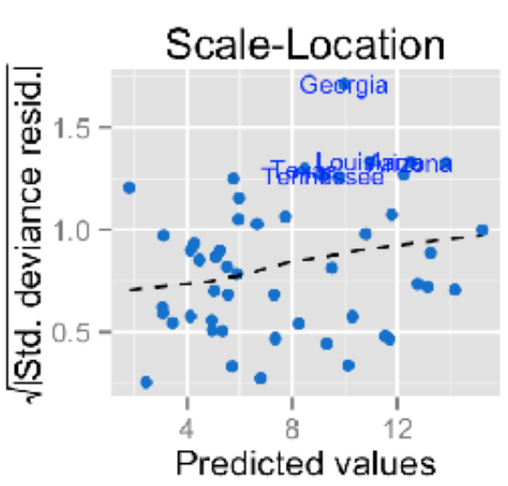


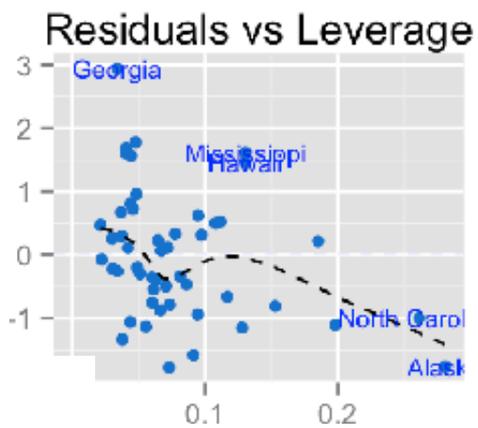
[Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.]

