

Sistemi za istraživanje i analizu podataka

PRIKUPLJANJE I PRIPREMA PODATAKA

DATA WRANGLING

predavač: Aleksandar Kovačević

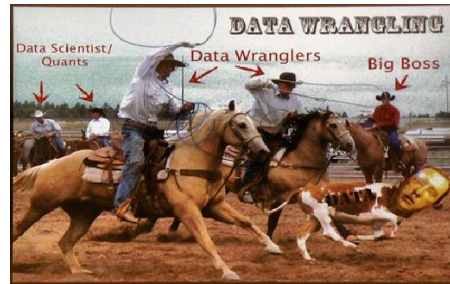
Želite da budete data scientist



Ovako ćete provesti većinu svog posla



Šta je *data wrangling*?



- **Cilj:** Izvući i standardizovati sirove podatke. Kombinovati više izvora podataka. Očistiti anomalije u podacima.
- **Strategija:** Kombinovati automatizaciju sa interaktivnim vizualizacijama kako bi se olakšalo čišćenje.

[illegible]

25 companies

- Healthcare
- Retail, Marketing
- Social networking
- Media
- Finance, Insurance

Various titles
Data analyst
Data scientist
Software engineer
Consultant
Chief technical officer

Kandel et al. "Enterprise Data Analysis and Visualization: An Interview Study. IEEE Visual Analytics Science & Technology (VAST), 2012
<http://db.cs.berkeley.edu/papers/vast12-interview.pdf>

Istraživanje o tome šta rade data sientisti

"I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any 'analysis' at all...

... Most of the time once you transform the data ... the insights can be scarily obvious."

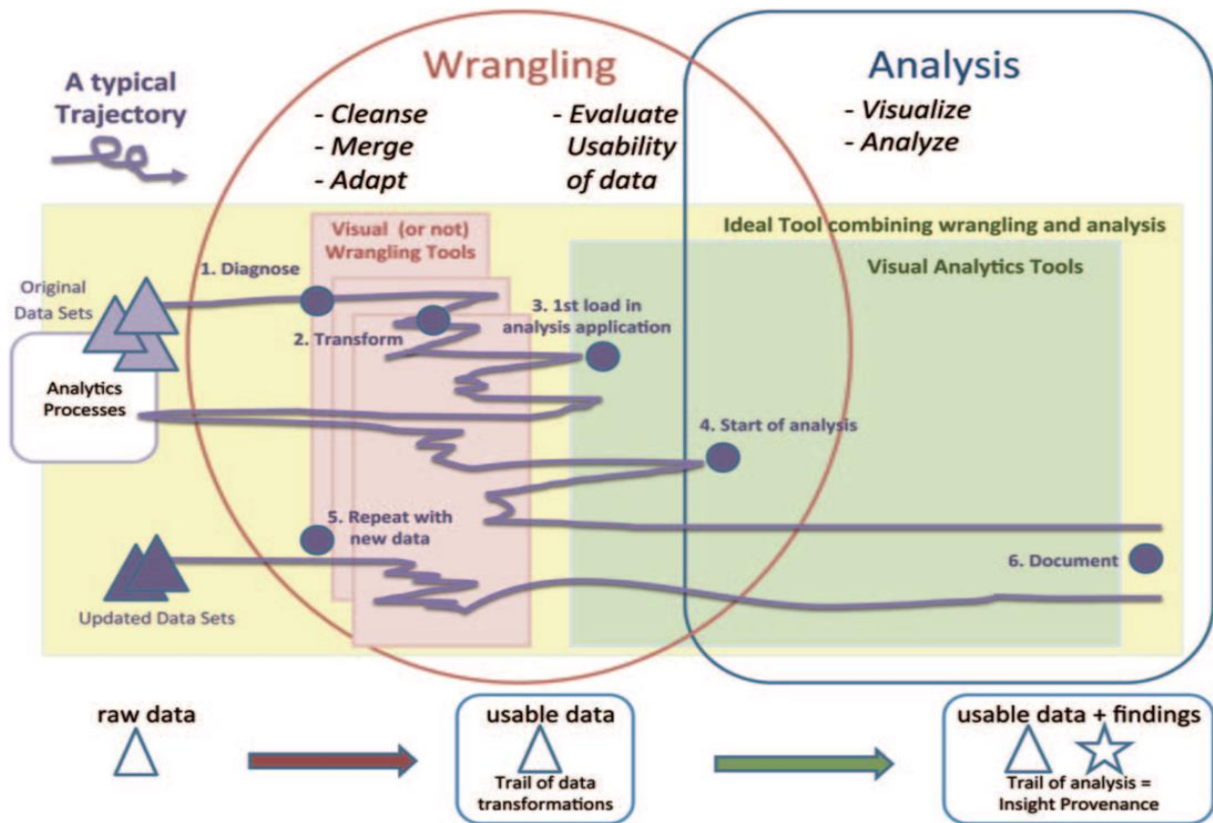
"Once you play with the data you realize you made an assumption that is completely wrong. It's really useful, it's not just a waste of time, even though you may be banging your head."

"In practice it tends not to be just data prep, you are learning about the data at the same time, you are learning about what assumptions you can make."

Još jedno istraživanje o tome šta rade data sientisti

Data Wrangling oduzima između 50% i 80% procesa analize podataka...

Izvor: anketiranje data science eksperata



<https://www.quora.com/What-is-the-research-source-behind-the-statement-Data-scientists-spend-up-to-80-of-their-time-on-data-wrangling-munging>

Još jedno „istraživanje“ o tome šta rade data sientisti



**Big Data
Borat**

@BigDataBorat

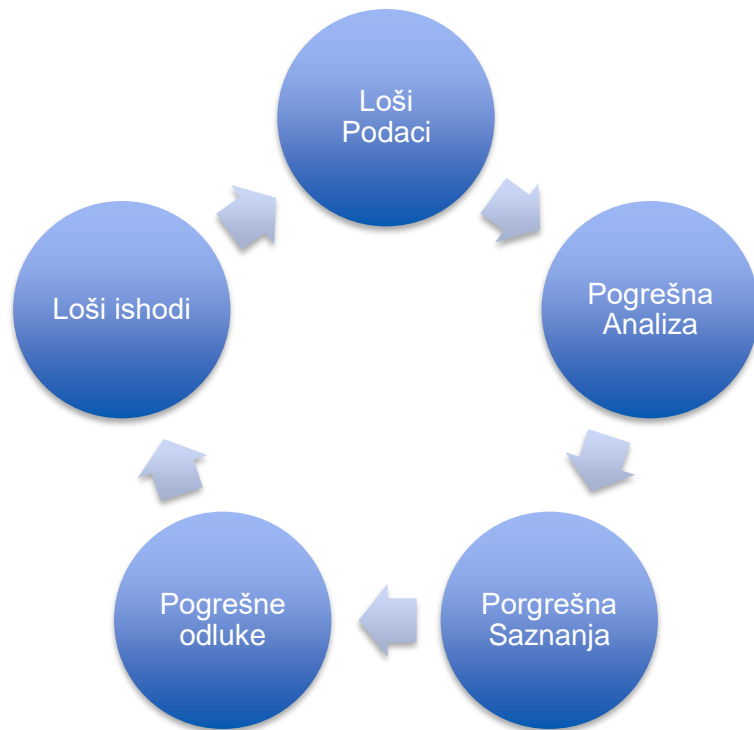


Following

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.



Zašto je priprema podataka važna?

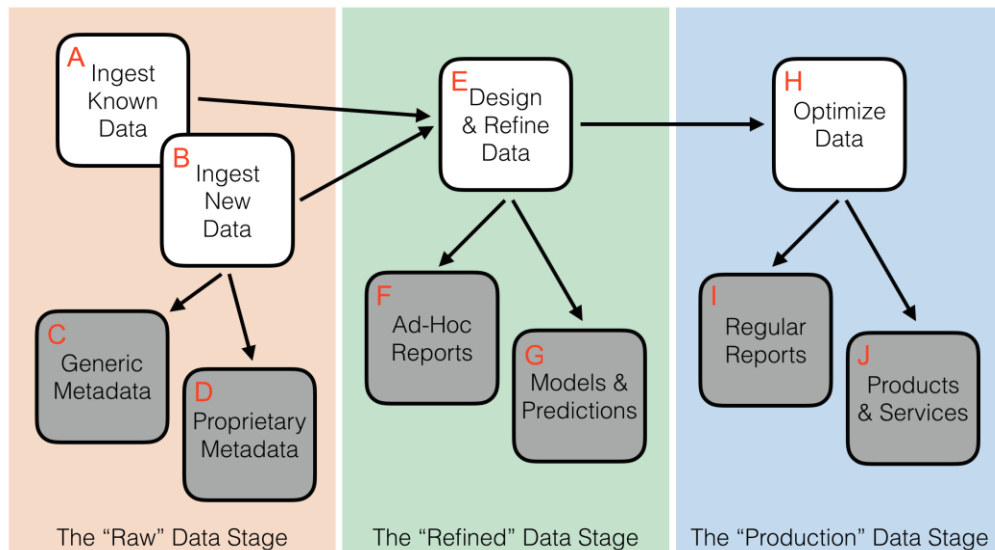


Šta se događa kada imamo loše podatke?

