Sistemi za istraživanje i analizu podataka

Uvod

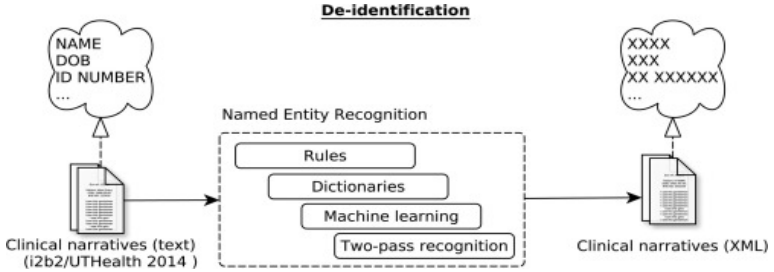
predavač: Aleksandar Kovačević

Par reči o predavaču

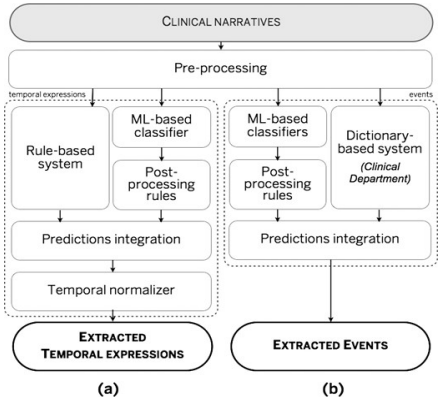
- Završio sam PMF u Novom Sadu.
- Magistrirao i doktorirao na FTN gde sam sada redovni profesor.
- Od 2007 godine bavim se mašinskim učenjem, data science, ali pre svega **NLP**.
- U nastavku prikazani su projekti na kojima sam radio.
- Cilj predavanja je, između ostalog, da Vam prenesem svoja iskustva iz prakse, pa Vas ti slajdovi mogu inspirisati na eventualna pitanja koja biste imali za mene.

NLP u medicinskom domenu na Engleskom i Srpskom jeziku

Anonimizacija medicinskih anamneza

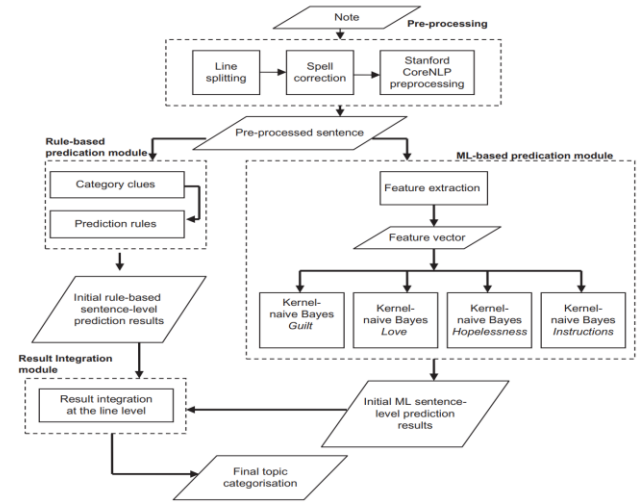


Ekstrakcija informacija iz anamneza

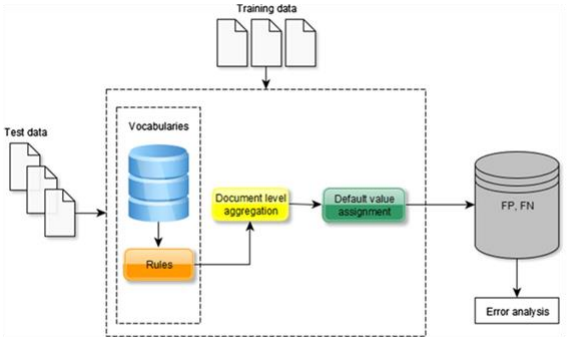


NLP

Detekcija sentimenta u psihijatriji

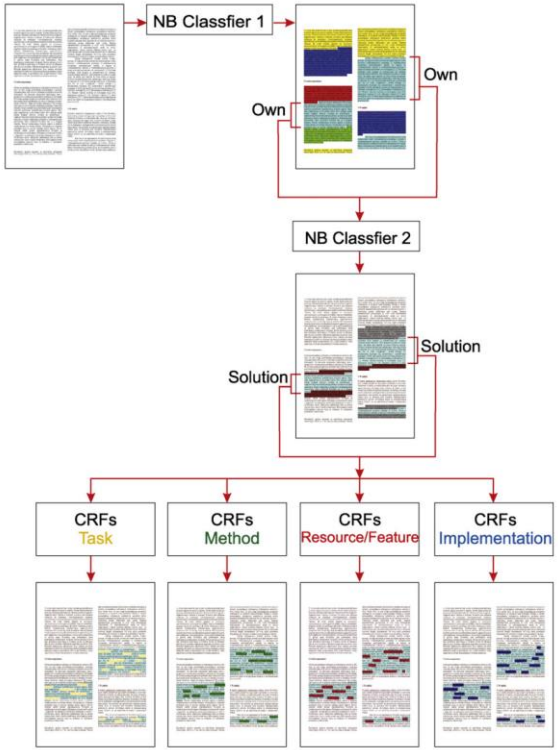


Identifikacija faktora rizika bolesti srca



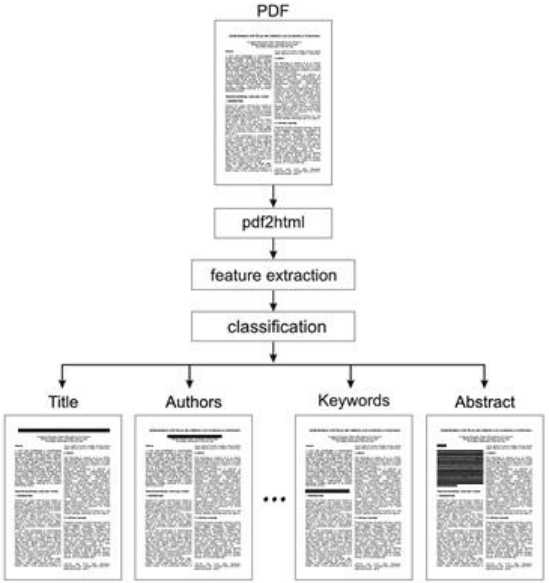
NLP naučnih radova Engleskom jeziku

Ekstrakcija metodologija iz naučnih radova



NLP

Ekstrakcija meta-podataka iz naučnih radova



Ekstrakcija bio-medicnskog softvera iz naučnih radova

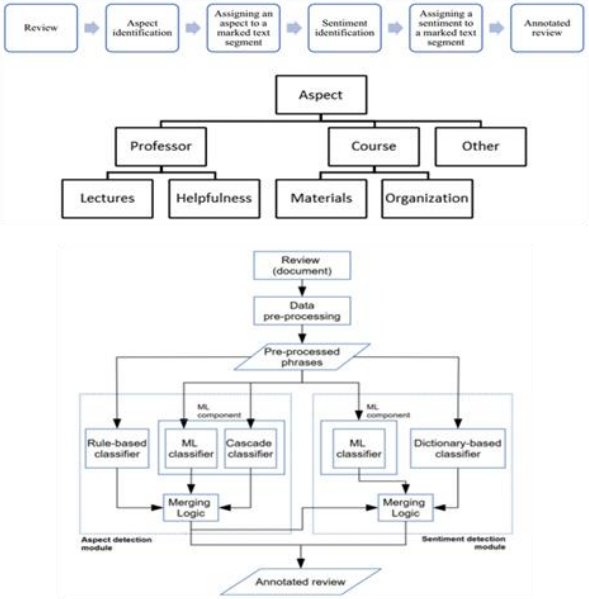
Mention Level		Document Level	
GO	2.08	R	0.29
R	1.17	GO	0.19
BLAST	0.62	BLAST	0.16
PDB	0.43	GenBank	0.13
KEGG	0.43	GEO	0.09
GenBank	0.35	KEGG	0.09
Ensembl	0.24	PDB	0.08
GEO	0.24	Ensembl	0.06
Pfam	0.20	Cluster	0.06
Cluster	0.18	UniProt	0.05

The mention level numbers provide the average mentions per document, and the document level number provides the fraction of the *bioinformatics* corpus to contain at least a single mention of that resource.

doi:10.1371/journal.pone.0157989.t005

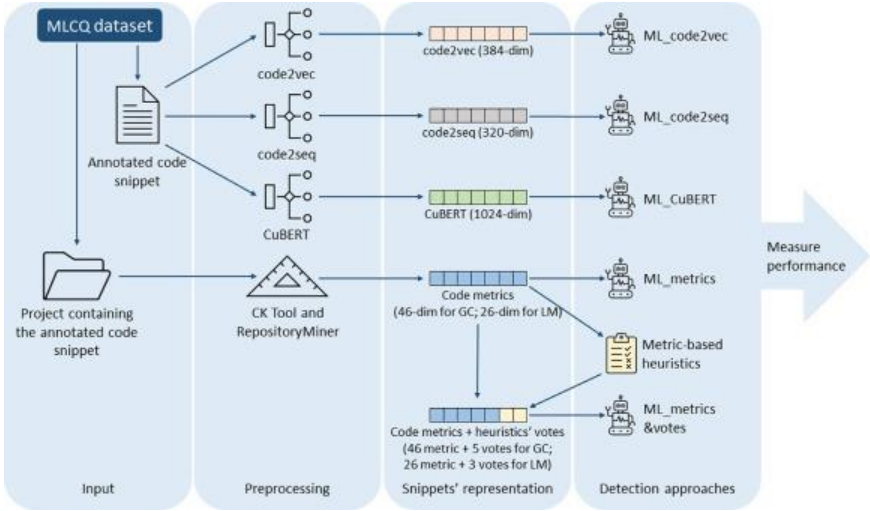
NLP drugi domeni

Detekcija sentimenta i aspekata iz studentskih komentara i anketa



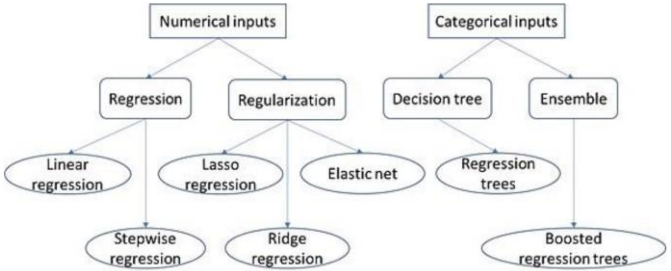
NLP

Detekcija loše napisanog koda (*code smell*) u programskim jezicima Java i C#



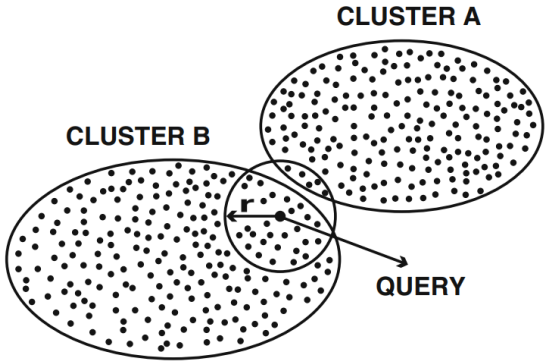
Data mining u domenima: zabave, sporta i farmacije

Analiza faktora uticaja oslobađanja aktivne supstance iz tableta



Data mining

Klasterovanje i pretraga muzike po žanru na osnovu obrade signala



Predikcija ishoda utakmica NBA lige

Terminologija

- Kada je ovaj kurs nastao (2009 godine) u svetu je bio aktuelan termin *data mining*.
- Termin *data mining* je vremenom evoluirao u termin *data sicence*, odnosno nauka o podacima.
- *Data science* je relativno nedefinisan termin (kao što ćete videti u nastavku prezentacije) ali je takođe i opšte poznat termin, pa ćemo ga iz tog razloga koristiti na ovom krusu.

Šta je *data science*?

- Postoji mnogo definicija, ali nijedna nije univerzalna.
- Oblast je multi-disciplinarna pa nije relano očekivati da će ikada doći do dogovora oko jedne definicije.
- U nastavku ćemo prvo prikazati nekoliko definicija.
- Nakon toga ćemo istaći razlike između *data science* i srodnih disciplina kao što su mašinsko učenje, statistika i procesiranje velikih količina podtaka.

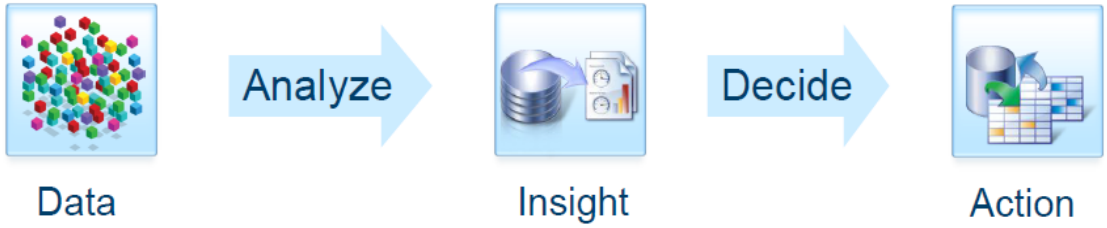
Šta je *data science*?

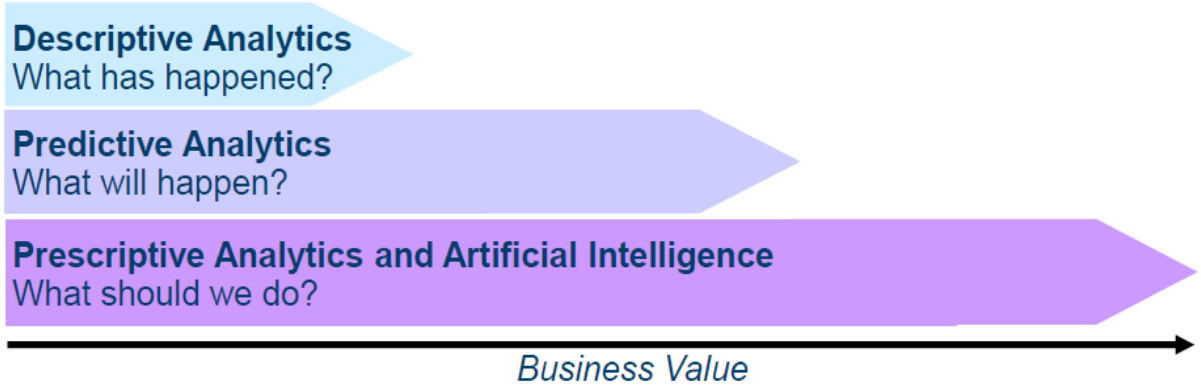
*„...proces **čišećenja**, **transformacije**, **istraživanja** i **modelovanja** podataka sa ciljem da se **otkriju korisne informacije** koje pomažu u procesu donošenja odluka...“*

Wikipedia

Šta je *data science*?

„*Proces **izvođenja novih zaključaka** ili **uvida iz podataka** sa ciljem **donošenja odluka**. “*





Oleksandr Romanko, University of Toronto

Šta je *data science*?

(1) Prikupiti podatke

pomoću računara, senzora, ljudi,…

(2) Uraditi nešto (korisno) sa podacima

doneti odluke, potvditi hipotezu, dobiti novi uvid, predvideti budućnost

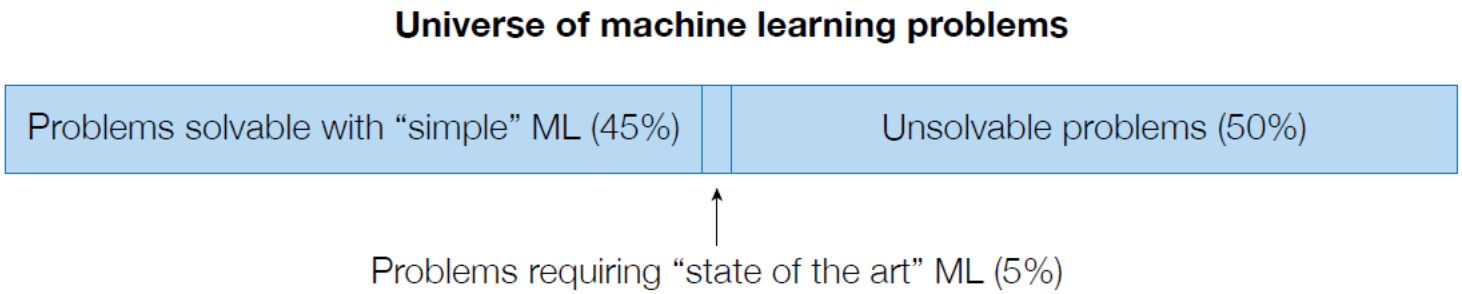
*Data Science = Preći iz **(1)** u **(2)***

Data science vs Mašinsko učenje (ML)

- *Data science* nije (samo) mašinsko učenje.
- Mašinsko učenje koristi računarstvo i statistiku, ali tradicionalno nema za cilj dobijanje novih uvida u podatke.
- ML je veoma fokusirano na algoritme i njihovo međusobno poređenje.
- Primena ML je značajan deo data science, ali ponekad je nabolji način da se reši neki problem iz realnog sveta jednostavna vizualizacija podataka.

