Compact Data Representations and their Applications

Moses Charikar
Princeton University

Lots and lots of data

- AT&T
- Information about who calls whom
- What information can be got from this data ?

- Network router
- Sees high speed stream of packets
- Detect DOS attacks ?
  fair resource allocation ?

Lots and lots of data

- Typical search engine
- A few billion web pages
- Many many queries every day
- How to efficiently process data ?
  - Eliminate near duplicate web pages
  - Query log analysis

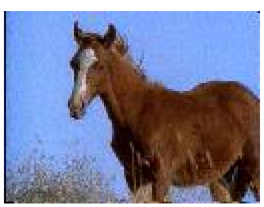
Sketching Paradigm

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

complex object

sketch

0	1	0	1	1	0	0	1	1	0



0	0	1	0	1	1	0	0	1	0

Why care about compact representations ?

- Practical motivations
  - Algorithmic techniques for massive data sets
  - Compact representations lead to reduced space, time requirements
  - Make impractical tasks feasible

- Theoretical Motivations
  - Interesting mathematical problems
  - Connections to many areas of research

Questions

- What is the data ?
- What functions do we want to compute on the data ?
- How do we estimate functions on the sketches ?

- Different considerations arise from different combinations of answers

- Compact representation schemes are functions of the requirements

What is the data ?

- Sets, vectors, points in Euclidean space, points in a metric space, vertices of a graph.

- Mathematical representation of objects (e.g. documents, images, customer profiles, queries).

Distance/similarity functions

- Distance is a general metric, i.e satisfies triangle inequality

- Normed space
  $x = (x_1, x_2, \ldots, x_d)$ $y = (y_1, y_2, \ldots, y_d)$

$$d(x, y) = \left(\sum_{i=1}^{d} | x_i - y_i |^p \right)^{1/p}$$

$$L_p \ norm \qquad L_1, L_2, L_\infty$$

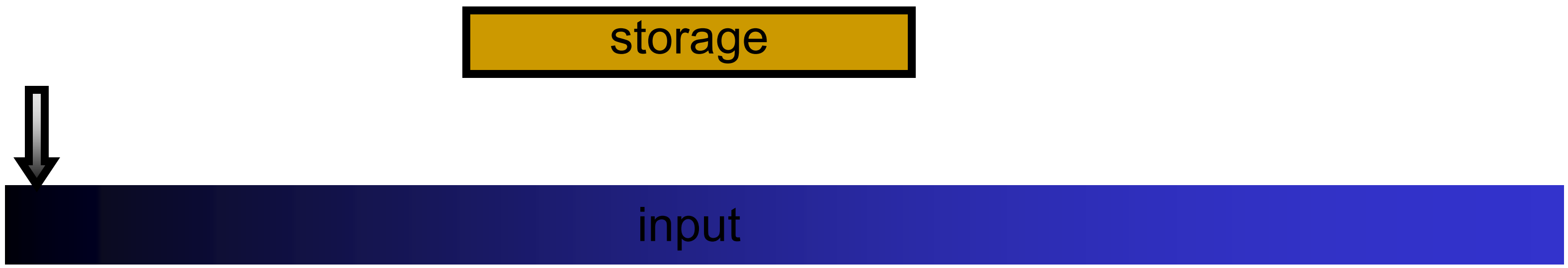
- Other special metrics
  (e.g. Earth Mover Distance)

Estimating distance from sketches

- Arbitrary function of sketches
  - Information theory, communication complexity question.
- Sketches are points in normed space
  - Embedding original distance function in normed space. [Bourgain '85] [Linial,London,Rabinovich '94]
- Original metric is (same) normed space
  - Original data points are high dimensional
  - Sketches are points low dimensions
  - Dimension reduction in normed spaces [Johnson Lindenstrauss '84]

Streaming algorithms

- Perform computation in one (or constant) pass(es) over data using a small amount of storage space



- Availability of sketch function facilitates streaming algorithm
- Additional requirements - sketch should allow:
  - Update to incorporate new data items
  - Combination of sketches for different data sets

Talk Outline:
Glimpse of Compact Representation Techniques

- Dimension reduction

- Similarity preserving hash functions
  - sketching vector norms
  - sketching sets of points:
    Earth Mover Distance (EMD)

Low Distortion Embeddings

- Given metric spaces (X_1,d_1) & (X_2,d_2), embedding f: $X_1 \to X_2$ has distortion D if ratio of distances changes by at most D

http://www.physast.uga.edu/~jss/1010/ch10/earth.jpg

f

http://humanities.ucsd.edu/courses/kuchtahum4/pix/earth.jpg

- "Dimension Reduction" –
  - Original space high dimensional
  - Make target space be of "low" dimension, while maintaining small distortion

