Introduction to Visualization and Data Presentation

Jordi Bieger

Reykjavik University
School of Computer Science
Center for Analysis and Design of Intelligent Agents

jbieger@gmail.com

October 7, 2016

Overview

Research

Critical!

A critical part of research is *communicating* your findings to an *audience*.

Communication Methods

- Text
- Math / Logic / Code
- **Tables**
- **Graphs**
- Diagrams
- Illustrations
- Animation
- ...

Communication Methods

- Text
- Math / Logic / Code
- **Tables**
- **Graphs**
- Diagrams
- Illustrations
- Animation
- ...

Communication Methods

- Text
- Math / Logic / Code
- **Tables**
- **Graphs**
- Diagrams
- Illustrations
- Animation
- ...

Communication Methods

- Text
- Math / Logic / Code
- **Tables**
- **Graphs**
- Diagrams
- Illustrations
- Animation
- ...

Communication Methods

- Text
- Math / Logic / Code
- **Tables**
- **Graphs**
- Diagrams
- Illustrations
- Animation
- ...

Communication Methods

- Text
- Math / Logic / Code
- **Tables**
- **Graphs**
- Diagrams
- Illustrations
- Animation
- ...

Communication Methods

- Text
- Math / Logic / Code
- **Tables**
- **Graphs**
- Diagrams
- Illustrations
- Animation
- ...

Communication Methods

- Text
- Math / Logic / Code
- **Tables**
- **Graphs**
- Diagrams
- Illustrations
- Animation
- ...

General remarks

Style

Always check the journal / conference and author instructions for the general style of tables and figures.

Early

Consider the kind of visualizations you want to use when you are designing your experiment.

General remarks - cont.

Captions

Captions should make it possible to understand completely what a table or figure shows.

Completeness

By highlighting and discussing the important parts of tables and figures, the text should be understandable just by reading the text.

Variable Types

Experimental variables:

1. **Independent**: Variable that is not changed by the other variables (e.g. age).
2. **Dependent**: Measured variable that is affected by others (e.g. cancer risk).

Data types:

1. **Nominal / Categorical**: Discrete data that cannot be ordered (e.g. people, countries). Operations: count, mode
2. **Ordinal**: Quantities with a natural order (e.g. Likert scale). Extra operations: order, median
3. **Interval**: Ordinal + the interval between each value is equal (e.g. Celsius, Fahrenheit). Extra operations: mean, add, subtract
4. **Ratio**: Interval + a natural zero point (e.g. elevation, money). Extra operations: multiply, divide

Tables

Table: Caption (often above table).

Stub	**Column heading**	**Column heading**
Row variable 1	x%	x%
Row variable 2	x%	x%
Row variable 3	x%	x%
Row variable 4	x%	x%
Total	x%	x%

Multivariate table

Attitude towards uranium mining by age and gender (hypothetical data)

	Number of respondents												
	<25		25–34		35–44		45–54		55+		Total		
Attitude towards uranium mining	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**T**
Strongly favourable	0	0	1	1	3	1	5	2	3	–	12	4	16
Favourable	0	0	1	2	3	2	3	1	0	0	7	5	12
Uncertain	0	0	0	0	1	1	2	2	0	0	3	3	6
Unfavourable	1	1	4	3	1	0	0	0	0	0	6	4	10
Strongly unfavourable	4	8	17	7	8	7	2	3	0	0	31	25	56
Total	5	9	23	13	16	11	12	8	3	0	59	41	100

Figure: Table 16.4 from the book.

Confusion matrix

		Predicted	
		True	*False*
Actual	*True*	tp	fn
	False	fp	tn

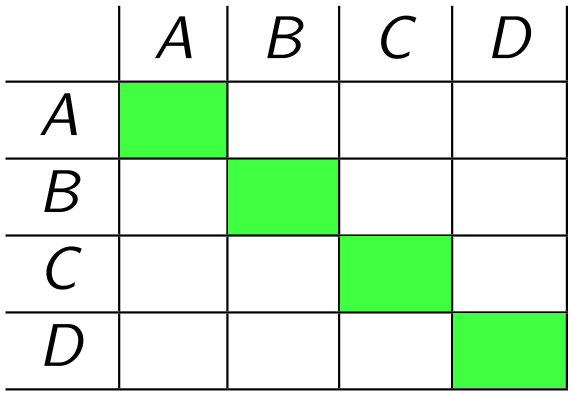


Table usage

Use a table when:

- Detailed data
- Large volume*
- No trend or pattern

Anscombe

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

→

$N = 11$
mean of X's $= 9.0$
mean of Y's $= 7.5$
equation of regression line: $Y = 3 + 0.5X$
standard error of estimate of slope $= 0.118$
$t = 4.24$
sum of squares $X - \overline{X} = 110.0$
regression sum of squares $= 27.50$
residual sum of squares of $Y = 13.75$
correlation coefficient $= .82$
$r^2 = .67$

Figure: From Anscombe (1973), "Graphs in Statistical Analysis" via VDQI (page 13)

Anscombe

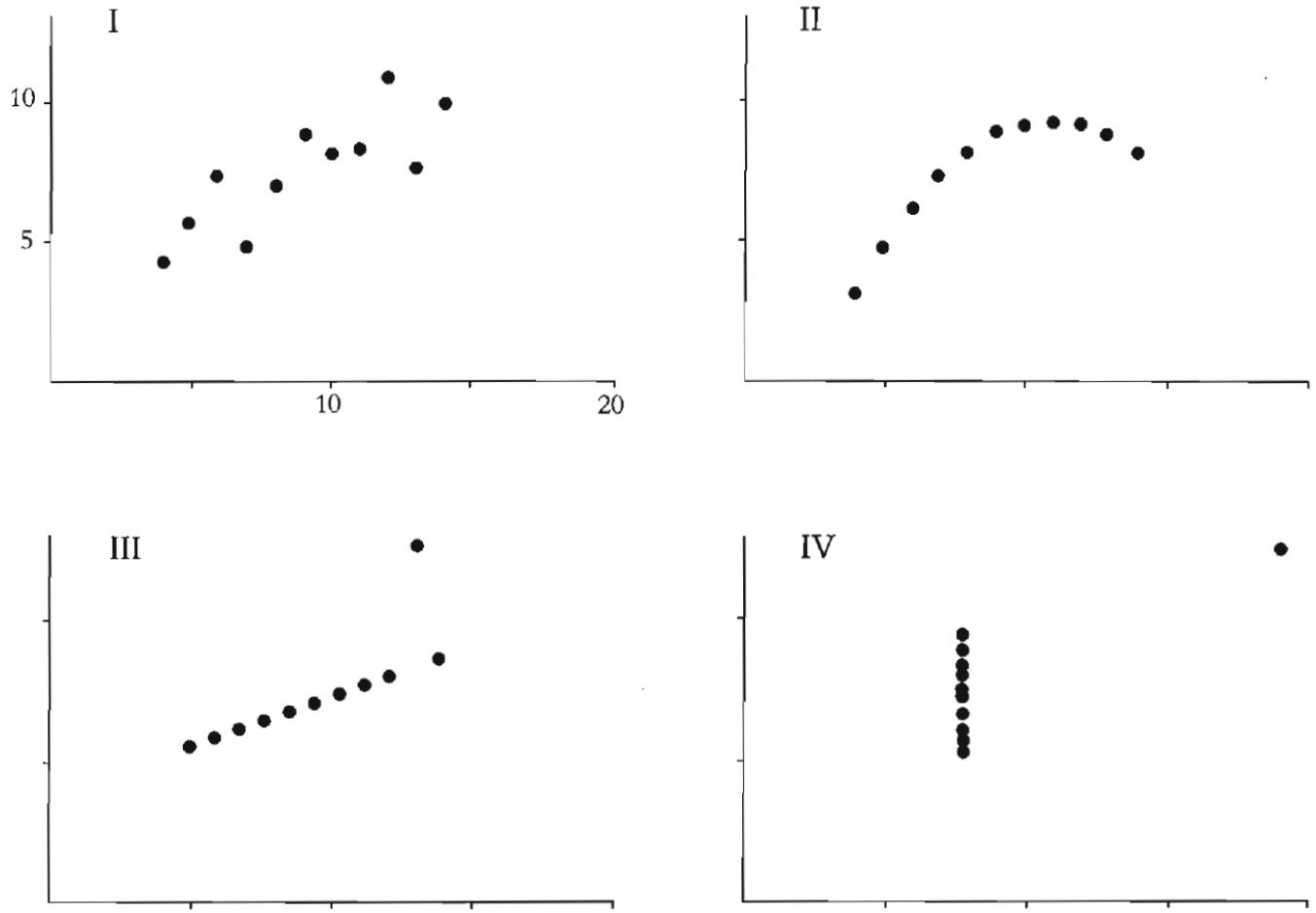


Figure: From Anscombe (1973), "Graphs in Statistical Analysis" via VDQI (page 14)

2D Chart Anatomy

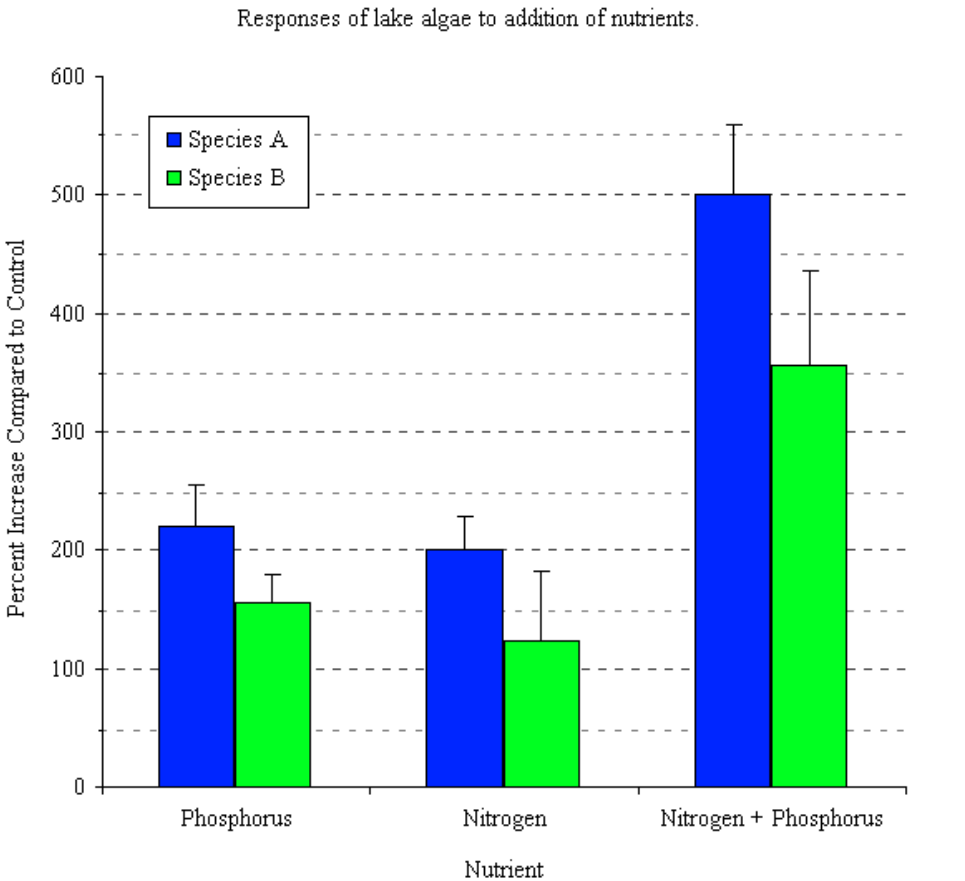


Figure 3. Responses of lake algae to addition of nutrients. Error bars are 95% confidence limits.

N = 50 lakes.

Figure: Adapted from **University of Wisconsin-La Crosse (2001)**.

2D Chart Anatomy

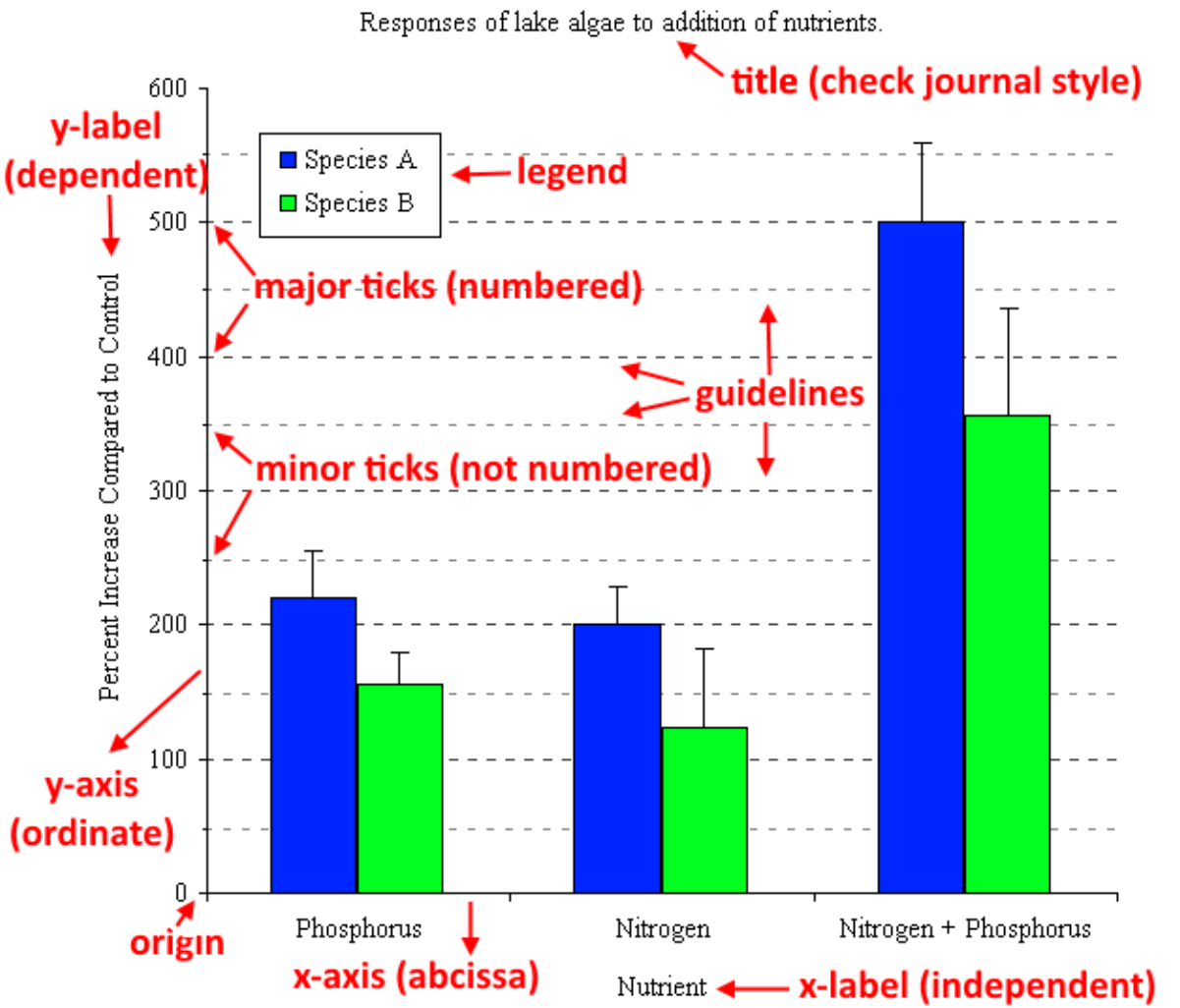


Figure 3. Responses of lake algae to addition of nutrients. Error bars are 95% confidence limits.

N = 50 lakes.

caption

2D Chart Anatomy - Axis Offset

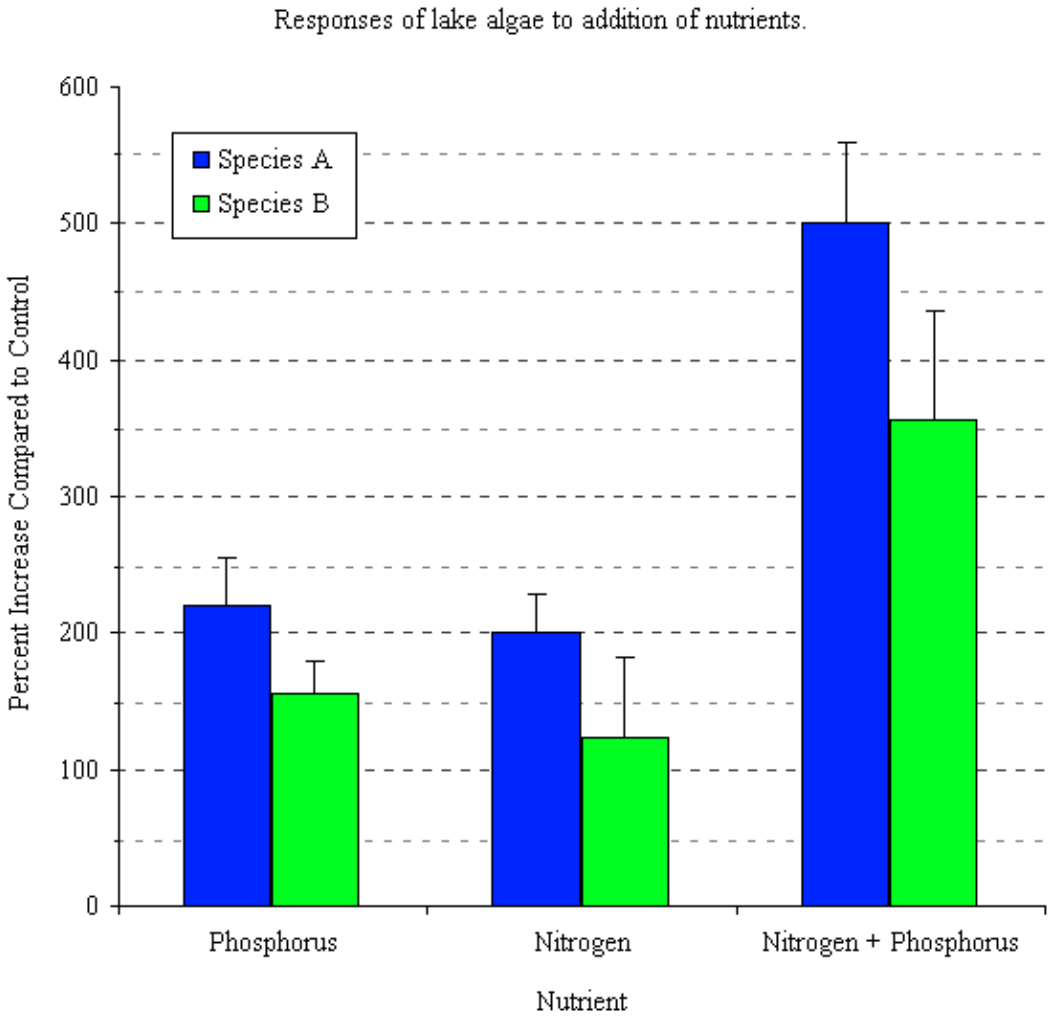


Figure 3. Responses of lake algae to addition of nutrients. Error bars are 95% confidence limits.

