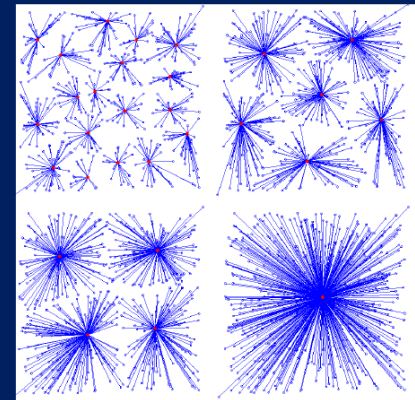


Shlukování

I. KOLINGEROVÁ



[Šed17]

Popis

2

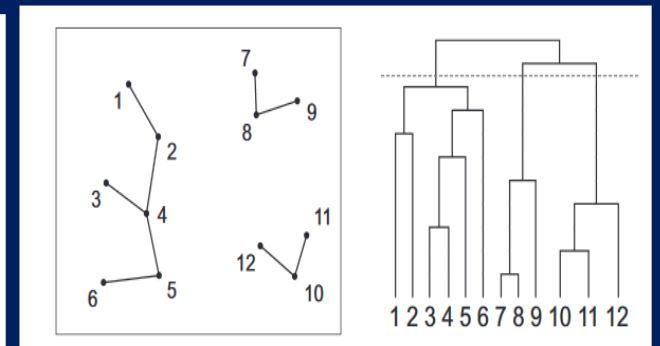
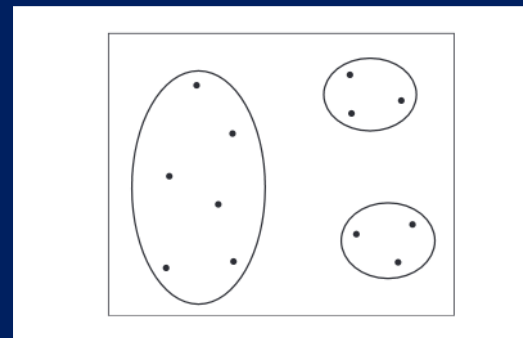
- ▶ **Shlukování:** rozdělování množiny objektů na skupiny tak, aby si objekty v jedné skupině byly podobnější než objekty z různých skupin
- ▶ Vstup často reprezentován body v E^d
- ▶ Často využívaný a velmi užitečný princip => řada algoritmů
- ▶ Prameny: např. [Ezu22, Tow20, Sci25, Ska09, Kuc]

- ▶ Pro analýzu dat
- ▶ Kdy vhodné:
 - ▶ Počáteční porozumění neznámým datům (potenciální vzory nebo trendy hodné dalšího zkoumání)
 - ▶ Segmentace dat – např. najít segmenty trhu se společnými vlastnostmi (nákupní zvyky, preference, demografická fakta), pro ně tržní strategie „na tělo“
 - ▶ Přidělení zdrojů – které skupiny nebo oblasti potřebují nejvíc pozornosti nebo zdrojů

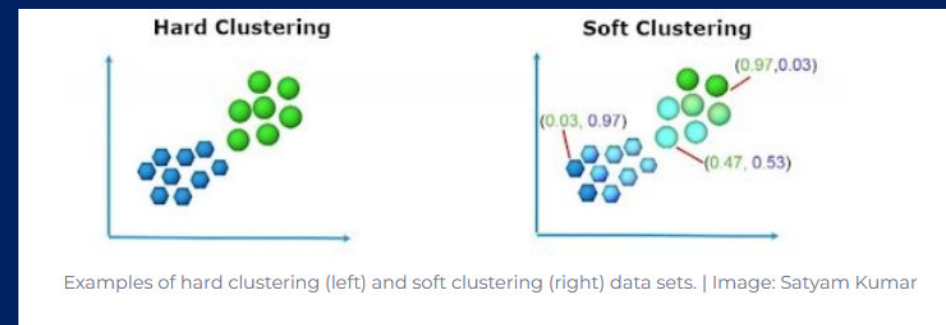
Základní třídy shlukovacích metod

4

- ▶ Hierarchické a nehierarchické
 - ▶ Jedna úroveň x více – vzniká tzv. dendrogram
- ▶ Deterministické a stochastické



