National Science Foundation
WHERE DISCOVERIES BEGIN

Homeland Security

FODAVA Partners

Leland Wilkinson (SYSTAT, Advise Analytics & UIC)

Anushka Anand, Tuan Nhon Dang (UIC)

Anomaly Discovery through Visual Characterizations of Point Sets Embedded in High-Dimensional Geometric Spaces

Georgia Tech, December 8-9, 2011

Research Goals

- Visual Anomaly Detection
- Applications
  - Threat detection.
  - Interactive visual analytics based on VMBT (Visual-Model-Based Transformations).
  - Model diagnosis.

FODAVA Products

- Scagnostics Explorer
- Autovis
- Time Seer
- Visual Classifier
- CHIRP Classifier
- Anomaly Detector

Visual-Model-Based Transformations

- Compute Transformation $X \longrightarrow S$
- Analyze Patterns in S
- Invert Transform to X

$$\begin{tikzcd}
\{P_i\} \arrow[r, "VMBT()"] & S \subset \mathbb{R}^q \arrow[d, "visualize()"] \\
X \subset \mathbb{R}^p \arrow[u, "project()"] \arrow[r, dotted] & \{V_j\}
\end{tikzcd}$$

Scagnostics

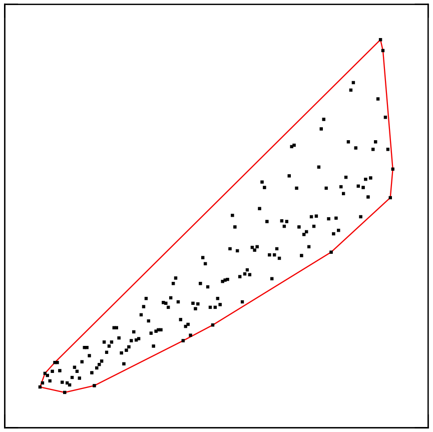
- Every visualization depends on a model (even EDA).
- Scagnostics (Scatterplot Diagnostics) is a Tukey (John and Paul) idea that offers such a model. Scagnostics help us to characterize 2D scatterplots (lots of them).
- VMBTs are a generalization of scagnostics.
- We do visual model-based transformations to see signals that are not picked up by classical statistical or data-mining methods.

Scagnostics

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

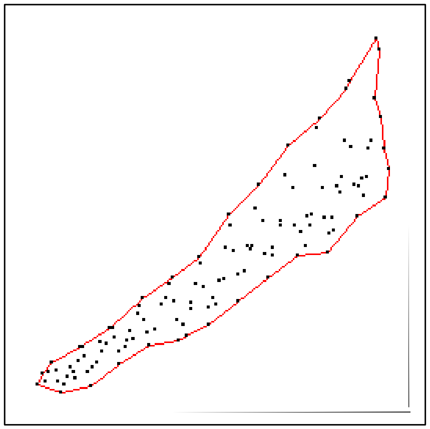
Scagnostics

Convex Hull



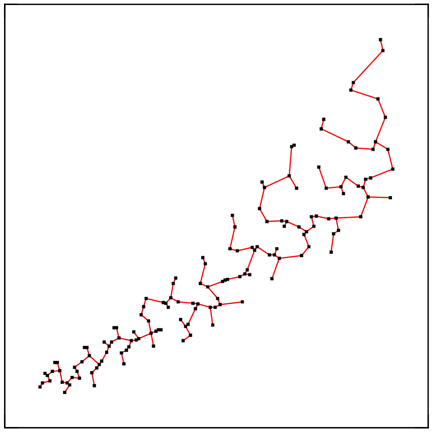
Scagnostics

Alpha Shape



Scagnostics

Minimum Spanning Tree

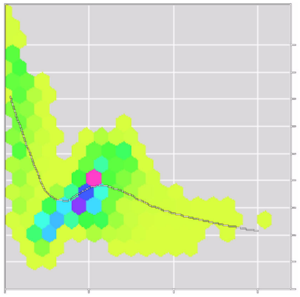


Scagnostics

- Bin
- Delete Outliers
- Compute Measures
  - Shape
  - Trend
  - Density

Scagnostics

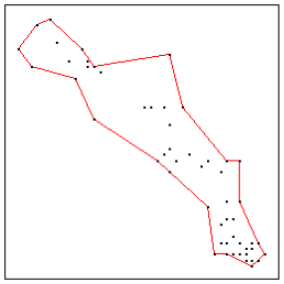
- We bin on a 40x40 hexagon grid.
- Until there are fewer than 250 nonempty cells, we recursively enlarge the bin size and re-bin.

