Strojno učenje in podatkovno rudarjenje

Nada Lavrač

Odsek za tehnologije znanja

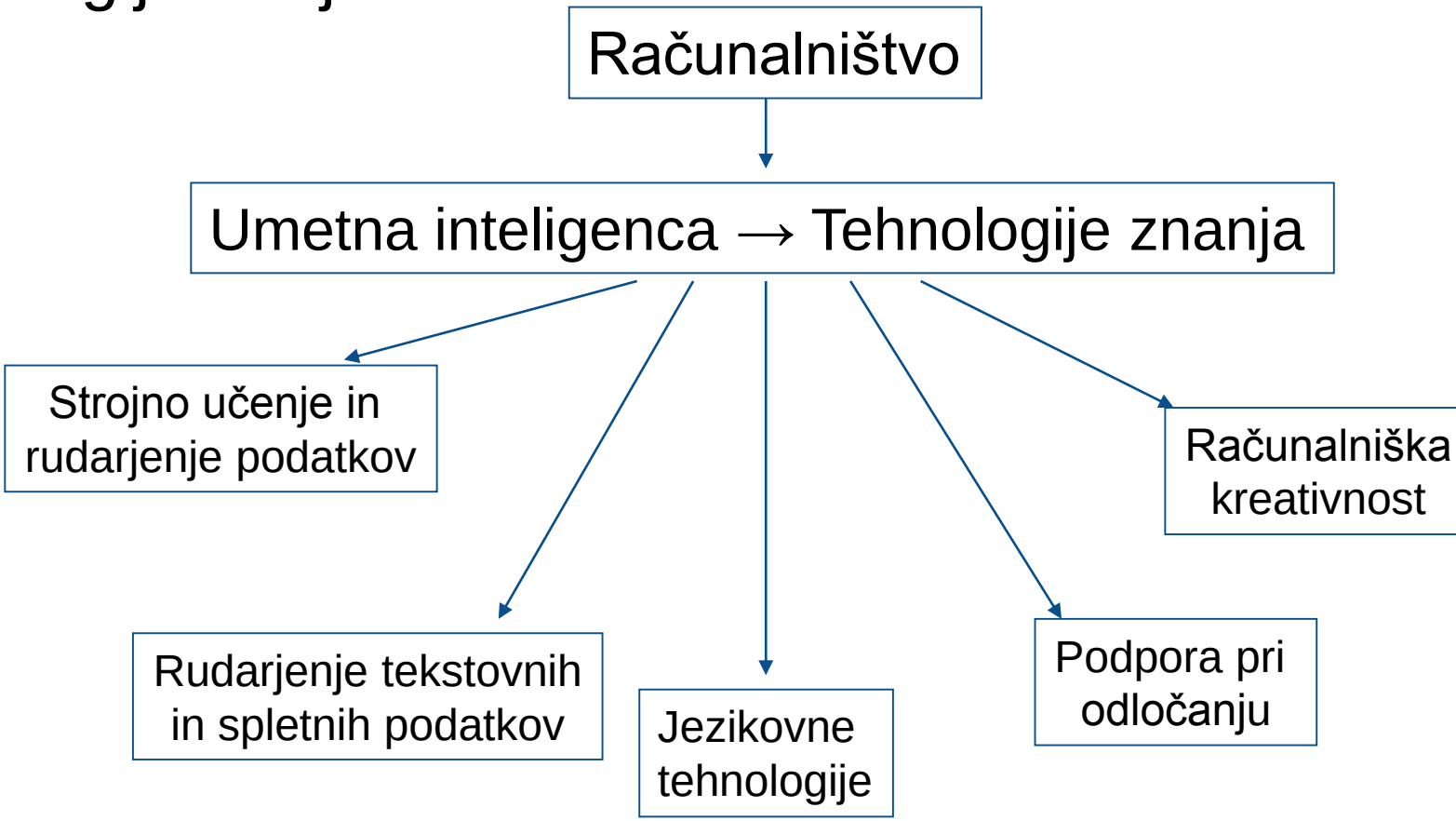
Institut Jožef Stefan

Odsek za tehnologije znanja IJS

- IJS (~ 700 raziskovalcev): naravoslovje in tehnike
- Odsek (~35 raziskovalcev): Raziskave in aplikacije tehnologij znanja

Odsek za tehnologije znanja IJS

- IJS (~ 700 raziskovalcev): naravoslovje in tehnike
- Odsek (~40 raziskovalcev): Raziskave in aplikacije tehnologij znanja



Kdo smo in kaj delamo

Tehnologi znanja IJS

Kdo smo in kaj delamo

- odkrivanje zakonitosti v podatkih
- rudarjenje tekstovnih in spletnih podatkov
- računalniška podpora pri odločanju
- jezikovne tehnologije in računalniško jezikoslovje
- računalniška kreativnost

Vsebina

- **Tehnologije znanja na IJS**
- **Uvod: Klasične tehnike strojnega učenja**
- **Rudarjenje podatkov**
  - Izbrani sistemi in algoritmi druge generacije
  - Izbrane biomedicinske aplikacije
- **Napredne tehnike rudarjenja podatkov**
  - Relacijsko in semantično podatkovno rudarjenje
  - Izbrane biomedicinske aplikacije
- **Tekoče delo in zaključki**

Klasične tehnike strojnega učenja

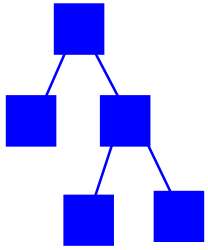
Oseba	Starost	Dioptrija	Astigmat.	Solzenje	Leče
O1	mlad	kratko	ne	zmanjšano	NE
O2	mlad	kratko	ne	normalno	MEHKE
O3	mlad	kratko	da	zmanjšano	NE
O4	mlad	kratko	da	normalno	TRDE
O5	mlad	daleko	ne	zmanjšano	NE
O6-O13	…	…	…	…	…
O14	pr_st_dal	daleko	ne	normalno	MEHKE
O15	pr_st_dal	daleko	da	zmanjšano	NE
O16	pr_st_dal	daleko	da	normalno	NE
O17	st_daleko	kratko	ne	zmanjšano	NE
O18	st_daleko	kratko	ne	normalno	NE
O19-O23	…	…	…	…	…
O24	st_daleko	daleko	da	normalno	NE

podatki

odkrivanje zakonitosti v podatkih

Strojno učenje

Rudarjenje podatkov



model, vzorci, …

Vhod: tabela podatkov, relacijska podatkovna baza, dokumenti, spletne strani

Izhod: klasifikacijski model, posamezni vzorci

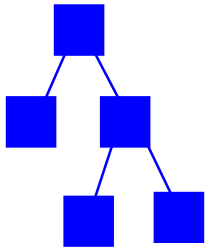
Klasične tehnike strojnega učenja

Oseba	Starost	Dioptrija	Astigmat.	Solzenje	Leče
O1	mlad	kratko	ne	zmanjšano	NE
O2	mlad	kratko	ne	normalno	MEHKE
O3	mlad	kratko	da	zmanjšano	NE
O4	mlad	kratko	da	normalno	TRDE
O5	mlad	daleko	ne	zmanjšano	NE
O6-O13	…	…	…	…	…
O14	pr_st_dal	daleko	ne	normalno	MEHKE
O15	pr_st_dal	daleko	da	zmanjšano	NE
O16	pr_st_dal	daleko	da	normalno	NE
O17	st_daleko	kratko	ne	zmanjšano	NE
O18	st_daleko	kratko	ne	normalno	NE
O19-O23	…	…	…	…	…
O24	st_daleko	daleko	da	normalno	NE

podatki

odkrivanje zakonitosti v podatkih

Strojno učenje
Rudarjenje podatkov

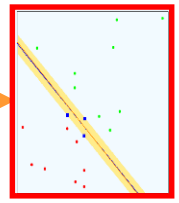


model, vzorci, …

Vhod: tabela podatkov, relacijska podatkovna baza, dokumenti, spletne strani

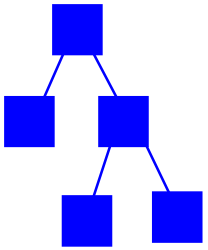
Izhod: klasifikacijski model, posamezni vzorci

nov neklasifician primer



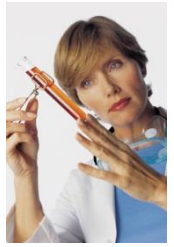
klasificiran primer

črna škatla
ni razlage



simbolični model
simbolični vzorci

razlaga



Klasične tehnike strojnega učenja

1. PODATKI

starost	dioptrija	astigmatizem	solzenje	kontaktne leče
17	kratkovidnost	ne	zmanjšano	ne
23	kratkovidnost	da	normalno	trde
22	daljnovidnost	ne	normalno	mehke
27	kratkovidnost	ne	normalno	mehke
19	kratkovidnost	da	zmanjšano	ne
23	kratkovidnost	da	normalno	trde
20	daljnovidnost	ne	zmanjšano	ne
25	daljnovidnost	ne	normalno	mehke
28	daljnovidnost	da	zmanjšano	ne
21	daljnovidnost	da	normalno	trde
35	kratkovidnost	ne	zmanjšano	ne
36	kratkovidnost	ne	normalno	mehke
40	kratkovidnost	da	zmanjšano	ne
...	kratkovidnost	da	normalno	trde
...	daljnovidnost	ne	zmanjšano	ne
…	daljnovidnost	da	zmanjšano	ne
…	daljnovidnost	da	normalno	ne
...	kratkovidnost	ne	zmanjšano	ne
...	kratkovidnost	ne	normalno	ne
...	kratkoviden	da	zmanjšano	ne
...	daljnovidnost	ne	zmanjšano	ne
...	daljnovidnost	ne	normalno	mehke
...	daljnovidnost	da	zmanjšano	ne
...i	daljnovidnost	da	normalno	ne

Klasične tehnike strojnega učenja

starost	dioptrija	astigma-tizem	solzenje	kontaktne leče
17	kratkovidnost	ne	zmanjšano	ne
23	kratkovidnost	da	normalno	trde
22	daljnovidnost	ne	normalno	mehke
27	kratkovidnost	ne	normalno	mehke
19	kratkovidnost	da	zmanjšano	ne
23	kratkovidnost	da	normalno	trde
20	daljnovidnost	ne	zmanjšano	ne
25	daljnovidnost	ne	normalno	mehke
28	daljnovidnost	da	zmanjšano	ne
21	daljnovidnost	da	normalno	trde
35	kratkovidnost	ne	zmanjšano	ne
36	kratkovidnost	ne	normalno	mehke
40	kratkovidnost	da	zmanjšano	ne
...	kratkovidnost	da	normalno	trde
...	daljnovidnost	ne	zmanjšano	ne
...	daljnovidnost	da	zmanjšano	ne
...	daljnovidnost	da	normalno	ne
...	kratkovidnost	ne	zmanjšano	ne
...	kratkovidnost	ne	normalno	ne
...	kratkoviden	da	zmanjšano	ne
...	daljnovidnost	ne	zmanjšano	ne

2. VZOREC

PRAVILO

ČE

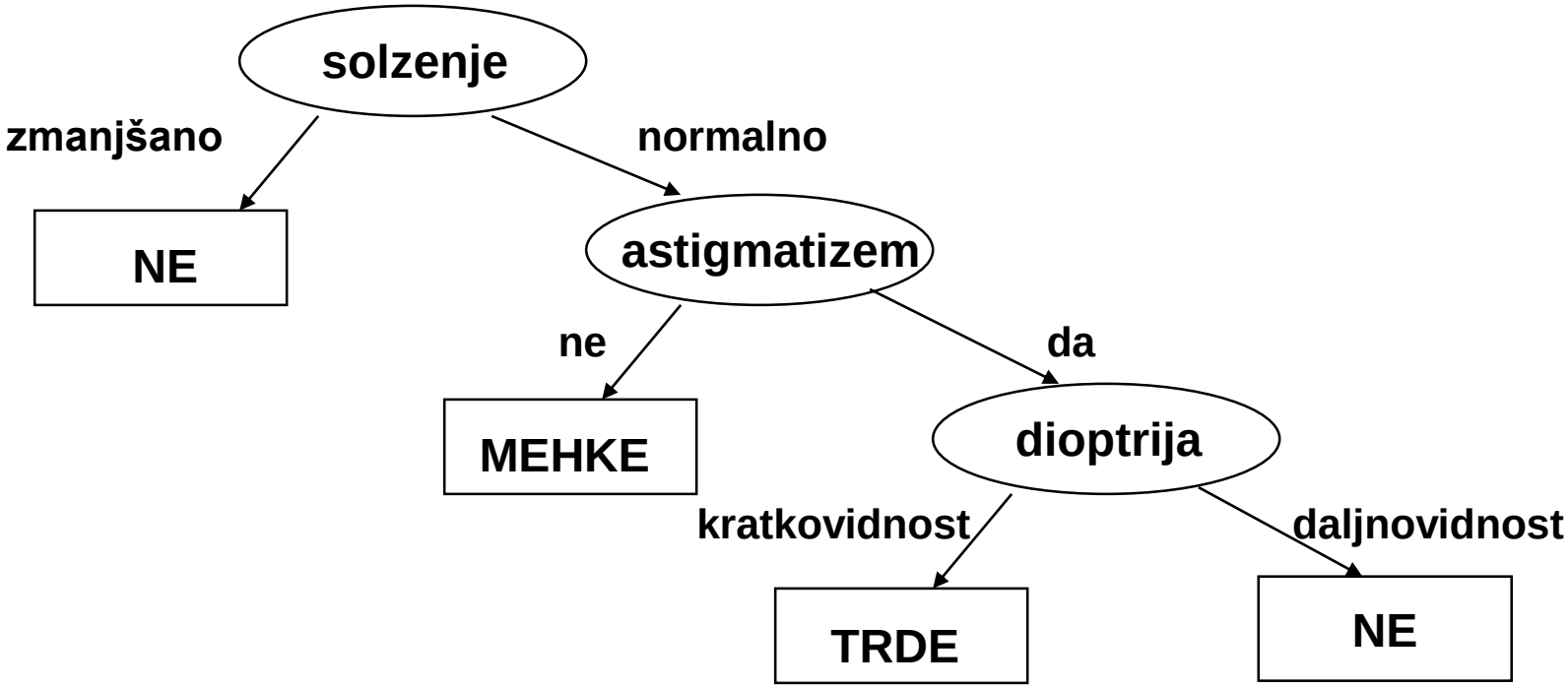
solzenje zmanjšano

POTEM

kontaktne leče=ne

Klasične tehnike strojnega učenja … Odkrivanje znanja v podatkih

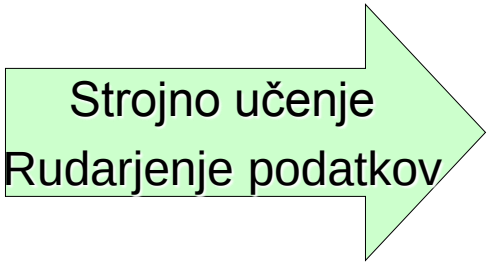
3. VZOREC / MODEL, ki ga potrdi človek = ZNANJE



4. UPORABA ZNANJA kot dodatno ekspertno mnenje

Klasične tehnike strojnega učenja

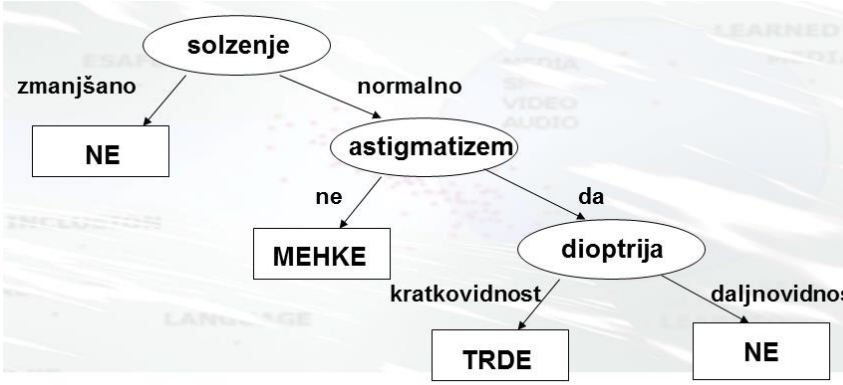
Oseba	Starost	Dioptrija	Astigmat.	Solzenje	Leče
O1	mlad	kratko	ne	zmanjšano	NE
O2	mlad	kratko	ne	normalno	MEHKE
O3	mlad	kratko	da	zmanjšano	NE
O4	mlad	kratko	da	normalno	TRDE
O5	mlad	daleko	ne	zmanjšano	NE
O6-O13	…	…	…	…	…
O14	pr_st_dal	daleko	ne	normalno	MEHKE
O15	pr_st_dal	daleko	da	zmanjšano	NE
O16	pr_st_dal	daleko	da	normalno	NE
O17	st_daleko	kratko	ne	zmanjšano	NE
O18	st_daleko	kratko	ne	normalno	NE
O19-O23	…	…	…	…	…
O24	st_daleko	daleko	da	normalno	NE



