Content-based Similarity Search of Large Feature-rich Datasets

Moses Charikar
Princeton University

Project Members:

Moses Charikar, Wei Dong, William Josephson,
Kai Li, Christine (Qin) Lv, Zhe Wang

Princeton Collaborators:

Perry Cook, Matt Hibbs, Matthew Hoffman,
Haakon Larsen, Emily Pronin, Jennifer Rexford, Olga Troyanskaya

http://www.cs.princeton.edu/cass/

Motivations

- Data has been growing exponentially
  - search is crucial
- Non-text data dominates data volume
  - Audio, video, digital photos, scientific sensor data
  - Noisy and high-dimensional
- Current tools are inadequate
  - limited to text based search

german shepherd - Google Image Search - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://images.google.com/images?svnum=10&hl=en&lr=&q=german+shepherd Go Links

Google german shepherd Go Bookmarks 1 blocked Check AutoLink AutoFill Send to Settings

Sign in

Google Images

Web **Images** Groups News Froogle Local Scholar **more »**

german shepherd Search Advanced Image Search Preferences

Moderate SafeSearch is on

Images Showing: All image sizes

Results **1** - **20** of about **71,800** for **german shepherd** [definition]. (**0.08** seconds)

... made To Fit **German Shepherd** Dog
300 x 300 pixels - 27k - jpg
www.fordogtrainers.com

German Shepherd, White
180 x 251 pixels - 25k - jpg
www.puppypurebred.com

In bringing a **German Shepherd** into ...
327 x 291 pixels - 21k - jpg
www.hohenwaldvet.com

german-shepherd-for-sale-kiara2.jpg ...
300 x 236 pixels - 13k - jpg
www.vomfieldcrestshepherds.com

German Shepherd Puppy in a Box
499 x 333 pixels - 19k - jpg
www.livingwilderness.com

I ve got a **german shepherd** dog now! ...
400 x 300 pixels - 18k - jpg
www.excess-luggage.com

Our **German Shepherd** Chloe
269 x 202 pixels - 20k - jpg
www.just-german-shepherds.com

German Shepherd pups
448 x 646 pixels - 29k - jpg
pics.hoobly.com

Done Internet

start german shep... Inbox - Micro... 2 Microsoft... My Documents 10:28 AM

Similarity Search Problem

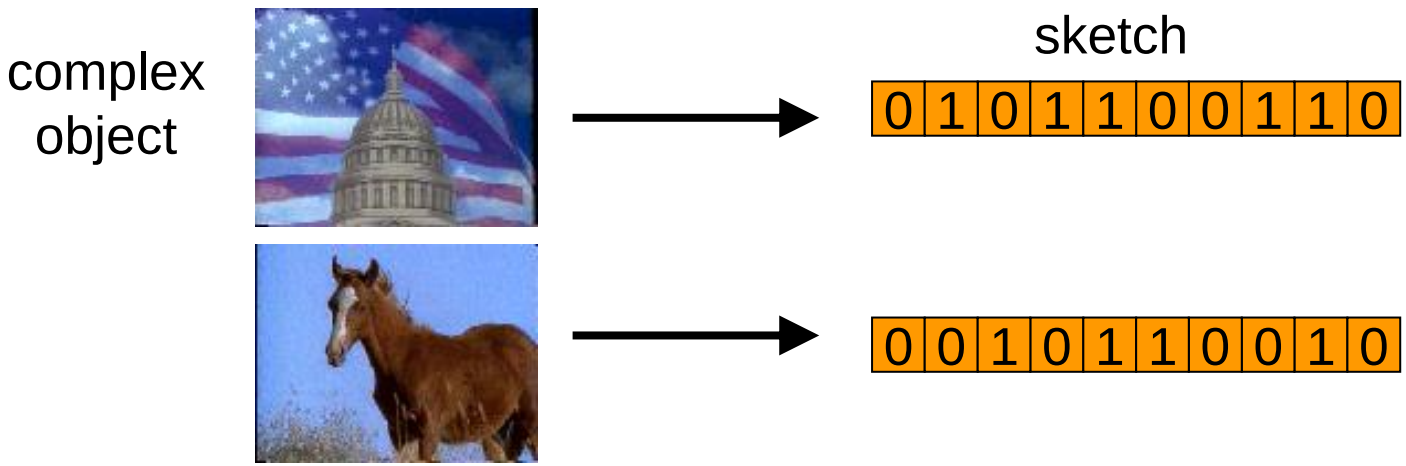
- Similarity search
  - Given a query object, finding similar objects (containing similar features)
- Distance function d(X, Y) between objects
  - K-nearest neighbor (KNN)
  - Approximate nearest neighbor (ANN)
- Challenges
  - Mathematical approximation of human perception of similarity
  - High-dimensional search is limited by “curse of dimensionality”

Content-Aware Search Systems Project

- How to extract information from data?
  - Segmentation and feature extraction
  - Leverage domain knowledge
- How to build scalable content-based search systems?
  - Similarity match instead of exact match
  - Novel data structures to overcome “Curse of dimensionality”
- Initial goals
  - Study compact data structures for high-quality and efficient content-based similarity search
  - Leverage recent progress of theoretical studies on sketches
  - Build a toolkit for research and search tool construction

