Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

email: soumyad@cse.iitk.ac.in

Study Materials for Lecture 10

- Reference papers and resoucres on slide footnotes
- Book: Visualization Analysis and Design by T. Munzner
  - Chapter 6: Rules of Thumb for Designing Visualizations
  - Chapter 12: Facet into Multiple Views
  - Chapter 14: Focus + Context
  - Chapter 13: Reduce Items and Attributes

Acknowledgements

- Some of the following slides are adapted from the excellent course materials and tutorials made available by:
  - Prof. Michelle Borkin (Northeastern University)

Data Types

Data Types

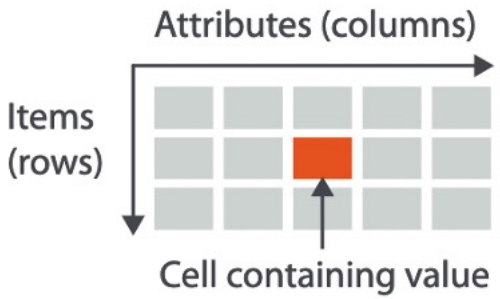
- Type = Structural or mathematical interpretation of the data

→ Items *(row, node)* → Attributes *(variable, data dimension)* → Links *(relationship)* → Positions *(spatial location)* → Grids *(sampling)*

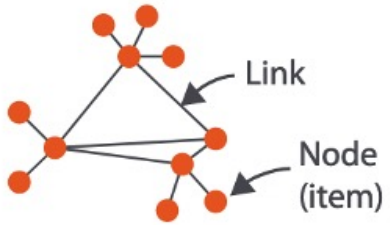
Data Types

- Dataset = collection of information/data

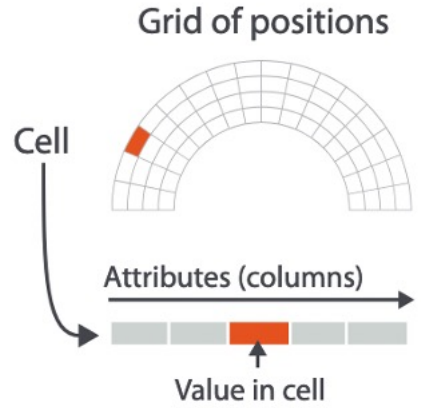
➔ Tables



➔ Networks

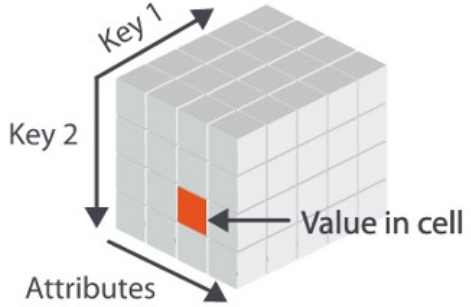


➔ Fields (Continuous)

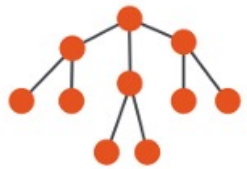


➔ Geometry (Spatial)

➔ *Multidimensional Table*



➔ *Trees*



Attribute Types

- Attribute Types:

→ Categorical *(nominal)*

e.g., fruit (apple, pear, etc.)
colleges (CCIS, CAMD, etc.)

→ Ordered

→ *Ordinal* *(ordered)*

e.g., months (J, F, M, etc.)
sizes (xs, s, m, l, xl)

→ *Quantitative* *(continuous)*

e.g., lengths (1", 2.5", 5")
population

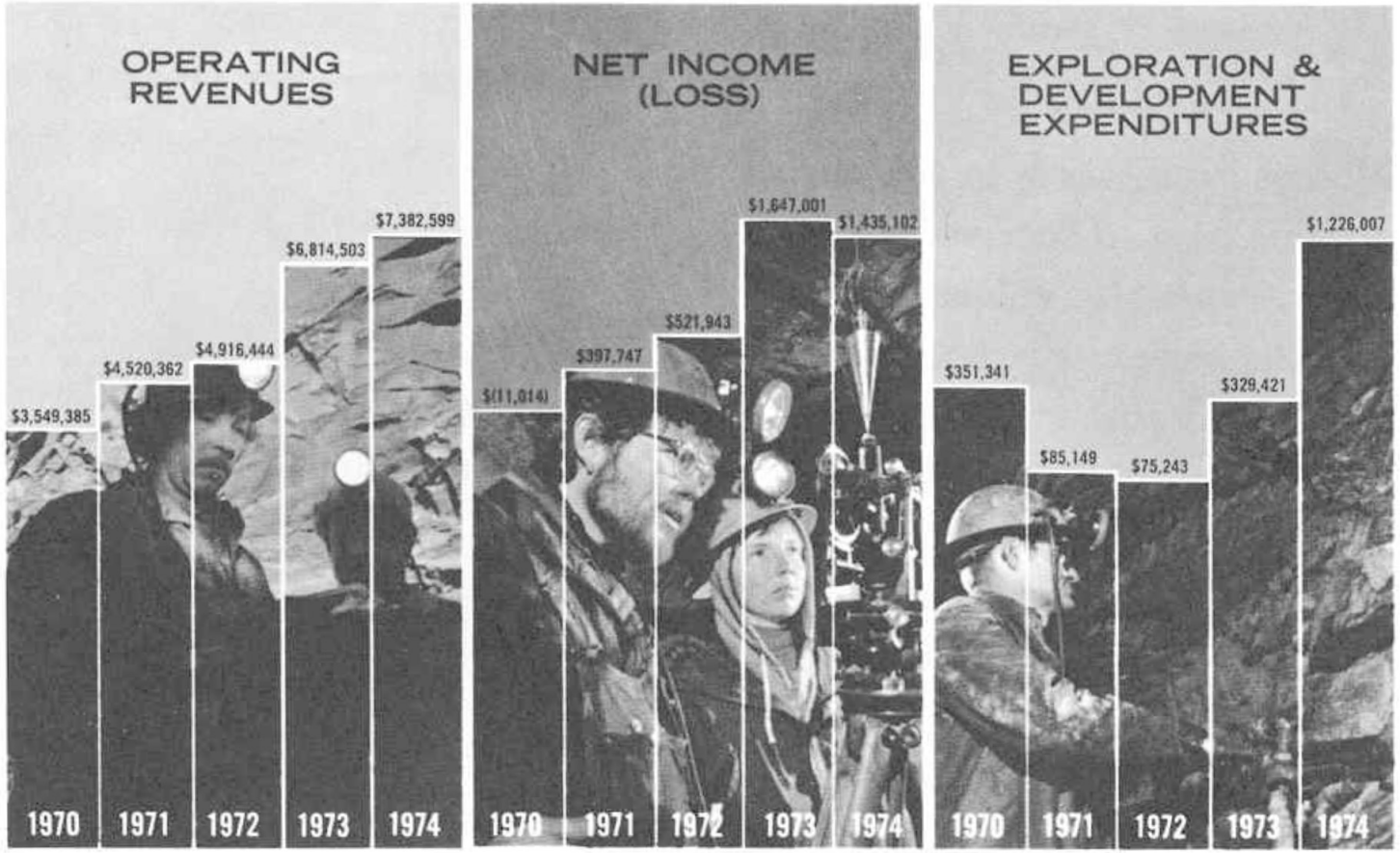
Rules of Thumb for Designing Visualizations

Graphical Integrity

"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data"

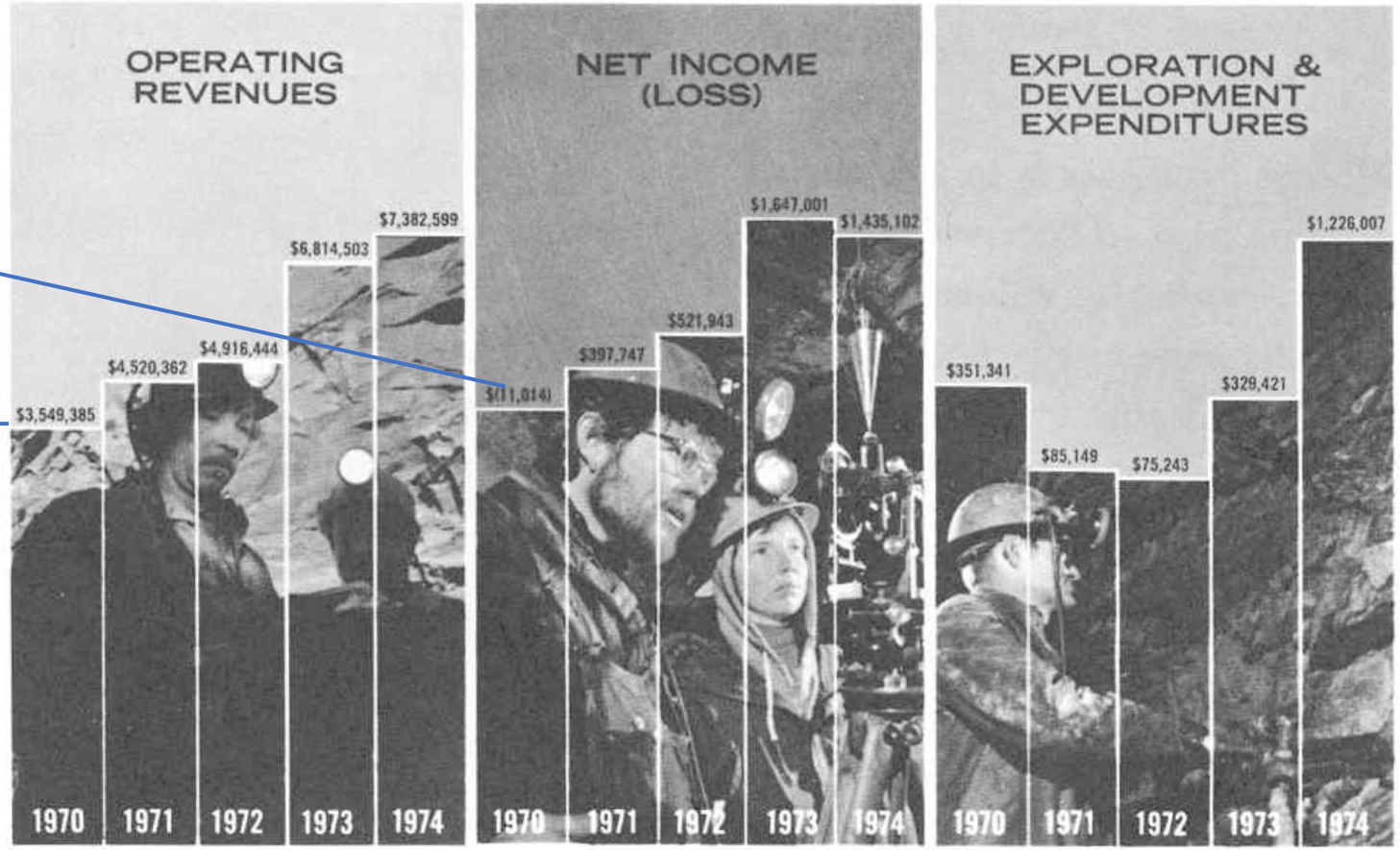
Tufte, "Visual Display of Quantitative Information" (1983)

Graphical Integrity



"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data."

Graphical Integrity



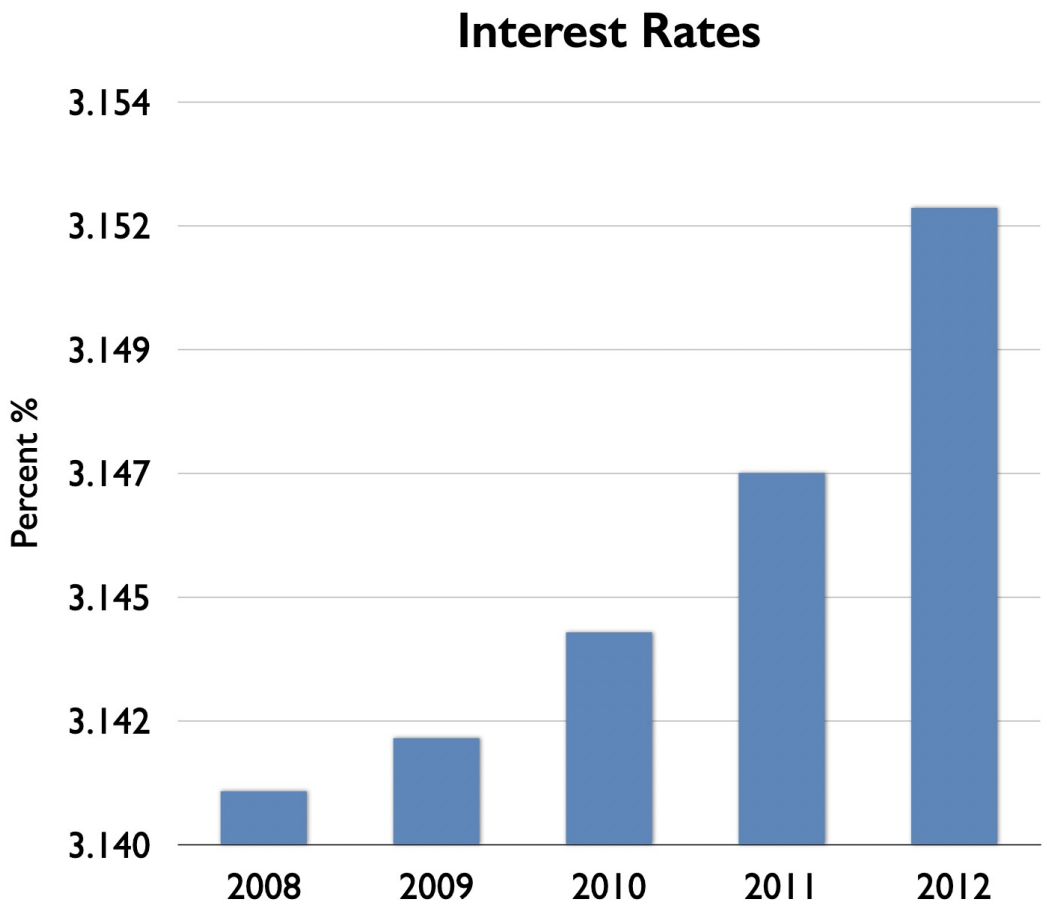
$11,014

$3,549,385

No Y-axis baseline!

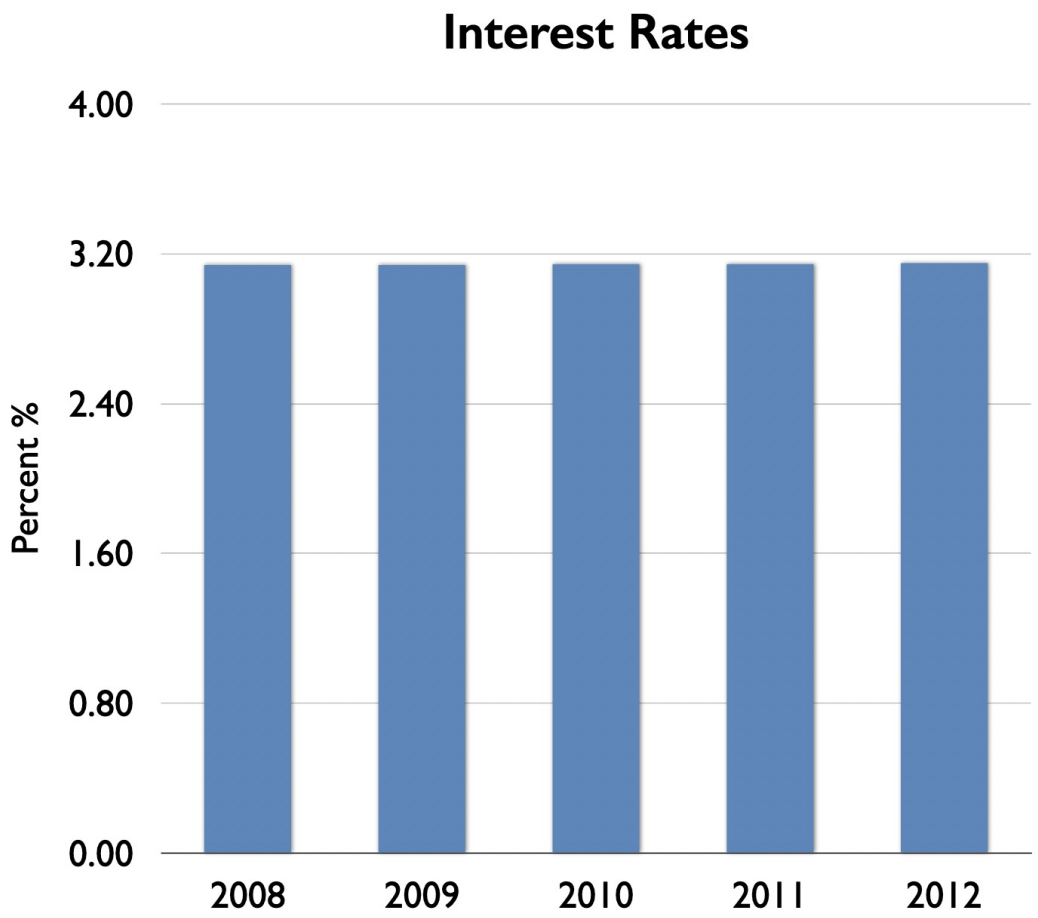
"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data."

