Uvod u nauku o podacima

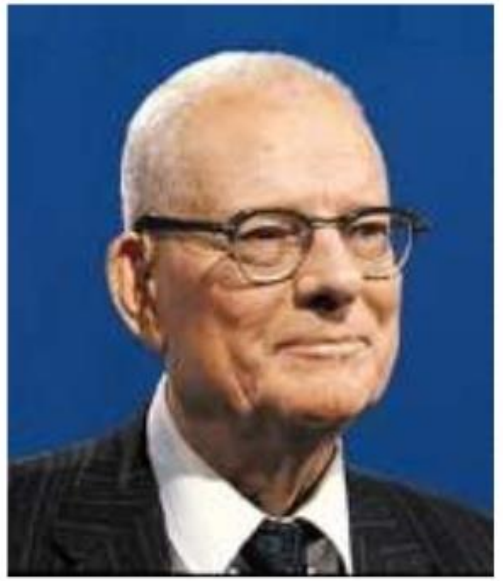
Milan M.Milosavljević

mmilosavljevic@singidunum.ac.rs

Poreklo oblasti

U boga verujemo, svi ostali neka donesu podatke.

DR. WILLIAM EDWARDS DEMING



Biodata:

- 1900 – 1993
- Graduated from University of Wyoming, University of Colorado (Master in Maths & Physics), University of Yale (PhD in Mathematical Physics)

2

Na webu se ova izjava pripisuje podjednako i Vilijemu Demingu i Robertu Hajdenu. Profesor Hajden tvrdi da ovo nije njegova izjava. Ironija je u tome da se ne mogu naći „podaci“ koji bi potvrdili da je Deming zaista njen autor.

Fenomen podataka

Peter Norvig:

Direkor istraživanja u Googlu

(1956 -)

