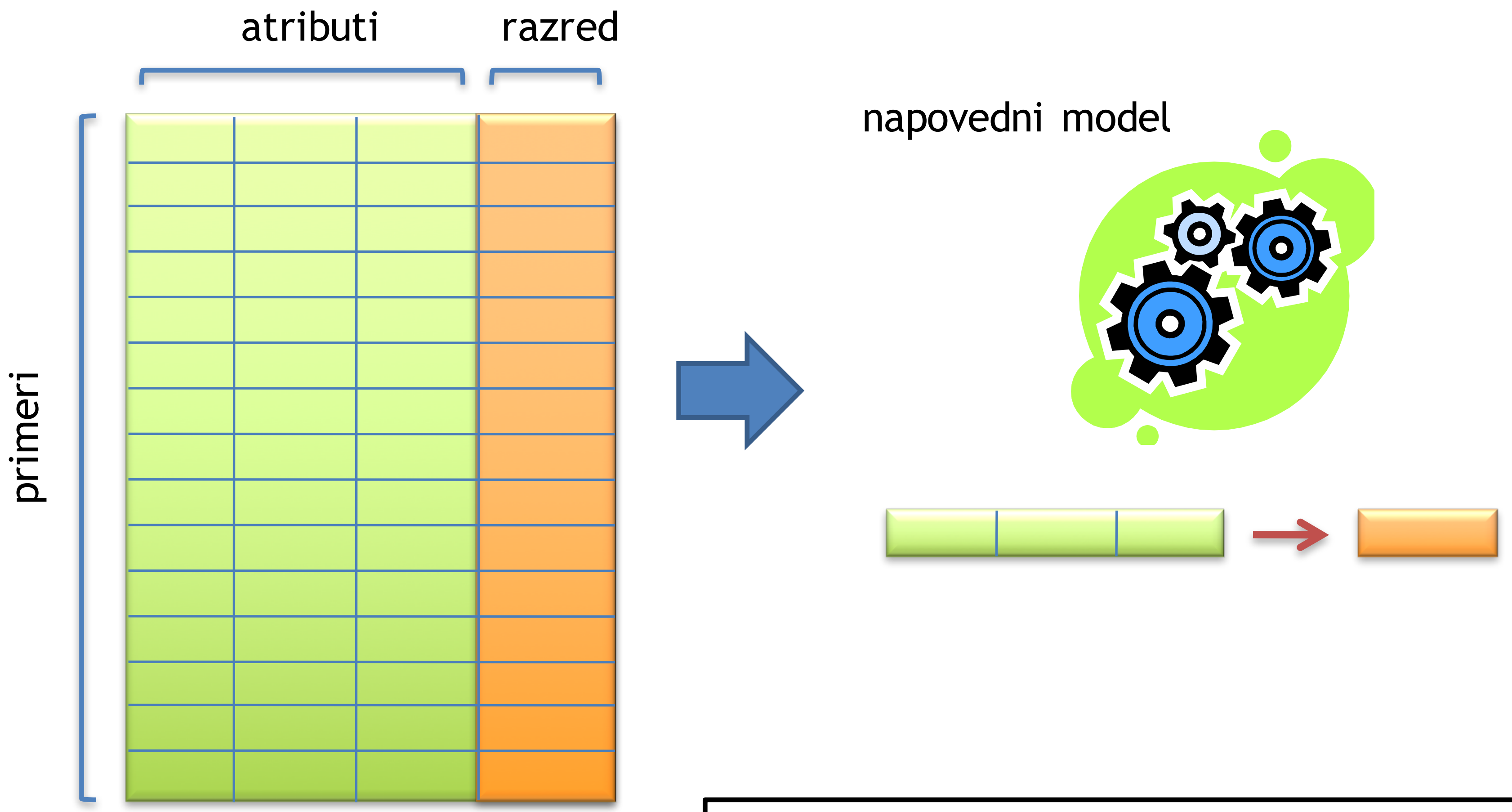
Prekletstvo dimenzionalnosti (curse of dimensionality)

Povzeto po predavanju prof. dr. Blaža Zupana.

Napovedno podatkovno rudarjenje (predictive data mining)

