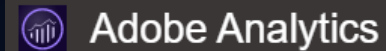


Analýza a meranie správania sa návštevníka vo webovej aplikácii

PETER TRUCHAN
MÁRIA BIELIKOVÁ

STU FIIT 14.06.2016

Takmer každá väčšia webová aplikácia meria správanie svojich návštevníkov.

The logo for Google Analytics, featuring the word "Google" in its multi-colored font followed by "Analytics" in a dark grey sans-serif font.The logo for clicktale, with the word "click" in white and "tale" in a lighter purple font, separated by a less-than sign, all on a solid purple rectangular background.The logo for compete, featuring a stylized blue and green circular icon to the left of the word "compete" in a bold, dark grey sans-serif font. Below "compete" is the text "Millward Brown Digital" in a smaller, lighter font.The logo for Adobe Analytics, featuring a small purple square icon with a white bar chart symbol to the left of the words "Adobe Analytics" in a white sans-serif font, all on a dark grey rectangular background.The logo for crazyegg, with the word "crazyegg" in a bold, dark grey sans-serif font. To the right of the text is a green egg-shaped icon with a white "C" inside, and a small trademark symbol (TM) is at the bottom right.The logo for Optimizely, featuring a stylized white "O" with a circular arrow inside, followed by the word "optimizely" in a white sans-serif font, all on a solid teal rectangular background.

Problematika merania a vyhodnocovania je veľmi obsiahla

- Získavanie dát
- Spracovanie dát
- Ukladanie dát
- Vizualizácia dát
- Vyhodnocovanie dát

Nedostatok výskumu v oblasti spracovania dát získaných z analytických skriptov

prediction	Google Scholar	ACM	IEEE	Springer
+ server logs	50,000	31,193	11,802	10,311
+ browser extension	223,000	37,849	1,802	3,275
+ analytical/browser scripts	18,100	17,773	2,423	4,738

Doména má špecifické vlastnosti

- Ochrana osobných údajov
- Množstvo dát
- Neúplnosť dát
- Chybovosť dát

Výstup práce

1. Spracovanie dát
2. Uloženie dát
3. Výpočet charakteristík návštevníka
4. Segmentácia návštev
5. Obohatenie existujúcich predikčných modelov o ďalší atribút

Rozdelenie implicitných charakteristík

- Charakteristiky závislé od aktuálnej activity
- Charakteristiky nezávislé od aktuálnej aktivity
- Časové charakteristiky
- Zoznam charakteristík stránky