Jednostavan model + puno podataka > kompleksni modeli

Masivni podaci - primeri

facebook

1B+ USERS
30+ PETABYTES

WIKIPEDIA
The Free Encyclopedia

32 million
pages

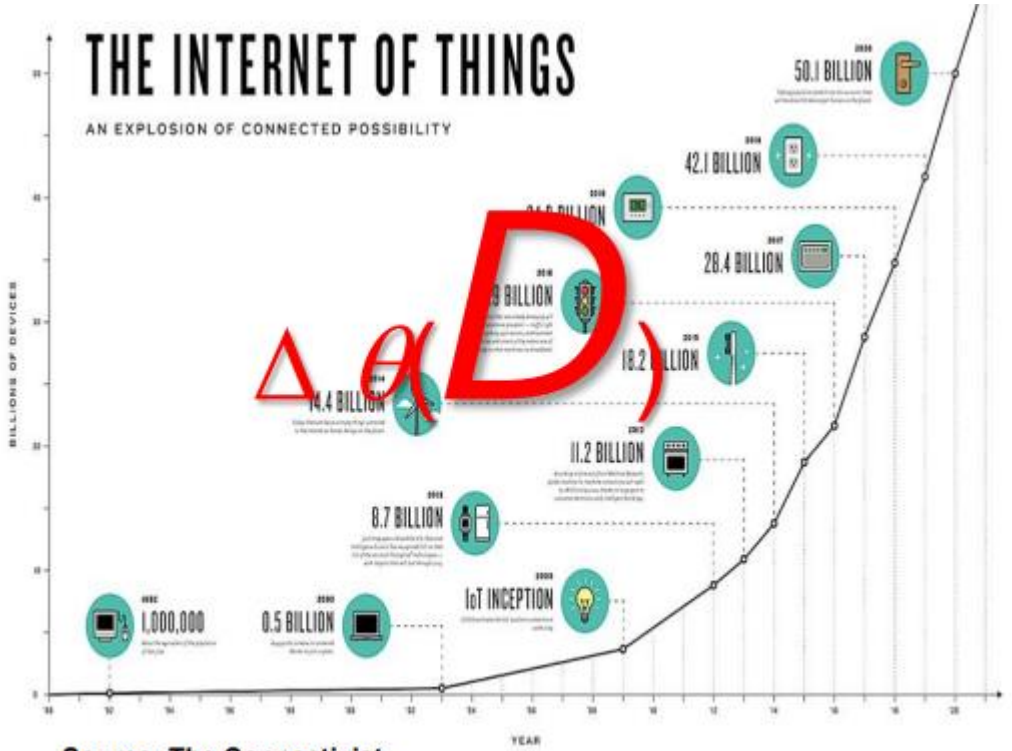
YouTube

100+ hours video
uploaded every minute

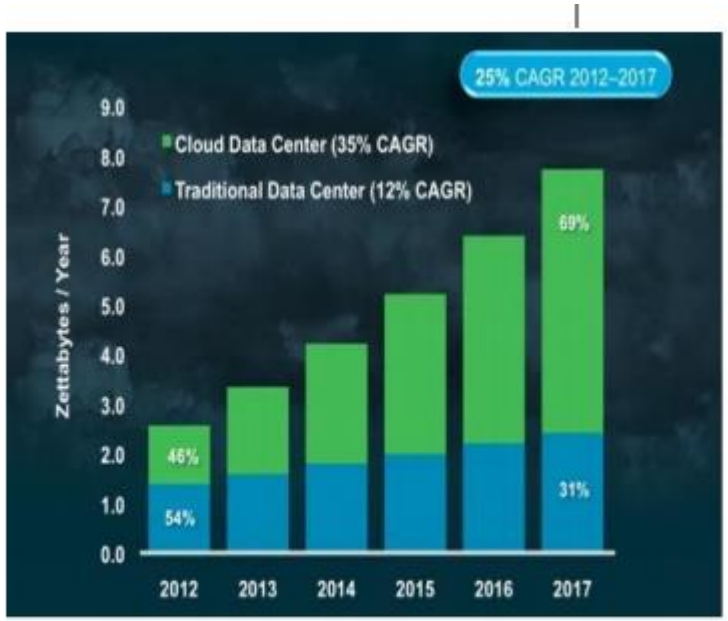
twitter

645 million users
500 million tweets / day

Primer1 – masivni podaci



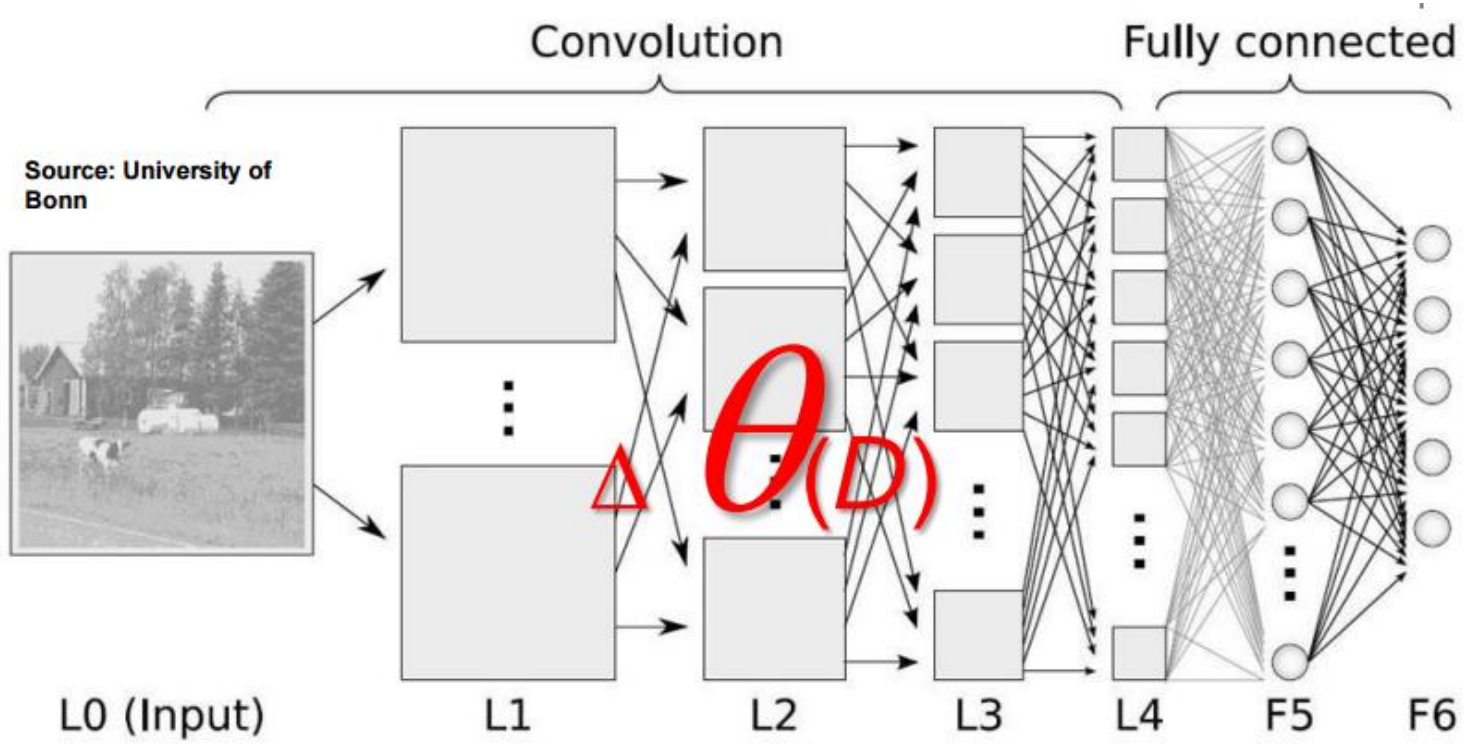
Source: The Connectivist



Source: Cisco Global Cloud Index

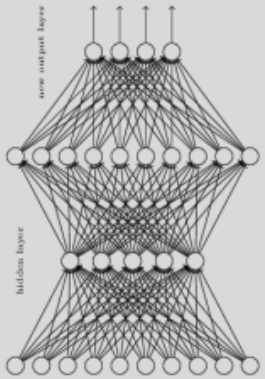
Familiar problem: data from 50B devices, data centers won't fit into memory of single machine

