Visual Data Mining - Techniques and Examples

Jürgen Symanzik

Utah State University, Logan, UT

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

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

Contents

- Visual Data Mining
  - Definitions
  - Software
  - Techniques
- Examples
  - Archaeological Data
  - Human Motion Data
  - Neuroanatomical Data

Data Mining

Ed Wegman:

> “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”

Visual Data Mining (1)

- Working Definition:
  - 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

Visual Data Mining (2)

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

Software: XGobi/ggobi

Swayne, Cook and Buja

- Interactive environment for exploring multivariate data
  - Linked views allow “linked brushing”
  - Univariate, Bivariate and Multivariate views of the data
  - Grand tour
  - Wide variety of methods
  - Open source
  - Free

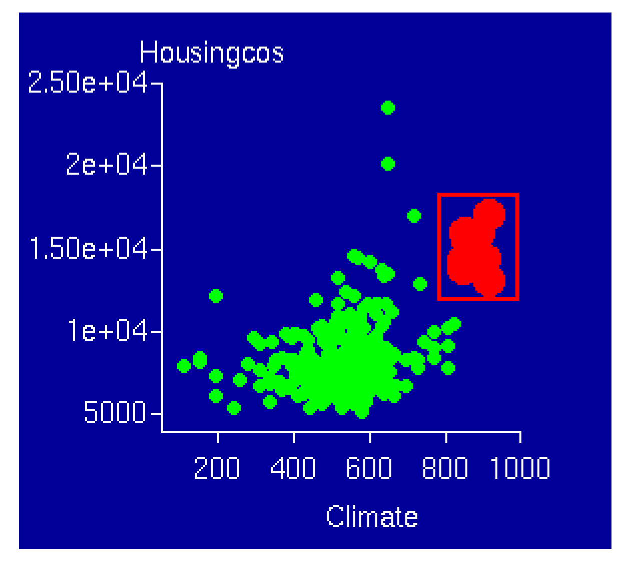
- Caveats
  - XGobi only on UNIX and Linux platforms
  - ggobi also available for PCs but not yet fully developed

Software: ExplorN

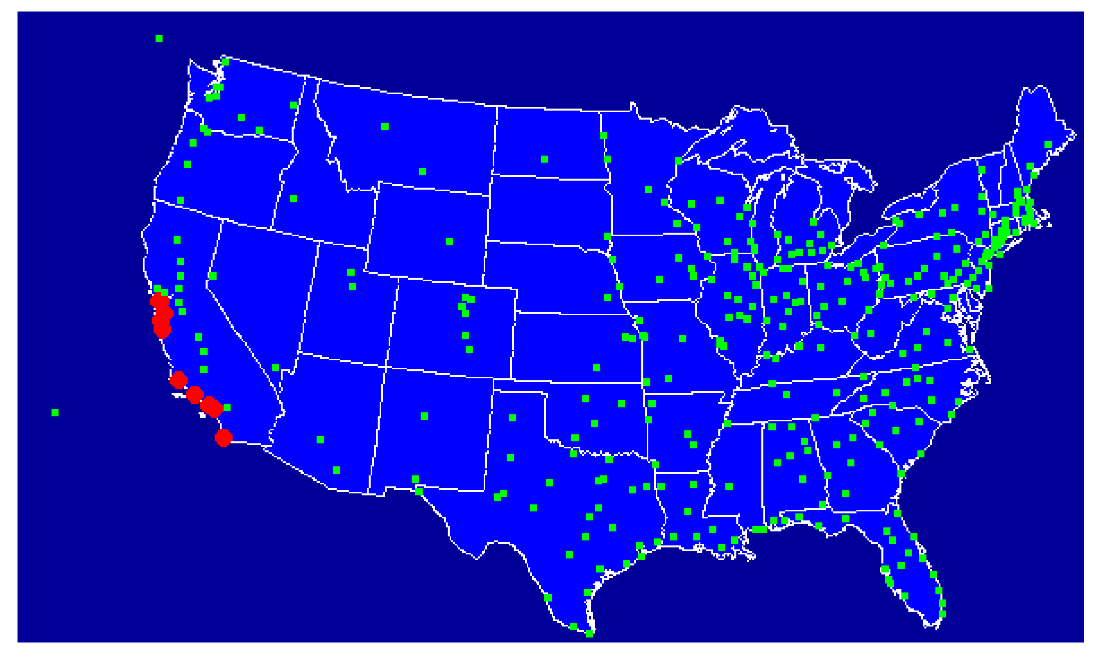
Carr, Wegman, Luo

- Interactive environment for exploring multivariate data (similar to XGobi)
  - *Advanced Parallel Coordinates Displays
  - *3D Surfaces
  - *Stereoscopic Displays
- Caveats
  - *Only on SGI platforms
  - *No interface

Tools: Linked Brushing

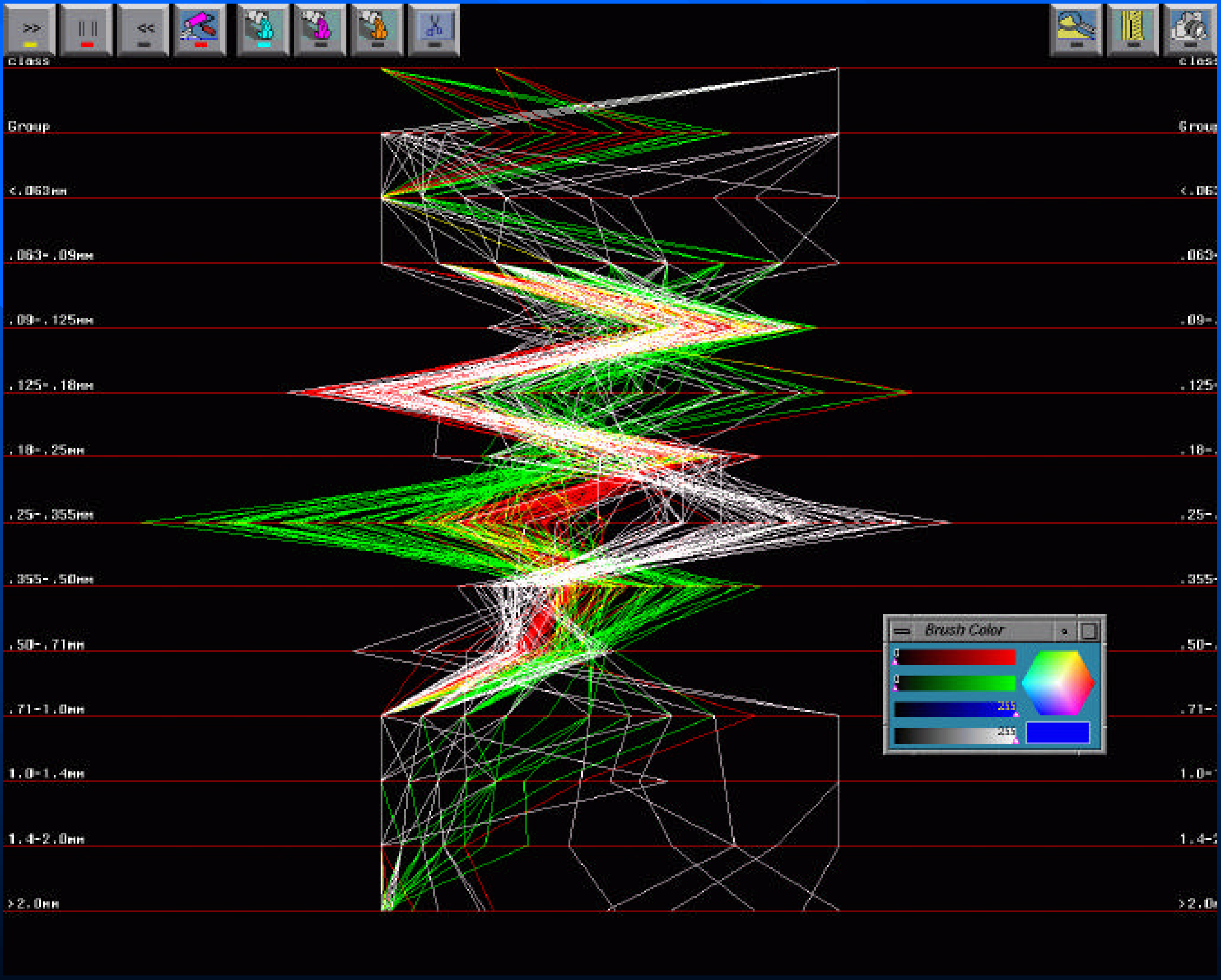


XGobi



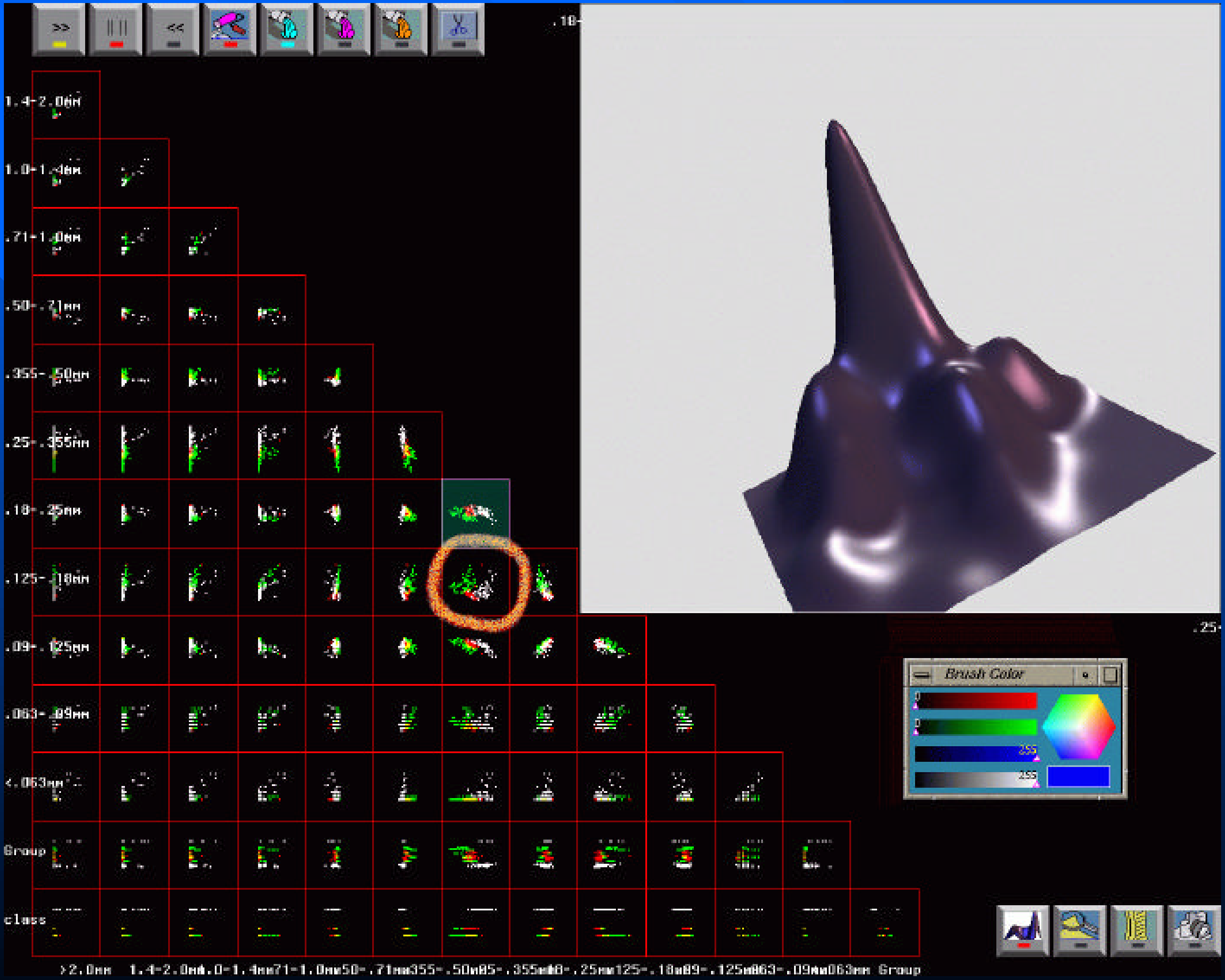
Tools: Parallel Coordinate Plots

ExplorN



Tools: Scatterplot Matrix

ExplorN



Tools: Grand Tour

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

Examples

- – Archaeological Data
- – Human Motion Data
- – Neuroanatomical Data

Example: Archaeological Data

Published as:

Wilhelm, A. F. X., Wegman, E. J., Symanzik, J. (1999): Visual Clustering and Classification: The Oronsay Particle Size Data Set Revisited, Computational Statistics: Special Issue on Interactive Graphical Data Analysis, Vol. 14, No. 1, 109-146.

