GR5293 Statistical Graphics

Spring 2022

Tues/Thurs 2:40 - 3:55pm

Hamilton 602 (starting 2/1)

Joyce Robbins, Dept. of Statistics

jtr13@columbia.edu

Keeping us all safe

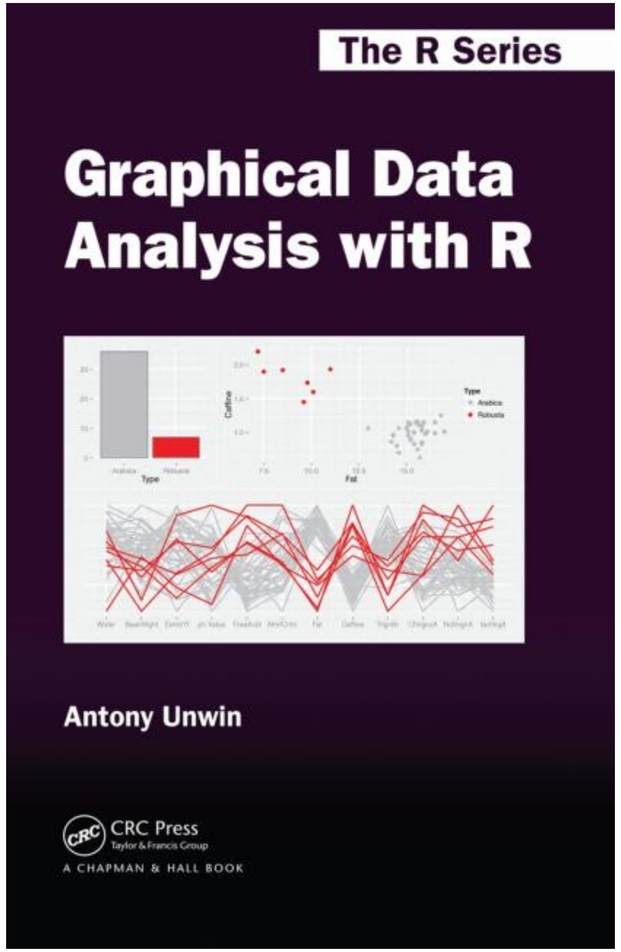
- Do not come to class if you don’t feel well or need to quarantine.
- Make sure you wear an approved mask (surgical, KN95, KF94, or N95) covering your nose and mouth in class
- Zoom access is available through the “Zoom Class Sessions” link on CourseWorks
- Recordings are available under the Class Recordings tab or Course Video Recordings (Panopto)

Agenda

- Syllabus
- Exploratory Data Analysis
- Visualization
- Critique
- Tools

SYLLABUS

Antony Unwin
2015
Graphical Data Analysis with R
CRC Press
ISBN 978-1498715232



Class requirements and grading

Your grade will be based on the following:

25% Problem Sets

10% Community Contribution

30% Final Exam

30% Final Project

5% Peer Review of Final Projects

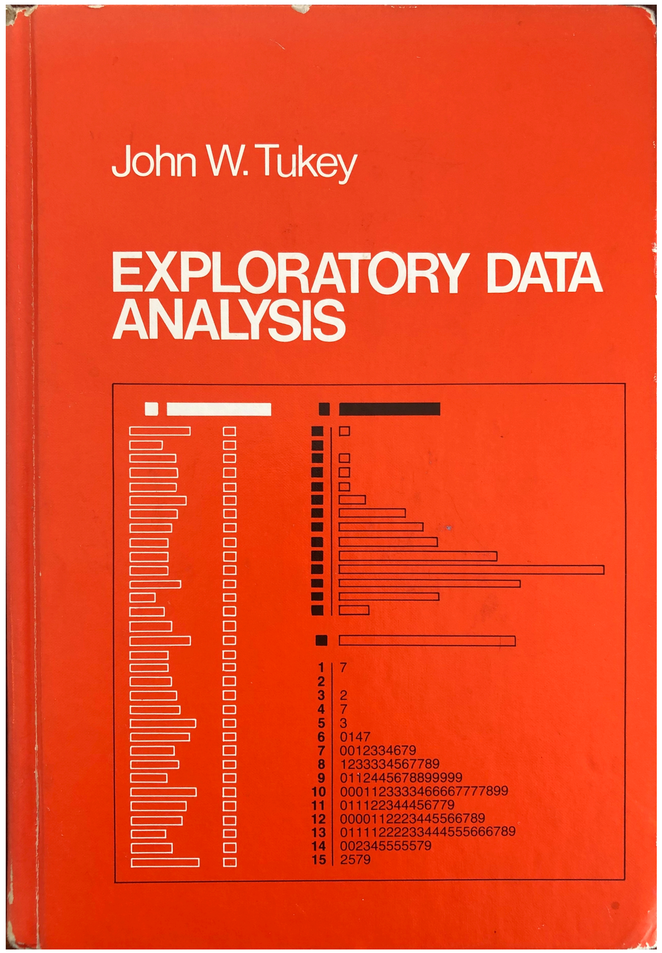
Community Contribution

There are many ways in which you can contribute:

- give a well-rehearsed 5 minute lightning talk in class (live or video) on a datavis topic (theory or tool)
- create a cheatsheet or other resource
- write a tutorial for a tool that's not well documented
- build a viz product (ex. htmlwidget) for class use
- create a web site for sharing class resources publicly
- organize and a lead a help session on a topic you've mastered
- other...

EXPLORATORY DATA ANALYSIS

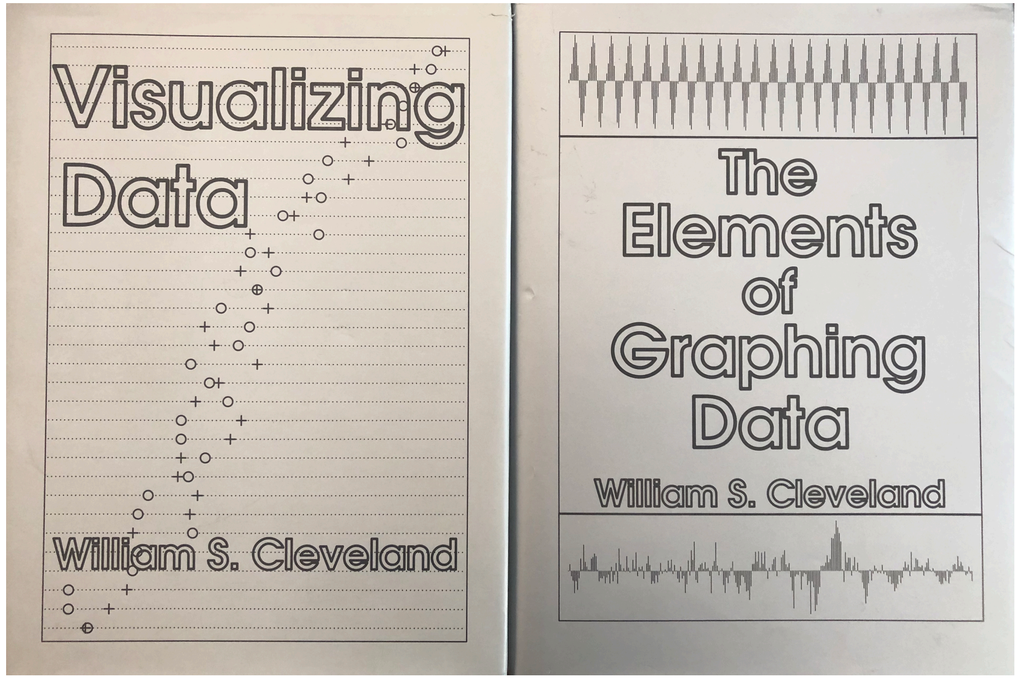
Tukey 1977



"Exploratory data analysis is detective work."

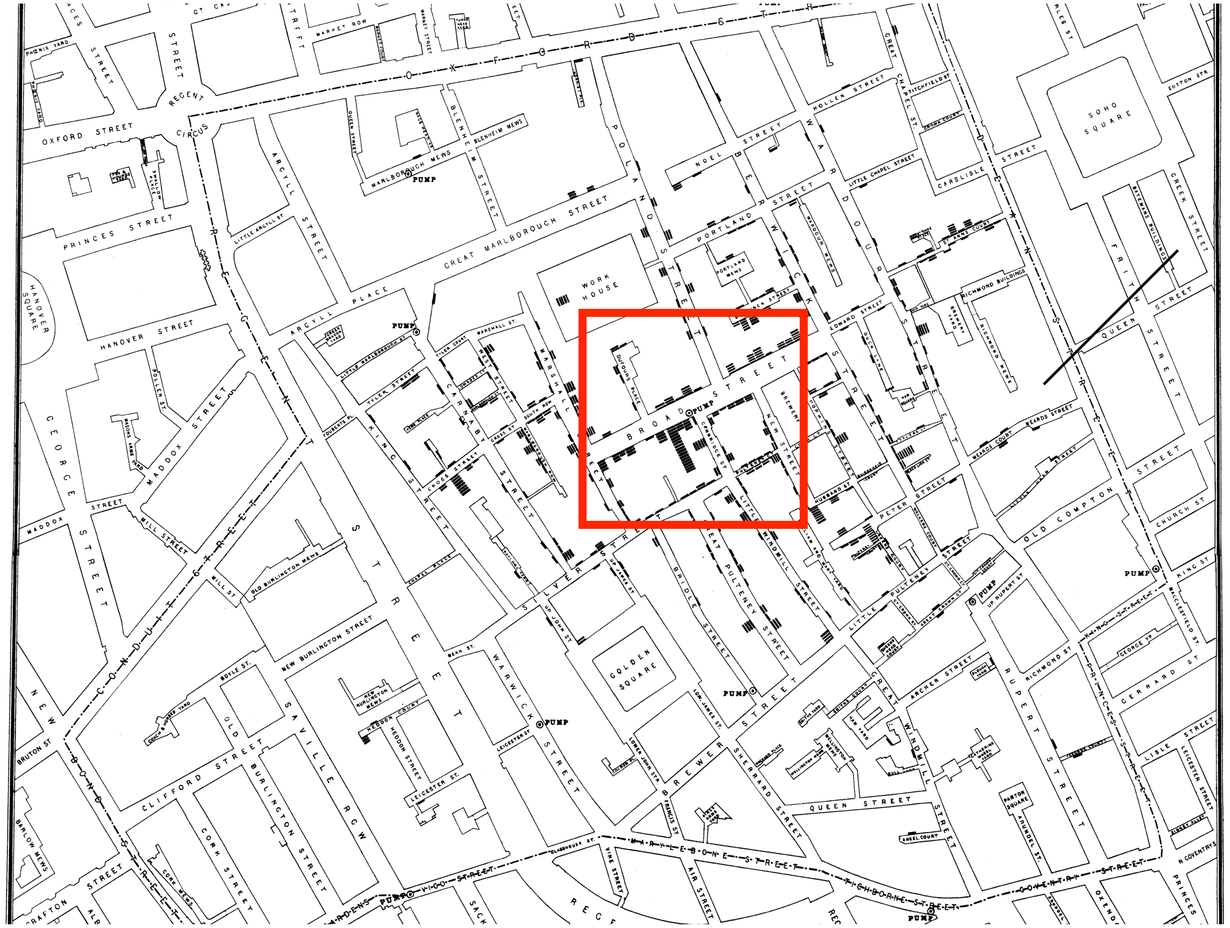
"Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone--as the first step."