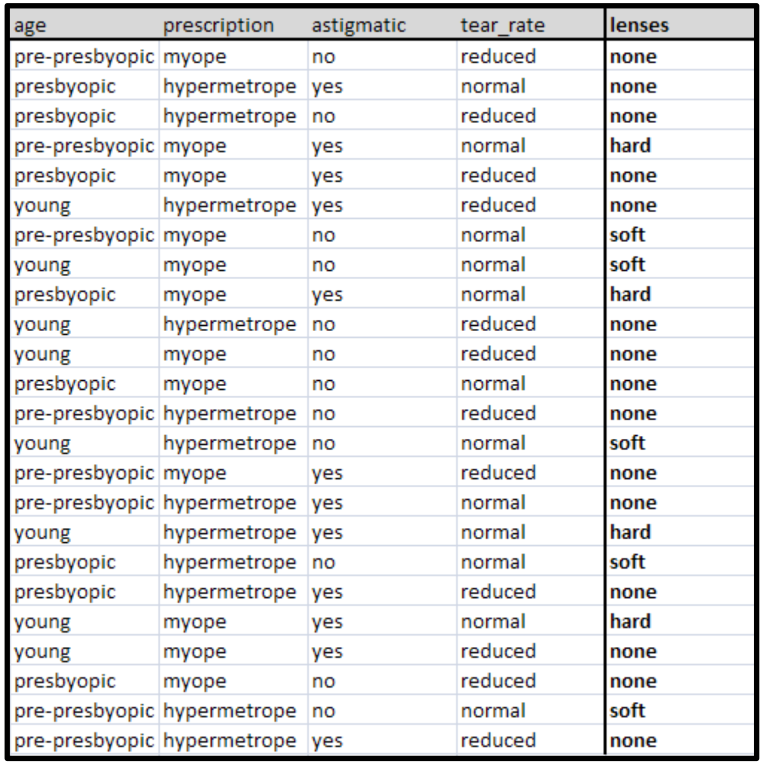


1. razumevanje (kompleksnega) sistema
2. točne napovedi

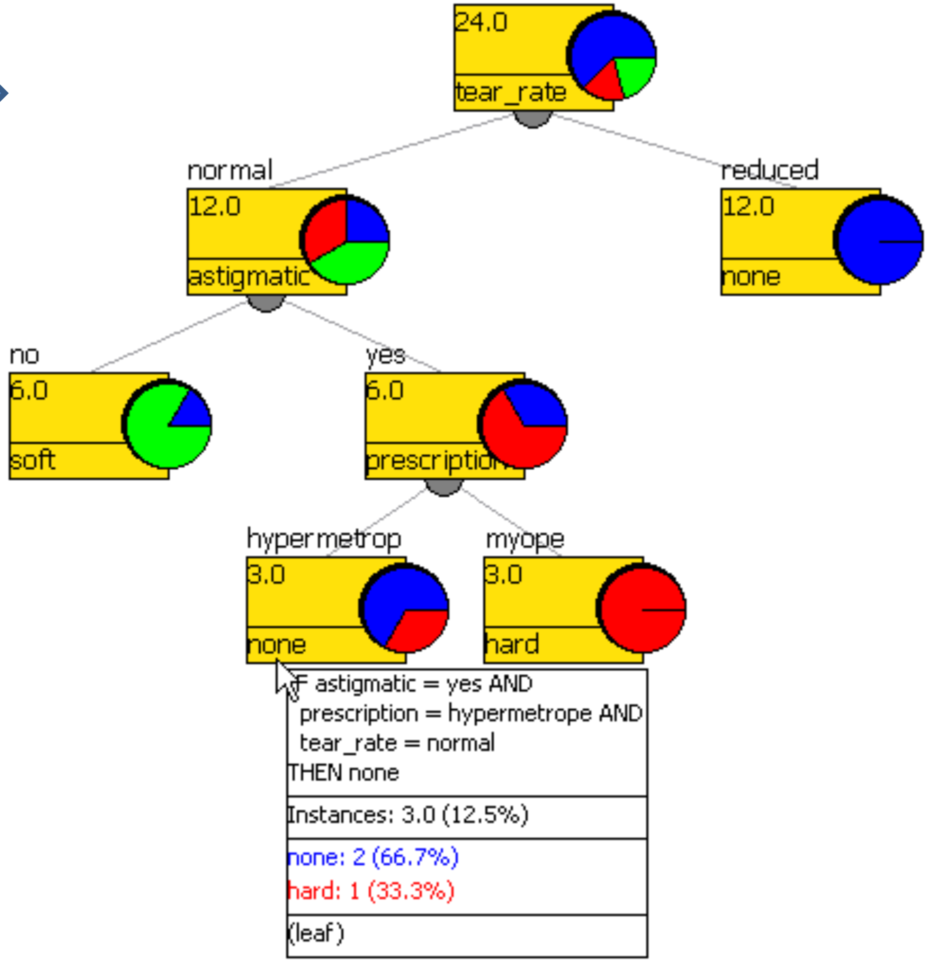
Lenses

4 atributi

24 primerov



age	prescription	astigmatic	tear_rate	lenses
pre-presbyopic	myope	no	reduced	**none**
presbyopic	hypermetrope	yes	normal	**none**
presbyopic	hypermetrope	no	reduced	**none**
pre-presbyopic	myope	yes	normal	**hard**
presbyopic	myope	yes	reduced	**none**
young	hypermetrope	yes	reduced	**none**
pre-presbyopic	myope	no	normal	**soft**
young	myope	no	normal	**soft**
presbyopic	myope	yes	normal	**hard**
young	hypermetrope	no	reduced	**none**
young	myope	no	reduced	**none**
presbyopic	myope	no	normal	**none**
pre-presbyopic	hypermetrope	no	reduced	**none**
young	hypermetrope	no	normal	**soft**
pre-presbyopic	myope	yes	reduced	**none**
pre-presbyopic	hypermetrope	yes	normal	**none**
young	hypermetrope	yes	normal	**hard**
presbyopic	hypermetrope	no	normal	**soft**
presbyopic	hypermetrope	yes	reduced	**none**
young	myope	yes	normal	**hard**
young	myope	yes	reduced	**none**
presbyopic	myope	no	reduced	**none**
pre-presbyopic	hypermetrope	no	normal	**soft**
pre-presbyopic	hypermetrope	yes	reduced	**none**



IF astigmatic = yes AND
prescription = hypermetrope AND
tear_rate = normal
THEN none

Instances: 3.0 (12.5%)

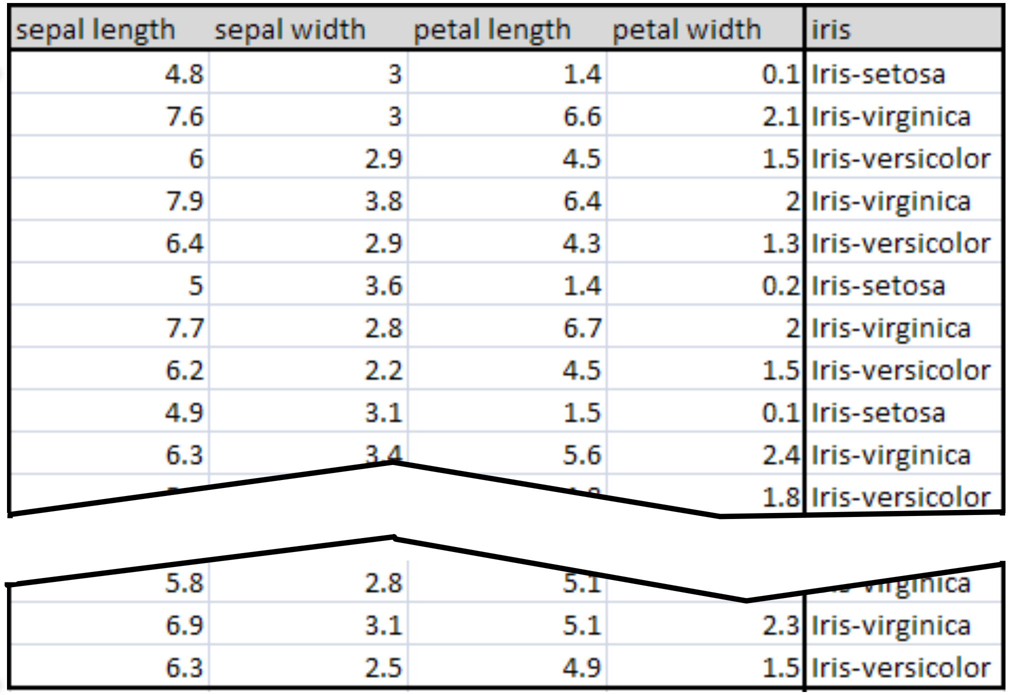
none: 2 (66.7%)
hard: 1 (33.3%)

(leaf)