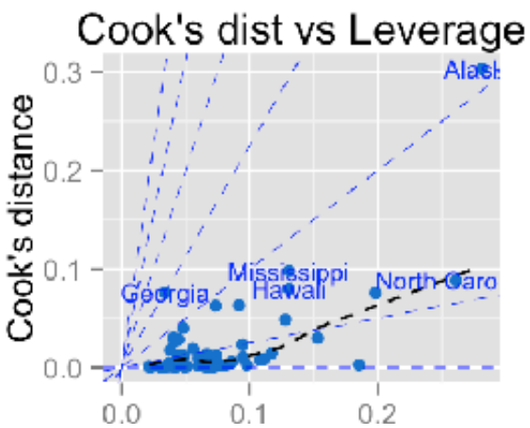
arima

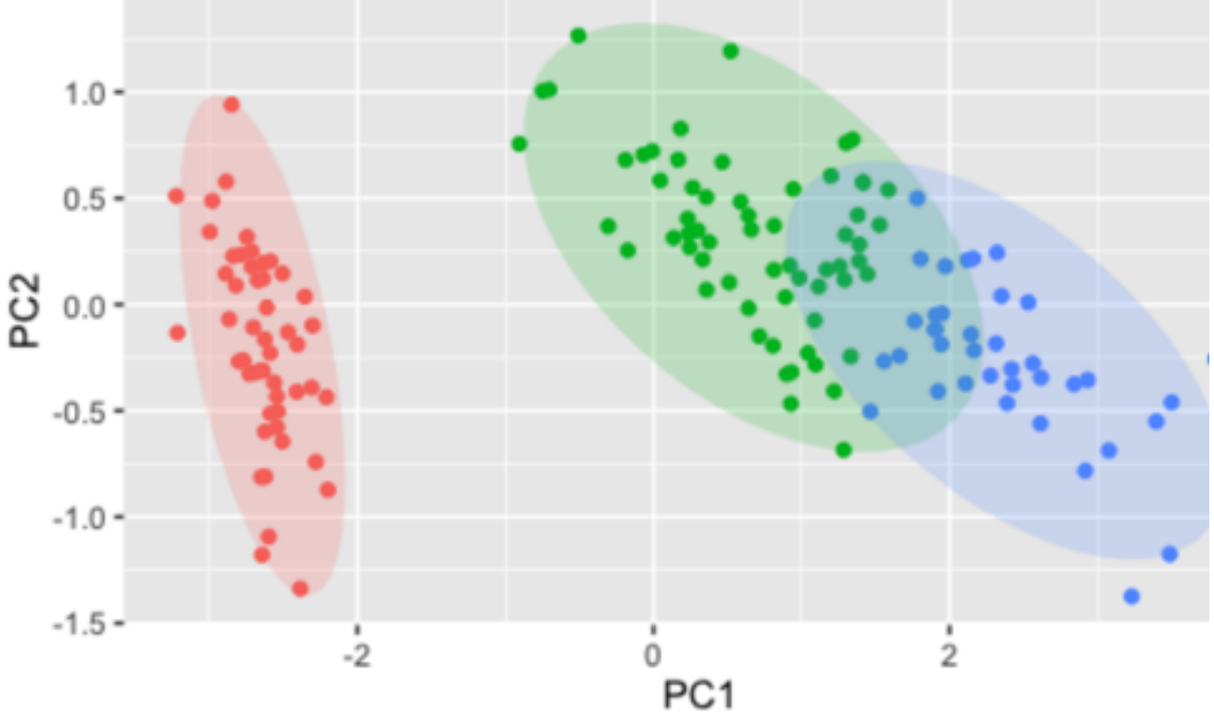