Primer 2: Gigantski modeli

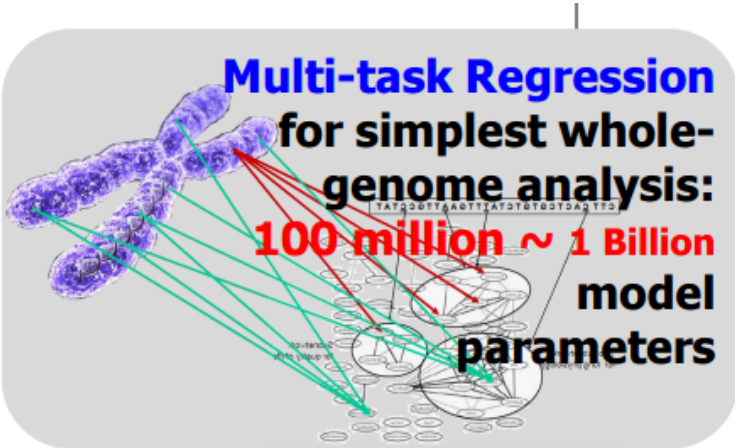


Maybe Big Data needs Big Models to extract understanding?
But models with >1 trillion params also won't fit!

Rastuća potreba za velikim savremenim programima mašinskog učenja



Google Brain
Deep Learning
for images:
1~10 Billion
model parameters



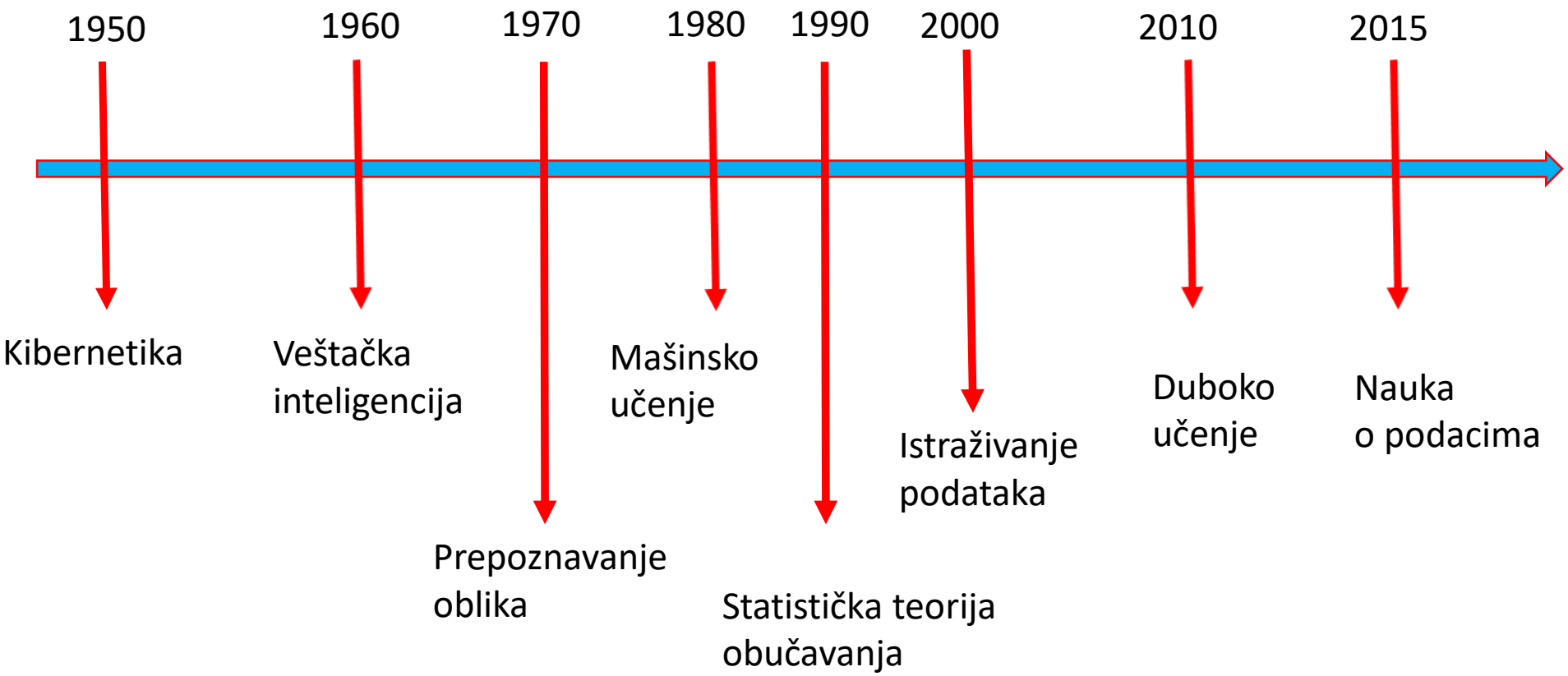
Multi-task Regression
for simplest whole-genome analysis:
100 million ~ 1 Billion
model parameters