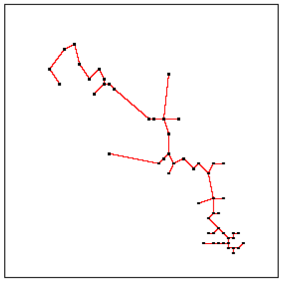


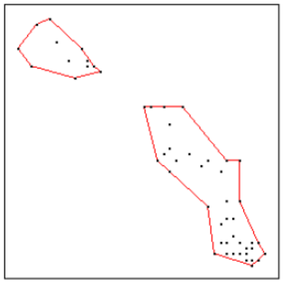
A 20 x 20 hex grid
on weather data

Scagnostics

- Peel MST using distribution of edge lengths.
- An outlier is MST vertex whose adjacent edges all have a large weight.
- We use a statistical test to identify large weights.

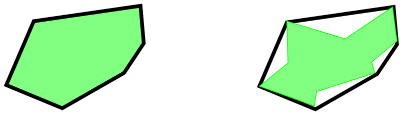




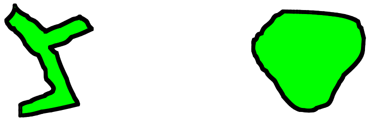


Scagnostics

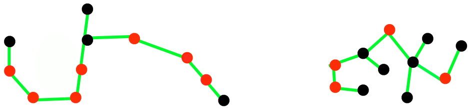
Convex: area of alpha shape divided by area of convex hull



Skinny: ratio of perimeter to area of the alpha shape

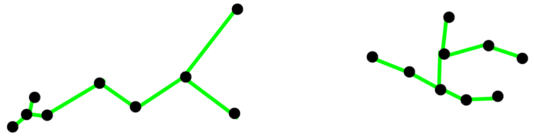


Stringy: ratio of 2-degree vertices in MST to number of vertices > 1-degree

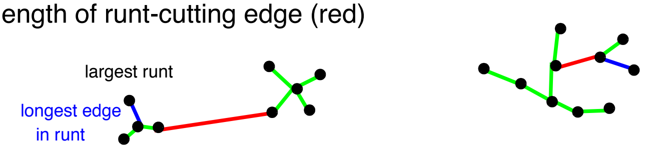


Scagnostics

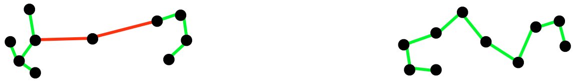
Skewed: ratio of $(Q90 - Q50) / (Q90 - Q10)$, where quantiles are on MST edge lengths



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

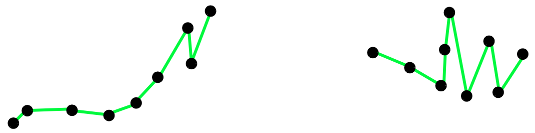


Outlying: proportion of total MST length due to edges adjacent to outliers

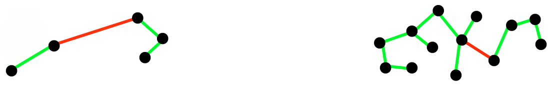


Scagnostics

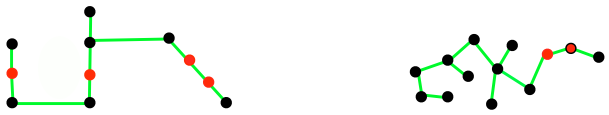
Monotonic: squared Spearman correlation coefficient



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



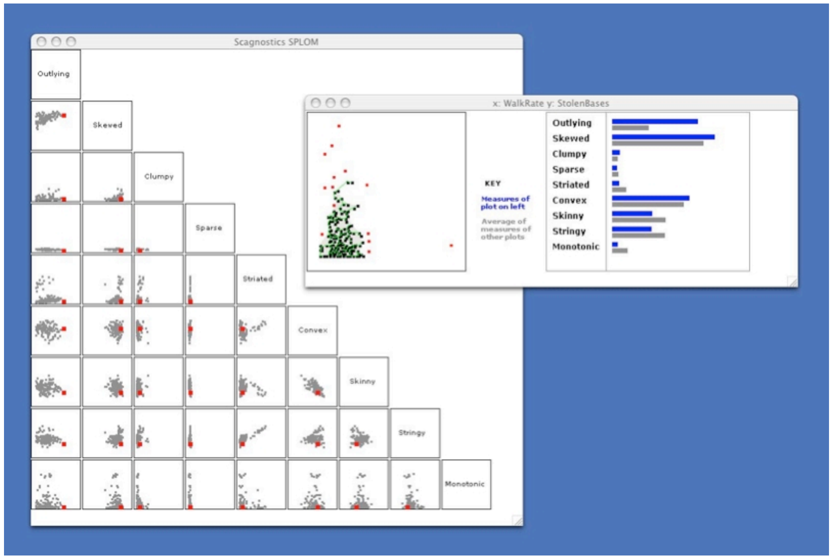
Striated: proportion of all vertices in the MST that are degree-2 and have a cosine between adjacent edges less than -.75