Data science vs Mašinsko učenje (ML)

- Mašinsko učenje, iskustva iz industrije*:*



J. Zico Kolter, Carnegie Mellon University

Data science vs Statistika

- Debata oko razlika između *data science* i statisitke se još uvek vodi.
- Razlog je to što su statističari, zahvaljujući medijskom preterivanju sa terminom *data science*, praktično preko noći izgubili svoj identitet.
- Iako se debata nikad neće završiti, kao što ćete videti u nastavku, neke razlike postoje.

Data science vs Statistika

- Tradicionalna statistika
  - Tradicionalno počinje sa konkretim ciljem, opservacijom iz realnog sveta, odnosno hipoteznom koja bi trebalo da se testira.
  - Podaci se prikupljaju jasno definisanim procesom koji je vezan za hipotezu (npr. kliničke studije za testiranje lekova).
  - U tom smislu statistika je više orijentisana ka hipotezi i samim statističkim alatima tj. modelima (model-oriented), a ne podacima.
- Data science
  - Vrlo malo (ili bez) hipoteza o podacima.
  - Podaci su obično unapred dostupni (ne prikupljaju se nakon hipoteze).
  - Analiza je obično vođena podacima (data-driven), a ne hipotezom.
  - Malo statističkih kurseva se bavi prikupljanjem podataka sa weba ili predprocesiranjem podataka.
- i najvažnija razlika „Statističari koriste R, a data scientisti Python....“

Data science vs Obrada velikih količina podataka

- Obrada velikih količina podataka je oblast koja se fokusira na skladištenje i procesiranje podataka.

- Procesiranje u ovom kontekstu znači relativno površna analiza,

  - Na primer praćenje saobraćaja na nekom web sajtu pomoću kontrolne table (*dashboard*).

- Ponekad, da bi stvarno razumeli neki fenomen ili rešili neki problem iz realnog sveta, potrebne su ogromne količine podataka.

- Međutim, vrlo često ali mnogo češće dovoljan je samo laptop, SQL i Python.

Ko je *data scientist*?

- Termin *data scientist* su 2008 godine osmislili *DJ Patil* (tada u *LinkedIn-u*) *i Jeff Hammerbacher* (*tada u Facebook-u*).

- Njihov cilj je bio da daju naziv novom tipu posla (profilu osobe) koji je počeo da se pojavljuje u silinkonskoj dolini.

Izvor: “Doing Data Science“, Rachel Schutt and Cathy O’Neil, OReilly

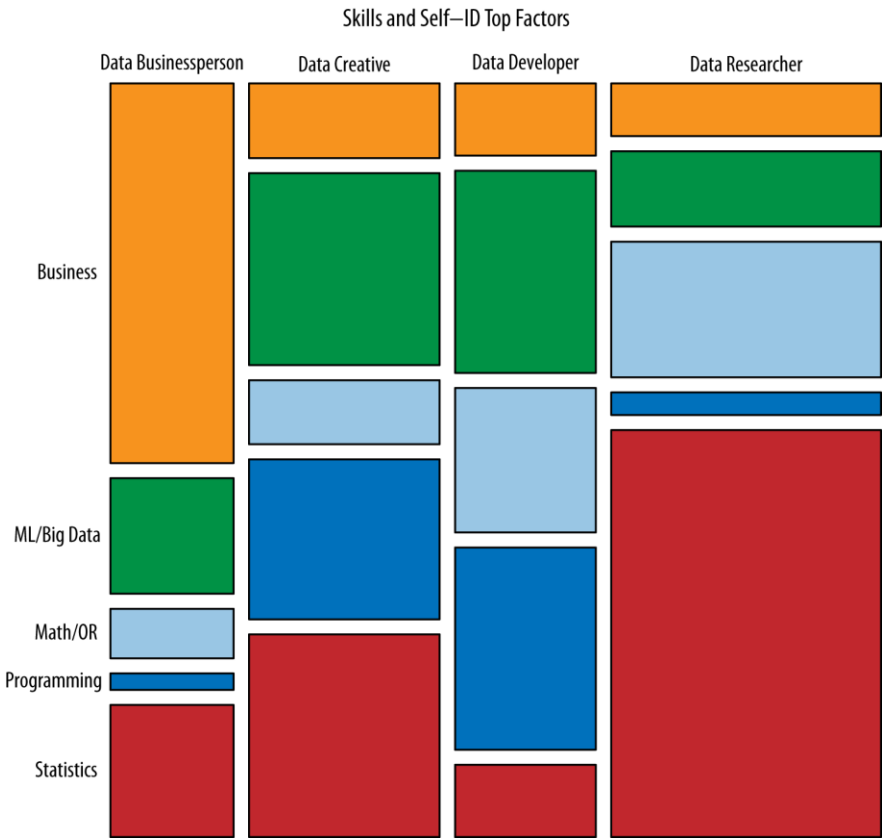
Ko je *data scientist*?

- „Analitičar podataka koji živi u Kaliforniji.”
- „…bili ko, ko radi sa podacima u nekoj organizaciji…“
- „…**redak hibrid**, programera koji ume da prikupi i obradi podatke iz različitih izvora i statističara koji ume da otkrije novo znanje iz tih podataka...“
- „…neko ko može da prikupi, očisti, istraži, modeluje i interpretira podatke, kombinujući programiranje, stastistiku i mašinsko učenje...“

Izvor: http://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/

Ko je *data scientist*?

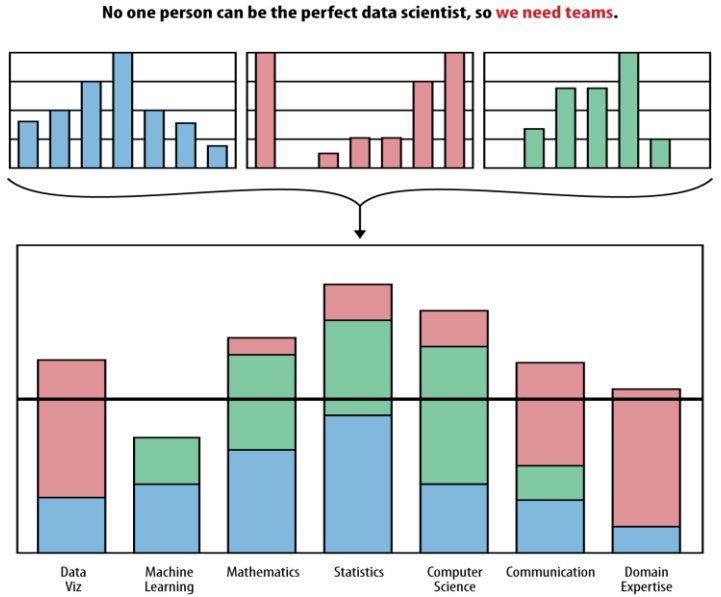
- „Analyzing the Analyzers“
  - anketirano je 250 *data sicentista* sa ciljem da utvrdi tačan opis profila *data scientista*.
- Rezultati pokazuju da ne postoji jedan jasno definisan uzak profil.
- Profil u obliku slova T.



Izvor: Jeffrey Stanton et al, Interdisciplinary Data Science Education, http://pubs.acs.org/doi/abs/10.1021/bk-2012-1110.ch006

Ko je *data scientist*?

- Profil u obliku slova T: širok spektar znanja (veština), ali ekspertiza samo u jednom polju.
- Suština je u tome da su potrebni *data science* timovi, a ne jedna osoba koja može da uradi sve.
- Profil tima zavisi od profila *data science* problema.



Izvor: “Doing Data Science“, Rachel Schutt and Cathy O’Neil, OReilly

Da li ima smisla biti *data scientist*?

MENU

Harvard Business Review

DATA

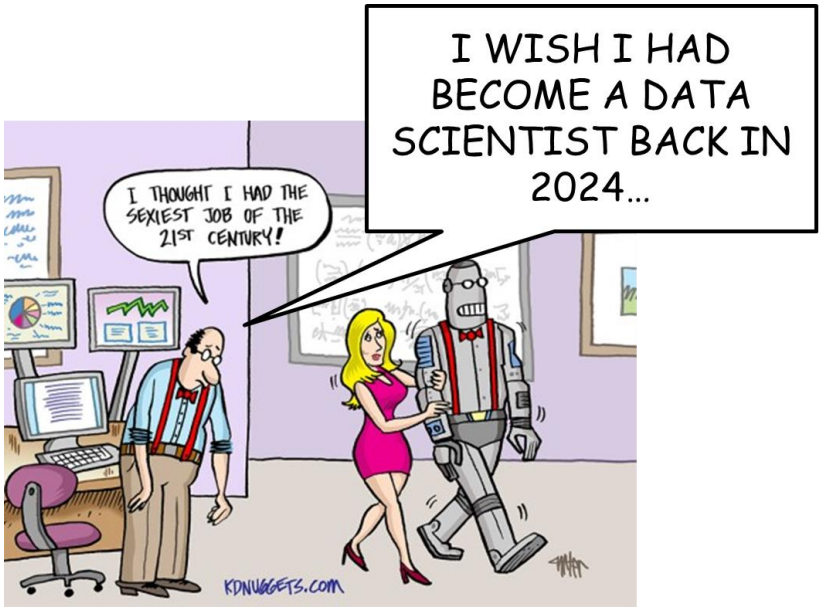
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

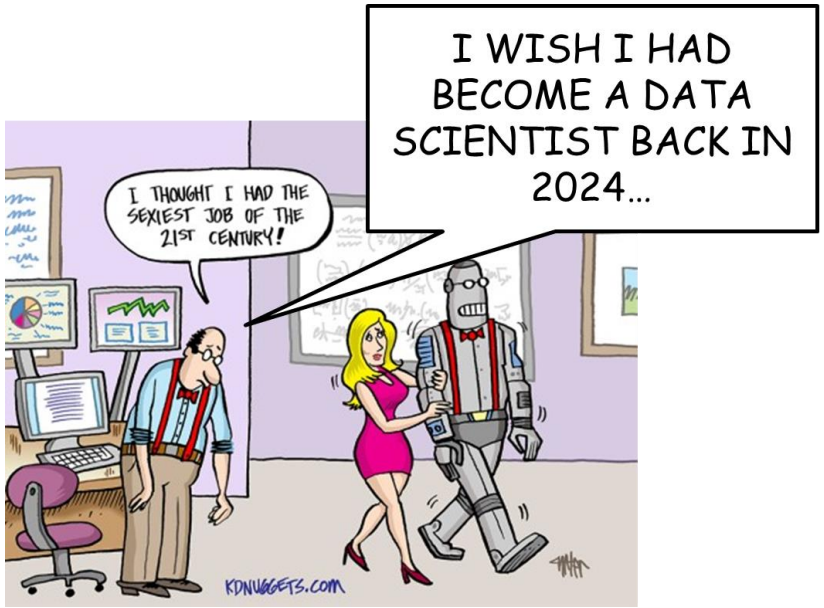
Da li ima smisla biti *data scientist u* 2024. godini?

- Po mom skromnom mišljenju, da.
- LLMs ne mogu da automatizuju kompletan data science projekata (još uvek).
- Mogu da automatizuju neke aspekte:
  - Elementarno čišćenje podataka
  - Prikaz sumarnih statistika i jednostavnih grafikona
  - Generisanje koda za isprobavanje predikitvnih modela....
- Ali....



Da li ima smisla biti *data scientist* 2024. godini?

- Ali, (još uvek) ne mogu da:
  - Uvide kompleksnije veze na grafikonima.
  - Podignu performanse prediktivnih modela npr. analizom grešaka, kada se iscrpe sve standardne tehnike.
  - Analiziraju rezultate klasterovanja kao što može iskusan data scientist.
  - Odaberu pravu tehniku i parametre za redukciju dimenzionalnosti da biste klijentu (ili sami sebi) prikazali kompleksan skup podataka
  - ....



- Promena će sigurno biti ali znanje i iskustvo koje ćete steći do tada biće Vam velika prednost.

Sadržaj predmeta

- U skladu sa „T-profilom“ **ovaj predmet se bavi širinom, a ne dubinom**.
- **„Koje metode, principi i alati postoje i kada ih i kako primeniti?“**, a ne „Kako mogu postati stručnjak u dubokom učenju za računarsku viziju u kontekstu detekcije objekata na slikama?“
- *Data science* je brzo razvijajuće i promenljivo polje.
- Opsesija jednim alatom ili tehnikom neće se isplatiti za nekoliko godina, ako ne i meseci.
- Budite spremni da istražujete i nastavite sa učenjem samostalno.
- **Cilj ovog kursa**: Omogućiti Vam da realizujete celokupan *data science* projekat od početka do kraja.

Sadržaj predmeta

- Prikupljanje i priprema podataka (*data wrangling*)
- Vizualizacija i eksplorativna analiza podataka
- Prediktivno modelovanje
- Procesiranje teksta (NLP)
- Procesiranje vremenskih serija
- Redukcija dimenzionalnosti

Sadržaj predmeta - napomene

- Kod predikitvnog modelovanja **fokus <u>nije</u> na teorijskom objašnjavanju tehnika.**
- Kod **nadgledanih** tehnika fokus je na:
  - Načinima za prikupljanje korpusa (*dataset*)
  - Procesu označavanja (anotiranja) korpusa i evaluaciji kvaliteta oznaka
  - Radu sa realnim skupovima podataka gde balans oznaka klasa nije ravnomeran
  - Analizi grešaka i poboljšanju kvaliteta modela
- Kod **nenadgledanih** tehnika fokus je na:
  - Odabiru pravilnih metoda
  - Evaluaciji rezultata
  - Tumačenju i sapoštavanju rezultata u vidu izveštaja koji je razumljiv klijetnu, npr. profilisanje mušterija prodavnice.

Polaganje predmeta

- Ispit sadrži dva dela:

1. Teorijski

2. Praktični

Polaganje predmeta – Teorijski deo

- Sastoji se od pitanja iz gradiva obrađenog na predavanjima

- Polaže se usmeno ili pismeno
  - Bićete obavešetni na vreme

Polaganje predmeta – Praktični deo

- Realizuje se kroz projekat
- Teme projekta sami birate uz kontrolu nastavnika
- Postoji više važnih tačaka tj. datuma koji su dati u nastavku.

Praktični deo, Projekat, Rokovi :

Prijava timova i (okvirnog) naslova teme	10.11.2024.	Na sajtu predmeta biće objavljen *google sheet* u koji treba da se upišete
Predaja predloga projekta	24.11.2024.	Asistentu na mail
Rok do koga ćete dobiti komentare za ispravku	08.12.2024.	\
Predaja revidiranog predloga projekta	22.12.2024.	Asistentu na mail
Rok do koga dobijate notifikaciju da li je projekat prihvaćen	29.12.2024.	\
Prva kontrolna tačka	Okvirno 20.01.2025.	Usmene konsultacije sa asistentom u zakazanom terminu
Druga kontrolna tačka	Okvirno 20.02.2025.	Usmene konsultacije sa asistentima i profesorom u zakazanom terminu
Predaja finalne verzije projekta	20.03.2025.	Asistentu na mail

Polaganje predmeta

- Ako ne ispoštujete datum predaje predloga projekta možete dobiti maksimalno ocenu 6 iz praktičnog dela.

- Za svaku kontrolnu tačku koju ne ispoštujete ocena iz praktičnog dela Vam se smanjuje za 1.

- Ako propustite jednu kontrolnu tačku maksimalna ocena iz praktičnog dela Vam je 9.

- Ako propustite obe kontrolne tačku maksimalna ocena iz praktičnog dela Vam je 8.

Organizacija kursa - Predavanja

- Učenje metoda za analizu podataka na dva načina:
  - Pomoću slajdova
  - Pomoću demonstrativnih primera

- Demostrativni primeri na predavanjima biće prikazani pomoću:
  - Jupyter notebook u Python jeziku
    - Pandas, scikit-learn, XGBoost, PyTorch....

