LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures

Keogh, Wei, Xi, Lee &Vlachos

Come, we shall learn of the indexing of shapes

*Set forth these figures as I have conceived their shape…**

Outline of Talk

- The utility of shape matching
- Shape representations
- Shape distance measures
- Lower bounding rotation invariant measures with the LB_Keogh
- Accuracy experiments
- Efficiency experiments
- Conclusions

*Paradise Canto XVIII 85

The Utility of Shape Matching I

…discovering insect mimicry, clustering petroglyphs, finding unusual arrowheads, tracking fish migration, finding anomalous fruit fly wings…

1st Discord (Castroville Cornerstang)

2nd Discord (Martindale point)

1st Discord

Drosophila melanogaster

The Utility of Shape Matching II

…automatically annotating old manuscripts, mining medical images, biometrics, spatial mining of horned lizards, indexing nematodes…

1st Discord

Shape Representations I

*For virtually all shape matching problems, rotation is **the** problem*

If I asked you to group these reptile skulls, rotation would not confuse you

There are two ways to be rotation invariant

1) Landmarking: Find the one "true" rotation
2) Rotation invariant features

Landmarking

Best Rotation Alignment

Owl Monkey (species unknown)

Owl Monkey Northern Gray-Necked

Orangutan

• **Domain Specific Landmarking**
Find some fixed point in your domain, eg. the nose on a face, the stem of leaf, the tail of a fish …

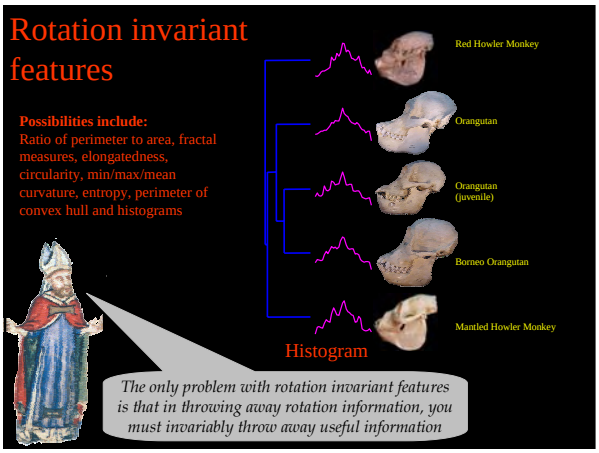
• **Generic Landmarking**
Find the major axis of the shape and use that as the canonical alignment

