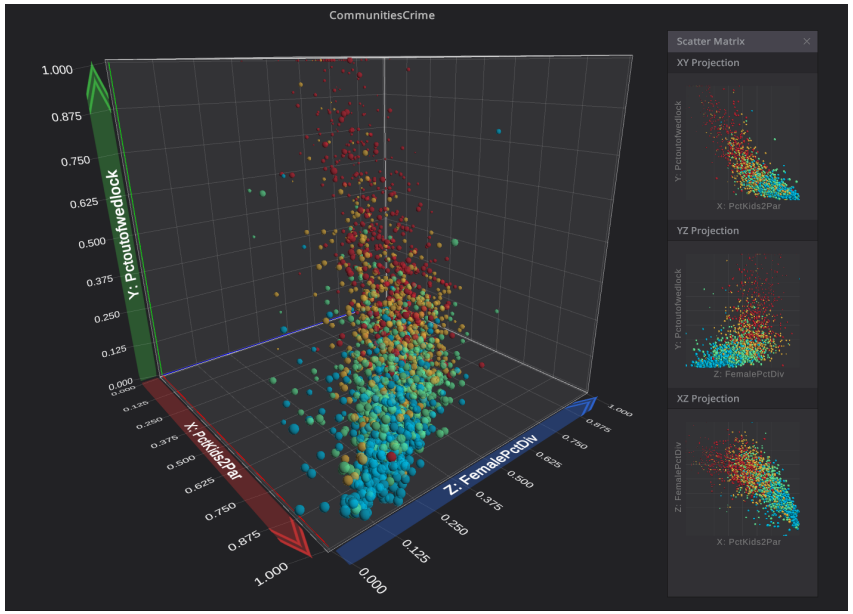
Title of the Lecture: *Finding Similar Items in Large Data Sets*

Who: Andrea Clementi (Dipartimento Ingegneria dell'Impresa, UniTV – clementi@mat.uniroma2.it)

When and Where: 19.05.2022, 11h00-13h00 - Aula G2B (Sogene)



Syllabus of the Lecture

- *Universal* Family of *Hash* Functions
- Modelling *Items* (e.g. *Docs*) as *Points* of *High-Dimensional Metric Spaces*
- *Similarity/Distance* among *Subsets*: *Jaccard* Similarity
- Convert *Items* to *Subsets*: *Shingling* Technique
- Convert large *Subsets* to short *Signatures* (while preserving J. Similarity): *Min-Hashing*
- Finding candidate for *similar Docs pairs*: *Locally-Sensitive Hashing*