# UFMG Workshop: Statistical Graphics for Spatial and Environmental Research

**Jürgen Symanzik**

**Utah State University, Logan, UT, USA**

***e-mail: symanzik@math.usu.edu**

**WWW: http://www.math.usu.edu/~symanzik**

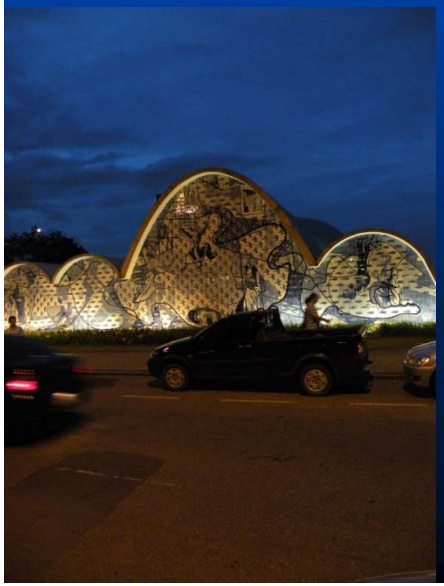
# Contents

- Talk 1: Mo, Nov 30, 2009:
  - Interactive and Dynamic Statistical Graphics – An Overview
- Talk 2: Tu, Dec 1, 2009:
  - Linked Micromap Plots – In Print and on the Web
- Talk 3: We, Dec 2, 2009:
  - Interactive and Dynamic Statistical Graphics – Special Applications

# Interactive and Dynamic Statistical Graphics - An Overview

**Jürgen Symanzik**

**Utah State University, Logan, UT, USA**



***e-mail: symanzik@math.usu.edu**

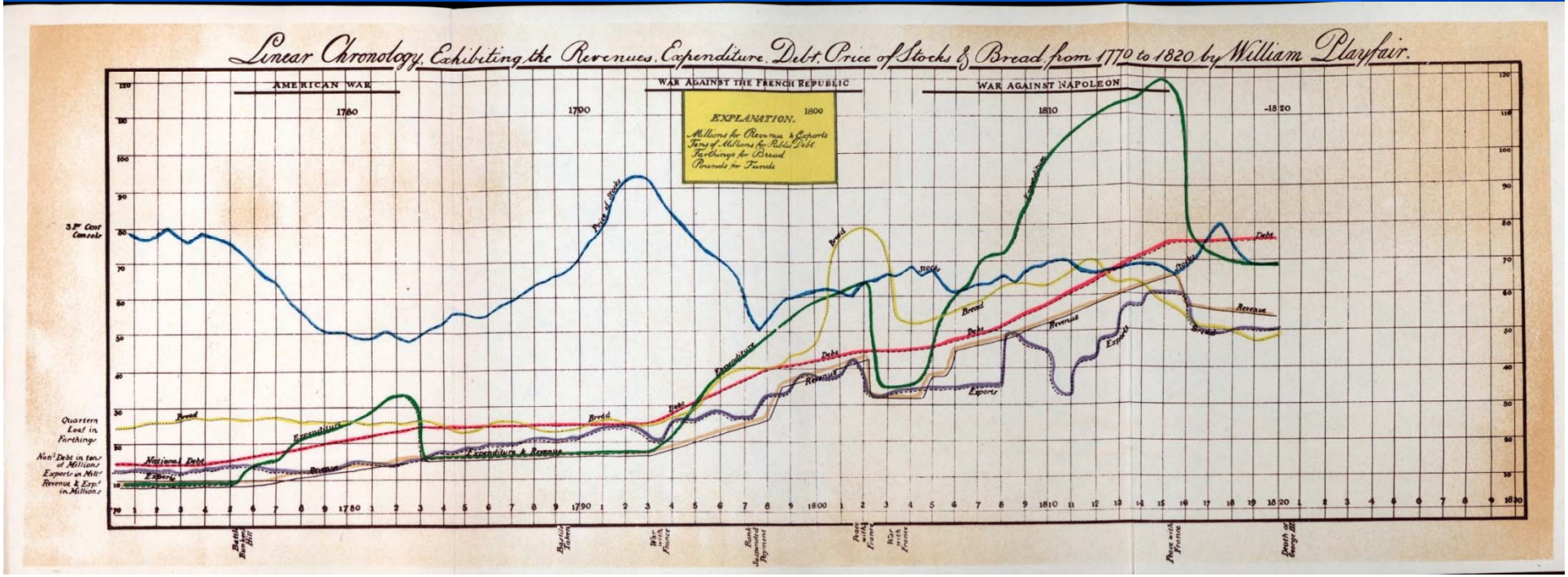
**WWW: http://www.math.usu.edu/~symanzik**

# Contents

- Data, Terms, Citations, and Definitions
- Main Concepts
- Graphical Software
- Live Demo
- Conclusion

# Historical 250th Birthday

- William Playfair: Born on September 22, 1759
- "Father" and "Inventor" of Statistical Graphics
- Modern plots still look very similar:



# Why Graphics ? (1)

- Try this in R:

```
require(stats)
data(anscombe)

##-- now some "magic" to do the 4 regressions in a loop:
ff <- y ~ x
for(i in 1:4)
{
  ff[2:3] <- lapply(paste(c("y","x"), i, sep=""), as.name)
  assign(paste("lm.",i,sep=""), lmi <- lm(ff, data= anscombe))
}

## See how close they are (numerically!)
sapply(objects(pat="lm.[1-4]$"), function(n) coef(get(n)))
lapply(objects(pat="lm.[1-4]$"), function(n) summary(get(n))$coef)
```

