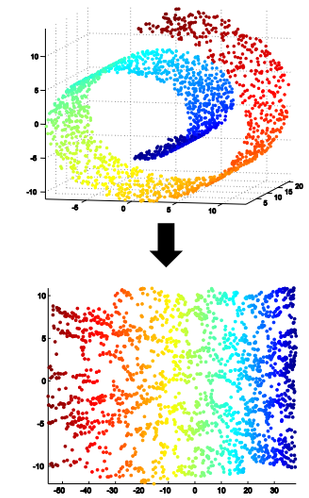
# Where Are We? Feedback So far?

Collection

RottenTomatoes API, SQL refresher

Cleaning

OpenRefine



Integration

Analysis

Dimensionality reduction:
PCA, MDS, LDA, Isomap, t-SNE

Visualization

Vis 101, D3, Tableau [HW2]

Presentation

Dissemination

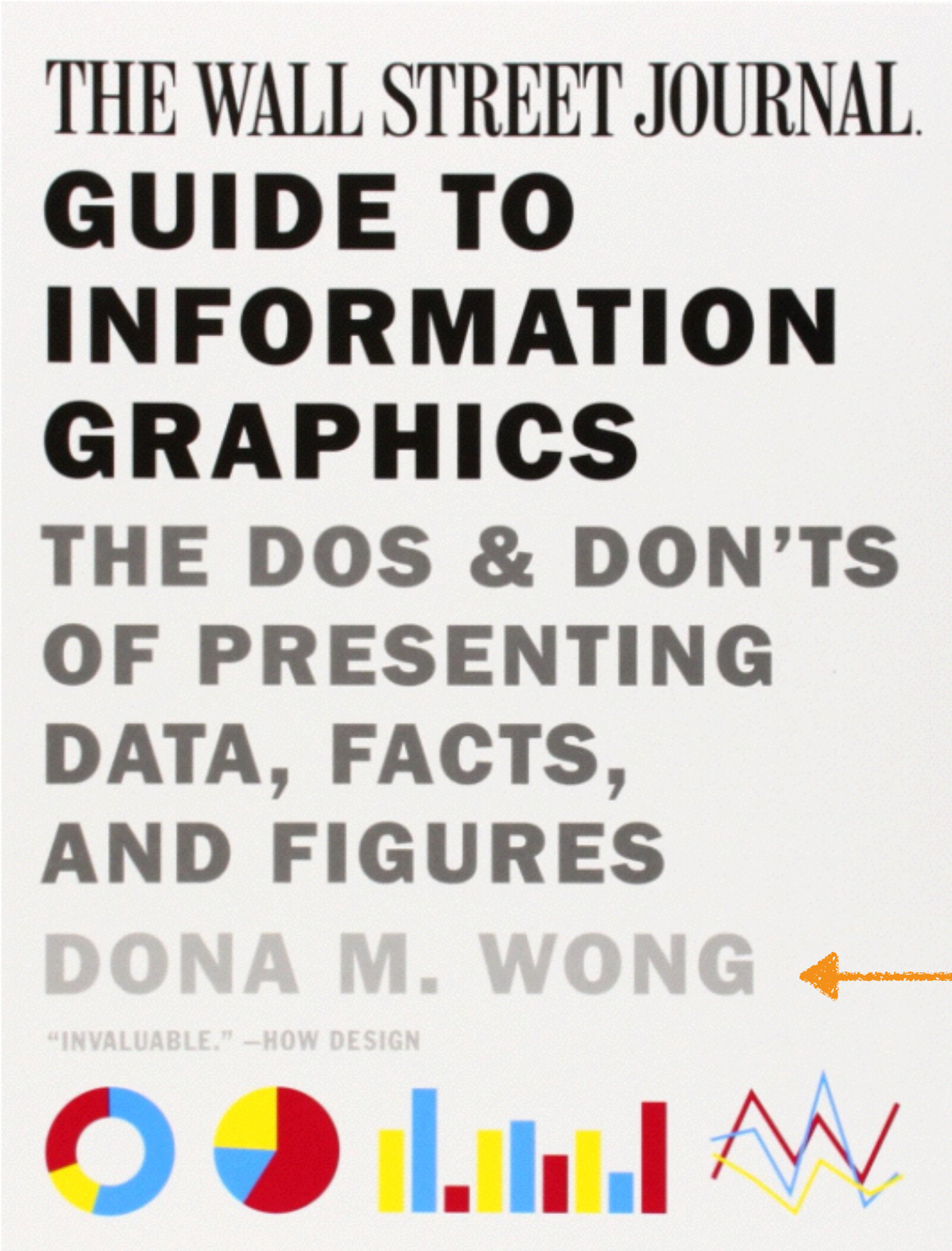
HW2 will add “Expected Time to Spend”

CSE 6242 / CX 4242

# 1. How to Fix Vis Issues?
# 2. Class Project

**Duen Horng (Polo) Chau**
Georgia Tech

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos



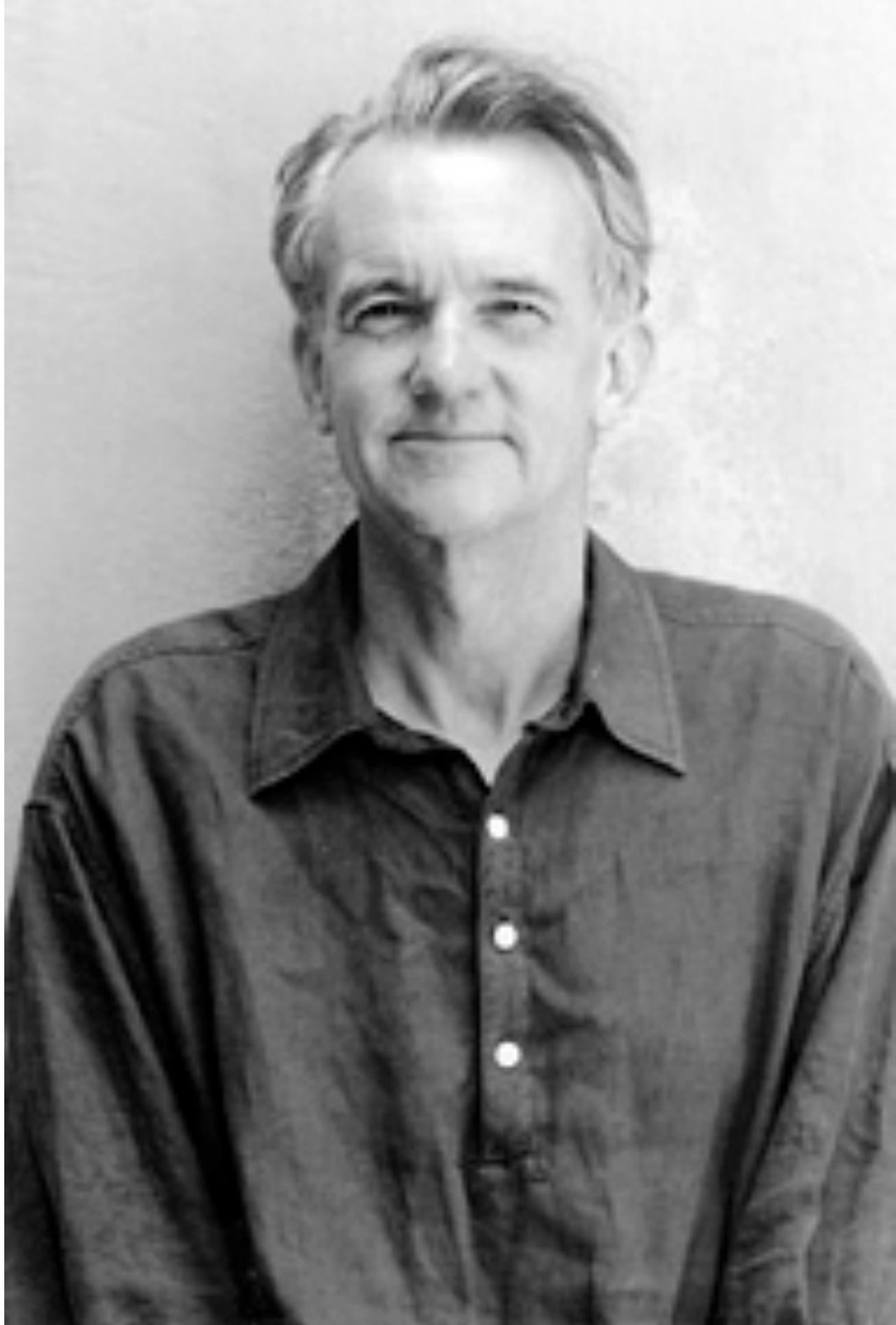

Student of Edward Tufte

# Edward Tufte

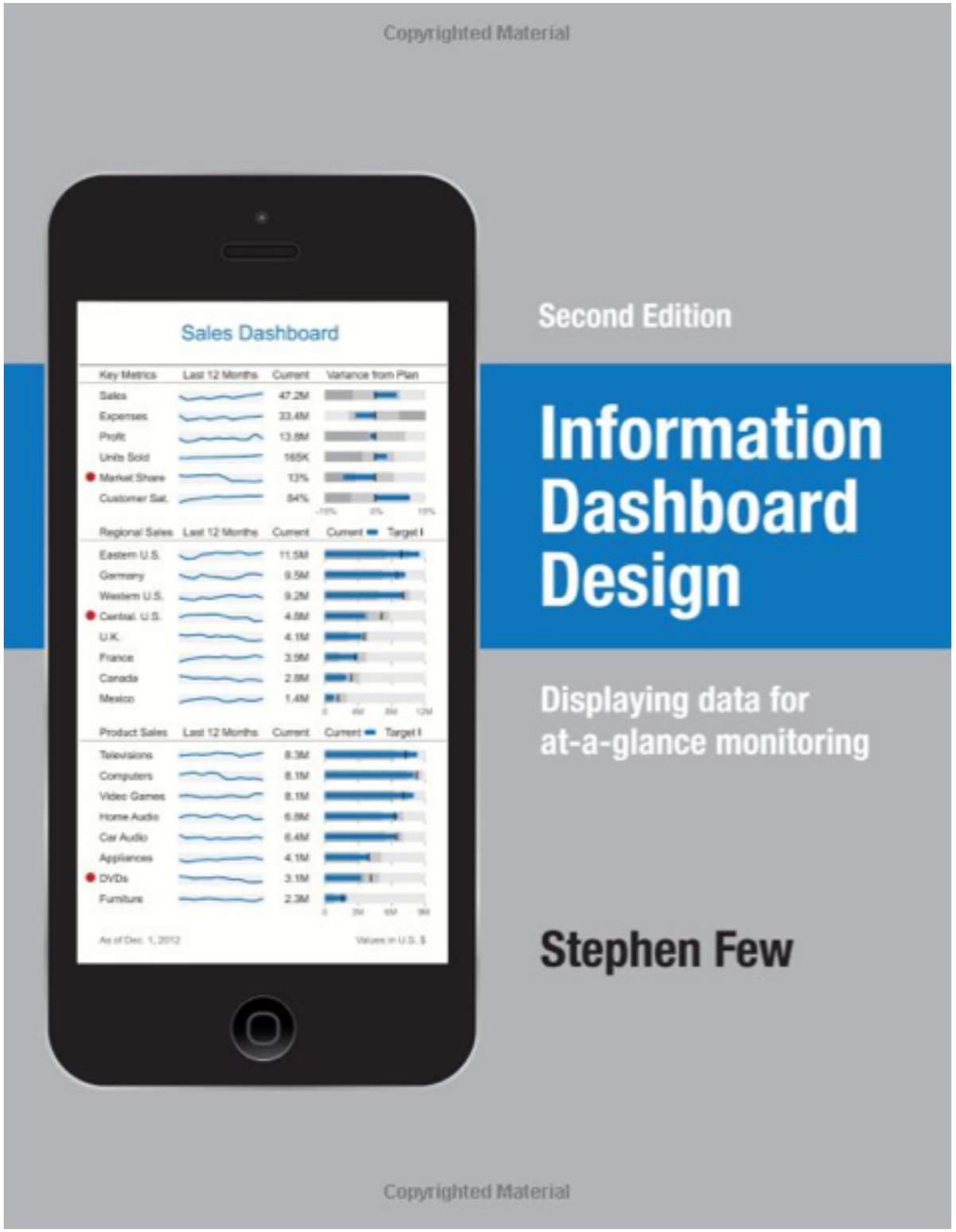
An American statistician and professor emeritus of political science, statistics, and computer science at Yale University.

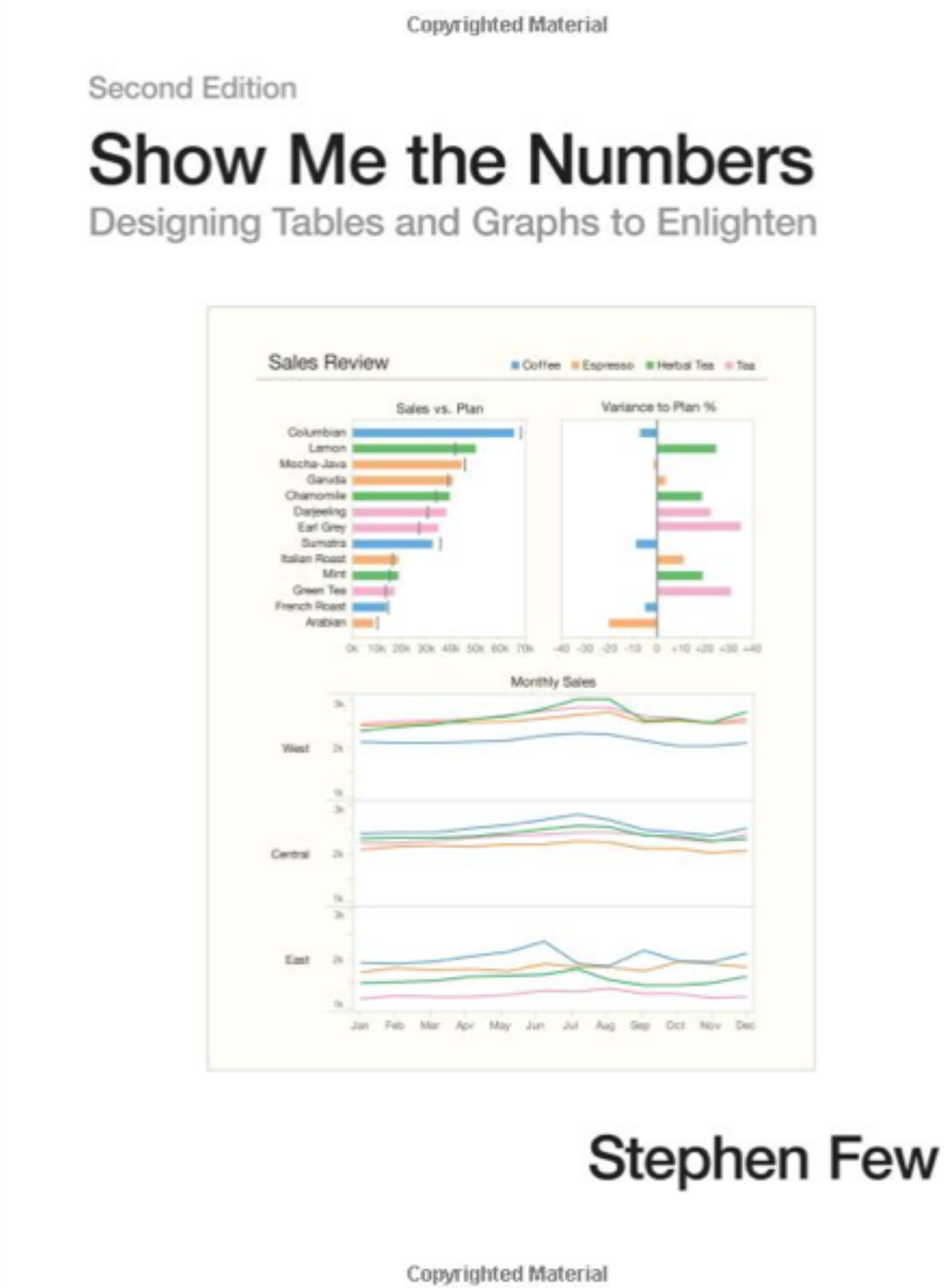
He is noted for his writings on information design and as a pioneer in the field of data visualization.

