Ferret: A Toolkit for Content-Based Similarity Search of Feature-Rich Data

Qin Lv

Joint work with:
William Josephson, Zhe Wang,
Moses Charikar, and Kai Li

Motivations

- Digital data is everywhere
  - Increases exponentially
- Feature-rich digital data dominates data volume
  - Audio, video, digital photos, scientific sensor data
  - Systems support for managing feature-rich data
- Techniques for text data do not apply
  - Feature-rich data are noisy and high-dimensional
- Domain efforts limited to small datasets

Current Search Techniques

- Search capability is becoming an integral part of modern operating systems
  - Mac OS X Tiger: Spotlight
  - Windows Vista
- Limited to text-based search
  - Web search engines: Google, Yahoo, Microsoft, …
  - Desktop search: Google, Yahoo, MSN, …
  - Text-based documents
    - Emails, word documents, PDF files, instant messages, …
  - Text-based annotations and attributes
    - Image annotations, music (title, artist, lyrics), …

dog - Google Image Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://images.google.com/images?svnum=10&hl=en&lr=&q=dog&btnG=Search

Sign in

Google Images

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

dog Search Advanced Image Search Preferences

Moderate SafeSearch is on

Images Showing: All image sizes

Results **1** - **20** of about **3,090,000** for **dog** [definition]. (**0.12** seconds)

Dog Image
1024 x 1536 pixels - 237k - jpg
www.genome.gov

... Retriever a breed of domestic **dog**
250 x 206 pixels - 50k - jpg
en.wikipedia.org

Coole Brillen für Hunde **dog**-goes. ...
360 x 264 pixels - 71k - jpg
www.dog-goes.com

Dog Sled Wild **Dog** Mushing Company
705 x 487 pixels - 97k - jpg
www.dog-sled.com

Cat & **Dog** Pictures
1024 x 768 pixels - 117k - jpg
www.mooseyscountrygarden.com

Dog Raccoon ?
302 x 362 pixels - 19k - jpg
www.chaosproject.com

Wild **Dog** Racing
800 x 600 pixels - 106k - jpg
www.dog-sled.com
[More results from www.dog-sled.com]

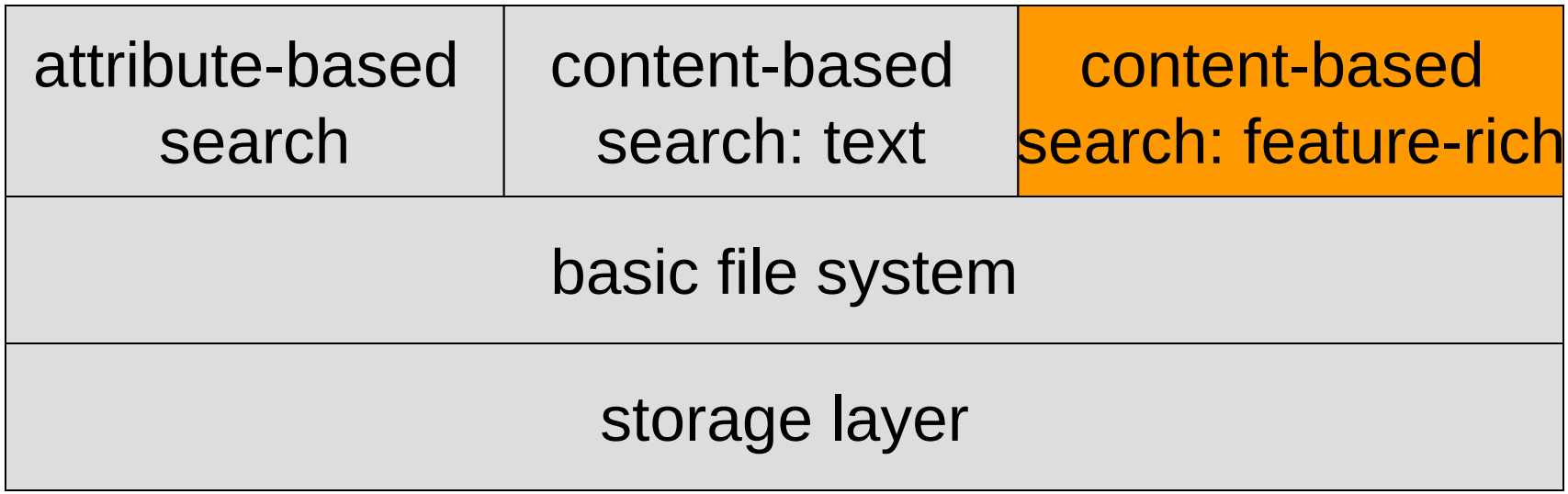
Labrador retriever **dog** food and ...
350 x 242 pixels - 9k - jpg
www.ronjun-eshop.com

