Applications of Technology and Large Data in Statistics Education and Statistical Graphics

Karsten Maurer

Iowa State University

June 4, 2015

Presentation Outline

Chapter 4: Binning Strategies and Related Loss for Binned Scatterplots

- Scatterplot Adaptations
- Binning and Loss Functions
- Impact of Binning Specification on Loss
- Results and Recommendations

Chapter 3: A `shiny` New Opportunity for Interaction with Big Data in Undergraduate Education

- Development
- Applications
- Student User Survey
- Results

Binning Strategies and Related Loss for Binned Scatterplots
Chapter 4

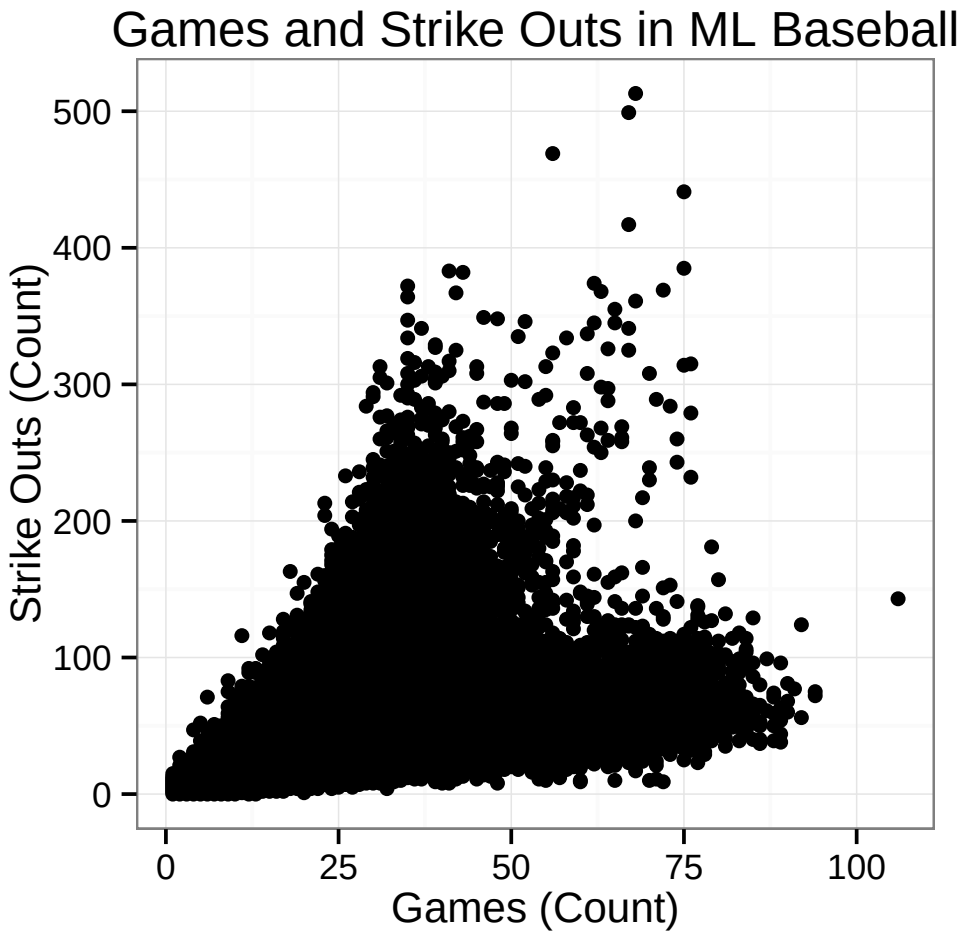
Bivariate Data Visualization

Example Data from Baseball

- Sean Lahman's Baseball database (`http://www.seanlahman.com/baseball-archive/`)
- Variables are Games Played and Strike Outs
- Observations: 42583 MLB pitcher seasons
- Data from 1871-2009

Goal: Visualize relationship between games played and strike outs

Traditional Scatterplot



Heavy over-plotting

Adaptations to Scatterplots

Lots of Recommendations for Dealing with Scatterplot Over-plotting

Change Point Rendering

- Open Points
- Alpha-Blending
- Generalized Scatterplots

Plotting Binned Aggregations

- Sunflower Plots
- Bubble Plots
- Binned Scatterplots

Adaptations to Scatterplots

Lots of Recommendations for Dealing with Scatterplot Over-plotting

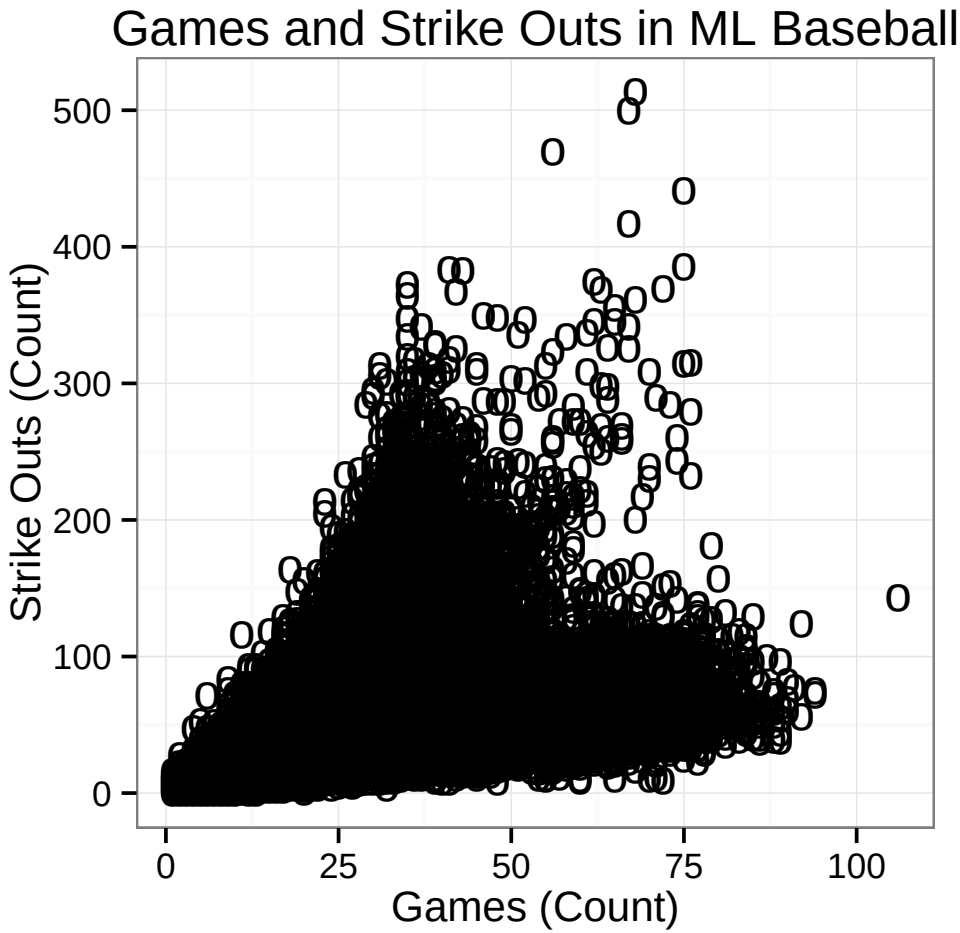
Change Point Rendering

- Open Points
- Alpha-Blending
- Generalized Scatterplots

Plotting Binned Aggregations

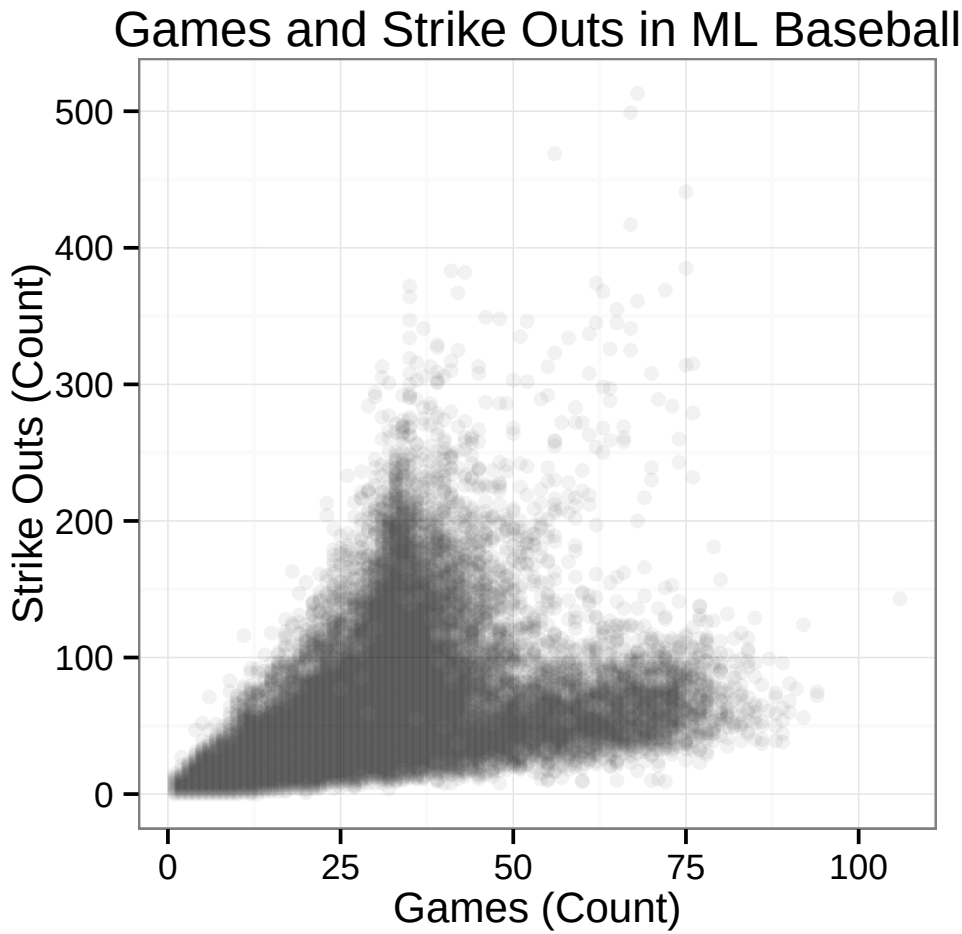
- Sunflower Plots
- Bubble Plots
- Binned Scatterplots

