Visually-Motivated Characterizations of Point Sets Embedded in High-Dimensional Geometric Spaces: Current Projects

Leland Wilkinson
SYSTAT
University of Illinois at Chicago (Computer Science)
Northwestern University (Statistics)

Robert Grossman
University of Illinois at Chicago (Computer Science)

Adilson Motter
Northwestern University (Physics)

SYSTAT

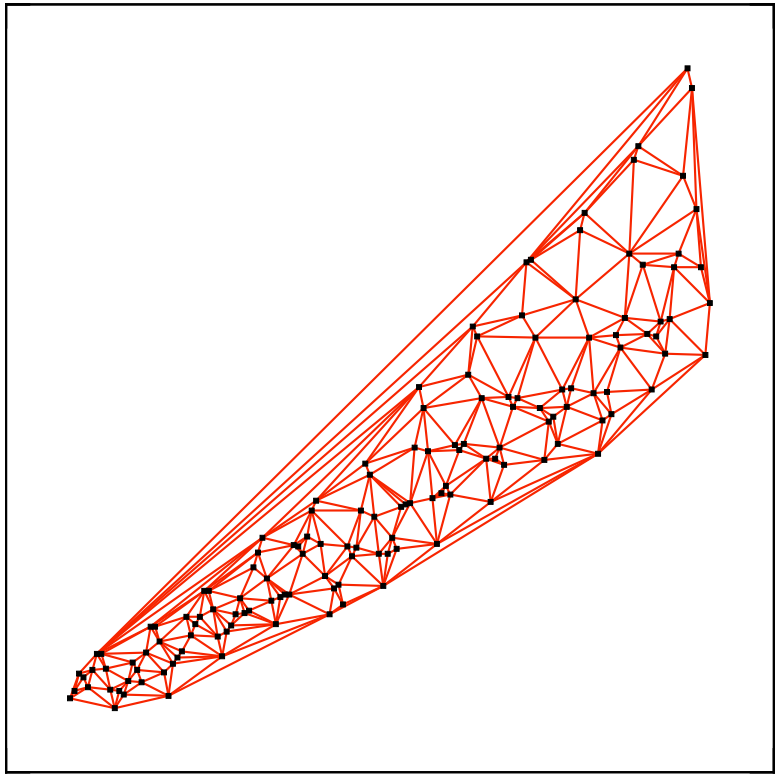
Subprojects

- Scagnostics
- Classification
- Venn/Euler Diagrams
- Treemaps

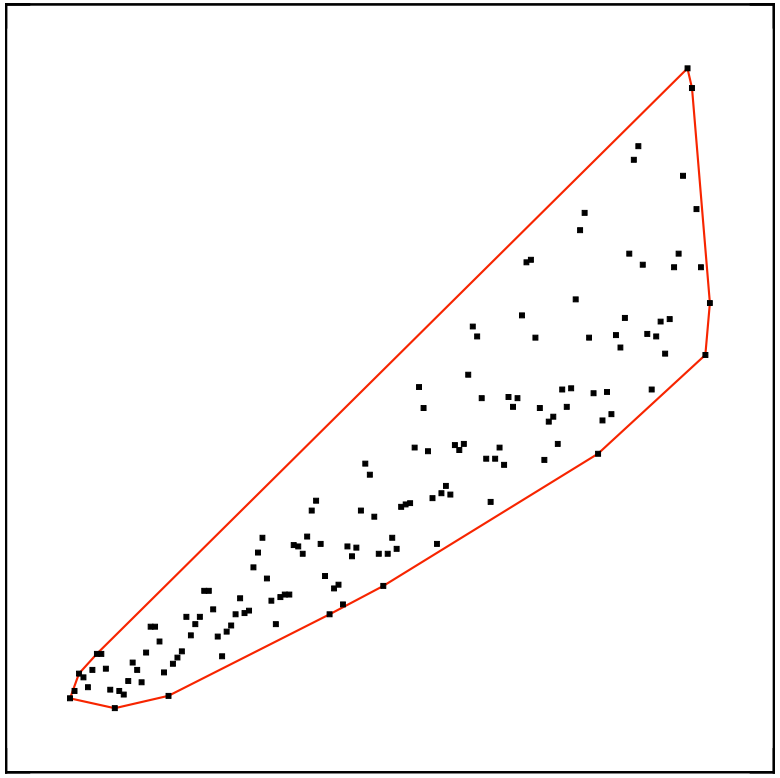
Scagnostics

- Wilkinson, Anand, and Grossman (2006) characterize a scatterplot (2D point set) with nine measures.
- We base our measures on three *geometric graphs*.
  - Convex Hull
  - Alpha Shape
  - Minimum Spanning Tree

