TheKappaStatistic PaulCzodrowski

Paul Czodrowski
Technische Universität Dortmund
44 PUBLICATIONS **1,486** CITATIONS

The *kappa* statistic: Taking care of background rates

Paul Czodrowski

Merck KGaA

Small Molecule Platform, Computational Chemistry

Darmstadt, Germany

Gordon Research Conference CADD,

Mound Snow, July 2013

Stop me if you think you‘ve heard this one before!

We want to train accurate models.

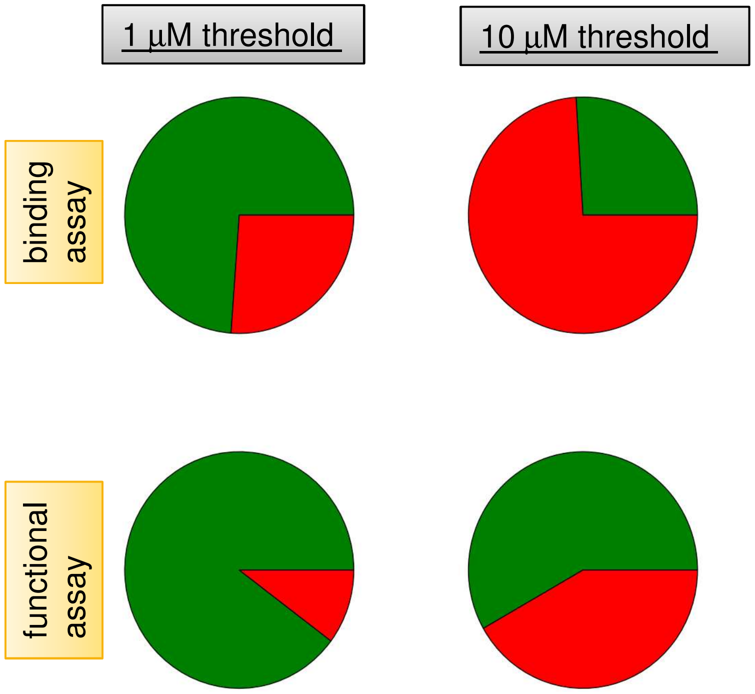
We do not want miss the positives.

For the positive predictions, no false-positives shall be predicted.

We want to be better than a trivial model!

Data used throughout the talk

ChEMBL



TN	FP
FN	TP

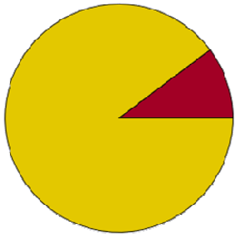
🟥 hERG active

🟩 hERG inactive

[Paper under review; all source code & data provided]

