Visualization Frameworks to Detect the Presence of Unusual Structures in High-Dimensional Data

Jeff Solka
NSWDD and George Mason University
February 19, 2004

Acknowledgements

- **Ed Wegman**
- **Dan Carr**
- **Helpful discussions, pictures, software**

Benefits of Visualization

- **Software debugging**
- **Algorithmic understanding**
- **New applications of algorithms**
- **Data mining and exploration**
- **Model mining and exploration**

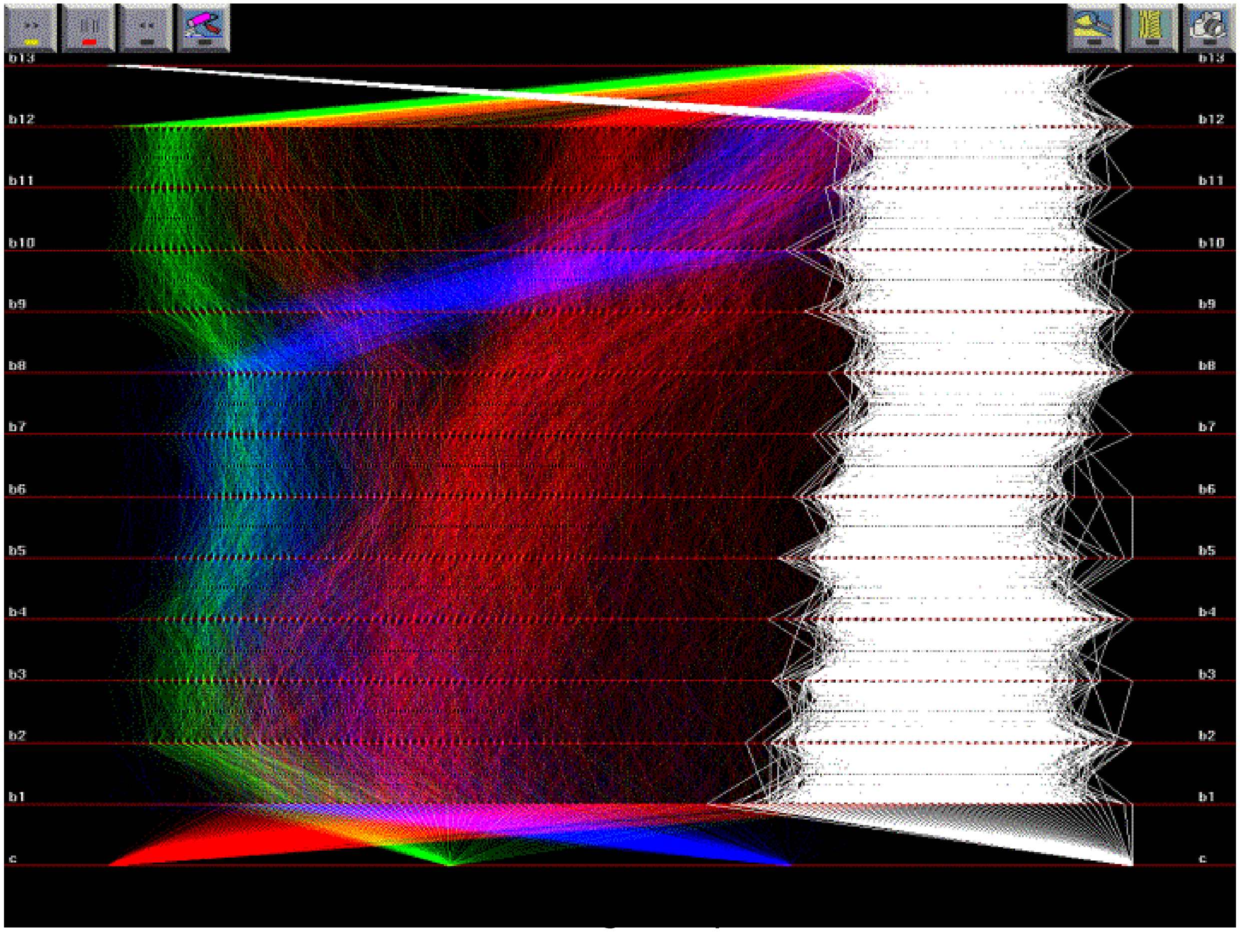
Biosurveillance Needs and Visualization Strategies

- High-dimensional Data Visualization
- Outlier/Anomaly Detection
- Cluster Detection
- Methods of Hypothesis Generation and Testing
- Spatial Statistics Smoothing Methods
- Disparate Data Fusion Methods
- Meta Data Access and Organization
- Modeling
- Collaborative Environments
- Denoising

- Parallel Coordinates
- Brushing on Linked Plots
- Grand Tour
- Dotplots
- Distributional Summaries
- Micromaps
- Ordered Data Images
- Ordered Interpoint Distance Images
- Integrated Visualization Environments
  - Overview and drill-down
  - Data analysis management

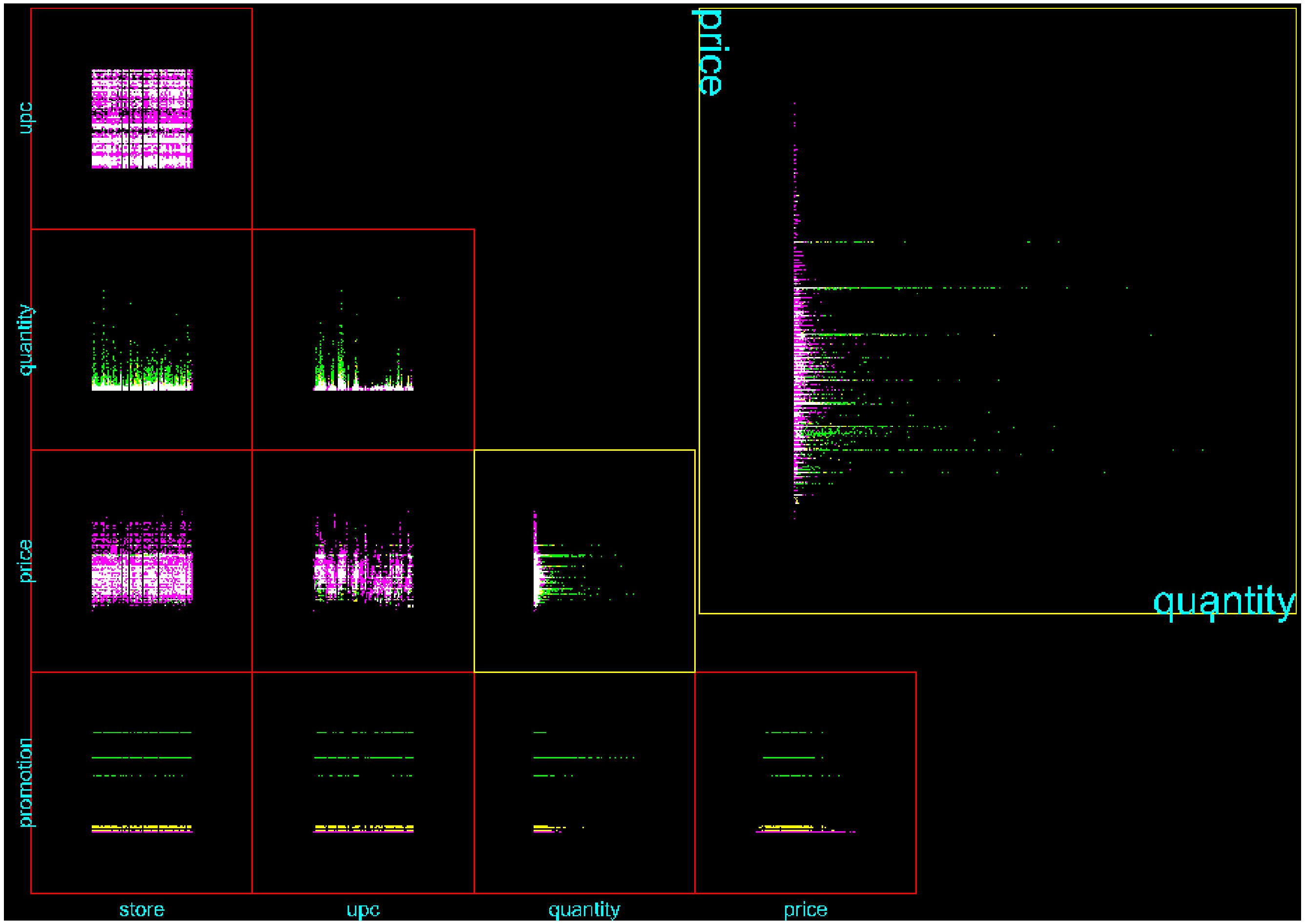
SALAD Data

Parallel coordinates for cluster identification and discriminant analysis facilitation



Scanner Data for Breakfast Cereals
(April 1999)

