



On Low Dimensional Embeddings & Similarity Search

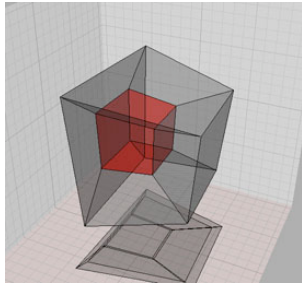
Yu-En Lu, Pietro Lió, and Steven Hand
University of Cambridge

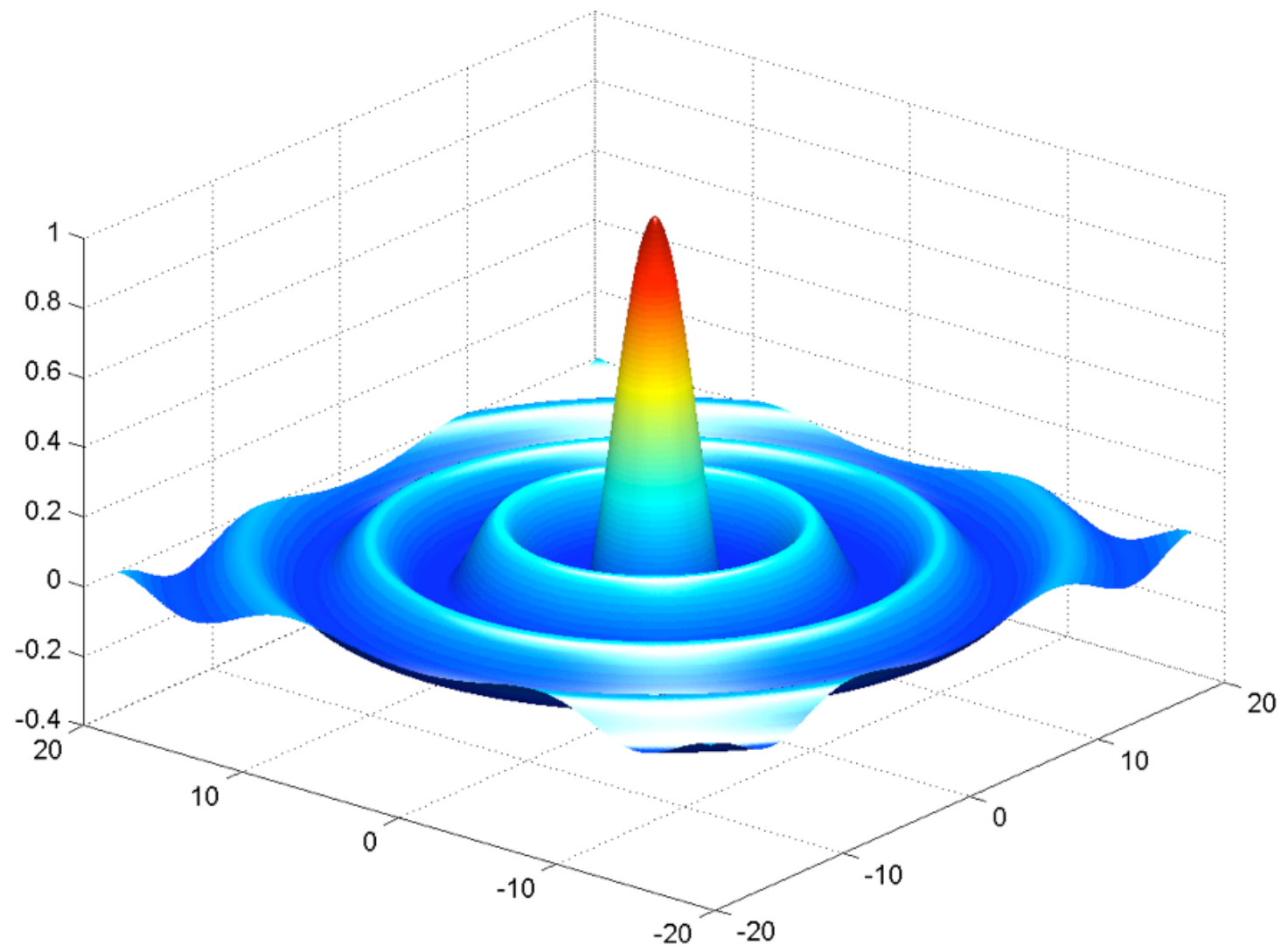
The Problem

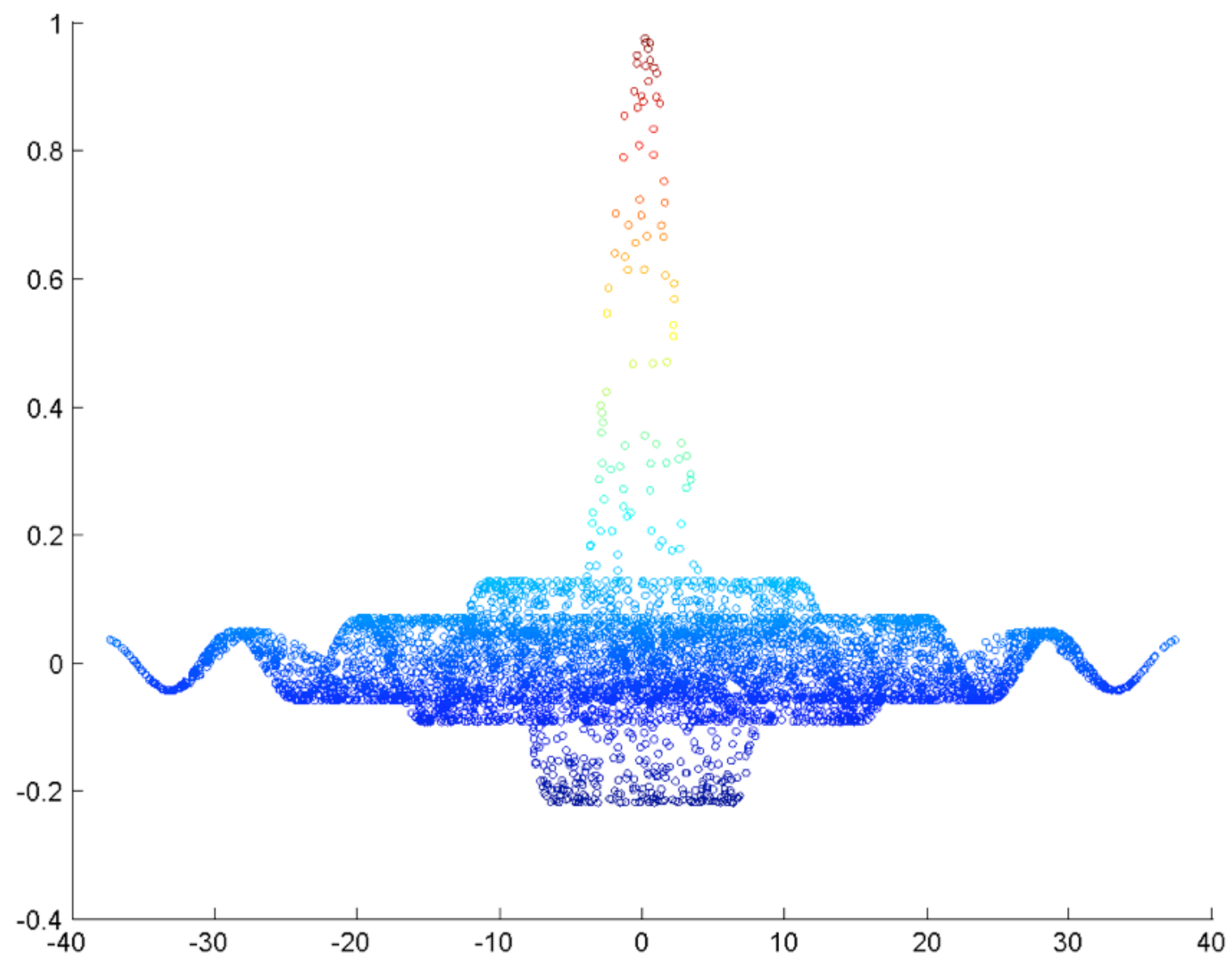
- Given n points in d -dimensional space equipped with L_2 norm, find:
- A image of k dimensions such that:
 - pair-wise distances are largely preserved in the image

