

# Primenjeno Mašinsko Učenje

Predavač: Aleksandar Kovačević

Deo predavanja preuzet sa: <https://dlab.epfl.ch/teaching/fall2017/cs401/>

# Čemu ovo predavanje?

---

## Machine Learning that Matters

---

Kiri L. Wagstaff

Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109 USA

KIRI.L.WAGSTAFF@JPL.NASA.GOV

[link](#)

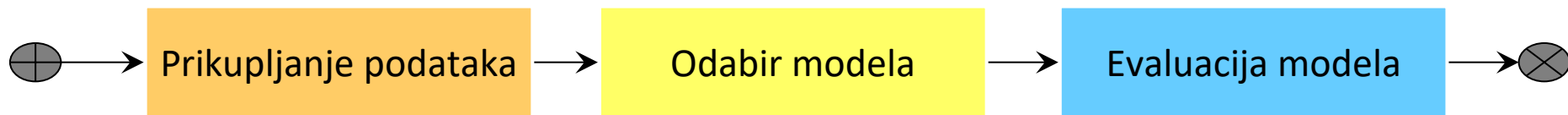
„Klasično“  
Mašinsko učenje

It is easy to sit in your office and run a Weka (Hall et al., 2009) algorithm on a data set you downloaded from the web.

It is very hard to identify a problem for which machine learning may offer a solution, determine what data should be collected, select or extract relevant features, choose an appropriate learning method, select an evaluation method, interpret the results, involve domain experts, publicize the results to the relevant scientific community, persuade users to adopt the technique, and (only then) to truly have made a difference (see Figure 1). An ML researcher might well feel fatigued or daunted just contemplating this list of activities. However, each one is a necessary component of any research program that seeks to have a real impact on the world outside of machine learning.

Primenjeno Mašinsko učenje

# Tipičan proces kod klasifikacije



# Prikupljanje podataka

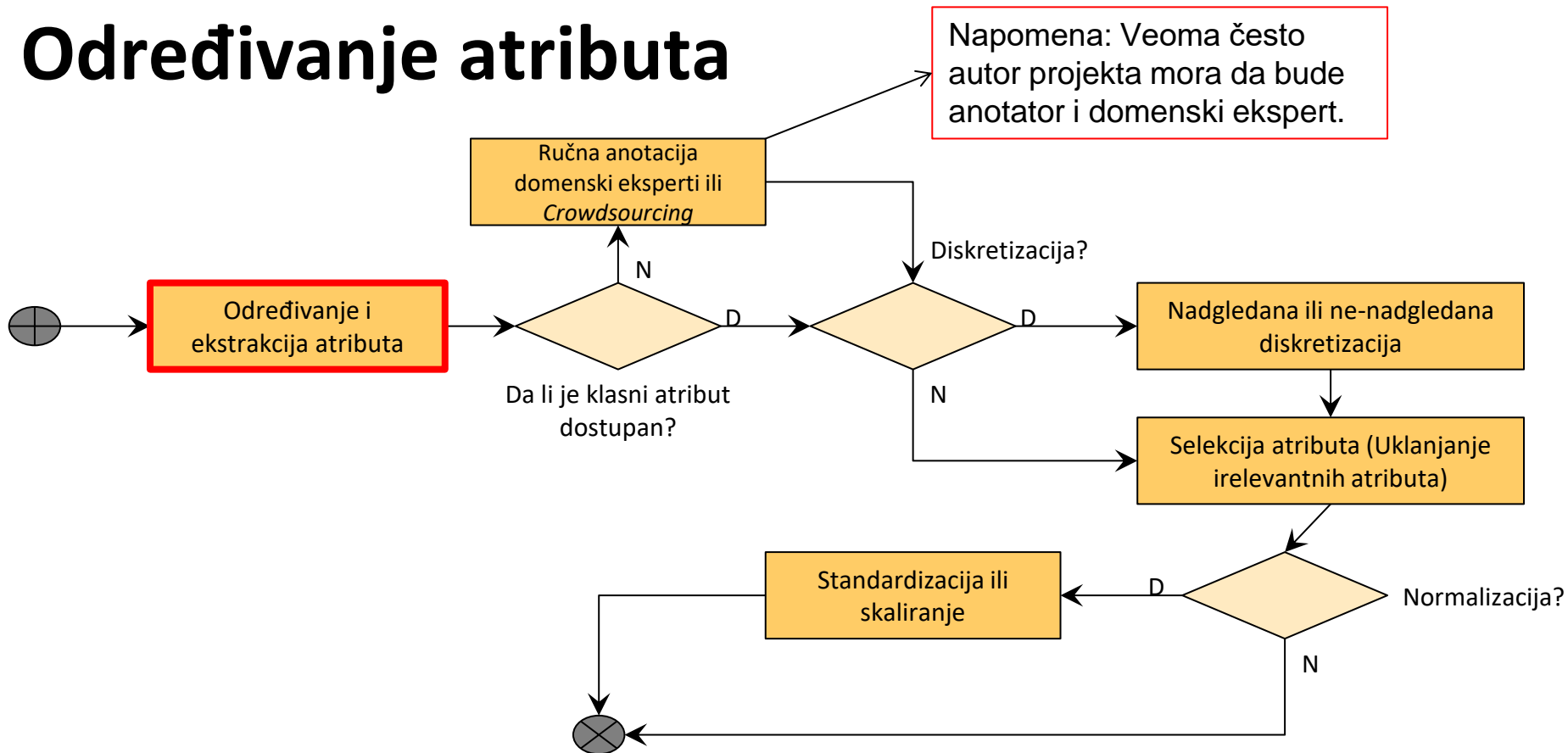
Prvi korak je prikupljanje podataka potrebnih za zadatak klasifikacije koji želimo da obavimo.

- Potrebno je da definišemo attribute (*features*) koji opisiju svaki slog (predmet, *item*) koji klasifikujemo.
- Potrebno je da definišemo klasni atribut (*class label*).

Domensko znanje je veoma važno u ovom koraku.

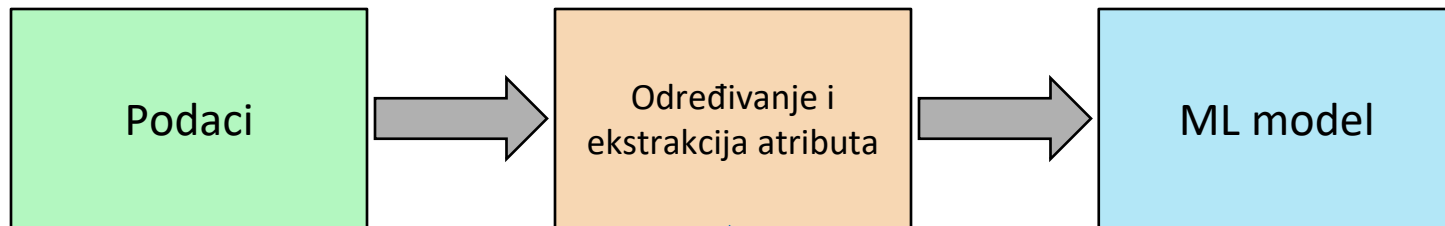
Šta ako je dodeljivanje klasnog atributa svakom slogu vremenski zahtevno (ili čak nemoguće u razumnom vremenu)?

# Određivanje atributa



# Kratka istorija ML

- Pre 2012\*, takođe veoma često i danas:

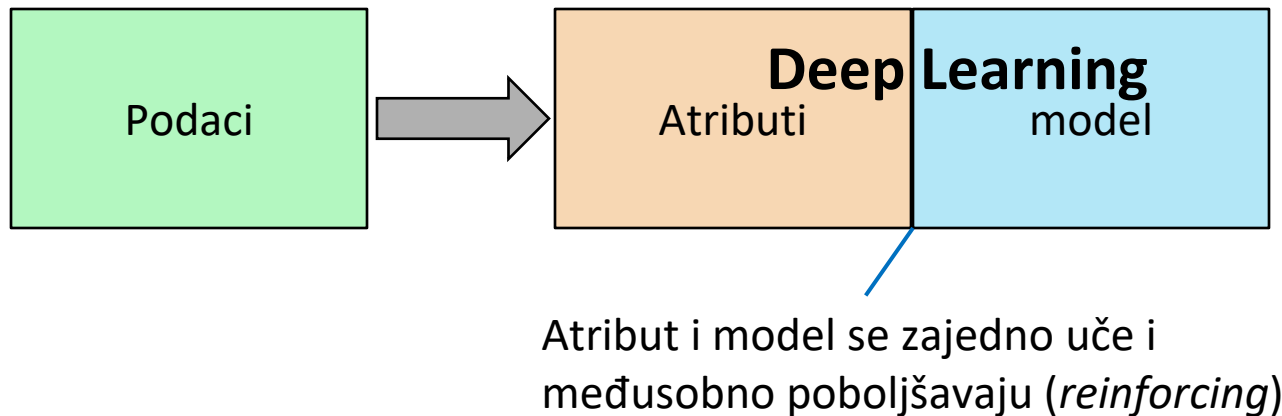


Najviše posla je kod ovog koraka.  
Konačan model je onoliko dobar,  
koliko je dobar skup atributa.

\* Pre publikacije Krizhevsky et al. - ImageNet CNN.

# Kratka istorija ML

- Posle 2012:



# Određivanje i ekstrakcija atributa

## Različite vrste atributa

- Numerički (godište, temperatura...)
- Kategorički (pol, boja očiju, )
- Ordinalni (kategorički atributi koji mogu da se porede npr. odgovori na anekatama, stepen stručne spreme, ...)

Novi atributi mogu se dobiti pomoću jednostavnih **statistika**

- *Određivanje i ekstrakcija atributa (Feature engineering)* je „umetnost“,
- Savet, pogledajte prvo koje attribute su drugi koristili za sličan zadatak (naučni radovi, Google, saveti iskusnijih kolega....)



**Primeri Određivanja i Ekstrakcija atributa  
(*feature engineering*)  
u SIAP projektima prethodnih generacija**

# Predikcija zarade filmova

Autori: Vladimir Ivković i Aleksa Mirković

Cilj:

- Predvideti zaradu filma pre prikazivanja
- Pronaći što više meta-podataka (atributa) o filmu iz različitih izvora
- Odrediti uticaj atributa na zaradu
- Razviti prediktivni model za zaradu

# Predikcija zarade filmova

- Preuzeti skup podataka je sadržao 16 atributa:
  - naslov filma, boja, broj kritika,
  - broj facebook svidanja filma, trajanje, i
  - imena i popularnost na facebook-u reditelja i glavnih glumaca,
  - zaradu filma, žanrove, broj korisnika koji su ocenili film,
  - broj lica na posteru filma, ključne reči,
  - jezik i državu filma, MPAA kategoriju,
  - budžet filma, godinu proizvodnje i ocenu na imdb-u.

# Predikcija zarade filmova

- Originalni skup podataka proširen je sa sledećim atributima:
  - Zarada nakon premijere, odnosno nakon prvog vikenda podaci o produkcionim kompanijama, nominacijama i nagradama na festivalima (<http://www.myapifilms.com/>)
  - Broj pregleda, lajkova, dislajkova i komentara na trejler (<http://www.youtube.com>)
  - Prosečna ocena i broj kritika sa Rotten Tomatoes, (<http://www.omdbapi.com/>)
  - Podaci o zaradama glumaca, reditelja i scenarista (<http://www.boxofficemojo.com/>)

# Predikcija zarade filmova

- Komentar:
  - Kod ovog projekta *Feature engineering* je pre svega bio zasnovan na domenskom znanju.
  - Samim tim najveći deo tereta bio je na *Data Wrangling* i EDA koracima
  - Tako je dobijeno 10+ korisnih atributa
  - Očigledno je da su *Feature engineering* i *Data Wr.* i EDA usko povezani i isprepletani procesi.
  - Već pomenuto:
    - To su procesi kod kojih ima najviše posla
    - To su procesi koji izdvajaju uspešnu primenu ML od neuspešne

# Još tri slična projekta

- Capital Bikeshare
- Procena broja stanovnika u Srbiji
- Klasterovanje heroja u *League of Legends*
- Kao kod prethodnog projekta akcenat je bio na
  - Domenskom znanju
  - Data Wrangling
  - EDA
- Za detalje pogledati Data Wrangling slajdove.

