Visualizing high-dimensional data:

Applying graph theory to data visualization

Wayne Oldford

based on joint work with

Catherine Hurley (Maynooth, Ireland)

Adrian Waddell (Waterloo, Canada)

WATERLOO | MATHEMATICS
STATISTICS AND ACTUARIAL SCIENCE

Challenge

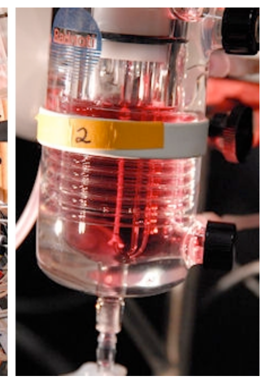
- p values on each of n individuals
- modern data: n, or p, or both, can be very large

Challenge

- p values on each of n individuals
- modern data: n, or p, or both, can be very large

Challenge

- p values on each of n individuals
- modern data: n, or p, or both, can be very large



Challenge

- p values on each of n individuals
- modern data: n, or p, or both, can be very large

Challenge

- p values on each of n individuals
- modern data: n, or p, or both, can be very large

Challenge

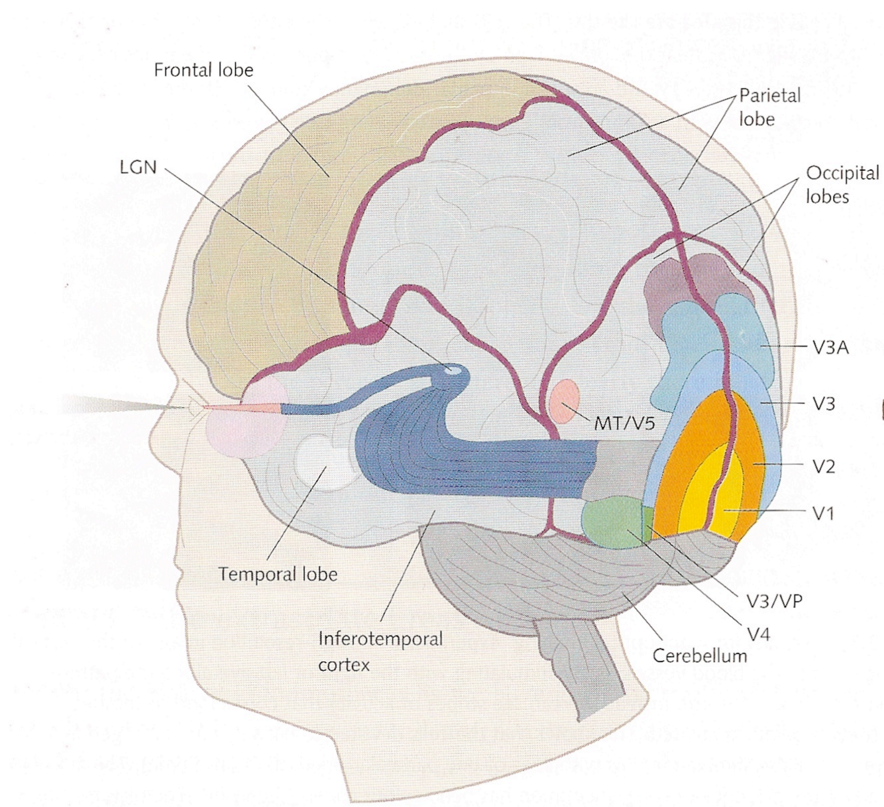
- p values on each of n individuals
- modern data: n, or p, or both, can be very large

- can have non-obvious variables, complex, unanticipated structure, ...

Data Visualization

powerful human visual system

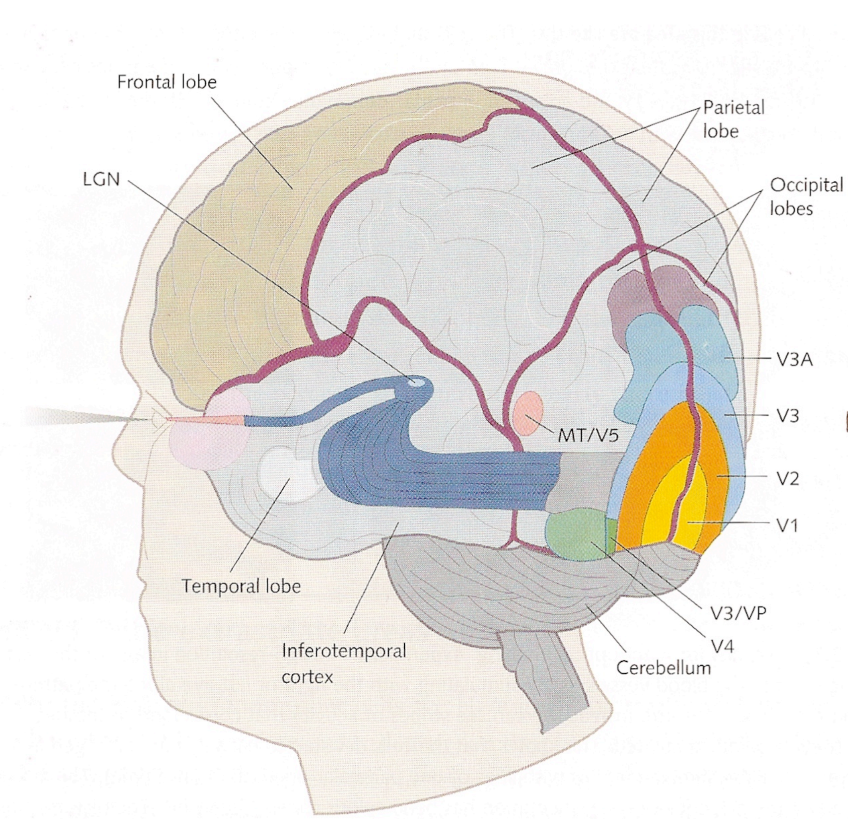
- use a variety of cues:
  - proximity, movement, shape, colour, texture, ...
- patterns, relations, like and unlike, ...
- recognition and discovery
- structure need not be anticipated



Data Visualization

powerful human visual system

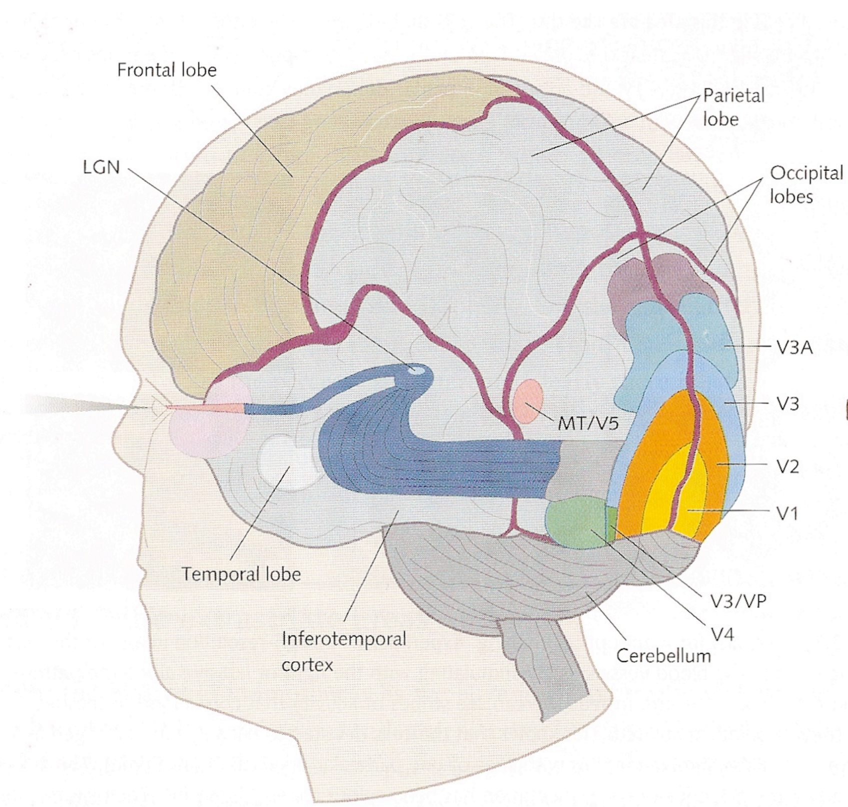
- use a variety of cues:
  - proximity, movement, shape, colour, texture, ...
- patterns, relations, like and unlike, ...
- recognition and discovery
- structure need not be anticipated



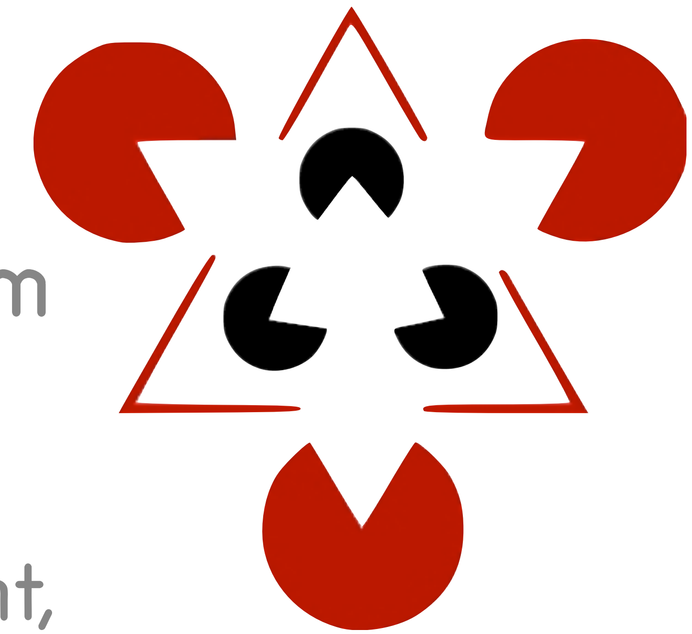
Data Visualization

powerful human visual system

- use a variety of cues:
  - proximity, movement, shape, colour, texture, ...
- patterns, relations, like and unlike, ...
- recognition and discovery
- structure need not be anticipated

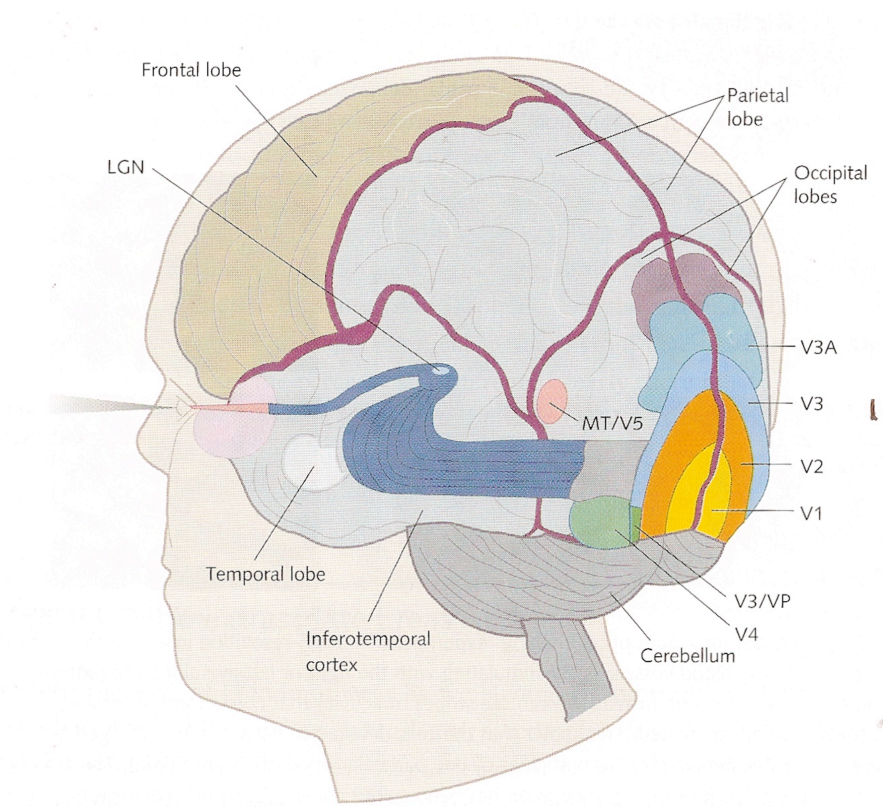


Data Visualization



powerful human visual system

- use a variety of cues:
    - proximity, movement, shape, colour, texture, ...
- patterns, relations, like and unlike, ...
- recognition and discovery
- structure need not be anticipated



Challenge

Large p

- visually, we're constrained to small p
  - locations: $p < 4$
  - use colour, shape, texture, movement, ...
- comprehension depends on only a few dimensions

... at a time

Challenge

Large p

- visually, we're constrained to small p
  - locations: $p < 4$
  - use colour, shape, texture, movement, ...
- comprehension depends on only a few dimensions

  ... at a time

- Approach: large number of low dimensional views

Challenge

Large p

- visually, we're constrained to small p
  - locations: p < 4
  - use colour, shape, texture, movement, ...
- comprehension depends on only a few dimensions

  ... at a time
- Approach: large number of low dimensional views
  - $\binom{p}{d}$ d-dimensional views, preferably highly interactive

Challenge

Large p

- visually, we're constrained to small p
  - locations: $p < 4$
  - use colour, shape, texture, movement, …
- comprehension depends on only a few dimensions

  … at a time
- Approach: large number of low dimensional views
  - $\binom{p}{d}$ d-dimensional views, preferably highly interactive
  - Which dimensions? How connected? How explored?

Axis systems

- Choice of coordinate axis layout
  - Radial (`PairViz` `R` package)
  - Parallel (`PairViz` `R` package)
  - Orthogonal (`RnavGraph` `R` package)

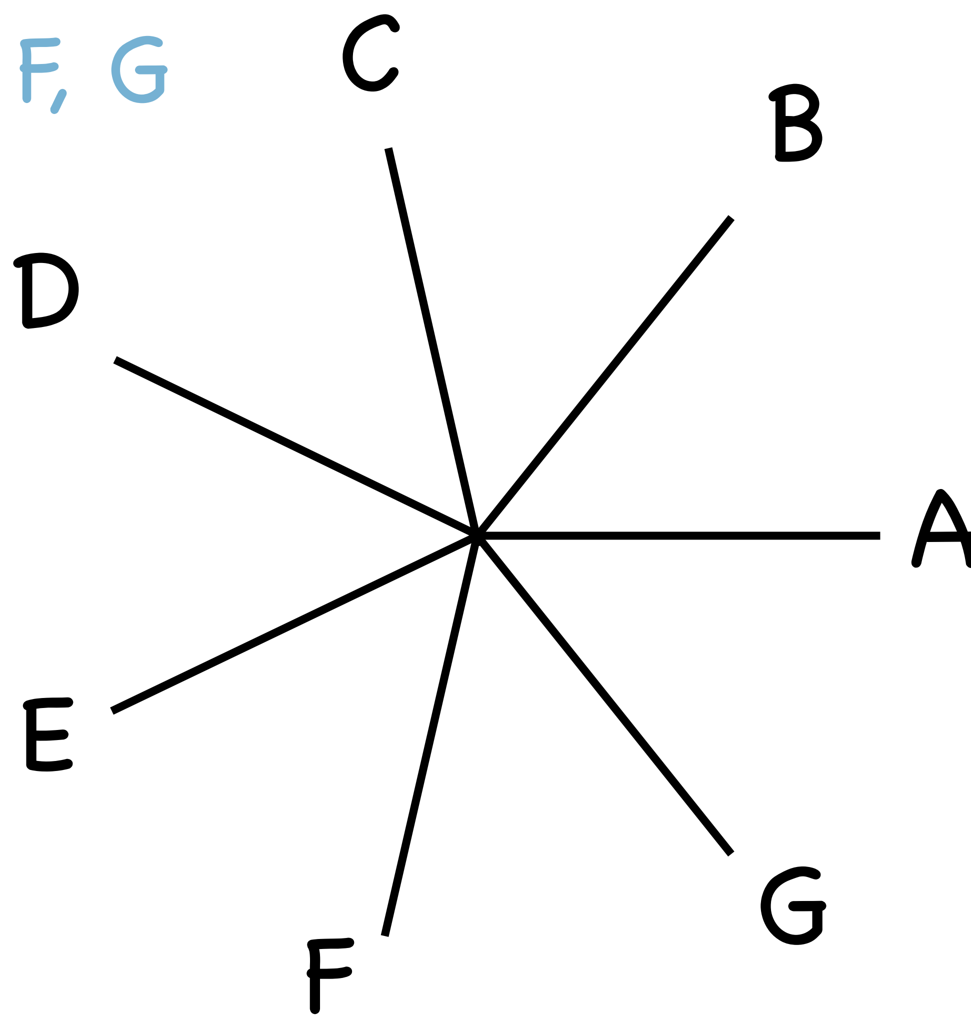
- Punchline
  - graph theory framework for exploratory data analysis looks very promising

Radial axes

7 dim: A, B, C, D, E, F, G

Radial axes

7 dim: A, B, C, D, E, F, G

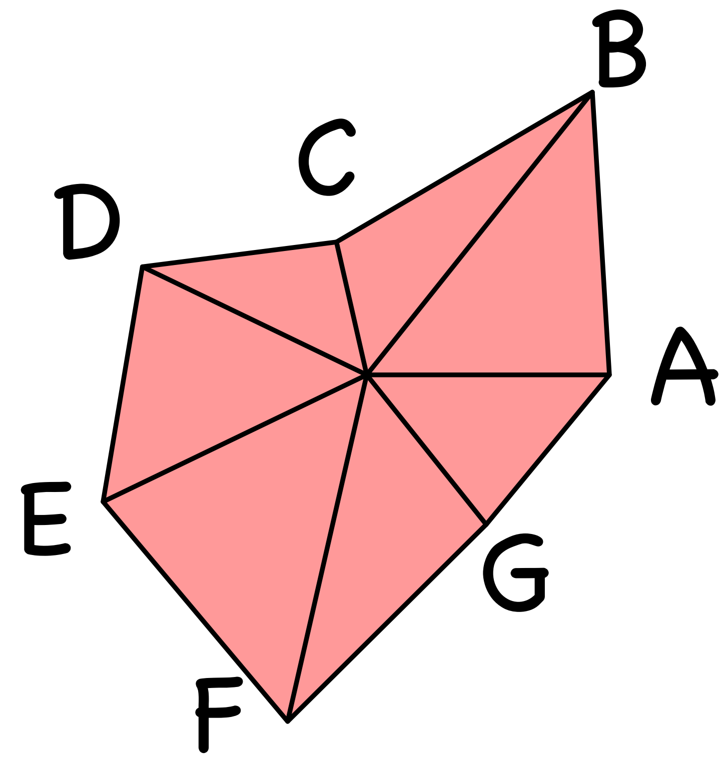


Equi-angular radial axes

Radial axes

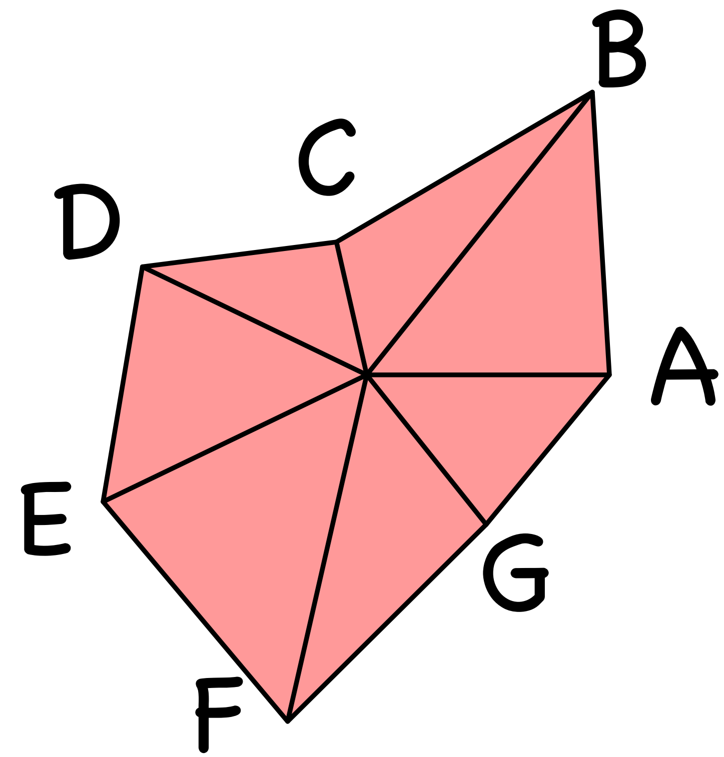
7 dim: A, B, C, D, E, F, G

Equi-angular radial axes

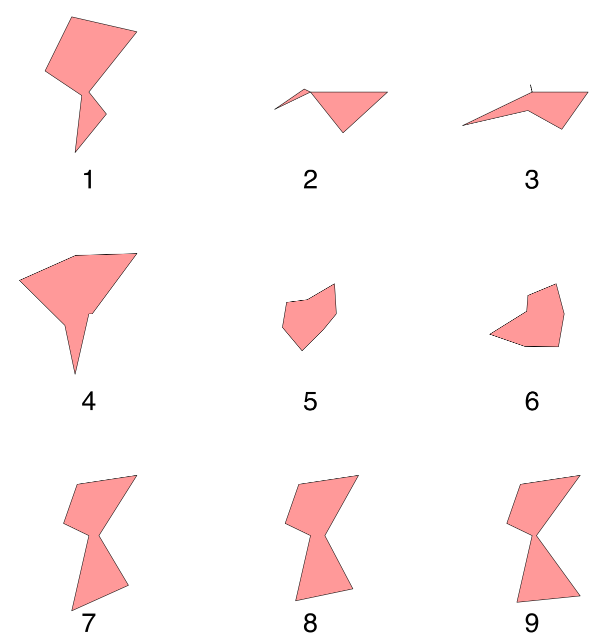


- Length of each radius is proportional to value of variate for that case.
- High dimensional data, each case represented by a star shaped glyph

Radial axes:



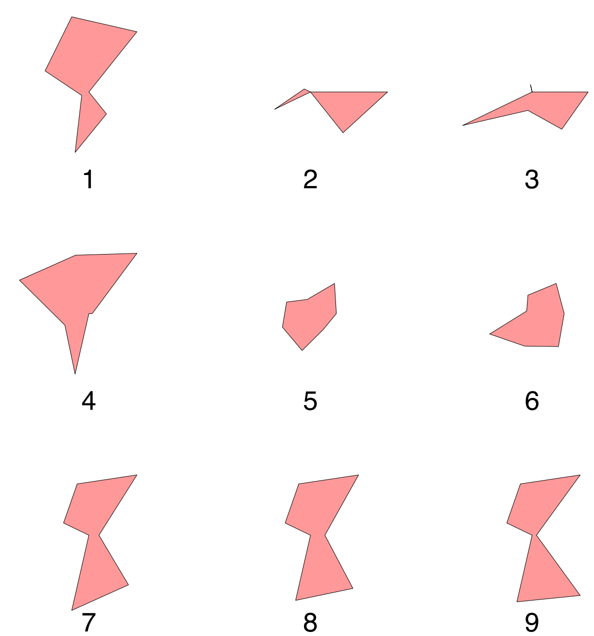
Radial axes:



- Compare cases by shape of glyph, here 9 cases in 7 dimensions
- Visually cluster high dimensional data by shape:

{7,8,9,1} {2,3} {4} {5,6} ?

Radial axes:

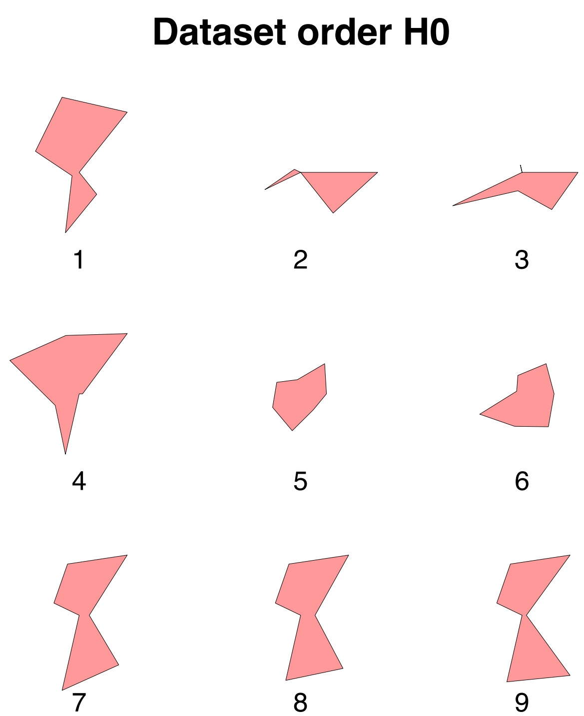


- Compare cases by shape of glyph, here 9 cases in 7 dimensions
- Visually cluster high dimensional data by shape:

  {7,8,9,1} {2,3} {4} {5,6} ?

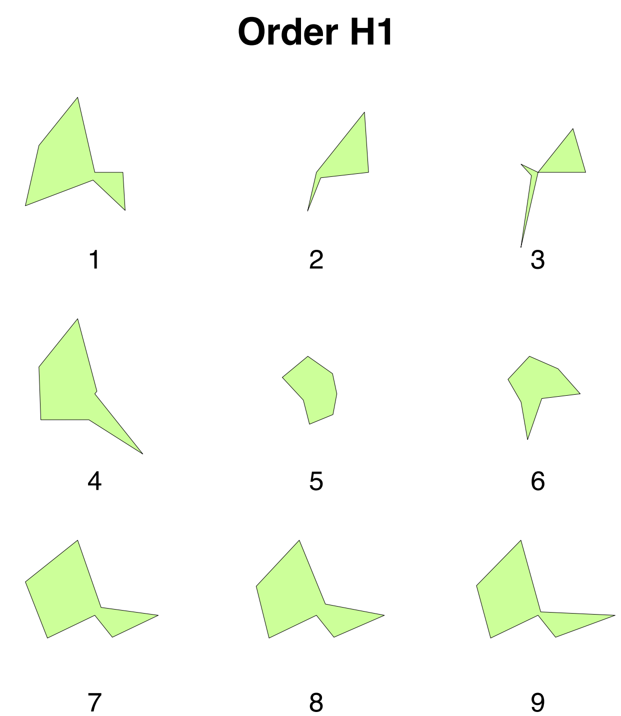
- What if the variables were assigned in a different order?

Radial axes: order effect



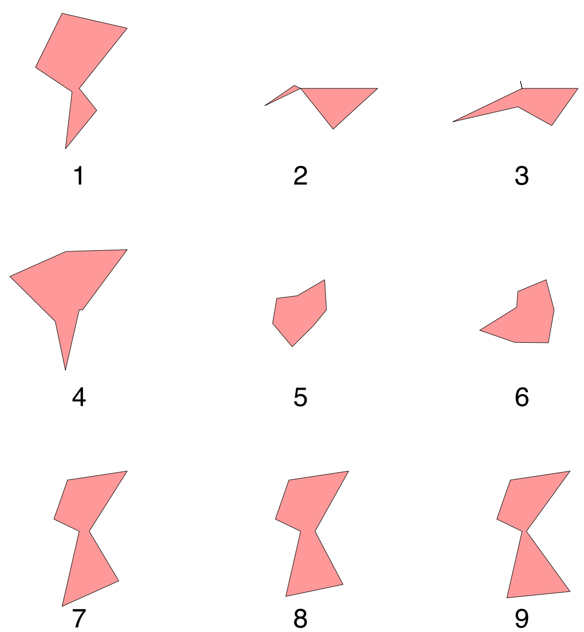
{7,8,9,1} {2,3} {4} {5,6} ?

{7,8,9,1} {2,3} {4} {5} {6} ?



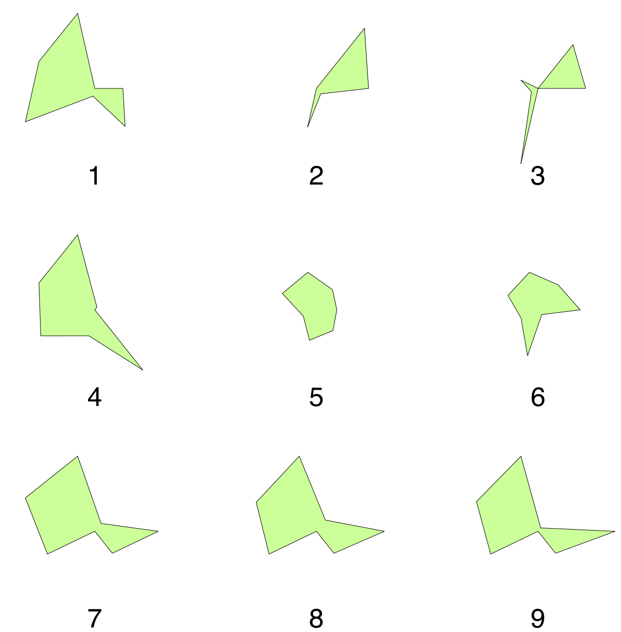
Radial axes: order effect

Dataset order H0



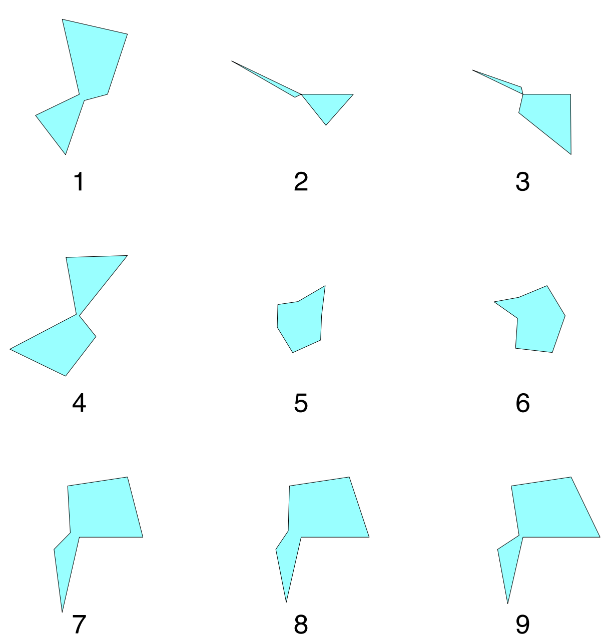
{7,8,9,1} {2,3} {4} {5,6} ?

Order H1



{7,8,9,1} {2,3} {4} {5} {6} ?

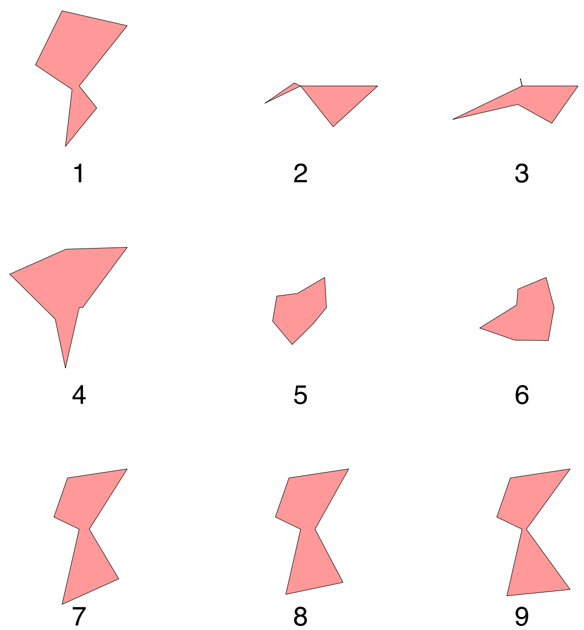
Order H2



{1,4} {2,3} {5,6} {7,8,9} ?

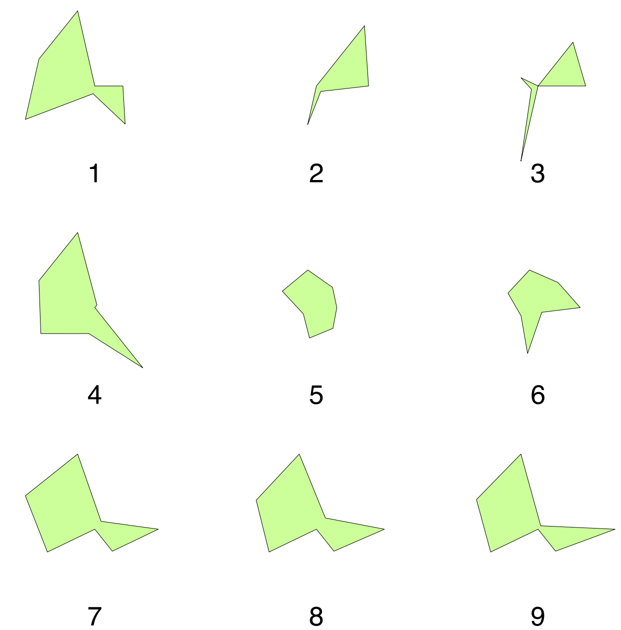
Radial axes: order effect

Dataset order H0



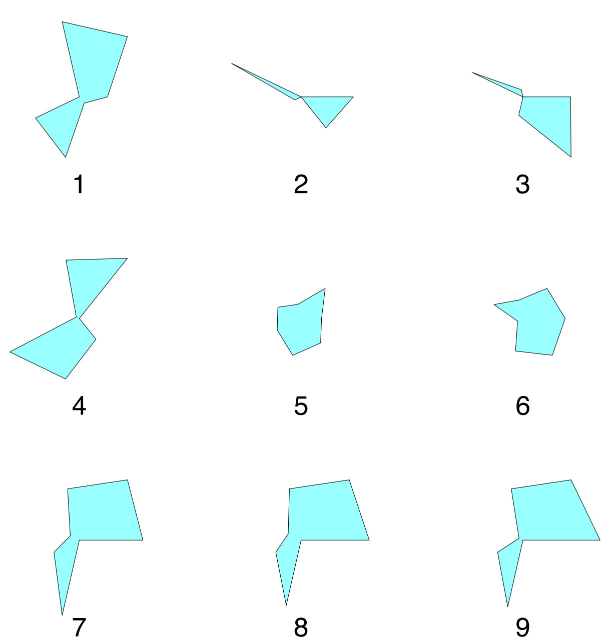
{7,8,9,1} {2,3} {4} {5,6} ?

Order H1



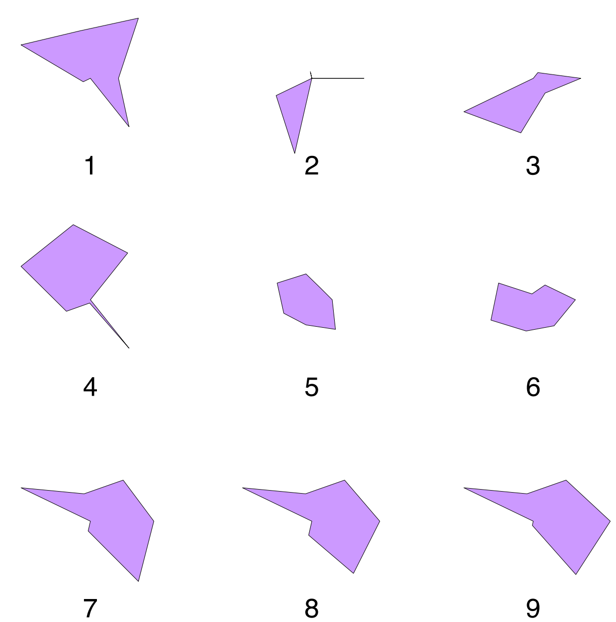
{7,8,9,1} {2,3} {4} {5} {6} ?

Order H2



{1,4} {2,3} {5,6} {7,8,9} ?

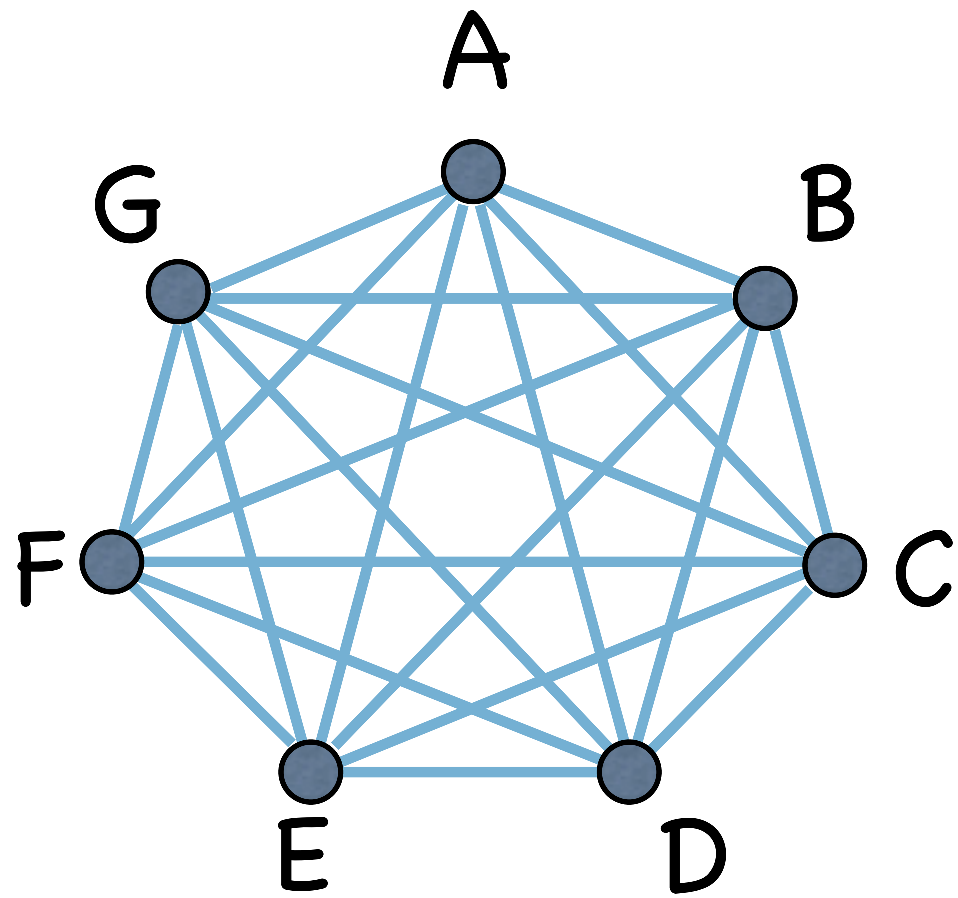
Order H3



{1} {2} {3} {4} {5} {6} {7,8,9} ?

Radial axes: reduced order effect

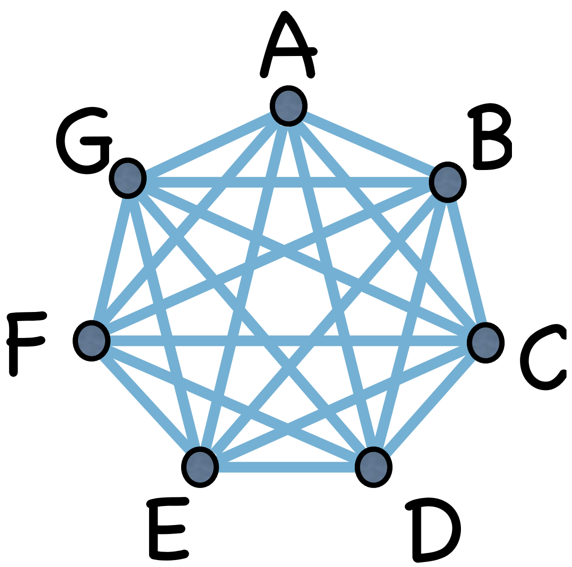
Instead, have all pairs of variables appear together



- Default was an arbitrary selection of pairs such that all nodes were visited once. If via a path (cycle), it is a Hamiltonian path (cycle).
- Visiting all edges would give all pairs in some order.

