



TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií ■

Shlukování 2

František Kynych
31. 10. 2024 | MVD





Část I.: Rozšíření K-Means algoritmu

K-Means inicializace

- Standardně v prvním kroku vybráno K náhodných centroidů
 - Algoritmus je potřeba provést několikrát
- Další možnosti
 - K-Means++ (2007)
 - První centroid je vybrán náhodně
 - Další centroid je nejdále od prvního centroidu, ...
 - Využití řazení
 - Podle nějaké metriky seřadíme body
 - Vzdálenost od centra, hustota, ...
 - Na základě seřazení vybereme centroidy
 - Prvních K bodů (+ zákaz bodů bližších než ϵ k již existujícím centroidům)
 - Každý bod na indexu $\frac{N}{K}$

K-Medoids clustering

- Průměrný bod je nahrazen za medoid
- Medoid
 - Centrální bod v clusteru
 - Vždy je reprezentován reálným objektem z dat
- Nejpoužívanější K-Medoids metoda – Partitioning Around Medoids (PAM, vznik 1990)
 1. Vybereme K medoidů
 2. Jednotlivé body přiřadíme do clusteru s nejbližšími medoidy
 3. Prohledáváme body v každém clusteru
 - Pokud nějaký bod v clusteru snižuje průměrnou vzdálenost, tak se stane medoidem
 4. Pokud se změnil alespoň jeden medoid, tak dále iterujeme od bodu 2.

K-Medoids clustering vylepšení

- Výpočetní náročnost PAM - $O(K(n - K)2)$
 - Problém s velkým množstvím dat
- **CLARA (1990)**
 - Řeší výpočetní náročnost PAM
 - Neprohledáváme všechny body v clusteru, ale pouze S bodů
 - $O(KS^2 + K(n - K))$
 - Špatné výsledky pokud je jeden nebo více počátečních medoidů daleko od nejlepších možných medoidů
- **CLARANS (1994)**
 - Vylepšené hledání nových medoidů