# Procena broja stanovnika za opštine u Republici Srbiji

## Izvori podataka

- Republički statistički zavod
- Podaci po opštinama i regionima
- Objavljaju se svake godine
- Sadrže brojne podatke o opštinama
  - procene broja stanovnika, procene prosečne starosti, vitalne događaje, broj školske dece, podatke o stambenoj izgradnji, podatke o zaposlenosti i zaradama i mnoge druge.

Za detalje pogledati Data Wrangling slajdove.

# Klasterovanje heroja u *League of Legends*

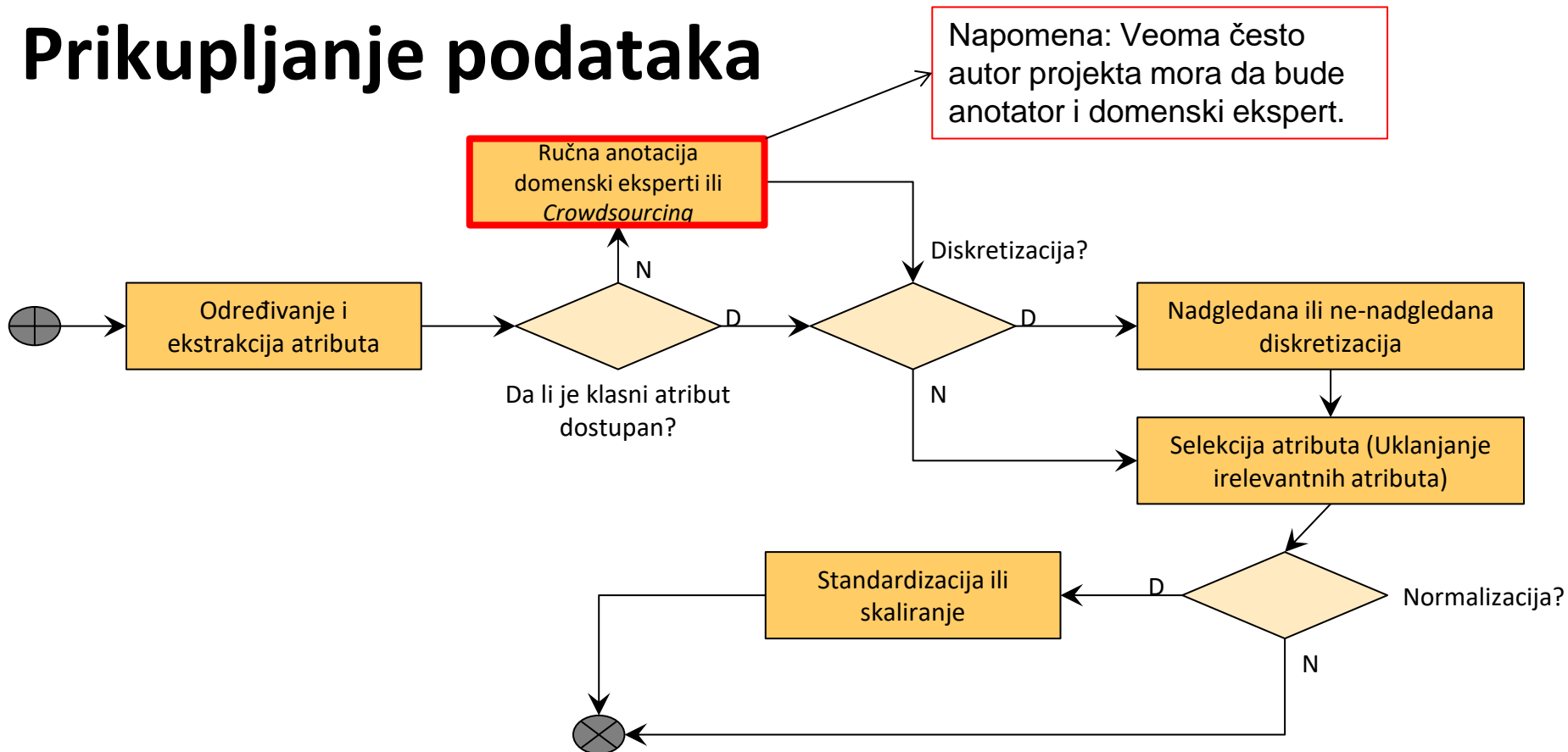
- Izvori podataka: LOL API
- Podaci na nivou meča
- Za ekstrakciju atributa bilo je potrebno:
  - Prikupljanje o svim predmetima koje su igrači kupovali za svoje heroje
  - Agregiranje tih podataka na nivo meča
  - Skaliranje na osnovu trajanja meča
  - Filtriranje ostalih podataka o meču (koji nisu o igračima)
- Jako puno posla do konačnog skupa od 8 atributa



# Klasterovanje heroja u *League of Legends*

- Konačan skup atributa:
  - Physical Damage Dealt
  - Magical Damage Dealt
  - Bonus Attack Damage
  - Bonus Ability Power
  - Bonus Health
  - Bonus Armor
  - Bonus Magic Resistance
  - Bonus Attack Speed

# Prikupljanje podataka



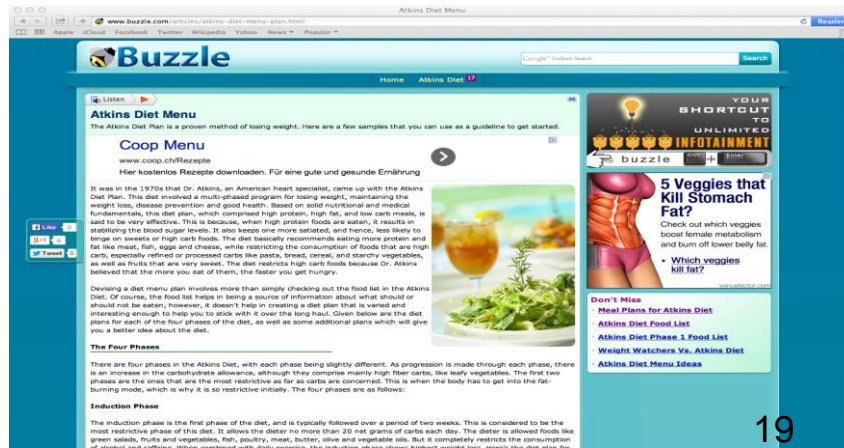
# Klasni atribut

Prikupljanje velike količine podataka je uglavnom lako.

Dodeljivanje klasnog atributa podacima (anotiranje) je vremenski zahtevno, teško i ponekad, nemoguće u razumnom vremenu.

Klasifikator koji klasifikuje web strane u pouzdane (*credible*) i nepouzidane.

Trebaju nam eksperti za dijete...



# Potencijalni anotatori

- Vi sami
- Prijatelji, kolege,...
- Domenski eksperti (\$\$\$)
- LLM-ovi – u poslednje vreme jako dobra opcija za anotiranje podataka.
  - Pogedati na primer: *A. Vujinović, N. Luburić, J. Slivka, and A. Kovačević, “Using ChatGPT to annotate a dataset: A case study in intelligent tutoring systems,” Mach. Learn. Appl., vol. 16, Jun. 2024, Art. no. 100557.*
- Crowdsourcing
  - Mogu biti i amateri i eksperti



Da li je web strana  
pouzdana?

1. Slanje zadatka



Korisnik

4. Prikupljanje  
odgovora



CC -C C -C



2. Prihvatanje  
zadatka

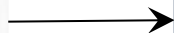


Radnici (Crowd)

3. Slanje odgovora

# Crowdsourcing

Agregacija odgovora (da li svima jednako verujemo? da li imamo informacije o radniku?)



Worker	Webpage	Credible
$W_1$	www.diet.com	C
$W_2$	www.diet.com	$\neg C$
$W_3$	www.diet.com	C
...	...	...



**Aggregation**



www.diet.com	C
--------------	---

# **Primeri određivanja klasnog atributa u SIAP projektima prethodnih generacija**

# Capital Bikeshare i Procena zarade filmova

- Klasni atribut (popularnost stanice, zarada filma) kreiran na osnovu:
  - Cilja projekta
  - Domeskog znanja autora
  - Agregacijom atributa (npr. br. Iznajmljenih i vraćenih bicikala)
  - Diskretizacijom uz pomoć grafičkog prikaza



# Klasfikacija tvitova po relevantnosti za određenu temu

Autor: Stanko Kuveljić

Cilj: Za tvitove sa tagom #cassandra odrediti koji se odnose na bazu podataka sa istim imenom, a koji na neku drugu temu

Izvor podataka: Twitter API

Metodologija:

- Bag-of-words

- Mašinsko učenje (klasifikacija)

# Klasfikacija tvitova po relevantnosti za određenu temu

Prikupljanje podataka je lako (Twitter API + upit za tag #cassandra)

Tako dobijamo ne označene podatke (*unlabelled*)

Šta je problem? Nemamo odvojene skupove tvitova po tome da li govore o bazi podataka ili nečem drugom

# Klasfikacija tvitova po relevantnosti za određenu temu

Rešenje?

Kreirati nenadgledani algoritam

Npr. Detektujemo reči „database“ i slično i onda tvrdimo da tvit govori o bazi ili

Koristimo nenadgledane tehnike za ekstrakciju tema (LDA) – detaljnije na predavanju o NLP

Šta je problem sa ovkavim rešenjima?

# Klasfikacija tvitova po relevantnosti za određenu temu

Jako puno posla je potrebno za kreiranje dobrog nenadgledanog algoritma (npr. kako odrediti i prikupiti indikatore?)

Dok kreiramo algoritam moraćemo da pogledamo tvitove koji su nam relevantni da bi znali šta tražimo

Kad to već radimo možemo i da kreiramo klasni atribut

# Klasfikacija tvitova po relevantnosti za određenu temu

Najozbiljniji problem je evaluacija

Kako proveriti da je algoritam dobar i ubediti nekog drugog u to (klijenta ili naučnu zajednicu) ako nemamo mere performasni?

Zašto ih nemamo? Zato što nam je potrebna klasa.

# Klasfikacija tvitova po relevantnosti za određenu temu

Rezime, ako nemamo odgovarajući obučavajući skup, moramo da ga kreiramo

Kreiramo ga ručnom anotacijom

Anotator redom za svaki tvit iz korpusa odlučuje da li pripada temi ili ne

Koliko nam podataka treba?

Što više to bolje. Treba pogledati veličine korpusa kod srodnih problema

# Klasfikacija tvitova po relevantnosti za određenu temu

Ko su anotatori?

Zavisi od tipa problema i dostupnih resursa (novca)

Ako je problem lak (kao ovaj) anotator može biti bilo ko (obično autor projekta)

Za teže probleme potrebni su nam eksperti. Npr. anotiranje medicinskih izveštaja ili pravnih dokumenata

Kod ovog projekta anotator je bio autor, kao i kolege iz frime u kojoj je zaposlen

Imati pomoć je mač sa dve oštrice:

- Anotacija brže ide

- Šta ako isti tvit dobije različite anotacije (ko je u pravu? Ko je objektivan?)

# Lovac na sendviče

Autor: Mihalio Isakov

Cilj: Prepoznavanje „botovskih“ komentara na Blic i B92 itd.