- ▶ Randomizovaná rozhodnutí nebo ne (náhodnost typicky nezaručí optimum, ale může být rychlejší a v průměru fungovat lépe)
- ▶ Hard a soft (fuzzy)
 - ▶ Každý objekt patří jen do jednoho shluku nebo ne
(u soft získány také váhy/pravděpodobnosti příslušnosti bodu do jednotlivých shluků)



Examples of hard clustering (left) and soft clustering (right) data sets. | Image: Satyam Kumar

Obr. [Kumar S., Ska09]

Základní postup

5

- ▶ **Přesné řešení:** Nalezení nejlepšího rozkladu vyzkoušením všech možných uspořádání shluků – většinou neproveditelné
- ▶ Typicky nějaký počáteční rozklad a jeho následná optimalizace
- ▶ Konec obvykle v lokálním optimu
- ▶ **Kritéria kvality:** podle cílů, typicky součet kvadrátů odchylek objektů od těžišť příslušných shluků, podobnost objektů ve shluku, míra separace shluků, rovnoměrnost rozložení objektů uvnitř shluku, rovnoměrnost rozložení objektů do shluků ...
- ▶ **Optimalizace:** typicky přesuny objektů mezi shluky
- ▶ **Stanovení vhodného počtu shluků:** pokud není dán zadáním a vybraný algoritmus to neumí, nutno vyzkoušet několik hodnot nebo výpočet z nějakého tzv. indexu (např. Goodman-Kruskal)
- ▶ Stanovení počátečního rozkladu: např. náhodně

Hlavní nehierarchické metody

6

- ▶ S konstantním nebo proměnným počtem shluků
- ▶ **Konstatní počet shluků:** MacQueenova metoda (K-means) a její varianty, facility location, fuzzy C-means
- ▶ **Proměnný počet shluků:** MacQueenova metoda se dvěma parametry, ISODATA (Iterative Self-Organizing Data Analysis Techniques), Wishartova metoda RELOC, Density-based spatial clustering (DBSCAN)
- ▶ **Výběr pro tuto přednášku:**
 - ▶ Deterministické K-means shlukování
 - ▶ Fuzzy C-means shlukování (FCS)

K-means shlukování

7

Vstup: množina vstupních bodů $\{x_i, i=1,2,\dots,N\}$,
požadovaný počet shluků k ,

Výstup: Pro každý bod jeho příslušnost ke clusteru

1. Vybrat k bodů jako počáteční středy shluků
2. Opakovat, dokud dochází k významnější změně středů shluků:
 - ▶ Vytvořit k shluků přiřazením bodů k nejbližšímu středu shluku
 - ▶ Přepočítat středy všech shluků

Výhody a nevýhody K-means

8

- ▶ **Výhoda:** jednoduchý, flexibilní algoritmus
- ▶ **Nevýhoda:** nutnost zadat k , citlivost na počáteční výběr shluků
 - vhodné opakovat s různou volbou

Varianta **Facility Location**: zkoušíme otvírat, zavírat shluky, zkoumáme výhodnost z hlediska celkového ohodnocení

Fuzzy C-means shlukování (FCS)

9

Vstup:

- ▶ množina vstupních bodů $\{x_k, k=1,2,\dots,N\}$,
- ▶ požadovaný počet shluků C ,
- ▶ „fuzziness parameter“ m (typicky 2, $1.25 < m < 2$), větší m – „více fuzzy“

Výstup: Pro každý bod pravděp. příslušnosti ke každému clusteru

1. Inicializovat pravděp. příslušnosti bodů x_i do shluků j v matici $U^{(0)}=[u_{ij}]$ (např. náhodně); $k=0$

2. V k -té iteraci:

- ▶ Spočítat středy shluků $C^{(k)}=[c_j]$ s užitím $U^{(k)}$ jako vážené sumy bodů (vztah 1)
- ▶ Update $U^{(k)}$ na $U^{(k+1)}$ (vztah 2)
- ▶ Pokud se $U^{(k)}$ a $U^{(k+1)}$ liší o méně než ϵ , konec, jinak inkrement k a další iterace