Klasične tehnike strojnega učenja

Oseba	Starost	Dioptrija	Astigmat.	Solzenje	Leče
O1	mlad	kratko	ne	zmanjšano	NE
O2	mlad	kratko	ne	normalno	MEHKE
O3	mlad	kratko	da	zmanjšano	NE
O4	mlad	kratko	da	normalno	TRDE
O5	mlad	daleko	ne	zmanjšano	NE
O6-O13	…	…	…	…	…
O14	pr_st_dal	daleko	ne	normalno	MEHKE
O15	pr_st_dal	daleko	da	zmanjšano	NE
O16	pr_st_dal	daleko	da	normalno	NE
O17	st_daleko	kratko	ne	zmanjšano	NE
O18	st_daleko	kratko	ne	normalno	NE
O19-O23	…	…	…	…	…
O24	st_daleko	daleko	da	normalno	NE

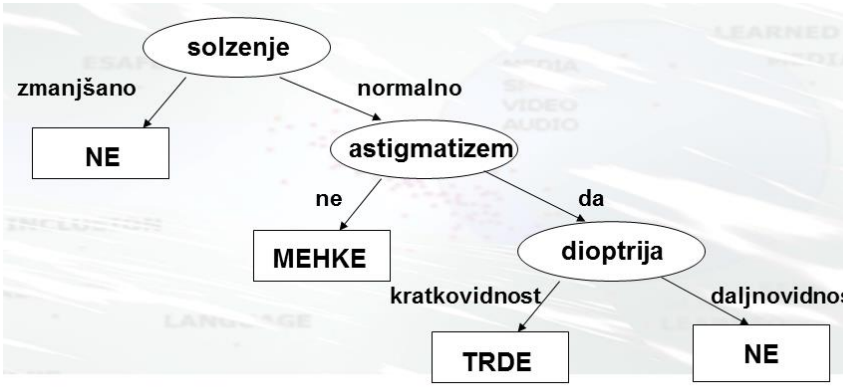
Strojno učenje
Rudarjenje podatkov



Klasične tehnike strojnega učenja

Oseba	Starost	Dioptrija	Astigmat.	Solzenje	Leče
O1	mlad	kratko	ne	zmanjšano	NE
O2	mlad	kratko	ne	normalno	MEHKE
O3	mlad	kratko	da	zmanjšano	NE
O4	mlad	kratko	da	normalno	TRDE
O5	mlad	daleko	ne	zmanjšano	NE
O6-O13	…	…	…	…	…
O14	pr_st_dal	daleko	ne	normalno	MEHKE
O15	pr_st_dal	daleko	da	zmanjšano	NE
O16	pr_st_dal	daleko	da	normalno	NE
O17	st_daleko	kratko	ne	zmanjšano	NE
O18	st_daleko	kratko	ne	normalno	NE
O19-O23	…	…	…	…	…
O24	st_daleko	daleko	da	normalno	NE

Strojno učenje
Rudarjenje podatkov

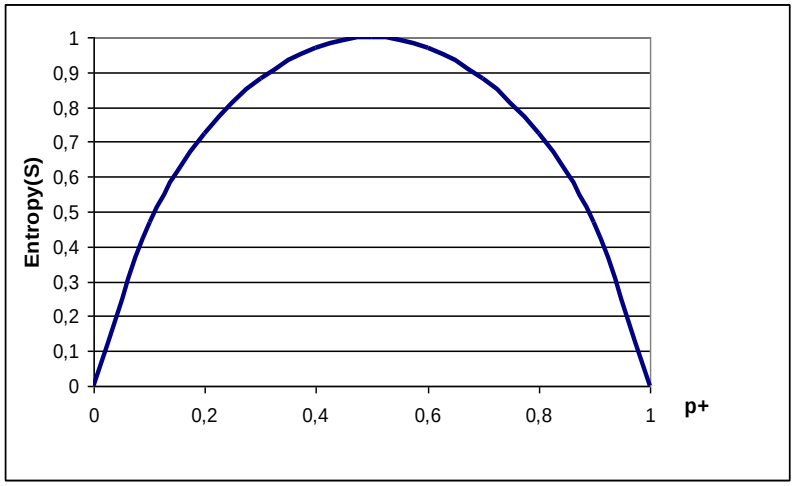


$$Gain(S,A) = E(S) - \sum_{v \in Values(A)} p_v \cdot E(S_v)$$

Hevristika za določanje najinformativnejšega atributa: Gain(S,A)
ocena zmanjšanja entropije množice S zaradi razbitja na podmnožice glede na vrednosti *v* atributa A

Klasične tehnike strojnega učenja
… Ocenjevanje informativnosti atributov

- **Glavna hevristika**: Kateri atribut izbrati kot test v danem vozlišču odločitvenega drevesa ? Atribut, ki je najkoristnejši za čimtočnejšo klasifikacijo primerov.
- Definiramo **statistično oceno informativnosti atributa**, ki meri kako dobro atribut ločuje med primeri, ki pripadajo različnim razredom
- **Informativnost** merimo kot **zmanjšanje entropije učne množice** primerov
- **Entropija** je mera "nečistosti" učne množice: $\mathbf{E(S) = - p_+ \log_2 p_+ - p_- \log_2 p_-}$
- **Najinformativnejši atribut**:
  - uporabi atribut v vozlišču drevesa, razbij S na $S_1, S_2, \ldots, S_v$
  - izberi A, ki maksimizira informacijski prispevek Max Gain(S,A)

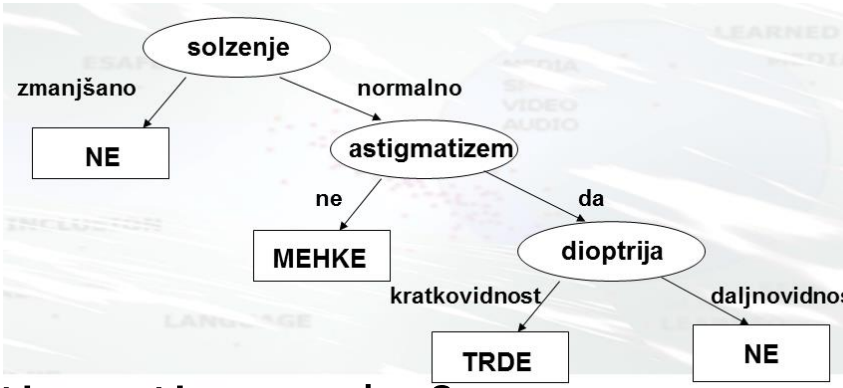


$$Gain(S,A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot E(S_v)$$

Klasične tehnike strojnega učenja

Oseba	Starost	Dioptrija	Astigmat.	Solzenje	Leče
O1	mlad	kratko	ne	zmanjšano	NE
O2	mlad	kratko	ne	normalno	MEHKE
O3	mlad	kratko	da	zmanjšano	NE
O4	mlad	kratko	da	normalno	TRDE
O5	mlad	daleko	ne	zmanjšano	NE
O6-O13	…	…	…	…	…
O14	pr_st_dal	daleko	ne	normalno	MEHKE
O15	pr_st_dal	daleko	da	zmanjšano	NE
O16	pr_st_dal	daleko	da	normalno	NE
O17	st_daleko	kratko	ne	zmanjšano	NE
O18	st_daleko	kratko	ne	normalno	NE
O19-O23	…	…	…	…	…
O24	st_daleko	daleko	da	normalno	NE

Strojno učenje
Rudarjenje podatkov



leče=NE ← solzenje=zmanjšano

leče=NE ← solzenje=normalno & astigmatizem=da & dioptrija=daljnovidnost

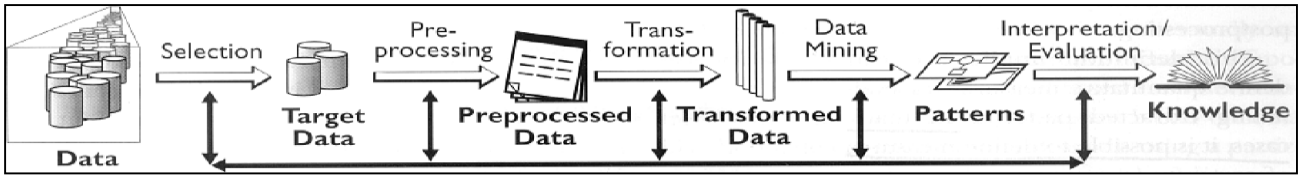
leče=MEHKE ← solzenje=normalno & astigmatizem=ne

leče=TRDE ← solzenje= normalno & astigmatizem=da & dioptrija=kratkovidnost

Druga generacija sistemov

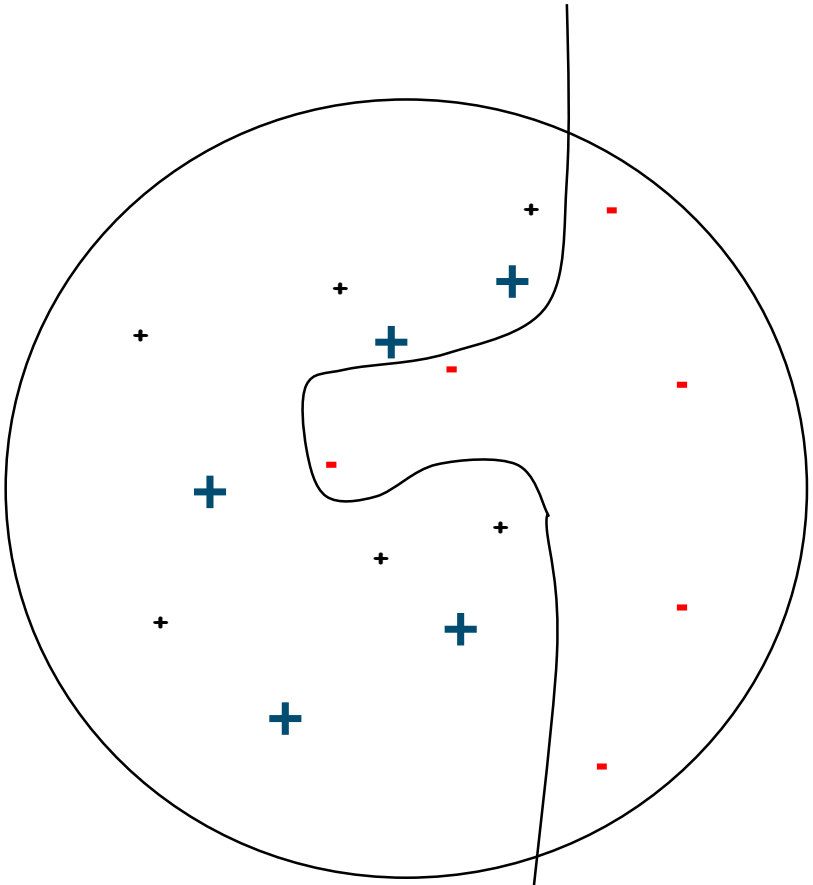
Po letu 1990 …

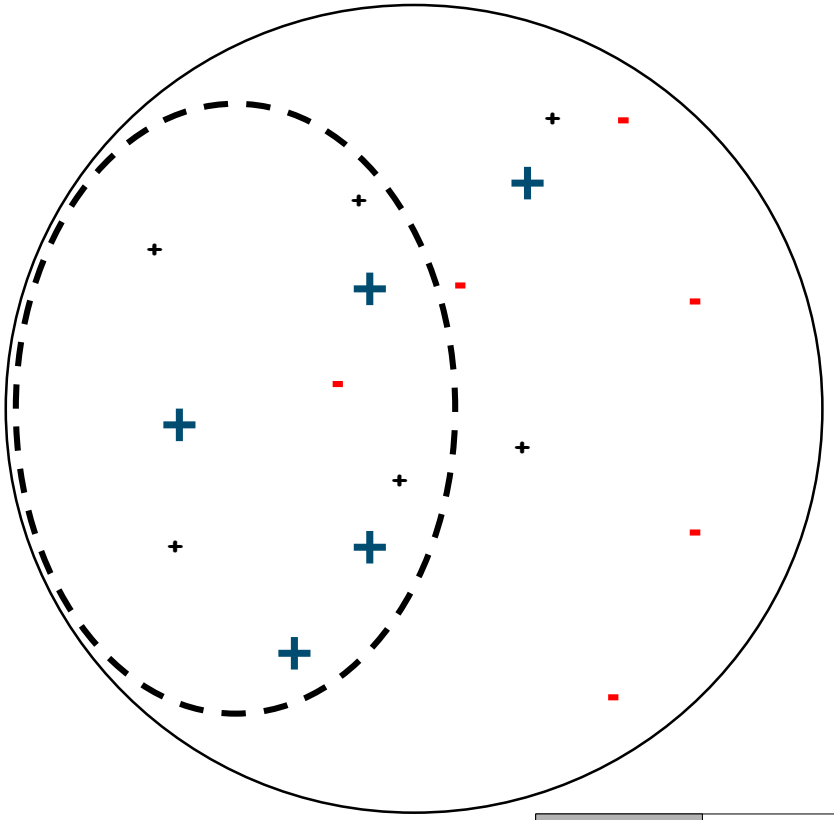
- **Odkrivanje zakonitosti v podatkih** iz velikih podatkovnih baz, opisanih z velikim številom atributov

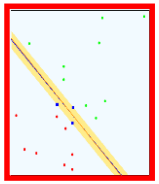


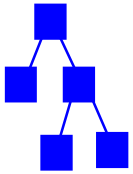
  - Industrijski standard: CRISP-DM (1997)
  - Nove konference o praktičnih aspektih rudarjenja podatkov in odkrivanja zakonitosti v podatkh: KDD, PKDD, …
- **Nove naloge in učinkoviti algoritmi:**
  - **Učenje klasifikacijskim modelov:** Bayesovske mreže, večrelacijsko učenje, statistično relacijsko učenje, metode podpornih vektorjev (SVM), …
  - **Odkrivanje vzorcev in učenje opisnih pravil:** učenje povezovalnih pravil, odkrivanje podskupin, …

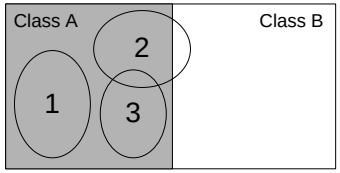
Klasifikacija vs. odkrivanje vzorcev







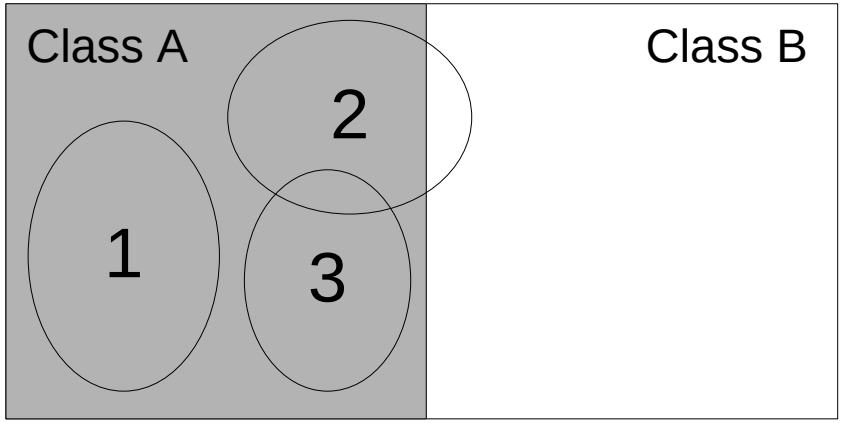




Odkrivanje podskupin

- Učenje opisnih pravil:
  - učenje posameznih pravil, ki opisujejo izbrano lastnost podatkov
- Algoritmi za odkrivanje podskupin se naučijo več neodvisnih pravil, vsako opiše podskupino primerov s podobnimi lastnostmi
  - odkrivanje in opis velikih signifikantnih podmnožic primerov

Oseba	Starost	Dioptrija	Astigmat.	Solzenje	Leče
O1	mlad	kratko	ne	zmanjšano	NE
O2	mlad	kratko	ne	normalno	DA
O3	mlad	kratko	da	zmanjšano	NE
O4	mlad	kratko	da	normalno	DA
O5	mlad	daleko	ne	zmanjšano	NE
O6-O13	…	…	…	…	…
O14	pr_st_dal	daleko	ne	normalno	DA
O15	pr_st_dal	daleko	da	zmanjšano	NE
O16	pr_st_dal	daleko	da	normalno	NE
O17	st_daleko	kratko	ne	zmanjšano	NE
O18	st_daleko	kratko	ne	normalno	NE
O19-O23	…	…	…	…	…
O24	st_daleko	daleko	da	normalno	NE



Primer: Odkrivanje rizičnih skupin pacientov

Vhod: Podatki o pacientih, opisanih z atributi **A** (anamneza), **B** (anamneza & laboratorij), in **C** (an., lab. & ECG)

Naloga: V dani populaciji poišči in opiši rizične skupine za CHD (dovolj velike podskupine z signifikantno različno distribucijo ciljnega razreda glede na distribucijo v celotni populaciji)

Izmed avtomatsko najdenih pravil je ekspert izbral tiste, ki imajo najboljši potencial za seznanjanje pacientov o rizičnosti:

A1: rizičnaSkupina ← moški & pozitivna družinska anamneza & starost > 46

A2: rizičnaSkupina ← ženska & indeks telesne teže > 25 & starost > 63

B1: rizičnaSkupina ← ...

