Predikcia správania sa návštevníka v prostredí webovej aplikácie

Bc. Peter Truchan
prof. Ing. Mária Bieliková, PhD.
PEWE
17.3.2016

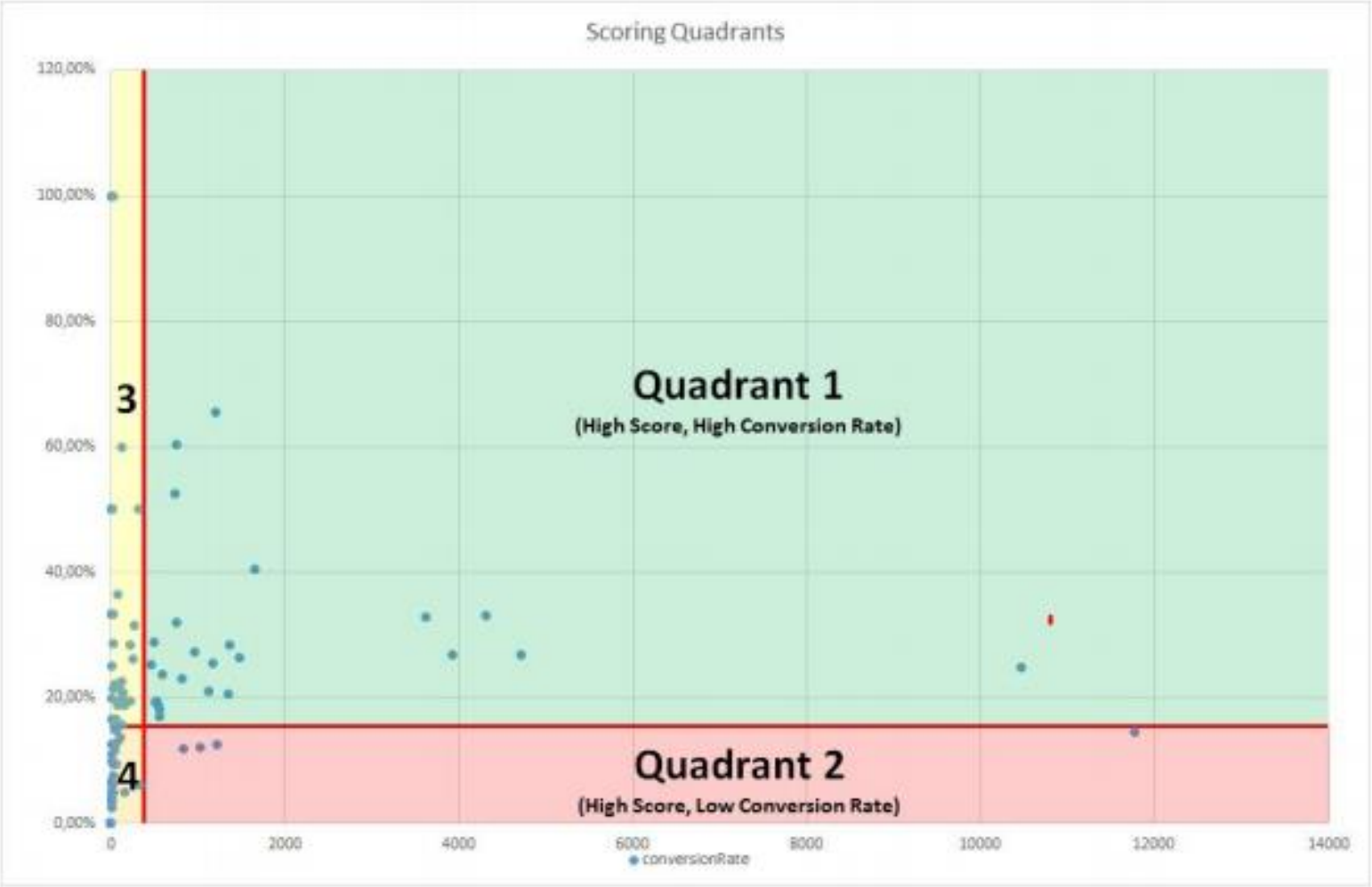
Motiváciou je identifikovať zaujímavé segmenty návštevníkov

Správanie sa návštevníka vo webovej aplikácii je možné predpovedať na základe jeho aktuálneho stavu, demografických, geografických, *socio-ekonomických*, technických charakteristík a najmä na základe jeho predchádzajúcich akcií a charakteristík.

Neuvažujeme vnútorný stav návštevníka – emócie, pocity.

