Big Data Visualization

Stats 285, Stanford
November 11, 2019

Leland Wilkinson

Chief Scientist, H2O
Adjunct Professor UIC

Leland.Wilkinson@h2o.ai
www.cs.uic.edu/~wilkinson

Big Data

File size is a useless statistic (giga, tera, peta, exa, zetta, …)

- Graph data
  - how many nodes and edges?
- Rectangular data
  - how many rows and columns?
  - how long are the strings?
  - how many distinct strings?
  - what is the precision of the numbers?
  - what is the file format?
- Image data
  - how many images?
  - resolution of the images?
- Text data
  - how many words?
  - what language?

Big Data

Many big data problems can be attacked with software or hardware

- distributed file systems
- GPUs
- Columnar in-memory databases

And some big data problems can be attacked with models

- deep learning
- stacking

These are the kind of things computer scientists think are “solutions”

But the problems big data presents to visualization involve other things

- human factors (perception, cognition, …)
- display limitations (pixels, rendering, …)
- real-time performance (constrained by human in the loop)

So let’s look at some of these problems peculiar to visualization

Big Data

Problems: Difficulties peculiar to big data visualization

Solutions:

- Architecture: Design of a big data visualization system
- Wrangling: Ways to make big data tractable for visualization
- Graphics: Graphics suited for big data exploration

Problems

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

Solutions

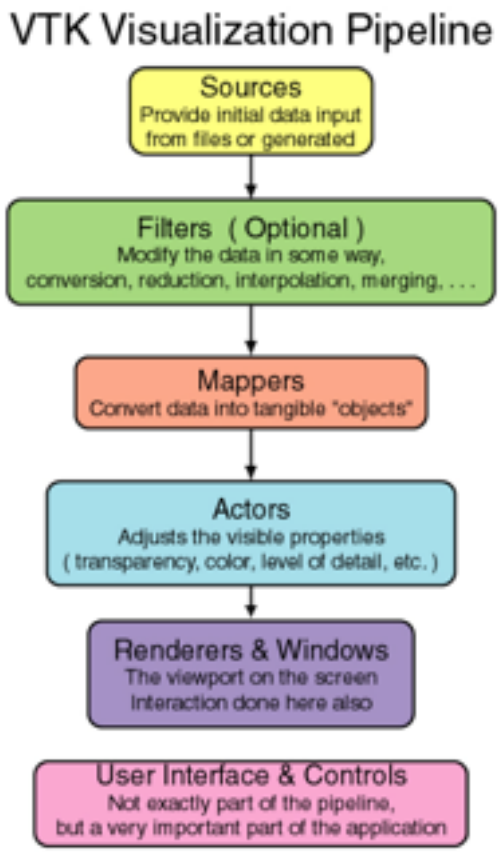
Architecture: Design of a big data visualization system

Wrangling: Ways to make big data tractable for visualization

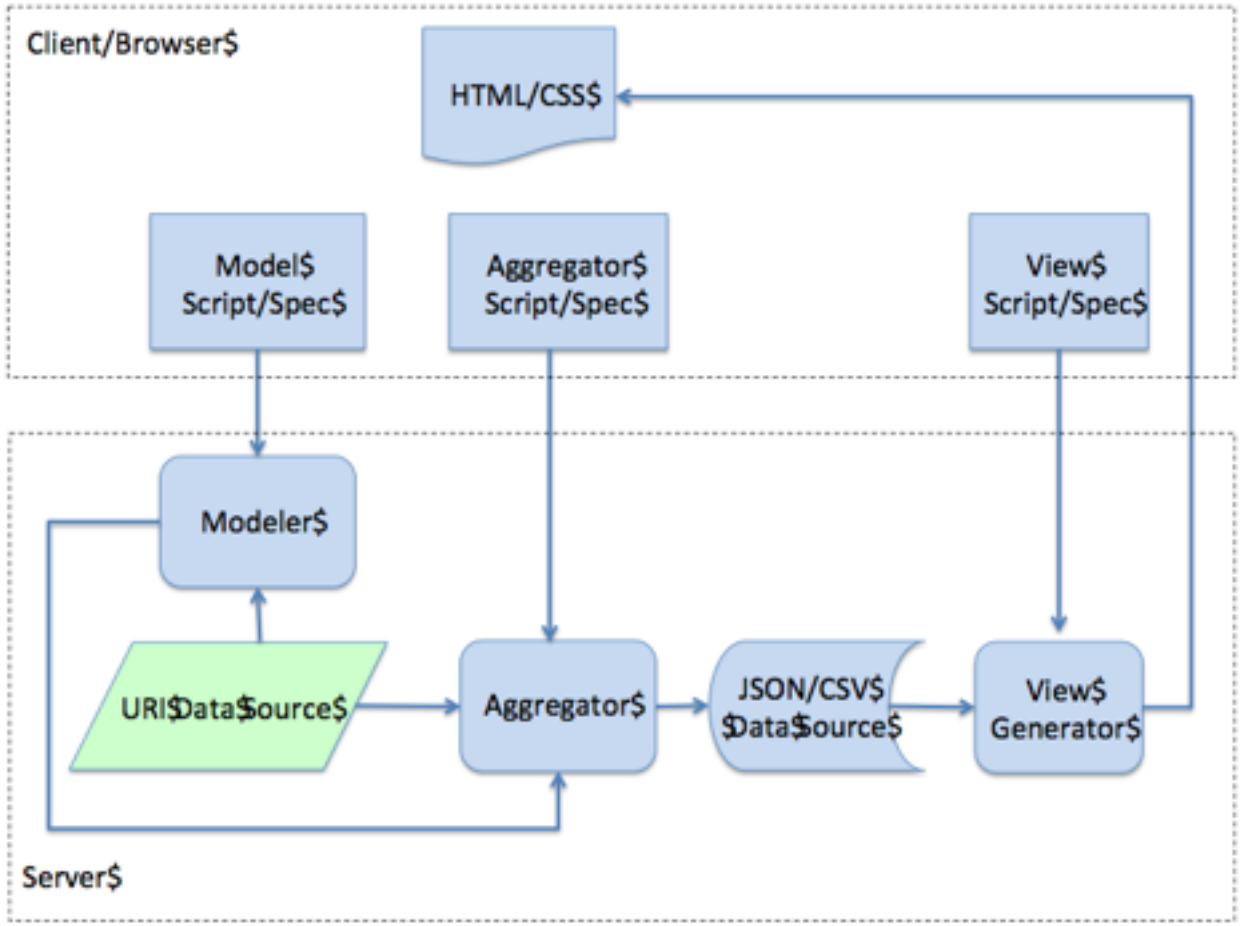
Graphics: Graphics suited for big data exploration

Architecture

Old



New



Wrangling

Aggregate (big n)

usually reduces accuracy when $k \ll n$

Reduce (big p)

usually violates triangle inequality when $d \ll p$

We have some flexibility because of limited range of precision in visualization

But that's not a hunting license

n number of rows in dataset

p number of columns in dataset

k number rows in aggregated dataset

d number of columns in reduced dataset

Aggregate

Do not aggregate if n is tractable : unless resolution demands it

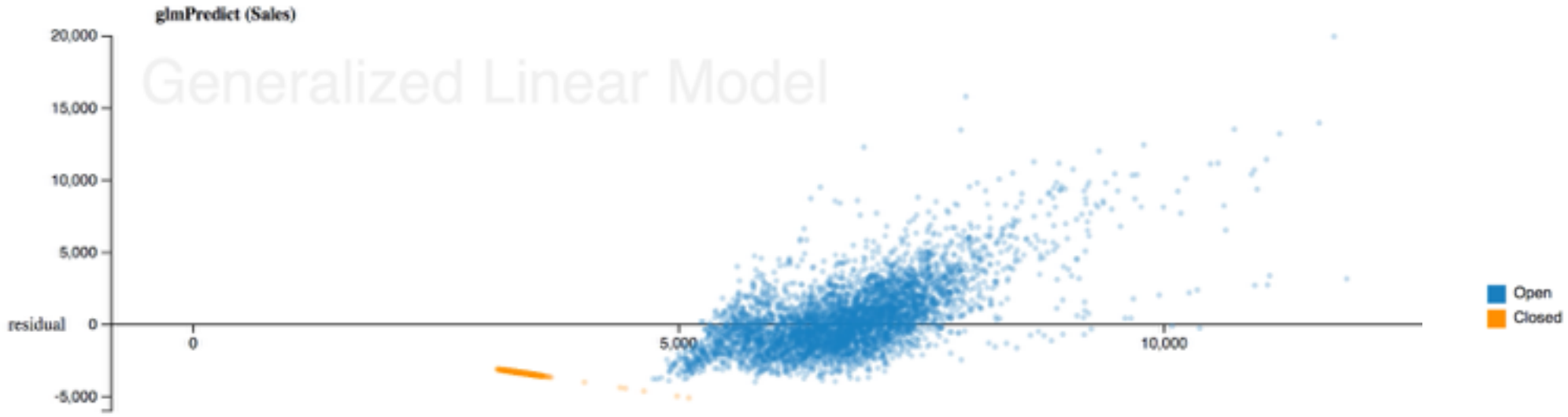
Do not sample: unless using bootstrapped visualization to represent error

Do use different algorithms: 1D, 2D, nD

Do Not Aggregate

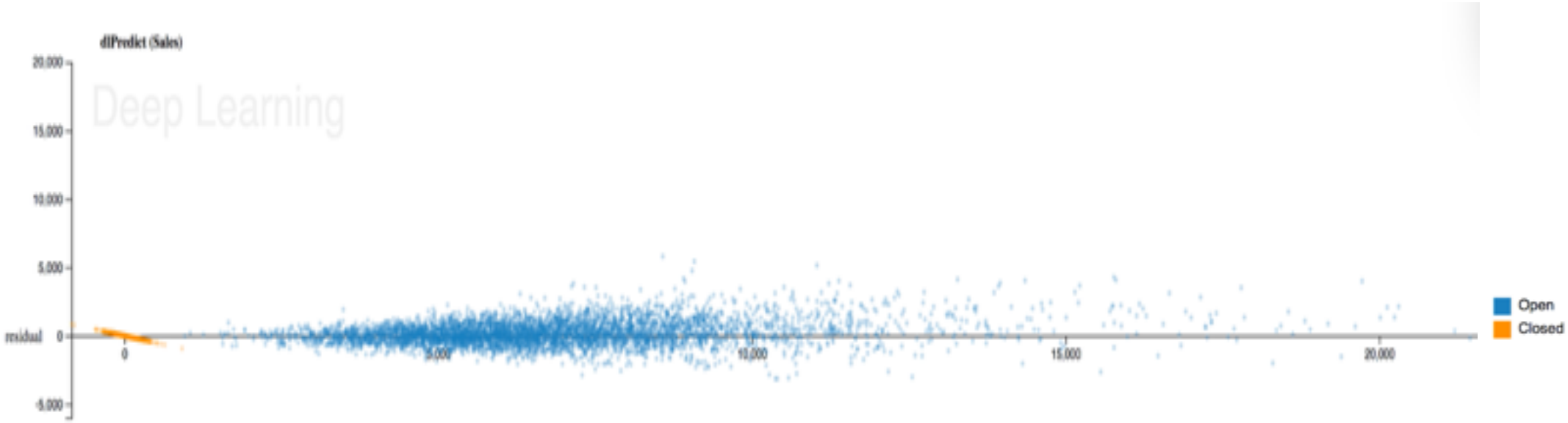
Residual plots.

Rossmann Stores Kaggle dataset (https://www.kaggle.com/c/rossmann-store-sales)



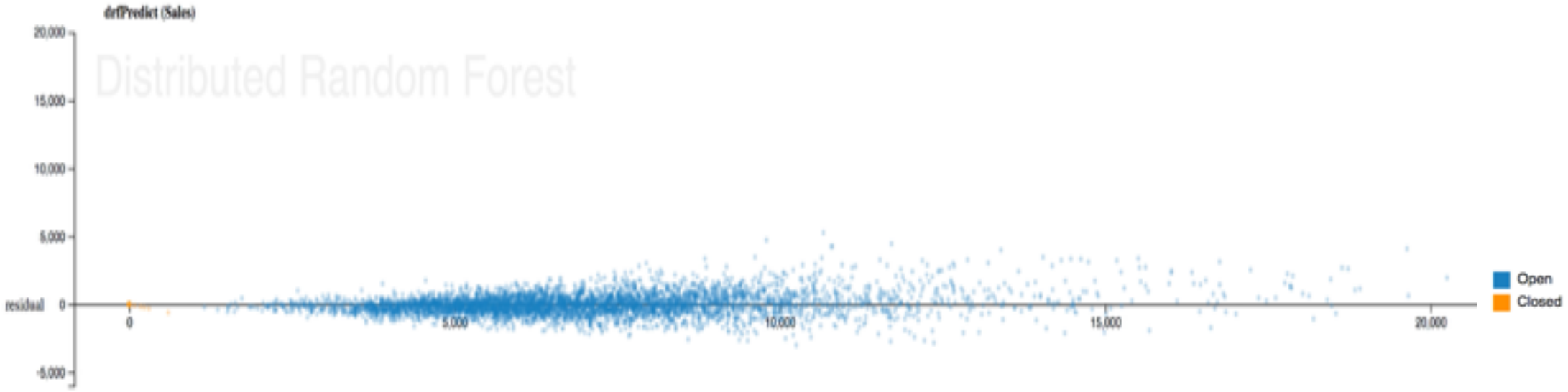
Do Not Aggregate

Residual plots.



