Graphical Methods: From Visual Data Mining to Modern Presentation Graphics

Jürgen Symanzik

Utah State University, Logan, UT, USA

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

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

Contents

- Terms, Citations, and Definitions
- Main Concepts and Live Demos
- Presentation Graphics via Micromaps
- Current Work at WUSTL
- Conclusion

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)

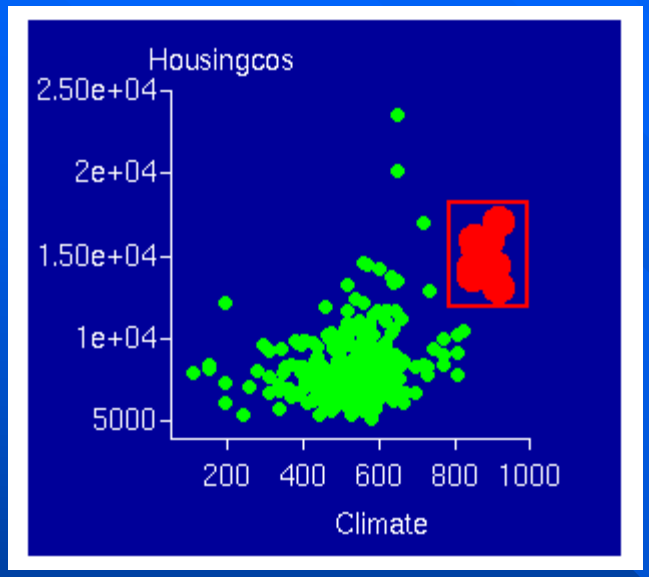
- Scatterplots* and Scatterplot Matrices
- Brushing* and Linked Brushing/Linked Views*
- Focusing, Zooming, Panning, Slicing, Rescaling, and Reformatting*
- Rotations* and Projections*
- Grand Tour*
- Parallel Coordinate Plots*

* Will be discussed in Presentation

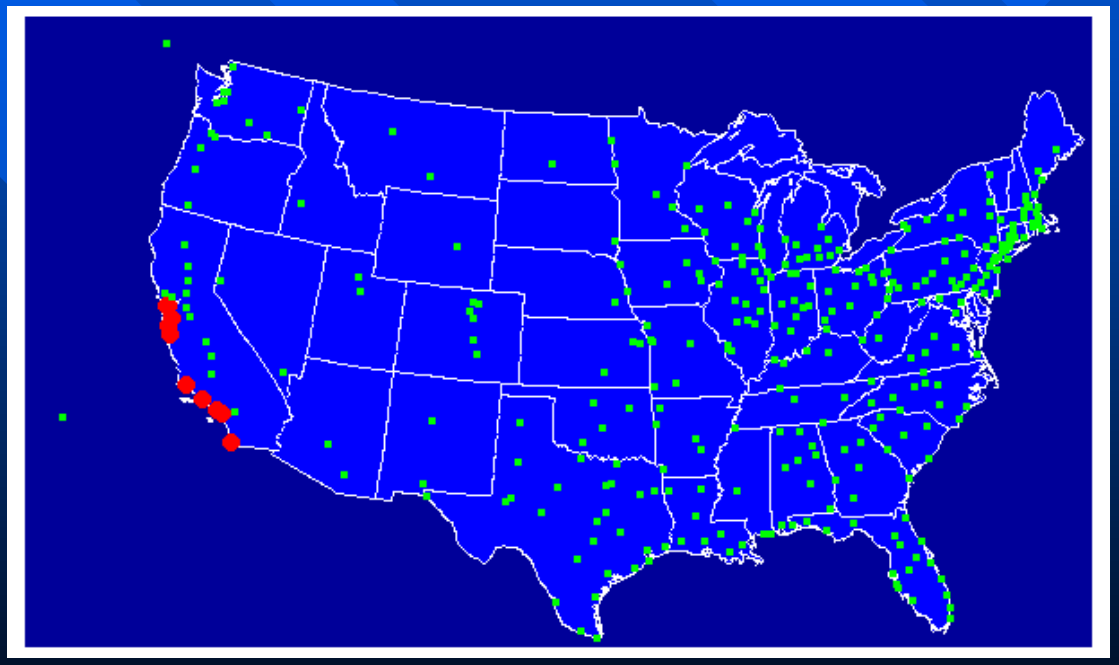
DSG Concepts (2)

- 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*

Scatterplots and Linked Brushing



XGobi



Graphical Software

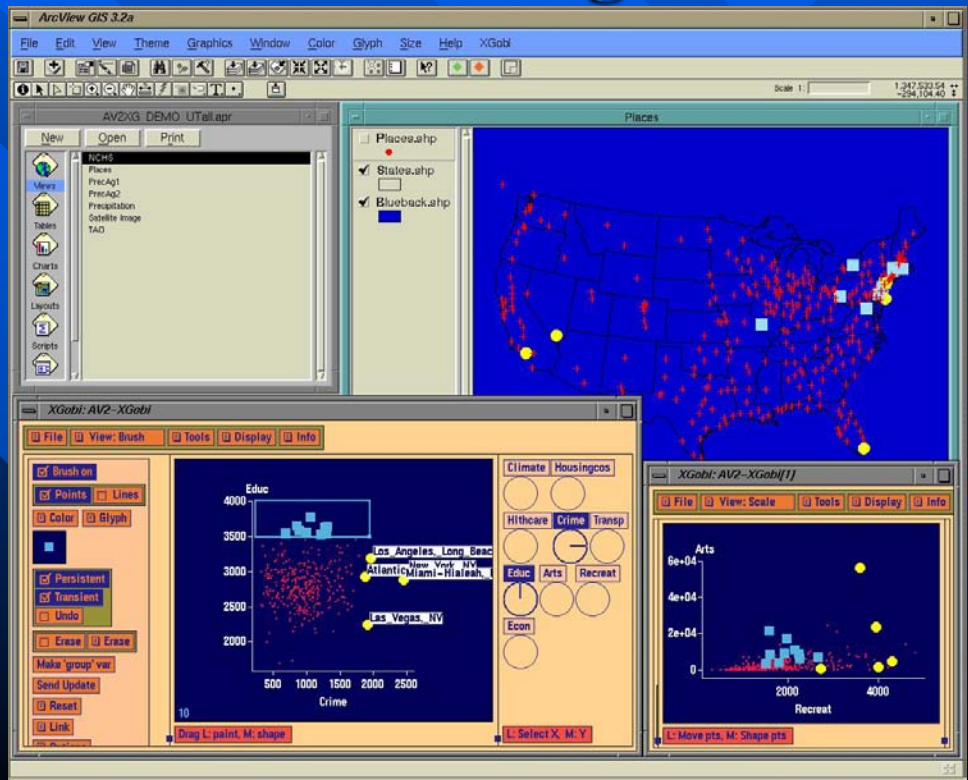
- Origin: PRIM-9 (Picturing, Rotation, Isolation and Masking in up to 9 Dimensions)
- DataViewer, XGobi*, and GGobi* Family
- REGARD, MANET, and Mondrian* Family
- EXPLOR4, HyperVision, ExplorN*, and CrystalVision* Family

DSG Software: DataViewer, XGobi, and GGobi

- 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, and Duncan Temple Lang

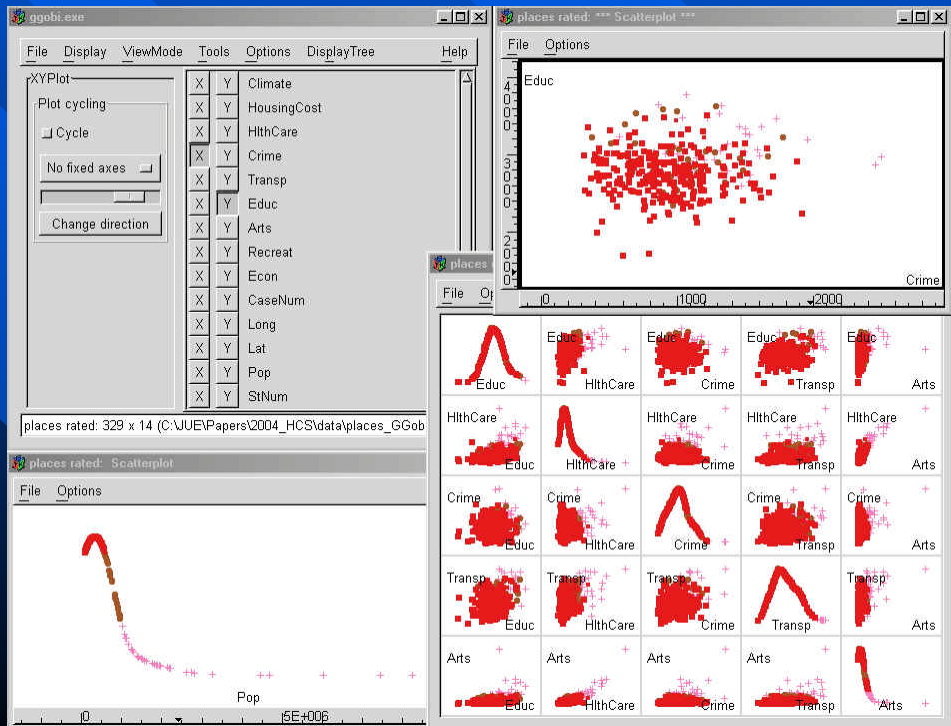
XGobi

- 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



DSG Software: EXPLOR4, HyperVision, ExplorN, and CrystalVision

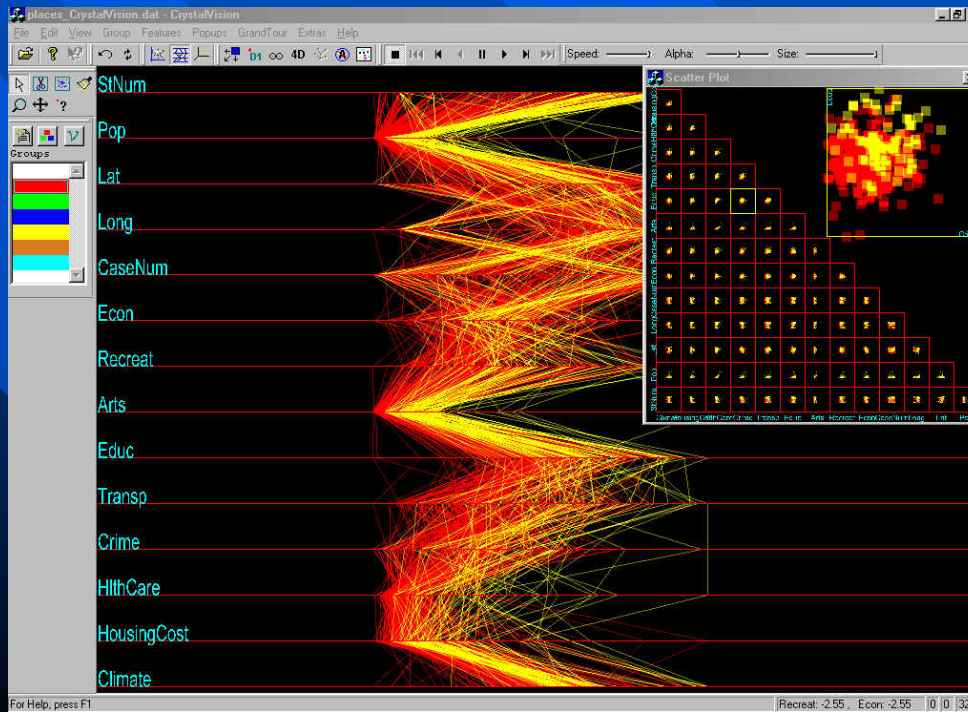
- Initiated in the late 1980's by Dan Carr and Ed Wegman at George Mason University
- Other main collaborators: Qiang Luo and Wesley L. Nicholson

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

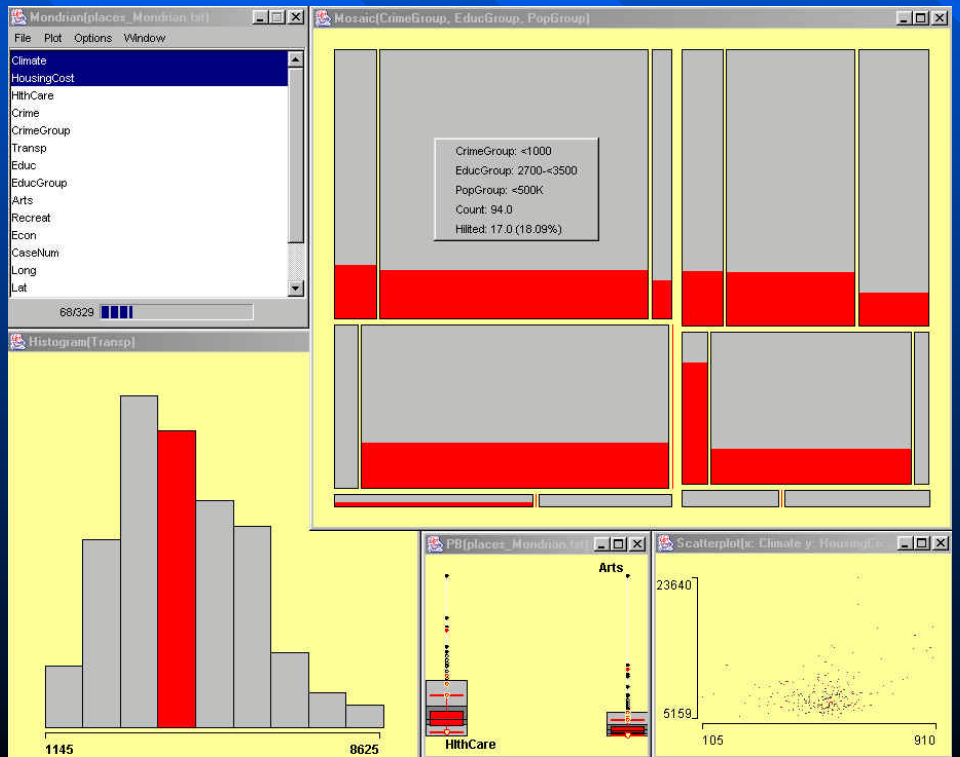


DSG Software: REGARD, MANET, and Mondrian

- 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
- Other main collaborators: Heike Hofmann, Martin Theus, Adalbert Wilhelm, and Graham Wills

Mondrian

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



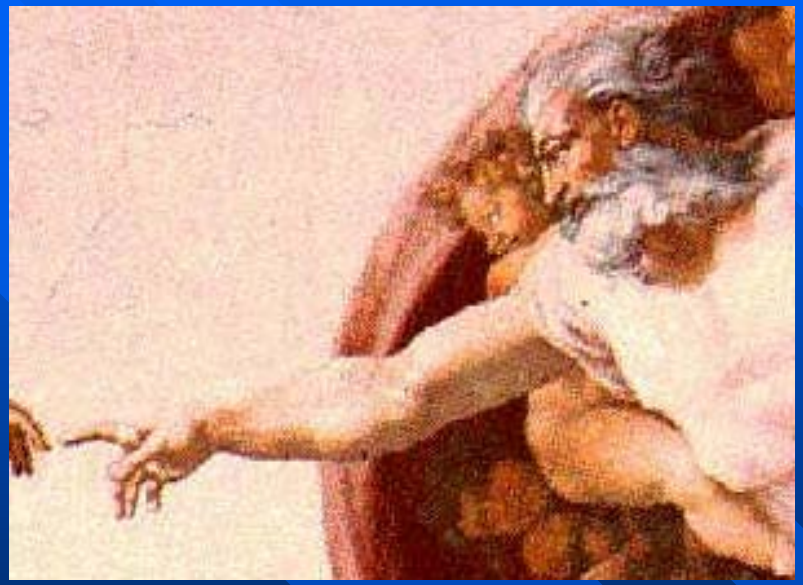
- **Example:**
  Human Hand & Arm Movements

- **Reference:**
  Vandersluis, J. P., Cooke, J. D., Ascoli, G. A., Krichmar, J. L., Michaels, G. S., Montgomery, M., Symanzik, J., Vitucci, B. (1998): Exploratory Statistical Graphics for an Initial Motion Control Experiment, *Computing Science and Statistics*, Vol. 30, 482-487.