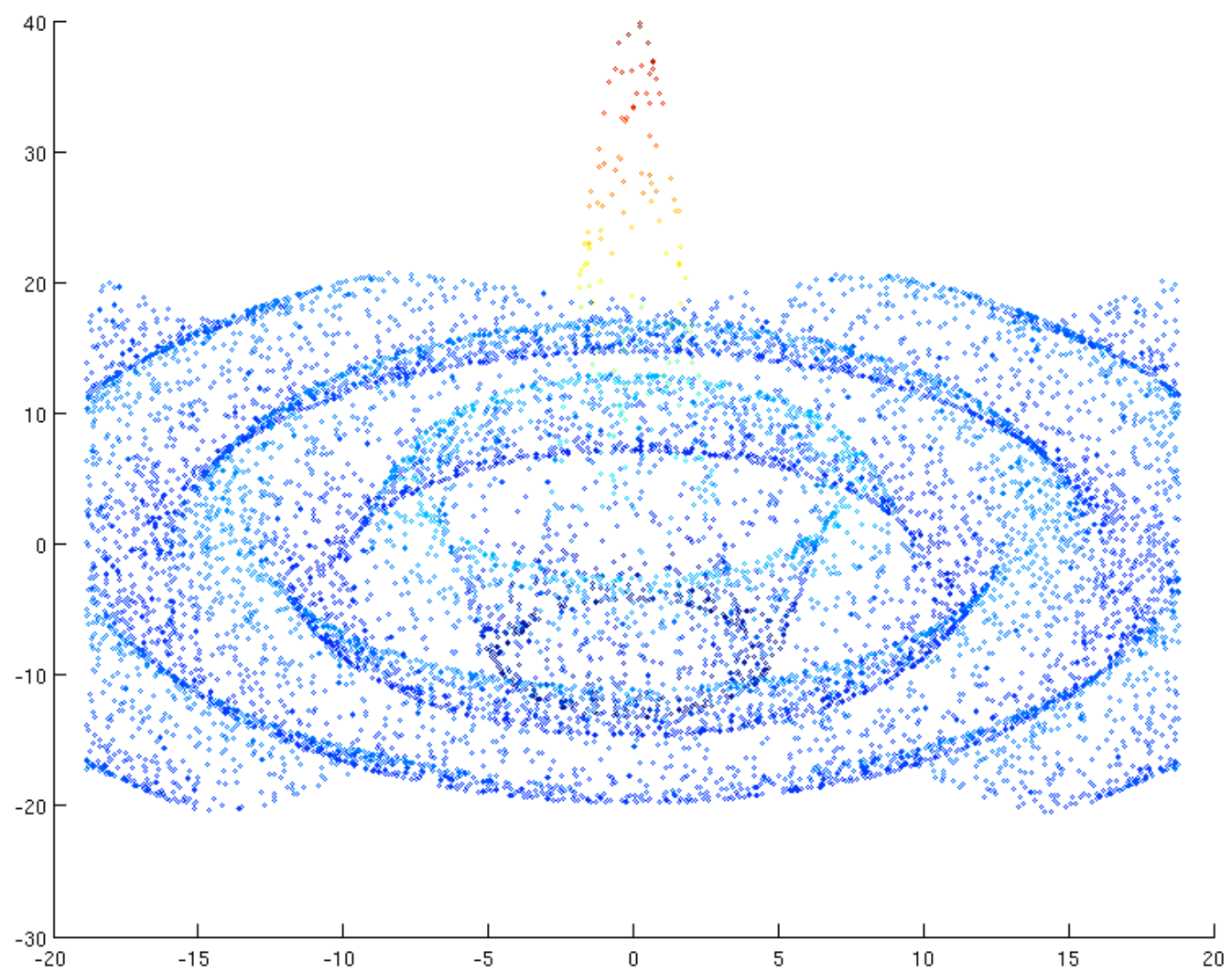
Why this matters

- Nearest neighbor queries: $\text{poly}(n,d)$ computation/storage complexity
- The tale of Qube- a multi-dimensional hypercubic overlay
- Similarity queries are fundamental primitives of any search engine









Projection: Matrix View

$$\sqrt{\frac{d}{k}} \begin{bmatrix} a_{1.} \\ a_{2.} \\ \vdots \\ a_{k.} \end{bmatrix} \times \begin{bmatrix} v_{.1} & v_{.2} & \cdots & v_{.n} \end{bmatrix}$$

$v_{\{ * j \}}$: n points in \mathbb{R}^d

$a_{\{ i * \}}$: k vectors of cardinality n

- Matrix A serve as the estimation vectors for V
- obviously picks of a_i crucial to projections

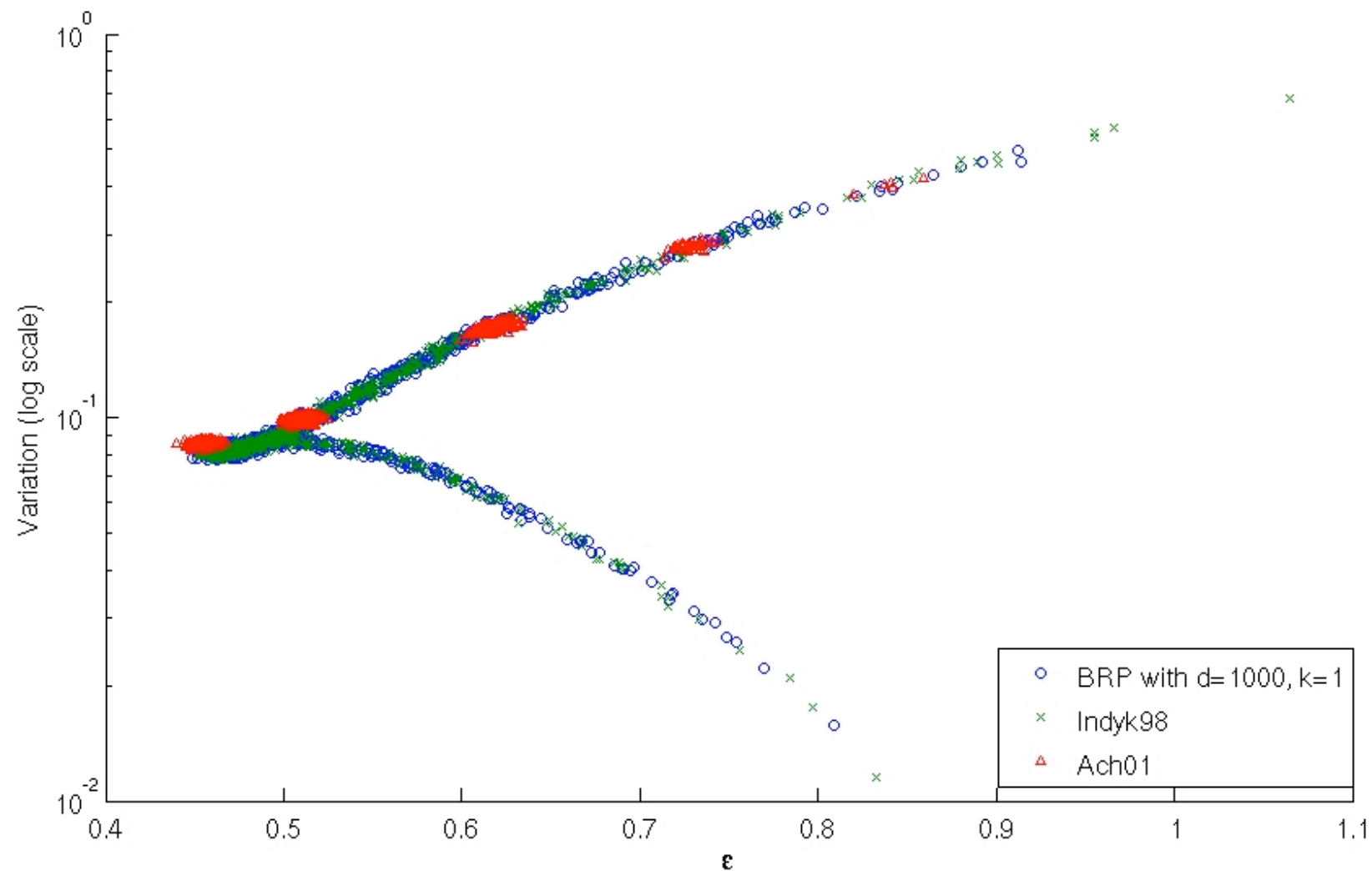
Random Projection Story

- First by Johnson & Lindenstrauss in 80s
- Indyk98 showed $N_d(0,1)$ can do the trick (Gaussian ensembles)
- Achlioptas01 used mildly sparse distributions (1 / 3 computation only)
- Li05 said very sparse ones can do too!

This Talk in 1 Slide

- Sparse is not good for you
- The devil is in the $O(\cdot)$
- Up to 40% reduction in that $O(\cdot)$ is possible for the same distortion
- And that's not just L_2 , cosine too!
- Show this on TREC (130K / 170K corpus) & Flickr images (250K corpus)

? Simplify \neq Neglect ?



The Formal Bits

THEOREM 2.1. *For any $u, v \in S \subset \mathcal{R}^d$ and $|S| = n$, an random instance of H_A in equation 1 where each row A_i is a random vector from the unit d -sphere yield distortion*

$$(1 - \epsilon)|u - v|^2 \leq |H_A(u) - H_A(v)|^2 \leq (1 + \epsilon)|u - v|^2$$

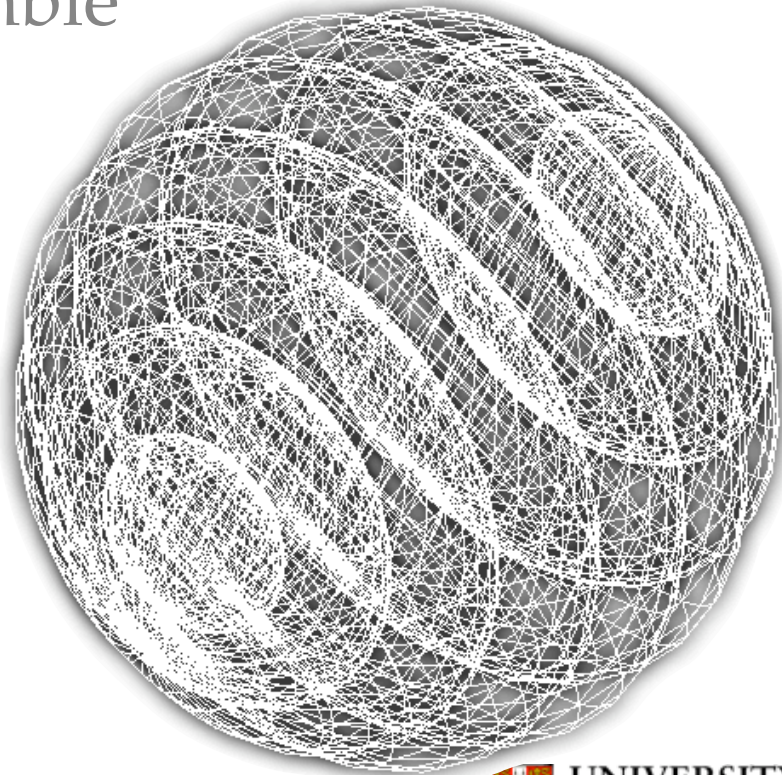
with probability at least $1 - n^{-1}$, when

$$k \geq \begin{cases} \frac{8 \log n - 2 \log 4\pi}{\epsilon^2} & \text{for } k \geq 30, \sqrt{k}\epsilon \geq 2 \\ \frac{6 \log n}{\epsilon^2(1 - 2/3\epsilon)} & \text{for } 30 > k > 0, k(1 - \epsilon) \geq 1, \\ & 0 \leq \epsilon \leq 1, n \geq 10 \end{cases}$$