Organizacija kursa – vežbe

- Na vežbama zajedno sa asistentima prolazite kroz proces prijave projekta, rada na projektu i odbrane projekta.
  - Nakon toga će asistenti pratiti vaš rad na projektu, dok će odbrana biti u drugom semestru.
  - O svim detaljima bićete obavešteni na posebnom predavanju.
- Prisustvo nije obavezno, ali ćete vam rad na projektu biti značajno lakši ako budete dolazili.

Organizacija kursa – vežbe

- Vežbe počinju od ponedeljka 28.10.2024.
- Termini vežbi:
  - Ponedeljak 18:00 - NTP 222
  - Četvrtak 19:45 - NTP 222
- Učionica može da primi otrpilike polovinu vas.
- Mi ćemo vas podeliti po broju indeksa, a ako želite možete se zameniti sa nekim i organizovati kako vam odgovara.

Pitanja?

Da li je *data „science”* ili *data sicence*?

- Svaka nauka jeste (ili bi bar trebalo da bude) zasnovana na podacima.

- Šta je drugačije kod *data science* (nauke o podacima) u odnosu na druge nauke?

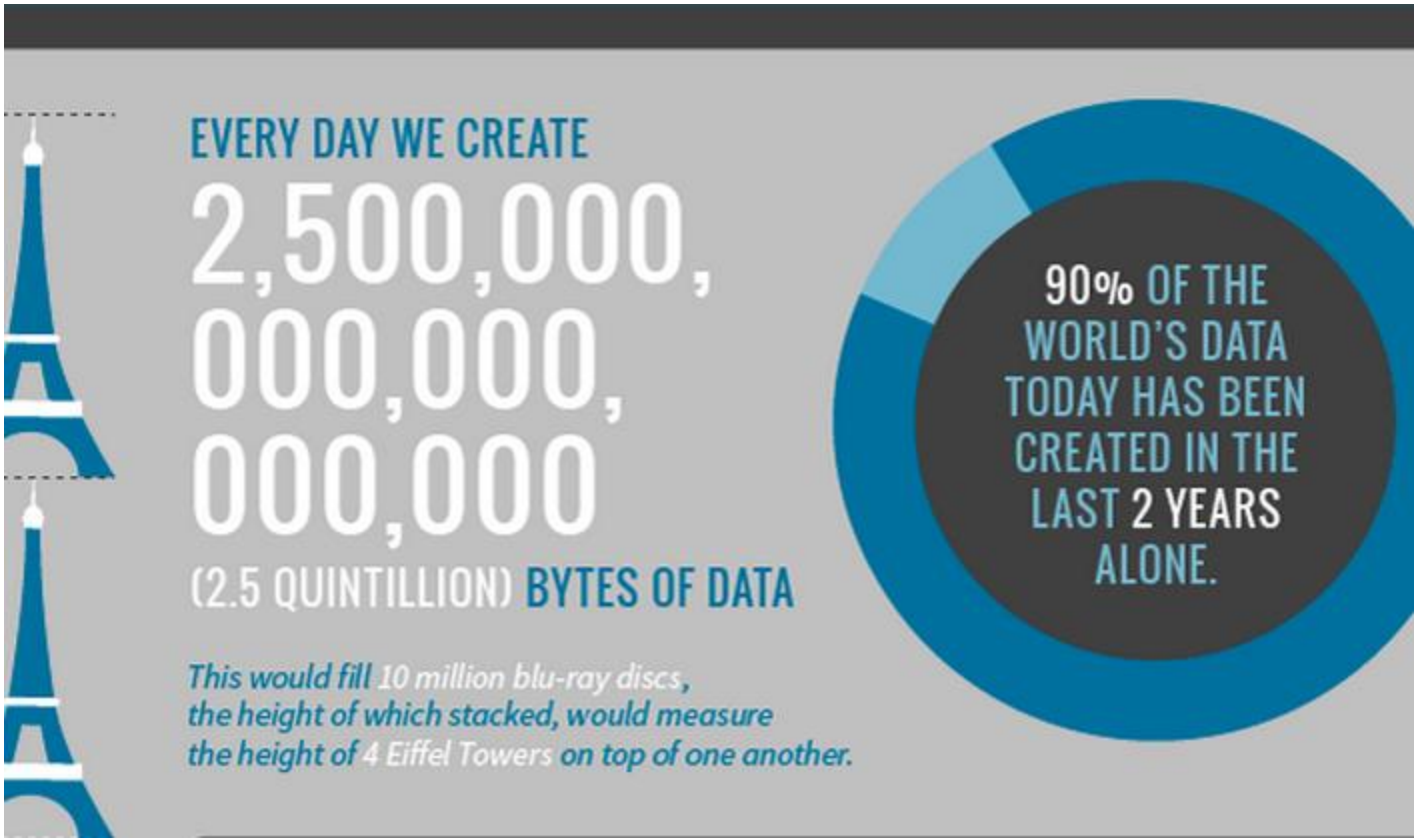
- Koje su okolnosti uopšte dovele do potrebe za naukom kod koje su podaci centralni?

Ogromne količine podataka (naglo) postaju dostupne

"Between the dawn of civilization and 2003, we only created **five exabytes** of information; now [2010] we're creating that amount **every two days.**"

Eric Schmidt, Google (2010)

Ogromne količine podataka postaju dostupne

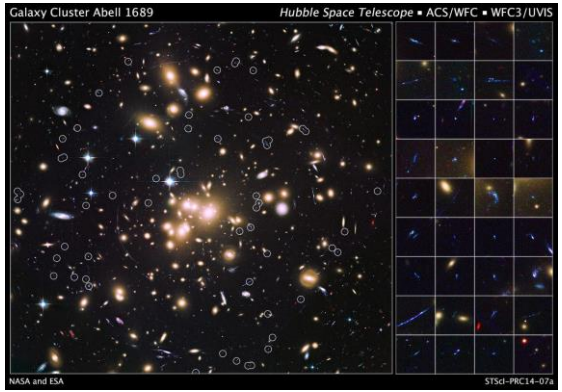


Izvor: https://wonderfulengineering.com/this-is-the-amount-of-data-that-is-being-generated-in-the-world-every-minute/

Ogromne količine podataka postaju dostupne

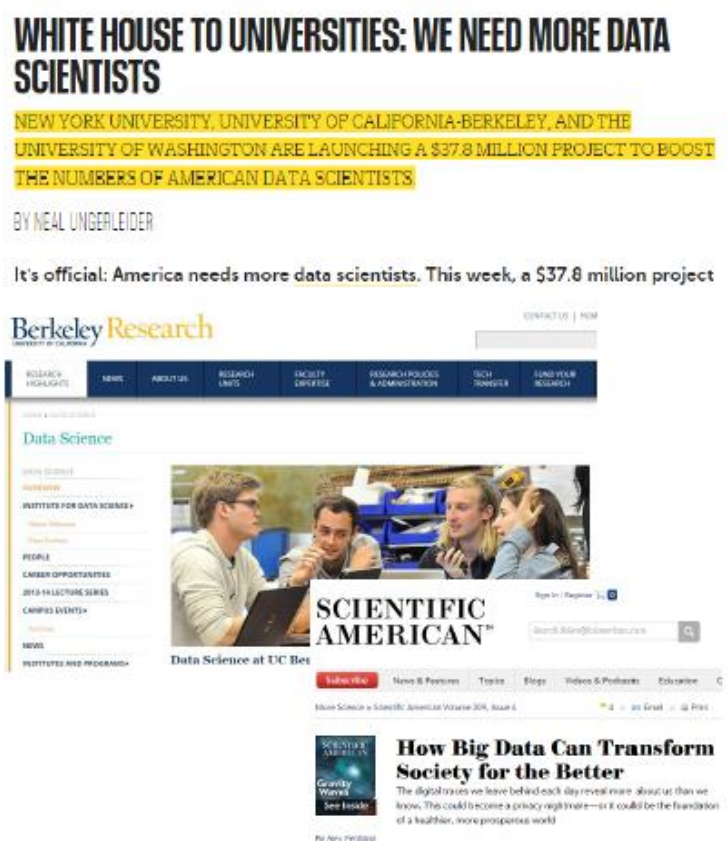
Izvor: https://ediscoverytoday.com/2023/04/20/2023-internet-minute-infographic-by-ediscovery-today-and-ltmg-ediscovery-trends/

Ogromne količine različitih vrsta podataka postaju dostupne

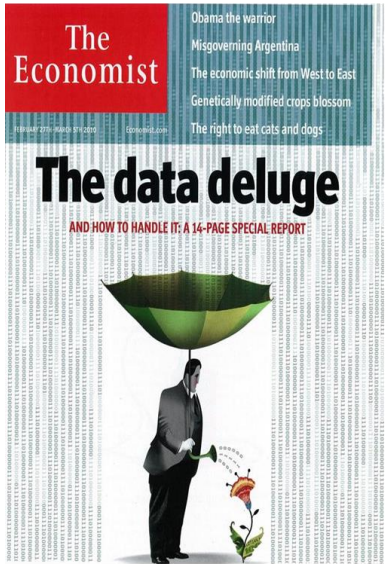


Text (web strane, društvene mreže, naučni radovi, interna dokumenta kompanija, e-mail), **grafovi** (web, društvene mreže, semantički web), **slike, video, audio, mape, logovi**, …

Države, komapanije i naučnici širom sveta shvataju značaj upotrebe dostupnih podataka



Države, komapanije i naučnici širom sveta shvataju značaj upotrebe dostupnih podataka



“Data is the new oil”

Collecting Big Data is challenging, but it’s nearly not enough

Value comes from the insights

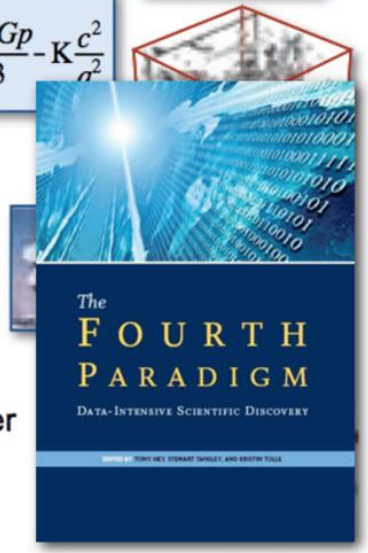
The “Internet” companies understood this paradigm perfectly (cfr. Google, Facebook, Netflix, etc.)

Nastaje potreba za novom naučnom metodologjiom

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



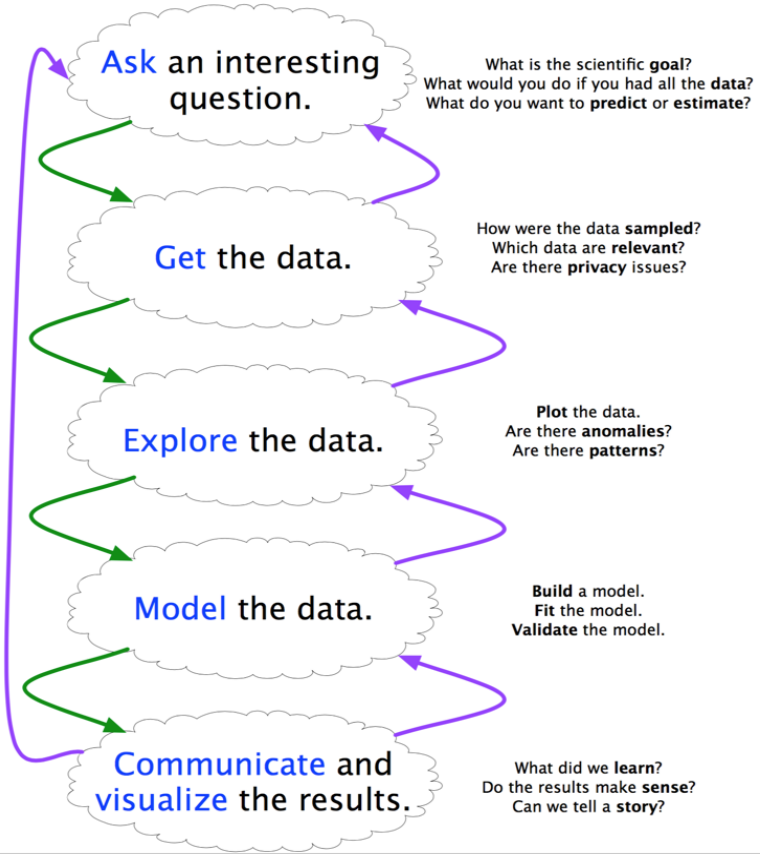
Jim Gray, Microsoft

Nauka 2.0 = sistematizacija *data science* procesa

- Nauka 1.0:
  - Fokusirana na opservacijama od kojih se formiraju hiopteze
  - **Hipoteza je centralna**, oko nje se prikupljaju podaci

- Nauka 2.0:
  - **Podaci postaju centralni**
  - Istraživanjem podataka formiraju se hipoteze

Data science proces kao naučna metodologija:



Primer nauke 1.0 – Otkriće planete Neptun

- **Opservacija**: Početkom 19. veka, astronomi su primetili da orbita Urana nije bila u skladu sa očekivanjima.

- **Pitanje**: Zašto je orbita Urana odstupala od predviđenog puta?

- **Hipoteza**: Odstupanja u orbiti Urana mogla su biti uzrokovana gravitacionim uticajem neotkrivene planete.

- **Eksperimentisanje**: Umesto tradicionalnog eksperimenta, ovo je uključivalo matematičke proračune i predikcije. Dva astronoma, Urbain Le Verrier iz Francuske i John Couch Adams iz Engleske, nezavisno su izračunali poziciju na kojoj bi ta nova planeta trebala da bude na osnovu gravitacionih uticaja.

- **Prikupljanje podataka**: Posmatranje noćnog neba kako bi se locirala predviđena pozicija nove planete.

Primer nauke 1.0 – Otkriće planete Neptun

- **Analiza podataka**: Poređenje pozicije utvrđene posmatranjem sa pozicijom koja je rezultat matematičkih modela.

- **Zaključak**: Nova planeta je pronađena na predviđenoj lokaciji, što je dovelo do otkrića Neptuna 1846. godine.

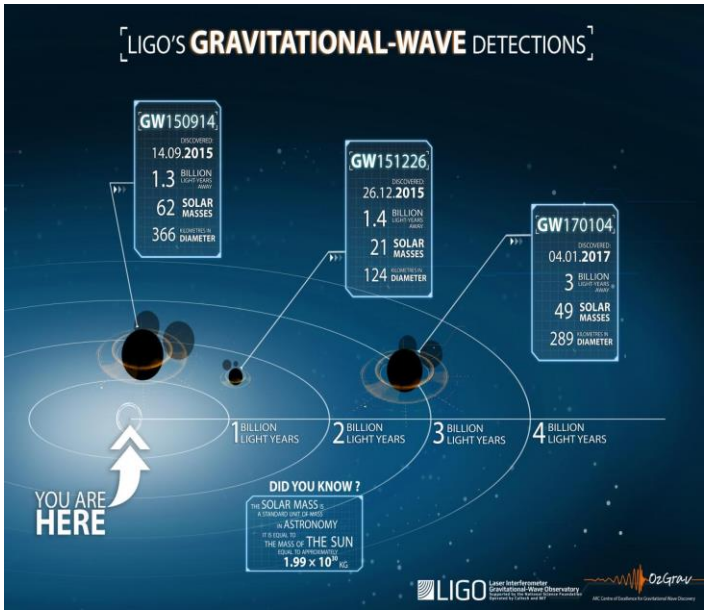
- **Izveštavanje**: Nalazi su objavljeni i potvrđeni dodatnim posmatranjima i recenzijama.

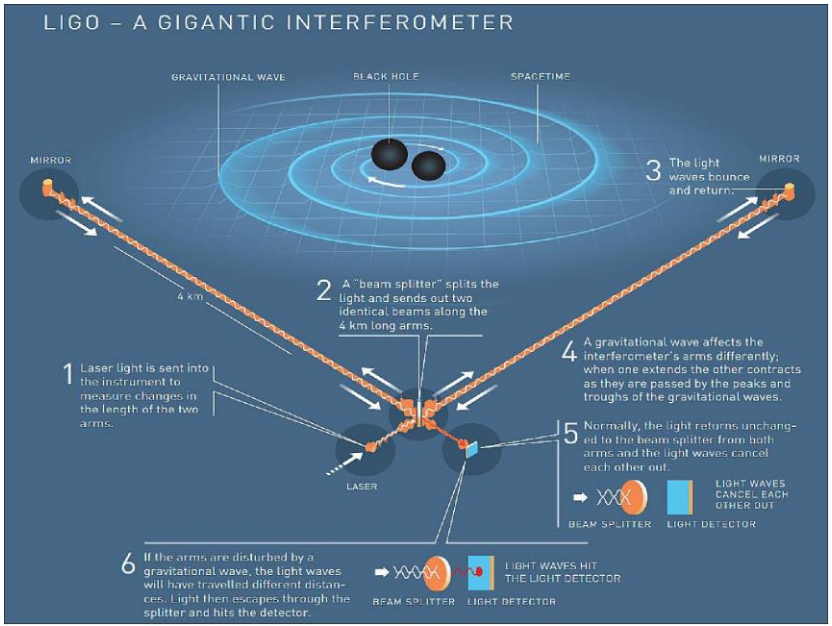
- **Primetite**: Opservacija i hipoteza su centralne, podaci su prikupljeni sa namerom da se potvrdi hipoteza.