Izvor podataka: komentari sa Blic i B92

Metodologija:

- Bag of words

- ML



# Lovac na sendviče

Ovo je primer *crowdsourcing* projekta


Autor je odlučio da anotiranje ne radi sam već,

Da razvije dodatak za *Chrome* koji će omogućiti korisnicima da sami jednostavnim klikom označe tvit kao „botovski“

# Lovac na sendviče

<https://chrome.google.com/webstore/detail/lovac-na-sendviče/>

<https://github.com/Mihailolsakov/love-for-sandwiches>



## Lovac na sendviče

offered by Index studios


★★★★★ (24) | [News & Weather](#) | 184 users

[+ ADD TO CHROME](#) 

OVERVIEW



REVIEWS


RELATED



**~Rade :** Bolje da si doneo zakon o poreklu imovine, i oduzeti lopovima sve, kao sto si obecao pre izbora ali ti mnogo LAZES to je uvidela cela Srbija. Vucicu ti samo ne lazes Amere i EU jer nesmes, a narod, ko pita narod!? Zbog vlasti i nekih tvojih bolesnih ambicija spreman si na sve.


100% BOT! nije bot


Ocena:   | [Odgovor](#)



**~Kikica :** Nema razloga da ne budemo ponosni na ovo gradjevinsko cudo koje ce krasiti nasu prestonicu.To ce biti i novi napredak za nase gradjevske firme koje su do sad bile bez posla kao i napredak cele industrije.



100% BOT! nije bot


Ocena:   | [Odgovor](#)



**~misa :** evo ovo je dokaz da vladu vodi samo jedan covek, i da je hrabrosti i morala kod ostalih ministara, svi bi oni trebali da daju ostavke, pa neka vucic, sam vodi vladu, kao sto je i dokazano,


100% BOT! nije bot

Ocena:   | [Odgovor](#)

 Compatible with your device

### Ocenite koliko je komentar na B92, N1 ili Blicu zaslužio sendviča!

Lovac na sendviče je Chrome ekstenzija koja omogućava da ocenite da li je komentar na B92, N1 ili Blicu plaćen, tj. da li ga je pisao stranački bot. Na komentarima na B92 i Blicu će se pojaviti dva dugmeta: "BOT!" i "nije bot", kojima možete da prijavite komentar kao originalan ili plaćen. Vaše ocene se akumuliraju i služe za praćenje stila pisanja stranačkih botova. Nakon što ocenite komentar, videćete i kako su drugi ljudi ocenili komentar.

 [Report Abuse](#)

#### Additional Information

Version: 0.4.2  
Updated: March 23, 2016  
Size: 102KiB  
Language: Српски

USERS OF THIS EXTENSION HAVE ALSO USED

36

# Lovac na sendviče

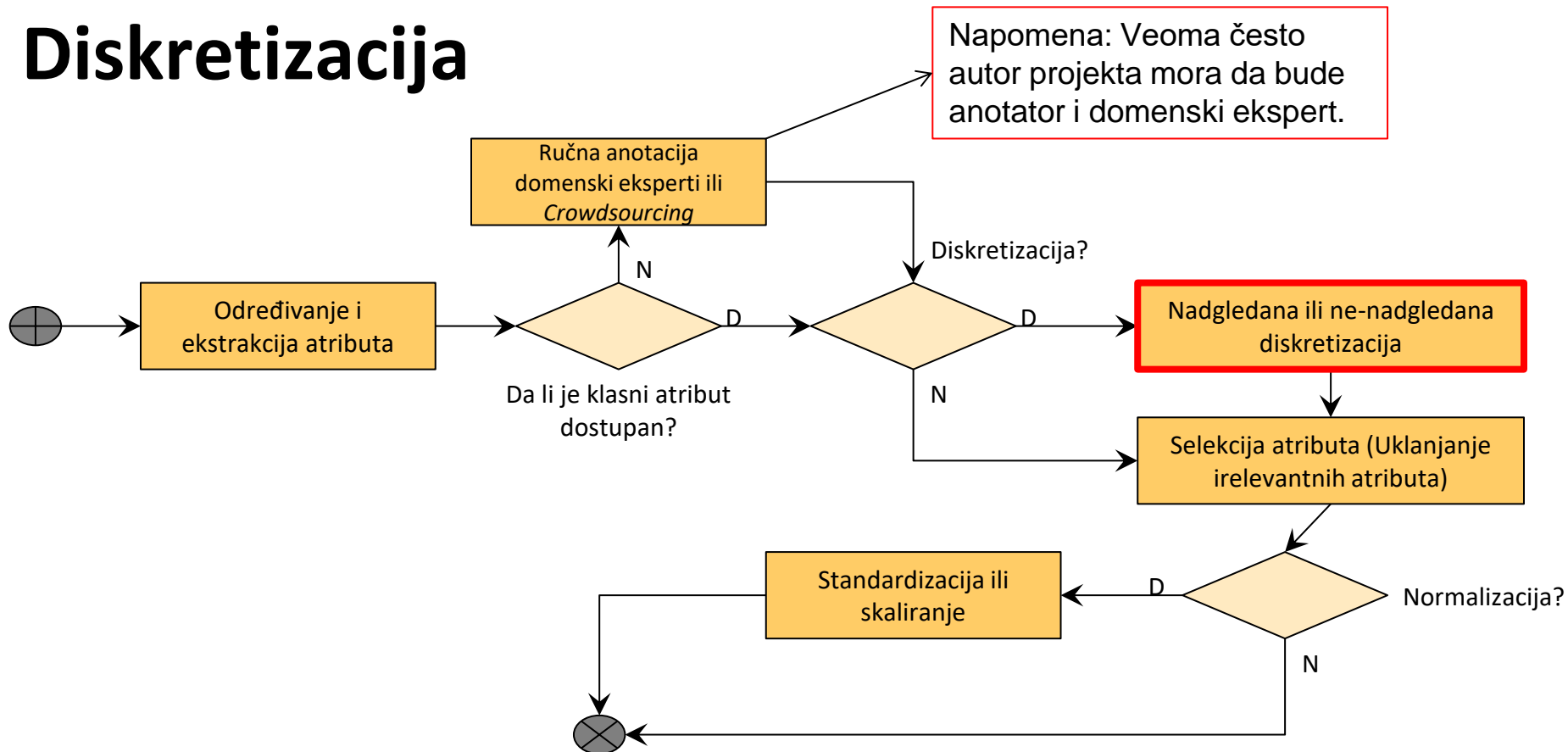
Brzina anotacije:

„U okviru od 24 sata bilo je poslato više od 2000 reakcija na 500 komentara. Aplikacija je dodatno reklamirana na Facebook-u u periodu od nedelju dana sa 20 dolara dnevno gde je bila viđena više od 200000 puta. U trenutku pisanja preko aplikacije je priklupljeno više od 12000 reakcija na 8000 komentara, od strane od više od 500 korisnika.“

Prednost: brzo prikupljanje velikog ob. skupa

Mana: pozdanost skupa (standardan problem kod ovog tipa anotiranja)

# Diskretizacija



# Diskretizacija

Zašto?

- Generalno diskretizacija nije nužna
  - Modeli kao što su stablo odlučivanja kojima trebaju diskretne vrednosti, diskretizaciju radi sami
- ali može biti korisna
  - kad sami želimo kontrolu nad diskretizacijom
  - kad model ne zahteva diskretne vrednosti, ali je prostor kontinualnih vrednosti preveliki
    - Npr. upotreba transformera ili RNN za rad sa podacima koji nisu tekst
    - Ti modeli generalno dobro rade kad su ulazi iz relativno ograničenog skupa (rečnika)
    - Ako to nije slučaj, diskretizacijom moramo sami ograničiti broj mogućih različitih ulaza

# Diskretizacija

## Nenadgledana

- Jednake širine



- Delimo raspon na predefinisan skup delova

- Jednake frekvencije



- Delimo raspon tako da svaki deo ima jednak broj slogova

- Klasterovanje



# Diskretizacija

## Nadgledana

- Radi se redom po susednim intervalima.
- Pomoću statističkih testova utvrđujemo da li nam informacija o tome da je primer u jednom ili drugom intervalu pomaže da odredimo njegovu klasu.
- Ako nam informacija ne pomaže, treba ih spojiti u jedan interval
- U suprotnom, treba da ostanu razdvojeni
- Test statističke nezavisnosti:  $\chi^2$  statistika i druge...
- Tema za razmišljanje: kako diskretizovati test skup da evaluacija bude realna?

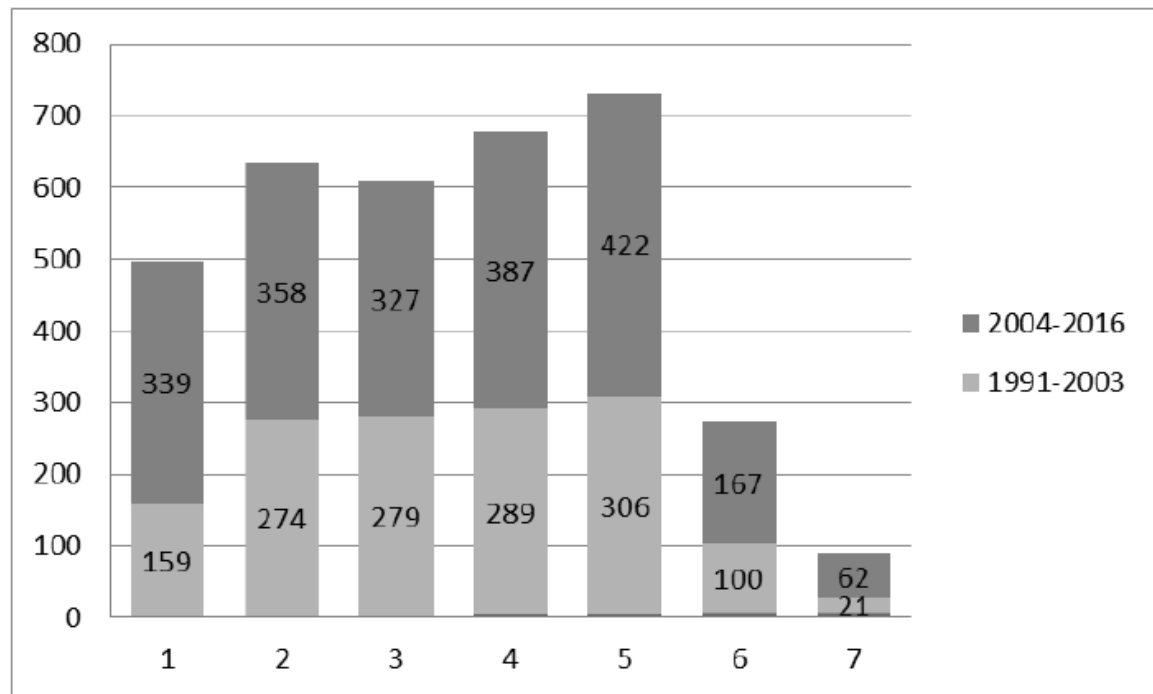
## **Primeri diskretizacije u svrhu kreiranja klasnog atributa u SIAP projektima prethodnih generacija**



# Capital Bikeshare i Procena zarade filmova

- Klasni atribut (popularnost stanice, zarada filma) kreiran na osnovu:
  - Agregacije atributa (npr. br. Iznajmljenih i vraćenih bicikala)
  - Diskretizacije uz pomoć grafičkog prikaza

