http://poloclub.gatech.edu/cse6242

CSE6242: Data & Visual Analytics

# Common visualization Issues & how to fix them


Duen Horng (Polo) Chau

Associate Professor, College of Computing

Associate Director, MS Analytics

Georgia Tech

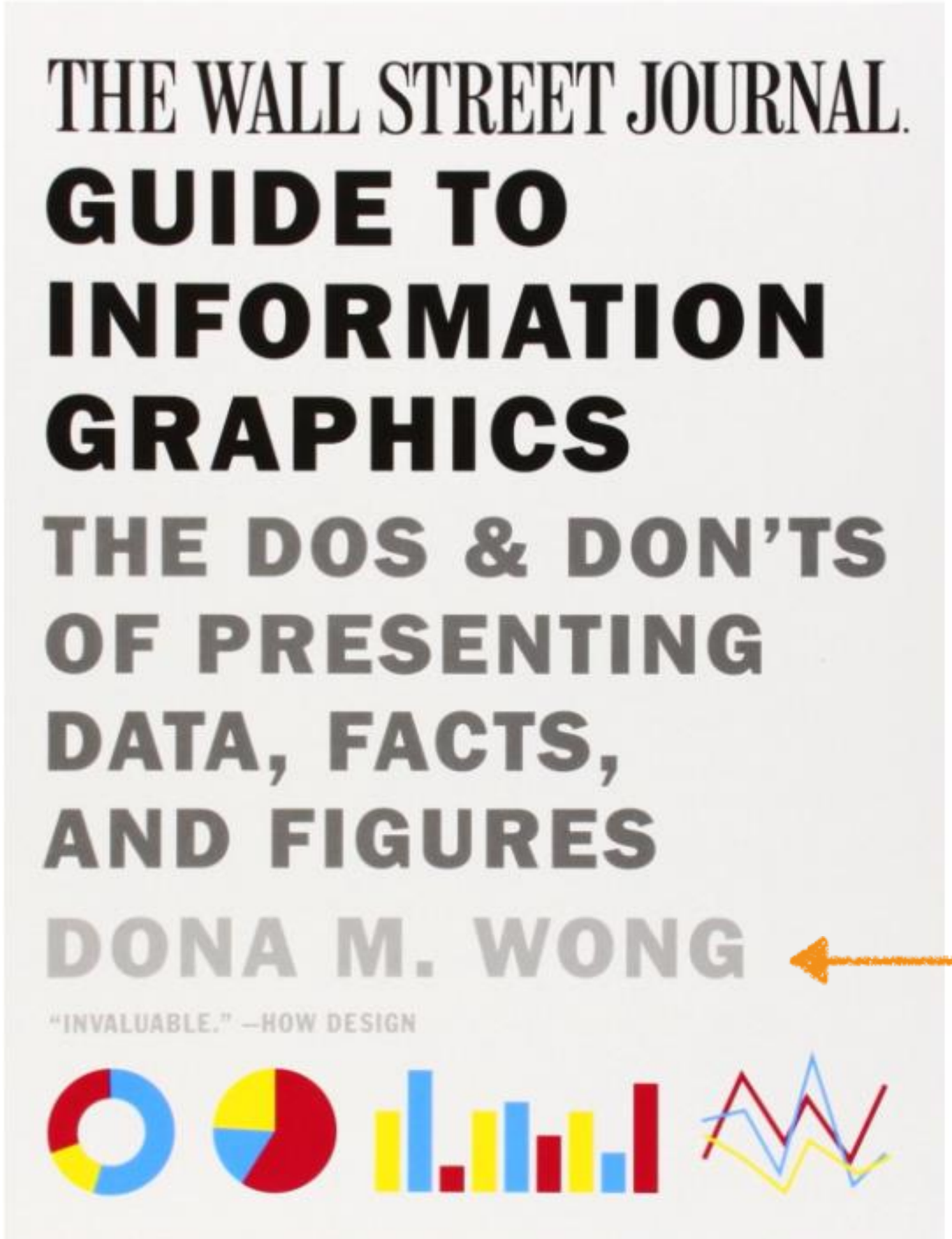
Mahdi Roozbahani

Lecturer, Computational Science & Engineering, Georgia Tech

Founder of Filio, a visual asset management platform


Partly based on materials by

Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos

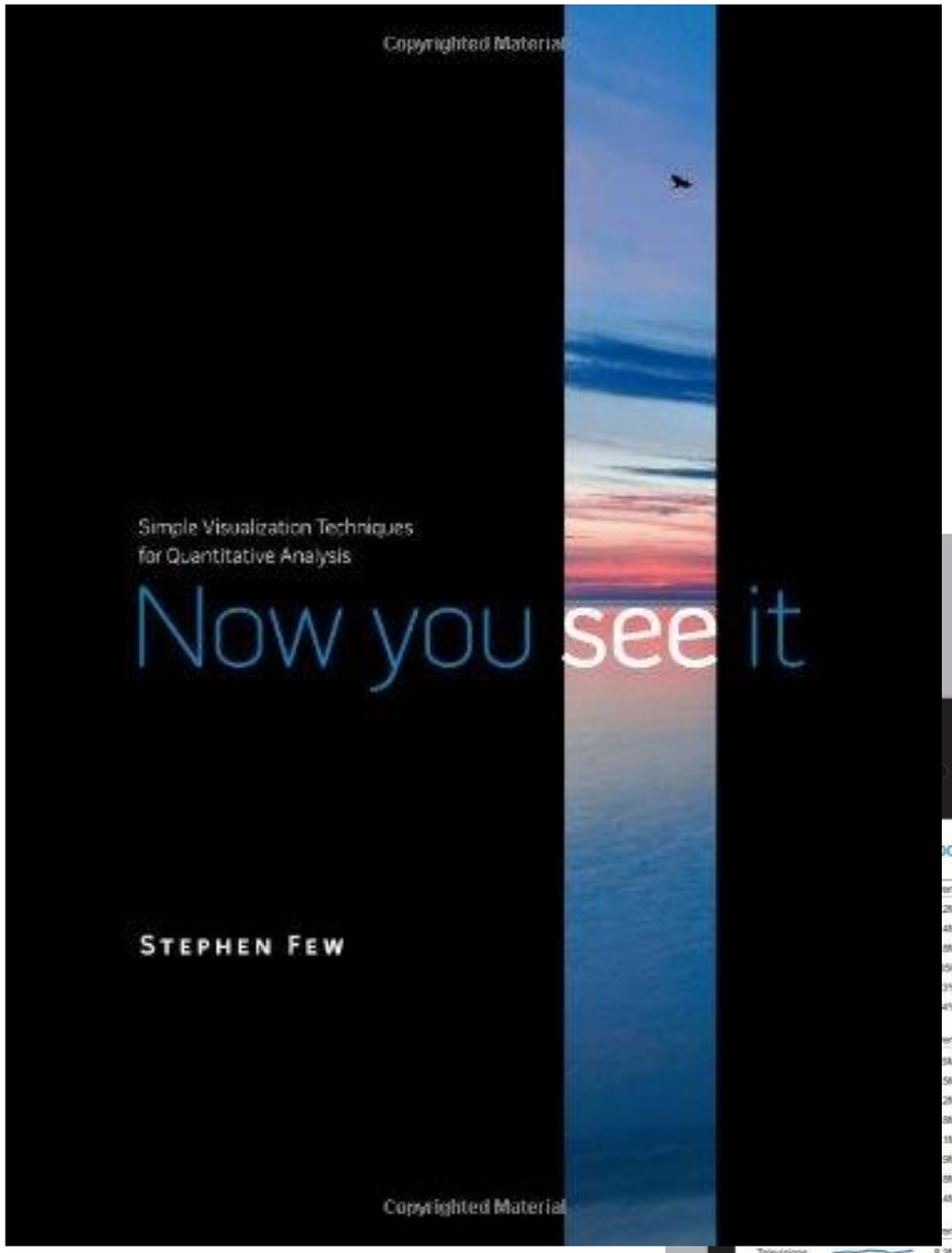


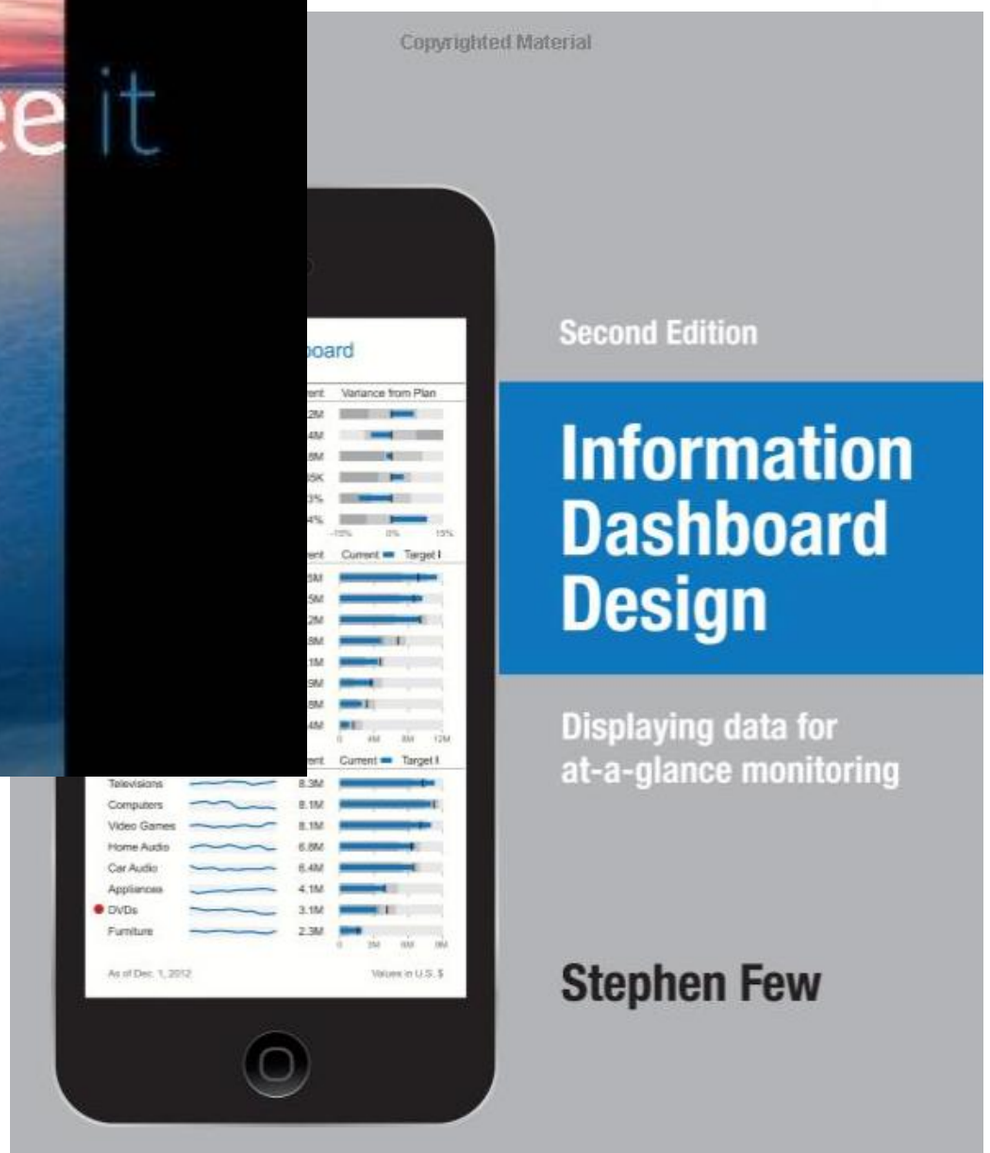

Student of Edward Tufte