Purpose of Experiments

- Rehabilitation of people after accidents
- Knowledge of adaptation of humans to perform mechanical tasks, e.g., arm movement
- Perfection of movements
  - Dancers
  - Ski jumpers
  - Piano players

Inspiration of Arm Movement



http://www.christusrex.org/www1/sistine/4-Genesis.html

Movements in Virtual Reality

- Data collected:
  - 3 Flock-of-Birds sensors attached to one hand, forearm, and biceps
  - x, y, z locations and pitch, roll, yaw recorded
  - Data collected at a rate of 60 Hz over 20 sec
  - 60 * 20 * 3 * 6 = 21,600 Measurements overall

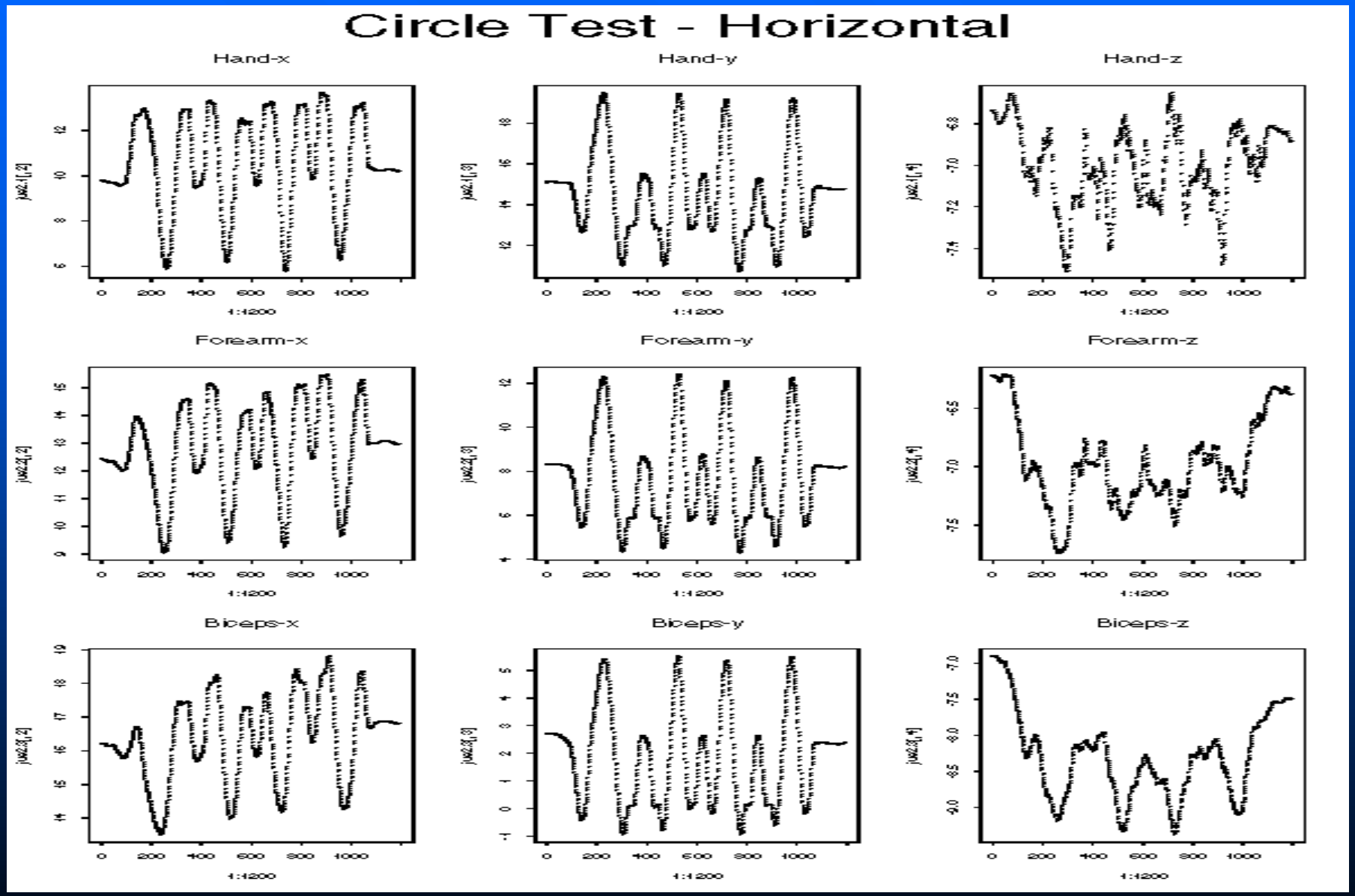
Visualization during Experiments

- Complicated setup - impossible to redo once finished
- Data plausible?
- Data correctly recorded?

Possible Problems

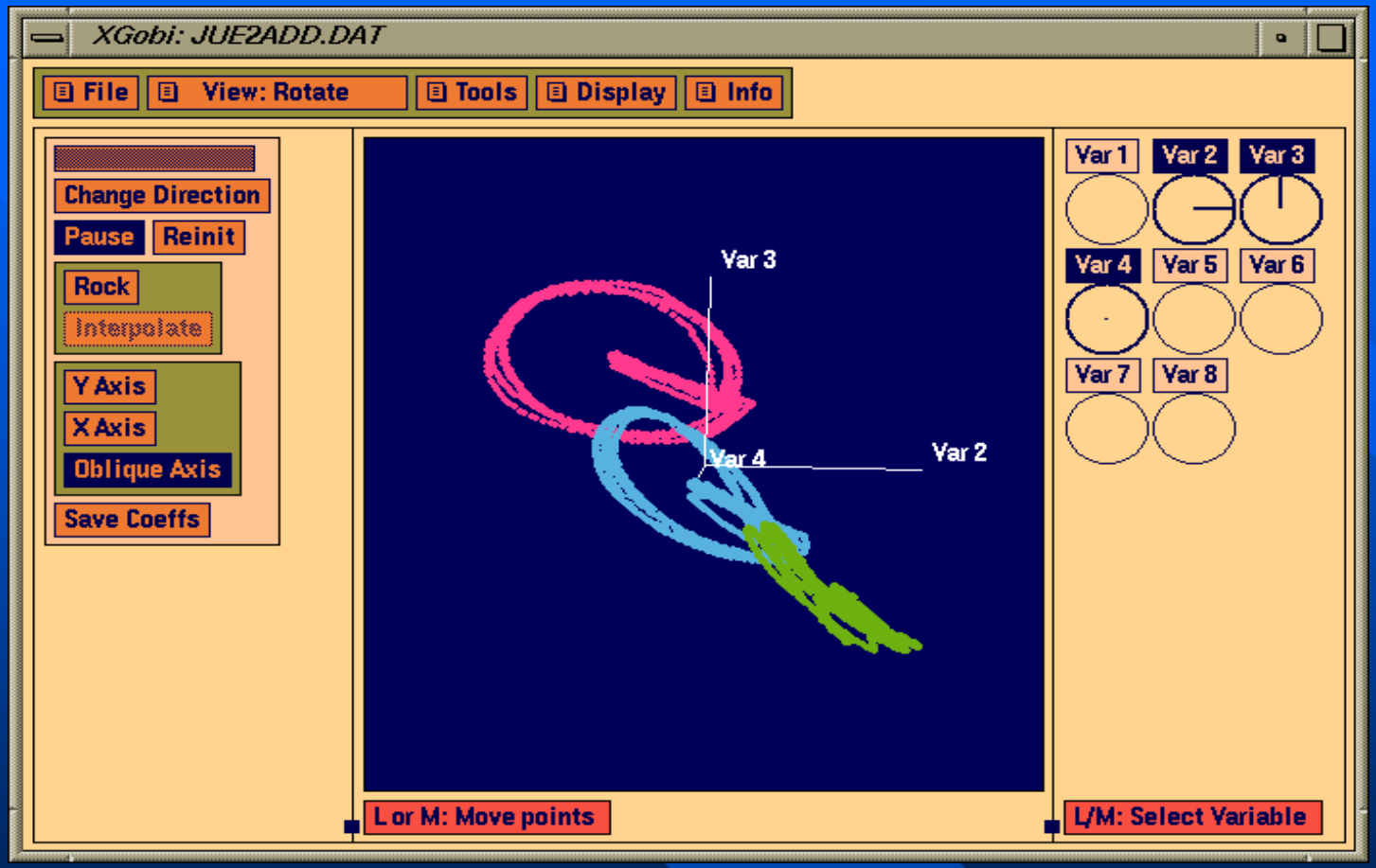
- Radio frequency reflections
- Changing fields
- Displacement of hardware attached to human body

Timeseries Plots (S-Plus)

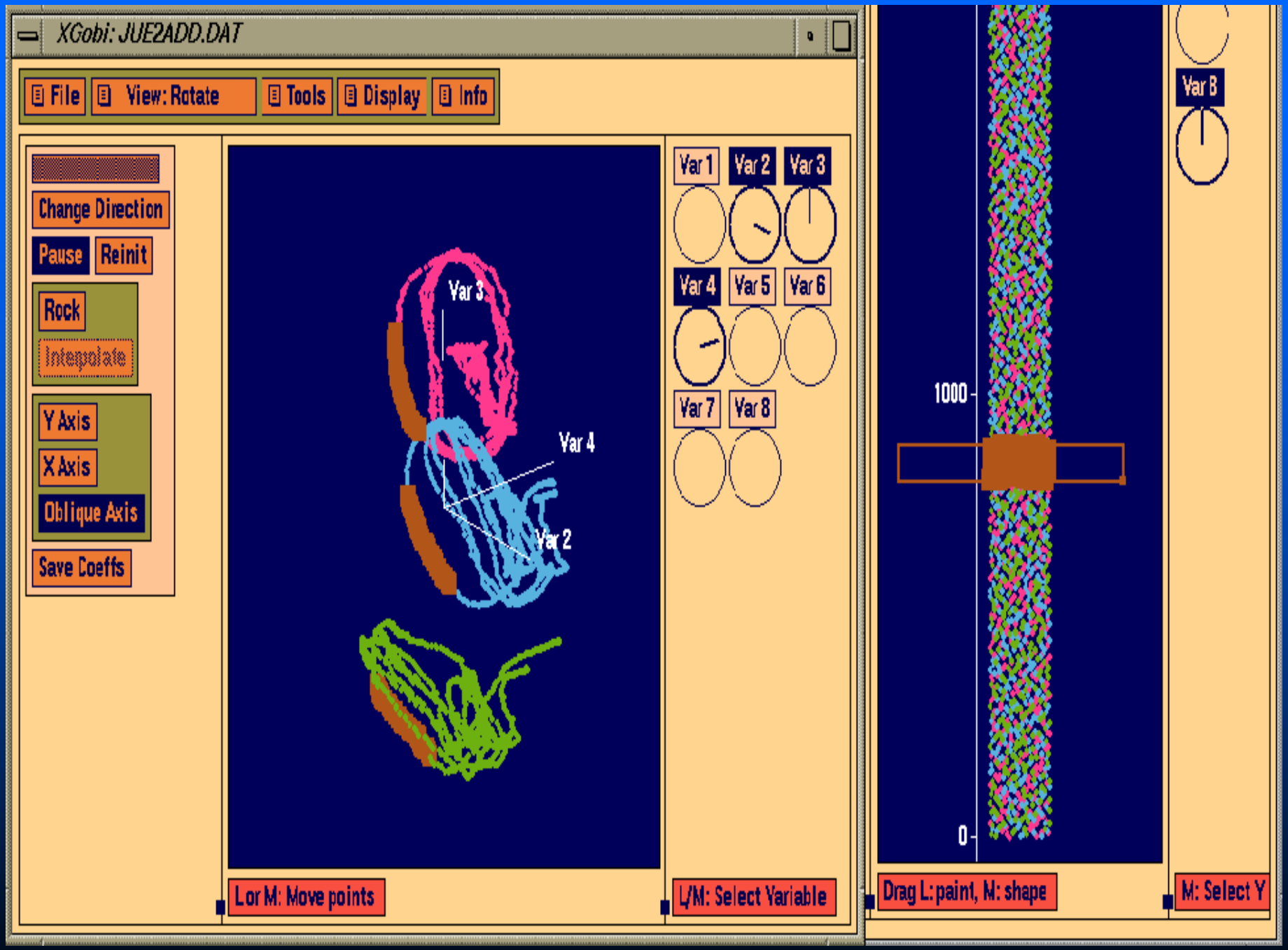


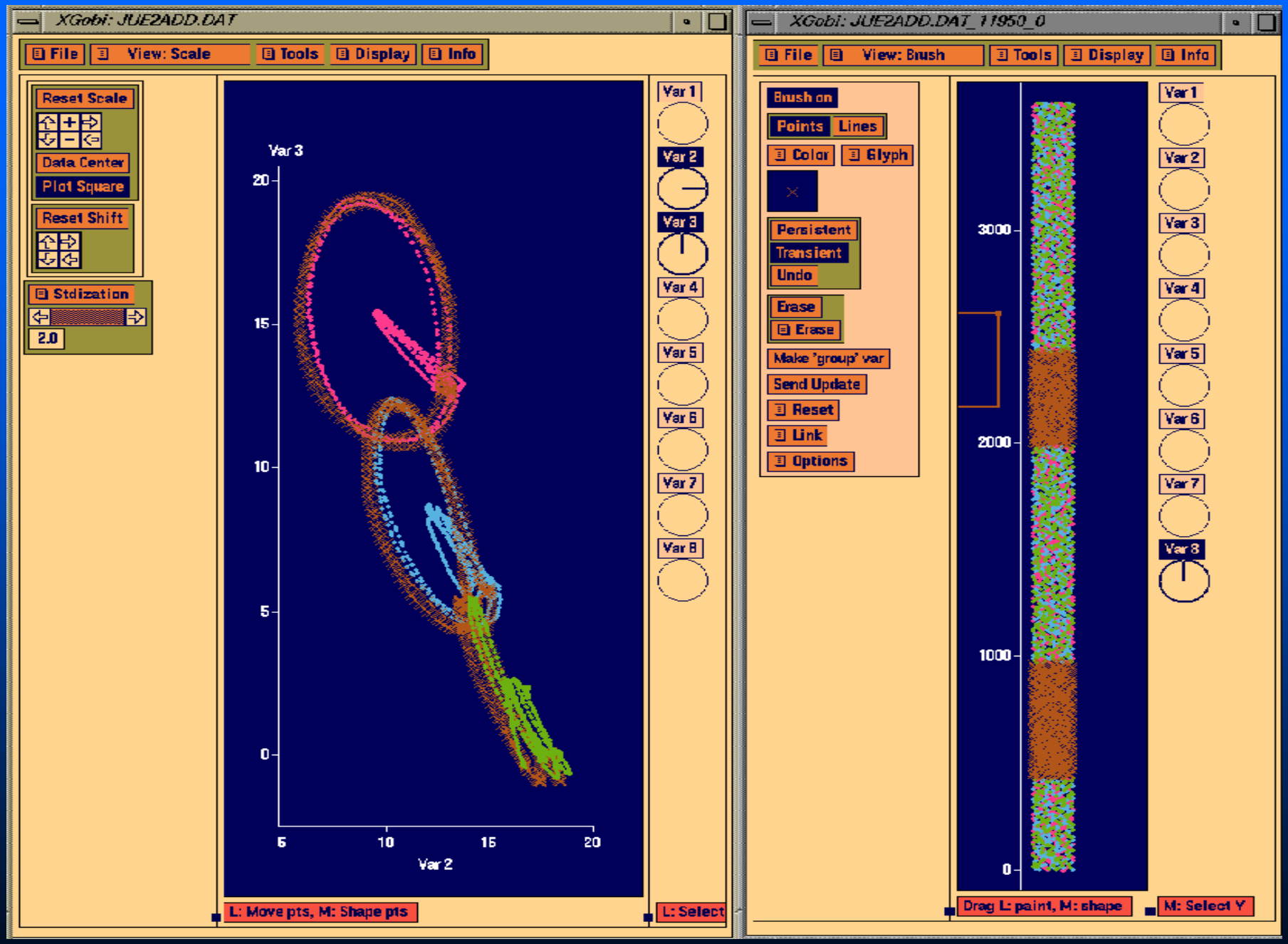
Circle Test - Angular

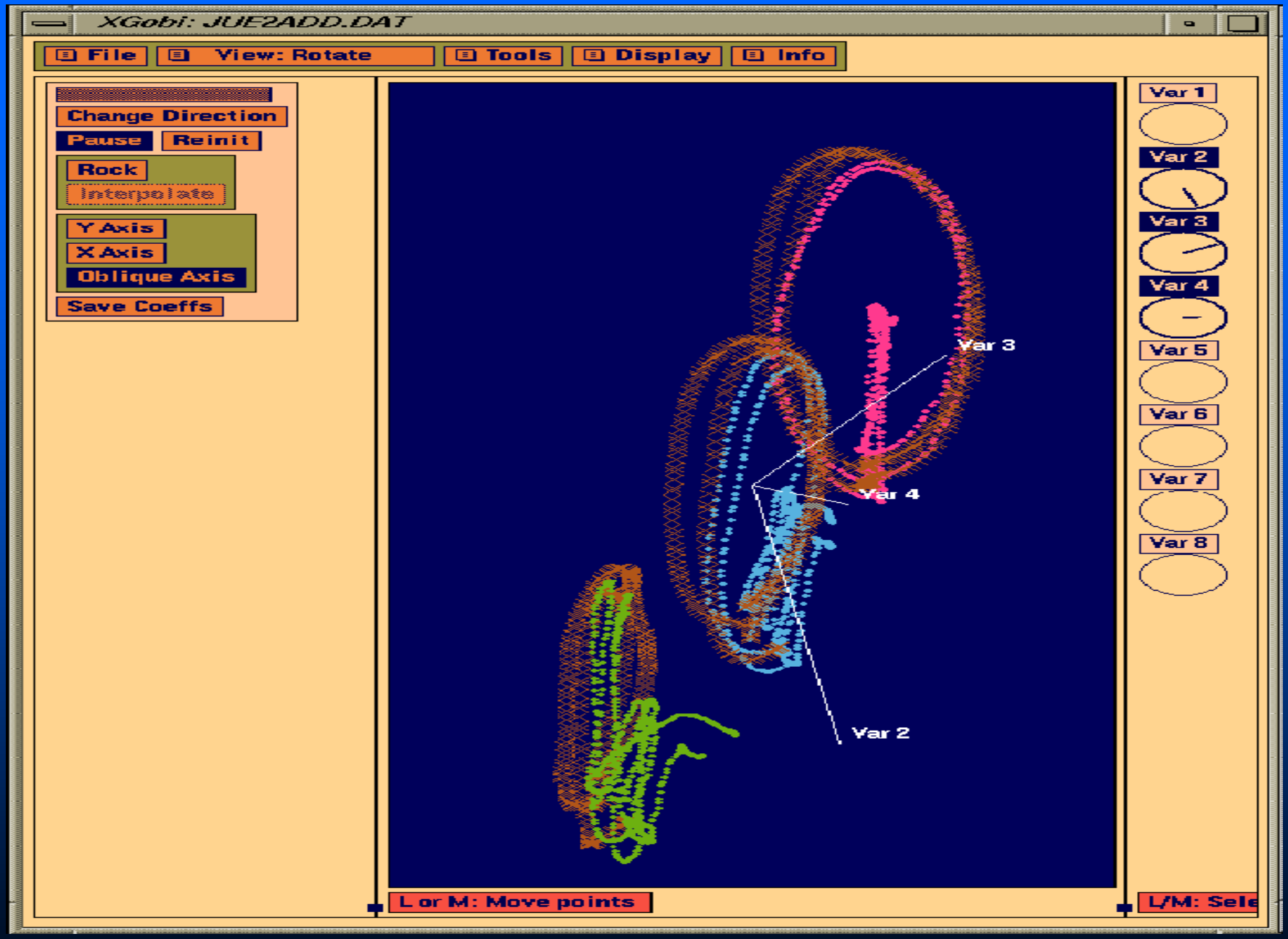
Scatterplots and Rotation (XGobi/GGobi)



