User's topological similarity

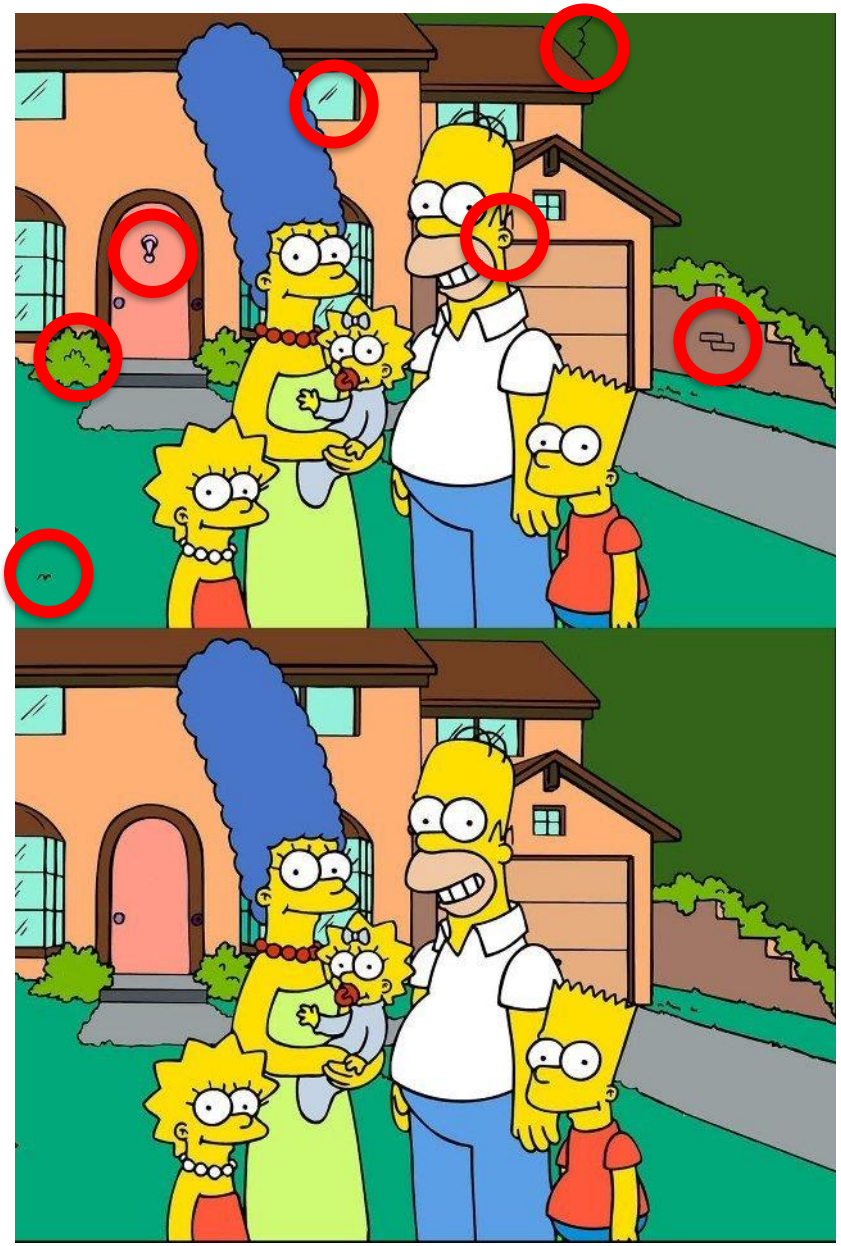
Serrano, J.[1]; Díez, F.[2]; Bellogin, A.[2]