Cleveland 1980s, 90s

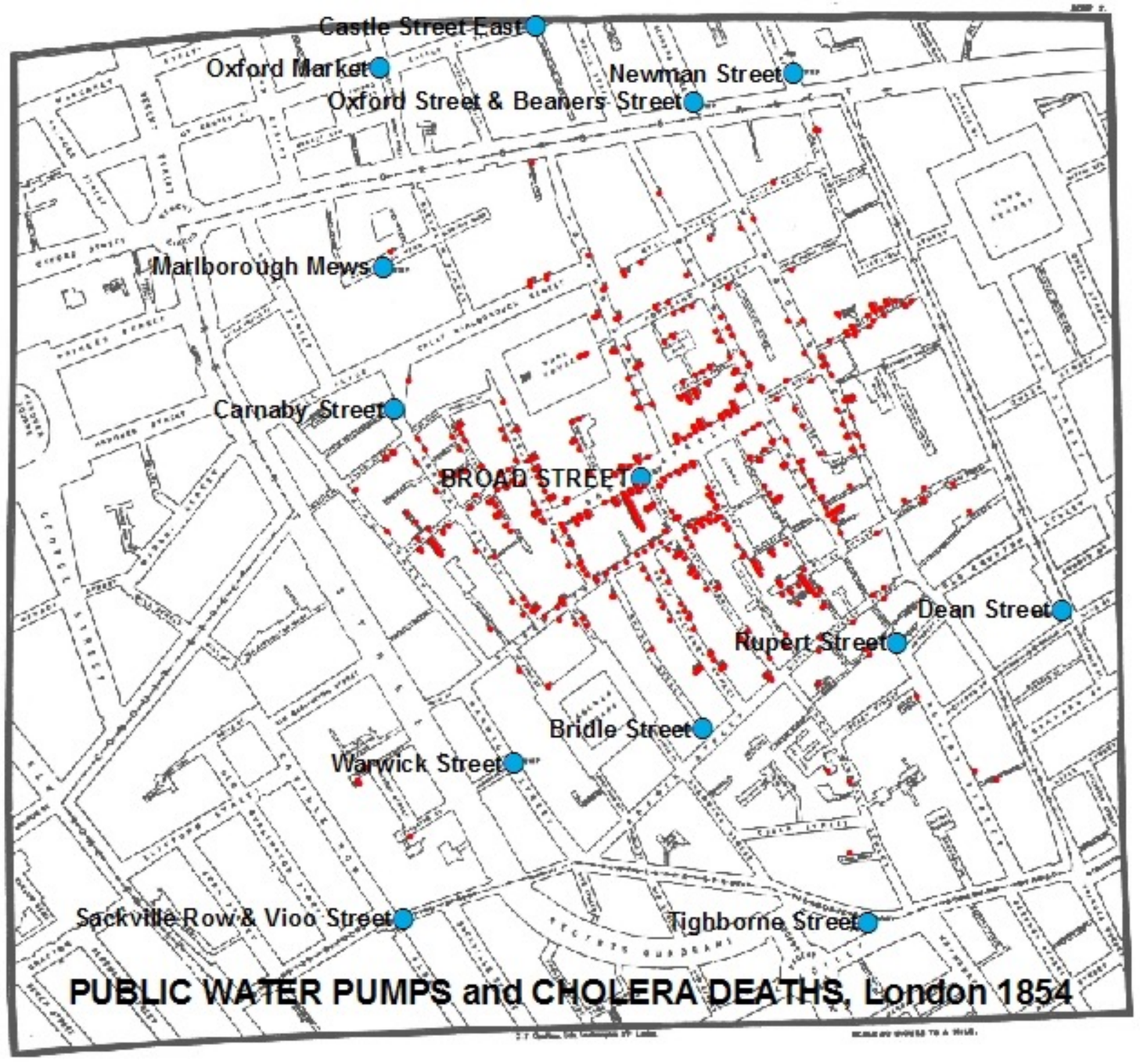


"Visualization is critical to data analysis. It provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones."

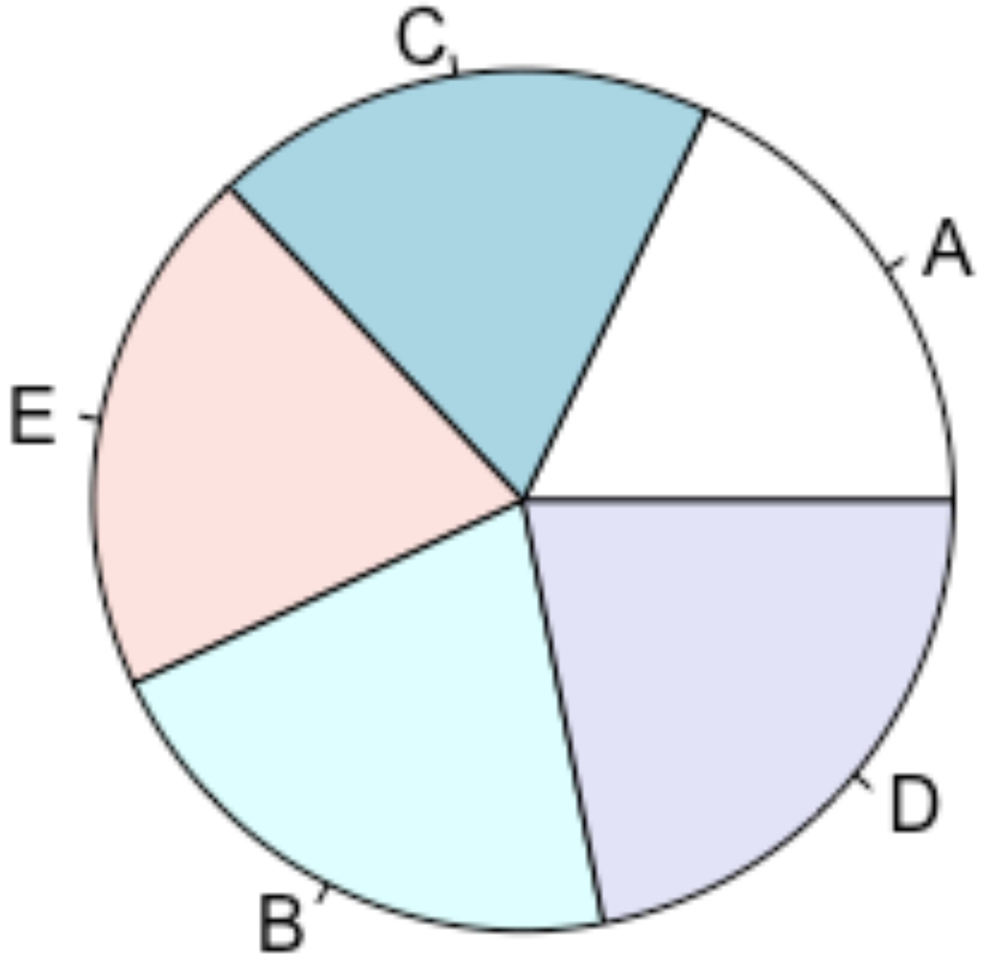
John Snow, Cholera Map 1854



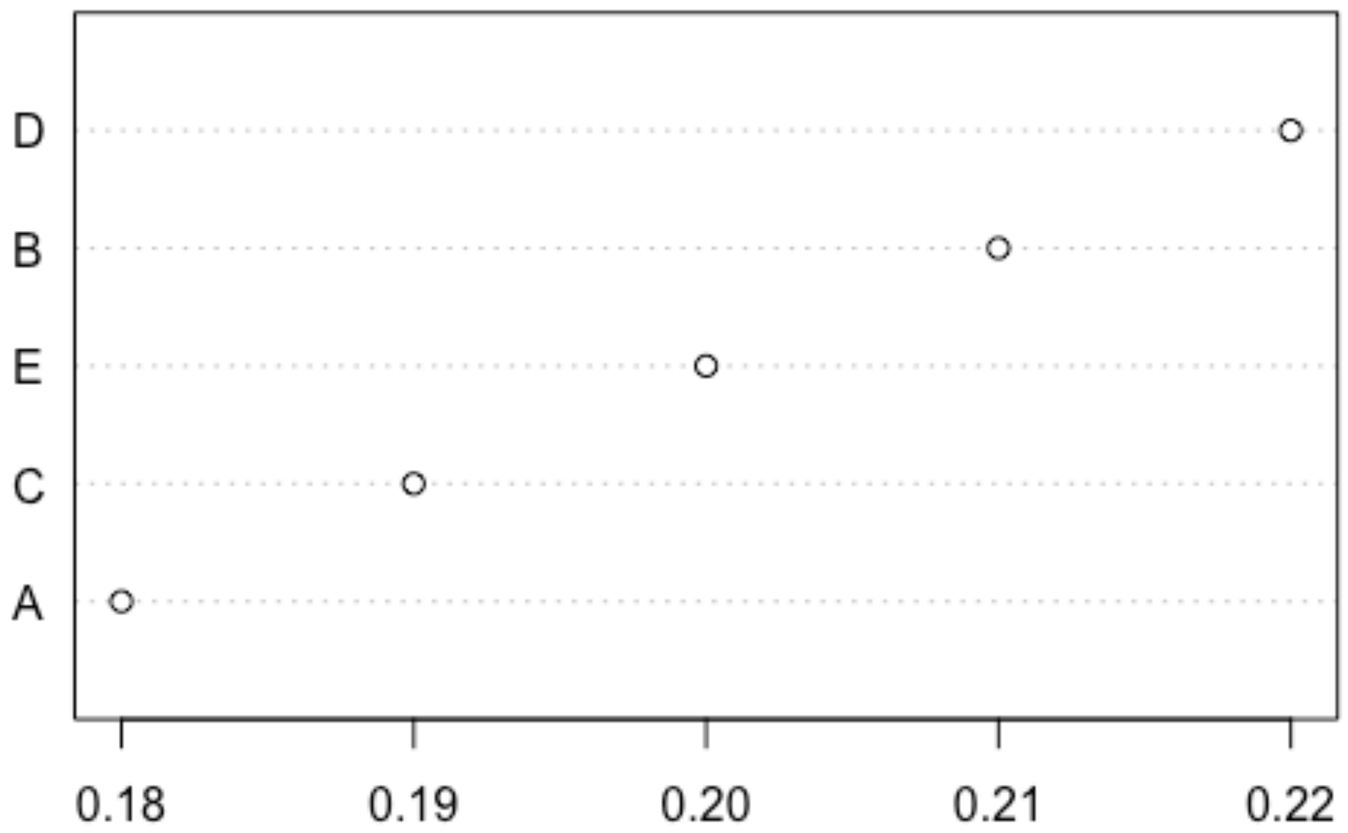
Broad St. pump


Castle Street East
Oxford Market
Newman Street
Oxford Street & Beaners Street
Marlborough Mews
Carnaby Street
BROAD STREET
Dean Street
Rupert Street
Bridle Street
Warwick Street
Sackville Row & Vioo Street
Tighborne Street
PUBLIC WATER PUMPS and CHOLERA DEATHS, London 1854