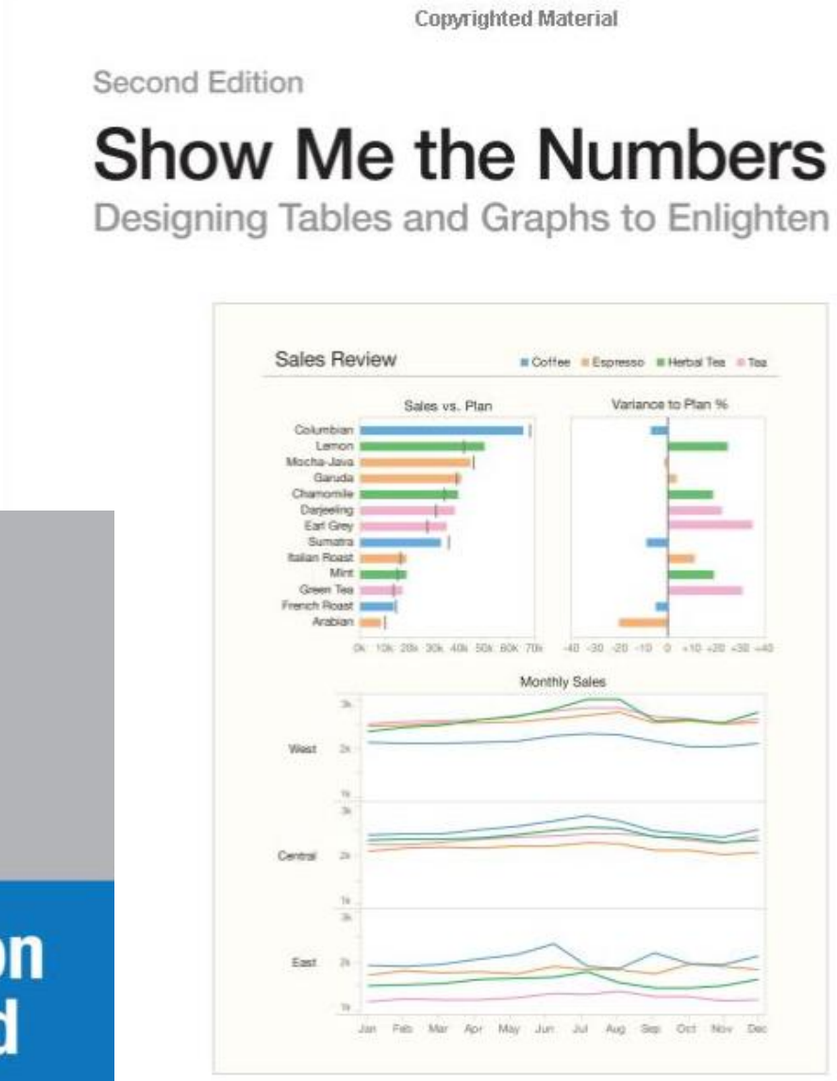


http://a.co/6BhlPfZ

# Also Highly Recommended:



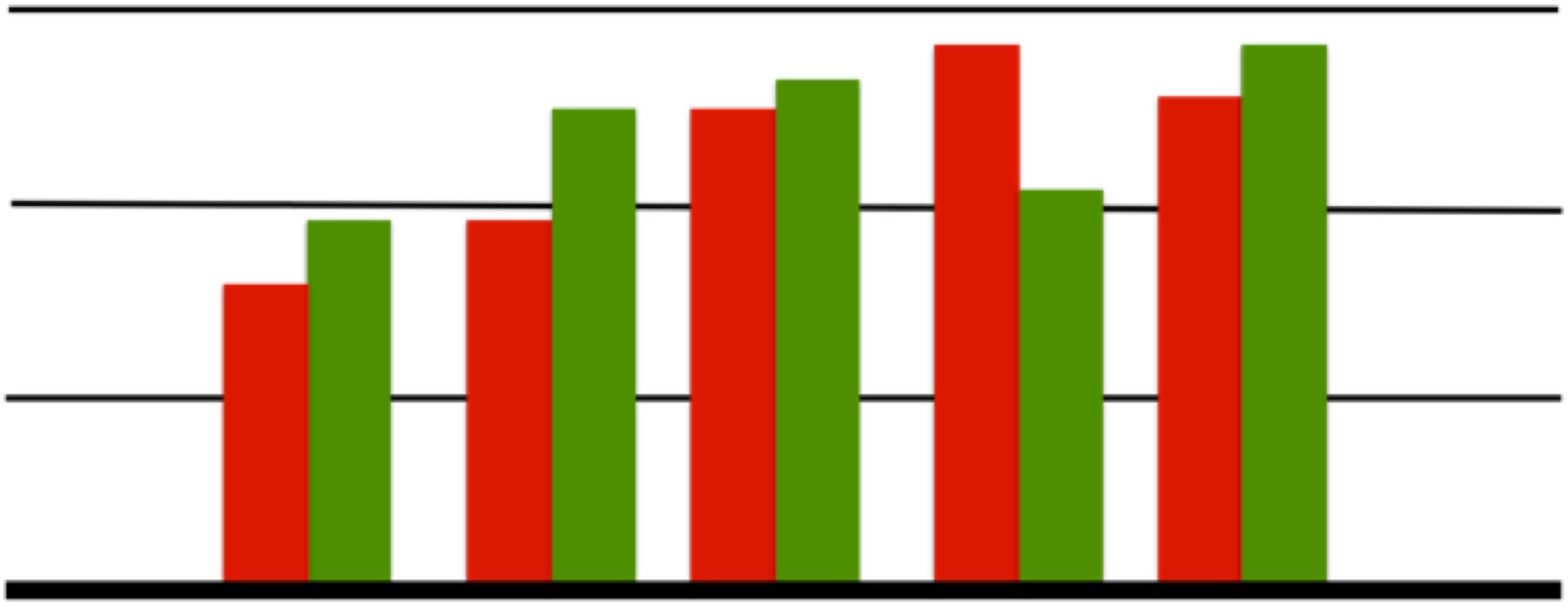



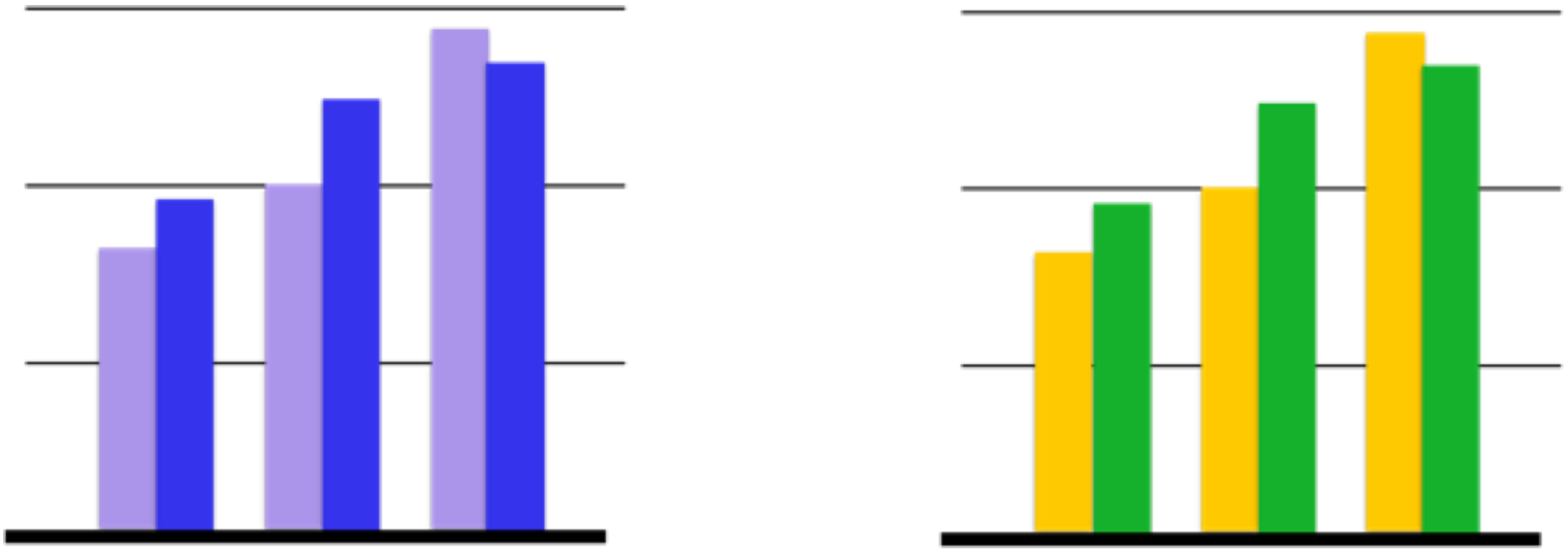

Stephen Few

Copyrighted Material

# Bar Charts



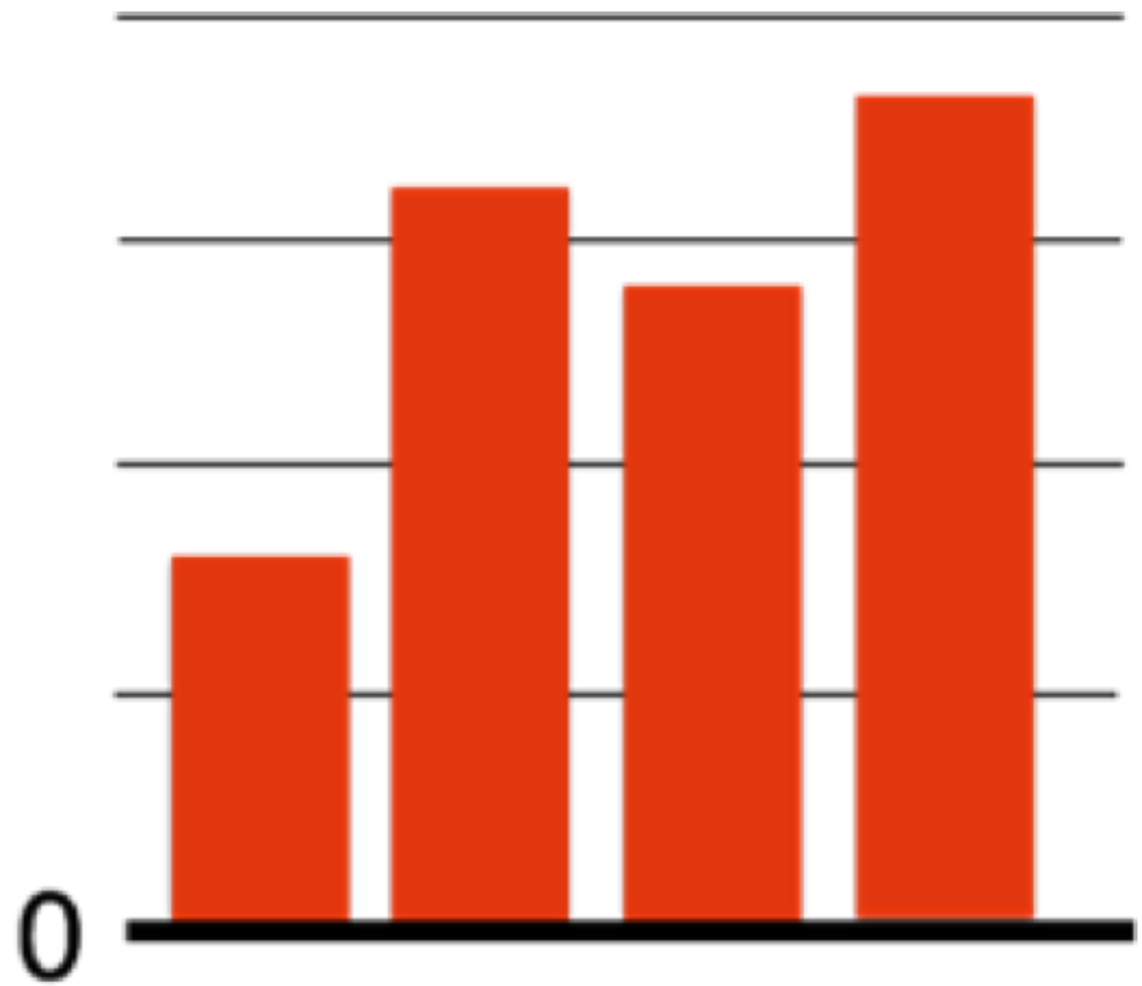
The color scheme reminds you of what?



# Better than Christmas

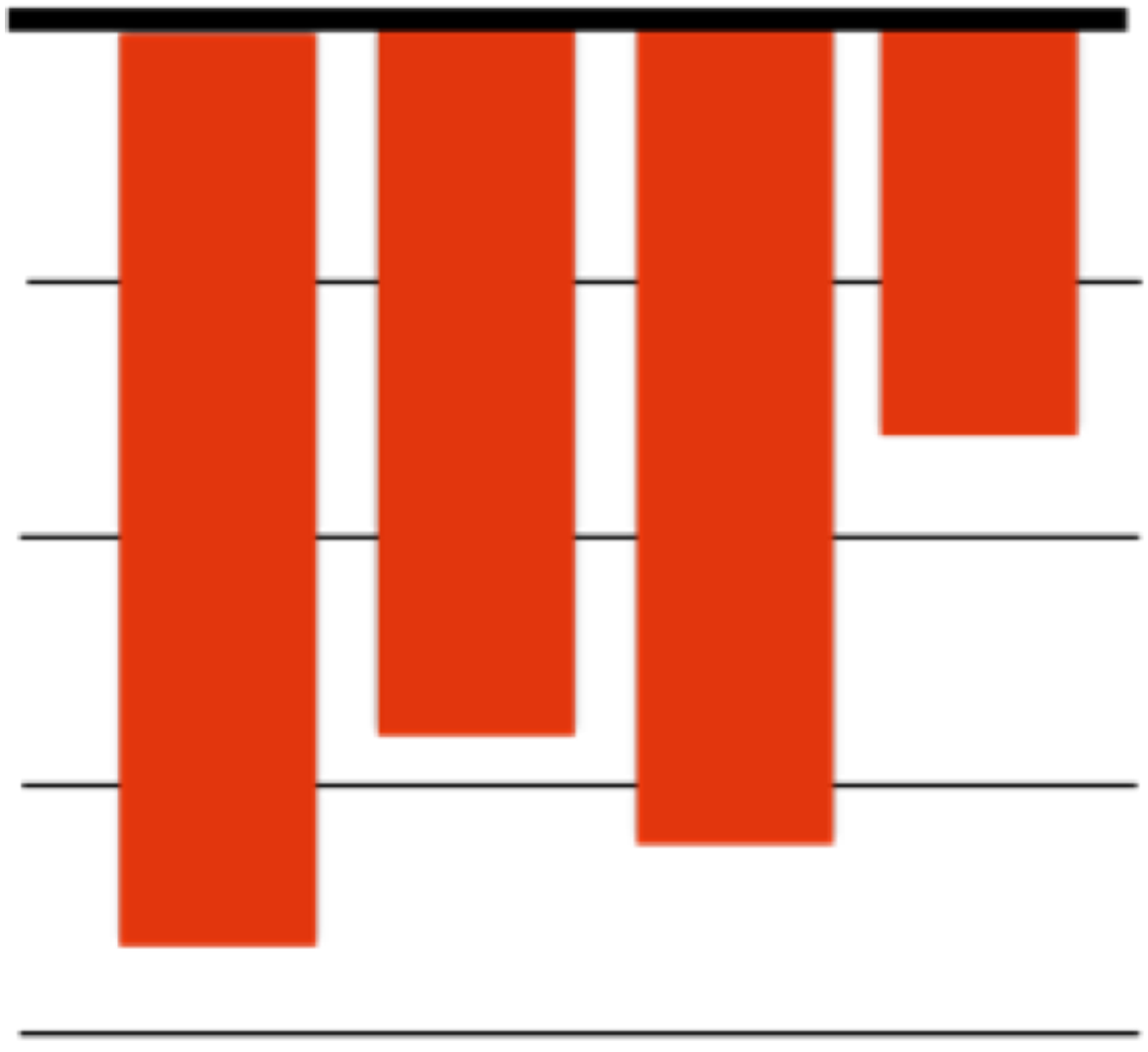
(Use color brewer to find good color schemes)

# Company Profits

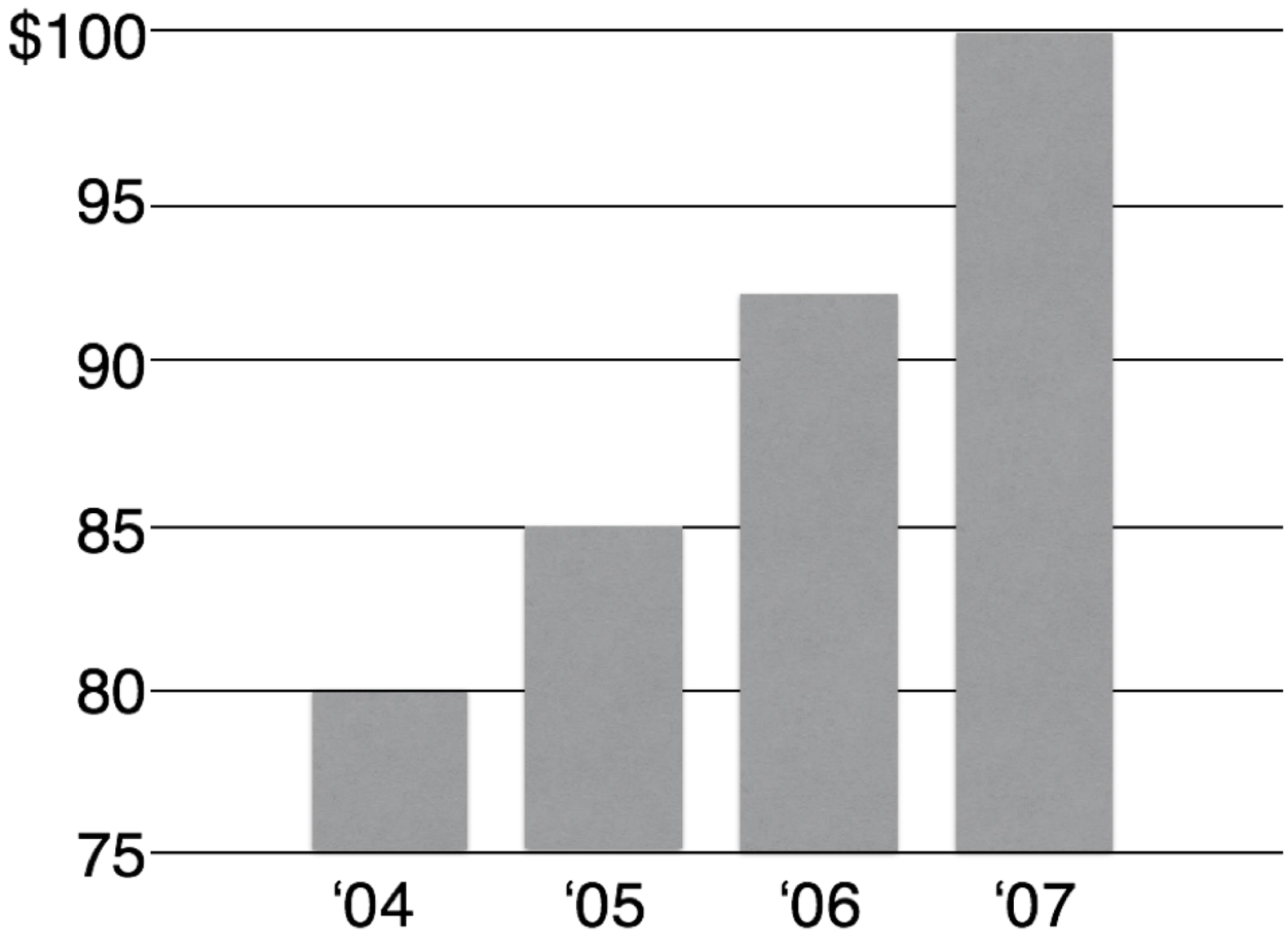


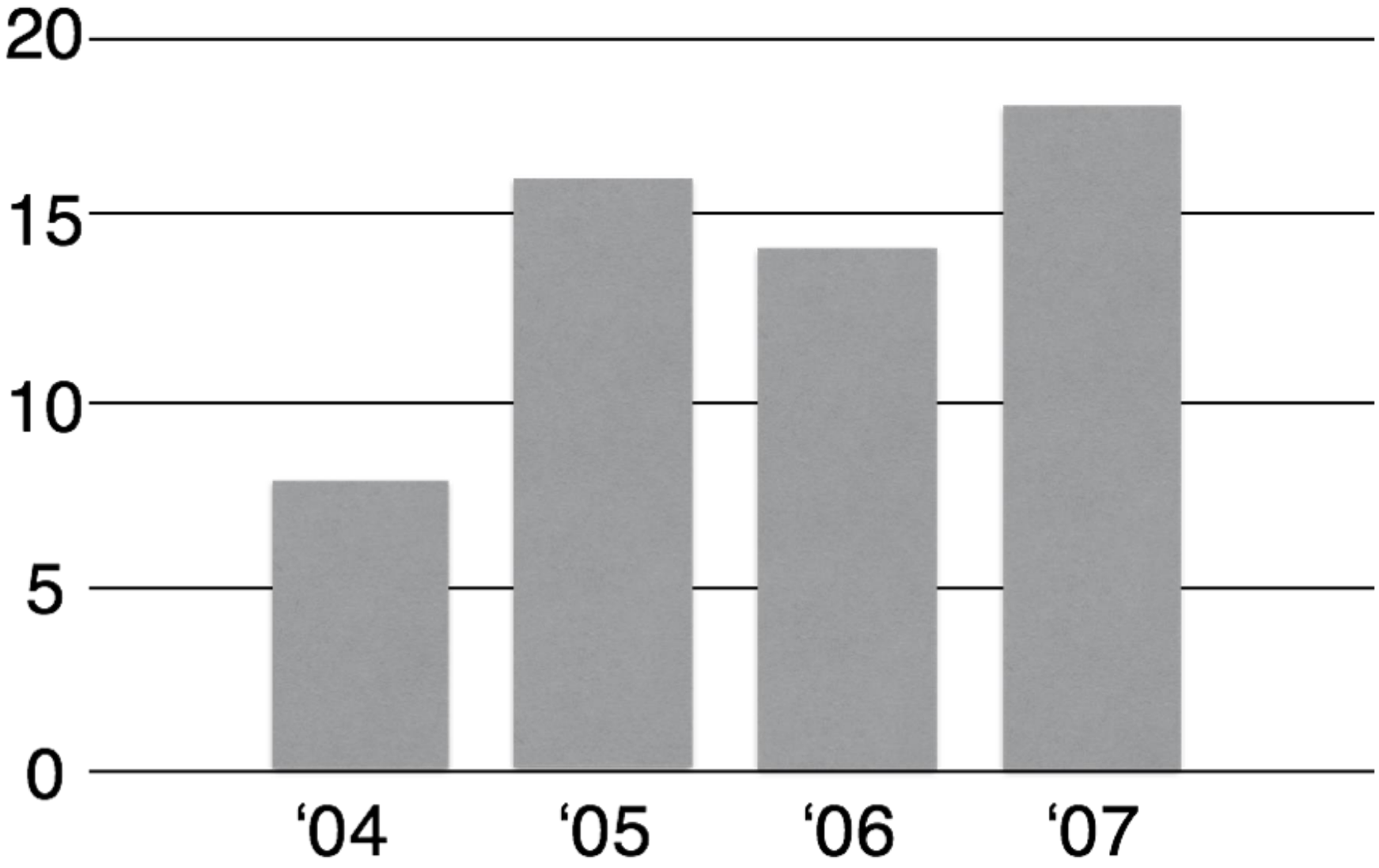

Don’t show profits in red!!

Think carefully about your color choices.



