Interactive Visual Analytics for Discovering Simpson's Paradox

Presenter
Chenguang (Shine) Xu
University of Oklahoma
chguxu@ou.edu

Sarah M. Brown
University of California, Berkeley
smb@sarahmbrown.org

Chris Weaver, Christan Grant
University of Oklahoma
{cweaver, cgrant}@ou.edu

OU Data Analytics Lab
https://oudalab.github.io

Outline

- Motivation

- What is SP

- Why detect SP

- How to detect SP

- Summary

Motivation

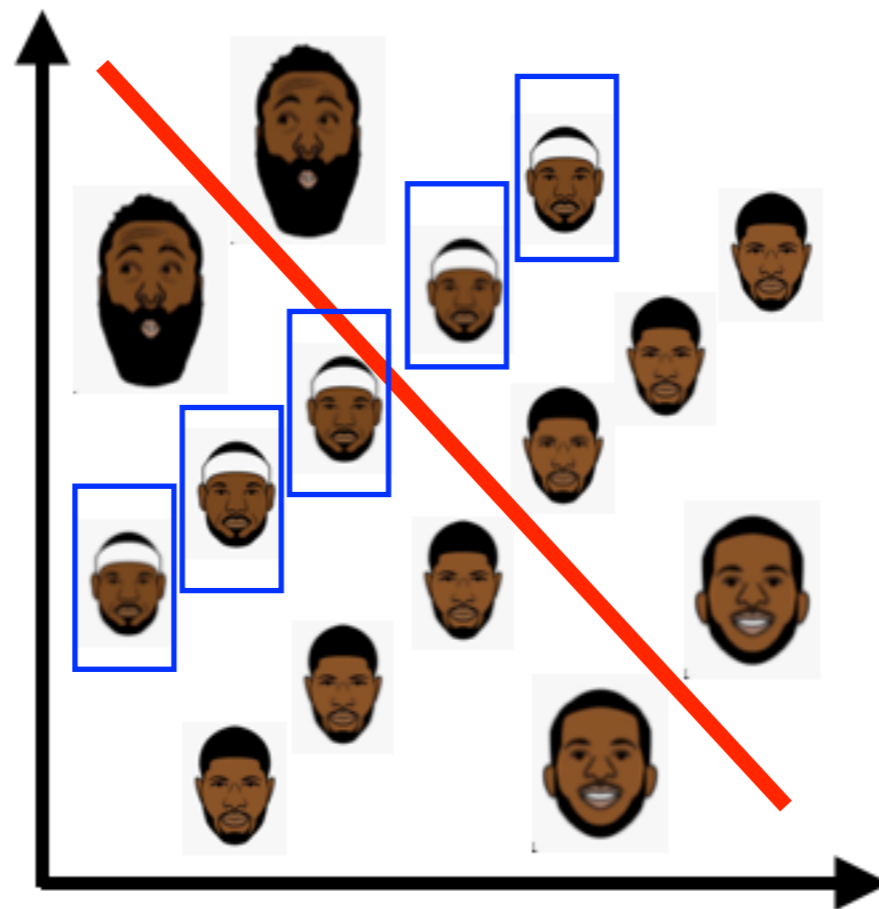- Fairness forensics, investigate possible bias in data
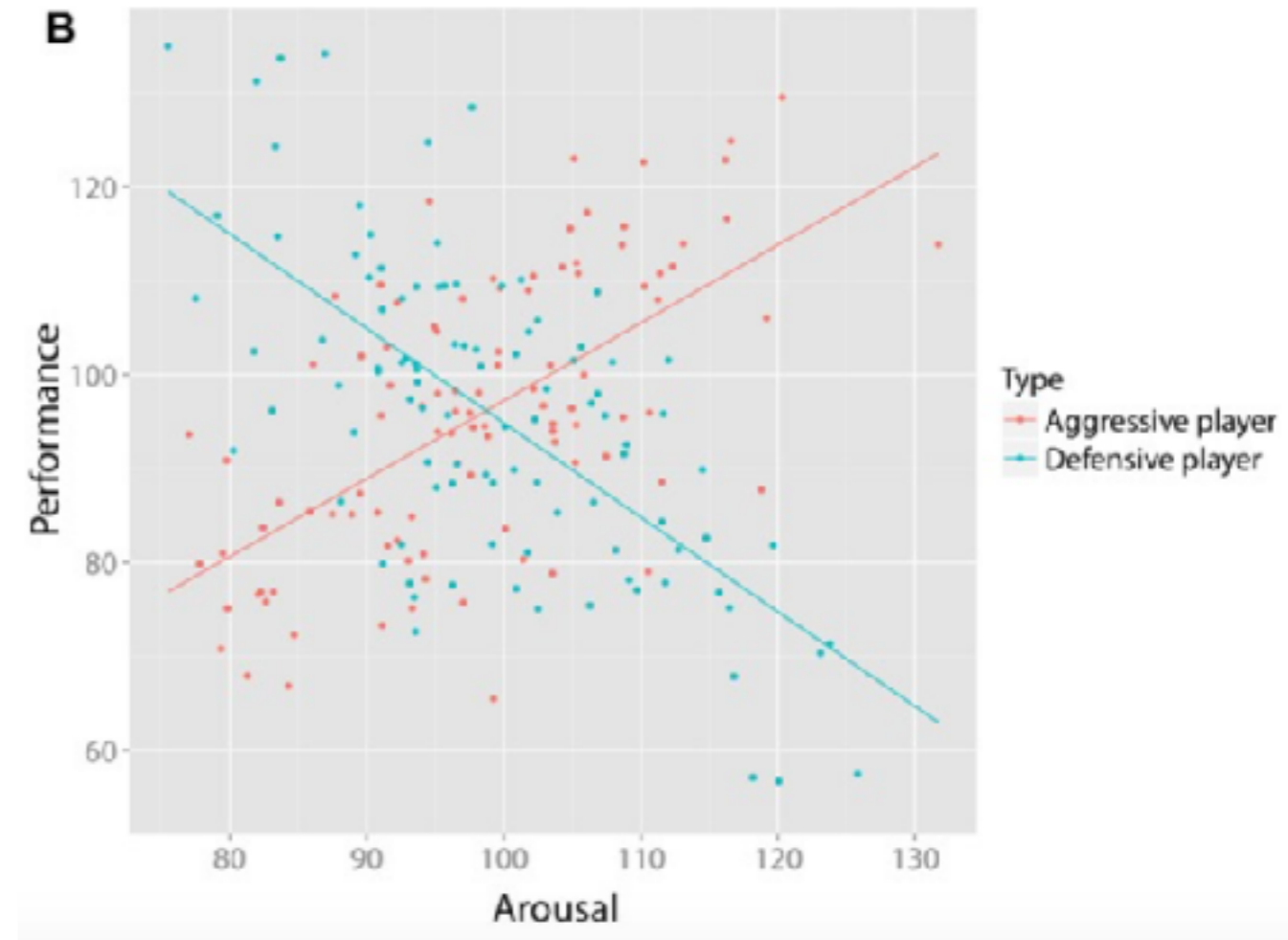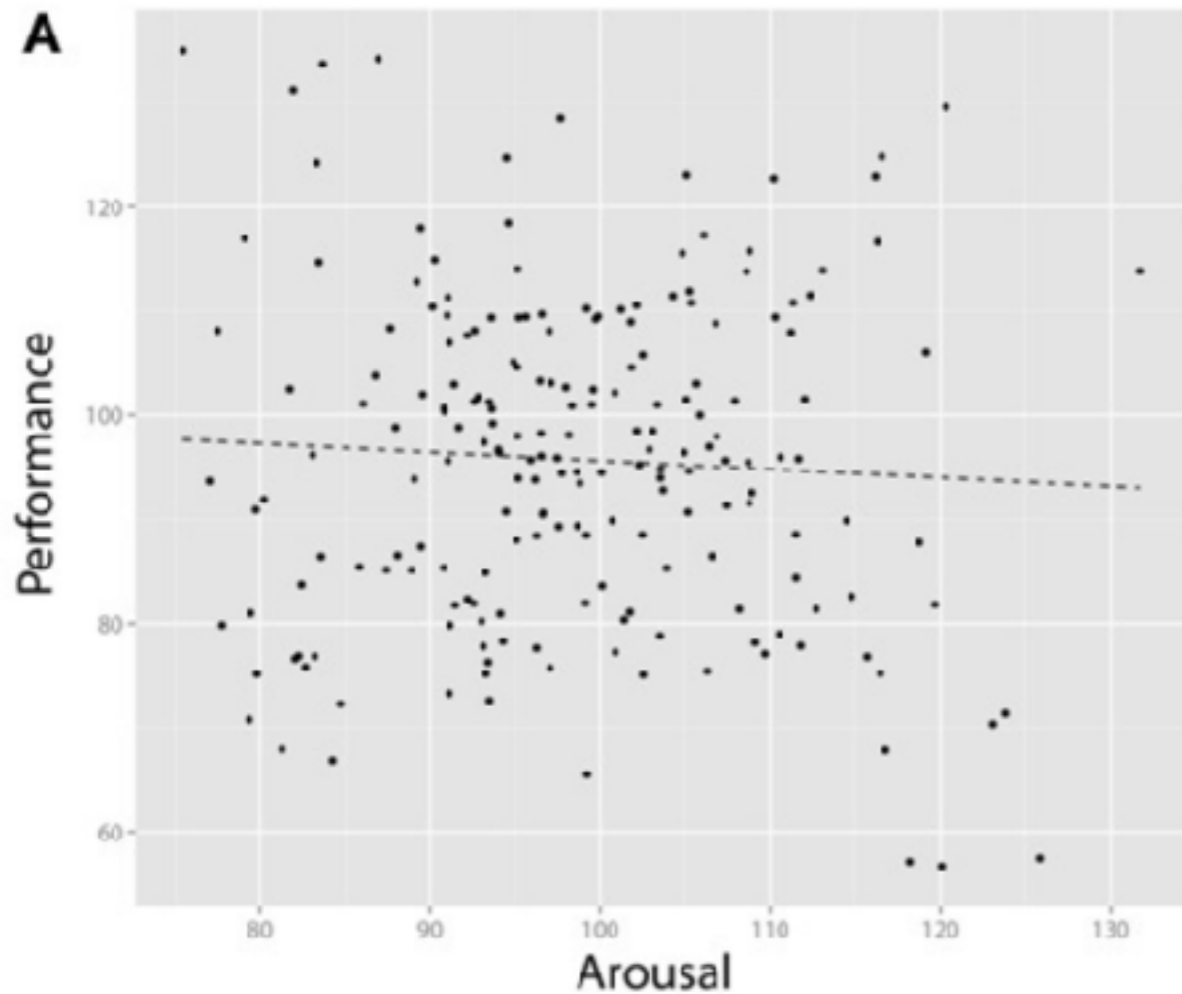