N = 50 lakes.

2D Chart Anatomy - Axis Offset

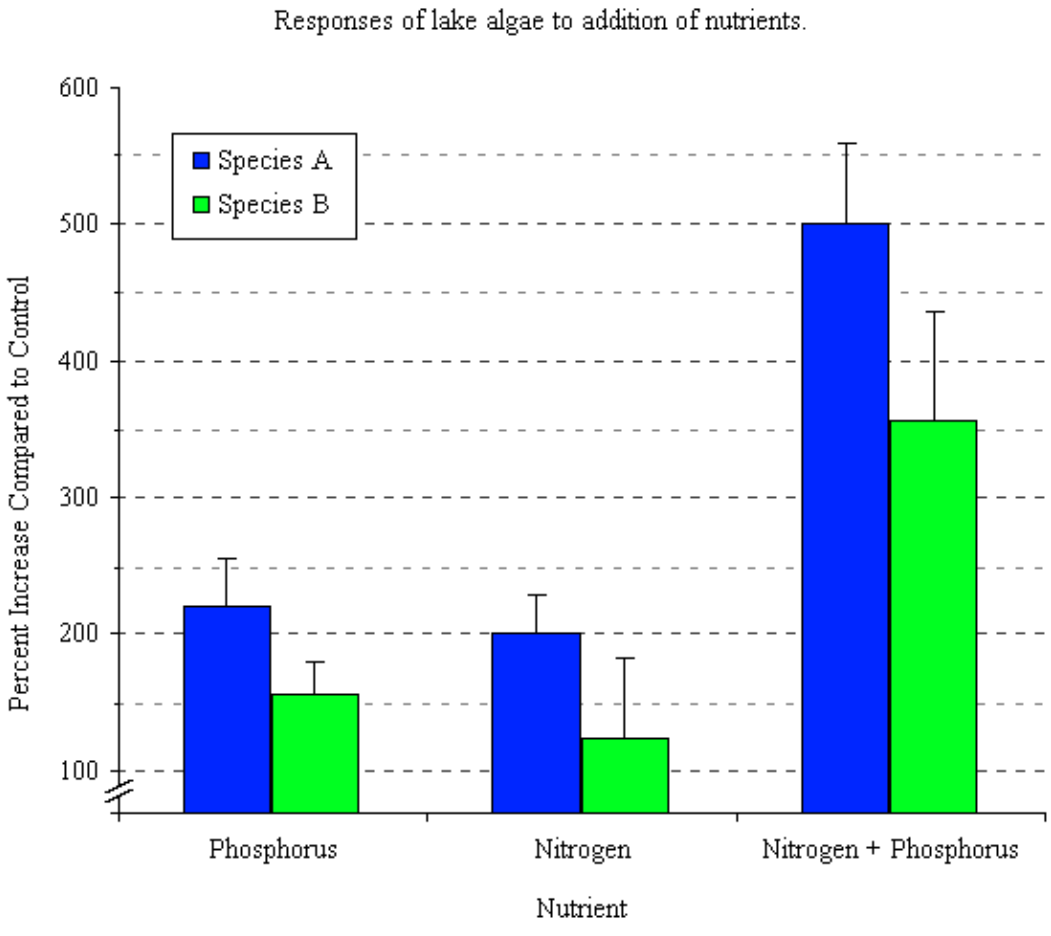


Figure 3. Responses of lake algae to addition of nutrients. Error bars are 95% confidence limits.

N = 50 lakes.

2D Chart Anatomy - Axis Offset

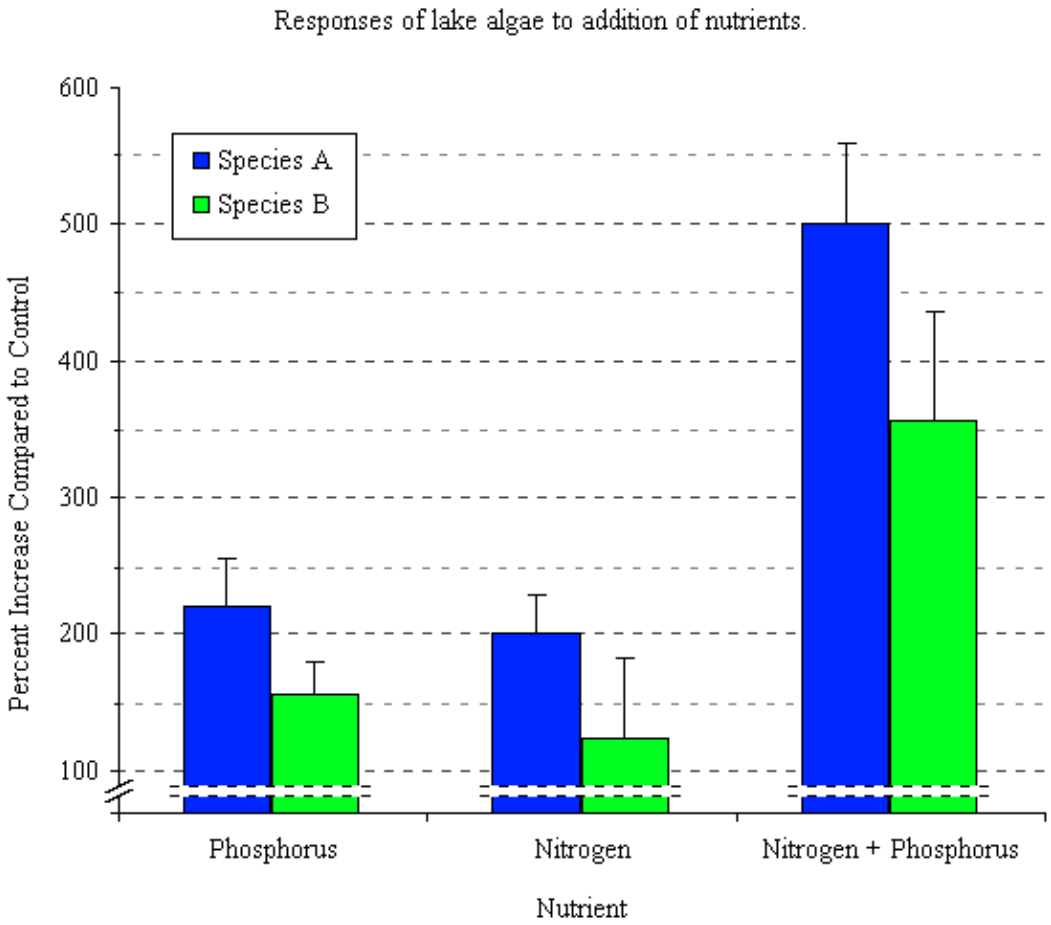
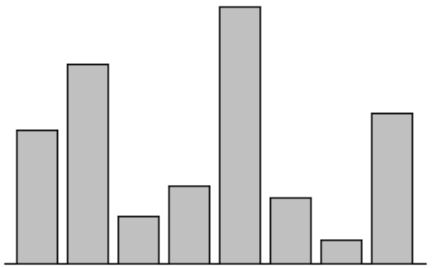


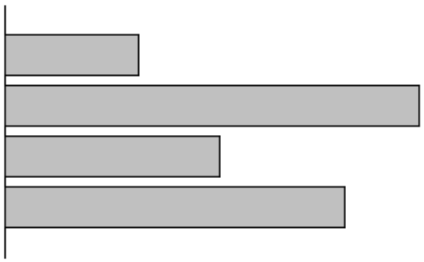
Figure 3. Responses of lake algae to addition of nutrients. Error bars are 95% confidence limits.

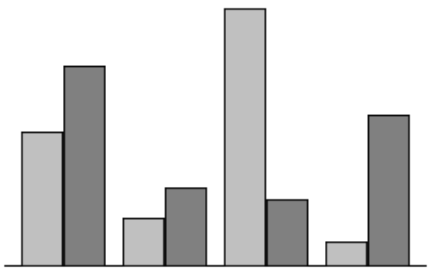
N = 50 lakes.

Bar charts

- Use when you want to compare how values of 1 or 2 discrete independent variables affect a numeric dependent variable or count.
- Actual numbers and/or error bars can be added on top of the bars.
- For ordinal data, a histogram may also be used.
- With 2 independent variables, a stacked bar chart can also be used, but this is not recommended for comparisons.

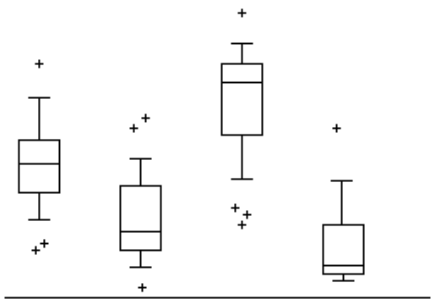






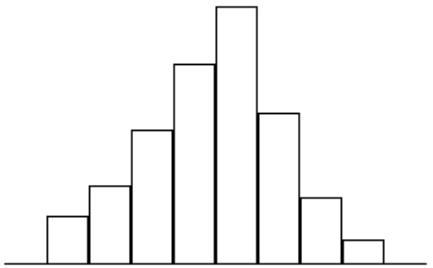
Box plots

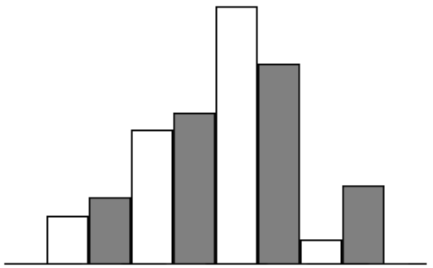
- Box plots are like bar charts with extra information.
- They generally show the 1st, 2nd and 3rd quartile of the data, the range and outliers.



Histograms

- Use when you want to show the distribution of items over a small number of values of a quantitative variable.
- For ordinal data, a bar chart may also be used.
- Ratio or interval data can be divided into buckets/intervals; otherwise you can use a area chart.
- With multiple variables, use a 3D effect or an overlapping area chart.
- The book calls drawing a line over the histogram a "frequency polygon".





Stem-and-leaf displays

A stem-and-leaf display let's you show fairly detailed distribution information in the shape of a histogram.

Example (Data)

```
37, 33, 33, 32, 29, 28, 28, 23,
22, 22, 22, 21, 21, 21, 20, 20,
19, 19, 18, 18, 18, 18, 16, 15,
14, 14, 14, 12, 12, 9, 6
```

Example from **Lane @ OnlineStatBook**.

Example (S&L display 1)

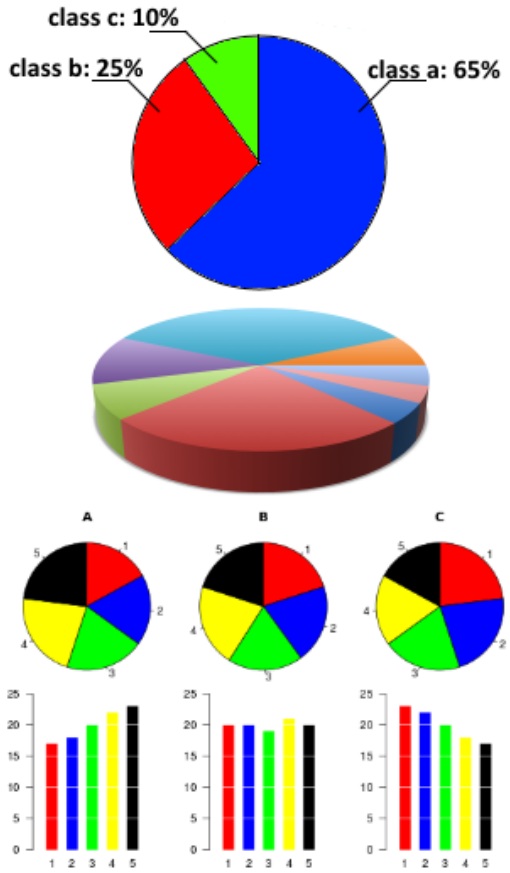
```
3|2337
2|001112223889
1|2244456888899
0|69
```

Example (S&L display 2)

```
3|7
3|233
2|889
2|001112223
1|56888899
1|22444
0|69
```

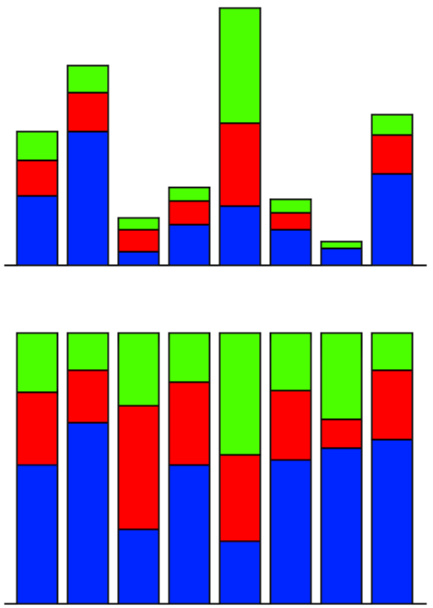
Pie charts

- Pie charts can show how some quantity (100%) is divided over various categories.
- Categories often sorted (beware continuity between pie charts).
- Beware perspective (for all charts).
- Difficult to compare categories. Easier to judge percentage of whole.
- Heavily criticized.



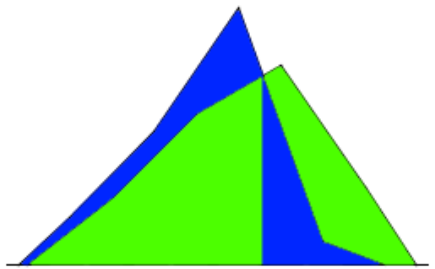
Stacked bar charts

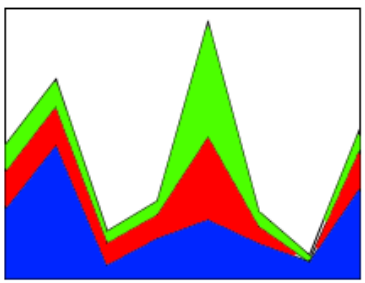
- Use to show the composition of a thing varying along some discrete dimension.
- Use a 100% bar chart if the absolute value doesn't matter.

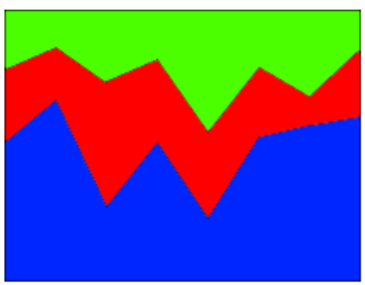


Area charts

- Area charts can be used to show distributions under a continuous independent variable.
- Stacked area charts can also be used to show how compositions vary with such a variable.

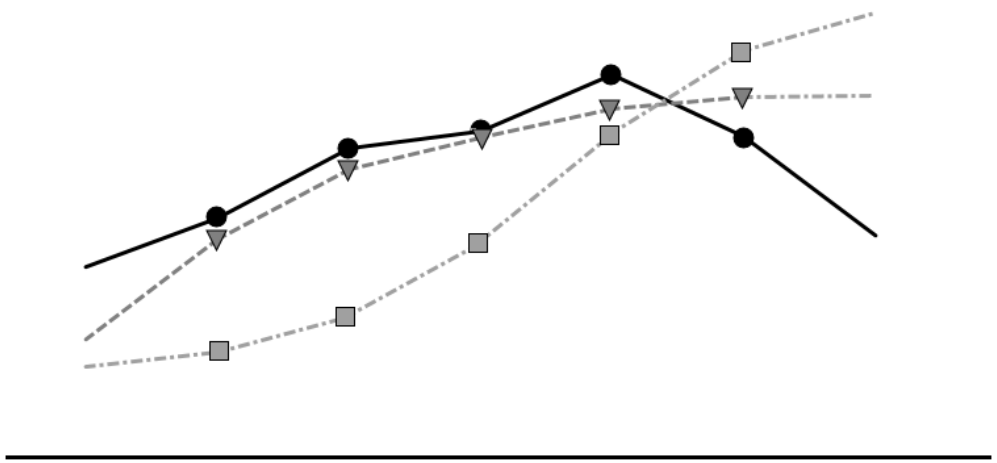






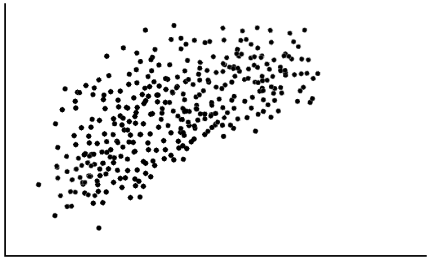
Line charts

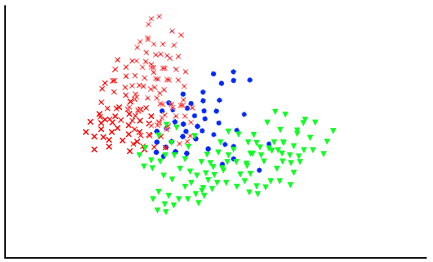
Line charts can be used to show how several numeric quantitative variables change with another variable (e.g. time).