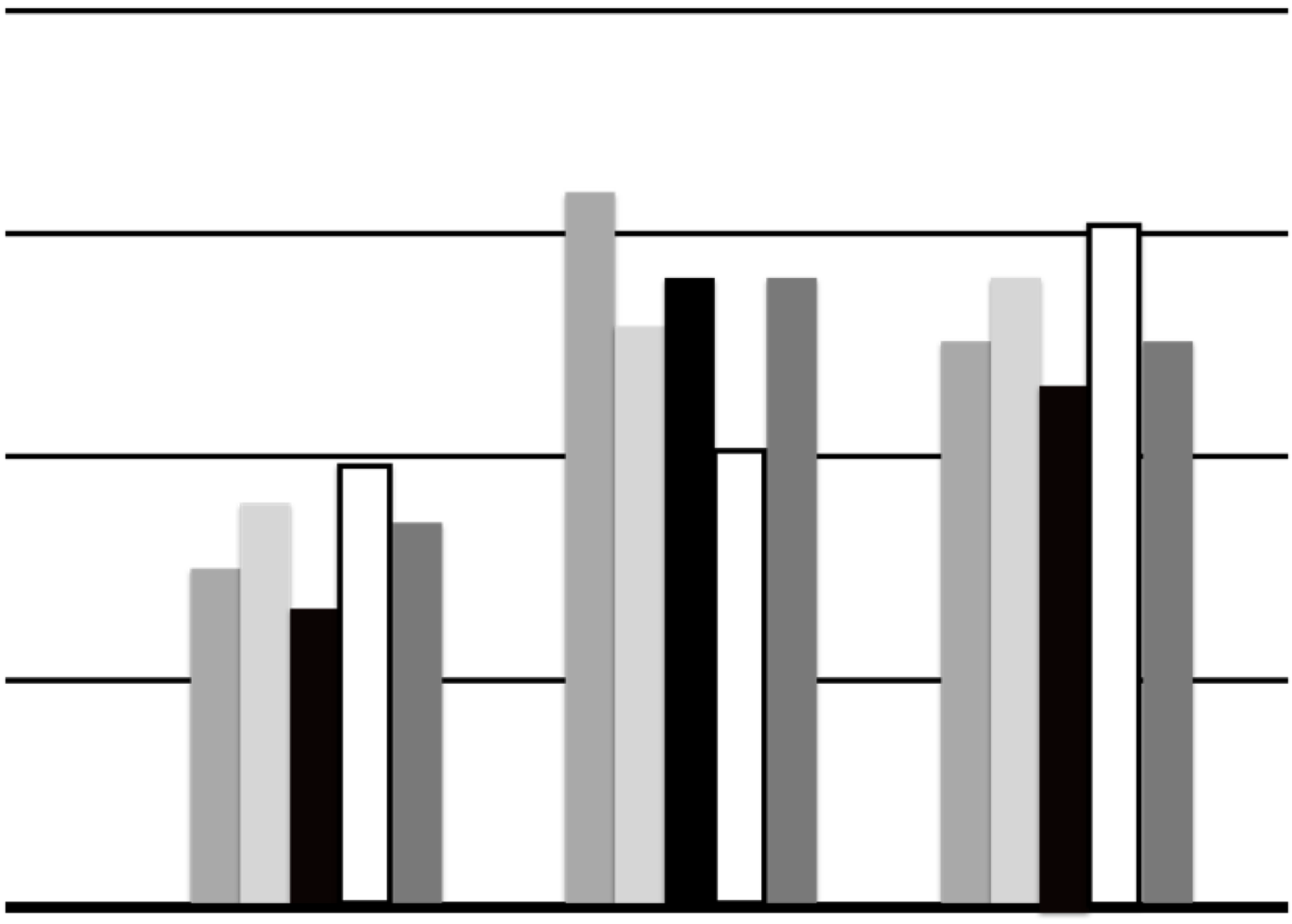
# Misleading Bar Charts

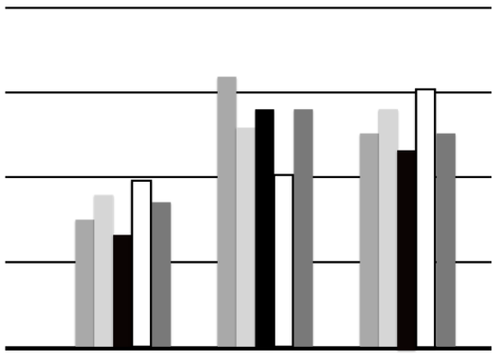




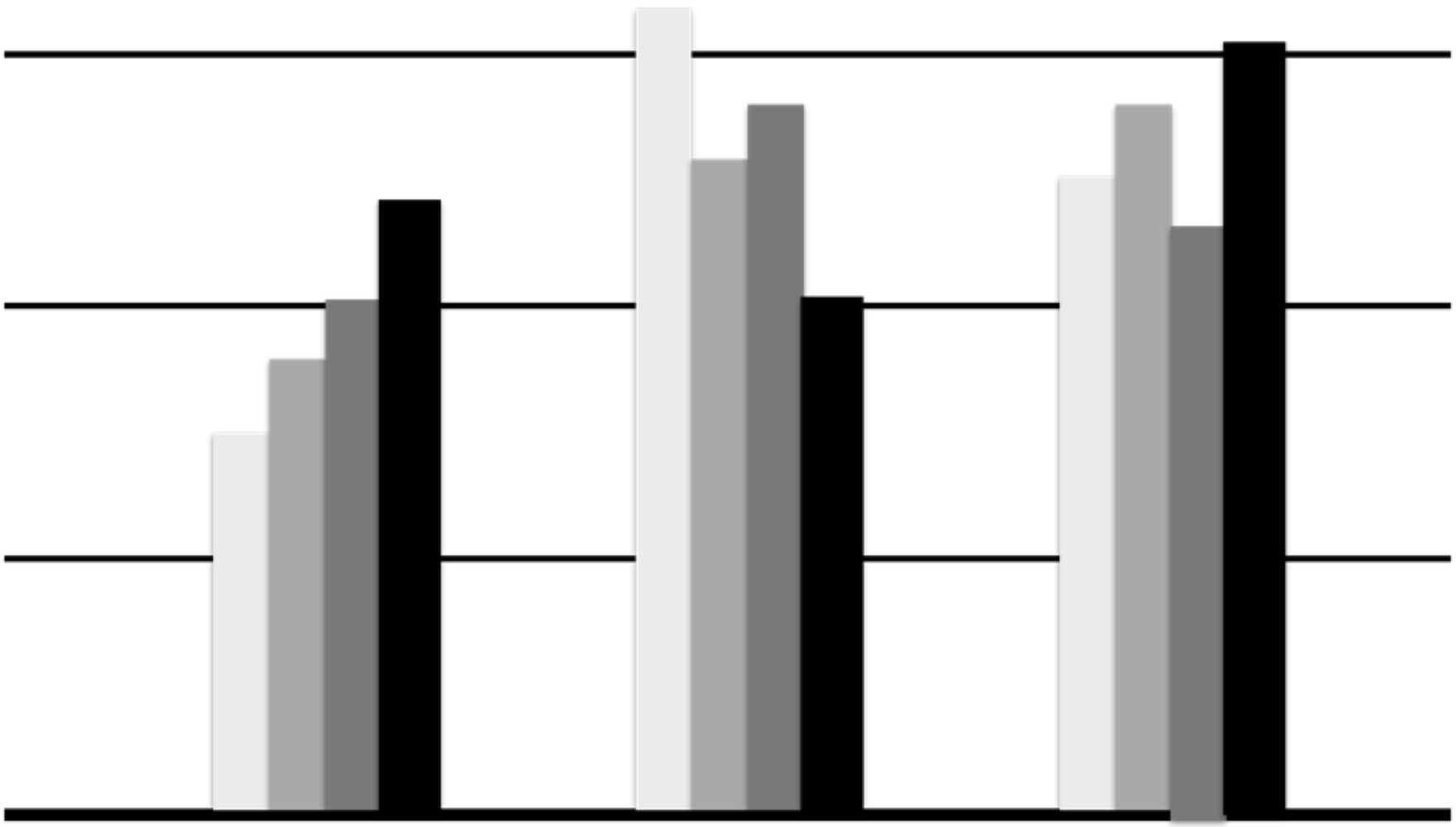

# Vertical axis of bar charts should start at 0, almost always

# Disorienting color bars

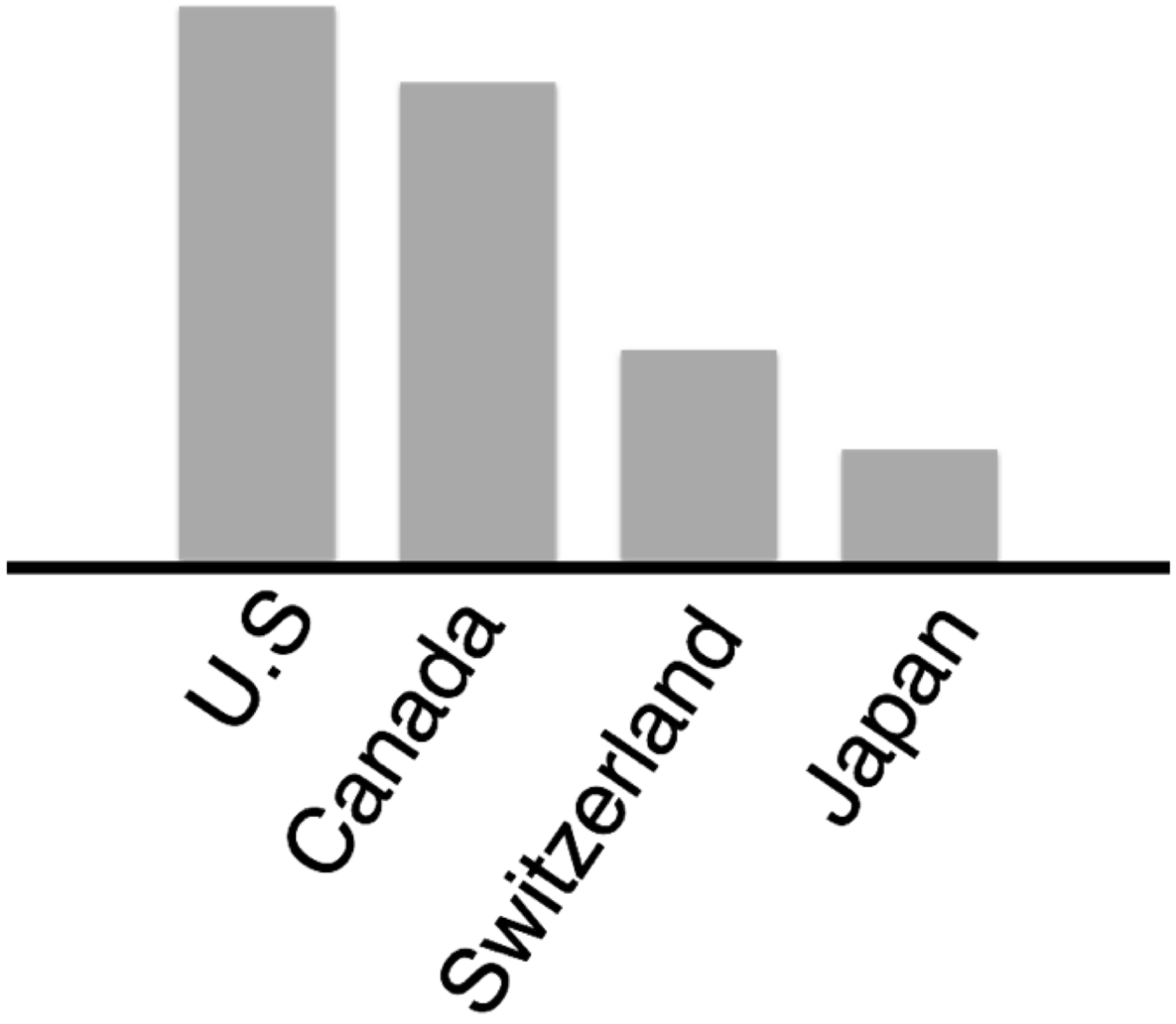




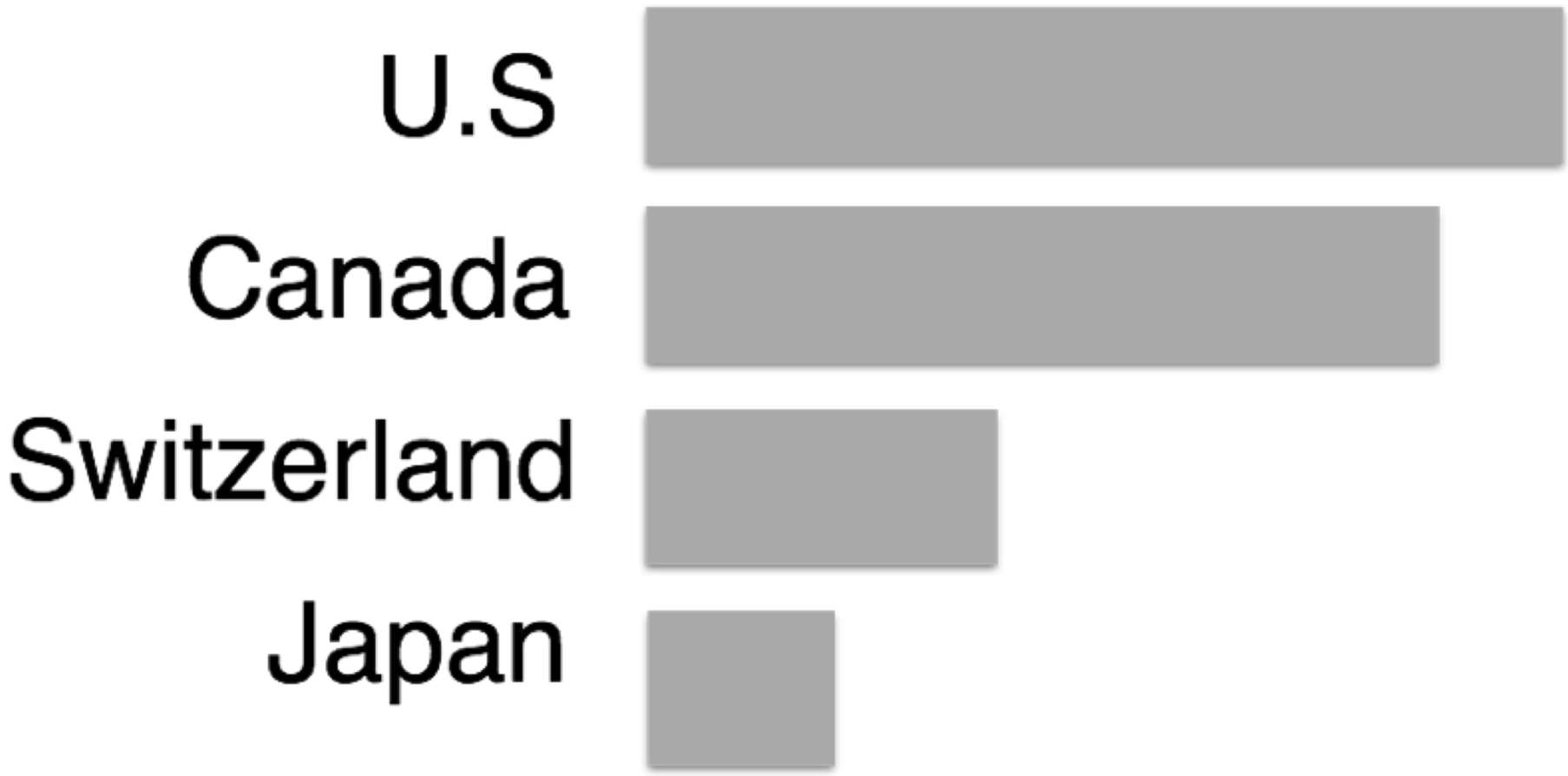
# Use gradation



# Avoid Tilted or Rotated Labels



# Bars Can be **Horizontal**



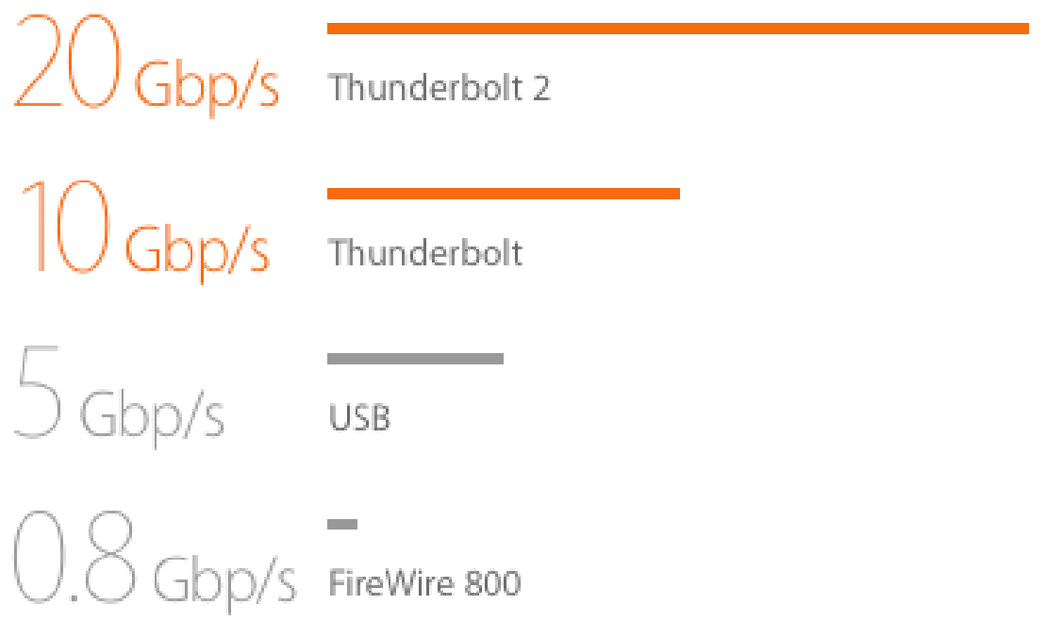

When labels are hard to read, try horizontal layout.

**Don’t settle for the default.**

Thunderbolt 2

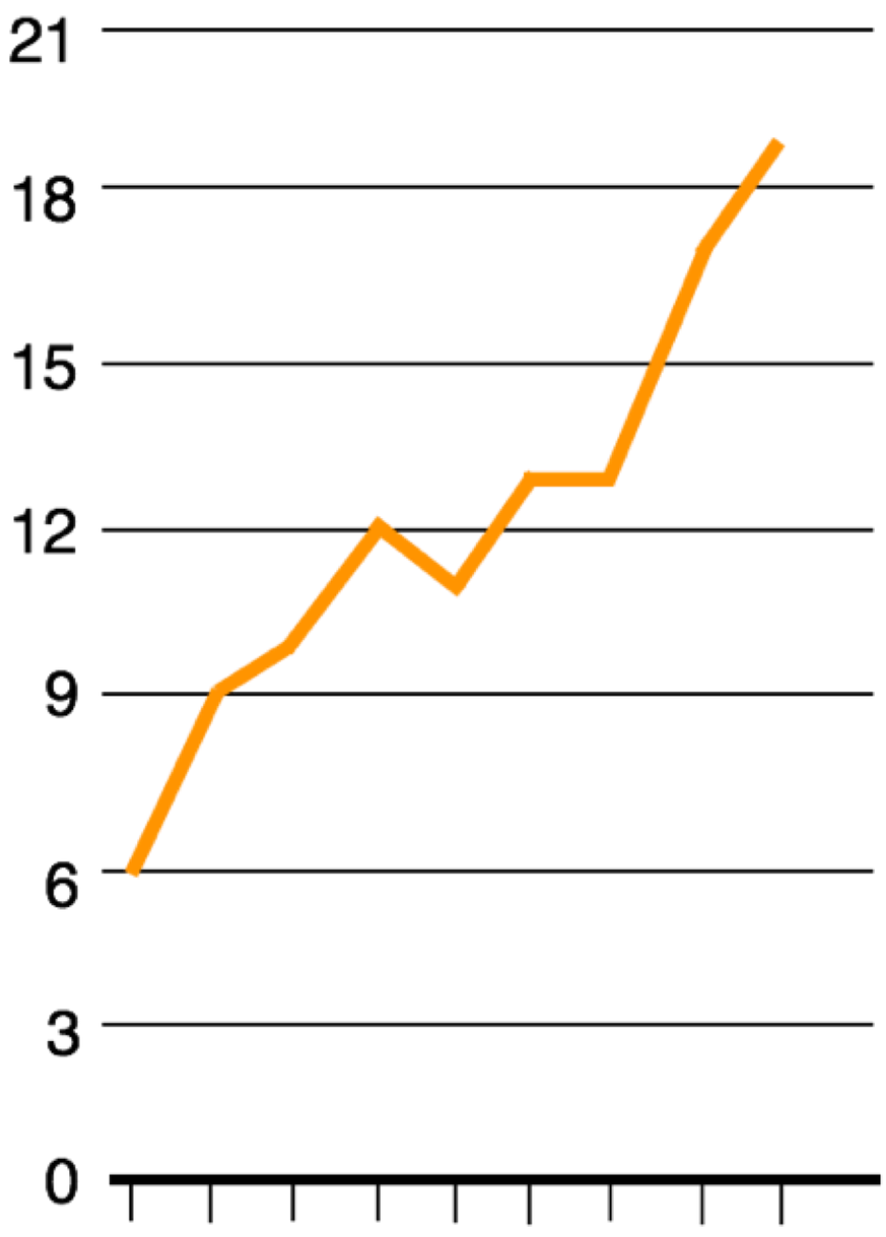
# Blazing-fast data transfer.