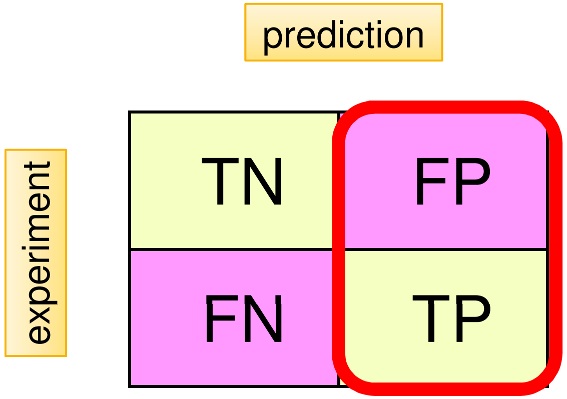
toy data

		prediction	
		inactive	active
experiment	inactive	TN	FP
	active	FN	TP



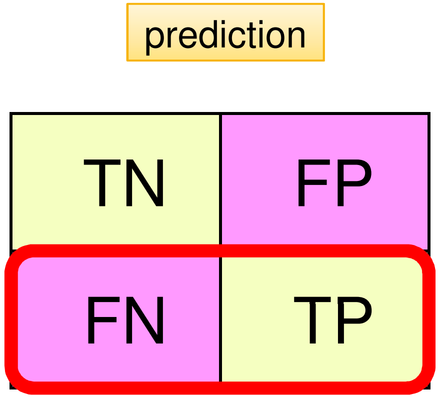
	pred	
exp	0	71
	0	623

Some figures of merit



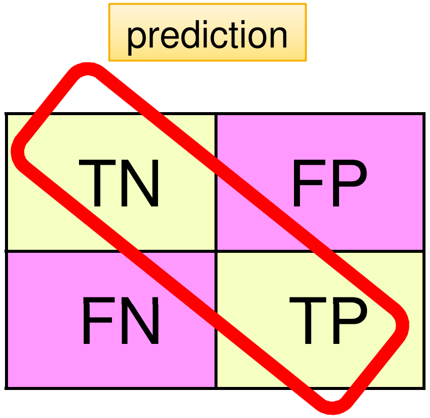
$$precision = \frac{TP}{TP + FP}$$

what fraction of positively labeled points are correctly **labeled**



$$recall = \frac{TP}{TP + FN}$$

what fraction of positive samples are correctly **identified**

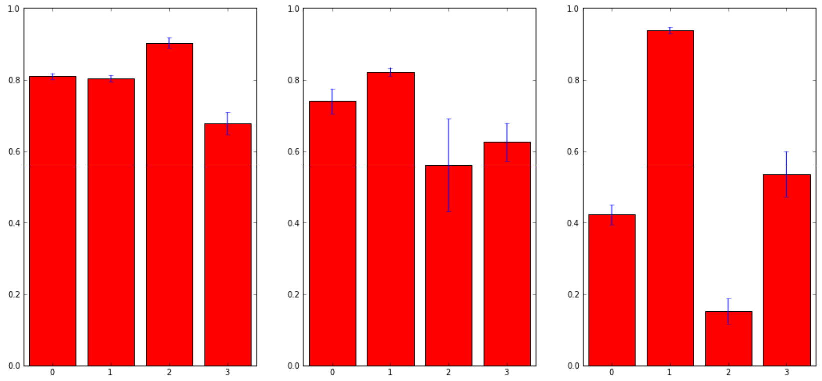


$$accuracy = \frac{TP + TN}{N}$$

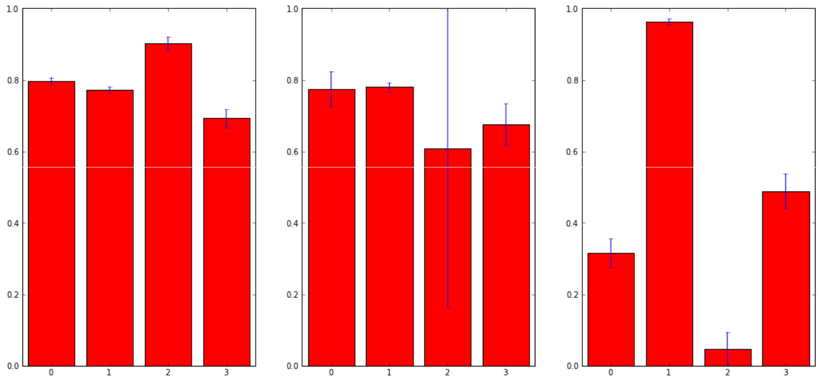
Binary classification: model performance

2 assays / 2 thresholds → 4 models

RF



SVM



accuracy	precision	recall
0.81±0.008	0.74±0.035	0.422±0.028
0.804±0.009	0.822±0.012	0.938±0.009
0.903±0.015	0.561±0.13	0.152±0.035
0.678±0.031	0.626±0.053	0.536±0.064

accuracy	precision	recall
0.797±0.01	0.775±0.049	0.316±0.04
0.773±0.009	0.781±0.012	0.962±0.009
0.903±0.018	0.608±0.444	0.046±0.048
0.693±0.026	0.676±0.059	0.489±0.049

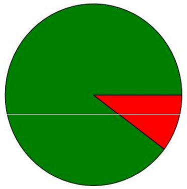
10 random train/test splits – mean/stdev values are given