Graphical Integrity



"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data."

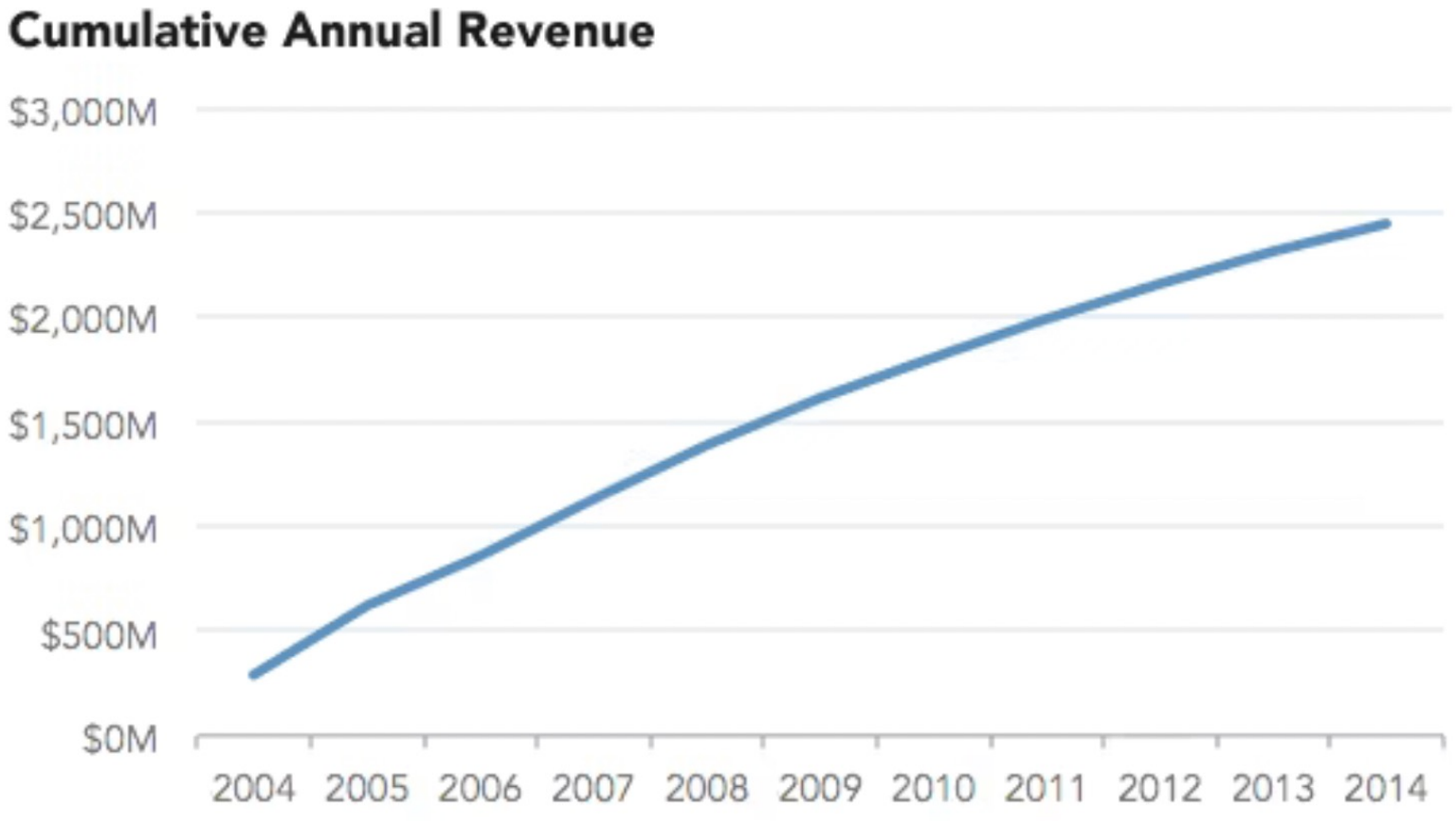
Graphical Integrity



Y-axis scale is important to show the context

"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data."

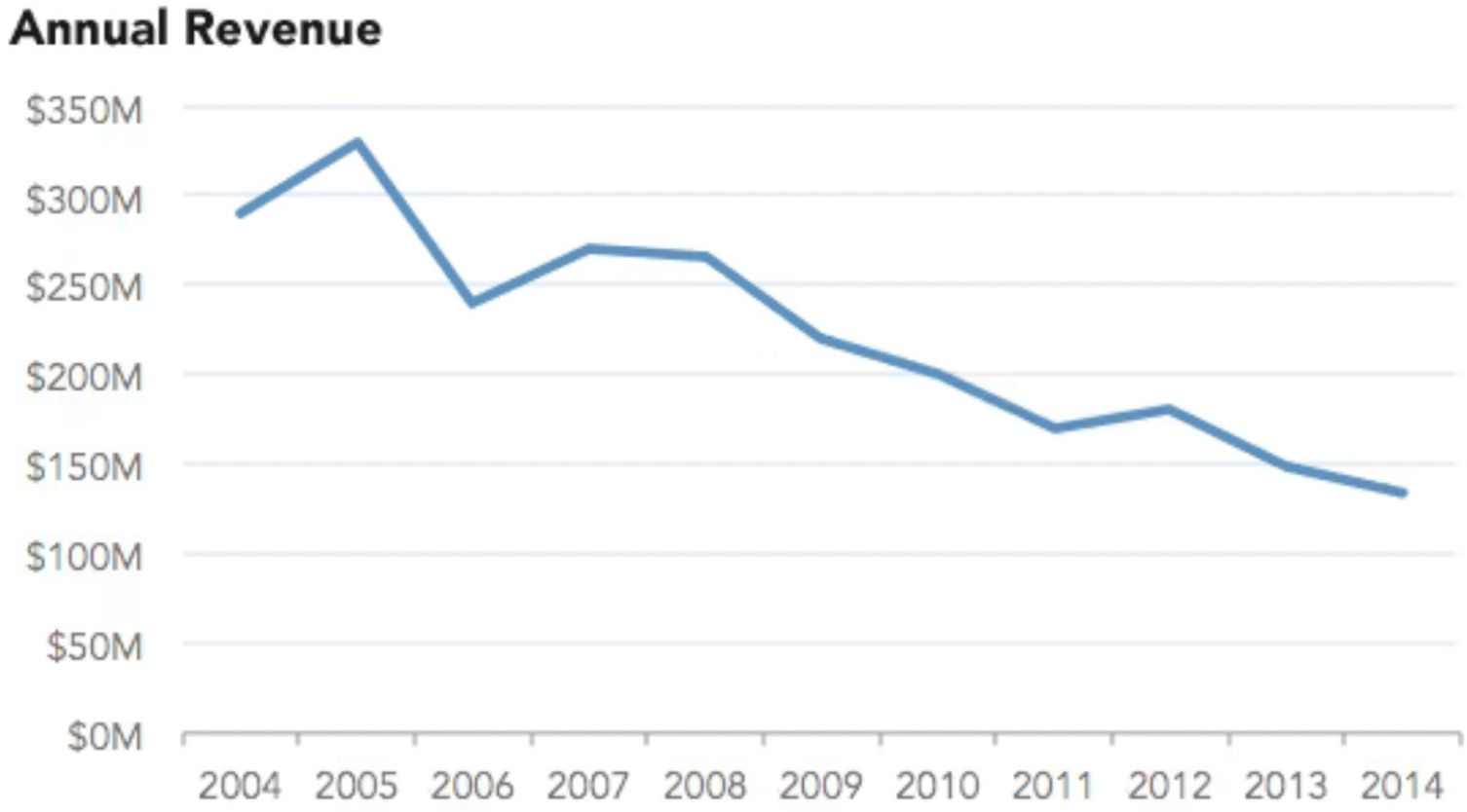
Graphical Integrity



Y-axis scale is important to show the context

"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data."

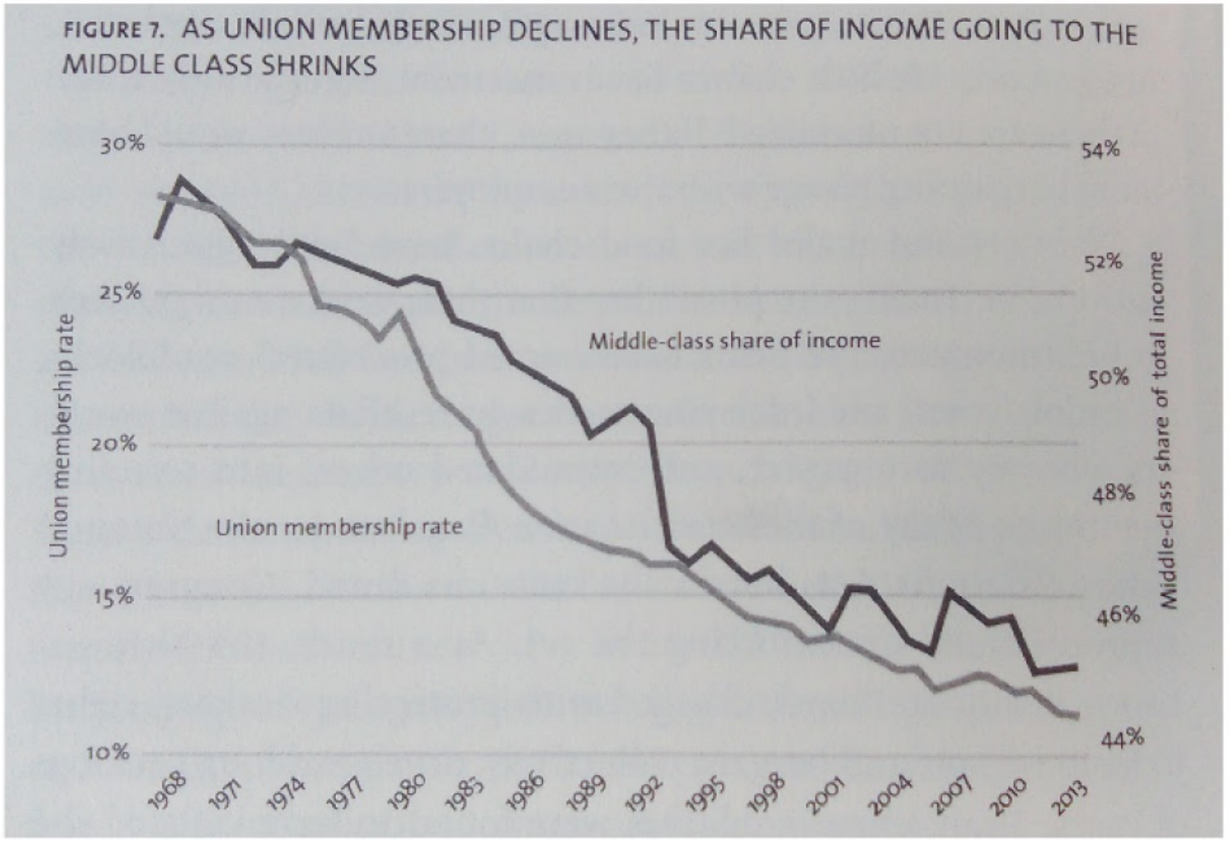
Graphical Integrity



Cumulative graphs can mislead

"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data."

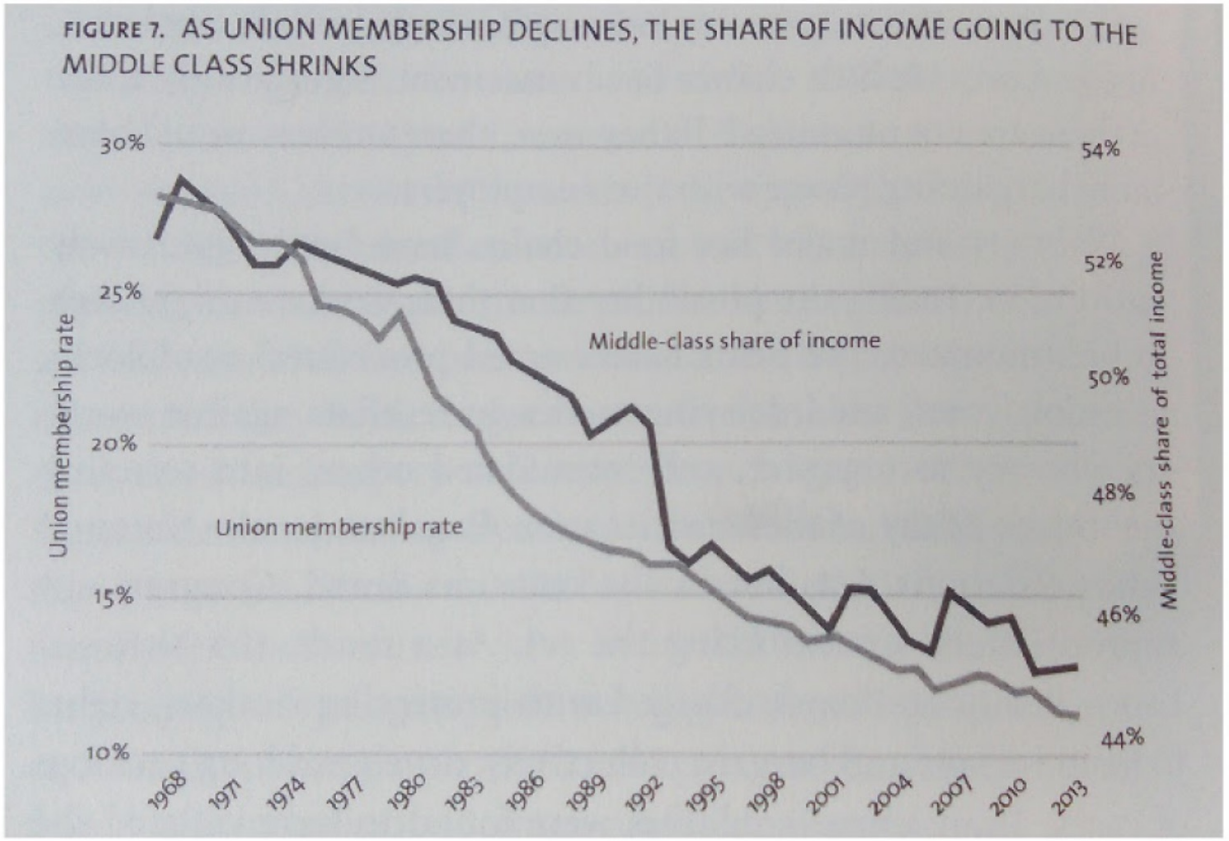
Graphical Integrity



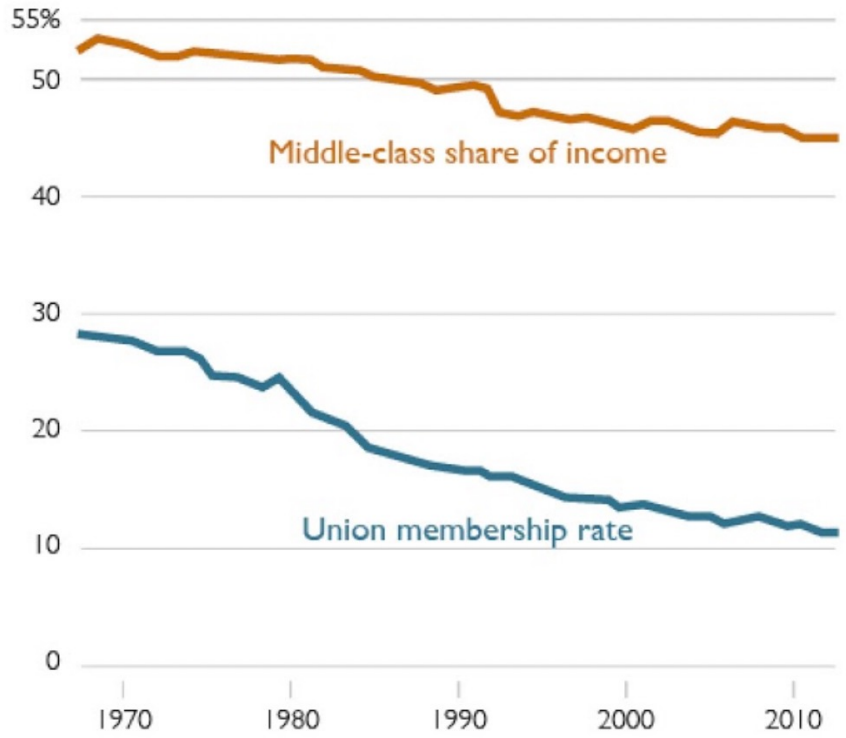
"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data."