Sketching Paradigm

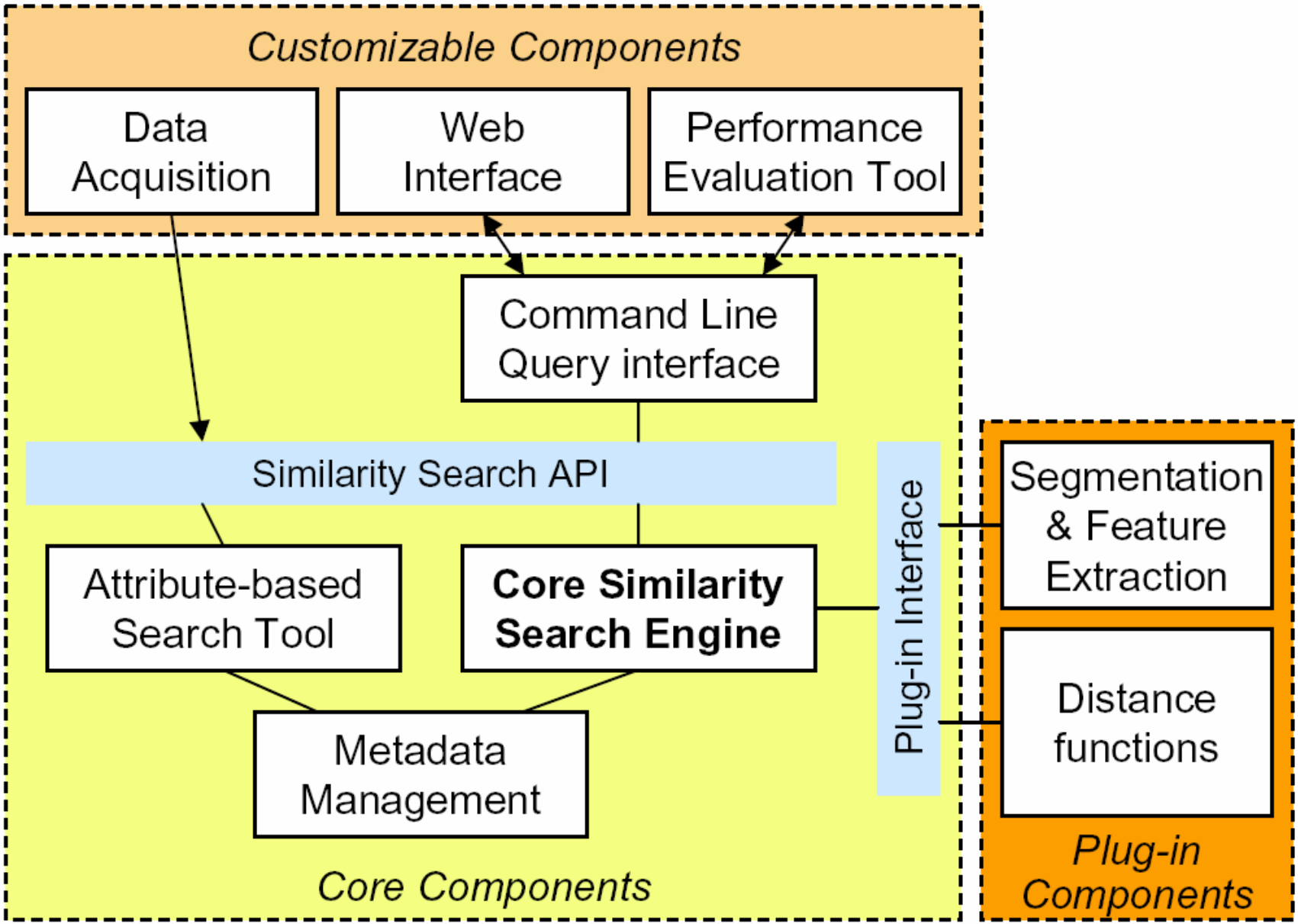
- Construct compact representation (sketch) of data such that
- Interesting functions of data can be ~~computed~~ estimated from compact representation



