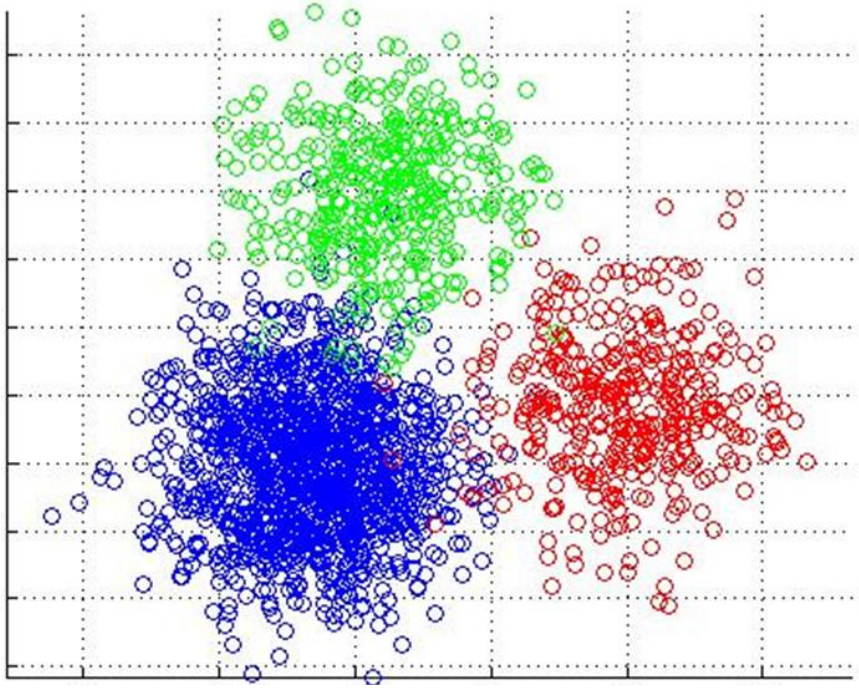
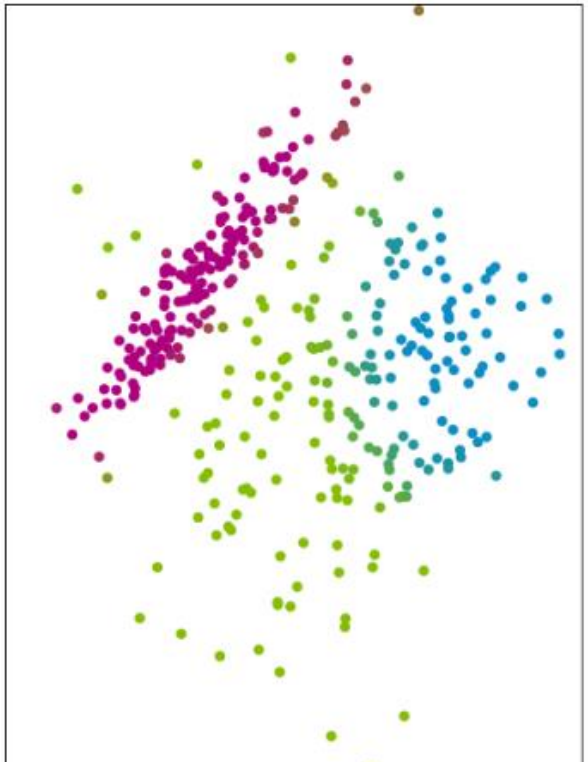
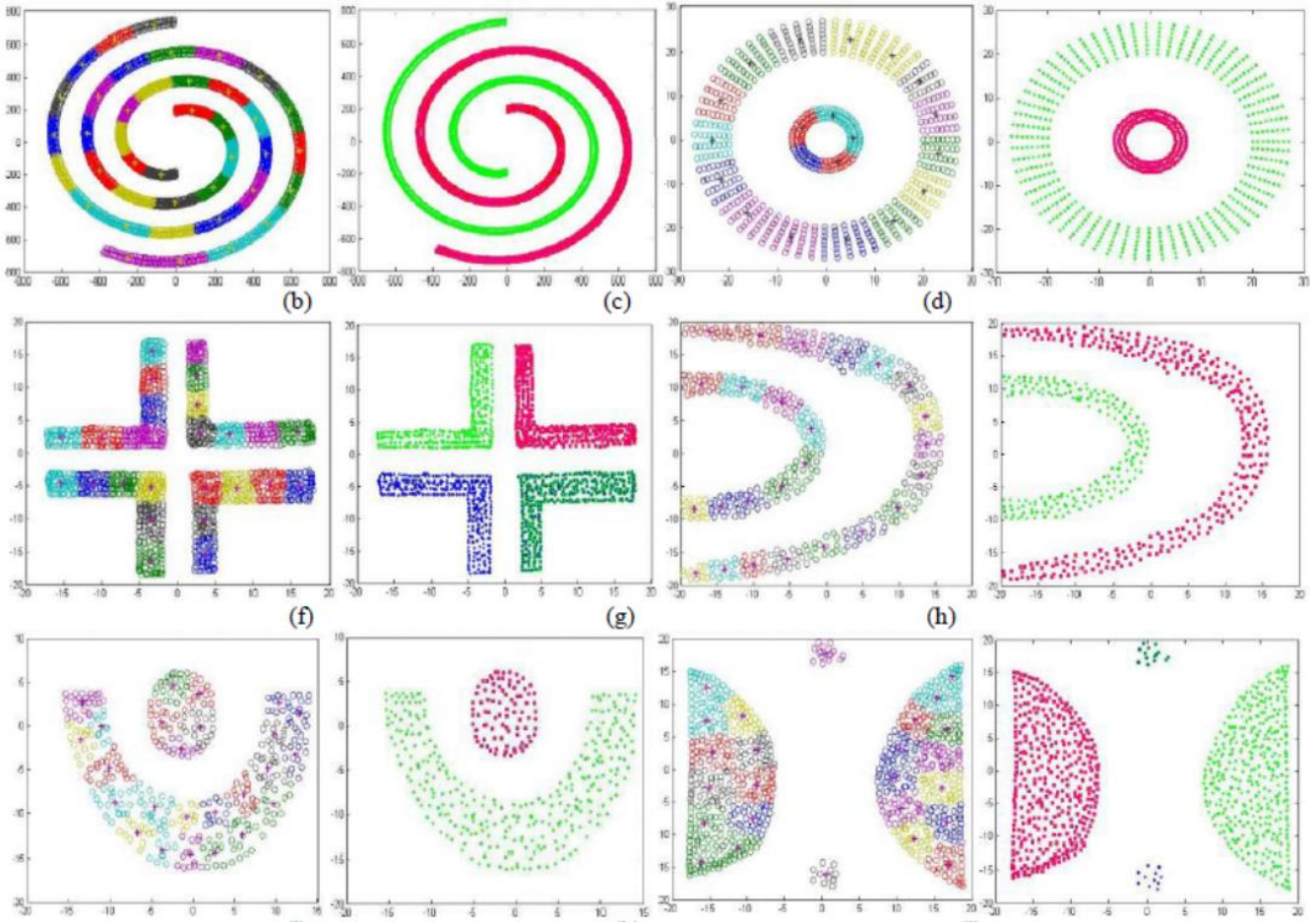