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m}$$

Vztah 1

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

Vztah 2

Výhody a nevýhody FCS

10

- ▶ **Výhody:**
 - ▶ Prvek může patřit do více shluků
 - ▶ „Nezaokrouhlujeme tak často“
- ▶ **Nevýhody:** podobné jako K-means
 - ▶ Nutnost nastavení počtu shluků předem
 - ▶ Citlivý na počáteční volbu středů shluků (můžeme uvíznout v lokálním minimu)
 - ▶ Časté využití euklidovské vzdálenosti nemusí odpovídat typu vah pro daný problém
 - ▶ Výrazně pomalejší než např. K-means

Hlavní hierarchické metody

11

- ▶ Slučující (aglomerativní) anebo rozdělující (diversivní)
 - ▶ Postup zdola nebo shora
- ▶ Výběr pro tuto přednášku:
 - ▶ Aglomerativní hierarchické shlukování

Aglomerativní hierarchické shlukování (1)

12

Vstup: množina vstupních bodů $\{x_k, k=1,2,\dots,n\}$,
požadovaný počet shluků nebo
povolená max. vzdálenost mezi shluky

Výstup: Pro každý bod příslušnost k nějakému shluku

- ▶ Úroveň $i-1$ na úroveň i : Slučujeme vždy dva shluky nejpodobnější podle nějakého kritéria
- ▶ Končíme při dosažení potřebného počtu shluků nebo na základě vzdálenosti mezi shluky (každé seskupení vzniká ve větší vzdálenosti od shluků z předchozího uskupení)
- ▶ **Metrika:** typicky vzdálenost těžišť shluků, ale i jiné volby (např. Wardova metoda – minimalizace čtverců odchylek bodů od těžiště jejich shluku – pro rovnoměrnou velikost shluků), volba metriky podle typu dat, viz např. [Kuc]

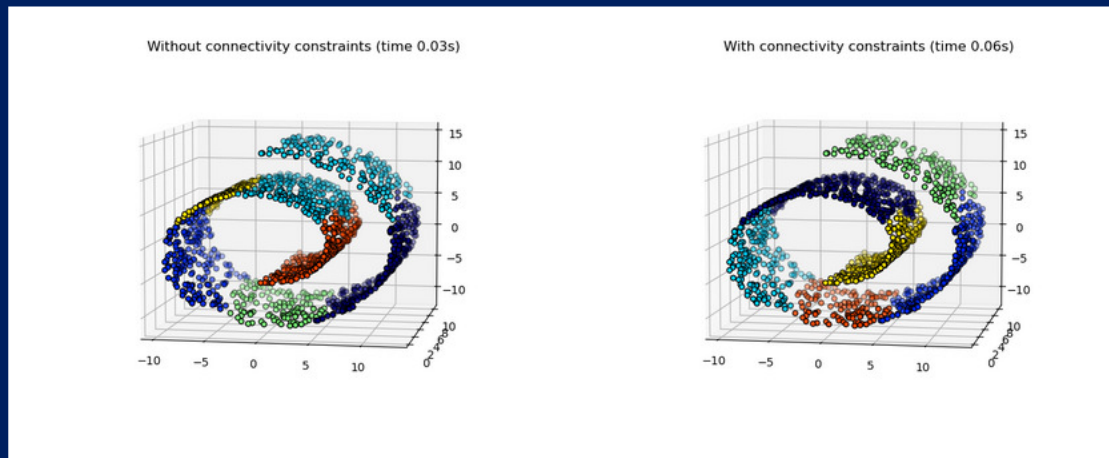
Nevýhoda:

- ▶ rozhodnutí z předchozí úrovně už nelze změnit
- ▶ většina metrik kromě Warda vede k posilování velkých shluků (rich get richer)

Aglomerativní hierarchické shlukování (2)

13

- ▶ Modifikace: přidání omezení povolené vzdálenosti (connectivity constraints) v podobě matice definující sousedy pro každý vzorek – povoleno spojit jen sousední shluky
- ▶ Umožnění určité lokální struktury, ale i urychlení