- “Dear Idiot’ Letter”
- “17,000 men are pregnant”
- “As the crow flies”

Faze u prikupljanju i skladištenju podataka

- **Sirovi podaci:** Prikupljanje i prva analiza (“unboxing”) podataka
- **Šta se radi:** Prikupljanje i obrada podataka
- **Ko radi:** Data scientist (Data wrangler)
- **Rafinirani podaci:** Skladištenje podataka
- **Šta se radi:** Data warehousing, baze podataka
- **Ko radi:** Data curators, IT engineers...
- **Produkcija:** Procesi za korišćenje rafiniranih podataka
- **Šta se radi:** Automatizovanje raznih načina na koje se koriste podaci (npr. izveštavanje, predikcija, sistemi za preporuku itd.)
- **Ko radi:** SW engineers, IT/ops (data scientisti razvijaju modele, a u ovoj fazi se oni koriste)



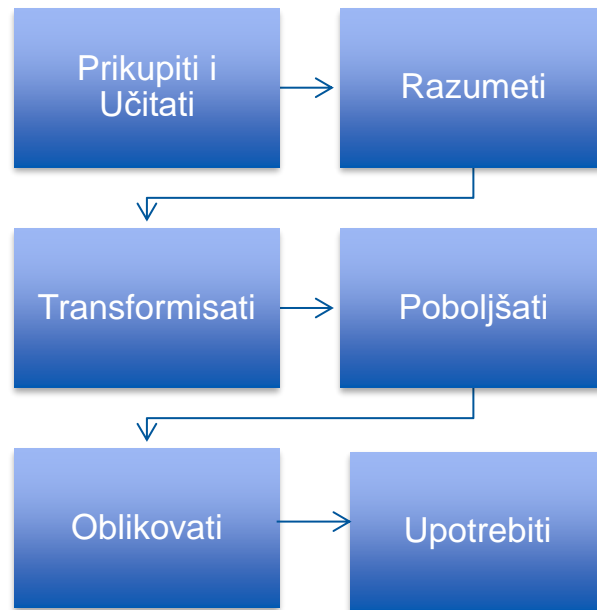
Rattenbury, et al. "Data Wrangling: Techniques and Concepts for Agile Analytics". To appear, O'Reilly Media, 2017.

Tema današnjeg predavanja

- Fokusiramo se na prvu fazu: “Raw→Refined”
- Prikupljanje podataka
- „Unboxing“
- Transformisanje podataka u oblik pogodan za analizu (*analytics-ready structure*)
- Procena kvaliteta podataka

Koraci u data wrangling procesu

- Iterativni proces obrade podataka
- Prikupiti
- Razumeti (istražiti)
- Transformisati
- Poboljšati sa drugim podacima (informacijama)
- Oblikovati
- Upotrebiti



An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem. – John Tukey

Koraci u data wrangling procesu – konkretnije 1/2


- Učitavanje i prikupljanje podataka (Data Ingestion)
 - CSV
 - PDF
 - API/JSON
 - HTML Web Scraping....
- Razumevanje (Istraživanje)
 - Prvi susret sa podacima (“unboxing“)
 - Pregled oblika podataka, vizualno istraživanje, sumarne statistike...
- Oblikovanje
 - Organizovanje podataka u oblik pogodan za metode koje želimo da primenimo...

Koraci u data wrangling procesu – konkretnije 2/2

- Čišćenje podataka
 - Nedostajuće vrednosti
 - Autlajeri
 - Pogrešno uneti podaci
 - Domenski pogrešni podaci....
- Poboljšavanje podataka (*Data Augmenting*)
 - Agregacija različitih izvora
 - Tačno ili Fuzzy uparivanje različitih reprezentacija istih podataka...

Učitavanje i prikupljanje podataka

- Učitavanje podataka u memoriju je generalno jednostavnije od prikupljanja.
- Postoji puno različitih formata podataka.
- Svi poznati data wrangling alati mogu da učitaju mnogo različitih formata.



Type	Data Description	Reader	Writer
text	CSV	<code>read_csv</code>	<code>to_csv</code>
text	Fixed-Width Text File	<code>read_fwf</code>	
text	JSON	<code>read_json</code>	<code>to_json</code>
text	HTML	<code>read_html</code>	<code>to_html</code>
text	Local clipboard	<code>read_clipboard</code>	<code>to_clipboard</code>
	MS Excel	<code>read_excel</code>	<code>to_excel</code>
binary	OpenDocument	<code>read_excel</code>	
binary	HDF5 Format	<code>read_hdf</code>	<code>to_hdf</code>
binary	Feather Format	<code>read_feather</code>	<code>to_feather</code>
binary	Parquet Format	<code>read_parquet</code>	<code>to_parquet</code>
binary	ORC Format	<code>read_orc</code>	
binary	Msgpack	<code>read_msgpack</code>	<code>to_msgpack</code>
binary	Stata	<code>read_stata</code>	<code>to_stata</code>
binary	SAS	<code>read_sas</code>	
binary	SPSS	<code>read_spss</code>	
binary	Python Pickle Format	<code>read_pickle</code>	<code>to_pickle</code>
SQL	SQL	<code>read_sql</code>	<code>to_sql</code>
SQL	Google BigQuery	<code>read_gbq</code>	<code>to_gbq</code>

Učitavanje i prikupljanje podataka

- Ponekad je potrebno podatke prikupiti iz izvora kao što su:

- PDF,
- Web strane,
- Slike
- Tekst

- Tada koristimo specijalizovane alate

textextract



BeautifulSoup

extricator

spaCy



Tesseract OCR



NLTK

Razumevanje podataka – “Unboxing“

- Prvi susret sa podacima
- Cilj nam je da vidimo kakve podatke stvarno imamo:
 - oblik (struktura) – u kom tačno obliku su podaci?
 - granularnost – da li je svaki red jedan entitet?
 - temporalnost – da li imamo vremenski atribut i u kom obilku?
 - sveobuhvatnost (*scope*) – da li su podaci kompletni?
 - kvalitet – koliko dobro podaci predstavljaju realnu pojavu koju analiziramo?

Razumevanje podataka – “Unboxing“

- Odgovori na neka od pitanja sa prethodnog slajda su subjektivni.
- Data science nije potpuno automatizovan proces i uključuje čoveka
- Odgovori na pitanja su takođe promenljivi jer je data wrangling deo data science ciklusa, odnosno to je faza na koju se često vraćamo.
 - Na primer, setite se jednog od sigurnijih načina da poboljšate svaki ML model (*more data always helps...*)

Čišćenje podataka

- Nedostajuće vrednosti
- Netačni podaci
- Nedosledne reprezentacije istih podataka
- Oko 75% problema zahteva intervenciju čoveka (eksperti, crowdsourcing itd.)
- Treba naći balans između čišćenja podataka i prevelikog čišćenja podatka (overly sanitized data)
 - Npr., anonimizacija – da li na taj način uklanjamo i korisne podatke?

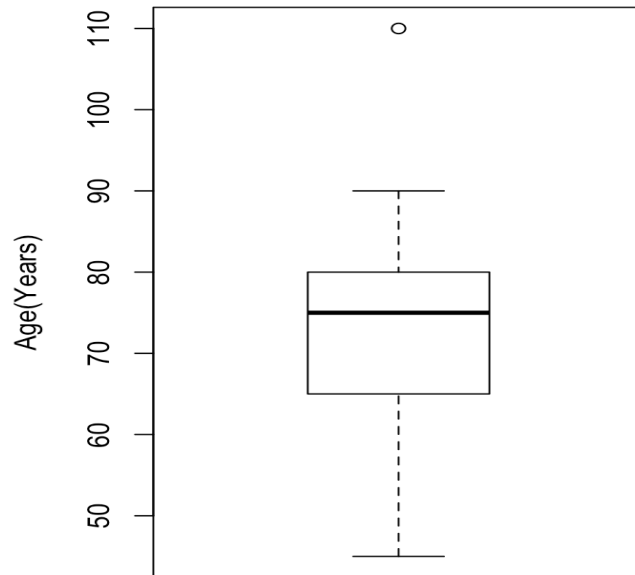
Čišćenje podataka

- Vizualizacije i standardne statistike mogu da otkriju nepravilnosti u “sirovim” podacima
- Različite reprezentacije podataka za različite probleme:
 - Autlajeri se obično mogu otkriti pomoću grafikona
 - Nedostajući podaci se mogu videti u tabelarnom prikazu
- Sve se komplikuje sa porastom količine i dimenzionalnosti podataka
 - Semplovanje je jedno rešenje
 - Postoji mnogo drugih tehnika...