The Main Theorem

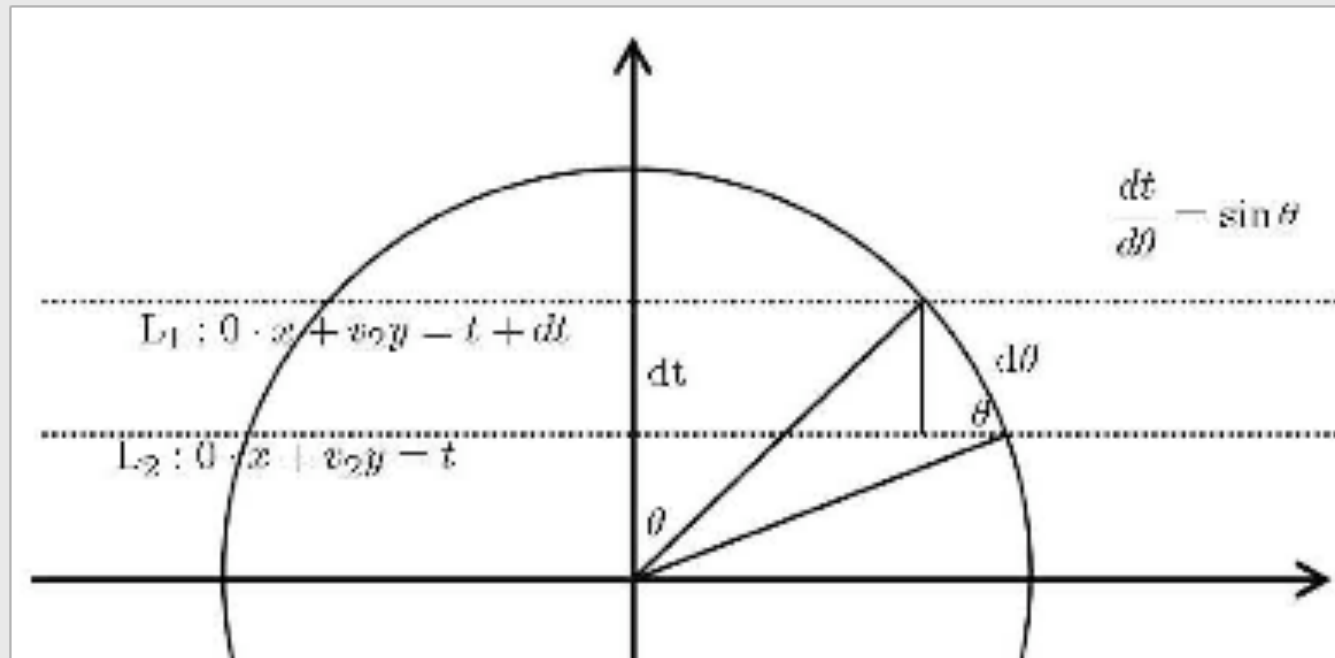
- If we pick a random matrix from the unit circular ensemble
- Then with probability $1 - 1/n$
- you get a “good” projection with distortion ε



Proof Strategy

- For each $a_i \cdot v$: show that distortion obeys a Beta distribution
- For $k < 30$, use Beta approximation due to Johannesson & Giri
- For $k > 30$, invoke Central Limit Thm

How we get Beta



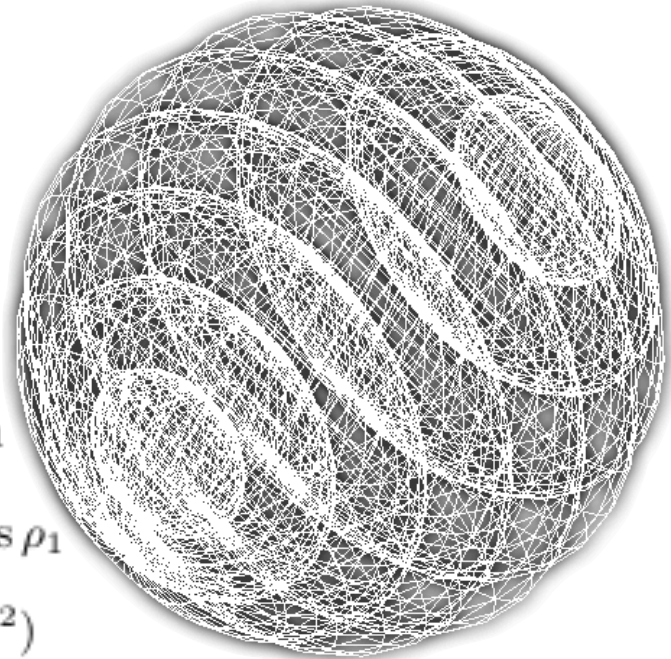
How we get Beta

LEMMA 3.1. *Let X be an uniformly random point on the surface of the unit d -dimension sphere and $v \in \mathcal{R}^d$. Then, we have*

$$\frac{(X \cdot v)^2}{|v|^2} \sim \beta\left(\frac{1}{2}, \frac{d-1}{2}\right)$$

where $\beta(\alpha, \beta)$ is the beta distribution.

$$\begin{aligned} \Pr[|X \cdot v| \leq t] &= \Pr[|X \cdot v| \leq \cos \rho_1] \\ &= \frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} 2 \int_0^{\rho_1} \sin^{d-2}(\rho_1) d\rho_1 \\ &= \frac{1}{\beta(\frac{1}{2}, \frac{d-1}{2})} \int (1 - \cos^2 \rho_1)^{\frac{d-3}{2}} d \cos \rho_1 \\ &= \frac{1}{\beta(\frac{1}{2}, \frac{d-1}{2})} 2t \cdot {}_2F_1\left(\frac{1}{2}, -\frac{d-3}{2}; \frac{3}{2}; t^2\right) \\ &= \frac{1}{\beta(\frac{1}{2}, \frac{d-1}{2})} \beta_{t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right) \\ &= I_{t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right) \end{aligned}$$





Business

- » Advertising
- » Autos
- » The Biz
- » Investing
- » Real Estate



INVESTING TIPS AND TOOLS

Quote:

Symbol or Name

Go

Finance Tools

- Home Equity
- Stock, Fund Quotes & Charts
- Key Media Stocks
- Currencies



Evaluation: Text

- » LA Land
- » Up to Speed

Business Tools

- » Business A-Z
- » Investor Tips
- » Law Resources
- » Money Library
- » Money Q & A

continue

By Martin Zimmerman | 1:36 PM PDT

The Dow closes 203 points lower after a burst of late selling. Investors are rattled after steep drops in overseas markets, especially in Asia, where Japan's Nikkei 225 index falls to a 26-year low.

Blog: IHOP, Applebee's parent gets a lift from move to sell stores

BUSINESS: LATEST AP NEWS

7/1 ARM

6.23%

6.45%

Powered by Interest.com



This is not just a car.
It's a vision of our future.