# Procena zarade filmova



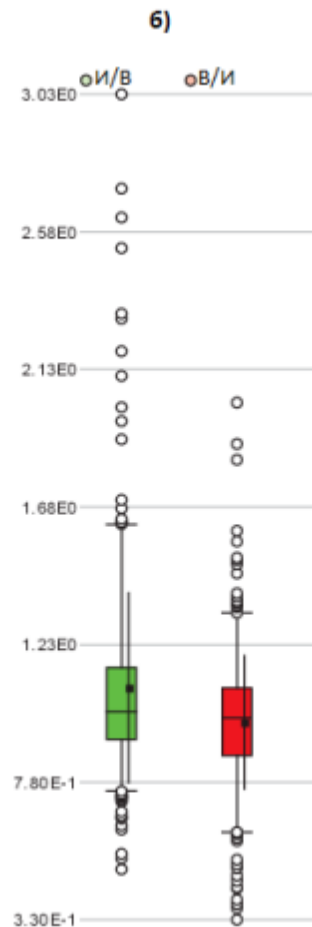
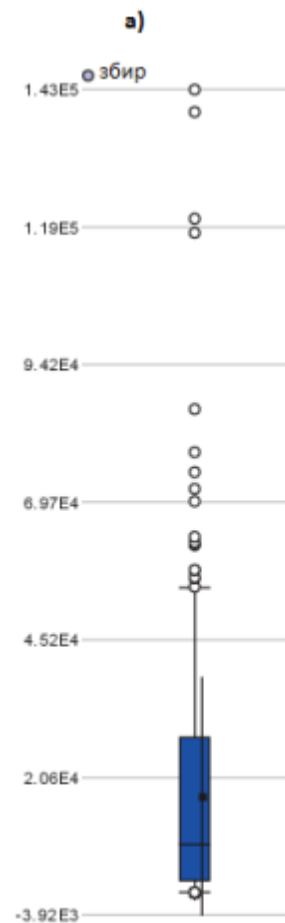
Za granice kategorija zarade uzete su vrednosti: \$1M, \$10M, \$25M, \$50M, \$125M, \$250M

*Slika 2. Zarada filma po kategorijama od 2004. do 2016. godine i od 1991. do 2003. godine*

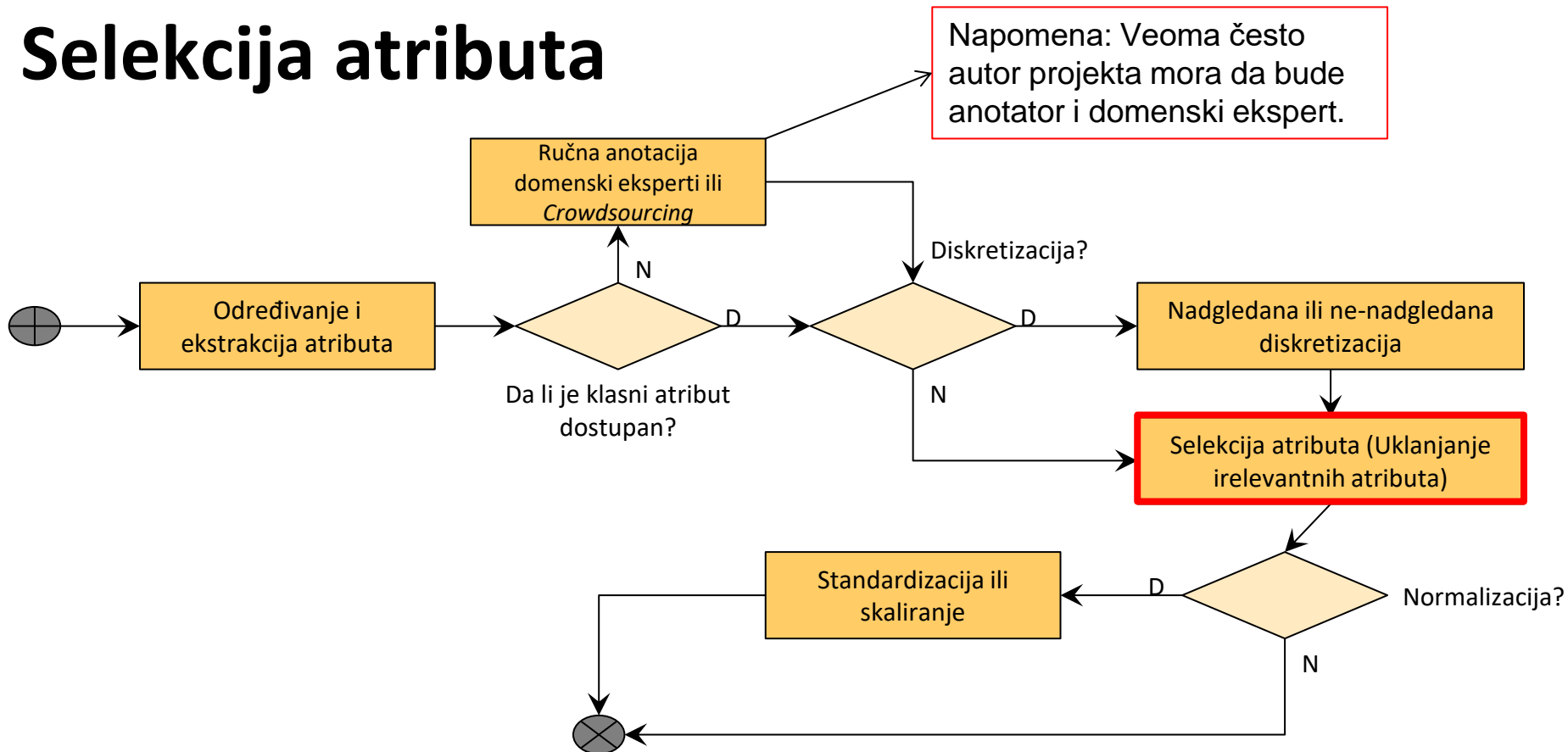
# Capital Bikeshare

Klasa je popularnost stanice, deli zbir iznajmljenih i vraćenih bicikala na intervale

оцена	доња граница	горња граница
0	0	0
1	0+	376
2	376+	2566
3	2566+	28646
4	28646+	54607
5	54607+	$\infty$



# Selekcija atributa



# Selekcija atributa (*Feature selection*)

Redukcija skupa  $N$  atributa na podskup  $M < N$  atributa sa najboljim mogućim performansama

Postoji  $2^N$  mogućih podskupova

Metode:

- **Filtriranje**
- **Omotač (*Wrapper*)**

# Selekcija atributa

Filtriranje: rangiranje metoda po prediktivnoj moći i odabir najboljih

- Prednosti

- Ne zavisi od klasifikacionog modela (objavlja se samo jednom u procesu razvoja modela)

- Mane

- Ne zavisi od klasifikacionog modela (nema interakcije sa modelom, različitim modelima mogu da odgovaraju različiti atributi)
- Prepostavka je da su atributi međusobno nezavisni

# Rangiranje atributa

## Numerički atributi

- Koeficijent korelacije (*Pearson*, linearni odnos)

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

- Zajednička informacija (*Mutual information*)

$$I(F; C) = H(C) - H(C | F) = H(F) + H(C) - H(F, C)$$

$$H(F) = - \sum_i P(f_i) \log_2 P(f_i)$$

$$H(F, C) = - \sum_i \sum_j P(f_i, c_j) \log_2 P(f_i, c_j)$$

Napomena:

$r$  i  $I$  se računaju za svaki atribut sa jedne strane i klasu sa druge strane. Atribut kod koga  $r$  i  $I$  predju neki prag ostaje u skupu podataka.

# Rangiranje atributa

## Kategorijalni atributi

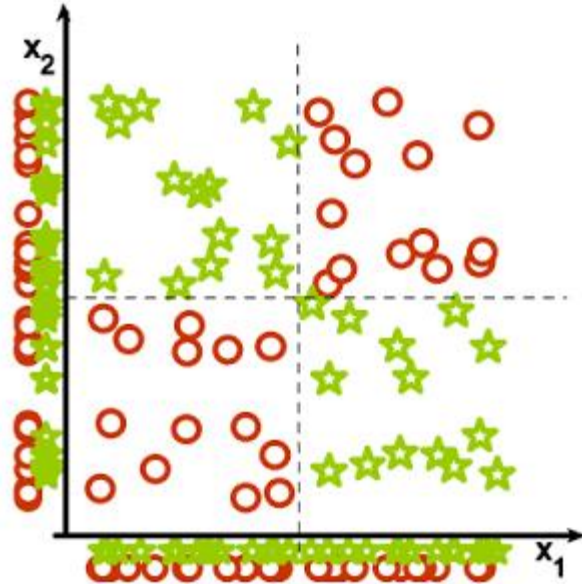
- **$\chi^2$  metod**

Drugačije od korelacije, hi-kvadrat test proverava da li su atribut i klasa nezavisni, ali ne pruža informaciju o pravcu ili jačini veze.



# Rangiranje atributa

Atributi koji su zajedno relevantni mogu delovati irelevantno ako se posmatraju odvojeno.



# Selekcija atributa – Wrapper (omotač) metode

**Ablacija:** interaktivno **uklanjanje** atributa, koristeći unakrsnu-validaciju (ili validacioni skup) za dodnošenje odluka. Zauzatatavljanje kad nema poboljšanja (ili atributa).

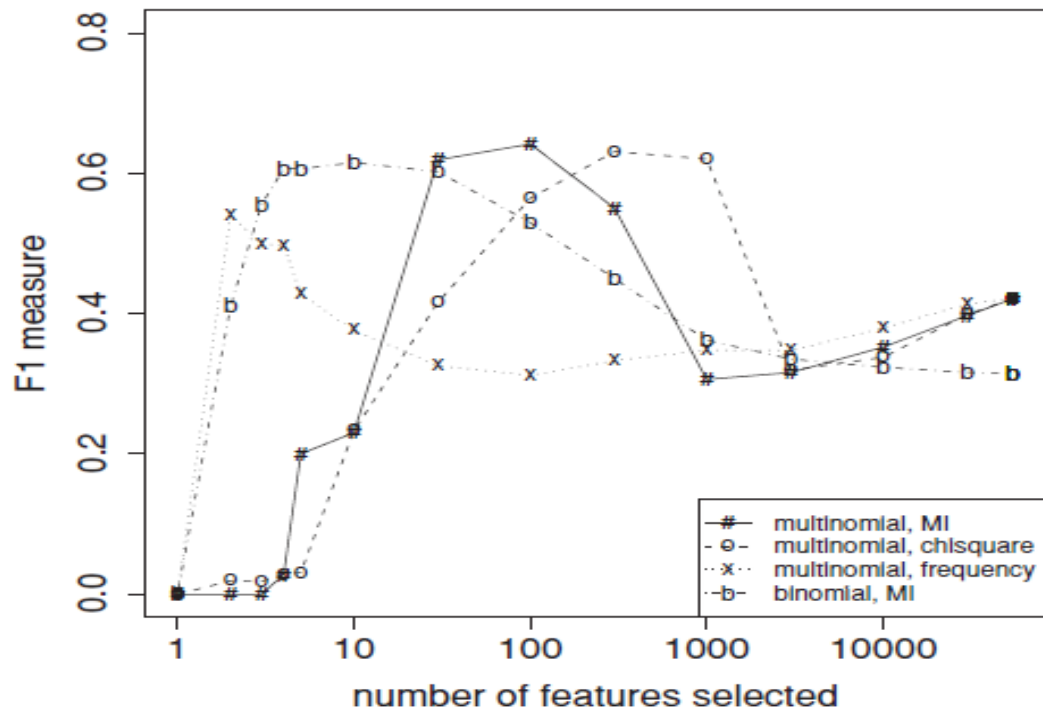
- **Prednosti**

- Intrerakcija sa klasifikacionim modelom
- Nema pretpostavke o nezavisnosti atributa

