Kombinácia automatický a vizuálnych metód dolovania dát

Tomáš Poláček

školiteľ: Mgr. Matej Novotný

Zadanie a cieľ práce

- Kombinácia automatických a vizuálnych metód dolovania dát
- Opísať klady a zápory oboch prístupov, navrhnúť spôsob ako prístupy skombinovať čo najlepšie a ilustrovať túto kombináciu na zvolenej implementácii

Úvod do dolovania dát

- Veľké množstvo nazhromaždených dát, ktoré treba analyzovať a spracovať
- Netriviálne získavanie implicitných, pred tým neznámych, a potencionálne užitočných informácii (Petr Berka, 2003)
- Metódy dolovania: automatické, vizuálne a kombinované

Automatické metódy

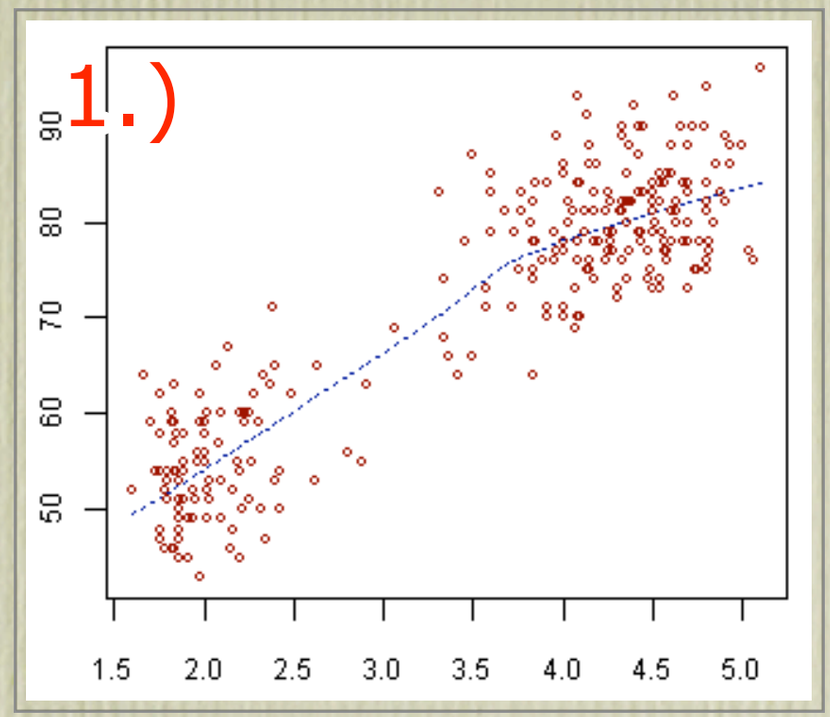
- Patria sem matematické metódy, štatistické metódy a metódy umelej inteligencie
- Väčšinou nie je možné ovplyvniť ich priebeh
- Výsledky sú ťažšie interpretovateľné analytikom a ťažšie prezentovateľné inej osobe
- Napríklad: K-Means, PCA, lineárna regresia ...