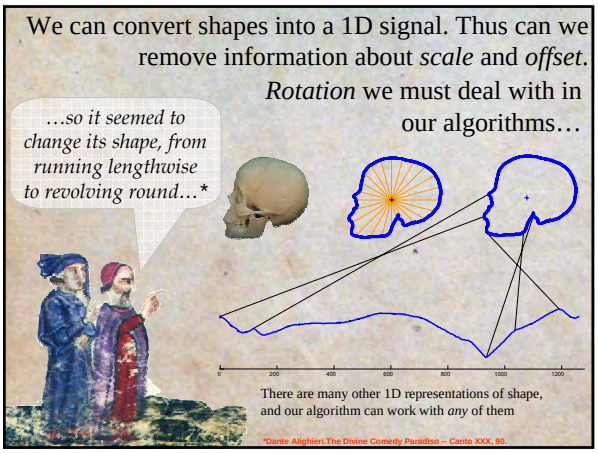
Generic Landmark Alignment

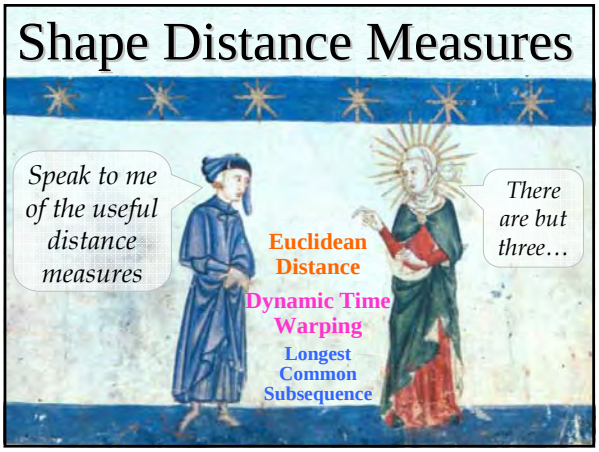
Generic Landmark Alignment

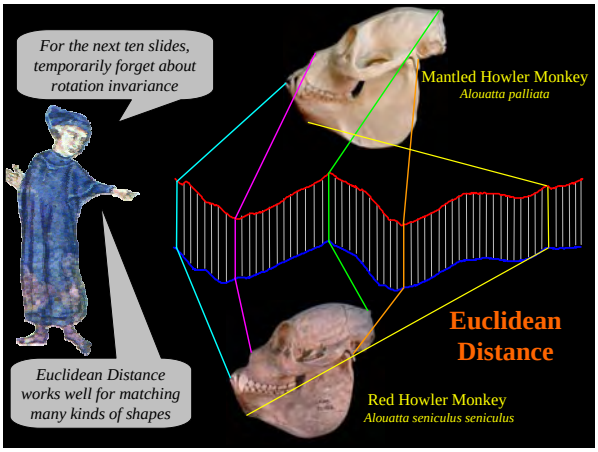
Best Rotation Alignment

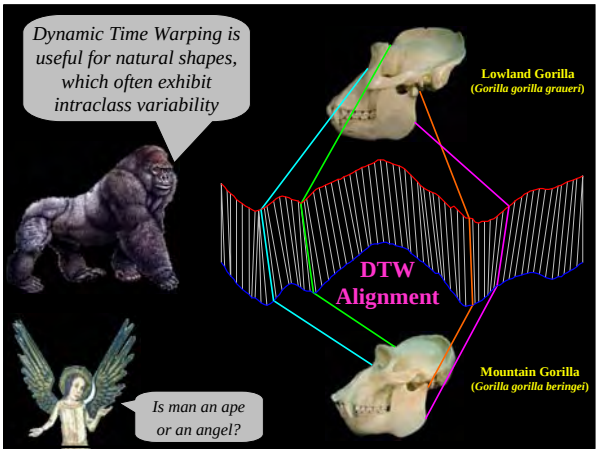
The only problem with landmarking is that it does not work

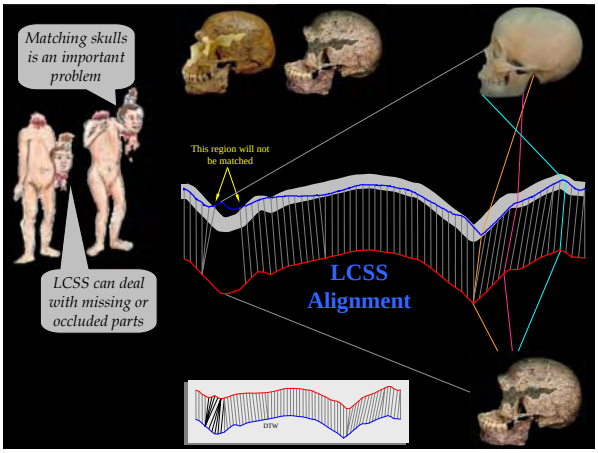

Rotation invariant features
Possibilities include:
Ratio of perimeter to area, fractal measures, elongatedness, circularity, min/max/mean curvature, entropy, perimeter of convex hull and histograms
Red Howler Monkey
Orangutan
Orangutan (juvenile)
Borneo Orangutan
Mantled Howler Monkey
Histogram
The only problem with rotation invariant features is that in throwing away rotation information, you must invariably throw away useful information

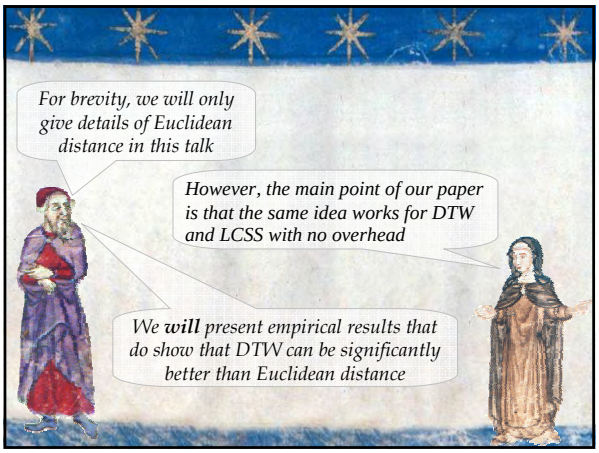

We can convert shapes into a 1D signal. Thus can we remove information about *scale* and *offset*. *Rotation* we must deal with in our algorithms…
…so it seemed to change its shape, from running lengthwise to revolving round…*
There are many other 1D representations of shape, and our algorithm can work with *any* of them
*Dante Alighieri.The Divine Comedy Paradiso – Canto XXX. 90.

