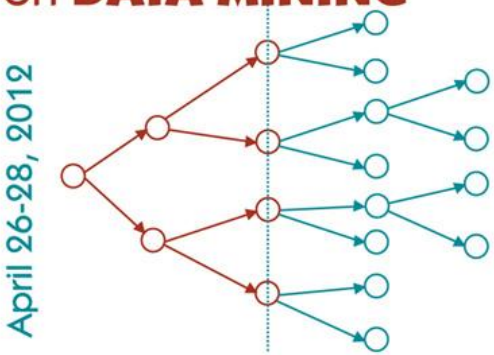
How to do good research, get it published

Eamonn Keogh

Computer Science & Engineering Department
University of California - Riverside
Riverside, CA 92521
eamonn@cs.ucr.edu

2012 SIAM International Conference on DATA MINING



April 26-28, 2012

Disney's Paradise Pier Hotel
Anaheim, California, USA

Disclaimers I

- I don't have a magic bullet for publishing
  - This is simply my best effort to the community, especially young faculty, grad students and "outsiders".
- For every piece of advice where I tell you "*you should do this*" or "*you should never do this*"…
  - You will be able to find counterexamples, including ones that won best paper awards etc.
- I will be critiquing some published papers (including some of my own), however I mean no offence.
  - Of course, these are *published* papers, so the authors could legitimately say I am wrong.

Disclaimers II

- These slides are meant to be *presented*, and then *studied* offline. To allow them to be self-contained like this, I had to break my rule about keeping the number of words to a minimum.
- You have a PDF copy of these slides, if you want a PowerPoint version, email me.
- I plan to continually update these slides, so if you have any feedback/suggestions/criticisms please let me know.

Disclaimers III

- Many of the *positive* examples are mine, making this tutorial seem self indulgent and vain.
- I did this simply because…
  - I know what reviewers said for my papers.
  - I know the reasoning behind the decisions in my papers.
  - I know when earlier versions of my papers got rejected, and why, and how this was fixed.

Disclaimers IIII

- Many of the ideas I will share are *very simple*, you might find them insultingly simple.
- Nevertheless, in my ten year experience as a reviewer/area chair, at least half of papers submitted to SDM/SIGKDD/ICDM/ have at least one of these simple flaws.

The Following People Offered Advice

- Geoff Webb
- Frans Coenen
- Cathy Blake
- Michael Pazzani
- Lane Desborough
- Stephen North
- Fabian Moerchen
- Ankur Jain
- Themis Palpanas
- Jeff Scargle
- Howard J. Hamilton
- Mark Last
- Chen Li
- Magnus Lie Hetland
- David Jensen
- Chris Clifton
- Oded Goldreich
- Victoria Stodden
- Michalis Vlachos
- Claudia Bauzer Medeiros
- Chunsheng Yang
- Xindong Wu
- Lee Giles
- Johannes Fuernkranz
- Vineet Chaoji
- Stephen Few
- Wolfgang Jank
- Claudia Perlich
- Mitsunori Ogihara
- Hui Xiong
- Chris Drummond
- Charles Ling
- Charles Elkan
- Jieping Ye
- Saeed Salem
- Tina Eliassi-Rad
- Parthasarathy Srinivasan
- Mohammad Hasan
- Vibhu Mittal
- Chris Giannella
- Frank Vahid
- Carla Brodley
- Ansaf Salleb-Aouissi
- Tomas Skopal
- Frans Coenen
- Sang-Hee Lee
- Michael Carey
- Vijay Atluri
- Shashi Shekhar
- Jennifer Windom
- Hui Yang
- Graham Cormode

My students: Jessica Lin, Chotirat Ratanamahatana, Li Wei, Xiaopeng Xi, Dragomir Yankov, Lexiang Ye, Xiaoyue (Elaine) Wang , Jin-Wien Shieh, Abdullah Mueen, Qiang Zhu, Bilson Campana

These people are *not* responsible for any controversial or incorrect claims made here

Outline

- The Review Process
- Writing a data mining paper
  - Finding problems/data
    - Framing problems
    - Solving problems
  - Tips for writing
    - Motivating your work
    - Clear writing
    - Clear figures
- The top ten reasons papers get rejected
  - With solutions

The Curious Case of Srikanth Krishnamurthy

- In 2004 Srikanth's student submitted a paper to MobiCom
- Deciding to change the title, the student resubmitted the paper, accidentally submitting it as a new paper
- One version of the paper scored 1,2 and 3, and was rejected, the other version scored a 3,4 and 5, and was accepted!

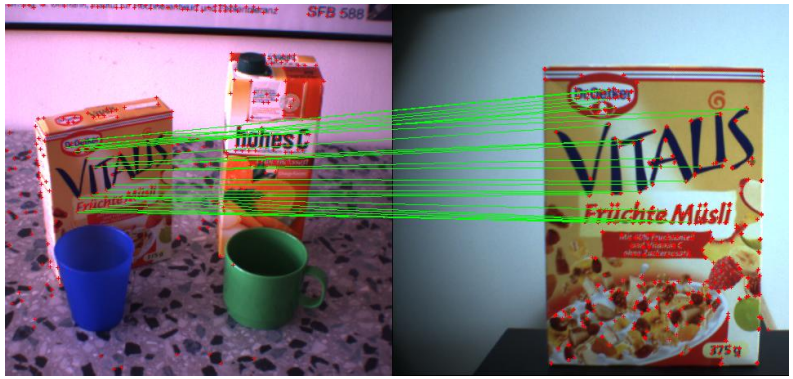
- This "natural" experiments suggests that the reviewing process is random, is it really that bad?

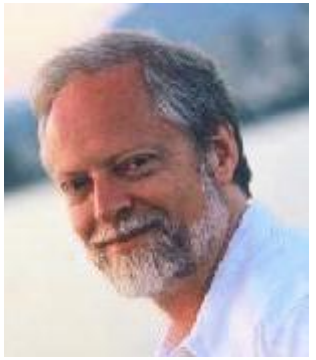
Reviewers do get it (very) wrong sometimes

David Lowe's work on the SIFT method has about 10,000 citations, it was the most highly cited paper in all of engineering sciences in 2005.

I did submit papers on earlier versions of SIFT to both ICCV 97 and CVPR 98 and both were rejected... David Lowe

Story from Yann LeCun



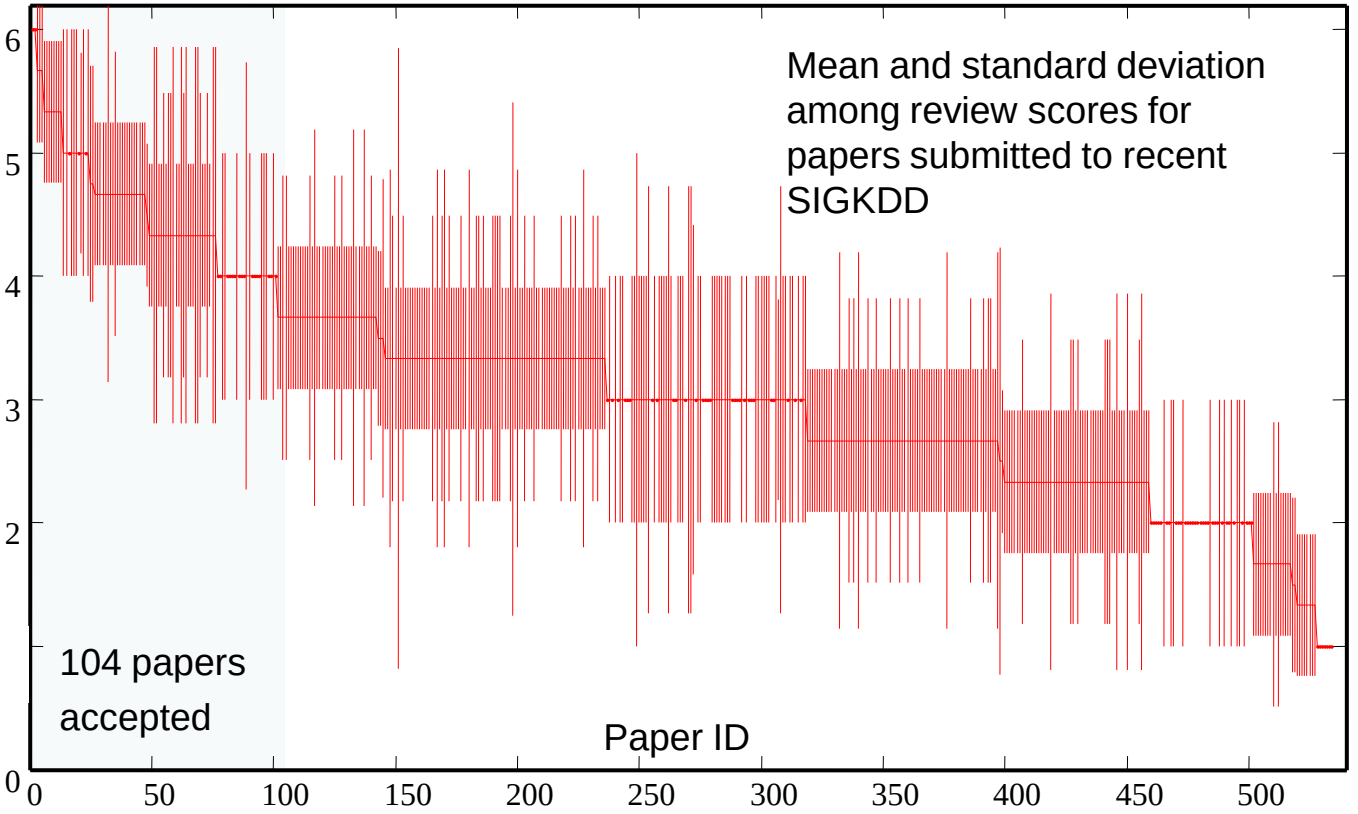


David Lowe

A look at the reviewing statistics for a recent SIGKDD

(I cannot say what year)

Mean number of reviews 3.02

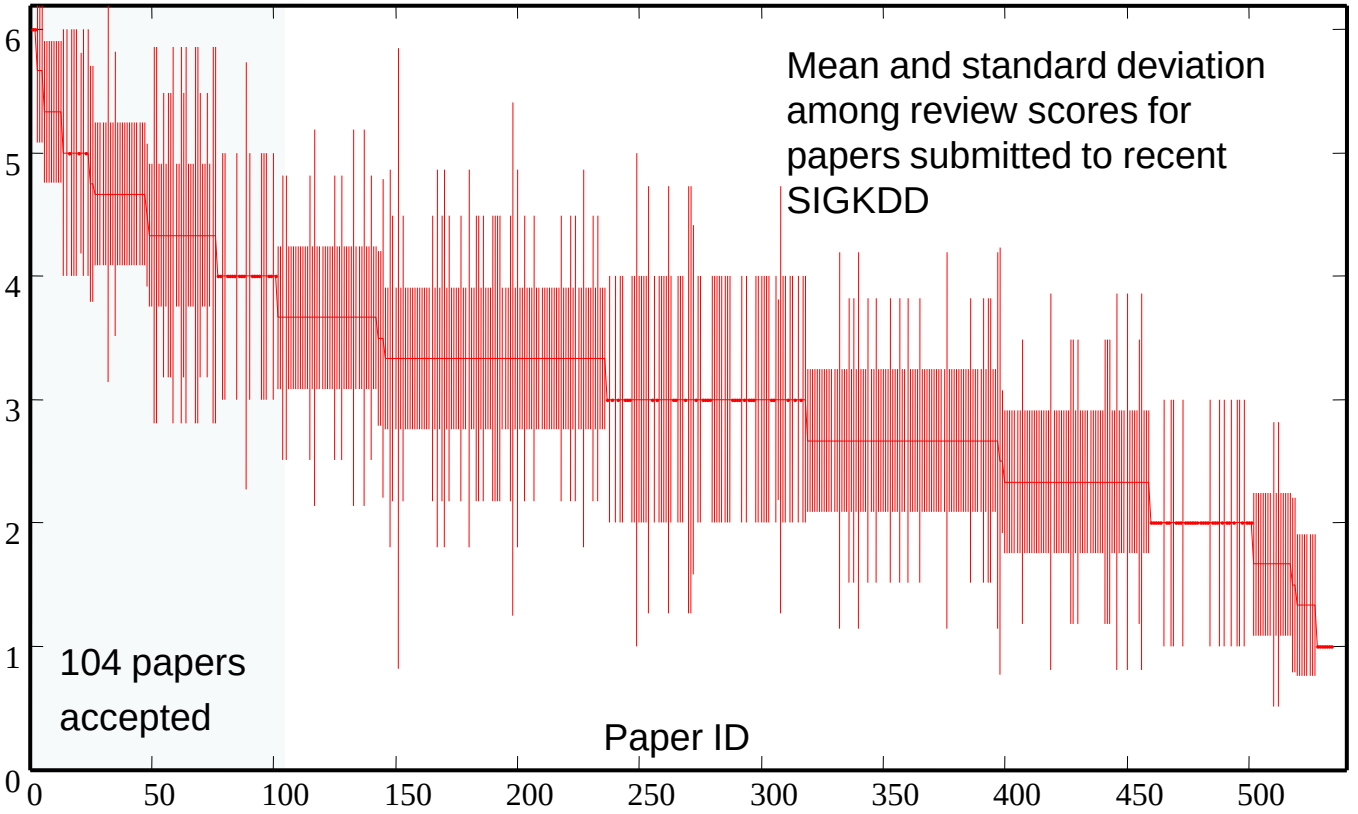


- Papers accepted after a discussion, not solely based on the mean score.
- These are final scores, after reviewer discussions.
- The variance in reviewer scores is **much larger** than the differences in the mean score, for papers on the boundary between accept and reject.
- In order to *halve* the standard deviation we must *quadruple* the number of reviews.

Conference reviewing is an imperfect system.

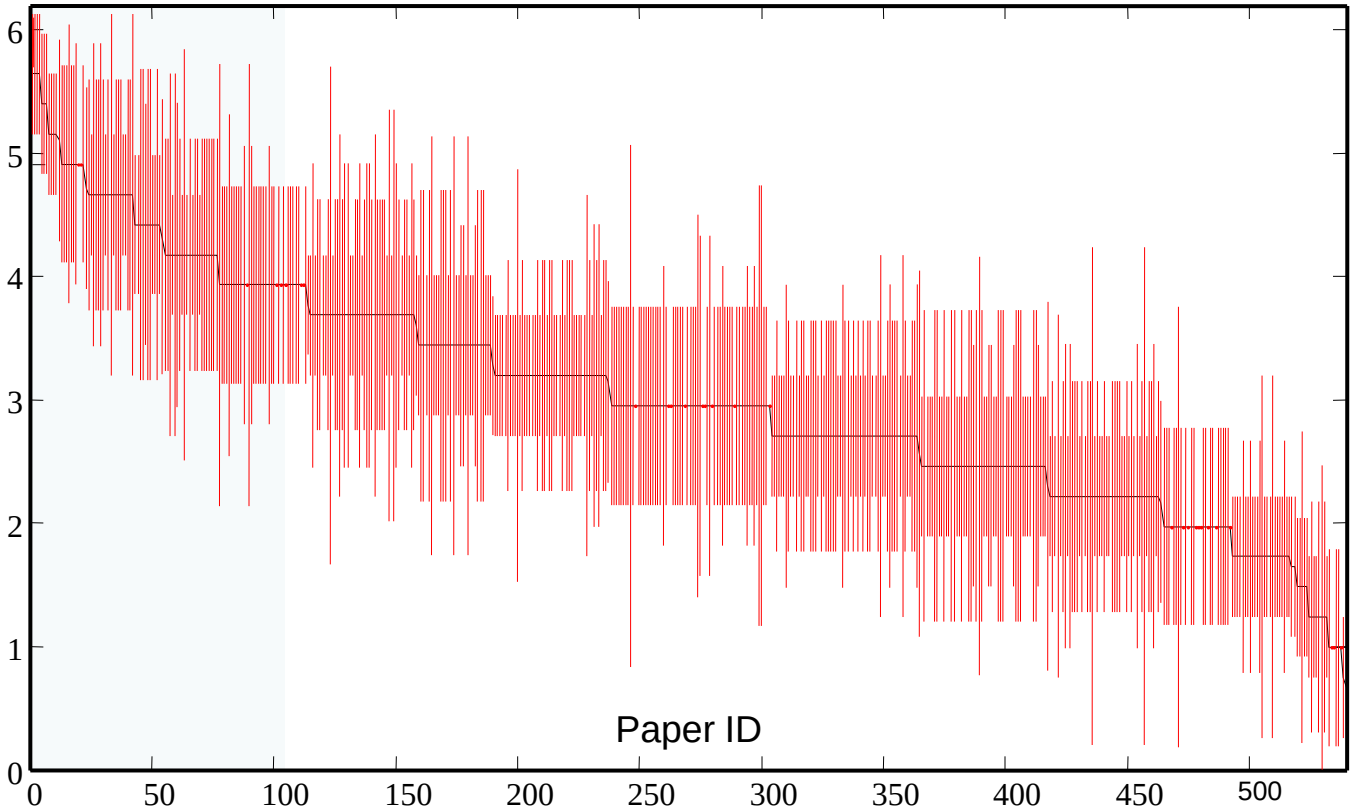
We must learn to live with rejection.

All we can do is try to make sure that our paper lands as far left as possible



- At least three papers with a score of 3.67 (or lower) must have been accepted. But there were a total of 41 papers that had a score of 3.67.
- That means there exist at least 38 papers that were rejected, that had the same or better numeric score as some papers that were accepted.
- **Bottom Line**: With very high probability, multiple papers will be rejected in favor of less worthy papers.

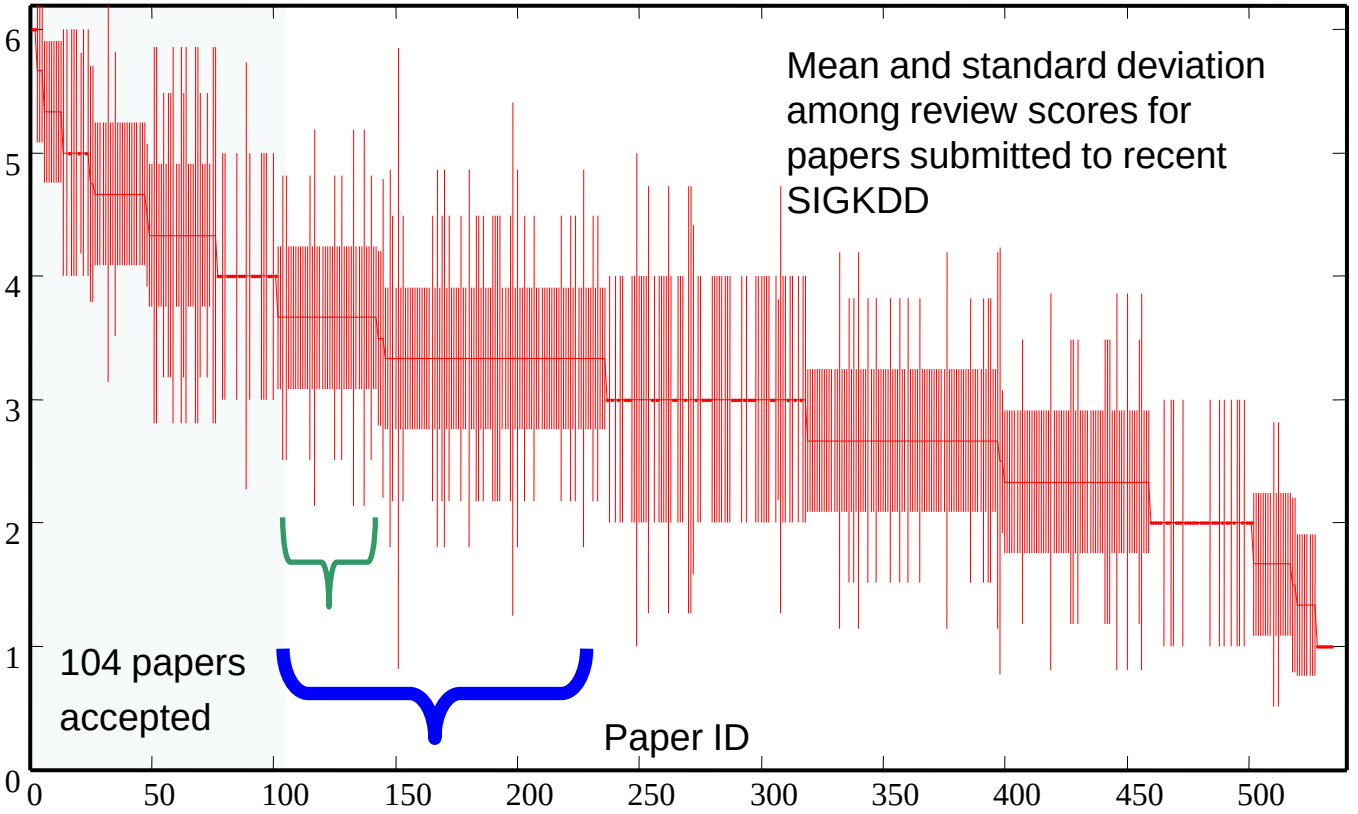
A sobering experiment



- Suppose I add one *reasonable review* to each paper.
- A *reasonable review* is one that is drawn uniformly from the range of one less than the lowest score to one higher than the highest score.
- If we do this, then on average, 14.1 papers move across the accept/reject borderline. This suggests a very brittle system.

But the good news is…

Most of us only need to improve a *little* to improve our odds a *lot*.



• Suppose you are one of the 41 groups in the **green** (light) area. If you can convince just one reviewer to increase their ranking by just one point, you go from near certain reject to near certain accept.

• Suppose you are one of the 140 groups in the **blue** (**bold**) area. If you can convince just one reviewer to increase their ranking by just one point, you go from near certain reject to a good chance at accept.

Idealized Algorithm for Writing a Paper

- Find problem/data
- Start writing (*yes*, start writing *before* and *during* research)
- Do research/solve problem
- Finish 95% draft
- Send preview to mock reviewers
- Send preview to the rival authors (virtually or literally)
- Revise using checklist.
- Submit

One month before deadline

What Makes a Good Research Problem?

- **It is important:** If you can solve it, you can make money, or save lives, or help children learn a new language, or...
- **You can get real data**: Doing DNA analysis of the Loch Ness Monster would be interesting, but…
- **You can make incremental progress**: Some problems are all-or-nothing. Such problems may be too risky for young scientists.
- **There is a clear metric for success**: Some problems fulfill the criteria above, but it is hard to know when you are making progress on them.

Finding Problems/Finding Data

- Finding a good problem can be the hardest part of the whole process.
- Once you have a problem, you will need data…
- As I shall show in the next few slides, finding problems and finding data are best integrated.

- However, the obvious way to find problems is the best, read *lots* of papers, both in SIGKDD and elsewhere.

Domain Experts as a Source of Problems

- Data miners are almost unique in that they can work with almost any scientist or business
- I have worked with anthropologists, nematologists, archaeologists, astronomers, entomologists, cardiologists, herpetologists, electroencephalographers, geneticists, space vehicle technicians etc
- Such collaborations can be a rich source of interesting problems.

Working with Domain Experts I

- Getting problems from domain experts might come with some bonuses
- Domain experts can help with the **motivation** for the paper
  - *..insects cause 40 billion dollars of damage to crops each year..*
  - *..compiling a dictionary of such patterns would help doctors diagnosis..*
  - *Petroglyphs are one of the earliest expressions of abstract thinking, and a true hallmark...*
- Domain experts sometimes have funding/internships etc
- Co-authoring with domain experts can give you credibility.

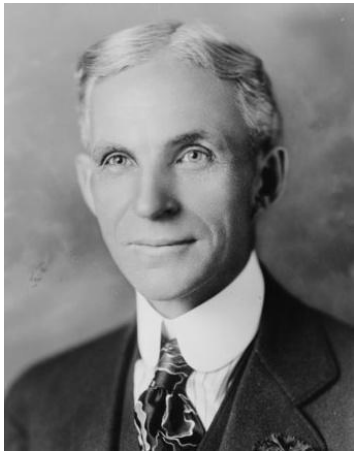
Augmenting the Generalized Hough Transform to Enable the Mining of Petroglyphs

Qiang Zhu Xiaoyue Wang Eamonn Keogh [1]Sang-Hee Lee
Dept. of Computer Science & Engineering, [1]Dept. of Anthropology
University of California, Riverside, CA 92521
{qzhu, xwang, eamonn}@cs.ucr.edu, sang-hee.lee@ucr.edu

SIGKDD 09

Working with Domain Experts II

If I had asked my customers what they wanted, they would have said a faster horse



Henry Ford

- Ford focused not on stated need but on **latent need**.
- In working with domain experts, don't just ask them what they want. Instead, try to learn enough about their domain to understand their **latent** needs.
- In general, domain experts have little idea about what is hard/easy for computer scientists.

Working with Domain Experts III

Concrete Example:

- I once had a biologist spend an hour asking me about sampling/estimation. She wanted to estimate a quantity.
- After an hour I realized that we did not have to estimate it, we could compute an *exact* answer!
- The exact computation did take three days, but it had taken several years to gather the data.
- Understand the **latent need**.

Finding Research Problems

- Suppose you think idea **X** is very good
- Can you extend **X** by…
  - Making it more accurate (*statistically significantly* more accurate)
  - Making it faster (usually an order of magnitude, or no one cares)
  - Making it an anytime algorithm
  - Making it an online (streaming) algorithm
  - Making it work for a different data type (including uncertain data)
  - Making it work on low powered devices
  - Explaining *why* it works so well
  - Making it work for distributed systems
  - Applying it in a novel setting (industrial/government track)
  - Removing a parameter/assumption
  - Making it disk-aware (if it is currently a main memory algorithm)
  - Making it simpler

Finding Research Problems (examples)

- Suppose you think idea X is a very good
- Can you extend **X** by…
  - Making it more accurate (*statistically significantly* more accurate)
  - Making it faster (usually an order of magnitude, or no one cares)
  - **Making it an anytime algorithm**
  - **Making it an online (streaming) algorithm**
  - **Making it work for a different data type** (including uncertain data)
  - Making it work on low powered devices
  - Explaining *why* it works so well
  - Making it work for distributed systems
  - Applying it in a novel setting (industrial/government track)
  - Removing a parameter/assumption
  - Making it disk-aware (if it is currently a main memory algorithm)

- The Nearest Neighbor Algorithm is very useful. I wondered if we could make it an ***anytime*** algorithm…. ICDM06 [b].
- Motif discovery is very useful for ***DNA***, would it be useful for ***time series***? SIGKDD03 [c]
- The bottom-up algorithm is very useful for batch data, could we make it work in an ***online*** setting? ICDM01 [d]
- Chaos Game Visualization of DNA is very useful, would it be useful for ***other kinds of data***? SDM05 [a]

[a] Kumar, N., Lolla N., Keogh, E., Lonardi, S. , Ratanamahatana, C. A. and Wei, L. (2005). Time-series Bitmaps: ICDM 2006
[b] Ueno, Xi, Keogh, Lee. Anytime Classification Using the Nearest Neighbor Algorithm with Applications to Stream Mining. ICDM 2006.
[c] Chiu, B. Keogh, E., & Lonardi, S. (2003). Probabilistic Discovery of Time Series Motifs. SIGKDD 2003
[d] Keogh, E., Chu, S., Hart, D. & Pazzani, M. An Online Algorithm for Segmenting Time Series. ICDM 2001

Finding Research Problems

- Suppose you think idea X is a very good
- Can you extend **X** by…
  - Making it more accurate (*statistically significantly* more accurate)
  - Making it faster (usually an order of magnitude, or no one cares)
  - **Making it an anytime algorithm**
  - **Making it an online (streaming) algorithm**
  - **Making it work for a different data type** (including uncertain data)
  - Making it work on low powered devices
  - Explaining *why* it works so well
  - Making it work for distributed systems
  - Applying it in a novel setting (industrial/government track)
  - Removing a parameter/assumption
  - Making it disk-aware (if it is currently a main memory algorithm)

- Some people have suggested that this method can lead to incremental, boring, low-risk papers…
  - Perhaps, but there are 104 papers in SIGKDD this year, they are not all going to be groundbreaking.
  - Sometimes ideas that seem incremental at first blush may turn out to be very exciting as you explore the problem.
  - An early career person might eventually go on to do high risk research, after they have a "cushion" of two or three lower-risk SIGKDD papers.

Framing Research Problems I

As a reviewer, I am often frustrated by how many people don't have a clear problem statement in the abstract (or the entire paper!)

Can you write a research statement for your paper in a single sentence?

- **X** is good for **Y** (in the context of **Z**).
- **X** can be extended to achieve **Y** (in the context of **Z**).
- The adoption of **X** facilitates **Y** (for data in **Z** format).
- An **X** approach to the problem of **Y** mitigates the need for **Z**.

(An **anytime algorithm** approach to the problem of **nearest neighbor classification** mitigates the need for **high performance hardware**) (Ueno et al. ICDM 06)

If I, as a reviewer, cannot form such a sentence for your paper after reading just the abstract, then your paper is usually doomed.

Tina Eliassi-Rad

I hate it when a paper under review does not give a concise definition of the problem

See talk by Frans Coenen on this topic
http://www.csc.liv.ac.uk/~frans/Seminars/doingAphdSeminarAI2007.pdf

Framing Research Problems II

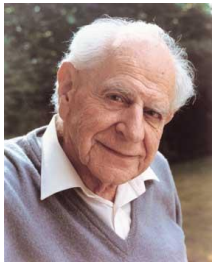
Your research statement should be **falsifiable**

A real paper claims:

> *To the best of our knowledge, this is most sophisticated subsequence matching solution mentioned in the literature.*

Is there a way that we could show this is not true?

Falsifiability (or **refutability**) is the logical possibility that an claim can be shown false by an observation or a physical experiment. That something is 'falsifiable' does not mean it is false; rather, that *if* it is false, then this can be shown by observation or experiment



Karl Popper

Falsifiability is the demarcation between science and nonscience

Framing Research Problems III

Examples of falsifiable claims:

- *Quicksort is faster than bubblesort.* (this may needed expanding, *if the lists are..*)
- *The X function lower bounds the DTW distance.*
- *The L2 distance measure generally outperforms L1 measure*

(this needs some work (under what conditions etc), but it is falsifiable)

Examples of unfalsifiable claims:

- *We can approximately cluster DNA with DFT.*
  - Any random arrangement of DNA could be considered a "clustering".
- *We present an alterative approach through Fourier harmonic projections to enhance the visualization. The experimental results demonstrate significant improvement of the visualizations.*
  - Since "*enhance*" and "*improvement* " are subjective and vague, this is unfalsifiable. Note that it *could* be made falsifiable. Consider:
    - *We improve the mean time to find an embedded pattern by a factor of ten.*
    - *We enhanced the separability of weekdays and weekends, as measured by..*

From the Problem to the Data

- At this point we have a concrete, falsifiable research problem
- Now is the time to get data!

  By "now", I mean months before the deadline. I have one of the largest collections of free datasets in the world. Each year I am amazed at how many emails I get a few days before the SIGKDD deadline that asks "*we want to submit a paper to SIGKDD, do you have any datasets that..* "
- Interesting, real (large, when appropriate) datasets *greatly* increase your papers chances.
- Having good data will also help do better research, by preventing you from converging on unrealistic solutions.
- Early experience with real data can feed back into the *finding and framing the research question* stage.

- Given the above, we are going to spend some time considering data..

Is it OK to Make Data?

There is a **huge** difference between…

We wrote a Matlab script to create random trajectories

and…

We glued tiny radio transmitters to the backs of Mormon crickets and tracked the trajectories

Photo by Jaime Holguin

Why is Synthetic Data so Bad?

Suppose you say "*Here are the results on our synthetic dataset*:"

	Our Method	Their Method
Accuracy	95%	80%

This is good right? After all, you are doing much better than your rival.

Why is Synthetic Data so Bad?

Suppose you say "*Here are the results on our synthetic dataset*:"

	Our Method	Their Method
Accuracy	**95%**	**80%**

But as far as I know, you might have created ten versions of your dataset, but only reported one!

Even if you did not do this consciously, you may have done it unconsciously.

At best, *your* making of *your* test data is a huge conflict of interest.

	Our Method	Their Method
Accuracy	80%	85%
Accuracy	75%	85%
Accuracy	90%	90%
Accuracy	**95%**	**80%**
Accuracy	85%	95%

Why is Synthetic Data so Bad?

Note that is does not really make a difference if you have real data but you modify it somehow, it is still synthetic data.

A paper has a section heading: **Results on Two Real Data Sets**

But then we read…

We add some noises to a small number of shapes in both data sets to manually create some anomalies.

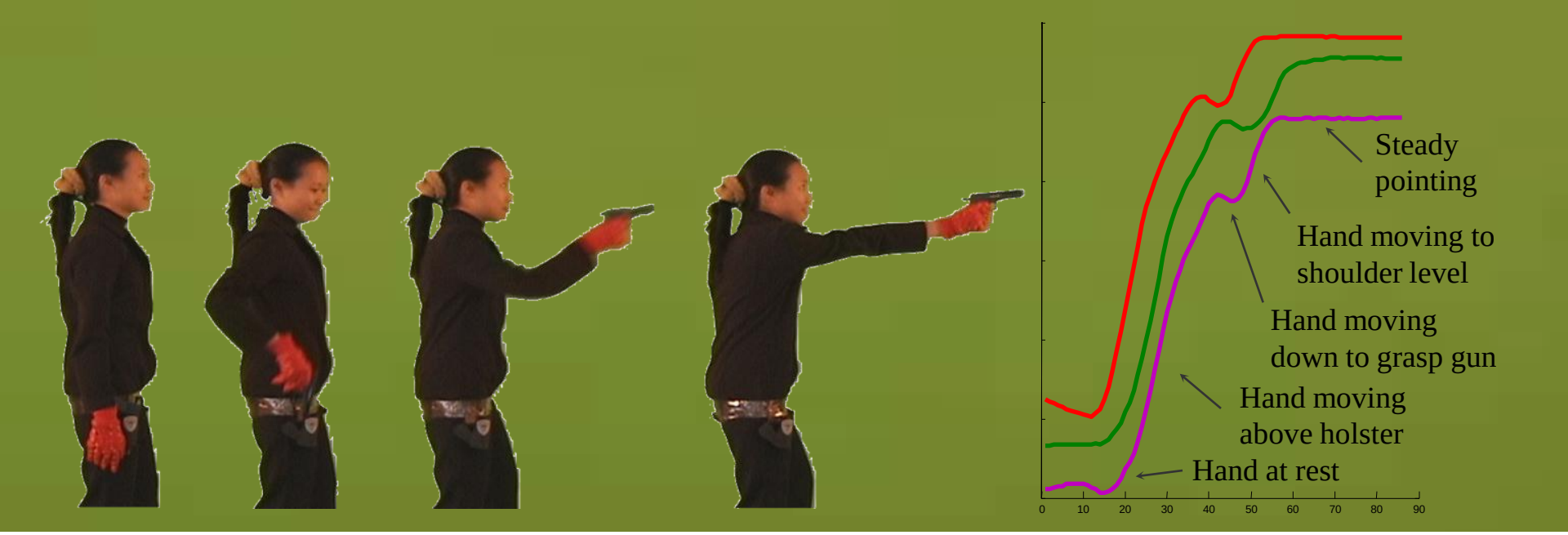
Is this still real data? The answer is *no*, even if they authors had explained how they added noise (which they don't).

Note that there are probably a handful of circumstances were taking real data, doing an experiment, tweaking the data and repeating the experiment is genuinely illuminating.

Synthetic Data can lead to a Contradiction

- Avoid the contradiction of claiming that the problem is very important, but there is no real data.
- If the problem is as important as you claim, a reviewer would wonder why there is no real data.
- I encounter this contradiction very frequently, here is a real example:

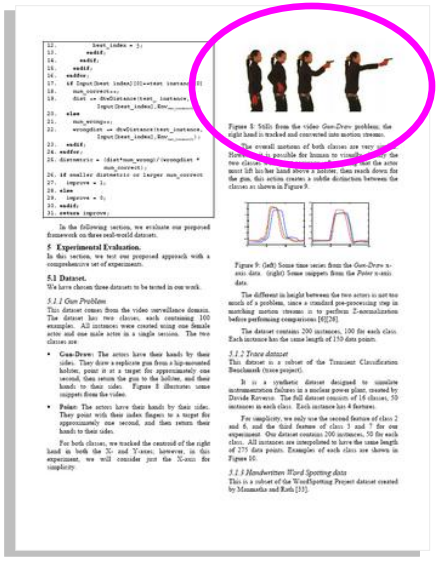
- **Early in the paper**: *The ability to process large datasets becomes more and more important…*
- **Later in the paper**: *..because of the lack of publicly available large datasets…*



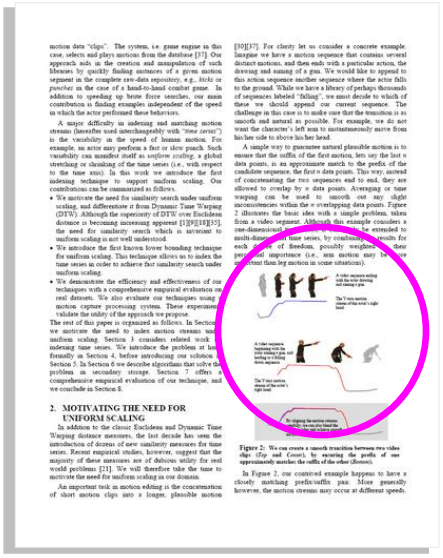
I want to convince you that the effort it takes to find or create real data is worthwhile.