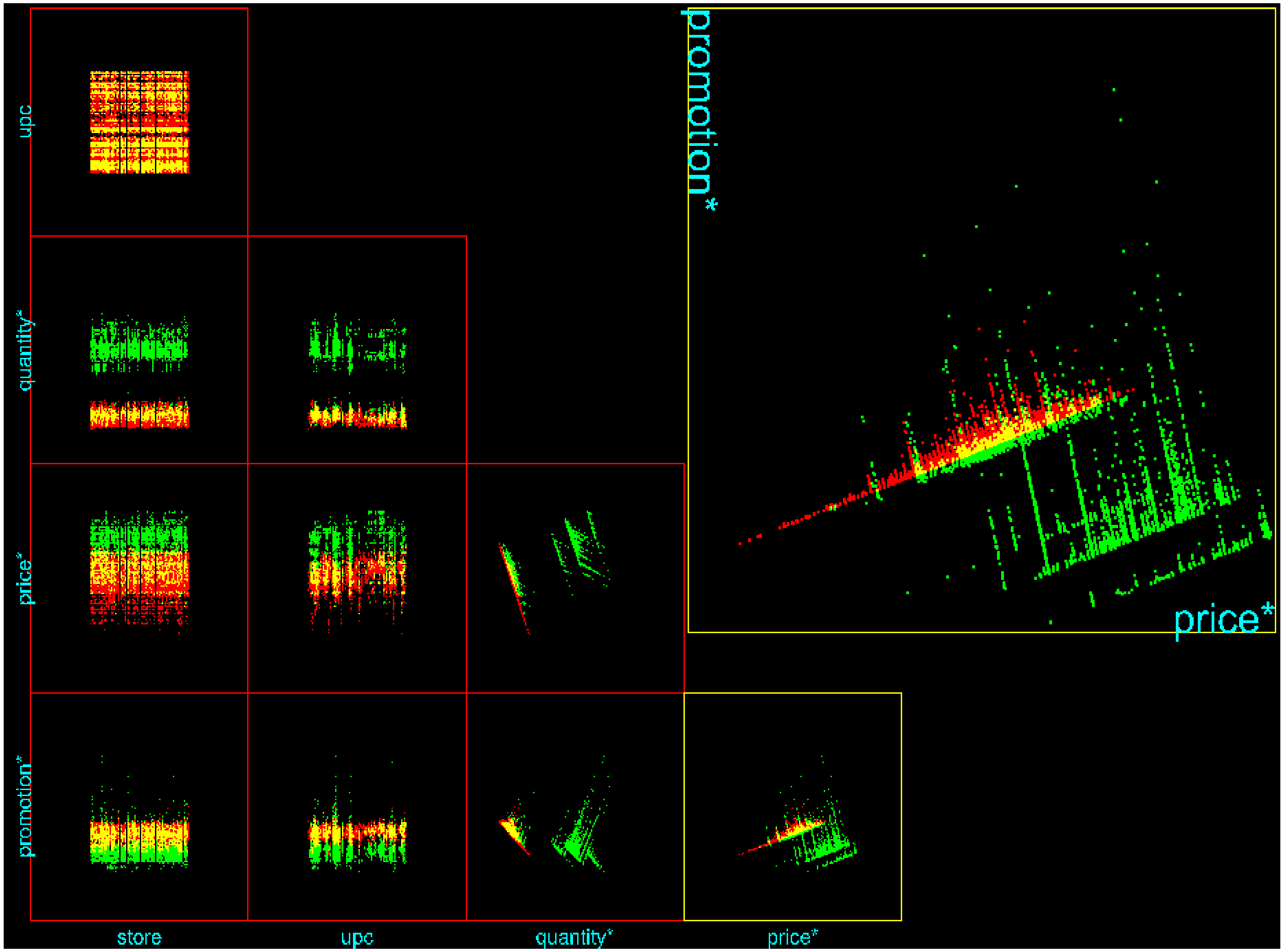
Reprinted with permission from "Visualizing Cereal World," Ed Wegman, George Mason University (ewegman@gmu.edu)



Promotion has huge impact on sales volume

Scanner Data for Breakfast Cereals
(April 1999)

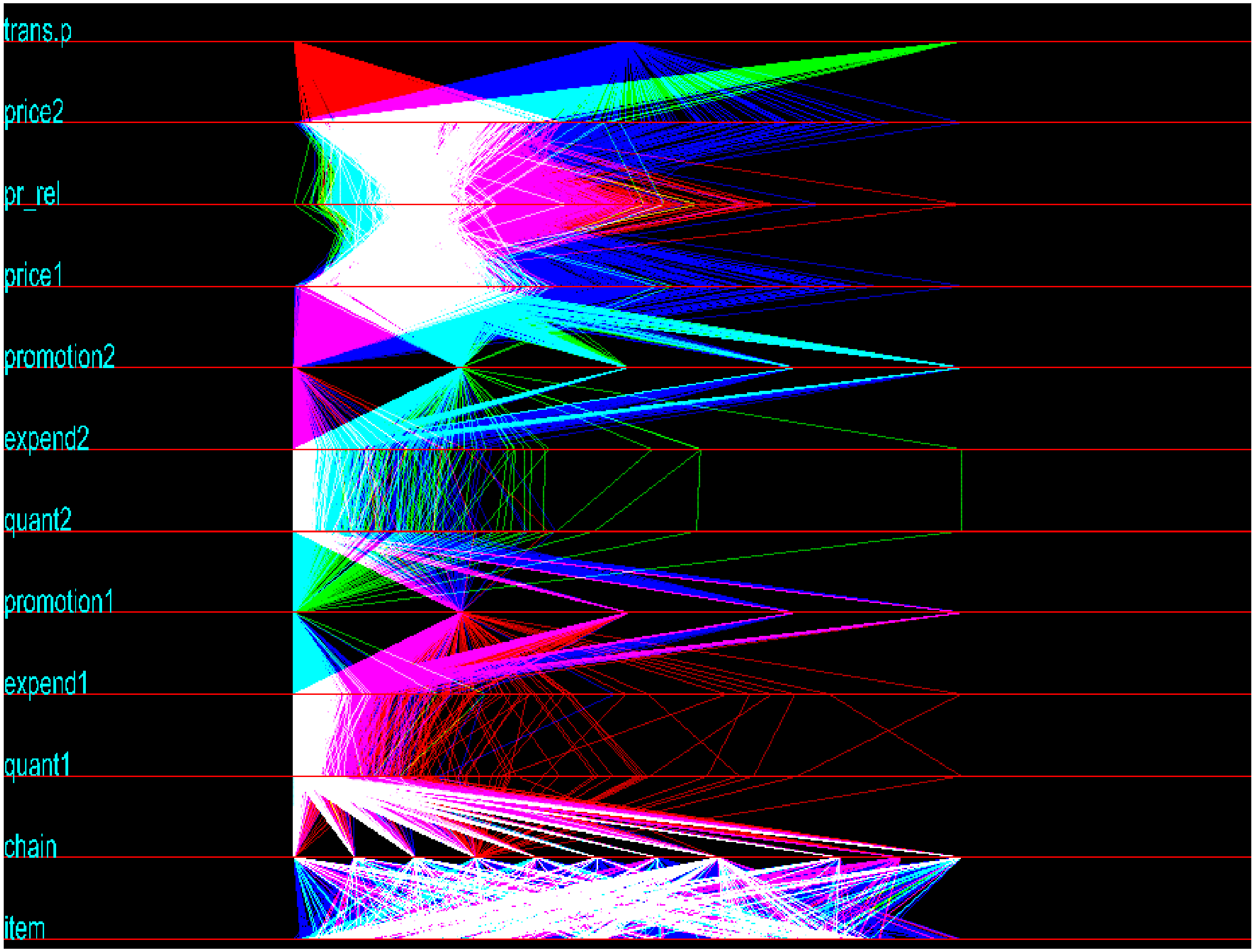
Reprinted with permission from "Visualizing Cereal World," Ed Wegman, George Mason University (ewegman@gmu.edu)



Aggressive promotion pays

Scanner Data for Breakfast Cereals

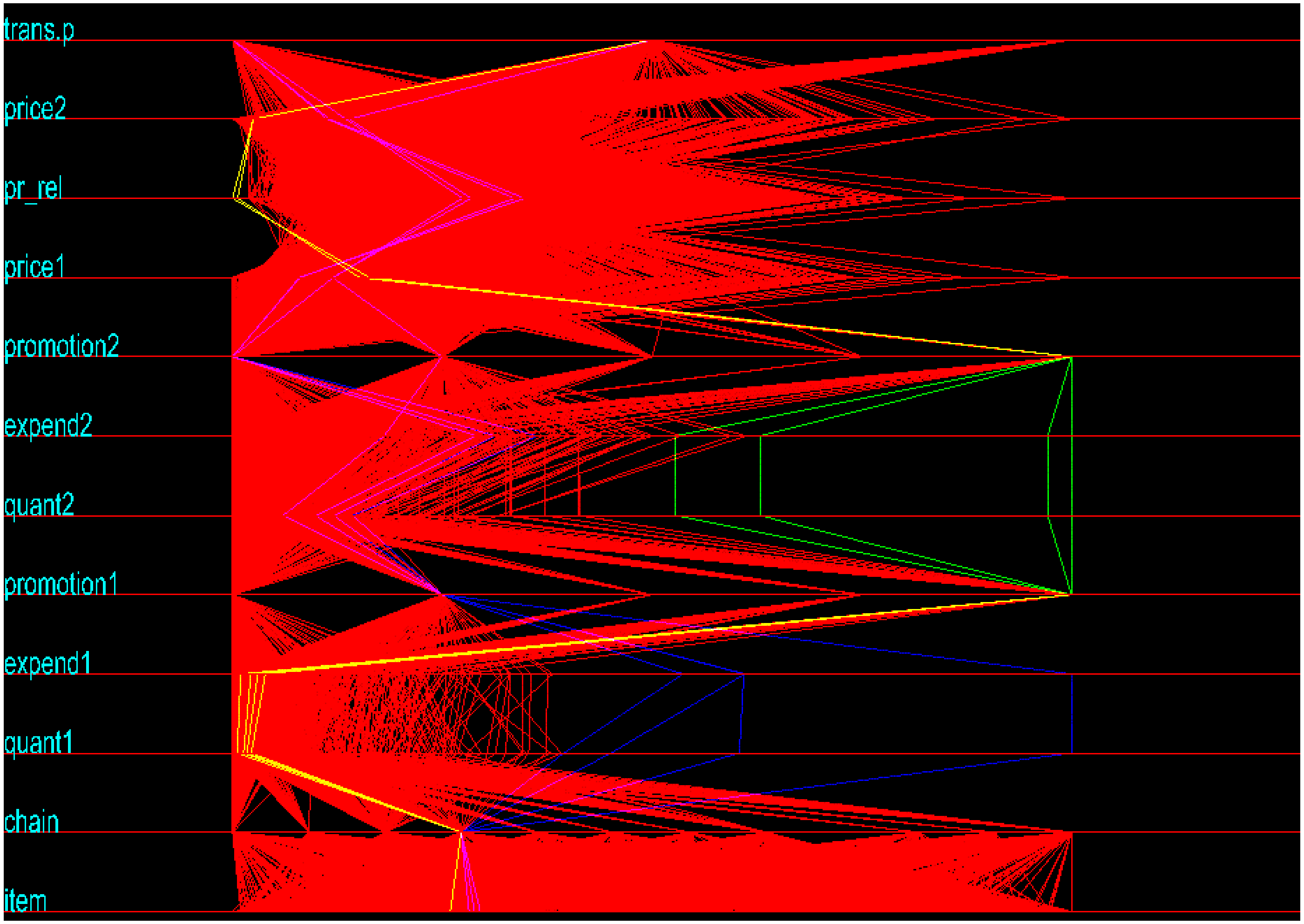
Reprinted with permission from "Visualizing Cereal World," Ed Wegman, George Mason University (ewegman@gmu.edu)