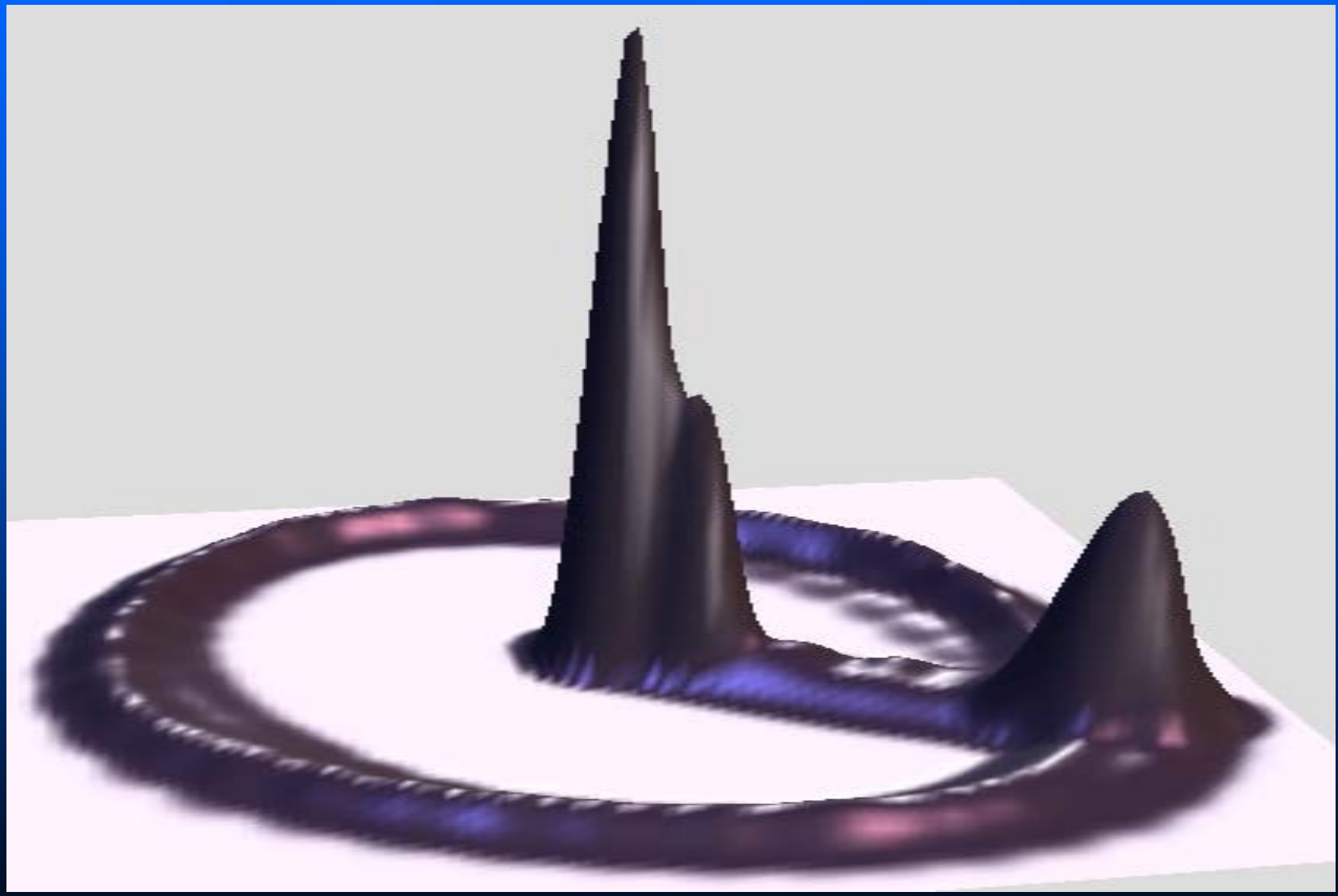
- GGobi Live Demo


XGobi: JUE2ADD.DAT
File
View: Rotate
Tools
Display
Info
Change Direction
Pause
Reinit
Rock
Interpolate
Y Axis
X Axis
Oblique Axis
Save Coeffs
Var 3
Var 4
Var 2
Var 1
Var 2
Var 3
Var 4
Var 5
Var 6
Var 7
Var 8
1000
0
Var 8
L or M: Move points
L/M: Select Variable
Drag L: paint, M: shape
M: Select Y


XGobi: JUE2ADD.DAT
File
View: Scale
Tools
Display
Info
Reset Scale
Data Center
Plot Square
Reset Shift
Stdization
2.0
Var 3
20
15
10
5
0
5
10
15
20
Var 2
Var 1
Var 2
Var 3
Var 4
Var 5
Var 6
Var 7
Var 8
L: Move pts, M: Shape pts
L: Select
XGobi: JUE2ADD.DAT_11950_0
File
View: Brush
Tools
Display
Info
Brush on
Points
Lines
Color
Glyph
Persistent
Transient
Undo
Erase
Erase
Make 'group' var
Send Update
Reset
Link
Options
3000
2000
1000
0
Var 1
Var 2
Var 3
Var 4
Var 5
Var 6
Var 7
Var 8
Drag L: paint, M: shape
M: Select Y


XGobi: JUE2ADD.DAT
File
View: Rotate
Tools
Display
Info
Change Direction
Pause
Reinit
Rock
Interpolate
Y Axis
X Axis
Oblique Axis
Save Coeffs
Var 1
Var 2
Var 3
Var 4
Var 5
Var 6
Var 7
Var 8
Var 3
Var 4
Var 2
L or M: Move points
L/M: Sele

Density Plots (ExplorN)

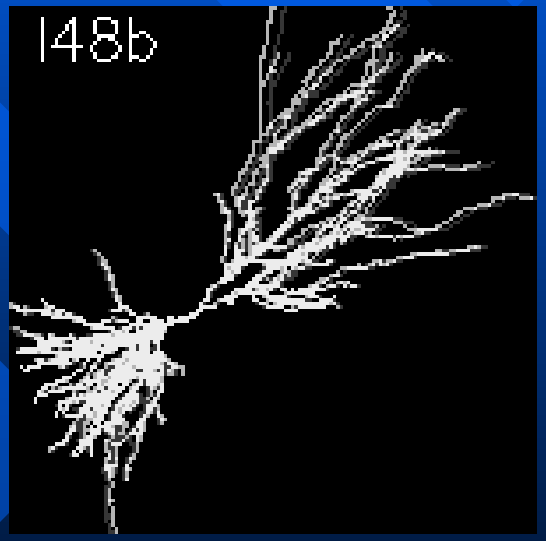


- **Example:**
  Firing Behavior of Pyramidal Brain Cells

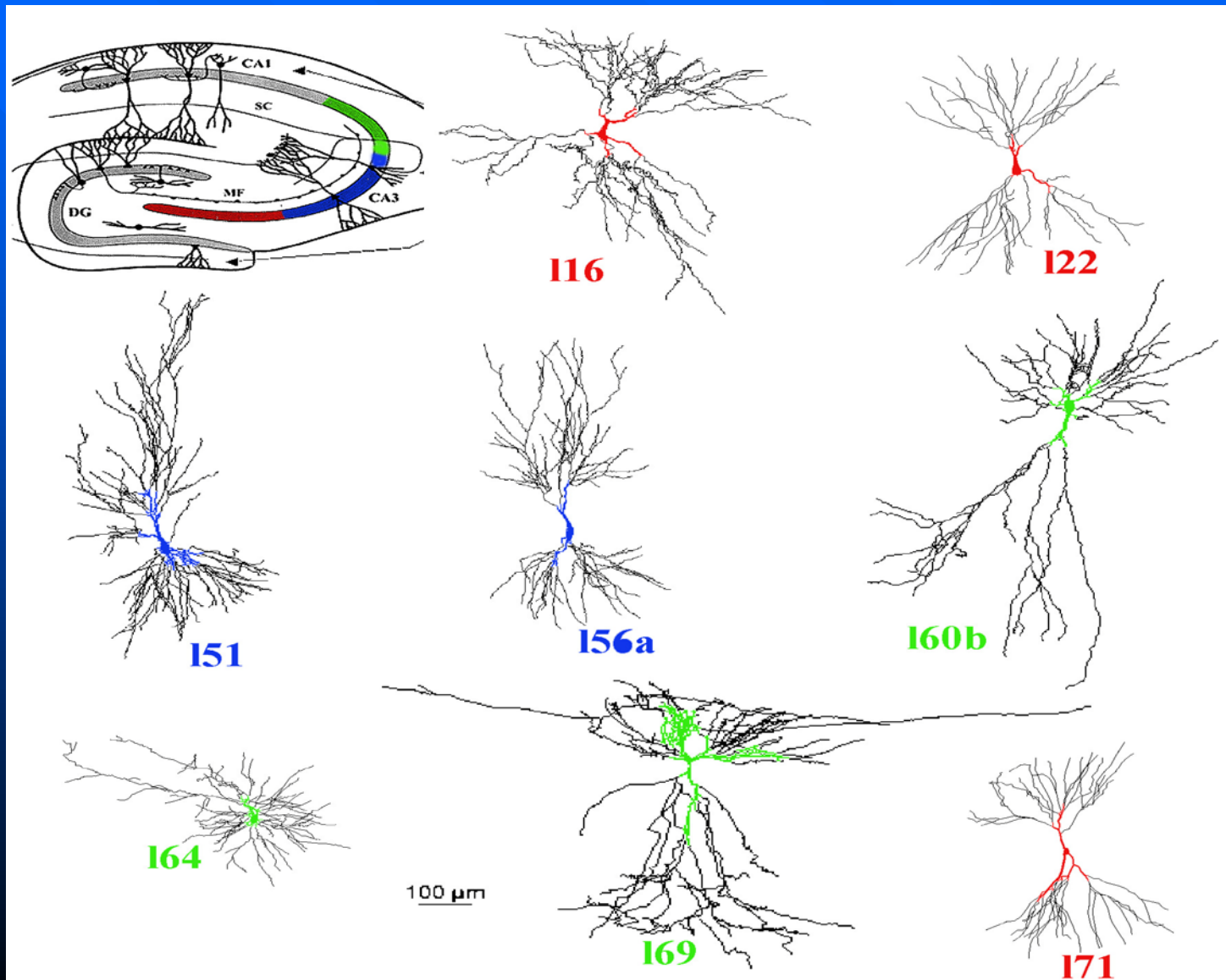
  Reference:
  Symanzik, J., Ascoli, G. A., Washington, S. S., Krichmar, J. L. (1999): Visual Data Mining of Brain Cells, *Computing Science and Statistics*, Vol. 31, 445-449.

Data Archives

Public Morphological Archive:
http://www.neuro.soton.ac.uk
~200 hippocampal neurons
(pyramidal, chandelier, etc.)



Pyramidal Brain Cells



Morphological Parameters

- Apical Dendrite
- Basal Dendrite

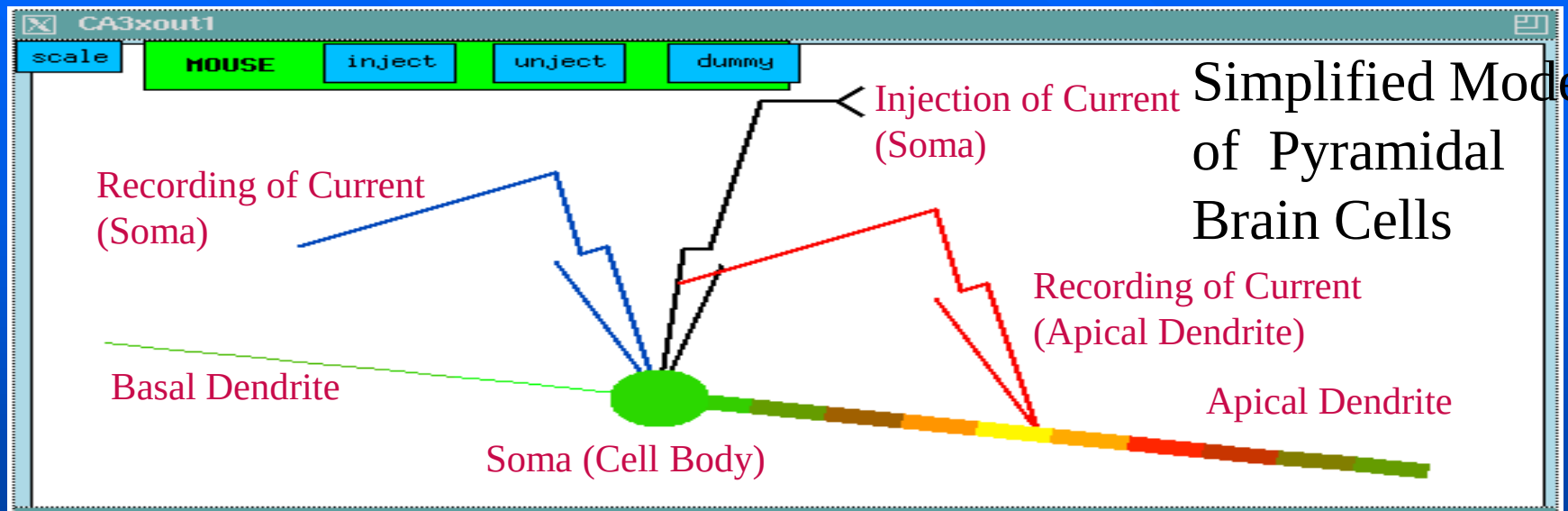
- Distance from Soma
  - 50 um
  - 100 um
  - 150 um
  - 200 um
  - Entire Dendrite Tree

