

Scatterplots for Massive Datasets

Martin Theus, Di Cook, Heike Hofmann

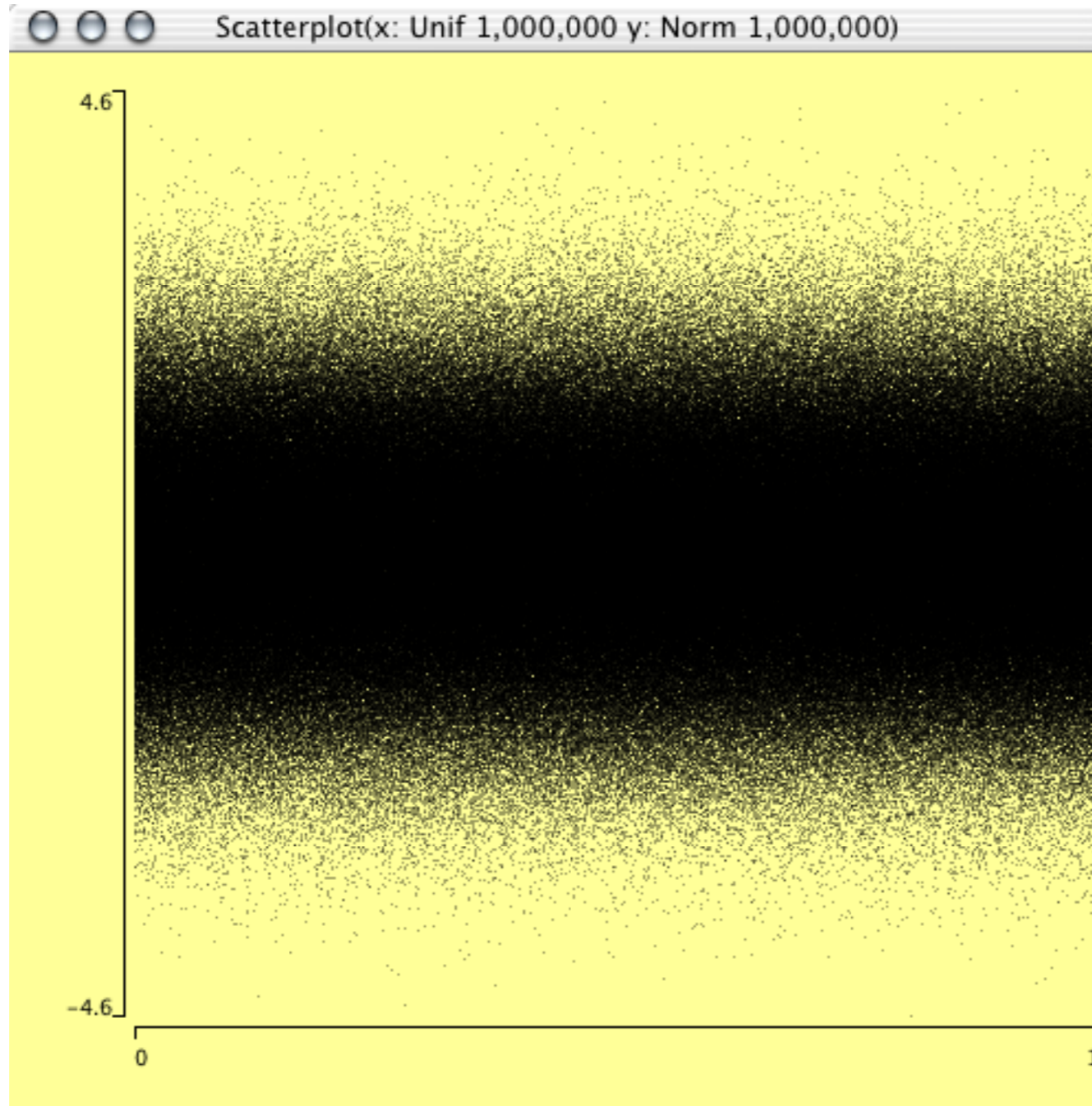
Department of Statistics
Iowa State University

Les Miller, Jing Zhang

Department of Computer Science
Iowa State University

Why binning scatterplots?

