

# Similarity searching

Filippo Geraci

May 14, 2019

# Similarity search - Motivations

## [Clustering - Wikipedia](#)

Il **Clustering** o analisi dei **cluster** o analisi di raggruppamento è un insieme di tecniche di analisi multivariata dei dati volte alla selezione e ...

[it.wikipedia.org/wiki/Clustering](http://it.wikipedia.org/wiki/Clustering) - 29k - [Copia cache](#) - [Pagine simili](#)

## Scenario

- Text retrieval systems provide a *similarity search utility*, that allows users to find efficiently documents that are the *most similar* to the query.

### Active bibliography (related documents): [More](#) [All](#)

- 1.2:** [The Zebra Striped Network File System - Hartman, Ousterhout \(1993\)](#) [\(Correct\)](#)
- 0.5:** [HYDRANET-FT: Network Support for Dependable Services - Shenoy, Satapati, Bettati \(2000\)](#) [\(Correct\)](#)
- 0.3:** [The Logical Disk: A New Approach to . . . - de Jonge, Kaashoek, Hsieh](#) [\(Correct\)](#)

### Similar documents based on text: [More](#) [All](#)

- 0.7:** [Zebra: A Striped Network File System - Hartman, Ousterhout \(1993\)](#) [\(Correct\)](#)
- 0.5:** [Zebra Zebra Zebra Zebra Zebra Zebra Zebra Zebra Zebra.. - Overview Of](#) [\(Correct\)](#)
- 0.3:** [Zebra-crossing Detection for the Partially Sighted - Se \(2000\)](#) [\(Correct\)](#)

### Related documents from co-citation: [More](#) [All](#)

- 30:** [The design and implementation of a log-structured file system - Rosenblum, Ousterhout - 1991](#)
- 28:** [Scale and Performance in a Distributed File System \(context\) - Howard, Kazar - 1988](#)
- 28:** [A Case for Redundant Arrays of Inexpensive Disks \(context\) - Patterson, Gibson et al. - 1988](#)

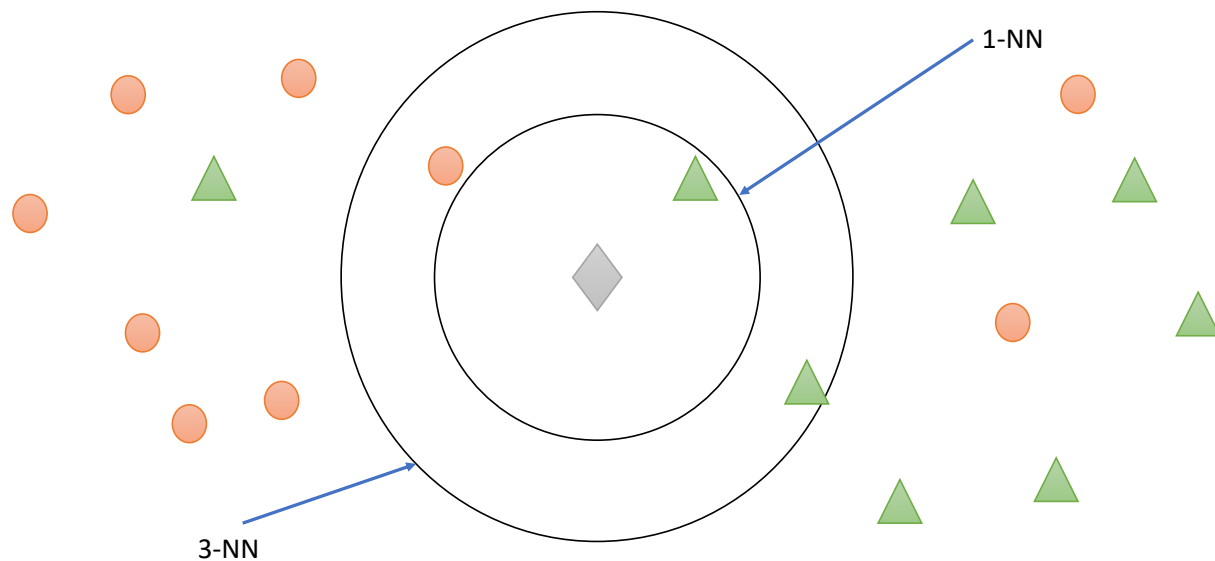
## The problem

- 1 Find the  $k$  elements more similar to a given query
- 2 In absence of information all the pairwise distances must be computed