- Length
- Diameter
- Area
- Asymmetry
- Bifurcations
- Terminations

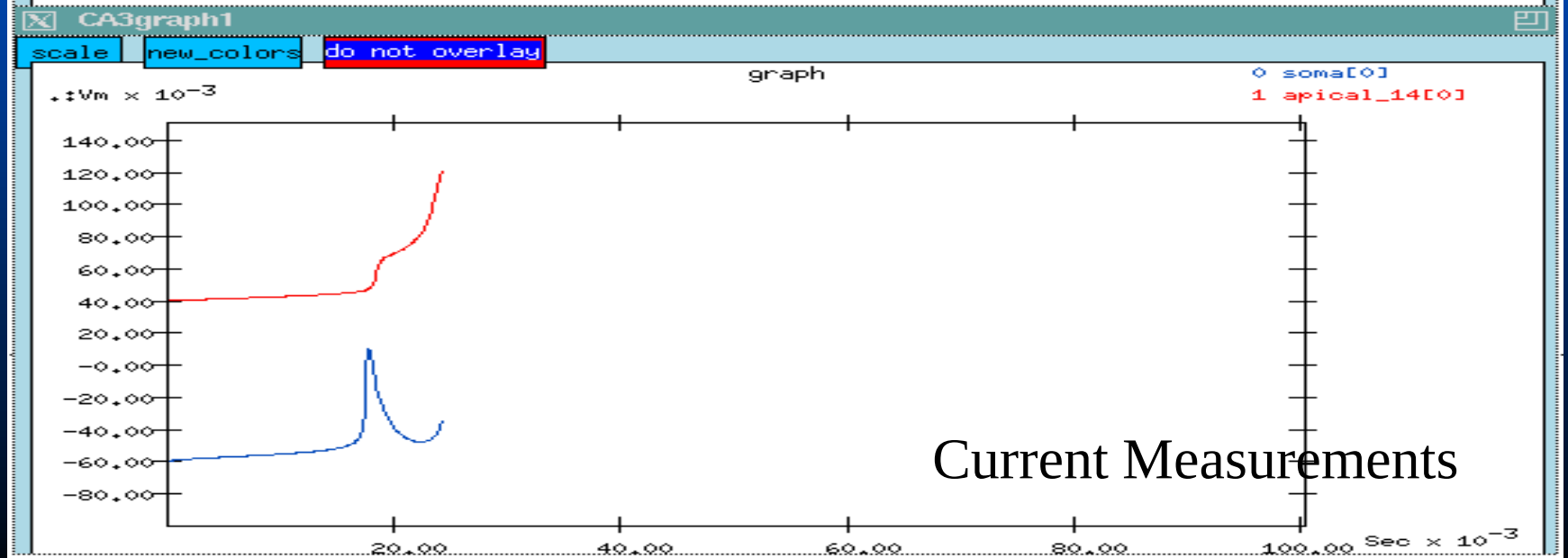
Aim of the Study

- Study the function of neurons by injecting current into a neuron and measure the neuron's response
- Here: Computational Simulator
- 16 sets of morphometric data used
- About 3 hours of computer time for 5 sec of neuron time on SGI Origin 200
- 10 injected currents per cell: 0.1 nA to 1.9 nA

Simulation

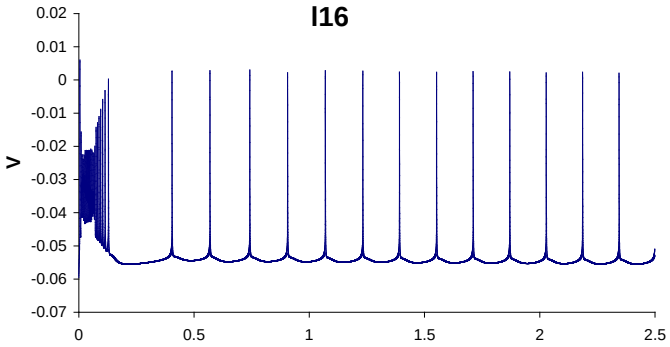


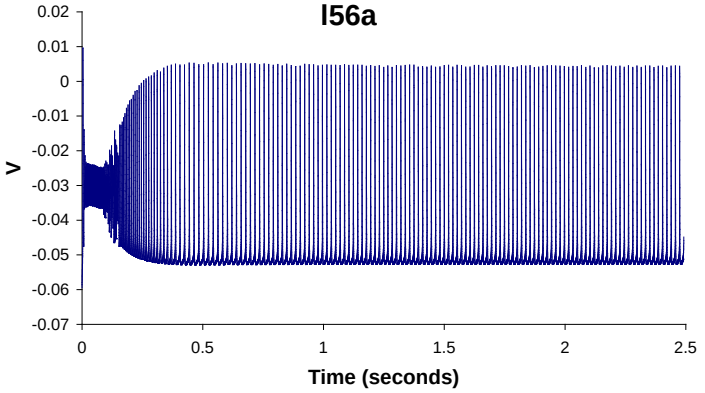
Simplified Model of Pyramidal Brain Cells

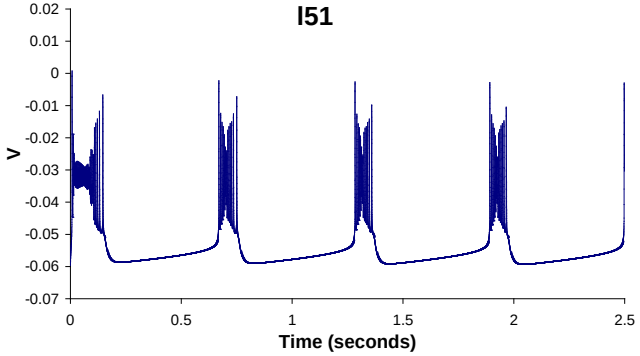


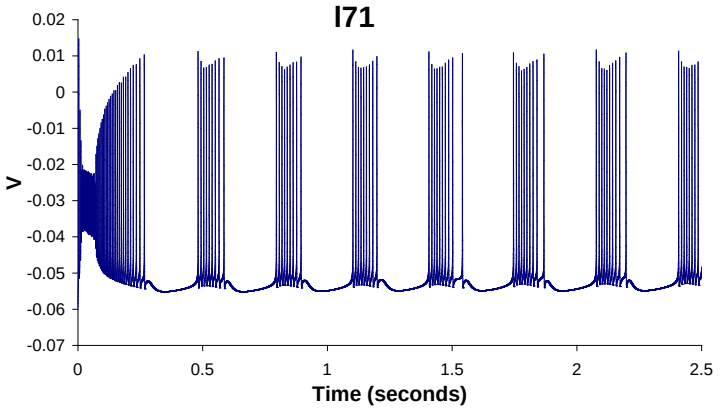
Current Measurements

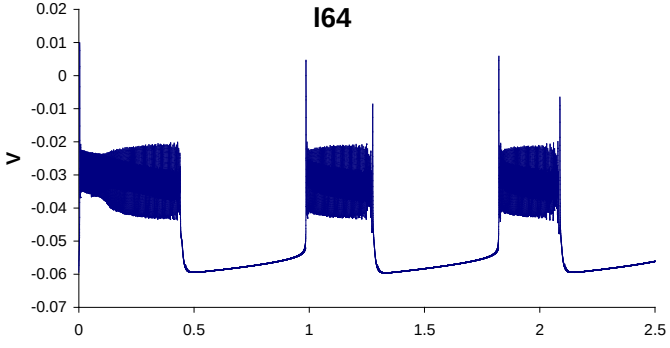
Simulated Physiological Response under 0.7 nA

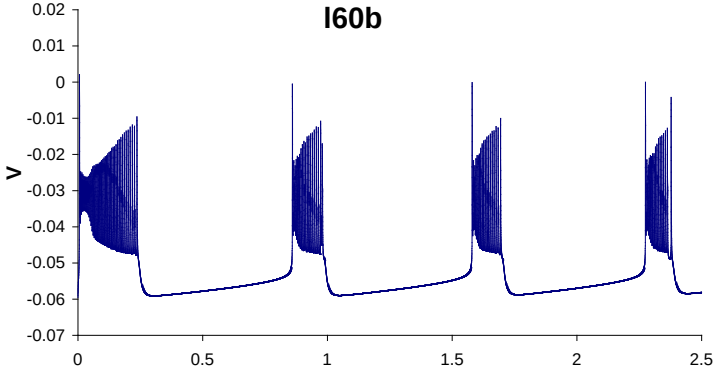






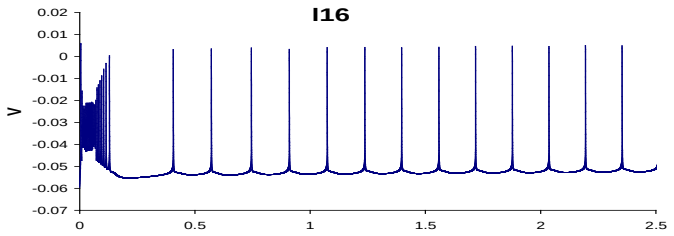




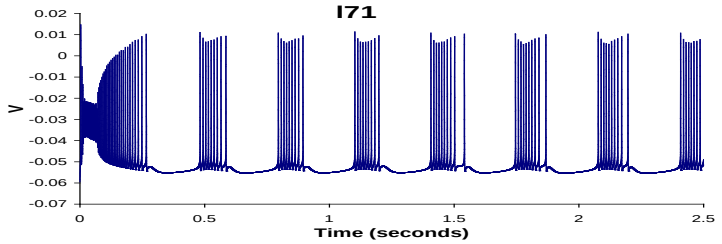


Response Parameters

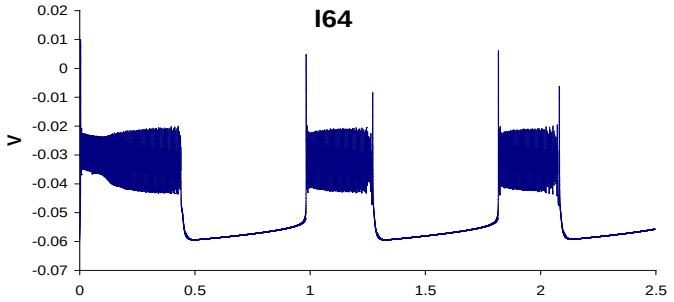
- Spiking:
  - Spike Rate (Hz)
  - Spike Transition (nA)



- Bursting:
  - Burst Rate (Hz)
  - Interburst Interval (sec)
  - Spikes per Burst (Hz)

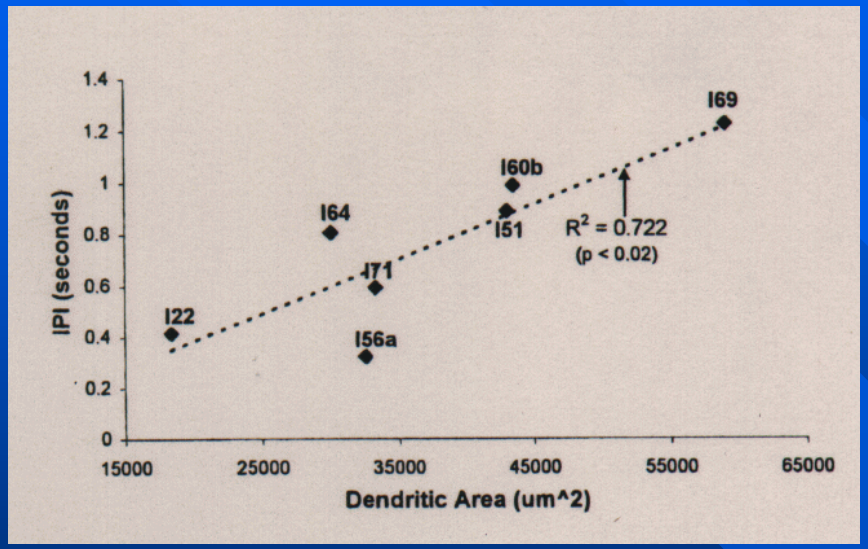


- Plateau:
  - Plateau Range (nA)
  - Plateau Rate (Hz)
  - Interplateau Interval (sec)
  - Spikes per Plateau (Hz)



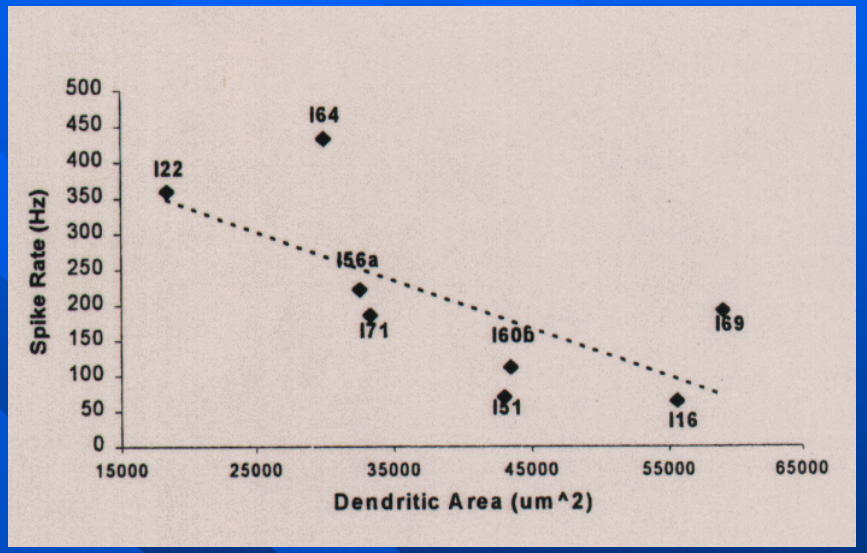
Influence of Dendritic Area on Firing Rate