- **Mane**

- Računski zahtevno
- Pohlepan (greedy) metod

# Primer, Broj atributa vs. F-mera



Primeri selekcije atributa u SIAP projektima  
prethodnih generacija

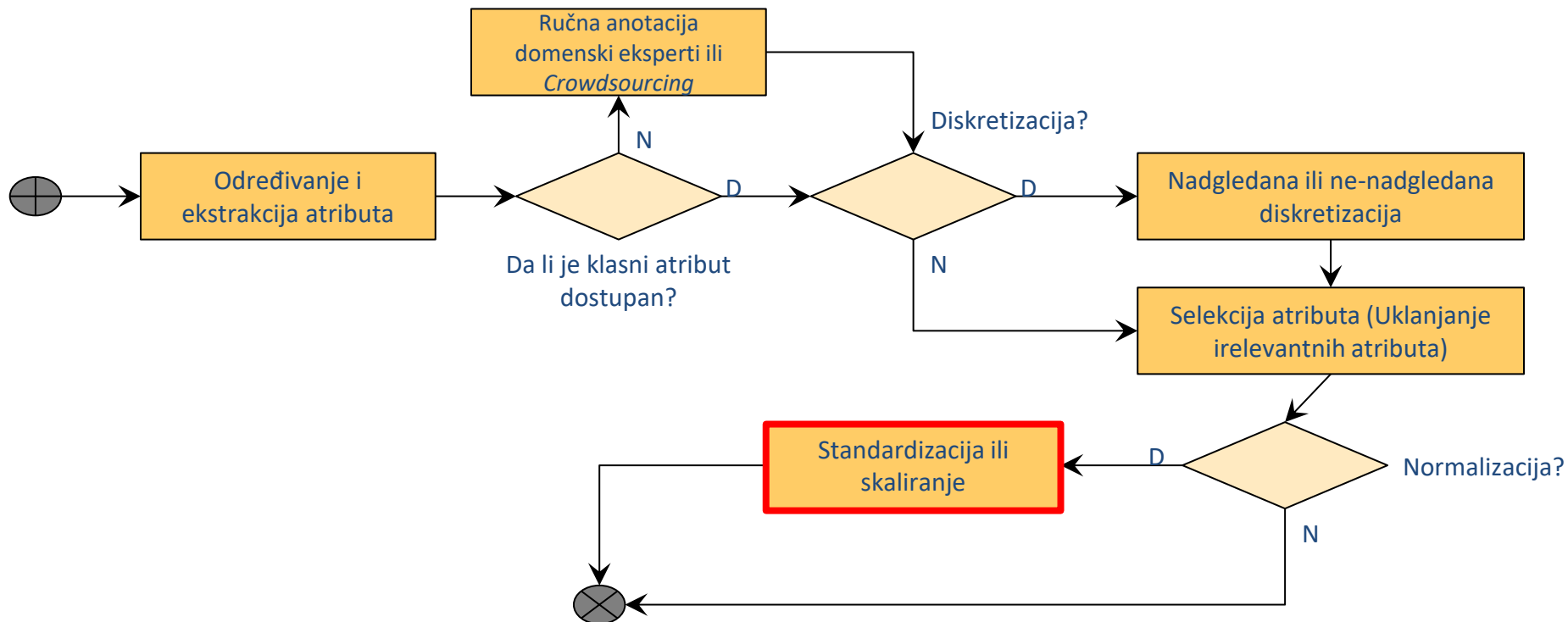
# Slekcijske osobine u SIAP projektima

Kod većine projekata korišćene su wrapper metode forward (dodavanje) ili backward (ablacija) selekcija u alatu RapidMiner (RM nudi i genetski algoritma za selekciju koji nije greedy)

Kod projekata sa malo atributa autori su uglavnom sami vršili odabir atributa na osnovu domenskog znanja

Kod text mining projekata atributi su rangirani uglavnom po frekvenciji

# Normalizacija atributa



# Normalizacija atributa

Neki klasifikatori ne funkcionišu dobro ako su vrednosti atributa u različitim rasponima

- Zarada: 10,000,000 (CHF)
- Broj zaposlenih: 300

Atributi sa velikim vrednostima obično dominiraju nad onim sa manjim pri formiranju modela.

# Standardizacija

$$x_i' = (x_i - \mu_i) / \sigma_i$$

Gde je  $\mu_i$  srednja vrednost atributa  $x_i$ , a  $\sigma_i$  je standardna devijacija

Novi atribut  $x_i'$  ima srednju vrednost 0 i standardnu devijaciju 1



# Min-max Skaliranje

$$x_i' = (x_i - m_i) / (M_i - m_i)$$

Gde su  $M_i$  i  $m_i$  max i min vrednosti atributa  $x_i$

Novi atribut  $x_i'$  biće u intervalu  $[0,1]$

U nekim alatima Min-max Skaliranje mapira vrednosti na  $[-1,1]$  pomoću vrlo slične formule.

# Standardizacija vs Skaliranje

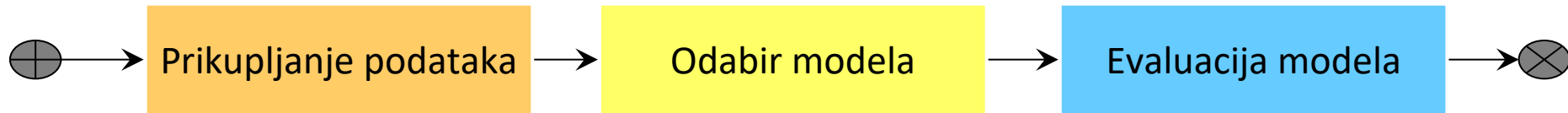
## Standardizacija:

- Pretpostavlja da su vrednosti atributa generisane po Gausovoj raspodeli

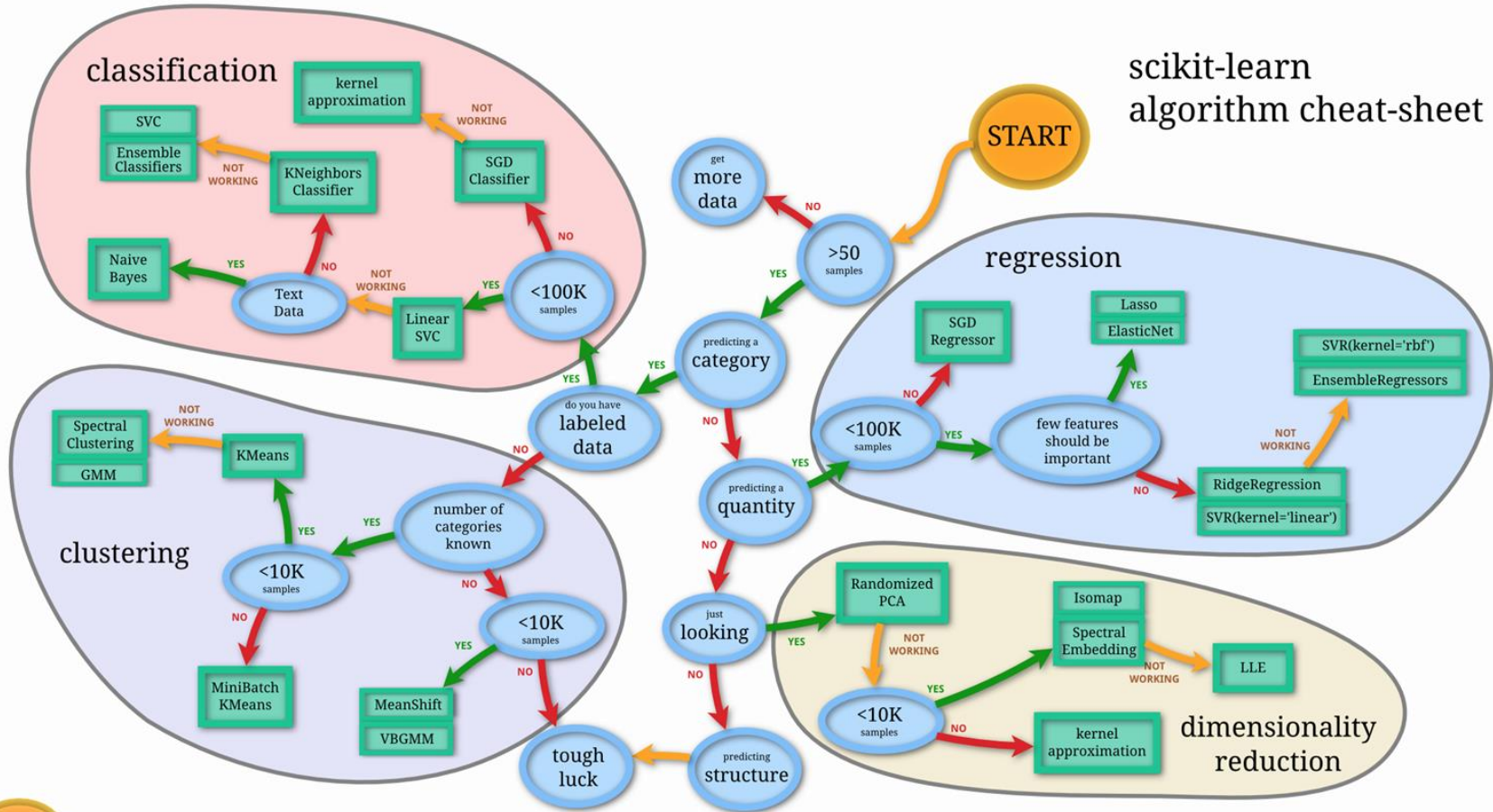
## Skaliranje:

- Ako podaci sadrže autlajere, “normalne” (tipične) vrednosti će biti skalirane na veoma mali interval

# Tipičan proces kod klasifikacije – Odabir modela



# scikit-learn algorithm cheat-sheet





Obavezno pogledati slična  
rešenja i videti koji se modeli  
koriste kod njih!

# Dobri izvori za pronalaženje SOTA modela

SOTA = State-of-the-art

- Slike, video, tekst...
  - <https://paperswithcode.com/sota>
  - Twitter – dosta lako možete pronaći uticajne ljude u oblasti i pratiti ih
  - Google naravno...
- Tabelarni podaci
  - *XGboost* je trenutno najbolji model, ali uvek probajte sve poznate modele.
  - *Catboost* je takođe trenutno vrlo akutelan.

# https://paperswithcode.com/task/sentiment-analysis

Browse SoTA > Natural Language Processing > Sentiment Analysis



## Sentiment Analysis


430 papers with code · [Natural Language Processing](#)

Edit

Sentiment analysis is the task of classifying the polarity of a given text.

## Benchmarks

Add a Result

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE	COMPARE
	SST-2 Binary classification	🏆 T5-3B	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer			See all
	SST-5 Fine-grained classification	🏆 BERT Large	Fine-grained Sentiment Classification using BERT			See all
	IMDb	🏆 NB-weighted-BON + dv-cosine	Sentiment Classification Using Document Embeddings Trained with Cosine Similarity			See all

# Odabir modela

Klasifikator obično ima parametre čije vrednosti treba odrediti („naštelovati“)

- Faktor regularizacije
- Neki prag
- Metrike za rastojanje
- Broj komšija u KNN
- ....

Potrebne su nam mere perfomansi.