Predchádzajúce prístupy korelačná matica

Predchádzajúce prístupy
Page Rank

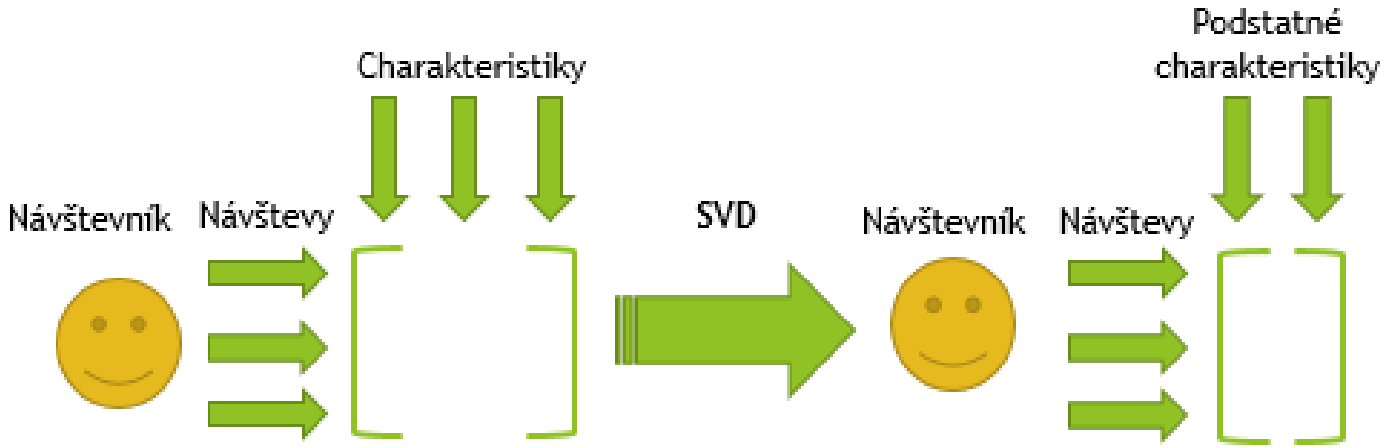


Predchádzajúce prístupy „tradičný“ machine learning

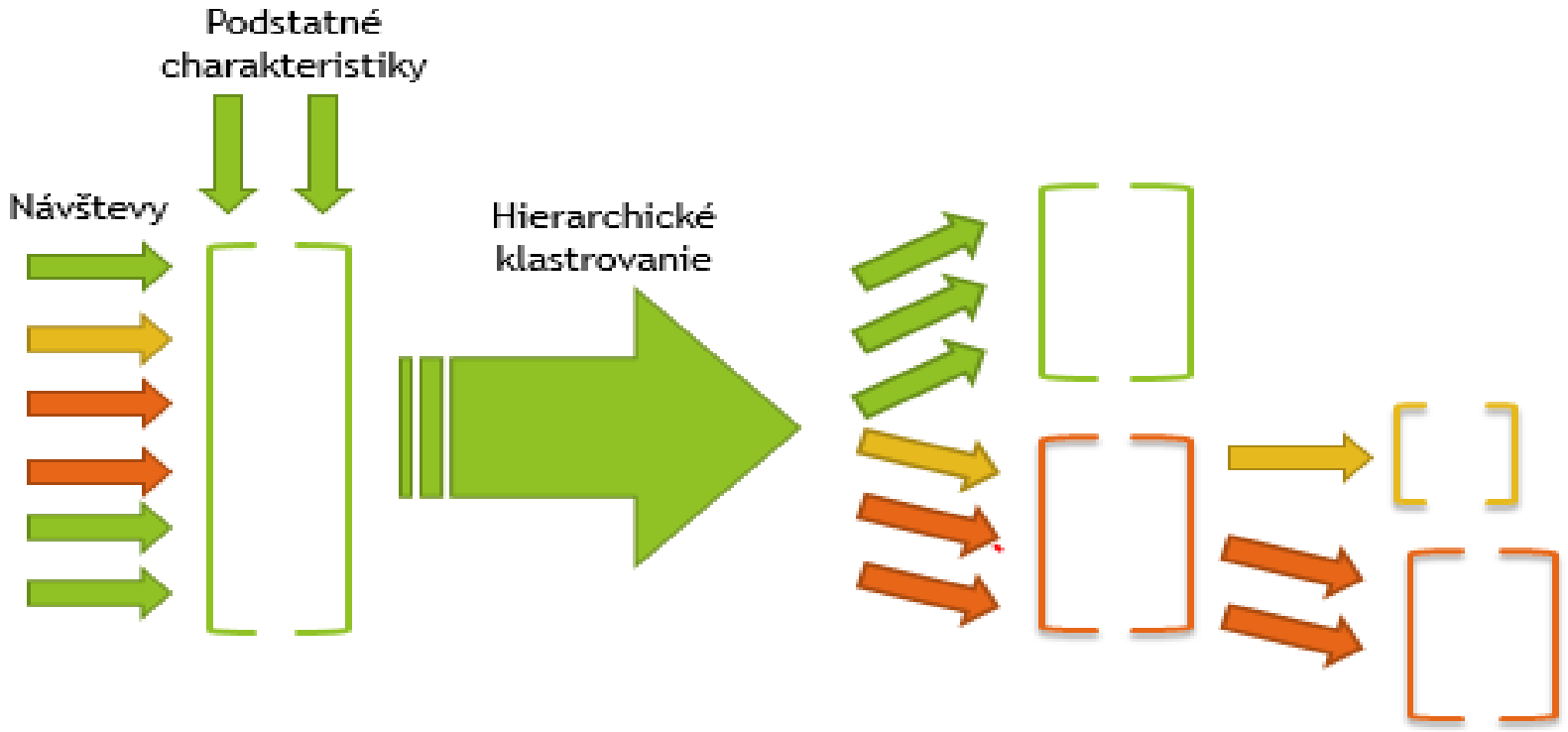
Aktuálna metóda

- Dáta
- Odstránenie korelujúcich metrík a charakteristík
- Segmentácia
- Sekvencia návštev
- Vizualizácia

Aplikácia na náš problém



Hierarchické klastrovanie BIRCH O(N)



Otázky

- „Veľké dáta“?
- **Ako na overenie metódy?**
- Markovova skrytá reťaz, Bayes, alebo iné metódy na reprezentáciu sekvenčných informácii.

Množstvo dát

- Viac ako 130 000 unikátnych klientov
- Viac ako 1 000 000 návštev
- Matica 1 000 000 * 1000 float hodnôt = GB
- RAM = 15GB SwapFile na SSD=25GB
- SPOLU=40GB
- **Vygenerovanie matice a naplnenie náhodnými hodnotami = O(N) = 220s**
- **Algoritmus s O(N^2) = 220 * 220 = 14 hodín**
- **Algoritmus s O(N^3) = 220 * 220 * 220 = 123 dní**

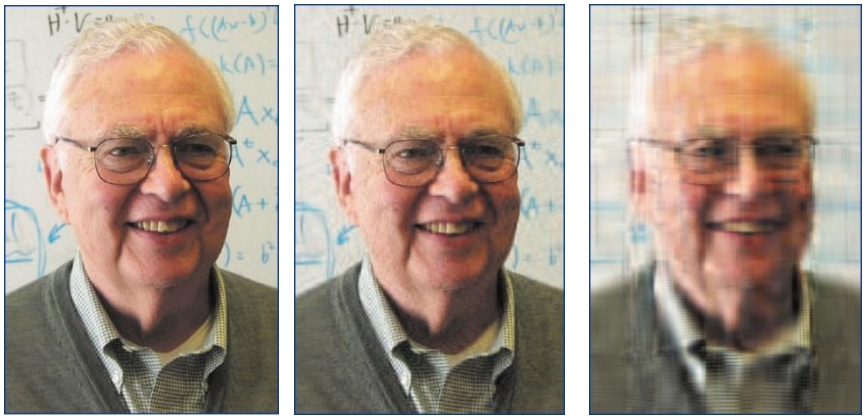
Problém mnoho dimenzií

Pohlavie - Binárne 0-muž, 1-žena, pribudla 1 dimenzia

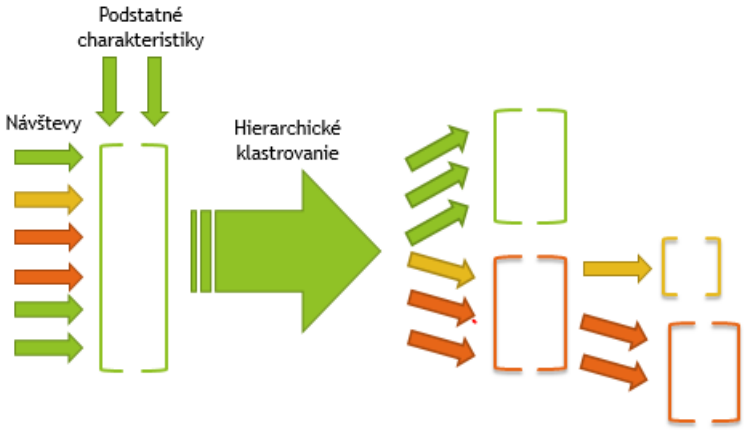
Charakteristika stránky (napr. produkt kt. sa týka) – pribudlo 1000 dimenzií