Interplateau Interval vs Dendritic Area



Current: 0.5 nA

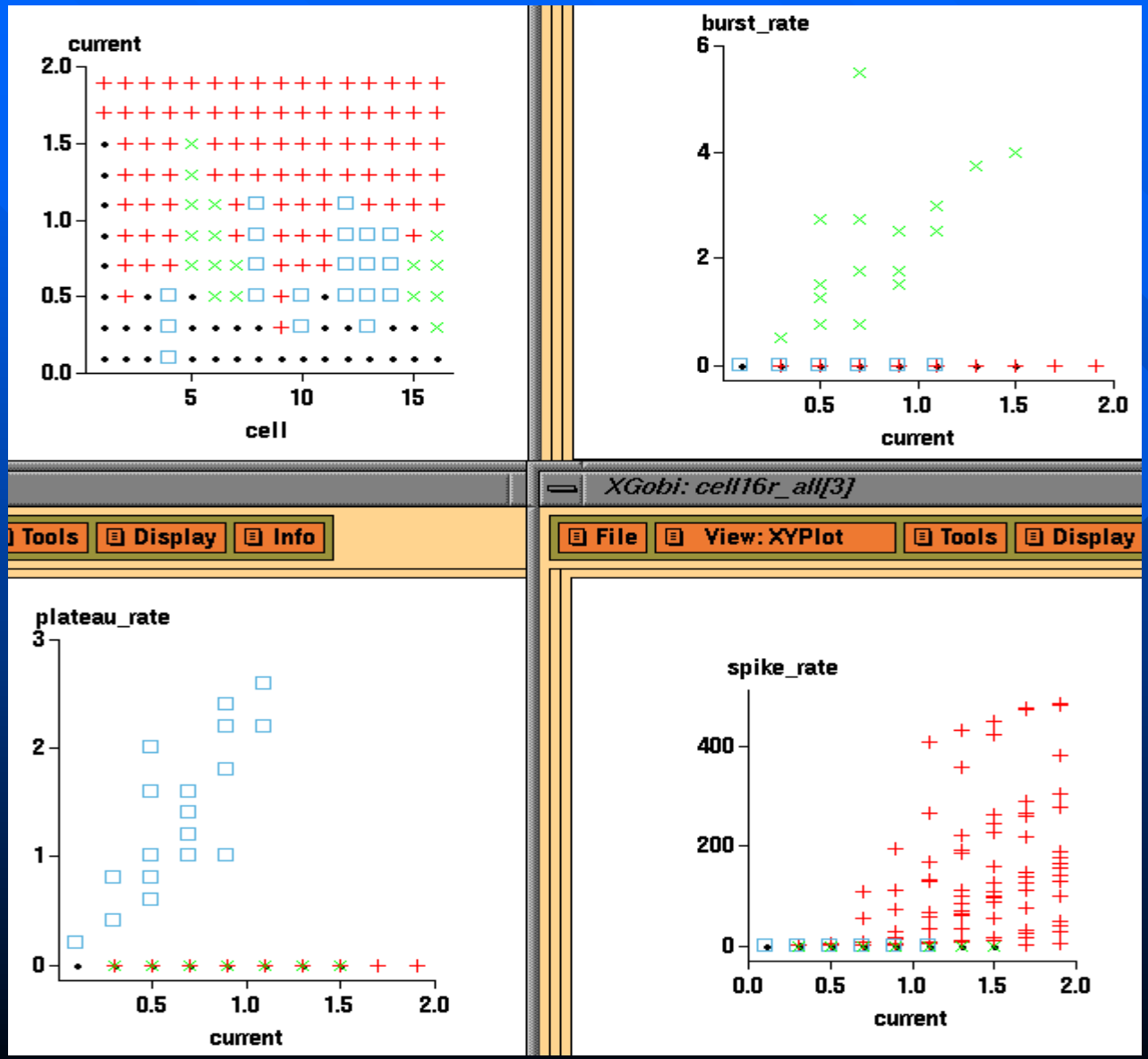
Spike Rate vs Dendritic Area



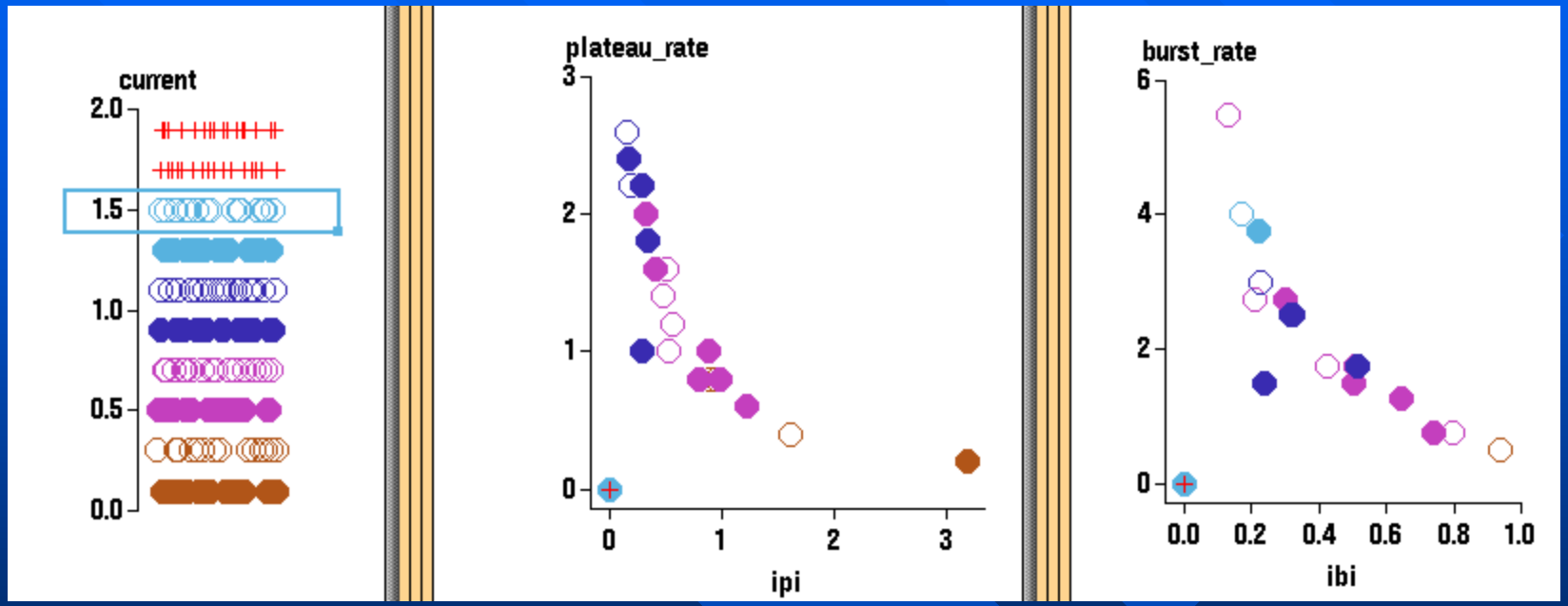
Current: 1.3 nA

- Smaller cells tend to be more excitable and have higher firing rates.

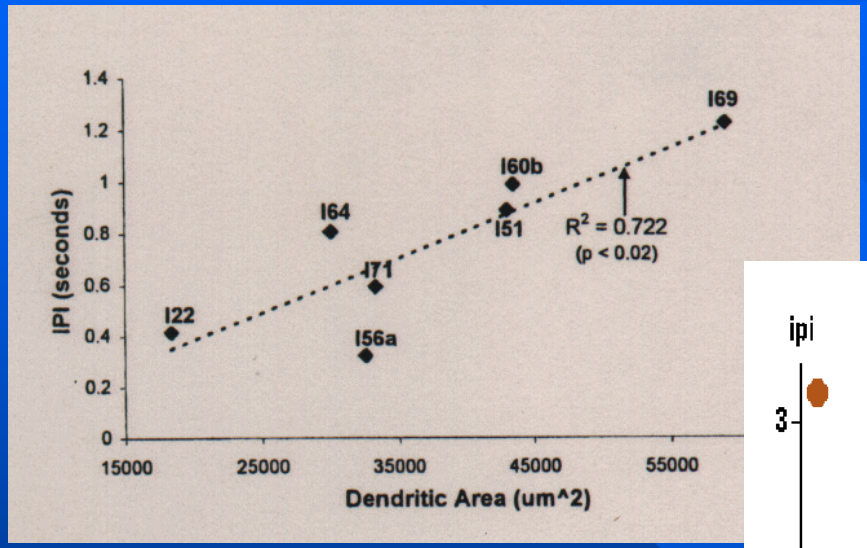
Visual Data Mining Using XGobi

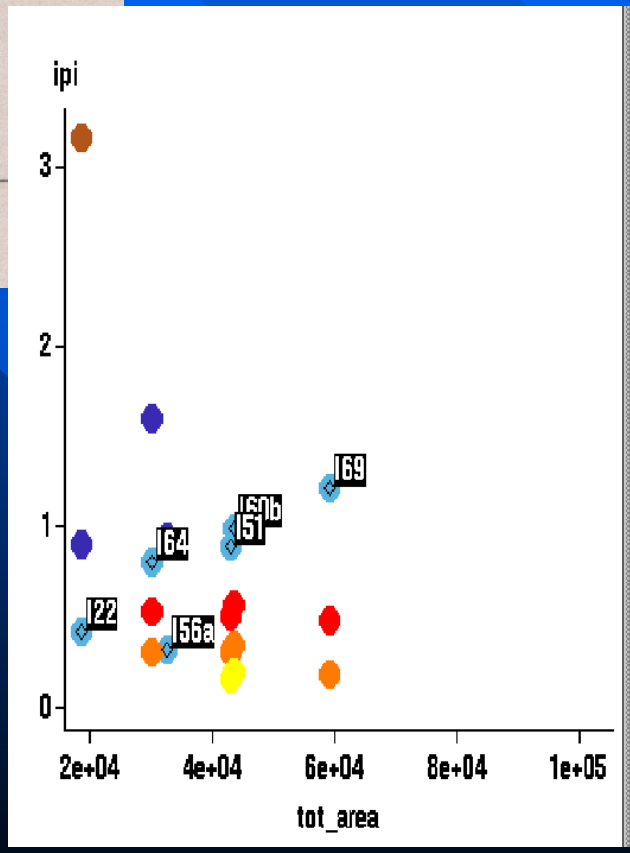


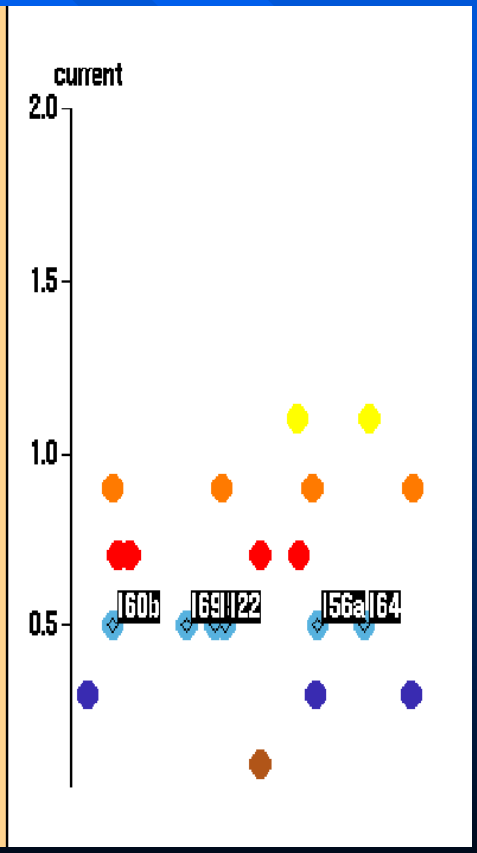
Visible Patterns



Interplateau Interval vs Dendritic Area ???







- **Example:**
  Image Grand Tour (IGT)

- **Reference:**
  Symanzik, J., Wegman, E. J., Braverman, A. J., Luo, Q. (2002): New Applications of the Image Grand Tour, *Computing Science and Statistics*, Vol. 34, CD.

Image Grand Tour (IGT)

- A method to address:
  - Multispectral imagery
  - Satellite imagery various look angles
  - Multiple time series
    - » Seismic Signals
    - » Electrocardiograms: see HealthRx CardioView ECG Analysis Software by J. Patrick Vandersluis, http://www.healthrx.com/patrick/
- Original Developers:
  Ed Wegman, Wendy Poston, Jeff Solka

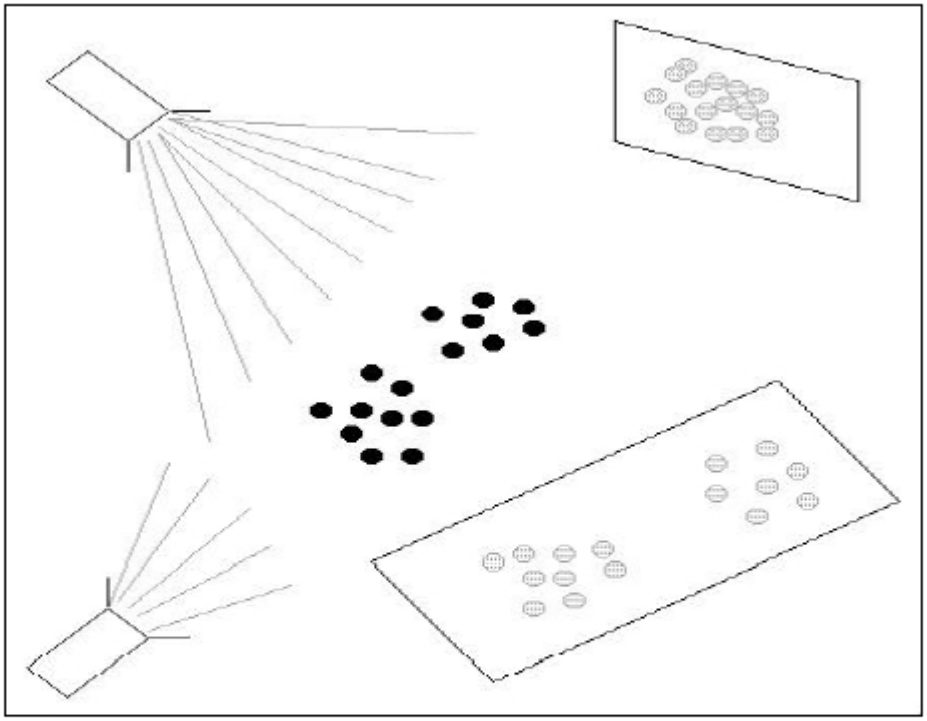
Grand Tour (1)

The Grand Tour is motivated by the idea of looking at an object from all possible points of view. From a high-dimensional data visualization point of view, we want to look at a point cloud from all perspectives.

Types of tours:

- 2-dimensional (Asimov,1985), (Buja and Asimov, 1985)
- d-dimensional (Wegman, 1991)

Grand Tour (2)



This simple diagram, borrowed from Guy Nason , University of Bristol, UK (http://www.stats.bris.ac.uk/~guy/pjg/SimplePP.html) illustrates the principle behind projection of 3-dimensional data onto 2-dimensional surfaces. Notice that from one projection only a single cluster is visible while from another, two clusters are clearly visible

Grand Tour (3)

Two key components of a grand tour:

- Space Filling
- Continuous

Smoothly rotate a d-dimensional coordinate system indexed by time so that it "occupies all possible orientations." Project data into rotated coordinate system. Look for "interesting" configurations.

Minefield Snapshot (1)

Minefield Snapshot (2)

Minefield Snapshot (3)

Minefield Movie (4)

IGT Movie: Minefield

Minefield Snapshot (5)

- **Example:**
  Presentation Graphics via Micromaps

- **Reference:**
  Symanzik, J., Gebreab, S., Gillies, R. Wilson, J. (2003): Visualizing the Spread of West Nile Virus, *2003 Proceedings*, American Statistical Association, Alexandria, Virginia, CD .

Micromaps

- Link of row-labeled univariate (or multivariate) statistical summaries to corresponding geographical region
- Focus on statistical display and not on maps
- Useful for
  - environmental data
  - agricultural data
  - medical data
  - economical data

History of Micromaps

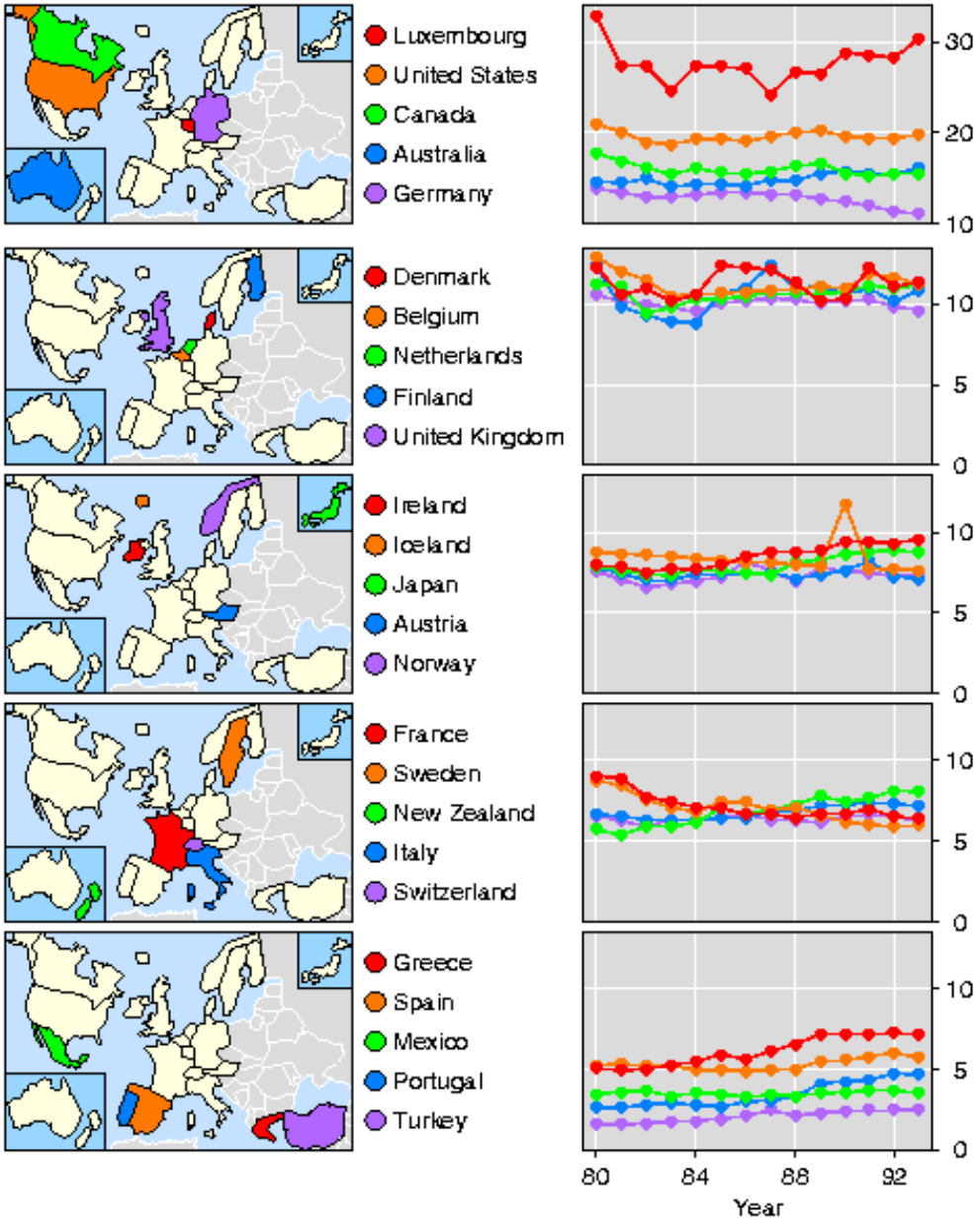
- First presented at 1995 American Statistical Association's annual meeting (Olsen, Carr, Courbois, Pierson)
- First references:
  - Carr, Pierson (1996) Emphasizing Statistical Summaries … with Micromaps, Stat. Computing & Stat. Graphics Newsletter, 7(3)
  - Carr, Olsen, Courbois, Pierson, Carr (1998) Linked Micromap Plots …, Stat. Computing & Stat. Graphics Newsletter, 9(1)