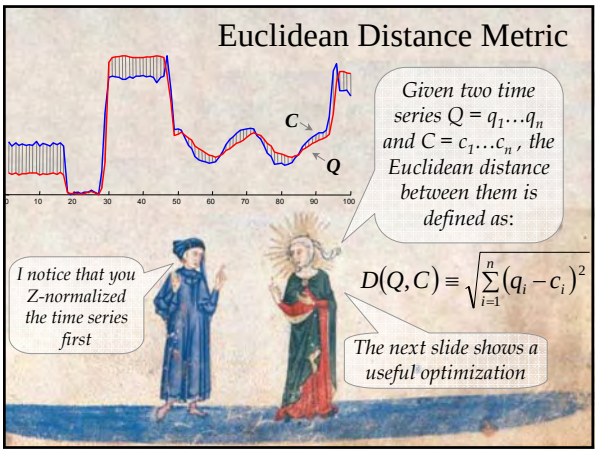

Shape Distance Measures
Speak to me of the useful distance measures
There are but three…
Euclidean Distance
Dynamic Time Warping
Longest Common Subsequence

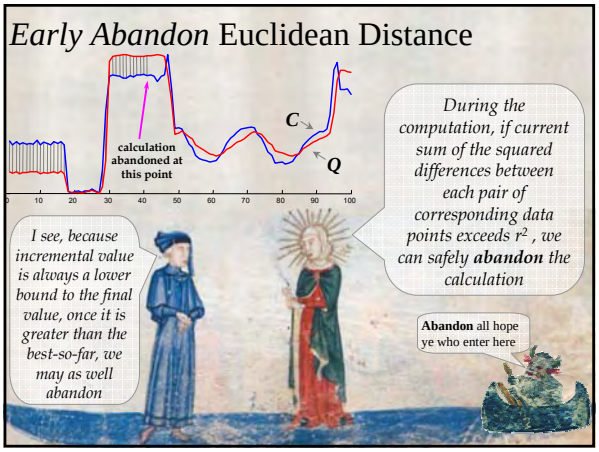

For the next ten slides, temporarily forget about rotation invariance
Mantled Howler Monkey
Alouatta palliata
Euclidean Distance
Euclidean Distance works well for matching many kinds of shapes
Red Howler Monkey
Alouatta seniculus seniculus

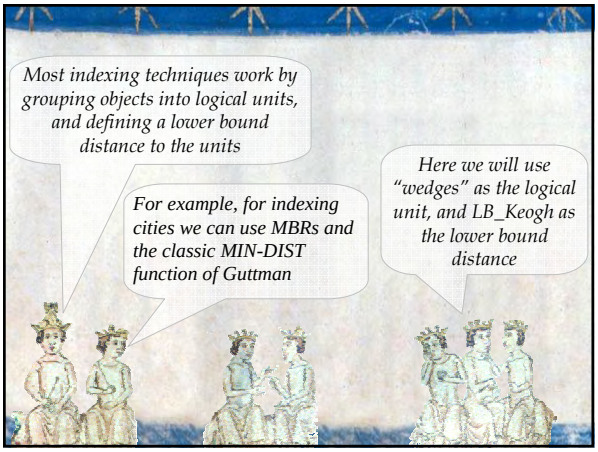

Dynamic Time Warping is useful for natural shapes, which often exhibit intraclass variability
Lowland Gorilla
(Gorilla gorilla graueri)
DTW Alignment
Mountain Gorilla
(Gorilla gorilla beringei)
Is man an ape or an angel?