Promotion of Cereals Creates Outliers

Scanner Data for Breakfast Cereals

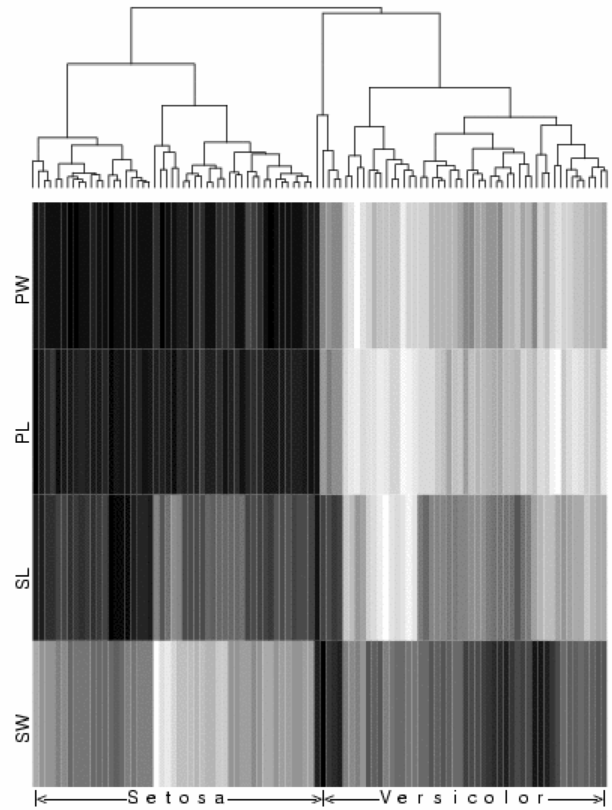
Reprinted with permission from "Visualizing Cereal World," Ed Wegman, George Mason University (ewegman@gmu.edu)



Outliers belong to same chain

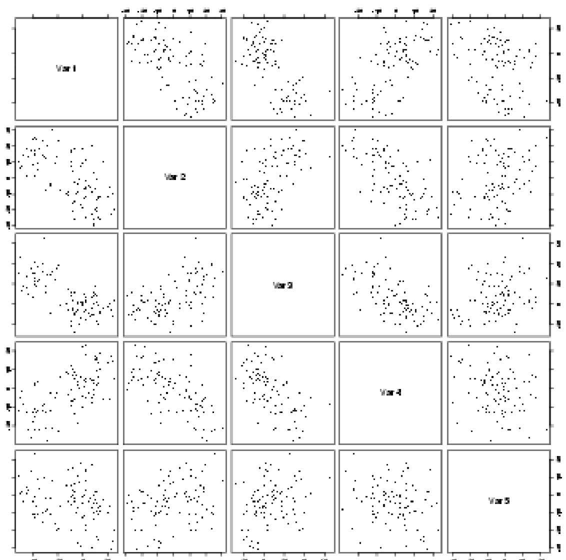
Data Color Histogram or Data Image

Orderings of the data matrix were first discussed in Bertin. Wegman in 1990 coined the term "data color histogram." Mike Minnotte and Webster West subsequently termed the phrase "data image" in 1998.

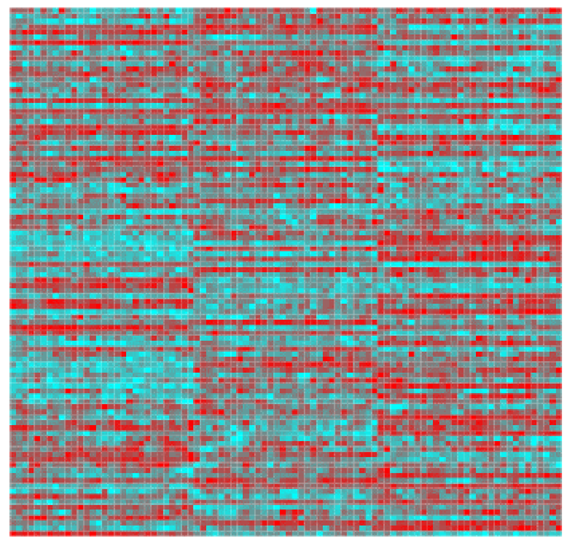


Data Image Reveals Obfuscated Cluster Structure

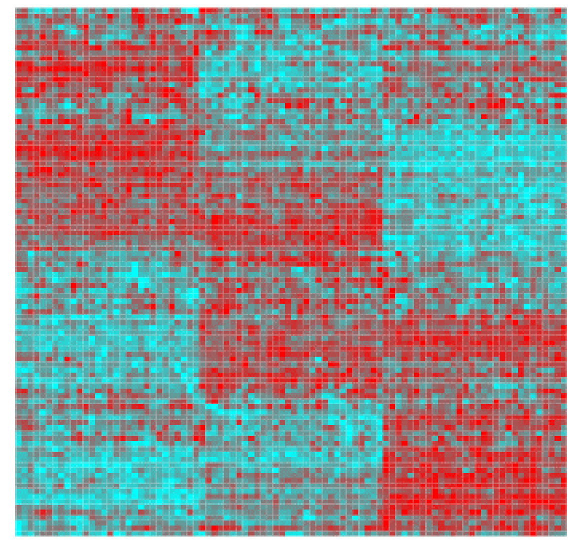
Subset of the pairs plot



Sorted on Observations

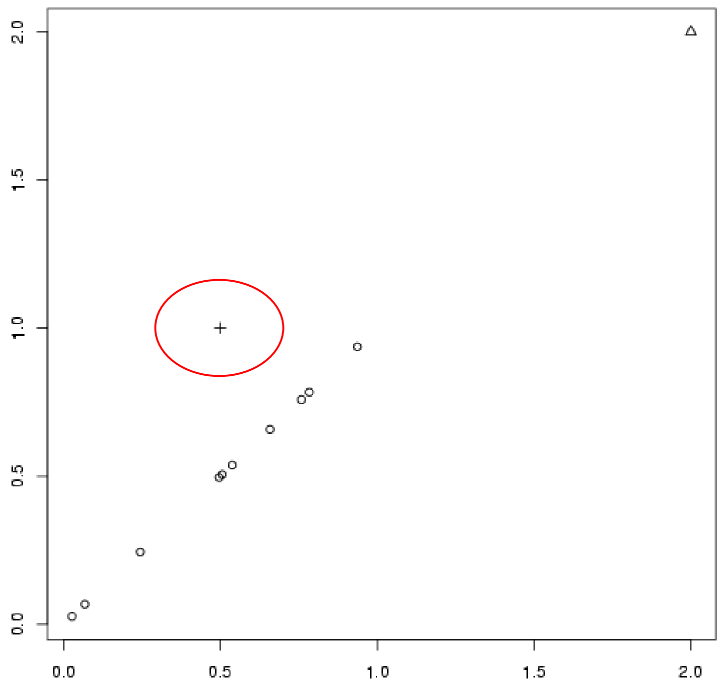


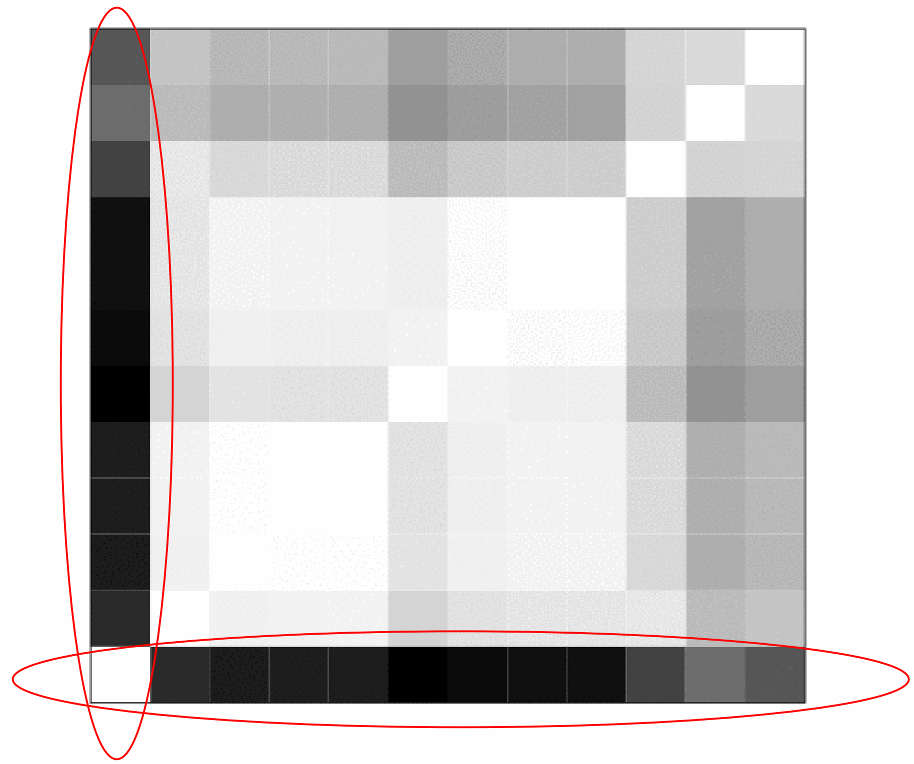
Sorted on Observations and Features



90 observations in R^{100} drawn from a standard normal distribution
The first and second 30 rows were shifted by 20 in their first and second dimensions respectively. This data matrix was then multiplied by a 100 x 100 matrix of Gaussian noise.