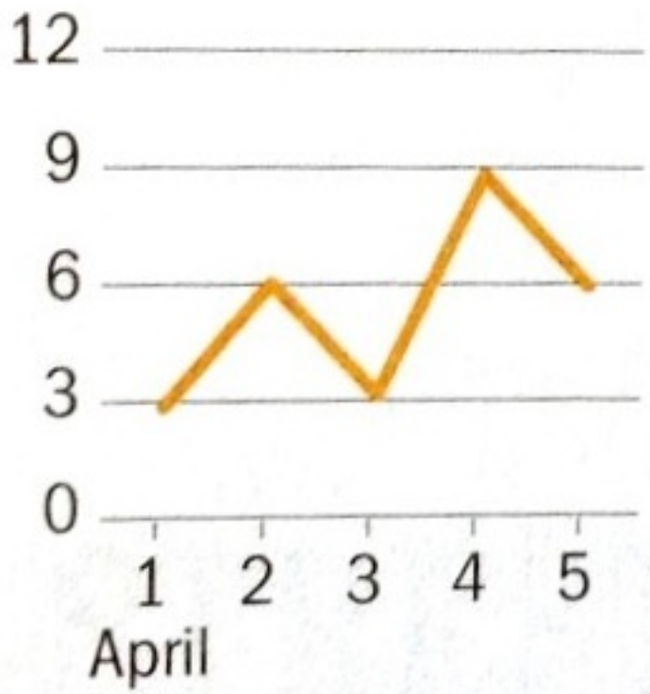
-Wikipedia

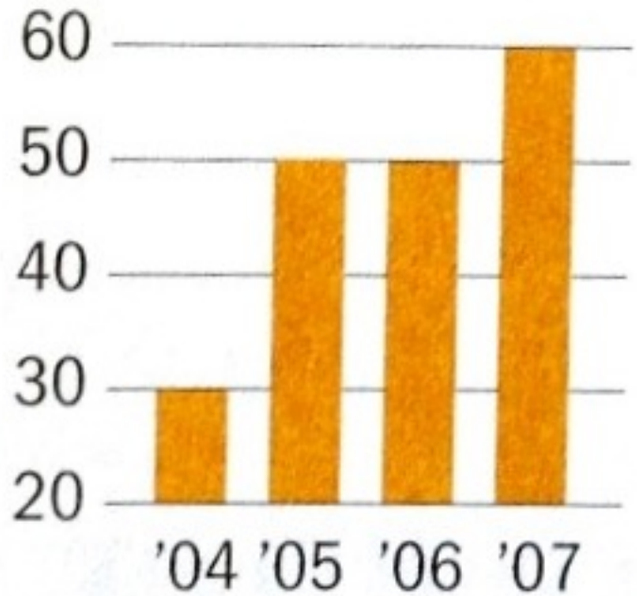


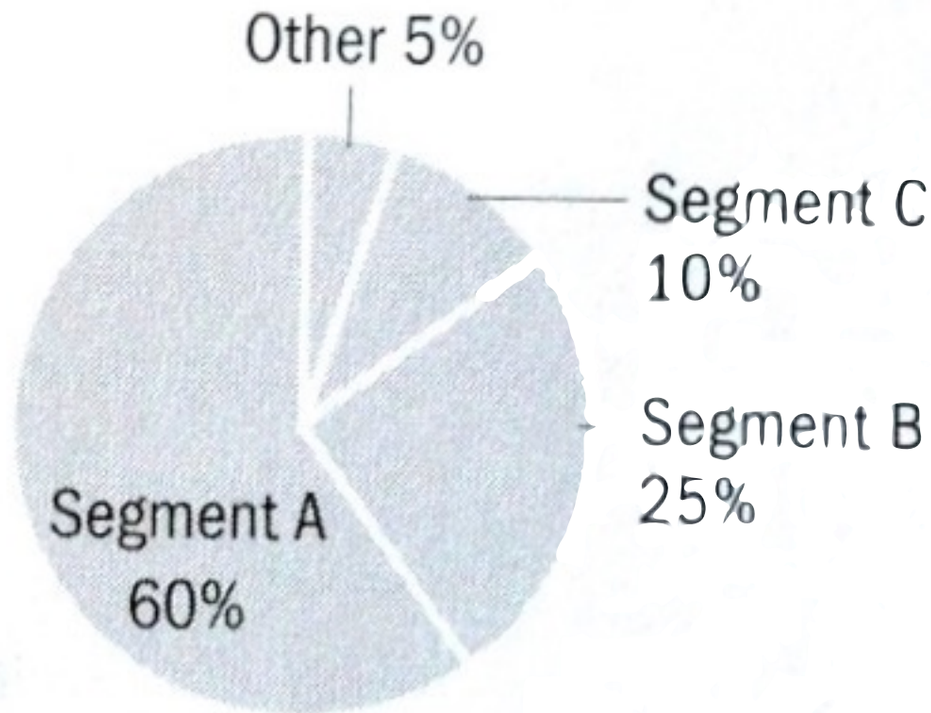
# Also Highly Recommended:



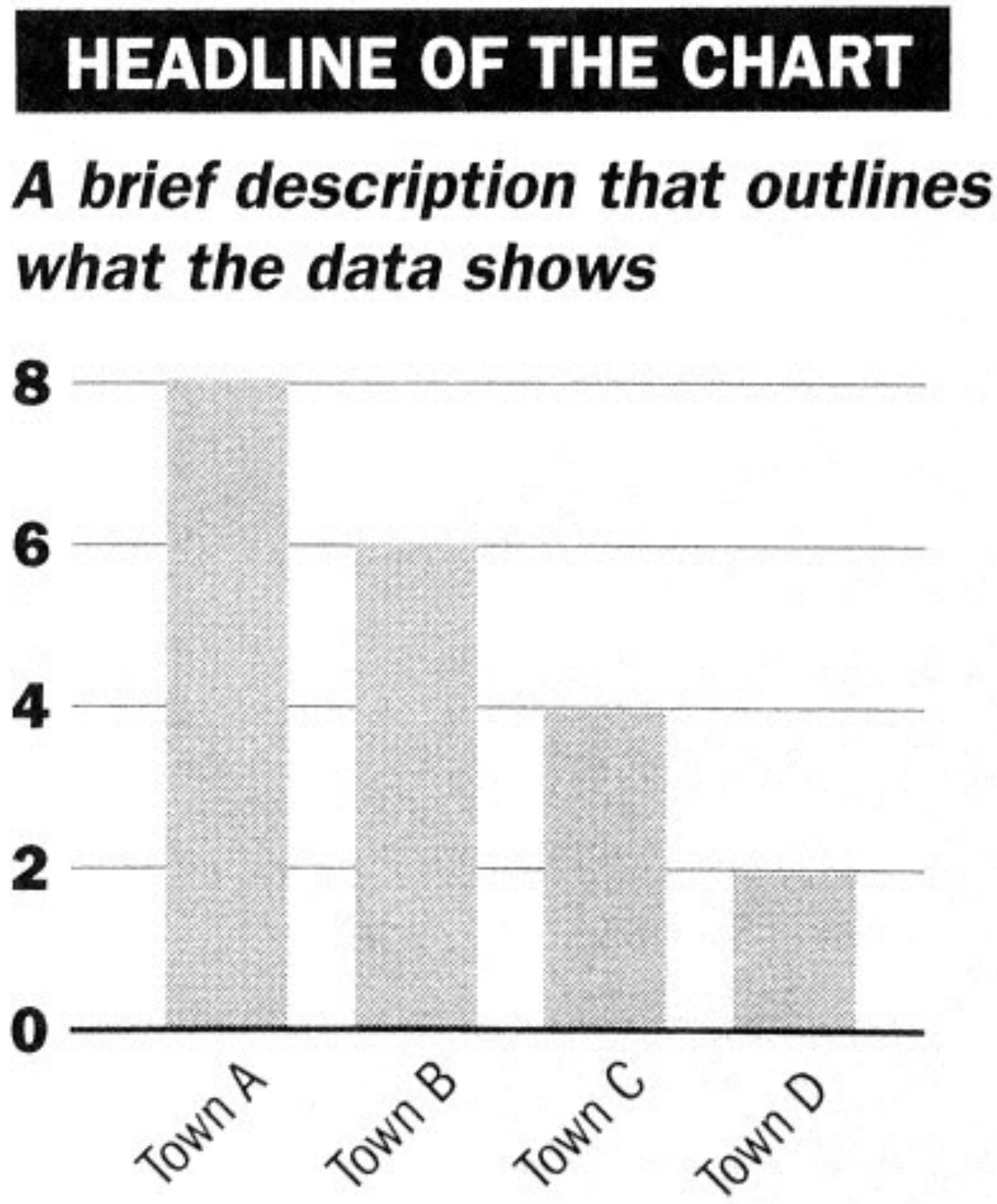





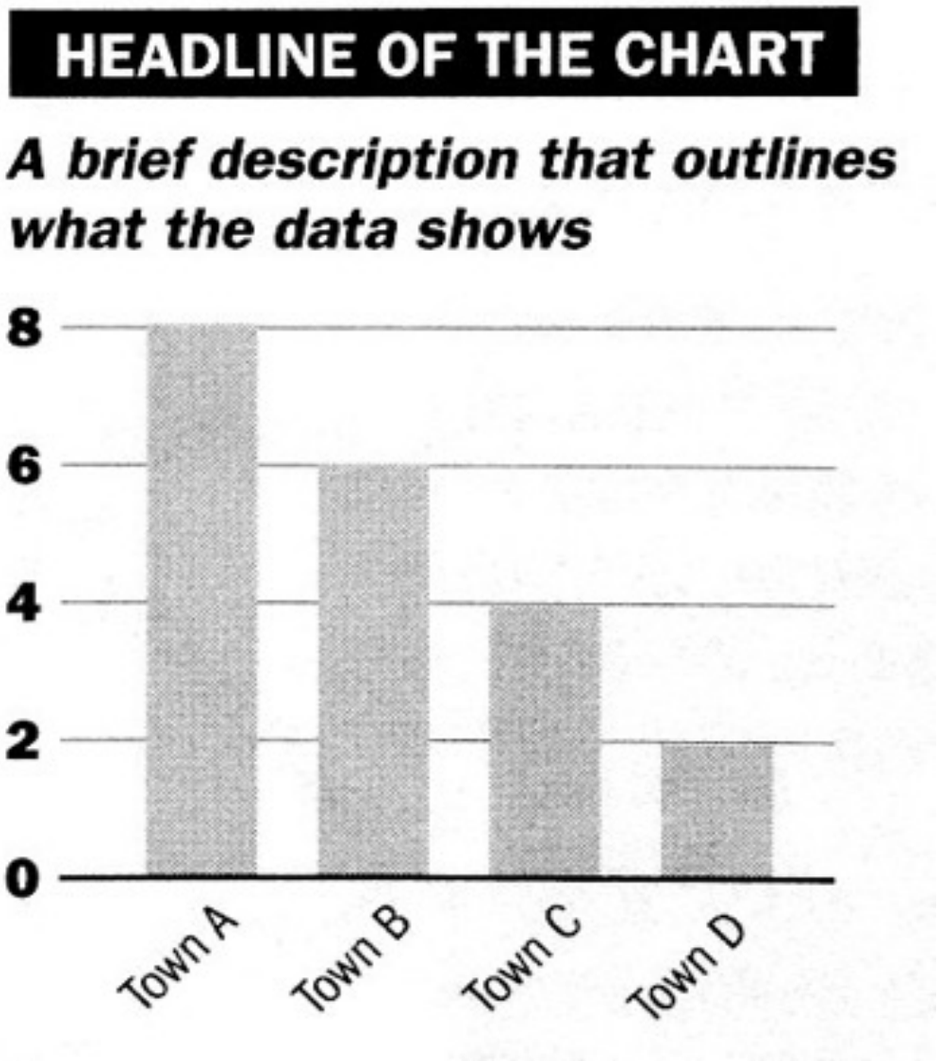



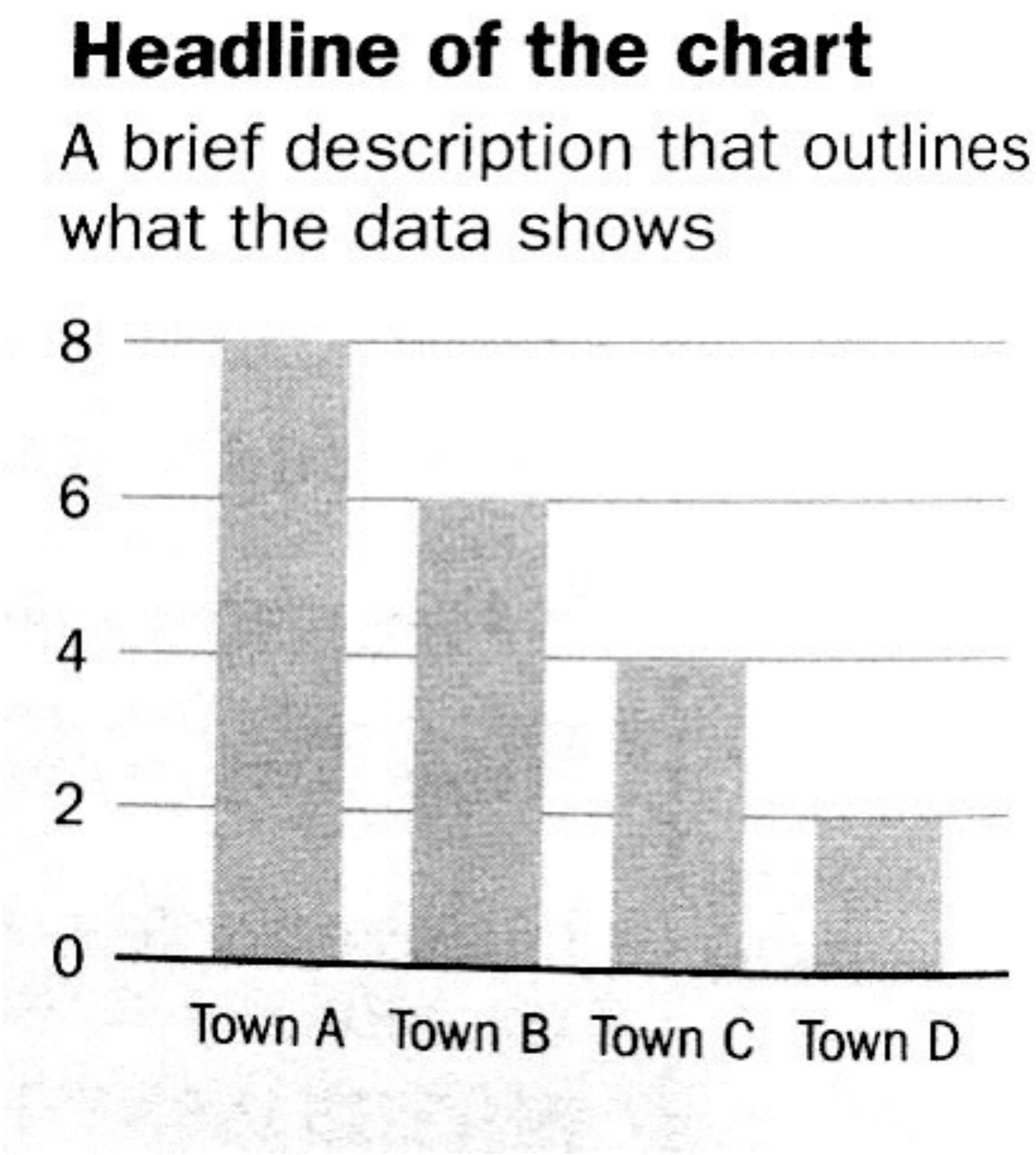




Good charts? How would you improve them?




How about this one?




Which is better?

# Tables

## What are they good for?

| Name | Data | Data | Data |
|---|---|---|---|
| Company A | 0.0 | 0.0 | 0.0 |
| Company B | 0.0 | 0.0 | 0.0 |
| Company C | 0.0 | 0.0 | 0.0 |
| Company D | 0.0 | 0.0 | 0.0 |

Can you improve this table's design?

# “When everyone is special, no one is special”

http://www.youtube.com/watch?v=A8I9pYCl9AQ

| Name | Data | Data | Data |
|---|---|---|---|
| Company A | 0.0 | 0.0 | 0.0 |
| Company B | 0.0 | 0.0 | 0.0 |
| Company C | 0.0 | 0.0 | 0.0 |
| Company D | 0.0 | 0.0 | 0.0 |

| Name | Data | Data | Data | Data | Data | Data |
|---|---|---|---|---|---|---|
| Company A | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Company B | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Company C | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Company D | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Company E | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Company F | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Company G | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Company H | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

A lot of “chart junk”.
Low “data to ink” ratio (Edward Tufte)

| Name | Data | Data | Data | Data | Data | Data |
|---|---|---|---|---|---|---|
| Company A | 0.0 | 0.0 | 0.0 | 12.0 | 0.0 | 0.0 |
| Company B | 0.0 | 0.0 | 0.0 | 11.0 | 0.0 | 0.0 |
| Company C | 0.0 | 0.0 | 0.0 | 10.0 | 0.0 | 0.0 |
| Company D | 0.0 | 0.0 | 0.0 | 9.0 | 0.0 | 0.0 |
| Company E | 0.0 | 0.0 | 0.0 | 8.0 | 0.0 | 0.0 |
| Company F | 0.0 | 0.0 | 0.0 | 7.0 | 0.0 | 0.0 |
| Company G | 0.0 | 0.0 | 0.0 | 6.0 | 0.0 | 0.0 |
| Company H | 0.0 | 0.0 | 0.0 | 5.0 | 0.0 | 0.0 |
| Company I | 0.0 | 0.0 | 0.0 | 4.0 | 0.0 | 0.0 |
| Company J | 0.0 | 0.0 | 0.0 | 3.0 | 0.0 | 0.0 |
| Company K | 0.0 | 0.0 | 0.0 | 2.0 | 0.0 | 0.0 |
| Company L | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |

Better? High “data to ink” ratio

# Aligning Numbers

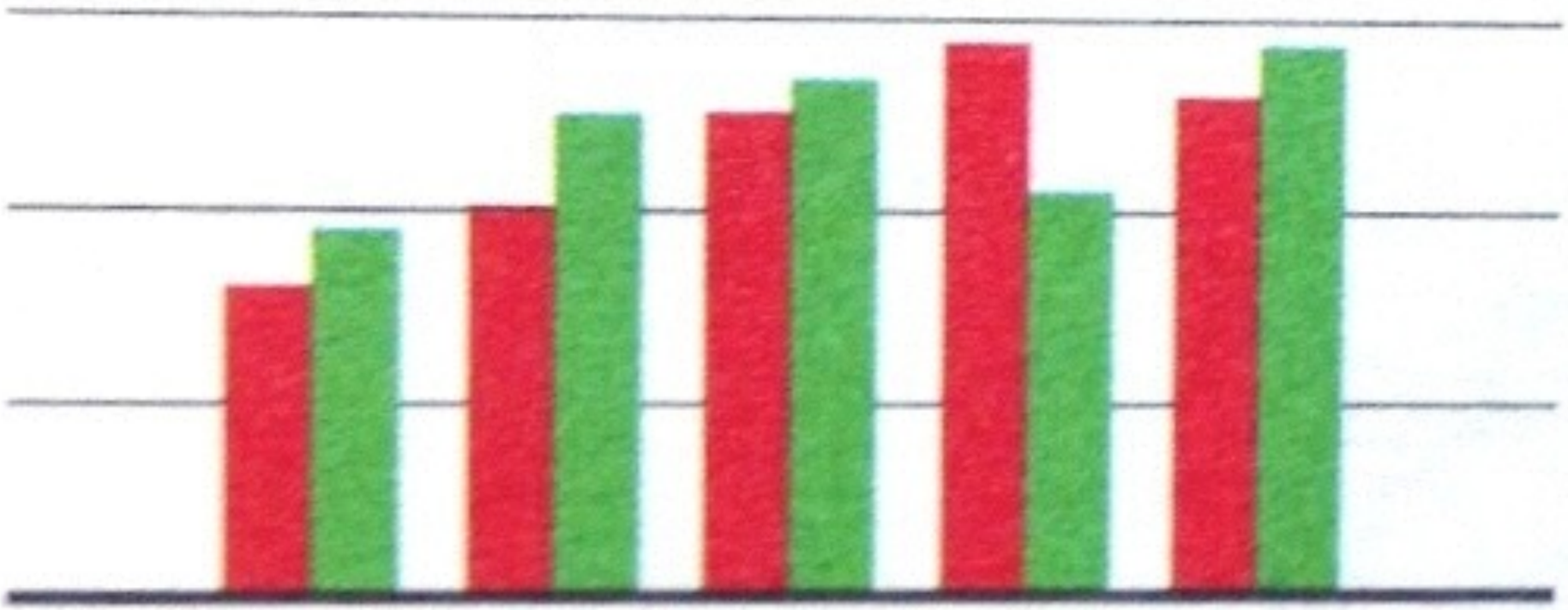
| Name | Data |
|---|---|
| Company A | 1000 |
| Company B | 900 |
| Company C | 80 |
| Company D | 7 |

| Name | Data |
|---|---|
| Company A | 10.82 |
| Company B | 9.49 |
| Company C | 8 |
| Company D | 7.4 |