Primer nauke 2.0 – LIGO, potraga za gravitacionim talasima

The Laser Interferometer Gravitational-Wave Observatory (LIGO)

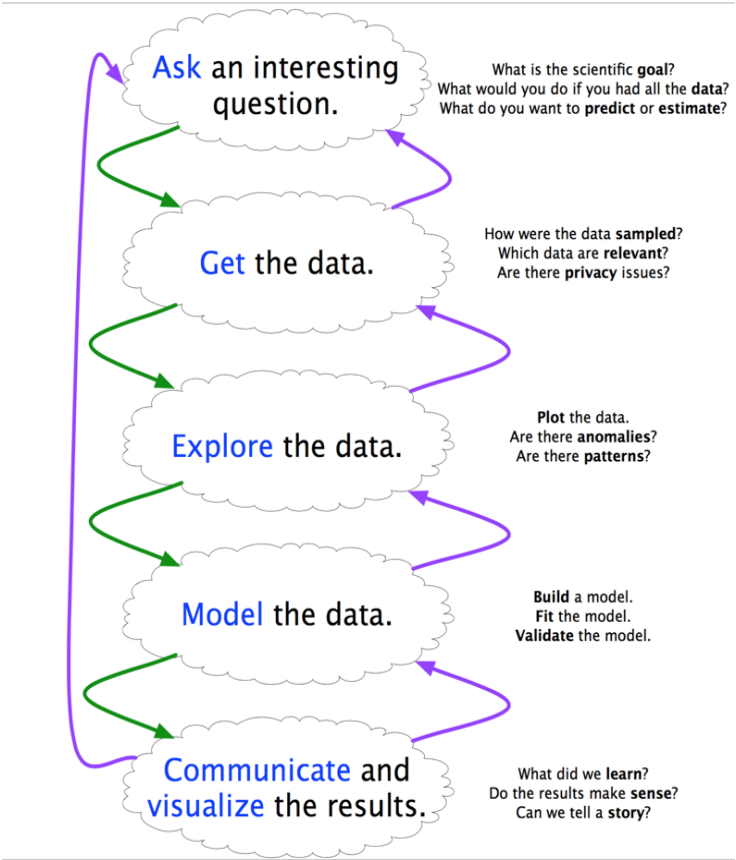
- Pretraga za gravitacionim talasima uključuje analizu podataka sa Laserskog Interferometra za Gravitacione talase (LIGO), koji detektuje male promene u udaljenosti izazvane prolaskom gravitacionih talasa.





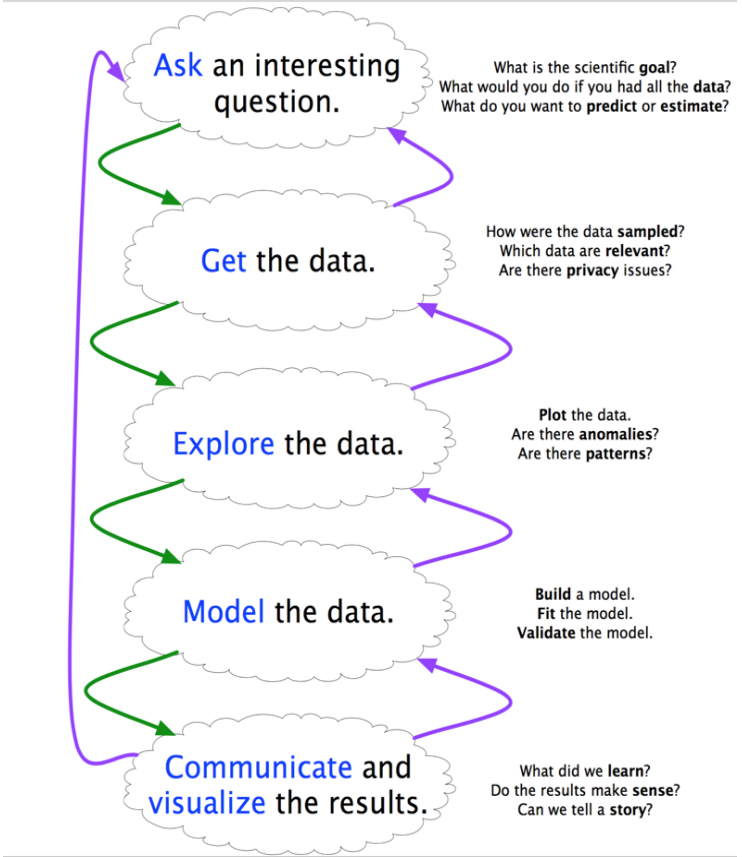
Primer nauke 2.0 – LIGO, potraga za gravitacionim talasima

- **Pitanje**: Da li stvarno postoje gravitacioni talasi?
- **Prikupljanje podataka**: LIGO prikuplja ogromne količine podataka pomoću osetljivih interferometara, koji mere minijaturna pomeranja laserskih zraka kada gravitacioni talasi prođu korz njih. **1 TB sirovih podataka dnevno!**
- **Čišćenje i obrada podataka**: Sirove podatke treba očistiti i obraditi kako bi se filtrirao šum i drugi nerelevantni signali.
- **Eksplorativna analiza podataka (EDA)**: Fizičari i *data scientisti* koriste statističke alate za istraživanje i vizualizaciju podataka, tražeći anomalije ili obrasce koji bi mogli ukazivati na prisustvo gravitacionih talasa.



Primer nauke 2.0 – LIGO, potraga za gravitacionim talasima

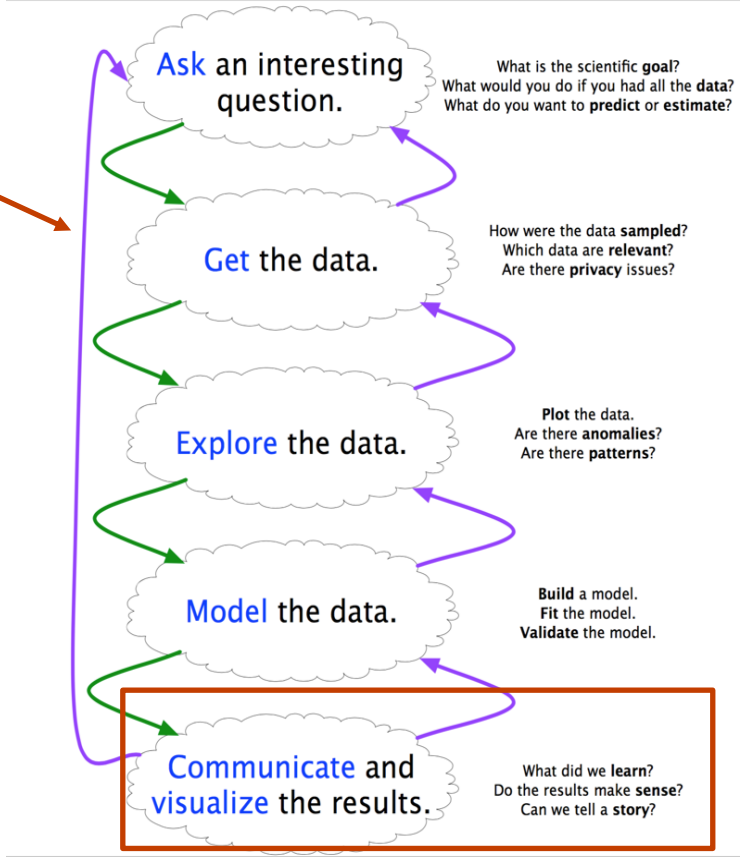
- **Kreiranje modela**: Modeli mašinskog učenja i statističke tehnike primenjuju se kako bi se razlikovali potencijalni signali gravitacionih talasa od šuma.
- **Validacija**: Identifikovani signali se upoređuju sa poznatim izvorima i obrascima šuma kako bi se proverila njihova autentičnost.
- **Tumačenje**: Nakon validacije, podaci se tumače kako bi se razumela svojstva gravitacionih talasa i događaji koji su ih proizveli, kao što su spajanja crnih rupa.



Primer nauke 2.0 – LIGO, potraga za gravitacionim talasima

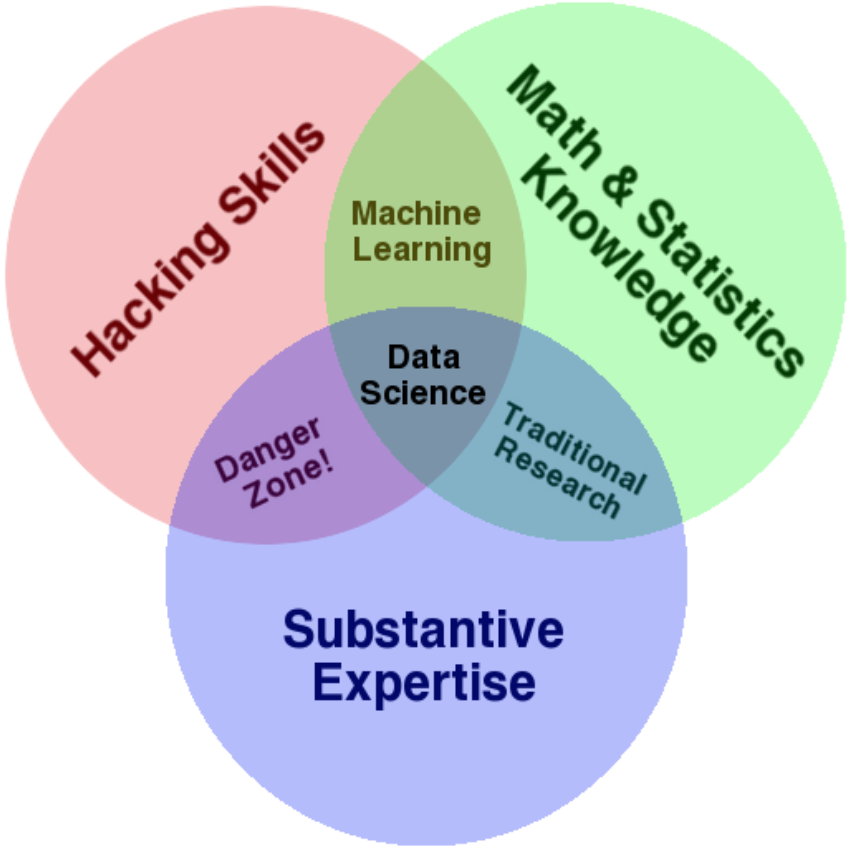
- **Iteracija**: Proces uključuje kontinuirano usavršavanje modela i tehnika dok se prikupljaju i analiziraju novi podaci.

- **Komunikacija**: Rezultati istraživanja dele se putem publikacija, prezentacija i interaktivnih alata sa naučnom zajednicom radi dalje validacije i istraživanja.



Naučnik 2.0 = *data scientist*

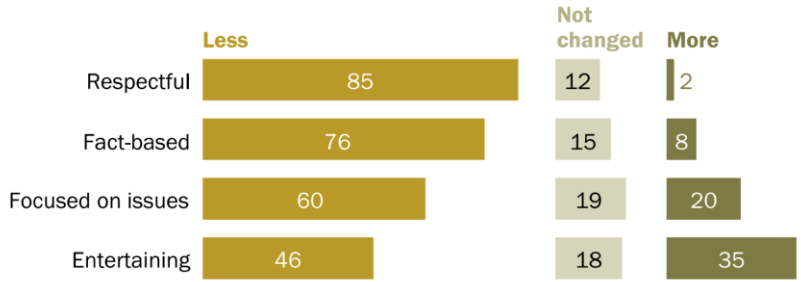
- Da li stvarno postoje osobe koje su u preseku sva 3 skupa sa Venovog dijagrama?
- Retko – obično su to domenski eksperti (npr. fizičari, biolozi...) koji već znaju statistiku, a naučili su da kodiraju i razvijaju ML modele.
- Realnije je da postoje timovi koji uključuju ljude iz preseka bar 2 skupa.
- Na projektu LIGO fizičari su radili sa ljudima koji znaju da programiraju i obučavaju ML modele.



Primer primene nauke 2.0 – politika u 21. veku

Most Americans say political debate in the U.S. has become less respectful, fact-based, substantive

% who say over the last several years the tone and nature of political debate in this country has become ...



% who say Donald Trump has changed the tone and nature of political debate in the U.S. ...

Note: No answer responses not shown.
Source: Survey of U.S. adults conducted April 29-May 13, 2019.

PEW RESEARCH CENTER

Pitanje:
Da li subjektivni utisci ispitanika ove ankete oslikavaju stvarno stanje politike u USA?

Na ovom predmetu naučićete sve potrebne tehnike da biste odgovorili na ovo pitanje (videti slajdove u nastvku)

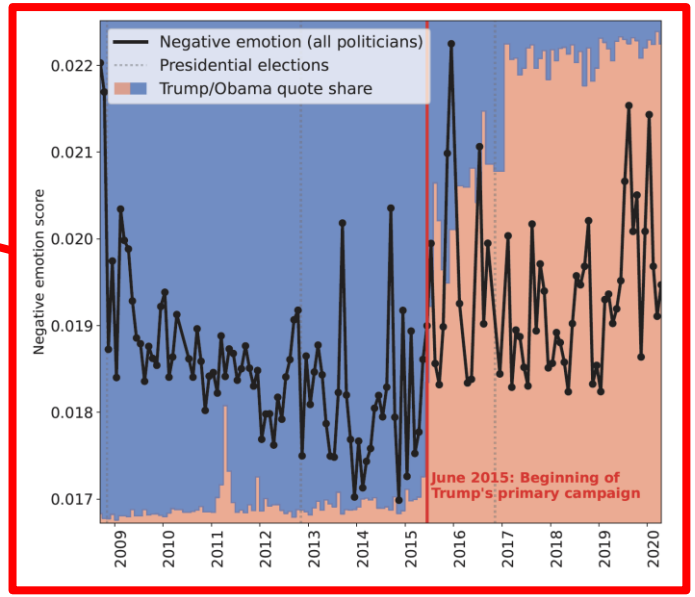
https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2019/06/PP_2019.06.19_Political-Discourse_FINAL.pdf

Sadražaj predmeta

- **Prikupljanje i priprema podataka**
- Vizualizacija i eksplorativna analiza podataka
- Prediktivno modelovanje
- Procesiranje teksta (NLP)
- Procesiranje vremenskih serija
- Redukcija dimenzionalnosti

Sadražaj predmeta

- Prikupljanje i priprema podataka
- **Vizualizacija i eksplorativna analiza podataka**
- Prediktivno modelovanje
- Procesiranje teksta (NLP)
- Procesiranje vremenskih serija
- Redukcija dimenzionalnosti

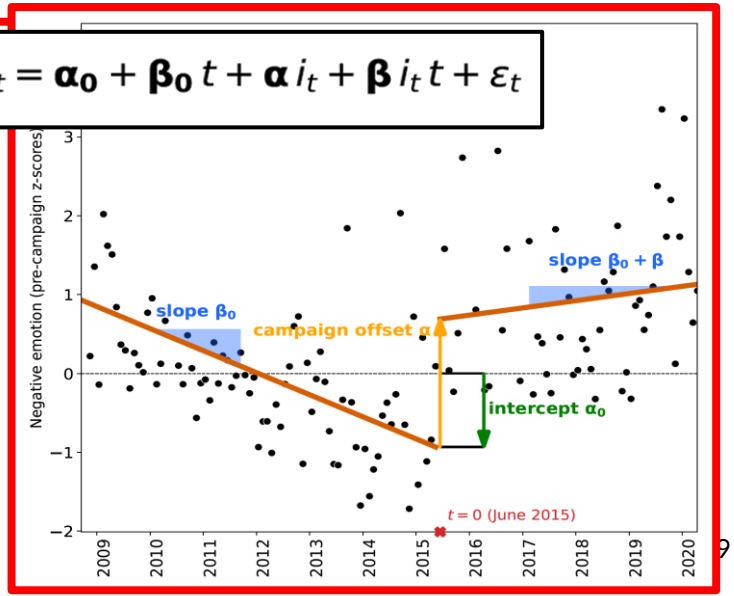


Sadražaj predmeta

- Prikupljanje i priprema podataka
- **Vizualizacija i eksplorativna analiza podataka**
- **Prediktivno modelovanje**
  - Linerana regresija
  - Statistkički testovi značajnosti
  - Modelovanje vremenskih serija
- Procesiranje teksta (NLP)
- Procesiranje vremenskih serija
- Redukcija dimenzionalnosti

