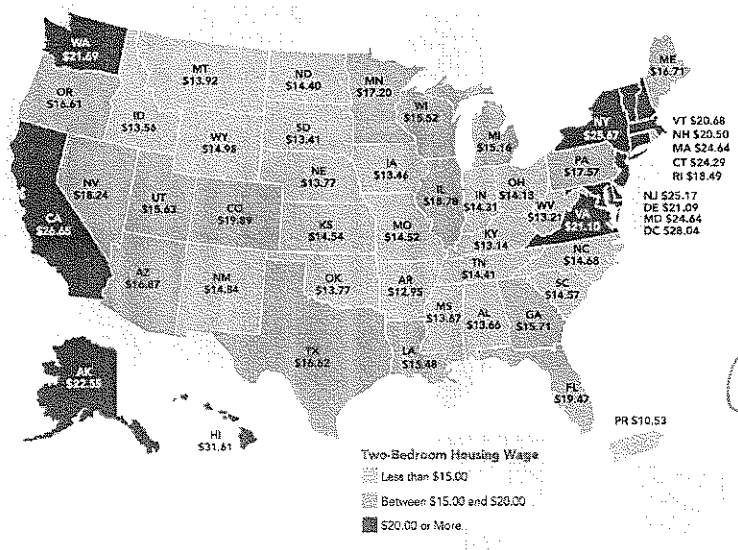
Today: Critiquing Statistical Graphics
Introduction to Data
Graphics Principles
Friday: Introduction to R and Reproducibility
Monday: No class (Martin Luther King, Jr. Day)

Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

January 13, 2016

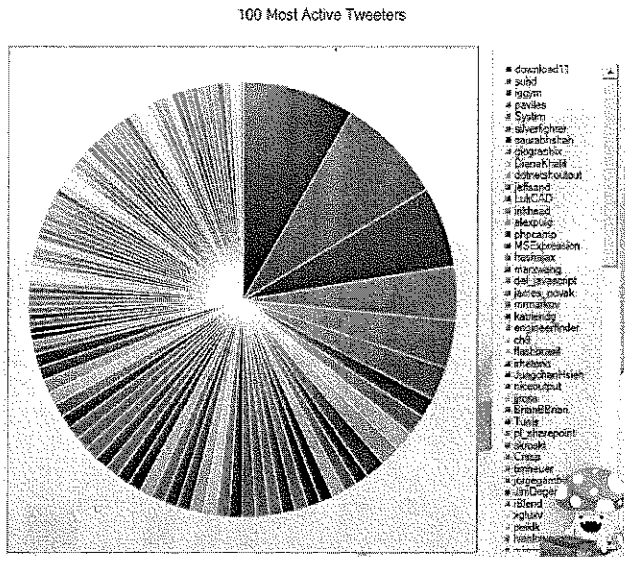
Hourly Wages to Afford Two-Bedroom Apartment



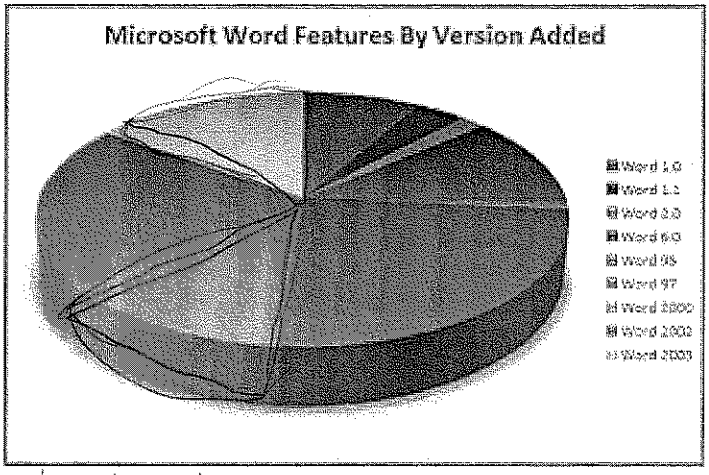
Improve w/ color scale OR just more options

Top 100 Tweeters

Pie = bad



Never Make 3-D Pie Charts



distortion

What Is Data?