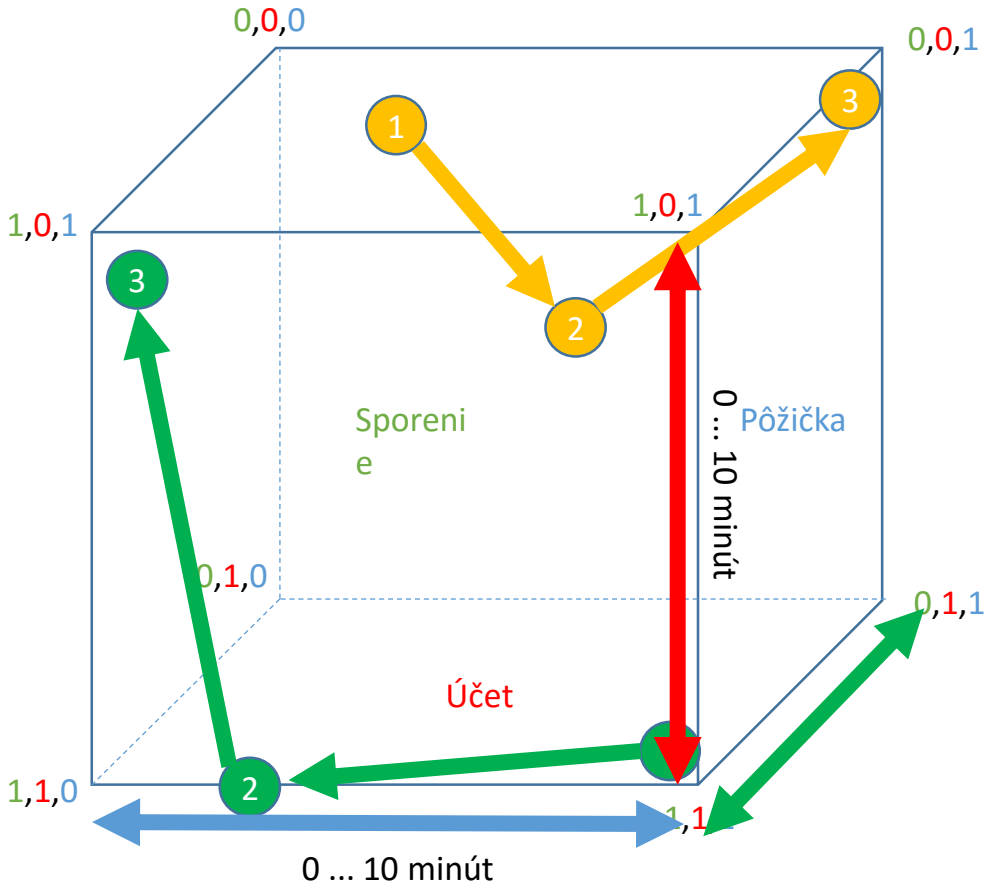
Vysvetlenie SVD

- Odstránenie korelujúcich charakteristík
- Ponechanie jedinečných hodnôt

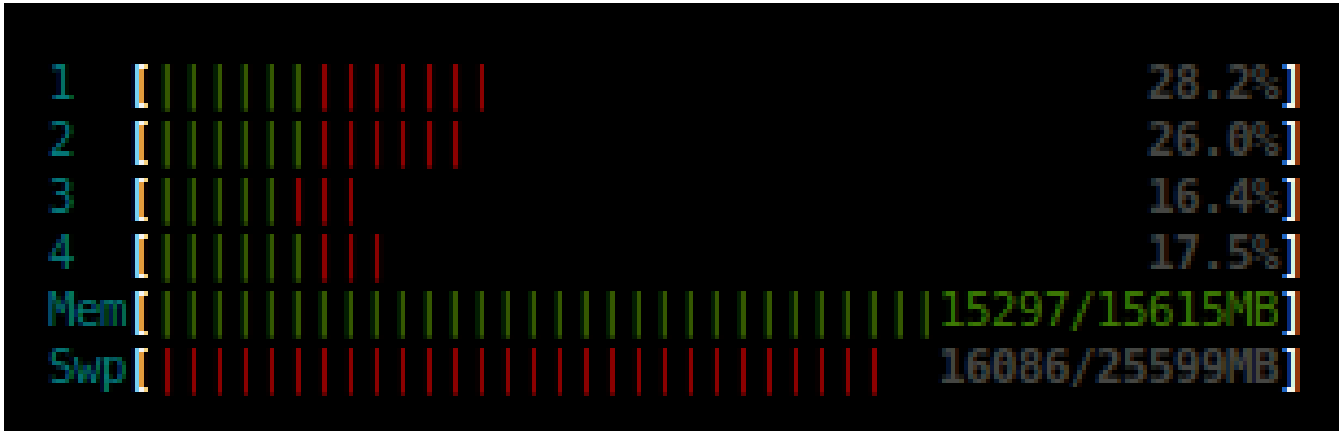


Redukcia dát (BIRCH O(N))

