

Dátová veda - z teórie do praxe

Iveta Stankovičová, Fakulta managementu UK Bratislava

eFOCUS



Dáta

- **Informačná explózia** – dnešná doba
- Dáta sú tvorené:
 - pre vedecké účely (historicky, pre dátové analýzy) – **experimentálne dáta**
 - ako operačné dáta (dnes, v biznise) – **príležitostné, resp. oportunistické dáta** (Huber 1977)

príležitostný - casual, occasional, opportunistic

Dáta 2

	Experimentálne dáta	Operatívne dáta
Účel	vedecký	operačný
Hodnota	vedecká	komerčná
Tvorba	aktívne kontrolovaný proces	pasívne napozorované
Veľkosť	malá	masívna
Čistota	čisté	špinavé
Stav	statické	dynamické

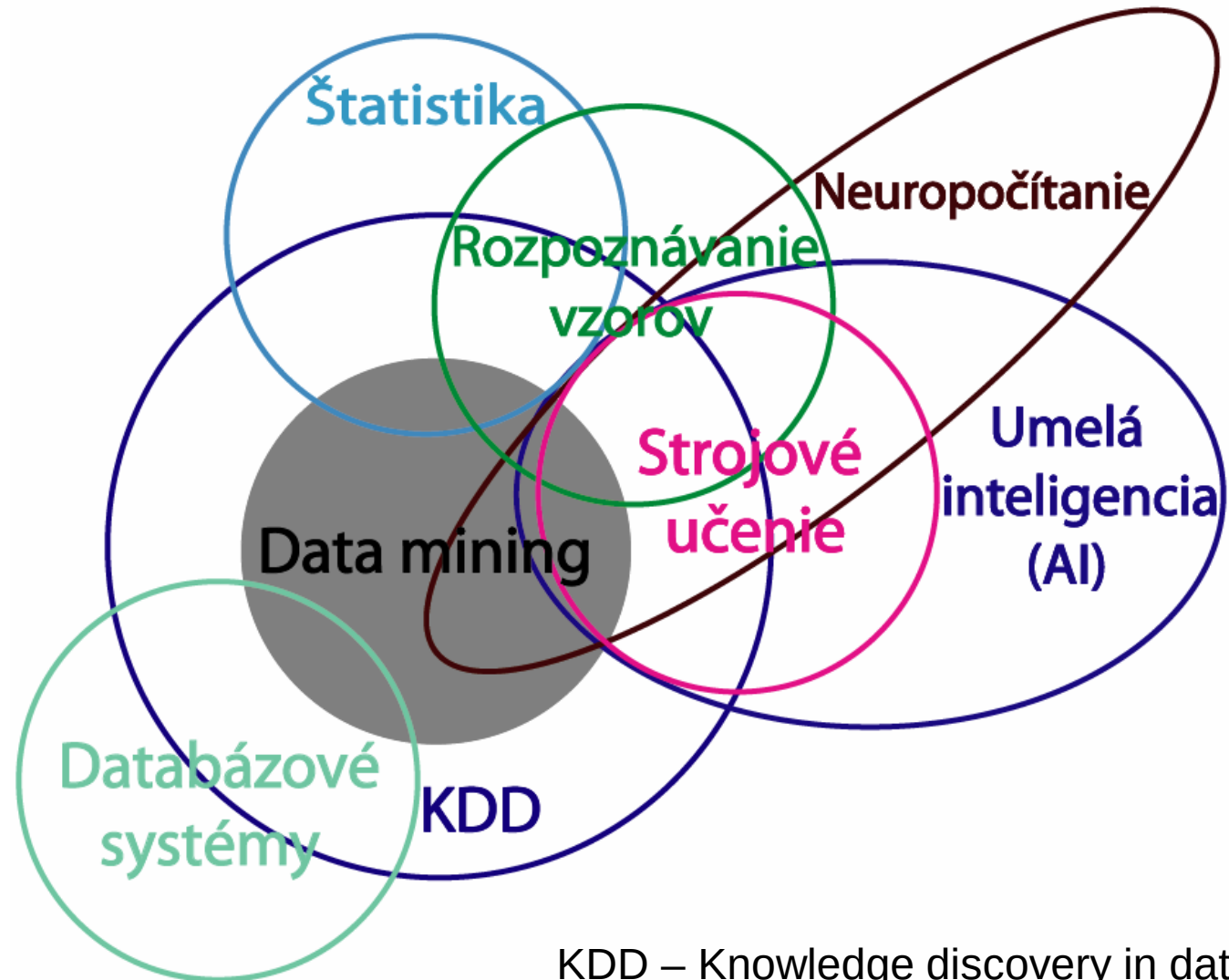
Dáta (Big Data) → Informácie

- Je nutné získavať informácie z veľkých objemov operačných údajov pre rozhodovanie sa manažérov (business decision support)
- Je nutné skúmať a modelovať vzťahy v dátach pomocou prediktívneho modelovania (základná úloha)