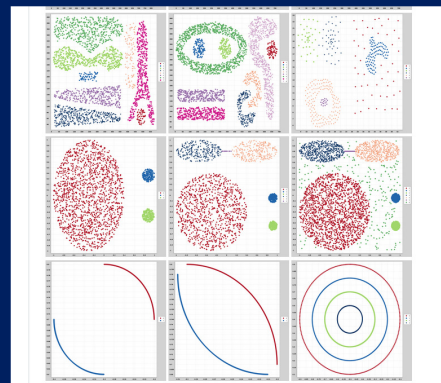
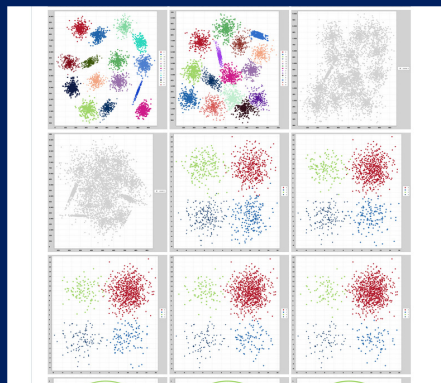


Obr. [Sci25]

Datové množiny pro testy

14

- ▶ Např.
 - ▶ **UC Irvine Machine Learning Repository** [UC]: iris, haberman, blood, glass, wine, seeds, cancer ...
 - ▶ Deric/clustering benchmark – umělá i reálná data [Der]



Obecné poznámky

15

- ▶ U dat bez jasné a jednoznačné struktury pravděpodobně z různých metod různé výsledky a naopak
- ▶ Typicky metody citlivé na outliers - vhodná jejich předběžná detekce a eliminace
- ▶ Hodnocení shluků: na základě vzdáleností bodů ve shluku (v „silném shlukování“ menší vzdálenosti – více homogenní data) a vzdálenosti bodů v různých shlucích (v „silném shlukování“ větší vzdálenosti – více heterogenní data)

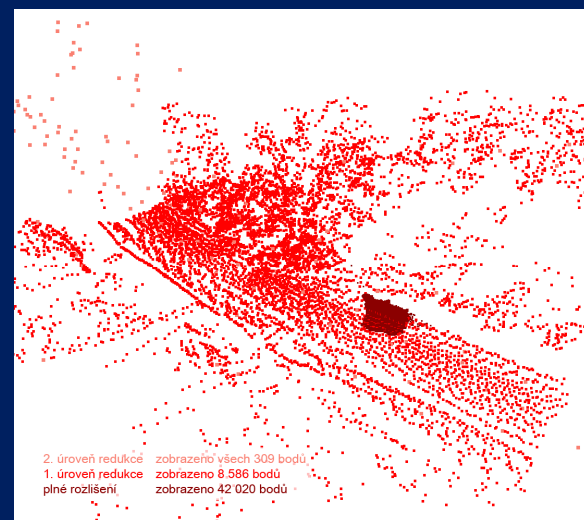
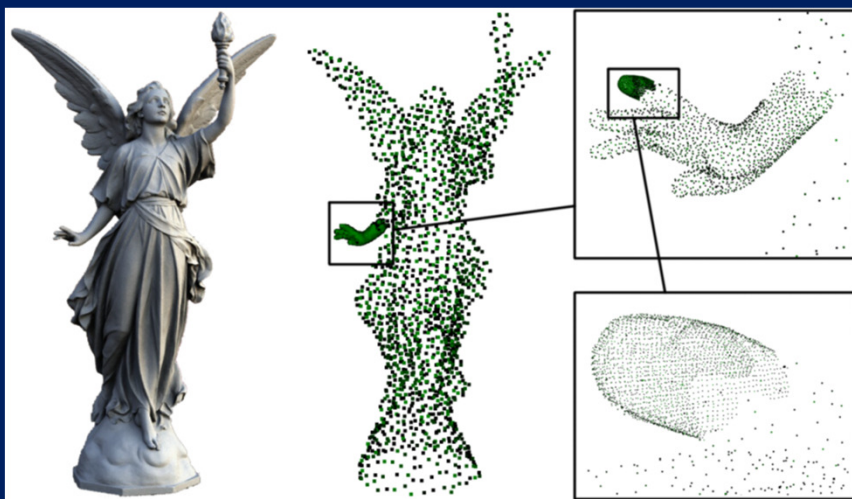
Domácí příklad využití shlukování (1)