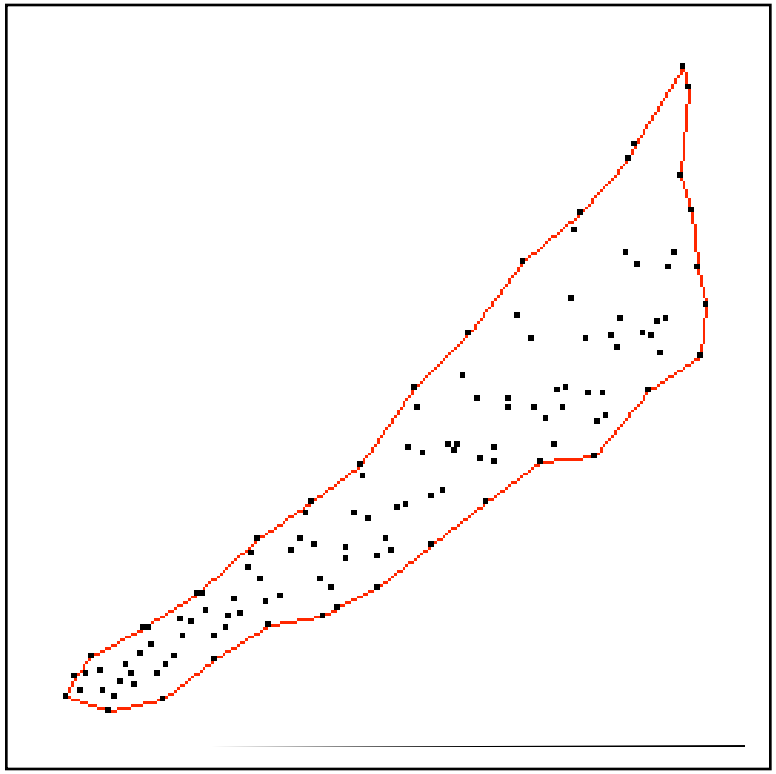
- Each geometric graph is a subset of the Delaunay Triangulation



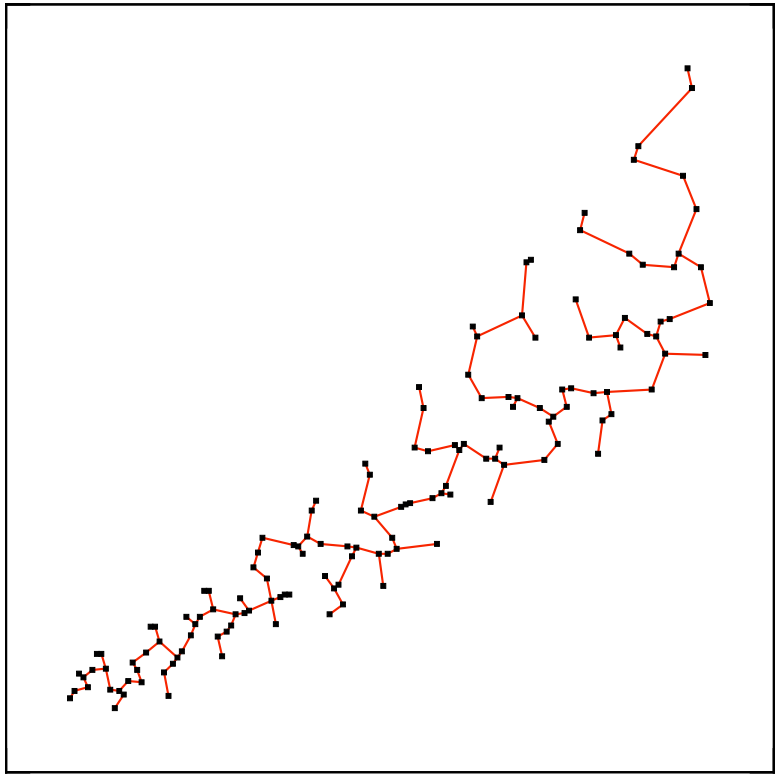
Convex Hull



Alpha Shape



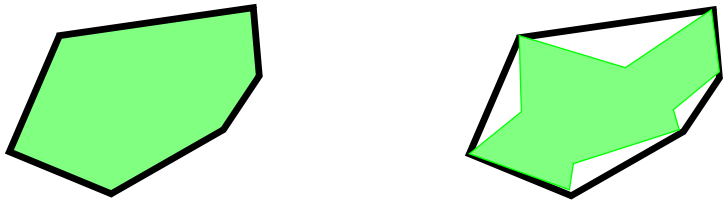
Minimum Spanning Tree



SYSTAT

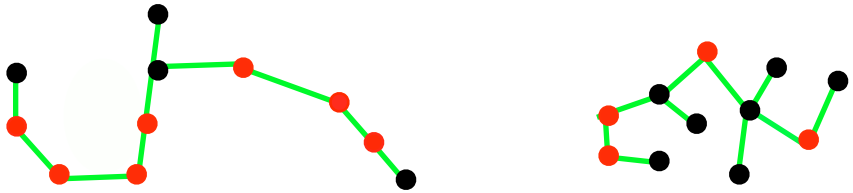
Shape

Convex: area of alpha shape divided by area of convex hull



Skinny: ratio of perimeter to area of the alpha shape

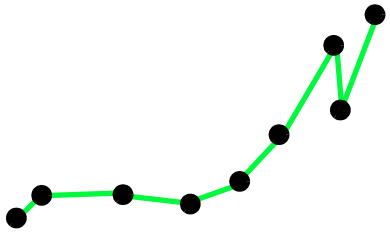
Stringy: ratio of 2-degree vertices in MST to number of vertices > 1-degree