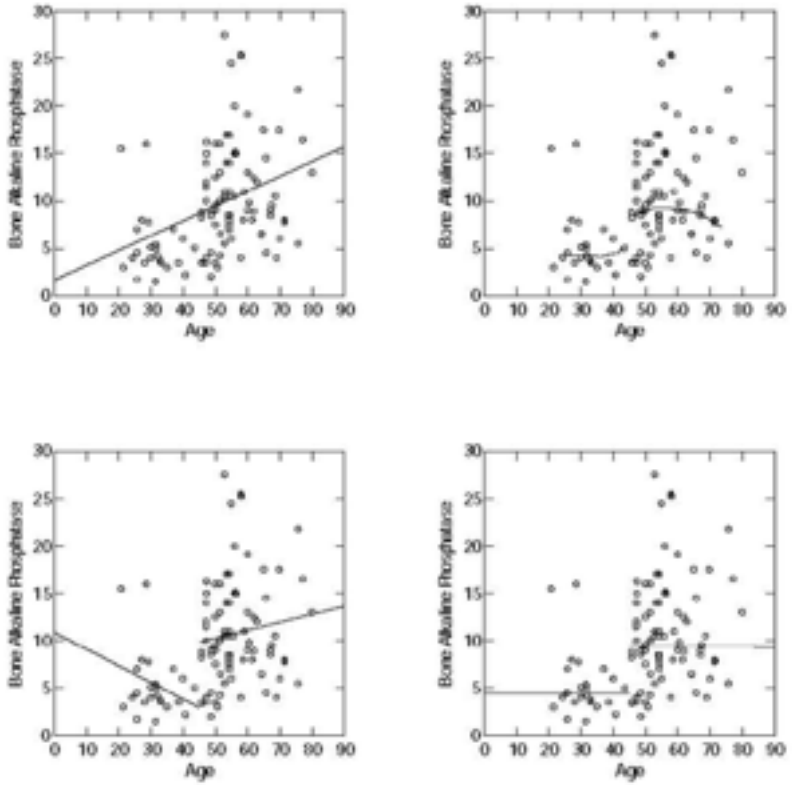
Do Not Aggregate

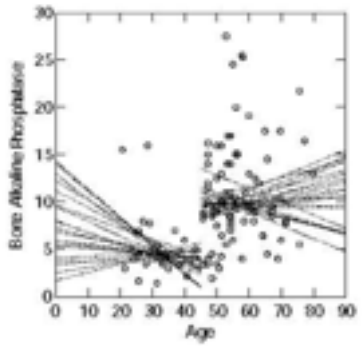
Residual plots.



Do Not Sample

Unless Bootstrapping





Gonnelli, S., Cepollaro, C., Montagnani, A., Monaci, G., Campagna, M.S., Franci, M.B., and Gennari, C. (1996). Bone alkaline phosphatase measured with a new immunoradiometric assay in patients with metabolic bone diseases. *European Journal of Clinical Investigation, 26*, 391– 396.

1D Aggregation/Quantization

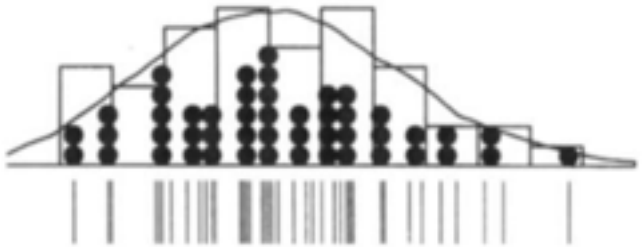
Dot plot algorithm

Wilkinson, L. (1999). Dot plots. *The American Statistician*, *53*, 276–281.

Sort data
Choose dot size and set first stack to min(x) location
For i = min(x) to max(x): add dot to stack at current stack or start new stack at x_i if no collision

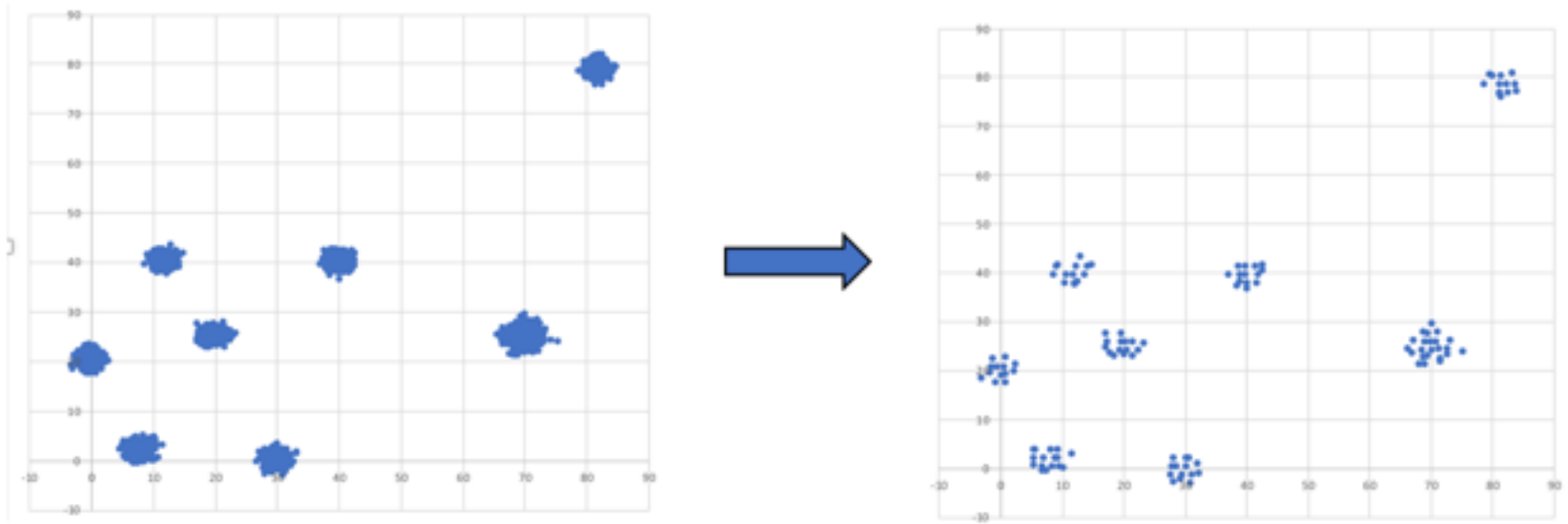
Histogramming algorithm

Choose small bin width (k = 100 bins works well for most display resolutions)
When finished, average the values in each bin to get a single centroid
Delete empty bins and return centroids and counts in each bin



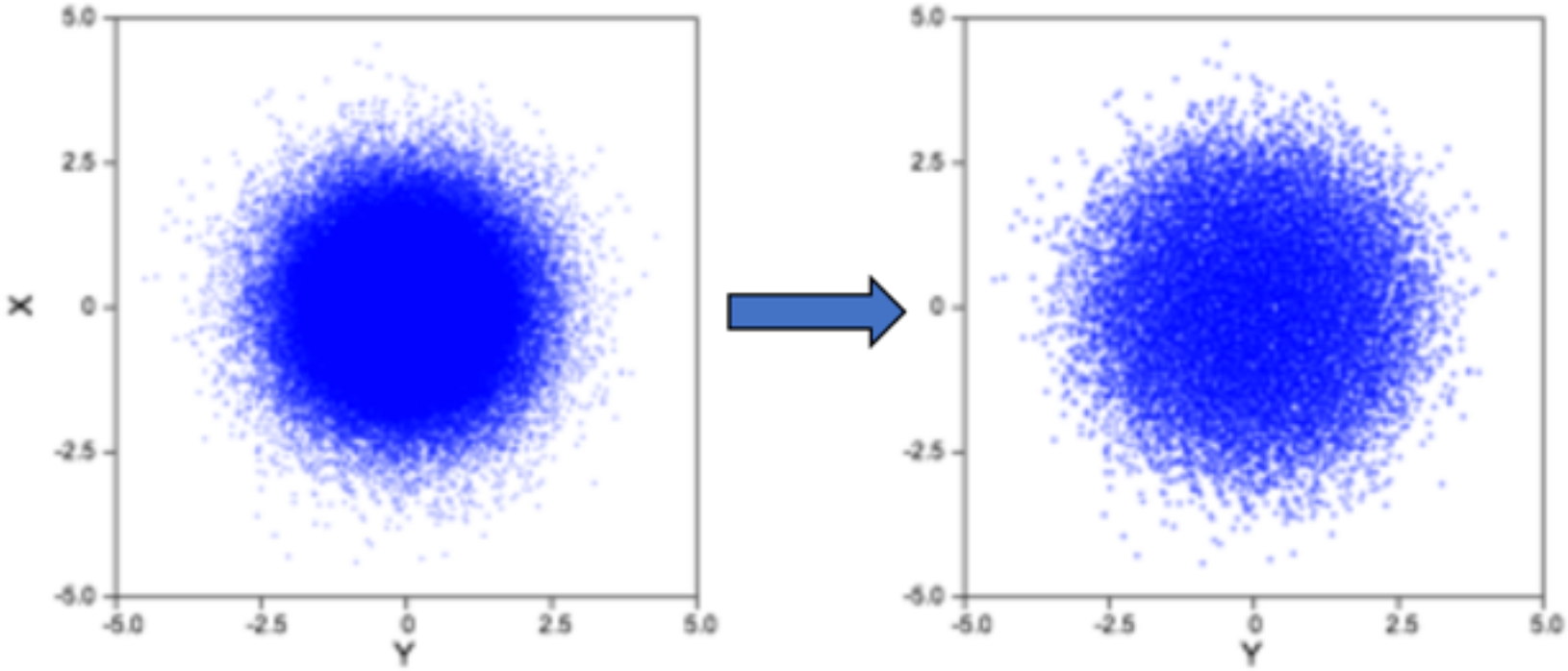
2D Aggregation

Histogramming/gridding algorithm



2D Aggregation

Histogramming/gridding algorithm



nD Aggregation

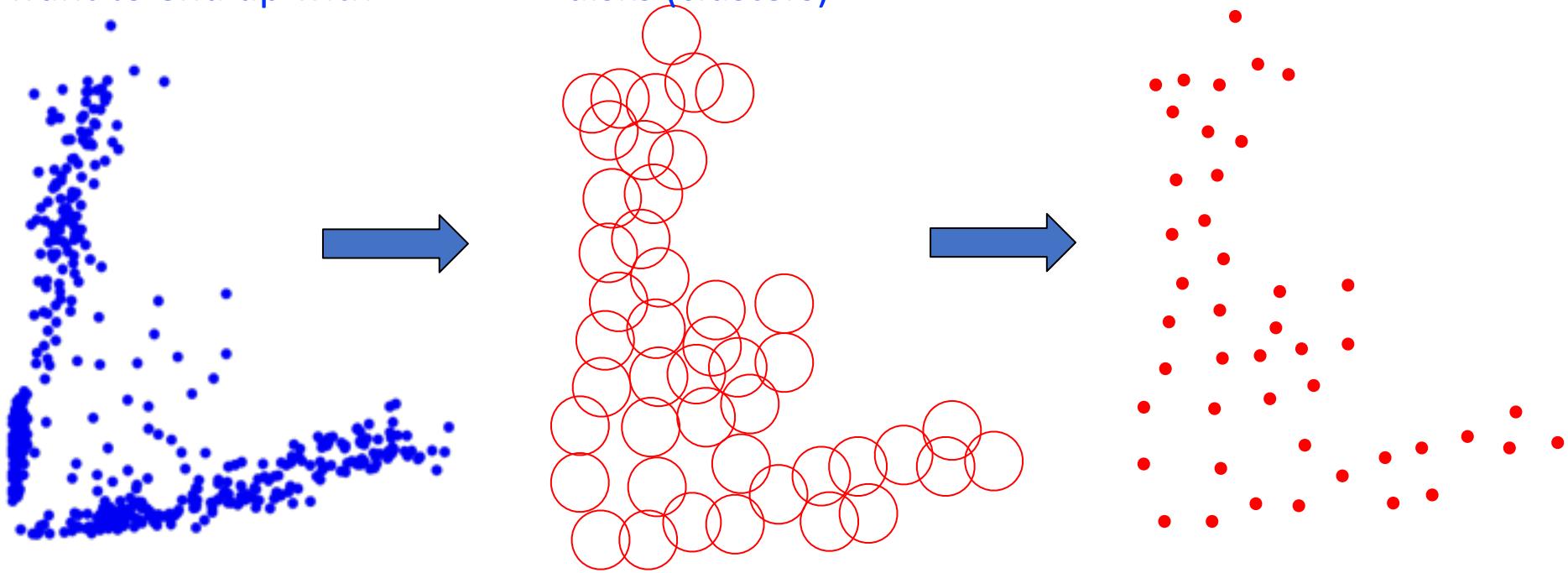
Leader Algorithm

Resembles a set cover (core sets)

In 1D, reduces to the Wilkinson dot plot algorithm

Disk (ball) radius r is a parameter that determines degree of aggregation

We want to end up with $k \approx 500$ disks (clusters)



nD Aggregation

Leader Algorithm

Each disk is centered on an *exemplar* (real data point, not a centroid as in k-means)