Graphical Integrity

Double the axes, double the mischief

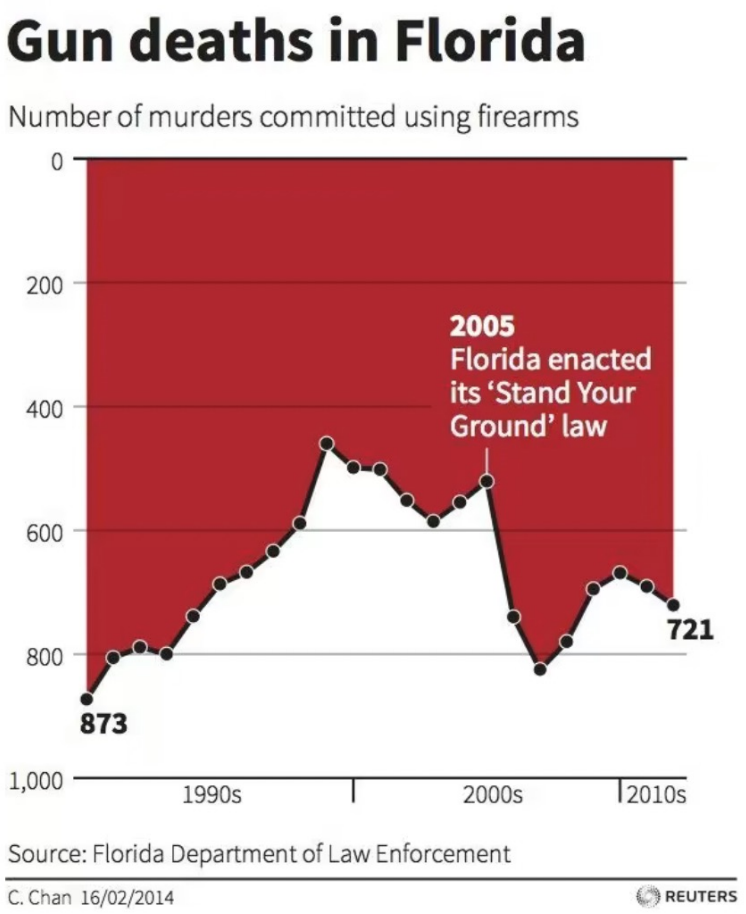


NEW VERSION



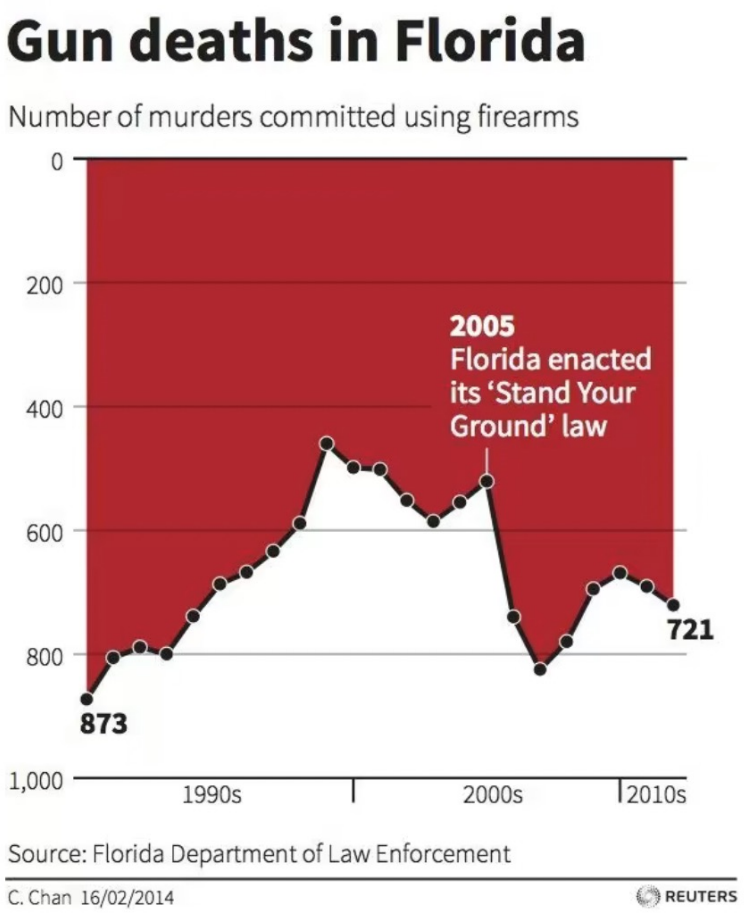
"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data."

Graphical Integrity



"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data."

Graphical Integrity



Y-axis is flipped!

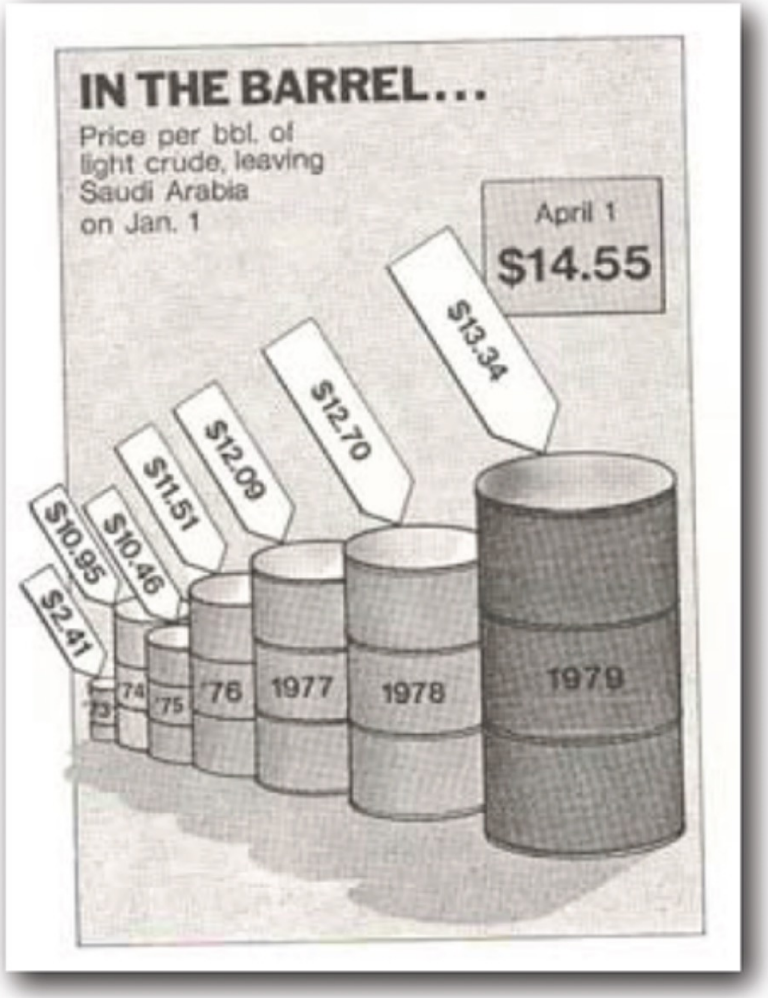
"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data."

Graphical Integrity

"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured."

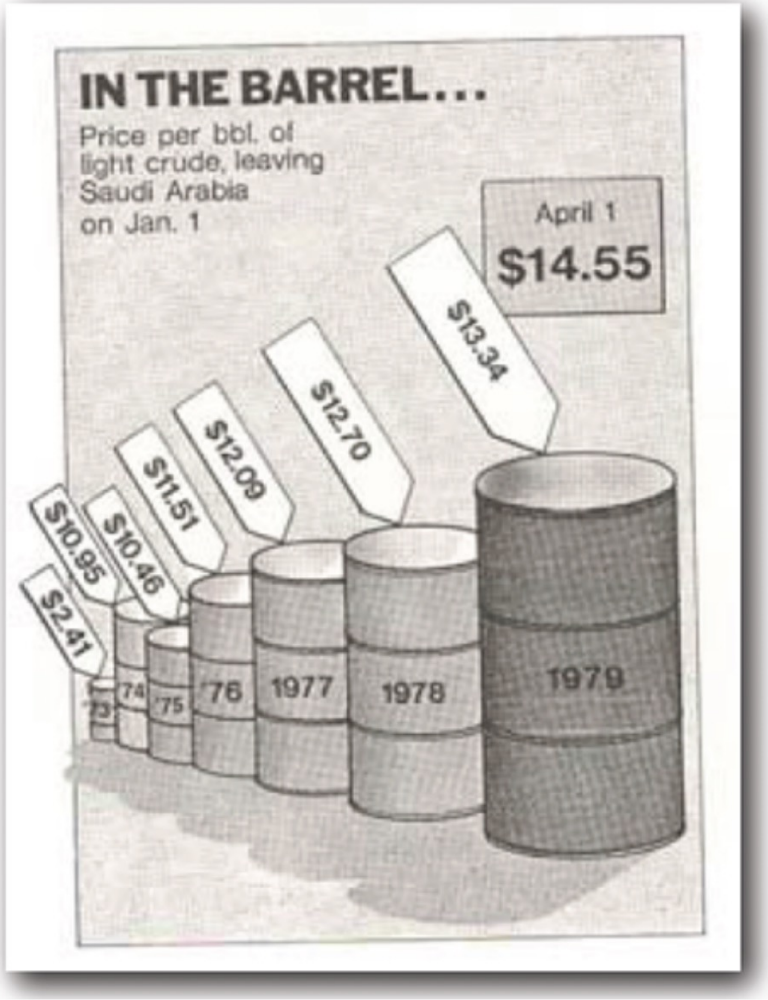
Tufte, "Visual Display of Quantitative Information" (1983)

Graphical Integrity



"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured."

Graphical Integrity

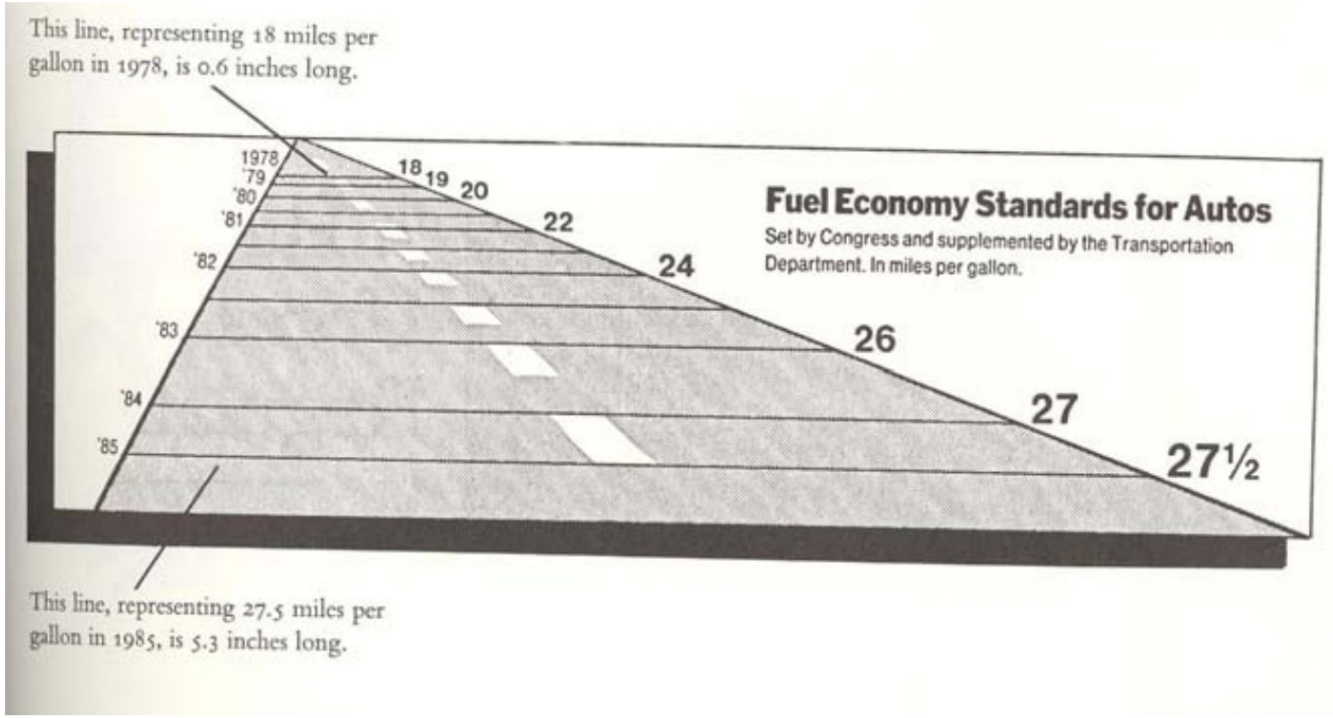


Inconsistent proportion of barrel sizes

"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured."

