

# **Content-based Similarity Search of Large Feature-rich Datasets**

Moses Charikar  
Princeton University

## **Abstract**

Digital data has been increasing at a phenomenal rate. Most of this data is not in text form, e.g. images, sounds and video. The utility of this data would be vastly increased by good tools to search and organize it.

However, currently available search tools are inherently text based and depend on text annotations for such feature-rich datasets. Performing content-based similarity search on such high dimensional data is a challenging problem. We have developed efficient general-purpose methods to search and index such datasets, building on recent theoretical work on constructing sketches - compact representations for data.

These methods have been implemented in a toolkit designed to help system builders quickly construct content-based similarity search systems for a wide variety of such datasets. In this talk, I will describe this toolkit, some of the search and indexing technology it incorporates, and our experiences in using it to build similarity search systems for four different data types: images, audio recordings, 3D shape models and genomic micro array data.