Each of the two ultrafast, ultraflexible Thunderbolt 2 ports significantly expand the capabilities of iMac. For example, you can connect high-performance peripherals and move data up to 40 times faster than with USB 2, and up to 25 times faster than with FireWire 800. You also have more than enough bandwidth — up to 20 Gbps — to daisy-chain multiple high-speed devices and still maintain maximum throughput.



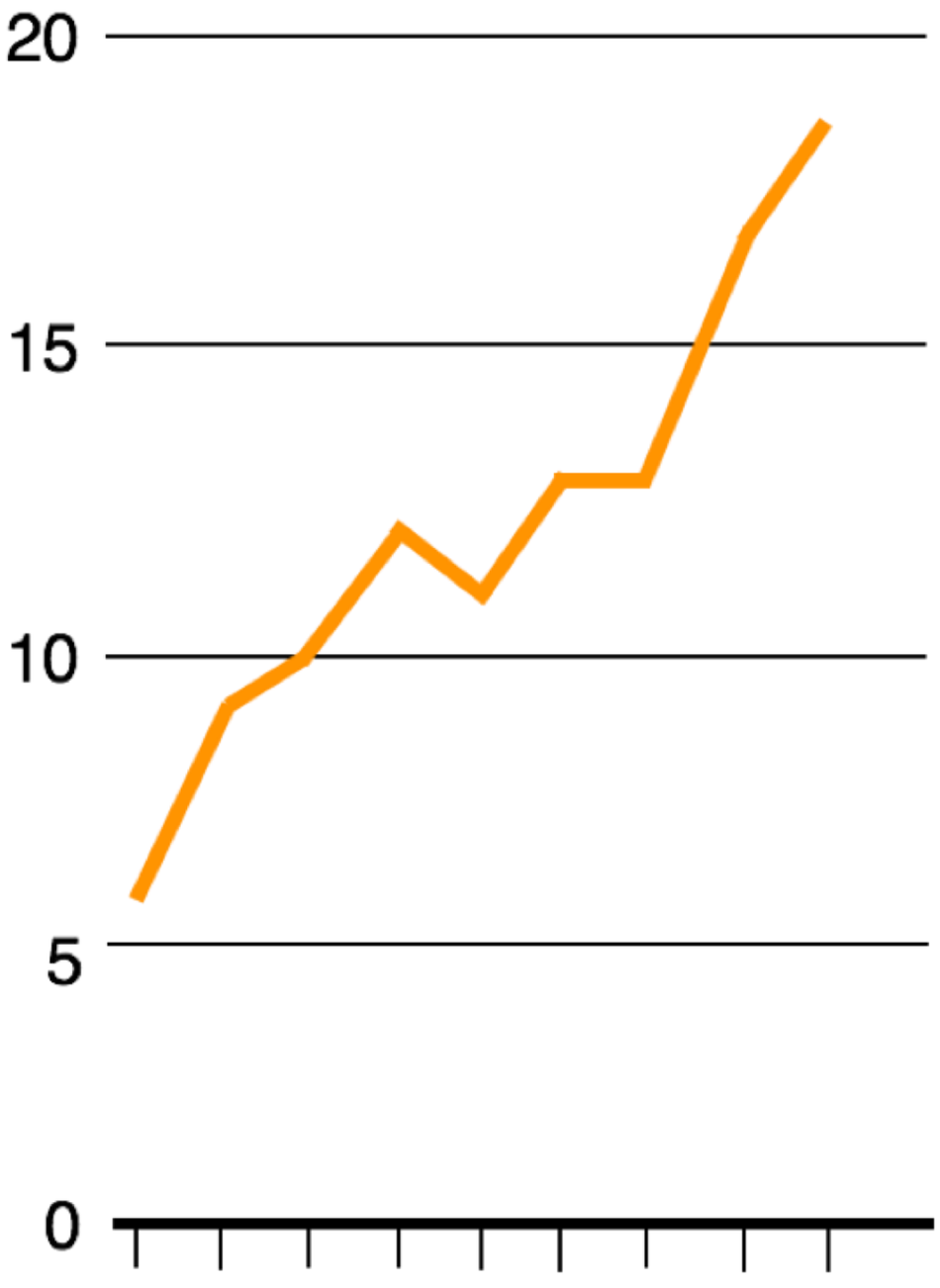

http://www.apple.com/imac/performance/

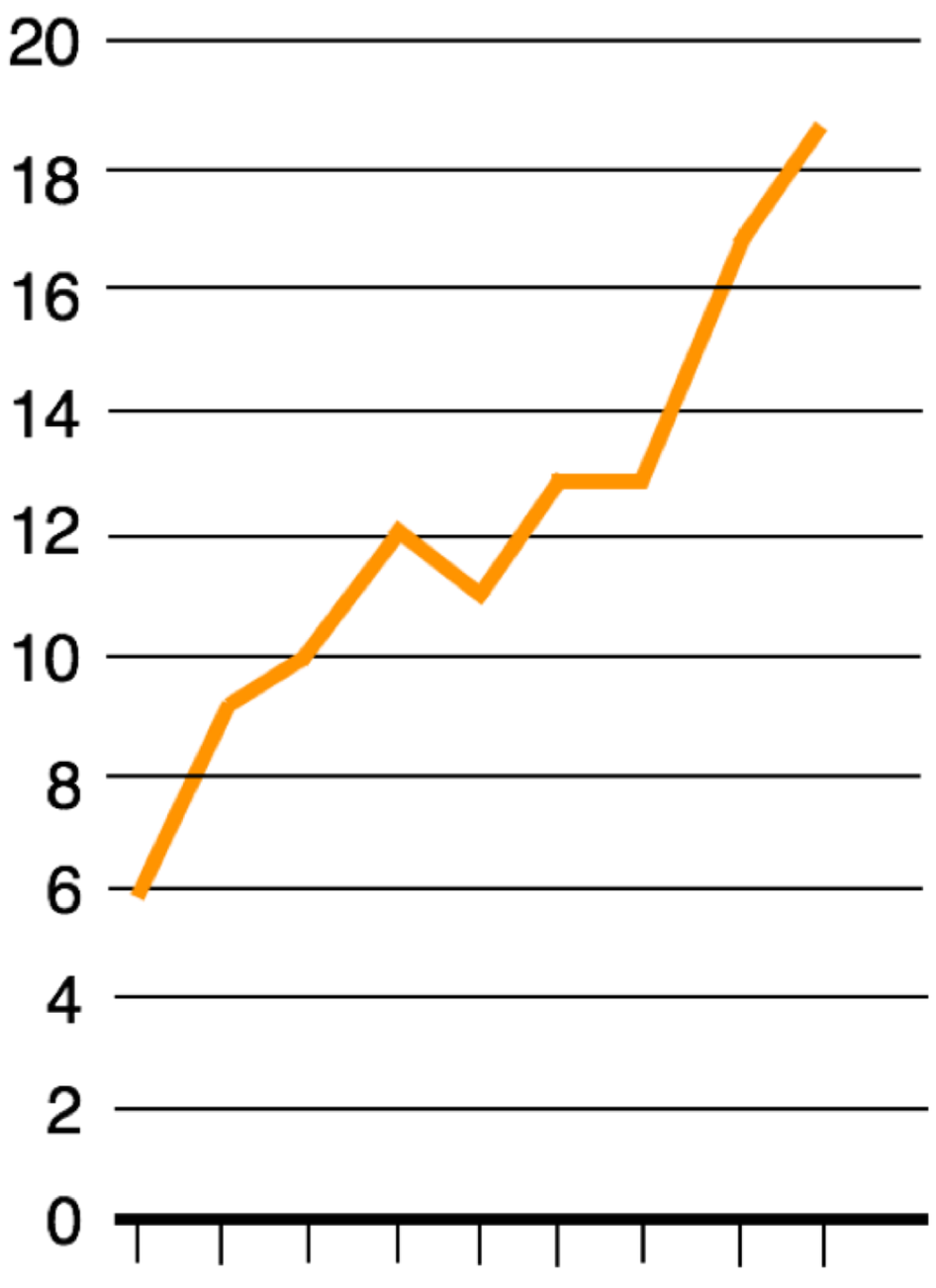
# Line Charts (a.k.a. fever lines)



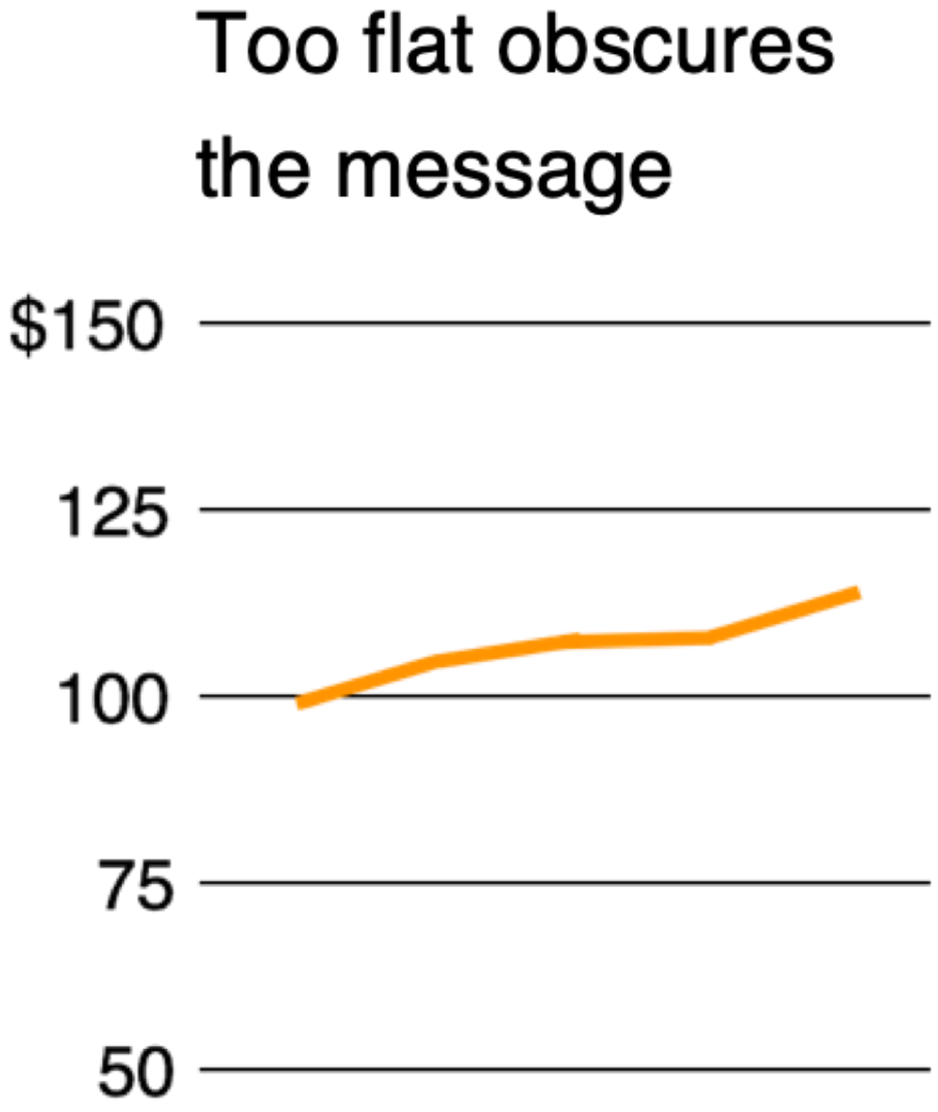

Can you improve the tick labels?

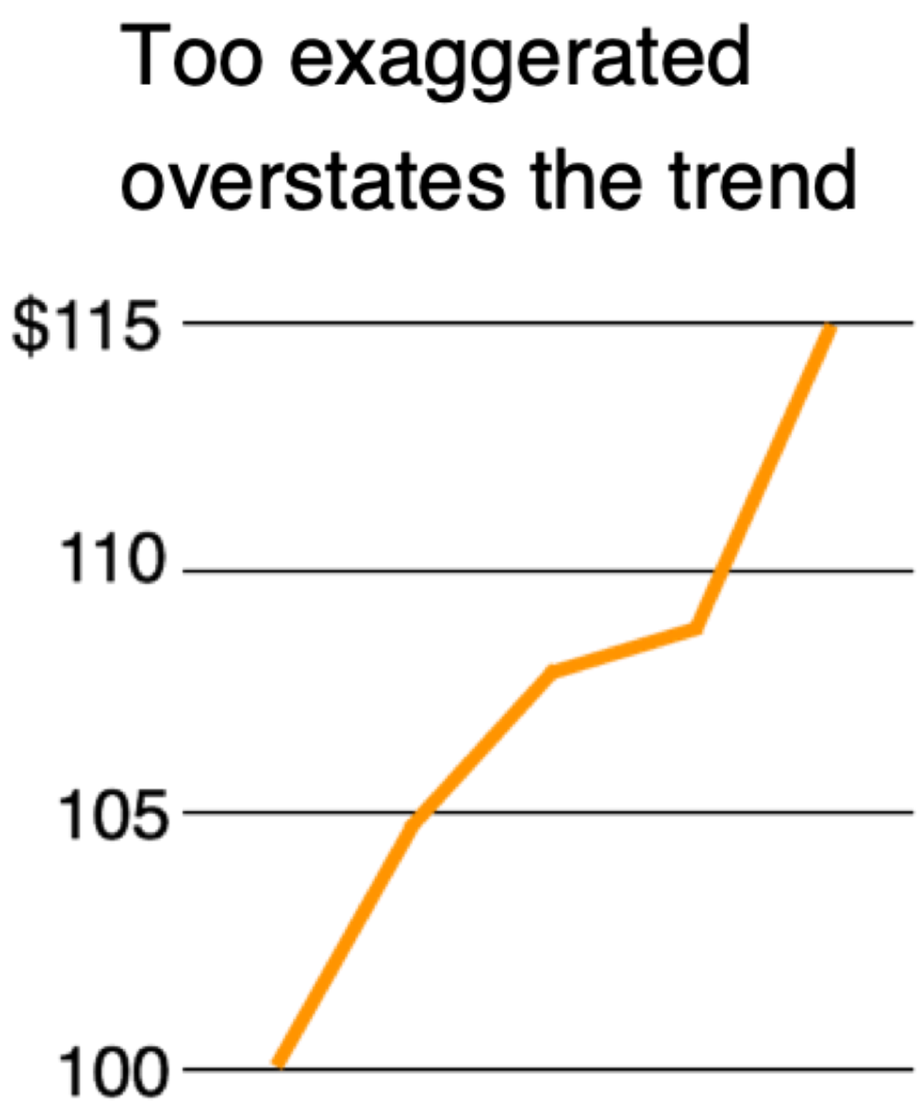
# Use ticks at **common** intervals (e.g., 2, 5, 10, etc.)



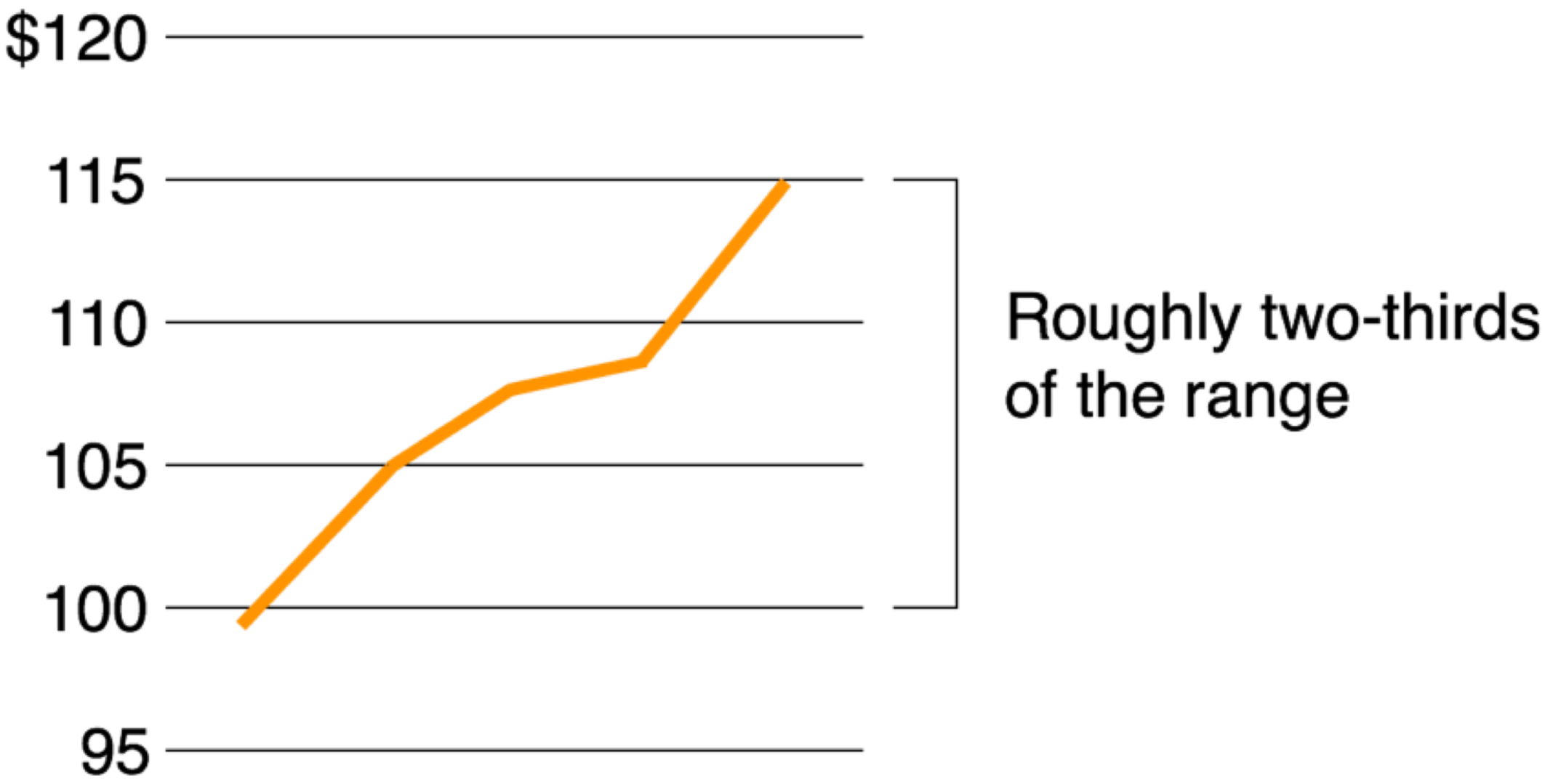
# Too flat or too steep?