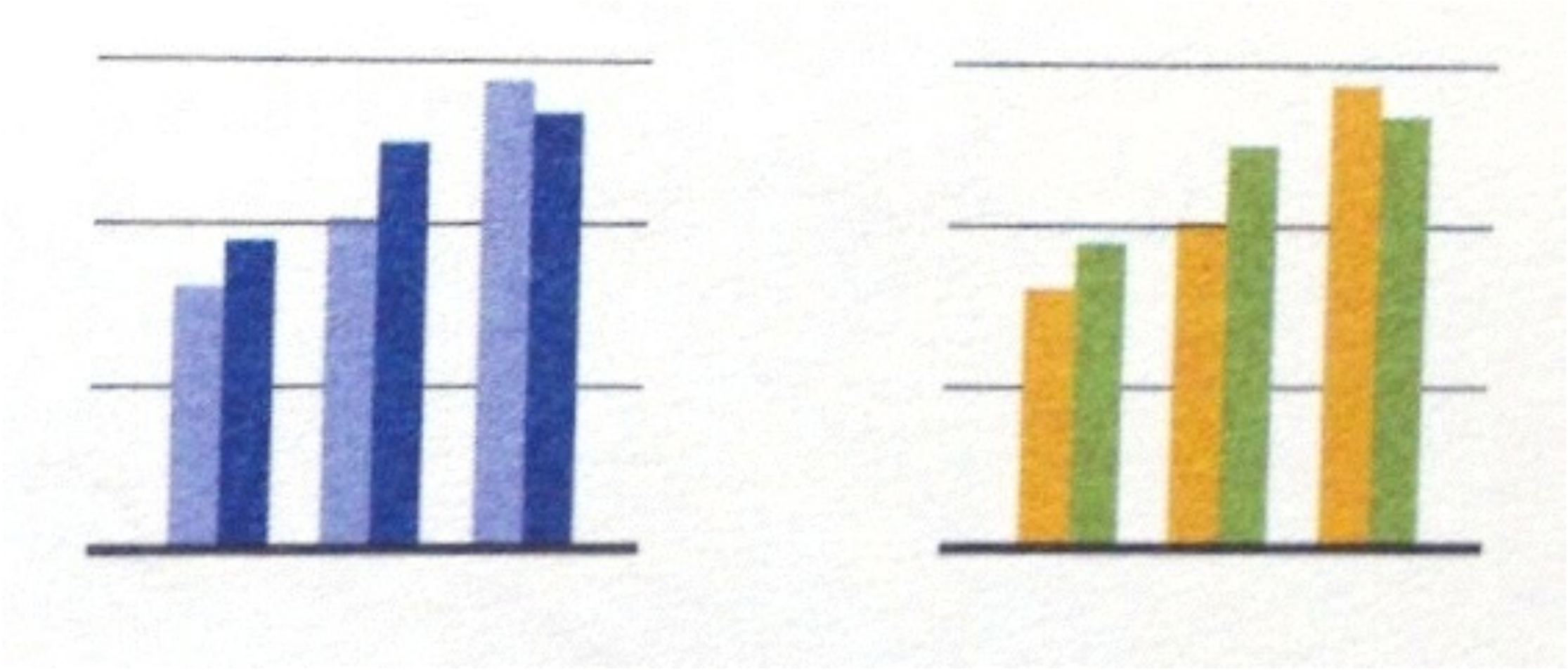
# Look good?

| Name | Data |
|---|---|
| Company A | 10.8 |
| Company B | 9.5 |
| Company C | 8.0 |
| Company D | 7.4 |

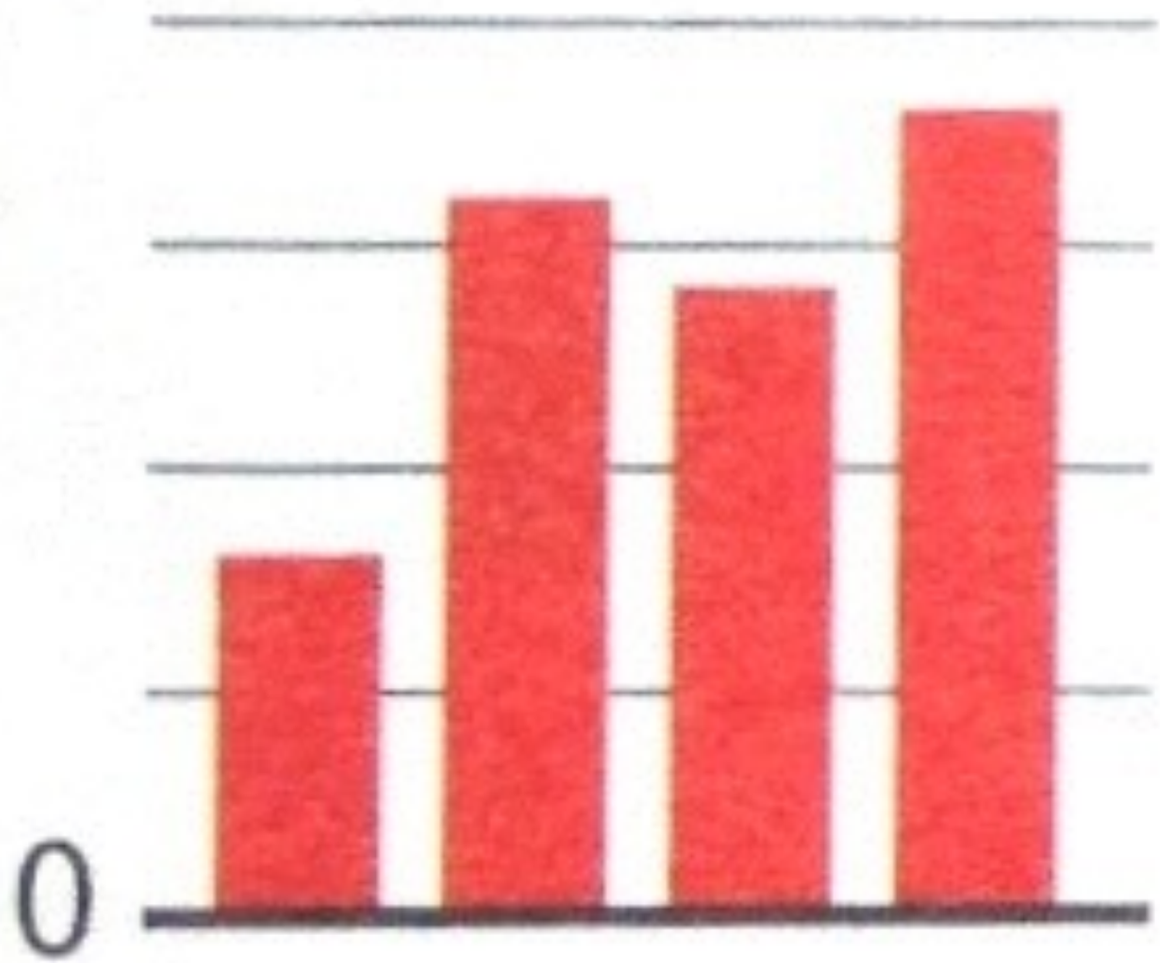
# Bar Charts



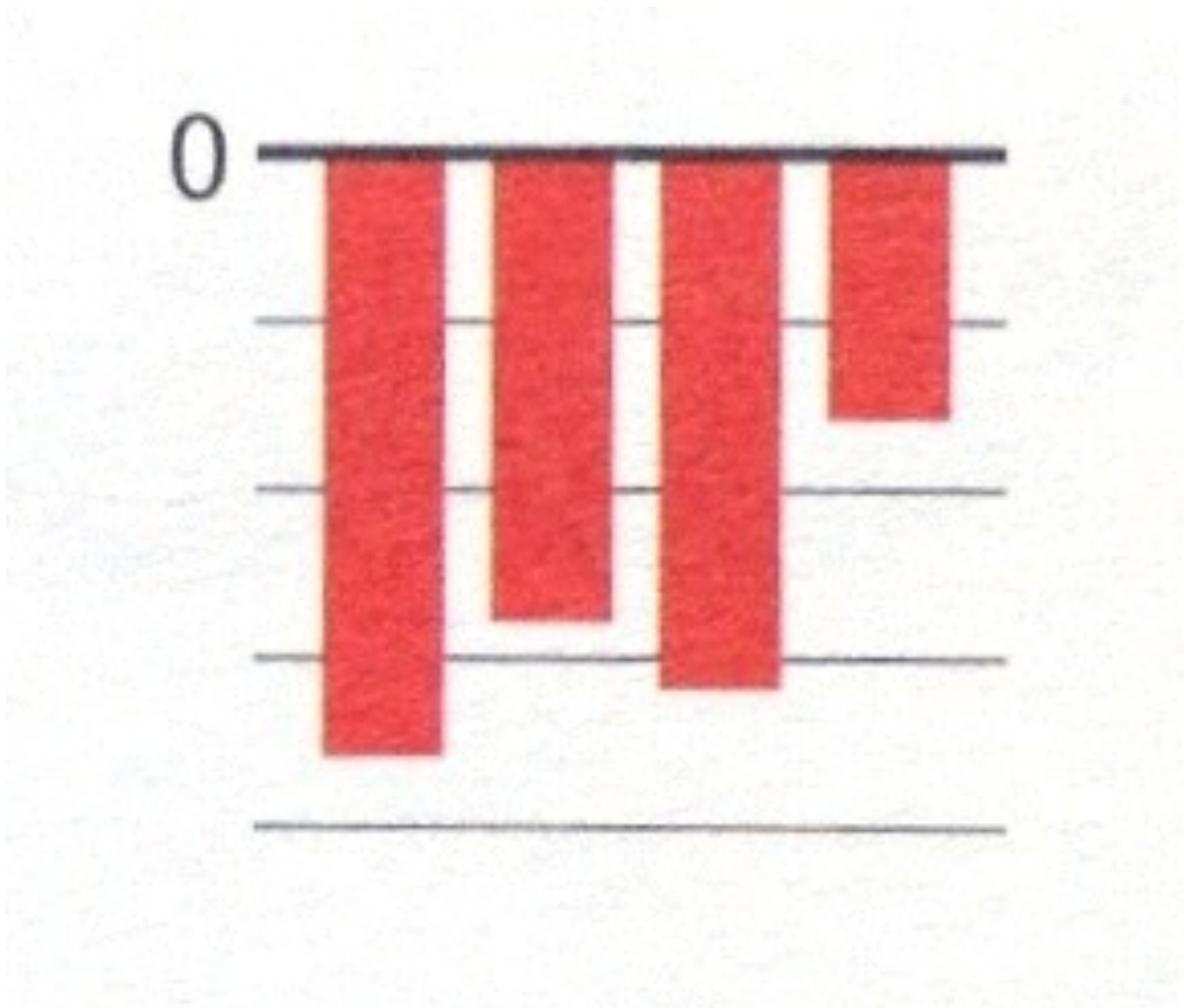
This reminds you of what?



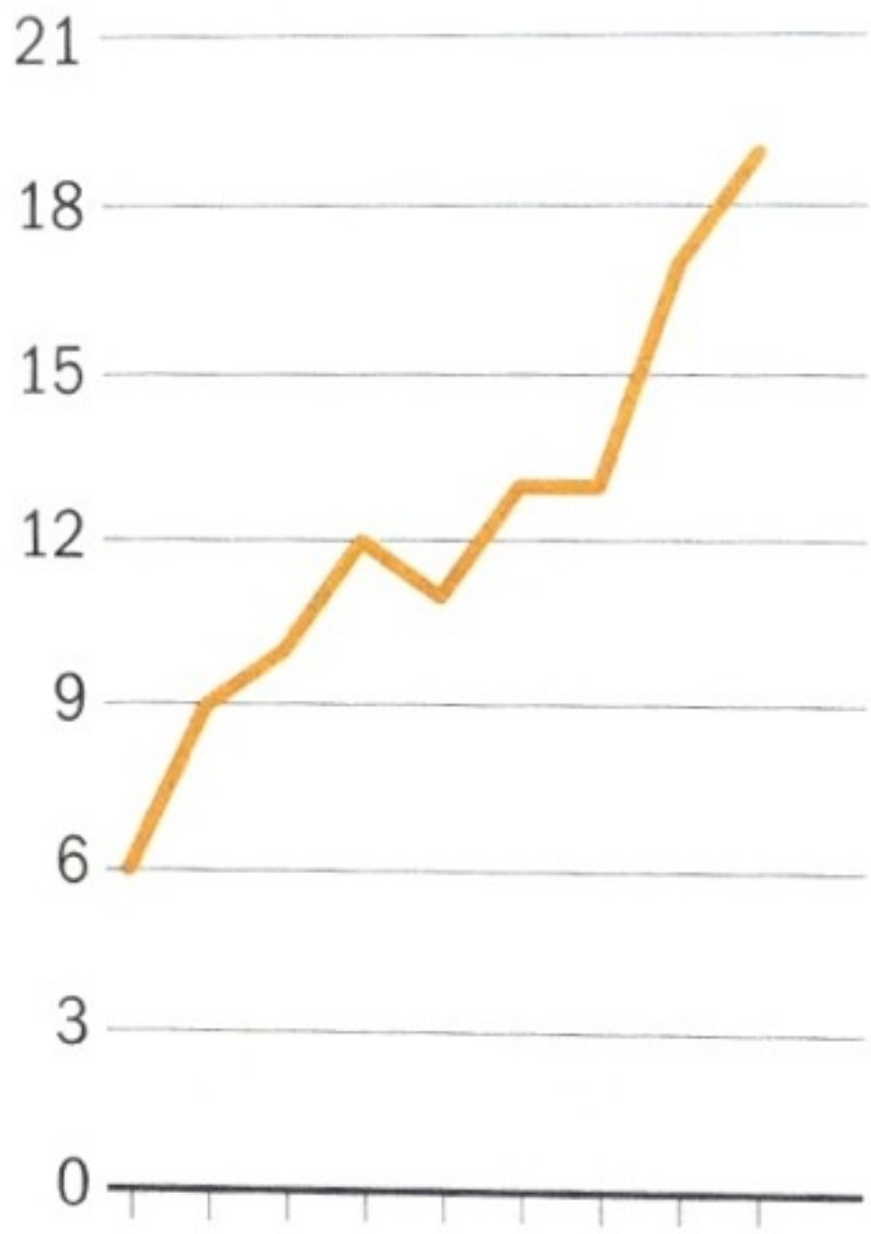
Better than Christmas.




Showing profits in red!!


0

# Line Charts




Does this look alright to you?

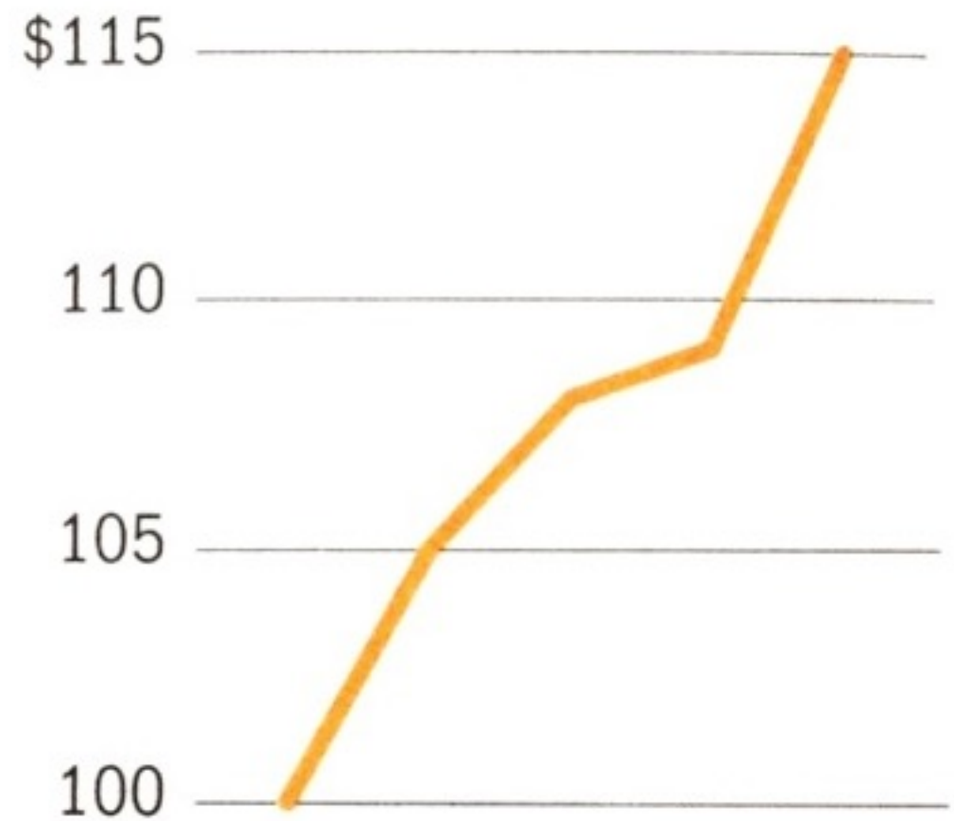


Use “ticks” at regular intervals (e.g., 2, 5, 10, etc.)