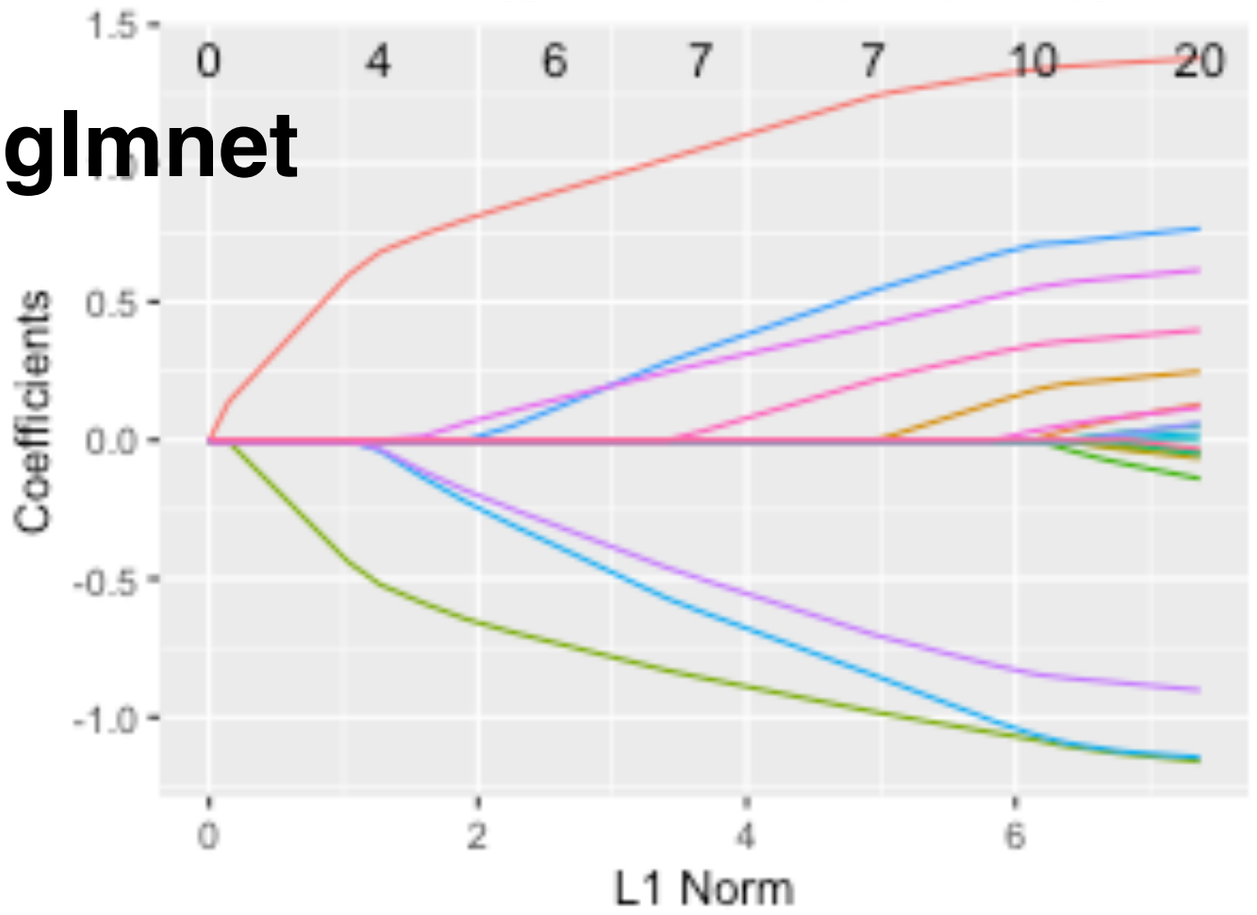


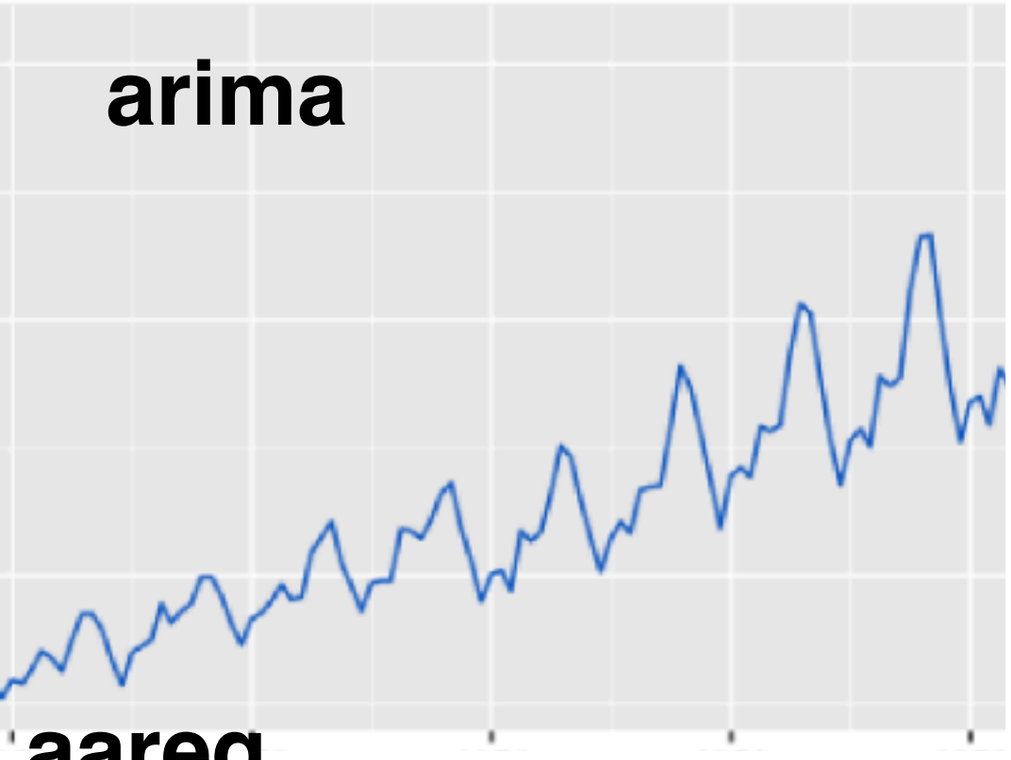


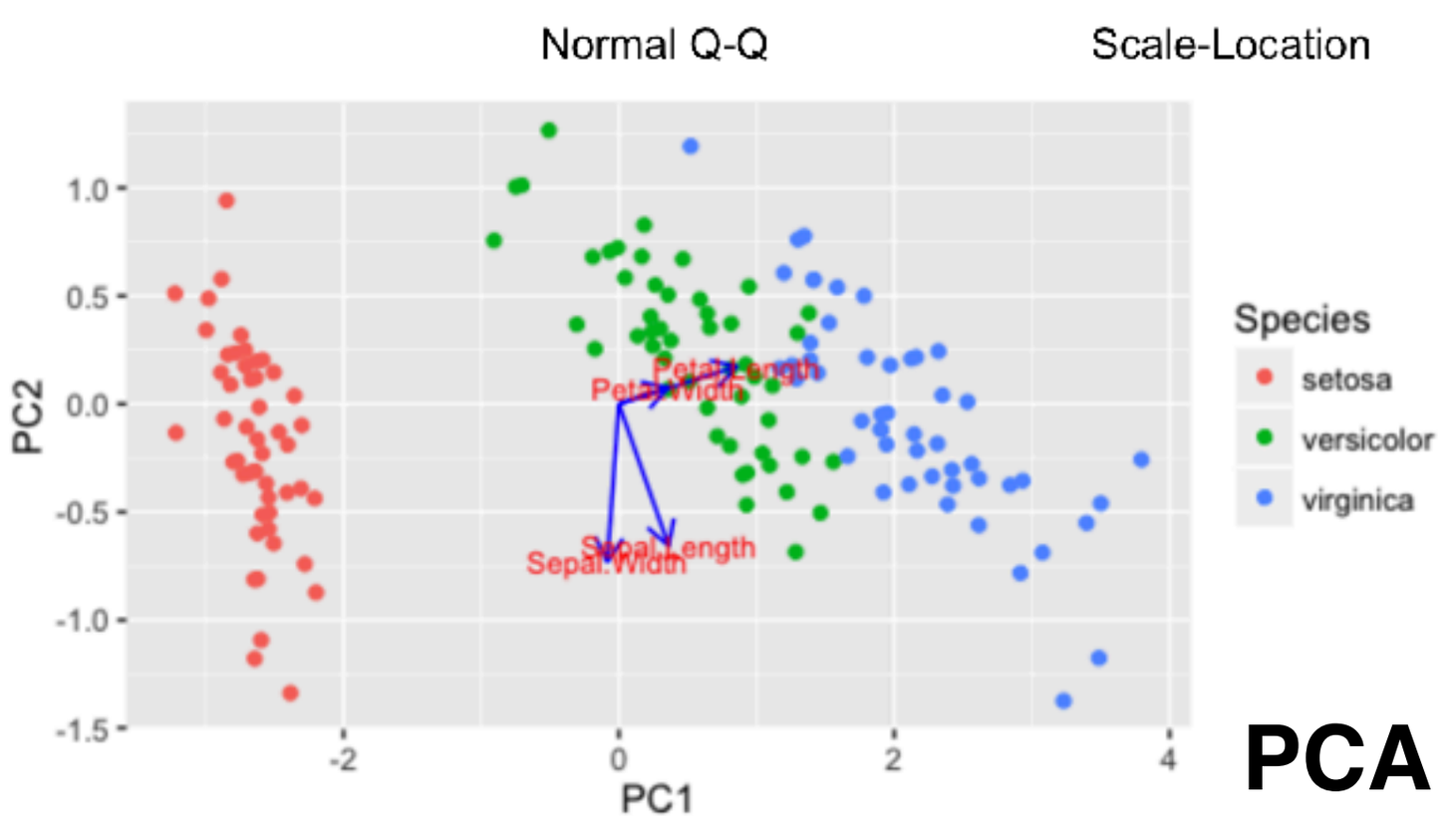


[Yuan Tang, Masaaki Horikoshi, and Wenxuan
Statistical Result of Popular R Packages." Th