Dimension Reduction in L_2

- n points in Euclidean space (L_2 norm) can be mapped down to $O((\log n)/\varepsilon^2)$ dimensions with distortion at most $1+\varepsilon$.
  [Johnson Lindenstrauss '84]

- Two interesting properties:
  - Linear mapping
  - Oblivious – choice of linear mapping does not depend on point set
  - Quite simple [JL84, FM88, IM98, DG99, Ach01]: Even a random +1/-1 matrix works…
- Many applications…

Dimension reduction for L_1

- [C,Sahai '02]
  Linear embeddings are not good for dimension reduction in L_1

- There exist $O(n)$ points in L_1 in n dimensions, such that any *linear mapping* with distortion δ needs n/δ^2 dimensions

Dimension reduction for L_1

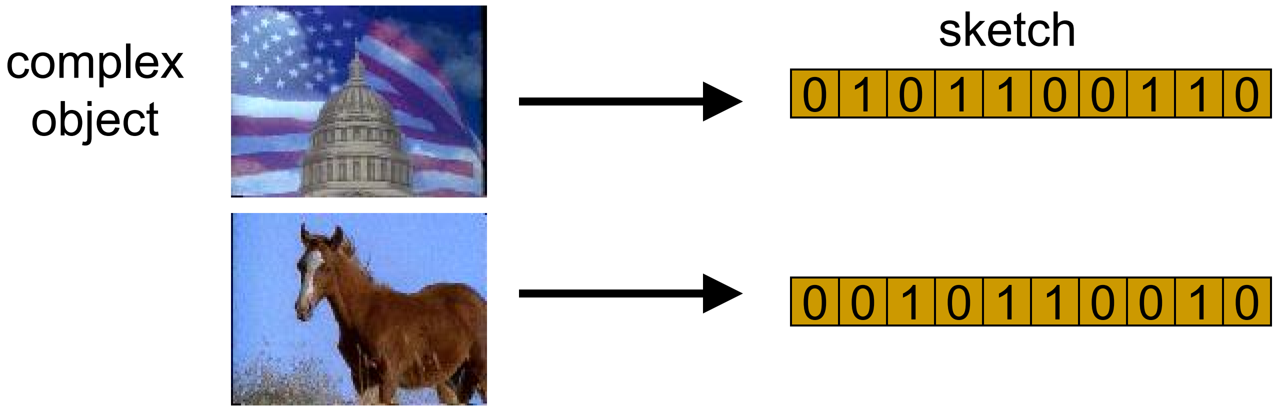
- [C, Brinkman '03]
  Strong lower bounds for dimension reduction in L_1
- There exist n points in L_1, such that *any embedding* with constant distortion δ needs n^{1/δ^2} dimensions
- Simpler proof by [Lee,Naor '04]
- Does not rule out other sketching techniques

Talk Outline

- Dimension reduction

- Similarity preserving hash functions
  - sketching vector norms
  - sketching sets of points:
    Earth Mover Distance (EMD)

Similarity Preserving Hash Functions



- Similarity function *sim(x,y),* distance *d(x,y)*
- Family of hash functions *F* with probability distribution such that

$$\Pr_{h \in F}[h(x) = h(y)] = sim(x, y)$$

$$\Pr_{h \in F}[h(x) \neq h(y)] = d(x, y)$$

Applications

- Compact representation scheme for estimating similarity

$$x \longrightarrow (h_1(x), h_2(x), \ldots, h_k(x))$$

$$y \longrightarrow (h_1(y), h_2(y), \ldots, h_k(y))$$

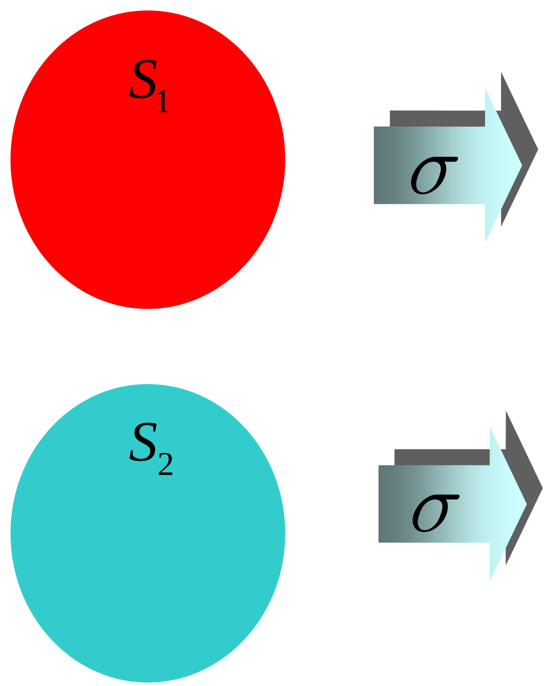
- Approximate nearest neighbor search
[Indyk,Motwani '98]
[Kushilevitz,Ostrovsky,Rabani '98]