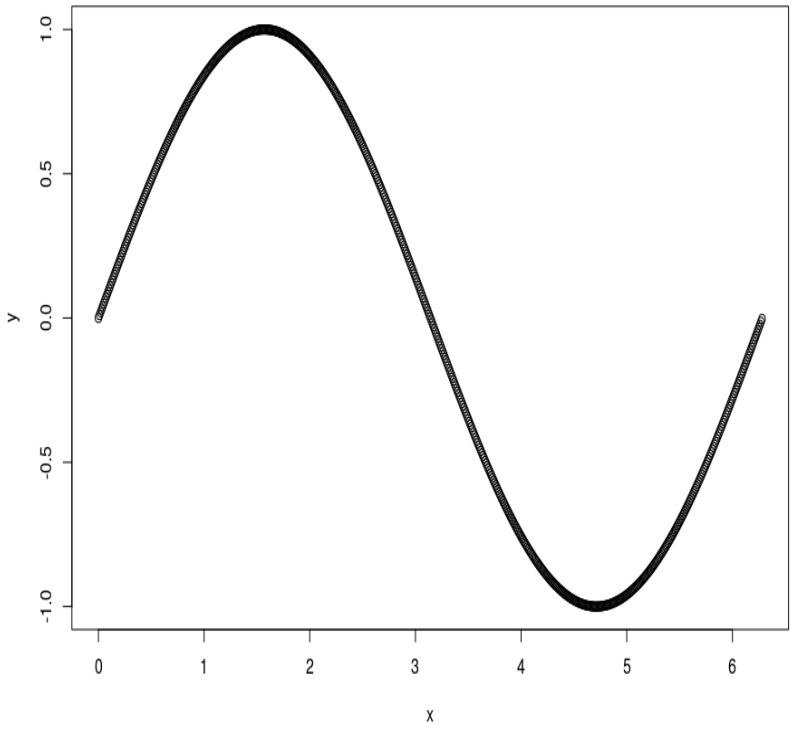


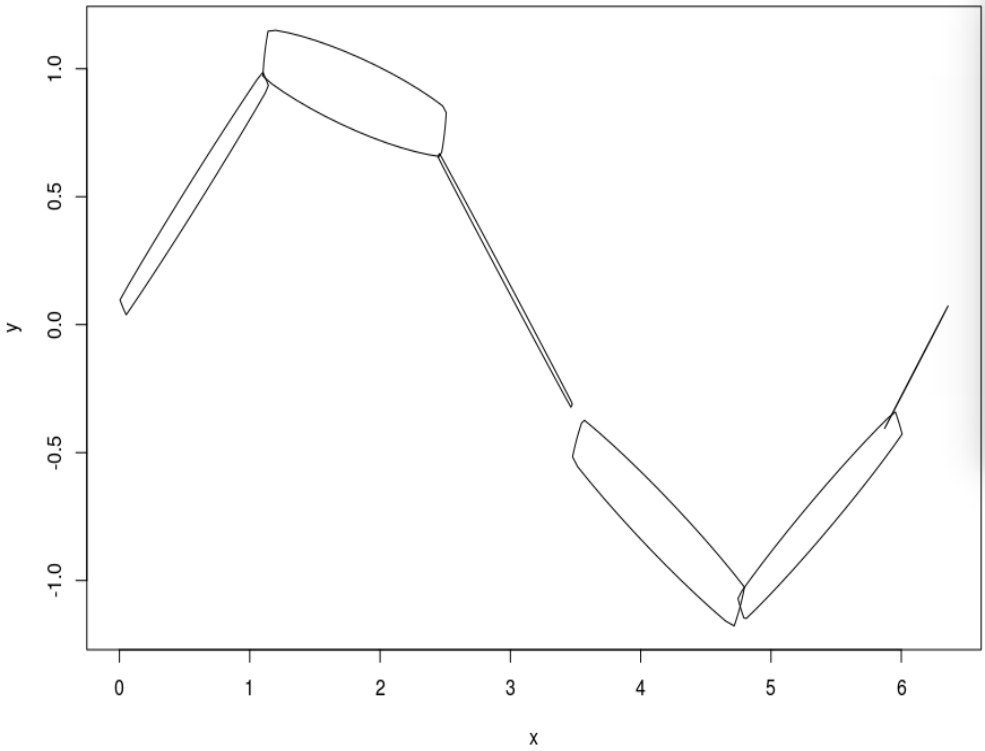


Pamäť po vygenerovaní matice 800 000 * 1000

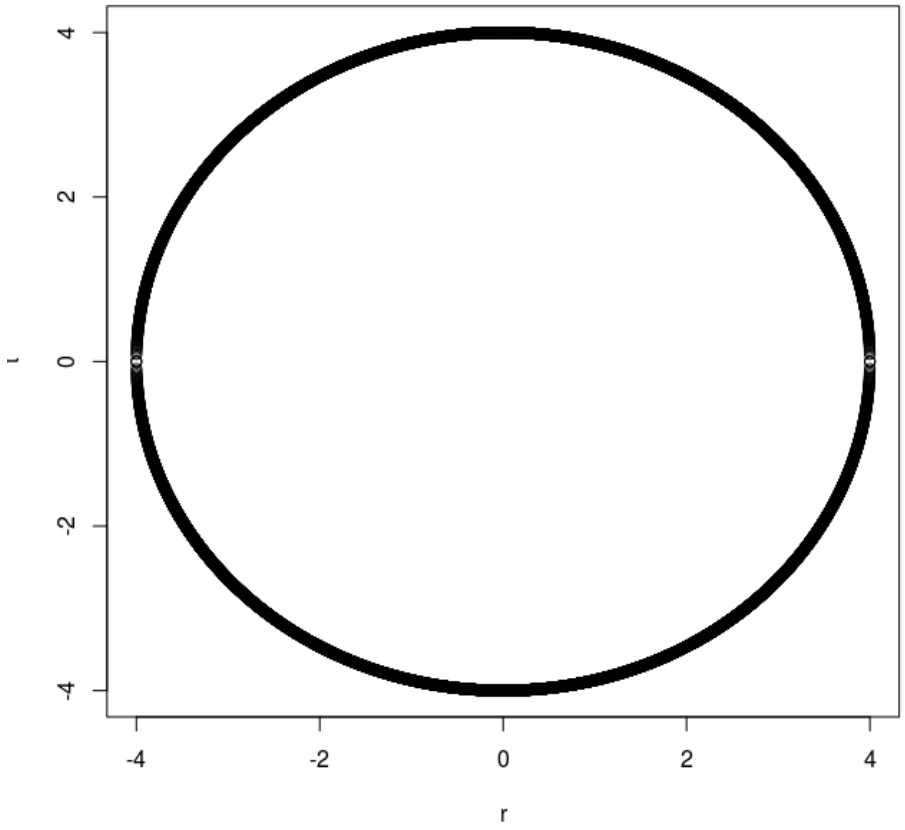


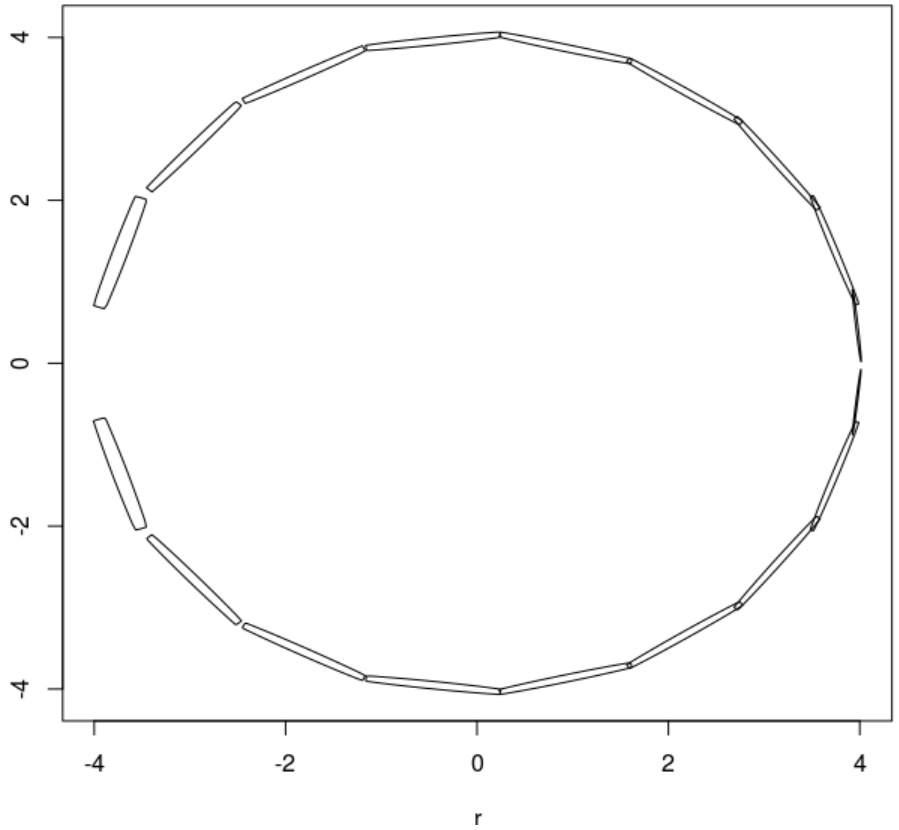
Klastrovanie BIRCH ukážky - sinusoida



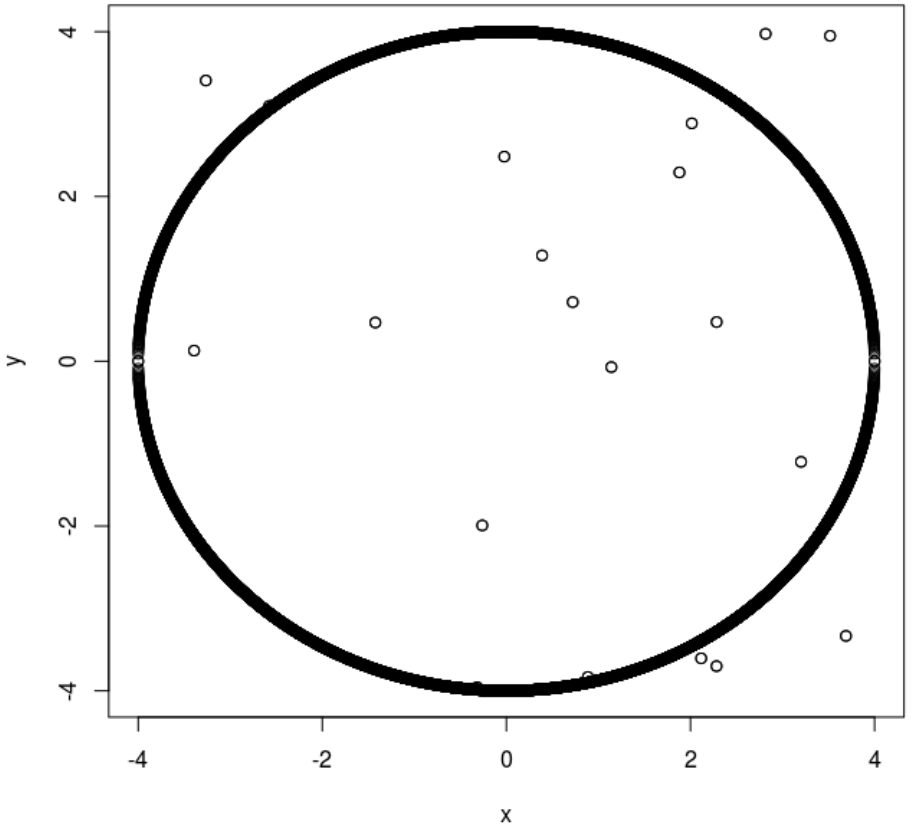


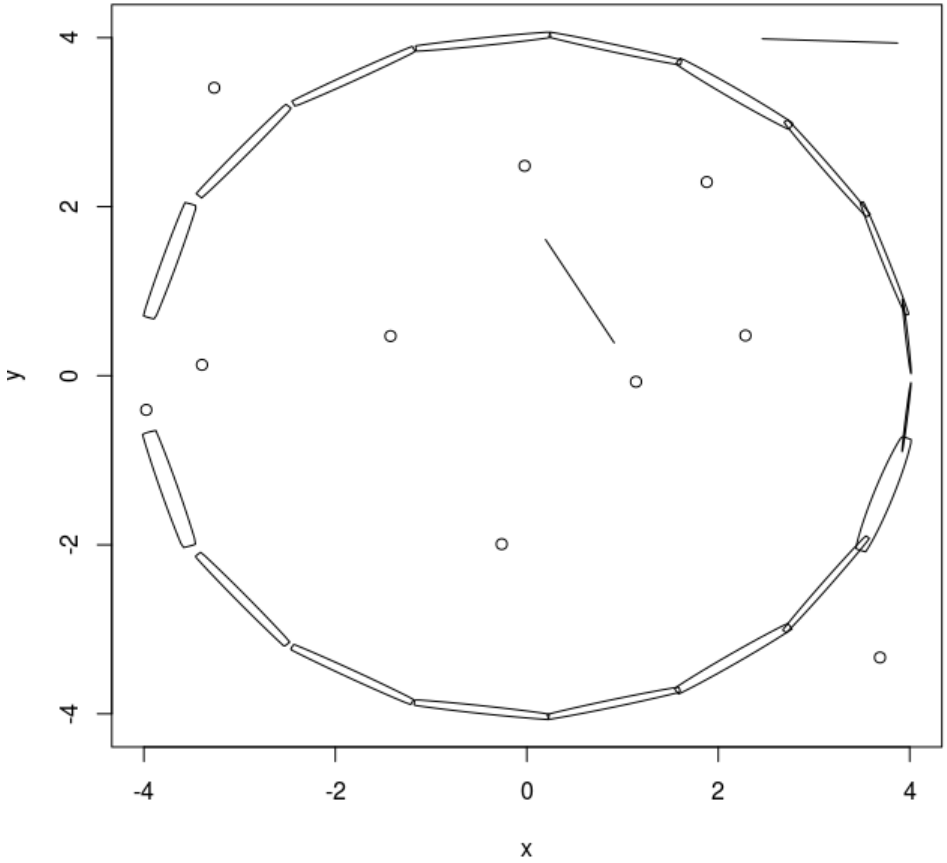
Klastrovanie BIRCH ukážky - Kruh



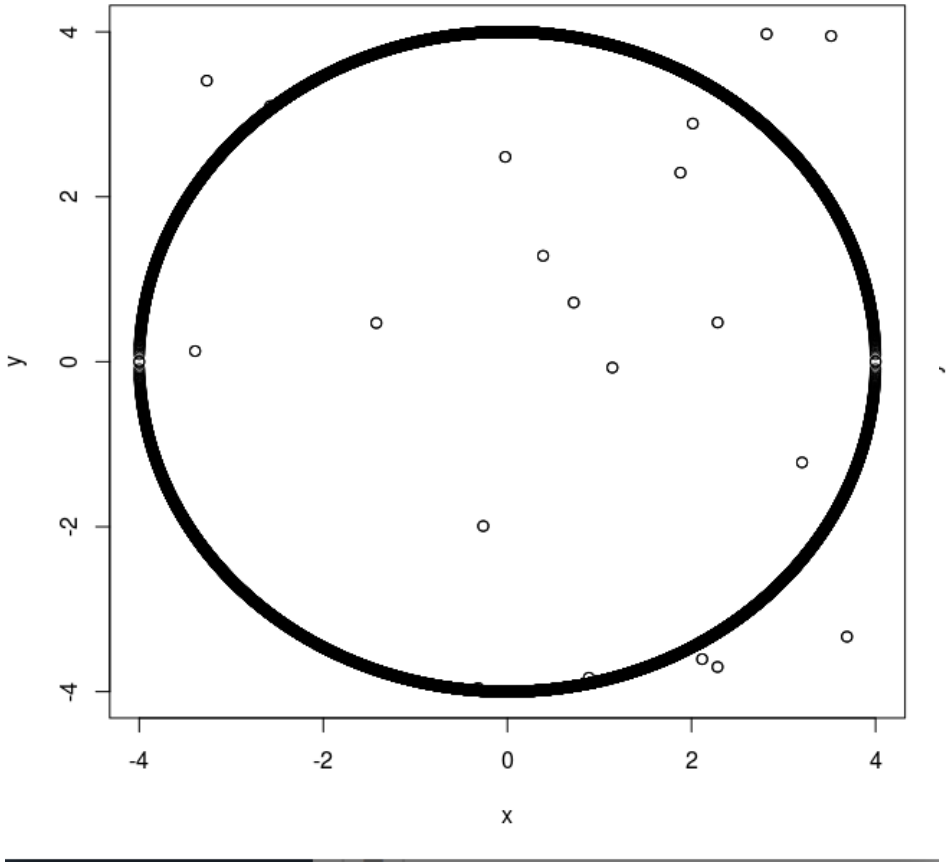


