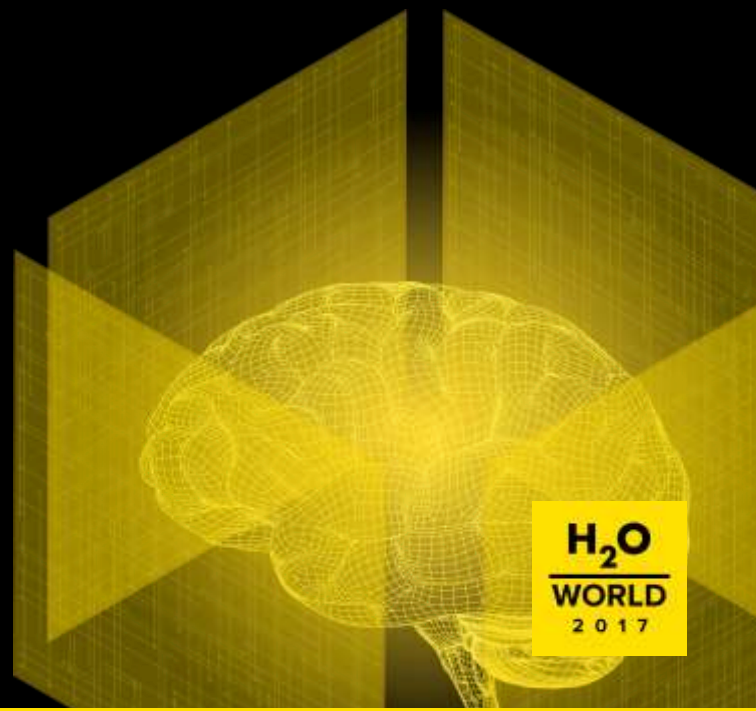
Automatic Visualization

Leland Wilkinson

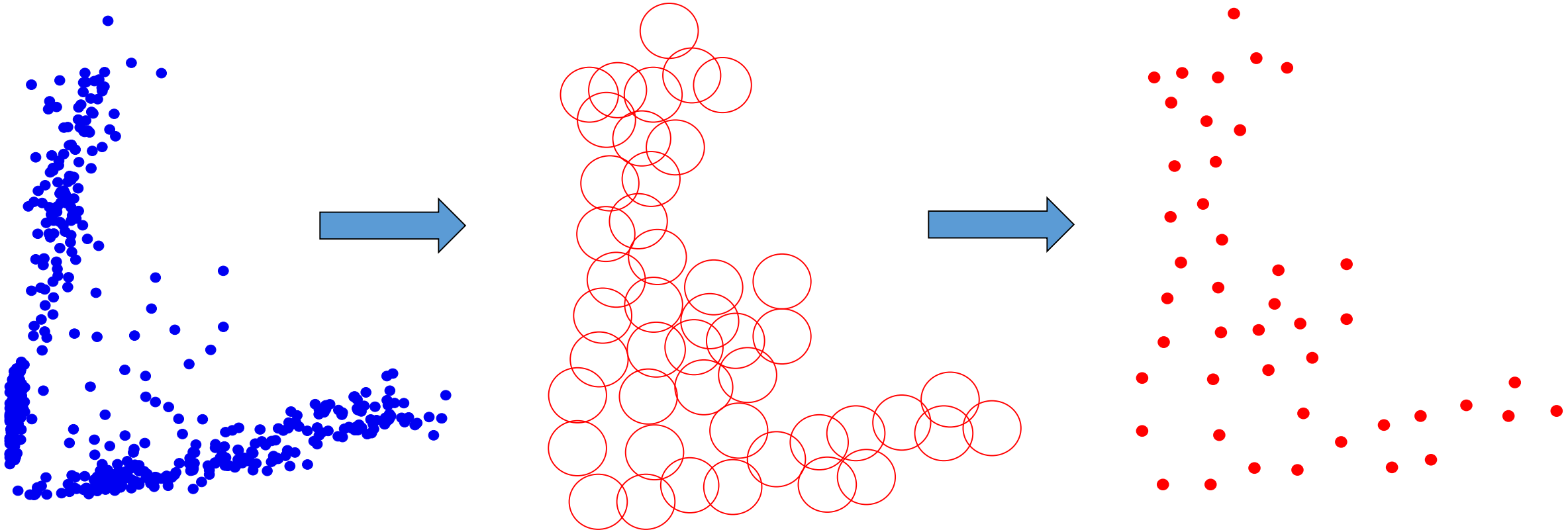
Chief Scientist, H2O
leland@h2o.ai
www.cs.uic.edu/~wilkinson



Visualizing Big Data

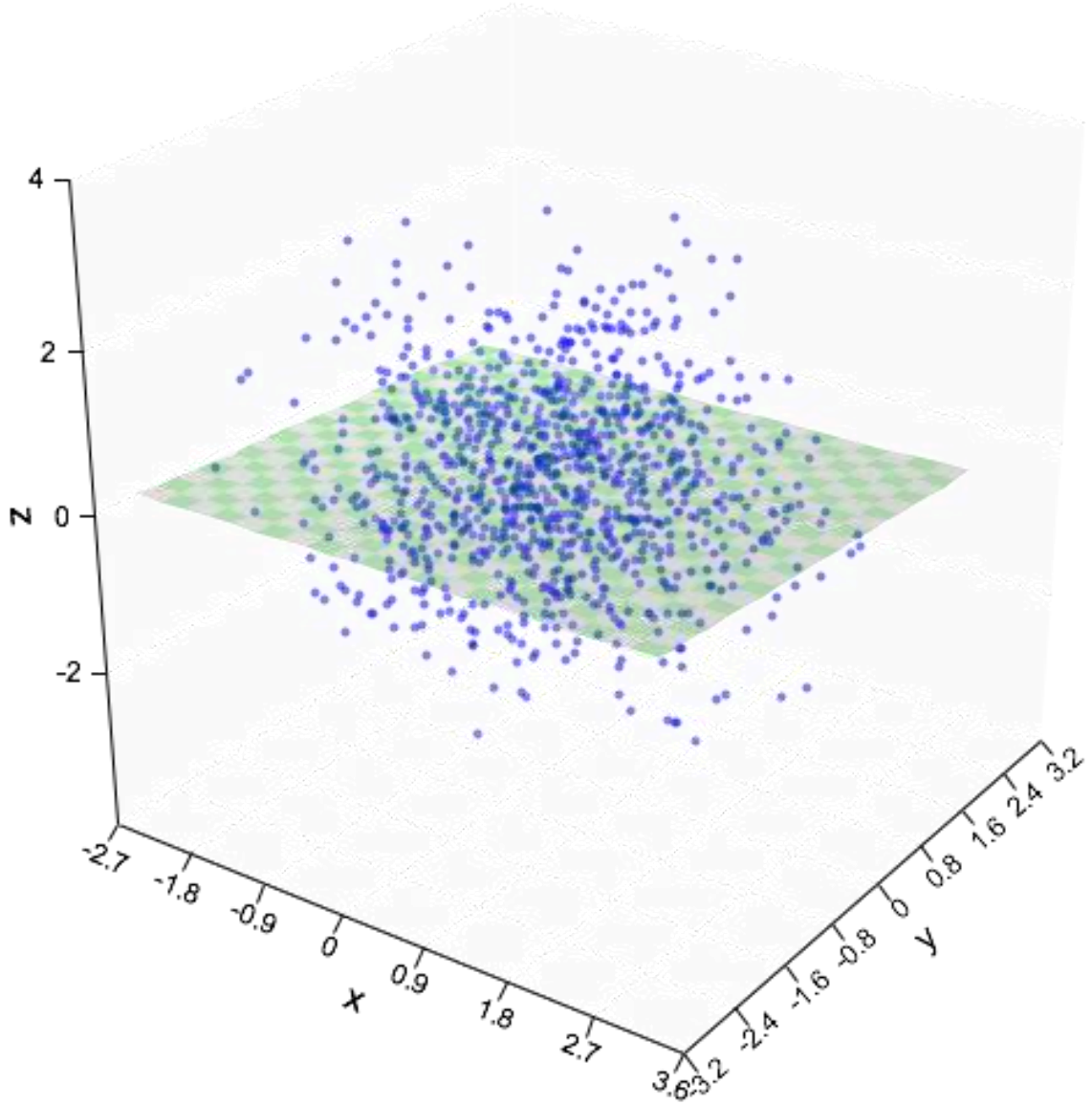
- Complexity: Many functions are polynomial or exponential
- Curse of Dimensionality: distances tend toward constant as $d \to \infty$
- Chokepoint: Cannot send big data over the wire
- Real Estate: Cannot plot big data on the client
- Cheesy solutions in 2D
  - Pixelate (too complex for higher dimensions)
  - Project (usually violates triangle inequality for $\mathbb{R}^d \mapsto \mathbb{R}^p, p \ll d$)
  - Image maps (OK for popups and simple links, not for EDA)
- Viable solutions
  - Aggregate (big n) to a few thousand rows
  - Project (big p) to a few dozen columns