Autlajeri – vrednosti koje odstupaju od ostalih

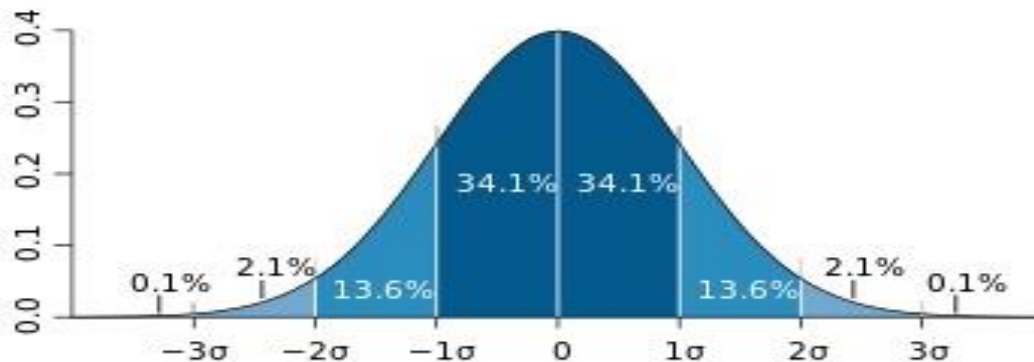
???
75
80
65
55
67
78
88
90
45
58
69
80
110

Age(Years)
75
80
65
55
67
78
88
90
45
58
69
80
110



Autlajeri – jedan atribut

- Prvi korak: pogledati histogram vrednosti i utvrditi da li podaci prate normalnu distribuciju
- Ako su vrednosti normalno distribuirane:
 - Jedna opcija:
 - izračunati srednju vrednost i standardnu devijaciju
 - svaka tačka koja je više od dve standardne devijacije udaljena od srednje vrednosti može se smatrati autlajerom



Autlajeri – jedan atribut

- Problem kod prethodnog pristupa je uticaj autlajera na srednju vrednost i st.dev. Ove vrednosti će biti „pomerene“. Robusnija opcija:

- koristiti medijan umesto srednje vrednosti, pa onda
- razmatrati podatke koji mnogo odstupaju od medijana
- izračunati *Median Absolute Deviation* (MAD) umesto st.dev.

$$\mathbf{MAD} = \mathbf{median}(|X_i - \mathbf{median}(X)|)$$

- Pravilo (iz literature): st.dev. = 1.4826 * MAD
- svaka tačka koja je više od 2*st.dev. udaljena od srednje vrednosti može se smatrati autlajerom

Autlajeri – jedan atribut

- Ako vrednosti nisu normalno distribuirane:
- Isolation forest
- k-nn metode (npr. upotreba DBScan klasterovanja)
 - tačke koji imaju malo komšija u okolini su potencijalni autlajeri
 - Relativna gustina = $\text{gustina tačke} / \text{prosečna gustina njeg. komšija}$

Autlajeri – više atributa

- Prvi korak: Pomoću stat. testova utvrditi da li podaci prate normalnu distribuciju (postoje gotove implementacije)
- Ako su vrednosti normalno distribuirane:
 - Koristimo sredju vrednost i matricu kovarijansi
 - Računamo Mahalanobis Depth of point (MDP)
 - MDP meri koliko tačka odstupa od norm. dist.
 - Tačke sa velikom vrednošću MDP su potencijalni autlajeri

Autlajeri – više atributa

- Ako vrednosti nisu normalno distribuirane:
- Isolation forest
- k-nn metode (npr. upotreba DBScan klasterovanja)
 - tačke koji imaju malo komšija u okolini su potencijalni autlajeri
 - Relativna gustina = $\text{gustina tačke} / \text{prosečna gustina njeg. komšija}$

Nedostajuće vrednosti

Antony Mc James IV, 123 Untidy Cir, Suite# 234, Cumbersome, TX, 76849
Miller Johnson, 1102 Messy Data St, Middletow, OH
Betty Flyier 6483 Phew Lane, Apt A4, FixMe, CA, 91103



Nedosajuće vrednosti zbog lošeg parsiranja

Customer	Address Line 1	Address Line 2	City	State	Zip
Antony Mc James IV	123 Untidy Cir	Suite# 234	Cumbersome	TX	76849
Miller Johnson	1102 Messy Data St	Middletow	OH		
Betty Flyier 6483 Phew Lane	Apt A4	FixMe	CA	91103	

Nedosajuće vrednosti zbog neprimenljivosti pitanja

Date of your most recent IMMUNIZATIONS:

Hepatitis A _____ Hepatitis B _____ Influenza (flu shot) _____ MMR _____ Tetanus (Td) _____
Pneumovax (pneumonia) _____ Meningitis _____ Varicella (chicken pox) shot or Illness _____
HPV _____ Zostavax (Shingles) _____

HEALTH MAINTENANCE SCREENING TESTS:

Lipid (cholesterol) & Sugar _____ Date _____ Abnormal? ☐ Yes ☐ No
Sigmoidoscopy _____ or Colonoscopy _____ Date _____ Abnormal? ☐ Yes ☐ No

Women: Mammogram Date _____ Abnormal? ☐ Yes ☐ No Pap Smear Date _____ Abnormal? ☐ Yes ☐ No
Have you ever had a DEXA Scan (for bone density) ☐ Yes ☐ No Abnormal? _____ Date _____

Women's Health History: Number of: pregnancies _____ deliveries _____ abortions _____ miscarriages _____
Age at start of periods: _____ Date of Last Period: _____ Age at end of periods: _____

Name and phone number of your OB/Gynecologist: _____

Men's Health History:
Have you had a blood test for: PSA (prostate) ☐ Yes ☐ No If yes, what date _____ Was it normal? _____

PERSONAL MEDICAL HISTORY: Please indicate whether YOU have had any of the following medical problems (with dates).

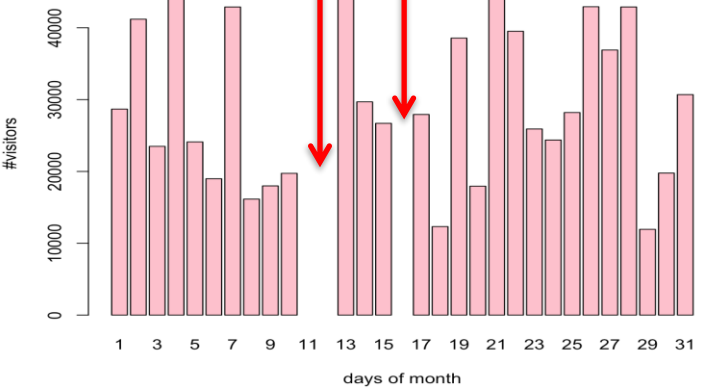
- Alcoholism _____
Anxiety / Depression _____
Cancer, specify type _____
Heart disease _____
Diabetes _____
Stroke _____
Urinary / Kidney Problems _____
- Arthritis / Rheumatologic _____
Bleeding or Clotting disorder or Blood Clot _____
Thyroid problems _____
High Blood Pressure _____
High Cholesterol _____
Other _____

Do you see any other doctors? If yes, please list their name, office phone number and specialty.



MISSING? DID SERVER CRASH?!

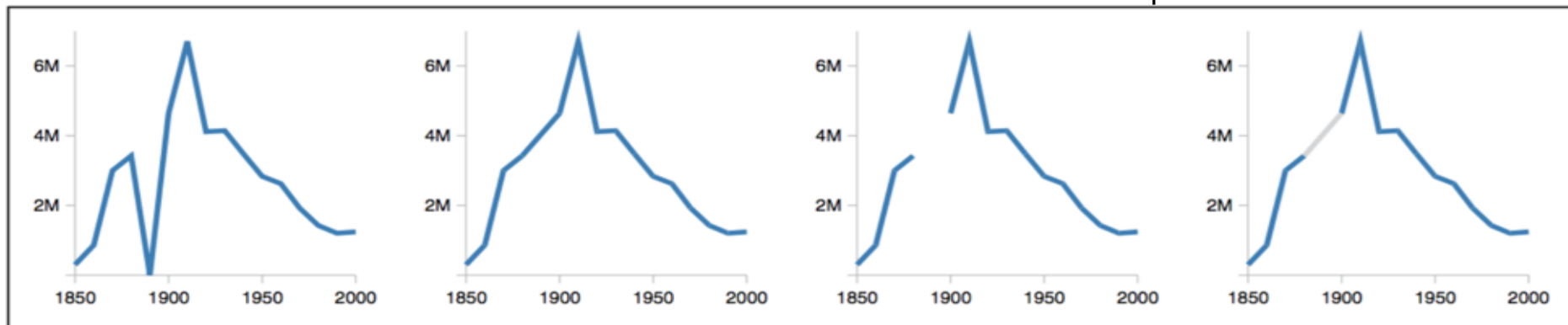
Missing @
Random
Missing completely



Nedostajuće vrednosti

- Domensko znanje je ključno!
- Postaviti vrednosti na 0? na prosek?
- Interpolirati na osnovu postojećih podataka?
- Izbaciti slogove/atribute?