# Fever Line

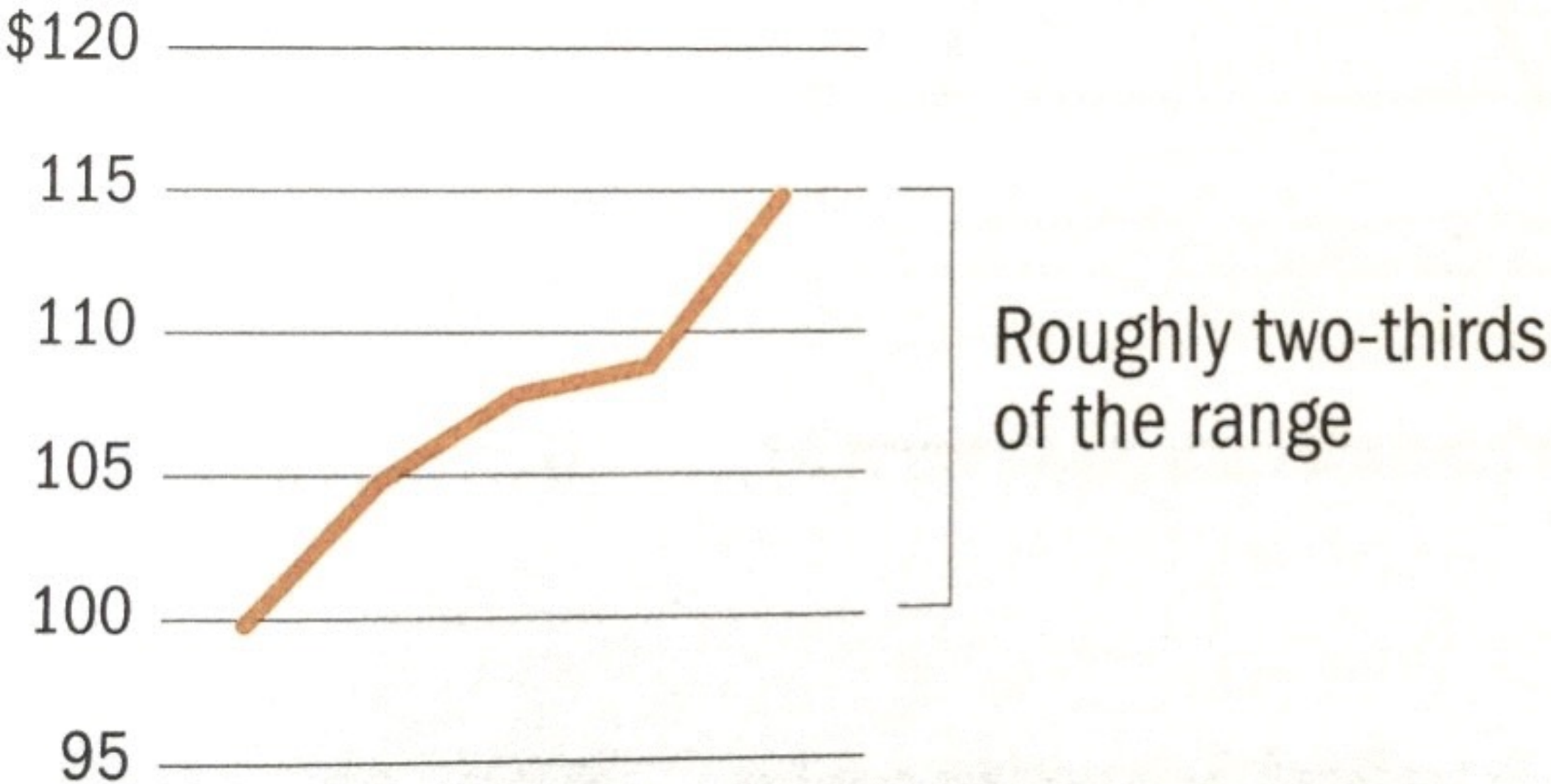
Too flat obscures the message


Too exaggerated overstates the trend

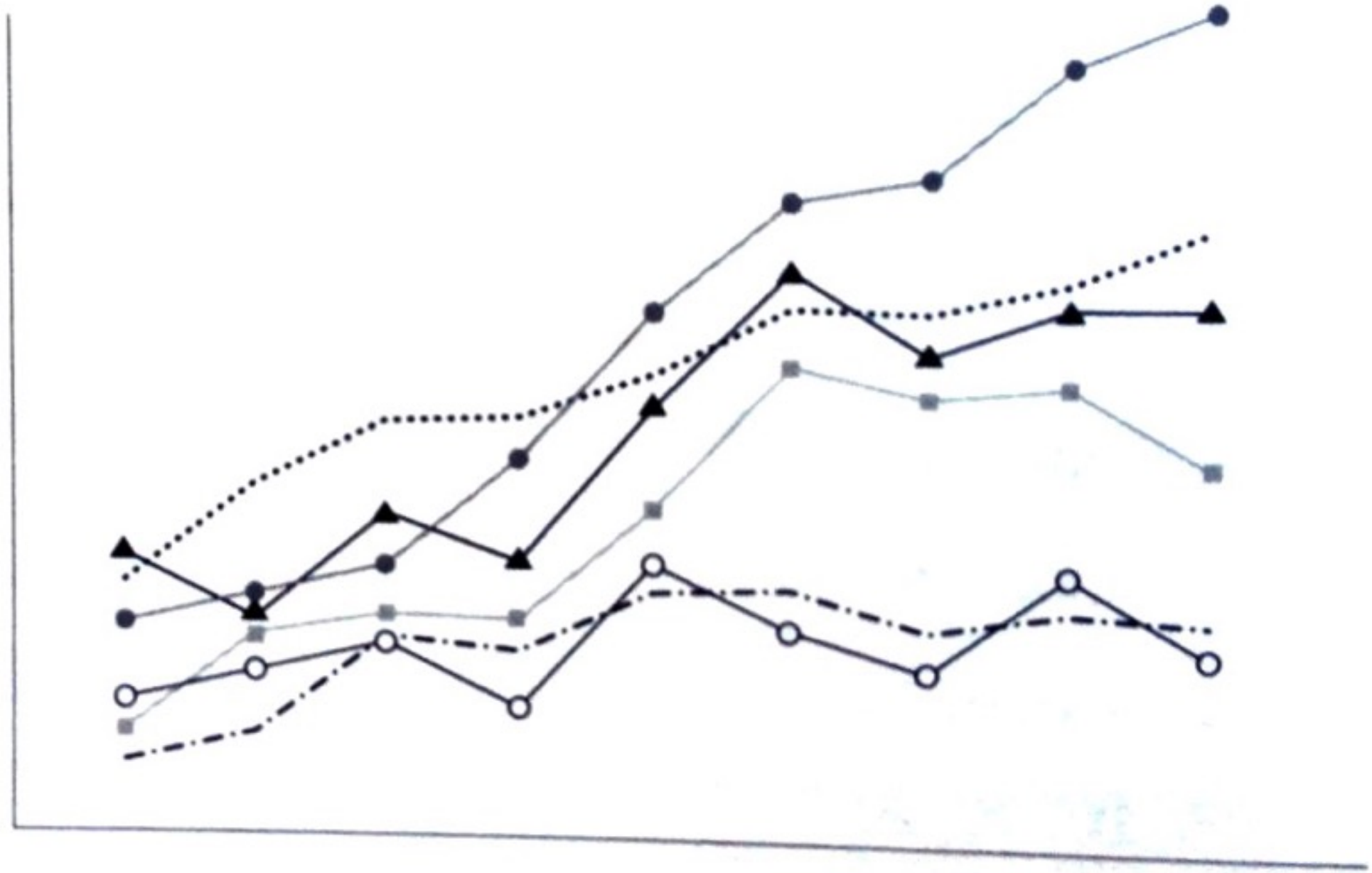



Note y-axis doesn't start at 0.
Why not as bad as in the case of bar chart?

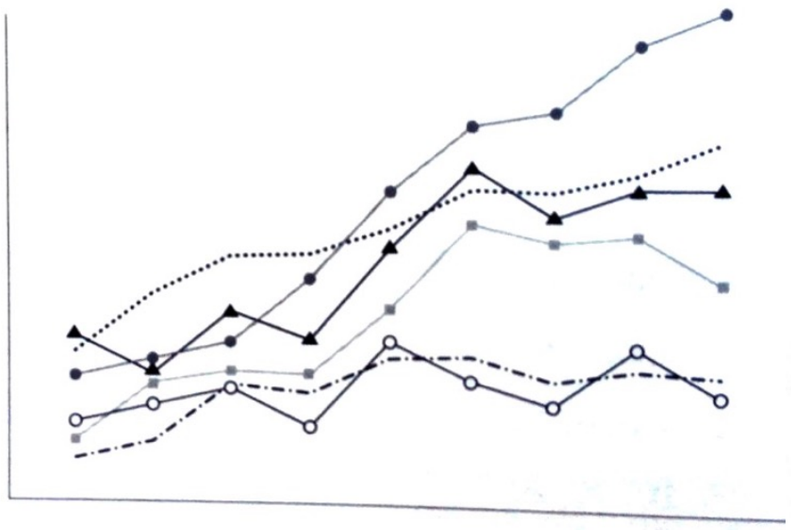
# Fever Line

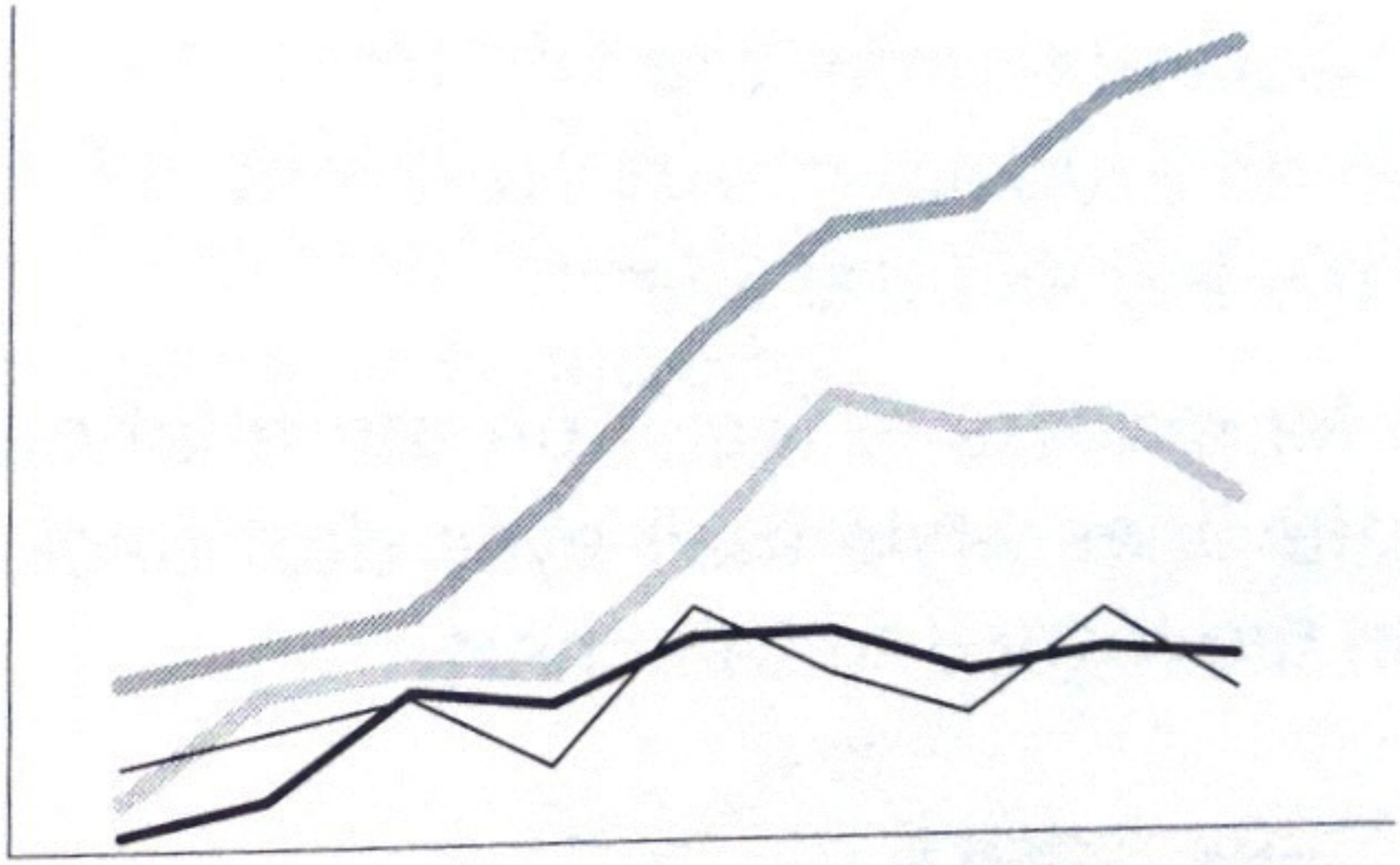


# Multiple Lines in one chart

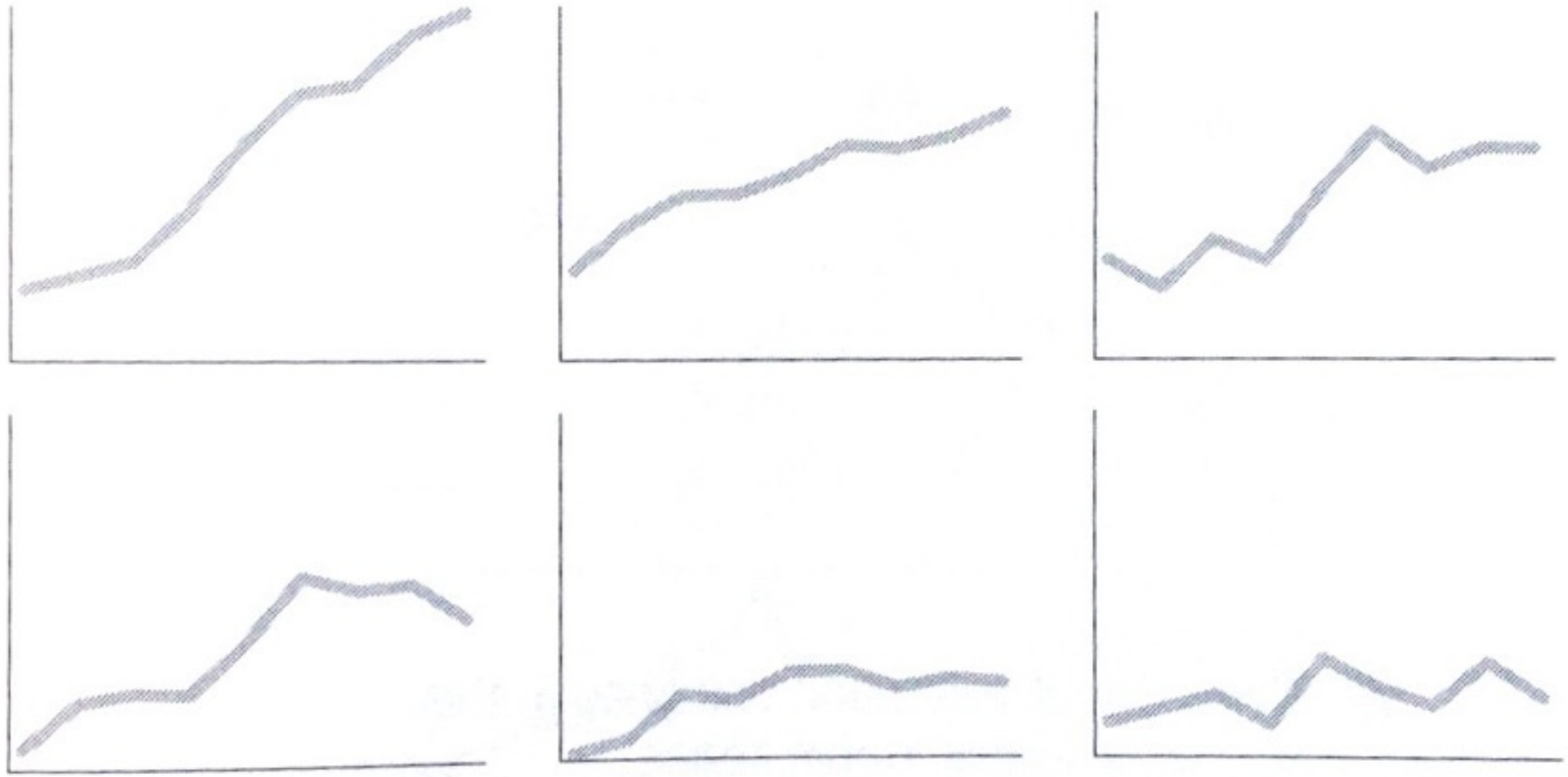


We see this often in academic papers. Better ways?





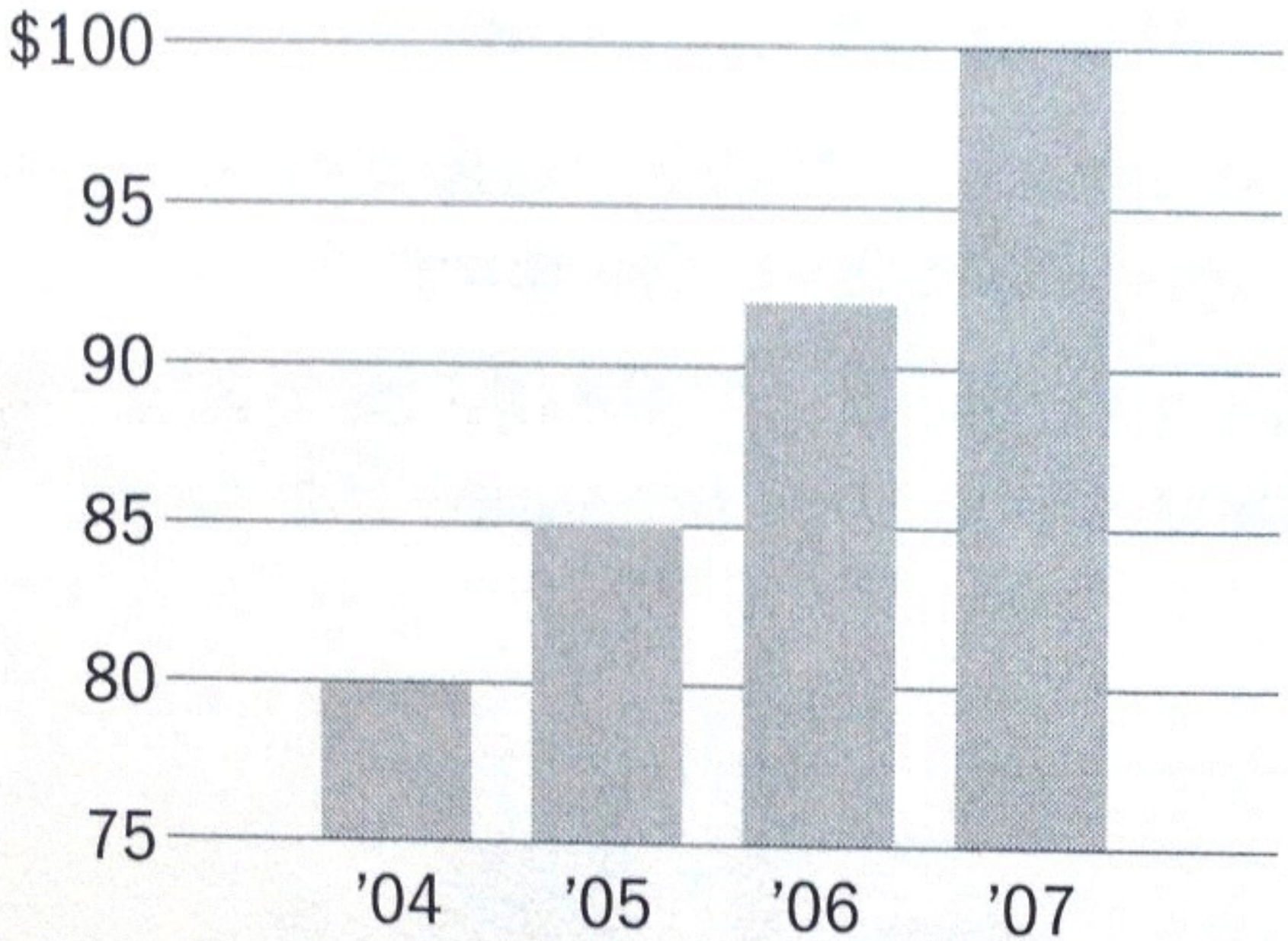
Which one is more effective? Why?
What if you have many lines you want to show?