glmnet

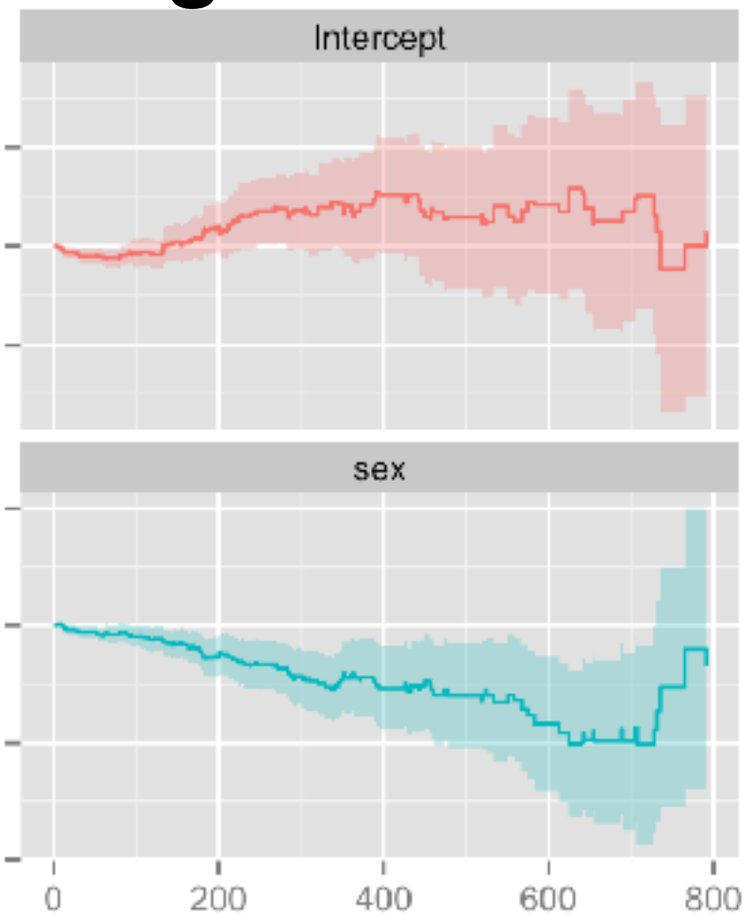


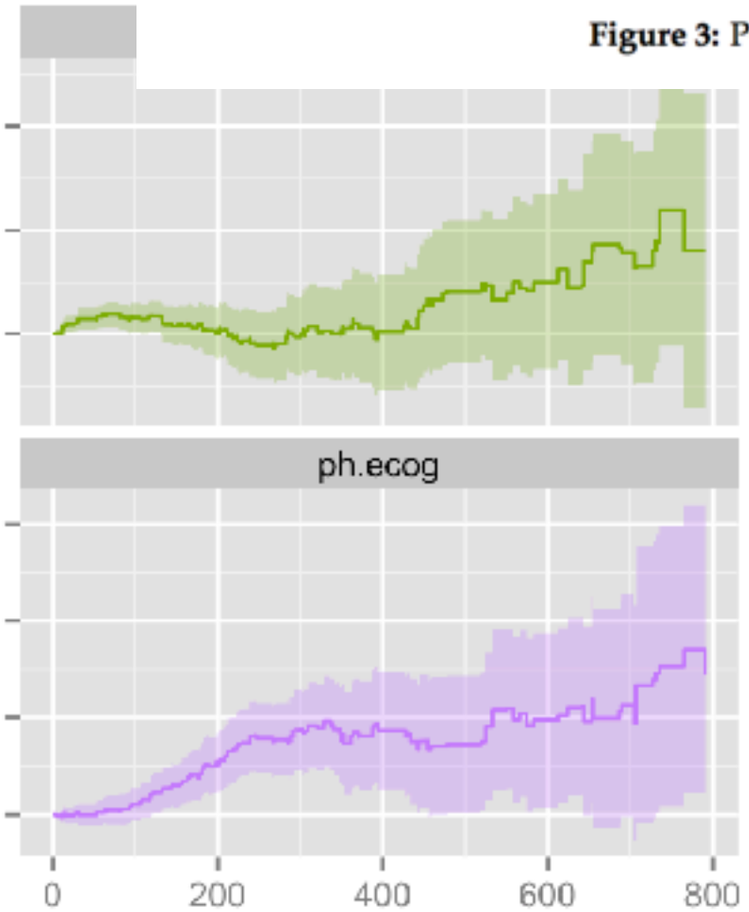


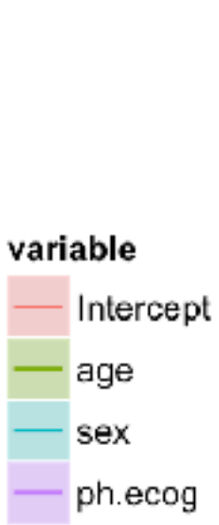


PCA

aareg







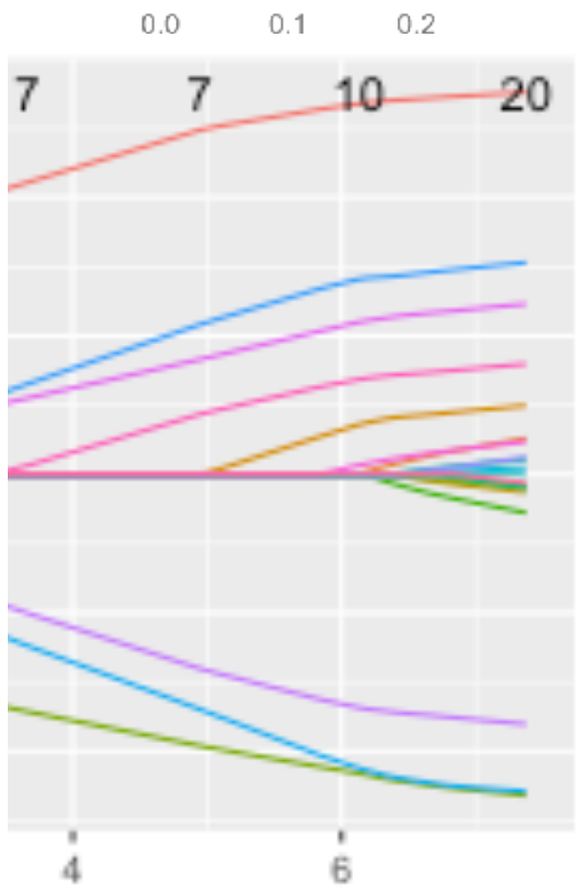


Figure 3: PCA with eigen-vectors and labels.