Open-Circle Scatterplot



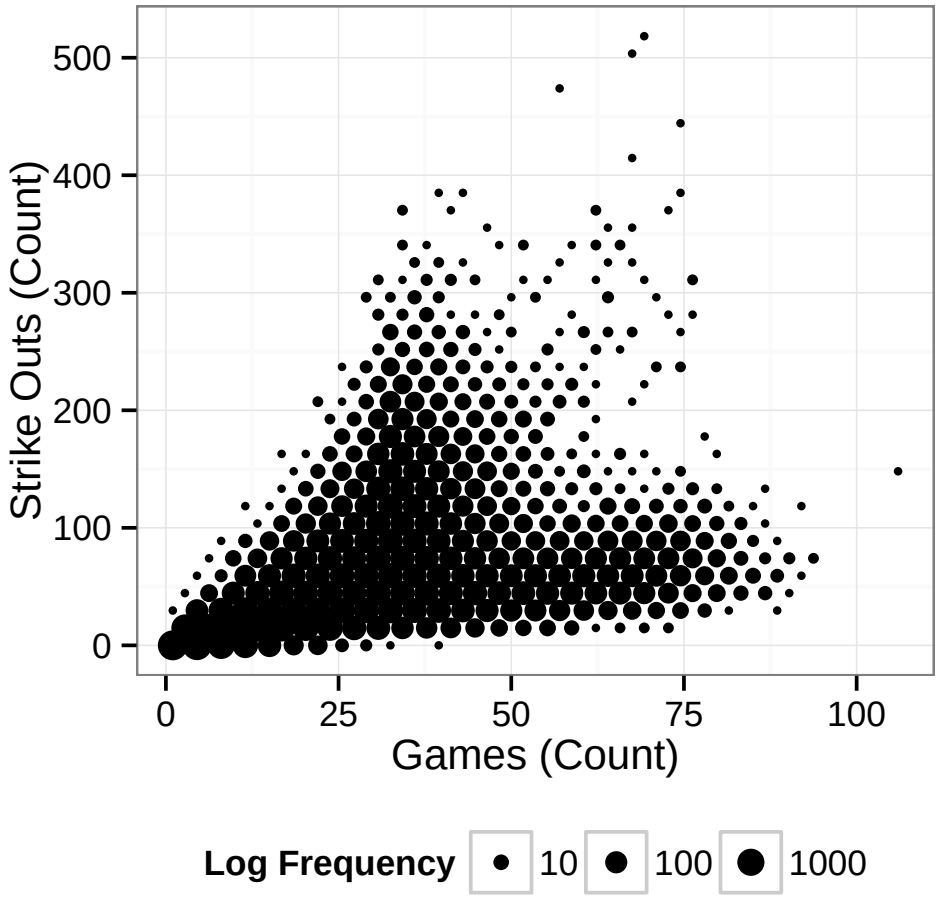
Helps slightly, but still allows for overplotting

Alpha-Blended Scatterplot



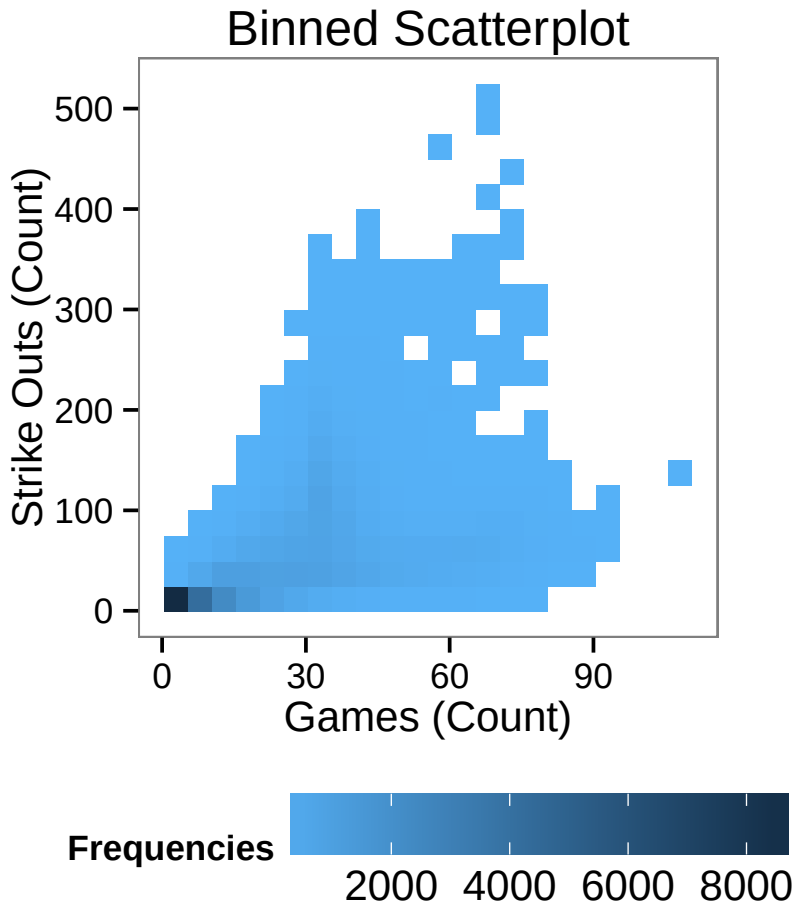
Fairly effective, but downplays outliers

Bubble Plot



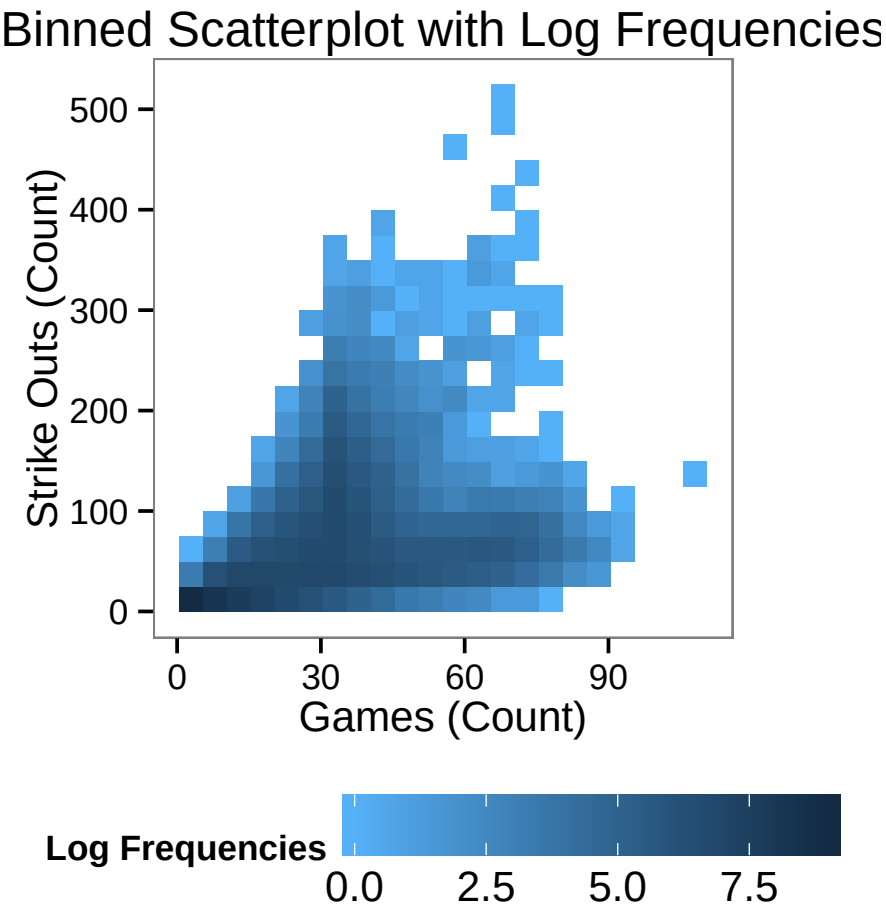
Mapping frequency to size of point at bin center

Binned Scatterplot



Mapping frequency to shade of tile

Binned Scatterplot



Mapping log(frequency) to shade of tile.

Binning Functions

Goal for Binned Scatterplot:

- Reduced Binned Data Triples (x^*, y^*, c)
- (x^*, y^*) = bin center
- c = bin frequency count

Need binning function to uniquely map points to bins

- $b(.) : (x, y) \rightarrow (x^*, y^*)$

Rectangular Binning Functions

Rectangular binning bins each dimension univariately

- $b(x_i, y_i) = (b_X(x_i), b_Y(y_i))$
- Forms rectangular binned grid

Generally $b_X(.) : x_i \rightarrow x_j^*$ is defined as

$$b_X(x_i) = \left\{ \begin{array}{ll} x_1^* & \text{for all } x_i = \beta_0 \\ x_j^* & \text{for all } x_i \in (\beta_{j-1}, \beta_j] \end{array} \right\} \quad (1)$$

where $(\beta_{j-1}, \beta_j]$ for $j \in \{1, \ldots, J\}$ are sequence of J adjacent intervals spanning range of X data.

Rectangular Binning Functions

Rectangular binning bins each dimension univariately

- $b(x_i, y_i) = (b_X(x_i), b_Y(y_i))$
- Forms rectangular binned grid

Generally $b_X(.) : x_i \rightarrow x_j^*$ is defined as

$$b_X(x_i) = \left\{ \begin{array}{ll} x_1^* & \text{for all } x_i = \beta_0 \\ x_j^* & \text{for all } x_i \in (\beta_{j-1}, \beta_j] \end{array} \right\} \tag{1}$$

where $(\beta_{j-1}, \beta_j]$ for $j \in \{1, \ldots, J\}$ are sequence of J adjacent intervals spanning range of X data.

Rectangular Binning Functions

Standard Rectangular Binning

- Use equally spaced intervals
- Bin width $= \omega_X$
- $\{\beta_j \mid \beta_j = \beta_{j-1} + \omega_X\}$
- $x_j^* = (\beta_{j-1} + \beta_j)/2$

Reduced Binned Data

x	y
-7.7325	-9.6340
-8.1176	-1.4529
-5.8996	-3.2033
-7.0375	-5.5563
-3.6354	-3.9315
-8.7639	0.9874
-2.9781	8.6802
0.8210	-8.6118
5.4477	-8.4555
4.6849	-5.6620
9.4785	1.1133
1.7579	5.3759

(a) Original Data, 12 rows

$b_X(x)$	$b_Y(y)$
-5	-5
-5	-5
-5	-5
-5	-5
-5	-5
-5	5
-5	5
5	-5
5	-5
5	-5
5	5
5	5

(b) Binned Data Centers, 12 rows

x^*	y^*	c
-5	-5	5
-5	5	2
5	-5	3
5	5	2

(c) Reduced Binned Data, 4 rows

Table: Original, Binned and Reduced Binned Data Tables, with data storage sizes. Binned using standard rectangular approach with origin $(\beta_{0,x}, \beta_{0,y})$ = (-10,-10) and bin widths $\omega_x = \omega_y = 10$.