(ir)relevance of accuracy/precision/recall

accuracy	precision	recall
0.90	1.00	0.03

Functional assay threshold = 1 µM

experimental situation



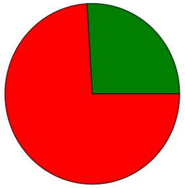
exemplary confusion matrix

experiment	prediction: inactive	prediction: active
inactive	247	0
active	29	1

accuracy	precision	recall
0.79	0.80	0.96

Binding assay threshold = 10 µM

experimental situation



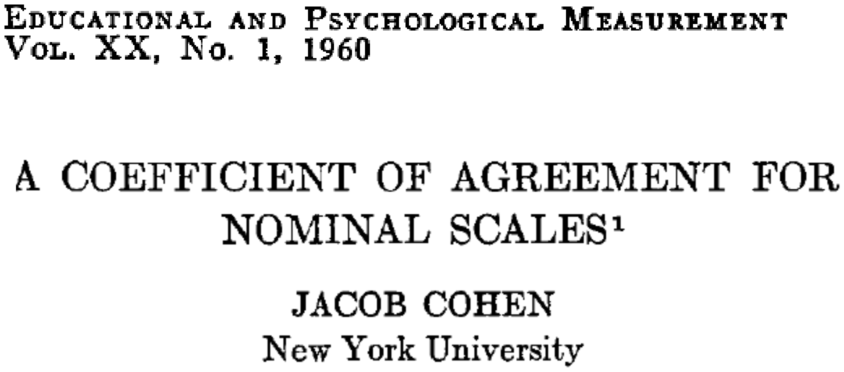
exemplary confusion matrix

experiment	prediction: inactive	prediction: active
inactive	97	276
active	44	1072

What is the background noise?

Jacob Cohen

Originally, Cohen developed κ
to estimate
the inter-rater reliability.

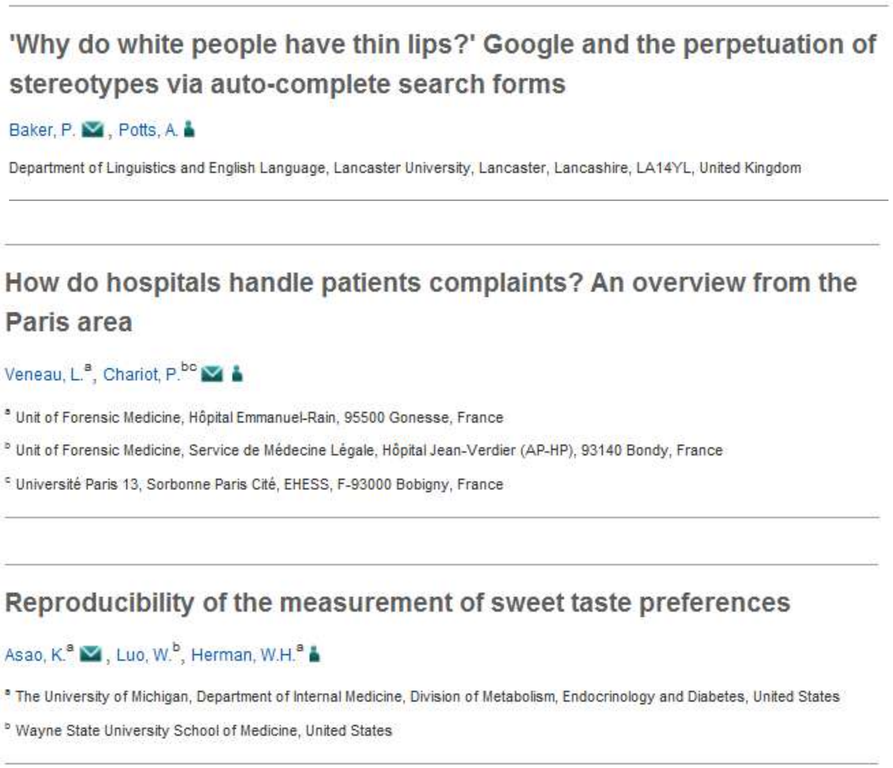


EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XX, No. 1, 1960

A COEFFICIENT OF AGREEMENT FOR NOMINAL SCALES[1]

JACOB COHEN
New York University

some recent studies using Cohens‘s κ



'Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms

Baker, P. ✉, Potts, A.

