New interactive visualization tools for exploring high dimensional data in R

Wayne Oldford & Adrian Waddell

UNIVERSITY OF
WATERLOO

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, ...

Challenge

Large p

- visually, we're constrained to small p

  ✦ locations: $p < 4$

- comprehension depends on only a few dimensions

  ... at a time

- need to connect many smaller dimensional views

Approach

exploratory

- interactively slice and dice
  - ✦ find interesting low dimensional structure
  - ✦ cut away uninteresting subsets
  - ✦ dynamic linking across data

- integrate with statistical methods and models
  - ✦ scagnostics, dimension reduction, model building, ...

- mix direct manipulation and scripting in analysis
  - ✦ whichever is more powerful/natural
  - ✦ extendability is essential

Loon R
demo uses

R Studio

... a whirlwind tour

These slides contain videos.

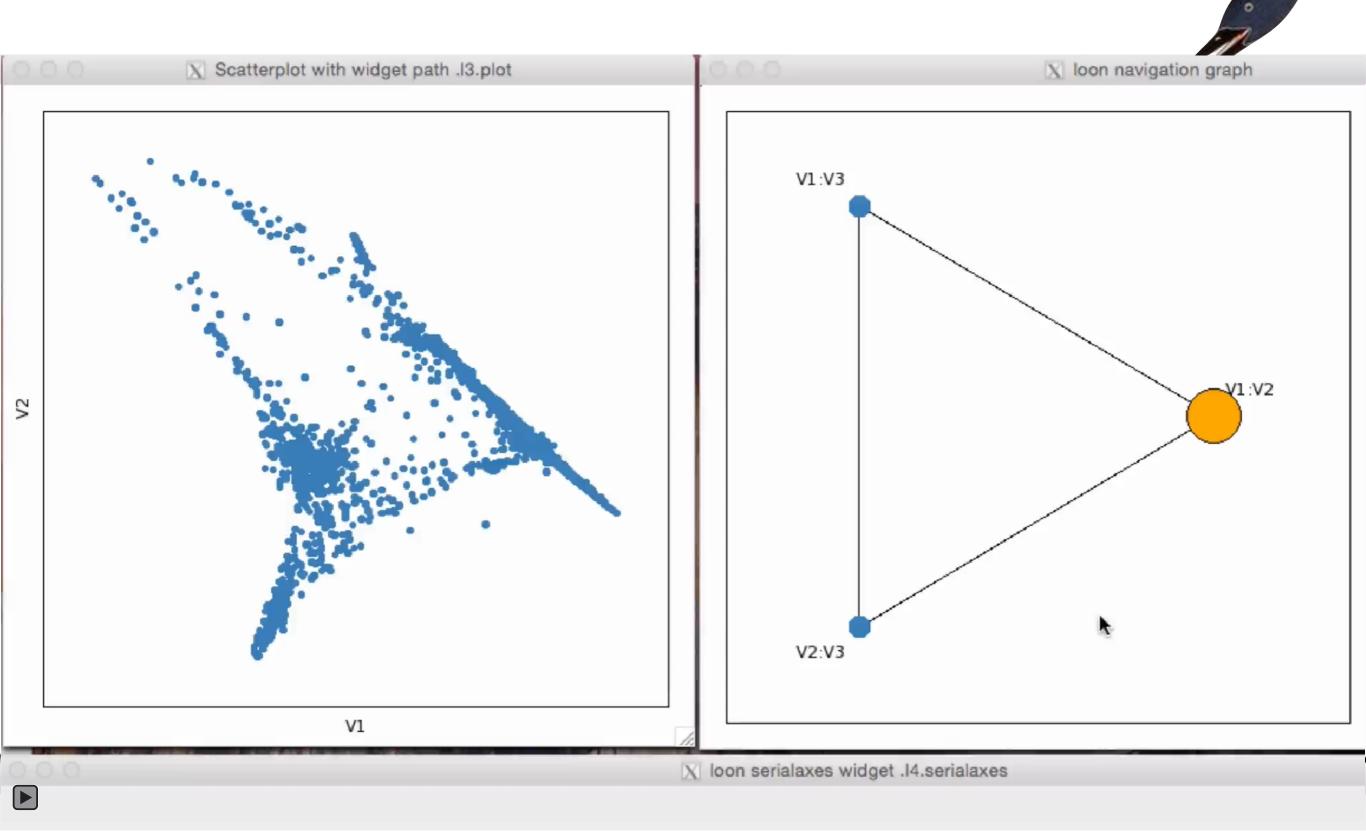If your PDF reader cannot display the videos on the following pages use the Adobe Acrobat reader:

get.adobe.com/reader
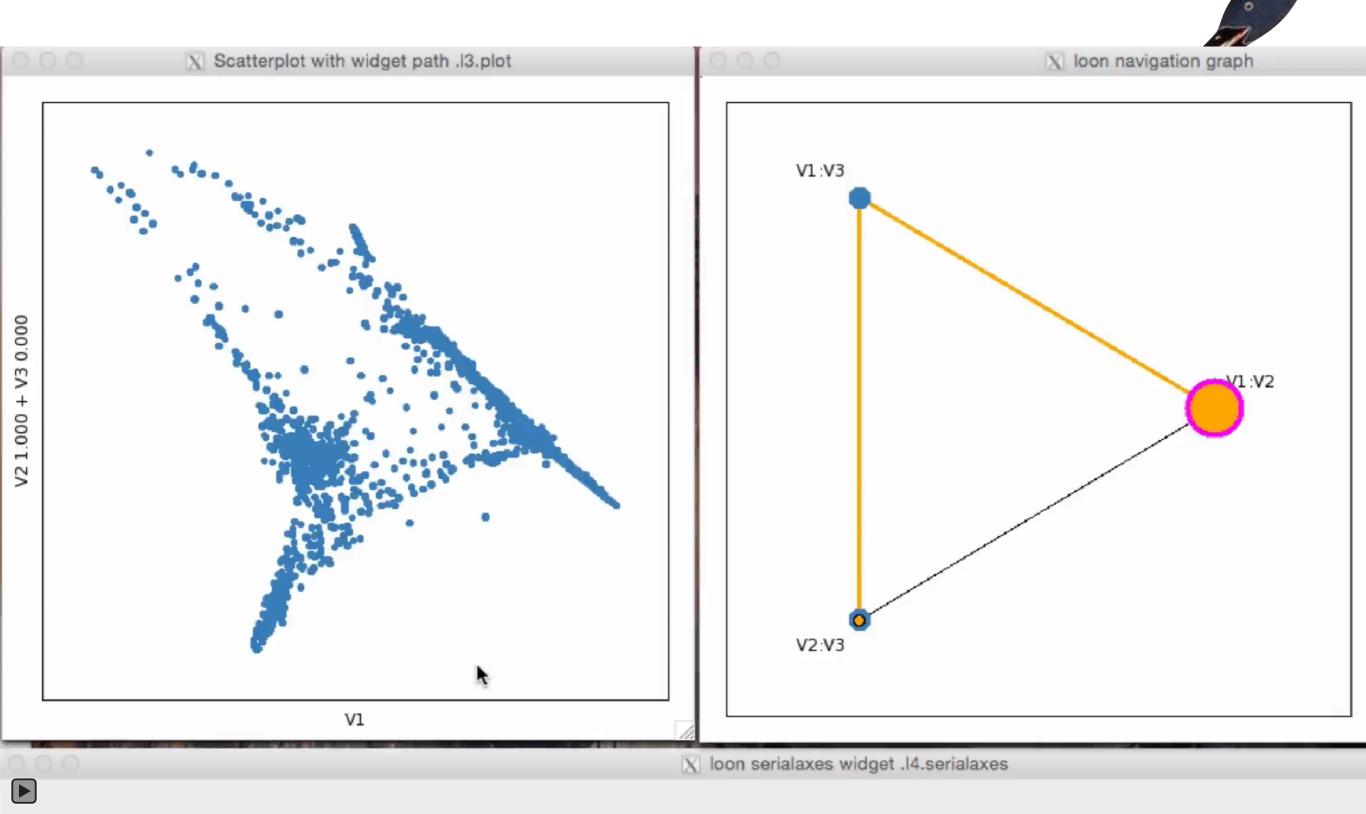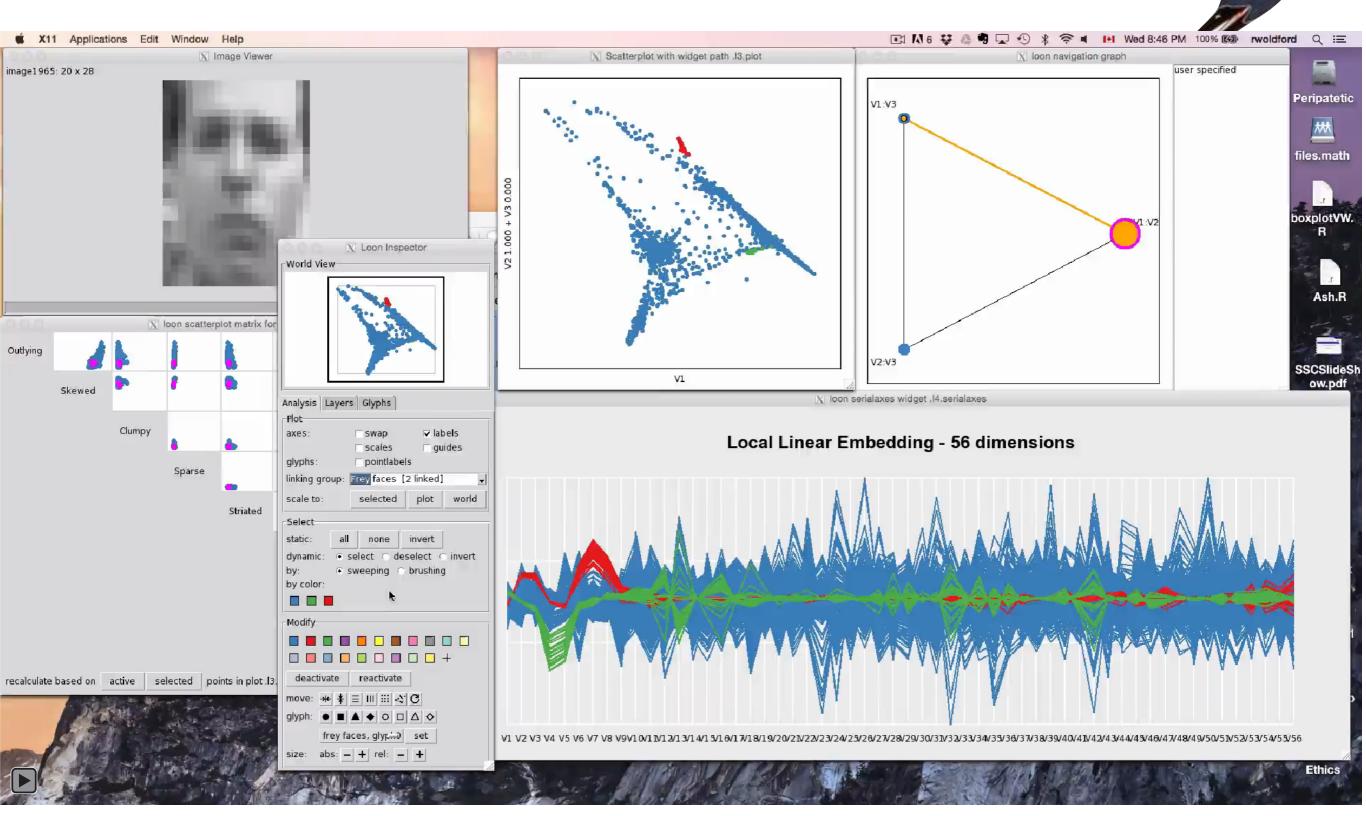
Example: images