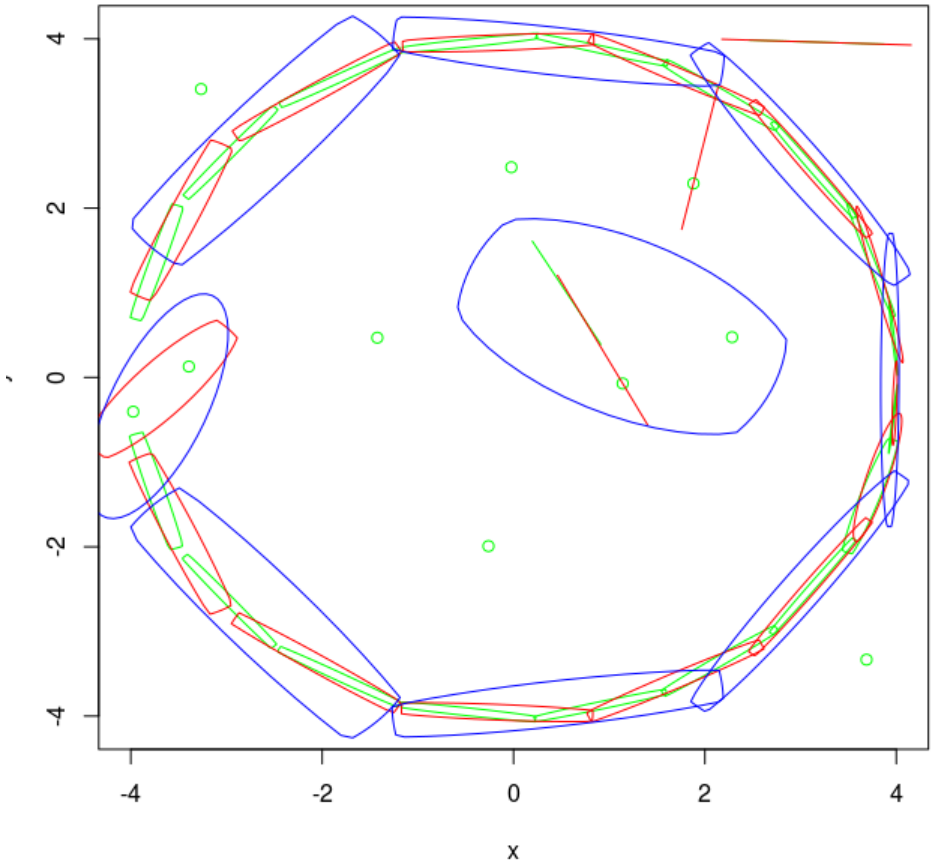
Klastrovanie BIRCH ukážky – Kruh + šum



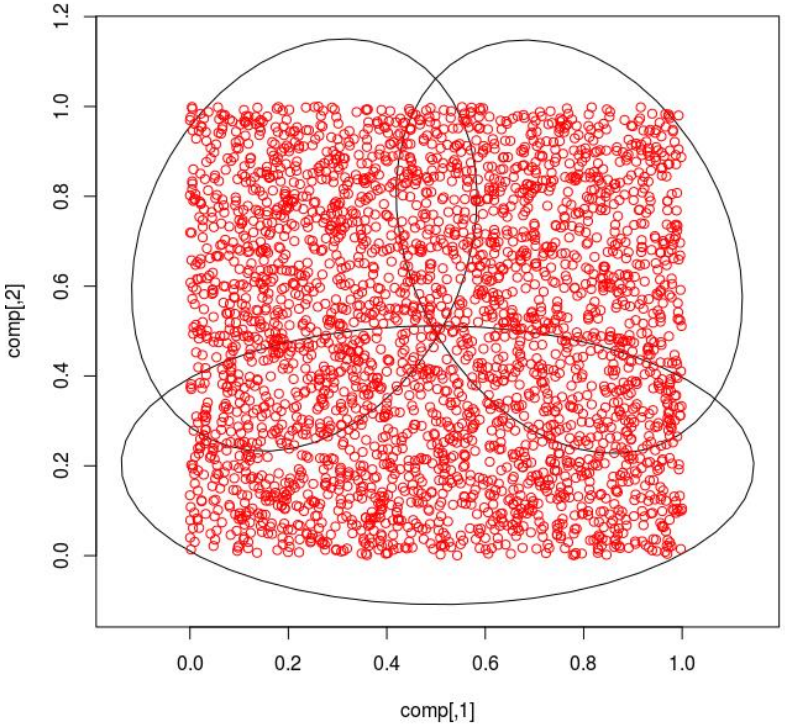


Klastrovanie BIRCH ukážky – Hierarchia



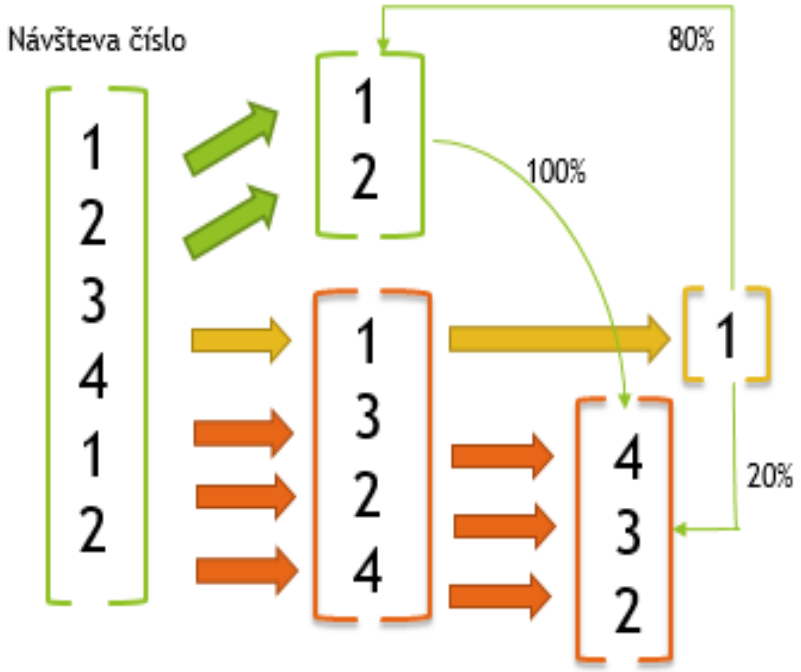


Klastrovanie BIRCH ukážky – Náhoda



Markovov skrytý model

- Neviditeľné medzistavy
- Možno stačí obyčajná markovova reťaz
- Skúsenosti?



Alternatívne postupy

- Učenie s učiteľom a nekonečná optimalizácia, pretrénovanie vs. Podtrénovanie – zameranie sa vždy len na jeden cieľ