[1] UIMP
Santander, España
serranopriego@posgrado.uimp.es

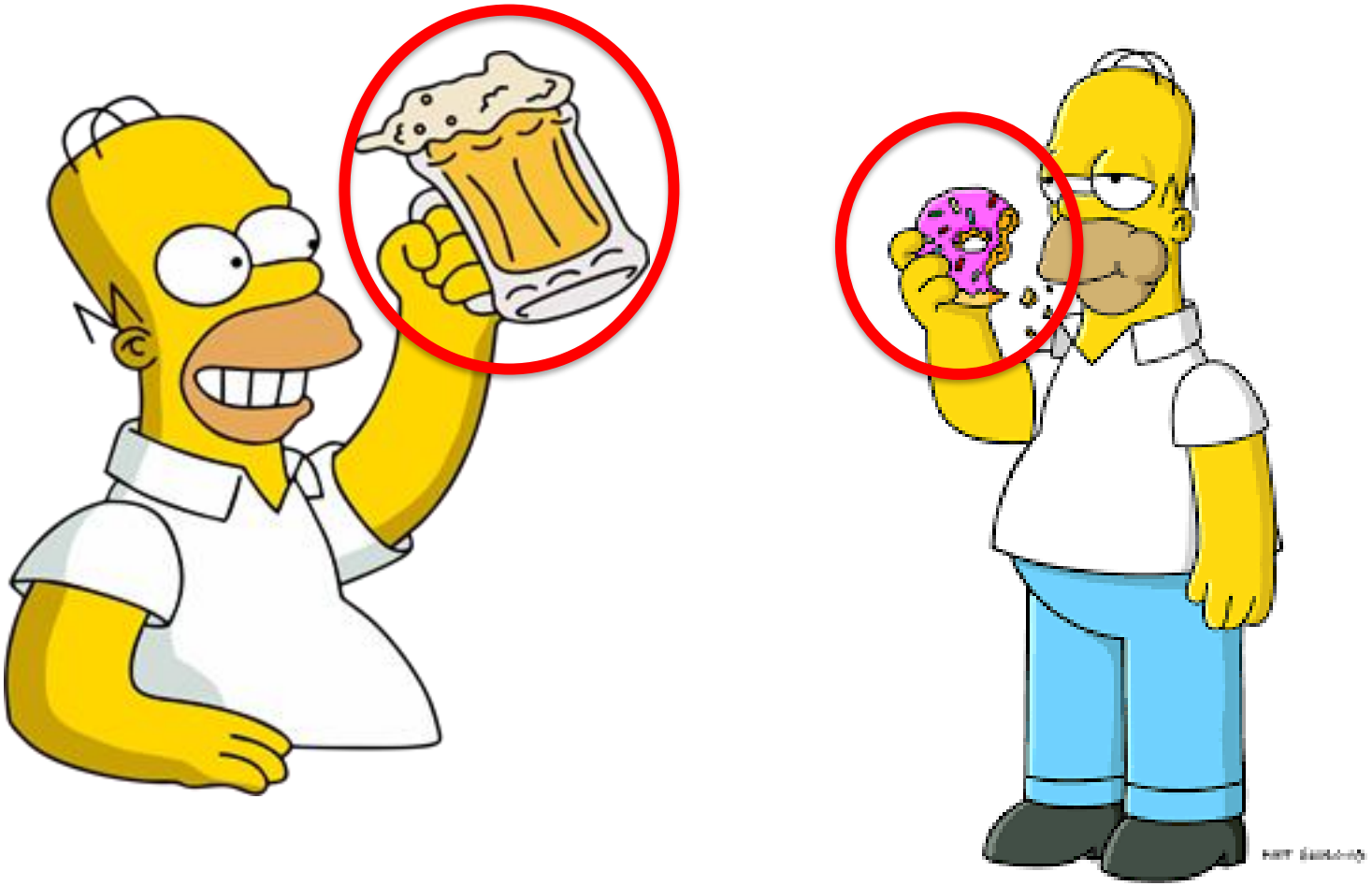
[2] Information Retrieval Group
Departamento de Ingeniería Informática
Universidad Autónoma de Madrid, Spain
fernando.diez@uam.es, alejandro.bellogin@uam.es

Let’s play

Find the 7 differences



Now … spot the similarity!