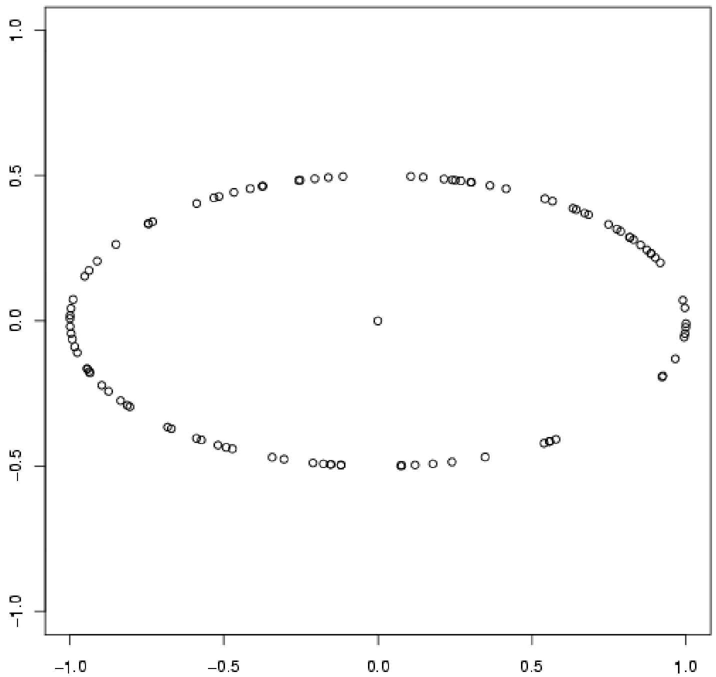
Sorted Data Image of the Interpoint Distance Matrix

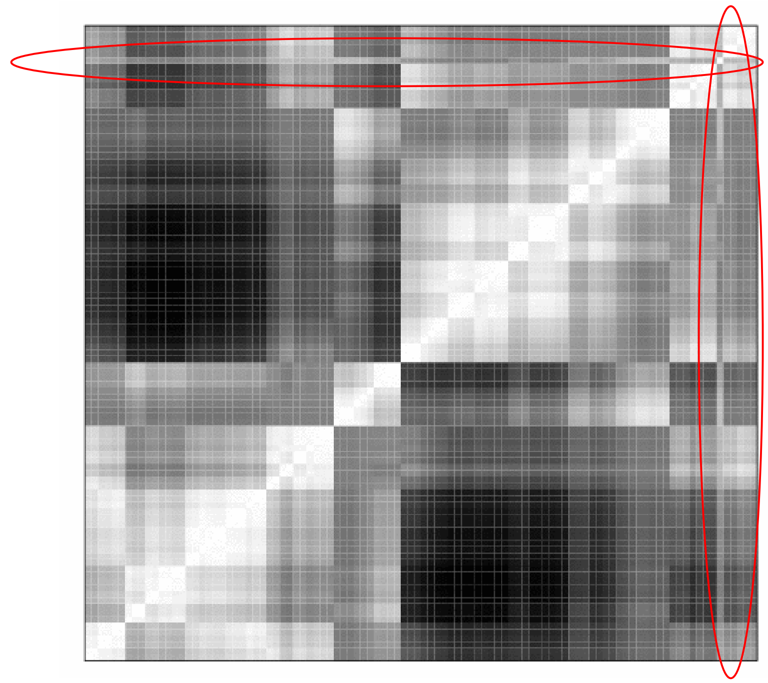




Outlier

Sorted Data Image of the Interpoint Distance Matrix Reveals "Subtle" Outliers





Outlier

Distance Measures on Manifolds

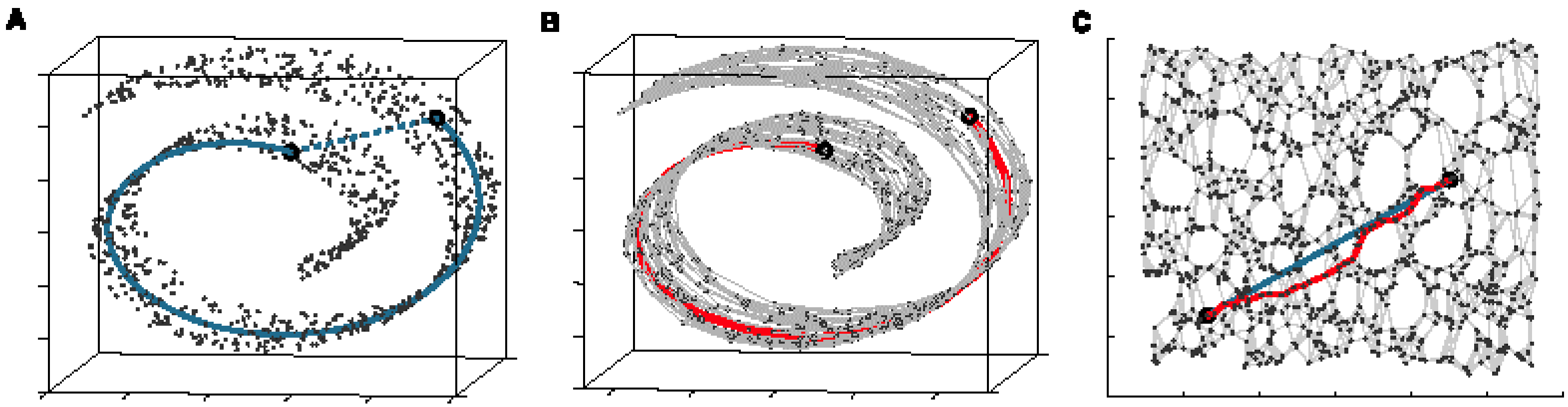
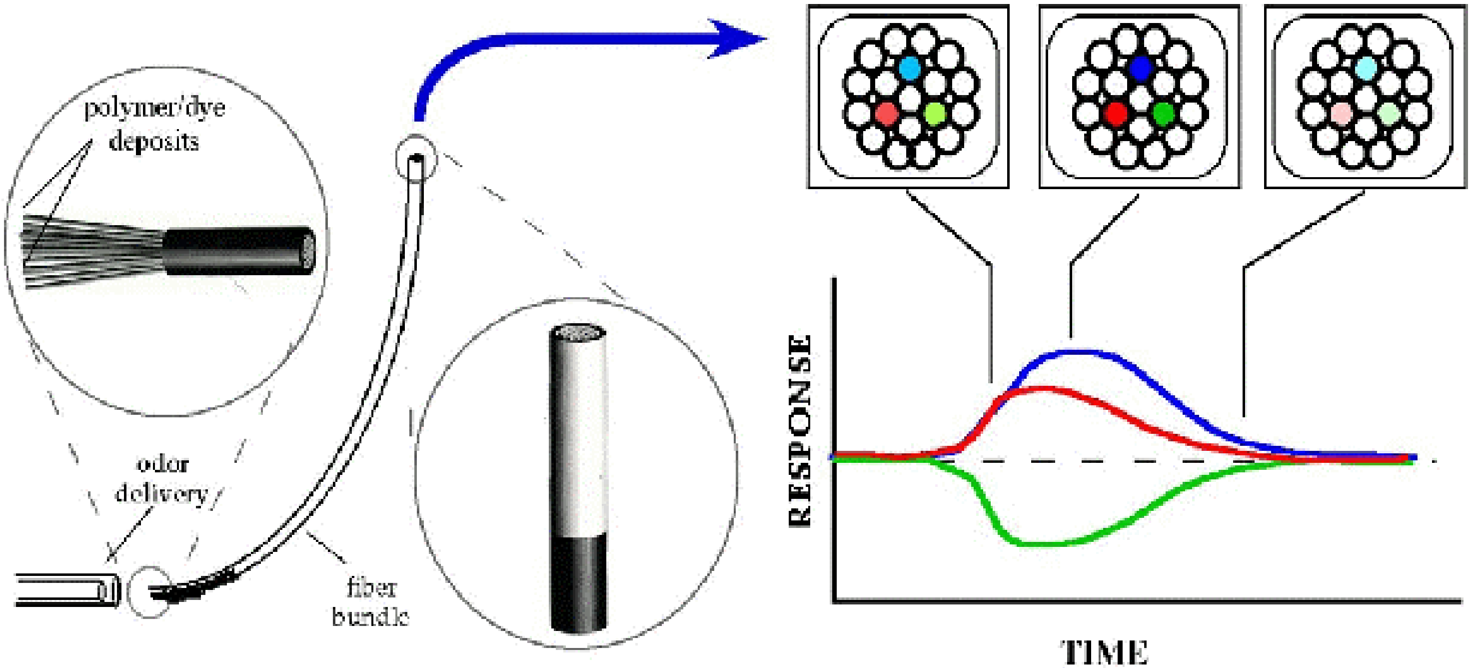


Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (**A**) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect their intrinsic similarity, as measured by geodesic distance along the low-dimensional manifold (length of solid curve). (**B**) The neighborhood graph *G* constructed in step one of Isomap (with $K = 7$ and $N = 1000$ data points) allows an approximation (red segments) to the true geodesic path to be computed efficiently in step two, as the shortest path in *G*. (**C**) The two-dimensional embedding recovered by Isomap in step three, which best preserves the shortest path distances in the neighborhood graph (overlaid). Straight lines in the embedding (blue) now represent simpler and cleaner approximations to the true geodesic paths than do the corresponding graph paths (red).