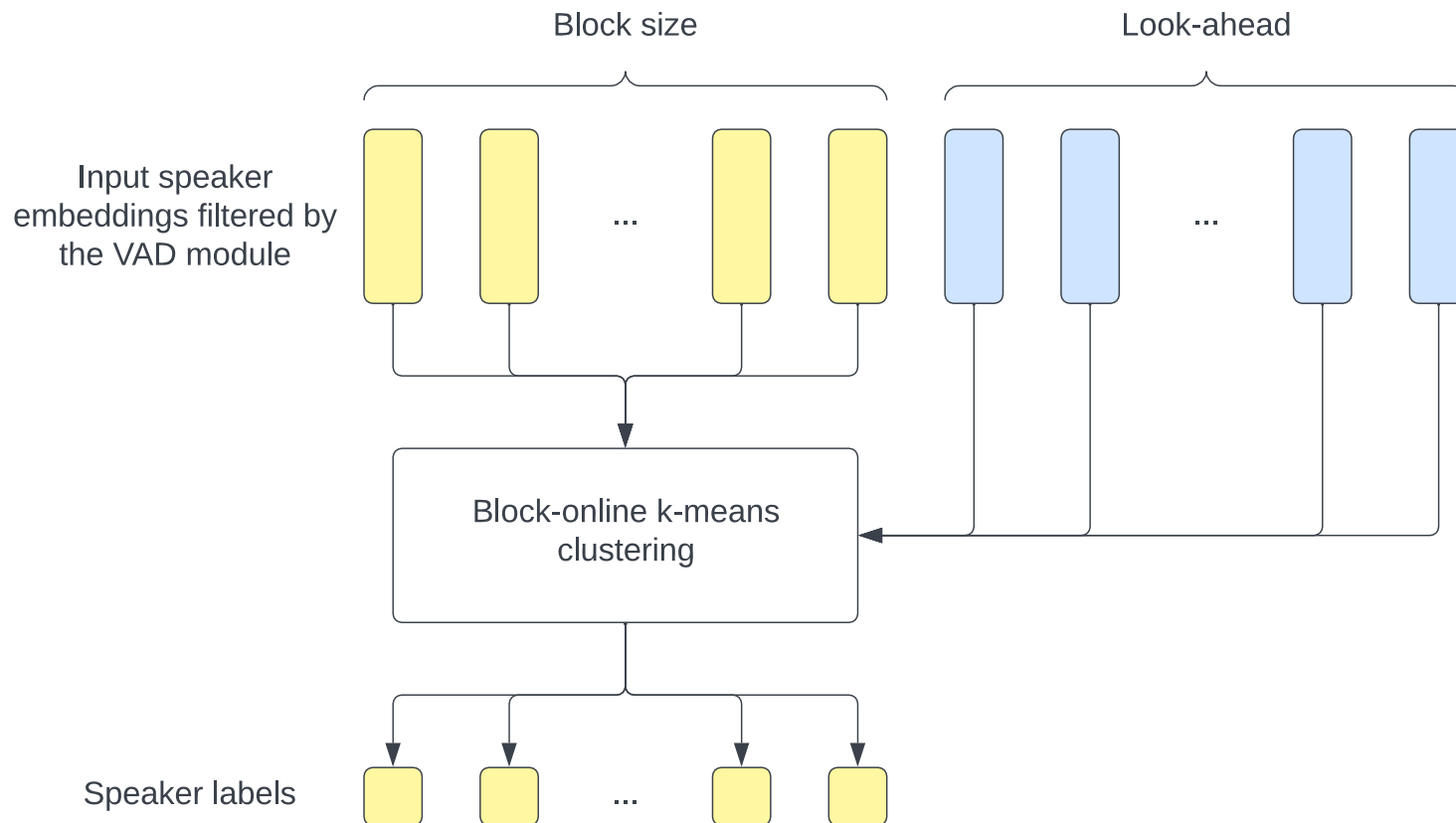
K-Medians

- **K-Medians**
 - Medián je méně náchylný na outliery
 - Mediány jsou použity jako centroidy (+ použití L1 normy jako vzdálenostní metriky)
 - Stejný postup jako u K-Means, pouze nový medián je vypočten jako medián jednotlivých příznaků (každé dimenze)

Sequential K-Means

- Podobný jako standardní K-Means algoritmus
- S přicházejícími body postupně aktualizujeme polohu centroidů
- Existuje i možnost, kdy předem nelze odhadnout počet shluků
 - Vytvoření prvního shluku s počátečními daty
 - Pro každý nový bod je potřeba se rozhodnout, zda bude zařazen do již existujícího shluku nebo bude vytvořen nový
 - Řešeno na základě vzdálenostní metriky

Block-online K-Means s look-ahead





Část II.: Shlukování založené na hustotě

Shlukování založené na hustotě

- Shluky jsou založeny na základě hustoty dat
 - Lokální kritérium
- Výhody
 - Nalezne shluky libovolného tvaru
 - Nevadí šum v datech
 - Stačí projít jednou

DBSCAN

- Density-Based Spatial Clustering Algorithm
- Shluk je definován jako maximální množina hustě spojených bodů