"Da li je povećanje negativnog govora od pojave Trampa stvarno ili slučajno poklapanje?"

$$y_t = \alpha_0 + \beta_0 t + \alpha i_t + \beta i_t t + \varepsilon_t$$



Sadražaj predmeta

- Prikupljanje i priprema podataka
- Vizualizacija i eksplorativna analiza podataka
- **Prediktivno modelovanje**
- **Procesiranje teksta (NLP)**
- Procesiranje vremenskih serija
- Redukcija dimenzionalnosti

Ako želite da pogledate kompletan projekat

Full paper available at https://www.nature.com/articles/s41598-023-36839-1

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

United States politicians' tone became more negative with 2016 primary campaigns

Jonathan Külz[1], Andreas Spitz[2], Ahmad Abu-Akel[3], Stephan Günnemann[1] & Robert West[4]✉

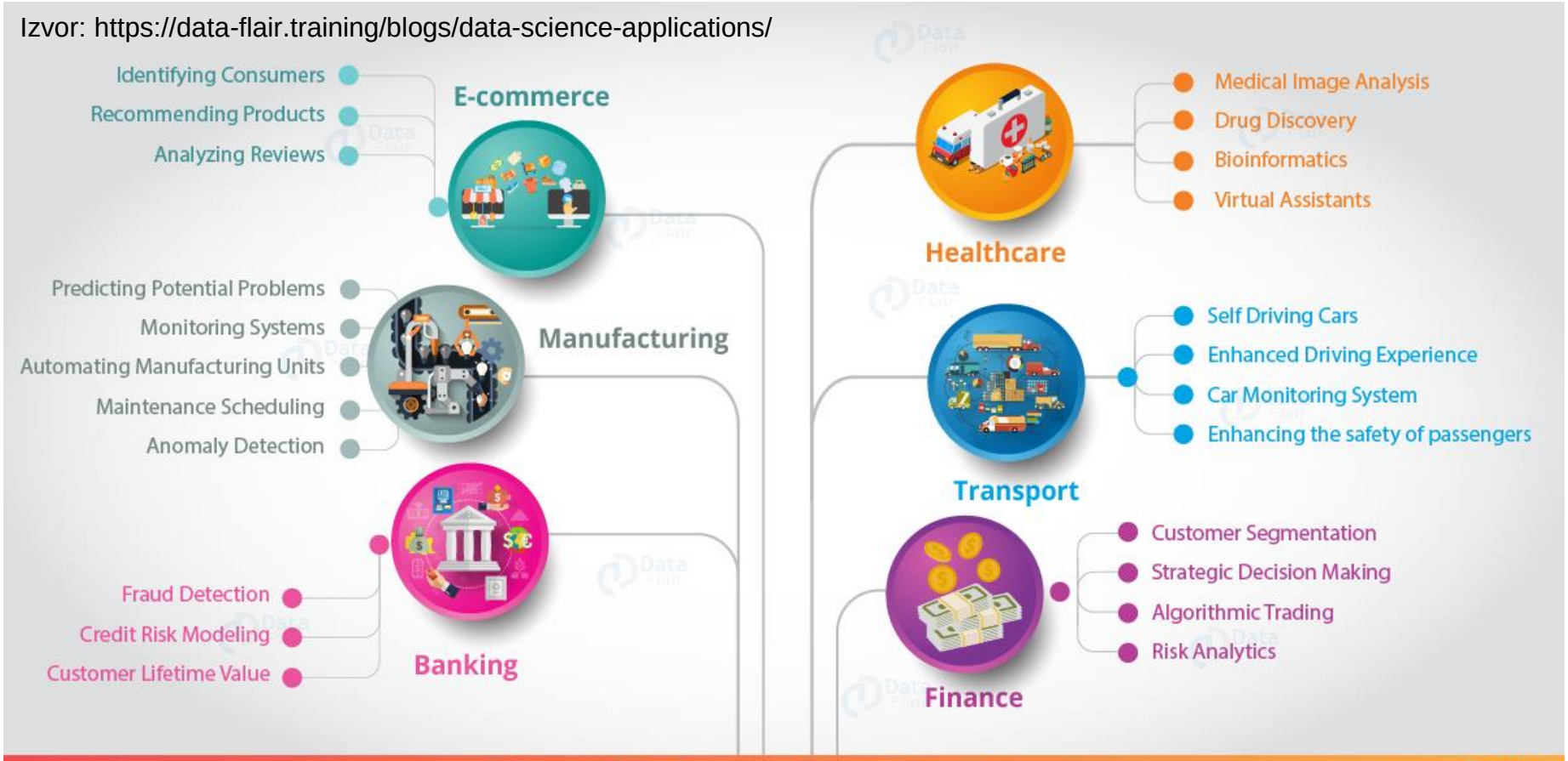
There is a widespread belief that the tone of political debate in the US has become more negative recently, in particular when Donald Trump entered politics. At the same time, there is disagreement as to whether Trump changed or merely continued previous trends. To date, data-driven evidence regarding these questions is scarce, partly due to the difficulty of obtaining a comprehensive, longitudinal record of politicians' utterances. Here we apply psycholinguistic tools to a novel, comprehensive corpus of 24 million quotes from online news attributed to 18,627 US politicians in order to analyze how the tone of US politicians' language as reported in online media evolved between 2008 and 2020. We show that, whereas the frequency of negative emotion words had decreased continuously during Obama's tenure, it suddenly and lastingly increased with the 2016 primary campaigns, by 1.6 pre-campaign standard deviations, or 8% of the pre-campaign mean, in a pattern that emerges across parties. The effect size drops by 40% when omitting Trump's quotes, and by 50% when averaging over speakers rather than quotes, implying that prominent speakers, and Trump in particular, have disproportionately, though not exclusively, contributed to the rise in negative language. This work provides the first large-scale data-driven evidence of a drastic shift toward a more negative political tone following Trump's campaign start as a catalyst. The findings have important implications for the debate about the state of US politics.

A vast majority of Americans—85% in a representative survey by the Pew Research Center[1]—have the impression that "the tone and nature of political debate in the United States has become more negative in recent years". Many see a cause in Donald Trump, who a majority (55%) think "has changed the tone and nature of political debate [...] for the worse", whereas only 24% think he "has changed it for the better"[1]. The purpose of the present article

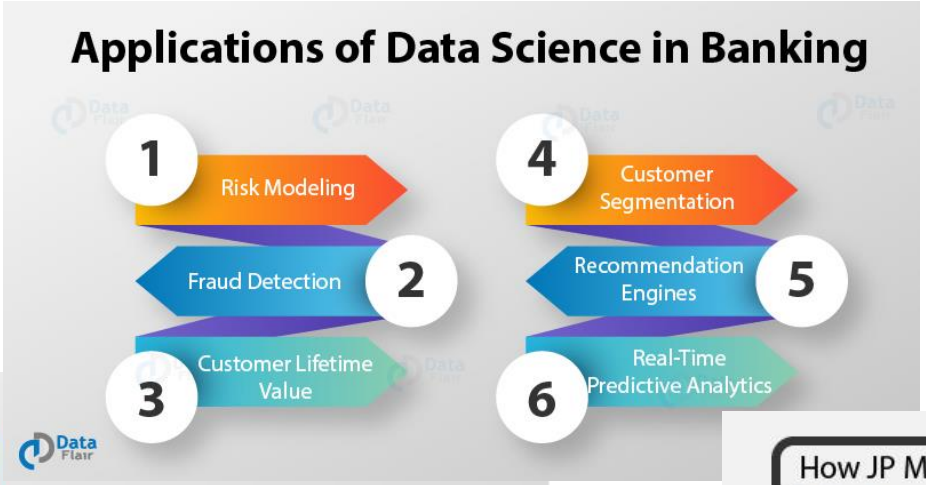
PRIMERI UPOTREBE *DATA SCIENCE* NA REALNIM PROBLEMIMA U RAZLIČITIM OBLASTIMA

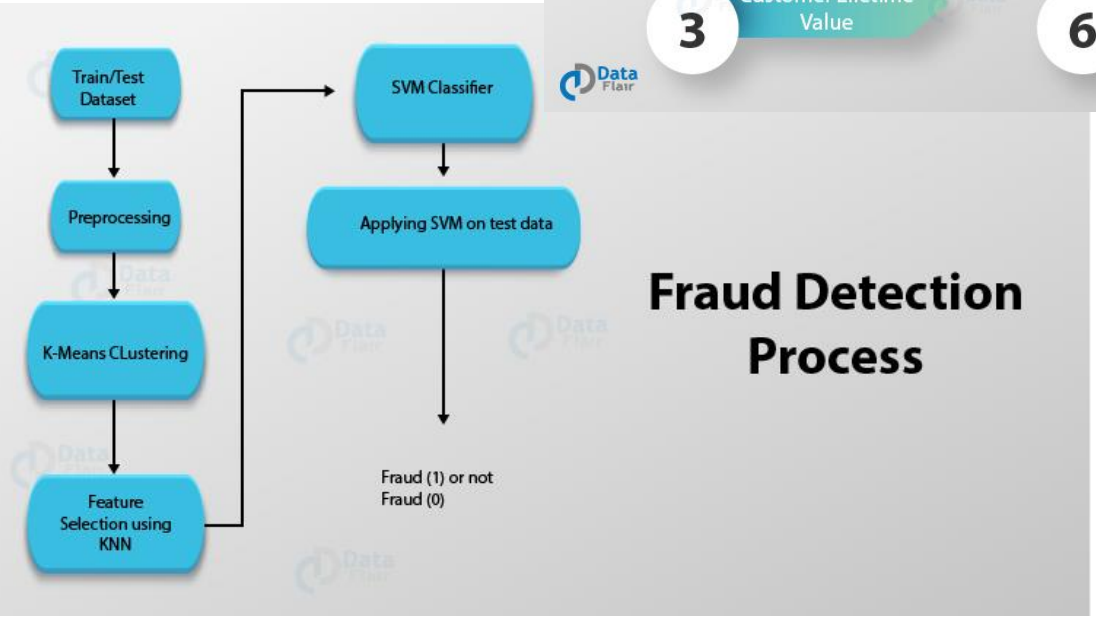
Primeri upotrebe *data science*

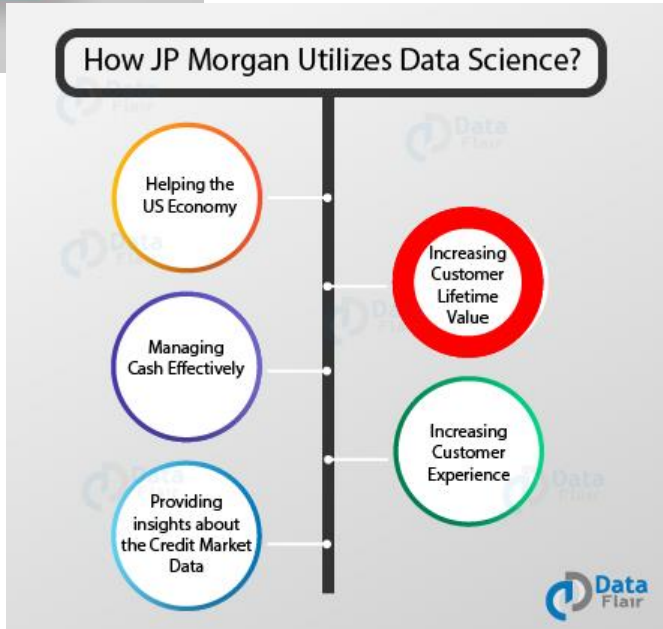
Izvor: https://data-flair.training/blogs/data-science-applications/



