

# G53FIV: Fundamentals of Information Visualization

## Lecture 5: Multivariate Data Visualization

Ke Zhou  
School of Computer Science  
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

# Announcement

- Change of Time on the Optional Lab Sessions from next week
- **Monday, 9:00 - 10:00, A32 Computer Science**
- Feb 18: Course Work Case Study (Optional)
  - one session at 9:00 - 10:00 at CS-A32.
  - one session at 14:00 - 15:00 as well at Business South, A25.
- Feb 25, Mar 4, Mar 11
  - Lab Computing Sessions (Optional) in CS-A32

# Announcement

- Issue of Course Work
- Please check Moodle for details
- Due date: **April 8 2019**
  - report of maximum 10 pages (3000 words).
  - R codes

# Last Lecture

## Data and Image

# Data Models

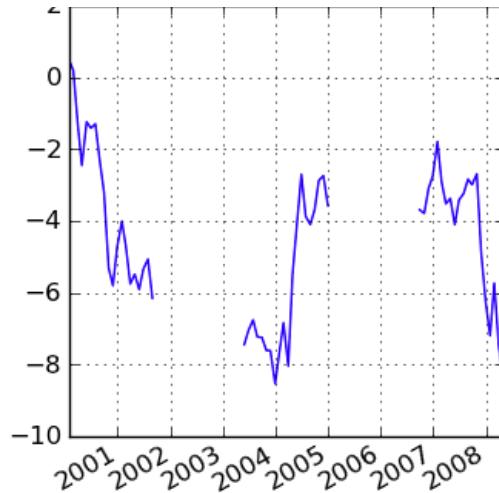
- N, O, Q?
- Dimension or Measure?

– Year	Q-Internal (O)	Dimension
– Age	Q-Ratio (O)	Depends
– Marital	N	Dimension
– Sex	N	Dimension
– People	Q-Ratio	Measure

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080

# Data Processing

- Data cleaning and filtering
  - for quality control
  - Remove (Outlier, missing data)
  - Modify (conversion of format, etc.)
- Data adjustment
  - Depends on your task and questions to ask
  - Relational algebra:
    - e.g. Aggregation, mean, sort, projection
  - Reformatting and Integration

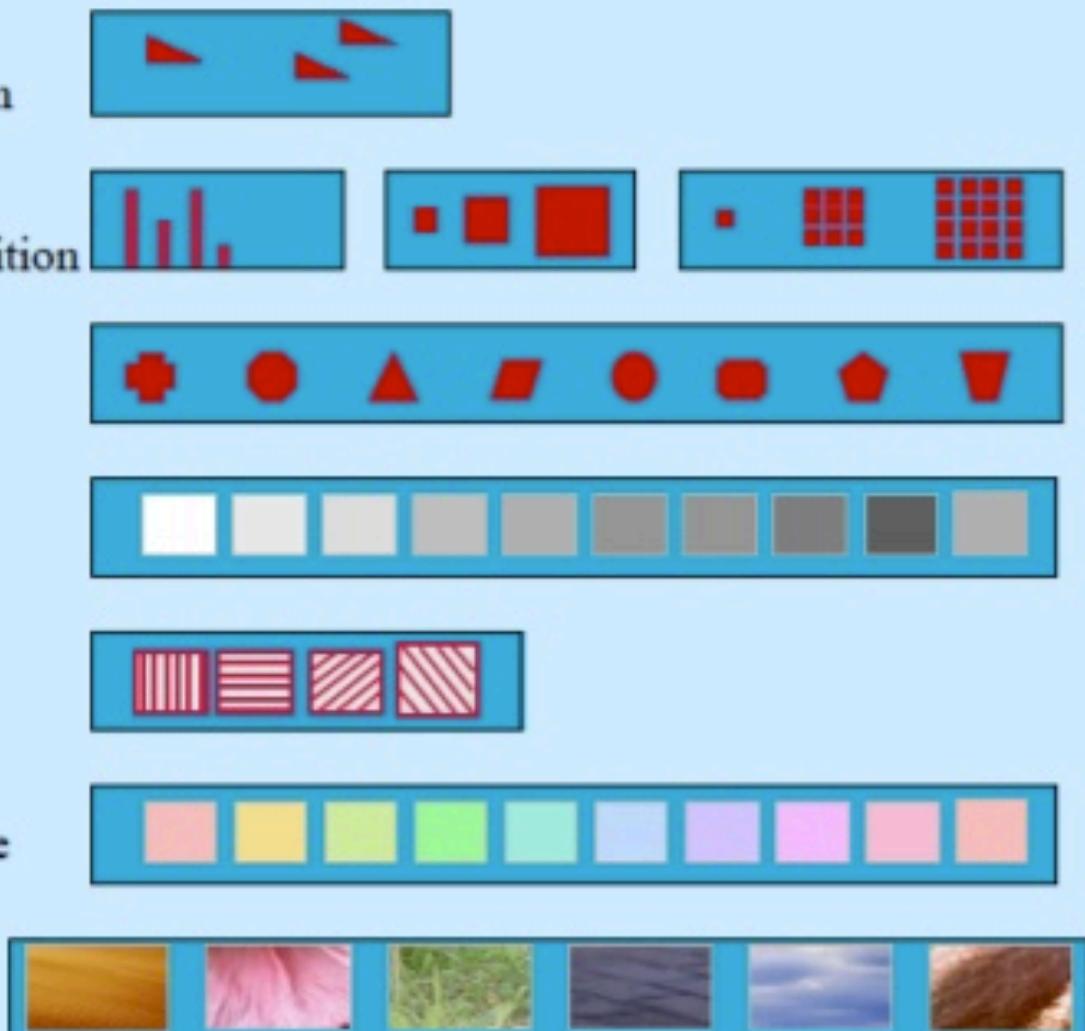


*We will learn later how  
to do these in R.*

# Image: Visual Encoding Variables

## Bertin's Semiology of Graphics (1967)

- **position**
  - changes in the x, y, (z) location
- **size**
  - change in length, area or repetition
- **shape**
  - infinite number of shapes
- **value**
  - changes from light to dark
- **orientation**
  - changes in alignment
- **colour**
  - changes in hue at a given value
- **texture**
  - variation in pattern
- **motion**



Graphic by: Sheelagh Carpendale

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

# Levels of Organization

	Nominal	Ordinal	Quantitative
Position	✓	✓	✓
Size	✓	✓	~
(Grey)Value	✓	✓	~
Texture	✓	~	✗
Color	✓	✗	✗
Orientation	✓	✗	✗
Shape	✓	✗	✗

✓ = Good

~ = OK

✗ = Bad

# Last Lecture

## Design and Graphs

# What Design Criteria to Follow?

- **Expressiveness**

- A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language **express** (1) **all the facts** in the set of data, and (2) **only the facts** in the data.

***Tell the truth***

- **Effectiveness**

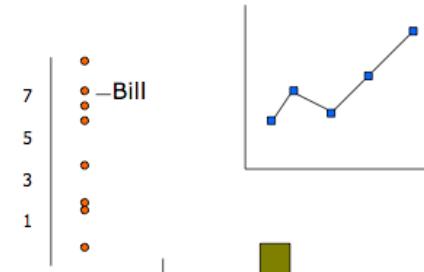
- A visualization is more effective than another visualization if the information conveyed by one visualization **is more readily perceived** than the information in the other visualization.

***Use proper encoding***

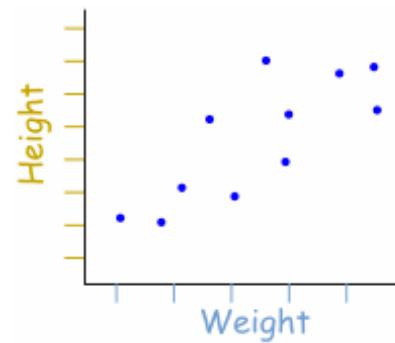
Mackinlay, Automating the design of graphical presentations of relational information, 1986.

# Graphs

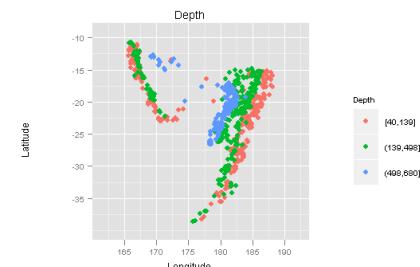
- Data Dimensions
  - 1 - Univariate data
  - 2 - Bivariate data
  - 3 - Trivariate data
  - >3 - Hypervariate data
- Data Types
  - Nominal, Ordinal, Quantitative
- Visualization Representations
  - Points, Lines, Bars, Boxes



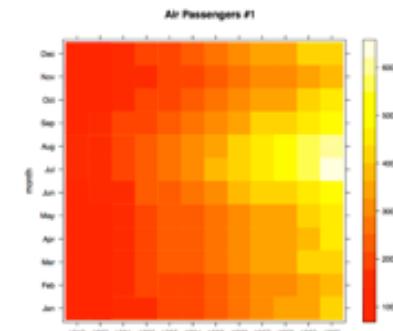
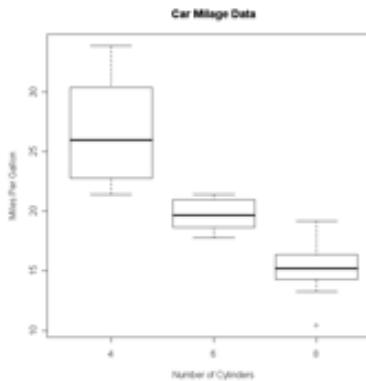
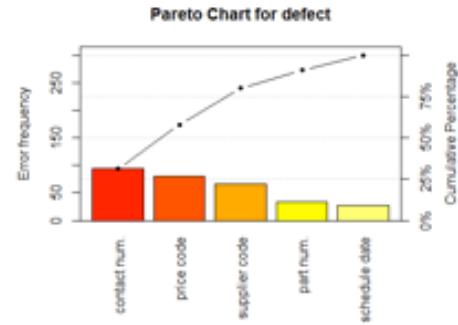
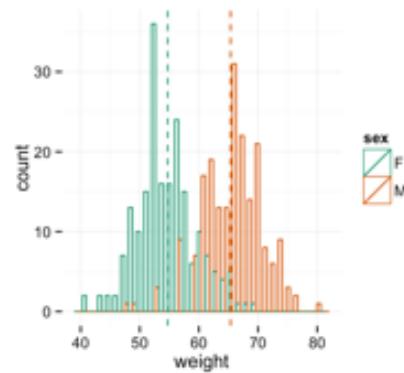
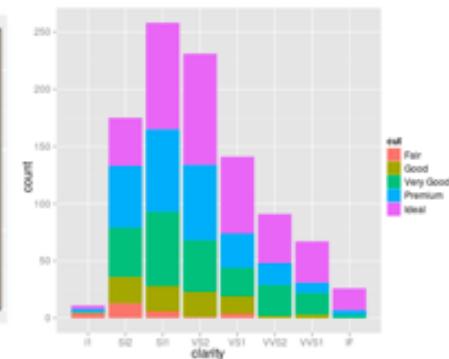
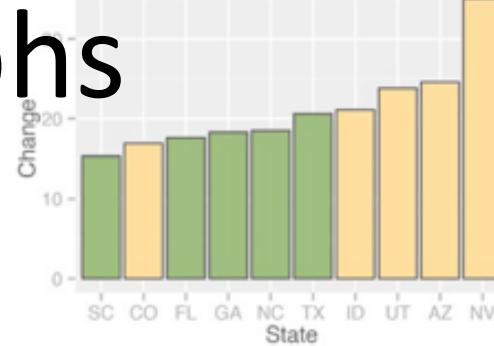
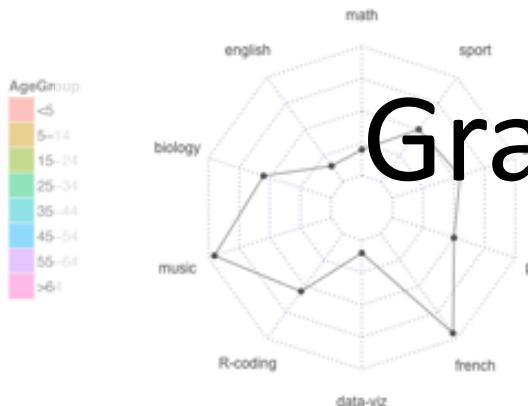
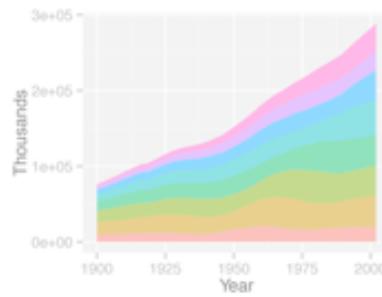
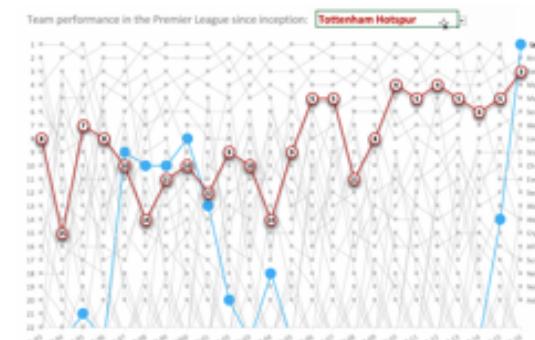
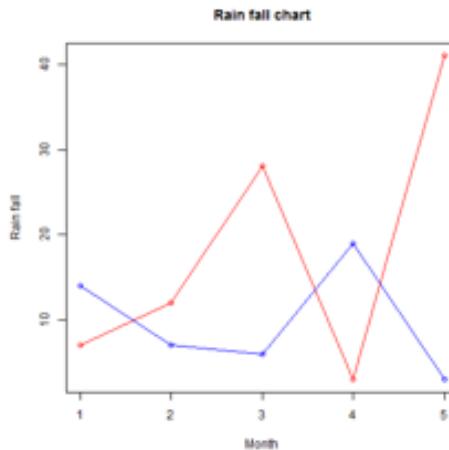
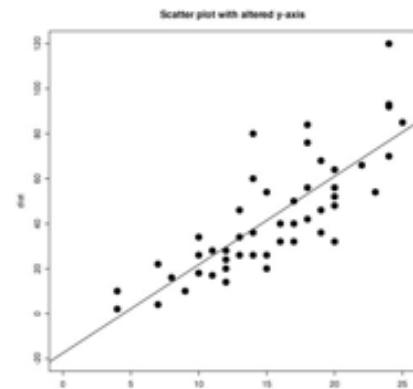
Univariate



Bivariate



Trivariate



# Overview

- Multivariate Data Visualization Design Challenge
  - hypervariate data – our focus
- Common Multivariate Data Visualization Techniques

# Design Challenge

- Data about dogs (hypervariate data)
  - Variety N
  - Group N
  - Size O
  - Smartness N
  - Popularity Q
  - Ranking Q
- Design a visualization for this multivariate data

# Intuition

- Fundamentally, we have 2 geometric (position) display dimensions
- For data sets with >2 variables, we must project data down to 2D
- Come up with visual mapping that locates each dimension into 2D plane

# Representation

- What are two main ways of presenting multivariate data sets?
  - Directly (textually): Tables
  - Symbolically (pictures): Graphs
- When use which?