16

- ▶ Hierarchická manipulace s velkými geodetickými daty [Ska13]
- ▶ **Motivace:** bodová mračna v geo oblasti rozsáhlá, požadována i generalizace dat pro různá „měřítka“, ale i dostupnost detailních dat
 - ▶ **Hlavní myšlenka:** Udržovat v paměti model v nízkém rozlišení, vybrané části ve větším detailu
 - ▶ **Postup:**
 1. Víceúrovňové shlukování data streamu, většina uložena na disk
 2. Interaktivní prohlížení dat s možností LOD a případnou dynamickou triangulací (tj. do triangulace jde přidávat a odebírat body)

Domácí příklad využití shlukování (2)

17

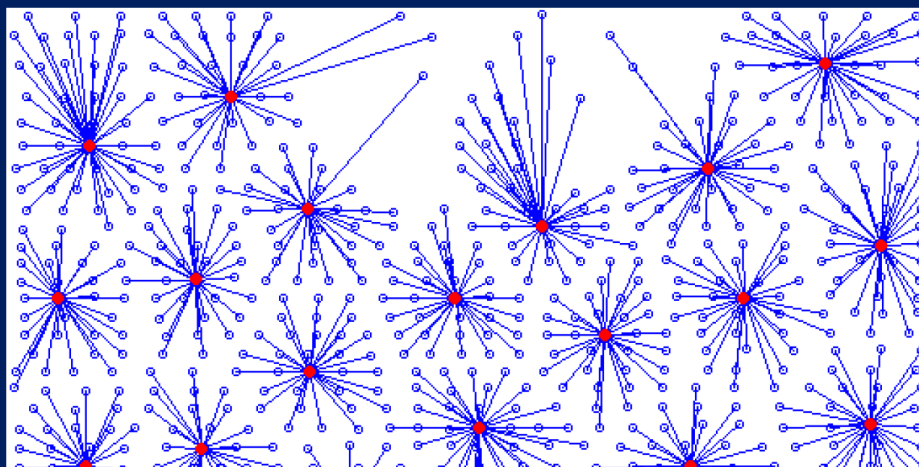


[Ska13]

Domácí příklad využití shlukování (3)

18

1. Shlukování – skupina podobných bodů (typicky geometricky blízkých) nahrazena jedním reprezentantem – středem shluku



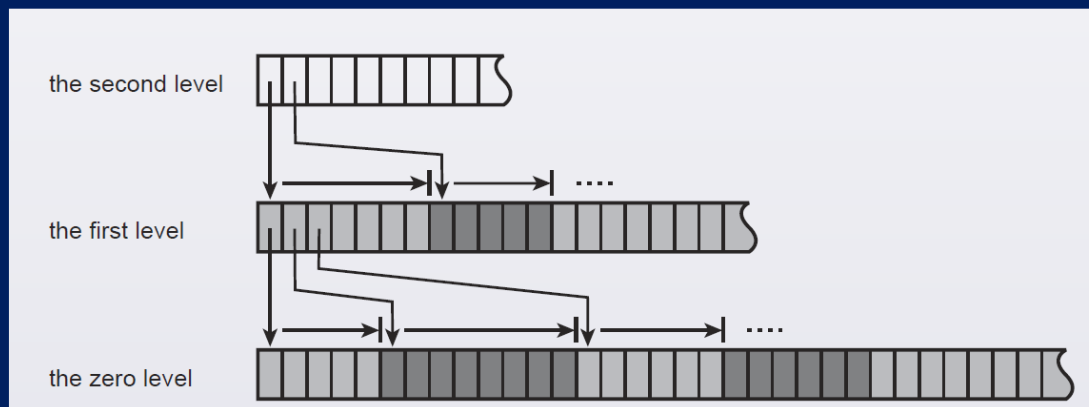
[Ska13]

[Ivo H]

Domácí příklad využití shlukování (4)