K-means



Gaussian Mixture Model

Klasterovanje



Klasterovanje

Definicija i primena

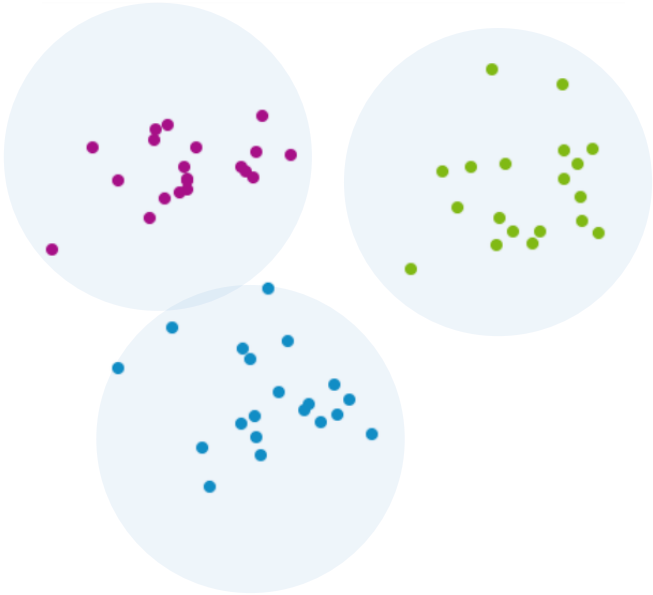
Klasterovanje: motivacioni primer

- Imamo grupu članaka koji nisu anotirani (nemamo oznake tema: sport, vesti, ...)

- Želimo da strukturiramo dokumete prema temi
  - Lakša pretraga
  - Preporuke

Šta je klasterovanje?

