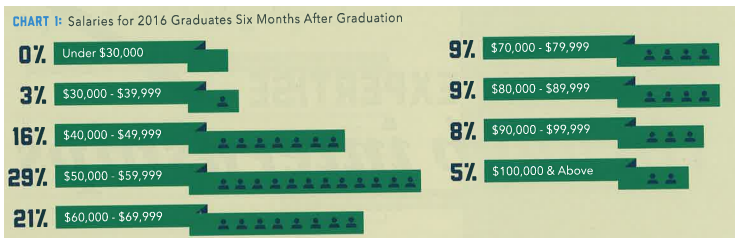# Lecture 1:
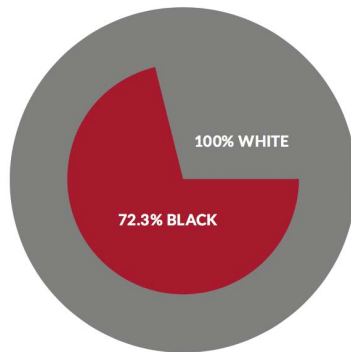# Welcome to Data Visualization Using R

January 22, 2018

Take This Class So You Won't Make This Graphic

From Trachtenberg's most recent magazine issue.

Or This One

EQUALITY INDEX OF BLACK AMERICA, 2016–2017

	REVISED 2016	2017
EQUALITY INDEX	72.2%	72.3%
Economics	56.2%	56.5%
Health	79.4%	80.0%
Education	77.4%	78.2%
Social Justice	60.9%	57.4%
Civic Engagement	100.6%	100.6%

"U.S. Metros Ranked on Black-White Income Inequality," *Next City*, May 2, 2017

Instead, Aspire to This

See Tufte for citation.

Overview

Course Administration

Tufte, Grandfather of Visualization

Getting Started with R

Course Administration

1. Syllabus
   - sign up for good/bad/ugly
2. Questions/issues with readings?
3. Make sure you're signed up for Piazza
4. Introductions
   - name and degree
   - why this course?
   - what you do now
   - what you'd like to do when you're done

Tufte

Edward Tufte

- A quantitative political scientist
- Writing in the mid-1970s
- Became interested in visualization by working with pioneering statistician John Tukey
- Remember that this is the pre-Excel era, in which data graphics are difficult to make

An Argument for Better Visualization

Because good visualizations tell the most compelling story

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

An Argument for Better Visualization

Because good visualizations tell the most compelling story
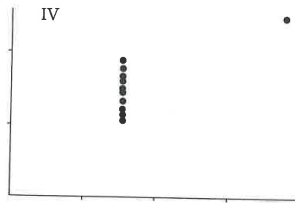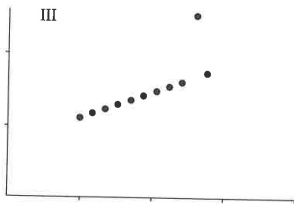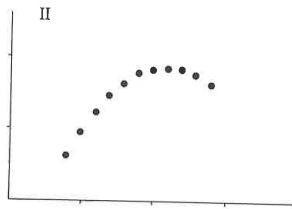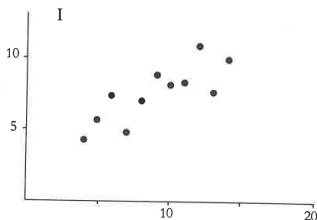
Tufte's Types of Graphs

1. Data maps
2. Time series
3. Space-time narrative designs
4. Relational graphs – the holy grail

Data Maps

- Describe the location of numbers
- This can be revealing or obfuscating
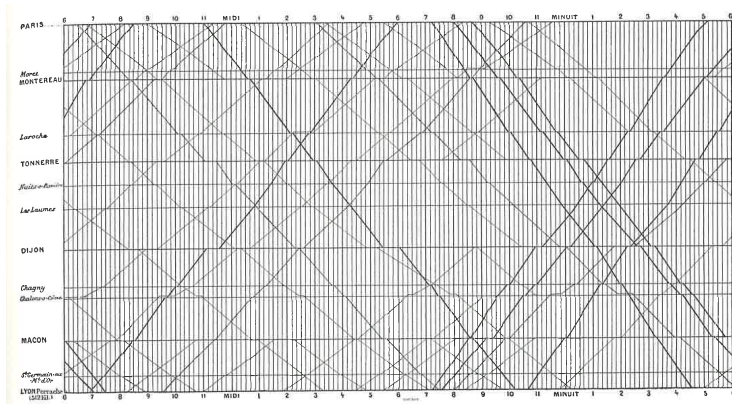- We will make these in this class

Data Map Example

Time Series

- Time on the horizontal axis
- Something else on the vertical axis
- One of the first types of data graphics

Train, Paris to Lyon

See Tufte for citation.