# Značaj optimizacije parametara

## COMPUTER SCIENCE

### Core progress in AI has stalled in some fields

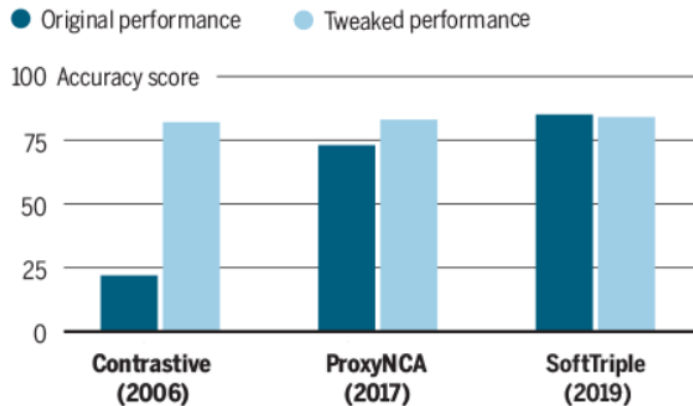
When tuned up, old algorithms can match the abilities of their successors

By Matthew Hutson

Artificial intelligence (AI) just seems to get smarter and smarter. Each iPhone learns your face, voice, and habits better than the last, and the threats AI poses to privacy and jobs continue to grow. The surge reflects faster chips, more data, and better algorithms. But some of the improvement comes from tweaks rather than the core innovations their inventors claim—and some of the gains may not exist at all, says Davis Blalock, a computer science graduate student at the Massachusetts Institute of Technology (MIT). Blalock and his colleagues compared dozens of approaches to

#### Old dogs, new tricks

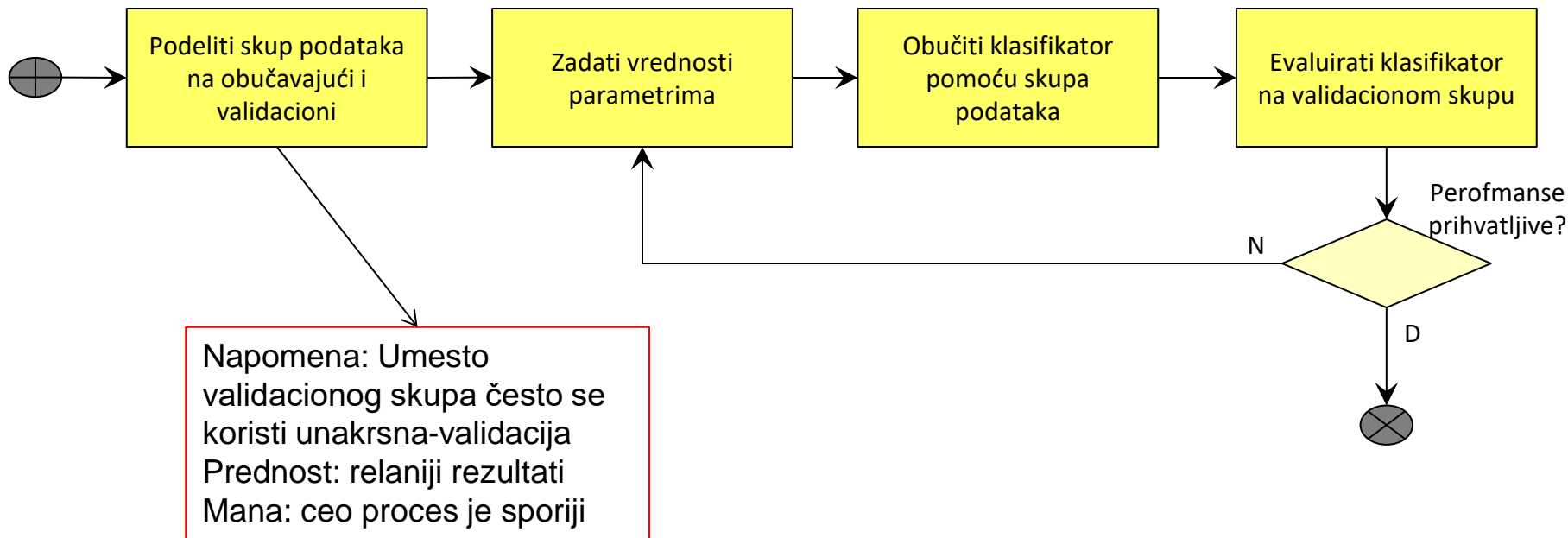
After modest tweaks, old image-retrieval algorithms perform as well as new ones, suggesting little actual innovation.



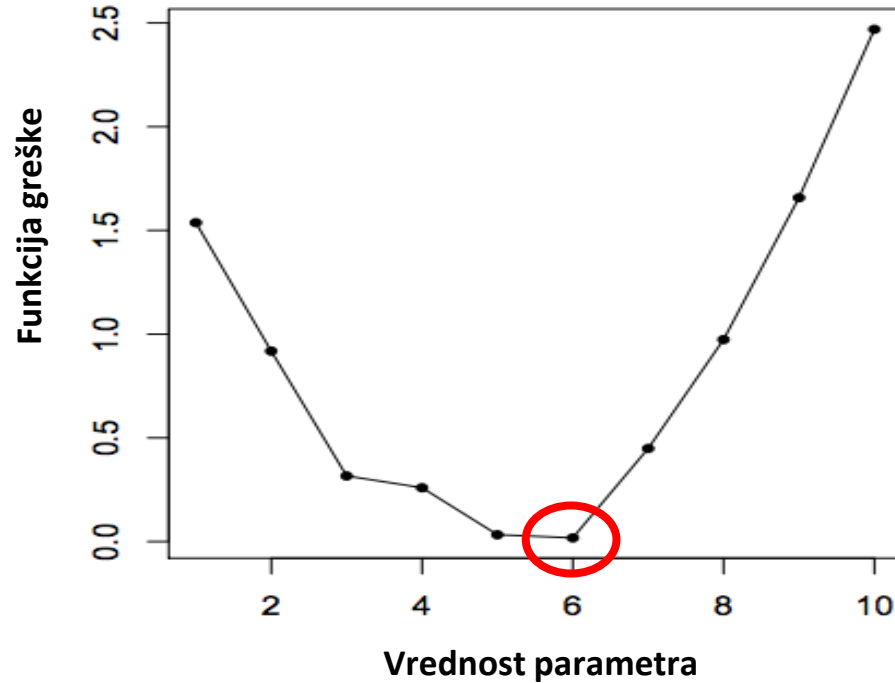
CREDITS: (GRAPHIC) X. LIU/SCIENCE; (DATA) MUSGRAVE ET AL.,  
[ARXIV: 2003.08505](https://arxiv.org/abs/2003.08505)

<https://arxiv.org/abs/2003.08505>

# Odabir modela



# Odabir modela



# Mere performansi za binarnu klasifikaciju

Kod binarne klasifikacije postoje četiri tipa greške:

- True Positive (pozitivni primeri klasifikovani kao pozitivni)
- True Negative (negativni primeri klasifikovani kao negativni)
- False Positive (negativni primeri klasifikovani kao pozitivni)
- False Negative (pozitivni primeri klasifikovani kao negativni)

## Napomene:

- 1.Koristimo engleske termine jer su rasprostranjeni u literaturi
- 2.Korisnik ili alat određuju koja se klasa smatra za pozitivnu ili negativnu

		Tačna klasa	
		A	B
Rezultat klasifikatora	A	TP	FP
	B	FN	TN

# Tačnost (*Accuracy*)

$$A = \frac{TP+TN}{TP + TN + FP + FN} = \frac{TP+TN}{N}$$

Prikladna mera kad su:

- Klase balansirane (približno jednak broj primera jedne i druge)
- Tipovi grešaka su nam jednako bitni (nebitni)

# Tačnost (neizbalansiran primer)

Klasifikator 1		Tačna klasa	
		Fraud	¬Fraud
Rezultat klasifikatora	Fraud	5	10
	¬Fraud	5	80

$$A = 85/100 = 0.85$$

Predikcija uvek = ¬Fraud		Tačna klasa	
		Fraud	¬Fraud
Rezultat klasifikatora	Fraud	0	0
	¬Fraud	10	90

$$A = 90/100 = 0.90$$

# Tačnost (*Accuracy*)

Klasifikator 1		Tačna klasa	
		A	B
Rezultat klasifikatora	A	45	20
	B	5	30

$$A = 75/100 = 0.75$$

Klasifikator 2		Tačna klasa	
		A	B
Rezultat klasifikatora	A	40	10
	B	10	40

$$A = 80/100 = 0.80$$

Koji klasifikator je bolji?

- A. Klasifikator 1
- B. Klasifikator 2
- C. Oba su jednako dobri

# Tačnost (*Accuracy*)

Klasifikator 1		Tačna klasa	
		Cancer	¬Cancer
Rezultat klasifikatora	Cancer	45	20
	¬Cancer	5	30

Klasifikator 2		Tačna klasa	
		Cancer	¬Cancer
Rezultat klasifikatora	Cancer	40	10
	¬Cancer	10	40

Koji klasifikator je bolji?

- A. Klasifikator 1
- B. Klasifikator 2
- C. Oba su jednako dobri



# Preciznost i Odziv

Preciznost (*Precision*)

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

Odziv (*Recall*)

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

# Preciznost i Odziv

Klasifikator 1		Tačna klasa	
		Cancer	¬Cancer
Rezultat klasifikatora	Cancer	45	20
	¬Cancer	5	30

$$P_1 = 45/65 = 0.69$$

$$R_1 = 45/50 = 0.9$$

Klasifikator 2		Tačna klasa	
		Cancer	¬Cancer
Rezultat klasifikatora	Cancer	40	10
	¬Cancer	10	40

$$P_2 = 40/50 = 0.8$$

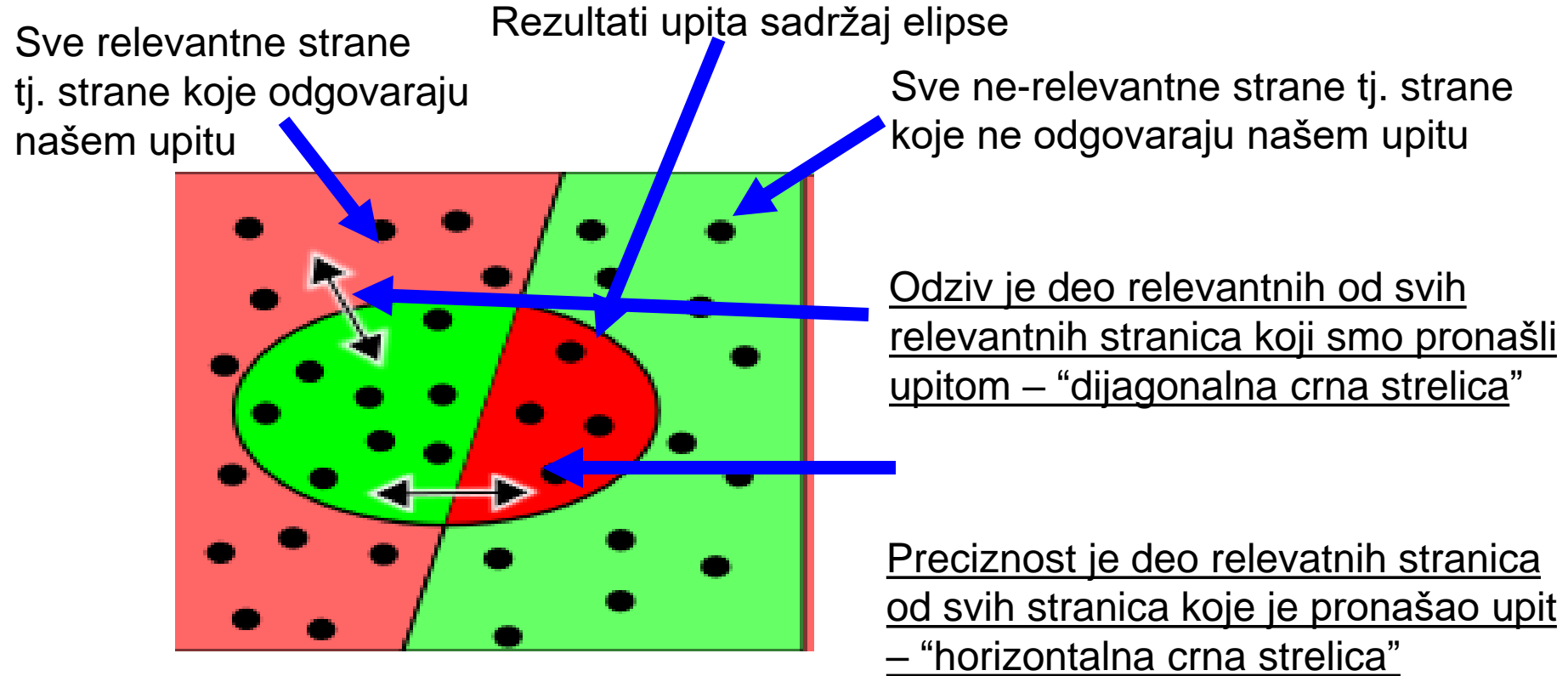
$$R_2 = 40/50 = 0.8$$

Model: "Svako ima rak"		Class	
		Cancer	¬Cancer
Classified	Cancer	50	50
	¬Cancer	0	0

$$P = 50/100 = 0.5$$

$$R = 50/50 = 1$$

# Preciznost i Odziv na primeru pretraživanja web strana



# F-mera (*F-measure* ili *F-score*)

Pogodno je imati jednu meru da bi lakše mogli da poredimo klasifikatore

F mera (ili F1 mera)

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

Takođe postoji mogućnost da preciznost i odziv **imaju različite težine**, ako nam je jedna mera važnija od druge. Nije uobičajeno.