# Table / Spreedsheet

- A spreadsheet (table) already does that
  - Each variable is positioned into a column
  - Data cases in rows
  - This is a projection (mapping)

Name	Economy	Cylinders	Displacement	Horsepower
Mazda RX4	21	6	160	110
Mazda RX4 Wag	21	6	160	110
Datsun 710	22.8	4	108	93
Hornet 4 Drive	21.4	6	258	110
Hornet Sportabout	18.7	8	360	175
Valiant	18.1	6	225	105
Duster 360	14.3	8	360	245
Merc 2400	24.4	4	146.7	62
Merc 230	22.8	4	140.8	95
Merc 280	19.2	6	167.6	123
Merc 280C	17.8	6	167.6	123
Merc 450SE	16.4	8	275.8	180
Merc 450SL	17.3	8	275.8	180
Merc 450SLC	15.2	8	275.8	180
Cadillac Fleetwood	10.4	8	472	205
	---	-	---	---

# Limitations

- Occupy large space
- Difficult to understand the relationships
- Hard to see the overall picture, focus and see the context
- Less effective in amplifying human perception and cognition

# When to use?

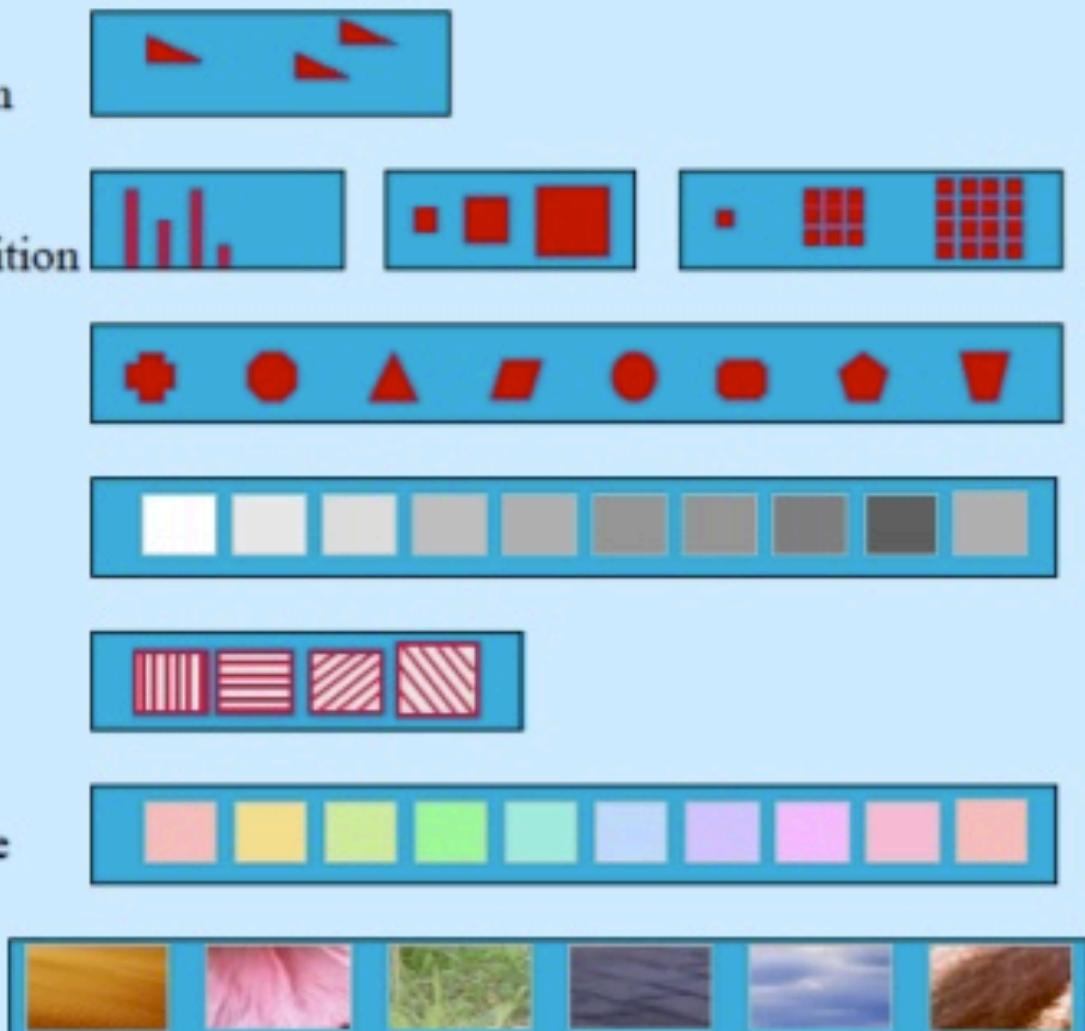
- Use tables when
  - The document will be used to **look up individual values**
  - The document will be used to **compare individual values**
  - **Precise values** are required
  - The quantitative info to be communicated involves **more than one unit of measure**
- Use graphs when
  - The message is contained in the **shape** of the values
  - The document will be used to **reveal relationships** among values
  - Especially useful when **the number of data points is huge**

(Optional Reading) Stephen Few. 2012. Show Me the Numbers: Designing Tables and Graphs to Enlighten (2nd ed.). Analytics Press, , USA.

# Image: Visual Encoding Variables

## Bertin's Semiology of Graphics (1967)

- **position**
  - changes in the x, y, (z) location
- **size**
  - change in length, area or repetition
- **shape**
  - infinite number of shapes
- **value**
  - changes from light to dark
- **orientation**
  - changes in alignment
- **colour**
  - changes in hue at a given value
- **texture**
  - variation in pattern
- **motion**

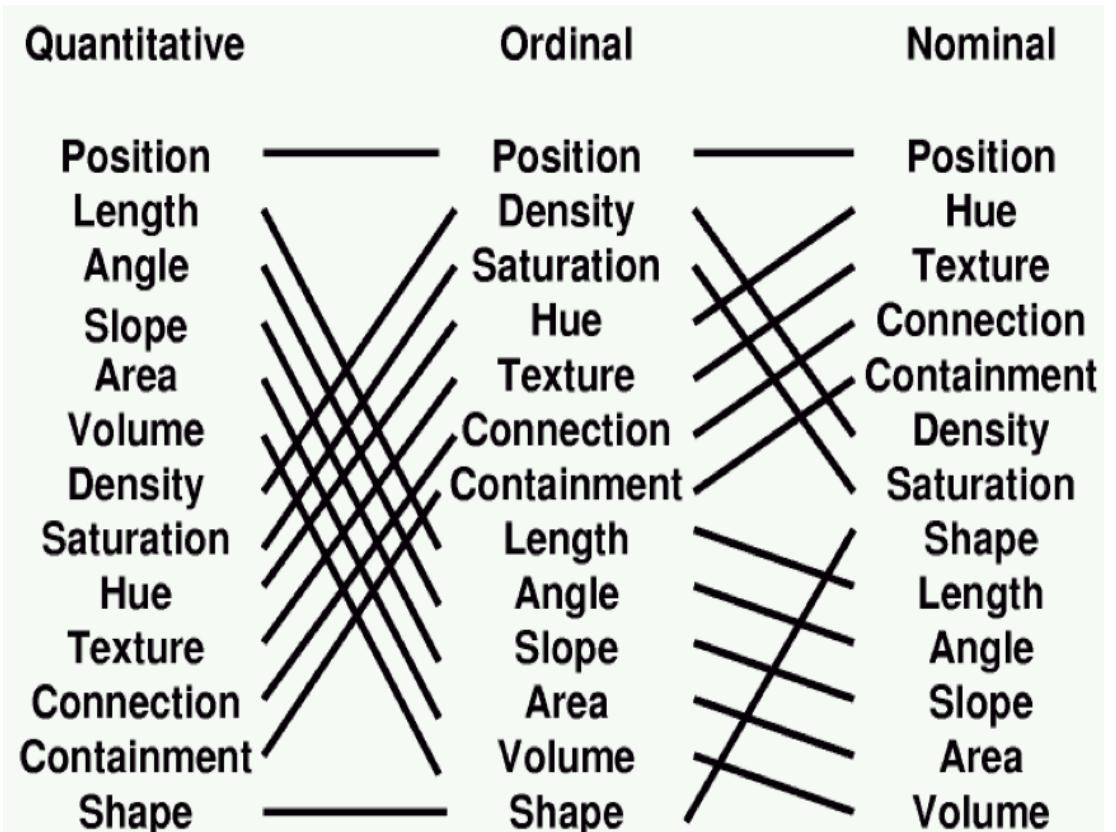


Graphic by: Sheelagh Carpendale

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

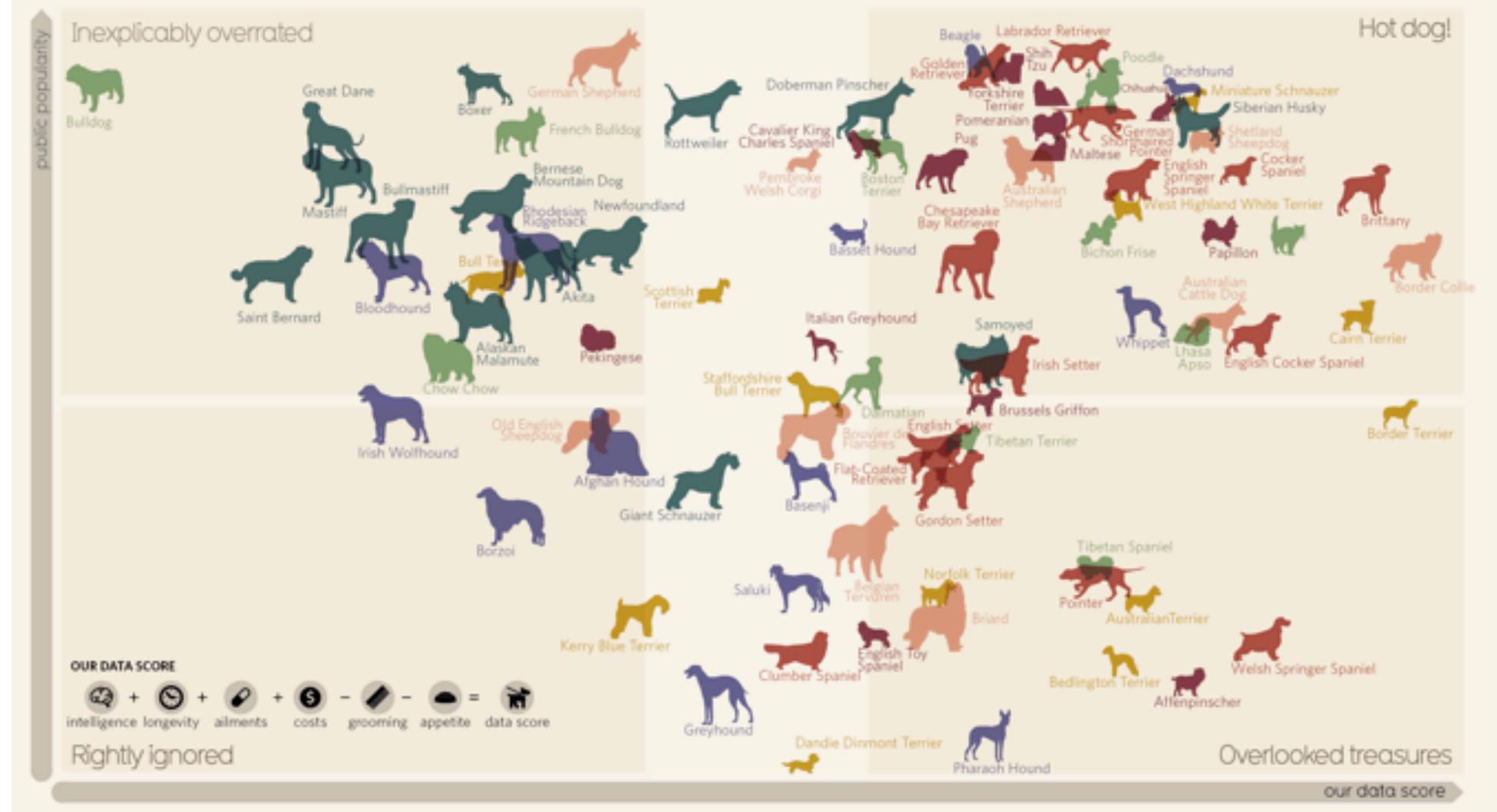
# Design Challenge

- Data about dogs (hypervariate data)
  - Variety N
  - Group N
  - Size O
  - Smartness N
  - Popularity Q
  - Ranking Q
- Design a visualization