- Just too many data!
- Overplotting obscures the structure in the data

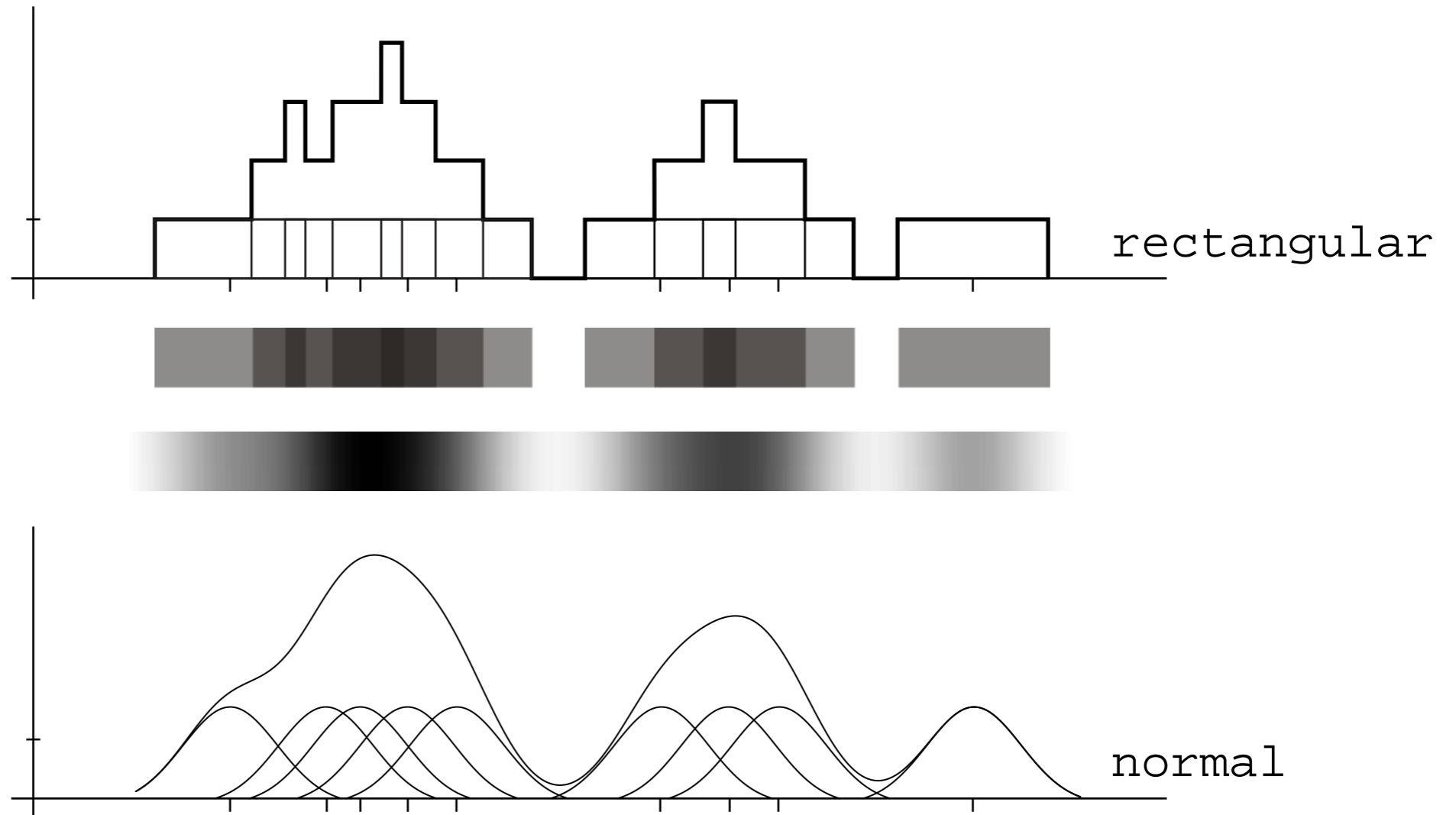


1,000,000
points

What solutions do exist?

- Tonal highlighting (i.e. 2-dim density estimator)
 - in memory
 - via alpha blending

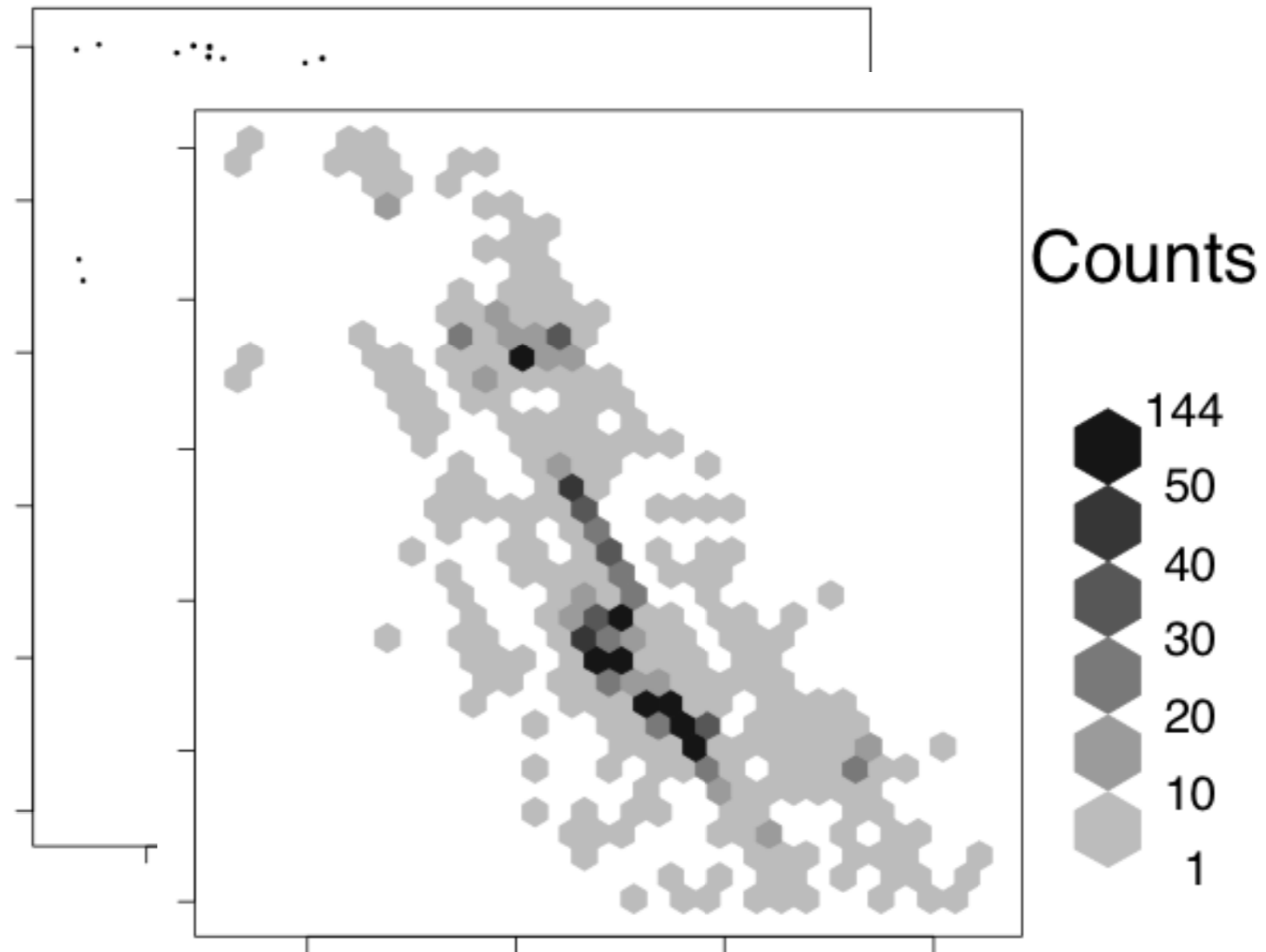
$$\hat{f}_{th} = \frac{1}{cn} \sum_{i=1}^n k\left(\frac{x - x_i}{c}\right) = \hat{f}_{kernel} = \frac{1}{cn} \sum_{i=1}^n k\left(\frac{x_i - x}{c}\right) \quad \text{for } k(x) = k(-x) \quad \forall x \in \mathbb{R}.$$



What solutions do exist?