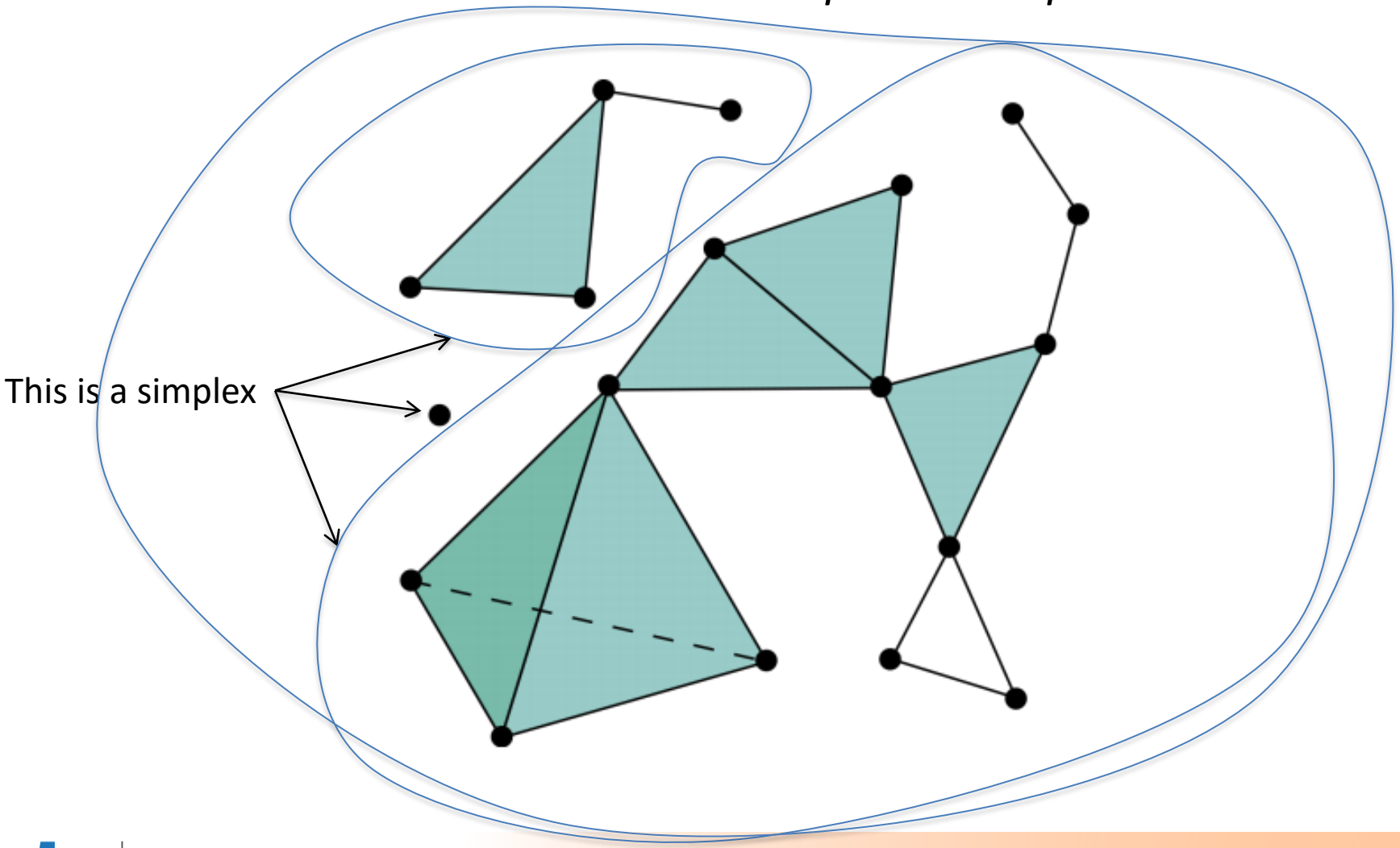


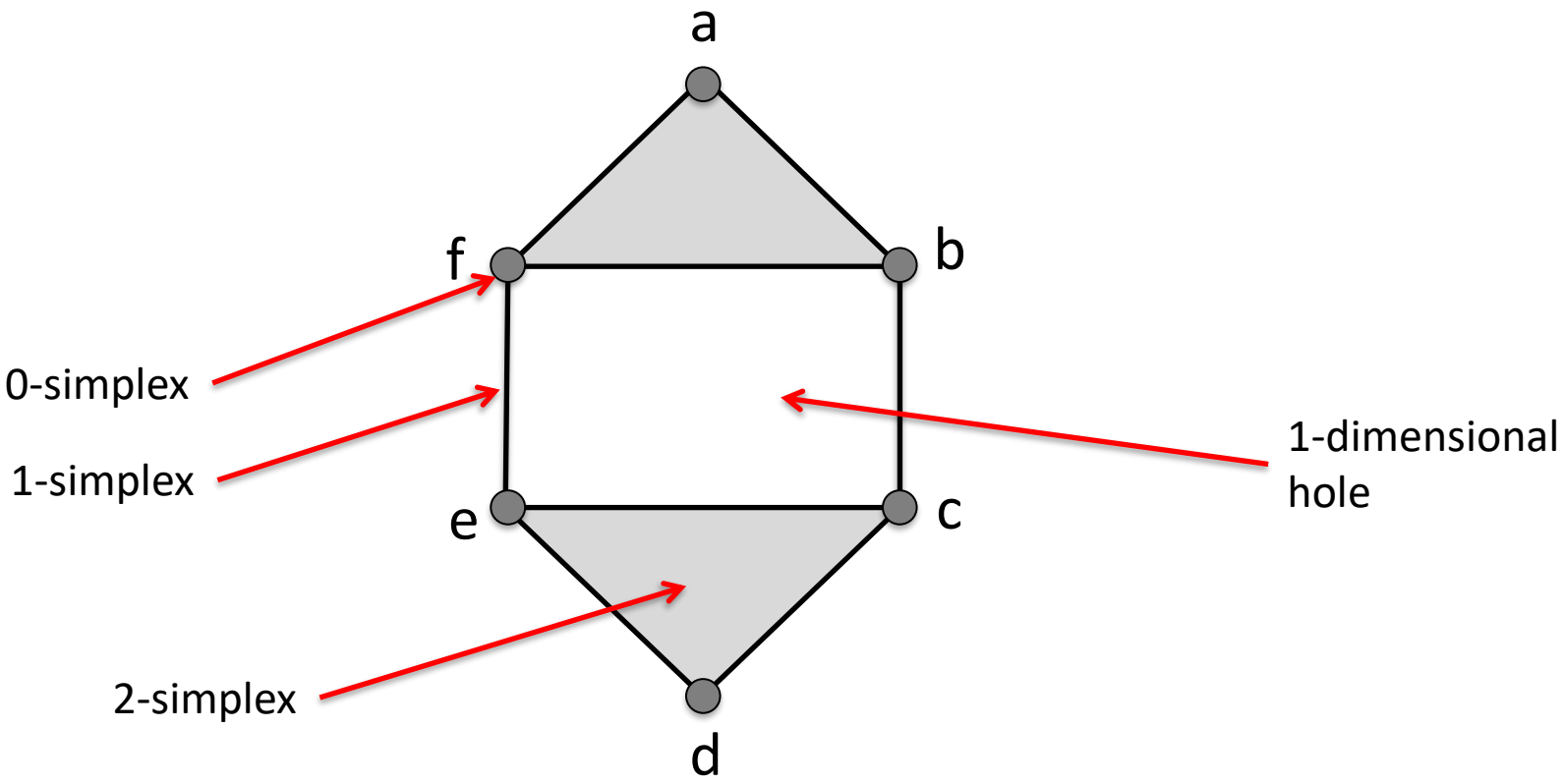
A cup and a doughnut are topologically equivalent!!

Basics

- This paper is about **topology** and **data** or **TDA** (Topological Data Analysis)
- Works on the understanding of the **shape of data**
- Recognizing typical shapes (**patterns**)
- Topology studies shapes that are **invariant** under small deformations