# Best in Show

The ultimate data-dog



- Iconic Representations: Glyph (graphical object) represents a data case
- Visual properties of glyph represent different variables

# Multivariate Data Visualization

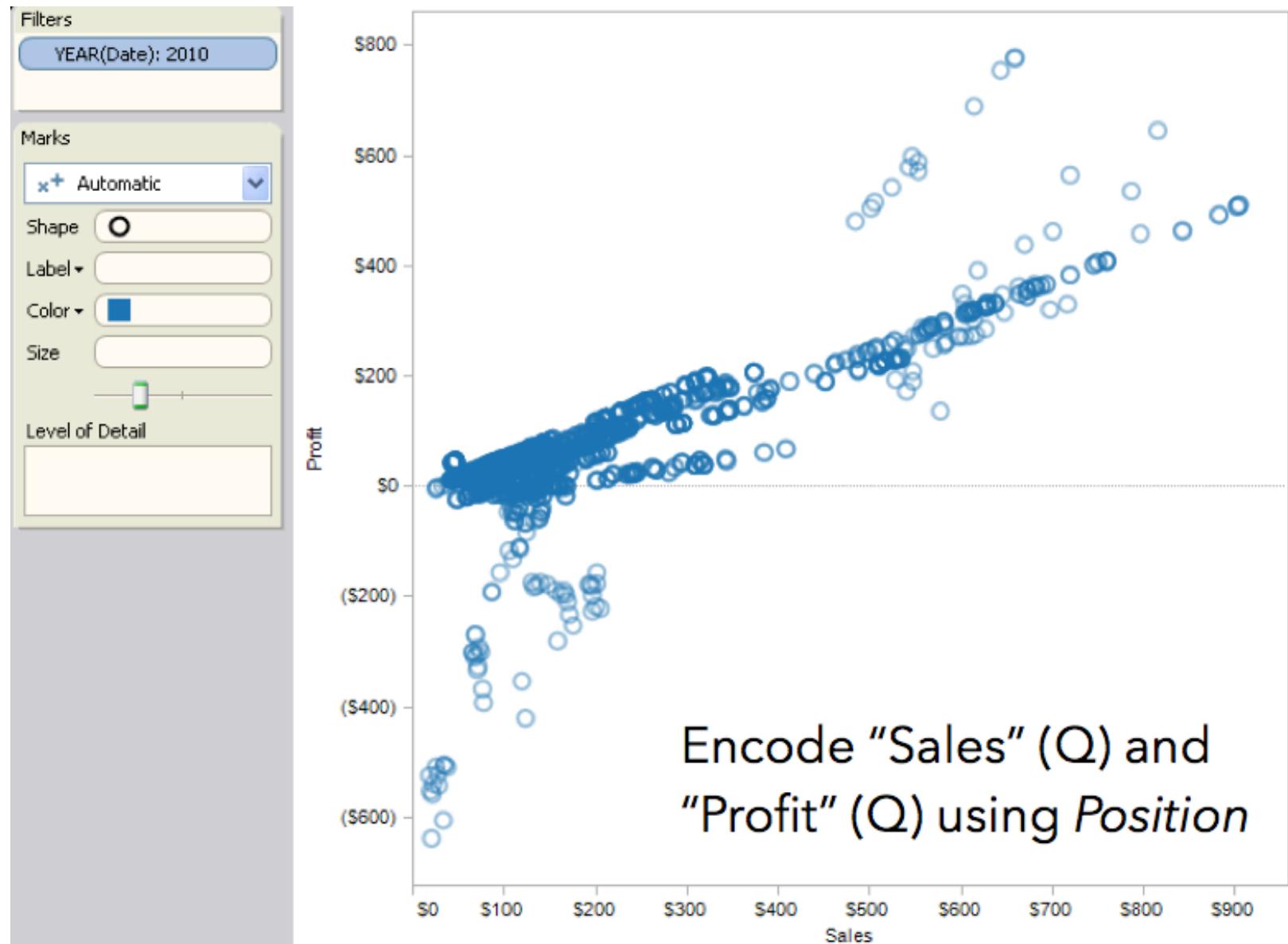
- Visual Encodings: 8 dimensions?
- Focus: techniques can generally handle all data sets

Visual Variables	Characteristics				
	Selective	Associative	Quantitative	Order	Length
<i>Position</i>	• .	••• .••	↑ . . . . .	↑ . . . . .	Theoretically Infinite
<i>Size</i>	• ●	•●●●		● > ● > ● > ●	Selection: ~5 Distinction: ~20
<i>Shape</i>					Theoretically Infinite
<i>Value</i>	○ ● ○ ○ ○ ○	○ ○ ○ ● ○ ○		○ < ○ < ○ < ○ < ○ < ●	Selection: <7 Distinction: ~10
<i>Color</i>	• .	● ● ● ● ● ●			Selection: <7 Distinction: ~10
<i>Orientation</i>	\\   /				Theoretically Infinite
<i>Texture</i>	○○○○	○○○○			Theoretically Infinite

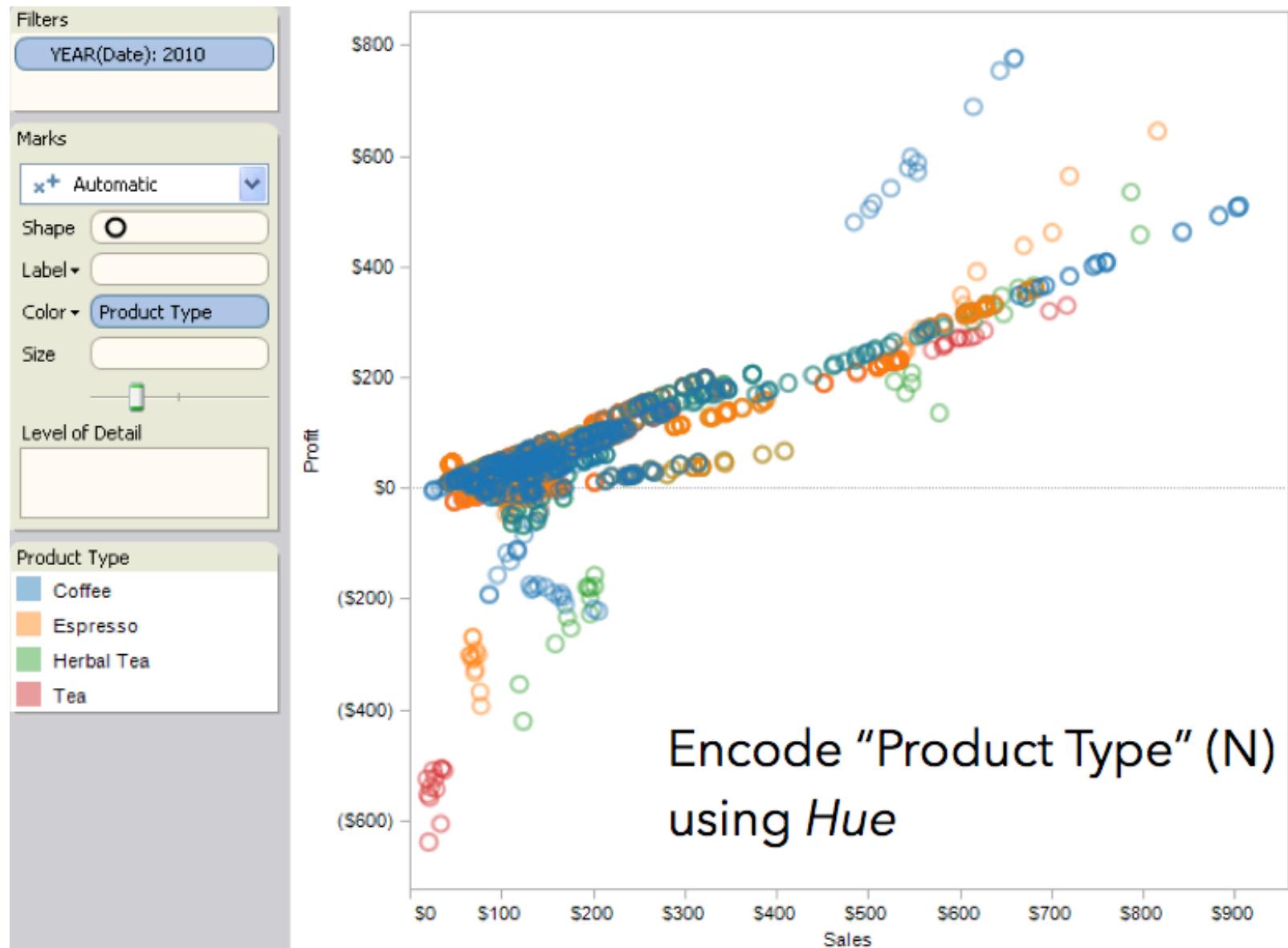
# An Example: Coffee Sales

- Sales: Q-Ratio
- Profit: Q-Ratio
- Marketing: Q-Ratio
- Product Type: N {Coffee, Espresso, Herbal Tea, Tea}
- Market: N {Central, East, South, West}