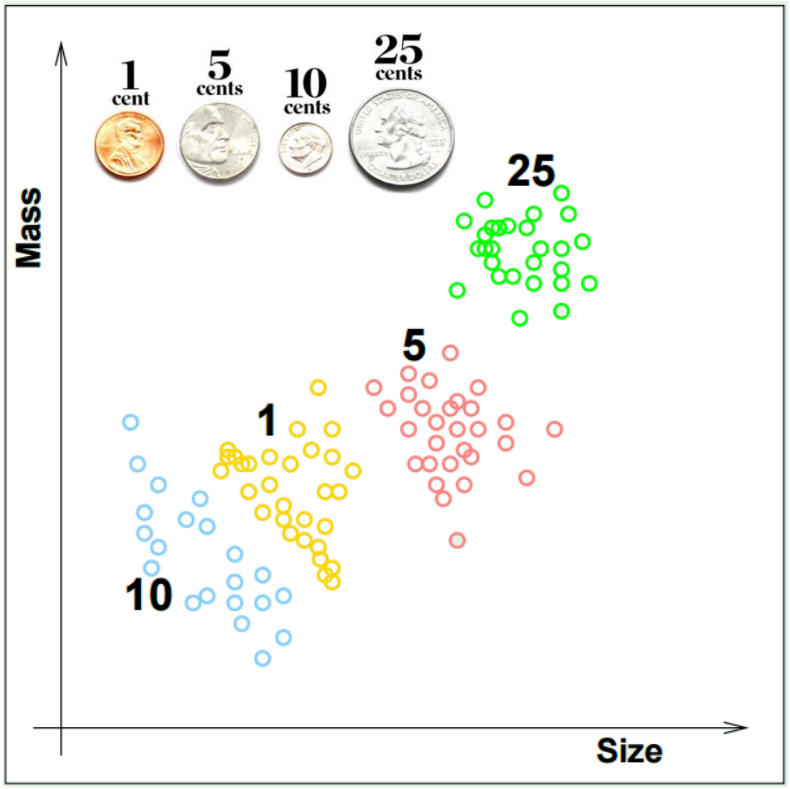
- Za dati skup opservacija,
- sa definisanom merom udaljenosti između opservacija,
- grupisati opservacije u određeni broj klastera,
- tako da članovi unutar jednog klastera:
  - budu što bliži članovima istog klastera
  - budu što dalji od članova drugih klastera



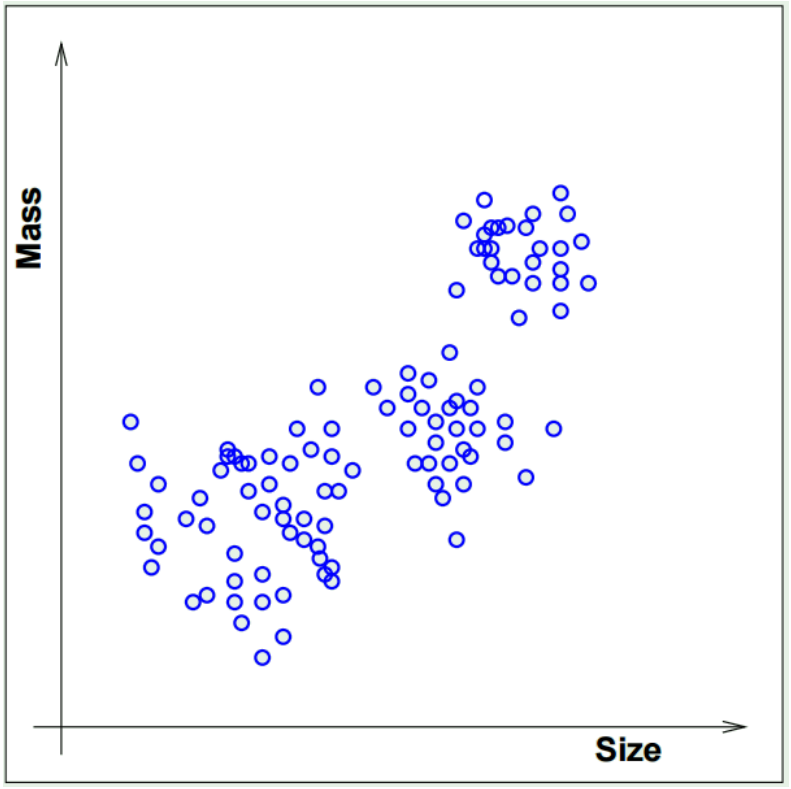
Nenadgledano obučavanje

- Klasterovanje je zadatak nenadgledanog obučavanja
  - Podaci nisu anotirani (nemaju oznaku klase)
  - Potrebno je da pronađemo strukturu posmatrajući samo ulaz *x*

