ISA 401: Business Intelligence & Data Visualization

15: Fundamentals of Data Visualization

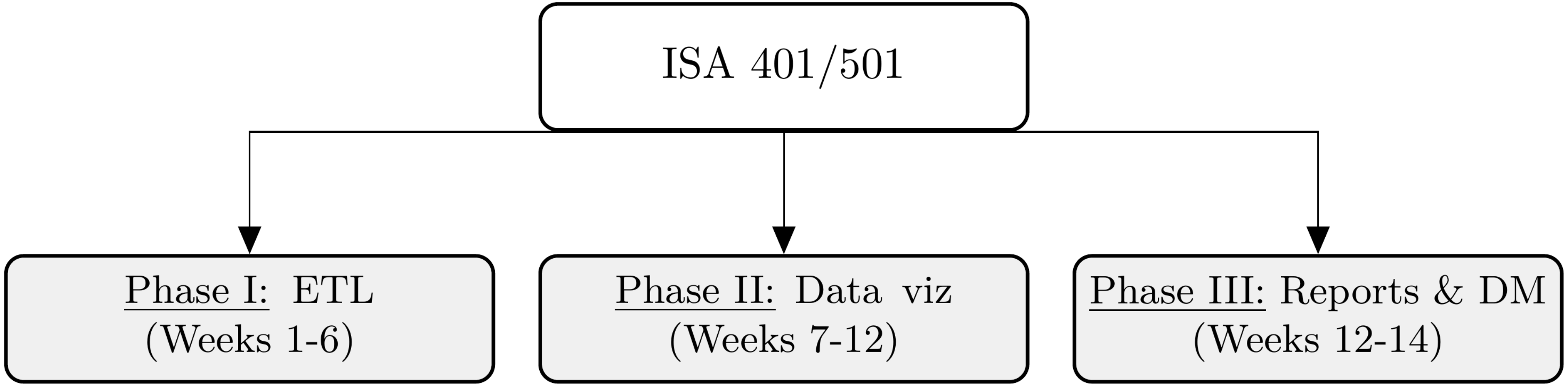
Fadel M. Megahed, PhD

Raymond E. Glos Professor in Business
Farmer School of Business
Miami University

@FadelMegahed
fmegahed
fmegahed@miamioh.edu
Automated Scheduler for Office Hours

Fall 2025

Refresher: Organization of this Course



How the ISA 401/501 course is organized.

Learning Objectives for Today's Class

- Explain the concept of "graphical excellence"
- Explain the theory of data graphics
- Optimize visual encoding based on data types
- Understand why color should be used sparingly and how to select appropriate colors (when color is a must)

Graphical Excellence

Non-graded activity: Terrible Charts

05:00

Activity | Russia's Defense Budget | White House Economy Growth | Tucker Carlson

> Over the next 5 minutes, please identify the **1-2 main problems** in the charts in the following tabs.

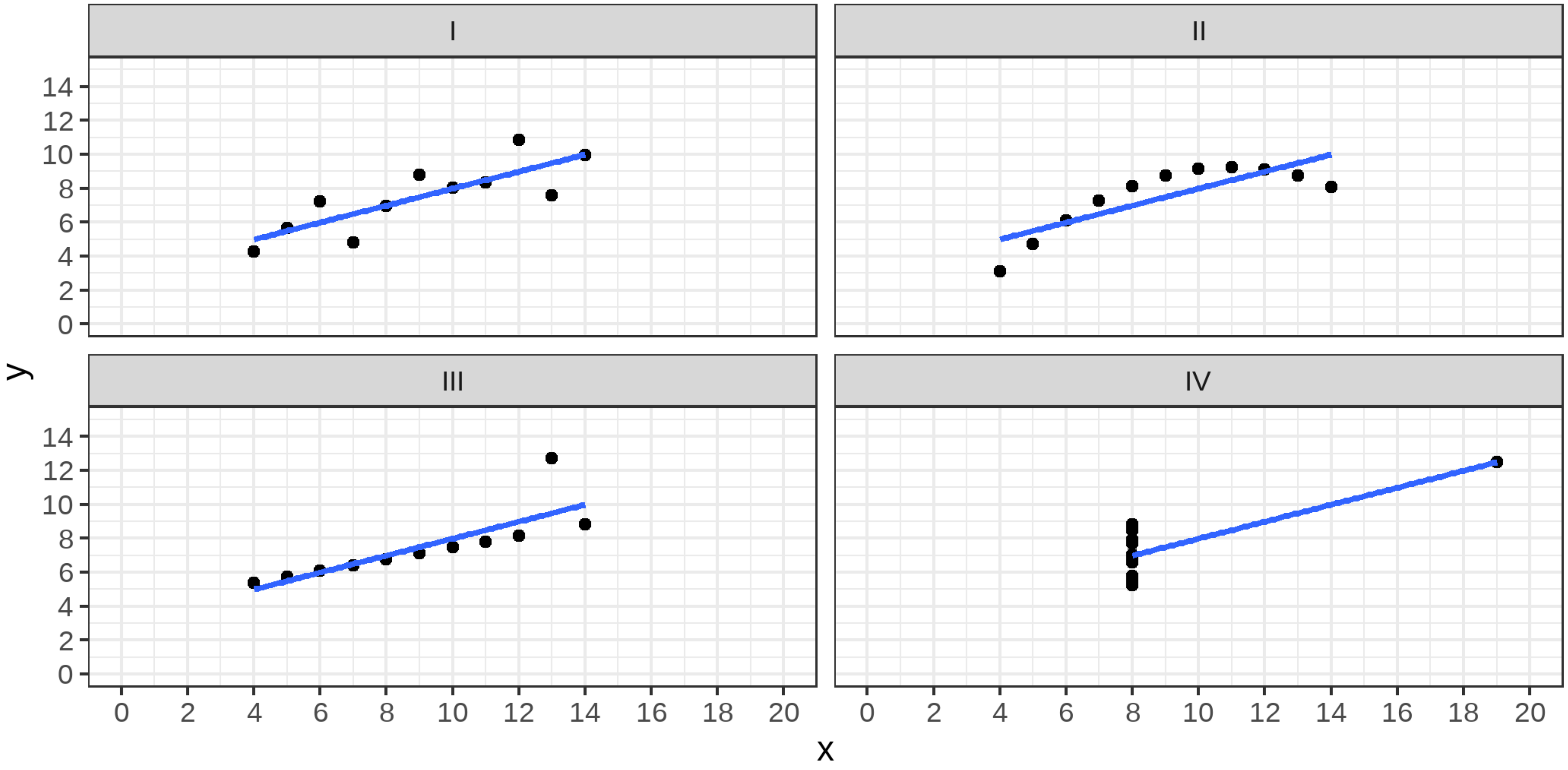
- Write down your answers in the editable area of each chart.
- Discuss your answers with your neighboring classmates.
- Be prepared to share these answers with class.