This is a *simplicial complex*



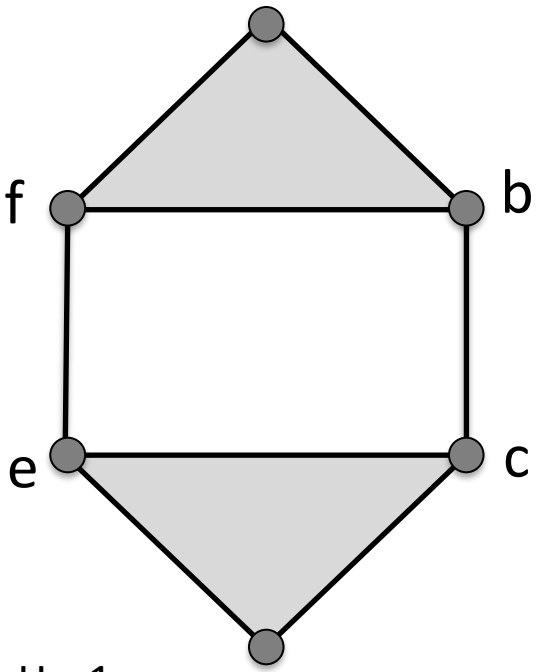


Betti numbers: characterizes simplicial complexes

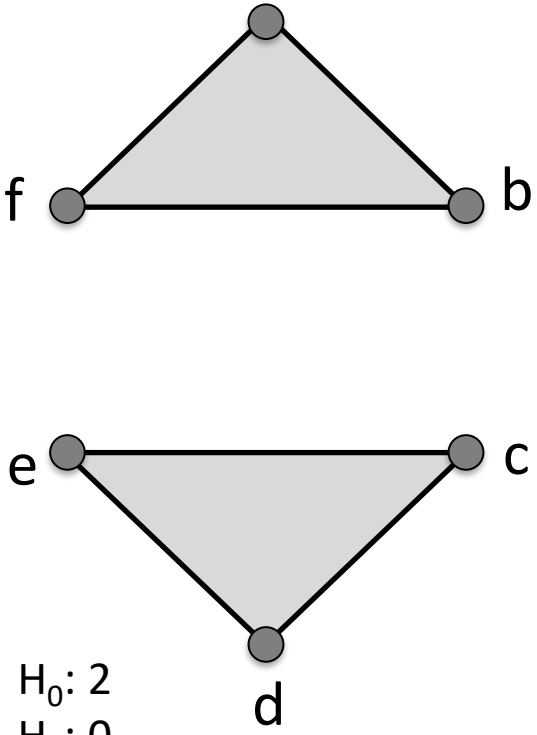
H_0: 1

H_1: 1

H_2: 0

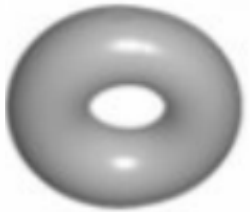


H_0: 1
H_1: 1
H_2: 0



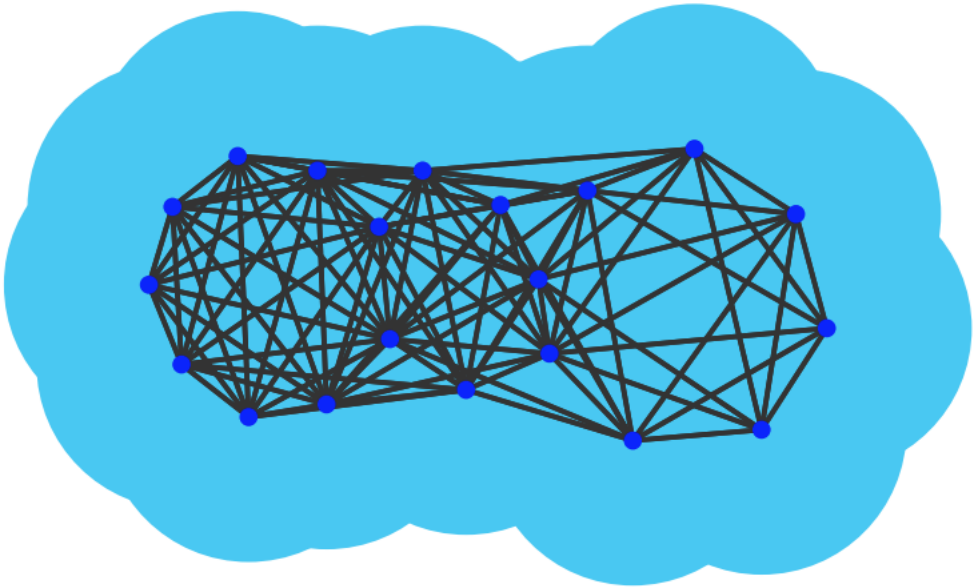
H_0: 2
H_1: 0
H_2: 0

H_0: 1