Looking for collaborators!

What is SP

Simpson's Paradox occurs when **subgroups** of a data set exhibit the ***opposite*** trend of the **whole** data set.
- Regression-based SP
- Rate-based SP

Regression-based SP

Kievit, Rogier A., et al. "Simpson's paradox in psychological science: a practical guide." *Frontiers in psychology* 4 (2013).

Rate-based SP

A study of gender bias among graduate school admissions to University of California, Berkeley, for the fall of 1973

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

https://en.wikipedia.org/wiki

Why Detect SP

Undetected SP can cause an unaware analyst to draw incorrect conclusions.

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Our Contribution

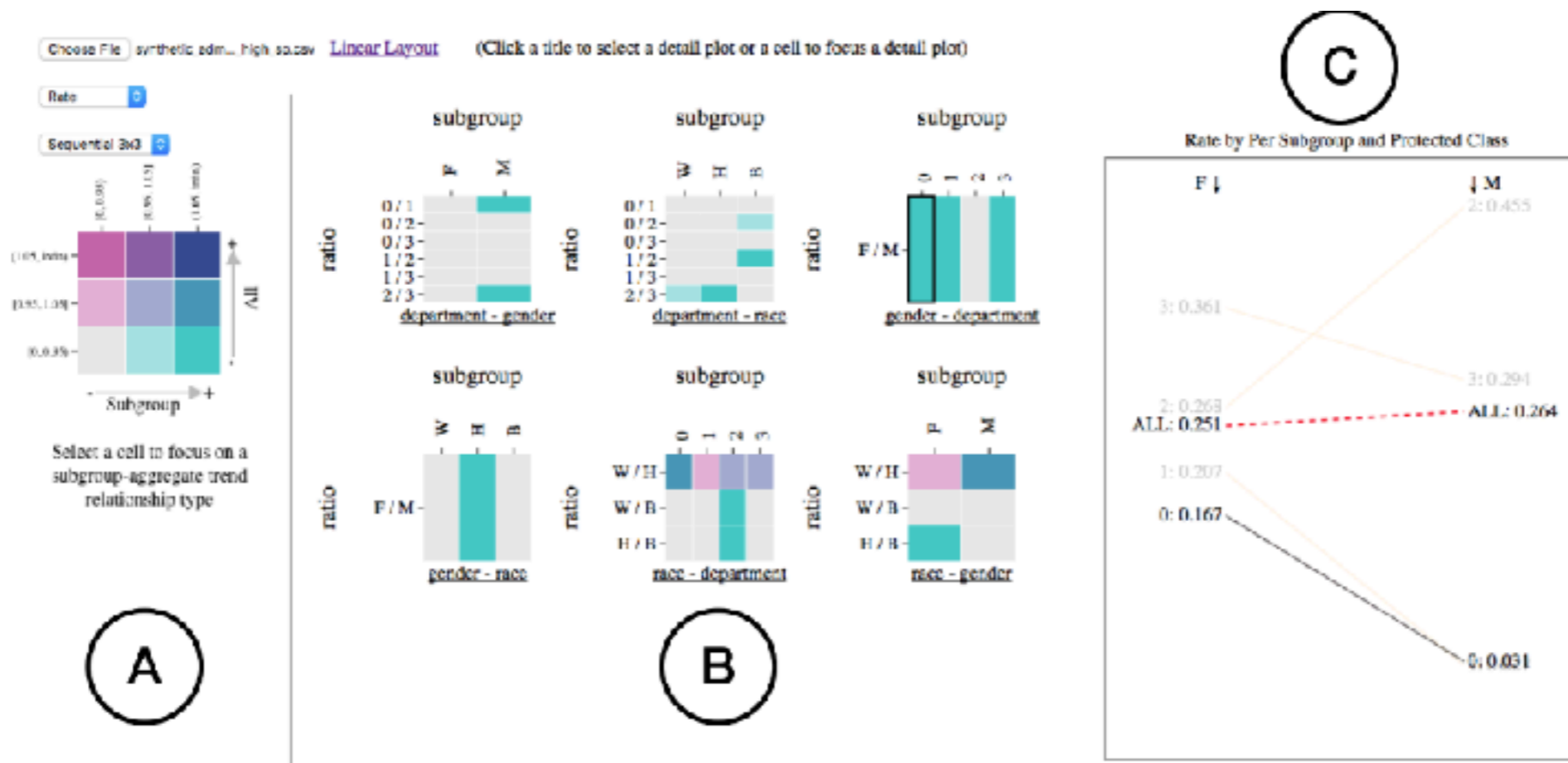Develop an interactive visual SP detecting website

Figure 2. An overview of interactive visual Analytics for rate based Simpson's paradox Detection. (A) User control panel (B) Bivariate scaled heatmaps for rate comparison matrices (C) A slope graph showing details.
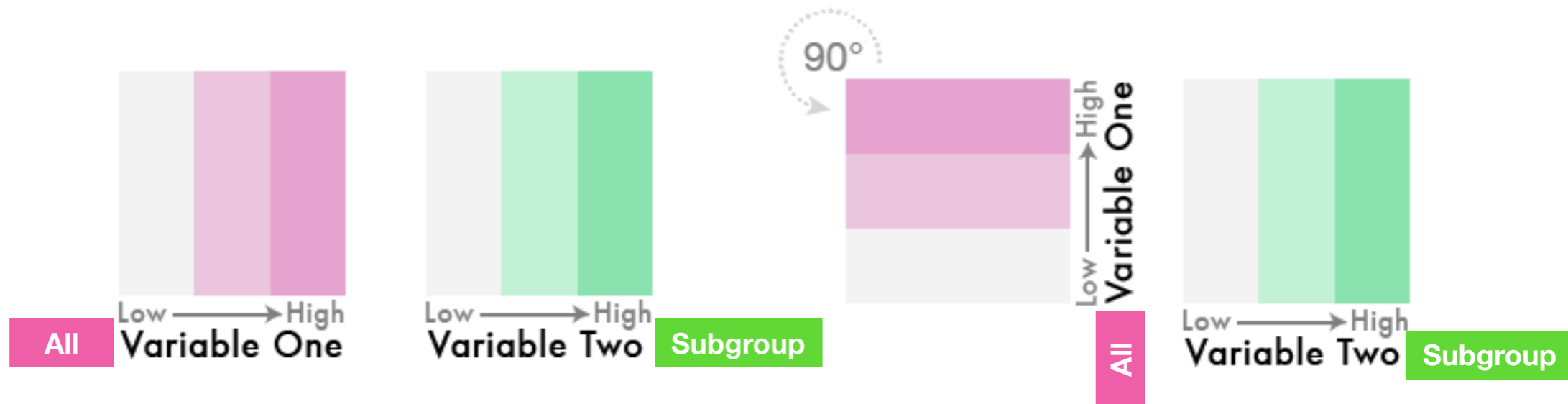
How to Detect SP