Graphical Excellence: What should Graphs Do?

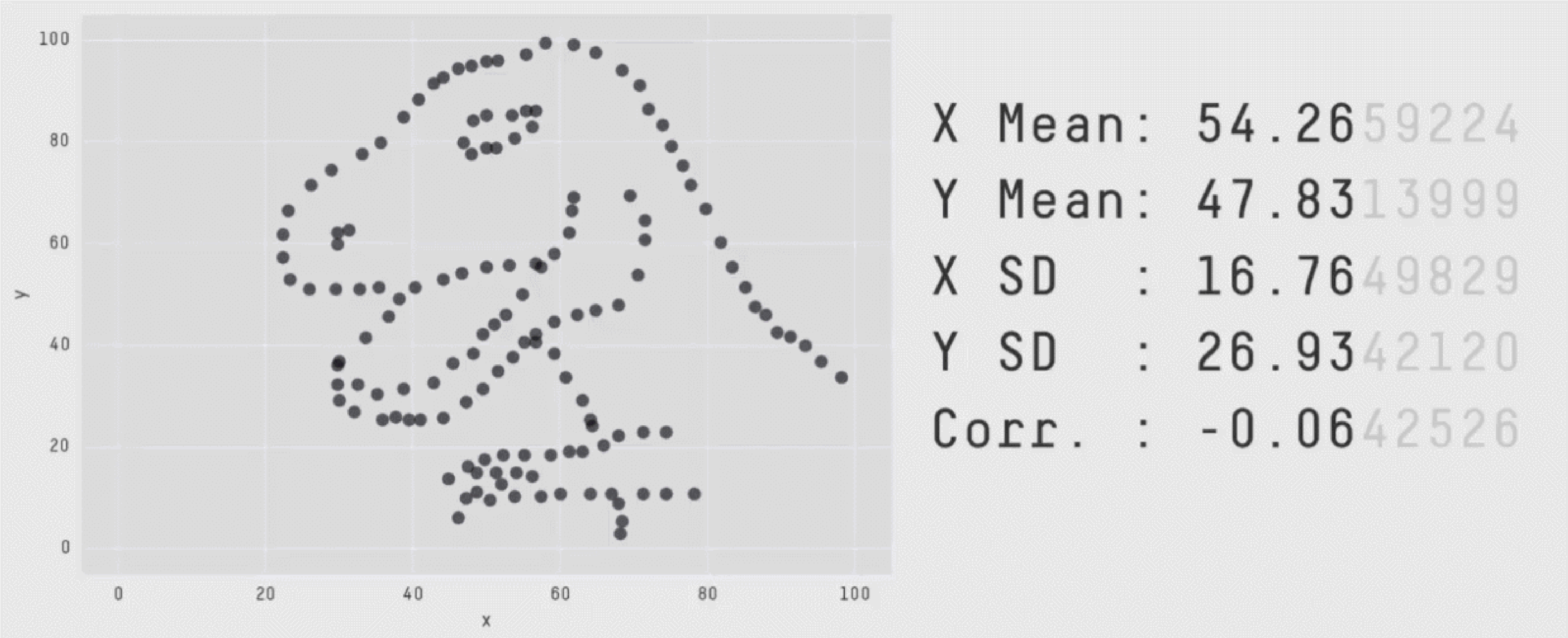
- **Show the data**
- Lead to thinking about the **substance** rather than something else
- Avoid **distorting** what the data have to say
- Present **many numbers in a small space**
- Make **large datasets coherent**
- Encourage the eye to **compare different pieces of the data**
- **Reveal the data at several levels of detail**, from a broad overview to the fine structure
- Serve **a purpose:** description, exploration, tabulation, decoration
- Be **closely integrated with the statistical & verbal descriptions of the data**

Source: Tufte, E. R. (2001). The visual display of quantitative information. Cheshire, Conn: Graphics Press, P. 13.