Perception



Perception



Anscombe's Quartet

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

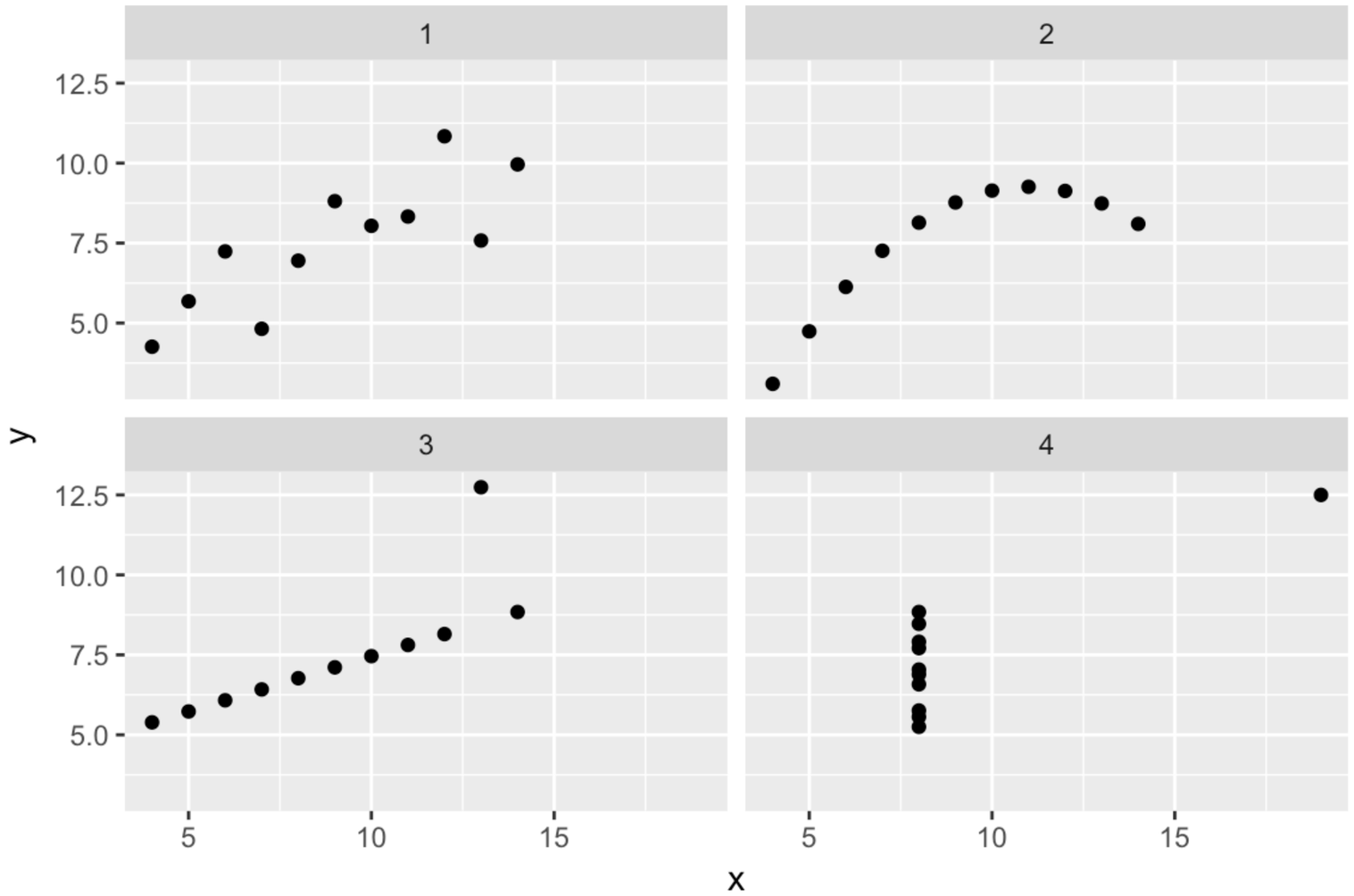
Numeric summary

Each of the four data sets yields the same standard output from a typical regression program, namely

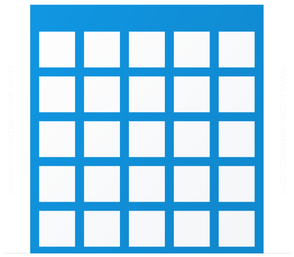
Number of observations $(n) = 11$
Mean of the x's $(\bar{x}) = 9.0$
Mean of the y's $(\bar{y}) = 7.5$
Regression coefficient (b_1) of y on $x = 0.5$
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x} = 110.0$
Regression sum of squares $= 27.50$ (1 d.f.)
Residual sum of squares of $y = 13.75$ (9 d.f.)
Estimated standard error of $b_1 = 0.118$
Multiple $R^2 = 0.667$

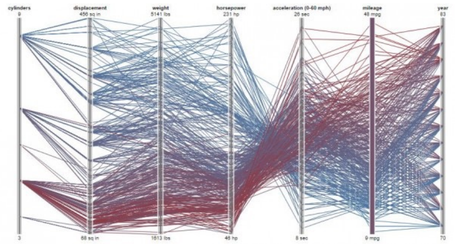
F.J. Anscombe, "Graphs in Statistical Analysis"

Graphical summary



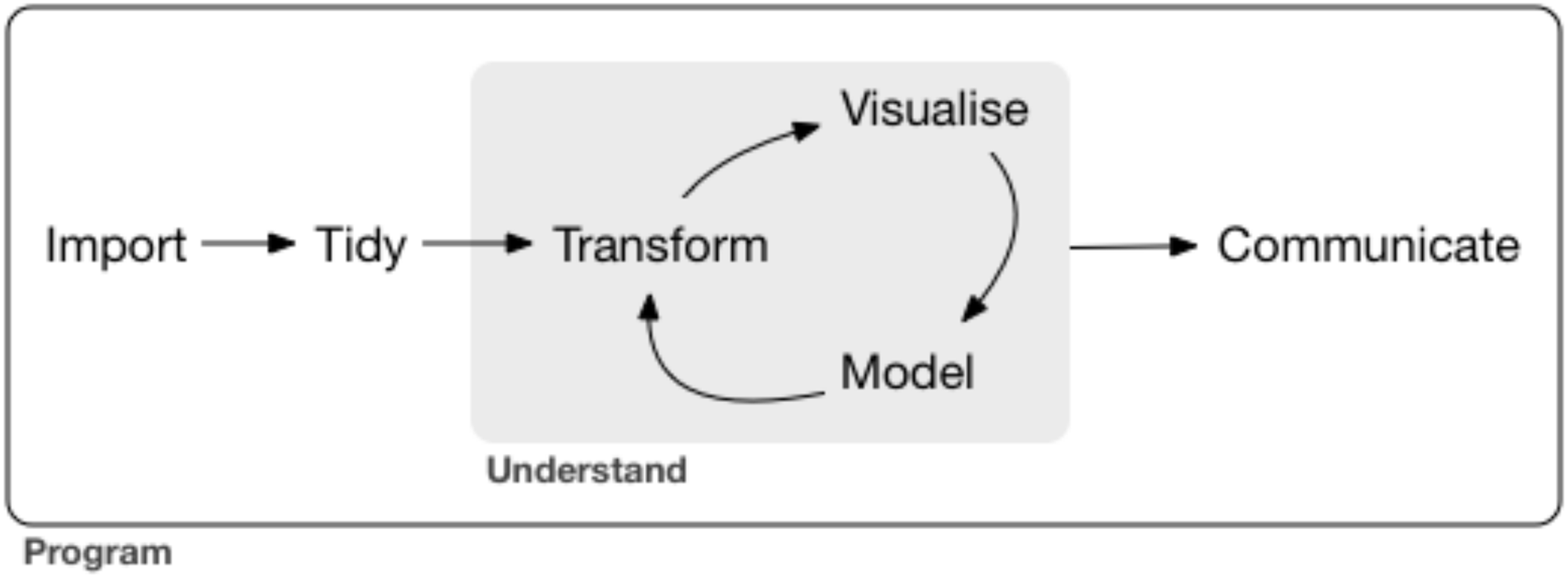
How do we gain insight?





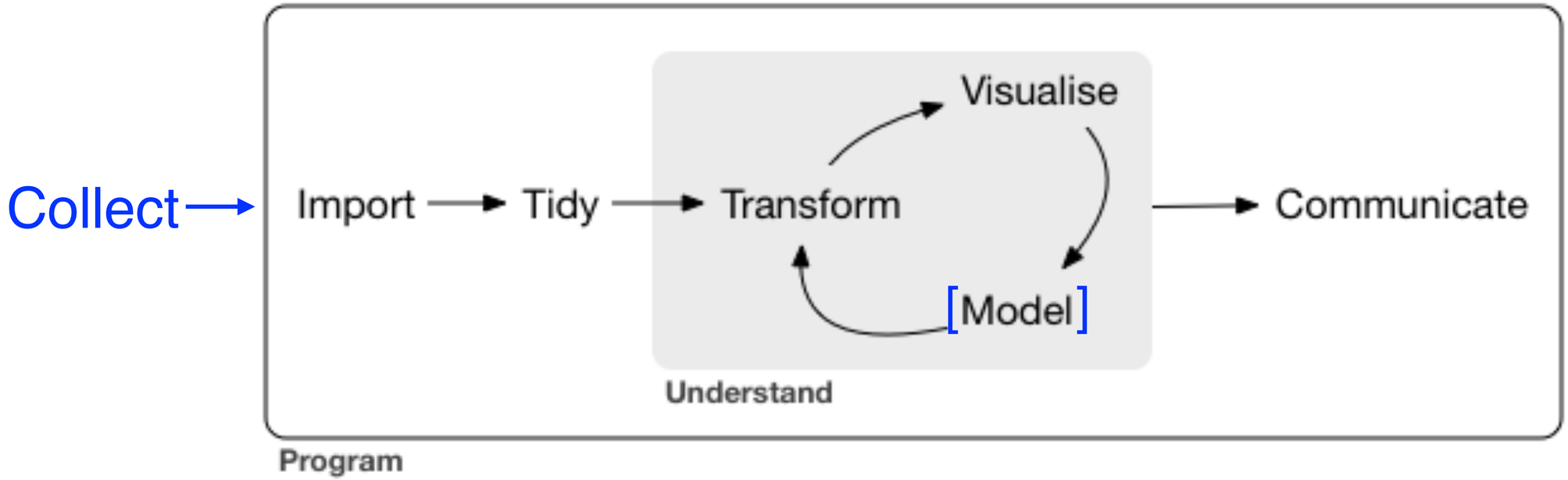
- Deep understanding of the dataset, where it came from, what its limitations are
- Experiment with different graphic forms, based on theory on what forms work well with different data types

Data science pipeline



Source: r4ds.had.co.nz/introduction.html

Data science pipeline



Source: r4ds.had.co.nz/introduction.html

"Visualization is a fundamentally human activity."