loon

travel from one 2d space to another
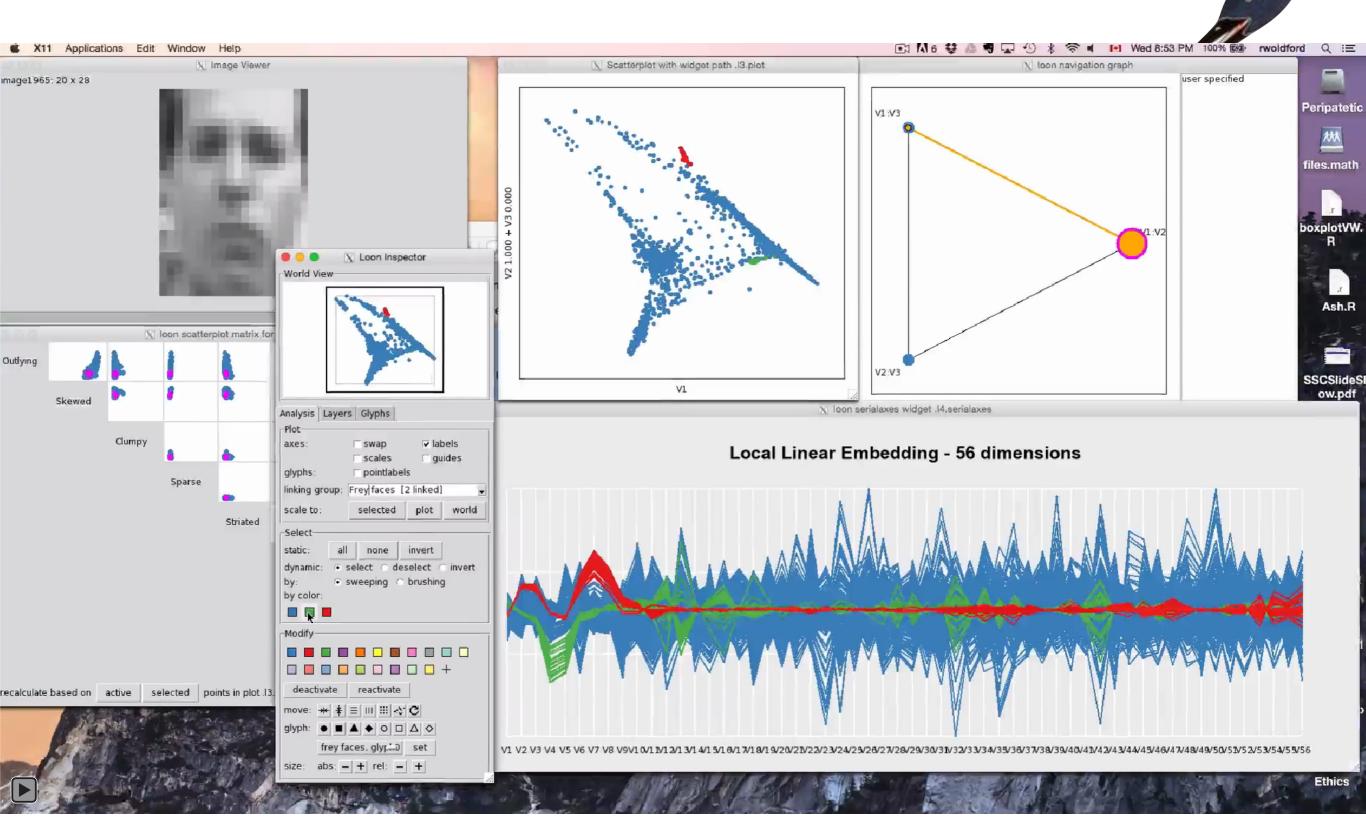
loon

link the scatterplot to a parallel coordinate plot

loon

Focus on the red group

loon

Focus on the green group

Challenge

Look at all scatterplots

- p ... overall dimensionality (images, p=56 and 560)

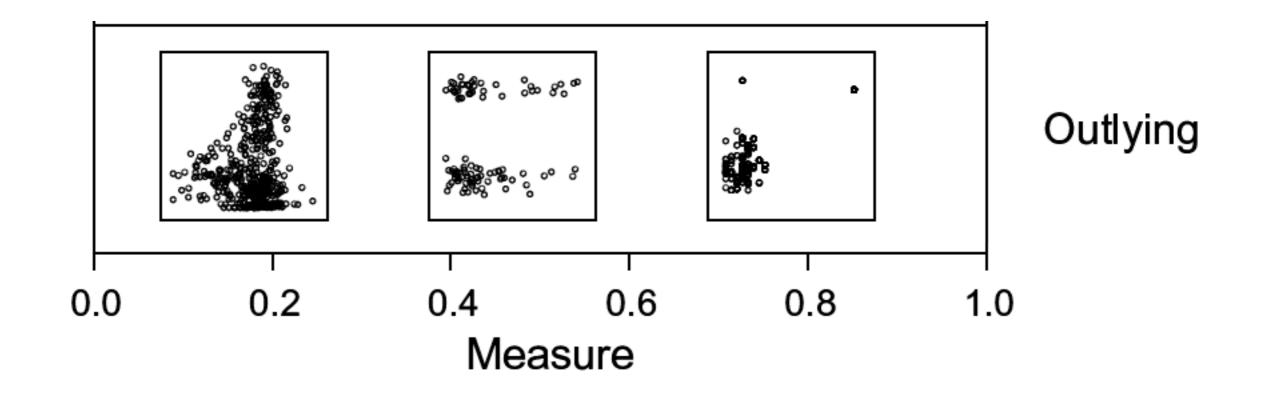| p	10	20	56	560
$\binom{p}{2}$	45	190	1,540	156,520