Each disk contains m *members*

Disk (ball) radius r depends on distribution of points

Simple strategy is to run algorithm with tiny r and then expand r in a golden search toward k = 500

Leader has worst case complexity $O(n^2)$

But if $k \ll n$ (which is our usual case), it will be much faster

Aggregating Categorical Variables

Multiple Correspondence Analysis

For each categorical variable:

dummy code (0/1) categories

compute first principal component on covariance matrix of dummy codes

numeric value is product of dummy codes of category and first principal component

Other categorical → continuous mappings could be used

Aggregate

ALL statistics on aggregated data MUST include frequency weights

MOMENTS (Python)

```
for x in data:
    if weights != None:
        wt = weights[i]
    if wt > 0:
        if x != None:
            xCount += 1
            xWeightedCount += wt
            xSum += x * wt
            xd = (x - xMean) * wt
            xMean += xd / xWeightedCount
            xSS += (x - xMean) * xd
```

LOESS (Java)

```
for (int k = left; k <= right; k++) {
   double xk = x[k];
   double yk = y[k];
   double dist = xk - xi;
   if (k < i)
      dist = xi - xk;
   double wt = tricube(dist * denom) * weights[k] * frequencies[k];
   double xkw = xk * wt;
   sumWeights += wt;
   sumX += xkw;
   sumXSquared += xk * xkw;
   sumY += yk * wt;
   sumXY += yk * xkw;
}
```

Reduce

Projection of a set of points in $\mathbb{R}^p$ to $\mathbb{R}^d$: such that $d_p(x,y) \underset{\sim}{\propto} d_d(x,y)$

- Principal Components, SVD: linear projection
- Random: random projection
- Discovering dimension d is problematic: don't plan on looking for elbow

Feature Extraction: replace points with derived features

Principal Components (SVD)

Singular Value Decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

Alternatively

$$\mathbf{S} = \mathbf{X}^T\mathbf{X}/N$$

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= (\mathbf{U}\mathbf{D}\mathbf{V}^T)^T\mathbf{U}\mathbf{D}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{D}^2\mathbf{V}^T\end{aligned}$$

Pick first k principal components

Random Projections

Replace principal components with random Gaussian elements

$$\mathbf{X}^{(d)}_{n\times d} = \mathbf{X}_{n\times p}\mathbf{W}_{p\times d}$$

$\mathbf{W}$ is matrix of random Gaussians

Johnson, W. B. and Lindenstrauss, J. (1984). Lipschitz mapping into Hilbert space. Contemporary Mathematics 26, 189–206.

Achlioptas, D. (2001). Database-friendly random projections. In PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, New York, 274– 281.

Li, P., Hastie, T. J., and Church, K. W. (2006). Very sparse random projections. In KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, 287–296.

$\mathbf{W}$ can be matrix of {1, 0, -1}

Discovering dimensionality of embedding

Elbow test on scree plot doesn't work on most real data

- Even piecewise regression with cutpoint as estimated parameter doesn't work

Kaiser method doesn't work either

- Factor correlation matrix
- Retain components with eigenvalues > 1

Kaiser H. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement, 20*. 141–151.

Horn's Parallel Analysis works better (but not always)

- Generate random data for problem of same size
- Compute eigendecomposition on random and real data
- Compute average eigenvalue over k samples of random data
- Retain real data components whose eigenvalues are greater than average of random eigenvalues

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30* (2). 179–185

Feature Selection

Scagnostics

Scagnostics

An idea of John and Paul Tukey

Never published, but discussed in a JSM talk

Given many scatterplots (too many to view in a scatterplot matrix)

How can we identify unusual scatterplots?

Their approach involved expensive computations

principal curves, kernels, etc.

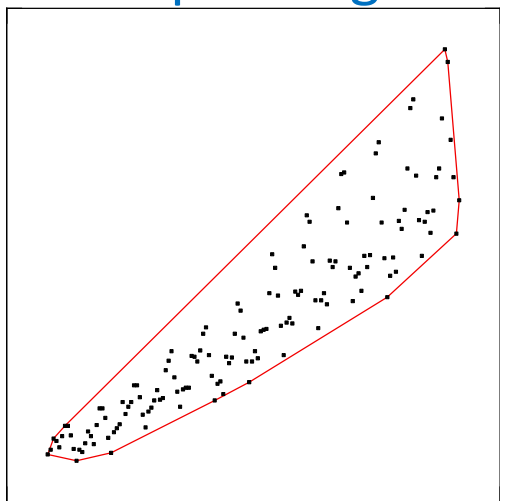
Scagnostics

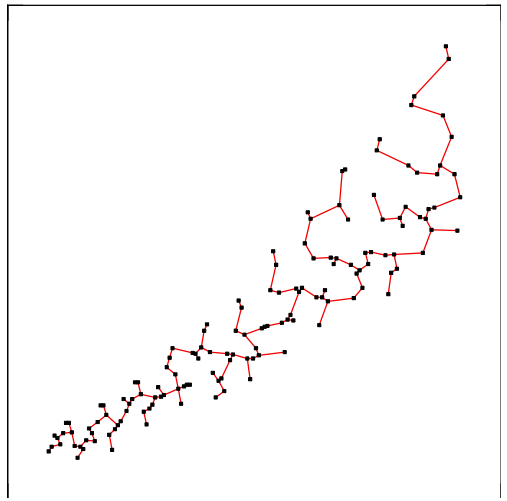
Wilkinson L., Anand, A., and Grossman, R. (2006). High-Dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. IEEE Transactions on Visualization and Computer Graphics, *12*(6) pp. 1363-1372.

We characterize a scatterplot with nine measures.

We base our measures on three geometric graphs.

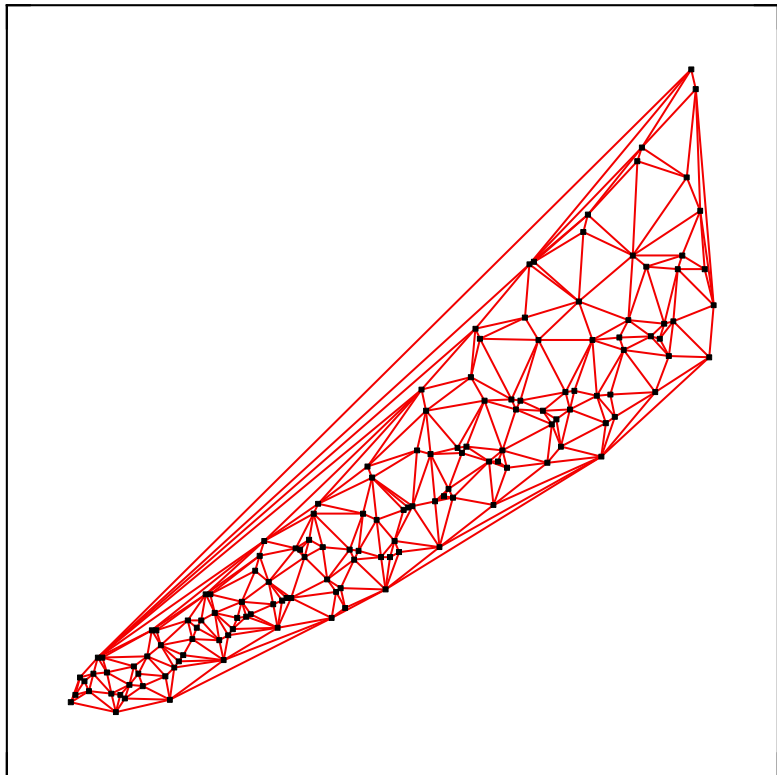
- Convex Hull
- Alpha Shape
- Minimum Spanning Tree





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

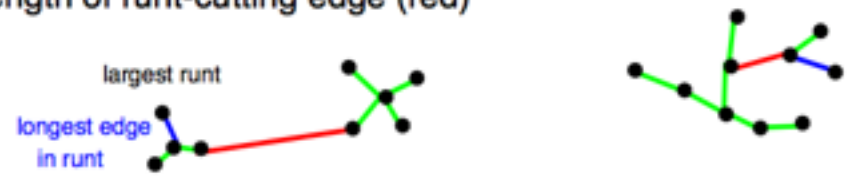
Stringy: ratio of 2-degree vertices in MST to number of vertices > 1-degree

Scagnostics

Density

Skewed: ratio of $(Q_{90} - Q_{50}) / (Q_{90} - Q_{10})$,
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

