National Science Foundation
WHERE DISCOVERIES BEGIN

Homeland Security

FODAVA Partners

Leland Wilkinson (SYSTAT & UIC)
Robert Grossman (UIC)
Adilson Motter (Northwestern)
Anushka Anand, Troy Hernandez (UIC)

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

Research Goals

- Visual-Model-Based Transformations
- Applications
  - Automated visualization based on VMBT.
  - Interactive visual analytics based on VMBT.

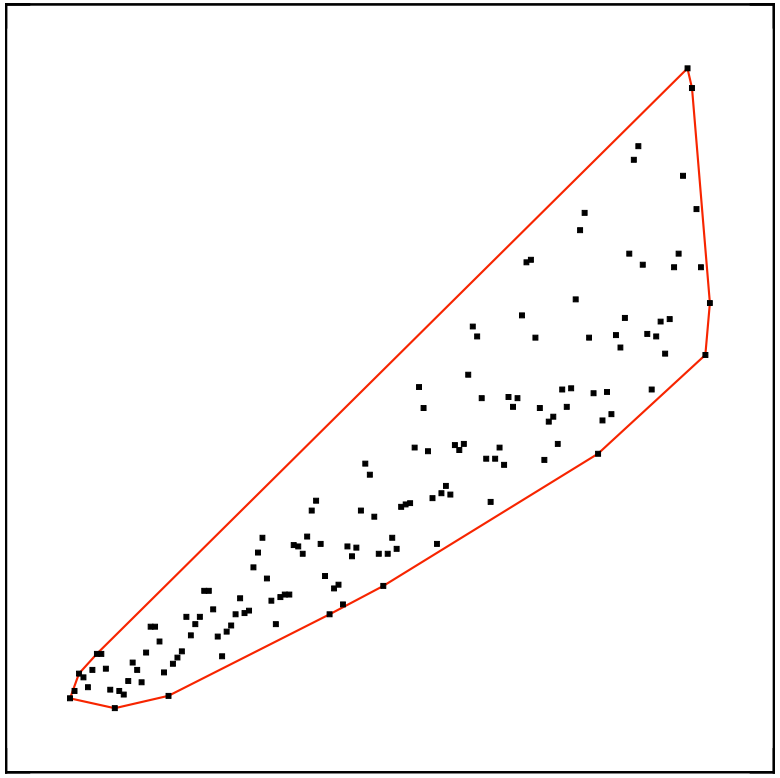
Background

- 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).