Topic Models
for news article analysis:
Up to 1 Trillion
model parameters

Collaborative filtering
for Video recommendation:
1~10 Billion
model parameters

NETFLIX

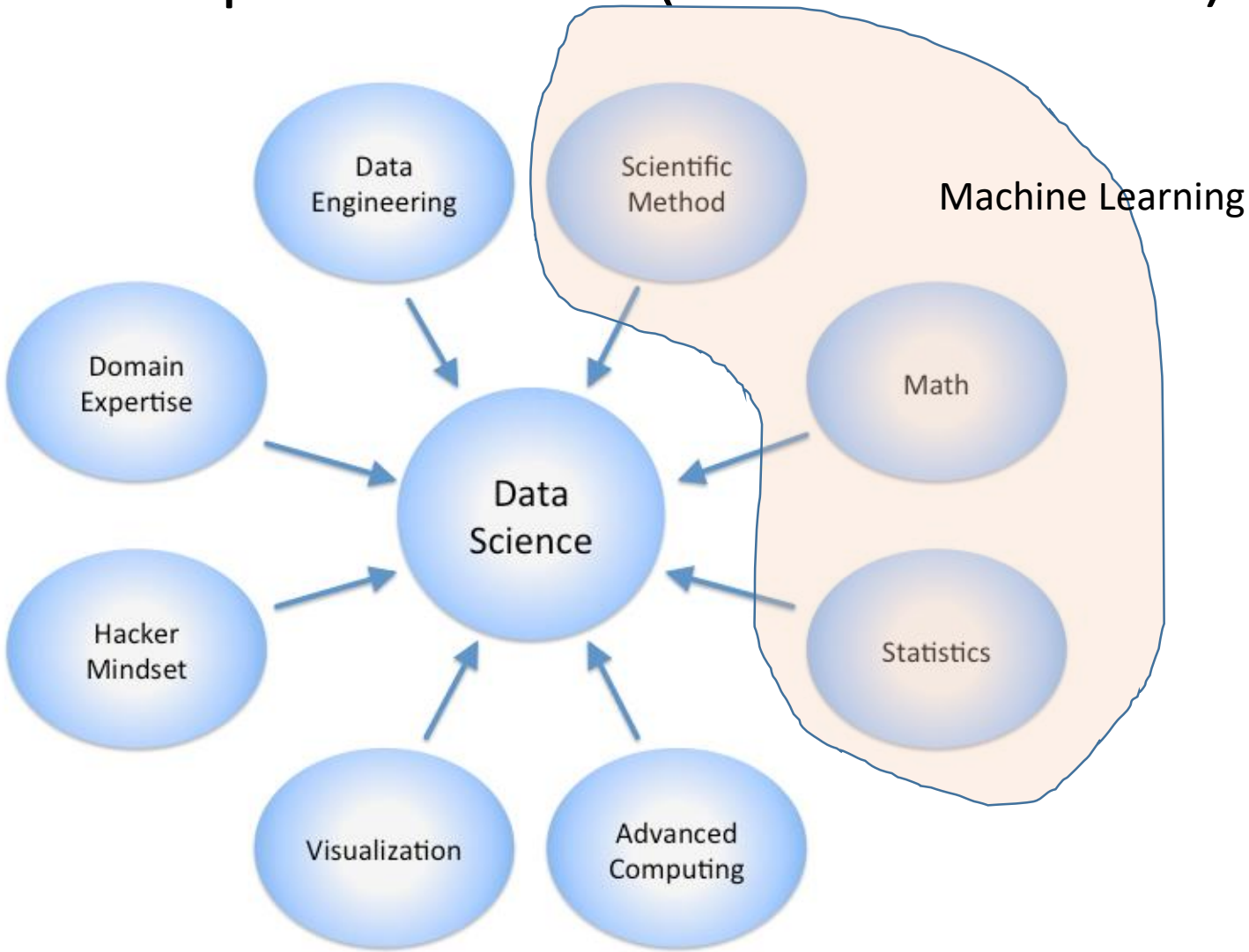
Poreklo oblasti



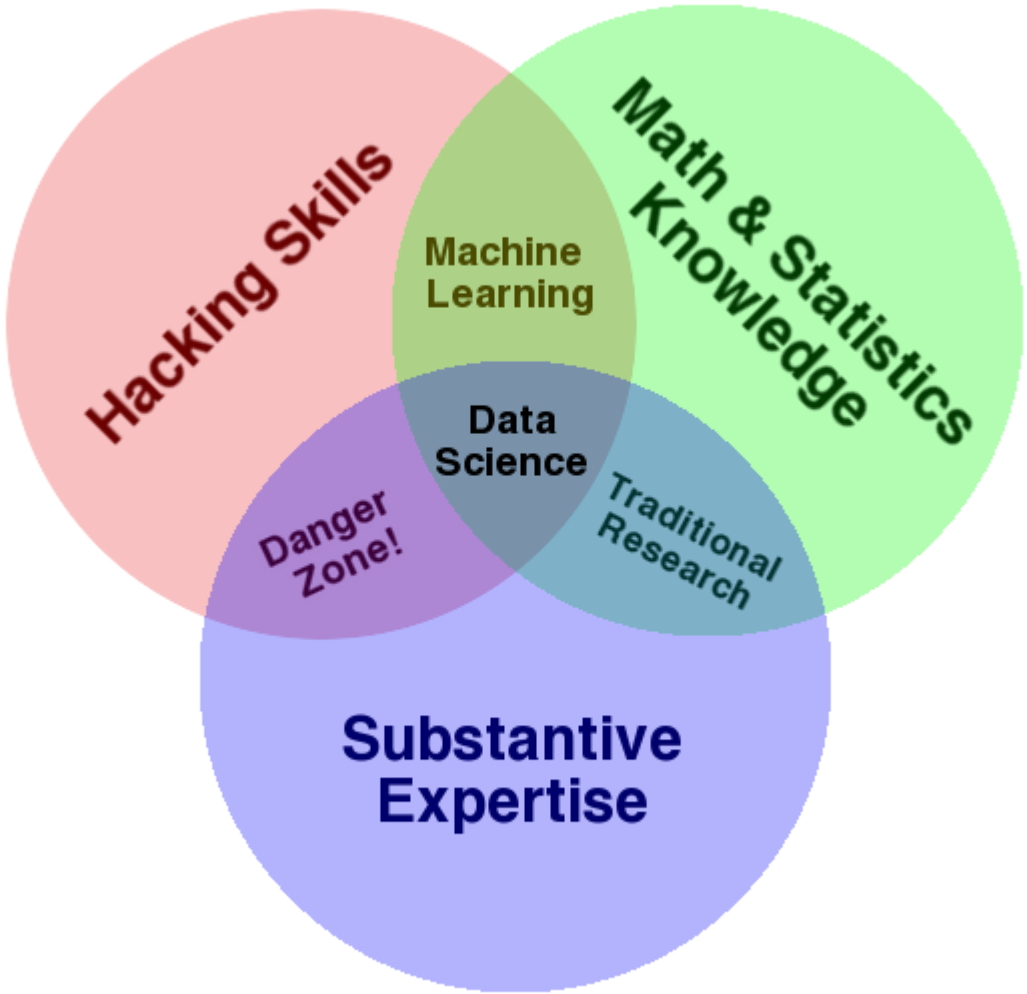
Nauka o podacima (Data science)

- Data science je multidisciplinarna oblast vezana za procese i sisteme ekstrakcije znanja ili sticanja uvida na osnovu struktuisanih i nestruktuisaih podataka u različitim formatima.
- Predstavlja nastavak nekih od oblasti analize podataka, kao što su: statistika, mašinsko učenje, istraživanje podataka, prediktivna analitika.
- U tom pogledu je bliska oblasti Knowledge Discovery in Databases (KDD) – otkrivanju znanja u bazama podataka.

Nauka o podacima (Data science)



Nauka o podacima (Data science)



Nauka o podacima (Data science)

- Mašinsko učenje je kolekcija algoritama i tehnika neophodnih za sintezu sistema koji uče na osnovu podataka.
- Algoritmi mašinskog učenja su veoma opšti, čvrsto zasnovani na matematici istatistici, i po pravilu ne uzimaju u obzir domensko znanje i preprocesiranje podataka.
- Stručnjaci iz domena nuke o podacima, odgovaraju na pitanje kako se nakon prikupljanja podataka, njihovog pročišćavanja (Data cleansing) i transformacije u pogodnu formu, na osnovu domenskog znanja biraju statističke metode i algoritmi mašinskog učenja u cilju rešavanja postavljenog problema.

Nauka o podacima (Data science)