Need to look at the most interesting ones

Interesting variable pairs
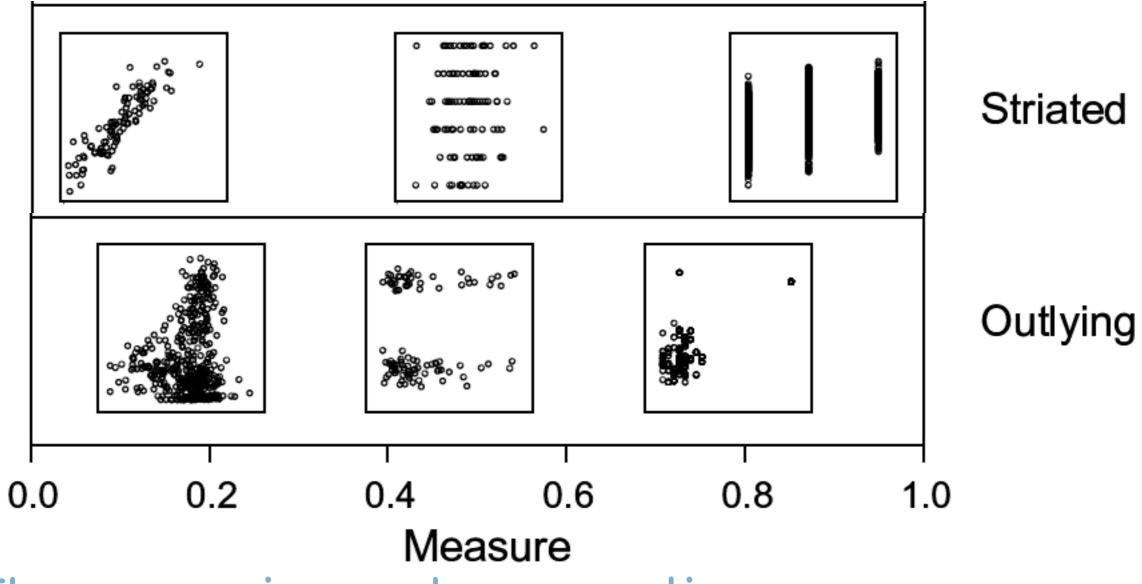
Cognostics (**Co**mputer aided dia**gnostics**)