Pakiet ggfortify

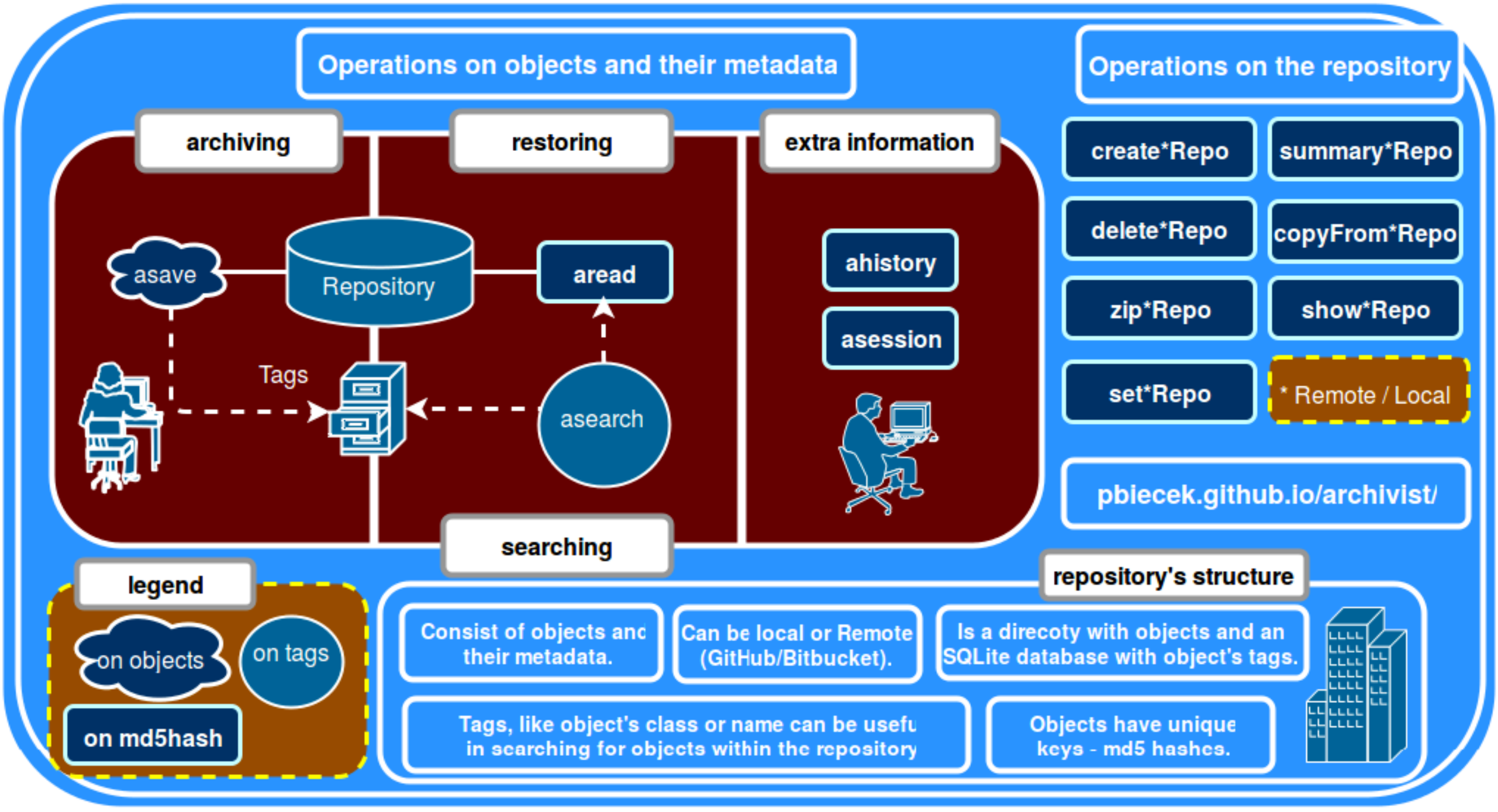
Table 1: Supported packages

package	supported types	package	supported types
base	"matrix", "table"	**sp**	"SpatialPoints", "SpatialPolygons", "Line", "Lines", "Polygon", "Polygons", "SpatialLines", "SpatialLinesDataFrame", "SpatialPointsDataFrame", "SpatialPolygonsDataFrame"
cluster	"clara", "fanny", "pam"	**stats**	"HoltWinters", "lm", "acf", "ar", "Arima", "stepfun", "stl", "ts", "cmdscale", "decomposed.ts", "density", "factanal", "glm", "kmeans", "princomp", "spec"
changepoint	"cpt"	**survival**	"survfit", "survvfit.cox"
dlm	"dlmFilter", "dlmSmooth"	**strucchange**	"breakpoints", "breakpointsfull"
fGarch	"fGARCH"	**timeSeries**	"timeSeries"
forecast	"bats", "forecast", "ets", "nnetar"	**tseries**	"irts"
fracdiff	"fracdiff"	**vars**	"varprd"
glmnet	"cv.glmnet", "glmnet"	**xts**	"xts"
KFAS	"KFS", "signal"	**zoo**	"zooreg"
lfda	"lfda", "klfda", "self"	**MASS**	"isoMDS", "sammon"
maps	"map"		

[Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.]