# Why Graphics ? (2)

- Try this in R (We should have done this first !!!):

```
op <- par(mfrow=c(2,2), mar=.1+c(4,4,1,1), oma= c(0,0,2,0))
for(i in 1:4)
{
  ff[2:3] <- lapply(paste(c("y","x"), i, sep=""), as.name)
  plot(ff, data =anscombe, col="red", pch=21, bg = "orange", cex = 1.2, xlim=c(0,20), ylim=c(3,13))
  abline(get(paste("lm.",i,sep="")), col="blue")
}
mtext("Anscombe's 4 Regression data sets", outer = TRUE, cex=1.5)
par(op)


# R code based on: http://pbil.univ-lyon1.fr/library/base/html/anscombe.html
```

# Terms

- Interactive & Dynamic Statistical Graphics (DSG)
- Exploratory Data Analysis (EDA)
- Exploratory Spatial Data Analysis (ESDA)
- Visual Data Mining (VDM)
- Visual Analysis/Visual Analytics (VA)
- Data Mining (DM)

# Citations

- John W. Tukey (1977):

  EDA *"is detective work - numerical detective work - or counting detective work - or graphical detective work."*

- Edward J. Wegman (2000):

  *"Data Mining is exploratory data analysis with little or no human interaction using computationally feasible techniques, i.e., the attempt to find interesting structure unknown a priori."*

# DSG/VDM (1)

- Working Definition for DSG/VDM:
  - Find structure (cluster, unusual observations) in large and not necessarily homogeneous data sets based on human perception using graphical methods and user interaction
  - Goal or expected outcome of exploration usually unknown in advance

# DSG/VDM (2)

- First uses of the term VDM:
  - Cox, Eick, Wills, Brachman (1997): Visual Data Mining: Recognizing Telephone Calling Fraud, *Data Mining and Knowledge Discovery*, 1:225-231.
  - Inselberg (1998): Visual Data Mining with Parallel Coordinates, *Computational Statistics*, 13(1):47-63.

# DSG Concepts (1)

- Scatterplots and Scatterplot Matrices
- Brushing and Linked Brushing/Linked Views
- Focusing, Zooming, Panning, Slicing, Rescaling, and Reformatting
- Rotations and Projections
- Grand Tour
- Parallel Coordinate Plots

# DSG Concepts (2)

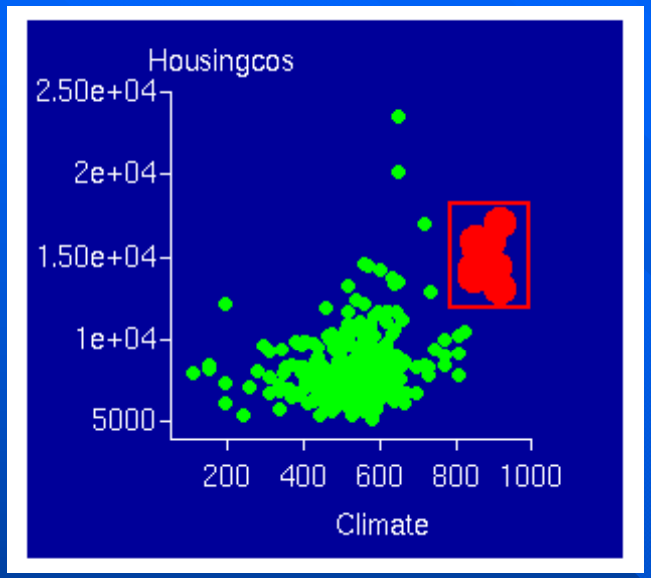
- Projection Pursuit and Projection Pursuit Guided Tours
- Pixel or Image Grand Tours
- Andrews Plots
- Density Plots, Binning, and Brushing with Hue and Saturation
- Special DSG techniques for Categorical Data

# Places Data

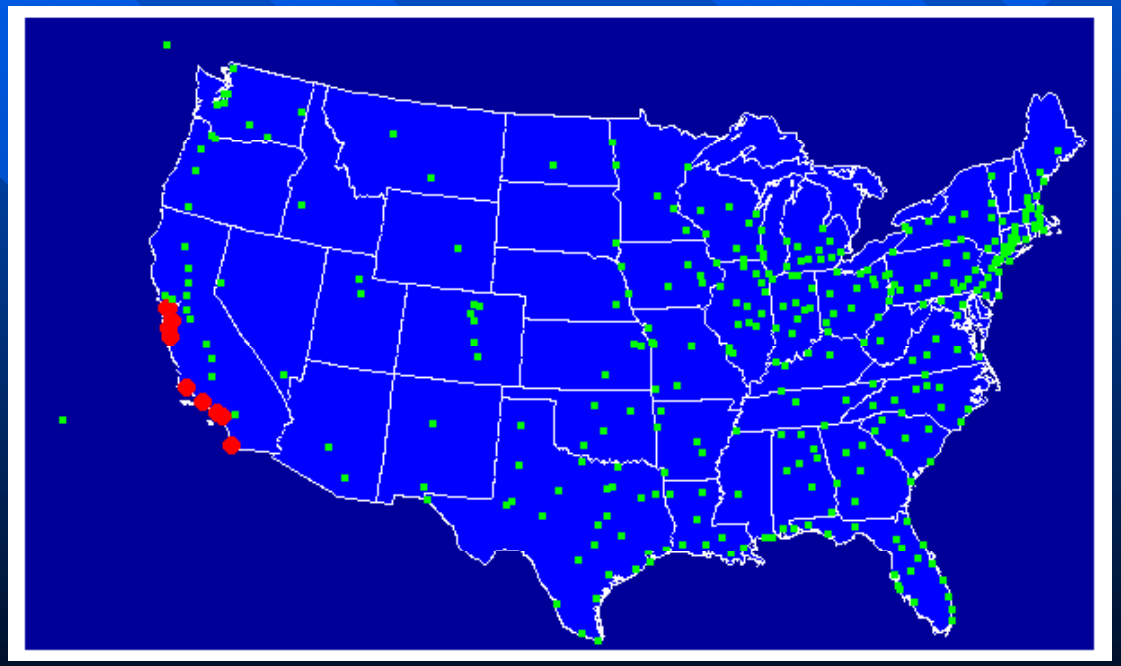
- "Places" data set:
  - 329 cities in the U.S.
  - 9 measures of livability (early 1980's):