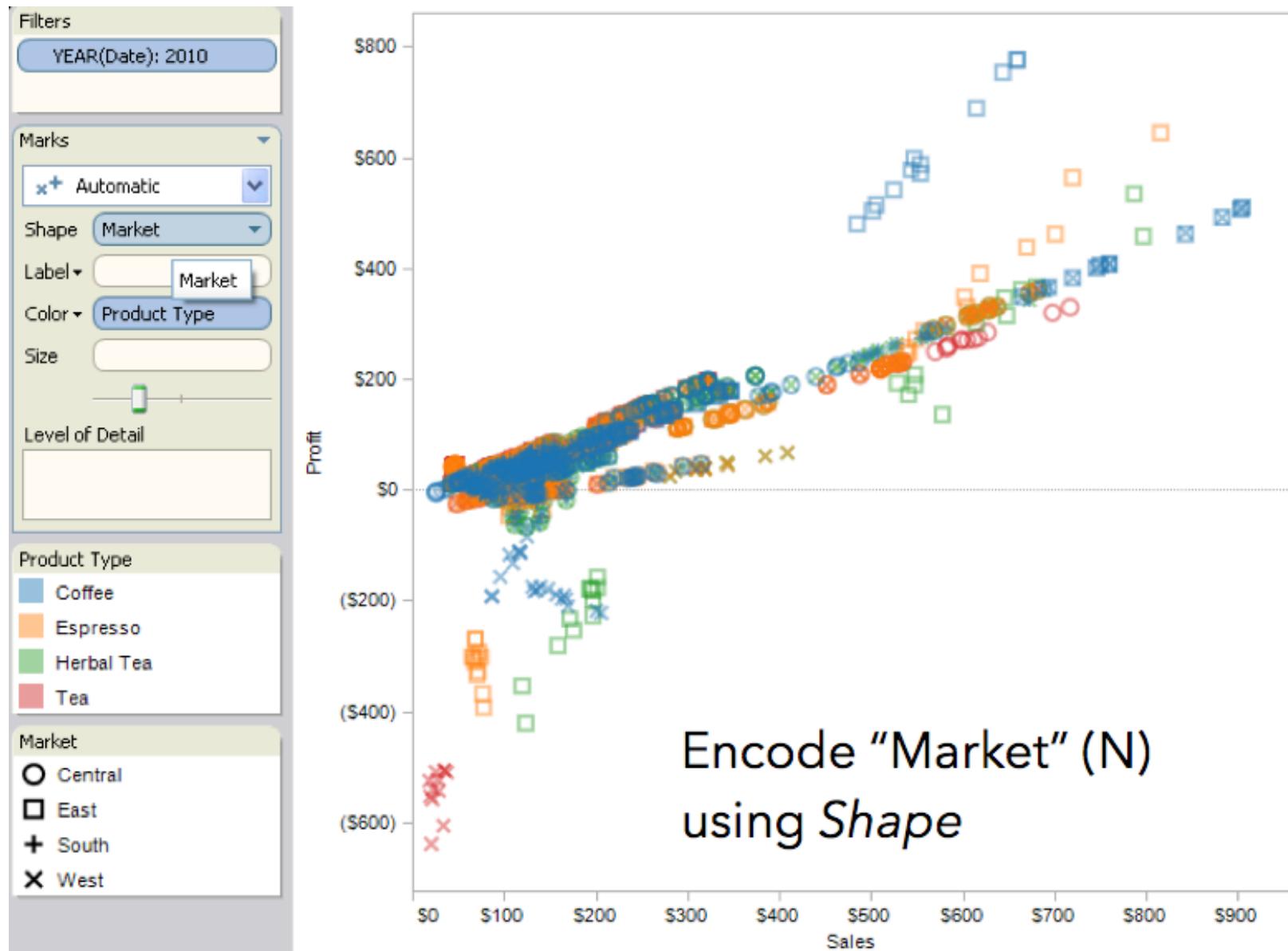
# First Two Variables



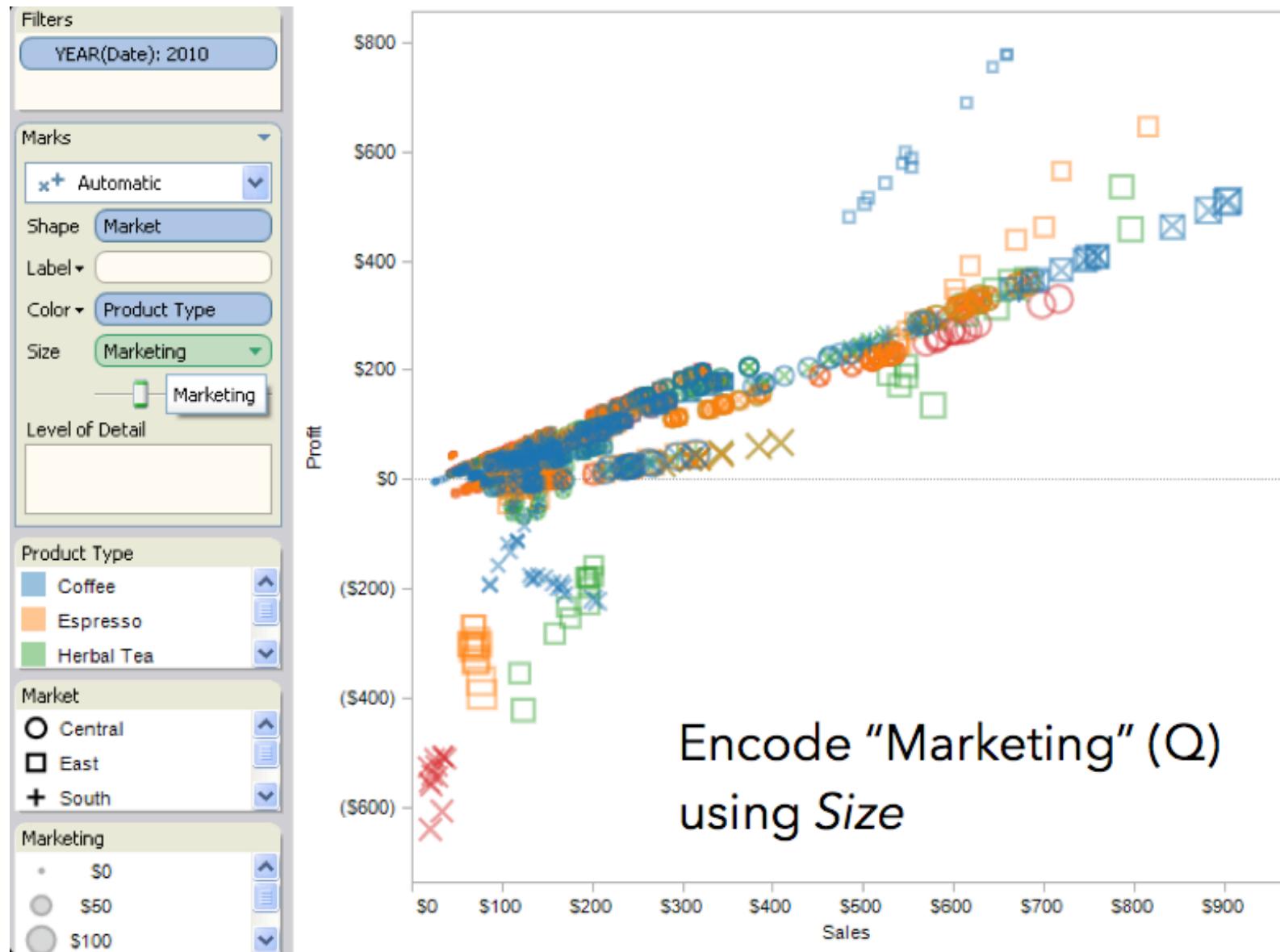
# Third Variable



# Fourth Variable



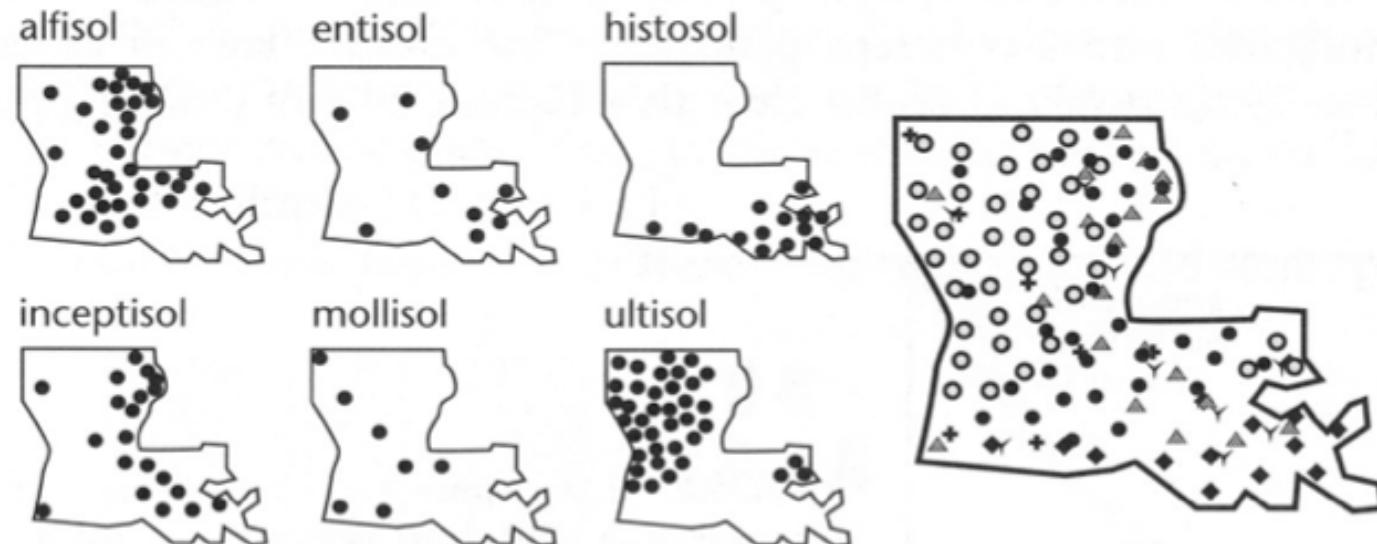
# Fifth Variable



# Small Multiples

“At the heart of quantitative reasoning is a single question: Compared to what? Small multiple designs, multivariate and data bountiful, answer directly by visually enforcing comparisons of changes, of the differences among objects, of the scope of alternatives. For a wide range of problems in data presentation, small multiples are the best design solution.”

Tufte, Envisioning Information

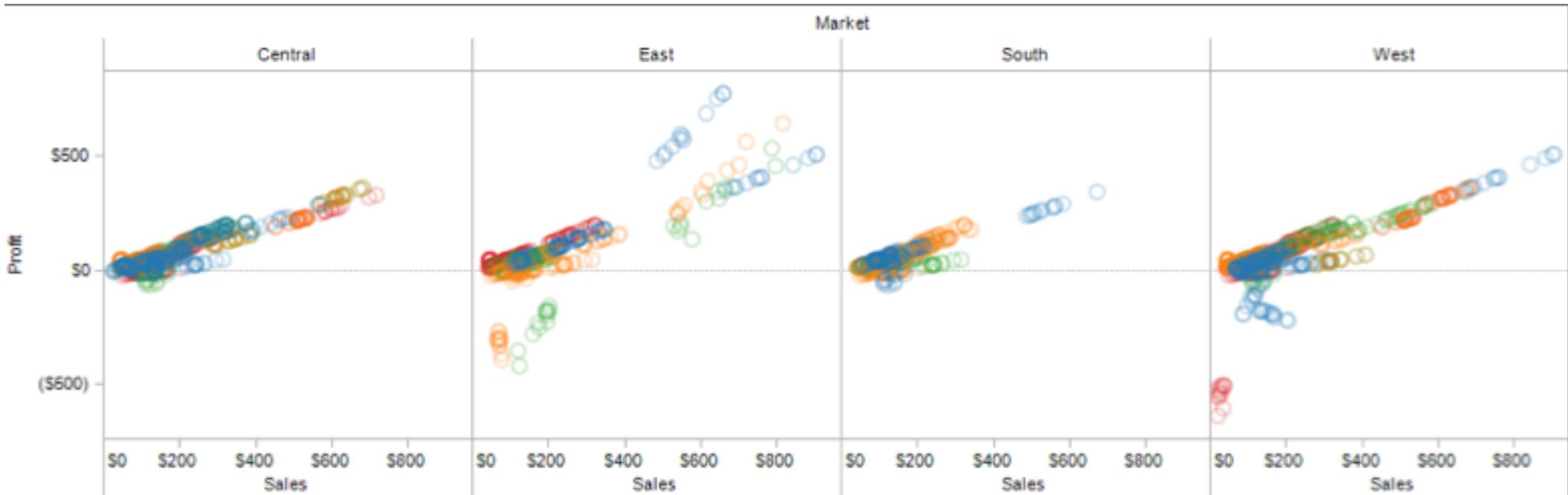


In The Visual Display of Quantitative Information (Textbook, Chapter 8)

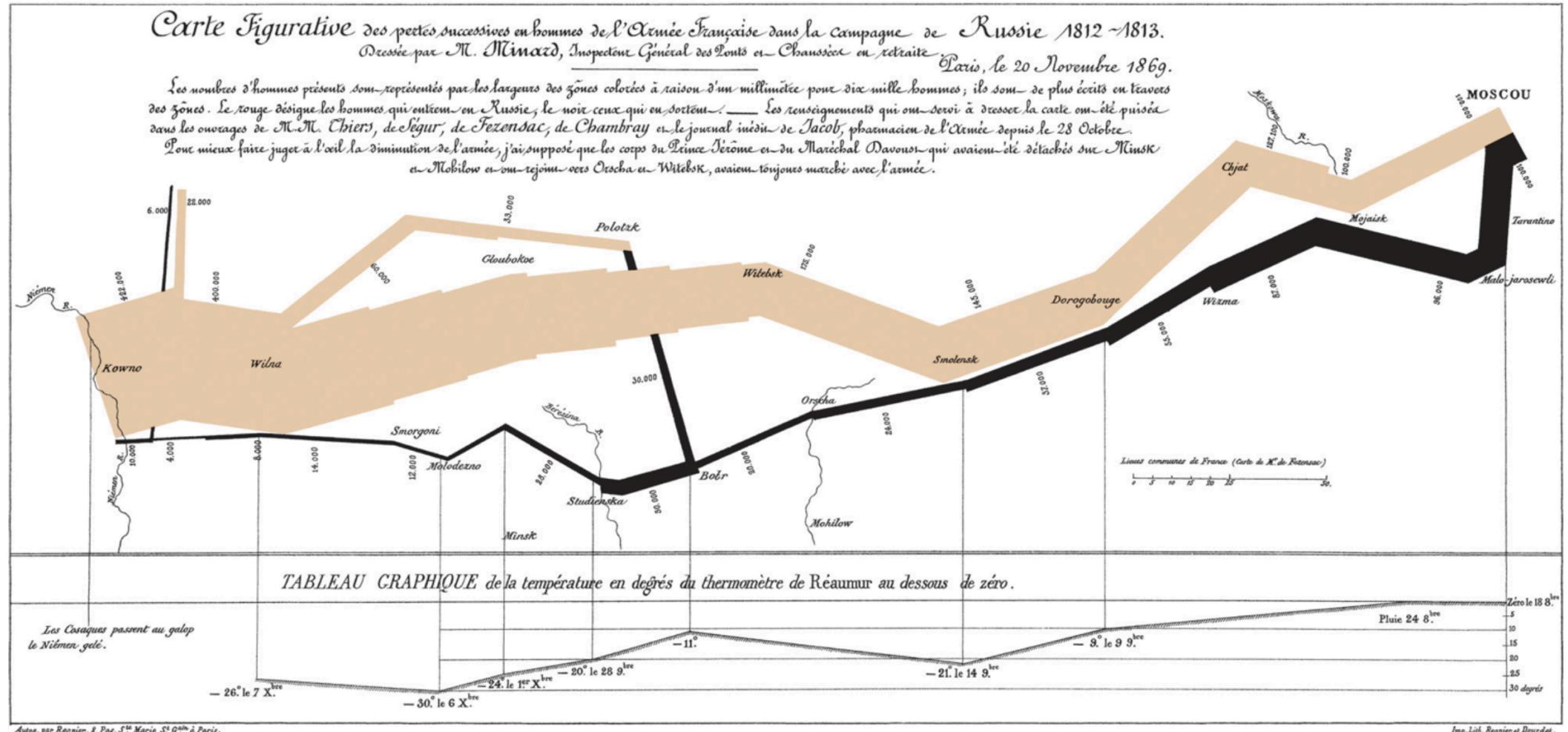
Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

# Trellis Display (Small Multiples)

- It subdivides space to enable comparison across multiple plots.
- Typically nominal or ordinal variables are used as dimensions for subdivision.