- Dva parametry

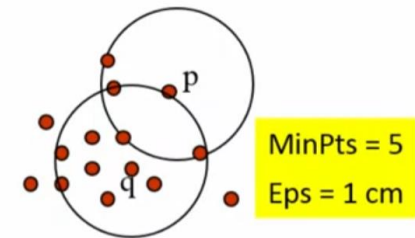
- **Epsilon (Eps)**

- Radius okolo bodu

- **MinPts**

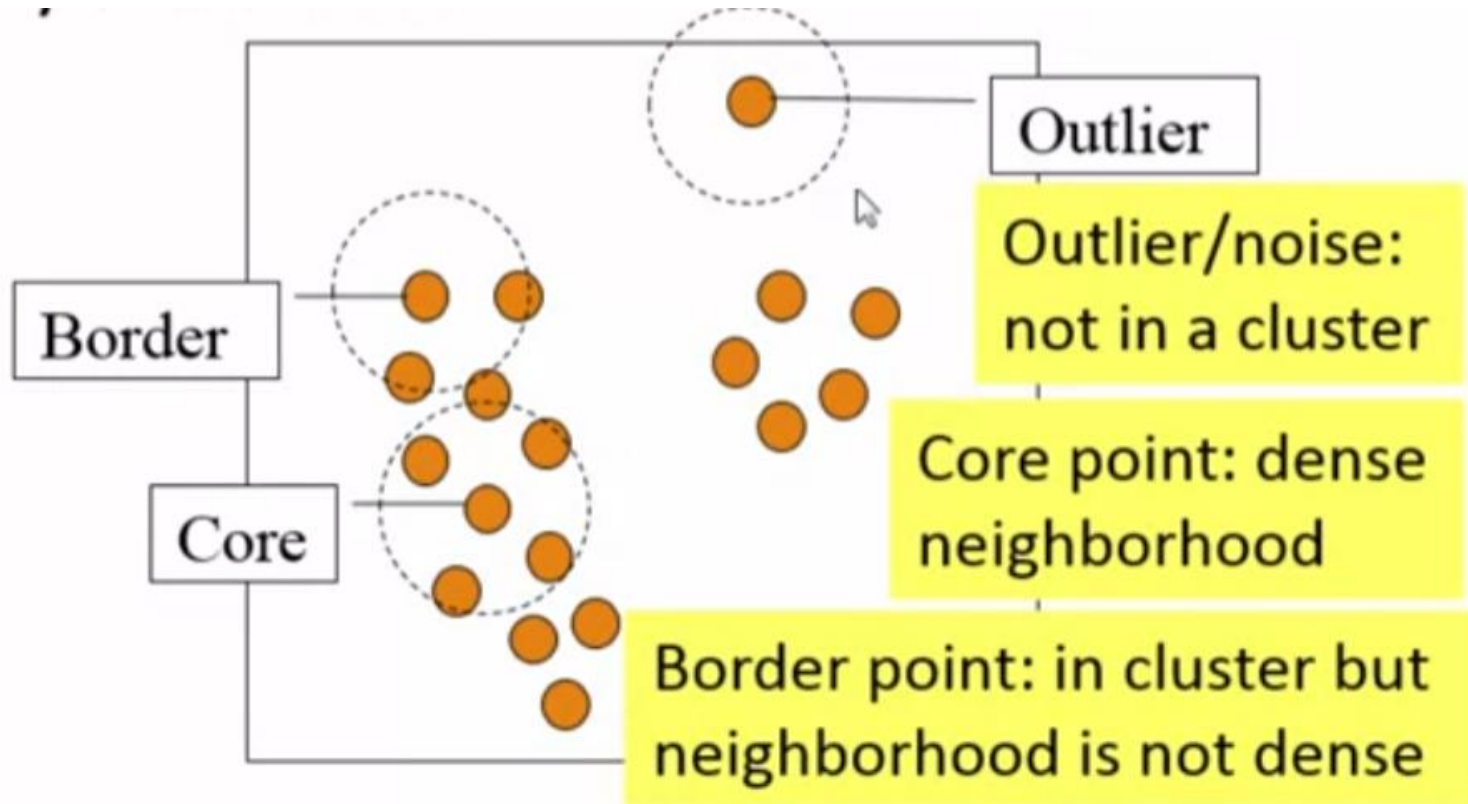
- Minimální počet bodů v okolí Eps

- Eps okolí bodu q : $N_{Eps}(q) = \{p \text{ patří do } D \mid d(p, q) \leq Eps\}$