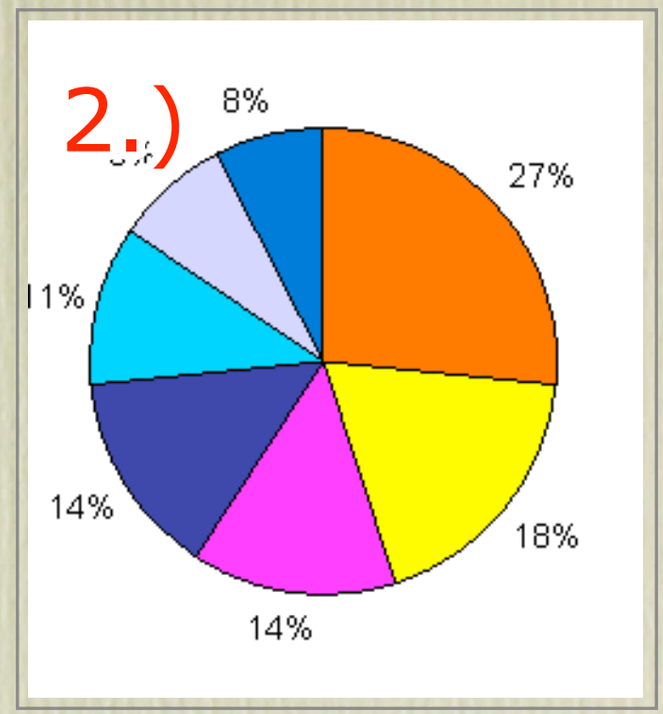
Vizuálne metódy

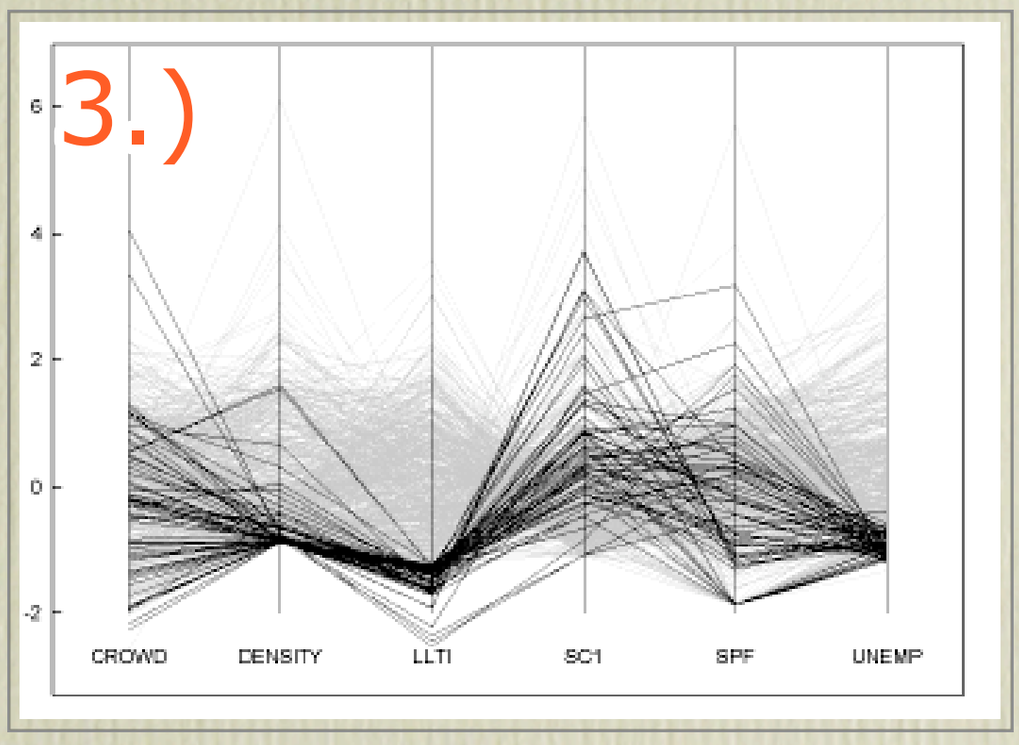
- Hľadanie skrytej informácie v dátach pomocou zraku
- Využitie trénovaného ľudského oka a jeho schopnosti prijímať obrovské množstvo informácií
- Vo veľkej miere sa využívajú skúsenosti analytika
- Vyžadujú viac interakcie od používateľa ako automatické metódy

Vizuálne metódy - príklady

- 1.) Scatter plot
- 2.) Koláčový graf
- 3.) Paralelné súradnice







Kombinované metódy

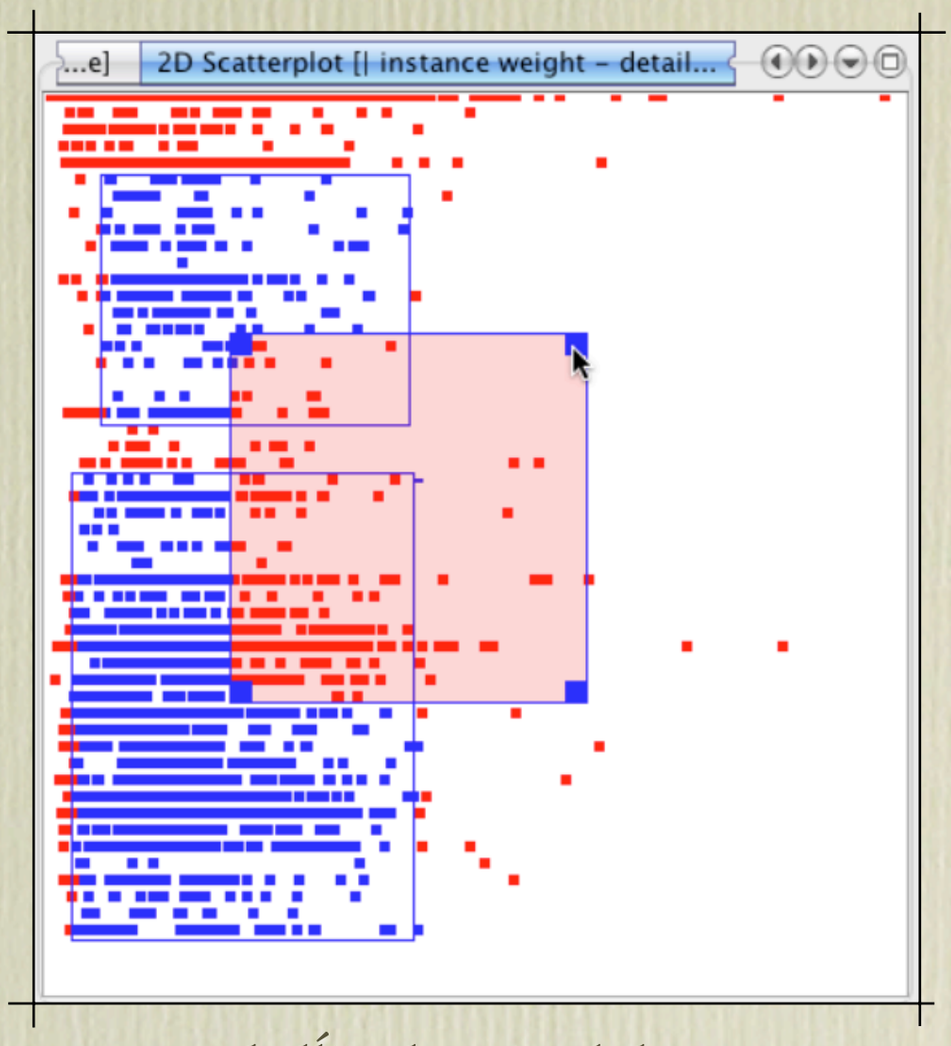
- Rôzne spôsoby kombinácie:
  - **vyvážená kombinácia** - obe metódy zastúpené rovnako, vzájomne sa dopĺňajú
  - **nevyvážená kombinácia** - jedna metóda je použitá ako základ a druhá ju doplňuje
- Očakávajú sa pozitívne výsledky

Ciele práce

- Rozšírenie a vylepšenie K-Means algoritmu pomocou kombinácie s vizuálnou metódou Scatterplot
- Zachytiť analytický proces, ktorý vzniká pri segmentácii dát, vizualizovať ho a umožniť jeho opätovné použitie
- Implementovať navrhnuté spôsoby kombinácie do aplikácie
- Ukázať či boli implementované vylepšenia úspešné a poskytujú lepšie výsledky