Iris

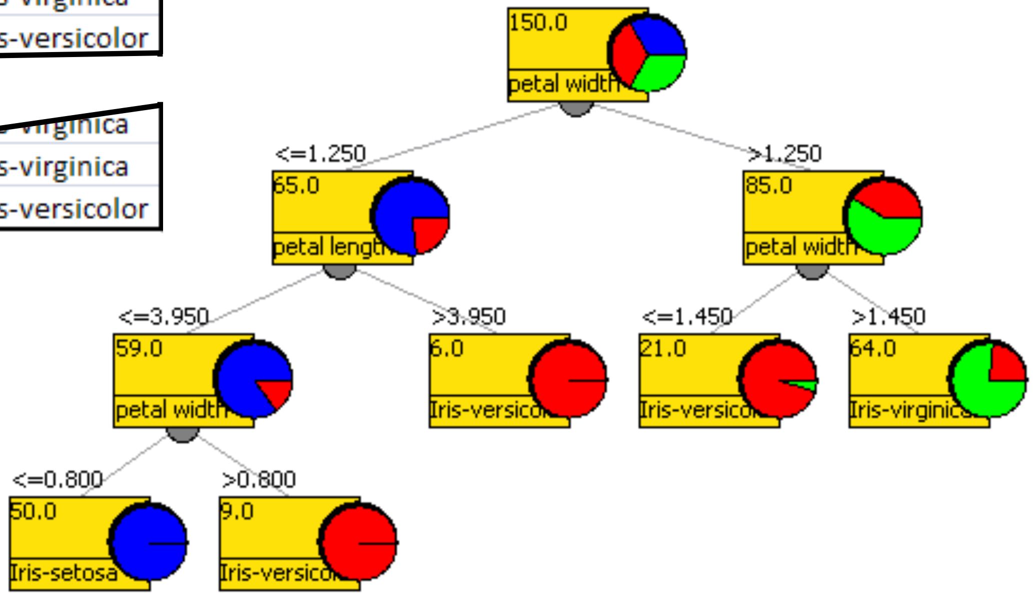
4 atributi

150 primerov

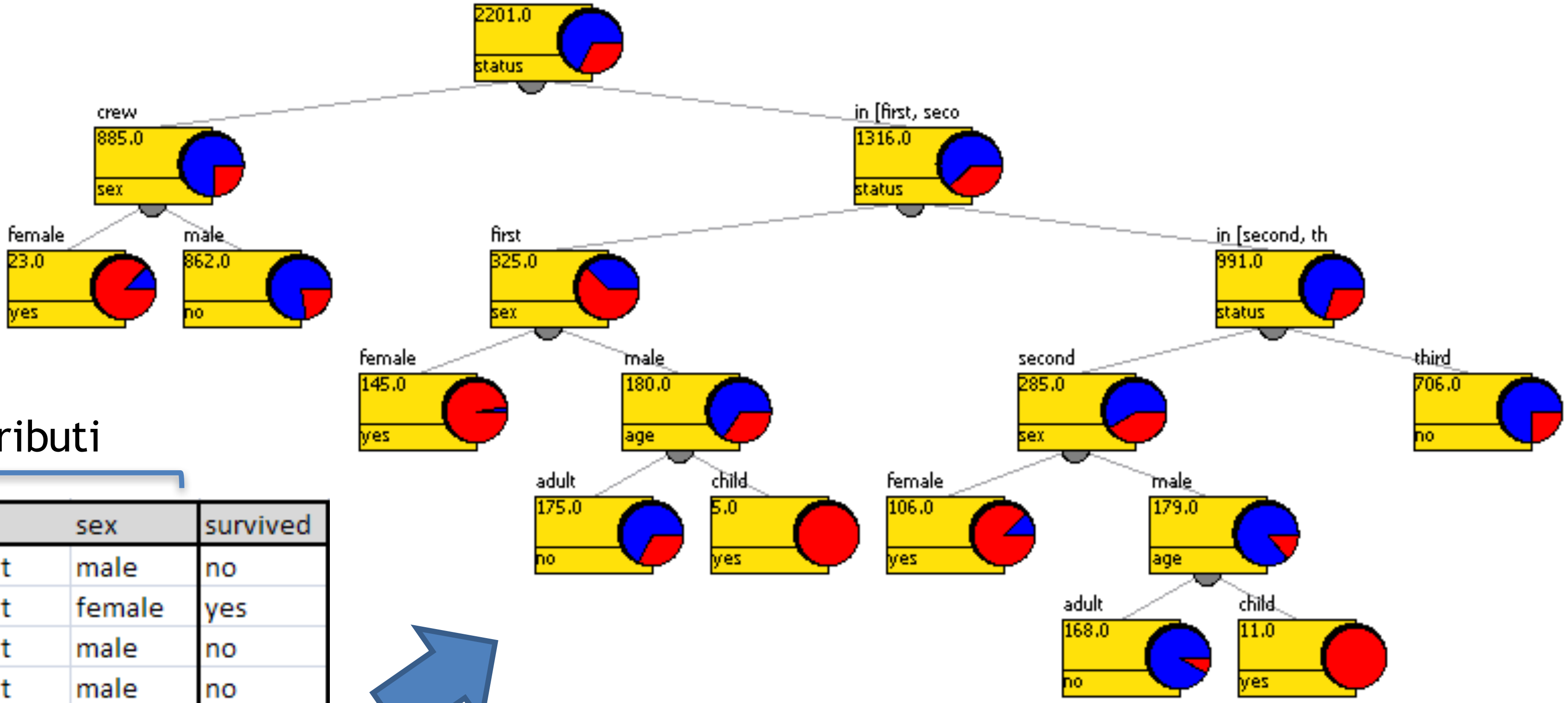


sepal length	sepal width	petal length	petal width	iris
4.8	3	1.4	0.1	Iris-setosa
7.6	3	6.6	2.1	Iris-virginica
6	2.9	4.5	1.5	Iris-versicolor
7.9	3.8	6.4	2	Iris-virginica
6.4	2.9	4.3	1.3	Iris-versicolor
5	3.6	1.4	0.2	Iris-setosa
7.7	2.8	6.7	2	Iris-virginica
6.2	2.2	4.5	1.5	Iris-versicolor
4.9	3.1	1.5	0.1	Iris-setosa
6.3	3.4	5.6	2.4	Iris-virginica
			1.8	Iris-versicolor
5.8	2.8	5.1		virginica
6.9	3.1	5.1	2.3	Iris-virginica
6.3	2.5	4.9	1.5	Iris-versicolor

Fischer RA (1936) Annual Eugenics, 7: 179-188.

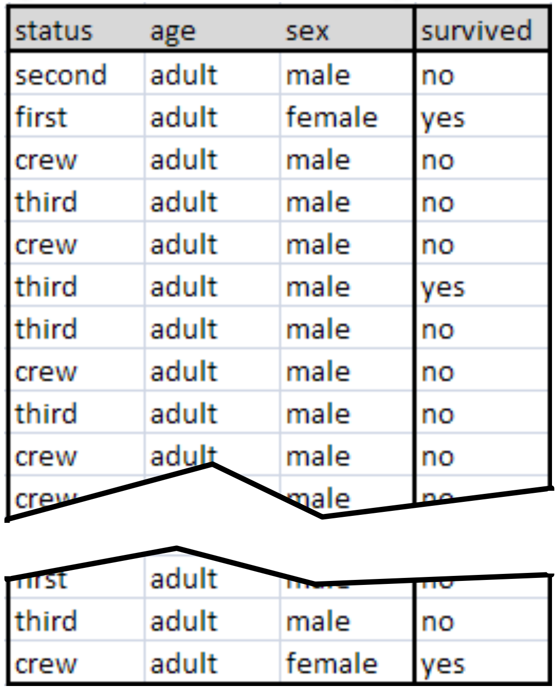


Titanic



3 atributi

2201 primerov



status	age	sex	survived
second	adult	male	no
first	adult	female	yes
crew	adult	male	no
third	adult	male	no
crew	adult	male	no
third	adult	male	yes
third	adult	male	no
crew	adult	male	no
third	adult	male	no
crew	adult	male	no
crew		male	no
first	adult	male	no
third	adult	male	no
crew	adult	female	yes