19

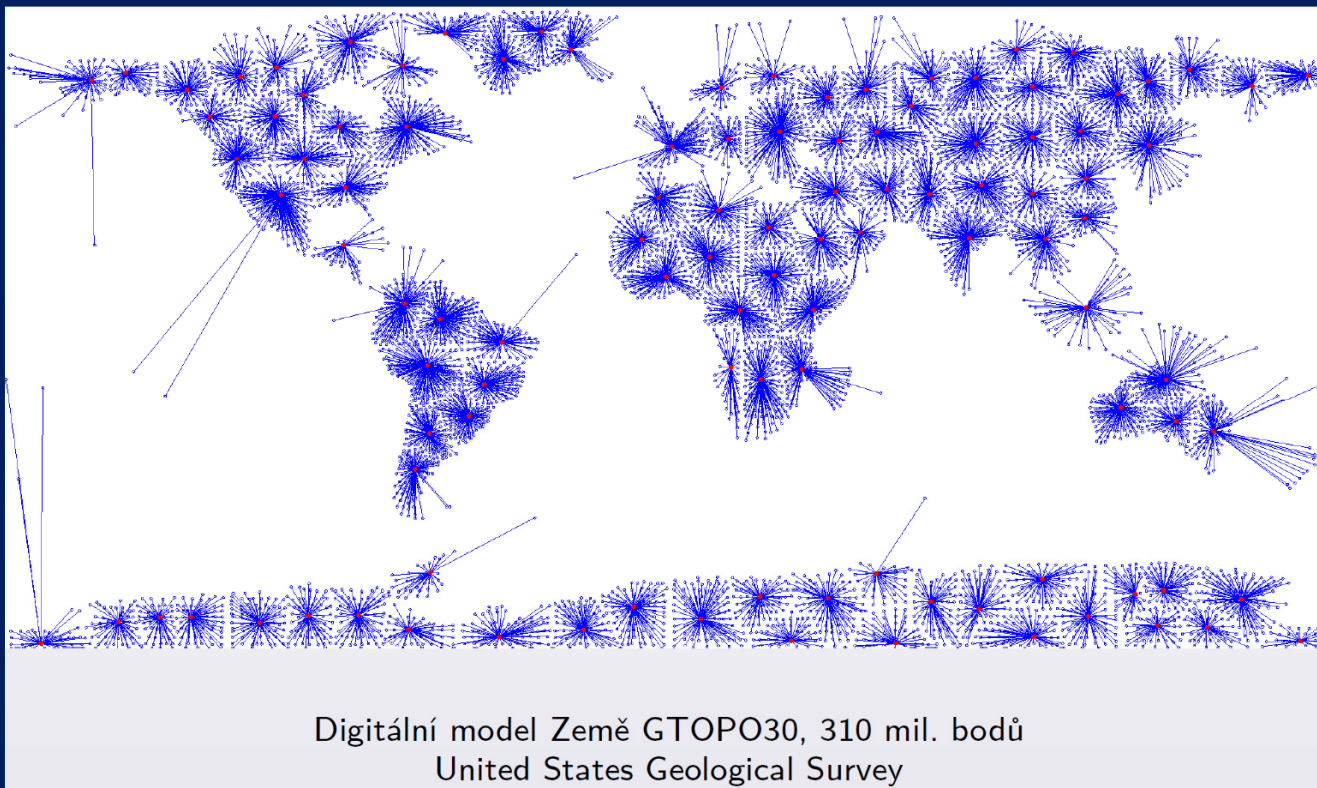
- ▶ Shlukování prováděno na data streamu, v jednom průchodu, po blocích, hierarchicky
- ▶ Výsledky uloženy do binárních souborů
- ▶ Efektivní přístup k jednotlivým bodům v libovolné úrovni hierarchie



[Ska13]

Domácí příklad využití shlukování (5)

20



[Ska13])

Domácí příklad využití shlukování (6)