Použité analytické nástroje a metódy

- K-Means algoritmus
- Obdĺžniková selekcia
- Scatterplot vizualizácia
- Stromové a grafové vizualizácie



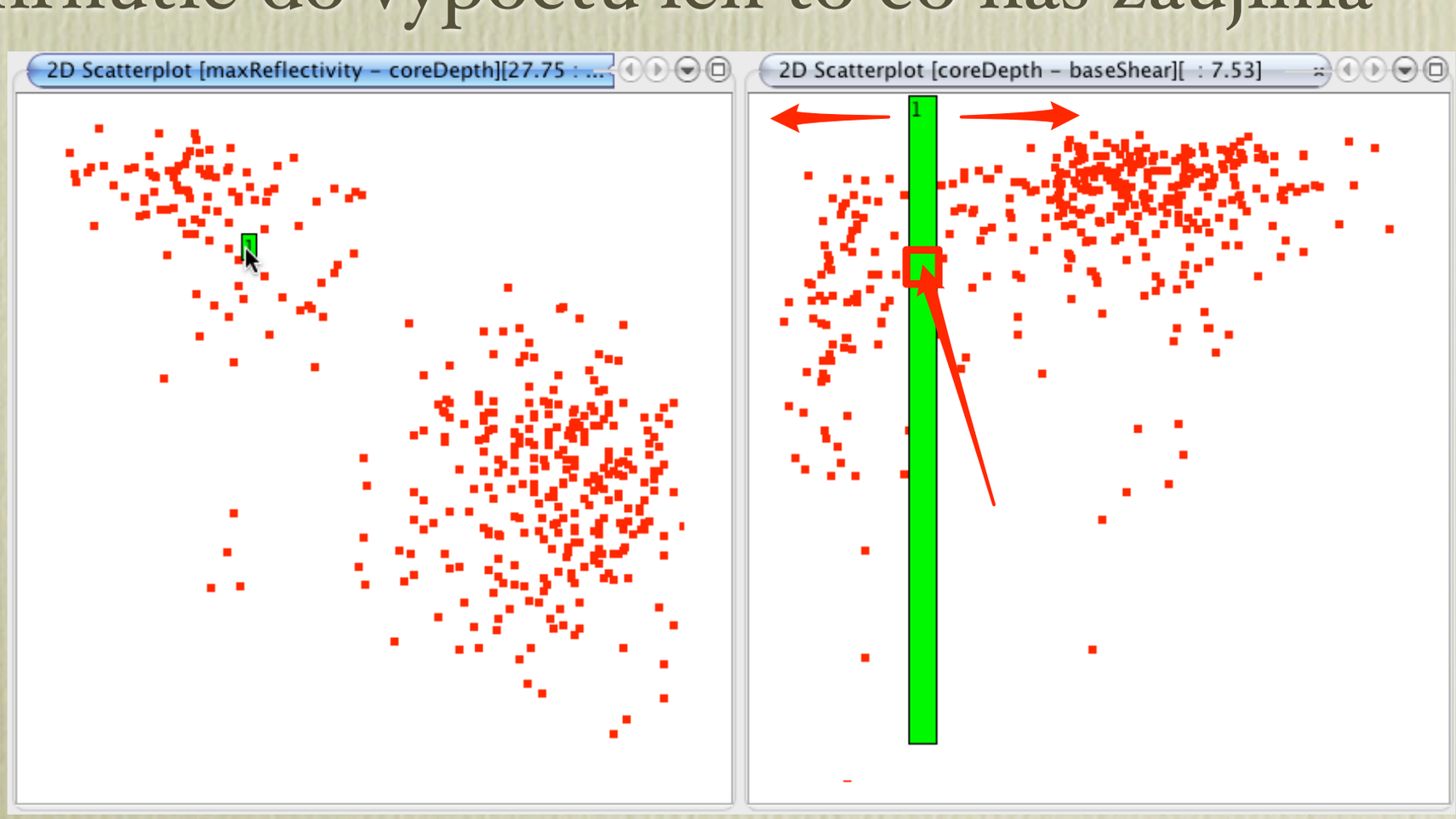
obdĺžniková selekcia

Vylepšenie K-Means

- Zadávanie počiatočných stredov zhlukov pre algoritmus
- Upravenie výpočtu algoritmu tak, aby počítal aj s oddeľujúcimi nadrovinami