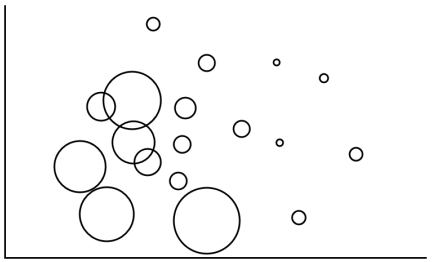


Scatter plots

- Scatter plots are useful for seeing the relationship between two quantitative variables.
- Bubble plots let you add another dimension.







Color

Journal

Always check the style of the journal!

Legibility

Keep everything legible!

Account for B&W

Even if you use color, make sure your figures are interpretable if someone prints them without or is color blind. (Don't refer to the color in the text.)

Bar chart color

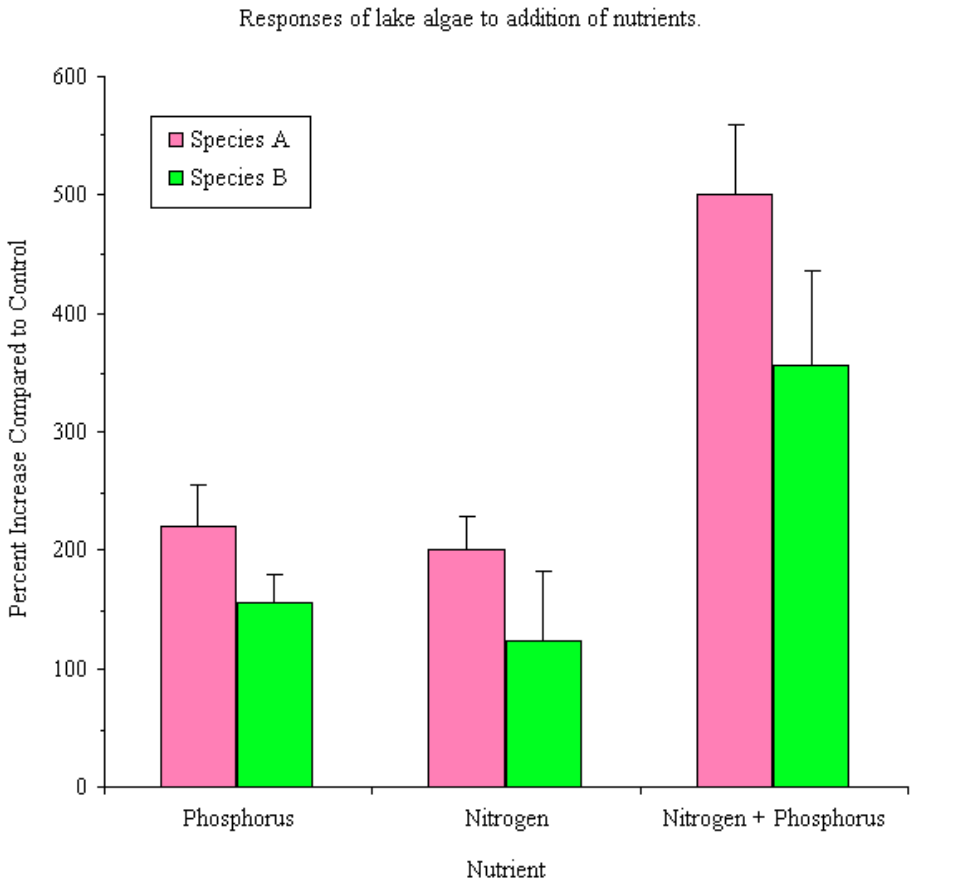


Figure 3. Responses of lake algae to addition of nutrients. Error bars are 95% confidence limits.

N = 50 lakes.

Figure: Adapted from **University of Wisconsin-La Crosse (2001)**.

Bar chart color

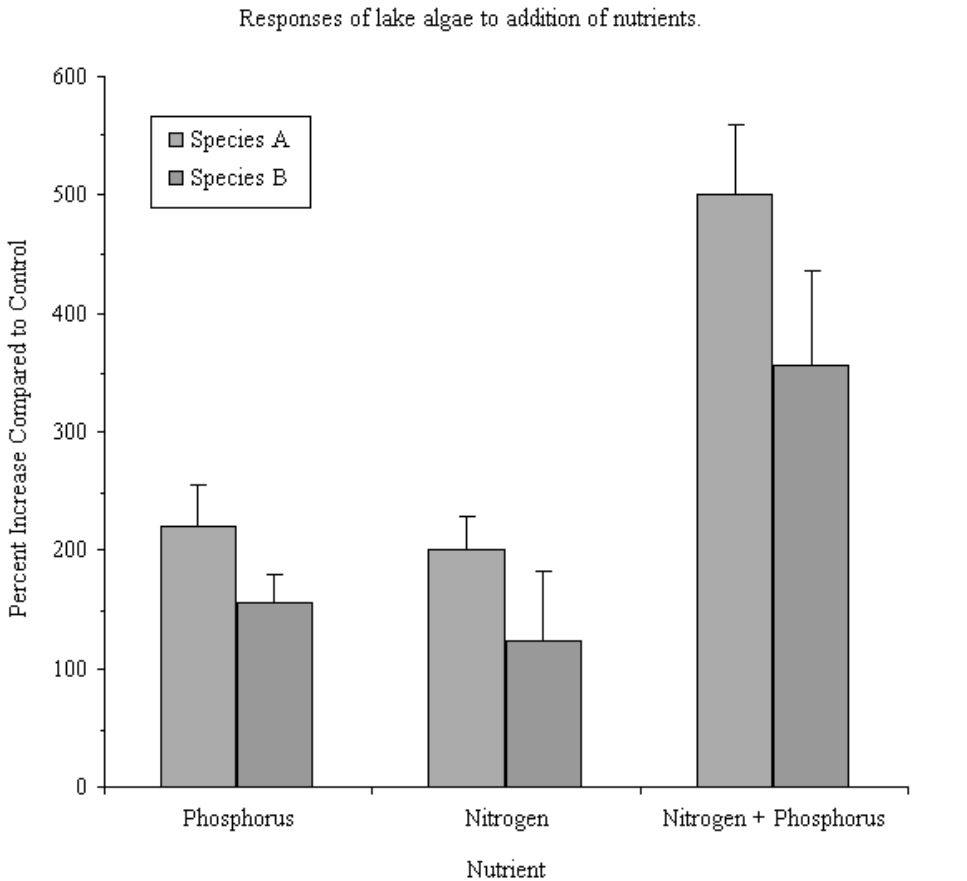


Figure 3. Responses of lake algae to addition of nutrients. Error bars are 95% confidence limits.

N = 50 lakes.

Figure: Adapted from **University of Wisconsin-La Crosse (2001)**.

Bar chart color - Hatching

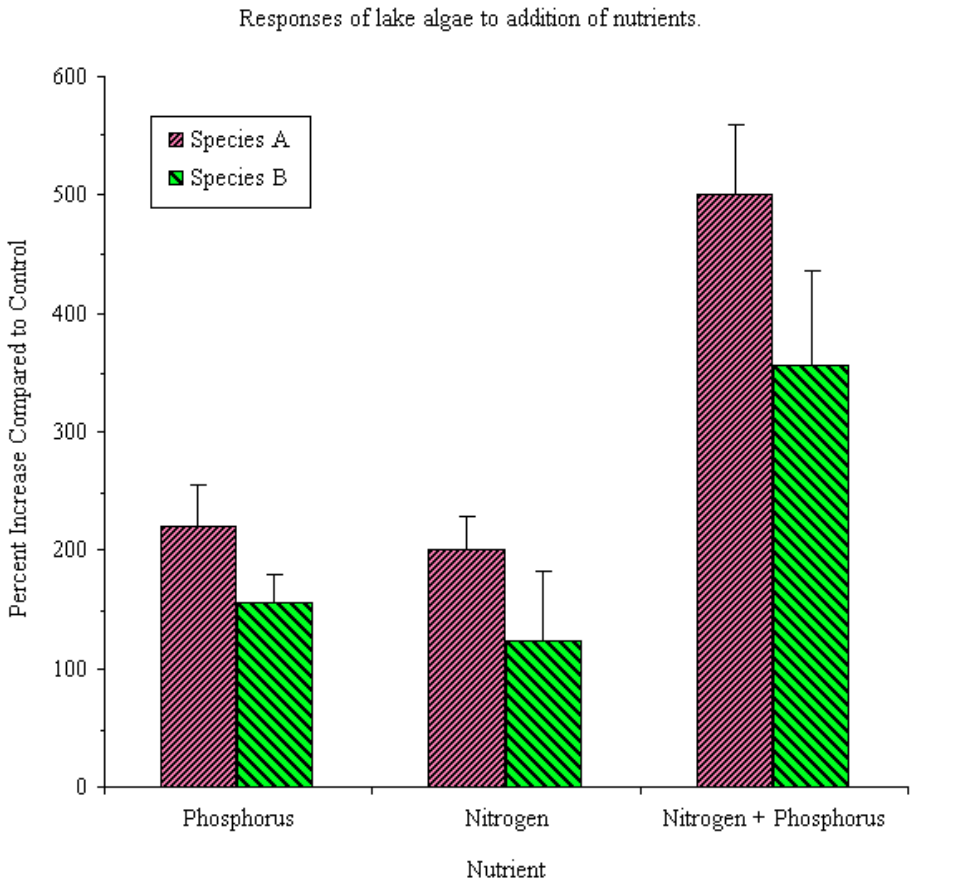


Figure 3. Responses of lake algae to addition of nutrients. Error bars are 95% confidence limits.

N = 50 lakes.

Figure: Adapted from **University of Wisconsin-La Crosse (2001)**.

Bar chart color - Hatching

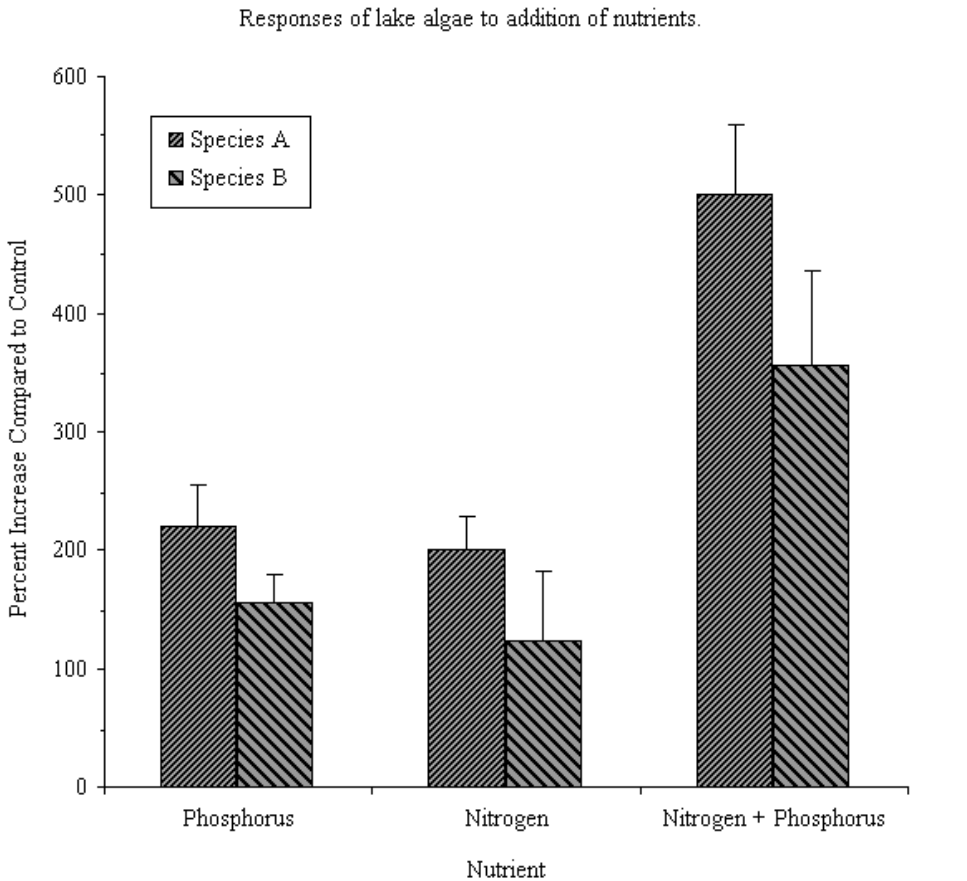


Figure 3. Responses of lake algae to addition of nutrients. Error bars are 95% confidence limits.

N = 50 lakes.

Figure: Adapted from **University of Wisconsin-La Crosse (2001)**.

Line chart color - Notches and Line Types

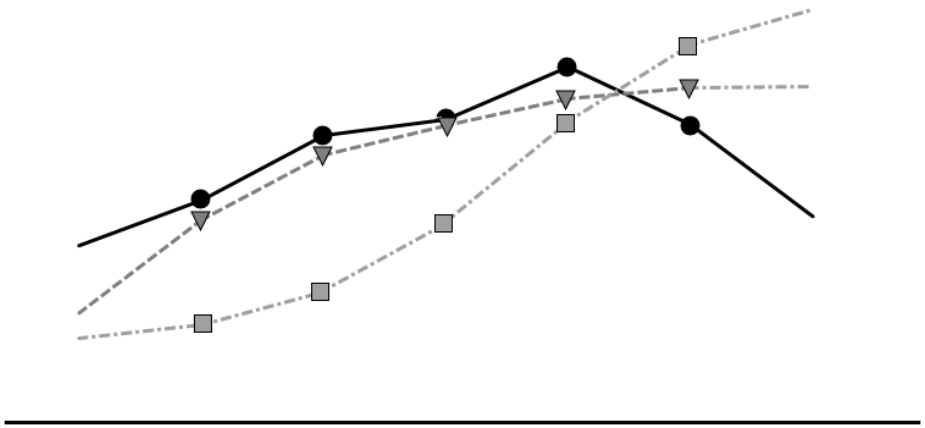
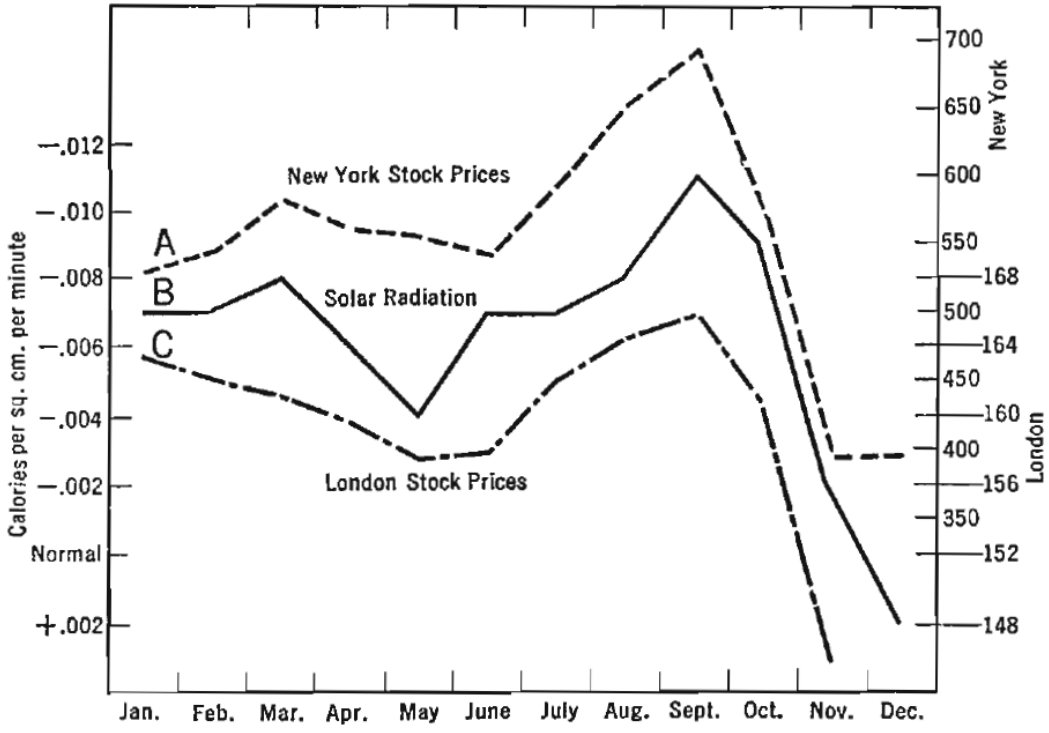


Figure: Use different notches and line types.

Multiple Y-Scales



SOLAR RADIATION AND STOCK PRICES

A. New York stock prices (Barron's average). B. Solar Radiation, inverted, and C. London stock prices, all by months, 1929 (after Garcia-Mata and Shaffner).

Figure: From Dewey & Dakin (1947), "Cycles: The science of prediction", p. 144 via VDQI (page 15)

Graphical Integrity

Graphical Integrity

The ability of a graph to provide a visual representation that is consistent with an underlying numerical representation that **accurately represents the world**.