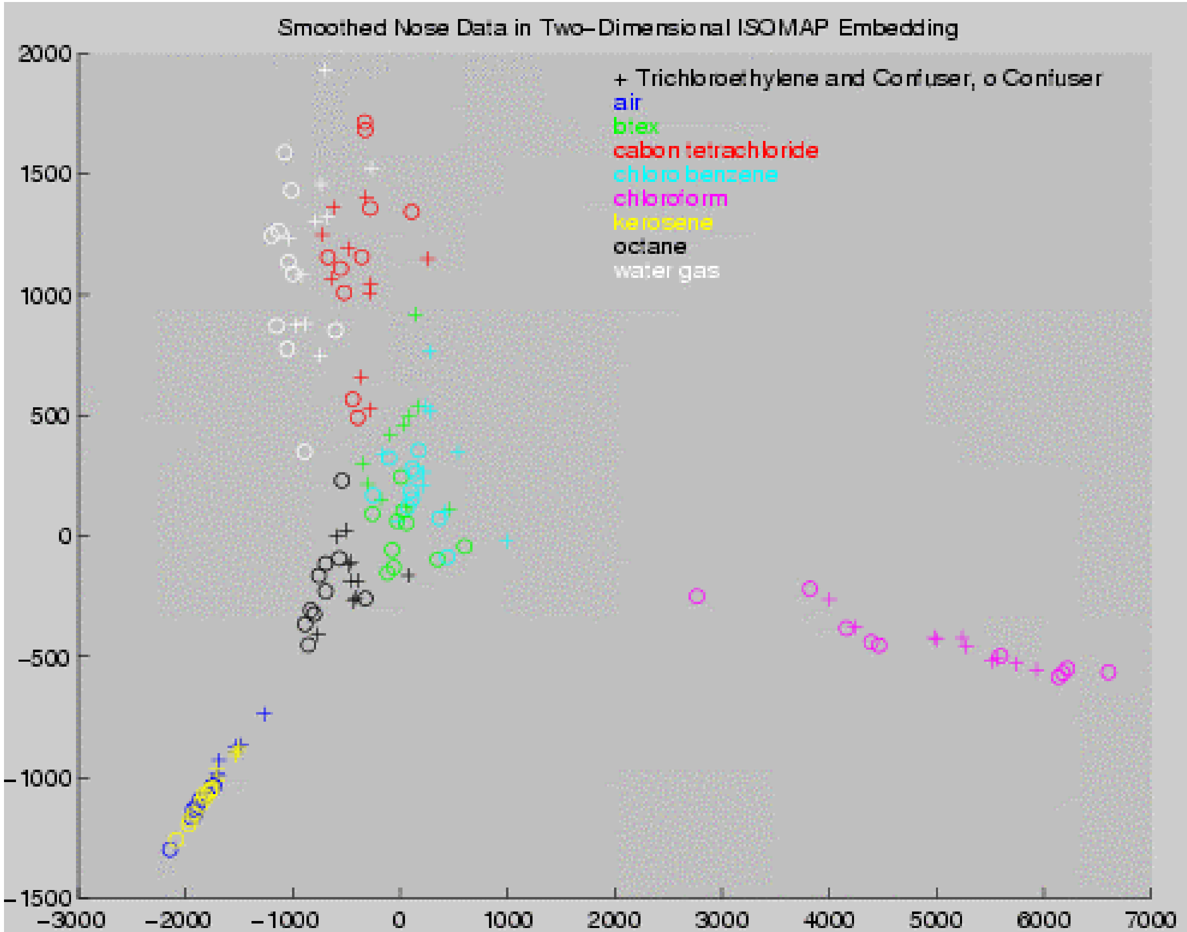
A Global Geometric Framework for Nonlinear Dimensionality Reduction, Joshua B. Tenenbaum, Vin de Silva, John C. Langford

Artificial Nose Data Set



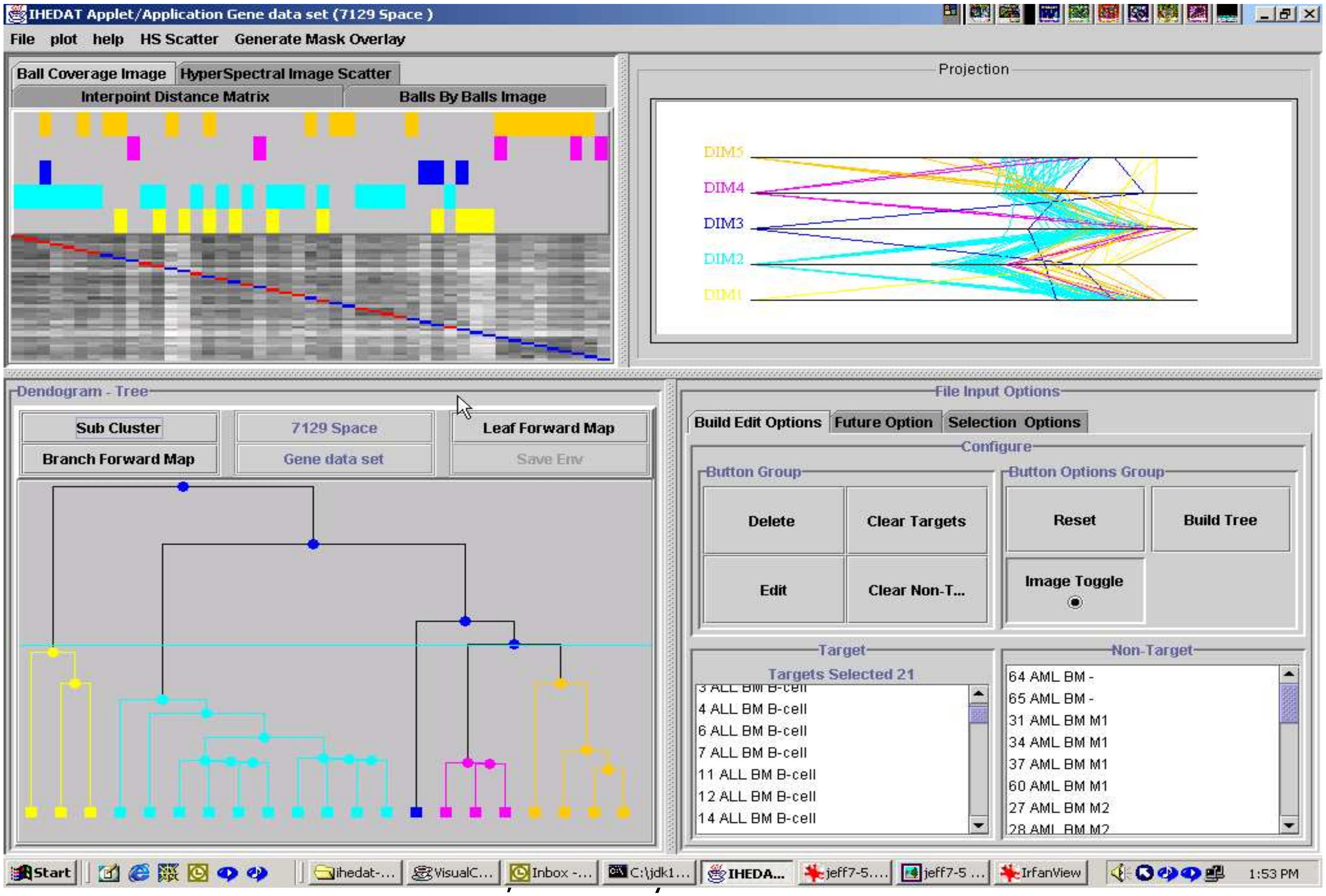
19 fibers x 2 wavelengths x 60 samples/sniff = R^{2280}