Big Data

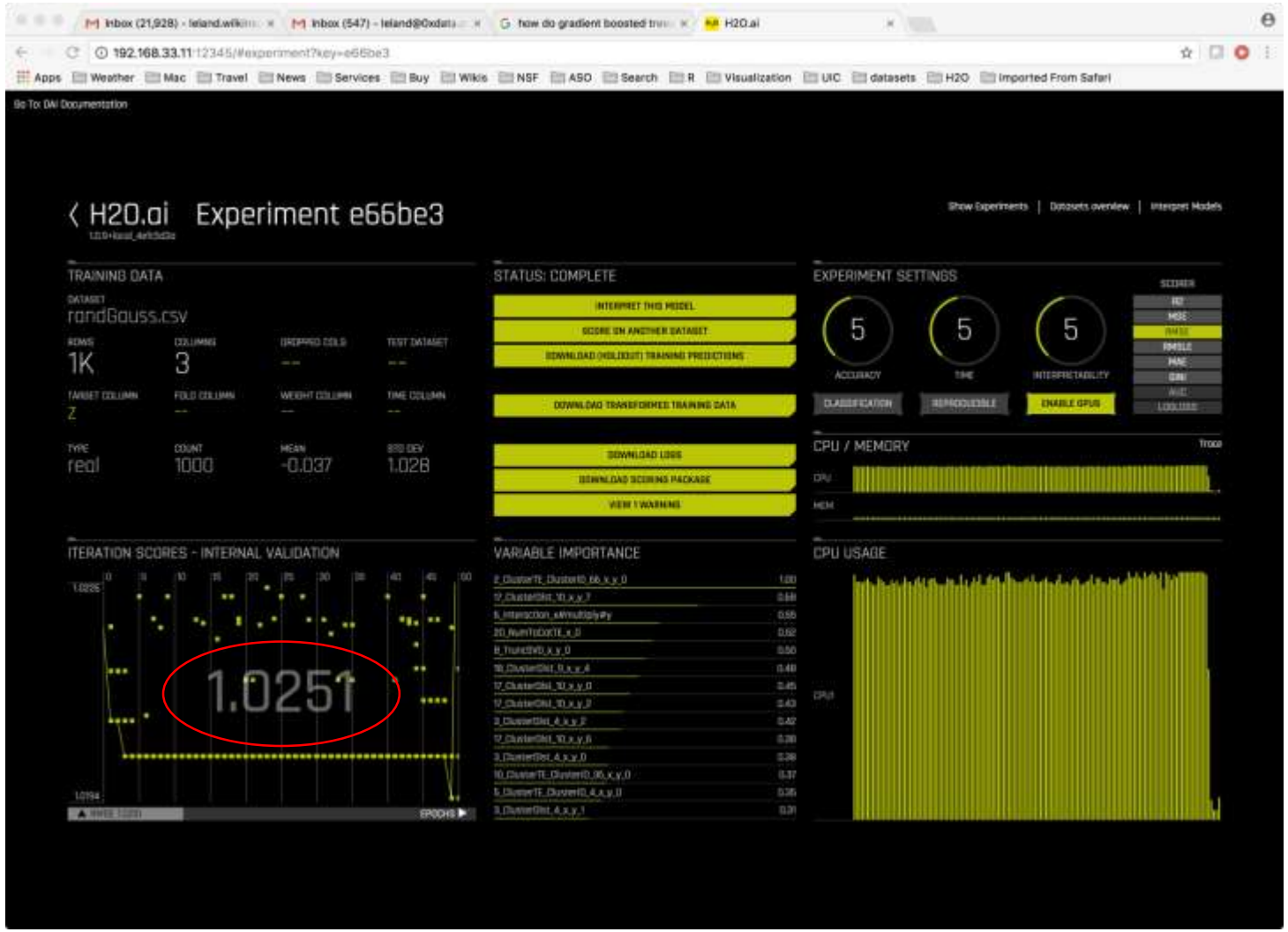


set cover (core sets)