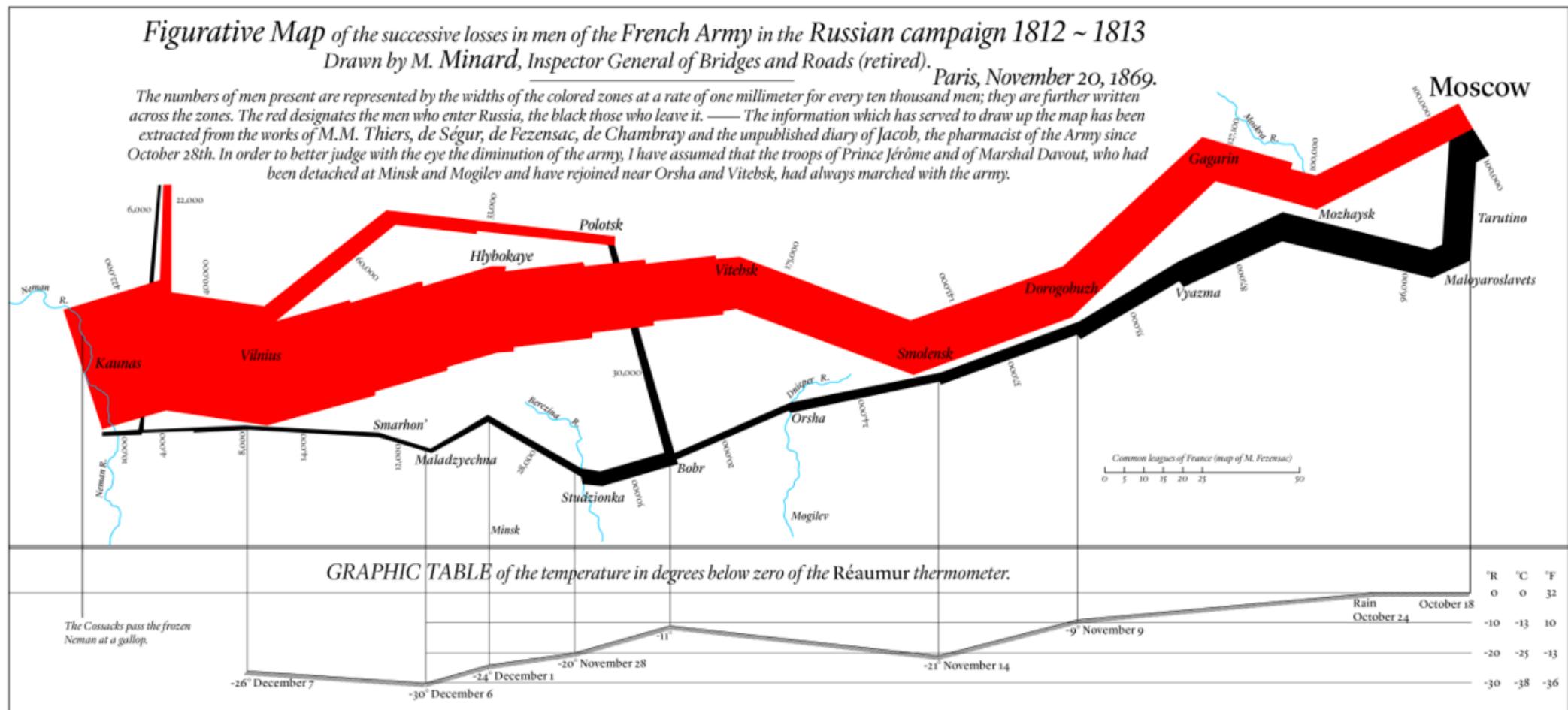


# Minard 1869: Napoleon's March



In The Visual Display of Quantitative Information (Textbook, Chapter 1)

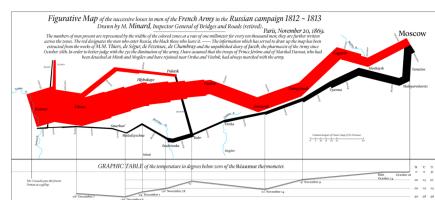
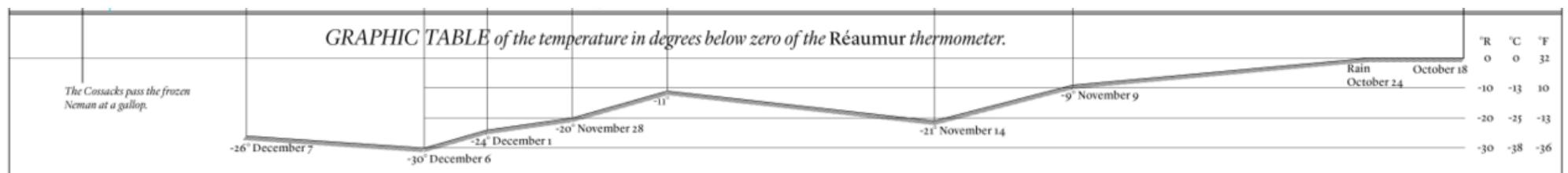
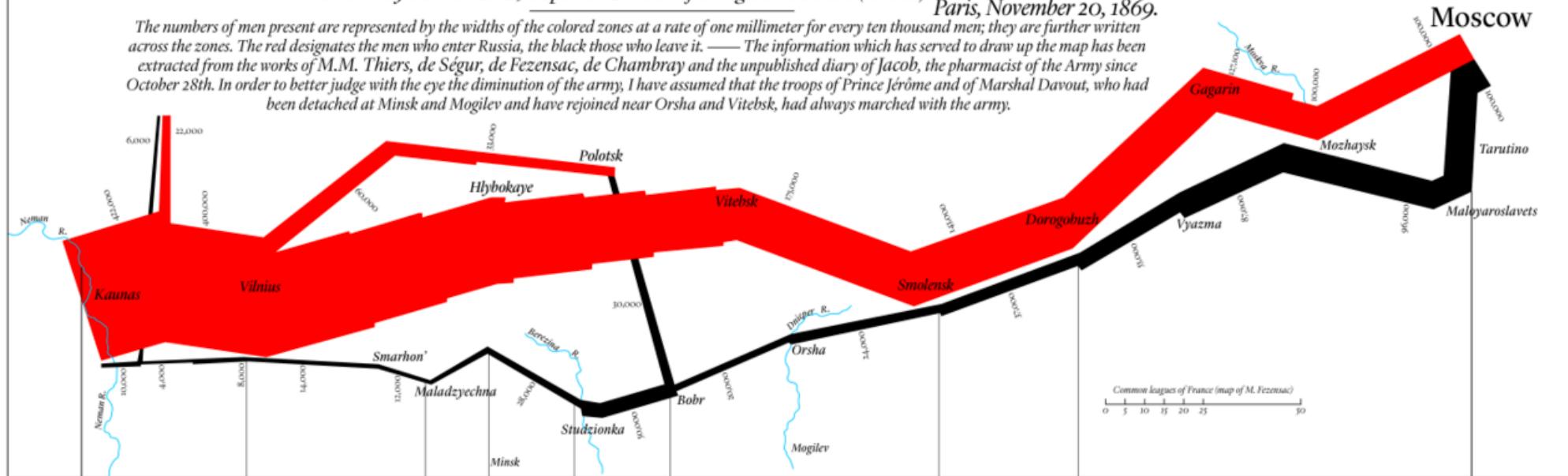
# Minard 1869: Napoleon's March



# Decomposition

*Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~ 1813  
 Drawn by M. Minard, Inspector General of Bridges and Roads (retired).*

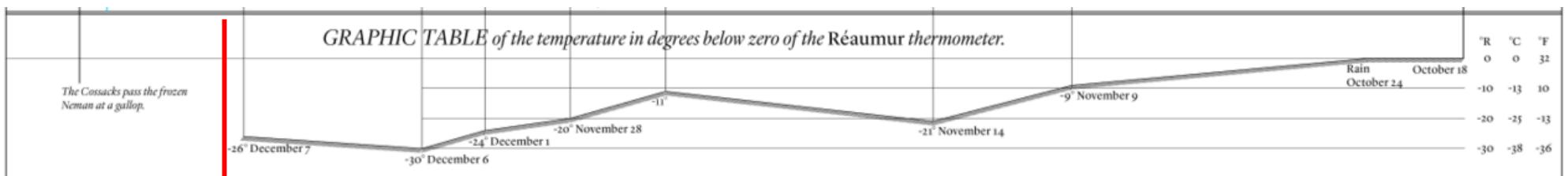
Paris, November 20, 1869.  
 The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red designates the men who enter Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M.M. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davout, who had been detached at Minsk and Mogilev and have rejoined near Orsha and Vitebsk, had always marched with the army.



# Decomposition

Y-axis: temperature (Q)

X-axis: longitude (Q) / time (O)



Temperature over space/time (Q x Q)

# Decomposition

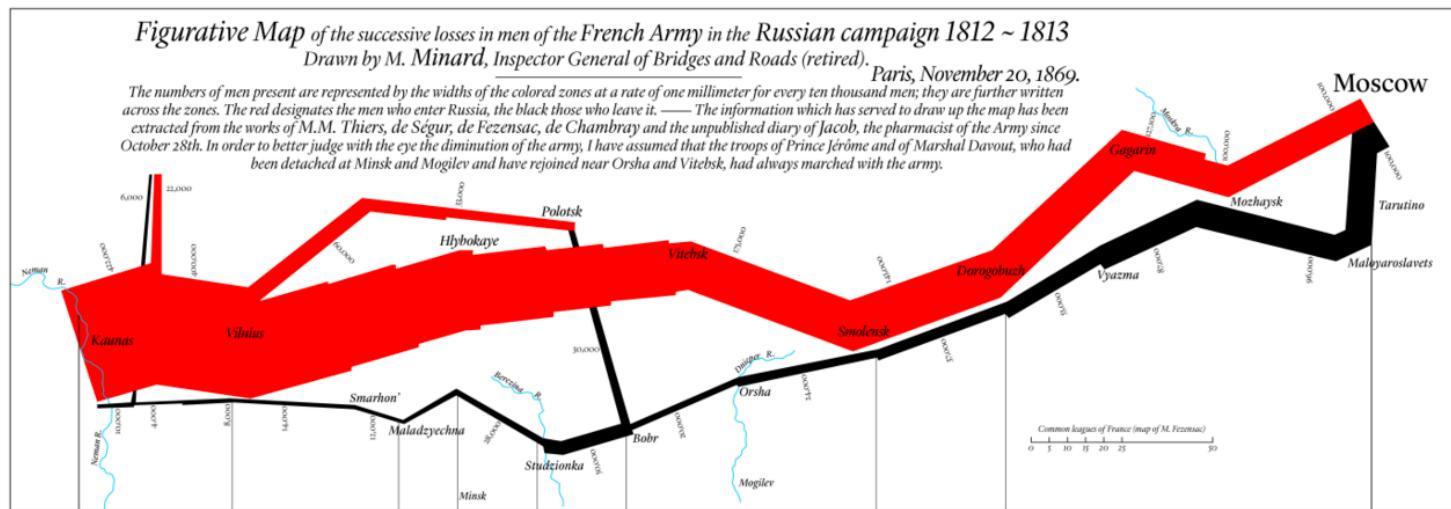
Y-axis: longitude (Q)



X-axis: latitude (Q)

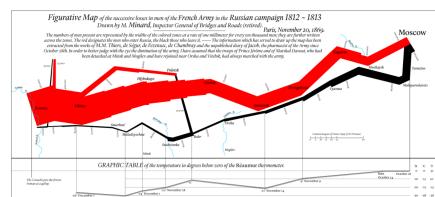
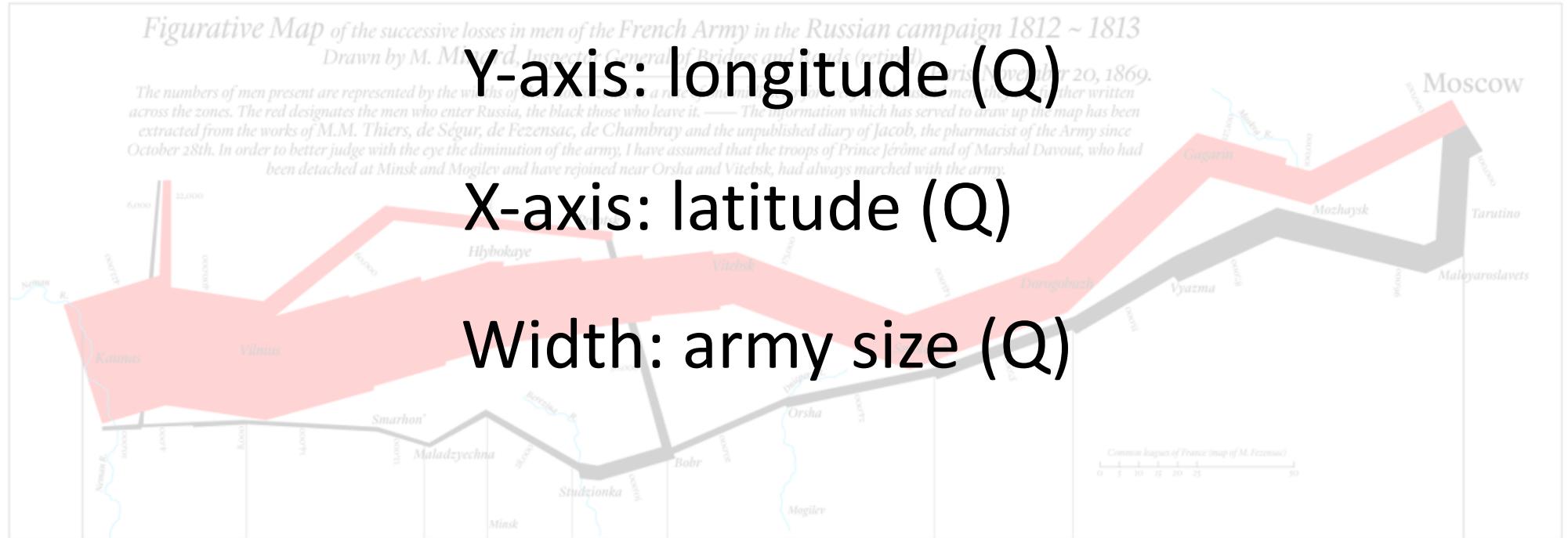


Width: army size (Q)