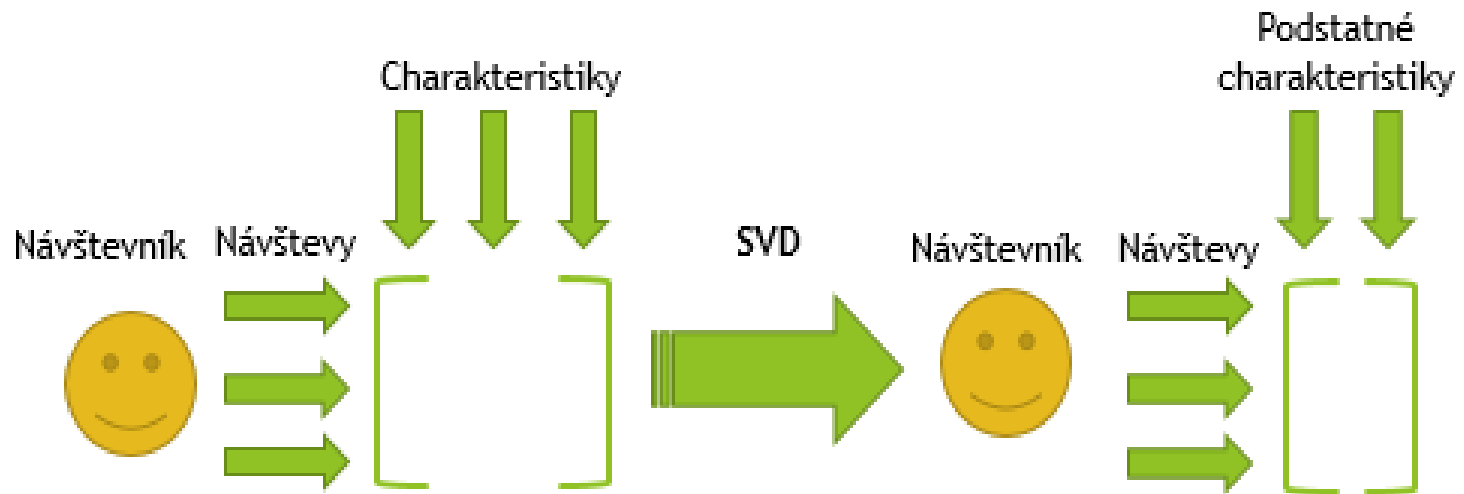
Matica návštev a ich charakteristík

	1. Charakteristika návštevníka			n. Charakteristika návštevníka
návštevník 1	návšteva 1	0.60	...	0.54
návštevník 1	návšteva 2	0.81	...	0.84
návštevník 1	návšteva 3	0.56	...	0.64
návštevník 2	návšteva 1	0.82	...	0.93
návštevník 2	návšteva 2	0.51	...	0.76
návštevník 3	návšteva 1	0.31	...	0.70
návštevník 3	návšteva 2	0.44	...	0.54
návštevník 3	návšteva 3	0.35	...	0.78

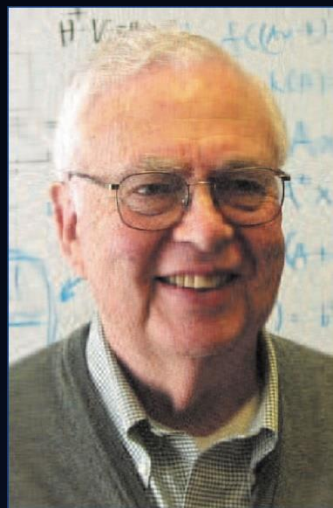
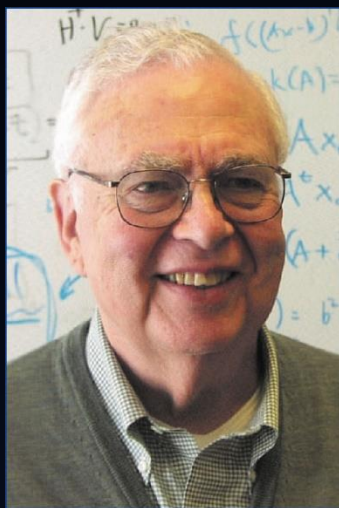
Charakteristika domény

- 130 000 používateľov
- 11 080 charakteristík
- 1100 stránok
- 9 000 000 návštev
- 3 roky merania

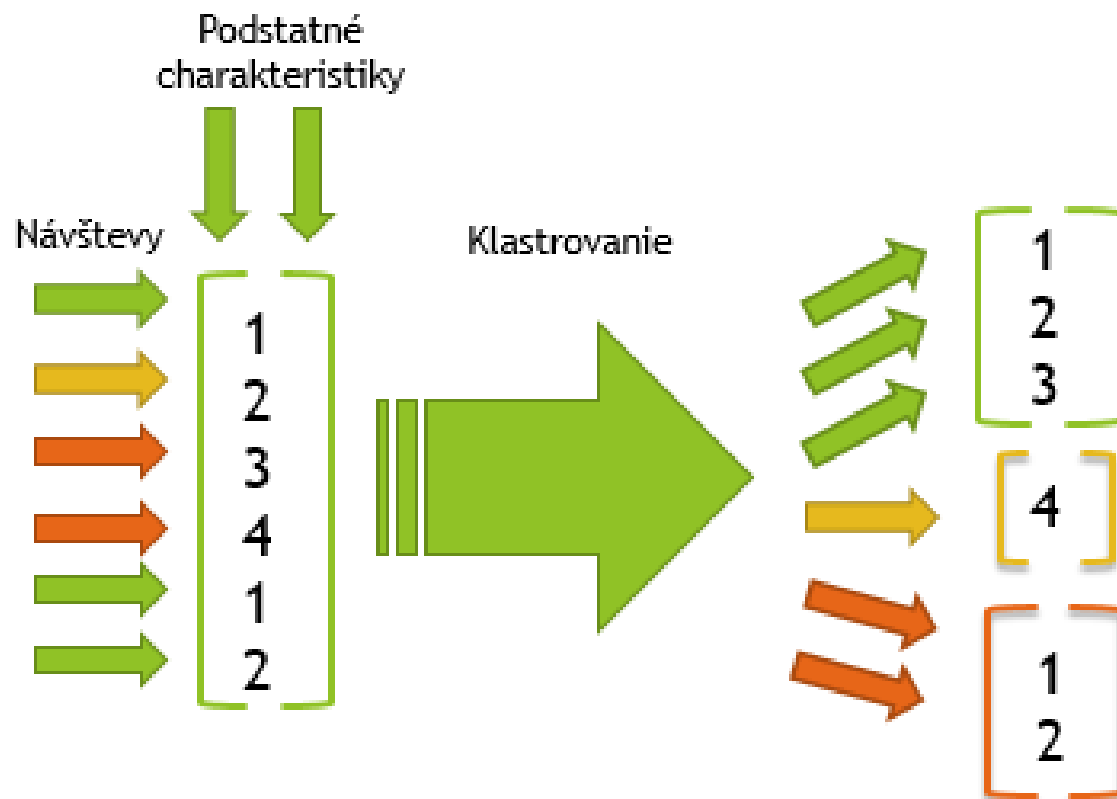
Hlavnou výzvou bola redukcia redundantných informácií a spracovanie veľkého množstva dát