<https://www.coursera.org/learn/cluster-analysis>

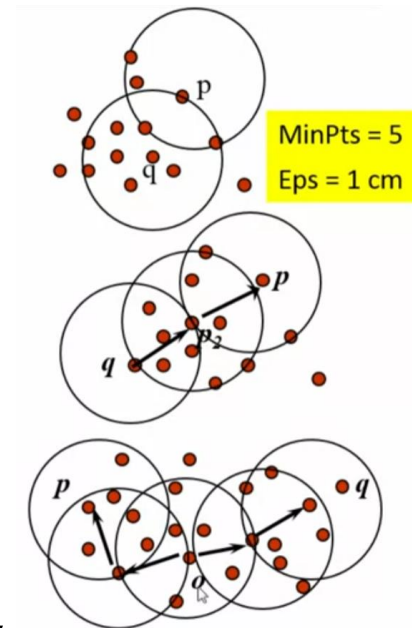
DBSCAN



<https://www.coursera.org/learn/cluster-analysis>

DBSCAN

- Přímá dosažitelnost
 - Bod p je přímo dosažitelný z bodu q , pokud:
 - p patří do $N_{Eps}(q)$ (bod p patří do okolí bodu q)
 - $|N_{Eps}(q)| \geq MinPts$ (bod q je core bod)
- Nepřímá dosažitelnost
 - Bod p je nepřímo dosažitelný z bodu q , pokud existuje řada bodů p_1, \dots, p_n ($p_1 = q, p_n = p$) taková, že p_{i+1} je přímo dosažitelný z p_i
- Propojenost
 - Body p a q jsou propojené, pokud existuje bod o , ze kterého jsou oba body (p a q) nepřímo dosažitelné



<https://www.coursera.org/learn/cluster-analysis>

DBSCAN

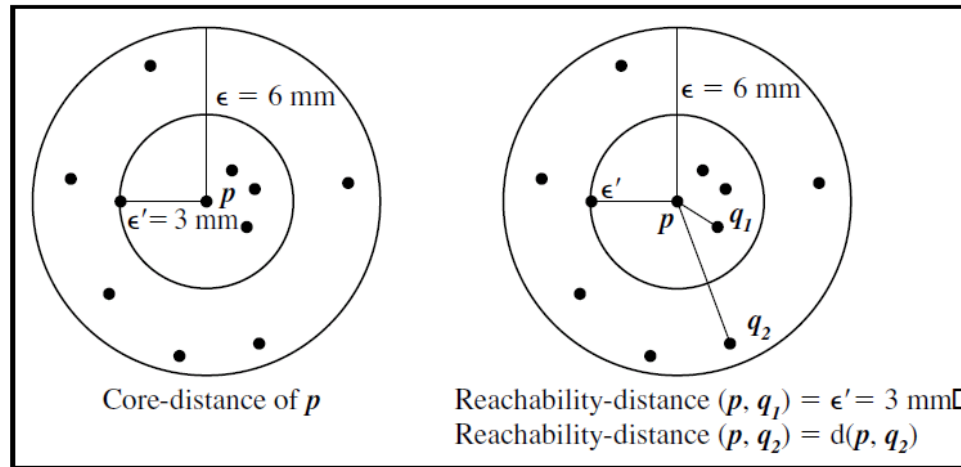
Algoritmus:

1. Vybrat počáteční bod p
2. Získat všechny nepřímo dosažitelné body
 - a) Pokud je bod p core bod \rightarrow cluster hotový
 - b) Pokud je p border bod, tak z něj ostatní body nemohou být nepřímo dosažitelné \rightarrow vybrat další počáteční bod z databáze
3. Pokračujeme, dokud jsme neprošli všechny body

OPTICS

- Ordering Points To Identify Clustering Structure
- Vytvořeno (téměř) stejnými autory jako DBSCAN
 - DBSCAN je citlivý na nastavení správných parametrů
- Parametr maximální Epsilon poskytujeme pouze pokud chceme zrychlit výpočet
- Vytvoření grafu dosažitelnosti, ze kterého lze extrahovat shluky

OPTICS



<https://www.coursera.org/learn/cluster-analysis>

- **Core distance** – minimální hodnota ϵ , se kterou bude v okolí *MinPts* bodů (aby se stal core bodem)
- **Reachability-distance** – minimální radius, díky kterému je p nepřímo dosažitelný z bodu q

$$\max(\text{core_distance}(q), d(q, p))$$

- Pokud je q core bod, jinak nedefinováno

OPTICS

- Graf dosažitelnosti
 - Čím hlubší je oblast, tím hustší je shluk
- Body patřící do shluku mají nízkou vzdálenost dosažitelnosti k jejich nejbližším sousedům, proto vytvoří údolí