- Ovaj proces može zahtevati u izvesnoj meri hakerske veštine u cilju dobijanja pravog smisla prikupljenih podataka.
- Vizualizacija podataka je značajan deo nauke o podacima stoga što se na taj način u proces interpretacije i razumevanja mogu uključiti i oni kojima su formalne metode statistike i mašinskog učenja nepoznate.
- Stručnjaci iz domena nauke o podacima moraju da poseduju znanja o tome koje algoritme mašinskog učenja treba da primene i na koji način. Samo poznavanje načina rada ovih algoritama nije neophodno. Naravno da dodatno poznavanje prirode i načina rada ovih algoritama predstavlja dodatnu vrednost.

Mašinsko učenje vs Istraživanje podataka

- Formalno, obe oblasti operišu sa gotovo istim skupom algoritama i tehnika.
- U čemu je razlika?
- Mašinsko učenje je fokusirano na predikciju zasnovanu na *poznatim* svojstvima naučenim na osnovu skupova podataka za obučavanje.
- Istraživanje podataka se fokusira na otkrivanje (prethodno) *nepoznatih* osobina podataka.

Mašinsko učenje vs nauka o podacima

Mašinsko učenje	Nauka o podacima
Razvoj novih (individualnih) modela	Ispitivanje više modela, formiranje i podešavanje hibridnih modela
Dokazivanje matematičkih svojstava modela	Razumevanje empirijskih svojstava modela
Poboljšanje/validacija na nekoliko, relativno čistiih, malih skupova podataka	Razvoj/korišćenje alata koji mogu da obradjuju masivne skupove podataka
Publikovanje rezultata	Preduzimanje akcije