Oronsay Sand Particles

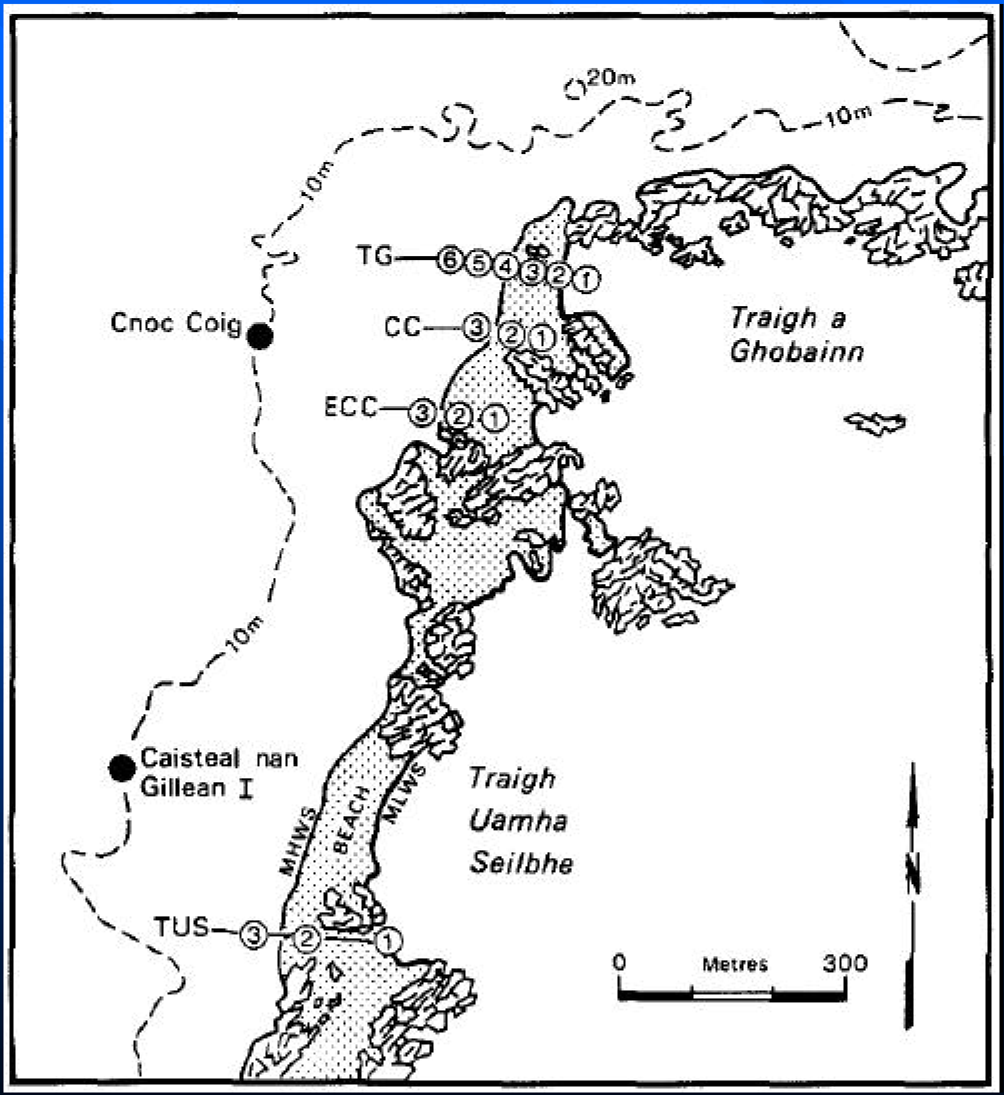
“The mesolithic shell middens on the island of Oronsay are one of the most important archeological sites in Britain. It is of considerable interest to determine their position with respect to the mesolithic coastline. If the sand below the midden were beach sand and the sand from the upper layers dune sand, this would indicate a seaward shift of the beach-dune interface.”

Flenley and Olbricht, 1993

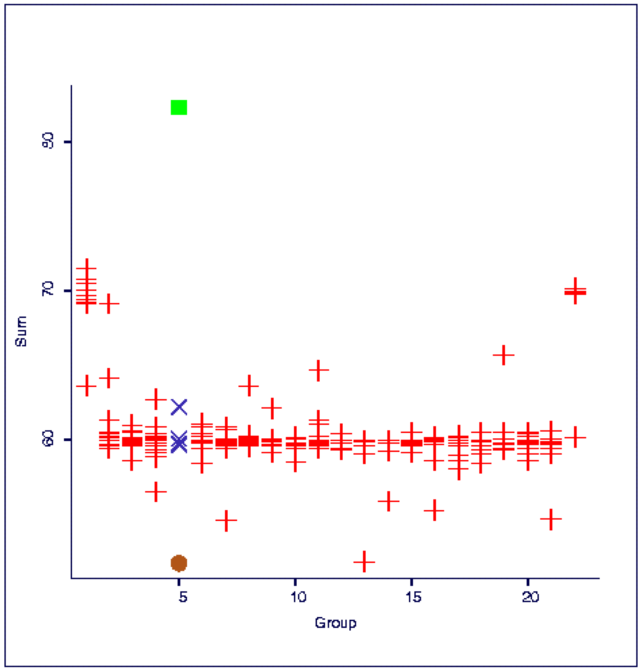
Objective of Study

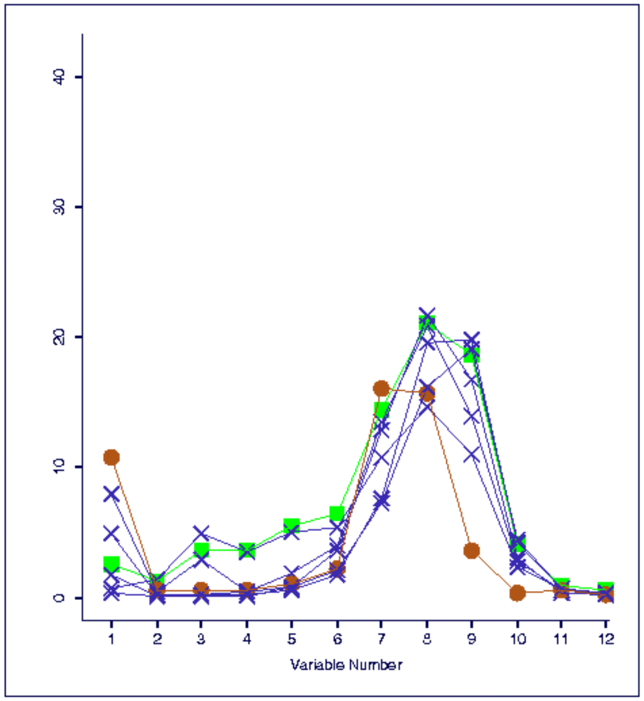
- Cluster samples of modern sand into "beach-like" or "dune-like" sand.
- Classify archeological sand samples as to whether they are beach sand or dune sand.

Oronsay - Geography



Oronsay - Data Problems





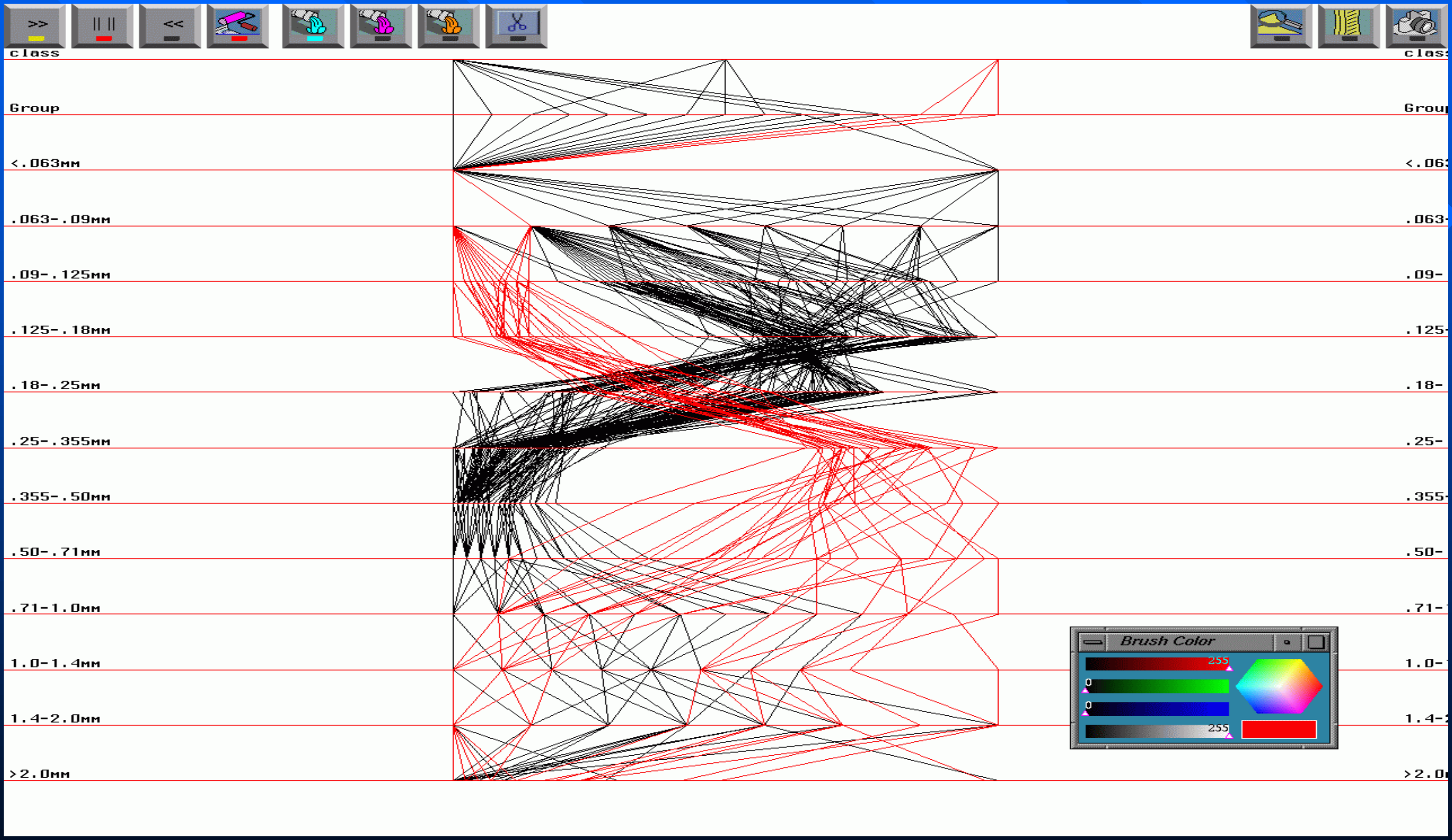
Oronsay - Parametric Analysis

- Historical strategy is to fit parametric distributions and compare modern and archeological sands based on parameters.
- Weibull, 1933; lognormal (breakage models), log-hyperbolic, log-skew-Laplace, 1937, Barndorff-Nielsen, 1977.
- Models 2 to 4 parameters, theory developed, practice problematic.