Show/Reveal the Data: Anscombe's Dataset



A Modern Version of Anscombe's Dataset

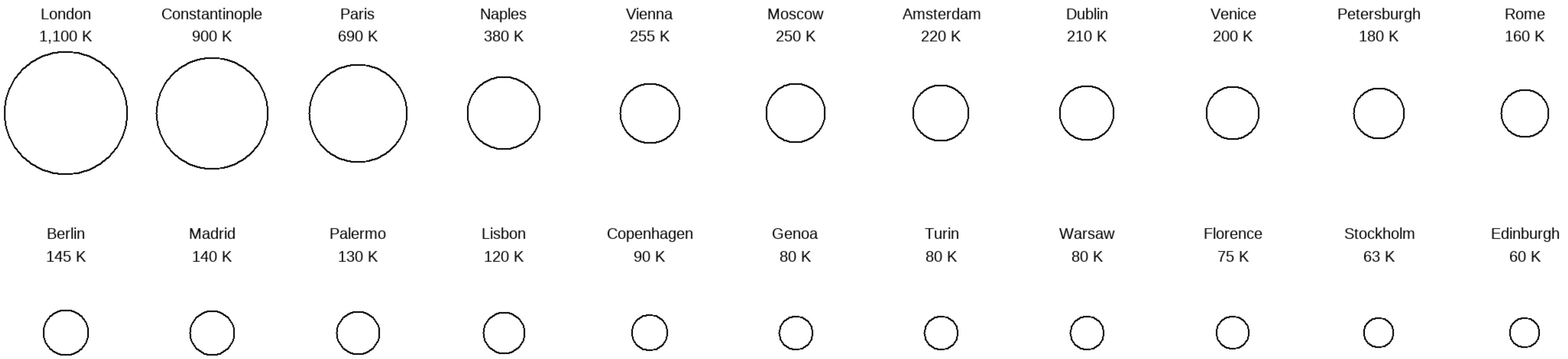


Substance

05:00

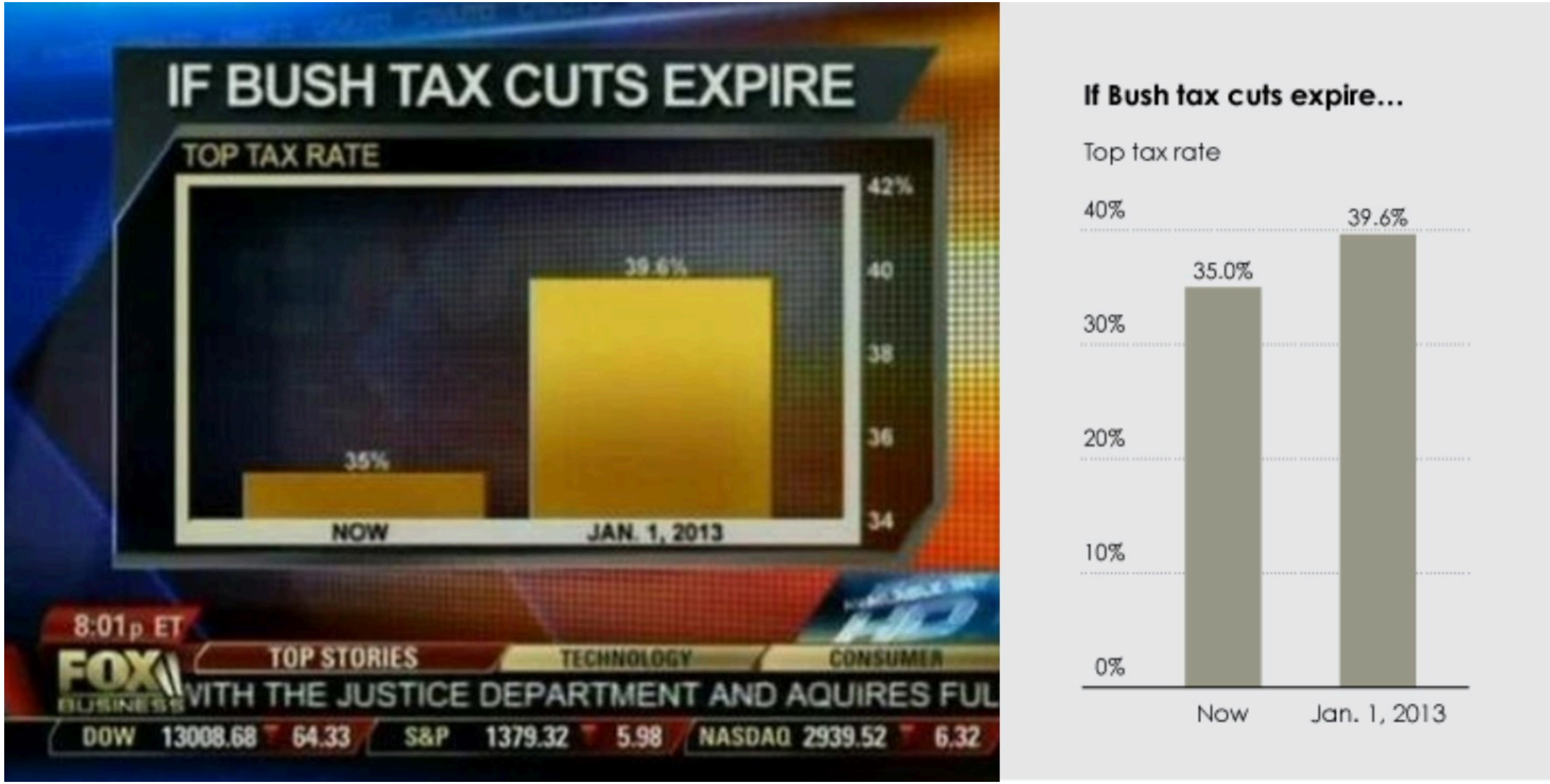
Activity | Your Solution | My Solution

> In 5 minutes, please **sketch** a better (non-bubble) chart than the one use by William Playfair for plotting the populations of 22 European cities at the end of the 1700s.



Source: Chart made by Fadel M. Megahed based on data from http://www.stat.uiowa.edu/~luke/data/Playfair