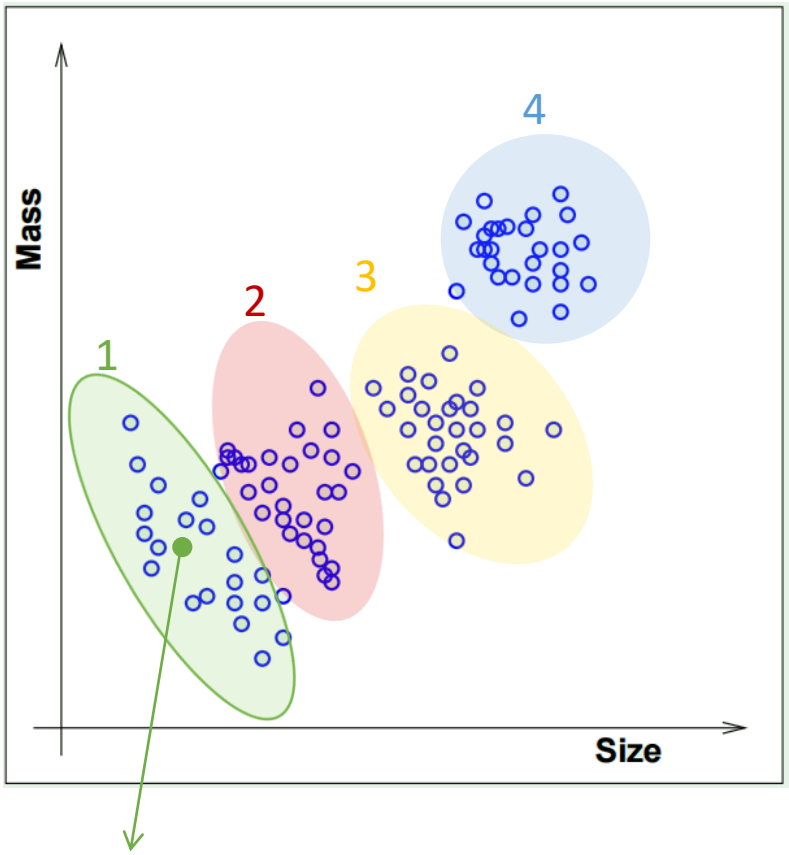
Nadgledano obučavanje



Nenadgledano obučavanje



Klasterovanje



- Ulaz:

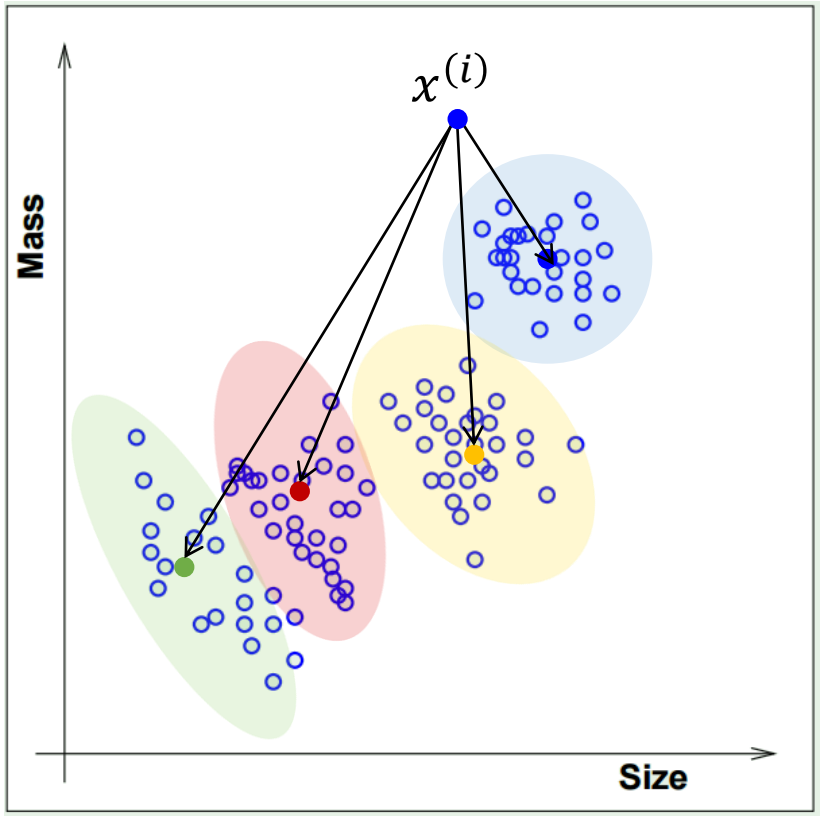
$$T = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$$

- Izlaz: oznake klastera

$$\{(x^{(1)}, z^{(1)}), \dots, (x^{(N)}, z^{(N)})\},$$
$$z \in \{1, .., 4\}$$

- Ne znamo kojoj klasi pripada koji ulaz ali ćemo pronaći strukturu (na osnovu sličnosti ulaza) i obeležiti ulaze koji su međusobno slični da pripadaju istom klasteru
- Klastere ćemo definisati putem centra (centroida) i oblika/rasprostiranja oko centroida