UNIVERSITY OF
CAMBRIDGE

TREC Corpus

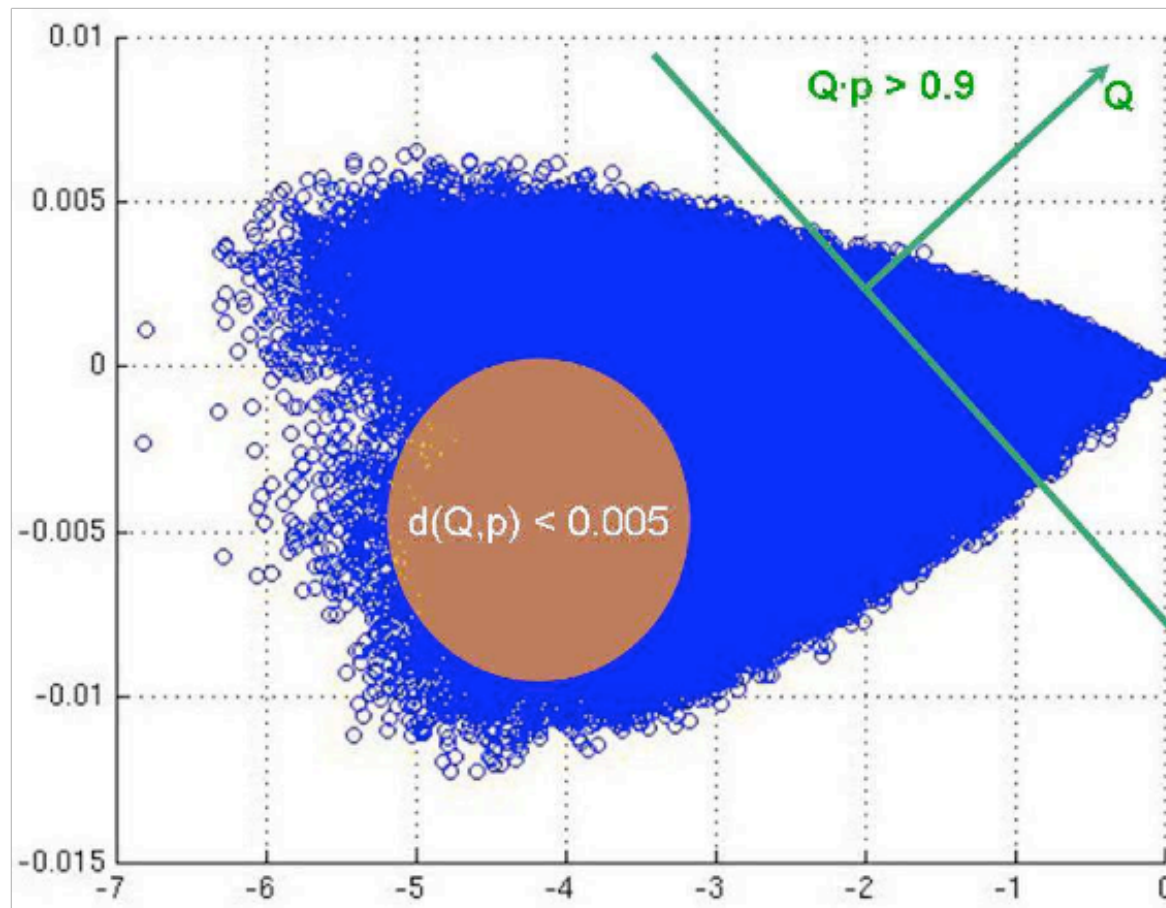


- TFIDF extraction via LEMUR toolkit
- Foreign Broadcast Information Service
- LA Times

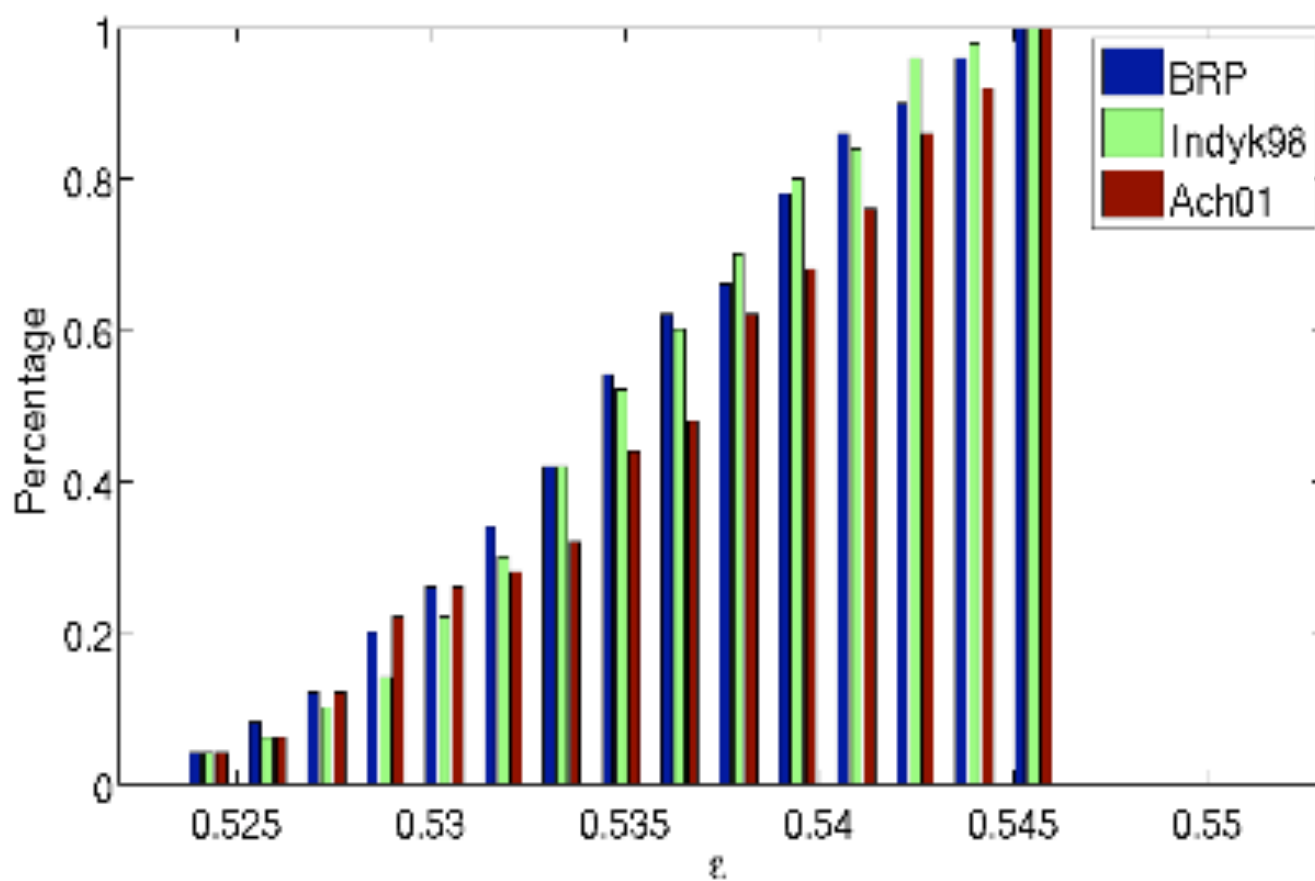
Evaluation Method

- Instantiate 10K instances of projections for both LA Times and FBIS, check mean and variance of:
 - L2 distance distortion
 - Cosine distortion (with SVD)
 - Agreement with latent semantic indexing (LSI)