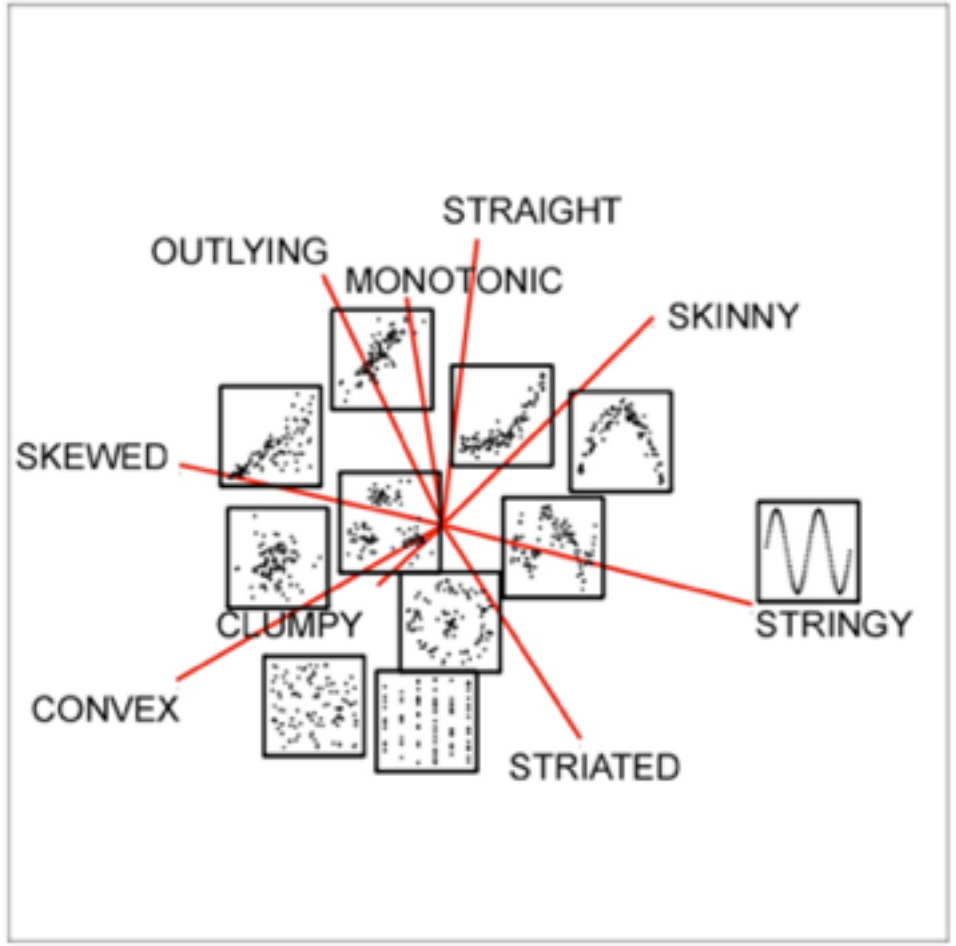
Scagnostics

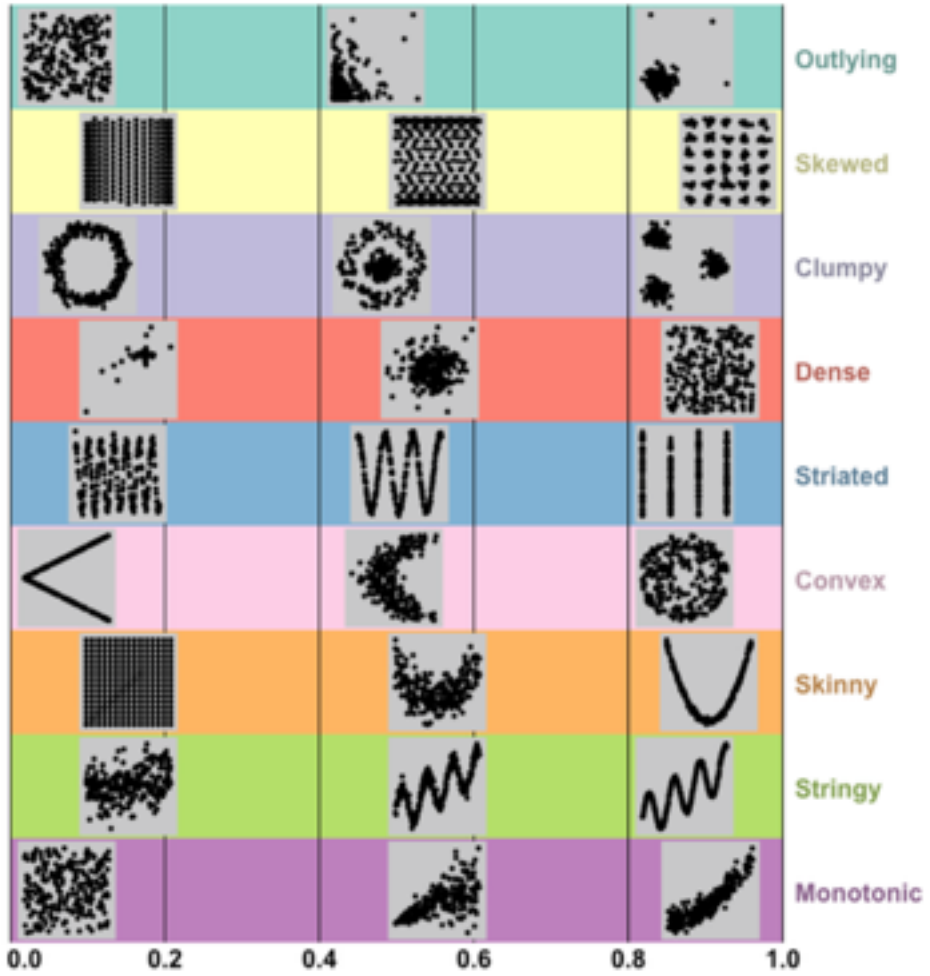
Density

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

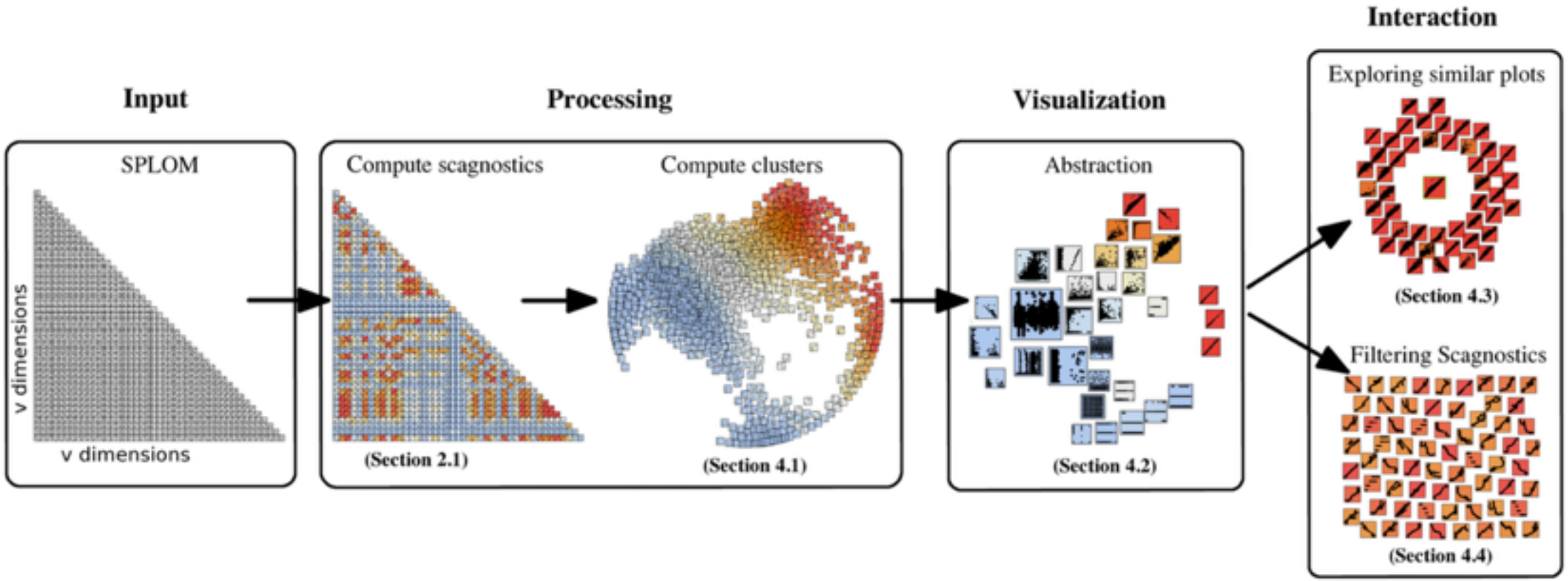
Striated: proportion of all vertices in the MST that are degree-2 and have a cosine between adjacent edges less than -.75

Scagnostics

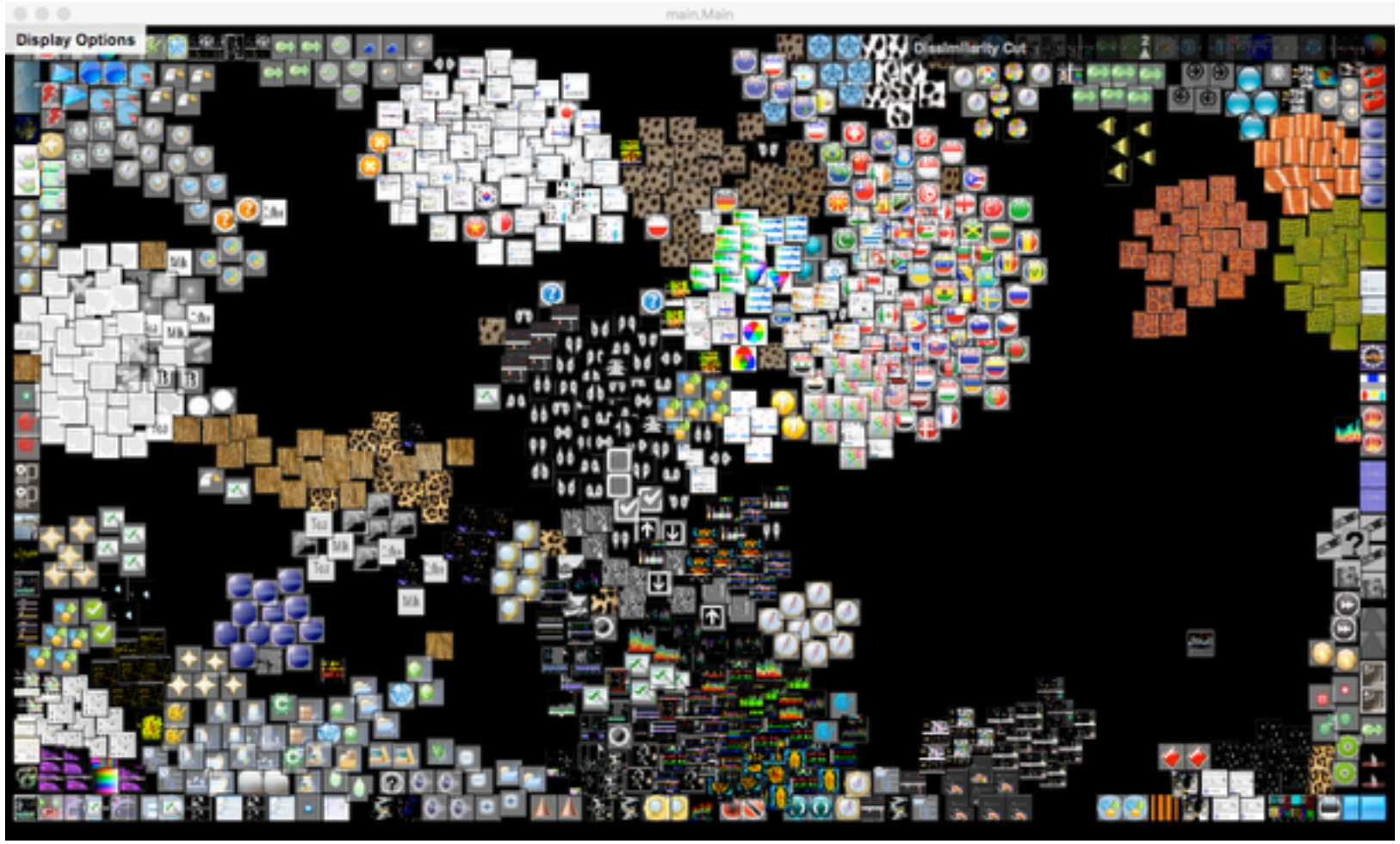




Scagnostics



Scagnostics



Graphics on Aggregated Data

1D

- Box plots
- Histograms
- Dot plots
- Kernel densities

2D

- Scatterplots

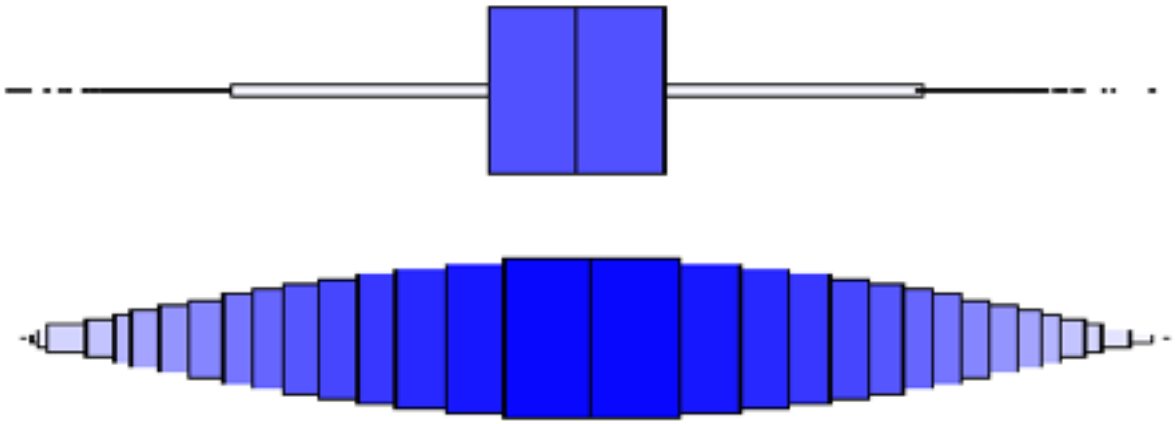
nD

- Scatterplot matrices
- Parallel coordinates
- Projections
- Graph layout

Box Plots

Not suitable for large n

- Tukey designed them for hand calculation on small batches
- Velleman and Hoaglin devised a computer program to plot them
- Tukey's fences were based on fractiles of normal distribution
- Since n is not part of the algorithm, outliers explode with big n



Hofmann, H., Wickham, H. & Kafadar, K. (2017) Letter-Value Plots: Boxplots for Large Data, *Journal of Computational and Graphical Statistics*, *26:3*, 469-477.

Histograms

Intro-stat book algorithms will get you into trouble

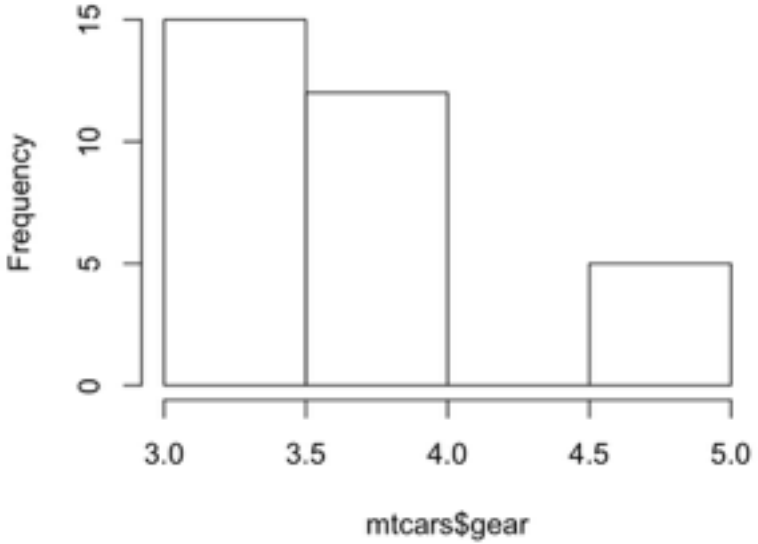
Details

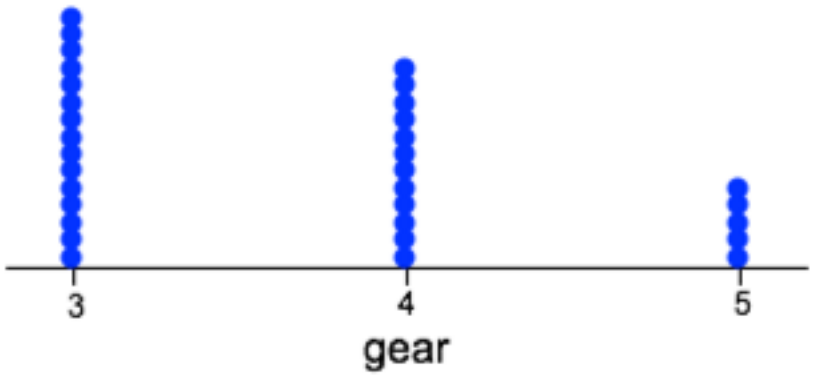
Choose number of bins (or, equivalently, bin width)

Sturges (1926), Doane (1976), Scott (1979) , Freedman and Diaconis (1981) , Stone (1985), Yu & Speed (1991)

Align scale tick values with edges of bins, so scale choice depends on bin widths