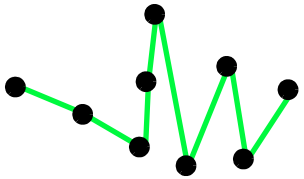


SYSTAT

Trend

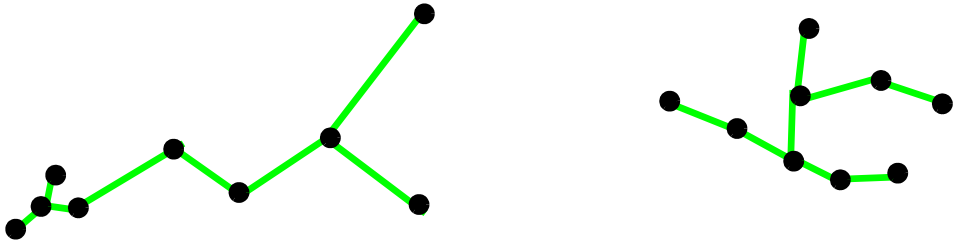
Monotonic: squared Spearman correlation coefficient



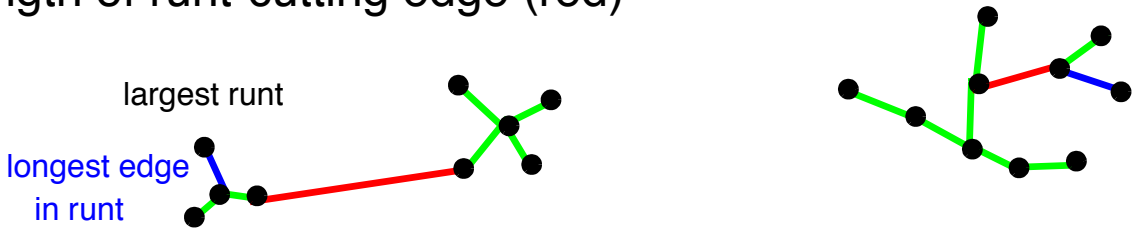


Density

Skewed: ratio of $(Q90 - Q50) / (Q90 - Q10)$, where quantiles are on MST edge lengths



Clumpy: 1 minus the ratio of the longest edge in the largest runt (blue) to the length of runt-cutting edge (red)