Graphical Integrity: Lie Factor

- Lie Factor = (Size of effect in graphic)/(Size of effect in data)
- Lie Factor = >1, overstating
- Lie Factor = 1, accurate
- Lie Factor = <1, understating



Graphical Integrity: Lie Factor

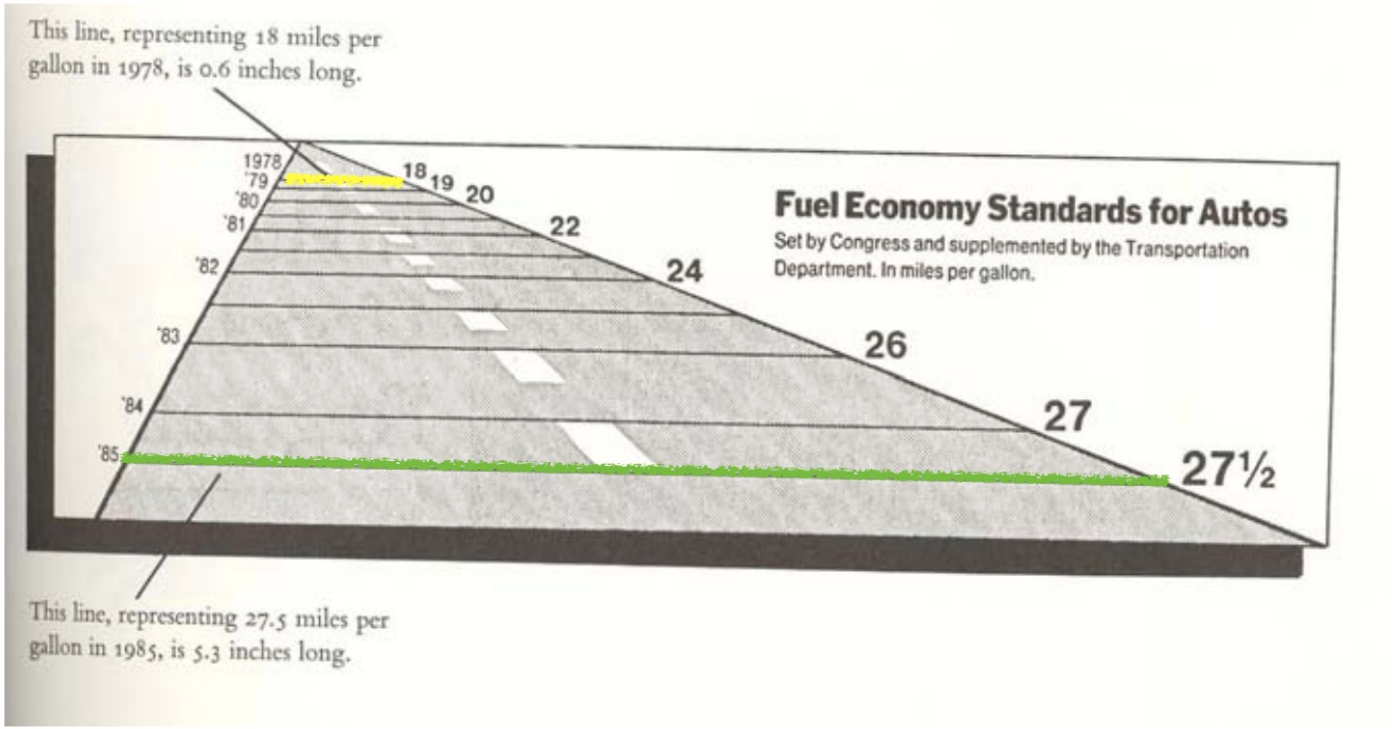


Image = $\frac{5.3" - 0.6"}{0.6"} = 7.83 = 783\%$

Data = $\frac{27.5 - 18}{18} = 0.53 = 53\%$

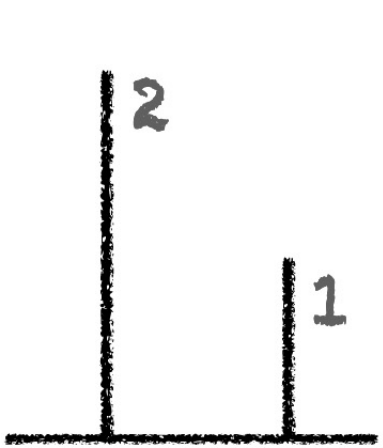
Lie Factor = $\frac{783\%}{53\%} = 14.8$

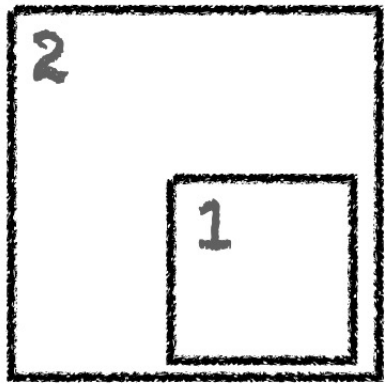
Lie Factor = >1, overstating

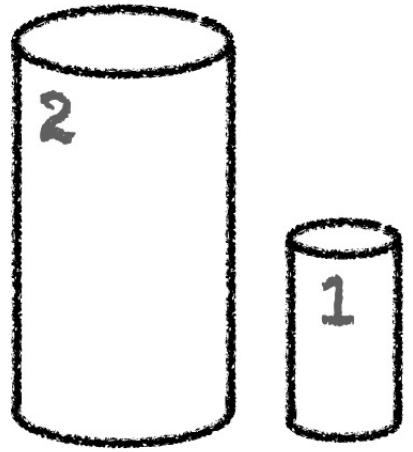
"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured."

Graphical Integrity: Lie Factor

Lie Factor = (Size of effect in graphic)/(Size of effect in data)



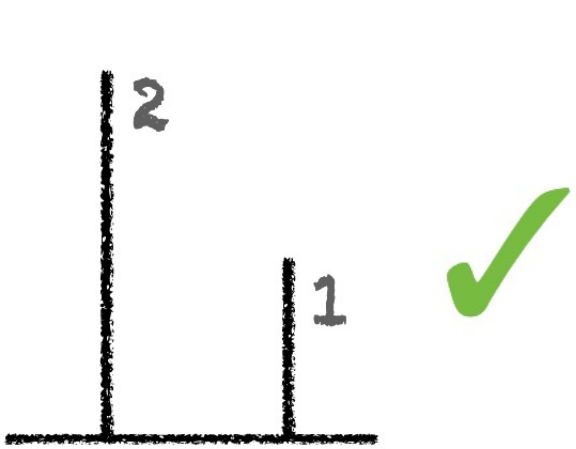


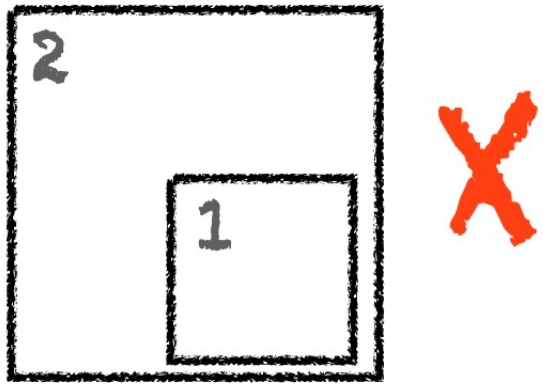


"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured."

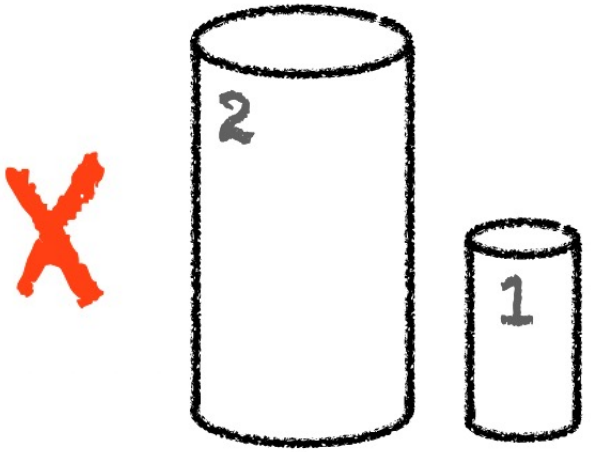
Graphical Integrity: Lie Factor

Lie Factor = (Size of effect in graphic)/(Size of effect in data)





Make sure area is proportional to data!



3D bar charts are bad

"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured."

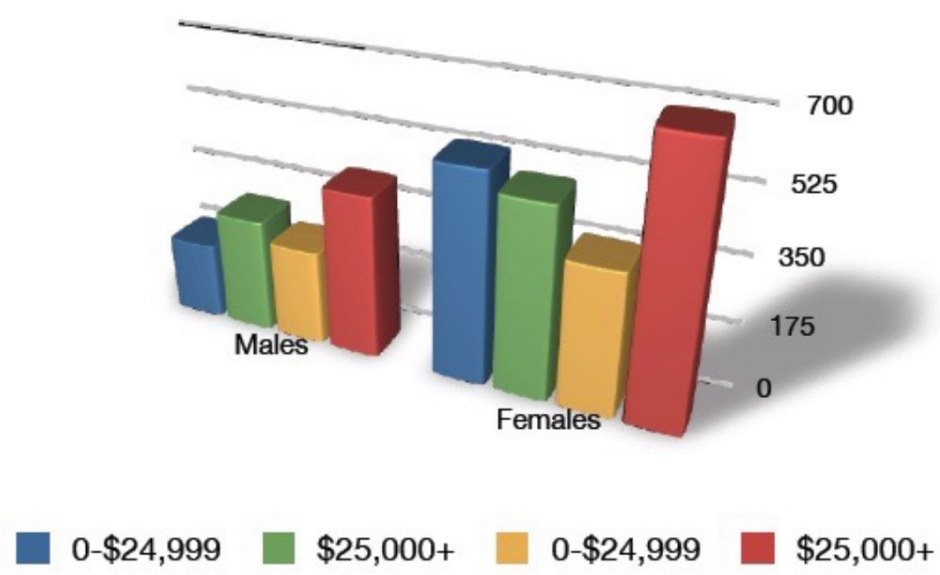
Graphical Integrity

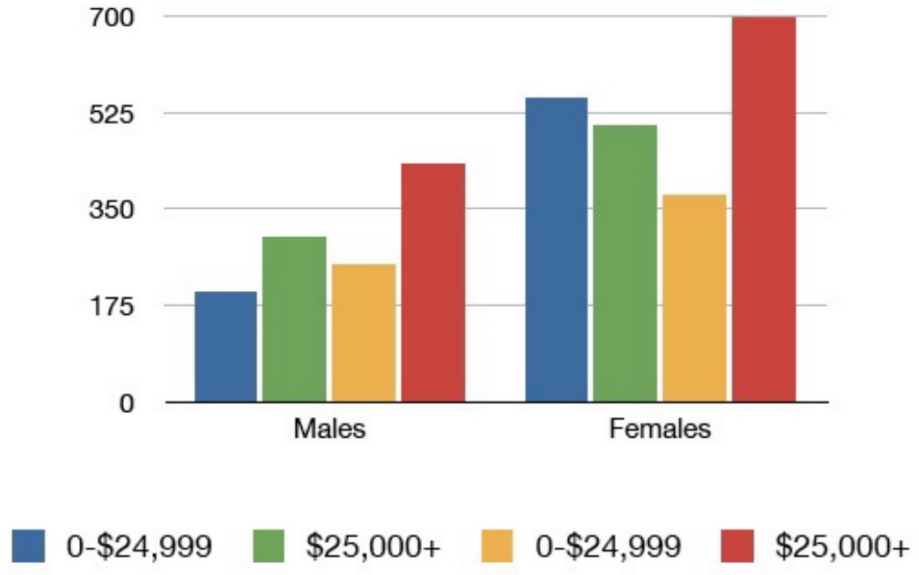
"The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data."

Tufte, "Visual Display of Quantitative Information" (1983)

Graphical Integrity

- No Unjustified 3D



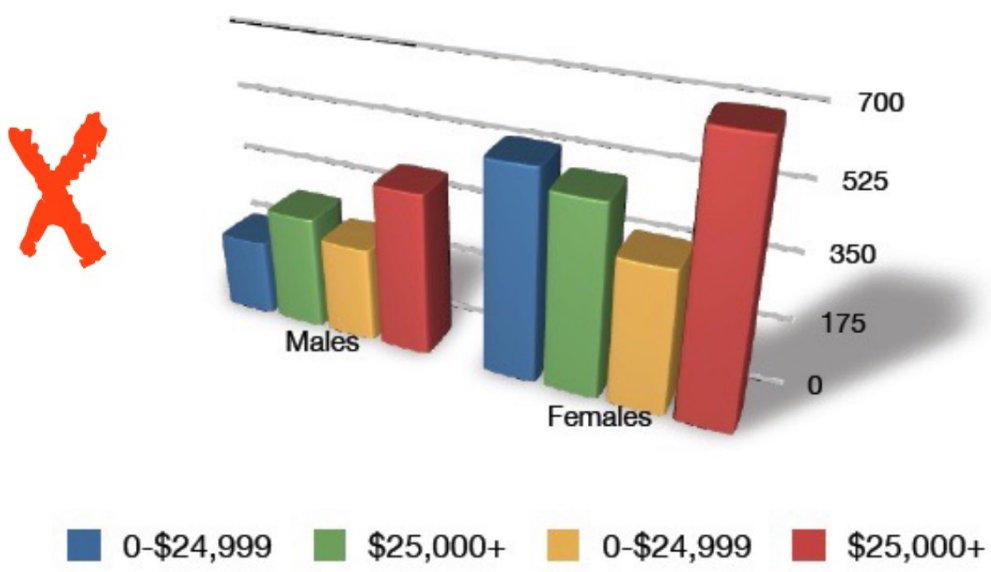


"The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data."

Graphical Integrity

- No Unjustified 3D

Dimensions in data: 2
Dimensions in plot: 3

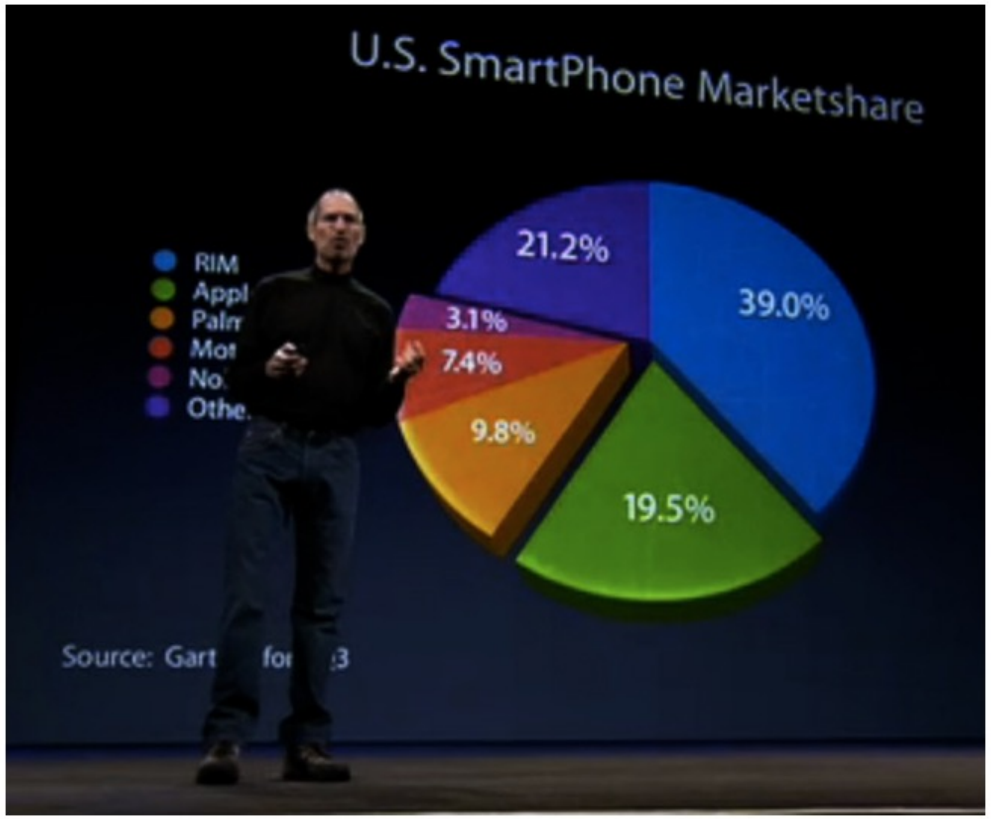


Dimensions in data: 2
Dimensions in plot: 2

"The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data."

Graphical Integrity

- No Unjustified 3D



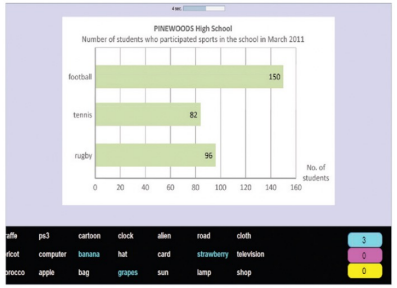
"The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data."