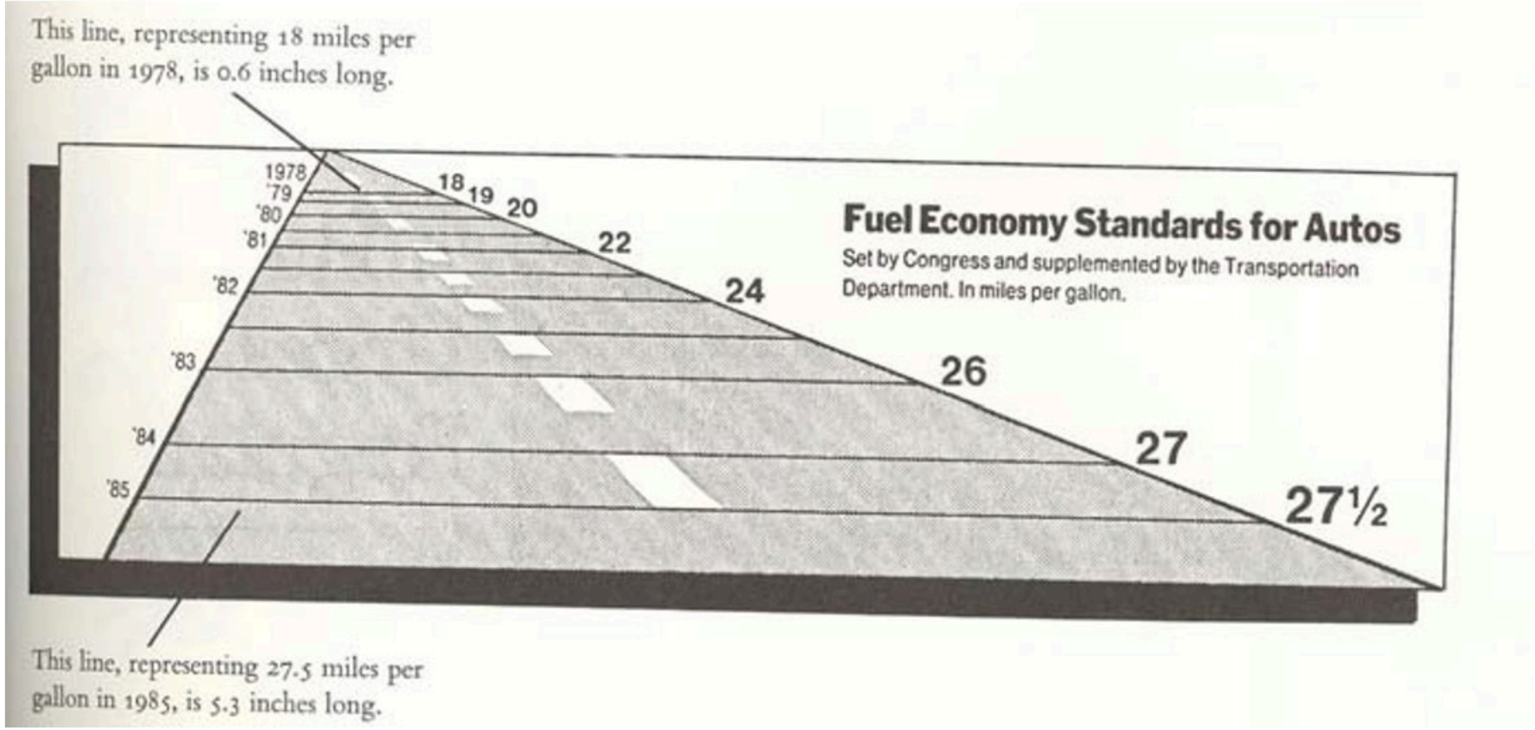
Avoid Distortion of Data



Source: Nathan Yau (2012). Fox News continues charting excellence.

Avoid Distortion of Data: The Lie Factor

$$\frac{\text{Size of effect shown in graphic}}{\text{Size of effect in data}}$$



Source: Graph from Tufte, E. R. (2001). The visual display of quantitative information. Cheshire, Conn: Graphics Press, P. 57.

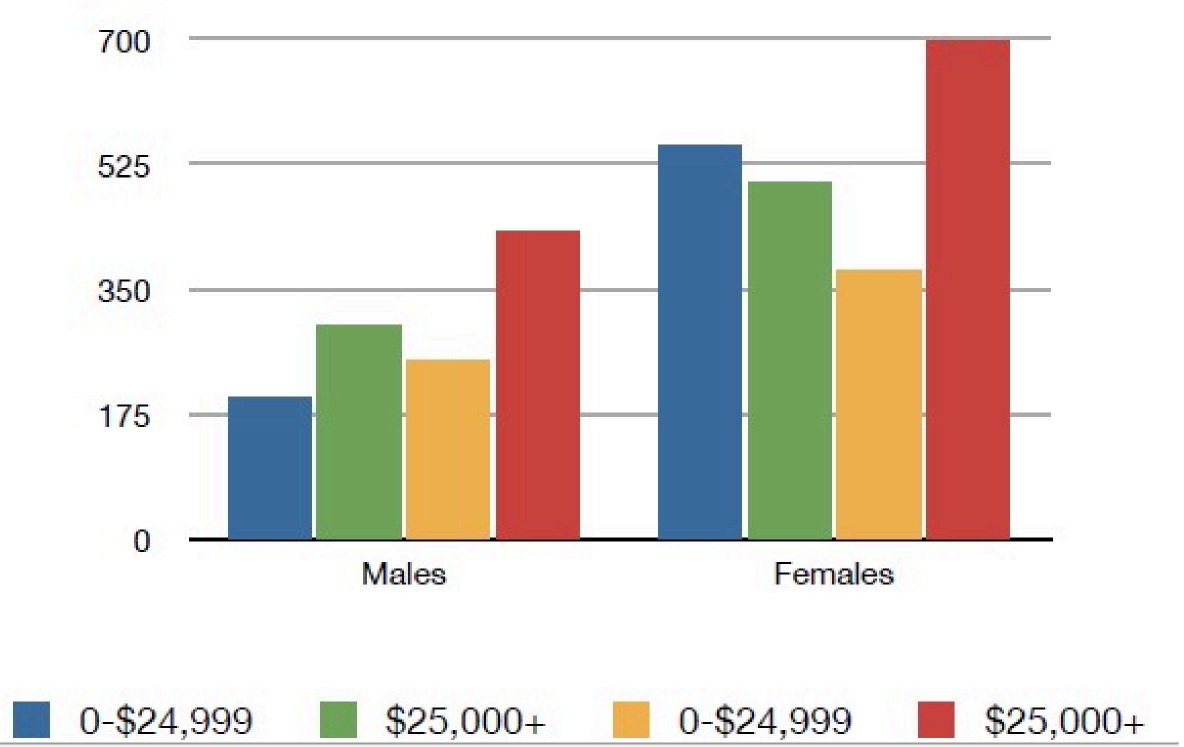
Graphical Integrity Principles: A Summary

- Clear, detailed, and thorough labeling and appropriate scales
- Size of the graphic effect should be directly proportional to the numerical quantities (“lie factor”)
- Show data variation, not design variation

Theory of Data Graphics

Definition of Data Ink

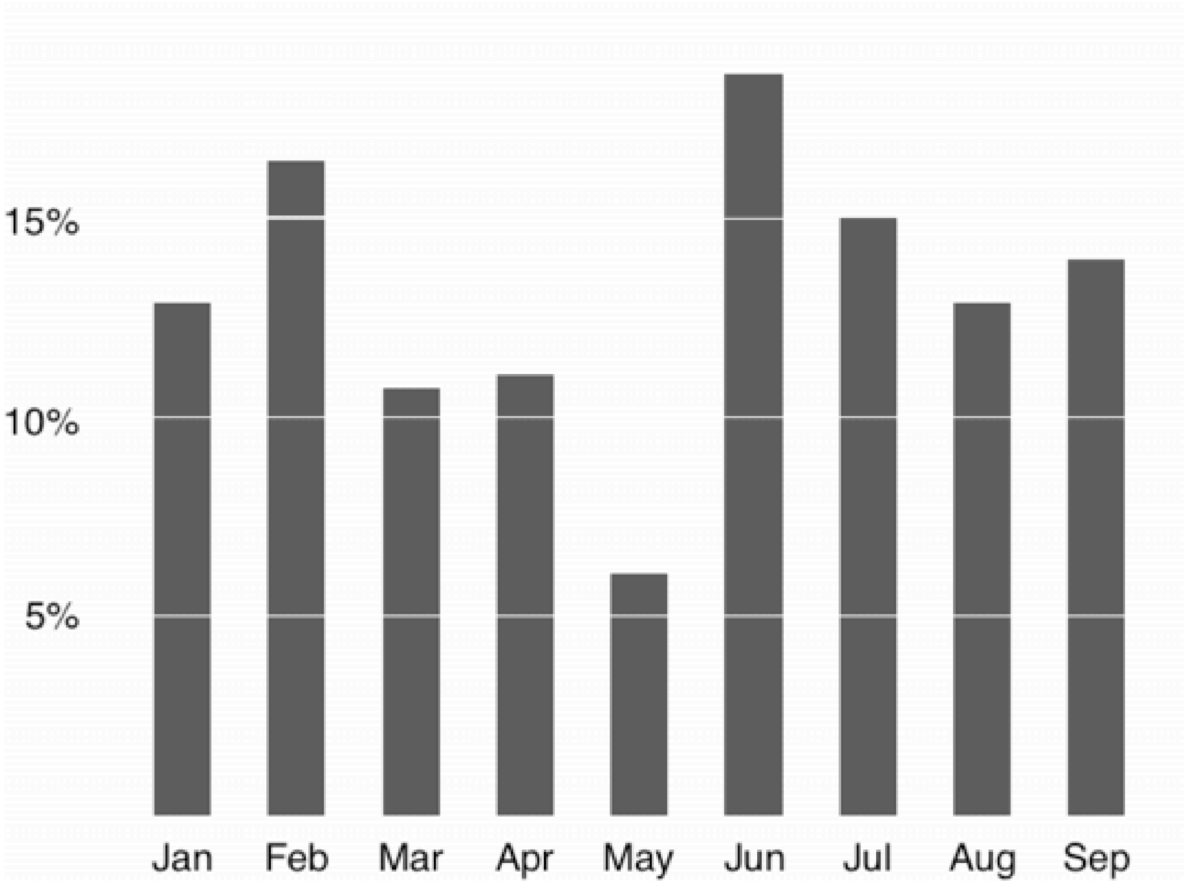
- **Data-ink refers to the non-erasable ink used for presenting the data**.
  - If data-ink would be removed from the image, the graphic would lose the content.