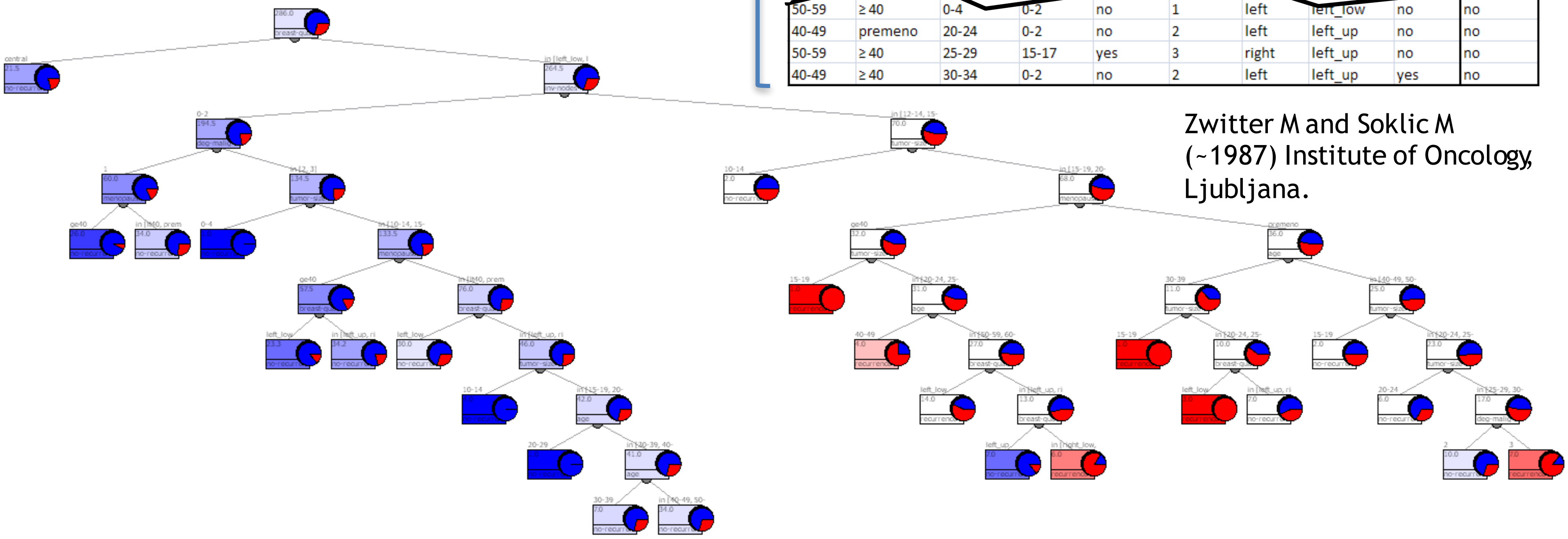
Breast cancer recurrence

9 atributov

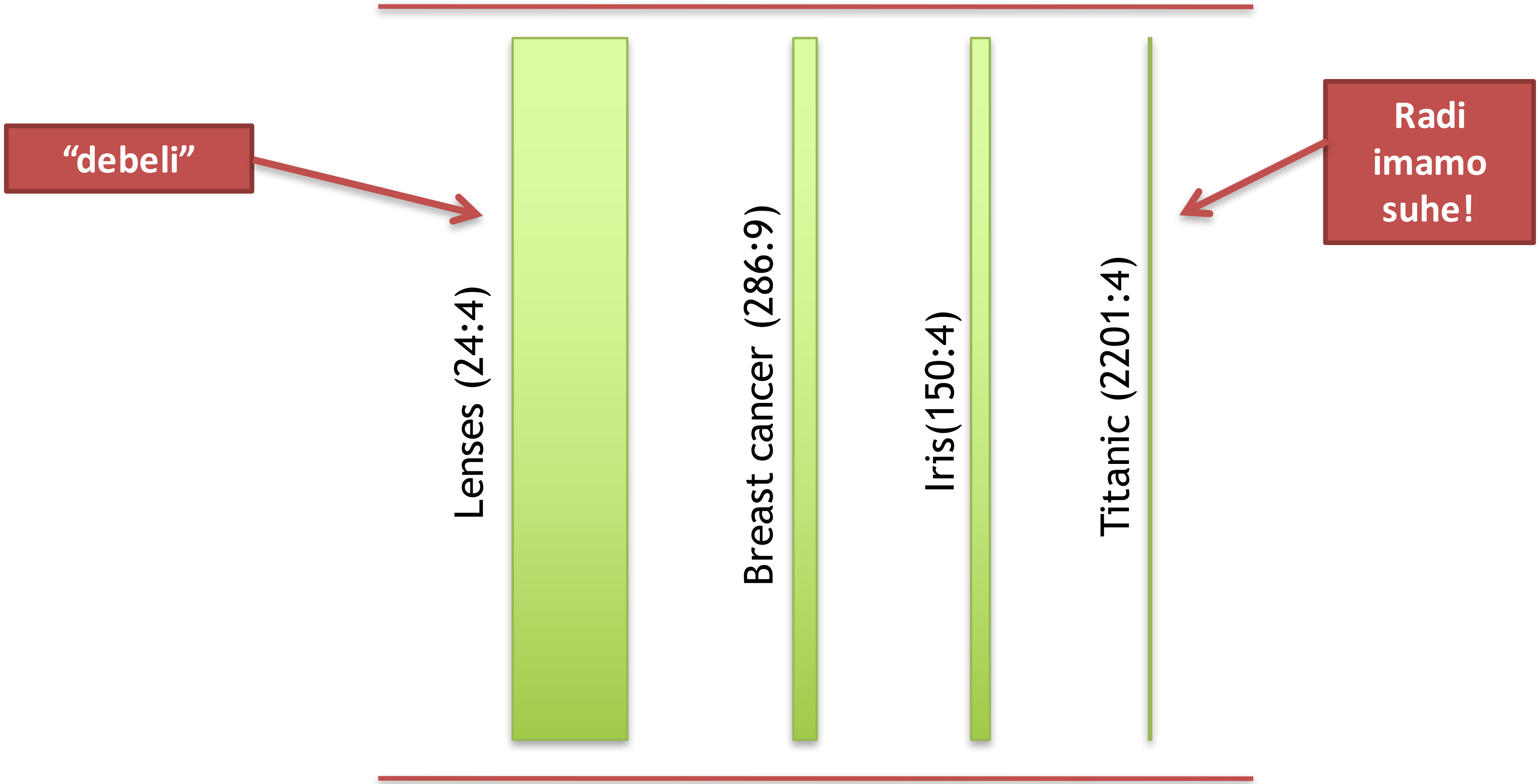
286 primerov

age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat	recurrence
40-49	premeno	25-29	0-2	no	2	right	left_low	no	yes
50-59	≥ 40	15-19	0-2	no	1	right	central	no	no
40-49	premeno	15-19	0-2	no	2	left	left_up	no	yes
40-49	≥ 40	40-44	15-17	yes	2	right	left_up	yes	no
30-39	premeno	40-44	0-2	no	2	right	right_up	no	no
40-49	premeno	30-34	15-17	yes	3	left	left_low	no	yes
40-49	premeno	10-14	0-2	no	1	right	left_up	no	no
40-49	premeno	35-39	9-11	yes	2	right	right_up	yes	no
50-59	premeno	25-29	0-2	no	2	left	right_up	no	yes
50-59	premeno	10-14	0-2	no	1	left	left_low	no	no
50-59	prem	50-54	9-11	yes		ht	left_up	no	yes
60-6		4	0-2				left_low	yes	no
	≥ 40				2	rig			
50-59	≥ 40	0-4	0-2	no	1	left	left_low	no	no
40-49	premeno	20-24	0-2	no	2	left	left_up	no	no
50-59	≥ 40	25-29	15-17	yes	3	right	left_up	no	no
40-49	≥ 40	30-34	0-2	no	2	left	left_up	yes	no

Zwitter M and Soklic M (~1987) Institute of Oncology, Ljubljana.

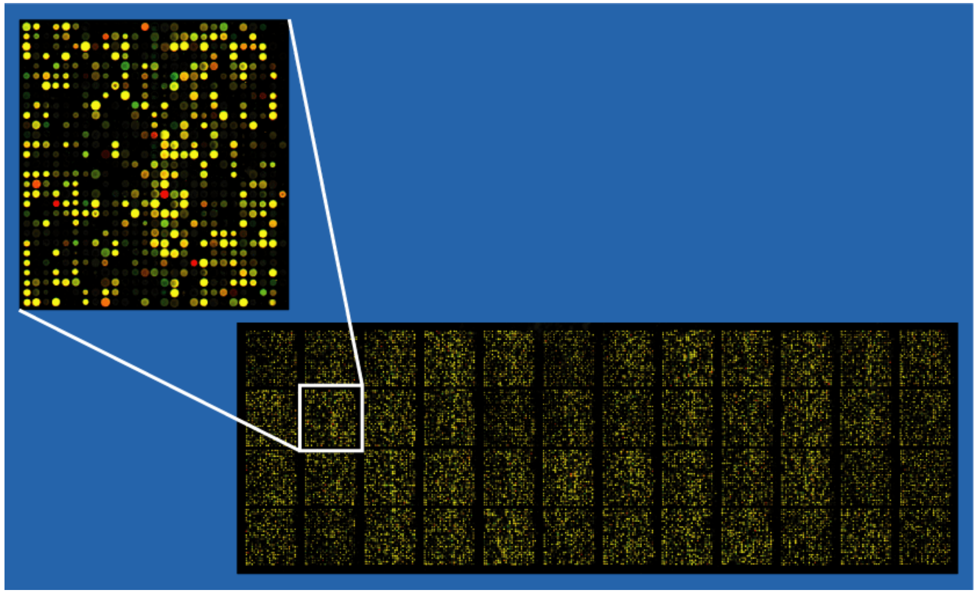


“Oblike” podatkov



Biološki podatki

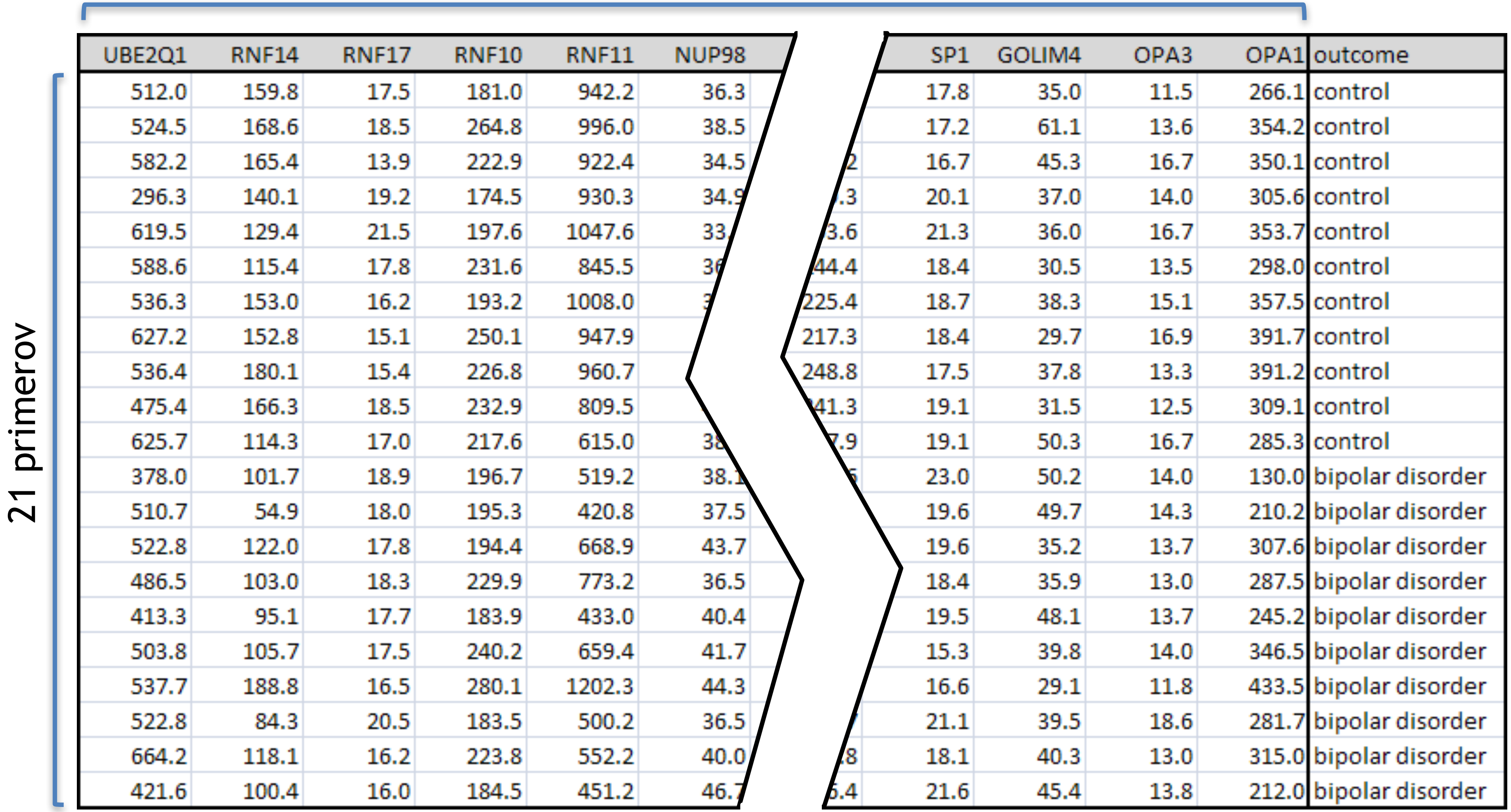
- več tisoč **<u>hkratnih meritev</u>**
- visoko-propustno (high throughput)
- različne tehnologije,
- primer, mikromreže DNA:



Primer: GEO GDS2191

14,903 atributov (genov)

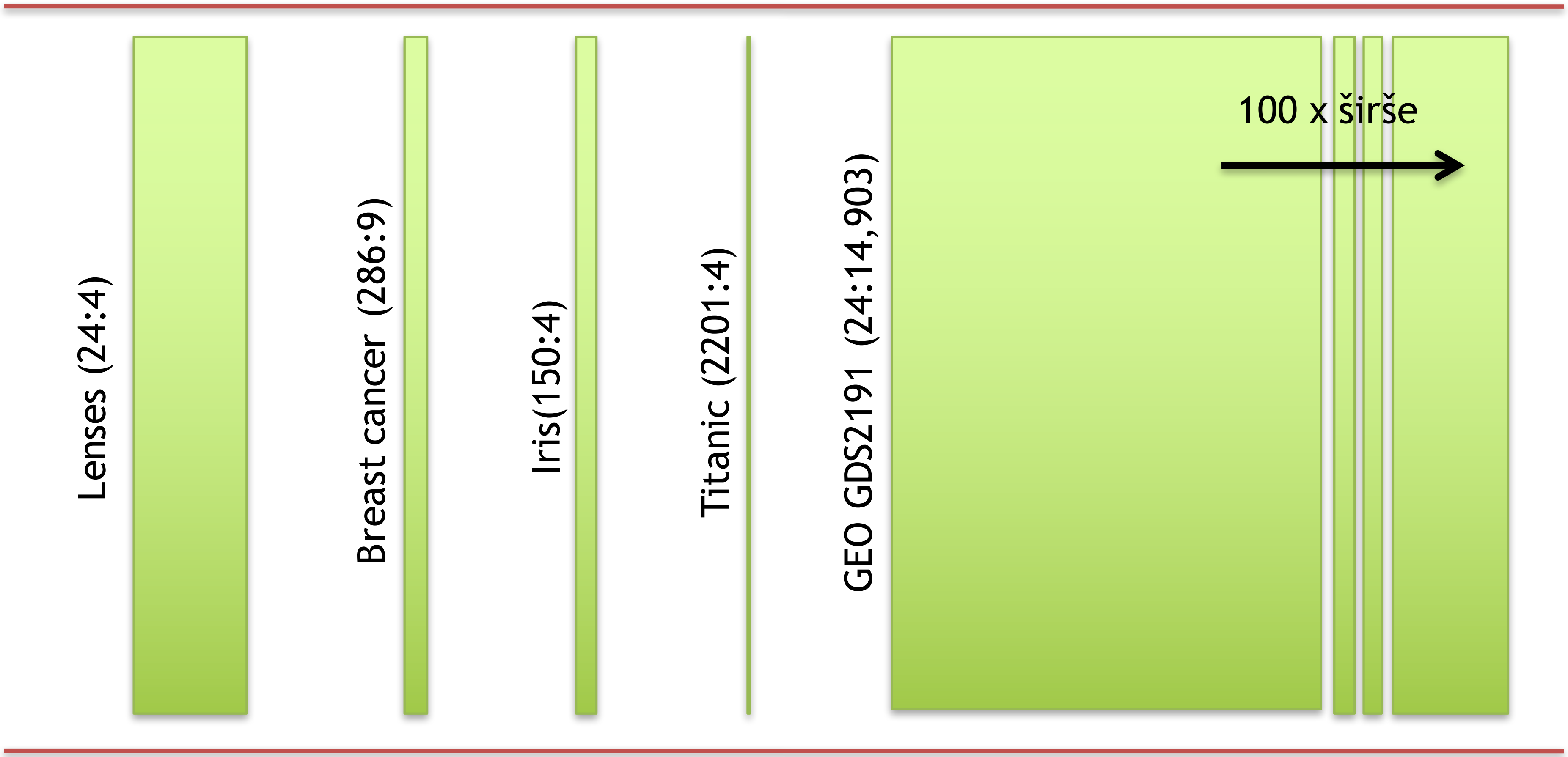
21 primerov



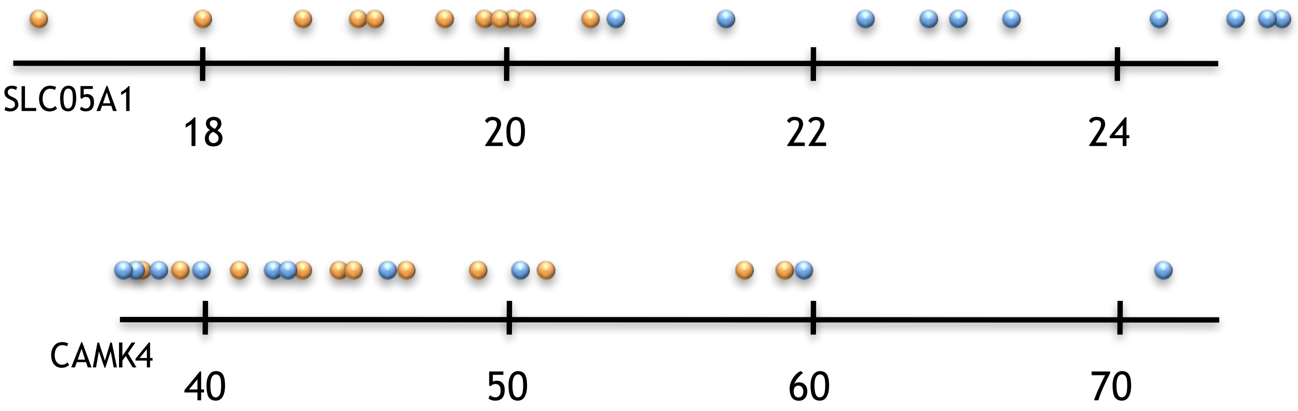
UBE2Q1	RNF14	RNF17	RNF10	RNF11	NUP98		SP1	GOLIM4	OPA3	OPA1	outcome
512.0	159.8	17.5	181.0	942.2	36.3		17.8	35.0	11.5	266.1	control
524.5	168.6	18.5	264.8	996.0	38.5		17.2	61.1	13.6	354.2	control
582.2	165.4	13.9	222.9	922.4	34.5	2	16.7	45.3	16.7	350.1	control
296.3	140.1	19.2	174.5	930.3	34.9	.3	20.1	37.0	14.0	305.6	control
619.5	129.4	21.5	197.6	1047.6	33.	3.6	21.3	36.0	16.7	353.7	control
588.6	115.4	17.8	231.6	845.5	36	44.4	18.4	30.5	13.5	298.0	control
536.3	153.0	16.2	193.2	1008.0	3	225.4	18.7	38.3	15.1	357.5	control
627.2	152.8	15.1	250.1	947.9		217.3	18.4	29.7	16.9	391.7	control
536.4	180.1	15.4	226.8	960.7		248.8	17.5	37.8	13.3	391.2	control
475.4	166.3	18.5	232.9	809.5		41.3	19.1	31.5	12.5	309.1	control
625.7	114.3	17.0	217.6	615.0	38	7.9	19.1	50.3	16.7	285.3	control
378.0	101.7	18.9	196.7	519.2	38.1	5	23.0	50.2	14.0	130.0	bipolar disorder
510.7	54.9	18.0	195.3	420.8	37.5		19.6	49.7	14.3	210.2	bipolar disorder
522.8	122.0	17.8	194.4	668.9	43.7		19.6	35.2	13.7	307.6	bipolar disorder
486.5	103.0	18.3	229.9	773.2	36.5		18.4	35.9	13.0	287.5	bipolar disorder
413.3	95.1	17.7	183.9	433.0	40.4		19.5	48.1	13.7	245.2	bipolar disorder
503.8	105.7	17.5	240.2	659.4	41.7		15.3	39.8	14.0	346.5	bipolar disorder
537.7	188.8	16.5	280.1	1202.3	44.3		16.6	29.1	11.8	433.5	bipolar disorder
522.8	84.3	20.5	183.5	500.2	36.5	7	21.1	39.5	18.6	281.7	bipolar disorder
664.2	118.1	16.2	223.8	552.2	40.0	8	18.1	40.3	13.0	315.0	bipolar disorder
421.6	100.4	16.0	184.5	451.2	46.7	5.4	21.6	45.4	13.8	212.0	bipolar disorder