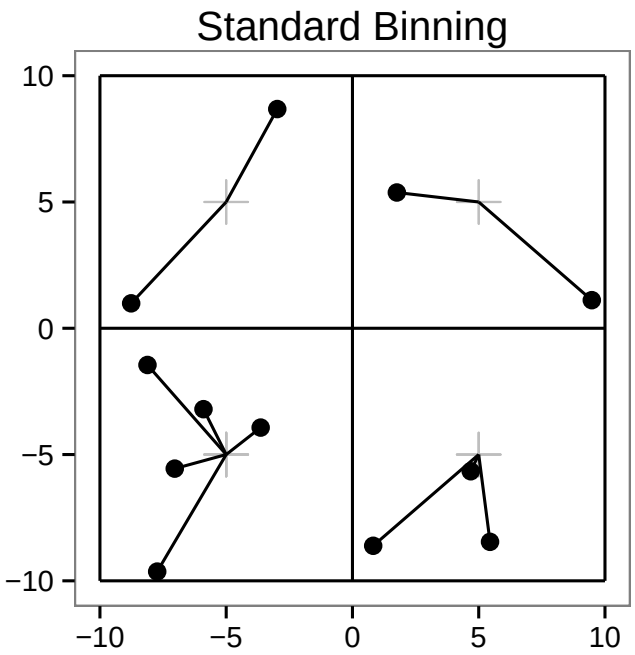
Rectangular Binning Functions

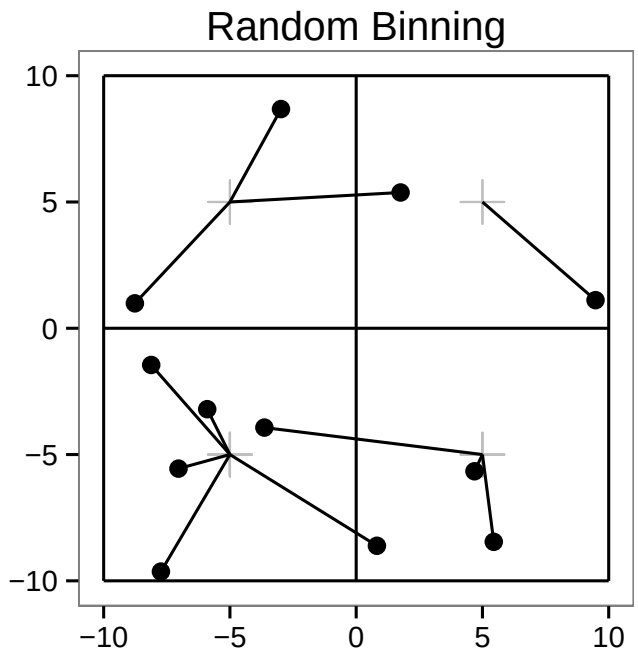
Random Rectangular Binning

- Define set of bin centers: $\{x_j^* \mid x_j^* > x_{j-1}^*\}$
- Assign points at random to surrounding bin centers
- Probabilities inversely proportional do distance from bin centers

$$b_X^r(x_i) = \left\{ \begin{array}{ll} x_j^* & \text{with probability } (x_{j+1}^* - x_i)/(x_{j+1}^* - x_j^*) \\ x_{j+1}^* & \text{with probability } (x_i - x_j^*)/(x_{j+1}^* - x_j^*) \end{array} \right\} \quad (2)$$

Standard vs. Random Rectangular Binning





Spatial Loss

Create loss function for spatial information

- Use Euclidean distance between points and bin centers
- Spatial loss for i^{th} observation:
  $L_i^S = \sqrt{(x_i - b_X(x_i))^2 + (y_i - b_Y(y_i))^2}$
- *Total* spatial loss: $L^S = \sum_{i=1}^{n} L_i^S$

Slight issue with spatial loss for random binning

- Many possible random assignments result in same reduced binned data
- Find assignment with minimum total spatial loss
- Call this *net* spatial loss

Spatial Loss

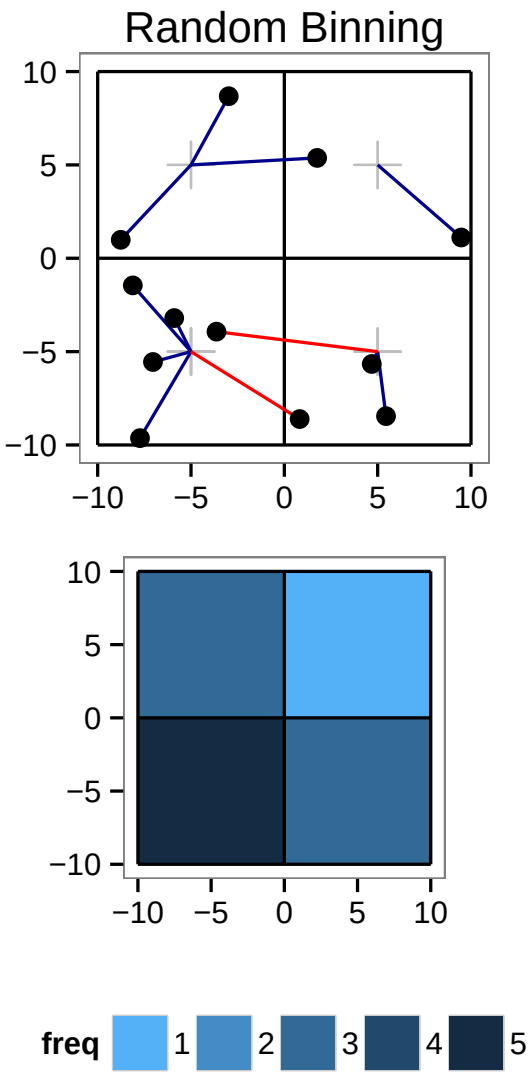
Create loss function for spatial information

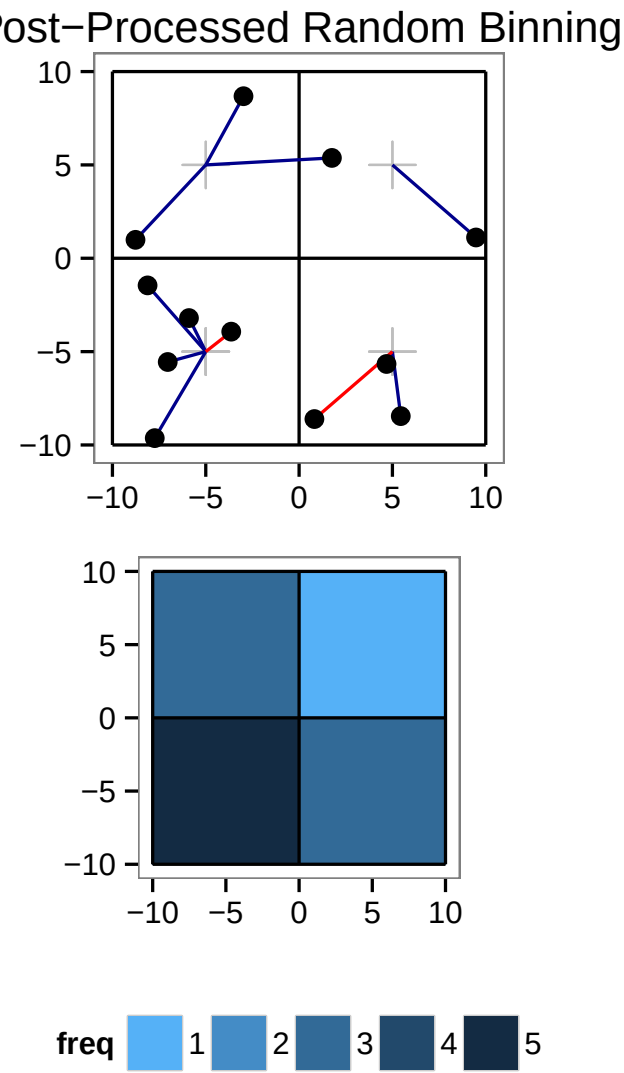
- Use Euclidean distance between points and bin centers
- Spatial loss for i^{th} observation:
  $L_i^S = \sqrt{(x_i - b_X(x_i))^2 + (y_i - b_Y(y_i))^2}$
- *Total* spatial loss: $L^S = \sum_{i=1}^{n} L_i^S$

Slight issue with spatial loss for random binning

- Many possible random assignments result in same reduced binned data
- Find assignment with minimum total spatial loss
- Call this *net* spatial loss

Post Processing Random Rectangular Binning





Rectangular Binning Specifications

Rectangular Binning Specification

- Binning function
- Bin dimensions
- Binning origin

Properties of Interest

- Net spatial loss
- Computation time
- Visually misleading features

Rectangular Binning Specifications

Rectangular Binning Specification

- Binning function
- Bin dimensions
- Binning origin

Properties of Interest

- Net spatial loss
- Computation time
- Visually misleading features

Rectangular Binning Specifications

Standard versus Random Rectangular Binning

- Standard binning computationally faster
- Standard binning is superior for spatial loss
- Random binning locally smooths bin frequencies

Small versus Large Bins

- Net spatial loss decreases as bin dimensions decrease
- Computation time increases as bin dimensions decrease
- Small bins allow fine bivariate structure to show through
- Large bins emphasize large scale bivariate structure