Start with the data

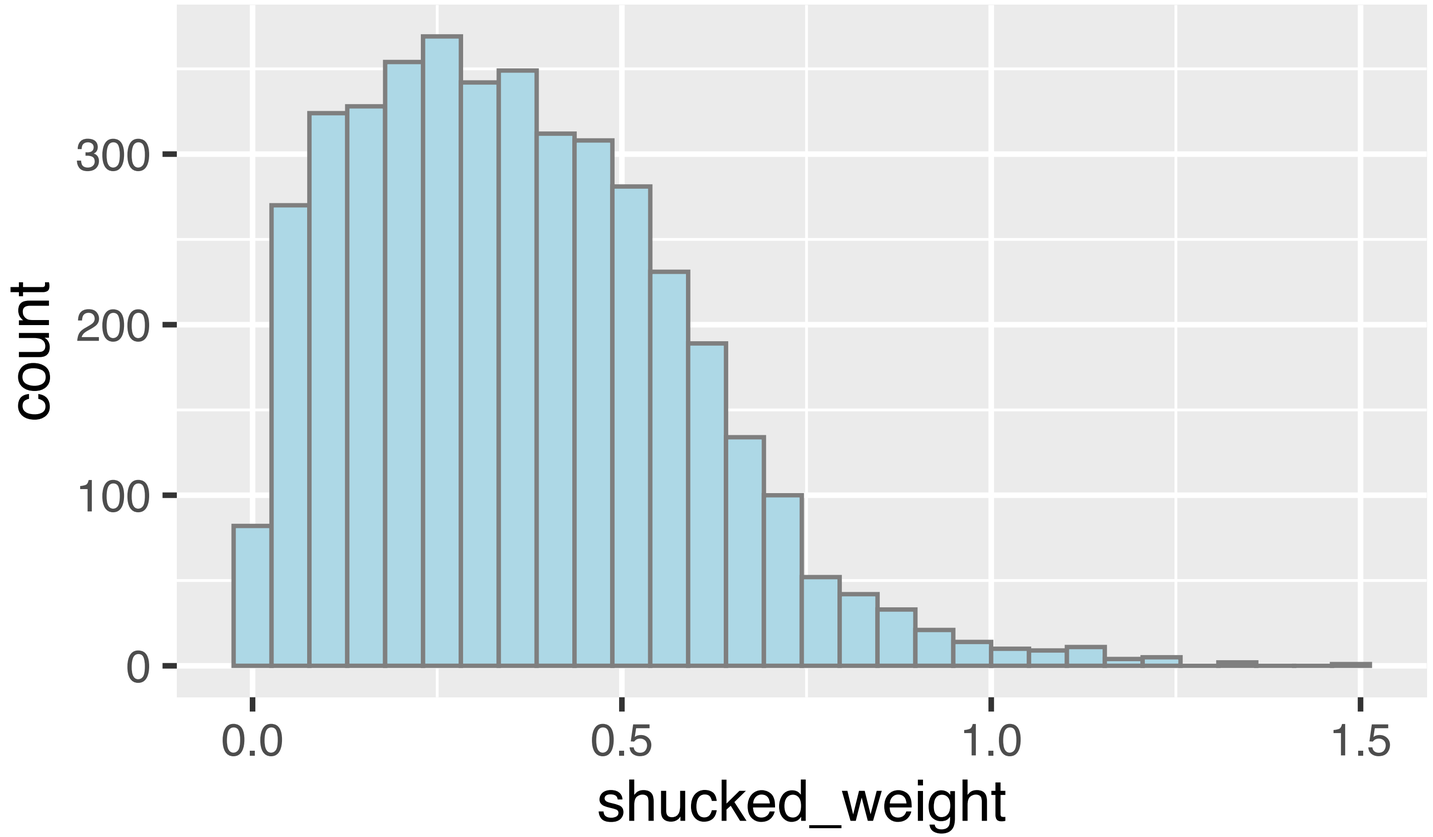
```
> library(ucidata)
> str(abalone)
#> 'data.frame':    4177 obs. of  9 variables:
#>  $ sex           : Factor w/ 3 levels "F","I","M": 3 3 1 3 2 2 1 ...
#>  $ length        : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 ...
#>  $ diameter      : num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 ...
#>  $ height        : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 ...
#>  $ whole_weight  : num  0.514 0.226 0.677 0.516 0.205 ...
#>  $ shucked_weight: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
#>  $ viscera_weight: num  0.101 0.0485 0.1415 0.114 0.0395 ...
#>  $ shell_weight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26...
#>  $ rings         : int  15 7 9 10 7 8 20 16 9 19 ...
```

Know what you’re studying

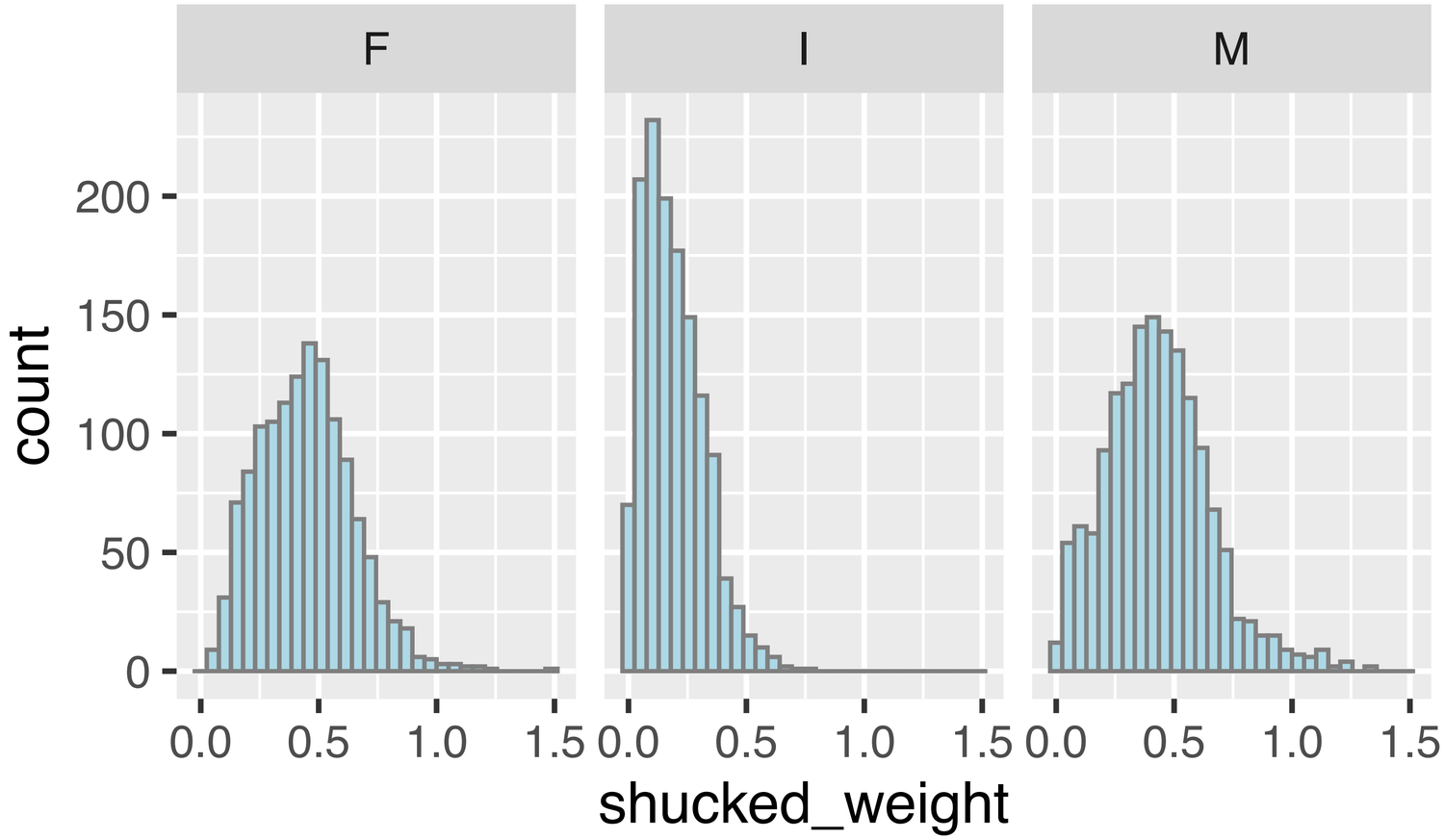
abalone???