Data Modelling = Data Mining (DM)

(cca 1963)

Interdisciplinarita DM



Vlastnosti Big Data

- veľmi veľký objem údajov (masívnosť)
- vysoká dimenzionalita údajov (mnohorozmernosť) alebo
- neštruktúrovanosť údajov (texty)

Infostat: Big Data je koncept, ktorý vznikol na základe obrovského množstva neštruktúrovaných dát, ktoré denne vznikajú a je ich problém spracovať tradičnými metódami spracovania dát. Základné charakteristiky Big Data:

- **Množstvo** (Volume) – veľké objemy dát, typicky začínajú na desiatkach terabytov
- **Rýchlosť** (Velocity) – rýchlosť ako sú dáta tvorené poprípade upravované
- **Pestrosť** (Variety) – rôznorodosť dátových formátov

Pokiaľ dáta spĺňajú aspoň 2 z uvedených charakteristík = Big Data.

Ciele analýzy Big Data

mnohorozmernosť údajov:

- vyvinúť účinné modely/metódy pre predikciu budúcich hodnôt
- pochopiť vzťahy medzi súčasnosťou a budúcnosťou

masívnosť údajov:

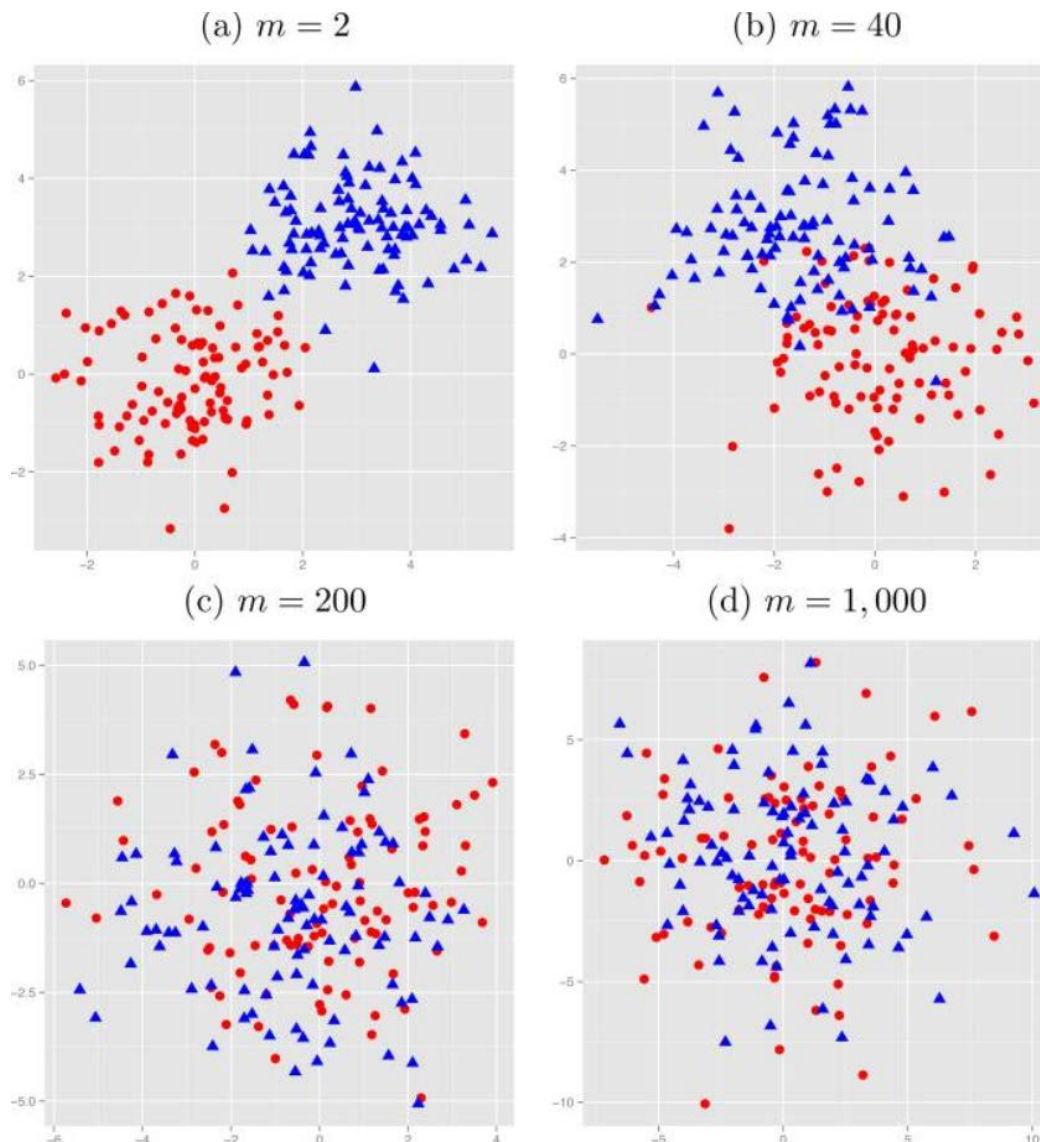
- prieskum skrytých štruktúr jednotlivých čiastkových populácií - špecifiká, individuálne rozdiely, odľahlé hodnoty
- extrakcia dôležitých spoločných vlastností z čiastkových populácií, aj keď existujú veľké individuálne rozdiely