Done

start Ferret.ppt dog - Google Image S... EN 7:28 PM

Princeton University

Ferret Toolkit: Design Goals

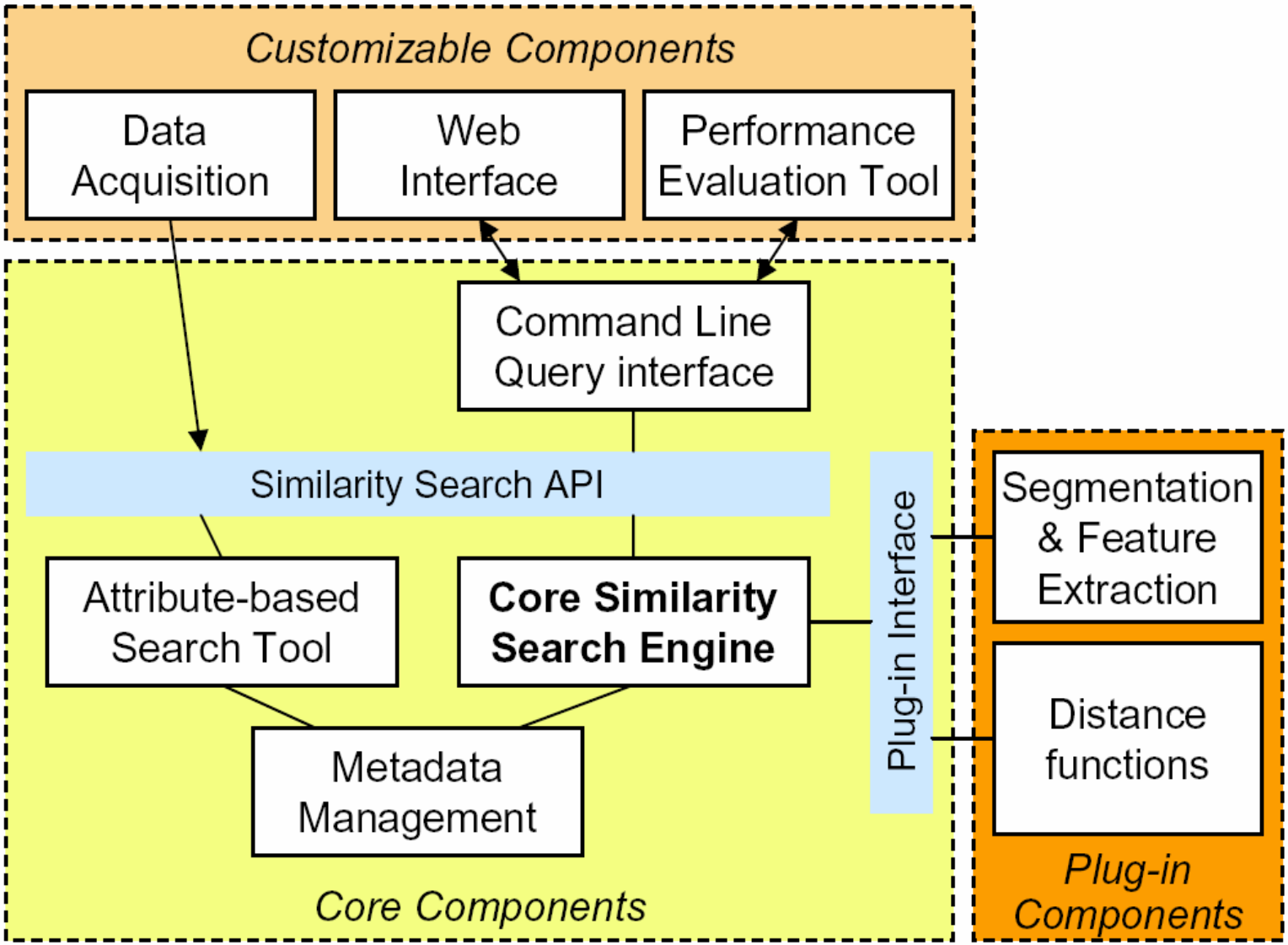


- Works with multiple feature-rich data types
  - Image, audio, 3D shape model, gene expression data
- High performance
  - Search quality, search speed, memory usage

Outline

- Motivations
- Ferret toolkit architecture design
- Similarity search problem
- Core similarity search engine
- Using the Ferret toolkit
- Evaluation results
- Conclusion & future work

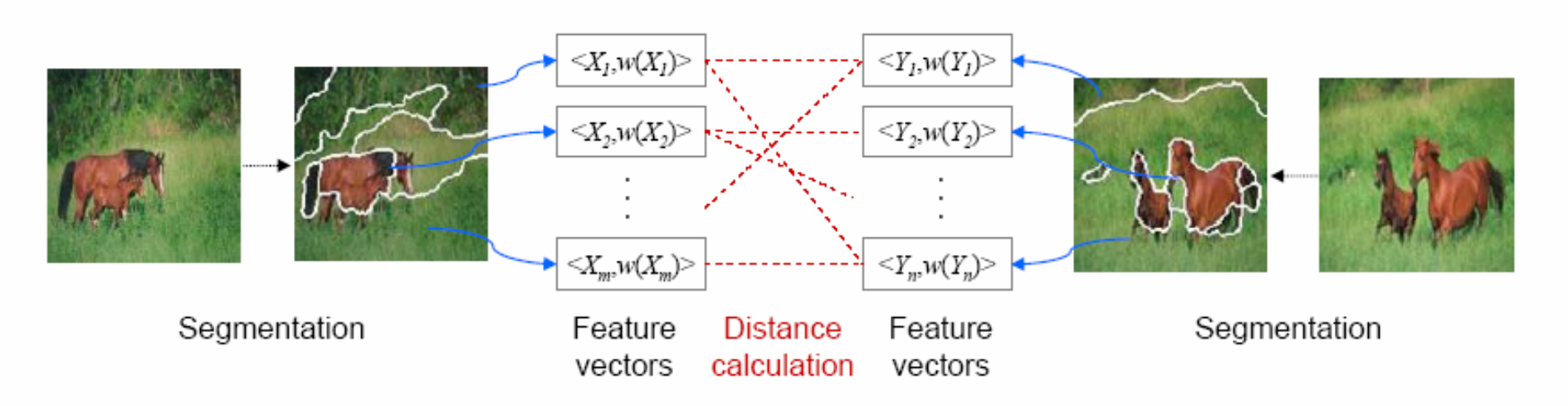
Ferret Toolkit Architecture Design

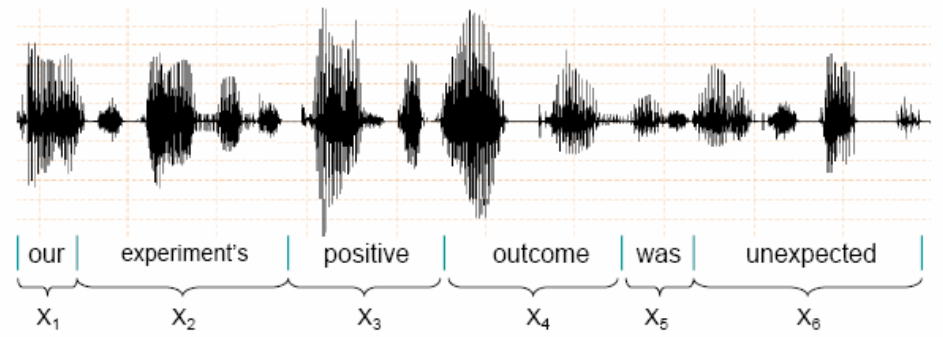


Similarity Search Problem

- Similarity search
  - Given a query object, find similar objects (*i.e.* containing similar features)
- Distance function *d (X, Y)*
  - Between objects
- Nearest neighbor search
  - K-nearest neighbor (KNN)
  - Approximate nearest neighbor (ANN)
- Hard problem for high dimensional search

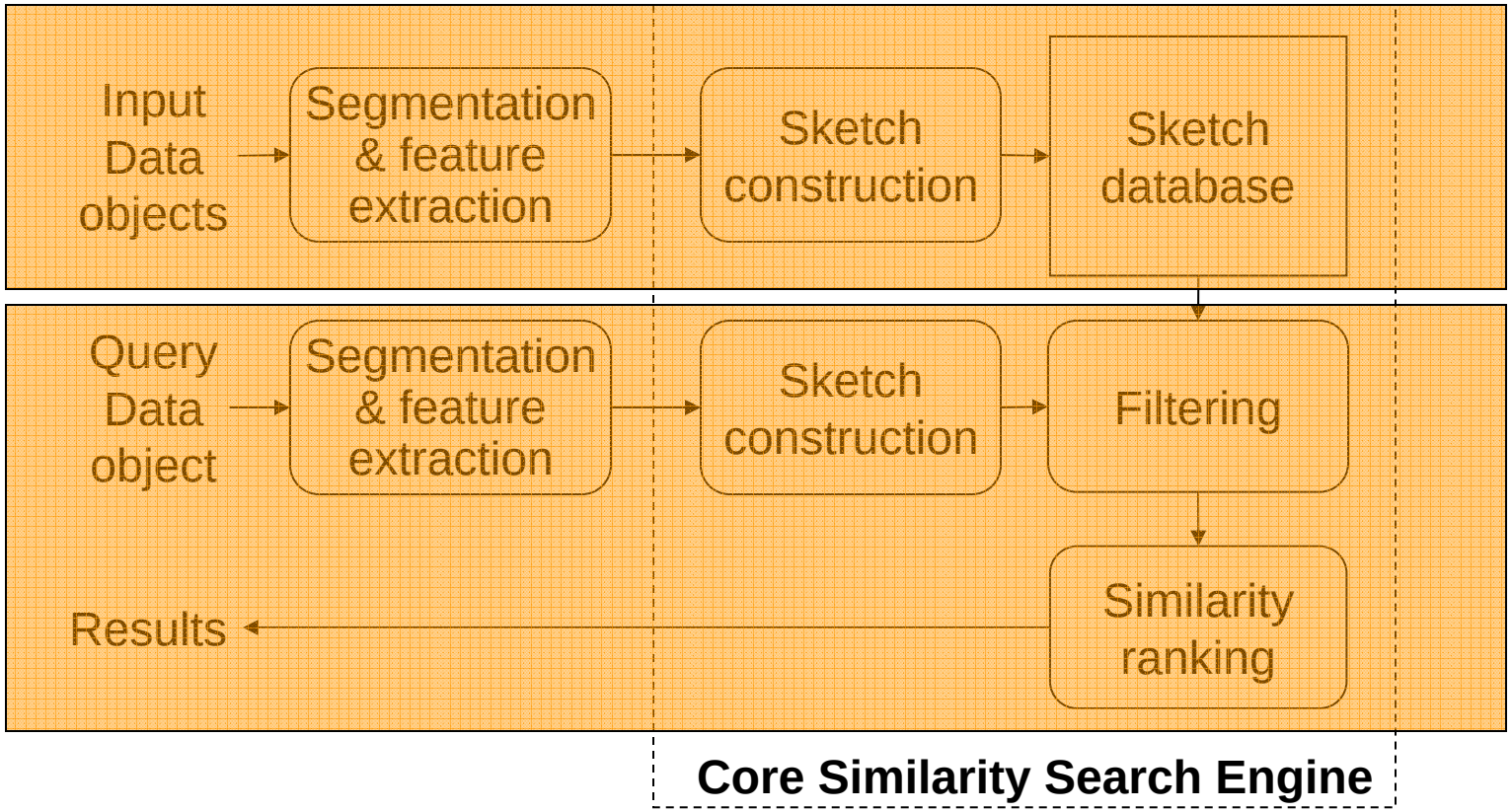
Object Representation & Distance Function



