



Exploratory Data Analysis

Visualizing your data

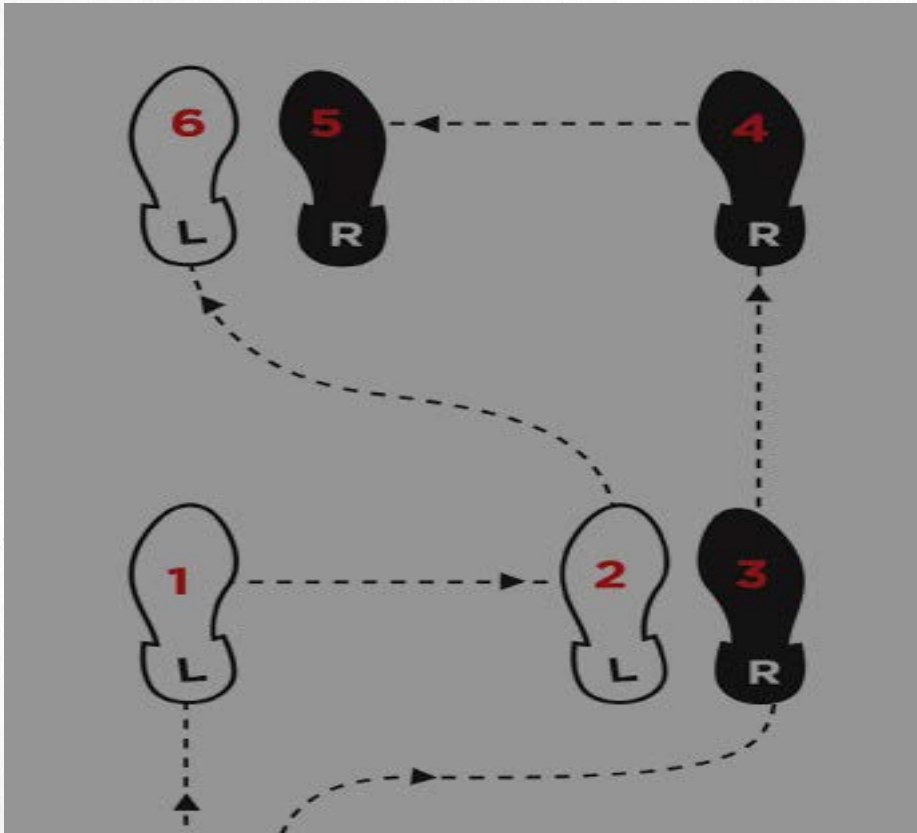


Data After Dark
OHSU BD2K Data Science Workshop

hannon McWeeney, PhD

14th January 2016

Exploratory Data Analysis (EDA)



1st step in
a 2-step process

Main Objectives

- ASSESS Assumptions
- SUPPORT Selection
- PROVIDE Basis

EDA **Features**



- Examines distributions + relationships
- Utilizes visualization + numerical summaries

EDA **FEATURES**



- Examines distributions + relationships
- Utilizes visualization + numerical summaries