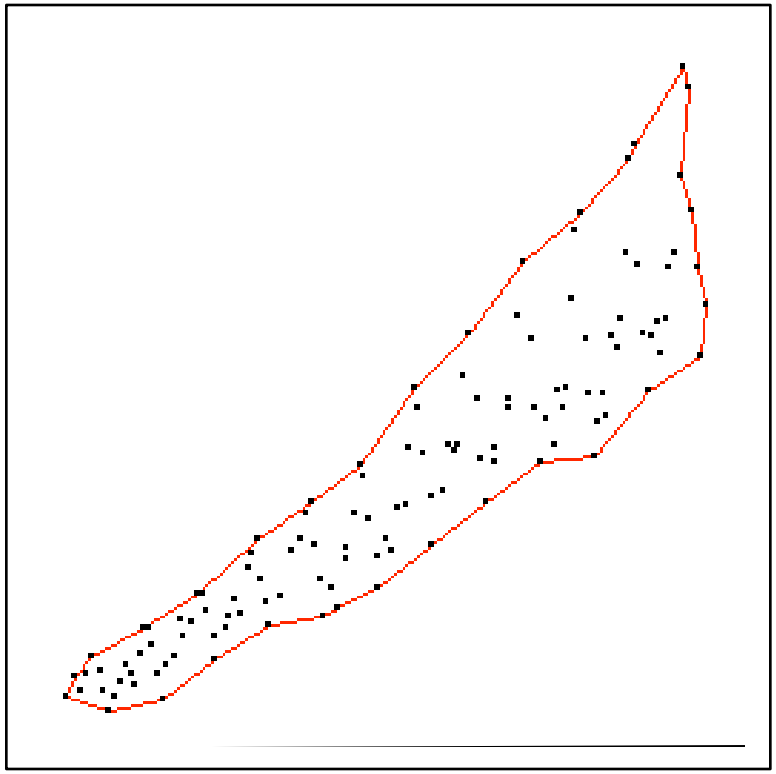
Scagnostics

- 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

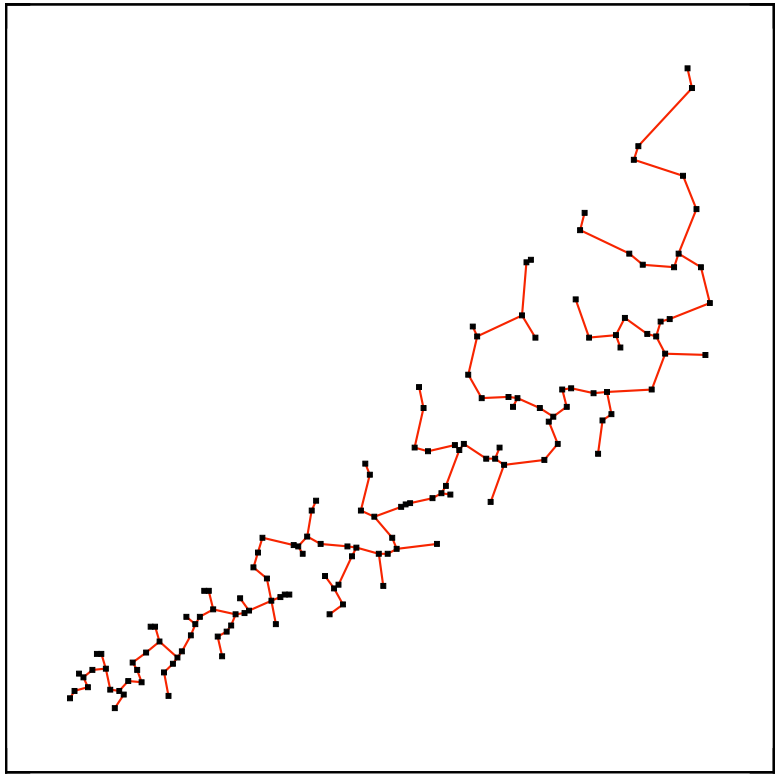
Convex Hull



Alpha Shape



Minimum Spanning Tree

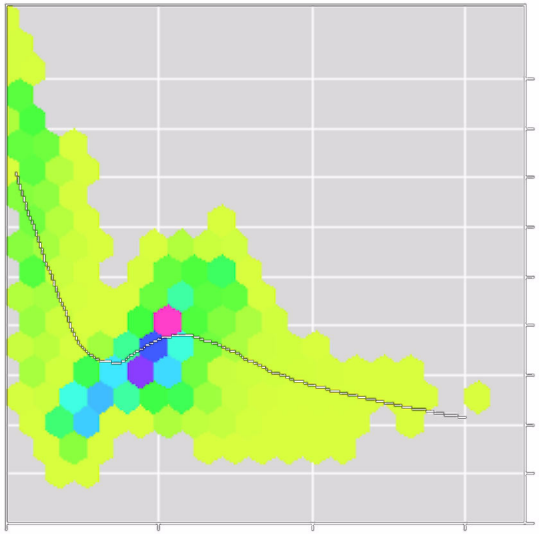


Computing

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

Hexagon Binning

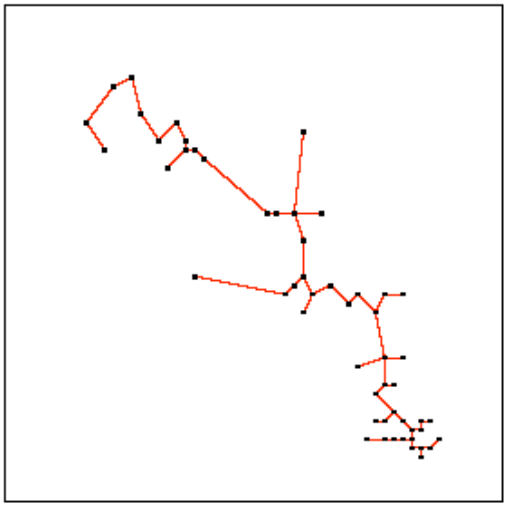
- We bin on a 40x40 hexagon grid.
- Until there are fewer than 250 nonempty cells, we recursively enlarge the bin size and re-bin.



A 20 x 20 hex grid
on weather data

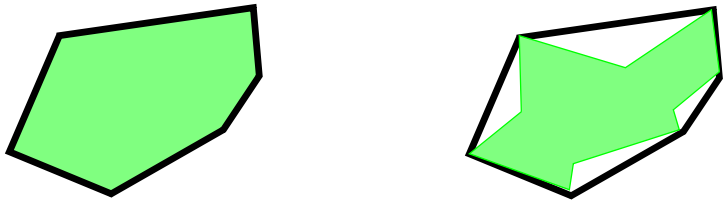
Delete Outliers

- 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.



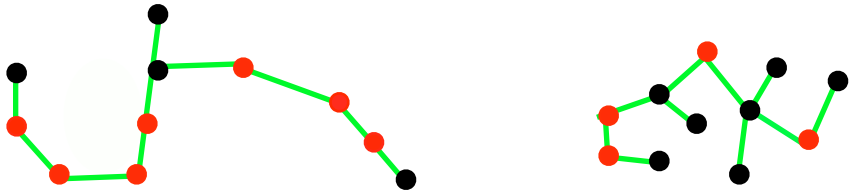
Shape

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



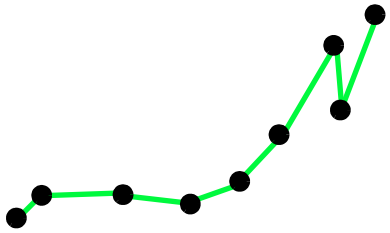
Skinny: ratio of perimeter to area of the alpha shape

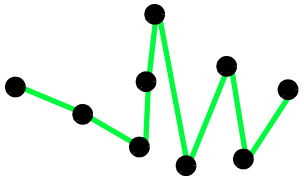
Stringy: ratio of 2-degree vertices in MST to number of vertices > 1-degree



Trend

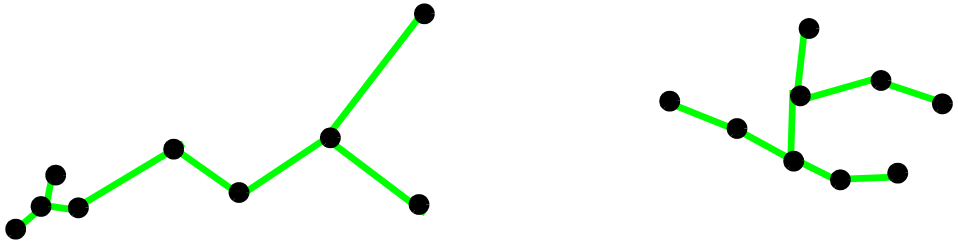
Monotonic: squared Spearman correlation coefficient