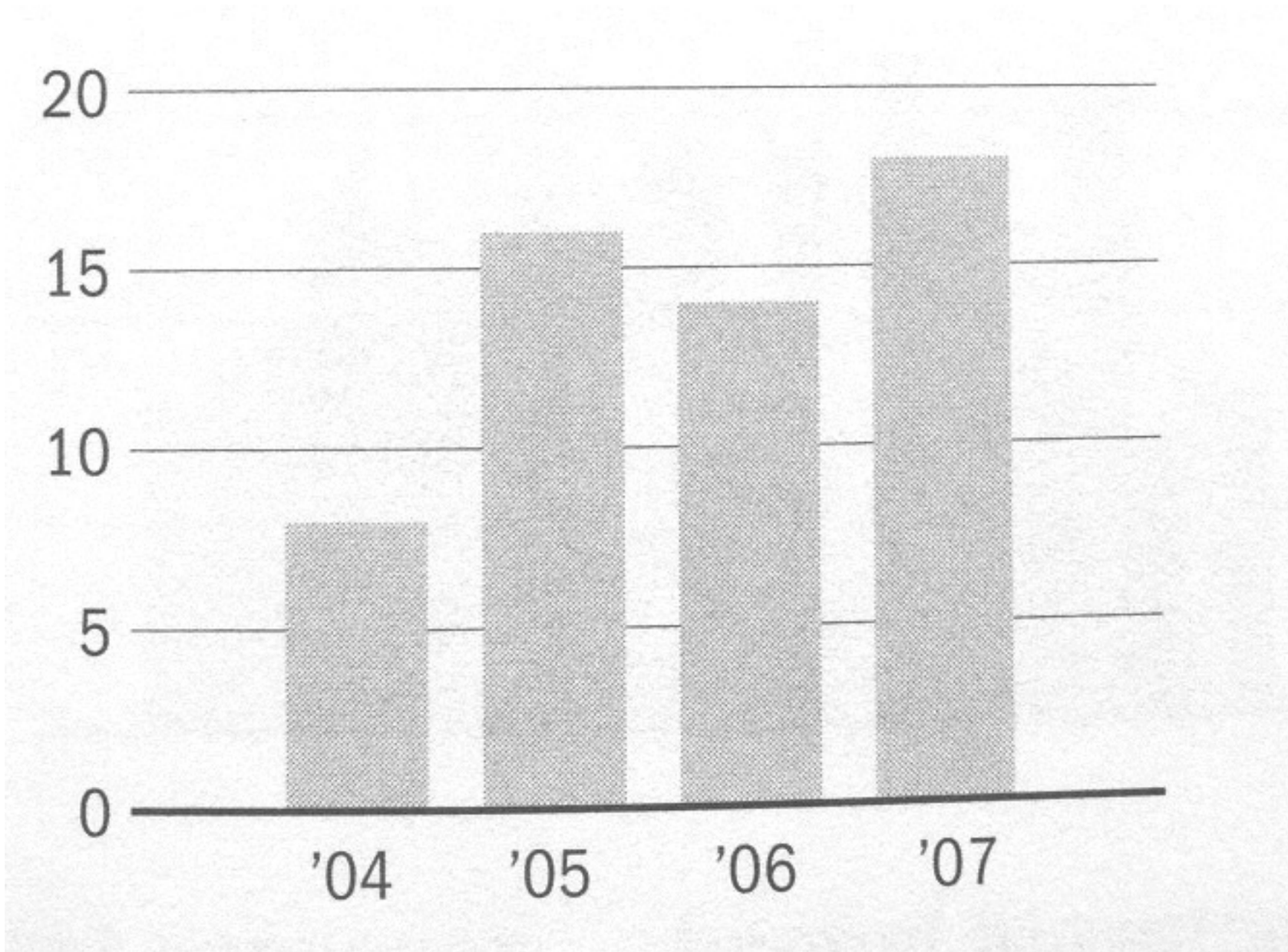


# “**Small Multiple**” - Edward Tufte
# Better than overlapping (sometimes)

“a series or grid of small similar graphics or charts, allowing them to be easily compared”