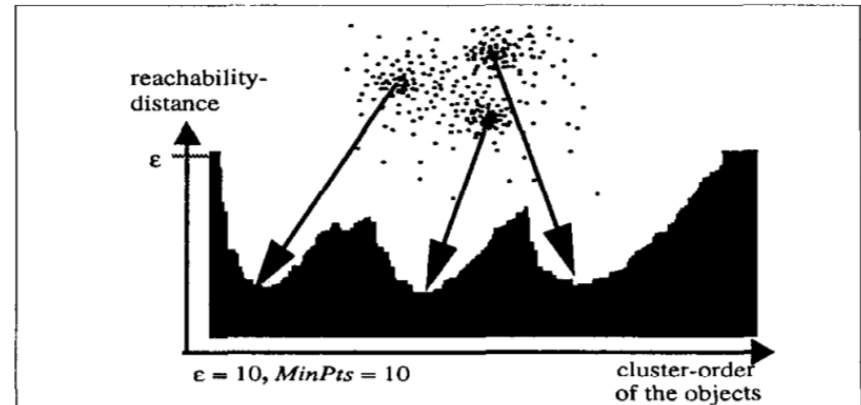


Figure 9. Illustration of the cluster-ordering



Figure 12. Reachability-plots for a data set with hierarchical clusters of different sizes, densities and shapes

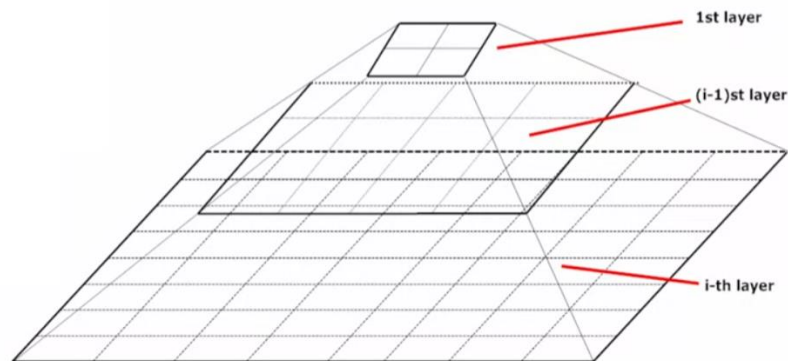
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.407.5572&rep=rep1&type=pdf>



Část III.: Shlukování založené na mřížce

STING

- Statistical Information Grid
- Vytváříme mřížkovou strukturu
 - Pro každou buňku jsou spočítány statistiky
 - Počet bodů, průměr, min, max, typ rozdělení
 - Postupně vytváříme další vrstvy



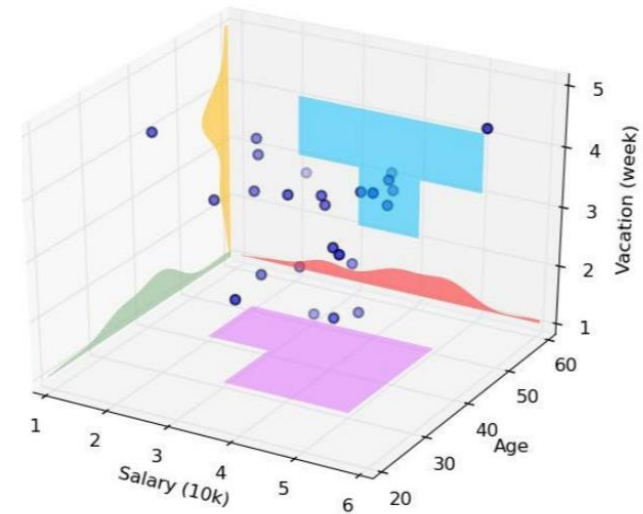
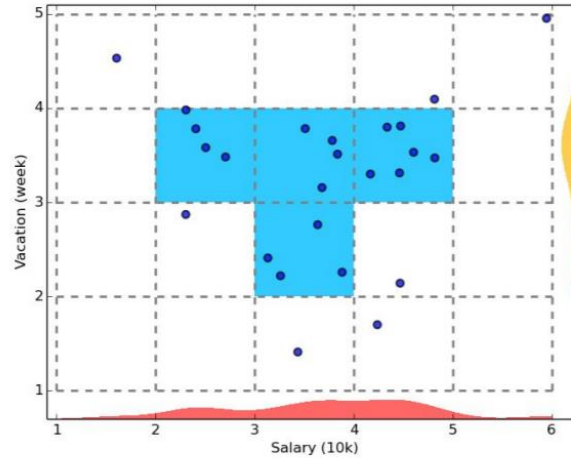
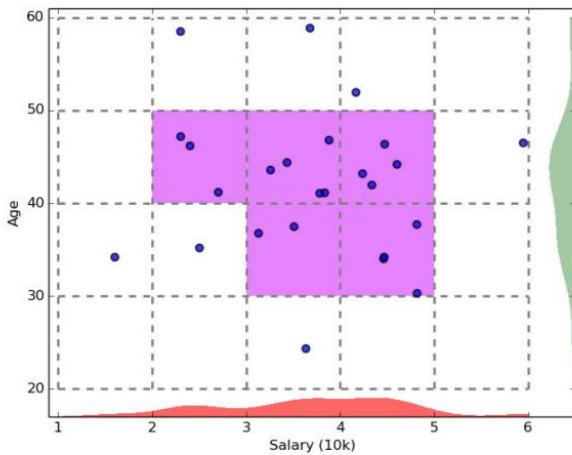
<https://www.coursera.org/learn/cluster-analysis>