Banarstvo, primer: *JP Morgan*

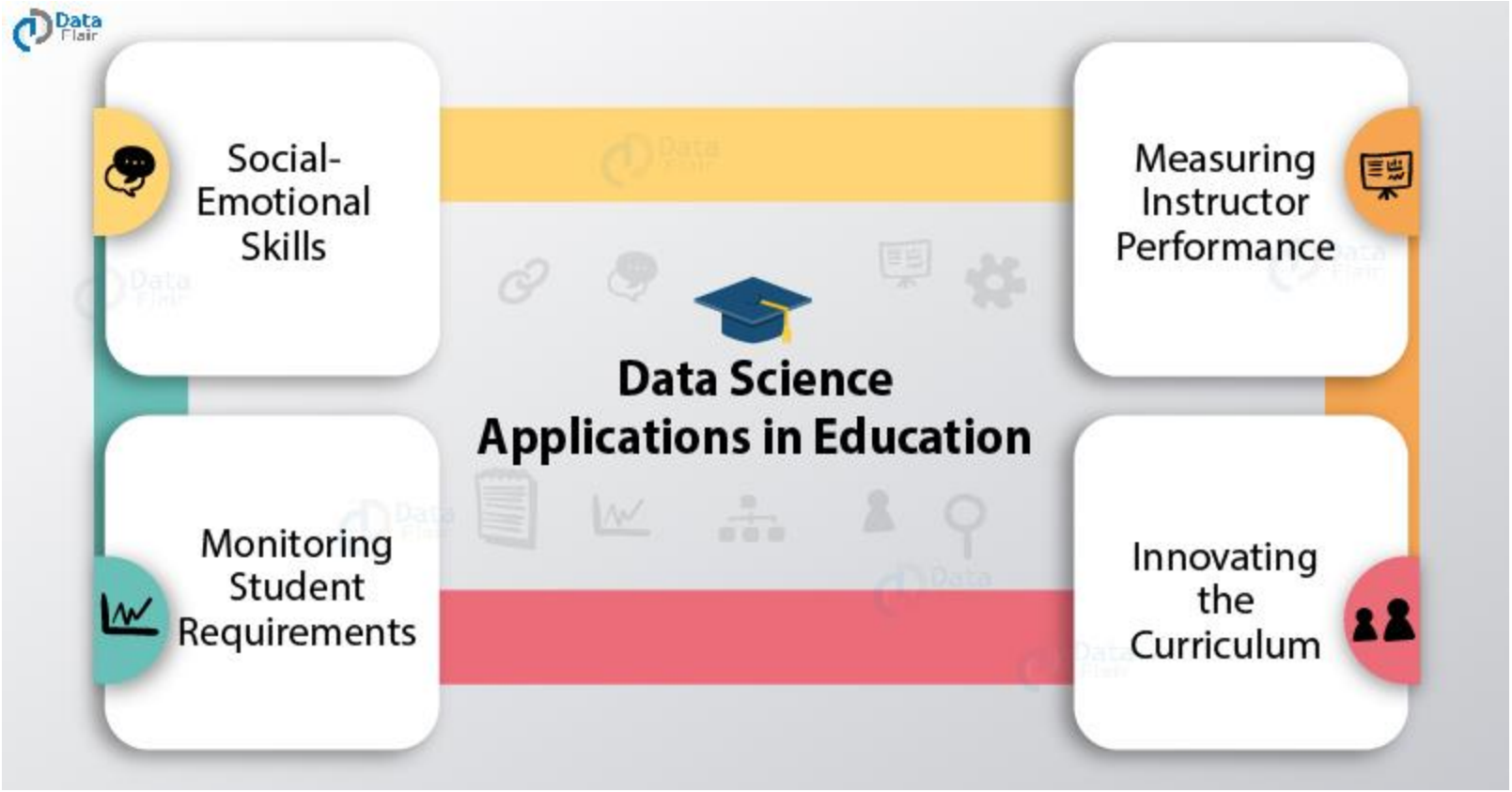






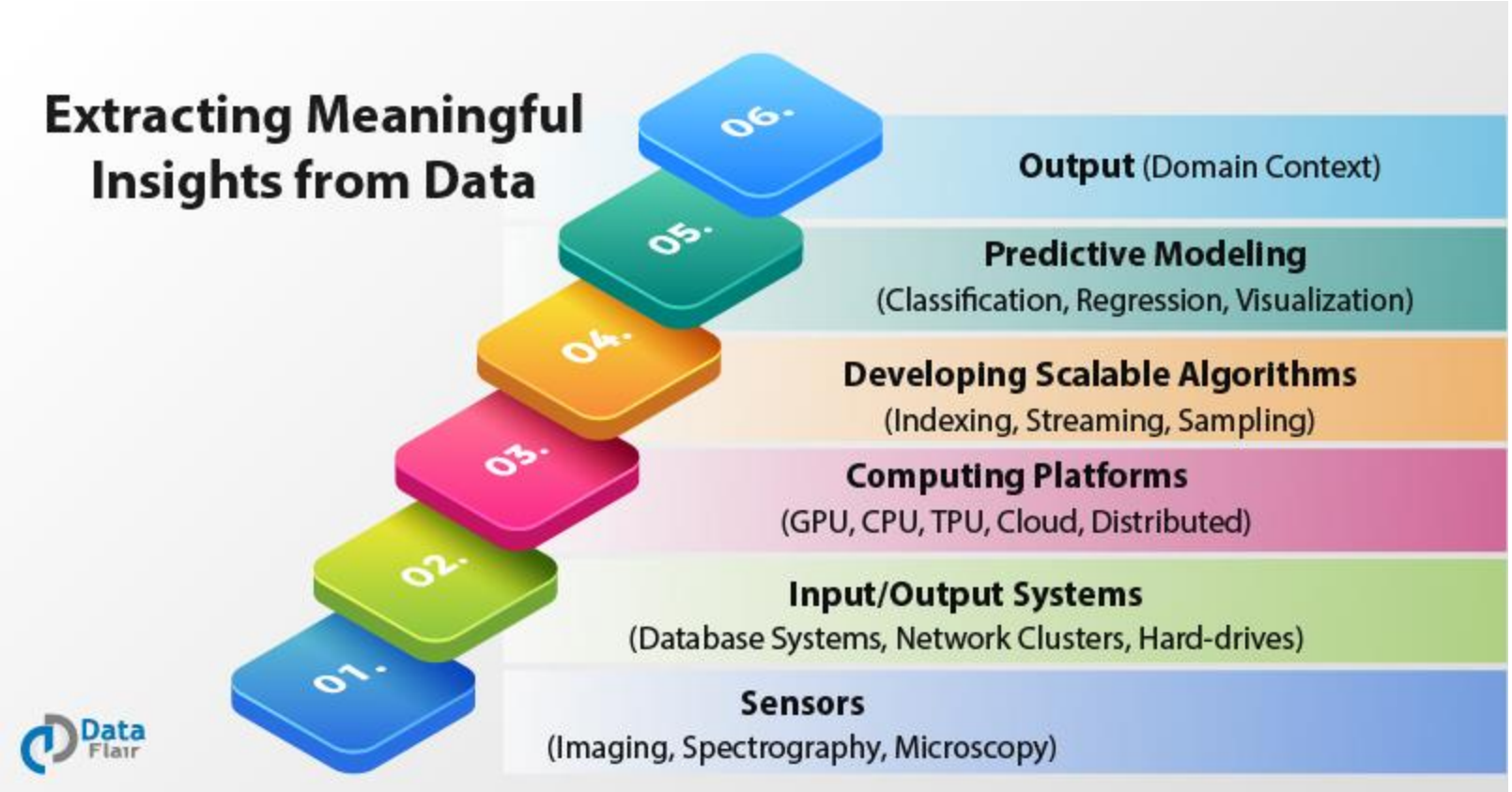
Izvor: https://data-flair.training/blogs/data-science-applications/

Obrazovanje, primer*: The University of Florida*



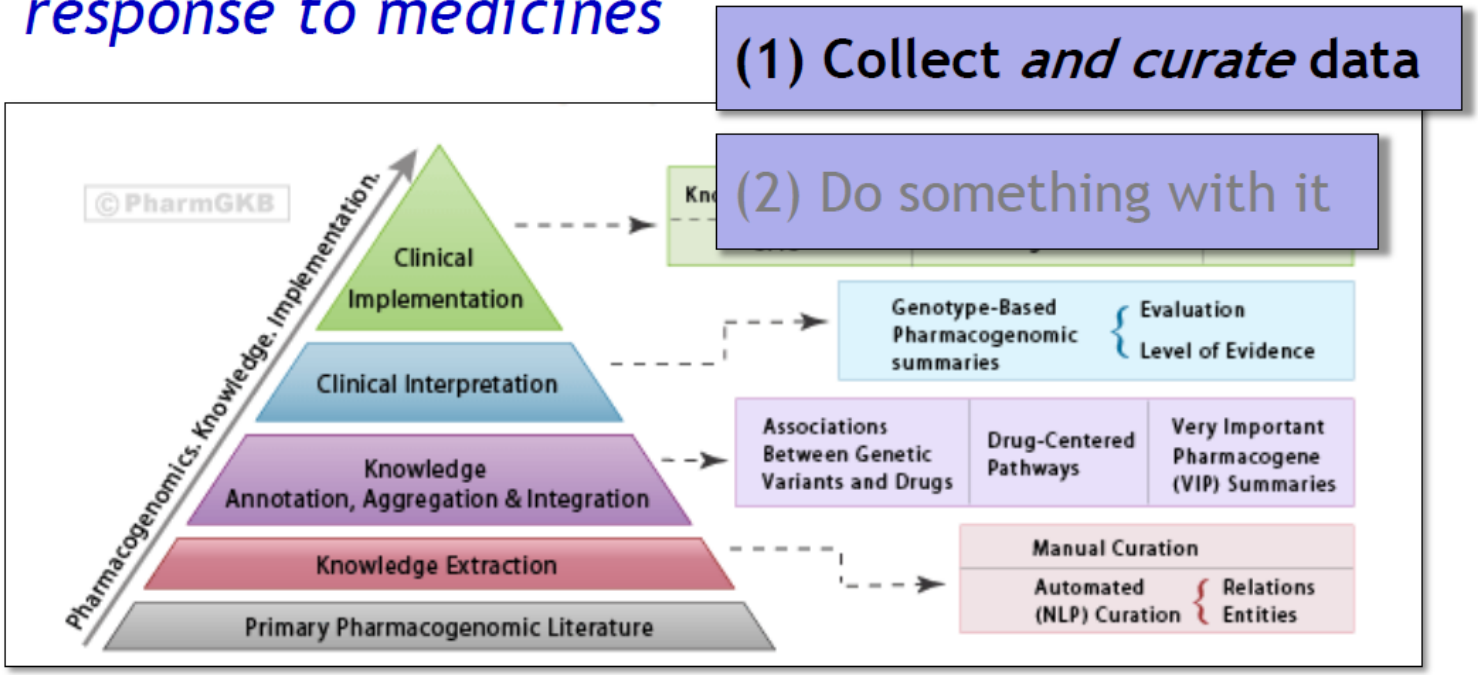
Izvor: https://data-flair.training/blogs/data-science-applications/

Zdravstvo


Extracting Meaningful Insights from Data
06.
Output (Domain Context)
05.
Predictive Modeling
(Classification, Regression, Visualization)
04.
Developing Scalable Algorithms
(Indexing, Streaming, Sampling)
03.
Computing Platforms
(GPU, CPU, TPU, Cloud, Distributed)
02.
Input/Output Systems
(Database Systems, Network Clusters, Hard-drives)
01.
Sensors
(Imaging, Spectrography, Microscopy)
Data Flair

Zdravstvo i Farmacija

PharmGKB collects, curates, and disseminates knowledge about how human genetics affects response to medicines



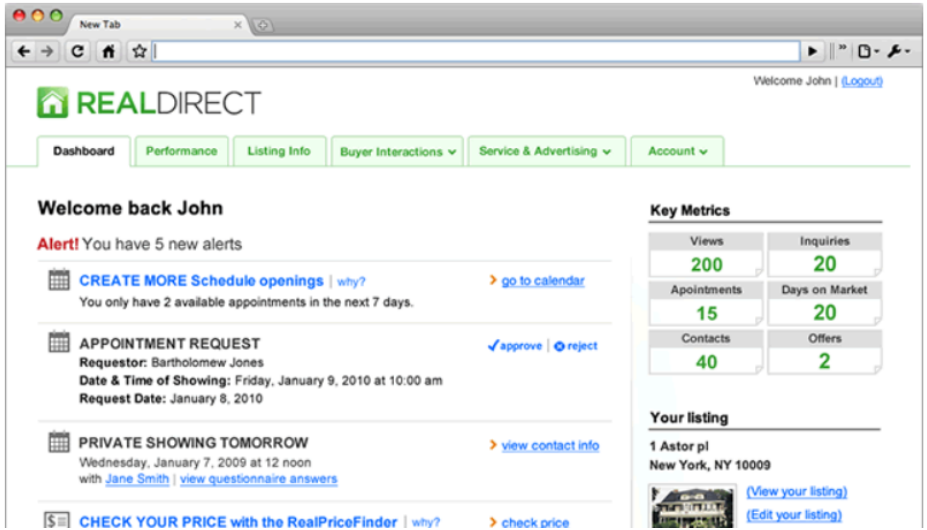
Izvor: “CS 102: Working with Data”, Stanford

Biznis, primer: nekretnine

5. Data Driven Recommendations and Alerts

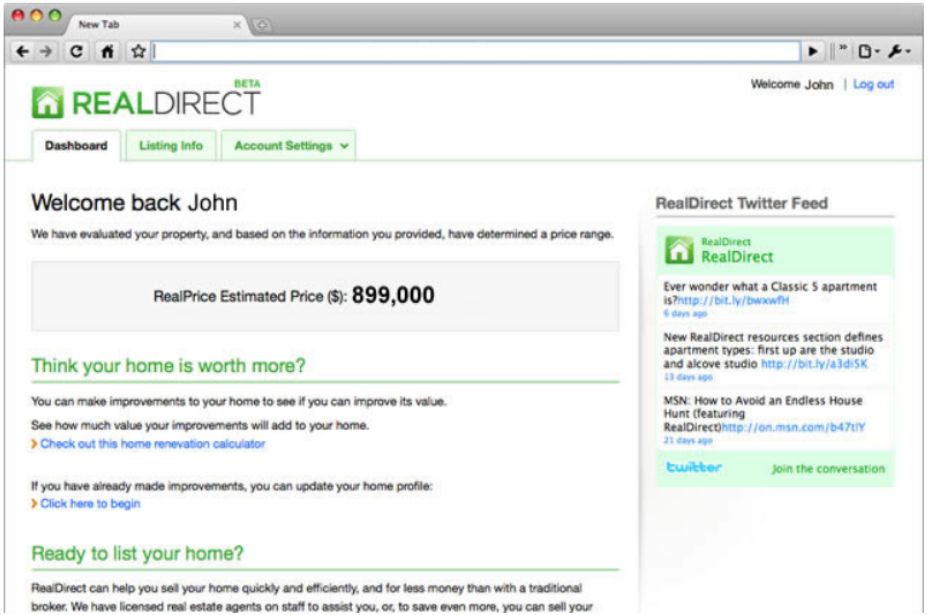
RealDirect pulls data from multiple sources to give you and/or your agent real time, personalized and actionable recommendations about what is working and not working in your sales process, and the ability to fix any problems with the click of a link.

Why It's Better: By utilizing real time data to make decisions, we catch problems before they can negatively impact your sale process, and provide that information in an actionable and impactful way.



1. Find your RealPrice

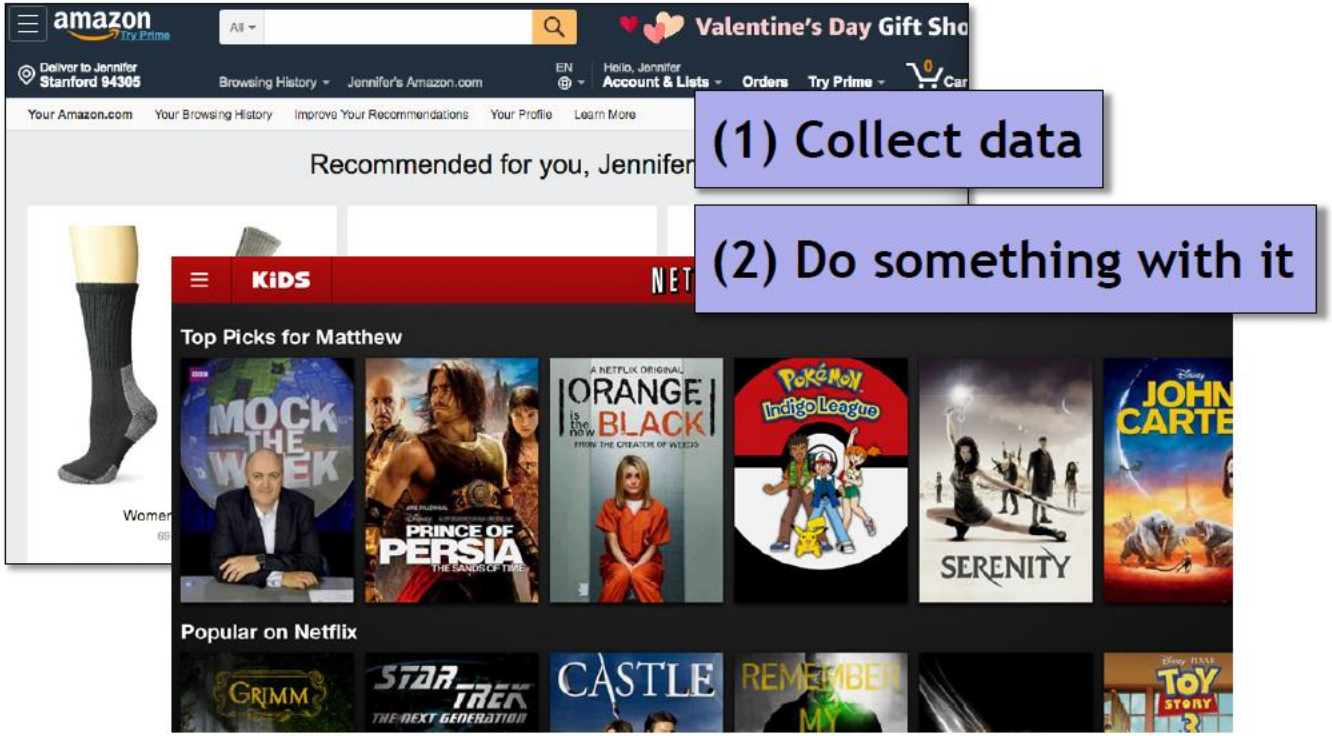
Fill out your RealPrice Profile and tell us about your home. We'll then send you a customized price recommendation for where to price your property if you were to list it today. We factor in considerations like location, size, and condition, but also ask you to provide photos so we can see how your home really compares to recent sales.



Izvor: https://www.realdirect.com/how_it_works

Biznis, prodaja

Biznis, sistemi za preporuku

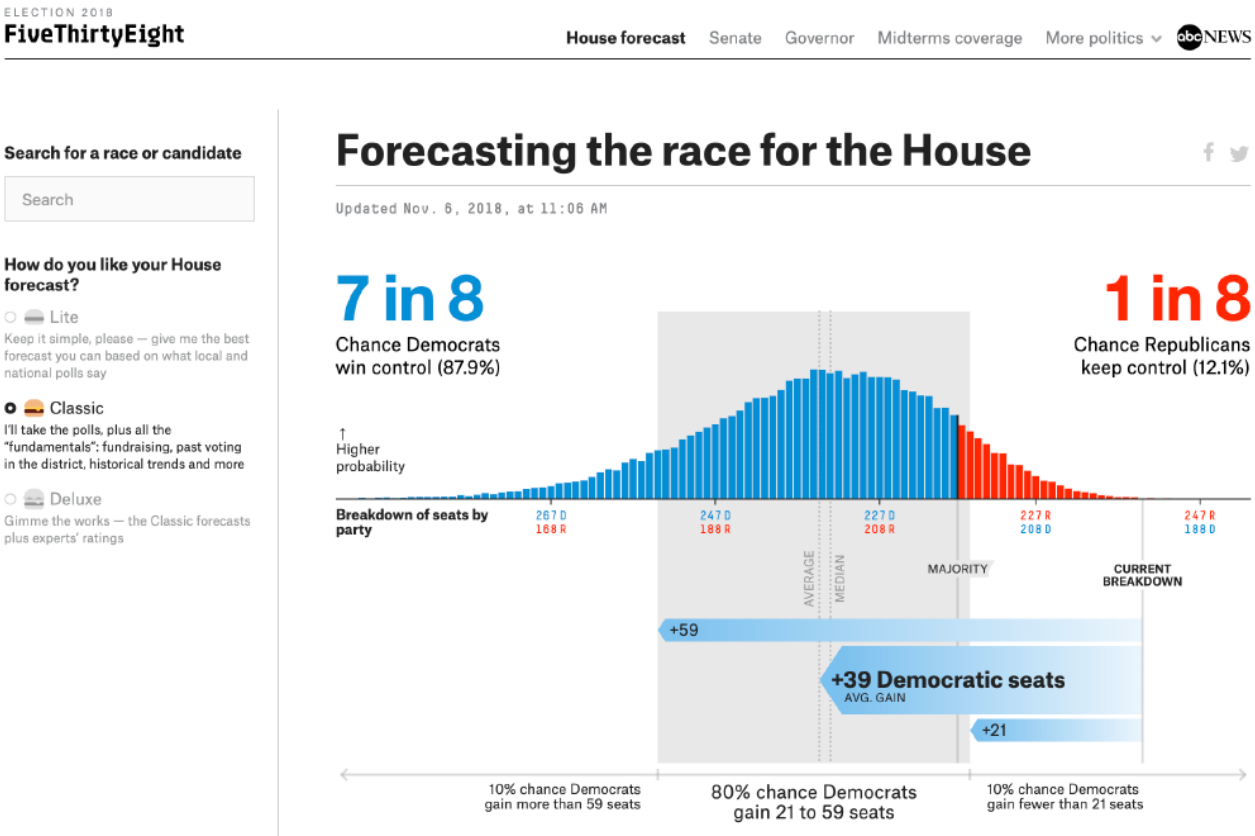


+ music, news, friends, romantic partners, and many more!