In 2003, I spent two full days recording a video dataset. The data consisted of my student Chotirat (Ann) Ratanamahatana performing actions in front of a green screen.

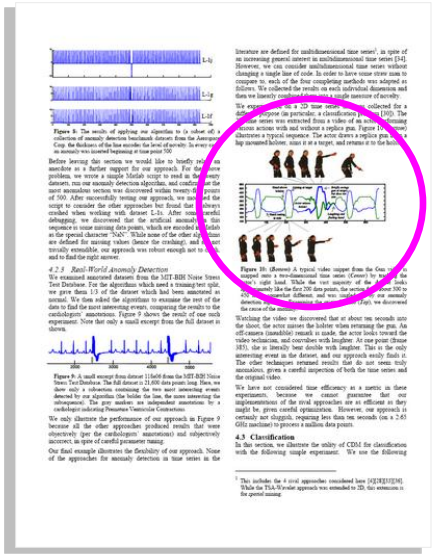
Was this a waste of two days?



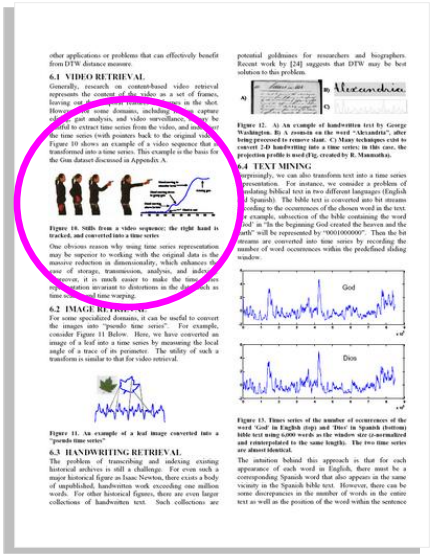
SDM 04



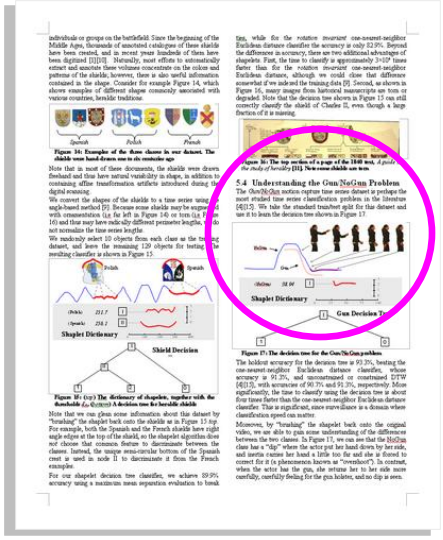
VLDB 04



SIGKDD 04



SDM 05



SIGKDD 09

I have used this data in at least a dozen papers, and one dataset derived from it, the GUN/NOGUN problem, has been used in well over 100 other papers (all of which reference my work!)

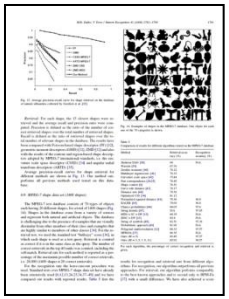
Spending the time to make/obtain/clean good datasets will pay off in the long run

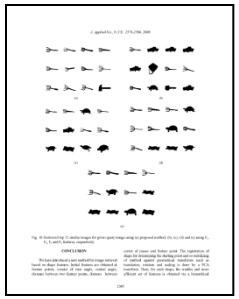
The vast majority of papers on shape mining use the MPEG-7 dataset.

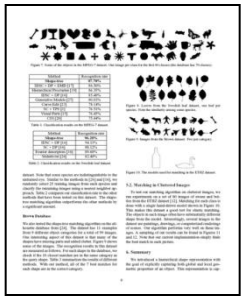
Visually, they are telling us : "*I can tell the difference between Mickey Mouse and spoon*".

The problem is not that I think this easy, the problem is I just don't care.

Show me data I care about







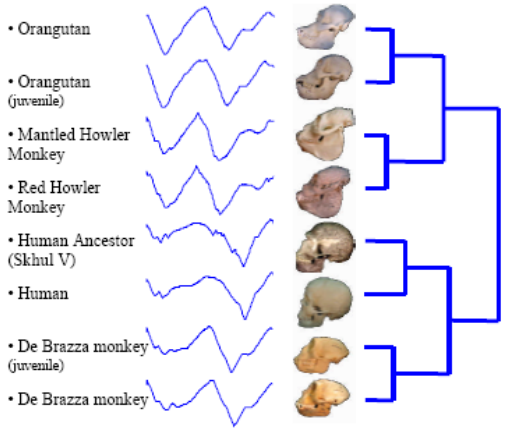


Figure 16: A group average hierarchal clustering of eight primate skulls based on the lateral view, using Euclidean distance

It is important to recall that Figure 16 shows a phenogram, *not* a phylogenetic tree. However on larger scale experiments in this domain (shown in [14]) we found that large subtrees of the dendrograms did conform to the current consensus on primate evolution.

While the Euclidean distance works very well on the relatively simple primate skulls, we found that considering a more (morphologically) diverse groups of animals, such as all reptiles, requires DTW as a distance measure. Consider Figure 17 which shows a hierarchical clustering of a very diverse set of reptiles. As with the primates, this is not the correct phylogenetic tree for these animals, once again however, the (uniquely colored) subtrees do correspond to current consensus on reptiles evolution based on DNA analysis and/or more complete morphological studies [10][11].

Note that we are *not* claiming that our shape matching techniques replace or even complement classic morphometrics in zoology. The point of these experiments is that if the shape matching techniques can produce intuitive results in a domain in which we know the correct relationships by other means, this suggests that algorithms may also produce meaningful results in shape problems for which there is more uncertainty, including projectile points (see [26] and Figure 15), petroglyphs, insect bite patterns in leaves [42], mammographic calcifications [43] etc.

It has recently been claimed that shape matching methods that only look at the contours of shapes (boundary based methods) are brittle to articulation distortion [33], however we believe that while this may be true for certain boundary based methods (i.e Hausdorff, Champer etc) the centroid based method we use is *very* robust to articulation distortions. To demonstrate this, we conducted a simple experiment/demonstration. We took three Lepidoptera, including the very similar and closely related *Actias maenaes* and *Actias philippinica*, and produced a copy of each. We then took these copies and "bent" the right hindwing. The clustering of the three originals and three copies under Euclidean distance, group average linkage is shown in Figure 18.

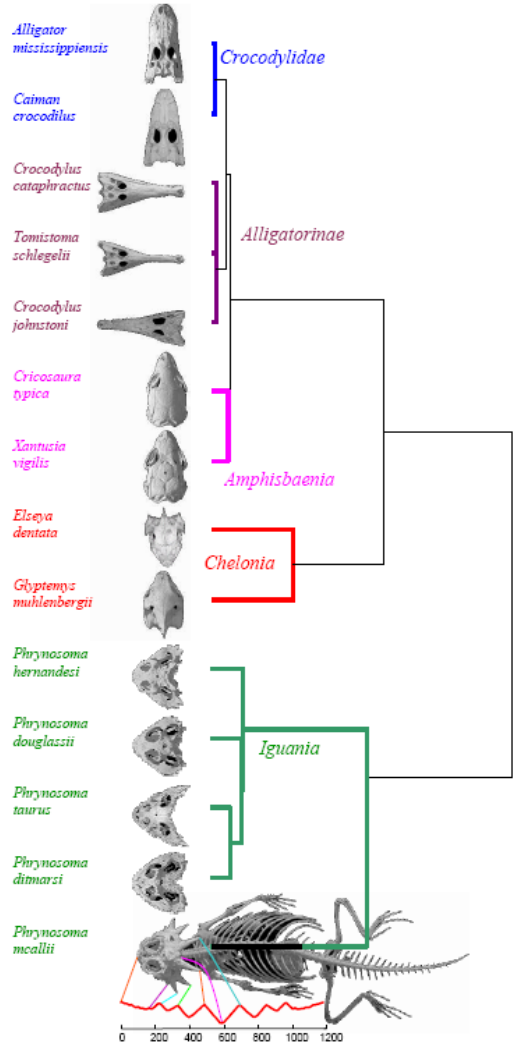


Figure 17: A group average hierarchal clustering of fourteen reptile skulls based on the superior view, using DTW distance

Real data motivates your clever algorithms: Part I

This figure tells me "*if I rotate my hand drawn apples, then I will need to have a rotation invariant algorithm to find them*"

In contrast, this figure tells me "*Even in this important domain, where tens of millions of dollars are spent each year, the robots that handle the wings cannot guarantee that they can present them in the same orientation each time. Therefore I will need to have a rotation invariant algorithm* "

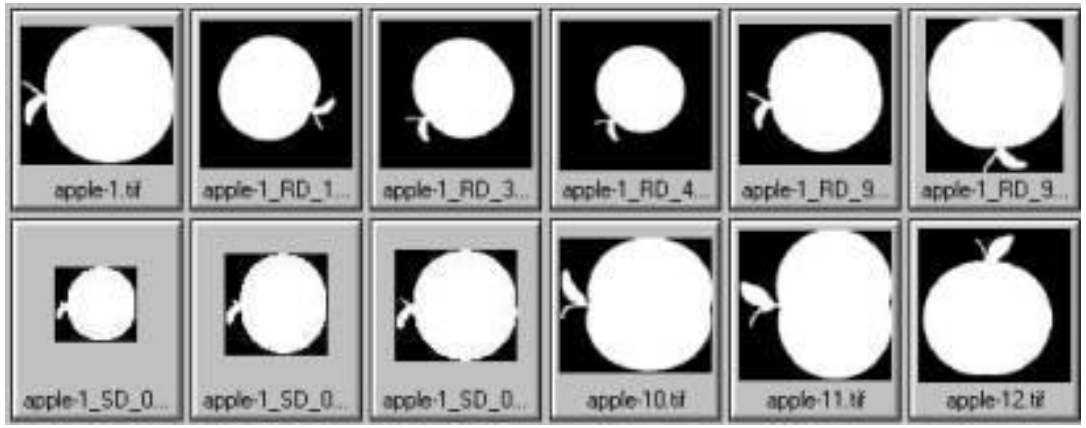


Figure 3: shapes of natural objects can be from different views of the same object, shapes can be rotated, scaled, skewed

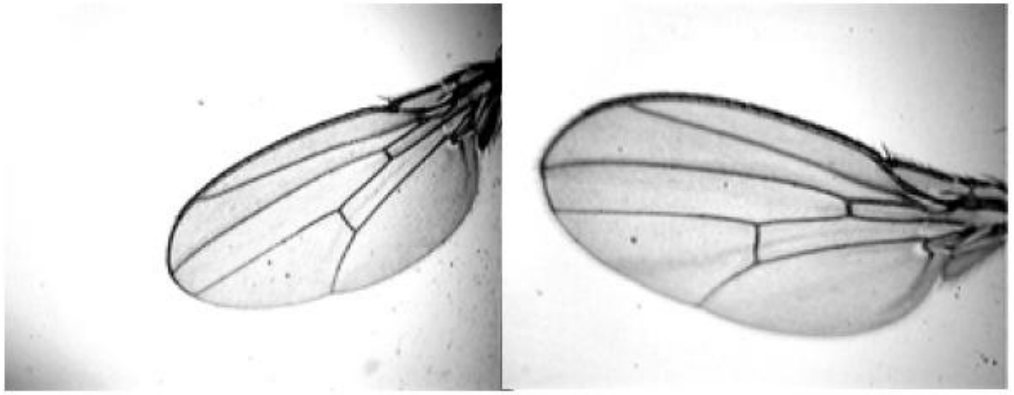


Figure 5: Two sample wing images from a collection of Drosophila images. Note that the rotation of images can vary even in such a structured domain

Real data motivates your clever algorithms: Part II

This figure tells me "*if I use Photoshop to take a chunk out of a drawing of an apple, then I will need an occlusion resistant algorithm to match it back to the original*"

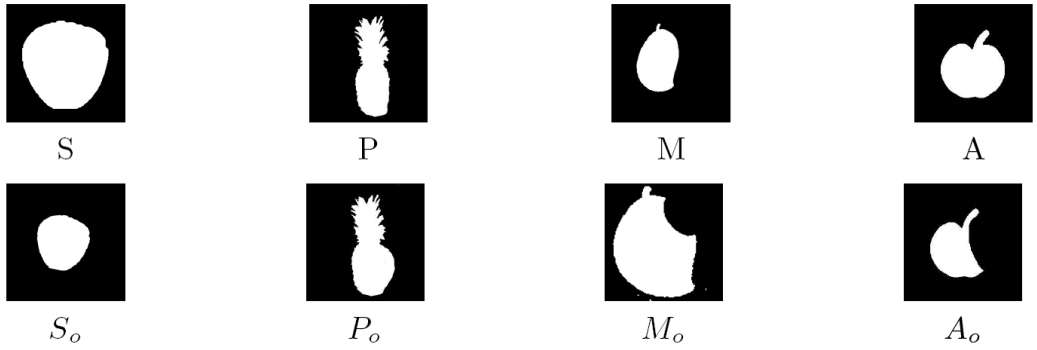


Fig. 6. Selected Original and Occluded Shapes

In contrast, this figure tells me "*In this important domain of cultural artifacts it is common to have objects which are effectively occluded by breakage. Therefore I will need to have an occlusion resistant algorithm* "

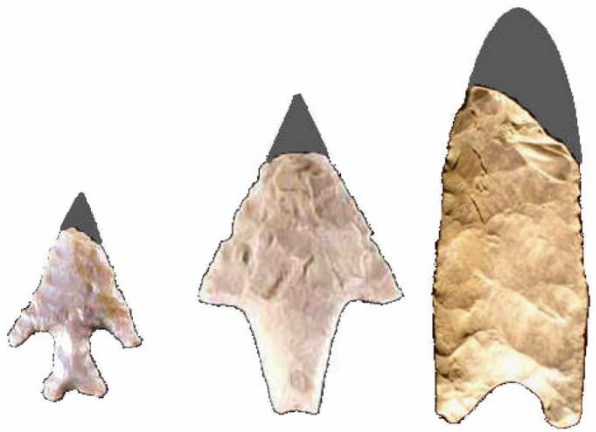
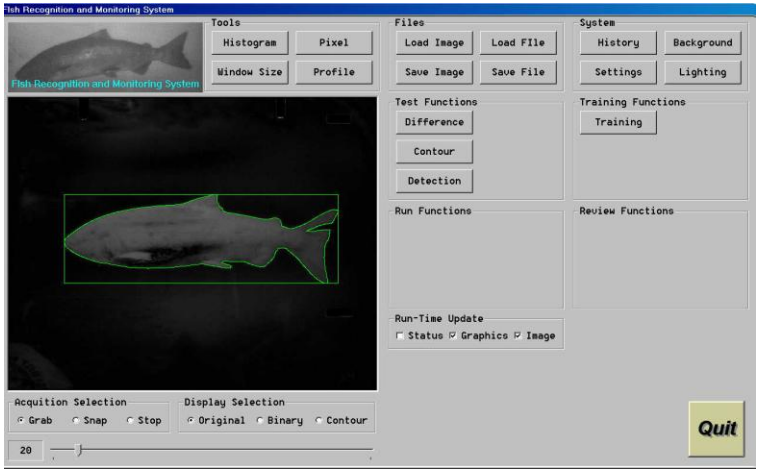


Figure 15: Project points are frequently found with broken tips or tangs. Such objects require LCSS to find meaningful matches to complete specimens.

Here is a great example. This paper is not technically deep.

However, instead of classifying synthetic shapes, they have a very cool problem (fish counting/classification) and they made an effort to create a very interesting dataset.

Show me data someone cares about



3. SHAPE EXTRACTION AND REPRESNETATION

3.1 Fish Contour Extraction

Subtraction of images acquired at different times can detect the motion of an object [12]. It is also a simple way to detect the presence of an object assuming a stationary camera position and constant illumination. Fig. 5 (a) shows an image taken without any objects. The only minor variation between frames of the same background is the water turbulence. Averaging of a few frames without objects provides a smooth background image as shown in Fig. 5 (a). Differences between the background image and the image taken at different time shown in Fig. 5 (b) can detect objects distinct from the background as shown in Fig. 5 (c). The difference image contains small pixel clusters (blobs) from water turbulence or image noise. They can be removed with a morphological opening operator. Size and the location of this binary blob can be used for the edge detection process. For small sized fish, edge detection processes can be initiated immediately after the entire fish is in the viewing window. For large fish, that are longer than the viewing window width, subsequent image processing tasks will have to be performed in two steps, one for the head portion of the fish and the other one for the tail. For field testing and data collection, we performed contour extraction for every image that has an object size larger than the set size threshold.

Under normal operation conditions, the difference image should have very good contrast between fish and background as shown in Fig. 5 (c). We used the Canny edge operator to extract the contour because of its contour following (edge tracking) feature. A non-maximum suppression technique based on gradient magnitude is used to *thin* the wide ridges around the local maxima to produce one-pixel wide edges. Once the gradient magnitudes are thinned, extended contour segments can be produced by following high gradient magnitudes from one neighborhood to another [13, 14]. Contour following is initiated only on edge pixels which have high gradient magnitude. However, once it starts, contours are tracked through lower gradient magnitude pixels. A closed fish contour can be detected with this method.

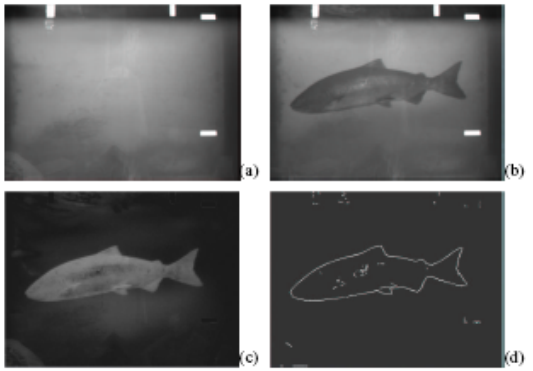


Figure 5. (a) Background image, (b) live image with fish, (c) difference image, and (d) fish contour.

3.2 Fish Detection and Tracking

The flowchart of the fish detection and tracking algorithms is shown in Fig. 6. We first calibrated the system to acquire a reference image (without fish) by averaging 10 or more frames of image to smooth out the image background noise. After the reference image was established, it was then subtracted from every frame of live image acquired from the camera to obtain a difference image. The acquired live image is sent to an edge detector and an edge threshold was applied to the edge image to determine the fish location. A difference threshold was applied to the difference image to obtain a binary image of the fish or object.

The binarized difference image and the binarized edge image were then sent to the contour detection subroutine to detect the fish/object contour. For the binarized edge image, we treated it as a mask, and "AND" it with the binarized difference image. Combining the contour and the mask from the edge image, we were able to extract the fish/object bounding box. This bounding box was checked to see if it falls in the area of interest. If this bounding box falls outside the area of interest, then the acquired image will be ignored and a new image will be acquired to repeat the same process. If the bounding box falls within the area of interest, then the image will then be fully processed.

D.J. Lee, R. Schoenberger, D. Shiozawa, X. Xu, and P. Zhan, "*Contour Matching for a Fish Recognition and Migration Monitoring System*", SPIE Optics East, Two and Three-Dimensional Vision Systems for Inspection, Control, and Metrology II, vol. 5606-05, Philadelphia, PA, USA, October 25-28, 2004

How big does my Dataset need to be?

It depends…

Suppose you are proposing an algorithm for mining Neanderthal bones. There are only a few hundred specimens known, and it is very unlikely that number will double in our lifetime. So you could reasonably test on a synthetic* dataset with a mere 1,000 objects.

However…

Suppose you are proposing an algorithm for mining Portuguese web pages (there are billions) or some new biometric (there may soon be millions). You do have an obligation to test on large datasets.

It is increasing difficult to excuse data mining papers testing on small datasets. Data is typically free, CPU cycles are essentially free, a terabyte of storage costs less than $100…

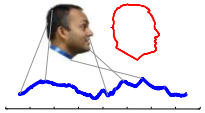
*In this case, the "synthetic" could be easer to obtain monkey bones etc.

Where do I get Good Data?

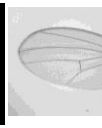
- From your domain expert collaborators:
- From formal data mining archives:
  - The UCI Knowledge Discovery in Databases Archive.
  - The UCR Time Series and Shape Archive.
- From general archives:
  - Chart-O-Matic
  - NASA GES DISC
- From creating it:









  - Glue tiny radio transmitters to the backs of Mormon crickets…
  - By a Wii, and hire a ASL interpreter to…
- Remember there is *no* excuse for not getting real data.

Solving Problems

- Now we have a problem and data, *all* we need to do is to solve the problem.
- Techniques for solving problems depend on your skill set/background and the problem itself, however I will quickly suggest some simple general techniques.
- Before we see these techniques, let me suggest you **avoid complex solutions**. This is because complex solutions...
    - …are less likely to generalize to datasets.
    - …are much easer to overfit with.
    - …are harder to explain well.
    - …are difficult to reproduce by others.
    - …are less likely to be cited.

Unjustified Complexity I

From a recent paper:

This forecasting model integrates a case based reasoning (CBR) technique, a Fuzzy Decision Tree (FDT), and Genetic Algorithms (GA) to construct a decision-making system based on historical data and technical indexes.

- Even if you believe the results. Did the improvement come from the CBR, the FDT, the GA, or from the combination of two things, or the combination of all three?
- In total, there are more than 15 parameters…
- How reproducible do you think this is?

Unjustified Complexity II

- There *may* be problems that really require very complex solutions, but they seem rare. see [a].
- Your paper is implicitly claiming "*this is the simplest way to get results this good*".
- Make that claim *explicit*, and carefully justify the complexity of your approach.

[a] R.C. Holte, *Very simple classification rules perform well on most commonly used datasets*, Machine Learning 11 (1) (1993). This paper shows that one-level decision trees do very well most of the time.

J. Shieh and E. Keogh *iSAX: Indexing and Mining Terabyte Sized Time Series*. SIGKDD 2008. This paper shows that the simple Euclidean distance is competitive to much more complex distance measures, once the datasets are reasonably large.

Unjustified Complexity III

Charles Elkan

Paradoxically and wrongly, sometimes if the paper used an excessively complicated algorithm, it is more likely that it would be accepted

If your idea is simple, *don't* try to hid that fact with unnecessary padding (although unfortunately, that does seem to work *sometimes*). Instead, *sell* the simplicity.

> *"...it reinforces our claim that our methods are very* ***simple*** *to implement.. ..Before explaining our* ***simple*** *solution this problem……we can objectively discover the anomaly using the* ***simple*** *algorithm..."* SIGKDD04

Simplicity is a strength, not a weakness, acknowledge it and claim it as an advantage.

Solving Research Problems

- **Problem Relaxation**:
- **Looking to other Fields for Solutions**:

We don't have time to look at all ways of solving problems, so lets just look at two examples in detail.

If there is a problem you can't solve, then there is an easier problem you can solve: find it.

George Polya

Can you find a problem analogous to your problem and solve that?

Can you vary or change your problem to create a new problem (or set of problems) whose solution(s) will help you solve your original problem?

Can you find a subproblem or side problem whose solution will help you solve your problem?

Can you find a problem related to yours that has been solved and use it to solve your problem?

Can you decompose the problem and "recombine its elements in some new manner"? (Divide and conquer)

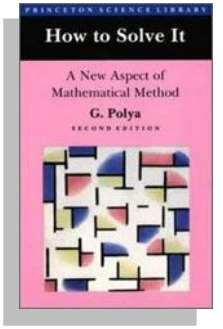
Can you solve your problem by deriving a generalization from some examples?

Can you find a problem more general than your problem?

Can you start with the goal and work backwards to something you already know?

Can you draw a picture of the problem?

Can you find a problem more specialized?



Problem Relaxation: If you cannot solve the problem, make it easier and then try to solve the easy version.

- If you **can** solve the easier problem… Publish it if it is worthy, then revisit the original problem to see if what you have learned helps.
- If you **cannot** solve the easier problem…Make it even easier and try again.

Example: Suppose you want to maintain the closest pair of real-valued points in a sliding window over a stream, in worst-case linear time and in constant space[1]. Suppose you find you cannot make progress on this…

Could you solve it if you..

- Relax to *amortized* instead of *worst-case* linear time.
- Assume the data is discrete, instead of real.
- Assume you have infinite space.
- Assume that there can never be ties.

[1]I am not suggesting this is an meaningful problem to work on, it is just a teaching example

Problem Relaxation: Concrete example, petroglyph mining

I want to build a tool that can find and extract petroglyphs from an image, quickly search for similar ones, do classification and clustering etc

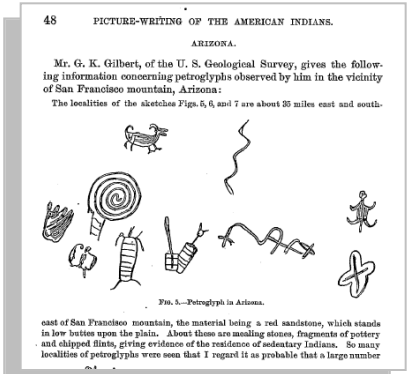
The extraction and segmentation is really hard, for example the cracks in the rock are extracted as features.

I need to be scale, offset, and rotation invariant, but rotation invariance is really hard to achieve in this domain.

What should I do? (continued next slide)

Problem Relaxation: Concrete example, petroglyph mining

- Let us relax the difficult segmentation and extraction problem, after all, there are thousands of segmented petroglyphs online in old books…
- Let us relax rotation invariance problem, after all, for some objects (people, animals) the orientation is usually fixed.
- Given the relaxed version of the problem, can we make progress? Yes! Is it worth publishing? Yes!
- Note that I am not saying we should give up now. We should still tried to solve the harder problem. What we have learned solving the easier version might help when we revisit it.
- In the meantime, we have a paper and a little more confidence.


48 PICTURE-WRITING OF THE AMERICAN INDIANS.

ARIZONA.

Mr. G. K. Gilbert, of the U. S. Geological Survey, gives the following information concerning petroglyphs observed by him in the vicinity of San Francisco mountain, Arizona:

The localities of the sketches Figs. 6, 6, and 7 are about 35 miles east and south-

FIG. 5.—Petroglyph in Arizona.

east of San Francisco mountain, the material being a red sandstone, which stands in low buttes upon the plain. About these are mealing stones, fragments of pottery and chipped flints, giving evidence of the residence of sedentary Indians. So many localities of petroglyphs were seen that I regard it as probable that a large number


Augmenting the Generalized Hough Transform to Enable the Mining of Petroglyphs

Qiang Zhu Xiaoyue Wang Eamonn Keogh Sang-Hee Lee

Dept. of Computer Science & Engineering, Dept. of Anthropology

University of California, Riverside, CA 92521

{qzhu, xwang, eamonn}@cs.ucr.edu, sang-hee.lee@ucr.edu

ABSTRACT

Categories and Subject Descriptors

General Terms

1. INTRODUCTION

2. BACKGROUND AND RELATED WORK

Note that we must acknowledge the assumptions/limitations in the paper

SIGKDD 2009

Looking to other Fields for Solutions: Concrete example, *Finding Repeated Patterns in Time Series*

• In 2002 I became interested in the idea of finding repeated patterns in time series, which is a computationally demanding problem.

• After making no progress on the problem, I started to look to other fields, in particular computational biology, which has a similar problem of DNA motifs..

• As happens Tompa & Buhler had just published a clever algorithm for DNA motif finding. We adapted their idea for time series, and published in SIGKDD 2002…

Tompa, M. & Buhler, J. (2001). Finding motifs using random projections. 5th Int'l Conference on Computational Molecular Biology. pp 67-74.

Looking to other Fields for Solutions

You never can tell were good ideas will come from. The solution to a problem on anytime classification came from looking at bee foraging strategies.

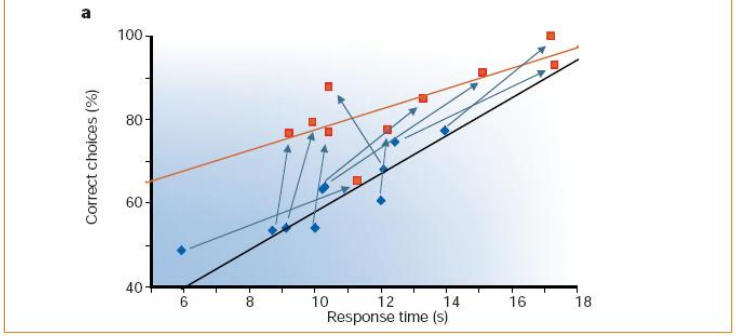


Figure 1 Bumblebees can choose wisely or rapidly, but not both at once. **a,** Interindividual correlation between response time and accuracy of bees discriminating between two virtual flower types. Each symbol denotes the average performance of one individual bee under one experimental condition. When targets were rewarded with sucrose solution and distractors contained no reward (plain water) (blue symbols and black regression line), bees investing more time made more accurate choices. When distractors were penalized with bitter quinine solution (red symbols and orange regression line), all bees improved their accuracy. Blue arrows link the average values for individual bees under the two experimental conditions. **b,** A blue virtual flower with a bumblebee imbibing sucrose solution from a Plexiglas platform.

Bumblebees can choose wisely or rapidly, but not both at once.. Lars Chittka, Adrian G. Dyer, Fiola Bock, Anna Dornhaus, Nature Vol.424, 24 Jul 2003, p.388

- We data miners can often be inspired by biologists, data compression experts, information retrieval experts, cartographers, biometricians, code breakers etc.
- Read widely, give talks about your *problems* (not *solutions*), collaborate, and ask for advice (on blogs, newsgroups etc)

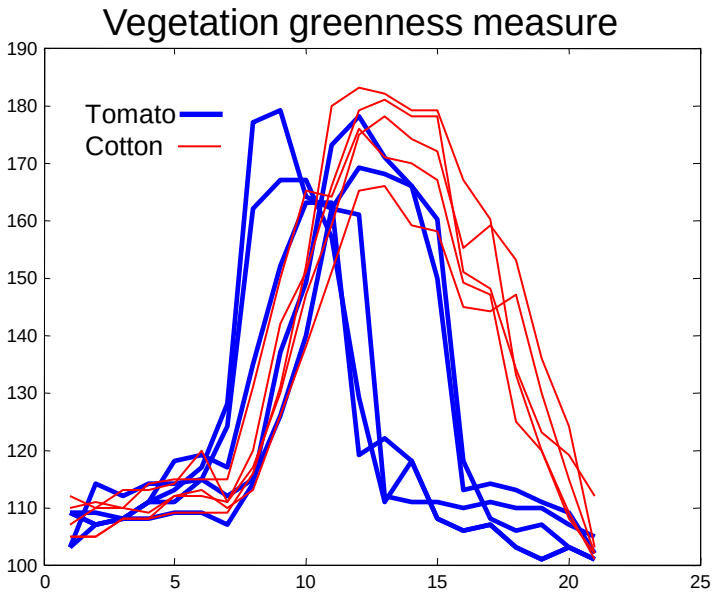
Eliminate Simple Ideas

When trying to solve a problem, you should begin by eliminating simple ideas. There are two reasons why:

- It may be the case that that simple ideas *really* work very well, this happens much more often than you might think.

- Your paper is making the implicit claim "*This is the simplest way to get results this good*". You need to convince the reviewer that this is true, to do this, start by convincing yourself.

Eliminate Simple Ideas: Case Study I (a)



In 2009 I was approached by a group to work on the classification of crop types in Central Valley California using Landsat satellite imagery to support pesticide exposure assessment in disease.

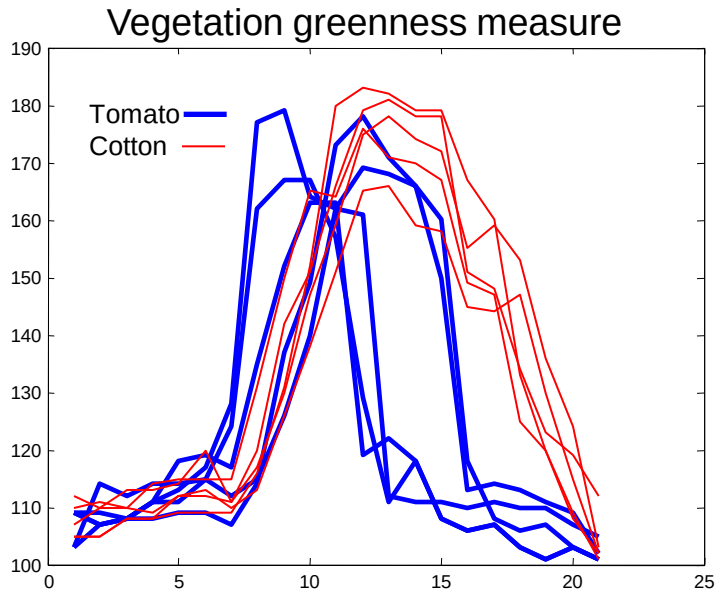
They came to me because they could not get DTW to work well..

At first glance this is a dream problem

- Important domain
- Different amounts of variability in each class
- I could see the need to invent a mechanism to allow **Partial Rotation Invariant Dynamic Time Warping** (I could almost smell the best paper award!)

But there is a problem….

Eliminate Simple Ideas: Case Study I (b)



It is possible to get perfect accuracy with a single line of matlab!

In particular this line: sum(x) > 2700

Lesson Learned: Sometimes *really* simple ideas work very well. They might be more difficult or impossible to publish, but oh well.
We should always be thinking in the back of our minds, is there a simpler way to do this?
When writing, we **must** convince the reviewer
This is the simplest way to get results this good

```
>> sum(x)
ans =   2845   2843   2734   2831   2875   2625   2642   2642   2490   2525

>> sum(x) > 2700
ans =   1   1   1   1   1   0   0   0   0   0
```

Eliminate Simple Ideas: Case Study II

A paper sent to SIGMOD 4 or 5 years ago tackled the problem of **Generating the Most Typical Time Series in a Large Collection**.

The paper used a complex method using wavelets, transition probabilities, multi-resolution properties etc.

The quality of the *most typical time series* was measured by comparing it to every time series in the collection, and the smaller the average distance to everything, the better.

SIGMOD Submission paper algorithm

(a few hundred lines of code, learns model from data)

…

```
X =  DWT(A + somefun(B))
Typical_Time_Series  = X + Z
```

Reviewers algorithm

```
(does not look at the data, and
takes exactly one line of code)

Typical_Time_Series  = zeros(64)
```

Under their metric of success, it is clear to the reviewer (without doing any experiments) that a constant line is the optimal answer for any dataset!

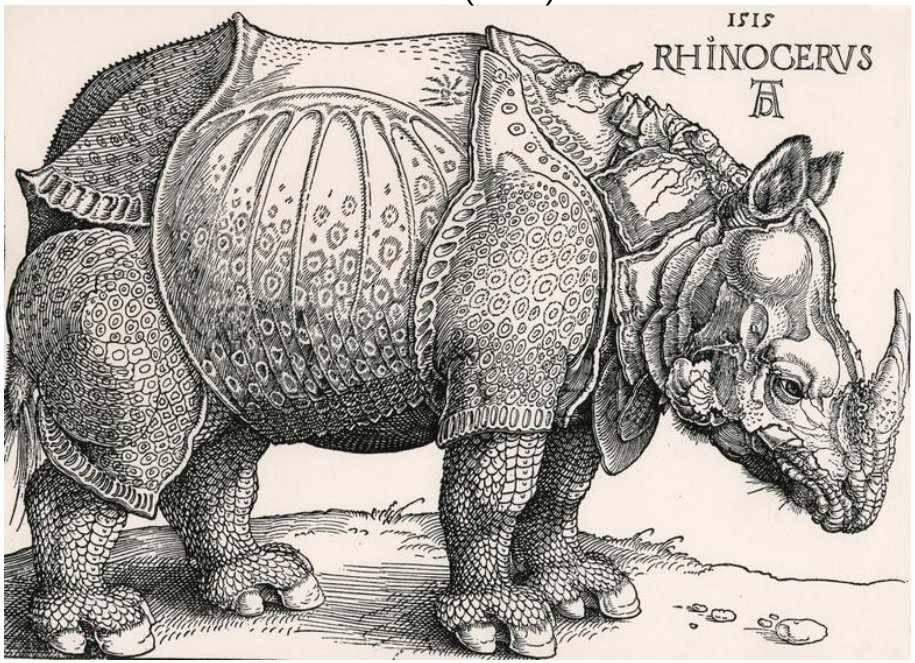
We should always be thinking in the back of our minds, is there a simpler way to do this? When writing, we **must** convince the reviewer *This is the simplest way to get results this good*

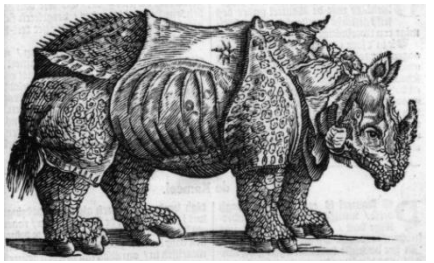
The Importance of being Cynical

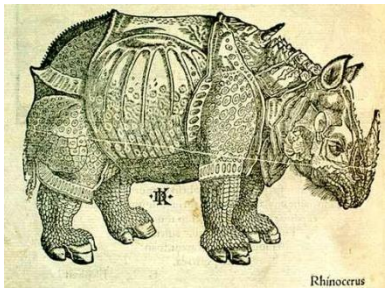
In 1515 Albrecht Dürer drew a Rhino from a sketch and written description. The drawing is remarkably accurate, except that there is a spurious horn on the shoulder.