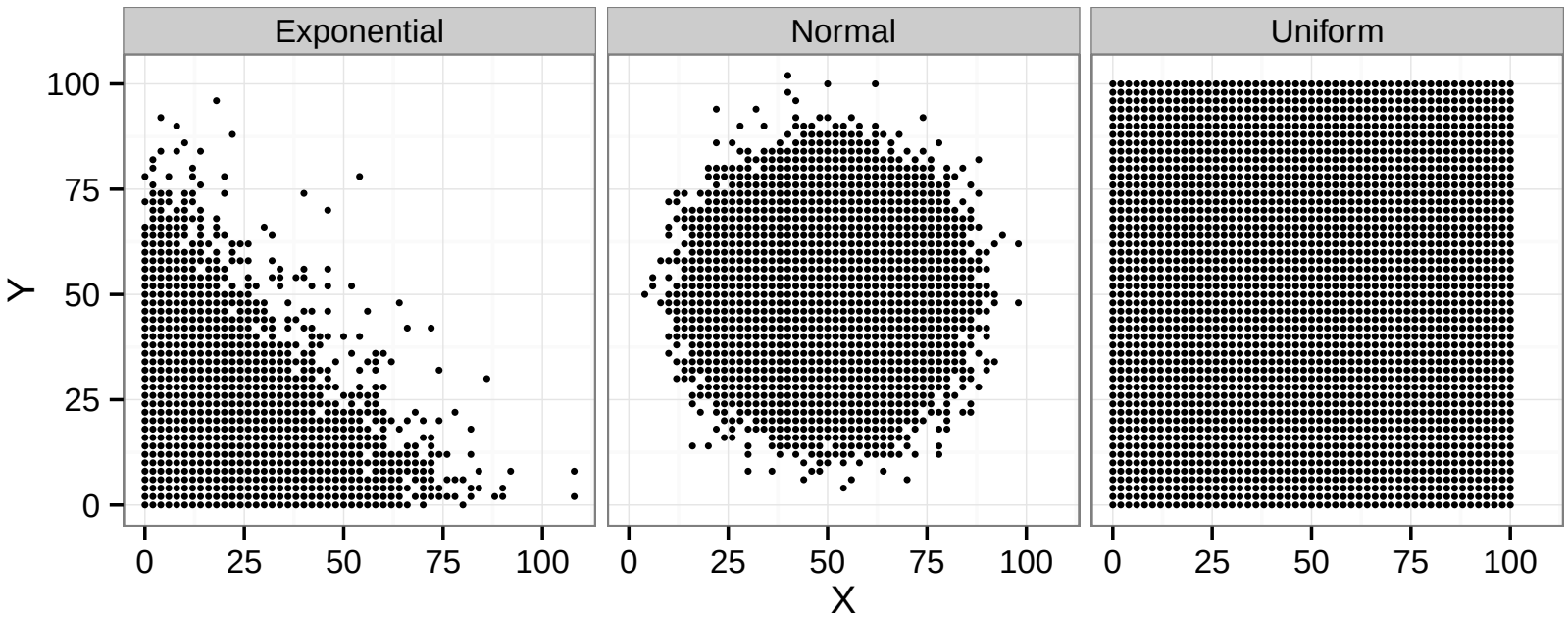
Simulation Study

Explore Properties with Three Simulated Data Sets
100000 observations from:

- Set I: x_i and $y_i \sim$ iid Exp($\lambda = 11$)
- Set II: x_i and $y_i \sim$ iid Normal($\mu = 50, \sigma^2 = 11^2$)
- Set III: x_i and $y_i \sim$ iid Uniform($a = 0, b = 100$)

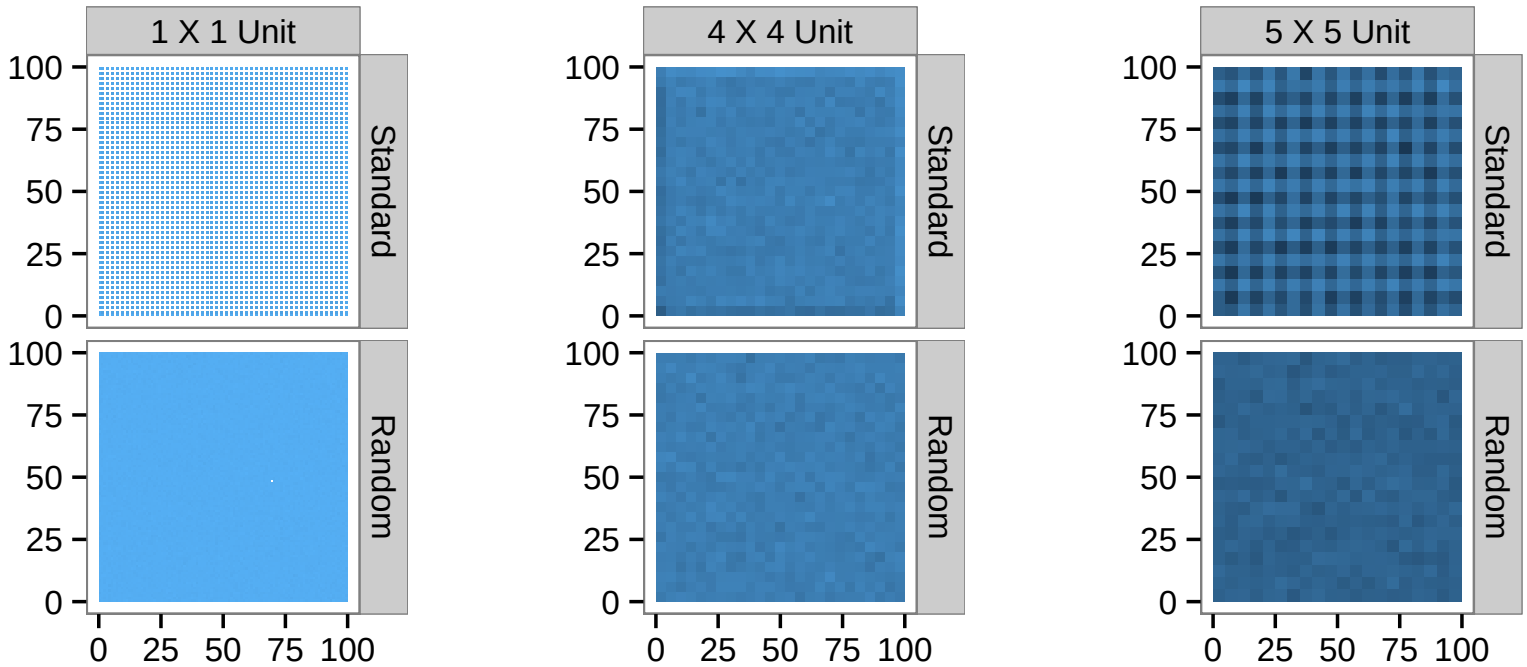
Data then rounded to nearest even integer to induce a *data resolution* of 2 units

Simulated Data



Scatterplots of simulated data recorded to nearest even integer

Selecting Bin Width



Binned Scatterplots of Uniform Data

Bin dimensions for standard binning should be integer multiples of the resolution of the data to avoid *artificial* stripes.

Selecting a Binning Origin

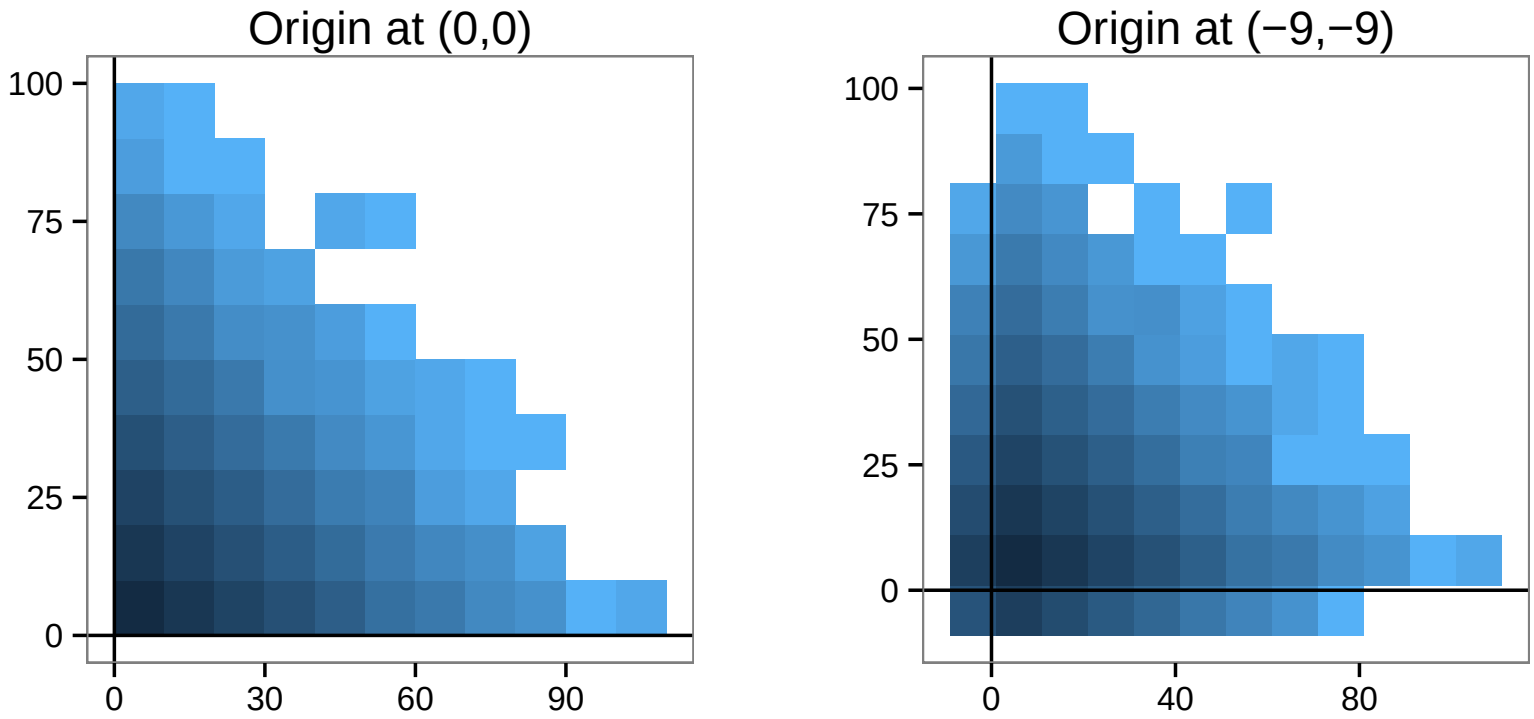
Specifying a binning origin to lower spatial loss

- Avoid heavy overhangs outside data boundaries
- Try to align bin centers with possible data values