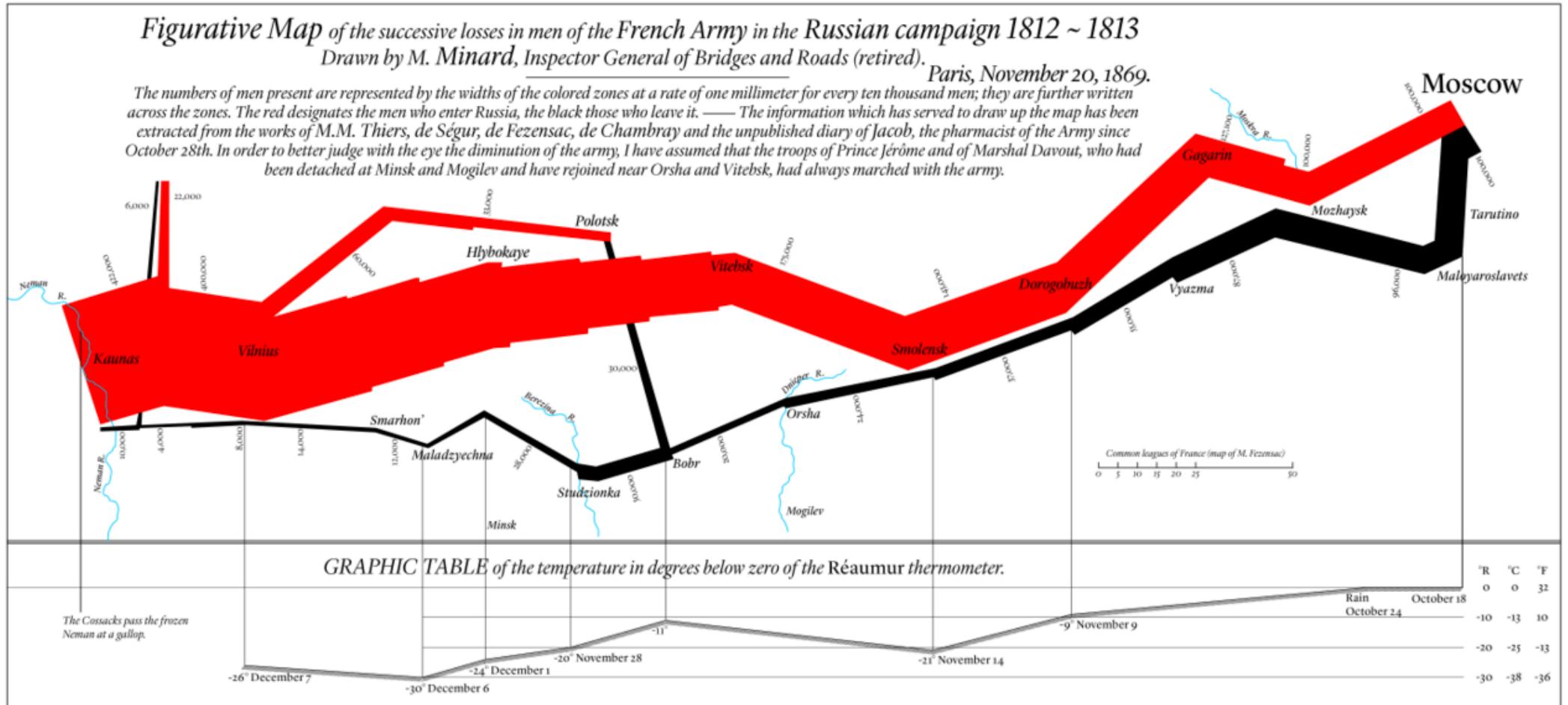
Army position ( $Q \times Q$ ) and army size (Q)

# Minard 1869: Napoleon's March



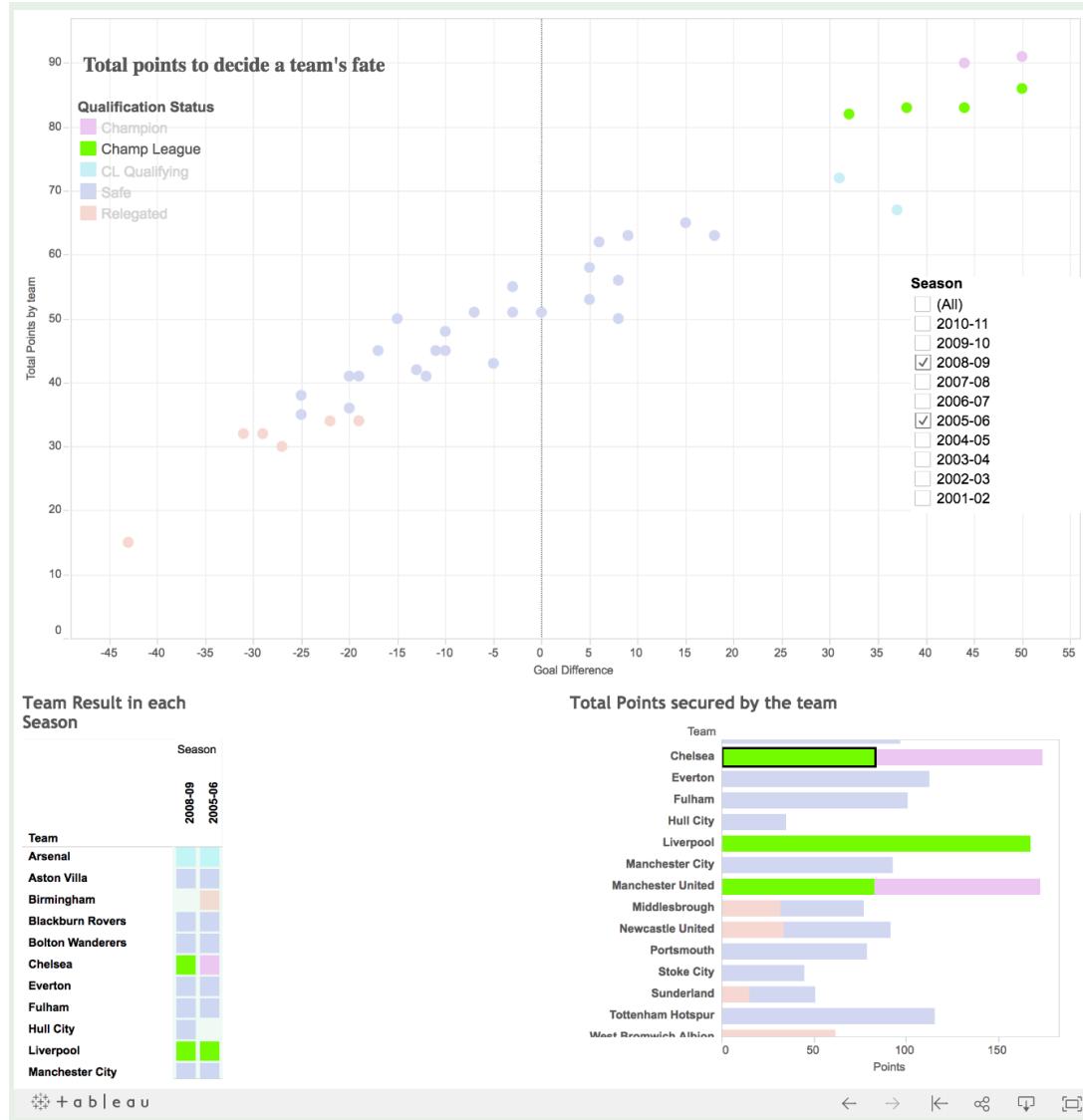
Depicts at least 5 quantitative variables.

# Minard 1869: Napoleon's March



Depicts at least 5 quantitative variables. Any others?

# Multiple Coordinated Views



# Multivariate Data Visualization

- Strategies:
  - Avoid “over-encoding”
  - Use space and small multiples intelligently
  - Reduce the problem space
  - Use interaction to generate relevant views
- Rarely does a single visualization answer all questions. Instead, the ability to generate appropriate visualizations quickly is key.

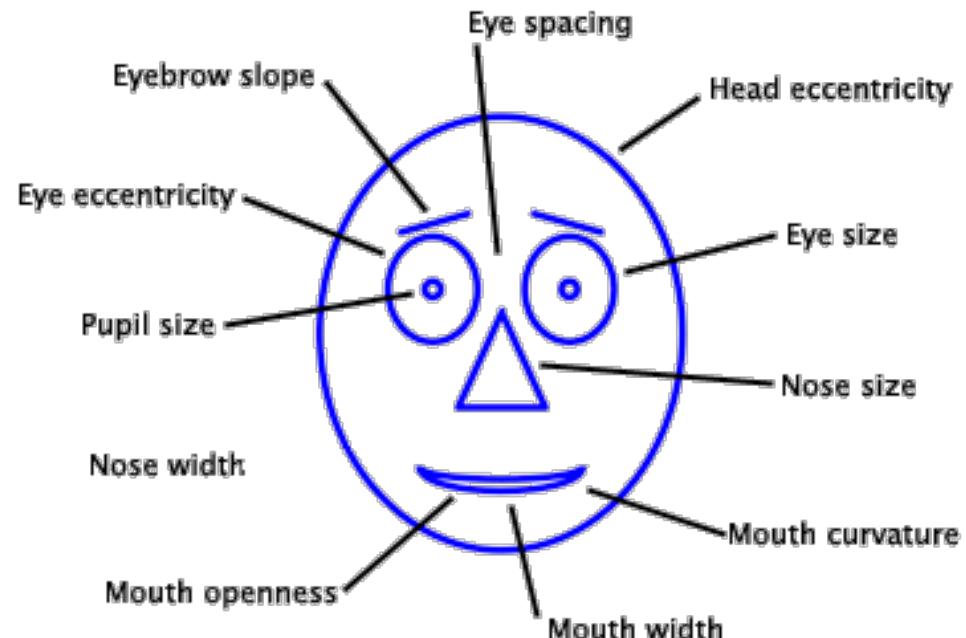
# Common Multivariate Data Visualization Techniques

- Chernoff Faces
- Table Lens
- Parallel Coordinates
- Mosaic Plot

# Chernoff Faces

# Chernoff Faces

- Observation: We have evolved a sophisticated ability to interpret faces.
- Idea: Encode different variables' values in characteristics of human face



In The Visual Display of Quantitative Information (Textbook, Chapter 7)

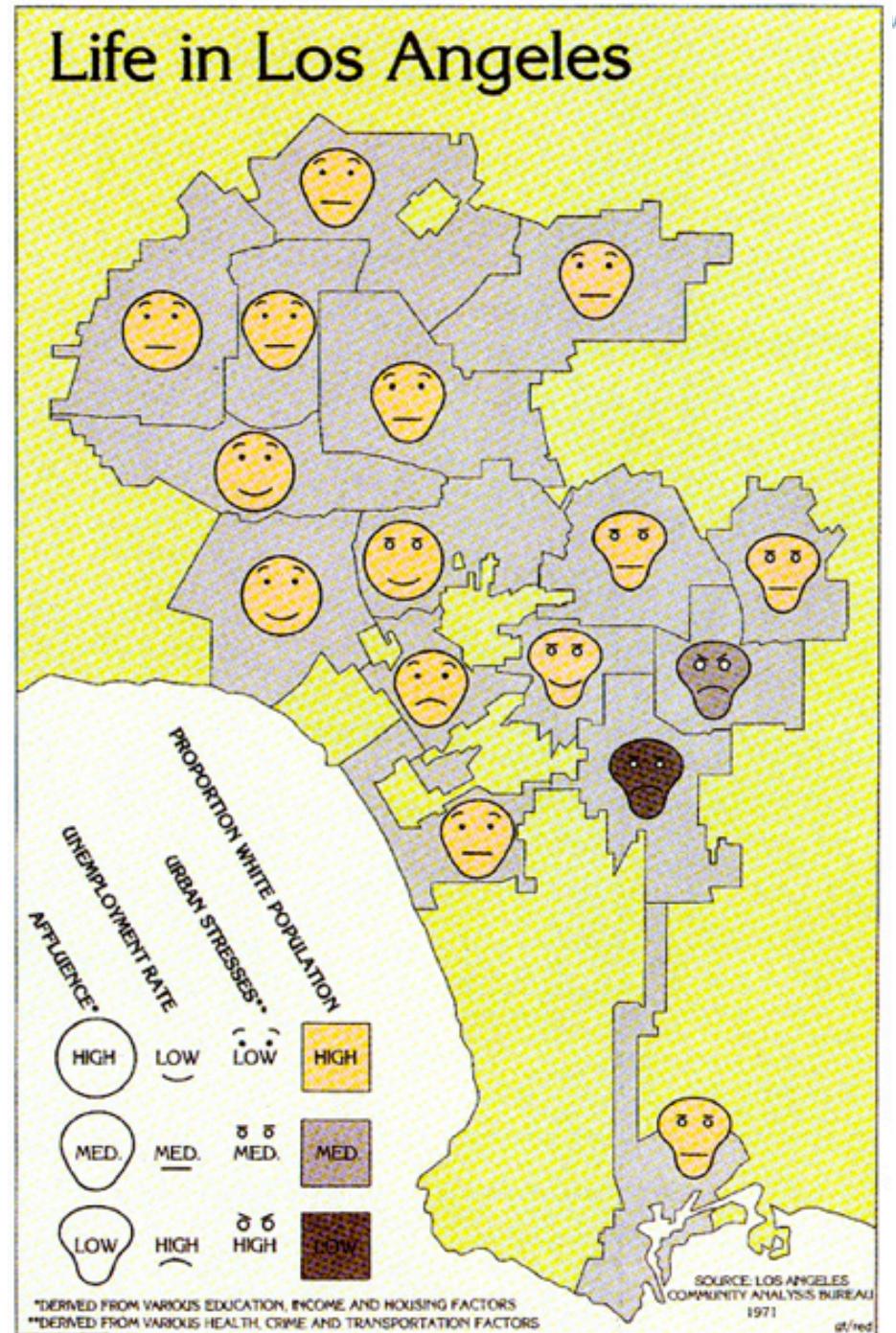
(Optional Reading) Chernoff, Herman. "The use of faces to represent points in k-dimensional space graphically." Journal of the American Statistical Association 68.342 (1973): 361-368.