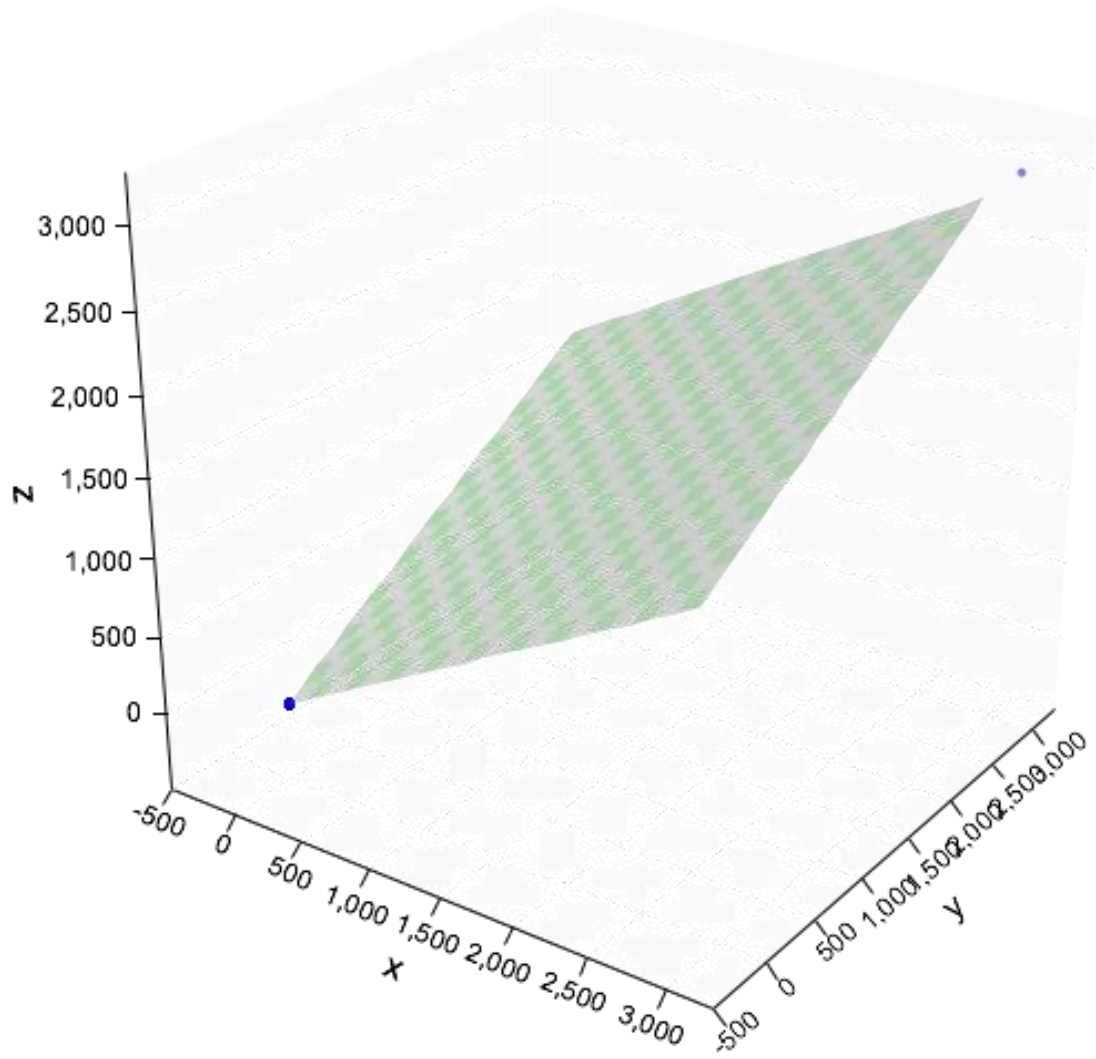
Outliers



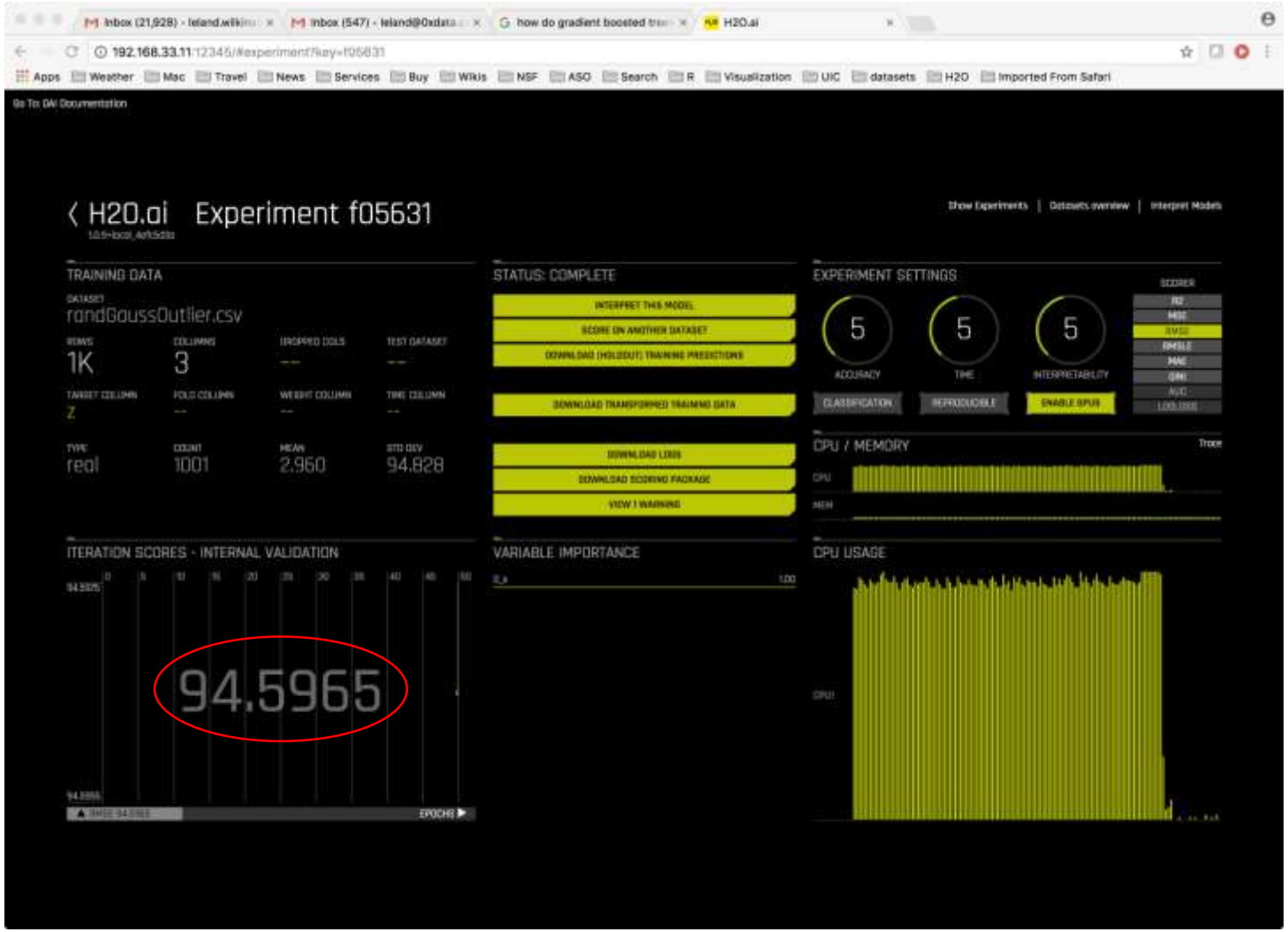
Outliers


H2O.ai
Experiment e66be3
TRAINING DATA
DATASET
randGauss.csv
ROWS
1K
COLUMNS
3
TYPE
real
COUNT
1000
MEAN
-0.037
STD DEV
1.028
STATUS: COMPLETE
EXPERIMENT SETTINGS
ACCURACY
TIME
INTERPRETABILITY
CPU / MEMORY
CPU USAGE
ITERATION SCORES - INTERNAL VALIDATION
1.0251
VARIABLE IMPORTANCE

Outliers



Outliers



H2O WORLD 2017

Outliers

- An anomaly is an observation inconsistent with a set of beliefs.
  - The anomaly depends on these beliefs
- An outlier is an observation inconsistent with a set of points.
  - The points are presumed generated by a probabilistic process in a vector space.
- All outliers are anomalies but not all anomalies are outliers
  - Some anomalies are logical or mathematical
  - Outliers are probabilistic
- Outlier detection has more than a 200 year history.
  - The goal was to reduce bias in models
  - The goal today is to learn interesting stuff from examining outliers
  - Statisticians no longer delete outliers. They use robust methods.

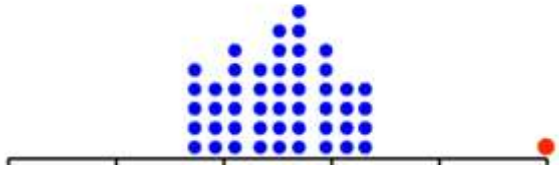
Outliers

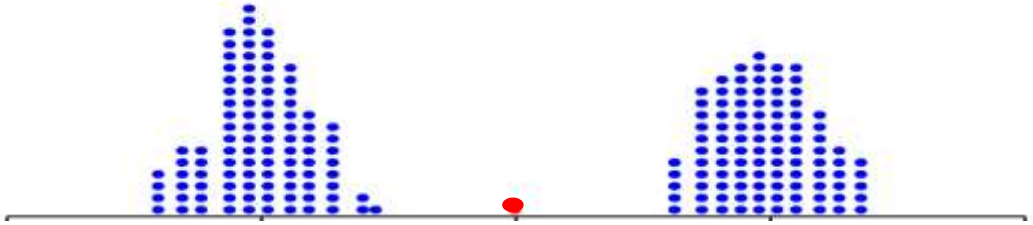
- Barnett & Lewis (1994), *Outliers in Statistical Data*.
- Rousseeuw & Leroy (1987). *Robust Regression & Outlier Detection*.
- Hartigan (1975) *Clustering Algorithms.*

Beauty is truth, truth beauty,—that is all
Ye know on earth, and all ye need to know.