Sketching Set Similarity:
Minwise Independent Permutations

[Broder,Manasse,Glassman,Zweig ‘97]

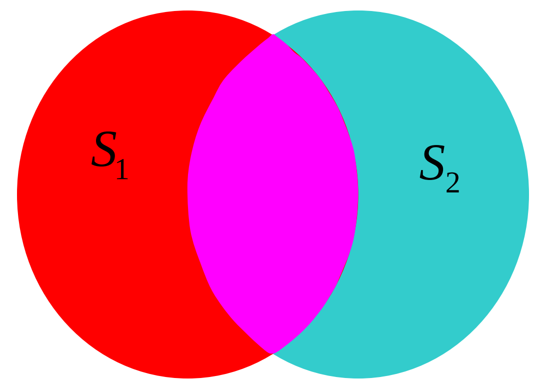
[Broder,C,Frieze,Mitzenmacher ‘98]

$$\text{similarity} = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$



$\min(\sigma(S_1))$

$\min(\sigma(S_2))$



$$\text{prob}(\min(\sigma(S_1)) = \min(\sigma(S_2))) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Existence of SPH schemes [C '02]

- *sim(x,y)* admits a similarity preserving hashing scheme if
  $\exists$ family of hash functions *F* such that

$$\Pr_{h\in F}[h(x)=h(y)]=sim(x,y)$$

- If *sim(x,y)* admits an SPH scheme then *1-sim(x,y)* is a distance metric isometrically embeddable in the Hamming cube.

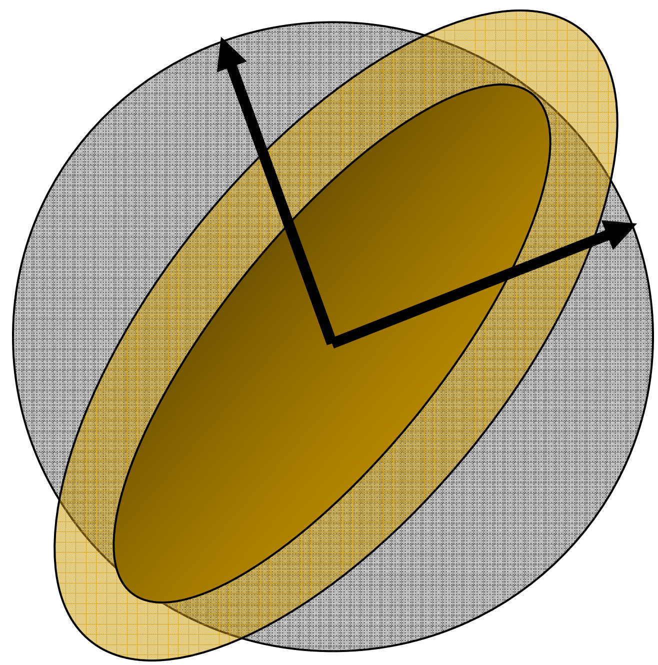
Random Hyperplane Rounding based SPH

- Collection of vectors

$$\mathrm{sim}(u,v) = 1 - \frac{\arccos(u,v)}{\pi}$$

- Pick random hyperplane through origin (normal *r*)

$$h_{\vec{r}}(\vec{u}) = \begin{cases} 1 & \text{if } \vec{r} \cdot \vec{u} \geq 0 \\ 0 & \text{if } \vec{r} \cdot \vec{u} < 0 \end{cases}$$



- Sketch is a bit vector
- [Goemans,Williamson '94]

Sketching L_1

- Design sketch for vectors to estimate L_1 norm
- Hash function to distinguish between small and large distances [KOR '98]
  - Map L_1 to Hamming space
  - Bit vectors $a=(a_1,a_2,\ldots,a_n)$ and $b=(b_1,b_2,\ldots,b_n)$
  - Distinguish between distances $\leq (1-\varepsilon)n/k$ versus $\geq (1+\varepsilon)n/k$
  - XOR random set of k bits
  - $\Pr[h(a)=h(b)]$ differs by constant in two cases