Know what you’re studying

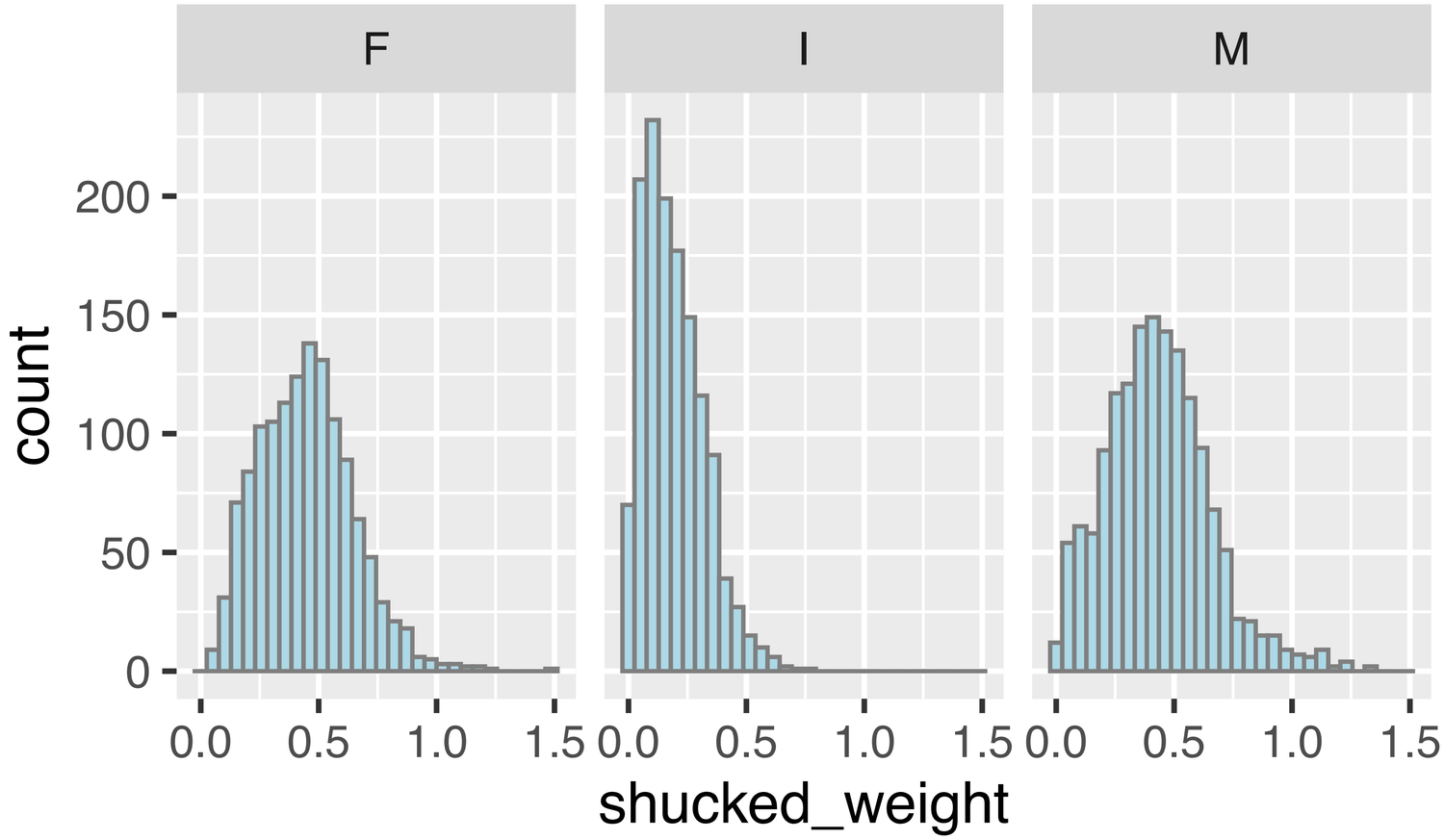
Single continuous variable



Single continuous variable faceted on factor (categorical) variable



Single continuous variable faceted on factor (categorical) variable

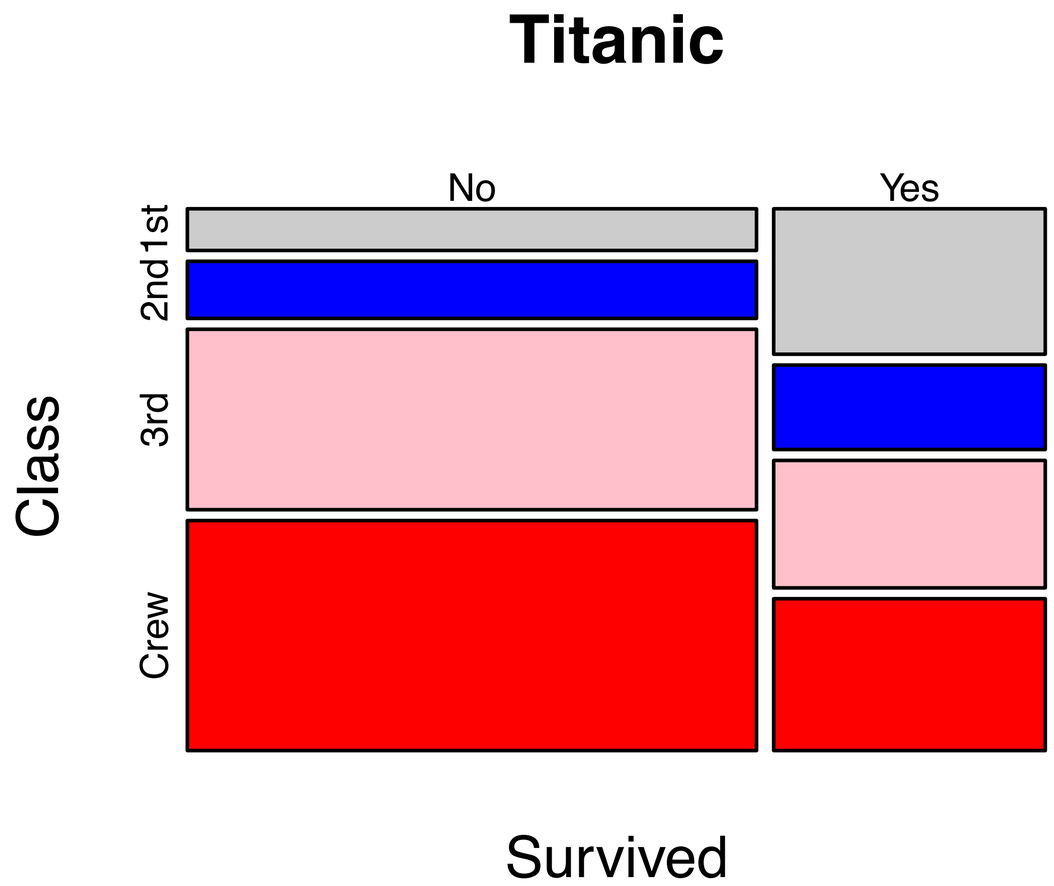


Multidimensional data

```
str(Titanic)
```

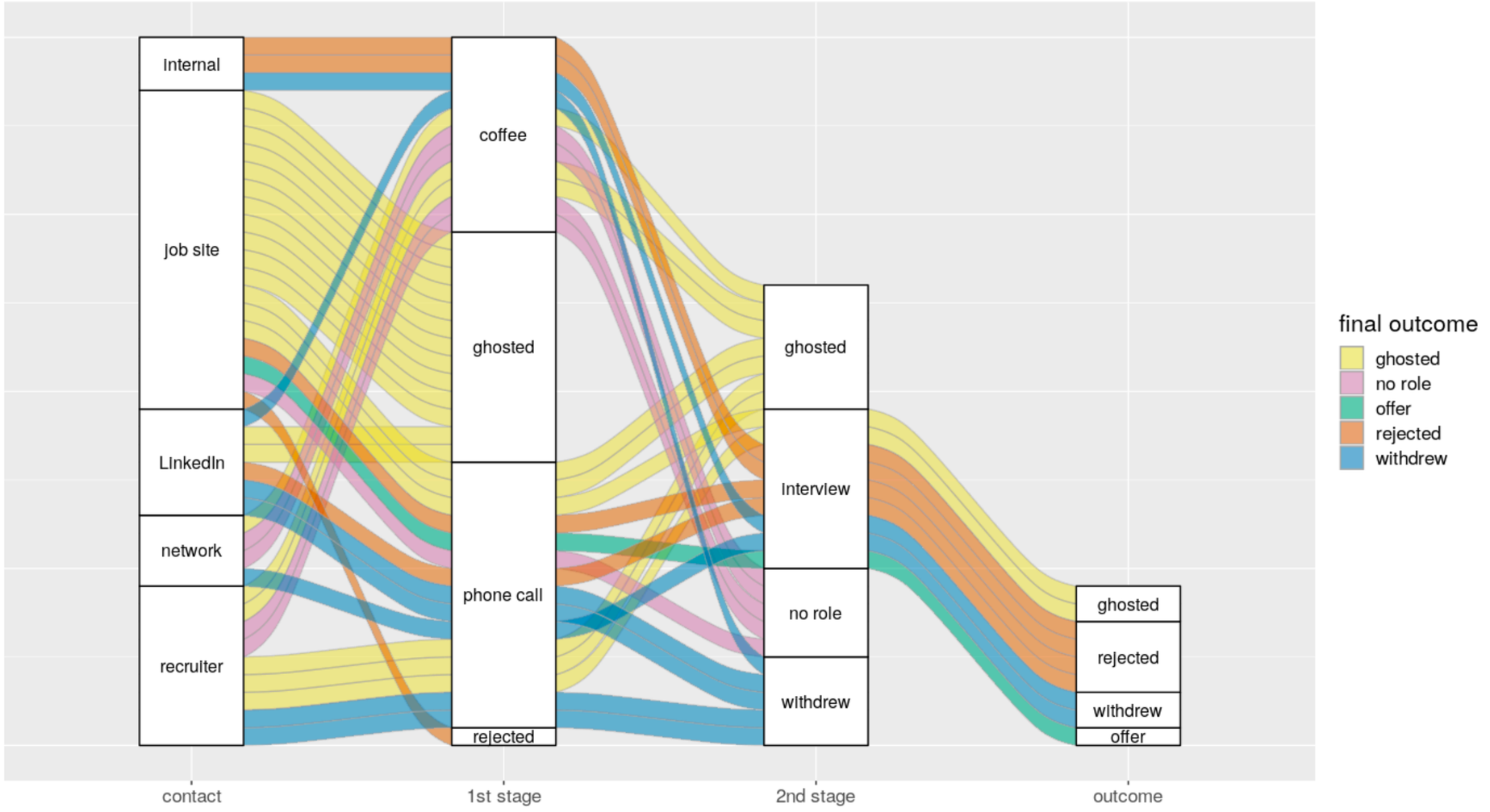
```
##  table [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 .
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

Two dimensional categorial data



mosaic plot

Multidimensional categorical data



alluvial diagram

https://mdneuzerling.com/post/my-data-science-job-hunt/

VISUALIZATION

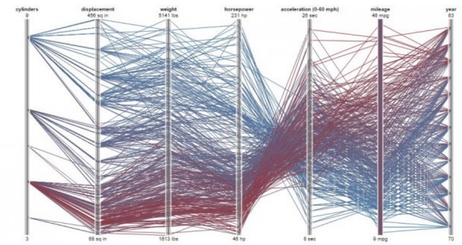
What is data visualization?

- relatively new field (but long history)
- multidisciplinary
- lack of consensus

Interdisciplinary influences

Statistics

Subject Expertise



Graphic Design

Psychology

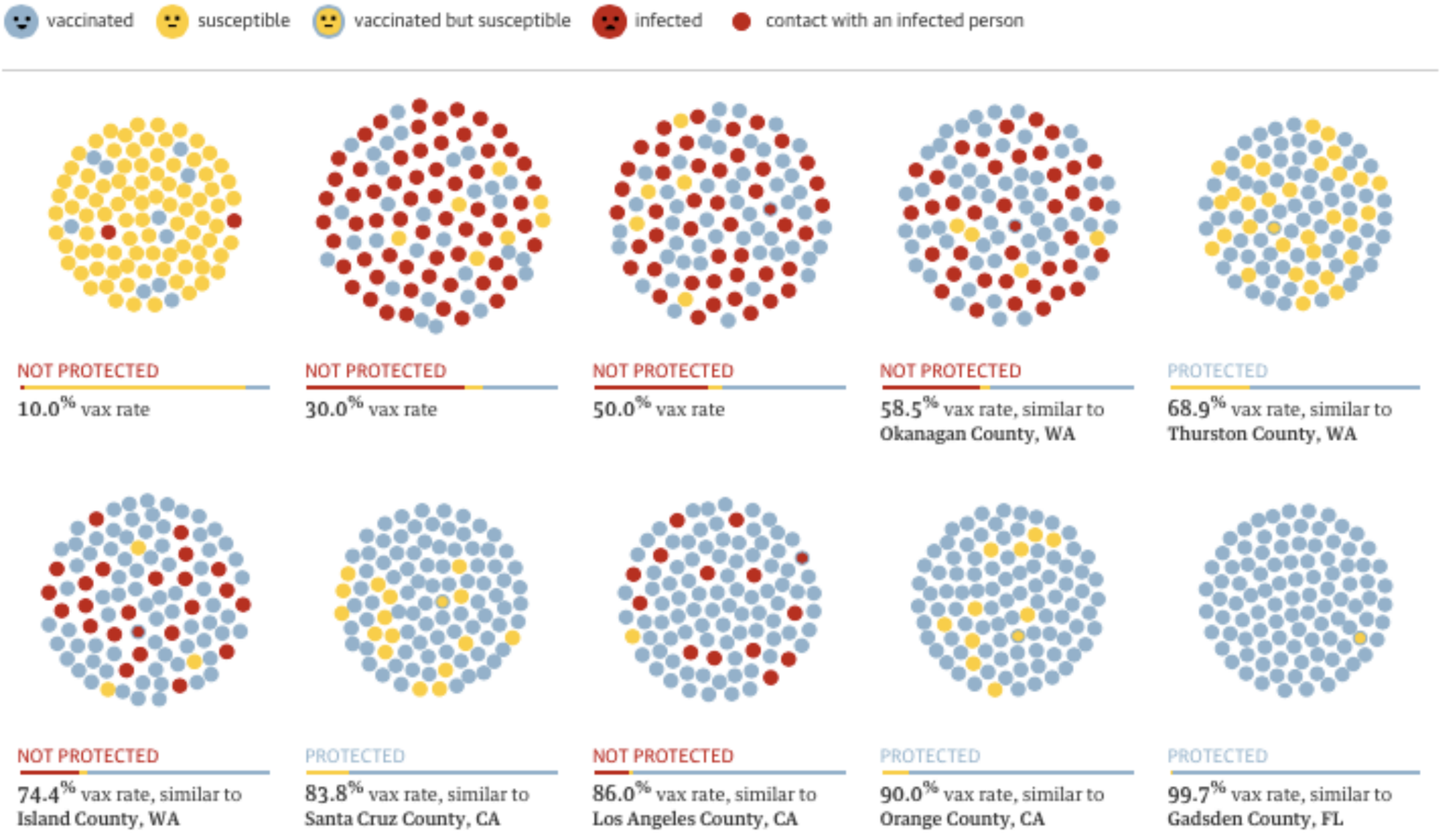
Computer Science

Journalism

Exploration vs. Visualization (Presentation)

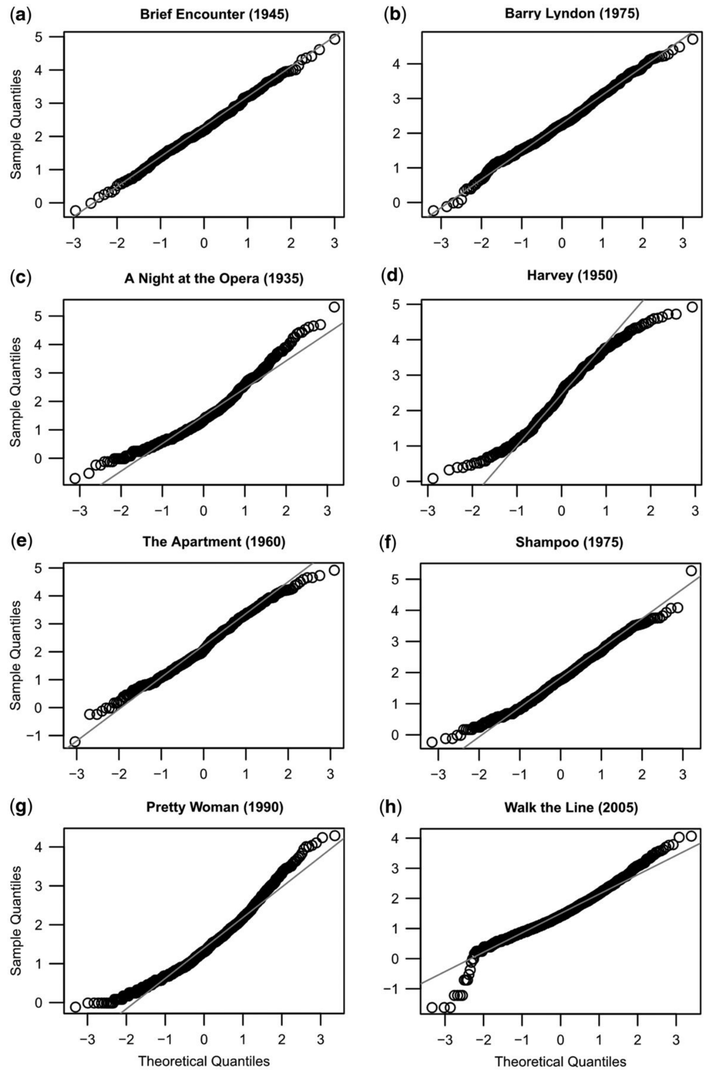