Micromap Examples

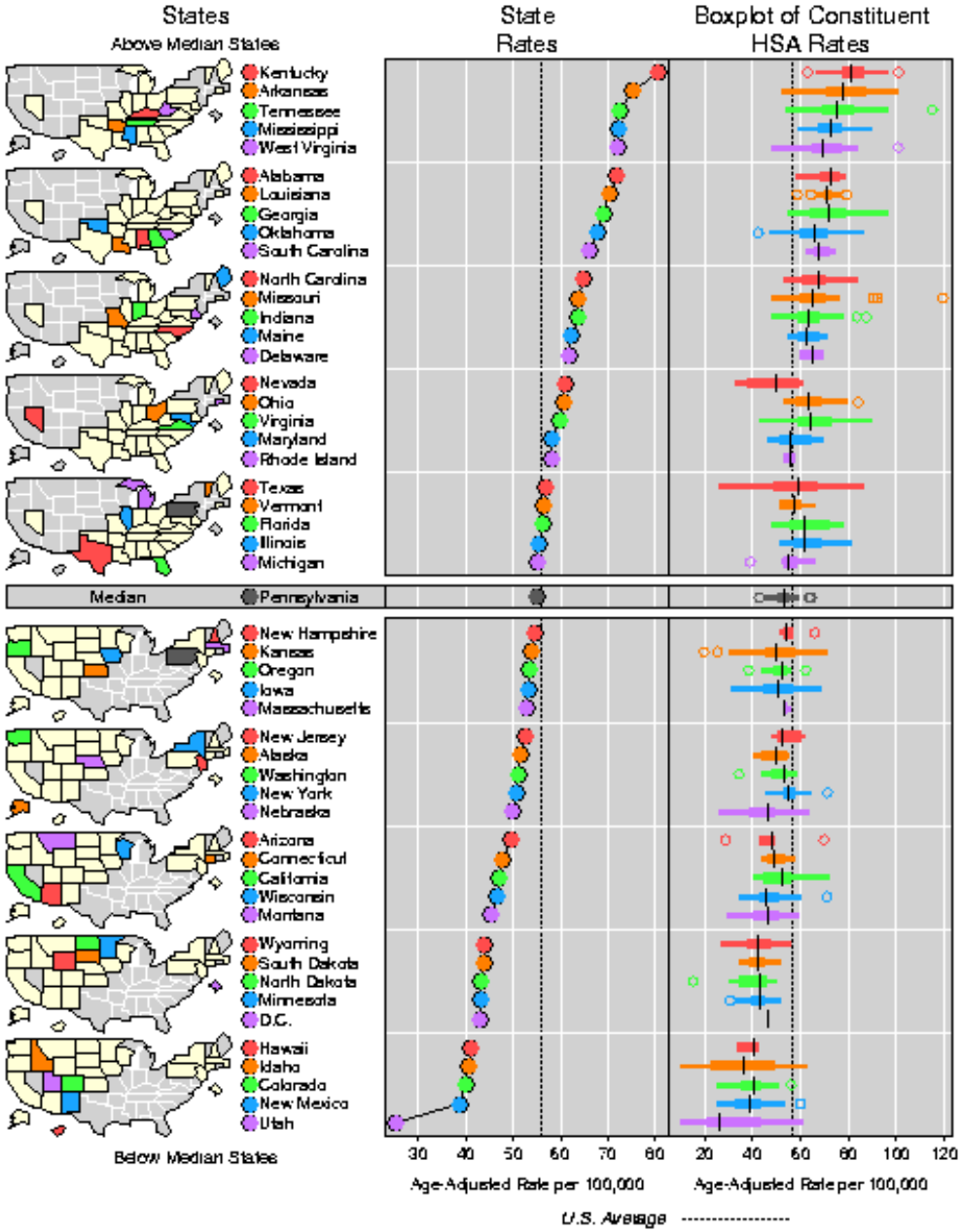
- Dan Carr's S-Plus functions available at
  - ftp://galaxy.gmu.edu/pub/dcarr/newsletter/micromap/
  - ftp://galaxy.gmu.edu/pub/dcarr/newsletter/lmplots/
- First 2 examples borrowed from Dan Carr

Annual CO2 Emissions From Energy Use

Units = Tons Per Person



Lung Cancer Mortality Rates By State
White Males, 1988-1992



Micromaps at NCI

- National Cancer Institute (NCI)
- *http://www.statecancerprofiles.cancer.gov/micromaps*
- Released in April 2003
- Cancer statistics:
  - Mortality and incidence counts and rates
  - Trends by sex and race/ethnicity
- Fully interactive
- Extensive usability testing

Wang, Chen, Carr, Bell, Pickle (2002): Geographic Statistics Visualization: Web-based Linked Micromap Plots, Computing in Science & Engineering 4(3):90-94.

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

Left Column Data

Area: US - state level
Data Group: Cancer Statistics
Cancer: Lung & Bronchus
Statistic: Incidence Rate
Race: All Races
Sex: Female
Age: All Ages

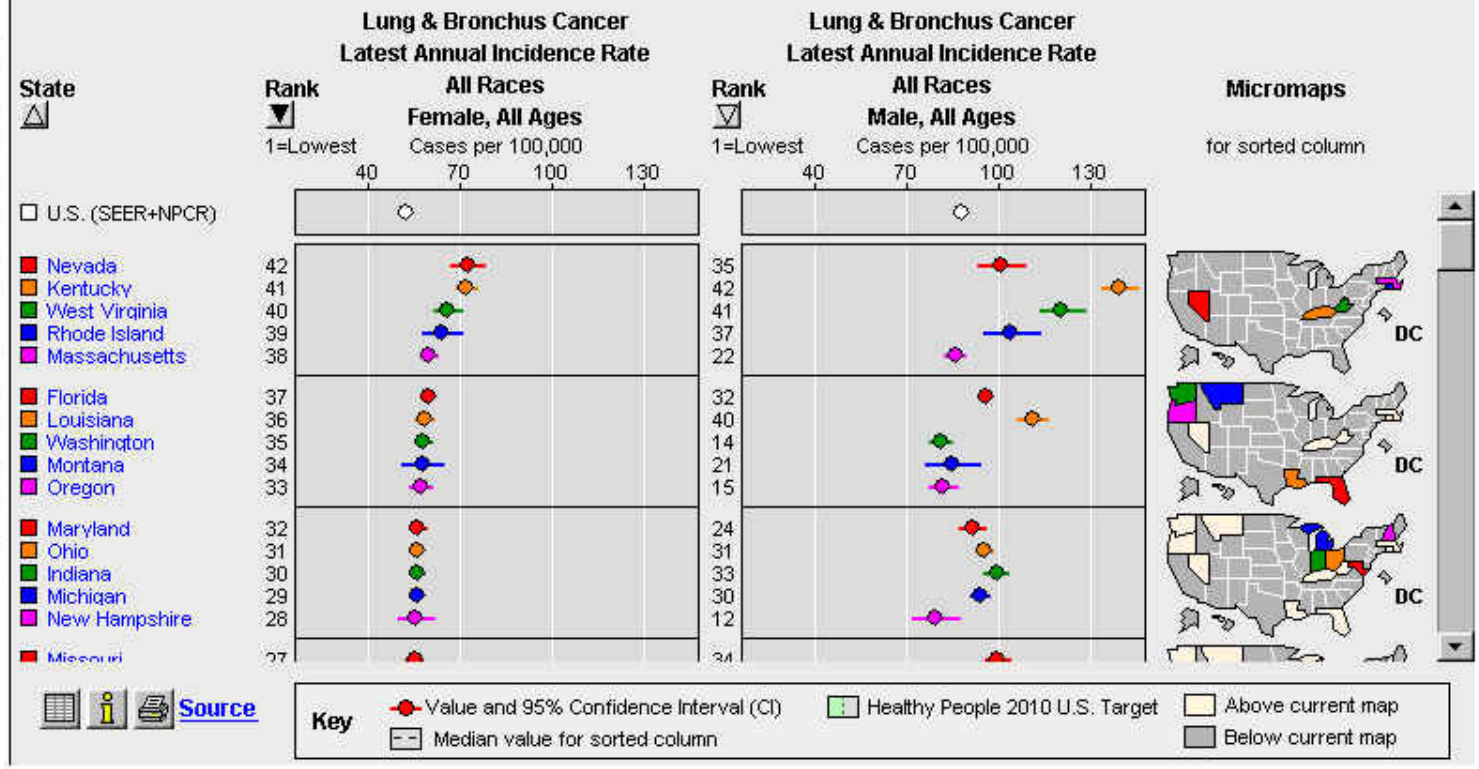
Right Column Data (optional)

Data Group: Cancer Statistics
Cancer: Lung & Bronchus
Statistic: Incidence Rate
Race: All Races
Sex: Male
Age: All Ages

Draw Clear

Overview

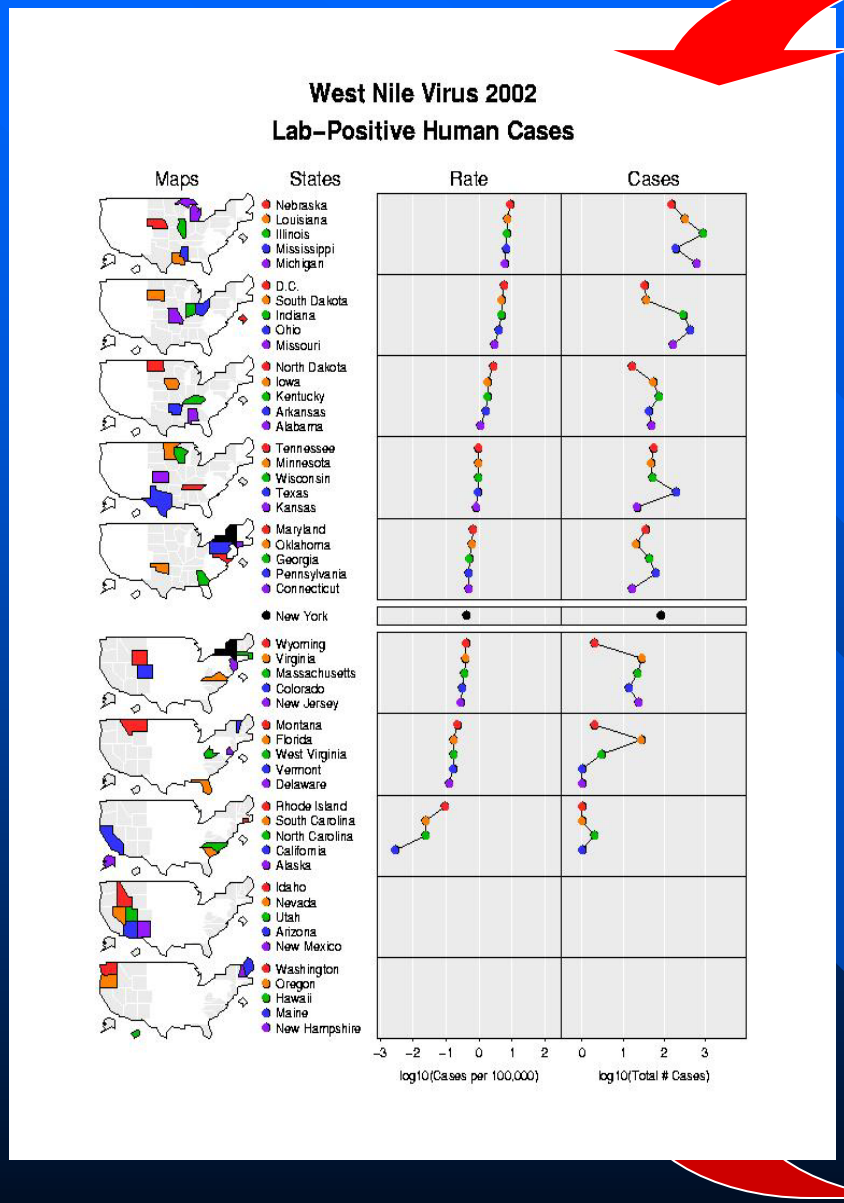
Options

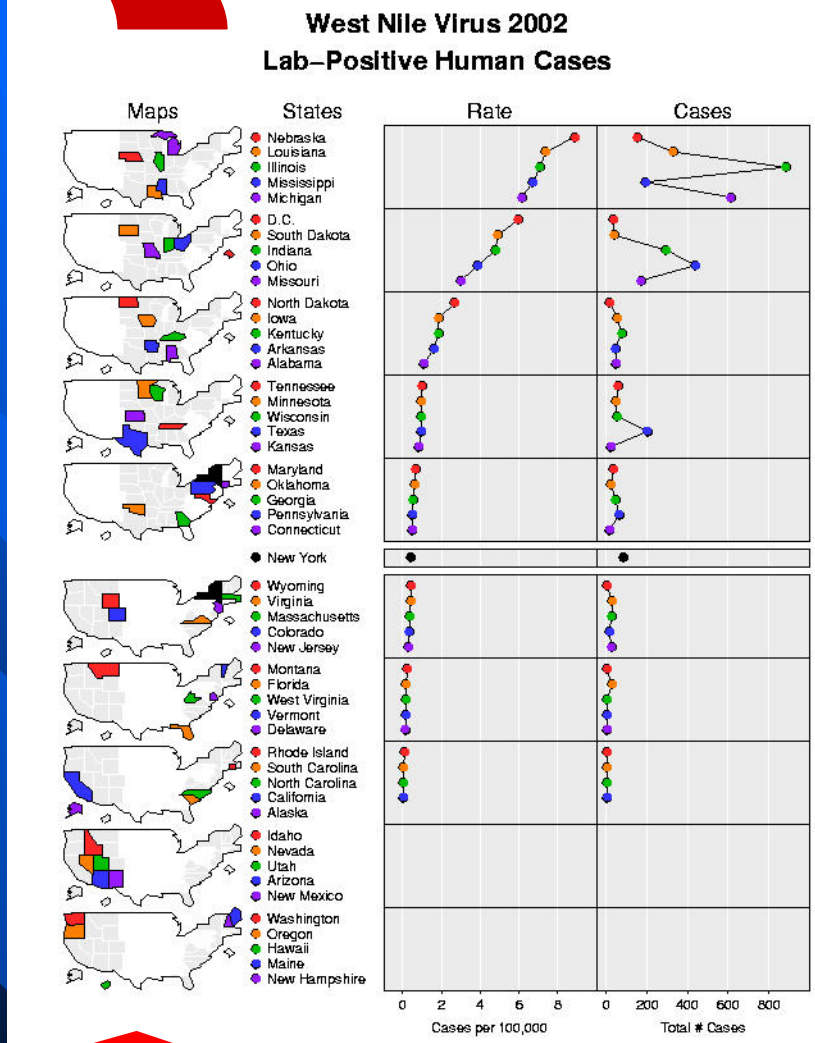


West Nile Virus (WNV)

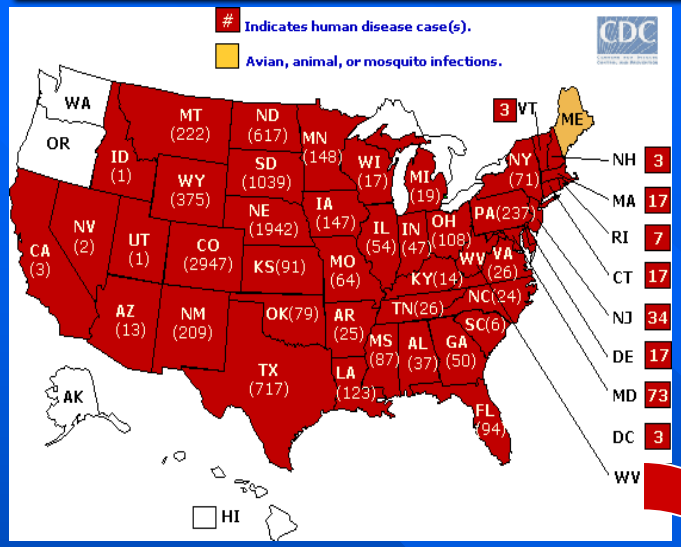
- Introduced to the US in 1999
- Spread across North America in 5 years
- Initial event - Culex mosquito transmits virus within avian populations
- Bridging Aedes albopictus transmits virus from birds to animals and humans