- Visual technique: Bivariate color scheme

- Interactive techniques:
  - Color Filtering
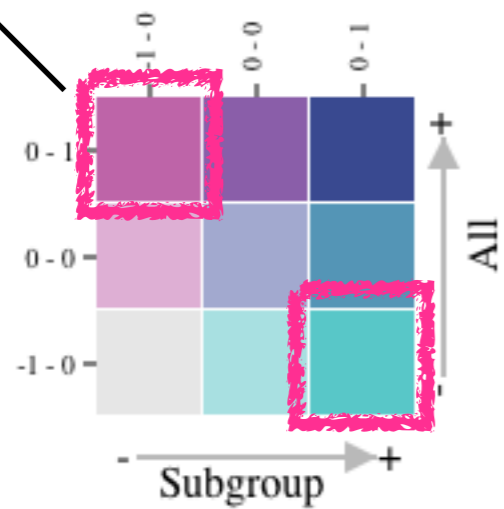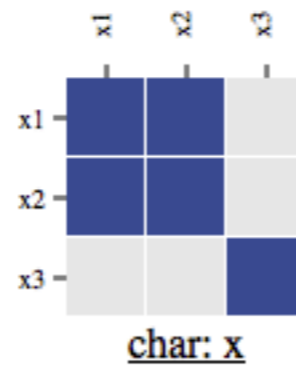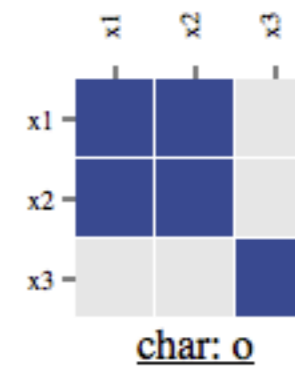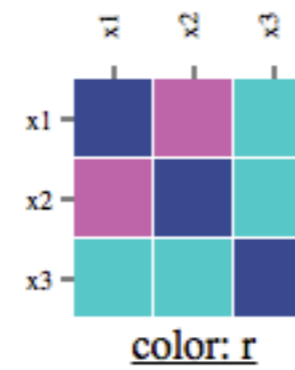 - Interact from overview to detail

Bivariate Color Scheme

Step 1

Step 2

90°

Step 3

Stevens, Joshua. Bivariate choropleth maps: A how-to guide. http://www.joshuastevens.net/cartography/make-a-bivariate-choropleth-map/, 2015