Department of Linguistics and English Language, Lancaster University, Lancaster, Lancashire, LA14YL, United Kingdom

How do hospitals handle patients complaints? An overview from the Paris area

Veneau, L.[a], Chariot, P.[bc] ✉

[a] Unit of Forensic Medicine, Hôpital Emmanuel-Rain, 95500 Gonesse, France

[b] Unit of Forensic Medicine, Service de Médecine Légale, Hôpital Jean-Verdier (AP-HP), 93140 Bondy, France

[c] Université Paris 13, Sorbonne Paris Cité, EHESS, F-93000 Bobigny, France

Reproducibility of the measurement of sweet taste preferences

Asao, K.[a] ✉, Luo, W.[b], Herman, W.H.[a]

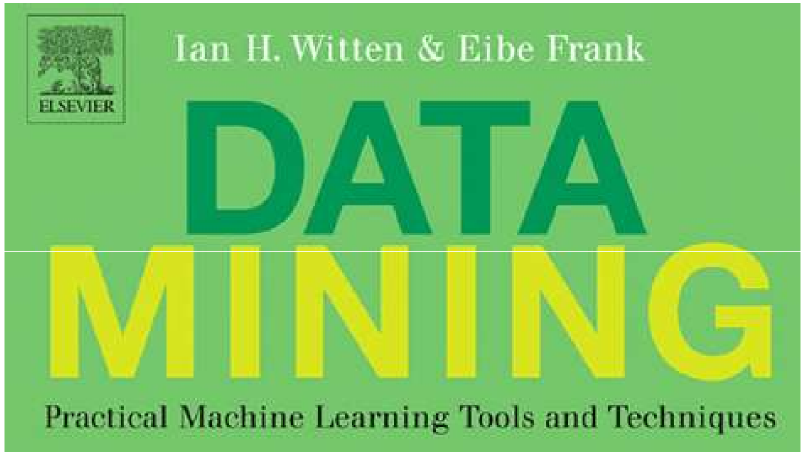
[a] The University of Michigan, Department of Internal Medicine, Division of Metabolism, Endocrinology and Diabetes, United States

[b] Wayne State University School of Medicine, United States

What is „the“ random choice?

	prediction			
experiment	88	10	2	100
	14	40	6	60
	18	10	12	40
	120	60	20	

Great introduction to kappa statistics



ELSEVIER

Ian H. Witten & Eibe Frank

DATA MINING

Practical Machine Learning Tools and Techniques

What is „the“ random choice?

	prediction			
experiment	88	10	2	100
	14	40	6	60
	18	10	12	40
	120	60	20	

	prediction			
experiment				100
				60
				40
	120	60	20	

What is „the“ random choice?

	prediction			
experiment	88	10	2	100
	14	40	6	60
	18	10	12	40
	120	60	20	

	prediction			
experiment	60	30	10	100
				60
				40
	120	60	20	

What is „the“ random choice?

	prediction			
experiment	88	10	2	100
	14	40	6	60
	18	10	12	40
	120	60	20	

	prediction			
experiment	60	30	10	100
	36	18	6	60
				40
	120	60	20	

What is „the“ random choice?

	prediction			
experiment	88	10	2	100
	14	40	6	60
	18	10	12	40
	120	60	20	

	prediction			
experiment	60	30	10	100
	36	18	6	60
	24	12	4	40
	120	60	20	

What is „the“ random choice?

		prediction		
experiment	88	10	2	100
	14	40	6	60
	18	10	12	40
	120	60	20	

		prediction		
experiment	60	30	10	h_{1x}
	36	18	6	h_{2x}
	24	12	4	h_{3x}
	h_{x1}	h_{x2}	h_{x3}	

$$\kappa = \frac{accuracy - baseline}{1 - baseline}$$

$$baseline = \sum_{i=1}^{k} \frac{h_{ix} \cdot h_{xi}}{N^2}$$

What is „the“ random choice?

experiment	prediction			
	88	10	2	100
	14	40	6	60
	18	10	12	40
	120	60	20	

κ=0.49±0.0.5

experiment	prediction			
	60	30	10	h_{1x}
	36	18	6	h_{2x}
	24	12	4	h_{3x}
	h_{x1}	h_{x2}	h_{x3}	