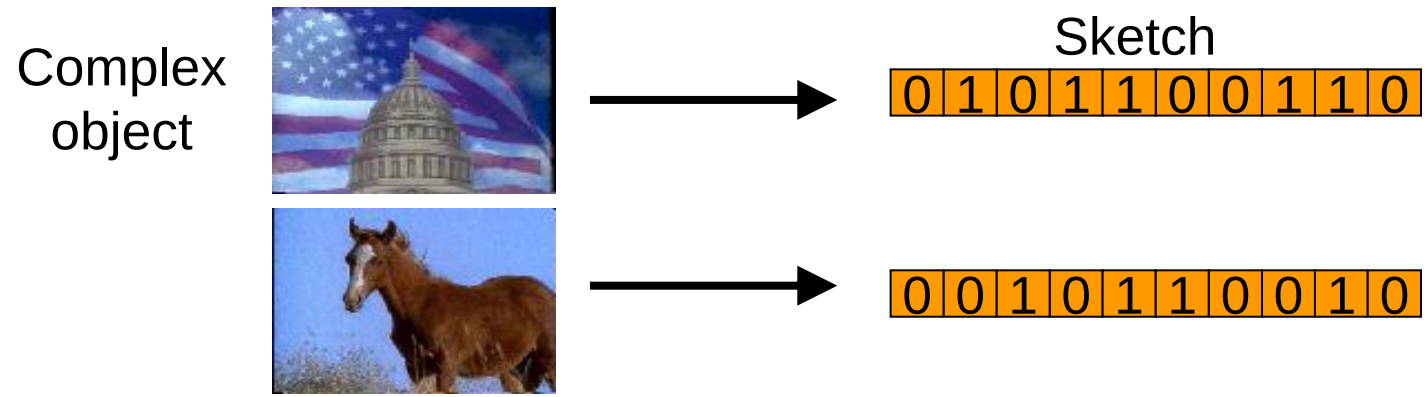


- Multi-feature representation
- Distance function
  - E.g. Earth Mover's Distance (EMD)

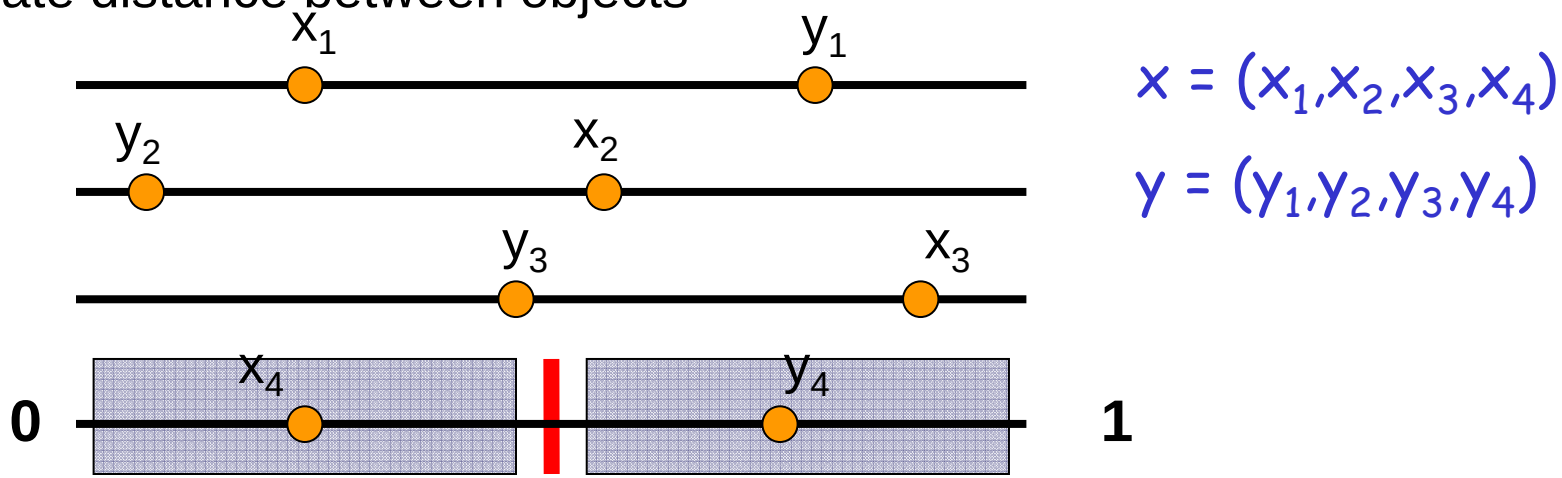
Core Similarity Search Engine



Sketch Construction



- Sketches
  - Compact data structures, estimate properties of original data
- Sketch distance
  - Hamming distance between bit vectors
  - Estimate distance between objects