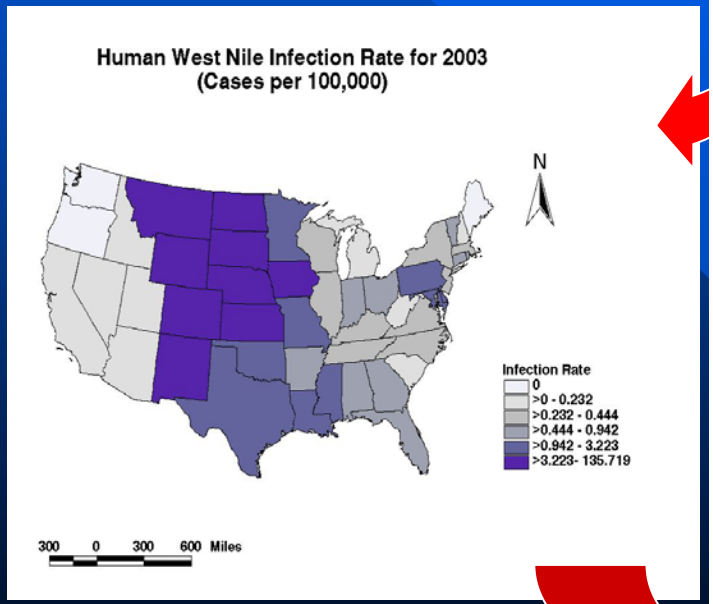
From 2002 CDC Web Page to Micromaps

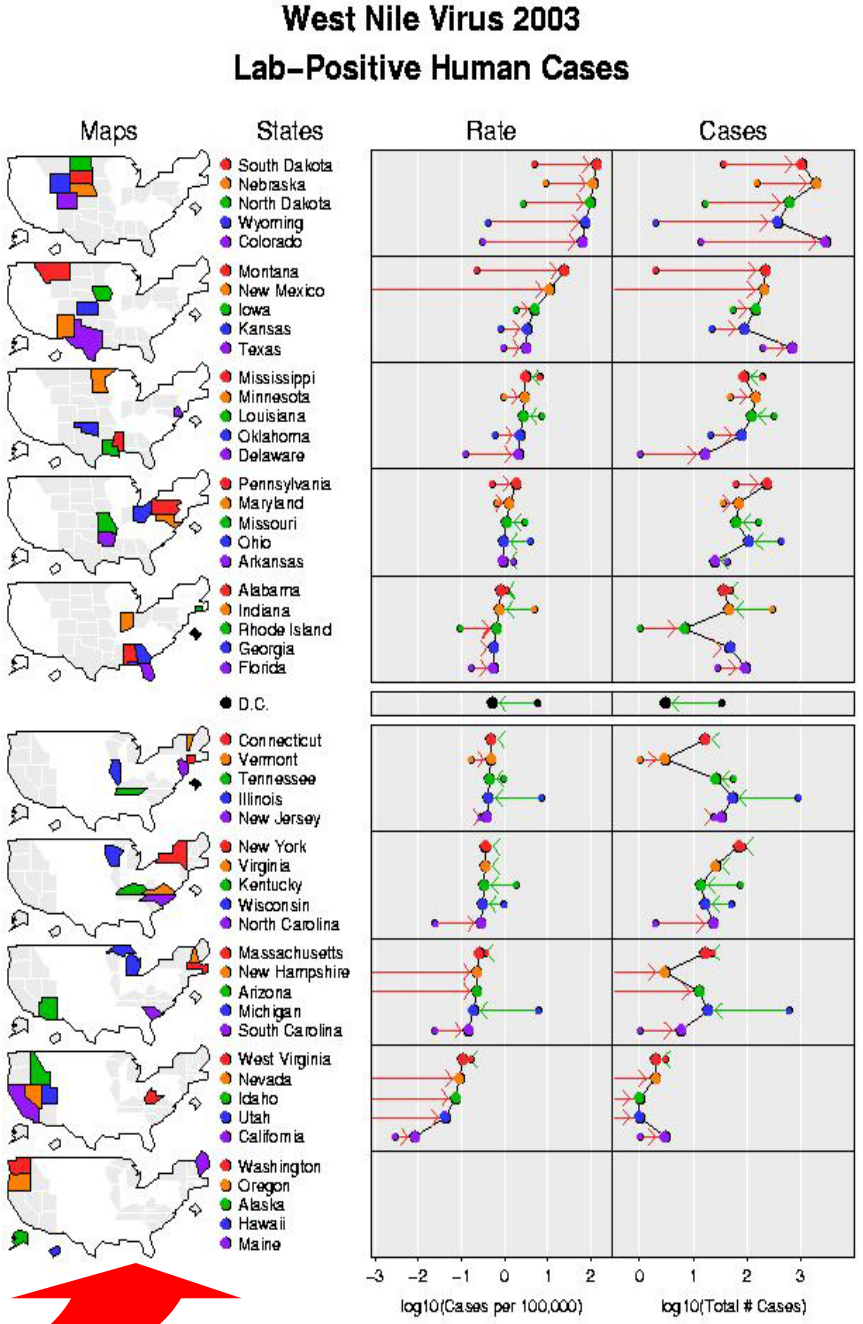




From 2003 CDC



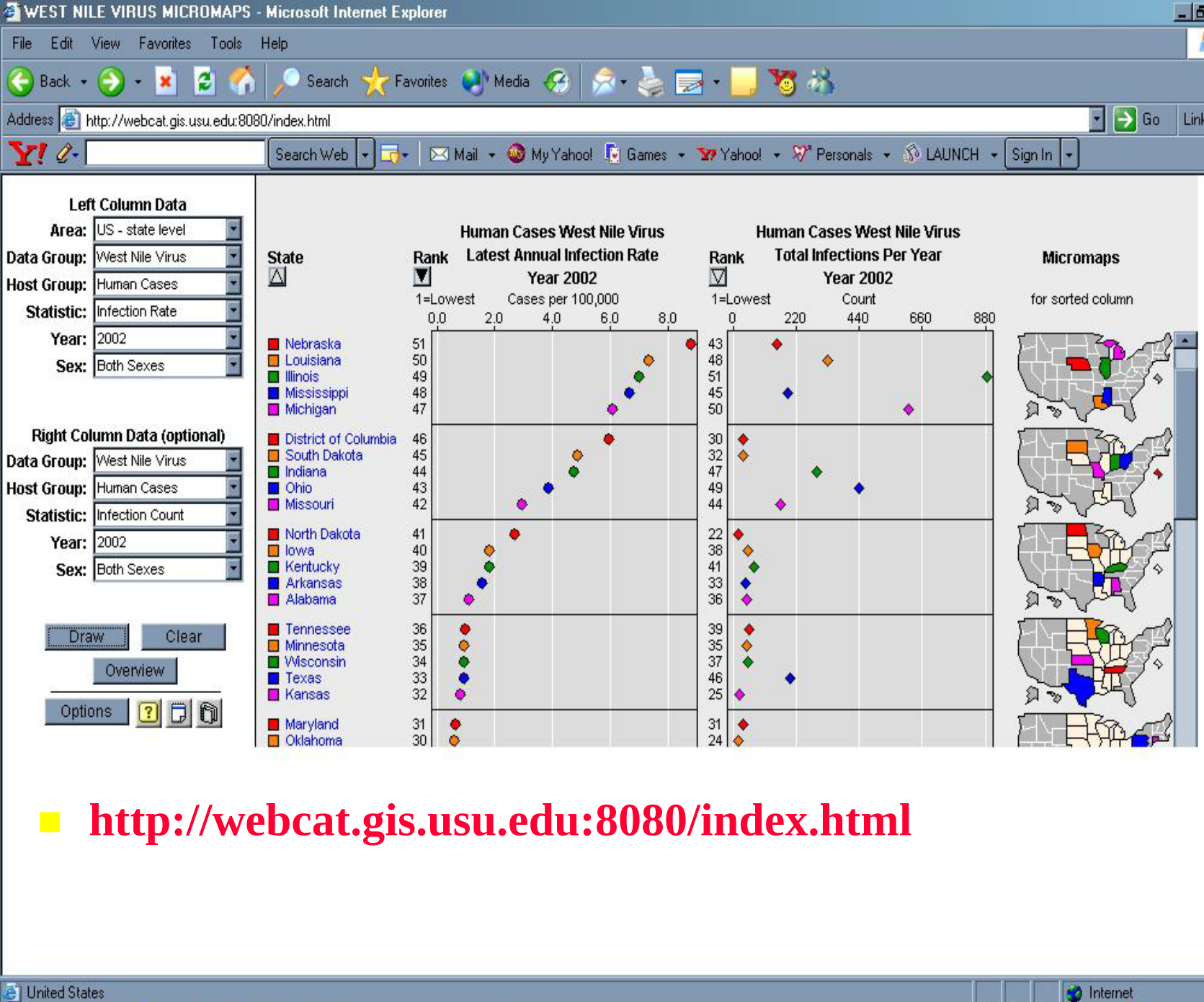




Web-Based Access to WNV Data

- Decision at Utah State University (USU):
  - Obtain NCI Java code for Web-based WNV micromaps
  - Upgrades for the display of WNV data

-  Live Web Demo at **http://webcat.gis.usu.edu:8080/index.html**



- **http://webcat.gis.usu.edu:8080/index.html**

WEST NILE VIRUS MICROMAPS - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://webcat.gis.usu.edu:8080/index.html Go Links

Search Web Mail My Yahoo! Games Yahoo! Personals LAUNCH Sign In

Left Column Data

Area: US - state level
Data Group: West Nile Virus
Host Group: Human Cases
Statistic: Infection Rate
Year: 2003
Sex: Both Sexes

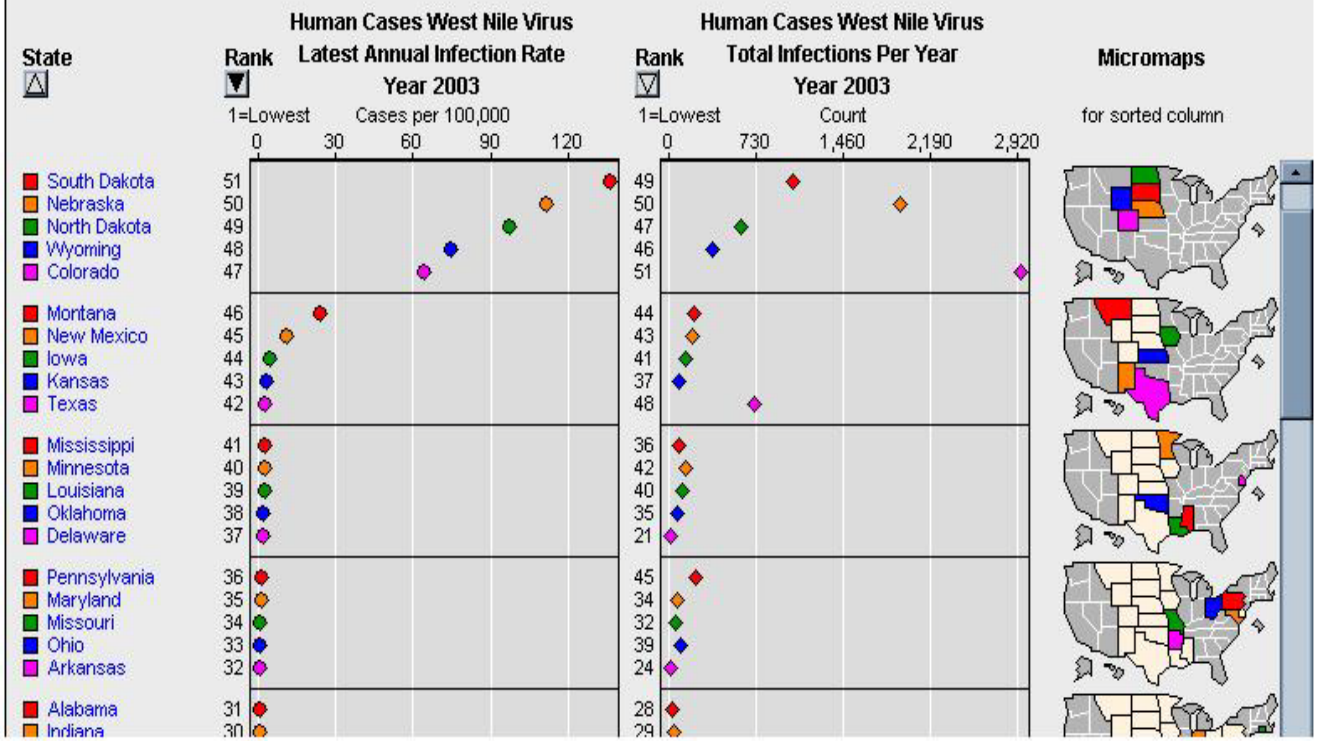
Right Column Data (optional)

Data Group: West Nile Virus
Host Group: Human Cases
Statistic: Infection Count
Year: 2003
Sex: Both Sexes

Draw Clear

Overview

Options



United States Internet

Start WEST NILE VIRUS ... Document1 - Microsoft W... Micromaps2002.jpeg - Paint 11:50 AM

WEST NILE VIRUS MICROMAPS - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://webcat.gis.usu.edu:8080/index.html Go Links

Search Web Mail My Yahoo! Games Yahoo! Personals LAUNCH Sign In

Left Column Data

Area: US - state level
Data Group: West Nile Virus
Host Group: Human Cases
Statistic: Infection Rate
Year: 2003
Sex: Both Sexes

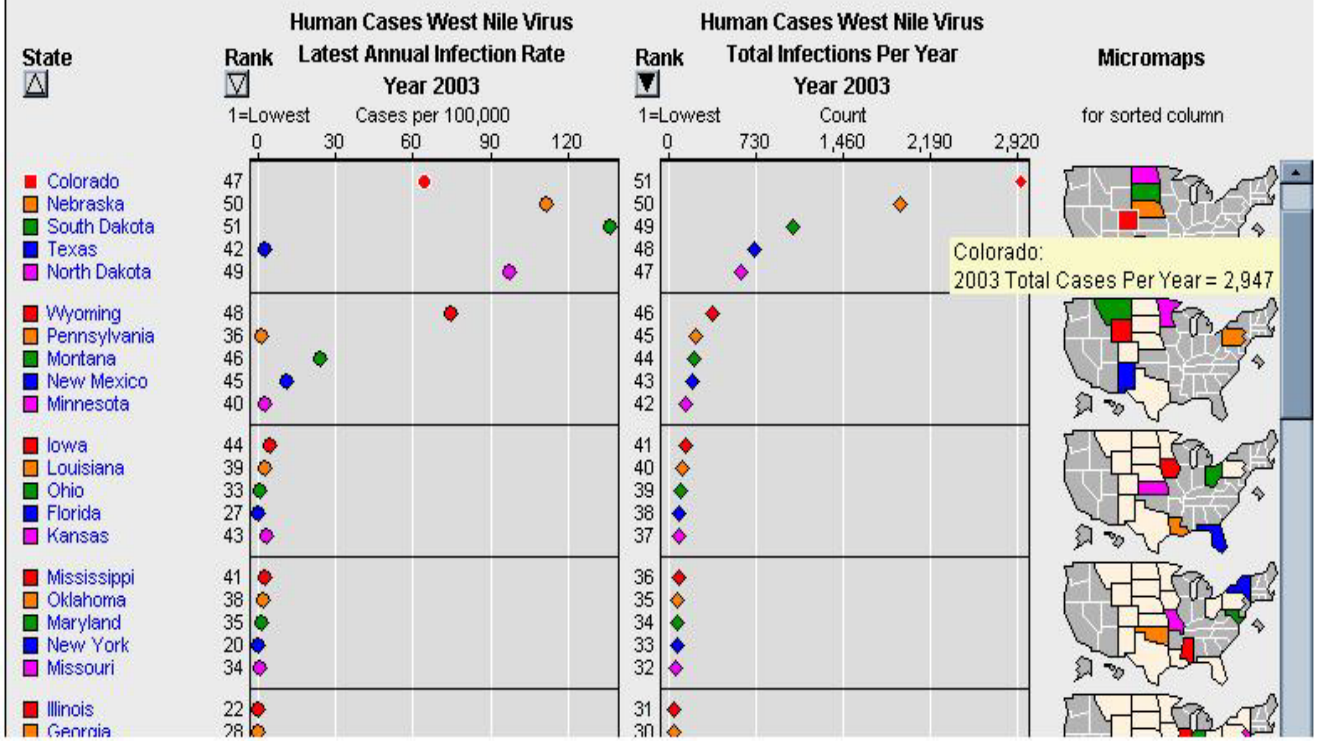
Right Column Data (optional)

Data Group: West Nile Virus
Host Group: Human Cases
Statistic: Infection Count
Year: 2003
Sex: Both Sexes

Draw Clear

Overview

Options



United States Internet

Start WEST NILE VIRUS ... Document1 - Microsoft W... Micromaps2003.jpeg - Paint 11:51 AM

- **Example:**
  Agreement in Carpal Tunnel Syndrome Surveys

  Work in Progress with:
  Bradley Evanoff, Ann Marie Dale, Jaime Strickland, Bethany Taylor, and Bill Shannon.