Artificial Nose ISOMAP Results

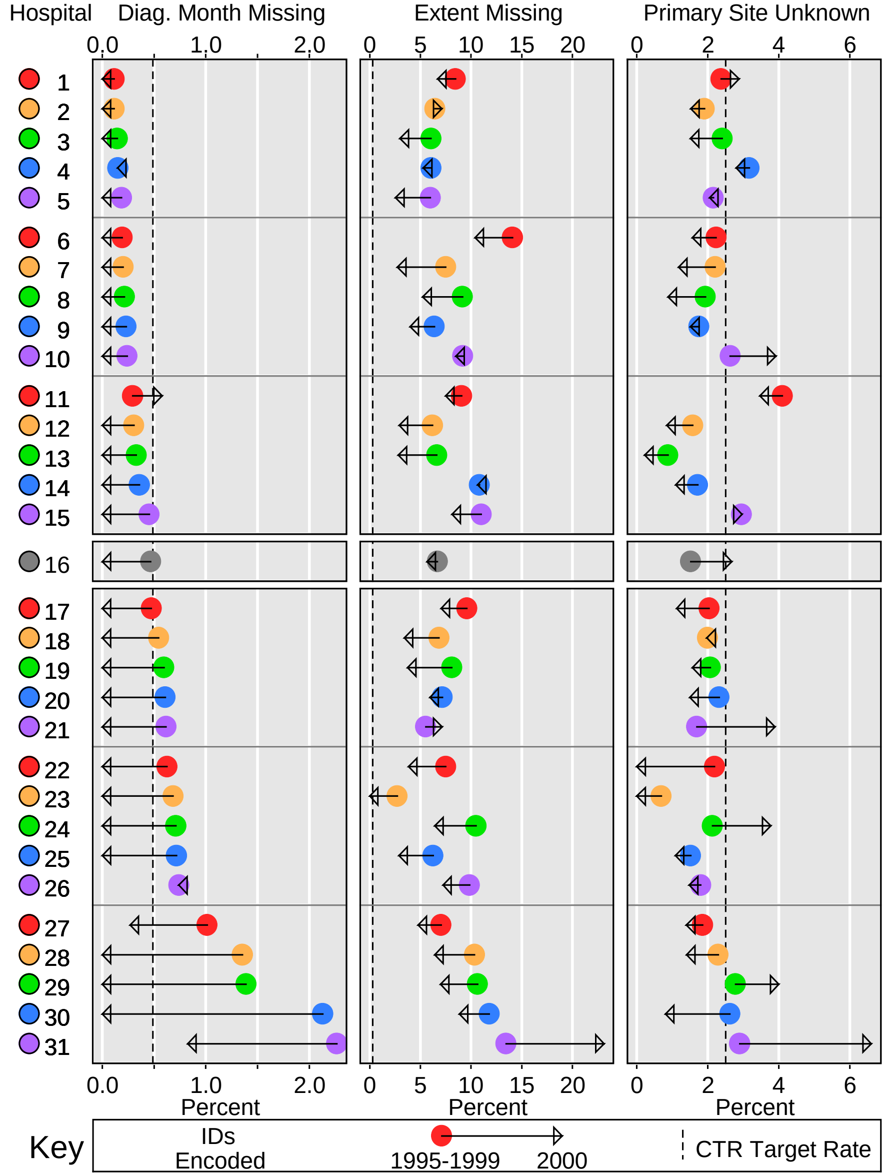


Interactive Hyperspectral Exploratory Data Analysis Tool

Integrated visualization environments can lead to new insights in the hands of skilled users.

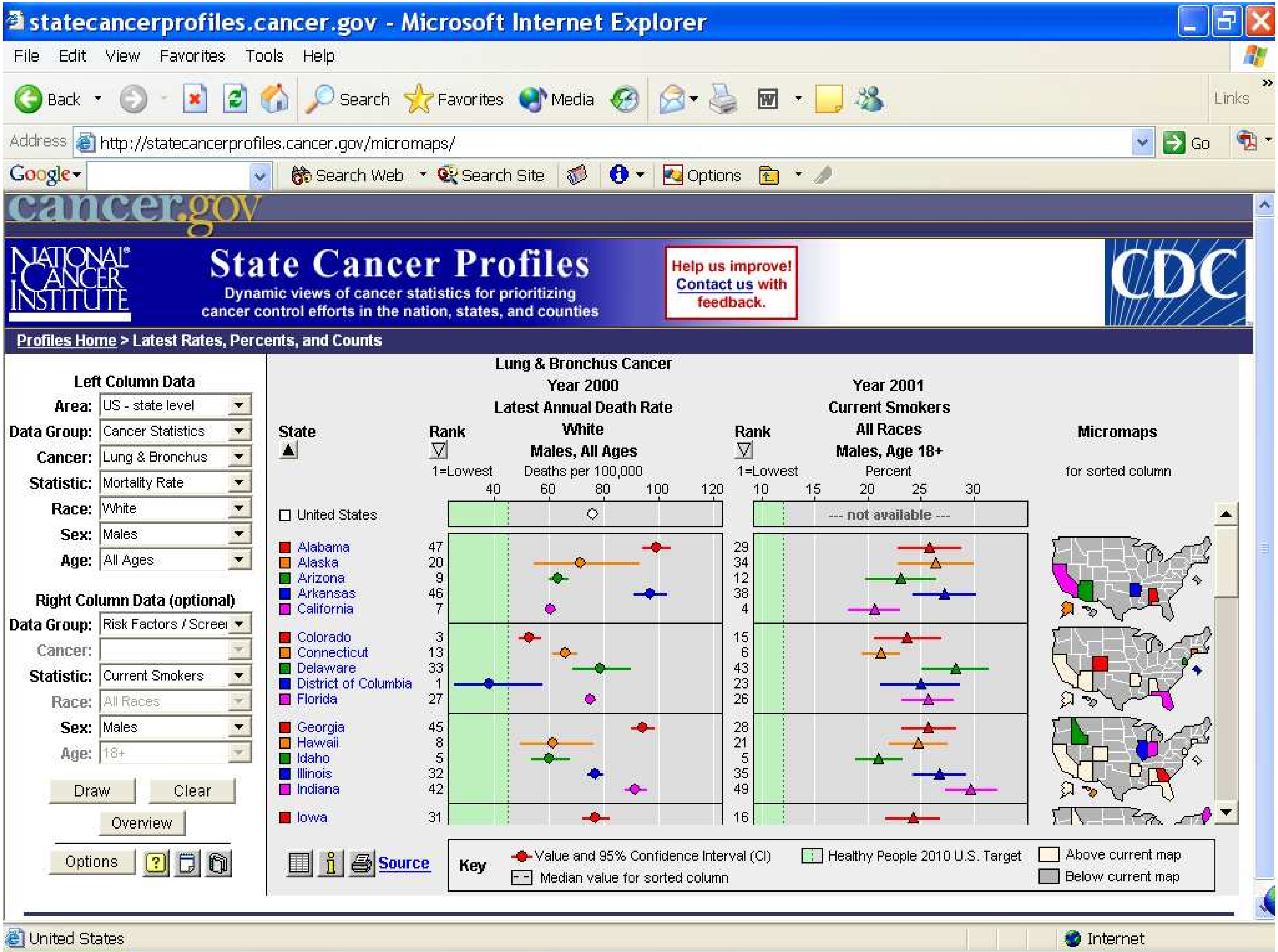


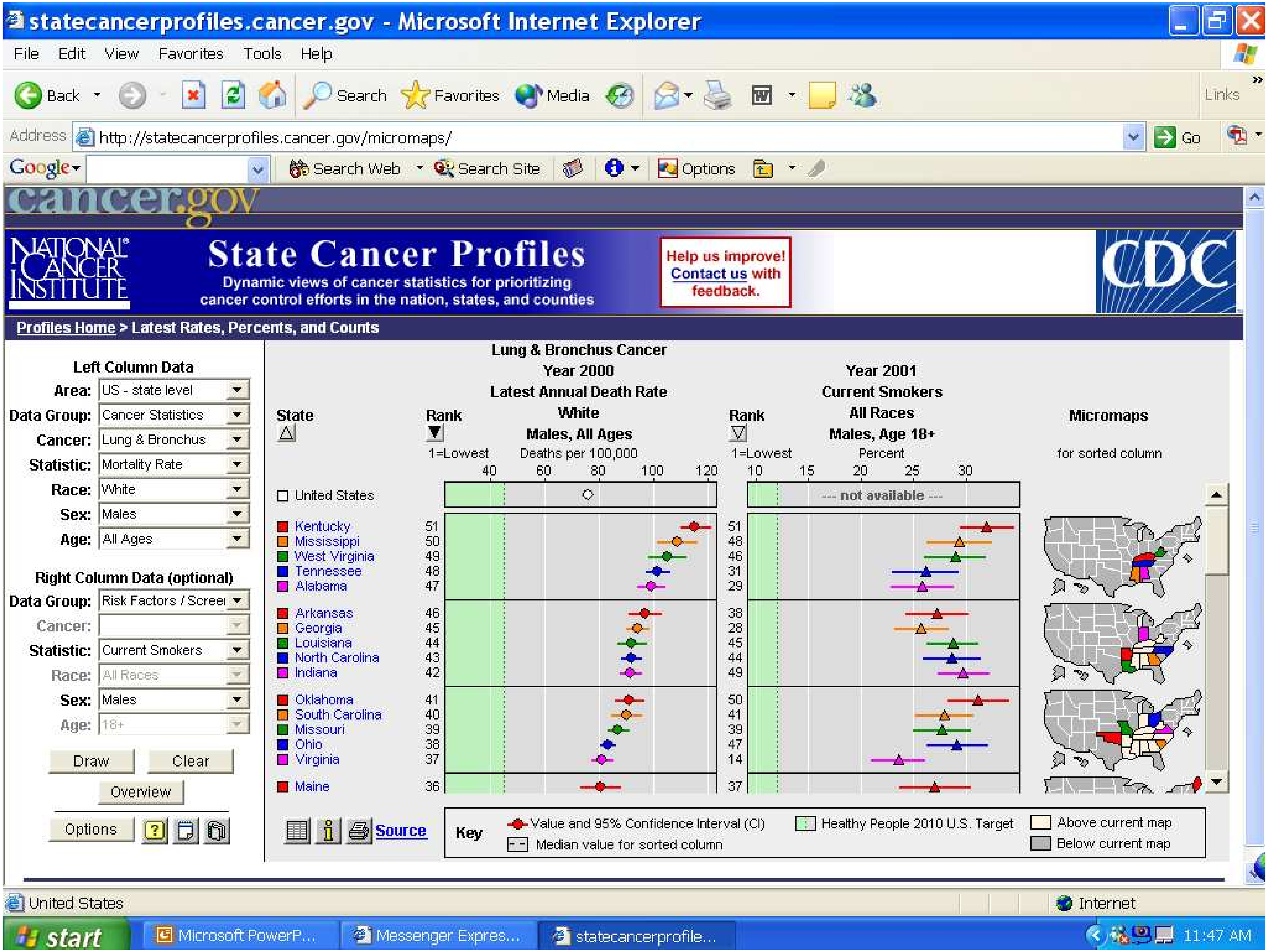
Connecticut Hospitals: Diagnosed Cancers
Data Quality Indicators: 1995 - 1999 To 2000