Outlying: proportion of total MST length due to edges adjacent to outliers

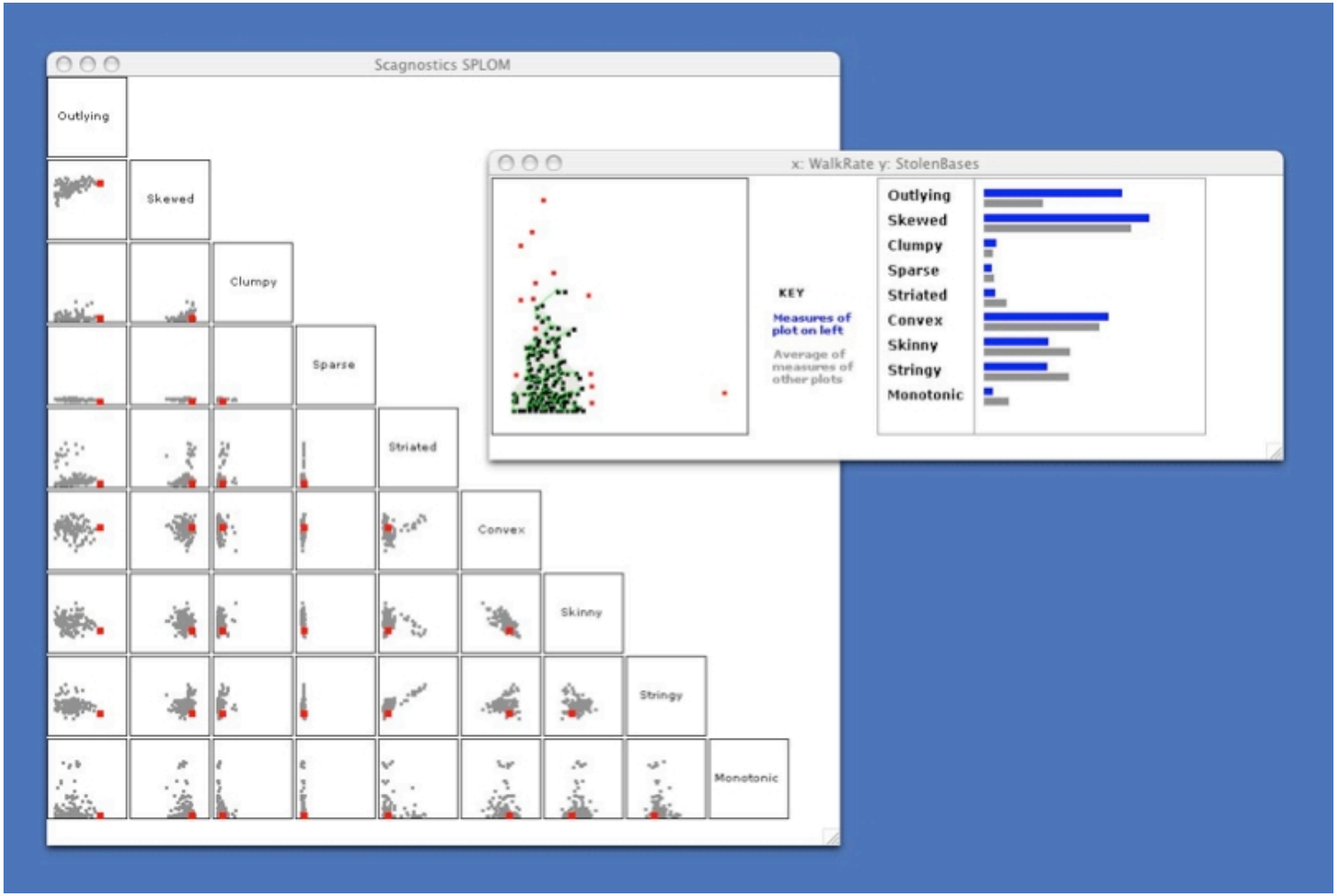
Density

Sparse: 90th percentile of distribution of edge lengths in MST

Striated: proportion of all vertices in the MST that are degree-2 and have a cosine between adjacent edges less than -.75

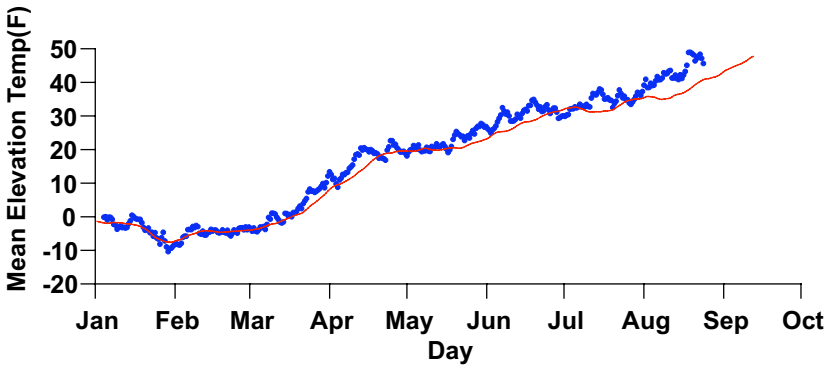
SYSTAT

Software (Wilkinson and Anand)

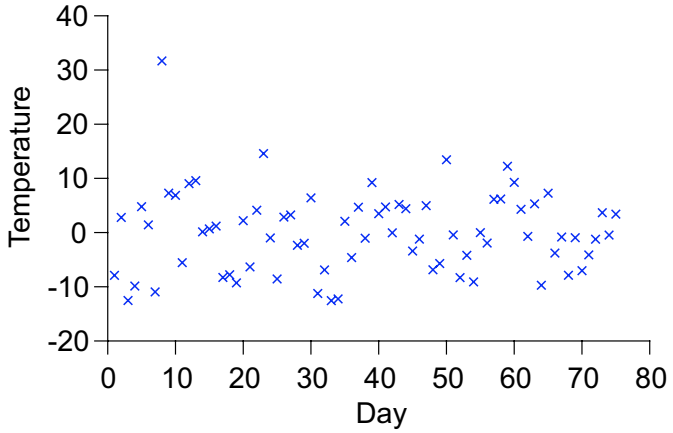


Fishing Expeditions in Visual Analytics

- We have used the empirical distribution of Scagnostics (Wilkinson and Wills, *JCGS*, 2008), the False Discovery Rate (FDR) statistic, and automated statistical modeling to develop algorithms for reducing false discoveries when people use visual-analytic software (Wilkinson and Wills, *IVS*, 2009).



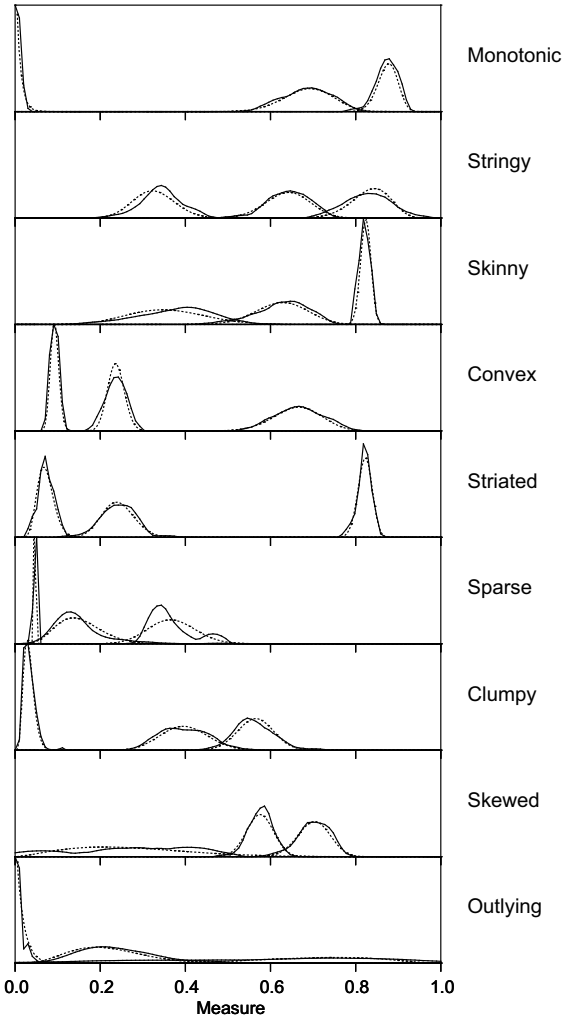
These data are fake. They are a random walk produced by a random number generator.