    Climate & Terrain, Housing Cost, Health Care & Environment, Crime, Transportation, Education, The Arts, Recreation, and Economics.
  - Published in Places Rated Almanac (Boyer and Savageau, 1981), copyrighted by Rand McNally
  - Latitude and longitude added by Paul Tukey

# Scatterplots and Linked Brushing

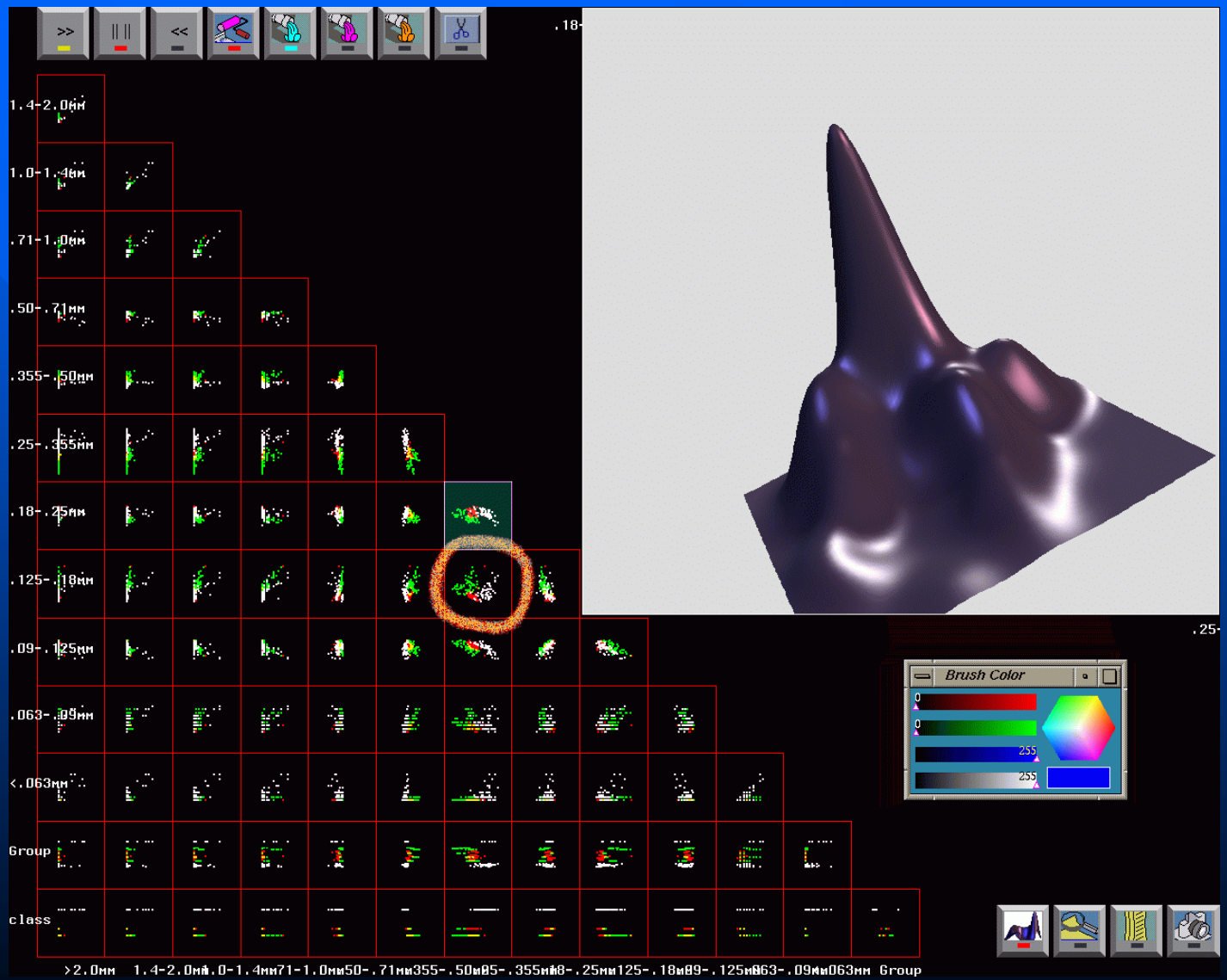



**XGobi**



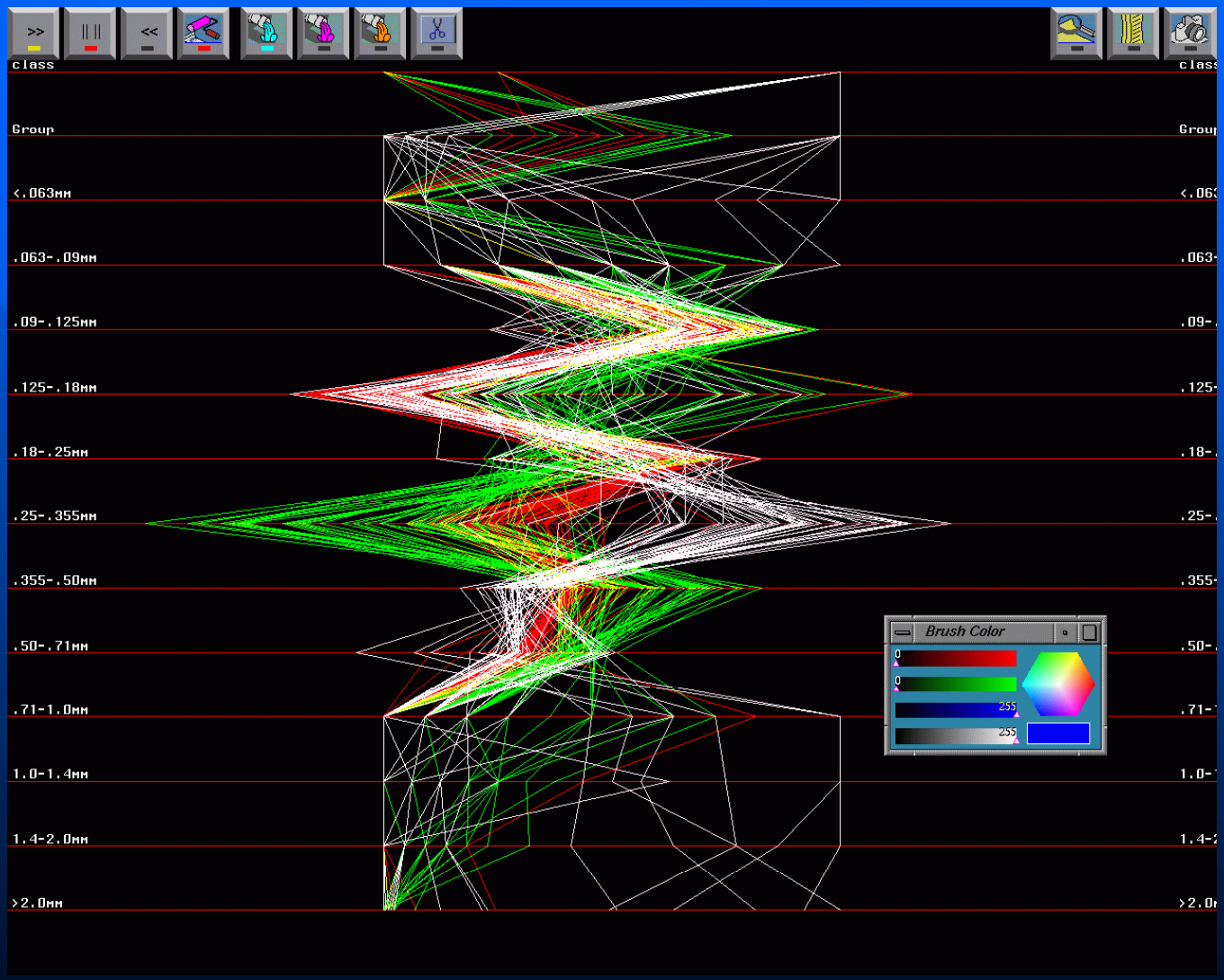
# Scatterplot Matrix and Density Plot

**ExplorN**



# Parallel Coordinate Plots

ExplorN



# Grand Tour

- Continuous random sequence of projections from n dimensions into 2 (or more) dimensions.

# Graphical Software

- Origin: PRIM-9
- REGARD, MANET, Mondrian, and iplots Family
- EXPLOR4, HyperVision, ExplorN, and CrystalVision Family
- DataViewer, XGobi, GGobi, and Rggobi Family

# Origin of DSG Software: PRIM-9

- "Picturing, Rotation, Isolation and Masking in up to 9 Dimensions"
- Initiated in the early 1970's by M. A. Fisherkeller, J. H. Friedman, and J. W. Tukey
- Main features:
  - Projections
  - Isolations and Masking

# DSG Software: REGARD, MANET, Mondrian, and iplots

- Initiated in the late 1980's by John Haslett and Antony Unwin at Trinity College, Dublin, Ireland
- Continued by Antony Unwin and collaborators at University of Augsburg, Germany
- Main collaborators: Heike Hofmann, Martin Theus, Simon Urbanek, Adalbert Wilhelm, and Graham Wills