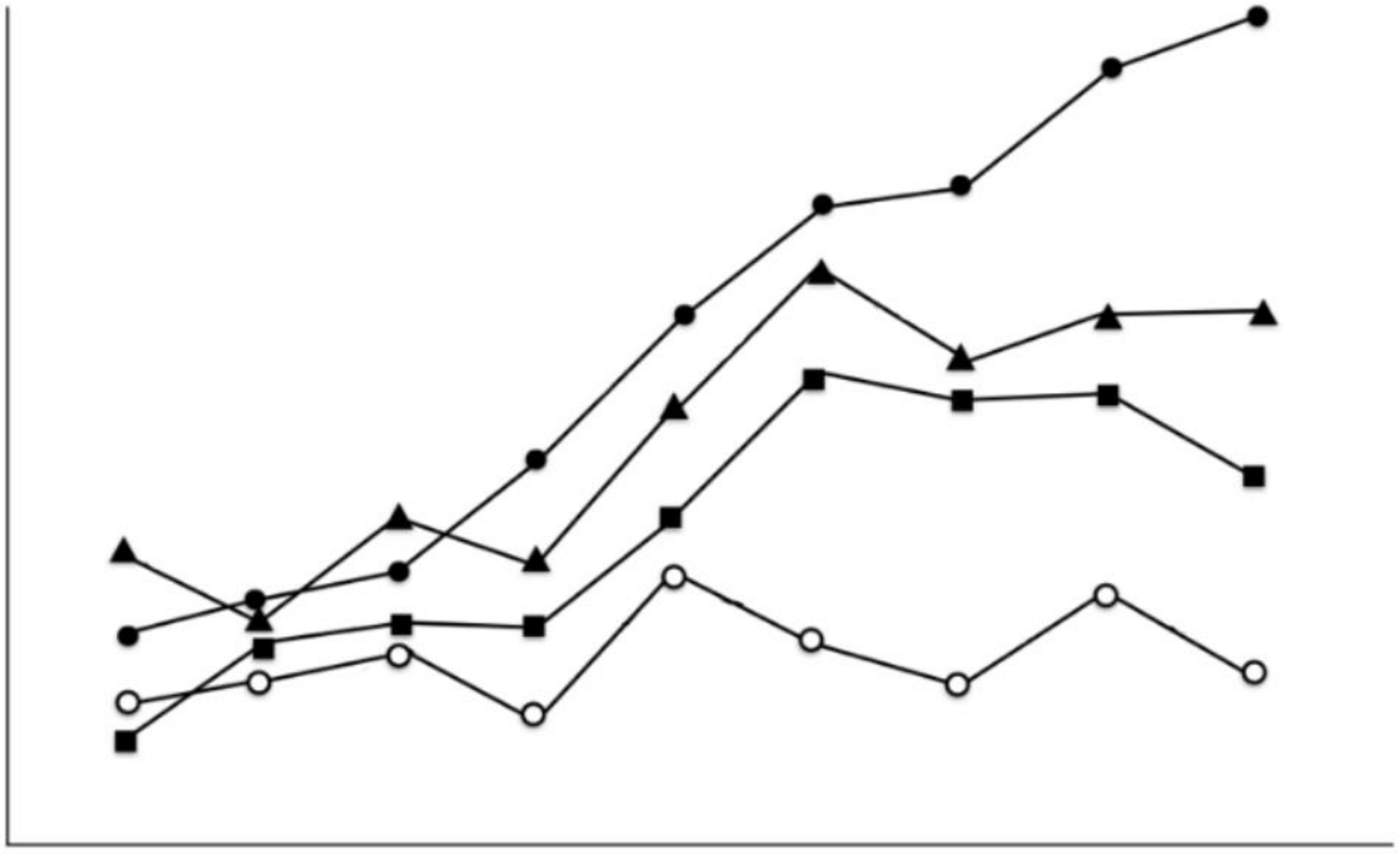



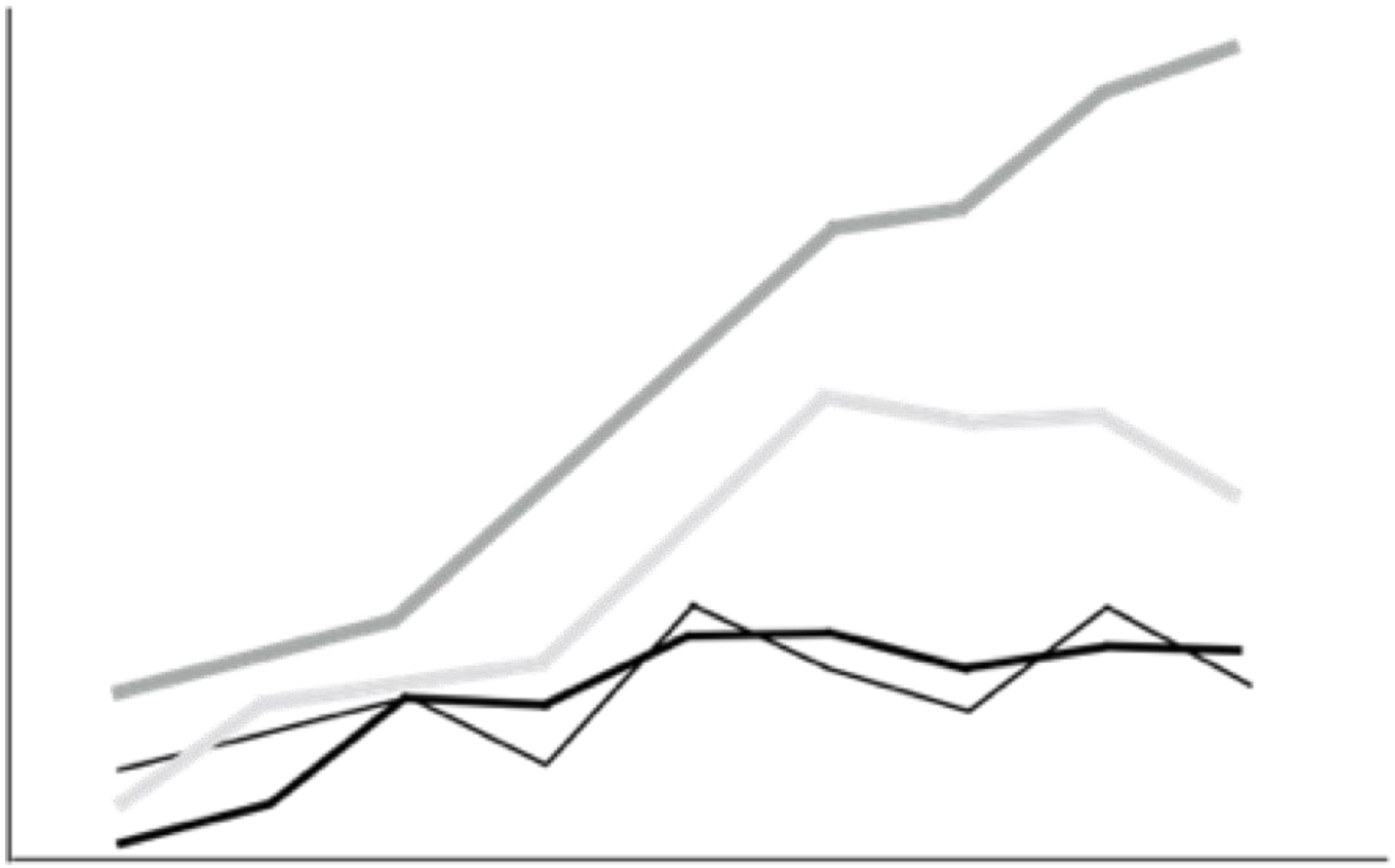

Note y-axis does not need to start at 0.
Why not as bad as in the case of bar chart?

# Rule of Thumb



# Multiple Patterned Lines in one chart

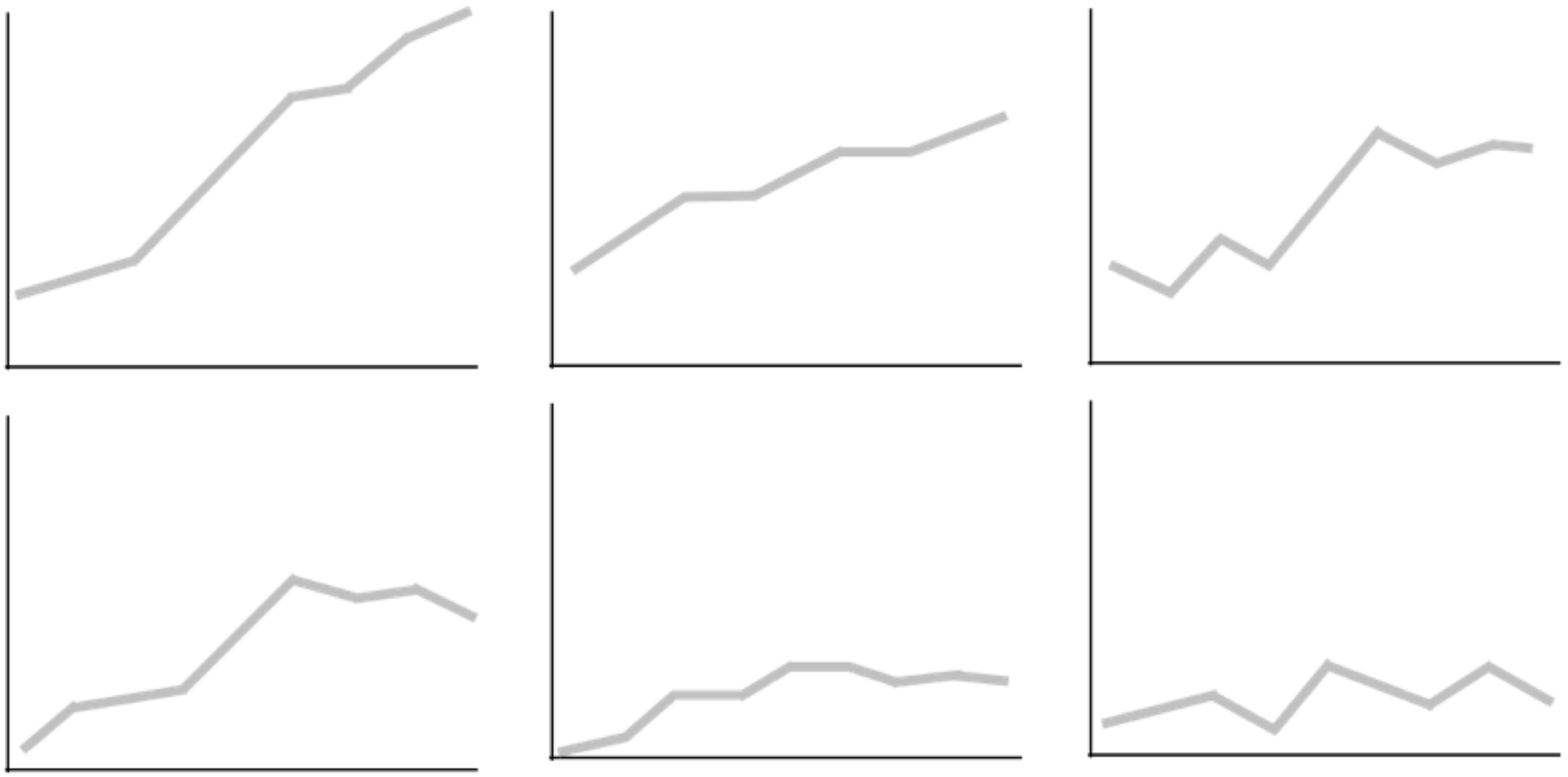




Better?

Note the “double encoding” of *line width* and *brightness*.

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,

# Tables

| 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** |

What can you improve?

# What’s the problem with making everything **bold** or *italic*?

# “Everyone is special” → “No one is”

https://youtu.be/1E9pKU_N15A

# When everyone is special, no one is!

| 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 |

# Higher “data to ink” ratio

# Problems?

| 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 |

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

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

# Beautiful Publication-quality LaTeX Tables

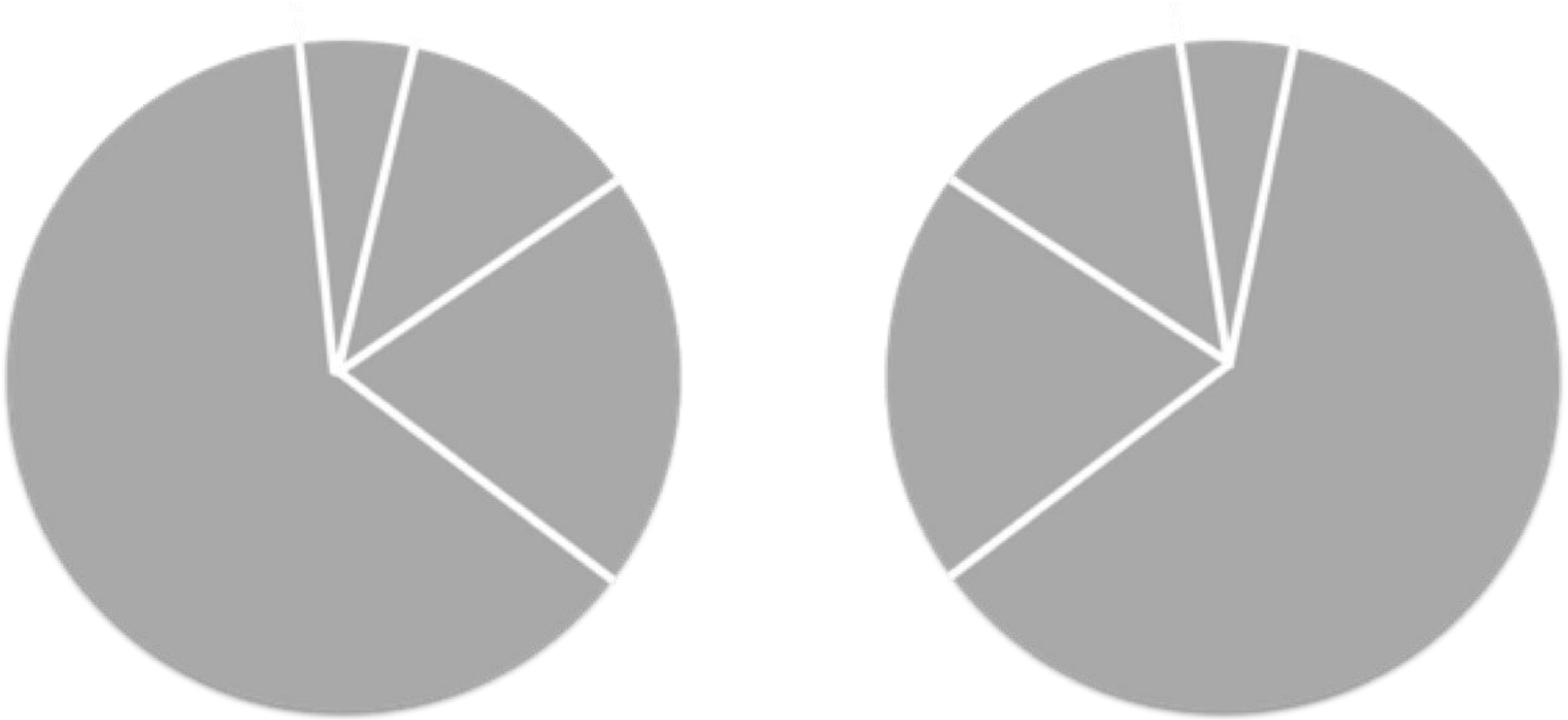
| slices | abs. error (%) | | abs. error (slices) | |
|---|---|---|---|---|
| | avg. | max. | avg. | max |
| $< 5000$ | 7.4 | 73.5 | 116 | 625 |
| 5000–10000 | 3.1 | 27.2 | 209 | 1807 |
| 10000–15000 | 2.4 | 15.6 | 297 | 2133 |
| $> 15000$ | 1.8 | 9.0 | 317 | 1609 |

https://tex.stackexchange.com/questions/112343/beautiful-table-samples

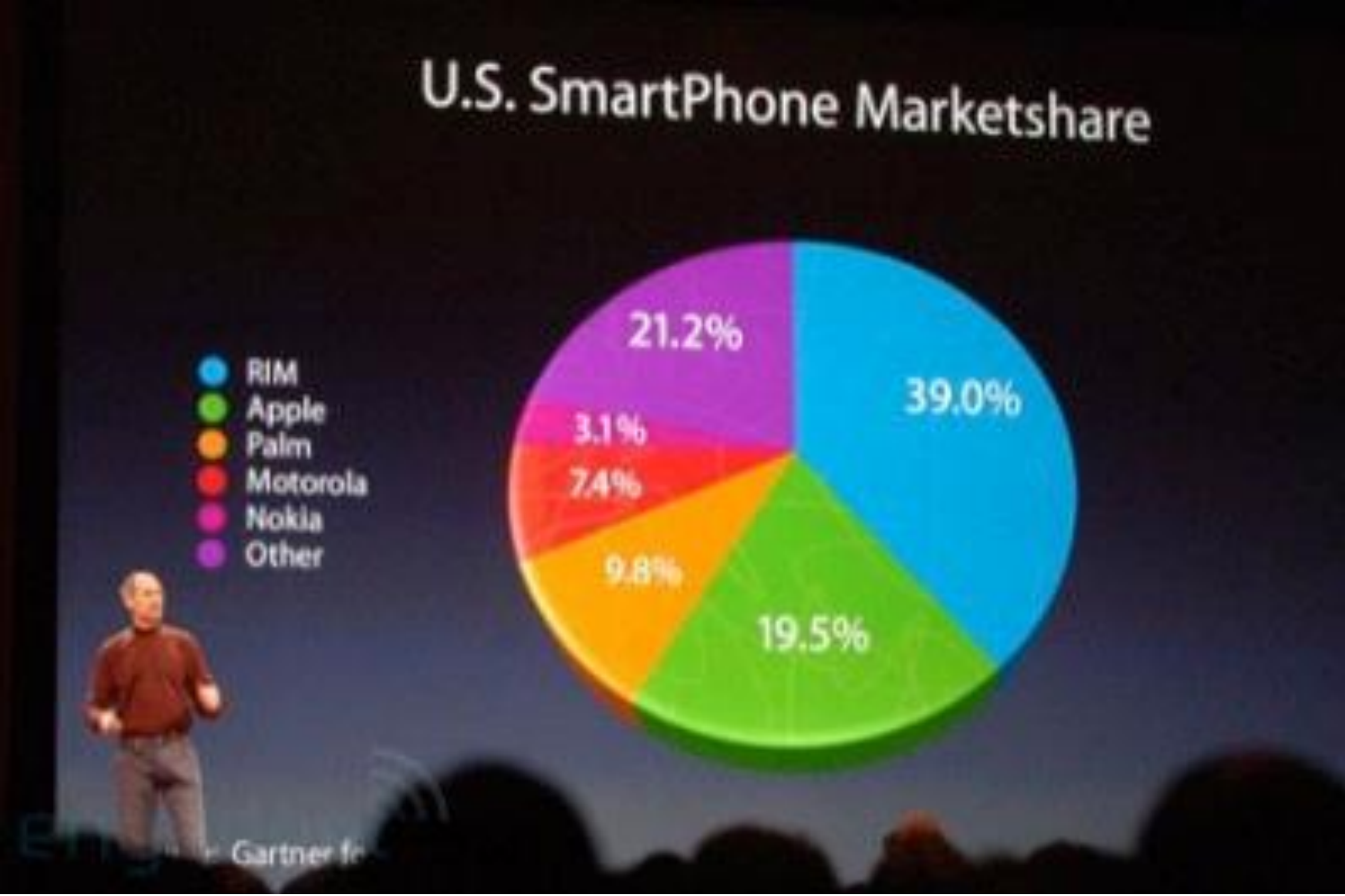
**Short guide:** https://www.inf.ethz.ch/personal/markusp/teaching/guides/guide-tables.pdf