Klasterovanje



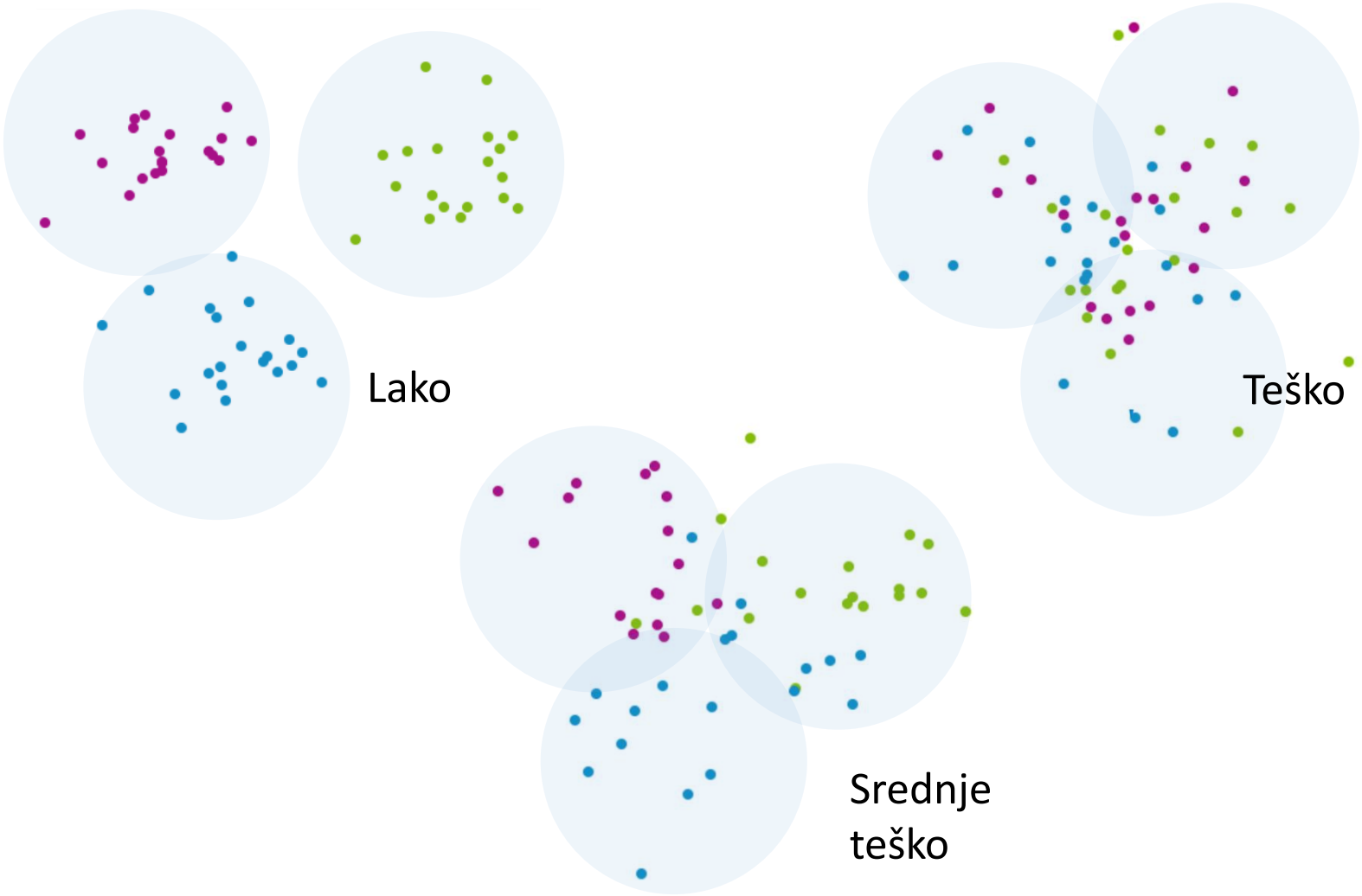
- Opservaciju $x^{(i)}$ ćemo dodeliti klasteru j ako:
  - je *score* pripadanja klasteru j veći od *score* pripadanja drugim klasterima
  - Često se zbog jednostavnosti *score* računa kao rastojanje od centra klastera (ignorišemo oblik klastera)

Klasterovanje

- Zadatak klasterovanja deluje teško
  - Imamo samo instance za koje su zabeležena određena obeležja, a labele su nam nepoznate
  - a od nas se očekuje da ih podelimo u kategorije (ne moramo da znamo šta kategorije predstavljaju)