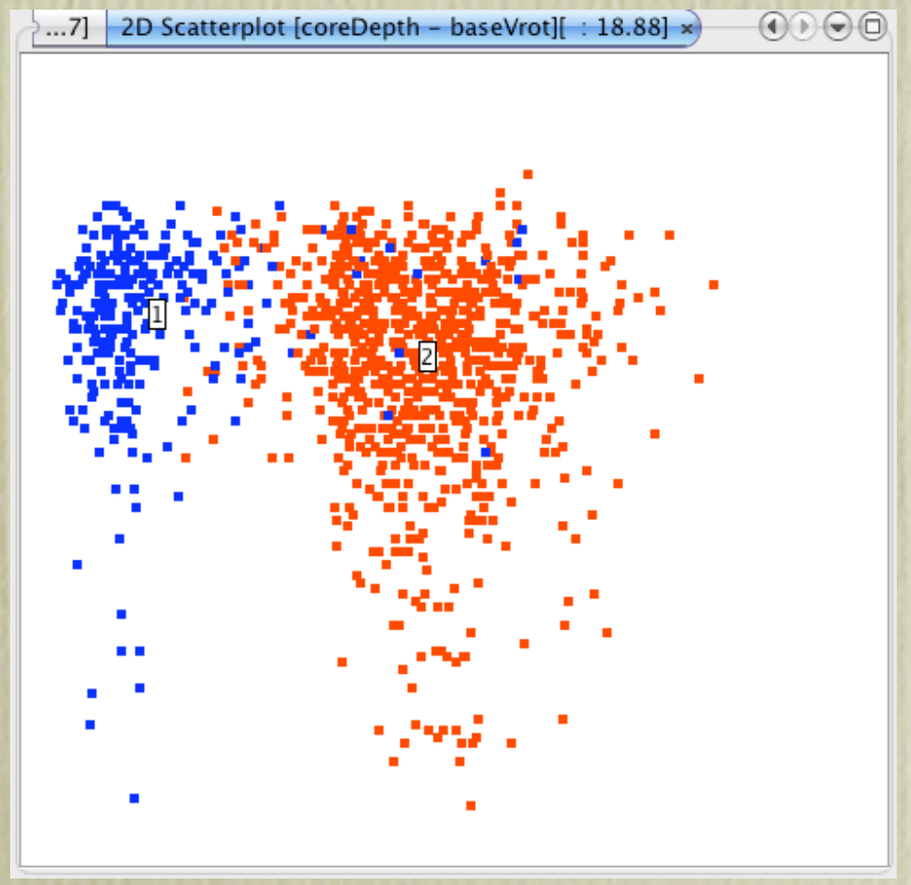
Zadávanie stredov zhlukov

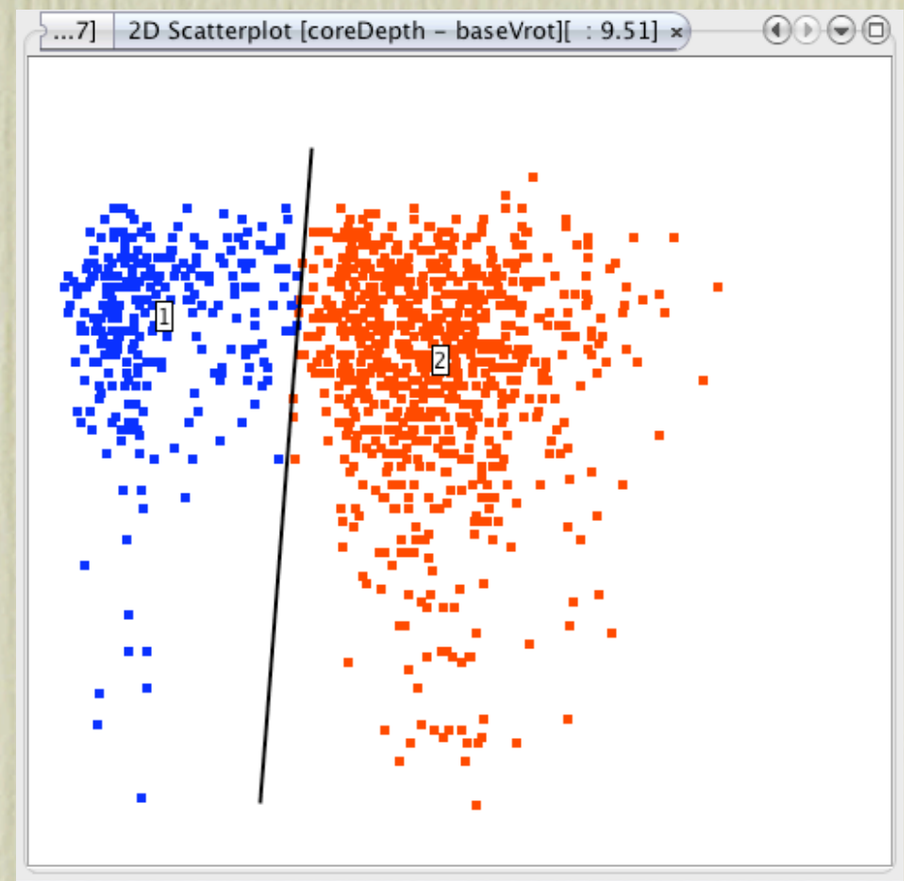
- Zadávanie stredov klikaním do grafu
- Zadávanie súradníc stredu v ďalších rozmeroch
- Odhadnutie počtu a pozície stredov zhlukov
- Zahrnutie do výpočtu len to čo nás zaujíma



Oddeľujúce nadroviny

- Zadávajú sa kreslením
- Dva významy:
  - "Odrezanie" outliers dát a pevné oddelenie zhlukov