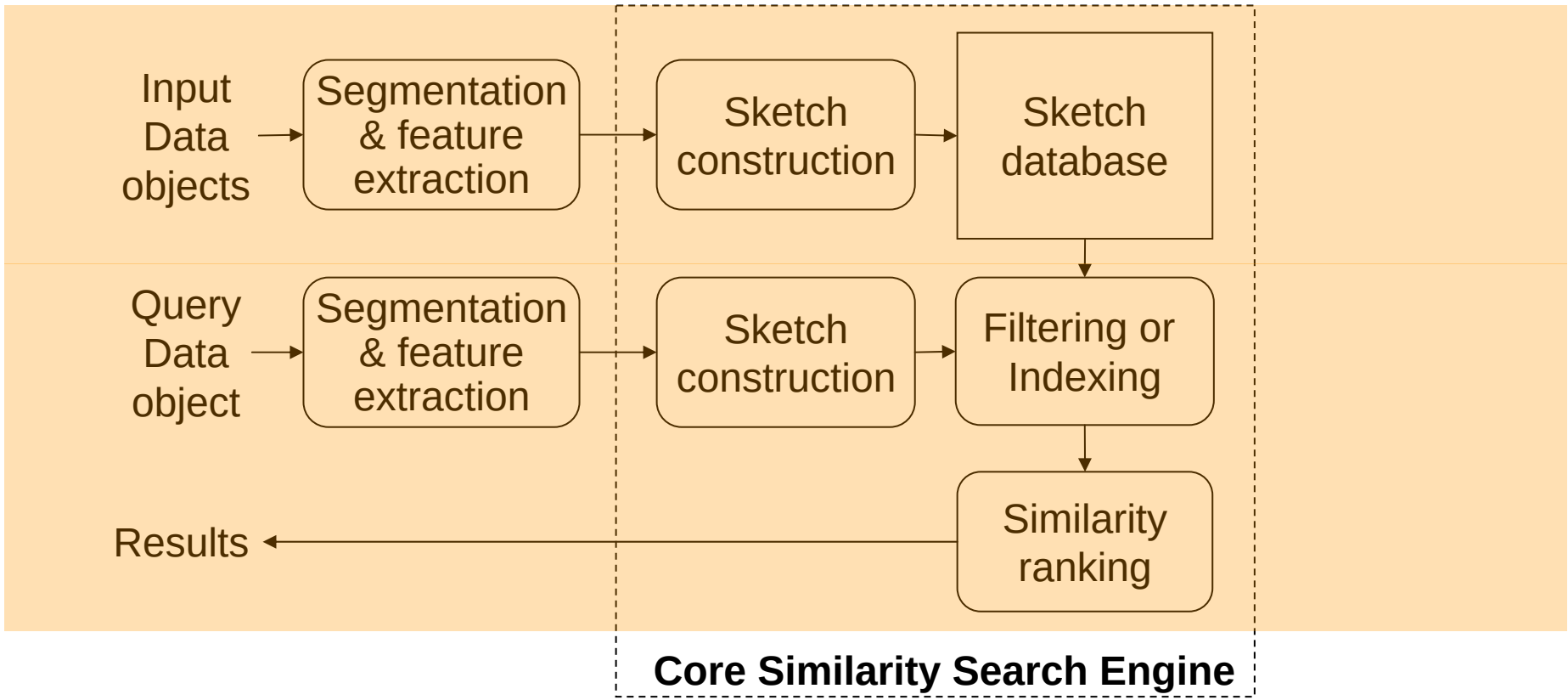
Project Status

- Ferret: a toolkit for content-based similarity search
  - Focus on handling large datasets and multiple data types
  - Software platform for research and collaboration
- Sketch construction
  - Reduce feature vector sizes by an order of magnitude
  - Experimented with 4 data types (images, audio, 3-D shapes and genomic data)
- Filtering and indexing
  - Sketch-based filtering
  - Locality sensitive hashing for indexing
    - Substantial improvement over previous proposals

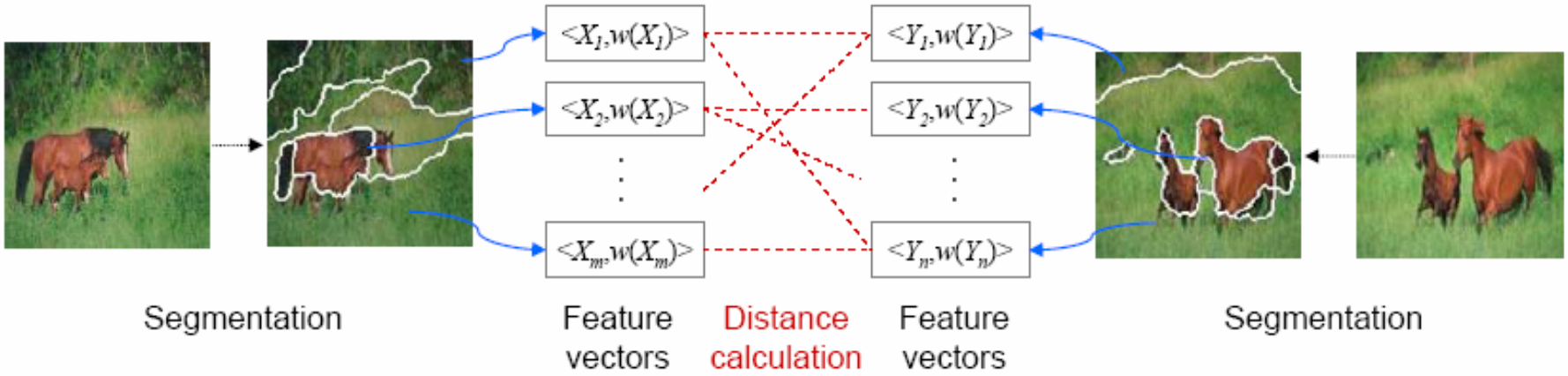
Ferret: A Content-Based Similarity Search Toolkit

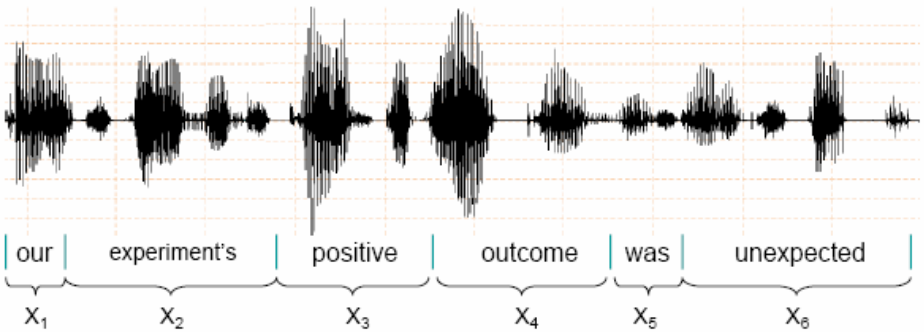


Core Similarity Search Engine



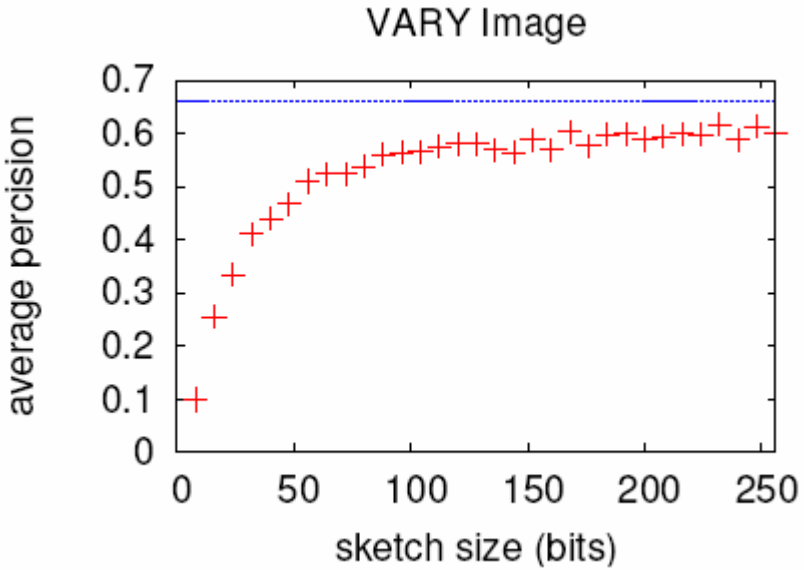
Object Representation & Distance Function