B2: rizičnaSkupina ← ...

C1: rizičnaSkupina ← ...

Primer: Odkrivanje rizičnih skupin pacientov

Rizična skupina A2:

rizičnaSkupina ← ženska & indeks telesne teže > 25 & starost > 63

Podporni faktorji za dodatno karakterizacija skupin – izračunani z uporabo χ^2 statističnega testa signifikance

Pozitivna družinska anamnreza in visok krvni pritisk. Ženske v tej skupini imajo tipično tudi povečano LDL vrednost holesterola in normalno a znižano vrednost HDL vrednost holesterola.

(Gamberger & Lavrač, JAIR 2002)

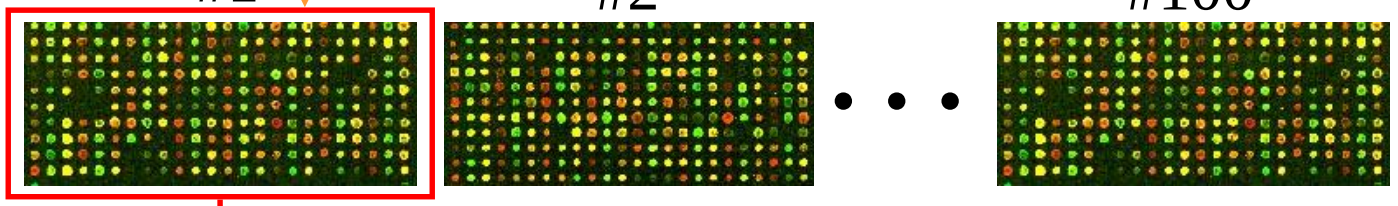
Primer: Analiza DNA mikromrež z odkrivanjem podskupin

- **Funkcijska genomika** je tipična domena za znanstveno odkrivanje zakonitosti, katere cilj je študij genov in njihovih funkcij. Za to domeno je značilno:
  - zelo veliko število atributov (genov) v primerjavi s številom primerov
  - tipična velikost: 7000-16000 atributov, 50-150 primerov

- **Primer problema:** Ločevanje med Akutno limfoblastično levkemijo (ALL, 27 primerov) in akutno mieloidno levkemijo (AML, 11 primerov), z dodatnimi 34 primeri v ločeni testni množici. Vsak primer je opisan z izraženostjo 7129 genov

http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi

Podatki DNA mikromrež: format za rudarjenje pdoatkov

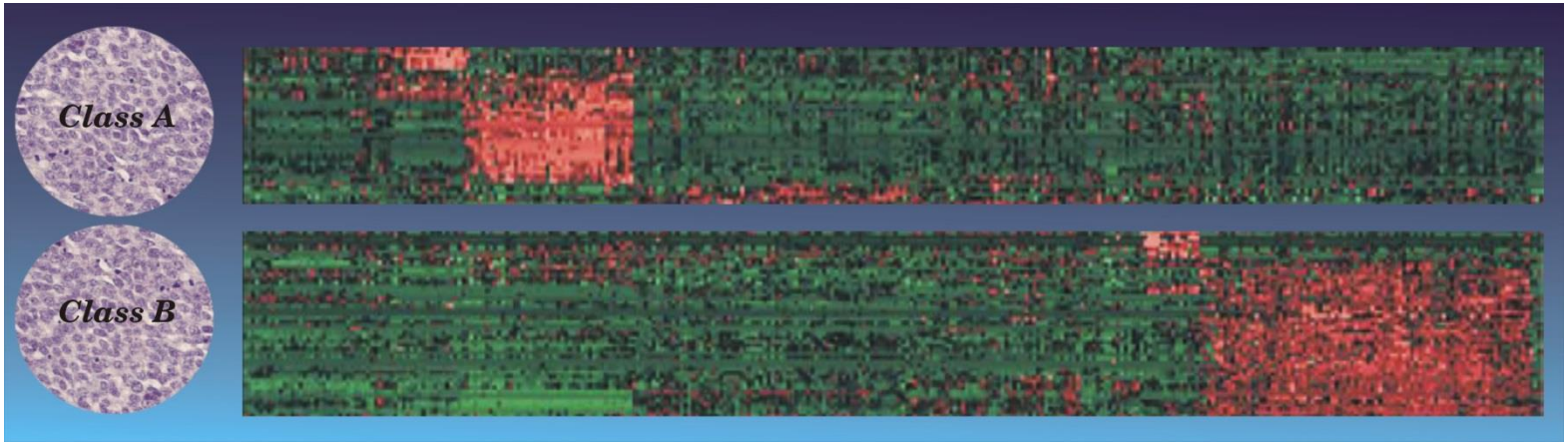


malo primerov

Patient #	Tumor Type	Gene #1	Gene #2	Gene #3	…	Gene #10,000
1	*A*	5.00	1.33	3.45	…	4.22
2	*A*	0.98	0.87	1.04	…	?
3	*B*	0.33	1.40	0.42	…	0.24
…	…	…	…	…	…	…
100	*B*	0.89	0.90	1.00	…	0.66

veliko atributov

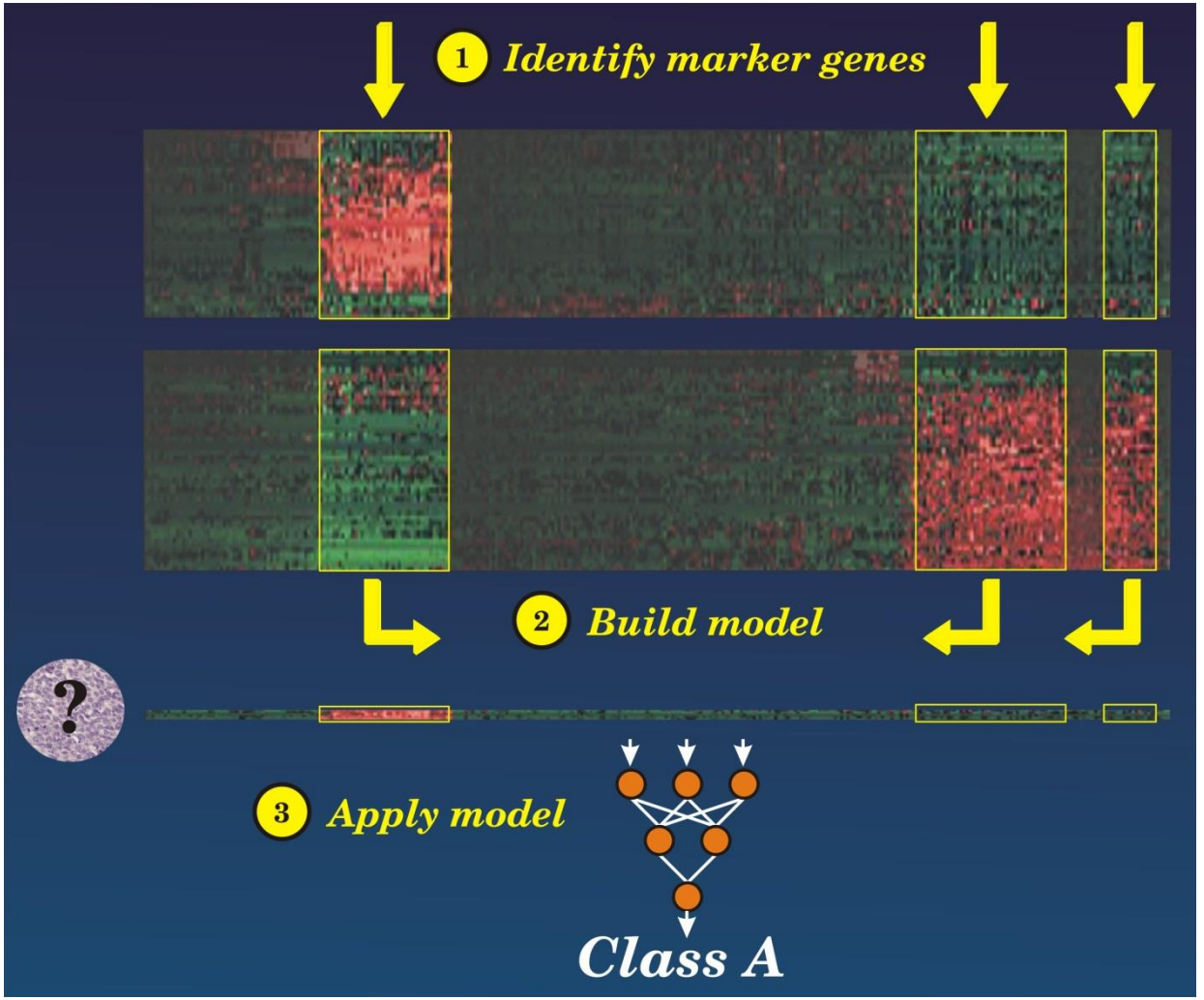
Podatki DNA mikromrež



- Dvo-razredni diagnostični problem ALL vs. AML
- Večrazredni diagnostični problem: 14 tipov raka, skupaj le 144 primerov v učni množici in 54 primerov v testni množici. Vsak primer opisan z izraženostjo 16063 genov.

http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi.

Klasična naloga:
Učenje klasifikacijskih modelov



Analiza DNA mikromrež z metodami odkrivanja podskupin

- Vzorci, odkriti z algoritmom SD za odkrivanje podskupin

AML ← gen_20056 = DIFF. IZRAŽEN &
gen_23984 = NI_DIFF. IZRAŽEN

Levkemija ← KIAA0128 = DIFF. IZRAŽEN &
prostoglandin_d2_synthase = NI_ DIFF. IZRAŽEN

- Možna je biološka **interpretacija** pravil, ki opisujejo podskupine

D. Gamberger, N. Lavrač, F. Železný, J. Tolar,
Journal of Biomedical Informatics, 2004

Analiza DNA mikromrež z metodami odkrivanja podskupin

- Vzorci, odkriti z algoritmom SD za odkrivanje podskupin

AML ← gen_20056 = DIFF. IZRAŽEN &
gen_23984 = NI_DIFF. IZRAŽEN

Levkemija ← KIAA0128 = DIFF. IZRAŽEN &
prostoglandin_d2_synthase = NI_ DIFF. IZRAŽEN

- Možna je biološka **interpretacija** pravil, ki opisujejo podskupine

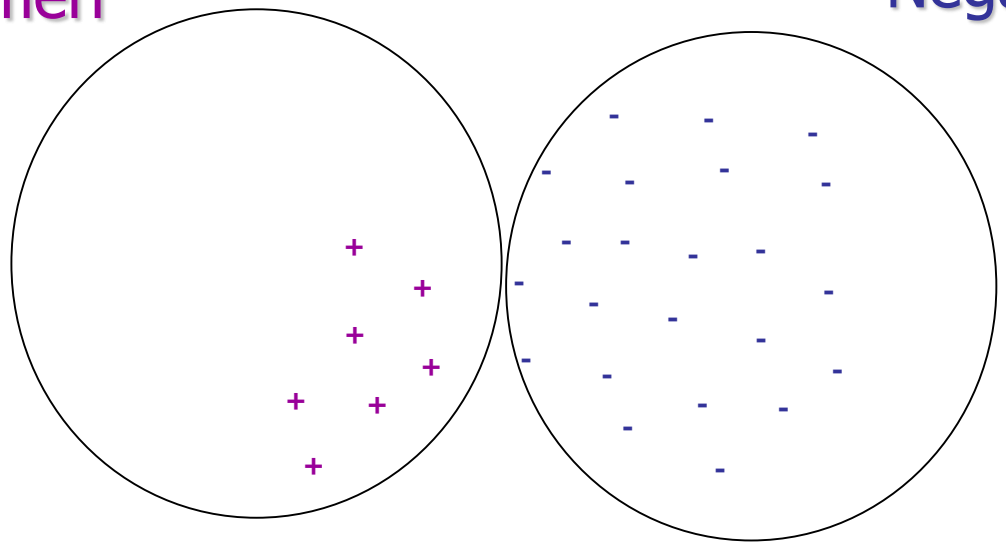
D. Gamberger, N. Lavrač, F. Železný, J. Tolar,
Journal of Biomedical Informatics, 2004

  - nove hevristike za učenje opisnih pravil
  - opisna pravila imajo manjšo klasifikacijsko točnost
  - **trade-off med točnostjo in interpretabilnostjo**