H_1: 1
H_2: 0



H_0: 1
H_1: 1
H_2: 0

Vietoris-Rips Complex computation

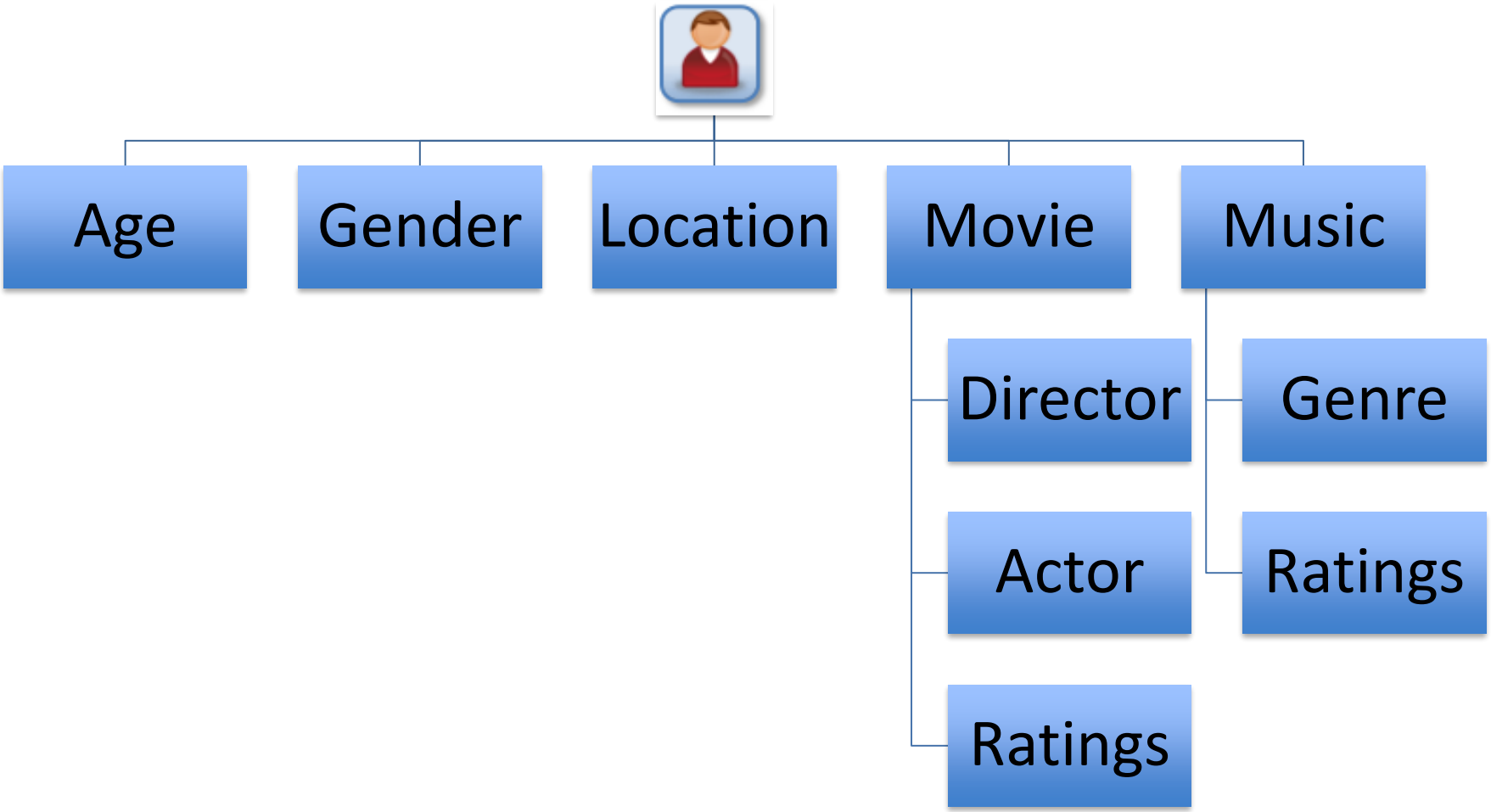


Vietoris-Rips complex characterizes the topology of a point set

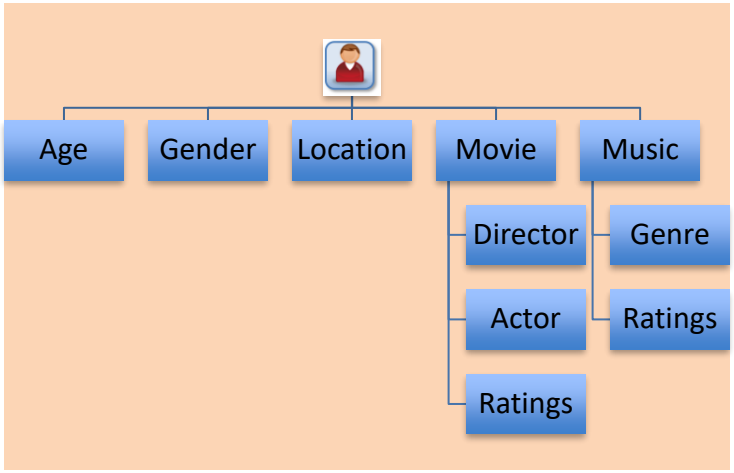
From users to barcodes

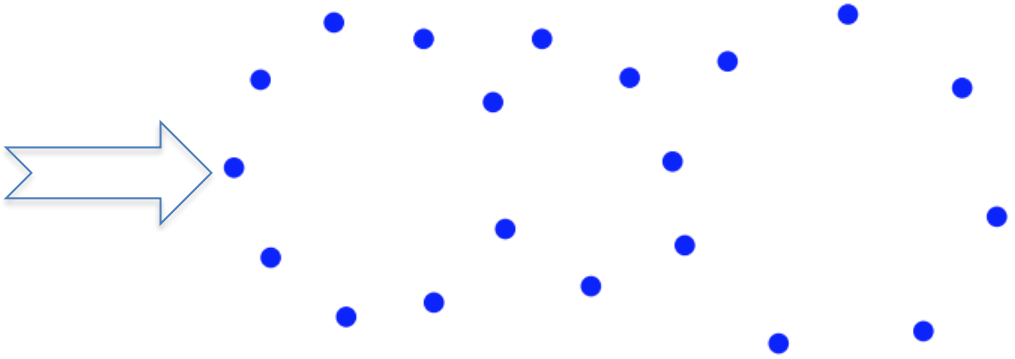
BARCODE BASED SIMILARITIES