ChartJunk Debate

- All elements in visualization that are not necessary for interpreting the information related to the data being shown

- Heavy or dark grid lines
- Unnecessary text
- Inappropriately complex or gimmicky font faces
- Ornamented chart axes
- Backgrounds or icons within data graphs
- Ornamental shading and unnecessary dimensions

An Empirical Study on Using Visual Embellishments in Visualization

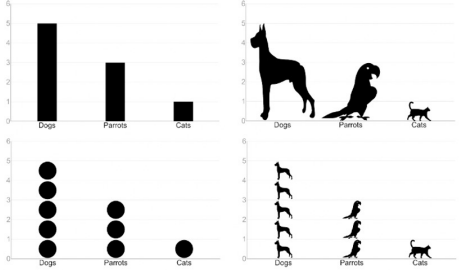


Borgo, et al. (2012)

What makes a visualization memorable?

Borkin, et al. (2013)
Borkin, et al. (2015)

ISOTYPE Visualization – Working Memory, Performance, and Engagement with Pictographs

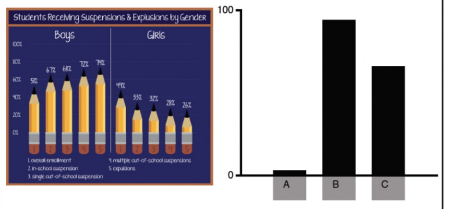


Haroz, et al. (2015)

Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts

Bateman, et al. (2010)

An Evaluation of the Impact of Visual Embellishments in Bar Charts



Skau, et al. (2015)

Tufte: Graphical displays should...

- Show the data
- Avoid distorting what the data have to say
- Encourage comparisons
- Reveal the data at several levels of detail
- Serve a reasonably clear purpose
- Be closely integrated with the statistical and verbal descriptions

Principled Approaches for Information Visualization

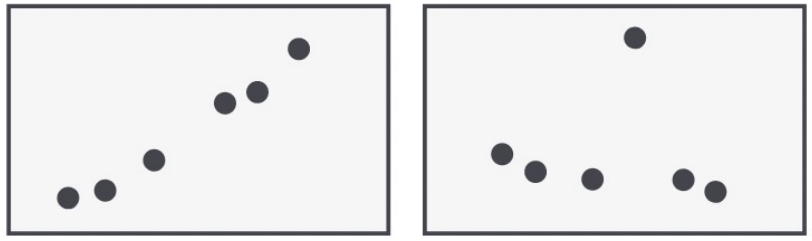
Principled Approaches for InfoVis

- Views & Facet & Linked Data
- Focus and Context
- Filtering and Aggregation

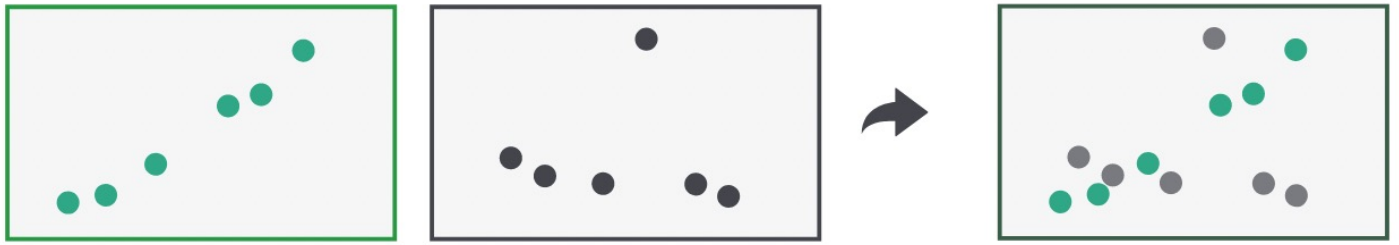
Views & Facet & Linked Data

Views and Facet

- Split visualization plots into multiple views or separate into multiple layers
- Benefit?
  - Complexity reduction
  - Rely on vision instead of memory retrieval!



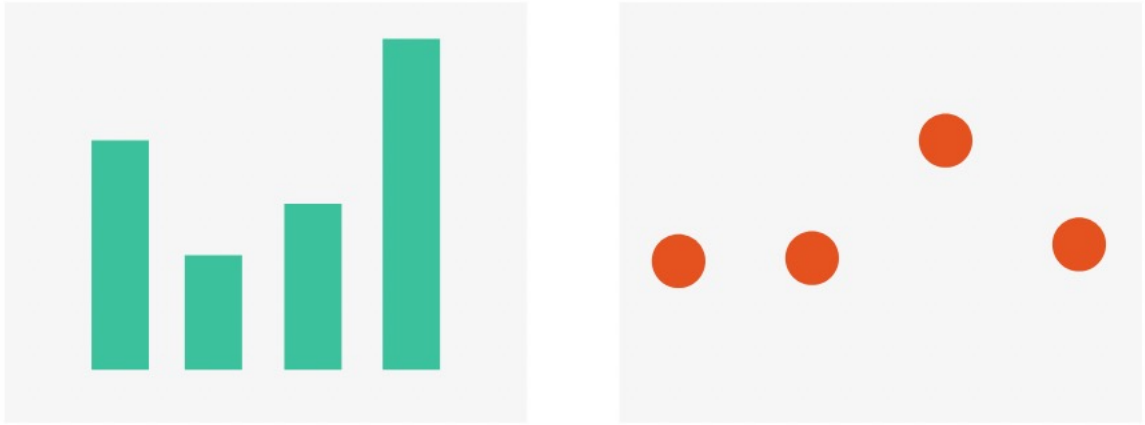
Side-by-side views



Superimpose as layers

Views and Facet: Side-by-Side

- Partition into Side-by-Side Views
  - Easy to compute and build
  - BUT: Multiple views take up more space!
- Side-by-side: Juxtapose
  - Small multiples
  - Multiform

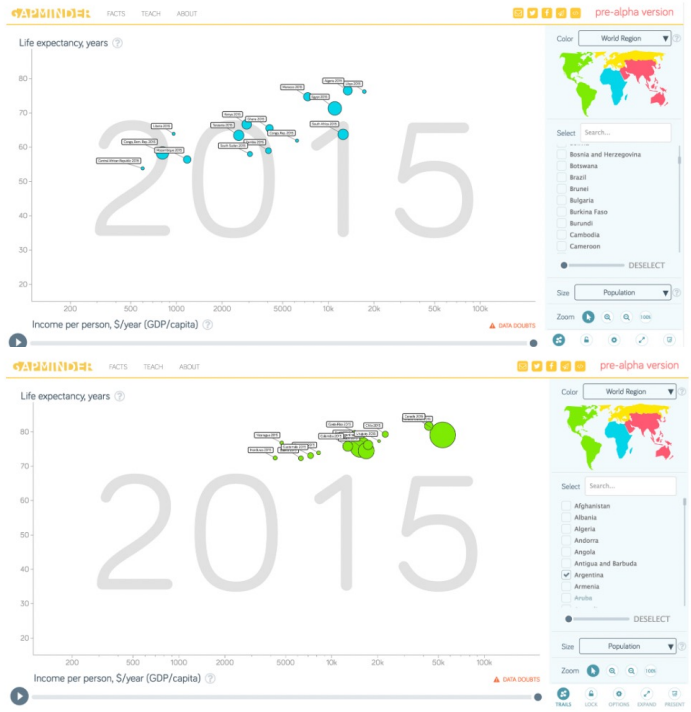


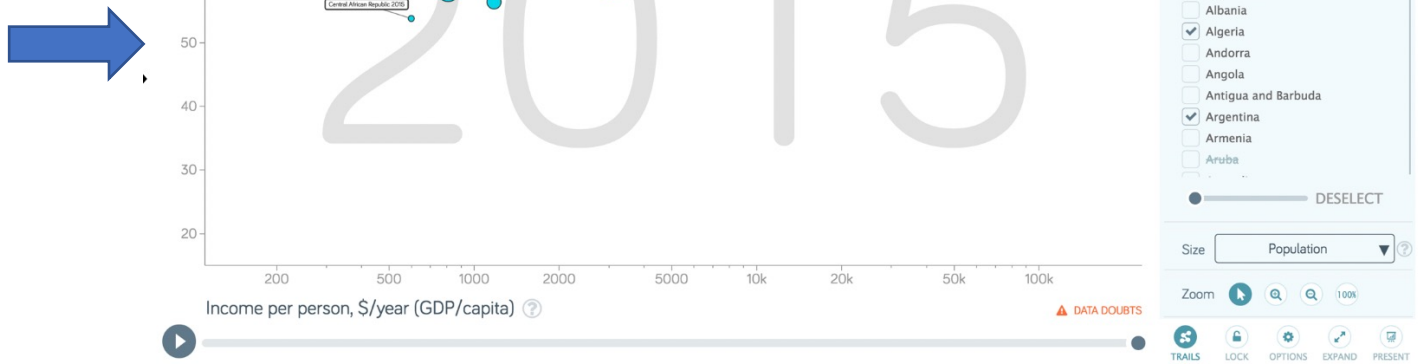
Multiform: Same data, different encoding

Small multiples

Views and Facet: Superimpose Layers

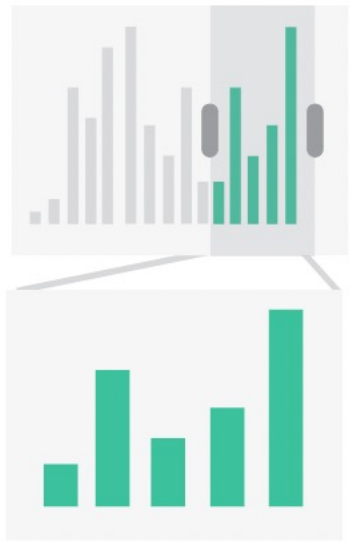
- Superimpose visualization layers to show data patterns
  - Less screen space required
  - still easy to compare
  - BUT: limits encoding options
  - Can get messy soon with multiple layers



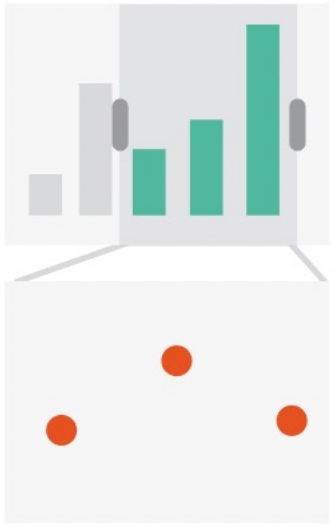


Views and Facet: Overview and Detail

- Provides detailed view of a subset
- Benefit: For large or complex data, a single view of the entire dataset cannot present fine details