Subjectivity

Peculiarities of human perception should be taken into account and accommodated rather than exploited. For example, perceived area of a circle = (actual area)^x where $x = .8 \pm .3$. Solution: clear labeling.

Lie Factor

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

Acceptable between .95 and 1.05.

Exercise

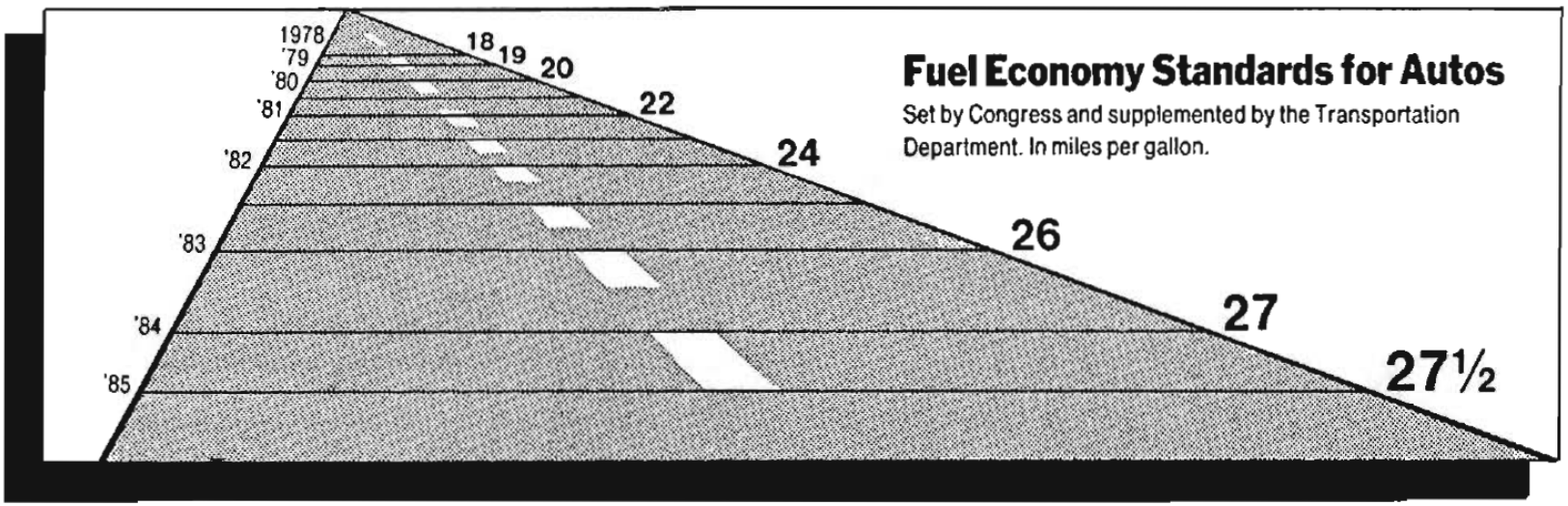


Figure: Adapted from New York Times, August 9 1978, p. D-2 via VDQI (page 57)

Exercise

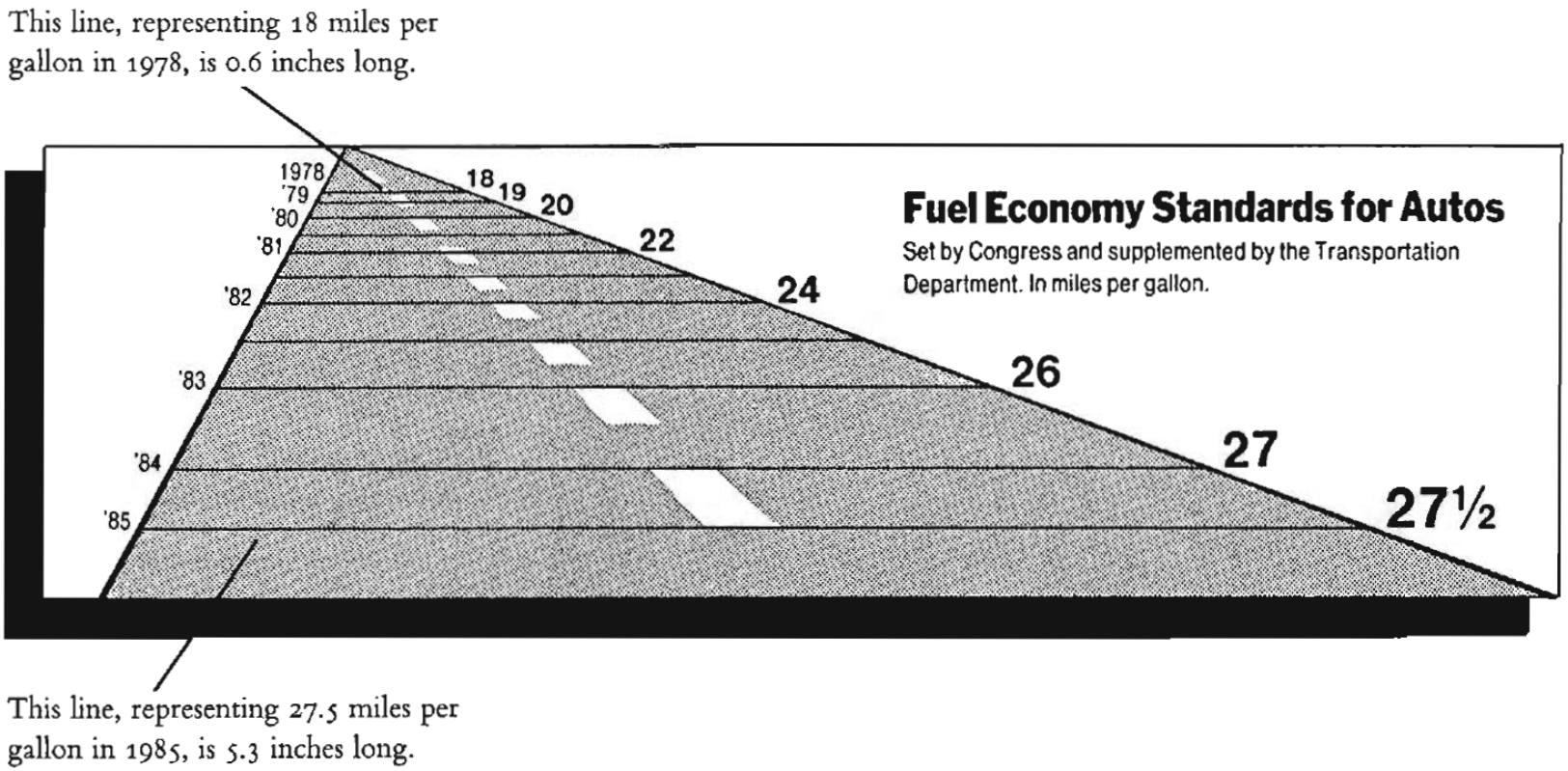
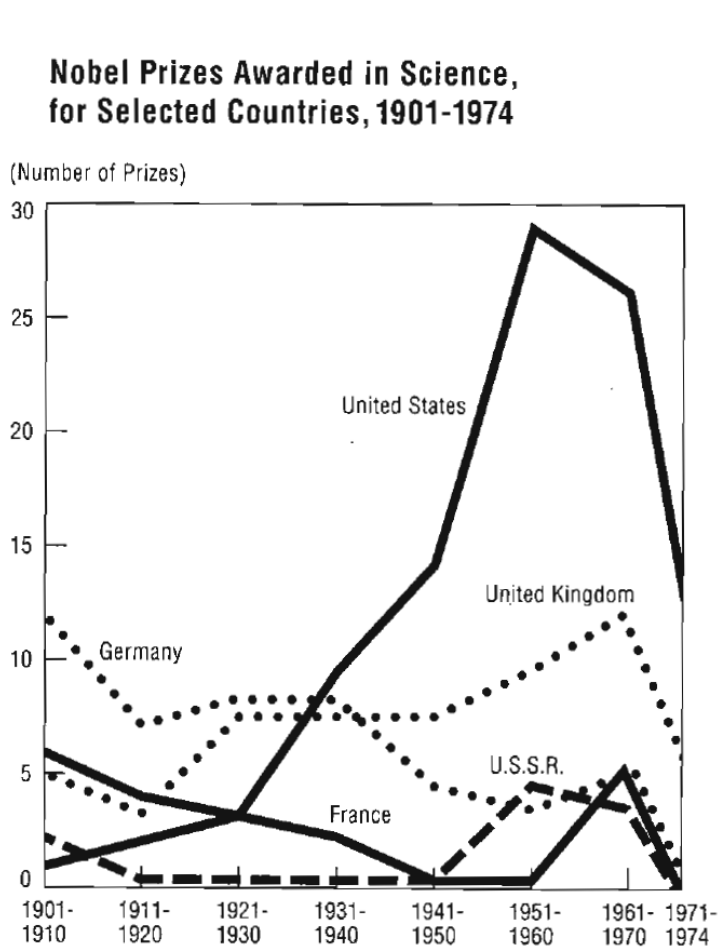


Figure: From New York Times, August 9 1978, p. D-2 via VDQI (page 57) Lie factor: 14.8 or 111

Design and Data Variation



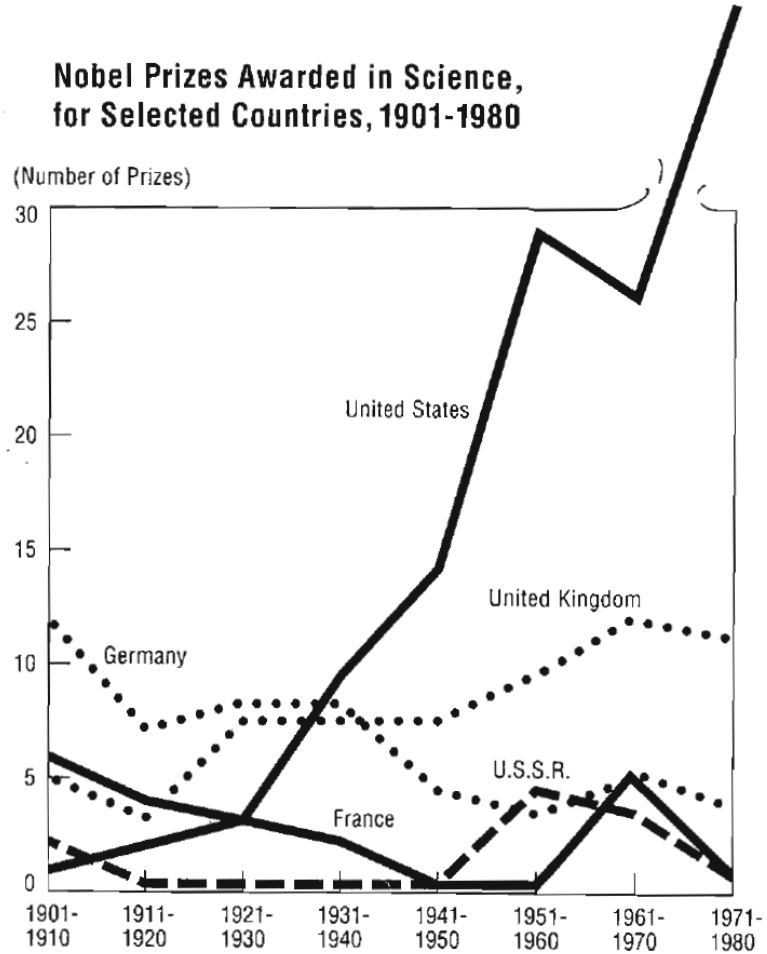


Figure: From National Science Foundation, Science Indicators, 1976 (Washington D.C., 1976) via VDQI (page 60)

2-D representation of 1-D data

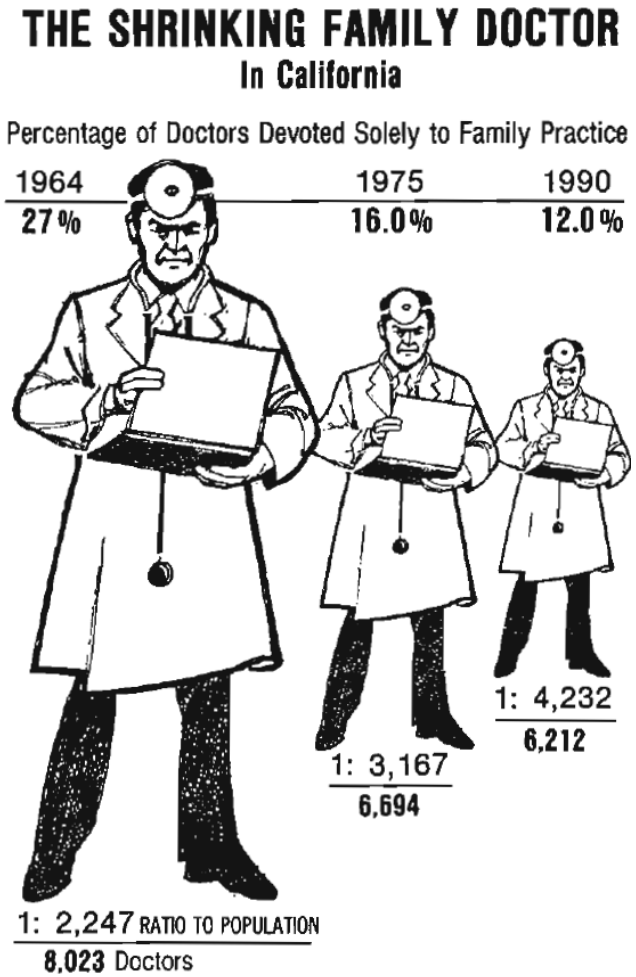


Figure: From Los Angeles Times, August 5 1979, p. 3 via VDQI (page 69) Lie factor: 2.8

3-D representation of 1-D data

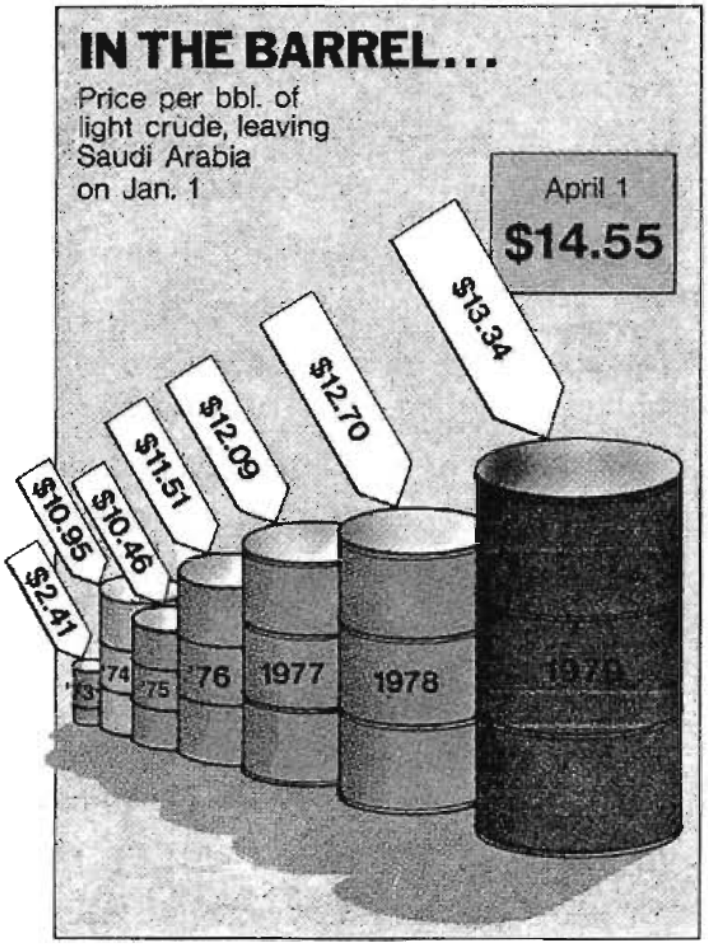


Figure: From Time, April 9 1979, p. 62 via VDQI (page 62) Lie factor: 9.4 or 59.4