Outliers

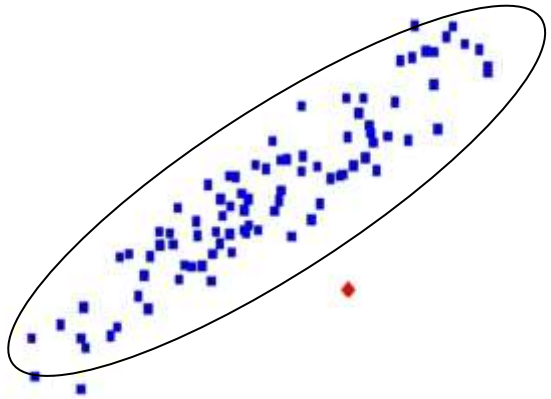
- Univariate outliers
  - Distance from Center Rule
  - Gaps Rule



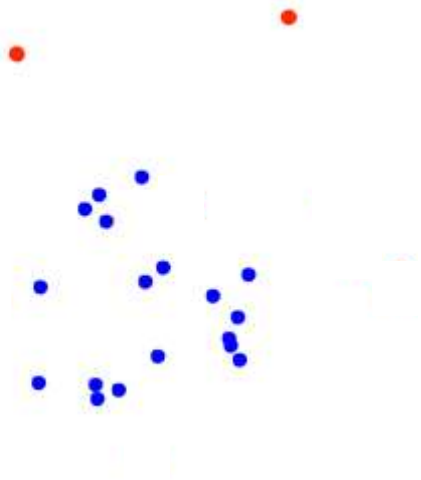


Outliers

- Multivariate outliers
  - Distance from Center Rule



  - Gaps Rule

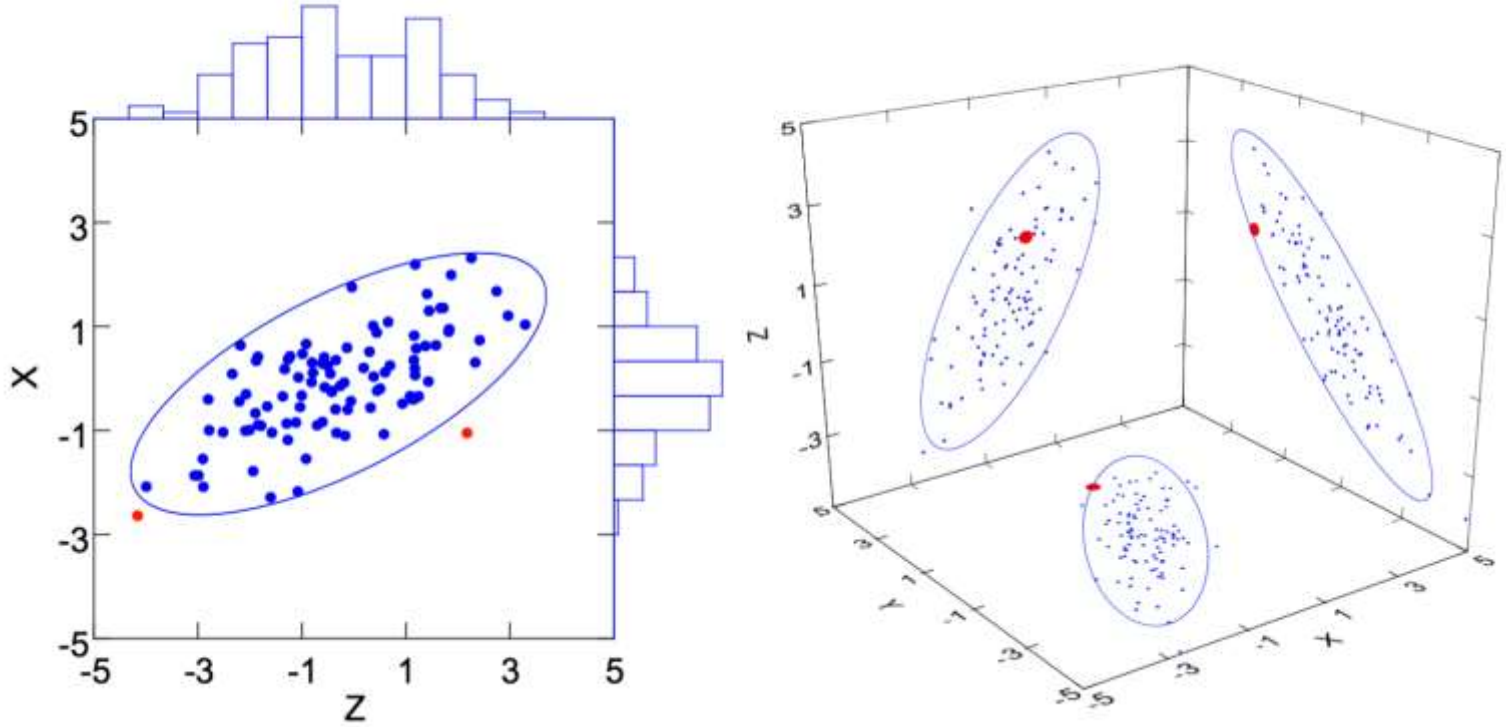


Outliers

1. Map categorical variables to continuous values (SVD).
2. If p large, use random projections to reduce dimensionality.
3. Normalize columns on [0, 1]
4. If n large, aggregate
   - If p = 2, you could use gridding or hex binning
   - But general solution is based on Hartigan's Leader algorithm
5. Compute nearest neighbor distances between points.
6. Fit exponential distribution to largest distances.
7. Reject points in upper tail of this distribution.