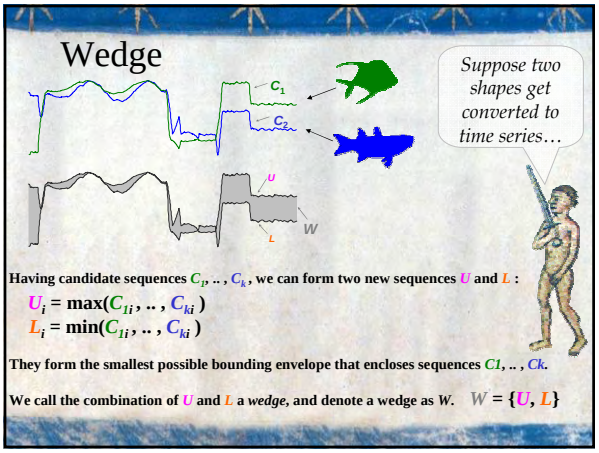

Matching skulls is an important problem
This region will not be matched
LCSS can deal with missing or occluded parts
LCSS Alignment
DTW

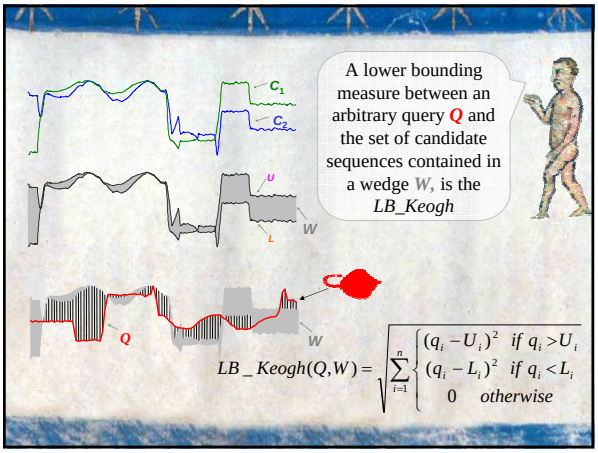

For brevity, we will only give details of Euclidean distance in this talk
However, the main point of our paper is that the same idea works for DTW and LCSS with no overhead
We **will** present empirical results that do show that DTW can be significantly better than Euclidean distance


Euclidean Distance Metric
C
Q
Given two time series $Q = q_1 \ldots q_n$ and $C = c_1 \ldots c_n$, the Euclidean distance between them is defined as:
$D(Q,C) \equiv \sqrt{\sum_{i=1}^{n} (q_i - c_i)^2}$
I notice that you Z-normalized the time series first
The next slide shows a useful optimization


Early Abandon Euclidean Distance
calculation abandoned at this point
C
Q
During the computation, if current sum of the squared differences between each pair of corresponding data points exceeds r^2, we can safely **abandon** the calculation
I see, because incremental value is always a lower bound to the final value, once it is greater than the best-so-far, we may as well abandon
Abandon all hope ye who enter here

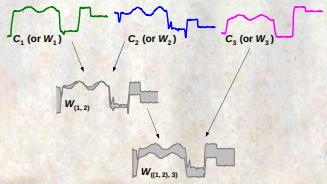

Most indexing techniques work by grouping objects into logical units, and defining a lower bound distance to the units
For example, for indexing cities we can use MBRs and the classic MIN-DIST function of Guttman
Here we will use "wedges" as the logical unit, and LB_Keogh as the lower bound distance

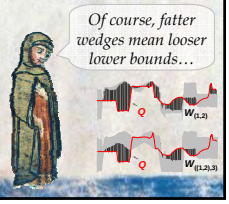

Wedge
C_1
C_2
U
L
W
Suppose two shapes get converted to time series…
Having candidate sequences $C_1, .., C_k$, we can form two new sequences U and L :
$U_i = \max(C_{1i}, .., C_{ki})$
$L_i = \min(C_{1i}, .., C_{ki})$
They form the smallest possible bounding envelope that encloses sequences $C1, .., Ck$.
We call the combination of U and L a *wedge*, and denote a wedge as W. $W = \{U, L\}$