Scagnostics Explorer

- Scatterplot matrix display
- Brushing
- Linking
- Anomaly Detection

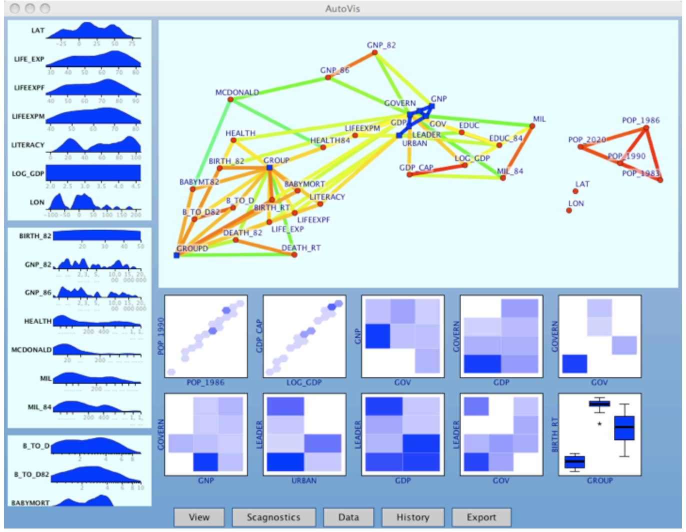
Scagnostics Explorer



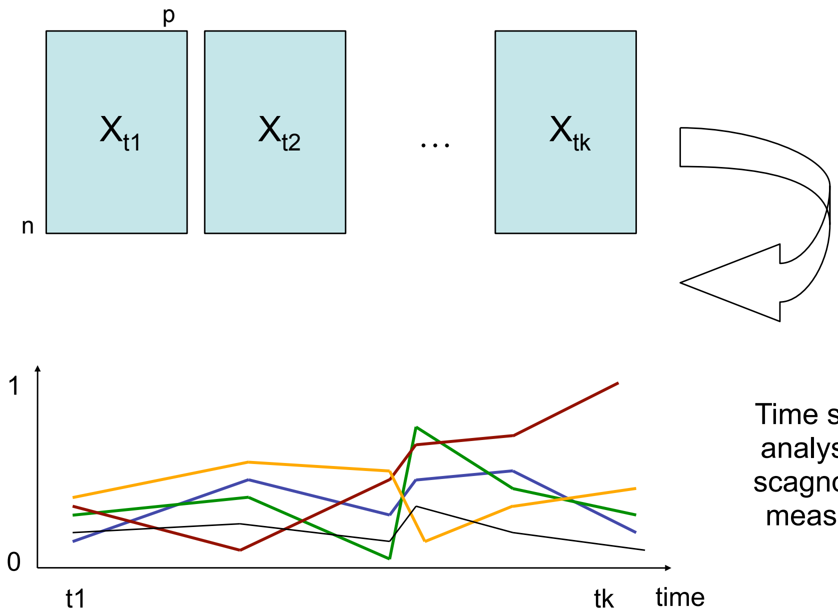
AutoVis

- Modeling: Grammar of Graphics
- Discovery: Scagnostics
- Filtering: Scagnostics Distributions
- Protection: False Discovery Rate