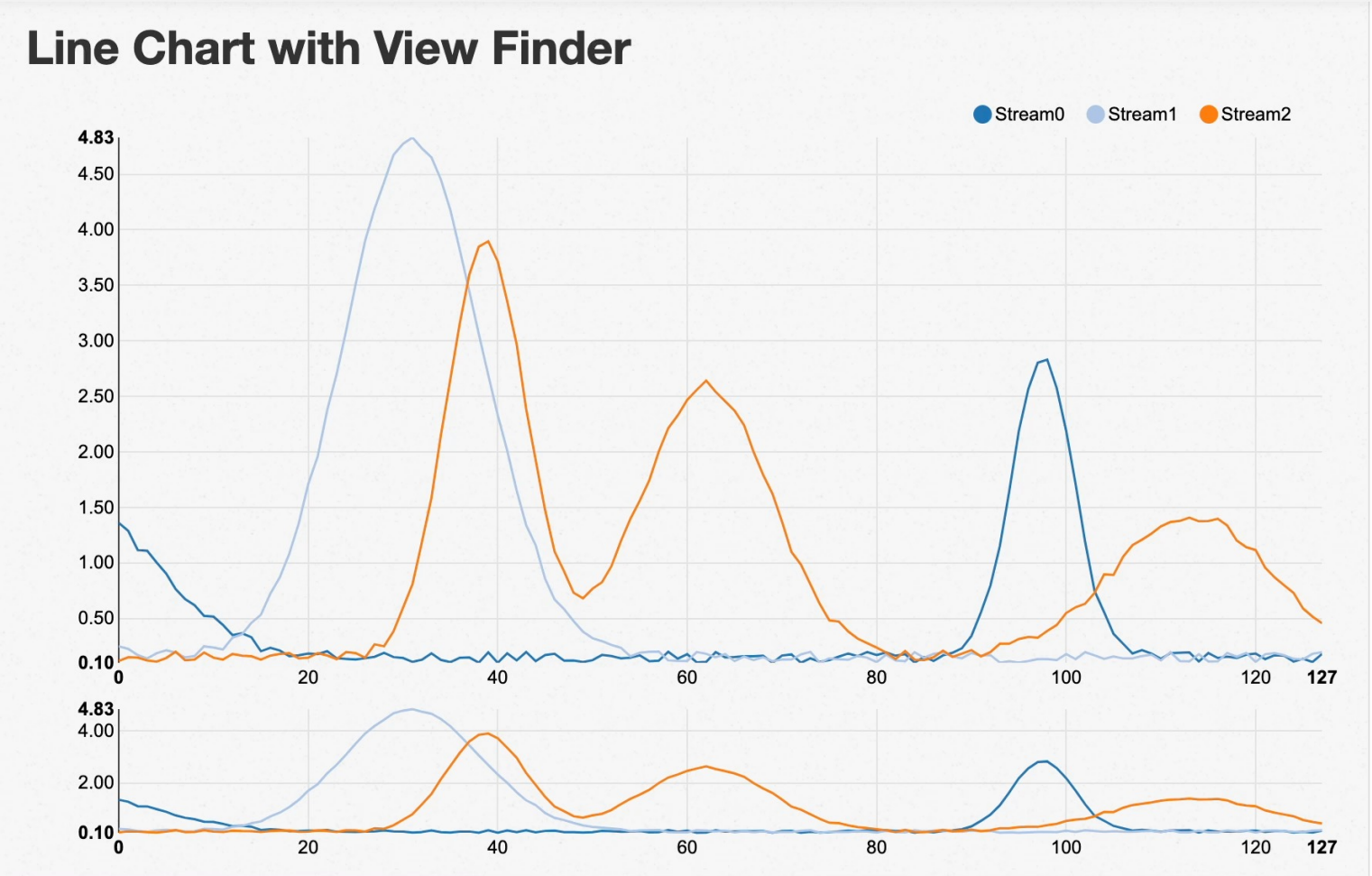
Overview/ Detail



Multiform, Overview/ Detail

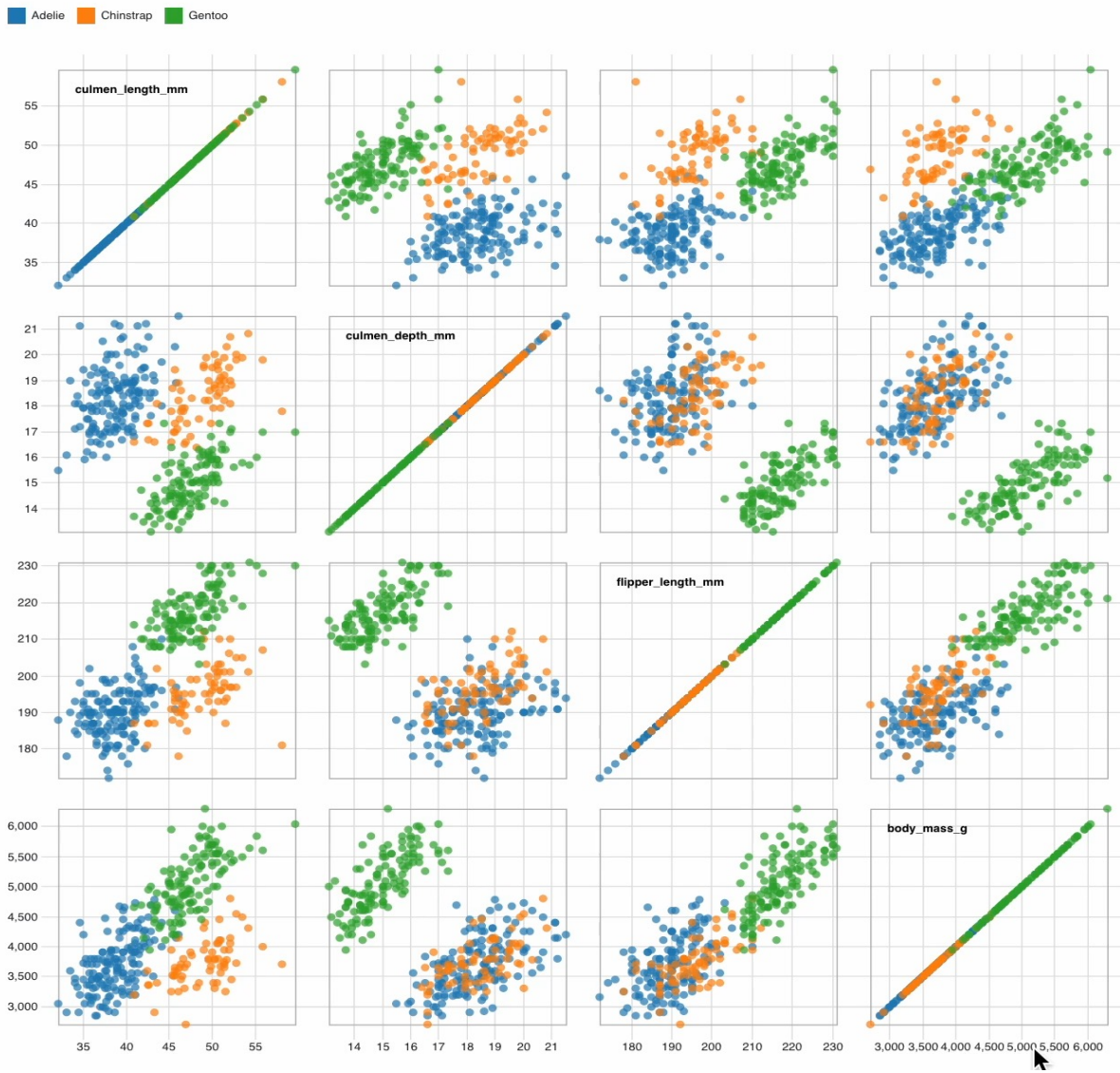
Views and Facet: Brush and Zoom

- Brush and Zoom



Views and Facet: Brush and Link

- Brush and Link
- Multiple views that are simultaneously visible and linked together such that actions in one view affect the others

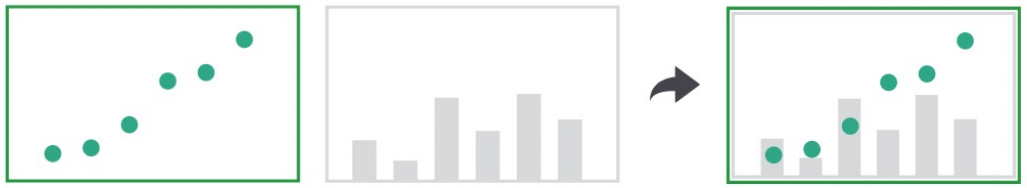


Focus and Context

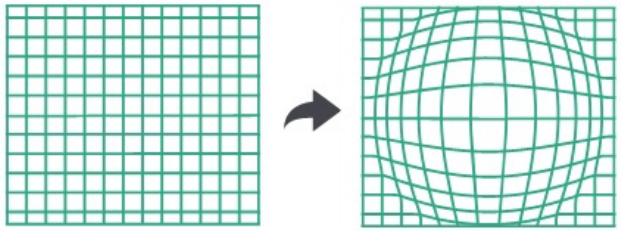
Focus and Context

- User selects region of interest (focus) through navigation or selection
- Provide context through aggregation, reduction, or layering
- Carefully pick what to show; hint at what you are not showing

Elide data



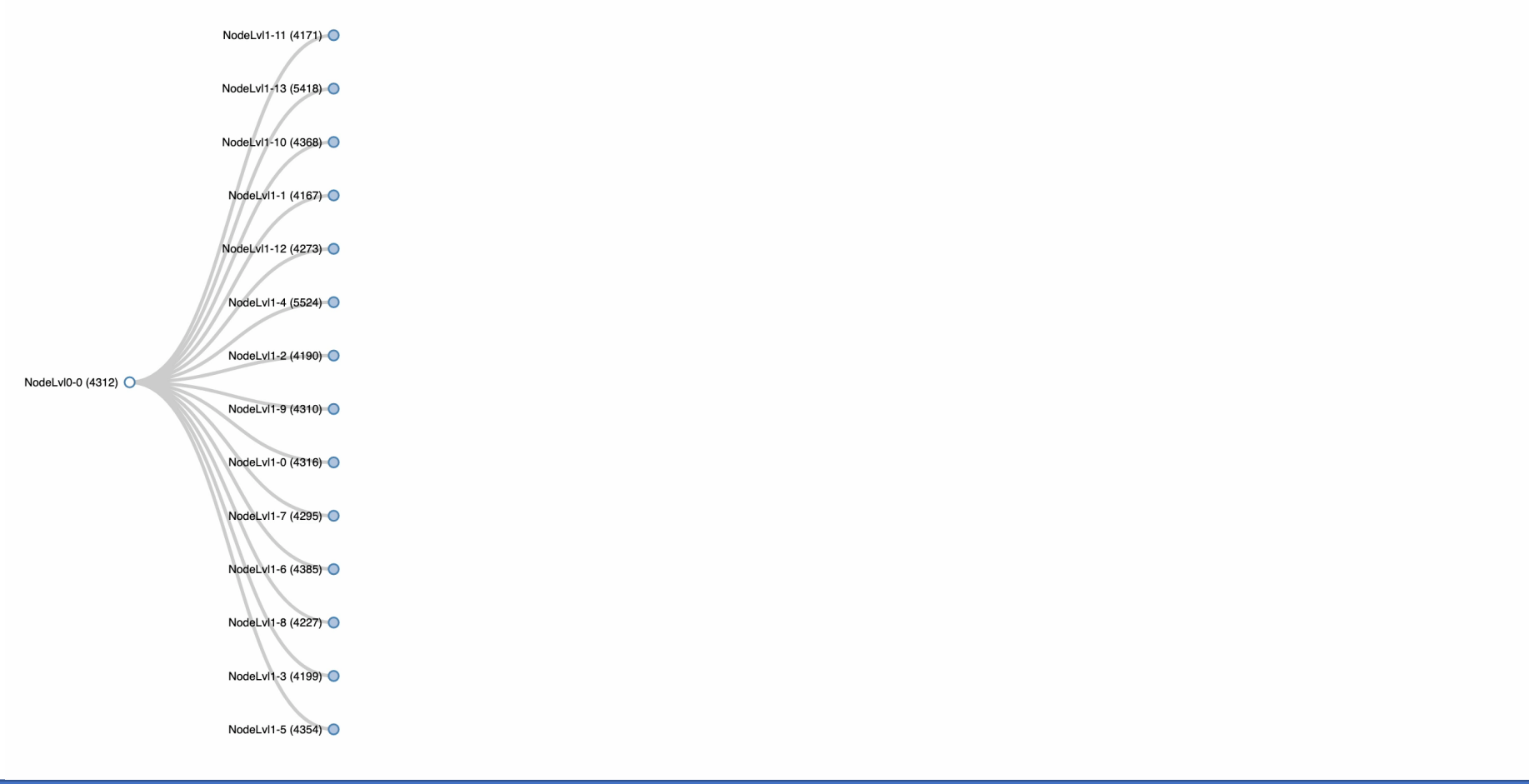
Superimpose layers



Distort geometry

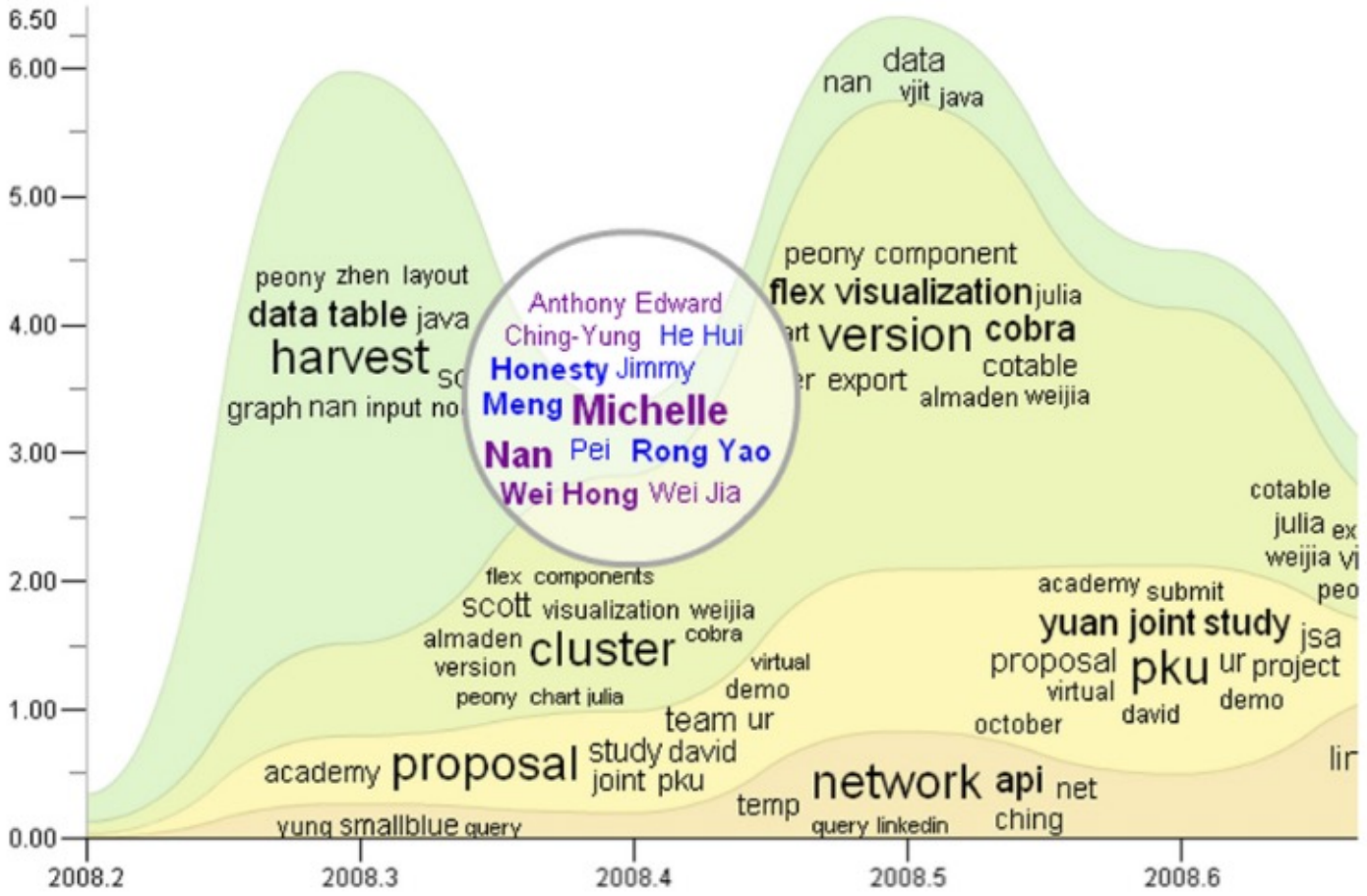
Focus and Context: Elision

- Focus items shown in detail, other items summarized for context – Collapsible tree



Focus and Context: Superimpose Layers

- Focus layer limited to a local region of view, instead of stretching across the entire view – Standard Lens



https://dl.acm.org/doi/pdf/10.1145/2089094.2089101

Focus and Context: Superimpose Layers

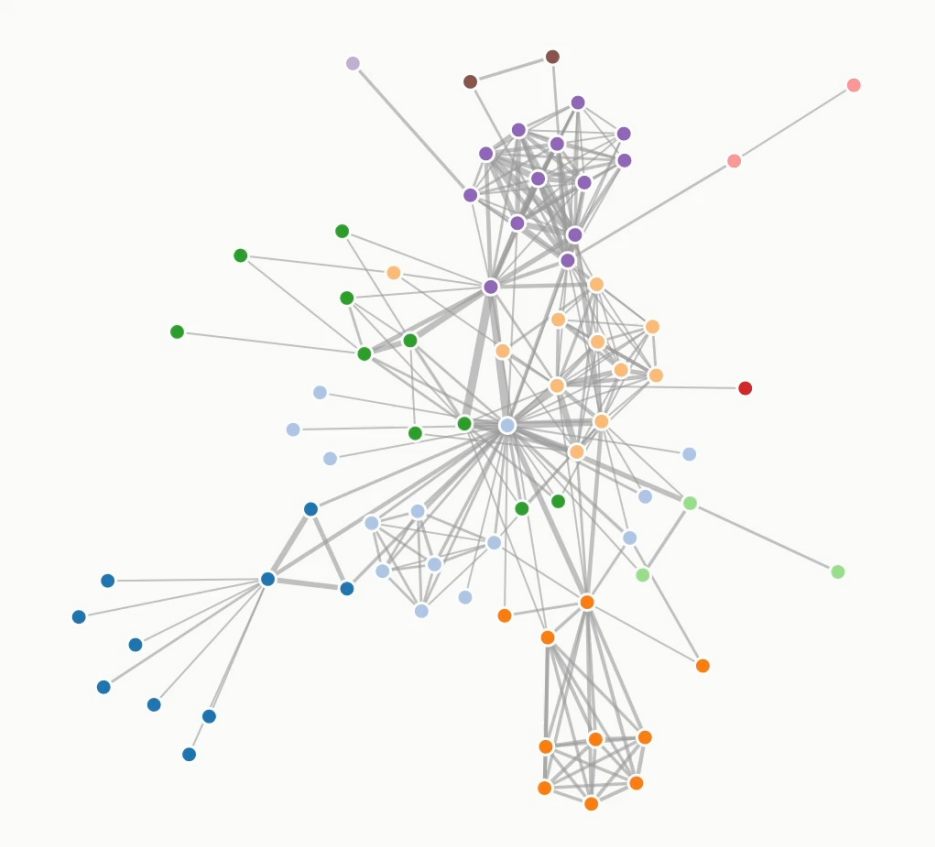
- Focus layer limited to a local region of view, instead of stretching across the entire view – Magic Lens