Source: Slide Adapted from Hanspeter Pfister. Lecture Notes for CS 171. Harvard University. http://cs171.org/

Focus on Data: Avoid Chartjunk

Chartjunk is the extraneous visual elements that distract from the message!!



Source: The image is from The Data-Ink Ratio Example

Other Subjective Design Principles

- **Aesthetics:** Attractive things are perceived as more useful than unattractive ones
- **Style:** Communicates brand, process, who the designer is
- **Playfulness:** Encourages experimentation and exploration
- **Vividness:** Can make a visualization more memorable

Source: Slide Adapted from Hanspeter Pfister. Lecture Notes for CS 171. Harvard University. http://cs171.org/. Information is based on comments by Pat Hanrahan in November 2007. Those can be found at http://www.perceptualedge.com/blog/?p=199

Data Models

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues "came out by that same door as they went in," and in order to have another try at agreement, the committee begged to be continued for another year.

For its final report (1940) the committee chose a common bone for its contentions, directing its arguments at a concrete example of a sensory scale. This was the Sone scale of loudness (S. S. Stevens and H. Davis. *Hearing*. New York: Wiley, 1938), which purports to measure the subjective magnitude of an auditory sensation against a scale having the formal properties of other basic scales, such as those used to measure length and weight. Again the 19 members of the committee came out by the routes they entered, and their views ranged widely between two extremes. One member submitted "that any law purporting to express a quantitative relation between sensation intensity and stimulus intensity is not merely false but is in fact meaningless unless and until a meaning can be given to the concept of addition as applied to sensation" (Final Report, p. 245).

It is plain from this and from other statements by the committee that the real issue is the meaning of measurement. This, to be sure, is a semantic issue, but one susceptible of orderly discussion. Perhaps agreement can better be achieved if we recognize that measurement exists in a variety of forms and that by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

A Classification of Scales of Measurement

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale.

Scales are possible in the first place only because there is a certain isomorphism between what we can do with the aspects of objects and the properties of the numeral series. In dealing with the aspects of objects we invoke empirical operations for determining equality (classifying), for rank-ordering, and for determining when differences and when ratios between the aspects of objects are equal. The conventional series of numerals yields to analogous operations: We can identify the members of a numeral series and classify them. We know their order as given by convention. We can determine equal differences, as $8-6=4-2$, and equal ratios, as $8/4=6/3$. The isomorphism between these properties of the numeral series and certain empirical operations which we perform with objects permits the use of the series as a *model* to represent aspects of the empirical world.

The type of scale achieved depends upon the character of the basic empirical operations performed.

Data Types (from S. Stevens, Theory of Scales)

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

Data Types: Explained

Nominal:

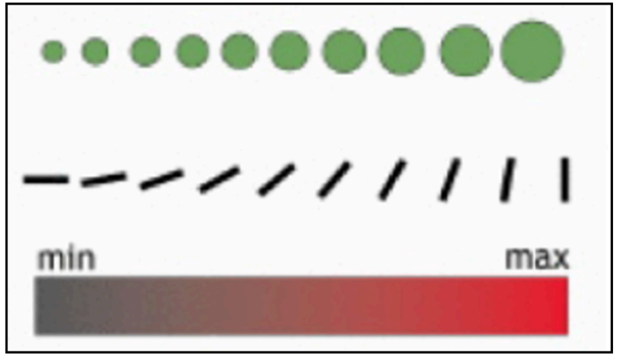
- Are $=$ or $\neq$ to other values
- Apples, bananas, oranges, etc.

Ordinal:

- Obey a $<$ relationship
- Small, medium and large

Quantitative:

- Can do math on them
- 50 inches, 53 inches, etc.



Quantitative Data Types (from S. Stevens)

Quantitative data can be further divided into:

Intervals (Location of Zero Arbitrary):

- Dates: Jan 19; Location: (Lat, Long)
- Only differences (i.e., intervals) can be compared.

Ratio (Zero Fixed):

- Measurements: Length, Weight, ...
- Origin is meaningful, we can compute ratios, proportions, differences, etc.

Non-Graded Activity: Data Terminology

02:00

Activity | Your Solution

> In 2 minutes, please identify an appropriate data type for each column below.

Order ID	Order Date	Order Priority	Product Container	Product Cost	Ship Date
1	1/1/2022	5 - low	Large box	25	1/5/2022
2	1/4/2022	4 - not specified	Small Box	36	1/7/2022
3	1/15/2022	2- high	Small Box	38	1/17/2022
3	1/15/2022	2- high	Small Box	41	1/17/2022
3	1/15/2022	2- high	Jumbo Box	44	1/17/2022
3	1/15/2022	2- high	Wrap Bag	33	1/17/2022
4	1/18/2022	1- urgent	Small Box	33	1/19/2022

Showing 1 to 7 of 11 entries

Previous 1 2 Next

Data vs. Conceptual Models

From data model

- 32.5, 54.0, -17.3, ...

