Interactive Statistical Graphics

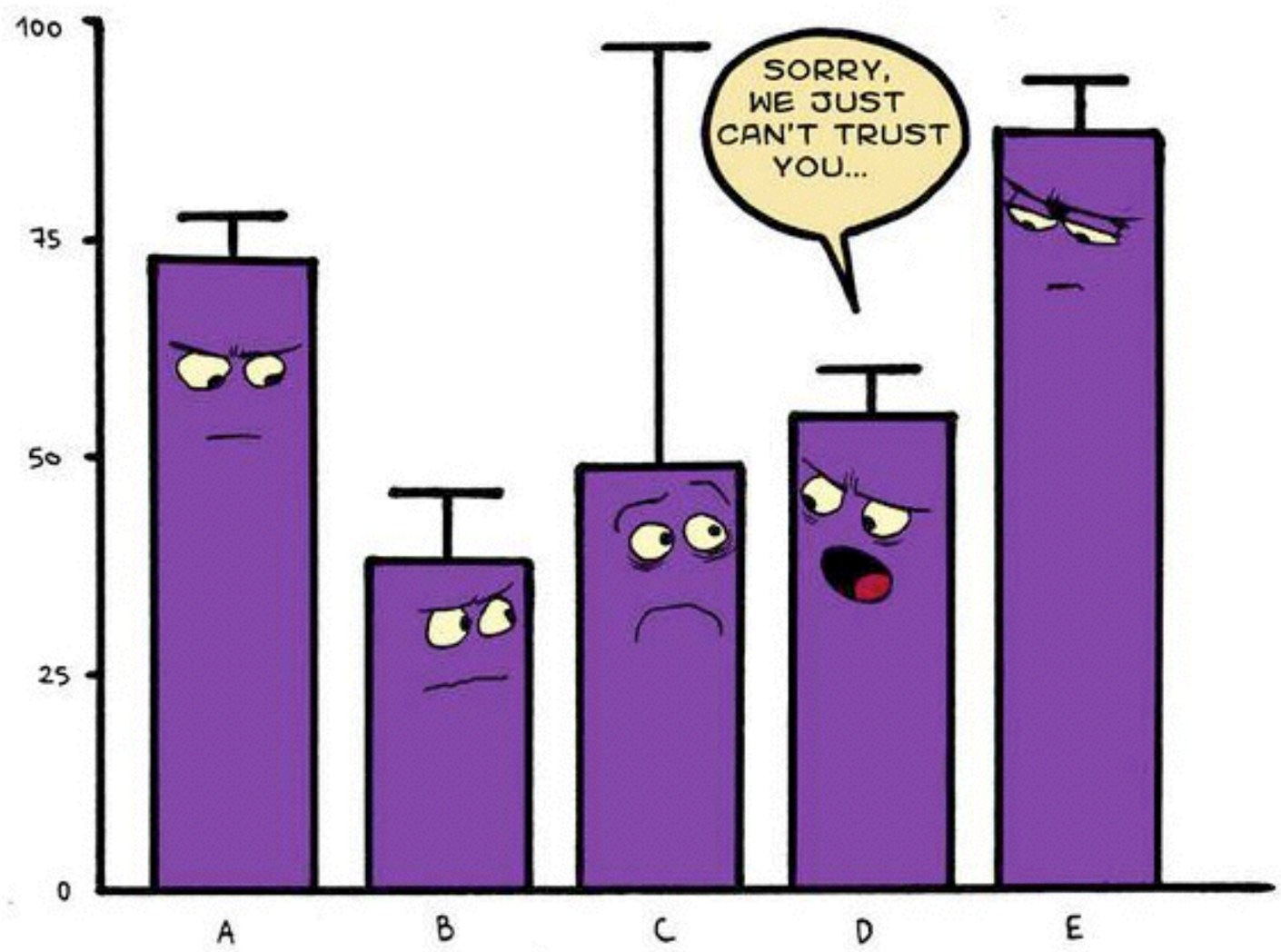
When Charts come to Life

martin@theusRus.de

www.theusRus.de

Telefónica Germany

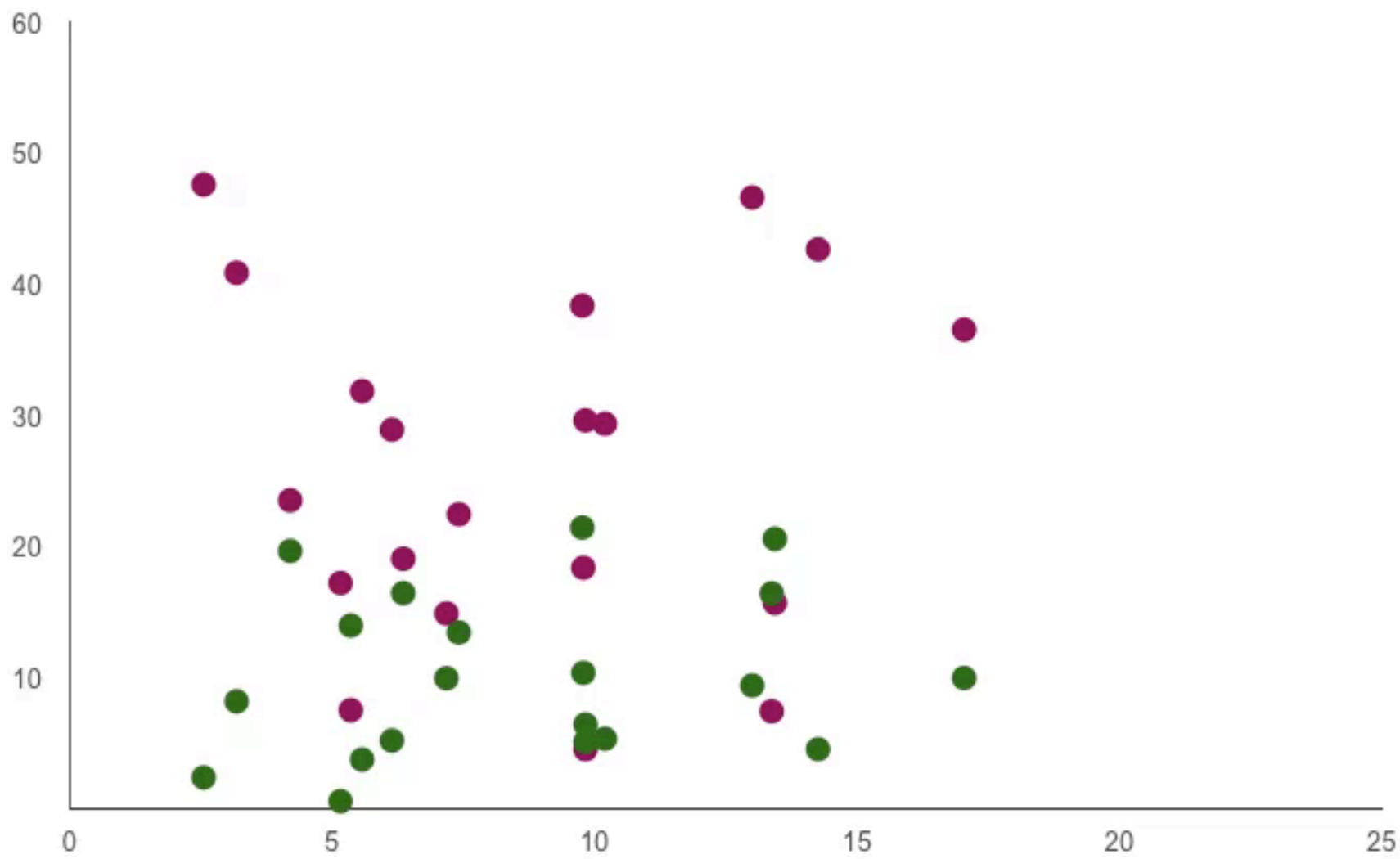
What I do not talk about ...



… still not what I mean.

Animation

This page describes how to animate modifications made to a chart, instead of applying them instantly.



Interactive Graphics ≠ Dynamic Graphics

- Interactive Graphics
  … uses various interactions with the plots to change selections and parameters quickly.

Interactive Graphics ≠ Dynamic Graphics

- Interactive Graphics
  … uses various interactions with the plots to change selections and parameters quickly.

- Dynamic Graphics
  … uses animated / rotating plots to visualize high dimensional (continuous) data.

Interactive Graphics ≠ Dynamic Graphics

- Interactive Graphics
  … uses various interactions with the plots to change selections and parameters quickly.

- Dynamic Graphics
  … uses animated / rotating plots to visualize high dimensional (continuous) data.

1973
PRIM-9
Tukey et al.

Interactive Graphics ≠ Dynamic Graphics

- Interactive Graphics
  … uses various interactions with the plots to change selections and parameters quickly.

- Dynamic Graphics
  … uses animated / rotating plots to visualize high dimensional (continuous) data.

1973
PRIM-9
Tukey et al.

1983
SPLOM
Becker et al.

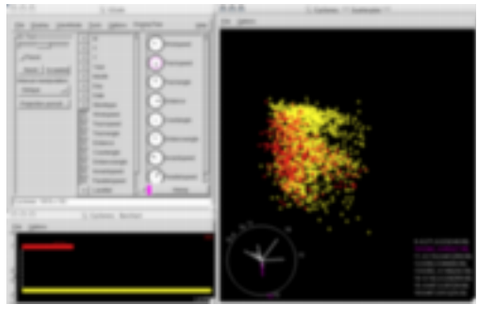
Interactive Graphics ≠ Dynamic Graphics

- Interactive Graphics
  … uses various interactions with the plots to change selections and parameters quickly.

- Dynamic Graphics
  … uses animated / rotating plots to visualize high dimensional (continuous) data.

1973
PRIM-9
Tukey et al.

1983
SPLOM
Becker et al.



1999
ggobi
Swayne et al.

Why do we use Graphics (not only in Statistics)?

- Classical ➢ **Presentation**
  The most common use of graphics is clearly in presenting qualitative or quantitative results to a broad audience

Why do we use Graphics (not only in Statistics)?

- Classical ➢ **Presentation**
  The most common use of graphics is clearly in presenting qualitative or quantitative results to a broad audience

- Statistical ➢ **Diagnostics**
  In statistics, graphics are often used to check the quality and properties of statistical procedures or models

Why do we use Graphics (not only in Statistics)?

- Classical ➢ **Presentation**
  The most common use of graphics is clearly in presenting qualitative or quantitative results to a broad audience

- Statistical ➢ **Diagnostics**
  In statistics, graphics are often used to check the quality and properties of statistical procedures or models

- Analytical ➢ **Exploration**
  During the exploration process of an analysis, graphics aid to generate insights and deduce properties and relationships

Why do we use Graphics (not only in Statistics)?

- Classical ➢ **Presentation**
  The most common use of graphics is clearly in presenting qualitative or quantitative results to a broad audience

- Statistical ➢ **Diagnostics**
  In statistics, graphics are often used to check the quality and properties of statistical procedures or models

- Analytical ➢ **Exploration**
  During the exploration process of an analysis, graphics aid to generate insights and deduce properties and relationships

- Essential ➢ **Data Cleaning**
  Whenever we get to work on raw (dirty) data, it is essential to find, understand and clean up artifacts and errors

Distinction of Talks ...

my talk

Distinction of Talks ...

Antony's talk

Elements of Interactive Statistical Graphics

- The 4 pillars of ISG
  - **Selection**
    selection of a subgroup of interest
  - **Highlighting**
    highlight a selected subgroup across all plots
  - **Query**
    query information on objects for non-obvious information
  - **Modification & “Statistification”**
    change plot parameters quickly and
    include statistical estimates and models easily

Elements of Interactive Statistical Graphics

- The 4 pillars of ISG
  - **Selection**
    selection of a subgroup of interest

    Link!

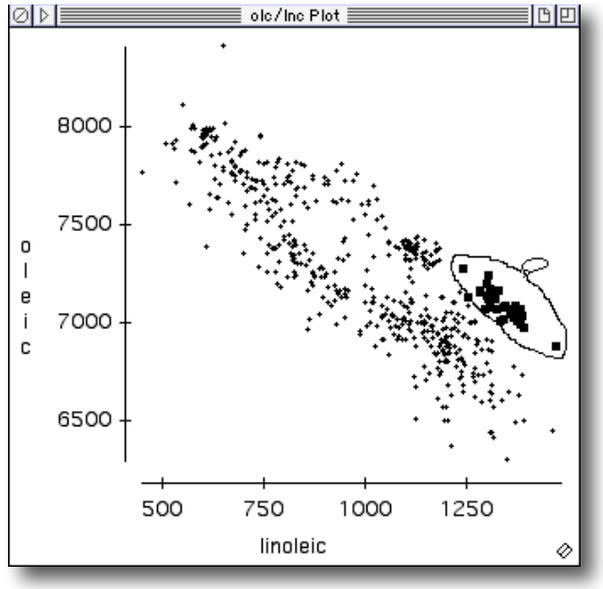
  - **Highlighting**
    highlight a selected subgroup across all plots
  - **Query**
    query information on objects for non-obvious information
  - **Modification & “Statistification”**
    change plot parameters quickly and
    include statistical estimates and models easily

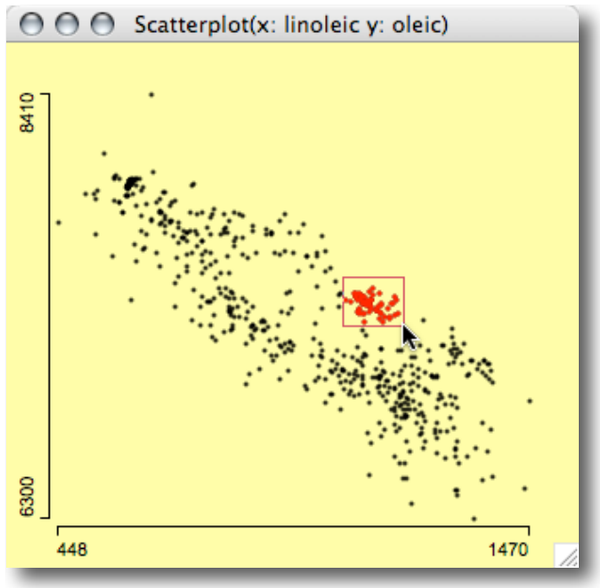
Selections

- Selections as such are not really interesting – but they are the necessary step to specify subsets of interest
- In an exploratory set-up we often want to look at the properties of specific subgroups, like

  "Find all customers, who paid less than 15% tip, on weekends!"