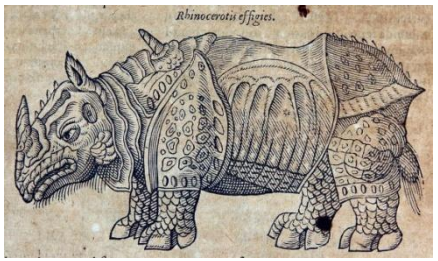
This extra horn appears on every European reproduction of a Rhino for the next 300 years.

Dürer's Rhinoceros (1515)









It Ain't Necessarily So

- Not every statement in the literature is true.
- Implications of this:
  - Research opportunities exist, confirming or refuting "known facts" (or more likely, investigating under what conditions they are true)
  - We must be careful not to assume that it is not worth trying X, since X is "*known*" not to work, or Y is "*known*" to be better than X
- In the next few slides we will see some examples

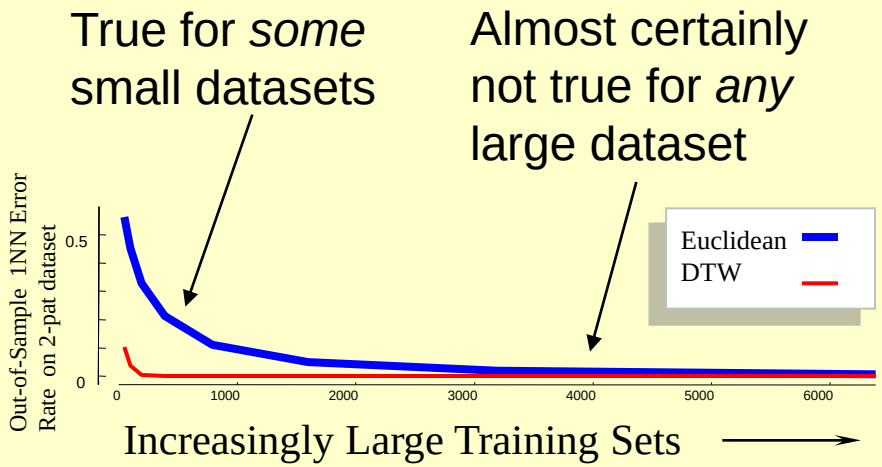
If you would be a real seeker after truth, it is necessary that you doubt, as far as possible, all things.

- In KDD 2000 I said "*Euclidean distance can be an extremely brittle distance measure*" Please note the "*can*"!

- This has been taken as gospel by many researchers
  - *However, Euclidean distance can be an extremely* ***brittle****..* Xiao et al. 04
  - *it is an extremely* ***brittle*** *distance measure…* Yu et al. 07
  - *The Euclidean distance, yields a* ***brittle*** *metric..* Adams et al 04
  - *to overcome the* ***brittle****ness of the Euclidean distance measure…* Wu 04
  - *Therefore, Euclidean distance is a* ***brittle*** *distance measure* Santosh 07
  - *that the Euclidean distance is a very* ***brittle*** *distance measure* Tuzcu 04

Is this really true?

Based on comparisons to 12 state-of-the-art measures on 40 different datasets, it is true on *some* small datasets, but there is no published evidence it is true on *any* large dataset (Ding et al VLDB 08)

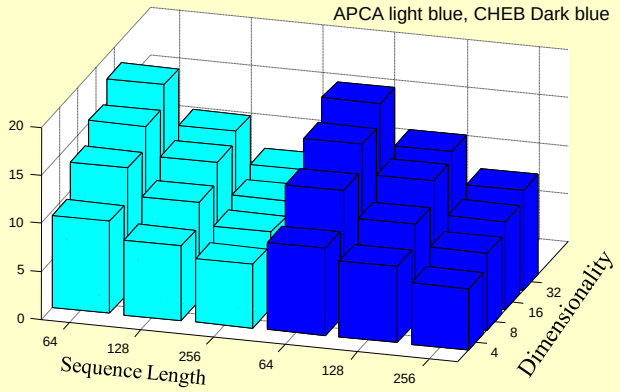


A SIGMOD Best Paper says..

Our empirical results indicate that Chebyshev approximation can deliver a 3- to 5-fold reduction on the dimensionality of the index space. For instance, it only takes 4 to 6 Chebyshev coefficients to deliver the same pruning power produced by 20 APCA coefficients

Is this really true?

No, actually Chebyshev approximation is slightly worse that other techniques (Ding et al VLDB 08)



The good results were due to a coding bug..

.. Thus it is clear that the C++ version contained a bug. We apologize for any inconvenience caused (note on authors page)

This is a problem, because many researchers have assumed it is true, and used Chebyshev polynomials without even considering other techniques. For example..

(we use Chebyshev polynomial approximation) because it is very accurate, and incurs low storage, which has proven very useful for similarity search. Ni and Ravishankar 07

In most cases, do *not* assume the problem is solved, or that algorithm X is the best, just because someone claims this.

A SIGKDD (r-up) Best Paper says..

(my paraphrasing) *You can slide a window across a time series, place all exacted subsequences in a matrix, and then cluster them with K-means. The resulting cluster centers then represent the typical patterns in that time series.*

Is this really true?

No, if you cluster the data as described above *the output is independent of the input* (random number generators are the only algorithms that are supposed to have this property). The first paper to point this out (Keogh et al 2003) met with tremendous resistance at first, but has been since confirmed in dozens of papers.

This is a problem, dozens of people wrote papers on making it faster/better, without realizing it does not work at all! At least two groups published multiple papers on this:

- Exploiting efficient parallelism for mining rules in time series data. Sarker et al 05
- Parallel Algorithms for Mining Association Rules in Time Series Data. Sarker et al 03
- Mining Association Rules from Multi-stream Time Series Data on Multiprocessor Systems. Sarker et al 05
- Efficient Parallelism for Mining Sequential Rules in Time Series. Sarker et al 06
- Parallel Mining of Sequential Rules from Temporal Multi-Stream Time Series Data. Sarker et al 06

In most cases, do *not* assume the problem is solved, or that algorithm X is the best, just because someone claims this.

Miscellaneous Examples

Voodoo Correlations in Social Neuroscience. Vul, E, Harris, C, Winkielman, P & Pashler, H.. Perspectives on Psychological Science. Here social neuroscientists criticized for overstating links between brain activity and emotion. This is an wonderful paper.

Why most Published Research Findings are False. J.P. Ioannidis. PLoS Med 2 (2005), p. e124.

Publication Bias: The "File-Drawer Problem" in Scientific Inference. Scargle, J. D. (2000), Journal for Scientific Exploration 14 (1): 91–106

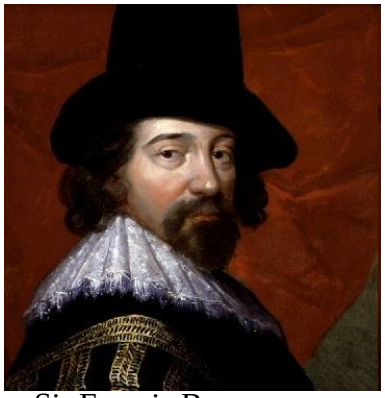
Classifier Technology and the Illusion of Progress. Hand, D. J.Statistical Science 2006, Vol. 21, No. 1, 1-15

Everything you know about Dynamic Time Warping is Wrong. Ratanamahatana, C. A. and Keogh. E. (2004). TDM 04

Magical thinking in data mining: lessons from CoIL challenge 2000 Charles Elkan

How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. Fanelli D, 2009 PLoS ONE4(5)

If a man will begin with certainties, he shall end in doubts; but if he will be content to begin with doubts he shall end in certainties.



Sir Francis Bacon
(1561 - 1626)

Non-Existent Problems

A final point before break.

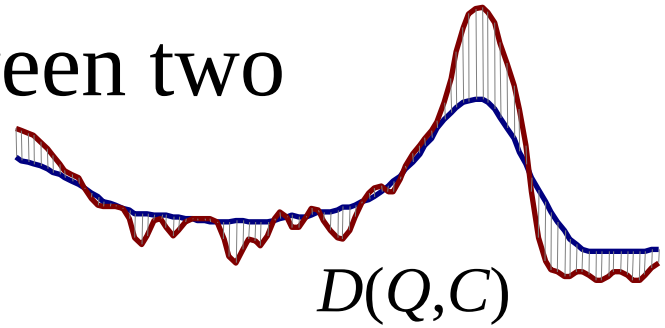
It is important that the problem you are working on is a *real* problem.

It may be hard to believe, but many people attempt (and occasionally succeed) to publish papers on problems that don't exist!

Lets us quickly spend 6 slides to see an example.

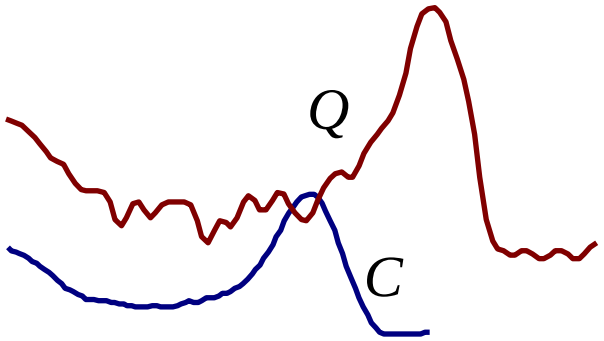
Solving problems that don't exist I

•This picture shows the visual intuition of the Euclidean distance between two time series of the same length



• Suppose the time series are of different lengths?

• We can just make one shorter or the other one longer..



C_new = resample(*C*, length(*Q*), length(*C*))

It takes one line of matlab code

Solving problems that don't exist II

But more than 2 dozen group have claimed that this is "wrong" for some reason, and written papers on how to compare two time series of different lengths (without simply making them the same length)

- "*(we need to be able) handle sequences of different lengths*" PODS 2005
- "*(we need to be able to find) sequences with similar patterns to be found even when they are of different lengths*" Information Systems 2004
- "*(our method) can be used to measure similarity between sequences of different lengths*" IDEAS2003

Solving problems that don't exist III

But an extensive literature search (by me), through more than 500 papers dating back to the 1960's failed to produce any theoretical or empirical results to suggest that simply making the sequences have the same length has any detrimental effect in classification, clustering, query by content or any other application.

Let us test this!

Solving problems that don't exist IIII

For all publicly available time series datasets which have naturally different lengths, let us compare the 1-nearest neighbor classification rate in two ways:

- After simply re-normalizing lengths (one line of matlab, no parameters)

- Using the ideas introduced in these papers to to support different length comparisons (various complicated ideas, some parameters to tweak) We tested the four most referenced ideas, and only report the best of the four.

Solving problems that don't exist V

The FACE, LEAF, ASL and TRACE datasets are the only publicly available classification datasets that come in different lengths, lets try all of them

Dataset	Resample to same length	Working with different lengths
Trace	0.00	0.00
Leaves	4.01	4.07
ASL	14.3	14.3
Face	2.68	2.68

A two-tailed t-test with 0.05 significance level for each dataset indicates that there is no statistically significant difference between the accuracy of the two sets of experiments.

Solving problems that don't exist VI

A least two dozen groups *assumed* that comparing different length sequences was a non-trivial problem worthy of research and publication.

But there was and still is to this day, zero evidence to support this!

And there is strong evidence to suggest this is **not** true.

There are two implications of this:

- Make sure the problem you are solving exists!
- Make sure you convince the reviewer it exists.

Coffee Break

Part II of

How to do good research, get it published in top venues

Eamonn Keogh

Writing the Paper

W. Somerset Maugham

There are three rules for writing the novel…

..Unfortunately, no one knows what they are.

Writing the Paper

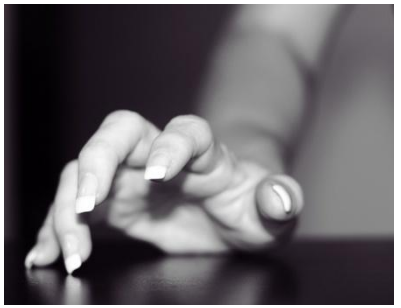

Samuel Johnson

What is written without effort is in general read without pleasure

- Make a working title
- Introduce the topic and define (informally at this stage) terminology
- Motivation: Emphasize why is the topic important
- Relate to current knowledge: what's been done
- Indicate the gap: what need's to be done?
- Formally pose research questions
- Explain any necessary background material.
- Introduce formal definitions.
- Introduce your novel algorithm/representation/data structure etc.
- Describe experimental set-up, explain what the experiments will show
- Describe the datasets
- Summarize results with figures/tables
- Discuss results
- Explain conflicting results, unexpected findings and discrepancies with other research
- State limitations of the study
- State importance of findings
- Announce directions for further research
- Acknowledgements
- References

Adapted from *Hengl, T. and Gould, M., 2002. Rules of thumb for writing research articles.*

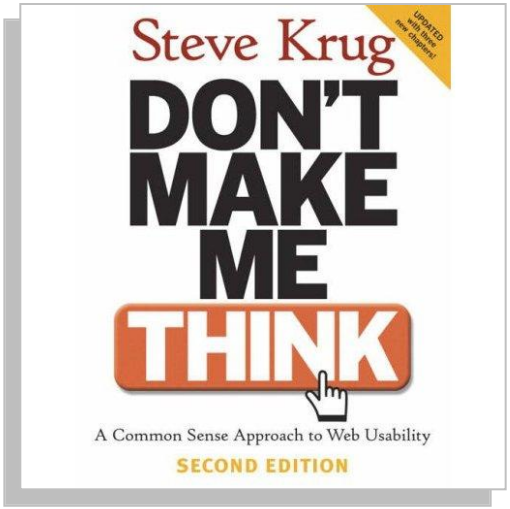
The Curse of Knowledge



- In 1990 Elizabeth Newton (Stanford), did an experiment with “tappers” and “listeners”.
- The “tappers” received a list of well-known songs that they had to tap out on a table to the “listeners”. The “listener” had to guess the song being “tapped.”
- The “tappers” were required to guess how often the “listeners” would guess a song correctly. The “tappers” guessed 50% when the reality was 2.5%. Why such a huge margin of error?

More details in: Made to Stick: Why Some Ideas Survive and Others Die. By Chip &Dan Heath

A Useful Principle



Steve Krug has a wonderful book about web design, which also has some useful ideas for writing papers.

A fundamental principle is captured in the title:

Don't make the reviewer of your paper think!

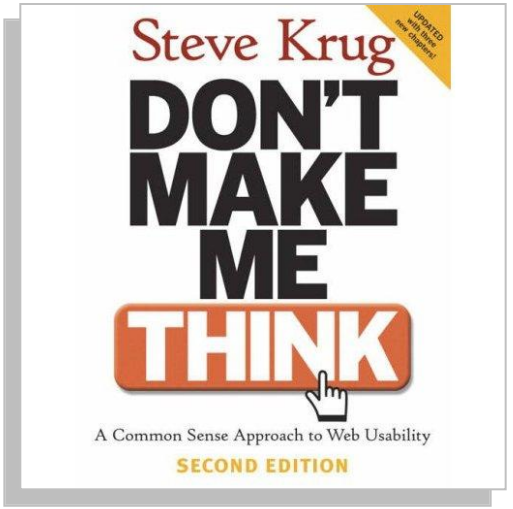
1) If they are forced to think, they may resent being forced to make the effort. The are literally not being paid to think.

2) If you let the reader think, they may think wrong!

With very careful writing, great organization, and self explaining figures, you can (and should) remove most of the effort for the reviewer

A Useful Principle

A simple concrete example:



This requires a lot of thought to see that 2DDW is better than Euclidian distance

This does not

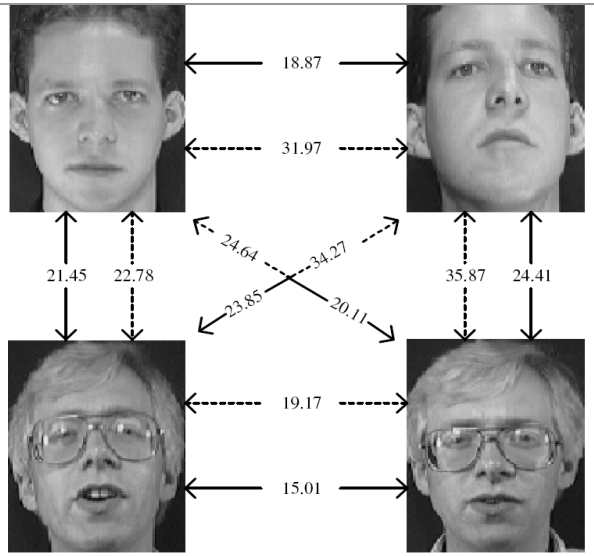


Figure 3. The distances of four faces by 2DDW and Euclidean norm. The 2DDW distances are shown on solid lines and Euclidean distances on dotted lines.

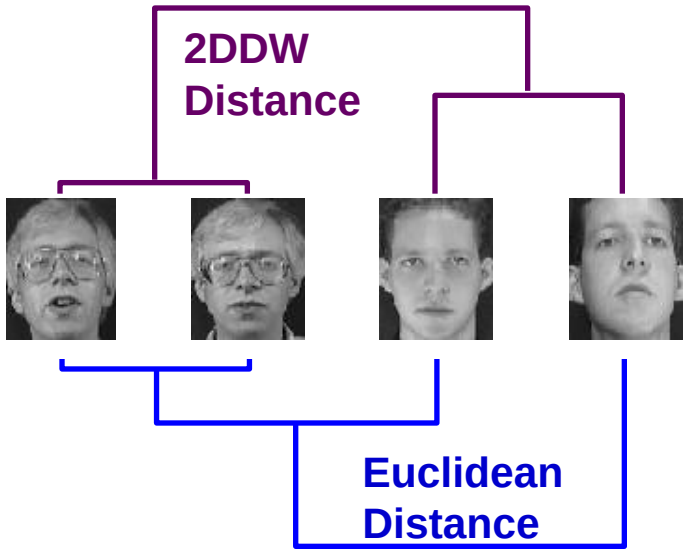


Figure 3: Two pairs of faces clustered using 2DDW (*top*) and Euclidean distance (*bottom*)

Keogh's Maxim

I firmly believe in the following:

If you can save the reviewer one minute of their time, by spending one extra hour of your time, then you have an obligation to do so.

Keogh's Maxim can be derived from first principles

- The author sends about **one** paper to SIGKDD
- The reviewer must review about **ten** papers for SIGKDD

- The benefit for the author in getting a paper into SIGKDD is hard to quantify, but could be tens of thousands of dollars (if you get tenure, if you get that job in Google…).
- The benefit for a reviewer is close to zero, they don't get paid.

Therefore: The author has the responsibly to do *all* the work to make the reviewers task as easy as possible.

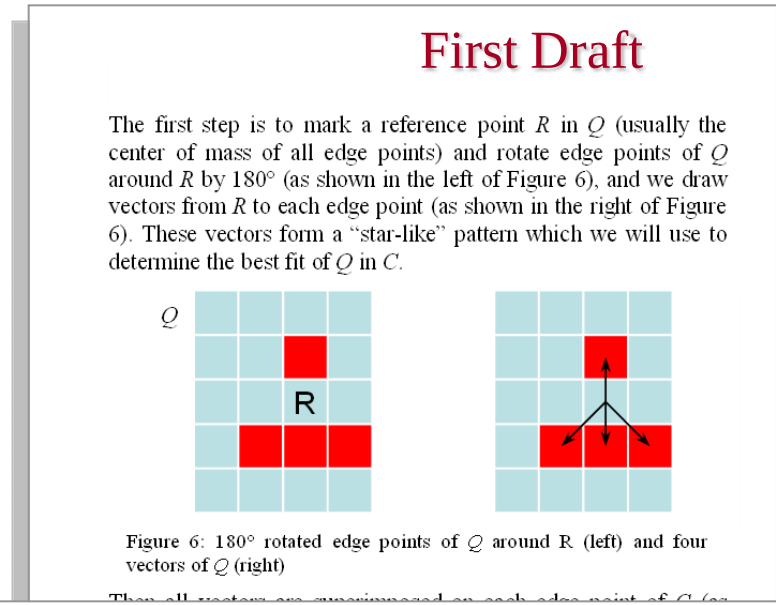
Alan Jay Smith

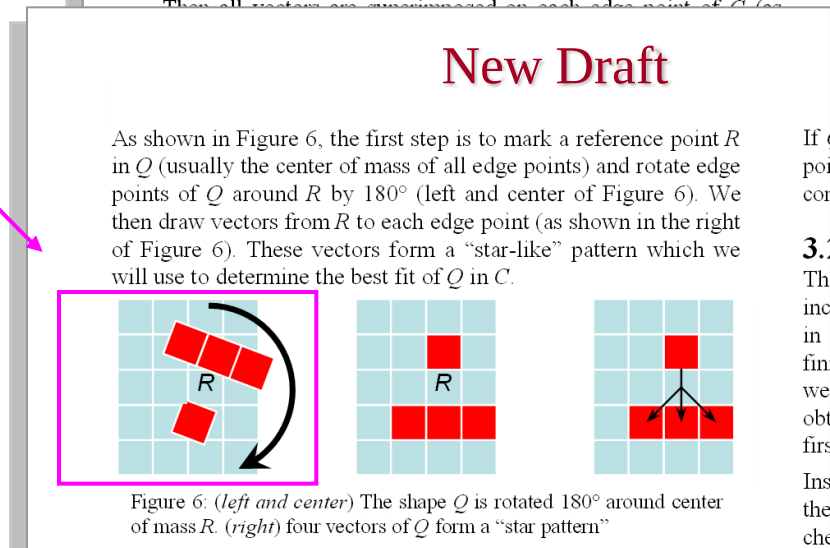
Remember, each report was prepared without charge by someone whose time you could not buy

A. J. Smith, "The task of the referee" IEEE Computer, vol. 23, no. 4, pp. 65-71, April 1990.

An example of Keogh's Maxim

- We wrote a paper for SIGKDD 2009
- Our mock reviewers had a hard time understanding a step, where a template must be rotated. They all eventually got it, it just took them some effort.
- We rewrote some of the text, and added in a figure that explicitly shows the template been rotated
- We retested the section on the same, and new mock reviewers, it worked much better.
- We spent 2 or 3 hours to save the reviewers tens of seconds.





I have often said reviewers make an initial impression on the first page and don't change 80% of the time

Mike Pazzani

This idea, that first impressions tend to be hard to change, has a formal name in psychology, *Anchoring*.

Others have claimed that *Anchoring* is used by reviewers

The Most Important Part of Your Paper: the Introduction

- The 1/3 – 2/3 Rule from a reviewer's perspective:
  - 1/3 time to read your introduction and make a decision
  - Remaining 2/3 time to find evidence for the decision
- [Take-Home Message #6] **A good introduction with a good motivation is half of your success!**

Xindong Wu

Another strategy people seem to use intuitively and unconsciously to simplify the task of making judgments is called ***anchoring****. Some natural starting point is used as a first approximation to the desired judgment. This starting point is then adjusted, based on the results of additional information or analysis. Typically, however, the starting point serves as an anchor that reduces the amount of adjustment, so the final estimate remains closer to the starting point than it ought to be.*

Richards J. Heuer, Jr. Psychology of Intelligence Analysis (CIA)

What might be the "natural starting point" for a SIGKDD reviewer making a judgment on your paper?

Hopefully it is not the author or institution: "*people from CMU tend to do good work, lets have a look at this…"*, "*This guys last paper was junk.."*

I believe that the title, abstract and introduction form an anchor. If these are excellent, then the reviewer reads on assuming this is a good paper, and she is looking for things to confirm this.

However, if they are poor, the reviewer is just going to scan the paper to confirm what she already knows,"*this is junk*"

I don't have any studies to support this for reviewing papers. I am making this claim based on my experience and feedback (*The title is the most important part of the paper*. Jeff Scargle). However there are dozens of studies to support the idea of anchoring when people make judgments about buying cars, stocks, personal injury amounts in court cases etc.

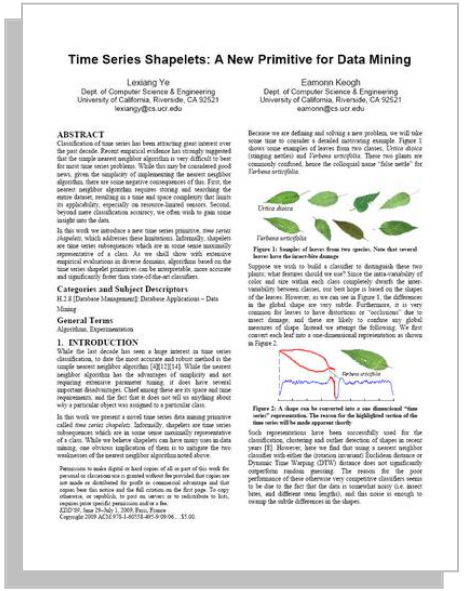
The First Page as an Anchor

Jennifer Windom

The introduction acts as an anchor. By the end of the introduction the reviewer *must* know.

- What is the problem?
- Why is it interesting and important?
- Why is it hard? why do naive approaches fail?
- Why hasn't it been solved before? (Or, what's wrong with previous proposed solutions?)
- What are the key components of my approach and results? Also include any specific limitations.
- A final paragraph or subsection: “Summary of Contributions”. It should list the major contributions in bullet form, mentioning in which sections they can be found. This material doubles as an outline of the rest of the paper, saving space and eliminating redundancy.

If possible, an interesting figure on the first page helps



Time Series Shapelets: A New Primitive for Data Mining

Lexiang Ye
Dept. of Computer Science & Engineering
University of California, Riverside, CA 92521
lexiangy@cs.ucr.edu

Eamonn Keogh
Dept. of Computer Science & Engineering
University of California, Riverside, CA 92521
eamonn@cs.ucr.edu

ABSTRACT

Categories and Subject Descriptors

General Terms

1. INTRODUCTION

This advice is taken almost verbatim from Jennifer.

Reproducibility

Reproducibility is one of the main principles of the scientific method, and refers to the ability of a test or experiment to be accurately reproduced, or replicated, by someone else working independently.

Reproducibility

- In a "bake-off" paper Veltkamp and Latecki attempted to reproduce the accuracy claims of 15 shape matching papers but discovered to their dismay that they could not match the claimed accuracy for *any* approach.
- A recent paper in VLDB showed a similar thing for time series distance measures.

The vast body of results being generated by current computational science practice suffer a large and growing credibility gap: it is impossible to believe most of the computational results shown in conferences and papers

David Donoho

Properties and Performance of Shape Similarity Measures. Remco C. Veltkamp and Longin Jan Latecki. IFCS 2006
Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. Ding, Trajcevski, Scheuermann, Wang & Keogh. VLDB 2008
Fifteen Years of Reproducible Research in Computational Harmonic Analysis- Donoho et al.

Two Types of Non-Reproducibility

- **Explicit**: The authors don't give you the data, or they don't tell you the parameter settings.

- **Implicit**: The work is so complex that it would take you weeks to attempts to reproduce the results, or you are forced to buy expensive software/ hardware/data to attempt reproduction.

Or, the authors *do* give distribute data/code, but it is not annotated or is so complex as to be an unnecessary large burden to work with.

We approximated collections of time series, using algorithms AgglomerativeHistogram and FixedWindowHistogram and utilized the techniques of Keogh et. al., in the problem of querying collections of time series based on similarity. Our results, indicate that the histogram approximations resulting from our algorithms are far superior than those resulting from the APCA algorithm of Keogh et. al.,The superior quality of our histograms is reflected in these problems by reducing the number of false positives during time series similarity indexing, while remaining competitive in terms of the time required to approximate the time series.

Explicit Non Reproducibility

This paper appeared in ICDE02. The "experiment" is shown in its entirety, there are no extra figures or details.

Which collections? How large? What kind of data? How are the queries selected?

What results?

superior by how much?, as measured how?

How competitive?, as measured how?

We approximated collections of time series, using algorithms AgglomerativeHistogram and FixedWindowHistogram and utilized the techniques of Keogh et. al., in the problem of querying collections of time series based on similarity. Our results, indicate that the histogram approximations resulting from our algorithms are far superior than those resulting from the APCA algorithm of Keogh et. al.,The superior quality of our histograms is reflected in these problems by reducing the number of false positives during time series similarity indexing, while remaining competitive in terms of the time required to approximate the time series.

I got a collection of opera arias as sung by Luciano Pavarotti, I compared his recordings to my own renditions of the songs. My results, indicate that my performances are far superior to those by Pavarotti. The superior quality of my performance is reflected in my mastery of the highest notes of a tenor's range, while remaining competitive in terms of the time required to prepare for a performance.

Implicit Non Reproducibility

From a recent paper:

> *This forecasting model integrates a case based reasoning (CBR) technique, a Fuzzy Decision Tree (FDT), and Genetic Algorithms (GA) to construct a decision-making system based on historical data and technical indexes*.

- In order to begin reproduce this work, we have to implement a Case Based Reasoning System and a Fuzzy Decision Tree and a Genetic Algorithm.
- With rare exceptions, people don't spend a month reproducing someone else's results, so this is effectively non-reproducible.
- Note that it is not the extraordinary complexity of the work that makes this non-reproducible (although it does not help), if the authors had put free high quality code and data online…

Why Reproducibility?

• We could talk about reproducibility as the cornerstone of scientific method and an *obligation* to the community, to your funders etc. However this tutorial is about *getting papers published*.

• Having highly reproducible research *will* greatly help your chances of getting your paper accepted.

• Explicit efforts in reproducibility instill confidence in the reviewers that your work is correct.

•Explicit efforts in reproducibility will give the (true) appearance of value.

As a bonus, reproducibility will increase your number of citations.

How to Ensure Reproducibility

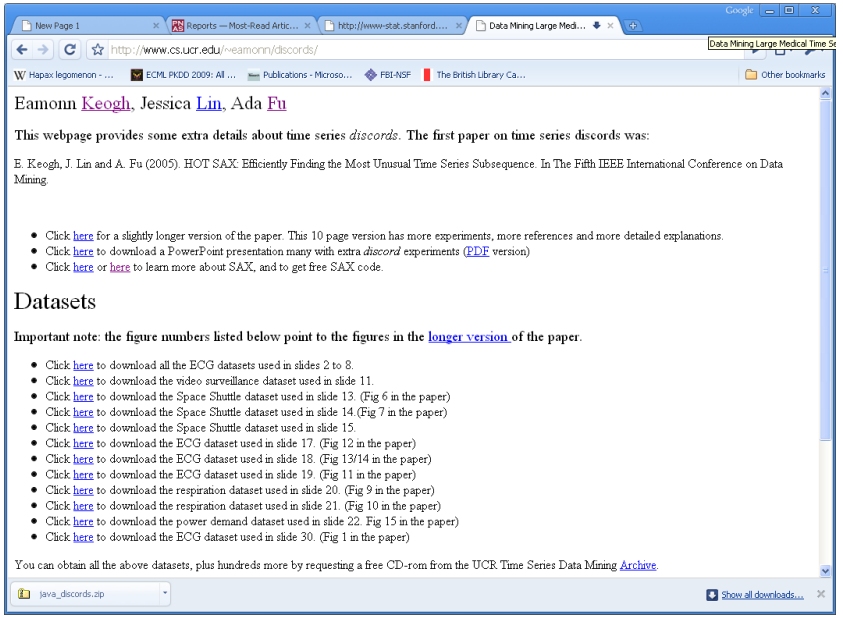
- Explicitly state *all* parameters and settings in your paper.
- Build a webpage with annotated data and code and point to it

  (Use an anonymous hosting service if necessary for double blind reviewing)

- It is too easy to fool yourself into thinking your work is reproducible when it is not. Someone other than you should test the reproducibly of the paper.

(from the paper)

Reproducible Results Statement: In the interests of competitive scientific inquiry, all datasets used in this work are available at the following URL [6]. This research was partly funded by the National Science Foundation under grant IIS-0237918.



For double blind review conferences, you can create a Gmail account or Google Docs account, place all data there, and put the account info in the paper.

How to Ensure Reproducibility

In the next few slides I will quickly dismiss commonly heard objections to reproducible research (with thanks to David Donoho)

- I can't share my data for privacy reasons.
- Reproducibility takes too much time and effort.
- Strangers will use your code/data to compete with you.
- No one else does it. I won't get any credit for it.

But I can't share my data for privacy reasons…

- My first reaction when I see this is to think it may not be true. If you a going to claim this, *prove* it.

  (Yes, *prove* it. Point to a webpage that shows the official policy of the funding agency, or university etc. Explain why your work falls under this policy)

- Can you *also* get a dataset that you can release?

- Can you *make* a dataset that you can publicly release, which is about the same size, cardinality, distribution as the private dataset, then test on *both* in you paper, and release the synthetic one?

Reproducibility takes too much time and effort

- First of all, this has not been my personal experience.

- Reproducibility can *save* time. When your conference paper gets invited to a journal a year later, and you need to do more experiments, you will find it much easier to pick up were you left off.

- Forcing grad students/collaborators to do reproducible research makes them much easier to work with.

Strangers will use your code/data to compete with you

- But competition means "*strangers will read your papers and try to learn from them and try to do even better*". If you prefer obscurity, why are you publishing?
- Other people using your code/data is something that funding agencies and tenure committees *love* to see.

Sometimes the competition is undone by their carelessness. Below (center) is a figure from a paper that uses my publicly available datasets. The alleged shapes in their paper are clearly not the real shapes (confusion of Cartesian and polar coordinates?). This is good example of the importance of the "*Send preview to the rival authors*". This would have avoided publishing such an embarrassing mistake.

Alleged Arrowhead and Diatoms

Actual Arrowhead

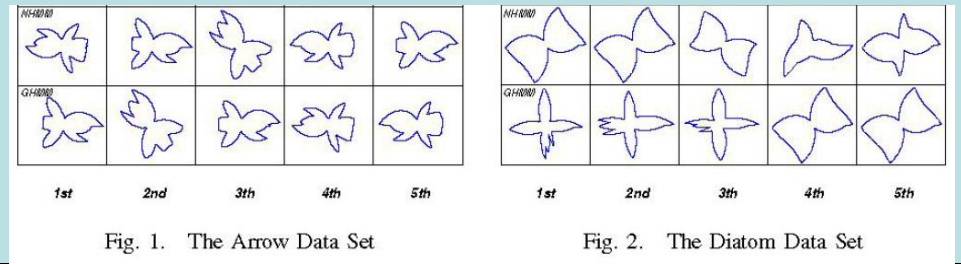


Fig. 1. The Arrow Data Set

Fig. 2. The Diatom Data Set

Actual Diatoms

No one else does it. I won't get any credit for it

- It is true that not everyone does it, but that just means that you have a way to stand above the competition.
- A review of my SIGKDD 2004 paper said (my paraphrasing, I have lost the original email).

The results seem to good to be true, but I had my grad student download the code and data and check the results, it really does work as well as they claim.

Parameters (are bad)

- The most common cause of **Implicit Non Reproducibility** is a algorithm with many parameters.
- Parameter-laden algorithms can seem (and often *are*) ad-hoc and brittle.
- Parameter-laden algorithms decrease reviewer confidence.
- For *every* parameter in your method, you *must* show, by logic, reason or experiment, that either…
  - There is some way to set a good value for the parameter.
  - The exact value of the parameter makes little difference.