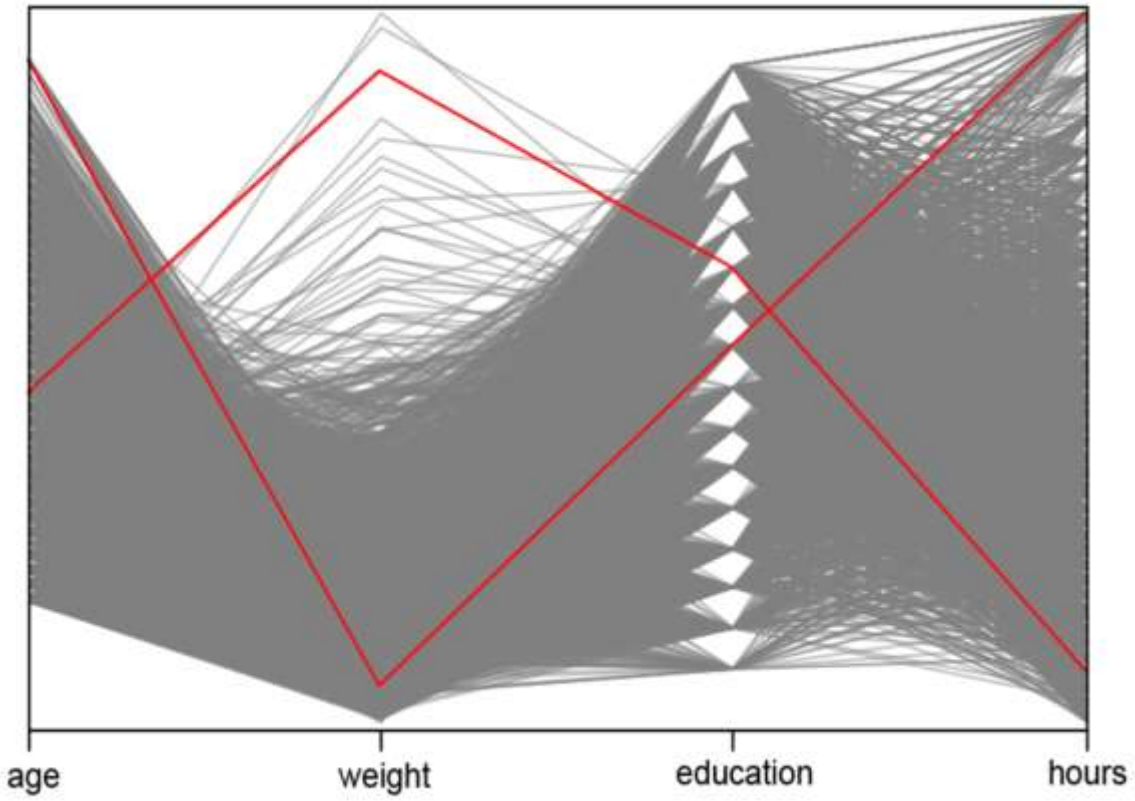
Outliers

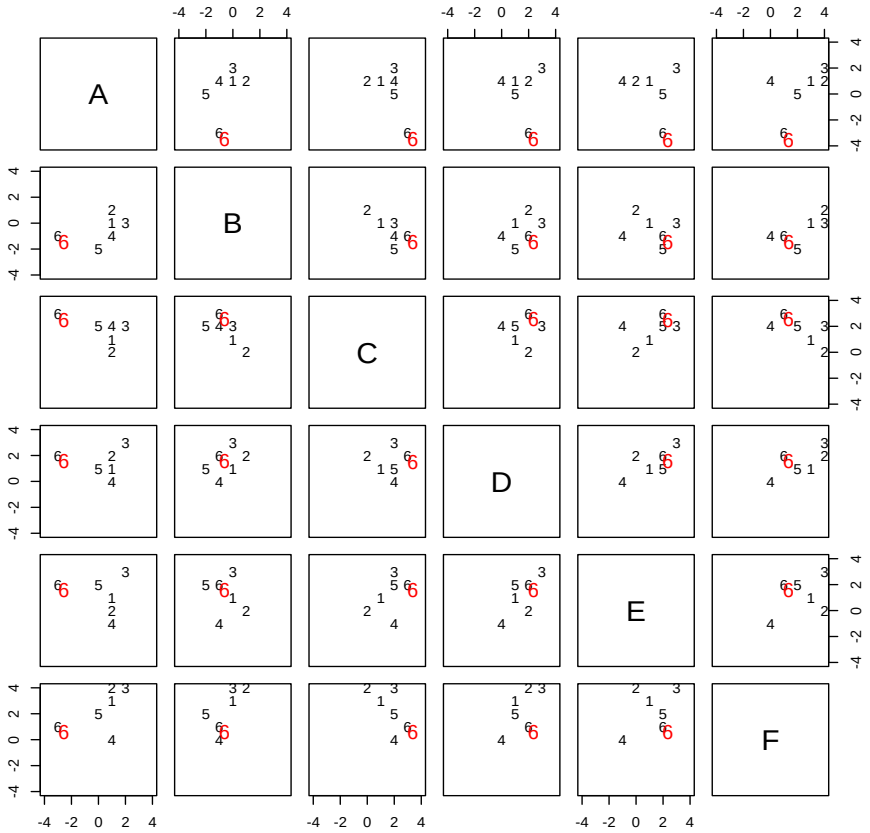
- Low-dimensional projections are not reliable ways to discover high-dimensional outliers.



Outliers

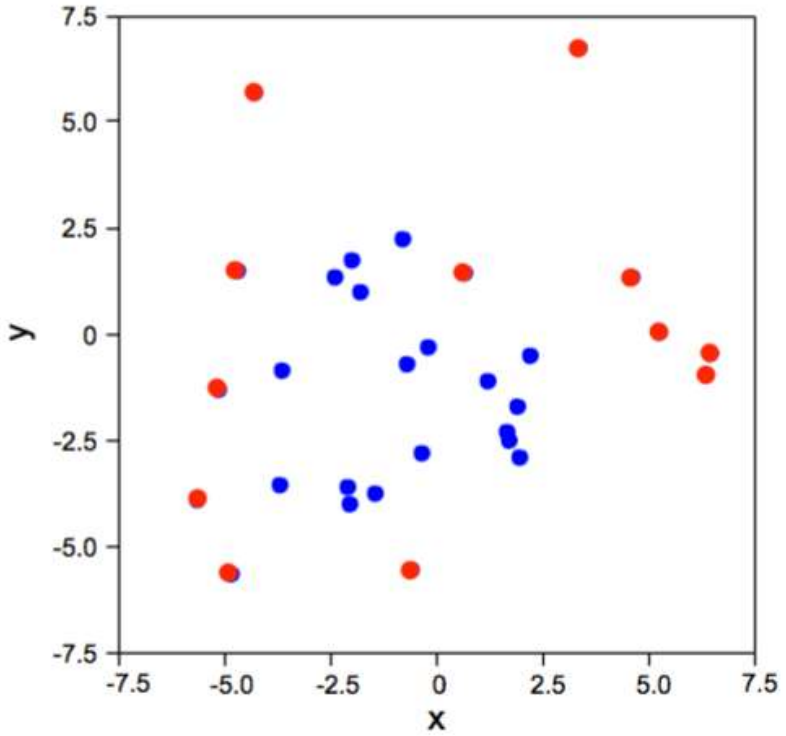
- Parallel coordinates, SPLOMs, and other multivariate visualizations are not reliable ways to discover high-dimensional outliers.

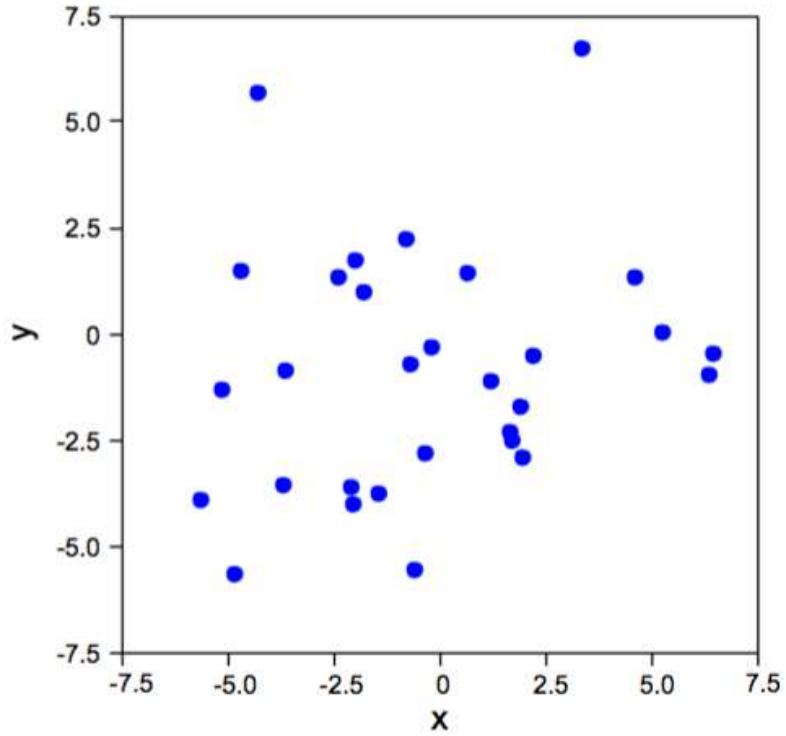




Outliers

- Popular ML algorithms are not reliable ways to identify outliers.