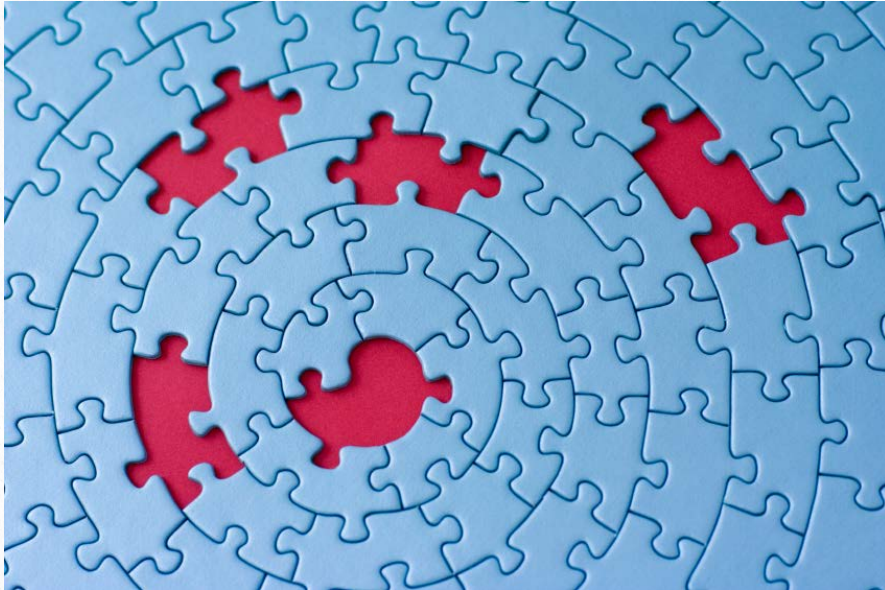
Context is **Key**

Context.

Everything is done in
framework of the
analysis plan!



Starting Point



Data File Assessment:

#variables

#subjects

Range

% Missing

Assessing the File

```
> summary(dat)
DLBCL.sample..LYM.number.      Analysis.Set Follow.up..years. Status.at.follow.up Subgroup
Min.   : 1.00      Training :160   Min.   : 0.000   Alive:102      ABC      : 73
1st Qu.: 91.75      Validation: 80   1st Qu.: 0.900   Dead :138      GCB      :115
Median :177.50      Median : 2.800   Mean    : 4.411   Type III: 52
Mean   :190.29      3rd Qu.: 7.100
3rd Qu.:284.25      Max.    :21.800
Max.    :439.00

  IPI.Group  Germinal.center.B.cell.signature Lymph.node.signature Proliferation.signature
High   : 32   Min.    :-2.61000             Min.    :-2.6500             Min.    :-1.700000
Low    : 82   1st Qu.: -0.91000             1st Qu.: -0.8675             1st Qu.: -0.410000
Medium :108   Median : -0.16000             Median : 0.0600             Median : -0.010000
missing: 1   Mean    :-0.03062             Mean    : 0.0065             Mean    : 0.005958
NA's   : 17   3rd Qu.: 0.86000             3rd Qu.: 0.8675             3rd Qu.: 0.412500
Max.    : 2.48000             Max.    : 2.9800             Max.    : 2.180000

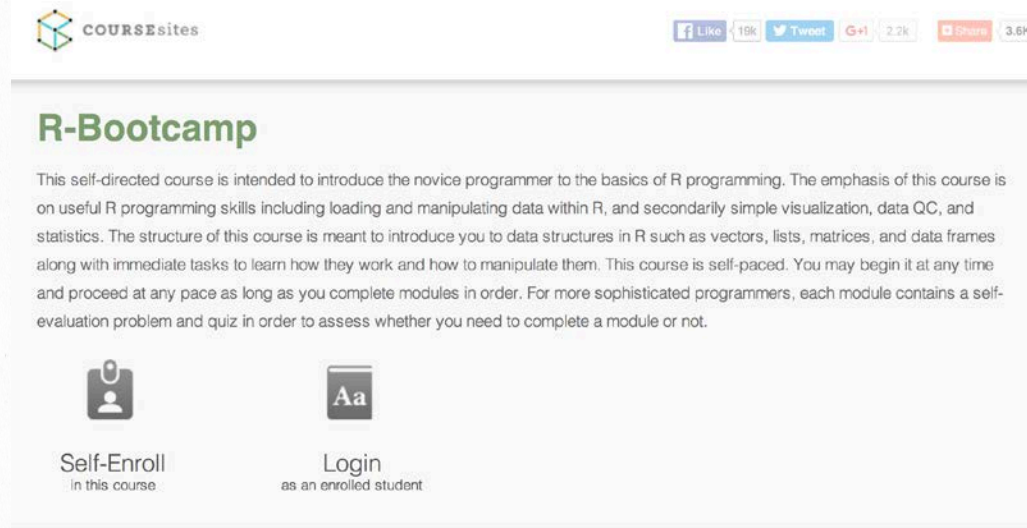
      BMP6      MHC.class.II.signature Outcome.predictor.score
Min.   :-1.87000   Min.   :-3.020000   Min.   :-1.700000
1st Qu.: -0.65250   1st Qu.: -0.537500   1st Qu.: -0.537500
Median : -0.13500   Median : 0.125000   Median : -0.085000
Mean   : -0.04362   Mean   : -0.006083   Mean   : -0.003208
3rd Qu.: 0.49250   3rd Qu.: 0.680000   3rd Qu.: 0.522500
Max.    : 2.69000   Max.    : 1.890000   Max.    : 2.360000
```