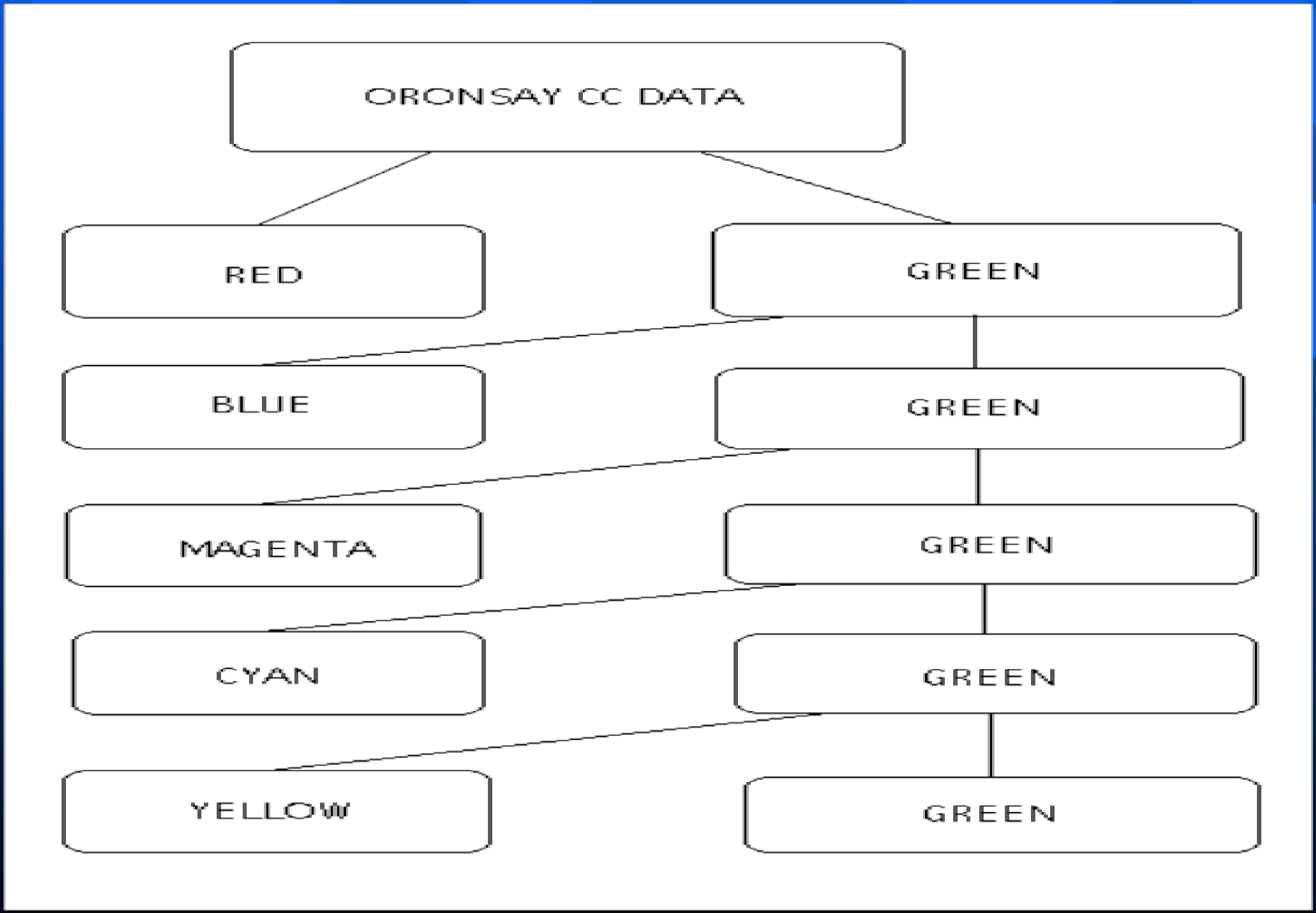
Oronsay - Visual Approach

- Multidimensional Parallel Coordinate Display Combined with Grand Tour.
- BRUSH-TOUR strategy
  - Clusters recognized by gaps in any horizontal axis.
  - Brush existing clusters with colors.
  - Execute grand tour until new clusters appear, brush again.
  - Continue until clusters are exhausted.

Beach & Dune Sand



Separation of Clusters



ORONSAY CC DATA

RED

GREEN

BLUE

GREEN

MAGENTA

GREEN

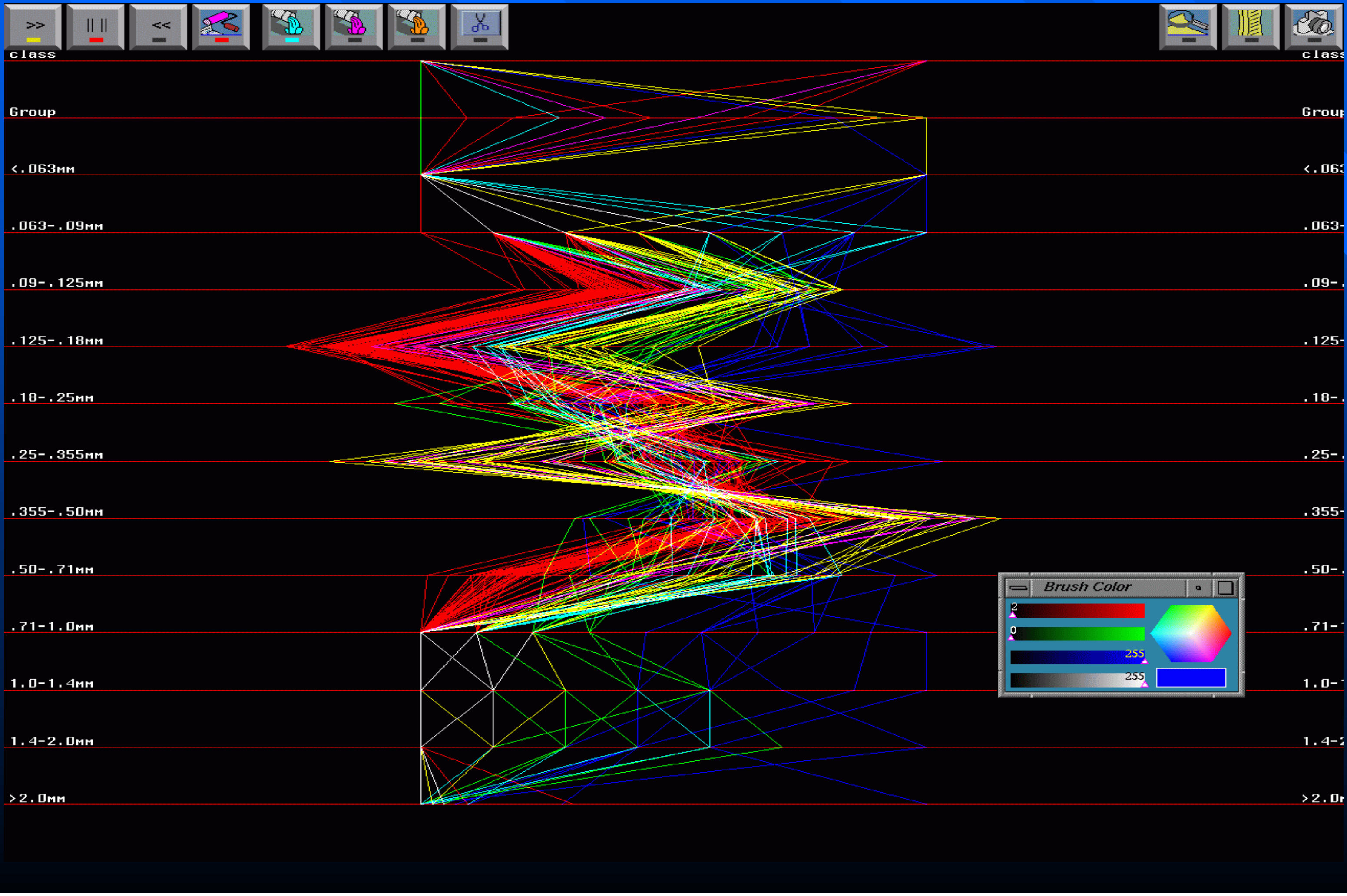
CYAN

GREEN

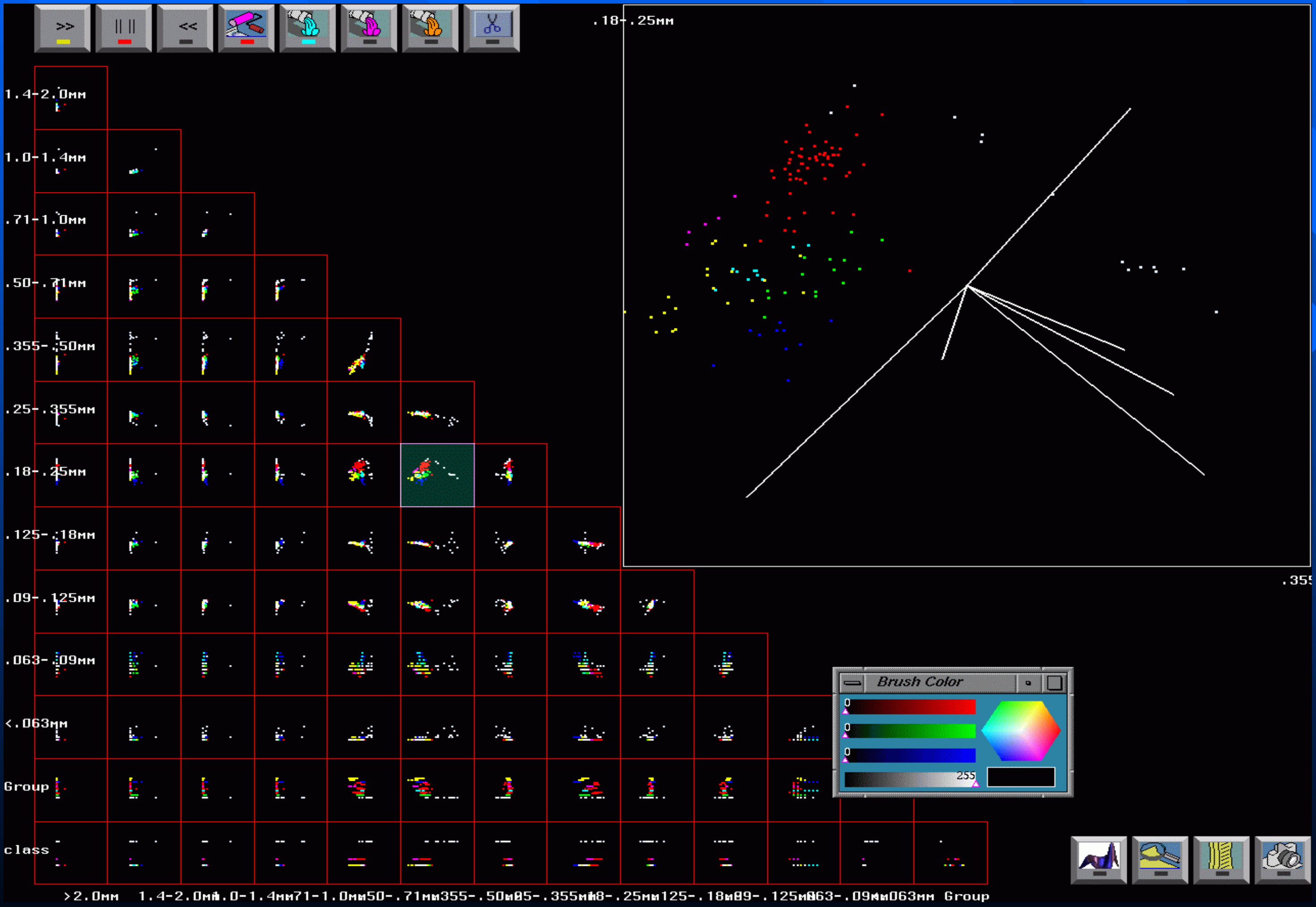
YELLOW

GREEN

Final Clustering



Scatterplots & Projection



Oronsay - Conclusions (1)

- Sands from the CC site and the CNG site have considerably different particle size distributions and cannot be effectively aggregated.
- Data at small and at large particle dimensions is too quantized to be used effectively.
- The visual based BRUSH-TOUR strategy is extremely effective at clustering.