- The flexibility with which we can select data directly determines how successful we may solve the exploratory analysis.
- Obviously we need different selection **tools** and selection **modes**





Selections

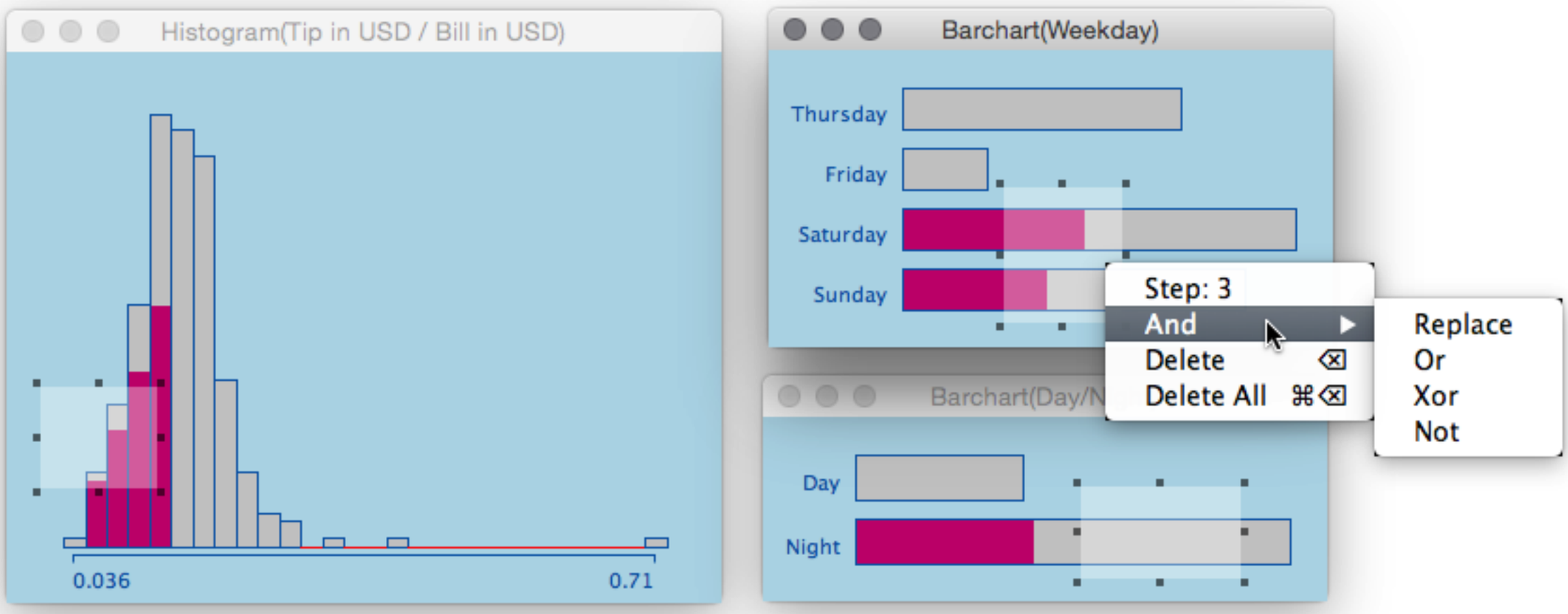
- **Tools** to select data:
  - **Pointer**
    … is used to select single points.
  - **Drag-Box**
    … selects rectangular regions in a graphics window.
  - **Brush**
    … allows a dynamic change (movement) of the selected region – usually a rectangle.
  - **Slicer**
    … selects intervals along an axis dynamically.
  - **Lasso**
    … allows the most flexible definition of the selection area.
    Startpoint and endpoint are always connected.

- **Modes** to select data:
  - **Simple / Standard / Default**
    … only points in the selected region are selected.
  - **Intersection / AND / ∩**
    … only points that already were selected and are within the new selection stay selected.
  - **Union / OR / ∪**
    … the newly selected points are added to the current selection.
  - **Toggle / XOR / ⊕**
    … selected points are deselected, unselected are selected.
  - **Negation / NOT / ¬**
    … points in the selection region are taken out of the current selection set.

Selections

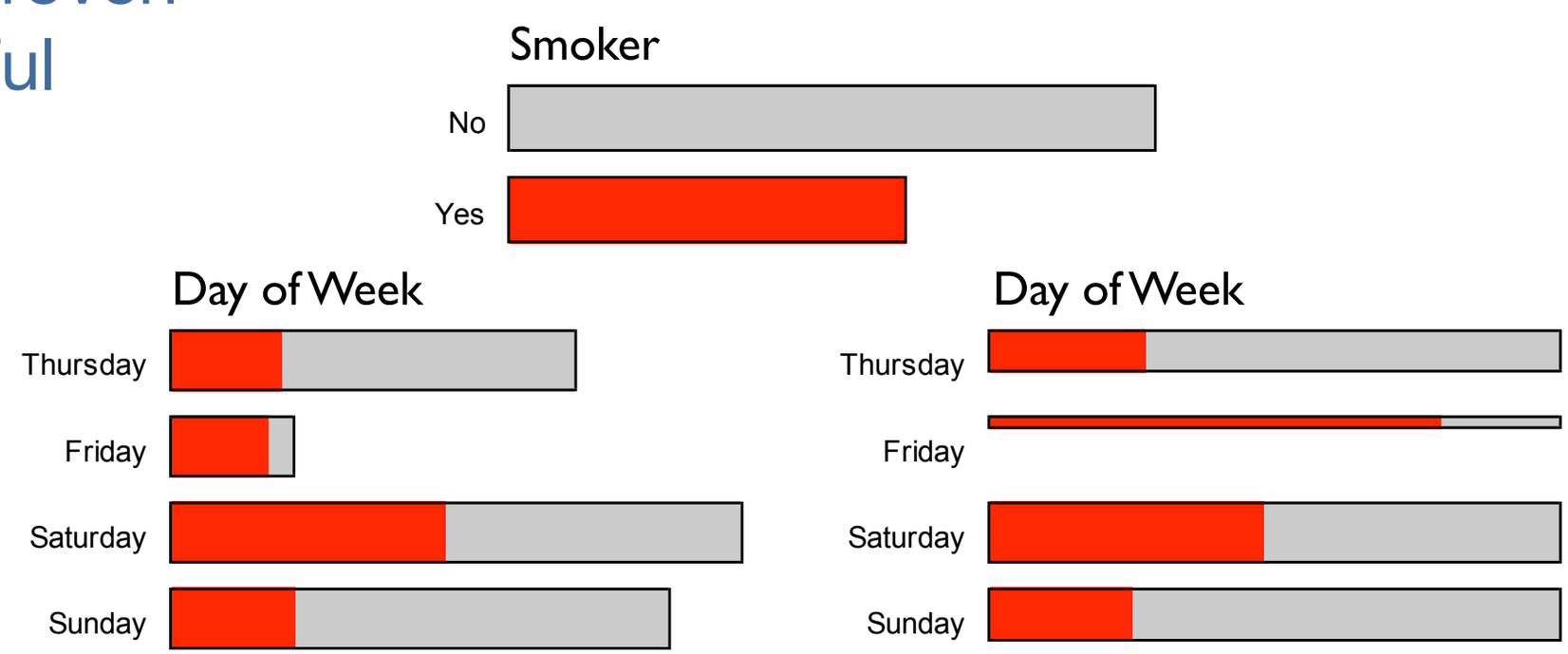
- **Selection Sequences** allow to select quite complex subsets

  "Find all customers, who paid less than 15% tip, at night, on weekends!"



Highlighting

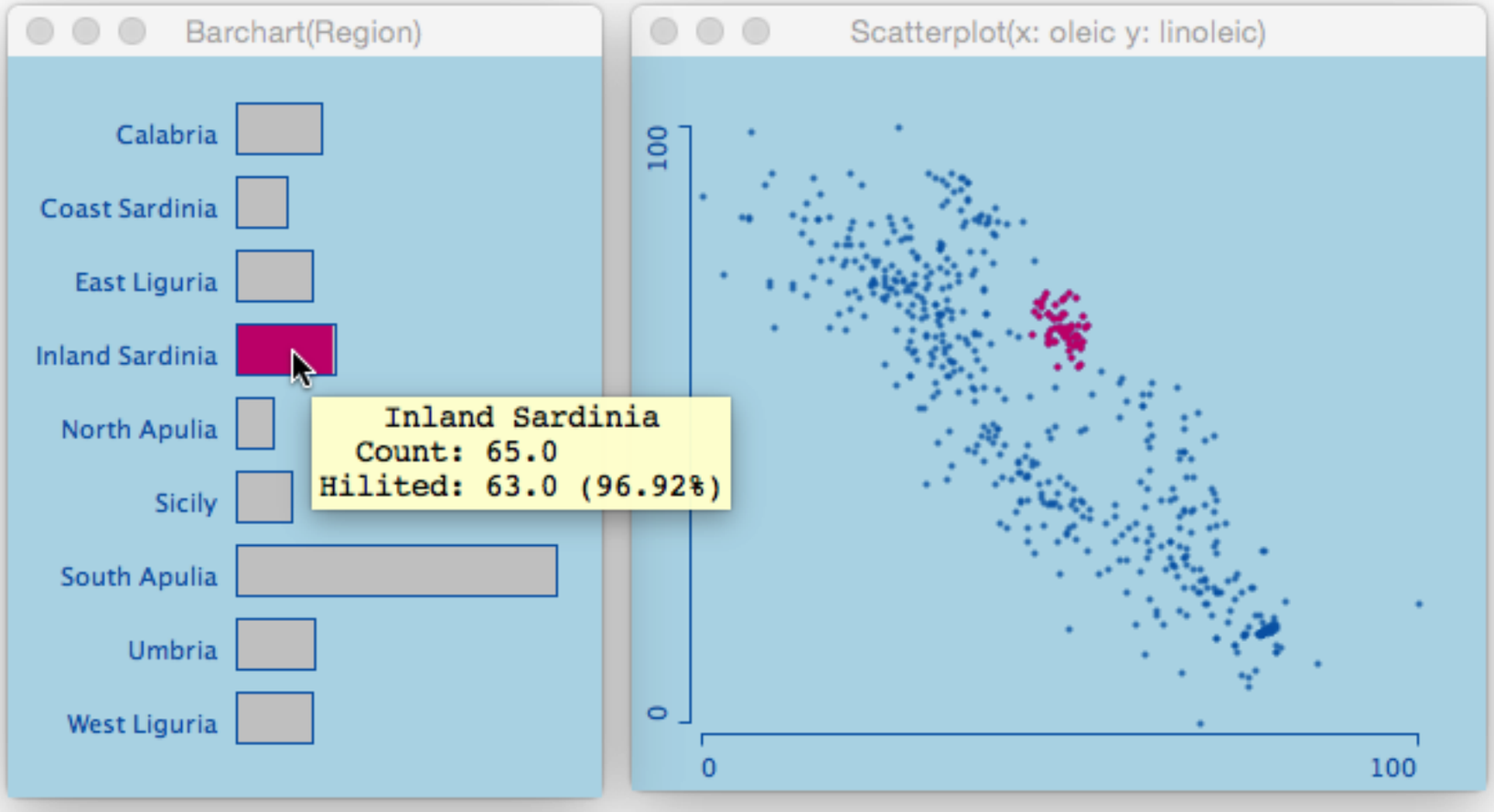
- Once a selection is defined, it needs to be propagated to all other plots,
- thus all plots need to know how to highlight a subgroup
- Highlighting may be
  - **transient** (only changes when a new selection is performed)
  - **persistent** (a new state explicitly must be assigned to the involved cases)
- A clear rule how highlighting is performed is desirable, but exceptions have proven to be quite powerful



Example:
Barchart/Spineplot

Queries

- Graphics are good at communicating qualitative information but fail to give exact quantities ⇒ need queries to get exact values
- Gridlines can help (only) for the variables within the plot
- Interactive graphics often display very little scale information (cf. Tufte's "data-ink-ratio")
- Example:



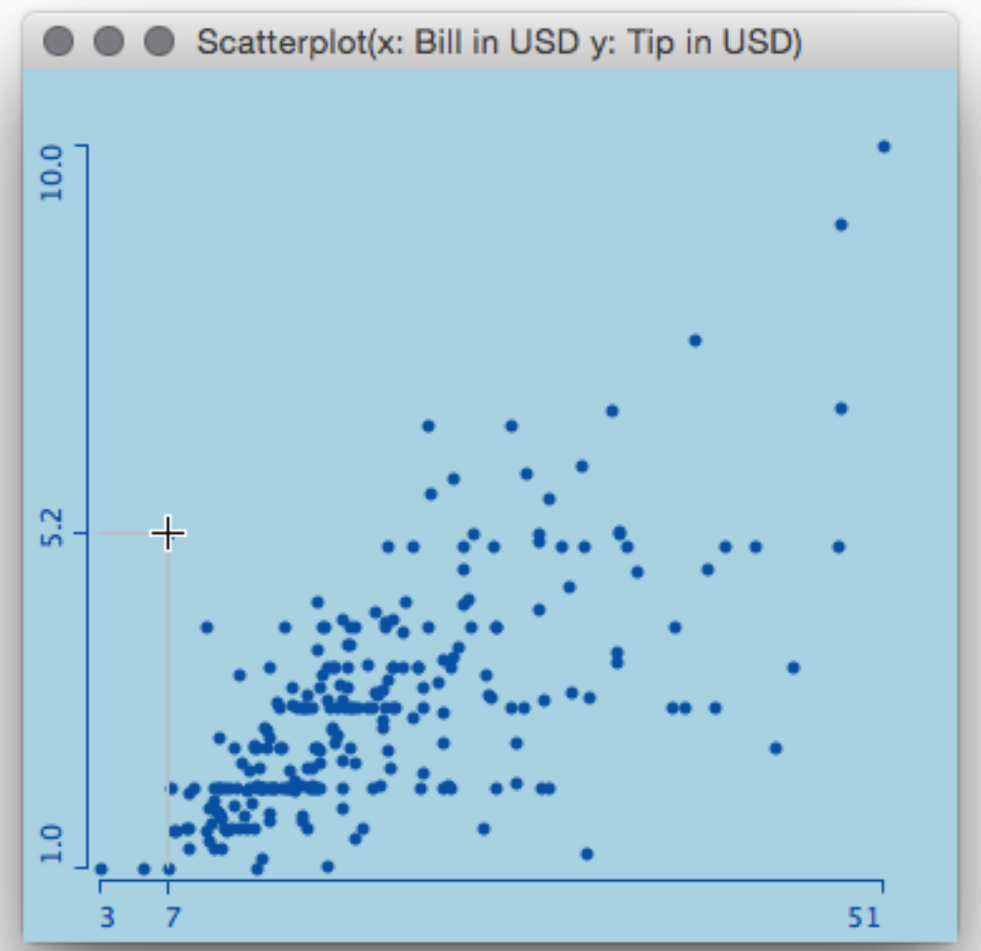
Queries

- The level of detail of a query should have optional granularities:
  - **orientation**, “what are the coordinates at the mouse pointer” (interactive grid)
  - **standard**, “what are the coordinates of a particular value”
  - **extended**, “what are the values for an object beyond the variables in the plot”

Queries

- The level of detail of a query should have optional granularities:
  - **orientation**, “what are the coordinates at the mouse pointer” (interactive grid)
  - **standard**, “what are the coordinates of a particular value”
  - **extended**, “what are the values for an object beyond the variables in the plot”

- Example: scatterplot

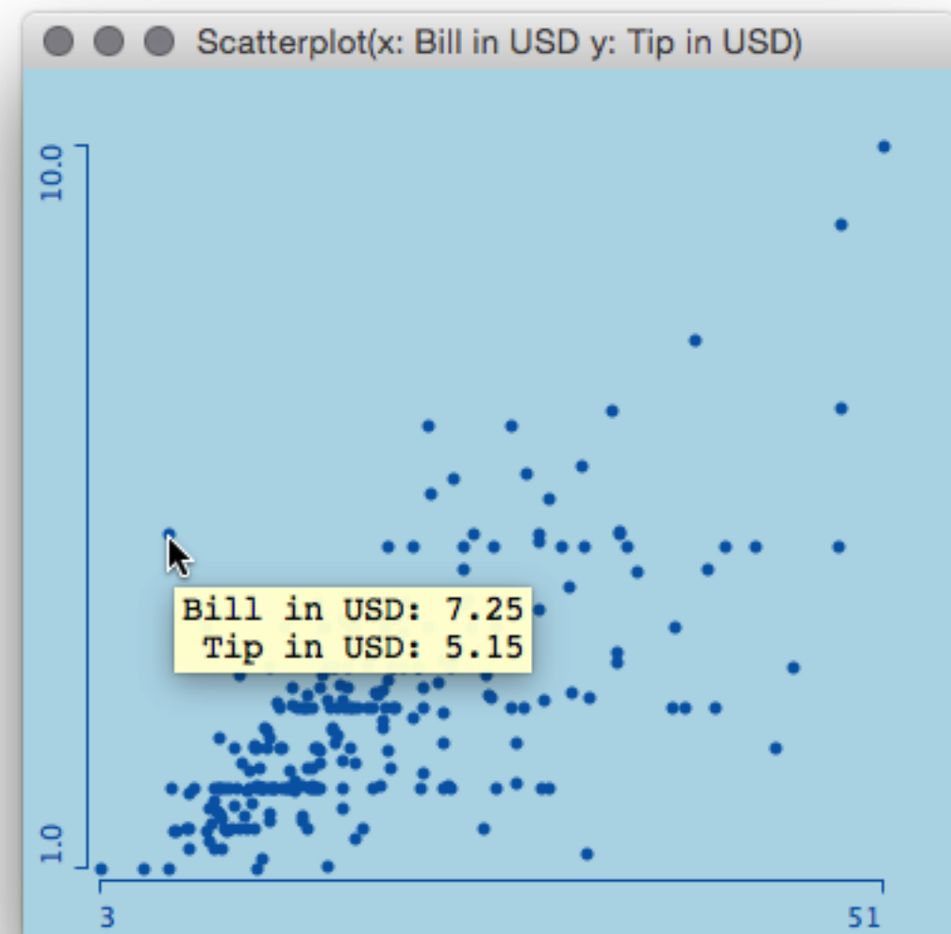


orientation

Queries

- The level of detail of a query should have optional granularities:
  - **orientation**, “what are the coordinates at the mouse pointer” (interactive grid)
  - **standard**, “what are the coordinates of a particular value”
  - **extended**, “what are the values for an object beyond the variables in the plot”

- Example: scatterplot

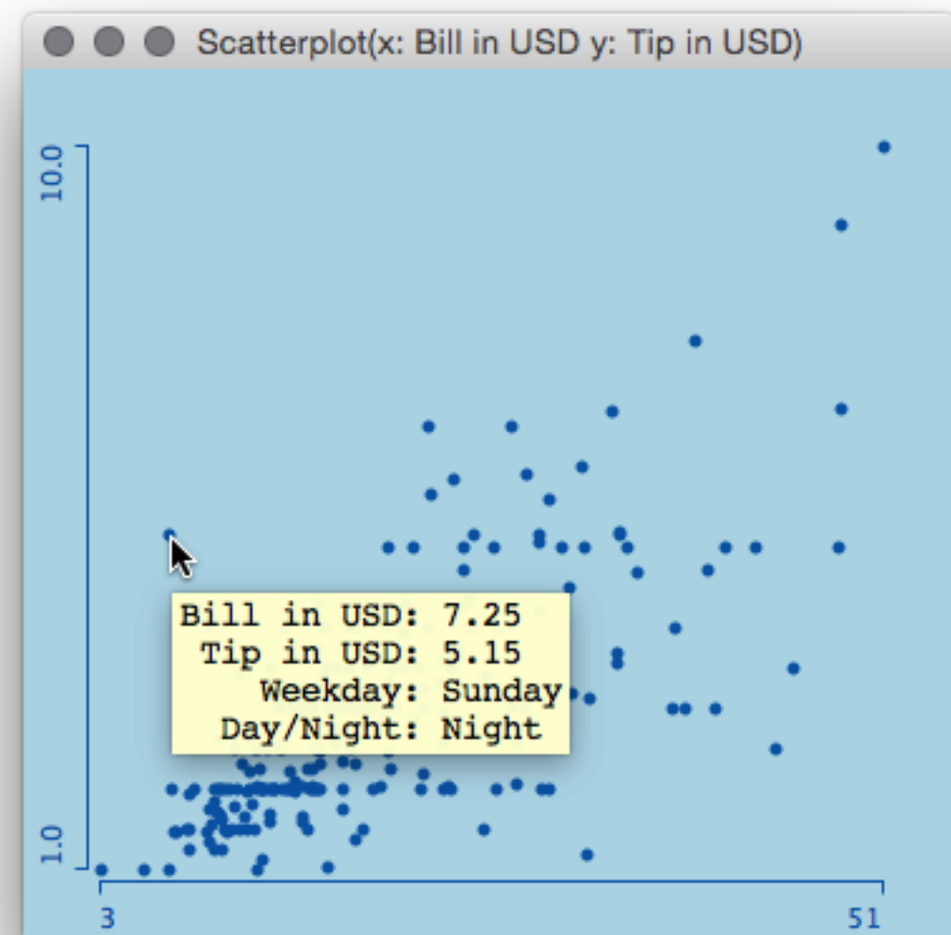


standard

Queries

- The level of detail of a query should have optional granularities:
  - **orientation**, “what are the coordinates at the mouse pointer” (interactive grid)
  - **standard**, “what are the coordinates of a particular value”
  - **extended**, “what are the values for an object beyond the variables in the plot”

- Example: scatterplot



extended

Changing Parameters

- Looking at the graphics functions in classical statistic systems, we find a large number of potential options to set
- Most of these options only apply to the “artistic quality” of the plots, i.e., fonts, colors, patterns, etc.
- For an exploratory analysis, we need to modify plot parameters, which relate to the statistical aspects of the graph
- Example: **Histogram**

  Two parameters:
  - anchor point
  - bin width / no. of bins

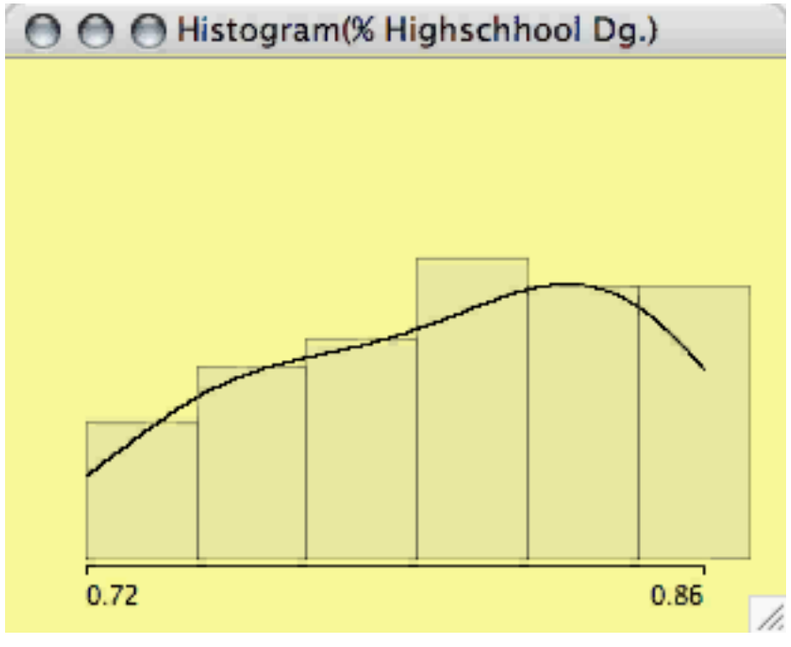
  Changes via:
  - Keyboard
  - numerical presets
  - numerical entry

Changing Parameters

- Looking at the graphics functions in classical statistic systems, we find a large number of potential options to set
- Most of these options only apply to the “artistic quality” of the plots, i.e., fonts, colors, patterns, etc.
- For an exploratory analysis, we need to modify plot parameters, which relate to the statistical aspects of the graph
- Example: **Histogram**

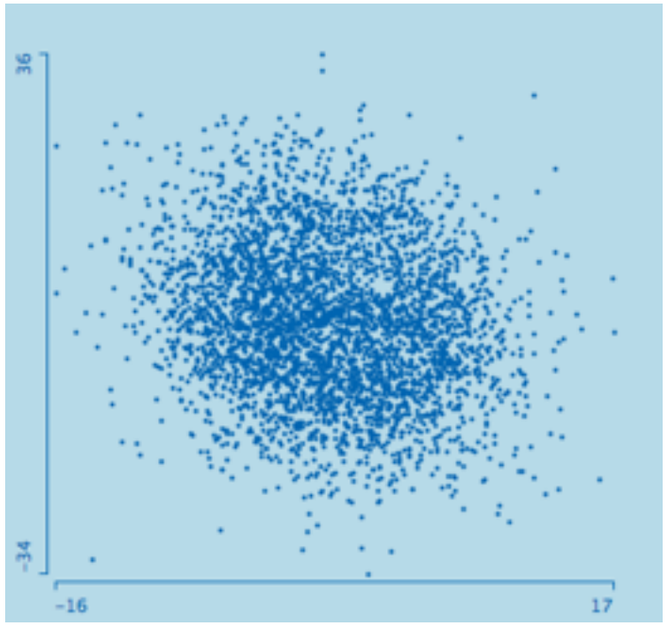
  Two parameters:
  - anchor point
  - bin width / no. of bins