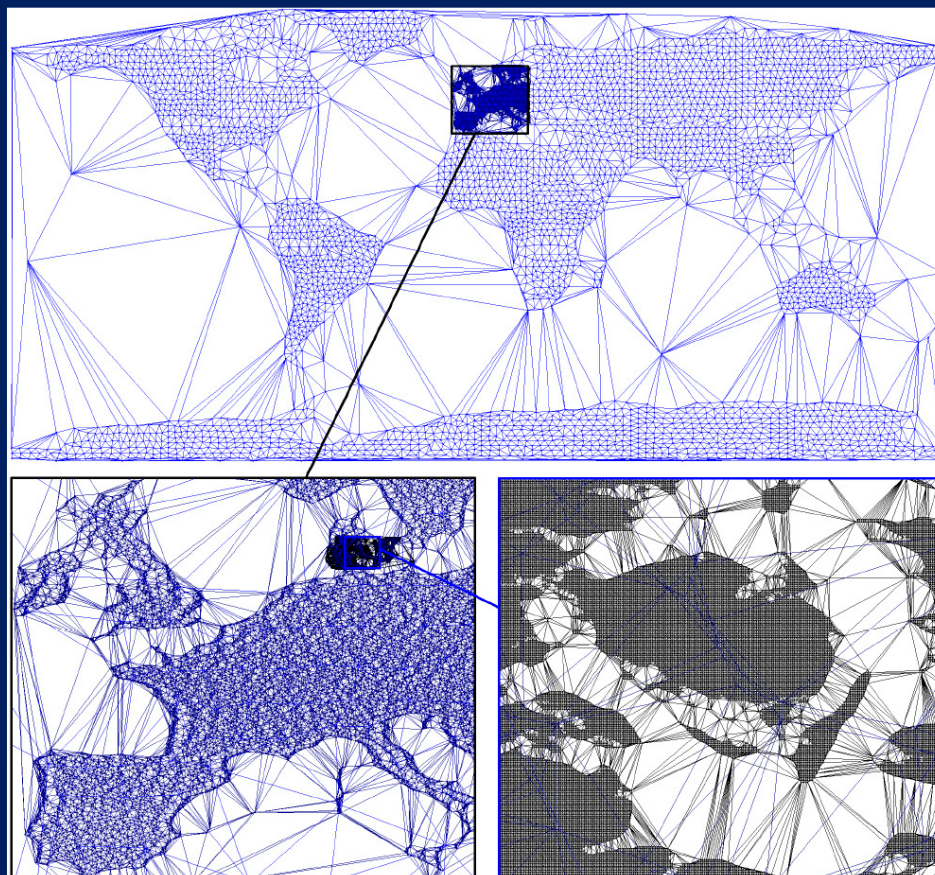
21

2. Interaktivní prohlížení dat

- ▶ Zobrazena nevyšší úroveň hierarchie bodů, místní zjemňování a naopak redukce podle aktuálního požadavku – shluky se v reálném čase rozbalují nebo zabalují
- ▶ Na začátku programu triangularizována nejvyšší úroveň dat, podle potřeby dynamické změny

Domácí příklad využití shlukování (7)