These data are real. They are temperature measurements of a cow over 80-days. The data are periodic.

SYSTAT

Distribution of Scagnostics



SYSTAT

False Discovery Rate (FDR)

Benjamini and Hochberg, *JRSS*, 1995

- Sort a list of "p" values and define $p_0 = 0$
  - $P = [p_1, \ldots, p_m]$
- Compute BH index

$$r_{BH} = max\left\{0 \le i \le m : p_i \le \alpha \frac{i}{m}\right\}$$

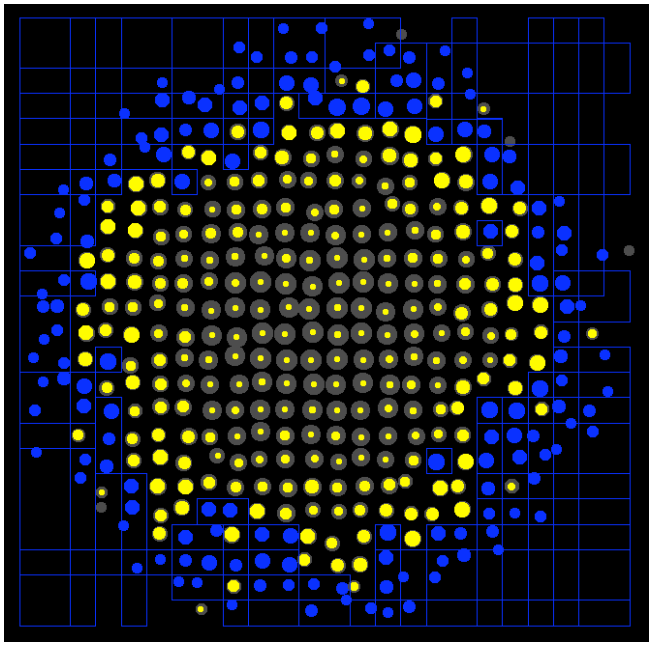
- Use adjusted "p" value (indexed by r_{BH}) as cutoff
- Based on uniform distribution of "p" values, so tests may be heterogeneous

Scagnostics on Graphs

- Adilson Motter is developing scagnostics for (*V, E*) graphs. This effort is similar to what we did for scatterplots. One wants a relatively small set of measures that are relatively independent and that nicely characterize a wide universe of real directed and undirected graphs. The goal is to classify real graphs, recognize anomalies, and locate exemplars in subgraphs of large graphs.

Classification

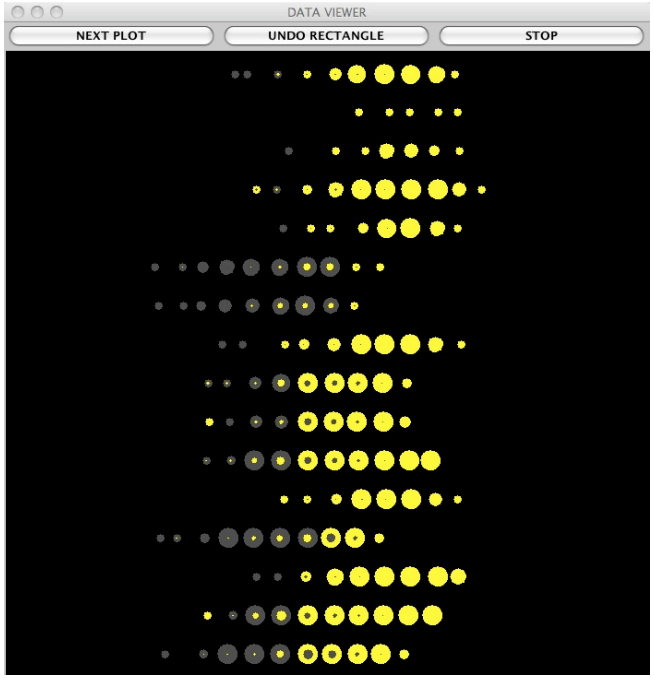
- Based on our analysis of how people visually classify 2D point sets, we have developed a high-dimensional, linear complexity, supervised classifier called Linf. This classifier, using the L-infinity metric, rivals or outperforms leading classifiers on 10 “difficult” datasets.



SYSTAT

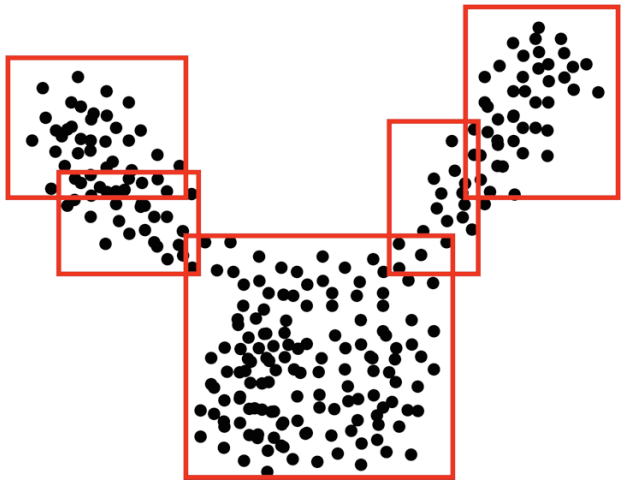
Visual Classification

- Anand, Wilkinson, and Tuan (*ICDM*, 2009) developed a visual classifier to analyze how people distinguish target point sets from background point sets. We designed the software to behave like a video game.