Jeff Hammerbacher-ov model

1. Identifikacija problema

2. Instrumentalizacija izvora podtaka

3. Prikupljanje podataka

4. Priprema podataka (integracija, transformacija, čišćenje, filtracija, agregacija)

5. Formiranje modela

6. Evaluacija modela

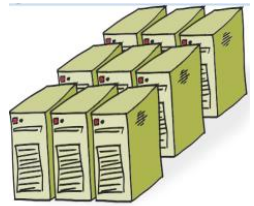
7. Komuniciranje rezultata

Praksa nauke o podacima

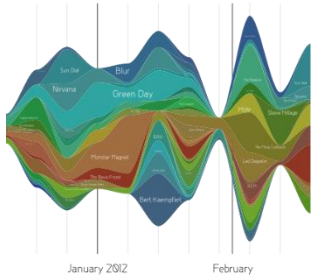
Kopanje po podacima

Čišćenje, preprocesiranje

Hipotetizacija modela



Ispitivanje na velikim uzorcima



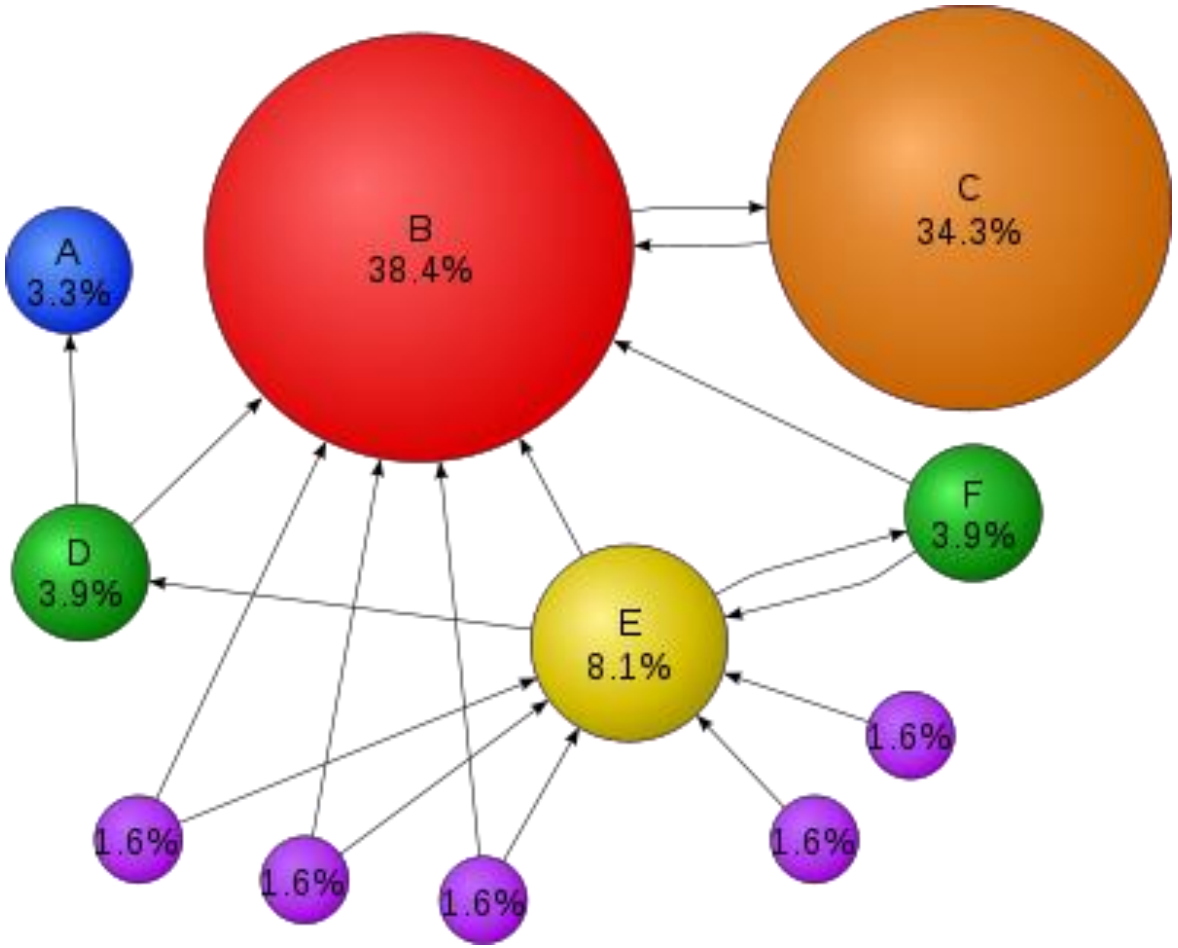
Evaluacija i Interpretacija

Šta je teško u nauci o podacima

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