AutoVis

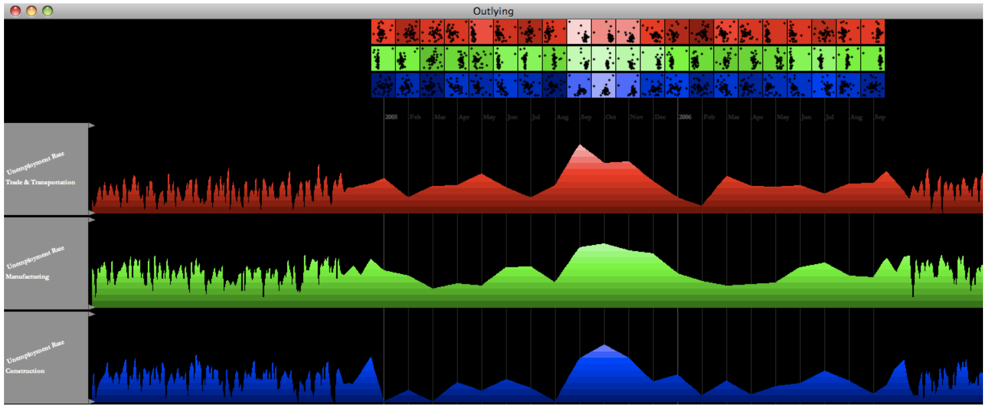

AutoVis
LAT
LIFE_EXP
LIFEEXPF
LIFEEXPM
LITERACY
LOG_GDP
LON
BIRTH_82
GNP_82
GNP_86
HEALTH
MCDONALD
MIL
MIL_84
B_TO_D
B_TO_D82
BABYMORT
GNP_82
GNP_86
MCDONALD
GOVERN
GNP
GDP
GOV
EDUC
MIL
LIFEEXPM
HEALTH
HEALTH84
LEADER
URBAN
EDUC_84
POP_1986
POP_2020
POP_1990
POP_1983
BIRTH_82
GROUP
LOG_GDP
GDP_CAP
MIL_84
BABYMT82
BABYMORT
LITERACY
B_TO_D
BIRTH_RT
B_TO_D82
LIFEEXPF
LIFE_EXP
DEATH_82
LAT
LON
GROUPD
DEATH_RT
POP_1990
POP_1986
GDP_CAP
LOG_GDP
GNP
GOV
GOVERN
GDP
GOVERN
GOV
GOVERN
GNP
LEADER
URBAN
LEADER
GDP
LEADER
GOV
BIRTH_RT
GROUP
View
Scagnostics
Data
History
Export

TimeSeer



Time series analysis of scagnostics measures

TimeSeer

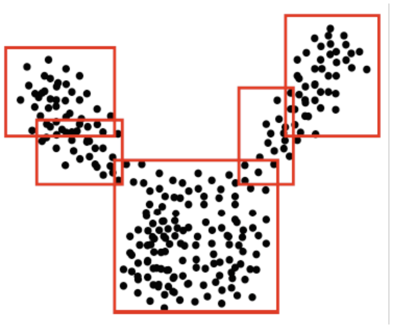

Outlying
2005
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2006
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Unemployment Rate
Trade & Transportation
Unemployment Rate
Manufacturing
Unemployment Rate
Construction

Visual Classifier

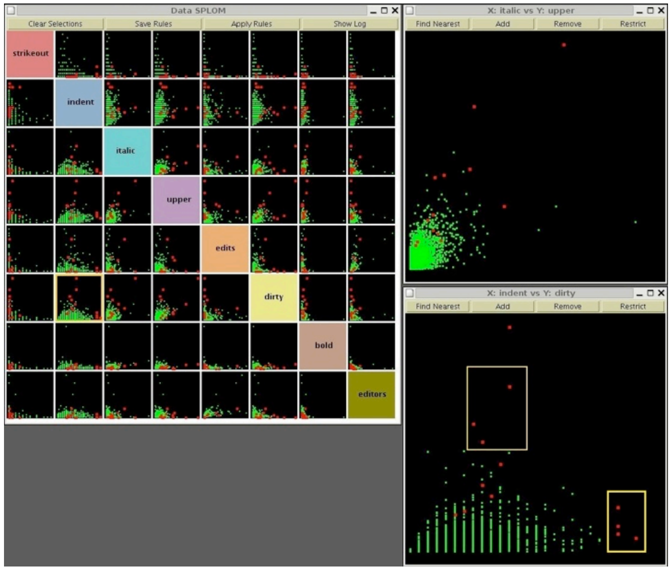
- Visually identify structure, formally define it so we can query unseen data for similar structure
- We use the union of open hypercubes to define the L^{∞} norm topology
  - Composite Hyper-rectangular Description Regions (CHDRs) – capture large-scale structure
  - 3-operator algebra on CHDRs – add, remove, restrict
  - Generate set-wise rules using gestures in the exploratory GUI
  - Log the rules and apply them to a test set

Visual Classifier

- Simple specification of neighborhoods
  - Visual brushing operations are translated into rules built from basic algebra on intervals
- Simple expressions to specify complex geometric objects – union of CHDRs



Visual Classifier



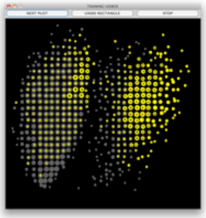
Visual Classifier

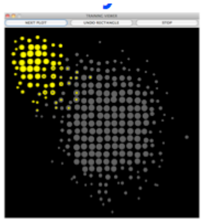
- Axis-parallel projections not effective for discovering multivariate structure. So, ...
- Compute list of {-1, 0, 1}-weighted random projections.
- Rank them on separation measure S (a nearest-centroid distance measure).
- Present these to user in GUI.