Long guide: http://cpansearch.perl.org/src/LIMAONE/LaTeX-Table-v1.0.6/examples/examples.pdf

# The Dreaded Pie Charts



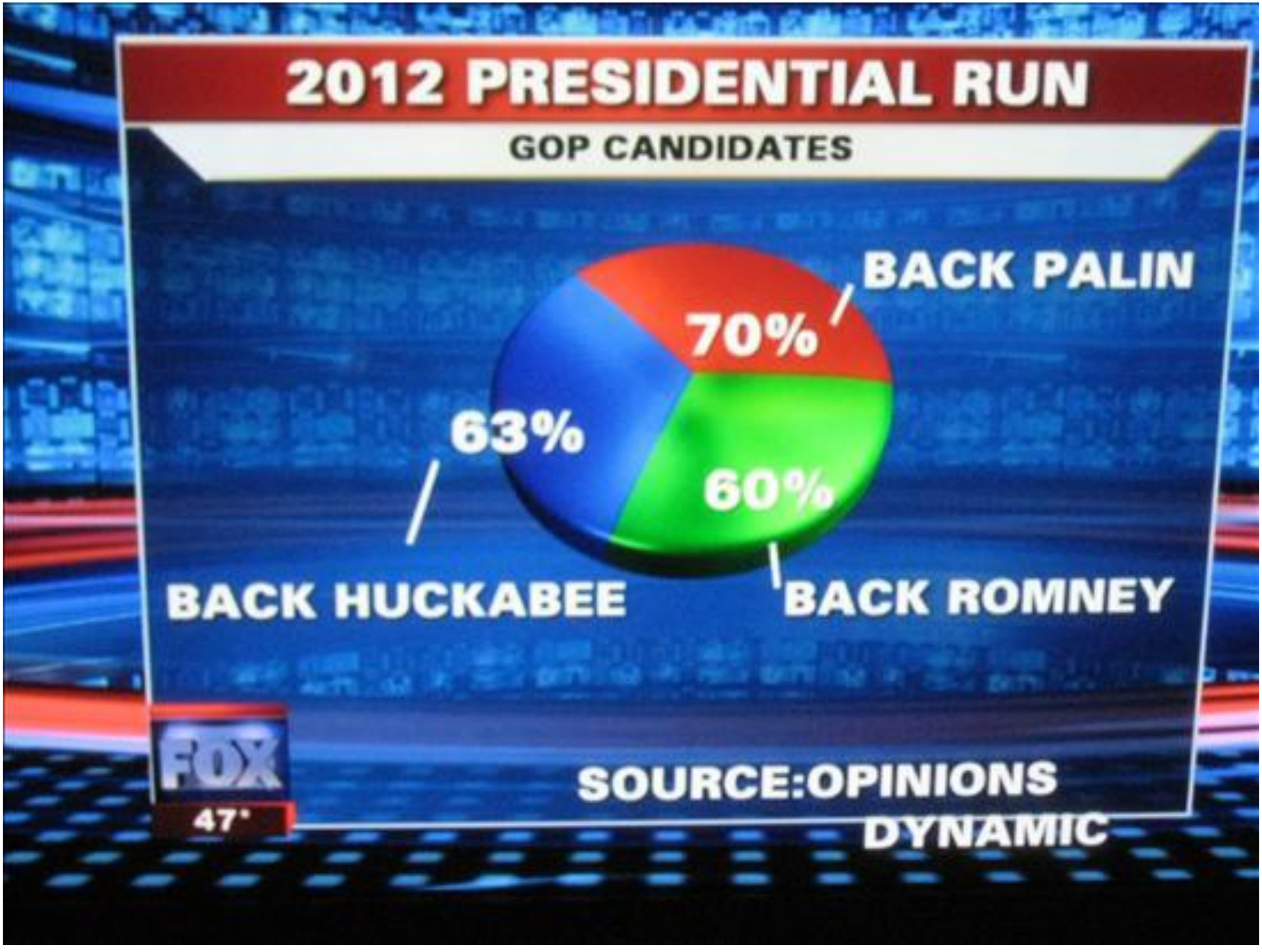
# Why people like to use pie charts?




http://www.wired.com/2008/02/macworlds-iphon/

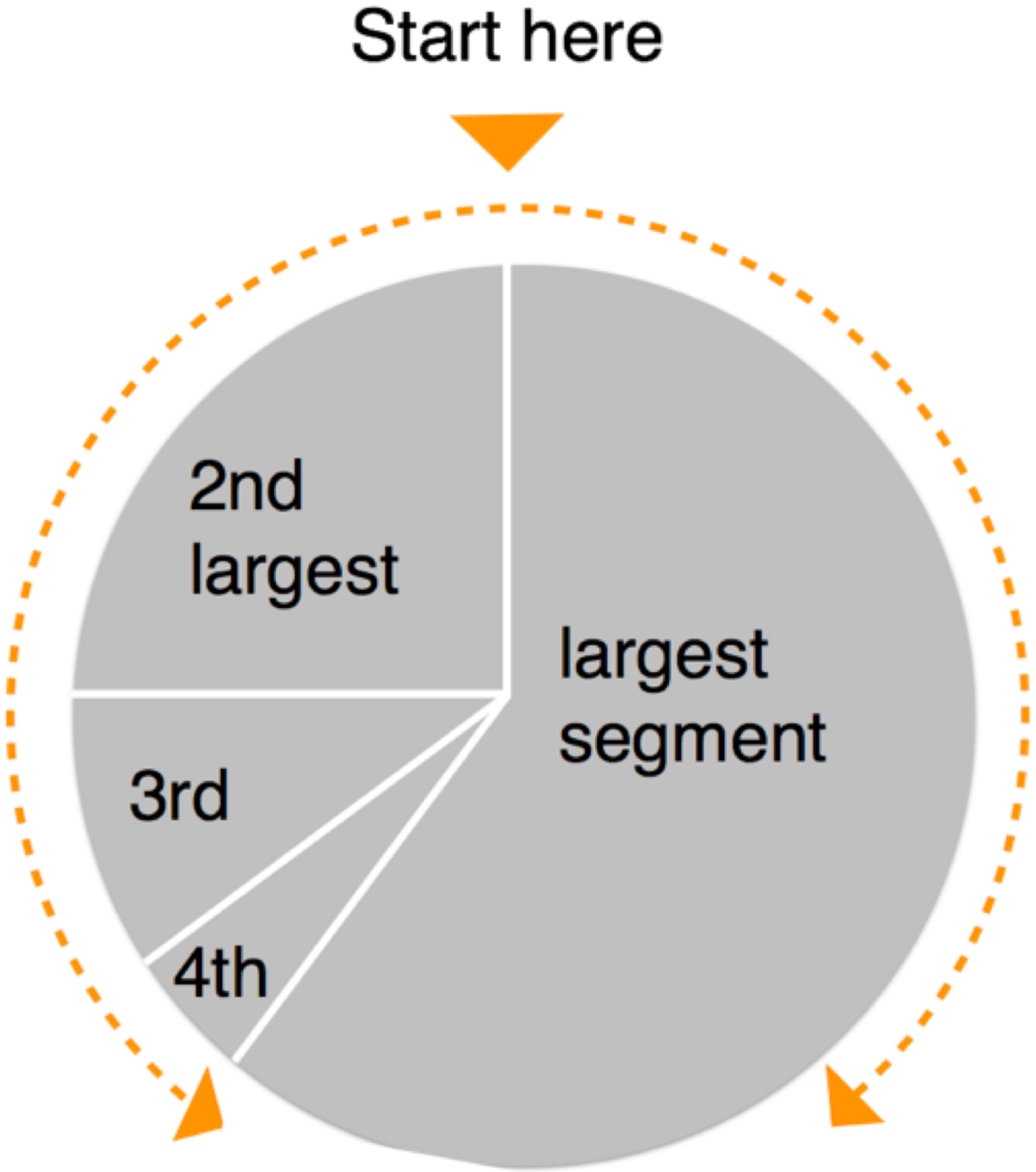
# WHAT 3-D PIE CHARTS ARE GOOD FOR.


http://flowingdata.com/2012/06/15/what-3-d-pie-charts-are-good-for/






http://wonkette.com/412361/all-193-of-republicans-support-palin-romney-and-huckabee