- Binned scatterplots

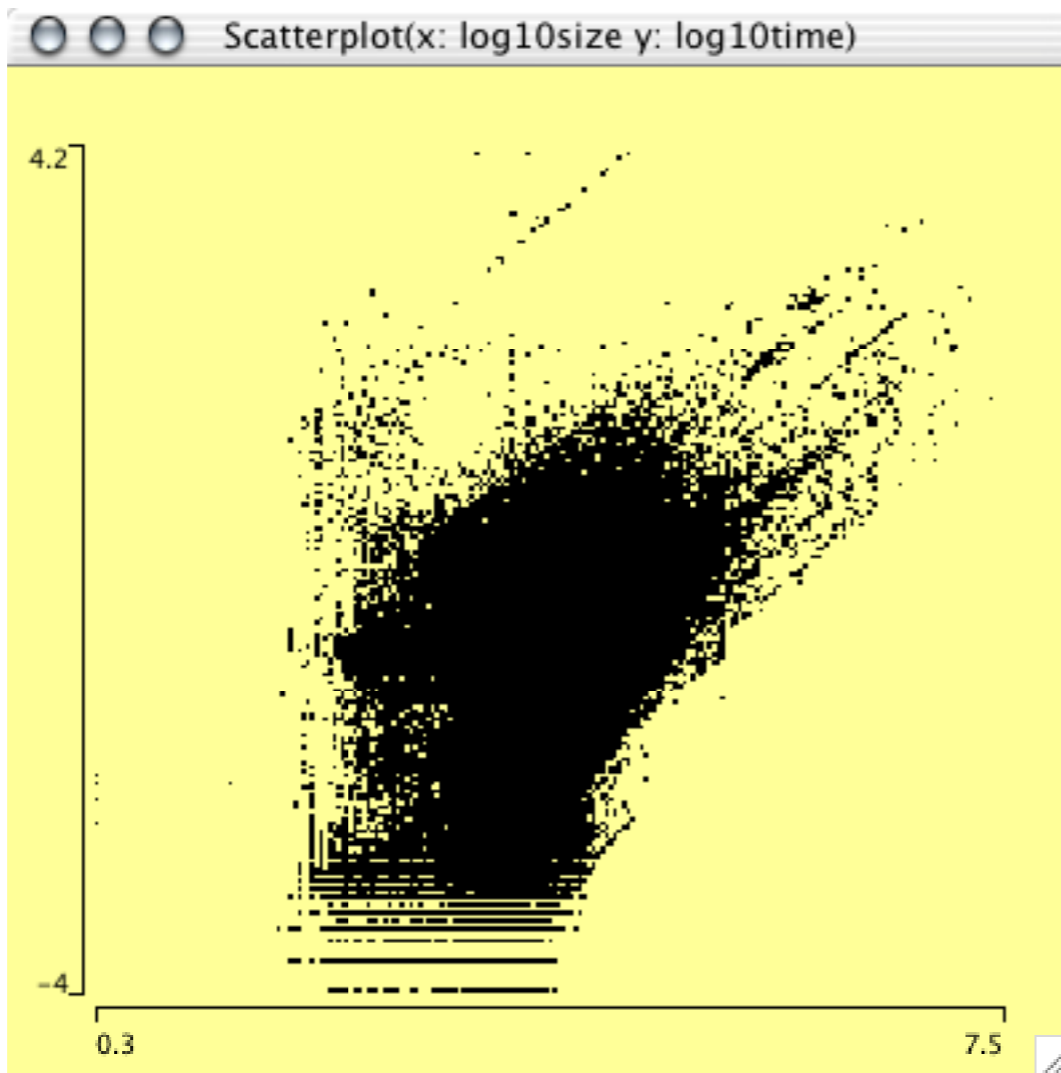
What you should not do:
Don't use glyph sizes



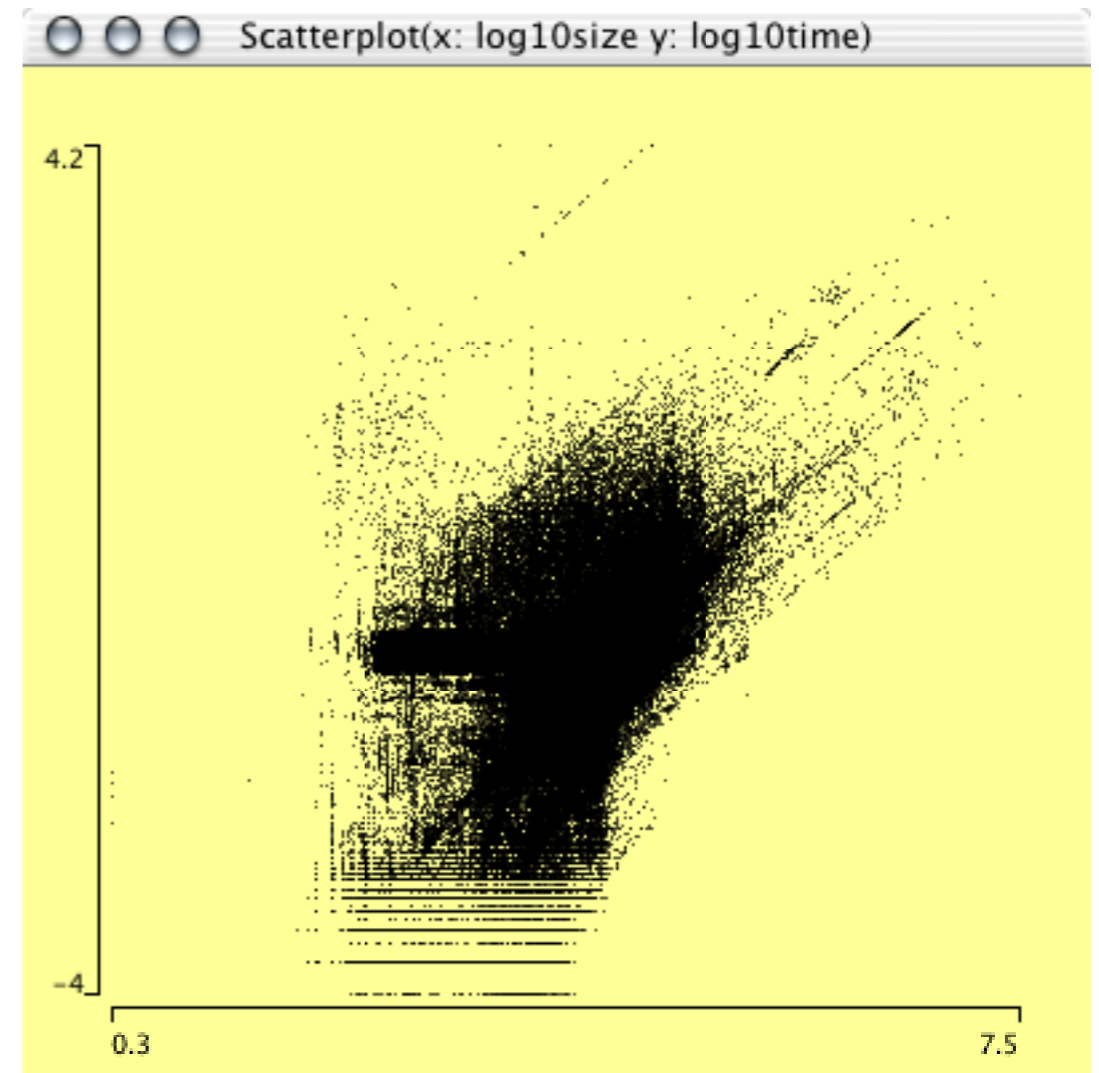
What you should do:
Use grayscales to map densities