# REGARD

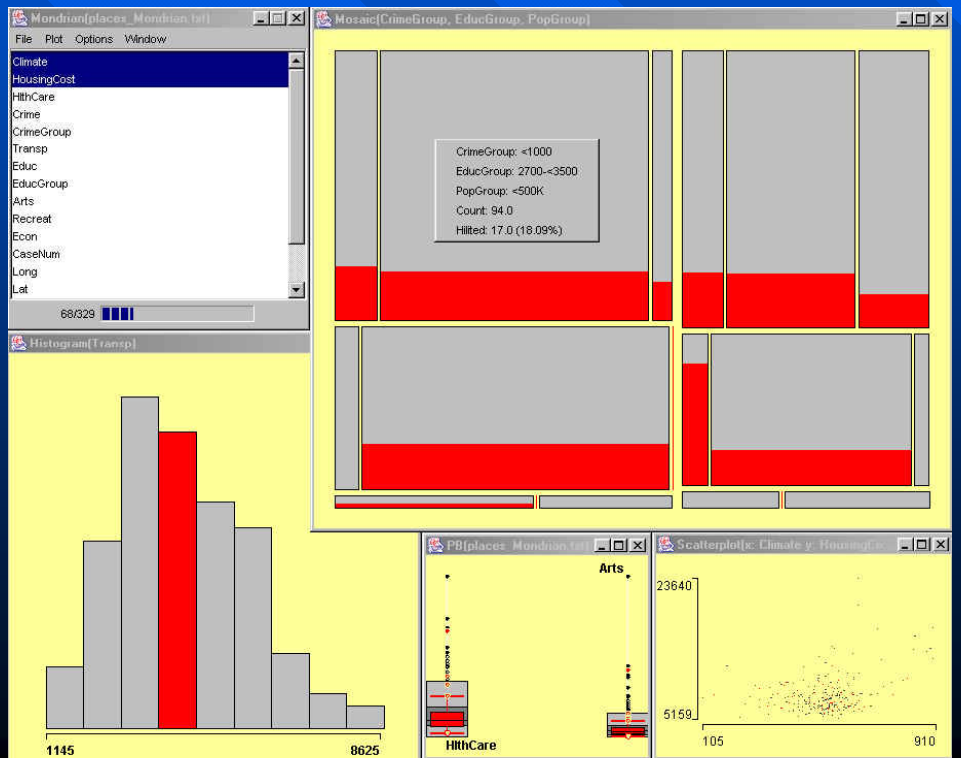
- “Radical Effective Graphical Analysis of Regional Data”
- Early 1990’s, Macintosh
- High interaction graphics tools for spatial data
- Map window that is linked to statistical displays

# MANET

- "Missings Are Now Equally Treated"
- Mid/Late 1990's, Macintosh
- *http://rosuda.org/manet/*
- Graphics for continuous and discrete data
- Keeps track of missing values in graphics

# Mondrian

- Early 2000's, JAVA
- *http://www.rosuda.org/Mondrian/*
- Visualization of categorical and geographic data



# iplots

- 2003, interactive graphics package for R
- *http://cran.r-project.org/web/packages/iplots/index.html*
- *http://rosuda.org/iplots/*
- Closely related to Mondrian

# DSG Software: EXPLOR4, HyperVision, ExplorN, and CrystalVision

- Initiated in the late 1980's by Dan Carr and Ed Wegman at George Mason University
- Other main collaborators: Qiang Luo and Wesley L. Nicholson

# EXPLOR4

- Late 1980's, VAX 11/780, Fortran
- Main features:
  - Rotations
  - Scatterplots & Scatterplot Matrix
  - Stereoscopic Views

# HyperVision

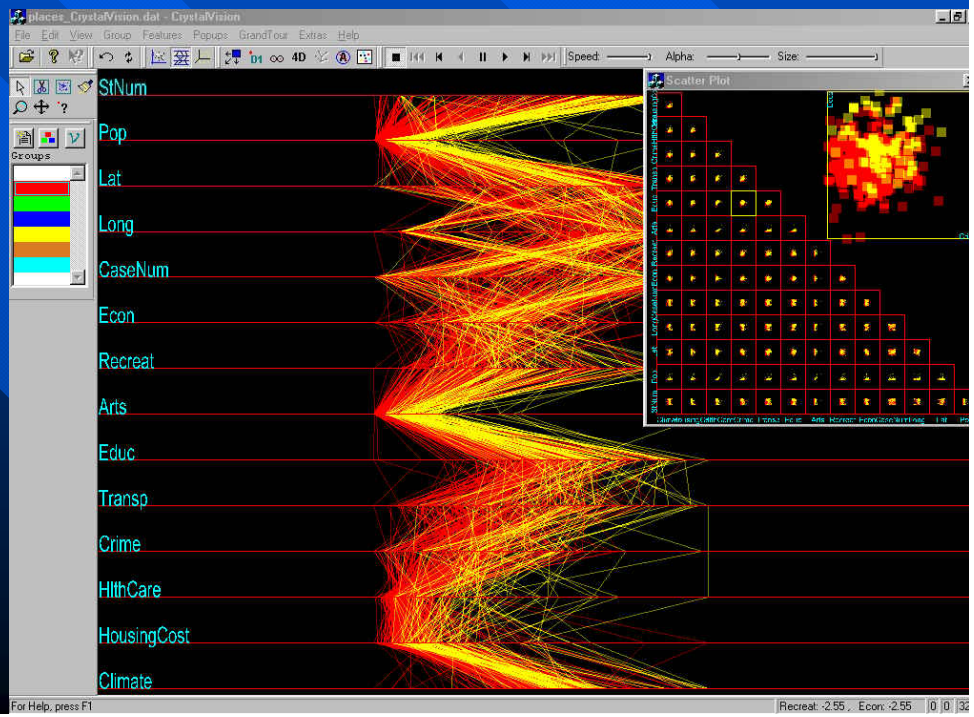
- Late 1980's, IBM RT & MS-DOS, Pascal
- Main features:
  - Real Time Rotations
  - 2D & 3D Scatterplots & Scatterplot Matrix
  - Parallel Coordinate Plots
  - Color Histograms