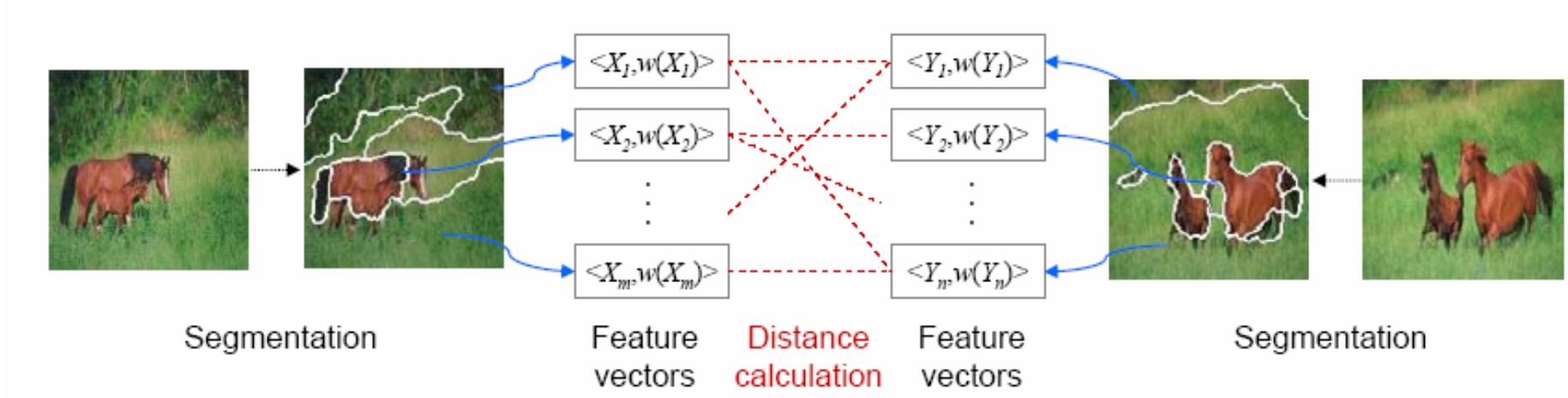
Filtering for Similarity Search

- Multi-feature representation
  - Computing object distance is expensive
- Filtering
  - Scans through the entire dataset
  - Uses a much faster distance function to filter out “bad” answers
    - Hamming distance of sketches
  - Computes object distance for a much smaller candidate set
- Criteria in picking candidate objects
  - Has at least one segment that is close enough to one of the major segments of the query object

Using the Ferret Toolkit

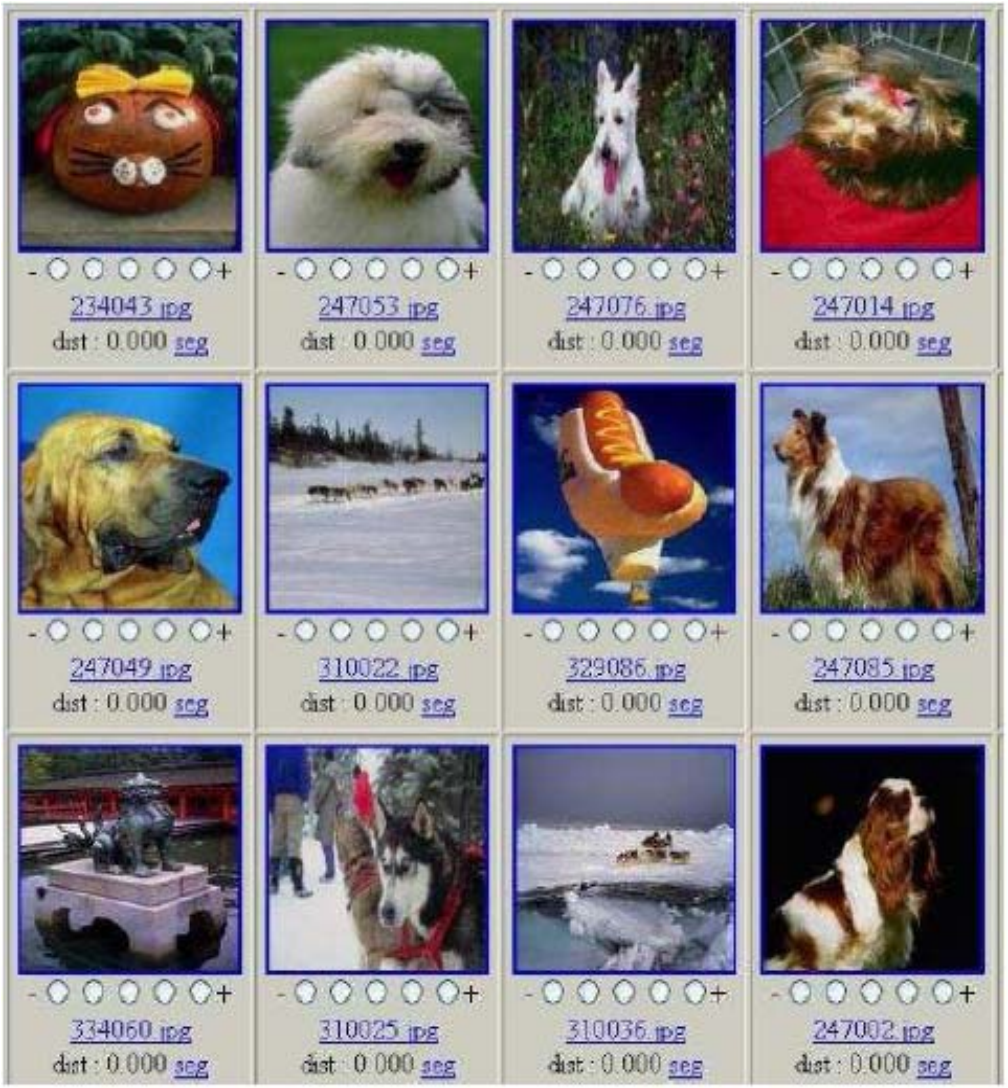
- Can the Ferret toolkit be applied to multiple data types?
  - Image data?
  - Audio data?
  - 3D shape models?
  - Gene expression data?

Image Similarity Search

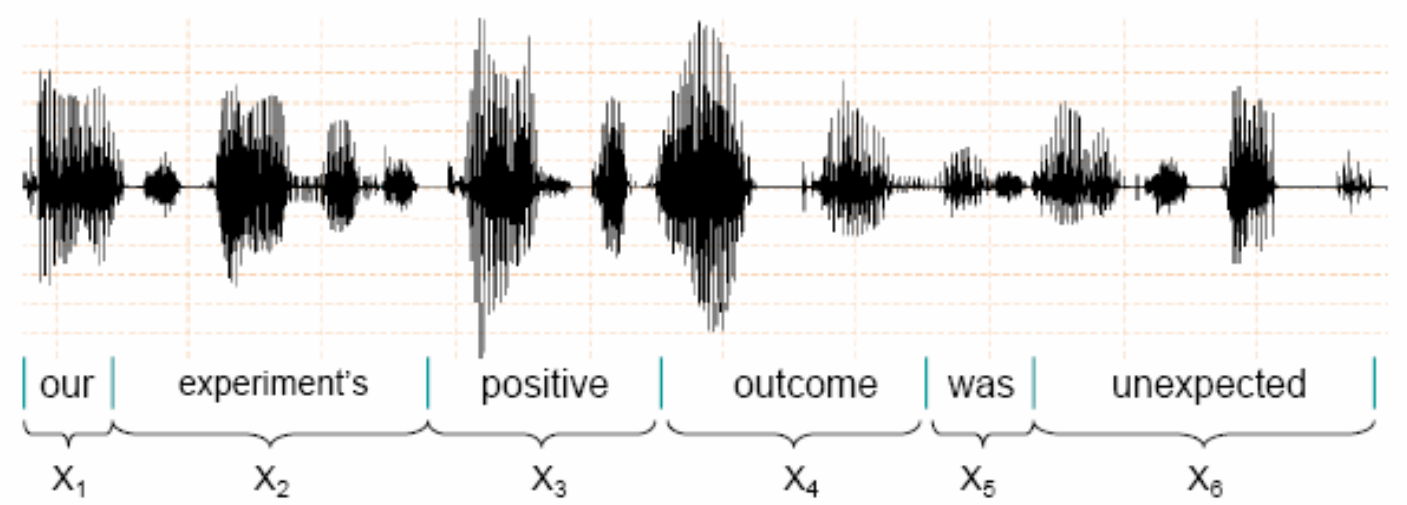


- Segmentation
  - JSEG segmentation tool from UCSB
- Feature extraction
  - 14-d features: 9-d color moments and 5-d bounding box
  - Segment weight: square root of segment size
- Distance functions
  - Segment distance: weighted ℓ_1 distance
  - Object distance: EMD