- Sometimes called "exploratory" vs. "explanatory"
- Not mutually exclusive
- Visualizations that offer insight are likely to be shared
- Still, focus is different when exploratory a dataset for the first time vs. presenting to an audience, particularly a less technical one

Watch how the measles outbreak spreads when kids get vaccinated – and when they don't



Run simulation again

Audience



Normal probability plots of log-transformed shot lengths for eight films

CRITIQUE

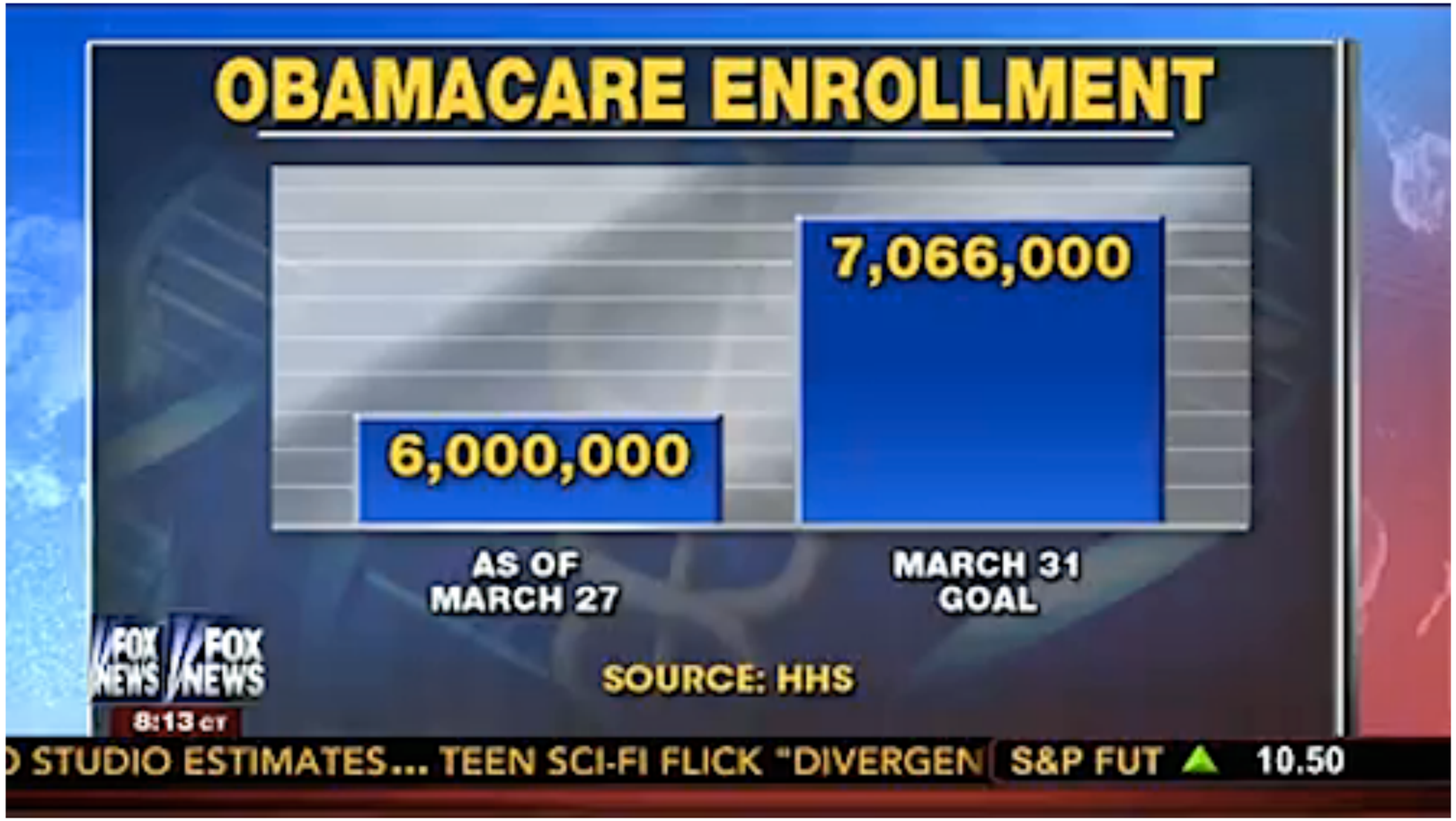
Growth of Data Visualization

"There is little evidence that the quality of the best graphics has improved over the last 100 years. I wonder if technology serves primarily as a **quantity**-multiplier, rather than a **quality**-multiplier." -Hadley Wickham

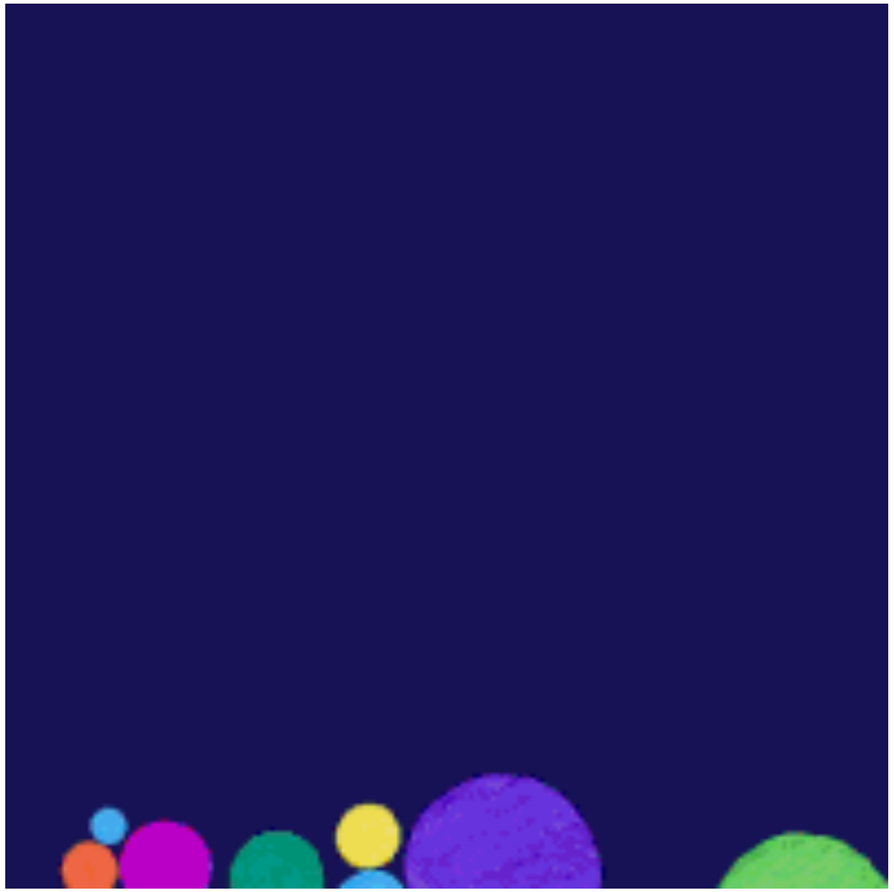
Evaluating Graphs

1. Wrong or misleading
2. Meaningless
3. Little added value
4. Good alternatives

Misleading Graph

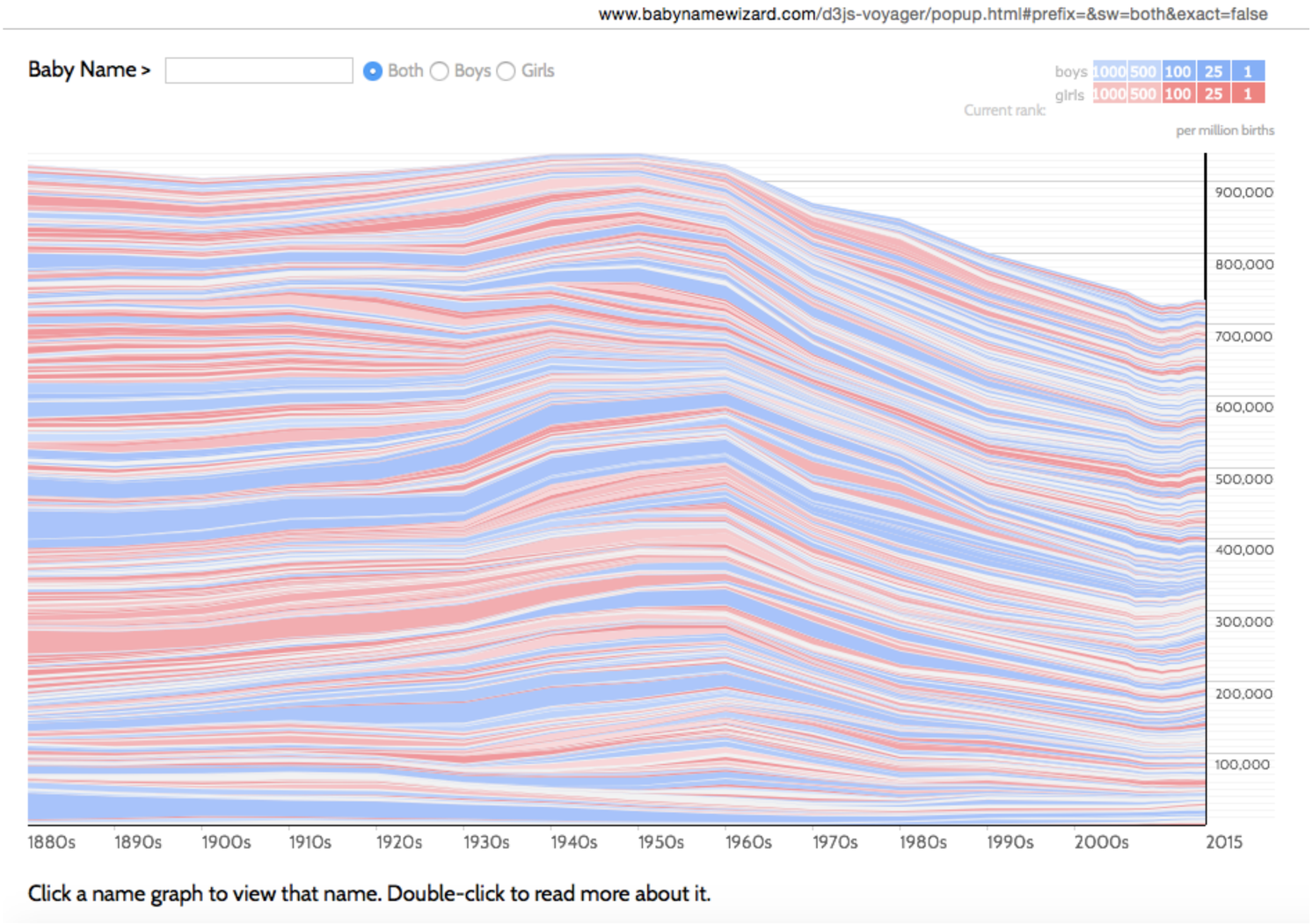


Data Visualization

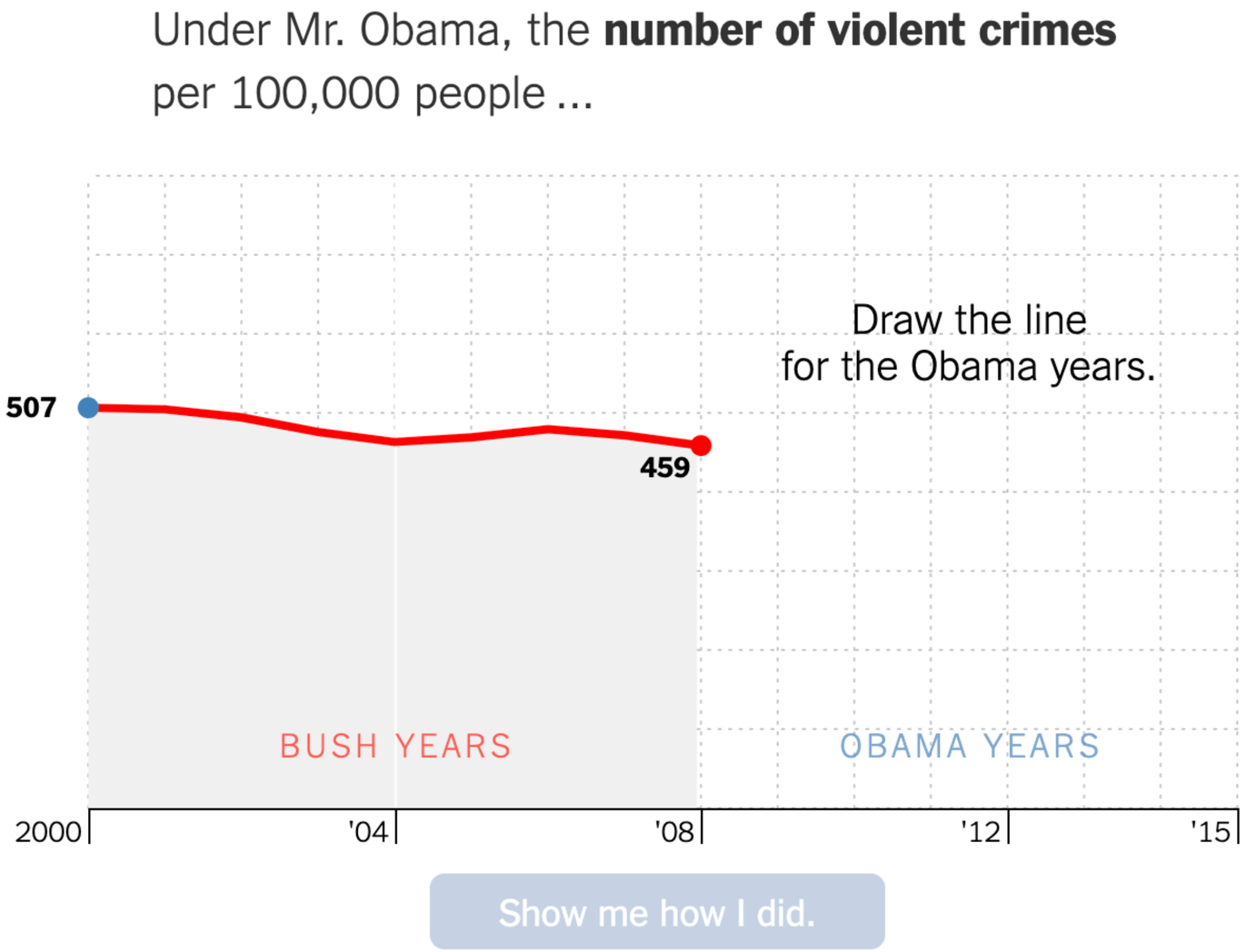


Registered Deaths

Baby Name Wizard



You-draw-it