Klasični pokrivni algoritem za učenje klasifikacijskih pravil

Pozitivni primeri

Negativni primeri

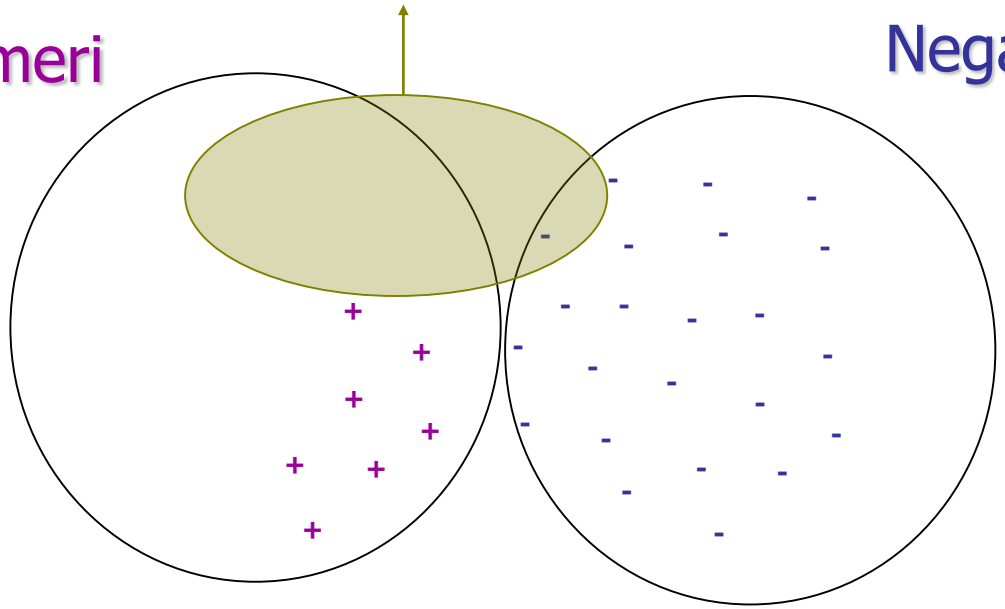


Klasični pokrivni algoritem za učenje klasifikacijskih pravil

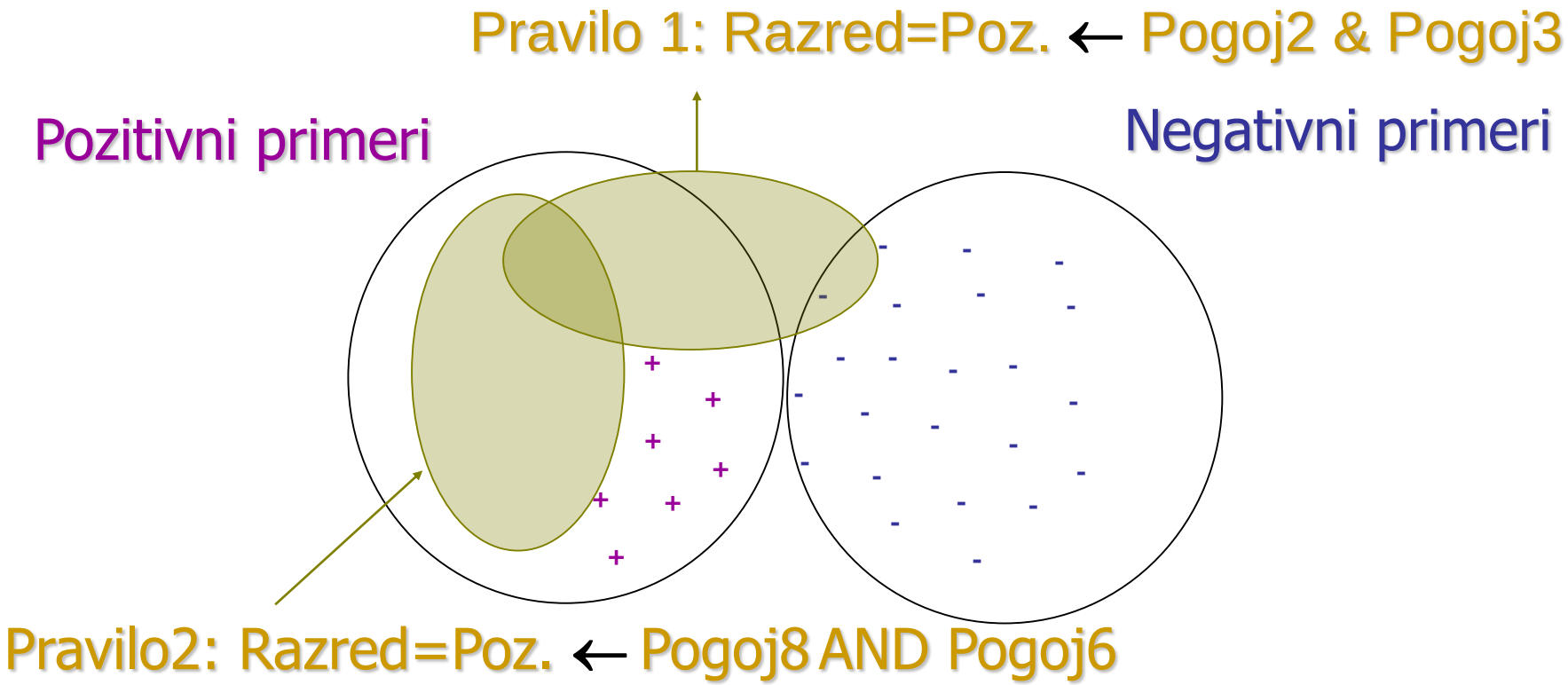
Pravilo 1: Razred=Poz. ← Pogoj2 & Pogoj3

Pozitivni primeri

Negativni primeri

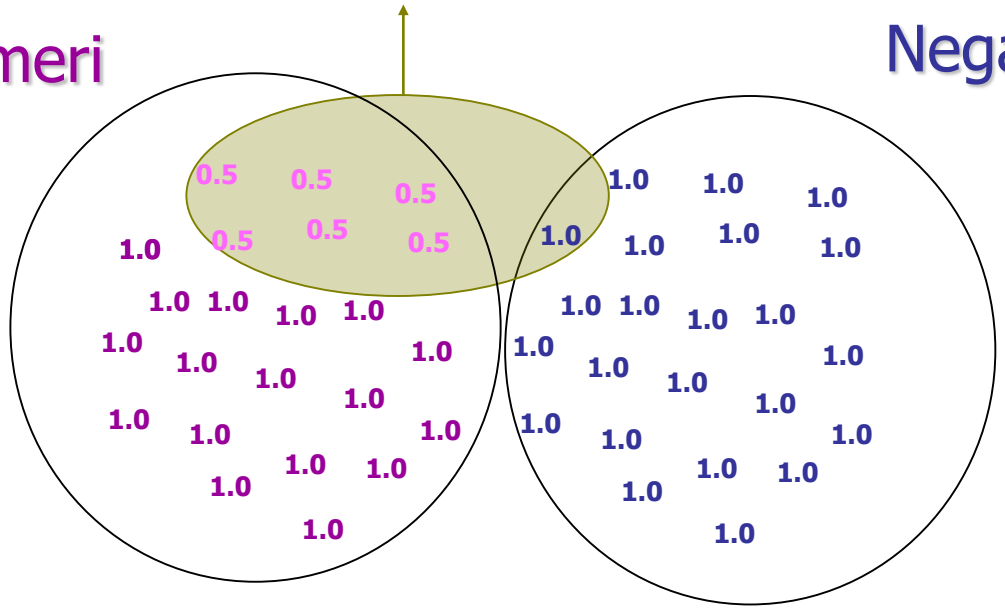


Klasični pokrivni algoritem za učenje klasifikacijskih pravil



Novi uteženi pokrivni algoritem in nove hevristike za učenje opisnih pravil

Pravilo 1: Razred=Poz. ← Pogoj2 & Pogoj3



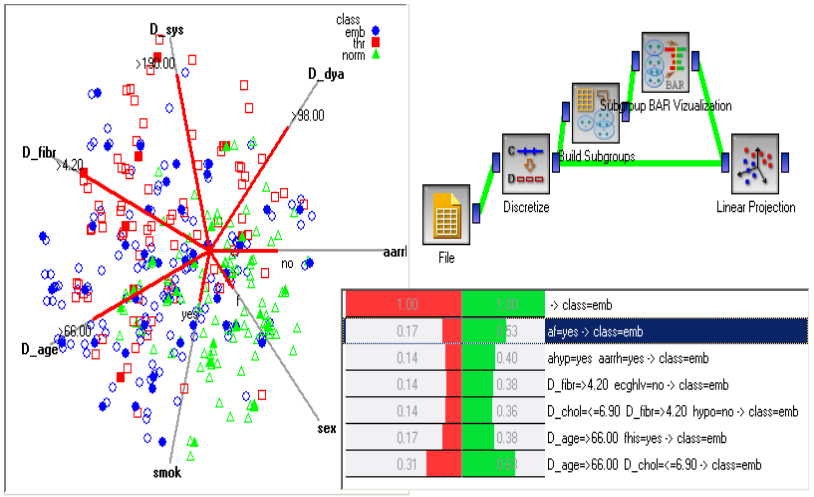
Hevristika - mera kvalitete pravila v SD: q(Razred ← Pogoji) = TP/(FP+g)

Hevristika v CN2-SD: q(Razred ← Pogoji) = ~
p(Pogoji) x (p(Razred | Pogoji) – p(Razred))
~ Pokritost x (Preciznost – Default)

Pokritost = delež pokritih primerov, **Preciznost** = delež pravilno pokritih pozitivnih primerov

Napredna orodja za rudarjenje podatkov v platformah Orange (FRI) in Orange4WS (IJS)

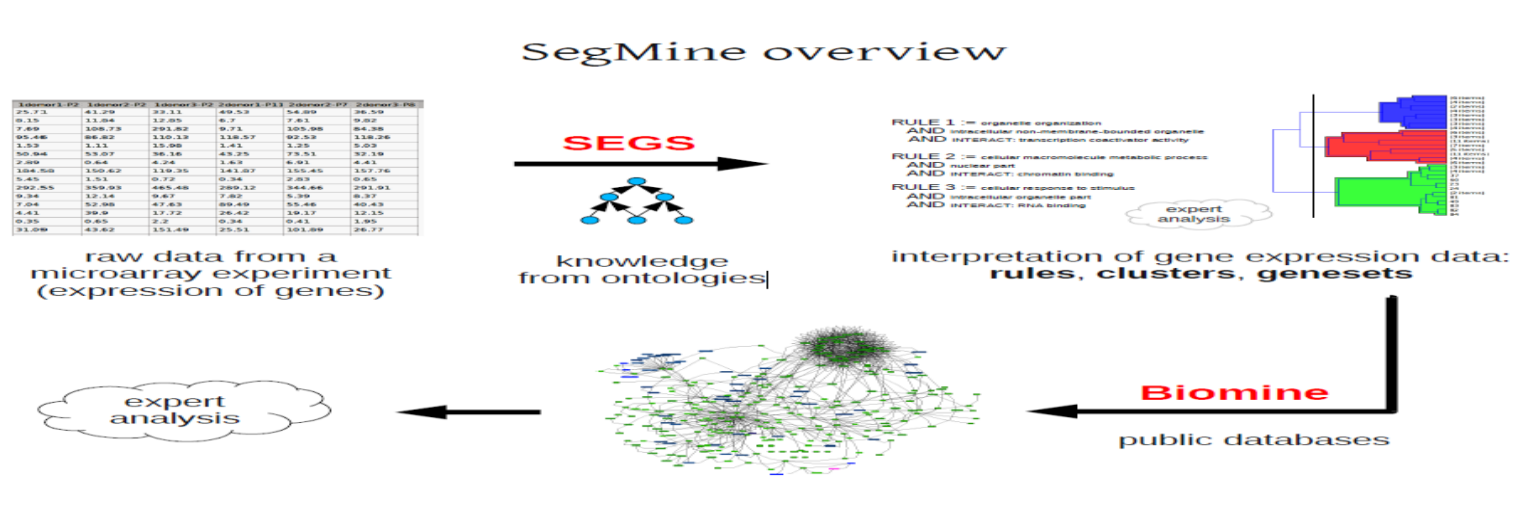
- **Orange**
  - klasifikacijski algoritmi
  - algoritmi za odkrivanje podskupin
  - vizualizacija podatkov
  - delotoki podatkovnega rudarjenja

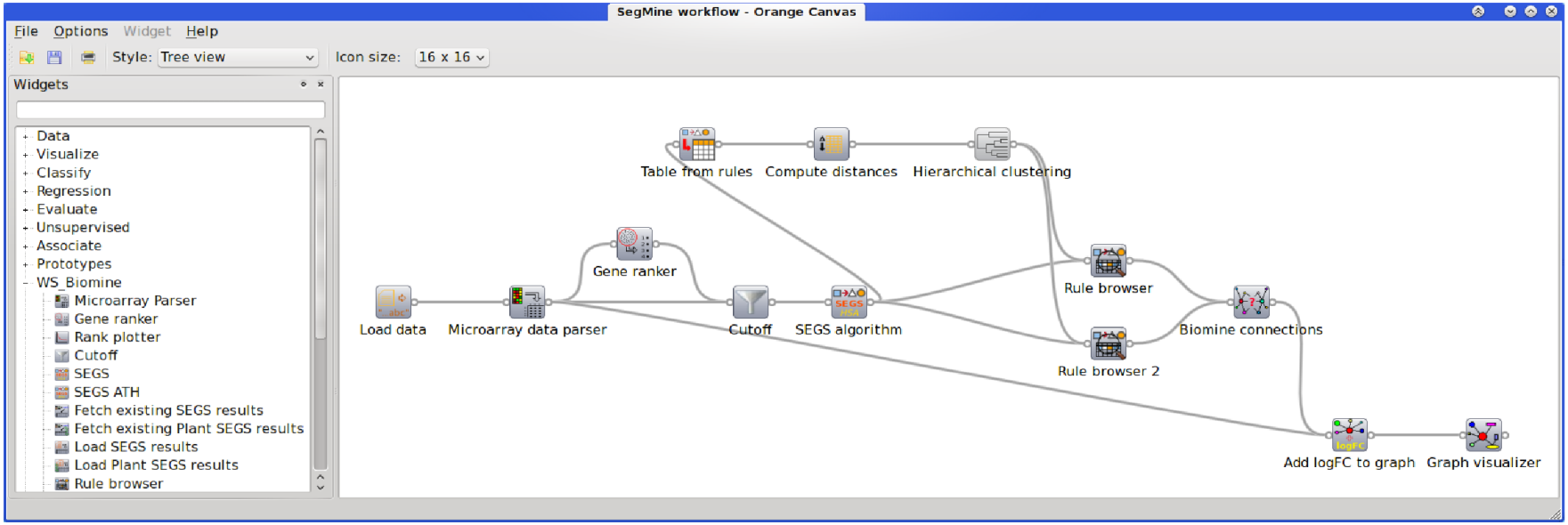


- **Algoritmi za odkrivanje podskupin v Orange in Orange4WS**

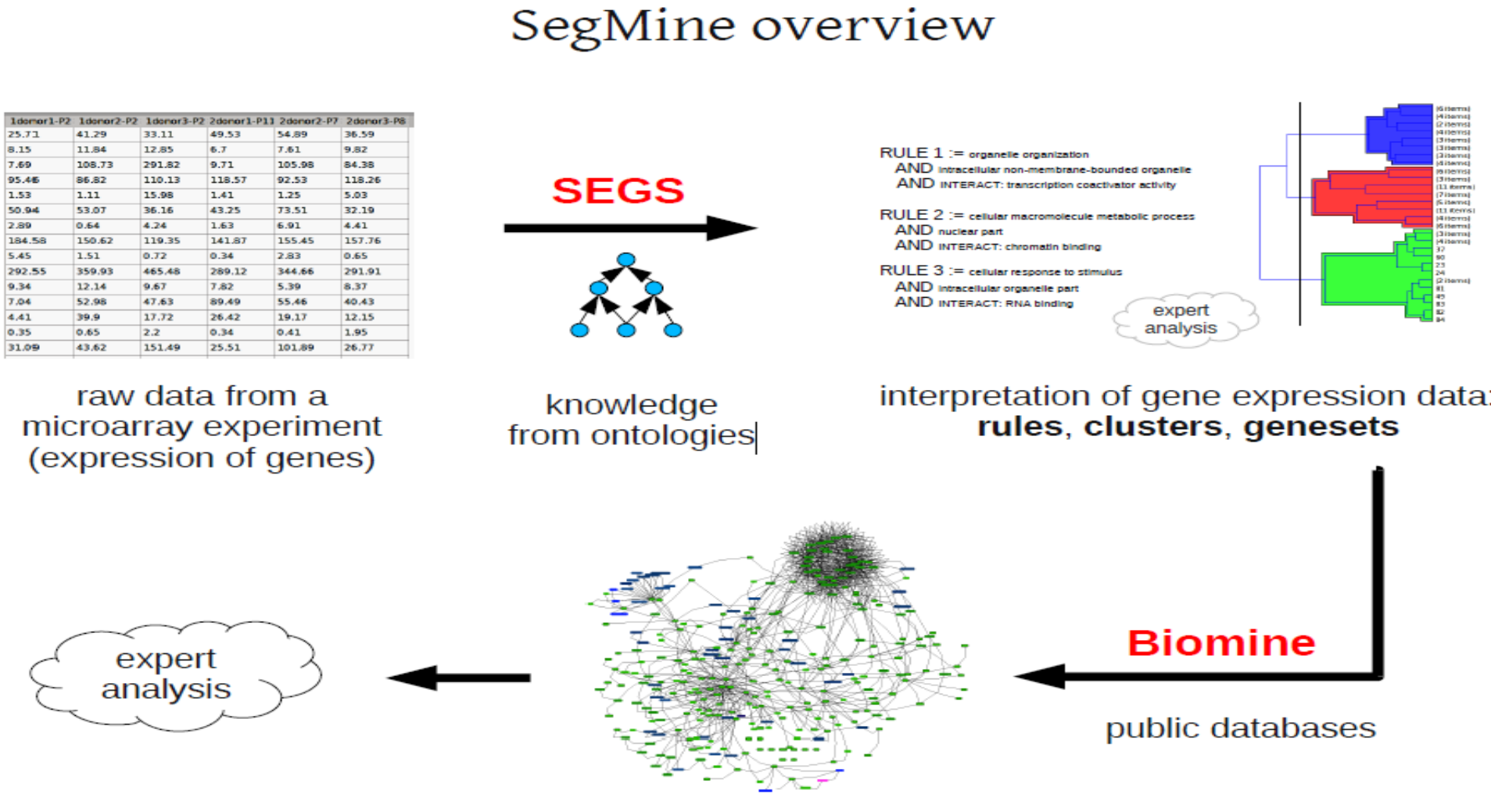
SD (Gamberger & Lavrač, JAIR 2002)
CN2-SD (Lavrač et al., JMLR 2004)
APRIORI-SD (Kavšek & Lavrač, AAI 2006)