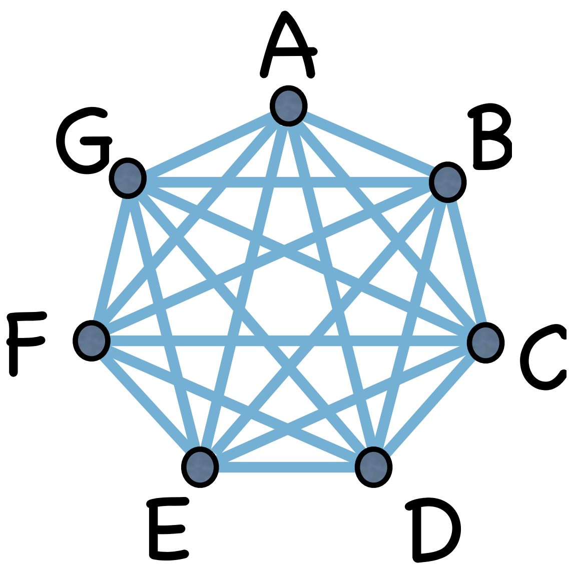
Radial axes: reduced order effect

An Eulerian

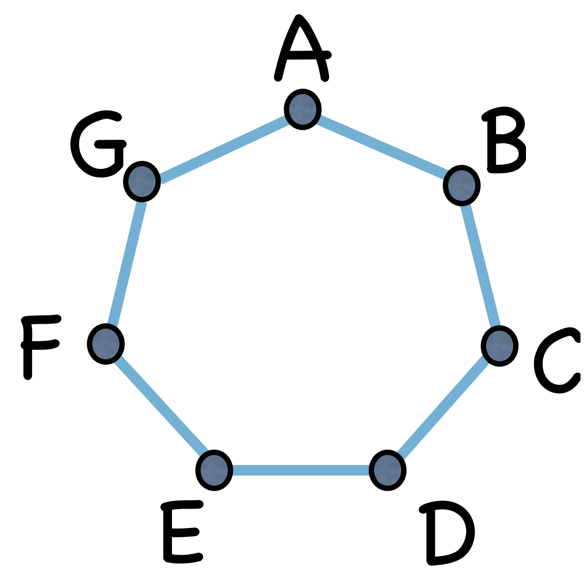


Radial axes: reduced order effect

An Eulerian

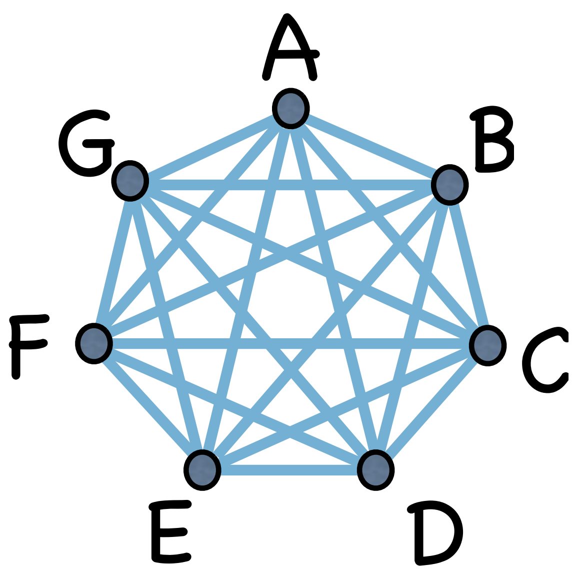


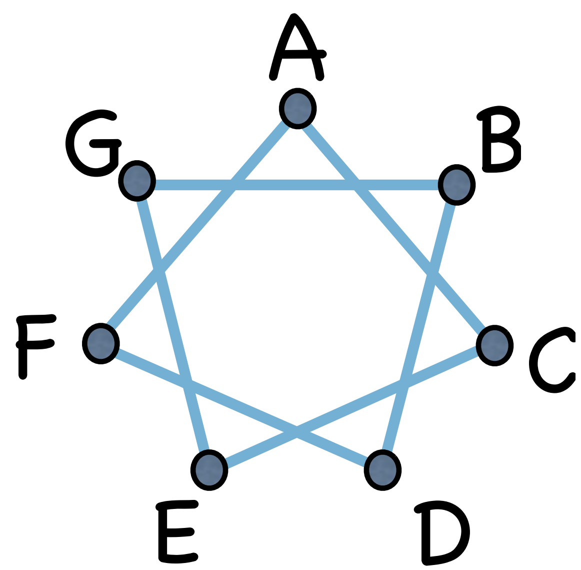
A Hamiltonian



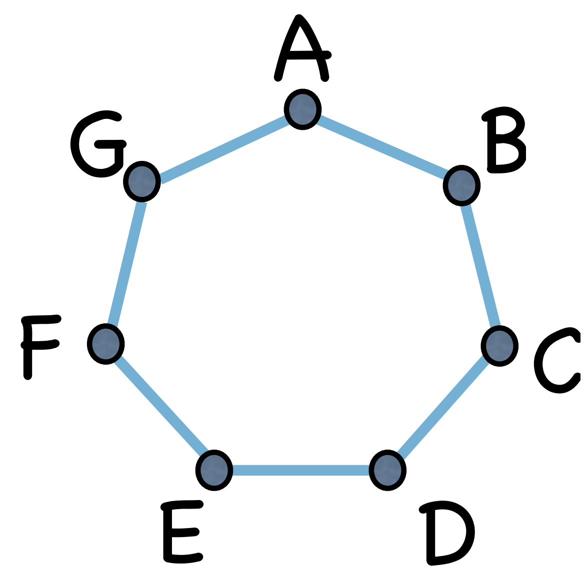
Radial axes: reduced order effect

An Eulerian



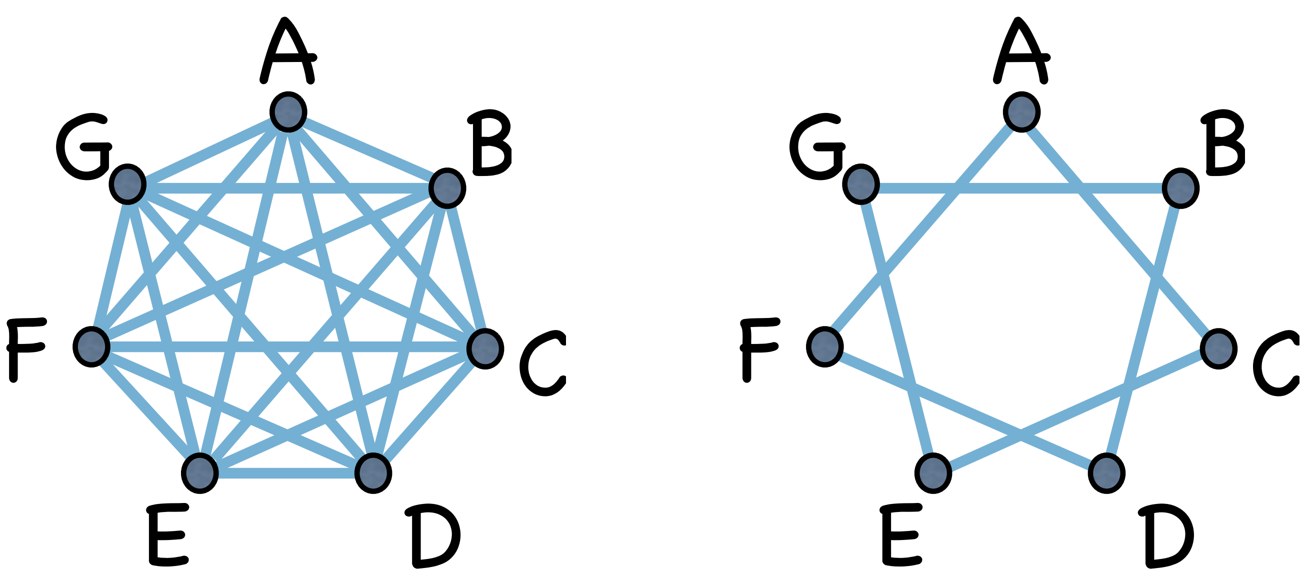


A Hamiltonian

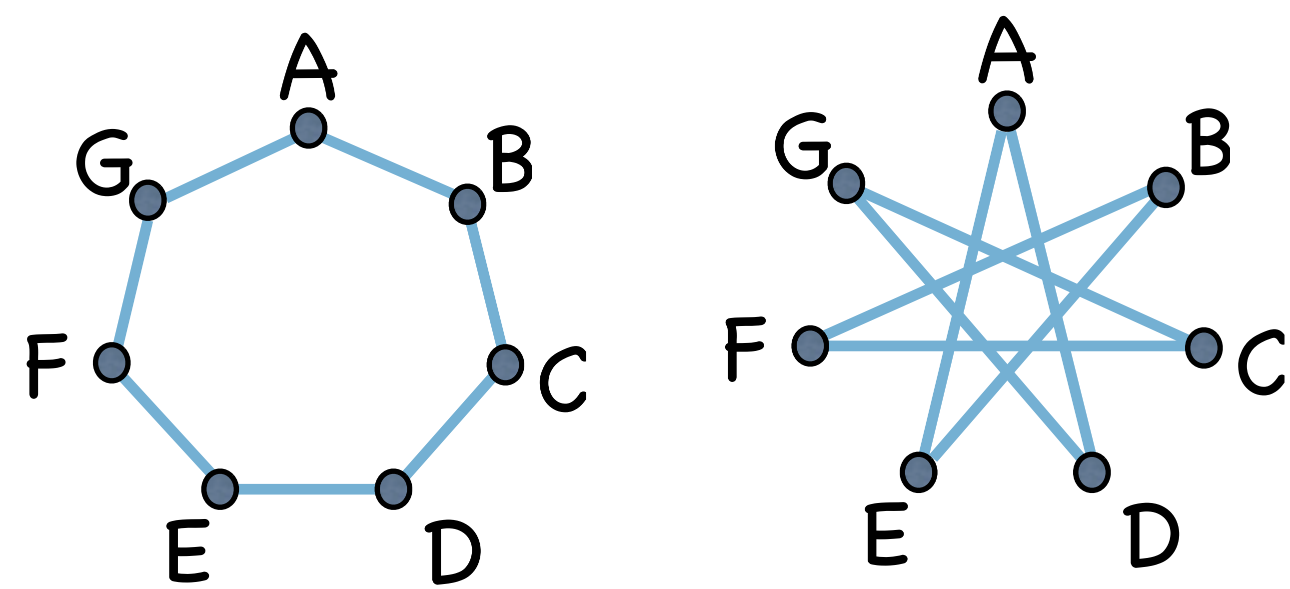


Radial axes: reduced order effect

An Eulerian

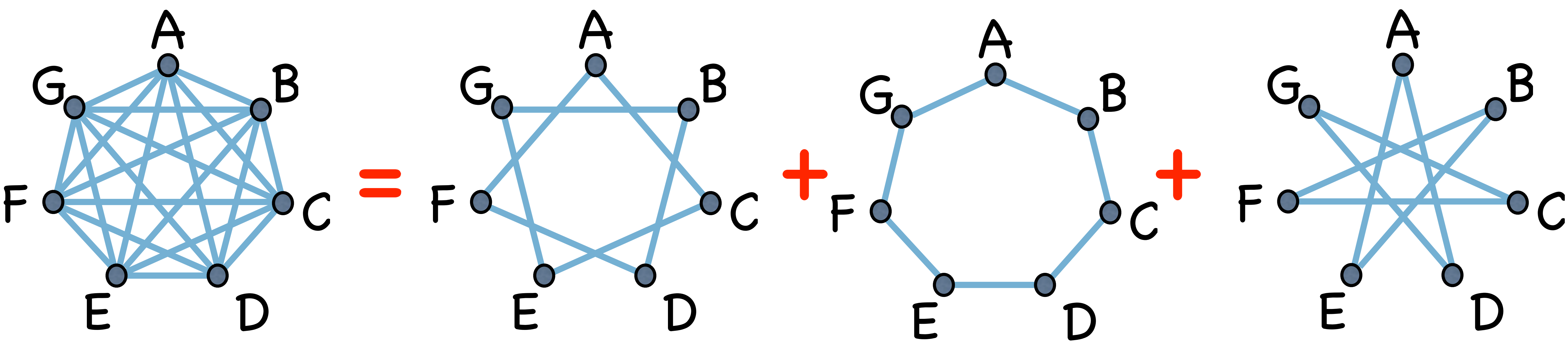


A Hamiltonian



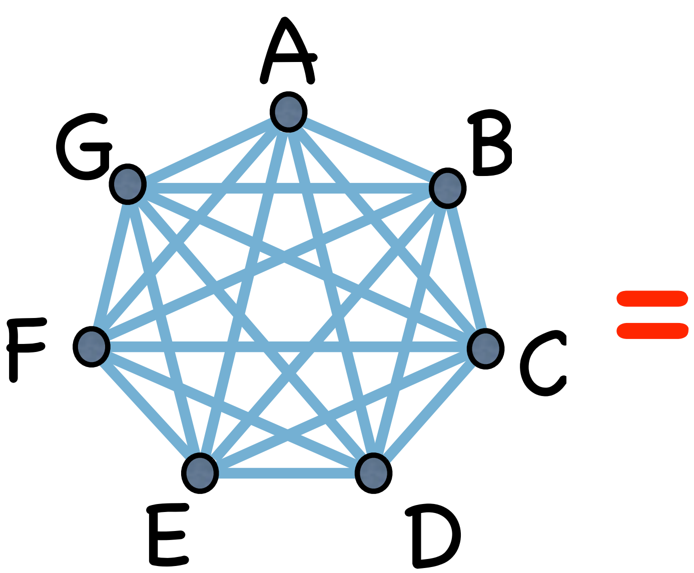
Radial axes: reduced order effect

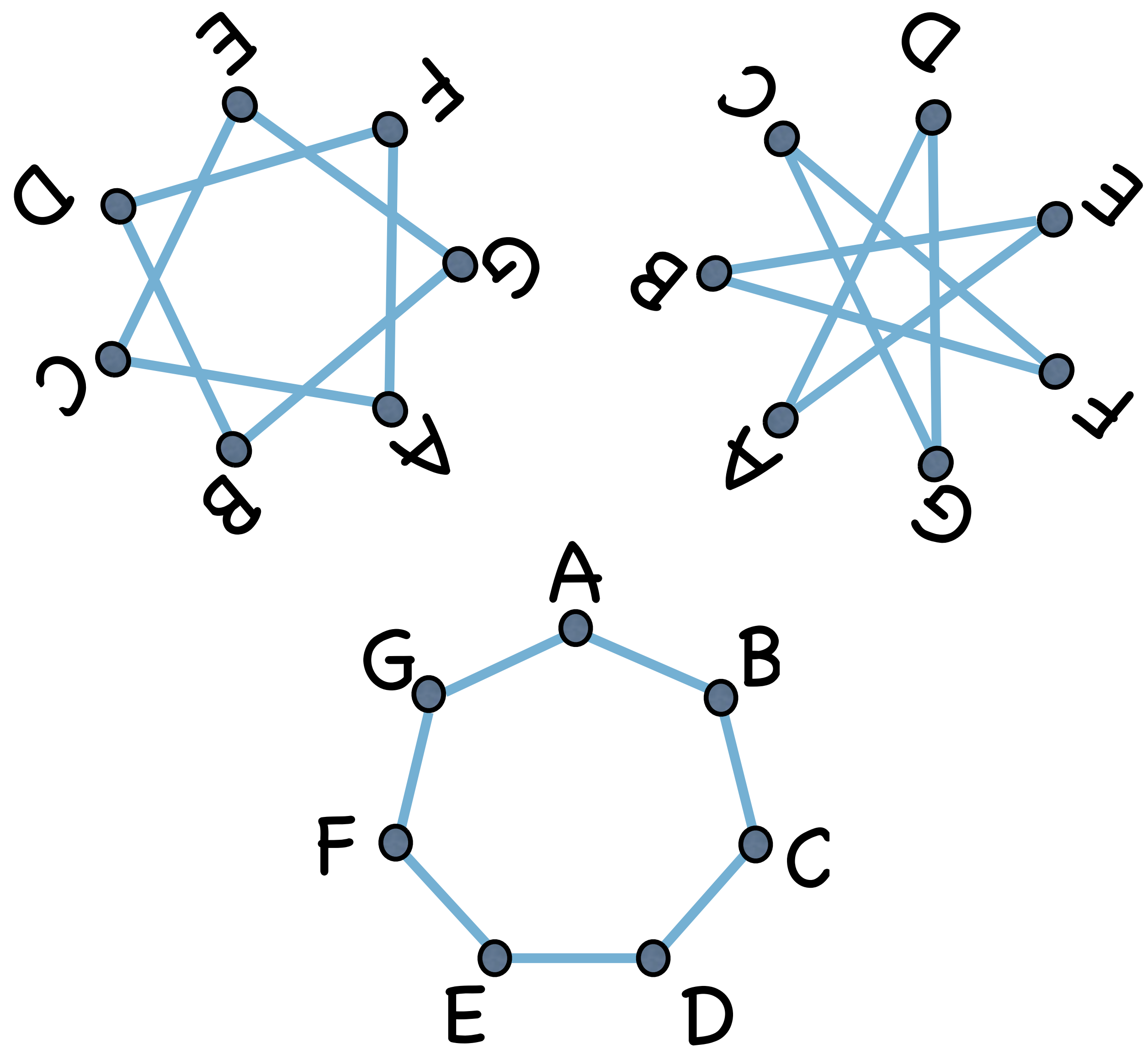
A Hamiltonian decomposition



Radial axes: reduced order effect

A Hamiltonian decomposition

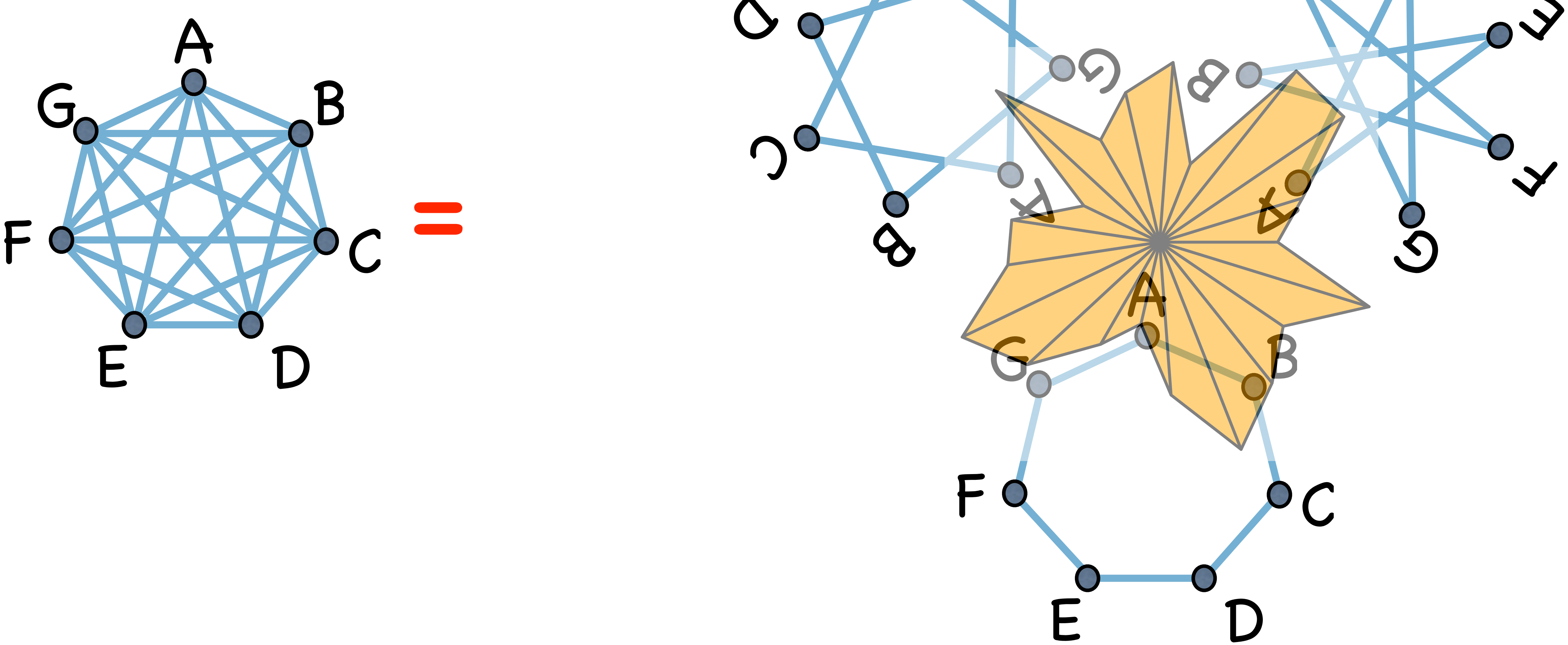




- Which when assembled form an Eulerian cycle composed of Hamiltonians

Radial axes: reduced order effect

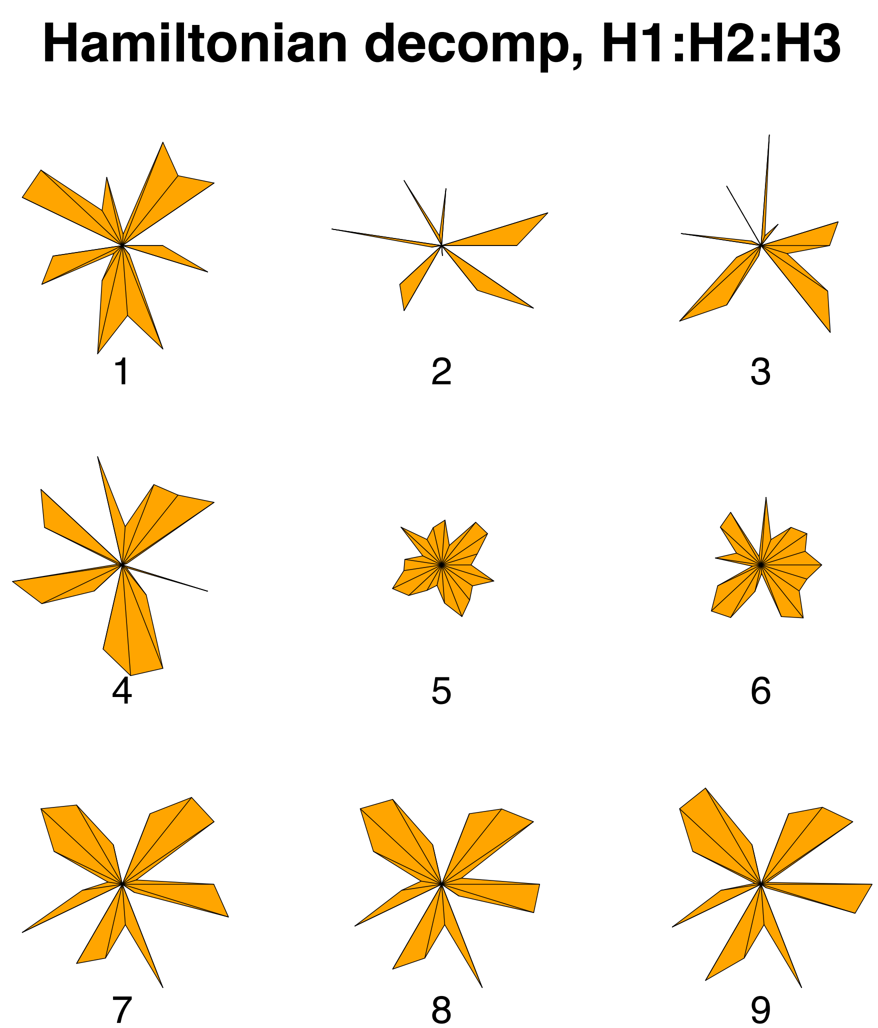
A Hamiltonian decomposition



- Which when assembled form an Eulerian cycle composed of Hamiltonians
- Could build a glyph from these cycles (21 radii instead of 7)

Radial axes: reduced order effect

A Hamiltonian decomposition

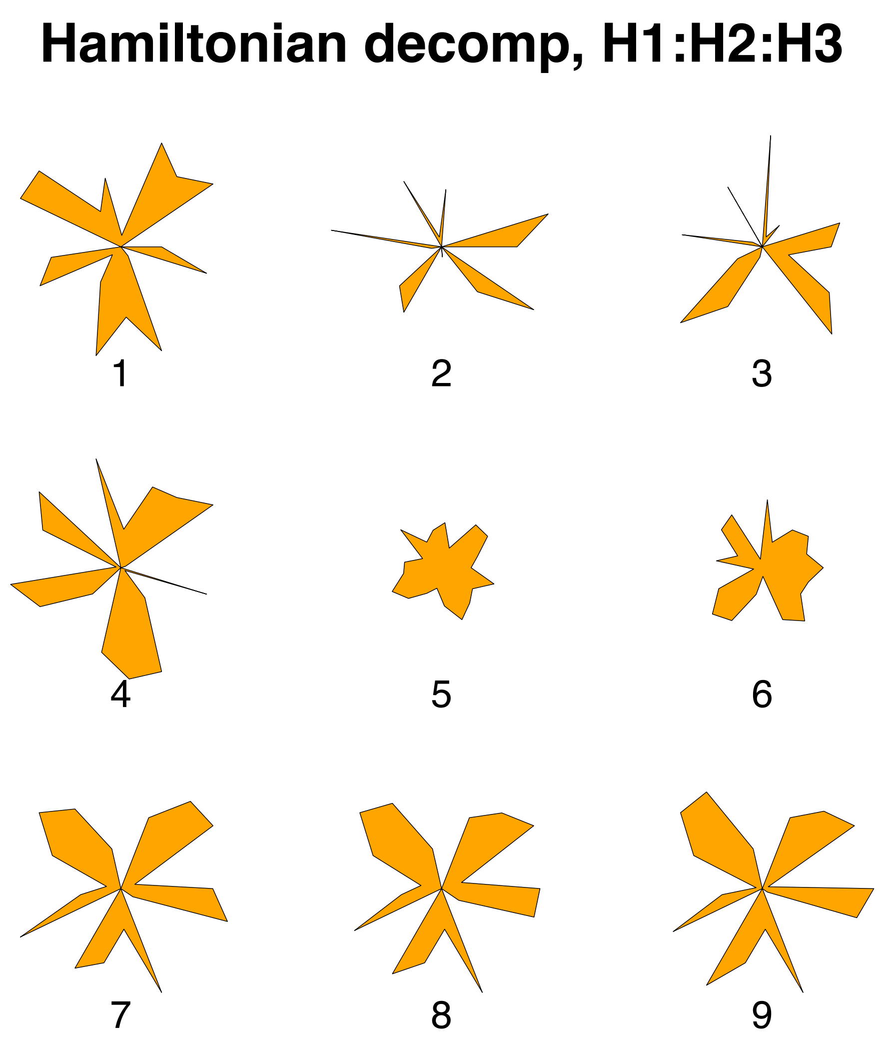


- Could build a glyph from these cycles (21 radii instead of 7)

Radial axes glyphs

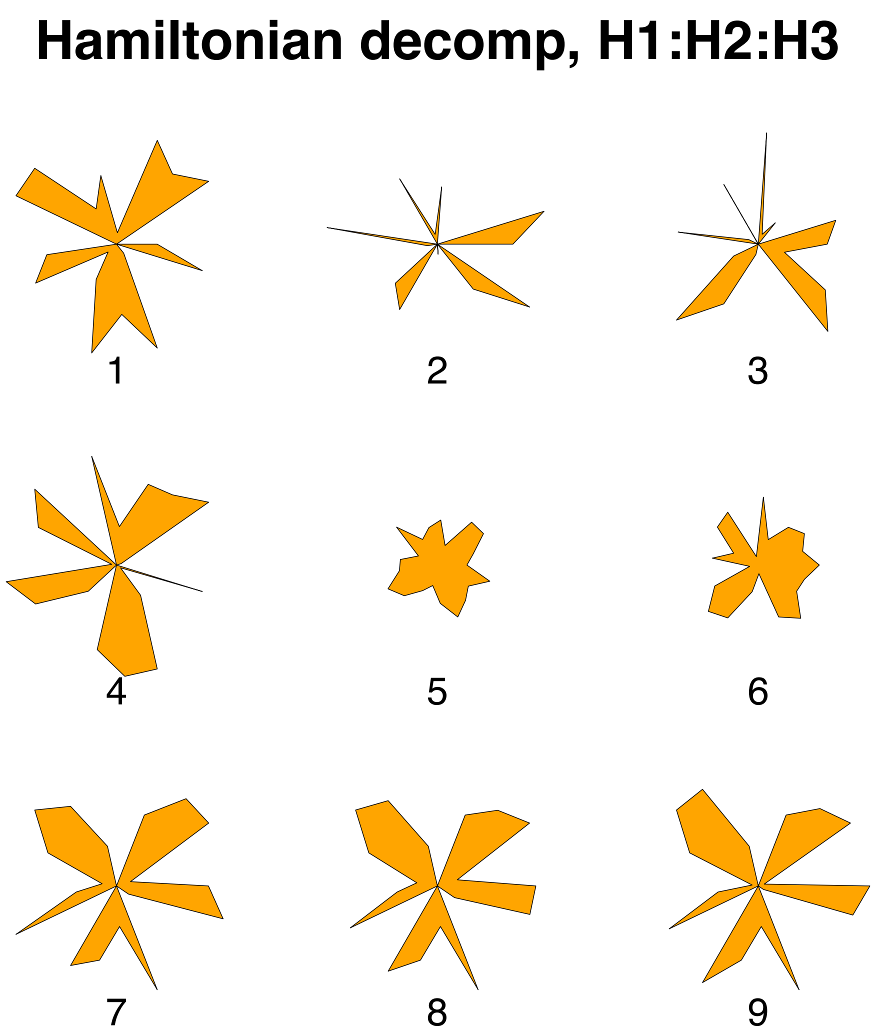
A Hamiltonian decomposition

{1,4} {2,3} {5,6} {7,8,9}

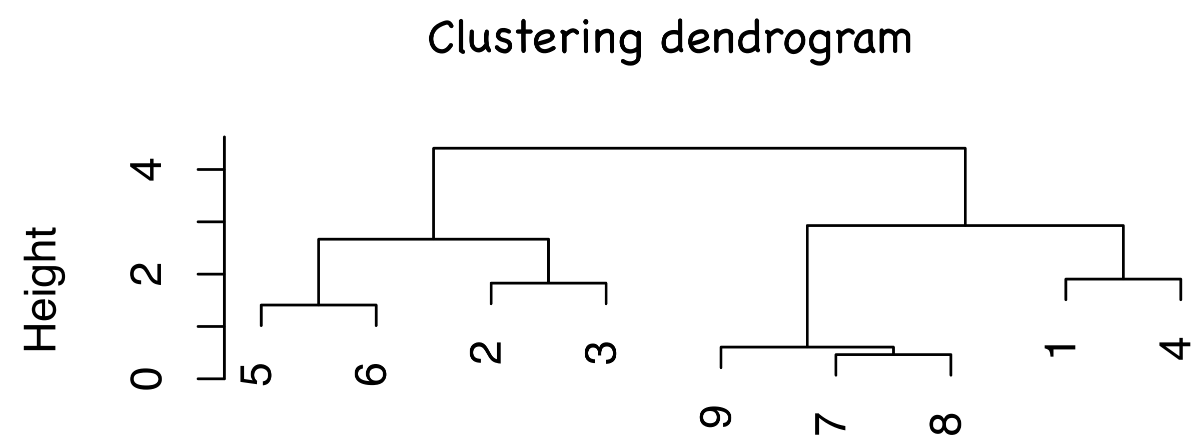


Radial axes glyphs

A Hamiltonian decomposition

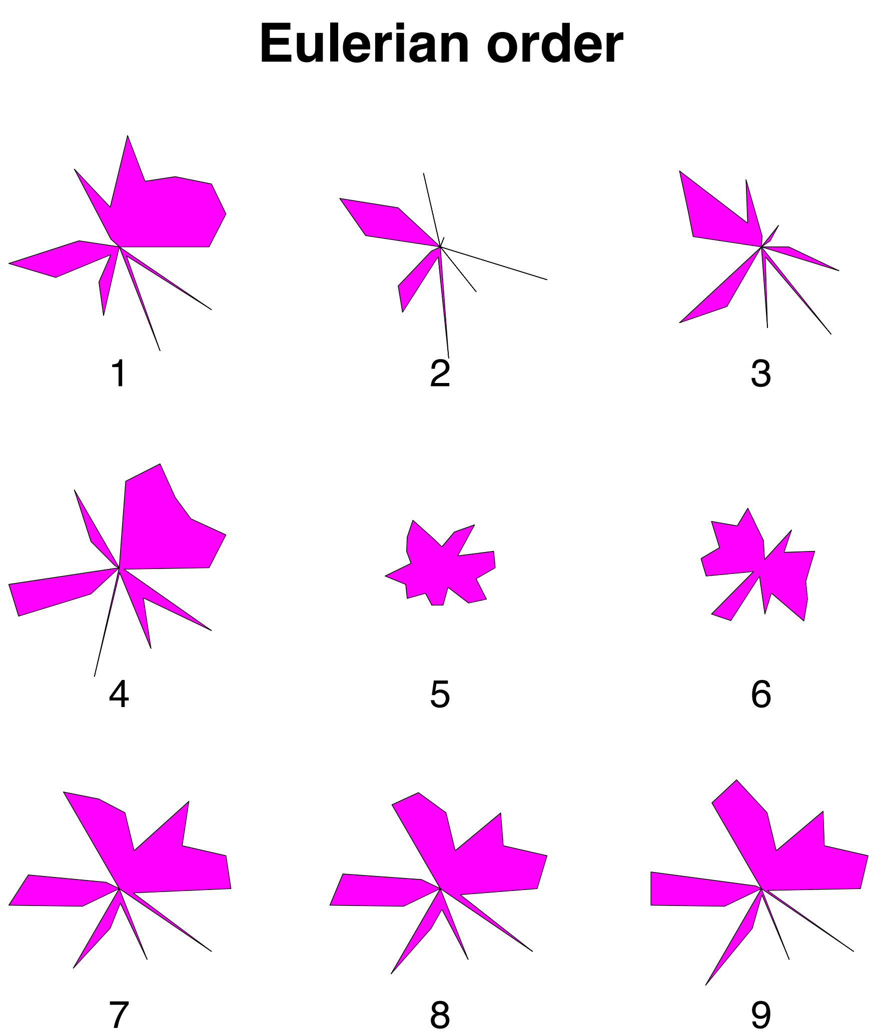


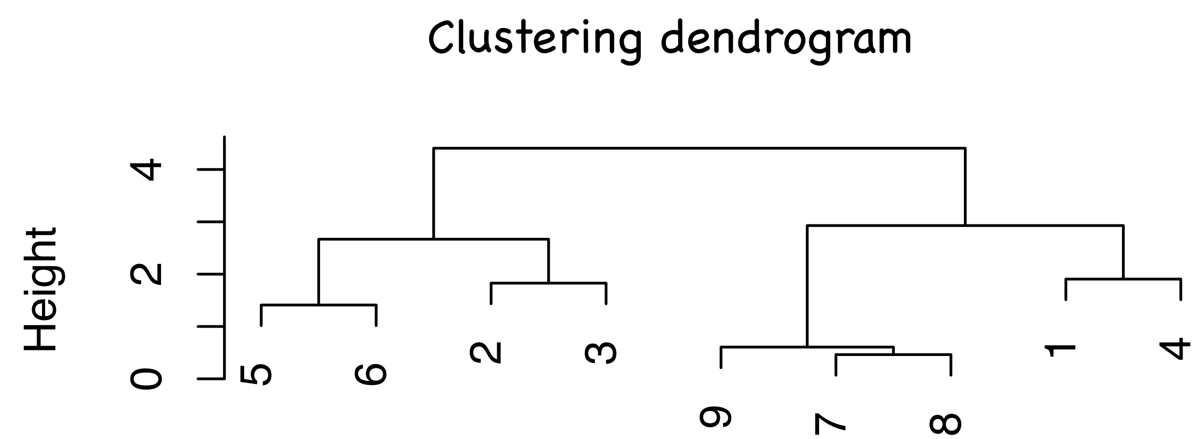
{1,4} {2,3} {5,6} {7,8,9}



Radial axes glyphs

A Greedy Eulerian (maximizing pairwise correlation)

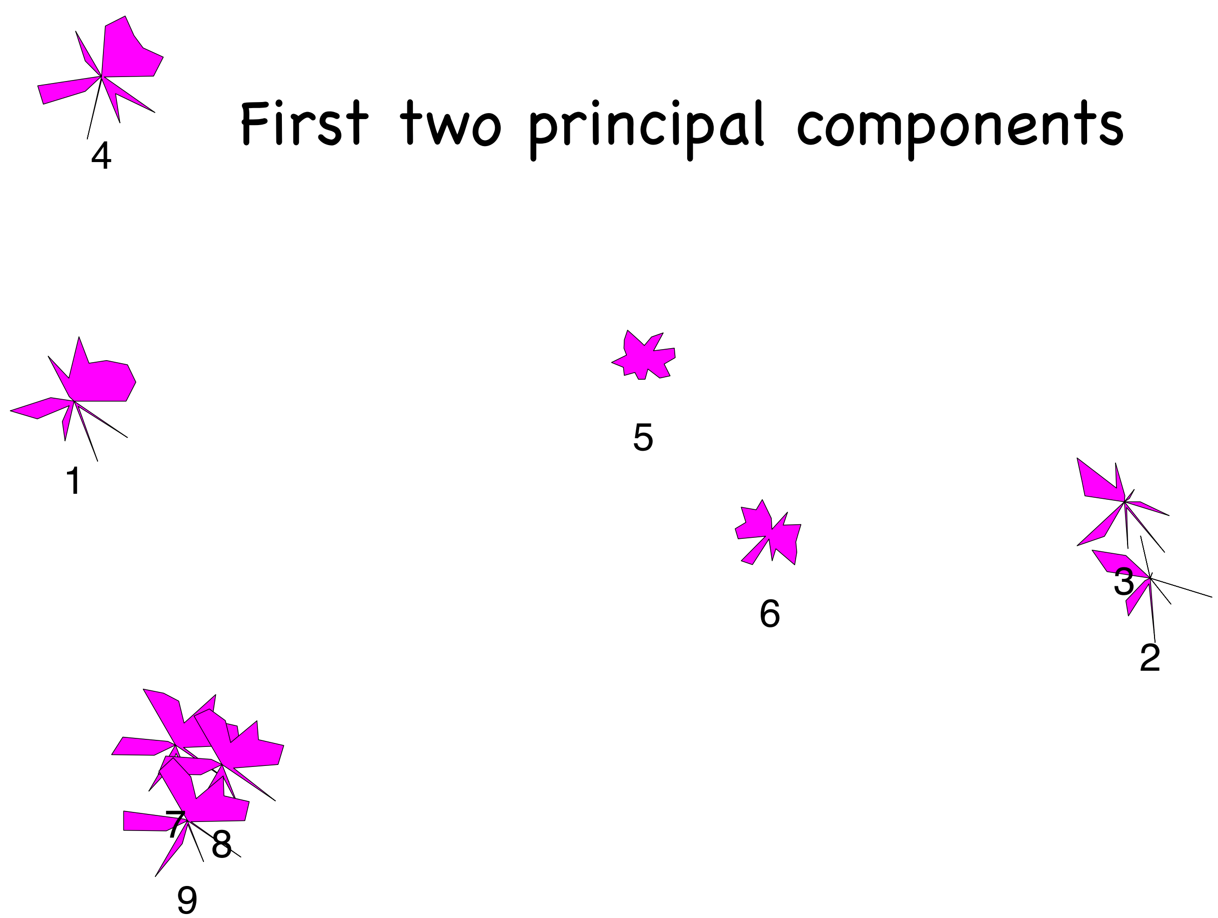


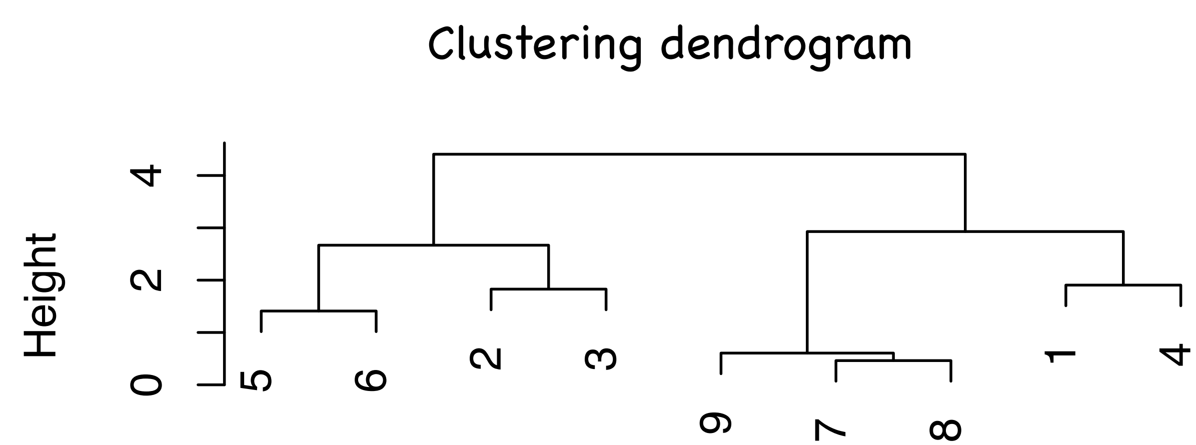


Radial axes glyphs

A Greedy Eulerian (maximizing pairwise correlation)

First two principal components

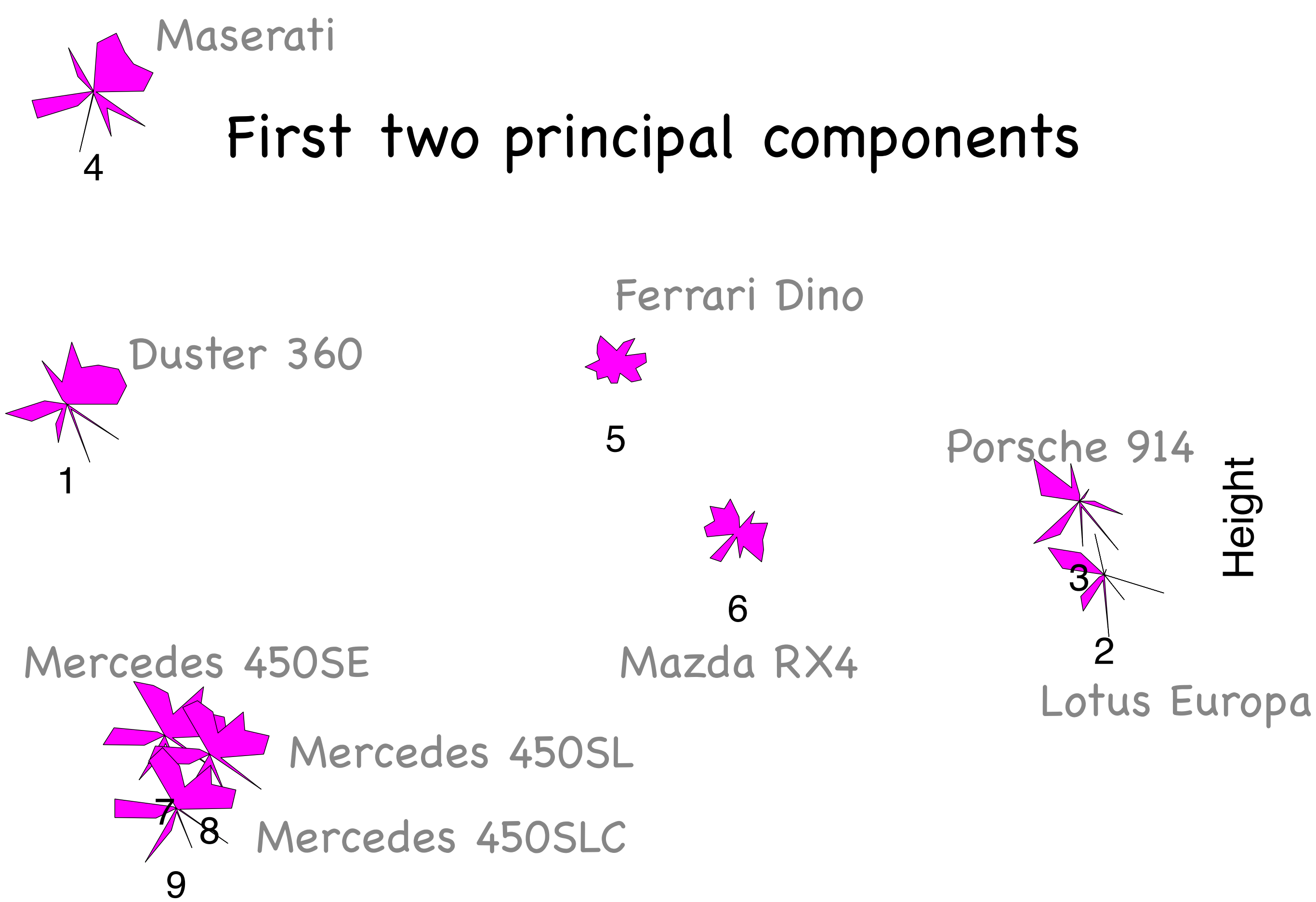


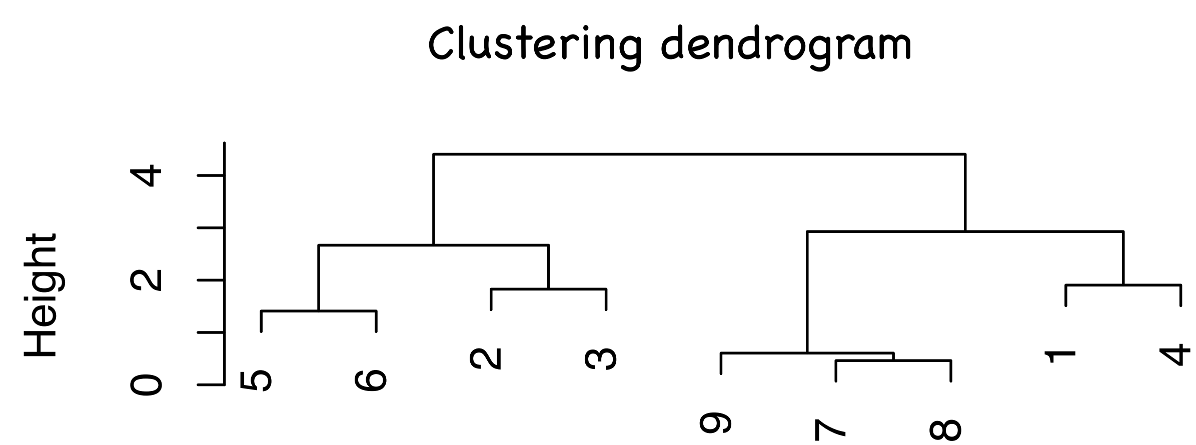


All pairs (Greedy Eulerian or Hamiltonian decomposition) reduce the effect of variable pair patterns, making star glyphs more reliable.

Radial axes glyphs

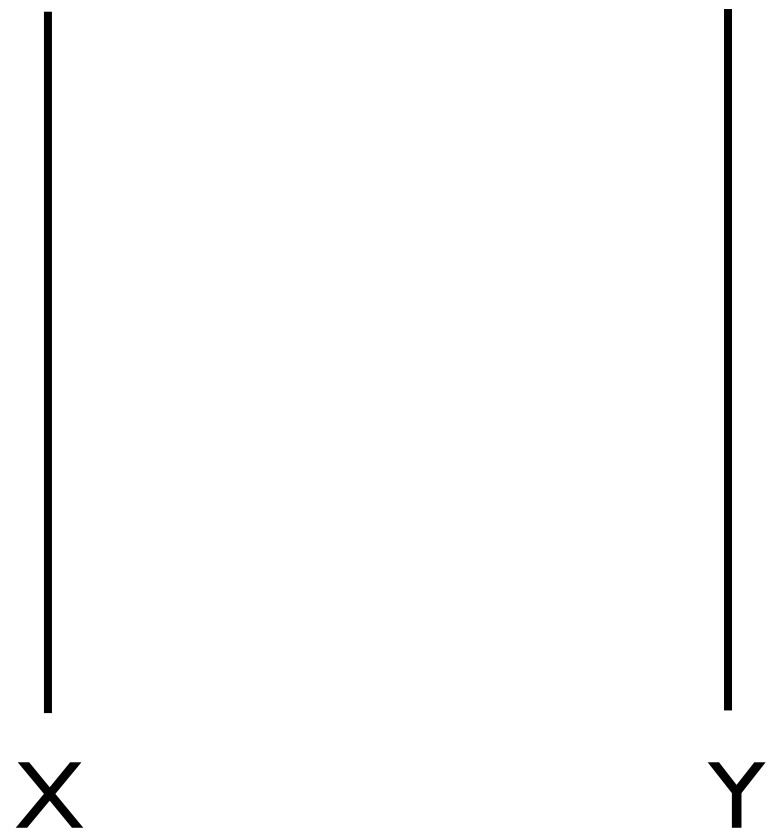
A Greedy Eulerian (maximizing pairwise correlation)



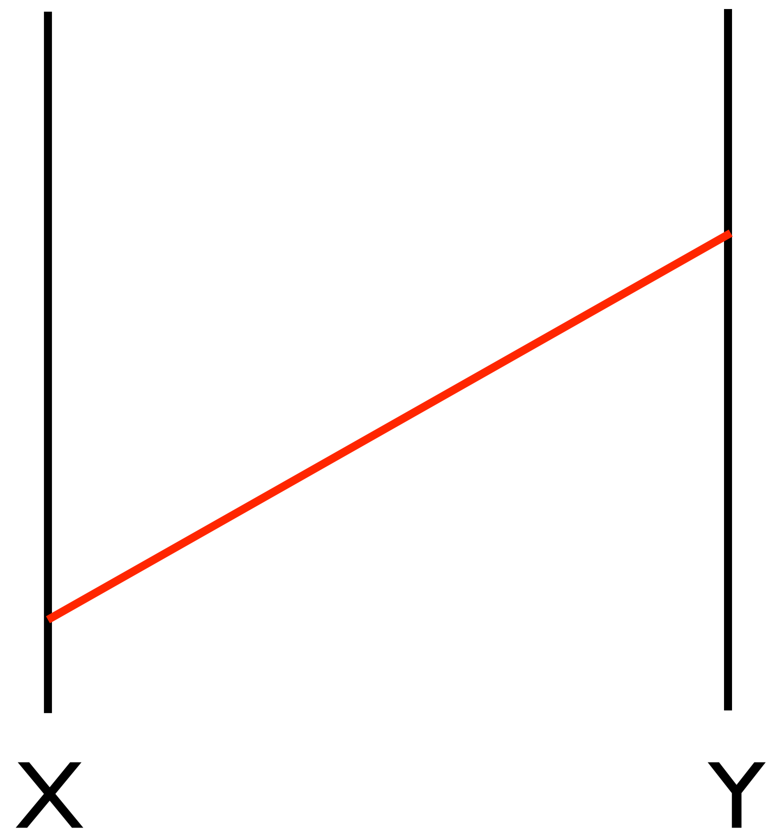


All pairs (Greedy Eulerian or Hamiltonian decomposition) reduce the effect of variable pair patterns, making star glyphs more reliable.