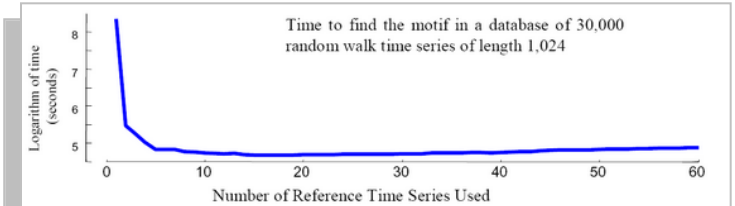


Figure 8: A plot of execution time vs. the number of reference points. Note that once the number of reference points is beyond say five, its exact value makes little difference. Note the log scale of the time axis

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk

John von Neumann

Unjustified Choices (are bad)

- It is important to explain/justify *every* choice, even if it was an arbitrary choice.
- **For example, this line frustrated me:** *Of the 300 users with enough number of sessions within the year, we randomly picked 100 users to study.* Why 100? Would we have gotten similar results with 200?

- **Bad:** *We used single linkage clustering...* Why single linkage, why not group average or Wards?
- **Good:** *We experimented with single/group/complete linkage, but found this choice made little difference, we therefore report only…*
- **Better:** *We experimented with single/group/complete linkage, but found this choice little difference, we therefore report only single linkage in this paper, however the interested reader can view the tech report [a] to see all variants of clustering.*

Important Words/Phrases I

- **Optimal**: Does *not* mean "very good"
  - *We picked the optimal value for X… No!* (unless you can prove it)
  - *We picked a value for X that produced the best..*
- **Proved**: Does *not* mean "demonstrated"
  - *With experiments we proved that our.. No!* (experiments rarely prove things)
  - *With experiments we offer evidence that our..*
- **Significant**: There is a danger of confusing the informal statement and the statistical claim
  - *Our idea is significantly better than Smiths*
  - *Our idea is statistically significantly better than Smiths, at a confidence level of…*

Important Words/Phrases II

- **Complexity**: Has an overloaded meaning in computer science
  - *The X algorithms complexity means it is not a good solution (complex= intricate)*
  - *The X algorithms time complexity is $O(n^6)$ meaning it is not a good solution*
- **It is easy to see**: First, this is a cliché. Second, are you sure it is easy?
  - *It is easy to see that P = NP*
- **Actual**: Almost always has no meaning in a sentence
  - *It is an actual B-tree -> It is a B-tree*
  - *There are actually 5 ways to hash a string -> There are 5 ways to hash a string*
- **Theoretically**: Almost always has no meaning in a sentence
  - *Theoretically we could have jam or jelly on our toast.*
- **etc** : Only use it if the remaining items on the list are obvious.
  - *We named the buckets for the 7 colors of the rainbow, red, orange, yellow etc.*
  - *We measure performance factors such as stability, scalability, etc.* *No!*

Important Words/Phrases III

- **Correlated:** In informal speech it is a synonym for "related"
  - *Celsius and Fahrenheit are correlated.* (clearly correct, perfect linear correlation)
  - *The tightness of lower bounds is correlated with pruning power.* *No!*
- **(Data) Mined**
  - *Don't say "We mined the data…", if you can say "We clustered the data.." or "We classified the data…" etc*

Important Words/Phrases IIII

- **In this paper:** Where else? We are reading *this* paper

From a single SIGMOD paper

- **In this paper**, we attempt to approximate..
- Thus, **in this paper**, we explore how to use..
- **In this paper**, our focus is on indexing large collections..
- **In this paper**, we seek to approximate and index..
- Thus, **in this paper**, we explore how to use the..
- The indexing proposed **in this paper** belongs to the class of..
- Figure 1 summarizes all the symbols used **in this paper**…
- **In this paper**, we use Euclidean distance..
- The results to be presented **in this paper** do not..
- A key result to be proven later **in this paper** is that the..
- **In this paper**, we adopt the Euclidean distance function..
- **In this paper**, we explore how to apply

DABTAU

DHT is used

DHT is used

and again

and again

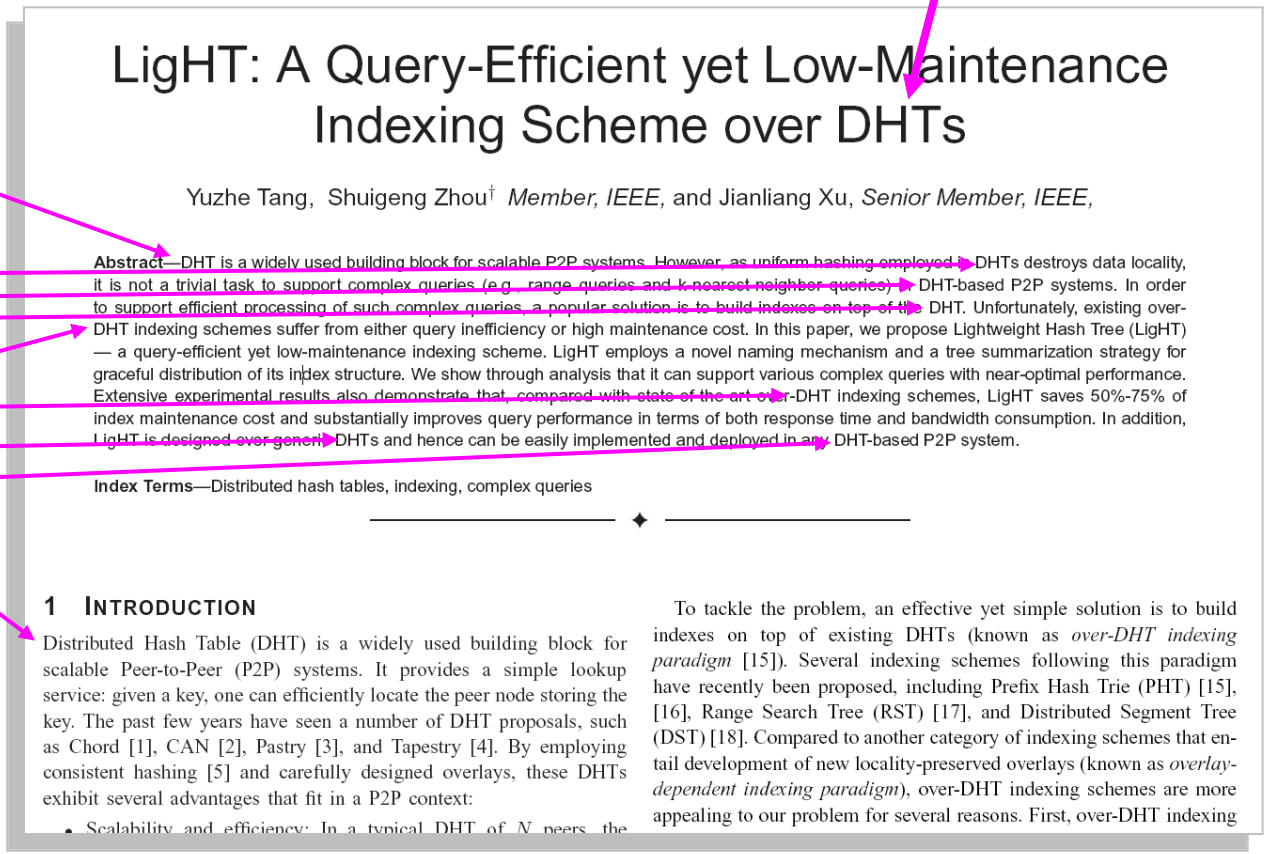
and again

and again

and again

and again

DHT is finally defined!



LigHT: A Query-Efficient yet Low-Maintenance Indexing Scheme over DHTs

Yuzhe Tang, Shuigeng Zhou† *Member, IEEE,* and Jianliang Xu, *Senior Member, IEEE,*

Abstract—DHT is a widely used building block for scalable P2P systems. However, as uniform hashing employed by DHTs destroys data locality, it is not a trivial task to support complex queries (e.g., range queries and k-nearest neighbor queries) in DHT-based P2P systems. In order to support efficient processing of such complex queries, a popular solution is to build indexes on top of the DHT. Unfortunately, existing over-DHT indexing schemes suffer from either query inefficiency or high maintenance cost. In this paper, we propose Lightweight Hash Tree (LigHT) — a query-efficient yet low-maintenance indexing scheme. LigHT employs a novel naming mechanism and a tree summarization strategy for graceful distribution of its index structure. We show through analysis that it can support various complex queries with near-optimal performance. Extensive experimental results also demonstrate that, compared with state-of-the-art over-DHT indexing schemes, LigHT saves 50%-75% of index maintenance cost and substantially improves query performance in terms of both response time and bandwidth consumption. In addition, LigHT is designed over generic DHTs and hence can be easily implemented and deployed in any DHT-based P2P system.

Index Terms—Distributed hash tables, indexing, complex queries

1 INTRODUCTION

Distributed Hash Table (DHT) is a widely used building block for scalable Peer-to-Peer (P2P) systems. It provides a simple lookup service: given a key, one can efficiently locate the peer node storing the key. The past few years have seen a number of DHT proposals, such as Chord [1], CAN [2], Pastry [3], and Tapestry [4]. By employing consistent hashing [5] and carefully designed overlays, these DHTs exhibit several advantages that fit in a P2P context:

- Scalability and efficiency: In a typical DHT of N peers, the

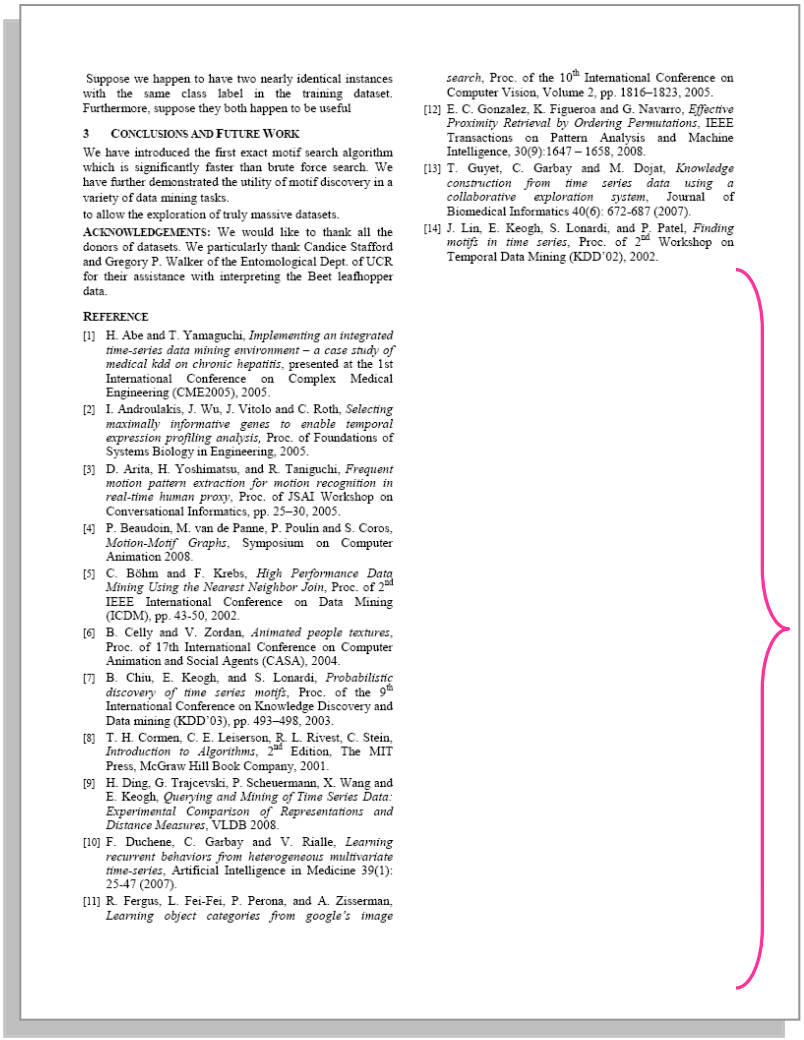
To tackle the problem, an effective yet simple solution is to build indexes on top of existing DHTs (known as *over-DHT indexing paradigm* [15]). Several indexing schemes following this paradigm have recently been proposed, including Prefix Hash Trie (PHT) [15], [16], Range Search Tree (RST) [17], and Distributed Segment Tree (DST) [18]. Compared to another category of indexing schemes that entail development of new locality-preserved overlays (known as *overlay-dependent indexing paradigm*), over-DHT indexing schemes are more appealing to our problem for several reasons. First, over-DHT indexing

It is very important that you DABTAU or your readers may be confused.
(Define Acronyms Before They Are Used)

But anyone that reviews for this conference will surely know what the acronym means!
Don't be so sure, your reviewer may be a first-year, non-native English-speaking grad student that got 15 papers dumped on his desk 3 days before the reviewing deadline.

You can only assume this for acronyms where we have forgotten the original words, like **laser**, **radar, Scuba**. Remember our principle, **don't make the reviewer think**.

Use *all* the Space Available



Suppose we happen to have two nearly identical instances with the same class label in the training dataset. Furthermore, suppose they both happen to be useful

3 CONCLUSIONS AND FUTURE WORK

We have introduced the first exact motif search algorithm which is significantly faster than brute force search. We have further demonstrated the utility of motif discovery in a variety of data mining tasks.

to allow the exploration of truly massive datasets.

ACKNOWLEDGEMENTS: We would like to thank all the donors of datasets. We particularly thank Candice Stafford and Gregory P. Walker of the Entomological Dept. of UCR for their assistance with interpreting the Beet leafhopper data.

REFERENCE

[1] H. Abe and T. Yamaguchi, *Implementing an integrated time-series data mining environment – a case study of medical kdd on chronic hepatitis*, presented at the 1st International Conference on Complex Medical Engineering (CME2005), 2005.

[2] I. Androulakis, J. Wu, J. Vitolo and C. Roth, *Selecting maximally informative genes to enable temporal expression profiling analysis*, Proc. of Foundations of Systems Biology in Engineering, 2005.

[3] D. Arita, H. Yoshimatsu, and R. Taniguchi, *Frequent motion pattern extraction for motion recognition in real-time human proxy*, Proc. of JSAI Workshop on Conversational Informatics, pp. 25–30, 2005.

[4] P. Beaudoin, M. van de Panne, P. Poulin and S. Coros, *Motion-Motif Graphs*, Symposium on Computer Animation 2008.

[5] C. Böhm and F. Krebs, *High Performance Data Mining Using the Nearest Neighbor Join*, Proc. of 2nd IEEE International Conference on Data Mining (ICDM), pp. 43-50, 2002.

[6] B. Celly and V. Zordan, *Animated people textures*, Proc. of 17th International Conference on Computer Animation and Social Agents (CASA), 2004.

[7] B. Chiu, E. Keogh, and S. Lonardi, *Probabilistic discovery of time series motifs*, Proc. of the 9th International Conference on Knowledge Discovery and Data mining (KDD'03), pp. 493–498, 2003.

[8] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, *Introduction to Algorithms*, 2nd Edition, The MIT Press, McGraw Hill Book Company, 2001.

[9] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang and E. Keogh, *Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures*, VLDB 2008.

[10] F. Duchene, C. Garbay and V. Rialle, *Learning recurrent behaviors from heterogeneous multivariate time-series*, Artificial Intelligence in Medicine 39(1): 25-47 (2007).

[11] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, *Learning object categories from google's image search*, Proc. of the 10th International Conference on Computer Vision, Volume 2, pp. 1816–1823, 2005.

[12] E. C. Gonzalez, K. Figueroa and G. Navarro, *Effective Proximity Retrieval by Ordering Permutations*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(9):1647 – 1658, 2008.

[13] T. Guyet, C. Garbay and M. Dojat, *Knowledge construction from time series data using a collaborative exploration system*, Journal of Biomedical Informatics 40(6): 672-687 (2007).

[14] J. Lin, E. Keogh, S. Lonardi, and P. Patel, *Finding motifs in time series*, Proc. of 2nd Workshop on Temporal Data Mining (KDD'02), 2002.

Some reviewer is going to look at this empty space and say..

They could have had an additional experiment

They could have had more discussion of related work

They could have referenced more of my papers

etc

The best way to write a great 9 page paper, is to write a good 12 or 13 page paper and carefully pare it down.

You can use Color in the Text

In the example to the right, color helps emphasize that the order in which bits are added/removed to a representation.

In the example below, color links numbers in the text with numbers in a figure.

Bear in mind that the reader may not see the color version, so you cannot *rely* on color.

The astute reader will have noted that once we have $\mathbf{T}^4$ we can derive $\mathbf{T}^2$ simply by ignoring the trailing bits in each of the four symbols in the SAX word. As one can readily imagine, this is a recursive property. For example, if we convert T to SAX with a cardinality of 8, we have SAX(T,4,8) = $\mathbf{T}^8$ = {110,110,011,000}. From this, we can convert to any lower resolution that differs by a power of two, simply by ignoring the correct number of bits. Table 3 makes this clearer.

Table 3: It is possible to obtain a reduced (by half) cardinality SAX word simply by ignoring trailing bits

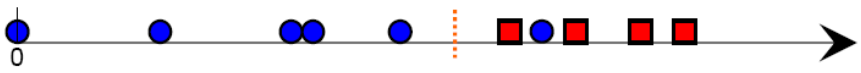
SAX(T,4,16) = $\mathbf{T}^{16}$ =	{1100,1101,0110,0001}
SAX(T,4,8) = $\mathbf{T}^8$ =	{110 ,110 ,011 ,000 }
SAX(T,4,4) = $\mathbf{T}^4$ =	{11 ,11 ,01 ,00 }
SAX(T,4,2) = $\mathbf{T}^2$ =	{1 ,1 ,0 ,0 }

As we shall see later, this ability to change cardinalities on the fly is a useful and exploitable property.

SIGKDD 2008

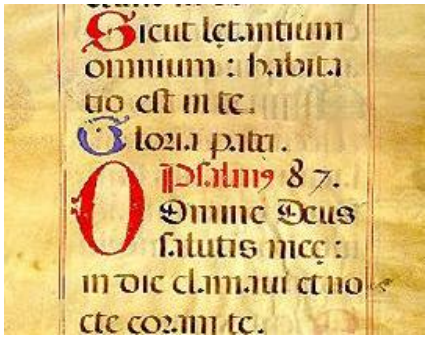
example. Suppose as shown in Figure 8, ten time series objects are arranged in a one-dimensional representation by measuring their distance to the best-so-far candidate. This happens to be a good case, with five of the six objects from class *A* (represented by circles) closer to the candidate than any of the four objects from class *B* (represented by squares). In addition, of the five objects to the right of the split point, only one object from class *A* is mixed up with the class *B*. The optimal split point is represented by a vertical dashed line, and the best-so-far information gain is:

[-(6/10)log(6/10)-(4/10)log(4/10)] - [(5/10)[-(5/5)log(5/5)]+(5/10)[-(4/5)log(4/5)-(1/5)log(1/5)]]=0.4228



SIGKDD 2009

People have been using color this way for well over a 1,000 years



Avoid Weak Language I

Compare

..with a dynamic series, it might fail to give accurate results.

With..

..with a dynamic series, it has been shown by [7] *to give inaccurate results.* (give a concrete reference)

Or..

..with a dynamic series, it will give inaccurate results, as we show in Section 7. (show me numbers)

Avoid Weak Language II

Compare

In this paper, we attempt to approximate and index a d-dimensional spatio-temporal trajectory..

With…

In this paper, we approximate and index a d-dimensional spatio-temporal trajectory..

Or…

In this paper, we show, for the first time, how to approximate and index a d-dimensional spatio-temporal trajectory..

Avoid Weak Language III

The paper is aiming to detect and retrieve videos of the same scene…

Are you *aiming* at doing this, or have you done it? Why not say…

In this work, we introduce a novel algorithm to detect and retrieve videos..

The DTW algorithm tries to find the path, minimizing the cost..

The DTW does not try to do this, it *does* this.

The DTW algorithm finds the path, minimizing the cost..

Monitoring aggregate queries in real-time over distributed streaming environments appears to be a great challenge.

Appears to be, or is? Why not say…

Monitoring aggregate queries in real-time over distributed streaming environments is known to be a great challenge [1,2].

Avoid Overstating

Don't say:

We have shown our algorithm is better than a decision tree.

If you really mean…

We have shown our algorithm can be better than decision trees, when the data is correlated.

Or..

On the Iris and Stock dataset, we have shown that our algorithm is more accurate, in future work we plan to discover the conditions under which our...

Use the Active Voice

It can be seen that…

"seen" by whom?

Experiments were conducted…

The data was collected by us.

We can see that…

We conducted experiments...

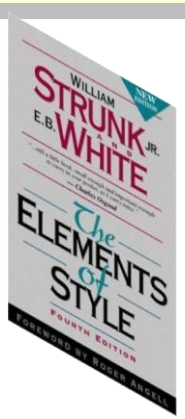
Take responsibility

We collected the data.

Active voice is often shorter

William Strunk, Jr

The active voice is usually more direct and vigorous than the passive



Avoid Implicit Pointers

Consider the following sentence:

"*We used DFT. It has circular convolution property but not the unique eigenvectors property. This allows us to…*"

What does the "*This*" refer to?

- The use of DFT?
- The convolution property?
- The unique eigenvectors property?

Check every occurrence of the words "it", "this", "these" etc. Are they used in an unambiguous way?

Avoid nonreferential use of "this", "that", "these", "it", and so on.

Jeffrey D. Ullman

Many papers read like this:

We invented a new problem, and guess what, we can solve it!

This paper proposes a new trajectory clustering scheme for objects moving on road networks. A trajectory on road networks can be defined as a sequence of road segments a moving object has passed by. We first propose a similarity measurement scheme that judges the degree of similarity by considering the total length of matched road segments. Then, we propose a new clustering algorithm based on such similarity measurement criteria by modifying and adjusting the FastMap and hierarchical clustering schemes. To evaluate the performance of the proposed clustering scheme, we also develop a trajectory generator considering the fact that most objects tend to move from the starting point to the destination point along their shortest path. The performance result shows that our scheme has the accuracy of over 95%.

When the authors invent the definition of the data, and they invent the problem, and they invent the error metric, and they make their own data, can we be surprised if they have high accuracy?

Motivating your Work

If there is a different way to solve your problem, and you do not address this, your reviewers might think you are hiding something

You should **very explicitly** say why the other ideas will not work. Even if it is obvious to you, it might not be obvious to the reviewer.

Another way to handle this might be to simply code up the other way and compare to it.

It is important to dismiss two apparent solutions to this problem before introducing our technique:

- *Why not replace the Euclidean distance with the Dynamic Time Warping (DTW) distance?* While DTW is a very useful tool for many data mining problems, it is not the solution here. For example, if we have a subsequence of length 500 that contains 10 heartbeats, and another subsequence of length 500 that contains 9 heartbeats, DTW is no more useful than Euclidean distance, because DTW must match *every* data point in each sequence, and there is no meaningful way to map 9 heartbeats to 10 heartbeats. What is required is *uniform scaling*, which compares the original 500 data points to a range of possible data points, say from 500 to 600, incorporating the second sequence.

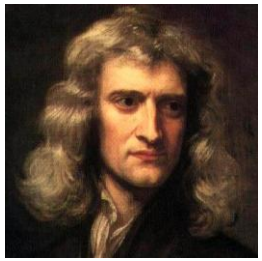
- *Why not search for shorter patterns, and after finding the shorter motifs, somehow "grow" them with invariance to uniform scaling?* This idea does seem attractive initially. In the example in Figure 2, if we shorten the required pattern length to 100 instead of 120, we do find a subsection of A and a subsection of B to be the best motif. The problem is that in most

Give Credit

nanos gigantium humeris insidente

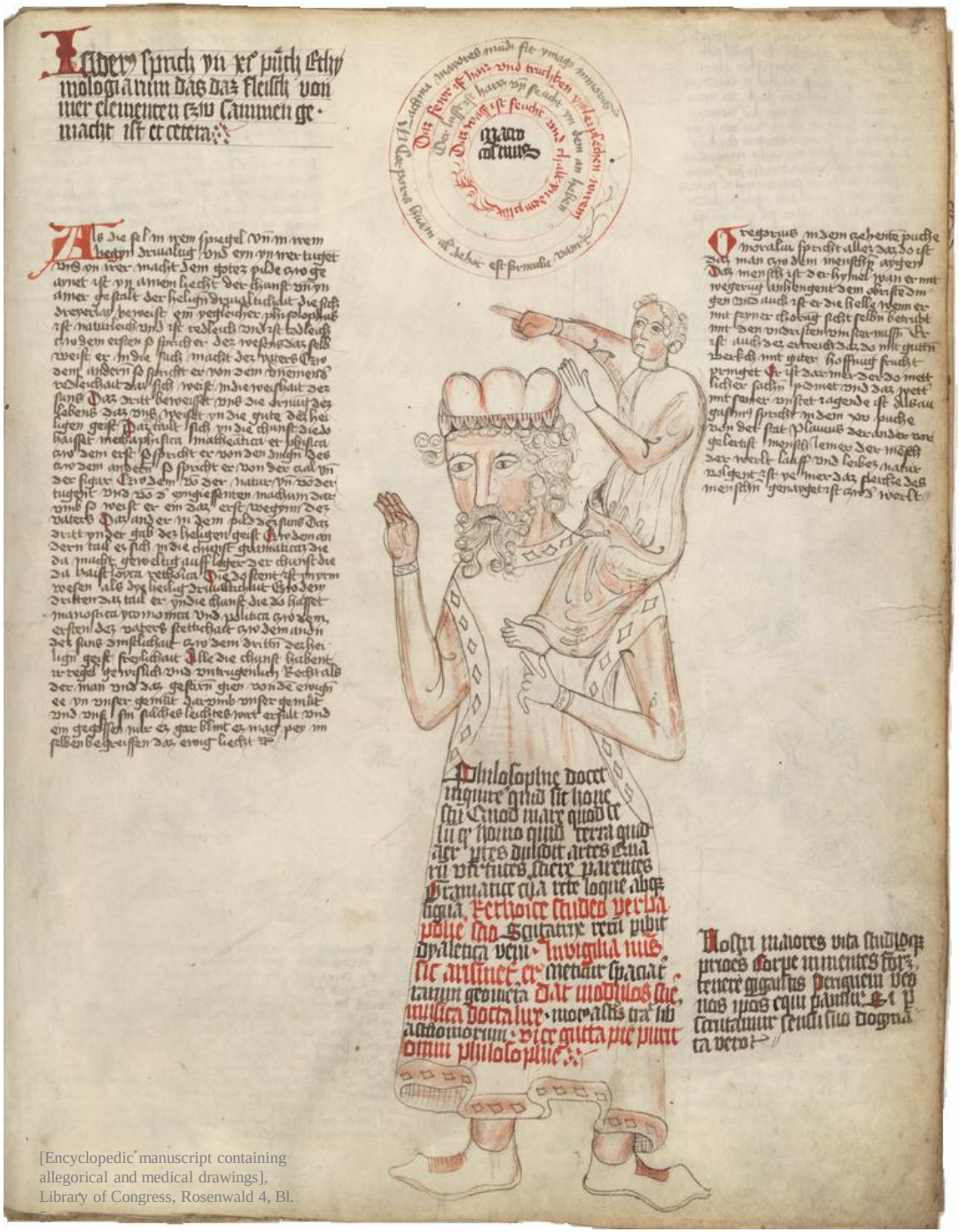
Dwarfs standing on the shoulders of giants

If you are using someone else's ideas as part of your solution to a problem, be sure to fully and explicitly credit their work, a vague reference is not sufficient.



Isaac Newton

If I have seen a little further it is by standing on the shoulders of Giants



[Encyclopedic manuscript containing allegorical and medical drawings], Library of Congress, Rosenwald 4, Bl. 7

Motivation

For reasons I don't understand, SIGKDD papers rarely quote other papers. Quoting other papers can allow the writing of more forceful arguments…

> However, no matter what representation is used, rotation invariance seems to be uniquely difficult to handle. For example [20] notes "*rotation is always something hard to handle compared with translation and scaling*".

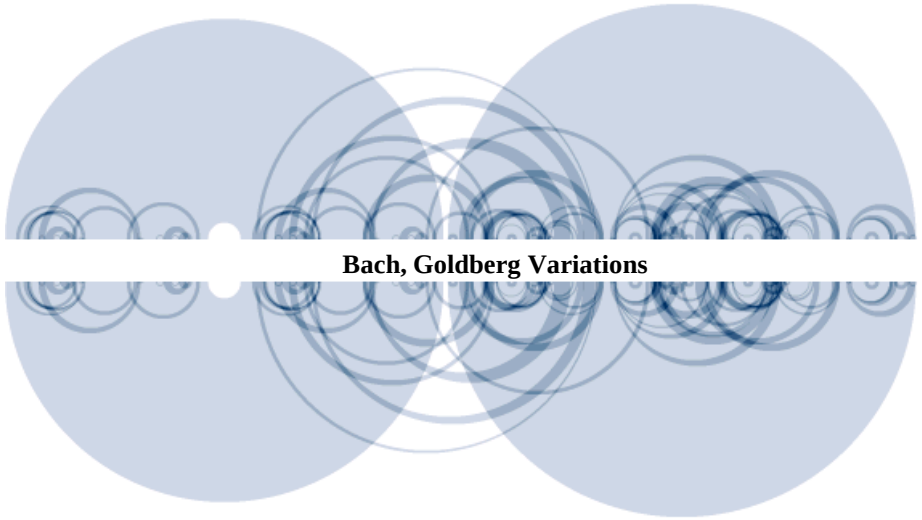
This is much better than..
Paper [20] notes that rotation is hard to deal with.

> For example, a recent paper suggested "*dynamic time warping incurs a heavy CPU cost…*" Surprisingly, as we will now show, the amortized CPU cost of DTW is

This is much better than..
That paper says time warping is too slow.

Motivation

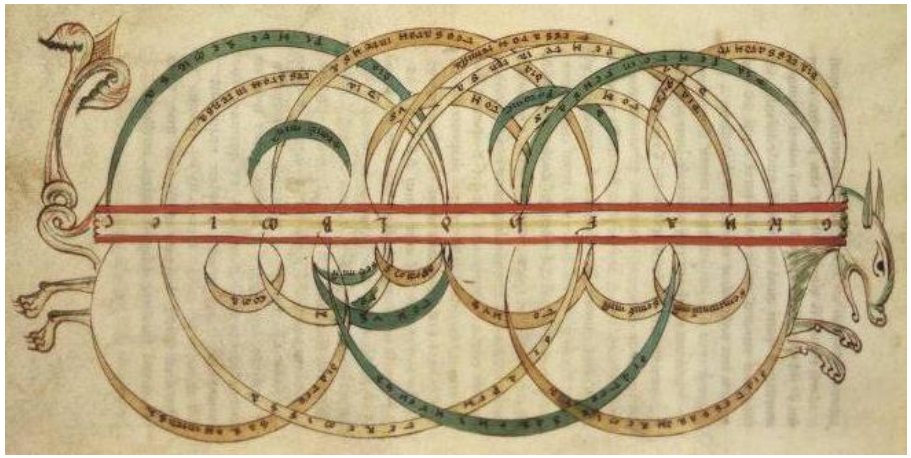
Martin Wattenberg had a beautiful paper in InfoVis 2002 that showed the repeated structure in strings…



If I had reviewed it, I would have rejected it, noting it had already been done, in 1120!

It is very important to convince the reviewers that your work is *original*.

- Do a detailed literature search.
- Use mock reviewers.
- Explain why your work is different (see Avoid "Laundry List" Citations)



De Musica: Leaf from Boethius' treatise on music. Diagram is decorated with the animal form of a beast. Alexander Turnbull Library, Wellington, New Zealand

Avoid "Laundry List" Citations I

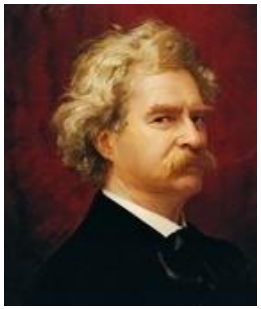
In some of my early papers, I misspelled Davood Rafiei's name *Refiei*. This spelling mistake now shows up in dozens of papers by others…

- Finding Similarity in Time Series Data by Method of Time Weighted ..
- Similarity Search in Time Series Databases Using ..
- Financial Time Series Indexing Based on Low Resolution …
- Similarity Search in Time Series Data Using Time Weighted …
- Data Reduction and Noise Filtering for Predicting Times …
- –Time Series Data Analysis and Pre-process on Large …
- –G probability-based method and its …
- –A Review on Time Series Representation for Similarity-based …
- –Financial Time Series Indexing Based on Low Resolution …
- –A New Design of Multiple Classifier System and its Application to…

This (along with other facts omitted here) suggests that some people copy "classic" references, without having read them.

In other cases I have seen papers that claim "*we introduce a novel algorithm X*", when in fact an essentially identical algorithm appears in one of the papers they have referenced (but probably not read).

Read your references! If what you are doing appears to contradict or duplicate previous work, explicitly address this in your paper.



Mark Twain

A classic is something that everybody wants to have read and nobody wants to read

Avoid "Laundry List" Citations II

One of Carla Brodley's pet peeves is laundry list citations:

"*Paper A says "blah blah" about Paper B, so in my paper I say the same thing, but cite Paper B, and I did not read Paper B to form my own opinion. (and in some cases did not even read Paper B at all....)*"

Carla Brodley

The problem with this is:

- You look like you are lazy.
- You look like you cannot form your own opinion.
- If paper A is wrong about paper B, and you echo the errors, you look naïve.

I dislike broad reference bundles such as *There has been plenty of related work [a,s,d,f,g,c,h]*
Often related work sections are little more than annotated bibliographies.

Claudia Perlich
Chris Drummond

A Common Logic Error in Evaluating Algorithms: Part I

Here the authors test the rival algorithm, DTW, which has no parameters, and achieved an error rate of 0.127.

They then test 64 variations of their own approach, and since there exists at least one combination that is lower than 0.127, they claim that their algorithm "*performs better*"

Note that in this case the error is explicit, because the authors published the table. However in **many** case the authors just publish the result "*we got 0.100*", and it is less clear that the problem exists.

"*Comparing the error rates of DTW (0.127) and those of Table 3, we observe that XXX performs better*"

	number of bins τ			
scale δ	8	16	32	64
1	0.628	0.590	0.563	0.536
2	0.223	0.157	0.123	0.130
3	0.257	0.158	0.140	0.127
4	0.223	0.158	0.140	0.130
5	0.167	0.143	0.128	0.123
6	**0.140**	**0.105**	**0.105**	**0.100**
7	0.172	0.137	0.126	0.268
8	0.182	0.137	0.108	0.298
9	0.190	0.137	0.117	0.253
10	0.122	0.137	0.105	0.248
11	0.168	0.130	0.118	0.298
12	0.153	0.148	0.125	0.298
13	0.197	0.150	0.137	0.268
14	0.243	0.168	0.143	0.298
15	0.243	0.188	0.143	0.298
16	0.243	0.188	0.143	0.298

Table 3: Error rates using XXX on time series histograms with equal bin size

A Common Logic Error in Evaluating Algorithms: Part II

To see why this is a flaw, consider this:

- We want to find the fastest 100m runner, between India and China.
- India does a set of trails, finds its best man, Anil, and Anil turns up expecting a race.
- China ask Anil to run by himself. Although mystified, he obliging does so, and clocks 9.75 seconds.
- China then tells all 1.4 billion Chinese people to run 100m.
- The best of all 1.4 billion runs was Jin, who clocked 9.70 seconds.
- China declares itself the winner!

Is this fair? Of course not, but this is exactly what the previous slide does.

Johannes Fuernkranz

Keep in mind that you should never look at the test set. This may sound obvious, but I cannot longer count the number of papers that I had to reject because of this.