U.S. census broj ljudi koji rade kao
“Farm Laborers”; podaci od 1890 su
nestali u požaru



Poboljšavanje podataka (Data Augmenting)

Uparivanje entiteta (Entity Resolution)

- Cilj: Identifikovati različite reprezentacije istog entiteta u podacima koje koristimo
 - identity reconciliation, record linkage, deduplication, fuzzy matching, object consolidation, coreference resolution
- Primeri:
 - Imena i prezimena, adrese, nazivi kompanija,poređenje proizvoda pri kupovini (comparison shopping), upraivanje naloga na društvenim mrežama, uparivanje autora naučnih radova...
- Ovaj korak je vrlo značajan jer korektna identifikacija entiteta menja frekvencije itd. i utiče na šablone koje otkrivamo.

Poboljšavanje podataka (Data Augmenting)

Uparivanje entiteta (Entity Resolution)

- Izazovi: Prirodno neodređeni podaci (imena, prezimena, adrese...), Skraćenice:

USA =

United States Army United States of America Ulhasnagar Sindhi Association Ultimate in Suspense and Action Unconditional Self-Acceptance Unconventional Stellar Aspect Under Secretary of the Army Underground Service Alert Underground Sewer Adapter Underwriting Service Assistant Unicycling Society of America	Union of South Africa Union Street Athletics Unionville-Sebewaing Area Unique Settable Attributes Unit Self Assessment United Scenic Artists University of South Alabama University of South Australia Unix System Admin Unstable Angina Unusually Sensitive Area
--	---

- Greške u unosu podataka, pravopisne greške itd.

Poboljšavanje podataka (Data Augmenting)

Uparivanje entiteta (Entity Resolution)

- Univerzalno rešenje ne postoji. Rešenja su mahom domenski zavisna.
- Generalni predlog metodologije 1/2:
 - Pronaći meru sličnosti za entitete:
 - Na primer, Levenstain (broj promena stringa potreban da se iz jednog dobije drugi)
 - Q-grami (pronaći sve pod-stringove dužine q , napraviti skup njih za svaki string, koristiti mere za sličnost skupova)
 - Često korišćena mera trenutno je *Jaro-Winkler* (postoje gotove implementacije)

Poboljšavanje podataka (Data Augmenting)

Uparivanje entiteta (Entity Resolution)

- Generalni predlog metodologije 2/2:
 - Kada imamo meru sličnosti:
 - Uparujemo entitete koji su slični iznad nekog praga ili
 - Koristimo klasterovanje
 - Aglomerativno se pokazalo kao najbolje
 - Šta je problem sa K-means ($K = ?$, puno malih klastera)
 - Nažalost, ako želimo stvarno čiste podatke ručna validacija uparenih entiteta je nezubežna
 - Jedno rešenje je crowdsourcing (Mechanical Turk servis npr.)

Alati za Data Wrangling – open source 1/2



IP[y]



pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Refine ^{OPEN} 

DATASCIENCE**TOOLKIT**

Alati za Data Wrangling – open source 2/2



BeautifulSoup

Tesseract

textract



OCRFeeder



<http://schoolofdata.org/>



<http://okfnlabs.org/>



Alati za Data Wrangling – komercijalni



Primeri Data Wrangling u SIAP projektima prethodnih generacija

Predikcija saobraćajnih gužvi na osnovu video materijala nadzornih kamera

- Autori: Mihaela Osmajić, Marija Golubović
- Cilj:
 - Automatski utvrditi prisustvo saobraćajne gužve na osnovu snimka video kamere
 - Pronaći i integrisati izvore video materijala i meteroloških podataka
 - Obučiti modele mašinskog učenja za predikciju gužvi
 - Analizirati rezultate i predložiti pravce daljeg poboljšanja pristupa

Predikcija saobraćajnih gužvi na osnovu video materijala nadzornih kamera

- Izvori podataka:

- *Skup podataka o vozilima je dobijen obradom video materijala kamere u realnom vremenu, postavljene u ulici Terzije u Beogradu.*

- Preuzeto sa YouTube kanala Beograd.com

- *Podaci o meteorološkim uslovima za kreiranje skupa podataka o istim preuzeti su sa sajta wunderground.com.*

- Informacije o temperaturi, pritisku, vetru, vlažnosti, padavinama za svakih pola sata u toku dana.

Predikcija saobraćajnih gužvi na osnovu video materijala nadzornih kamera

- **Data wrangling 1/2**: Skup podataka o vozilima je kreiran kroz proces koji je uključivao detekciju vozila, praćenje njihovog kretanja, procenu brzine i obradu dobijenih podataka.

- Video materijal je ručno sečen na na intervale od sat vremena
- Bilo je potrebno je odstraniti nerelevantne kadrove koji bi samo crpeli resurse i vreme. Ovo je učinjeno korišćenjem unapred kreiranih maski koje su prepoznavale takve kadrove.
- Da bi se izbegao problem sa obradom velikog broja frejmova, video materijal je ubrzan.
- Za detekciju vozila korišćen je YOLO model verzije 8, dok je za pracenje kretanja vozila korišćen DeepSort algoritam. Podatak o brzini vozila za jedan frejm dobijen na sledeći način:

$$v = t * u(p1,p2) / pm \quad (1)$$

- v - brzina vozila
- u - udaljenost između pozicija vozila na dve frejma, izračunata euklidskim rastojanjem
- pm - konstanta koja predstavlja broj piksela po metru, faktor konverzije piksela u metre
- t - konstanta koja predstavlja vreme potrebno za pređeni put od 1m

Predikcija saobraćajnih gužvi na osnovu video materijala nadzornih kamera

- **Data wrangling 2/2**: Obrada skupa podataka o vozilima na nivou frejma je podrazumevala pre svega ekstrakciju instanci koje predstavljaju automobile, autobuse i kamione, svi ostali identifikovani objekti i podaci o njima nam nisu relevantni.
 - Relativna srednja brzina vozila je podrazumevala usrednjavanje brzina i normalizaciju najvećom brzinom među podacima.
 - Analizom podataka smo uvidele da podatak je podatak o broju vozila za svaku grupu linearno zavistan od ukupnog broja vozila, što bi značilo da nam je jedino važna informacija o ukupnom broju vozila, ne i o grupi.
- **Integracija izvora podataka**: Skup podataka o meteorološkim uslovima i skup podataka spojeni su u odnosu na datum i vremenski interval, nakon čega su kreirana tri nova obeležja na osnovu datuma: dan u nedelji, mesec i godina.
- Kreiranje ciljnog obeležja: Određivanje vrednosti ciljne labela izvršeno je primenom k-means klasterovanja.
 - Podaci su klasterovani u dva klastera gde je jedan proglašen za prisustvo gužve, a drugi ne.

Traffic Accident Severity Prediction in the City of New York

- Autori: Milica Škipina, Đorđe Batić
- Cilj:
 - Pronaći i integrisati što više izvora podataka sa informacijama relevantnim za saobraćajne nesreće
 - Obučiti modele mašinskog učenja za predikciju težine nesreće
 - Analizirati rezultate i faktore koji utiču na težinu nesreće

Traffic Accident Severity Prediction in the City of New York

- Izvori podataka:

- *Motor Vehicle Collisions*

- Tri tabele sa detaljima o nesreći, vozilima i osobama koji su učestvovali u nesreći.