Oronsay Conclusions (2)

- Midden sands are neither modern beach sands nor modern dune sands.
- Midden sands are more similar to modern dune sands.
- This result does not support the seaward-shift-of-the-beach-dune-interface hypothesis, but suggests the middens were always in the dunes

Example: Human Motion Data

Published as:

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

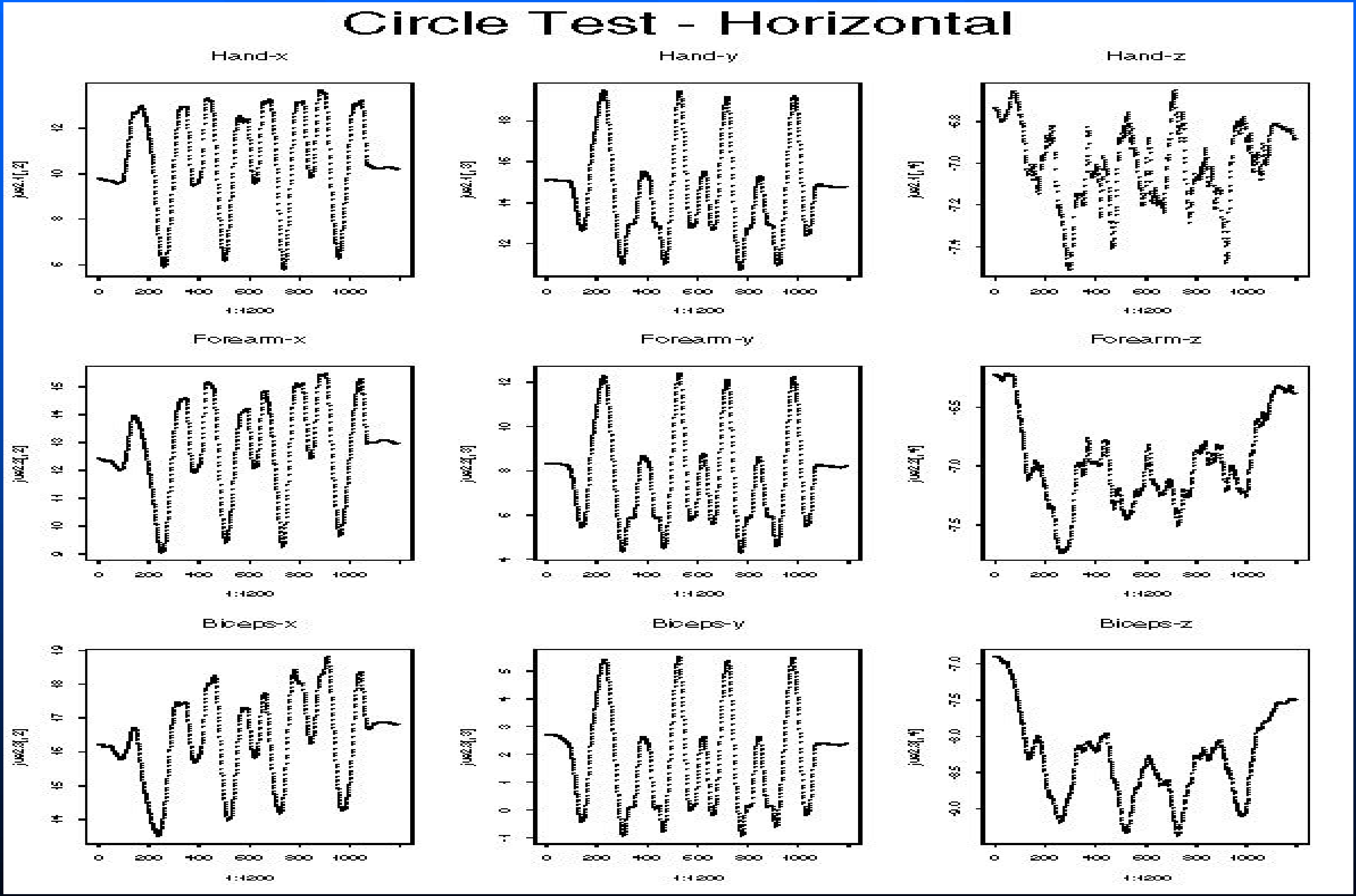
Aim of Preliminary Experiments

- Get used to Sensors & other Hardware.
- How does Visualization help to understand the data?
- Need: Visualization during Experiments
  - Complicated setup - impossible to redo once finished
  - Data plausible?
  - Data correctly recorded?

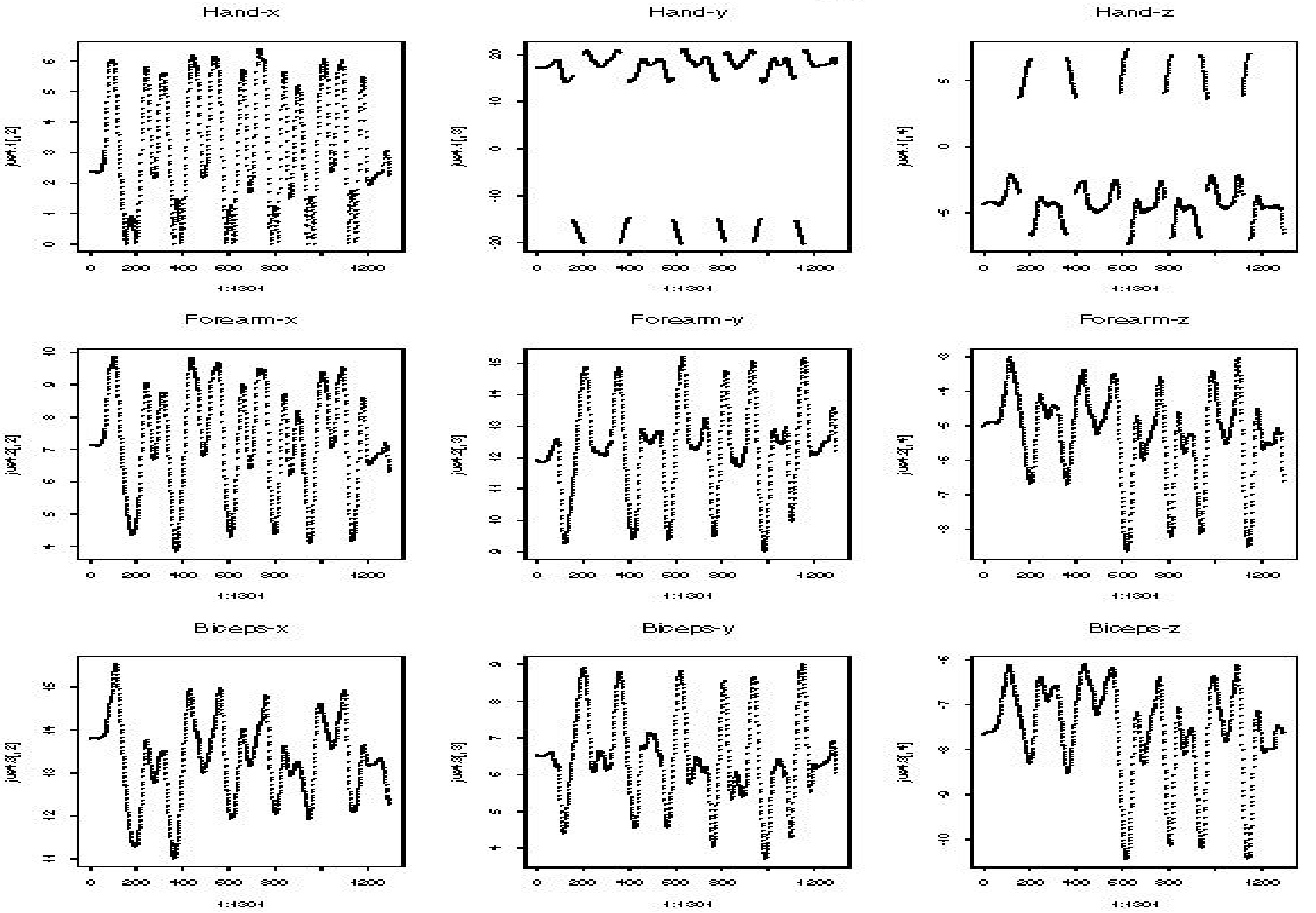
Data Collected

- Small to Medium Size Data Set:
  - – 60 to 100 Hz
  - – 30 to 120 sec
  - – 6 x 3 FOBs sensors
  - – Here: 25,000 to 40,000 Measurements

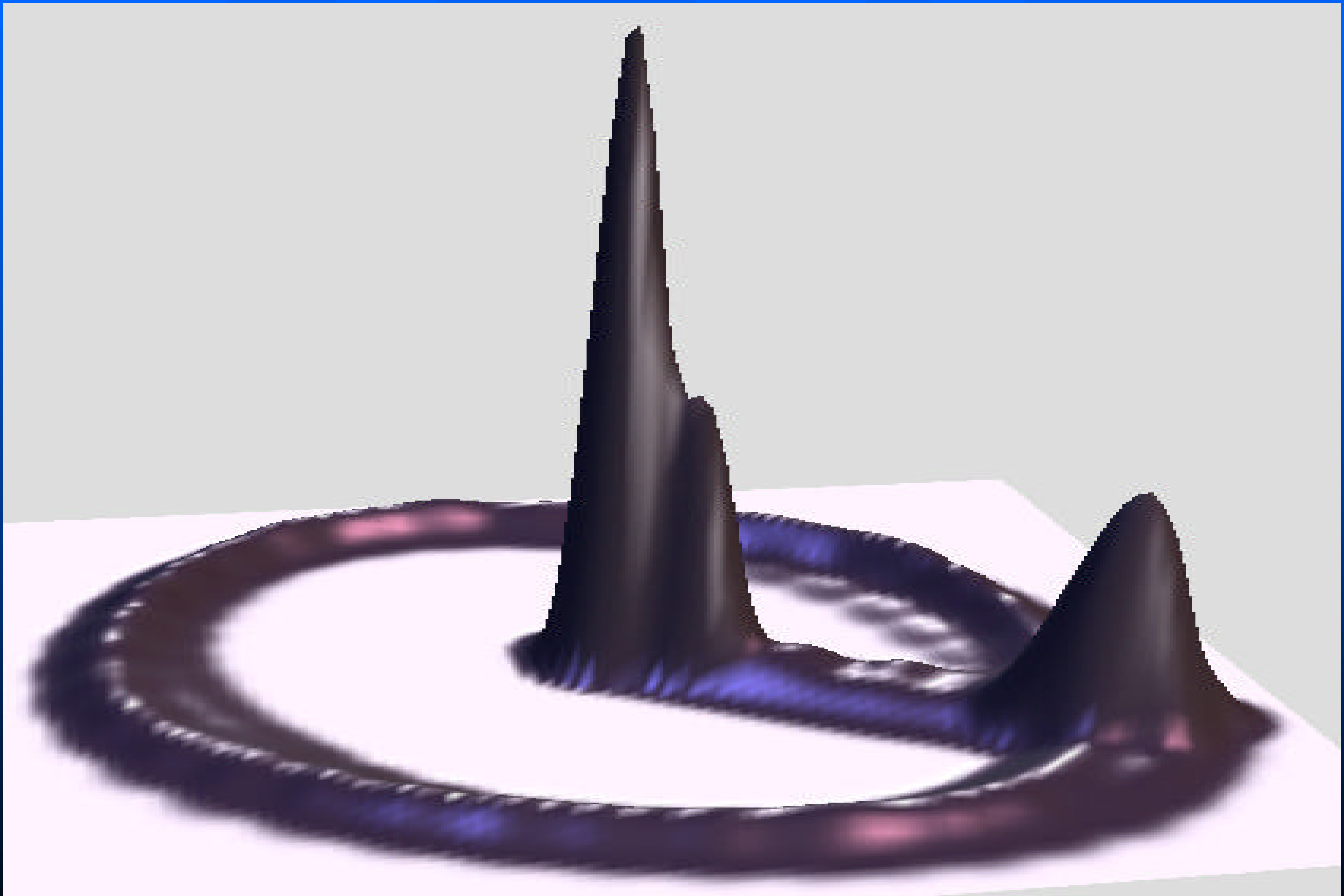
Timeseries Plots (S-Plus)