Izvor: “CS 102: Working with Data”, Stanford

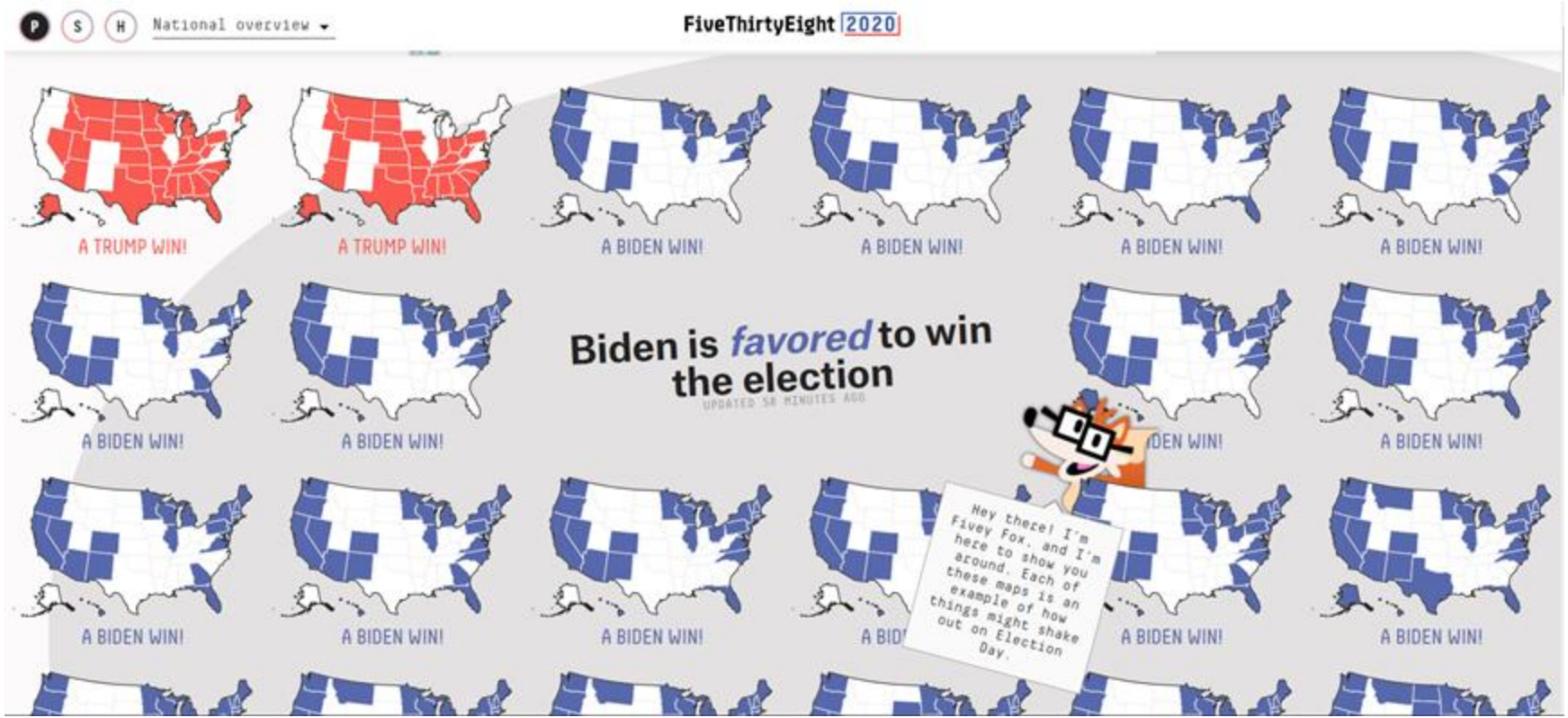
Politika

FiveThirtyEight



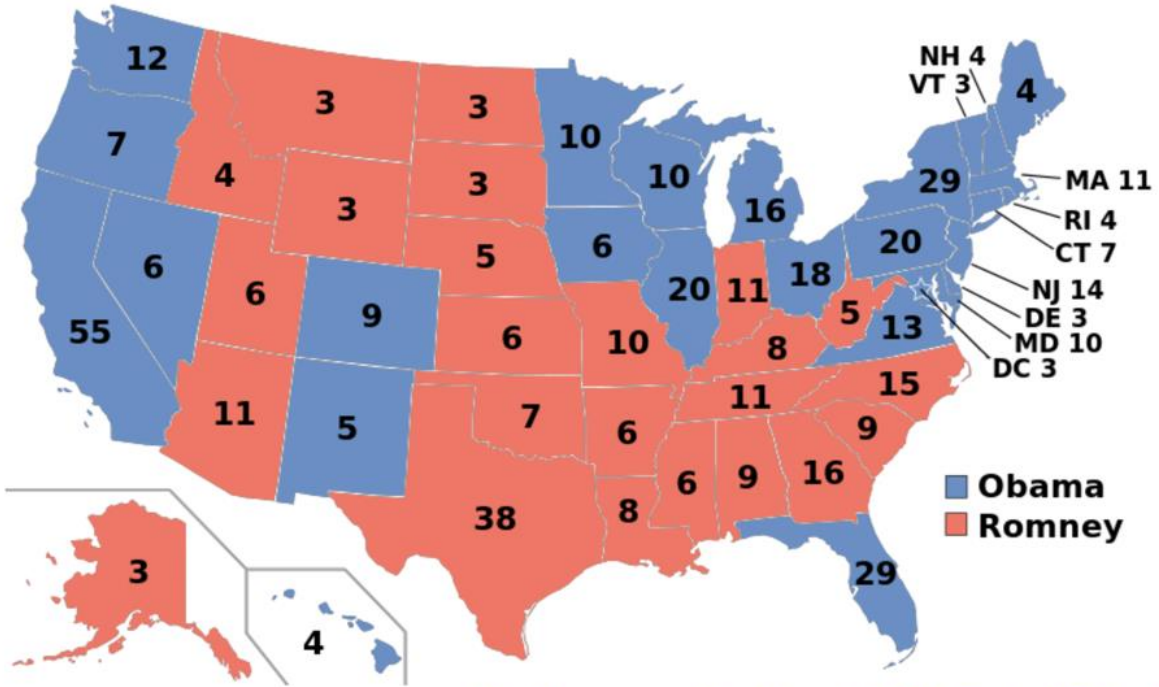
Izvor: https://projects.fivethirtyeight.com/2018-midterm-election-forecast/house/

Politika


P S H National overview
FiveThirtyEight 2020
A TRUMP WIN!
A TRUMP WIN!
A BIDEN WIN!
A BIDEN WIN!
A BIDEN WIN!
A BIDEN WIN!
A BIDEN WIN!
A BIDEN WIN!
Biden is favored to win the election
UPDATED 58 MINUTES AGO
A BIDEN WIN!
A BIDEN WIN!
A BIDEN WIN!
A BIDEN WIN!
A BIDEN WIN!
A BIDEN WIN!
A BIDEN WIN!
A BIDEN WIN!
Hey there! I'm Fivey Fox, and I'm here to show you around. Each of these maps is an example of how things might shake out on Election Day.

Politika izbori u USA 2008 i 2012

"Silver, who made his name by using cold hard math to call 49 out of 50 states in the 2008 general election and all 50 in 2012"



http://commons.wikimedia.org/wiki/File:ElectoralCollege2012.svg
(public domain)

Politika izbori u USA 2008 i 2012

"In the 21st century, the candidate with the **best data,** merged with the best messages dictated by that data, **wins.**"

Andrew Rasiej, Personal Democracy Forum

"... the biggest win came from **good old SQL** on a Vertica data warehouse and from **providing access to data to dozens of analytics staffers** who could follow their own curiosity and distill and analyze data as they needed."

Dan Woods, CITO Research

Izbori u USA 2016

How I Acted Like A Pundit And Screwed Up On Donald Trump

Trump's nomination shows the need for a more rigorous approach.

By Nate Silver

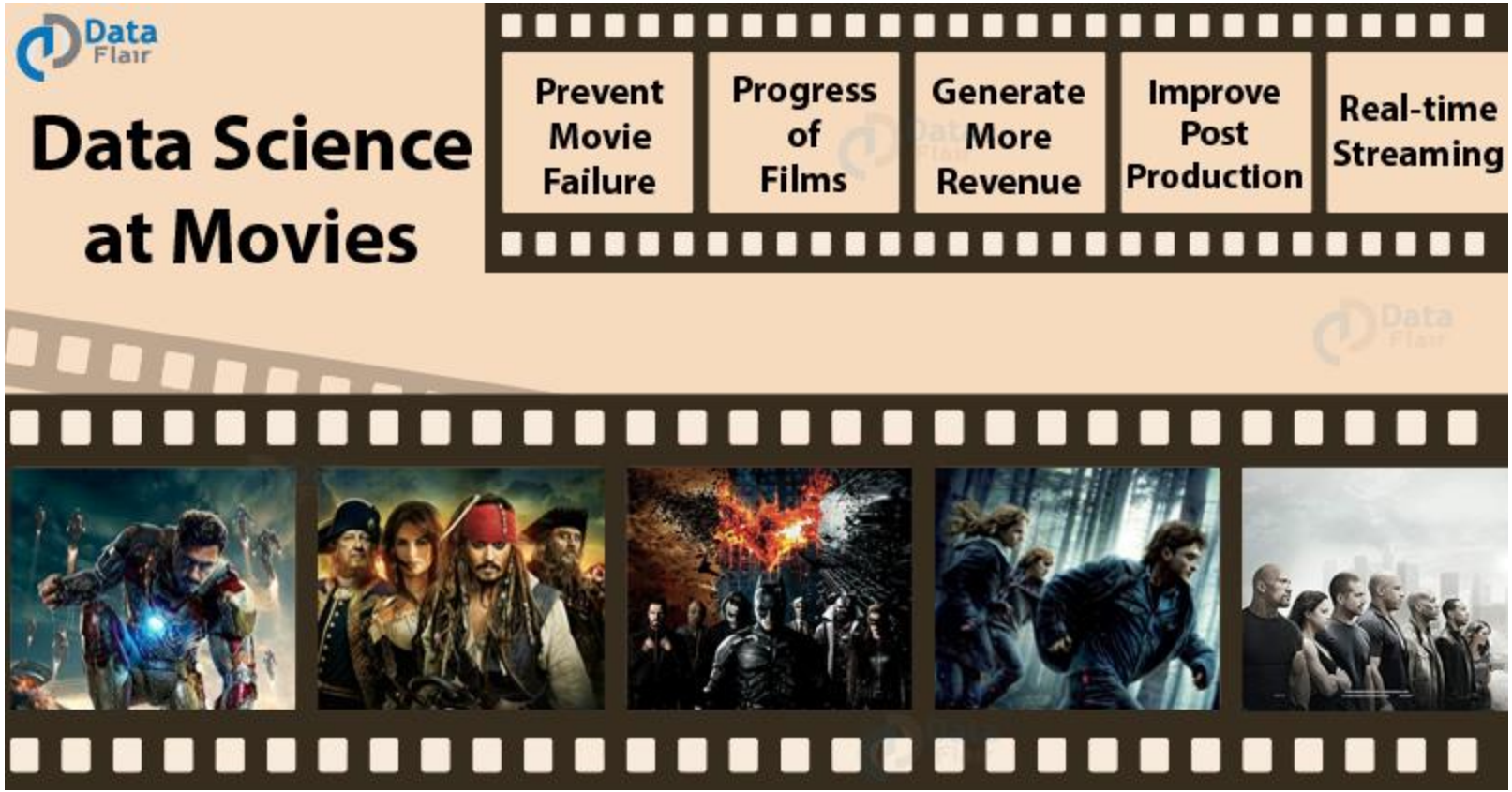
Filed under 2016 Election

Published May 18, 2016

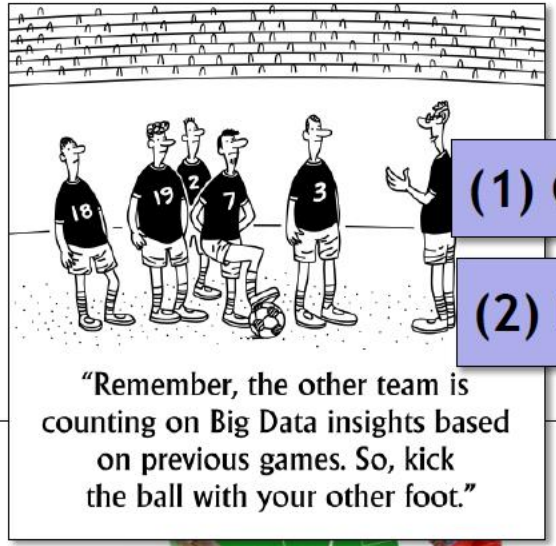
Polls whiz kid Nate Silver and presidential candidate Donald Trump.

Photo illustration by *Slate*. Images by Slaven Vlasic/Getty Images and Ethan Miller/Getty Images.