Under Mr. Obama, the **number of violent crimes** per 100,000 people ...
Draw the line for the Obama years.
507
459
BUSH YEARS
OBAMA YEARS
2000
'04
'08
'12
'15
Show me how I did.

TOOLS

Tools

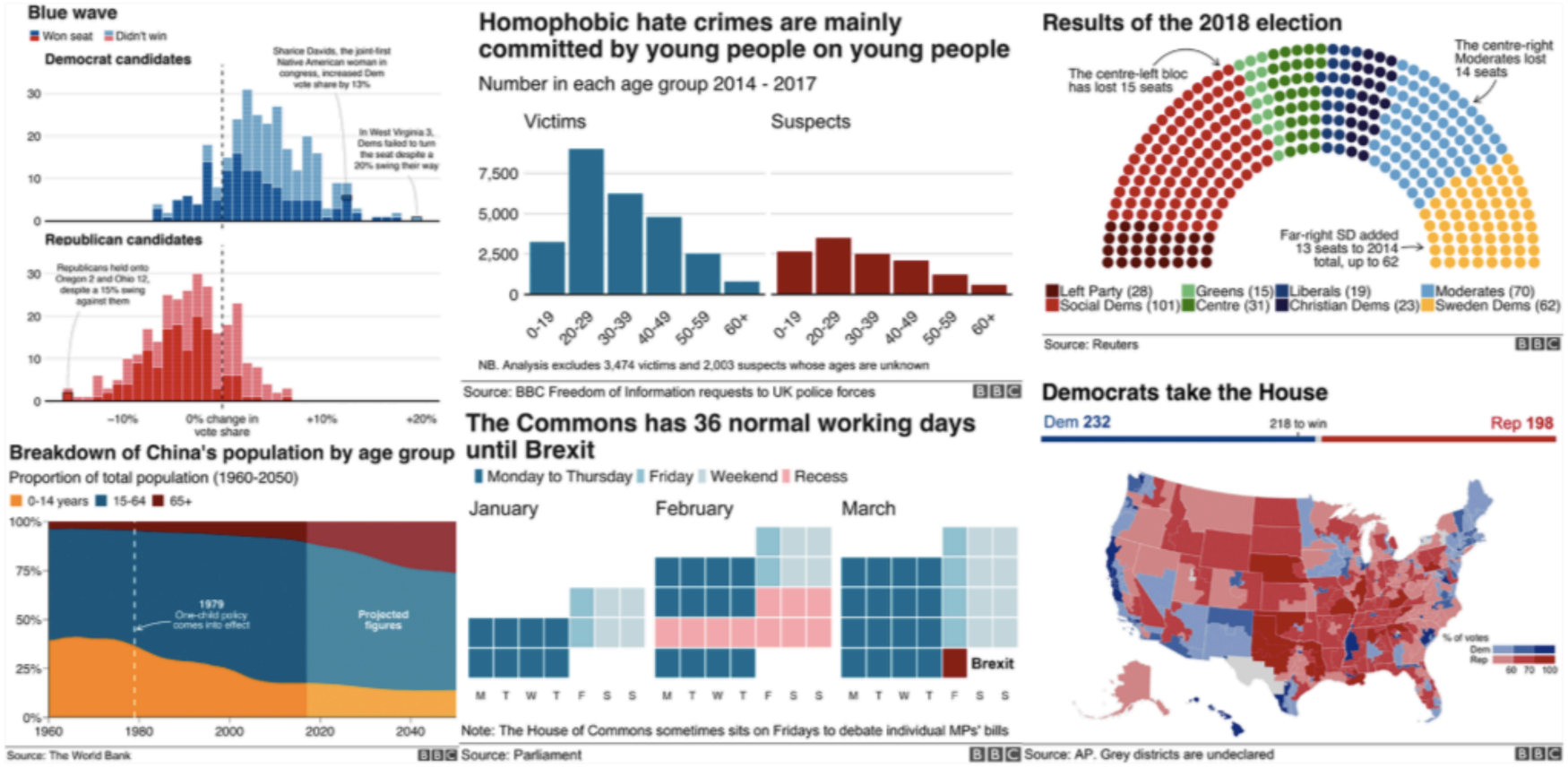
1. static: R (base graphics / ggplot2)
2. interactive: Plotly, htmlwidgets, Shiny, D3 + ...
3. version control: Git / GitHub
4. communication: Rmarkdown + ...

Why R?

- virtually unlimited graphical options
- opinionated graphics
- analytical tools (> 18,000 CRAN packages)
- community
- reproducibility
- ease of workflow: everything in one document
- free and open source

Why R? (Not only for EDA...)

How the BBC Visual and Data Journalism team works with graphics in R



https://medium.com/bbc-visual-and-data-journalism/how-the-bbc-visual-and-data-journalism-team-works-with-graphics-in-r-ed0b35693535

Why not R?

- learning curve
- lack of GUI for graphics
- interactive graphics are not native

Graphics in R

pie {graphics} R Documentation

Pie Charts

Description

Draw a pie chart.

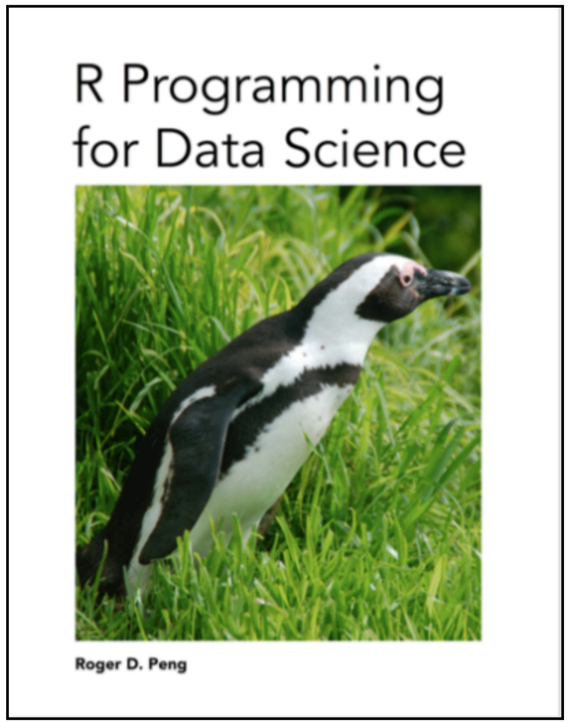
Usage