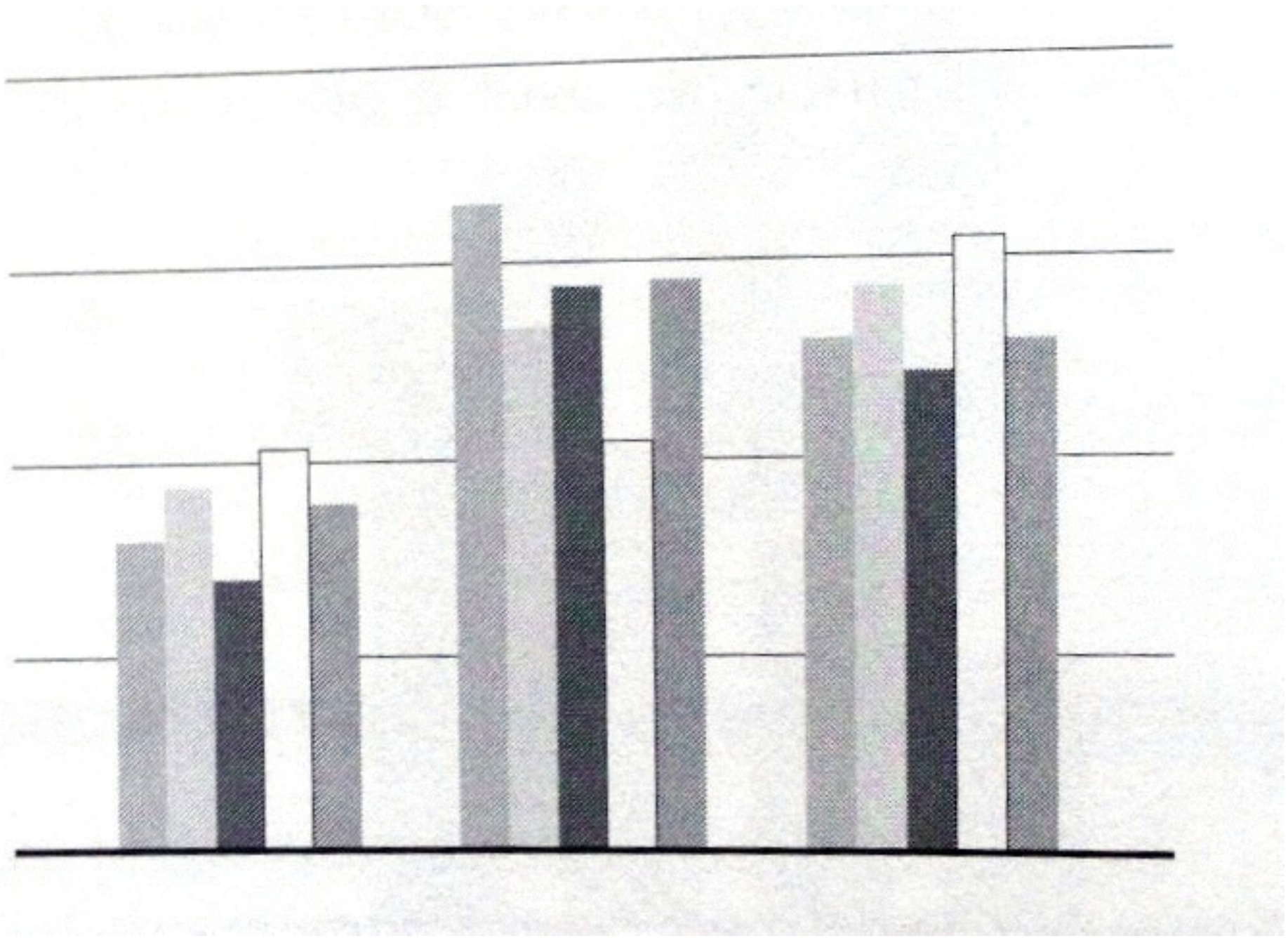
# Misleading Bar Charts



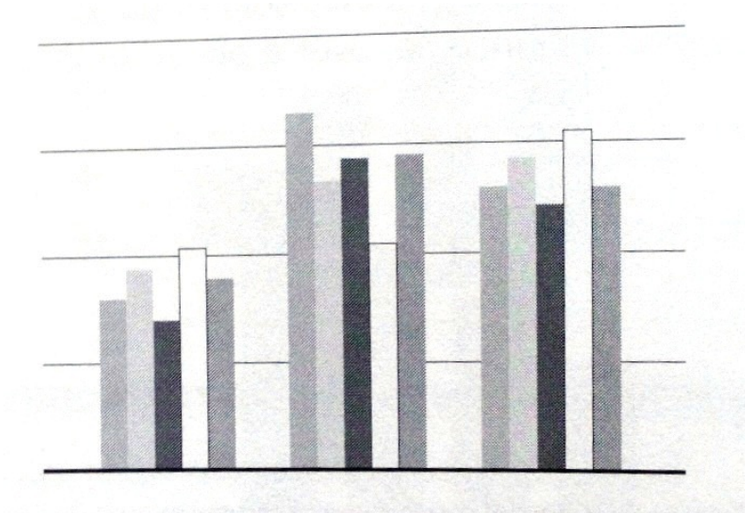


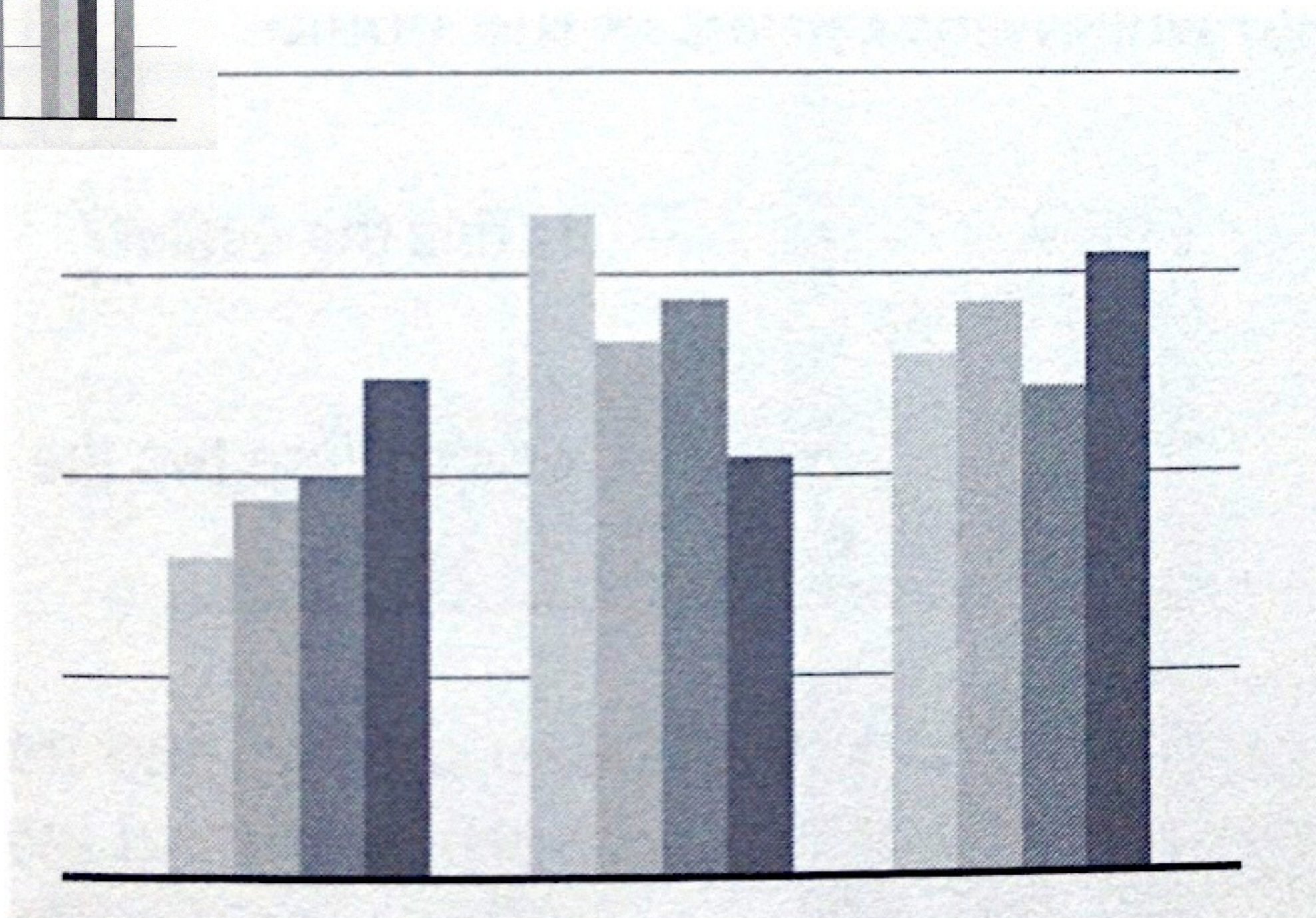

Vertical axis of bar charts start at “0” if possible

# Disorienting color bars

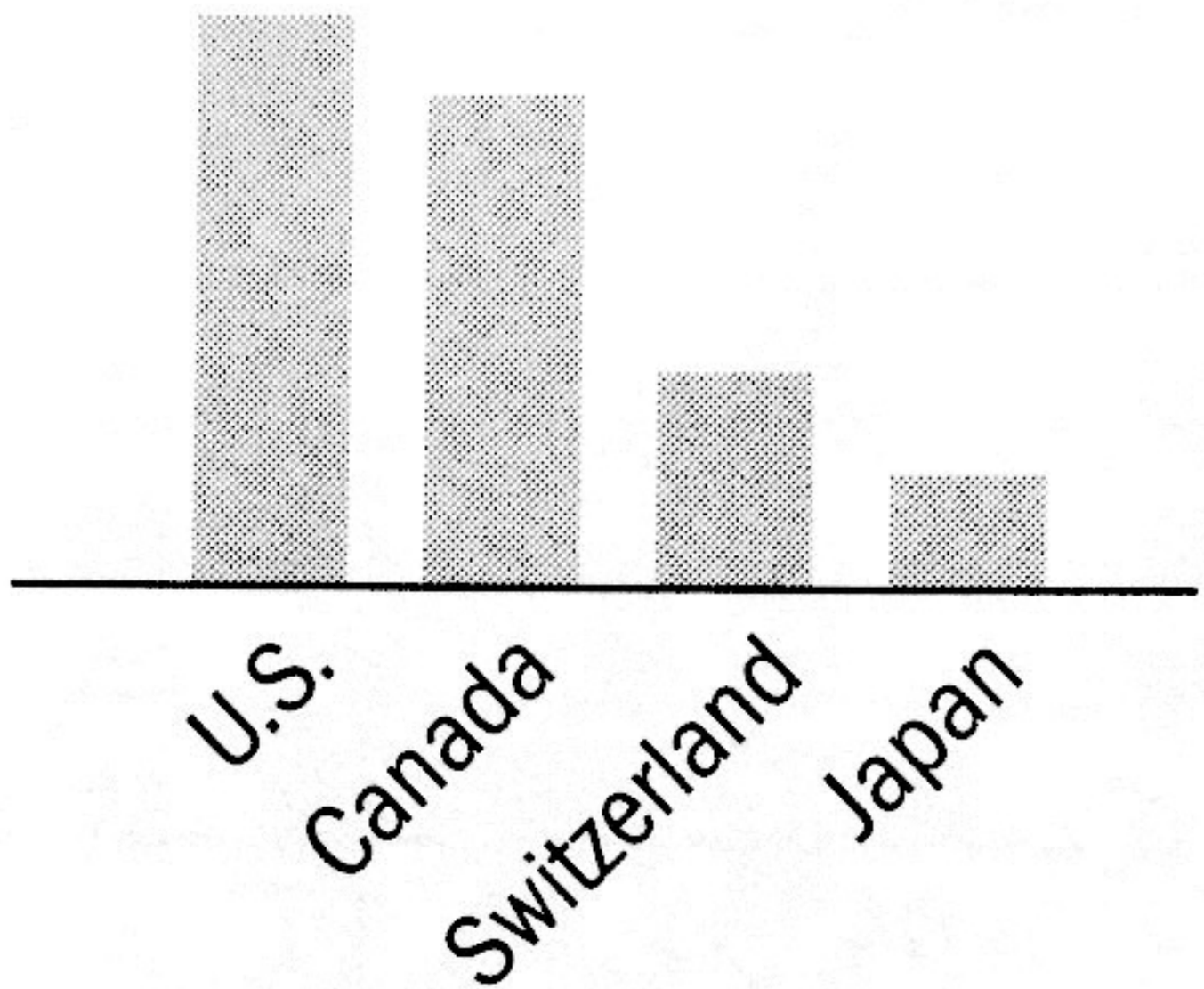


# Better?

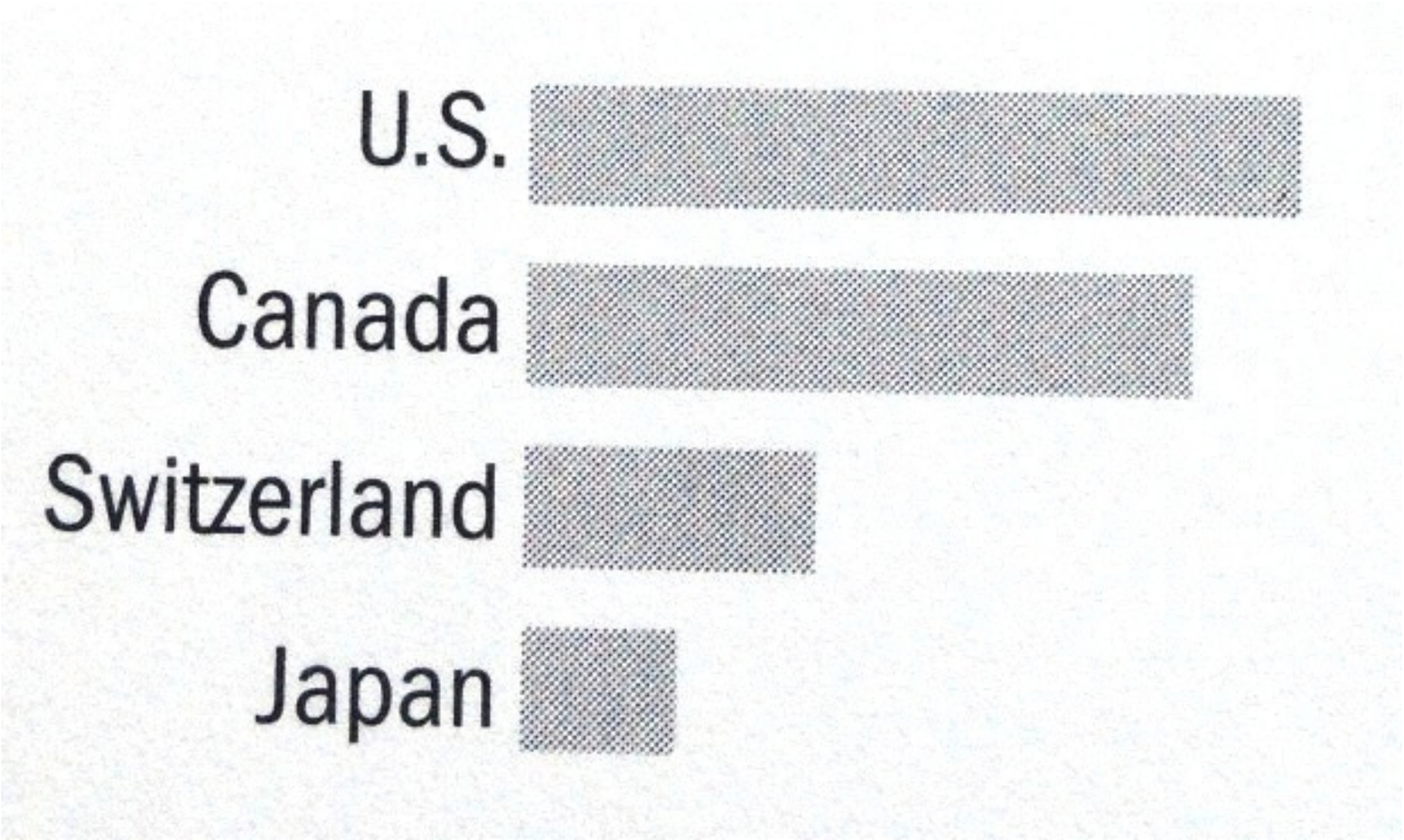




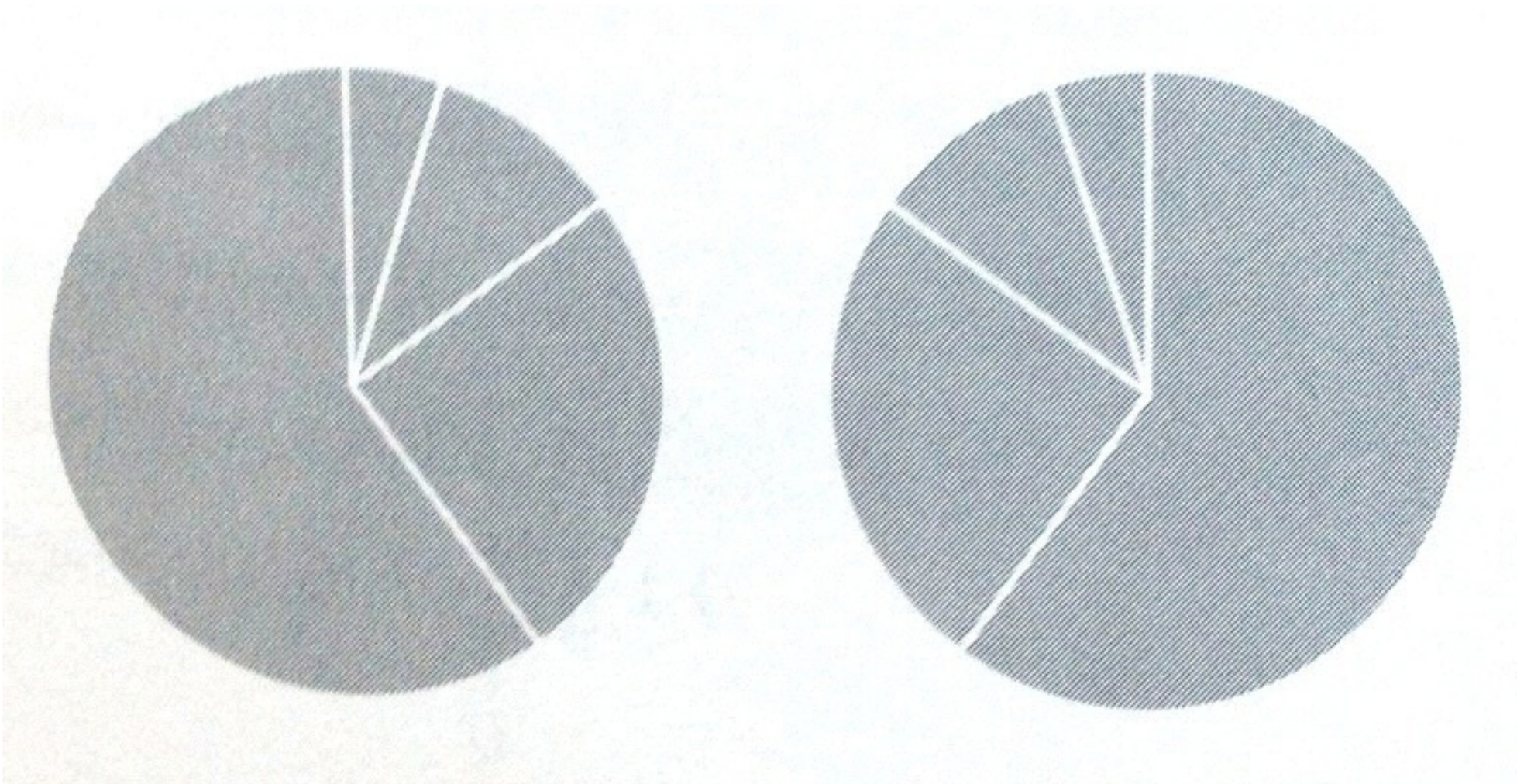
# Exercise For Your Necks



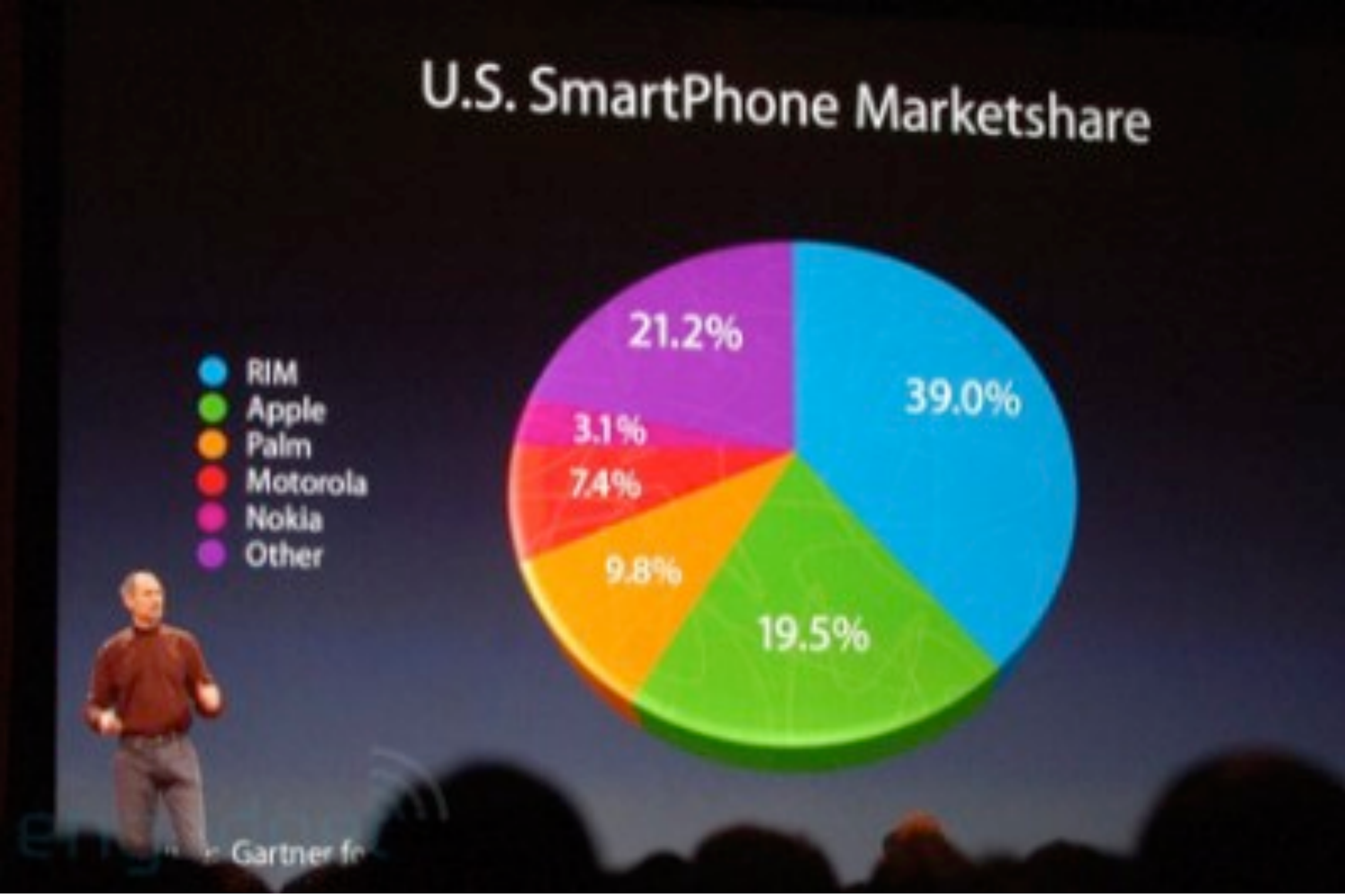
# Bars Can be Horizontal



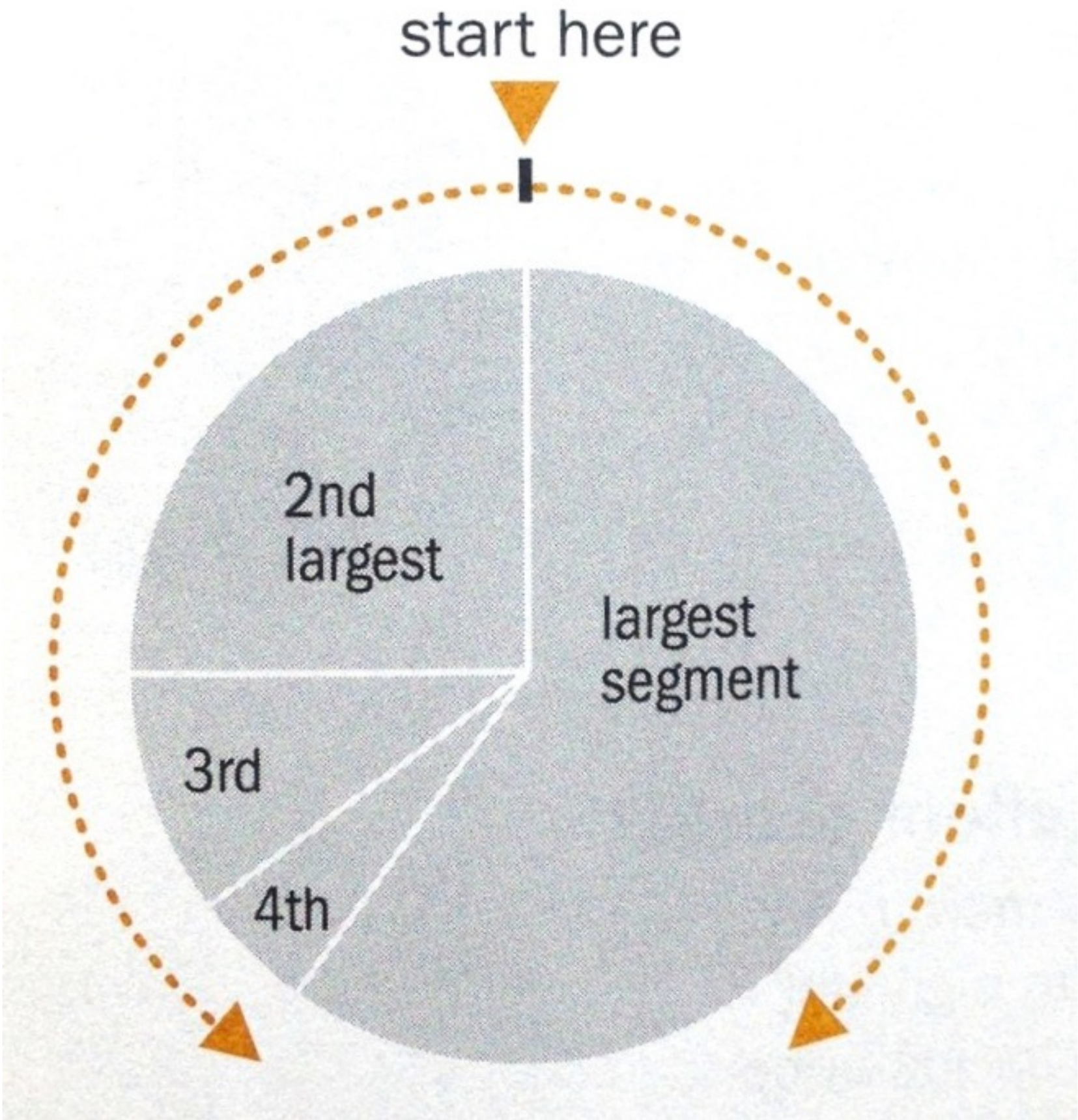
# The Dreaded Pie Charts



Why people like to use pie charts?