Image Similarity Search

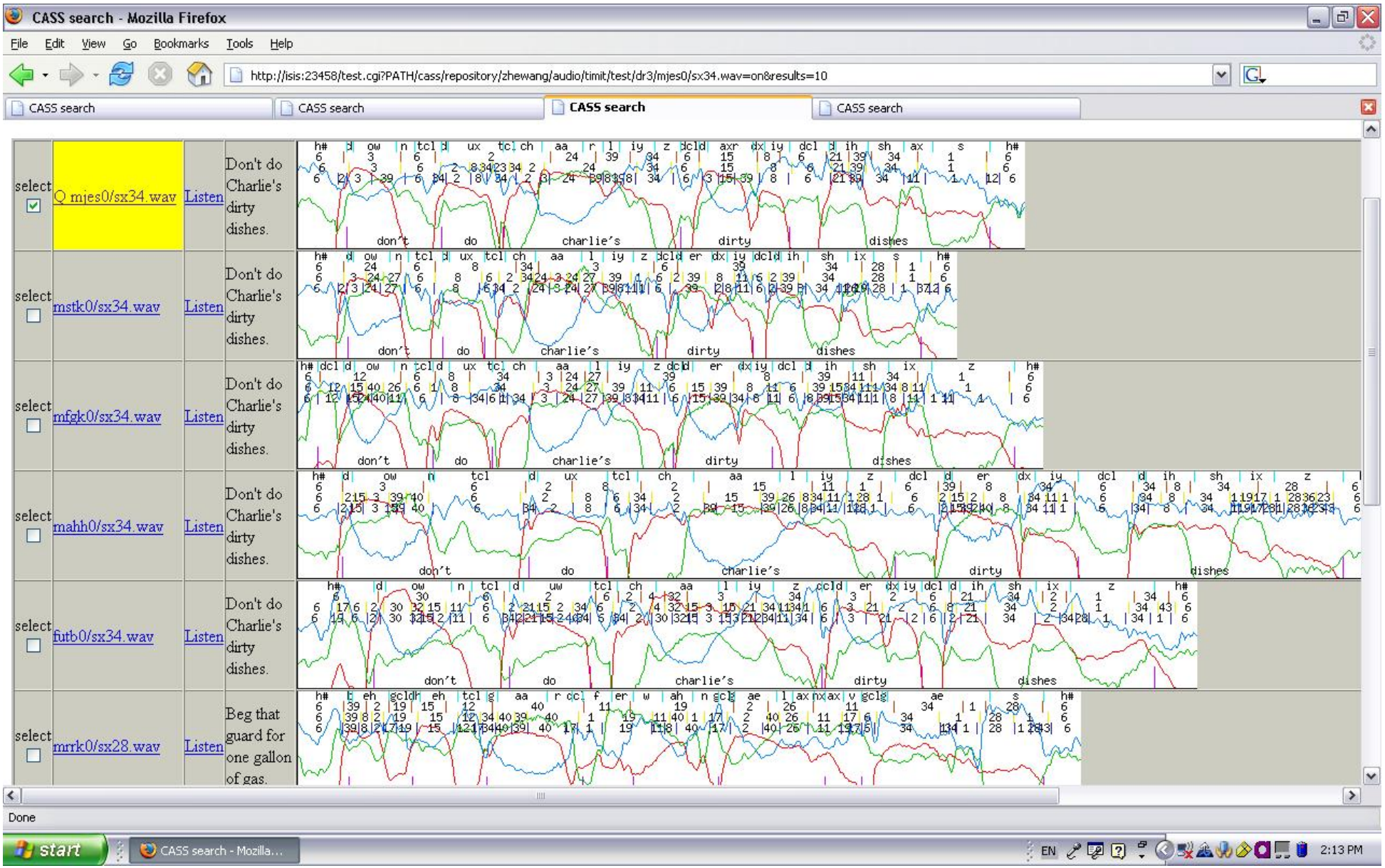


Audio Similarity Search

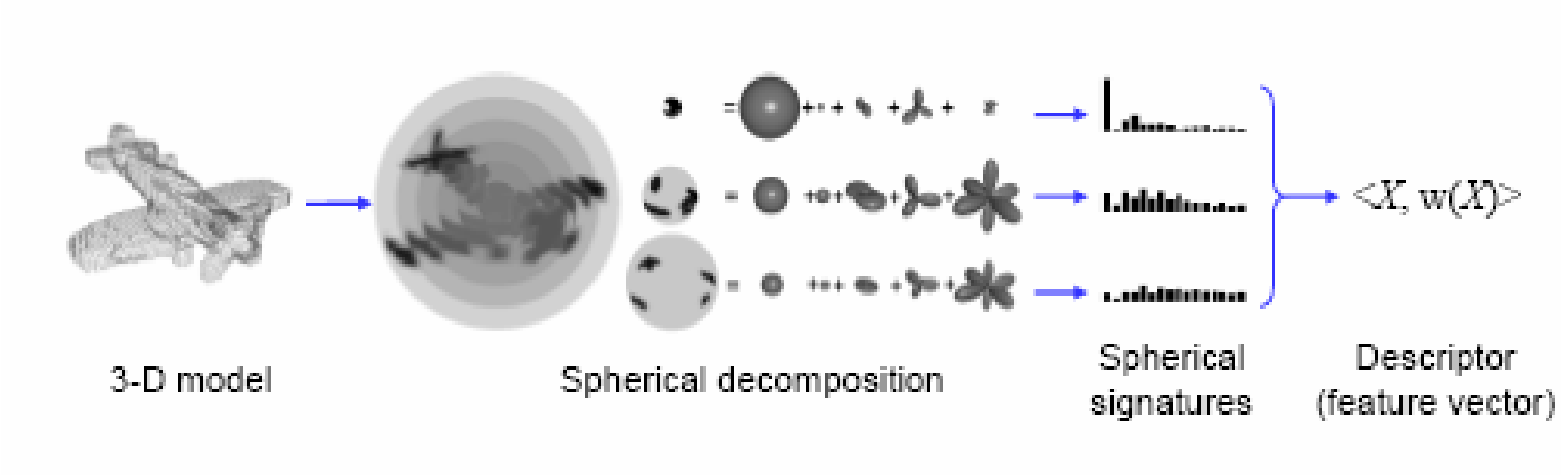


- Segmentation
  - Utterance level segmenter, human marked word boundary
- Feature extraction
  - 32 windows x 6 MFCC parameters = 192 features
  - Segment weight: proportional to segment length
- Distance functions
  - Segment distance: ℓ_1 distance
  - Object distance: EMD