Start here
2nd largest
largest segment
3rd
4th

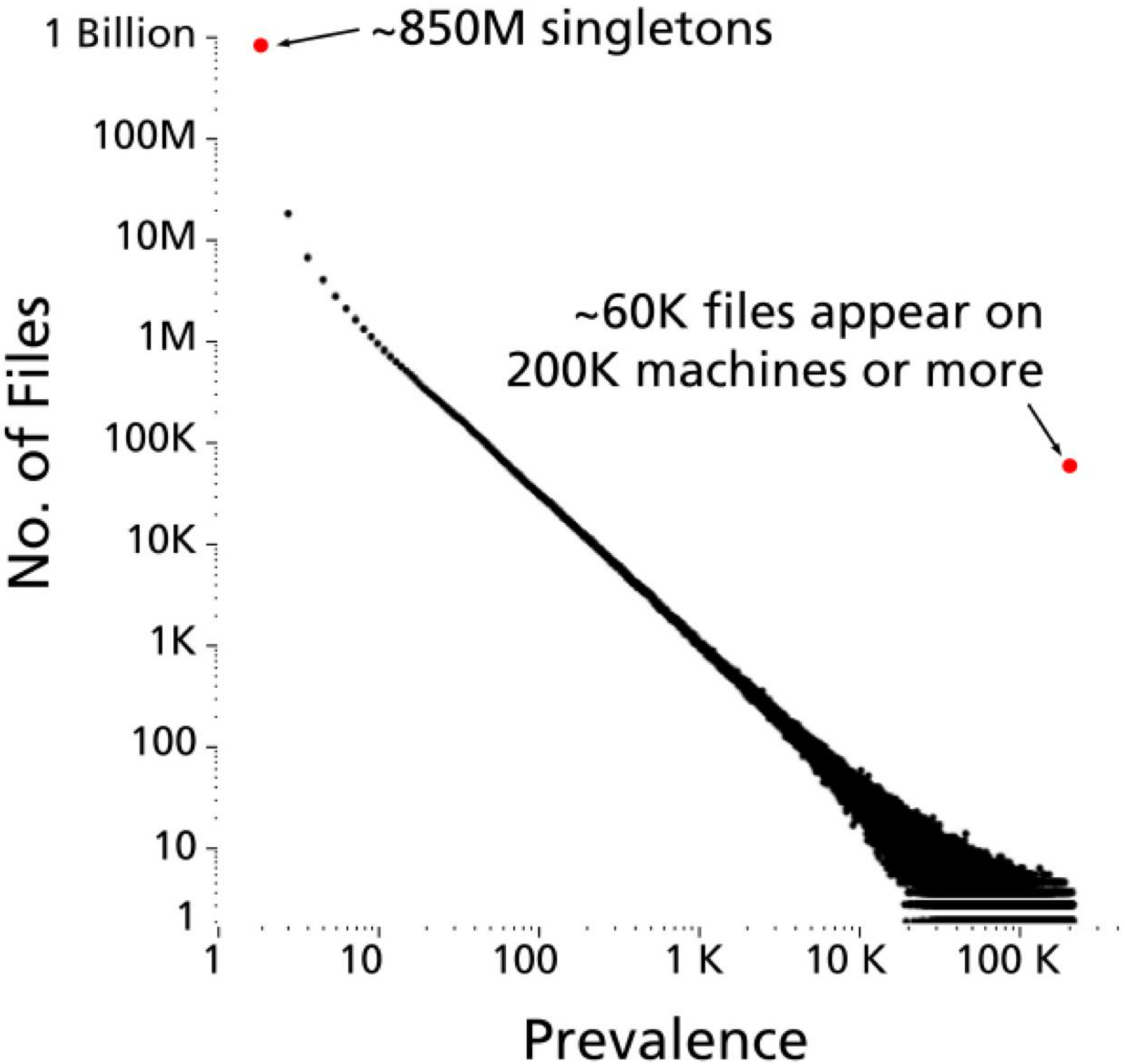

http://infosthetics.com/archives/2008/09/funniest_pie_chart_ever.html

# **Log scale** instead of linear scale

Include numbers from different orders of magnitude

# Example

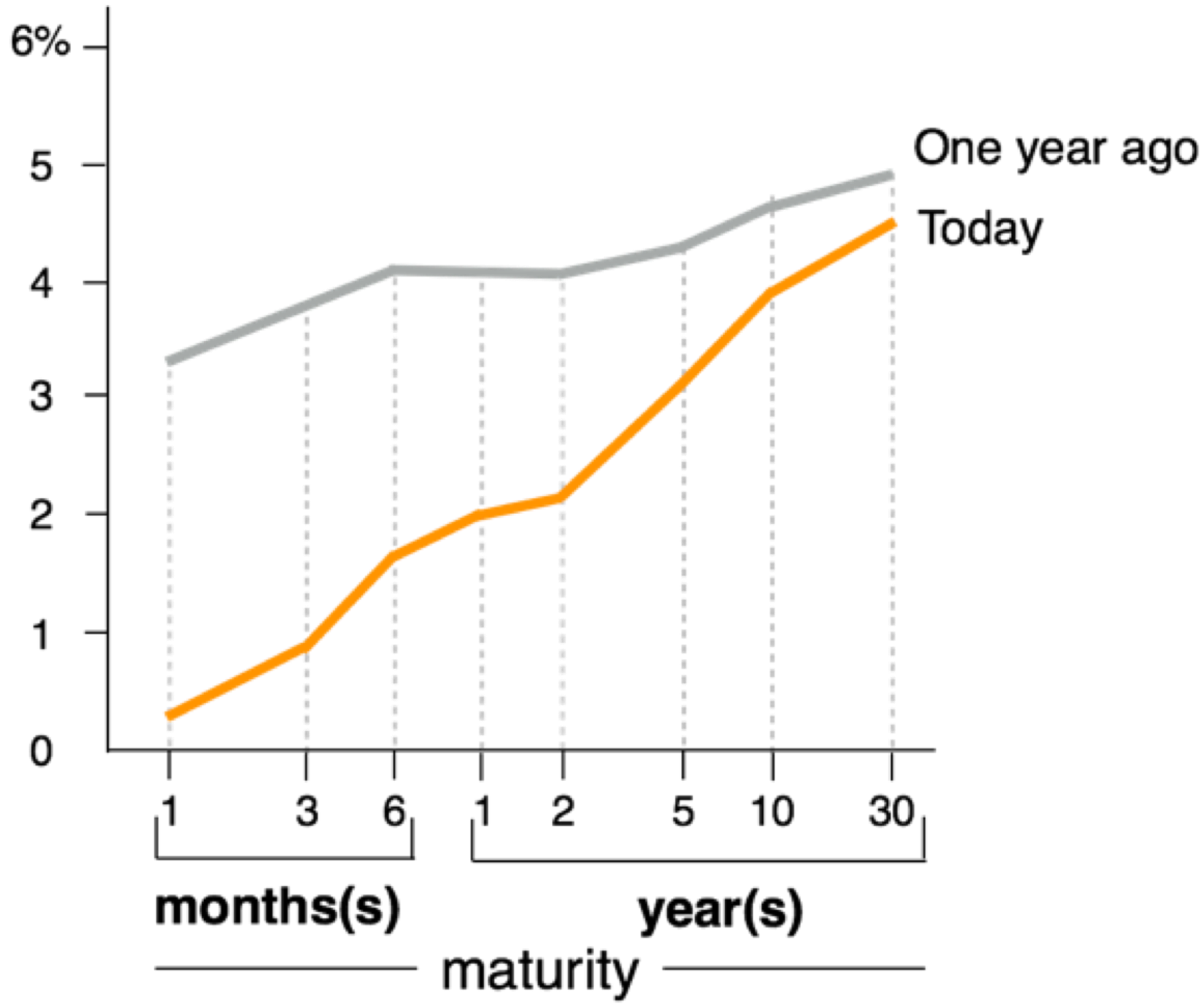
log-log



# “log” also works well for time

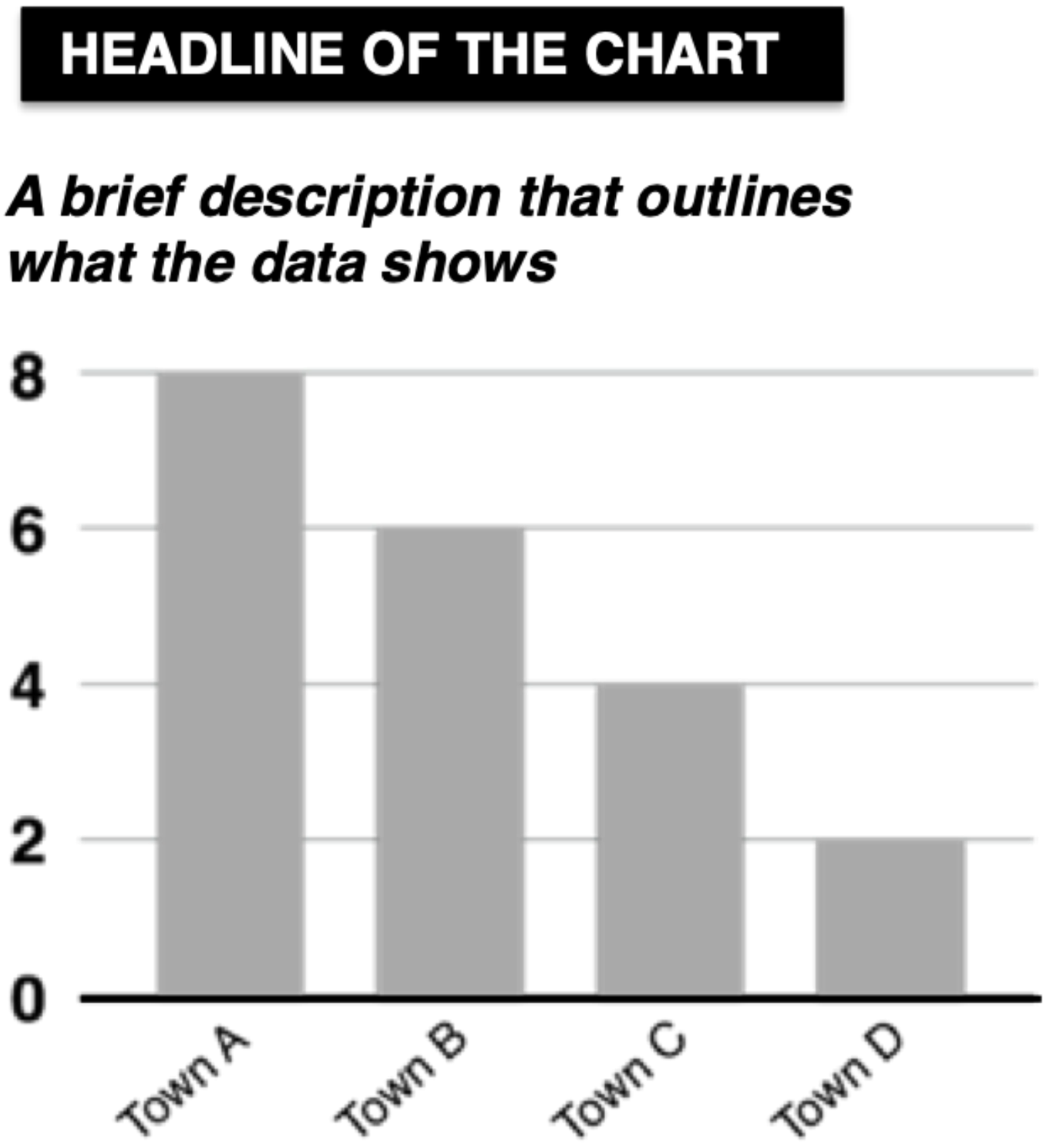
Example

The yield curve of Treasury bills, notes and bonds



# In-class Exercise.

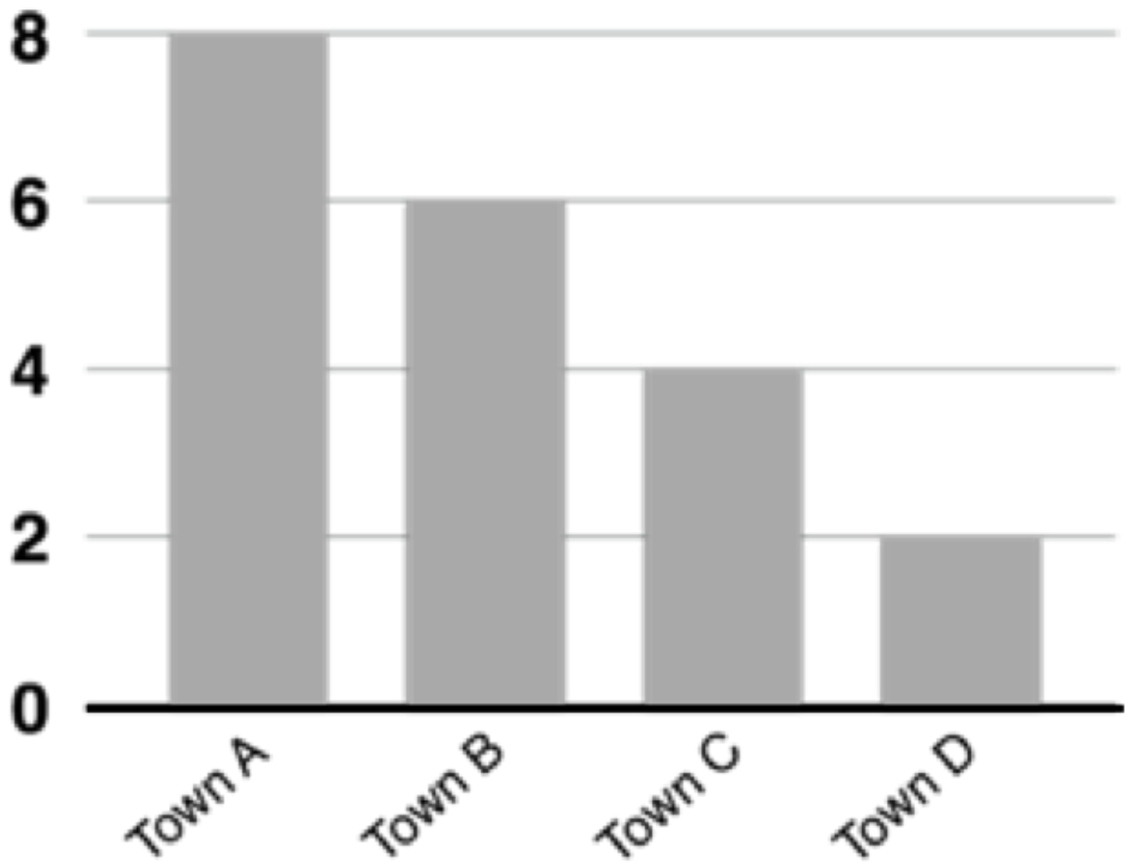
Applying what you have just learned.




Can you improve its visual design?

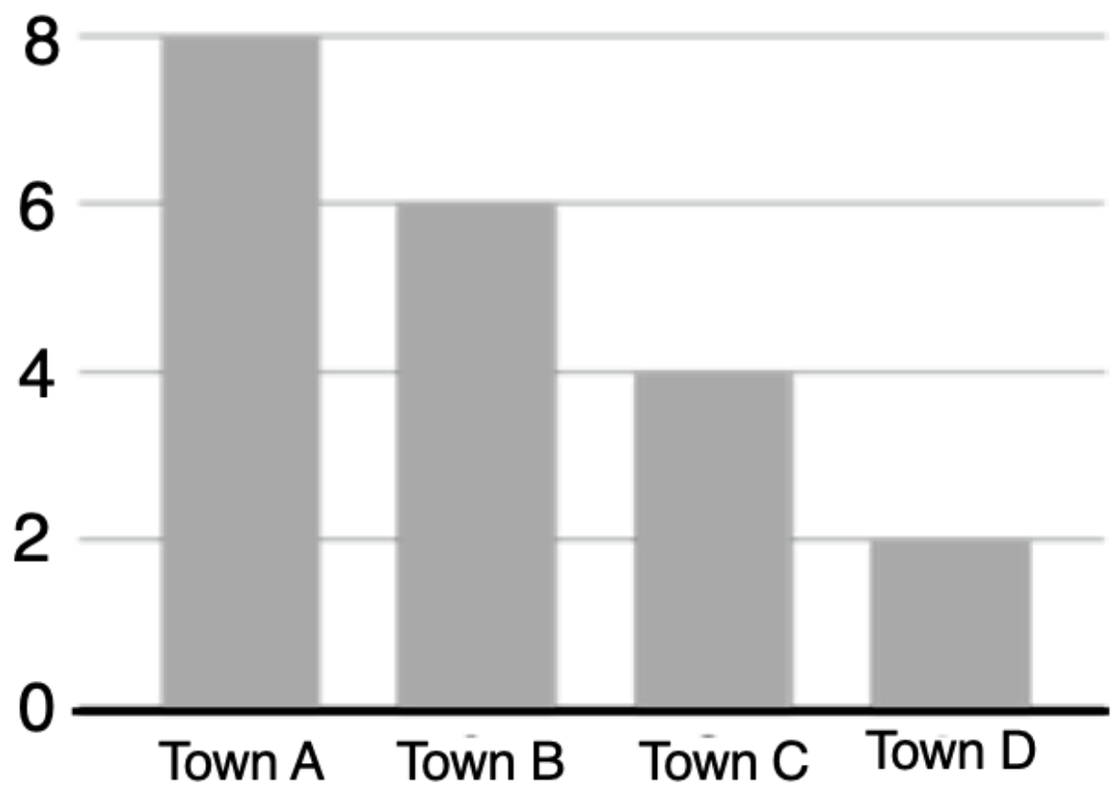
**HEADLINE OF THE CHART**

***A brief description that outlines what the data shows***




**Headline of the chart**

A brief description that outlines what the data shows




# Which is better?

# How to fix the defaults

http://www.darkhorseanalytics.com/blog/clear-off-the-table

| Role | Name | Year of the... | Debut | Number of Fans | Takedown Rate |
|---|---|---|---|---|---|
| Face (The Hero) | The Ultimate Warrior | Tiger | May-2011 | 97320.00 | 86.2 |
| Face (The Hero) | Hulk Hogan | Oxen | Jan-2008 | 988551.00 | 61.978 |
| Face (The Hero) | Macho Man Randy Savage | Monkey | Feb-2008 | 157618.00 | 59.29 |
| Face (The Hero) | Hacksaw Jim Duggan | Pig | Mar-2008 | 30300.00 | 53.4332 |
| Face (The Hero) | Superfly Jimmy Snuka | Dragon | Mar-2008 | 12341.00 | 52.7 |
| Heel (The Bad Guy) | Rowdy Roddy Piper | Rooster | Jun-1968 | 71645.00 | 45.4 |
| Heel (The Bad Guy) | The Million Dollar Man Ted DiBiase | Rat | Apr-1975 | 449342.00 | 43.7689 |
| Heel (The Bad Guy) | Mr. Perfect Curt Henning | Rat | May-1980 | 13773.00 | 38 |
| Heel (The Bad Guy) | Jake the Snake Roberts | Snake | Jul-1975 | 5609.00 | 37.99 |
| Jobber (The Unknown) | Brad Smith | Sheep | Aug-2008 | 1103.00 | 36.316 |
| Jobber (The Unknown) | Ted Duncan | Sheep | Aug-2008 | 200.00 | 33.61 |
| Jobber (The Unknown) | Joey the Uber Nerd Cherdarchuk | Snake | Aug-2008 | 5.00 | 21.0196 |

# How to fix the defaults

http://www.darkhorseanalytics.com/blog/clear-off-the-table

| Role | Name | Year of the... | Debut | Thousands of Fans | Takedown Rate |
|---|---|---|---|---|---|
| Face (The Hero) | The Ultimate Warrior | Tiger | May-2011 | 97.3 | 86.2 |
| | Hulk Hogan | Oxen | Jan-2008 | 988.6 | 62.0 |
| | **Macho Man Randy Savage** | **Monkey** | **Feb-2008** | **157.6** | **59.3** |
| | Hacksaw Jim Duggan | Pig | Mar-2008 | 30.3 | 53.4 |
| | Superfly Jimmy Snuka | Dragon | Mar-2008 | 12.3 | 52.7 |
| Heel (The Bad Guy) | Rowdy Roddy Piper | Rooster | Jun-1968 | 71.6 | 45.4 |
| | The Million Dollar Man Ted DiBiase | Rat | Apr-1975 | 449.3 | 43.8 |
| | Mr. Perfect Curt Henning | Rat | May-1980 | 13.8 | 38.0 |
| | Jake the Snake Roberts | Snake | Jul-1975 | 5.6 | 38.0 |
| Jobber (The Unknown) | Brad Smith | Sheep | Aug-2008 | 1.1 | 36.3 |
| | Ted Duncan | Sheep | Aug-2008 | 0.2 | 33.6 |
| | Joey the Uber Nerd Cherdarchuk | Snake | Aug-2008 | 0.0 | 21.0 |

# Practitioners' Guide

**Colors**: start with black & white, then add colors, carefully

Forces you to focus on content and layout

**Fonts**: **sans-serif** generally easier to read

(On Mac: Helvetica is great start)

**Animation**: start with **no** animation, then add meaningful ones

# Practitioners' Guide:
# **Use Pictures and Videos**

"Pictures" include tables, diagrams, charts, etc.

- Pictures often more succinct & memorable
- People like pictures and love movies

**And show them ASAP!**

Once people fall asleep, it's hard to wake them up!
If you have good stuff, show them now.

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

Example

# GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation

Minsuk Kahng, Nikhil Thorat, Duen Horng (Polo) Chau, Fernanda B. Viégas, and Martin Wattenberg

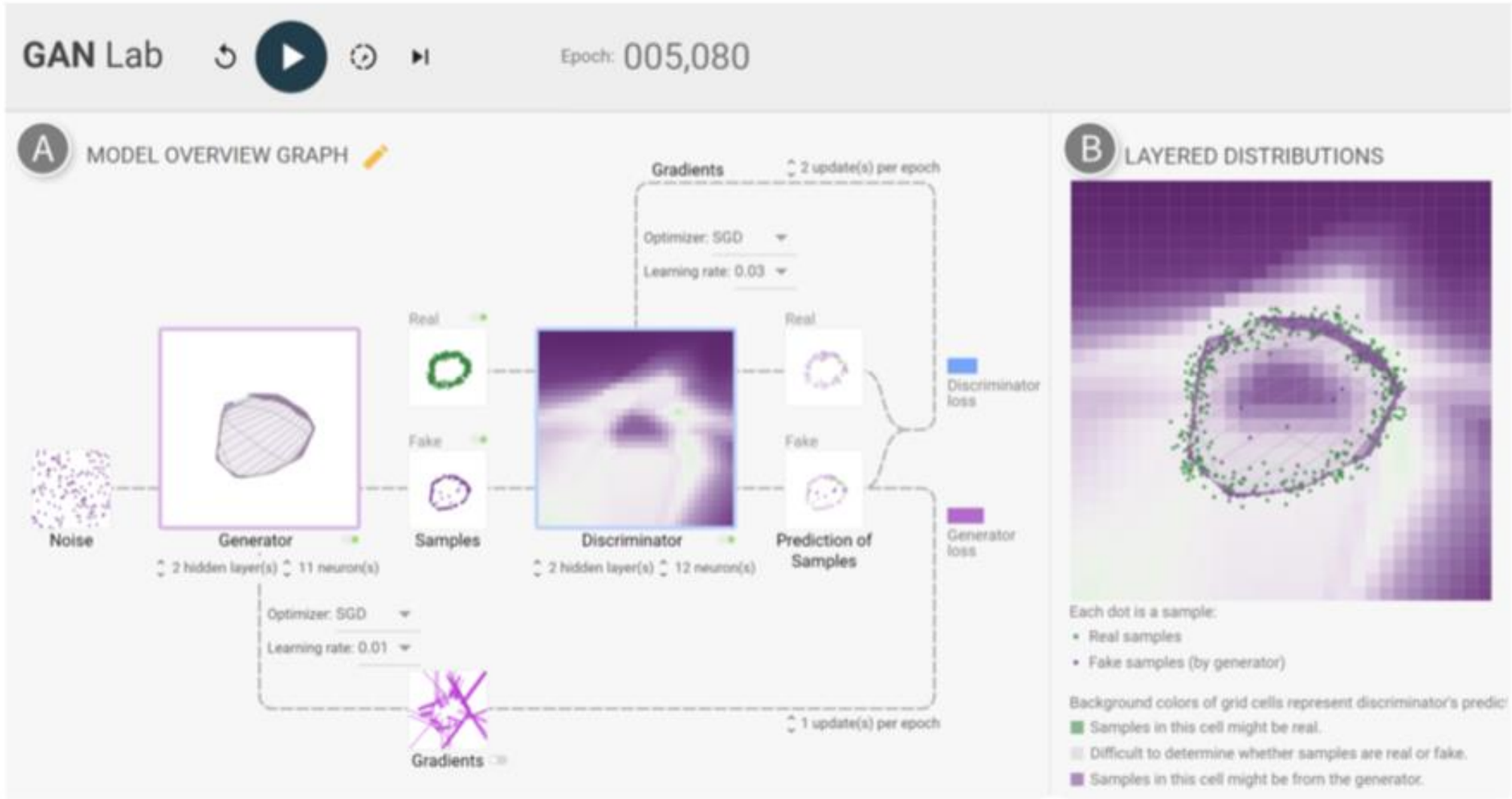



Fig. 1. With GAN Lab, users can interactively train Generative Adversarial Networks (GANs), and visually examine the model training process. In this example, a user has successfully used GAN Lab to train a GAN that generates 2D data points whose challenging distribution resembles a ring. **A.** The *model overview graph* summarizes a GAN model's structure as a graph, with nodes representing the generator and discriminator submodels, and the data that flow through the graph (e.g., fake samples produced by the generator). **B.** The *layered distributions* view helps users interpret the interplay between submodels through user-selected layers, such as the discriminator's classification heatmap, real samples, and fake samples produced by the generator.

**Abstract**—Recent success in deep learning has generated immense interest among practitioners and students, inspiring many to learn about this new technology. While visual and interactive approaches have been successfully developed to help people more easily

Example

# SHIELD: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression

Nilaksh Das[1], Madhuri Shanbhogue[1], Shang-Tse Chen[1], Fred Hohman[1], Siwei Li[1], Li Chen[2], Michael E. Kounavis[2], Polo Chau[1]
[1]Georgia Institute of Technology, Atlanta, GA, USA
{nilakshdas,madhuri.shanbhogue,schen351,fredhohman,robertsiweili,polo}@gatech.edu
[2]Intel Corporation, Hillsboro, OR, USA
{li.chen,michael.e.kounavis}@intel.com

## ABSTRACT

The rapidly growing body of research in adversarial machine learning has demonstrated that deep neural networks (DNNs) are highly vulnerable to adversarially generated images. This underscores the urgent need for practical defense techniques that can be readily deployed to combat attacks in real-time. Observing that many attack strategies aim to perturb image pixels in ways that are visually imperceptible, we place JPEG compression at the core of our proposed SHIELD defense framework, utilizing its capability to effectively "compress away" such pixel manipulation. To immunize a DNN model from artifacts introduced by compression, SHIELD "vaccinates" the model by retraining it with compressed images, where different compression levels are applied to generate multiple vaccinated models that are ultimately used together in an ensemble defense. On top of that, SHIELD adds an additional layer of protection by employing randomization at test time that compresses different regions of an image using random compression levels, making it harder for an adversary to estimate the transformation performed. This novel combination of vaccination, ensembling, and randomization makes SHIELD a fortified multi-pronged defense. We conducted extensive, large-scale experiments using the ImageNet dataset, and show that our approaches eliminate up to 94% of black-box attacks

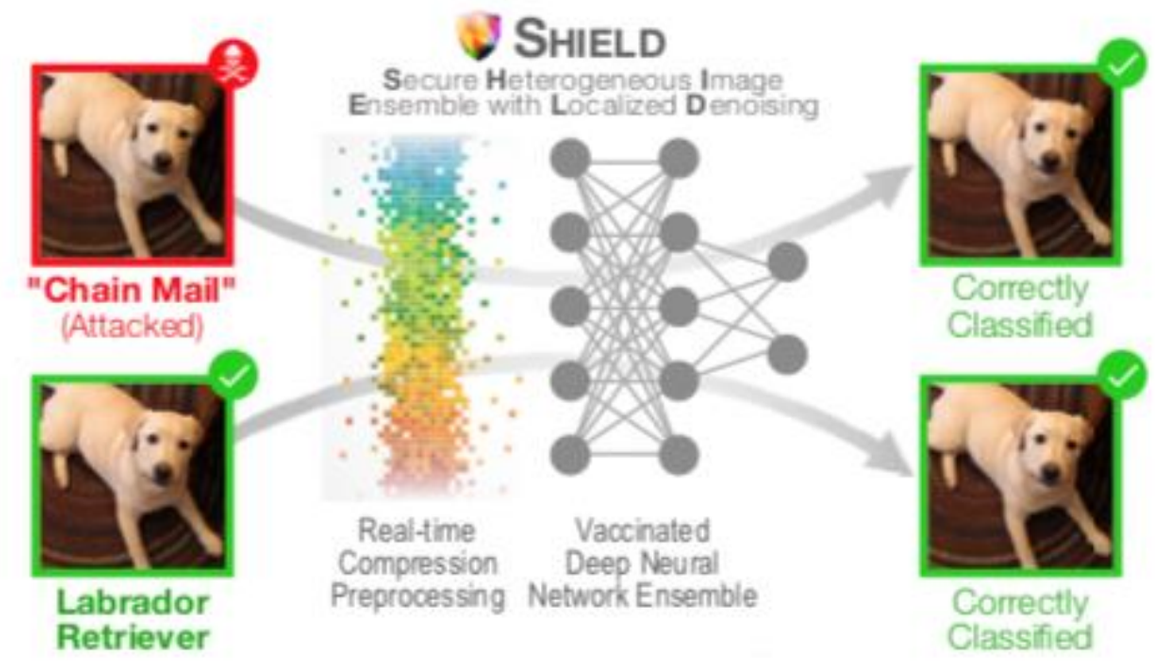



**Figure 1: SHIELD Framework Overview. SHIELD combats adversarial images (in red), by removing perturbation in real time using Stochastic Local Quantization (SLQ) and an ensemble of vaccinated models robust to compression transformation for both adversarial and benign images. Our approach eliminates up to 94% of black-box attacks and 98% of gray-box attacks delivered by some of the most recent, strongest attacks, such as *Carlini-Wagner's L2* and *DeepFool*.**

# Practitioners' Guide:
# **Additional Tips for Researchers**

Crown-jewel pictures are important

- Overview of what readers is going to get — **cut to the chase (don't tease!)**
- People skim and look at "interesting" things first
- Reviewers are busy and sleepy 😴 (read 5-10 papers per conference) — it's refreshing to read an interesting paper

How to do it?