Analysis of postmortem orbitofrontal cortex from 10 adults with bipolar disorder. Results provide insight into the pathophysiology of the disease.

"Oblike" podatkov



Ocenjevanje in izbira atributov



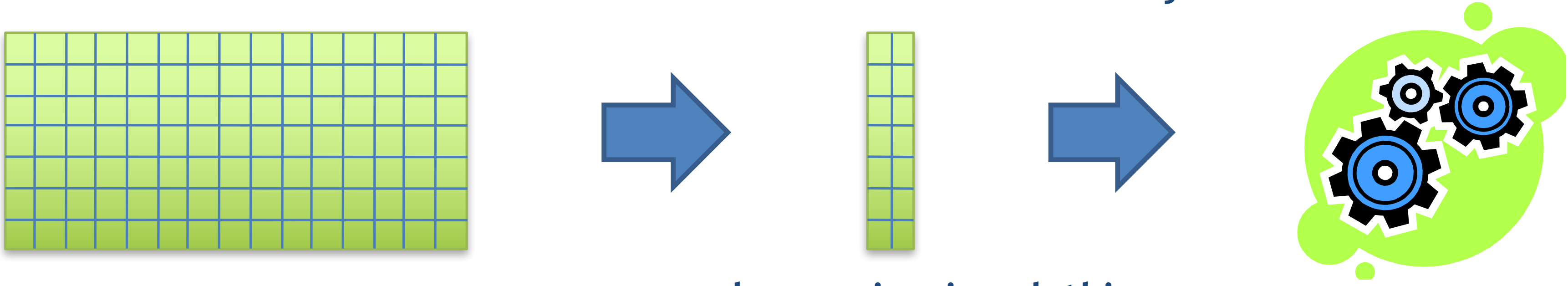
Signal-to-Noise Ratio
(Golub, Science 99)

$$S2N = \frac{|\ !(🔵) - !(🟠)\ |}{|\ ?(🔵) + ?(🟠)\ |}$$

SLCO5A1	CAMK4	outcome	
18.0	39.2	control	🟠
20.0	51.3	control	🟠
19.8	57.6	control	🟠
18.6	37.9	control	🟠
19.9	46.9	control	🟠
20.5	44.5	control	🟠
19.6	43.9	control	🟠
20.0	44.6	control	🟠
19.0	48.4	control	🟠
19.1	41.2	control	🟠
16.8	58.1	control	🟠
22.3	38.6	bipolar disorder	🔵
22.7	71.3	bipolar disorder	🔵
20.3	42.7	bipolar disorder	🔵
25.7	42.5	bipolar disorder	🔵
26.8	50.9	bipolar disorder	🔵
26.5	59.4	bipolar disorder	🔵
23.5	40.4	bipolar disorder	🔵
26.0	46.4	bipolar disorder	🔵
21.4	37.3	bipolar disorder	🔵
24.2	38.1	bipolar disorder	🔵
1.49	0.003		

Scenarij

pogost v prvih študijah, kjer so uporabljali mikromreže DNA



ocenjevanje, rangiranje in izbira atributov

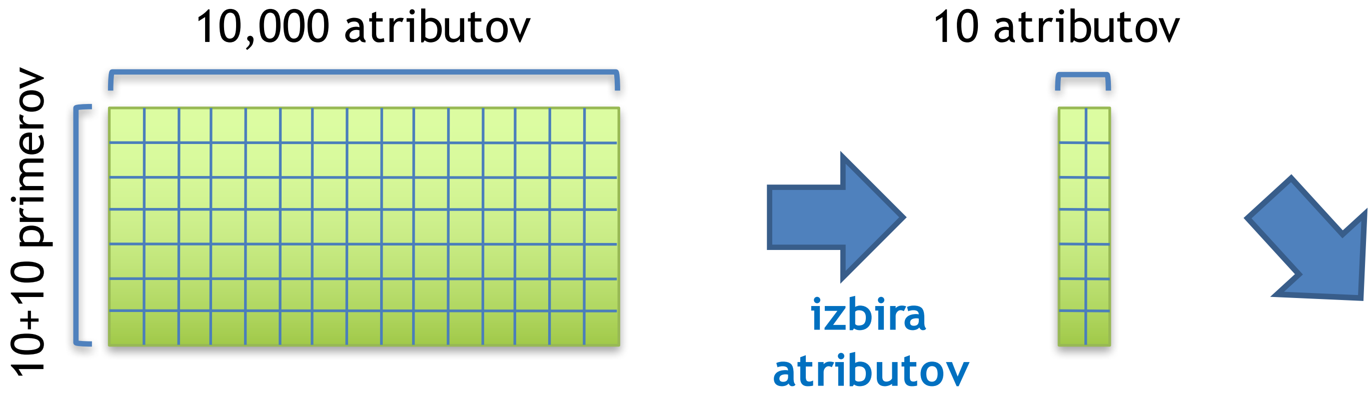
modeliranje

podatki

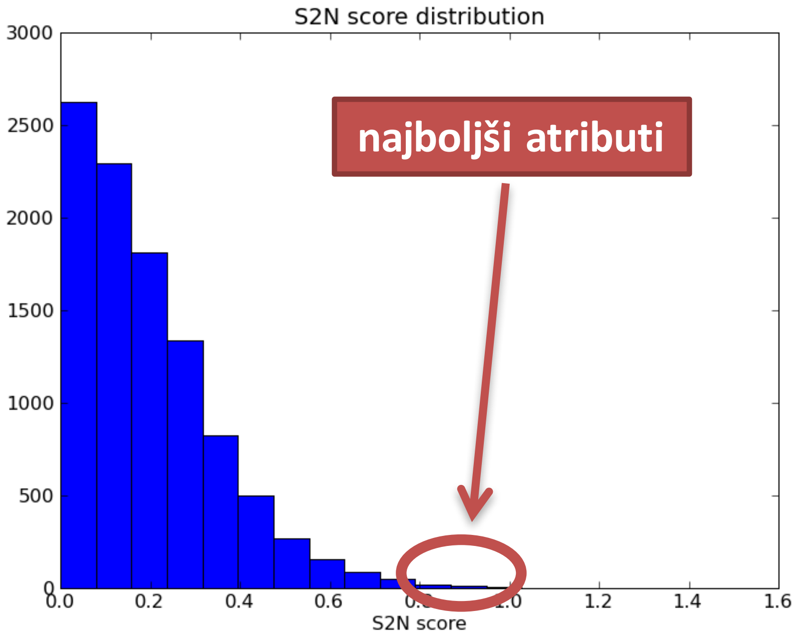
predprocesirani podatki, le izbrani atributi