Audio Similarity Search

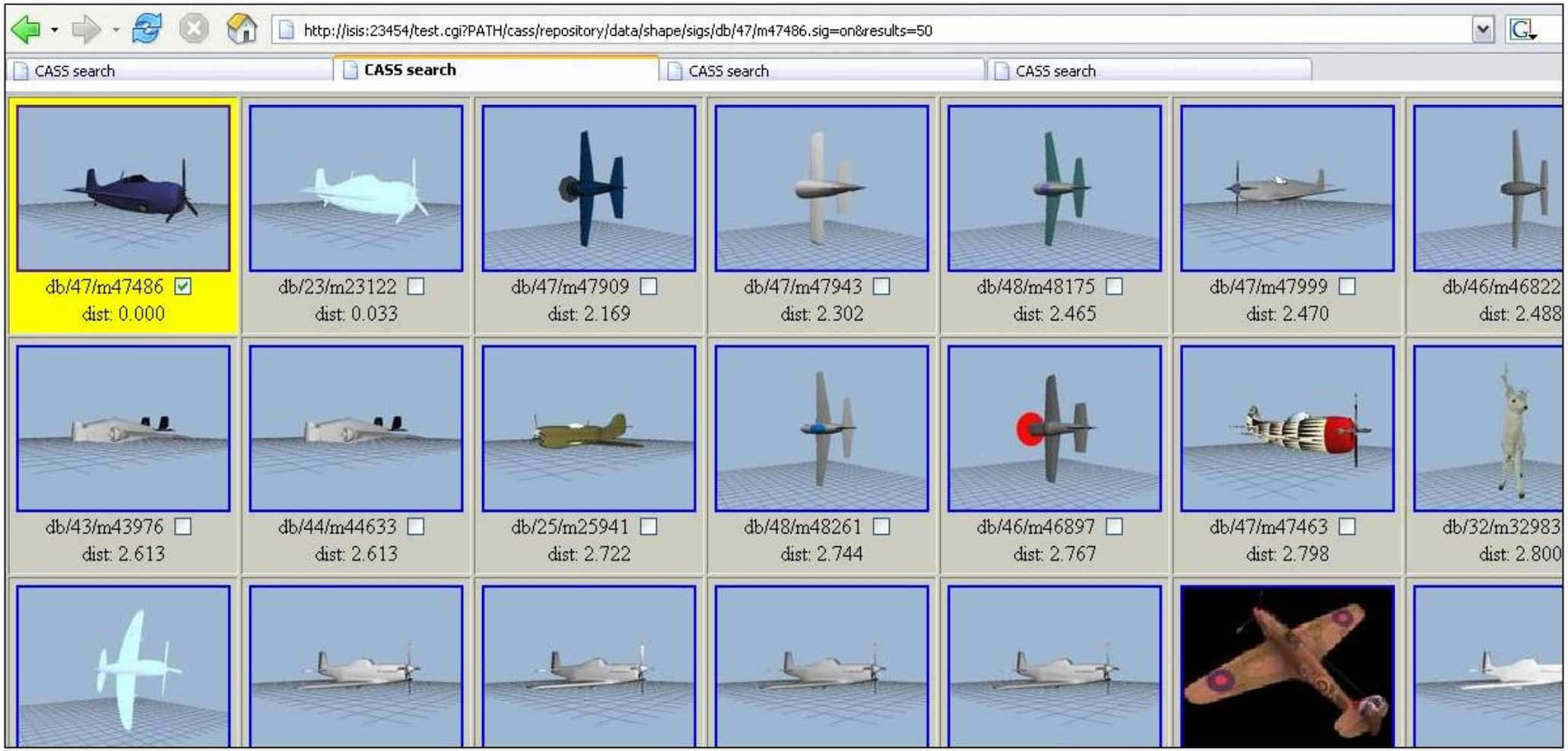


3D Shape Similarity Search

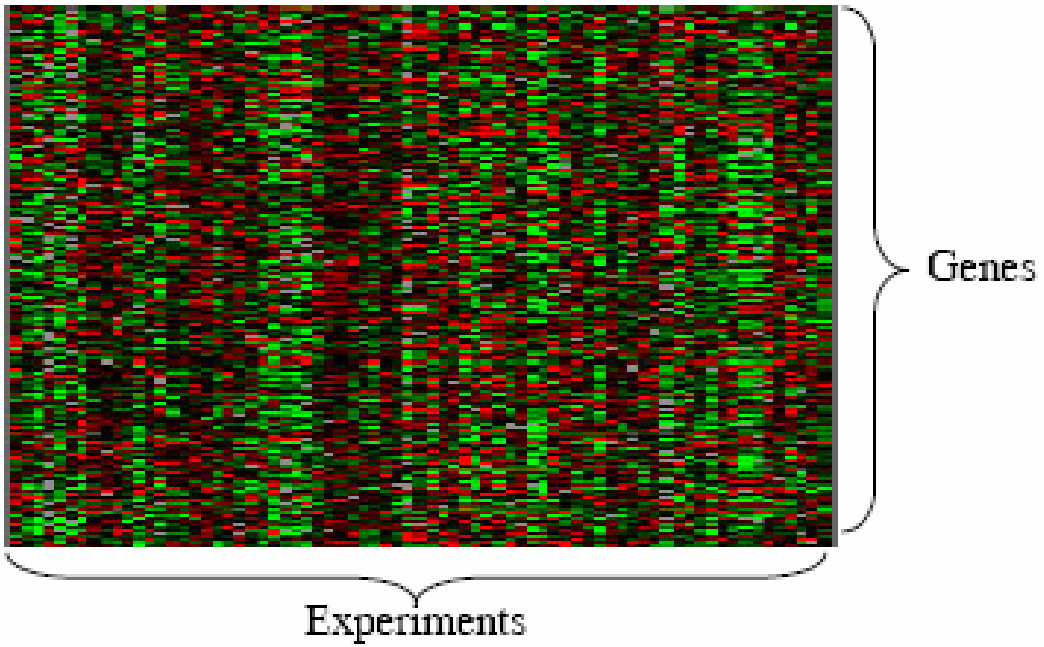


- Segmentation
  - 32 decomposing spheres
- Feature extraction
  - Spherical harmonic descriptor (SHD)
  - 32 x 17 = 544 dimensions
- Distance functions
  - Segment distance: ℓ_1 distance
  - Object distance: same as segment distance