Money

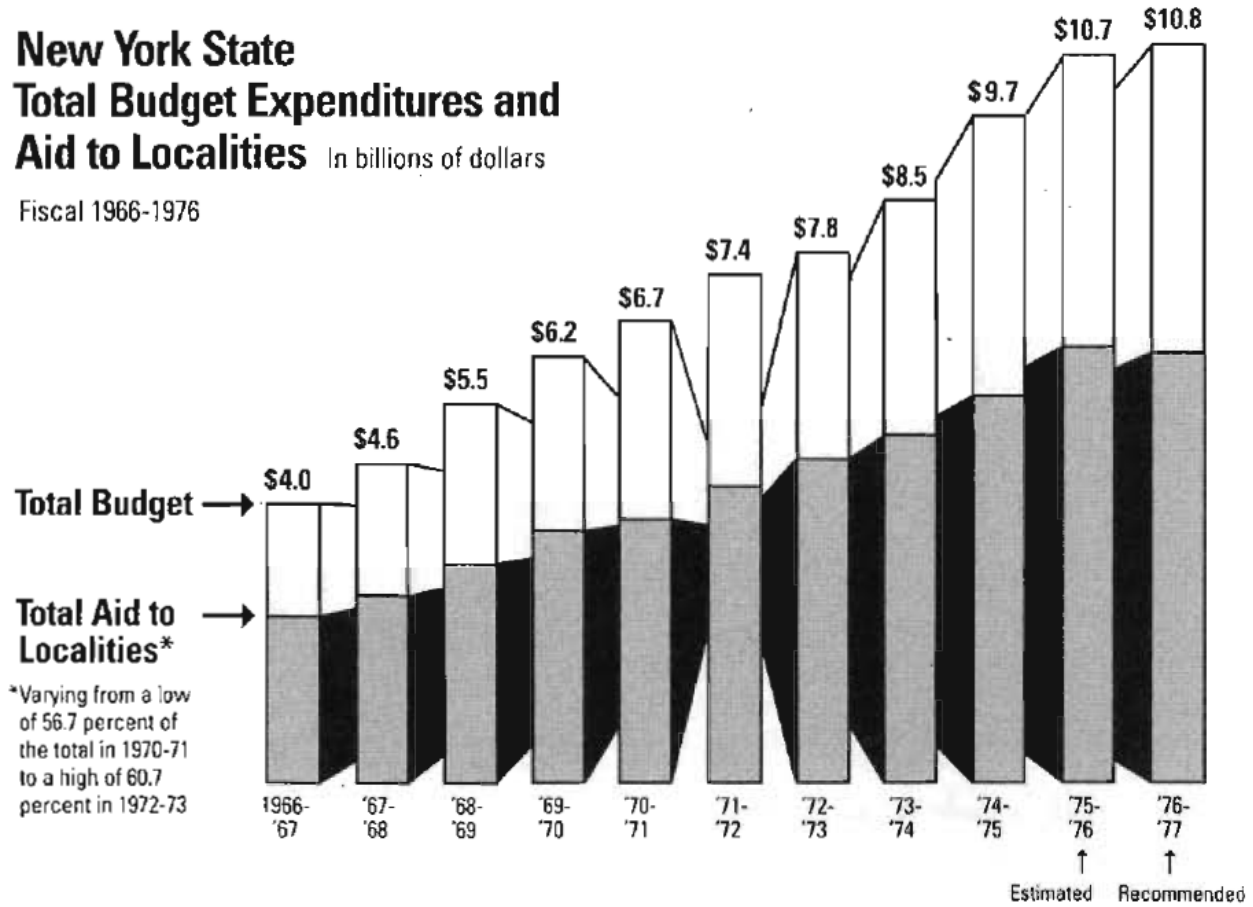


Figure: From New York Times, February 1 1976, p. IV-6 via VDQI (page 66)

Correct for inflation and other factors

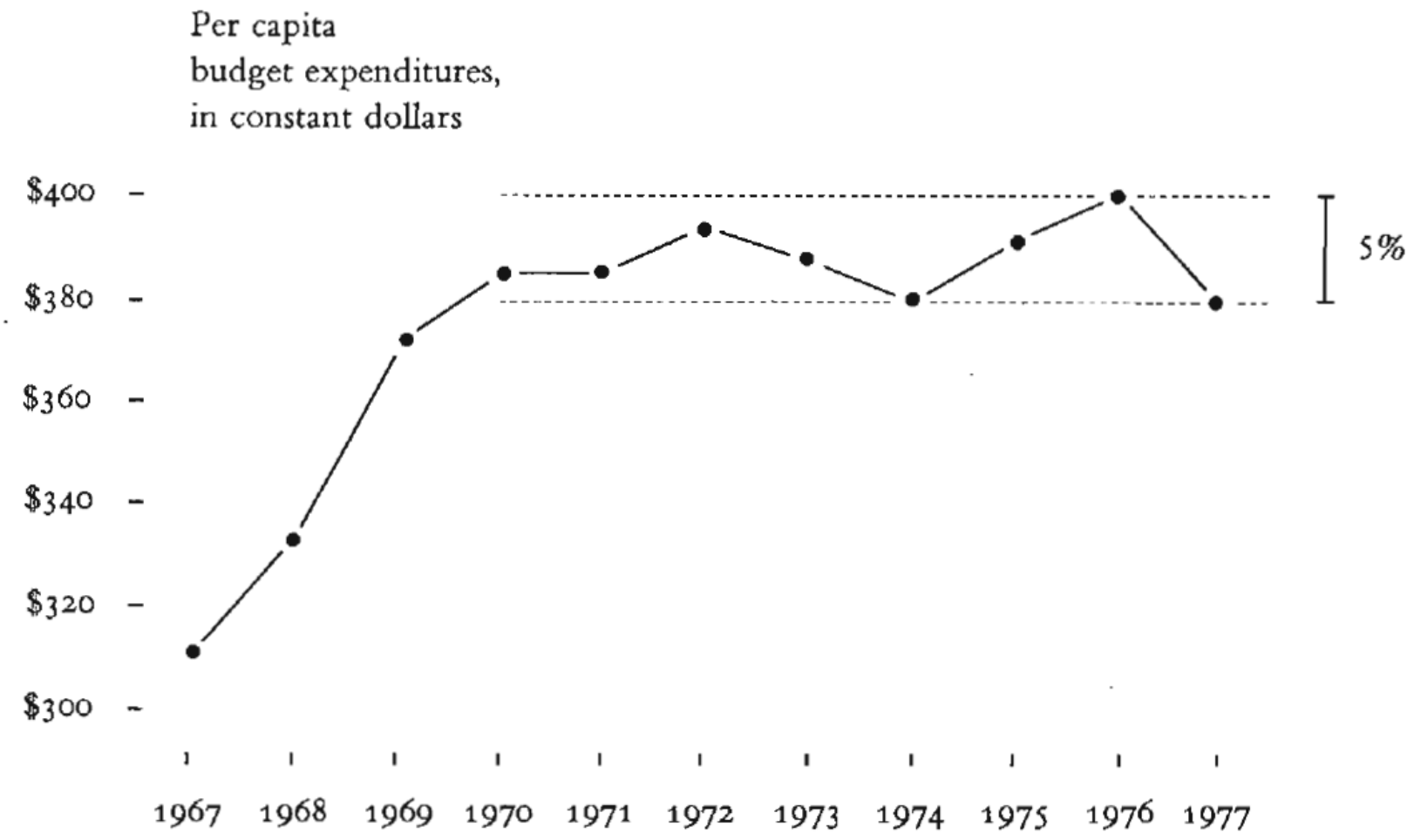


Figure: From VDQI (page 68)

Context

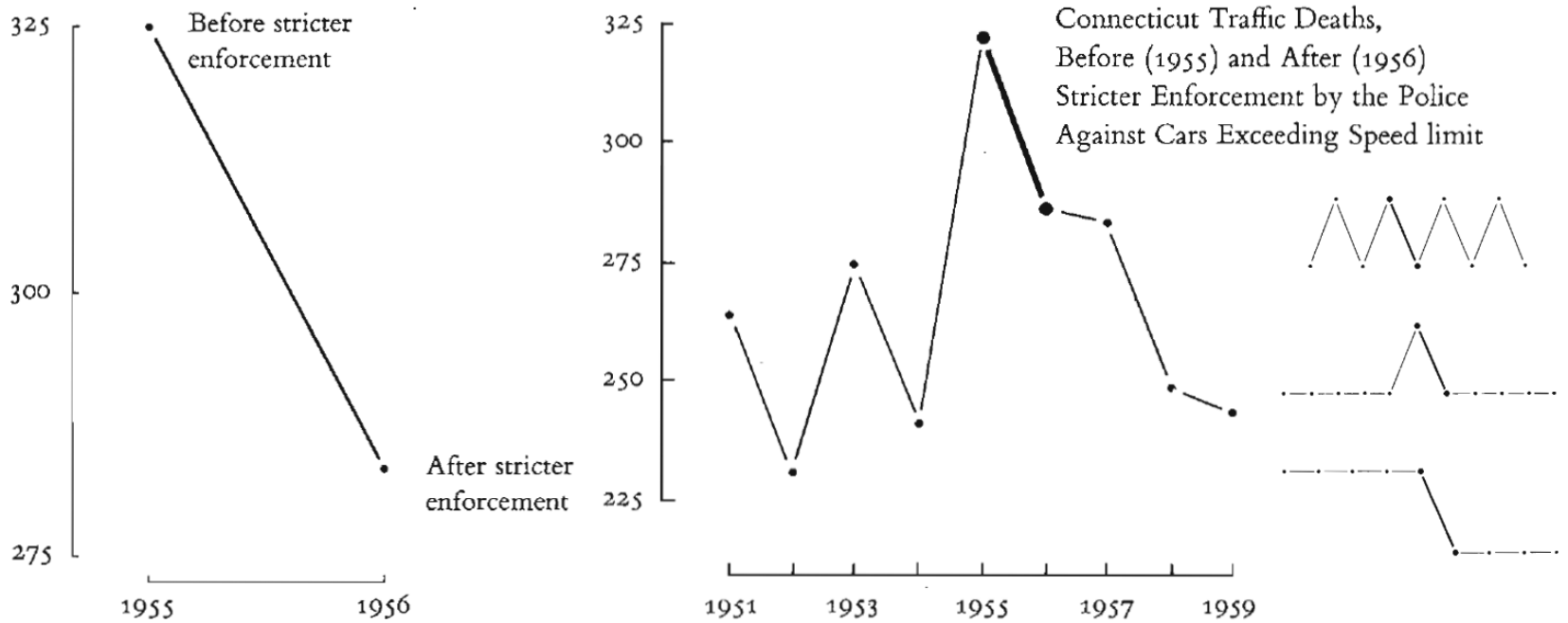


Figure: From Campbell & Ross (1970), "The Connecticut Crackdown on Speeding: Time Series Data in Quasi-Experimental Analysis" via VDQI (page 74)

Context

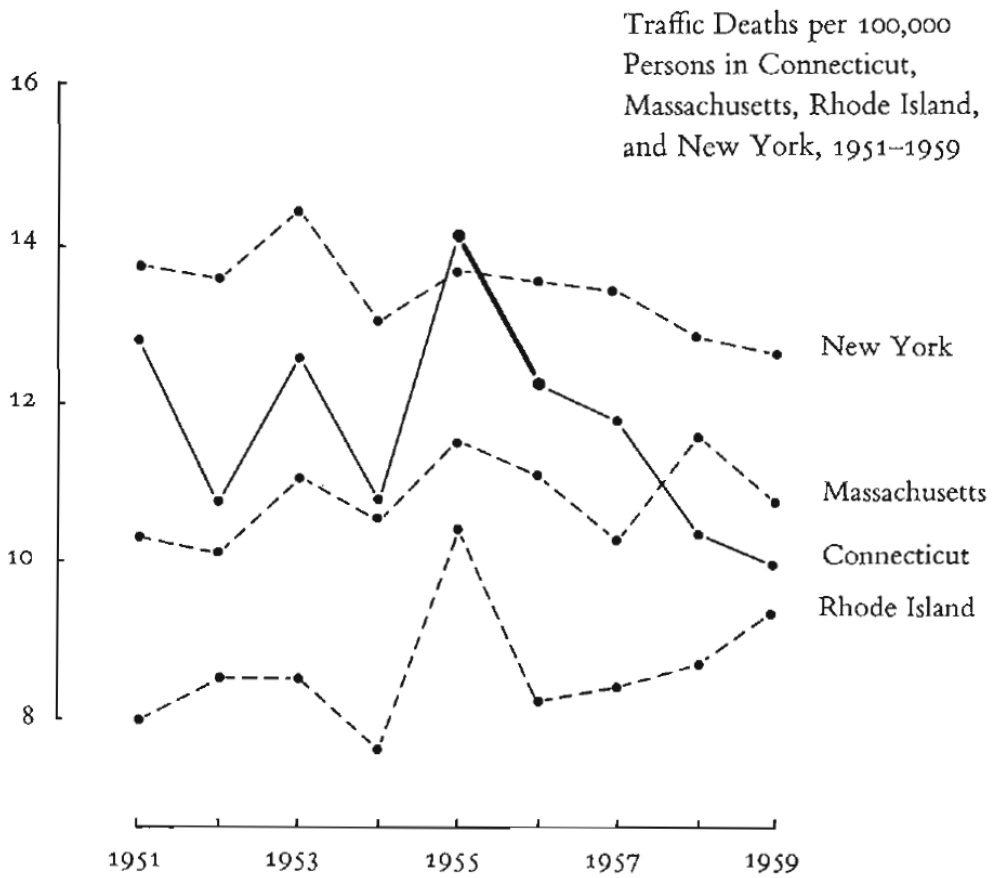
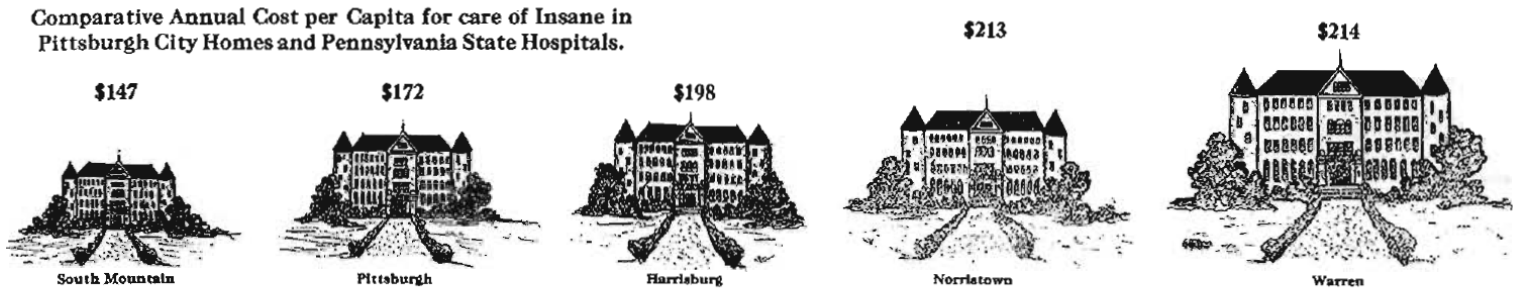


Figure: From Campbell & Ross (1970), "The Connecticut Crackdown on Speeding: Time Series Data in Quasi-Experimental Analysis" via VDQI (page 75)

Numbers have order *and* magnitude



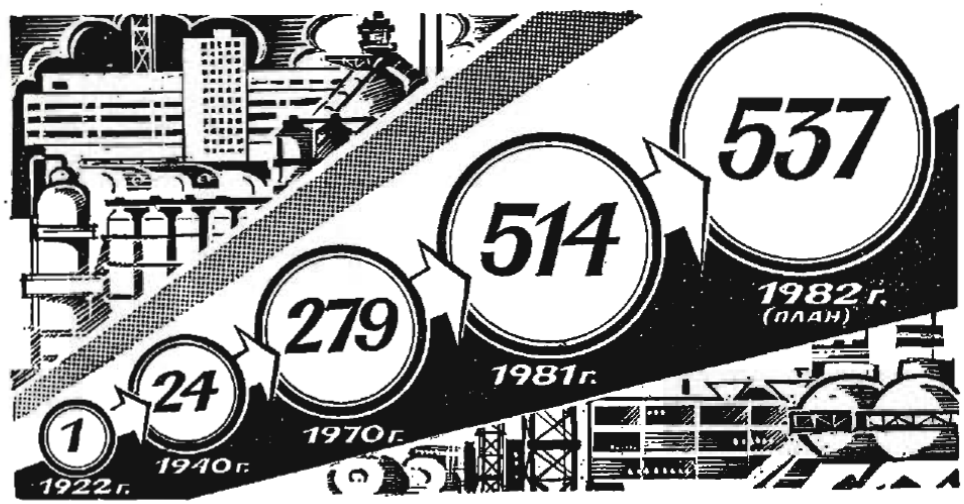


Figure: From Pittsburgh Civic Commission (1911), "Report on Expenditures of the Department of Charities" and Pravda, May 24 1982 p.2 via VDQI (page 55 and 76)

Truncated Y-Axis

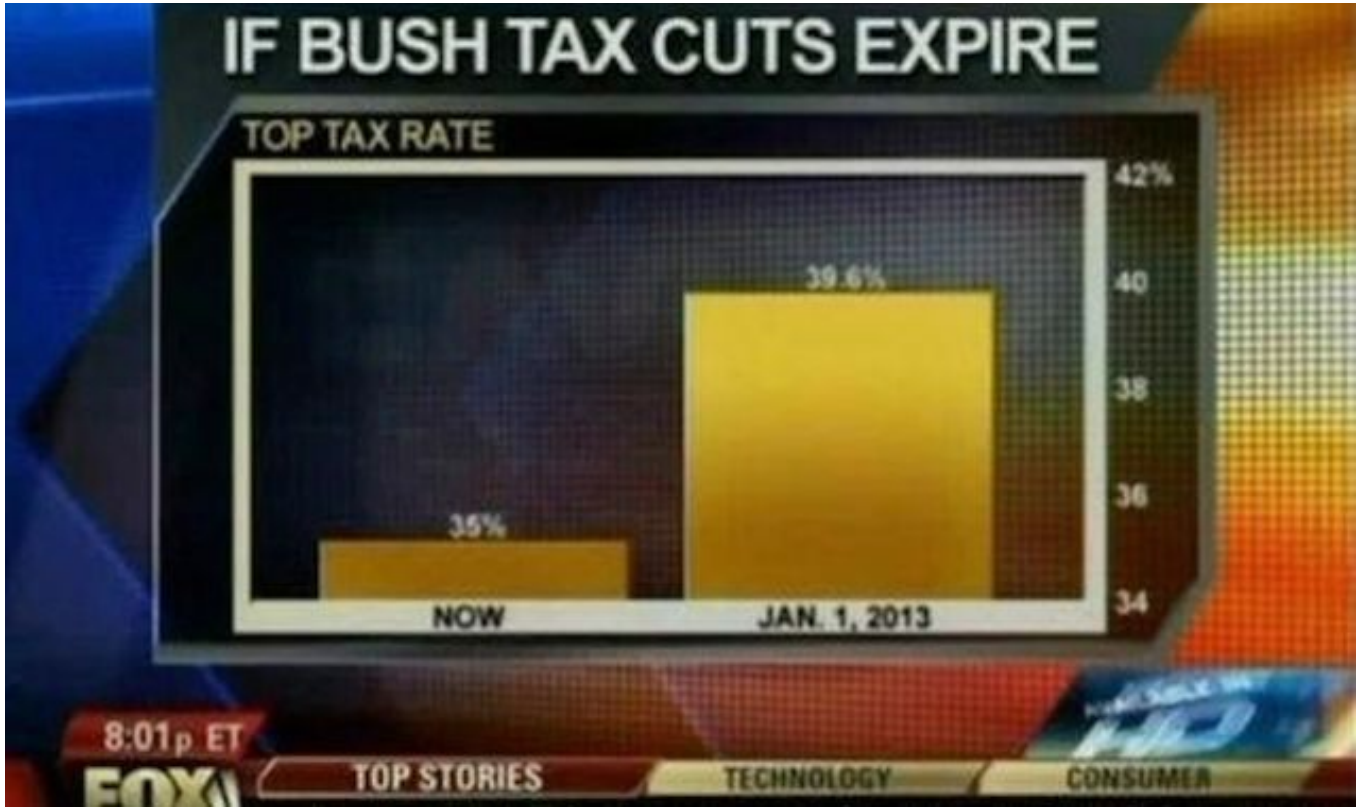


Figure: Via **Parikh @ Gizmodo**.