model

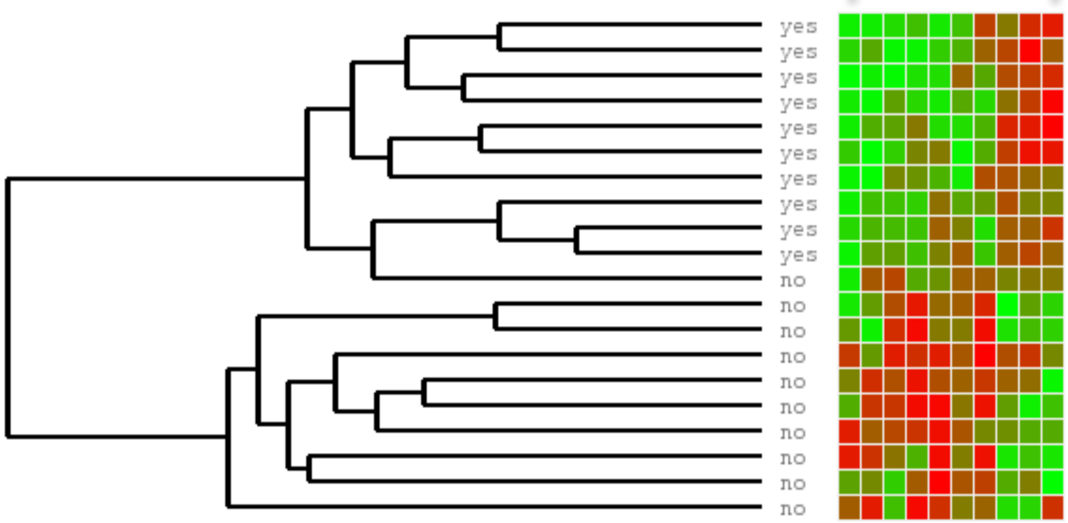
Primer analize



hierarhično razvrščanje



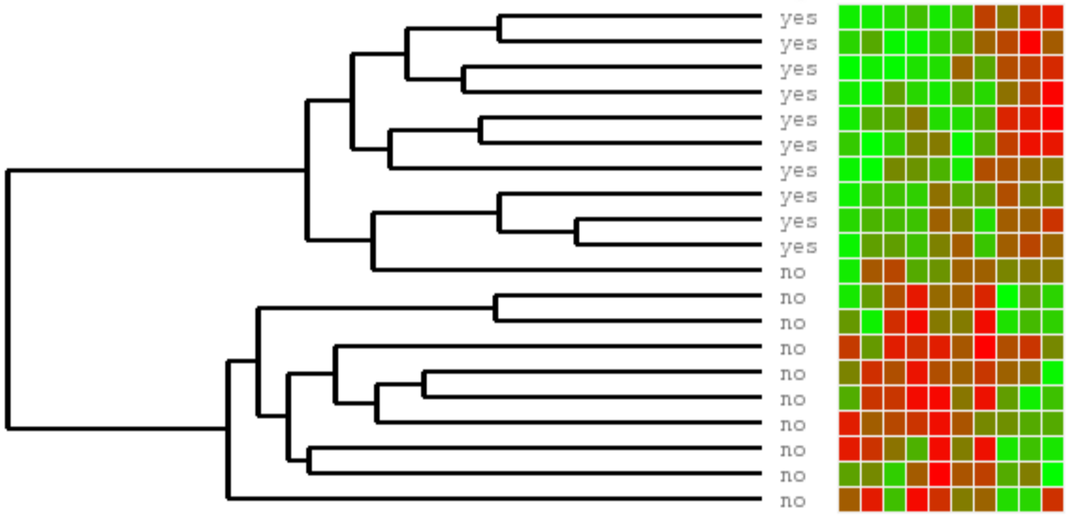
10 atributov



Leave-one-out s SVM-jem na predprocesirani množici: točnost = 100 %

Deluje perfektno!

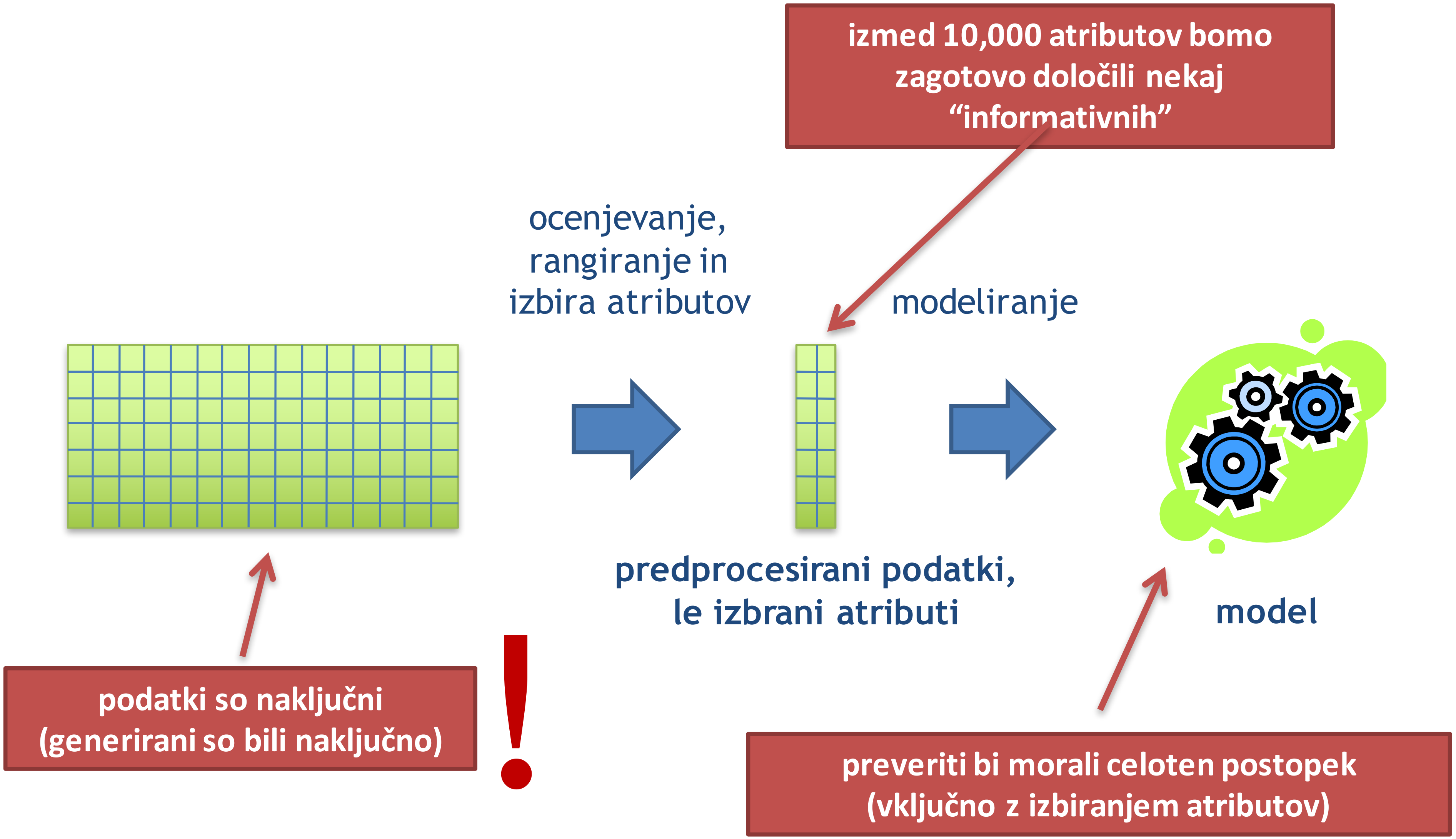
razredi lepo sovpadajo s skupinami



Leave-one-out s SVM-jem na predprocesirani množici: točnost = 100 %

boljše kot tako res ne gre!