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

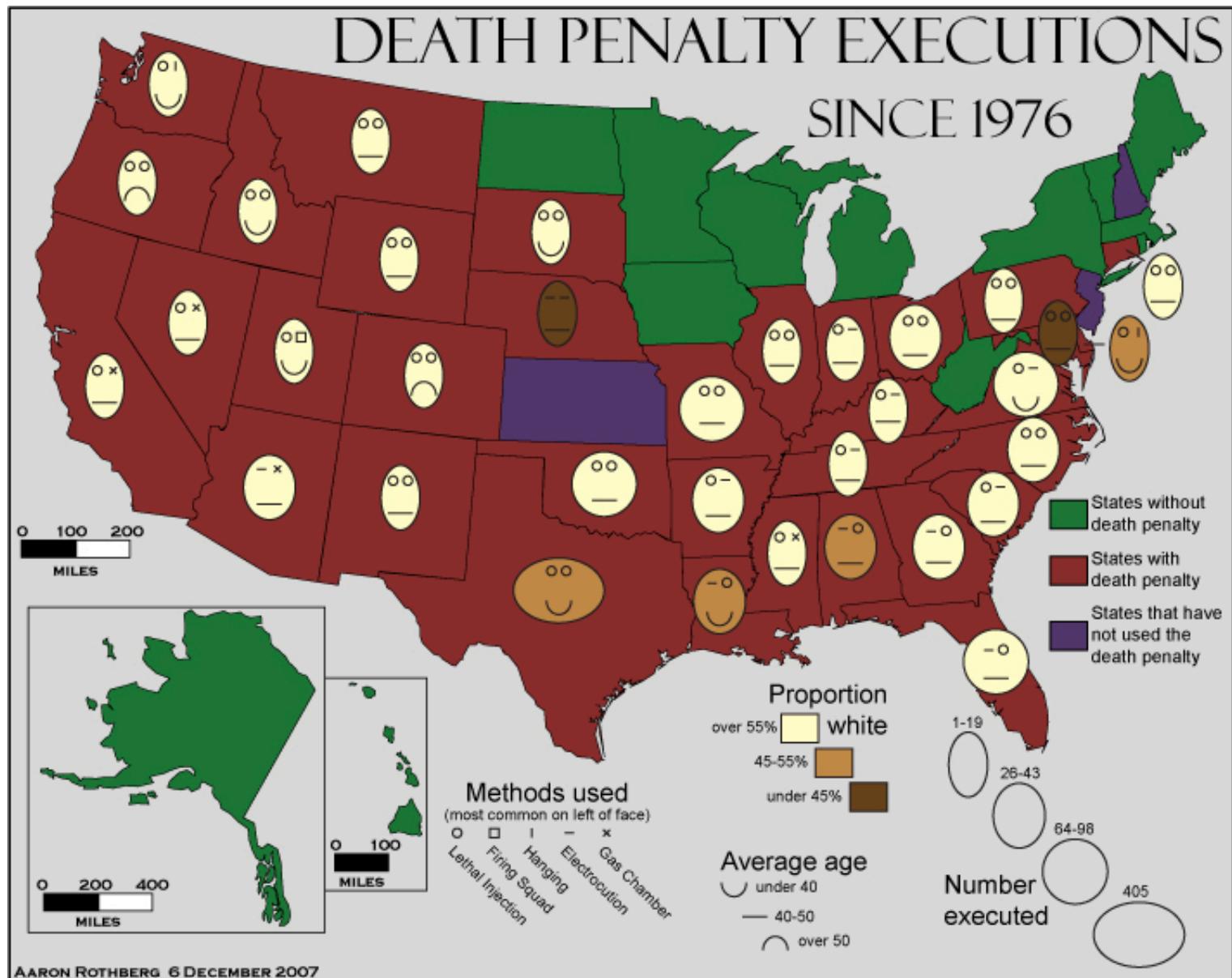
# Example

“It is probably one of the most interesting maps I’ve created because the expressions evoke an emotional association with the data.”

Eugene Turner



# Critiques



# Table Lens

# Table Lens

- Spreadsheet is certainly one hypervariate data presentation
- Idea: Make the text more visual and symbolic
- Just leverage basic bar chart idea

Rao, Ramana, and Stuart K. Card. "The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information." Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 1994.

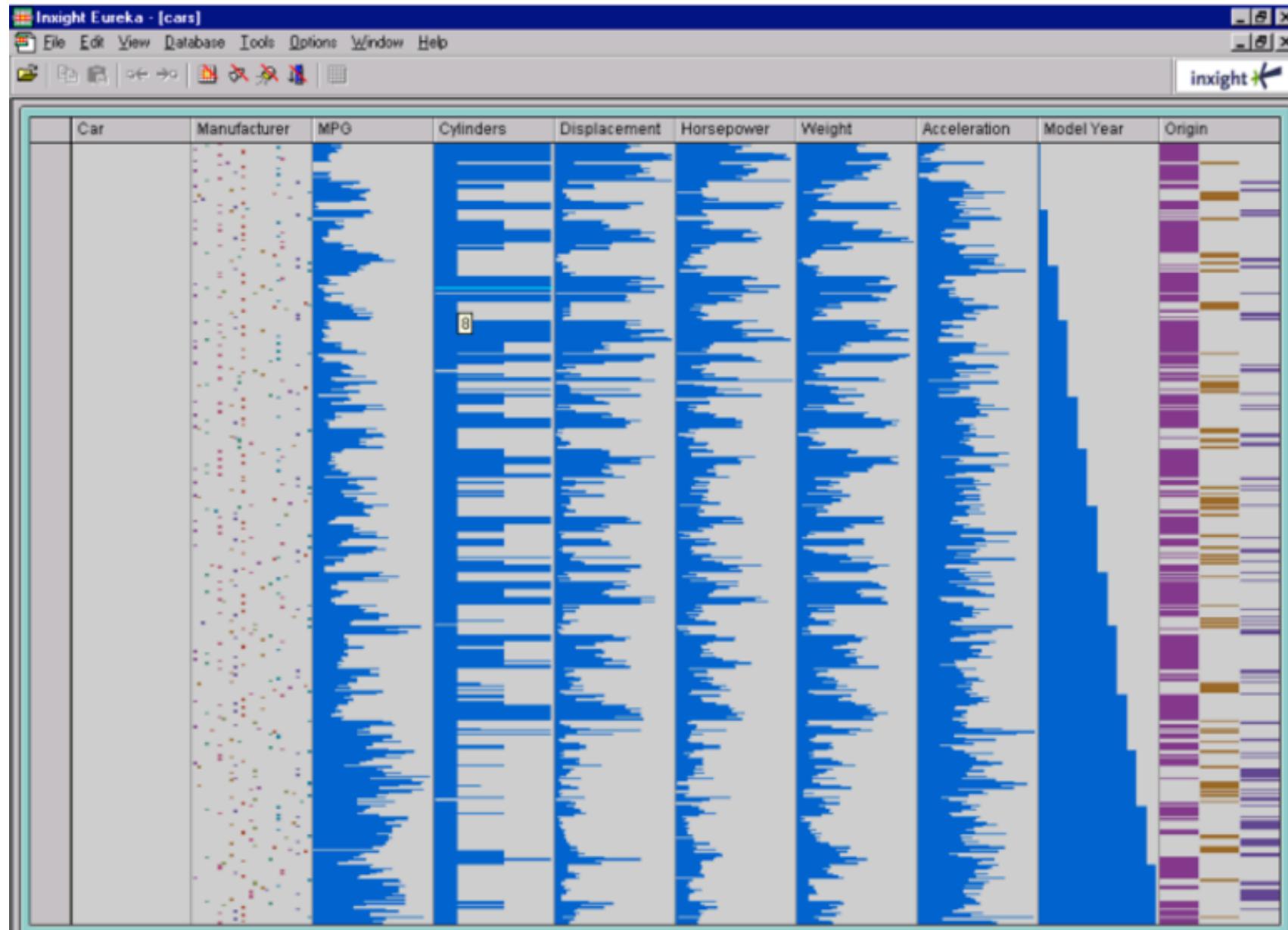
# Visual Mapping

- Basic idea:  
Change  
quantitative  
values to bars
- What do you  
do for nominal  
data?

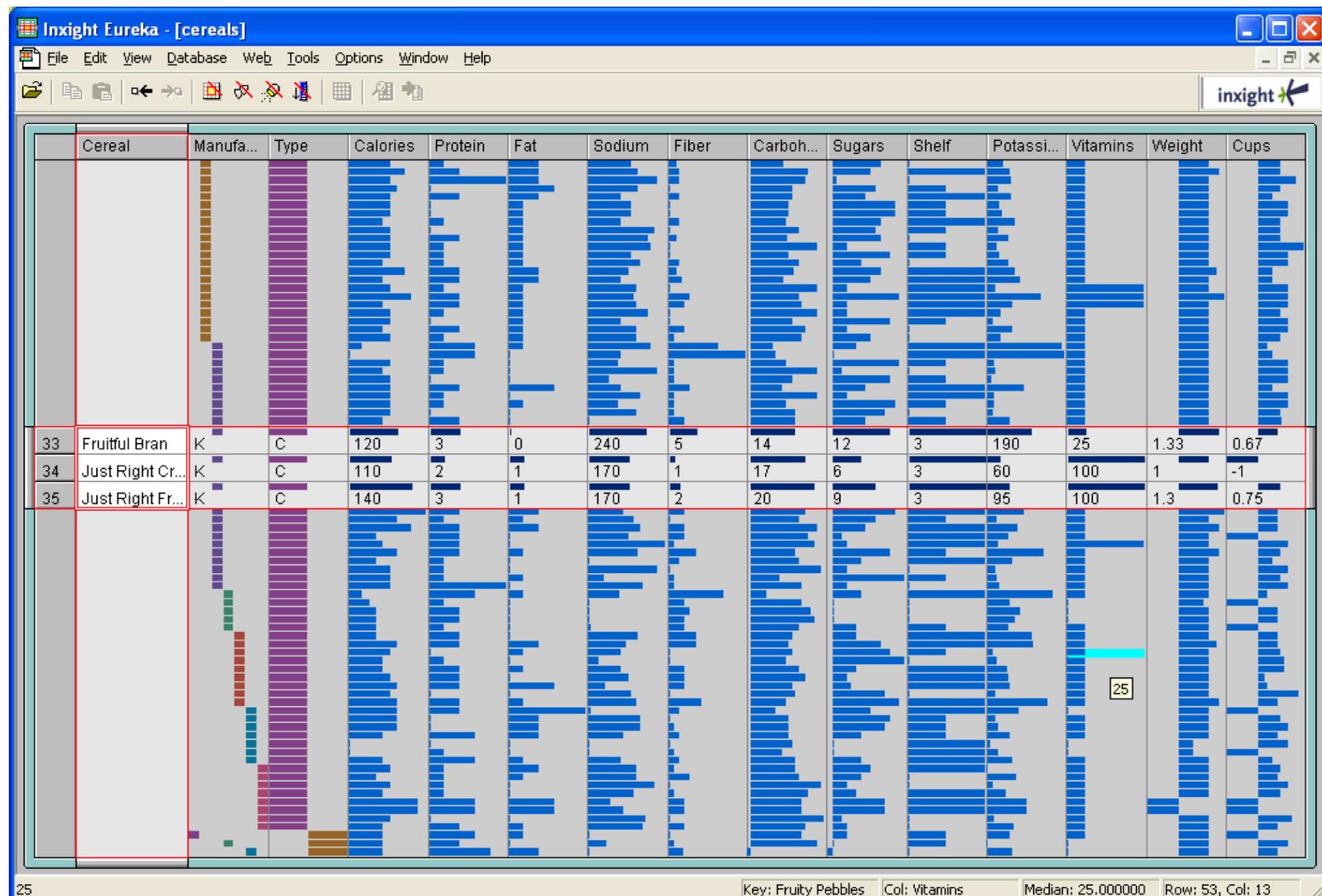
	A	B	C	D	E	F	G	H	I
1	Cereal	Manufactur	Type	Calories	Protein	Fat	Sodium	Fiber	Carbo
2	Frosted Mini-Wheats	K	C	100	3	0	0	0	3
3	Raisin Squares	K	C	90	2	0	0	0	2
4	Shredded Wheat	N	C	80	2	0	0	0	3
5	Shredded Wheat 'n'Bran	N	C	90	3	0	0	0	4
6	Shredded Wheat spoon s	N	C	90	3	0	0	0	3
7	Puffed Rice	Q	C	50	1	0	0	0	0
8	Puffed Wheat	Q	C	50	2	0	0	0	1
9	Maypo	A	H	100	4	1	0	0	0
10	Quaker Oatmeal	Q	H	100	5	2	0	2.7	
11	Strawberry Fruit Wheats	N	C	90	2	0	15	3	
12	100% Natural Bran	Q	C	120	3	5	15	2	
13	Golden Crisp	P	C	100	2	0	45	0	
14	Smacks	K	C	110	2	1	70	1	
15	Great Grains Pecan	P	C	120	3	3	75	3	
16	Cream of Wheat (Quick)	N	H	100	3	0	80	1	
17	Corn Pops	K	C	110	1	0	90	1	
18	Muesli Raisins, Dates, & R	C	C	150	4	3	95	3	
19	Apple Jacks	K	C	110	2	0	125	4	