Výzvy pre analýzy Big Data

mnohorozmernosť údajov + masívnosť údajov prináša:

- hromadenie nepresností a šumov
- zdanlivá korelácia
- náhodná homogenita
- nestabilita tradičných algoritmov – potreba vyvíjať nové
- kombinácia údajov z rôznych zdrojov a z rôznych časov – heterogenita, štatistické chyby ... – vývoj viac adaptívnych a robustných postupov a metód
- výpočtové nároky - na HW a SW – zvyšovanie nákladov

Výzvy pre analýzy Big Data – obr. 1



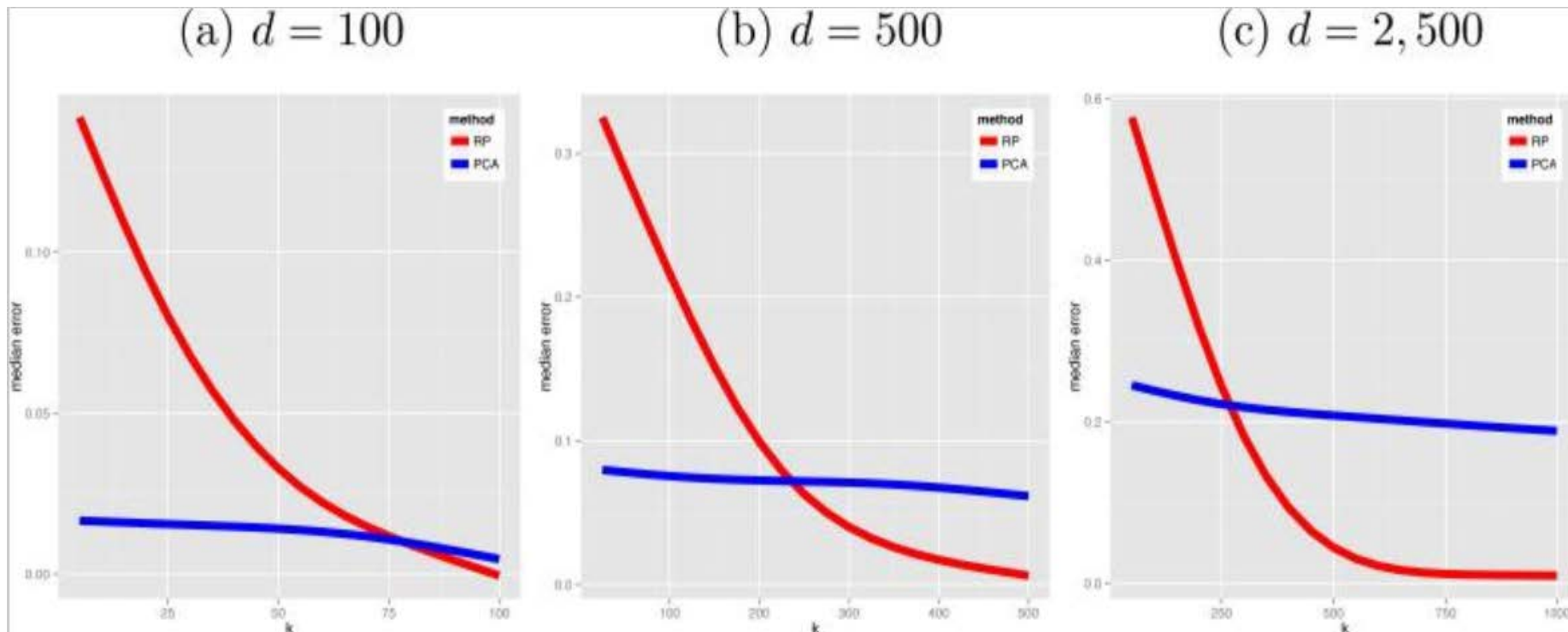
Scatter plots of projections of the observed data ($n = 100$ from each class) onto the first two principal components of the best m -dimensional selected feature space.