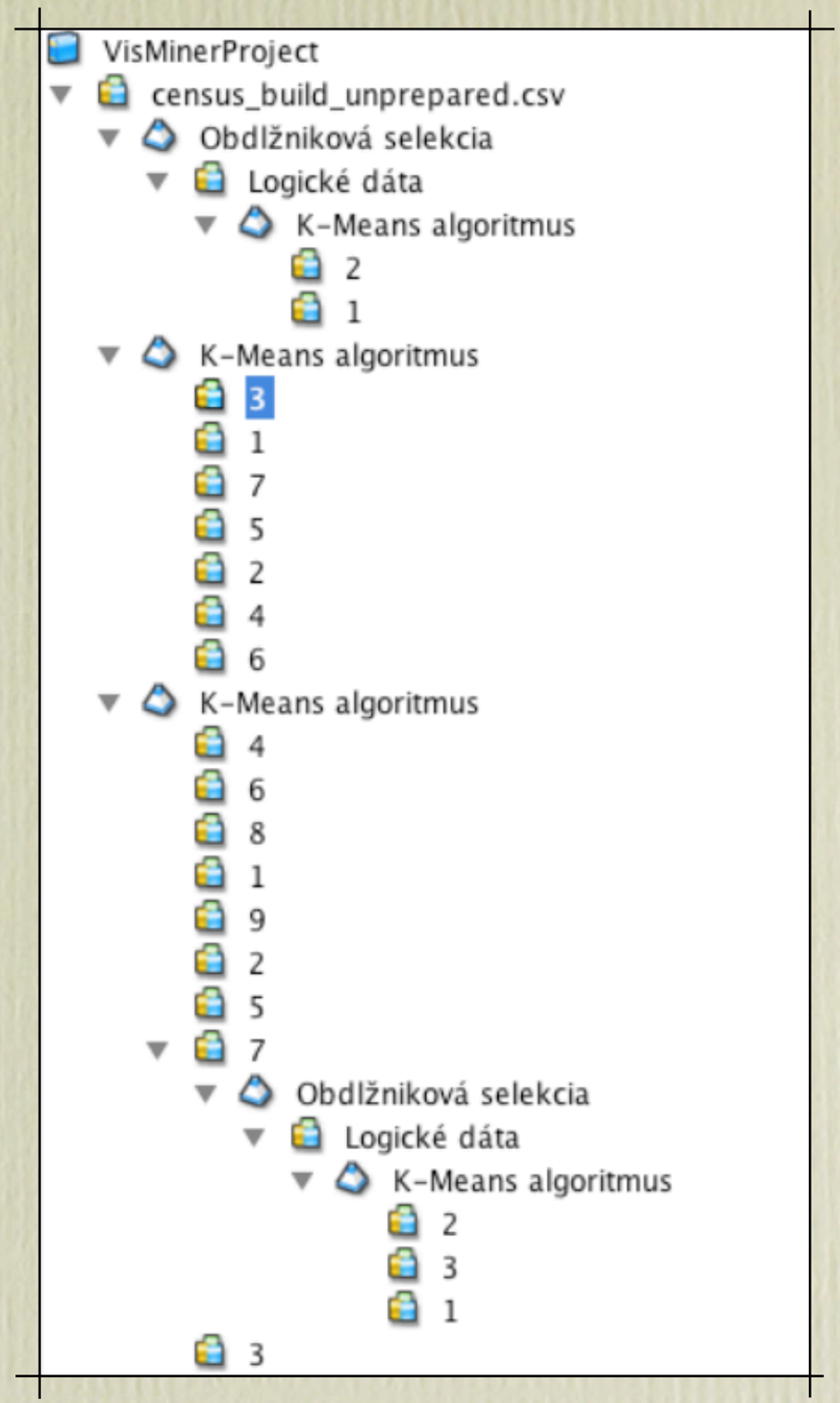


Analytické sedenie

- Záznam uvažovania analytika nad dátami
- Postupnosť krokov práce analytika s dátami
- Vysokoúrovňové programovanie
- Štruktúra sedenia:
  - stromový charakter
  - striedanie dvoch typov vrcholov: dátové vrcholy a filtrovacie vrcholy

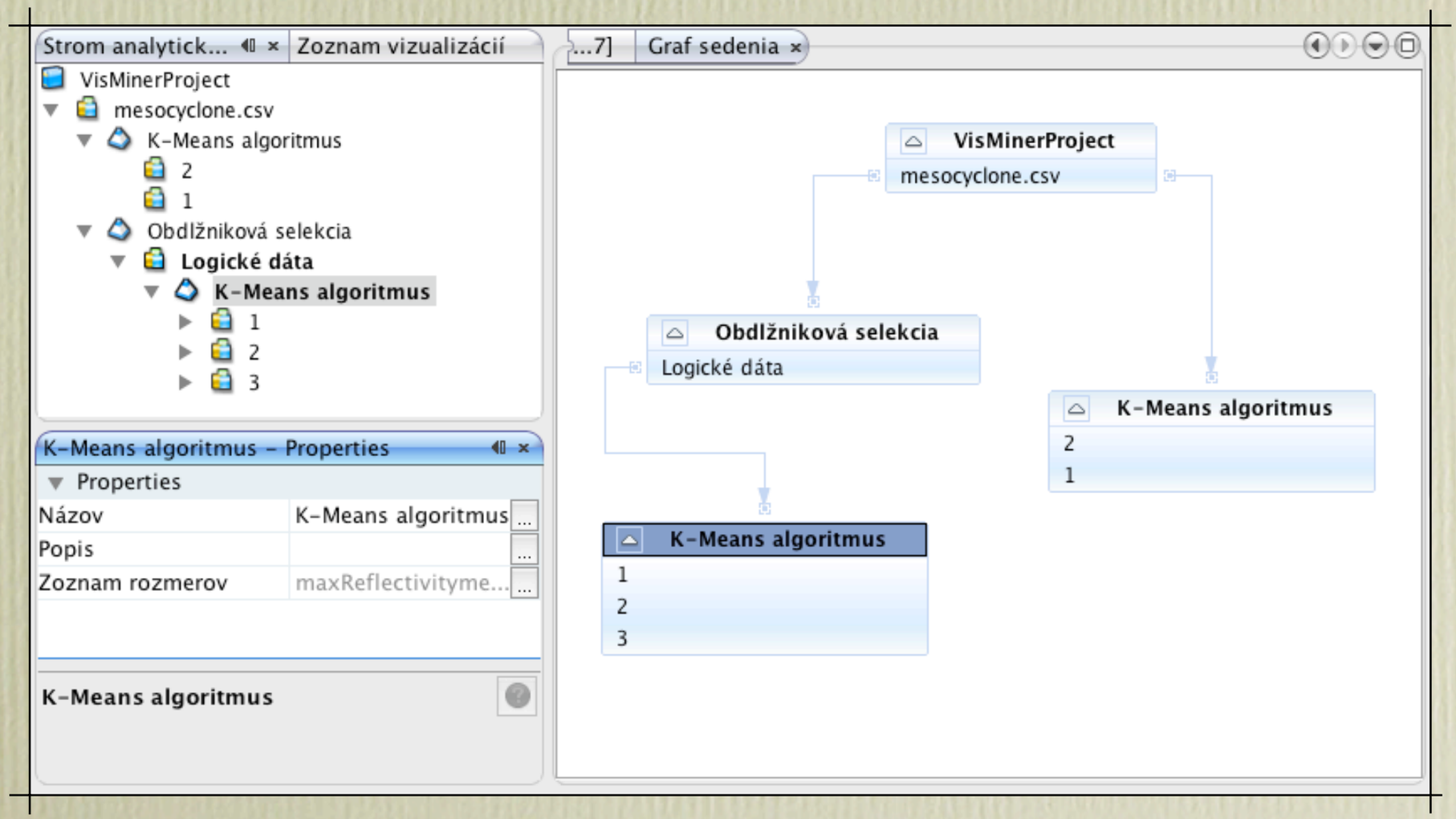
Analytické sedenie - štruktúra

- Dátové vrcholy:
  - Fyzické dáta
  - Logické dáta
- Filtrovacie vrcholy:
  - obdĺžniková selekcia
  - K-Means algoritmus