- *IBM Weather API*

- Podaci vremenskim prilikama skrejpovani sa <https://www.ibm.com/weather>

- *LION Single Line Street Base Map*

- Linear Integrated Ordered Network (LION) - informacije o ulicama Njujorka ali podeljenim na segmente gde jedna ulica može da ima više segmenata

- *Traffic Volume Counts*

- Informacije o intenzitetu saobraćaja

Traffic Accident Severity Prediction in the City of New York

- Data wrangling:
- Učitavnje podataka – relativno lako za 3 strukturirana izvora podataka
- Prikupljanje dodatnih podataka – skrejpovanje i čišćenje na taj način dobijenih podataka može bude vremenski zahtevan posao.
- Integracija, Upraivanje (matching):
 - Podatke o svakoj nesreći trebalo je upariti sa svim ostalim izvorima podataka (vremenske prilike, ulice i intezitet saobraćaja).
 - Uparivanje nije jednostavno i zahteva analiziranje atributa svakog izvora podataka
- Čišćenje podataka

Traffic Accident Severity Prediction in the City of New York

- Posao oko integracije:
 - Filtriranje skupa podataka o nesrećama po datumu da bi se mogao upariti sa podacima o intezitetu saobraćaja
 - Puno posla oko upravljanja podataka o segmentu ulice sa podacima o nesreći
 - Svaka nesreća uparuje se sa segmentom ulice koji joj je najbliži. Udaljenost se računa pomoću geografskih koordinata. Prvo se vrši uparivanje imena ulice nesreće i segmenta.
 - Deo skupa podataka o nesrećama nije imao naziv ulice. Za takve primere naziv ulice je dobijen od najbližeg segmenta. Autori su odabrali 0.001 kao prag za udaljenost iznad koga se smatra da se ne može pronaći najbliži segment da bi se dodelilo ime ulice.

Traffic Accident Severity Prediction in the City of New York

- Čišćenje podataka:
 - U skupu podataka o nesrećama uklonjeni su redovi kod kojih nije bilo podataka o broju povređenih ili preminulih kao i o lokaciji nesreće
 - U skupu podataka o nesrećama bio je veliki broj ulica kojima nije uneto ime – postupak rešavanja opisan je na prethodnom slajdu

Traffic Accident Severity Prediction in the City of New York

- Kreiranje novih i transformacija postojećih atributa 1/4:
 - Atribut za vreme dat u obliku 'DD-MM-YYYY HH:mm' razdvojen je na četiri atributa: godina, mesec, dan i sat.
 - Mesec, dan i sat u ciklični pa su transformisani u dve dimenzije na sledeći način:

$$x_{sin} = \sin\left(\frac{2 * \pi * x}{max(x)}\right)$$

$$x_{cos} = \cos\left(\frac{2 * \pi * x}{max(x)}\right)$$

Traffic Accident Severity Prediction in the City of New York

- Kreiranje novih i transformacija postojećih atributa 2/4:
 - Lokacija je data kao geografska širina i dužina – koordinate su transformisane u tro-dimenzioni prostor na sledeći način:

$$x = \cos(lat) * \cos(lon)$$

$$y = \cos(lat) * \sin(lon)$$

$$z = \sin(lat)$$

Traffic Accident Severity Prediction in the City of New York

- Kreiranje novih i transformacija postojećih atributa 3/4:
 - One-hot enkoding za kategoričke attribute
 - Kategorički atributi sa puno vrednosti su uprošćeni. Na primer za atribut *Condition*:
 - CLD for normal or cloudy conditions (e.g. 'Mostly Cloudy', 'Cloudy', 'Partly Cloudy', 'Mostly Cloudy / Windy',...),
 - LVS for low visibility conditions (e.g. 'Fog', 'Haze', 'Thunder', 'Drizzle and Fog',...)
 - RD for dangerous road conditions (e.g. 'Rain', 'Heavy Rain', 'Snow', 'Snow and Sleet',...).

Traffic Accident Severity Prediction in the City of New York

- Kreiranje novih i transformacija postojećih atributa 4/4:
 - Nesreće u dvosmernim ulicama reprezentovane su sa dva reda zbog protoka saobraćaja u jednom i u drugom smeru.
 - Izvršeno je spajanje u jedan red i uzet je prosek protoka.

Capital Bikeshare

- Autori: Gorana Gojić, Angelina Vujanović, Radovan Turović
- Cilj:
 - Analizirati podatke sistema za iznajmljivanje bicikala
 - Prikazati najinteresantnija saznanja
 - Kreirati model za predikciju potencijalnih lokacija za nove stanice za iznajmljivanje

Capital Bikeshare

- Izvori podataka:
 - <https://www.capitalbikeshare.com/>
 - Podaci o iznajmljivanjima (vožnjama), CSV format
 - Podaci o stanicama, XML
 - <https://www.openstreetmap.org/>,
http://wiki.openstreetmap.org/wiki/Overpass_API
 - Podaci o geografskim objektima bez semantike, XML
 - OSM tag finder, <https://tagfinder.herokuapp.com/>
 - Semantika OSM tagova (tip objekta: muzej, turistički, biznis, kupovina itd.), JSON

Capital Bikeshare

- Data wrangling:
- Učitavnje (parsiranje) CSV, XML, JSON
- Integracija, Upraivanje (matching):
 - Podaci o vožnjama sa podacima o stanicama
 - Podaci o stanicama sa OSM podacima
 - OSM podaci sa podacima o vrstama objekata
- Čišćenje podataka

Capital Bikeshare

- Problemi kod integracije:
- Podaci o vožnjama sa podacima o stanicama. Jedan primer:
 - Različiti nazivi stanica u podacima sa vožnjama u odnosu na podatke o stanicama
 - To je Entity Resolution (Matching) problem
 - Rešen je na dva načina:
 - Upotrebom drugog atributa za uparivanje (naziv terminala)
 - Ručnim korekcijama

Capital Bikeshare

- Nedostajući podaci
- Podaci o stanicama sa OSM podacima
 - Jako puno objekata od 20,000 sadržalo je samo geo. koordinate bez drugih tagova
 - Rešenje: obrisati takve objekte iz skupa
- Filtriranje atributa
 - Jako puno neinformativnih atributa (datumi, komentari korisnika itd.)
 - Rešenje: ručno odrediti skup takvih atributa
 - Automatski ih otkloniti
 - Ukloniti sve objekte koji su posle filtriranja ostali bez atributa

Capital Bikeshare

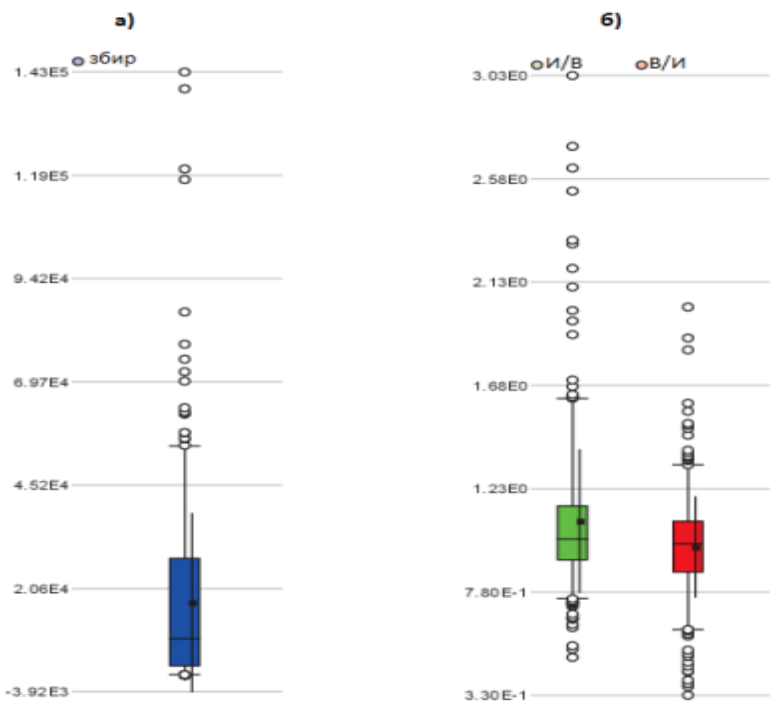
- Problemi sa čišćenjem podataka
- Podaci o stanicama sa OSM podacima
 - Rezultat OSM API je preko 20,000 objekata
 - Nisu svi objekti interesantni za analizu (kante za otpatke, trafo stanice,...)
 - Rešenje:
 - filtriranje objekata po tipu
 - ručno je određen skup tipova objekata (cluture, shopping, business itd.)

Capital Bikeshare

- Dalji postupak
- Jako puno flitriranja, Na primer:
 - Pronalaženje objekata u krugovima različitog radijusa (značaja) oko stanica
- Kreiranje novih atributa:
 - Rang objekta – zavisi od radiusa, što bliže stanici to bolje
 - Popularnost stanice – agregacija broj iznajmljivanja i vraćanja bicikala

Capital Bikeshare

- Korak razumevanja i vizualizacije podataka
- Na primer, kako diskretizovati broj iznajmljenih i vraćenih bicikala
 - Razlog, kako dobiti korisne kategorije za popularnost stanice (šta znači najmanja, a šta najveća popularnost itd.)
- Da li ima smisla kreirati kategorije za autlajere?