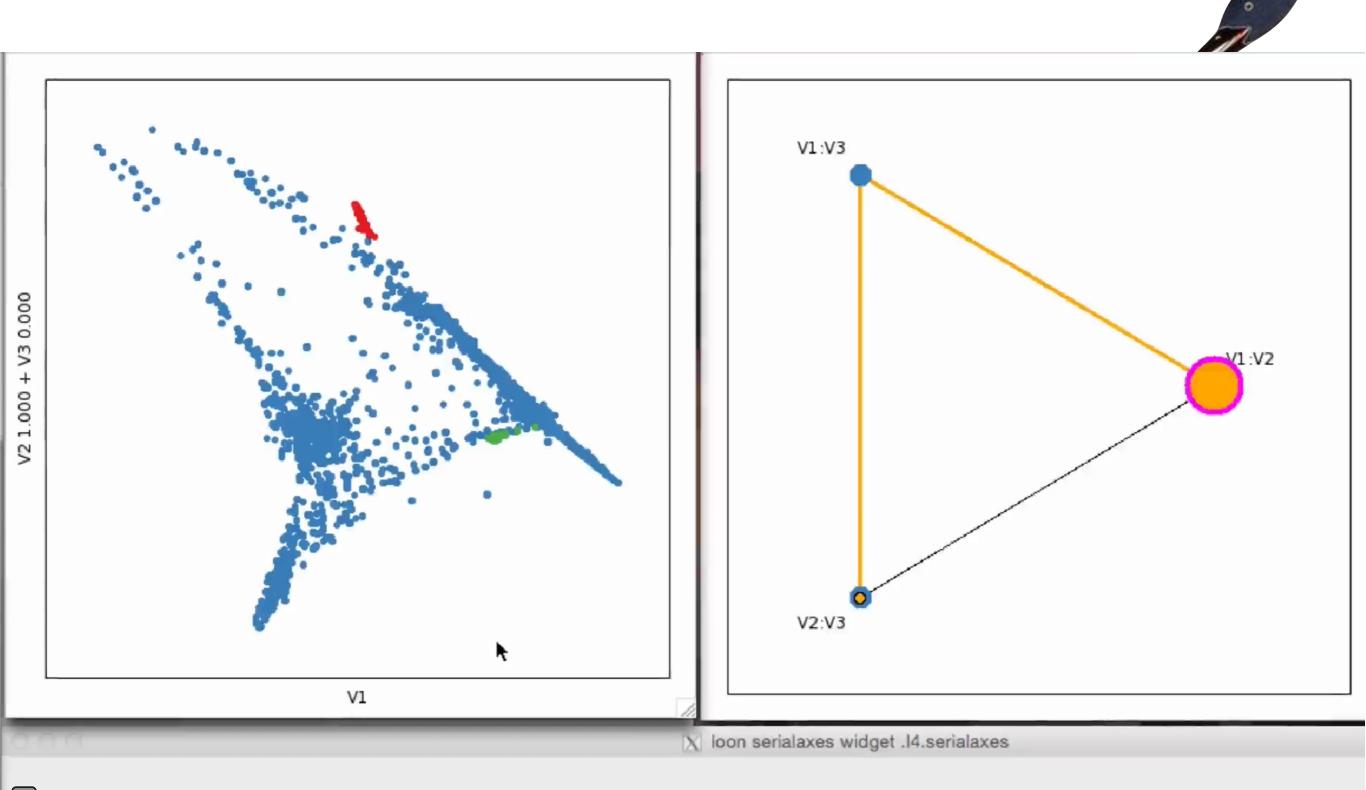Scagnostics ... **Sca**tterplot co**gnostics** (Tukey & Tukey, 1985)
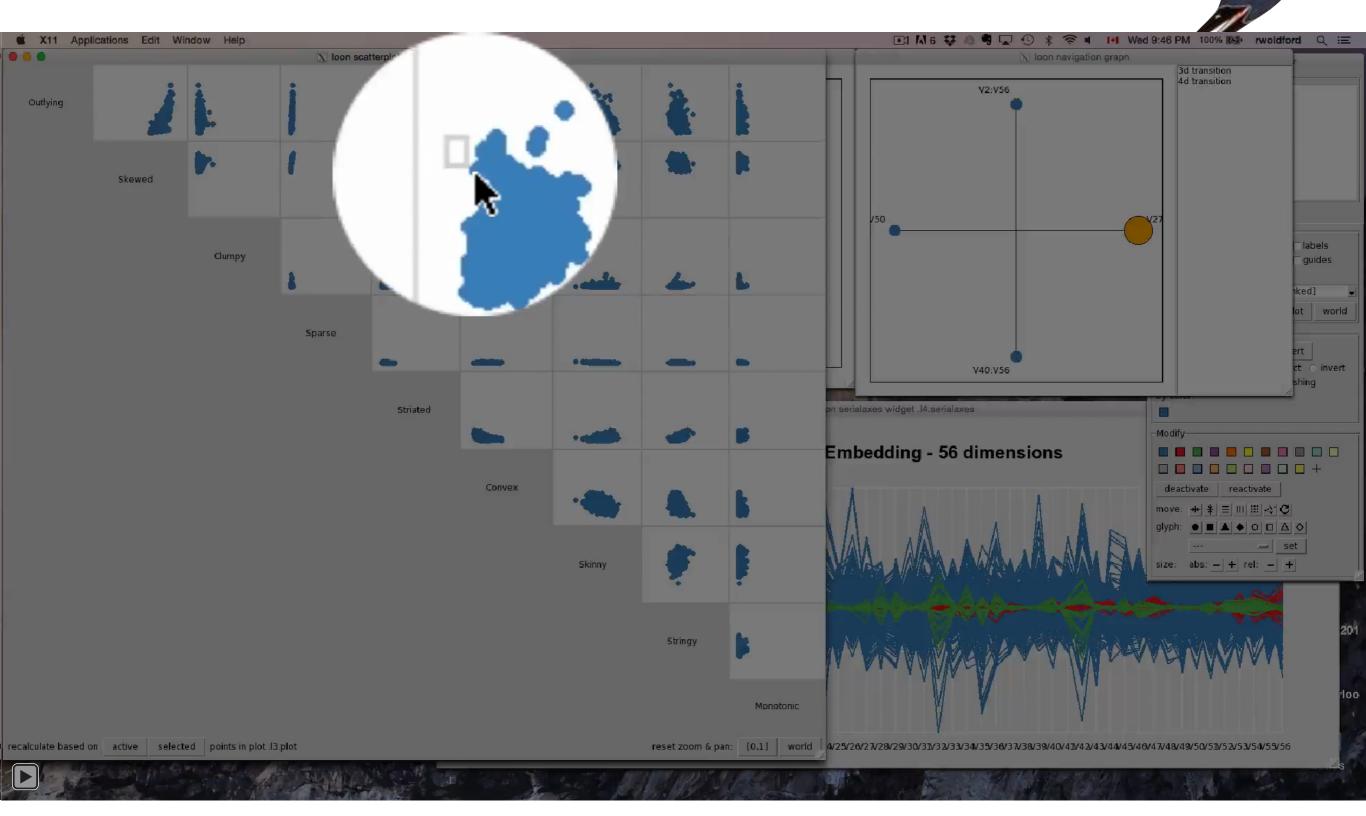
Wilkinson et al (2006+) provide such measures