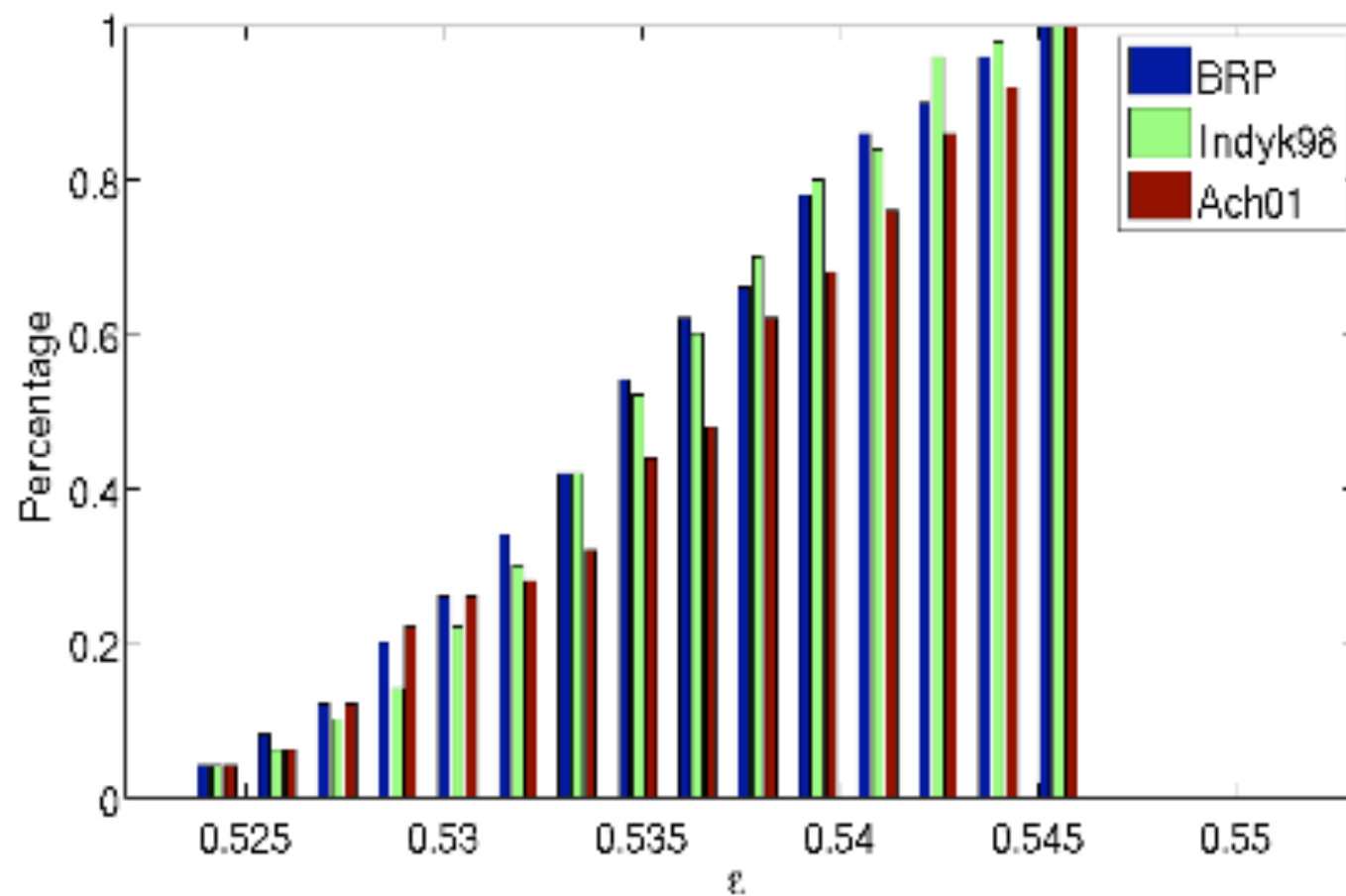
LA Times by LSI



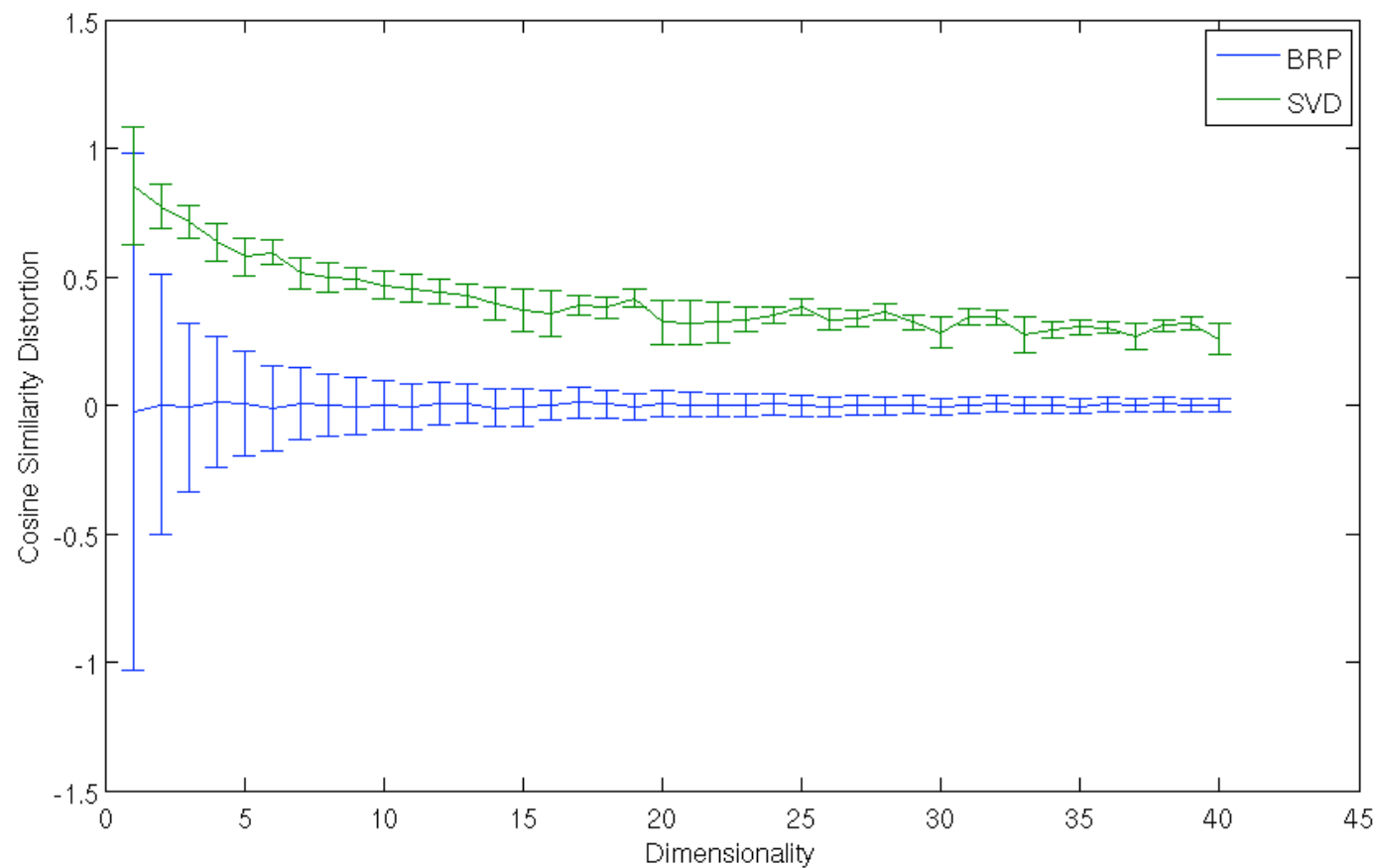
FBIS: L2 distance



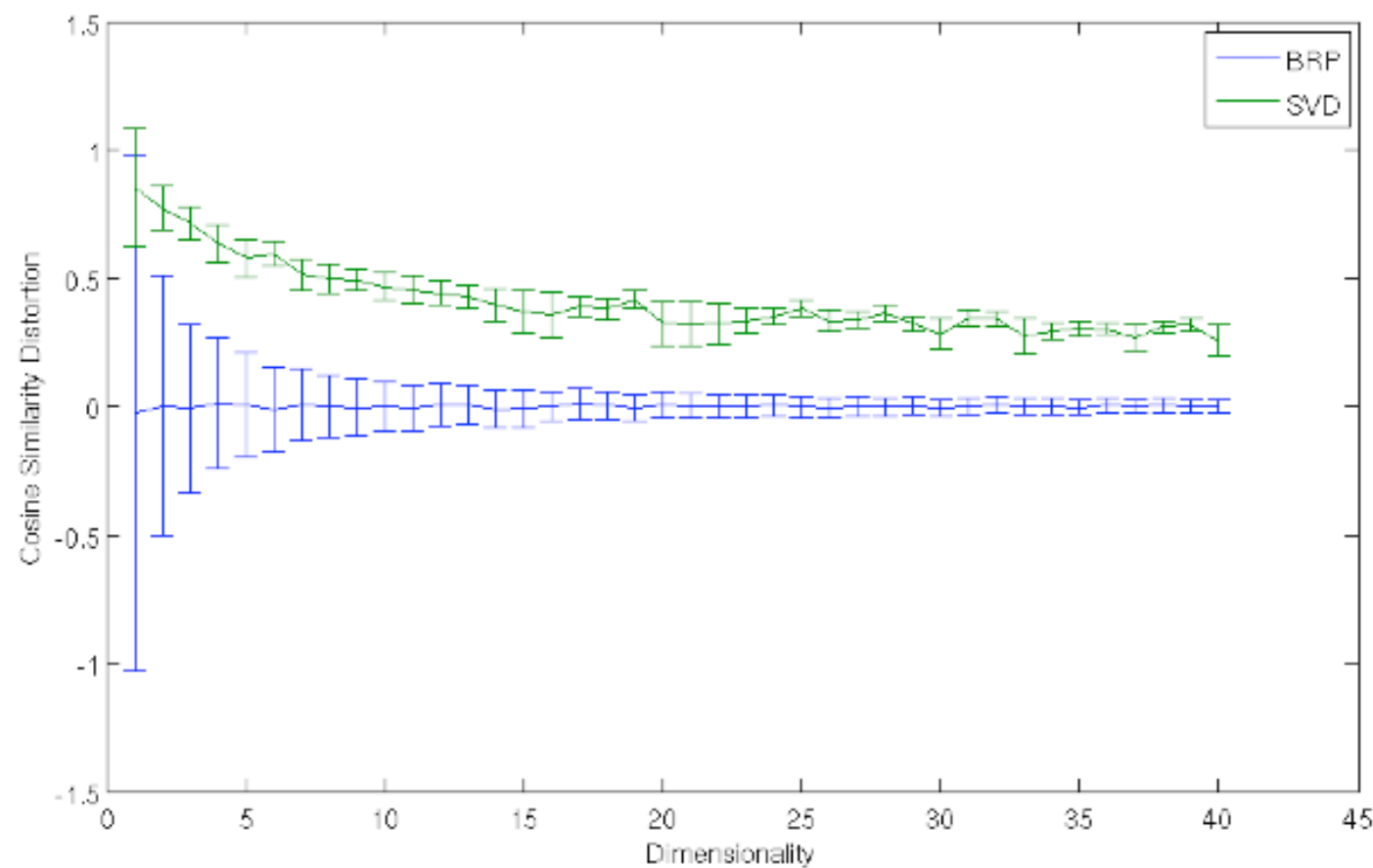
LA Times: L2 dist



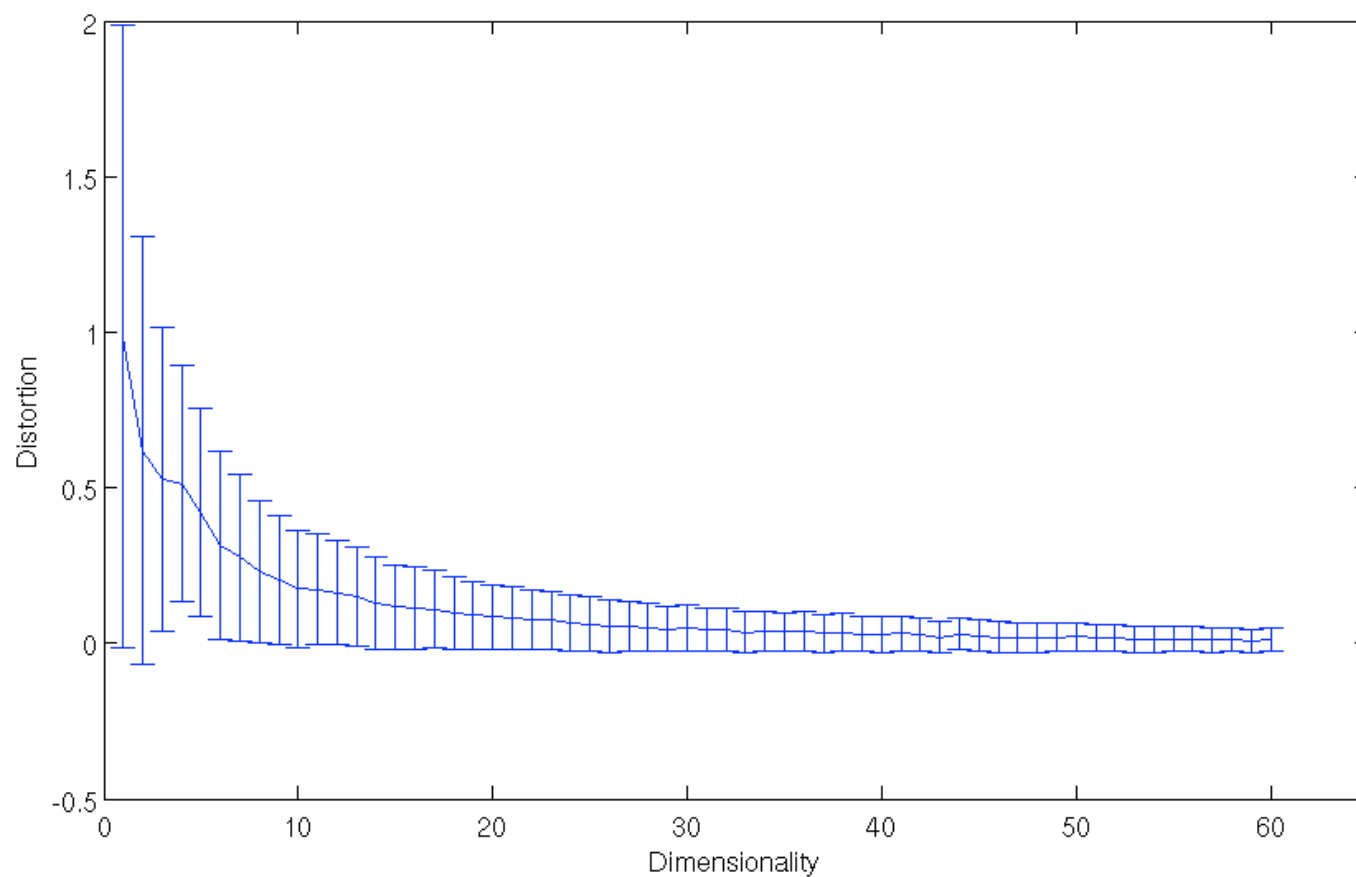
FBIS: Cosine



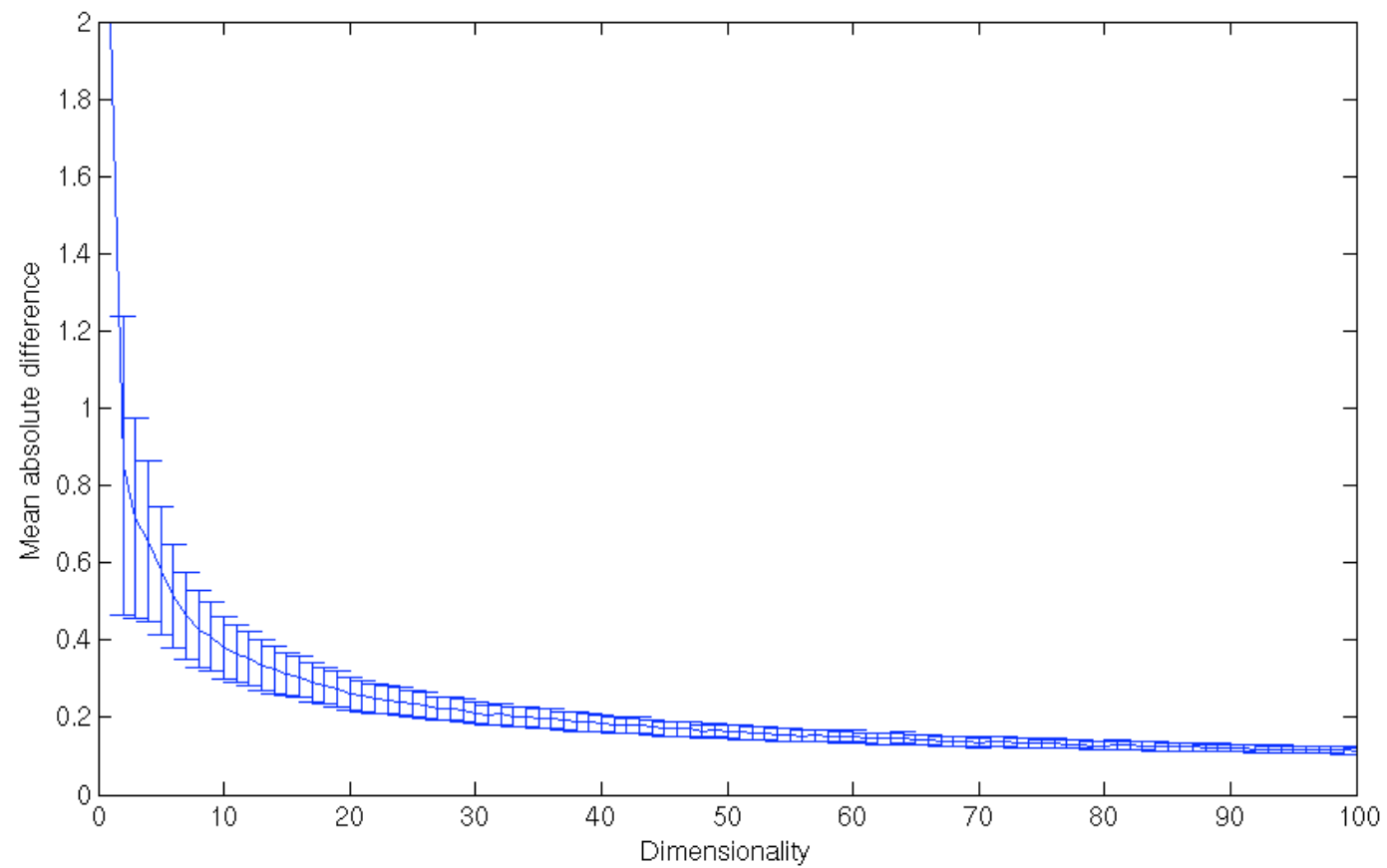
LA Times: Cosine



LA Times: LSI



FBIS: LSI



Picasa Web Albums - E

◀ ▶ ↺ ✂ +

http://picasaweb.google.com/lu.yuen





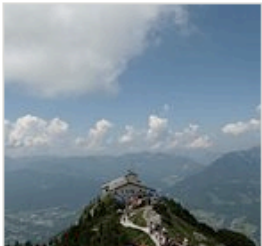
RSS Google






Apple Yahoo! Google Maps YouTube Wikipedia News (272) Popular Research

ITA | International Tech... The Lemur Toolkit Picasa Web Albums - E

My Photos > My Public Gallery Albums (16)