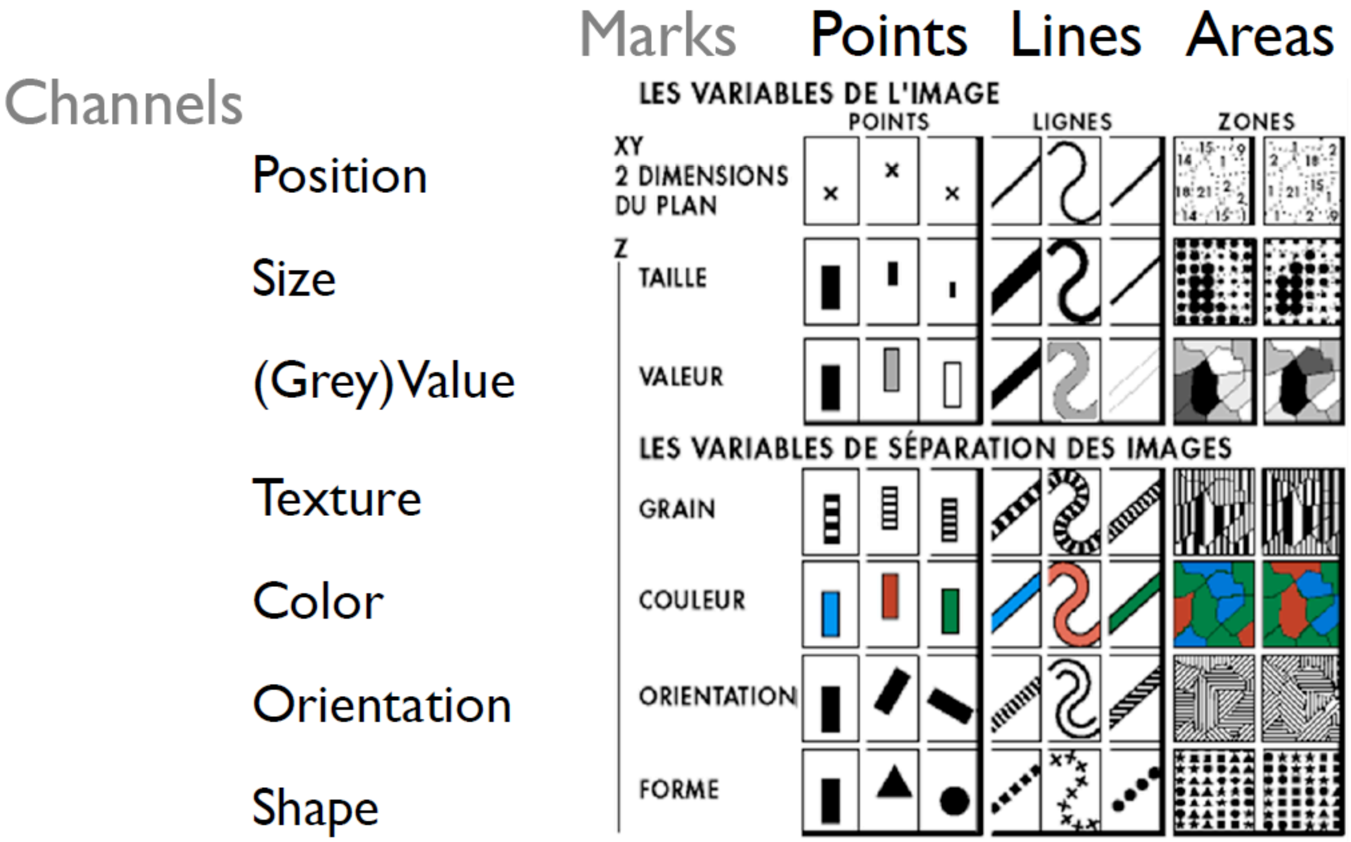
Using a conceptual model:

- Temperature

To data type:

- Continuous to x significant digits i.e. quantitative
- Hot, warm, cold i.e. ordinal
- Burned vs. not burned i.e. nominal

Image Model: Visual (Encoding) Variables



Source: Bertin (1967), Seminology of Graphics.