A projected data with “●” indicates the first class and “▲” indicates the second class.

Zdroj:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4236847/figure/F1/>

Výzvy pre analýzy Big Data – obr. 2



Plots of the median errors in preserving the distances between pairs of data points versus the reduced dimension k in large scale microarray data. Here “RP” stands for the random projection and “PCA” stands for the principal component analysis.

Zdroj: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4236847/figure/F11/>

Štatistické metódy pre Big Data

3 skupiny metód - algoritmov:

1. na základe prevzorkovania (resampling-based)
2. „rozdeľ a panuj“ („divide and conquer“) -
3. on-line aktualizácie pre dátové toky (online updating)

Štatistické metódy pre Big Data 1

Metódy/algoritmy na základe prevzorkovania (resampling-based):

- Bags of little bootstrap
- Leveraging
- Mean log-likelihood
- Bayesian inference
- Markov chain Monte Carlo (MCMC) ...

Štatistické metódy pre Big Data 2

Metódy/algoritmy na základe pravidla „rozdeľ a panuj“ („divide and conquer“) – 3 kroky:

- 1) veľké dátové súbory sa rozdelia do K blokov
 - 2) každý blok sa spracuje zvlášť (prípadne paralelne)
 - 3) finálne riešenie pre všetky údaje sa dostane agregáciou čiastkových riešení z každého bloku údajov
-
- Aggregated estimating equations
 - Majority voting
 - Screening with ultrahigh dimension
 - Parallel MCMC

Štatistické metódy pre Big Data 3

Metódy/algoritmy na základe on-line aktualizácie pre dátový tok (online updating):

- údaje prichádzajú na spracovanie po častiach alebo v tokoch
– model, resp. analýza sa musí postupne aktualizovať a je to potrebné urobiť bez skladovania (uchovávanía) údajov

Nároky na analytikov Big Data 1

Tradičné prístupy:

- prístupy „pokus – omyl“
- statické analýzy dát

Prístupy pre Big Data:

- Machine Learning – systémy schopné sa učiť z dát (štatistika, algoritmy, optimalizácia)
- prediktívne analýzy v reálnom čase
- vysoká prispôsobivosť systému

Nároky na analytikov Big Data 2

- **Študovať teóriu – metódy/algoritmy a aj podmienky ich použitia**
 - matematika, štatistika, algoritmy, optimalizácia, umelá inteligencia, učenie sa strojov, genetické algoritmy, ...
- **Poznať oblasť aplikácie**
 - ciele analýzy, vznik a obsah údajov, výber údajov, výber premenných, ...
- **Vedieť si vybrať a pripraviť dáta**
 - kde a ako sú dáta uložené, SQL, metódy transformácie údajov, úprava premenných, odvodené premenné, ...
- **Vedieť pracovať s HW a SW**
 - ovládať viac ako 1 SW, vedieť vybrať vhodné HW a SW

Použité zdroje

- Jianqing Fan, Fang Han, and Han Liu: **Challenges of Big Data Analysis.**
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4236847/>
- Chun Wang, Ming-Hui Chen, Elizabeth Schifano, Jing Wu, and Jun Yan: **Statistical methods and computing for big data.**
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5041595/>
- Cheng Meng, Ye Wang, Xinlian Zhang, Abhuyday Mandal, Ping Ma: **Effective Statistical Methods for Big Data Analytics.**
[http://faculty.franklin.uga.edu/amandal/sites/faculty.franklin.uga.edu.amandal/files/Effective Statistical Methods for Big Data Analytics.pdf](http://faculty.franklin.uga.edu/amandal/sites/faculty.franklin.uga.edu.amandal/files/Effective%20Statistical%20Methods%20for%20Big%20Data%20Analytics.pdf)
- Infostat, 2014. Robert Suja: **Big Data.**
http://www.infostat.sk/web2015/sk/publikacie/Big_Data.pdf

Ďakujem za pozornosť