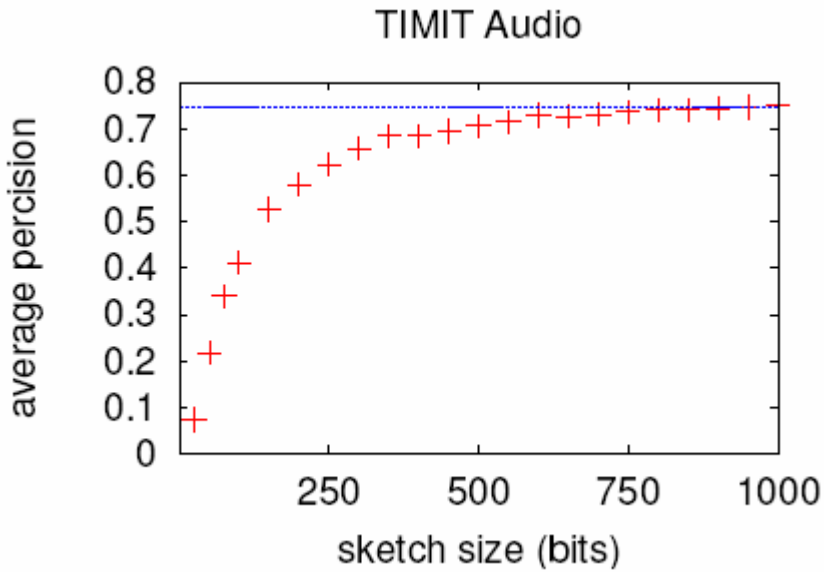




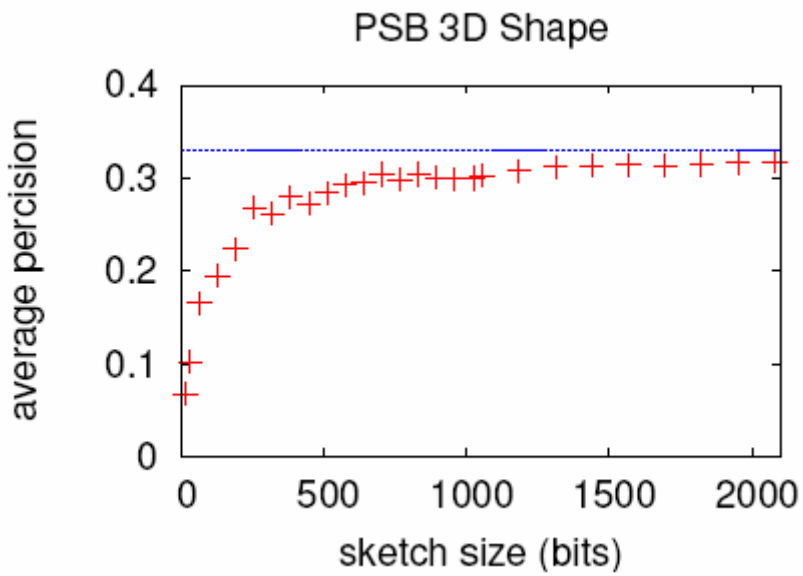
- Multi-feature representation
- Distance function
  - E.g. Earth Mover's Distance (EMD)
- Sketching techniques

Search Quality vs. Sketch Size





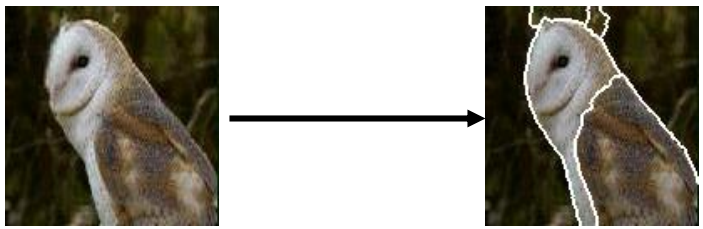
	Sketch Size	Sketch Size
VARY Image	64 bits (7:1)	88 bits (5:1)
TIMIT Audio	250 bits (6:1)	450 bits (3:1)
PSB 3D Shape	200 bits (87:1)	600 bits (29:1)



Multi-Feature Filtering

- Two-step query processing
  - Filtering step
    - Quickly filters out “bad” answers
    - Generates a small candidate set
  - Similarity ranking step
    - Uses multi-feature object distance function
    - Computes distance for each candidate object
    - Returns *k* nearest objects
- Criteria in picking candidate objects
  - Has at least one segment that is close enough to one of the top segments of the query object