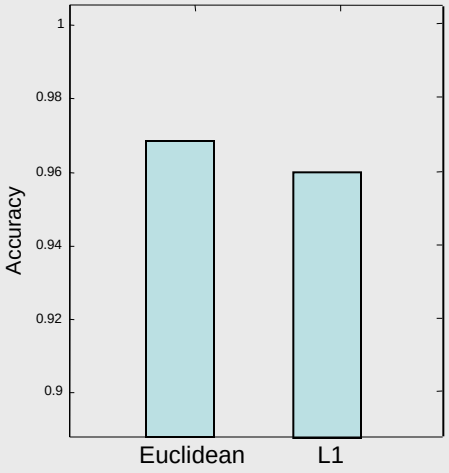
Claudia Perlich

ALWAYS put some variance estimate on performance measures (do everything 10 times and give me the variance of whatever you are reporting)

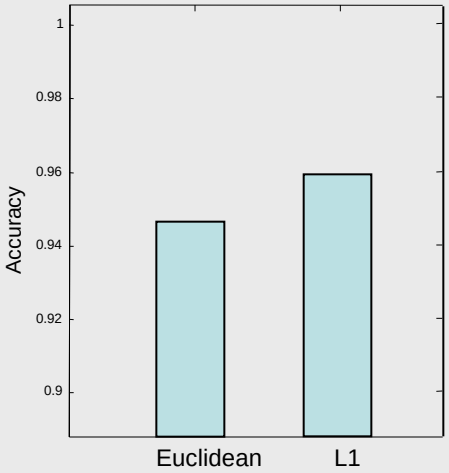
0.8933 0.9600
0.9733 0.9600
0.9867 0.9733
0.9333 0.9467
0.9200 0.9600
0.9200 0.9467
0.9600 1.0000
0.9600 0.9467
0.9467 0.9733
0.9200 0.9600
0.9067 0.9600
0.9067 0.9733
0.9600 0.9867
0.9600 0.9733
0.9200 0.9333
0.9200 0.9333
0.9600 0.9600
0.9467 0.9733
0.9467 0.9600
0.8933 0.9600
0.9200 0.9733
0.9200 0.9200
0.9467 0.9333
0.9200 0.9600
0.9333 0.9733
0.9333 0.9867
0.9867 0.9867
0.9200 0.9733
0.9733 0.9733
0.9333 0.9733
0.9067 0.9333
0.9467 0.9600
0.9333 0.9200
0.9467 0.9467
0.9333 0.9333
0.9600 0.9867
0.9733 0.9867
0.9333 0.9467
0.9600 0.9867
0.9467 0.9600
0.9600 0.9867
0.9733 0.9733
0.9467 0.9867
0.9600 0.9600
0.9467 0.9467
0.9467 0.9600
0.9600 0.9733
0.9333 0.9733
0.9467 0.9733
0.9200 0.9600

Suppose I want to know if Euclidean distance or L1 distance is best on the CBF problem (with 150 objects), using 1NN…

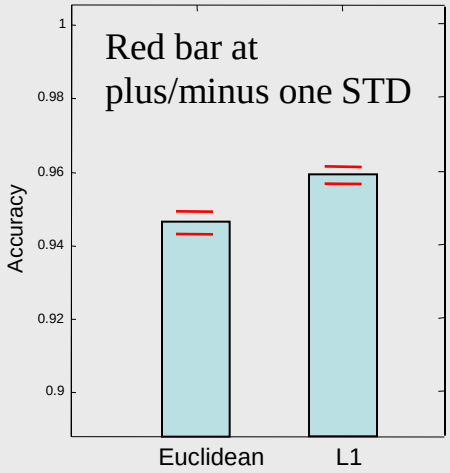
Bad: Do one test



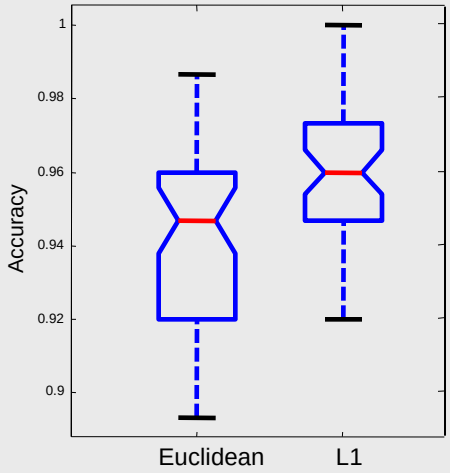
A littler better: Do 50 tests, and report mean



Better: Do 50 tests, report mean and variance



Much Better: Do 50 tests, report confidence

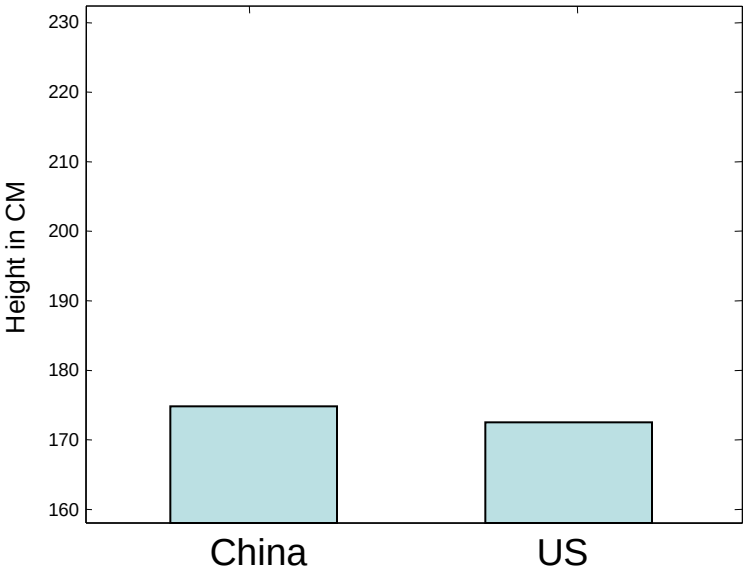


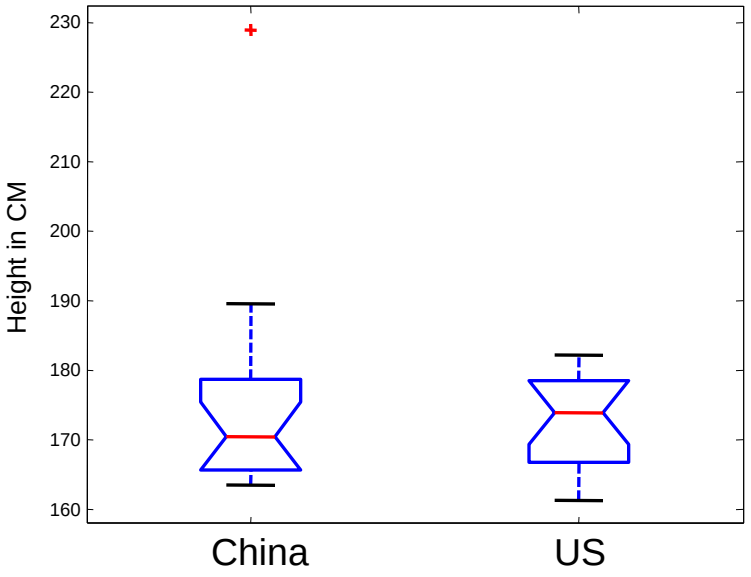
Variance Estimate on Performance Measures

229.00	166.26
170.31	167.08
163.61	166.60
179.06	161.40
170.52	175.32
164.91	173.31
168.69	180.39
164.99	182.37
184.31	177.39
189.76	167.75
170.95	179.81
168.47	174.83
164.25	171.04
178.09	177.40
178.53	166.41
166.31	180.62
Mean	
175.74	173.00
STD	
16.15	6.45

Suppose I want to know if American males are taller than Chinese males. I randomly sample 16 of each, although it happens that I get Yao Ming in the sample…

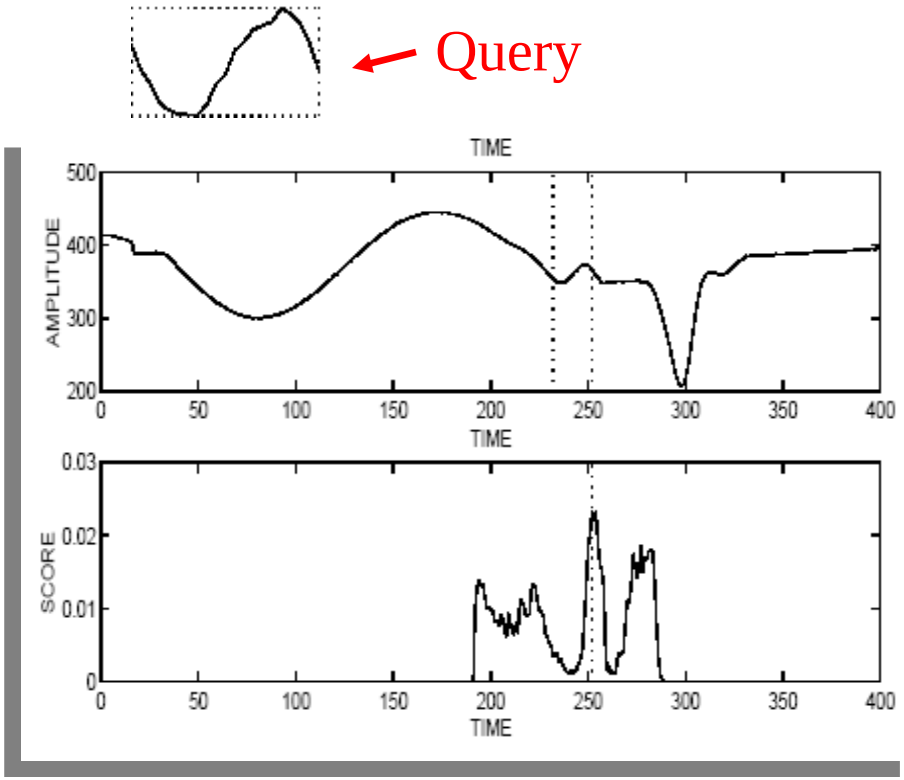
Plotting just the *mean* heights is very deceptive here.



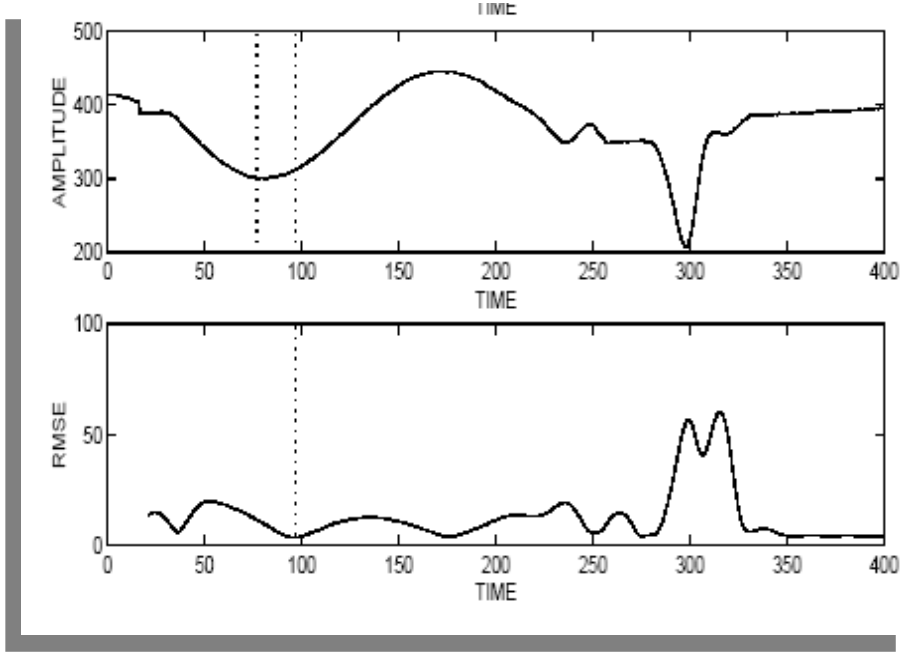


Be fair to the Strawmen/Rivals

In a KDD paper, this figure is the main proof of utility for a new idea. A query is suppose to match to location 250 in the target sequence. Their approach *does*, Euclidean distance *does not*….



SHMM (larger is better match)

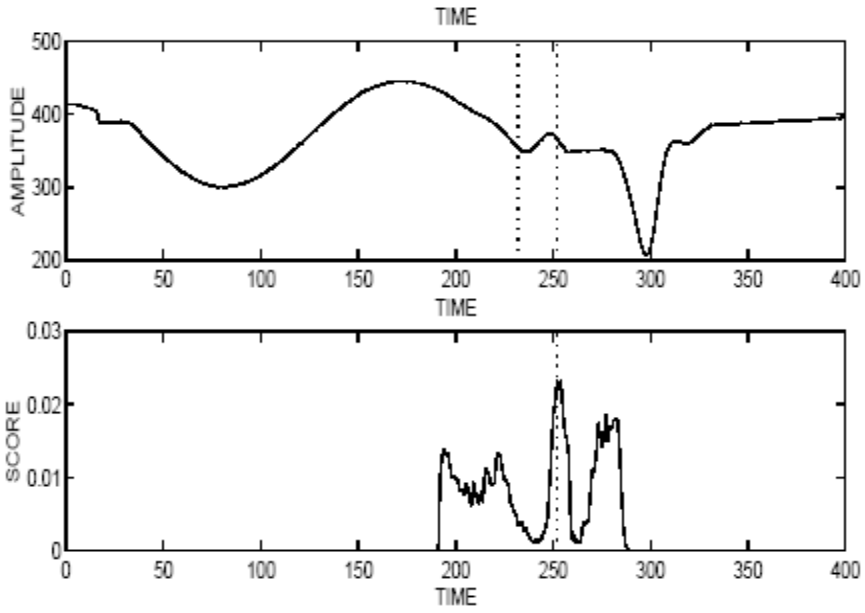


Euclidean Distance (Smaller is better match)

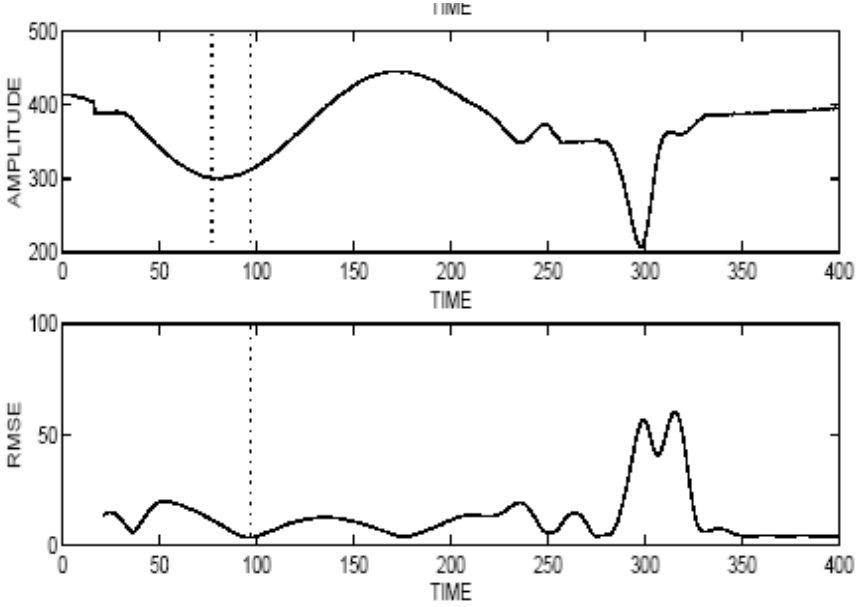
The authors would NOT share this data, citing confidentiality (even though the entire dataset is plotted in the paper) So we cannot reproduce their experiments… or can we?

I wrote a program to extract the data from the PDF file…

Query



SHMM (larger is better match)



Euclidean Distance (Smaller is better match)

If we simply normalize the data (as dozens of papers explicitly point out) the best match for Euclidean distance is at… location 250!

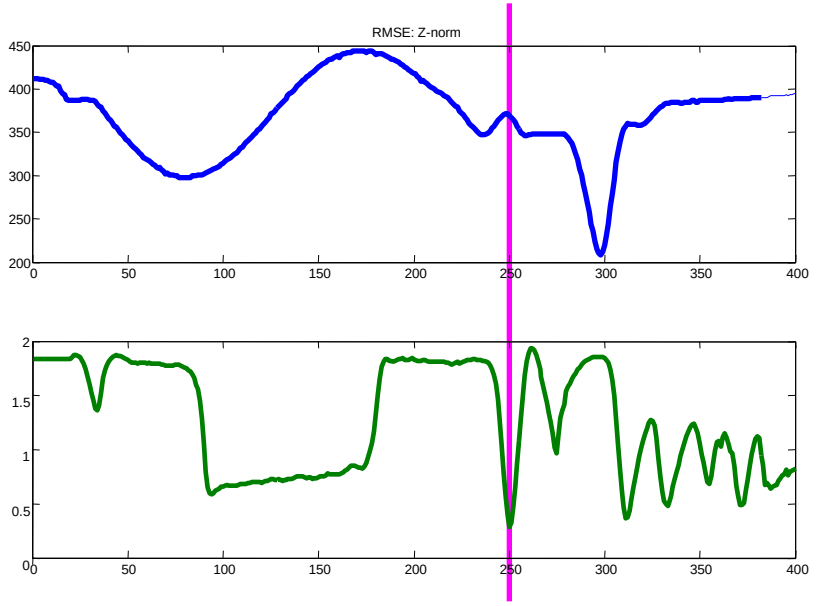
So this paper introduces a method which is:

1) Very hard to implement
2) Computationally demanding
3) Requires lots of parameters

To do the same job as 2 lines of parameter free code.

Be fair to the Strawmen/Rivals

Because the experiments are ***not reproducible***, no one has noticed this. Several authors wrote follow-up papers, simply assuming the utility of this work.



(Normalized) Euclidean Distance

Plagiarism can be obvious..

2006 paper

Suppose we have two time series *X*1 and *X*2, of length *t*1 and *t*2 respectively, where:
To align two sequences using DTW we construct an *t*1-by-*t*2 matrix where the (*i*-th, *j*th)
element of the matrix contains the distance d(x1i,x2j) between the two points x1i and x2j (With Euclidean distance, d(x1i, x2j) =(x1i−x2j)2). Each matrix element (*i,j*) corresponds to the alignment between the points x1i and x2j
A warping path W, is a contiguous (in the sense stated below) set of matrix elements that de fines a mapping between *X*1 and *X*2. The *k*-th element of W is defined as *wk* = (*i, j*)*k*, *W* = *w*1,*w*2, ,*wk*, ,*wK*
max(*t*1, *t*2) ≤ *K* < *m* + *n* − 1
The warping path is typically subject to several constraints.
• Boundary Conditions: w1 = (1,1) and wK = (t1, t2), simply stated, this requires the warping path to start and finish in diagonally opposite corner cells of the matrix.
• Continuity: Given *wk* = (a,b) then *wk*¡1 = (*a0, b0*) where *a*−*a0* ≤ 1 and *b*−*b0* ≤ 1. This restricts the allowable steps in the warping path to adjacent cells(including diagonally adjacent cells).
• Monotonicity: Given *wk* = (a,b) then *wk*¡1 = (*a0, b0*) where *a* − *a0* ≥ 1 and *b* − *b0* ≥ 0. This forces the points in W to be monotonically spaced in time.
There are exponentially many warping paths that satisfy the above conditions, however we are interested only in the path which minimizes the warping cost,
The K in the denominator is used to compensate for the fact that warping paths may have different

1999 paper

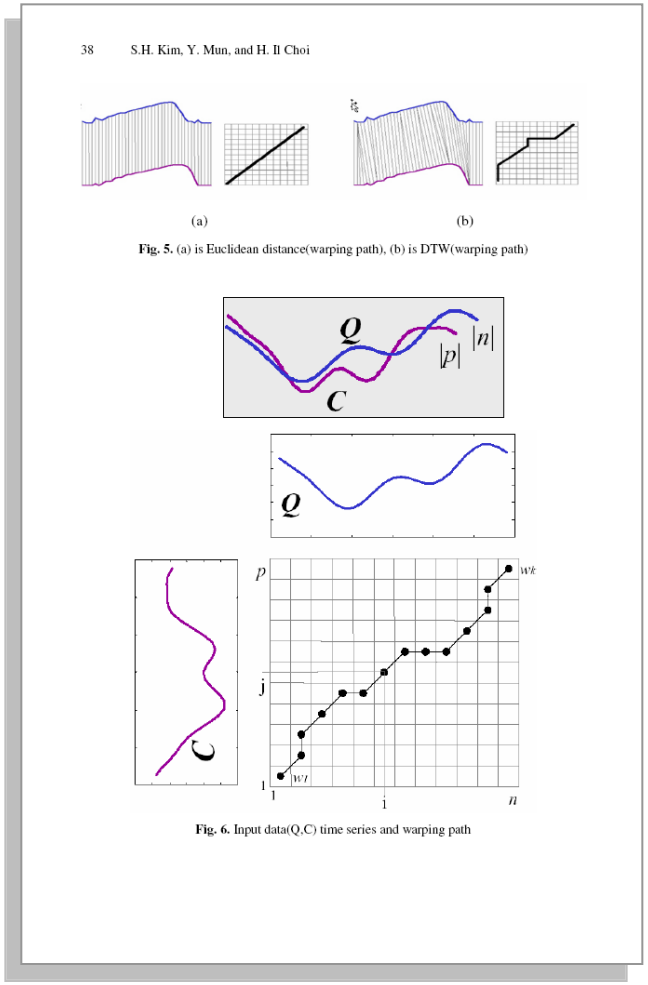
Suppose we have two time series *Q* and *C*, of length *n* and *m* respectively, where:
To align two sequences using DTW we construct an n-by-m matrix where the (ith, jth) element of the matrix contains the distance d(qi,cj) between the two points qi and cj (With Euclidean distance, d(qi,cj) = (qi - cj)2). Each matrix element (i,j) corresponds to the alignment between the points qi and cj. A warping path *W*, is a contiguous (in the sense stated below) set of matrix elements that definesa mapping between *Q* and *C*. The *k*th element of *W* is defined as *wk* = (i,j)k so we have:
W = *w*1, *w*2, …,*wk*,…,*wK* max(*m*,*n*) □□K < *m*+*n*-1
The warping path is typically subject to several constraints.
□□Boundary Conditions: *w*1 = (1,1) and *w*K = (*m*,*n*), simply stated, this requires the warping path to start and finish in diagonally opposite corner cells of the matrix.
□□Continuity: Given *wk* = (*a*,*b*) then *wk*-1 = (*a'*,*b'*) where *a*–*a'* □1 and *b*-*b'* □□1.
This restricts the allowable steps in the warping path to adjacent cells (including diagonally adjacent cells).
□ Monotonicity: Given wk = (a,b) then wk-1 = (a',b') where a–a' ³ 0 and b-b' ³ 0.This forces the points in W to be monotonically spaced in time.
There are exponentially many warping paths that satisfy the above conditions, however we are interested only in the path which minimizes the warping cost:
The *K* in the denominator is used to compensate for the fact that warping paths may have different

..or it can be subtle. I think the below is an example of plagiarism, but the 2005 authors do not.

2005 paper: **As with most** data mining **problems**, data **representation** is one of the major elements to reach an **efficient and effective solution.** … **pioneered by Pavlidis** et al… **refers to the** idea of representing **a time series of length n** using **K straight lines**

2001 paper: As with most computer science **problems**, **representation** of the data is the key to **efficient and effective solutions.**… **pioneered by Pavlidis… refers to the** approximation of **a time series** T**, of length n**, with **K straight lines**

Figures also get Plagiarized



One page of a ten page paper. All the figures are taken without acknowledgement from Keogh's tutorial

This particular figure gets stolen a lot.

Here by two medical doctors

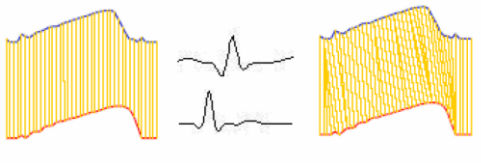


Figure 3. One to one alignment on time axis vs. non-linear alignment (warped time axis). Nonlinear curve alignment is important in pattern recognition of ECG signals. Wavelet analysis dose not allow this type of flexibility in pattern recognition and matching.

Here in a Chinese publication (the author did flip the figure upside down!)

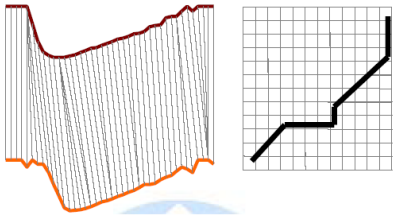


圖 2-4 時間軸扭曲，非線性比對資料

Here in a Portuguese publication..

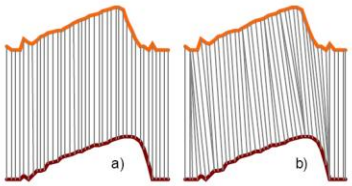


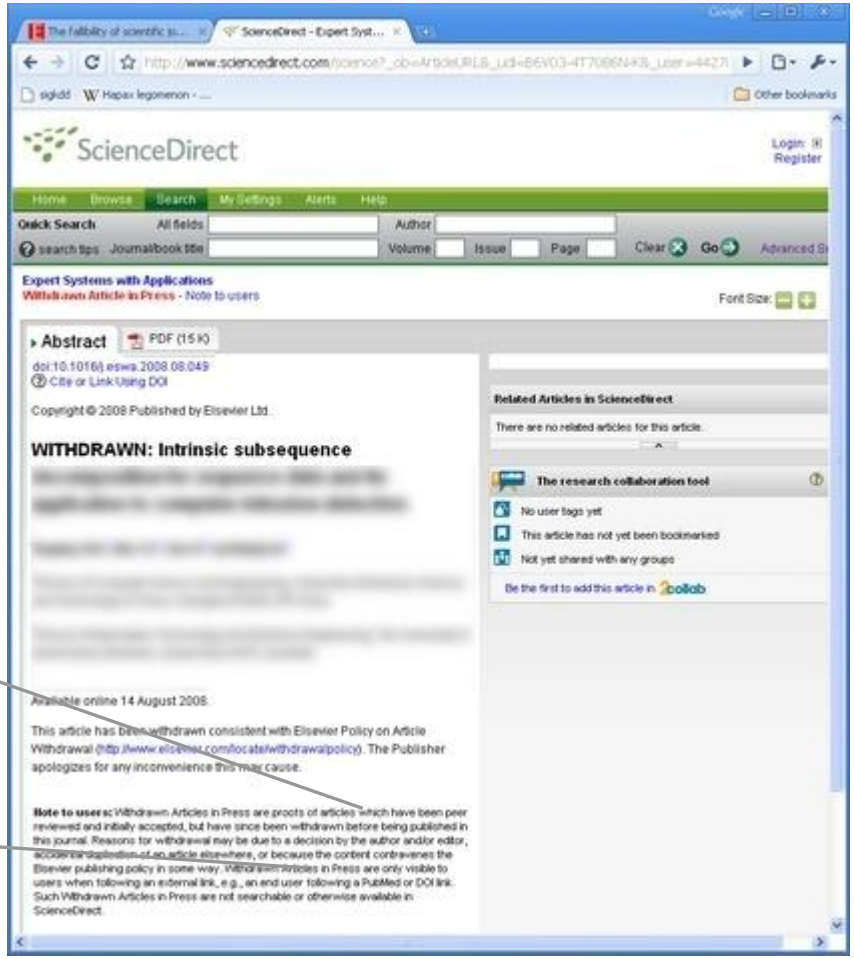
Figura 2. Comparação entre séries: a) convencional; b) com DTW

What Happens if you Plagiarize?

The *best* thing that can happen is the paper gets rejected by a reviewer that spots the problem.

If the paper gets published, there is an excellent chance that the original author will find out, at that point, they *own* you.

Note to users: Withdrawn Articles in Press are proofs of articles which have been peer reviewed and initially accepted, but have since been withdrawn..



Making Good Figures

- I personally feel that making good figures is very important to a papers chance of acceptance.
- The first thing reviewers often do with a paper is scan through it, so images act as an *anchor*.
- In some cases a picture really is worth a thousand words.

See papers of Michail Vlachos, it is clear that he agonizes over every detail in his beautiful figures.
See the books of Edward Tufte.
See Stephen Few's books/blog (www.perceptualedge.com)

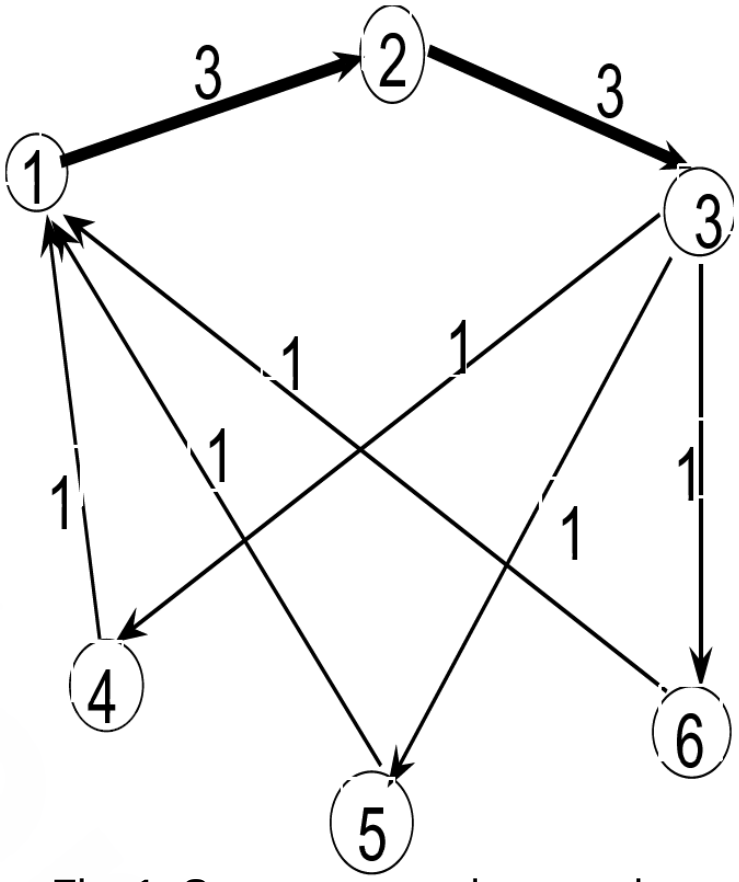


Fig. 1. Sequence graph example

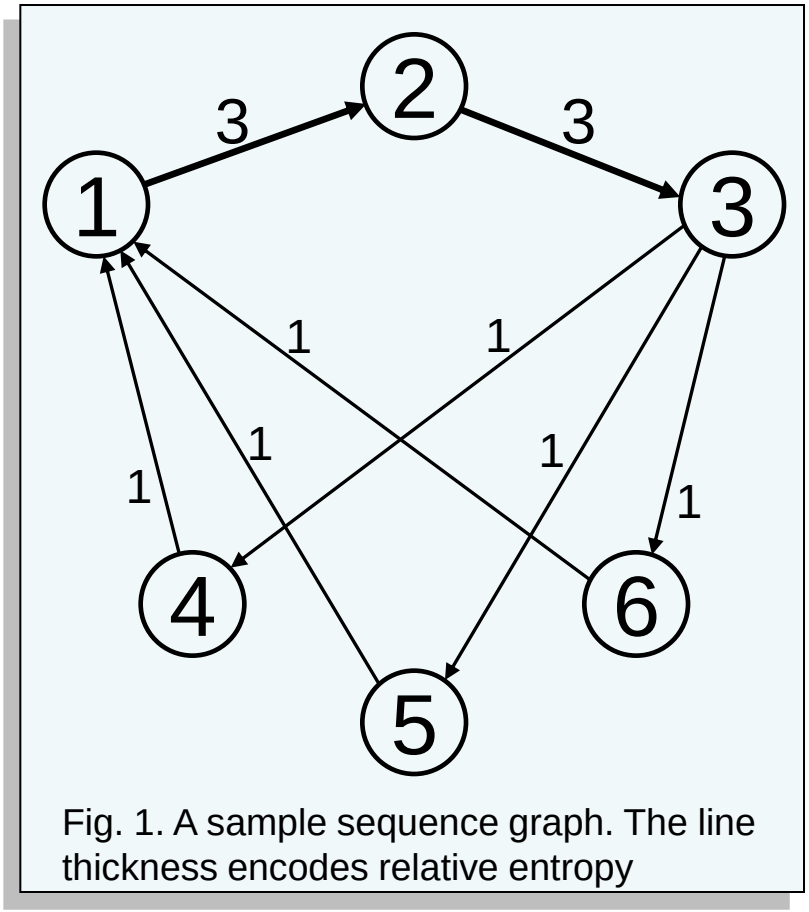


Fig. 1. A sample sequence graph. The line thickness encodes relative entropy

What's wrong with this figure? Let me count the ways…
None of the arrows line up with the "circles". The "circles" are all different sizes and aspect ratios, the (normally invisible) white bounding box around the numbers breaks the arrows in many places. The figure captions has almost no information. Circles are not aligned…

On the right is my redrawing of the figure with PowerPoint. It took me 300 seconds

This figure is an insult to reviewers. It says, "*we expect you to spend an unpaid hour to review our paper, but we don't think it worthwhile to spend 5 minutes to make clear figures*"

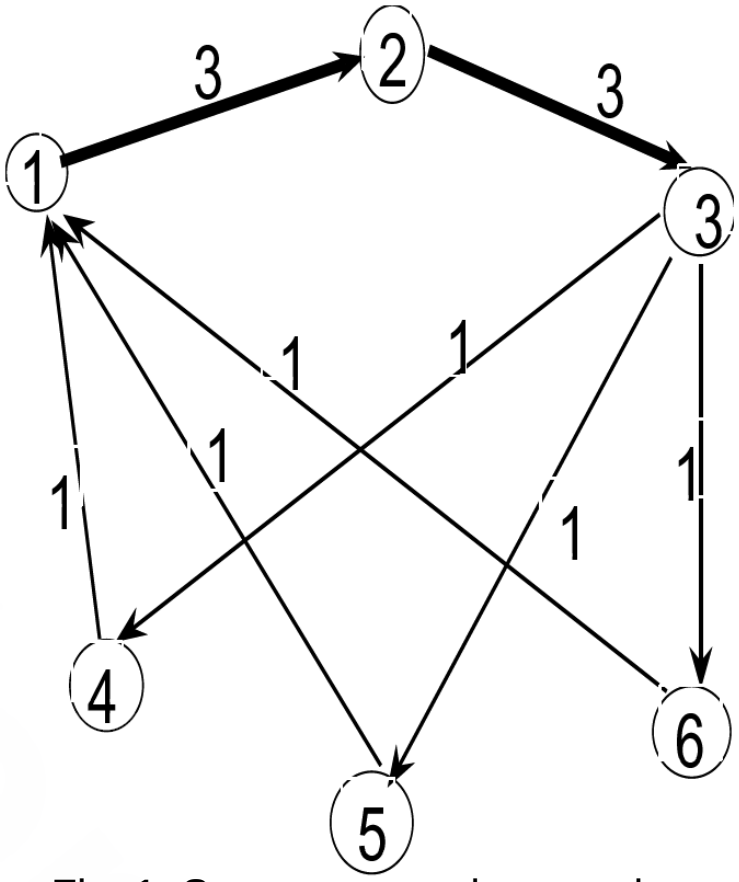
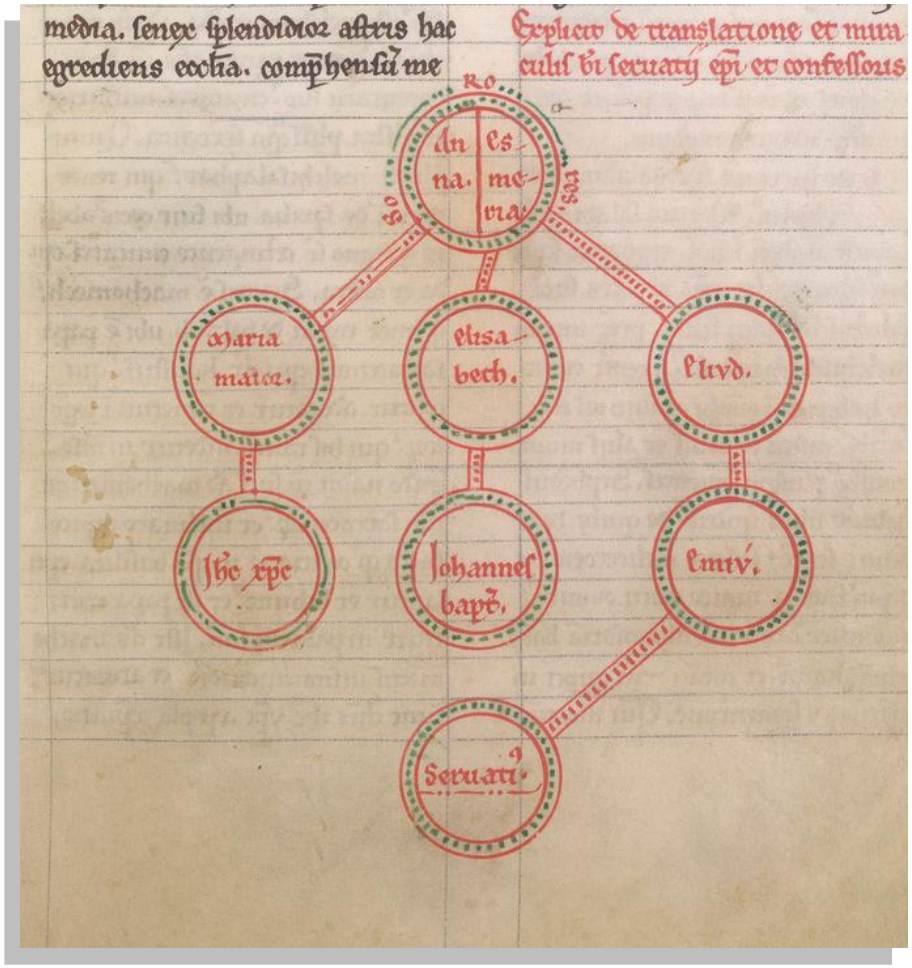


Fig. 1. Sequence graph example



Note that there are figures drawn seven hundred years ago that have much better symmetry and layout.