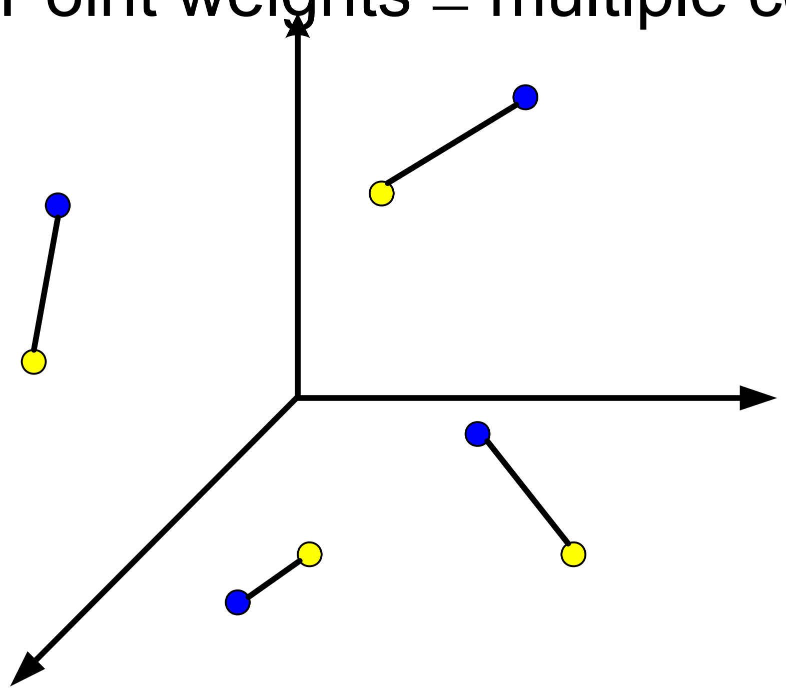
Sketching L_1 via stable distributions

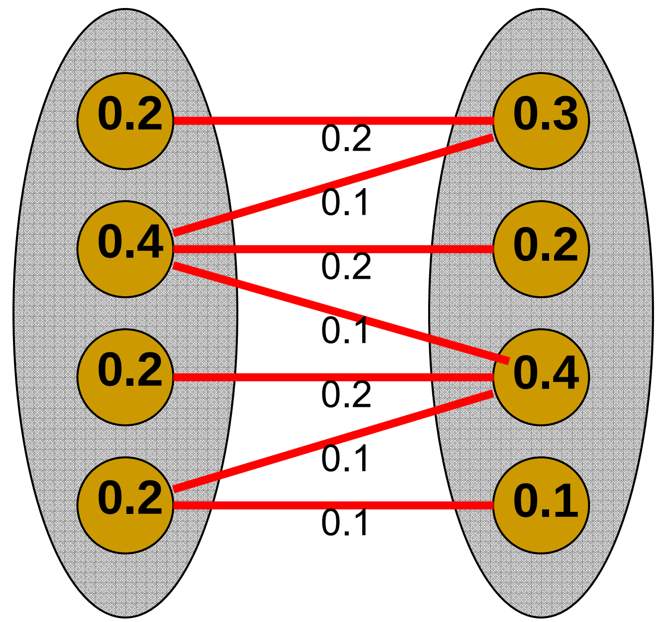
- $a=(a_1,a_2,\ldots,a_n)$ and $b=(b_1,b_2,\ldots,b_n)$
- Sketching L_2
  - $f(a) = \Sigma_i a_i X_i \quad f(b) = \Sigma_i b_i X_i$
    X_i independent Gaussian
  - $f(a)-f(b)$ has Gaussian distribution scaled by $|a-b|_2$
  - Form many coordinates, estimate $|a-b|_2$ by taking L_2 norm

- Sketching L_1
  - $f(a) = \Sigma_i a_i X_i \quad f(b) = \Sigma_i b_i X_i$
    X_i independent Cauchy distributed
  - $f(a)-f(b)$ has Cauchy distribution scaled by $|a-b|_1$
  - Form many coordinates, estimate $|a-b|_1$ by taking median
    [Indyk '00] -- streaming applications

Earth Mover Distance (EMD): Bipartite/Bichromatic matching

- Minimum cost matching between two sets of points.
- Point weights ≡ multiple copies of points