κ=0.0±0.05

$$\kappa = \frac{accuracy - baseline}{1 - baseline}$$

$$baseline = \sum_{i=1}^{k} \frac{h_{ix} \cdot h_{xi}}{N^2}$$

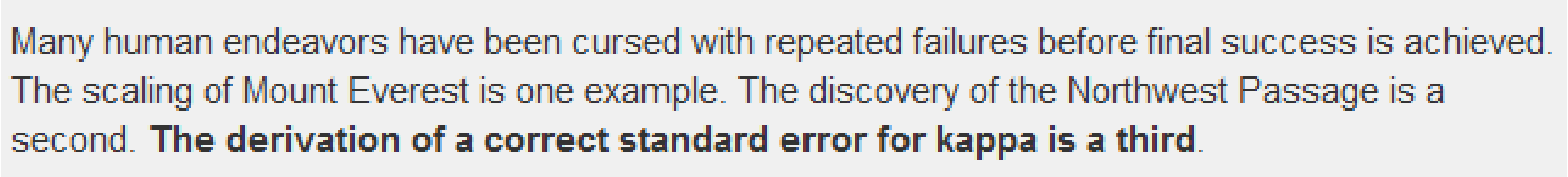
Error bars included!

$$var(\hat{\kappa}) = \frac{1}{n}\{\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2-\theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4-4\theta_2^2)}{(1-\theta_2)^4}$$

in which

$$\theta_1 = \frac{1}{n}\sum_{i=1}^{k} n_{ii} \qquad \theta_3 = \frac{1}{n^2}\sum_{i=1}^{k} n_{ii}(n_{i+}+n_{+i})$$

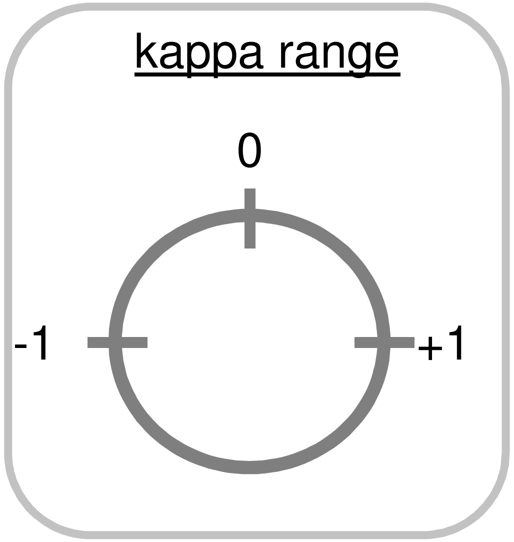
$$\theta_2 = \frac{1}{n^2}\sum_{i=1}^{k} n_{i+}n_{+i} \qquad \theta_4 = \frac{1}{n^3}\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}(n_{j+}+n_{+i})^2$$



Many human endeavors have been cursed with repeated failures before final success is achieved. The scaling of Mount Everest is one example. The discovery of the Northwest Passage is a second. **The derivation of a correct standard error for kappa is a third**.

Fleiss, J.L., Cohen, J., Everitt, B.S. (1969) Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* ***72(5), 323-327.***

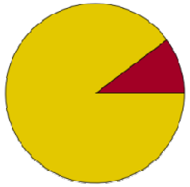
What is a good kappa value?



< 0.2	poor
0.21– 0.4	fair
0.41 – 0.60	moderate
0.61 – 0.80	good
> 0.81	very good

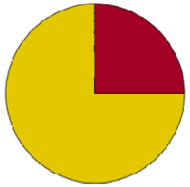
Landis, J.R. and Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics* ***33, 159-74.***

κ: If I take the majority vote…



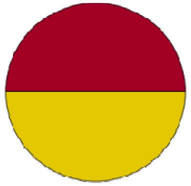
	pred	
exp	0	71
	0	623

kappa	0.0 ± 0.0
precision	0.90
recall	1.0
accuracy	0.90



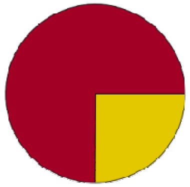
	pred	
exp	0	174
	0	520

kappa	0.0 ± 0.0
precision	0.75
recall	1
accuracy	0.75



	pred	
exp	347	0
	347	0

kappa	0.0 ± 0.0
precision	0.0
recall	0.0
accuracy	0.5



	pred	
exp	520	0
	174	0

kappa	0.0 ± 0.0
precision	0.0
recall	0.0
accuracy	0.75

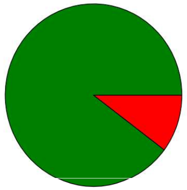
You can‘t fool κ!