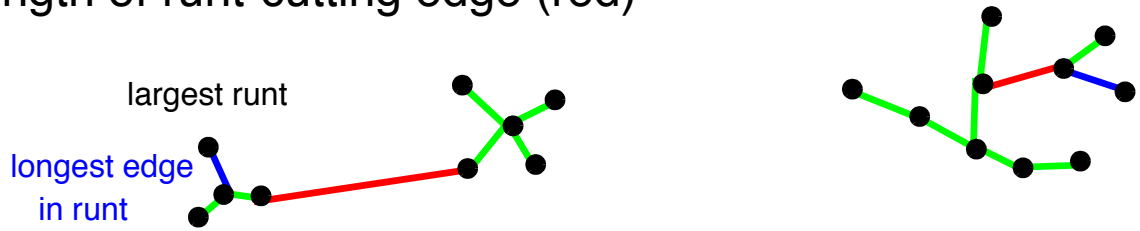


Density

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

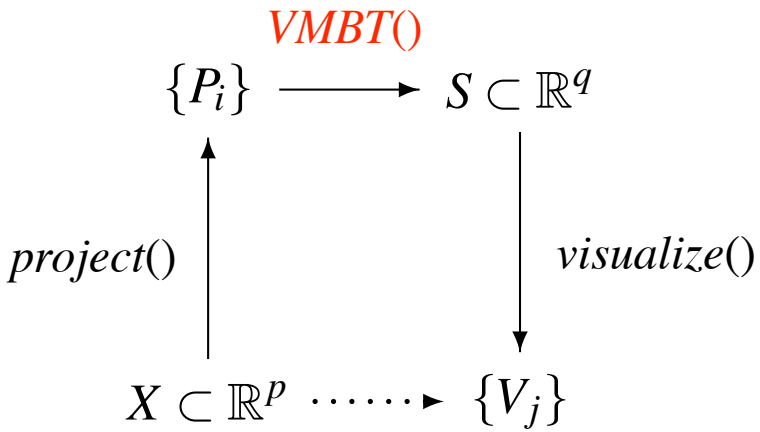
Density

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

Visual-Model-Based Transformations

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



VMBT Applications

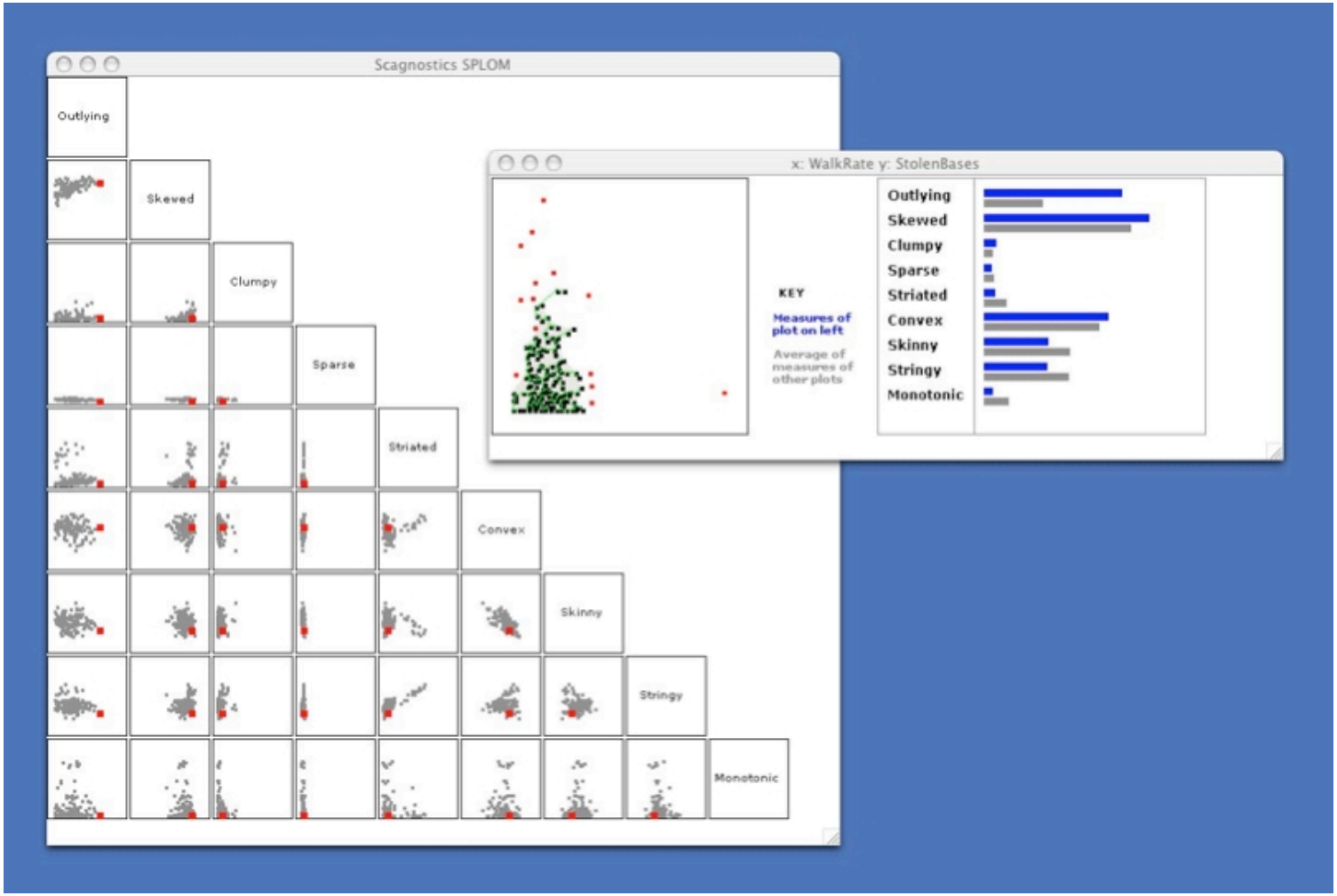
- Scagnostics Explorer
- AutoVis (Automated Visualization)
- Time/Space Explorer
- Visual Classifier