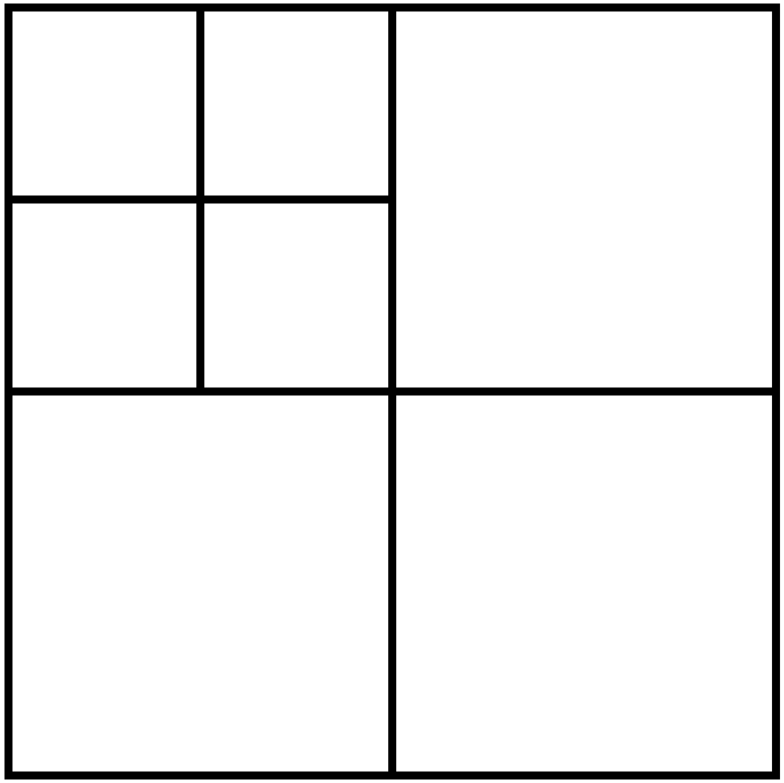


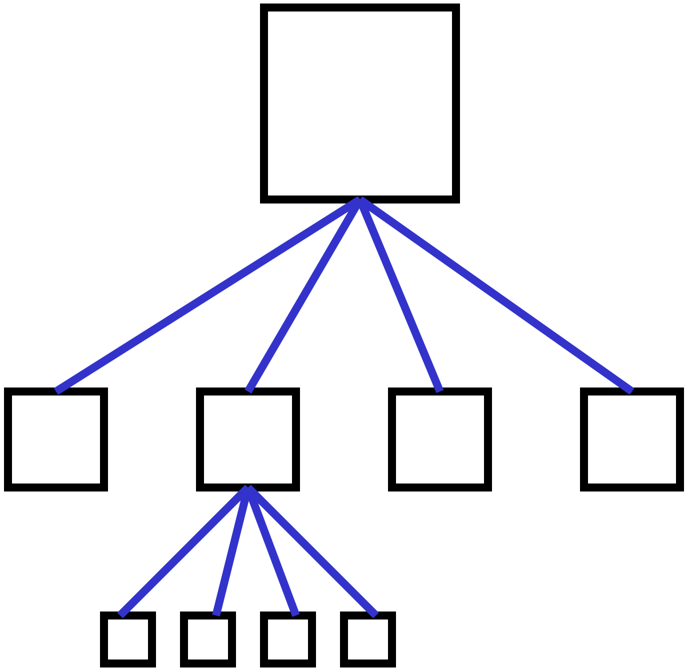


Fast estimation of bipartite matching [Agarwal,Varadarajan ’04]

Goal: Sketch point set to enable estimation of min cost matching

Tree approximations for Euclidean points





distortion $O(d \log \Delta)$ [Bartal '96, CCGGP '98]

EMD approximation [C'02, Indyk,Thaper '03]

- Construct vector from recursive decomposition
- Coordinate for each region in decomposition
  - number of points in the region
- L_1 difference of vectors for P and Q gives estimate of EMD(P,Q)

Image Similarity: Matching Sets of Features [Grauman, Darrell]

Pyramid match: a new similarity measure over sets of vectors that efficiently forms an implicit partial matching

- linear time complexity
- positive-definite function (a **kernel**)

Demonstrated effectiveness for retrieval, recognition, and regression tasks with local image features