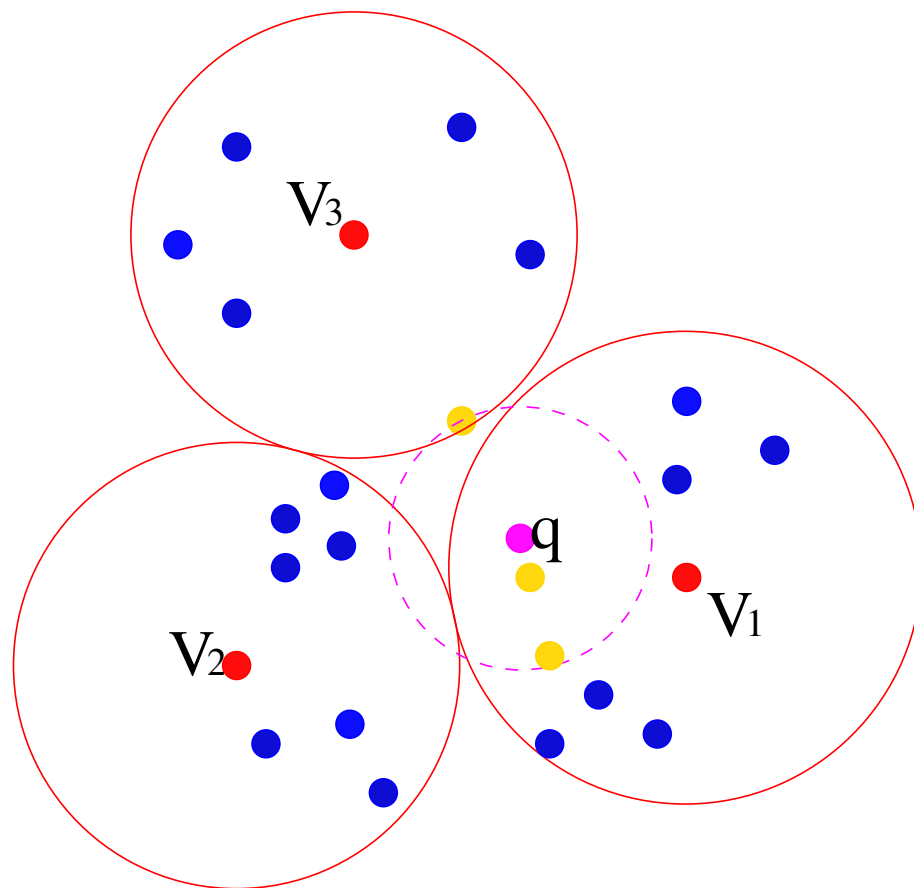
## Caveats

- 1 The database can be typically huge.
- 2 A faster and approximate answer might be acceptable.
- 3 The system evolves in the time (new items can be added to the database).
- 4 A single distance computation can be expensive for certain data (i.e. texts)

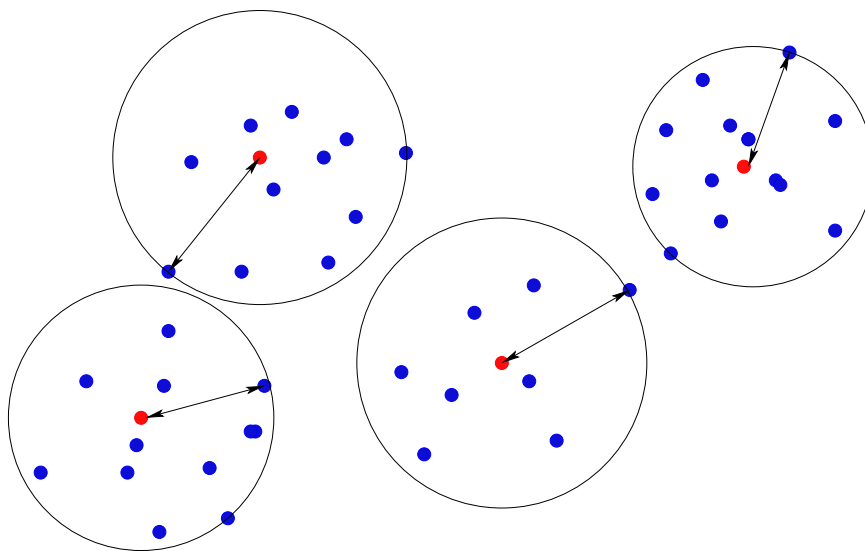
# Similarity searching and classification



# An approximate solution with clustering



# Analogies with the $k$ -center problem



- The closer a query is to a cluster center the more probable is that its nearest neighbors are in the cluster.
- Cluster radius bounds the distance among the query and its nearest neighbors in the cluster.