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

# Figures should be self-contained

Why?

- Don’t make people go back and forth between text & figure
- Bad figures means **bad first impression** (reject!)

How to fix?

- Succinctly describe your main (take-away) messages

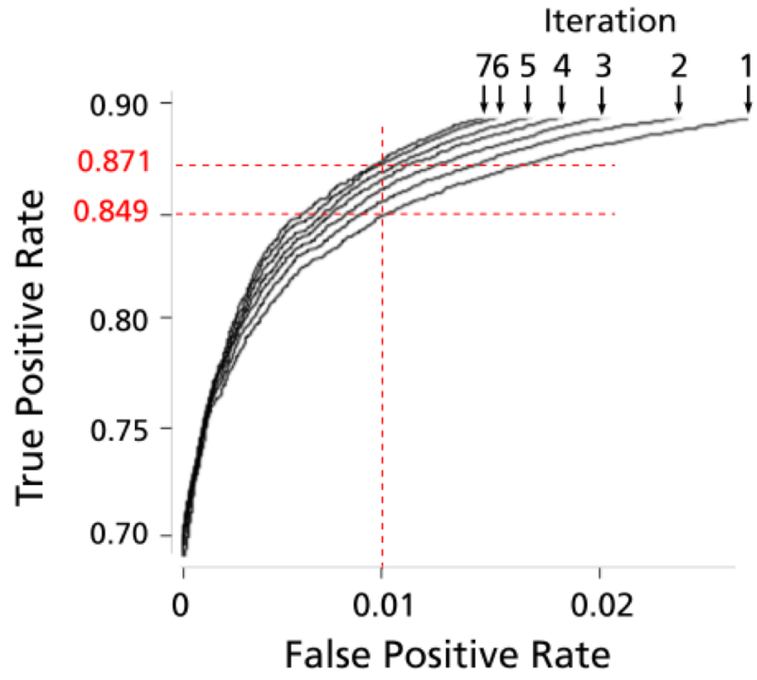



Figure 8: ROC curves of 7 iterations; true positive rate incrementally improves.

Example

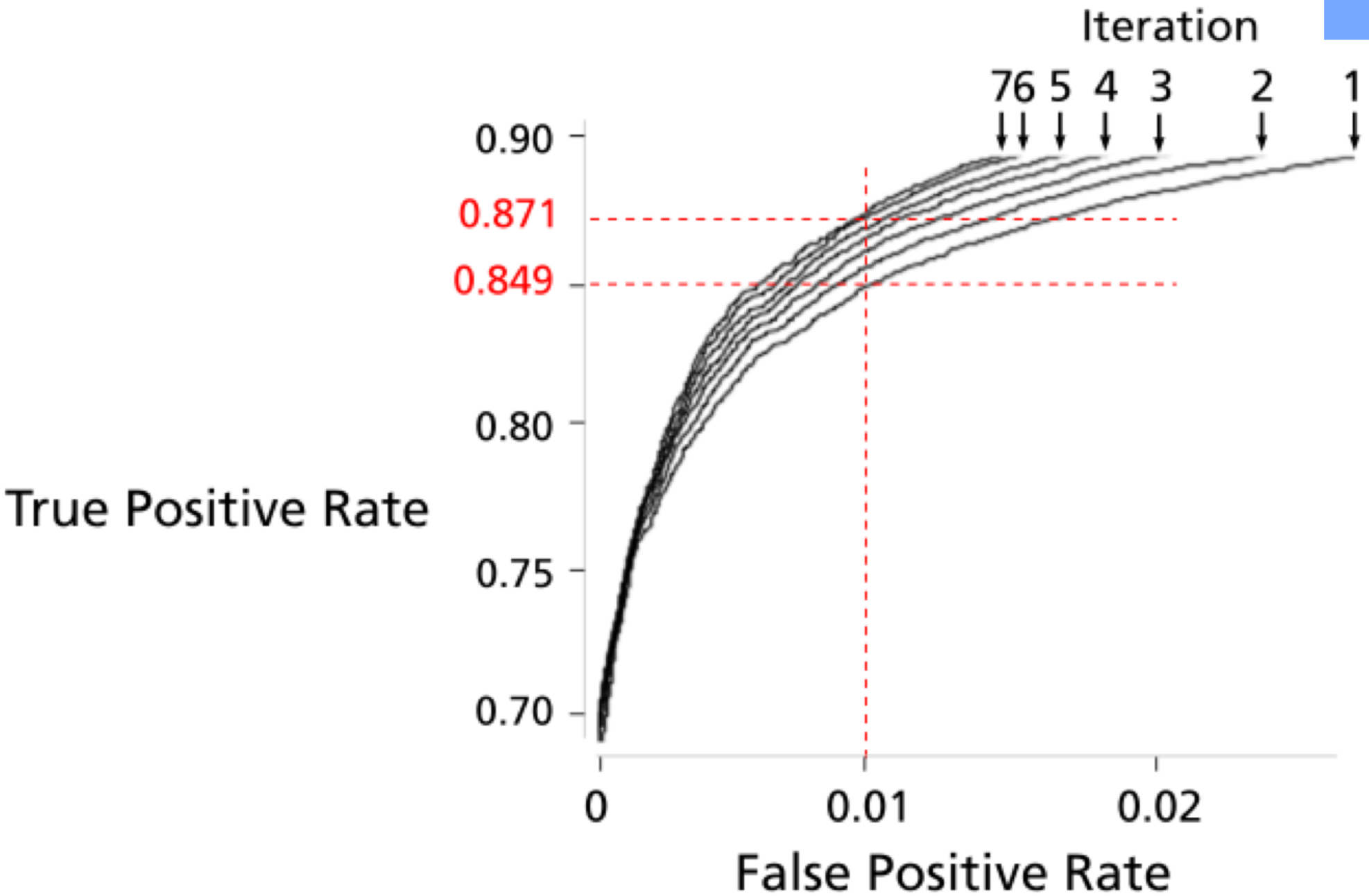



Figure 8: ROC curves of 7 iterations; true positive rate incrementally improves.

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

# More generally, how to write **“good”** papers?

## Heuristics for Scientific Writing (a Machine Learning Perspective)

It’s January 28th and I should be working on my paper submissions. So should you! But why write when we can **meta-write**? ICML deadlines loom only twelve days away. And KDD follows shortly after. The schedule hardly lets up there, with ACL, COLT, ECML, UAI, and NIPS all approaching before the summer break. Thousands of papers will be submitted to each.

The tremendous surge of interest in machine learning along with ML’s democratization due to open source software, YouTube coursework, and the availability of preprint articles are all exciting happenings. But every rose has a thorn. Of the thousands of papers that hit the arXiv in the coming month, many will be unreadable. Poor writing will damn some to rejection while others will fail to reach their potential impact. Even among accepted and influential papers, careless writing will sow confusion and damn some papers to later criticism for sloppy scholarship (you better hope Ali Rahimi and Ben Recht don’t win another test of time award!).

But wait, there’s hope! Your technical writing doesn’t have to stink. Over the course of my academic career, I’ve formed strong opinions about how to write a paper (as with all opinions, you may disagree). While one-liners can be trite, I learned early in my PhD from Charles Elkan that many important heuristics for scientific paper writing can be summed up in

http://approximatelycorrect.com/2018/01/29/heuristics-technical-scientific-writing-machine-learning-perspective/

## Catchy Titles Are Good: But Avoid Being Cute

**Jacob O. Wobbrock**
The Information School | DUB Group
University of Washington
Seattle, WA USA 98195
wobbrock@uw.edu

**ABSTRACT**
The most important rule of Abstracts is that they describe the work, not the paper. Include, at most, one sentence of motivation. Save the rest of your motivation for the Introduction. Effective Abstracts focus on two things: (1) Describing what was *done*. (2) Describing what was *found* (key results). Be specific about your key findings. Instead of “many” say “84%”. Keep the Abstract to one paragraph and fewer than 200 words.

**Author Keywords**
Authors’ choice; of terms; separated; by semicolons; commas, within terms only; this section is required.

**ACM Classification Keywords**
Example: H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

See http://acm.org/about/class/1998 for the full list of ACM classifiers. This section is required.

**INTRODUCTION**
The Introduction delivers the motivation for your paper. It explains *why* you did the work you did. This is the primary function of the Introduction.

I have found a 5-point structure for Introductions to be particularly effective. (Here, I build on advice I received as a Ph.D. student from Prof. Scott E. Hudson.)

- State of the world…
- The big BUT…
- Therefore, we did…
- The key findings are…
- The contributions of this work are…

The state of the world is a description of issues in whatever “world” is relevant to your topic. Drawing on popular press can be effective here if recent news items or data from


**Figure 1.** Choose a telling figure for your paper that is placed at the top of the right-hand column on the first page. I like to make figure caption text Arial 8 pt. so that it stands apart from the body text but is small. I place my figures in-line as single “characters” in their own paragraphs, and captions as their own separate paragraphs, rather than placing either in floating text boxes, which jump around. I give captions 12 pt. after-paragraph line spacing. This figure is from [13].

One of the worst ways to motivate your work is using the “absence from the literature” argument. “Studies to date have not …,” or “The literature is thus far silent on…,” or “Researchers have not yet examined how…” Such sentences are fine to add *after* you have established a problem or opportunity worthy of pursuit in its own right. But as stand-alone motivational statements, absence-from-the-literature does not zing. Maybe the literature is silent on an issue because that issue is not important.

After the big BUT, you will describe what you did, now in a bit more detail than in the Abstract. Devoting a paragraph to

http://faculty.washington.edu/wobbrock/pubs/Wobbrock-2015.pdf

# Use legible fonts.

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

For printed materials, print them out and check!

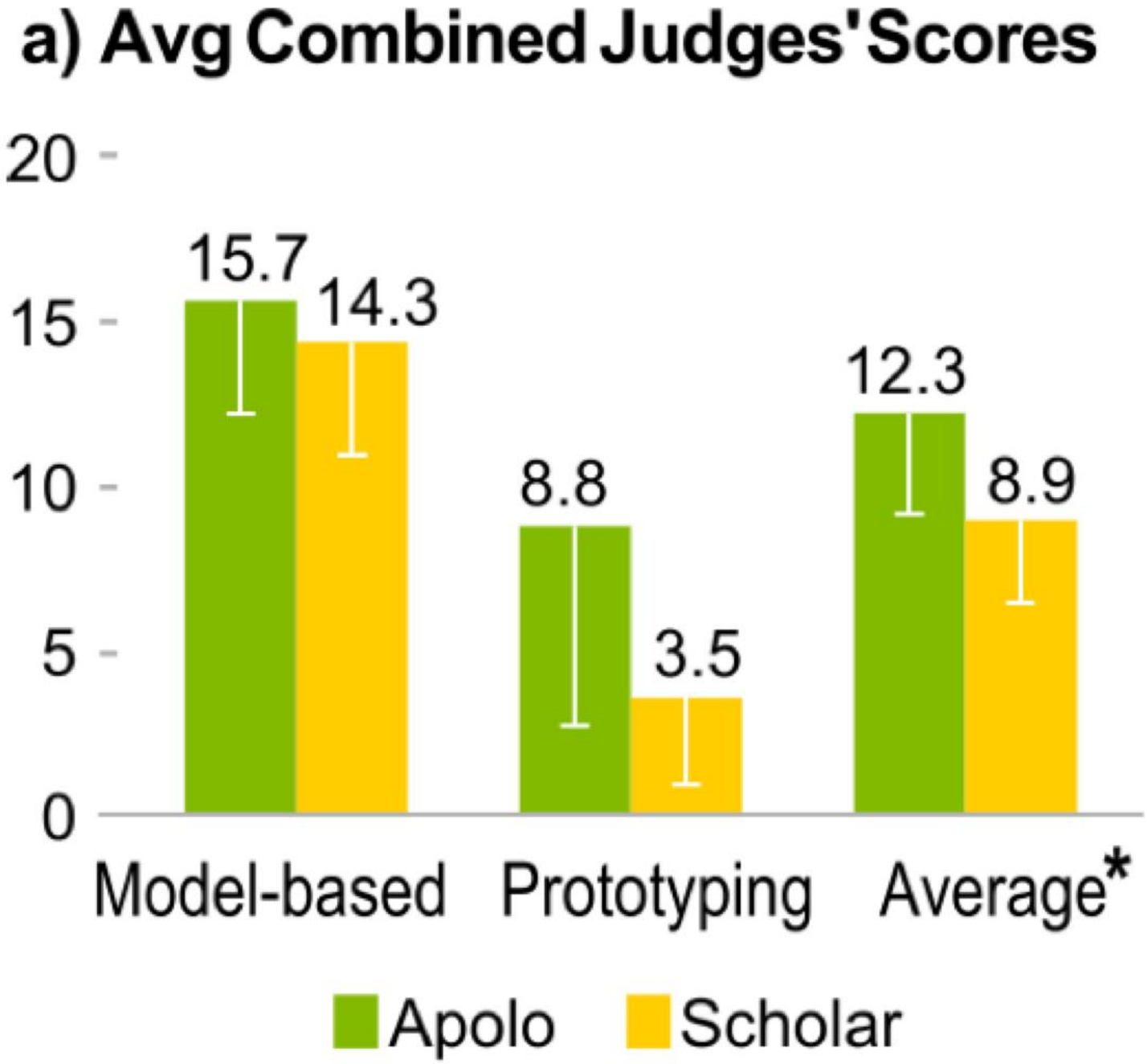
Rule of thumb: about 7 lines of text on a slide.

# Redesign figures for presentation

Designing for print is different from designing for the screen

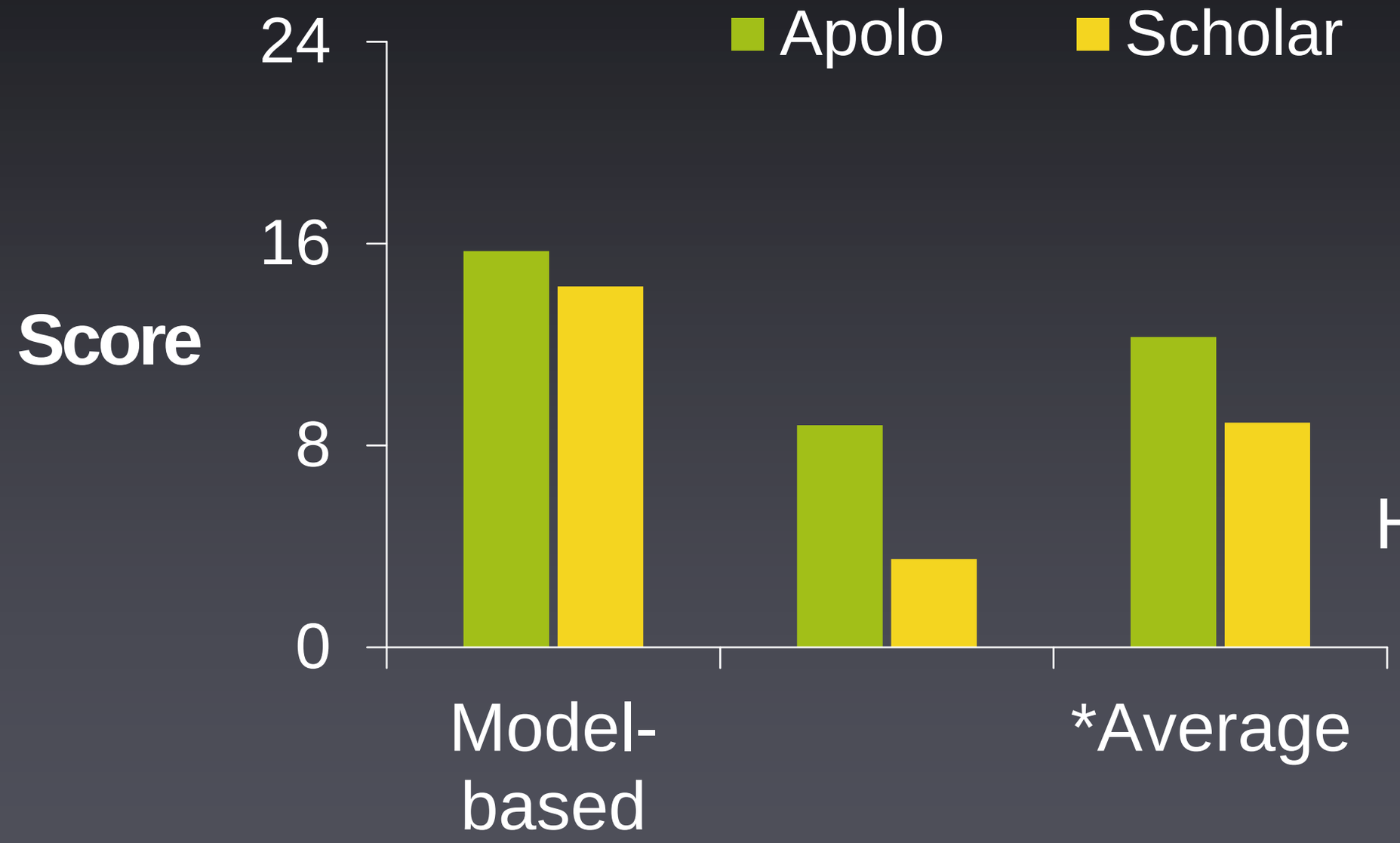
- Resolution (which is higher?)
- Levels of details (people mostly want a few “take-away” messages from your talk)

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




Higher is better.
**Apolo** wins.

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

# **Great Work** destroyed by **Poor Presentation**

Bad color schemes

Bad, tiny 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=IpvgfmEU2Ck&feature=player_embedded