  Changes via:
  - Keyboard
  - numerical presets
  - numerical entry



Changing Parameters

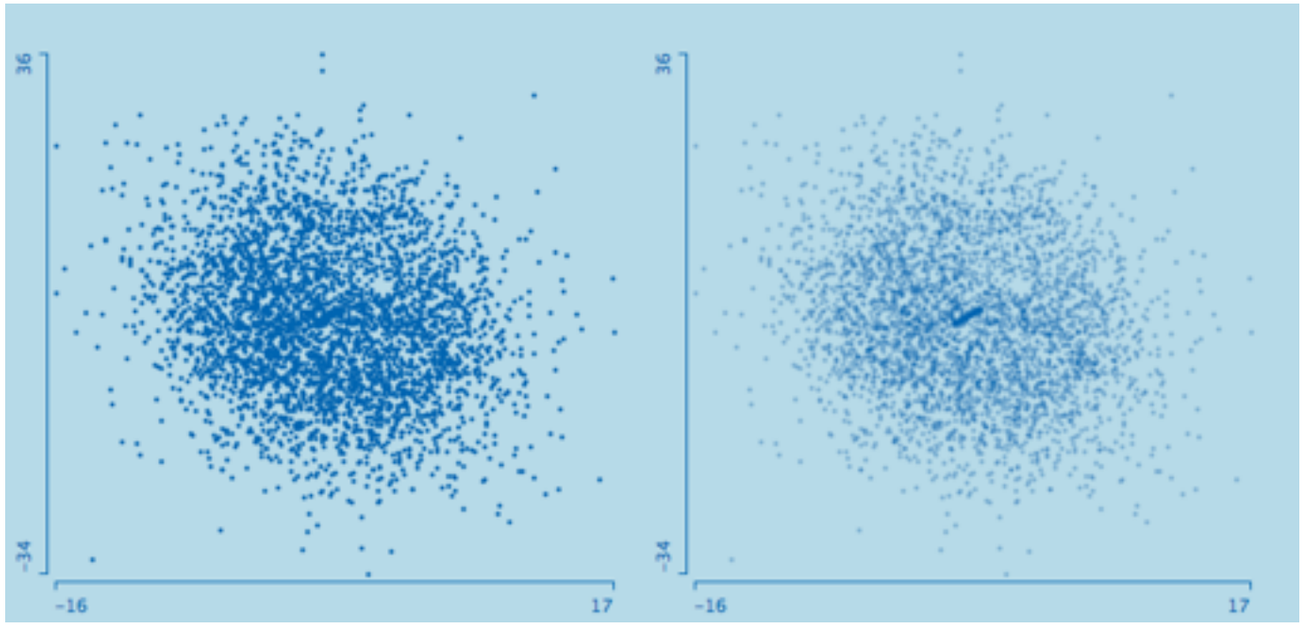
- Example: α-blending



- Example: Zooming

Changing Parameters

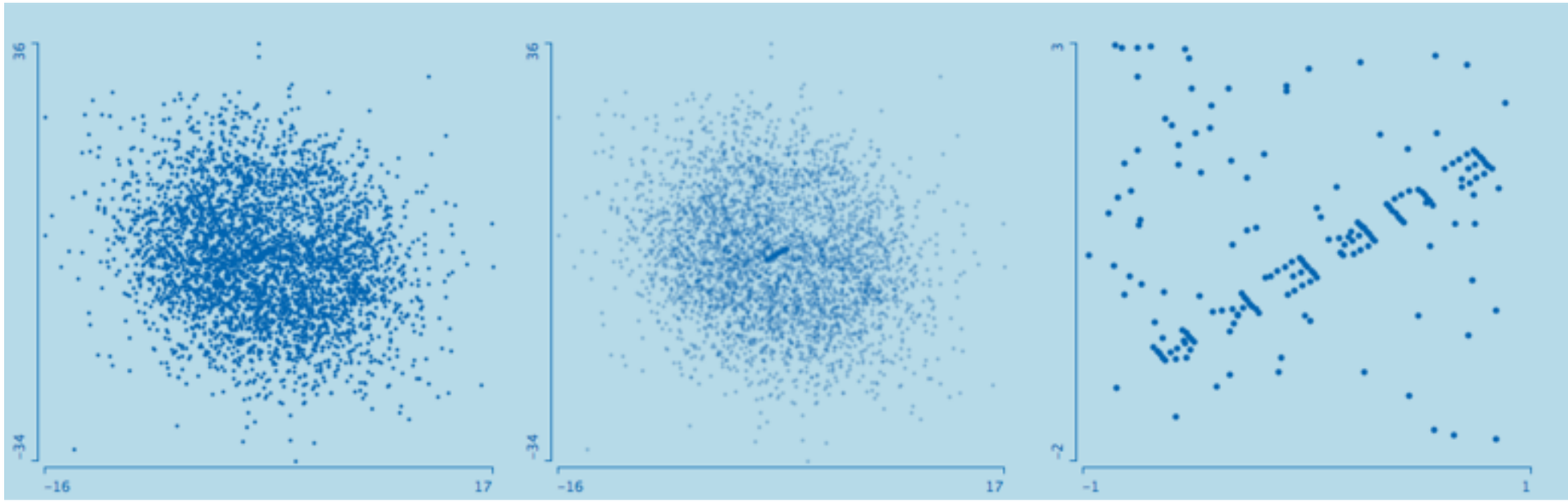
- Example: α-blending



- Example: Zooming

Changing Parameters

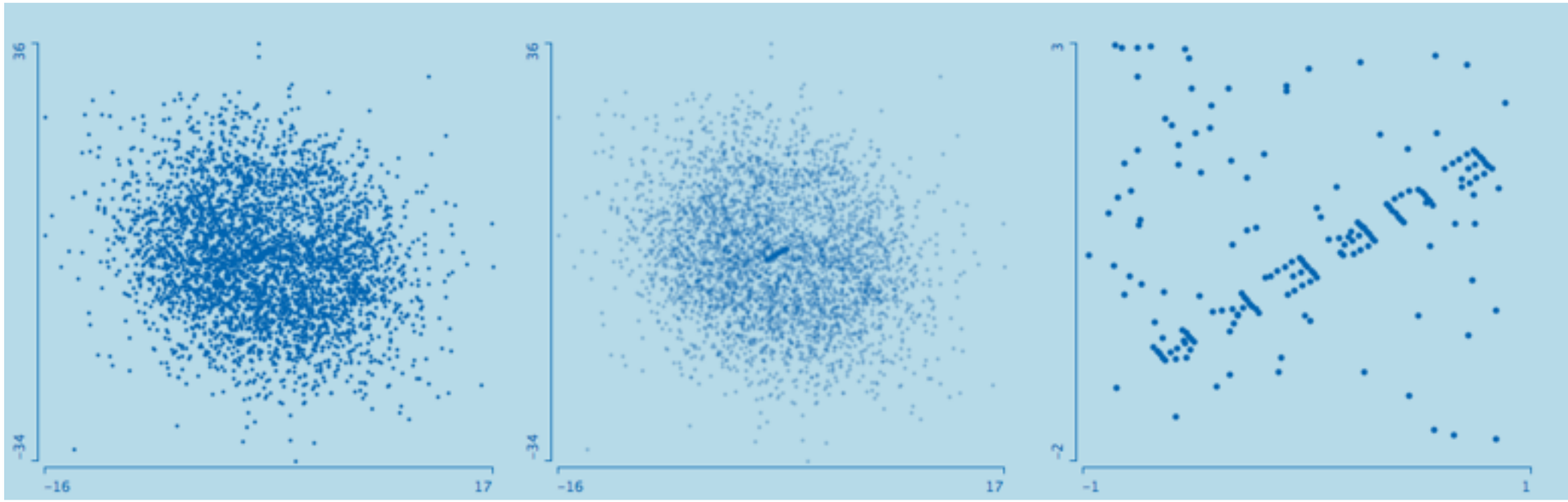
- Example: α-blending



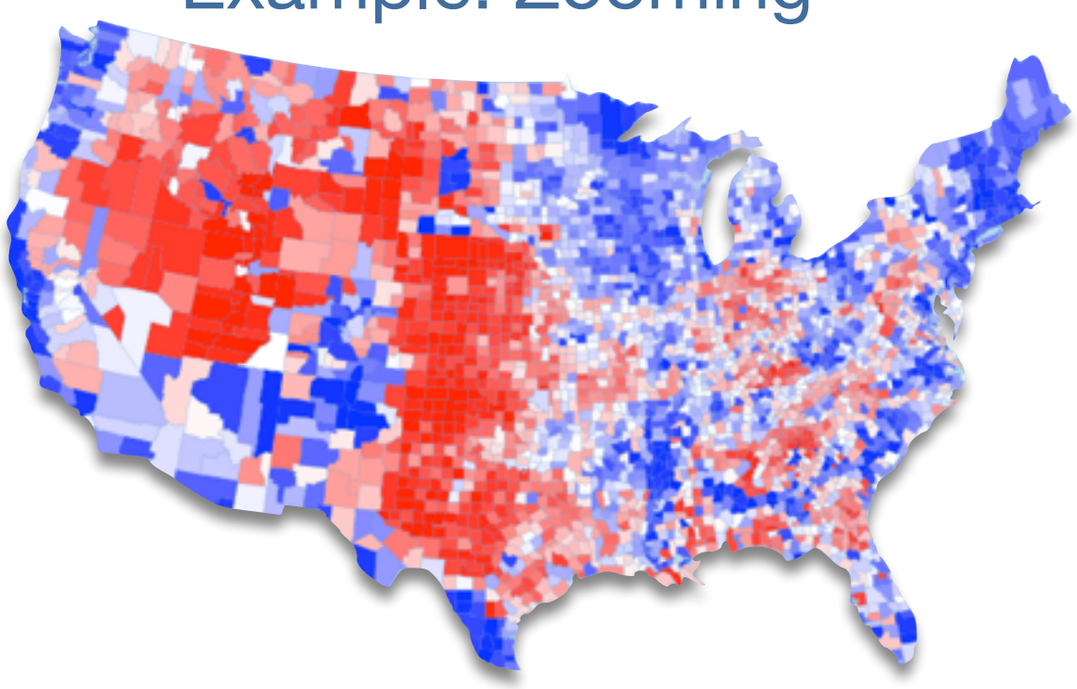
- Example: Zooming

Changing Parameters

- Example: α-blending

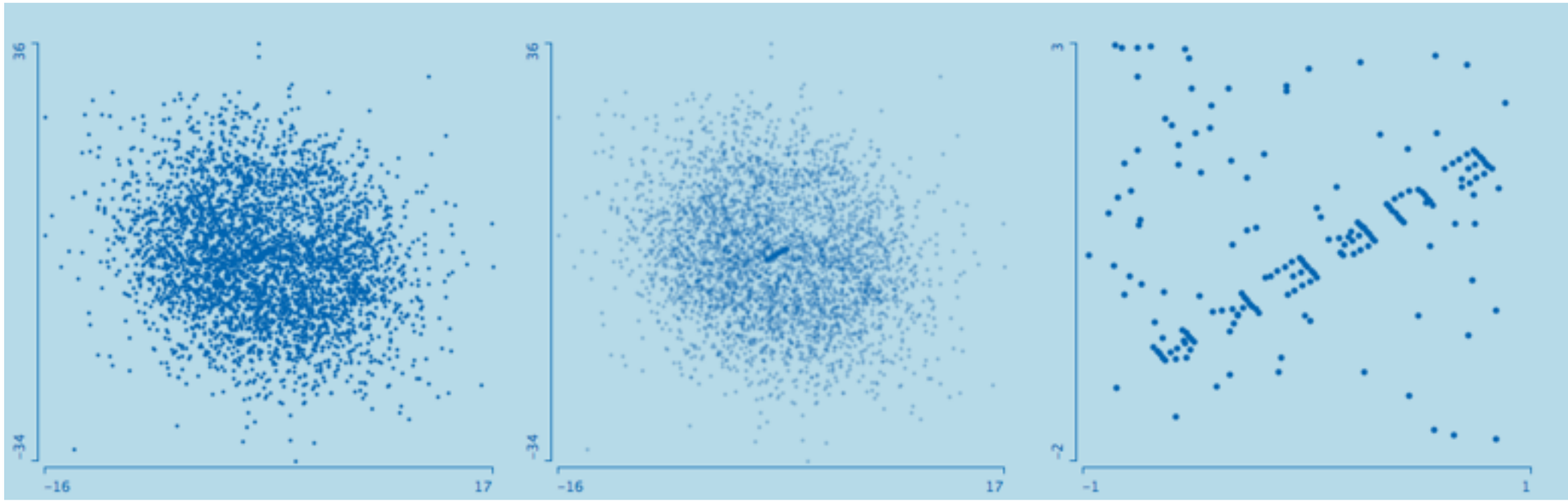


- Example: Zooming

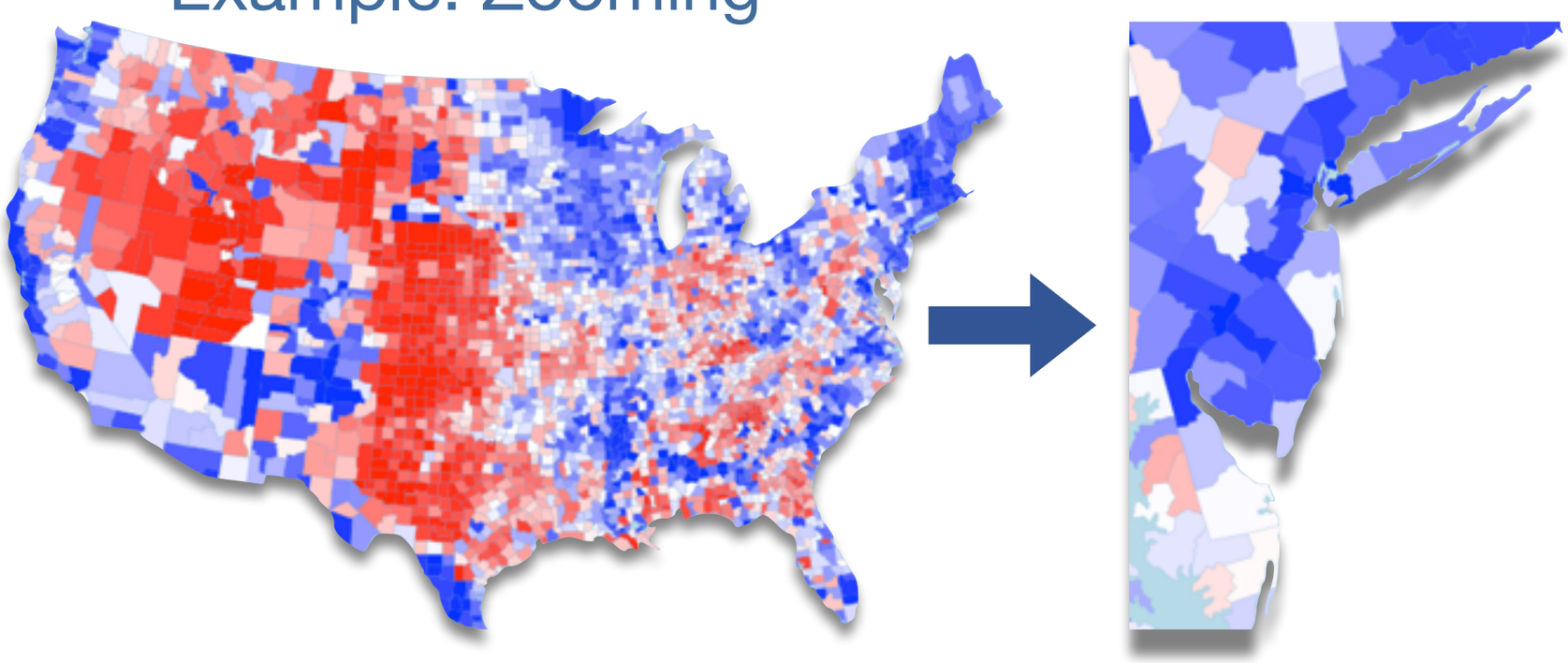


Changing Parameters

- Example: α-blending

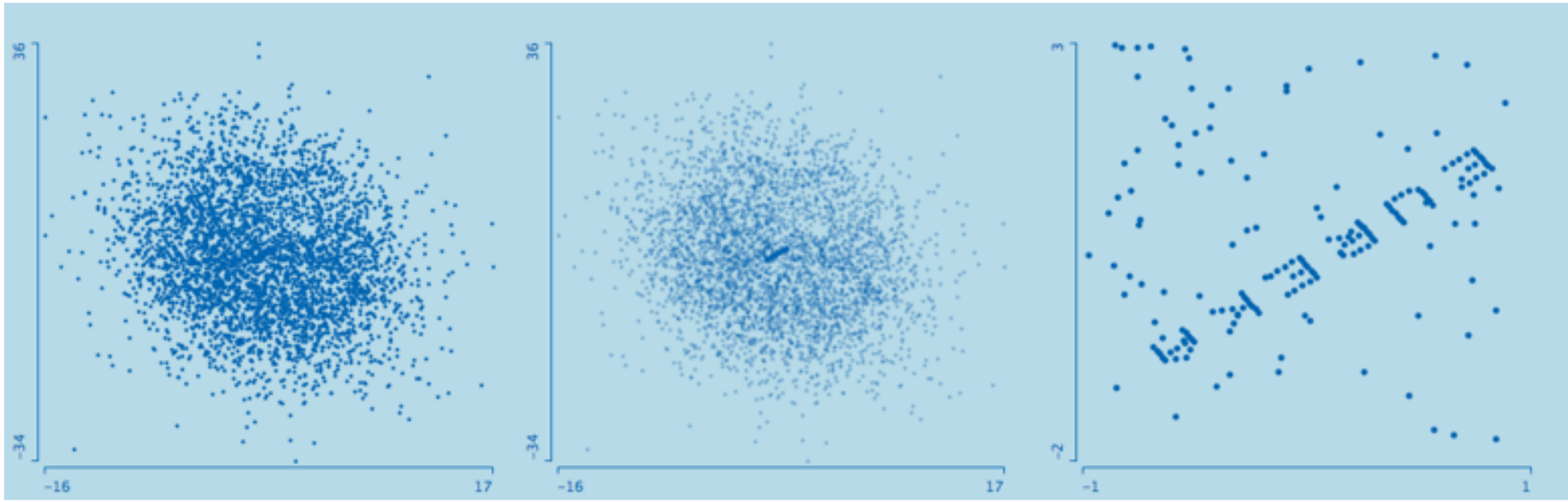


- Example: Zooming

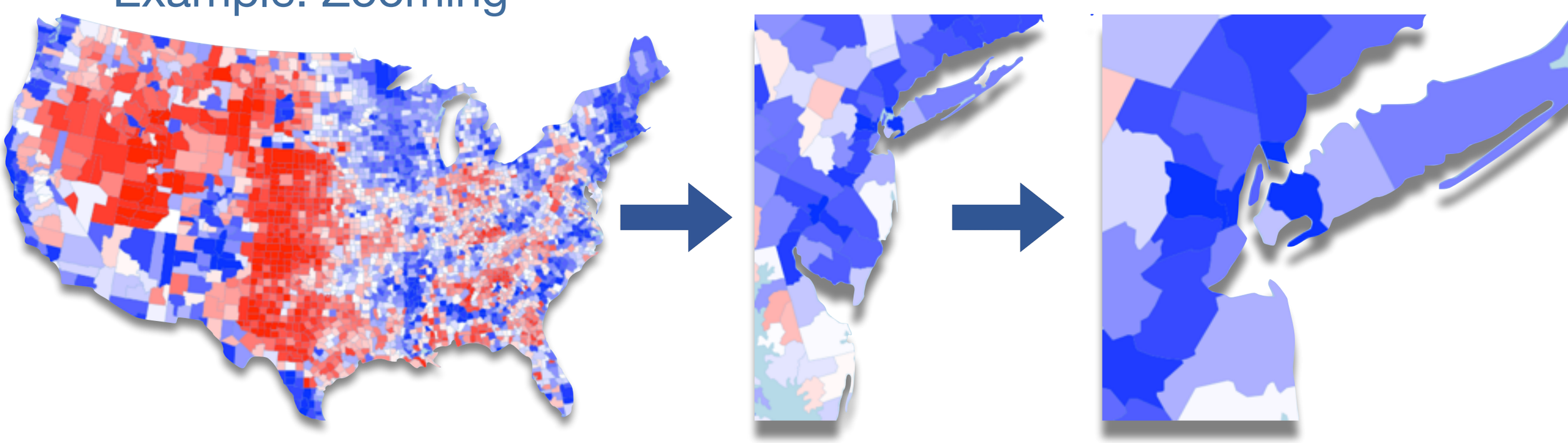


Changing Parameters

- Example: α-blending

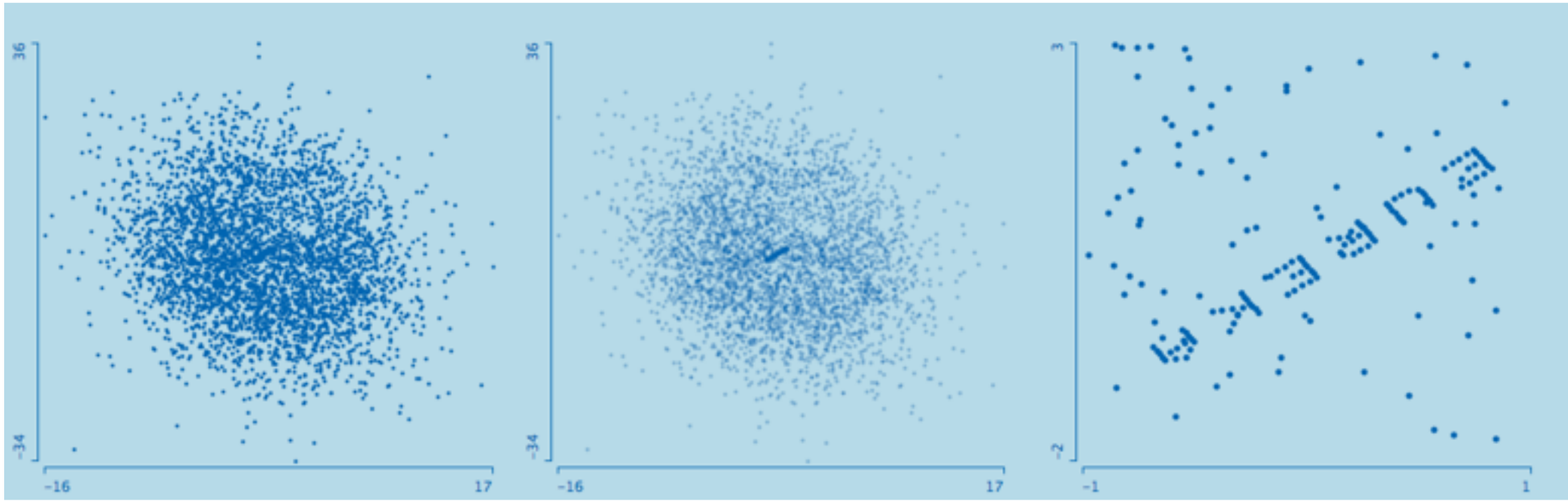


- Example: Zooming

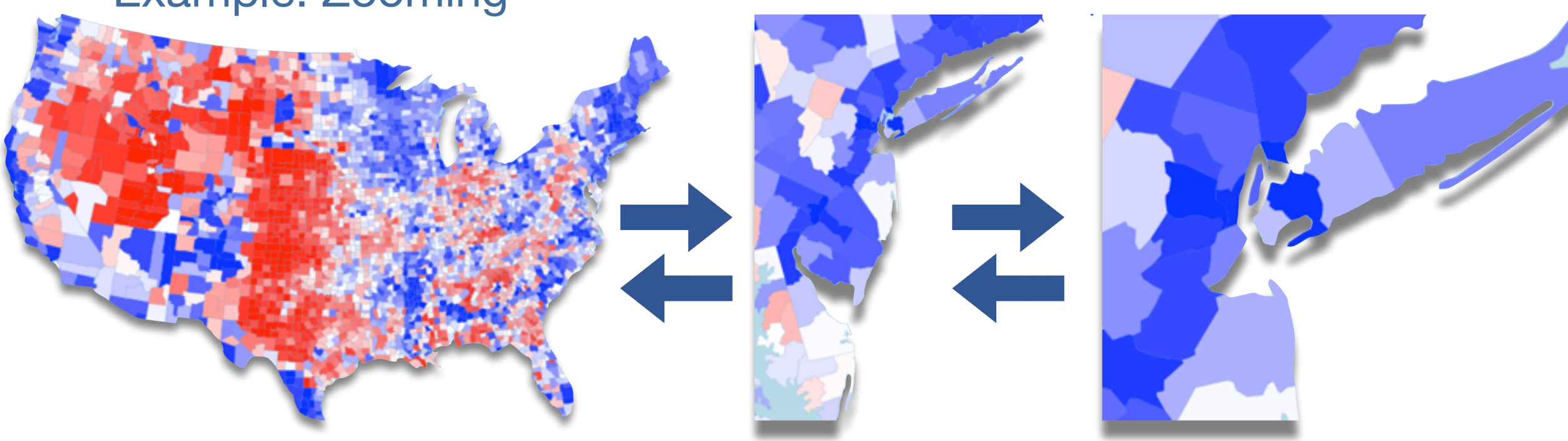


Changing Parameters

- Example: α-blending



- Example: Zooming



Changing Parameters

- The two inherently multivariate plots are
  - parallel coordinate plots (for continuous data)