# Example of Table Lens



# Focus and Context



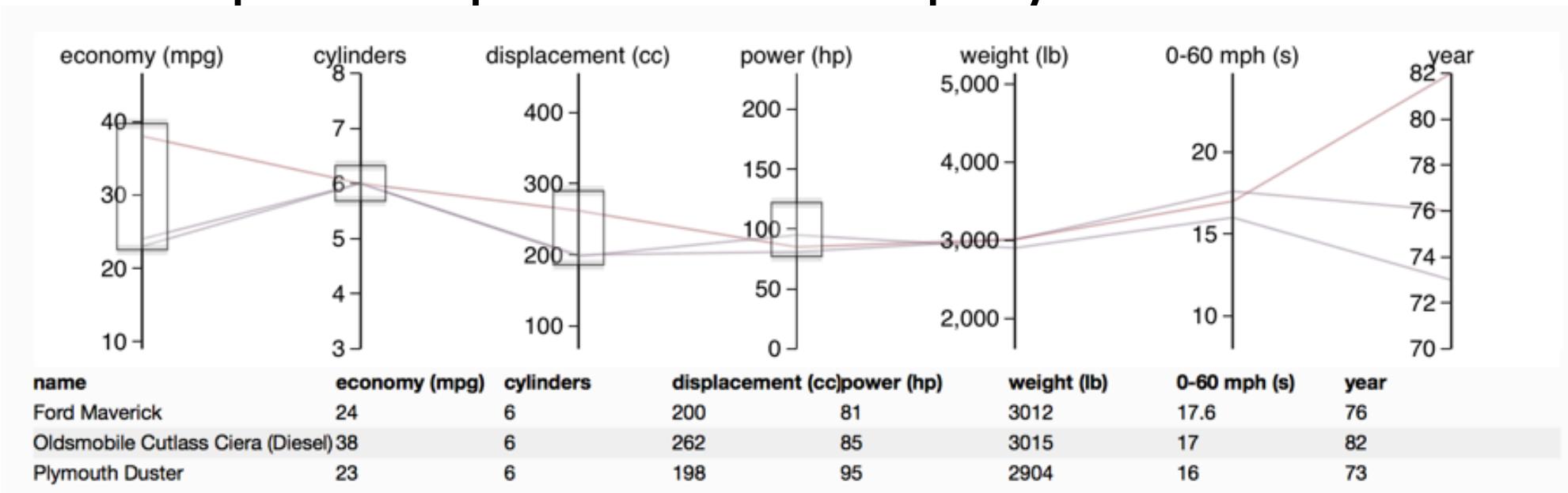
# Video Demo

- [http://www.open-video.org/details.php?  
videoid=8304](http://www.open-video.org/details.php?videoid=8304)
- Space advantage
- Fluid navigation
- Direct exploration

# Parallel Coordinates

# Parallel Coordinates

- Encode variables along a horizontal row
- Vertical line specifies different values that variable can take
- Data point represented as a polyline



# Live Demo

<https://syntagmatic.github.io/parallel-coordinates>

To learn more:

Heinrich, Julian, and Daniel Weiskopf. "State of the Art of Parallel Coordinates." Eurographics (STARs). 2013.

<http://www.parallelcoordinates.de>

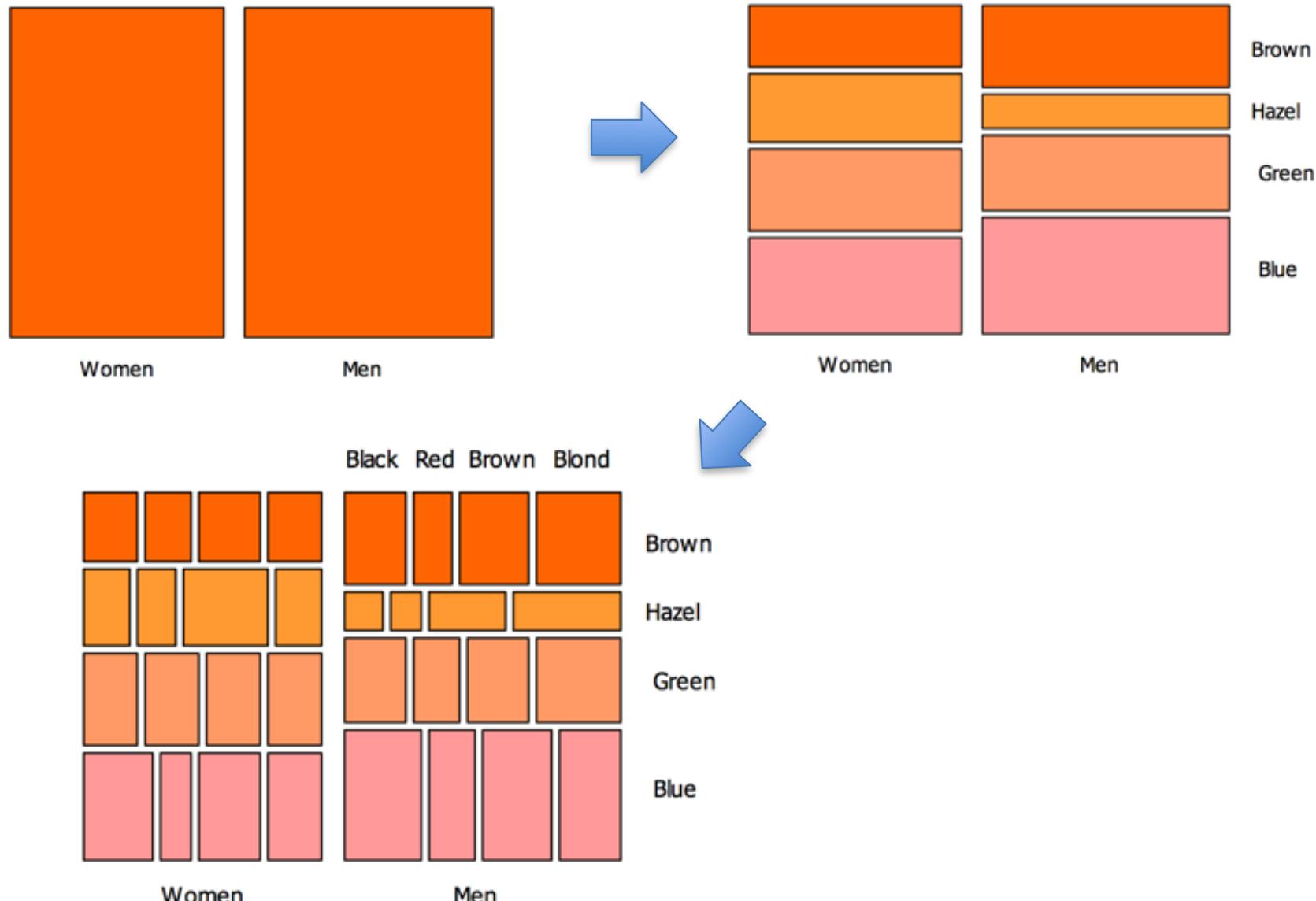
# Mosaic Plot

# Multivariate Categorical Data

- How about multivariate categorical data?
- Students
  - Gender: Female, male
  - Eye color: Brown, blue, green, hazel
  - Hair color: Black, red, brown, blonde, gray
  - Home country: USA, China, Italy, India, ...

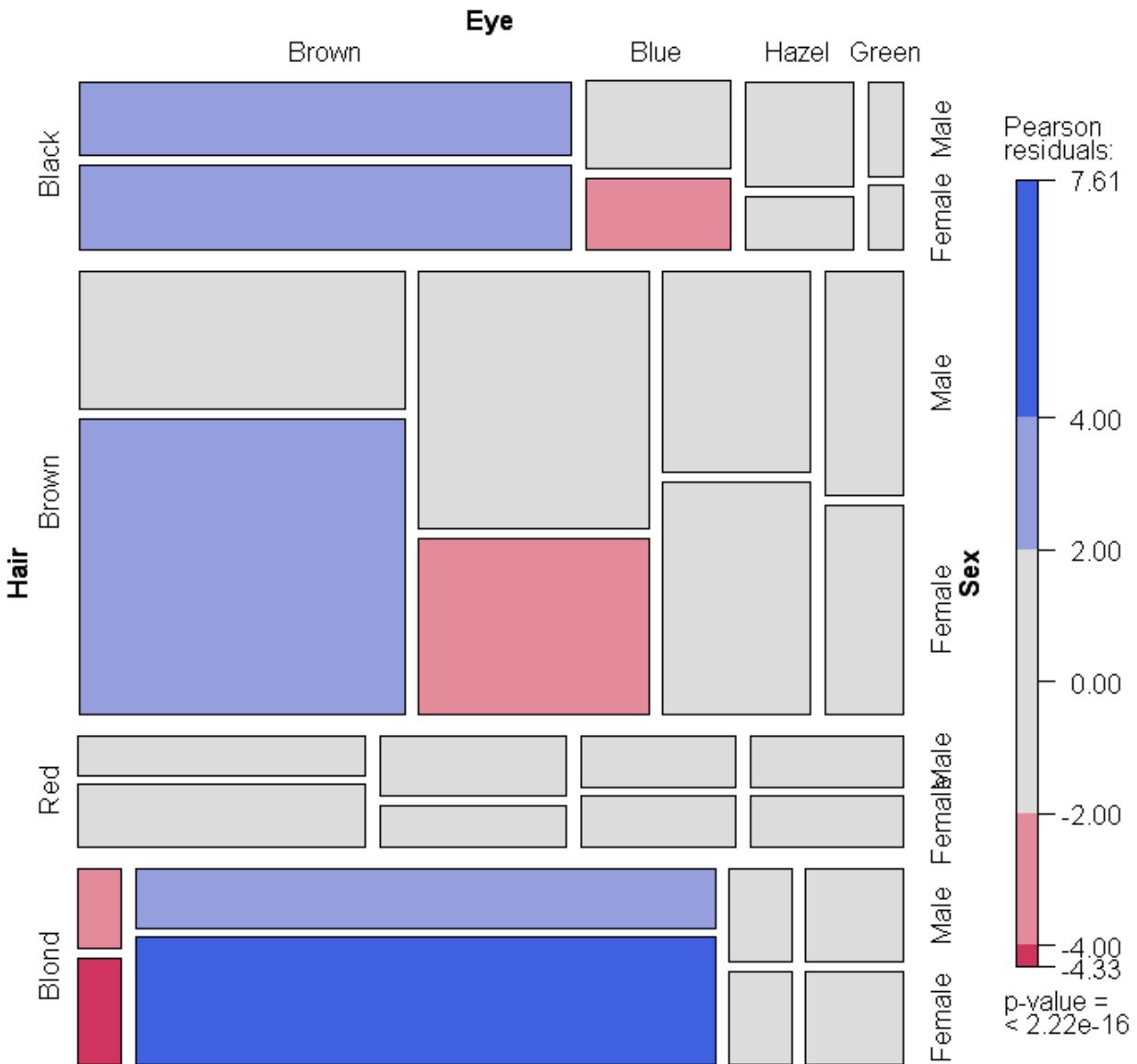
Friendly, Michael. "Mosaic displays for multi-way contingency tables."  
Journal of the American Statistical Association 89.425 (1994): 190-200.

# Mosaic Plot Decomposition



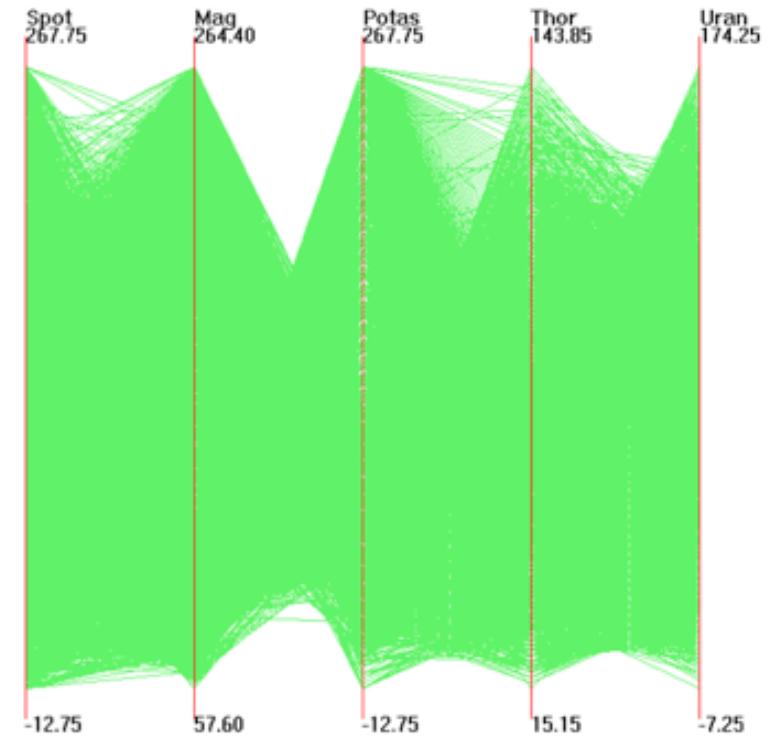
# Mosaic Plot

- Hair
- Sex
- Eye
- Level of the Pearson residual



# Data Overload

- Most of the techniques we've examined work for a modest number of data cases or variables
- What happens when you have lots and lots of data cases and/or variables?



Out5d dataset(5 dimensions, 16384 items)

We will address this in other lectures.

# Summary

- Table vs. Graphs
- Visual encodings (Bertin's semiology)
  - Limitation of possible number of variables
- Reduce Problem Space
  - Small Multiples
  - Multiple Views
- Common Visualizations