Peter Damian, Paulus Diaconus, and others, Various saints lives: Netherlands, S. or France, N. W.; 2nd quarter of the 13th century

Lets us see some more examples of poor figures, then see some principles that can help

This figure wastes 80% of the space it takes up.

In any case, it could be replace by a short English sentence: "*We found that for selectivity ranging from 0 to 0.05, the four methods did not differ by more than 5%*"

Why did they bother with the legend, since you can't tell the four lines apart anyway?

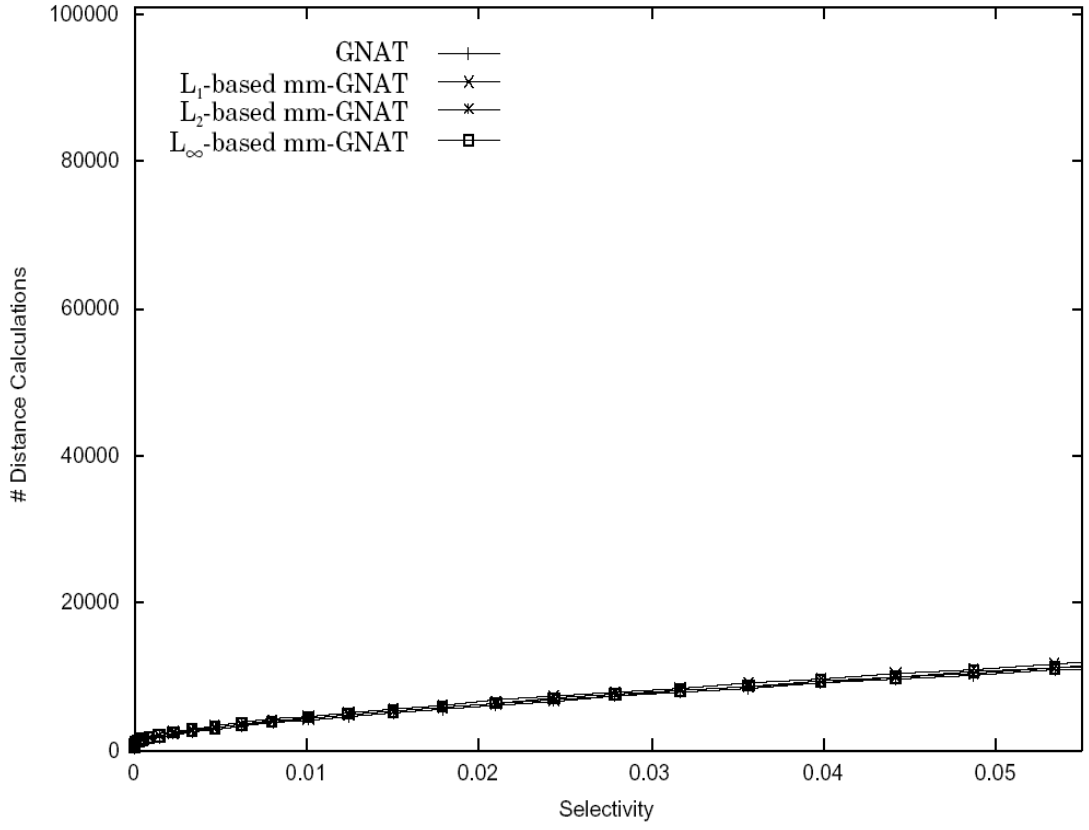
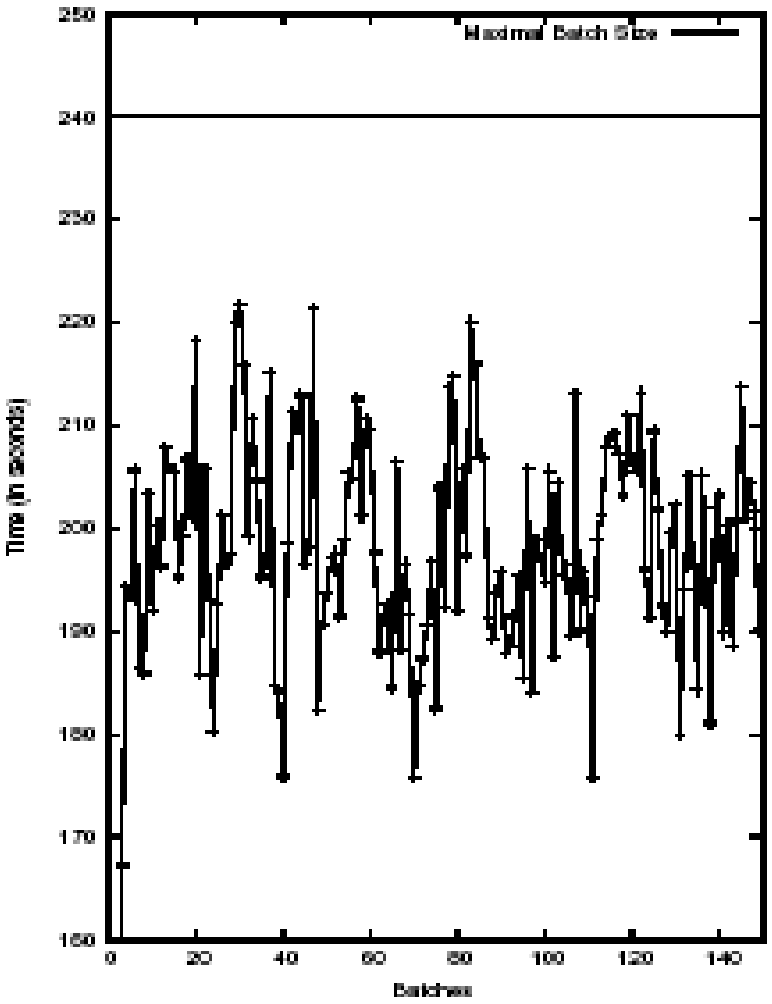


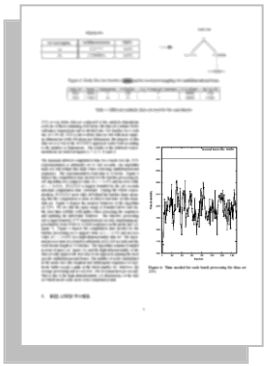
Figure 9. Selectivity versus number of distance calculations (search by L_∞ norm , DB_4)

This figure wastes almost a quarter of a page.

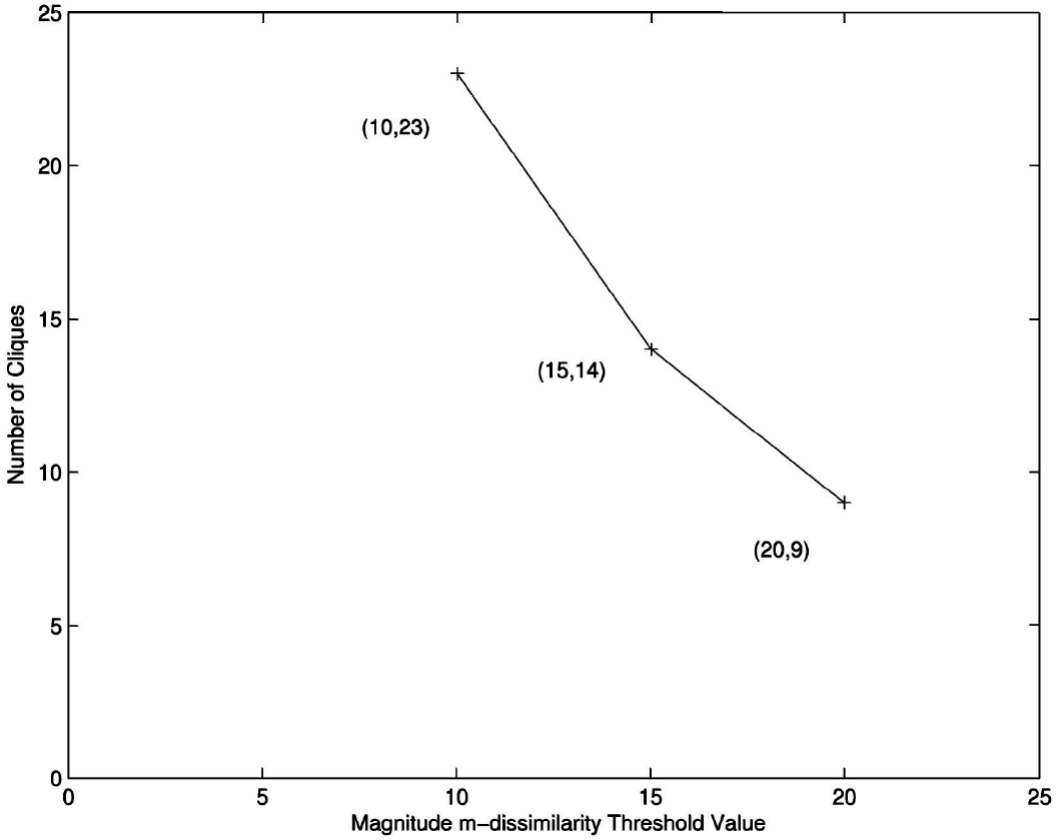
The ordering on the X-axis is arbitrary, so the figure could be replaced with the sentence "*We found the average performance was 198 with a standard deviation of 11.2*".

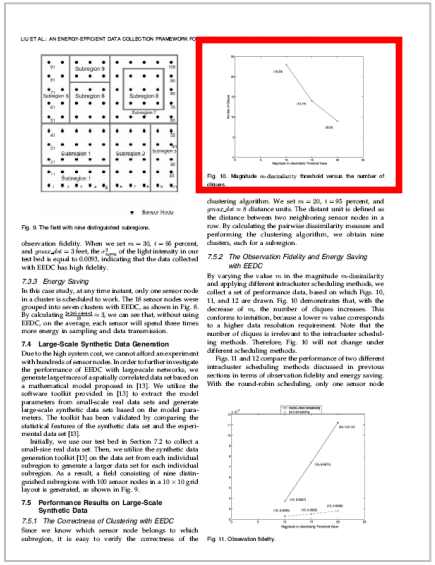
The paper in question had 5 similar plots, wasting an entire page.





The figure below takes up 1/6 of a page, but it only reports 3 numbers.





The figure below takes up 1/6 of a page, but it only reports 2 numbers!

Actually, it really only reports one number! Only the relative times really matter, so they could have written "*We found that FTW is 1007 times faster than the exact calculation, independent of the sequence length*".

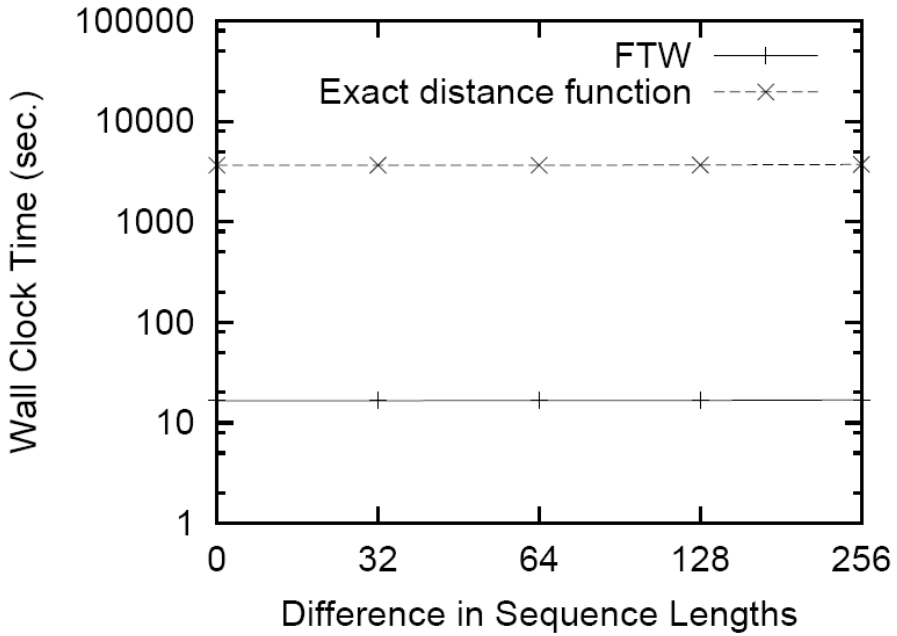
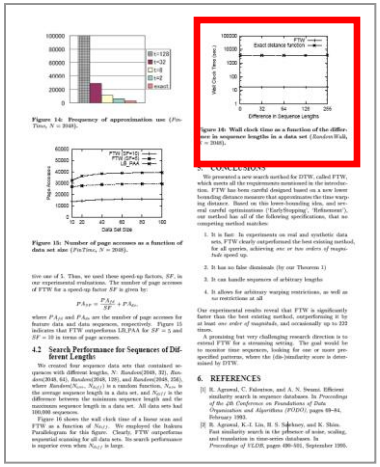


Figure 16: Wall clock time as a function of the difference in sequence lengths in a data set (*RandomWalk*, $N = 2048$).



Both figures below describe the classification of time series motions…

It is not obvious from this figure which algorithm is best. The caption has almost zero information

You need to read the text very carefully to understand the figure

Redesign by Keogh

At a glance we can see that the accuracy is very high. We can also see that DTW tends to win when the...

Similarities between similar motions are computed for the 100 motions, and similarity between each motion and other 99 dissimilar motions are also computed. Fig. 5 shows the similarities of more accurate motions and the highest similarities between each motion and the other 99 different motions. For more accurate motions, similarities between similar motions are higher than those between the same motion and all other dissimilar motion, achieving 100% recognition rate for the 100 different motions. For the less accurate motions, 92% similar motions have higher similarities than dissimilar motions.

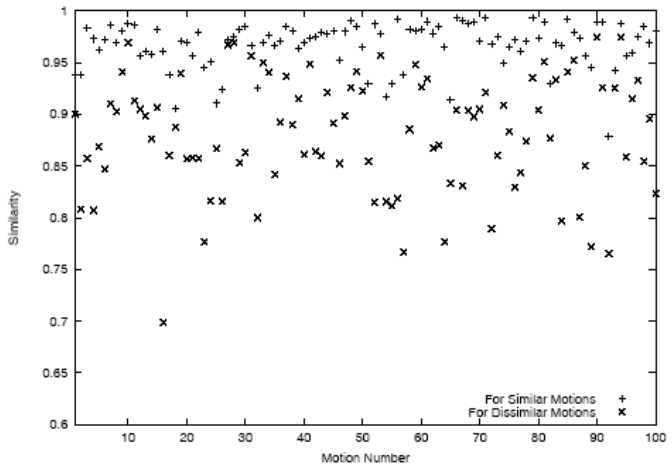


Fig. 5. Motion Similarities

The data is plotted in Figure 5. Note that any *correctly* classified motions must appear in the upper left (gray) triangle.

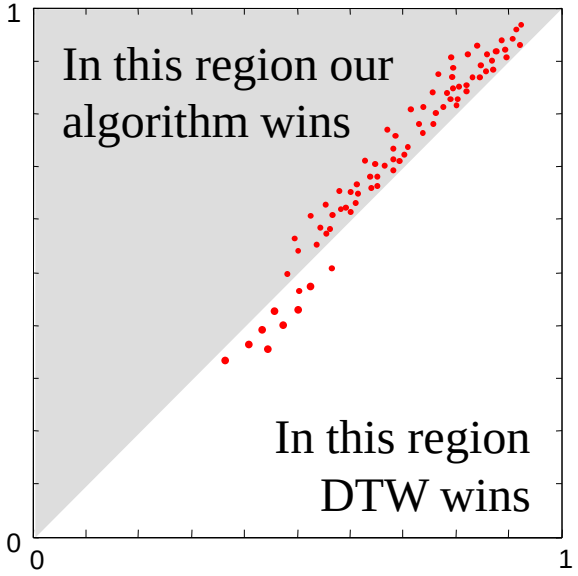


Figure 5. Each of our 100 motions plotted as a point in 2 dimensions. The X value is set to the distance to the nearest neighbor from the *same* class, and the Y value is set to the distance to the nearest neighbor from any *other* class.

Both figures below describe the performance of 4 algorithms on indexing of time series of different lengths…

This figure takes 1/2 of a page.

This figure takes 1/6 of a page.

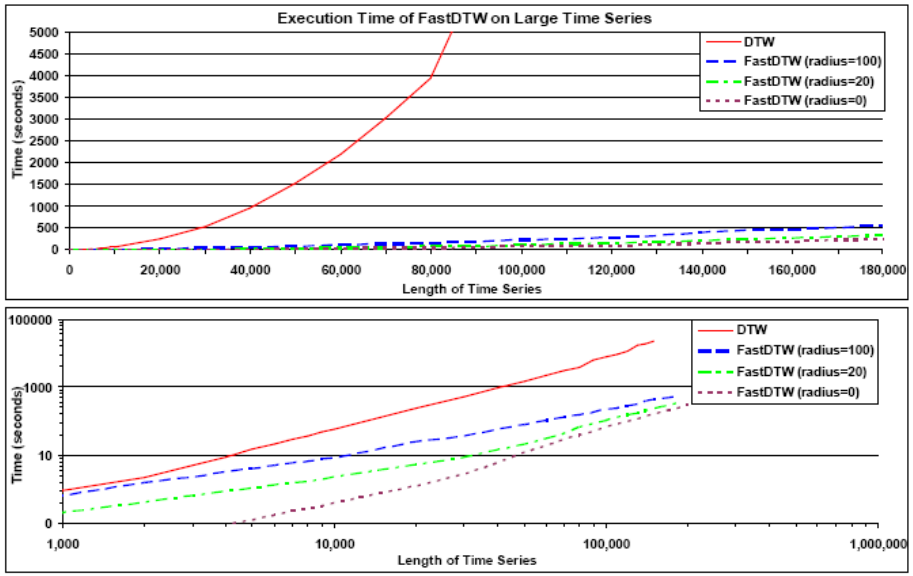


Figure 11. The eficiency of FastDTW and DTW on large time series. The top figure is scaled normally, and the bottom figure has log-log scaling.

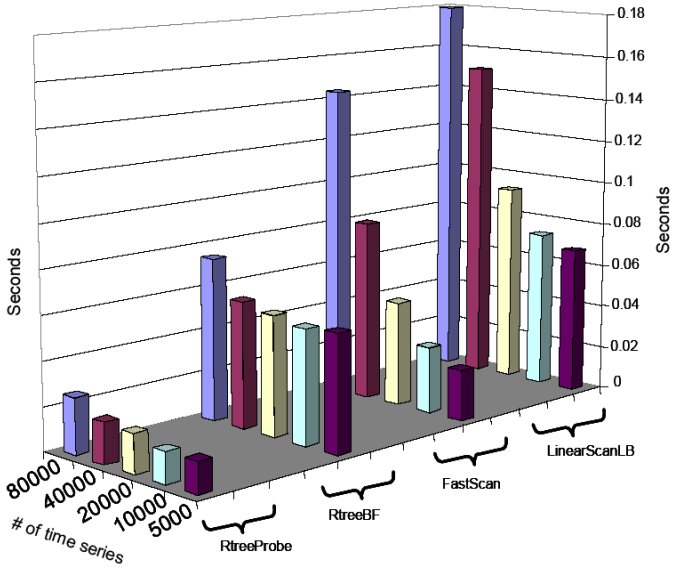
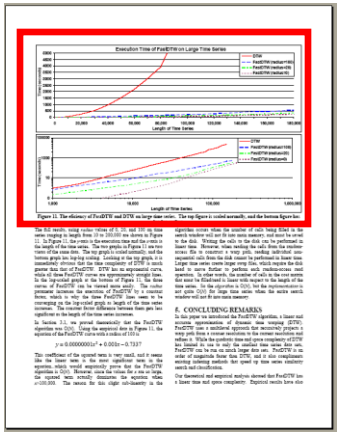
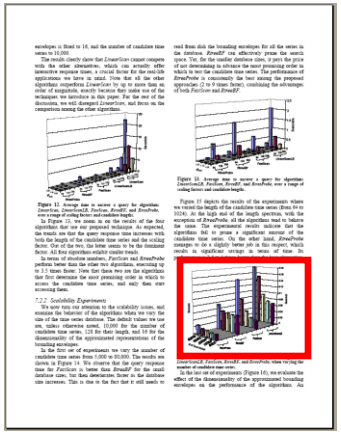


Figure 14. Average time to answer a query for algorithms *LinearScanLB*, *FastScan*, *RtreeBF*, and *RtreeProbe*, when varying the number of candidate time series.





This should be a bar chart, the four items are unrelated

(in any case this should probably be a table, not a figure)

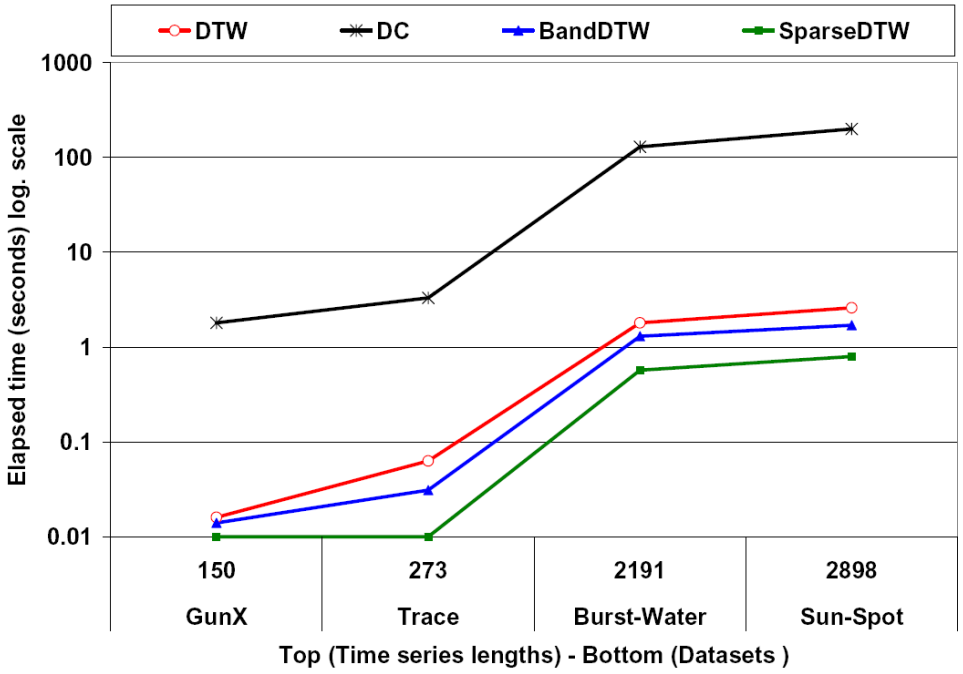


Figure 5: Elapsed time using real life datasets.

This pie chart takes up a lot of space to communicate 3 numbers (Better as a table, or as simple text)

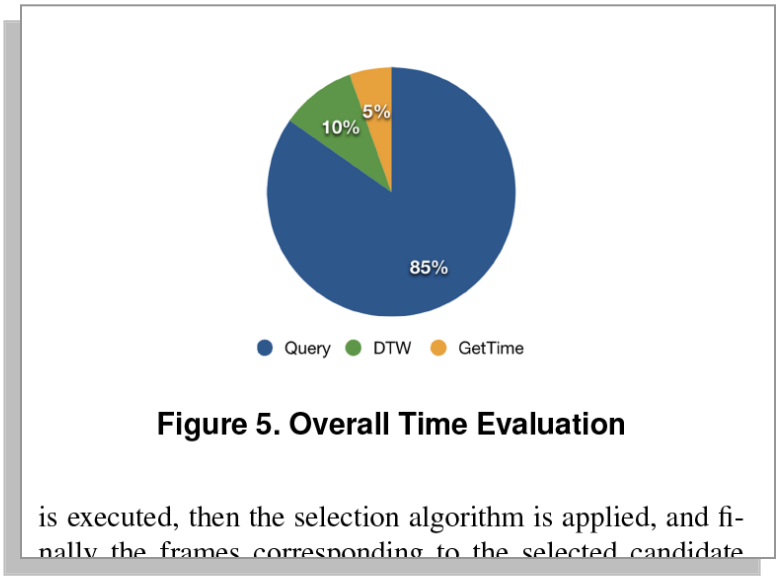
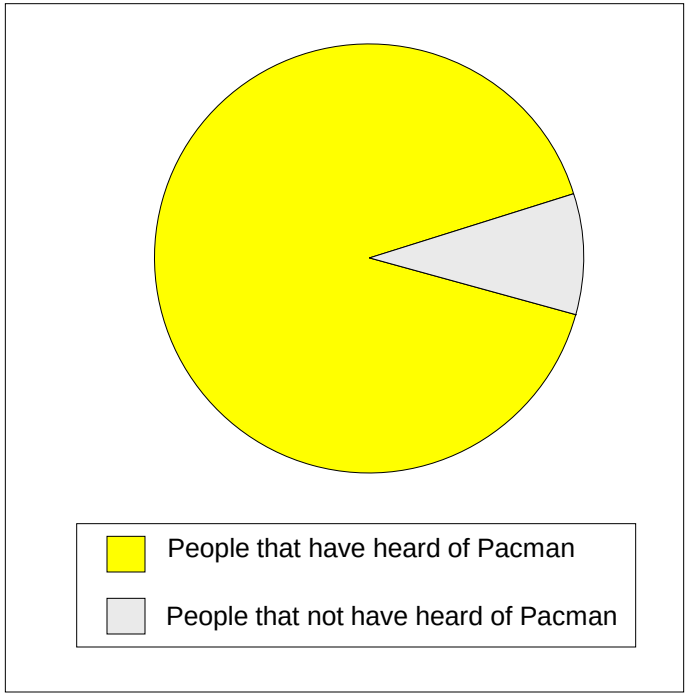


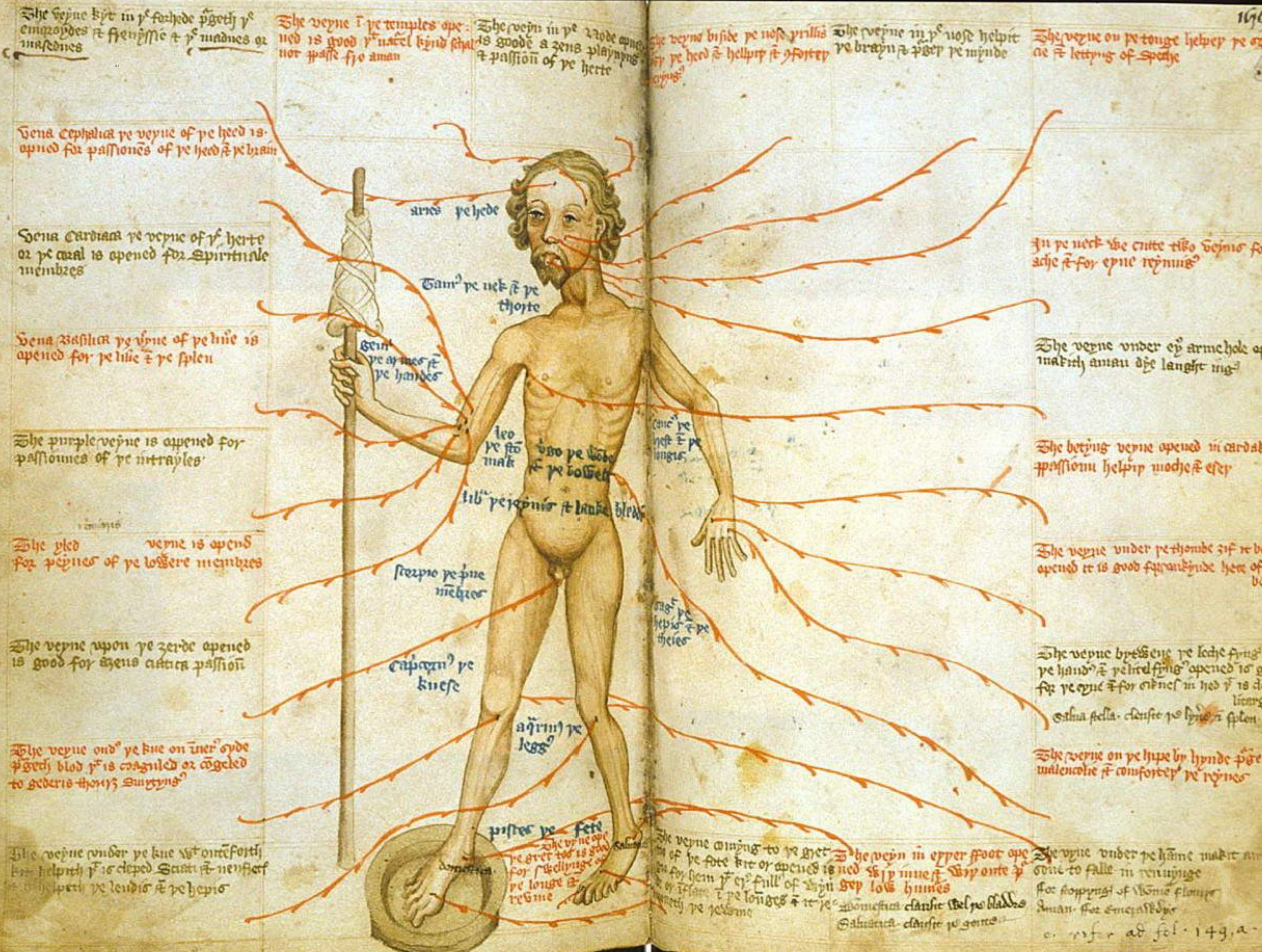
Figure 5. Overall Time Evaluation

is executed, then the selection algorithm is applied, and fi-
nally the frames corresponding to the selected candidate



Principles to make Good Figures

- Think about the point you want to make, should it be done with words, a table, or a figure. If a figure, what kind?
- Color helps (but you cannot depend on it)
- Linking helps (sometimes called brushing)
- Direct labeling helps
- Meaningful captions helps
- Minimalism helps (Omit needless elements)
- Finally, taking great care, taking pride in your work, helps


The veyne kyt in ye forhede purgeth ye emeroydes & frenysse & ye madnes or masednes
The veyne in ye temples opened is good yt natel kynd schal not passe fro aman
The veyn in ye nose opened is goode azens playnyng & passion of ye herte
The veyne biside ye nose purgith ye heed & helpith & avoidith
The veyne in ye nose helpit ye brayn & kepith ye mynde
The veyne on ye tonge helpith ... & letting of speche
Vena cephalica ye veyne of ye heed is opened for passiones of ye heed & ye brain
Vena cardiaca ye veyne of ye herte or ye coral is opened for spirituale membres
Vena basilica ye veyne of ye lure is opened for ye liue & ye splen
The purple veyne is opened for passiounes of ye intrayles
The yled veyne is opend for peynes of ye lowere membres
The veyne upon ye zerde opened is good for azens ciatica passion
The veyne und' ye kne on iner syde purgeth blod yt is coagulled or congeled
aries ye hede
Taur' ye nek & ye throte
Gemi' ye armes & ye handes
Canc' ye brest & ye longes
leo ye stomak
virgo ye wombe & ye bowelles
lib' ye reynes & lanke bledd
Scorpio ye priue membres
Sagit' ye hepis & ye theies
Capricorn' ye knese
Aquari' ye legges
pisces ye fete
In ye neck we cutte two veynes for ache & for eyne reynings
The veyne under ye armehole ... makith aman ... laught ...
The betyng veyne opened in cardiak passioun helpith moche & esy
The veyne under ye thombe ... opened it is good ...
The veyne bytwene ye leche fyngr & ye hand & ye litel fyngr opened is good for ye eyne & for ... in hed ...
Salua tella ... ye lyuer & splen
The veyne on ye hipe by hynde ... malencolie & confortep ye reynes
The veyne cominge to ye grete too of ye fote ... opened ... full of ... ye longes & ...
The veyn in epper foot opened ...
The veyne under ye hamme makit ... to falle in ...
For stopping of ... aman for ...
The veyne under ye kne ... helpith ... ye lendis & ye hepis
c. refer ad fol. 149 a

Direct labeling helps

It removes one level of indirection, and allows the figures to be self explaining

(see Edward Tufte: Visual Explanations, Chapter 4)

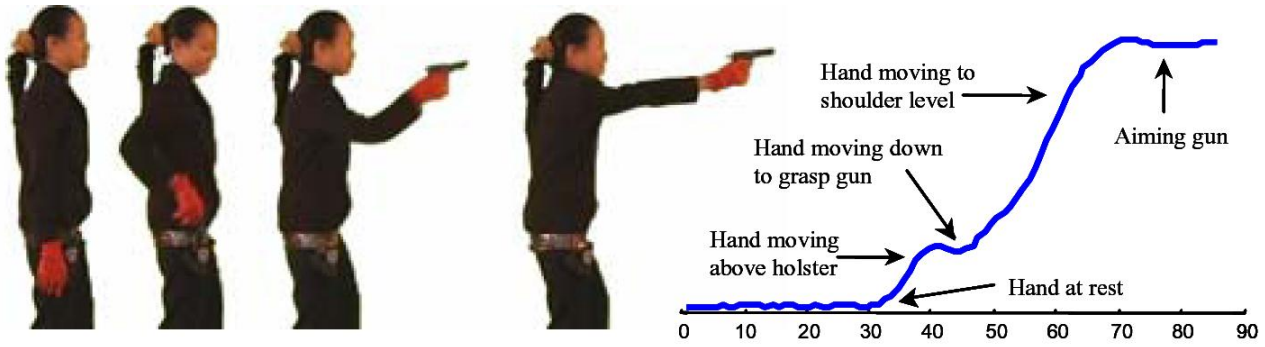


Figure 10. Stills from a video sequence; the right hand is tracked, and converted into a time series

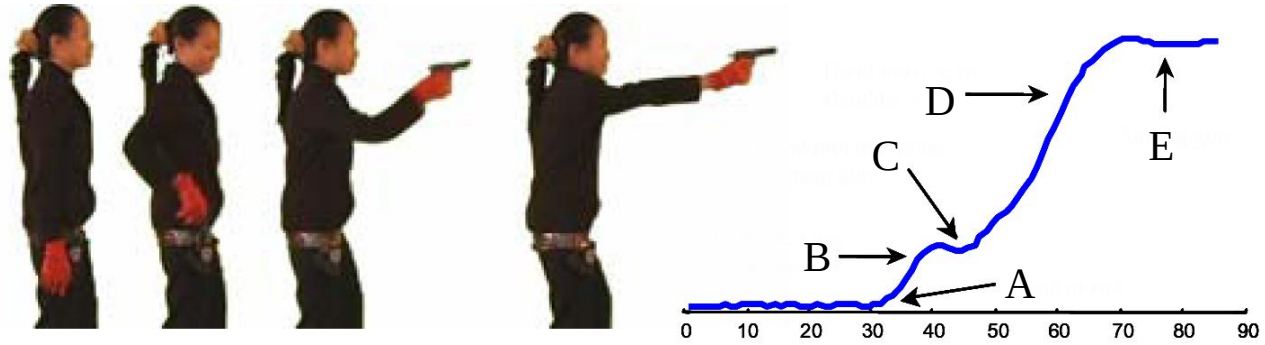


Figure 10. Stills from a video sequence; the right hand is tracked, and converted into a time series: A) Hand at rest: B) Hand moving above holster. C) Hand moving down to grasp gun. D Hand moving to shoulder level, E) Aiming Gun.

Linking helps interpretability I

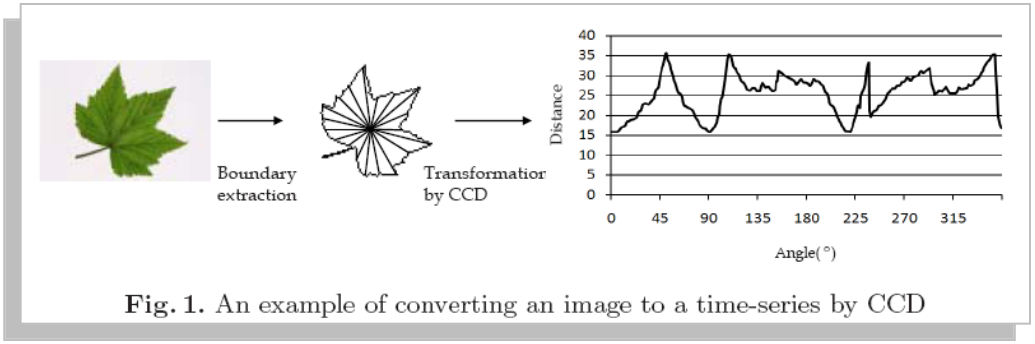
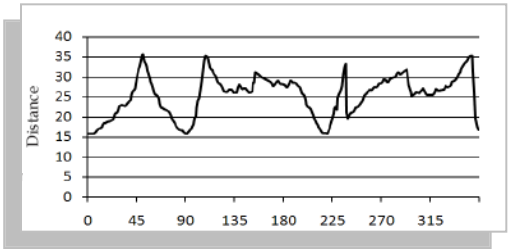


Fig. 1. An example of converting an image to a time-series by CCD

How did we get from here 

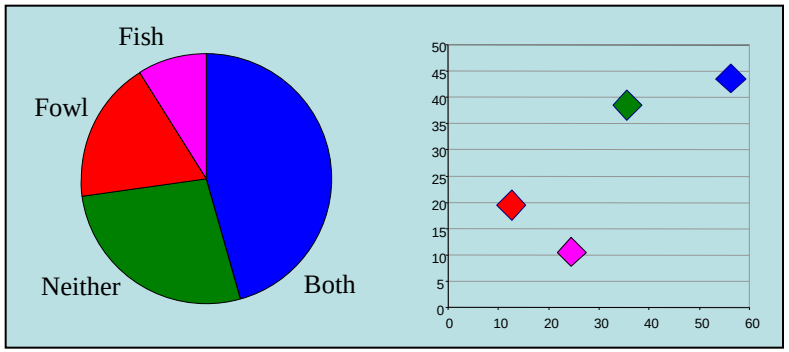
To here? 