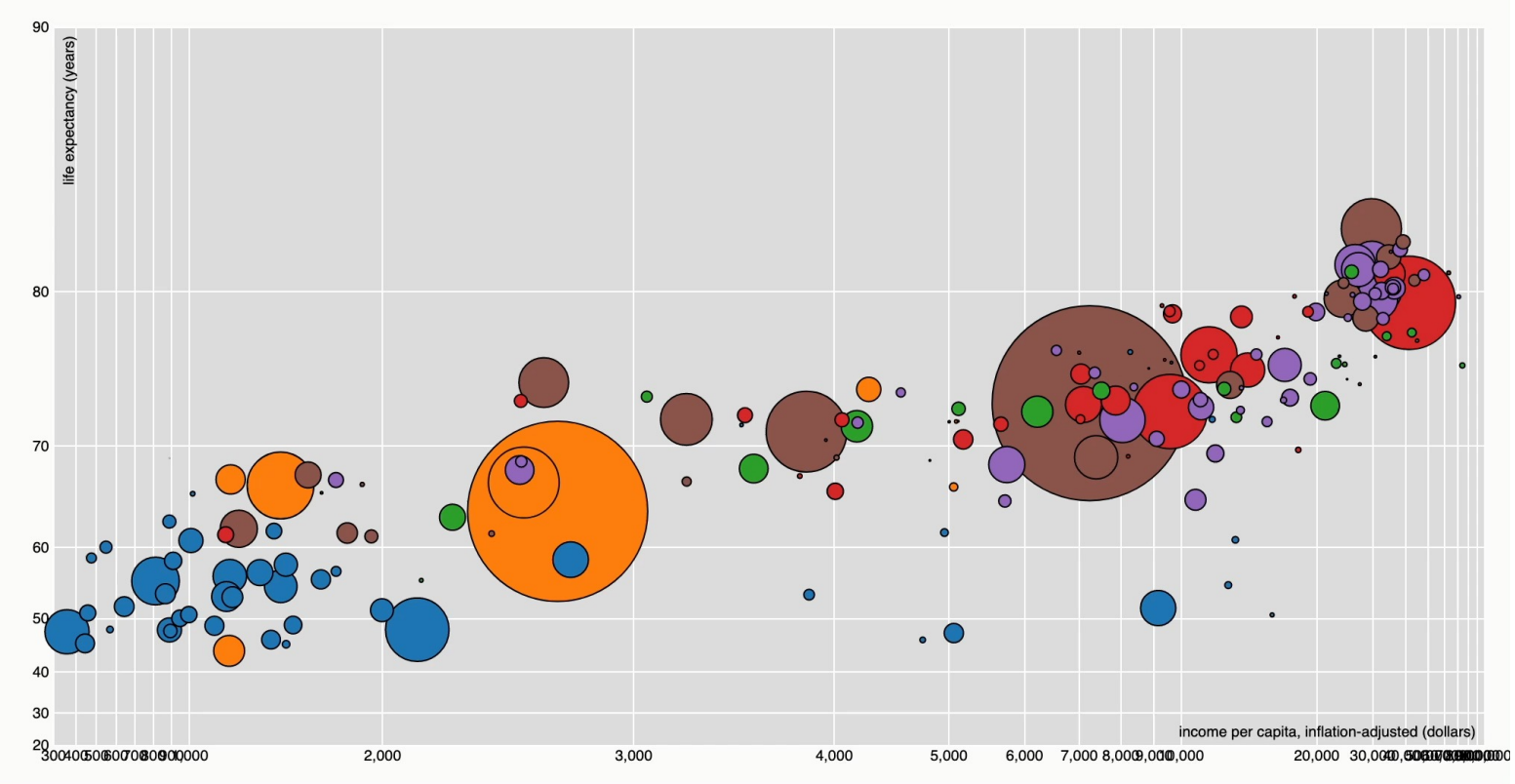
Focus and Context: Superimpose Layers

- Focus layer limited to a local region of view, instead of stretching across the entire view – FingerGlass

Focus and Context: Distort Geometry

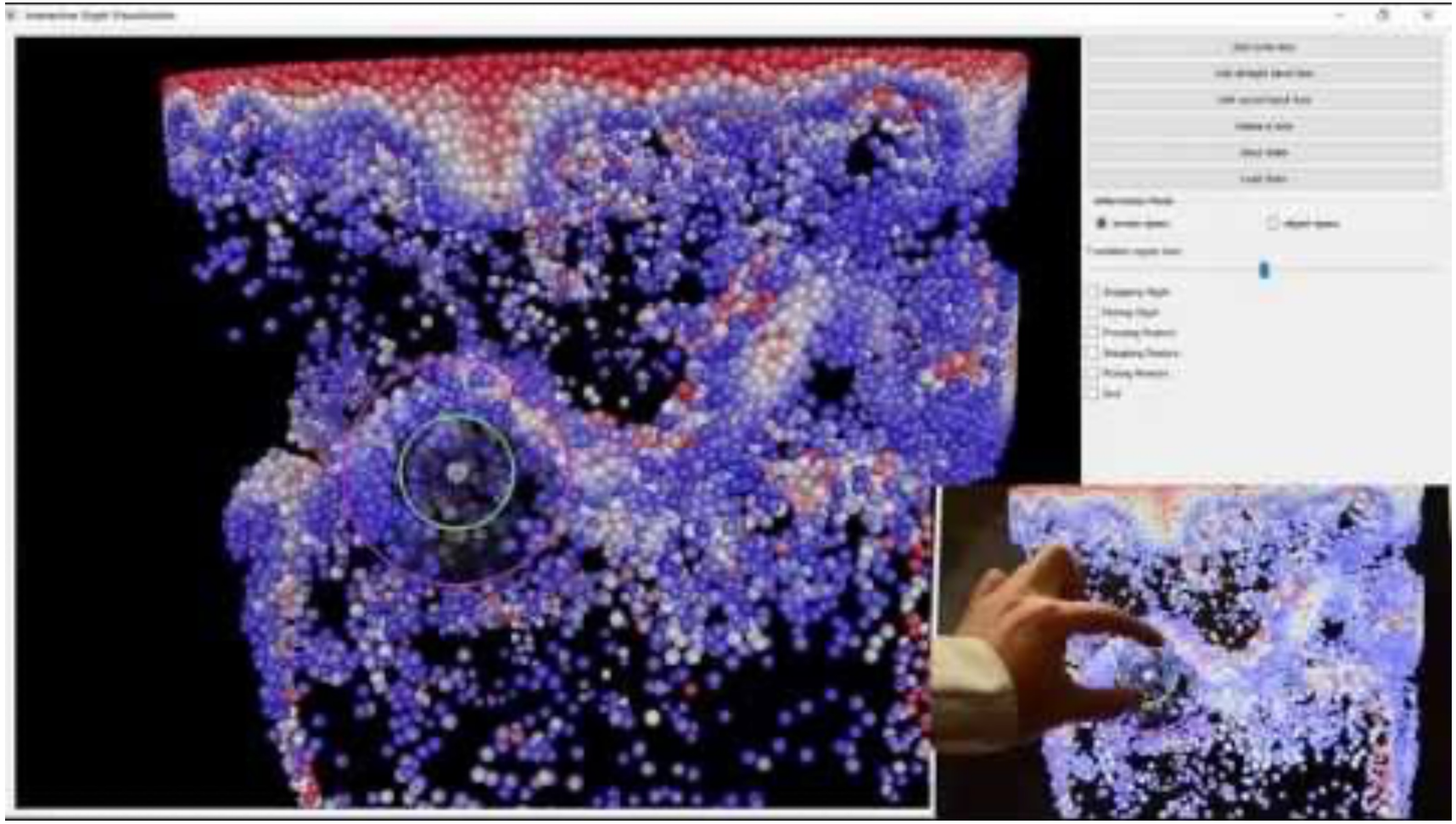
- Use geometric distortion of the contextual regions to make room for the details in the focus region(s) – Fisheye Lens





Focus and Context: Distort Geometry

- Use geometric distortion of the contextual regions to make room for the details in the focus region(s) – GlyphLens



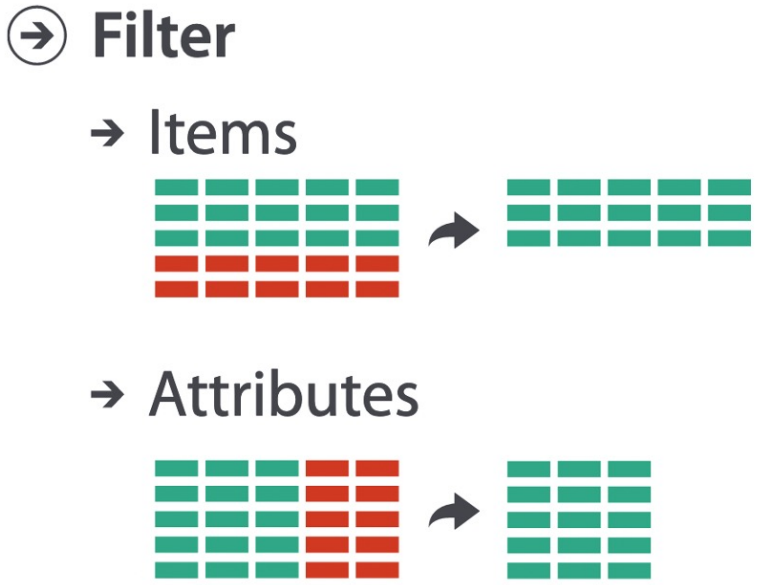
Filtering and Aggregation

Filtering and Aggregation

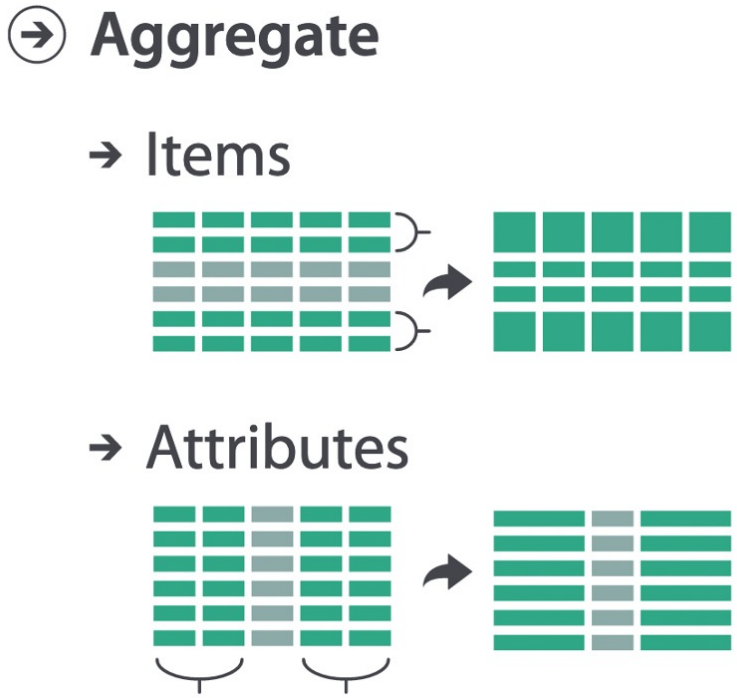
- Purpose: Complexity reduction in exploratory visual data analysis
- Reduce amount of data shown
  - Strategy for complexity reduction
  - Be careful not to hide important details
  - Can reduce items and/or attributes