Space-Time Narrative Designs

- Move over space and time at the same time
- A time series plus

Space-Time Narrative Example

Which dimensions?

Space-Time Narrative Example

Which dimensions?

1. army size
2. army location, N/S
3. army location, E/W
4. direction of movement
5. temperature
6. by date

Relational Graphics

- One variable on the vertical, another on the horizontal
- A conceptual advance in graphics
- A more sophisticated way of thinking

Relational Graphics Example
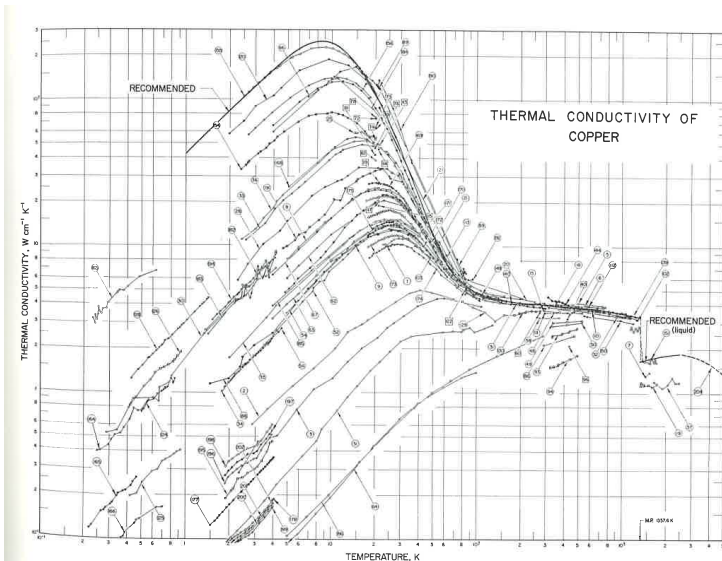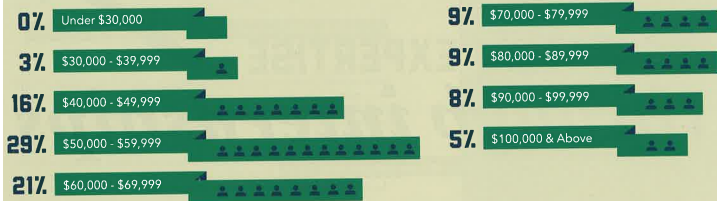
Tufte's Main Causes of Distortion in Graphics

1. Data are bad
   - should be per capita and are not
   - data are not consistent over time
   - don't adjust for inflation
2. Graphics are rotten
   - size doesn't match the numbers
   - colors and styles are misleading
   - graphic fails to highlight key point
3. Graphics are irrelevant
   - too much extraneous stuff

Size and Number Don't Match

CHART 1: Salaries for 2016 Graduates Six Months After Graduation

0% Under $30,000

3% $30,000 - $39,999

16% $40,000 - $49,999

29% $50,000 - $59,999

21% $60,000 - $69,999

9% $70,000 - $79,999

9% $80,000 - $89,999

8% $90,000 - $99,999

5% $100,000 & Above

A Continuing Problem: Graphics are Irrelevant

The Company Cafeteria was used by 9 Out of 10 Employees during the Fiscal Year 1949

0   100%
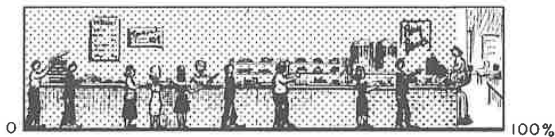
Source: COMPANY REPORTS

Tufte's Six Rules of Graphic Integrity, 1 to 3 of 6

1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.

2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.

3. Show data variation, not design variation.

Tufte's Six Rules of Graphic Integrity, 4 to 6

4. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.

5. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.

6. Graphics must not quote data out of context.

R

What is R?

- A programming language
- Developed by statisticians from New Zealand
- Open source, and therefore free
- Based on "S," developed by Bell Labs

Today's Goals

- When you leave today, you will be able to
  - run a R script
  - create a R Markdown file
- Download the R tutorial for this class now.
- You'll continue work at home on your own and turn in a problem set next lecture

Next Lecture

- Turn in PS 1
- Read Few Chapters 5 and 7
- R Graphics Cookbook, Chapters 2 and 15 – see readings list for specific pages
- I'll hand out instructions for class project