22

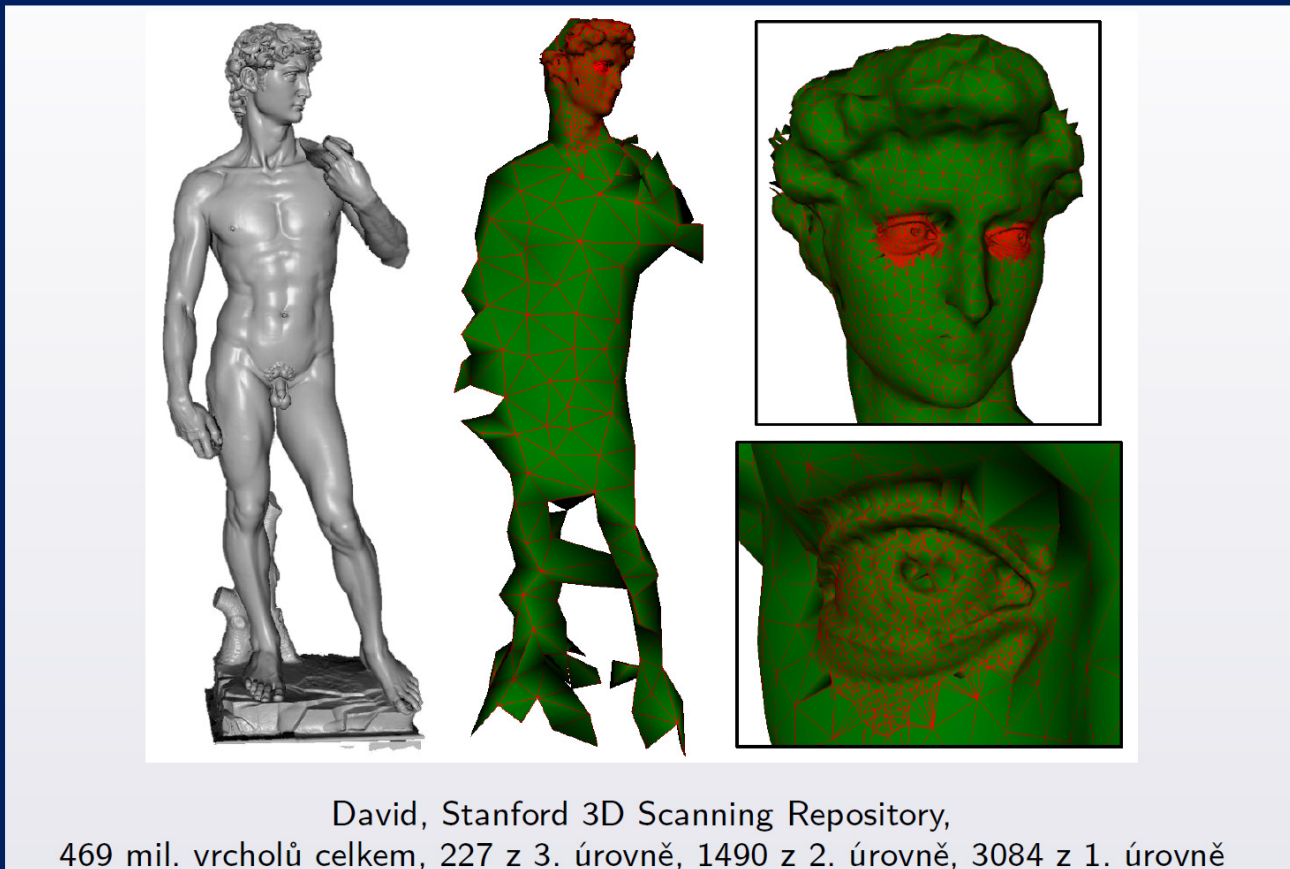


Digitální model Země GTOPO30, USGS,
310 mil. vrcholů celkem, 3600 vrcholů na nejvyšší úrovni

[Ska13]

Domácí příklad využití shlukování (8)

23



V 3D je triangulace složitější problém, naše řešení zde ne zcela dotaženo

[Ska13]

- ▶ [Der] Deric/clustering benchmark, <https://github.com/deric/clustering-benchmark?tab=readme-ov-file>
- ▶ [Ezu22] Ezugwu A.E., Ikotun A.M., Oyelade O.O. et. Al.: A comprehensive survey of clustering algorithms: State of the art machine learning applications, taxonomy, challenges, and future research prospects, Engineering Applications of Artificial Intelligence 110 (2022) 104743, Elsevier
- ▶ [Gee] Geeks for Geeks, <https://www.geeksforgeeks.org/machine-learning/ml-fuzzy-clustering/>
- ▶ [Kuc] Kučera J.: Shluková analýza, https://is.muni.cz/th/172767/fi_b/5739129/web/web/main.html
- ▶ [Sci25] Scikit- Learn, 2.3 Clustering, 2007-25, <https://scikit-learn.org/stable/modules/clustering.html#dbscan>
- ▶ [Ska09] Skála J.: Algorithms for Manipulation with Large Geometric and Graphic Data, TR, KIV FAV ZČU, 2009, <https://www.kiv.zcu.cz/cs/Research/Technical-reports.html>
- ▶ [Ska13] Skála J.: Algoritmy pro manipulaci s velkými geometrickými daty, disertační práce, FAV ZČU, Plzeň, 2013, na projektu spolupracovali I. Kolingerová, V. Čada a další členové KGM FAV
- ▶ [Šed17] Šedivý T.: Pomocné programové vybavení a experimenty pro vizualizaci modelů terénu, bakalářská práce, ZČU Plzeň, 2017
- ▶ [Tow20] Towards Data Science, 2020, <https://towardsdatascience.com/k-means-dbscan-gmm-agglomerative-clustering-mastering-the-popular-models-in-a-segmentation-c891a3818e29/>
- ▶ [UC] UC Irvine Machine Learning Repository, <https://archive.ics.uci.edu/>