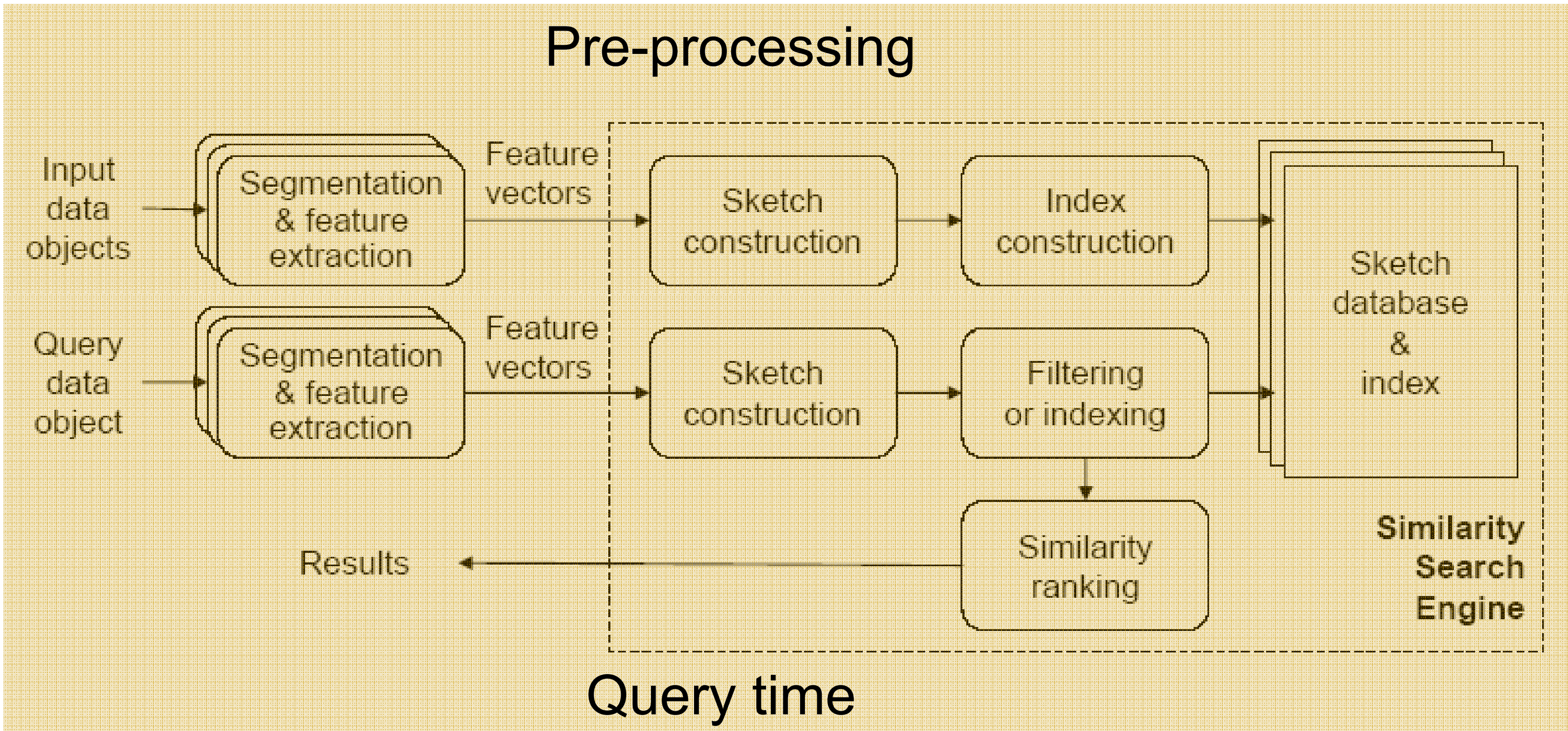
Content Based Similarity Search

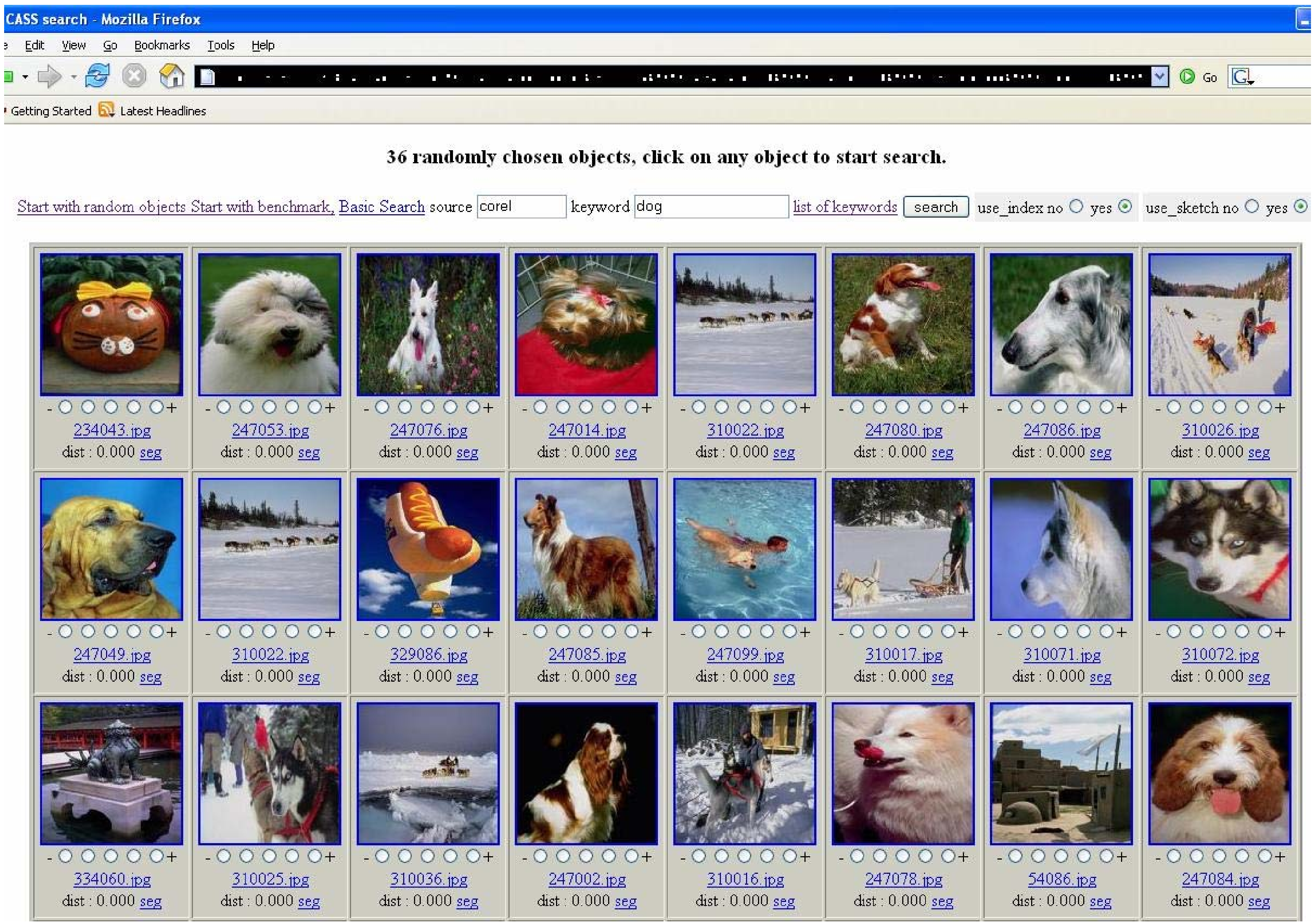
with Qin Lv, William Josephson, Zhe Wang, Perry Cook, Matthew Hoffman, Kai Li