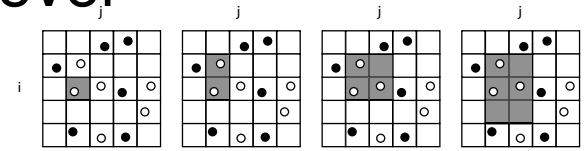
Description Regions

- Weighted L-infinity norm
  - $||x||_\infty = sup(w_1|x_1|, w_2|x_2|, \ldots w_n|x_n|)$
- Hypercube Description Region (HDR)
  - The set of points less than a fixed distance from a single point using the L-infinity norm.
- Composite Hypercube Description Region (CHDR)
  - Union of HDRs.



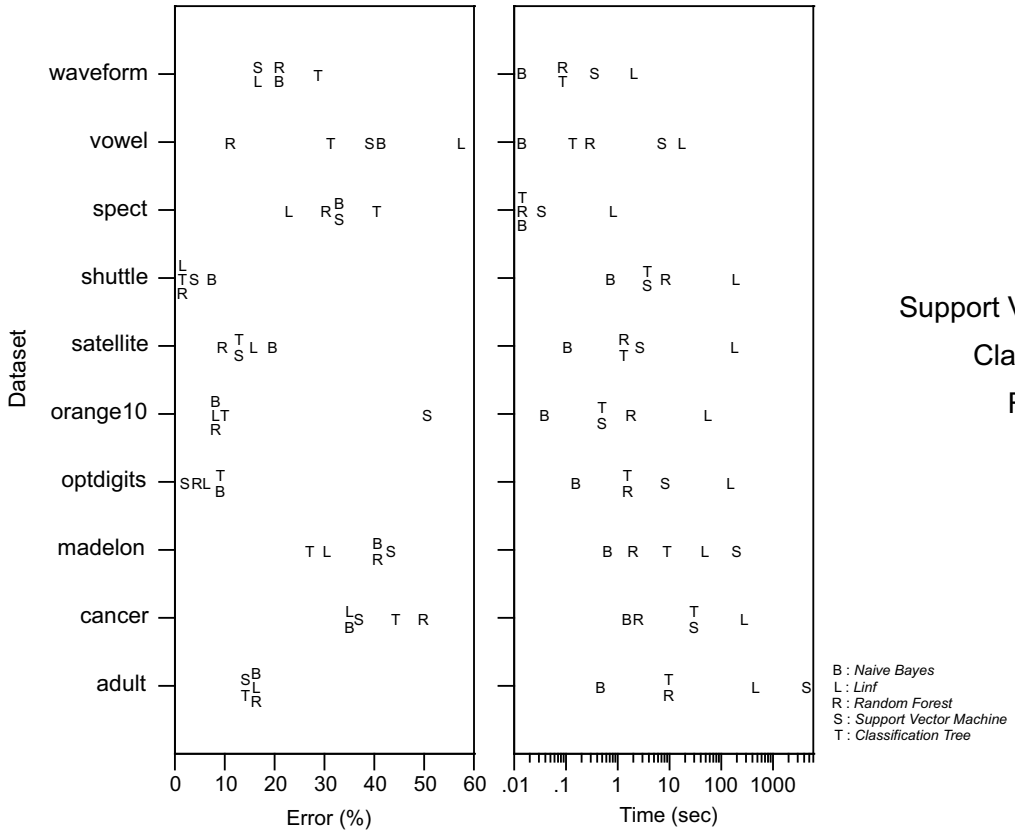
The Linf Algorithm

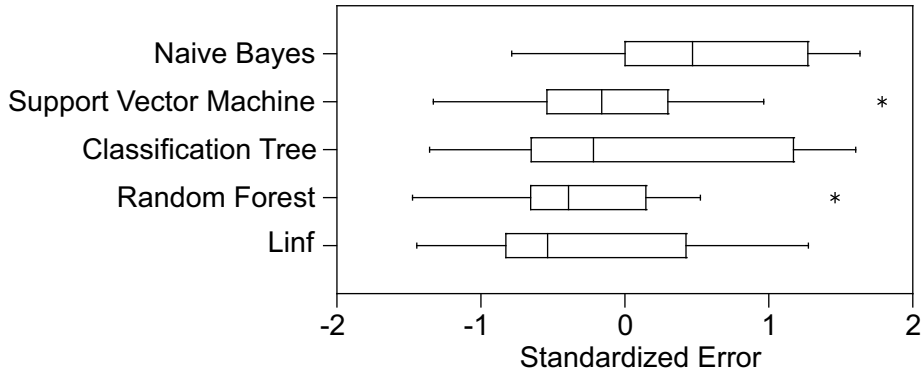
1. Transform
   - $t(x) = sgn(x)sqrt(abs(x))$
2. Project
   - Pick next target class (one against all).
   - Compute 25 random, integer-weighted, 2D projections.
   - Pick best 5 projections on a class-separation measure.
3. Bin
   - $b = 2log_2(n)$
   - Compute purity of target class instances in each bin.
4. Cover