A reliable default based on the data resolution

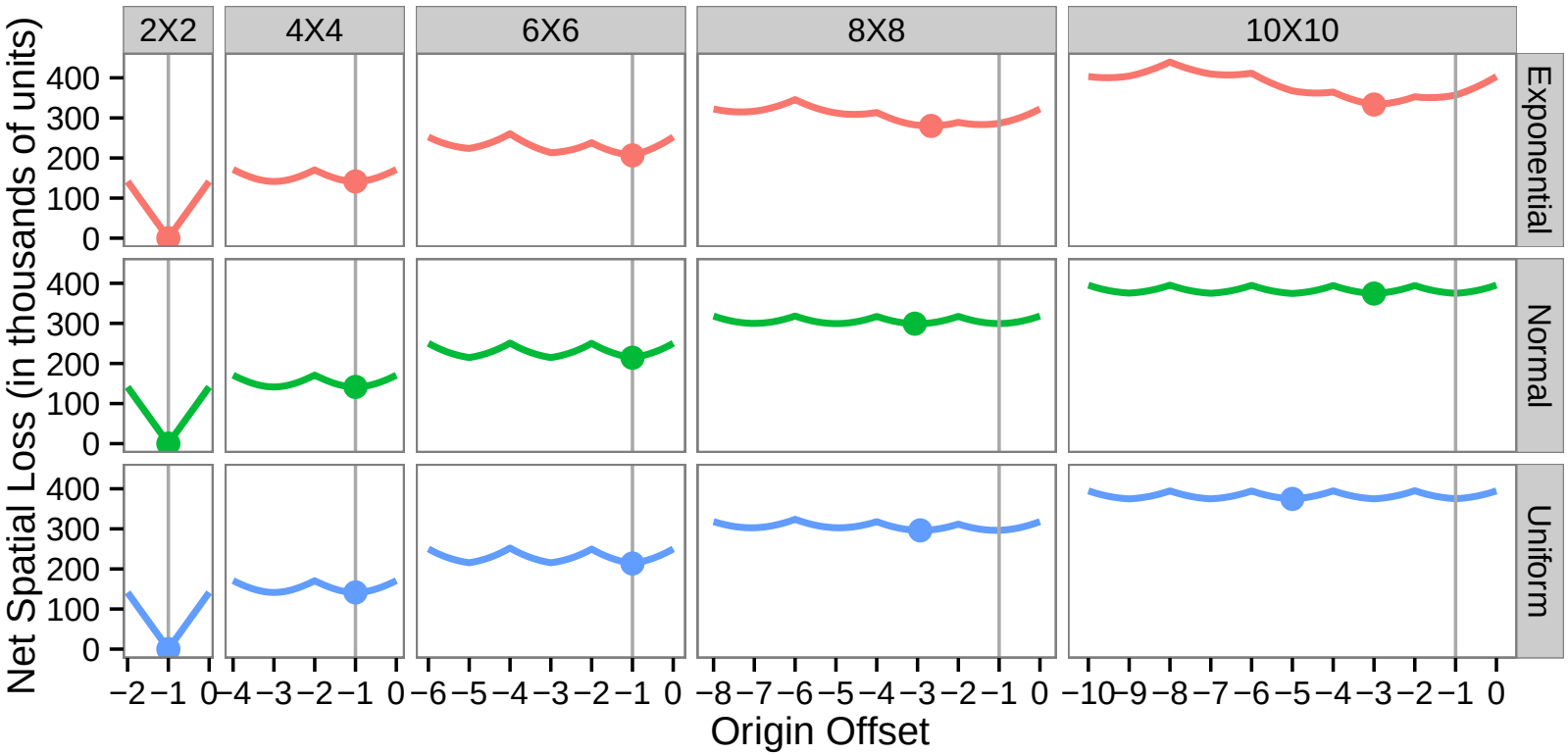
- Recorded to nearest α_x units in the X dimension
- Recorded to nearest α_y units in the Y dimension
- $x_{(1)}$ and $y_{(1)}$ are the minimum data values in each dimension
- Set binning origin at $(x_{(1)}, y_{(1)}) - (\alpha_x/2, \alpha_y/2)$

Selecting a Binning Origin



Exponential data binned with 10X10 square bins.
Net spatial loss for binning origin at (0,0) is 7% lower than (-9,-9).

Selecting a Binning Origin



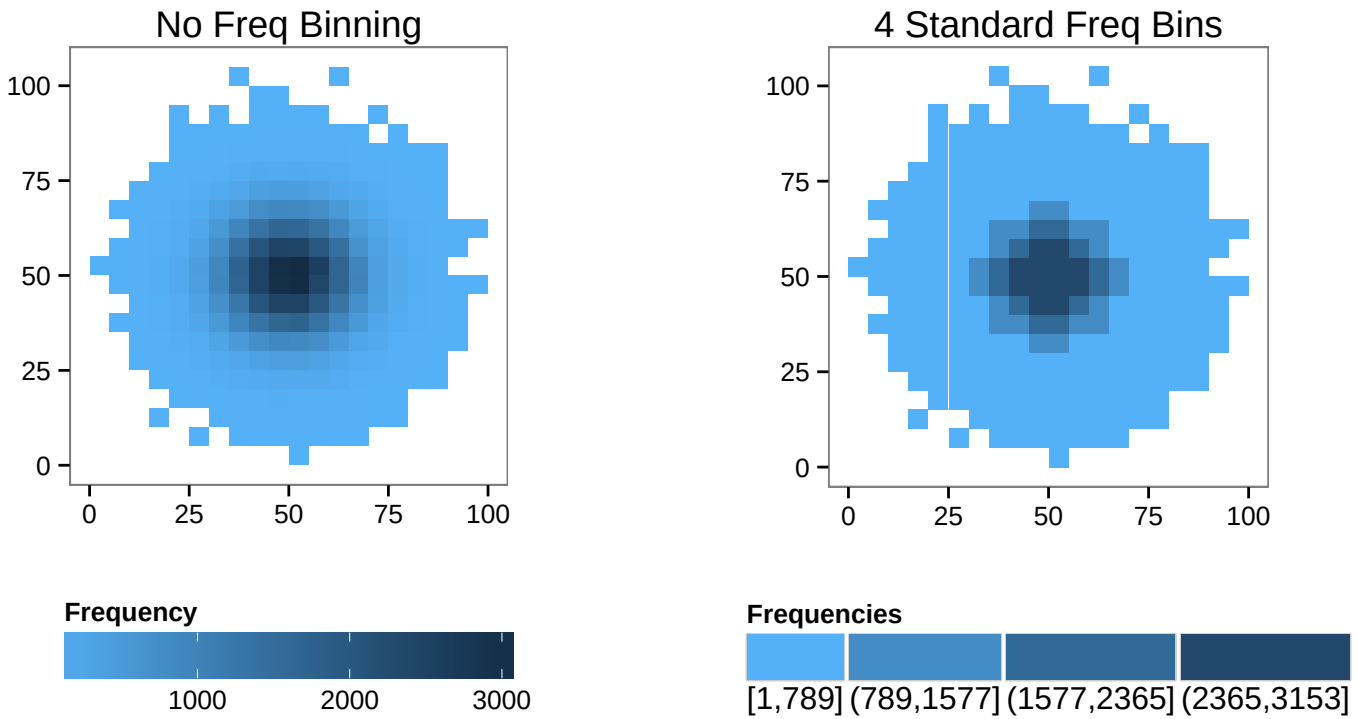
Net spatial loss for simulated data (with 2X2 resolution)
Binning origin at (-1,-1) is good default option for all bin sizes

Frequency Binning

Bin count mapped to shade can be rendered continuously in HCL color space

Why would we discretize the frequency shading?

- Make inability to accurately visually map bin shades to plot keys explicit



Frequency Binning

How do we discretize the frequency shading?

- Add second stage of binning
- Use one dimensional binning algorithms on bin counts, c_ℓ
- Shade ordinally based on binned frequency groups
- May transform counts before frequency binning (i.e. log counts)

Frequency Binning

Standard Frequency Binning

- Create k frequency groups using equal bin width

Quantile Frequency Binning

- Create k frequency groups based on empirical quantiles of raw bin counts
- Roughly equal number of tiles in each shade

Shade interpretability analogous to: "Histogram vs Boxplot"

Frequency Binning

Standard Frequency Binning

- Create k frequency groups using equal bin width

Quantile Frequency Binning

- Create k frequency groups based on empirical quantiles of raw bin counts
- Roughly equal number of tiles in each shade

Shade interpretability analogous to: "Histogram vs Boxplot"

Frequency Binning

Standard Frequency Binning

- Create k frequency groups using equal bin width

Quantile Frequency Binning

- Create k frequency groups based on empirical quantiles of raw bin counts
- Roughly equal number of tiles in each shade

Shade interpretability analogous to: "Histogram vs Boxplot"

Frequency Loss

Frequency Loss

- bin frequencies, c_ℓ, for spatial bins $\ell \in \{1, \ldots, \mathscr{L}\}$
- binned frequencies, $b_C(c_\ell)$

$$L^F = \sum_{\ell=1}^{\mathscr{L}} (c_\ell - b_c(c_\ell))^2 \tag{3}$$

Log Frequency Loss

- If bin frequencies, c_ℓ, are log transformed
- binned log frequencies, $b_C(\log(c_\ell))$

$$L^{\log F} = \sum_{\ell \in \mathscr{L}^*} (\log(c_\ell) - b_c(\log(c_\ell)))^2 \tag{4}$$

Frequency Loss

Frequency Loss

- bin frequencies, c_ℓ, for spatial bins $\ell \in \{1, \ldots, \mathscr{L}\}$
- binned frequencies, $b_C(c_\ell)$

$$L^F = \sum_{\ell=1}^{\mathscr{L}} (c_\ell - b_c(c_\ell))^2 \tag{3}$$

Log Frequency Loss

- If bin frequencies, c_ℓ, are log transformed
- binned log frequencies, $b_C(\log(c_\ell))$

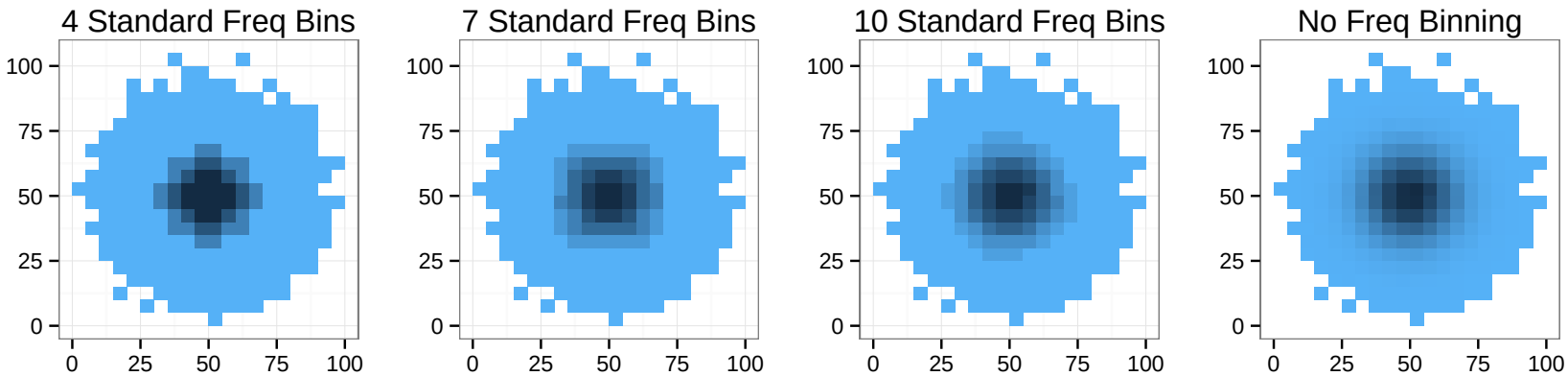
$$L^{\log F} = \sum_{\ell \in \mathscr{L}^*} (\log(c_\ell) - b_c(\log(c_\ell)))^2 \tag{4}$$