Bivariate Color for SP
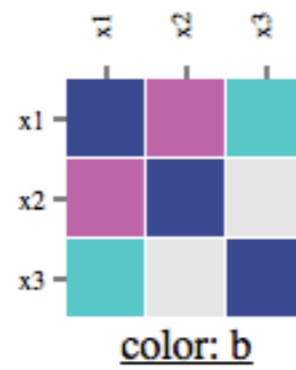
Bivariate color selector
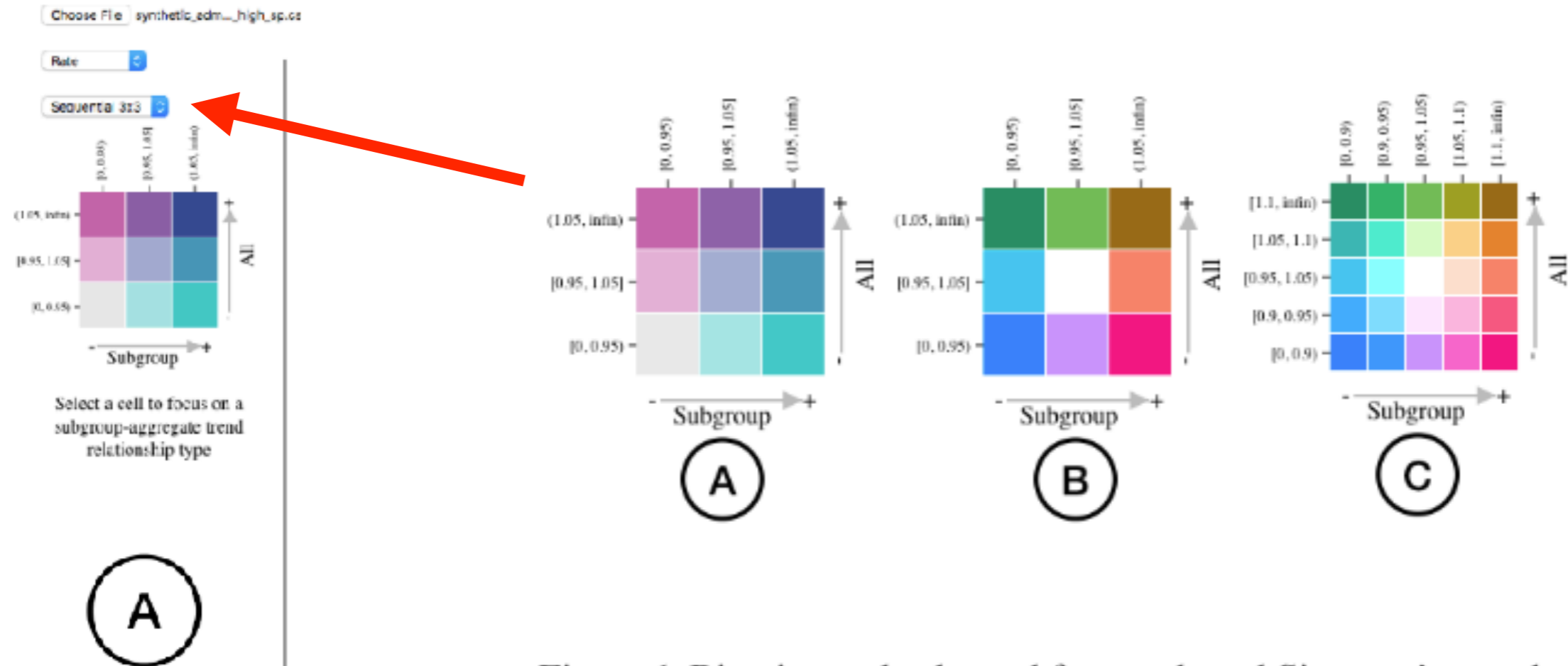
Figure 1. Bivariate color legend for rate based Simpson's paradox. (A) Sequential 3×3 (B) Diverging 3×3 (C) Diverging 5×5.