Multi-Feature Filtering



query image

image segments

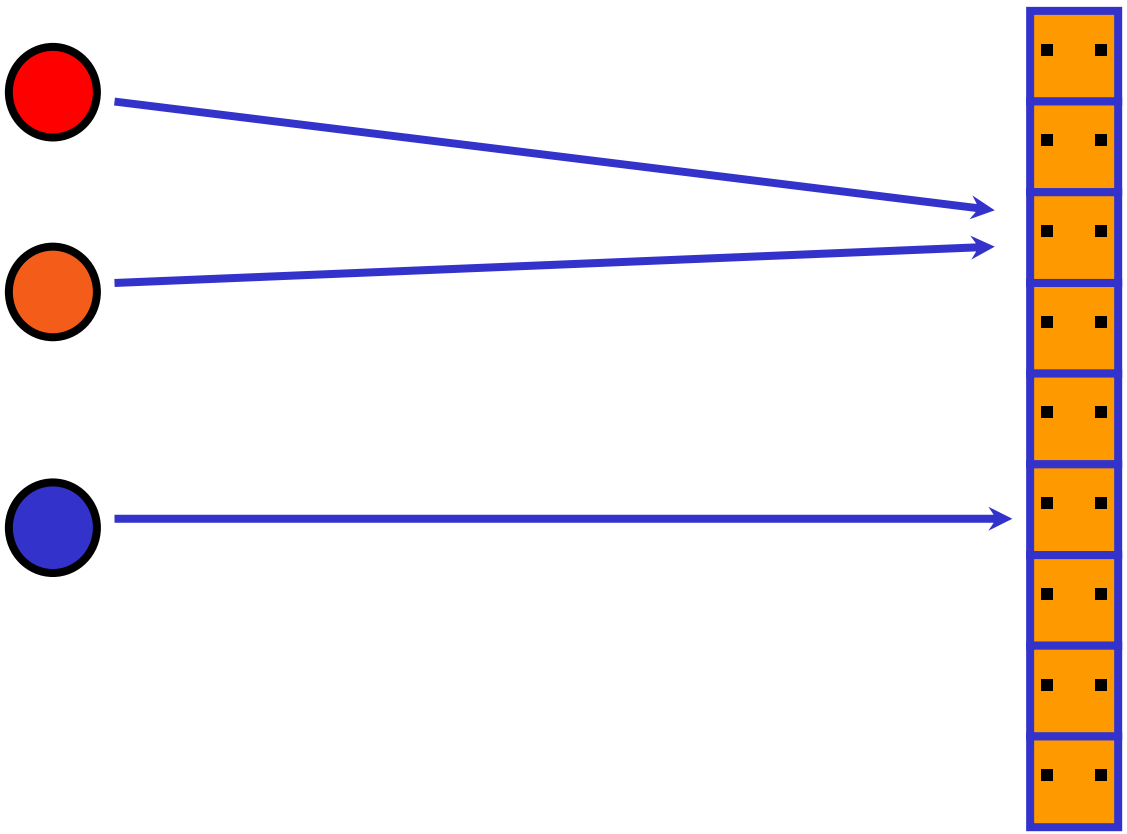
r query segments

n similar segments each

Similarity Ranking

candidate images

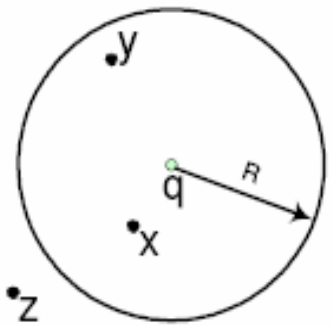
Locality Sensitive Hashing (LSH)



Similar objects more likely to hash to the same bucket

LSH based indexing

- [IM98, GIM99, DIIM04]
- *M* hash functions per table $g_i(v) = (h_{i,1}(v), \cdots, h_{i,M}(v))$
- *L* hash tables $G = \{g_1, \cdots, g_L\}$