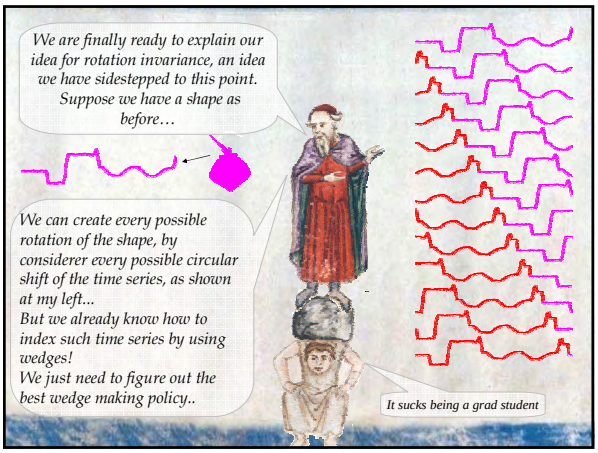

C_1
C_2
U
L
W
A lower bounding measure between an arbitrary query **Q** and the set of candidate sequences contained in a wedge W, is the *LB_Keogh*
Q
W
$LB\_Keogh(Q,W) = \sqrt{\sum_{i=1}^{n} \begin{cases} (q_i - U_i)^2 & \text{if } q_i > U_i \\ (q_i - L_i)^2 & \text{if } q_i < L_i \\ 0 & \text{otherwise} \end{cases}}$

Generalized Wedge

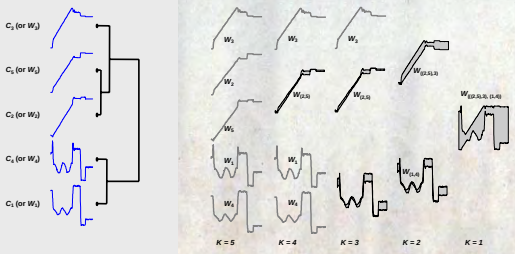
- Use $W_{(1,2)}$ to denote that a wedge is built from sequences C_1 and C_2.
- Wedges can be hierarchally nested. For example, $W_{((1,2),3)}$ consists of $W_{(1,2)}$ and C_3.