5. Repeat
   - Store best cover in scoring list.
   - Repeat steps 2-4 until no training instances left.

Results

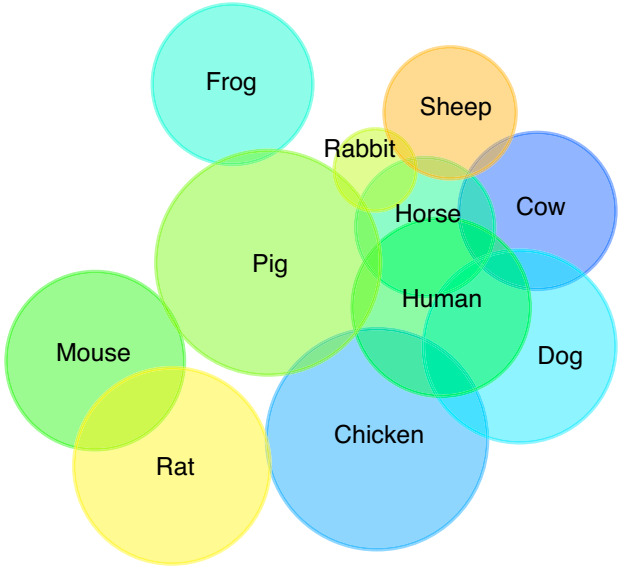

waveform
vowel
spect
shuttle
satellite
orange10
optdigits
madelon
cancer
adult
Dataset
0 10 20 30 40 50 60
Error (%)
.01 .1 1 10 100 1000
Time (sec)
B : Naive Bayes
L : Linf
R : Random Forest
S : Support Vector Machine
T : Classification Tree


Naive Bayes
Support Vector Machine
Classification Tree
Random Forest
Linf
-2 -1 0 1 2
Standardized Error

Venn/Euler Diagrams

- How do we fit a Venn/Euler diagram to empirical data involving more than 3 sets? These diagrams are widely used in the bioinformatics community. We have developed an R function called venneuler() and have posted it on CRAN for general use.

Euler diagram for 11 animals based on gene lists from the Agilent DNA oligo microarray database. The analysis was based on 247,412 gene symbols and 11 animal names. The stress for this solution is .06, with corresponding correlation of 0.97 between circle and intersection areas and counts of genes. The computation was completed in under 30 seconds on a 2.5 GHz MacBook Pro running the Java 1.5 Virtual Machine in 2GB of allocated memory.



SYSTAT

The `venneuler()` Algorithm

1. Make list of m intersections among n sets ($m = 2^n$)

$$P = [\emptyset, X_1, \ldots, X_n, X_1 \cap X_2, \ldots, X_1 \cap X_2 \cap \ldots X_n]$$

2. Make list of disk intersections (size disks proportional to $|X_i|$).

$$Q = [\emptyset, D_1, \ldots, D_n, D_1 \cap D_2, \ldots, D_1 \cap D_2 \cap \ldots D_n]$$

3. Make list of disjoint counts and list of disjoint areas.

$$\mathbf{c} = (|P_1^-|, \ldots, |P_m^-|)$$
$$\mathbf{a} = Area(Q_1^-, \ldots, Q_m^-)$$

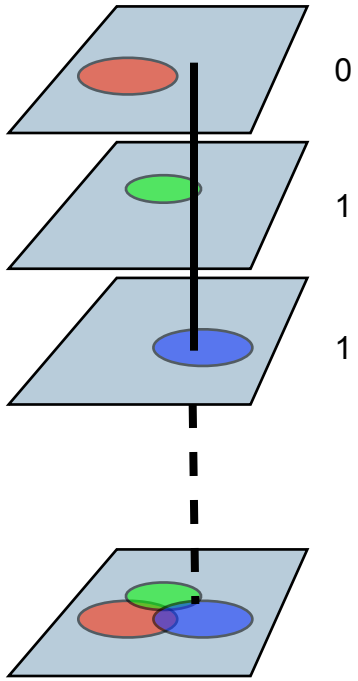
4. Estimate model.

$$\mathbf{a} = \beta\mathbf{c} + \boldsymbol{\varepsilon}$$
$$\hat{\beta} = \mathbf{a}'\mathbf{c}/\mathbf{c}'\mathbf{c}$$