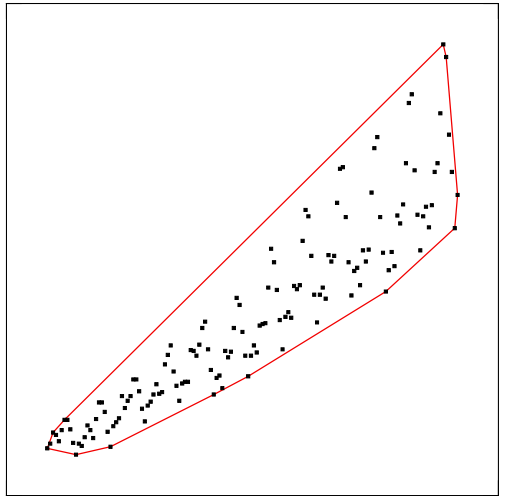


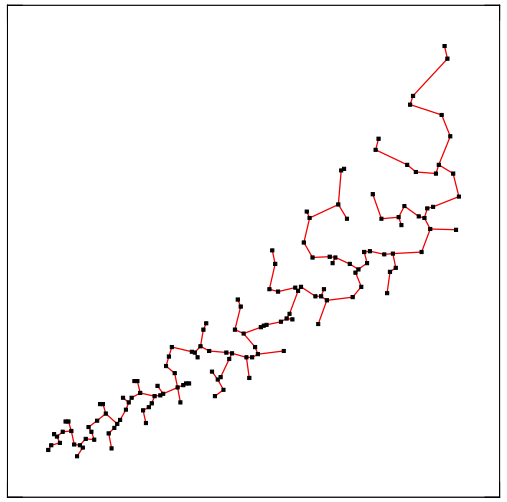


H2O WORLD 2017

Scagnostics

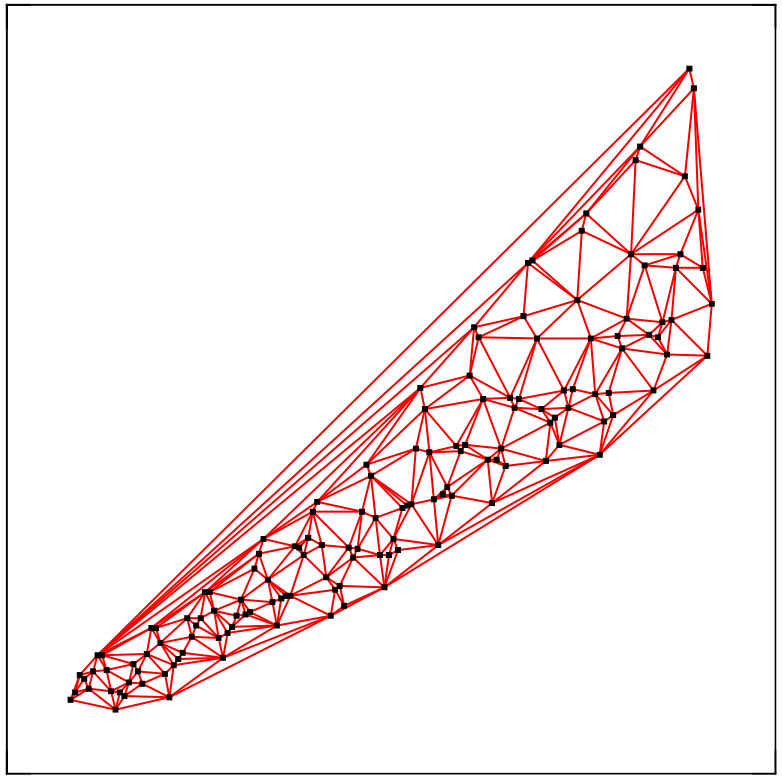
- We characterize a scatterplot (2D point set) with nine measures.
- We base our measures on three geometric graphs.
  - Convex Hull
  - Alpha Shape
  - Minimum Spanning Tree





Scagnostics

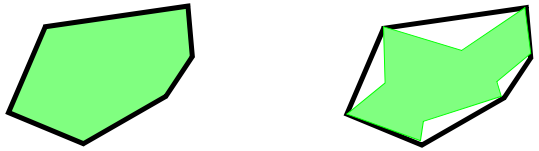
- Each geometric graph is a subset of the Delaunay triangulation



Scagnostics

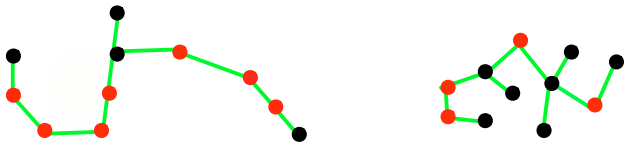
Shape

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



Skinny: ratio of perimeter to area of the alpha shape

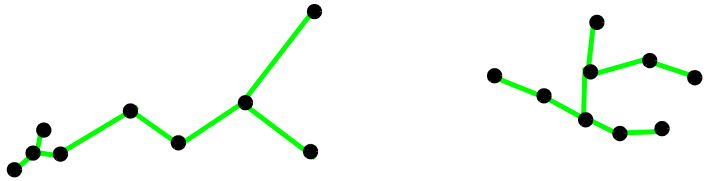
Stringy: ratio of 2-degree vertices in MST to number of vertices > 1-degree



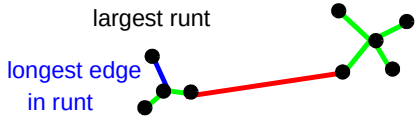
Scagnostics

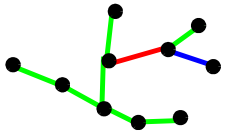
Density

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

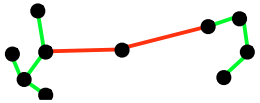


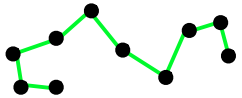
Clumpy: 1 minus the ratio of the longest edge in the largest runt (blue) to the length of runt-cutting edge (red)