Neki primeri primene mašinskog učenja

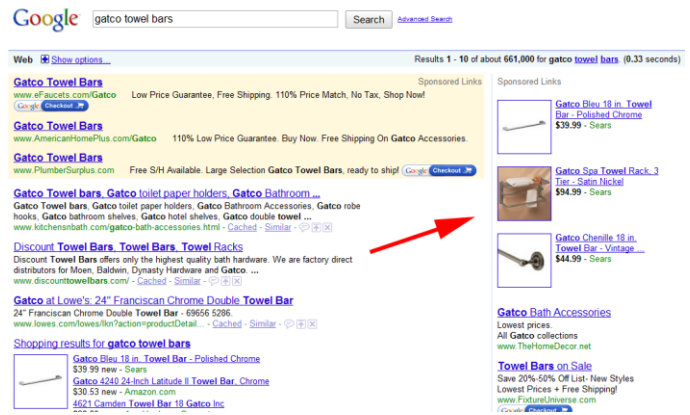
Pagerank: web kao bihevijoralni skupovi

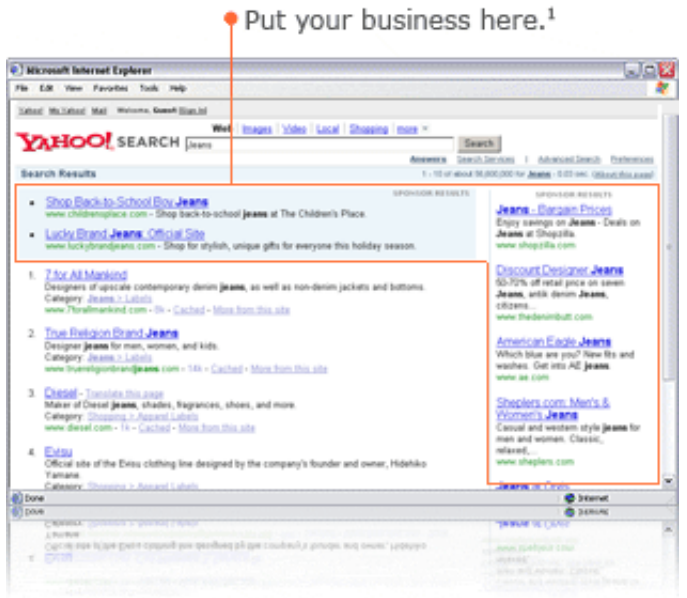


http://www.internetlivestats.com/total-number-of-websites/

Googl-ova farma server od
2 miliona mašina (procena)

1998 – sponzorisana pretraga





Pripisuje se Bill Gross-u, Idealabs

2002

Sponzorisana pretraga

- Googl/ov prihod od 50 milijardi $/godišnje od marketinga, predsytavlja 97% ukupnih prihoda kompanije.

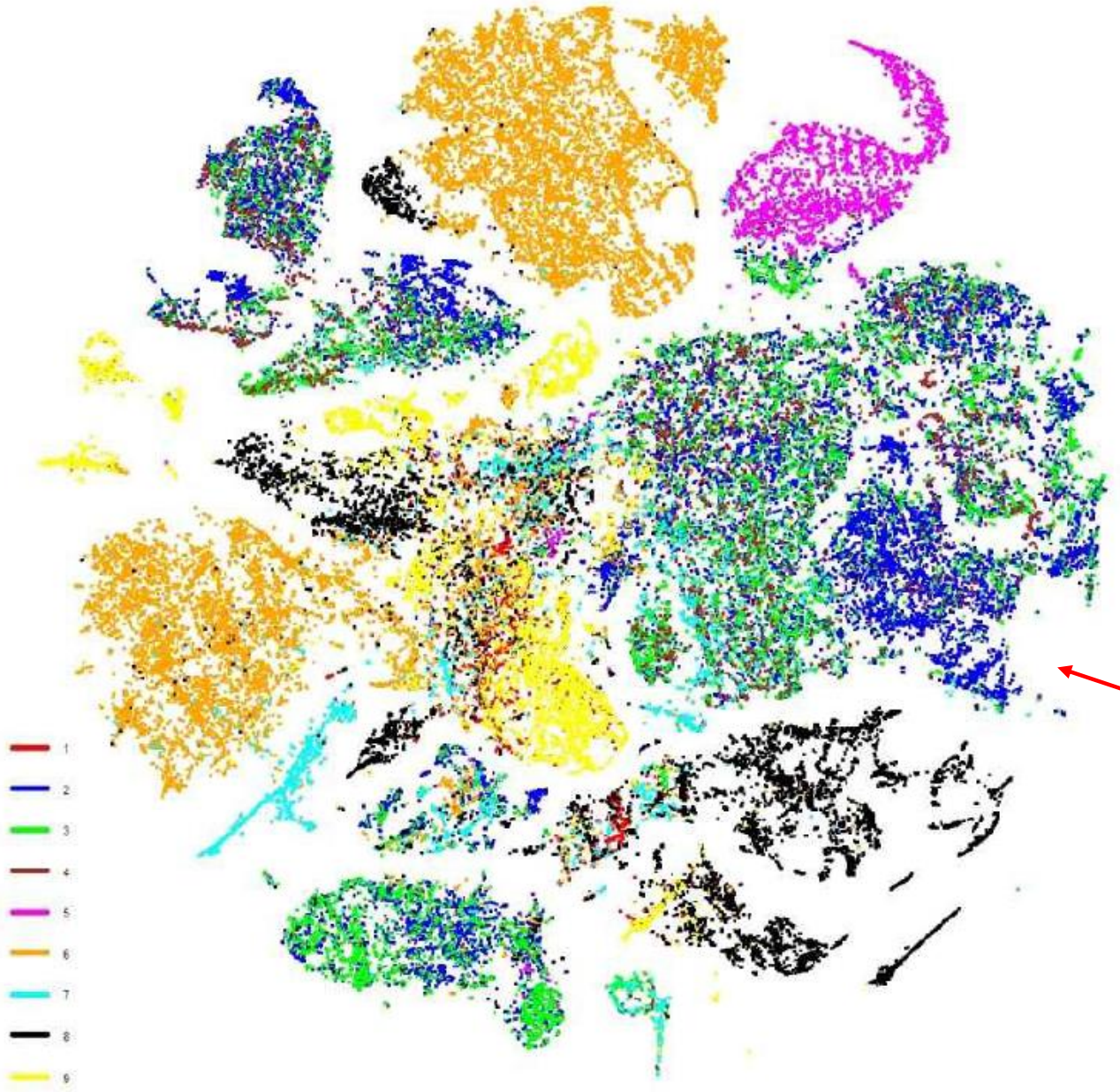
- Sponzorisana pretraga koristi aukciju – u okviru koje se nadmeću marketari koji pokušavaju da osvoje pristup korisnicima.