How does kappa perform for the hERG models?

accuracy	precision	recall
0.90	1.00	0.03

κ=0.074±0.0.76

Functional assay
threshold = 1 µM



„experimental situation“

exemplary confusion matrix

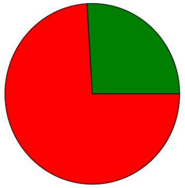
experiment		inactive	active
	inactive	247	0
	active	29	1

prediction

accuracy	precision	recall
0.79	0.80	0.96

κ=0.262±0.027

Binding assay
threshold = 10 µM



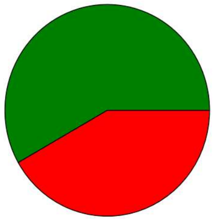
„experimental situation“

exemplary confusion matrix

experiment		inactive	active
	inactive	97	276
	active	44	1072

prediction

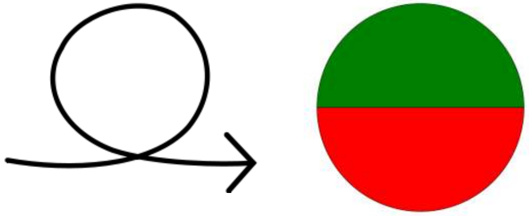
κ: influence of balancing



Functional assay threshold = 10 µM

		prediction	
experiment	inactive	134	35
	active	52	56
		inactive	active

κ=0.321±0.059

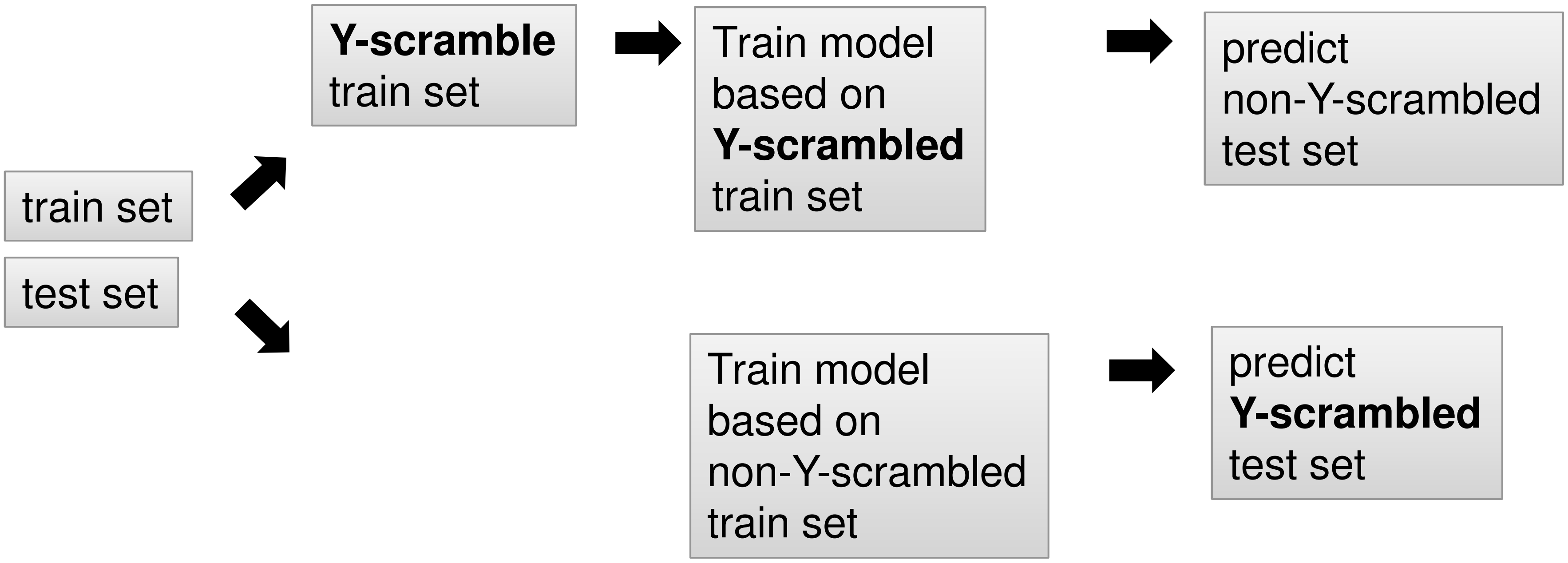


		prediction	
experiment	inactive	84	26
	active	56	66
		inactive	active

κ=0.301±0.060

- **Balancing the data set only has minor influence on κ.**
- **However, for largely imbalanced data sets, there is a stronger influence on κ.**