  - mosaic plots (for categorical data)

Changing Parameters

- The two inherently multivariate plots are
  - parallel coordinate plots (for continuous data)
  - mosaic plots (for categorical data)
- Both of these plots are not very powerful for exploratory work as long as they are not implemented interactively.

Changing Parameters

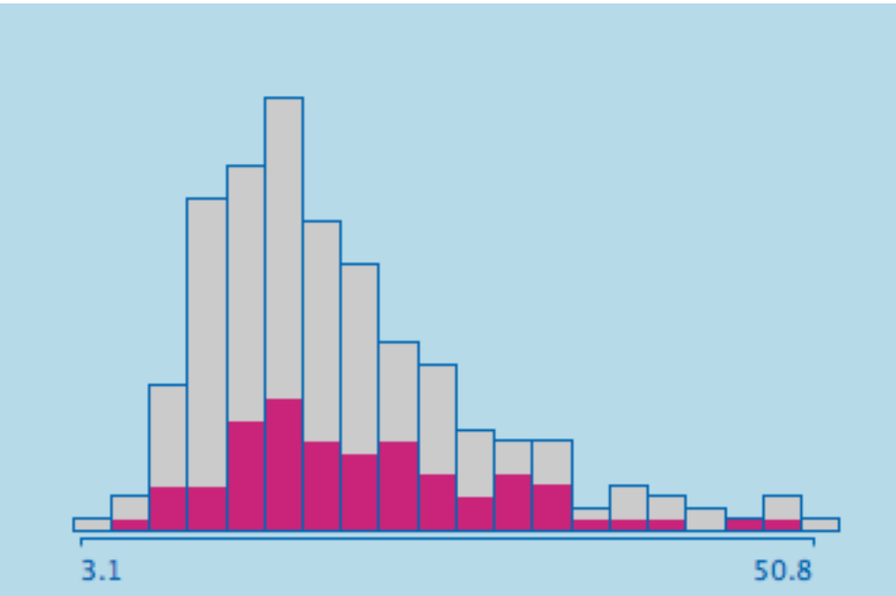
- The two inherently multivariate plots are
  - parallel coordinate plots (for continuous data)
  - mosaic plots (for categorical data)
- Both of these plots are not very powerful for exploratory work as long as they are not implemented interactively.
- Essential (but not exhaustive) interactive features are
  - Parallel coordinates
    - rearrangement of axes (manual, automatic permutations)
    - scaling of axes (common, individual, inversion)
    - alignment of axes (mean, median, constants)
    - sorting (min, max, mean, median, range, std.dev.)
  - Mosaic plots
    - include and exclude variables
    - permute variable order
    - (censored) zooming