Scagnostics Explorer

- Scatterplot matrix display
- Brushing
- Linking
- Anomaly Detection

Scagnostics Explorer

(Wilkinson and Anand)

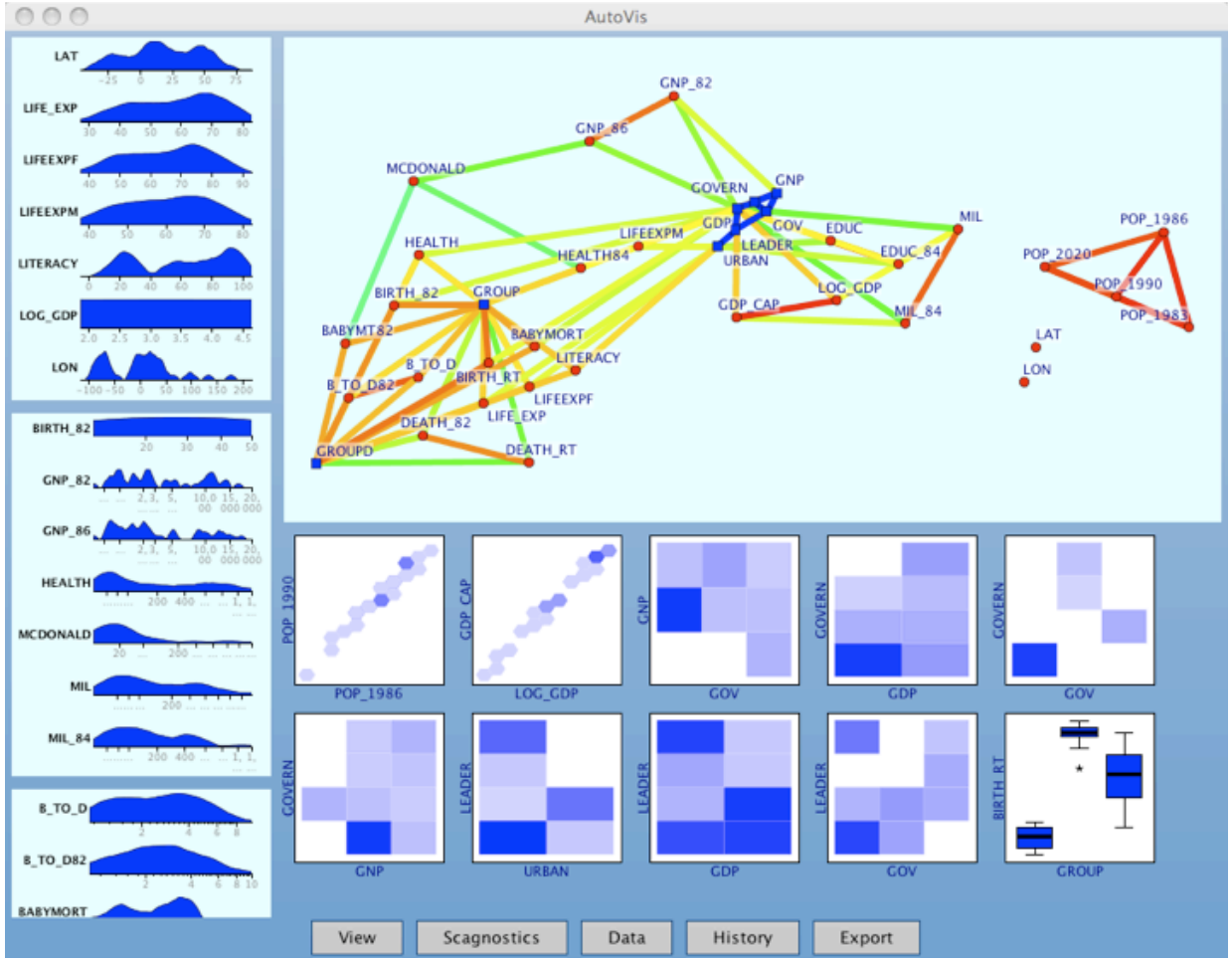


AutoVis

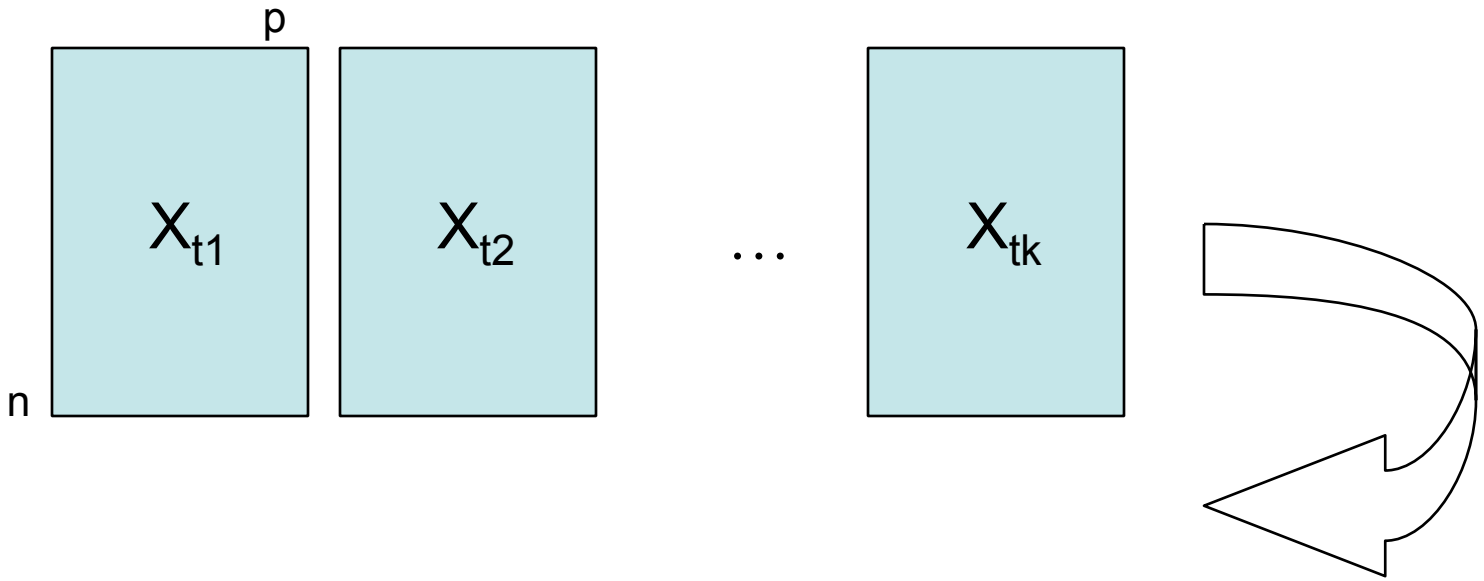