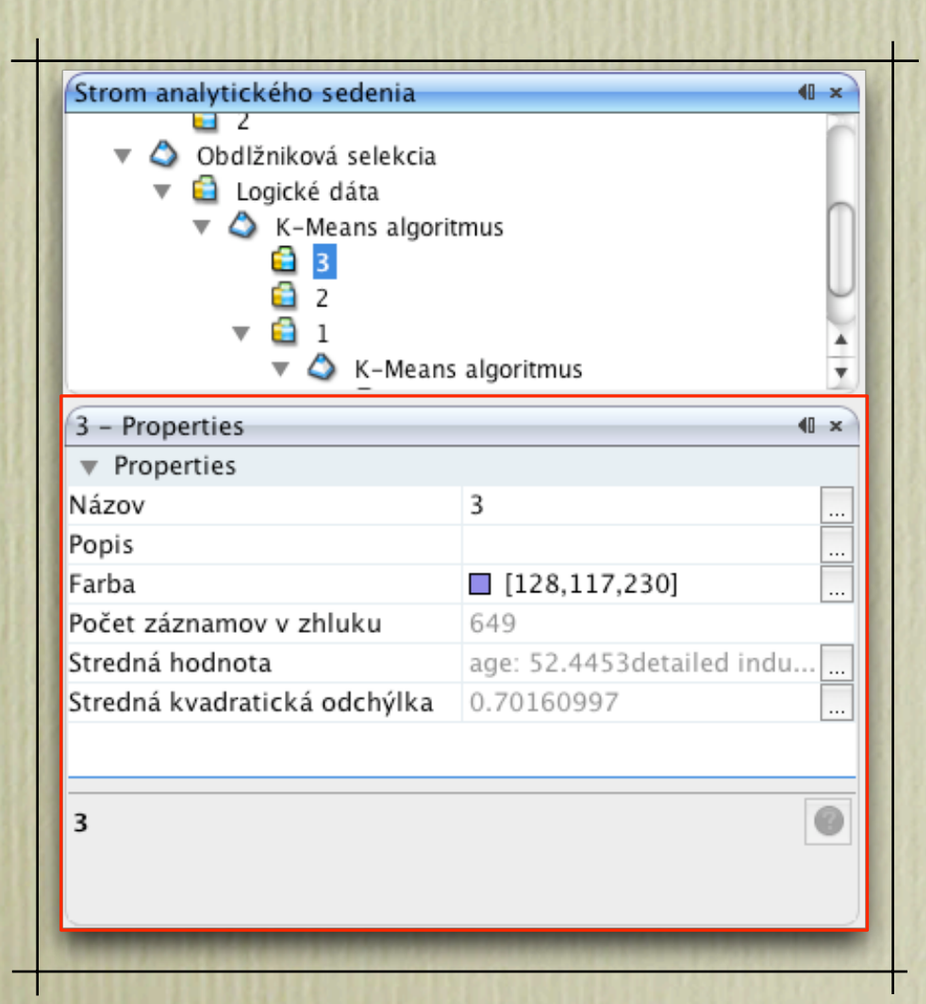
Vizualizácia sedenia

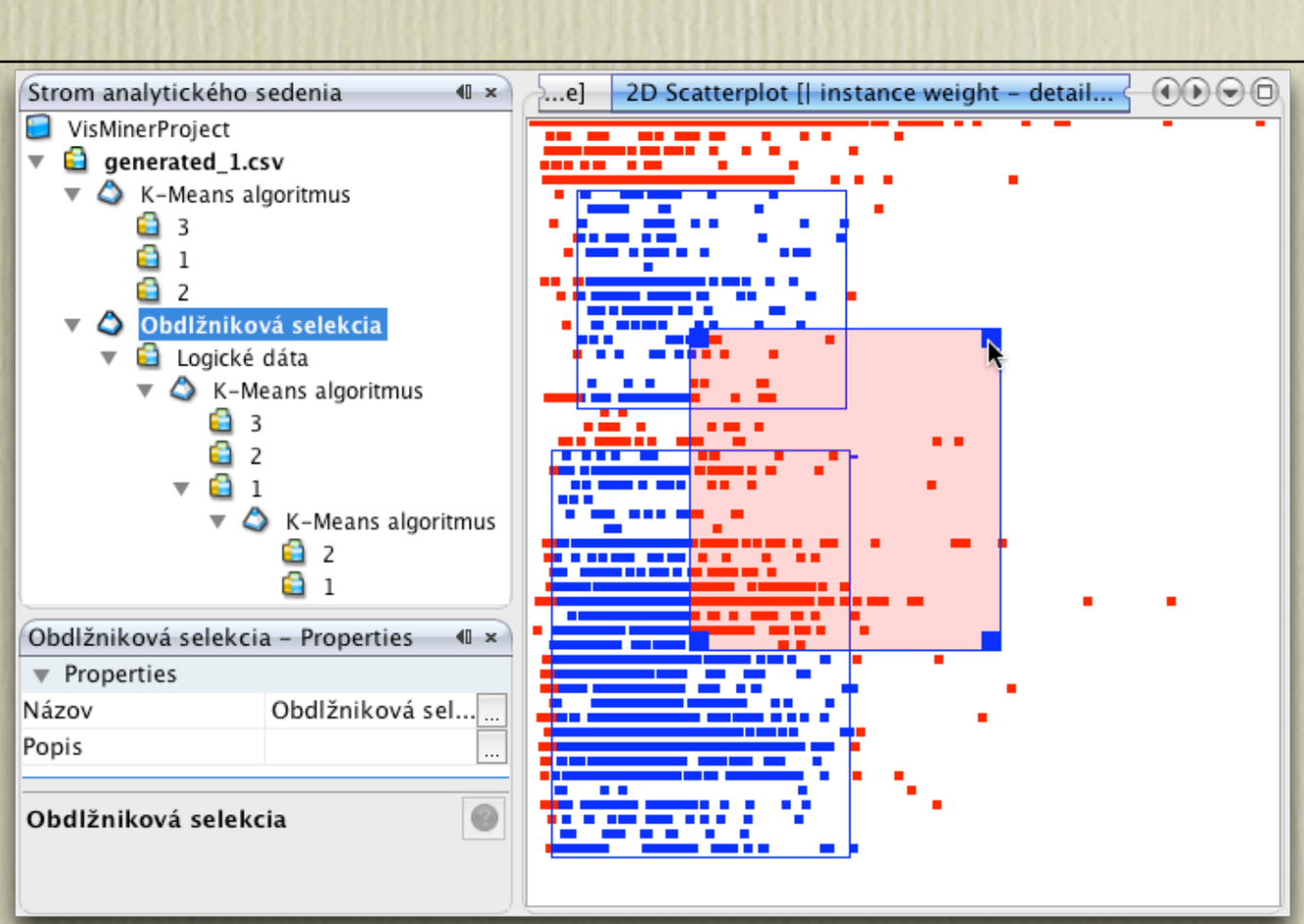
- Vizualizácia pomocou stromu a grafu
- Prezentovanie postupu ako vznikli výsledky (Visual Storytelling)



Interakcia so sedením

- Vrcholy v strome sedenia umožňujú interakciu
  - Modifikácia parametrov vrchola
  - Mazanie vrcholu zo sedenia





Opätovné použitie sedenia

- Opätovné použitie na kompatibilné dáta
- Uloženie sedenia do XML súboru

Výsledky - zhrnutie

- Vylepšenie K-Means algoritmu kombináciou s vizuálnym zadávaním počiatočných stredov zhlukov a oddeľujúcich nadrovín.