Information organized in some fixed easy-to-understand way

Ex) Tweets

field goal attempts $\sum_{\text{missed}}^{\text{made}}$ → Sat sh %

Temperature

Censuses / Surveys → collect info on population, demographics, etc

How Do We Describe Data? mean, median, mode

Two measurements used to describe datasets:

n = # of obs. people, subjects, objects, etc → rows

p/d = # of covariates, variables, questions, etc → columns

Data is usually in matrix form:

		V_1	V_2	...	V_p
O_1	X_1				
O_2	X_2				
O_3	X_3				
⋮	⋮				
O_n	X_n				

rows → observations

single row has all answers to all Qs from a single person

columns → variables

single column has all answers from a single Q for all people

Types of Data → String, integer

Categorical → qualitative, describes qualities of obs

Ordered: strongly disagree, disagree, neutral, agree, SA
- educ. level

unordered: "nominal" → race, colors (sort of)
names / general text, gender

Continuous:

↳ real-valued, quantitative, numerical data

Notation $X = \{X_1, X_2, \ldots, X_d\}$,

data/variable

$X_i \in \mathbb{R}$

$X \in \mathbb{R}^d$

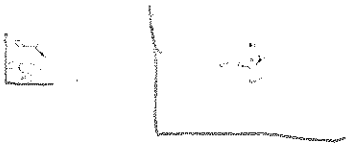
→ double, int, float

Graphics and Their Goals (from Tufte) → father of graphics

Graphics: visually display measured quantities by combining points, lines, coordinate system, numbers, symbols, words, shading, color

Goals: show data!

- induce viewer to think about substance, not graphical methodology
- avoid **distorting** the data
- present numbers in small space
- make large, complicated datasets more coherent
- encourage comparison of different pieces of data
- reveal data at several levels of detail
- describe, explore, tabulate, or decorate
- be closely integrated with statistical/verbal descriptions of dataset



Graphs that do not meet these goals are not successful

Graphs leading viewers to make misleading conclusions should be avoided

"Decorating" / Data-Ink

Graphics should not draw the viewer's attention away from the data.
Extras get in the way.

Note: Decoration does not refer to appropriate graph labeling.
Labels should always be clear, detailed, and thorough.
Label key parts of the data. Add text explanations if necessary.

Data Ink should primarily present information about the data:
the non-erasable, non-redundant core of a graphic

Tufte suggests using the *data-ink ratio*:

$$DI = \frac{\text{data ink}}{\text{total ink on graphic}}$$

% of ink devoted to non-redundant / useful information.

11/13

Ideally → Maximize DI (max = 1)
won't quite get to 1, because of axes, grid lines, etc

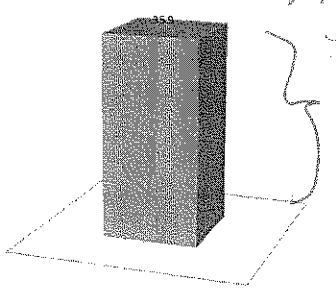
"Decorating" / Data-Ink

Two ways to increase the proportion of data-ink:

Remove non-data-ink:
↳ Ink that does not depict statistical info

In class Wednesday 1/20, hands on map graphic

Remove redundant data-ink: → Ink that is unnecessarily redundant/ repetitive.

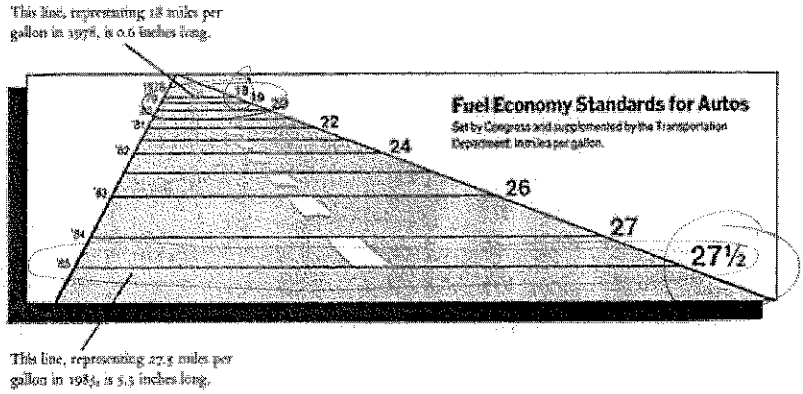


Indications of height:
1) height of front-left line on bar
2) height of front-right line on bar
3) ---- back-right ----
4) position of number
5) value of number
6) position of top line (front)
7) position of top line (back)
8) height of shading
9) etc....