Metodologija SegMine implementirana v platformi Orange4WS (BMC Bioinformatics 2011)

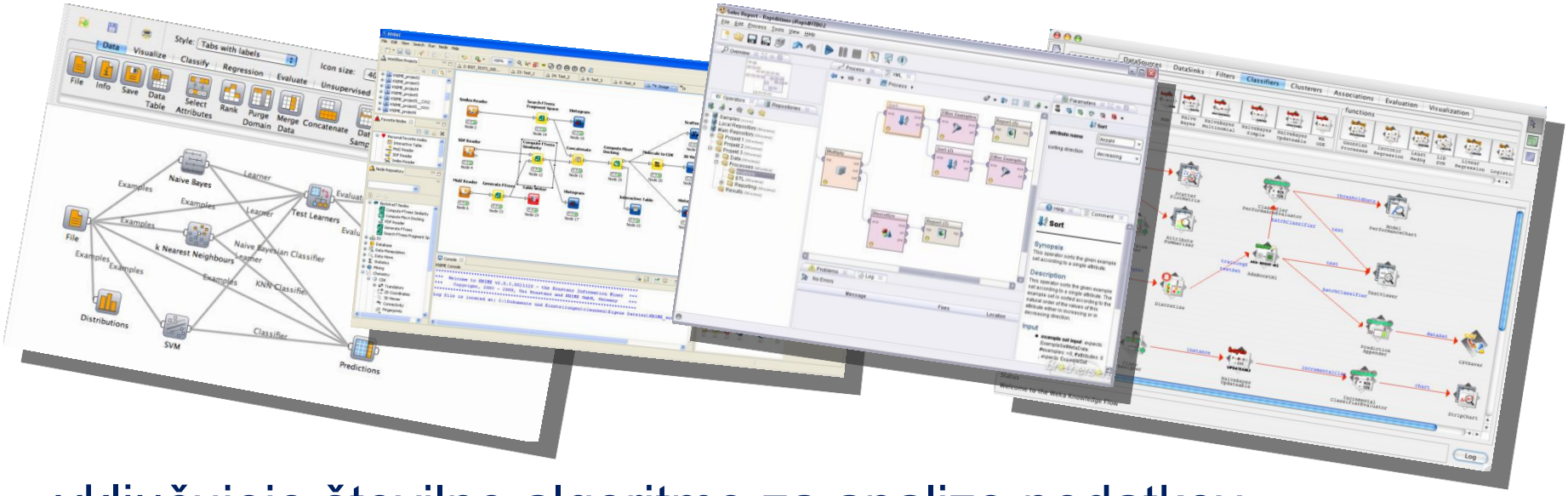




Primer kompleksne metodologije podatkovnega rudarjenja: SegMine



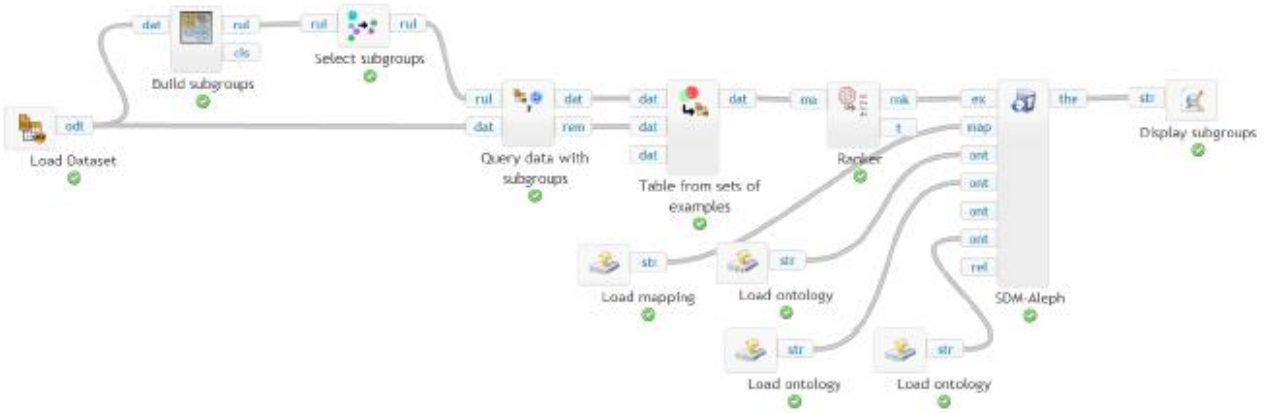
Platforme WEKA, KNIME, RapidMiner, Orange (FRI) in Orange4WS (IJS)



- vključujejo številne algoritme za analizo podatkov
- omogočajo analizo in vizualizacijo podatkov
- omogočajo gradnjo delotokov
- **Orange4WS omogoča tudi vključevanje spletnih servisov**

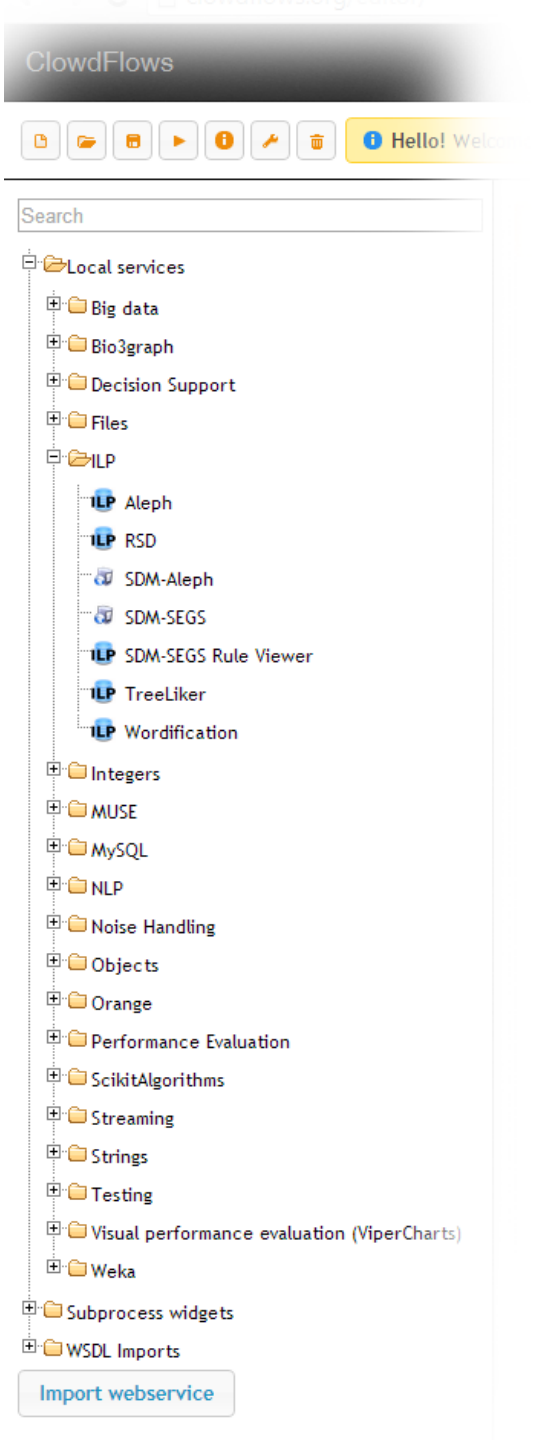
IJS platforma za rudarjenje podatkov ClowdFlows

- Platforma tretje generacije za kreiranje in izvajanje kompleksnih postopkov rudarjenja podatkov
  - algoritmi so spletni servisi (v “oblaku”)
  - ni potrebna instalacija platforme
  - delotoki so dostopni vsakomur kar iz brskalnika s preprostim klikom na spletni naslov: npr. **http://clowdflows.org/workflow/910/**



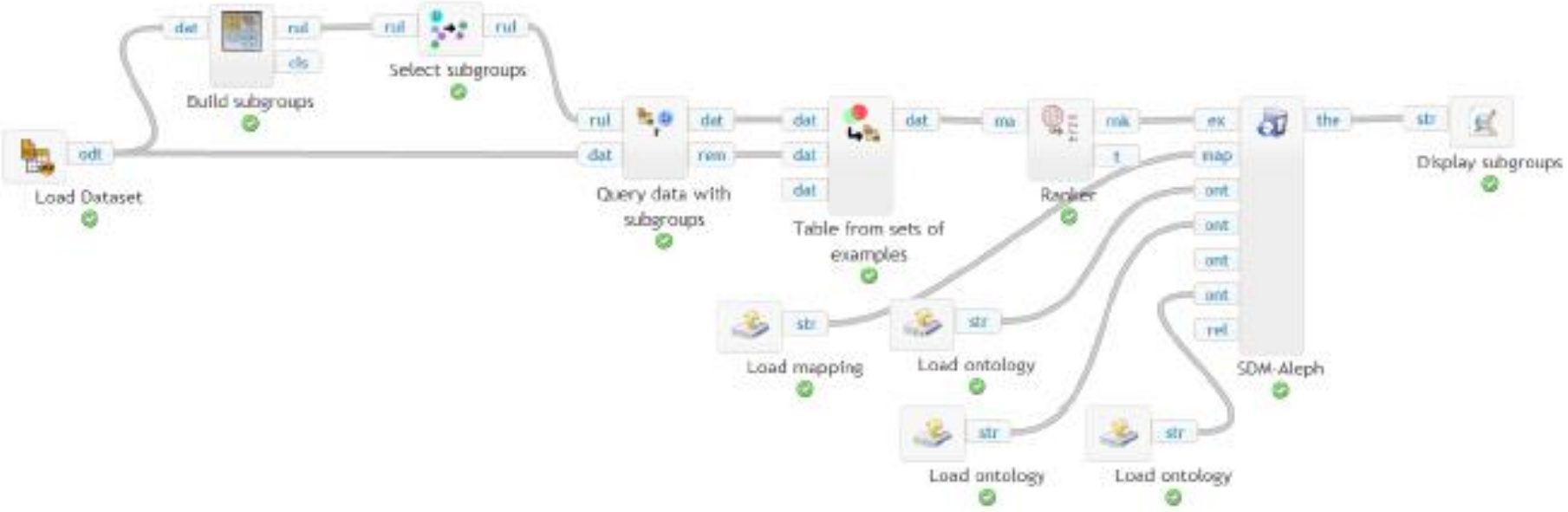
Platforma ClowdFlows

- **Velik repozitorij algoritmov**
  - analiza relacijskih podatkovnih baz
  - vsu algoritmi iz platforme Orange
  - WEKA algoritmi kot spletni servisi
  - vizualizacija podatkov in rezultatov analiz
  - analiza tekstovnih podatkov
  - analiza socialnih omrežij
  - analiza velikih tokov podatkov
- **Velik repozitorij delotokov**
  - omogča dostop do naše tehnološke dediščine



Primer: Semantično rudarjenje podatkov

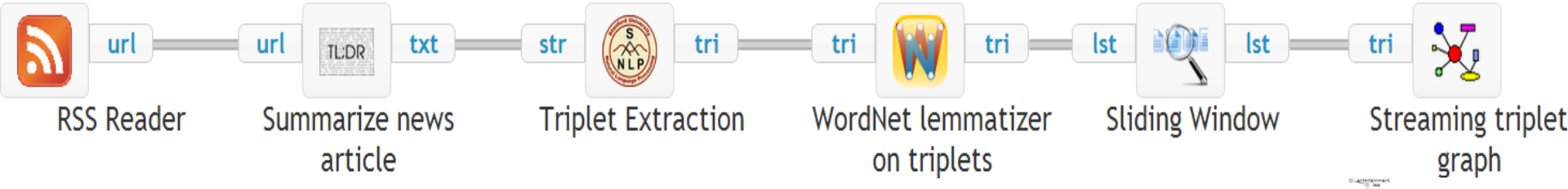
- Odkrivanje zanimivih podskupin v podatkih ter njihova razlaga s pomočjo biološkega predznanja - ontologij



- http://clowdflows.org/workflow/910/

Primer: “Big Data”

- Analiza velikih količin podatkov v realnem času. Primer: izgradnja semantičnega grafa iz tokov spletnih novic. http://clowdflows.org/workflow/1729/.

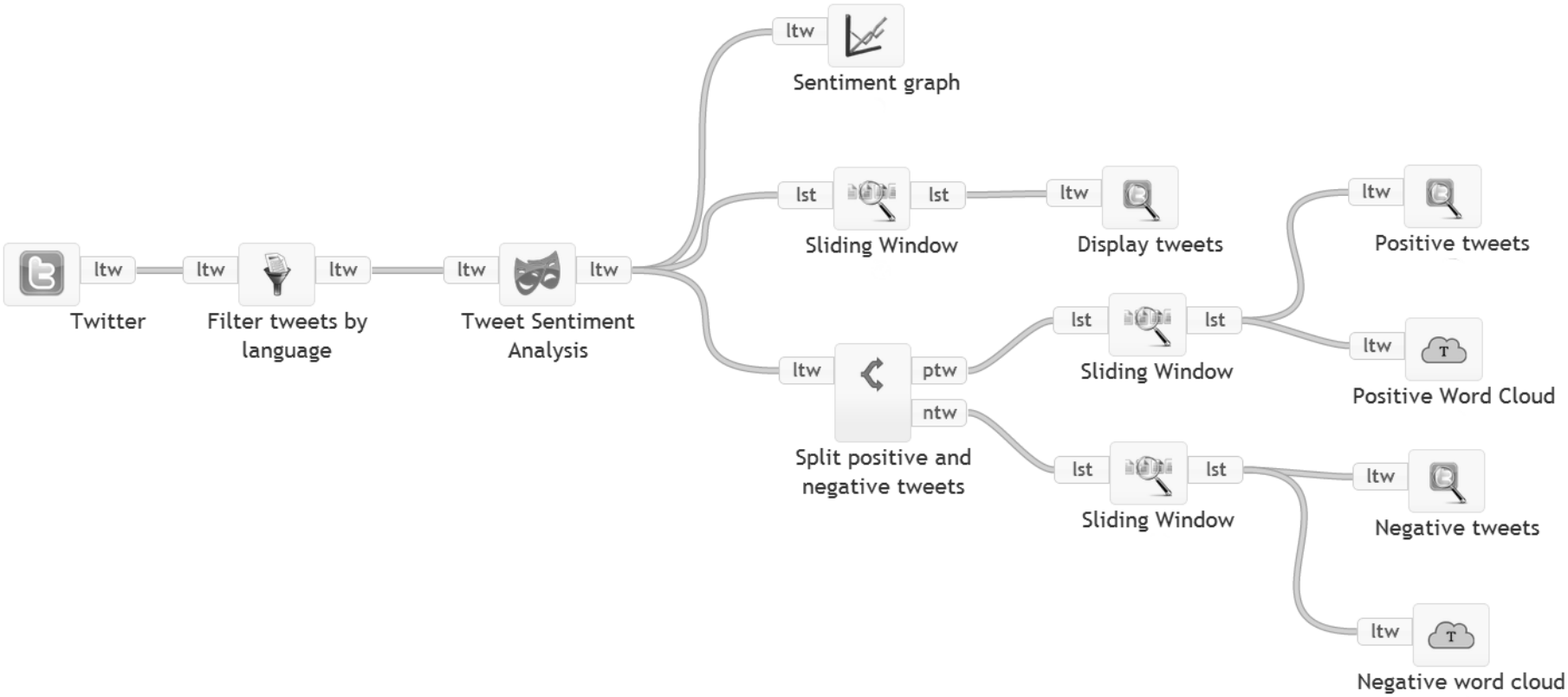


- Primer: monitoring novic z vizualizacijo grafa zgrajenega iz športnih novic (vir: CNN RSS feeds)