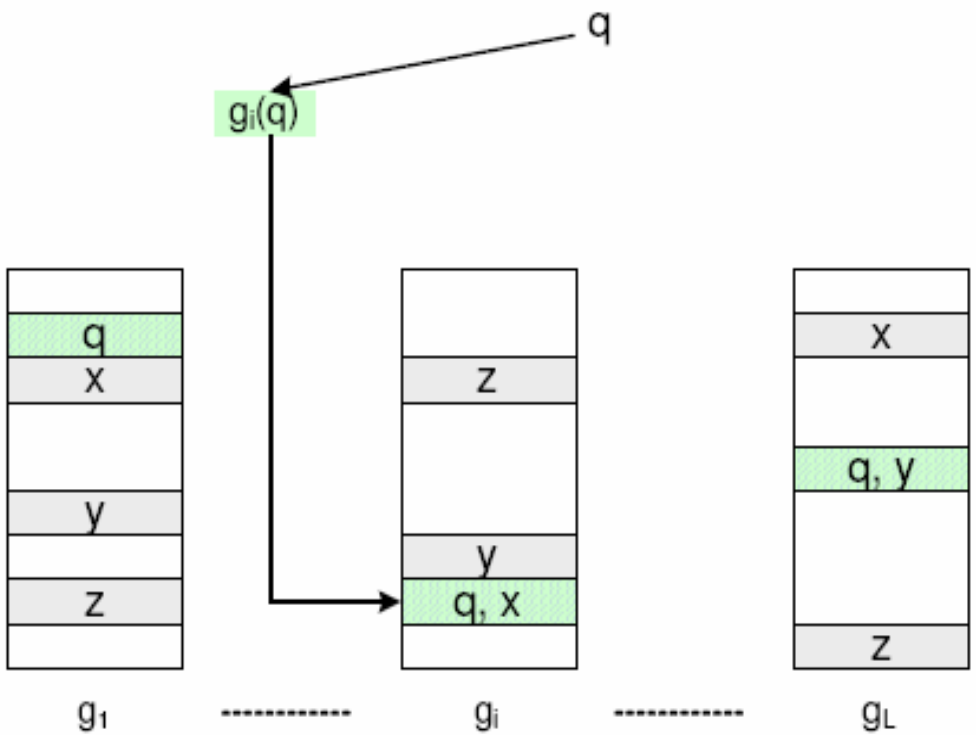
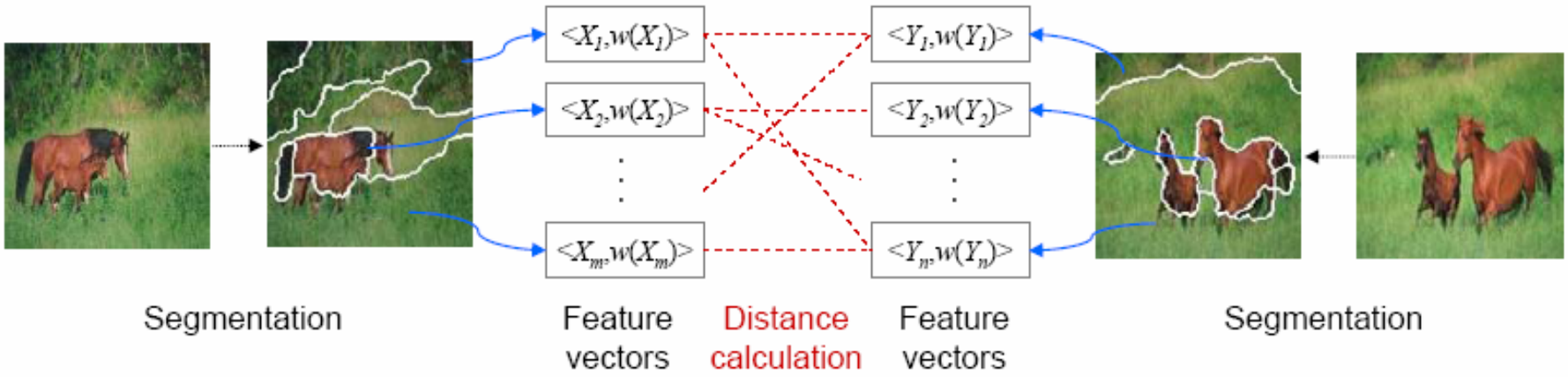
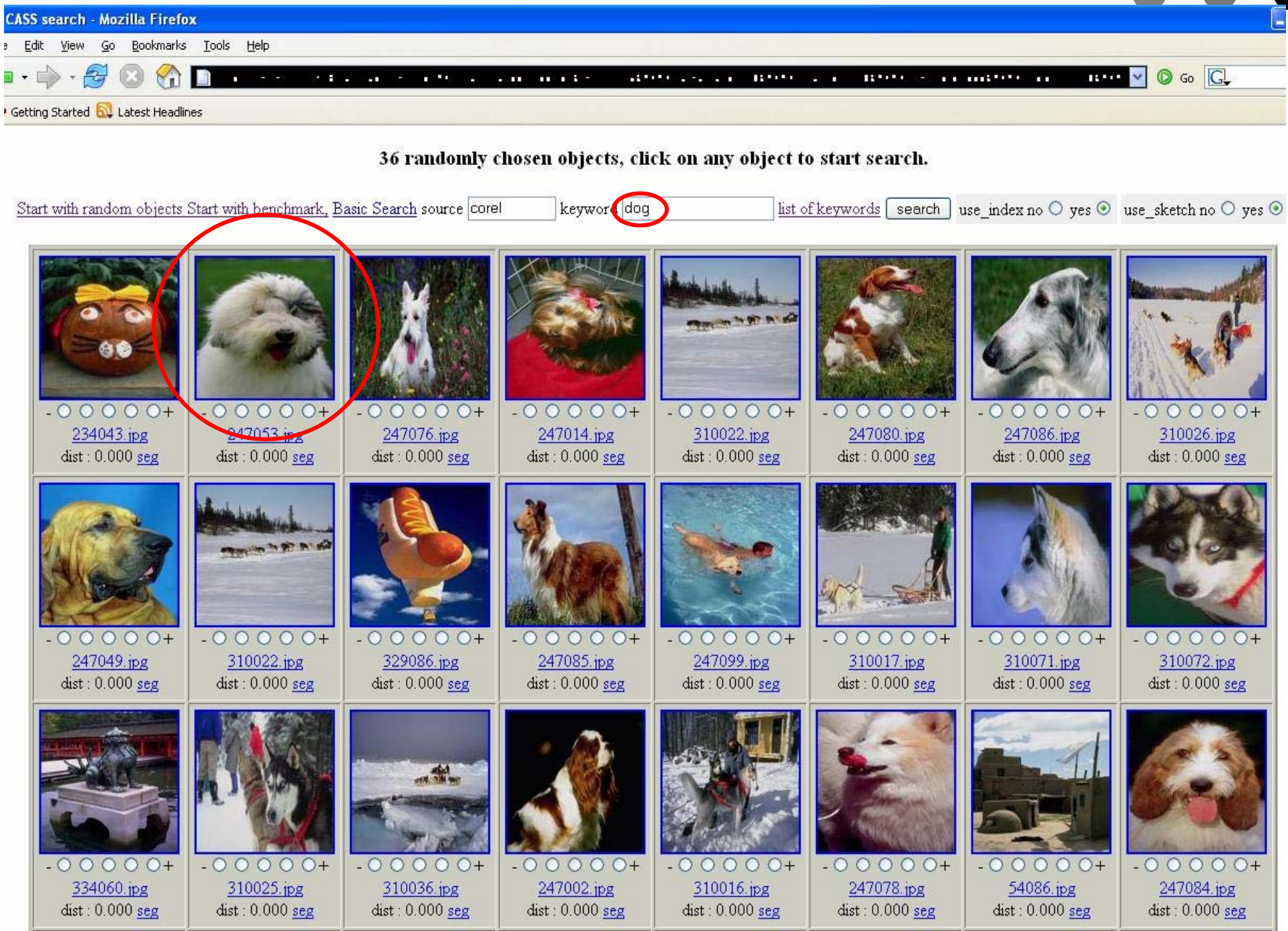