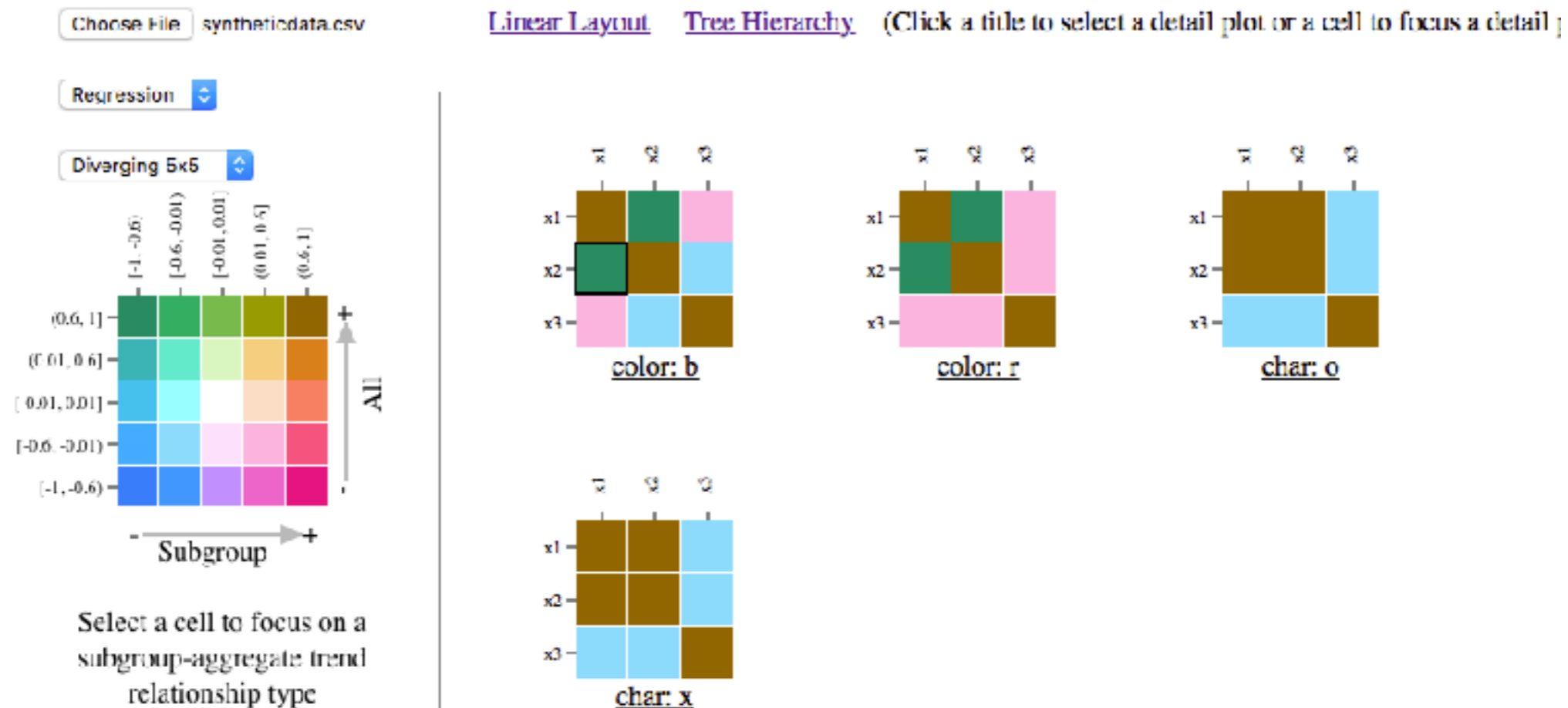
Bivariate Color for Matrices
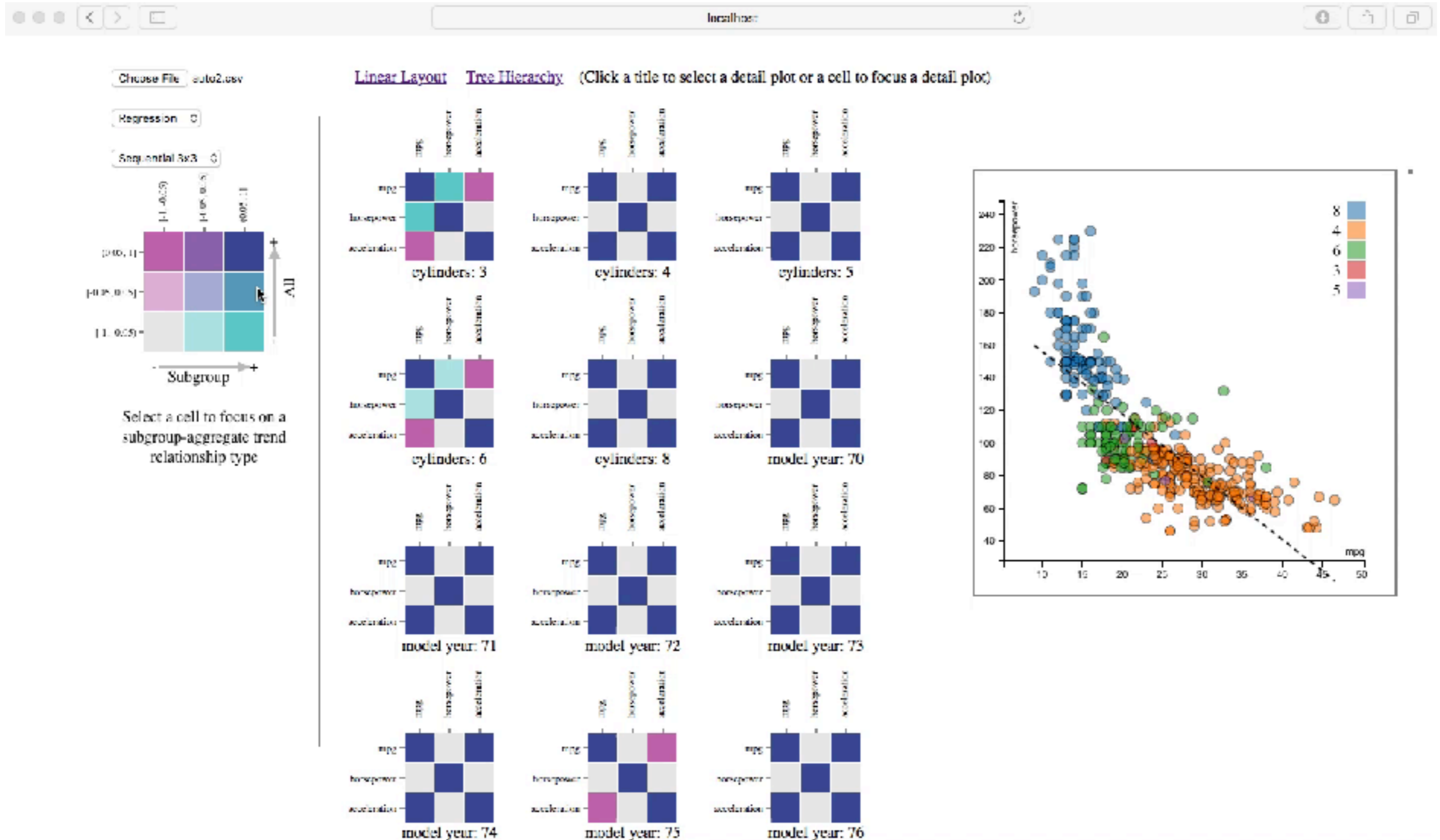
Bivariate color for rate comparison matrices
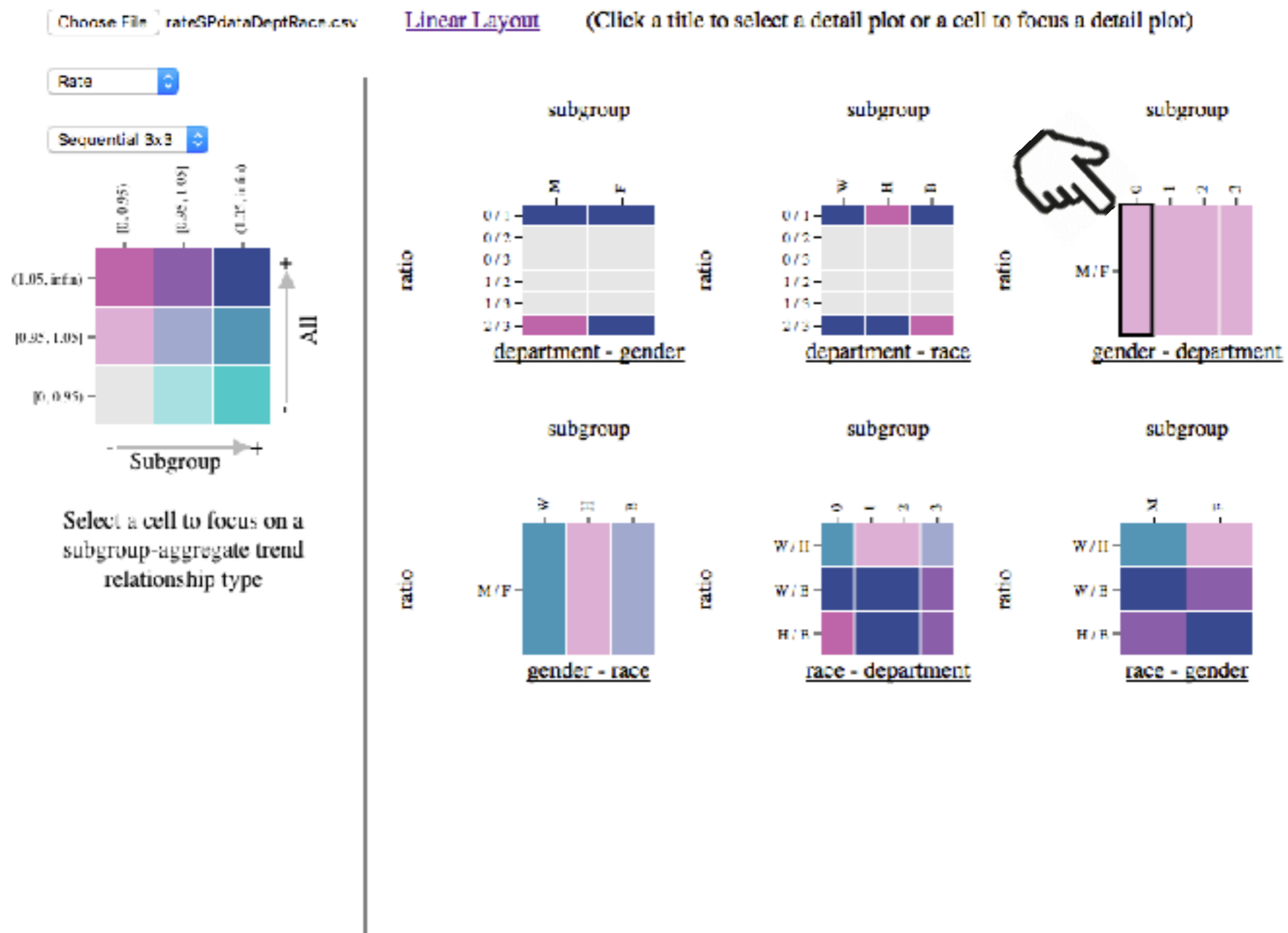
Bivariate Color for Matrices (cont.)