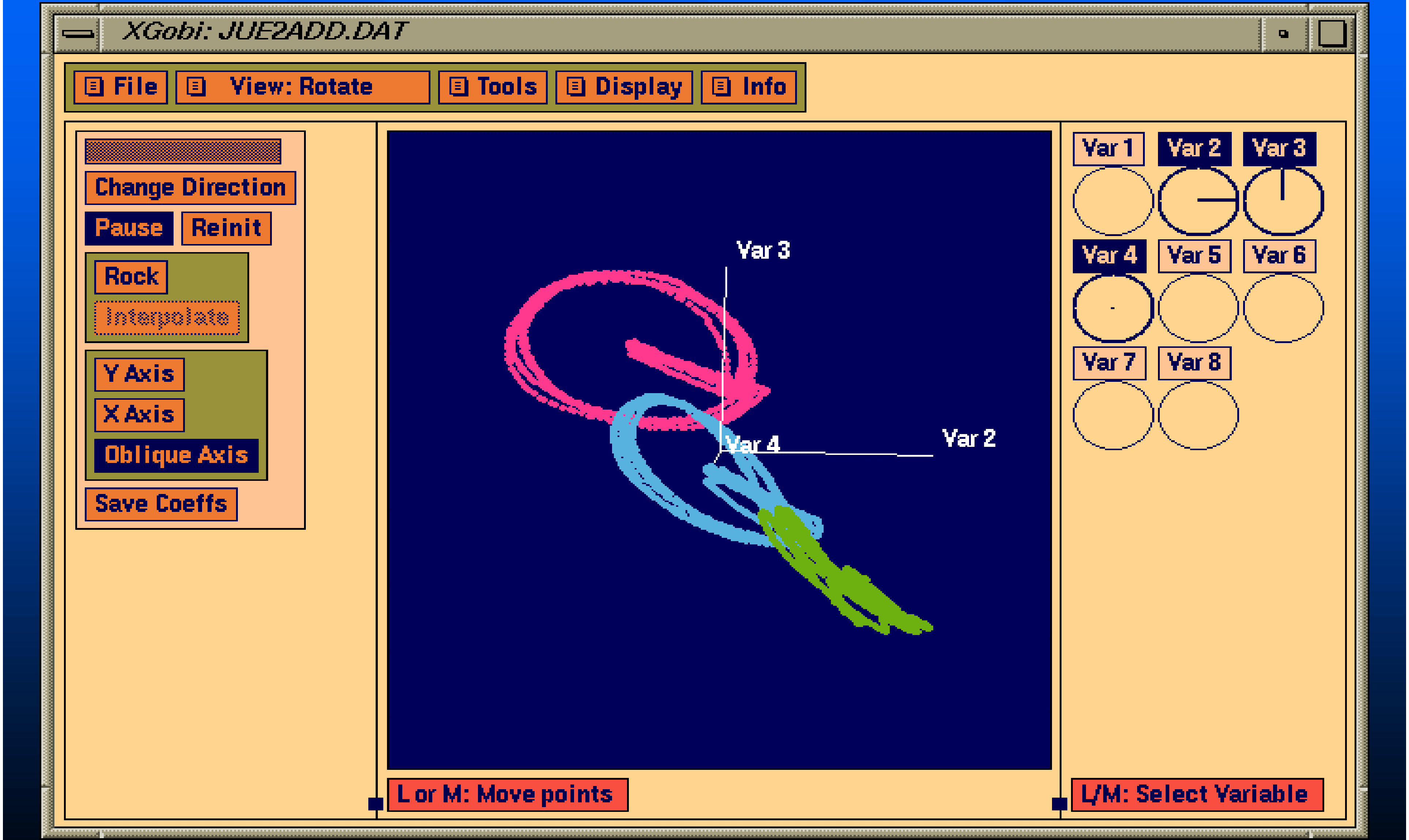
Circle Test - Angular

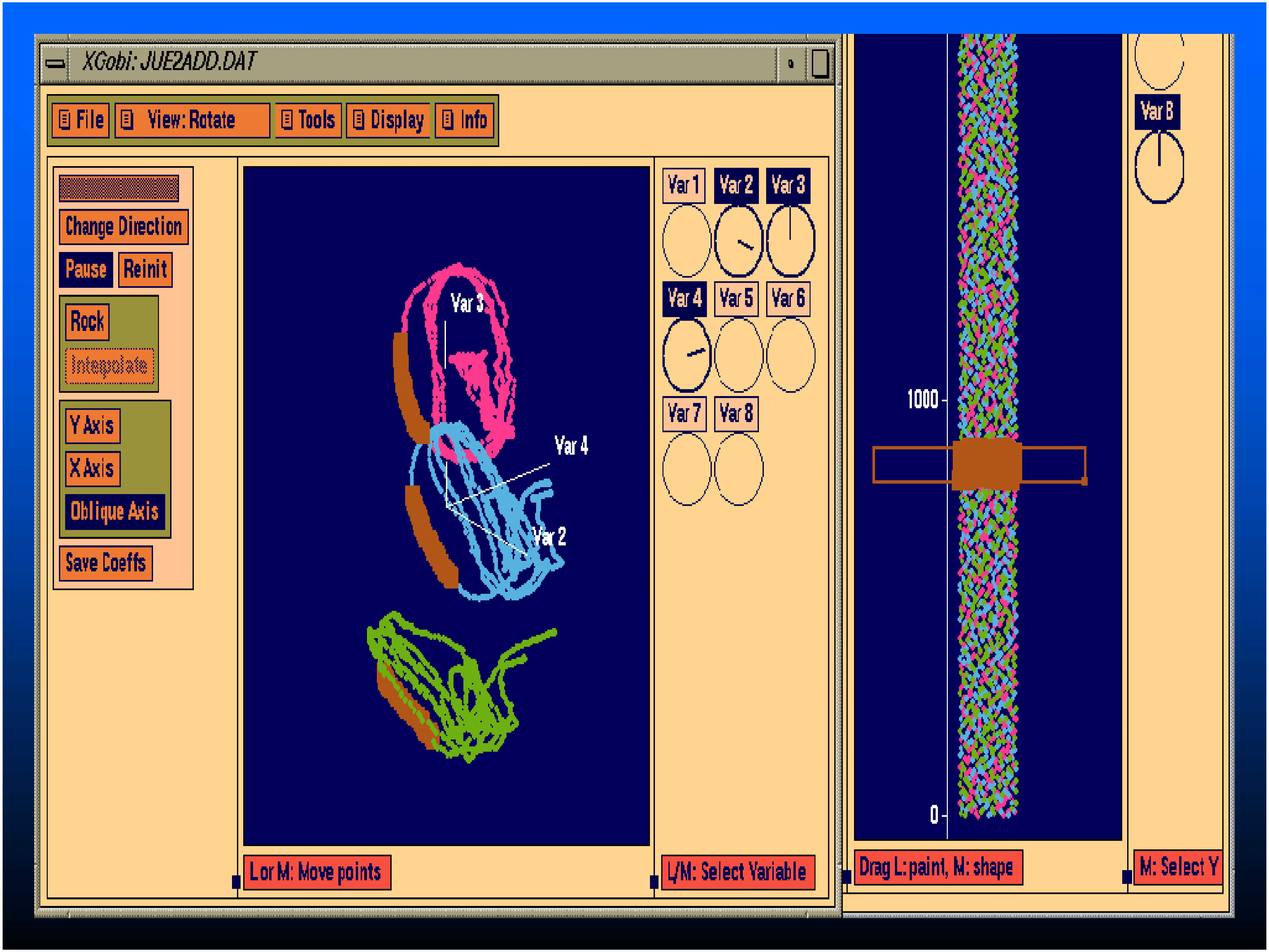


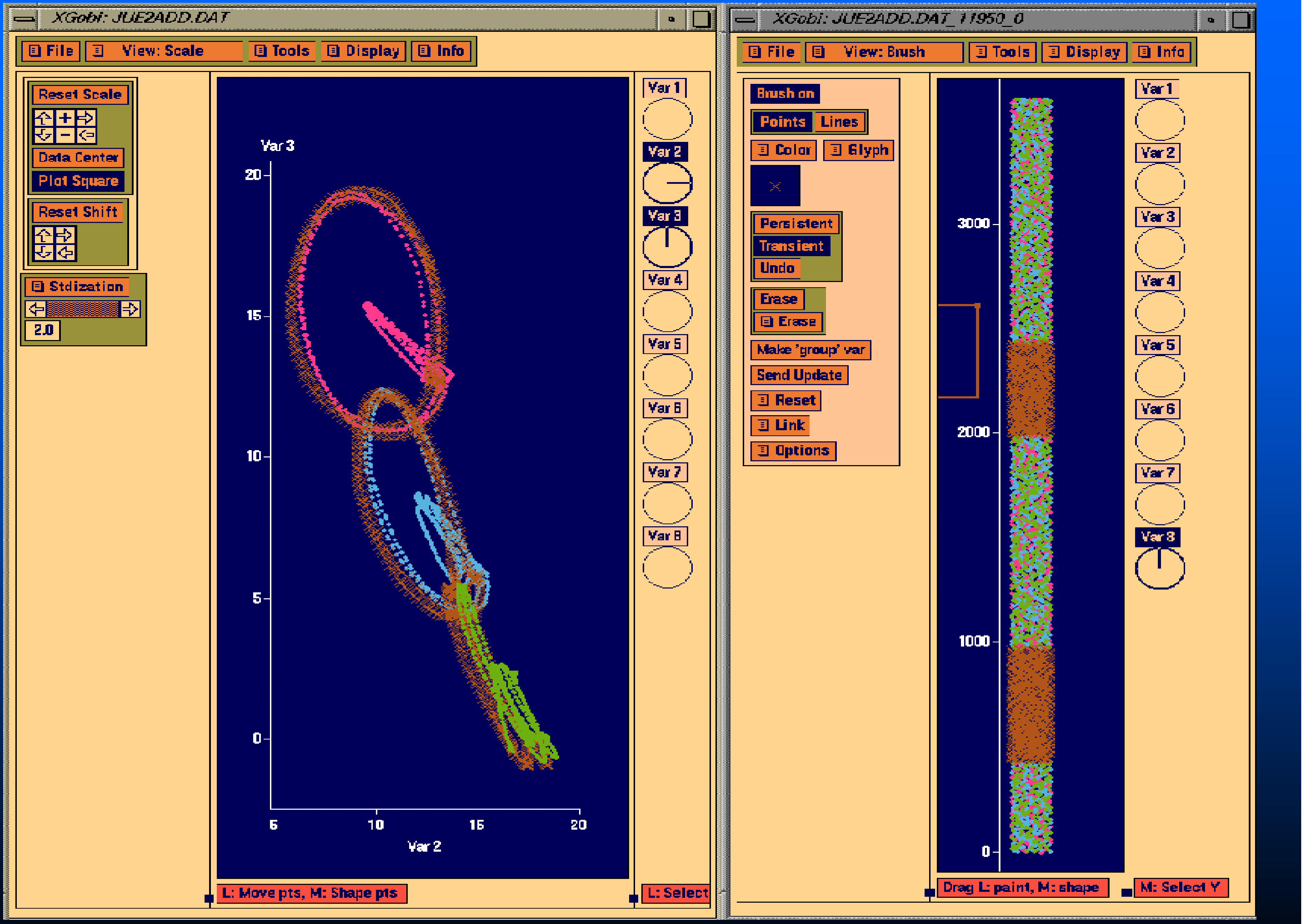
Density Plots (ExplorN)