Changing Parameters

- The two inherently multivariate plots are
  - parallel coordinate plots (for continuous data)
  - mosaic plots (for categorical data)
- Both of these plots are not very powerful for exploratory work as long as they are not implemented interactively.
- Essential (but not exhaustive) interactive features are
  - Parallel coordinates
    - rearrangement of axes (manual, automatic permutations)
    - scaling of axes (common, individual, inversion)
    - alignment of axes (mean, median, constants)
    - sorting (min, max, mean, median, range, std.dev.)
  - Mosaic plots
    - include and exclude variables
    - permute variable order
    - (censored) zooming
- Linking with these plots increases dimensionality even more

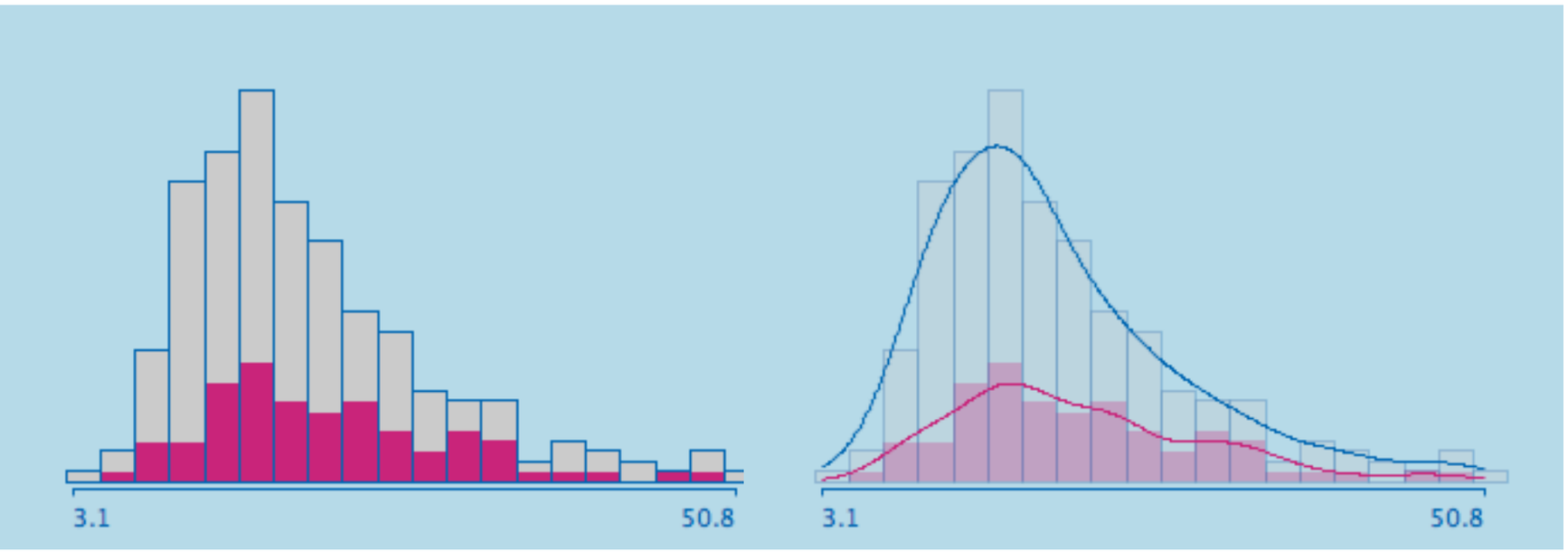
Statistification of Graphical Displays

- Example: Density Estimation



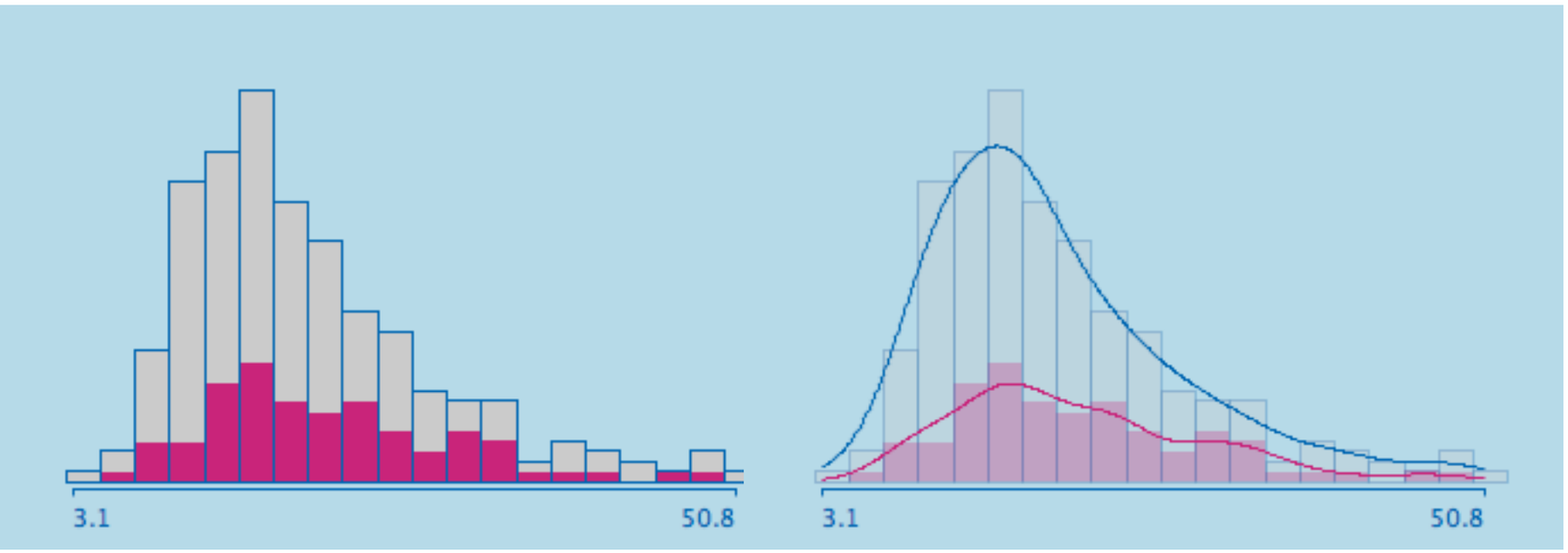
Statistification of Graphical Displays

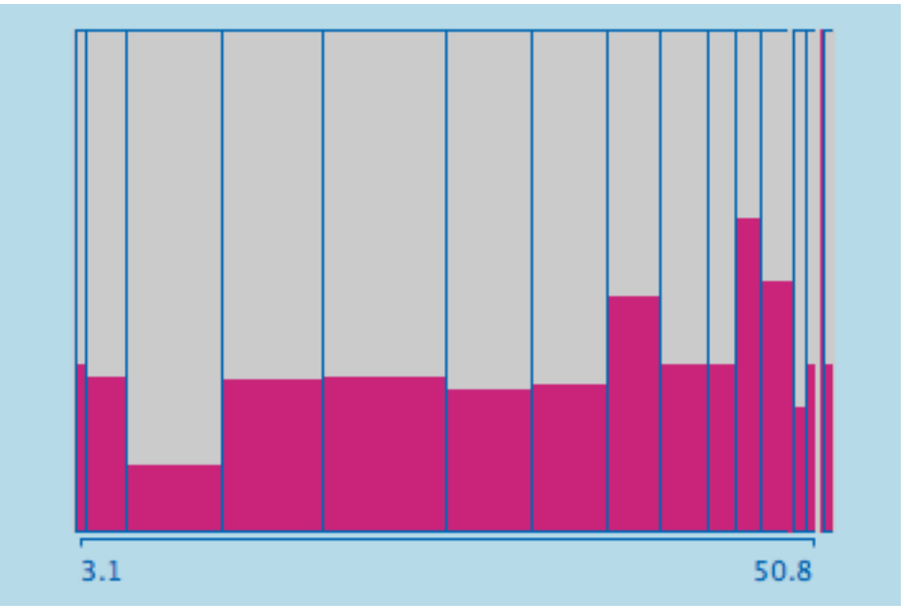
- Example: Density Estimation



Statistification of Graphical Displays

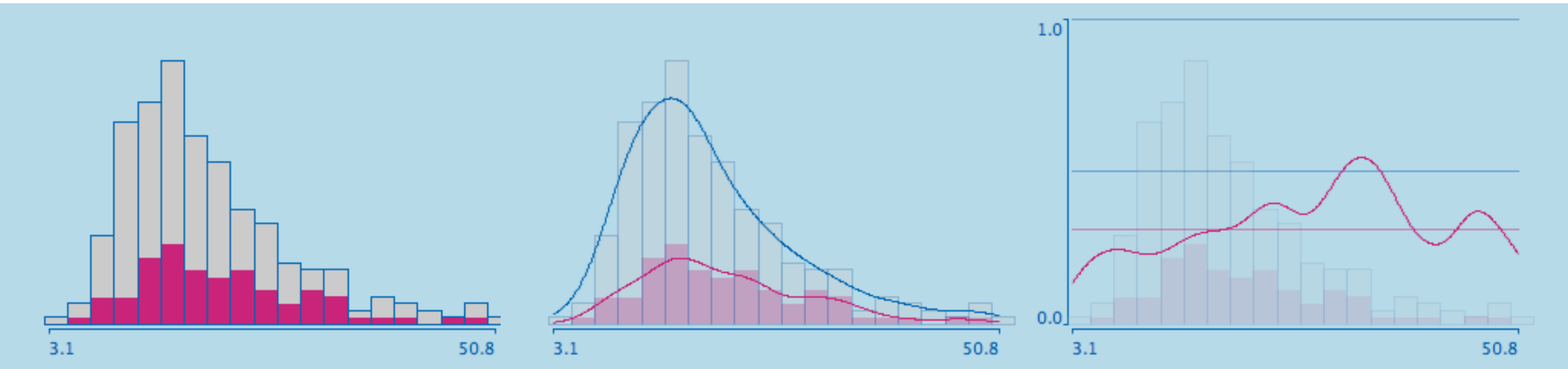
- Example: Density Estimation

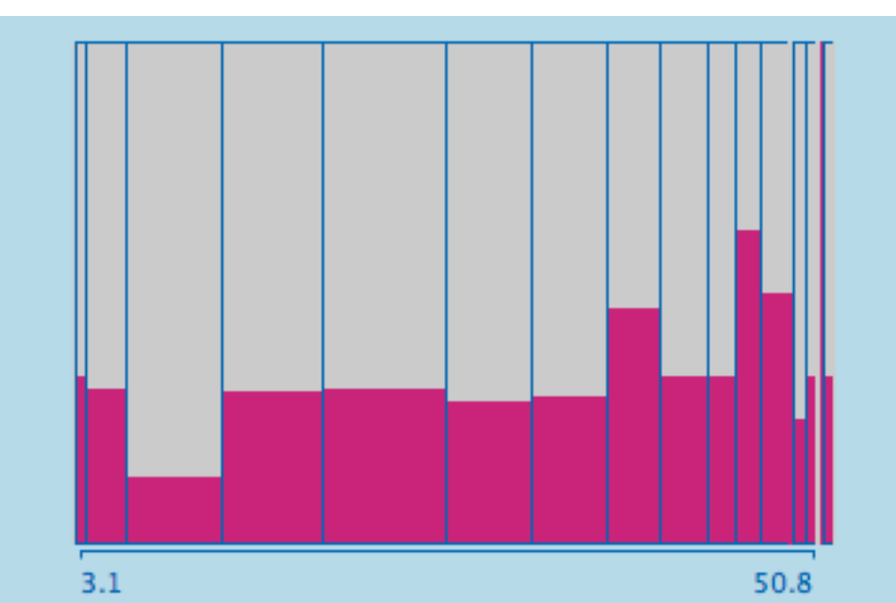




Statistification of Graphical Displays

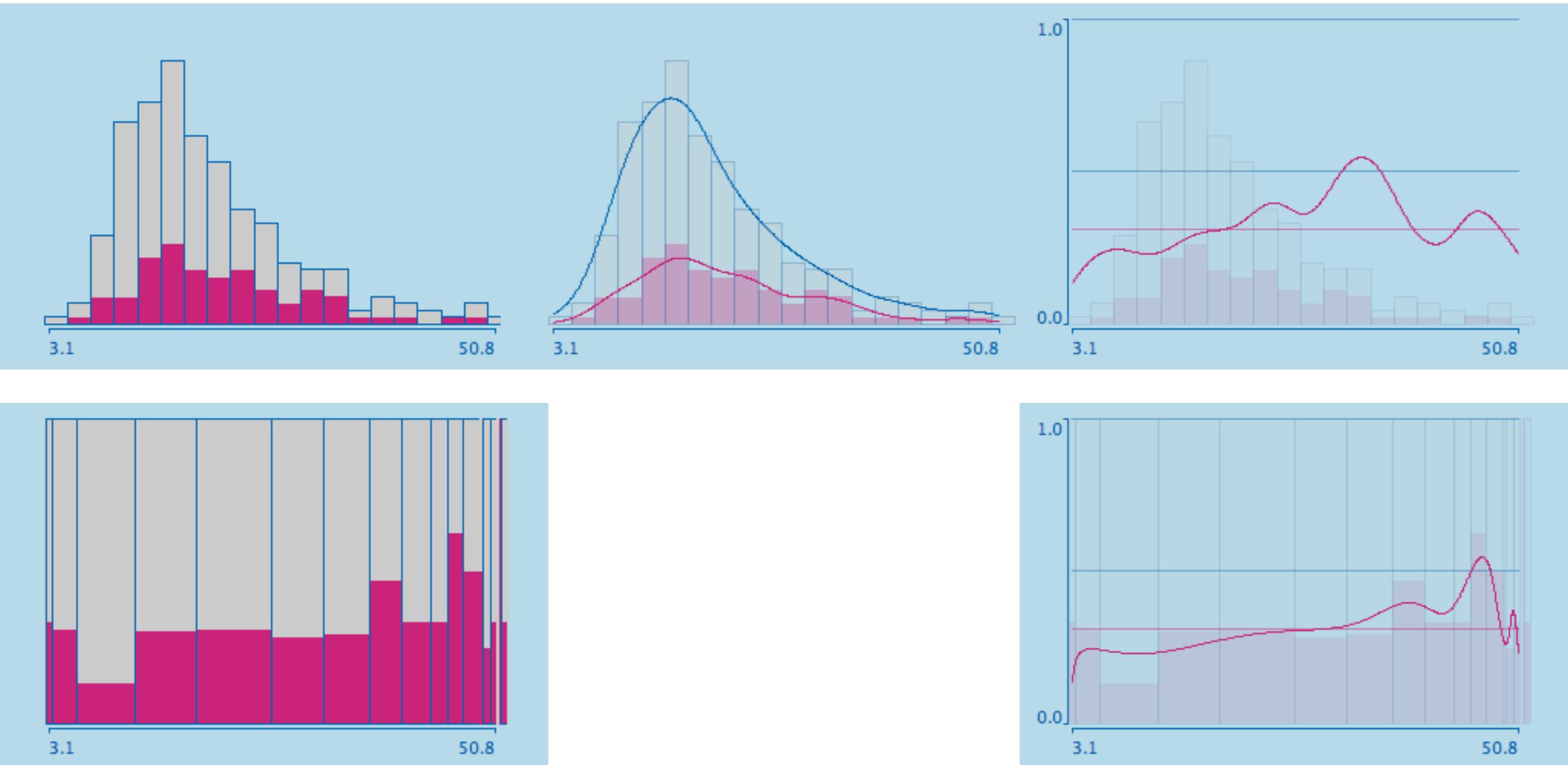
- Example: Density Estimation





Statistification of Graphical Displays

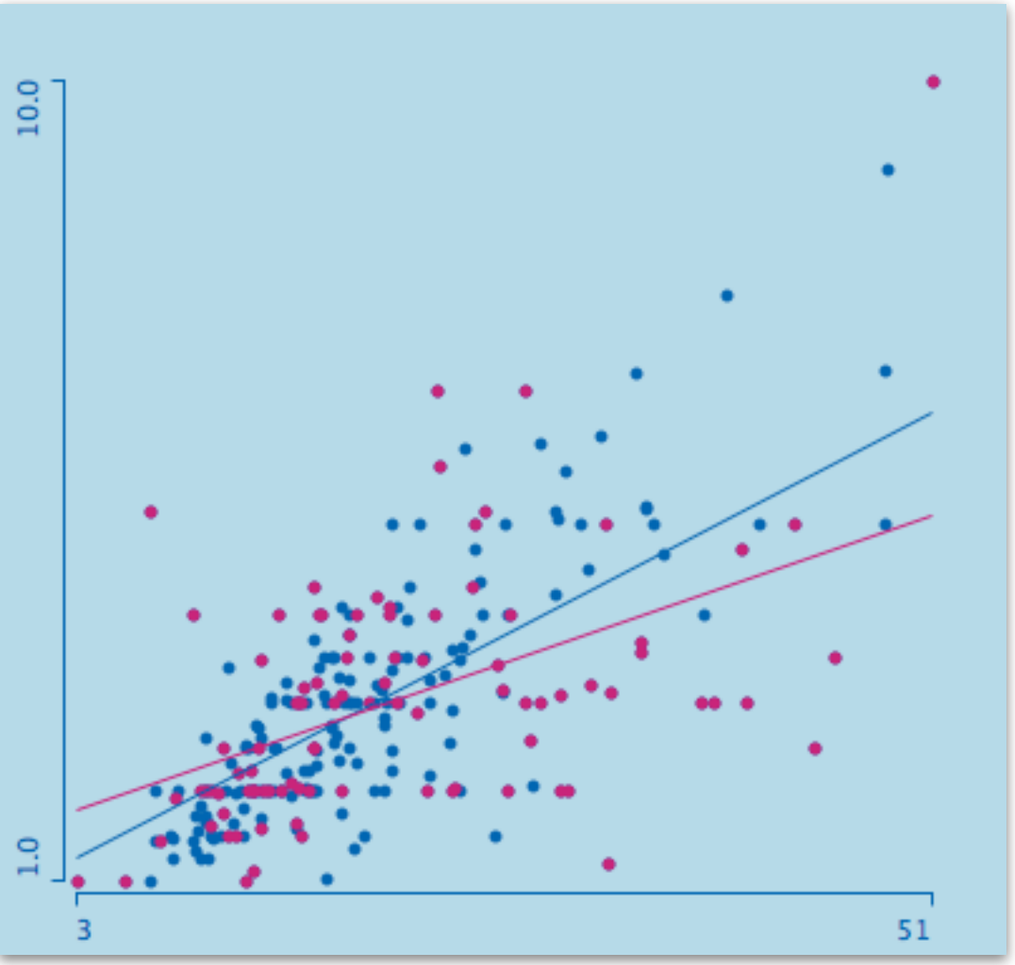
- Example: Density Estimation