R Commands:
Summary()
Dim()



OHSU Resources

- **R Boot-camp:** Created by Dr. Ted Laderas, Division of Bioinformatics and Computational Biology , Department of Medical Informatics and Clinical Epidemiology



The screenshot shows the 'R-Bootcamp' course page on the 'coursesites' platform. At the top, there is a 'COURSESites' logo and social media sharing buttons for Facebook (Like 10k), Twitter (Tweet), Google+ (2.2k), and YouTube (Share 3.6K). The course title 'R-Bootcamp' is displayed in green. Below the title, a paragraph describes the course as a self-directed introduction to R programming for novices, covering data loading, manipulation, and visualization. It mentions that the course is self-paced and includes evaluation problems and quizzes. At the bottom, there are two buttons: 'Self-Enroll in this course' with a person icon and 'Login as an enrolled student' with a book icon.

<https://www.coursesites.com/s/ Rbootcamp>
(Coursesites registration required)



Benefits of a Data dictionary*

- Improved data quality
- Improved trust in data integrity
- Improved documentation and control
- Reduced data redundancy
- Reuse of data
- Consistency in data use
- Easier data analysis
- Improved decision making based on better data
- Simpler programming
- Enforcement of standards

*From Ahima.org



First Example

Data Field	Name	Definition	Data Type	Format	Field Size	Values	Source System	Date First Entered	Why Item Is Included
Admission Date	ADMIT_DATE	The date the patient is admitted to the facility as an inpatient	date	mmddyyyy	8	Admission date cannot precede birth date or 2007 No hyphens or slashes	Patient Census	2/23/2008	Allows analysis of patients and services within a specific period that can be compared with other periods or trended
Census	CENSUS	The number of inpatients present in the facility at any given time	numeric	x to xx	3	Any whole number from 0 to 999	Patient Census	2/23/2008	Provides analysis of budget variances, aids future budgetary decisions, and allows quicker response to negative trends
Ethnicity	PT_ETHNIC	Patient's ethnicity Must be reported according to official Office of Management and Budget categories	alphanumeric	Ex; letter must be uppercase	2	E1 = Hispanic or Latino Ethnicity E2 = Non-Hispanic or Latino Ethnicity	Patient Census; Practice Management	2/23/2008	Patient demographics aid marketing and planning future budgets and services
Infant Patient	INFANT_PT	A patient who has not reached 1 year of age at the time of discharge	alphanumeric	Age in months = xD to xxD OR xM to xxM	3	Must be > 0 AND < 1 year	Patient Census; Practice Mgt	2/23/2008	Patient age affects types of services required and payer sources