Interesting node pairs

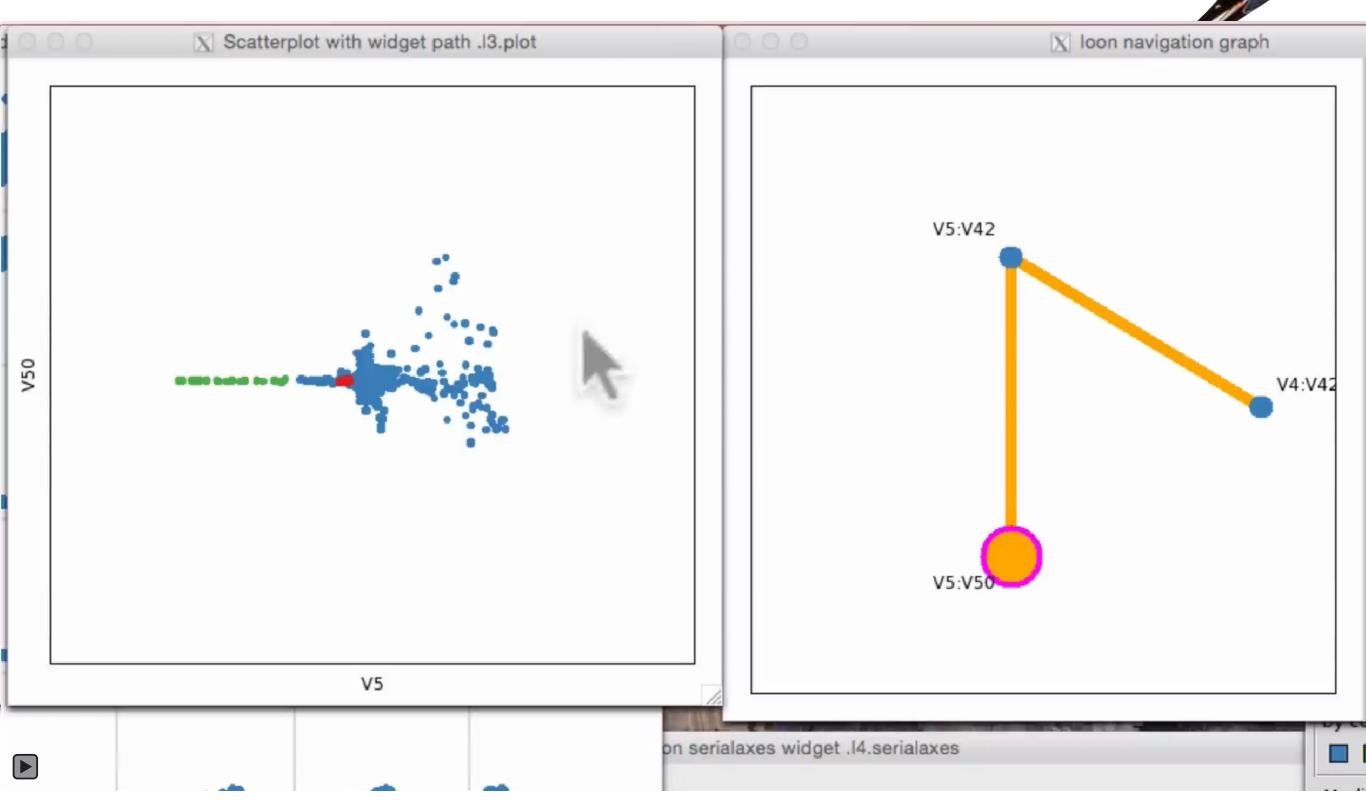There are nine such scagnostics measures

loon

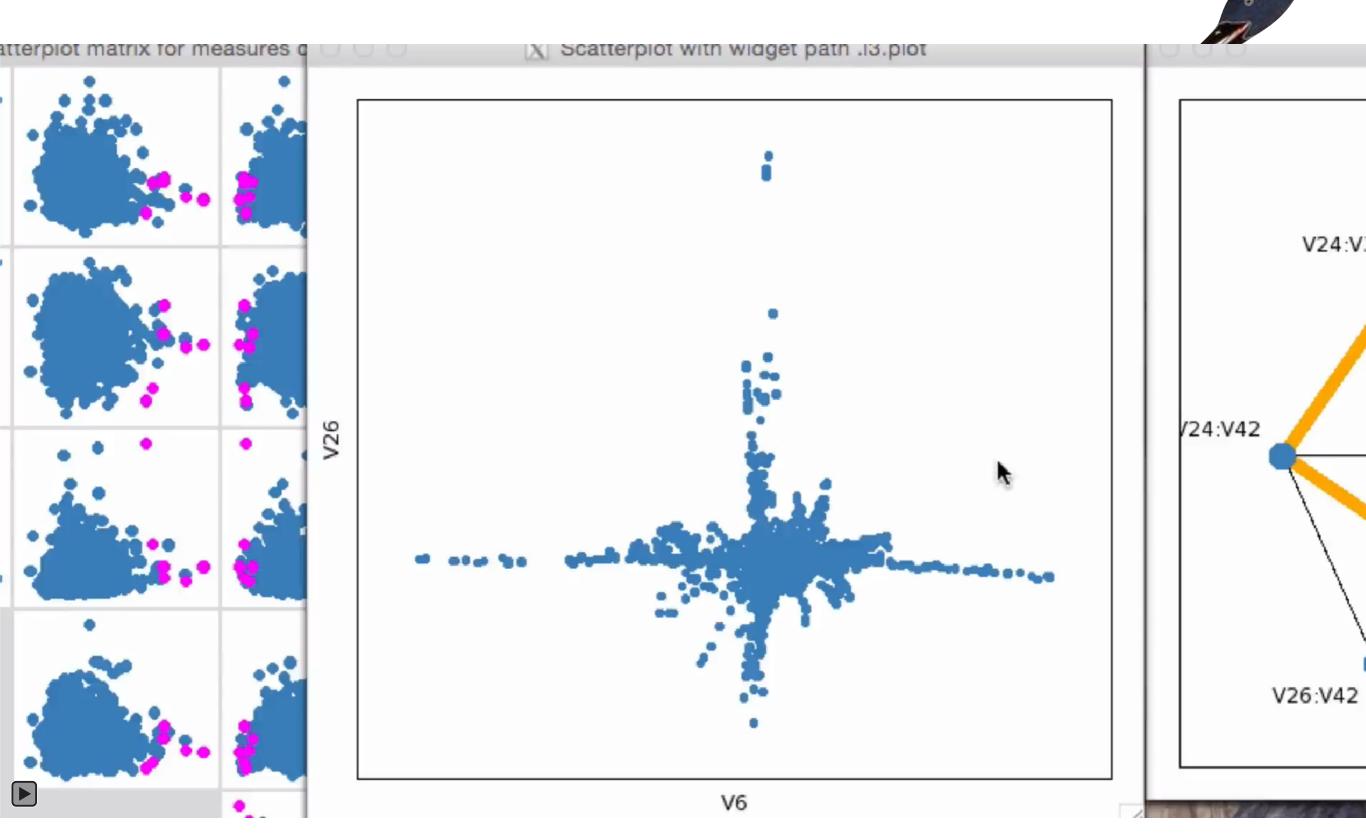*every point in the scatterplot matrix is a scatterplot!*