5. Move disks and repeat 2-5 to minimize error, using gradient:

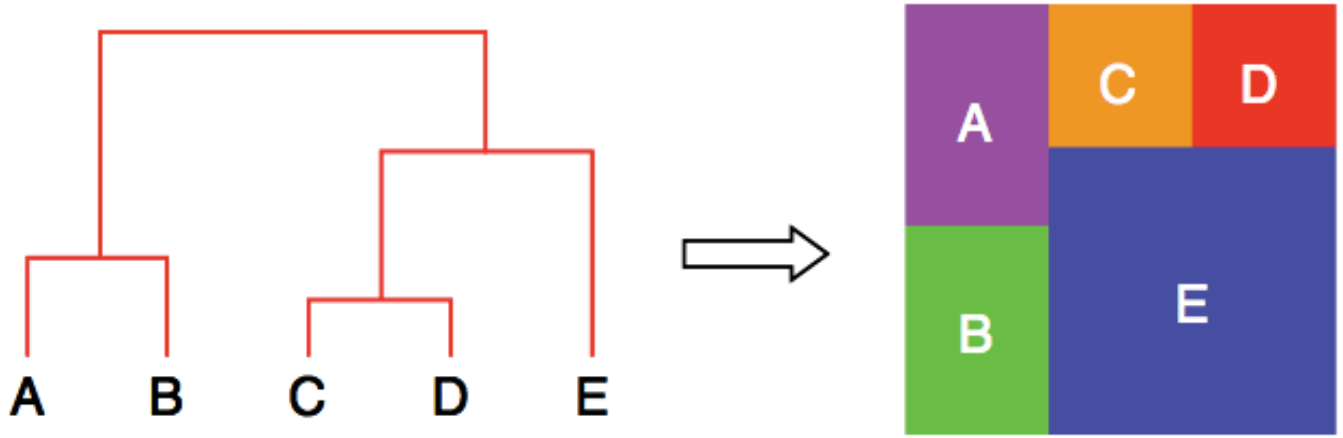
$$\nabla F(x,y)_i = \sum_{k=1}^{m} \sum_{i \neq j} \{(x_i - x_j)(a_k - \hat{a}_k), (y_i - y_j)(a_k - \hat{a}_k)\}$$

Calculating Areas

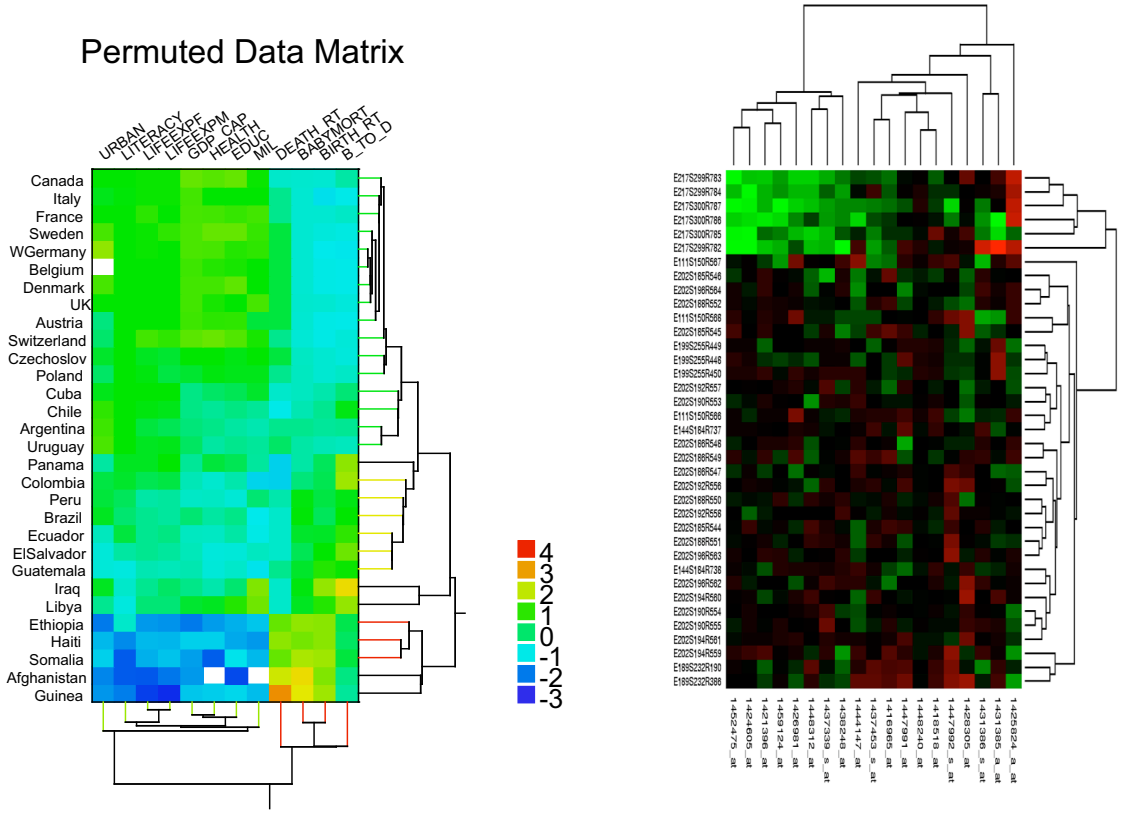


Treemaps

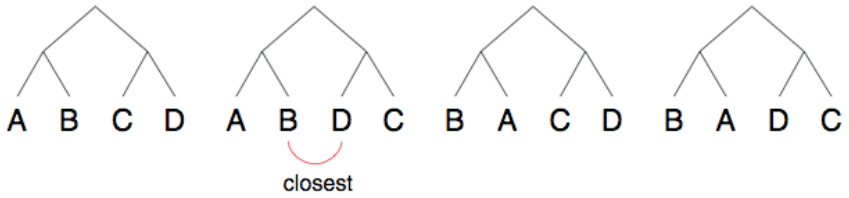
- We are investigating the ordering of treemaps.



Seriating a Tree



Wilkinson and Friendly, *TAS*, 2009



Fisher-Anderson Iris Data