Parallel Axes

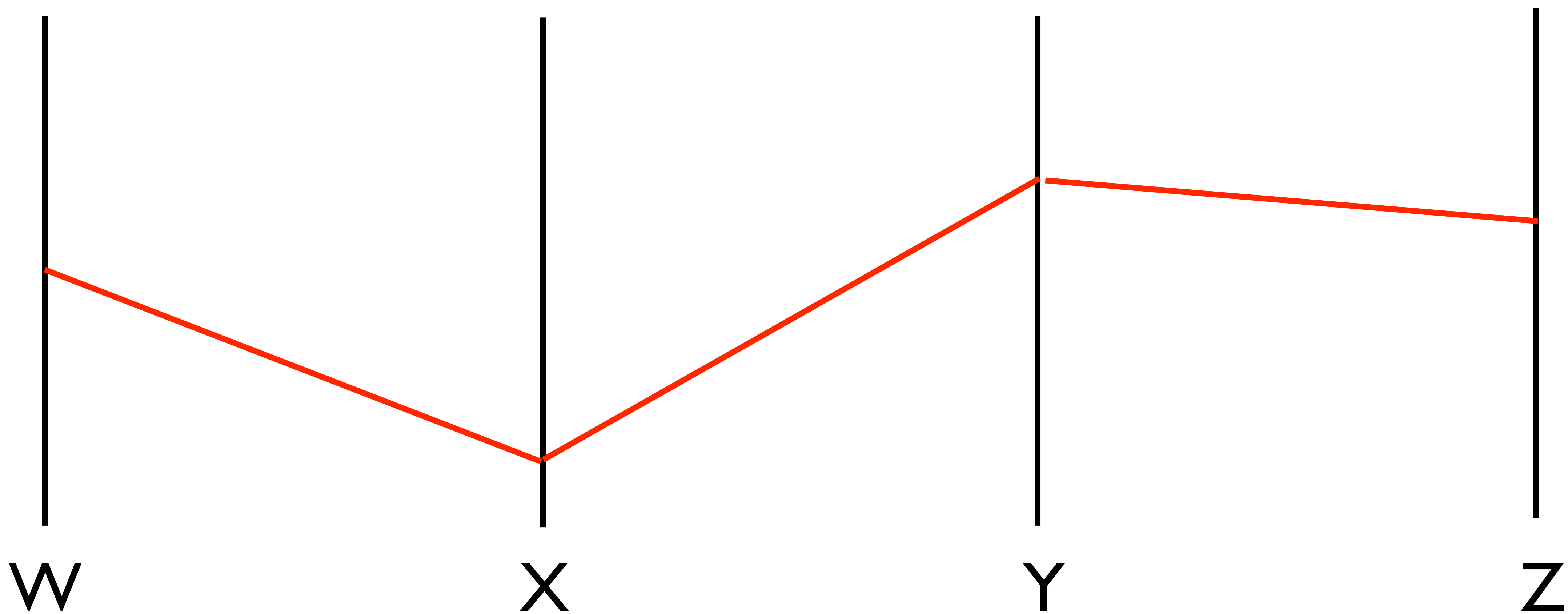


Parallel Axes



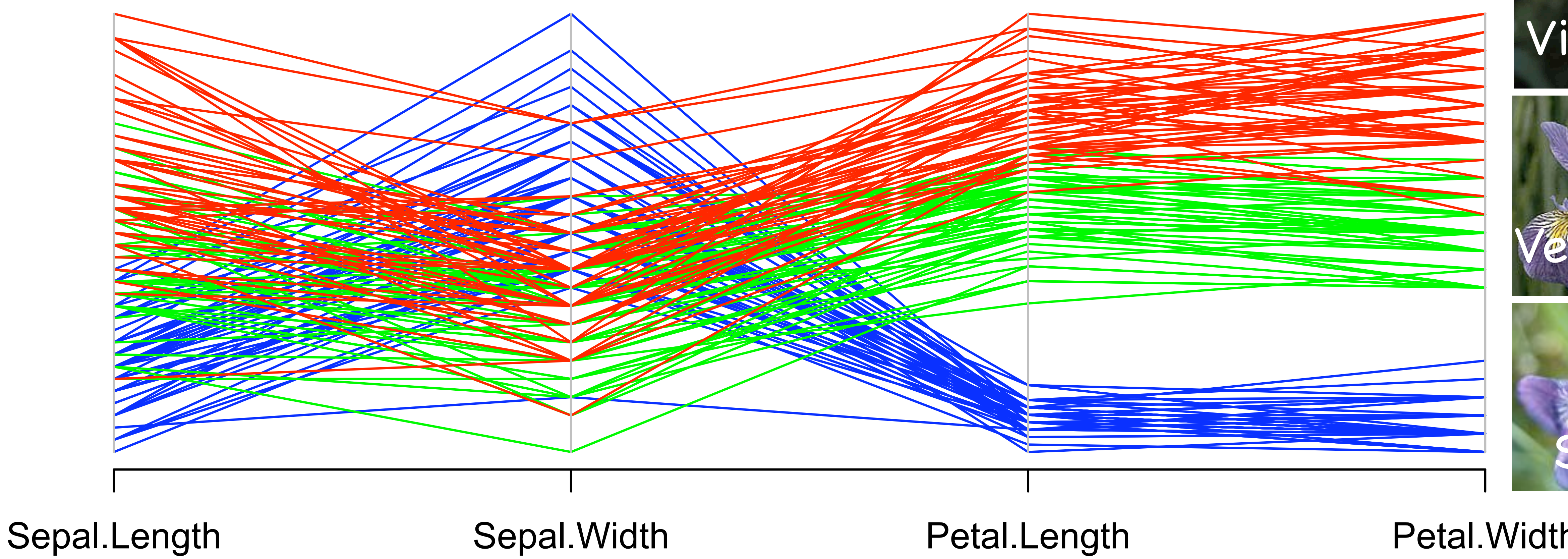
Each point is a "curve" in two dimensions

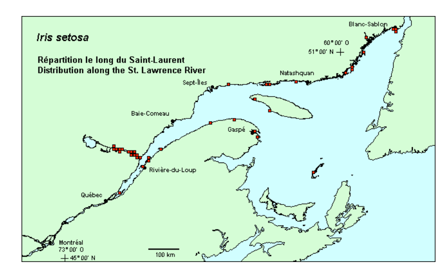
Parallel Coordinates



Each point is a "curve" in two dimensions
Have as many dimensions as the display permits

Parallel Coordinates

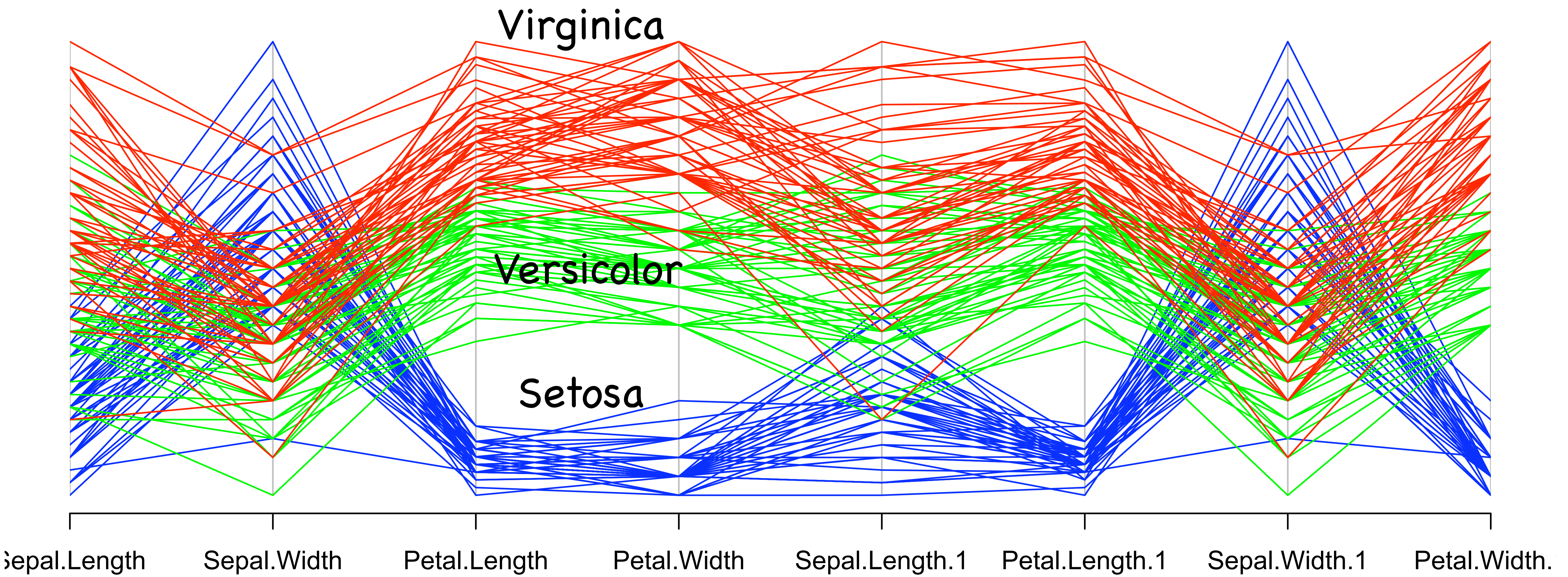




Gaspé Irises:

Each flower is a "curve" in two dimensions positioned by that flower's measurement on all variables.

Parallel Coordinates



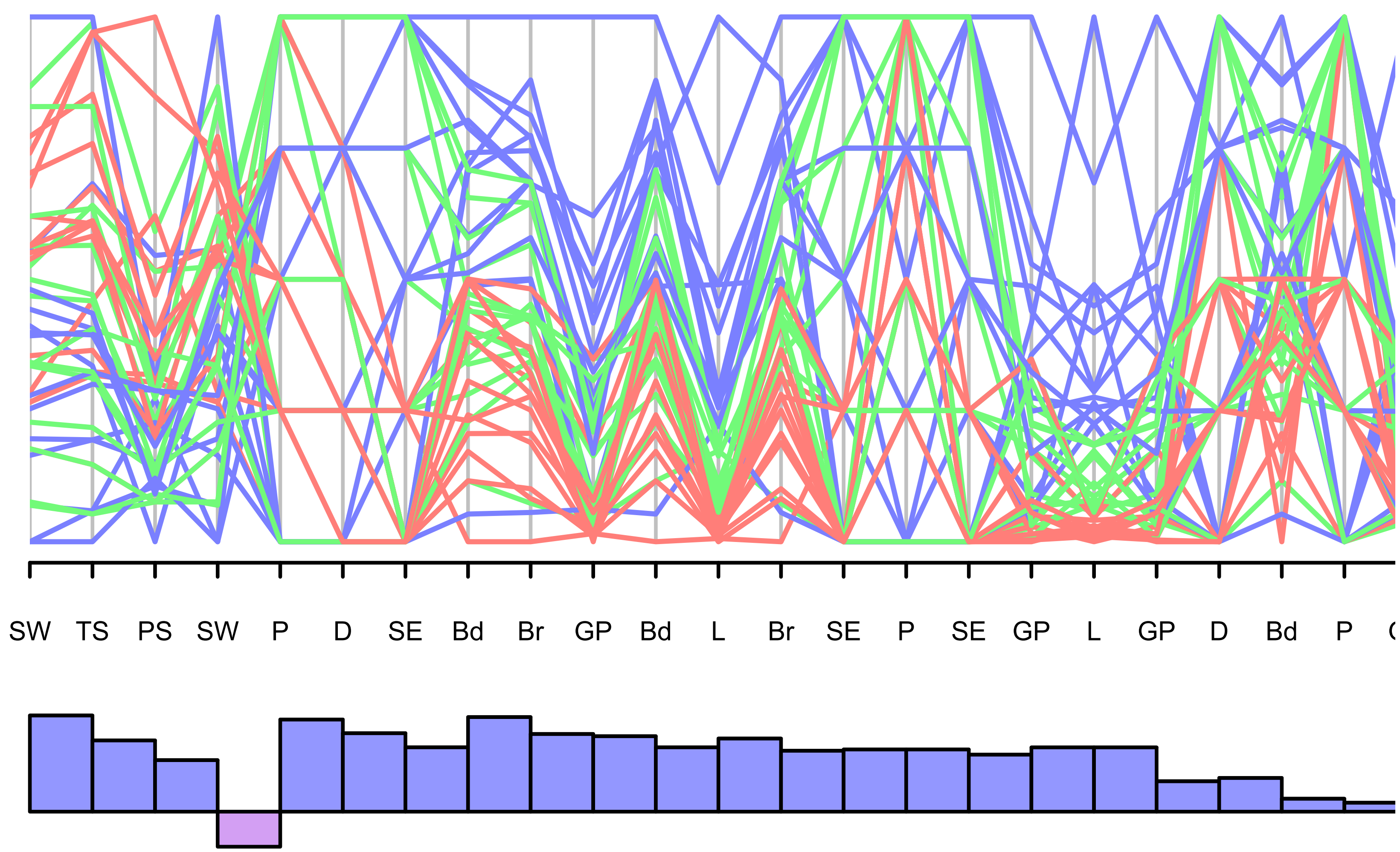
By following an Eulerian Tour on the complete graph, every pair of variables will appear side by side.

Parallel Coordinates

Greedy Eulerian: Focus on most positively correlated

`Sleep1` data (`alr3` R package): 10 variables on 62 mammals

Br = log brain weight,
SW = Slow wave non-dreaming sleep,
TS = Total sleep,
P = Predation index,
SE = Sleep exposure index,
Bd = log body weight,
PS = Paradoxical dreaming sleep,
D = Danger index,
L = Max life span,
GP = Gestation time



Scagnostics

Cognostics (Computer aided diagnostics)

Scagnostics ... Scatterplot cognostics

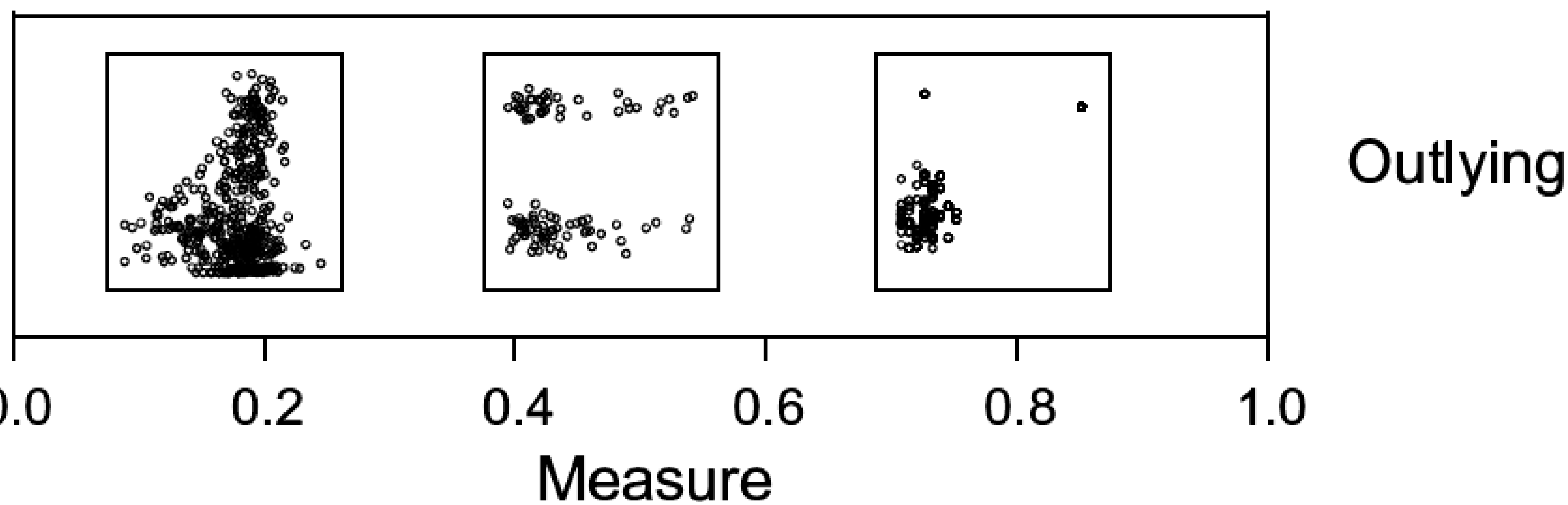
Wilkinson et al (2006) (from idea proposed by Tukey & Tukey (1985))

Scagnostics

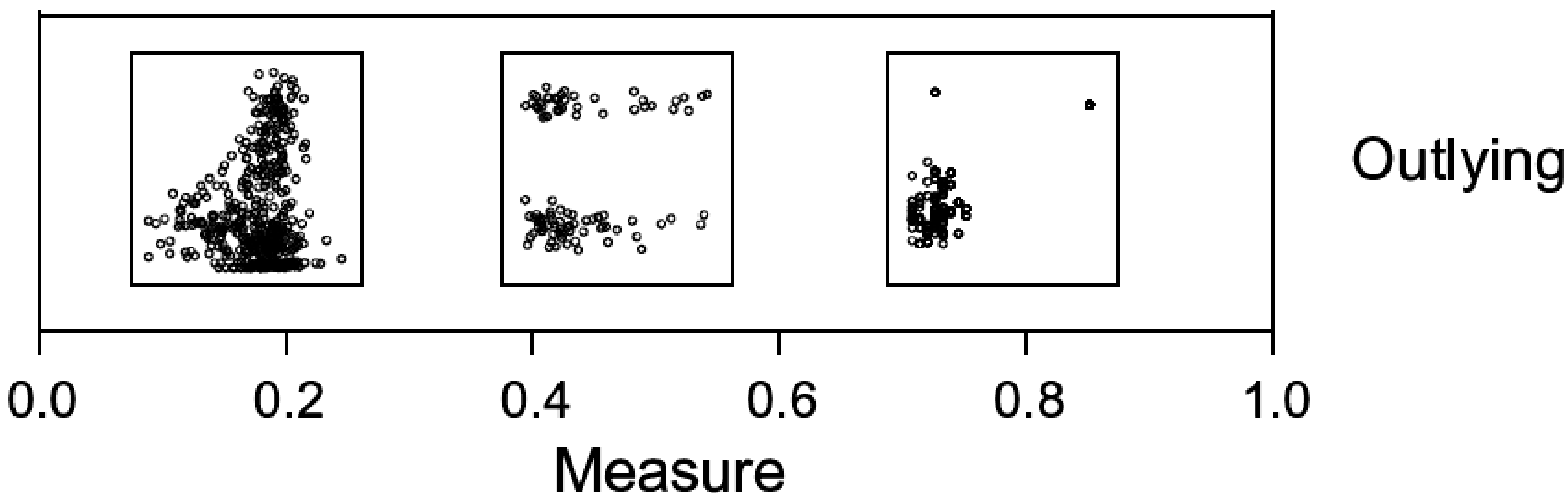
Cognostics (Computer aided diagnostics)

Scagnostics ... Scatterplot cognostics

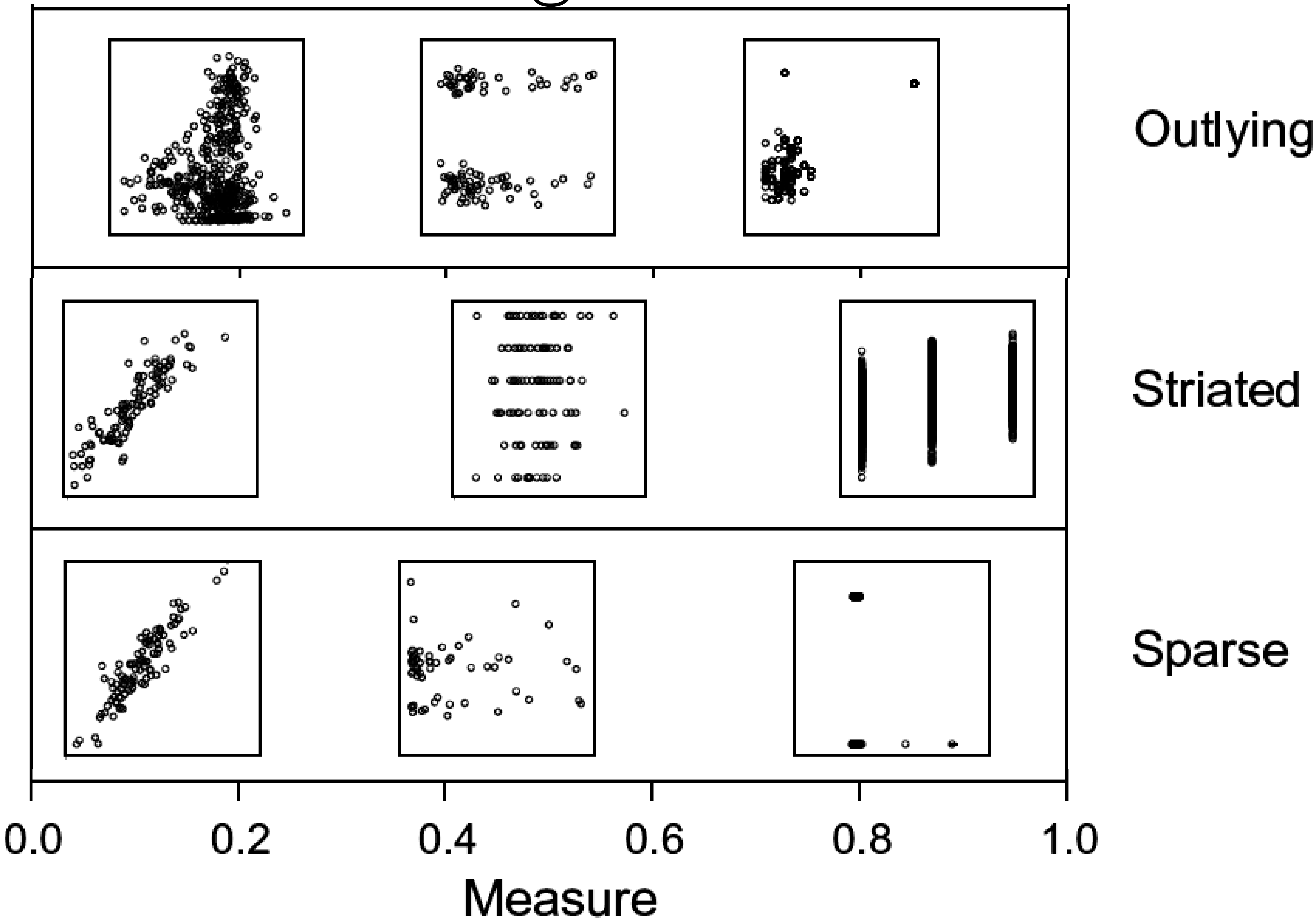
Wilkinson et al (2006) (from idea proposed by Tukey & Tukey (1985))



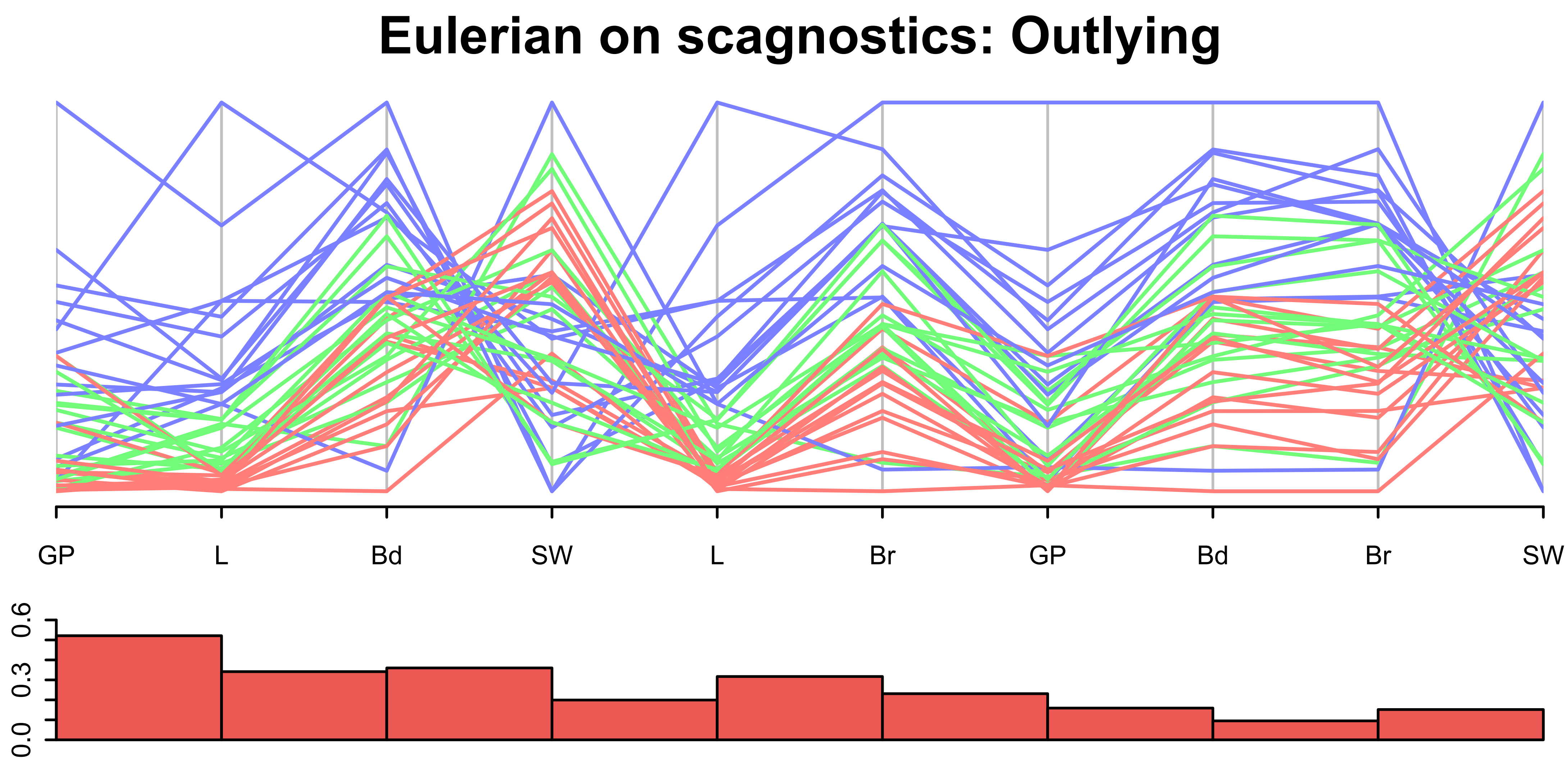
Scagnostics



Scagnostics



Parallel Coordinates

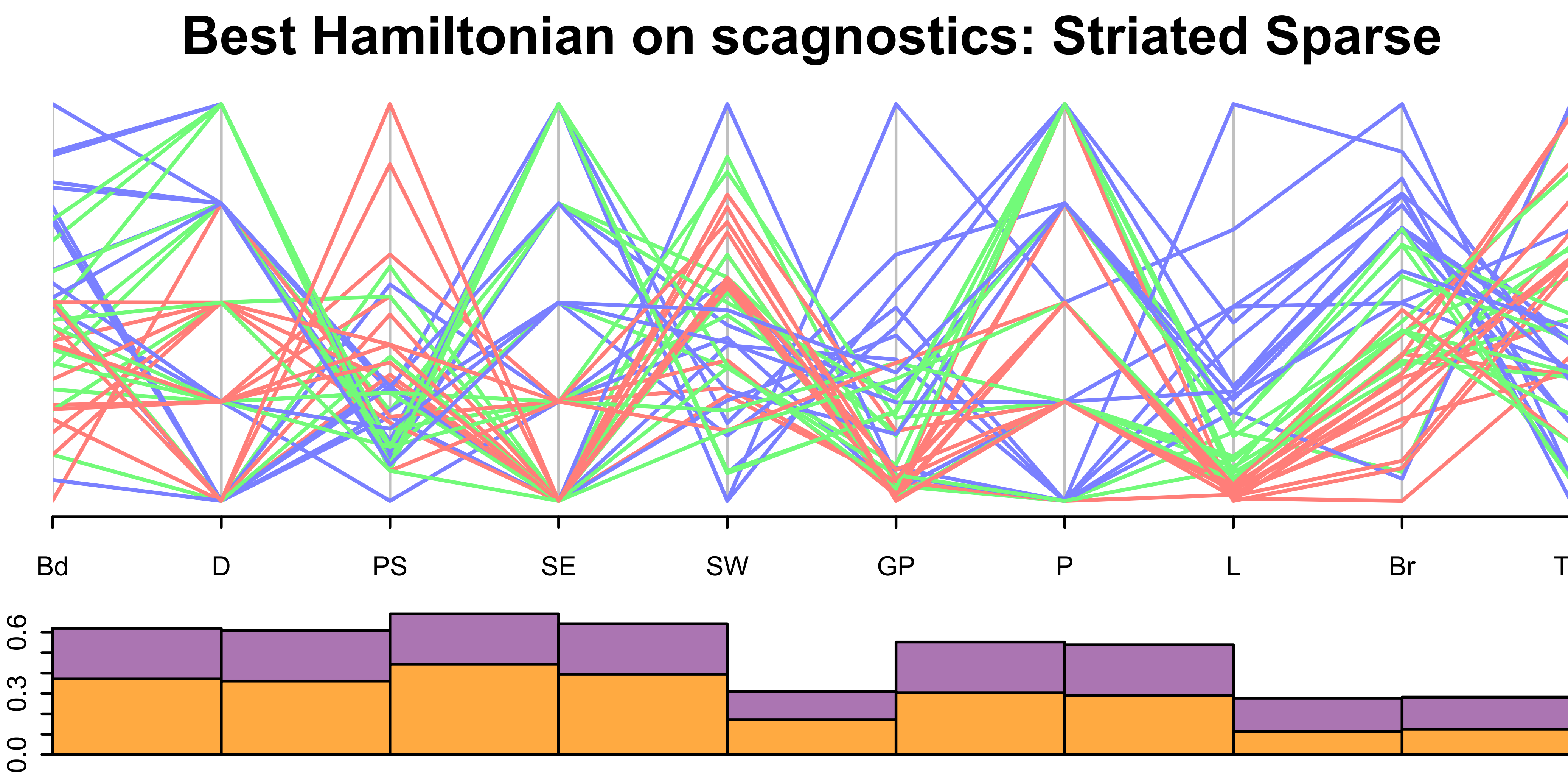


Focus on features of interest (Wilkinson et al 2006,

from Tukey & Tukey 1985 Cognostics ... scagnostics ...)

Eulerian may pick up a subset of the variables

Parallel Coordinates



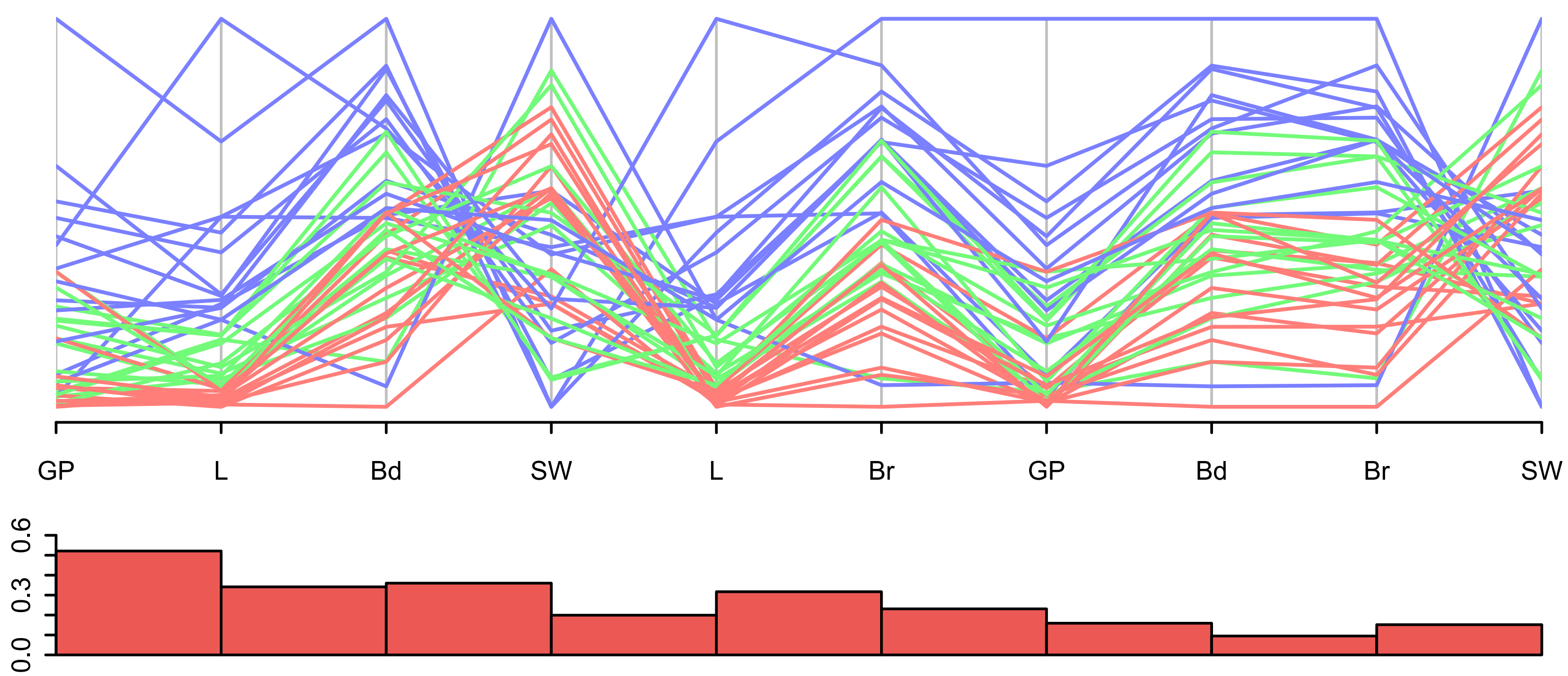
Focus on features of interest (Wilkinson et al 2006,

from Tukey & Tukey 1985 Cognostics ... scagnostics ...)

Hamiltonian ensures all variables appear

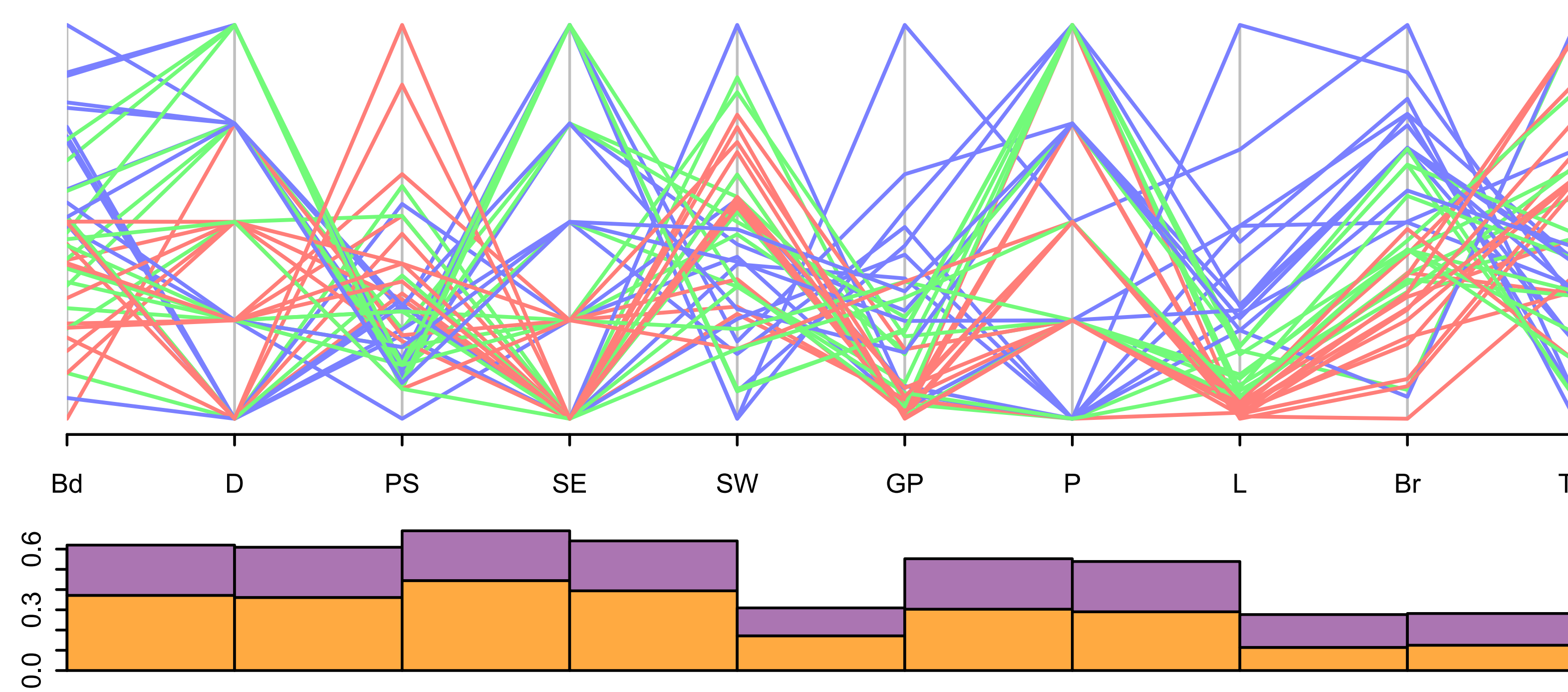
Parallel Coordinates

Eulerian on scagnostics: Outlying



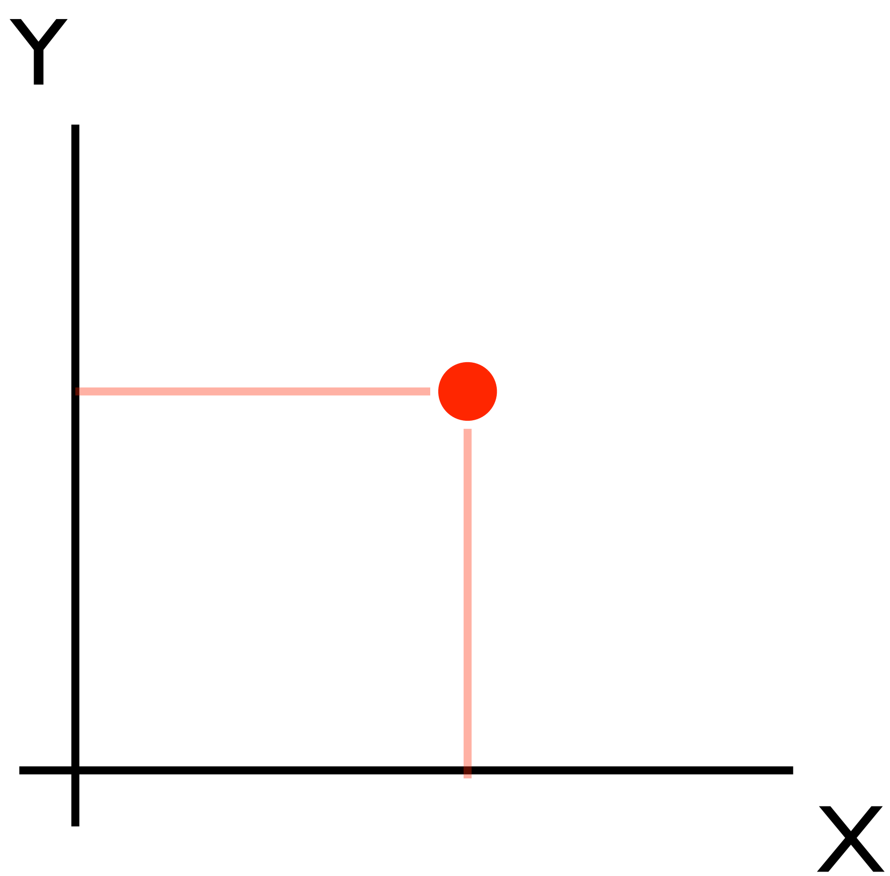
Same data

Best Hamiltonian on scagnostics: Striated Sparse

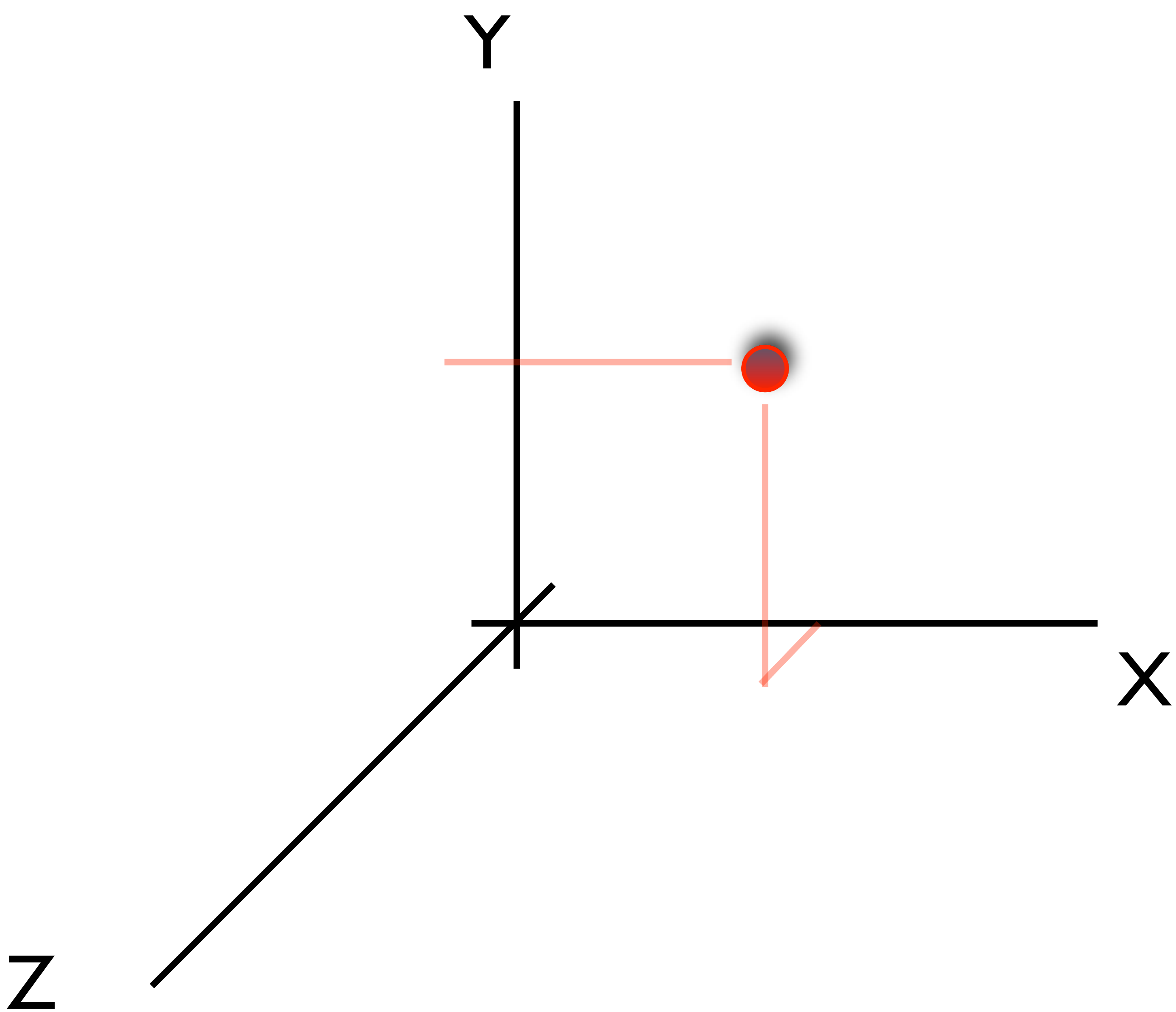


Different measures

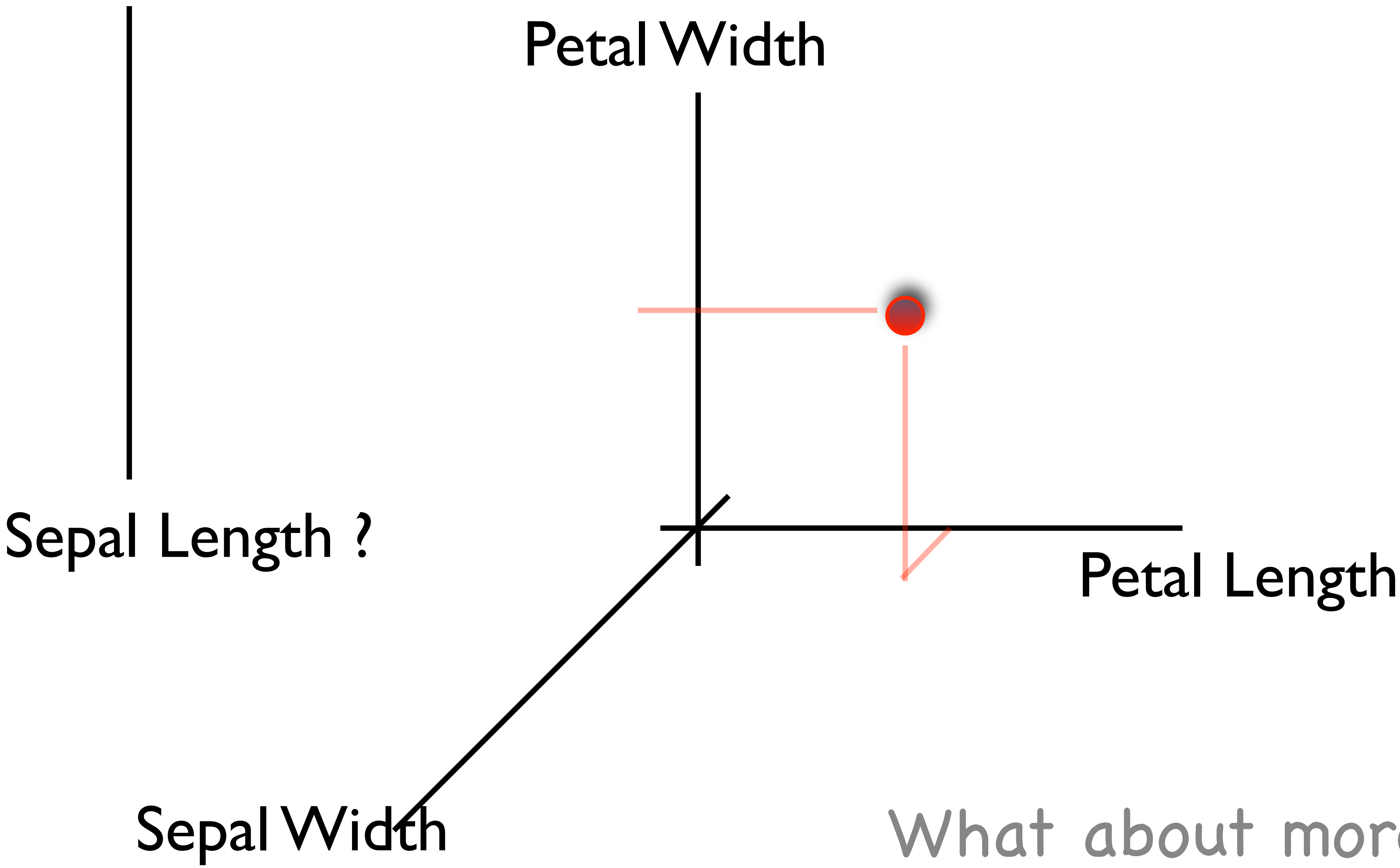
Orthogonal axes



Orthogonal axes



Orthogonal axes



What about more than 3 dimensions?