- Page rank na úrovni stránky – ako ktorá stránka prispela ku konverzii
- Visitor rank na úrovni interakcii so systémom

Vyhodnotenie

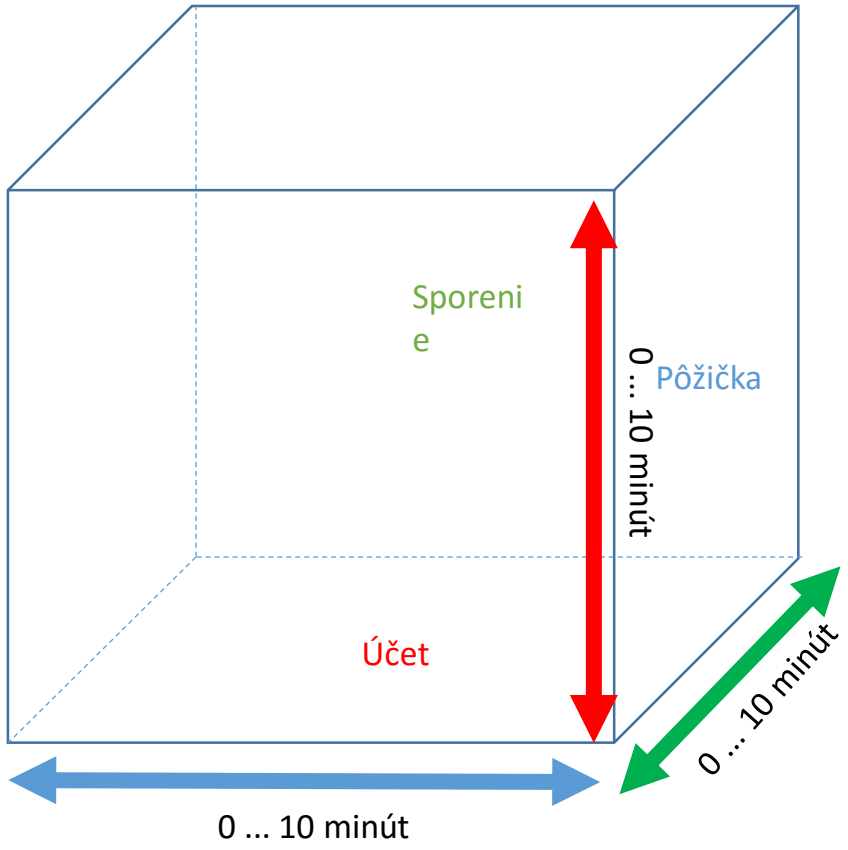
- Predpovedanie navštívených stránok v ďalšej návšteve => z toho aj akcií a charakteristík ďalšej návštevy
- Churn prediction
- Cielená zmena správania návštevníka
- Predpovedanie dopadu zmien v systéme
- ...