Frequency Binning Specification

3 Main Choices

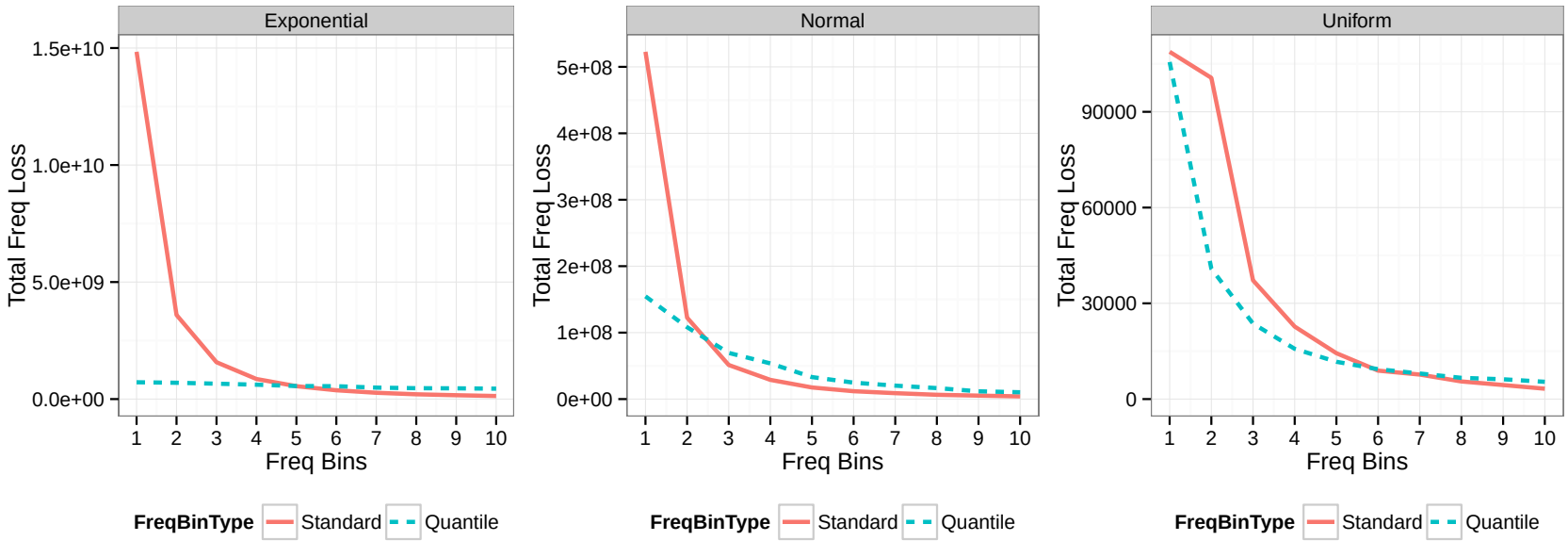
- Number of frequency bins (k)
- Transform counts (yes/no)
- Binning algorithm (standard/quantile)

Number of Frequency Bins



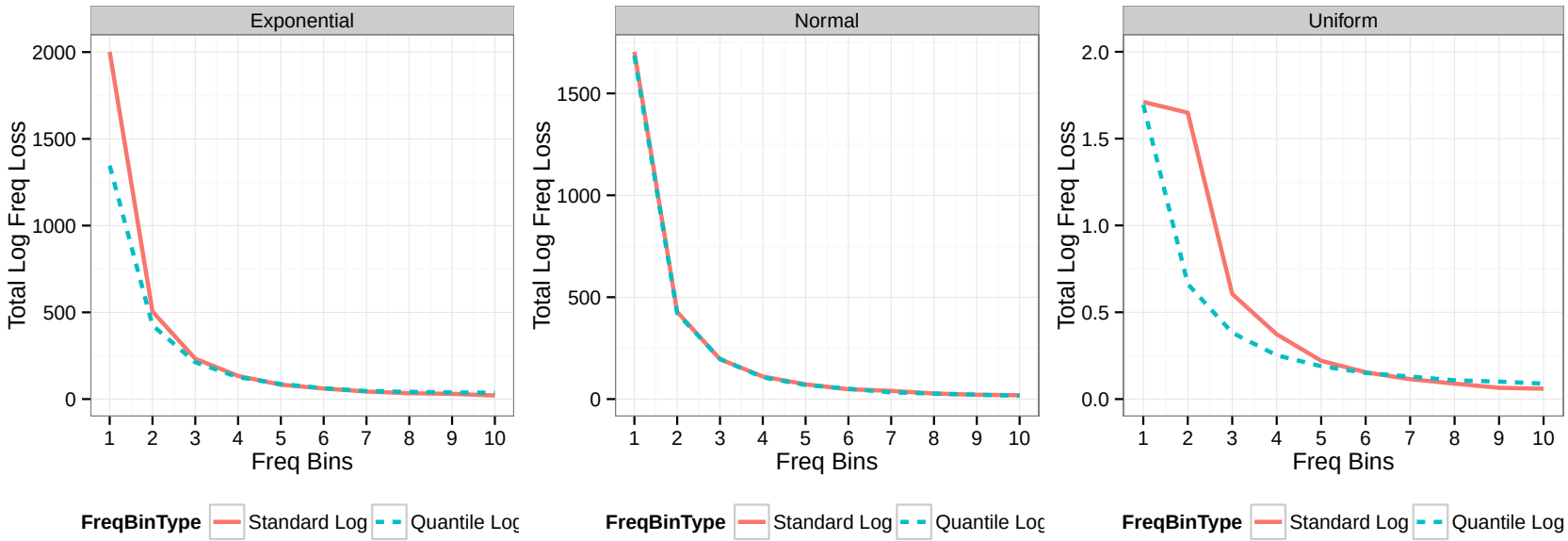
How many different shades can we distinguish simultaneously?
Research suggests 7 is max for the typical person $\Rightarrow$ use k $\leq$ 7

Frequency Loss



Number of Frequency Bins vs. Frequency Loss

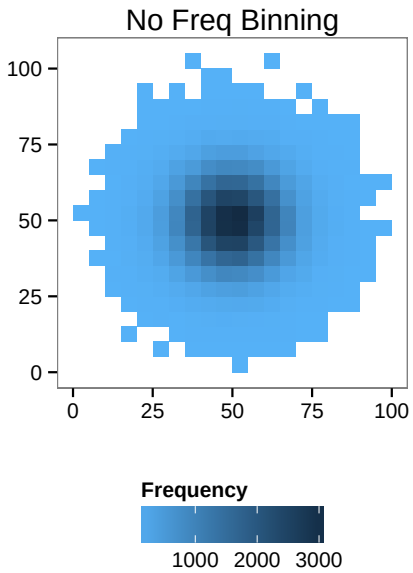
Log Frequency Loss

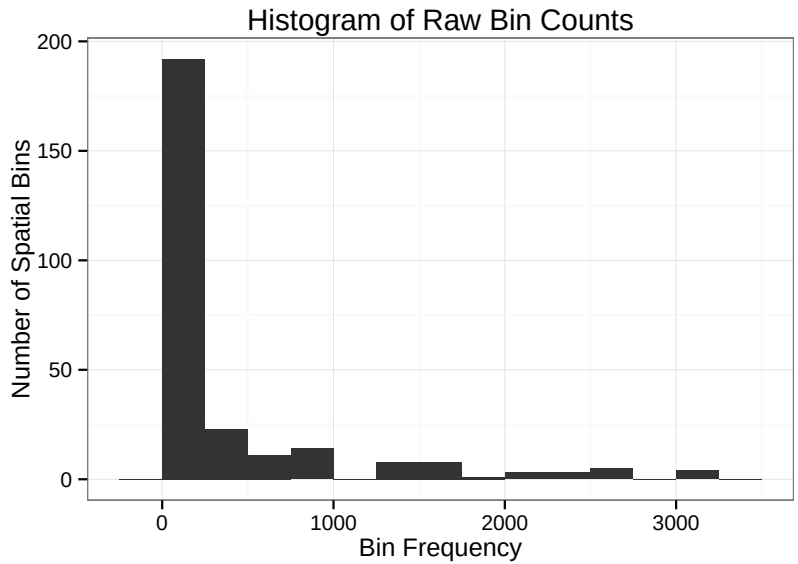


Number of Log Frequency Bins vs. Log Frequency Loss

Dealing With Skewed Bin Counts

Distribution of bin counts are commonly skewed

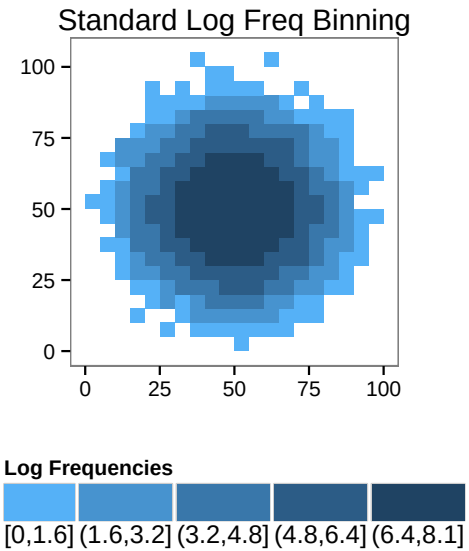


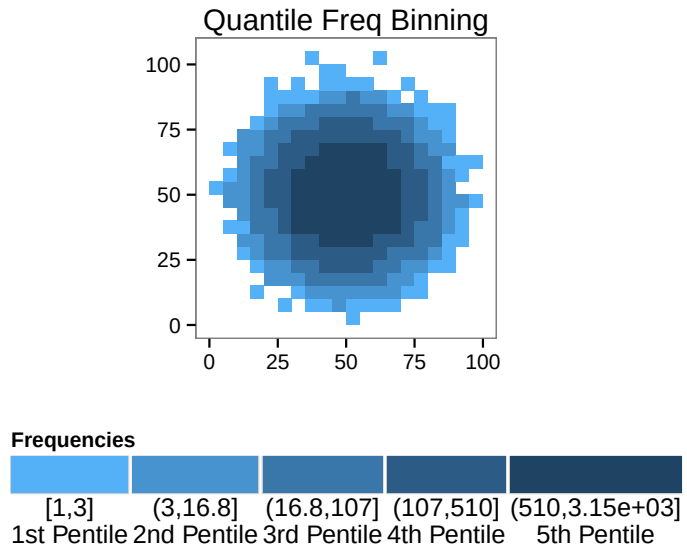


Unable to see differences in low frequency bins

Dealing With Skewed Bin Counts

Transforming bin counts can reduce skew in bin densities





Quantile frequency binning adds a second option for overcoming skew in visualization