3D Shape Similarity Search

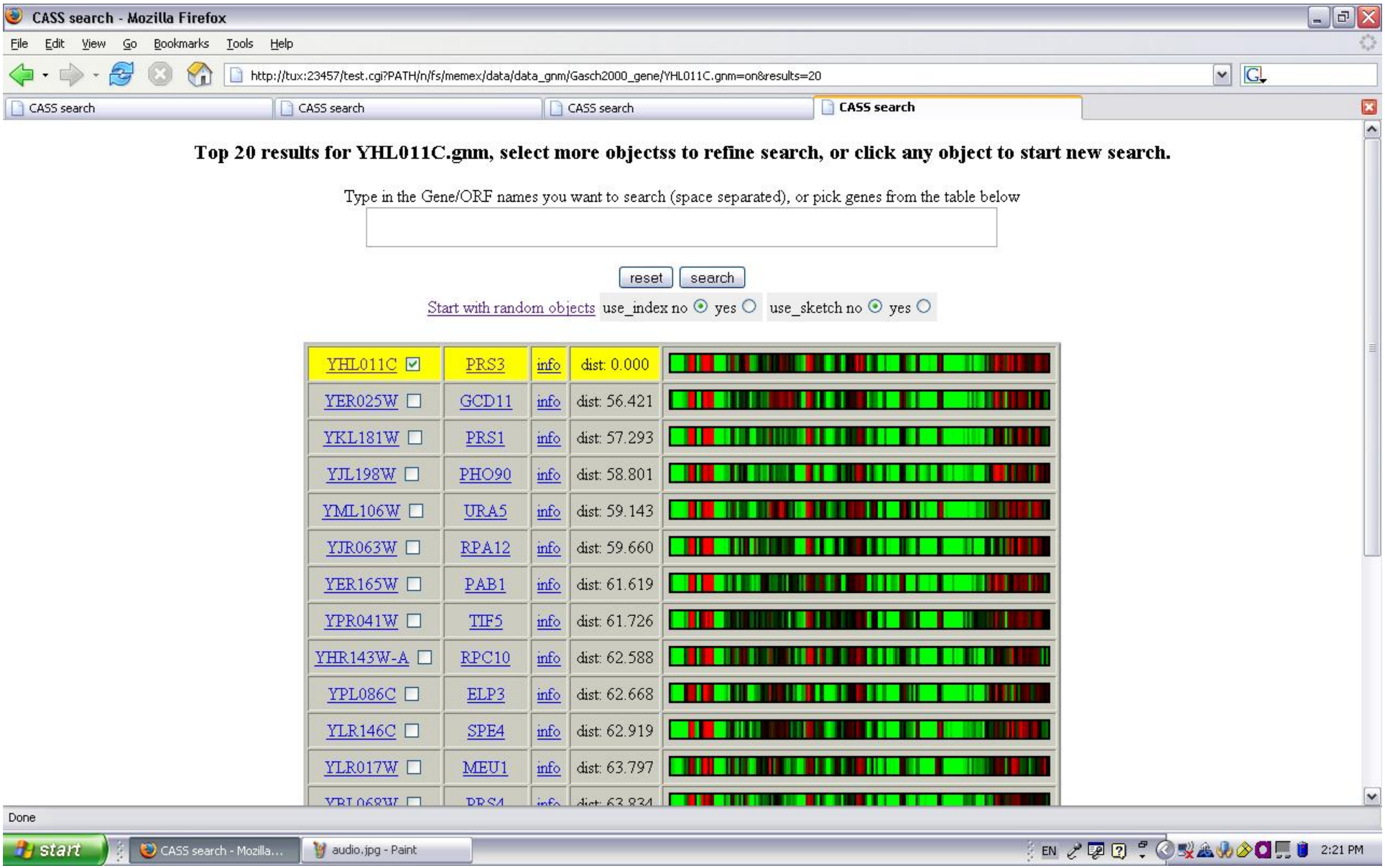


Gene Expression Similarity Search



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

Gene Expression Similarity Search



Evaluations

- Can the systems built with Ferret toolkit achieve high-quality similarity search results at a high speed?
- How small can the sketches be as the metadata of the similarity search systems?
- How much benefit can we get by using sketching and filtering?

Benchmarks

- Search quality benchmark suite
  - VARY image: 10,000 images, 32 sets
  - TIMIT audio: 6,300 sentences, 450 sets
  - PSB shape: 1,814 3D shape models, 92 sets
- Search speed benchmark suite
  - Mixed image dataset: 660,000 images
  - TIMIT audio: 6,300 sentences
  - Mixed shape dataset: 40,000 3D shape models

Search Quality Metrics

Given a query *q* with *k* similar objects:

- 1st-tier recall
  - Percentage of similar objects returned within rank *k*
- 2nd-tier recall
  - Percentage of similar objects returned within rank *2k*
- Average precision

$$\text{AveragePrecision} = \frac{1}{k}\sum_{i=1}^{k}\frac{i}{rank_i}$$

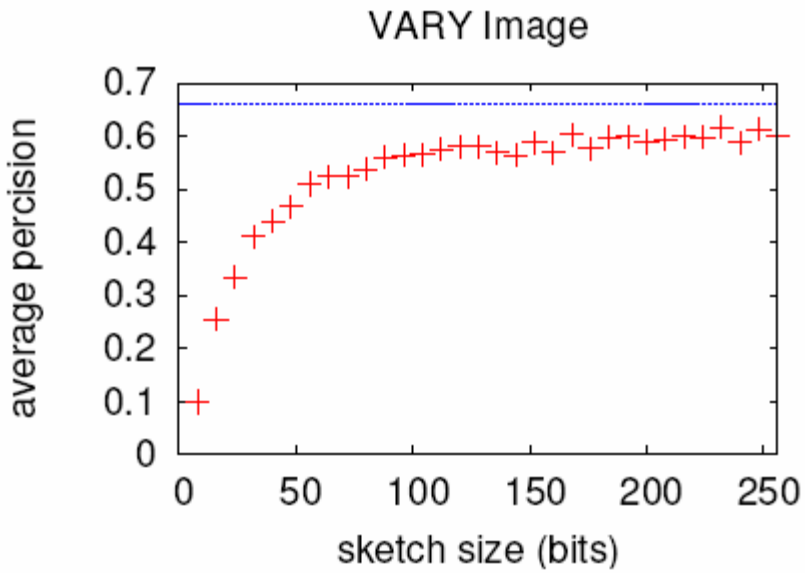
  - Example: k = 5, return 4 good results ranked at 1, 2, 5, 10
  - Average precision = (1/1 + 2/2 + 3/5 + 4/10) / 5 = 0.6

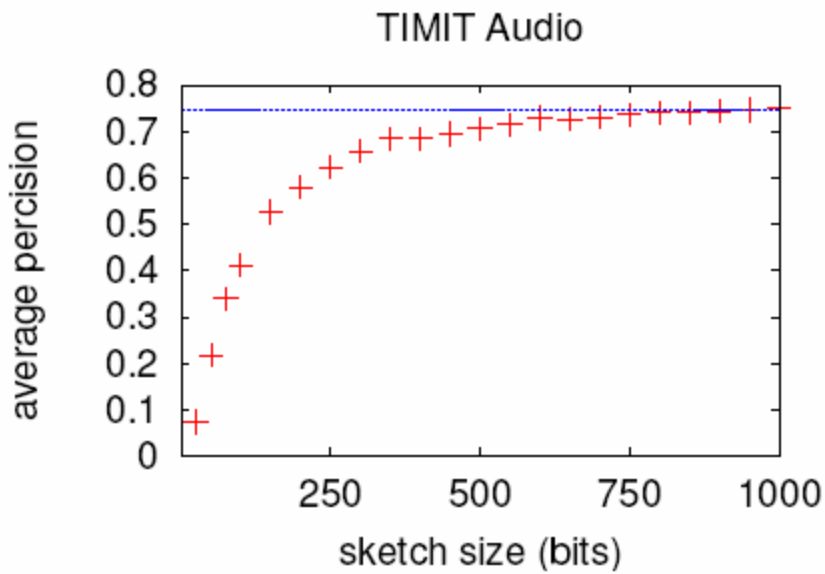
Search Quality & Search Speed

	Method	Average Precision	1st-tier	2nd-tier	Vector Size (bits)	Size Ratio
VARY Image	Ferret	0.59	0.54	0.63	96	**5:1**
	SIMPLIcity	0.41	0.41	0.47	264	
TIMT Audio	Ferret	0.44	0.42	0.49	600	**10:1**
PSB 3D Shape	Ferret	0.32	0.30	0.41	800	**22:1**
	SHD	0.33	0.32	0.43	17472	