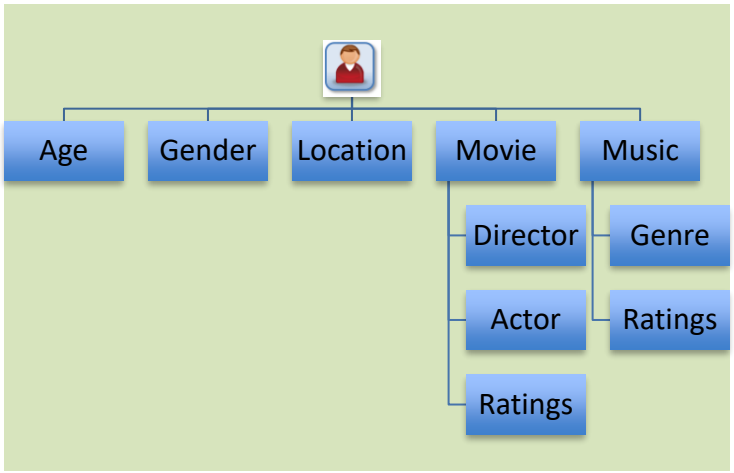
From users to barcodes

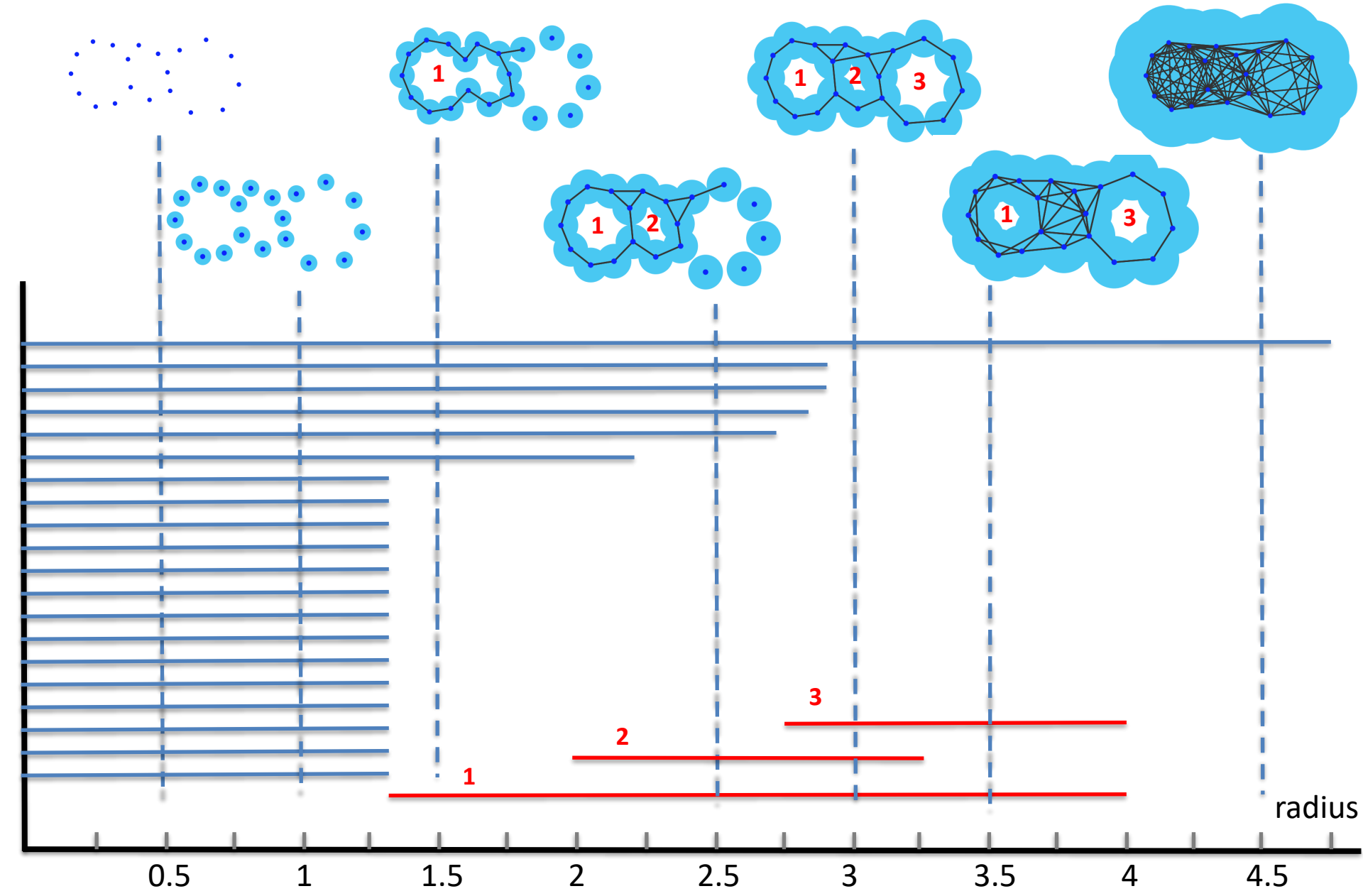


User's topologies







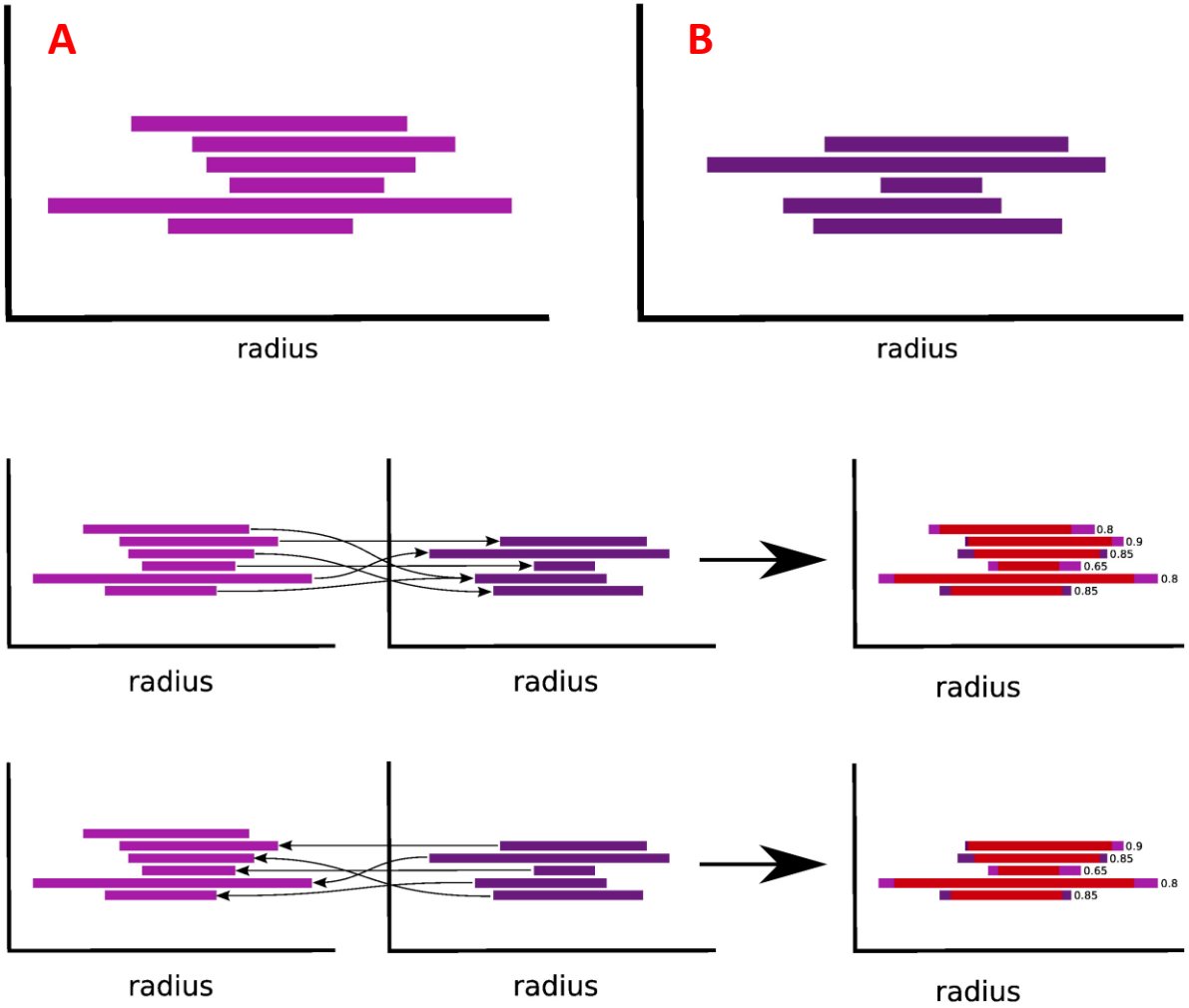

1
1
2
3
1
2
1
3
3
2
1
radius
0.5
1
1.5
2
2.5
3
3.5
4
4.5

Barcode **O**verlapping (**E**xtended) similarity metric

$$S_{BO}(A,B)=\frac{1}{|A|+|B|}\left(\sum_{a\in