How much information do we lose?

- Which plot is binned?



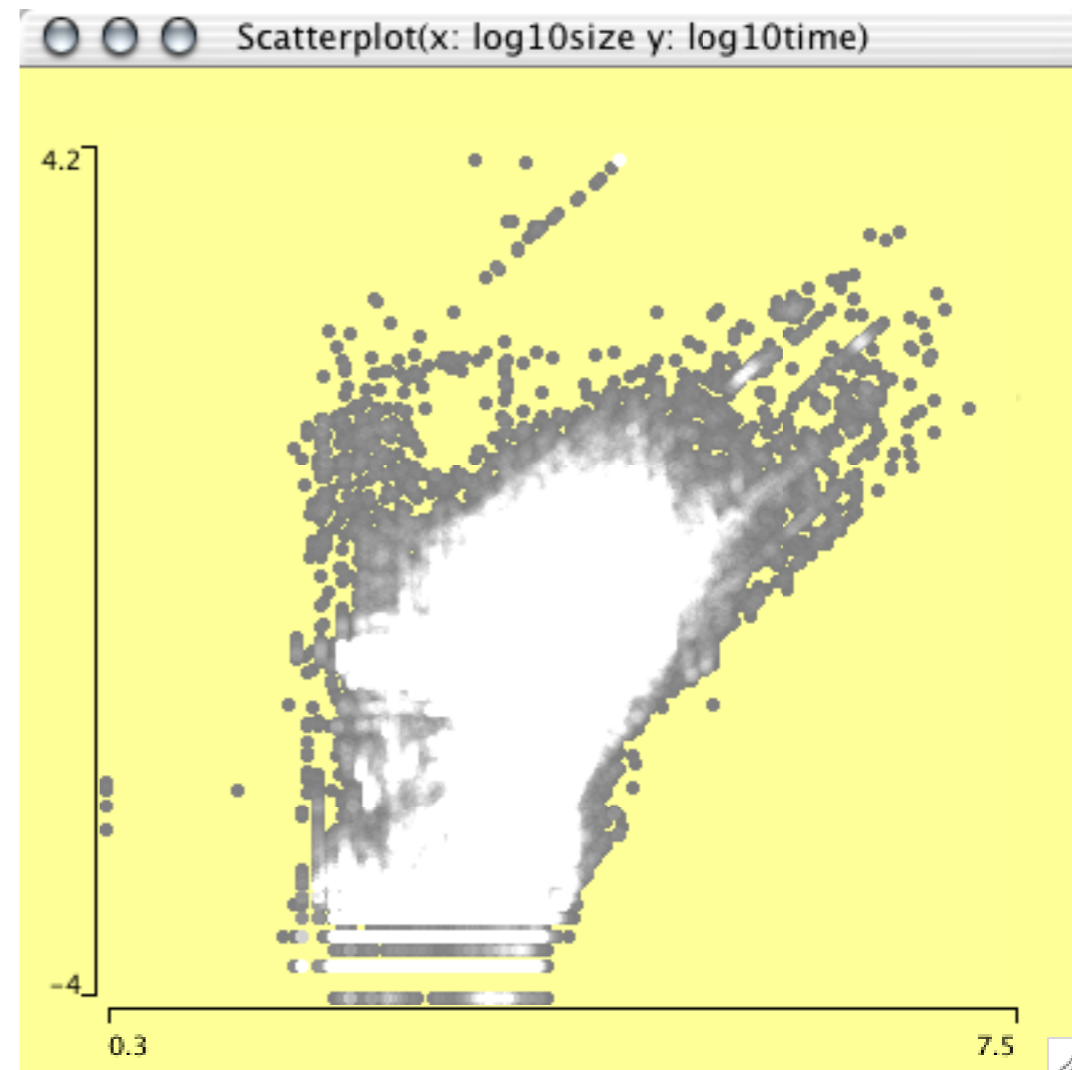
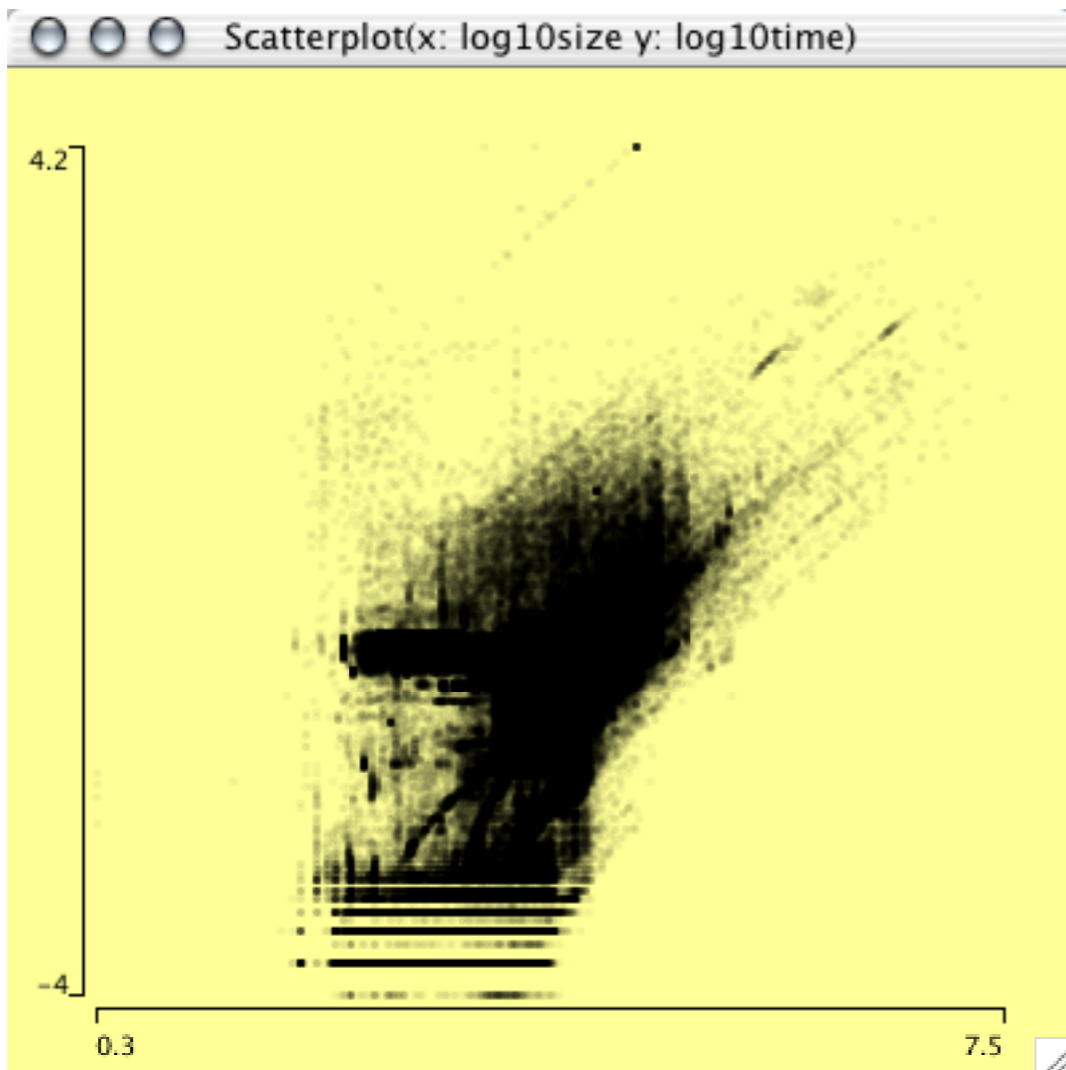
binned to a 256x256 grid