- Zachytenie vizualizácie a opätovné použitie analytického sedenia.
- Splnili sa očakávania pozitívneho prínosu kombinácie automatických metód s vizuálnymi

Ukážka aplikácie

- zadávania počiatočných stredov zhlukov
- zadávanie oddeľujúcich nadrovín
- výpočet algoritmu a vizualizácia výsledkov

VisMiner 200804211638

Súbor Pohľad Okno Nápoveda

Strom analytického...

VisMinerProject

mesocyclone.csv

K-Means algorit

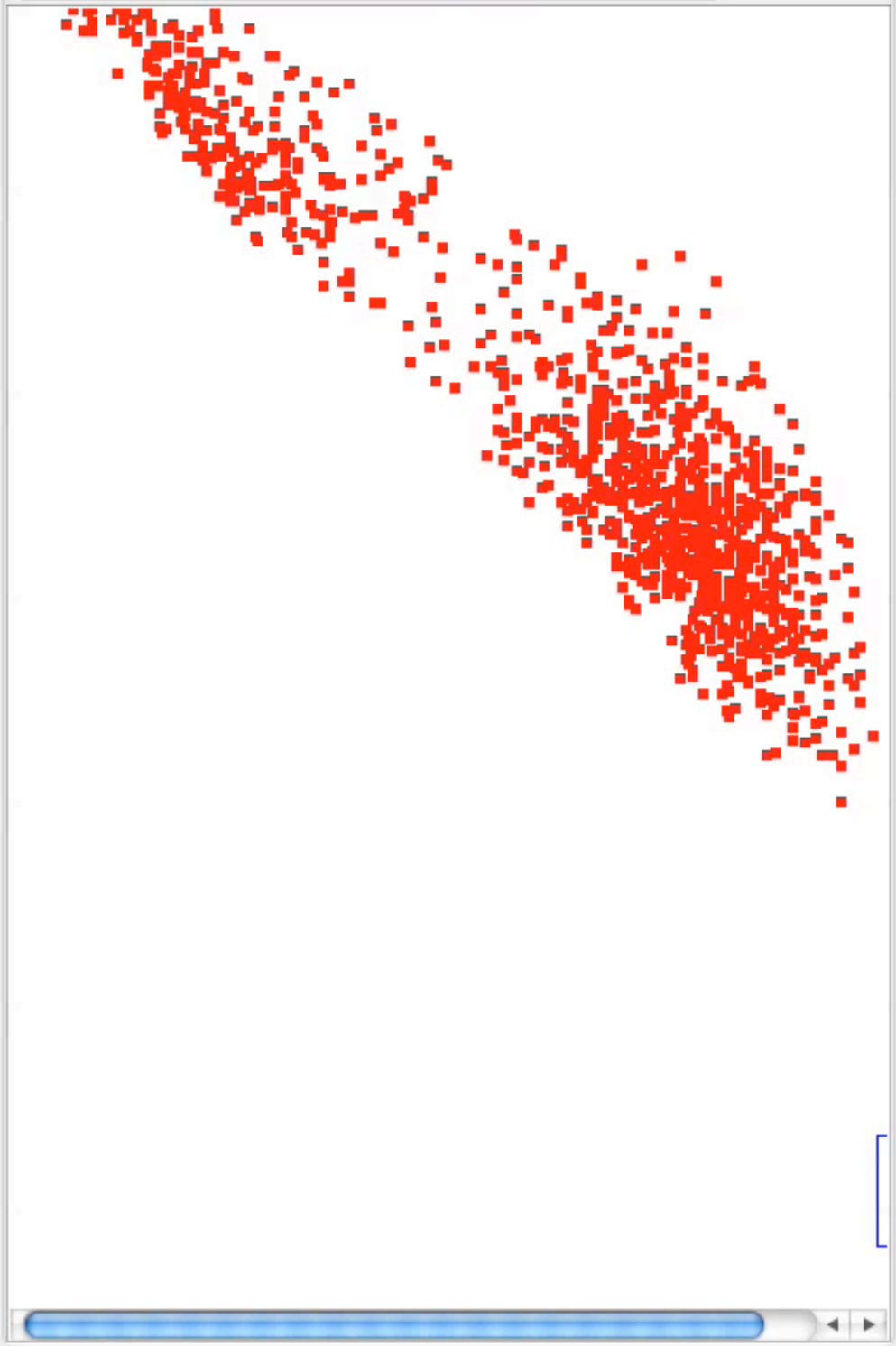
K-Means algoritm...