- Modeling: Grammar of Graphics
- Discovery: Scagnostics
- Filtering: Scagnostics Distributions
- Protection: False Discovery Rate

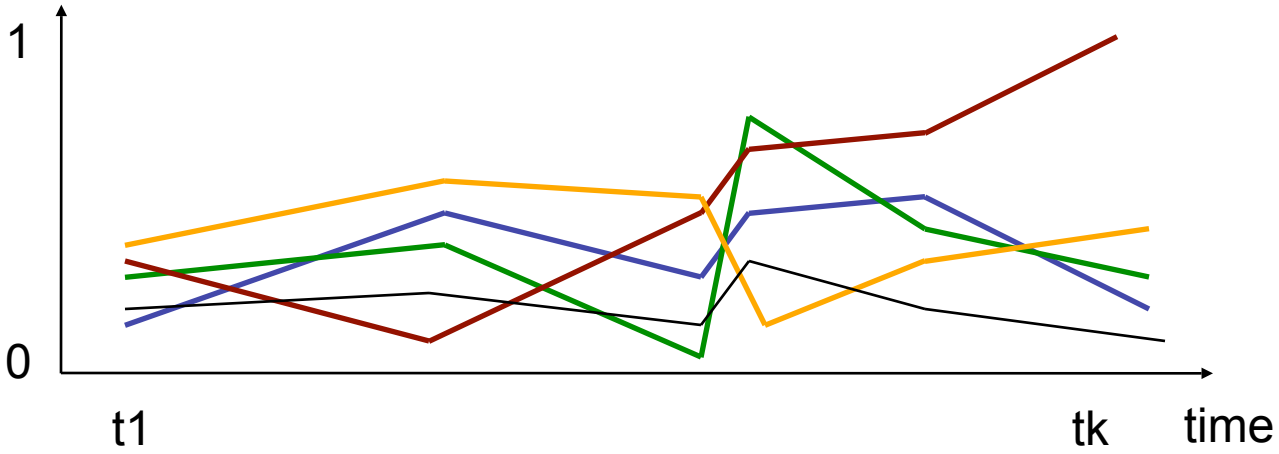
AutoVis Software

(Wills and Wilkinson)