Practical Takeaways for Binned Scatterplots

Spatial Binning

- Standard binning algorithm superior to random binning in spatial loss and computation speed
- Smaller bins superior for precisely displaying spatial information
- Make bin dimensions integer multiples of data resolution
- Quality default for origin offset based on data resolution

Frequency Binning

- Use 4 to 7 frequency bins
- Use standard log freq or quantile freq binning for desired interpretation

Practical Takeaways for Binned Scatterplots

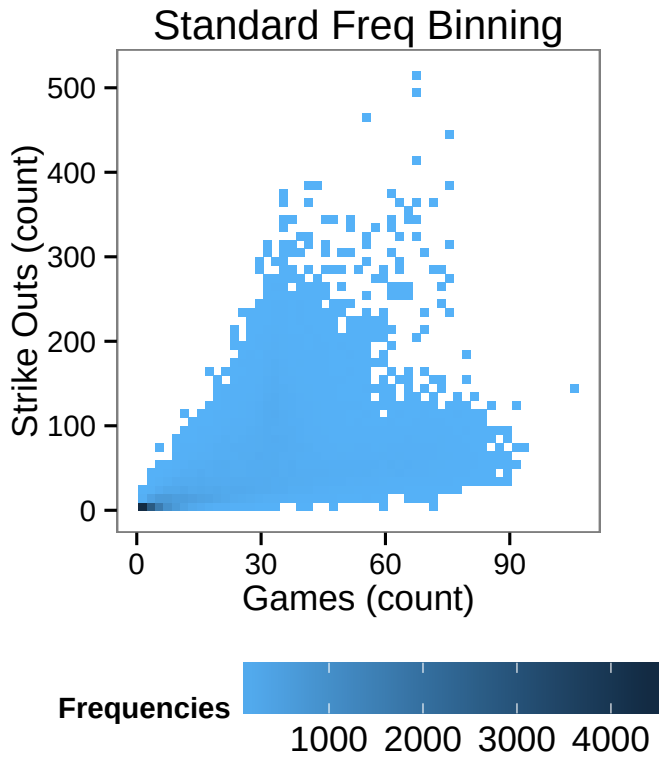
Spatial Binning

- Standard binning algorithm superior to random binning in spatial loss and computation speed
- Smaller bins superior for precisely displaying spatial information
- Make bin dimensions integer multiples of data resolution
- Quality default for origin offset based on data resolution

Frequency Binning

- Use 4 to 7 frequency bins
- Use standard log freq or quantile freq binning for desired interpretation

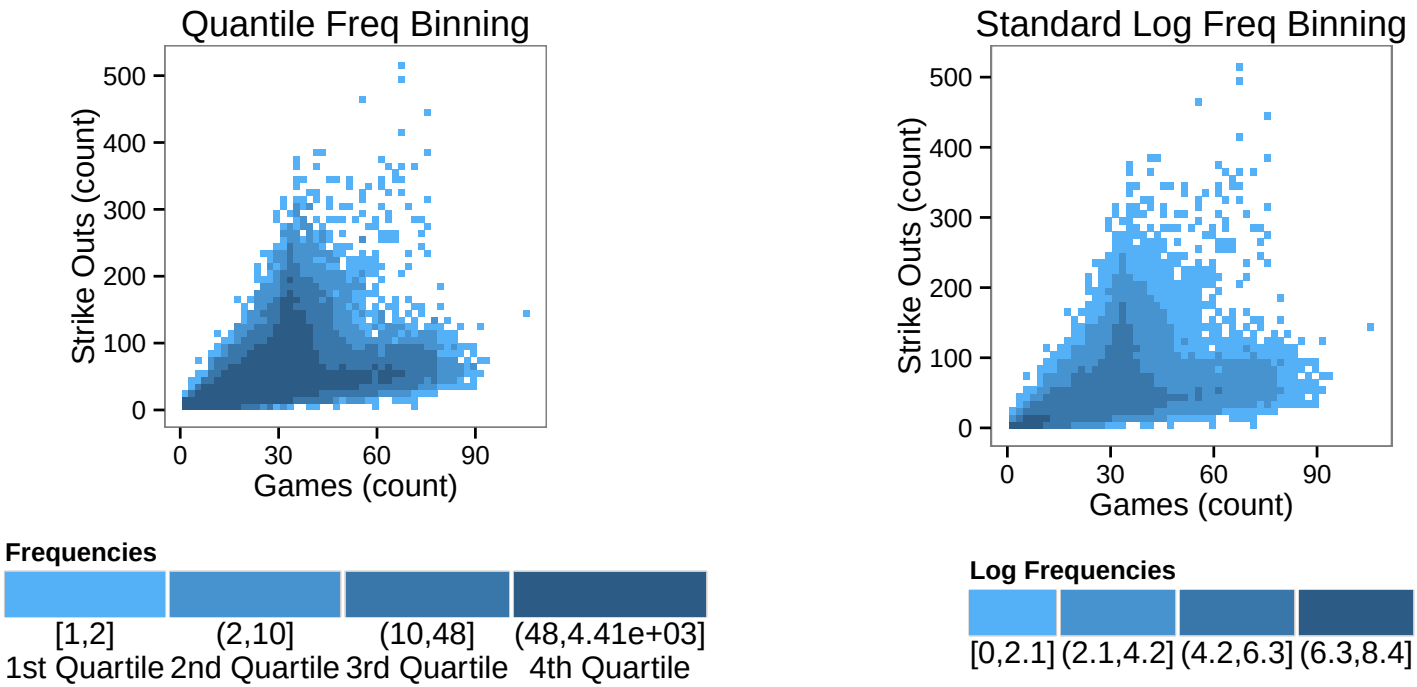
Application to Baseball Data



Use Standard Rectangular Binning

- Want approx 50 bins in each dimension $\Rightarrow$ use 2X10 bins
- Binning origin at (0.5,-0.5) $\Rightarrow$ net spatial loss is 3% lower than for origin at (1,0)

Application to Baseball Data



After frequency binning with 4 shade groups

A `shiny` New Opportunity for Interaction with Big Data in Undergraduate Education

Chapter 3

Shiny Database Sampler

Goals

- Connect intro students with big data
- Bypass technical hurdles of big data

Technical Components

- *Data storage*: MySQL
- *Computation*: `R`, `shiny`, `RMySQL`
- *User interface*: accessed through web browser

Interface Design

- Quality characteristics of software design
- Cognitive load theory
- Human centered design
- Pedagogical value and curricular integration

Shiny Database Sampler

Goals

- Connect intro students with big data
- Bypass technical hurdles of big data

Technical Components

- *Data storage*: MySQL
- *Computation*: `R`, `shiny`, `RMySQL`
- *User interface*: accessed through web browser

Interface Design

- Quality characteristics of software design
- Cognitive load theory
- Human centered design
- Pedagogical value and curricular integration

Shiny Database Sampler

Goals

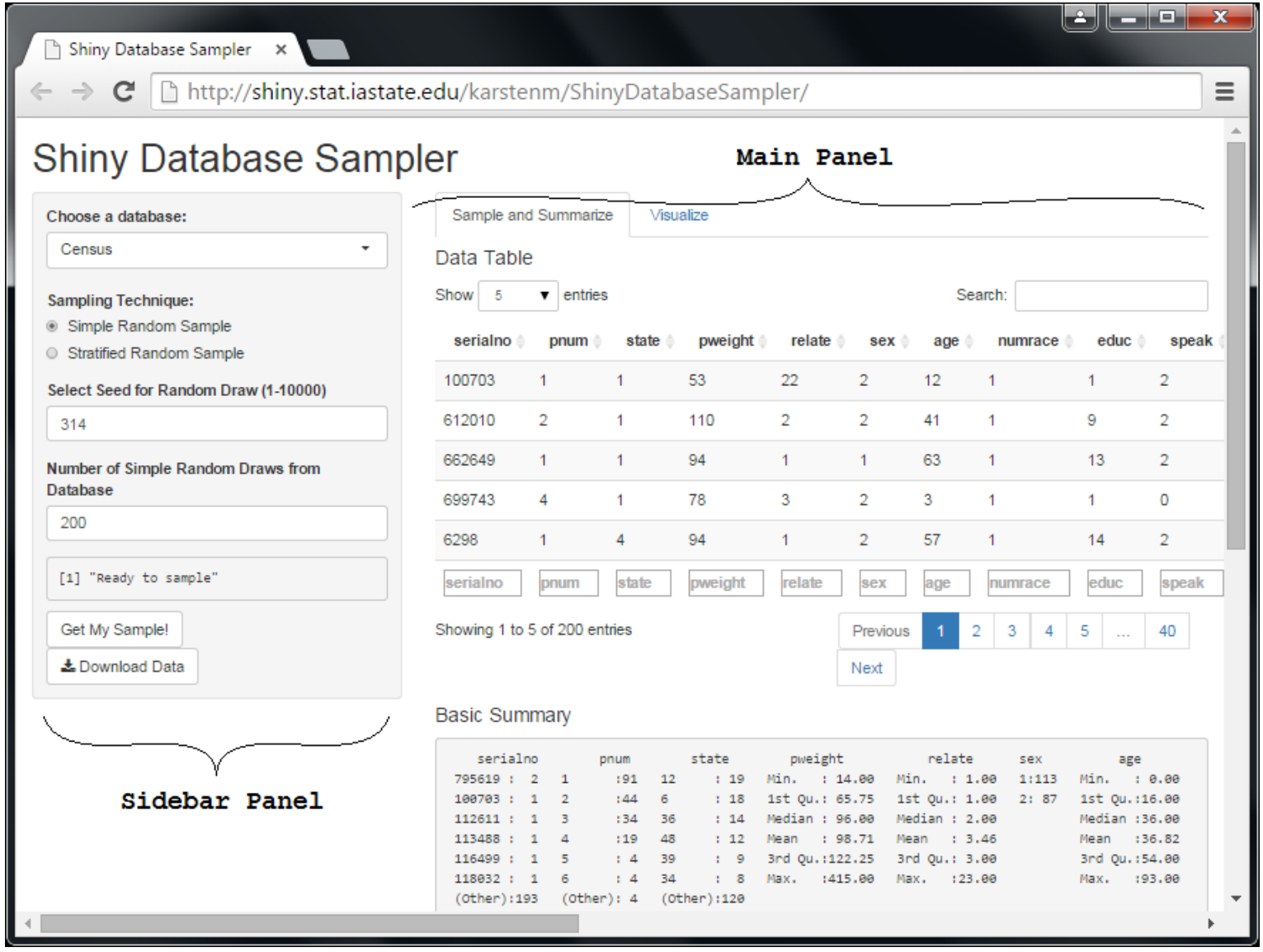
- Connect intro students with big data
- Bypass technical hurdles of big data