A}\sup_{b\in B}\frac{a\cap b}{a\cup b}+\sum_{b\in B}\sup_{a\in A}\frac{a\cap b}{a\cup b}\right)$$

$$S_{BOE}=\begin{cases}S_{BO}(A,B) & A\neq\emptyset, B\neq\emptyset\\ 1 & A=\emptyset, B=\emptyset\\ 0 & any\ other\ case\end{cases}$$

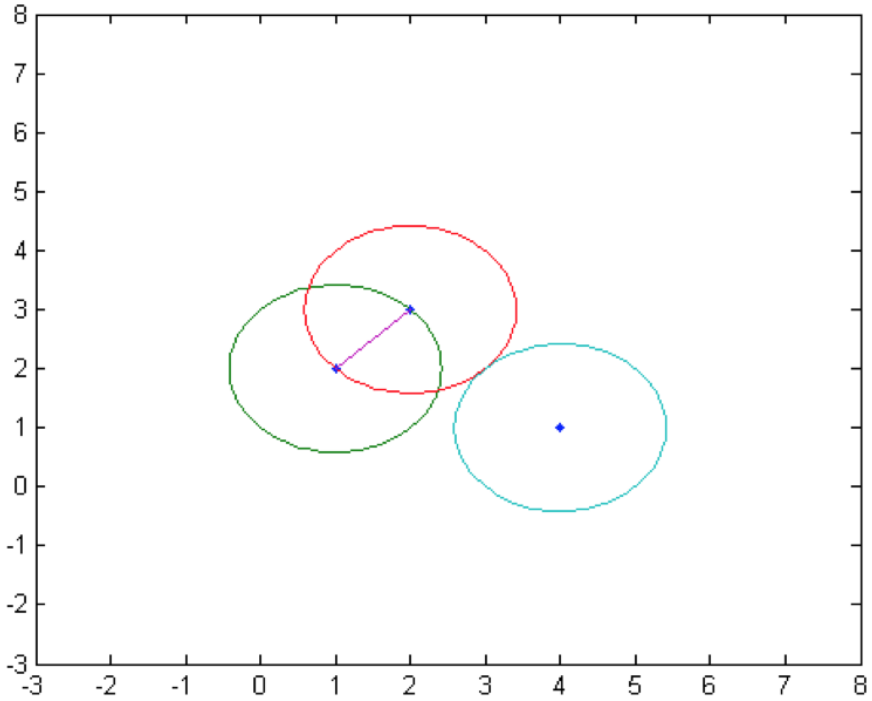
User's barcode comparisons S_{BO} similarity computation