Image Similarity Search

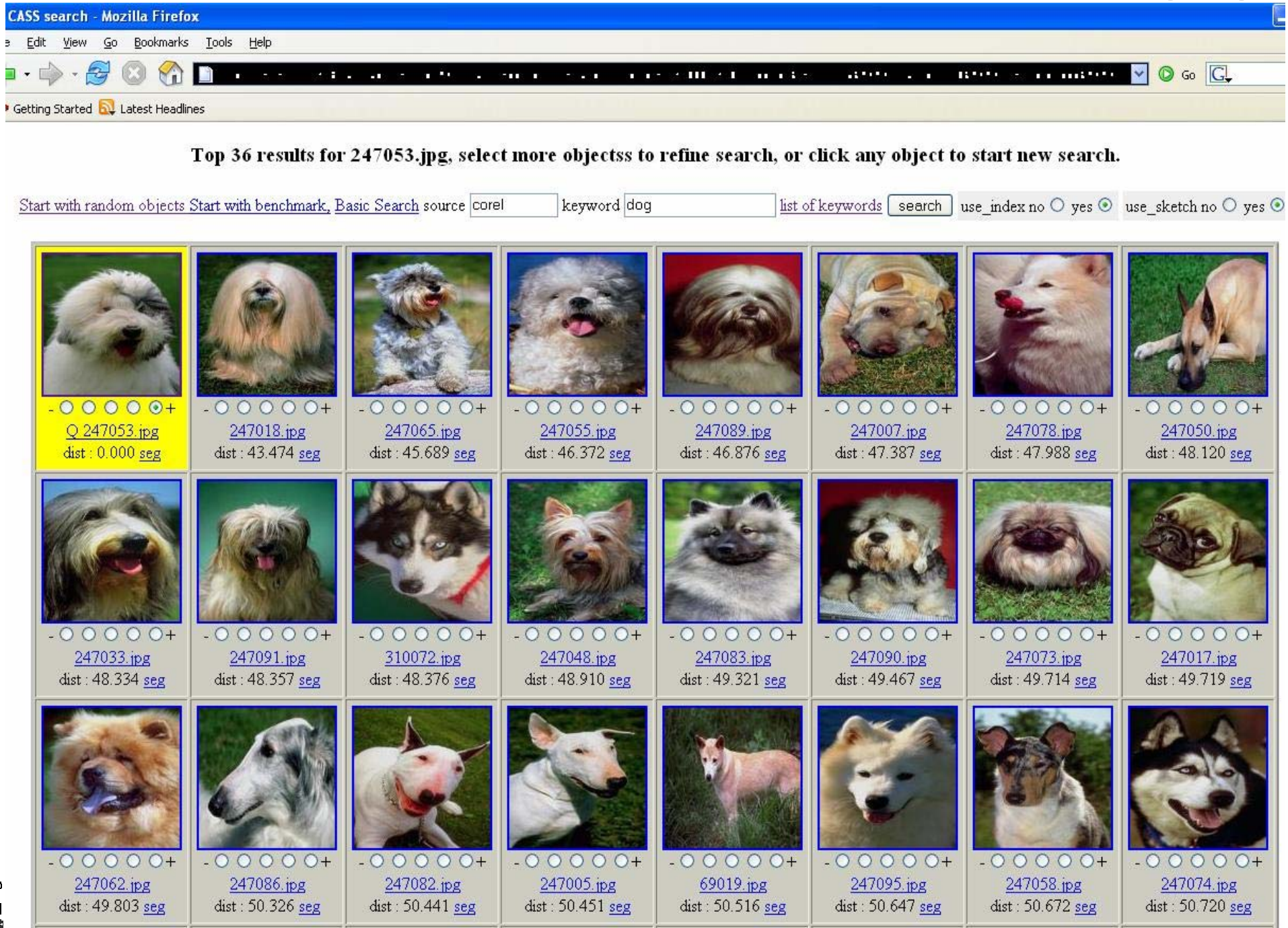


- Segmentation
  - JSEG segmentation tool from UCSB
- Feature extraction
  - 14-d features: 9-d color moments and 5-d bounding box
  - Segment weight: square root of segment size
- Distance functions
  - Segment distance: weighted ℓ_1 distance
  - Object distance: EMD