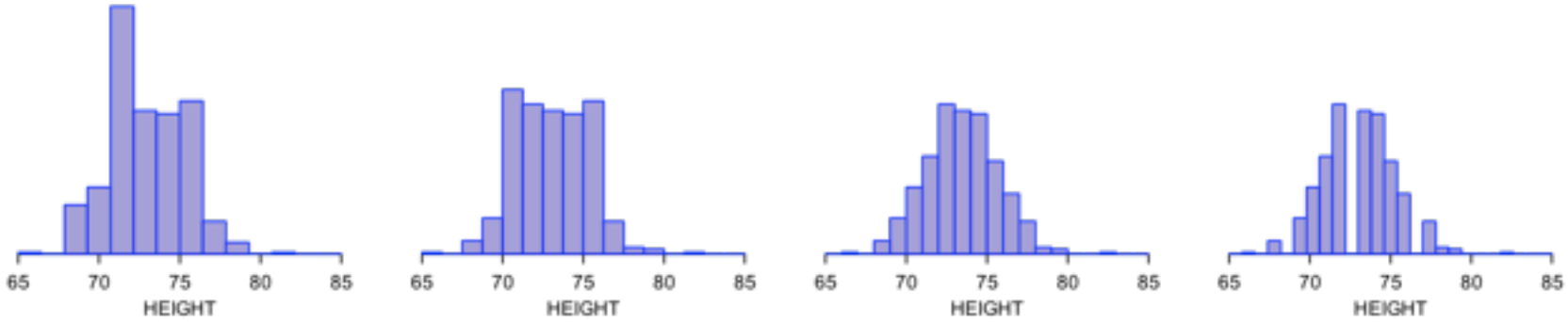
A histogram program should recognize when you feed it discrete values



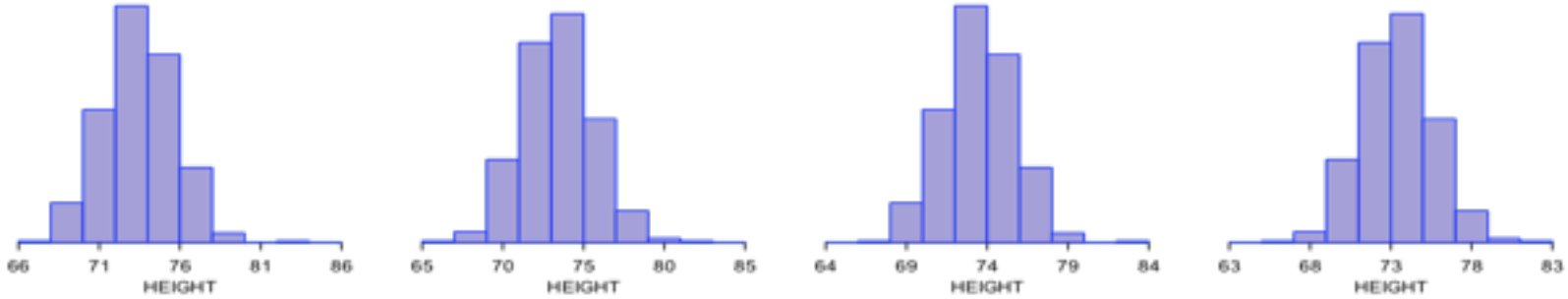


Histograms

Change bin width



Change location

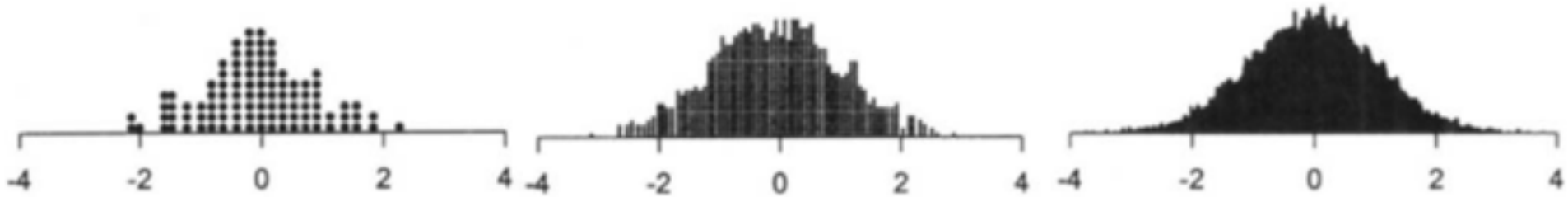


Dot Plots

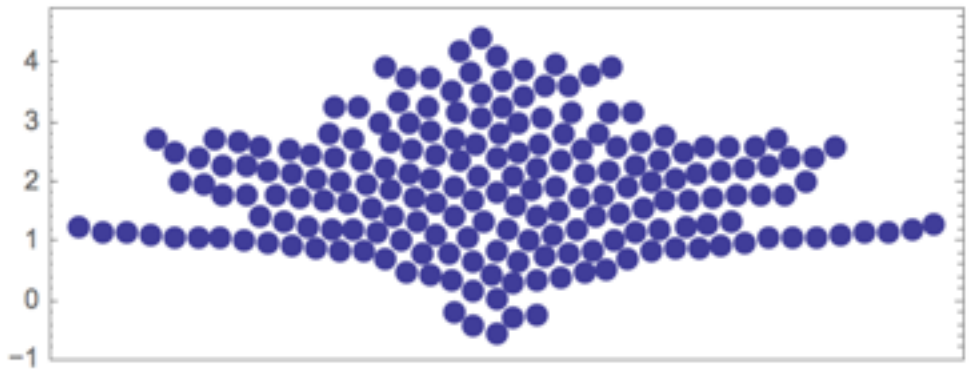
Originally for very small batches, but works on large

Facilitates brushing and linking

Wilkinson, L. (1999). Dot plots. *The American Statistician, 53*, 276–281.



Beeswarm plots ??



Kernel Density Plots

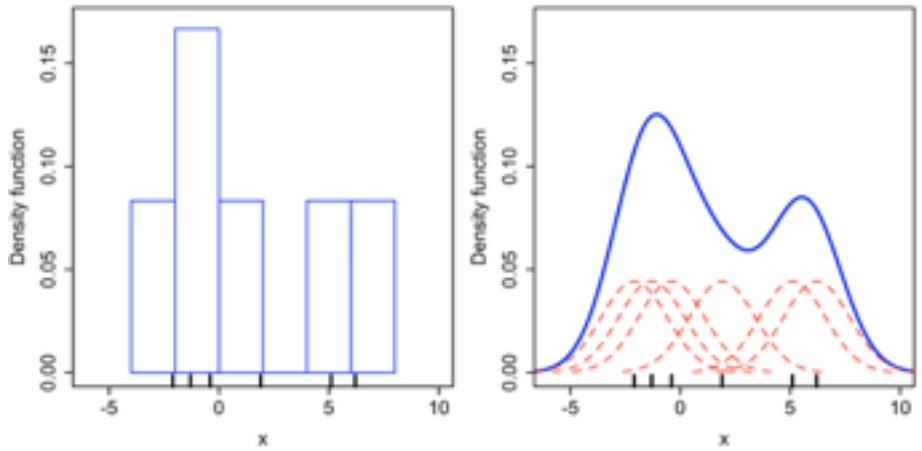
A convolution

Kernel is in red on lower right graphic (normal kernel), h is a bandwidth (smoothness) parameter

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Parzen, E. (1962). "On Estimation of a Probability Density Function and Mode." *The Annals of Mathematical Statistics*. 33(3): 1065–1076.

Dotplots are similar to kernel density estimation with a counting kernel rather than a continuous probability density function (h is the dot width).



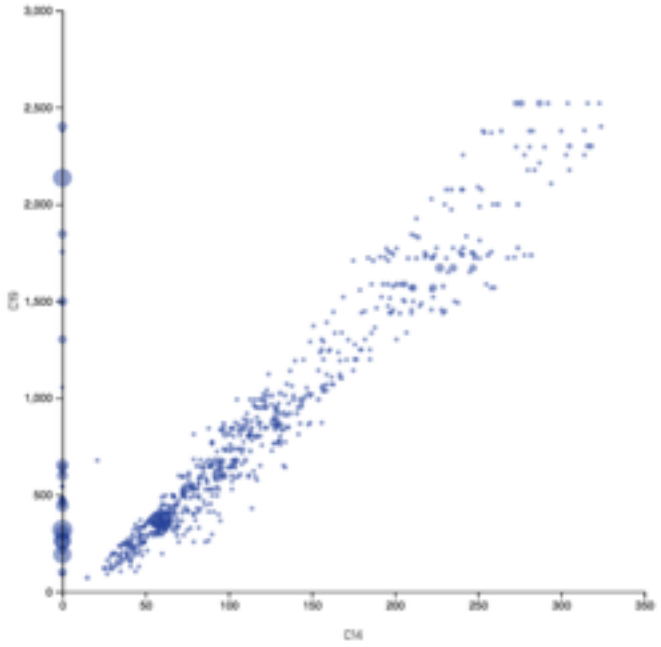
Scatterplots

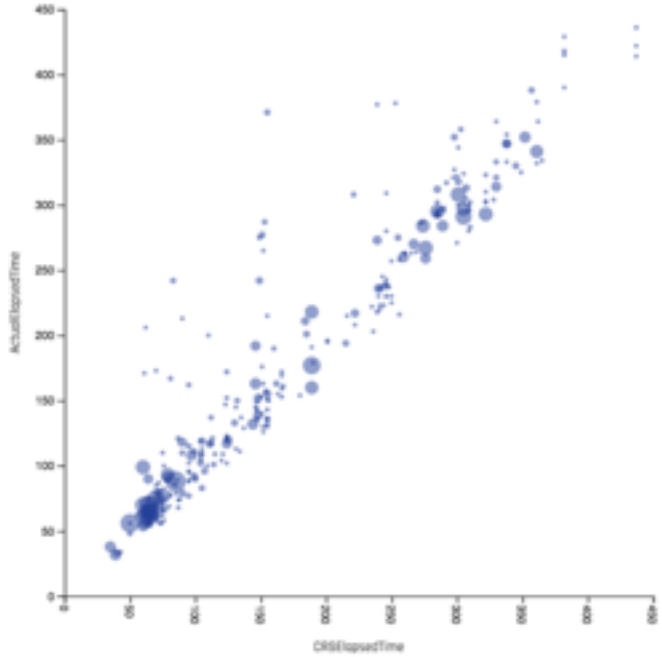
For aggregated data, we need to represent counts at each aggregated point

color, size, shape …

For size aesthetic, we sometimes call these bubble plots

but goal is to clamp sizes so the result looks like a plot of the raw data





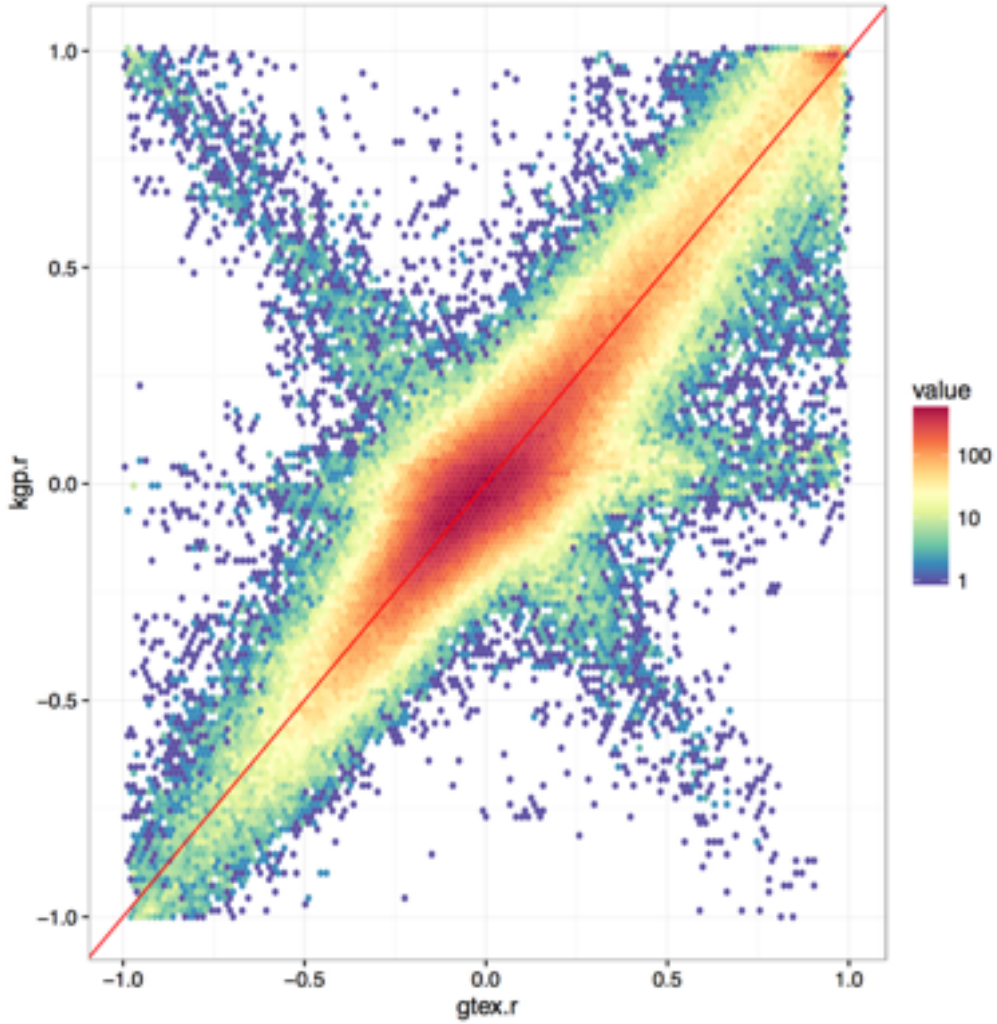
Scatterplots

An alternative is to use hex binning

this example is from ggplot2

Yukio Kosugi, Jun Ikebe, Nobuyuki Shitara, and Kintomo Takakura (1986). Graphical Presentation of Multidimensional Flow Histogram Using Hexagonal Segmentation. *Cytometry 7,* 291-294.