- **CAVE: Don‘t forget the remainder when balancing!**

κ: no signal on Y-scrambled data!



train set

test set

Y-scramble train set

Train model based on **Y-scrambled** train set

predict non-Y-scrambled test set

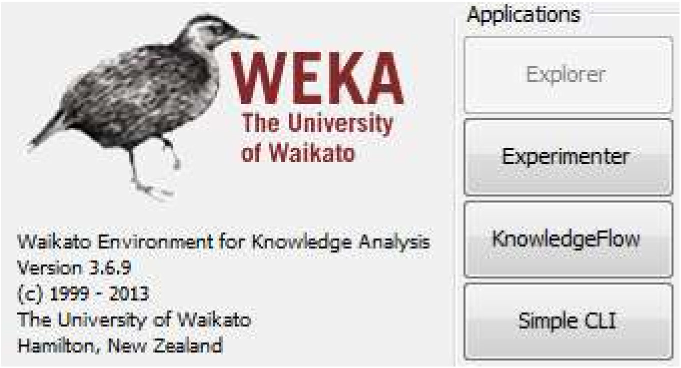
Train model based on non-Y-scrambled train set

predict **Y-scrambled** test set

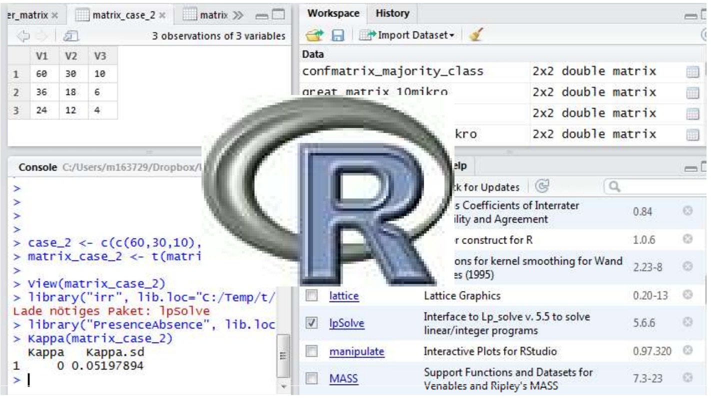
κ is ≈ 0.01

Other figures of merit show a signal!

κ: availability



Waikato Environment for Knowledge Analysis
Version 3.6.9
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand



	V1	V2	V3
1	60	30	10
2	36	18	6
3	24	12	4

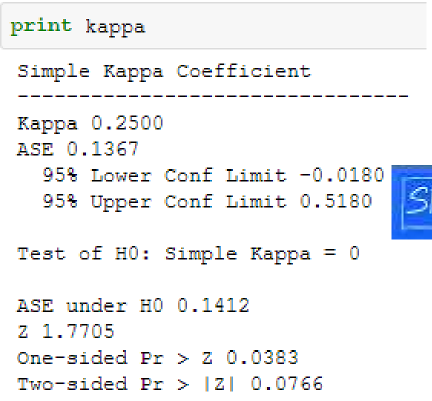
confmatrix_majority_class	2x2 double matrix
great_matrix_10mikro	2x2 double matrix
	2x2 double matrix
kro	2x2 double matrix

```
>
>
>
>
> case_2 <- c(c(60,30,10),
> matrix_case_2 <- t(matri
>
> View(matrix_case_2)
> library("irr", lib.loc="C:/Temp/t/
Lade nötiges Paket: lpSolve
> library("PresenceAbsence", lib.loc
> Kappa(matrix_case_2)
  Kappa    Kappa.sd
1     0 0.05197894
> |
```