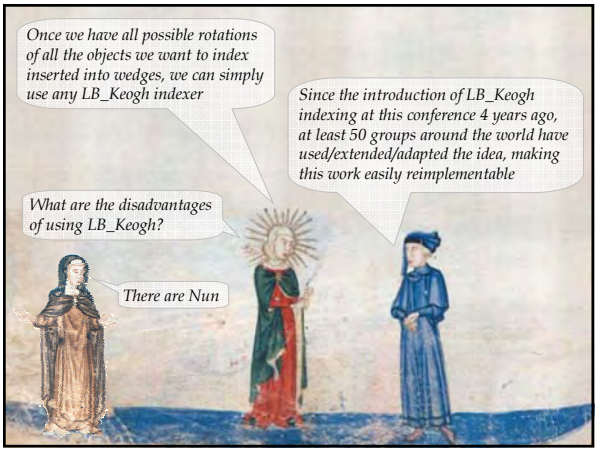


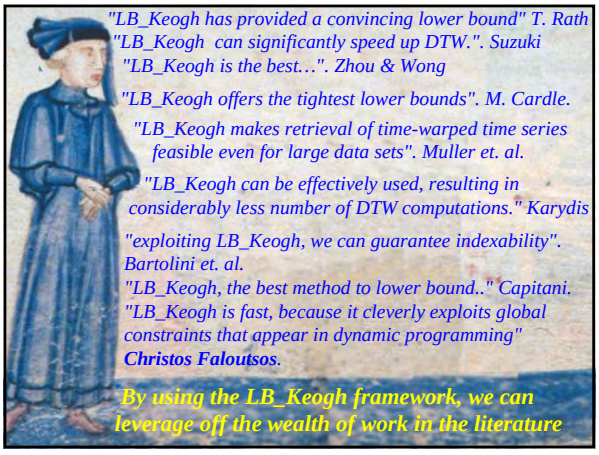


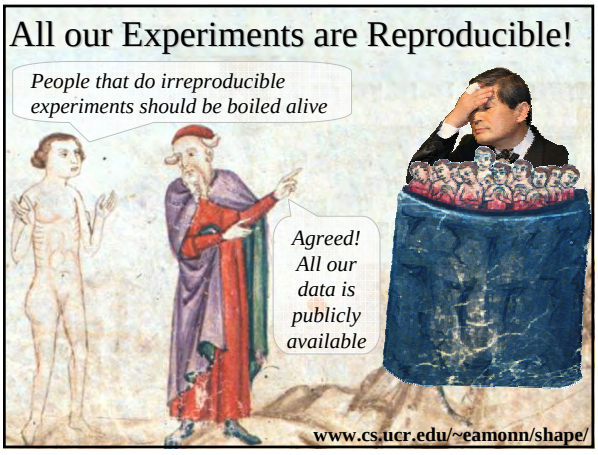
Hierarchal Clustering



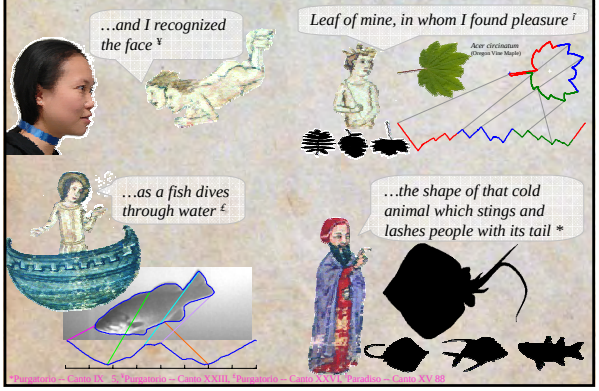
Which wedge set to choose ?







We tested on many diverse datasets



Name	Classes	Instances	Euclidean Error (%)	DTW Error (%) {R}	Other Techniques
Face	16	2240	3.839	**3.170**{3}	
Swedish Leaves	15	1125	13.33	**10.84**{2}	
Chicken	5	446	19.96	19.96{1}	20.5 Discrete strings
MixedBag	9	160	4.375	4.375{1}	Chamfer 6.0, Hausdorff 7.0
OSU Leaves	6	442	33.71	**15.61**{2}	
Diatoms	37	781	27.53	27.53{1}	26.0 Morphological Curvature Scale Spaces
Plane	7	210	0.95	**0.0**{3}	0.55 Markov Descriptor
Fish	7	350	11.43	**9.71**{1}	36.0 Fourier /Power Cepstrum

Note that DTW is sometimes worth the little extra effort

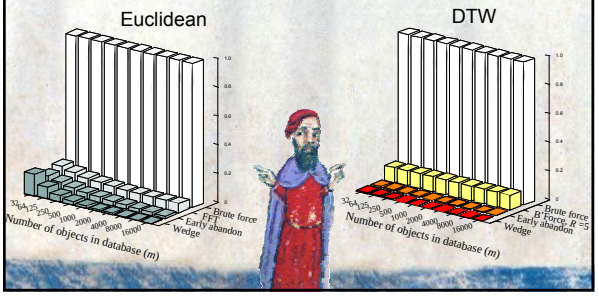
Implementation details should not matter, for example the results reported should be the same if reimplemented in Ret Hat Linux

We therefore use a cost model that is independent of hardware/software/buffer size etc. See the paper for details

We compare to brute force, and were possible a Fourier based approach (it can't handle DTW)