# ExplorN

- Mid 1990's, SGI
- *ftp://www.galaxy.gmu.edu/pub/software/*
- Interactive environment for exploring multivariate data:
  - Advanced Parallel Coordinates Displays
  - 3D Surfaces
  - Stereoscopic Displays

# CrystalVision

- Early 2000's, PCs
- *ftp://www.galaxy.gmu.edu/pub/software/*
- Main features:
  - Parallel coordinate plots
  - Scatterplots
  - Grand tour animations



# DSG Software: DataViewer, XGobi, GGobi, and Rggobi

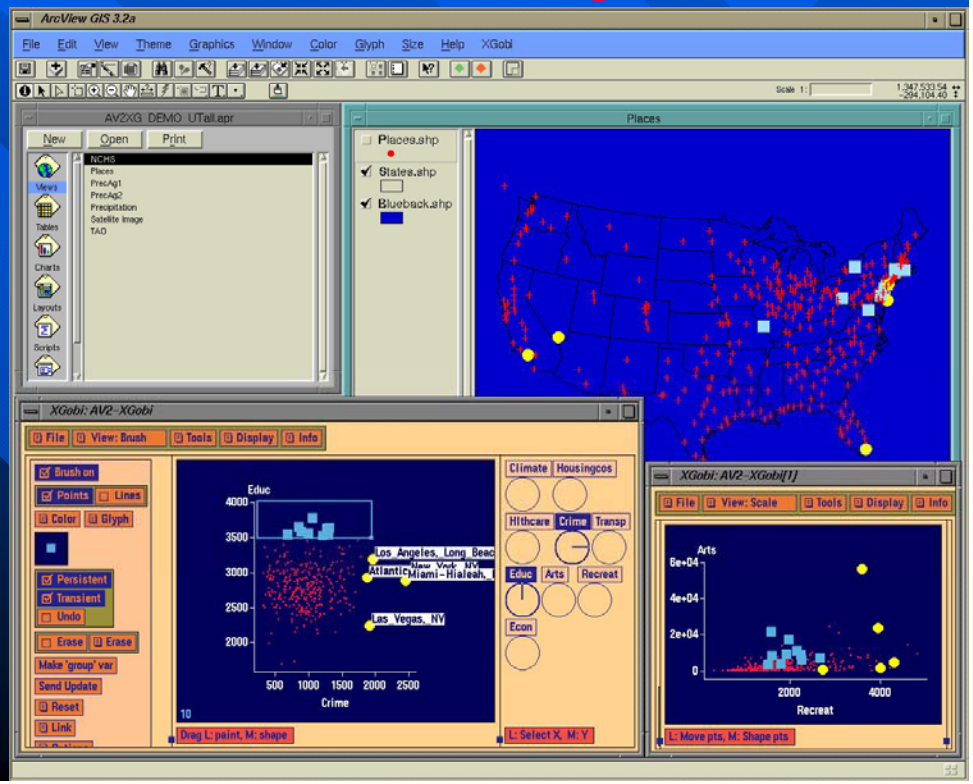
- Initiated in the mid 1980's by Andreas Buja, Deborah F. Swayne, and Dianne Cook at the University of Washington, Bellcore, AT&T Bell Labs, and Iowa State University
- Other main collaborators: Catherine Hurley, John A. McDonald, Duncan Temple Lang, Hadley Wickham

# DataViewer

- Mid 1980's, Symbolics Lisp Machine
- Main features:
  - Linked windows
  - Focusing
  - Projections such as 3D rotations and grand tour