Time/Space Explorer

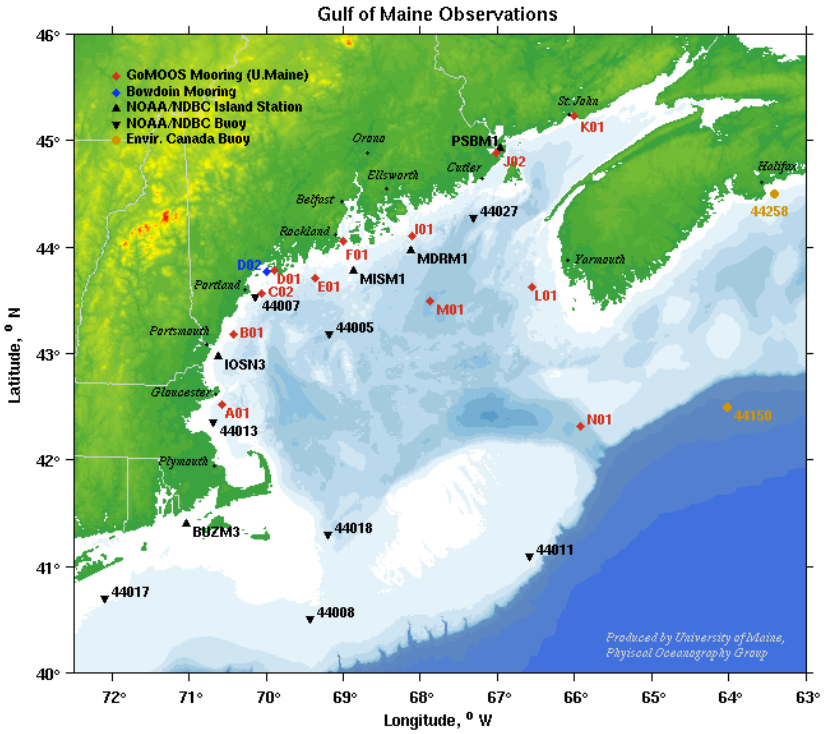




Time series analysis of scagnostics measures

Spatial-temporal data

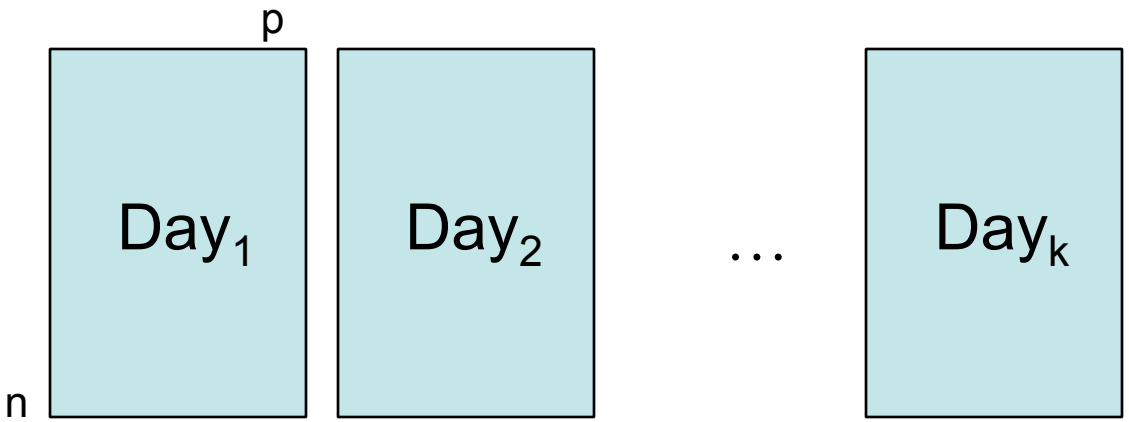
- GoMOOS Moored Buoy Program
  - 12 buoys in the Gulf of Maine
  - Historical data
  - Sensor readings
    - Pressure
    - Temperature
    - Wind speed
    - Visibility



(http://gyre.umeoce.maine.edu/buoyhome.php/)

Spatial-temporal pattern discovery

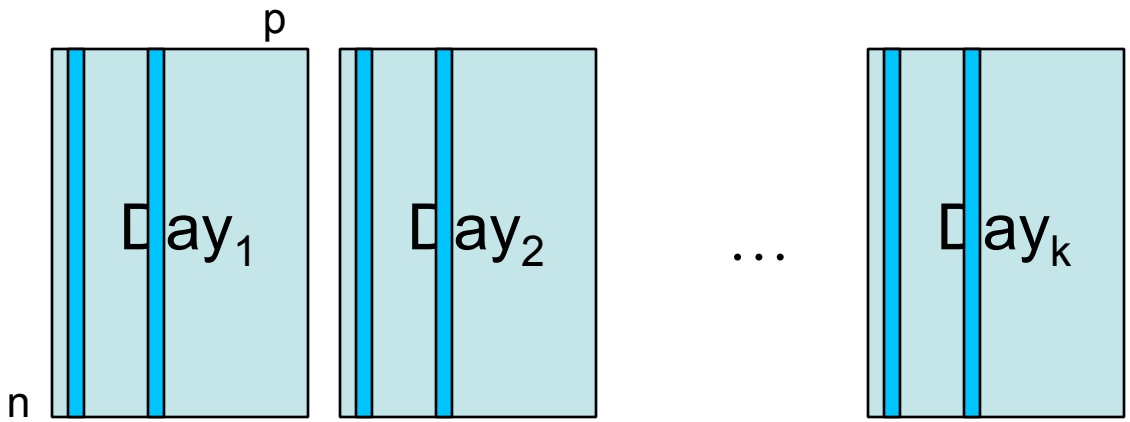
- Local behavior – for each buoy



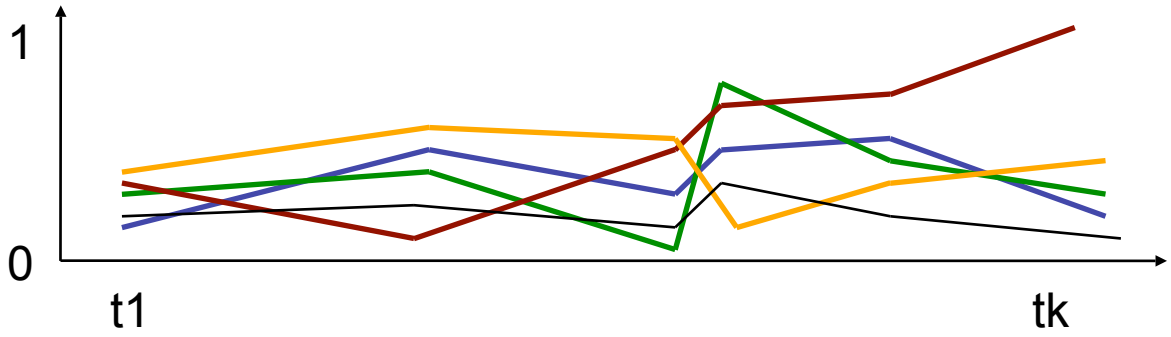
Each bivariate distribution is characterized by the scagnostics