Carr, D.B., Littlefield, R.J., Nicholson, W.L., and Littlefield, J.S. (1987). Scatterplot matrix techniques for large N. *Journal of the American Statistical Association, 82*, 424–436.

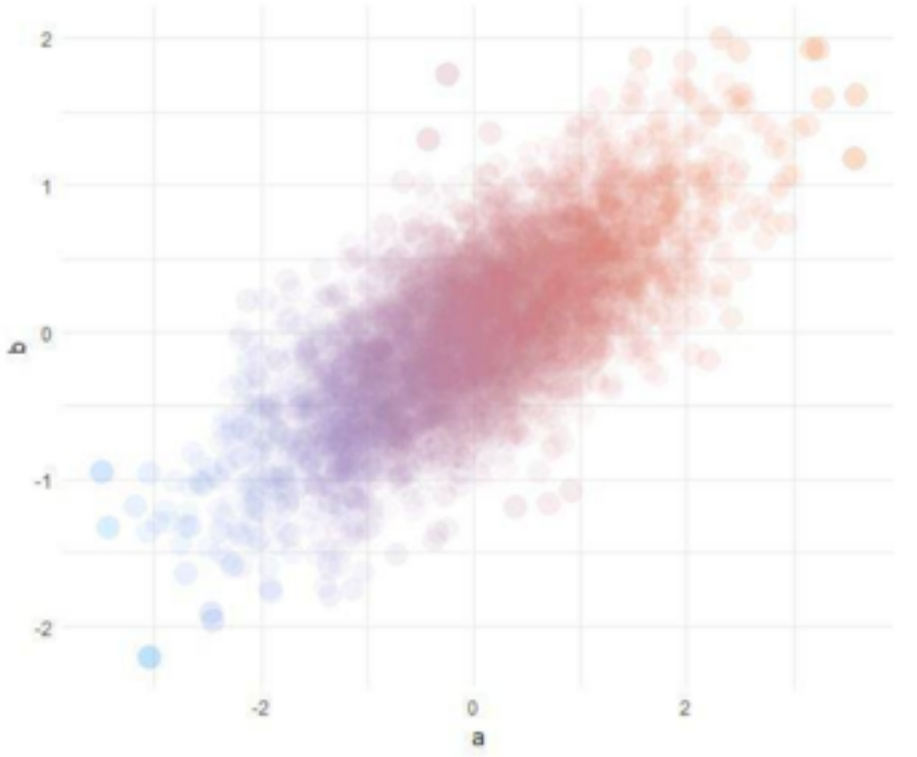


Scatterplots

Or alpha blending

again, ggplot2

```
ggplot(d, aes(a, b, color = pc, alpha = 1/density)) +
  geom_point(shape = 16, size = 5, show.legend = FALSE) +
  theme_minimal() +
  scale_color_gradient(low = "#0091ff", high = "#f0650e") +
  scale_alpha(range = c(.05, .25))
```

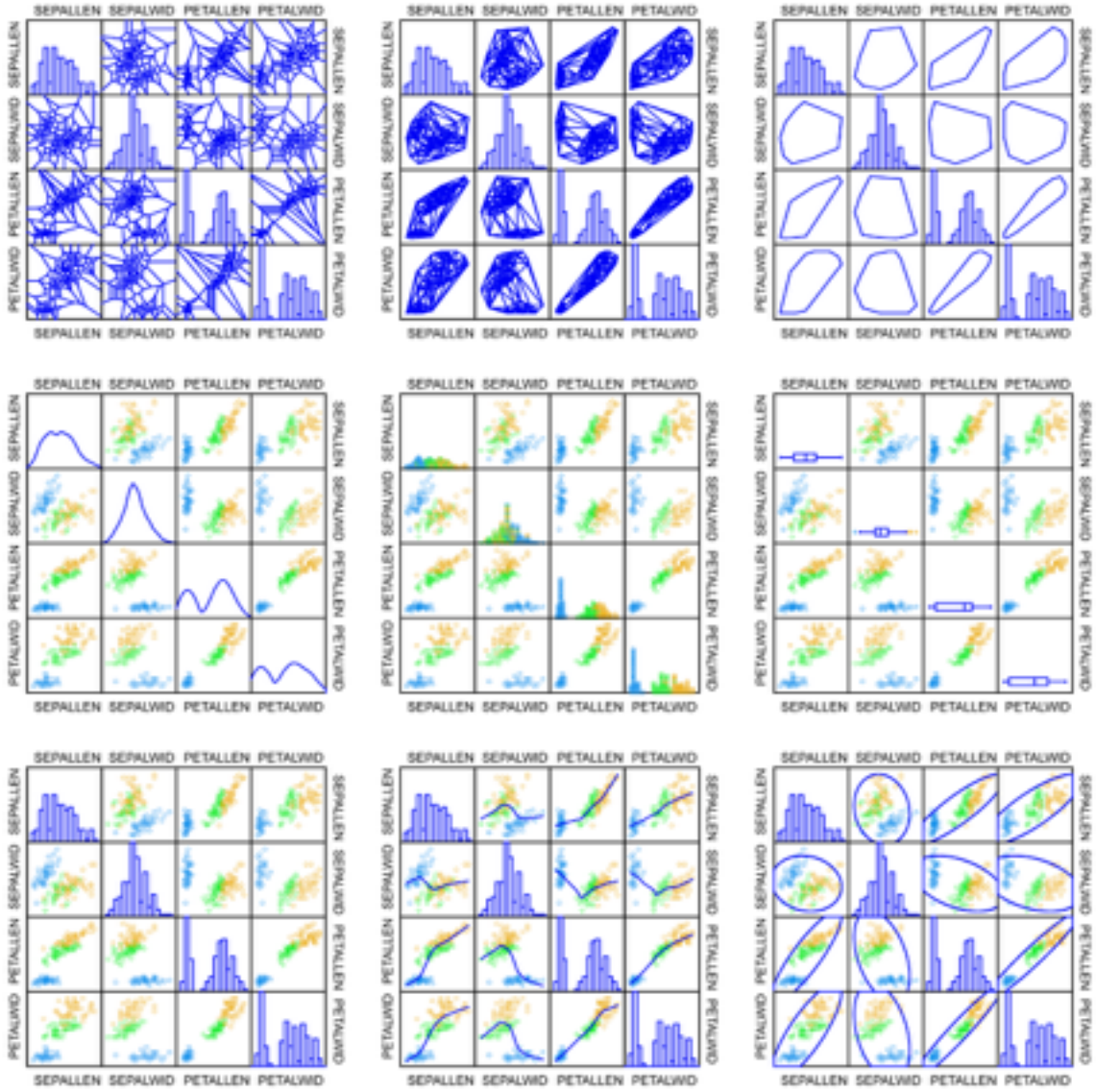


Scatterplot Matrices

Put anything in there

Hartigan, J.A. (1975). Printer graphics for clustering. *Journal of Statistical Computation and Simulation, 4,* 187–213.

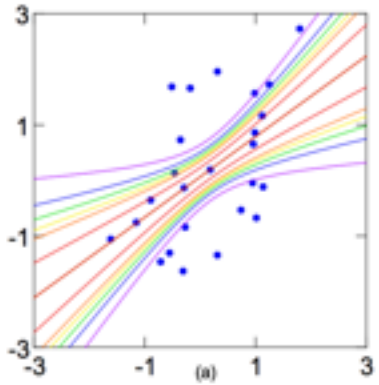
Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Monterey, CA: Wadsworth.

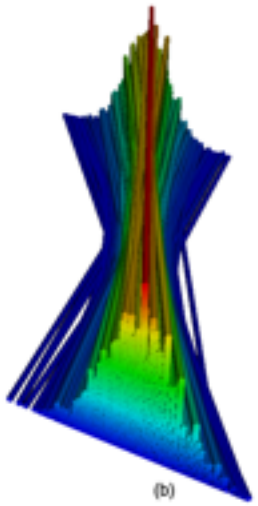


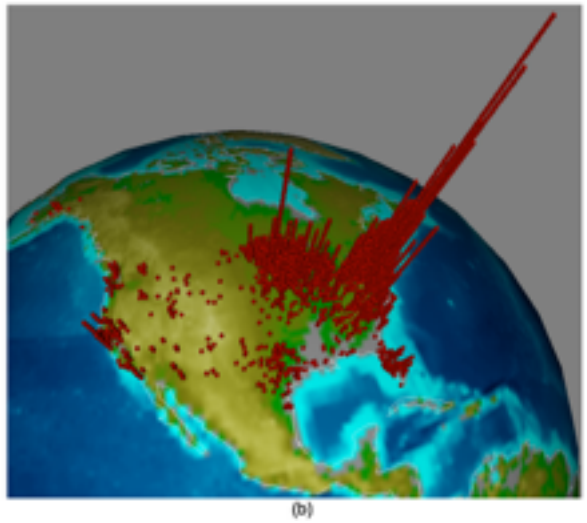
3D

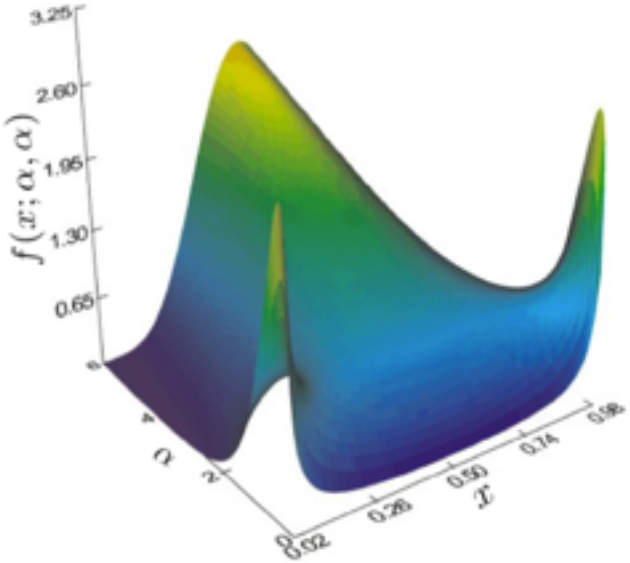
Don’t be afraid

Dang, T. N., Wilkinson, L., and Anand, A. (2010). Stacking graphic elements to avoid over-plotting. *Proceedings of the IEEE Symposium on Information Visualization 2010*, October, 23-25, Salt Lake City, UT.





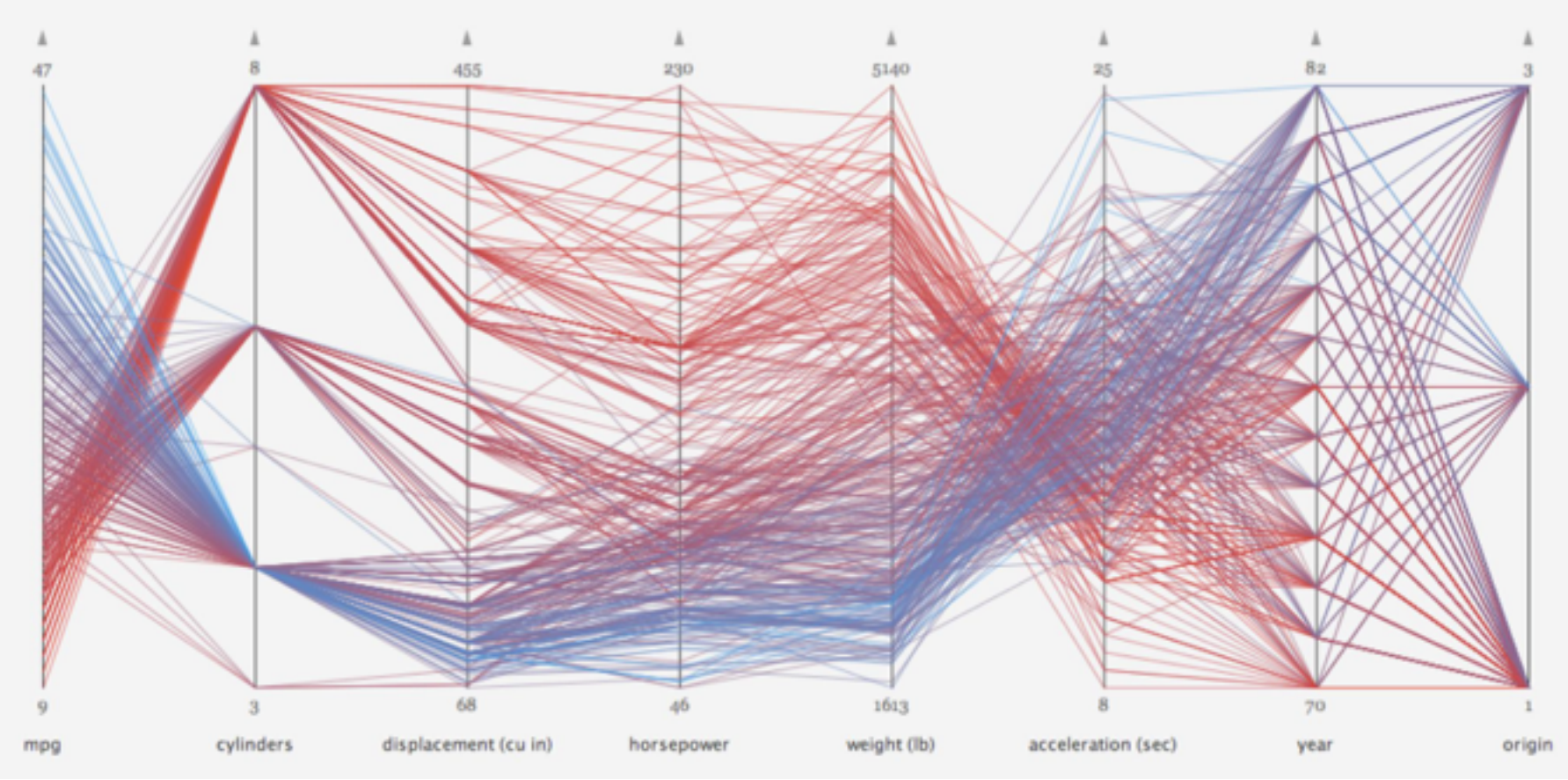




Parallel Coordinates

Continuous and Categorical Variables

Inselberg, A. (2009). *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. New York: Springer.