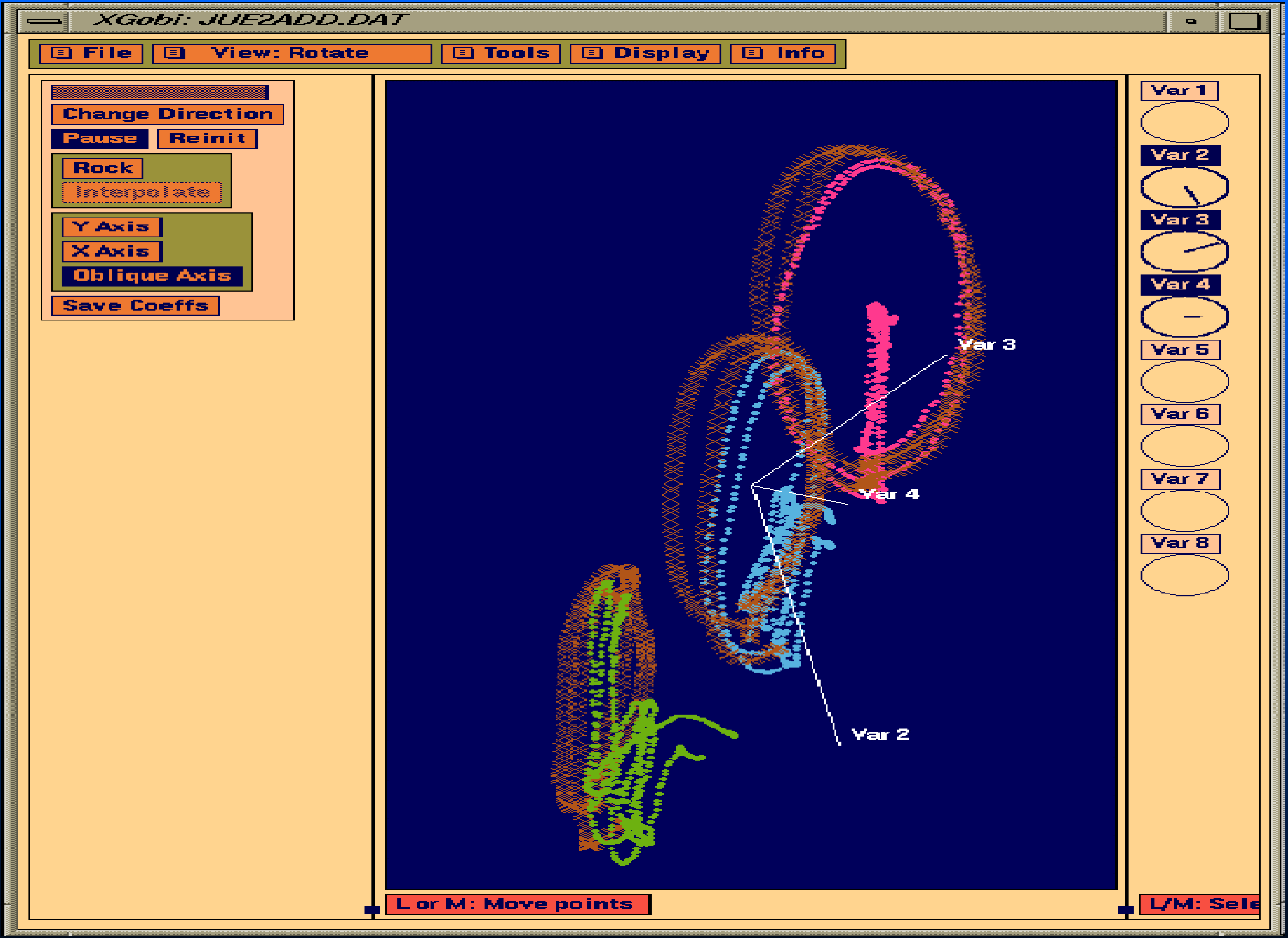


Scatterplots and Rotation (XGobi)




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 3
Var 4
Var 2
Var 1
Var 2
Var 3
Var 4
Var 5
Var 6
Var 7
Var 8
L or M: Move points
L/M: Sele

Motion - Conclusions

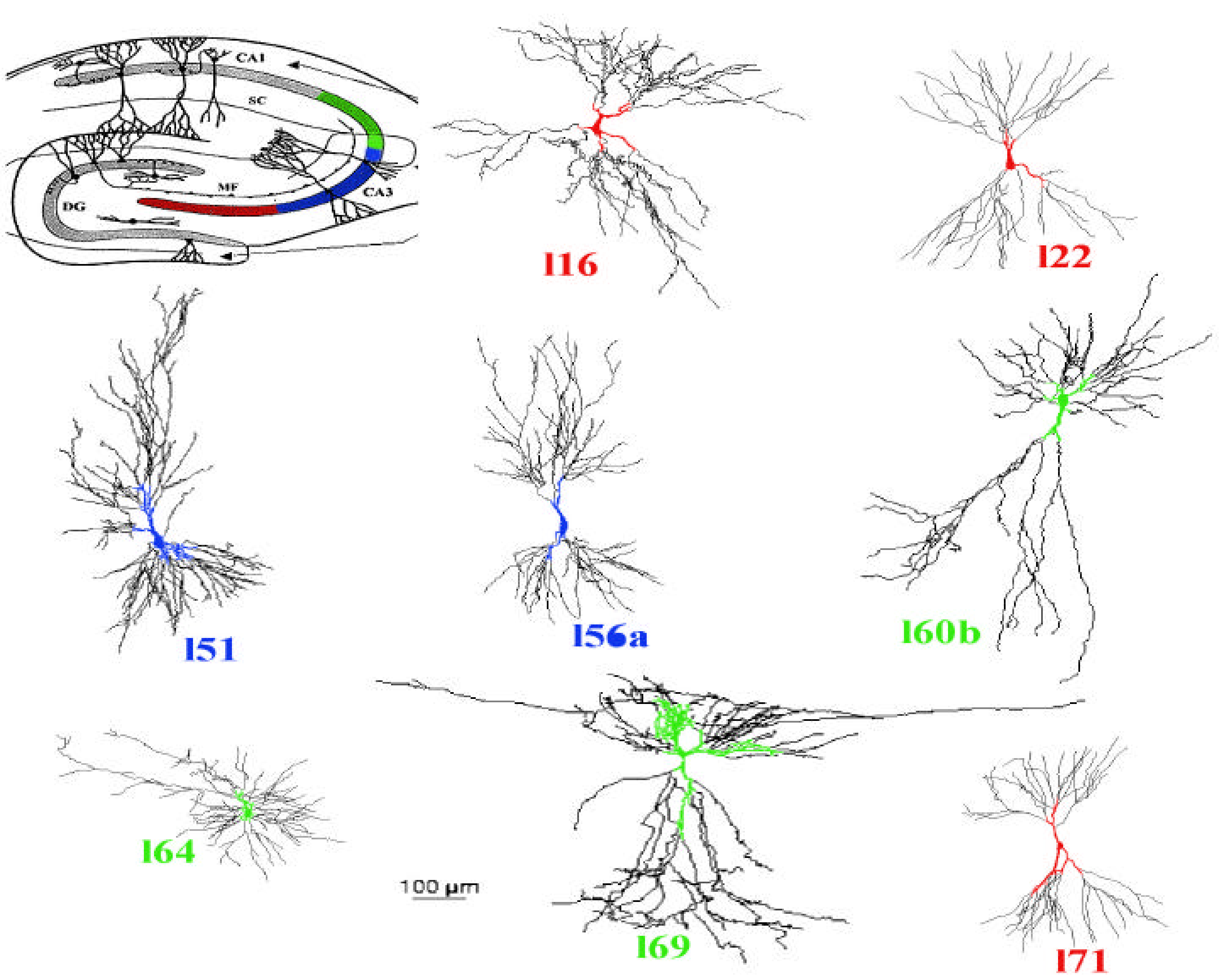
- Visualization helps to immediately check the correctness of the data.
- Realistic 3D Visualization helps to detect unexpected behavior.

Example: Neuroanatomical Data

Published as:

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.

Pyramidal Brain Cells



Morphological Parameters

- Apical Dendrite
- Basal Dendrite

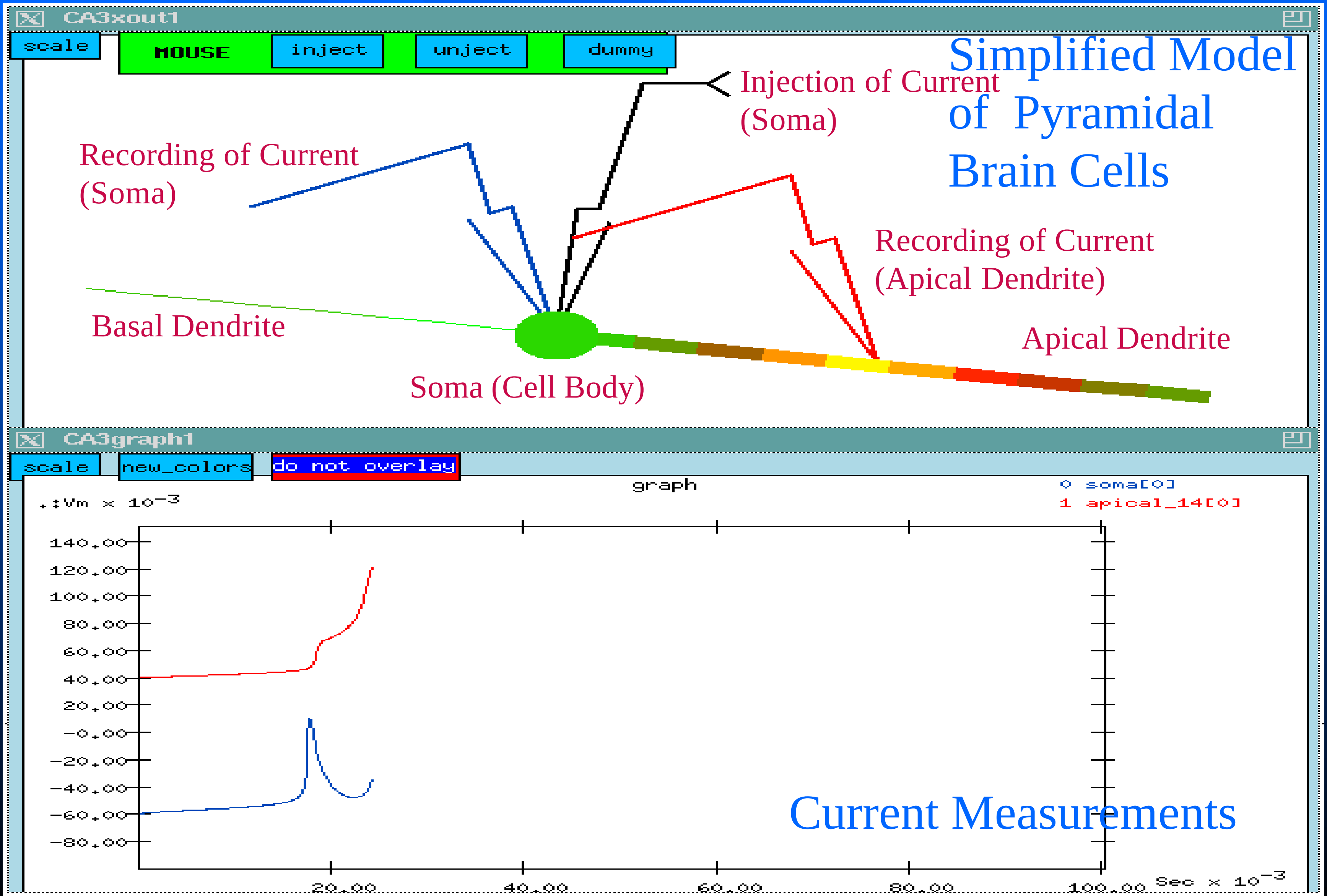
- Distance from Soma
  - 50 um
  - 100 um
  - 150 um
  - 200 um
  - Entire Dendrite Tree