Different Y-Axis

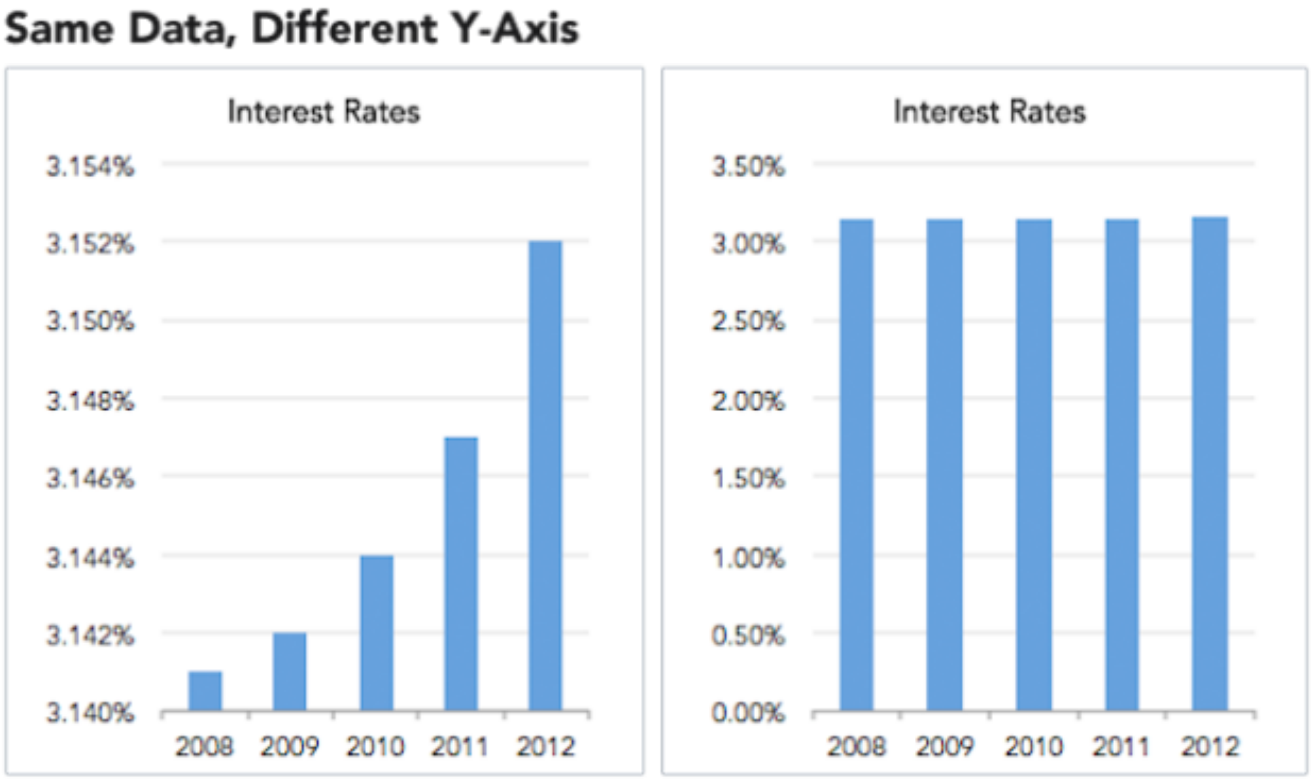


Figure: From **Parikh @ Gizmodo**.

Different Y-Axis

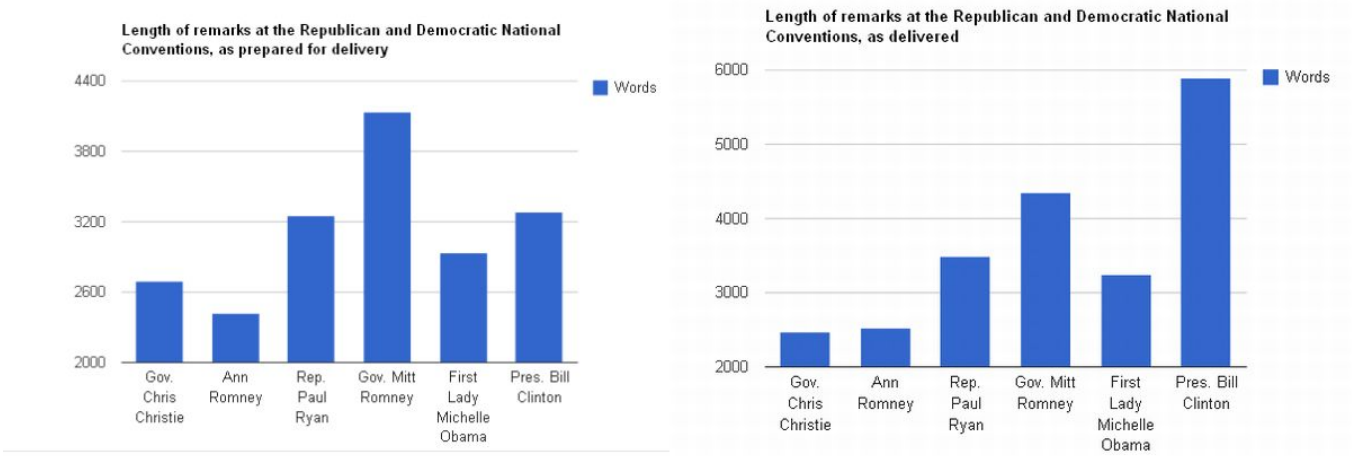


Figure: From **Cliff @ Washington Post**.

Different Order

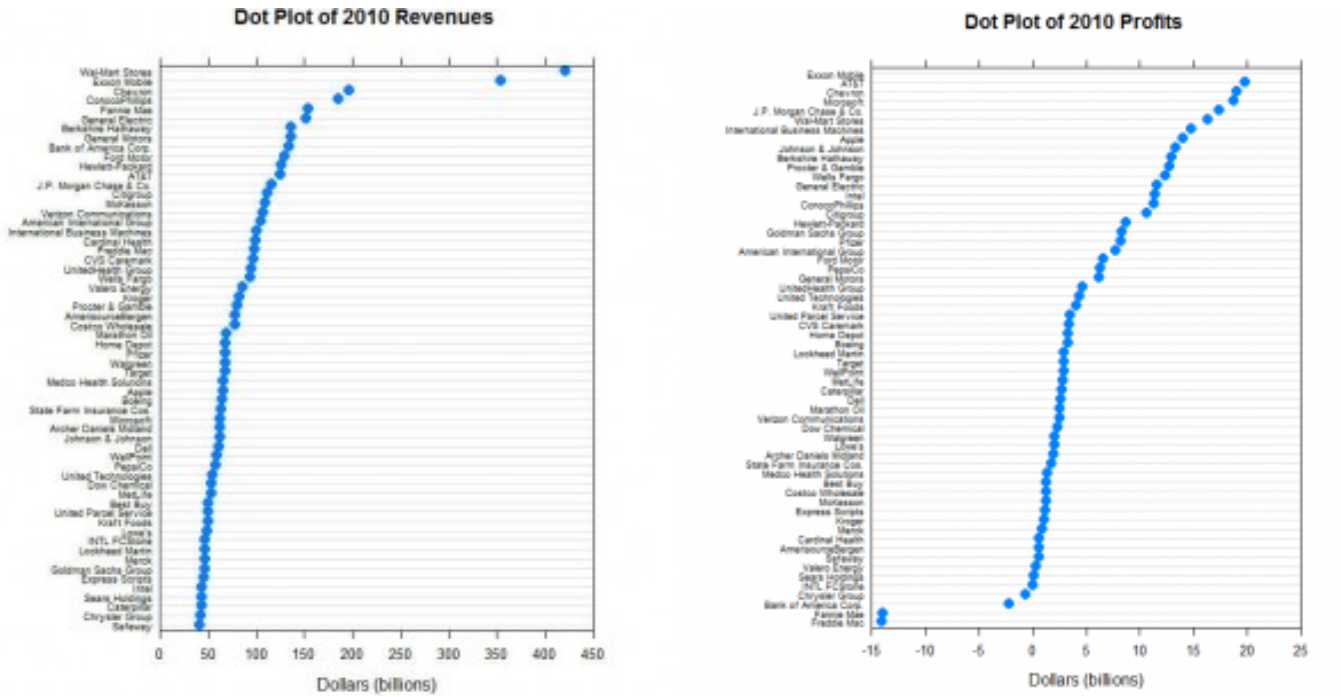


Figure: Via **Robbins @ Forbes**.

Different Order

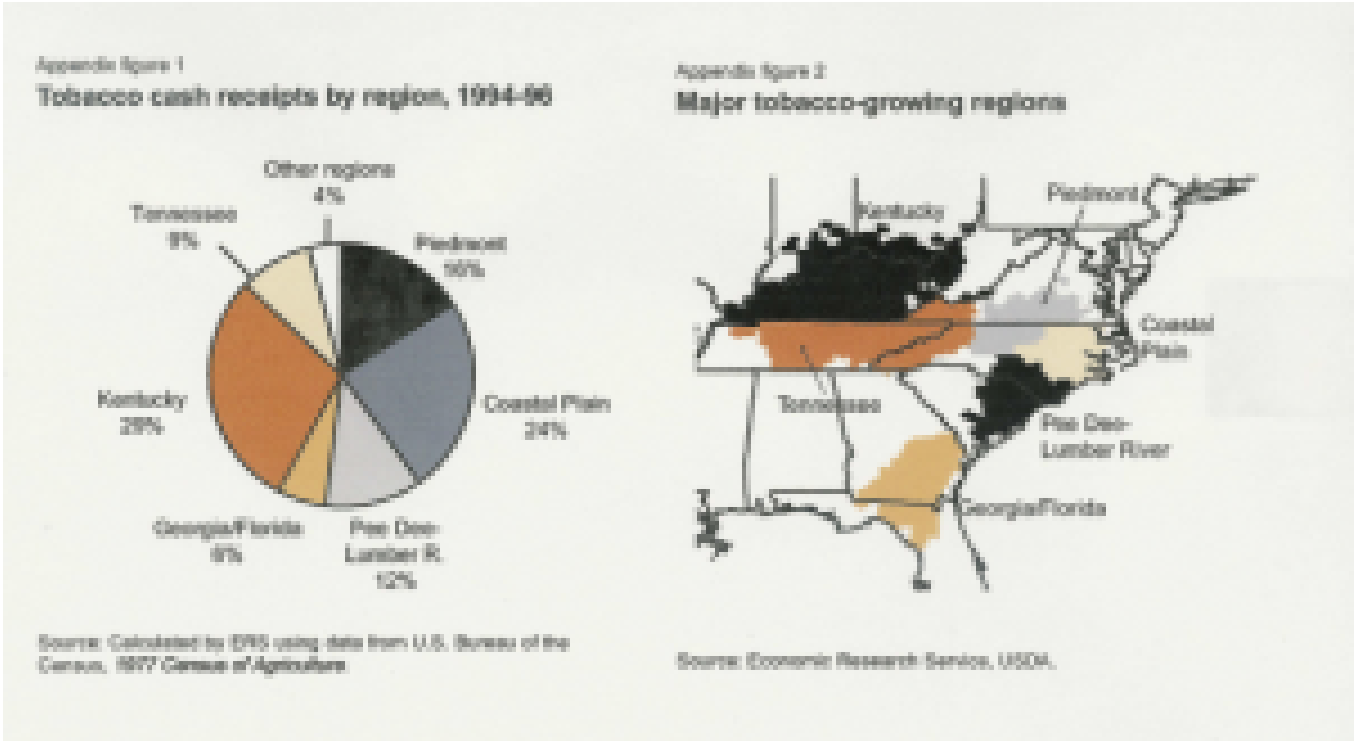


Figure: Via **Robbins @ Forbes**.

Poorly used cumulative graph

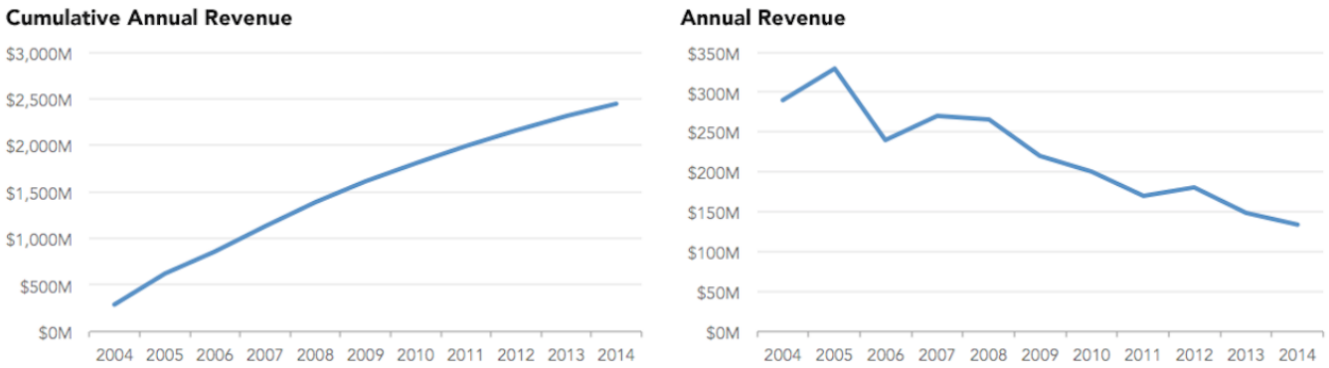


Figure: From **Parikh @ Gizmodo**.

Ignoring conventions and expectations

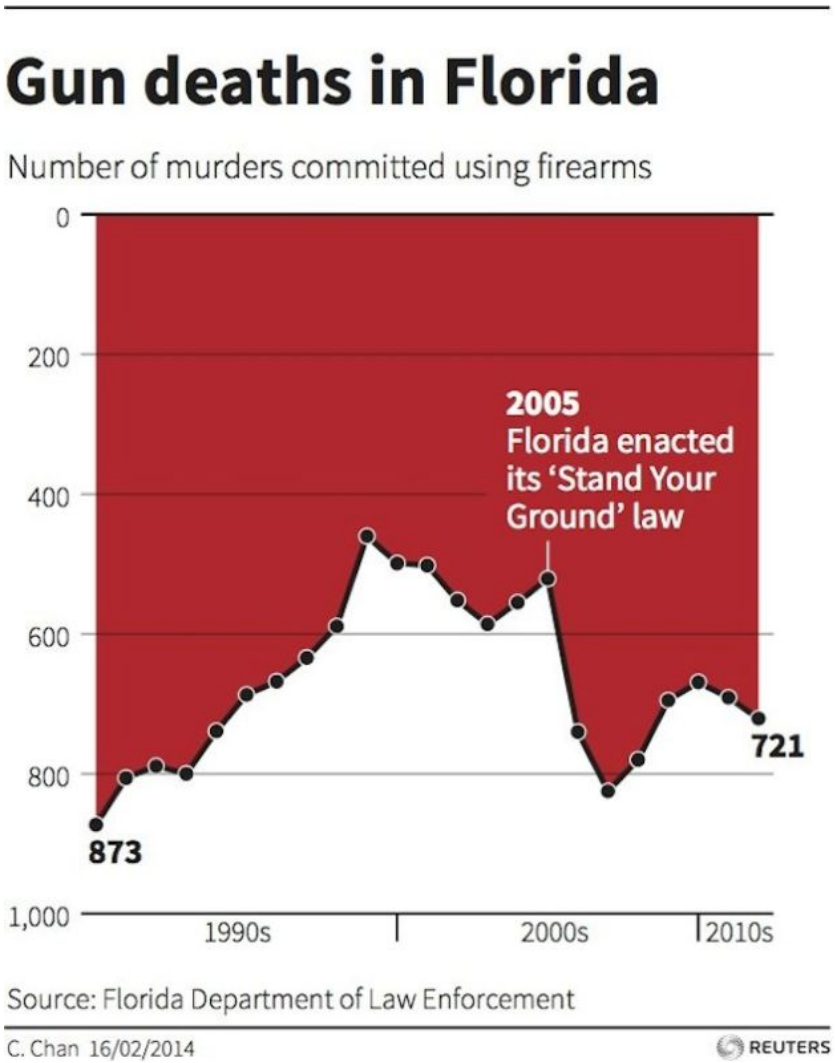


Figure: Via Parikh @ Gizmodo.

Data-Ink

Data-ink

Data-ink is the non-erasable core of a graphic; the non-redundant ink arranged in response to variation in the numbers presented.

