ds dnevi slovenske informatike 20 let

"Dvajset let pozneje"

Kongresni center Grand hotel Bernardin
Portorož | Slovenija | 15. - 17. april 2013

20. konferenca
Dnevi slovenske informatike

„BIG DATA“

Velika priložnost za uradno statistiko?

Tomaž Špeh

17. 04. 2013

slovensko društvo informatika

MEDNARODNO LETO STATISTIKE
SODELUJOČA ORGANIZACIJA

Ozadje in motivacija

- Kaj predstavlja big data?
- Ali bi morala uradna statistika izkoristiti big data za proizvodnjo statističnih rezultatov, katerih?
- Kakšne so nevarnosti, če bo uradna statistika ostala neodzivna?
- Kakšni so izzivi uradne statistike z vidika varovanja informacijske zasebnosti, organizacije dela, kadrov, metodoloških in tehnoloških izzivov?
- Možni primeri uporabe big data tehnologij v uradni statistiki

Uradna statistika

- Uradna statistika je nepogrešliv del informacijskih sistemov v sodobnih demokratičnih družbah, ki zagotavlja državi, gospodarstvu in javnosti uporabne in nepristranske podatke o ekonomskih, demografskih in socialnih pojavih ter o okolju

- Izvajalci: državni statistični uradi (+pooblaščeni izvajalci), Eurostat, Svetovna banka, OECD, IMF in Združeni narodi

- Globalni standardi, metodologije in sistemi za proizvodnjo dobrih, mednarodno primerljivih in zanesljivih statističnih podatkov

- Zbiranje, statistična obdelava in objavljanje statističnih podatkov

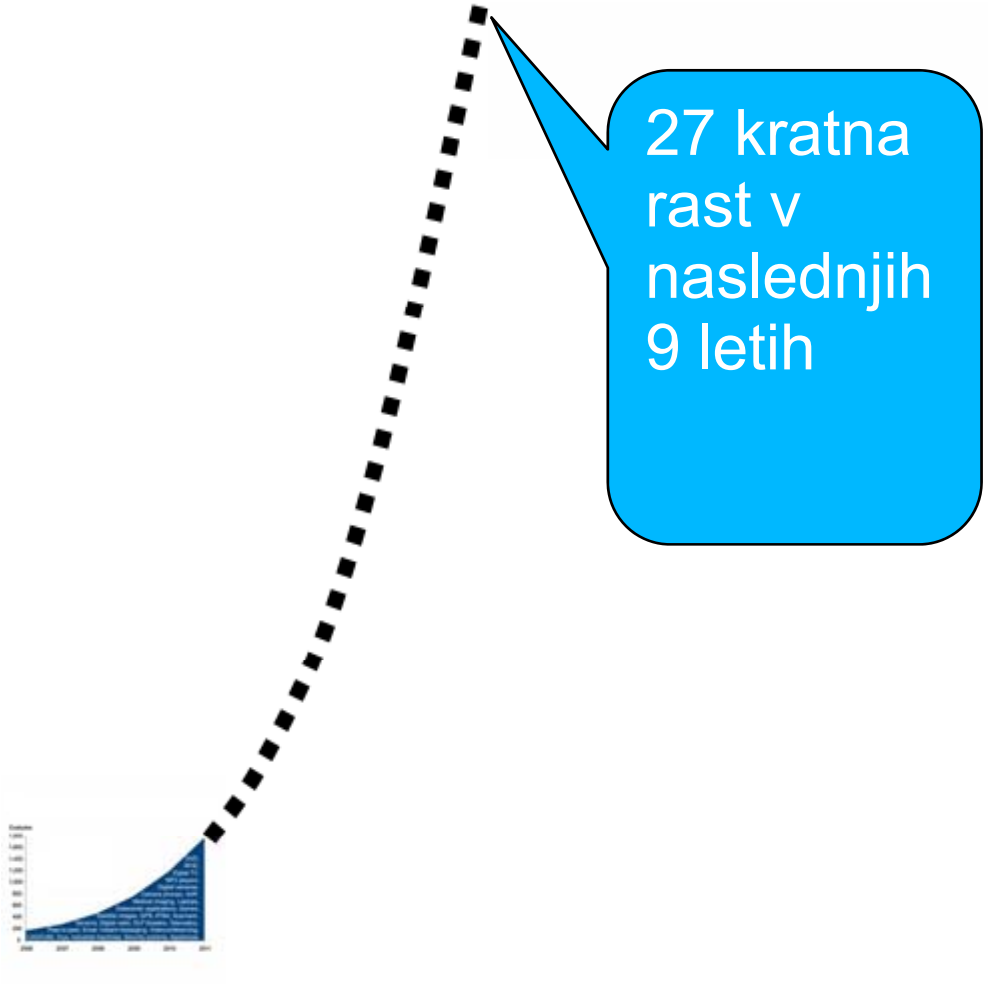
- Osnovni koncepti so populacije, vzorci ter vzorčenje te in verjetnost domnev.