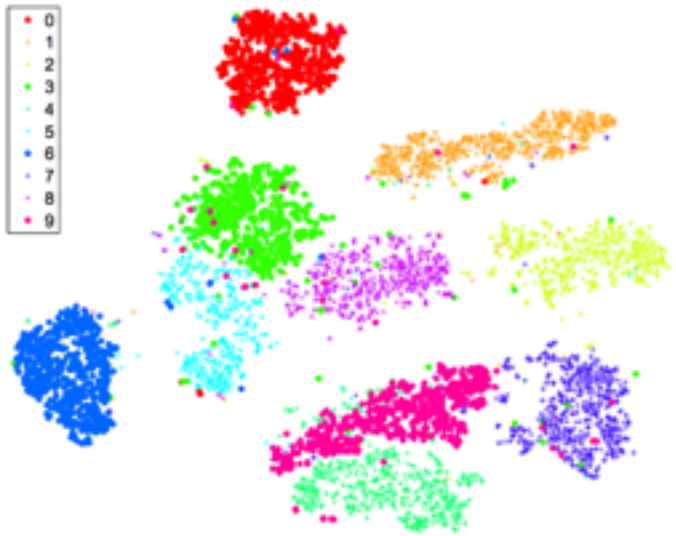


Manifold Learning

tSNE

L.J.P. van der Maaten and G.E. Hinton (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9, 2579-2605.

https://lvdmaaten.github.io/tsne/



tSNE

Sammon mapping

ISOMAP

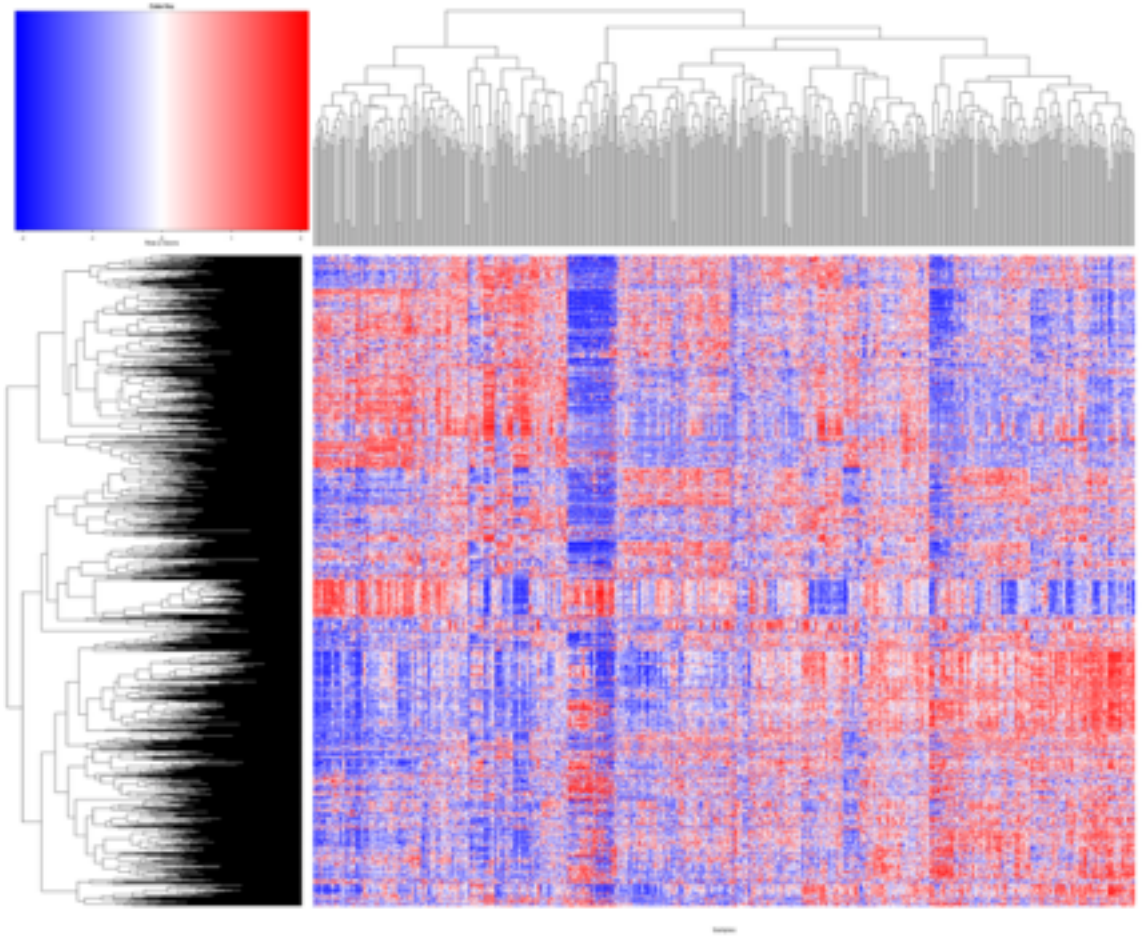
MNIST data: http://yann.lecun.com/exdb/mnist/index.html.

Heatmaps

Bohdan Bohdanovich Khomtchouk, Unpublished, Stanford.

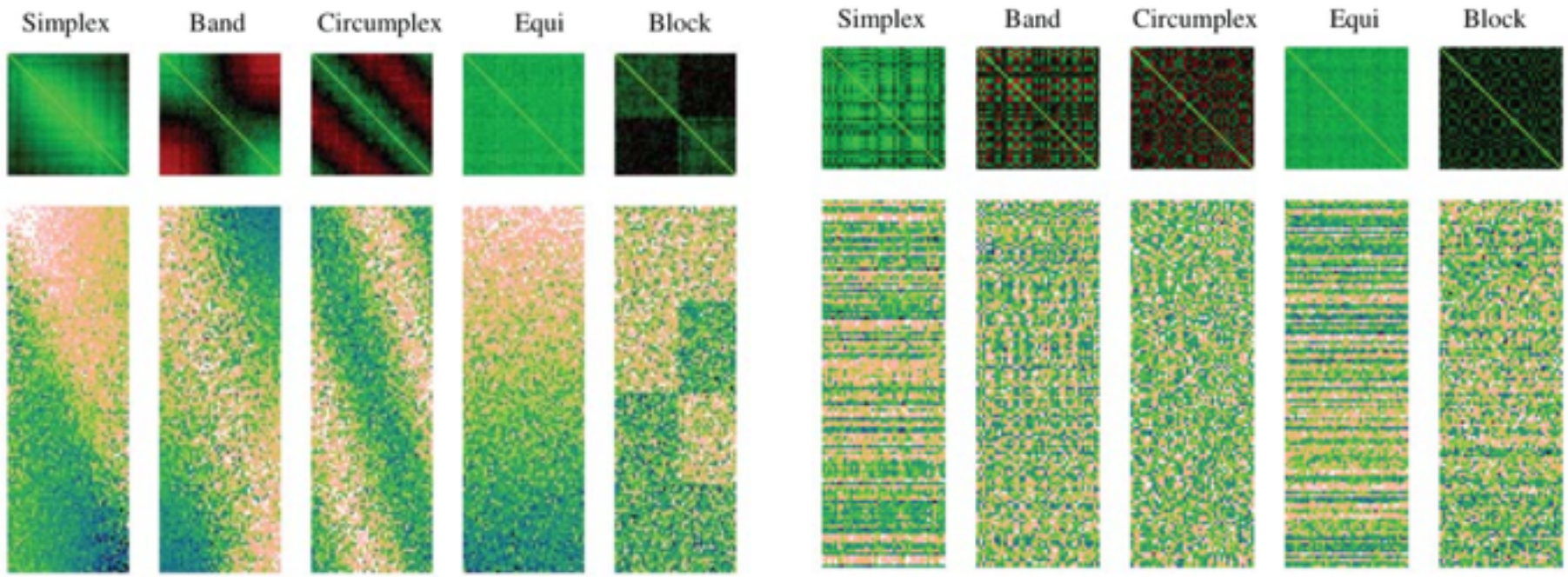
Behrisch, M., Schreck, T., and Pfister, H.P. (2019). GUIRO: User-Guided Matrix Reordering, VisWeek, TVCG.

Wilkinson, L. and Friendly, M (2009). The History of the Cluster Heat Map. *The American Statistician 63*(2), 179-184.

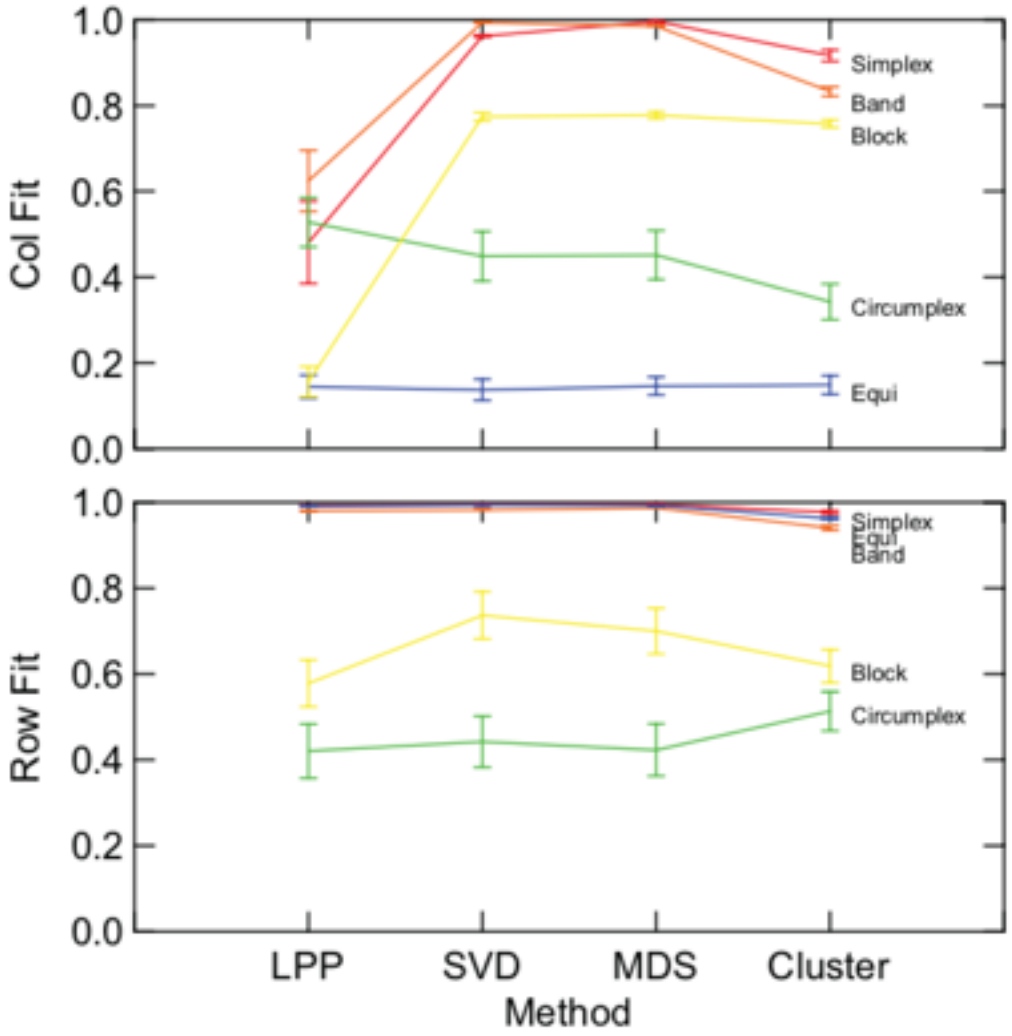


Heatmaps

Wilkinson, L. (2008). *The Grammar of Graph*ics. New York:Springer.



Heatmaps


Col Fit
1.0
0.8
0.6
0.4
0.2
0.0
Simplex
Band
Block
Circumplex
Equi
Row Fit
1.0
0.8
0.6
0.4
0.2
0.0
Simplex
Equi
Band
Block
Circumplex
LPP
SVD
MDS
Cluster
Method

3D Spinning

Prim-9

Donoho, A.W., Donoho, D.L., and Gasko, M. (1988). MacSpin: Dynamic graphics on a desktop computer. In W.S. Cleveland and M.E. McGill, (Eds.), *Dynamic Graphics for Statistics* (pp. 331–351). Belmont, CA: Wadsworth.