Source: AHIMA.ORG



2nd Example

CDE Public Id	Case Report Form Question Text	CDE Name XML Tag Name	Definition	Valid Values	Disease Type
2625735v1	Did the patient receive ATRA prior to registration	ATRA Agent Prior Clinical Trial Registration Administered Ind-2 xmlTag: atra_exposure	the yes/no indicator whether all-trans retinoic acid, a naturally-occurring acid of retinol, was administered prior to registration or enrollement in a controlled study performed in human subjects and intended to discover, evaluate, and/or verify safety, effectiveness, clinical and pharmacological effects, and adverse reactions.	No Yes	LAML
2003586v6	Route	Access Route of Administration Text Code xmlTag: route_of_administration	A text code or name to represent an access route for the administration of agents or substances.	Transmucosal Intraventricular Intravesical Urethral Intrauterine Oral and IV Otic ENDOTR Interstl Unknown ID IV IVI CIV IA SC IT IP IH IHl PO RT PR Transdermal Sublingual IM NASAL SWSP SWSW TOP NG Inhalatn G Tube Oph Each Oph Left Oph Rt INTUM Vaginal	HNSC BLCA BRCA CESC COAD GBM LUSC KICH KIRC KIRP LAML SKCM LGG LIHC LUAD SARC OV THCA PRAD READ UCEC STAD PAAD
3121502v1	Cytogenetic Risk Group (CALGB criteria)	Acute Myeloid Leukemia CALGB Cytogenetics Risk Category xmlTag: acute_myeloid_leukemia_calgb_cytogenetics_risk_category	Text term to classify the risk of developing acute myelogenous leukemia (AML) based on cytogenetic testing.	N/A - Remission Poor Intermediate/Normal Favorable	LAML

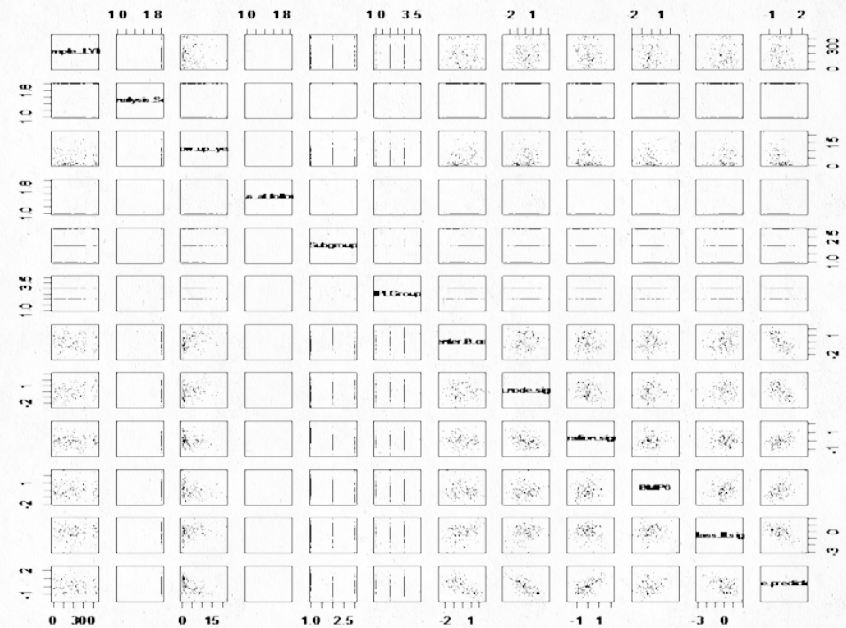
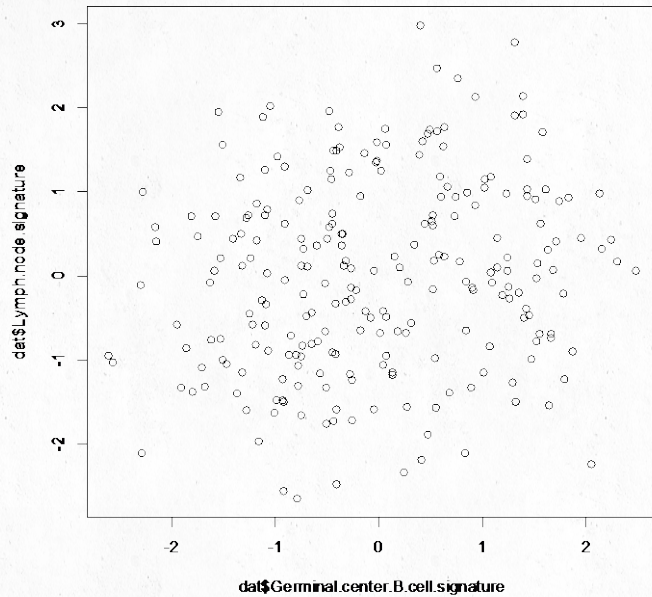
<https://tcga-data.nci.nih.gov/docs/dictionary/>