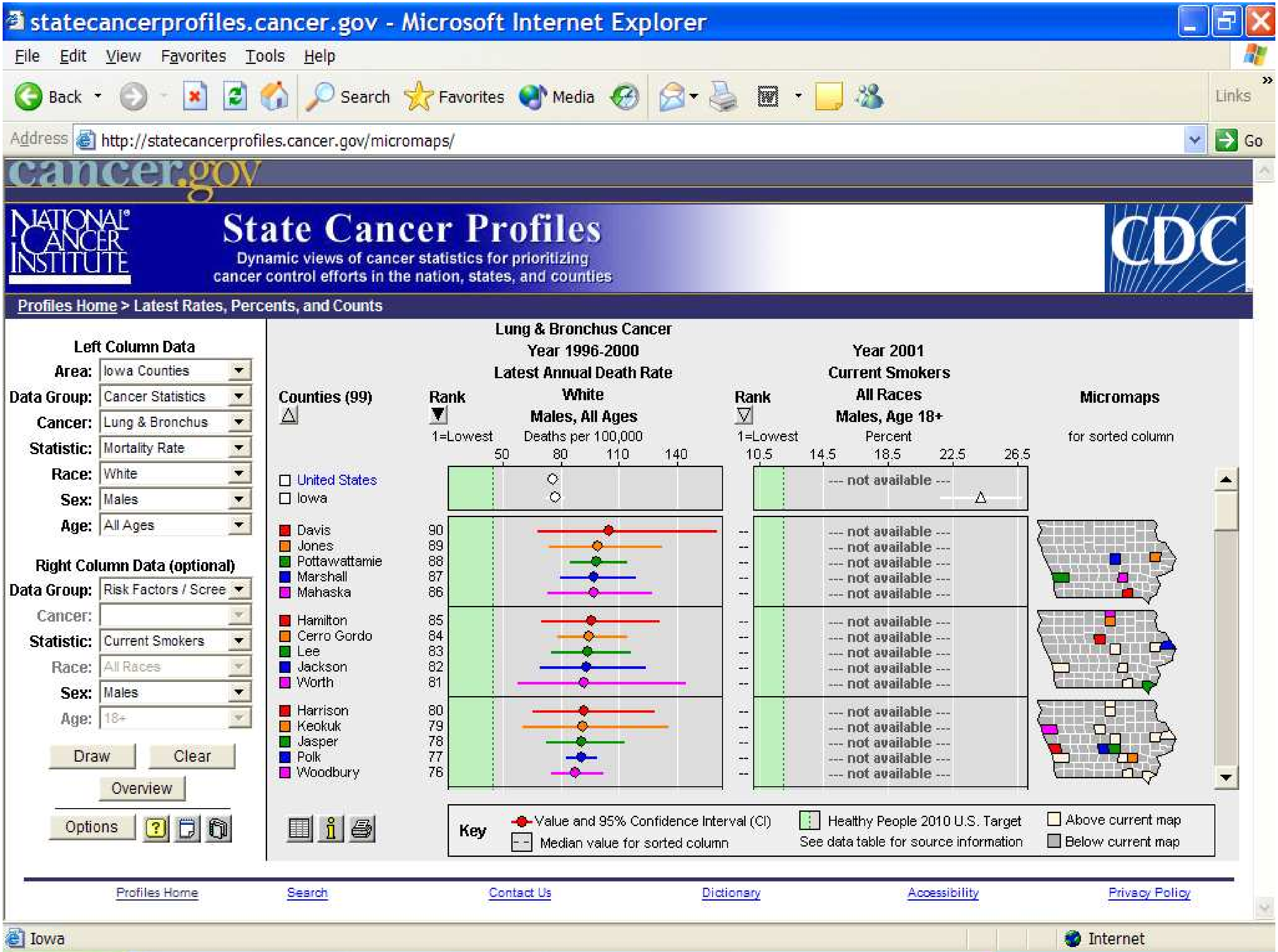


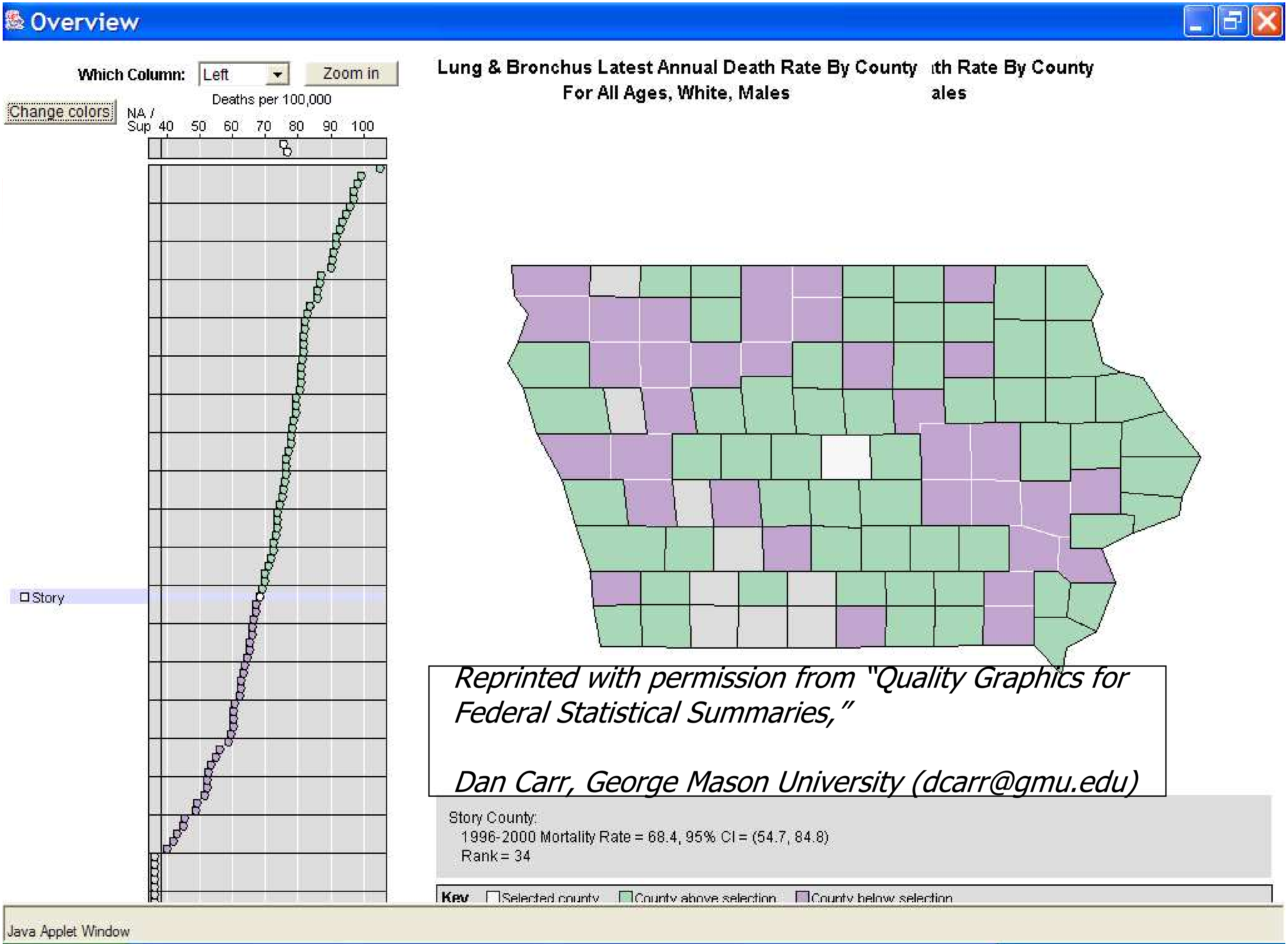
Reprinted with permission from "Quality Graphics for Federal Statistical Summaries,"

Dan Carr, George Mason University (dcarr@gmu.edu)


statecancerprofiles.cancer.gov - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Links
Address http://statecancerprofiles.cancer.gov/micromaps/
Go
Search Web Search Site Options
cancer.gov
NATIONAL CANCER INSTITUTE
State Cancer Profiles
Dynamic views of cancer statistics for prioritizing cancer control efforts in the nation, states, and counties
Help us improve! Contact us with feedback.
CDC
Profiles Home > Latest Rates, Percents, and Counts
Left Column Data
Area: US - state level
Data Group: Cancer Statistics
Cancer: Lung & Bronchus
Statistic: Mortality Rate
Race: White
Sex: Males
Age: All Ages
Right Column Data (optional)
Data Group: Risk Factors / Screening
Cancer:
Statistic: Current Smokers
Race: All Races
Sex: Males
Age: 18+
Draw
Clear
Overview
Options
Lung & Bronchus Cancer
Year 2000
Latest Annual Death Rate
White
Males, All Ages
Deaths per 100,000
40 60 80 100 120
Year 2001
Current Smokers
All Races
Males, Age 18+
Percent
10 15 20 25 30
State
Rank
1=Lowest
Micromaps
for sorted column
United States
--- not available ---
Alabama 47 29
Alaska 20 34
Arizona 9 12
Arkansas 46 38
California 7 4
Colorado 3 15
Connecticut 13 6
Delaware 33 43
District of Columbia 1 23
Florida 27 26
Georgia 45 28
Hawaii 8 21
Idaho 5 5
Illinois 32 35
Indiana 42 49
Iowa 31 16
Source
Key
Value and 95% Confidence Interval (CI)
Median value for sorted column
Healthy People 2010 U.S. Target
Above current map
Below current map
United States
Internet