Bivariate color for correlation matrices
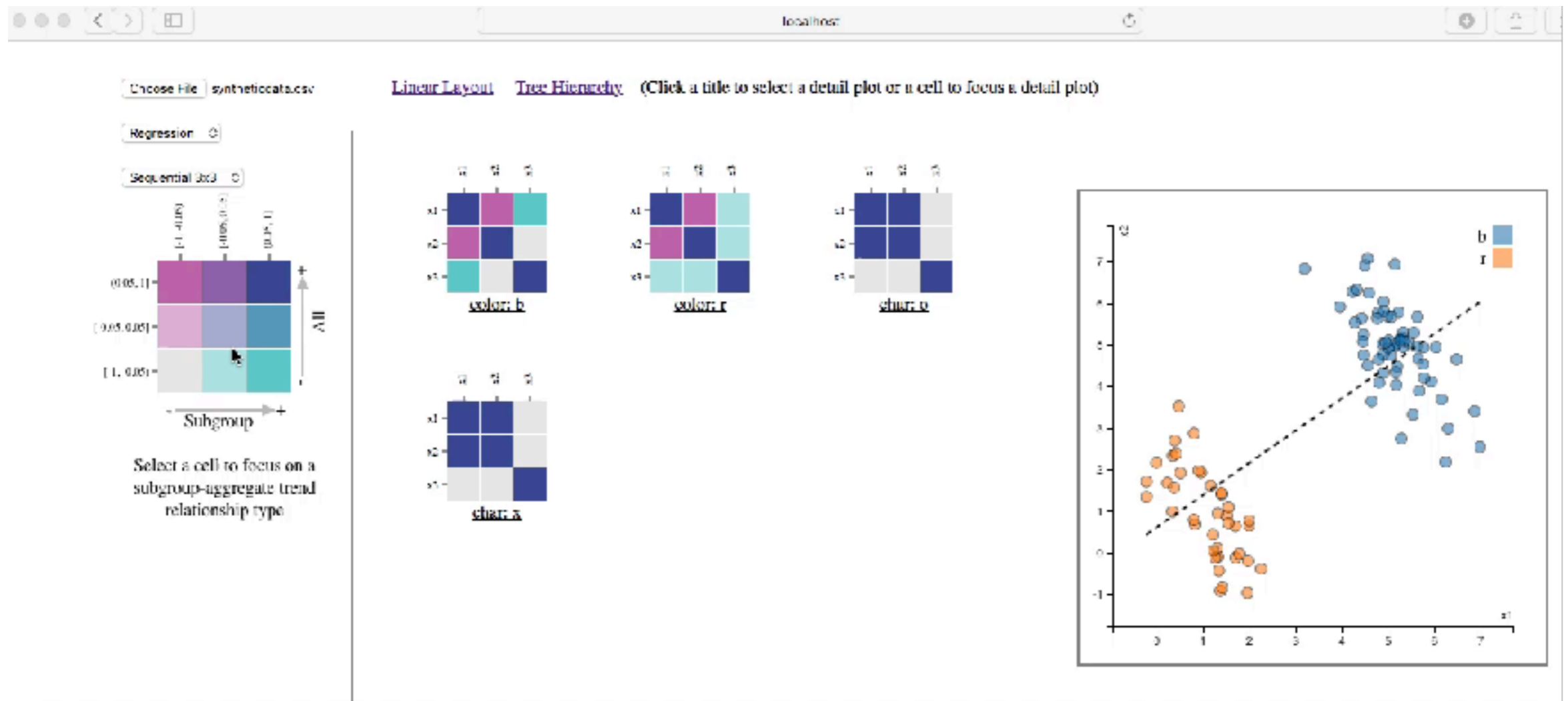
Color Filtering