binned to 400x400 pixel

Choice of gray scale mapping matters!

- Which plot do you like better?



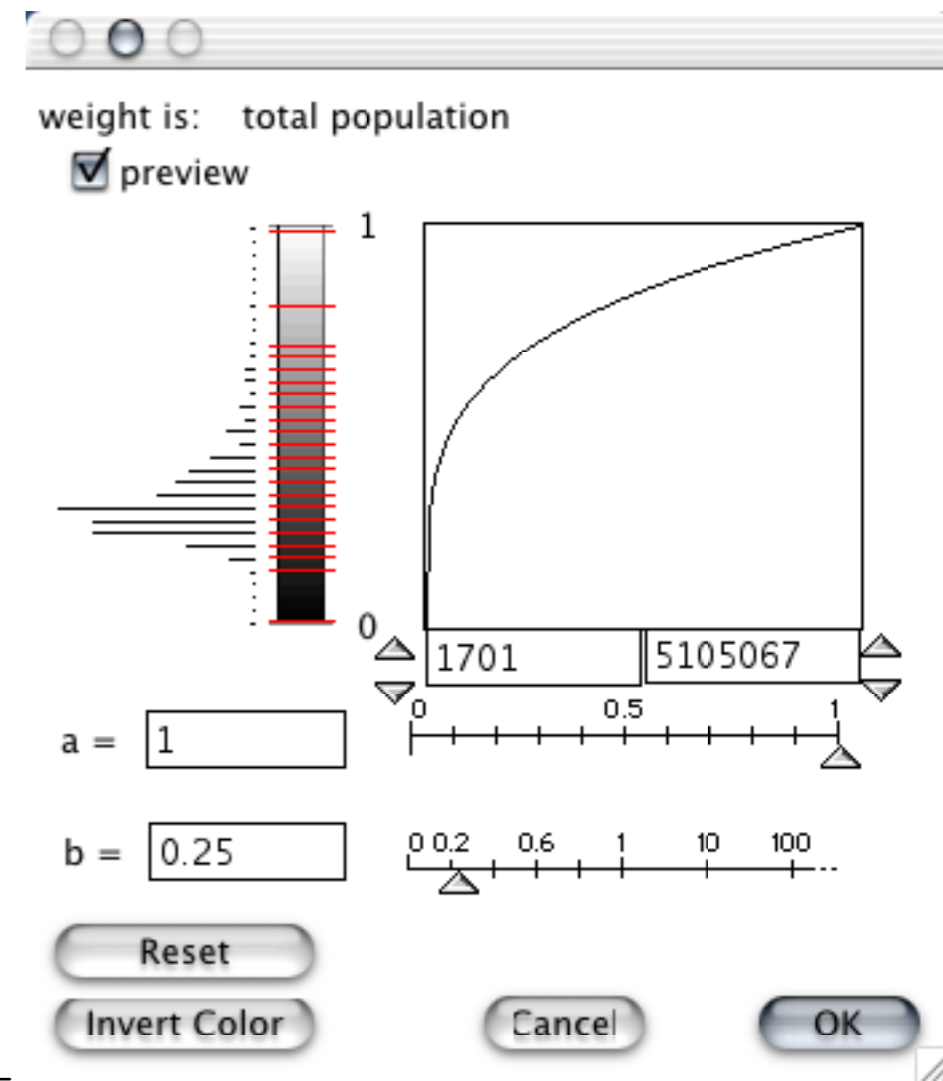
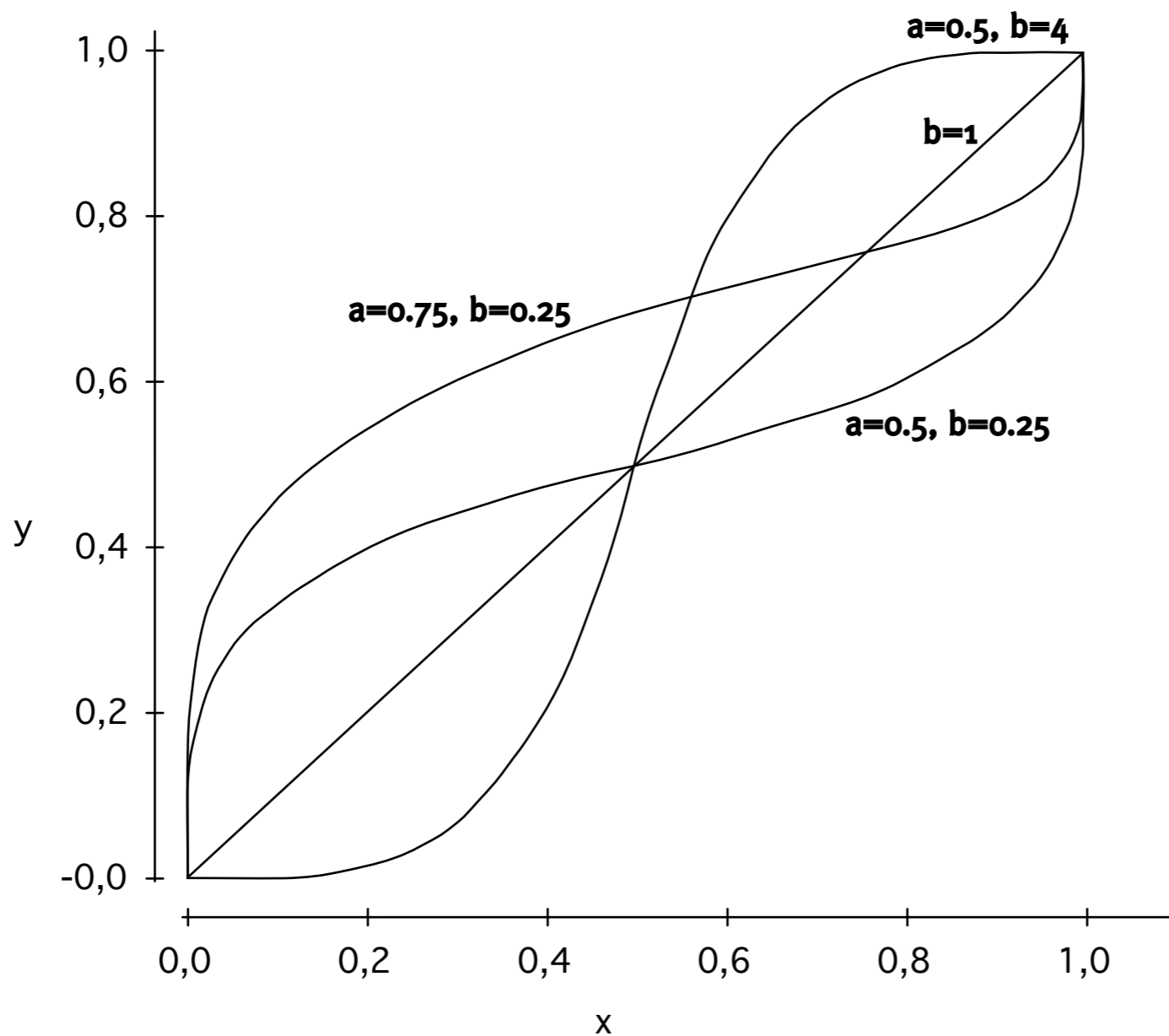
(this one is stolen from MANET)