оцена	доња граница	горња граница
0	0	0
1	0+	376
2	376+	2566
3	2566+	28646
4	28646+	54607
5	54607+	∞

Capital Bikeshare

- Rezime i rezultati
- Jako puno posla u data wrangling fazi
- Nakon kreiranja konačnog skupa podataka:
 - Izračunati su koeficijenti korelacije popularnosti sa ostalim atributima
 - Kreiran je model koji predviđa popularnost stanice na osnovu ostalih atributa (visoka popularnost = potencijalno mesto za novu stanicu)
 - Prethodne dve stavke = nekoliko: klikova u RapidMiner alatu ili linija koda u *sci-kit* biblioteci
 - Uporedite to sa Wrangling delom!

Capital Bikeshare

- Rezime i rezultati
- Potvrđene hipoteze koje je Capital Bikeshare objavio o navikama svojih putnika:
 - ✓ **H1** *"The top bikeshare trip purposes overall were for personal/non-work trips."*
 - ✓ **H2** *"A large share of members used bikeshare for their trip to work."*
 - ✓ **H3** *"Capital BikeShare also served as a feeder service to reach transit stops."*
- Relativno dobri prediktivni rezultati za kategorije kod kojih ima mogo stanica
- Pобоljšanje rezultata potencijalno boljom diskretizacijom
- Još jedan rezultat projekta – objavljen rad na međunarodnoj naučnoj konferenciji!

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

- Izvori podataka:

- <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

- Podaci o 5000+ filmova sa IMDB, veliki broj meta-podataka (npr. broj Facebook lajkova i sl.), CSV

- <http://www.myapifilms.com/>

- Zarada nakon premijere, odnosno nakon prvog vikenda podaci o produkcionim kompanijama, nominacijama i nagradama na festivalima, JSON

- <http://www.youtube.com>

- Broj pregleda, lajkova, dislajkova i komentara na trejler, JSON

Predikcija zarade filmova

- Izvori podataka:
 - <http://www.omdbapi.com/>
 - Prosečna ocena i broj kritika sa Rotten Tomatoes, JSON
 - <http://www.boxofficemojo.com/>
 - Podaci o zaradama glumaca, reditelja i scenarista , web scrapping (parsiranje HTML), python - urllib

Predikcija zarade filmova

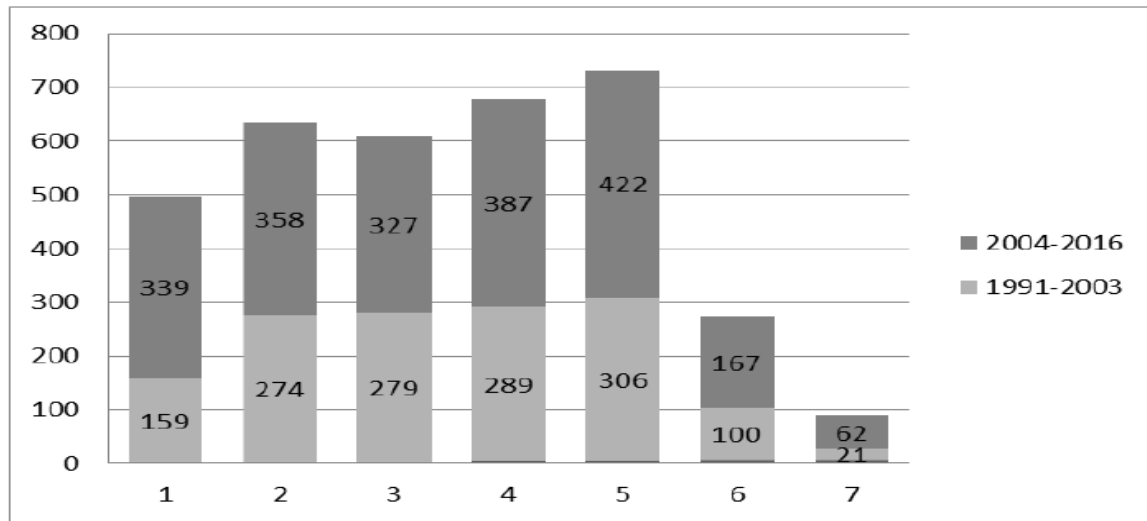
- Čišćenje podataka:
- Žanr – csv (commedy, animation, family...) -> one hot encoding
- Jezik filma agregiran na engleski i ostali
- Kompanije kao kategorijalni atribut (kompanija koja ima više od 40 filmova predstavlja posebnu kategoriju...)
- Broj nominacija i nagrada na Oskaru, Zlatnom globusu, posebno i zbirno broj nominacija i nagrada na ostalim festivalima.
- Kao stepen uticajnosti pojedinih glumaca, reditelja i scenarista uzet je njihov finansijski uspeh u karijeri, odnosno ukupna zarada svih filmova u kojima su učestvovali.

Predikcija zarade filmova

- Veliki broj (1600+) filmova sa nedostajućim vrednostima (podaci o zaradama itd).
 - Zbog nemogućnosti pribavljanja podataka nisu dalje razmatrani
- Zarade koji nisu u dolarima su konvertovane u tu valutu
- Filtrirani su filmovi posle 1991.
 - Tada snimljen prvi filma sa budžetom od 100M\$, pa se ta godine može uzeti kao prekretnica u pogledu finansija u filmu.
- Youtube i Facebook atributi imaju smisla od 2004
 - Skup podataka podeljen na 2 dela (pre i posle 2004)
 - Delovi su zasebno analizirani

Predikcija zarade filmova

- Korak razumevanja i vizualizacije podataka
- Slično kao kod Capital Bikeshare, kako diskretizovati zaradu i tako dobiti kategorijalnu klasu?



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

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

Predikcija zarade filmova

- Rezime i rezultati
- Kao i kod prethodnog projekta na Data Wrangling je utrošeno mnogo vremena
- Veoma dobri rezultati klasifikacije (RMSE za ordinalnu regresiju)

	Period	SVM	RDF	GBS	NN
Pre	1991-2003	1.293	1.659	0.978	1.371
	2004-2016	1.235	2.02	0.984	1.229
Posle	1991-2003	1.139	1.290	0.765	1.094
	2004-2016	1.080	1.009	0.579	1.101

Tabela 2. Rezultati dobijeni klasifikacijom

Predikcija zarade filmova

- Rezime i rezultati
- Relativno dobri rezultati regresije (modified R2)

	Period	GLM	NN
Pre	1991-2003	0.602	0.59
	2004-2016	0.744	0.667
Posle	1991-2003	0.723	0.703
	2004-2016	0.806	0.716

Tabela 1. Rezultati dobijeni regresijom

Procena broja stanovnika za opštine u Republici Srbiji

- Autori: Miloš Savić, Rajko Ilić i Vladimir Baumgartner
- Cilj:
 - Prikupiti što više različitih podataka korisnih za predikciju broja stanovnika (broj dece u školama, migracije itd.)
 - Kreirati prediktivni model
 - Različiti modeli: vremenske serije (na osnovu prethodnih brojeva stanovnika), regresija, klasifikacija...

Procena broja stanovnika za opštine u Republici Srbiji

- Izvori podataka:

1. Republički statistički zavod, podaci po naseljima, XLS

1_Uparedni pregled broja stanovnika.xls [Read-Only] [Compatibility Mode]												
	B	C	D	E	F	G	H	I	J	K	L	M
1	Упоредни преглед броја становника 1948, 1953, 1961, 1971, 1981, 1991, 2002 и 2011. године											
2	Comparative overview of the number of population in 1948, 1953, 1961, 1971, 1981, 1991, 2002 and 2011											
3	Регион Област Град – општина Насеље	Број становника / Number of population									Region Area City – Municipality Settlement	
4	РЕПУБЛИКА СРБИЈА	6527583	6978119	7641962	8446726	9313686	7822795	7498001	7186862		REPUBLIC OF SERBIA	
5	Градска	1717478	1980083	2574244	3505997	4390358	4214698	4218479	4271872		Urban	
6	Остала	4810105	4998036	5067718	4940729	4923328	3608097	3279522	2914990		Other	
7	СРБИЈА – СЕВЕР	2274602	2430477	2797161	3161920	3504855	3616115	3608116	3591249		SRBIA – SEVER	
8	Градска	1092884	1220689	1547383	1968387	2301491	2426482	2427598	2491575		Urban	
9	Остала	1181718	1209788	1249778	1193533	1203364	1189633	1180518	1099674		Other	
10	Београдски регион	634003	731837	942190	1209360	1470073	1602226	1576124	1659440		Beogradski region	
11	Градска	437053	521114	721183	990272	1206235	1310920	1274924	1344844		Urban	
12	Остала	196950	210723	221007	219088	263838	291306	301200	314596		Other	
13	Београдска област (Град Београд)	634003	731837	942190	1209360	1470073	1602226	1576124	1659440		Beogradska oblast (Grad Beograd)	
14	Градска	437053	521114	721183	990272	1206235	1310920	1274924	1344844		Urban	

Procena broja stanovnika za opštine u Republici Srbiji

- Izvori podataka:

- 2. Republički statistički zavod, XLS i PDF

- 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.