Example: Italian olive oils

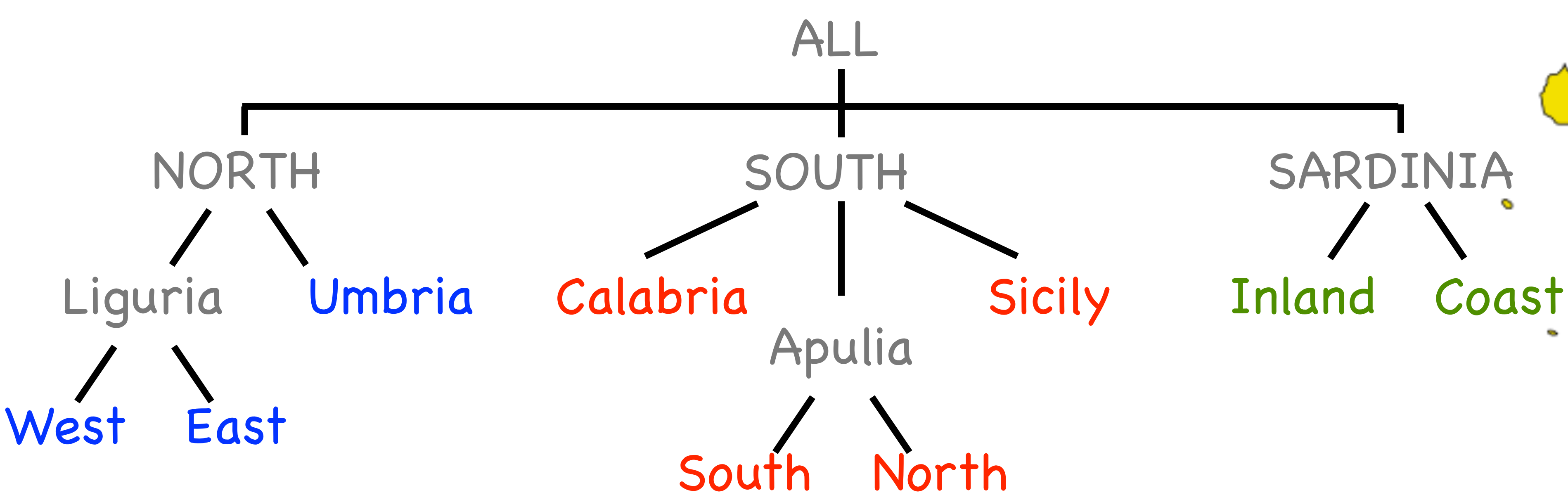
Different regions of Italy:

- NORTH (Umbria, East-Liguria, West-Liguria)
- SOUTH (Calabria, Sicily, North-Apulia, South-Apulia)
- SARDINIA (Inland, Coast)

Example: Italian olive oils

Different regions of Italy:

- NORTH (Umbria, East-Liguria, West-Liguria)
- SOUTH (Calabria, Sicily, North-Apulia, South-Apulia)
- SARDINIA (Inland, Coast)



Example: Italian olive oils

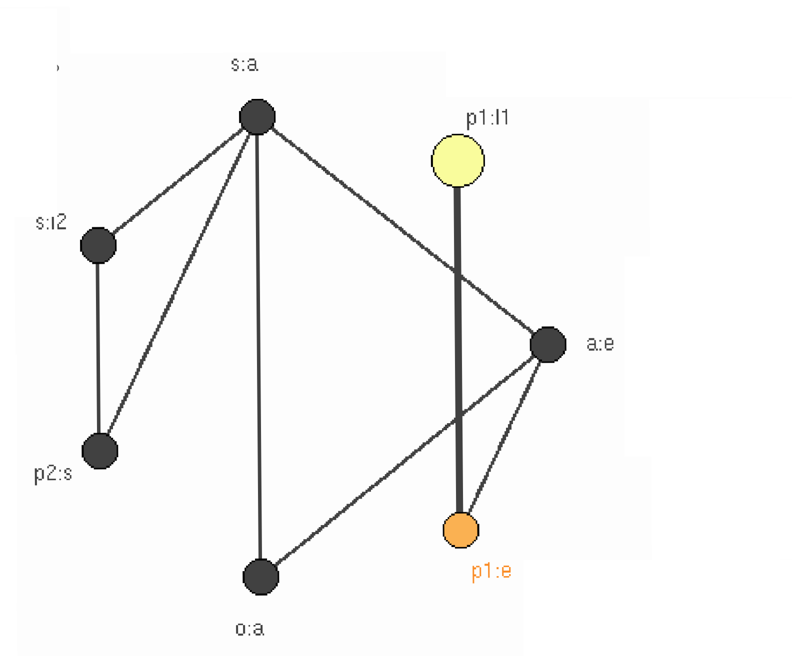
Measurements:

- n = 572 different olive samples
- concentrations of p=8 fatty acids:
    - arachidic, eicosenoic, linoleic (l1), linolenic (l2), oleic, palmitic (p1), palmitoleic (p2), and stearic.

Navigation Graphs

Connecting low-d spaces

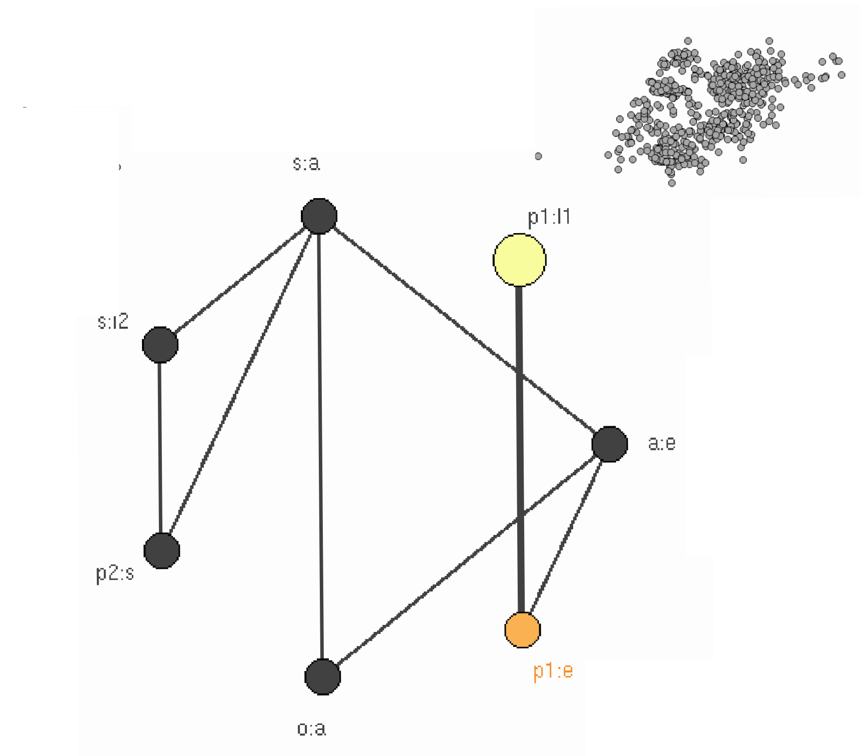
- node = variable pair
- edges connect nodes that share a variable
- could display scatterplot at each node
- edges are 3D transitions
- high dimensional space is explored by moving from one 2D space to another through 3D (or 4D) transitions
- track/map exploration
- suggest routes



Navigation Graphs

Connecting low-d spaces

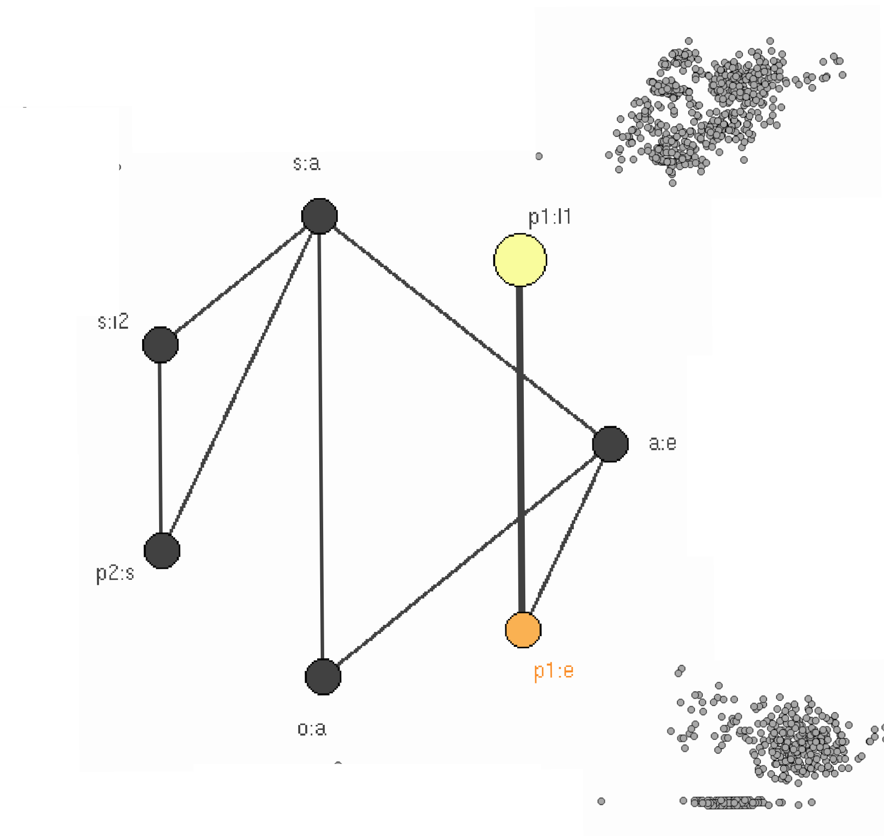
- node = variable pair
- edges connect nodes that share a variable
- could display scatterplot at each node
- edges are 3D transitions
- high dimensional space is explored by moving from one 2D space to another through 3D (or 4D) transitions
- track/map exploration
- suggest routes



Navigation Graphs

Connecting low-d spaces

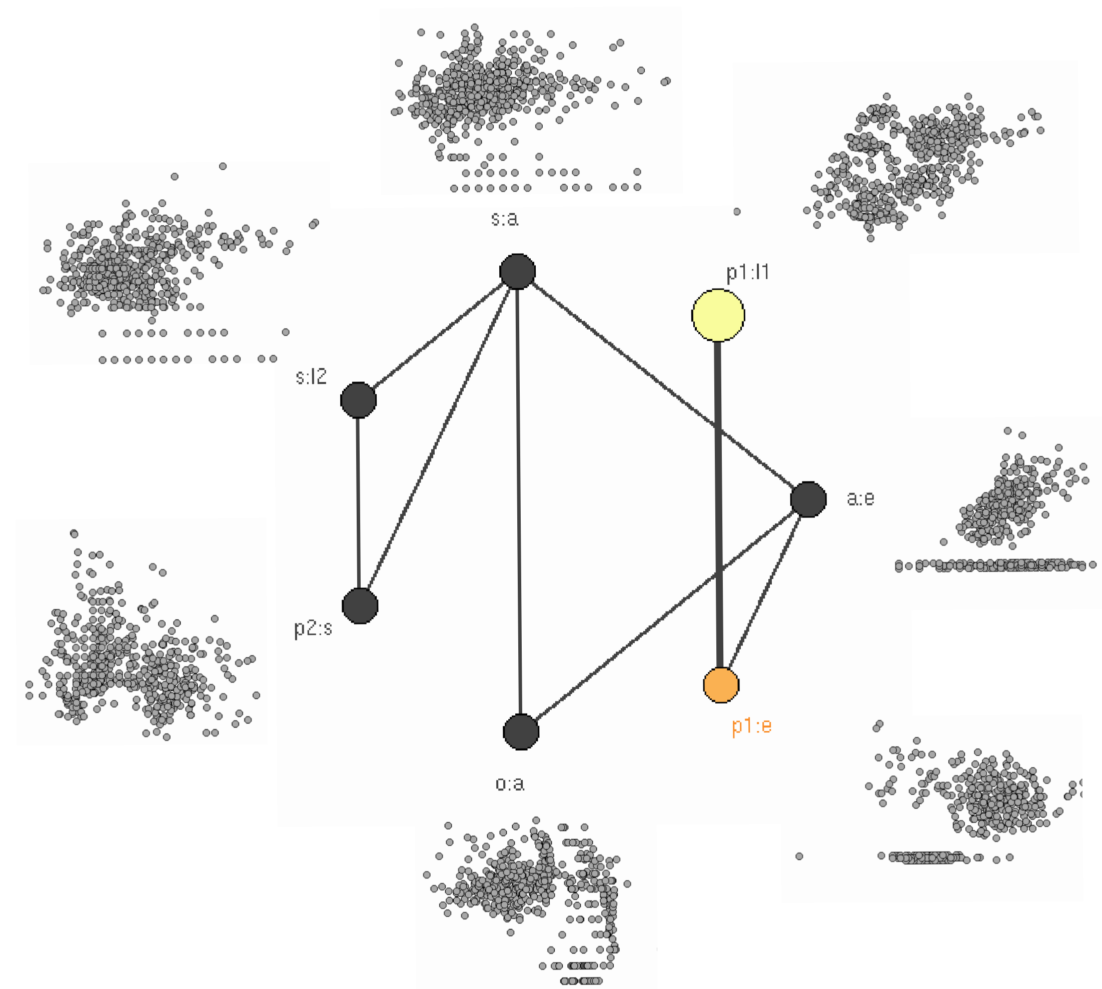
- node = variable pair
- edges connect nodes that share a variable
- could display scatterplot at each node
- edges are 3D transitions
- high dimensional space is explored by moving from one 2D space to another through 3D (or 4D) transitions
- track/map exploration
- suggest routes



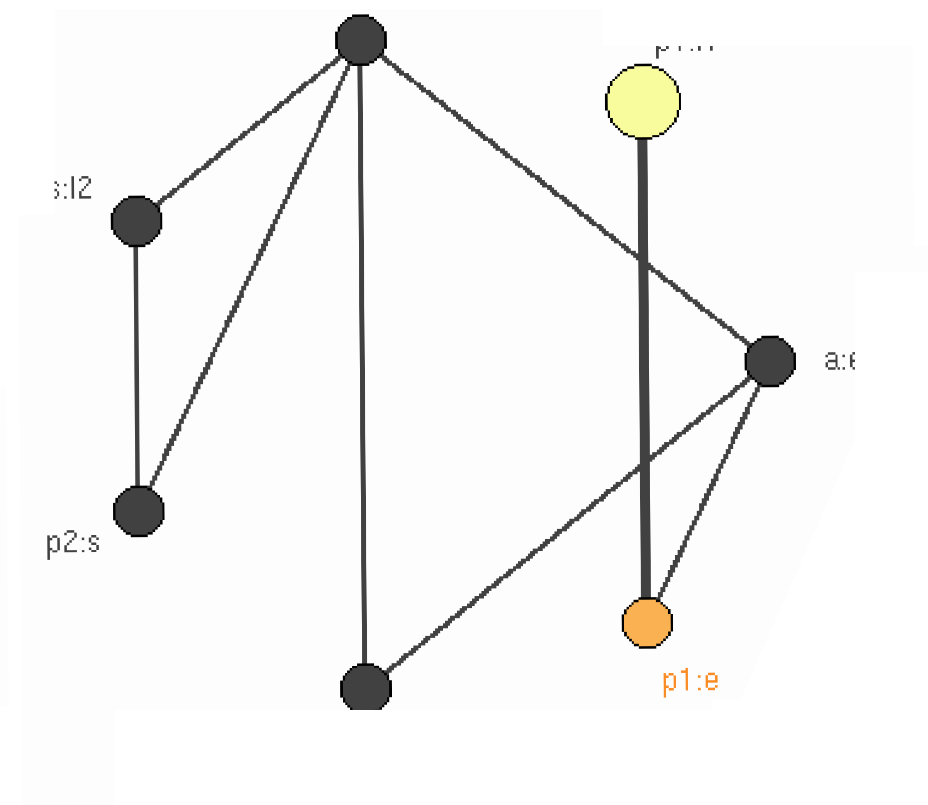
Navigation Graphs

Connecting low-d spaces

- node = variable pair
- edges connect nodes that share a variable
- could display scatterplot at each node
- edges are 3D transitions
- high dimensional space is explored by moving from one 2D space to another through 3D (or 4D) transitions
- track/map exploration
- suggest routes