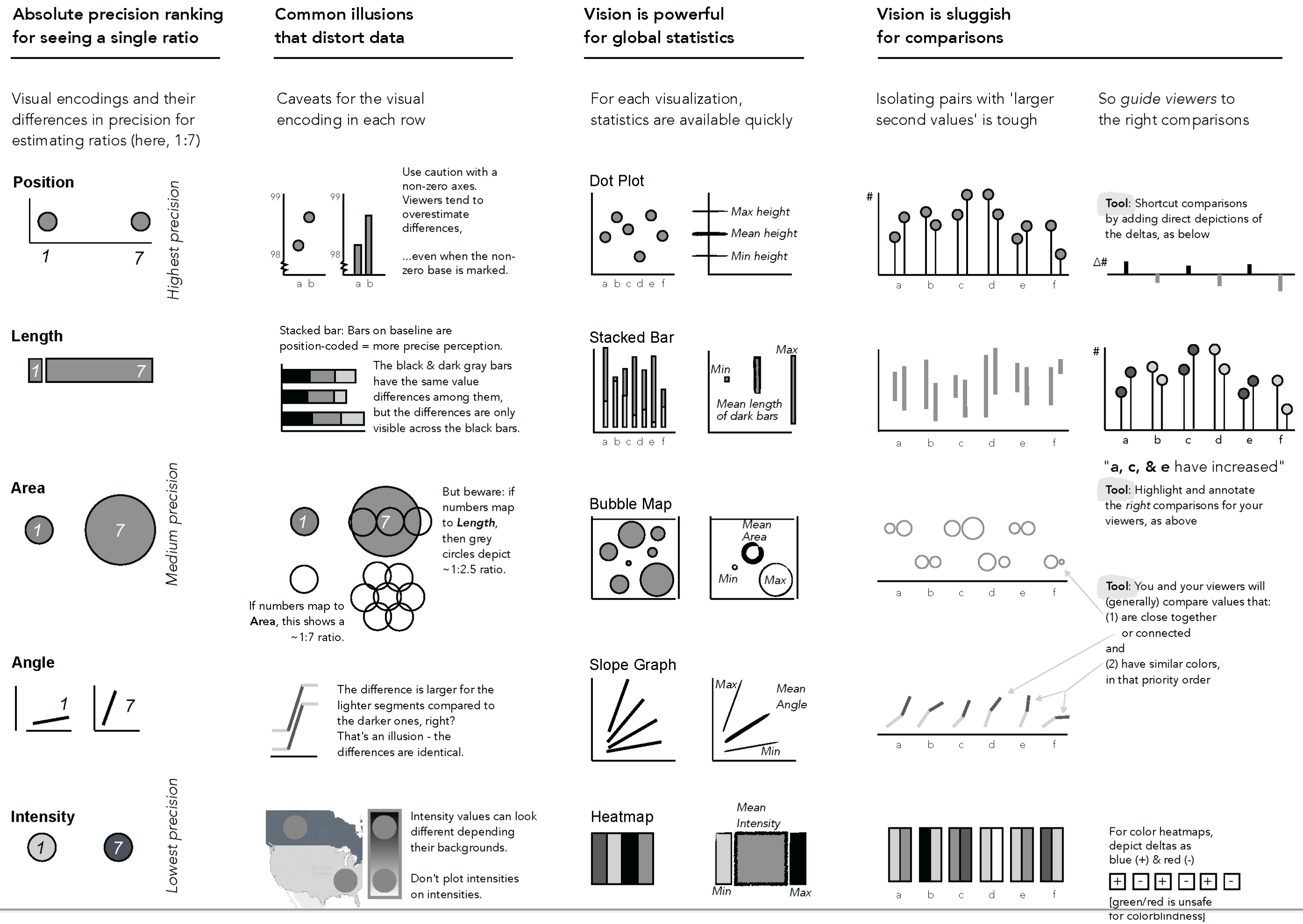
Mapping to Data Types

	Nominal	Ordinal	Quantitative
Position	✔	✔	✔
Size	✔	✔	~
(Grey)Value	✔	✔	~
Texture	✔	~	✖
Color	✔	✖	✖
Orientation	✔	✖	✖
Shape	✔	✖	✖

✔ = Good
~ = OK
✖ = Bad

Source: Slide Adapted from Hanspeter Pfister. Lecture Notes for CS 171. Harvard University. http://cs171.org/

Visual Channels and their Precision



Source: Franconeri, et al. (2021). The Science of Visual Data Communication: What Works. To view the full size image, click here.