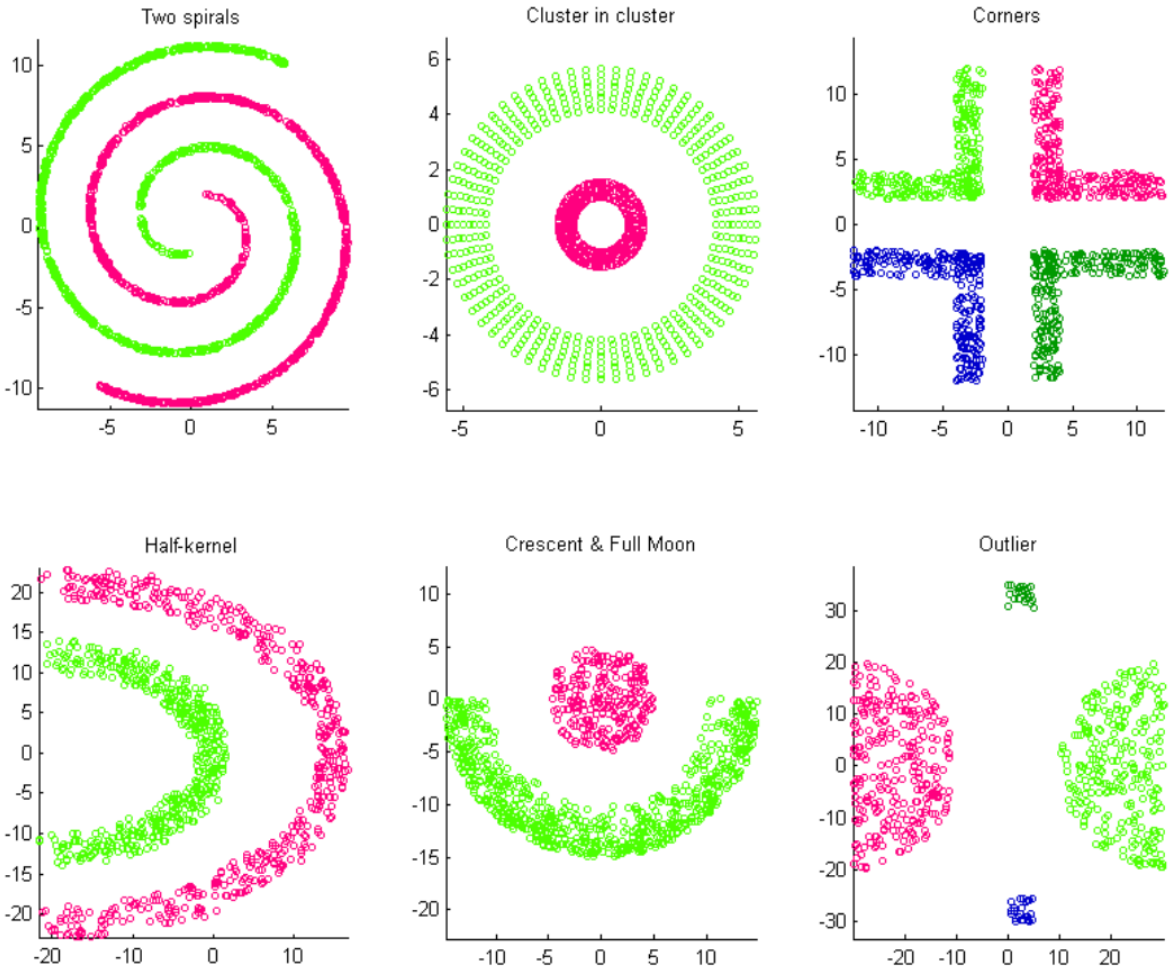
- Ipak, postoje dve stvari koje nam omogućavaju da ovo uradimo
  1. Struktura koja postoji u samim podacima
  2. Definicija šta je klaster (koju strukturu pokušavamo da pronađemo u podacima, npr. elipsoidni klasteri)