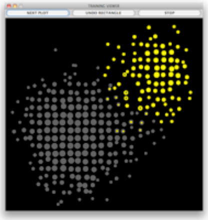
Achlioptas, D., “Database-friendly random projections.” In Proc. of ACM SIGMOD Symposium on Principles of Database Systems, 2001, pp. 274–281.
Li, P., Hastie, T. J., and Church, K. W. “Very sparse random projections.” In Proc. of ACM Conference on Knowledge Discovery and Data Mining, 2006, pp. 287–296.

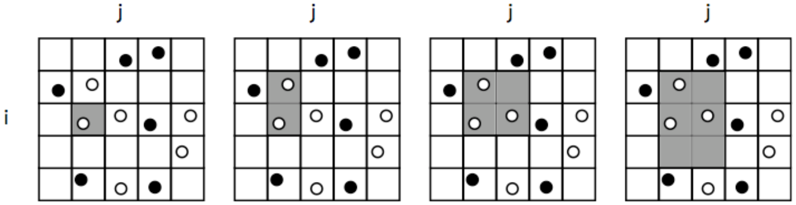
CHIRP

- Store bins in byte array.
- Automatic covering algorithm.

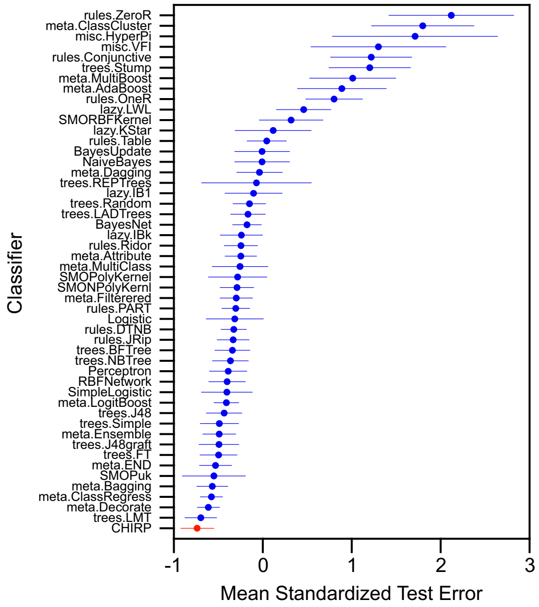








CHIRP

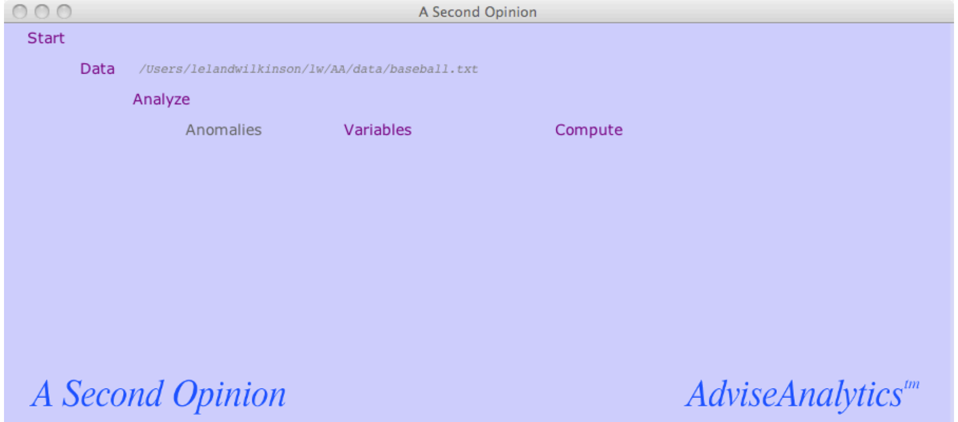


Wilkinson, Anand, Dang. KDD (2011)

Anomaly Detector

- New outlier detection methods
  - Use upper tail of gap distribution instead of σ
  - MST
  - k-means outliers
- Find outliers in VMBT space.
- Display results in data space.

Anomaly Detector



/Users/lelandwilkinson/lw/AA/data/baseball.txt

Future work

- High-dimensional (multivariate) time-series scagnostics.
- Scagnostics on projections (scalable Projection Pursuit).

Thank you

Leland Wilkinson
UIC
Systat Software Inc.
Advise Analytics Inc.
leland.wilkinson@gmail.com
http://www.cs.uic.edu/~wilkinson/

Anushka Anand
UIC CS
aanand2@uic.edu

Tuan Nhon Dang
UIC EVL
tdang@cs.uic.edu