Dostępność modeli

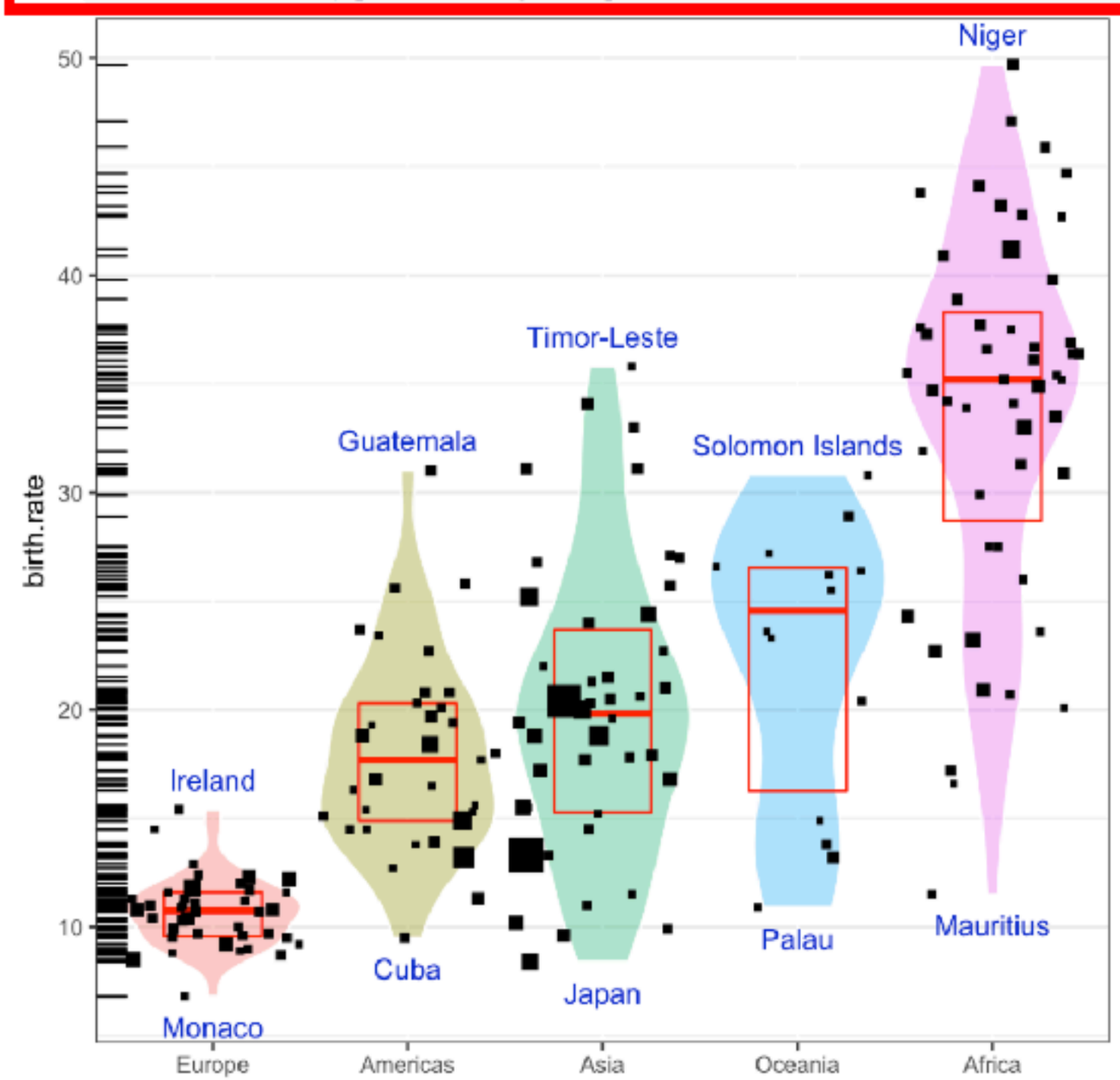
archivist - reproducible and recordable research



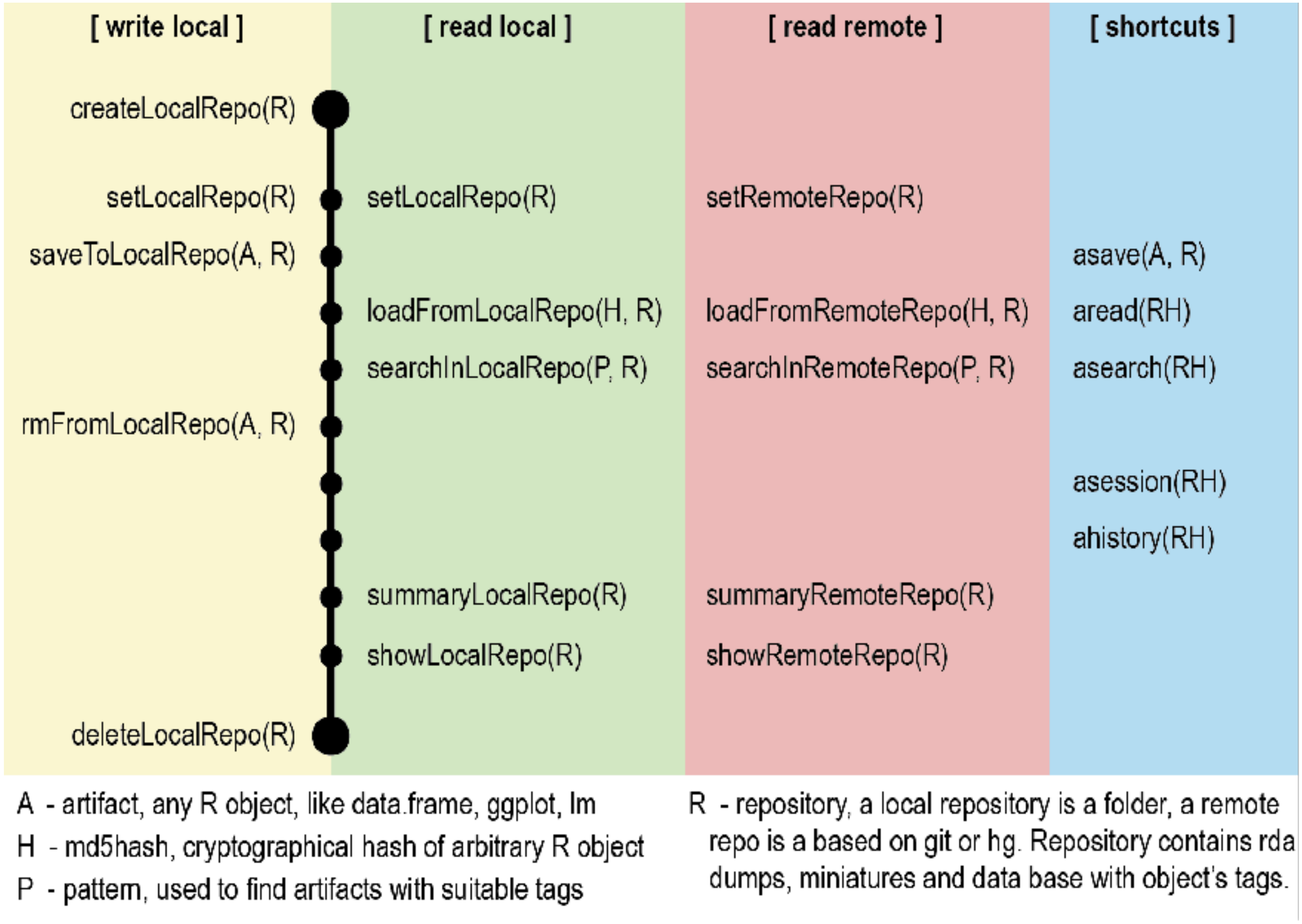
Przemysław Biecek, Marcin Kosiński (2015) https://github.com/pbiecek/archivist

```
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')
```