# Precision and recall

Klasifikator 1		Tačna klasa	
		Cancer	¬Cancer
Rezultat klasifikatora	Cancer	45	20
	¬Cancer	5	30

$$F_1 = 2 * (0.69 * 0.9) / (0.69 + 0.9) = 0.78$$

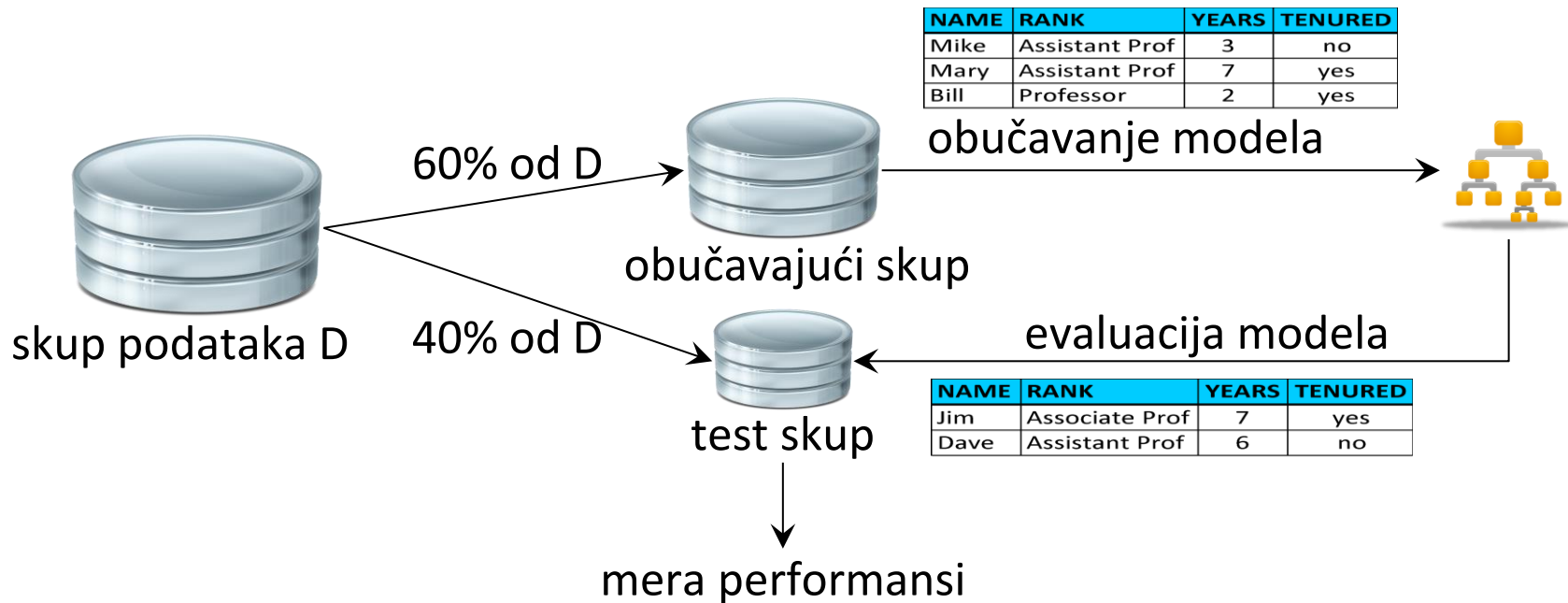
Klasifikator 2		Tačna klasa	
		Cancer	¬Cancer
Rezultat klasifikatora	Cancer	40	10
	¬Cancer	10	40

$$F_2 = 2 * (0.8 * 0.8) / (0.8 + 0.8) = 0.8$$

Model: "Svako ima rak"		Class	
		Cancer	¬Cancer
Classified	Cancer	50	50
	¬Cancer	0	0

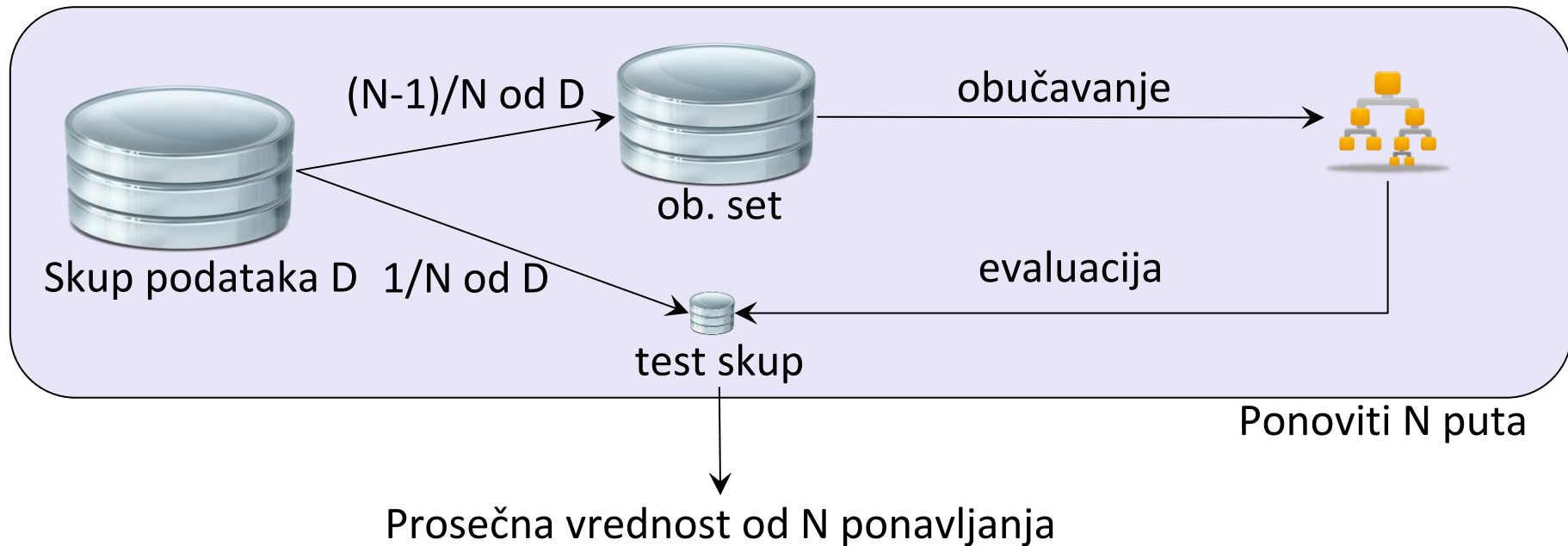
$$F = 2 * (0.5 * 1) / (0.5 + 1) = 0.66$$

# Obučavajući i test skup



# Leave-one-out unakrsna validacija

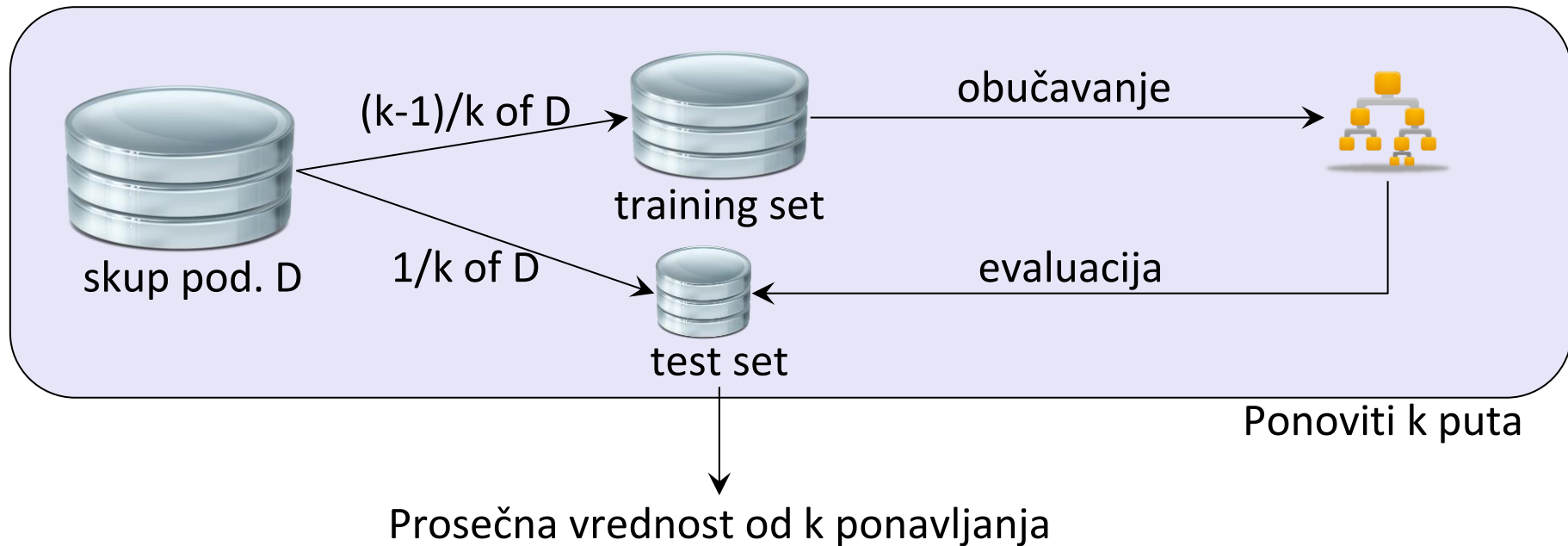
Obučavajući skup D sadrži N primera.



Test skup uvek sadrži samo jedan primer. Idemo redom po svim primerima pa računamo prosek na kraju. Veoma sporo...

# K-fold cross validation, k - unakrsna validacija

Obučavajući skup D sadrži N primera. k je deo od D. Na primer k=5 je peti deo D.





# Bijas (očekivana greška) i Varijansa

## *Bias and Variance*

Bijas – očekivana greška modela u primeni

Varijansa – očekivana varijabilnost rezultata modele kada se primeni na različite skupove podataka

Mogu se odrediti upoređivanjem performansi modela na obučavajućem i validacionom skupu.

Poređenje se vrši postepenim povećavanjem obučavajućeg skupa.