Ni res!



izmed 10,000 atributov bomo zagotovo določili nekaj "informativnih"

ocenjevanje, rangiranje in izbira atributov

modeliranje

predprocesirani podatki, le izbrani atributi

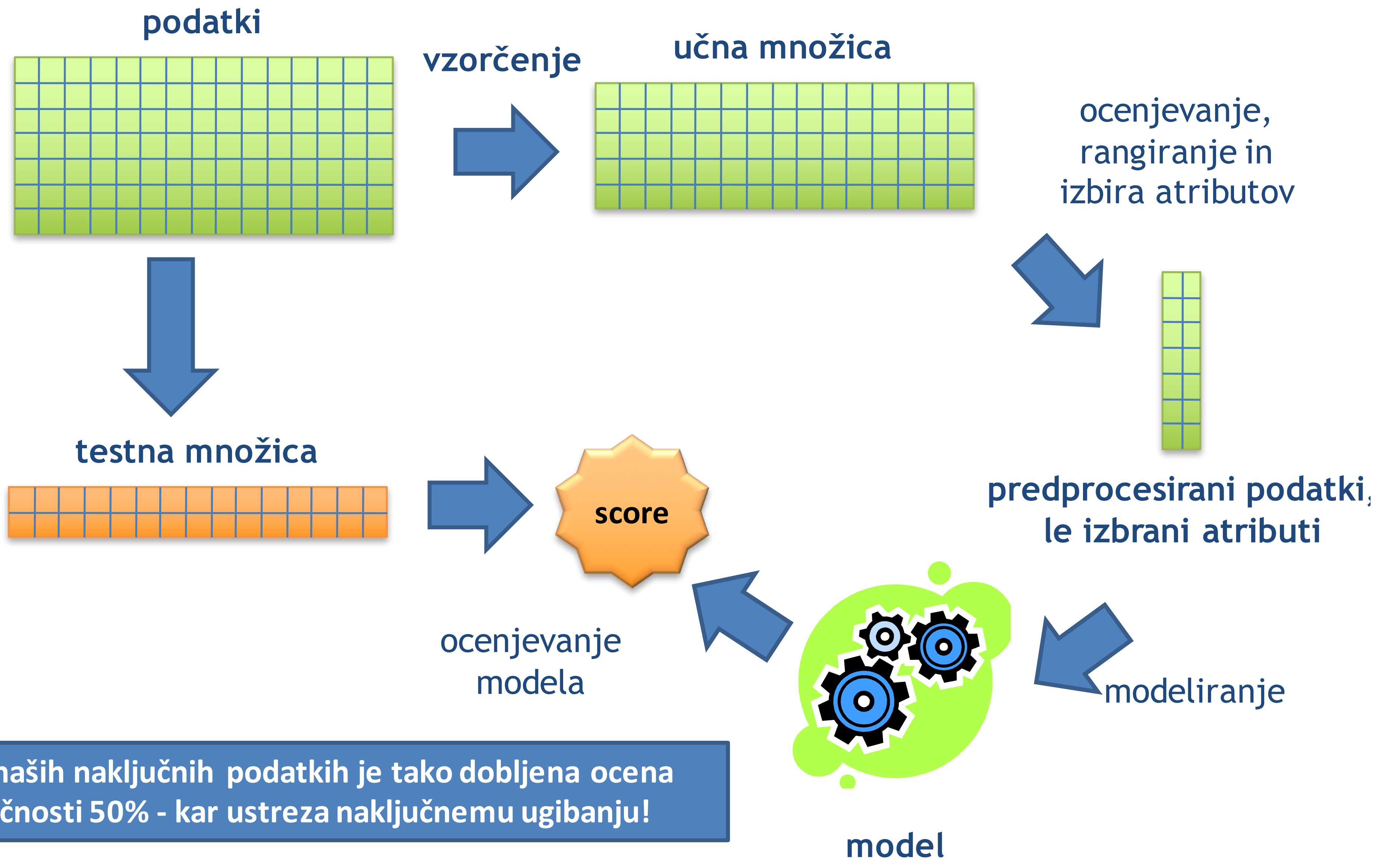
model

podatki so naključni (generirani so bili naključno)

!

preveriti bi morali celoten postopek (vključno z izbiranjem atributov)

Kvaliteta napovedi: pravilni način



Na naših naključnih podatkih je tako dobljena ocena točnosti 50% - kar ustreza naključnemu ugibanju!

Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification

Richard Simon, Michael D. Radmacher, Kevin Dobbin, Lisa M. McShane

Journal of the National Cancer Institute, Vol. 95, No. 1, January 1, 2003

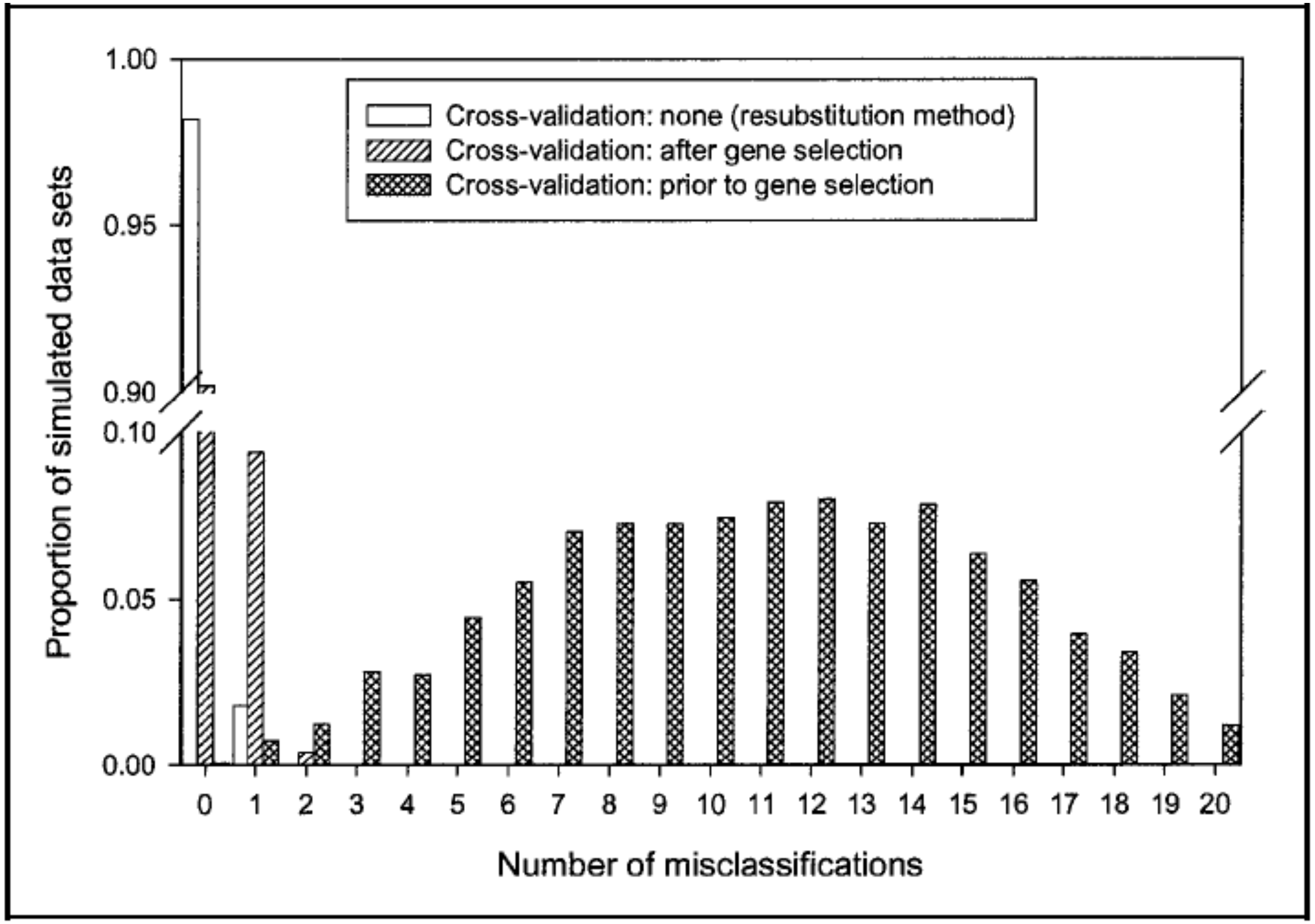


Fig. 1. Effect of various levels of cross-validation on the estimated error rate of a predictor derived from 2000 simulated datasets. Class labels were arbitrarily assigned to the specimens within each dataset, so poor classification accuracy is expected. Class prediction was performed on each dataset as described in the supplemental information (http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue1/index.shtml and http://linus.nci.nih.gov/~brb), varying the level of leave-one-out cross-validation used in the prediction. **Vertical bars** indicate the proportion of simulated datasets (of 2000) resulting in a given number of misclassifications for a specified cross-validation strategy.