Overview to Details

Interactive with slope graph for rate-based SP

Overview to Details (cont.)

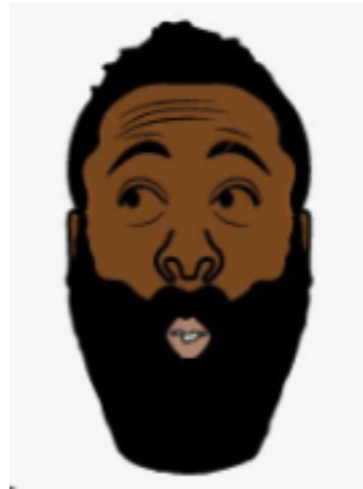Interactive with scatterplot for Regression SP

Summary

- Present an interactive interface that facilitates visual detection of SP

- Introduce bivariate-scale heat maps to indicate subgroup-aggregate trend relationship

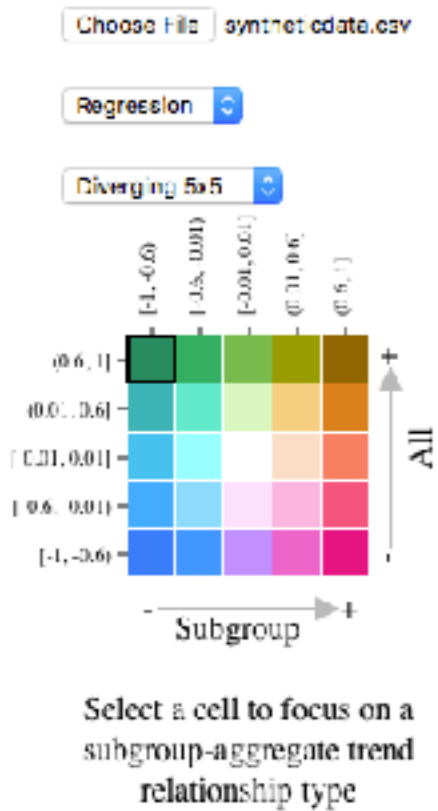- Explore SP from overview to details

References

[1] Armstrong, Zan and Wattenberg, Martin. Visualizing sta-tistical mix effects and simpson's paradox.*IEEE trans-actions on visualization and computer graphics*, 20(12):2132–2141, 2014

[2] Bickel, Peter J, Hammel, Eugene A, O'Connell, J William,et al. Sex bias in graduate admissions: Data from berkeley.*Science*, 187(4175):398–404, 1975.

[3] Stevens,Joshua.Bivariatechoroplethmaps:Ahow-toguide.http://www.joshuastevens.net/cartography/make-a-bivariate-choropleth-map/, 2015.

[4] Trumbo, Bruce E. A theory for coloring bivariate statisticalmaps.*The American Statistician*, 35(4):220–226, 1981.

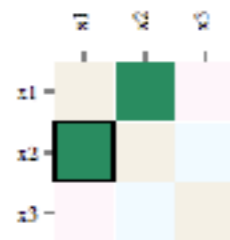[5] Xu, Chenguang, Brown, Sarah M, and Grant, Christan. De-tecting simpson's paradox.*AAAI, 2018.*
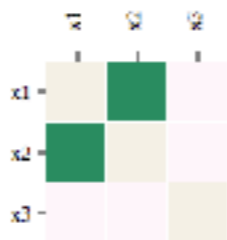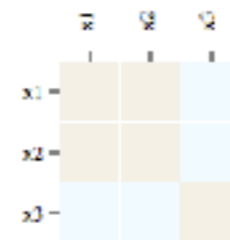
Question?