Filmska industrija

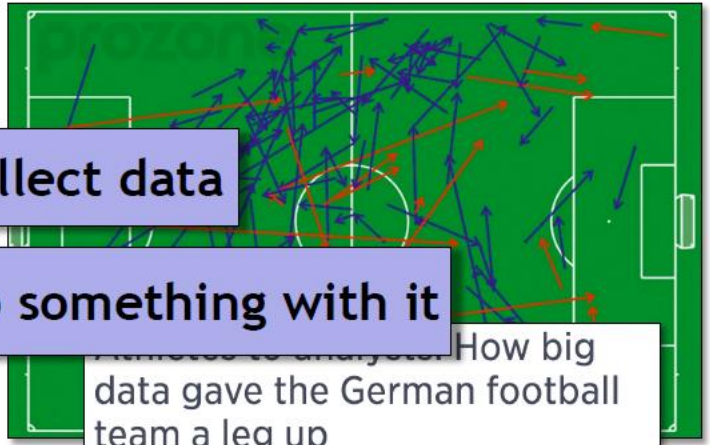


Sport



(1) Collect data

(2) Do something with it



Izvor: “CS 102: Working with Data”, Stanford

Drugi primeri: interesantni eksperimenti

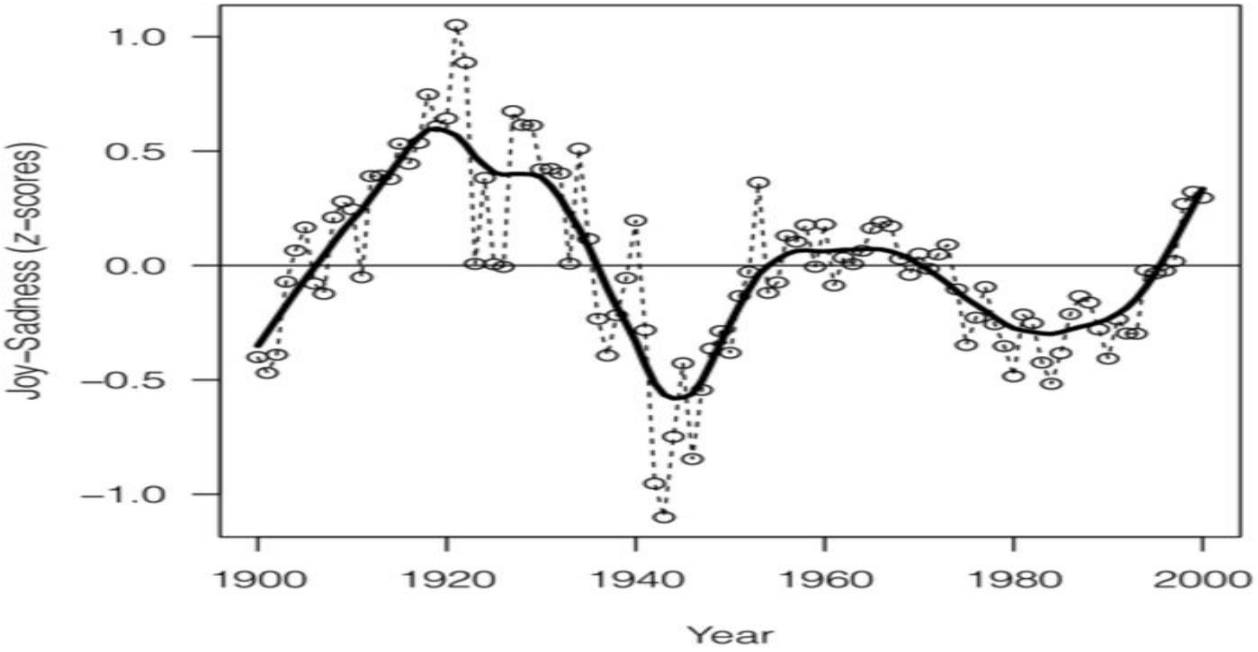
Evolution of Emotions over the 20th Century

Convert all the digitized books in the 20th century into n-grams (with Google n-grams)

Label each 1-gram (word) with a mood score (with WordNet Affect)

Count the occurences of each mood word

Drugi primeri: interesantni eksperimenti



Acerbi A, Lampos V, Garnett P, Bentley RA (2013) The Expression of Emotions in 20th Century Books. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030

Drugi primeri: interesantni eksperimenti

Gendered Language in Teacher Reviews

This interactive chart lets you explore the words used to describe male and female teachers in about 14 million reviews from RateMyProfessor.com.

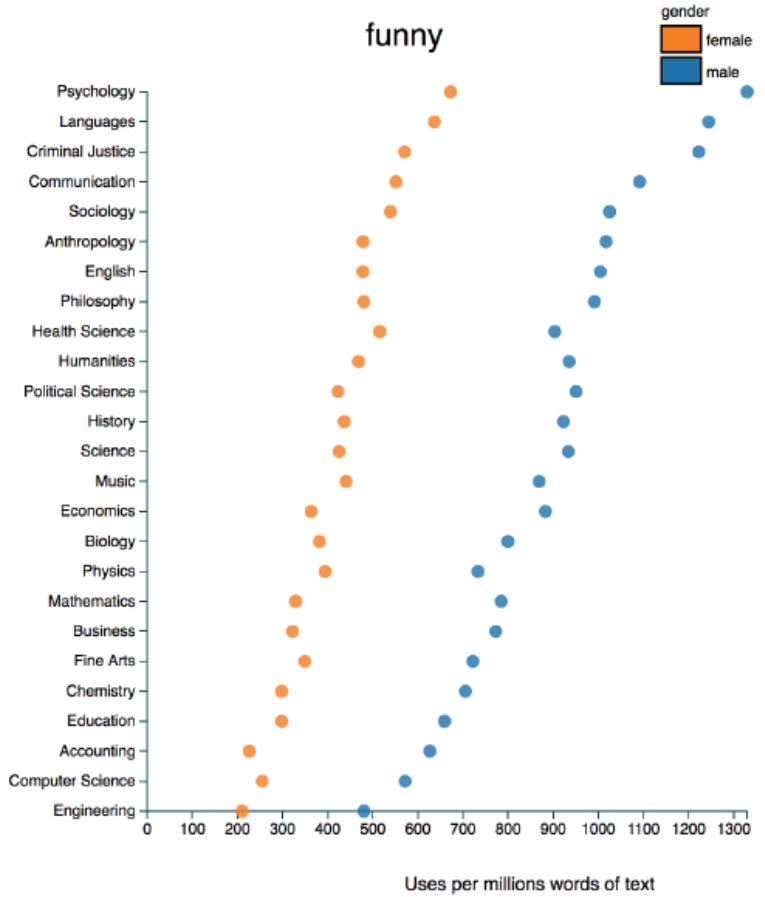
You can enter any other word (or two-word phrase) into the box below to see how it is split across gender and discipline: the x-axis gives how many times your term is used per million words of text (normalized against gender and field). You can also limit to just negative or positive reviews (based on the numeric ratings on the site). For some more background, see here.

Not all words have gender splits, but a surprising number do. Even things like pronouns are used quite differently by gender.

Search term(s) (case-insensitive): use commas to aggregate multiple terms

funny

All ratings | Only positive | Only negative



Izvor: http://benschmidt.org/profGender/

Drugi primer, humanost

Figure 2: Example of metal roof in center of satellite image.

Figure 3: Example of thatched roof in center of satellite image.

Abelson, Varshney, and Sun. “Targeting Direct Cash Transfers to the Extremely Poor,” 2012

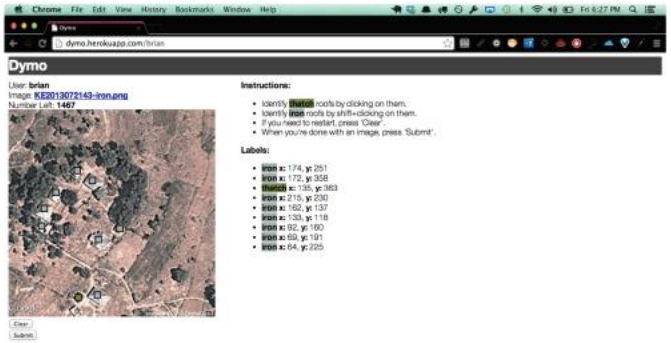


Figure 6: Screen shot of application deployed for crowdsourced labeling of roofs in satellite images.

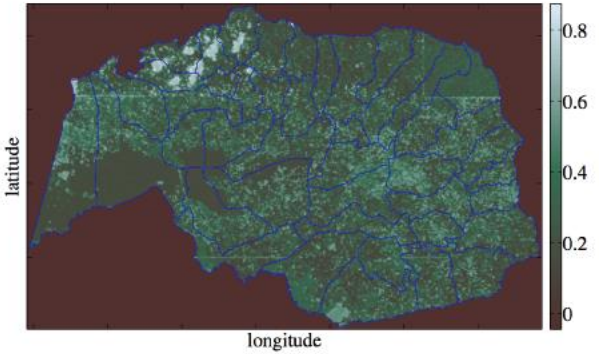
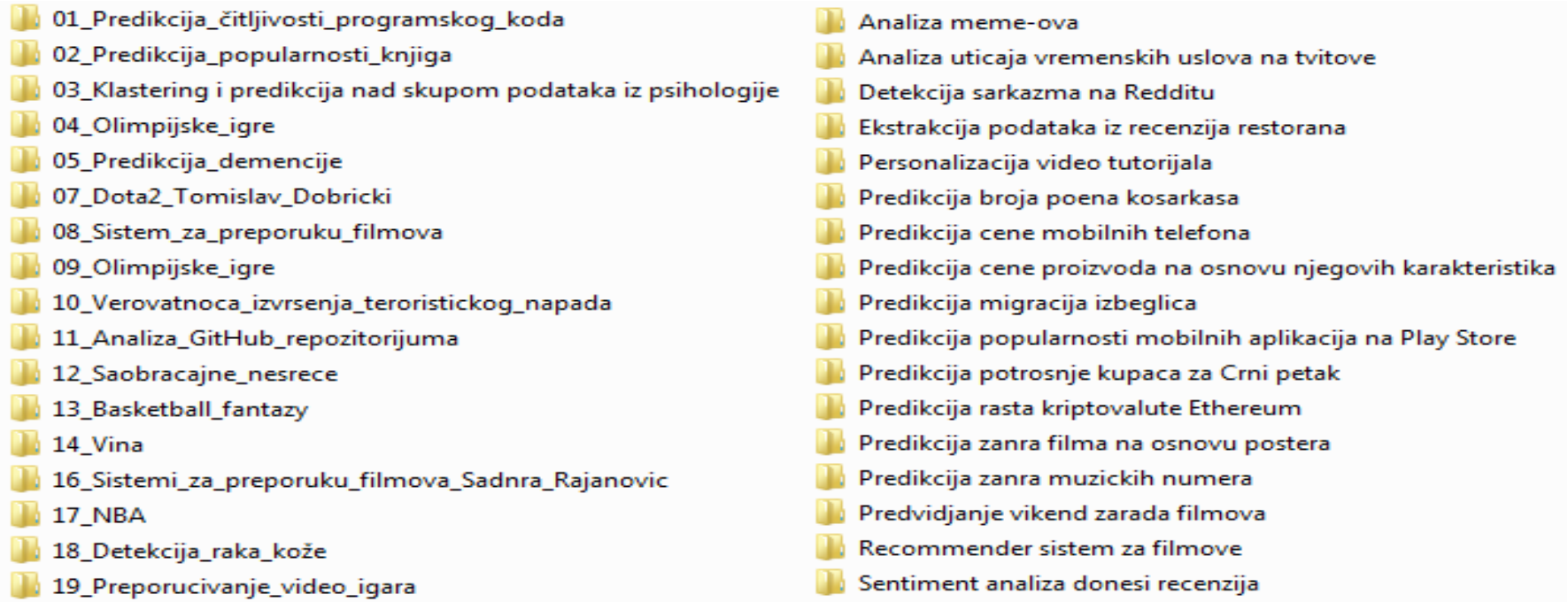


Figure 11: Heat map of proportion of roofs that are metal in the region of interest.

Izvor: J. Zico Kolter, Carnegie Mellon University

Neki od dosadašnjih projekata

Neki od projekata od pre nekoliko godina

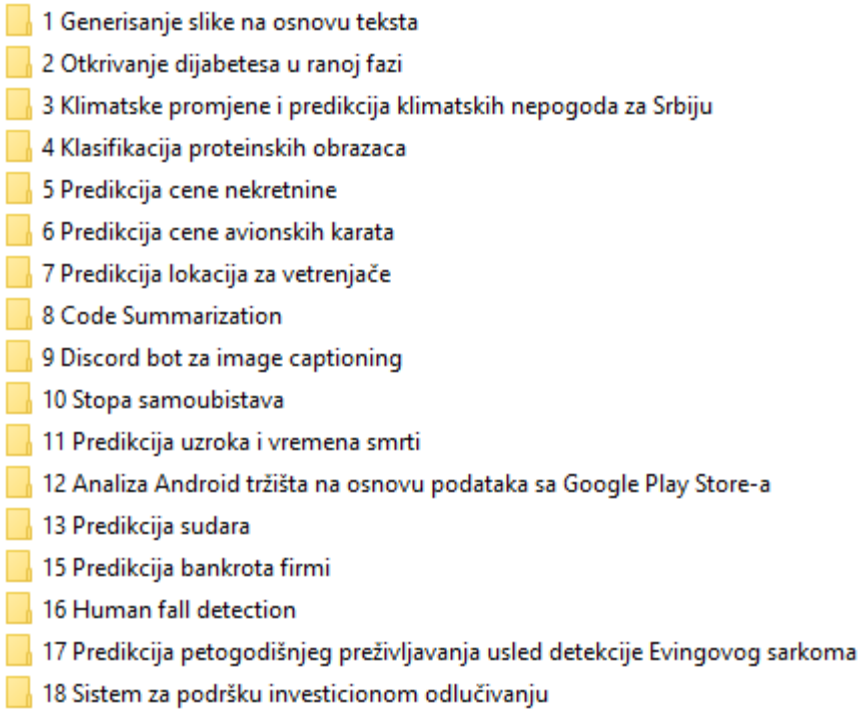


Neki od projekata iz 2019. godine

- Analiza Spotify pesama - Nevena Rokvic.pdf
- Multi-hop question-answering system on graph based Wikipedia - Fi
- Pravni NER - Katarina Glorija Grujic, Aleksandar Cvejic, Andrija Cvejic
- Predikcija cene nafte - Veljko Vojinovic.pdf
- Predikcija MVP igraca NBA lige - Nikola Malencic, Nikola Djordjevic,
- Predikcija subjektivnih aspekata pitanja i odgovora - Nikolina Radici
- Predikcija visine temperature i analiza uticaja vremenskih prilika na b
- YouTube trending analiza - Marko Pejic, Stefan Ruvceski, Mirko Ivic.p

- 13 9Gag
- 14 Detekcija respiratornih bolesti na osnovu disanja
- 15 Analiza sumskih pozara Brazila
- 16 Poljoprivredni proizvodi
- 17 Oruzani sukobi
- 18 Predikcija cene smestaja Airbnb
- 20 Anomalije u vremenskoj seriji podataka elektro mreže
- 21 Premijer liga
- 22 Fantasy Premier League
- 23 Mobilni telefoni
- 24 Generisanje sadržaja pesme
- 25 Diabetes
- 26 Prepoznavanje zanra i izvodjaca pesme na osnovu teksta

Neki od projekata iz 2020. godine



Neki od projekata iz 2022 godine

1. Predikcija rizičnosti investicije u kriptovalute
2. Predikcija cena spekulativnih kriptovaluta upotrebom Support Vector Regression algoritma
3. Određivanje rase mačaka i pasa
4. Analiza objavljenih članaka u cilju utvrdivanja da li je vest istinita ili lažna
5. Analiza IT firmi na osnovu dostupnih podataka sa platformi za oglasavanje poslova
6. Preporuka vesti uzimajuci u obzir da su vesti istinite
7. Analiza globalnog zagrevanja i uticaj na porast nivoa mora
8. Određivanje položaja na mapi političkog spektra korisnika društvene mreže "Twitter" na osnovu sadržaja koji se nalazi na njegovoj profilnoj stranici
9. Analiza zagađenosti vazduha u gradovima
10. Predikcija da li se tweet odnosi na pravu katastrofu ili ne
11. Predviđanje rizika izumiranja životinjskih vrsta
12. Predikcija ishoda profesionalnih timskih mečeva u igrici CS:GO
13. Analiza faktora koji utiču na globalni osećaj sreće
14. Human Activity Recognition
15. Detekcija indikatora loše dizajniranog koda upotrebom graf neuronskih mreža
16. Detekcija zagađenosti (vazduh, voda, zemljište)
17. Predikcija vremensih uslova na Marsu
18. Predviđanje/utvrđivanje rizika pojave požara u prirodi na osnovu istorijskih podataka o vremenskim i klimatskim uslovima, sateliski detektovanim termalnim anomalijama i podacima o pokrivenosti Zemlje šumom
19. Predikcija zemljotresa
20. Klasifikacija teksta iz domena finansijskih vesti

Neki od projekata od prošle godine

- Analiza _saobraćaja_Srbije_upotrebom_snimaka_video_kamera
- Comparison of RNN stock prediction models with
- Dušan_Lazić_Milan_Sekulić_Predviđan je_cene_korišćenih_automobila
- Klasifikacija EEG signala na osnovu sentimenta emocija ispitanika
- Predikcija cena avionskih letova
- Predikcija cene akcija na osnovu sentimenta
- Predikcija cijena avionskih karata
- Predikcija karcinoma dojke na osnovu uzoraka
- Predikcija potrošnje električne energije na osnovu
- Predikcija S&P 500 indeksa
- Predikcija saobraćajnih gužvi na osnovu video materijala nadzornih kamera
- Predikcija_cena_smart_telefona_izvestaj
- Primena algoritama mašinskog učenja za procenu

- Analiza sentimenta recenzija zaposlenih o svojim kompanijama
- Analiza sentimenta Rotten Tomatoes recenzija
- Klasifikacija stakla
- Predikcija email spam-a
- Predikcija ishoda Formula 1 trka i čitave sezone
- Predikcija popularnosti oglasa za posao na Linkedin
- Predviđanje popularnosti TikTok pesme
- Sentiment analiza komentara kulinarskih recepata
- Детекција малициозних Phishing e-mail порука
- Класификација вести по садржају
- Класцификација текста по детектованој емоцији
- Предикција ризика обољења од Алцхајмерове болести