- For both cases: we need a non-white background
- Interactive controls are very important

How to control the mapping!

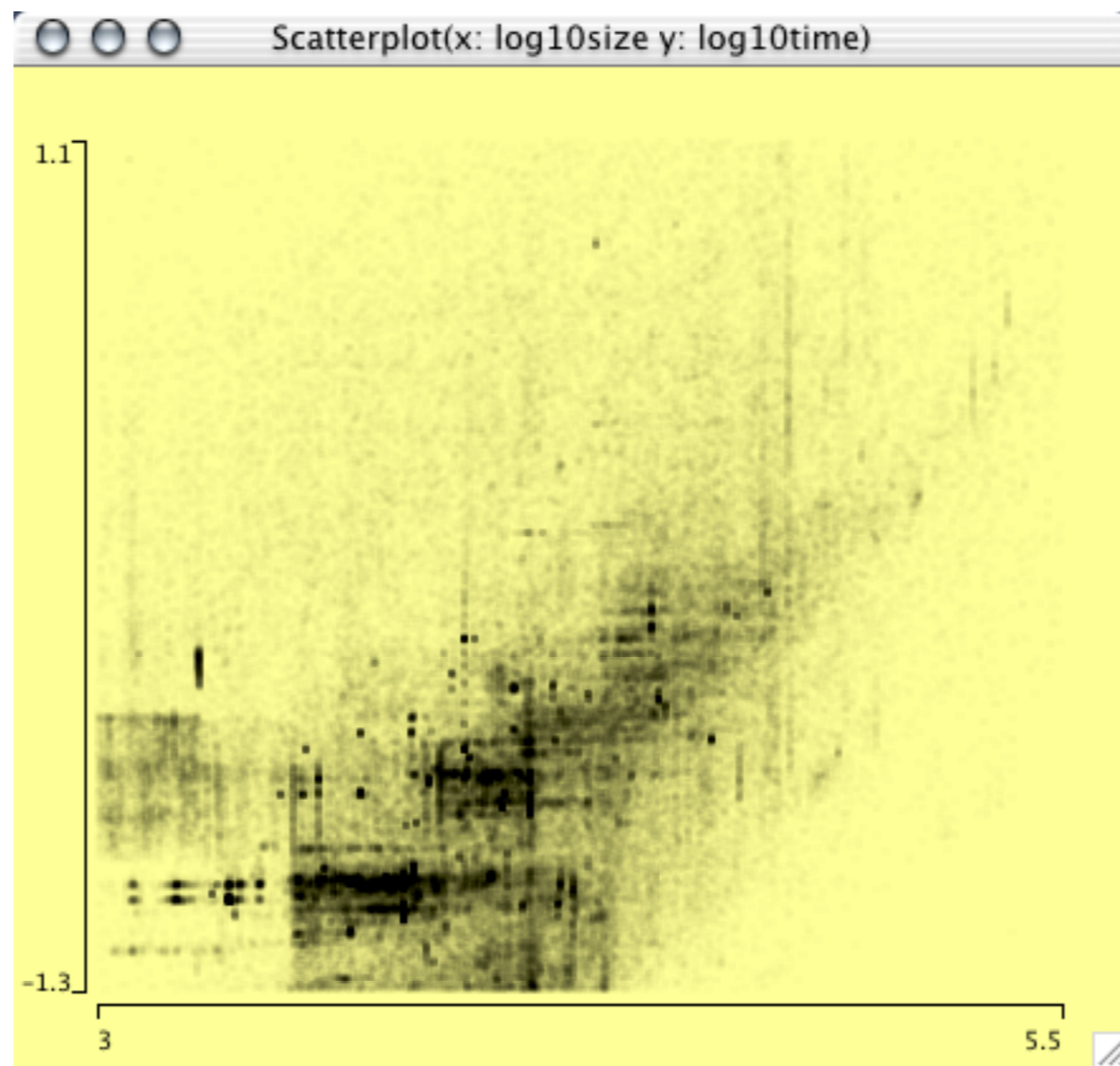
- 2-parameter, S-shaped link function:

$$f(x) := \begin{cases} a \cdot \left(\frac{x}{a}\right)^b & \text{for } x \leq a \\ 1 - (1 - a) \cdot \left(\frac{1-x}{1-a}\right)^b & \text{for } x \geq a \end{cases} \quad \text{for } a \in [0, 1] \text{ and } b > 0$$



More Interactivity: Zooming!

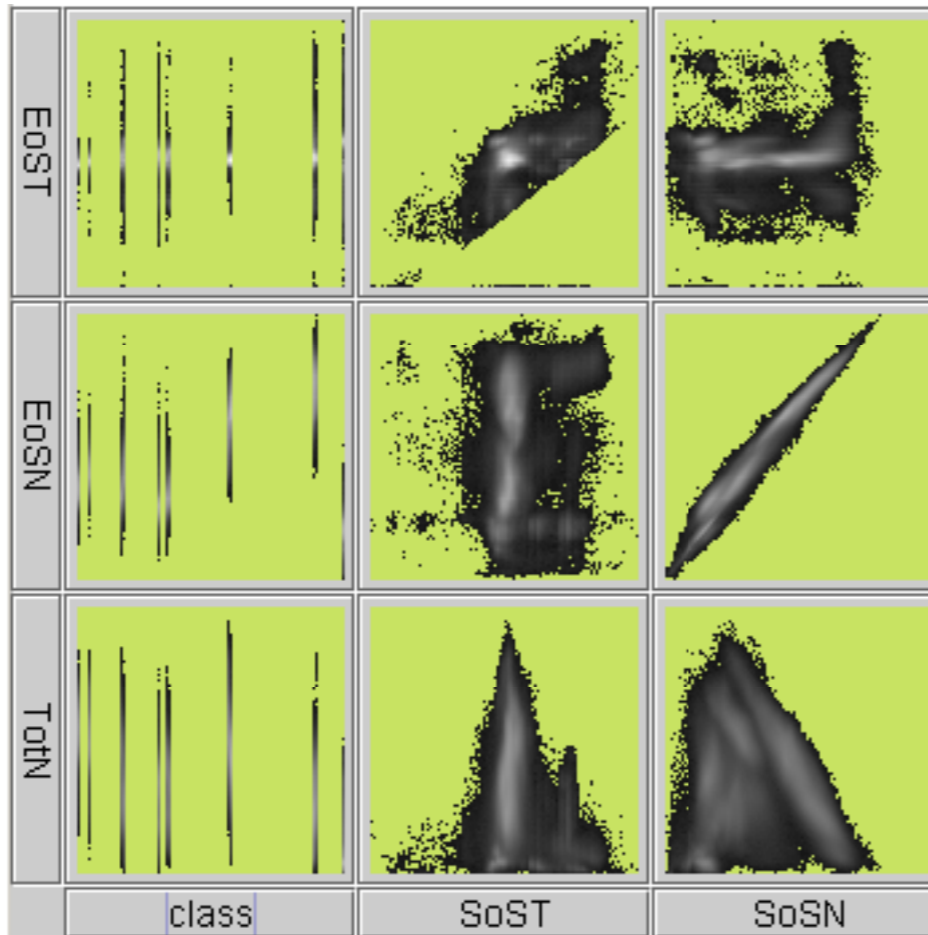
- Structure may only be visible at different resolutions