Navigation Graphs



RNavgraph

... R implementation

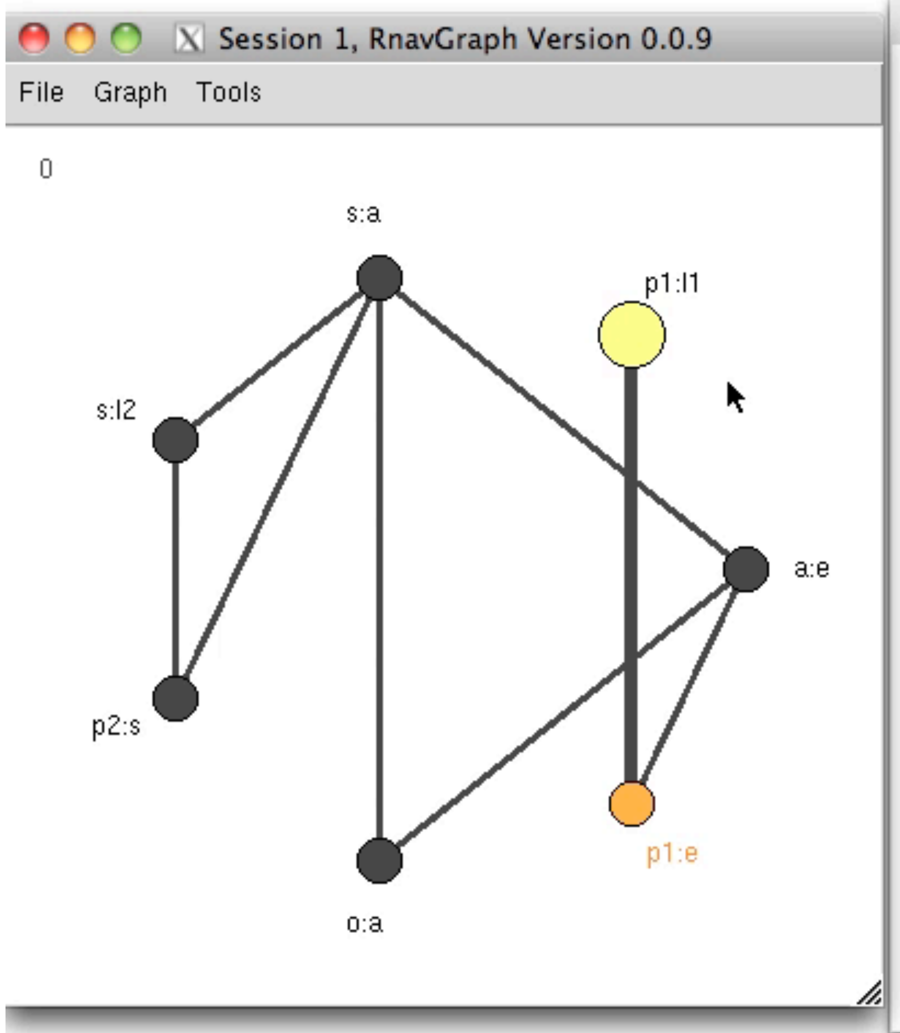
Example: Italian olive oils

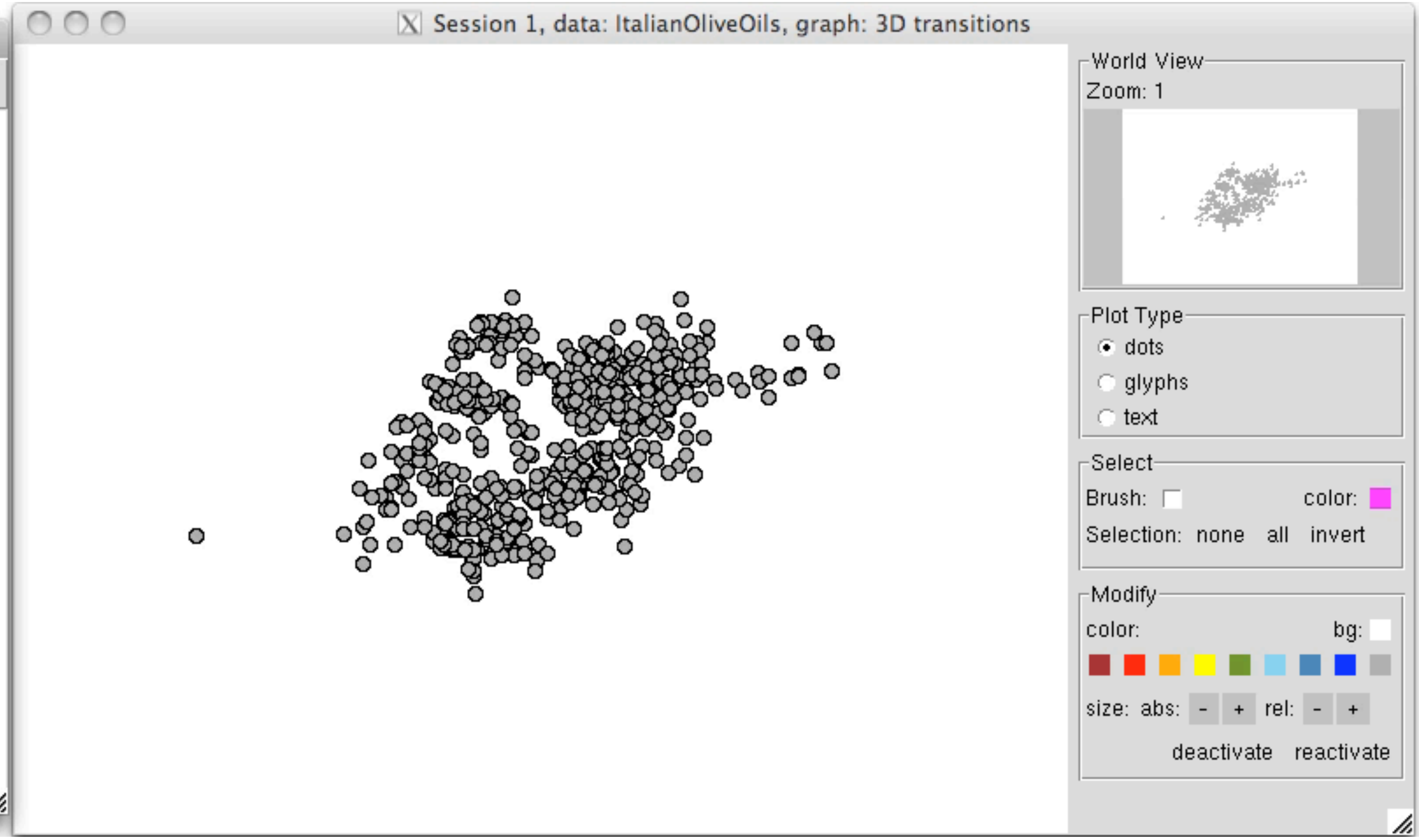
Interactive

3d transition graph

Interactive scatterplot

Example: Italian olive oils





Interactive

3d transition graph

Interactive scatterplot

Example: Italian olive oils

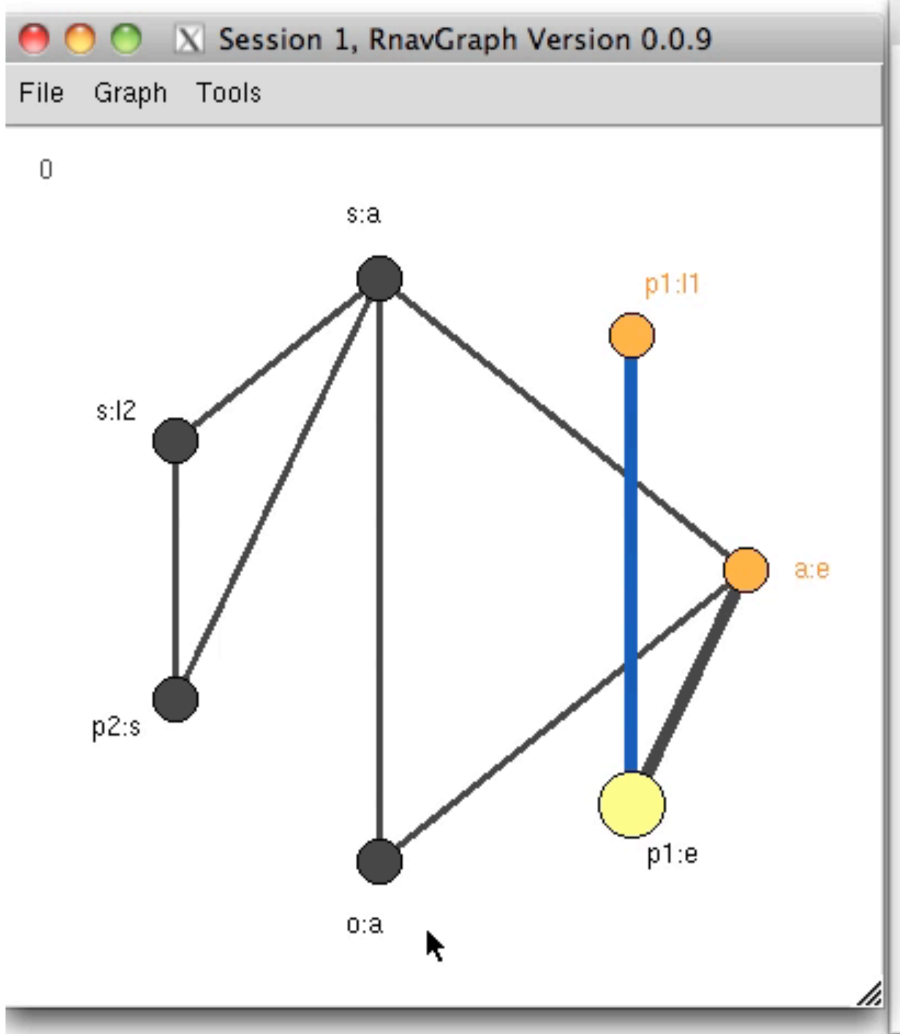
Interactive

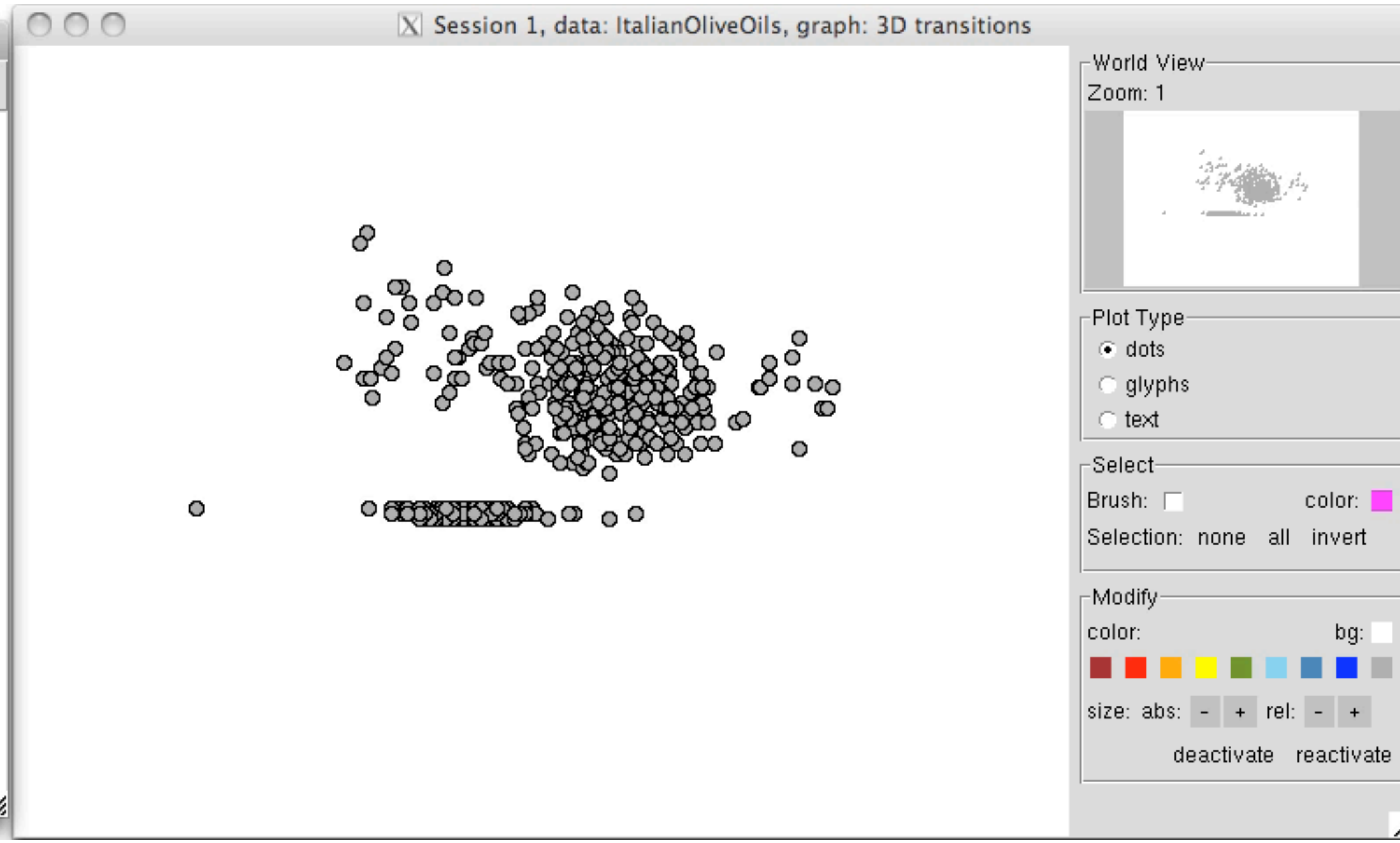
3d transition graph

Interactive scatterplot

Move back and forth by hand

Example: Italian olive oils





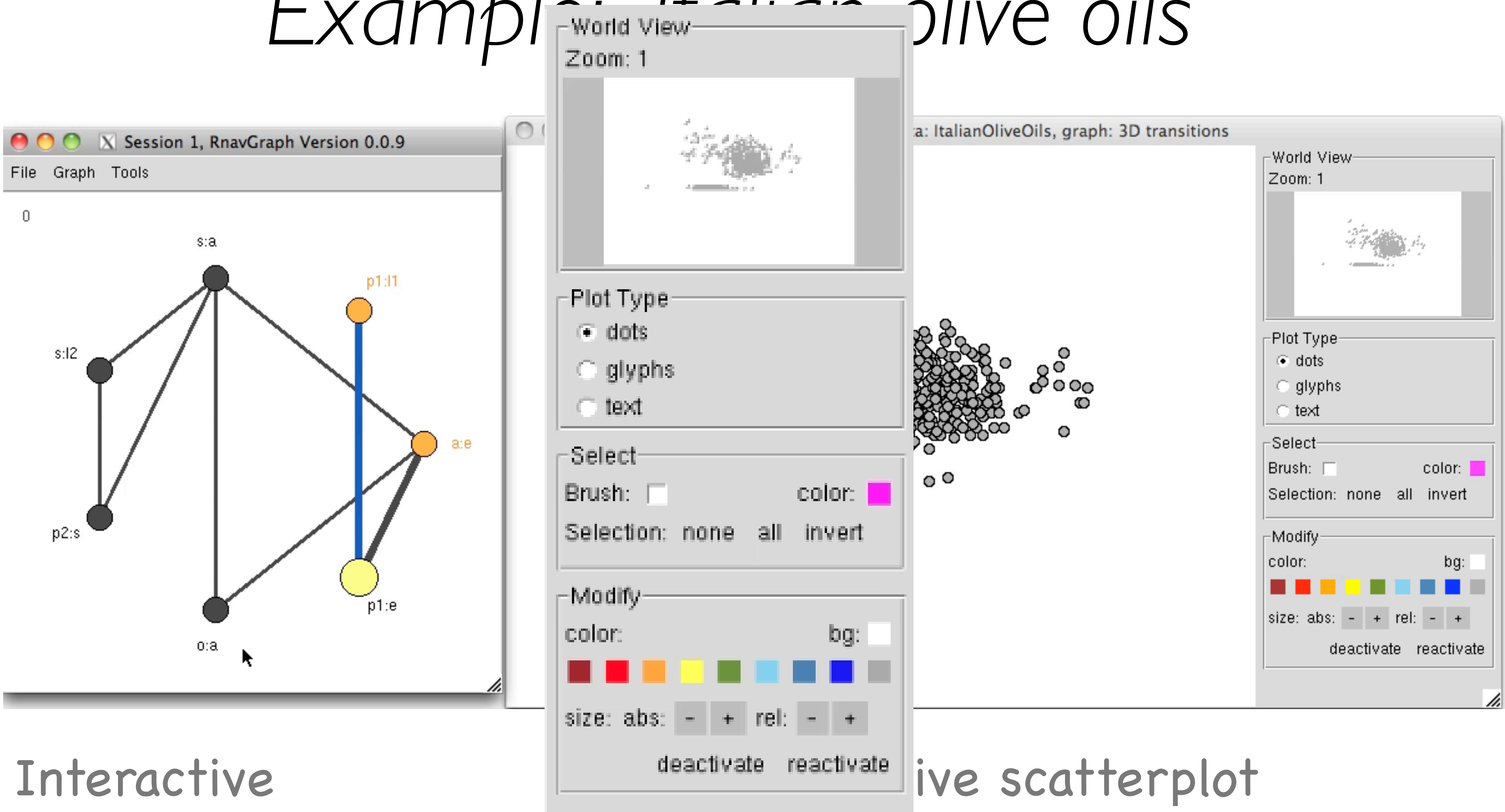
Interactive

3d transition graph

Interactive scatterplot

Move back and forth by hand

Example: Italian olive oils



Interactive

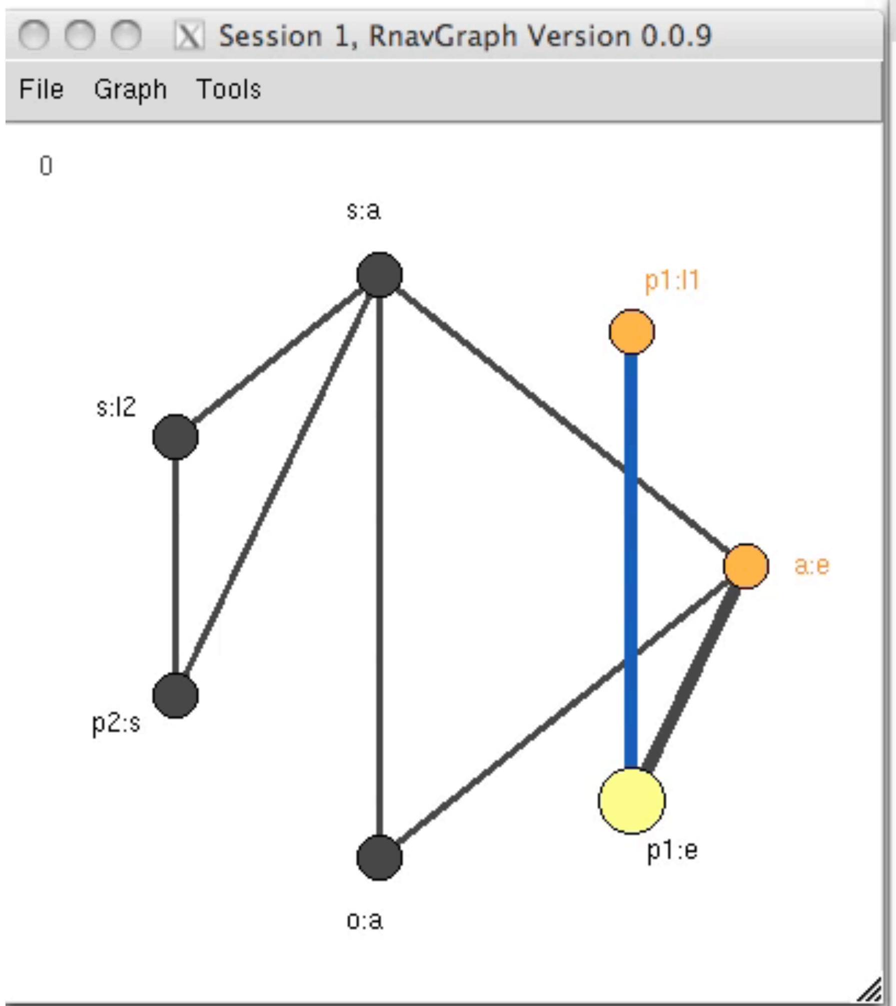
3d transition graph

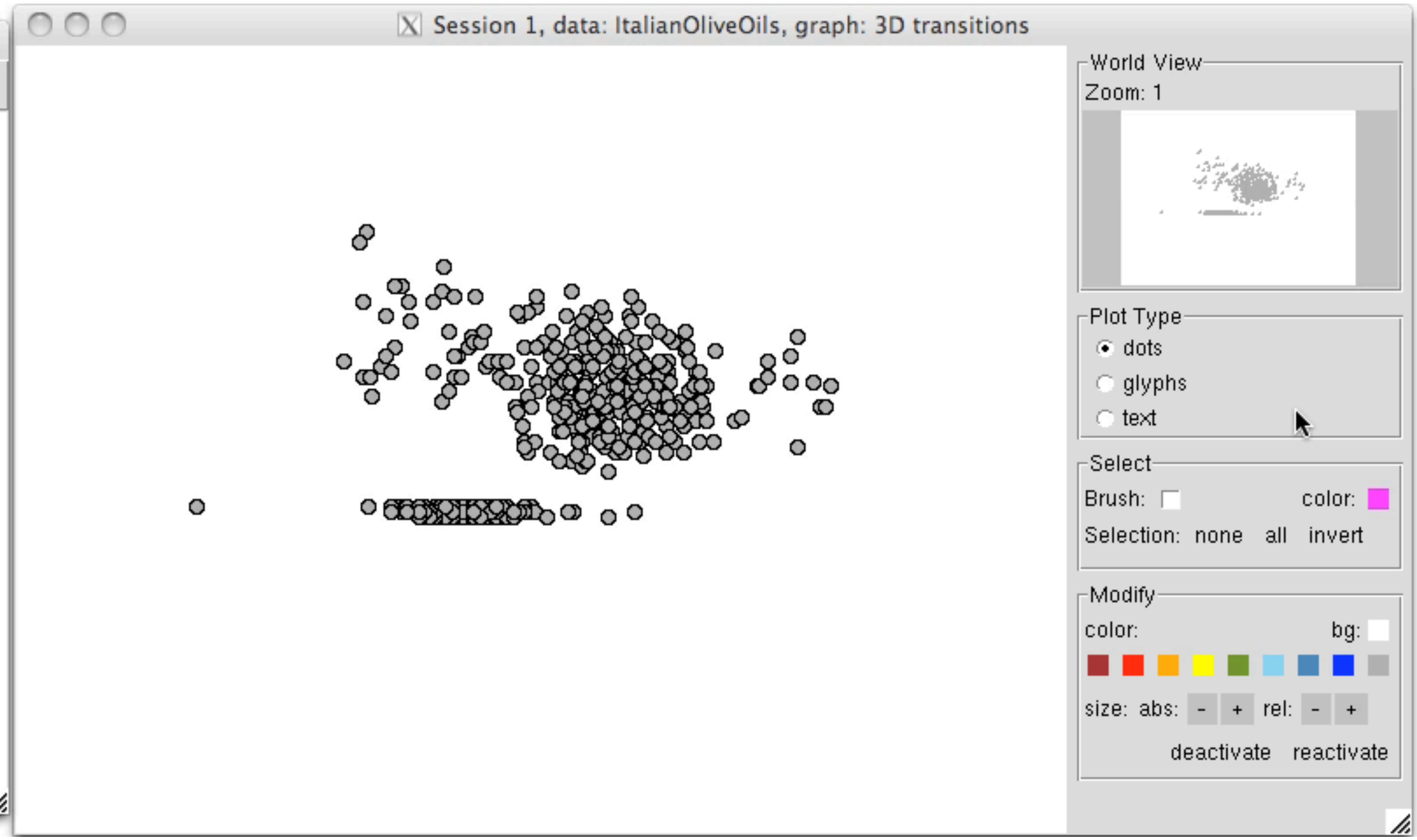
ive scatterplot

Scatterplot control panel

offers interactive features

Example: Italian olive oils





Interactive

3d transition graph

Interactive scatterplot

Brushing

Example: Italian olive oils

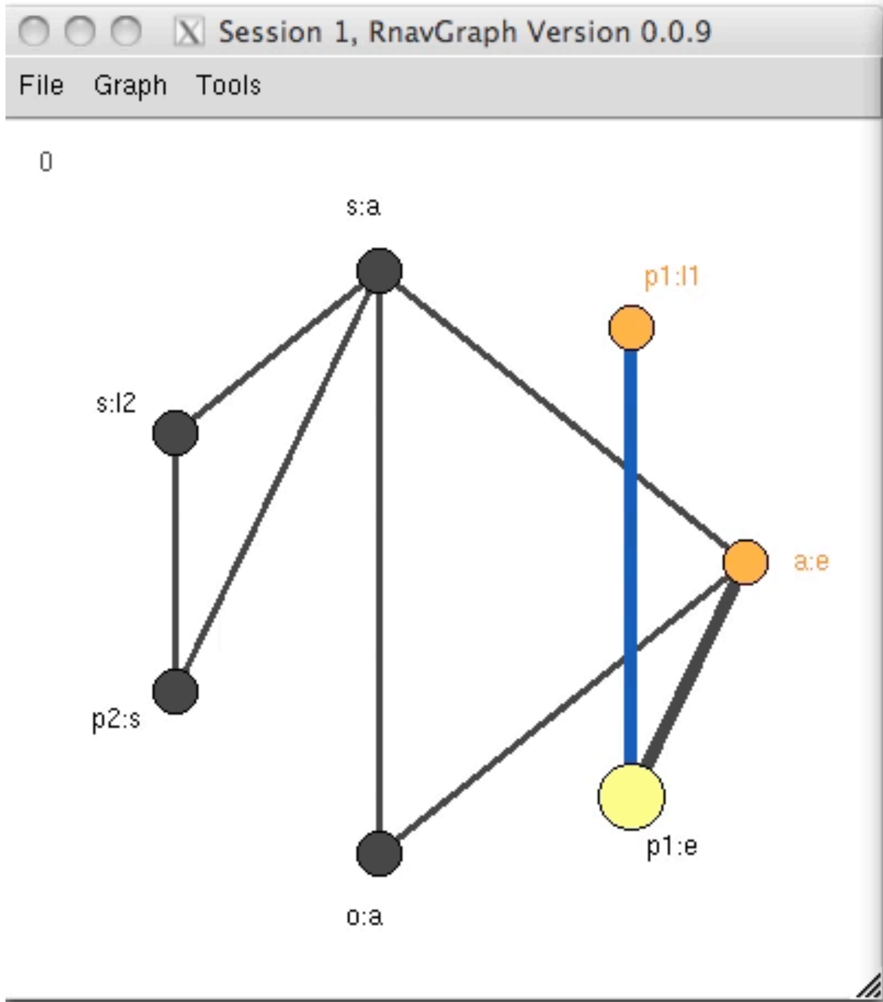
Interactive

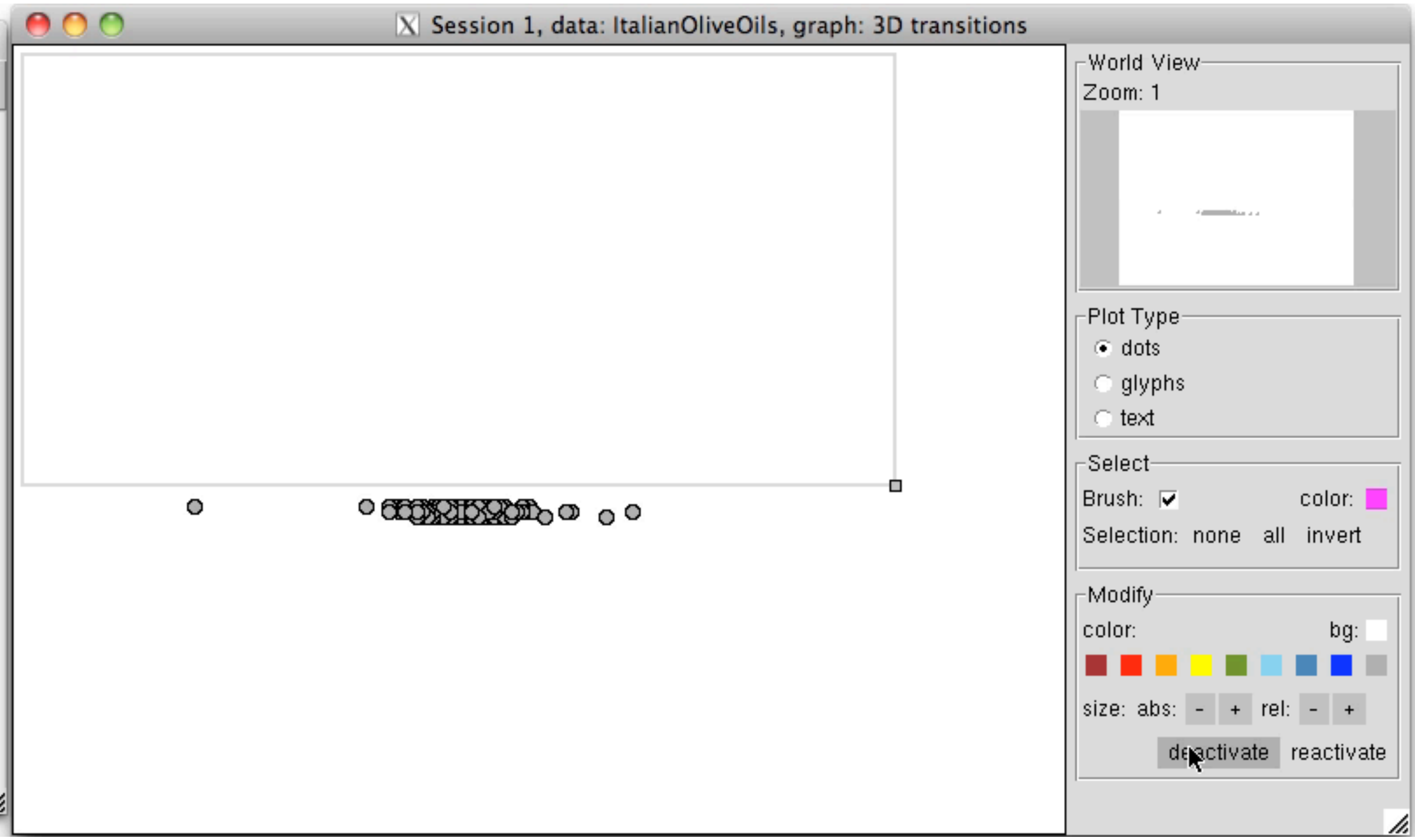
3d transition graph

Interactive scatterplot

Deactivate selected points

Example: Italian olive oils





Interactive

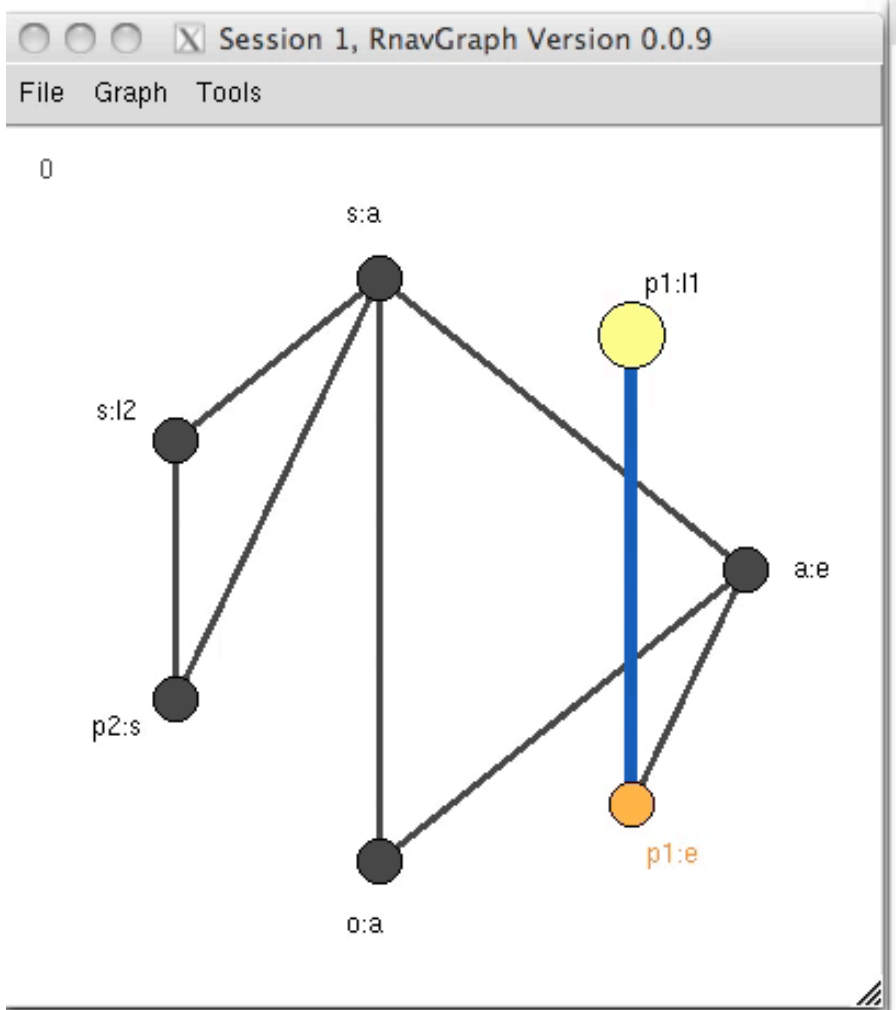
3d transition graph

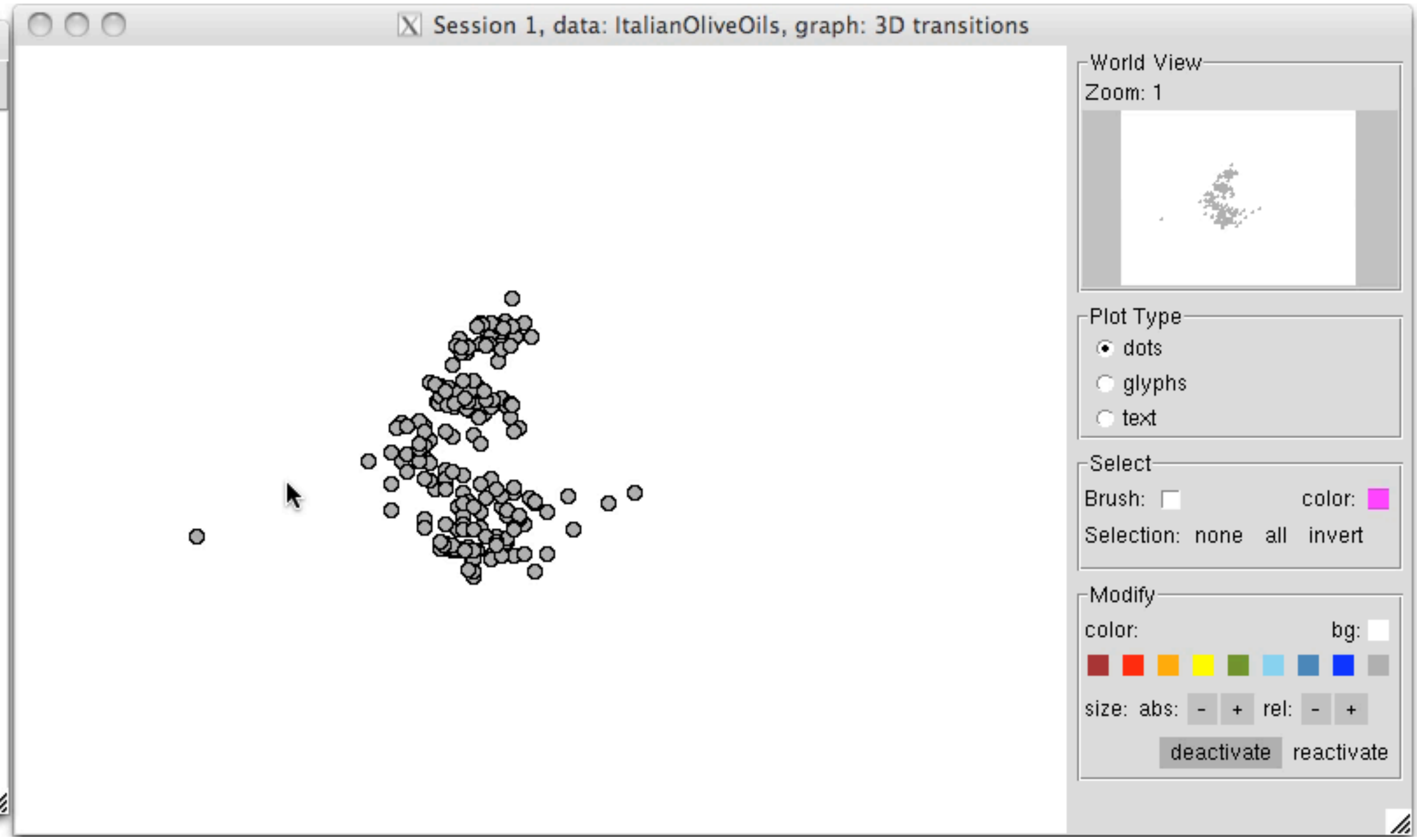
Interactive scatterplot

Deactivate selected points

Return to starting position

Example: Italian olive oils





Interactive

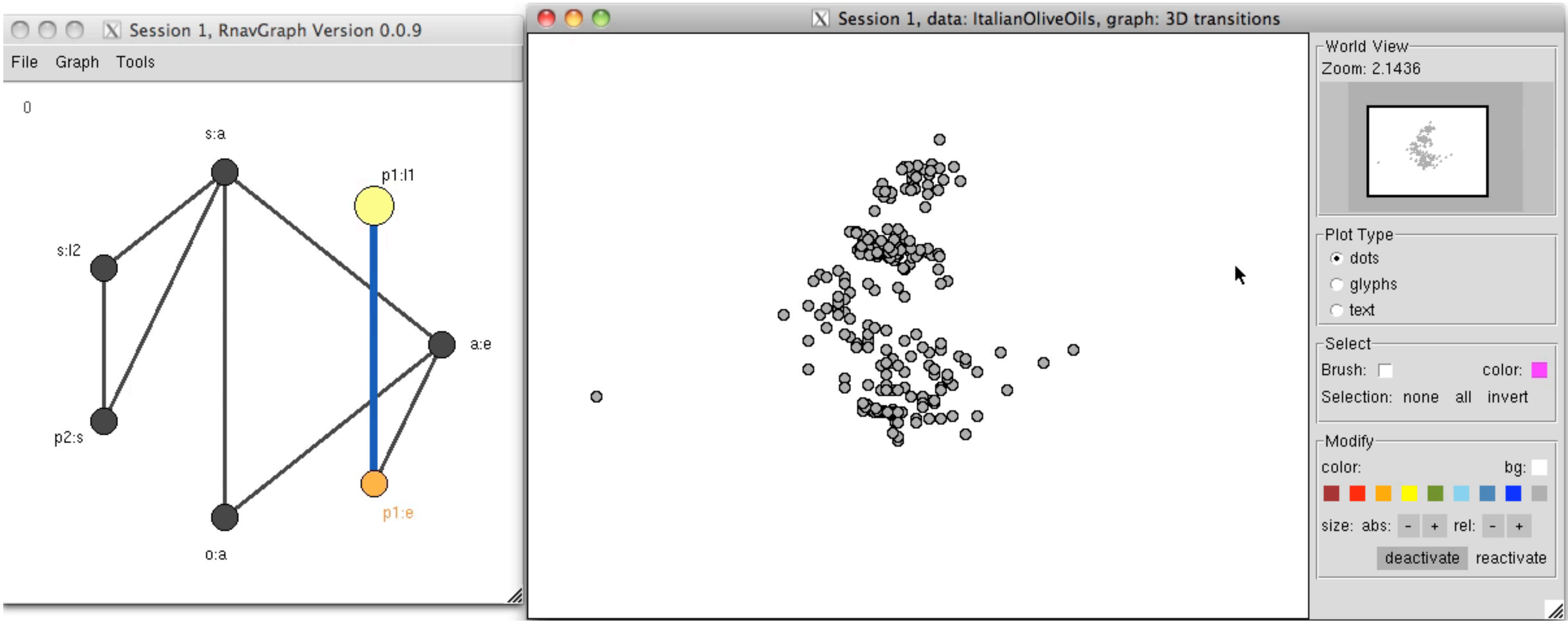
3d transition graph

Interactive scatterplot

Zoom and relocate

Note “World View” changes

Example: Italian olive oils



Interactive

3d transition graph

Interactive scatterplot

At least 3 groups;

Colour two of them.

Example: Italian olive oils

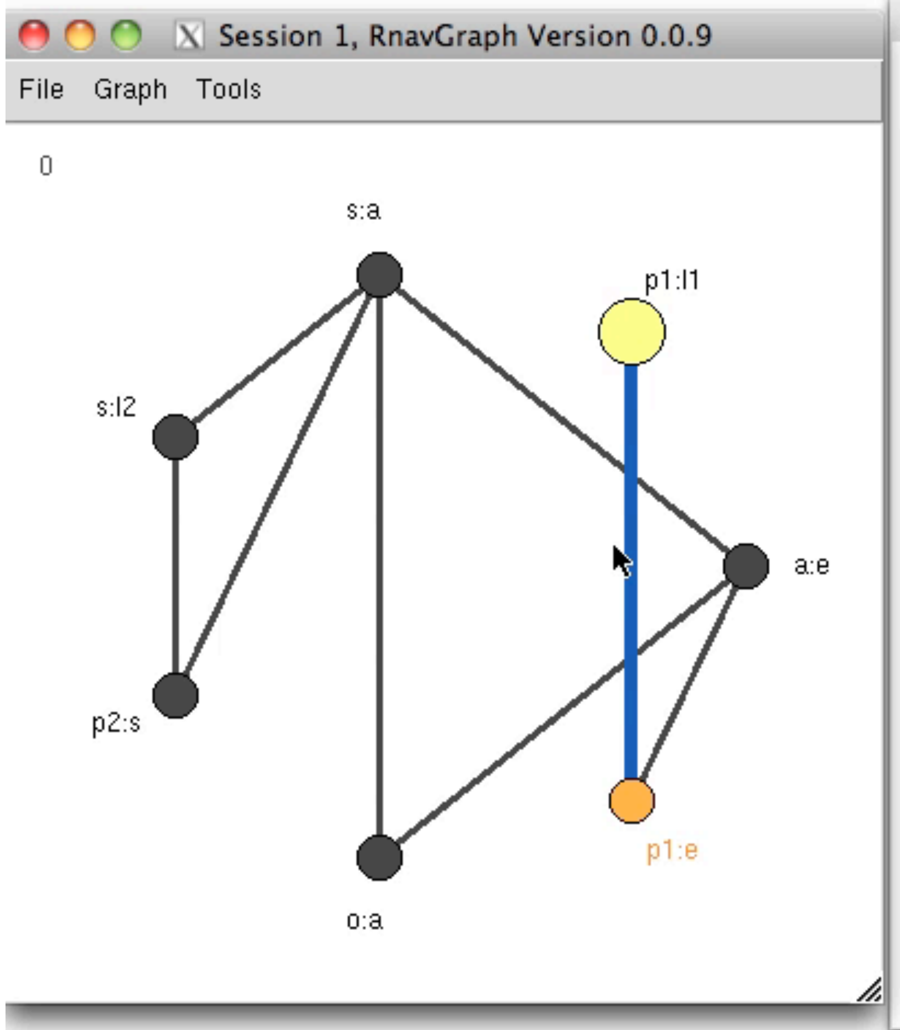
Interactive

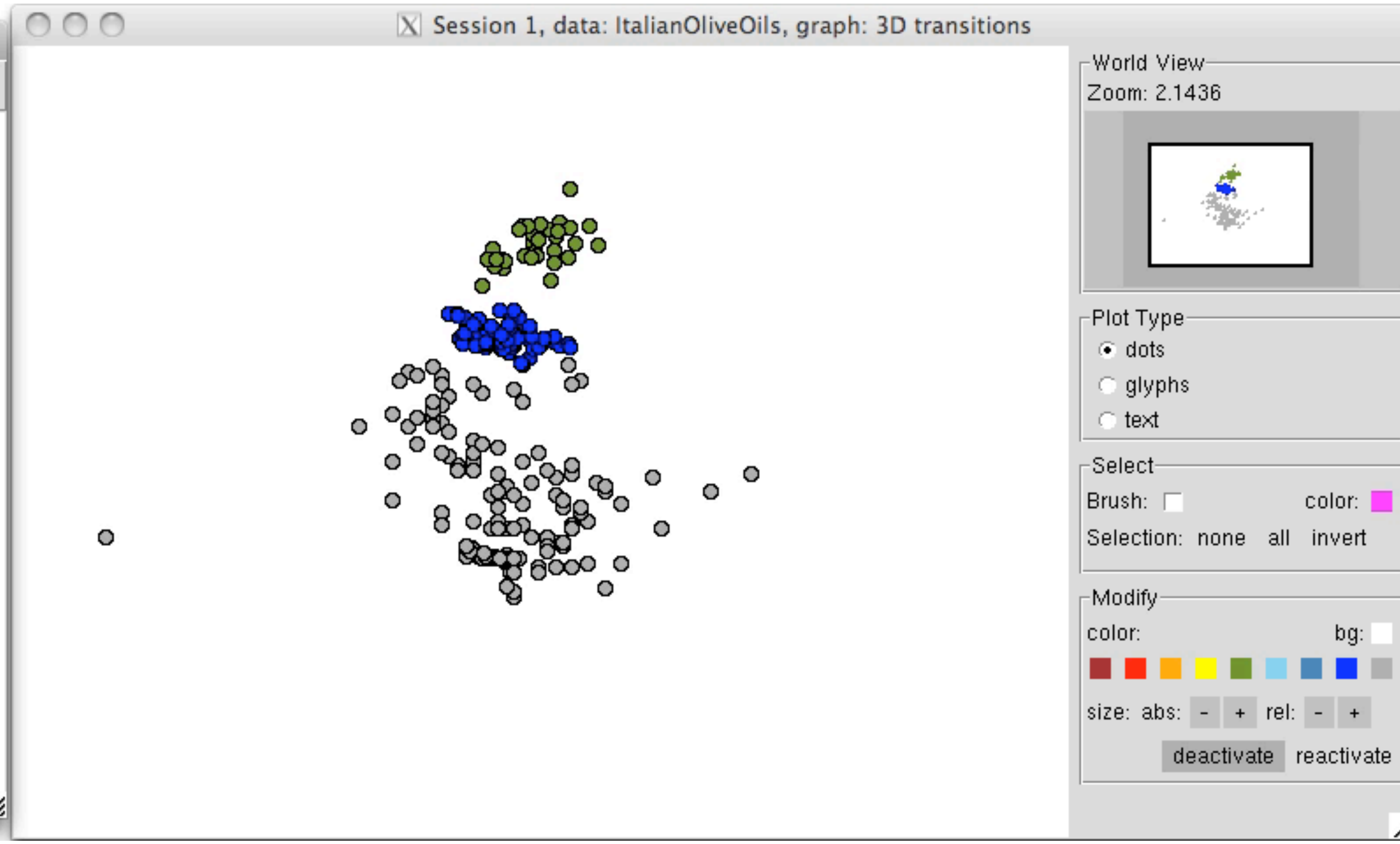
3d transition graph

Interactive scatterplot

Could also select a whole path to traverse

Example: Italian olive oils





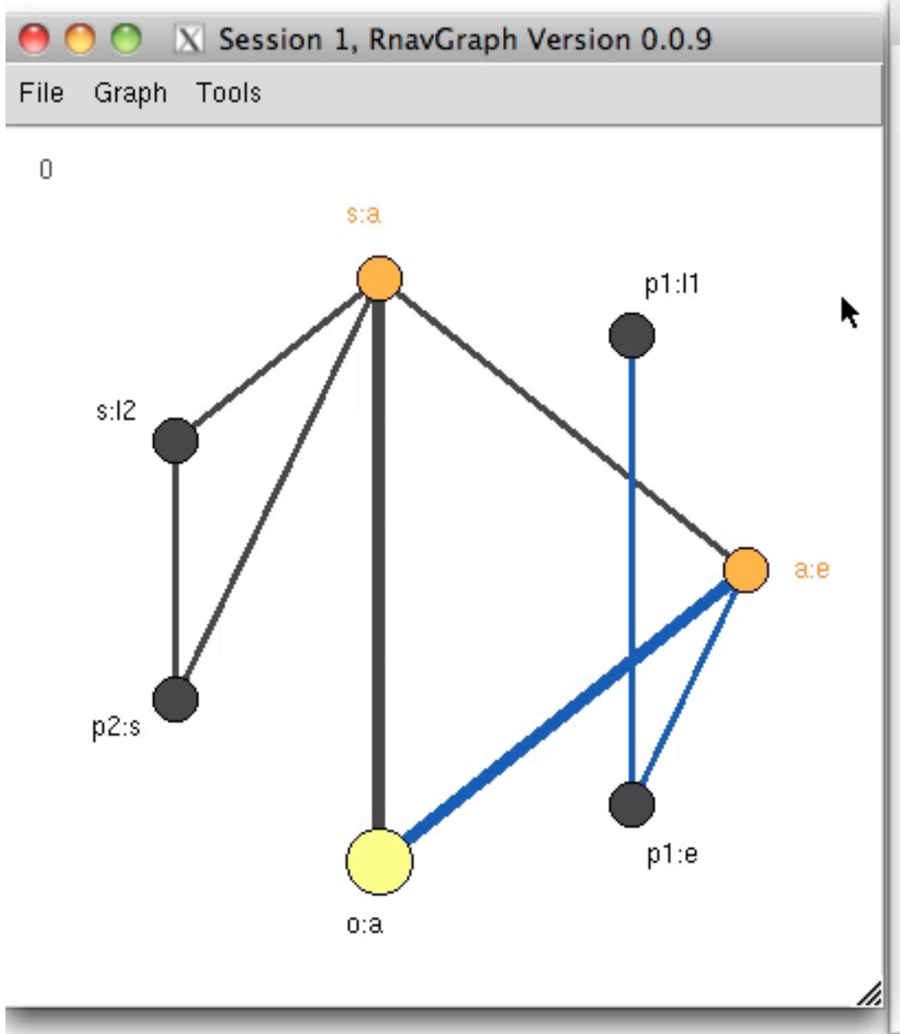
Interactive

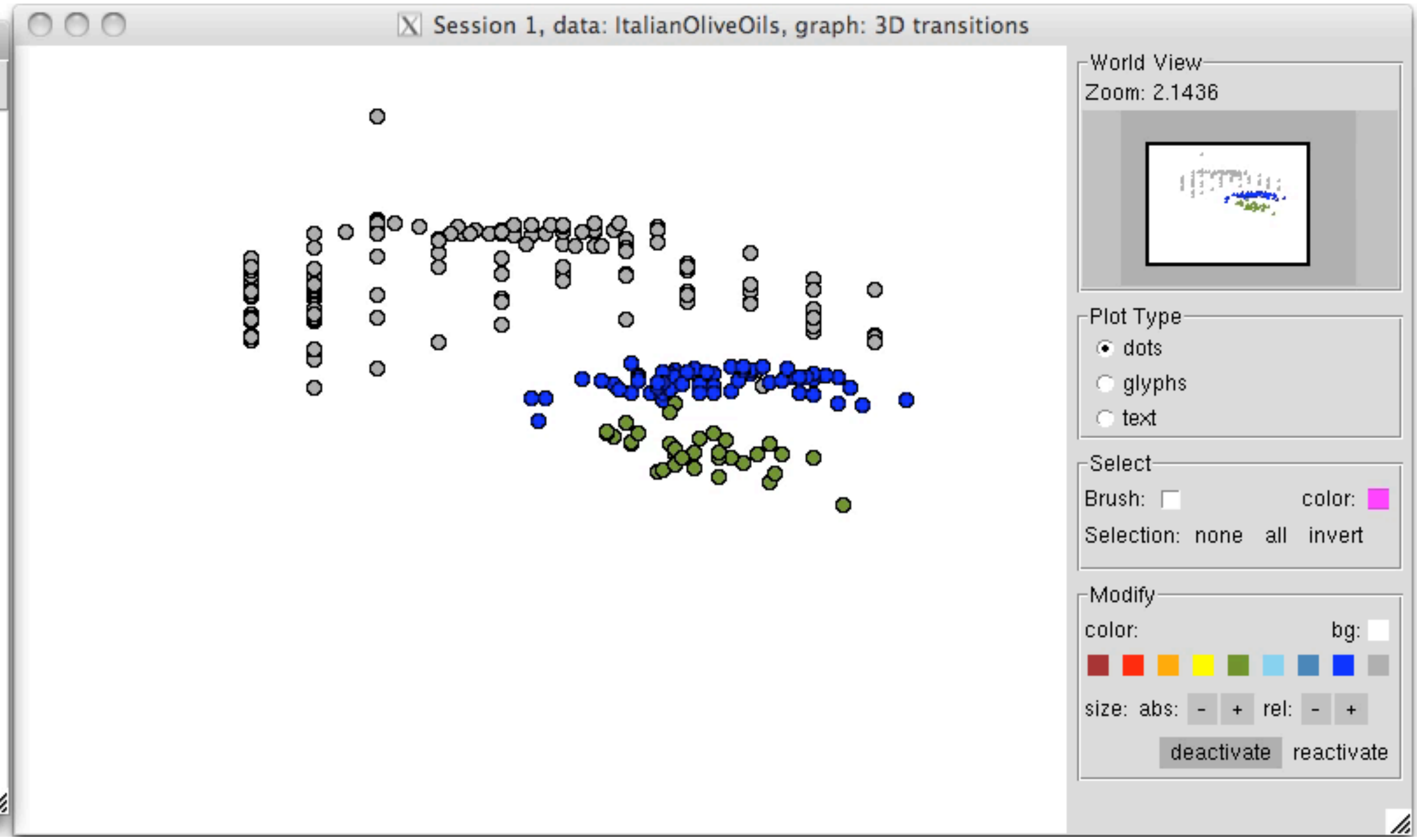
3d transition graph

Interactive scatterplot

Could also select a whole path to traverse

Example: Italian olive oils





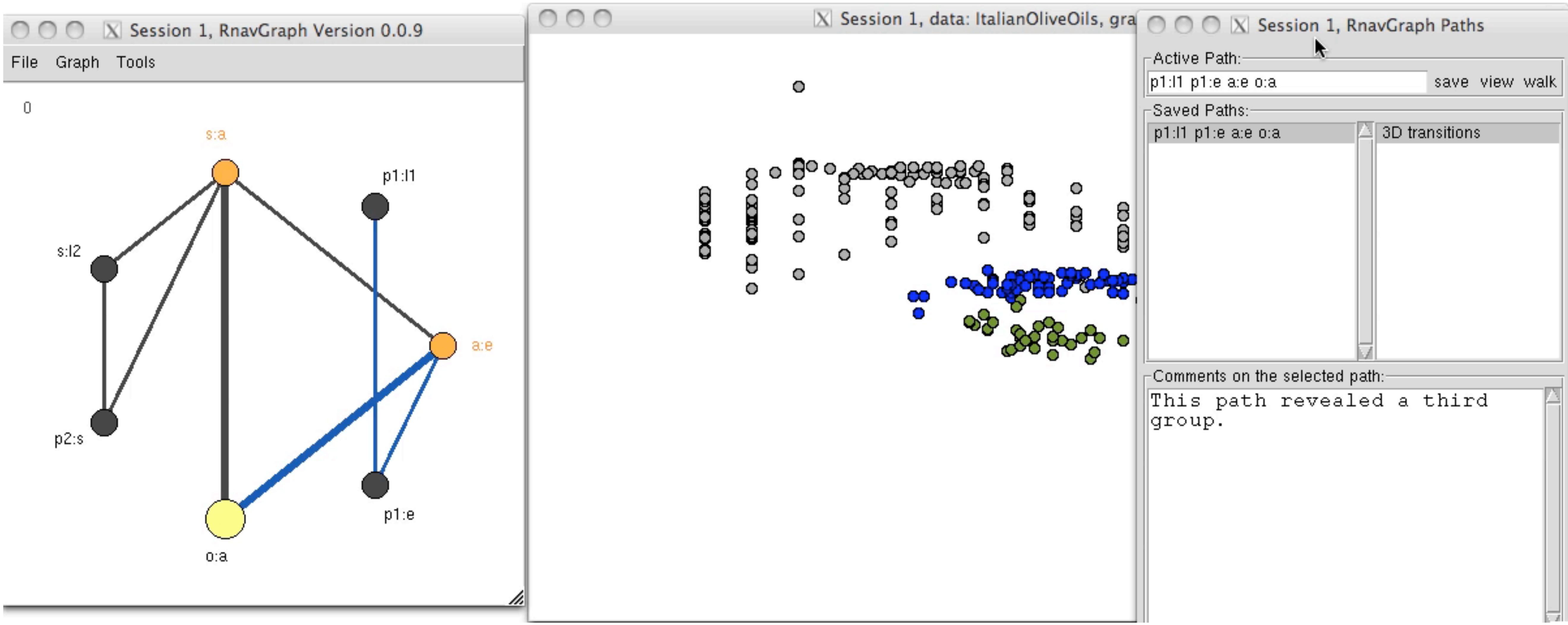
Interactive

3d transition graph

Interactive scatterplot

Paths can be saved, annotated, viewed, and walked again.

Example: Italian olive oils



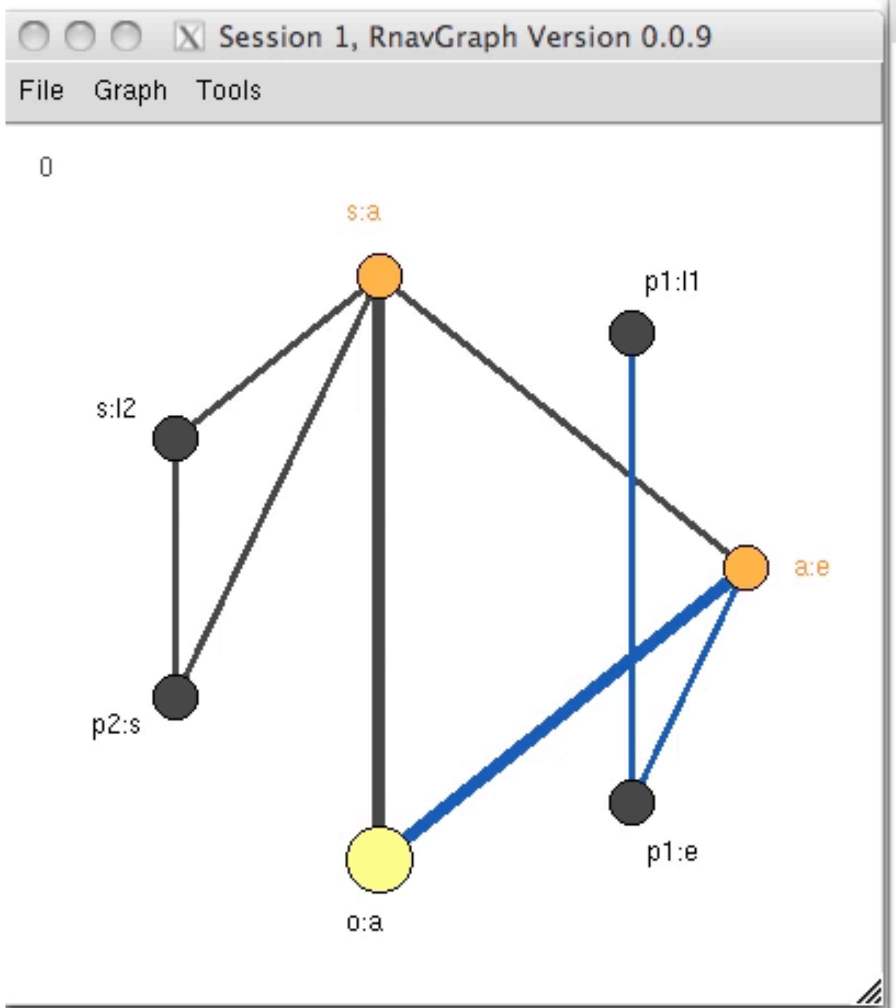
Interactive

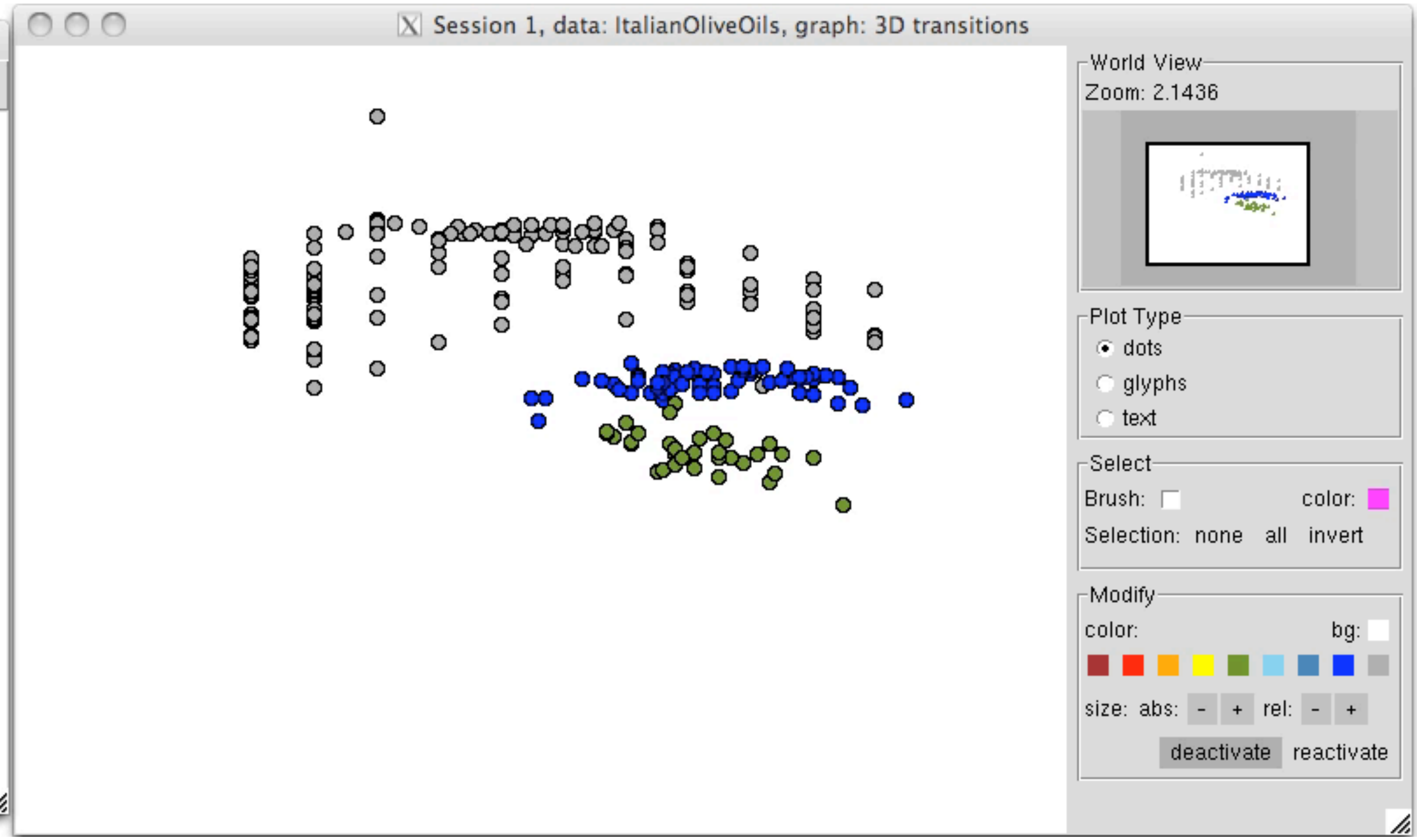
3d transition graph

Interactive scatterplot

Paths can be saved, annotated, viewed, and walked again.

Example: Italian olive oils





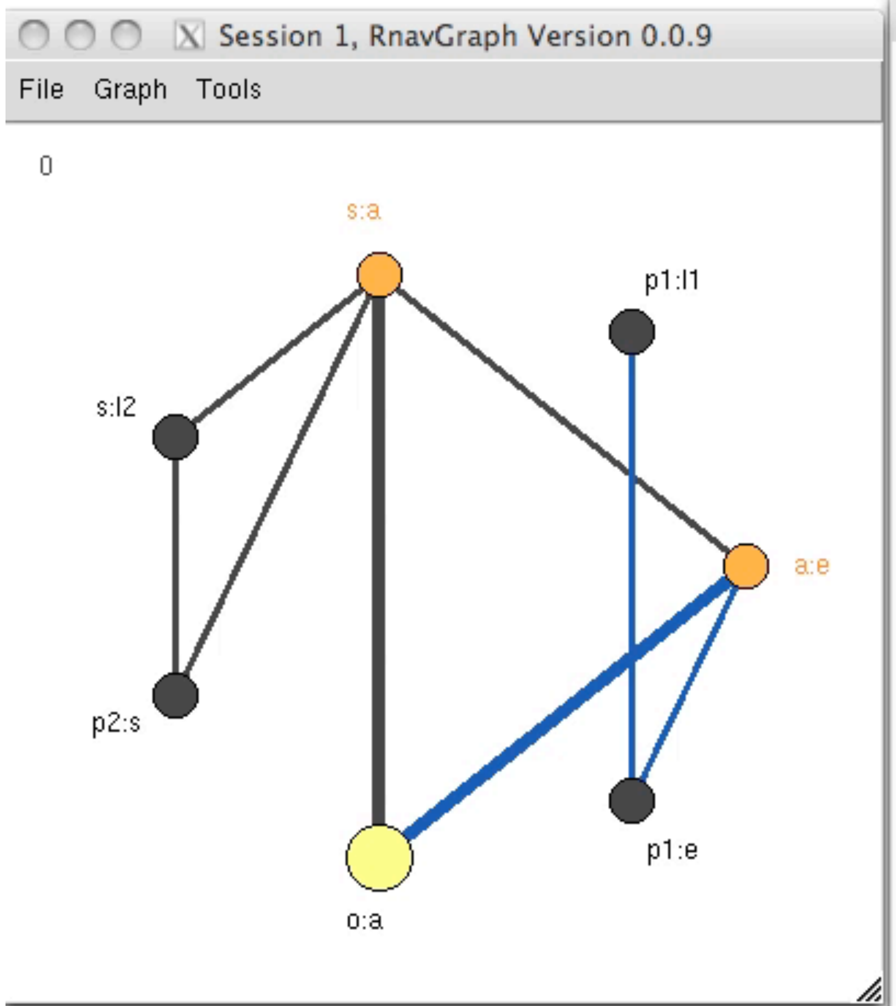
Interactive

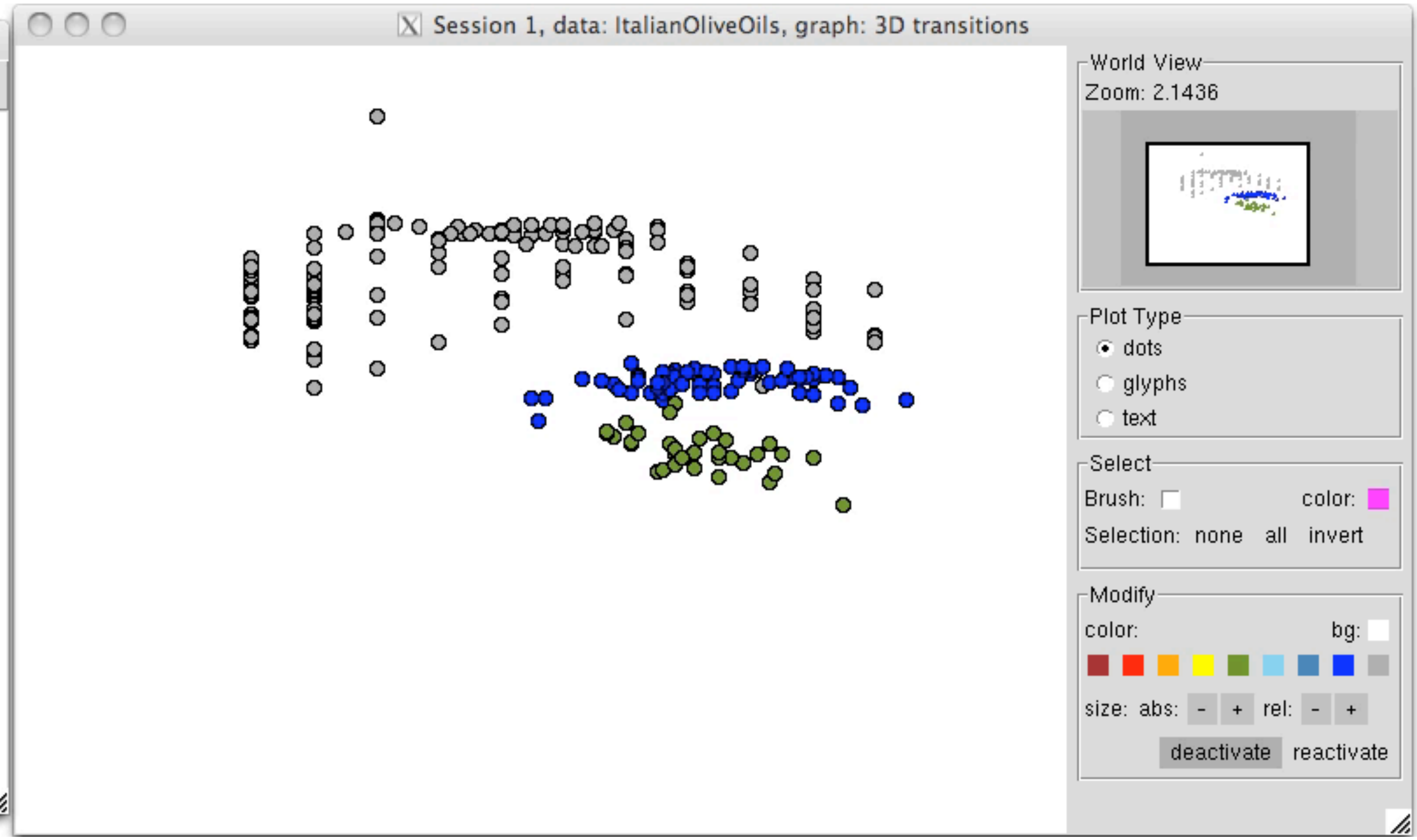
3d transition graph

Interactive scatterplot

Appears to be a third horizontal group ... zoom etc.

Example: Italian olive oils





Interactive

3d transition graph

Interactive scatterplot

Appears to be a third horizontal group ... zoom etc.