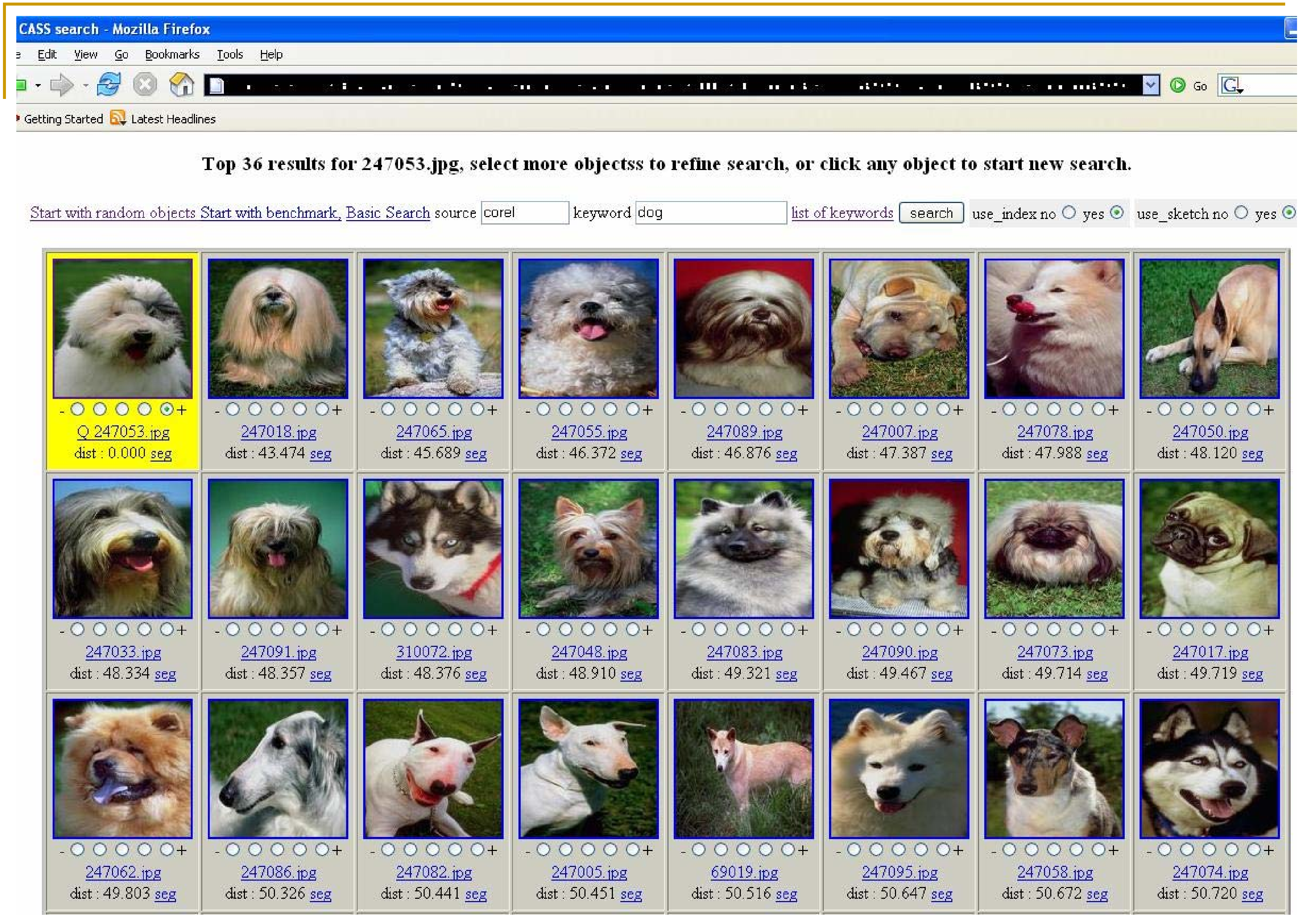
- Traditional search tools inadequate for high dimensional data
  - Exact match
  - Keyword-based search
- Need content-based similarity search