Redundantné informácie sme odstránili
pomocou upraveného algoritmu singulárnej
dekompozície matice

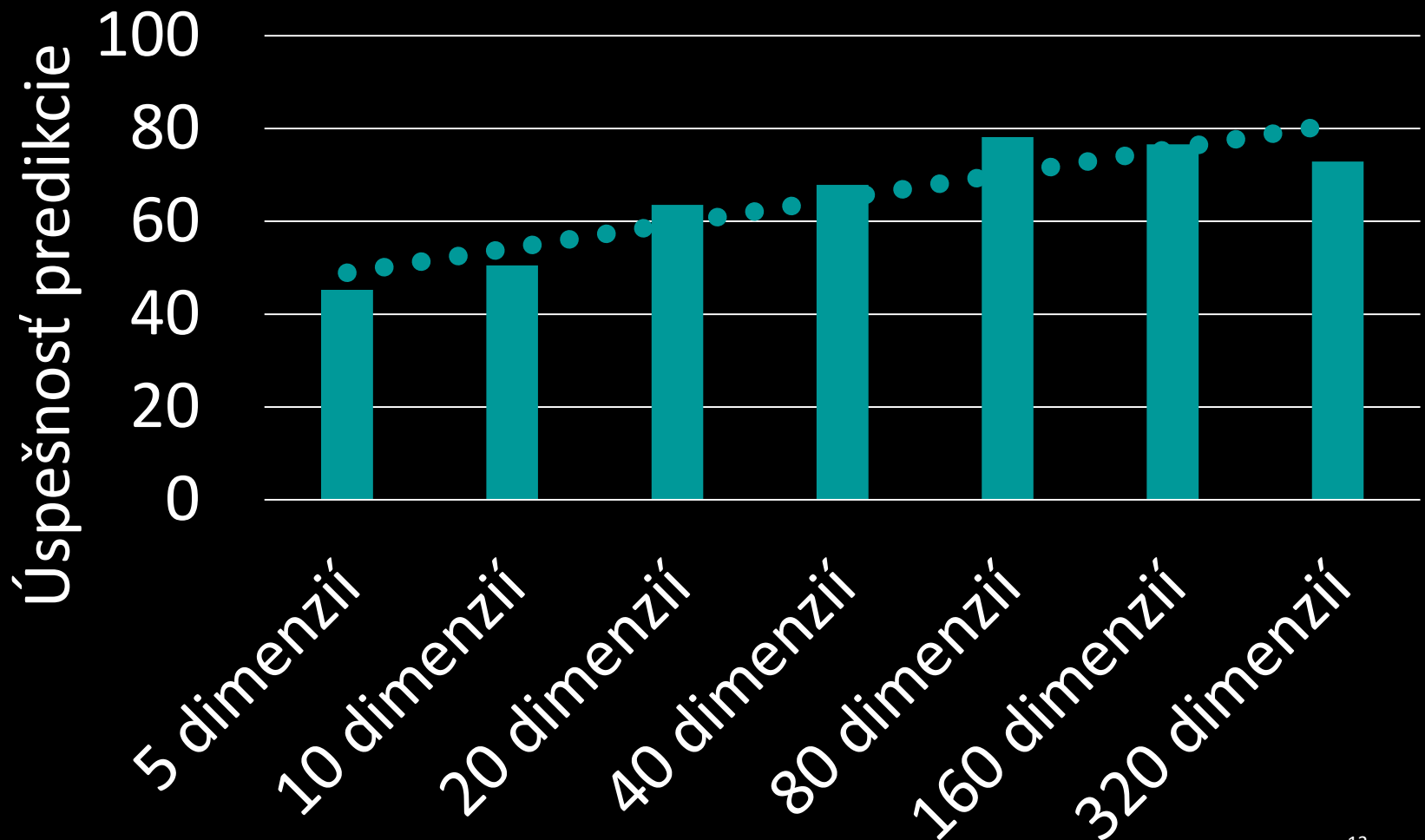


Klastrovacia úloha



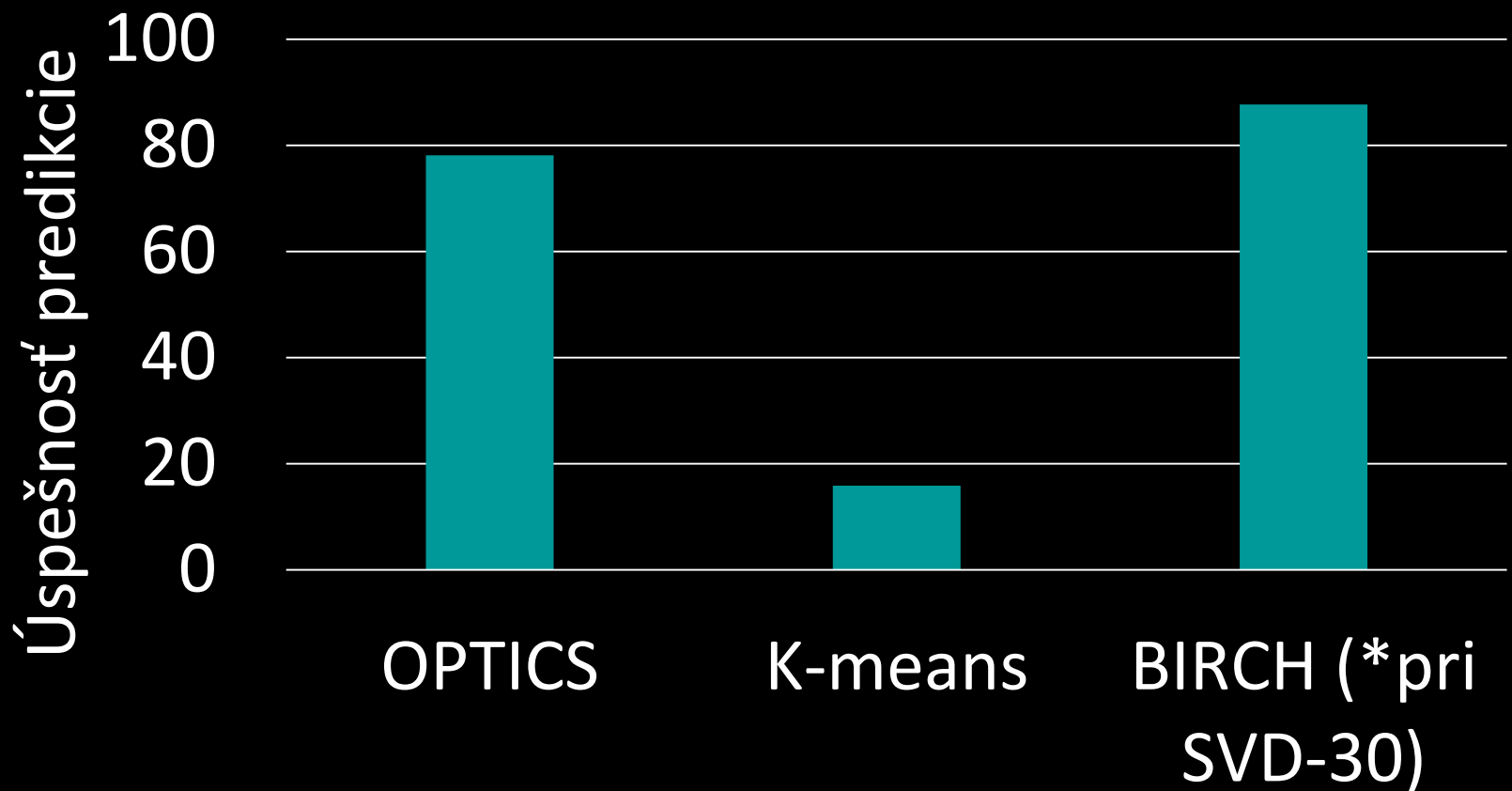
Vývoj úspěšnosti (R5) na základe počtu výstupných dimenzií SVD

1 000 000 návštev, DBSCAN, TraMineR



Porovnanie klastrovacích algoritmov

SVD-80, 1 000 000 vzorka dát



Porovnanie algoritmov na določenie sekvencií SVD 80 dimenzií, 500 000 návštev, DBSCAN



VÝSLEDOK PRÁCE

- 130 000 používateľov
- 11 080 charakteristík
- 1100 stránok
- 9 000 000 návštev
- 3 roky merania

-
- Zmerané, spracované, vyčistené, stiahnuté charakteristiky
 - Implementovaný model segmentácie návštevníkov do skupín
 - Overenie riešenia

Vývoj úspěšnosti s přibývajícím objemom dát 8000 dimenzí, SVD-80, DBSCAN, TraMineR