DISPLAYING Data: GRAPHICS

- Provide visual information
- Examine relationships; distribution

Assessing: Relationships



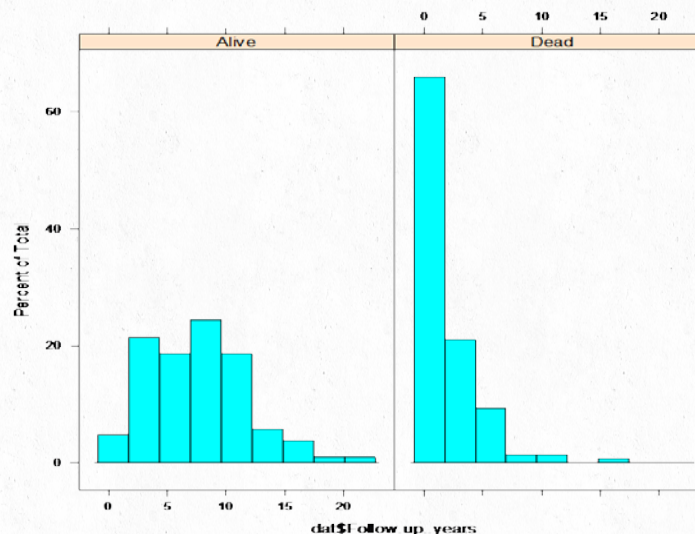
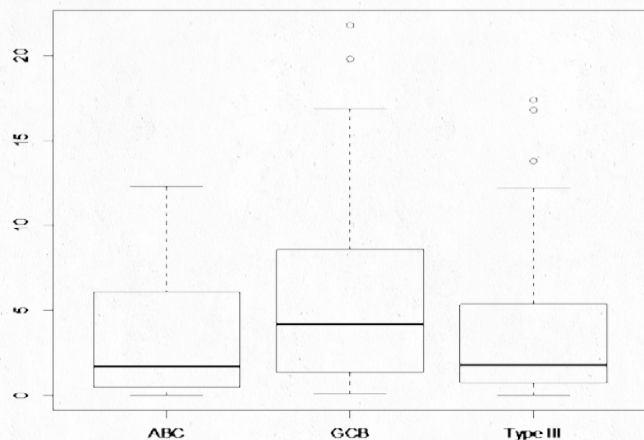
R Commands:

`Plot()`

`Cor()`



Assessing: Distributions



R Commands:

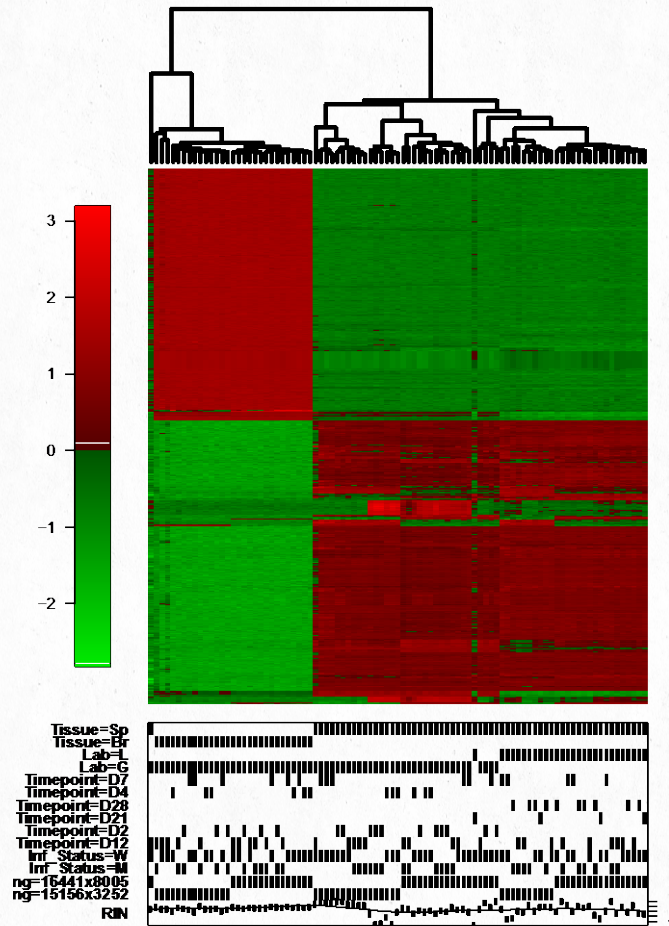
Hist() (also histogram() in lattice library)

boxplot()