loon

high outlying and low convexity

loon

focus on structure of a subset

loon

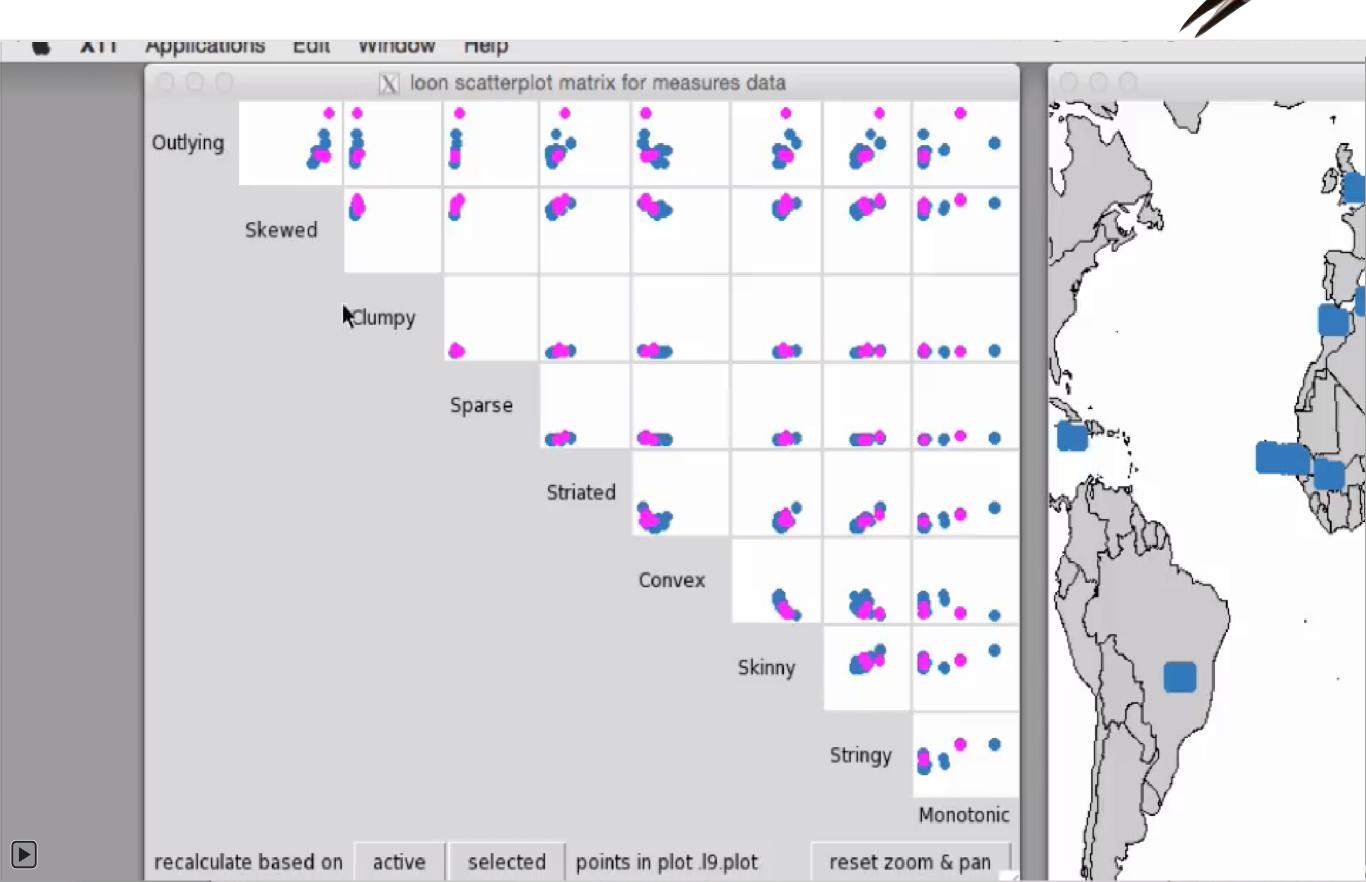*focus on structure of a subset*

Geographical ethnomusicology

1,059 tracks covering 33 countries/areas

traditional, ethnic, or `world' music only

no `western' music … too global

geographic location (lat and long)

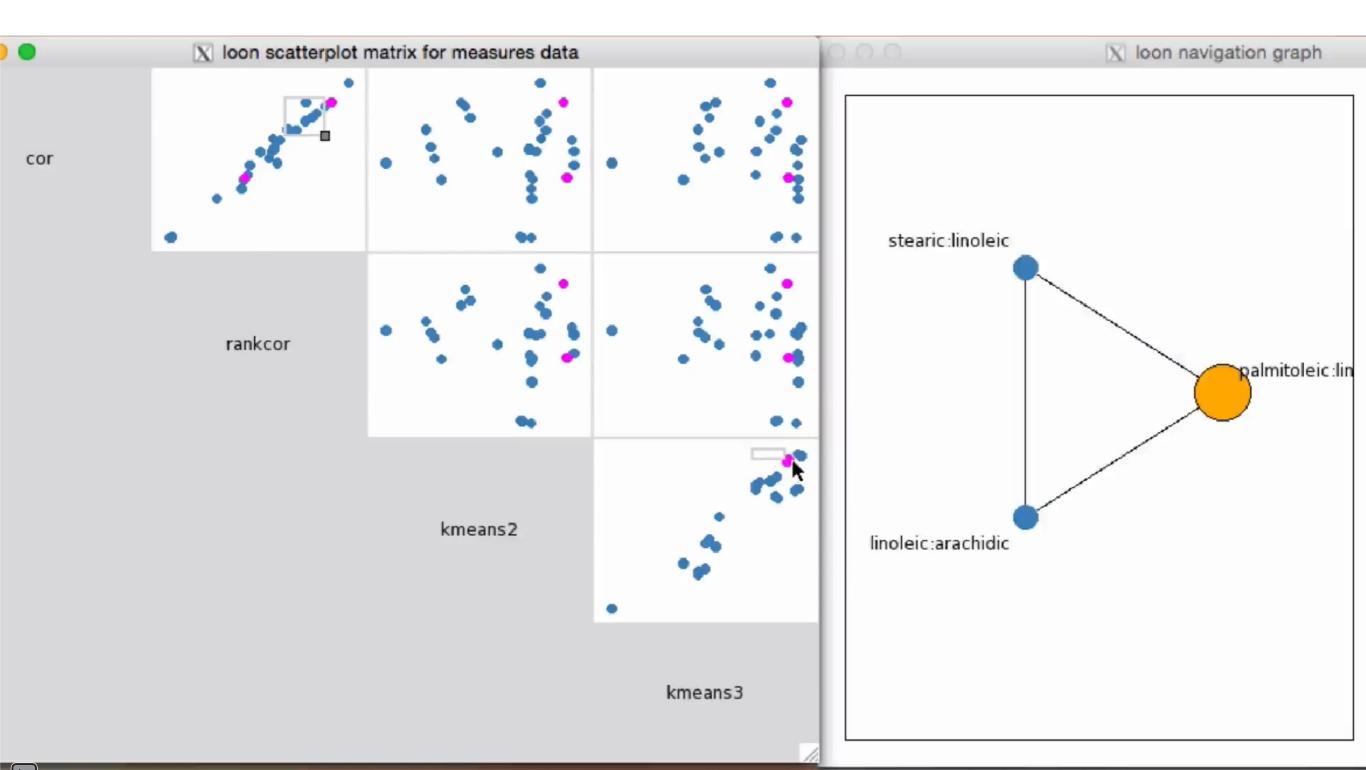68 audio descriptors from each track
(using MARSYAS default settings)