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





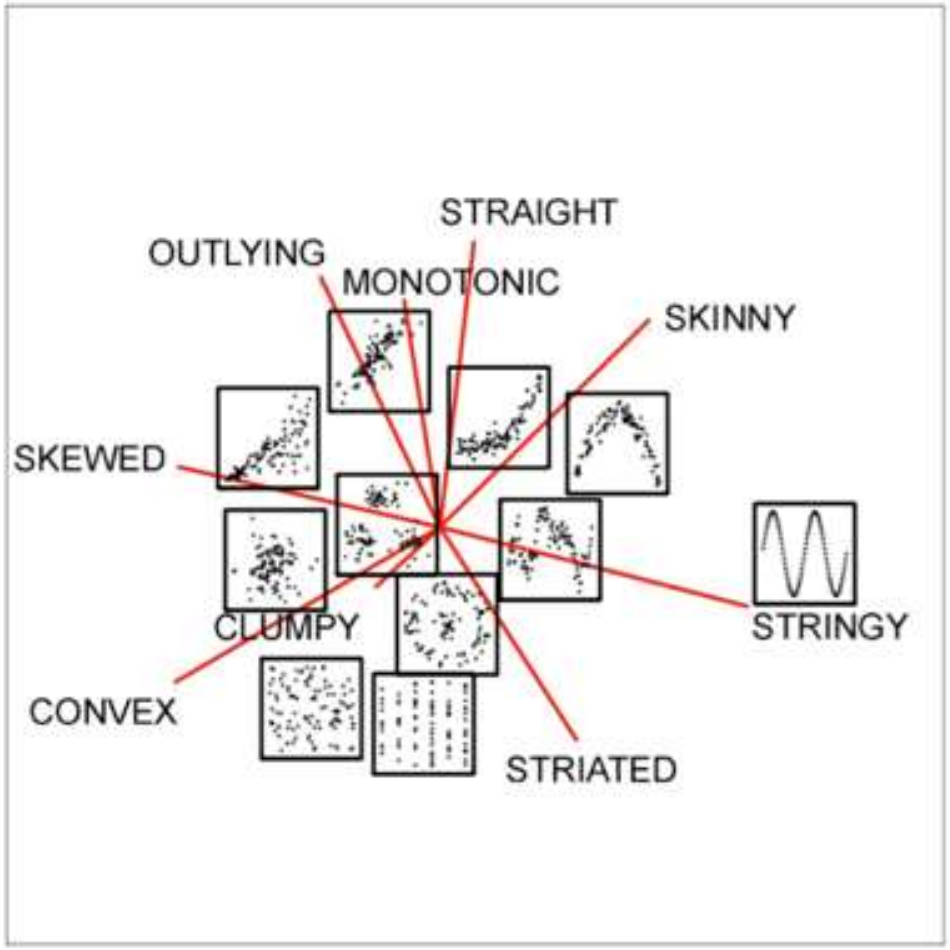
Scagnostics

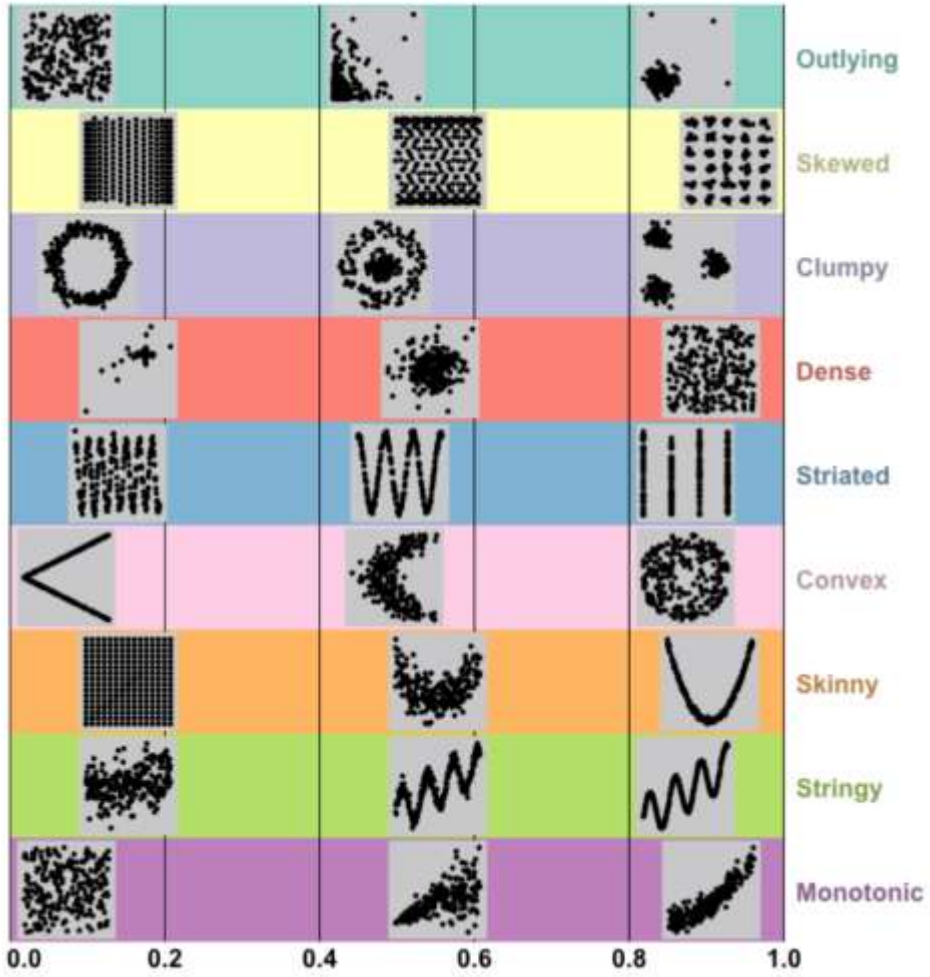
Density

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

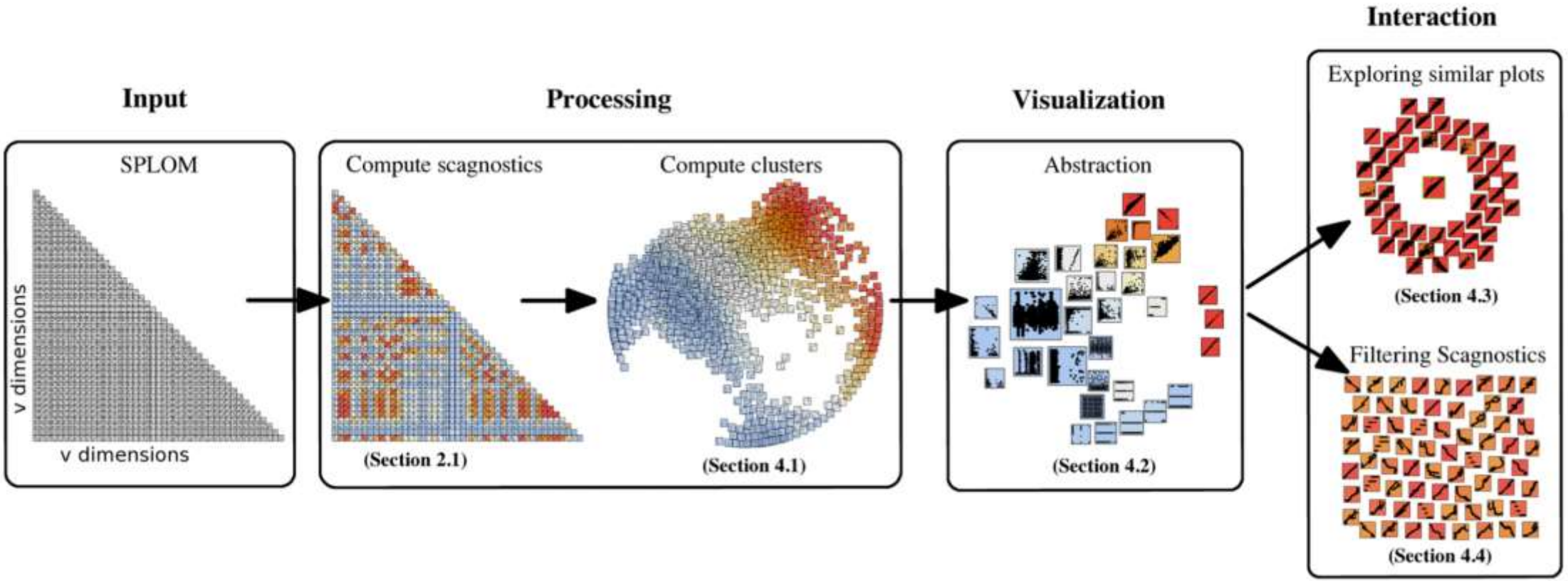
Striated: proportion of all vertices in the MST that are degree-2 and have a cosine between adjacent edges less than -.75