CLIQUE

- Clustering in Quest
- Založený na mřížce
 - Rozděluje prostor mřížkou a počítá počet bodů v buňce
- Založený na hustotě
 - Shluk je vytvořen z husté množiny sousedících bodů
 - Za podmínky, že je počet bodů v buňce větší, než vstupní parametr modelu
- Automaticky nalezne podprostor, který umožňuje lepší shlukování než původní prostor dat – založeno na Apriori principu

CLIQUE

- Bottom-up přístup



https://list01.biologie.ens.fr/www/d_read/machine_learning/SubspaceClustering/CLIQUE_algorithm_grid-based_subspace_clustering.pdf



Část III.: Vyhodnocení shlukování

Měření kvality shlukování

- Externí (Supervised)
 - Porovnání s předem daným požadovaným výsledkem např. na základě vzdálenosti
- Interní (Unsupervised)
 - Vyhodnocení správnosti shlukování na základě toho, jak dobře jsou shluky separované a kompaktní
 - Silhouette koeficient
- Relativní
 - Porovnání výsledků algoritmu s různým nastavením vstupních parametrů

Měření kvality shlukování

- Purity (čistota)
 - Měří zastoupení dominantních členů jednotlivých tříd ve shlucích (hodnota 0 až 1)
 - Pro shluk i :

$$purity_i = \frac{1}{n_i} \max_j \{n_{ij}\}$$

- Pro celý výsledek

$$purity = \sum_{i=1}^r \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^r \max_j \{n_{ij}\}$$

- Např.: $purity_1 = \frac{30}{50}$; $purity_2 = \frac{20}{25}$; $purity_3 = \frac{25}{25}$;

$$purity = \frac{30 + 20 + 25}{100} = 0.75$$

- Problém – výpočet čistoty je stejný u obou tabulek

$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	30	20	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	50	25	100

<https://www.coursera.org/learn/cluster-analysis>

Měření kvality shlukování

- Maximum matching
 - Třída může patřit pouze jednomu shluku
 - Váha $w(e_{ij}) = n_{ij}$; $w(M) = \sum_{e \in M} w(e)$
 - Maximum weight matching

$$match = \operatorname{armax}_M \left\{ \frac{w(M)}{n} \right\}$$

- Např.: zelená $\rightarrow match = purity = 0.75$
 oranžová \rightarrow

- Možnost $\rightarrow \frac{w_1(M)}{n} = \frac{30 + 5 + 25}{100} = 0.6$
- Možnost $\rightarrow \frac{w_2(M)}{n} = \frac{20 + 20 + 25}{100} = \mathbf{0.65}$
 $match = \mathbf{0.65}$

$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	30	20	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	50	25	100

<https://www.coursera.org/learn/cluster-analysis>

Užitečná literatura / kurzy

- [How much can k-means be improved by using better initialization and repeats?](#)
 - Článek k porovnání různých k-means přístupů (2019)