Zhrnutie

- Hypotéza – správanie závisí od aktuálneho stavu a predchádzajúceho správania sa
- Identifikácia podstatných charakteristík
- Umiestnenie návštev v priestore podľa ich charakteristík
- Redukcia na zhluky návštev s podobnými charakteristikami
- Časová následnosť návštev a predikcia ďalších krokov
- Overenie a vyhodnotenie

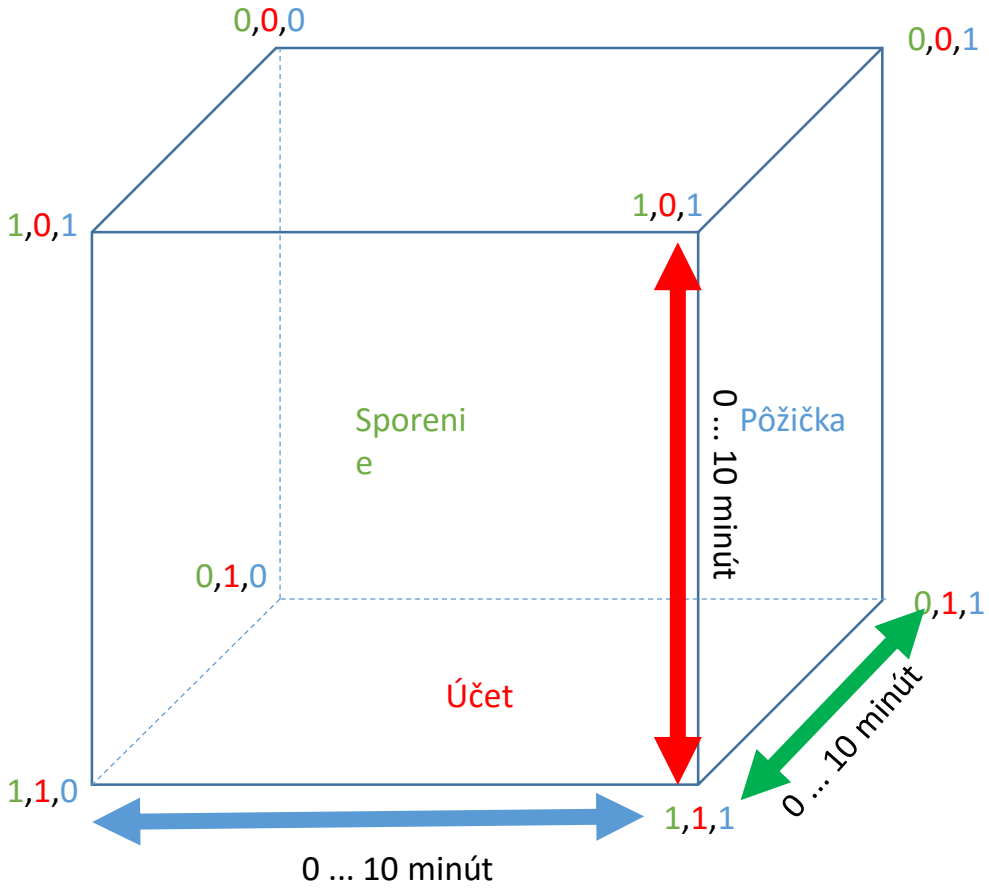
Umiestnenie návštev do priestoru

- Predstavme si 3 stránky
- Sporenie, Účet a Pôžička



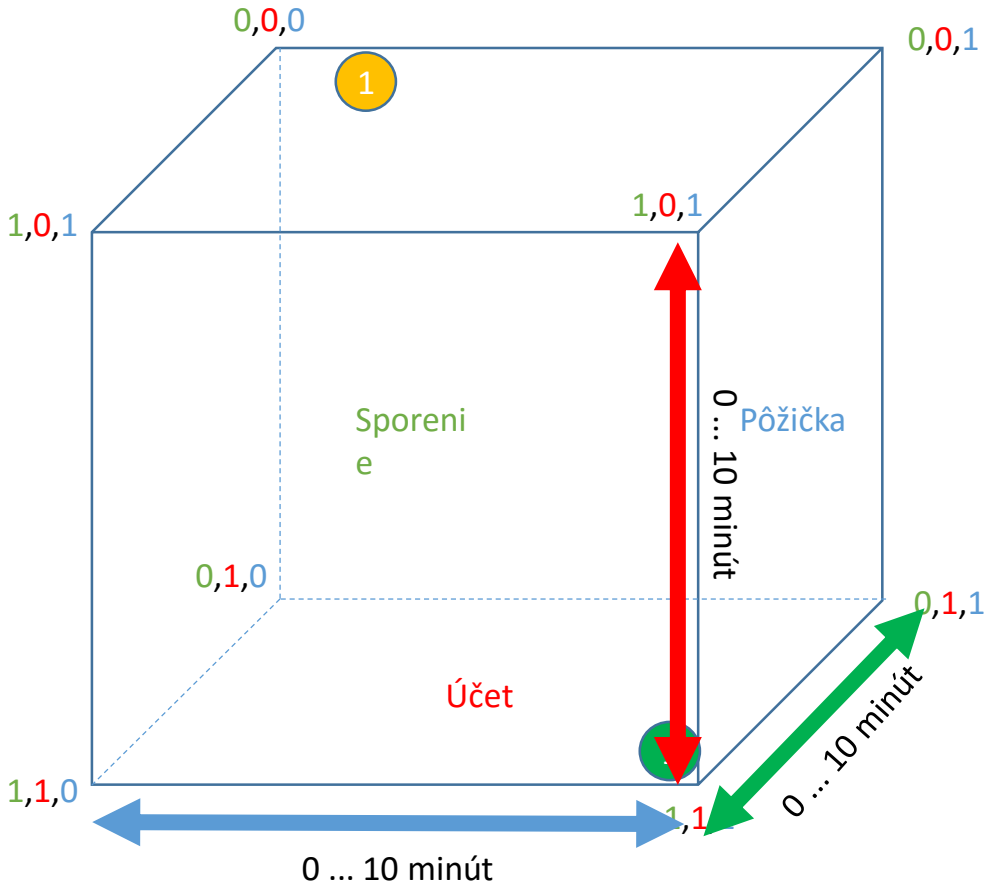
Umiestnenie návštev do priestoru

- Predstavme si 3 stránky
- Sporenie, Účet a Pôžička
- Každý na stránke strávil pri prvej návšteve istý čas (1 = 10min, 0 = nenavštívil stránku)