		s Coefficients of Interrater ility and Agreement	0.84	
		r construct for R	1.0.6	
		ons for kernel smoothing for Wand es (1995)	2.23-8	
☐	lattice	Lattice Graphics	0.20-13	
☑	lpSolve	Interface to Lp_solve v. 5.5 to solve linear/integer programs	5.6.6	
☐	manipulate	Interactive Plots for RStudio	0.97.320	
☐	MASS	Support Functions and Datasets for Venables and Ripley's MASS	7.3-23	

… but without the error bars!

`irr` package
`PresenceAbsence` package

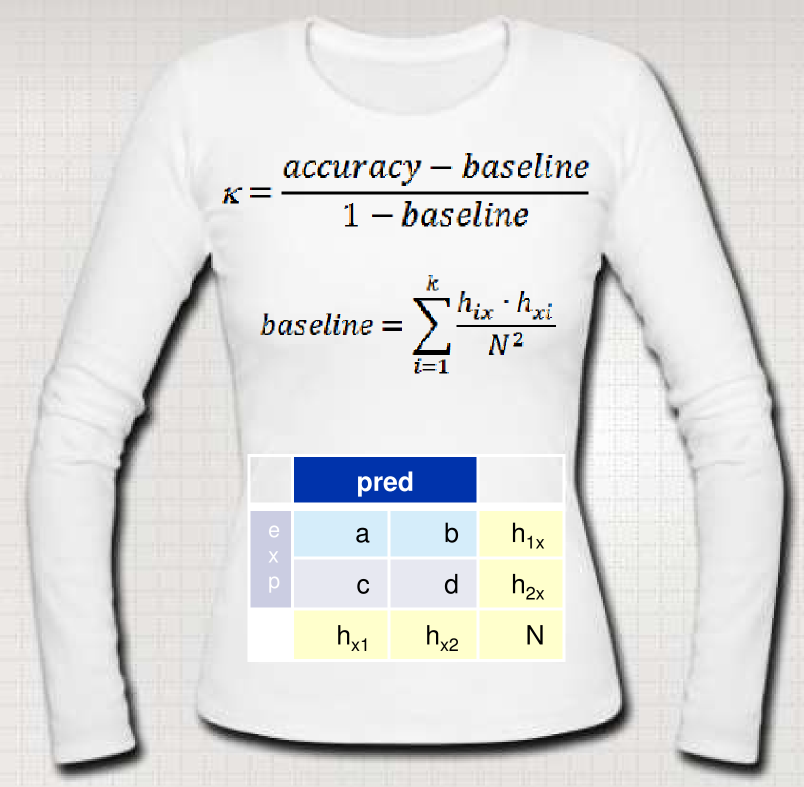


```
print kappa

Simple Kappa Coefficient
--------------------------------
Kappa 0.2500
ASE 0.1367
  95% Lower Conf Limit -0.0180
  95% Upper Conf Limit 0.5180

Test of H0: Simple Kappa = 0

ASE under H0 0.1412
Z 1.7705
One-sided Pr > Z 0.0383
Two-sided Pr > |Z| 0.0766
```

SM statsmodels

κ-onclusions

The principles fit on a sweater



$$\kappa = \frac{accuracy - baseline}{1 - baseline}$$

$$baseline = \sum_{i=1}^{k} \frac{h_{ix} \cdot h_{xi}}{N^2}$$

		pred		
exp		a	b	h_{1x}
		c	d	h_{2x}
		h_{x1}	h_{x2}	N

ac-κ-nowledgment

Anthony Nicholls

Christian Kramer

Greg Landrum

Kim Branson

Anja von Heydebreck

Daniel Kuhn

Friedrich Rippmann

Gerhard Barnickel

Martin Held