Share










2007-10-13-18, SOSP, ...
(38) 📍
Oct 14, 2007


2007-09-24-28, Wien, ...
(27) 📍
Sep 25, 2007

2007-09-15, Norfolk (11) 📍
Sep 15, 2007

2007-08-30-09-08, Spain
(94) 📍
Aug 31, 2007

2007-07-21, Botanic G...
(40) 📍
Jul 21, 2007



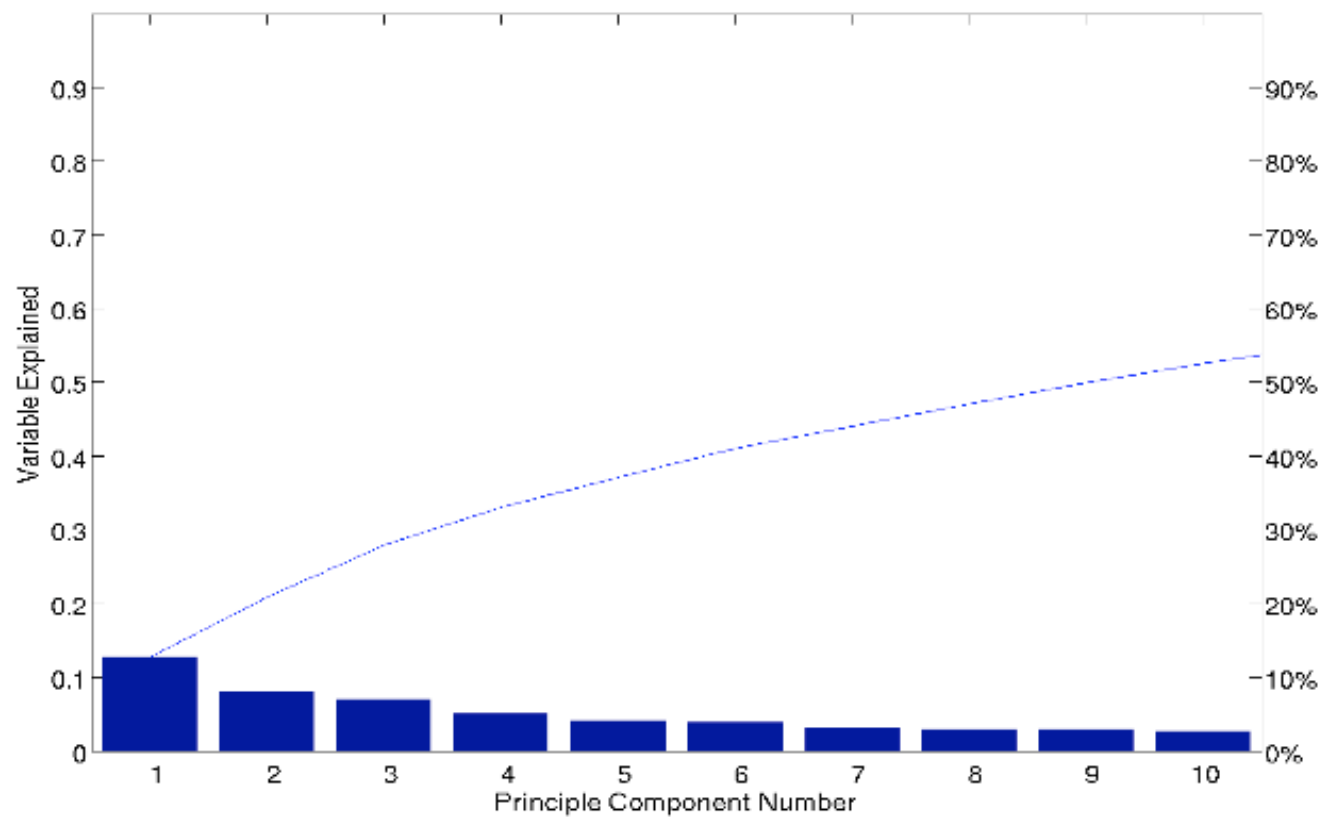
 UNIVERSITY OF
CAMBRIDGE

Evaluation: Images

Evaluation Method

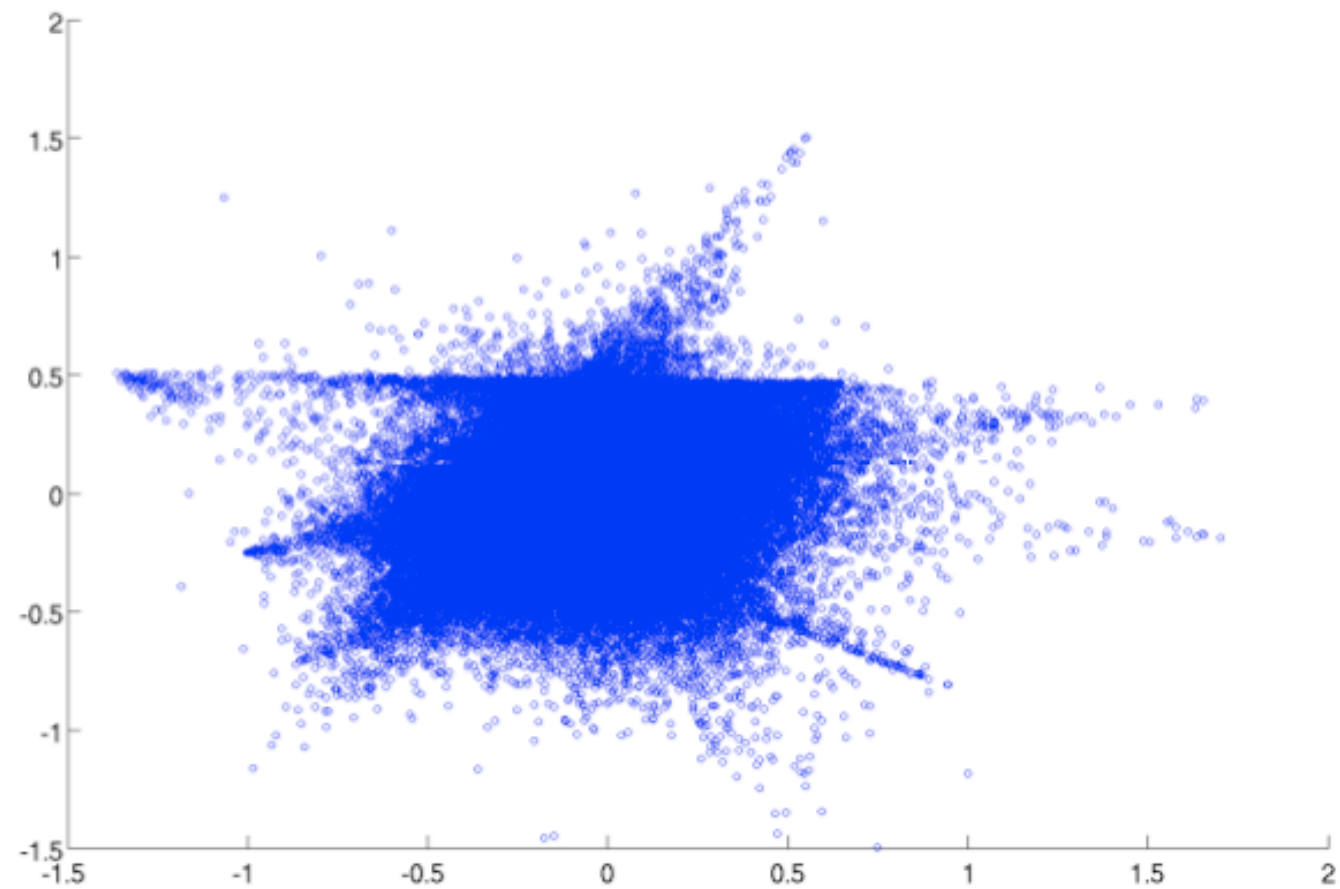
- Again, measure cosine and L2 distance distortions to features of images from:
 - The author's fotos (color histogram)
 - A flickr.com crawl of 250K images with 166 standard features: saturation, texture...etc.

PCA of Images

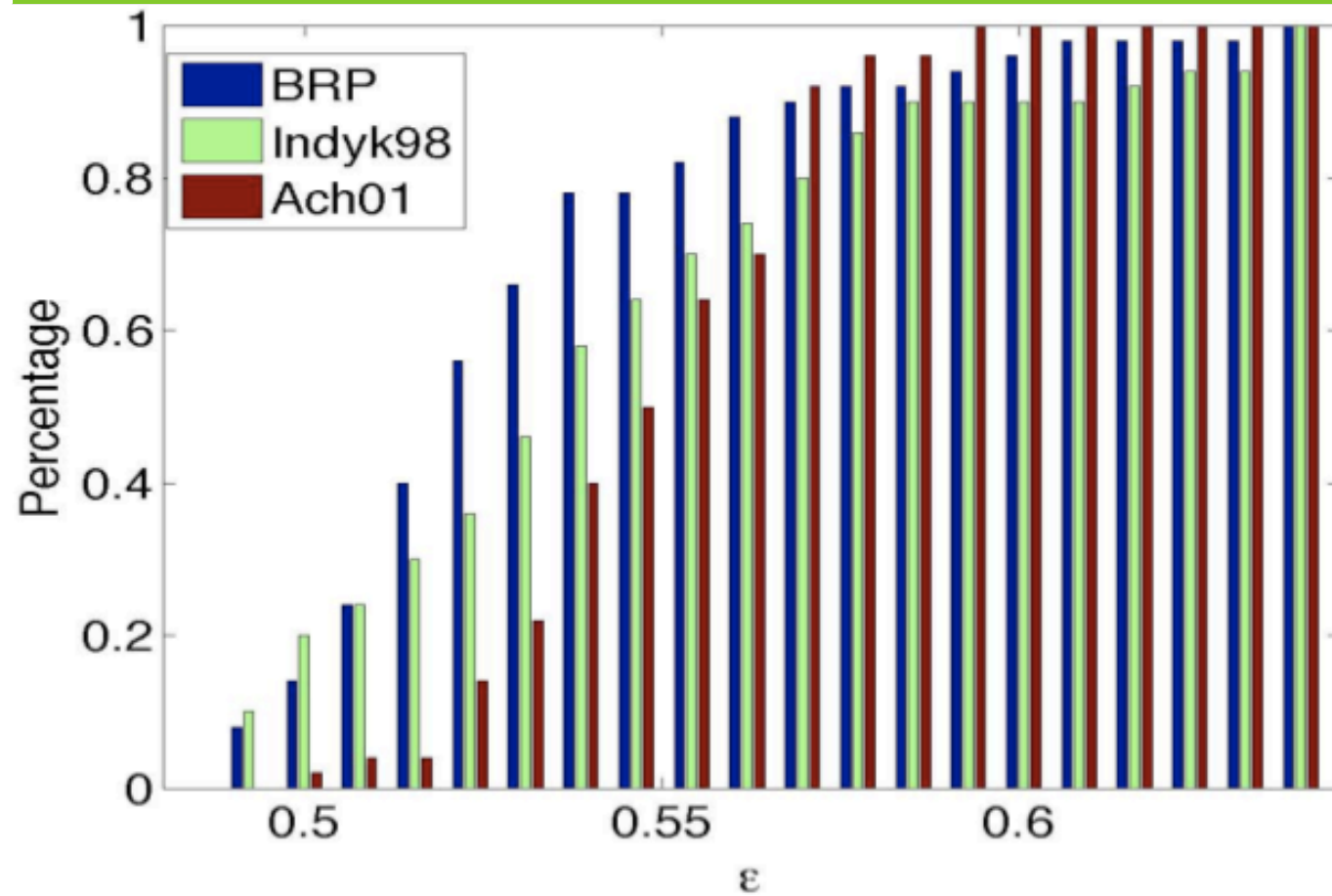


(b) PCA Analysis of the flickr dataset.