It is not clear from the above figure.

See next slide for a suggested fix.

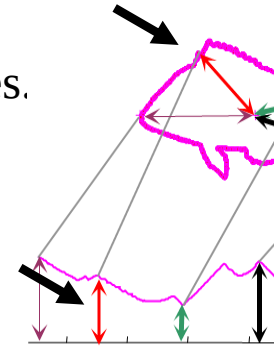
What is Linking?

Linking is connecting the same data in two views by using the same color (or thickness etc). In the figures below, color links the data in the pie chart, with data in the scatterplot.



Linking helps interpretability II

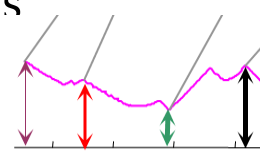
In this figure, the color of the arrows inside the fish **link** to the colors of the arrows on the time series.



This tells us exactly how we go from a shape to a time series.

Note that there are other **links**, for example in II, you can tell which fish is which based on color or link thickness **linking**.

Minimalism helps: In this case, numbers on the X-axis do not mean anything, so they are deleted.



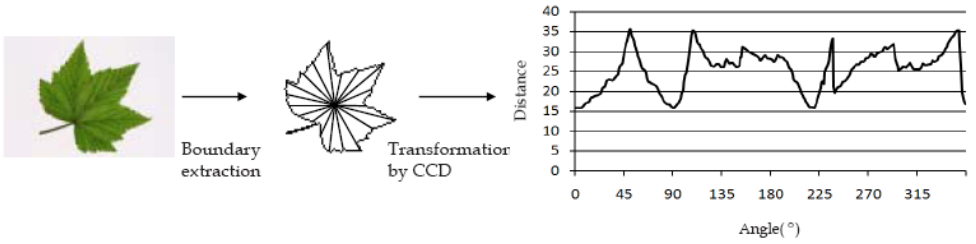
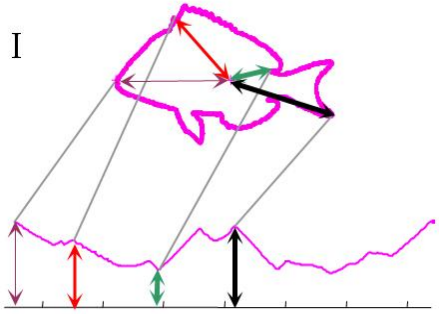
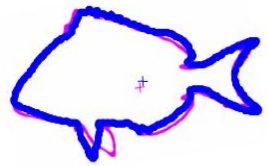


Fig. 1. An example of converting an image to a time-series by CCD





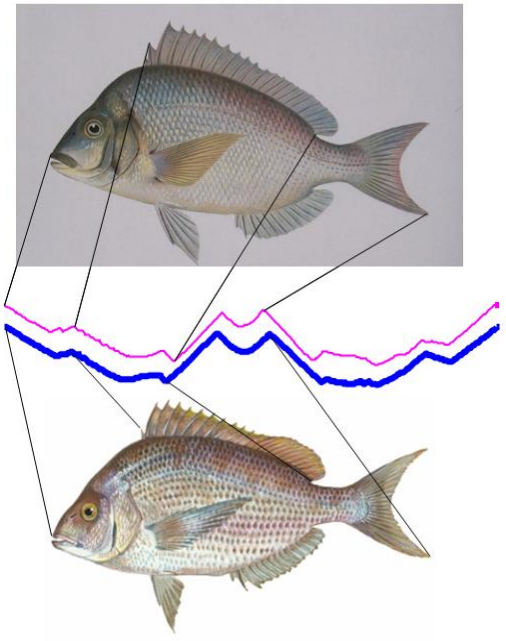
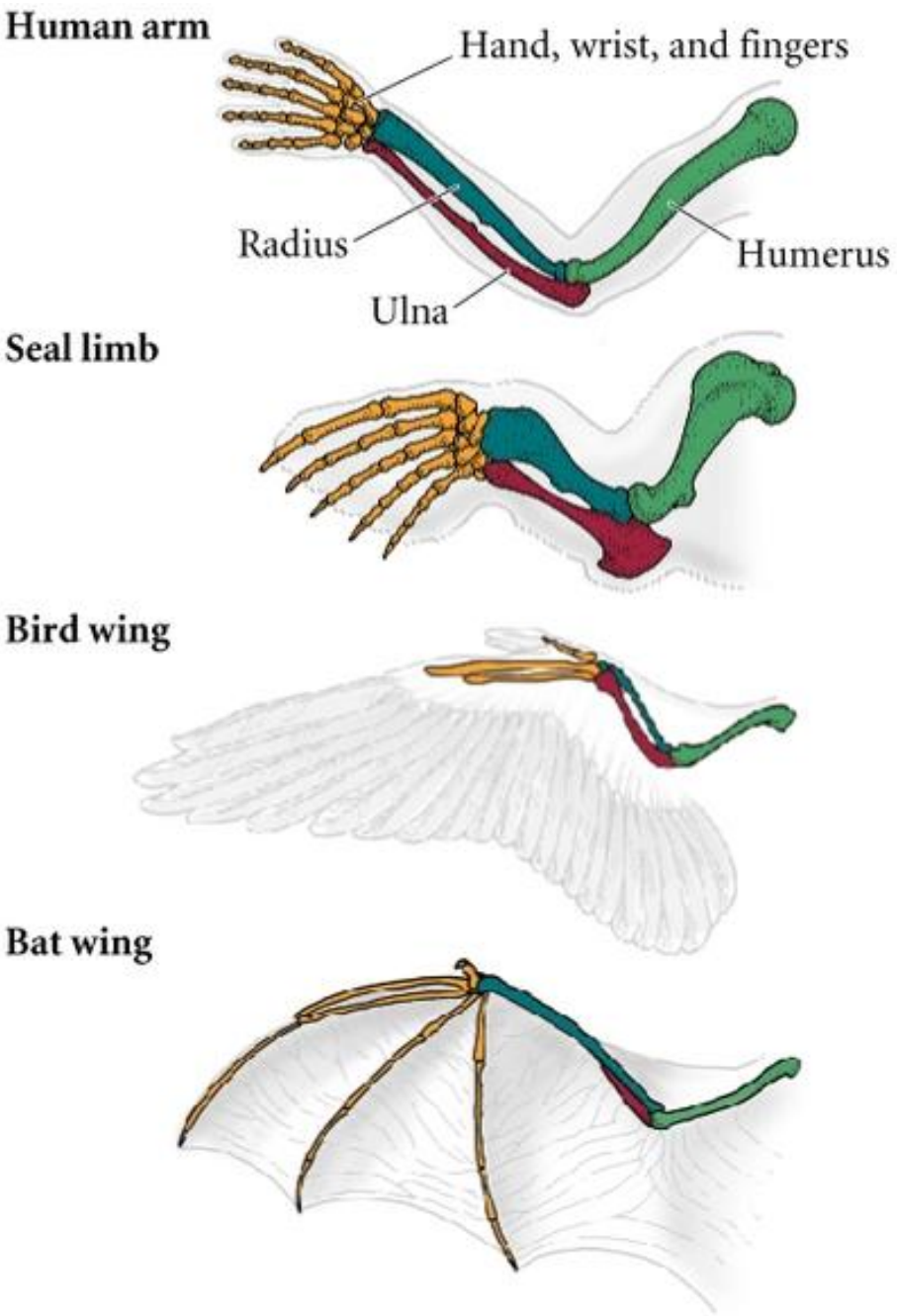
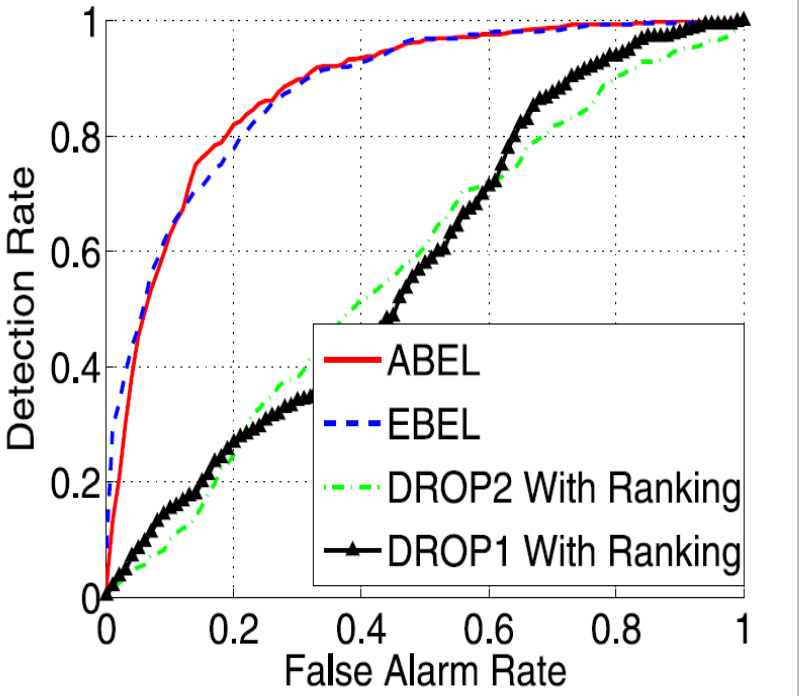
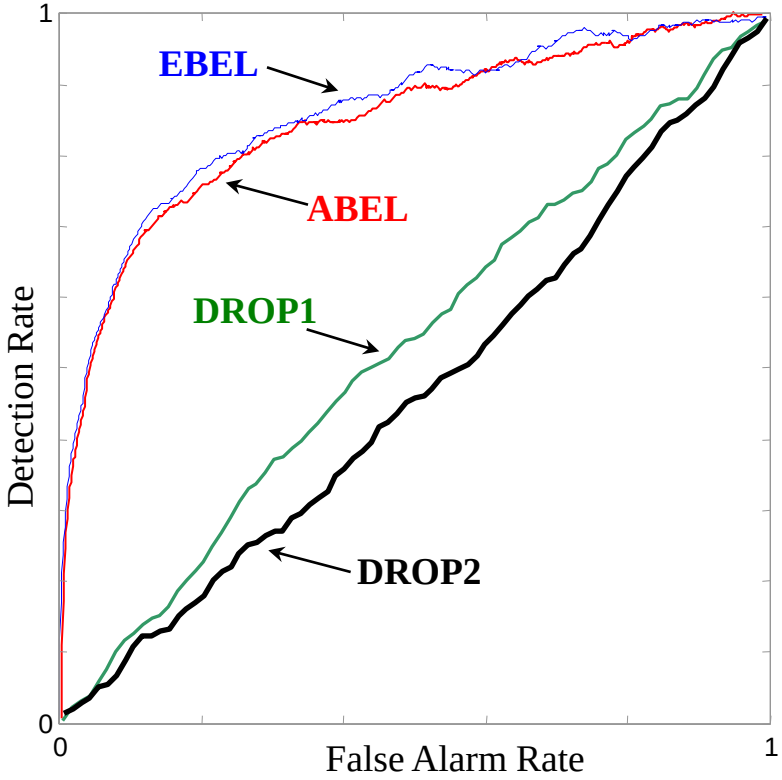


Figure 8: A visual intuition of the conversion of a two-dimensional shape to a one-dimensional "time series". II) Two shapes that are similar in the shape space will also be similar in the time series shape. III) Here we compare an 1890 chromolithograph [5] to a modern photograph of *Stenotomus chrysops* (common name: Scup or Porgy)

A nice example of linking







- Don't cover the data with the labels! You are implicitly saying "*the results are not that important*".
- Do we need all the numbers to annotate the X and Y axis?
- Can we remove the text "With Ranking"?

Direct labeling helps

Note that the line thicknesses differ by powers of 2, so even in a B/W printout you can tell the four lines apart.

Minimalism helps: delete the "with Ranking", the X-axis numbers, the grid…

Covering the data with the labels is a common sin

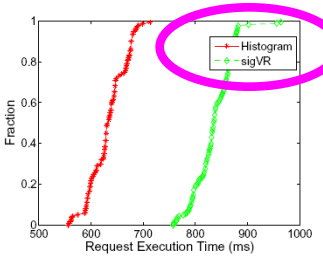
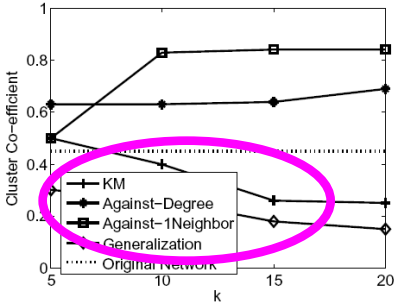
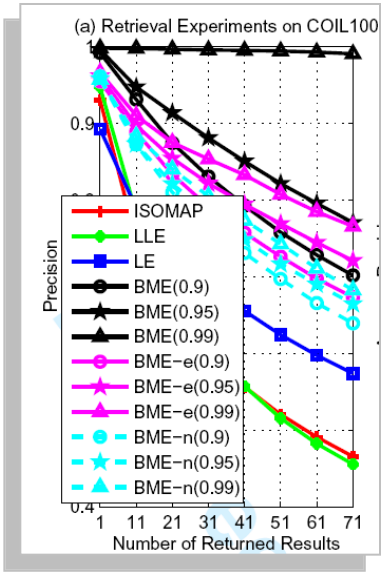
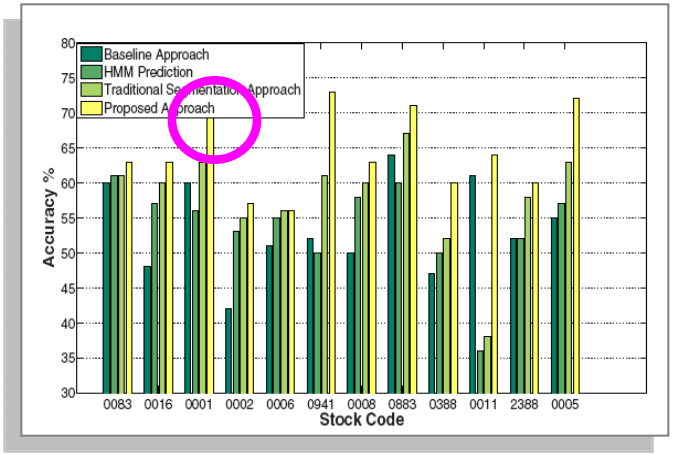


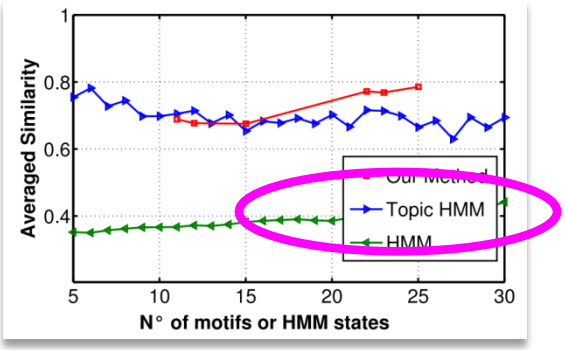
Figure 5.12: Signature matching computation time

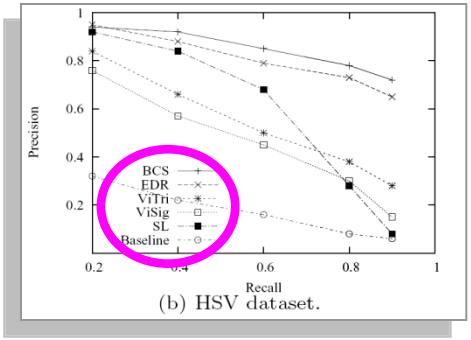


(a) in Prefuse network

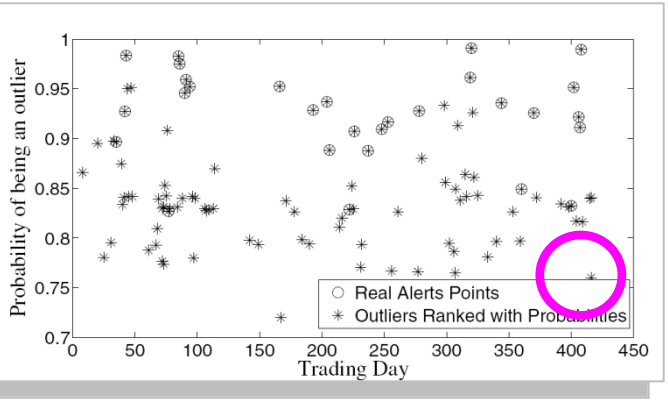


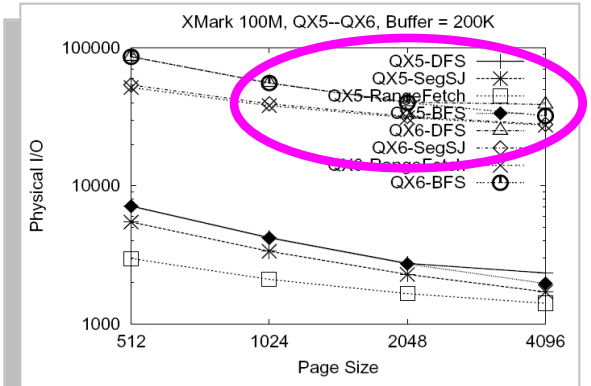




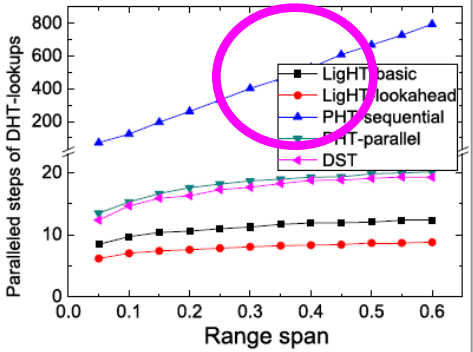


(b) HSV dataset.





(c) PIO vs PageSize: QX5 and QX6



(c) Varying range span

These two images, which are both use to discuss an anomaly detection algorithm, illustrate many of the points discussed in previous slides.

Color helps - Direct labeling helps - Meaningful captions help

The images should be as self contained as possible, to avoid forcing the reader to look back to the text for clarification multiple times.

Note that while Figure 6 use color to highlight the anomaly, it also uses the line thickness (hard to see in PowerPoint) thus this figure works also well in B/W printouts

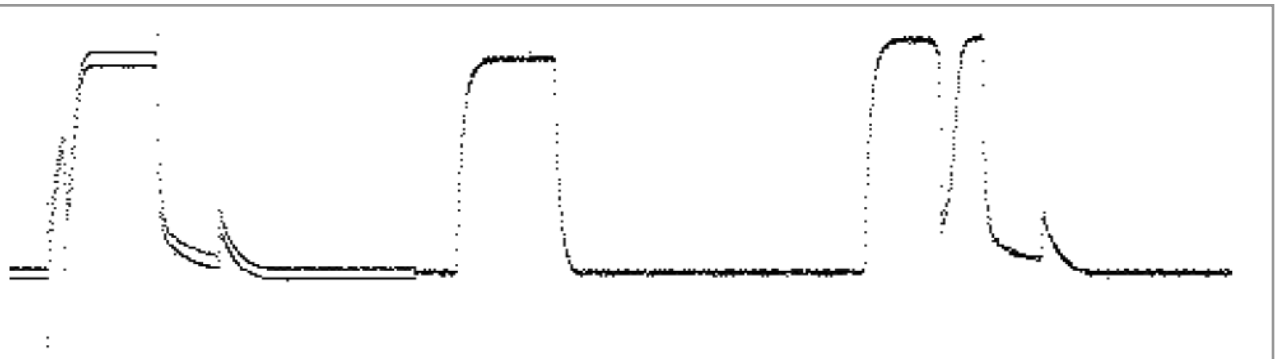


Fig. 9. Concatenation of TEK 0, 10, and 16.

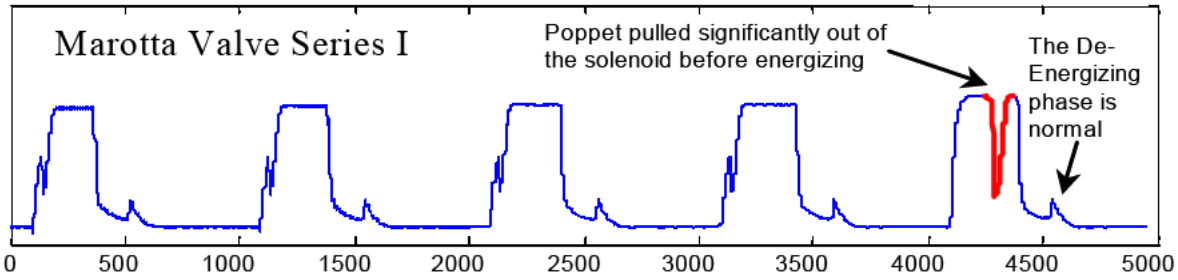
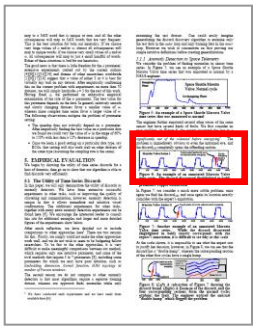
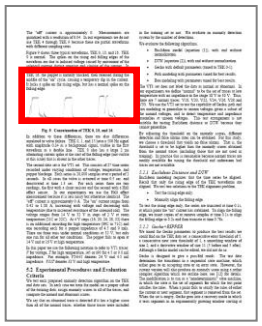


Figure 6: An example of an annotated Marotta Valve time series. The discord discovered (highlighted in bold) exactly corresponds with the expert's annotation of a premature Poppet withdrawal





Thinking about the point you want to make, helps

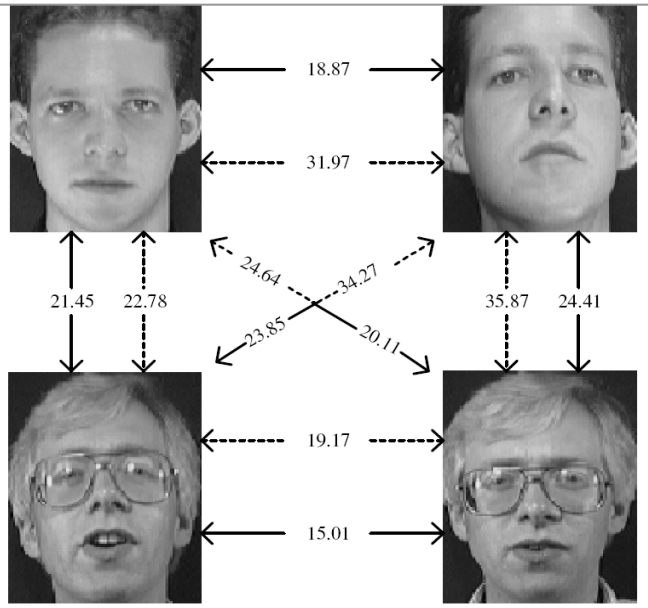


Figure 3. The distances of four faces by 2DDW and Euclidean norm. The 2DDW distances are shown on solid lines and Euclidean distances on dotted lines.

From looking at this figure, we are suppose to tell that 2DDW produces more intuitive results than Euclidean Distance.

I have a lot of experience with these types of things, and high motivation, but it still took me 4 or 5 minutes to see this.

Do you think the reviewer will spend that amount of time on a single figure?

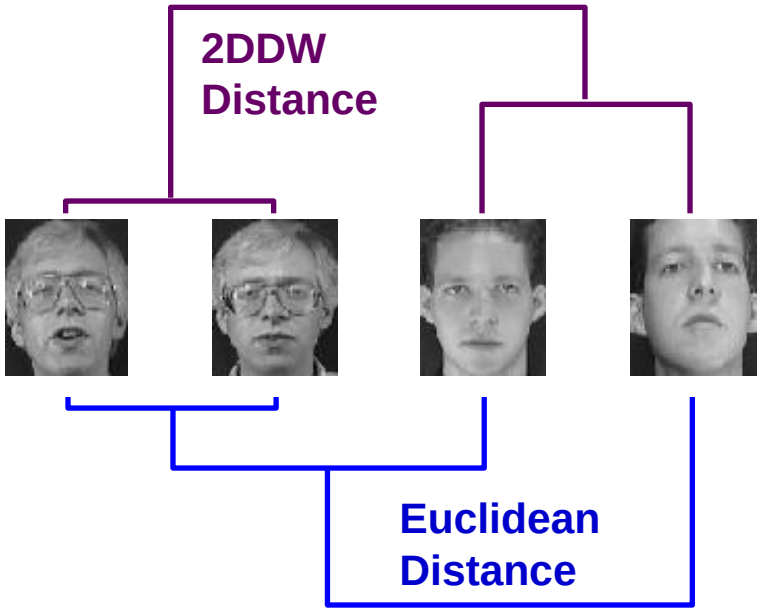


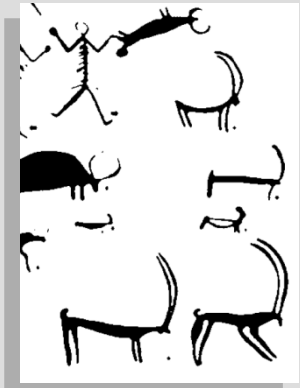
Figure 3: Two pairs of faces clustered using 2DDW (*top*) and Euclidean distance (*bottom*)

Looking at this figure, we can tell that 2DDW produces more intuitive results than Euclidean Distance in 2 or 3 seconds.

Paradoxically, this figure has less information (hierarchical clustering is *lossy* relative to a distance matrix) but communicates a lot more *knowledge*.

Contrast these two figures, both of which attempt to show that petroglyphs can be clustered meaningfully.

- **Thinking** about the…, helps
- **Color** helps
- **Direct** labeling helps
- **Meaningful** captions helps



To figure out the utility of the similarity measures in this paper, you need to look at text and two figures, spanning four pages.

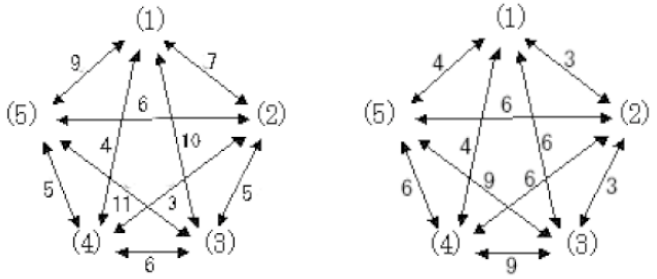


Fig. 7. Mutual distances among five ibexes drawn in the same figures, Fig. 4(g) (left) and Fig. 4(m) (right). The average distances within these figures are 6.6 and 5.6, respectively.

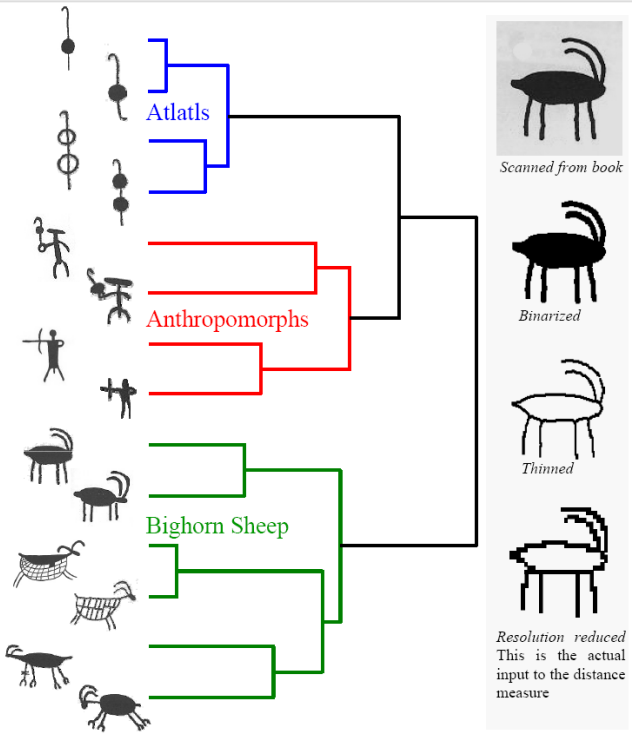


Figure 11: (*left*) A group-average linkage hierarchical clustering of typical Southwestern USA petroglyphs, with the $D_{clustering}$ measure. (*right*) While the dendrogram to the left shows the full resolution images for clarity, the images input to the distance measure have binarized, thinned and scaled to fit in a 30 by 30 bounding rectangle

Using the labels "**Method1**" **Method2**" etc, gives a level of indirection. We have to keep referring back to the text (on a different page) to understand the content.

Direct labeling helps

The four significant digits are ludicrous on a data set with 300 objects.

Len	Method 1	Method 2	Method 3	Method 4	Method 5
128	0.7767	0.9589	0.9589	0.7772	0.77
256	0.7144	0.9567	0.9411	0.8622	0.7433
512	0.6683	0.9419	0.9508	0.9408	0.7781
Avg	**0.7198**	**0.9525**	**0.9503**	**0.8601**	**0.7638**

Table 3: Similarity Results for CBF Trials

Redesigned by Keogh

Length	Sequential Sparsification	Linear Sparsification	Quadratic Sparsification	Wavelet Sparsification	Raw Data
128	0.77	0.95	0.95	0.77	0.77
256	0.71	0.95	0.94	0.86	0.74
512	0.66	0.94	0.95	0.94	0.77
Avg	**0.71**	**0.95**	**0.95**	**0.86**	**0.76**

Table 3: Similarity Results for CBF Trials

This paper offers 7 significant digits in the results on a dataset a few thousand items

Table VII. Result for Stream Segment of Length 6

U			σ	Ψ
-0.5314	0.7800	0.3304	937.7485	
-0.5797	-0.0504	-0.8133	133.1651	0.99991
-0.6178	-0.6237	0.4789	20.4448	

This paper offers 9 significant digits in the results on a dataset a few hundred items

Table II. Part of the 10-6

Curve no.	Real saturation point		Computed saturation point	
1	289	576998	228.88	602278.75
2	265	4425584	252.762	4498063.5
3	293	640952	234.418	674823.063
4	353	61662912	256.503	61871660
5	325	11038734	251.434	11018625
6	277	666613	252.498	686351.813
7	309	375608	216.822	372590.375

Spurious digits are not just unnecessary, they are a lie! They imply a precision that you do not have. At best they make you look like an amateur.

Pseudo code

As with real code, it is probably better to break very long puesdocode into several shorter units

Algorithm 1 Temporal Hashing Algorithm (*TH*) and Temporal Hashing with Dropping Algorithm (*THwD*)

```
1: Initialize the temporal table TT, location queue LQ, history distance H
2: Initialize flag flag = 1 for TH, and flag = 2 for THwD
3: for Each incoming location < (x, y), t > that belongs to trajectory P(i) do
4:     if flag == 2 then
5:         Get Dormant Time DP(i)
6:         if DP(i) − t < w then
7:             if q(i) == null then
8:                 Initialize q(i)
9:             else
10:                q(i).update((x,y))
11:            end if
12:        else
13:            store P[t]
14:        end if
15:    else
16:        q(i).update((x,y))
17:    end if
18: end for
19: if t <= w − 1 then
20:    for Each query Q(j) do
21:        for Each trajectory P(i) do
22:            compute MinDist(j), update history distance H and CINN(j)
23:        end for
24:    end for
25: else if t == w then
26:    for Each query Q(j) do
27:        for Each trajectory P(i) do
28:            compute MinDist(j), update history distance H and CINN(j)
29:            compute hash value δt using Formula (3)
30:            if δt == w − 1 then
31:                compute hash value δt using Formula (6)
32:                if (flag == 2 and δt ≥ w) then
33:                    Drop location queue and the historical distances of P(i).
34:                end if
35:            end if
36:        end for
37:    end for
38:    for Each trajectory P(i) do
39:        TT.map(i, min((t + δt))
40:    end for
41: else
42:    for Each query Q(j) do
43:        for Each trajectory P(k) in TT(t) do
44:            compute MinDist(j), update history distance H and CINN(j)
45:            compute hash value δt according to Formula (3) and (6) similarly
46:            if (flag == 2 and δt ≥ w) then
47:                Drop location queue and the historical distances of P(k).
48:            end if
49:        end for
50:    end for
51:    for Each trajectory P(k) in TT(t) do
52:        TT.map(k, min(t + δt))
53:    end for
54: end if
```

The most Common Problems with Figures

1. Too many patterns on bars
2. Use of both different symbols and different lines
3. Too many shades of gray on bars
4. Lines too thin (or thick)
5. Use of three-dimensional bars for only two variables
6. Lettering too small and font difficult to read
7. Symbols too small or difficult to distinguish
8. Redundant title printed on graph
9. Use of gray symbols or lines
10. Key outside the graph
11. Unnecessary numbers in the axis
12. Multiple colors map to the same shade of gray
13. Unnecessary shading in background
14. Using bitmap graphics (instead of vector graphics)
15. General carelessness

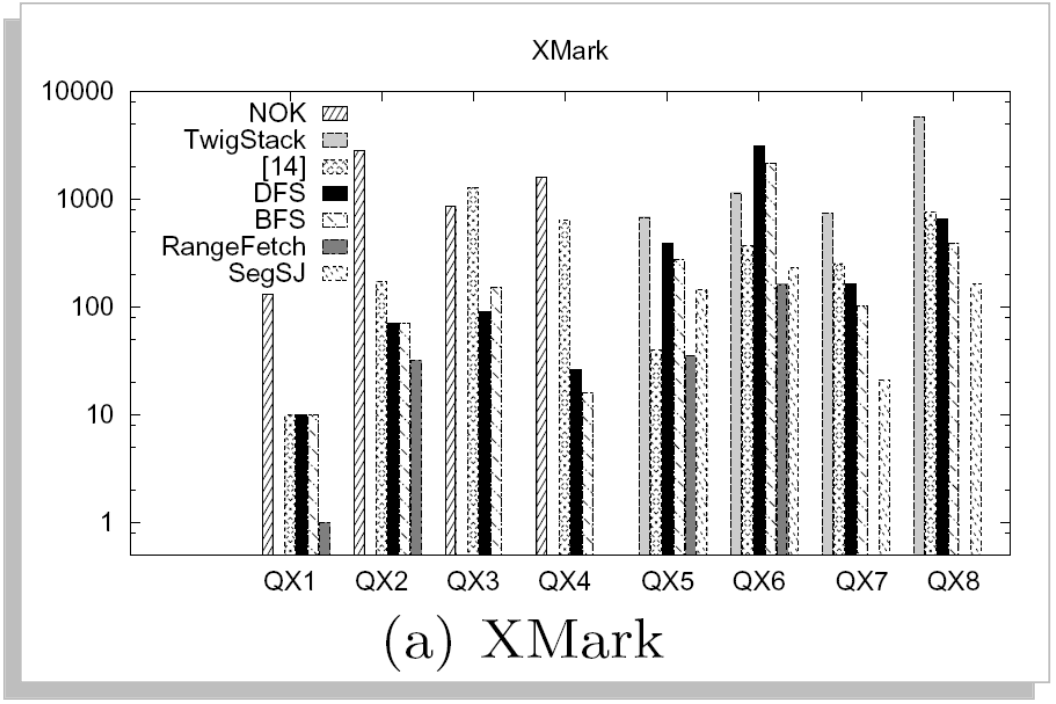
Eileen K Schofield: Quality of Graphs in Scientific Journals: An Exploratory Study. Science Editor, 25 (2), 39-41

Eamonn Keogh: *My Pet Peeves*

1. Too many patterns on bars

Here the problem is compounded by the tiny size of the key. The area of each key-box is about $2mm^2$

The key drawn to scale.