Data-Ink Ratio

$$\begin{aligned}\text{Data-ink ratio} &= \frac{\text{data-ink}}{\text{total ink used to print the graphic}} \\ &= \text{proportion of a graphic's ink devoted} \\ &\quad \text{to the non-redundant display of data-information} \\ &= 1.0 - \text{proportion of a graphic that can be erased} \\ &\quad \text{without loss of data-information.}\end{aligned}$$

Examples

Data-Ink

- Lines in a line graph, bars in a bar graph, dots in a scatter plot, etc.
- Labels
- Data values

Non-Data-Ink

- Axes
- Ticks
- Grid lines
- Decorations

Maximize Data-Ink-Ratio

- Depict more data
- Erase non-data-ink
- Erase redundant data-ink

Within reason!

Exercise

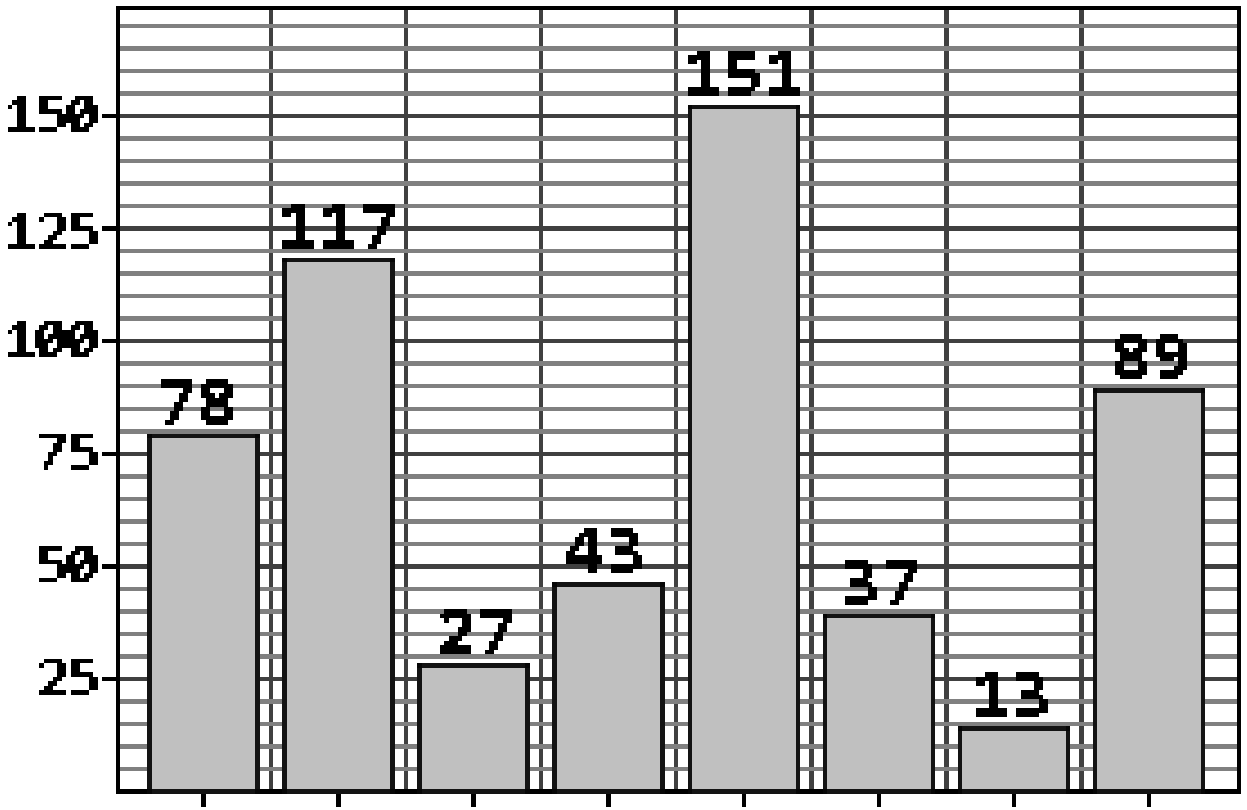


Figure: See VDQI (page 96 and 126-128)

How can we increase the data-ink-ratio?

Exercise 2

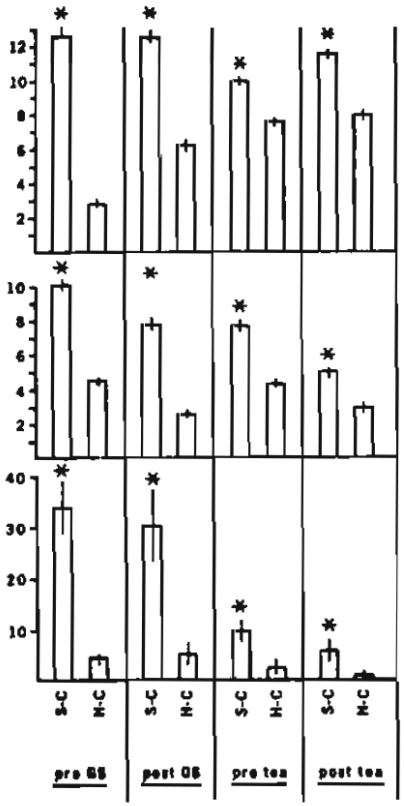


Figure: From Kuznicki & McCutcheon (1979) via VDQI (page 100)

Exercise 2

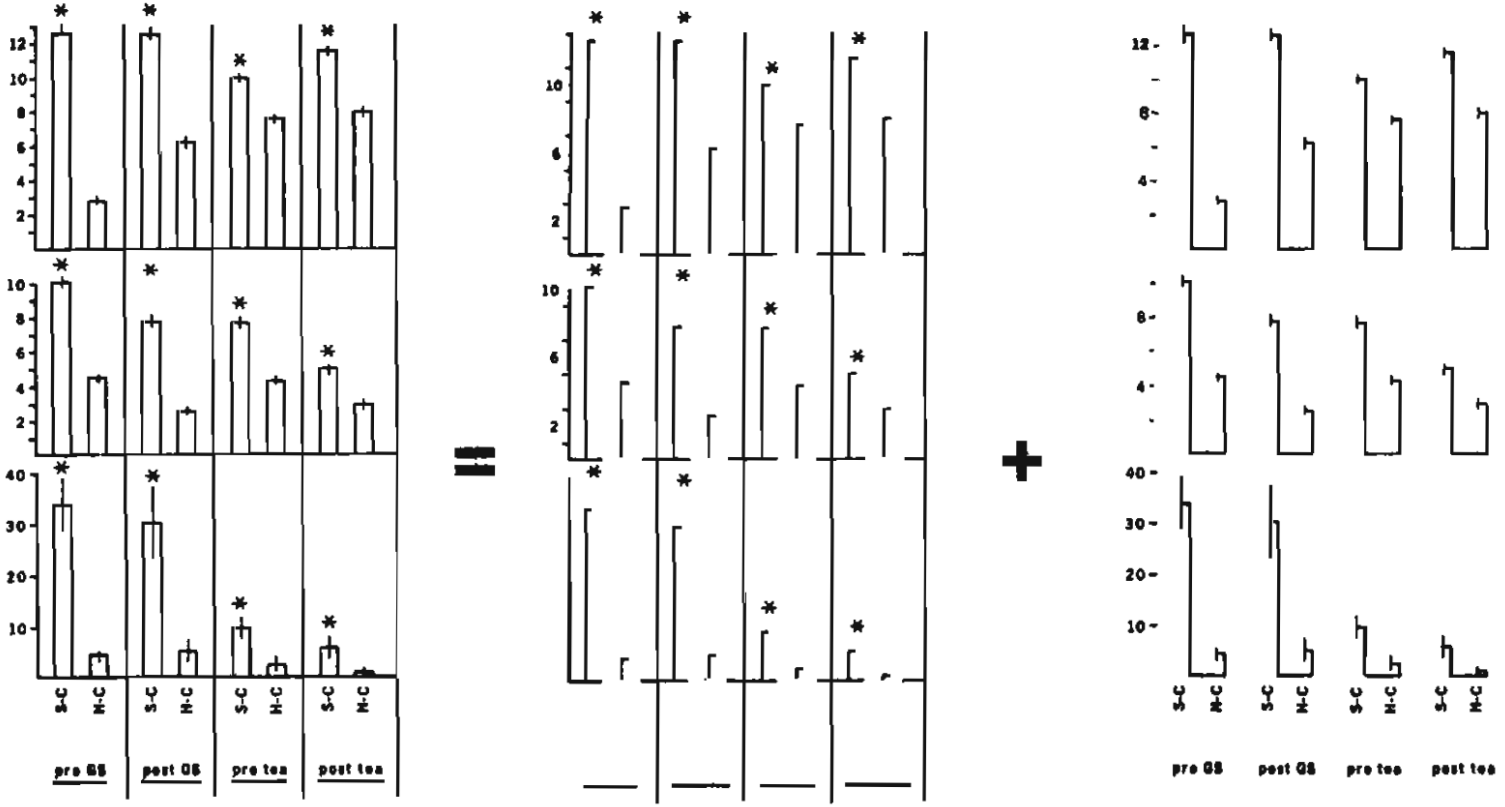


Figure: From VDQI (page 102)

Sparklines

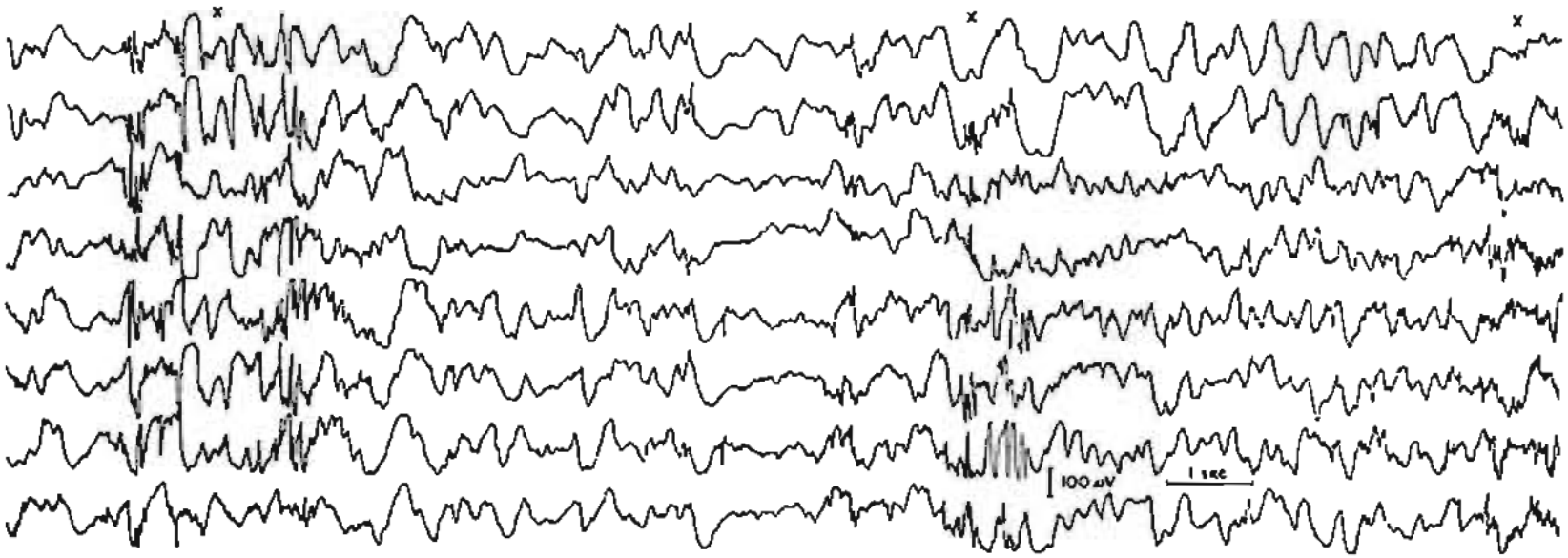


Figure: From Kooi (1971), "Fundamentals of Electroencephalography" via VDQI (page 93)

Boxplots

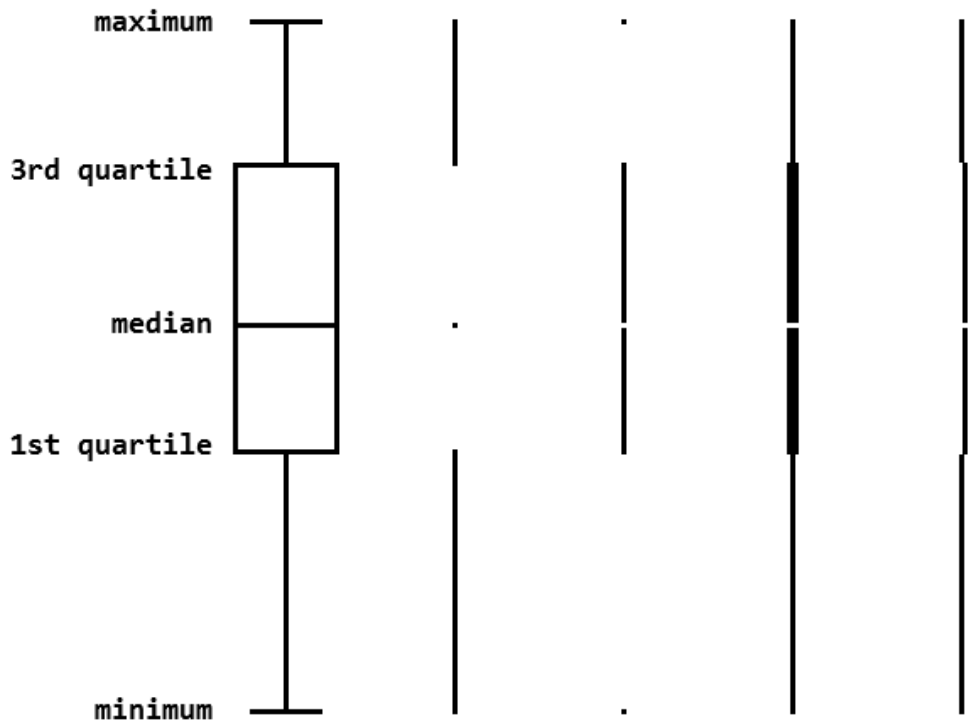


Figure: After VDQI (page 123-125)

Range-Frame

Figure: From VDQI (page 132)

Range-Frame

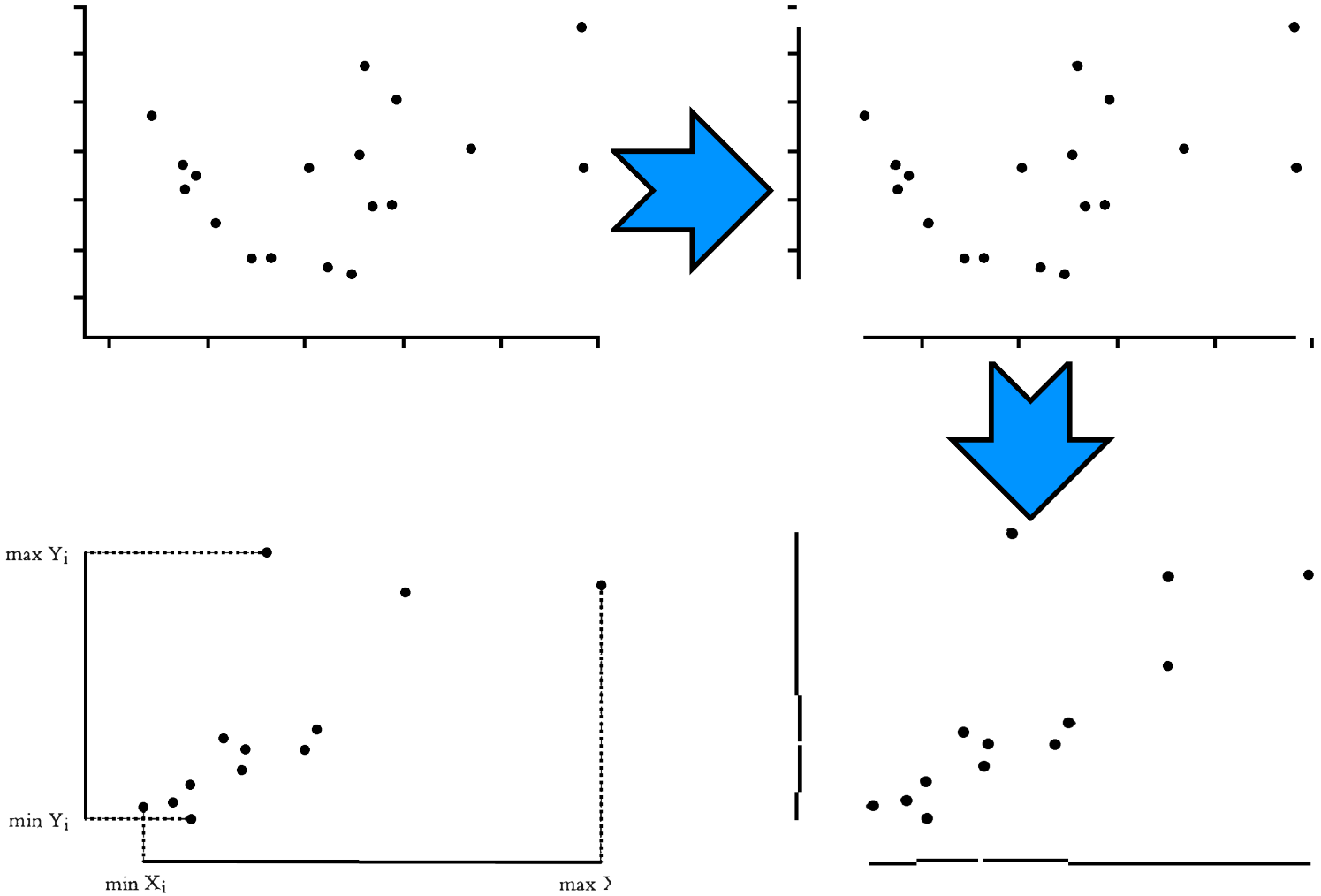


Figure: From VDQI (page 130-132)