- Generic search engine for different data types
  - images, audio, 3D shapes, …

Similarity Search Engine Architecture




CASS search - Mozilla Firefox
Edit View Go Bookmarks Tools Help
[redacted]
Go
Getting Started Latest Headlines
36 randomly chosen objects, click on any object to start search.
Start with random objects Start with benchmark, Basic Search source corel keyword dog list of keywords search use_index no yes use_sketch no yes
234043.jpg dist : 0.000 seg
247053.jpg dist : 0.000 seg
247076.jpg dist : 0.000 seg
247014.jpg dist : 0.000 seg
310022.jpg dist : 0.000 seg
247080.jpg dist : 0.000 seg
247086.jpg dist : 0.000 seg
310026.jpg dist : 0.000 seg
247049.jpg dist : 0.000 seg
310022.jpg dist : 0.000 seg
329086.jpg dist : 0.000 seg
247085.jpg dist : 0.000 seg
247099.jpg dist : 0.000 seg
310017.jpg dist : 0.000 seg
310071.jpg dist : 0.000 seg
310072.jpg dist : 0.000 seg
334060.jpg dist : 0.000 seg
310025.jpg dist : 0.000 seg
310036.jpg dist : 0.000 seg
247002.jpg dist : 0.000 seg
310016.jpg dist : 0.000 seg
247078.jpg dist : 0.000 seg
54086.jpg dist : 0.000 seg
247084.jpg dist : 0.000 seg


CASS search - Mozilla Firefox
Edit View Go Bookmarks Tools Help
Go
Getting Started Latest Headlines
Top 36 results for 247053.jpg, select more objectss to refine search, or click any object to start new search.
Start with random objects Start with benchmark, Basic Search source corel keyword dog list of keywords search use_index no yes use_sketch no yes
Q 247053.jpg dist : 0.000 seg
247018.jpg dist : 43.474 seg
247065.jpg dist : 45.689 seg
247055.jpg dist : 46.372 seg
247089.jpg dist : 46.876 seg
247007.jpg dist : 47.387 seg
247078.jpg dist : 47.988 seg
247050.jpg dist : 48.120 seg
247033.jpg dist : 48.334 seg
247091.jpg dist : 48.357 seg
310072.jpg dist : 48.376 seg
247048.jpg dist : 48.910 seg
247083.jpg dist : 49.321 seg
247090.jpg dist : 49.467 seg
247073.jpg dist : 49.714 seg
247017.jpg dist : 49.719 seg
247062.jpg dist : 49.803 seg
247086.jpg dist : 50.326 seg
247082.jpg dist : 50.441 seg
247005.jpg dist : 50.451 seg
69019.jpg dist : 50.516 seg
247095.jpg dist : 50.647 seg
247058.jpg dist : 50.672 seg
247074.jpg dist : 50.720 seg

Conclusions

- Compact representations at the heart of several algorithmic techniques for large data sets
  - Compact representations tailored to applications
  - Effective for different kinds of data