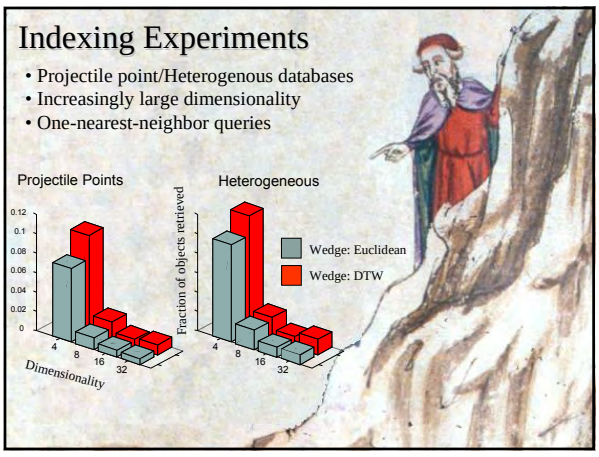
Main Memory Experiments

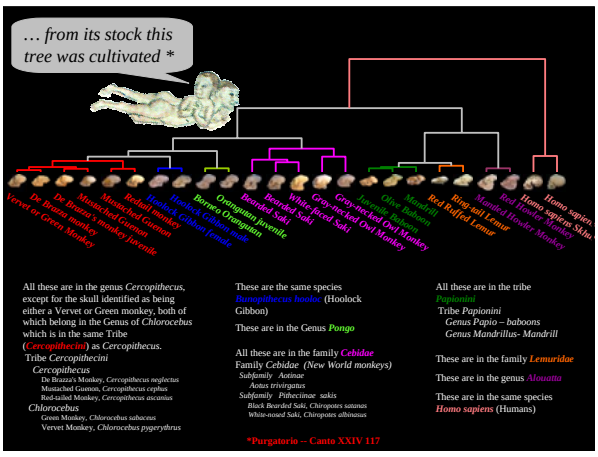
- Projectile point database
- Increasingly larger datasets
- One-nearest-neighbor queries

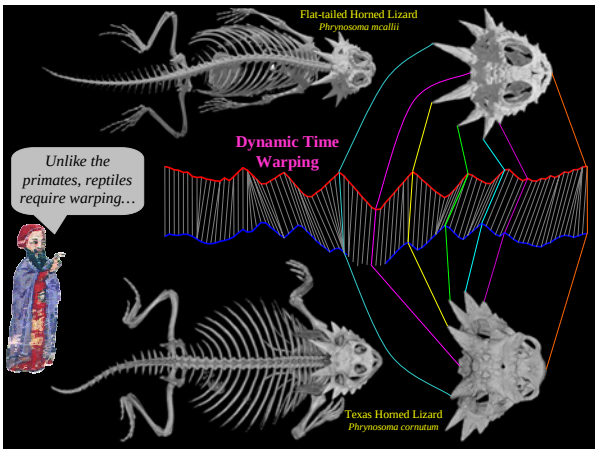


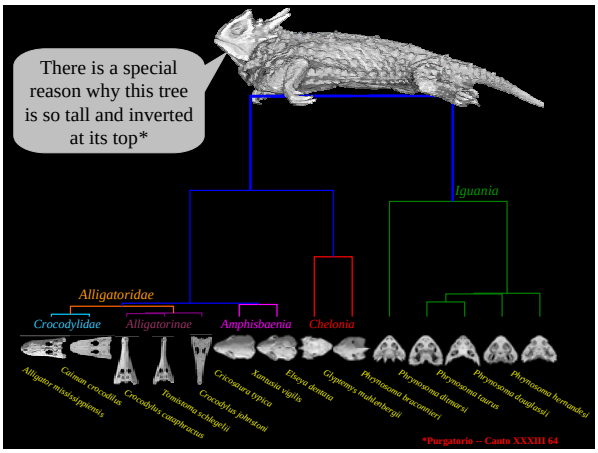
Indexing Experiments

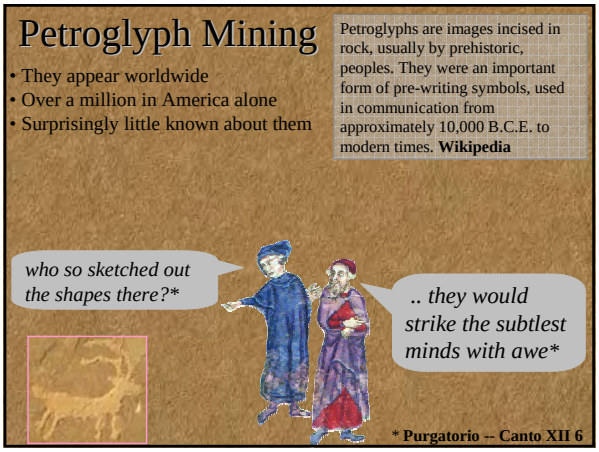
- Projectile point/Heterogenous databases
- Increasingly large dimensionality
- One-nearest-neighbor queries