Dot-Dash-Plot

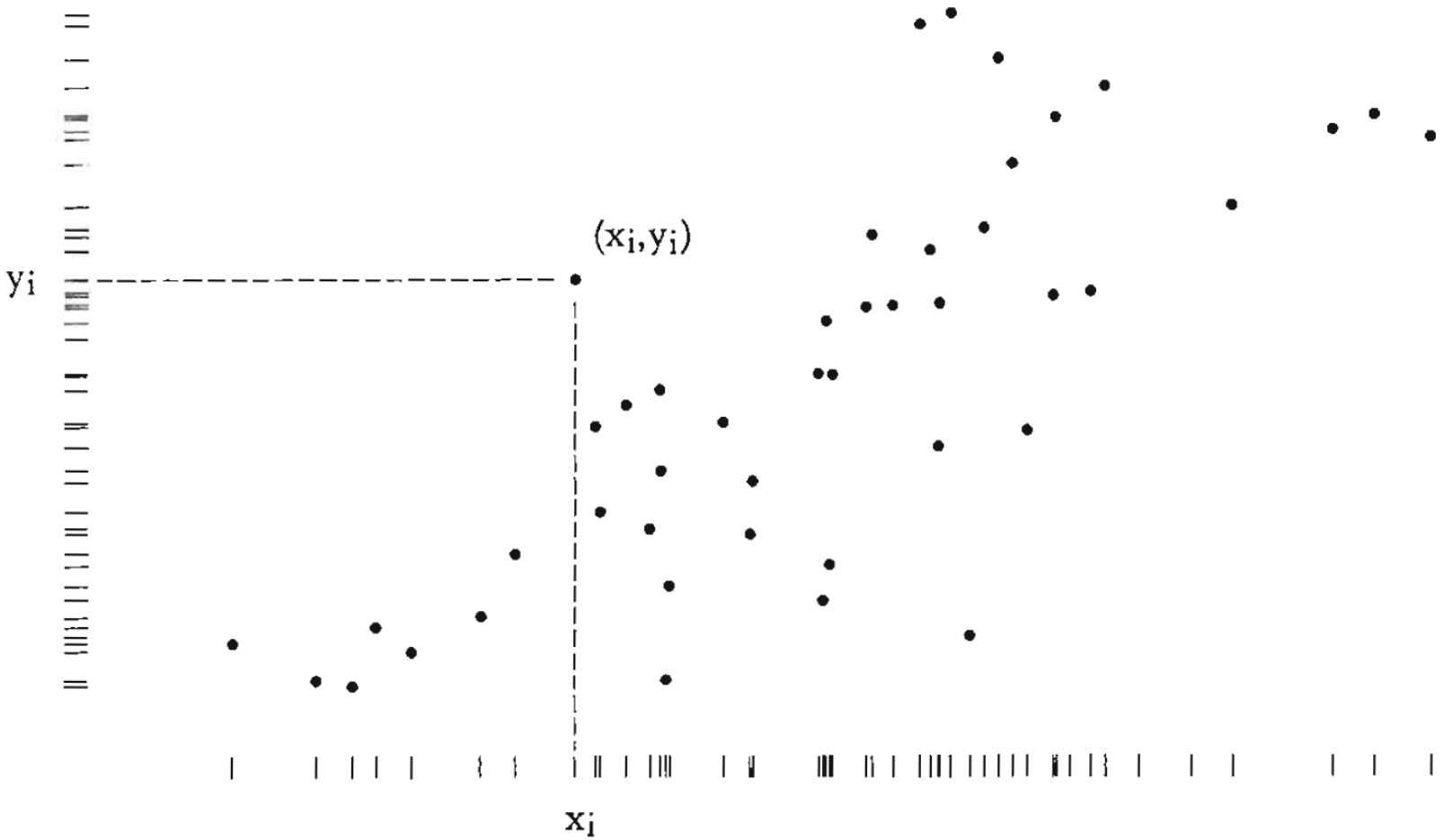


Figure: From VDQI (page 133)

Distribution on axes

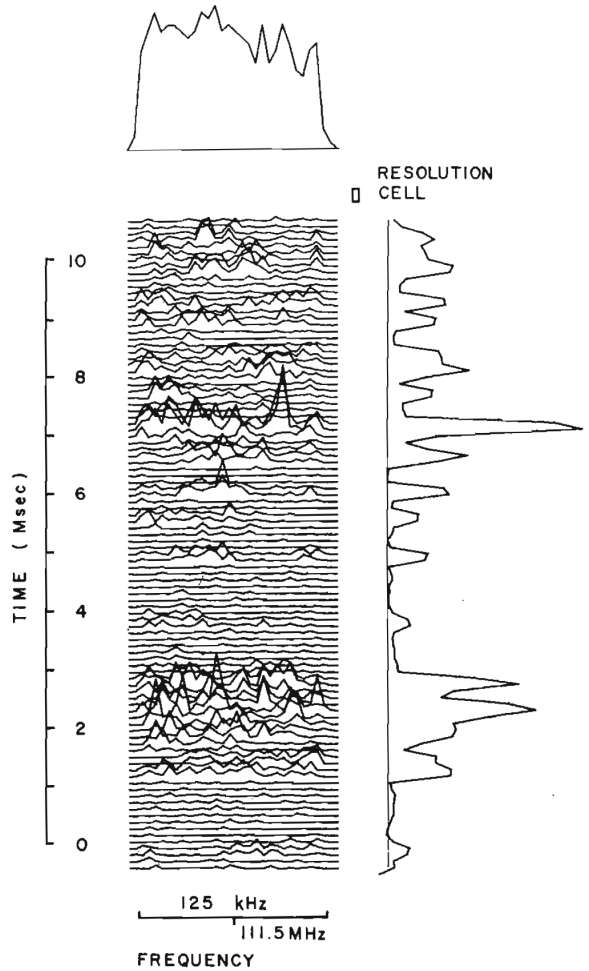


Figure: From Hawkins & Rickett (1975), "Pulsar Signal Processing", p. 108 via VDQI (page 134)

Data Density

$$\text{data density of a graphic} = \frac{\text{number of entries in data matrix}}{\text{area of data graphic}}$$

Maximize Data Density

- Depict more data
- Shrink the graphic
- Use multifunctioning graphical elements

Within reason!

Small Multiples

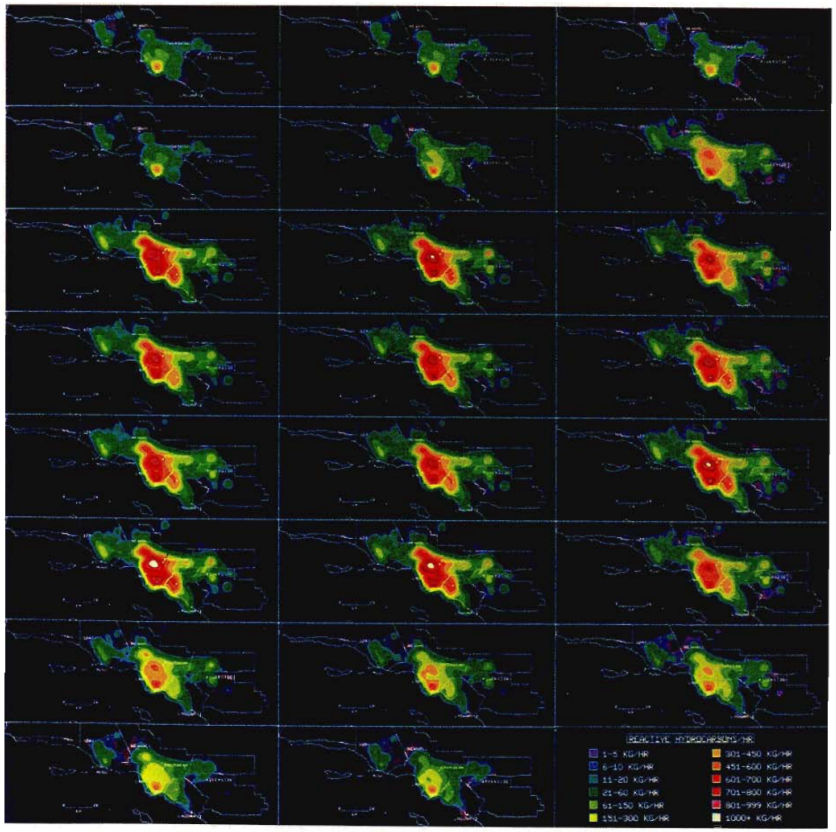


Figure: From video based on McRae, Goodin & Seinfeld (1982), "Development of a Second-Generation Mathematical Model for Urban Air Pollution" via VDQI (page 170)

Multifunctioning Graphical Elements

Advice

Mobilize every graphical element, perhaps several times over, to show the data.

Stem-and-leaf display

A stem-and-leaf display let's you show fairly detailed distribution information in the shape of a histogram.

Example (Data)

```
37, 33, 33, 32, 29, 28, 28, 23,
22, 22, 22, 21, 21, 21, 20, 20,
19, 19, 18, 18, 18, 18, 16, 15,
14, 14, 14, 12, 12, 9, 6
```

Example from **Lane @ OnlineStatBook**.

Example (S&L display 1)

```
3|2337
2|001112223889
1|2244456888899
0|69
```

Example (S&L display 2)

```
3|7
3|233
2|889
2|001112223
1|56888899
1|22444
0|69
```

Number Plots

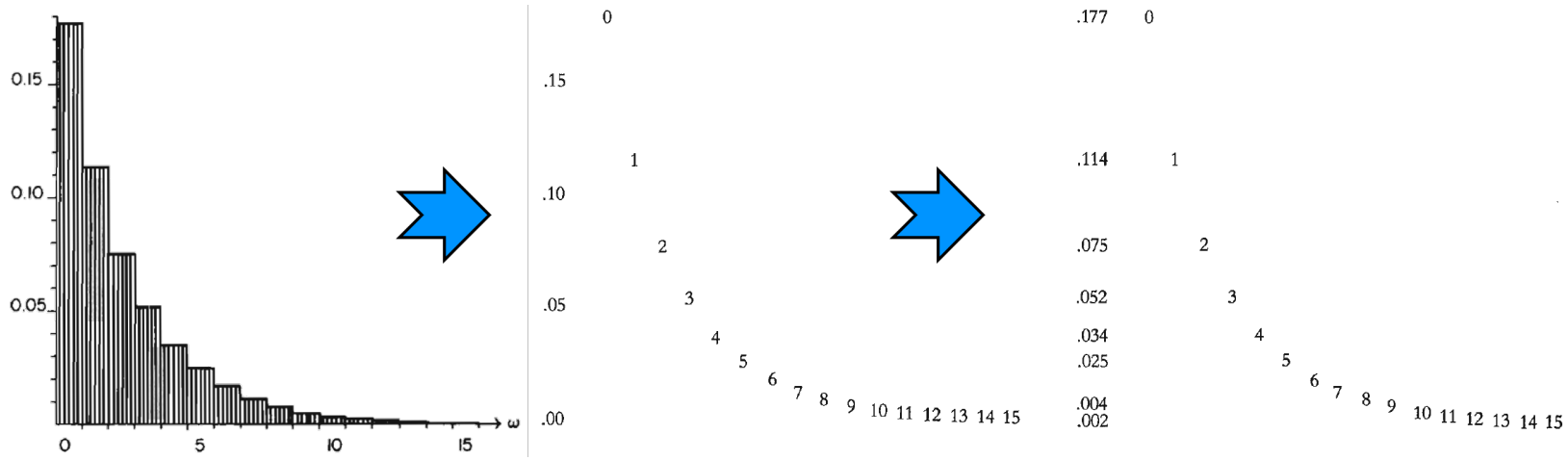


Figure: From stylesheet of the Journal of the American Statistical Association (left) and VDQI (page 150-151)

Quiver Plot

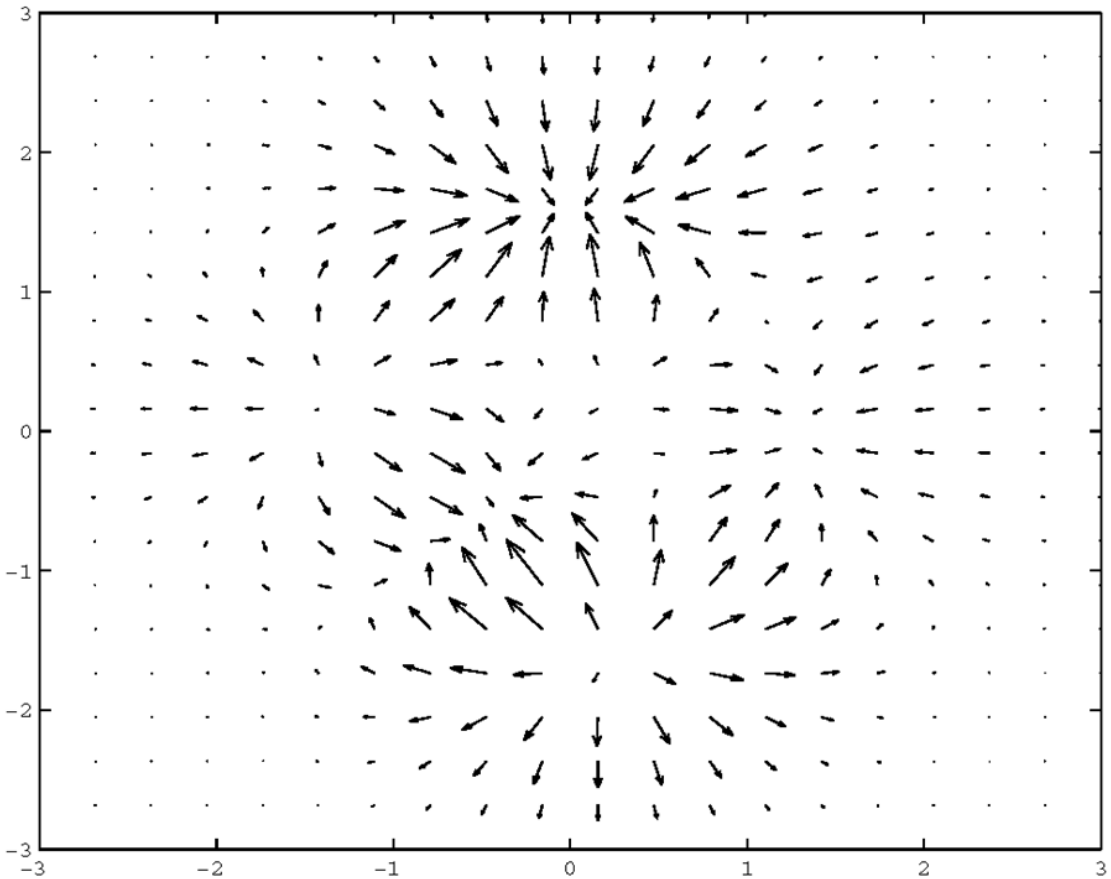


Figure: From **what-when-how**

Chernoff Faces

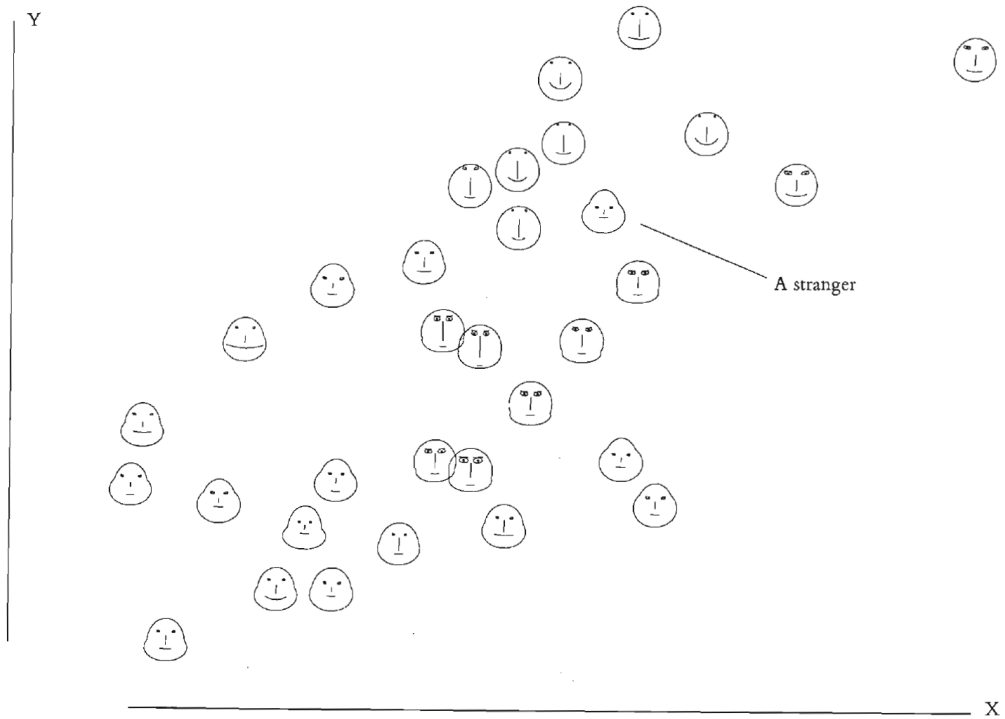


Figure: From Wainer & Thissen (1981), "Graphical Data Analysis" via VDQI (page 142)

See also Chernoff (1973), "The Use of Faces to Represent Points in k-Dimensional Space Graphically" and **Wikipedia**.

Questions?