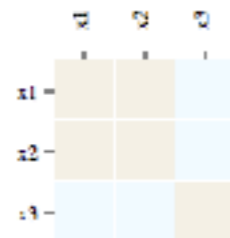# Color Filtering

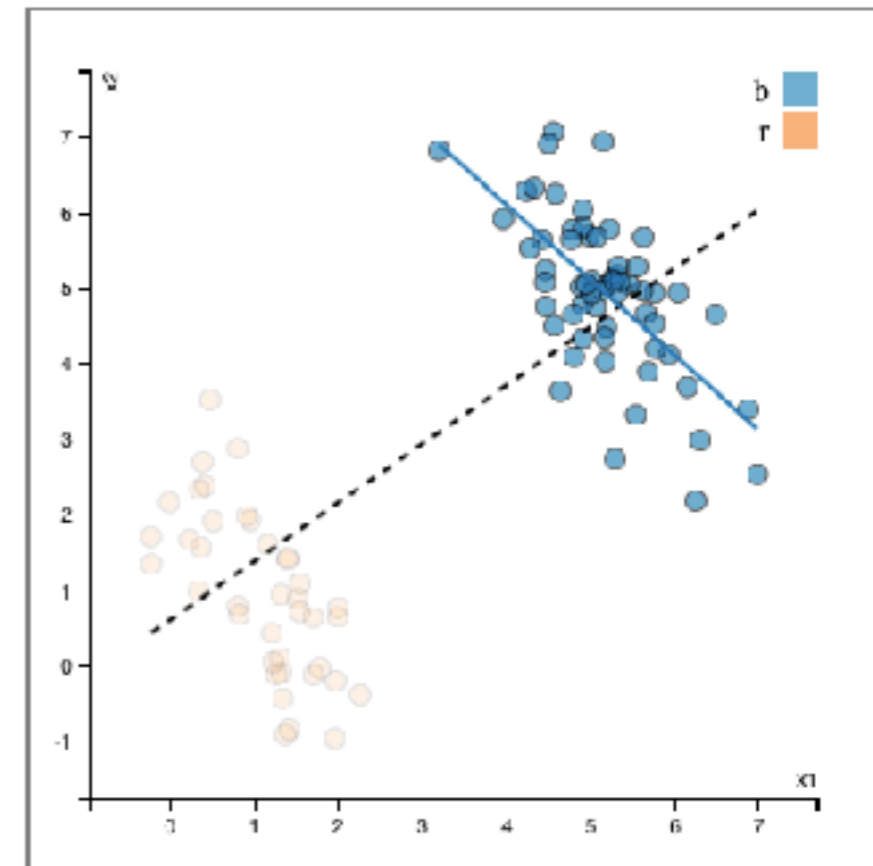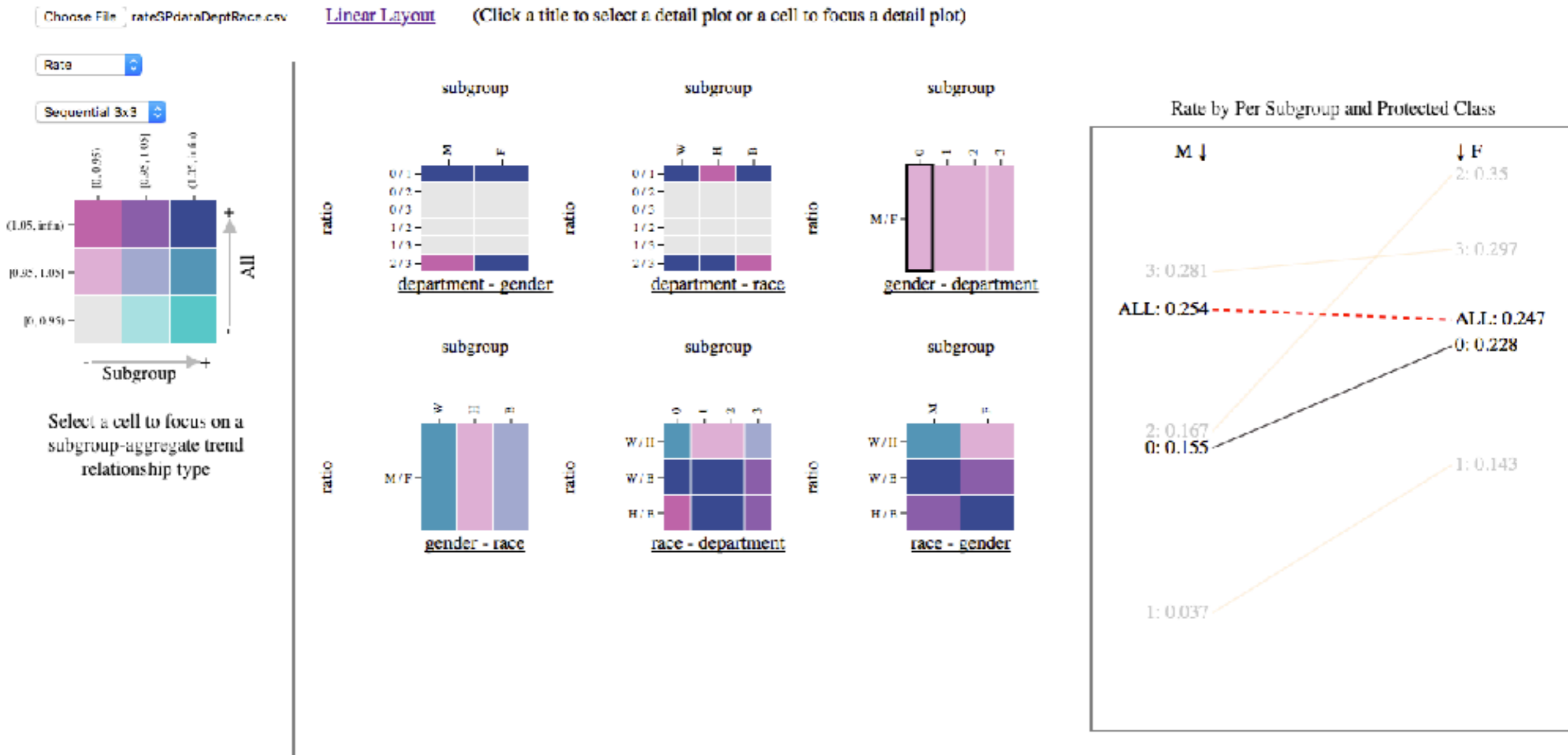# Overview to Details

- Interactive with slope graph for rate-based SP

Overview to Details (cont.)

- Interactive with scatterplot for Regression SP