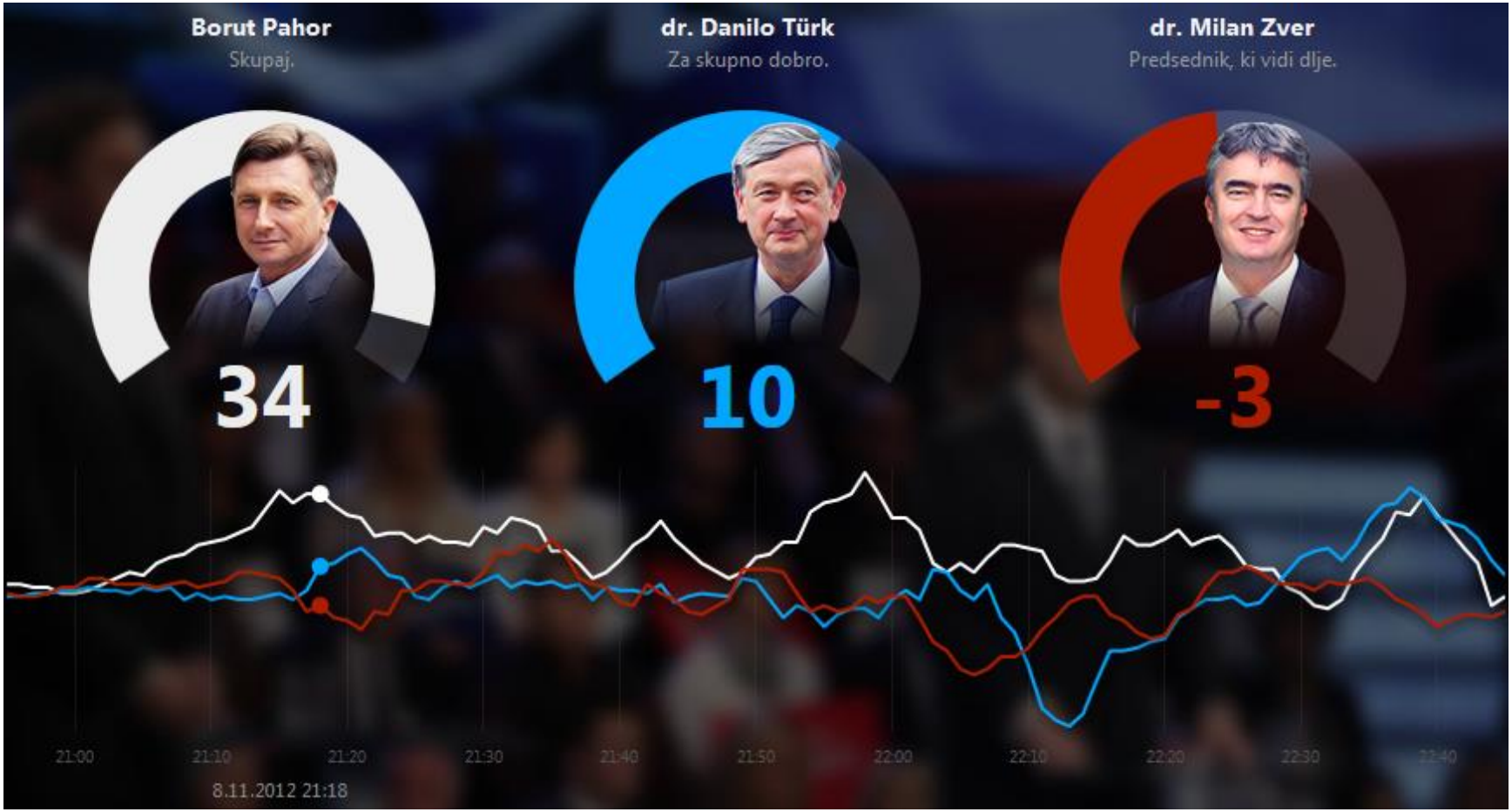
http://clowdflows.org/streams/data/31/15524/.

Primer: “Big Data”

- Analiza pozitivnega/negativnega sentimenta v tvitih v realnem času: http://clowdflows.org/workflow/1041/.

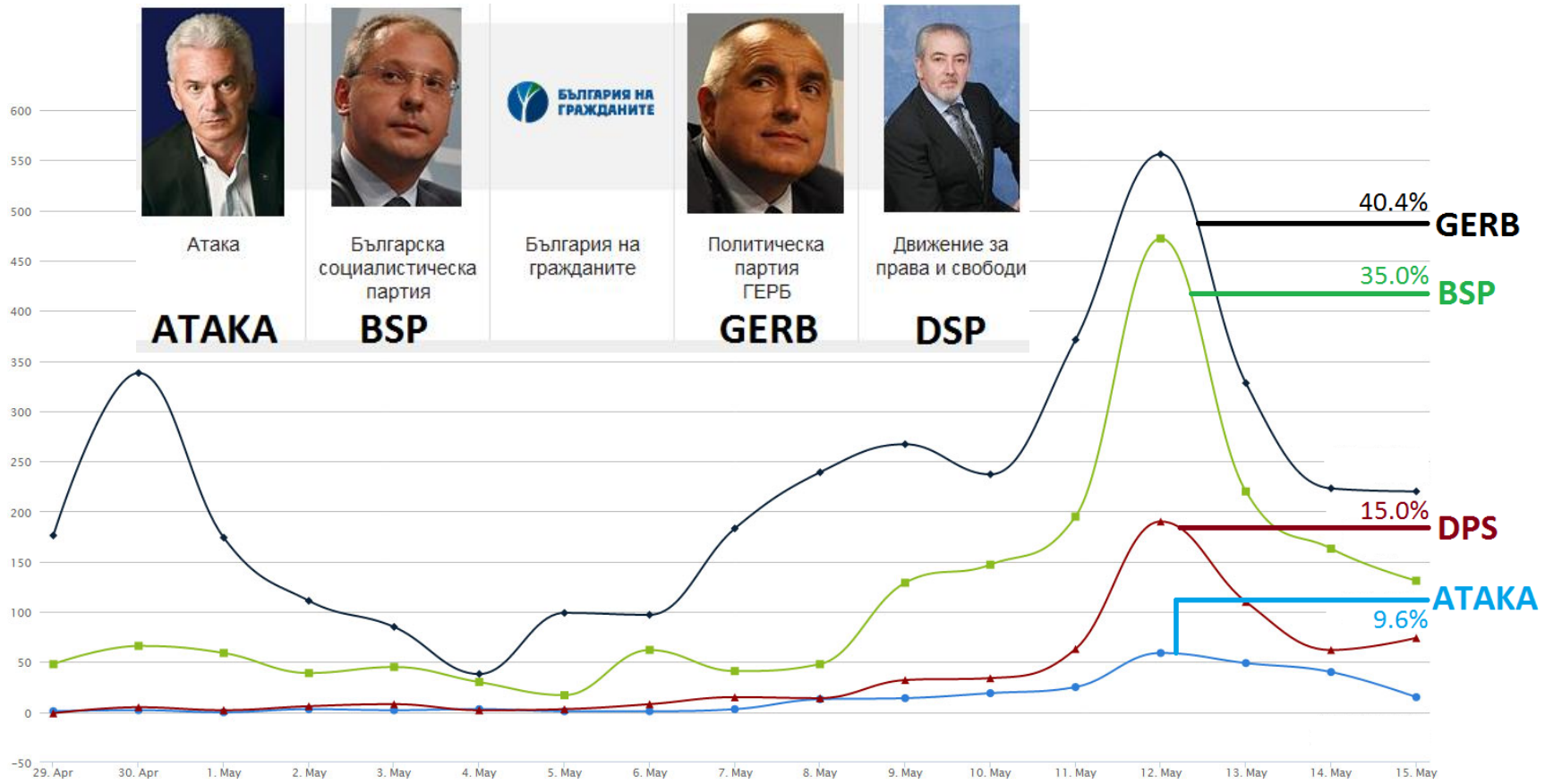


Analiza sentimenta v tvitih: Predsedniške volitve 2012



GamaSystem in IJS, Pop TV, 2012

Parlamentarne volitve v Bolgariji



GamaSystem in IJS, 2013

Vsebina predavanja

- **odkrivanje zakonitosti in anomalij v podatkih**
- **rudarjenje tekstovnih in spletnih podatkov**
- računalniška podpora pri odločanju
- jezikovne tehnologije in računalniško jezikoslovje
- **računalniška kreativnost**

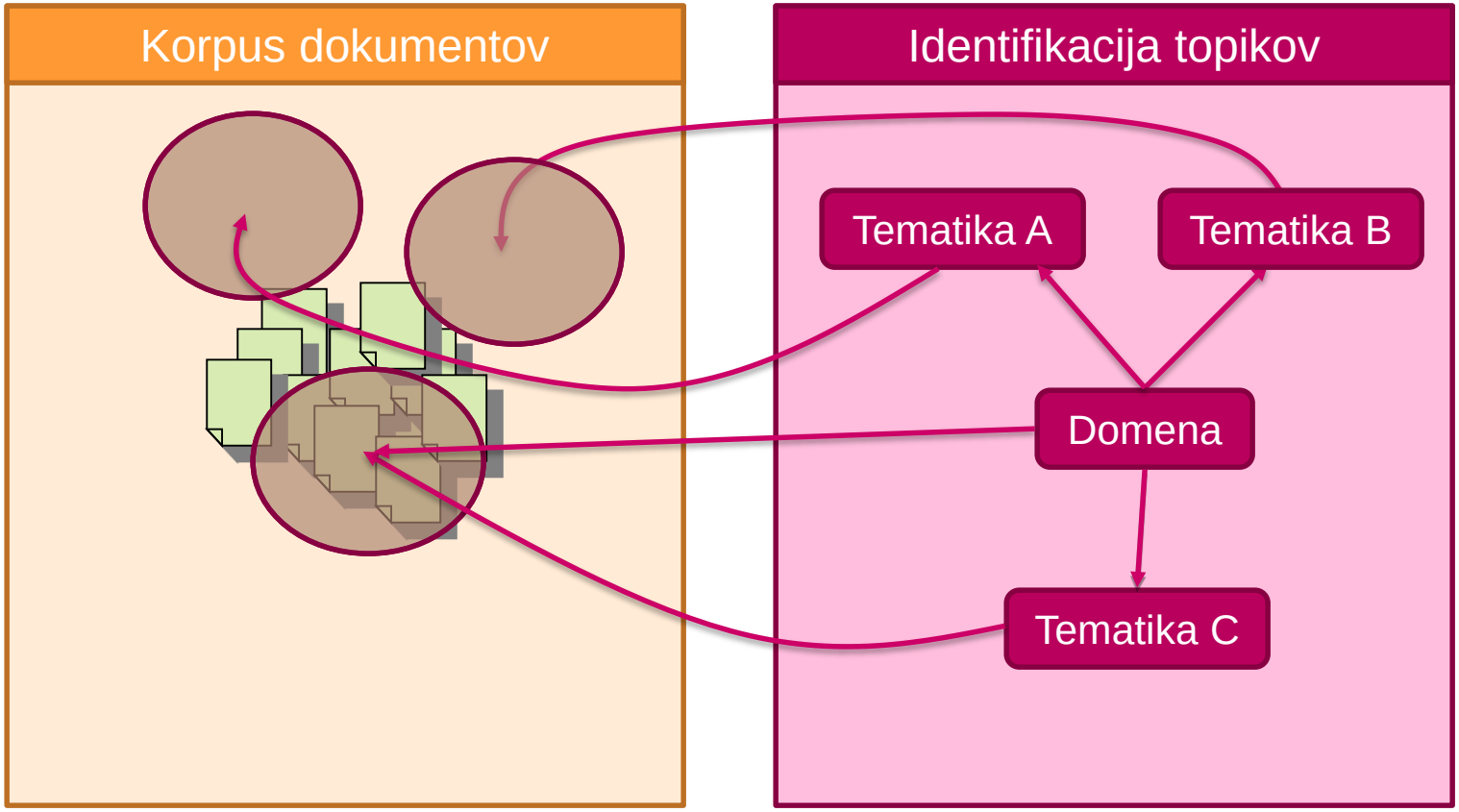
Tekstovno rudarjenje : Primeri aplikacij

- Analiza sentimenta v tvitih
- Klasifikacija dokumentov
- Ugotavljanje avtorstva dokumentov
- Razvrščanje dokumentov v skupine glede na njihovo podobnost
- Pomoč pri preiskovanju svetovnega spleta
- Analiza profilov uporabnikov spleta
- Iskanje anomalij in osamelcev
- Iskanje skritih povezav med znanstvenimi domenami
- …

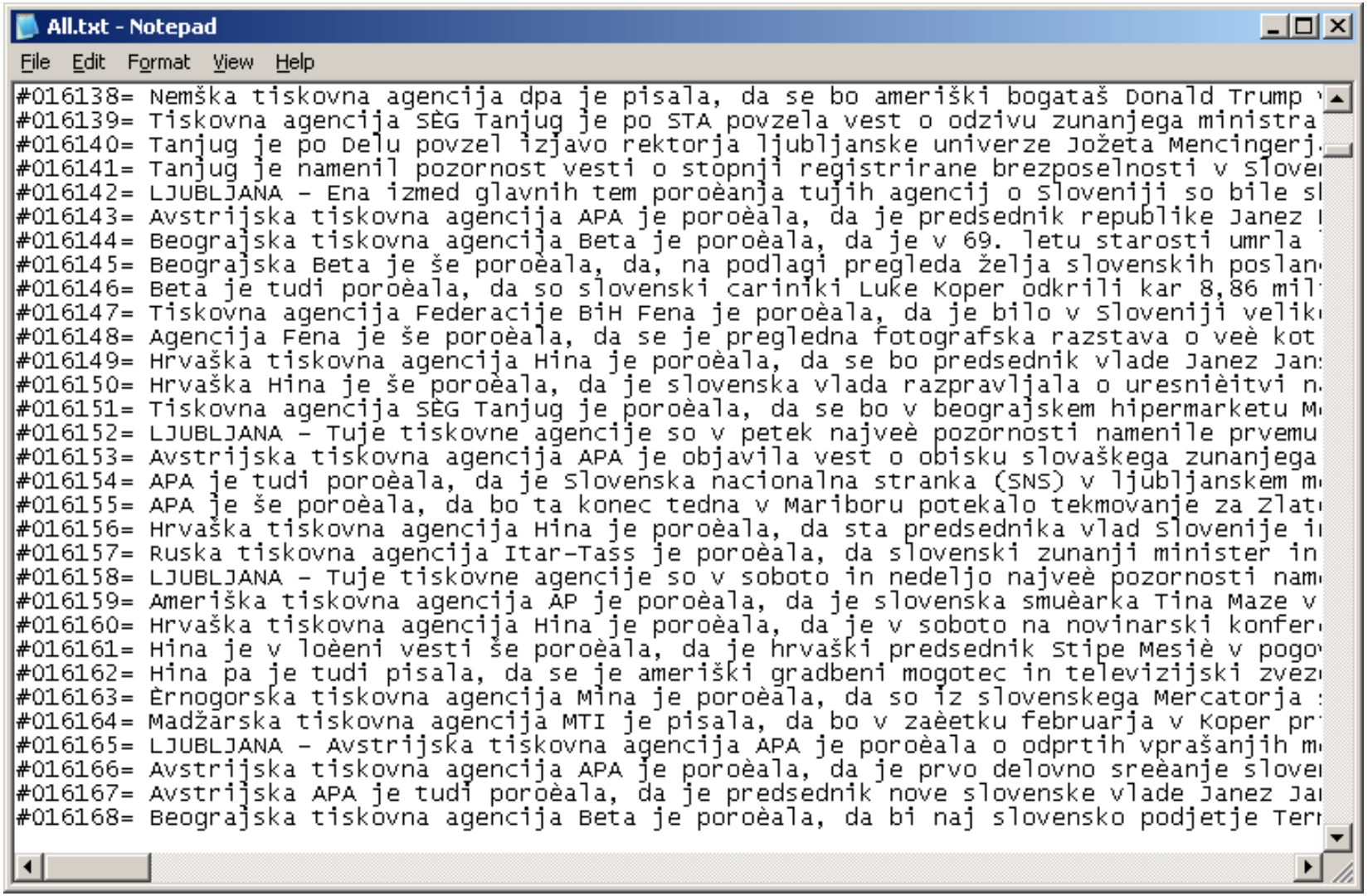
Primer: Analiza javnega mnenja z razvrščanjem novic STA v skupine