```
pie(x, labels = names(x), edges = 200, radius = 0.8,
    clockwise = FALSE, init.angle = if(clockwise) 90 else 0,
    density = NULL, angle = 45, col = NULL, border = NULL,
    lty = NULL, main = NULL, ...)
```

Arguments

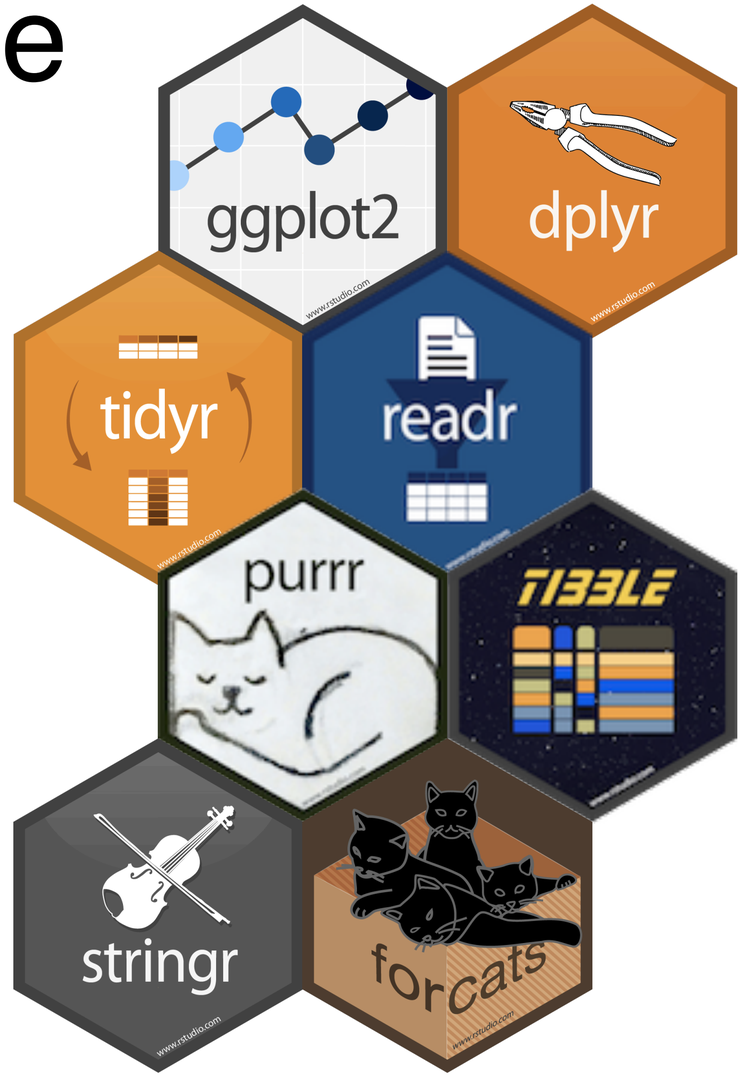
`x`	a vector of non-negative numerical quantities. The values in `x` are displayed as the areas of pie slices.
`labels`	one or more expressions or character strings giving names for the slices. Other objects are coerced by `as.graphicsAnnot`. For empty or `NA` (after coercion to character) labels, no label nor pointing line is drawn.
`edges`	the circular outline of the pie is approximated by a polygon with this many edges.
`radius`	the pie is drawn centered in a square box whose sides range from *-1* to *1*. If the character strings labeling the slices are long it may be necessary to use a smaller radius.
`clockwise`	logical indicating if slices are drawn clockwise or counter clockwise (i.e., mathematically positive direction), the latter is default.

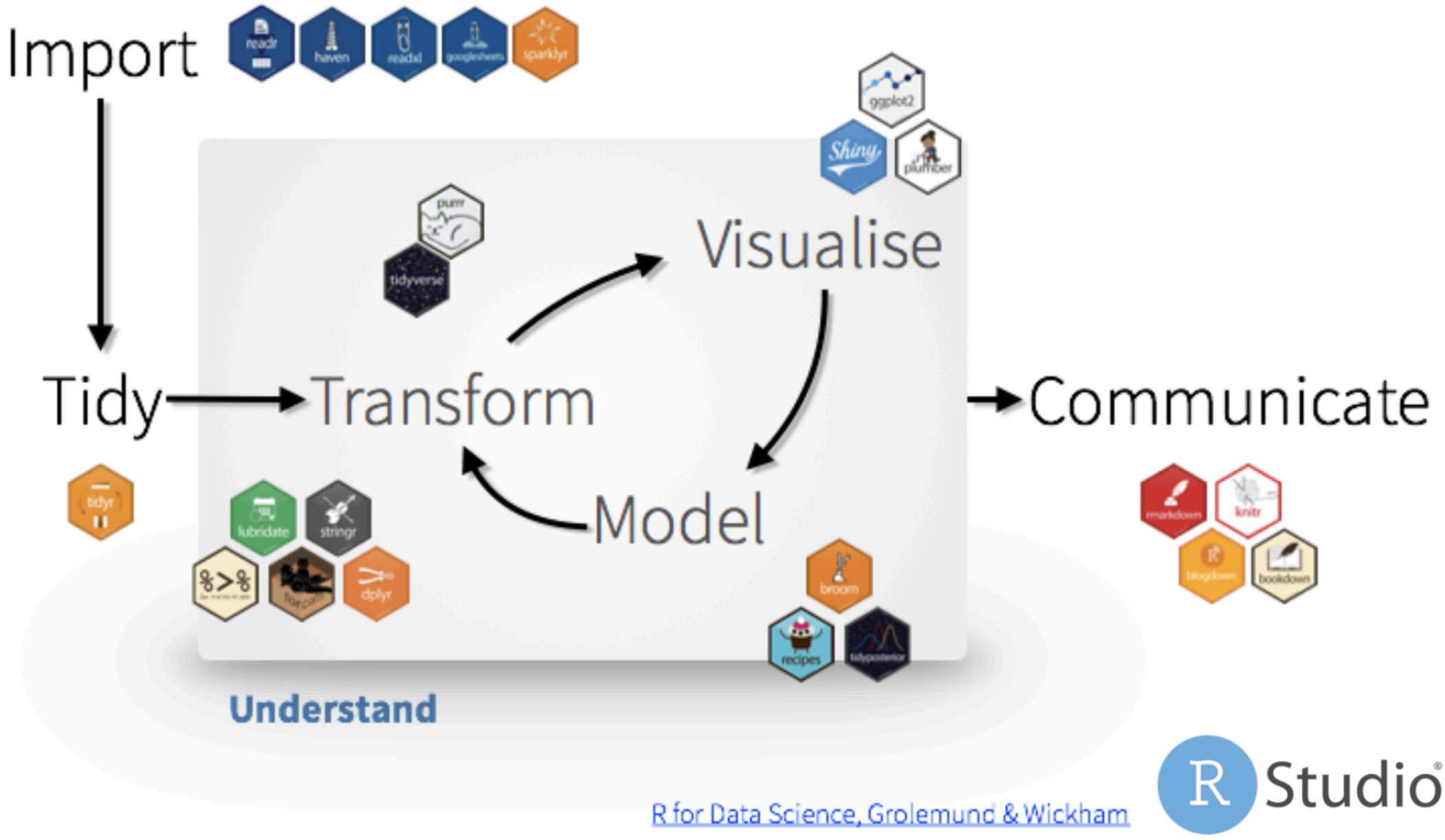
Learning R



https://leanpub.com/rprogramming

Base R vs. Tidyverse

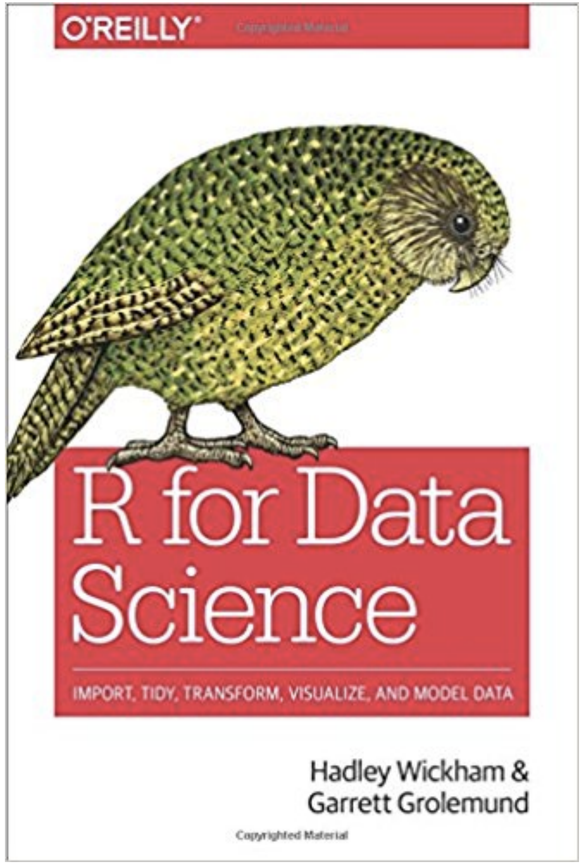




R for Data Science, Grolemund & Wickham

RStudio®

Tidyverse



r4ds.had.co.nz

Getting Started -- Top 10 R Essentials

1. Install R (*r4ds*) – You need to have this installed but you won't open the application since you'll be working in RStudio. If you already installed R, make sure you're current! The latest version of R (as of 2022-01-18) is R 4.1.2 "Bird Hippie" released on 2021/11/01.

2. Install RStudio (*r4ds*) – Download the free, Desktop version for your OS. Working in this IDE will make working in R much more enjoyable. As with R, stay current. RStudio is constantly adding new features. The latest version (as of 2022-01-18) is RStudio 2021.09.2+382 ("Ghost Orchid") release notes

(*r4ds* = R for Data Science by Garrett Grolemund and Hadley Wickham, free online)

Getting Started -- Top 10 R Essentials

3. Get comfortable with RStudio – In this chapter of Bruno Rodriguez's *Modern R with the Tidyverse*, you'll learn about panes, options, getting help, keyboard shortcuts, projects, add-ins, and packages. Be sure to try out:

- Do some math in the console
- Create an R Markdown file (`.Rmd`) and render it to `.html`
- Install some packages like `tidyverse` or `MASS`
- Another great option for learning the IDE: Watch Writing Code in RStudio (*RStudio webinar*)

Getting Started -- Top 10 R Essentials

4. Learn "R Nuts and Bolts" – Roger Peng's chapter in *R Programming* will give you a solid foundation in the basic building blocks of R. It's worth making the investing in understanding how R objects work now so they don't cause you problems later. Focus on **vectors** and especially **data frames**; matrices and lists don't come up often in data visualization. Get familiar with R classes: **integer, numeric, character,** and **logical**. Understand how **factors** work; they are very important for graphing.

5. Tidy up (*r4ds*) – Install the tidyverse, and get familiar with what it is. *We will discuss differences between base R and the tidyverse in class.*

Getting Started -- Top 10 R Essentials

6. Learn ggplot2 basics (*r4ds*) – In class we will study the grammar of graphics on which **ggplot2** is based, but it will help to familiarize yourself with the syntax in advance. Avail yourself of the "Data Visualization with ggplot2" cheatsheet by clicking "Help" "Cheatsheets…" within RStudio.

7. Learn some RMarkdown – For this class you will write assignments in R Markdown (stored as `.Rmd` files) and then render them into pdfs for submission. You can jump right in and open a new R Markdown file (*File > New File > R Markdown…*), and leave the `Default Output Format` as `HTML`. You will get a R Markdown template you can tinker with. Click the "knit" button and see what happens. For more detail, watch the RStudio webinar Getting Started with R Markdown

Getting Started -- Top 10 R Essentials

8. Use RStudio projects (*r4ds*) – If you haven't already, drink the Kool-Aid. Make each problem set a separate project. You will never have to worry about `getwd()` or `setwd()` again because everything will just be in the right places. Or watch the webinar: "Projects in RStudio"

9. Learn the basic dplyr verbs for data manipulation (*r4ds*) – Concentrate on the main verbs:
`filter()` (rows), **`select()`** (columns), **`mutate()`**, **`arrange()`** (rows), **`group_by()`**, and **`summarize()`**. Learn the pipe **`%>%`** operator.

Getting Started -- Top 10 R Essentials

10. Know how to tidy your data – The `pivot_longer()` function from the **tidyr** package – successor to `gather()` – will help you get your data in the right form for plotting. More on this in class. Check out these super cool animations, which follow a data frame as it is transformed by `tidyr` functions.

General advice: don't get caught up in the details. Keep a list of questions and move on.

Sources

"John Snow, Cholera Map" https://www1.udel.edu/johnmack/frec682/cholera/

"Data Science Process" diagram: Hadley Wickham and Garrett Grolemund, R for Data Science, 1.1 r4ds.had.co.nz/introduction.html

"Growth of Data Visualization": Hadley Wickham, 2013, "Graphical Criticism: Some Historical Notes", p. 43 www.tandfonline.com/doi/full/10.1080/10618600.2012.761140

"Perception Studies", "Cleveland Dot Plot": Naomi Robbins, 2013, *Creating More Effective Graphs*, Ch. 1., Ch. 3

"Wrong / Misleading Graphs" http://www.mediaite.com/tv/fox-news-airs-seriously-misleading-obamacare-graphic/

"Florence Nightingale's Coxcomb Diagram, 1858" https://understandinguncertainty.org/coxcombs

"Nightingale's Data, Redrawn" Andrew Gelman and Antony Unwin, 2012. "Infovis and Statistical Graphics: Different Goals, Different Looks" http://www.stat.columbia.edu/~gelman/research/published/vis14.pdf