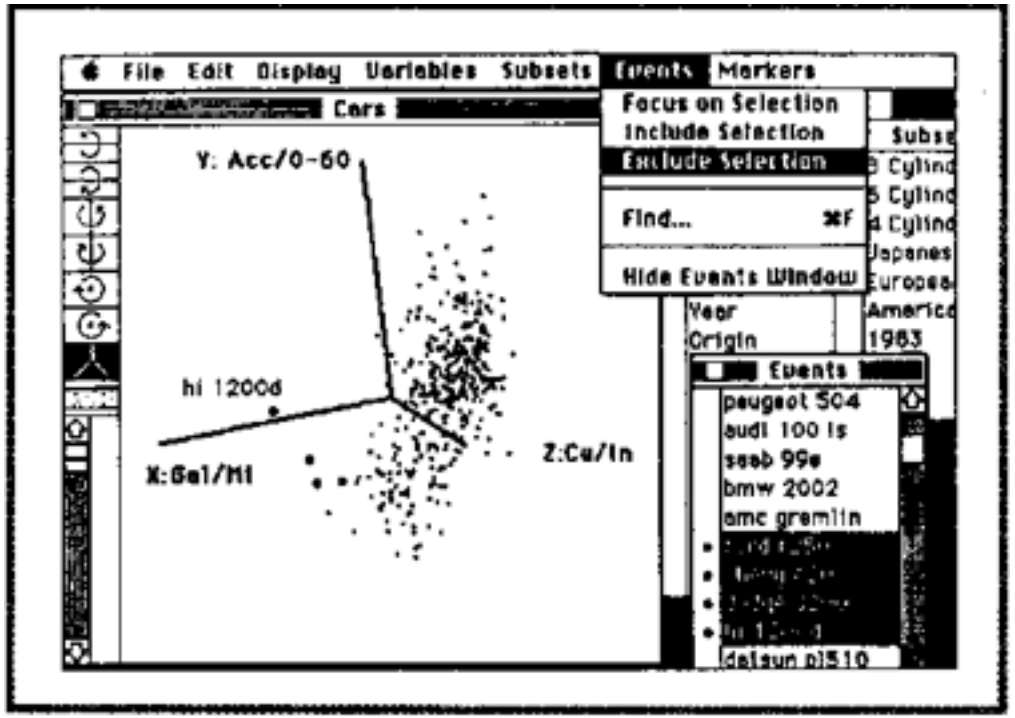
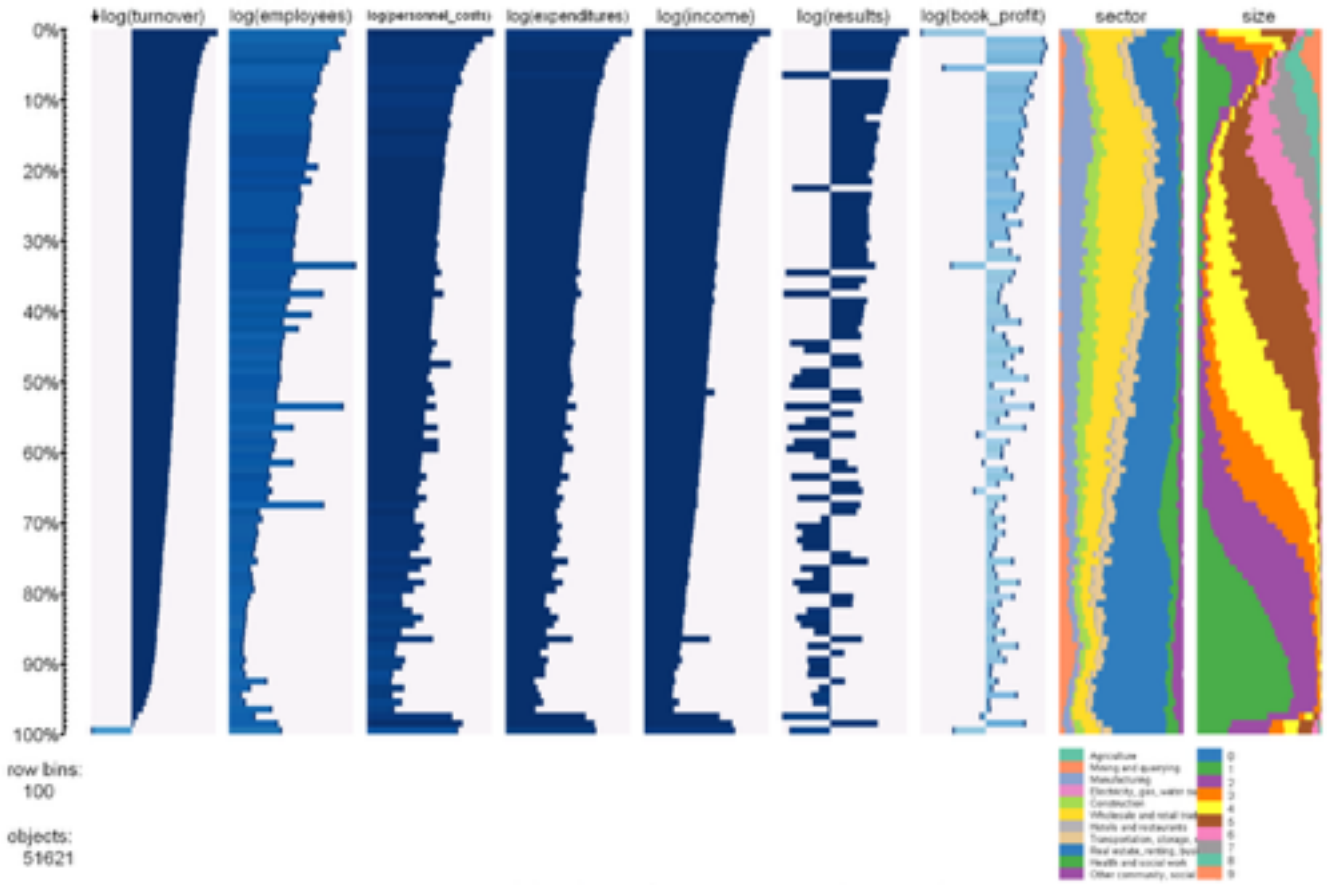


Table Plot

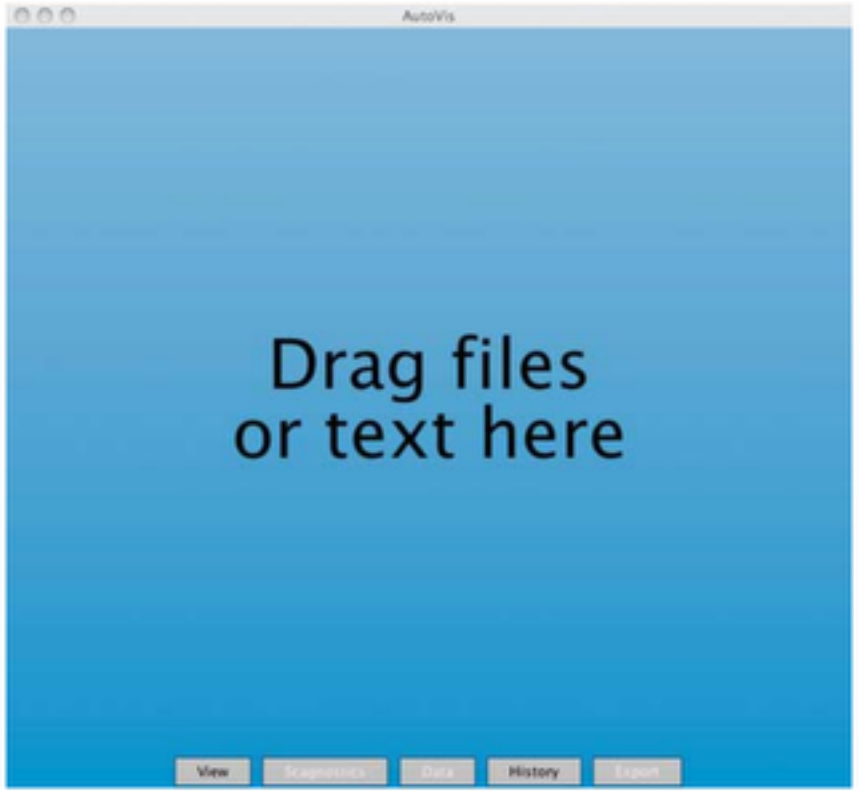
Malik, Unwin, Gribov (2010).

Malik, W.A. et al. (2010) An Interactive Graphical System for Visualizing Data Quality: Tableplot Graphics. In H. Loracek-Junge & C. Weihs (eds.), *Classification as a Tool for Research*, Proceedings of the 11th IFCS Conference. Berlin: Springer, 331-339.

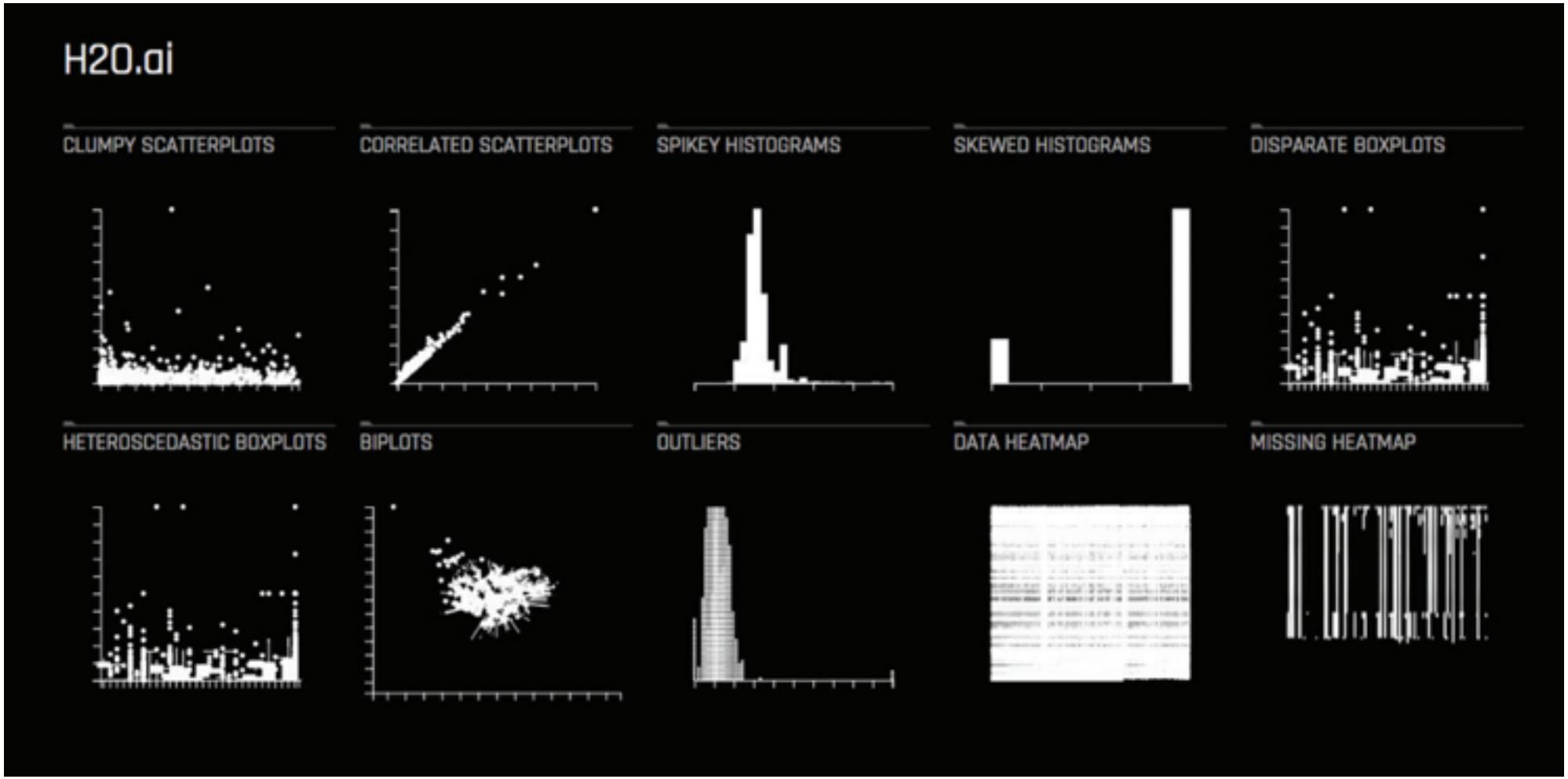


AutoVis

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



H2O AutoViz


H2O.ai
CLUMPY SCATTERPLOTS
CORRELATED SCATTERPLOTS
SPIKEY HISTOGRAMS
SKEWED HISTOGRAMS
DISPARATE BOXPLOTS
HETEROSCEDASTIC BOXPLOTS
BIPLOTS
OUTLIERS
DATA HEATMAP
MISSING HEATMAP

References

Tukey, J. W. (1962). The Future of Data Analysis. *Ann. Math. Statist. 33* (1), 1-67.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statist. Sci. 16* (3), 199-231.

Friedman, J. (2001). Data Mining and Statistics: What's the connection? *Proc. 29th Symposium on the Interface*.

Cleveland, W.S. (2001). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review, 69* (1), 21-26.

Donoho, D. (2017) 50 Years of Data Science. *Journal of Computational and Graphical Statistics* 26 (4), 745-766.

Thank You!