- 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

Injection of Current (Soma)

Recording of Current (Soma)

Recording of Current (Apical Dendrite)

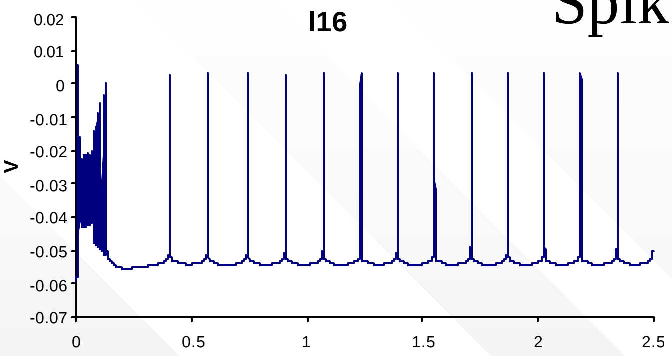
Basal Dendrite

Apical Dendrite

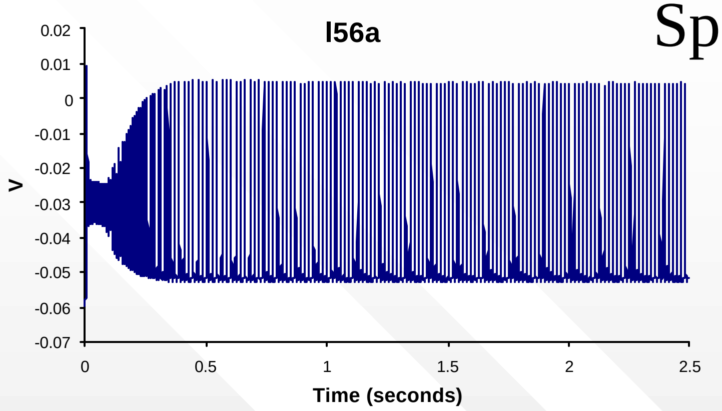
Soma (Cell Body)

Current Measurements

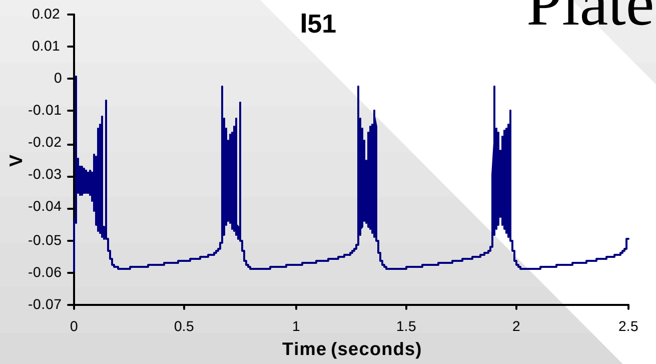
Simulated Physiological Response under 0.7 nA



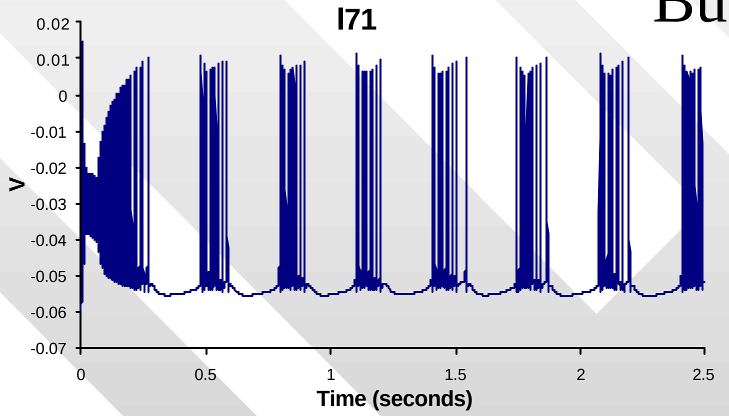
Spike



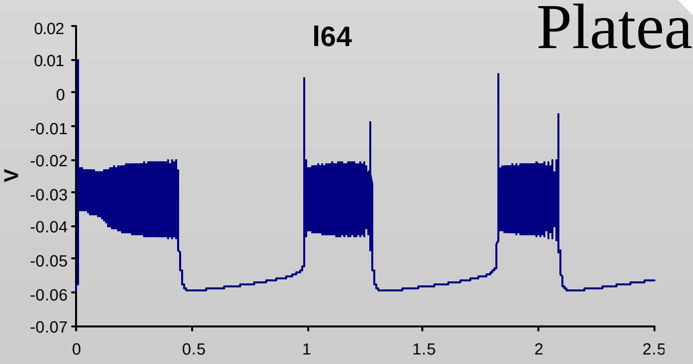
Spike



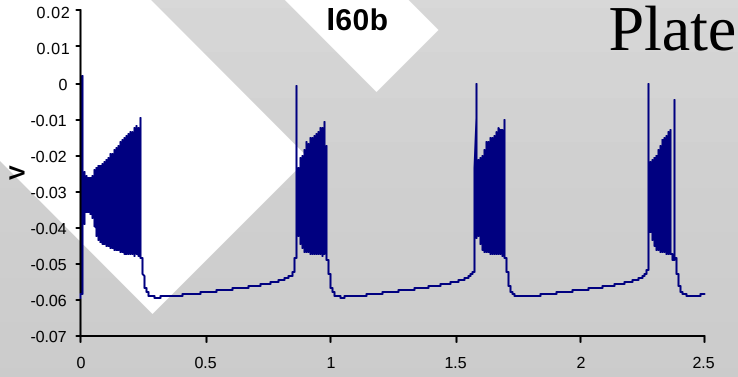
Plateau



Burst



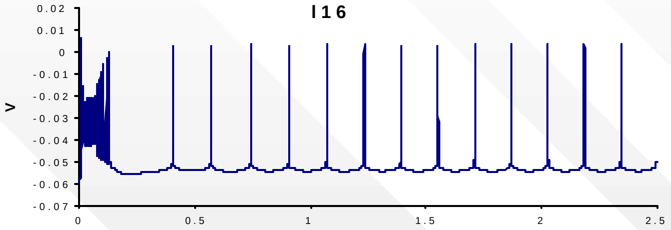
Plateau

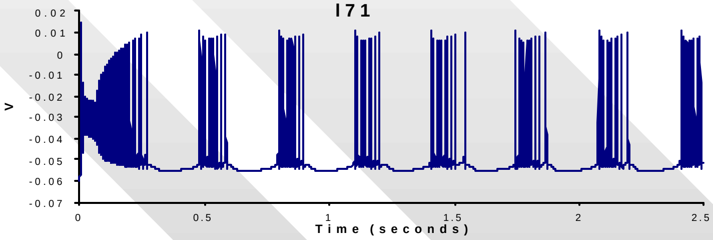


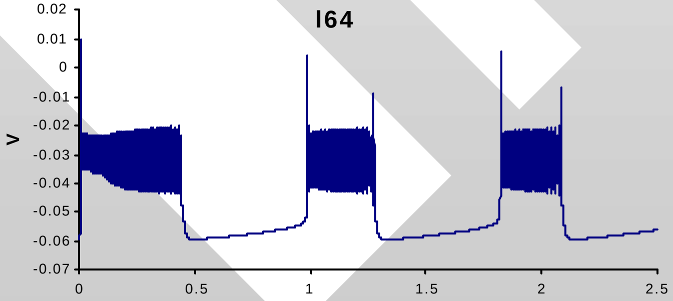
Plateau

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)

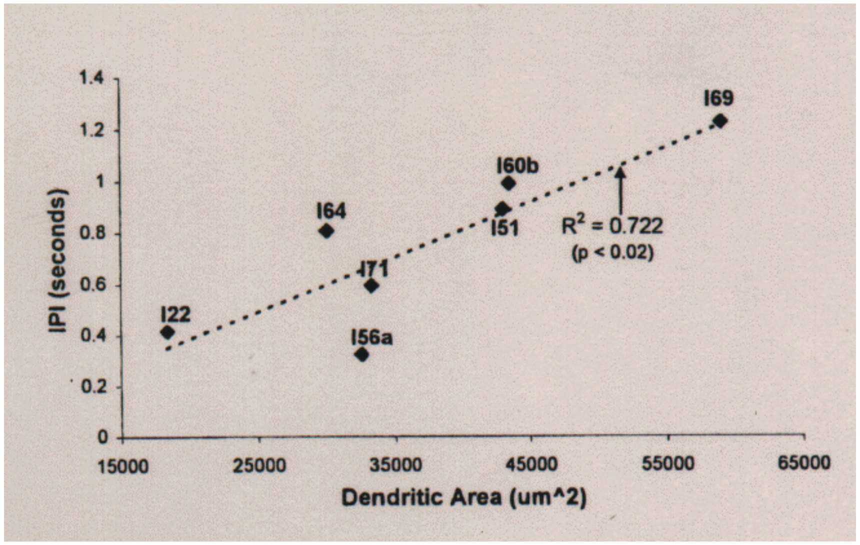






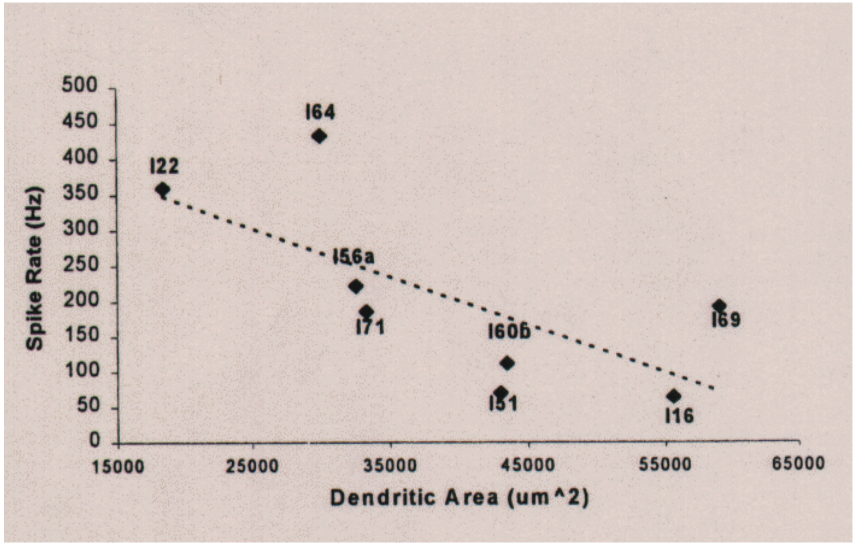
Influence of Dendritic Area on Firing Rate