Struktura koja postoji u podacima



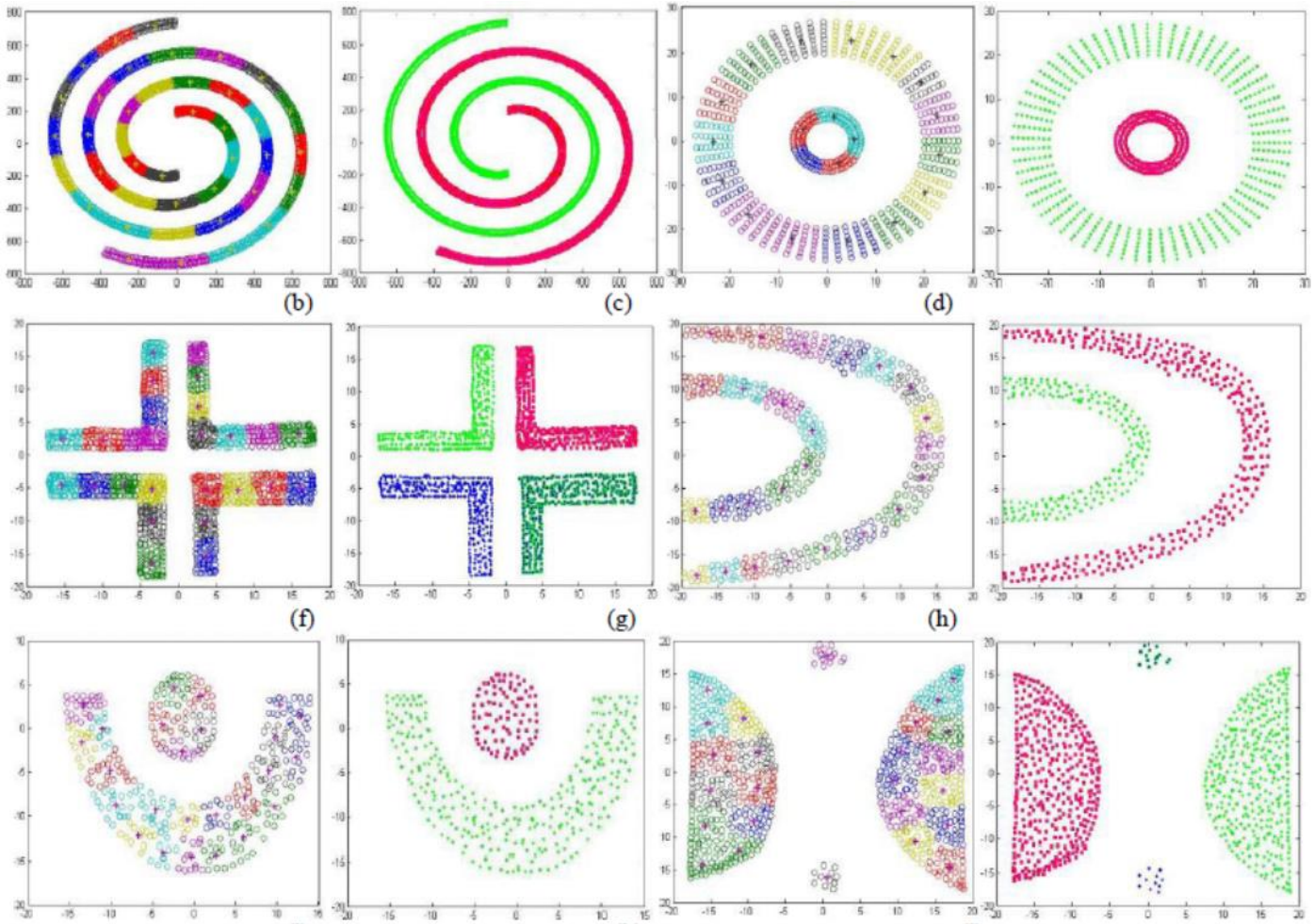
Definicija klastera

- Jako važno sa aspekta performansi jeste kako definišemo klaster



Definicija klastera

- Jako važno sa aspekta performansi jeste kako definišemo klaster



Primene klasterovanja

- Klasterovanje slika
  - Google image search

Pink flower

Ocean

Cat

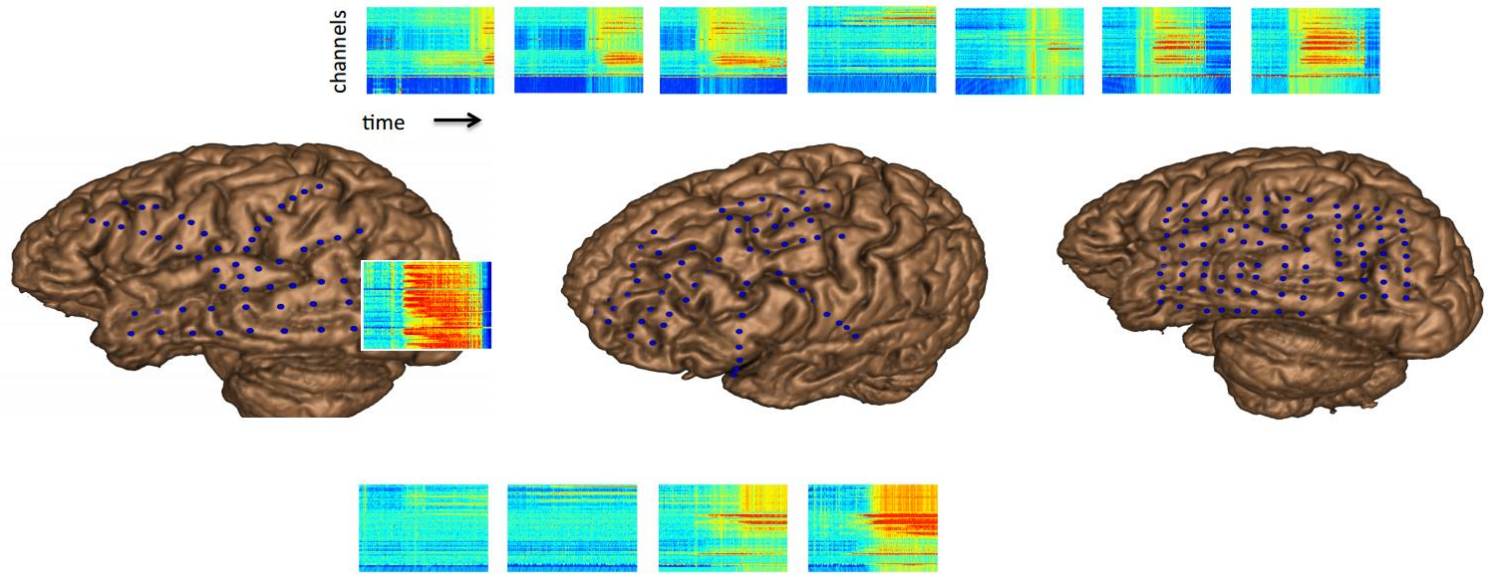
Sunset

Primene klasterovanja

- Strukturiranje rezultata pretrage
- Reči upita mogu imati više značenja
  - Npr. „cardinal“