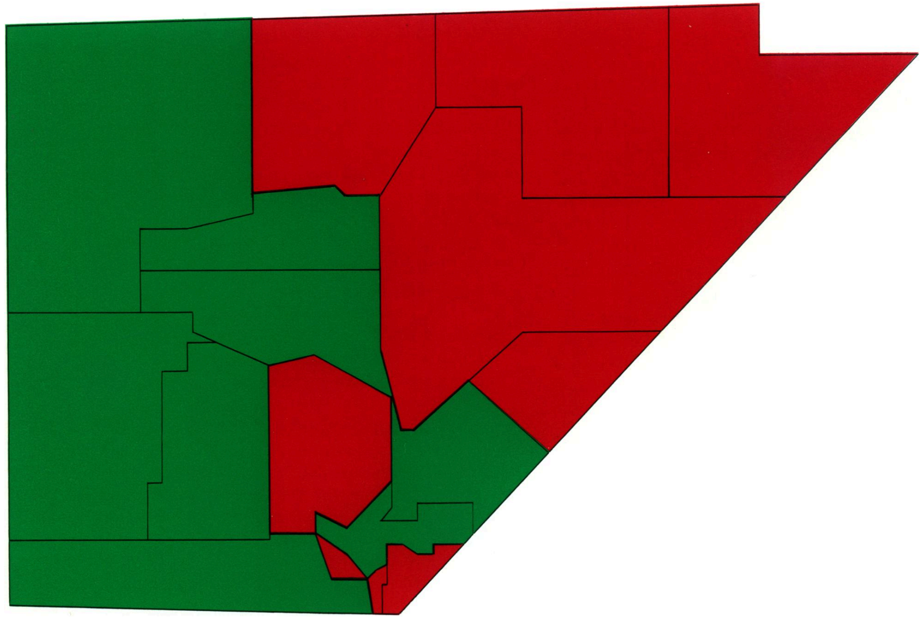
Color Should Be Used Sparingly

Color

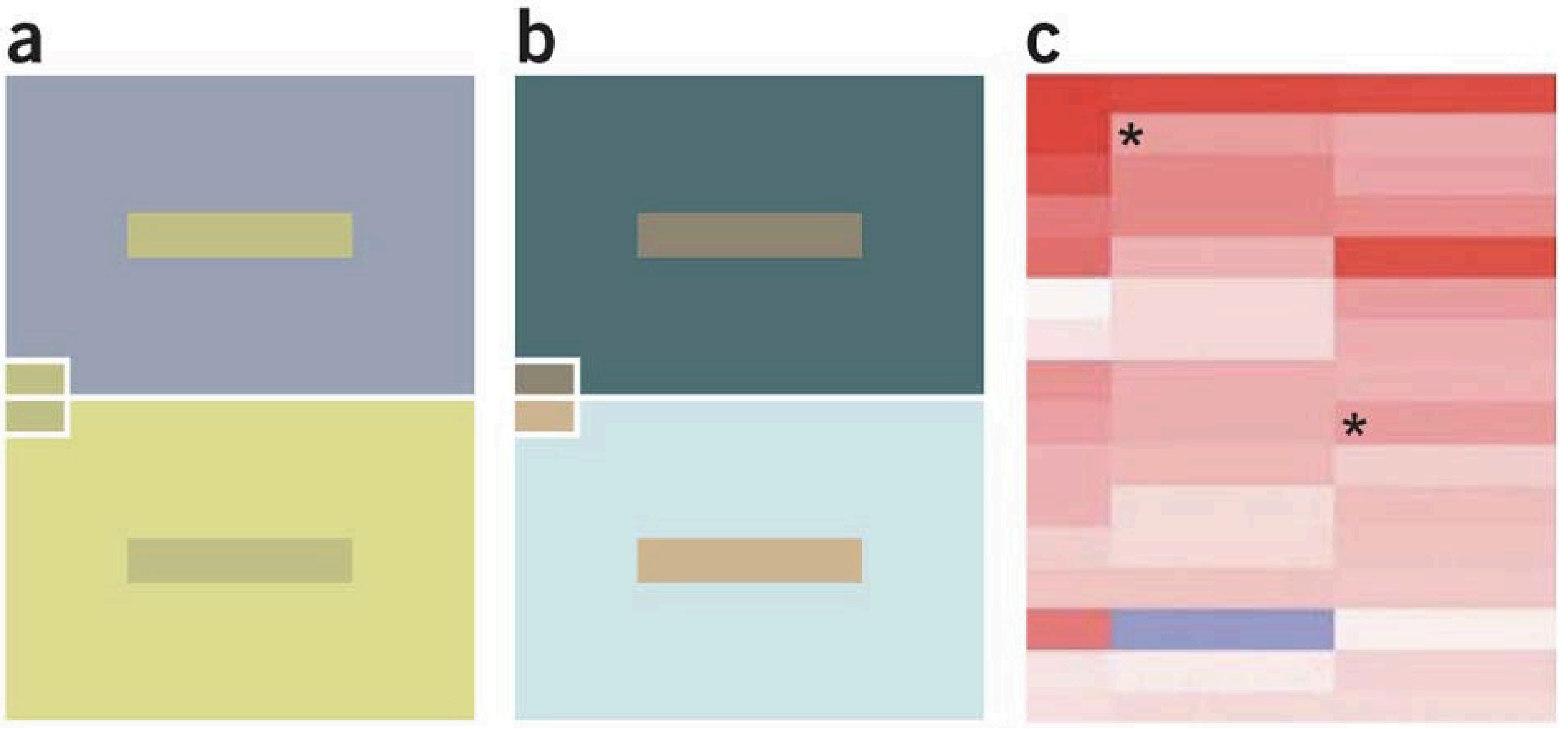
01:00

Activity | Your Solution | Results from 1983 Experiment

- The following map of Nevada has been colored to indicate various geological features in each county.
- Estimate the larger land area-more red, more green, or the same-and mark your answer on the mentimeter poll in the next panel.
- **Please work fairly quickly, as if you were trying to gain an overall impression from a map**.



Simultaneous Contrast Affects Perception



(a,b) The same color can look different (a), and different colors can appear to be nearly the same by changing the background color (b)[1]. (c) The rectangles in the heat map indicated by the asterisks (*) are the same color but appear to be different.

Source: Slide Adapted from Hanspeter Pfister. Lecture Notes for CS 171. Harvard University. http://cs171.org/

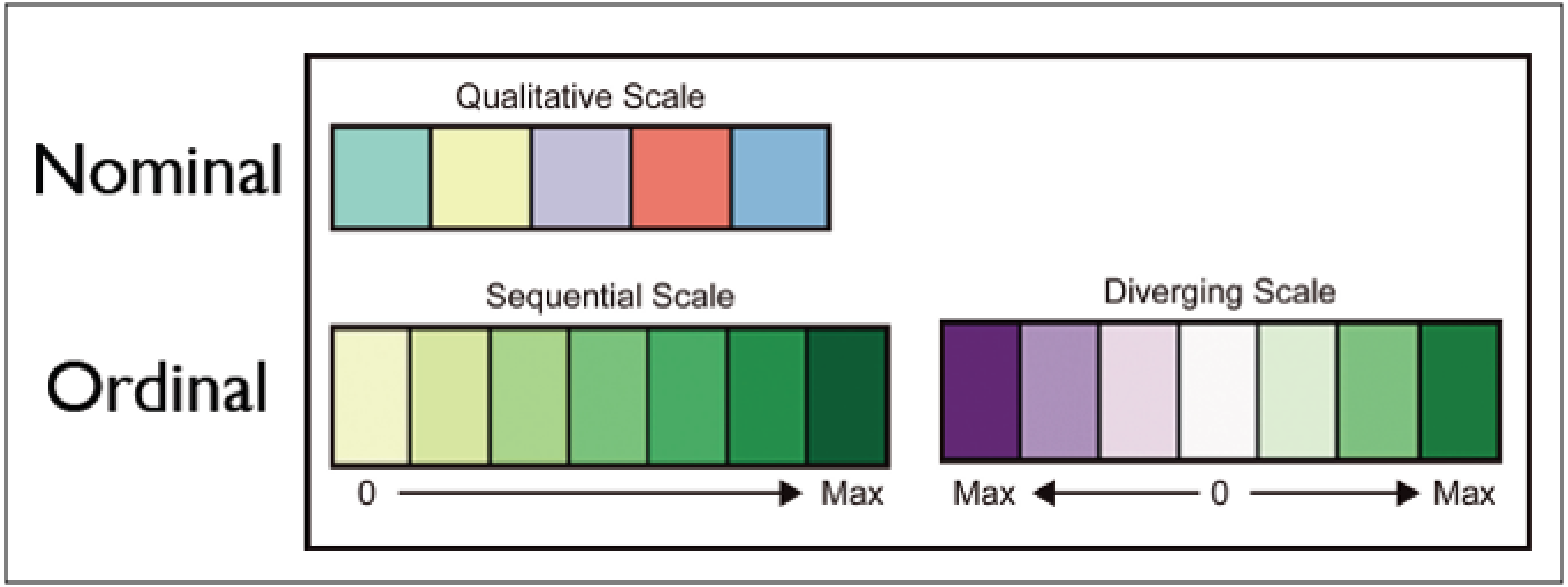
Color Blindness

> "About 1 in 12 men are color blind" -- NIH's At a glance: Color Blindness

Color Blindness Can Distort a Person's Reading/Interpretation of a Chart

- **Personal Check:** To have a feel for color blindness (if you are not color blind), you can take this color blind test
- Given the high prevalence of color blind individuals, your charts **should** accommodate for color-blindness. **How?**
  - Use color sparingly
  - Use color friendly palettes, e.g., see https://colorbrewer2.org/

Color Brewer: Color Scales and their Selection



Source: Brewer, Cynthia A. “Color use guidelines for data representation.”Proceedings of the Section on Statistical Graphics, American Statistical Association. 1999.

Recap

Summary of Main Points

By now, you should be able to do the following:

- Explain the concept of "graphical excellence"
- Explain the theory of data graphics
- Optimize visual encoding based on data types
- Understand why color should be used sparingly and how to select appropriate colors (when color is a must)

Things to Do Prior to Next Class

Please go through the following two supplementary readings and complete assignment 08.

- The Lie Factor and the Baseline Paradox; especially noting what the authors mean by "baseline", how the lie factor may be ignored in time-series applications, and/or in applications involving a "ratio" scale.
- Useful junk? The effects of visual embellishment on comprehension and memorability of charts, which presents an experimental counter against Tufte's argument for simplicity (by quantifying vividness and recall of data from the more artistic charts). Note they define **"ratio"** different from how we have defined in class.