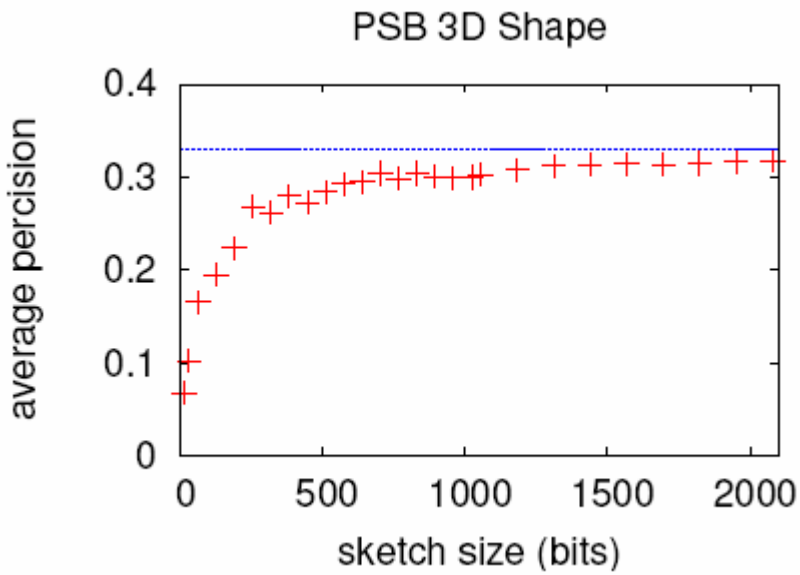
	#Data Objects	#Vectors / Object	Search Time (s)
Mixed Image	660,000	10.8	2.0
TIMIT Audio	6,300	8.6	0.09
Mixed 3D Shape	40,000	1	0.01

Search Quality vs. Sketch Size

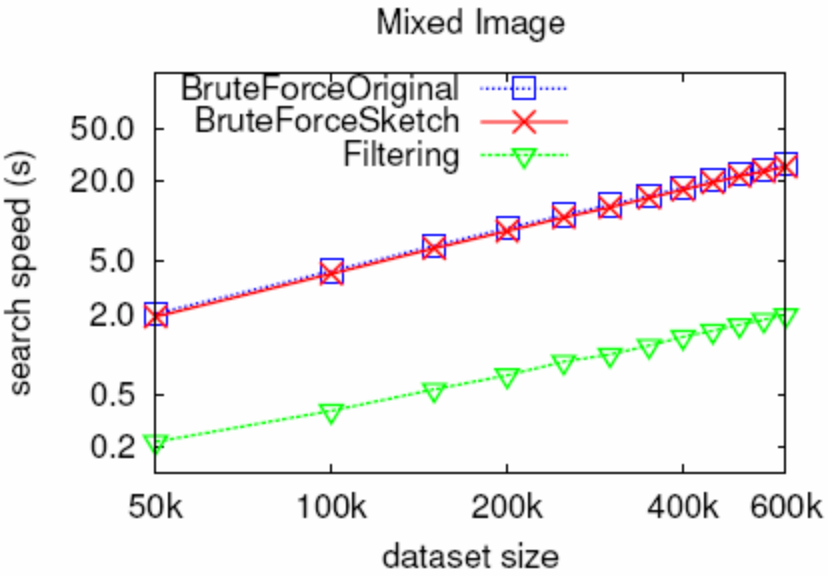


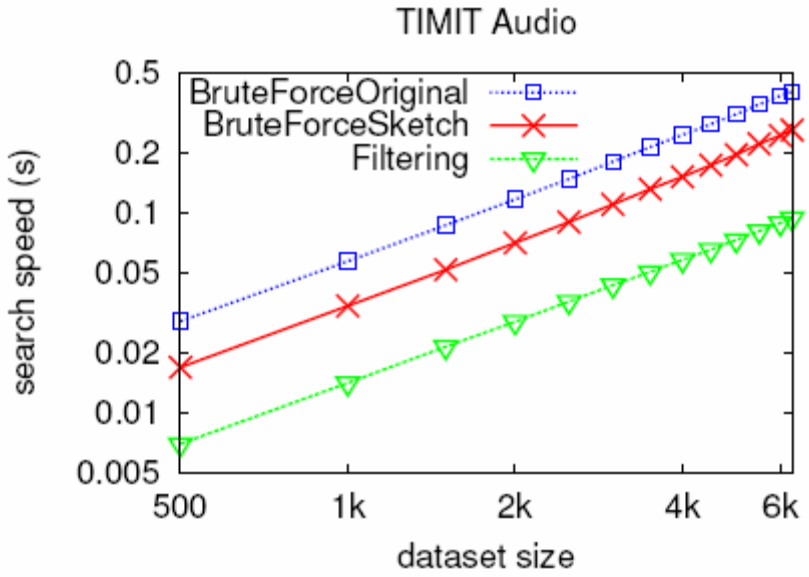


	Sketch Size	Sketch Size
VARY Image	64 bits (7:1)	88 bits (5:1)
TIMIT Audio	250 bits (6:1)	450 bits (3:1)
PSB 3D Shape	200 bits (87:1)	600 bits (29:1)

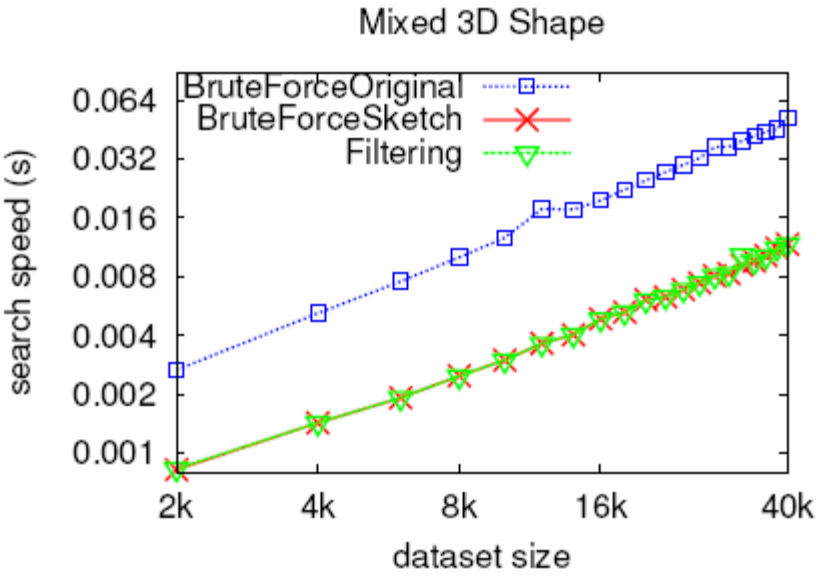


Brute-Force, Sketching, Filtering





- BruteForceOriginal
  - Linear scan using original feature vectors
- BruteForceSketch
  - Linear scan using segment sketches
- Filtering
  - Filtering using segment sketches



Conclusion & Future Work

- Ferret toolkit for content-based similarity search
  - Used for image, audio, 3D shape, genomic data
- Achieves high search quality at reasonably high search speed
- Using sketches greatly reduces metadata size with minimal quality degradation
- Future work
  - Integrate with attribute-based search
  - Indexing techniques
  - More effective and efficient distance functions and corresponding sketching techniques
  - More data types: video, sensor data

Thanks!

- CASS: Content-Aware Search Systems
  - http://www.cs.princeton.edu/cass
  - Try our image similarity search tool for Windows
    - http://www.cs.princeton.edu/cass/software