And that outlier

Example: Italian olive oils

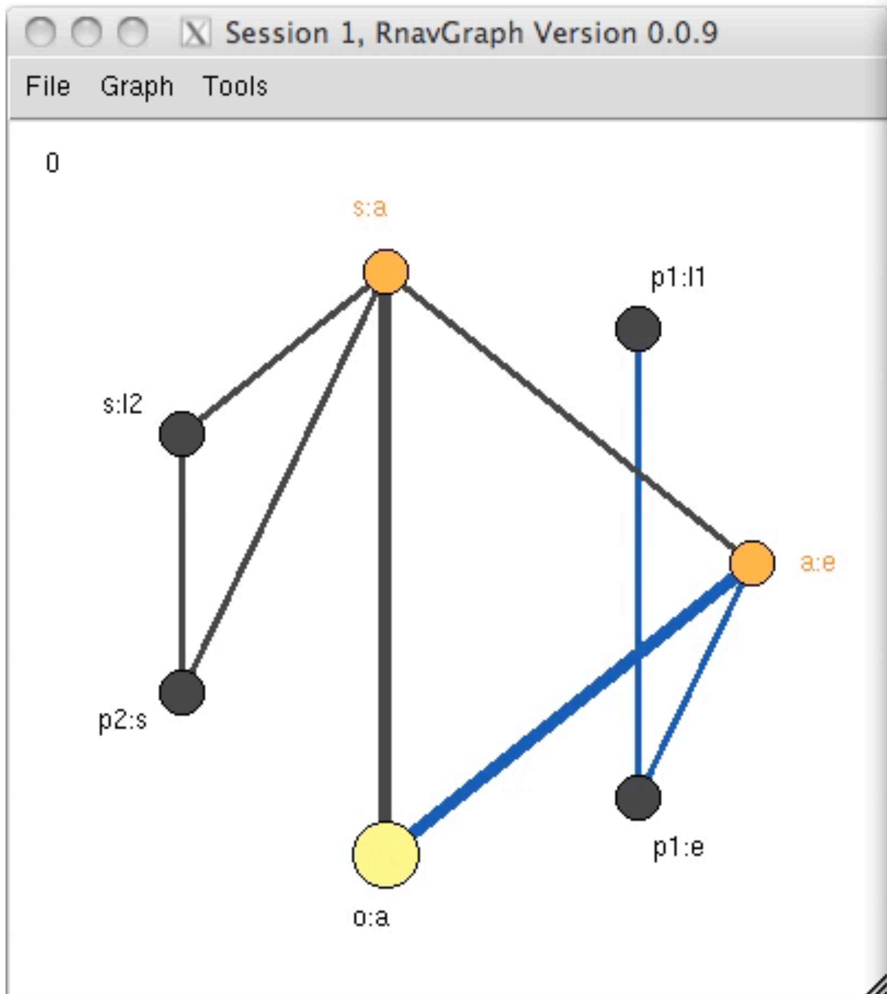
Interactive

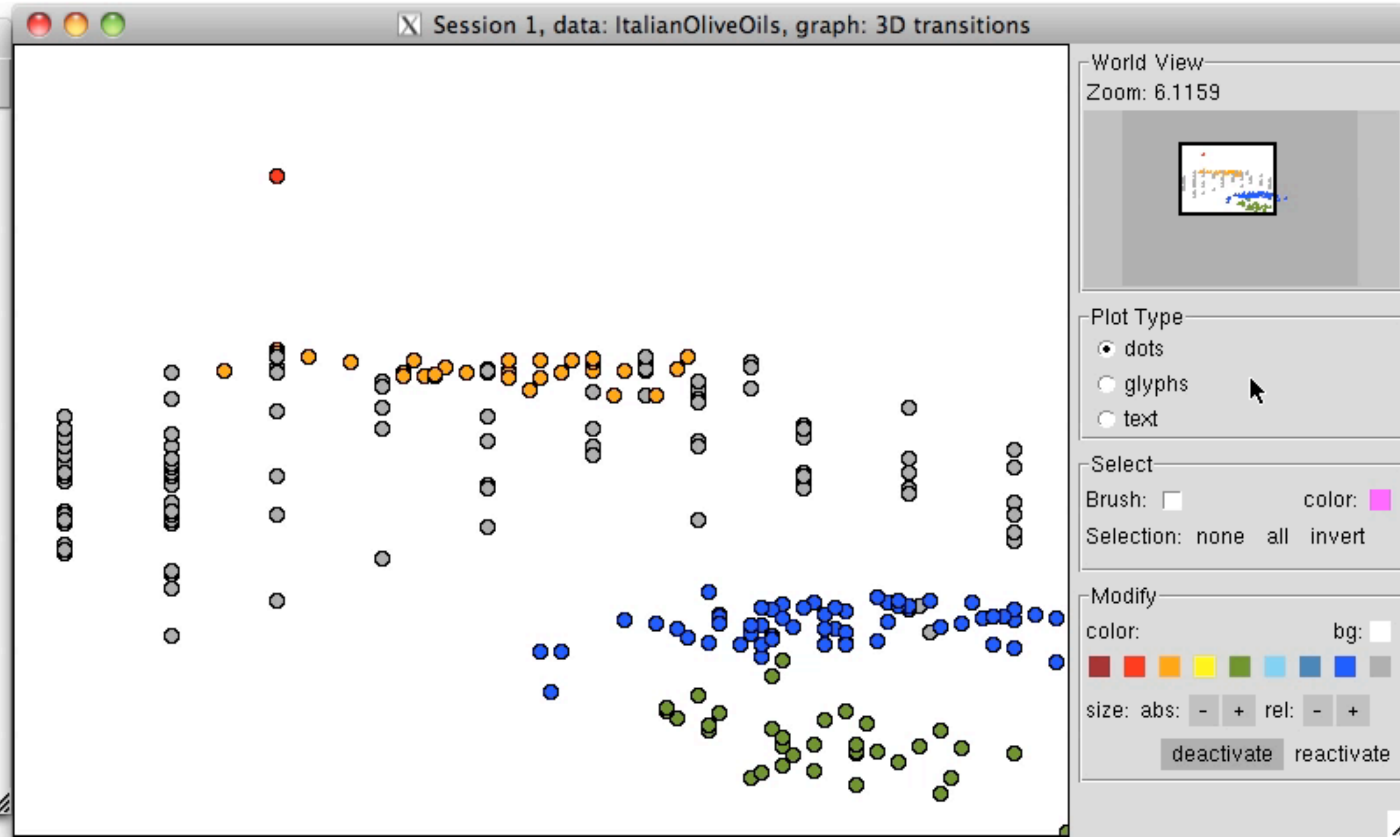
3d transition graph

Interactive scatterplot

Colour group orange, outlier red.

Example: Italian olive oils





Interactive

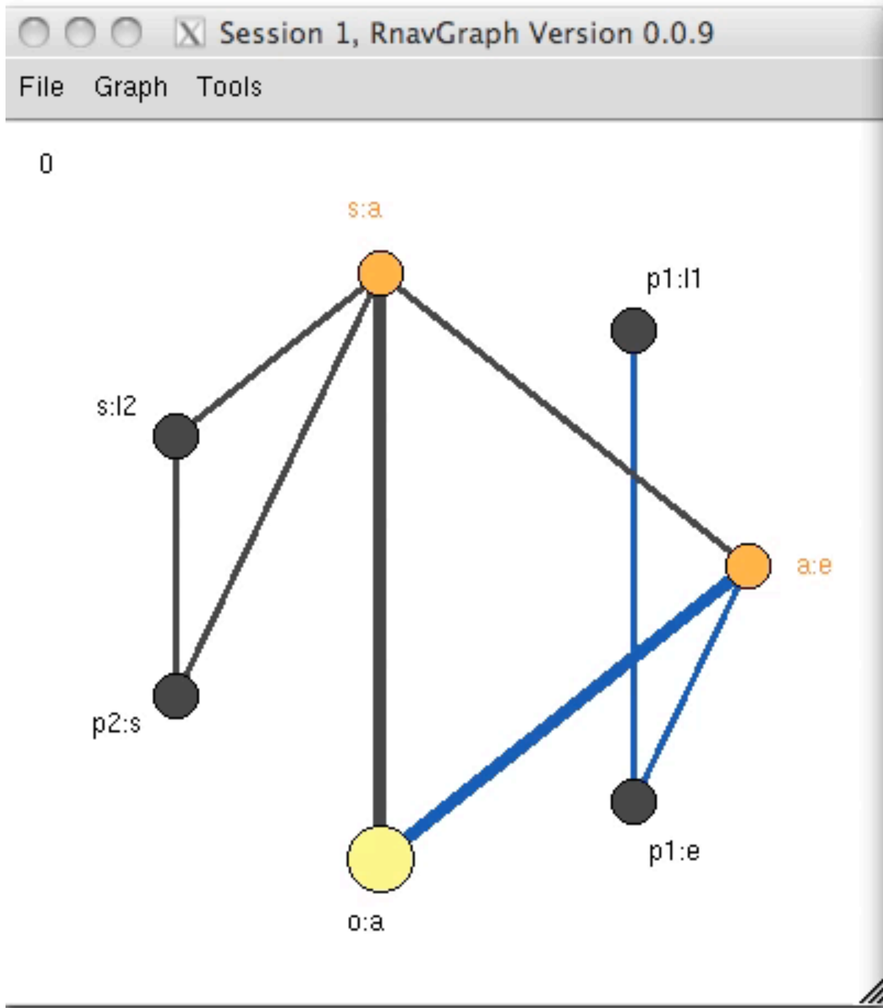
3d transition graph

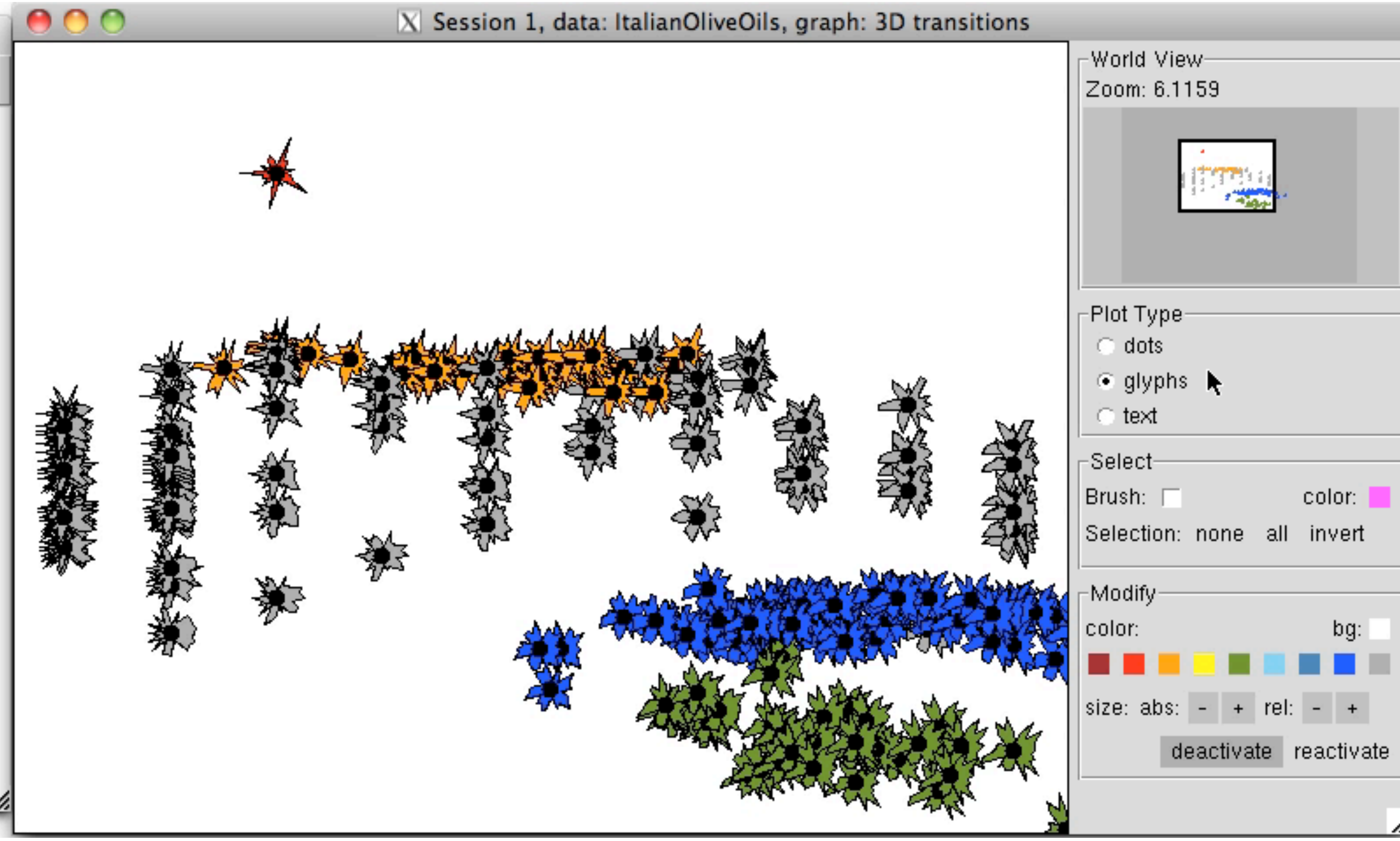
Interactive scatterplot

Colour group orange, outlier red.

Can switch to glyphs

Example: Italian olive oils





Interactive

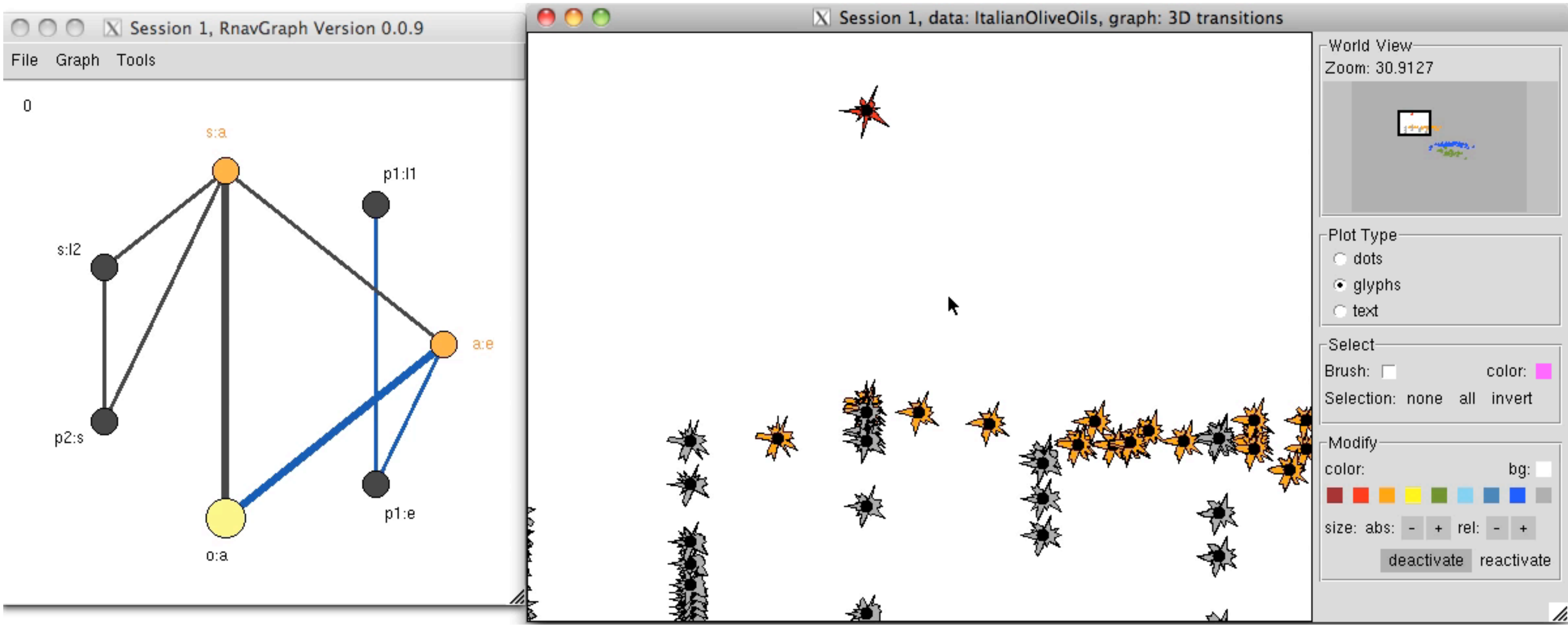
3d transition graph

Interactive scatterplot

Colour group orange, outlier red.

Focus on a region

Example: Italian olive oils



Interactive

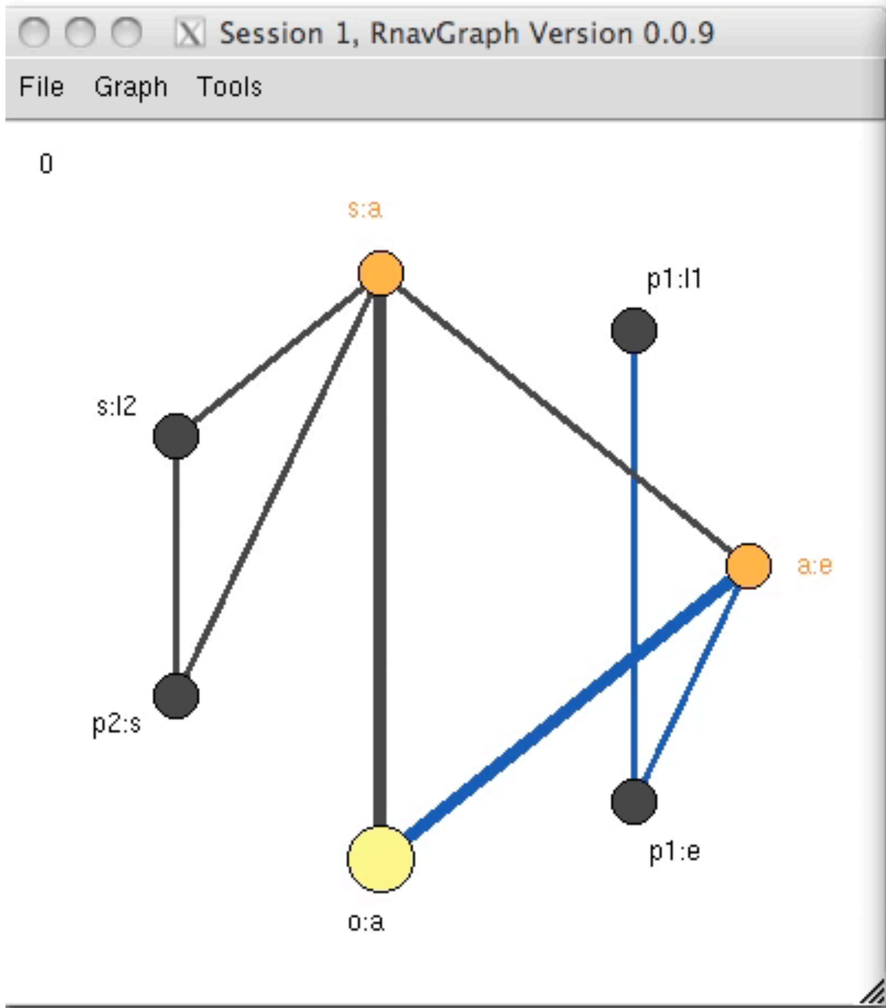
3d transition graph

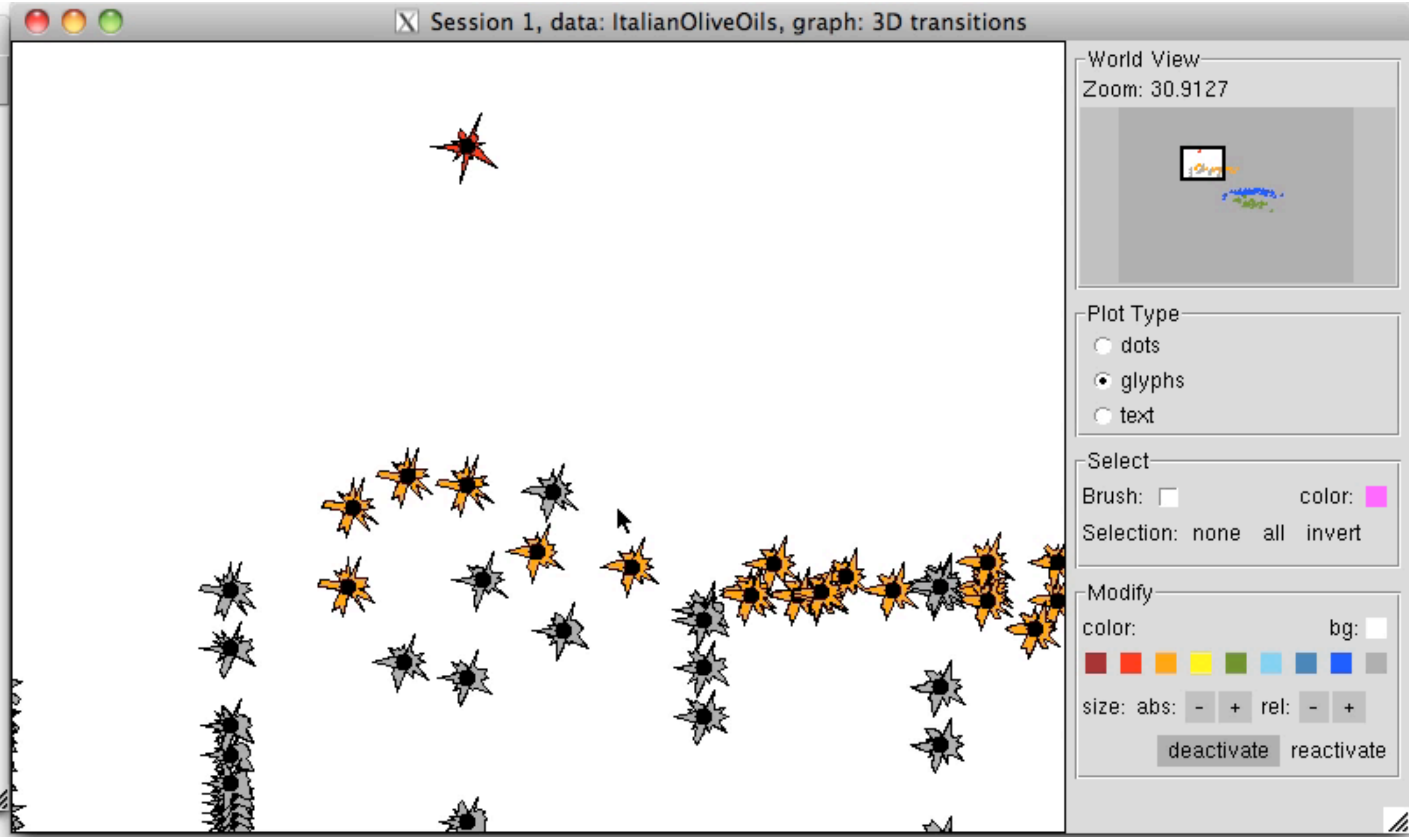
Interactive scatterplot

Colour group orange, outlier red.

Move to compare shapes

Example: Italian olive oils





Interactive

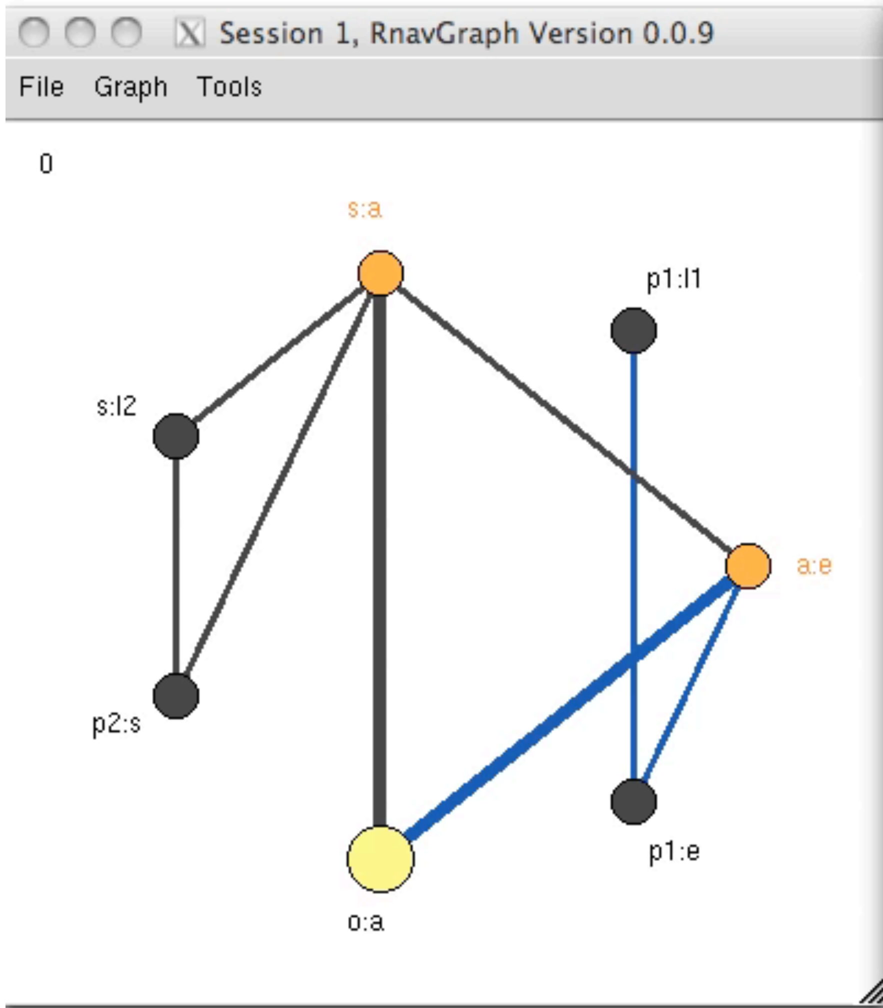
3d transition graph

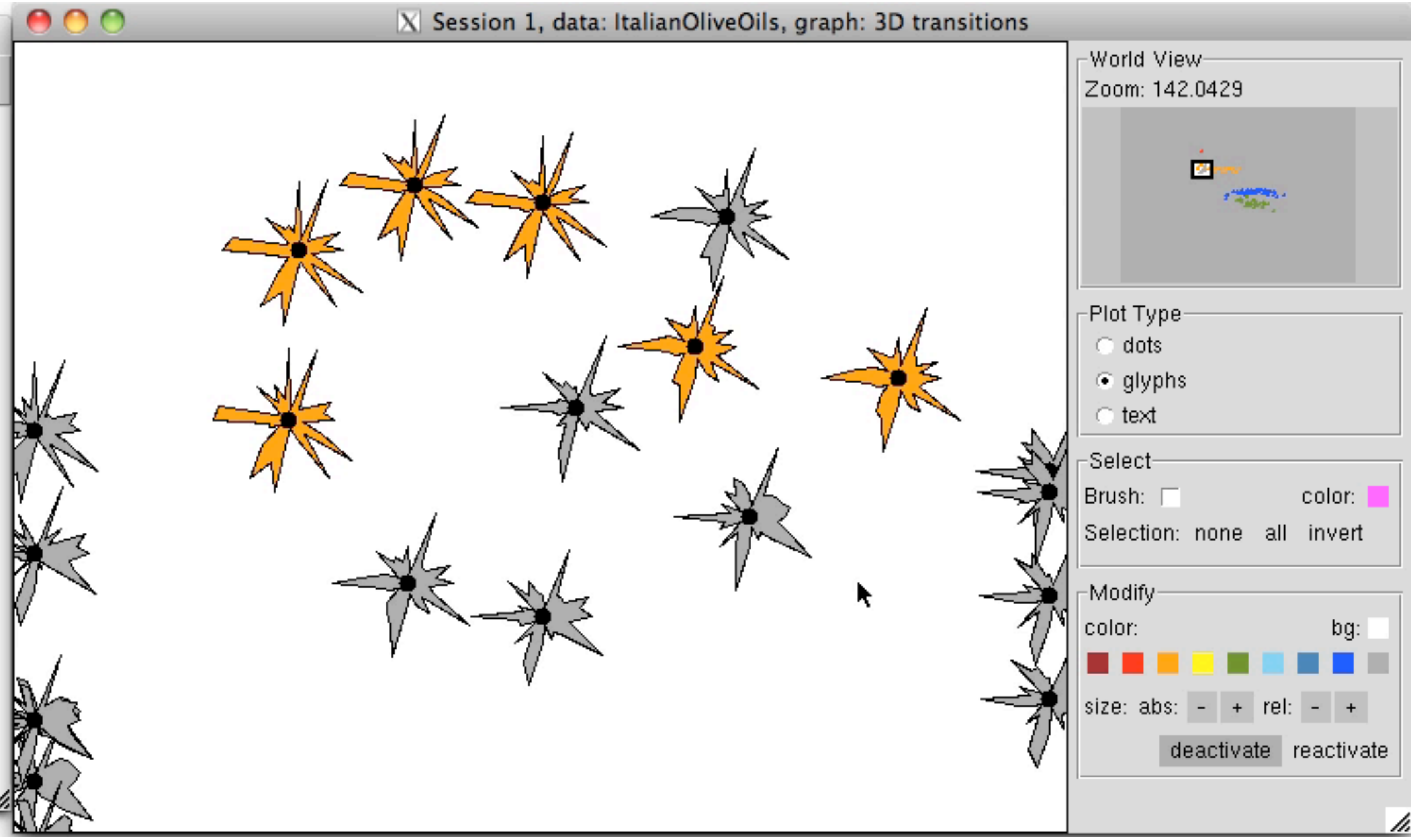
Interactive scatterplot

Colour group orange, outlier red.

Enlarge to compare shapes

Example: Italian olive oils





Interactive

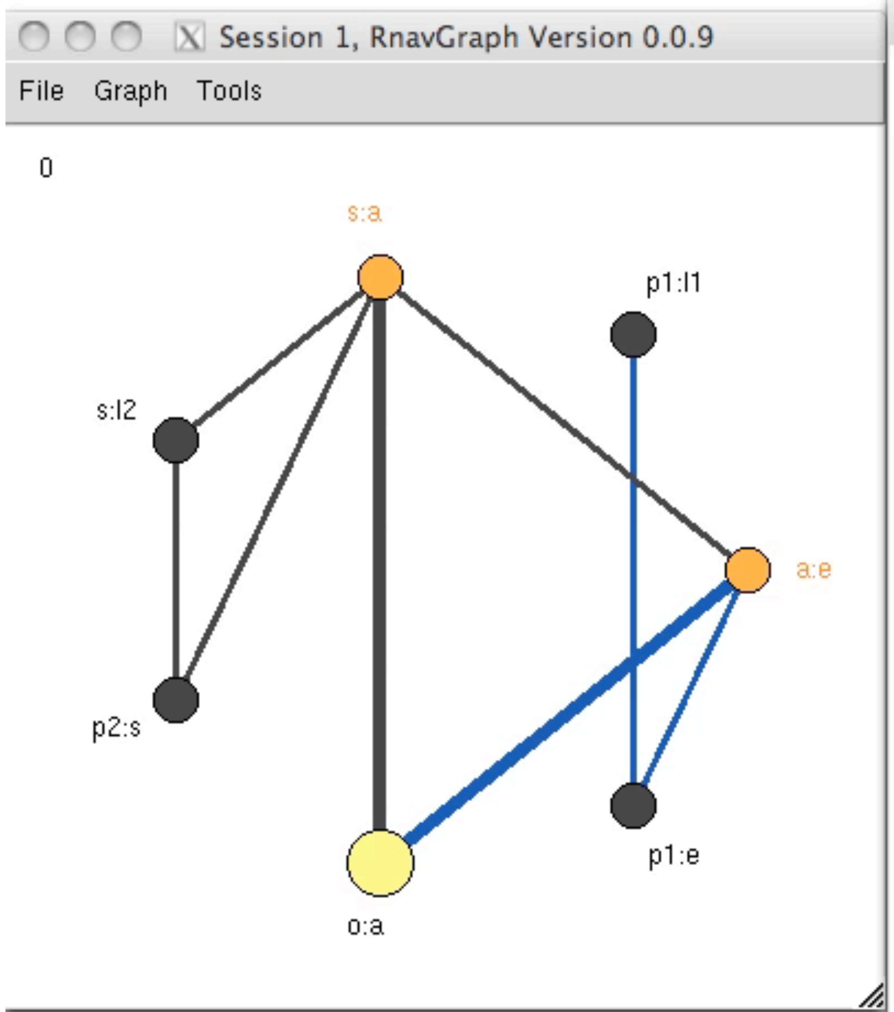
3d transition graph

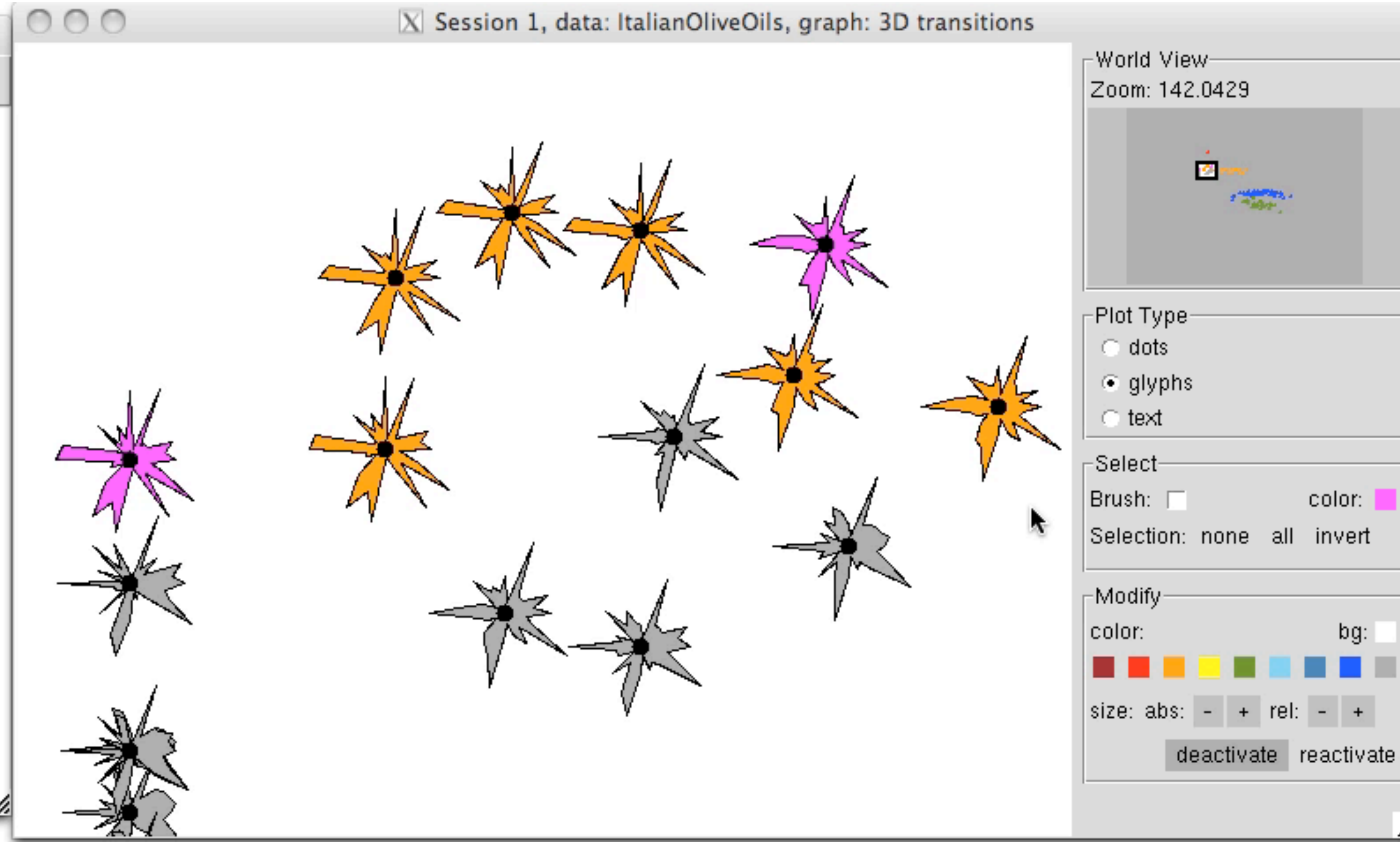
Interactive scatterplot

Colour group orange, outlier red.

Identify possible orange?

Example: Italian olive oils





Interactive

3d transition graph

Interactive scatterplot

Colour group orange, outlier red.

Can actually check here

Example: Italian olive oils

Continue in this way:

- bring back deactivated points
- identify groups, reassign points
- note natural hierarchical clustering
- save grouping by colour in R

Challenge

Large p => large graphs

- p ... overall dimensionality (olive, p=8)
  - $\binom{p}{2}$... potential 2d nodes (28)
  - $\binom{p}{3}$... potential 3d edges (56)

p	5	10	20	50
$\binom{p}{2}$	10	45	190	1225
$\binom{p}{3}$	10	120	1140	19600

Challenge

Large p => large graphs

- p ... overall dimensionality (olive, p=8)
  - $\binom{p}{2}$ ... potential 2d nodes (28)
  - $\binom{p}{3}$ ... potential 3d edges (56)

p	5	10	20	50
$\binom{p}{2}$	10	45	190	1225
$\binom{p}{3}$	10	120	1140	19600

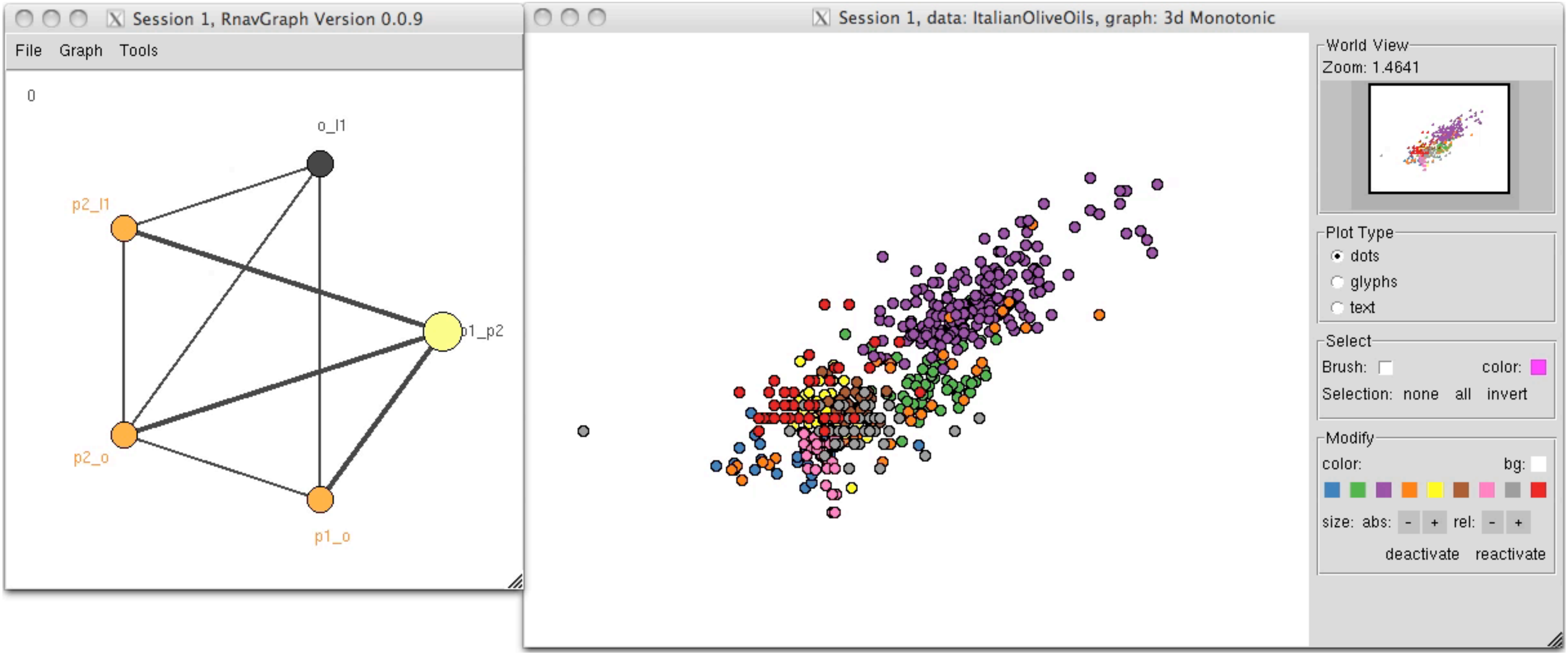
Need to start with small, but interesting, graphs

Interesting node pairs

For each scagnostic, calculate the value for every pair.

View only those pairs with high scores (e.g. top fraction of scores).