- Podatki: STA agencijske novice o Sloveniji povzete v tujih medijih
- Uporaba sistema za pol-avtomatsko razvrščanje v skupine
- V vsakem koraku postopka razcepimo množico dokumentov na k podskupin
  - Podatki znotraj skupine naj bodo čimbolj podobni
  - Podatki med skupinami naj bodo čim bolj različni
- Pri razvrščanju uporabimo mero podobnosti ali razdalje med dokumenti

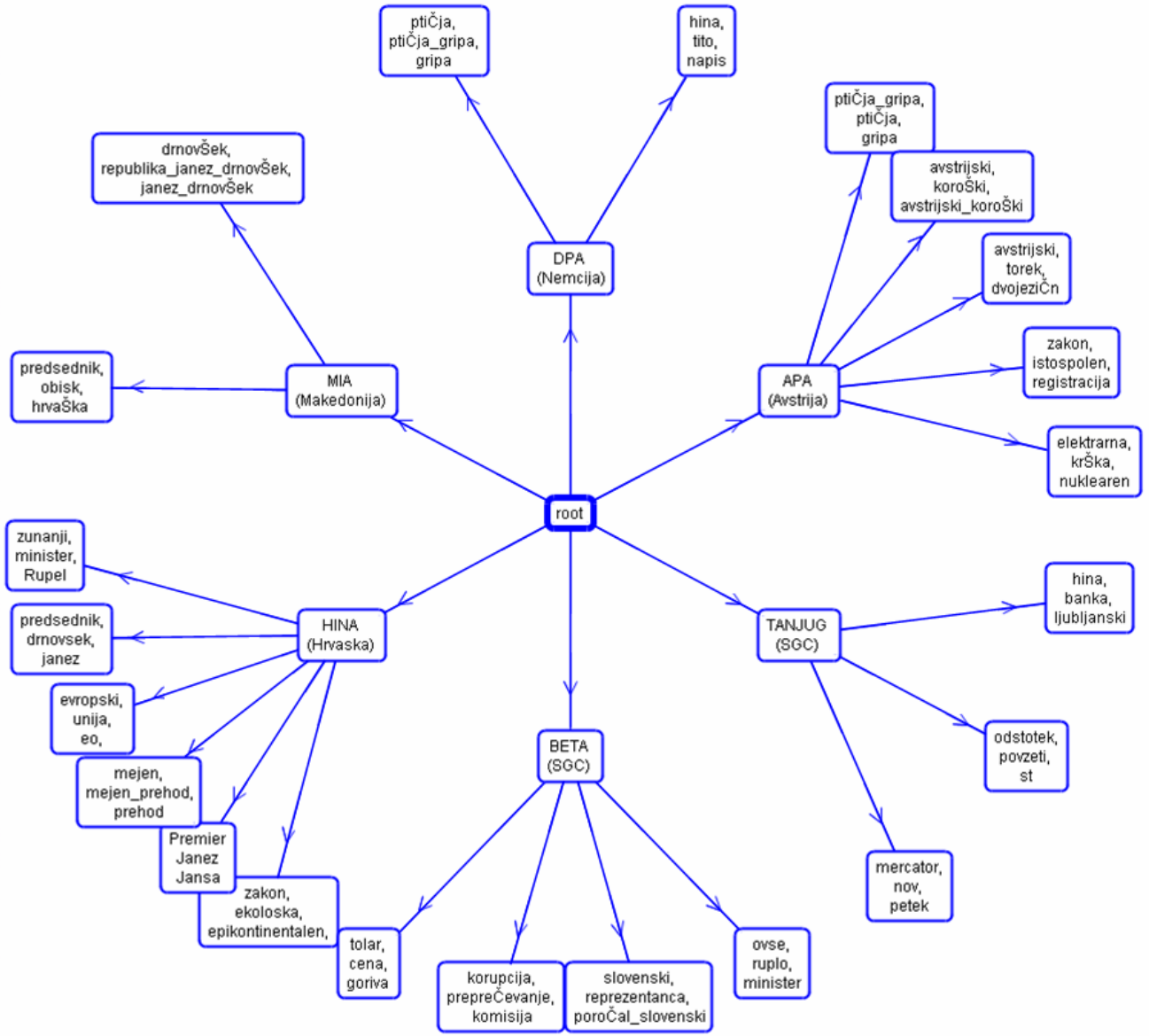
Razvrščanje člankov v skupine



Vhod v program



Analiza STA novic o Sloveniji v tujini



Znanstvena literatura kot vir znanja

Primer:

- Biomedicinska bibliografsta podatkovna baza PubMed
- US National Library of Medicine
- Več kot 21M citatov
- Več kot 5.600 revij
- 2.000–4.000 referenc dodanih na vsak delovni dan!

Iskanje bisociativnih povezav med članki iz različnih podpodročij

Argument 1 (magnesium literature)

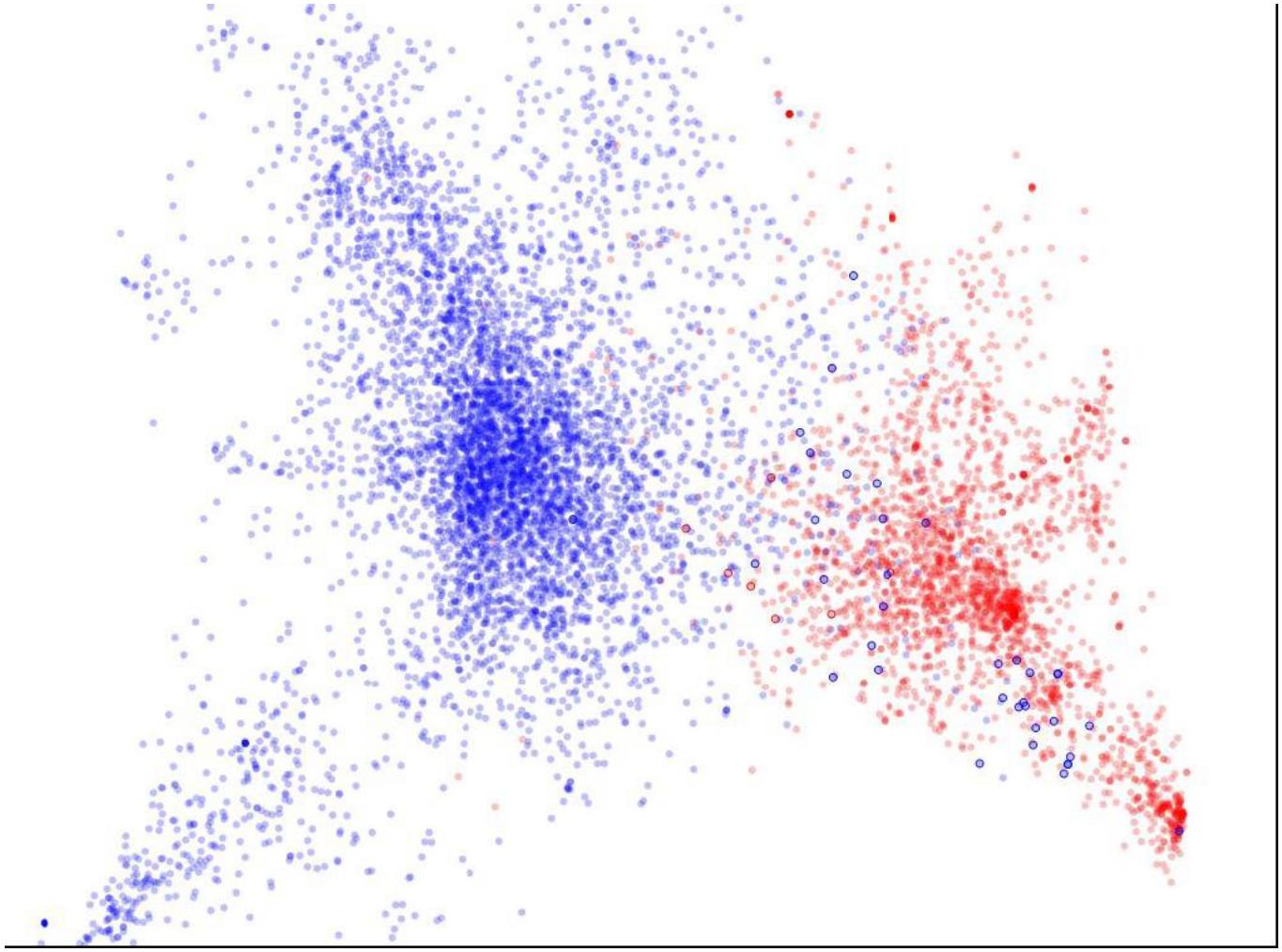
- Mg is a natural calcium channel blocker.
- Stress and Type A behavior can lead to body loss of Mg.
- Magnesium has anti-inflammatory properties.
- . . .

Argument 2 (migraine literature)

- Calcium channel blockers can prevent migraine attacks.
- Stress and Type A behavior are associated with migraine.
- Migraine may involve sterile inflammation of the cerebral blood vessels.
- . . .

Iskanje izjem in anomalij v podatkih

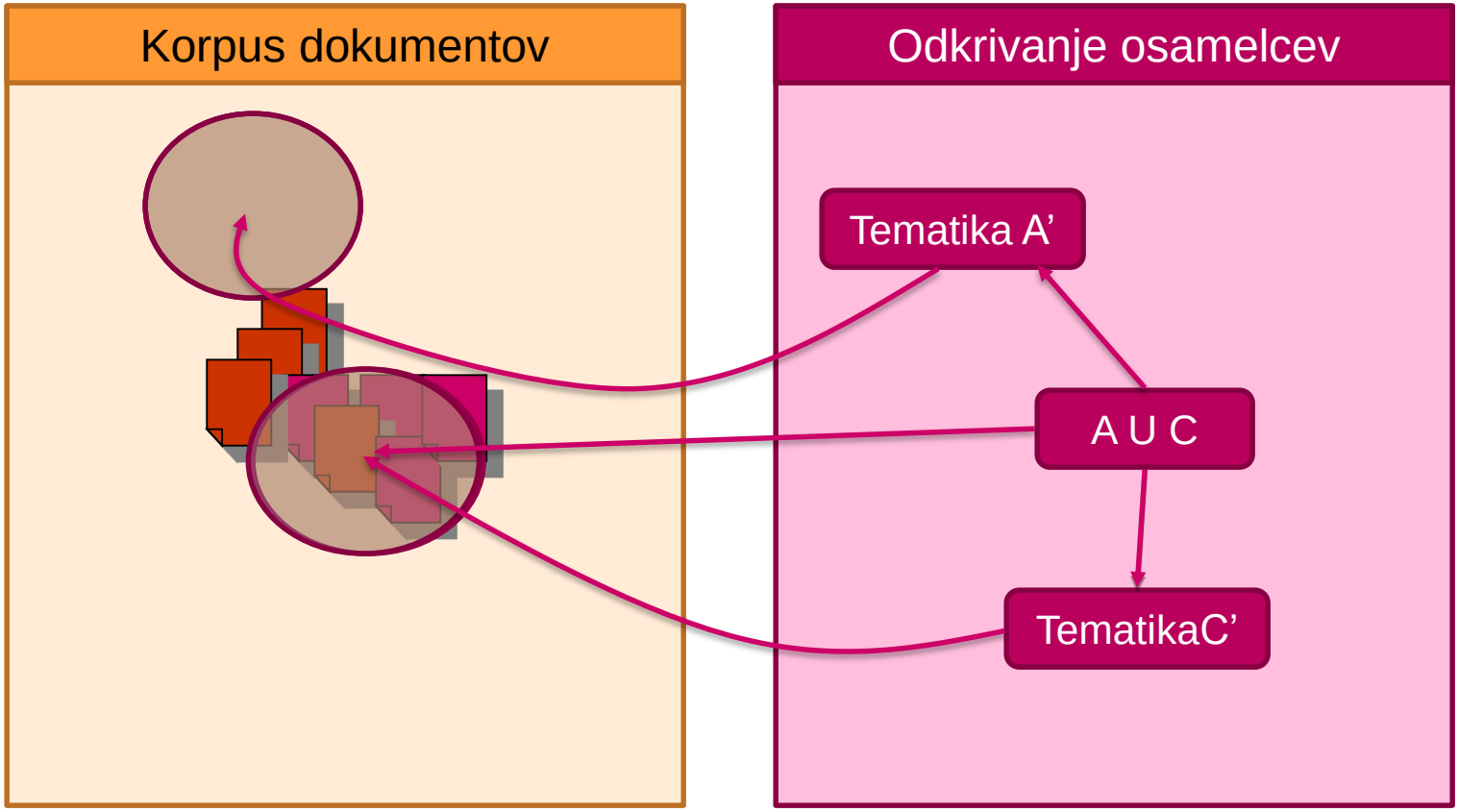
Iskanje izjem v korpusih besedil



2-dimenzionalna projekcija dokumentov z dveh različnih domen

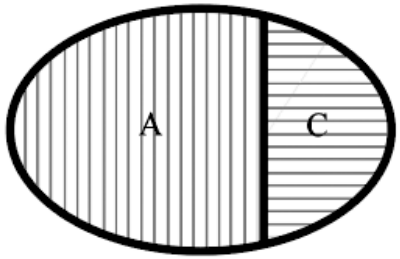
Iskanje nenavadnih člankov v korpusih

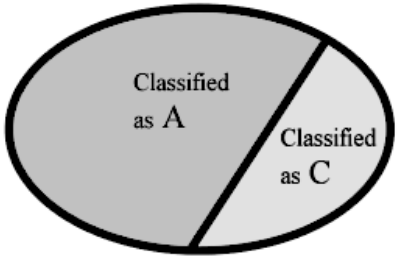
Uporaba razvrščanja v skupine za iskanje izjem v korpusih besedil

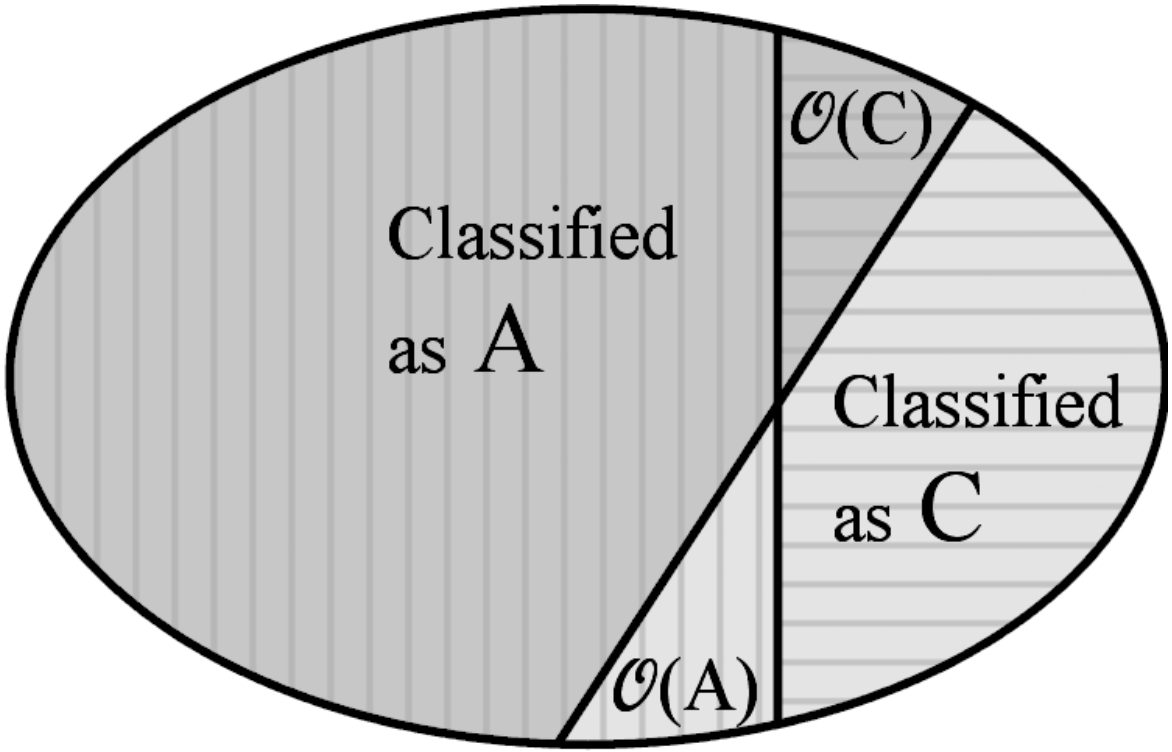


Klasifikacijski pristop za iskanje izjem v korpusih

- Iz podatkov se naučimo klasifikacijski model in ga uporabimo nad dokumenti iz dveh ločenih korpusov