http://www.guardian.co.uk/technology/blog/2008/jan/21/liesdamnliesandstevejobs

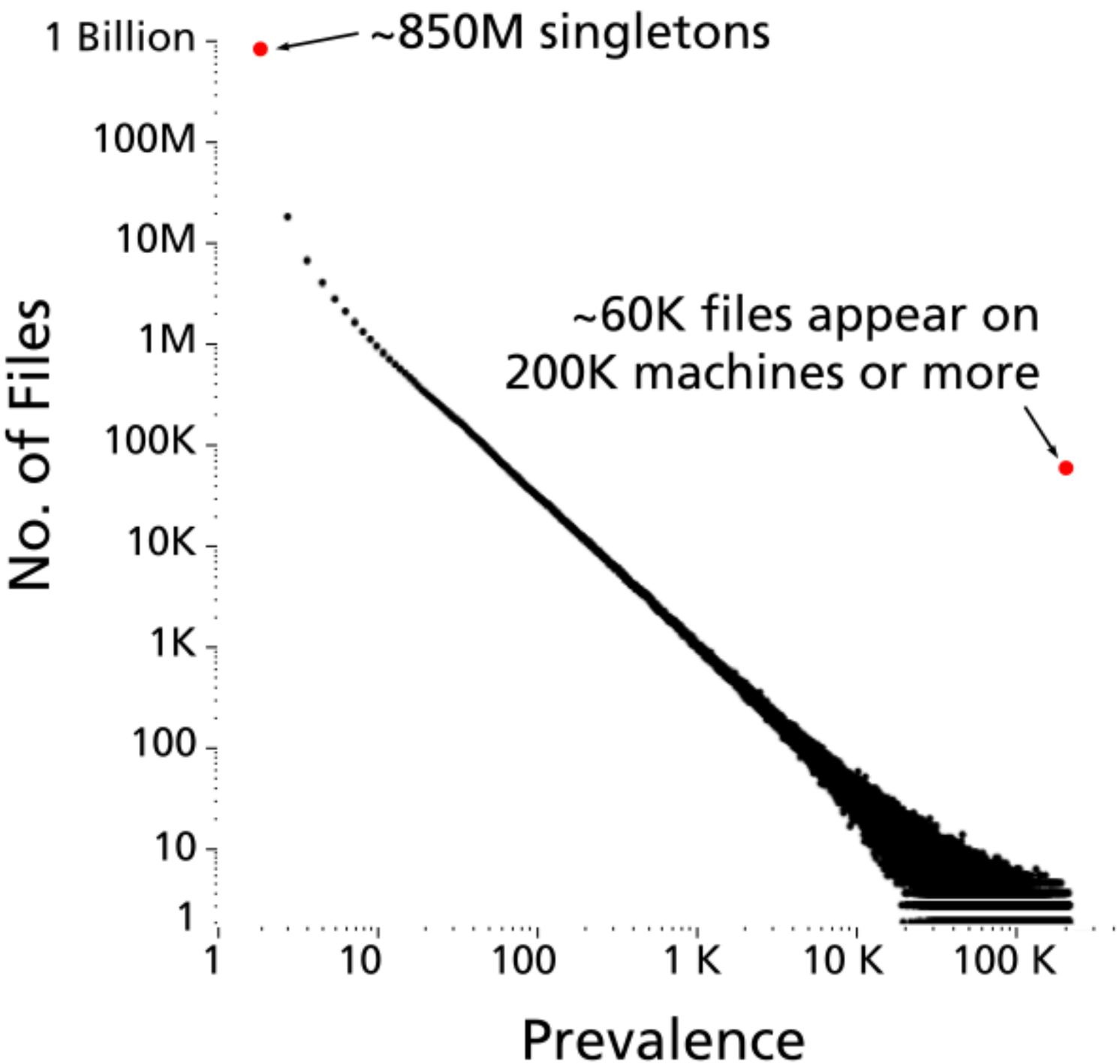

start here
2nd largest
largest segment
3rd
4th

**Log scale** instead of linear scale

Include numbers from different orders of magnitude

# Example

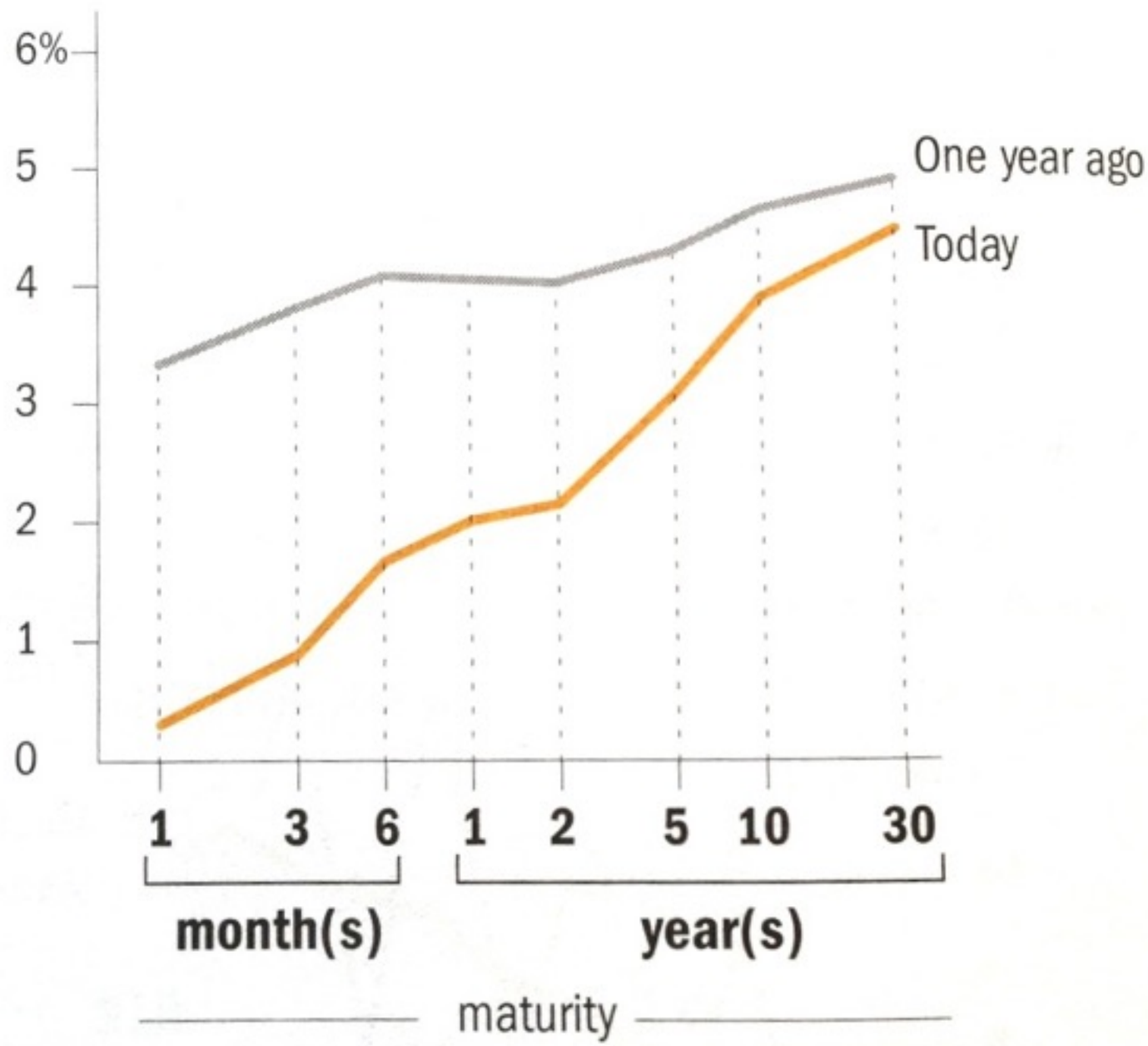
log-log



# “log” also works well for time

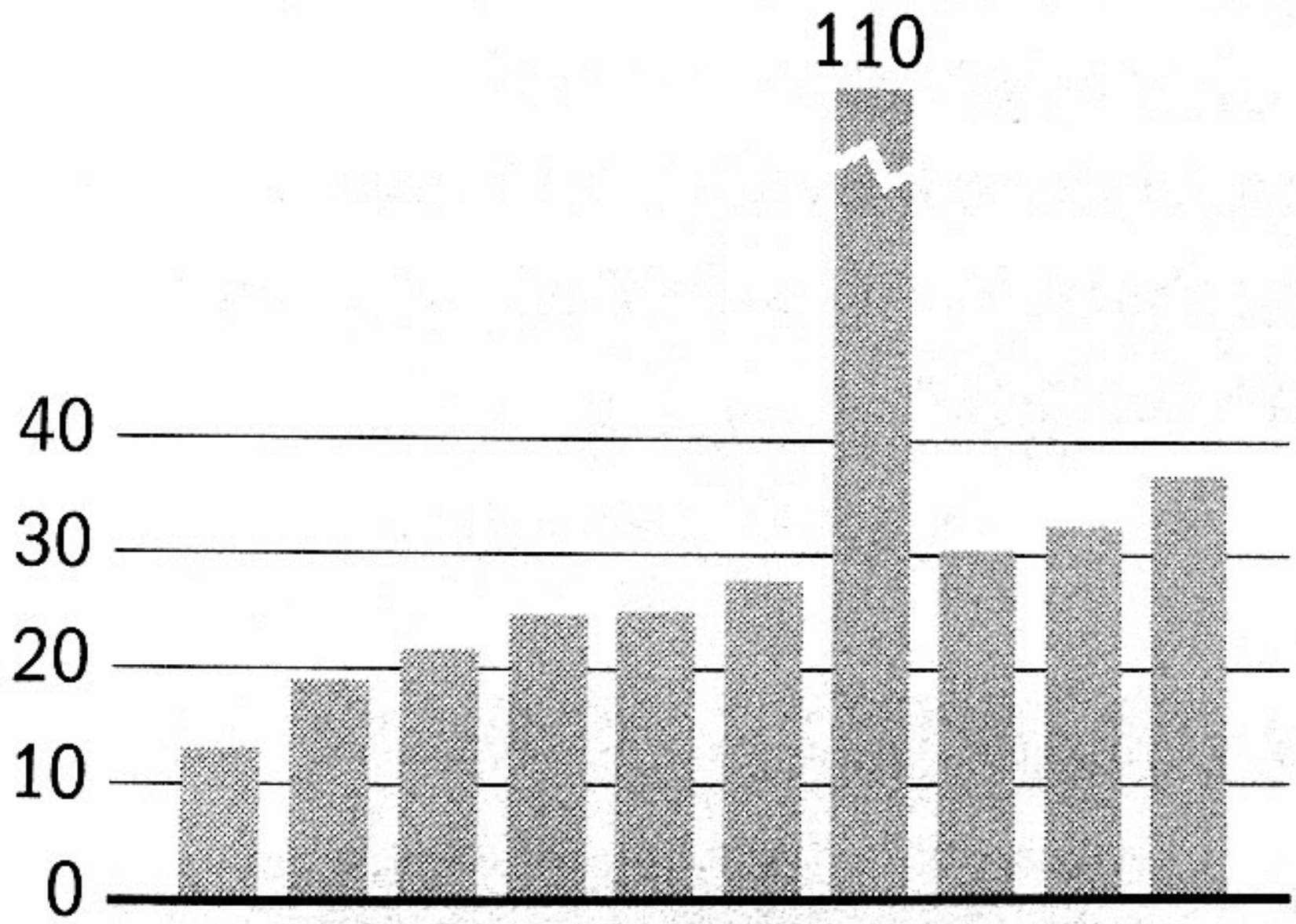
**Example**

The yield curve of Treasury bills, notes and bonds



OK for outliers that are *really* different

# Use broken bars sparingly



# Destroying your great results with poor powerpoint

Bad color schemes

Bad fonts

Too much animation

**Too much data**

can you read this?

100 times faster!

Don McMillan: Life After Death by PowerPoint

http://www.youtube.com/watch?v=lpvgfmEU2Ck&feature=player_embedded

# Destroying your great results with poor powerpoint

How to fix?

- **Color schemes**: start with black & white, add colors if needed
- **Fonts**: sans-serif font looks nicer
  - On Mac: Helvetica is always good
  - On Windows: Arial?
- **Too much animation**: start with **no** animation, then add if appropriate
- **Too much data:** don't just copy figures from paper and past them on the slides!

Don McMillan: Life After Death by PowerPoint

http://www.youtube.com/watch?v=lpvgfmEU2Ck&feature=player_embedded

# Suggestions: use pictures whenever appropriate

"Pictures" include most *non-text* elements: tables, diagrams, charts, etc.

Why?

- "A picture is worth a thousand words"
- People like pictures and love movies.
- Picture is often more succinct, memorable

# Figures should be self-contained

Why?

- Don’t make people go back and forth between text and figure
- People skim; look at “interesting” things first
- Especially academia, many busy reviewers look at figures first
- Bad figures -> bad first impression
  (lower chance of paper acceptance)

How to fix?

- Succinctly describe your main messages
  (what you want the readers to learn)

# Example

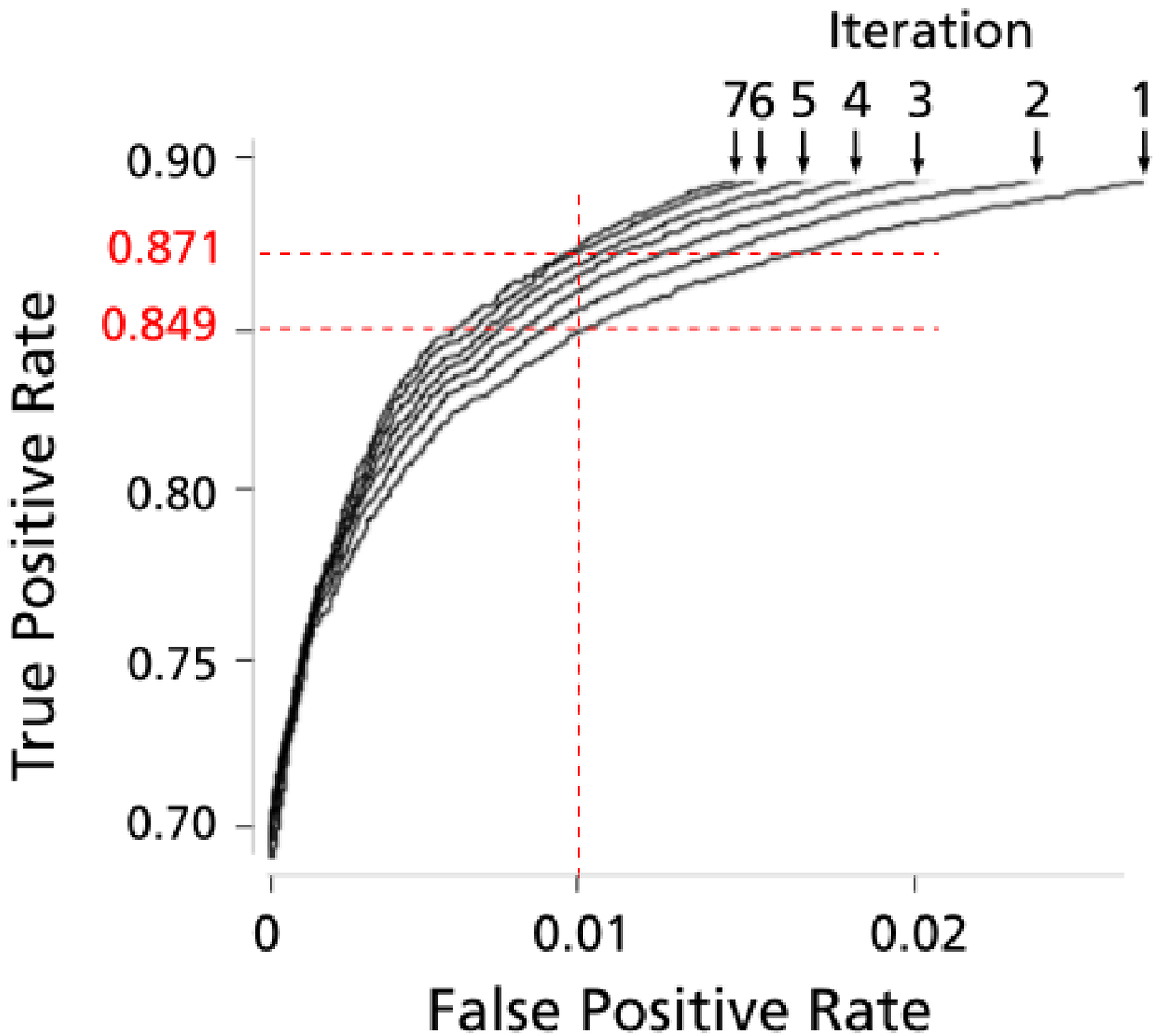



Figure 8: ROC curves of 7 iterations; true positive rate incrementally improves.

http://www.cs.cmu.edu/~dchau/polonium_sdm2011.pdf

Example

SIGGRAPH2007 Computer Graphics Proceedings, Annual Conference Series, 2007

# Scene Completion Using Millions of Photographs

James Hays Alexei A. Efros
Carnegie Mellon University

Original Image Input Scene Matches Output

Figure 1: Given an input image with a missing region, we use matching scenes from a large collection of photographs to complete the image.

## Abstract

What can you do with a million images? In this paper we present a new image completion algorithm powered by a huge database of photographs gathered from the Web. The algorithm patches up holes in images by finding similar image regions in the database that are not only seamless but also semantically valid. Our chief insight is that while the space of images is effectively infinite, the space of semantically differentiable scenes is actually not that large. For many image completion tasks we are able to find similar scenes which contain image fragments that will convincingly complete the image. Our algorithm is entirely data-driven, requiring no annotations or labelling by the user. Unlike existing image completion methods, our algorithm can generate a diverse set of results for each input image and we allow users to select among them. We demon-

There are two fundamentally different strategies for image completion. The first aims to reconstruct, as accurately as possible, the data that *should have been* there, but somehow got occluded or corrupted. Methods attempting an accurate reconstruction have to use some other source of data in addition to the input image, such as video (using various background stabilization techniques, e.g. [Irani et al. 1995]) or multiple photographs of the same physical scene [Agarwala et al. 2004; Snavely et al. 2006].

The alternative is to try finding a plausible way to fill in the missing pixels, hallucinating data that *could have been* there. This is a much less easily quantifiable endeavor, relying instead on the studies of human visual perception. The most successful existing methods [Criminisi et al. 2003; Drori et al. 2003; Wexler et al. 2004; Wilczkowiak et al. 2005; Komodakis 2006] operate by extending

# Crown-jewel figure on first page

(nice to have)

Why?

- Give an overview of what readers is going to get -- cut to the chase
- Again, people like to see interesting things

How to do it?

- Use your most impressive figure
- Can be similar to another shown later

Example

# Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning

**Duen Horng "Polo" Chau, Aniket Kittur, Jason I. Hong, Christos Faloutsos**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{dchau, nkittur, jasonh, christos}@cs.cmu.edu

**ABSTRACT**
Extracting useful knowledge from large network datasets has become a fundamental challenge in many domains, from scientific literature to social networks and the web. We introduce Apolo, a system that uses a mixed-initiative approach—combining visualization, rich user interaction and machine learning—to guide the user to incrementally and interactively explore large network data and make sense of it. Apolo engages the user in bottom-up sensemaking to gradually build up an understanding over time by starting small, rather than starting big and drilling down. Apolo also helps users find relevant information by specifying exemplars, and then using a machine learning method called Belief Propagation to infer which other nodes may be of interest. We evaluated Apolo with twelve participants in a between-subjects study, with the task being to find relevant new papers to update an existing survey paper. Using expert judges, participants using Apolo found significantly more relevant papers. Subjective feedback of Apolo was also very positive.

**Author Keywords**
Sensemaking, large network, Belief Propagation

**ACM Classification Keywords**
H.3.3 Information Storage and Retrieval: Relevance feedback; H.5.2 Information Interfaces and Presentation: User Interfaces

**General Terms**
Algorithms, Design, Human Factors

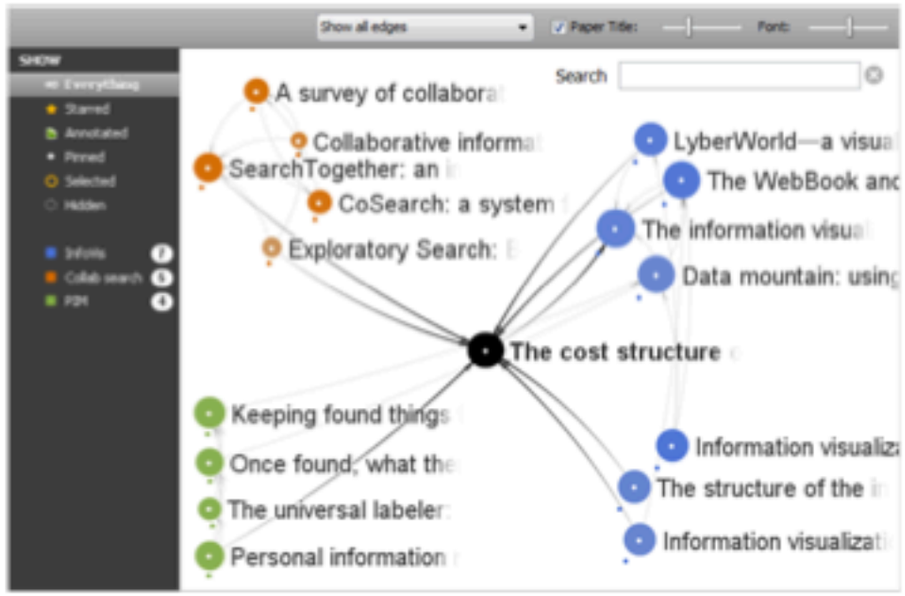



**Figure 1. Apolo displaying citation network data around the article *The Cost Structure of Sensemaking*. The user gradually builds up a mental model of the research areas around the article by manually inspecting some neighboring articles in the visualization and specifying them as exemplar articles (with colored dots underneath) for some ad hoc groups, and instructs Apolo to find more articles relevant to them.**

representation or schema of an information space that is useful for achieving the user's goal [31]. For example, a scientist interested in connecting her work to a new domain must build up a mental representation of the existing literature in the new domain to understand and contribute to it.