5. Use of three-dimensional bars for only two variables

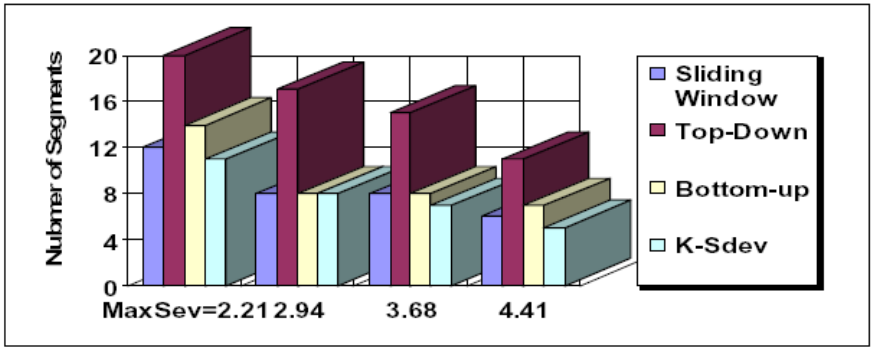
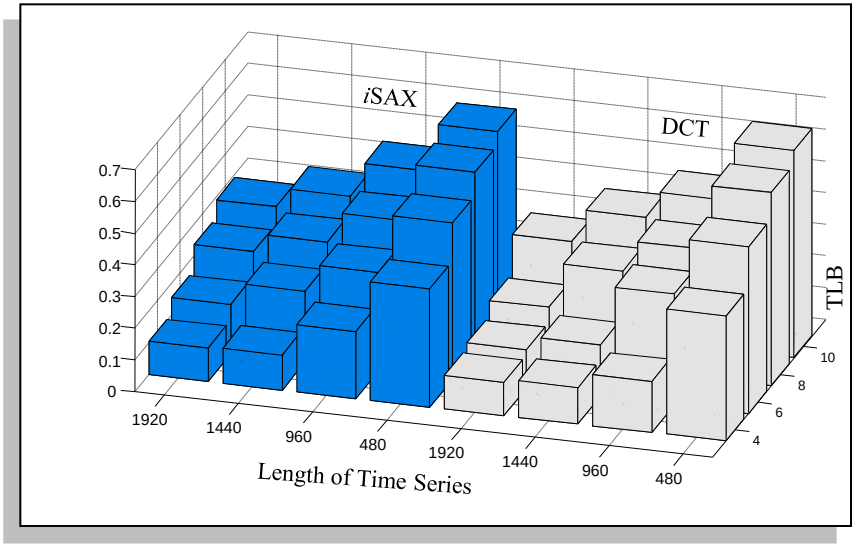


Fig. 4 Comparison with three generic segmentation algorithms

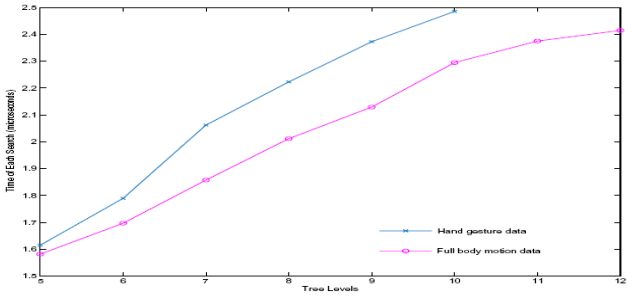
Why is this chart in 3D?

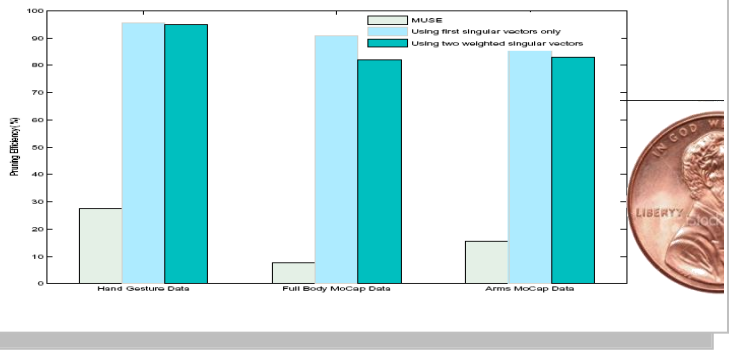


3D *is* fine when needed

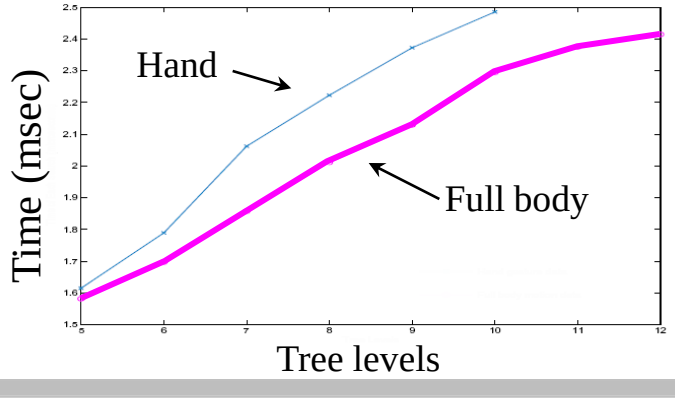
6. Lettering too small and font difficult to read

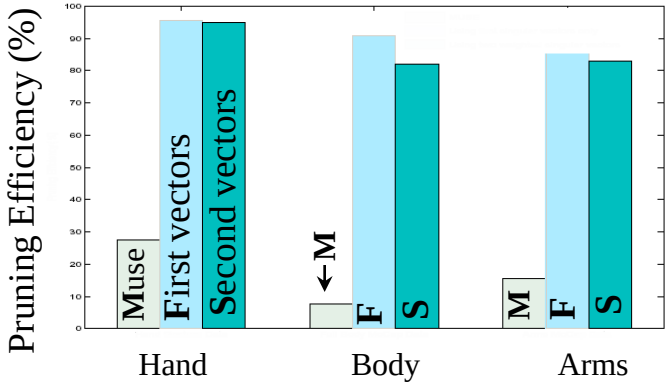
Here the font size on the legend and key is about 1mm. (coin for scale)





All the problems are trivial to fix



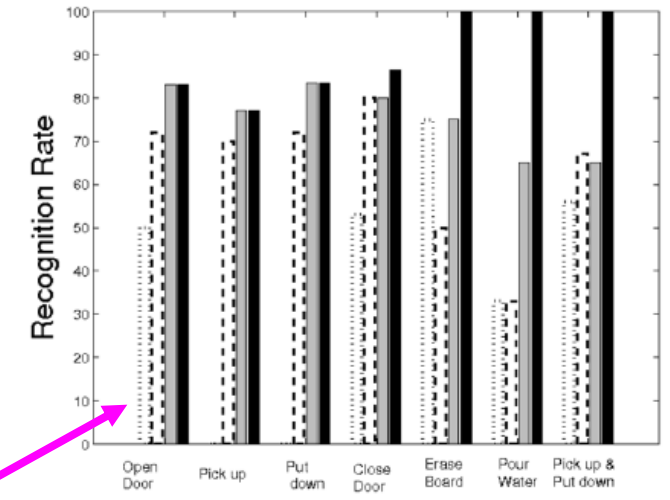


10. Key outside the graph

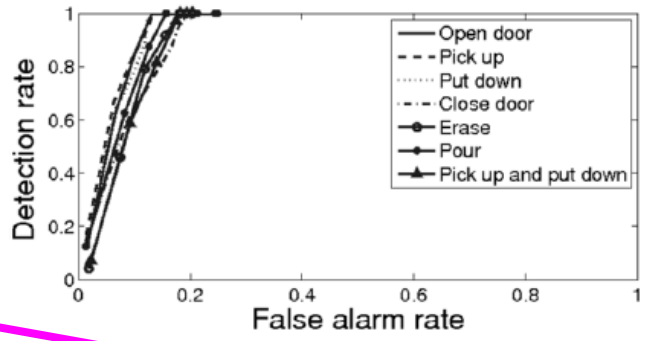
Here the problem is not that the key is in *text* format (although it does not help). The problem is the distance between the key and the data.

Data

Key



(a)

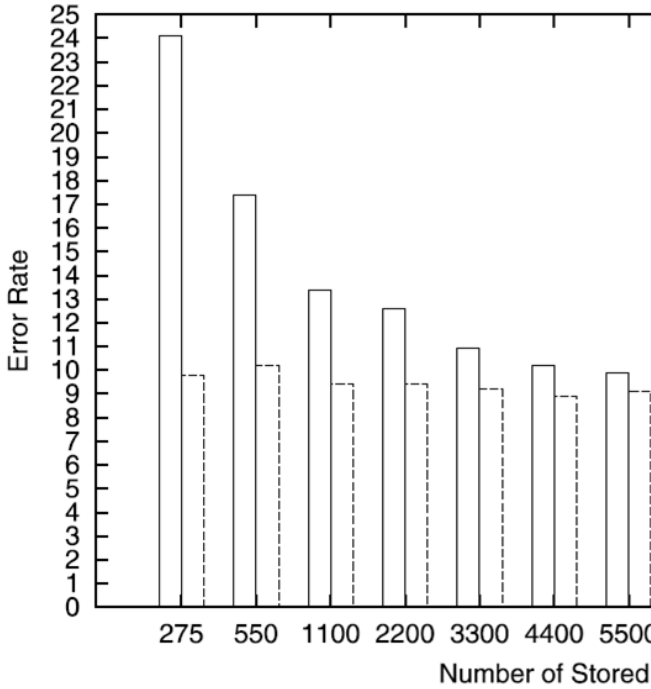


(b)

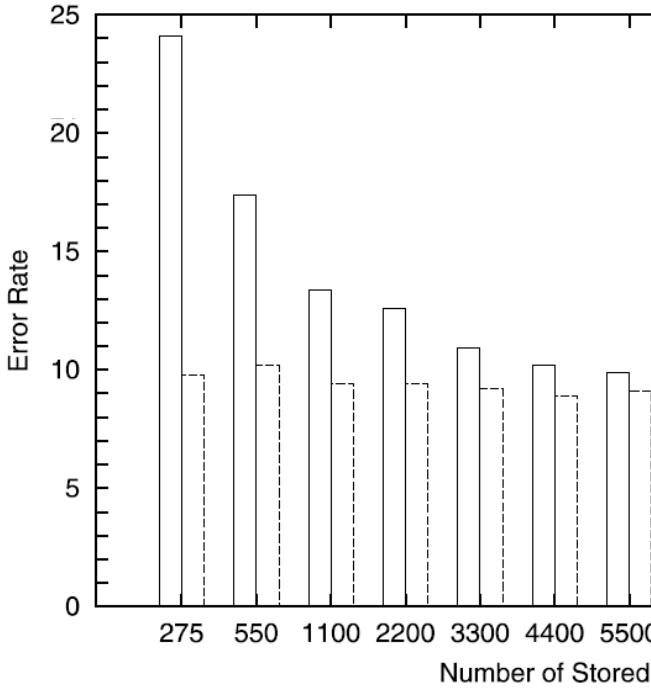
Fig. 11. (a) Recognition rate for UCF database. Dotted line: UCF results [12]. Dashed line: Overall similarity obtained by integrating P similarity scores from coarse-to-fine scales. Gray bar: Conditionally optimal p^*. Black bar: Jointly optimal $(N, p)^*$. (b) ROC curves for different activities in the conditionally optimal case.

11. Unnecessary numbers in the axis

Do we really need every integer from zero to 25 in this chart? (if “yes”, then make a table, not a figure)



In this version, I can still find, say “23”, by locating 20 and counting three check marks.



This problem is more common in the X-axis

12. Multiple colors map to the same shade of gray

This image works fine in color…

In B/W however, multiples colors map to the same shades of gray.

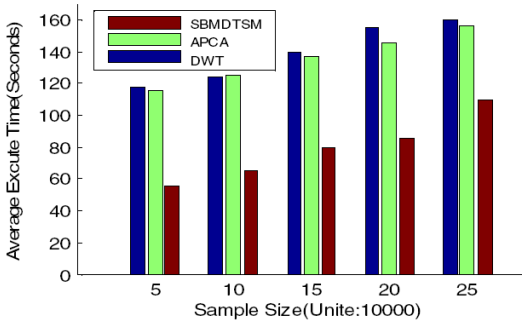


Figure 6. Average execute time

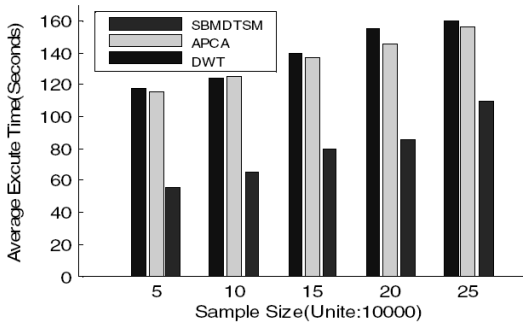
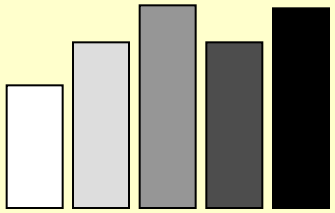
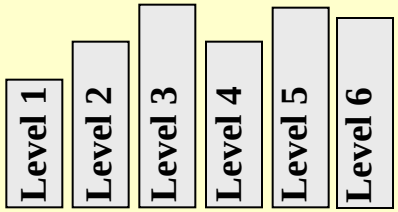
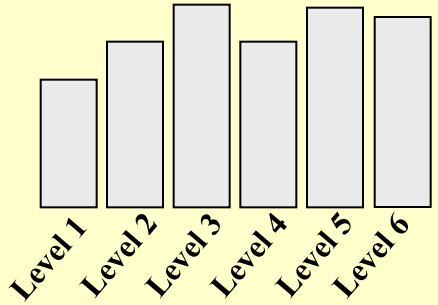


Figure 6. Average execute time

Note that we *can* easily represent upto 5 things with shades of gray. We can also *directly label* bars.

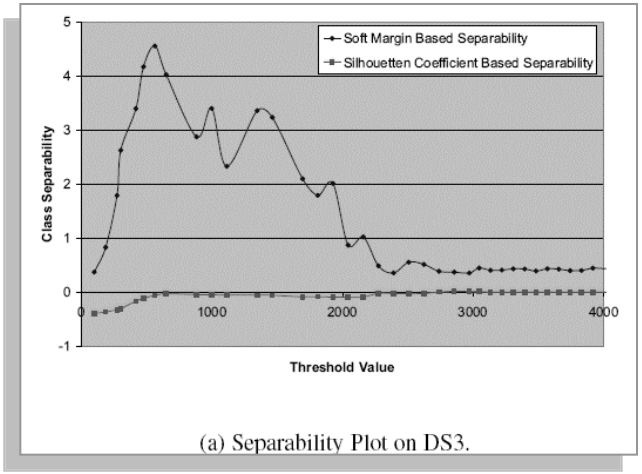






13. Unnecessary shading in background

All the other problems (Multiple colors map to the same shade of gray, etc) are compounded by having a shaded background.



(a) Separability Plot on DS3.

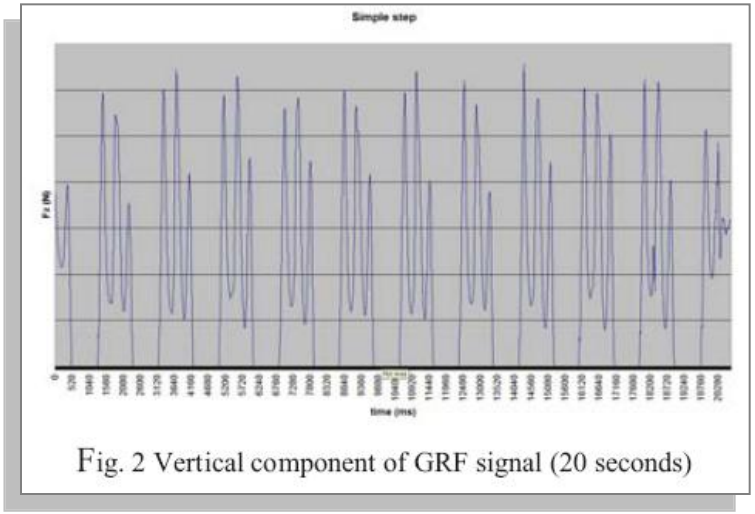


Fig. 2 Vertical component of GRF signal (20 seconds)

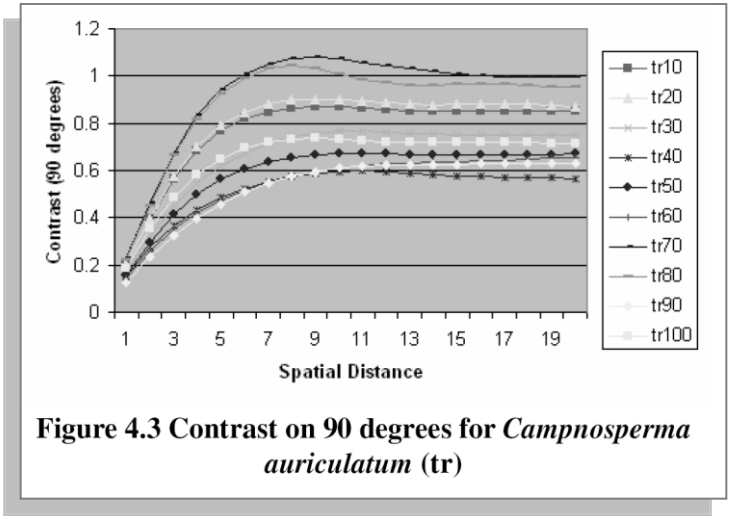
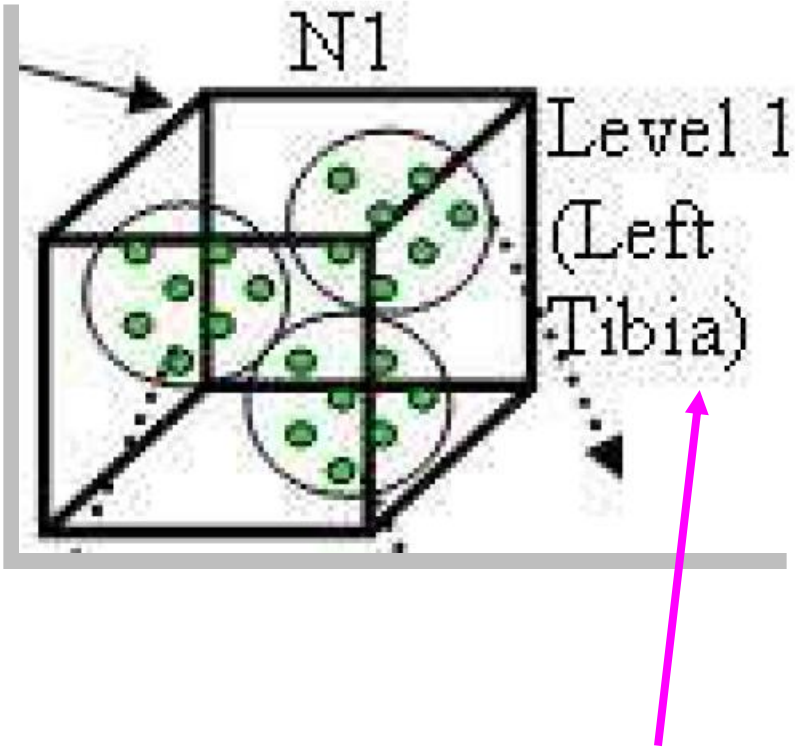
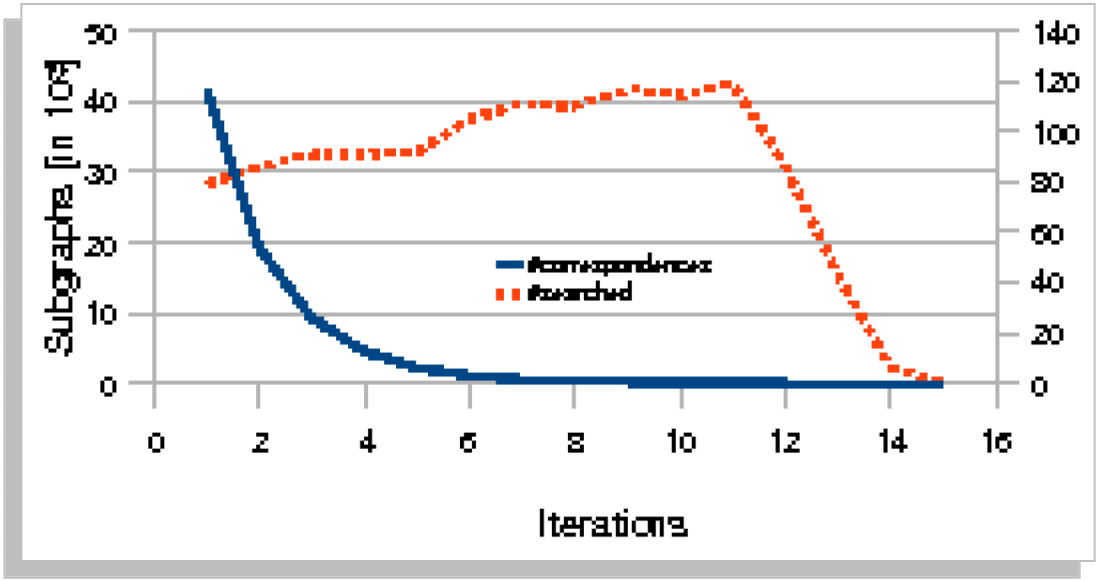


Figure 4.3 Contrast on 90 degrees for *Campnosperma auriculatum* (tr)

14 Using bitmap graphics



Below is a particularly bad example, compounded by a tiny font size, however even the best bitmaps look amateurish and can hard to read.
Use *vector* graphics.



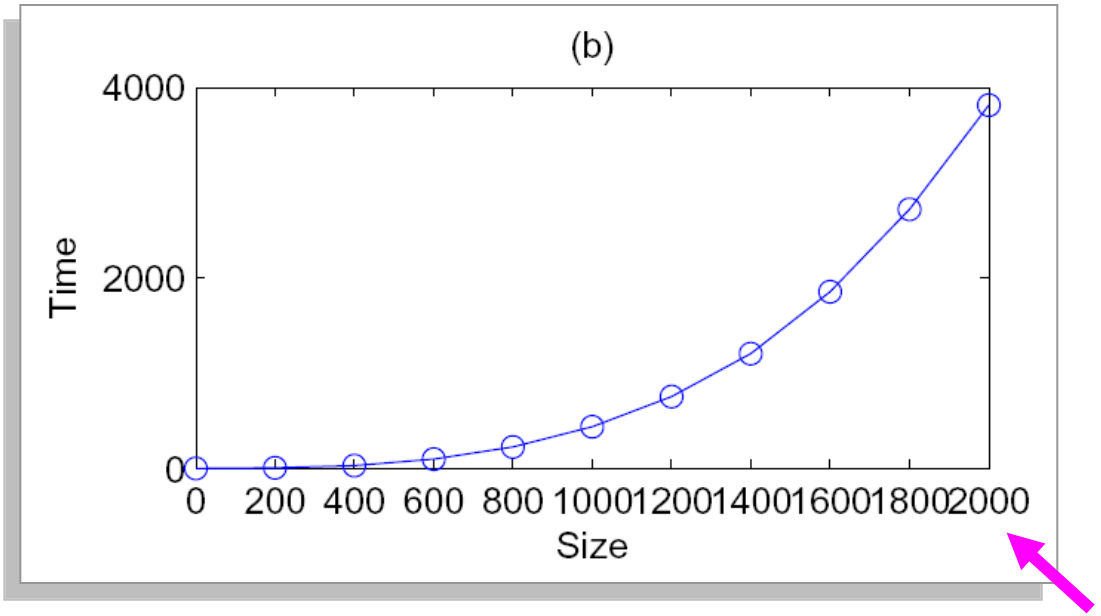
Bitmap graphics often have *compression artifacts*, resulting in noise around sharp lines.

15 General Carelessness

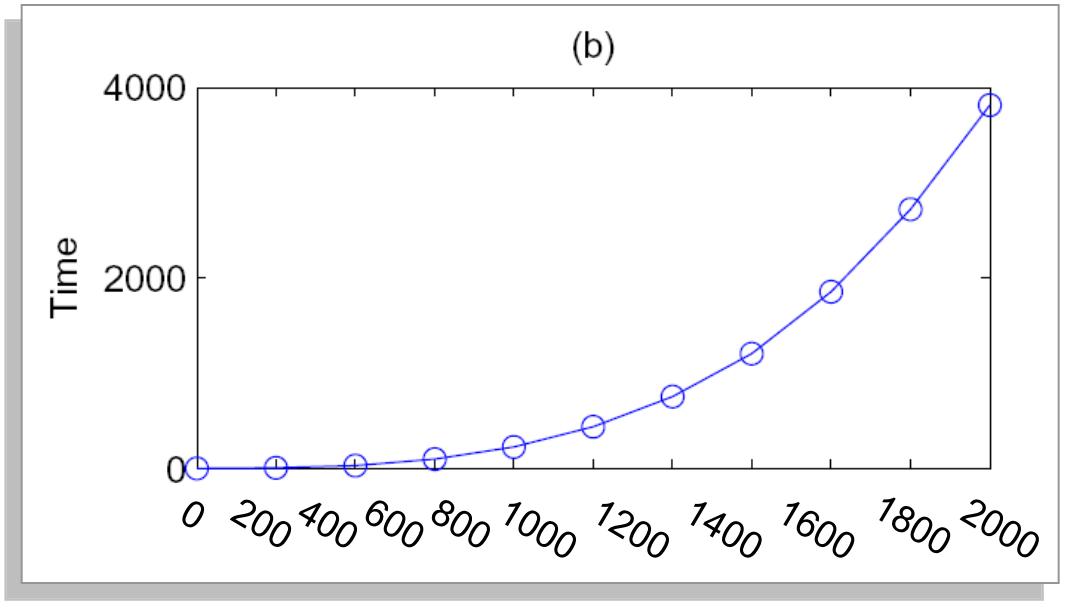
Why did the authors of this graphic not spend the 30 seconds it took to fix this problem?

Such careless figures are an insult to reviewers.

Original



Fixed



Top Ten Avoidable Reasons Papers get Rejected, with Solutions

To catch a thief, you must think like a thief

Old French Proverb

To convince a reviewer, you must think like a reviewer

Always write your paper imagining the most cynical reviewer looking over your shoulder*. This reviewer does not particularly like you, does not have a lot of time to spend on your paper, and does not think you are working in an interesting area. But he/she *will* listen to reason.

*See *How NOT to review a paper: The tools and techniques of the adversarial reviewer* by Graham Cormode

This paper is out of scope for SDM

- In some cases, your paper may really be irretrievably out of scope, so send it elsewhere.
- Solution
  - Did you read and reference SDM papers?
  - Did you frame the problem as a SDM problem?
  - Did you test on well known SDM datasets?
  - Did you use the common SDM evaluation metrics?
  - Did you use SDM formatting? (“look and feel”)
  - Can you write an explicit section that says: *At first blush this problem might seem like a signal processing problem, but note that..*

The experiments are not reproducible

- This is becoming more and more common as a reason for rejection and some conferences now have official standards for reproducibility
- Solution
  - Create a webpage with all the data/code and the paper itself.
  - Do the following sanity check. Assume you lose all files. Using *just* the webpage, can you recreate all the experiments in your paper? (it is easy to fool yourself here, *really really* think about this, or have a grad student actually attempt it).
  - Forcing yourself to do this will eliminate 99% of the problems

This is too similar to your last paper

- If you really are trying to "double-dip" then this is a justifiable reject.
- Solution
  - Did you reference your previous work?
  - Did you explicitly spend at least a paragraph explaining how you are *extending* that work (or, are *different to* that work).
  - Are you reusing all your introduction text and figures etc. It might be worth the effort to redo them.
  - If your last paper measured, say, accuracy on dataset X, and this paper is also about improving accuracy, did you compare to your last work on X? (note that this does not exclude you from *additional* datasets/rival methods, but if you don't compare to your previous work, you look like you are hiding something)

You did not acknowledge this weakness

- This looks like you either don't know it is a weakness (you are an idiot) or you are pretending it is not a weakness (you are a liar).
- Solution
  - Explicitly acknowledge the weaknesses, and explain why the work is still useful (and, if possible, how it might be fixed)

"*While our algorithm only works for discrete data, as we noted in section 4, there are commercially important problems in the discrete domain. We further believe that we may be able to mitigate this weakness by considering…*"

You unfairly diminish others work

- Compare:
  - "*In her inspiring paper Smith shows.... We extend her foundation by mitigating the need for...*"
  - "*Smith's idea is slow and clumsy.... we fixed it.*"
- Some reviewers noted that they would not explicitly tell the authors that they felt their papers was unfairly critical/dismissive (such subjective feedback takes time to write), but it would temper how they felt about the paper.
- Solution
  - Send a preview to the rival authors: "*Dear Sue, we are trying to extend your idea and we wanted to make sure that we represented your work correctly and fairly, would you mind taking a look at this preview…*"

There is a easier way to solve this problem. You did not compare to the X algorithm

- Solution
  - Include simple strawmen (“*while we do not expect the hamming distance to work well for the reasons we discussed, we include it for completeness*”)
  - Write an explicit explanation as to why other methods won’t work (see below). But don’t just say “*Smith says the hamming distance is not good, so we didn’t try it*”

It is important to dismiss two apparent solutions to this problem before introducing our technique:

- *Why not replace the Euclidean distance with the Dynamic Time Warping (DTW) distance?* While DTW is a very useful tool for many data mining problems, it is not the solution here. For example, if we have a

You do not reference this related work. This idea is already known, see Lee 1978

- Solution
  - Do a *detailed* literature search.
  - If the related literature is huge, write a longer tech report and say in your paper "*The related work in this area is vast, we refer the interested reader to our tech-report for a more detailed survey*"
  - Give a draft of your paper to mock-reviewers ahead of time.
  - Even if you have accidentally rediscovered a known result, you might be able to fix this if you know ahead of time. For example "*In our paper we reintroduced an obscure result from cartography to data mining and show…*"

(In ten years I have rejected 4 papers that rediscovered the Douglas-Peuker algorithm.)

You have too many parameters/magic numbers/arbitrary choices

- Solution
  - For *every* parameter, either:
    - Show how you can set the value (by theory or experiment)
    - Show your idea is not sensitive to the exact values
  - Explain *every* choice.
    - If your choice was arbitrary, state that explicitly. *We used single linkage in all our experiments, we also tried average, group and Wards linkage, but found it made almost no difference, so we omitted those results for brevity* (*but the results are archive in our tech report*).
    - If your choice was not arbitrary, justify it. *We chose DCT instead of the more traditional DFT for three reasons, which are…*

Not an interesting or important problem. Why do we care?

- Solution
  - Did you test on real data?
  - Did you have a domain expert collaborator help with motivation?
  - Did you *explicitly* state why this is an important problem?
  - Can you estimate value? "*In this case switching from motif 8 to motif 5 gives us a nearly $40,000 in annual savings!* Patnaiky et al. SIGKDD 2009"
  - Note that estimated value does not have to be in dollars, it could be in crimes solved, lives saved etc

The writing is generally careless. There are many typos, unclear figures

This may seem unfair if your paper has a good idea, but reviewing carelessly written papers is frustrating. Many reviewers will assume that you put as much care into the experiments as you did with the presentation.

- Solution
  - Finish writing well ahead of time, pay someone to check the writing.
  - Use mock reviewers.
  - Take pride in your work!

Tutorial Summary

- Publishing in top tier venues such as SDM can seem daunting, and can be frustrating…
- But you can do it!

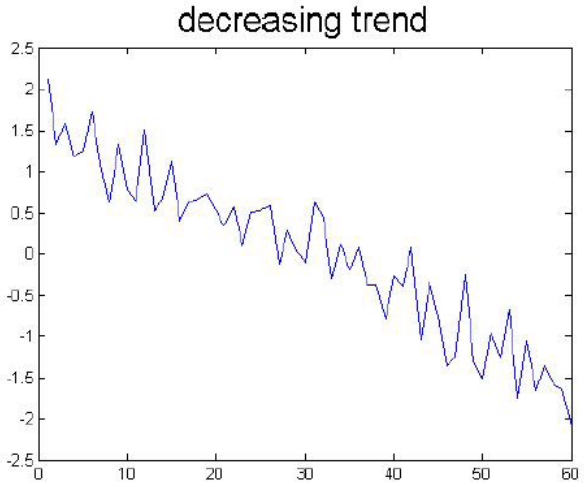
- Taking a systematic approach, and being self-critical at every stage will help you chances greatly.
- Having an external critical eye (mock-reviewers) will also help you chances greatly.

The End

Appendix A:

Why mock reviewers can help

A mock reviewer might have spotted that "upward shift" was misspelled, or that "Negro" is not a good choice of words, or…



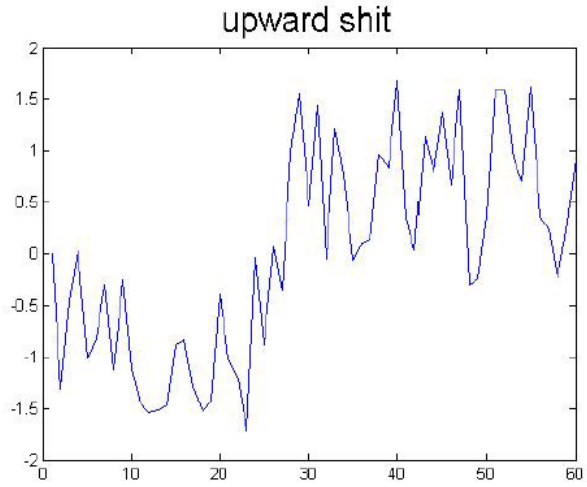
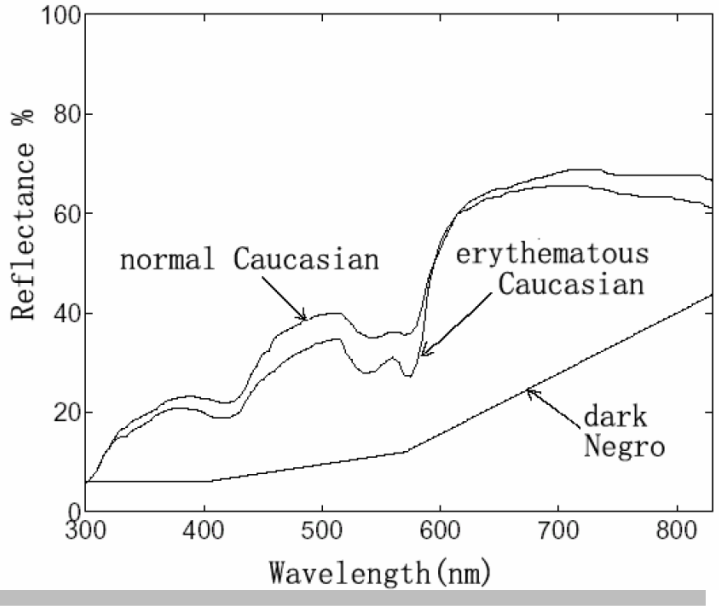


Fig. 4. An example time series from each class of the



FART Neural Network based Probabilistic Motif Discovery in Unaligned Biological Sequences

M. Hemalatha, P. Ranjit Jeba Thangaiah and K. Vivekanandan, Member IEEE

Abstract – **Finding Motif in bio-sequences is the most primitive operation in computational Biology. There are many computational requirements for a motif discovery algorithm such as computer memory space requirement and computational complexity. To overcome the complexity of motif discovery, an alternative solution is proposed by integrating genetic algorithm and Fuzzy Art machine learning approaches for eliminating multiple sequence alignment**

Importance of these patterns for biology comes from the role of motifs at protein DNA binding sites. Furthermore, finding similar sequences can be used to reveal unknown evolutionary relationships between different species.

II. MATERIALS AND METHODS

Appendix B: Be concrete

SAX is a kind of statistical **algorithm**…

No, SAX is a data **representation**

Finally, Dynamic Time Warping **metric** was…

The same dynamic time warping **metric** was used to compare clusters…

… or dynamic time warping **metric** and to retrieve the last sensor data…

No, Dynamic Time Warping is a **measure**, not a metric

Appendix C:

The owner of a small company needed to get rid of an old boiler that his company had replaced with a shiny new one. Not wanting to pay disposal fees, and thinking that someone else could use it, he dragged it out onto the street and put a "Free" sign on it. To his dismay, a week later it was still there. He was about to call a disposal company when his foreman said "*I can get rid of it in one day*".

The foreman replaced the "**Free**" sign with one that said "*For Sale, $1,500*". That night, the boiler was stolen.

The moral? *Imply value* for your paper.

- A biologist, an engineer and a mathematician were crossing the border into Scotland from England on a train when they saw a field with a black sheep in it.
- The biologist said, "*Look, in Scotland the sheep are black*."
- The engineer replied, "*No, in Scotland some of the sheep are black*."
- The mathematician rolled his eyes to heaven and said, very patiently, "*In Scotland, there exists at least one sheep which is black on at least one side*."