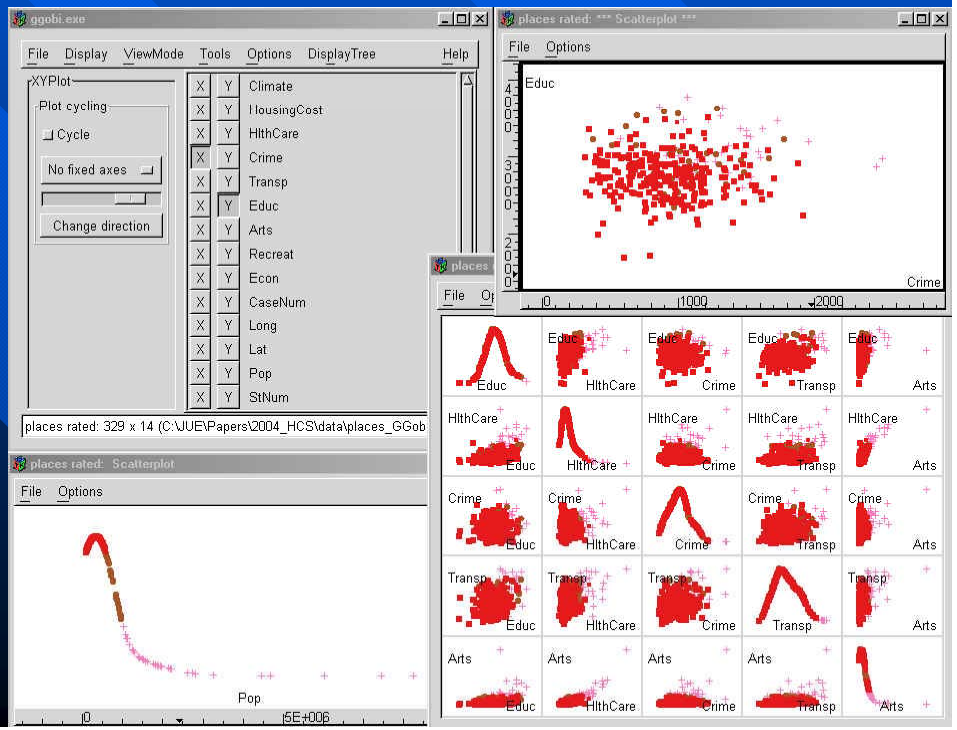
# XGobi

- Early 1990's through early 2000's
- UNIX and Linux platforms
- *http://www.research.att.com/areas/stat/xgobi/*
- Main features:
  - Linked views
  - Linked brushing
  - Univariate, bivariate, and multivariate views
  - Grand tour
  - Links to other software



# GGobi

- Early 2000's
- PCs, UNIX and Linux platforms
- *http://www.ggobi.org/*
- Main features:
  - Very similar to XGobi
  - Multiple plot windows
  - Uses GTK+ graphical toolkit



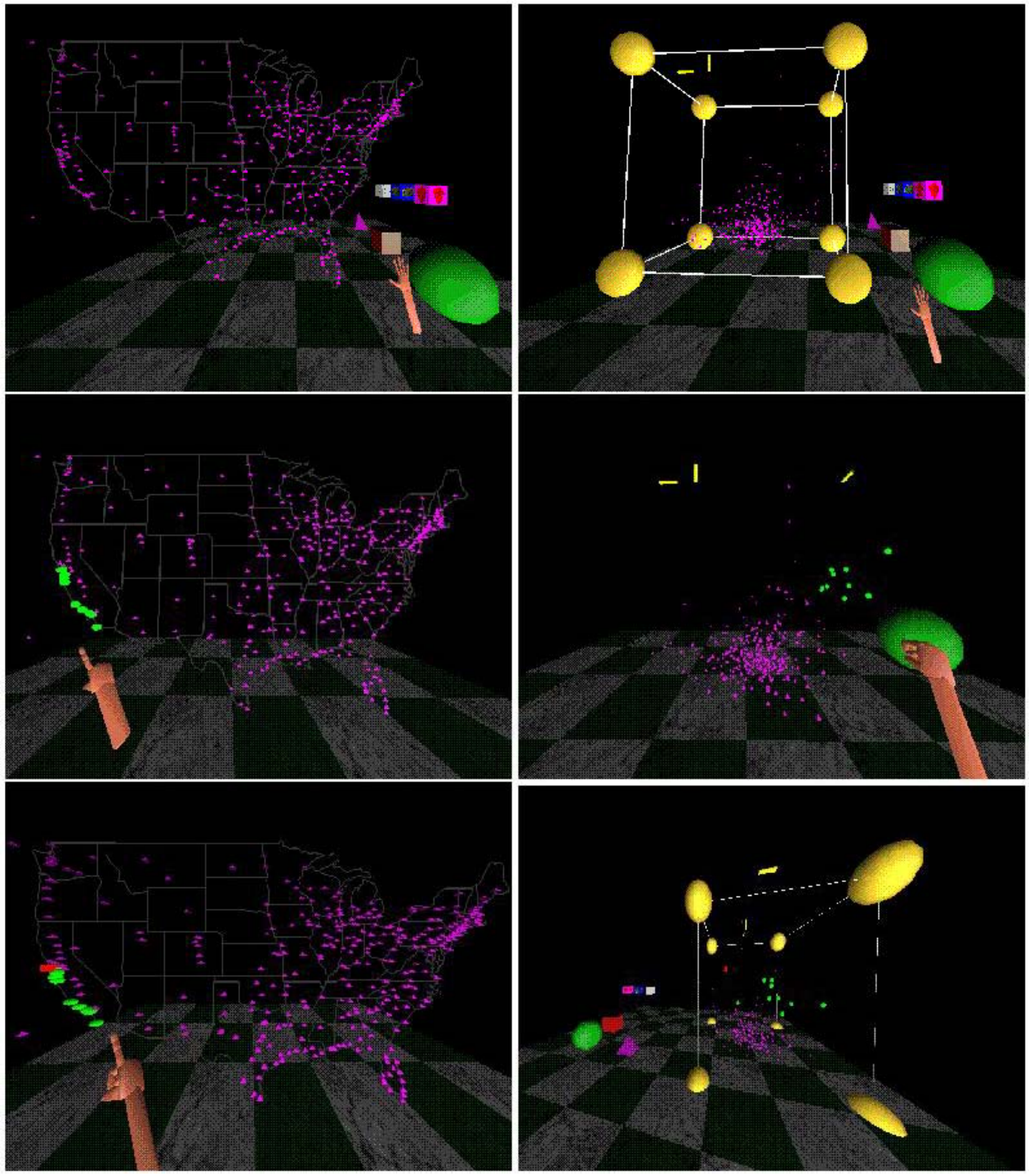
# Rggobi

- Early 2000's
- PCs, UNIX and Linux platforms
- *http://cran.r-project.org/web/packages/rggobi/index.html*
- Command-line Interface between R and GGobi