MEDNARODNO LETO STATISTIKE

Kaj predstavlja big data?

- Natančna definicija (še) ne obstaja
- Obseg, hitrost, raznolikost, spremenljivost (angleško 4V)
- PB, EB podatkov in več, zelo majni časovni intervali
- Kompleksnost (nepopolni metapodatki, zahteva po čiščenju, povezovanju, enoličnem določanju...)
- Podatki, ki jih ni mogoče obdelati s klasičnimi, relacijskimi programskimi orodji in strojno opremo
- Spletni iskalniki, spletne strani, družbena omrežja, telekomunikacije, mobilne naprave in senzorji, itd.

Živimo v eksponentnem svetu!



50,000 EB do 2020

Zakaj?

Kakšne so nevarnosti, če bo uradna statistika ostala neodzivna?

- Google: Google price index, Google trends, Public data explorer,
- Pricestat.com,
- Podatki v realnem času, podatkovno rudarjenje po velikih količinah podatkov, prilagojene metodologije
- Privatni sektor se zaveda vrednosti podatkov, povezovanje podatkov
- Privatni sektor poskuša izračunati lastne statistične proizvode, take, ki so bili še do nedavnega izključno domena uradne statistike

Vplivi uporabe big data na procese v uradni statistiki

STRUKTURA OPISA PROCESA, SURS, december 2011

1	2	3	4	5	6	7	8
ANALIZA POTREB IN ZAHTEV	NAČRTOVANJE IN PRIPRAVA STATISTIČNEGA RAZISKOVANJA	IZBOR ENOT OPAZOVANJA	ZBIRANJE PODATKOV	STATISTIČNA OBDELAVA PODATKOV	ANALIZA PODATKOV	DISEMINACIJA IN KOMUNICIRANJE Z UPORABNIKI	DOKUMENTIRANJE IN EVALVACIJA RAZISKOVANJA
Ugotavljanje potrebe po podatkih	Načrtovanje sredstev in določitev seznama aktivnosti z roki	Priprava podatkovnih virov za izgradnjo vzorčnega okvira	Priprava na zbiranje	Urejanje administrativnih podatkovnih virov	Analiza časovnih vrst	Posodabljanje izhodov	Izdelava dokumentacije o raziskovanju
Preučitev virov	Definiranje rezultatov raziskovanja	Priprava vzorčnega okvira	Zbiranje podatkov in komuniciranje s poročevalskimi enotami	Urejanje podatkov na mikroravni	Analiza ustreznosti ter potrditev rezultatov	Predstavitev rezultatov	Zbiranje informacij za oceno kakovosti
Preverjanje metodologije	Priprava metodologije izbora enot opazovanja	Izbor enot opazovanja	Prevzem administrativnih virov	Integracija različnih podatkovnih virov	Interpretacija rezultatov	Objava	Izvedba ocene postopkov in procesov
	Priprava metodologije statistične obdelave podatkov	Izdelava adresarja	Zajem podatkov	Vstavljanje podatkov (imputacije)		Podpora uporabnikom	
	Organizacija sodelovanja z zunanjimi institucijami in načrtovanje prevzema administrativnih virov			Deflacija			
	Načrtovanje in testiranje vprašalnika			Uteževanje			
	Priprava gradiv za komuniciranje s poročevalsko enoto			Izračun statističnih ocen (agregacija)			
	Načrtovanje in izvedba pilotnega raziskovanja			Urejanje podatkov na makroravni			
				Priprava tabel			
				Statistična zaščita podatkov			

- Definiranje ciljne populacije, statistik in statističnih domen (tradicionalni pristop - določitev vzorčnega načrta in zbiranje podatkov, big data pristop – podatki nastali za drug namen se uporabijo za pripravo statistik)

- Potrebni novi procesi: podatkovno rudarjenje, strojno učenje in napredne vizualizacijske metode – narediti big data majhne in jih povezati s tradicionalnimi viri

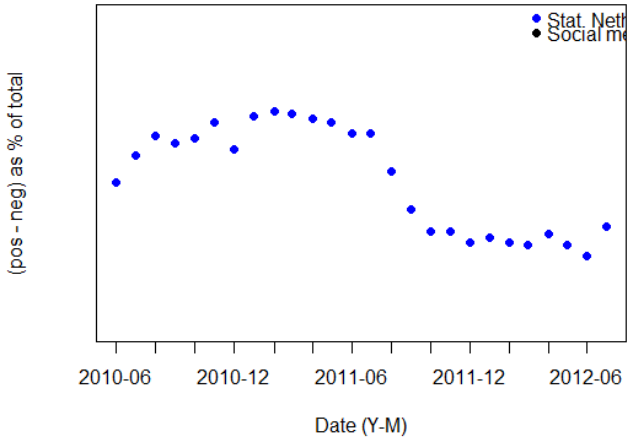
Izzivi pri uporabi BIG DATA v uradni statistiki