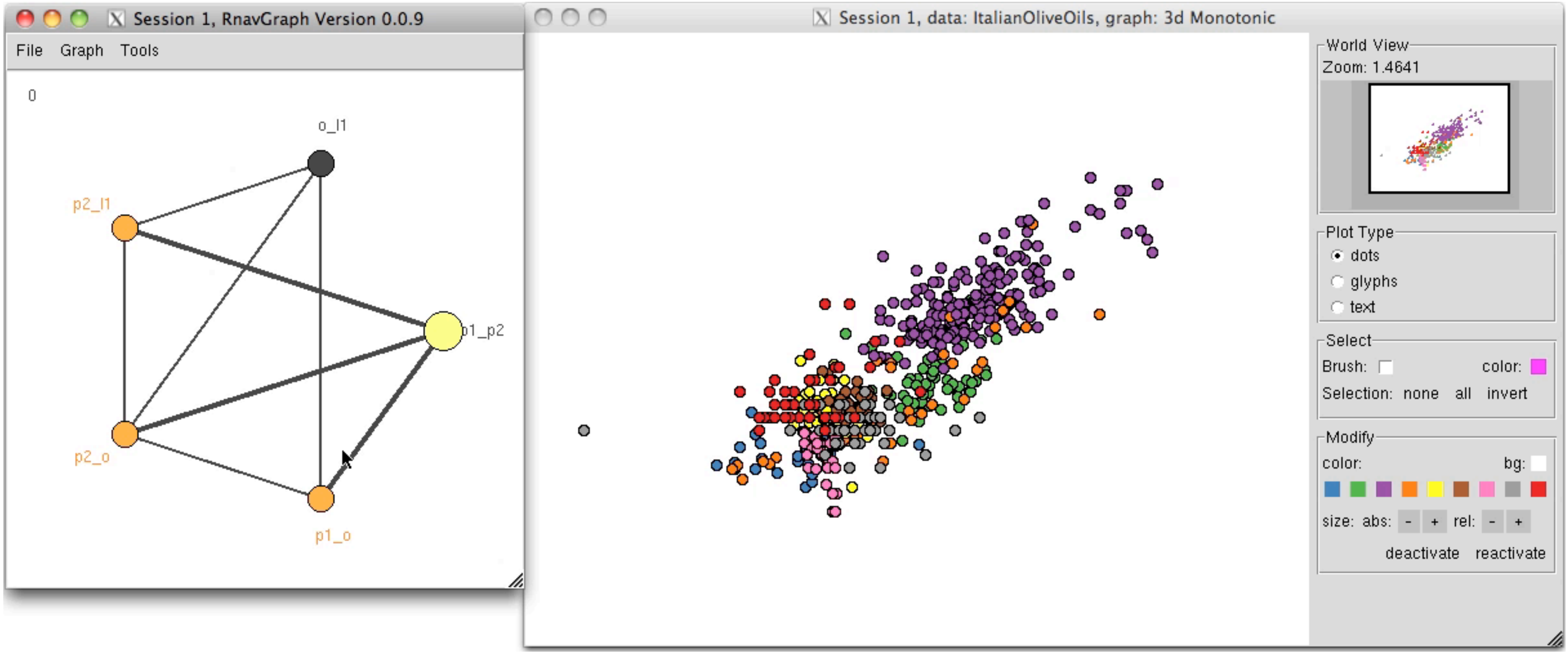
Scagnostics: Italian olive oils



3D Monotonic

Groups coloured by regions

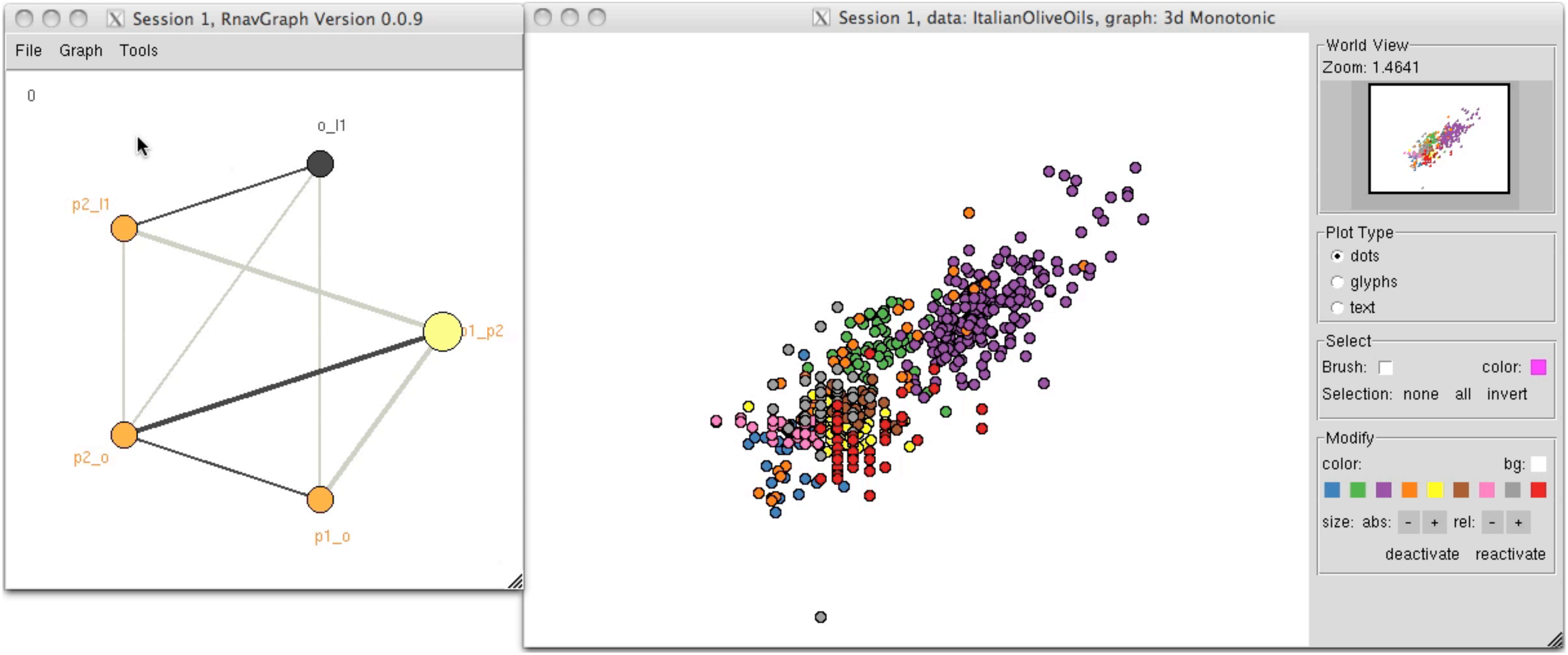
Scagnostics: Italian olive oils



3D Monotonic

Groups coloured by regions

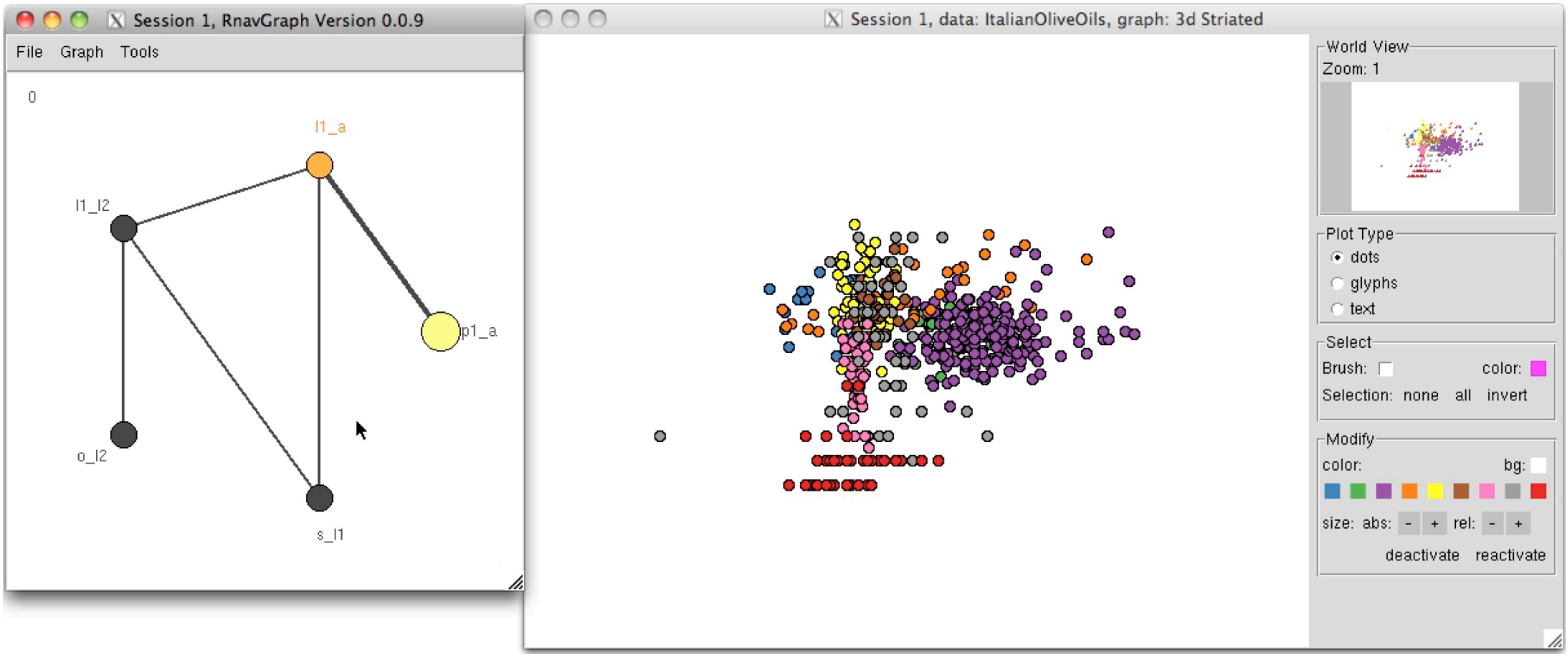
Scagnostics: Italian olive oils



Switch to 3D Striated

Groups coloured by regions

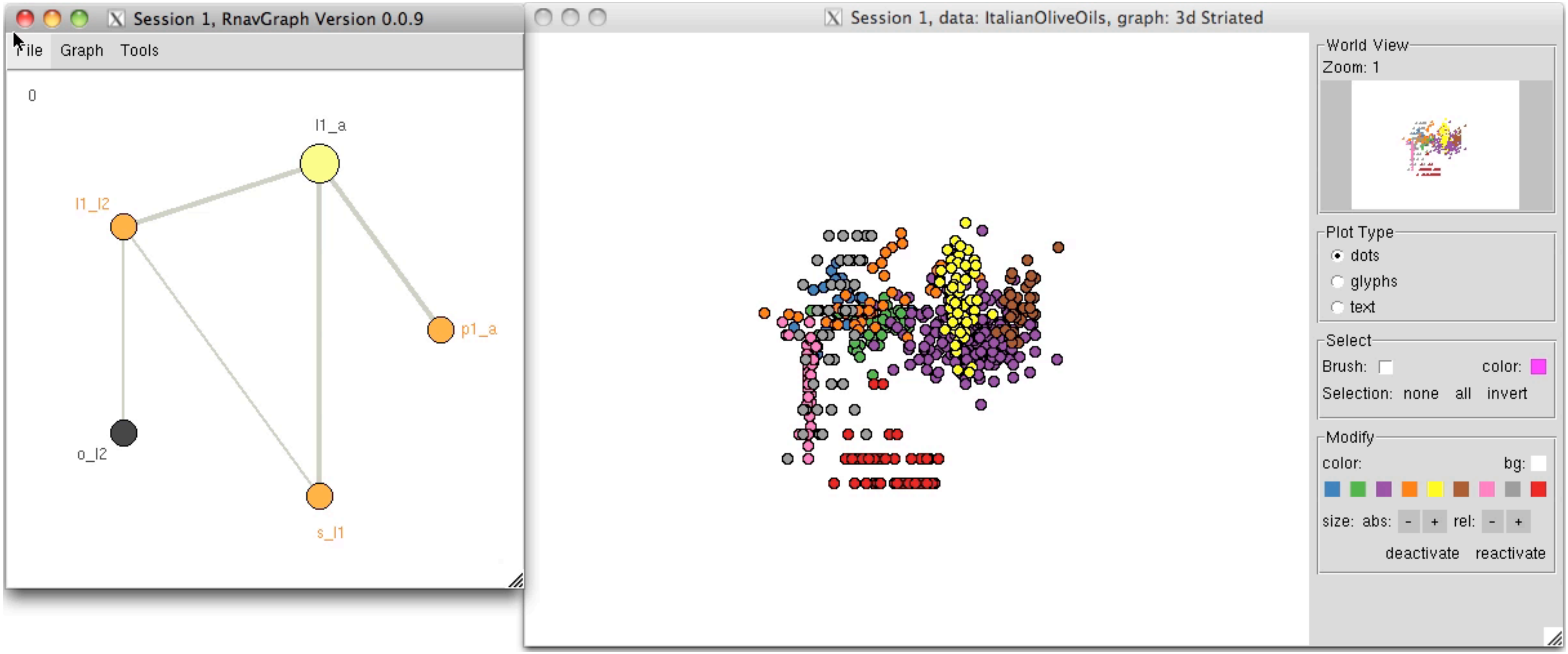
Scagnostics: Italian olive oils



3D Striated

Groups coloured by regions

Scagnostics: Italian olive oils



3D Non-Convex

Groups coloured by regions

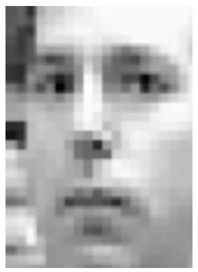
Challenge

Large p => large graphs

- scagnostics work well, but when p is very large, so is $\binom{p}{2}$
- dimensionality reduction methods could be employed.