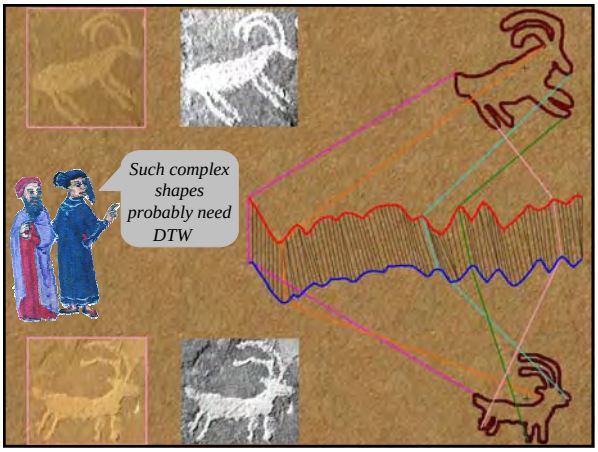


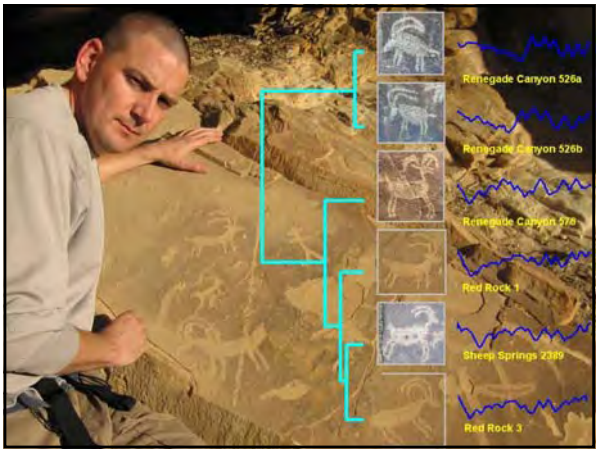


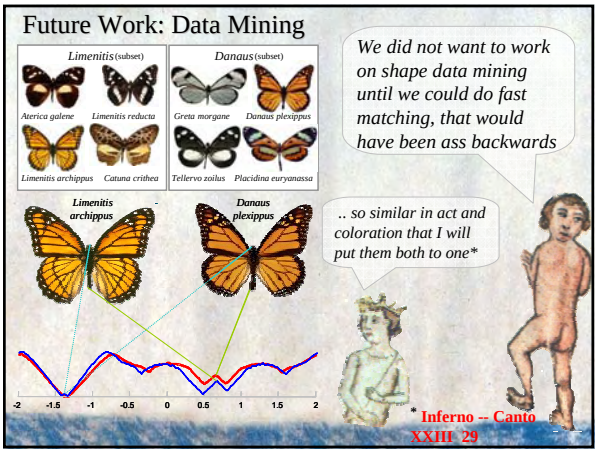

Flat-tailed Horned Lizard
Phrynosoma mcallii
Dynamic Time Warping
Unlike the primates, reptiles require warping…
Texas Horned Lizard
Phrynosoma cornutum


There is a special reason why this tree is so tall and inverted at its top*
Iguania
Alligatoridae
Crocodylidae
Alligatorinae
Amphisbaenia
Chelonia
*Purgatorio -- Canto XXXIII 64


Petroglyph Mining
• They appear worldwide
• Over a million in America alone
• Surprisingly little known about them
Petroglyphs are images incised in rock, usually by prehistoric, peoples. They were an important form of pre-writing symbols, used in communication from approximately 10,000 B.C.E. to modern times. Wikipedia
who so sketched out the shapes there?*
.. they would strike the subtlest minds with awe*
* Purgatorio -- Canto XII 6


Such complex shapes probably need DTW


Renegade Canyon 526a
Renegade Canyon 526b
Renegade Canyon 378
Red Rock 1
Sheep Springs 2389
Red Rock 3


Future Work: Data Mining
Limenitis (subset)
Danaus (subset)
Aterica galene
Limenitis reducta
Greta morgane
Danaus plexippus
Limenitis archippus
Catuna crithea
Tellervo zoilus
Placidina euryanassa
We did not want to work on shape data mining until we could do fast matching, that would have been ass backwards
Limenitis archippus
Danaus plexippus
.. so similar in act and coloration that I will put them both to one*
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
* Inferno -- Canto XXIII 29


Questions?
Feel free to email us with questions
Eamonn Keogh: Project Leader
eamonn@cs.ucr.edu
Li Wei: Lower Bounding
wli@cs.ucr.edu
Michail Vlachos: Public Nudity and Index Structures
vlachos@us.ibm.com
Sang Hee Lee: Anthropology and Primatology
shlee@ucr.edu
Xiaopeng Xi: Image Processing
xxi@cs.ucr.edu