Properties

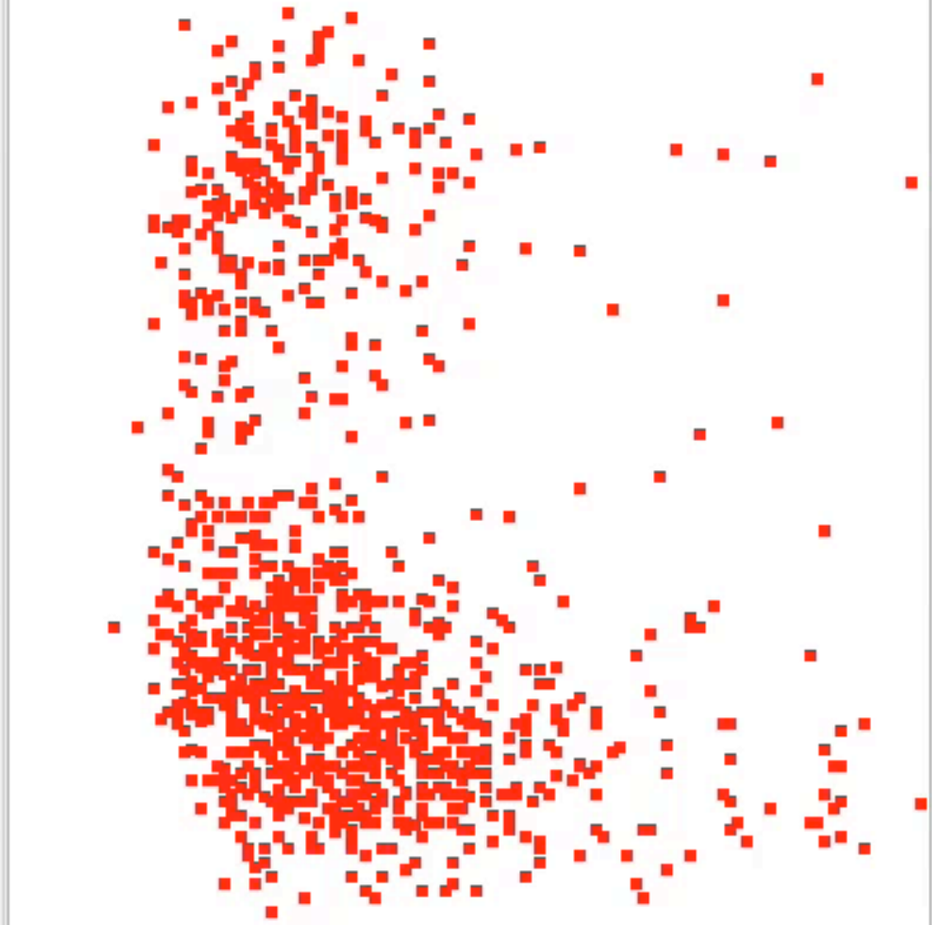
Názov	K-Mean...
Popis	
Počet kôl alg	2

K-Means algoritmus

...ov | 2D Scatterplot [maxReflectivity - meanRe...



2D Scatterplot [baseVrot - maxReflectivity][8...



Selekcia dát

Zoznam dostupných algoritmov

Zoznam vizualizácií

Ďakujem za pozornosť

Otázky oponenta

- **otázka:** Ako by sa zmenila funkčnosť algoritmu, keď by algoritmus K-means počítal zhluky vo všetkých rozmeroch, nielen v tých, ktoré majú zadané stredy?

- **odpoveď:**

  - dopočítanie chýbajúcich súradníc stredov pomocou strednej hodnoty

  - obmedzenie K-means bolo zámerné: zahrnutie do výpočtu len to, čo používateľ chce

- **otázka:** Vysvetliť význam obdĺžnikovej selekcie, ktorú som v kapitole 5 nevyužil pri dokumentácii vytvoreného postupu.

- **odpoveď:**

  - poskladanie selekcie dát z rôznych oblastí

  - vytvorenie logických dát

  - zahrnutie len tých dát do výpočtu, o ktoré ma analytik záujem

- **otázka:** Ako sa vybrali vizualizácie zvolené pre demonštráciu vytvoreného postupu?

- **odpoveď:**

    - so zámerom čo najlepšie demonštrovať funkcie aplikácie

- **otázka:** Porovnanie kvality výsledných zhlukov pri náhodnom zadaní počiatočných stredov a pri použití ručne zadaných stredov.

- **odpoveď:**

  - kvalita zhlukov rovnaká

  - počet kôl výpočtu K-means sa výrazne skrátil (pri ukážkových dátach a dvoch zhlukoch viac než o 70%)