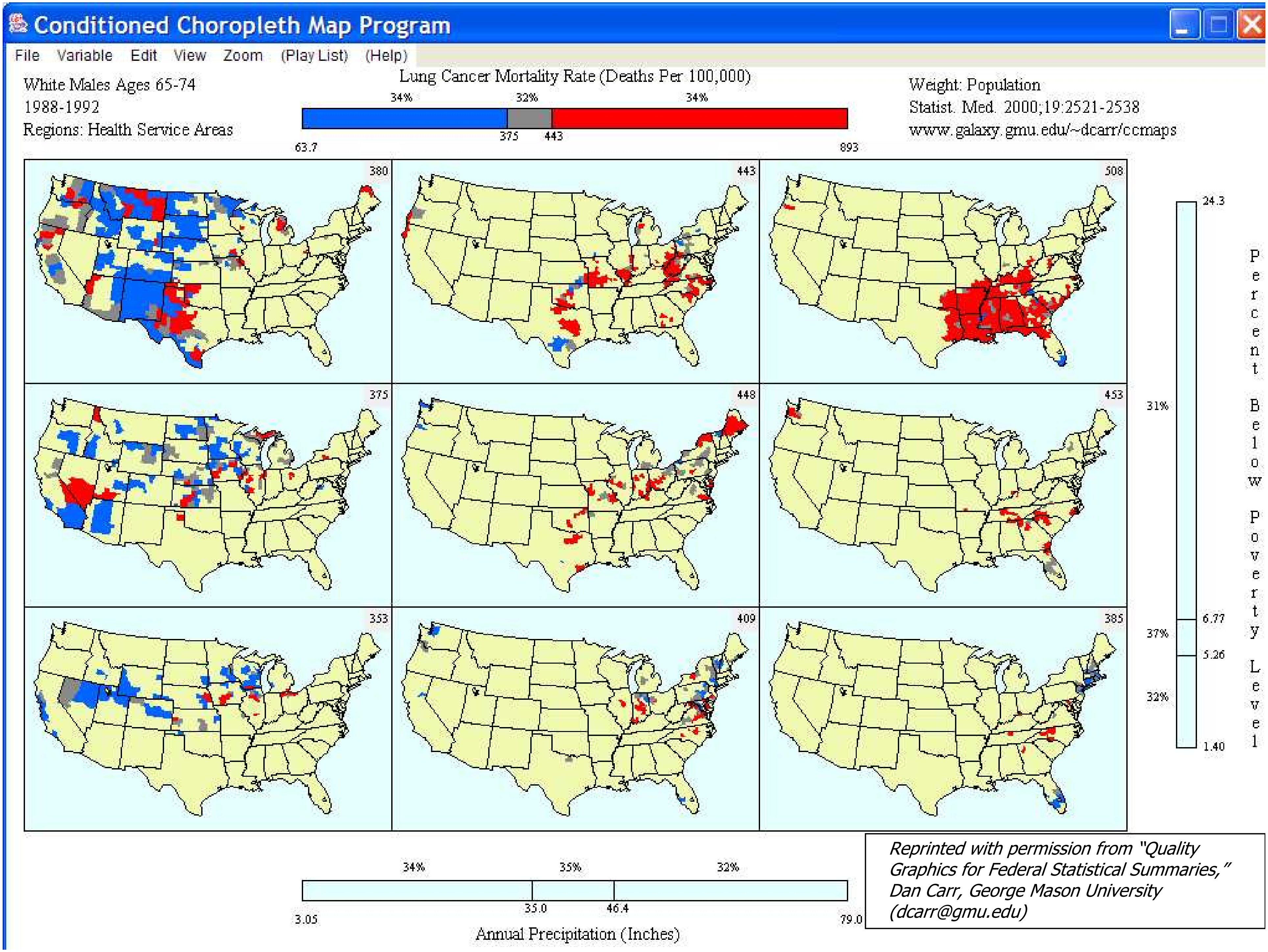

statecancerprofiles.cancer.gov - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Links
Address http://statecancerprofiles.cancer.gov/micromaps/ Go
Google Search Web Search Site Options
cancer.gov
NATIONAL CANCER INSTITUTE
State Cancer Profiles
Dynamic views of cancer statistics for prioritizing cancer control efforts in the nation, states, and counties
Help us improve! Contact us with feedback.
CDC
Profiles Home > Latest Rates, Percents, and Counts
Left Column Data
Area: US - state level
Data Group: Cancer Statistics
Cancer: Lung & Bronchus
Statistic: Mortality Rate
Race: White
Sex: Males
Age: All Ages
Right Column Data (optional)
Data Group: Risk Factors / Screen
Cancer:
Statistic: Current Smokers
Race: All Races
Sex: Males
Age: 18+
Draw
Clear
Overview
Options
Lung & Bronchus Cancer
Year 2000
Latest Annual Death Rate
White
Males, All Ages
Deaths per 100,000
Year 2001
Current Smokers
All Races
Males, Age 18+
Percent
State
Rank
1=Lowest
Micromaps
for sorted column
40 60 80 100 120
10 15 20 25 30
United States
--- not available ---
Kentucky 51 51
Mississippi 50 48
West Virginia 49 46
Tennessee 48 31
Alabama 47 29
Arkansas 46 38
Georgia 45 28
Louisiana 44 45
North Carolina 43 44
Indiana 42 49
Oklahoma 41 50
South Carolina 40 41
Missouri 39 39
Ohio 38 47
Virginia 37 14
Maine 36 37
Source
Key
Value and 95% Confidence Interval (CI)
Median value for sorted column
Healthy People 2010 U.S. Target
Above current map
Below current map
United States
Internet
start
Microsoft PowerP...
Messenger Expres...
statecancerprofile...
11:47 AM


statecancerprofiles.cancer.gov - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Links
Address http://statecancerprofiles.cancer.gov/micromaps/ Go
cancer.gov
NATIONAL CANCER INSTITUTE
State Cancer Profiles
Dynamic views of cancer statistics for prioritizing cancer control efforts in the nation, states, and counties
CDC
Profiles Home > Latest Rates, Percents, and Counts
Left Column Data
Area: Iowa Counties
Data Group: Cancer Statistics
Cancer: Lung & Bronchus
Statistic: Mortality Rate
Race: White
Sex: Males
Age: All Ages
Right Column Data (optional)
Data Group: Risk Factors / Scree
Cancer:
Statistic: Current Smokers
Race: All Races
Sex: Males
Age: 18+
Draw
Clear
Overview
Options
Lung & Bronchus Cancer
Year 1996-2000
Latest Annual Death Rate
White
Males, All Ages
Deaths per 100,000
Counties (99)
Rank
1=Lowest
50 80 110 140
Year 2001
Current Smokers
All Races
Males, Age 18+
Percent
10.5 14.5 18.5 22.5 26.5
Micromaps
for sorted column
United States
Iowa
--- not available ---
Davis 90
Jones 89
Pottawattamie 88
Marshall 87
Mahaska 86
Hamilton 85
Cerro Gordo 84
Lee 83
Jackson 82
Worth 81
Harrison 80
Keokuk 79
Jasper 78
Polk 77
Woodbury 76
Key
Value and 95% Confidence Interval (CI)
Median value for sorted column
Healthy People 2010 U.S. Target
See data table for source information
Above current map
Below current map
Profiles Home
Search
Contact Us
Dictionary
Accessibility
Privacy Policy
Iowa
Internet



Reprinted with permission from "Quality Graphics for Federal Statistical Summaries,"

Dan Carr, George Mason University (dcarr@gmu.edu)



Reprinted with permission from "Quality Graphics for Federal Statistical Summaries," Dan Carr, George Mason University (dcarr@gmu.edu)

Recommendations

- **Build visualization into all of the software systems from the ground up**
- **Seek ways to exploit the human visual system as an aid in pattern identification/hypothesis generation**
- **Look to exploit multi-technique synthesis where appropriate**
- **Don't relegate oneself to simple Euclidean geometry let the inherent geometric structure in the data come to the forefront**
  - **Incorporate relevant barriers into distance calculations**
- **Provide the ability for the analyst to annotate and log their analysis**
- **Additional funding is needed to develop more user friendly gui development programs**
- **Use statistical summaries to facilitate hotspot identification**
- **Algorithms to prioritize the presentation of data views to the user in order to facilitate subtle pattern discovery**

References - I

- Jacques Bertin, Semiologie Graphique, ISBN 2 –1732-1277-4, 1967.
- Dan Carr, http://statecancerprofiles.cancer.gov/micromaps, 2003.
- David J. Marchette and Jeffrey L. Solka, "Using data images for outlier detection," *Computational Statistics and Data Analysis*, pp. 541-552, 2003.
- Michael C. Minnotte and R. Webster West, "The data image a tool for exploring high-dimensional datasets," *1998 Proceedings of the ASA Section on Statistical Graphics*, 1998.
- Carey E. Priebe, Jeffrey L. Solka, David J. Marchette, and B. Ted Clark, "Class cover catch digraphs for latent class discovery in gene expression monitoring by DNA microarrays," *Computational Statistics and Data Analysis*, 43, pp. 621-632, 2003.

References - II

- Jeffrey L. Solka, Edward J. Wegman, and David J. Marchette, "Data mining strategies for the detection of chemical warfare agents," in *Statistical Data Mining and Knowledge Discovery*, (Hamparsum Bozdogan, ed.), 57-92, Boca Raton: Chapman and Hall/CRC, 2003.

- Jeffrey L. Solka, Barton T. Clark, and Jeffrey L. Solka, "A visualization framework for the analysis of hyperspectral data," *International Journal of Image and Graphics*, Vol. 2, No. 1, pp. 145-161, 2002.

- Joshua B. Tenenbaum, Vin de Silva, John C. Langford A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science 2001?

- Edward J. Wegman, "Hyperdimensional data analysis using parallel coordinates," *Journal of the American Statistical Association*, 85:664-675, 1990.

- Edward J. Wegman and Alan Dorfman, "Visualizing cereal world," *Computational Statistics and Data Analysis*, 43, pp. 633-649, 2003.