- Zakonske podlage za dostop do podatkovnih baz: Zakon o državni statistiki in veljavni programi statističnih raziskovanj (srednjeročni in letni)
- Varovanje informacijske zasebnosti in statistične zaupnosti.
- Organizacijski: novi procesi in standardi, pridobivanje in izobraževanje zaposlenih

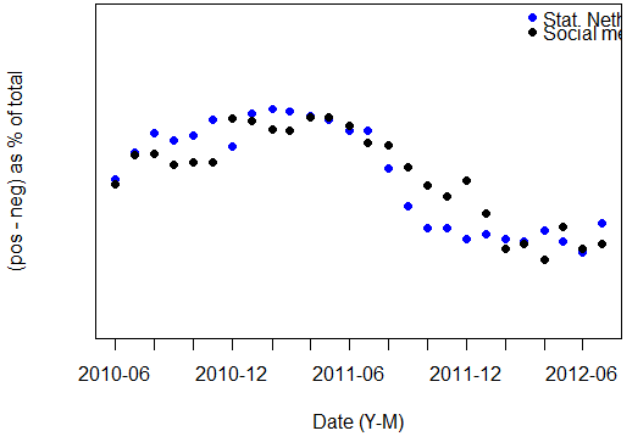
Možnosti za uporabo big data v uradni statistiki

- Kvaliteta in primernost za uporabo: natančnost, pravočasnost, popolnost, dolgoročna stabilnost, dostopnost

Klasični vzorčni pristop

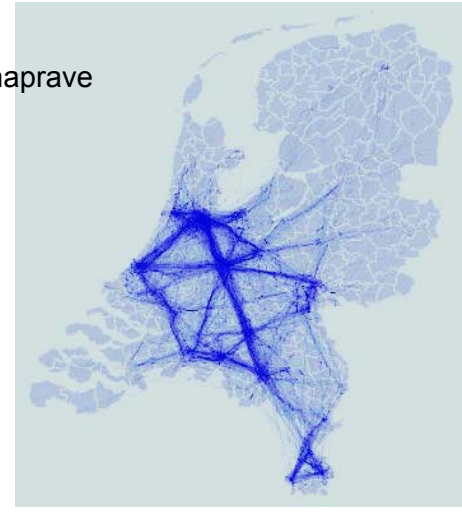


Big data pristop : twitter



Prometni senzorji

Mobilne naprave



CBS Nizozemska

Tehnološki izzivi pri uporabi big data v uradni statistiki

- Uvedba porazdeljenih datotečnih sistemov (Hadoop, HDFS)
- Paralelno procesiranje, ki ni vezano na programski jezik (MapReduce, vmesniki za programska orodja)
- Podatkovne baze z ohlapnimi strukturami podatkovnih modelov, ki se lahko hitro prilagodijo spreminjajoči strukturi podatkov, Not only SQL baze: Hbase, Hive, BigTable (na račun točnosti!)

Uporabnost big data tehnologij

```
/*mapiranje nove Hive tabele na obstoječe HDFS podatke*/
proc sql;
connect to Hadoop (server=hxpduped);
exec( create external table hadoop1( x double, y string, z double) row format
delimited fields terminated by '01' stored as textfile location
'/tmp/hadoop1_hdfs_file') b
y hadoop;
quit;
```

```
libname x hadoop (server=hxpduped);
data x.hadoop1;
Set
work.my_stat_data;
run;
```

- Kompleksne tehnologije, ki zahtevajo dobro usposobljene IT strokovnjake za vzpostavitev in vzdrževanje, oblačna infrastruktura
- Statistiki morajo pridobiti znanja o novih poizvedovalnih in analitičnih pristopih za delo z nosql bazami
- Analiza uporabnosti takih zbirk podatkov je lahko zelo zahtevna

Zaključek

1. Big data podatki omogočajo:

- Nove statistike, ki jih uradne statistike še ne izračunavajo
- Hitrejši odzivi na spremembe
- Možna finančna razbremenitev uradnih statistik
- Razbremenitev enot opazovanja

2. Big data tehnologije omogočajo:

- Učinkovitejše in hitrejše izvajanje procesno zahtevnih statističnih postopkov (record linkage)
- Nove načine izvajanja logičnih kontrol in urejanja podatkov (izpeljava logičnih kontrol na podlagi vzorcev big data podatkov, odprava nekonsistentnosti klasičnih virov s povezovanjem z big data podatki)

3. Potrebno bo vzpostaviti nove procese in metodologije in osvojiti nove tehnologije

MEDNARODNO LETO STATISTIKE

Hvala za vašo pozornost !

Vprašanja?

Pripombe?

Predlogi?

tomaz.speh@gov.si