Technical Components

- *Data storage*: MySQL
- *Computation*: `R`, `shiny`, `RMySQL`
- *User interface*: accessed through web browser

Interface Design

- Quality characteristics of software design
- Cognitive load theory
- Human centered design
- Pedagogical value and curricular integration



User Interface

Applications

Lab Assignment

- Groups assignment
- Questions on hypothetical sampling goals
- Use Shiny Database Samplers to gather and explore samples

Course Projects

- Group project
- Research question with focus on bivariate associations
- Shiny Database Sampler use by some for mock surveys

Survey

Student User Survey

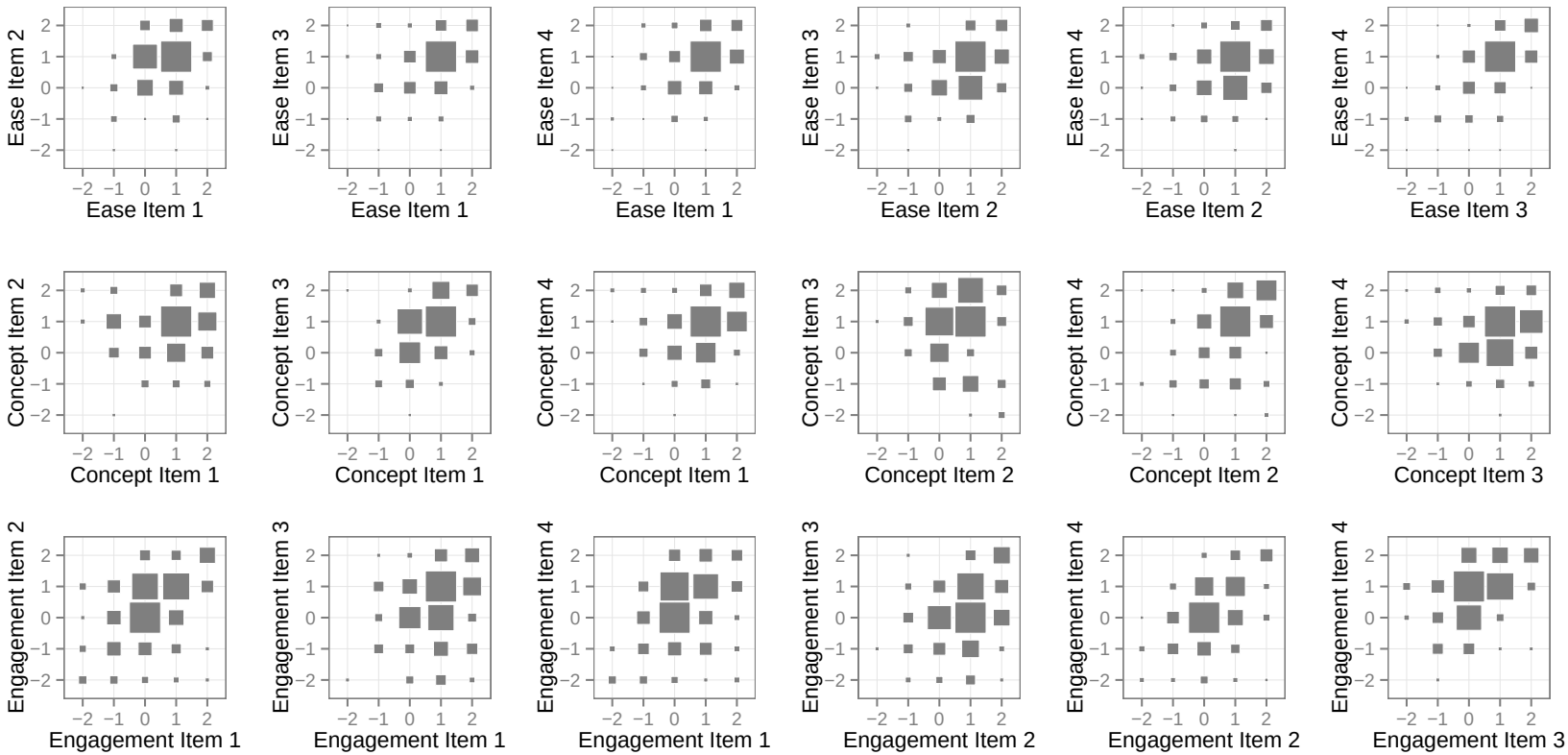
- 265 students from stat 104
- Survey followed lab assignment
- Likert scaled responses to 3 sets of 4 items
- Item Sets designed to focus on 3 attributes of interest
  - Ease of use
  - Connection to sampling concepts
  - Engagement with data
- Two per item set written in negative, then reverse scored

Survey

Topic Set	Item	Polarity
Ease	*I found the web tool easy to use*	+
	The layout of the web tool was intuitive	+
	Using the web tool was difficult	−
	Learning to use the web tool was hard	−
Concept	*The web tool helped me understand sampling concepts*	+
	I understand sampling ideas less after using the web tool	−
	Sampling techniques are clearer after using the web tool	+
	The web tool made me less sure how to randomly sample	−
Engagement	*I did not enjoy working with the Census data*	−
	I thought the Census data was boring	−
	Knowing that the Census data was from real people made it more interesting	+
	I liked analyzing the Census data	+

Table: Survey Items

Survey Internal Consistency



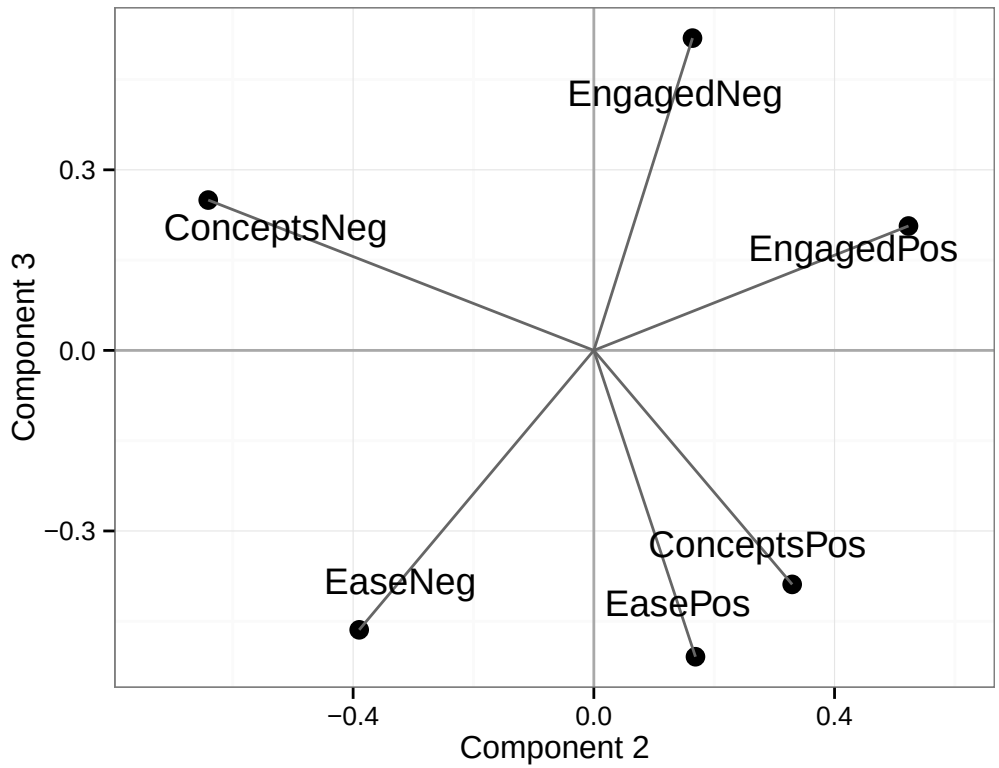
Fluctuation Diagrams of All Item Pairs within Topic Sets

Internal Consistency

Set	Estimate	95% Confidence Interval
Ease	0.70	(0.613 , 0.759)
Concept	0.53	(0.410 , 0.637)
Engagement	0.72	(0.643 , 0.776)

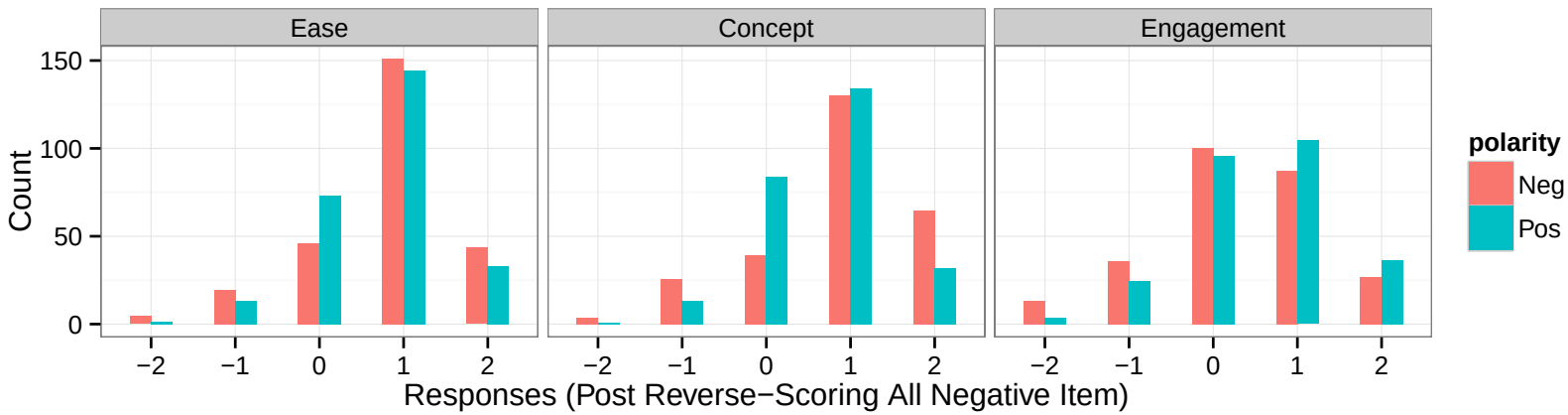
Table: Cronbach's α Estimates for each item set with 95% central bootstrap confidence interval based on 10,000 bootstrap samples

Survey



Item Pair Loadings on Components 2 and 3 from the Principal Component Analysis with Six Topic/Polarity Item Pairs

Survey



Item Set Response Distributions by Polarity

Results

Ease of use

- Students found the interface easy to use

Connection to sampling concepts

- Student views were neutral toward the learning benefit
- However they did not believe it hurt their understanding

Engagement with data

- Students were moderately engaged with the census data

Thanks!

Thank you for attending!

Any questions?