- Drugim rečima, nadmetanje **modela** potrošača – odnosno njihovih verodostojnosti odziva na datu reklamu – kao i odredjivanje odgovarajuće ponude.

- Trenutno se obavlja preko 30 milijardi pretraga svakog meseca.

https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

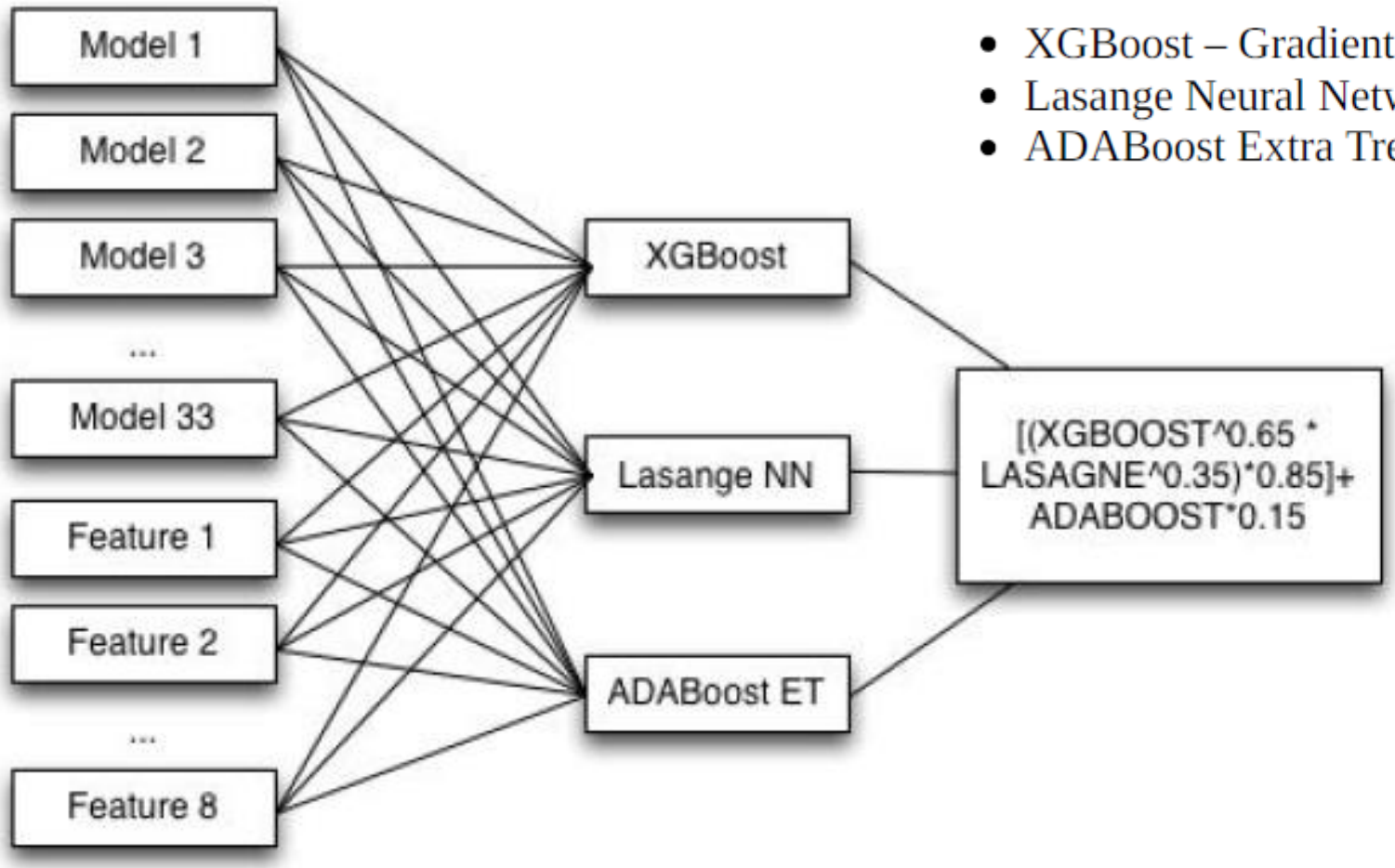
t-SNE



Veliki značaj vizualizacije velikih podataka (projektovanje u dve dimenzije sa minimalnom greškom)

Projekcija 93 dimenzionog prostora opisa proizvoda u dve dimenzije

Figure 16.4: Challenge Winning Ensemble



- XGBoost – Gradient boosting
- Lasange Neural Network – Deep learning
- ADABoost Extra Trees