Scagnostics

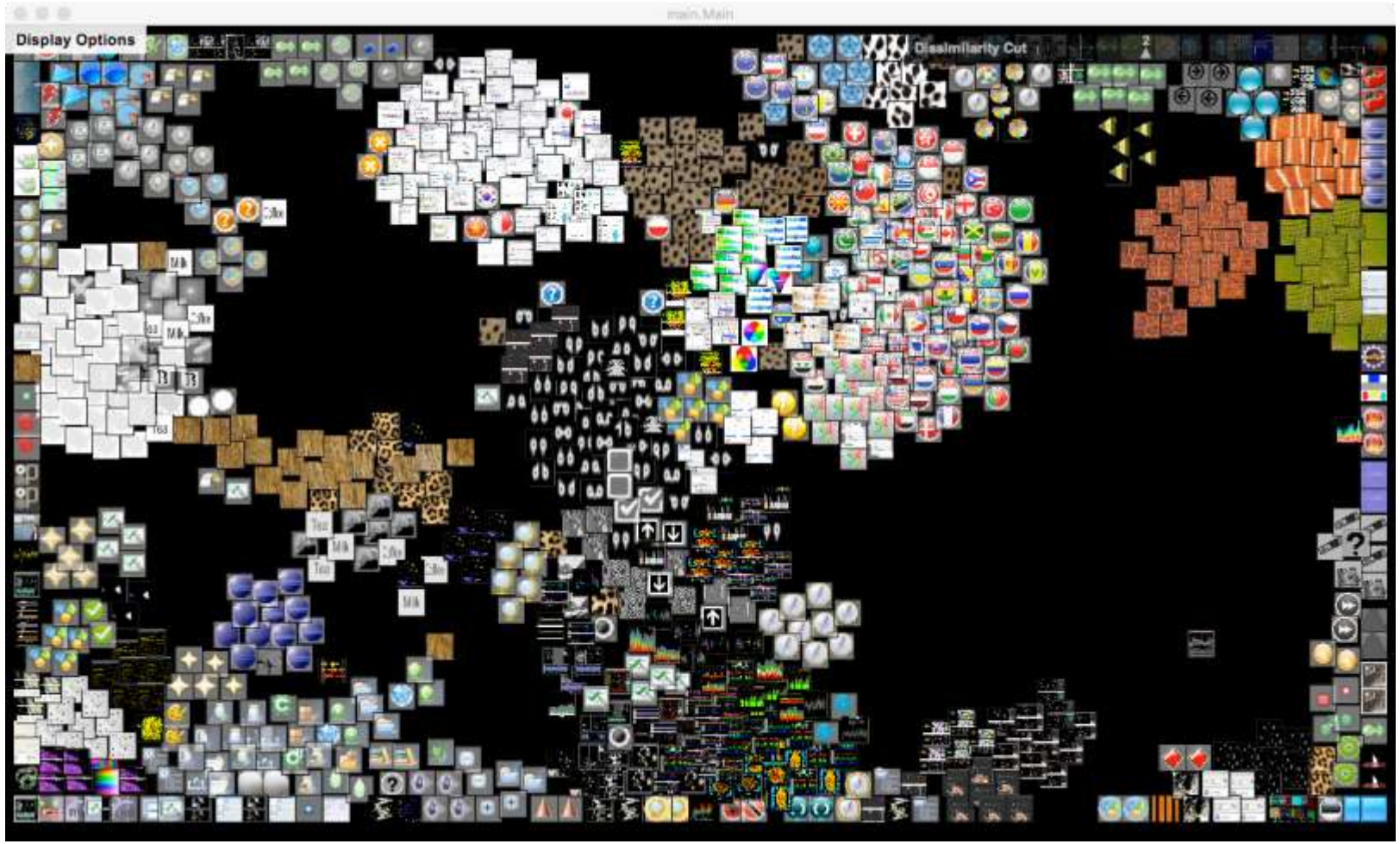




Scagnostics



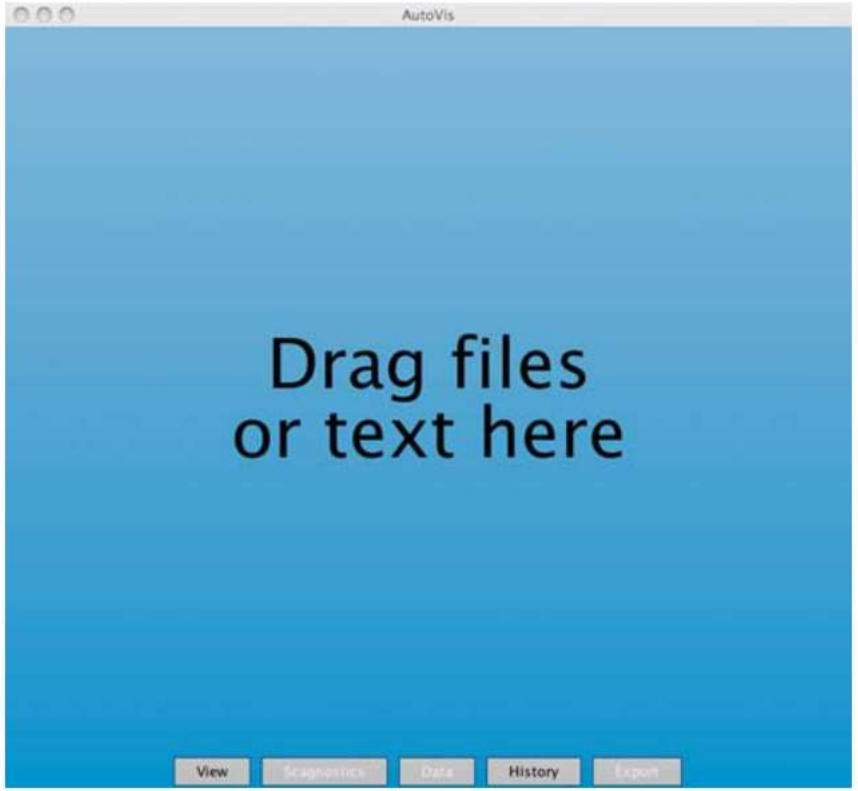
Scagnostics



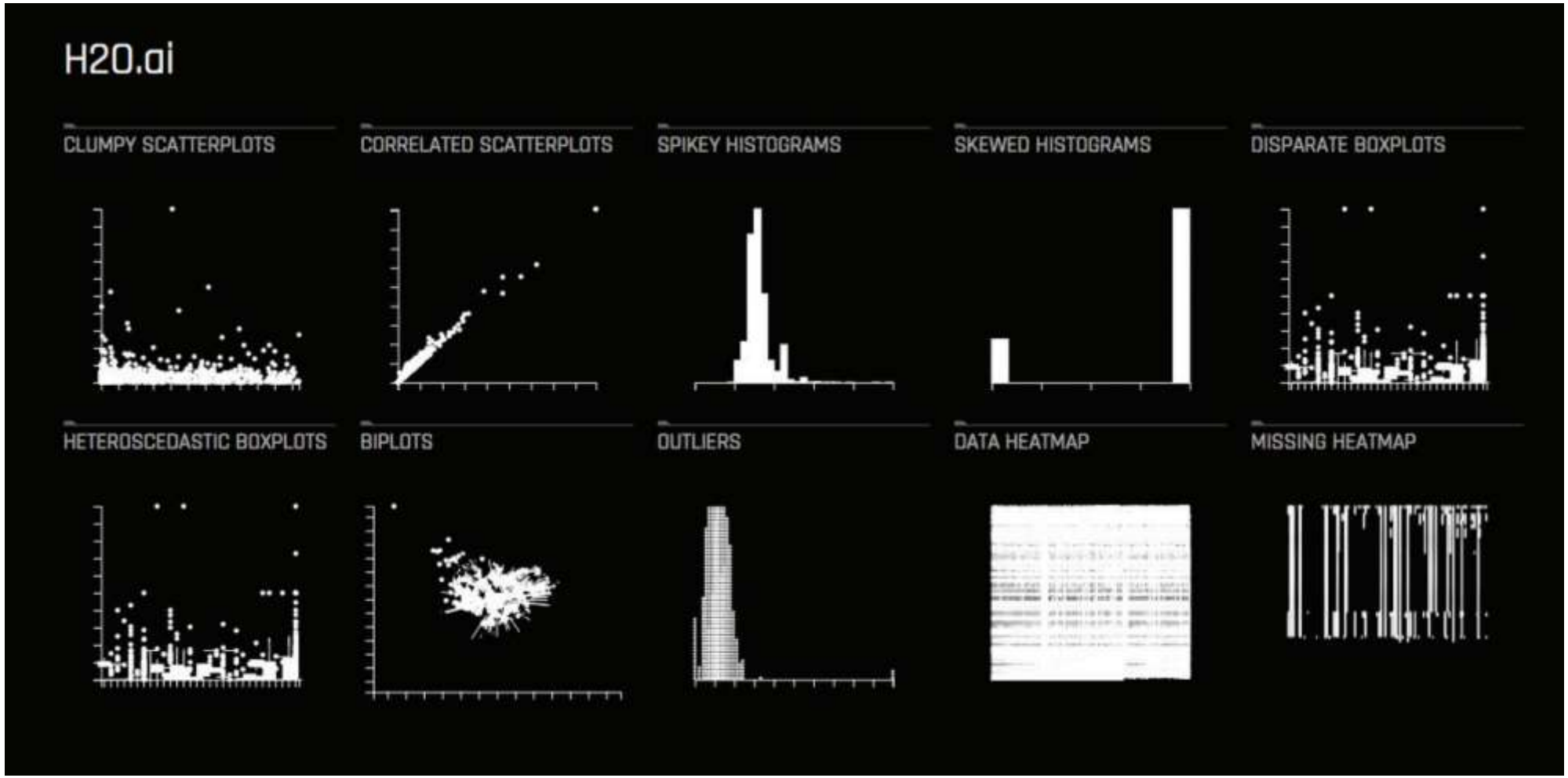
H2O WORLD 2017

AutoVis

Graham Wills and Leland Wilkinson. 2010. AutoVis: automatic visualization. *Information Visualization* 9, 1 (March 2010), 47-69.



H2O AutoViz



H2O WORLD 2017

Future Plans

1. Add brushing to graphics
2. Create case-weight vector for DAI (0 = exclude)
3. Suggest additional features to pass to DAI
4. Animate visualizations
5. Add natural language explanations to graphics.

Thank You!