Procena broja stanovnika za opštine u Republici Srbiji

- Učitavanje (ekstrakcija) podataka
- Za period od 1972 - 1996. godine postoje samo PDF-ovi koji su slike tj. iz njih nije moguće selektovati tekst.
- Opcije: OCR ili ručno prekucavanje
- Nijedna nije idealna, autori su se odlučili za ručno prekucavanje

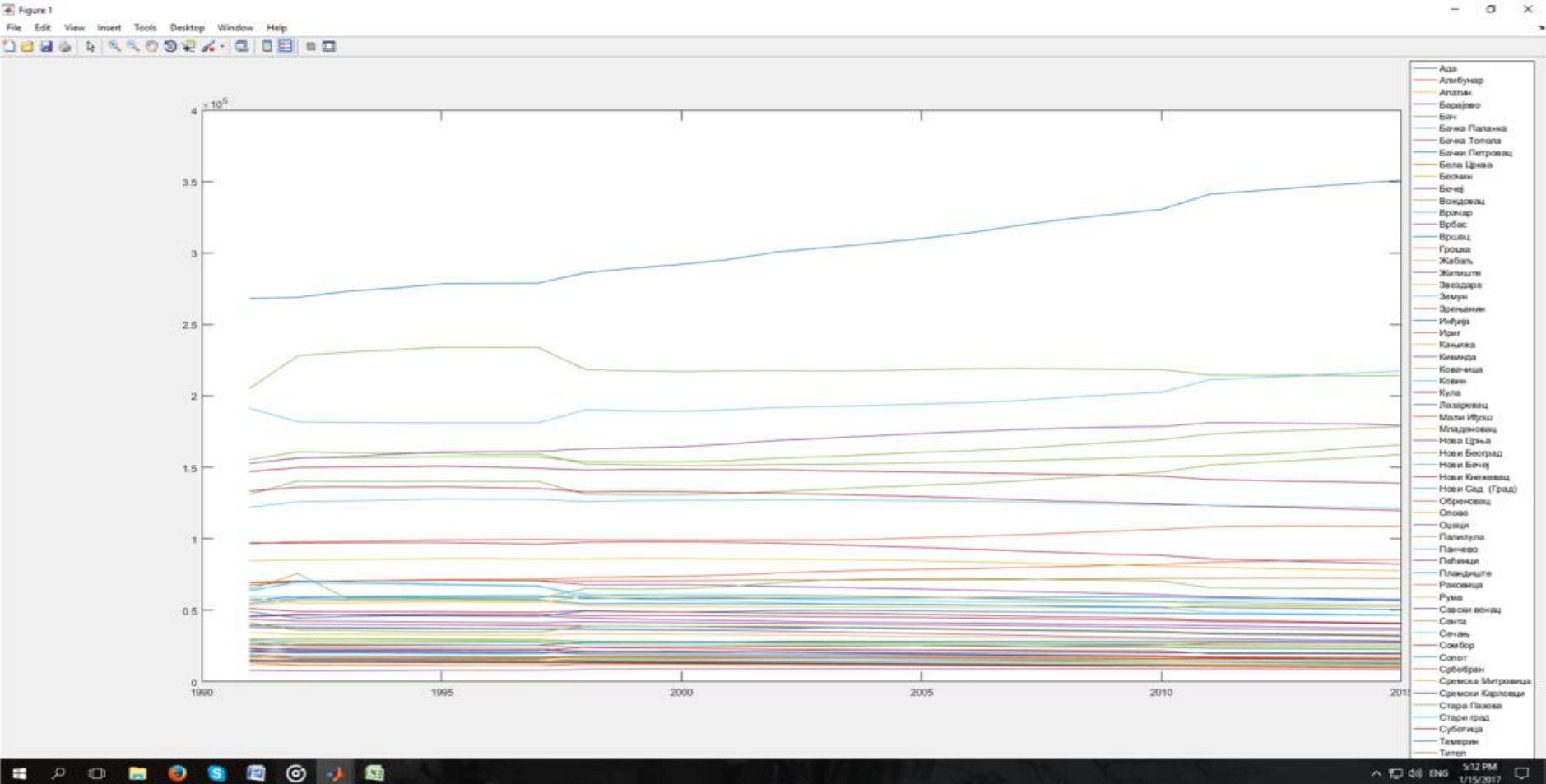
Procena broja stanovnika za opštine u Republici Srbiji

- Drugi problem, samo preuzimanje PDF dokumenata
 - Na sajtu elektronske biblioteke RZS mogu se pronaći različite vrste dokumenata kao što su obrasci, publikacije, arhivski primeri publikacija iz 19. veka.
 - Pretraga se mora vršiti ručno prema raznim meta-podacima (tipu dokumenta, godini izdanja, ključnoj reči ili statističkoj oblasti)
- Treći problem, obim PDF dokumenata.
 - Publikacija Opštine i regioni samo za jednu godinu ima više stotina strana.
 - U toj ogromnoj količini podataka je potrebno proceniti koji podaci su potencijalno korisni, ako koji ne.
- Autori su ručno filtrirali podatke na osnovu svog poznavanja demografije.

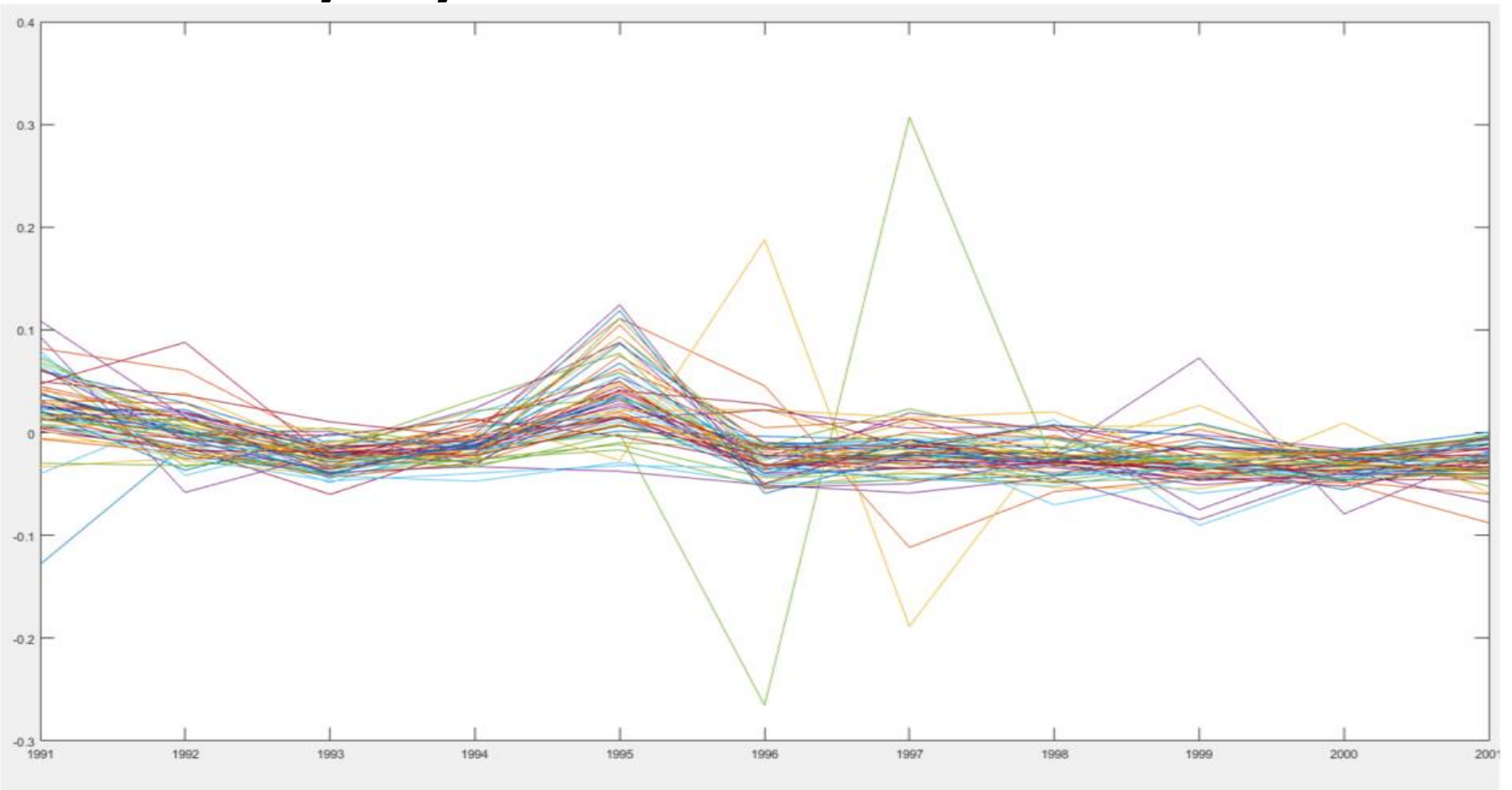
Procena broja stanovnika za opštine u Republici Srbiji

- Vizualizacija i čišćenje podataka
- Analizom vremenskih serija uočeno je da postoje godine kada su postojale velike promene broja stanovnika.
- Ratovi, migracije, promena metodologije popisa (ili procene)...
- Vremenske serije za sam broj stanovnika nisu pogodne za analizu jer je teško uočiti razlike iz godine u godinu. Zbog toga je bolje napraviti vremenske serije za procentualne promene po godinama.