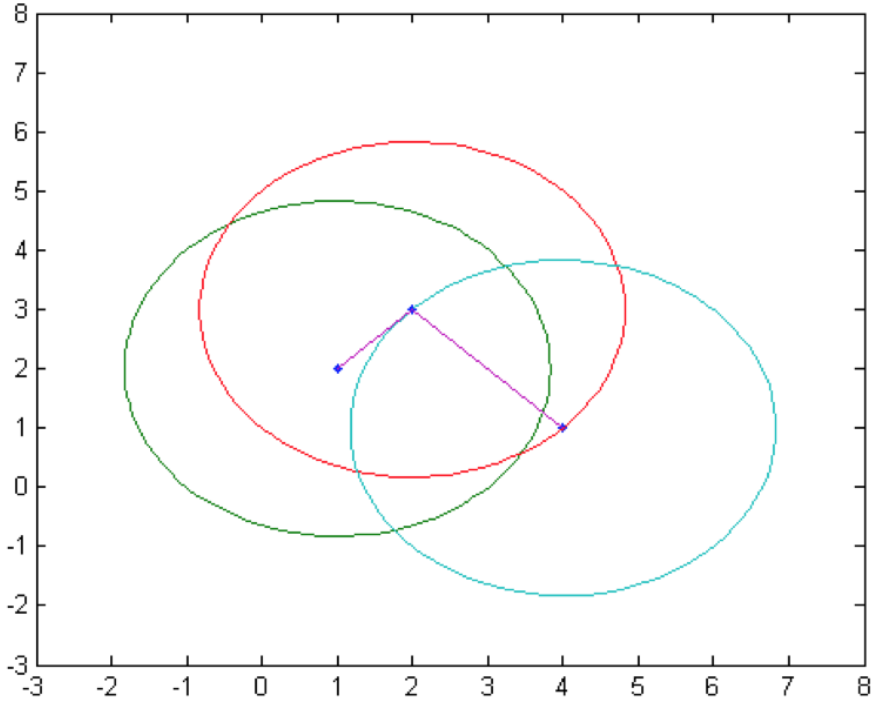
Experiments

- Dataset: Movilens 100K
  - 943 users. 1682 items
  - Users represented by means of pairs `(item,value)`
  - # user ratings < 50
- Javaplex library
  - To compute invariants as well as similarities among users' barcodes
- Standard kNN algorithm

Users' profiles simplicial complexes

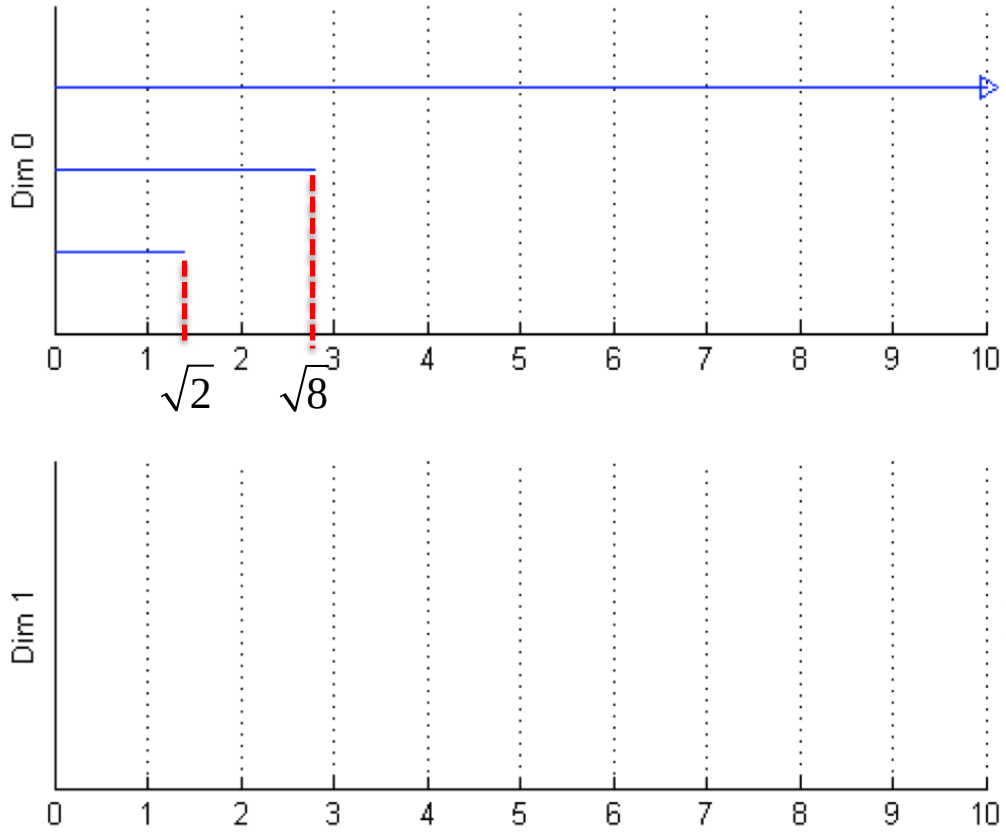


$r = \sqrt{2}$



$r = \sqrt{8}$

Associated users' barcodes computed



No holes

Results

Similarity	S_{BOE} 50 20	S_{BOE} 100 15	Cosine	Pearson	Random
RMSE	1.1616	1.1482	**1.0791**	1.5728	1.1217

Conclusions and future work

- Competitive behaviour against standard Cosine or Pearson similarities
- *Cold start* problem
- Accomplish new experiments using different datasets, graph based algorithms, optimization, etc.
- Improve the computation of barcode similarities, highly dependant on dataset size

Merci, Grazie, Gracias, Thank you!