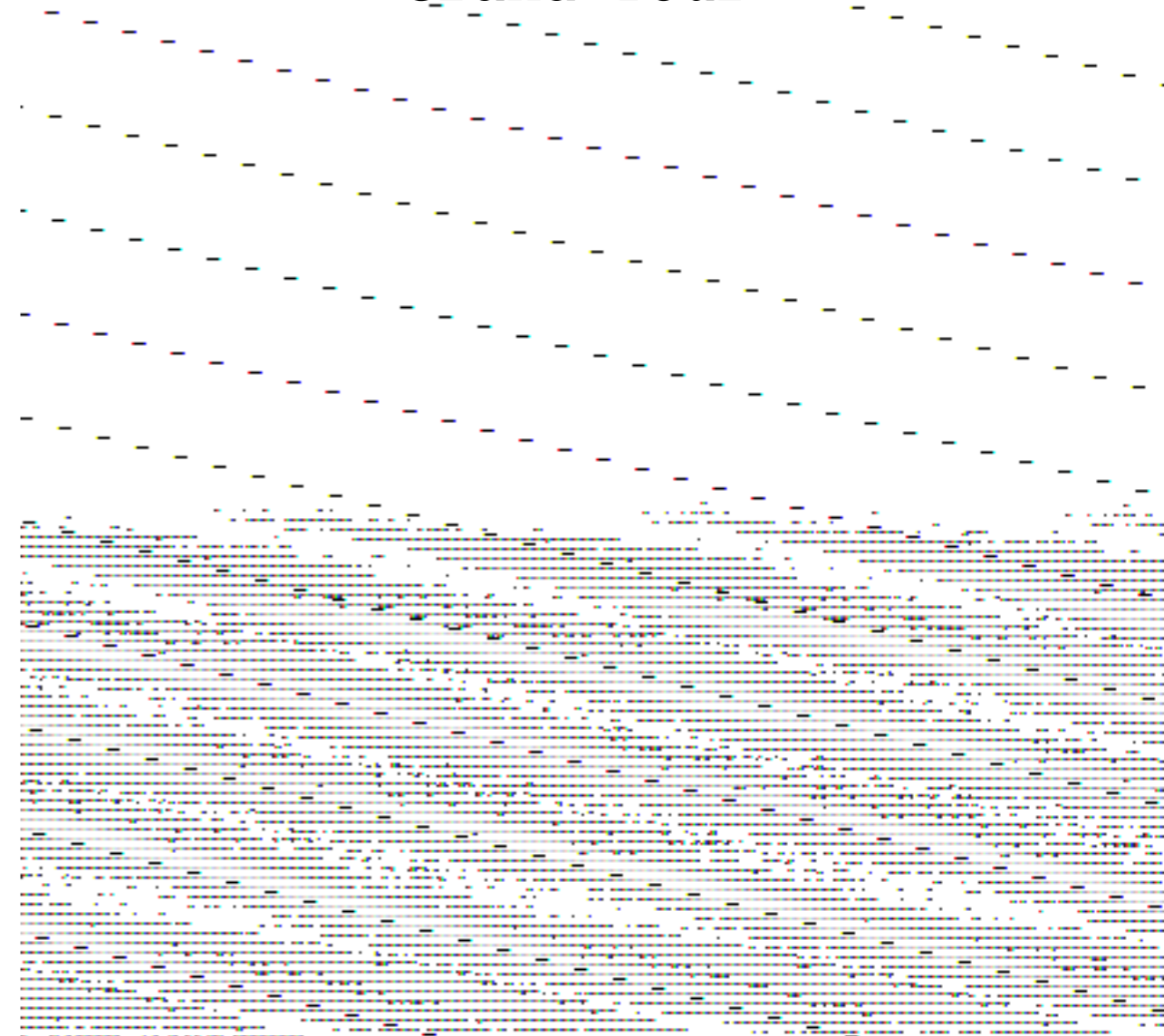
Ultimate binning: binning to pixel

- Idea: Represent data only in binned form
- Advantage: No loss of information
- Challenge: Mapping between data and indices
- Software: Limn

Scatterplot Matrix:



Grand Tour:



Open Questions

- How to handle highlighting?

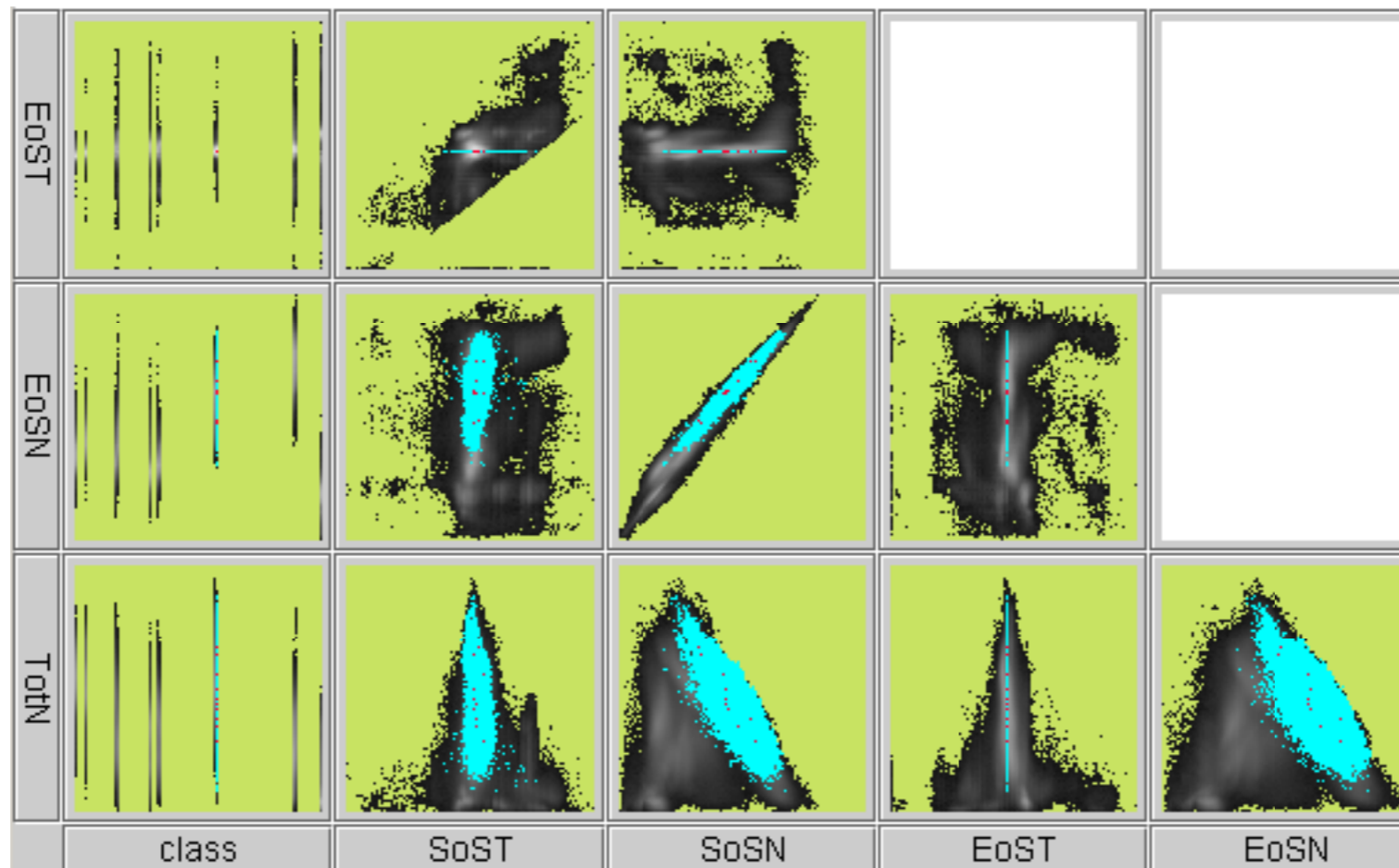
alpha-blending
works fine for
grayscales ..



... but is hard
to interpret
with colors.



First version: no density representation in the highlighted data



Summary

- With large data, we need modifications to ordinary scatterplots
- Overplotting masks the structure in plots
- Alpha blending can be the cheapest solution, if the graphics hardware is used
- Binning reduces the graphics complexity to a constant
- Offline binning still allows interactive and dynamic graphics of very large data ($>10^7$)
- Assignment of grayscales depends on the data
- Interactivity is a key feature
- Software:
 - Limn (JAVA)
 - MANET (OS X)
 - Mondrian (JAVA)