Interplateau Interval vs Dendritic Area



Current: 0.5 nA

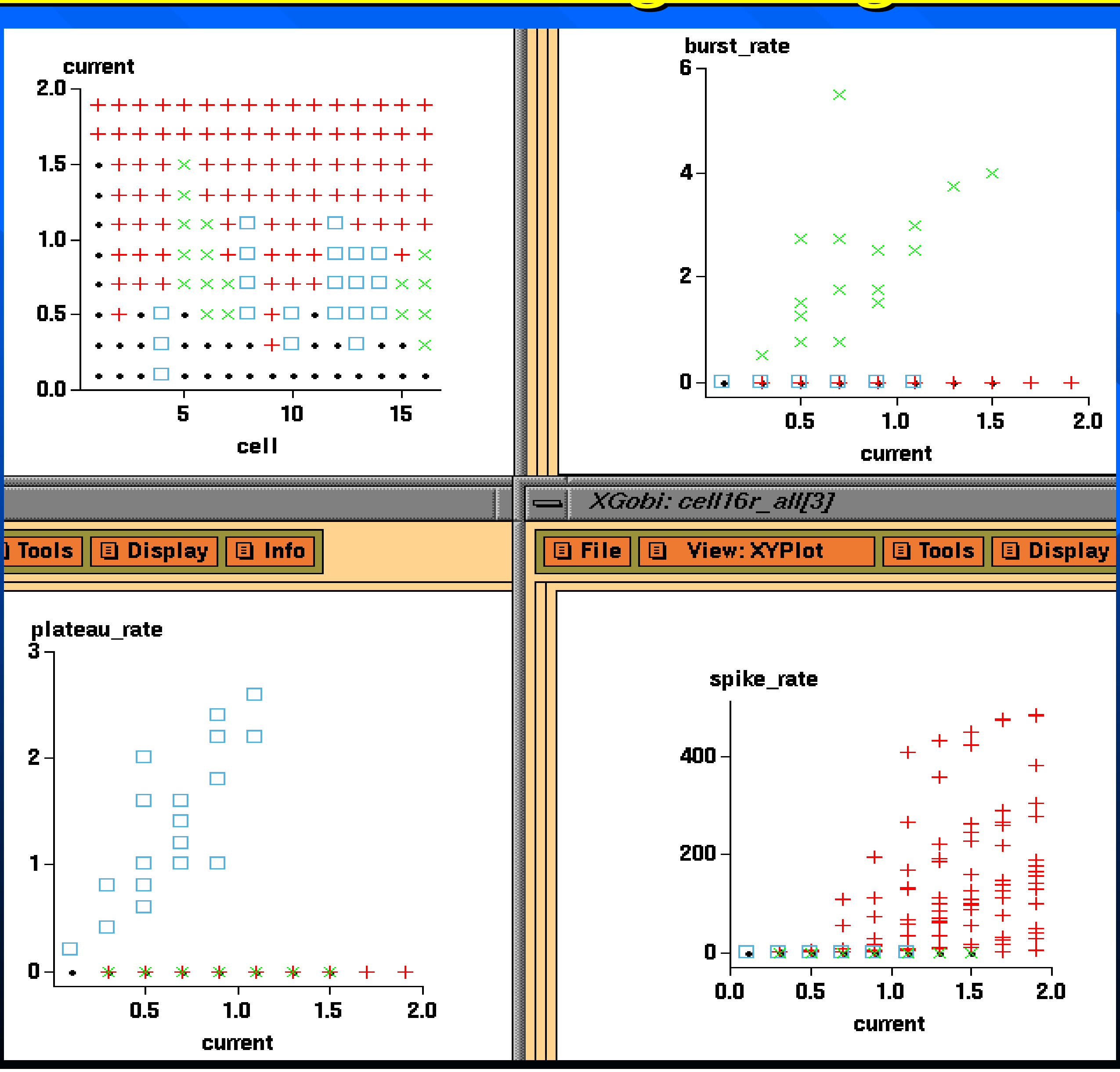
Spike Rate vs Dendritic Area



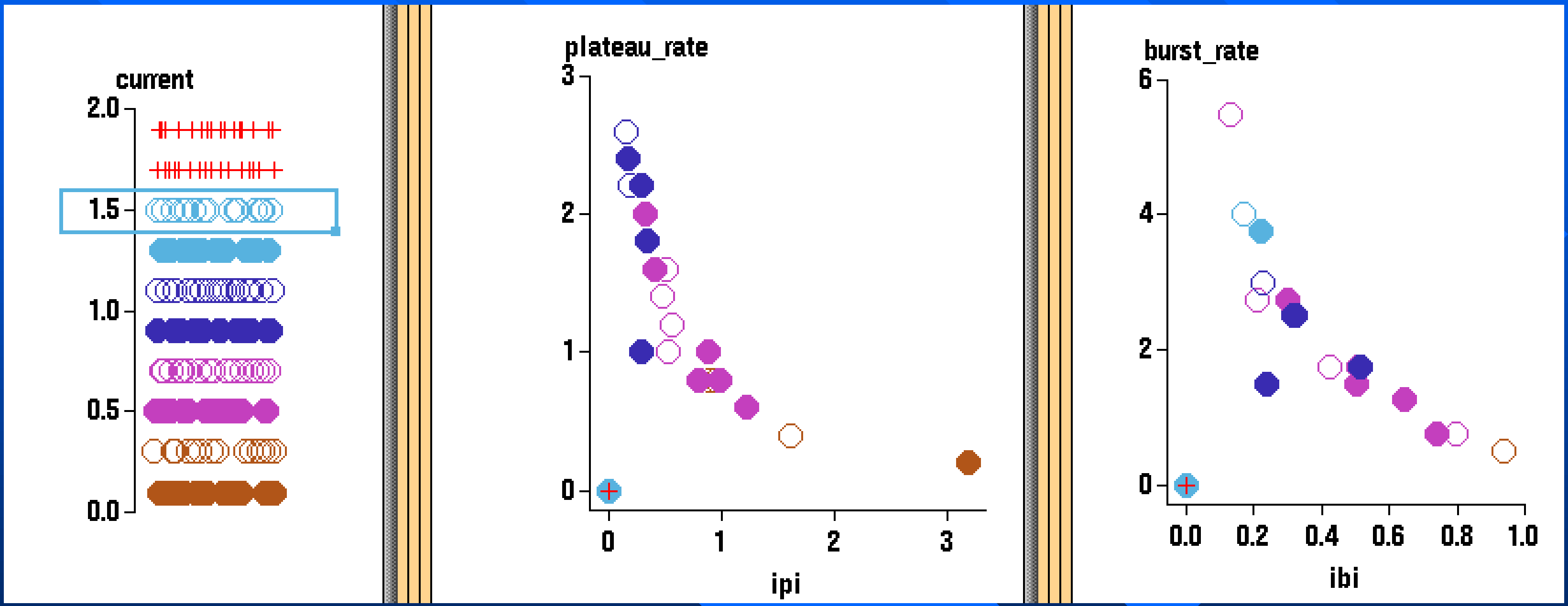
Current: 1.3 nA

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

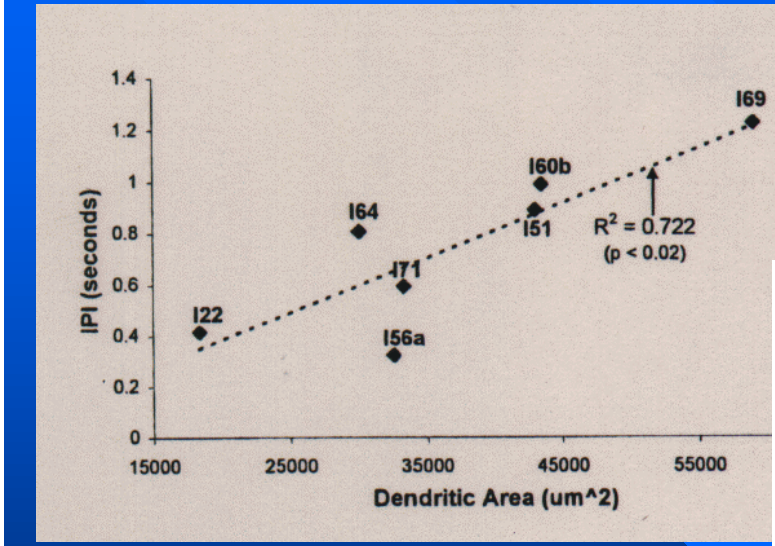
Visual Data Mining Using XGobi

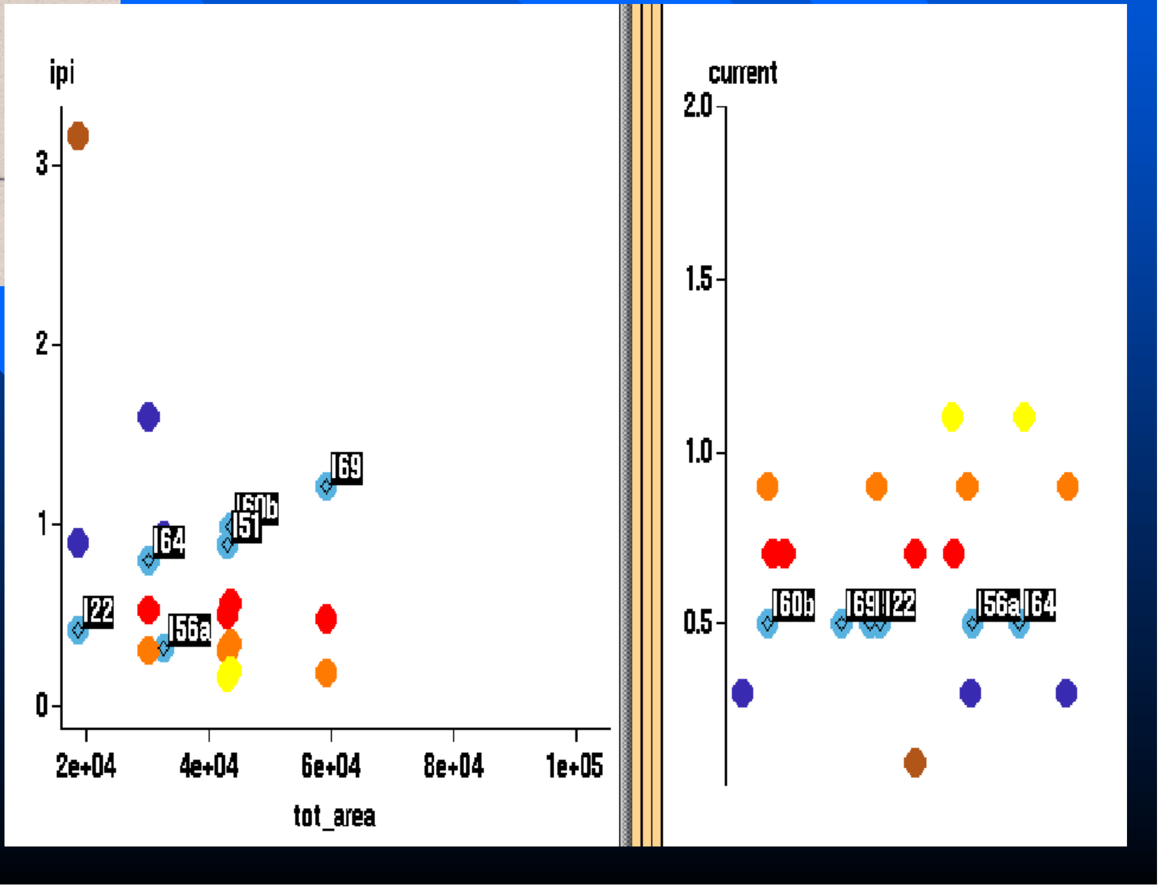


Visible Patterns



Interplateau Interval vs Dendritic Area ???





Brain Cells - Conclusions

- Visualization suggests which cells to simulate/analyze next.
- Some prior assumptions may not hold or only hold under additional restrictions.

Overall Conclusion

- Visual approach effective to see unexpected structure in data.
- Combination of different techniques most effective.
- Can be used for almost all types of data (another major application: Remote Sensing).

Contact

- Jürgen Symanzik
  - symanzik@sunfs.math.usu.edu
- Website
  - www.math.usu.edu/~symanzik