Primene klasterovanja

- Grupisanje pacijenata prema medicinskom stanju
  - Bolja karakterizacija grupa u populaciji i bolesti
  - Primer: klasterizacija epileptičnih napada – možemo pregledati snimke toka događaja pri napadu odrediti tipove napada/tipove pacijenata prema sličnosti događaja



Primene klasterovanja

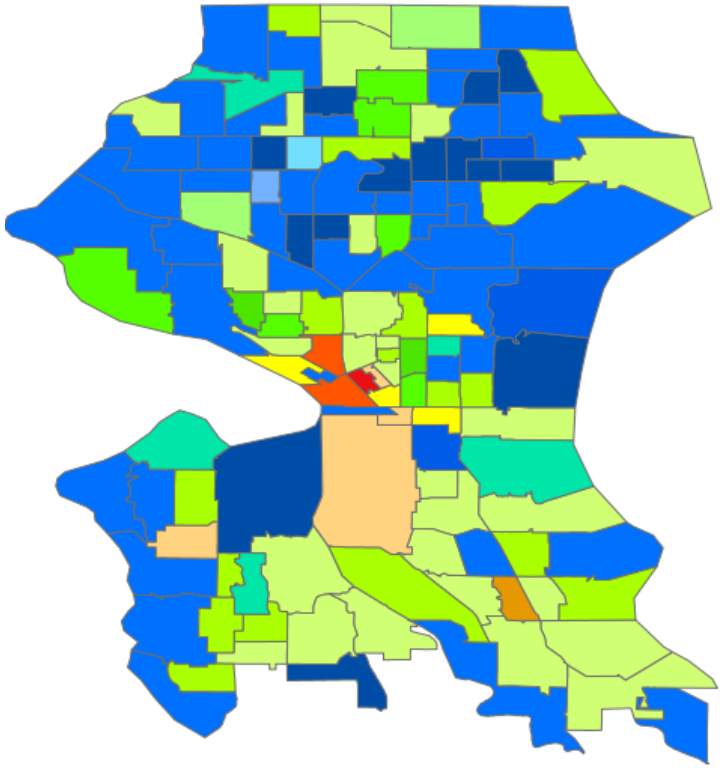
- Grupisanje proizvoda na Amazonu
  - Otkrivanje kategorija proizvoda na osnovu istorije kupovina
  - Otkrivanje grupa korisnika sa sličnim navikama u kupovini
  - Sistemi za preporuku

~~"furniture"~~

"baby"

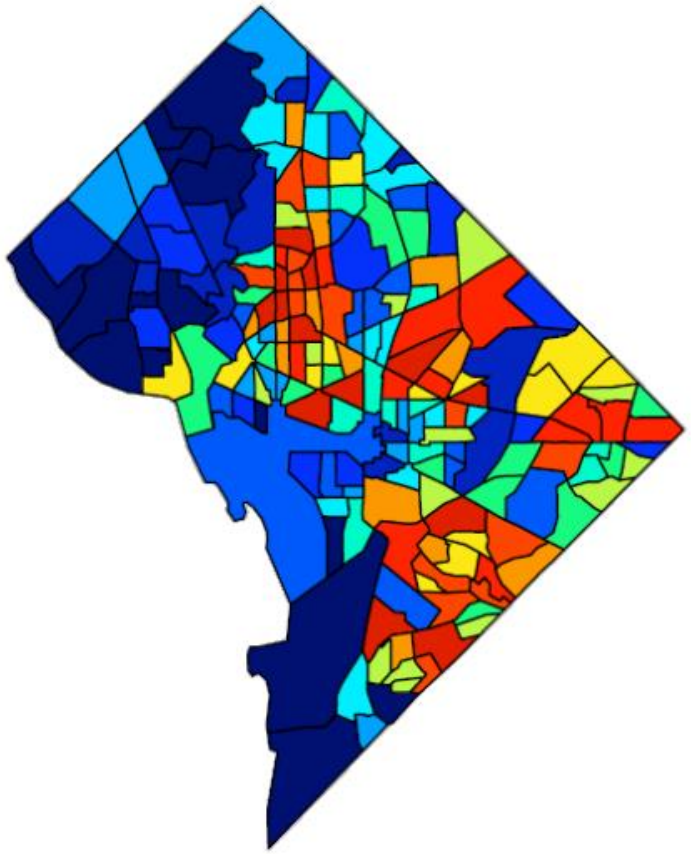
Primene klasterovanja

- Pronalaženje sličnih susedstava
- Zadatak 1: Proceniti cenu nekretnine na manjem regionalnom nivou
  - Izazov: Ima malo (ili nema) istorije prodaja u datom regionu
  - Rešenje: klasterovati regije na one sa istorijski sličnim trendom i iskoristiti informacije prilikom otkrivanja lokalnog trenda



Primene klasterovanja

- Pronalaženje sličnih susedstava
- Zadatak 2: Predviđanje zločina sa nasiljem radi bolje raspodele policije
  - Klasterovati regije i deliti informacije među njima
  - Dovodi do pobojšanja predikcija u poređenju sa nezavisnim ispitivanjem svake regije



Washington, DC