Example: images

Frey: 1,965 movie frames



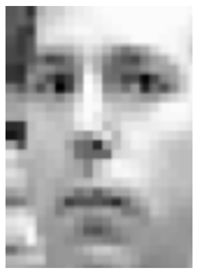
Example: images

Frey: 1,965 movie frames

28 x 20 array

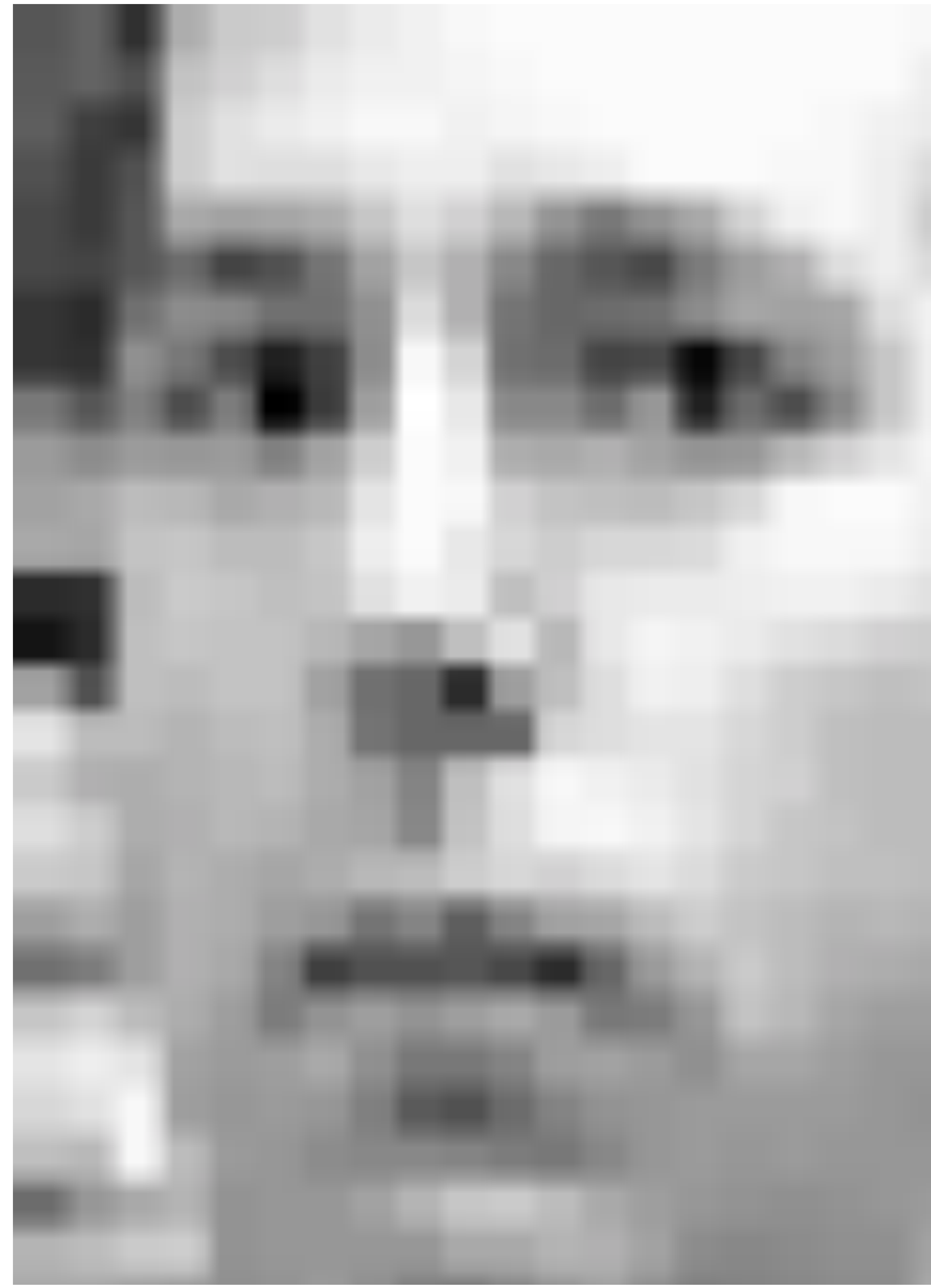
Example: images

Frey: 1,965 movie frames




28 x 20 array

560 dimensions



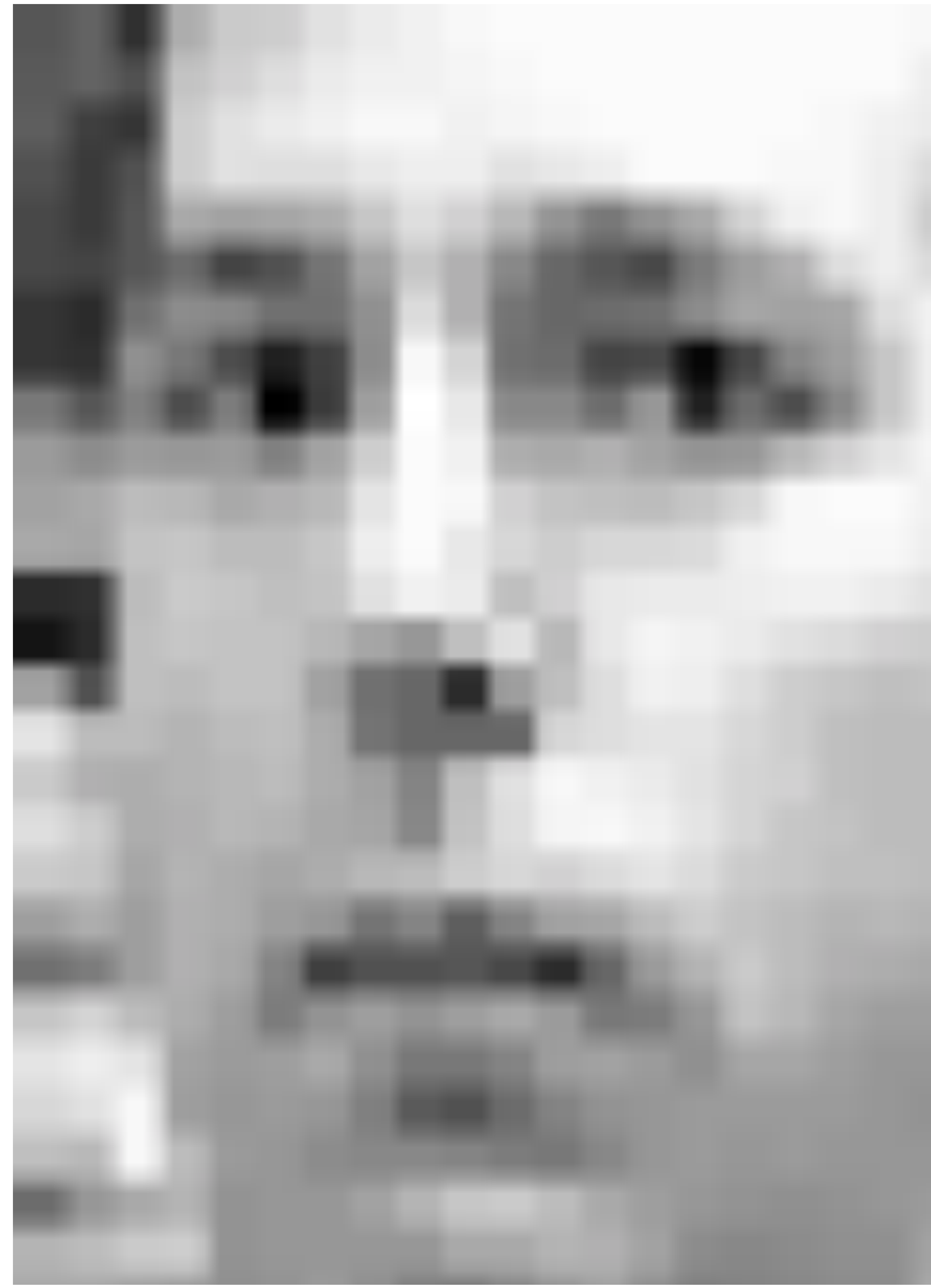
Example: images

Frey: 1,965 movie frames

28 x 20 array

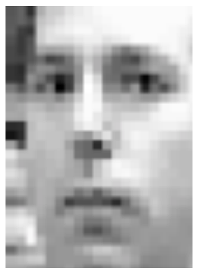
560 dimensions

explore via low dimensional spaces



Example: images

Frey: 1,965 movie frames




560 dimensions

Example: images

Frey: 1,965 movie frames

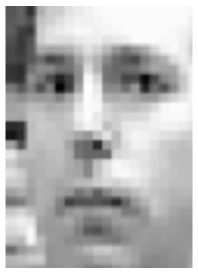
560 dimensions

Using LLE: local linear embedding

k=12 neighbours

Example: images

Frey: 1,965 movie frames



560 dimensions

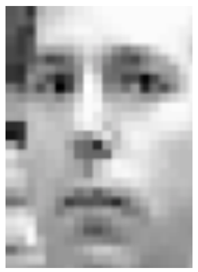
Using LLE: local linear embedding

k=12 neighbours

reduce to 5

Example: images

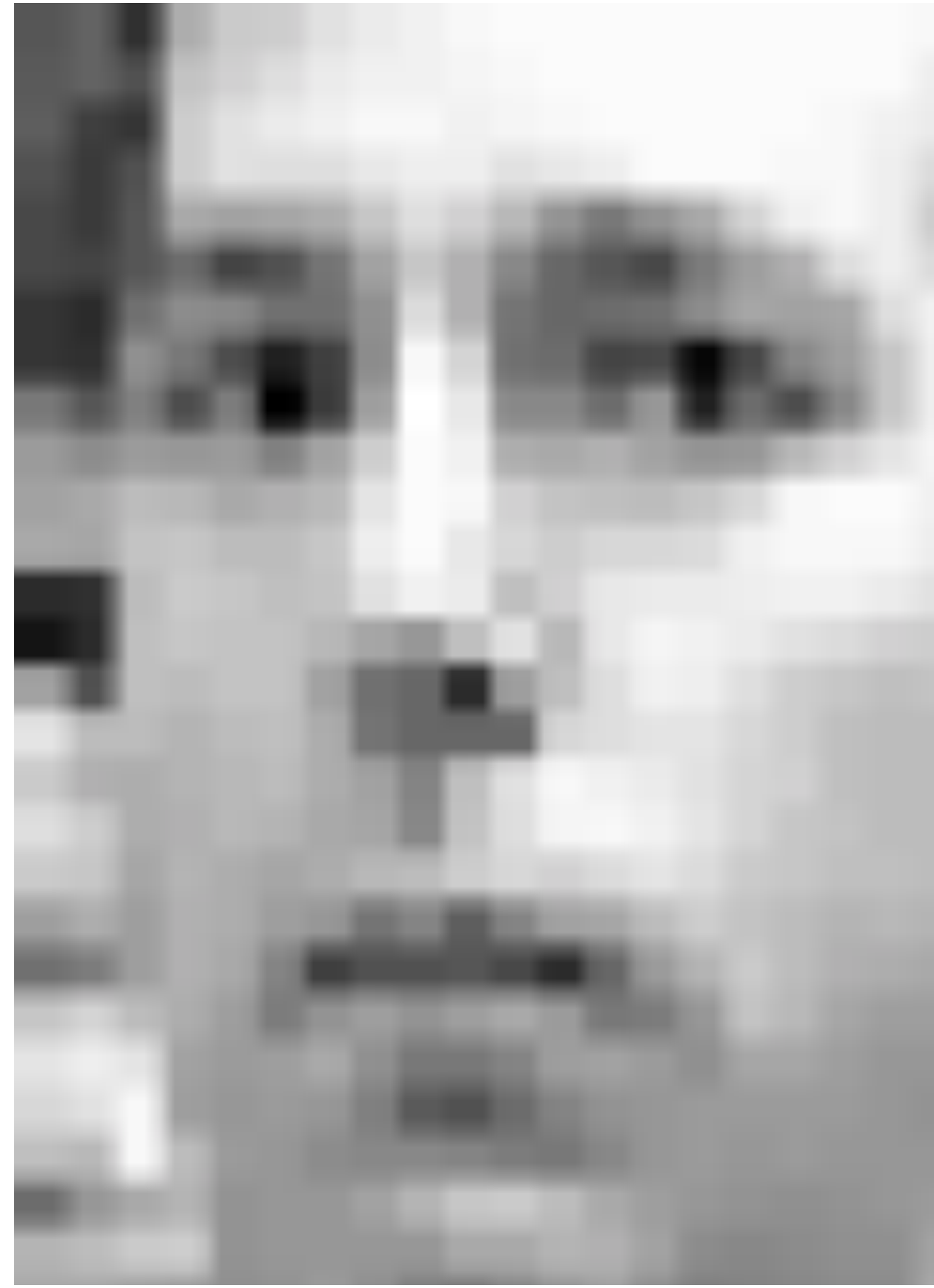
Frey: 1,965 movie frames



560 dimensions

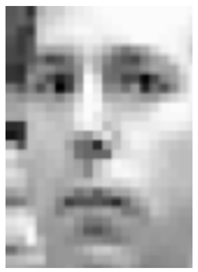
reduce to 5

interactive low-d view



Example: images

Frey: 1,965 movie frames



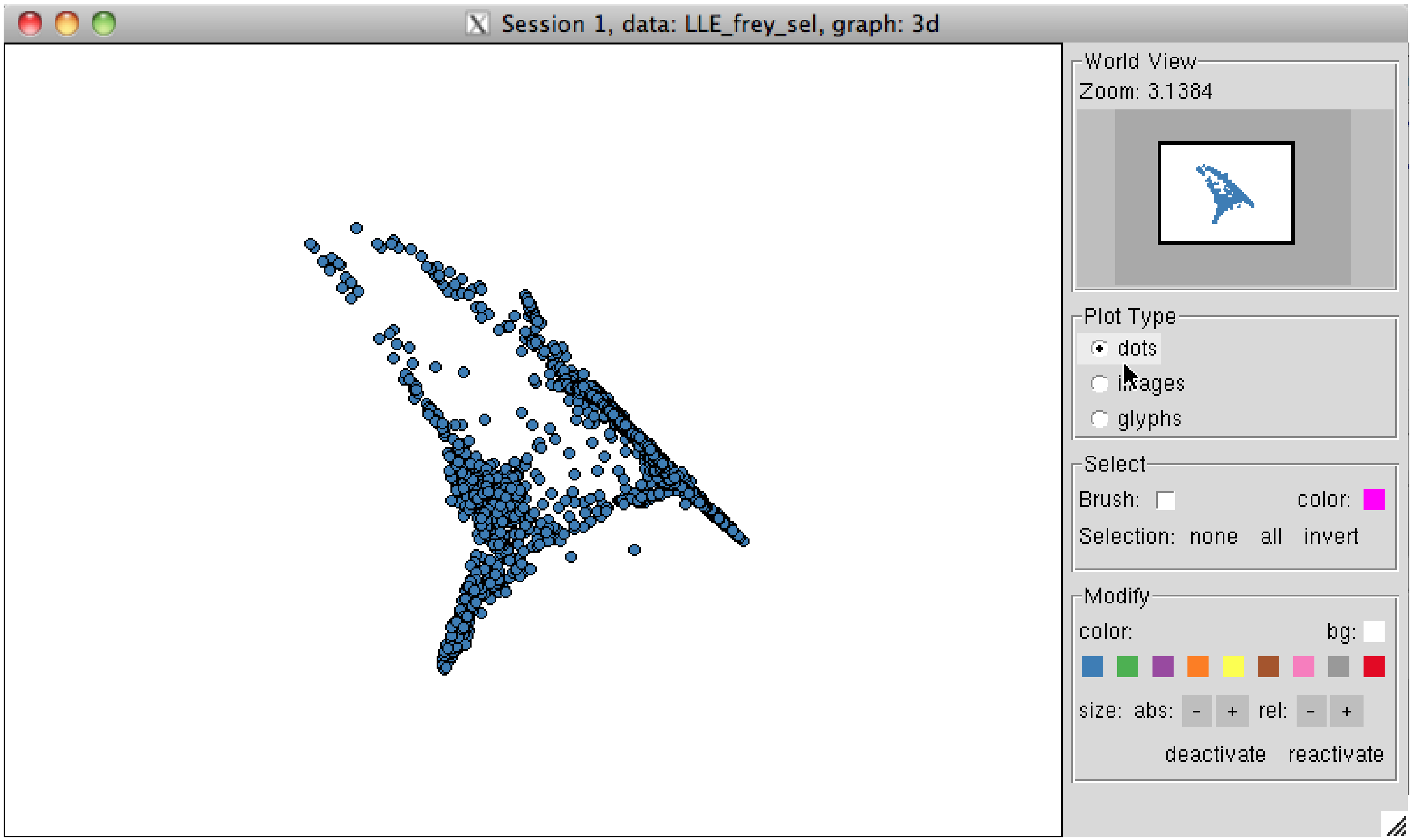
560 dimensions

reduce to 5

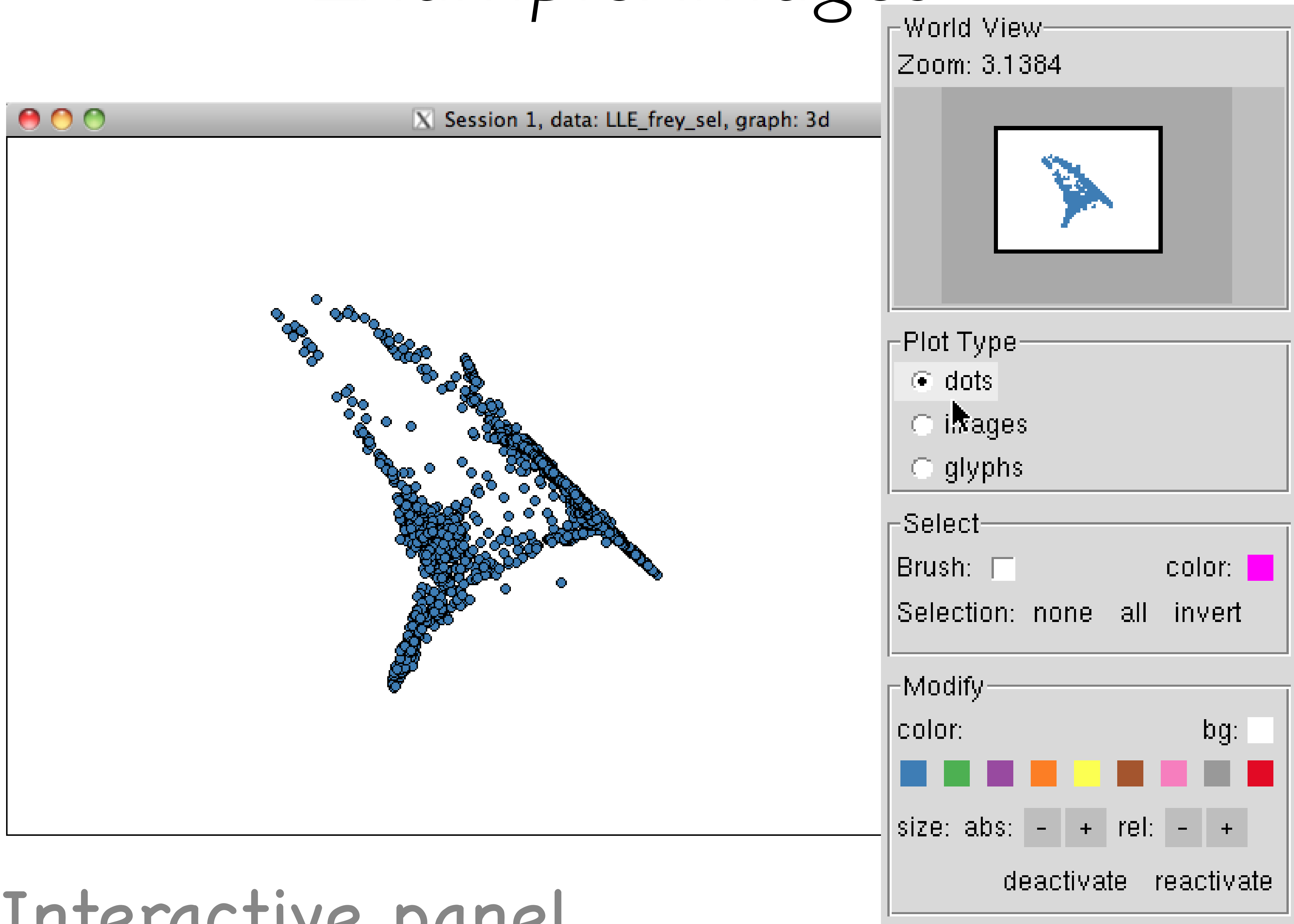
interactive low-d view

connect low-d views

Example: images

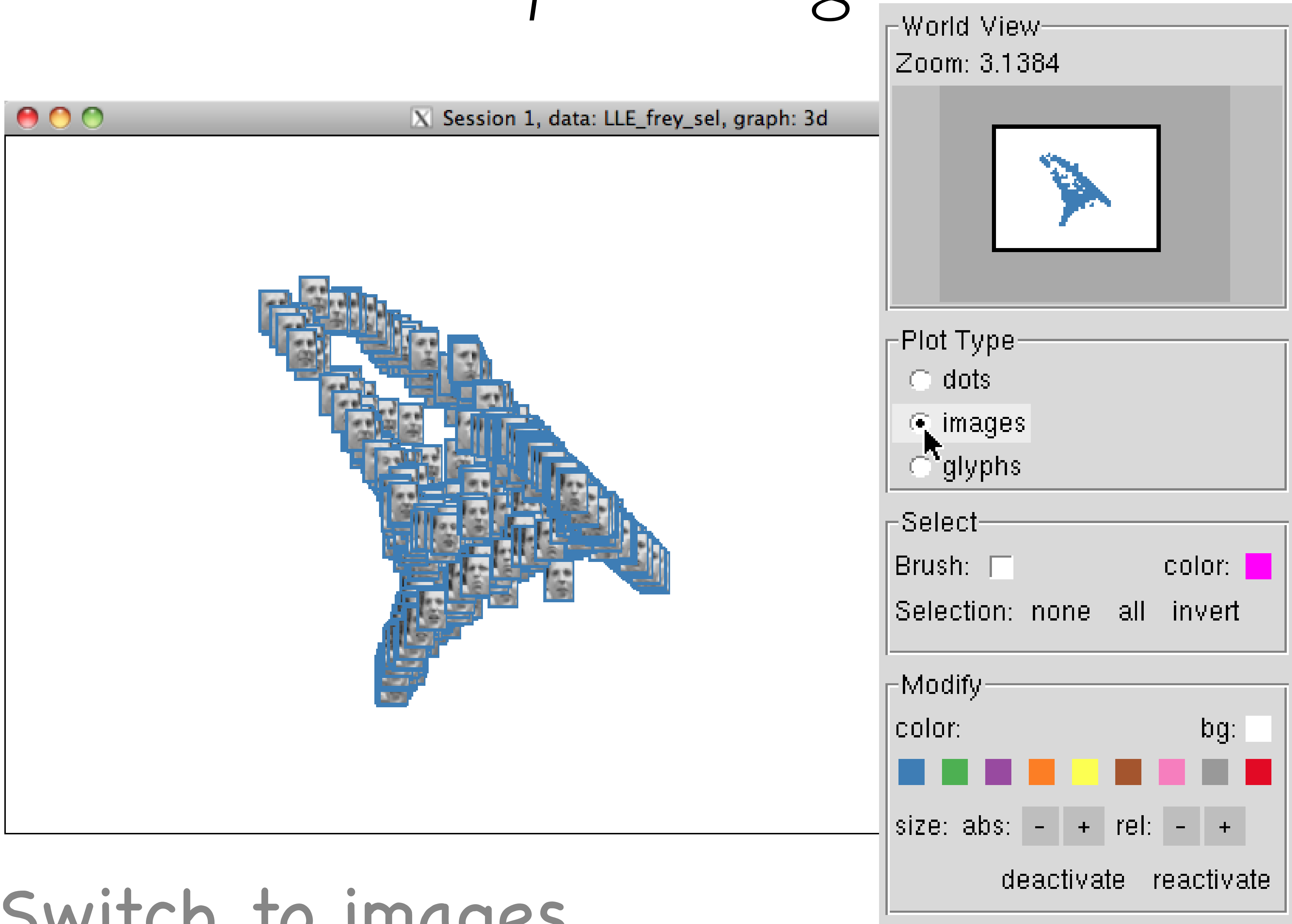


Example: images



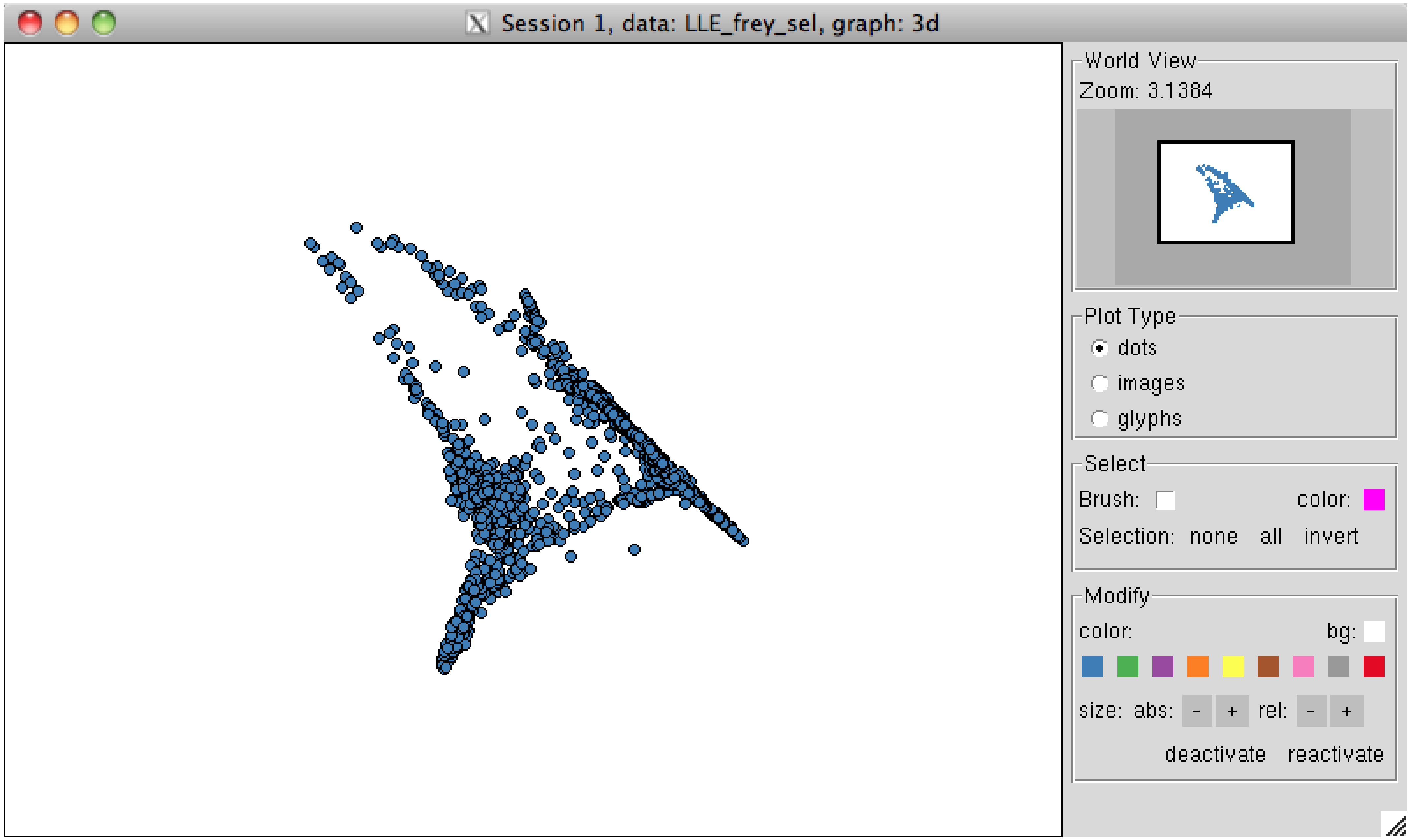
Interactive panel

Example: images



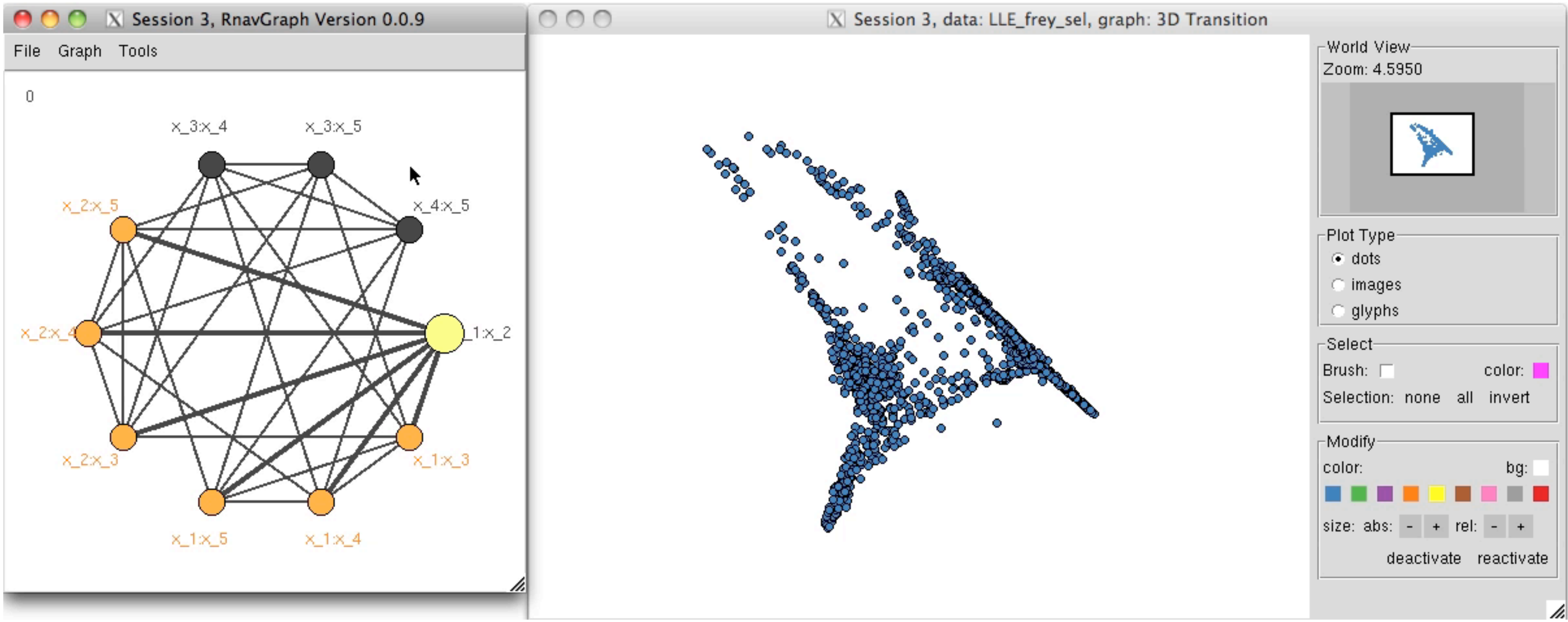
Switch to images

Example: images



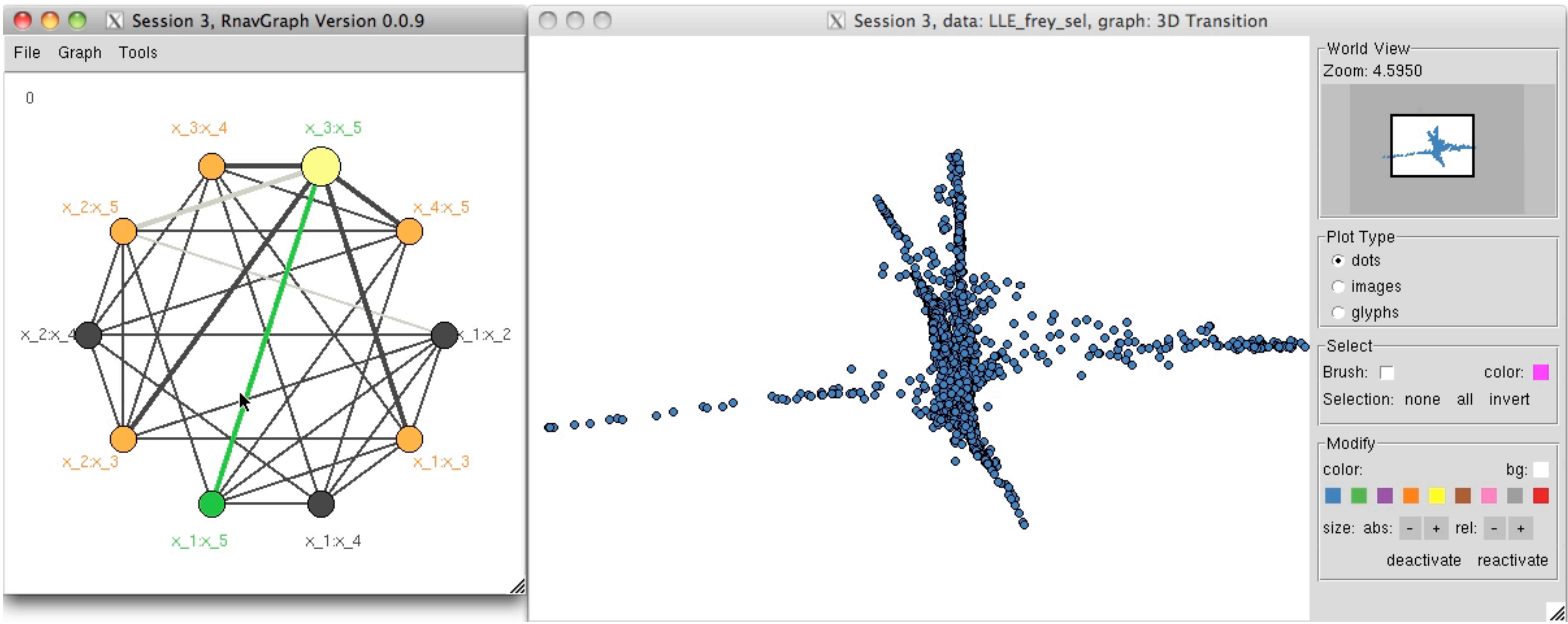
Back to dots

Example: images



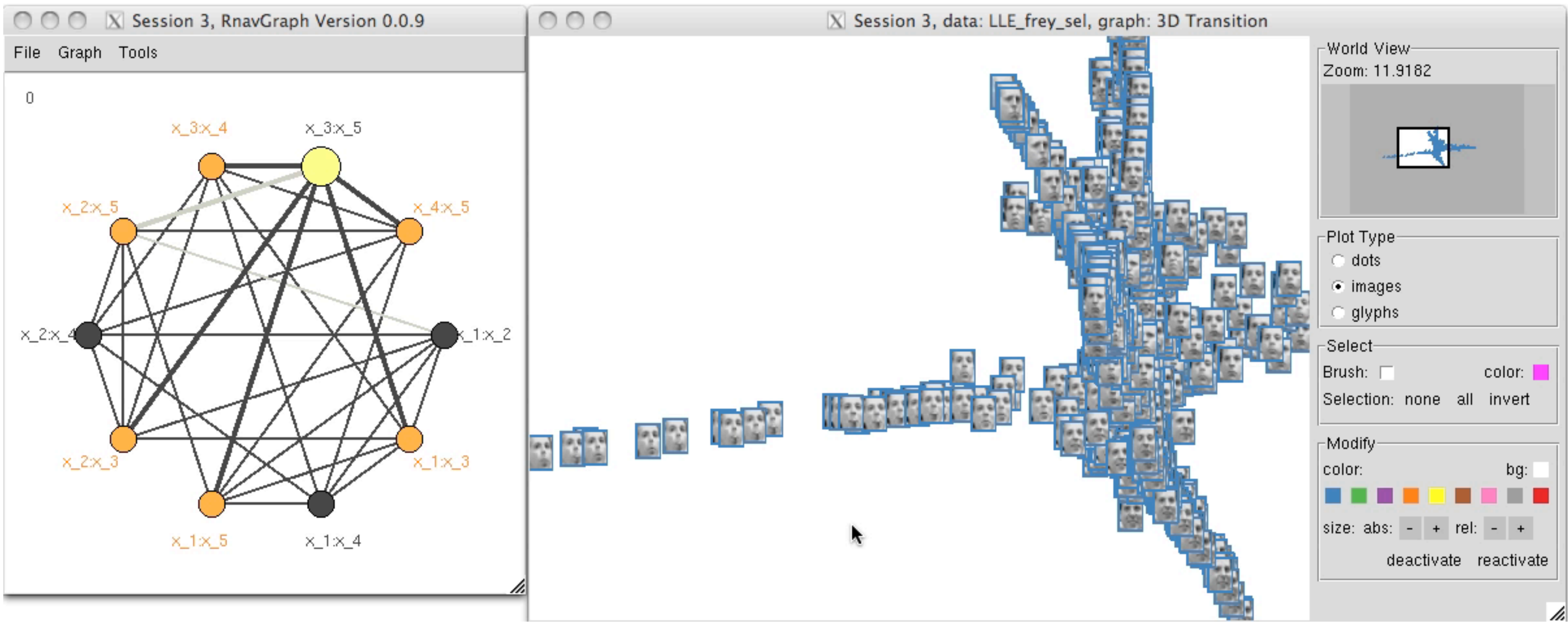
Lots of structure ... explored in 5d

Example: images



Lots of structure ... explored in 5d

Example: images

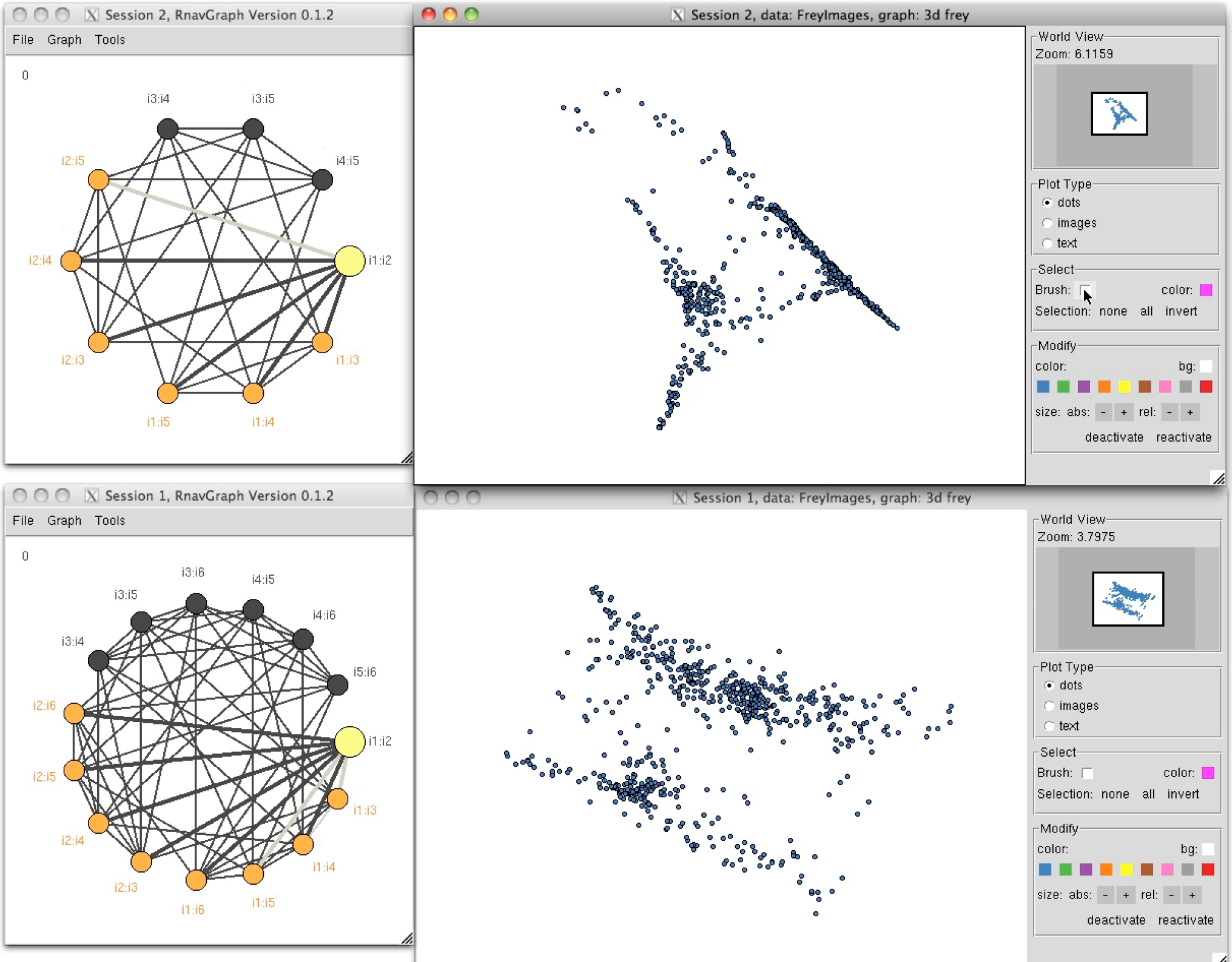


Lots of structure ... explored in 5d

Can link across NavGraph Sessions

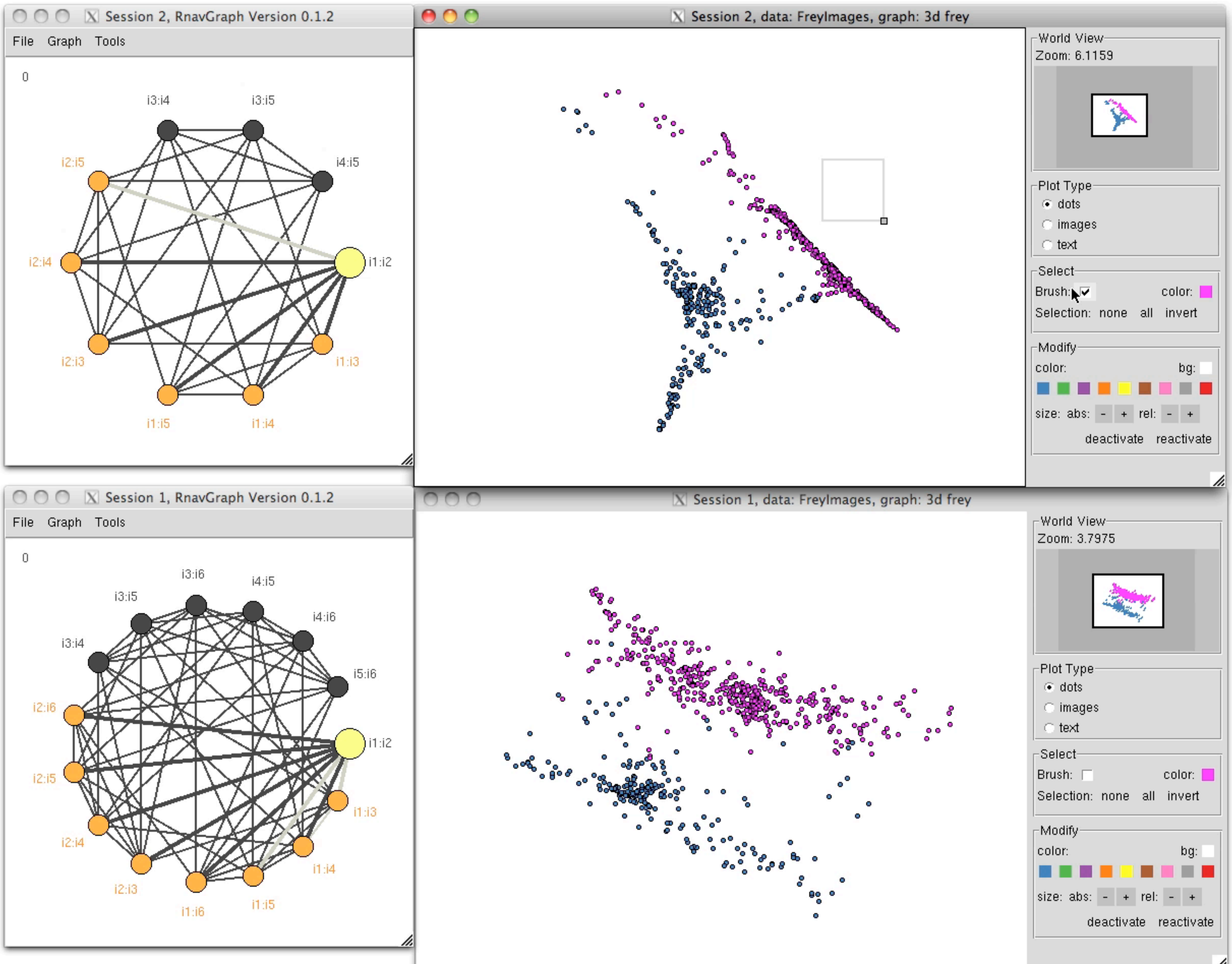
Here LLE and ISOMAP embeddings

Can link across NavGraph Sessions



Here LLE and ISOMAP embeddings

Can link across NavGraph Sessions

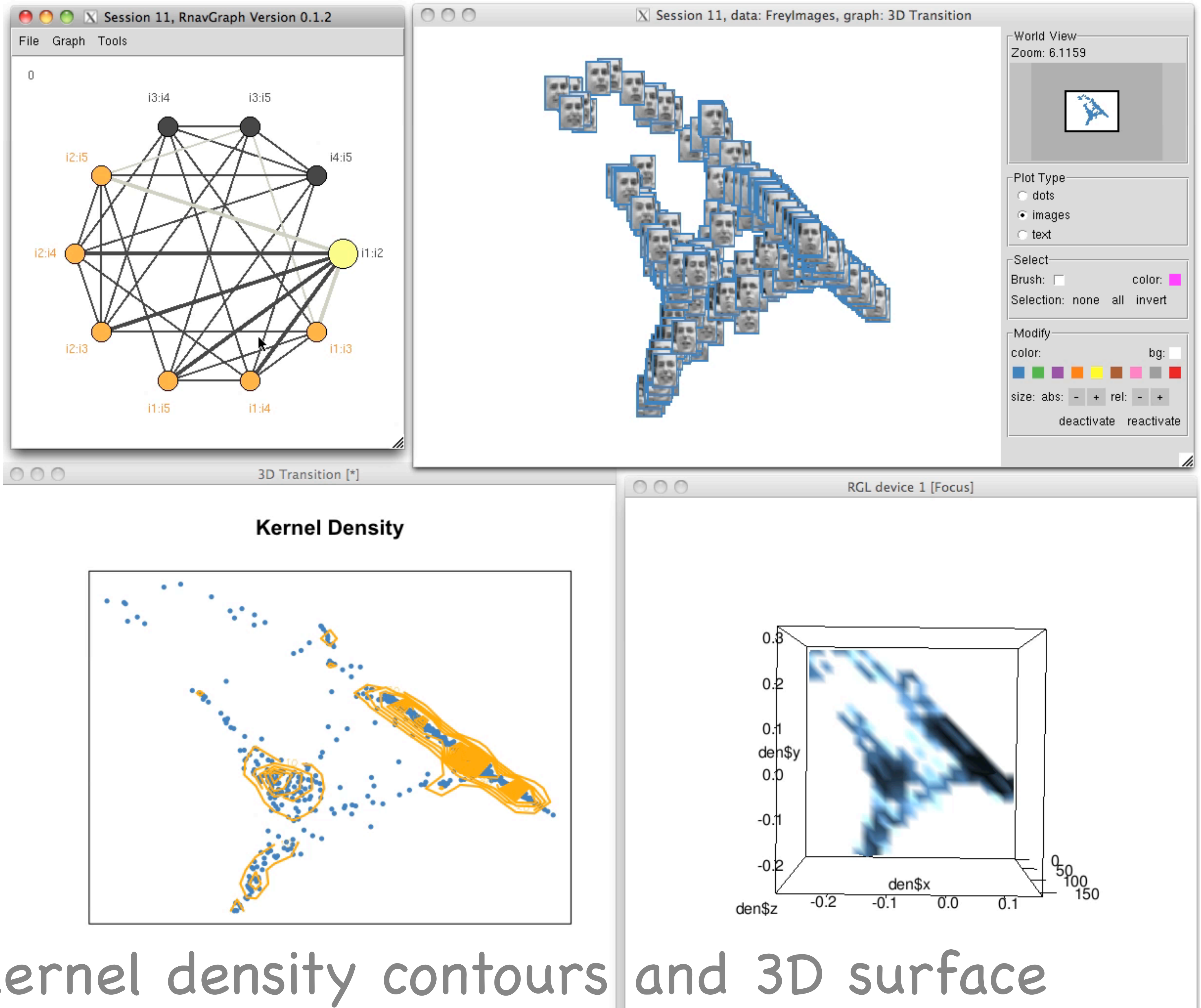


Here LLE and ISOMAP embeddings

Multiple visualizations

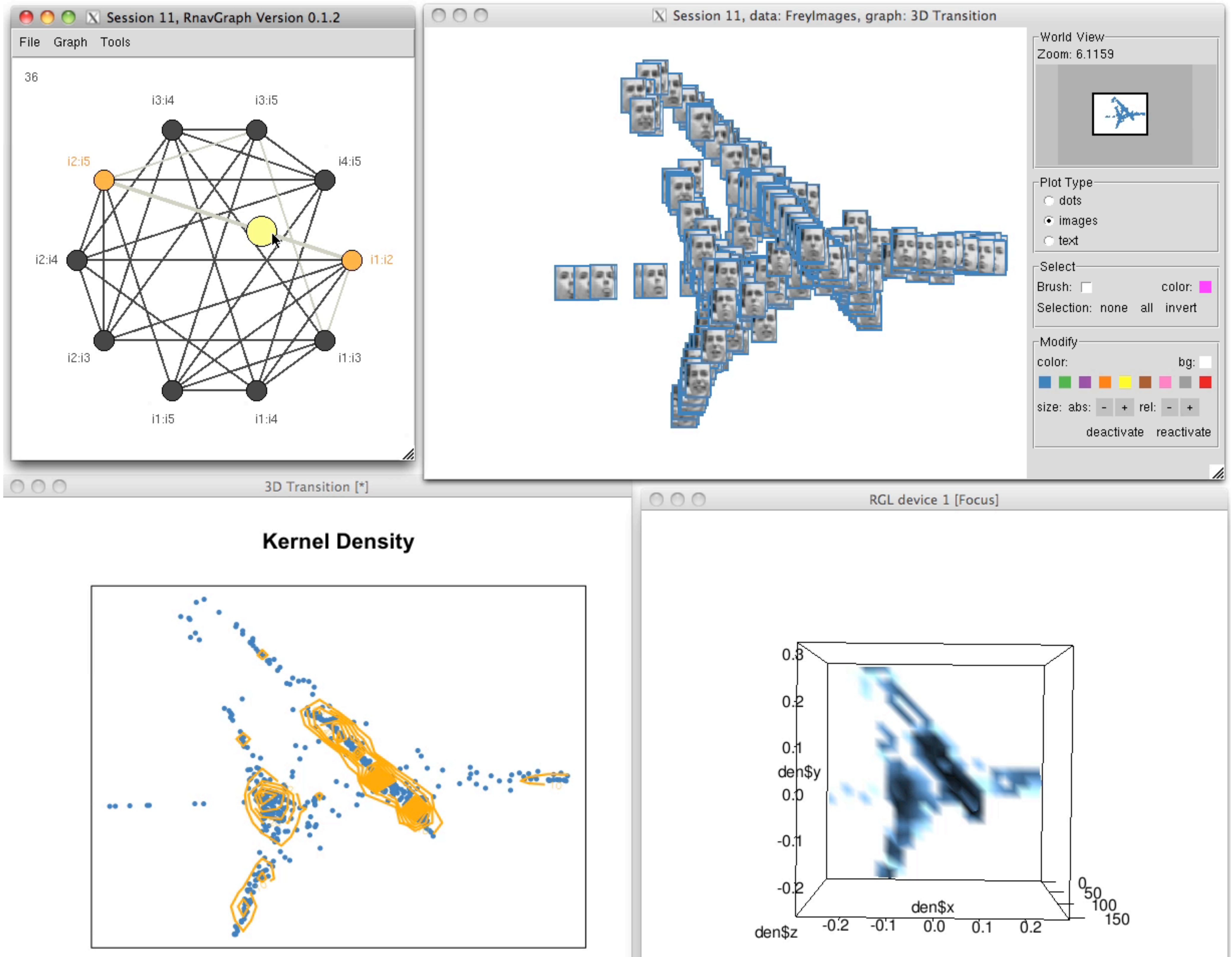
Kernel density contours and 3D surface

Multiple visualizations



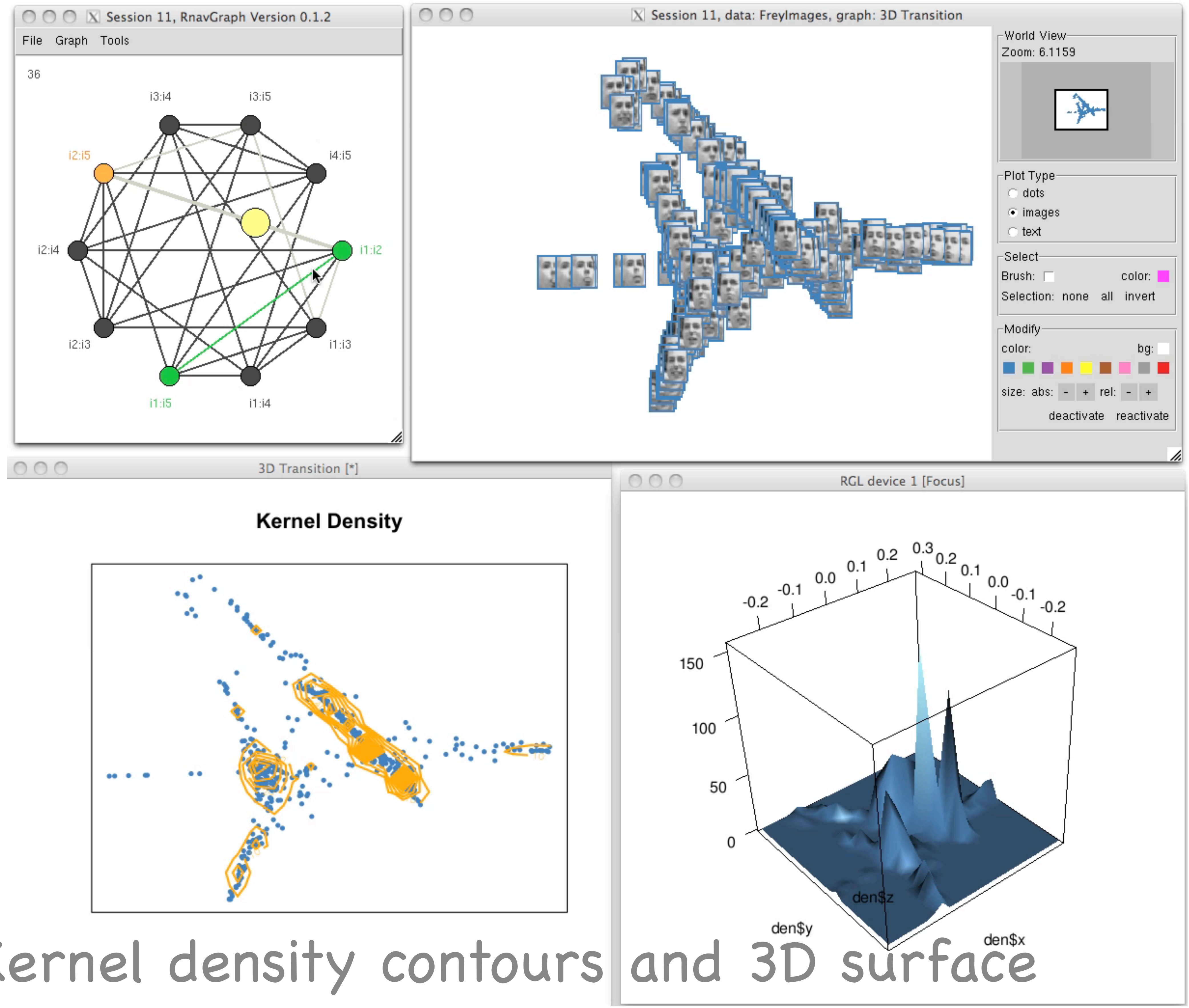
Kernel density contours and 3D surface

Multiple visualizations



Kernel density contours and 3D surface

Multiple visualizations



Kernel density contours and 3D surface

Summary

Graph theory structure

- organizes order of axes (e.g. radial, parallel, orthog.)
- use interesting orders (correlations, scagnostics, etc.)
- organizes ANY display order (e.g. multiple comparisons)
- graphs become maps to navigate high-dimensional space
- graph walks are low dimensional trajectories
- can focus on interesting trajectories
- capitalizes on visual ability
- graphs easily constructed; graph theory/algorithms exist

Summary

Try it yourself

- R packages:
  - `PairViz` Hurley & Oldford
  - `RnavGraph` Waddell & Oldford

Thank you

Thank you

Questions?

Papers

Hurley & Oldford:

- Graphs as navigational infrastructure for high dimensional data spaces (**Comp Stats 2011**)
- Pairwise display of high dimensional information via Eulerian tours and Hamiltonian decompositions (**JCGS, 2010**)
  - Eulerian tour algorithms for data visualization and the `PairViz` package (**Comp Stats 2011**)
  - `PairViz` R package ... available on CRAN.

Oldford & Waddell:

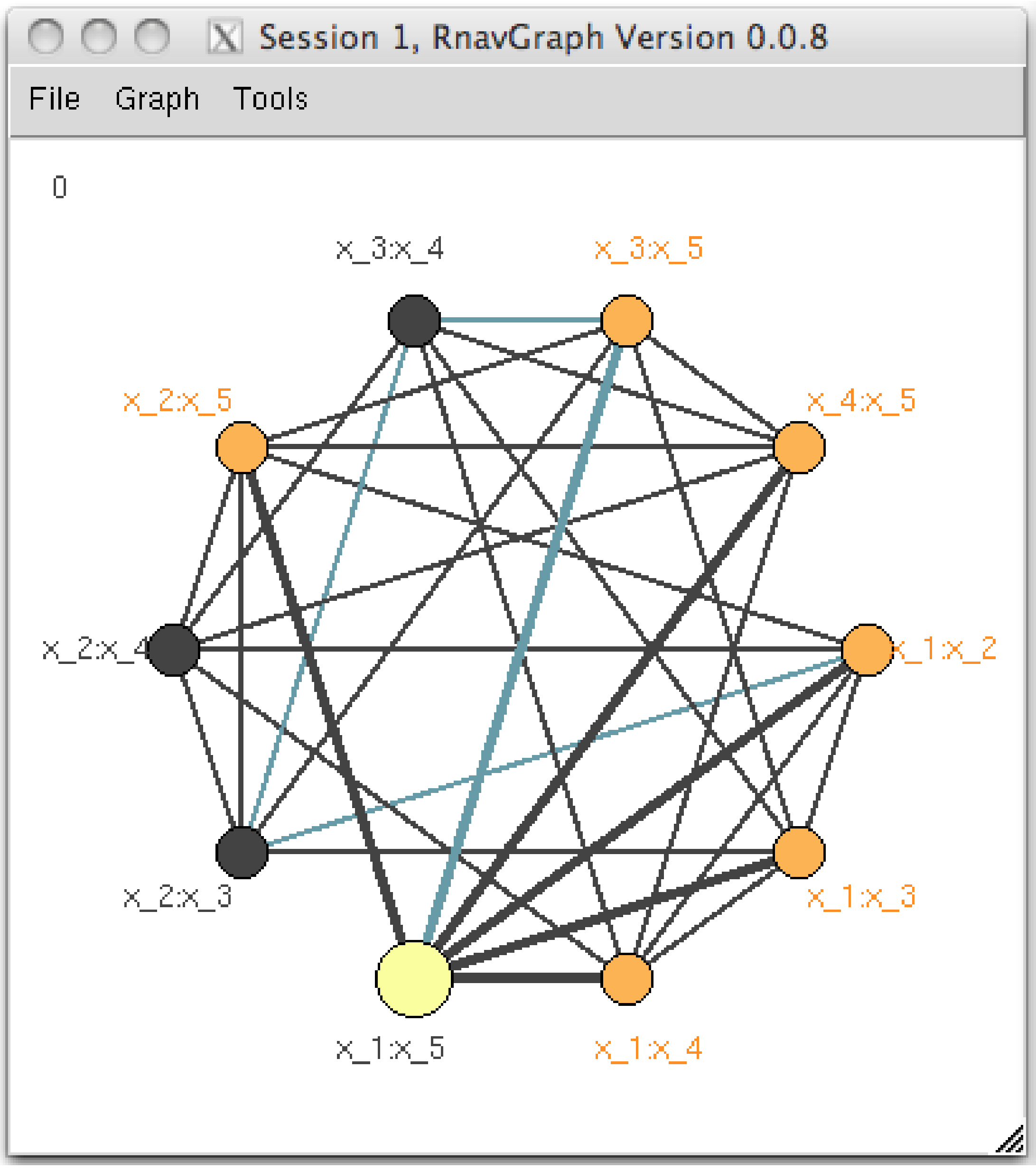
- Visual clustering of high-dimensional data by navigating low-dimensional spaces (**ISI Dublin, 2011**)
- RnavGraph: A visualization tool for navigating through high dimensional data (**ISI Dublin, 2011**)
- `RnavGraph` R package ... available on CRAN

Oldford & Zhou:

- Tree Ensemble Reduced Clustering via a Graph Algebraic Framework. submitted

Aside: 4d transitions

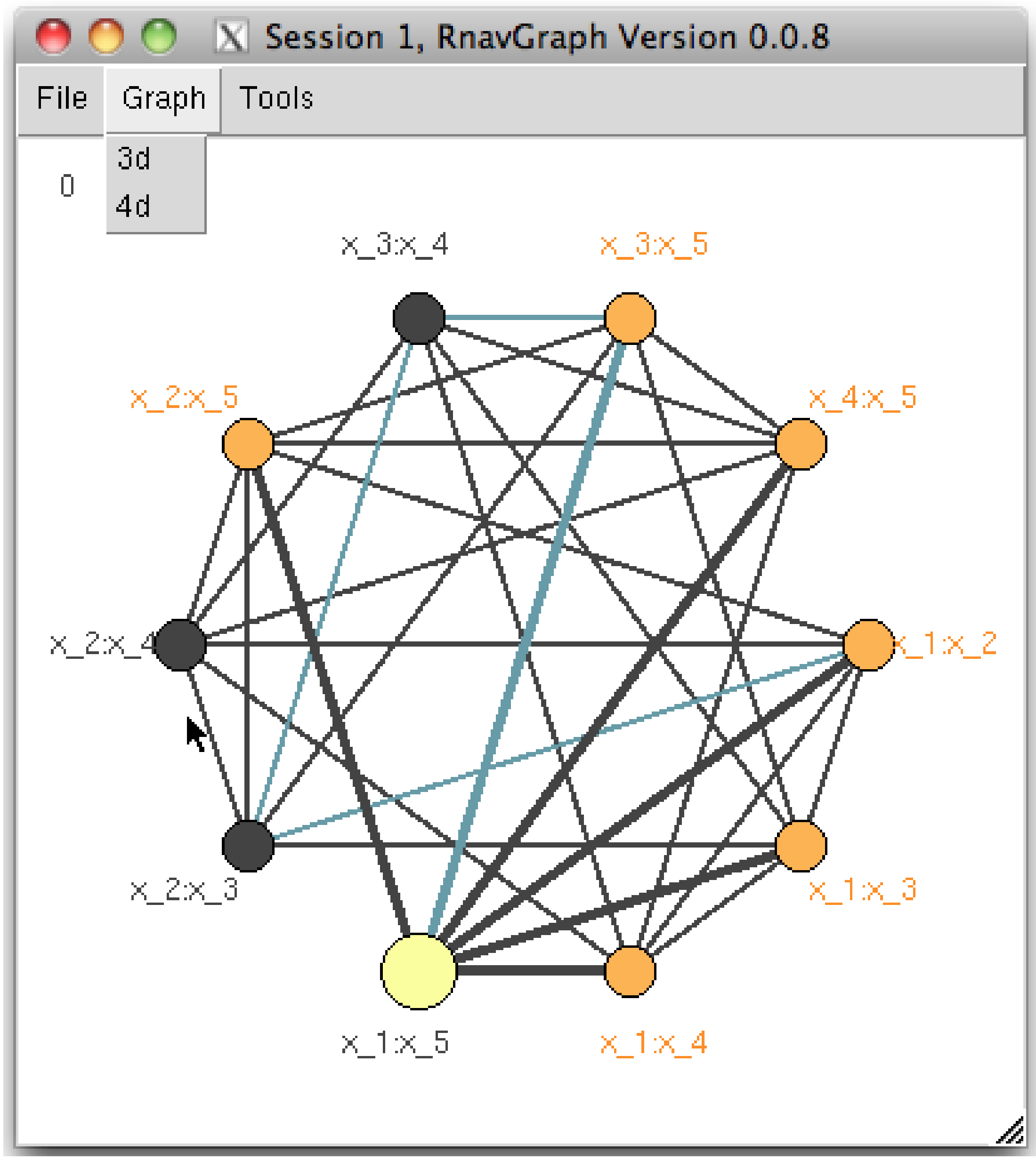
3d and 4d transition graphs



3d transition graph

Aside: 4d transitions

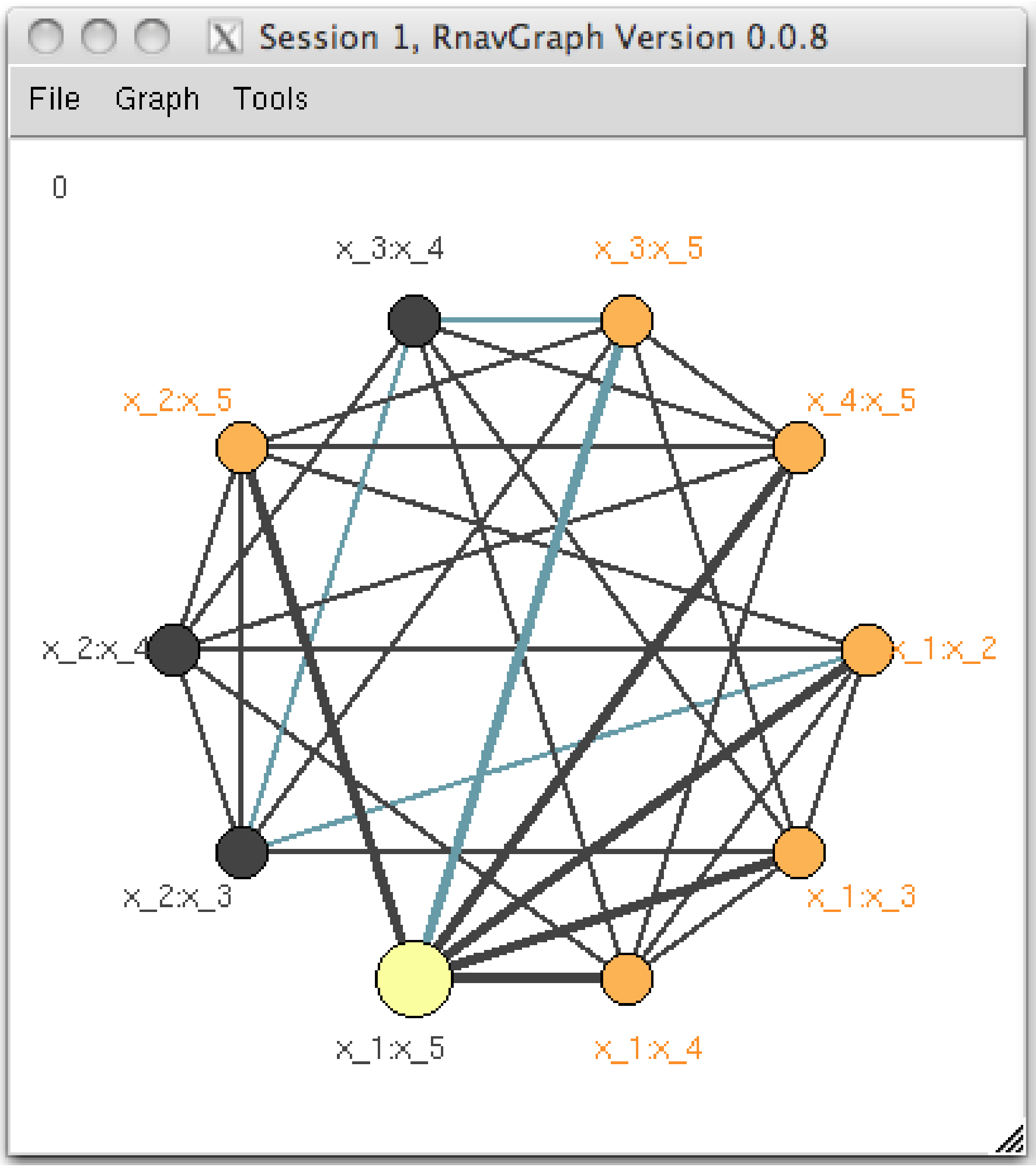
3d and 4d transition graphs



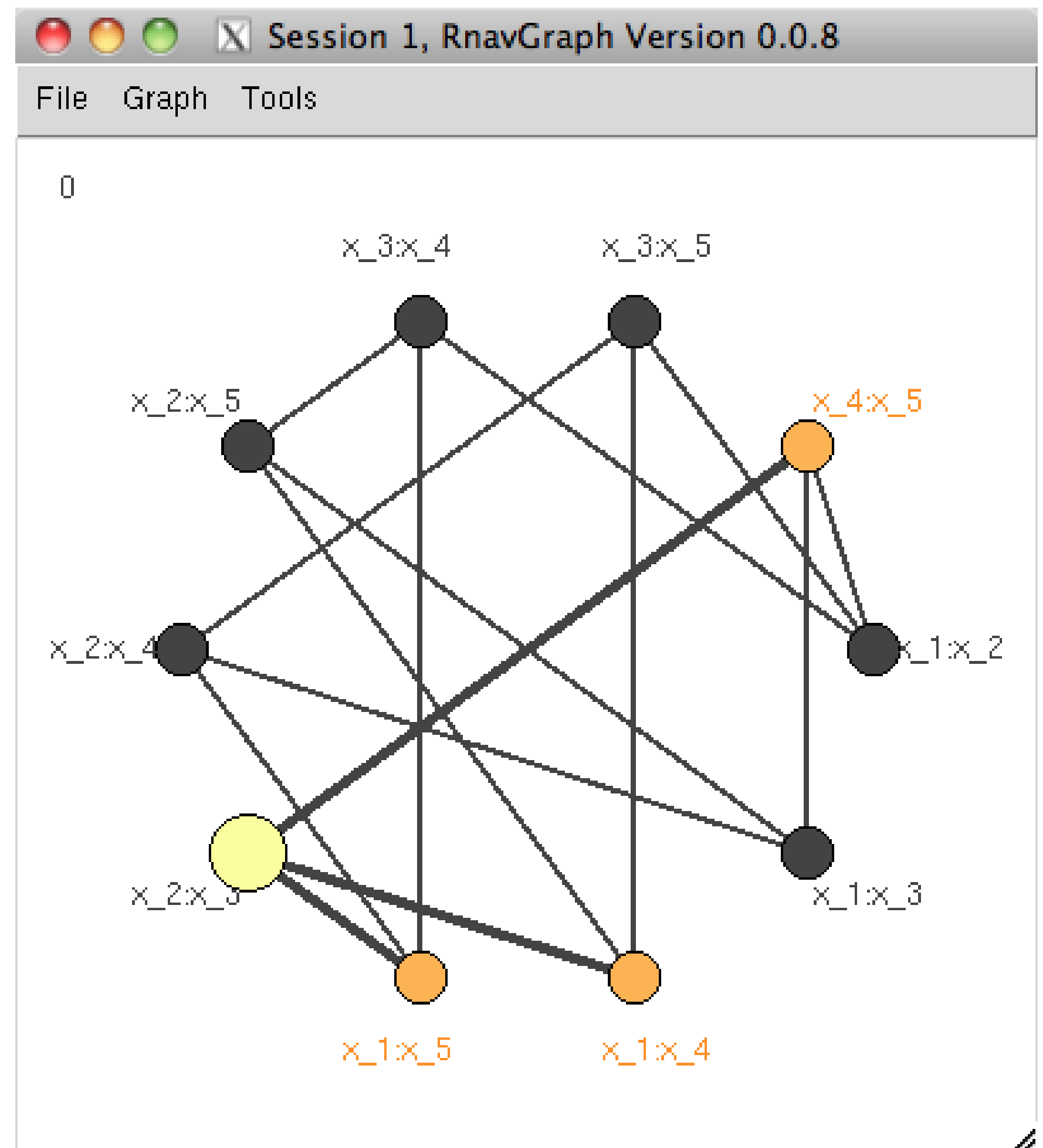
3d transition graph

Aside: 4d transitions

3d and 4d transition graphs



3d transition graph

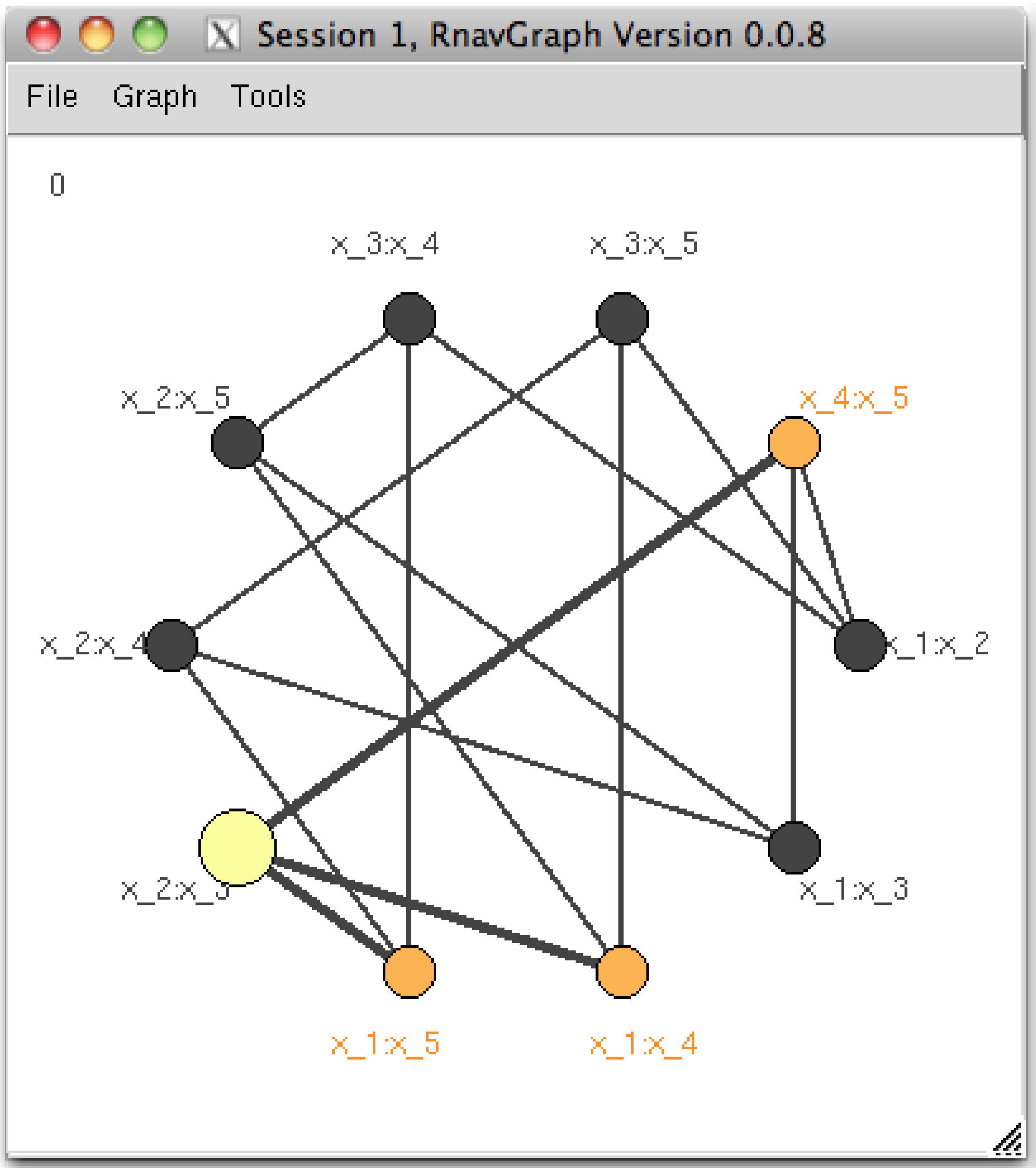


its complement

a 4d transition graph

Aside: 4d transitions

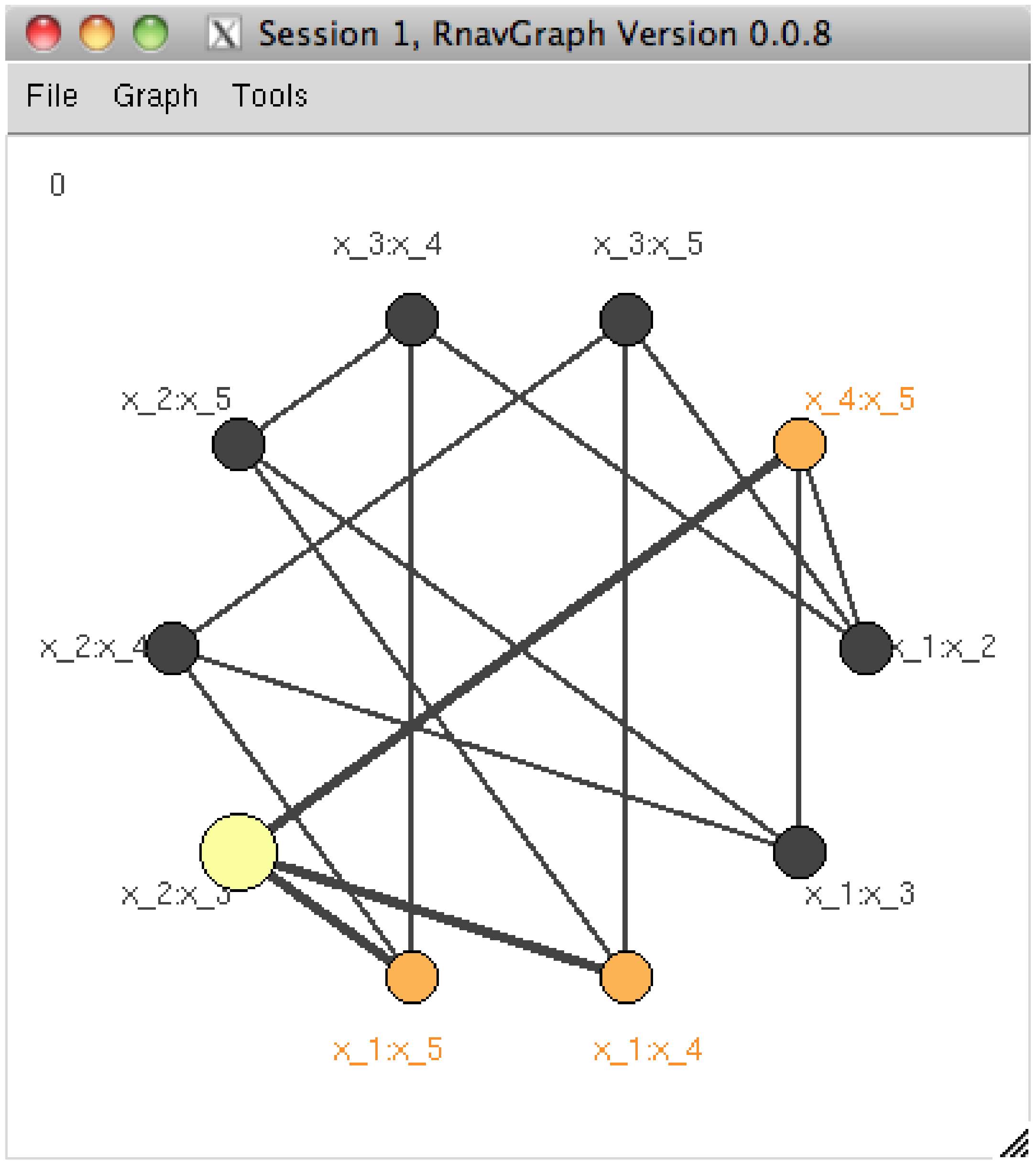
3d and 4d transition graphs



a 4d transition graph

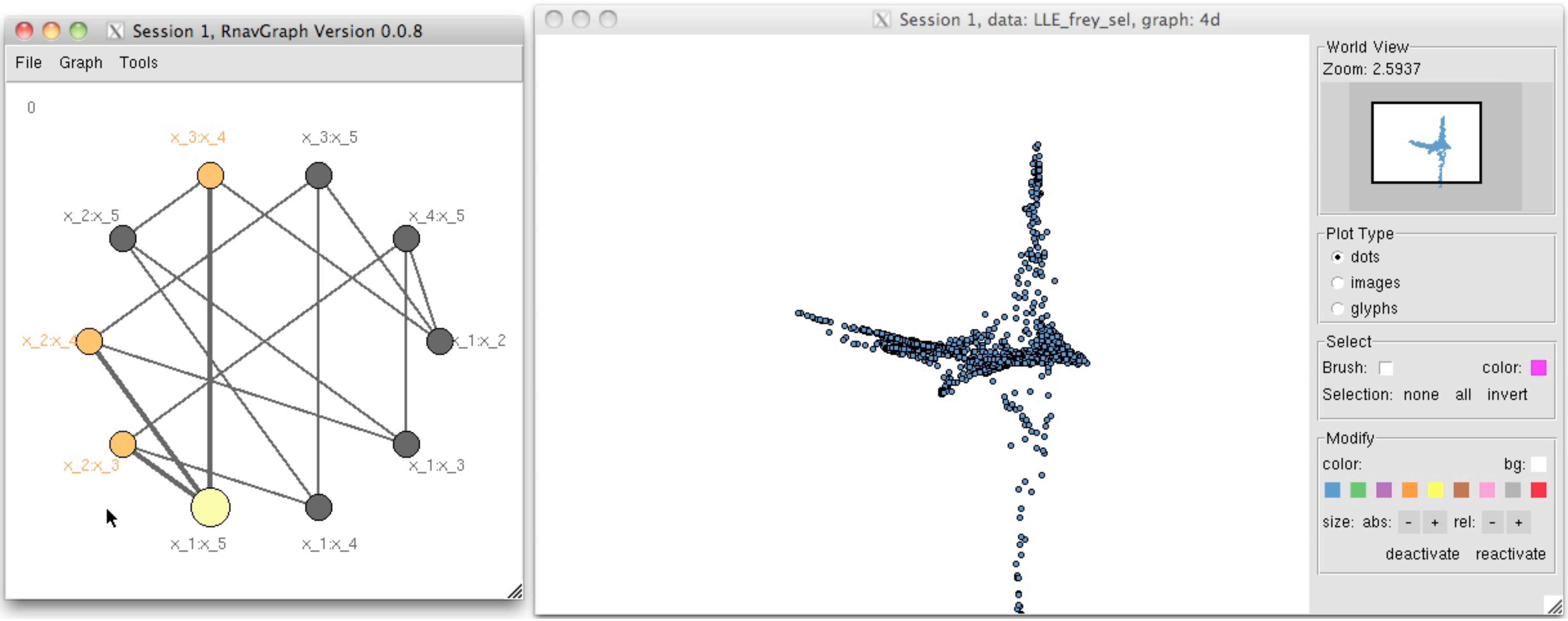
Aside: 4d transitions

3d and 4d transition graphs



Aside: 4d transitions

3d and 4d transition graphs



4d navGraph

Observe the 4d transition

NOT a rigid rotation