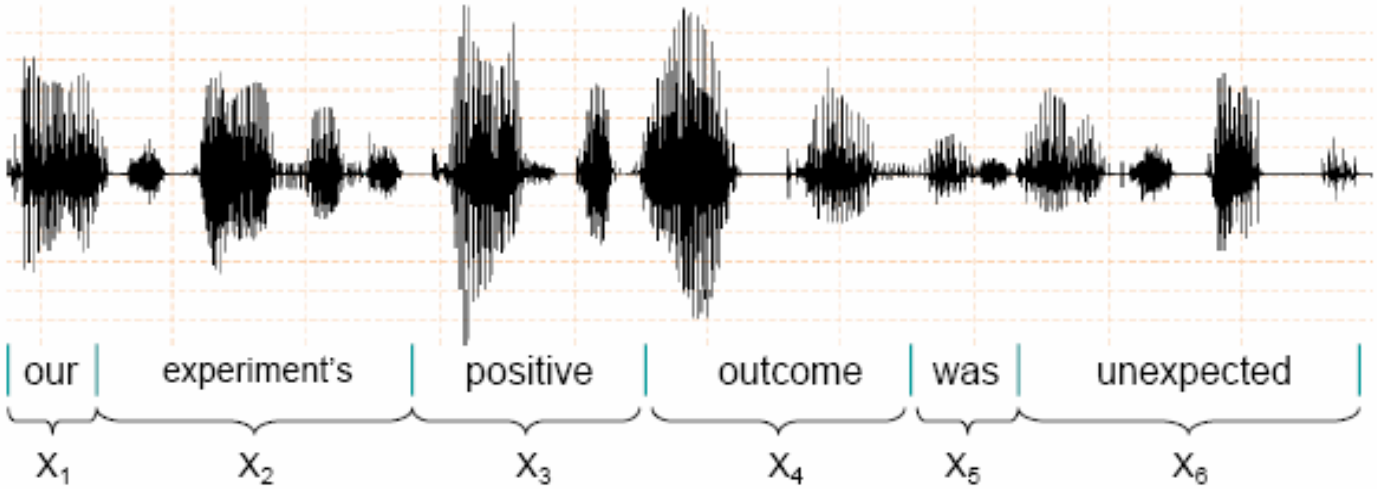
Attribute-Based Search of Corel Dataset



Content-Based Search Results

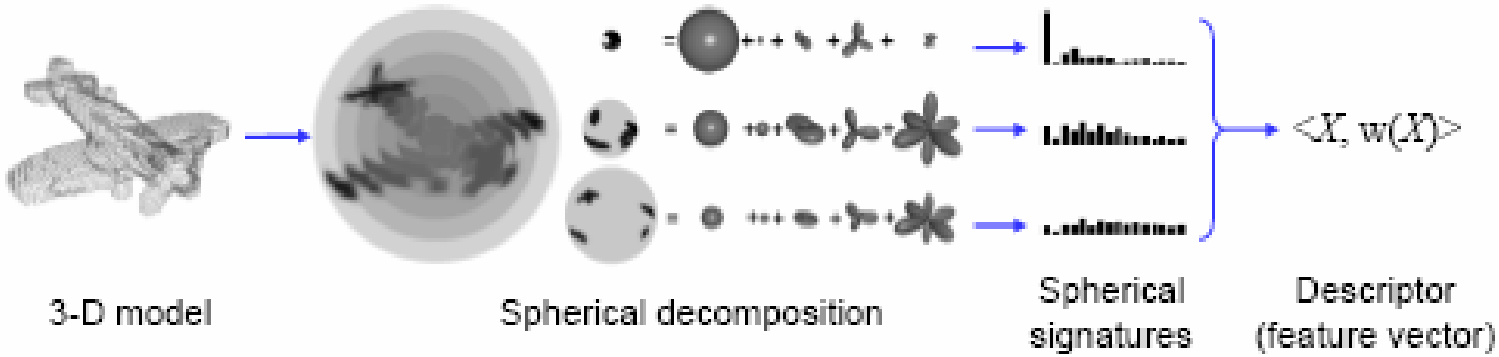


Audio Similarity Search



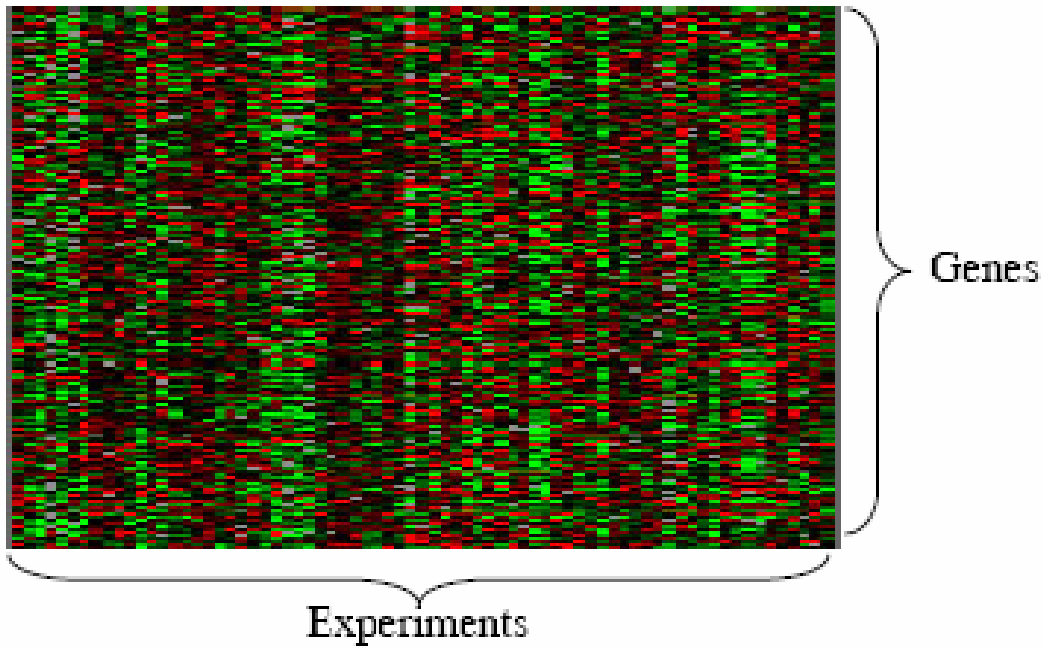
- Segmentation
  - Utterance level segmenter, human marked word boundary
- Feature extraction
  - 32 windows x 6 MFCC parameters = 192 features
  - Segment weight: proportional to segment length
- Distance functions
  - Segment distance: ℓ_1 distance
  - Object distance: EMD

3D Shape Similarity Search



- Segmentation
  - 32 decomposing spheres
- Feature extraction
  - Spherical harmonic descriptor (SHD)
  - 32 x 17 = 544 dimensions
- Distance functions
  - Segment distance: ℓ_1 distance
  - Object distance: same as segment distance

Gene Expression Similarity Search



- Segmentation
  - Gene expression microarray data: one gene per row
- Feature extraction
  - Gene expression values
- Distance function
  - Pearson correlation, spearman correlation, ℓ_1 distance

Summary

- Clear need for content-aware search systems
- Ferret toolkit
  - A step towards building content-aware search systems
  - Software infrastructure for research and collaborations
- Many interesting research problems
- Some current directions
  - Parallel implementation
    - Multicore architecture as well as cluster environment
  - Usable end system for public release

http://www.cs.princeton.edu/cass/