Distortion

Visual representation of data is inconsistent with numerical representation

In other words: **The graph doesn't match the data**



Optimal: $LF \approx 1$

$LF > 1 \rightarrow$ enhance the effect

$LF < 1 \rightarrow$ decrease the effect

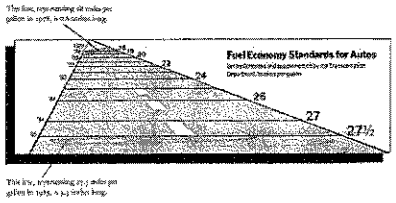
Lie Factor

Tufte suggests optimizing the Lie Factor:

$$LF = \frac{\text{size of "effect" in graphic}}{\text{in data}}$$

"effect" = change in amount of some feature or variable

Fuel Economy Standards Example:



Actual % increase (in data)

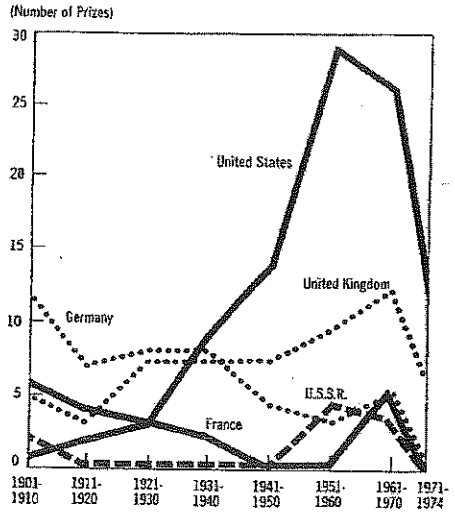
$$\frac{|27.5 - 18|}{18} \approx 0.528$$

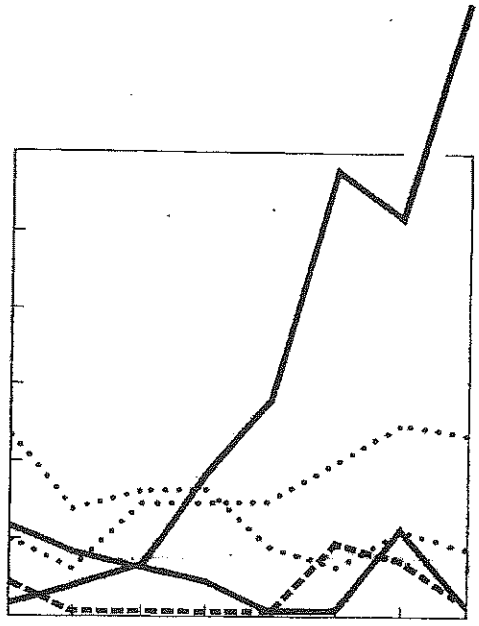
Graphical increase (in graph)

$$\frac{|5.3\text{ in} - 0.6\text{ in}|}{0.6\text{ in}} = 7.83$$

$$LF = \frac{7.83}{0.528} = 14.83$$

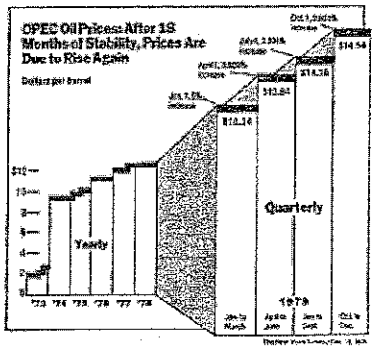
Nobel Prizes Awarded in Science, for Selected Countries, 1901-1974





Graph Principles (Tufte)

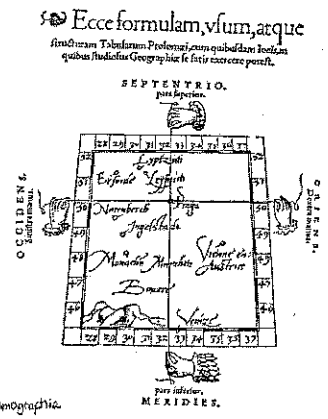
Minimize **Design Variation:** Changes in Design may imply Changes in Data



New York Times, December 19, 1978, p. Ew7.

Graph Principles (Tufte)

Maximize **Data-Ink Ratio:** Ink in graphic should primarily show data



1548 Cosmographia