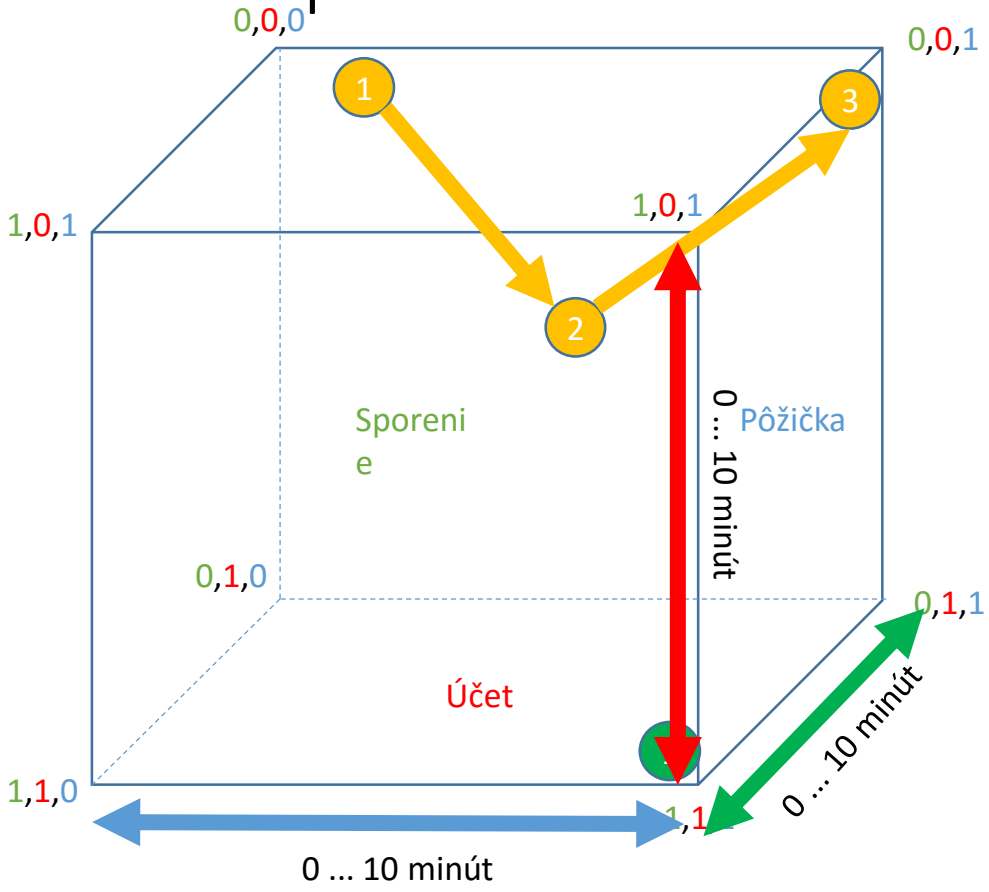
Umiestnenie návštev do priestoru

- 0 = nenavštívil stránku, 1= strávil 10 min. na stránke
- Žltý pri prvej návšteve strávil na stránke pôžičky jednu minútu, na stránke účtu 0 minút, na stránke sporenia 10 minút
- Zelený navštívil všetky 3 stránky a všade strávil 10 minút



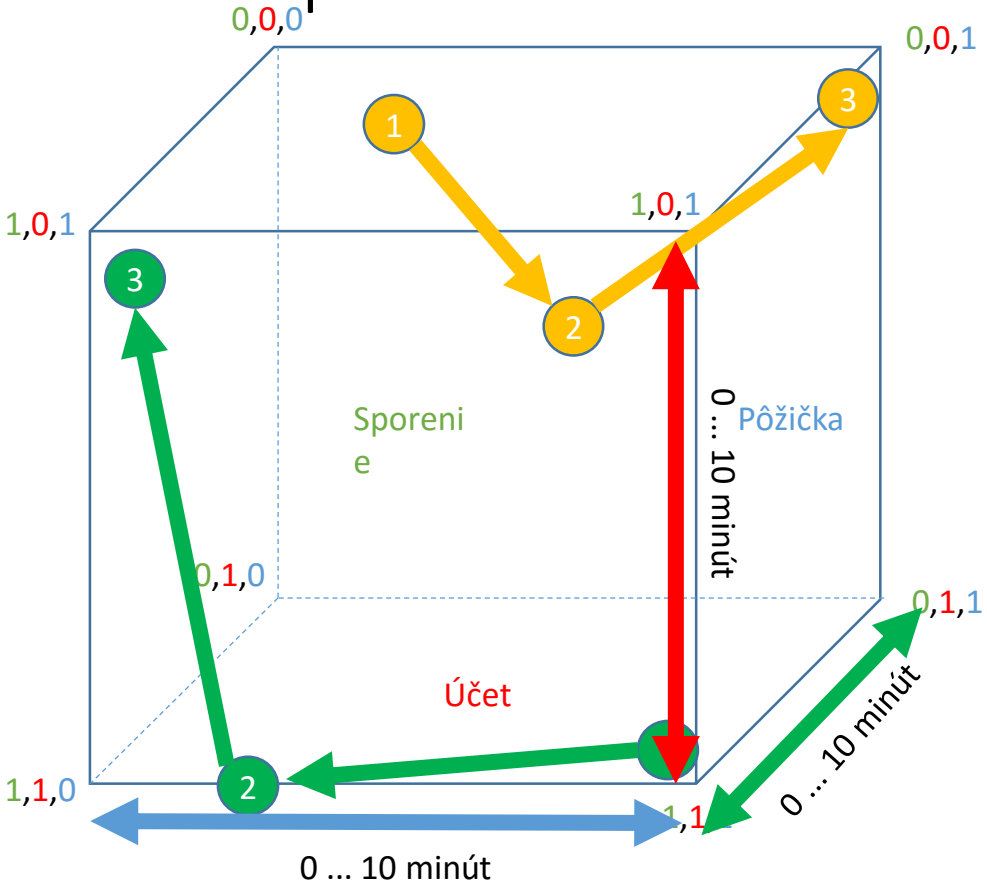
Umiestnenie návštev do priestoru

- 0 = nenavštívil stránku, 1= strávil 10 min. na stránke
- **Žltý pri druhej návšteve strávil na stránke sporenia šesť minút**
- **Pri tretej návšteve sporenie 10 minút**
- Zelený ešte neprišiel druhý krát



Umiestnenie návštev do priestoru

- 0 = nenavštívil stránku, 1= strávil 10 min. na stránke
- Zelený pri druhej návšteve strávil na účte a sporení 10 minút a na pôžičke 8 minút



Uvažujme že ich máme v priestore a vieme ako sa pohybujú medzi skupinami

- Niekoľko skupín bude takých, že väčšina návštev v nich dokončila žiadosť o pôžičku
- Ak sa pozrieme na skupinu ktorá viedla ku nim (predchádzajúce návštevy), zistíme, že 60% dokončilo žiadosť, 20% sa už nikdy nevrátilo a 20% žiadosť nedokončilo a putovalo do inej skupiny v ďalšej návšteve.
- Vieme optimalizovať, pozrieť sa na tých ktorí mali najväčší potenciál dokončiť žiadosť a navrhnúť čo s tým ďalej urobíme.