Spatial-temporal pattern discovery

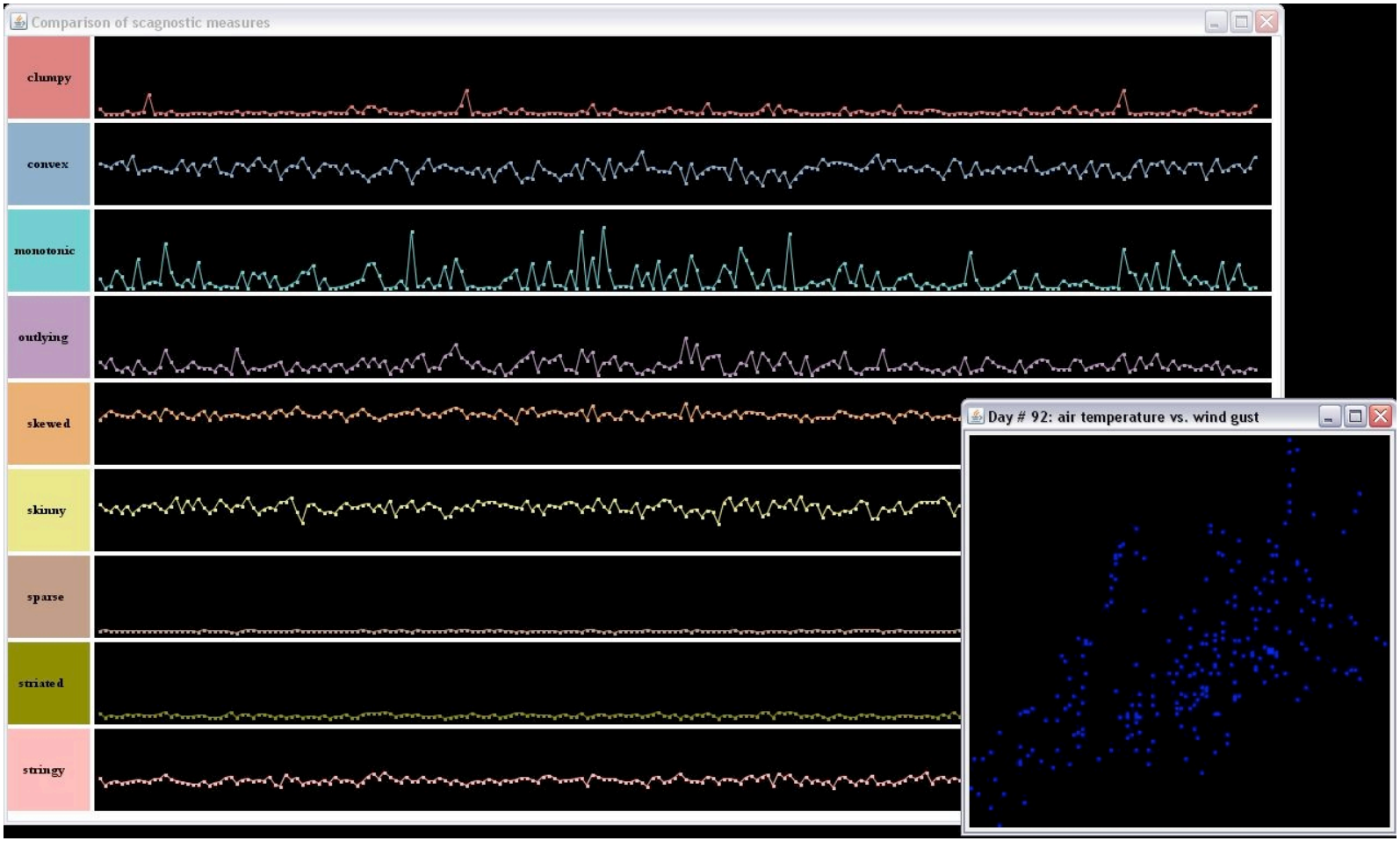
- Local behavior – for each buoy



Each bivariate distribution is characterized by the scagnostics



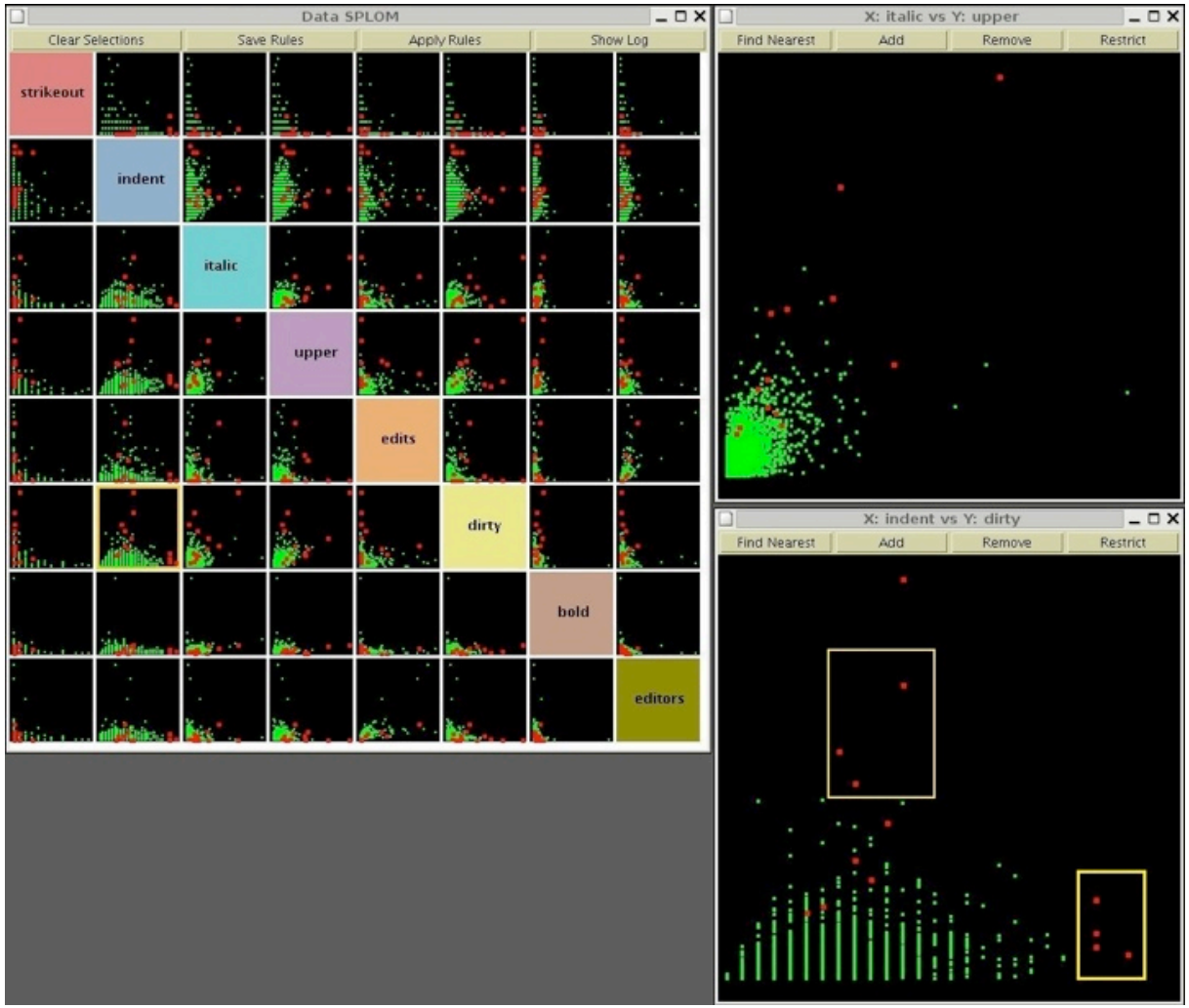
Scagnostics time series for air temperature(C) vs. wind gust(m/s)



Visual Classifier

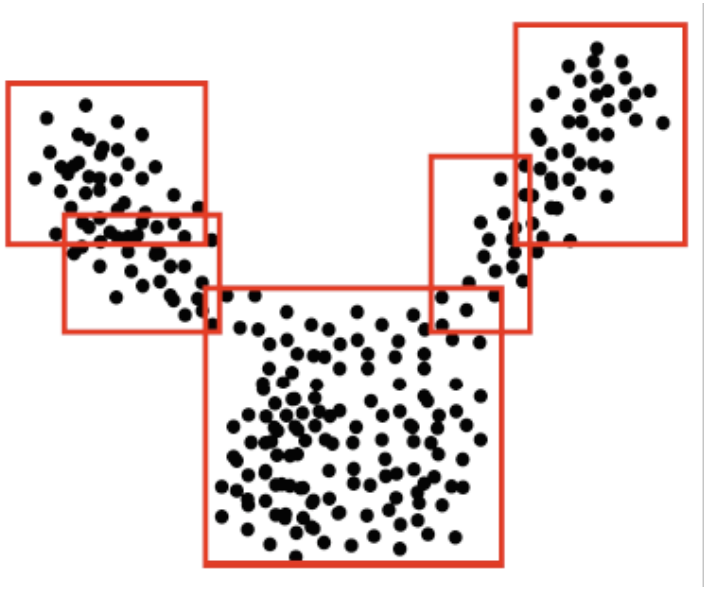
- Visually identify structure, formally define it so we can query unseen data for similar structure
- We use the union of open hypercubes to define the L^∞ norm topology
  - Composite Hyper-rectangular Description Regions (CHDRs) – capture large-scale structure
  - 3-operator algebra on CHDRs – add,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

Visual Classifier



Benefits

- 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



Challenges

- Can we build a classifier that is more than a black box – allows for insights into variables (unlike random forests, neural nets, SVMs, etc.)?
- How far can we go with the human eye – can we compete with best classifiers out there?
- How trained does a user have to be?
- How to select a “good” subset of dimensions to view when the number of dimensions is large?

How will this influence FODAVA?

- Exploration
  - what aspects of our data should we examine before we build a model?
- Anomaly Detection
  - can we detect more than just outliers?
- Guided Search
  - what is interesting to an analyst?

Plans for developing FODAVA?

- Proselytize visual models that are motivated by
  - vision (the capabilities of the visual system motivate our models)
  - data (real data motivate our models)