Vremenske serije broj stanovnika



Vremenske serije broj stanovnika



Procena broja stanovnika za opštine u Republici Srbiji

- Kometari vezani za prethodni grafikon
- Tri velike anomalije – za 1991, 1998. i 2011. godinu.
- Da bi se ove anomalije razumele, treba biti upoznat sa činjenicom da je 2002. godine došlo do promene načina računanja broja stanovnika.
- Najbitnija promena je da su ranije u stanovništvo uključivana i lica koja su boravila u inostranstvu duže od godinu dana, a od 2002. godine ne.

Procena broja stanovnika za opštine u Republici Srbiji

- Podaci za 1991. godinu su preuzeti iz jedne od novijih RZS-ovih publikacija, i to je procena broja stanovnika po novoj metodologiji koja isključuje lica koja rade u inostranstvu, i uključuje izbegla lica u stanovništvo.
- Ova procena nije dobra, jer uzrokuje nagle skokove i padove na grafiku.
- (primeri za to su mesta koja su bila pogođena izbegličkom krizom 1990-ih, na osnovu ove procene bi se moglo zaključiti da su izbeglice došle pre 1991. što nije tačno)
- 1998. godine je procena rađena na osnovu pomenute nove metodologije
- 2011. godine je sproveden popis i RZS je tada prilagodio svoje procene podacima sa tog popisa.

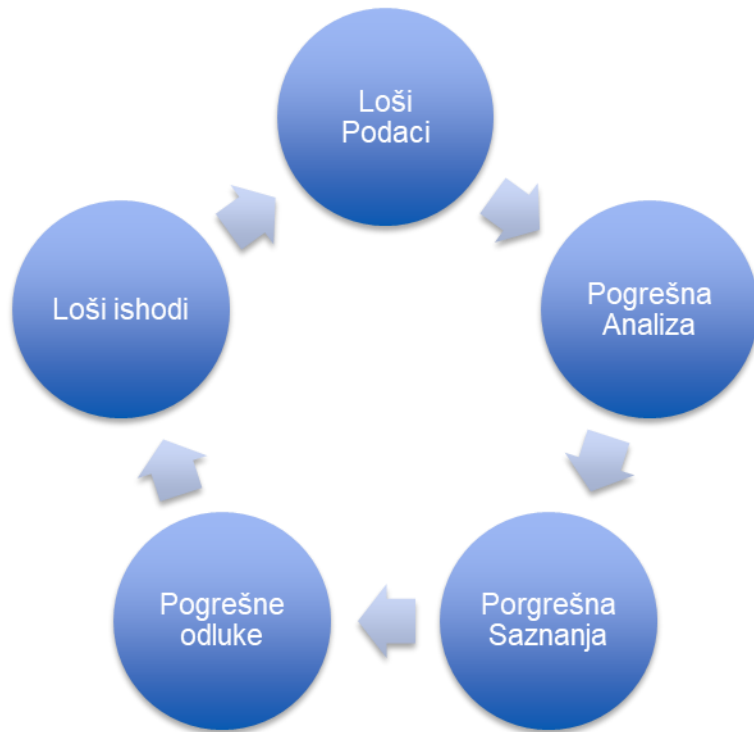
Procena broja stanovnika za opštine u Republici Srbiji

- Rezime:

- Potrebno je izdvojiti jako puno pažnje i vremna za razumevanje podataka
- Potrebno je i domensko znanje i dodatno istraživanje
- Rezultat: kvalitetni podaci i dobri modeli
- Npr. Posebno modelovanje za 1991, 1998. i 2011

Procena broja stanovnika za opštine u Republici Srbiji

- Bez prethodno opisanog data wrangling procesa rezultati projekta bili bi loši



Klasterovanje heroja u *League of Legends*

- Autor: Milan Keča
- Cilj:
 - klasterovanje heroja u LOL na osnovu načina na koji su se heroji igrali.
 - Za prikupljanje skupa podataka korišćen je LOL API.
 - Za klasterovanje korišćeni su K-Means++, Spectral Clustering i EM algoritmi.

Klasterovanje heroja u *League of Legends*

- Izvori podataka: LOL API
- Svi igrači koji su rangu Challenger ili Master na serveru EUW
- Za svakog igrača je uzeto poslednjih 20 mečeva > 30 min.
- Ukupno prikupljeno 477 mečeva, što znači 4770 primera igranja heroja.
- Svaki meč je imao podatke o svim igračima, ali nije imao podatke o itemima koje su igrači pravili, već samo ID svakog itema koji je igrač kupio.

Klasterovanje heroja u *League of Legends*

- Problemi sa podacima (uz terminologiju koju je koristio autor)
- Za meč nije bilo podataka o itemima koje su igrači pravili, već samo ID svakog itema koji je igrač kupio.
- Bilo je potrebno prikupiti statičke podatke o svim itemima koje API pruža.
- Problem: neki statsevi koje item pruža smatraju specijalnim efektima.

Klasterovanje heroja u *League of Legends*

- Pomoću API za iteme nije moguće uvek zaključiti koje statseve item daje.
- Na primer, jedan item daje 35 attack damage-a i 15 special effect 1.
- Special effect 1 predstavlja drugačiji stat za svaki item i nije moguće dobiti potpunu sliku o statsevima koje item pruža.
- Kako bi se ovaj problem rešio, potrebno je bilo iz same igre LOL prepisati skup podataka o itemima i jasno definisati šta itemi daju.
- (Kolega je morao da igra LOL da bi uradio projekat :))

Klasterovanje heroja u *League of Legends*

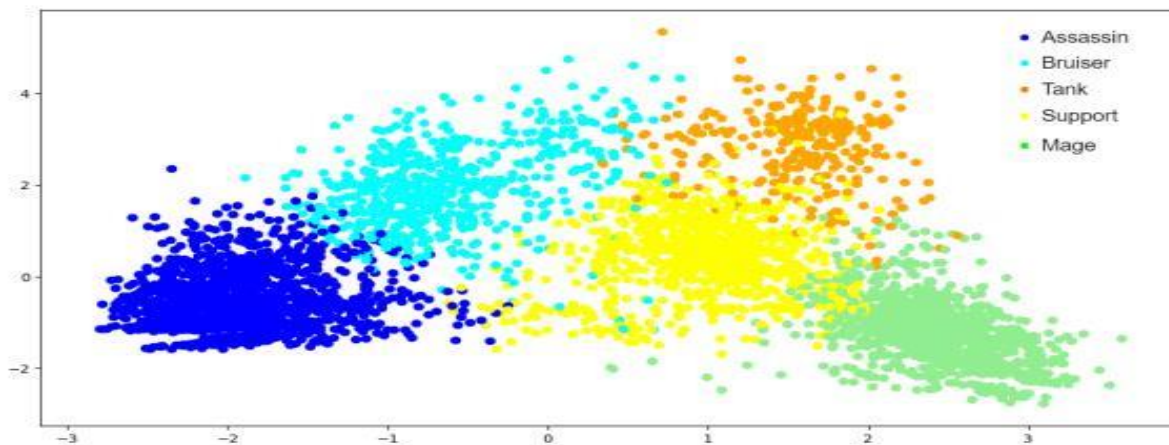
- Dalji koraci (uz terminologiju koju je koristio autor)
- za svakog igrača su izvučeni podaci koji su razmatrani, a zatim za sve iteme koje je igrač kupio izvučeni statsevi.
- Sve vrednosti su zatim skalirane u zavisnosti od vremena trajanja meča, kako bi svi igrači imali donekle ravnopravne vrednosti.
- Svaki meč sadrži podatke o meču, kao što su trajanje meča, pobednik, vreme početka, itd. Svi ovakvi generalni podaci o meču su izbačeni, jer nisu bitni za klasifikaciju heroja.

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

Klasterovanje heroja u *League of Legends*

- Rezime
- Dosta posla prilikom obrade (filtriranja) podataka
- Bez toga ne bi bilo moguće dobiti nikakve rezultate



- Detaljnija diskusija rezultata biće na predavanju o klasterovanju