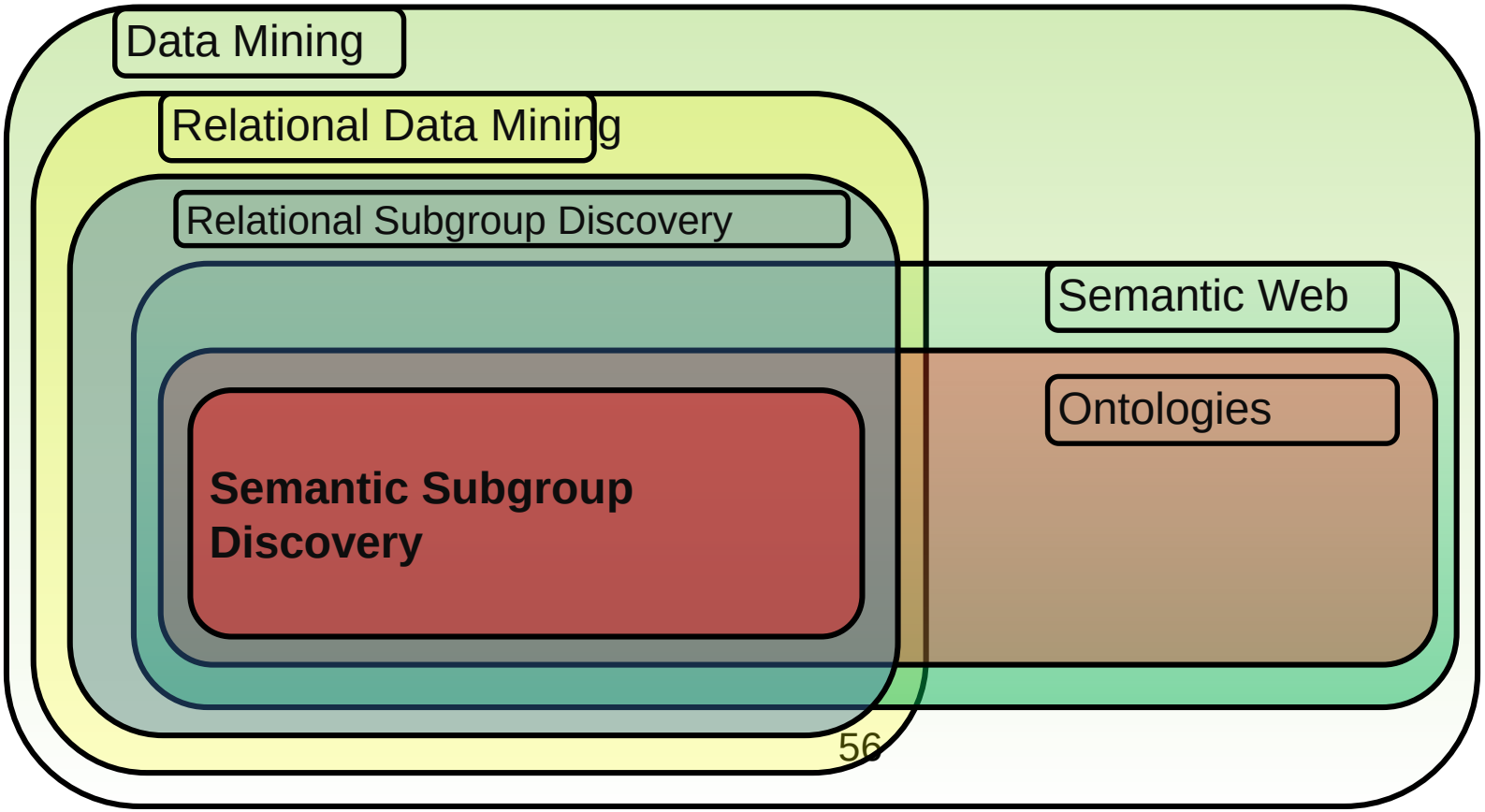
Rangiranje izjem v člankih o Kenijskih volitvah, objavljenih v lokalnih in zahodnih medijih

- **Članek 352**
  Ta članek so naknadno odstranili iz korpusa, ker ni govoril o politični in socialni problematiki, temveč o oropanem britanskem turistu..

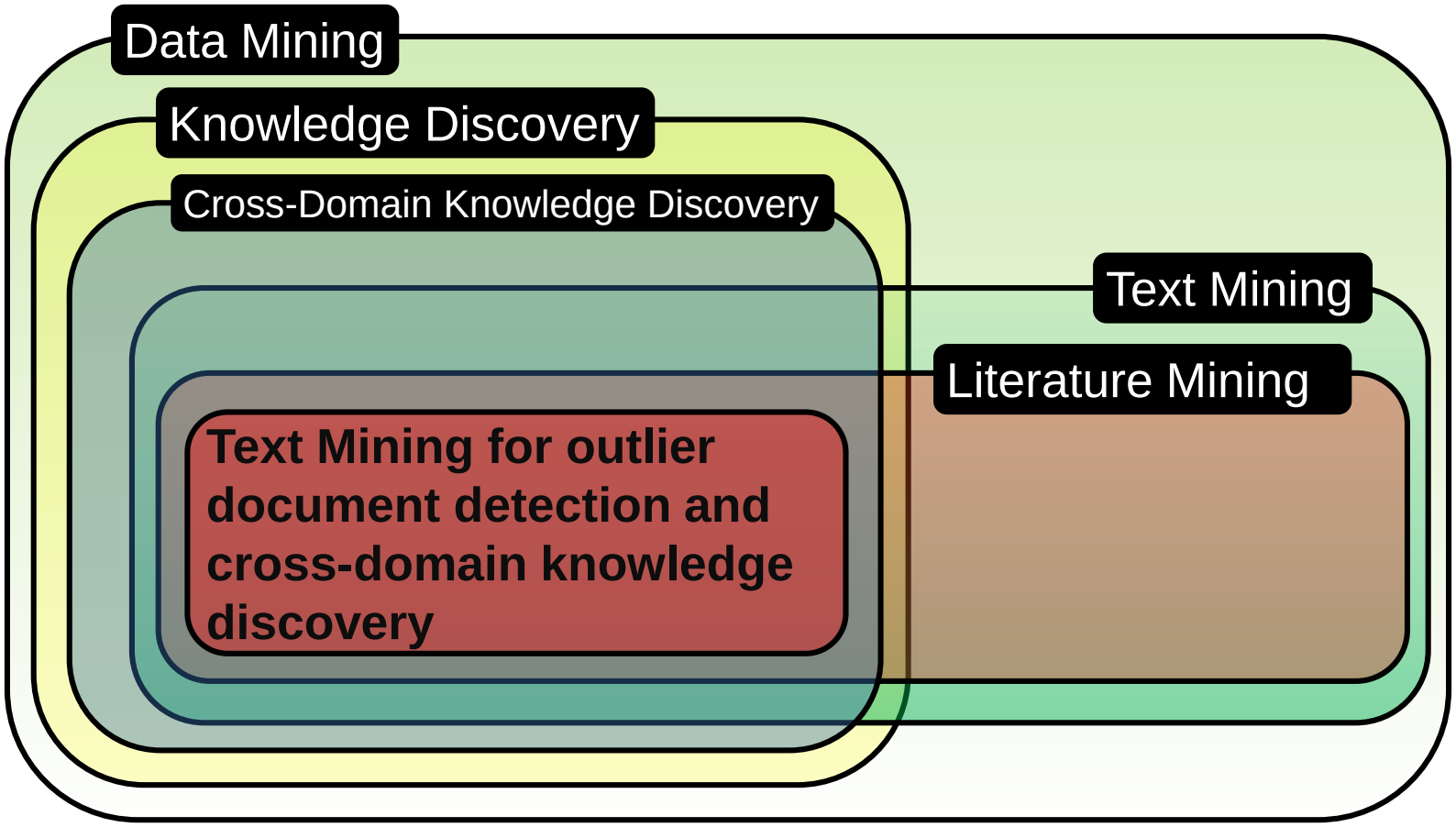
- **Članek 173**
  Ta članek je bil objavljen v lokalnem mediju a je bil avtomatsko klasificiran kot zahodni članek. Članek je dejansko napisal zahodni novinar, ki nima ustrezne senzitivnosti za pisanje (npr. omenja besede “pleme”, “plemenska pripadnost” itd. v negativnem kontekstu) – avtor je torej uporabljal zahodni stil pisanja.

Povzetek in tekoče delo

Semantično odkrivanje podskupin v podatkih (Vavpetič et al., 2012)



Povzetek in tekoče delo



Vsebina predavanja

- **odkrivanje zakonitosti in anomalij v podatkih**
- **rudarjenje tekstovnih in spletnih podatkov**
- računalniška podpora pri odločanju
- jezikovne tehnologije in računalniško jezikoslovje
- **računalniška kreativnost**

Kaj je računalniška kreativnost

- Računalniška kreativnost (umetna kreativnost, mehanska kreativnost ali kreativno računstvo) je multidisciplinarno področje znanosti na preseku umetne inteligence, kognitivne psihologije, filozofije in umetnosti.
- Cilji področja so modeliranje in simulacija človeške ustvarjalnosti kot tudi računalniško kreativno ustvarjanje in/ali izboljšanje človeške lastne kreativnosti s pomočjo robota/računalnika

Naši tekoči projekti s področja računalniške kreativnosti

- **MUSE – Machine Understanding For Interactive Storytelling (STREP, 2012-2015)**
  - Koordinatorka: Sien Moens, KU Leuven
- **PROSECCO – Promoting The Scientific Exploration Of Computational Creativity (CSA, 2013-2016)**
  - koordinator Tony Veale, UCD, Dublin
- **Concrete – Concept Creation Technology (STREP, 2013-2016)**
  - Koordinator Geraint Wiggins, QMUL, London
- **WHIM – The What-if Machine (STREP, 2013-2016)**
  - Koordinator Simon Colton, Goldsmith Uni., London
- **Skupaj ~ 1.1 Mio EUR evropskih sredstev** pridobljenih za naše delo v teh evropskih projektih

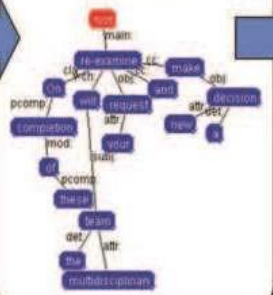
Primer projekta: MUSE

Patient Guidelines

On completion of the preparatory phase, the multidisciplinary team makes one of three decisions:
1 The operation can go ahead. The team will then give you more information on the operative technique chosen. If you have decided to have the operation, you will be given an operation date and a request for your health insurance fund to agree to help with the operation costs (to find out more :www.ameli.fr).
2 Your preparation for the operation is not sufficient. You will have to undertake additional preparations.
On completion of these, the multidisciplinary team will re-examine your request and make a new decision.
3 Surgery is not suitable in your case. The multidisciplinary team will explain the reasons why and offer you another treatment (non-surgical).

NLP Analysis

Connexor Analysis



Feature Structures Extraction

(:MAIN "re-examine"
((:OBJ "decision"
(:ATTR "new"))
((:OBJ "request"
(:ATTR "you"))
((:SUBJ "team"
(:ATTR
"multidisciplinary"))
("completion")))))

Plan-based Narrative Engine

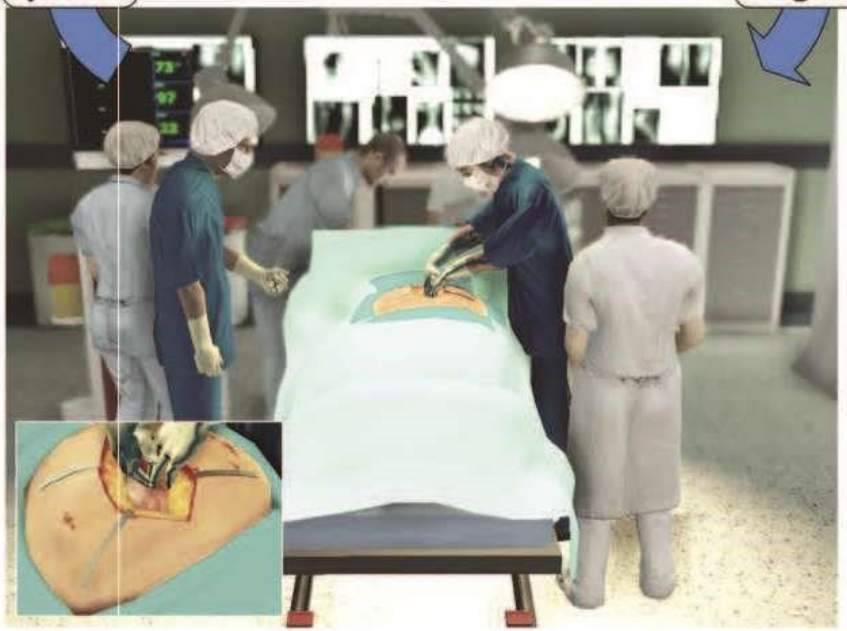
Narrative Actions

World State

... (...) (**can-provide-info** drnicholson anaesthesia) (...)
(**can-provide-info** drdowell digestive-surgery) (...) (**can-provide-info** drrichards nutrition)
(...) (**at-location** mrroberts reception) (**at-location** drsmith or-room) (...)
(**at-location** msjones patient-room) (...) ...

update

staging



3D Visualization (Unreal® Engine - UDK)

Primer projekta: MUSE

- **MUSE – Machine Understanding for Interactive Storytelling** (2012-2015)
  - 3D animacija teksta
  - z uporabo metod jezikovnih tehnologij omogočimo razumevanje teksta
  - rendering za 3D vizualizacijo zgodbe
  - planiranje za iskanje možnih razpletov zgodbe (what-if)
  - mehanizmi za igranje “resnih iger” v 3D svetu omogočajo uporabniku vodeno sledenje igri in interakcijo z virtualnim svetom
- **Uporabnost:** izobraževanje, medicina, letališka varnost, …, industrija iger

Področja računalniške kreativnosti

- Kreiranje pripovedi, zgodb, metafor, analogij, šal, neologizmov, poezije, glasbe, slik in drugih vizualnih artefaktov, marketinških oglasov, ...
- Primer: robot lahko napiše zgodbo, **Kaj-če stroj** (**What-If Machine**) pa lahko spremeni njen potek
- Uporabnost v industriji in znanosti

Naši raziskovalni cilji

- **Naš cilj:** Kreativno reševanje problemov in kreativno odkrivanje povezav, vzorcev in znanja iz podatkov.
  - Bisociativno preiskovanje biomedicinske literature za odkrivanje novih meddomenskih povezav
  - Kreativno kombiniranje algoritmov in avtomatsko generiranje novih
- **Naš pogled:**
  - Ni nujno, da je robot tudi sam kreativen, lahko je le inteligenten pomočnik
- **Večinski pogled:**
  - Računalnik/robot naj bo avtonomen kreator

Povzetek in zahvala sodelavcem odseka za soustvarjanje raziskovalnih poti v novo in še neraziskano