For the above scientist, she may forage to find papers that she thinks are relevant, and build up a representation of how

# Suggestion: Design in grayscale first

Then add color

If it doesn’t look good in black and white, it’s not gonna look good with color

(Why iPhone comes in black or white first?)

# Suggestion: Use legible fonts

## If people can’t see it, they won’t appreciate it

For printed materials, print them out and check!

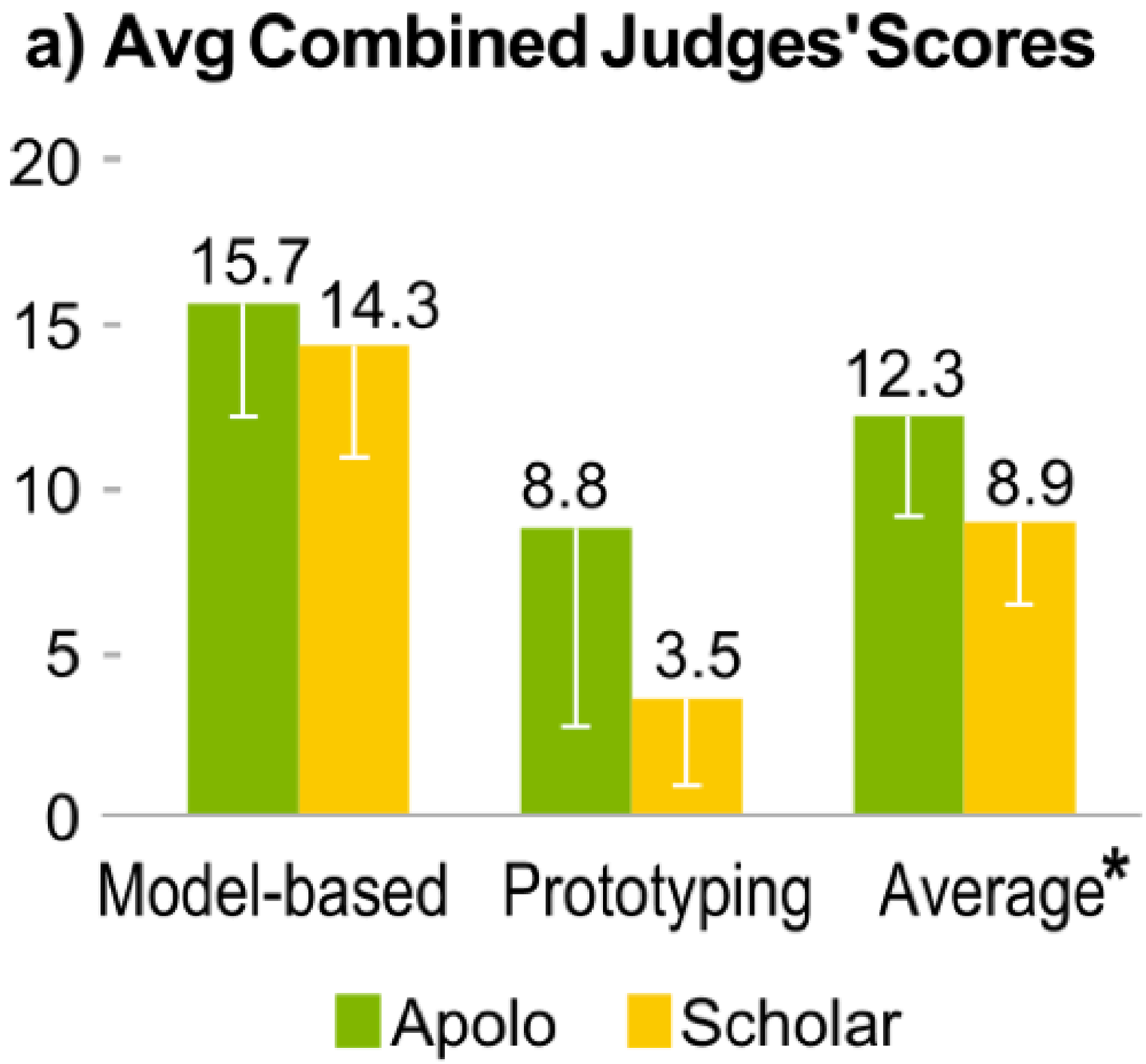
For slides, rule of thumb is about 7 lines of text per slide.

# Suggestion: you probably need to redo your figure for slides

Designing for print is different from designing for the screen

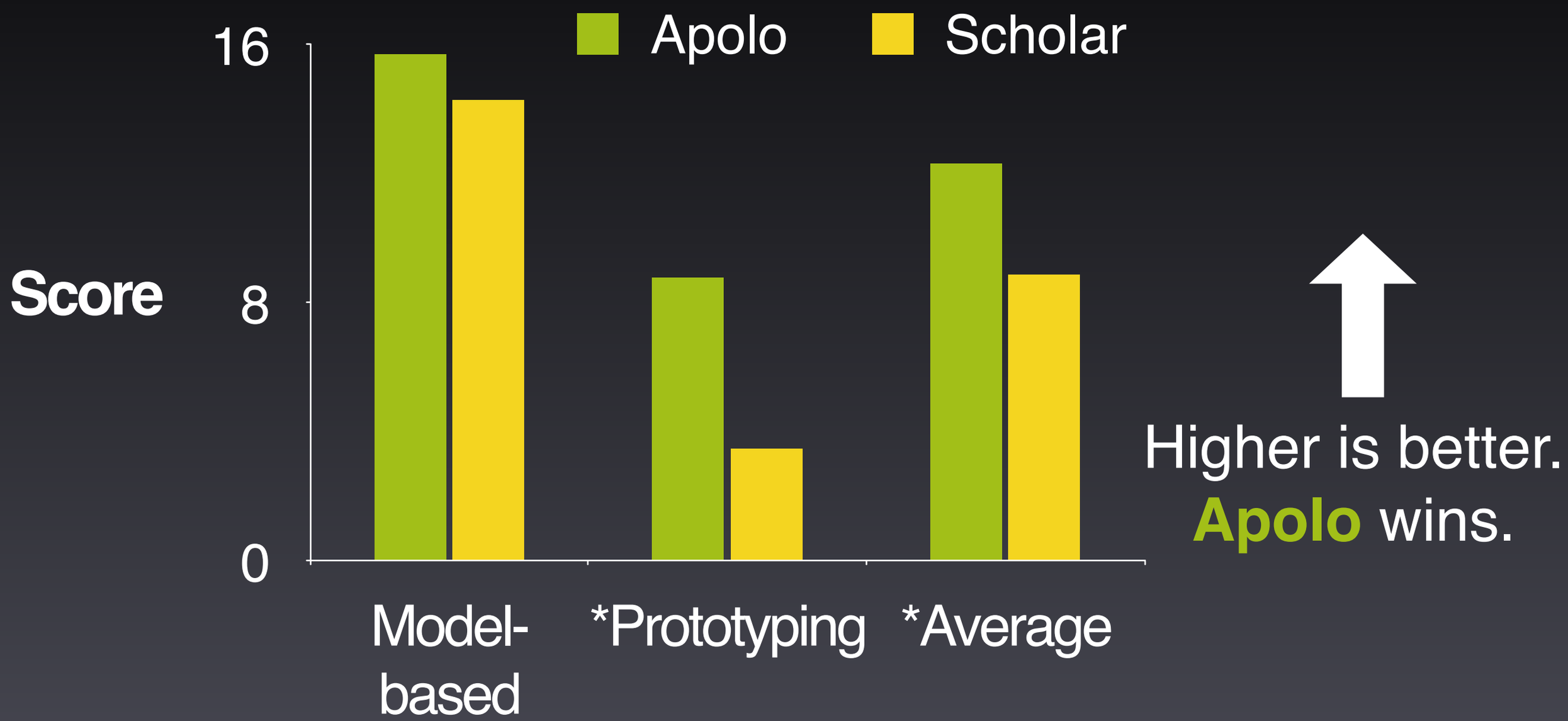
- Resolution (which is higher?)
- Levels of details (people mostly want a few “take-away” messages from your talk)

Example


a) Avg Combined Judges'Scores
20
15
10
5
0
15.7
14.3
8.8
3.5
12.3
8.9
Model-based
Prototyping
Average*
Apolo
Scholar

Example

# Judges' Scores




** Statistically significant, by two-tailed t test, $p < 0.05$*

# Good tools for creating data visualization

(beyond Excel)

# R

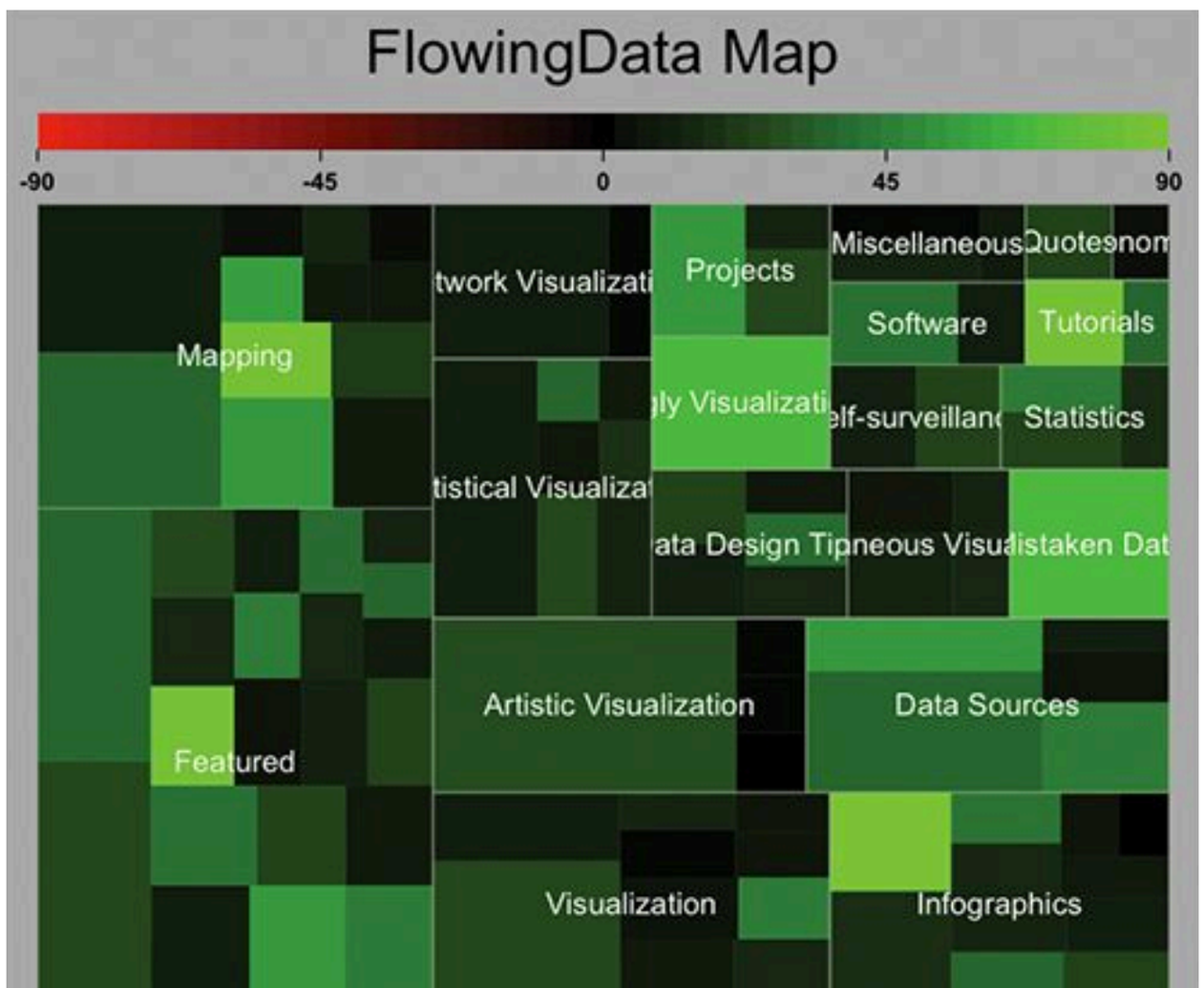
http://www.r-project.org

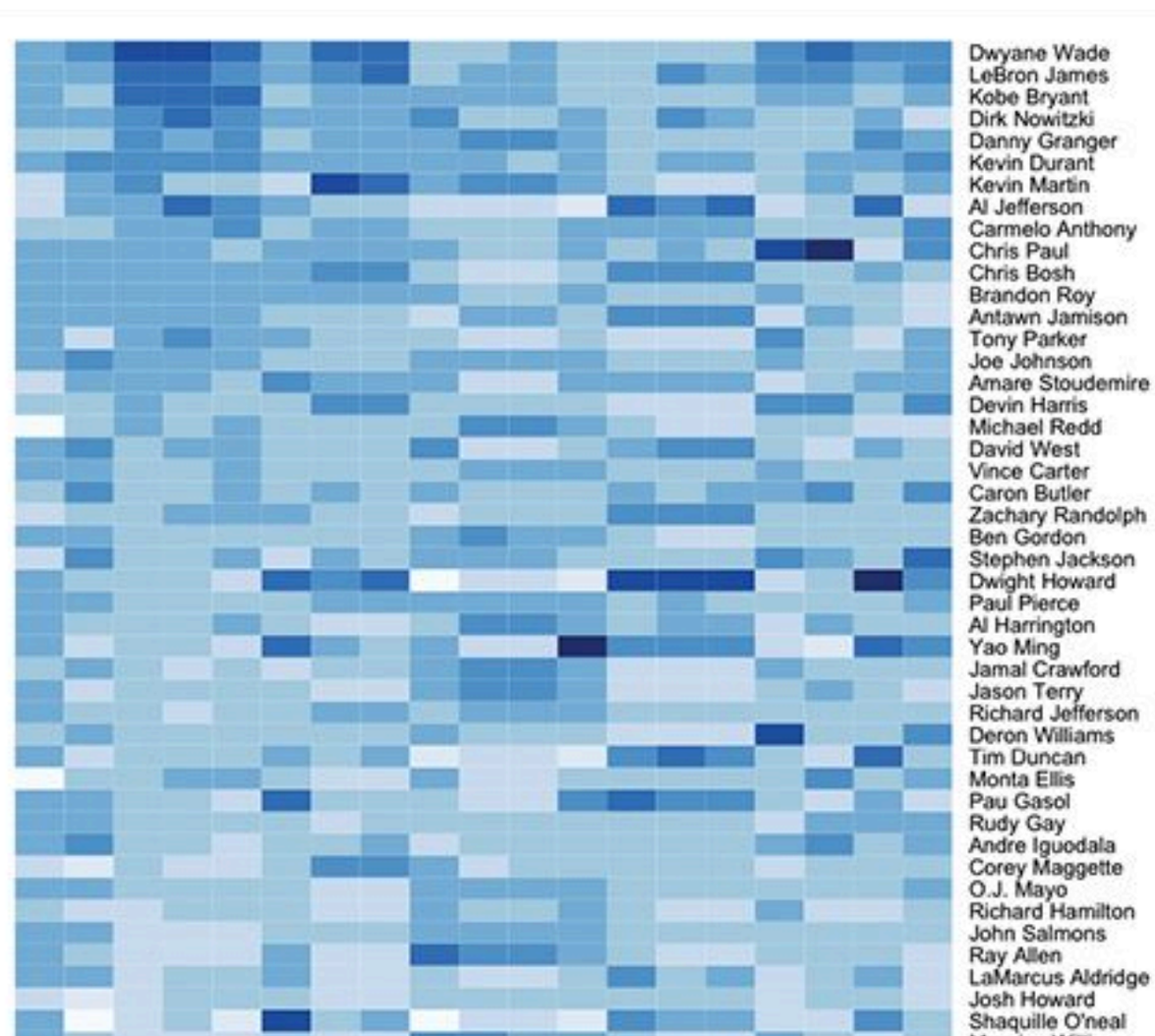
http://www.cc.gatech.edu/~lebanon/notes/quickIntroToR.pdf

Free!

Powerful. Can create any kinds of visualization available.

But results may not be pretty (need editing). Need to program.



# D3

http://d3js.org

Also free!

Create web-based visualization. Robust. Can create many kinds of visualization.

Need to learn javascript, CSS (+SVG)

“**Future-proof**” (likely to stay for many years)

Great interactive tutorial
http://vogievetsky.github.com/IntroD3/#1

# Processing

http://processing.org

“Java for designers”. Simplified Java.

Can create interactive visualization, images, and more.

Can be used as a library in normal Java app.

Many tutorials, examples.

# Illustrator / Inkscape / Xara

http://inkscape.org

The ultimate way to create visualization.

Or to edit / perfect visualization.

**Inkscape** is free!

**Illustrator** is powerful but expensive

**Xara** is the best alternative for Illustrator, on windows (less expensive, faster, easy to use)

# Design Principles

Bar chart's vertical axis should start at "0"! (Don't lie)

Follow conventions (e.g., red for negative values)

Data is the king

- minimize distraction (bold appropriately)
- Visual encodings should be meaningful

Design for legibility

- font choices, don't rotate vertical axis label

# Design Principles

Design for ease of comparison

- Use “small multiple” / panel chart
- E.g., use line thickness instead of patterns (dot, dash, etc.)
- E.g., align numbers by decimal points

Maximize data-ink ratio

# Design Principles
(what not to do)

**3D pie chart** (or 3D anything)

**Bar chart** not starting at 0

- Why not OK?
  People compare using bars' heights

**Wrong aspect ratio**

- Flatten or steepen trends

# Project

Description is out

High-level schedule

- Proposal (writeup + short presentation)
- Progress report
- Final report (writeup + poster presentation)

# George Heilmeier

Former Director of DARPA

# Heilmeier Questions

## Preflight checklist for successful projects

1. What are you trying to do?
   Articulate your objectives using absolutely no jargon.
2. How is it done today, and what are the limits of current practice?
3. What's new in your approach and why do you think it will be successful?
4. Who cares?
5. If you're successful, what difference will it make?
6. What are the risks and payoffs?
7. How much will it cost?
8. How long will it take?
9. What are the midterm and final "exams" to check for success?

http://en.wikipedia.org/wiki/George_H._Heilmeier
http://smlv.cc.gatech.edu/2010/10/17/heilmeiers-questions/