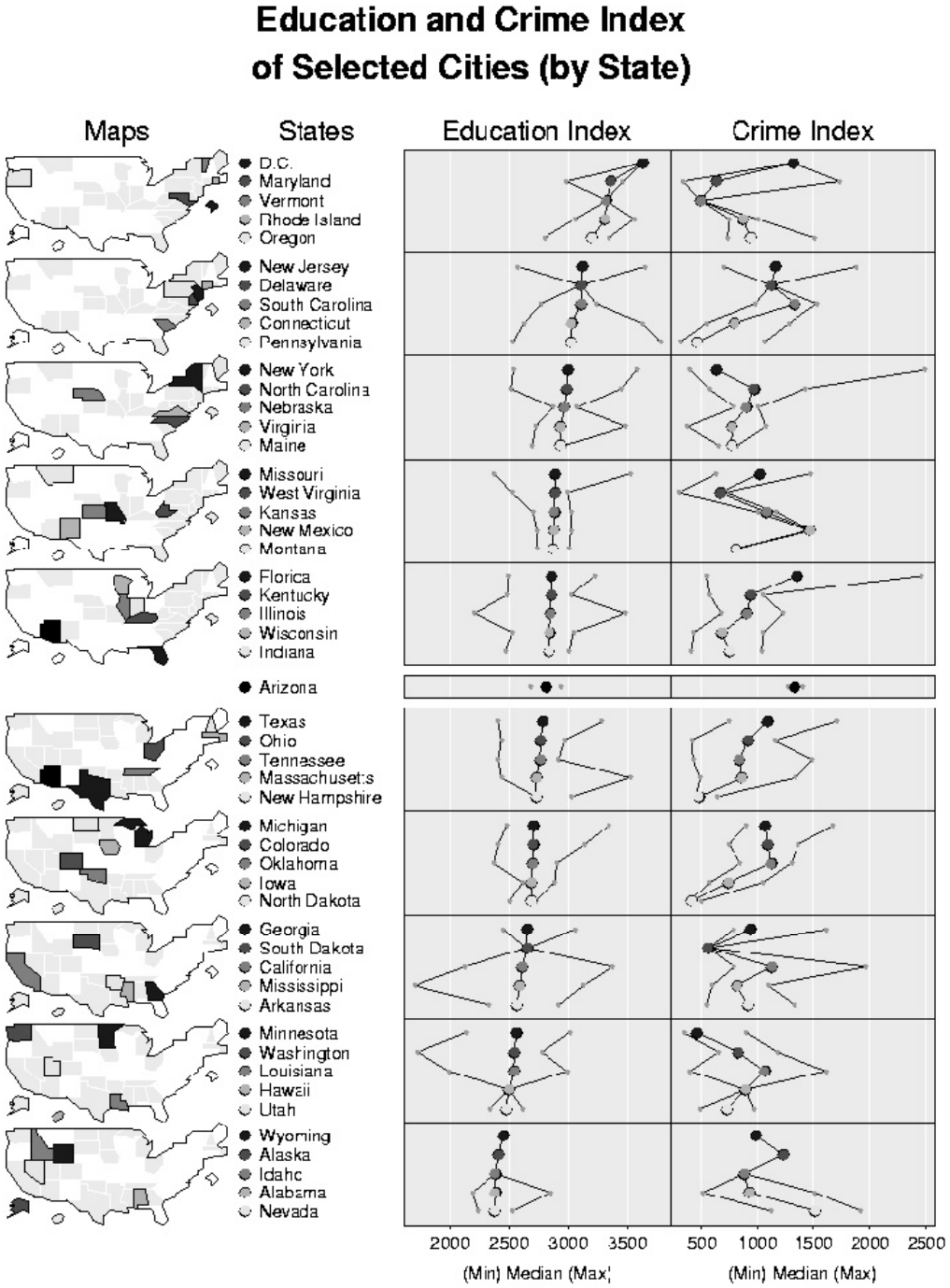
# Live Demo

-  GGobi
-  CrystalVision
-  Mondrian

# Places Data in VRGobi



# Places Data as Micromaps

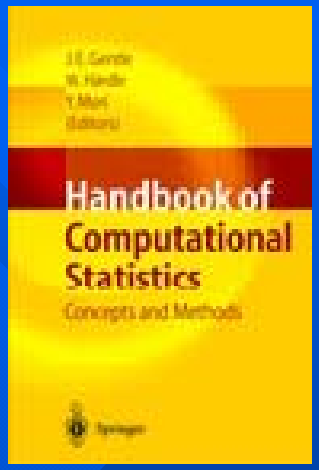


# Conclusions

- Visual approach effective to see unexpected structure in data
- Combination of different techniques most effective
- Can be used for almost all types of data

**Main References:**

Symanzik, J. (2004): Interactive and Dynamic Graphics, In: Gentle, J. E., Härdle, W., Mori, Y. (Eds.), Handbook of Computational Statistics - Concepts and Methods, Springer, Berlin/Heidelberg, 293-336.




Symanzik, J. (2010): Interactive and Dynamic Statistical Graphics, In: Lovric, M. (Ed.), International Encyclopedia of Statistical Science, Springer, Forthcoming.

*Questions ???*