Statistification of Graphical Displays

- Example: Scatterplot Smoothers

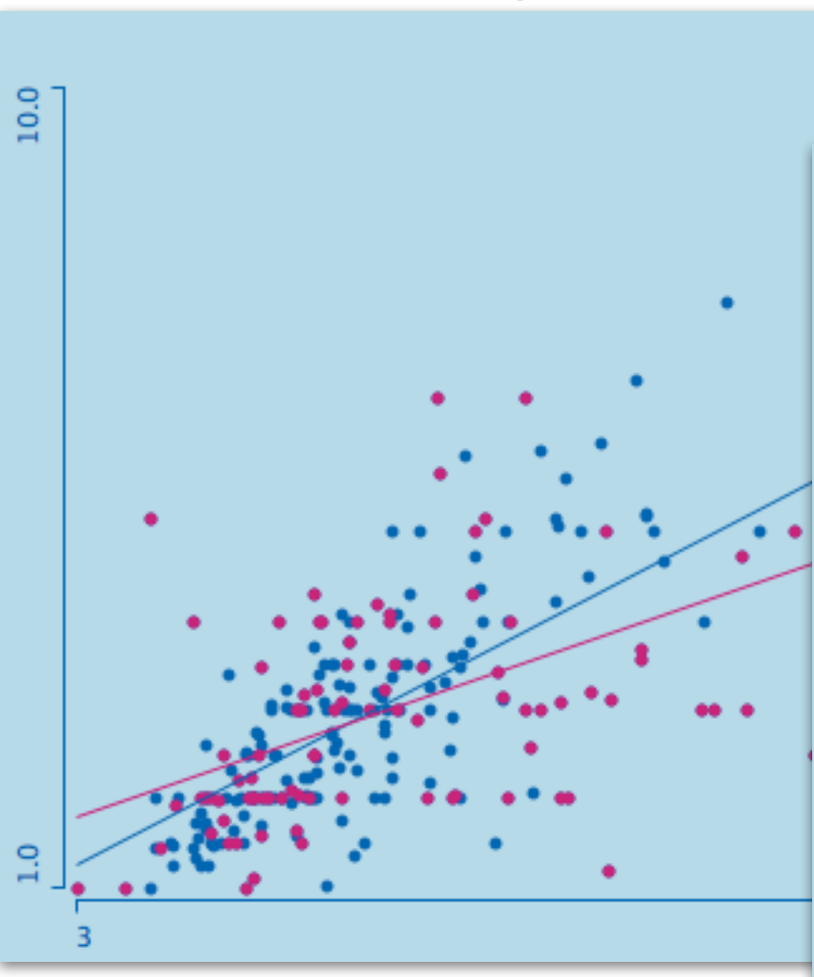
linear regression



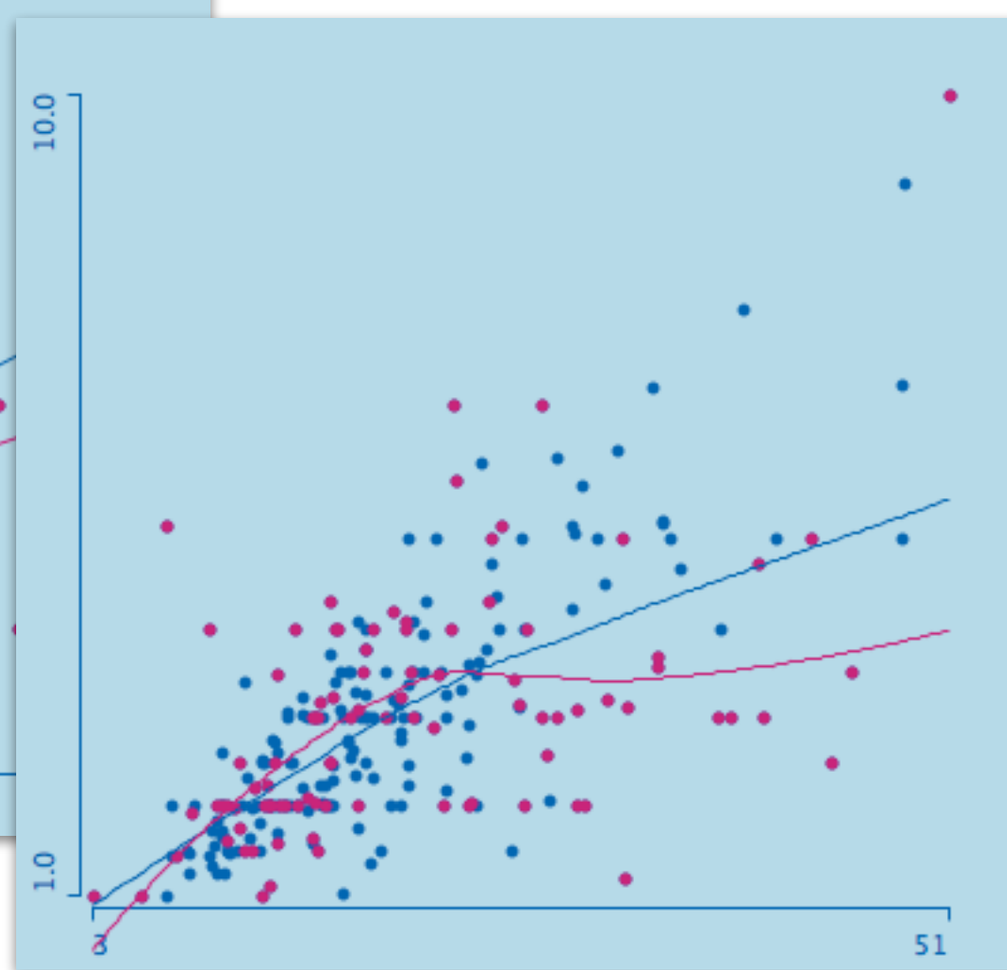
Statistification of Graphical Displays

- Example: Scatterplot Smoothers

linear regression



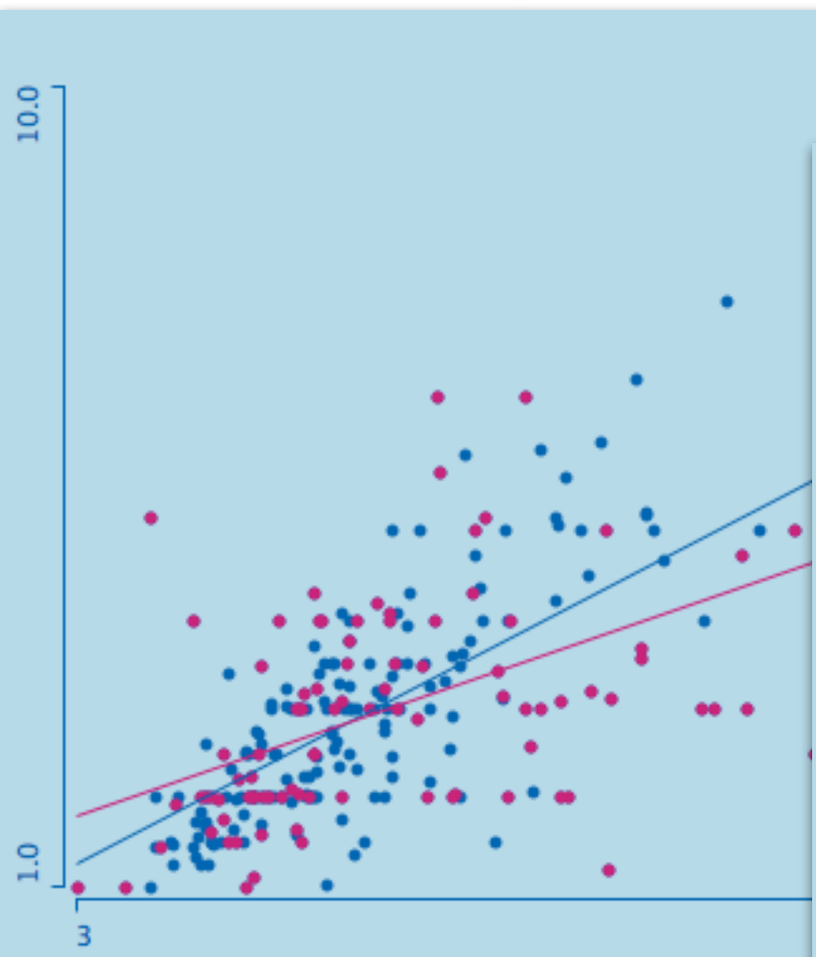
loess



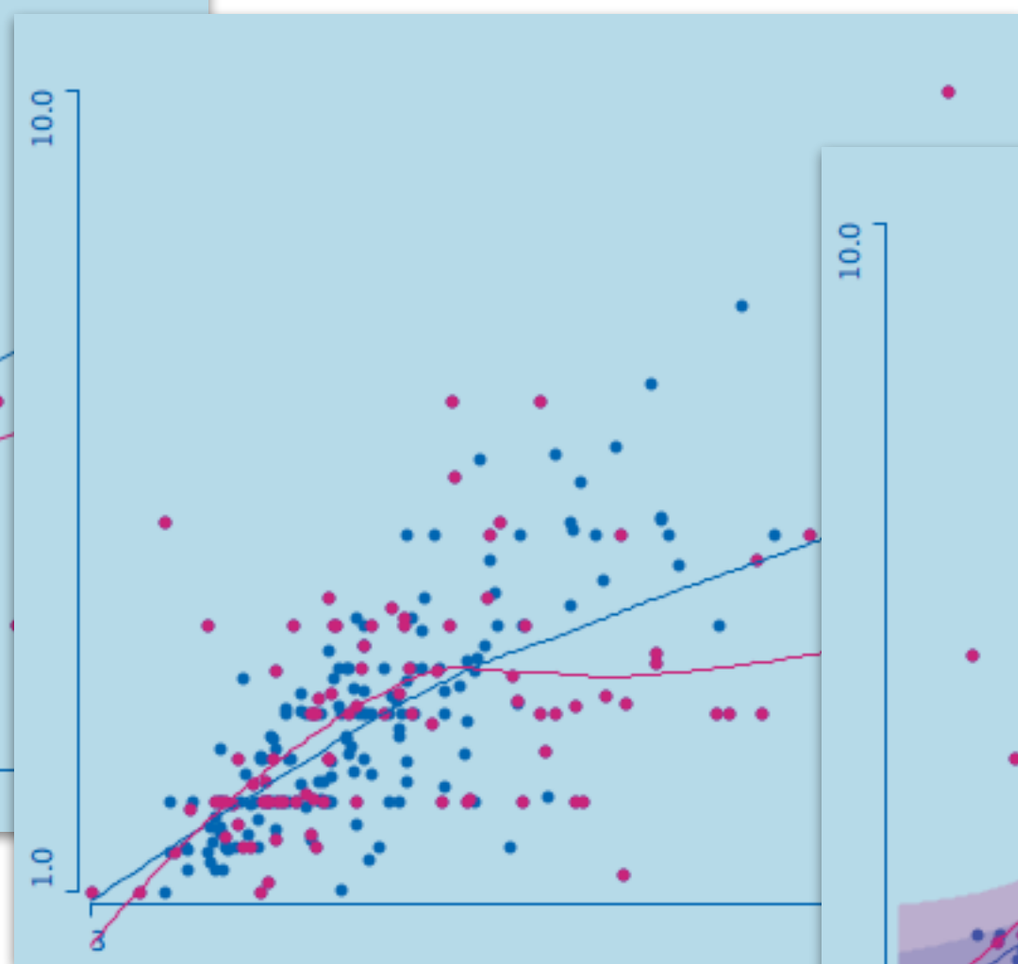
Statistification of Graphical Displays

- Example: Scatterplot Smoothers

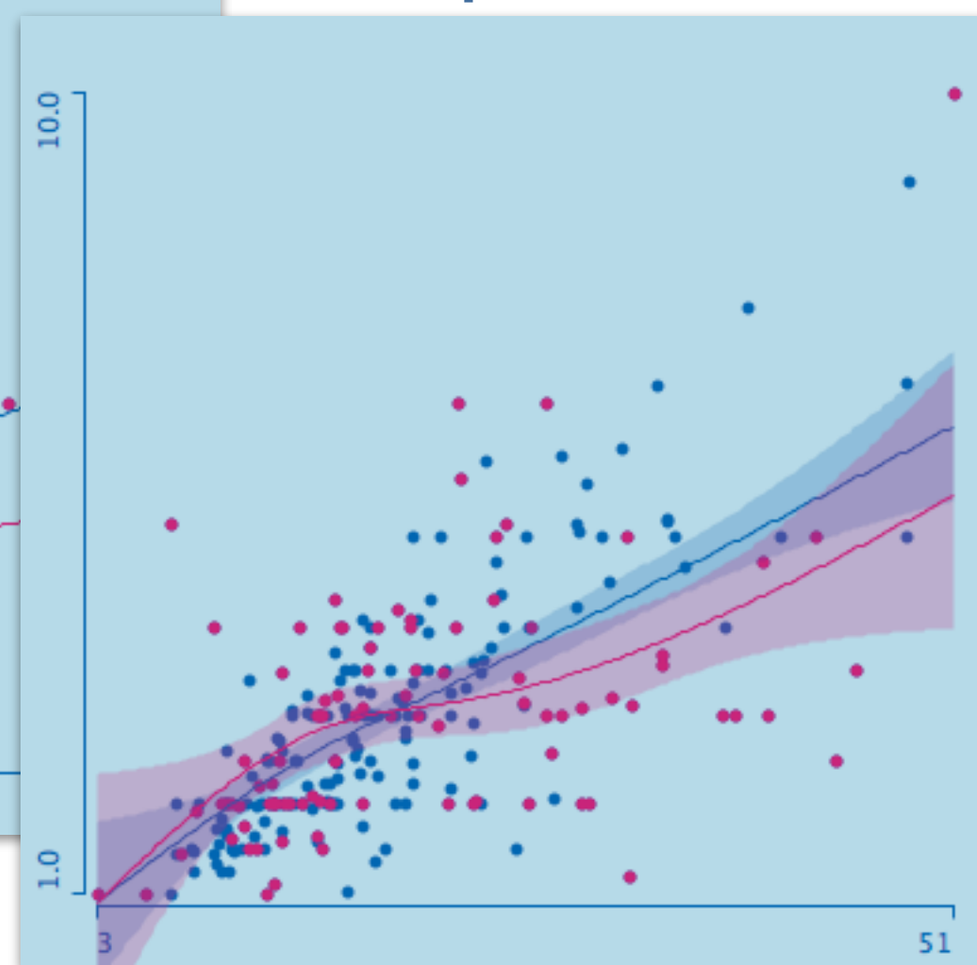
linear regression



loess



splines

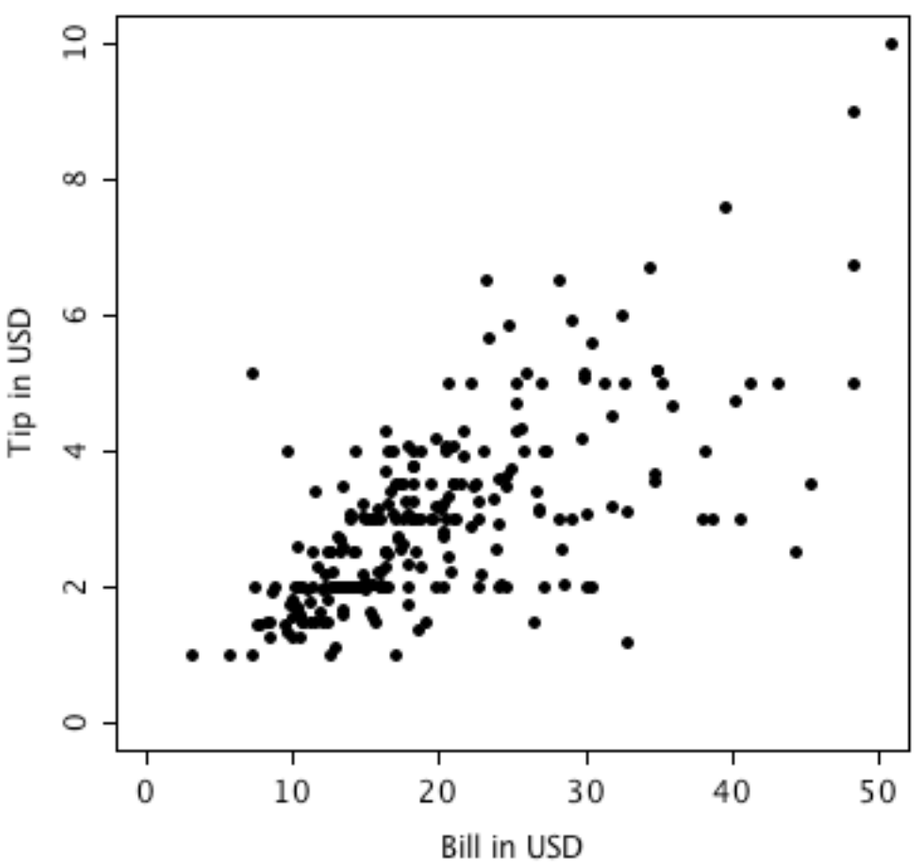


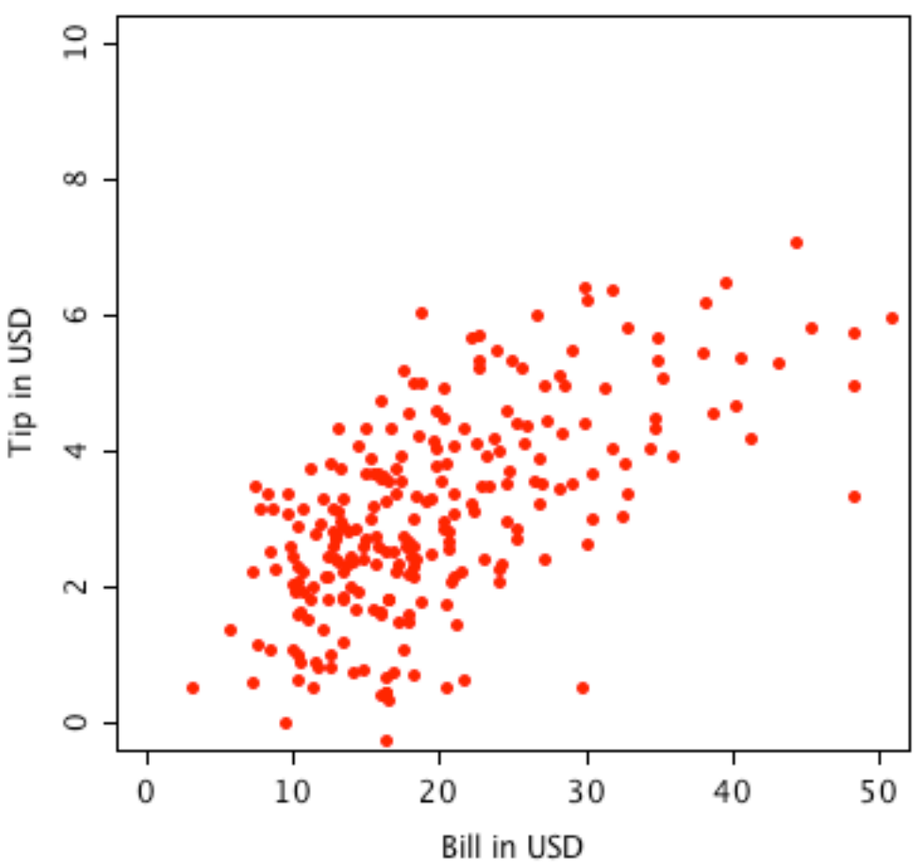
Statistification: Graphical Inference

- Basic Idea:
  "Look the sampled data of the model like my raw data?"
- Once we "know" how our raw data "looks like", we can compare it to the data we sample from a chosen model (many times …)
- Example: simple linear model for **`Tip ~ Billsize`**

Statistification: Graphical Inference

- Basic Idea:
  "Look the sampled data of the model like my raw data?"
- Once we "know" how our raw data "looks like", we can compare it to the data we sample from a chosen model (many times …)
- Example: simple linear model for **`Tip ~ Billsize`**





Summary

- Given the right tools, graphics can efficiently be used to
  - clean data
  - explore data, and
  - diagnose models
- The most important tools and techniques are
  - selection with linked highlighting
  - rapid change of parameters