Survey Questions (1)

- 125 subjects (S) are asked 8 question of the form:

  "On average, how much time do you spend each day lifting, carrying, pushing or pulling objects weighing more than 2 pounds?" [lift]
- Possible answers are:
  - Not at all (1)
  - Less than 5 minutes (2)
  - Between 5 minutes and ½ hour (3)
  - More than ½ hour but less than 1 hour (4)
  - 1-2 hours (5)
  - 2-4 hours (6)
  - More than 4 hours (7)

Survey Questions (2)

- Questions are related to:
  - Lifting [lift]
  - Vibrating tools [vibr]
  - Assembly line [asem]
  - Rotating [rota]
  - Bending [bend]
  - Gripping [grip]
  - Finger or thumb as a pushing tool [thmb]
  - Finger pinching [pnch]
- Of interest is agreement of assessments given by subjects (S) and experts (E) who watch videos of the subjects at work

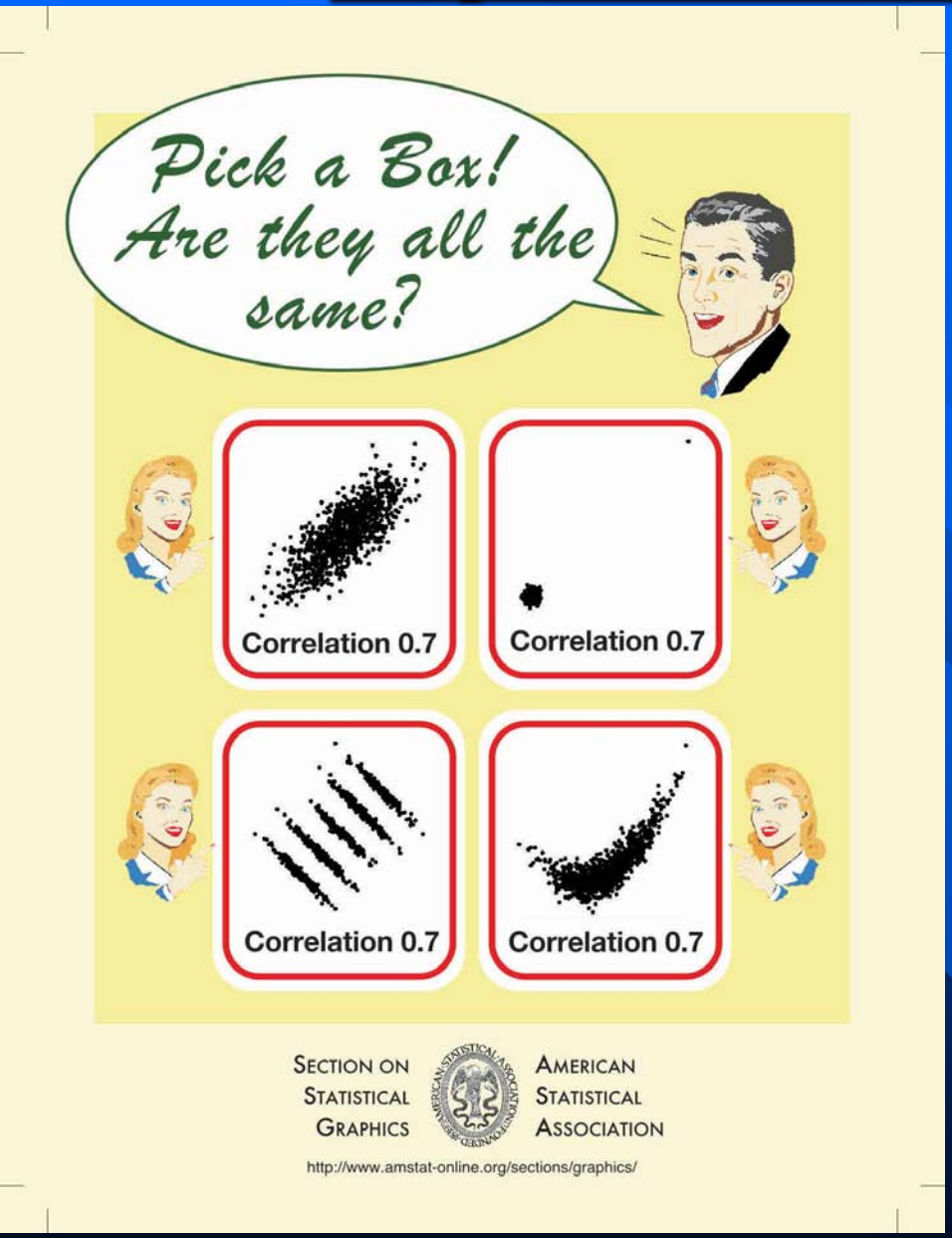
Survey Outcome (1)

- "kappa" represents agreement between subjects and experts, based on difference between how much agreement is present compared to how much agreement would be expected by chance alone (possible range is -1.0 to 1.0)
- Interpretation:
  - < 0 : Less than chance agreement
  - 0.01 – 0.20: Slight agreement
  - 0.21 – 0.40: Fair agreement
  - 0.41 – 0.60: Moderate agreement
  - 0.61 – 0.80: Substantial agreement
  - 0.81 – 0.99: Almost perfect agreement

Survey Outcome (2)

- Reported “kappa” values are:
  - 0.53 [lift], 0.49 [grip], 0.44 [vibr] -> “moderate”
  - 0.24 [pnch], 0.18 [bend] -> “slight/fair”
  - 0.02 [rota], -0.04 [asem], -0.09 [thmb] -> “less than chance/slight”

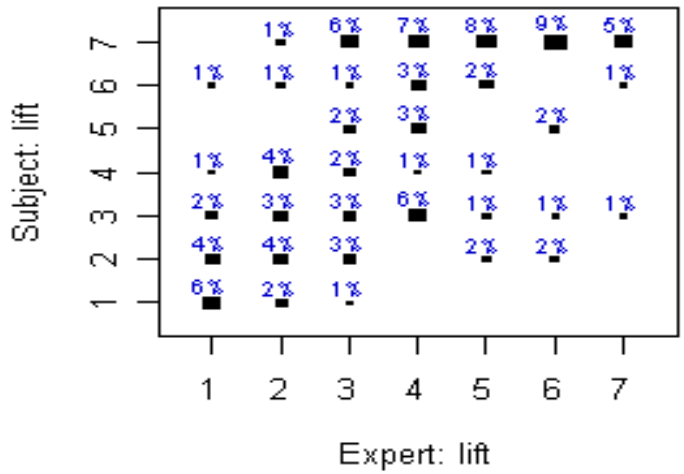
Inspiration of Agreement Plots

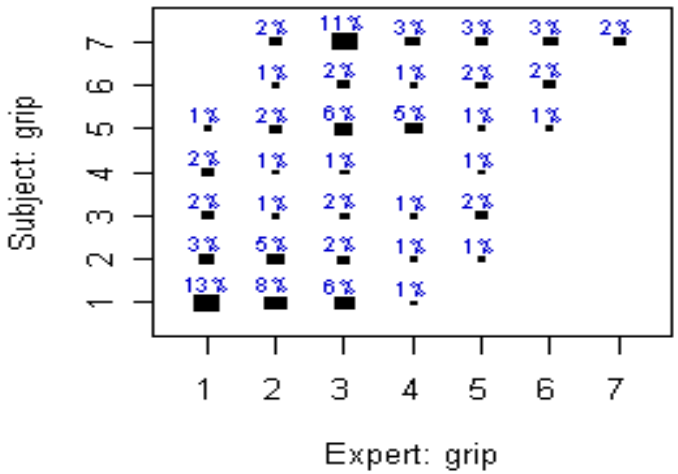


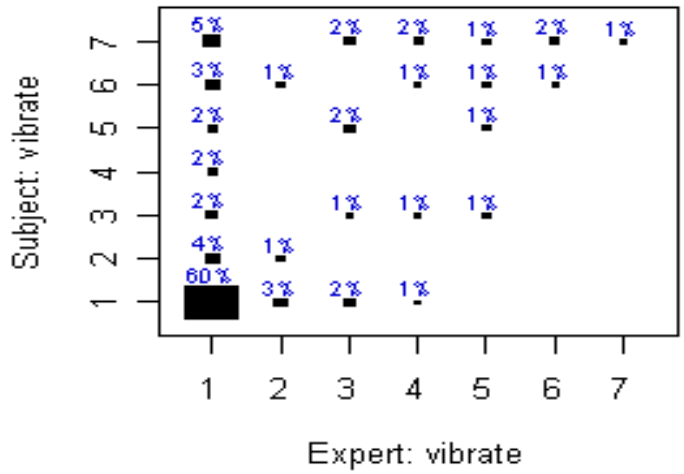
American Statistical Association Statistical Graphics Section Poster Series (~2004)

http://www.public.iastate.edu/~dicook/Stat.Graphics/posters.html

Agreement Plots (1)

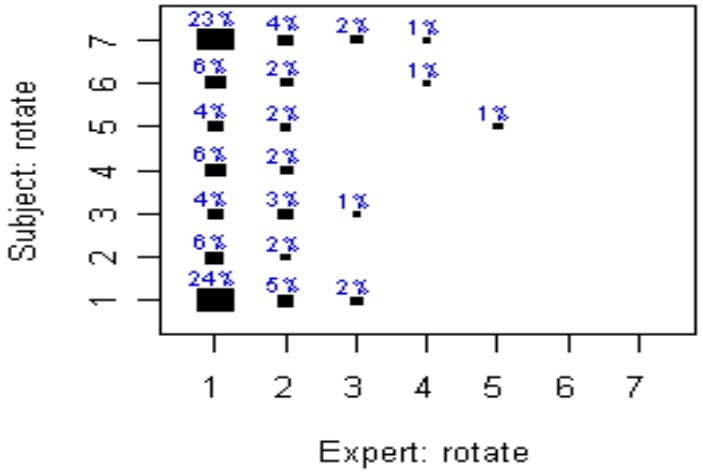


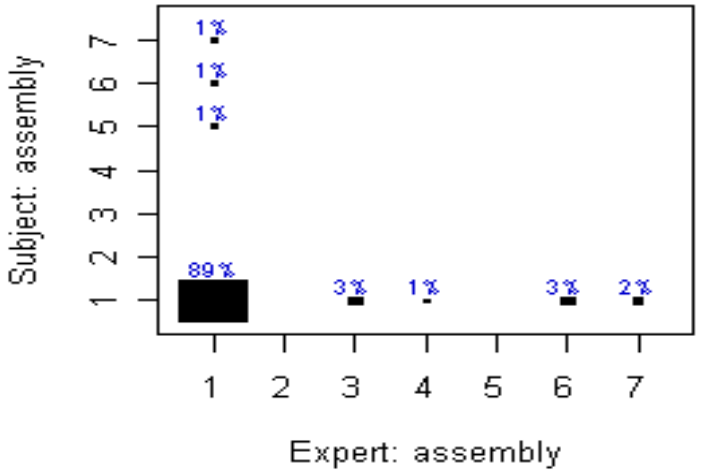


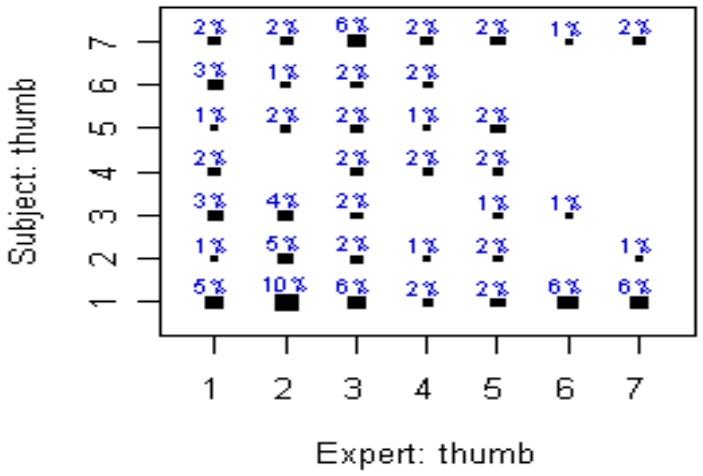


Recall kappa values:
0.53 [lift], 0.49 [grip], 0.44 [vibr]

Agreement Plots (2)

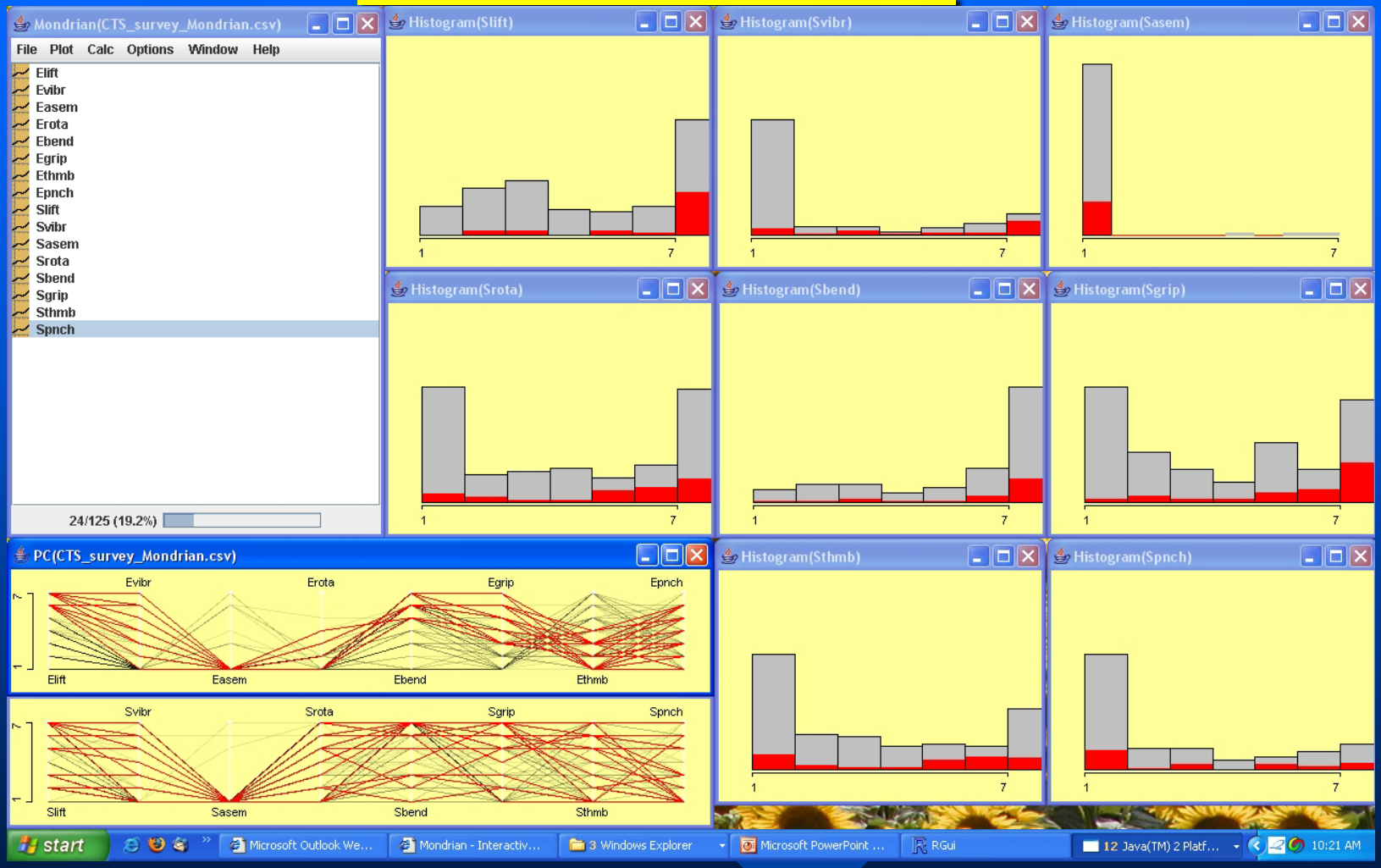






Recall kappa values:
0.02 [rota], -0.04 [asem], -0.09 [thmb]

Linked Histograms & Parallel Coordinate Plots (Mondrian)



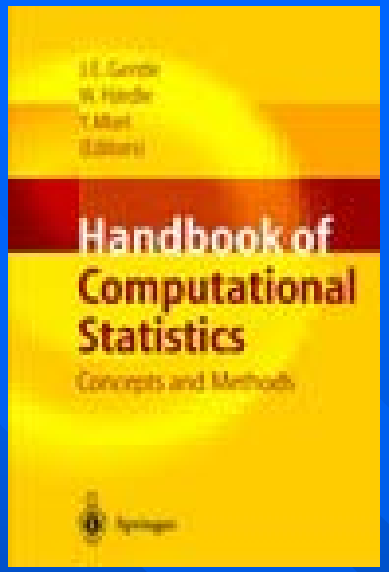
- Mondrian Live Demo

Conclusions

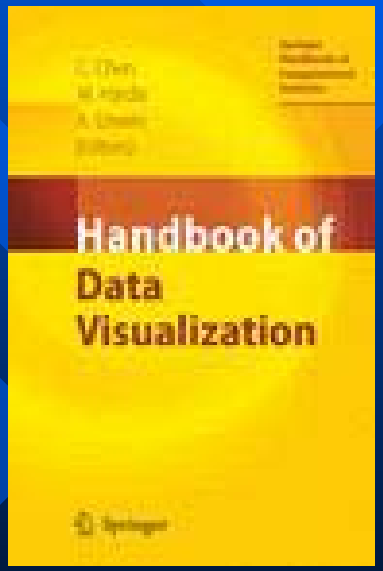
- Visual approach effective to see unexpected structure in data
- Combination of different techniques most effective
- Can be used for almost all types of data

Additional Reading:

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., Carr, D. B. (2007): Interactive Linked Micromap Plots for the Display of Geographically Referenced Statistical Data, In: Chen, C., Härdle, W., Unwin, A. (Eds.), Handbook of Computational Statistics (Volume III) - Data Visualization, Springer, Berlin/Heidelberg, Forthcoming.



Questions ???