Fang Zhou, Claire Q and Ross. D. King
Predicting the Geographical Origin of Music,
ICDM, 2014

loon

scagnostics,   RDRToolbox
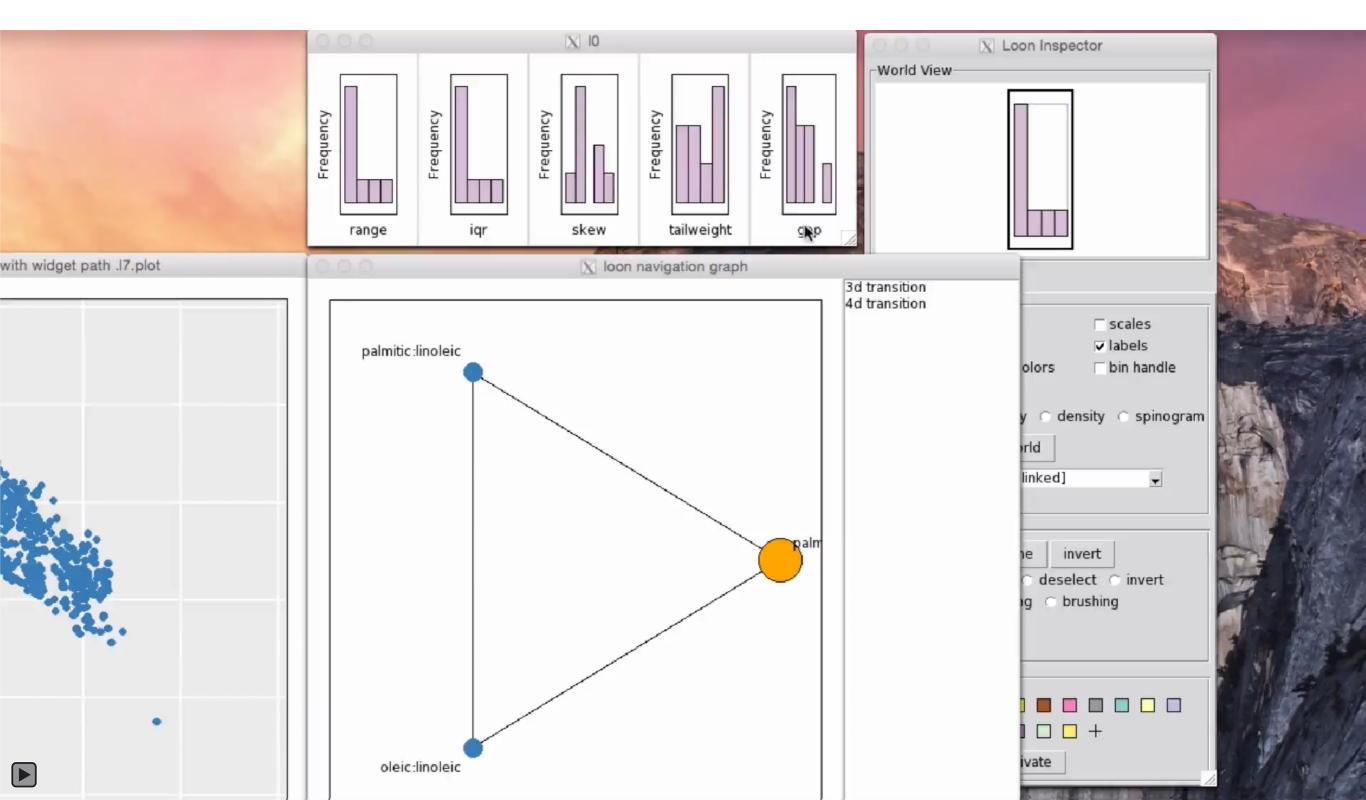
Build your own

User defined 2d-measures

Build your own

User defined 1d-measures

Summary

exploratory ... data diving

- interactively slice and dice

  - linkable interactive displays and inspector

  - deactivate/reactivate cases — focus on subsets

  - zoom, pan, ...

- integrate with statistical methods

  - apply to selected subsets and update displays

- mix direct manipulation and scripting in analysis

  - use R functionality and methods

  - Loon provides event hooks to call back to any R function

request a beta version:

navgraph.com/loon

exploratory ... data diving