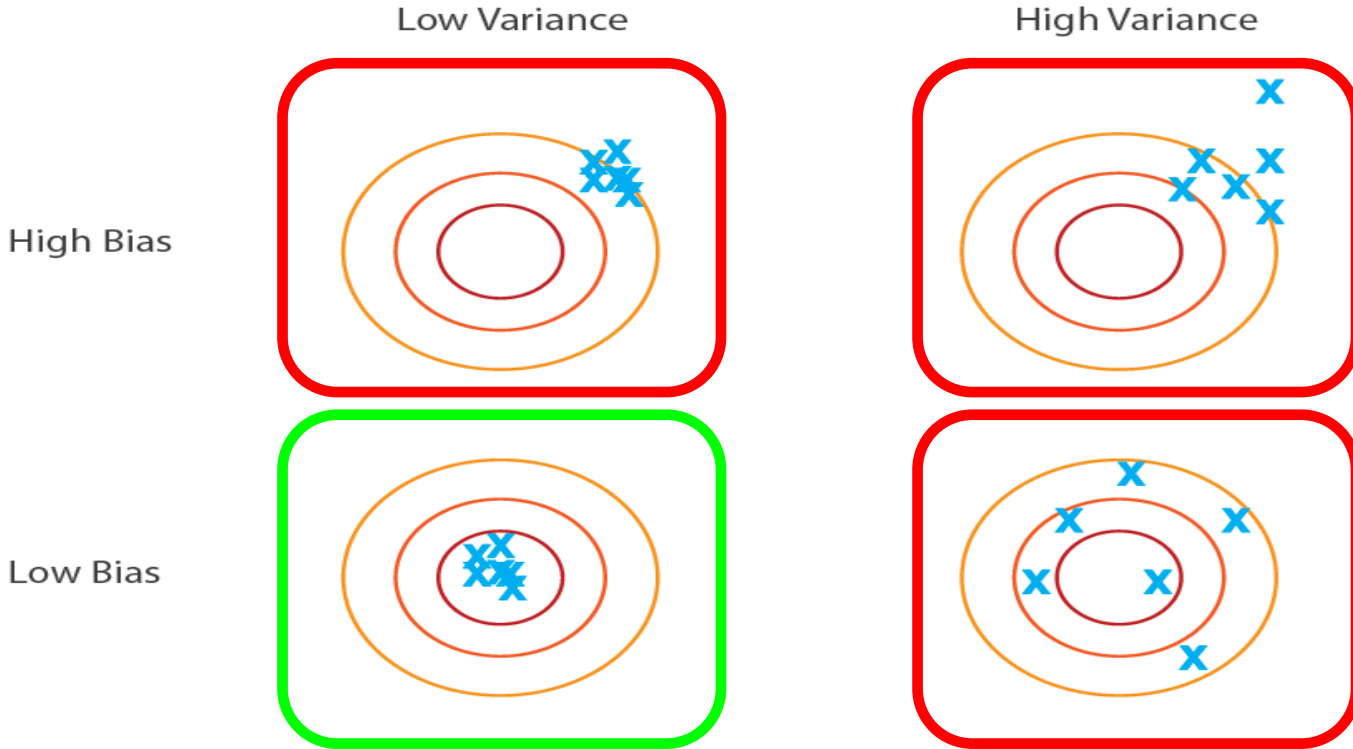
# Bijas (očekivana greška) i Varijansa

## *Bias and Variance*

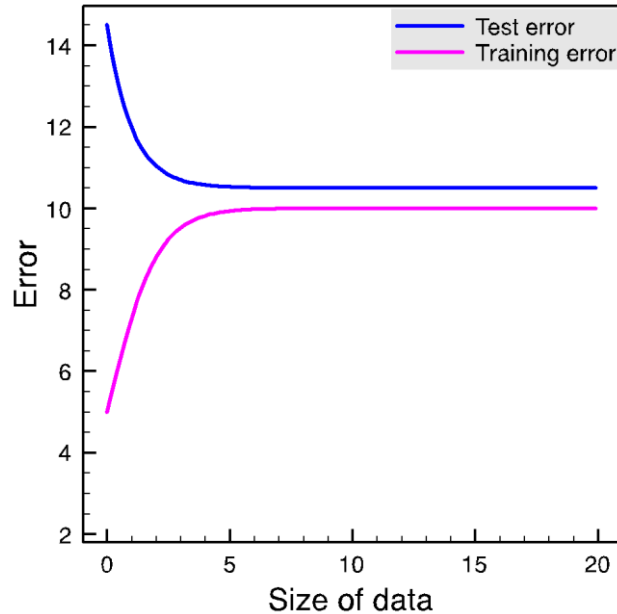
Is crtavanje grafikona poređenja performansi na ob. i validacionom skupu bi generalno trebalo raditi uvek pri razvoju klasifikatora.

U idelanom slučaju hoćemo mali bijas (malu grešku na obučavajućem skup) i malu varijansu (malu grešku na validacionom skupu).

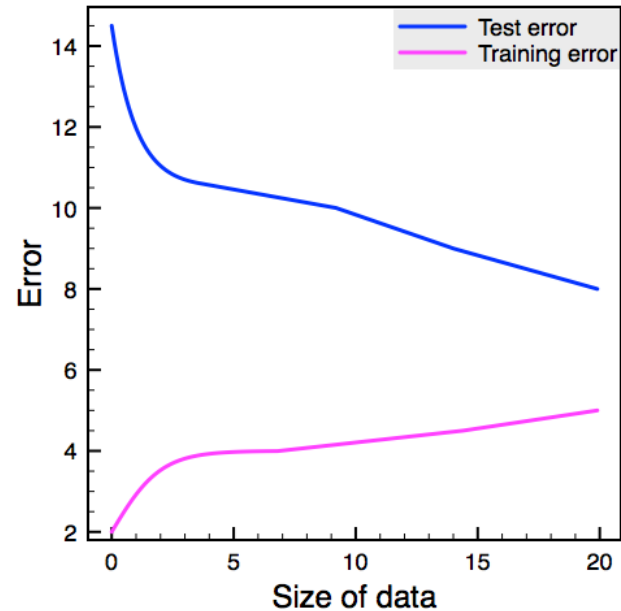
# Bijas (očekivana greška) i Varijansa



# Kada povećavanje ob. skupa pomaže



Veliki bijas

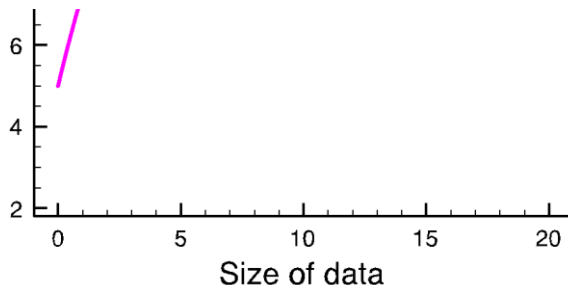


Velika varijansa

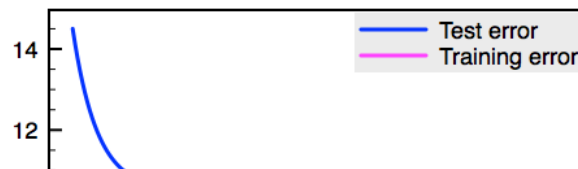
# Kada povećavanje ob. skupa pomaže



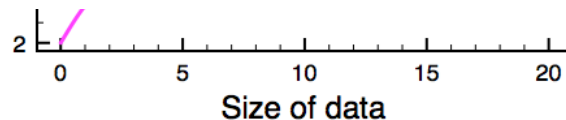
Savet: odabrati kompleksniji (fleksibilniji model)



Veliki bijas

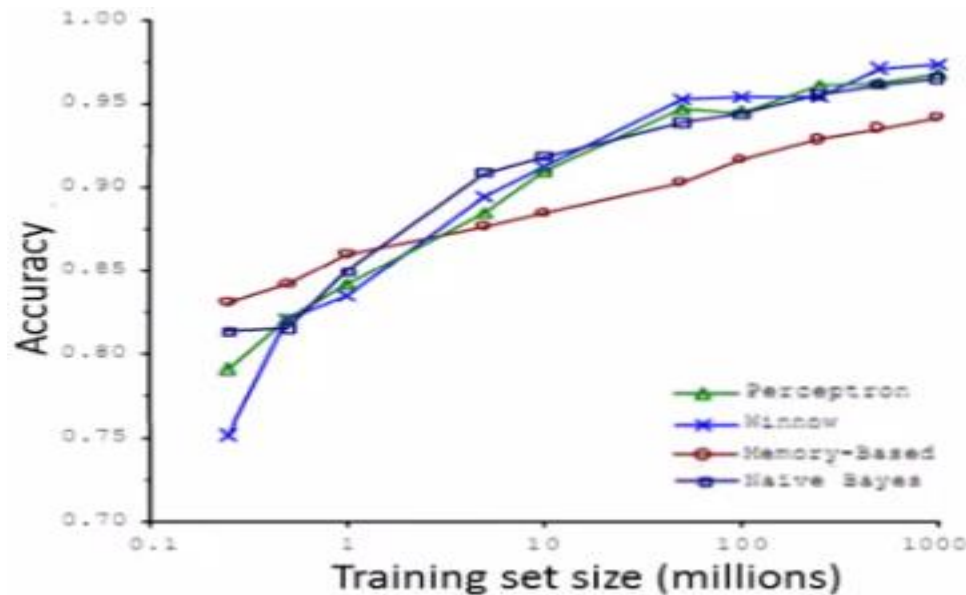


Savet: 1. povećati obučavajući skup.  
2. Probati jednostavniji model.



Velika varijansa

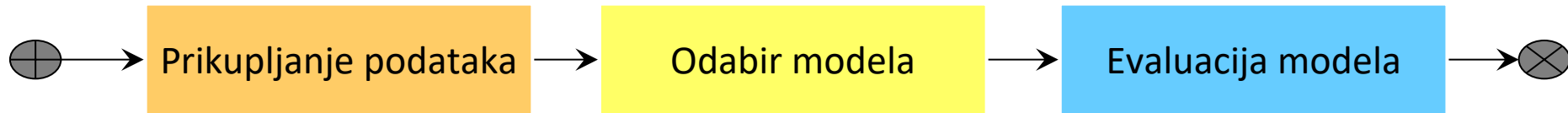
# Kada povećavanje ob. skupa pomaže



Data > Algorithms

Vreme je pokazalo da ova tvrdnja stoji...  
uz dodatak Data+Hardware

# Tipičan proces kod klasifikacije – Odabir modela



# Evaluacija modela

Metode evaluacije i mere performansi koje smo do sada pokazali koriste se za:

- Selekciju atributa

- Selekciju modela

- Evaluaciju modela (konačnu ocenu kvaliteta modela)



# Evaluacija modela, važna napomena

Kada radimo proces selekcije atributa i „šteloovanja“ parametara model evaluriamo na dva načina:

validacioni skup (deo ob. skupa)

unakrsna validacija

Dobijene mere performanse na ovaj način nisu realne (previše su optimistične)!

Nisu realne u smislu da ih ne možemo upotrebiti kao konačnu procenu kvaliteta našeg modela.

# Evaluacija modela, važna napomena

Zašto su mere dobijene u selekciji modela i atributa nerealne?

Zato što modele i parametre bирамо тако да добијемо najbolje rezultate

Tačnije mi znamo klase u validacionom skupu (ili unakrs. val.) i prilagođavamo model njima.

Podaci na koje ćemo primentiti model u praksi nemaju klasu tako da se ne možemo prilagoditi njima

Dakle treba nam realna ocena modela.

# Evaluacija modela, važna napomena

Dakle treba nam realna ocena modela

Realnu ocenu modela dobićemo tako što ćemo ga evaluirati na delu skupa podataka koji ni na koji način nije učestvovao u selekciji modela i atributa

Taj skup obično izdvajamo nakom procesa određivanja atributa. (ali pre selekcije atributa!)

Taj skup se u praksi (i literaturi) naziva test skup, dok je validacioni onaj koji učestvuje u procesu razvoja modela.

# Evaluacija modela, važna napomena

Još par napomena

Test skup mora da prođe isti proces standardizacije (i svog drugog pred-procesiranja kao obučavajući skup)

Test skup ne sme da učestvuje u procesu unakrsne validacije kad se ona koristi za selekciju atributa i modela  
Unarksna validacija se radi na preostalom delu obučavajućeg skupa