Filtering vs Aggregation

- Filter: Eliminate data elements

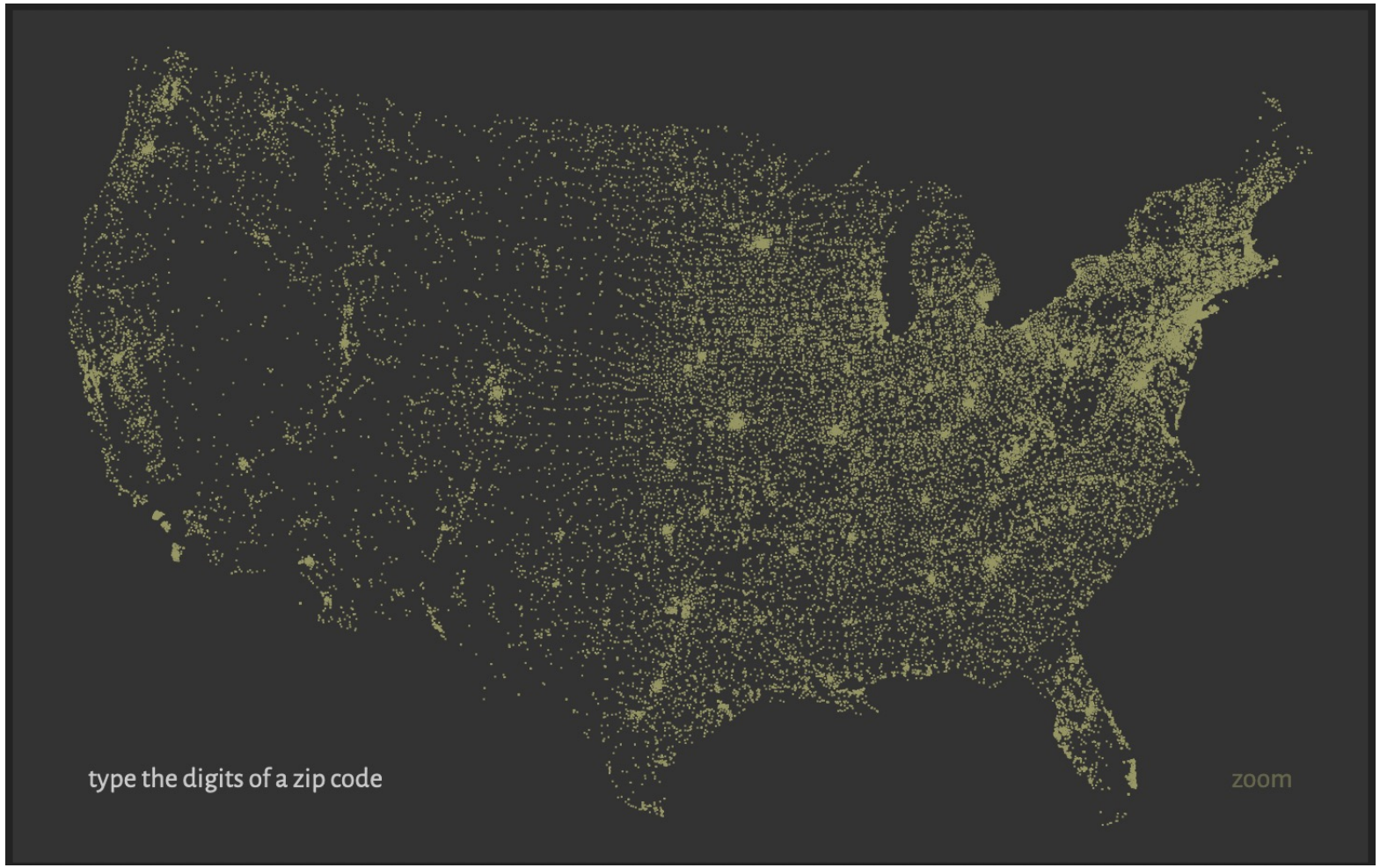


- Aggregate: Create new elements from multiple raw elements



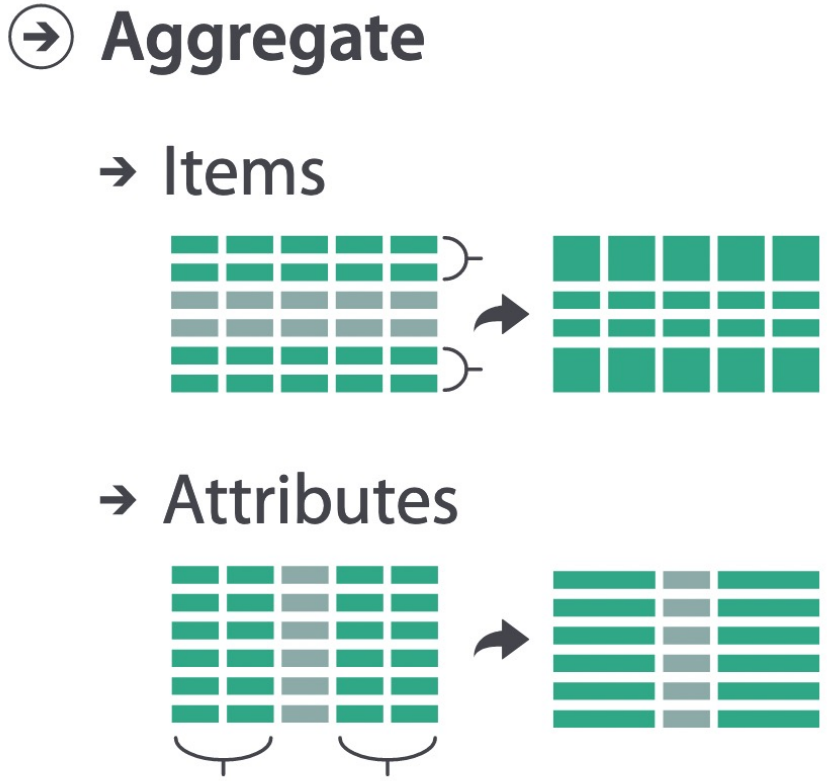
Filtering: Dynamic Query

http://benfry.com/zipdecode/



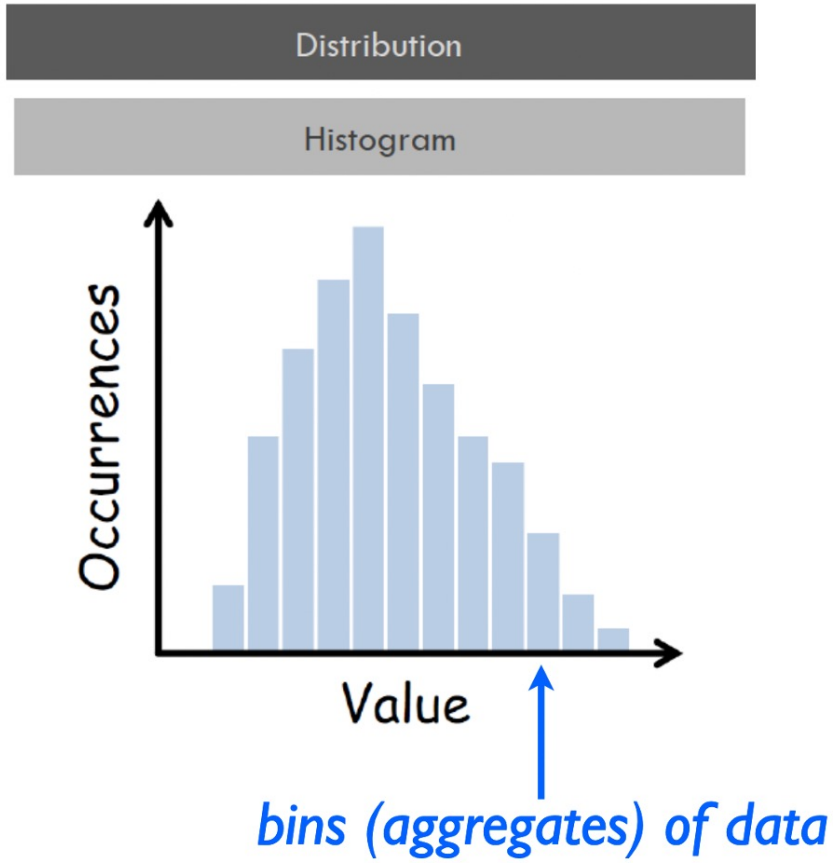
Aggregation

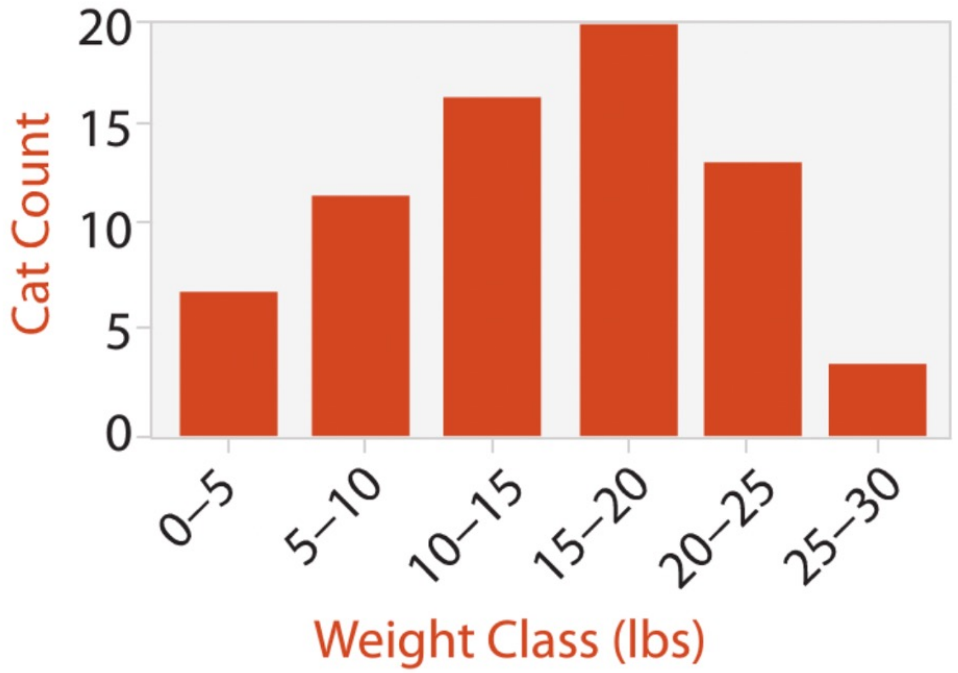
- Aggregate = Create new element representing multiple raw elements



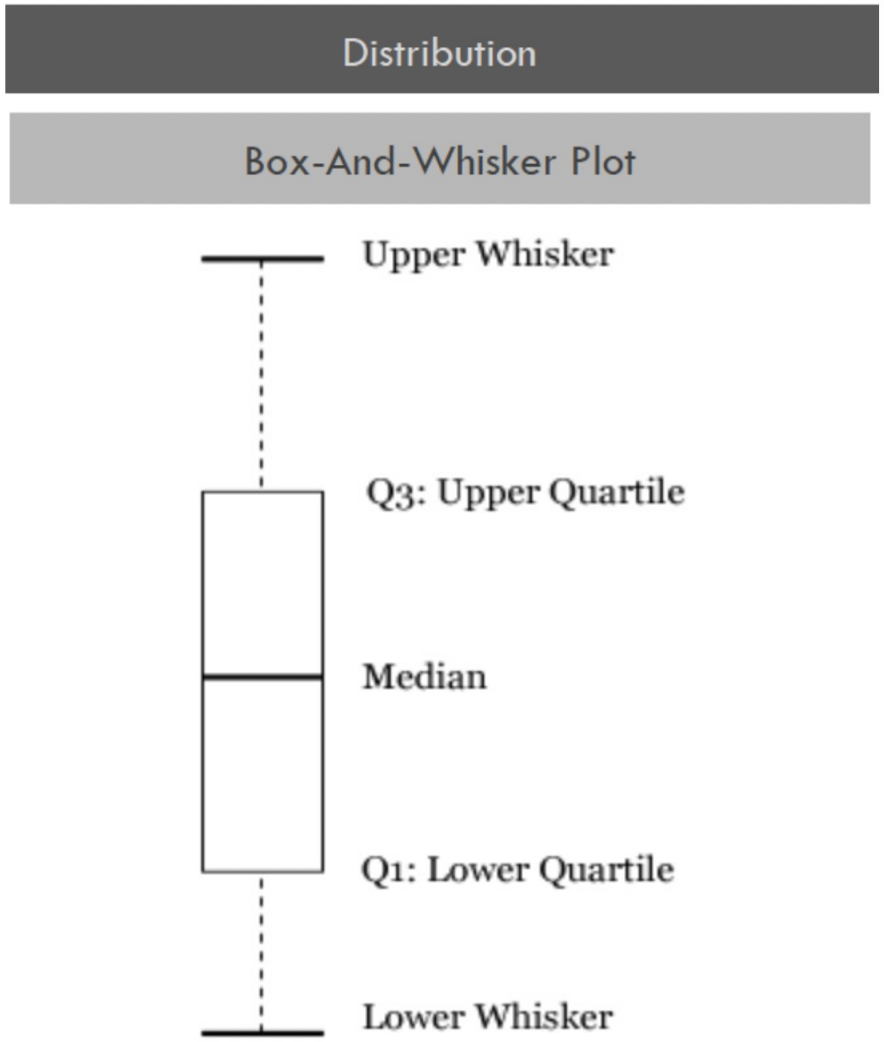
- How to Aggregate?
  - Item aggregation
  - Attribute aggregation (i.e., dimensionality reduction)
  - Spatial aggregation

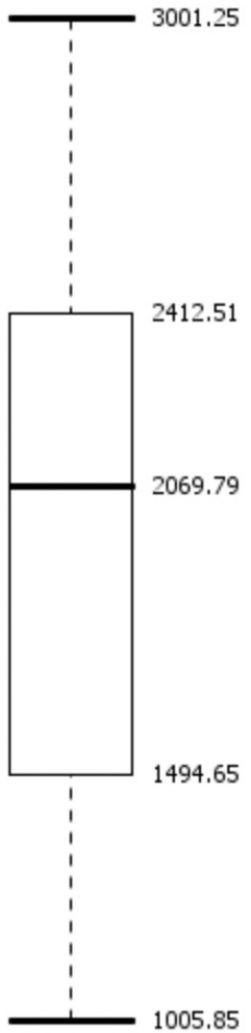
Aggregate Items: Histograms





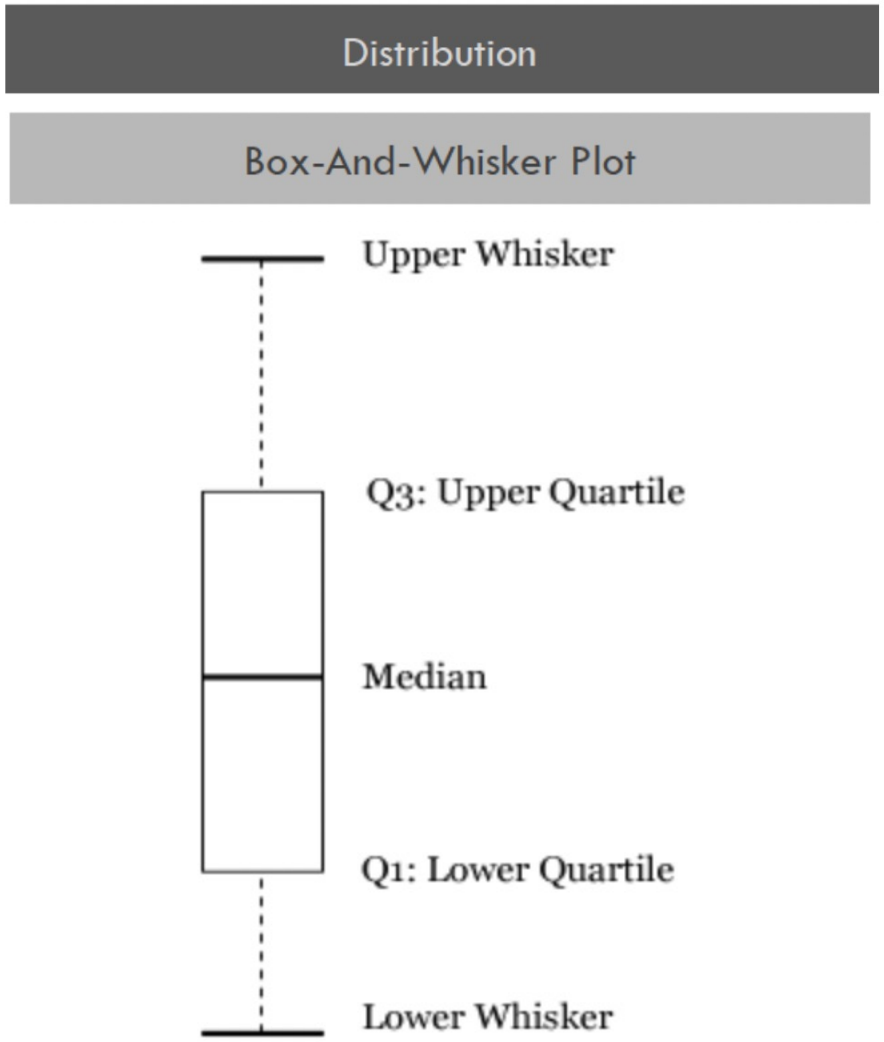
Aggregate Items: Box and Whisker Plot





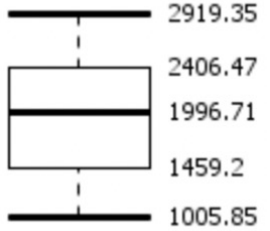
Citizen Nr.	Income	Key Value
25	3,001.25	Maximum
24	2,996.45	
23	2,919.35	
22	2,787.02	
21	2,784.72	
20	2,696.83	
19	2,412.51	Q3: 0.75 * 25 = 18.75 = 19.
18	2,400.43	
17	2,367.84	
16	2,333.37	
15	2,285.53	
14	2,214.87	
13	2,069.79	Median: 0.5 * 25 = 12.5 = 13.
12	1,923.62	
11	1,819.22	
10	1,773.34	
9	1,597.54	
8	1,589.48	
7	1,494.65	Q1: 0.25 * 25 = 6.25 = 7.
6	1,423.74	
5	1,391.92	
4	1,334.88	
3	1,184.53	
2	1,125.78	
1	1,005.85	Minimum

Aggregate Items: Box and Whisker Plot



+10345.67

Mean = 2,303.43,
Median = 1996.71



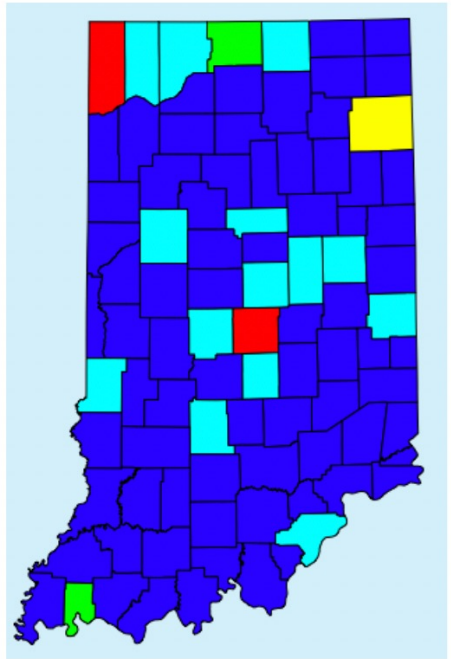
Citizen Nr.	Income	Key Value
24	10,345.67	Maximum
23	2,919.35	Upper Bound
22	2,787.02	
21	2,784.72	
20	2,696.83	
19	2,412.51	
18	2,400.43	Q3: (18. + 19.)/2 = 2,406.47
17	2,367.84	
16	2,333.37	
15	2,285.53	
14	2,214.87	
13	2,069.79	
12	1,923.62	Median: (12. + 13.) / 2 = 1,996.71
11	1,819.22	
10	1,773.34	
9	1,597.54	
8	1,589.48	
7	1,494.65	
6	1,423.74	Q1: (6. + 7.) / 2 = 1,459.2
5	1,391.92	
4	1,334.88	
3	1,184.53	
2	1,125.78	
1	1,005.85	Minimum / Lower Bound

Spatial Aggregation

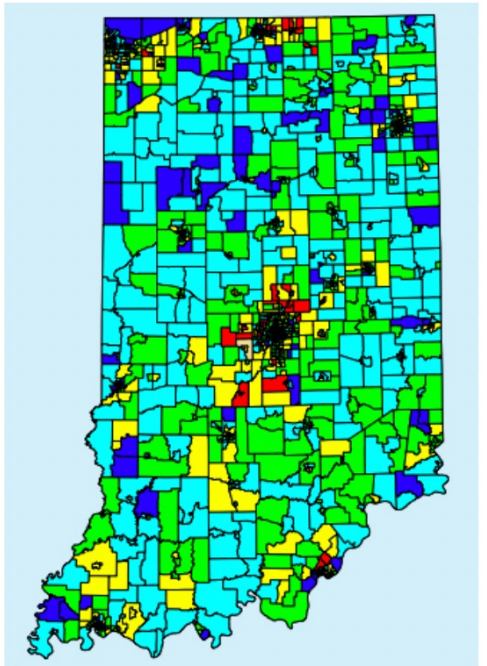
- Analysis of data using aggregated units
- Selecting the correct aggregate units of analysis is critical

Modifiable areal unit problem (MAUP): Boundary definition can dramatically change data analysis

Maps of household population in US state Indiana



By County



By Census Tract

Not as uniform as county map implies!