  - incorporation of statistical estimates and models
- Tools matter
  - Tableau
  - SAS *JMP*
  - shiny / RStudio (“interact with static graphics”)
  - Mondrian

About Mondrian

- Mondrian is a general purpose **graphical data analysis tool**
- It is based on the experiences and tries to expand the concepts and ideas of
  - **DataDesk** (Paul F. Velleman, 1985)
  - **MANET** (Unwin et al., 1994)
- The basic building blocks of Mondrian are
  - **uni- and multivariate plots** for variables measured on various scales (including geographical maps)
  - **selection**, and
  - linked **highlighting**
  - fast **parameter** changes
  - **link to R** to add statistical procedures of various kinds
- Mondrian can be used **free of charge**, is **open source** and runs equally well on **Windows**, **MacOS** and **Linux** computers

About Mondrian

- Mondrian is a general purpose **graphical data analysis tool**
- It is based on the exp to expand the concepts and ideas of
  - **DataDesk** (Paul F. Vell
  - **MANET** (Unwin et a
- The basic building bl re
  - **uni- and multivariate p** red on various scales (including geographical
  - **selection**, and
  - linked **highlighting**
  - fast **parameter** changes
  - **link to R** to add statistic s kinds
- Mondrian can be used **free of charge**, is **open source** and runs equally well on **Windows**, **MacOS** and **Linux** computers