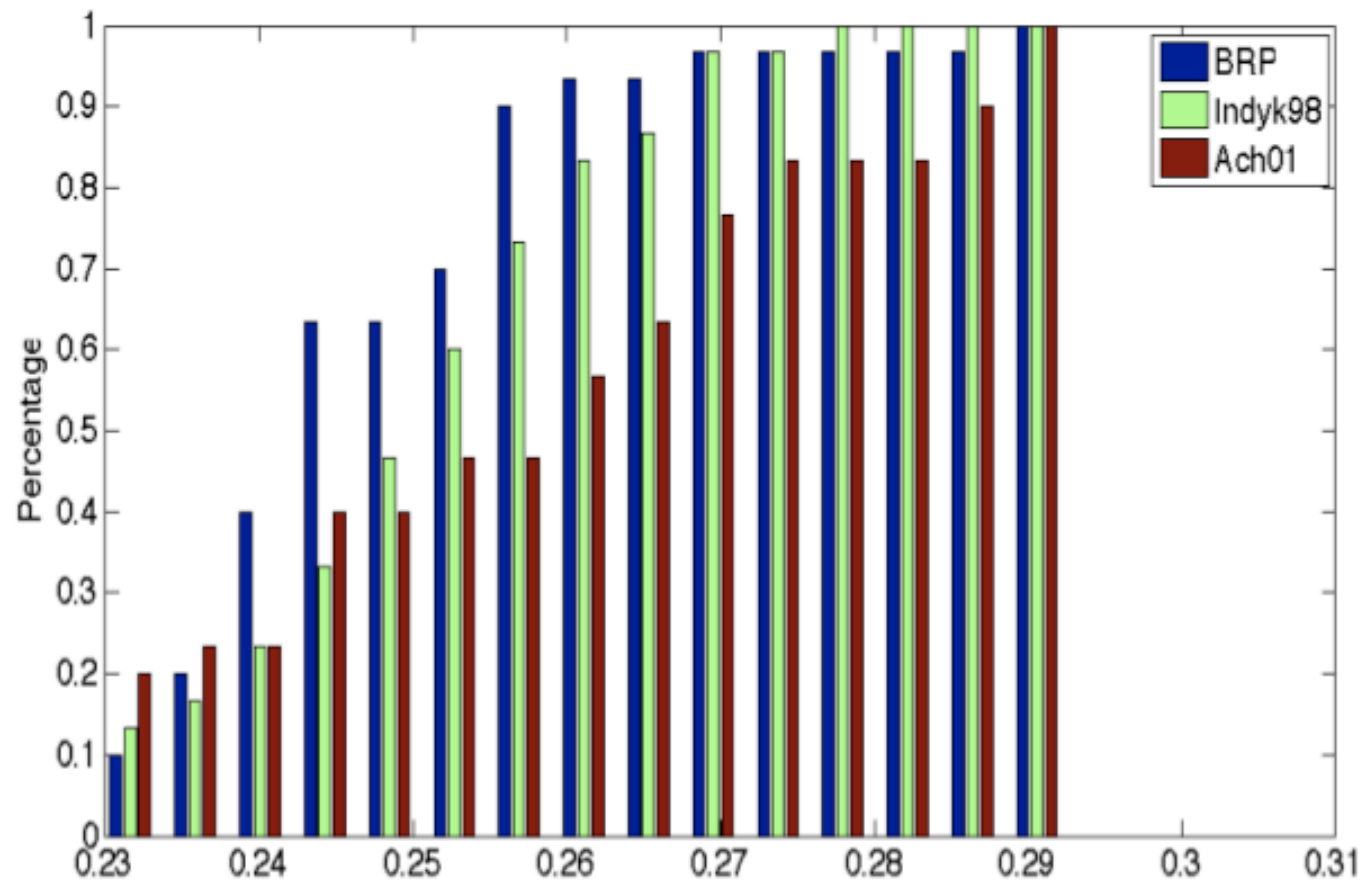
An Image of Images



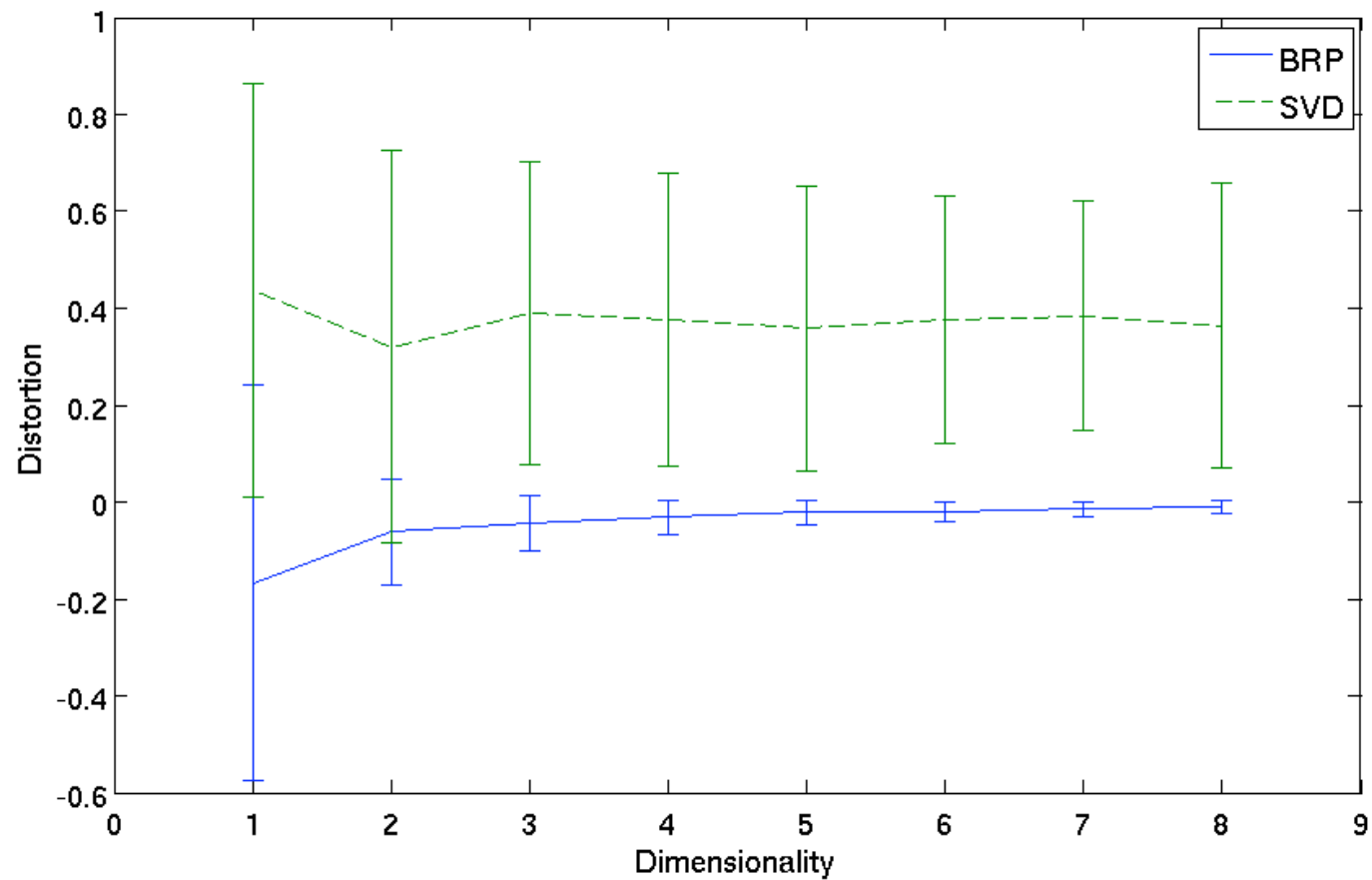
Distortion $k=1$



Distortion: $k=5$



BRP vs SVD

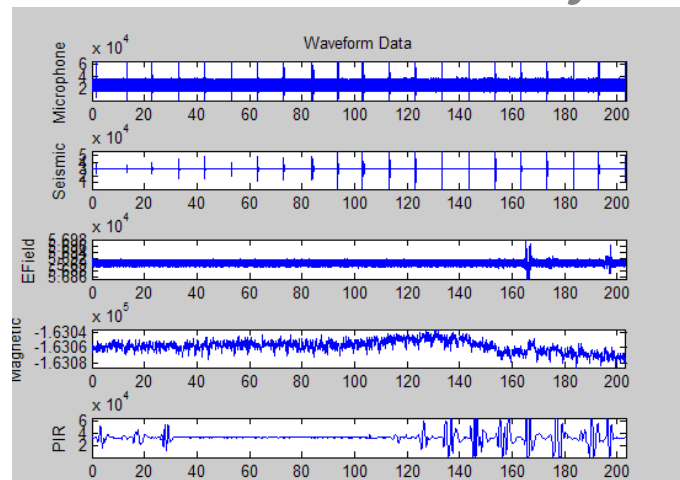


Conclusion

- Random projections are sharp up to L2 and cosine similarity measures
- Beta random projection
 - Picking circular ensembles from random matrices would suffice
 - consistently good performance!

Future Work

- Multi-dimensional sensory fusion
- Streaming DBs
- Curious results in computational chemistry



Acknowledgements

- Cambridge NLP group

(N1/n1_n
(N1/n1_n1-coord
(N1/n1
(N1/ap_n1
(AP/a1 (A1/a Natural_JJ) (N1/n Language_NN1)))
(N1/cj-end_n1 and_CC (N1/n Information_NN1)))
Processing_NN1)
Group_NN1)



• Daniel Blank & the Media group in University of Bamberg

- US-UK ITA for partly sponsoring this research



• Anonymous referees for useful comments and feedbacks

Thank you! Questions?

Yu-En.Lu@cl.cam.ac.uk

Introduction