archivist - reproducible and recordable research

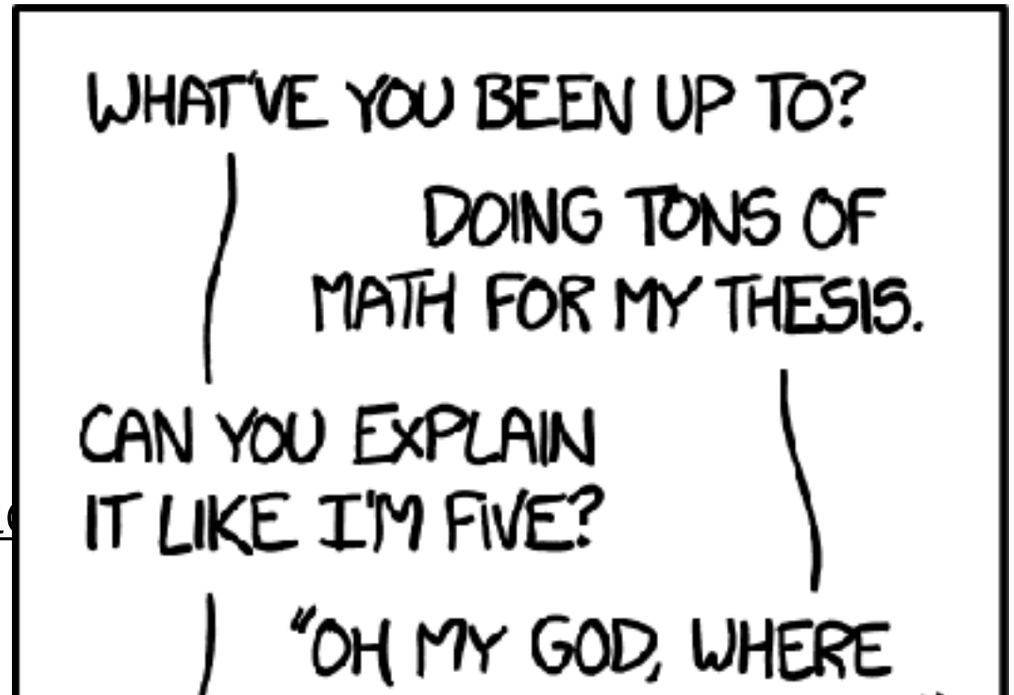


Przemysław Biecek, Marcin Kosiński (2015) https://github.com/pbiecek/archivist

ELI5 is a Python library which allows to visualize and debug various Machine Learning models
http://eli5.readthedocs.io/en/latest/index.html

Ideas on interpreting machine learning.
Patrick Hall, Wen Phan, SriSatish Ambati (2017)
https://www.oreilly.com/ideas/ideas-on-interpreting-ma

Visualizing statistical models: Removing the blindfold.
Hadley Wickham, Dianne Cook, Heike Hofmann (2015)
Statistical Analysis and Data Mining
http://had.co.nz/stat645/model-vis.pdf



https://imgs.xkcd.com/comics/like_im_five.png