Displaying Data: **TABLES**

- Layout
- “Stand alone”
- Comparison of interest/focus

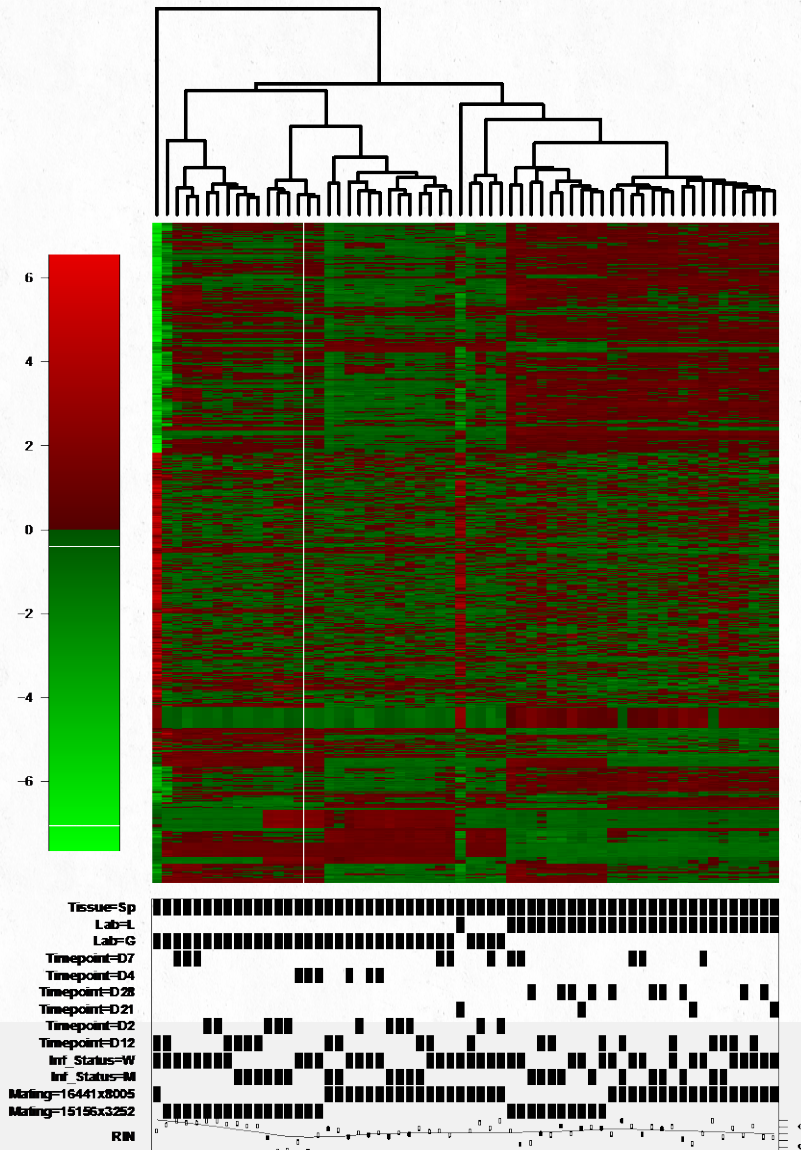
Sample Inspection



R
Commands:
Heatmap()



Sample Inspection



Display data accurately and clearly

Good and bad data visualization

INDUSTRY



EMPLOYMENT

- Specialized Design Services
- Advertising, Public Relations & Related Services
- Newspaper, Periodical, Book, and Directory Publishers
- Printing and Related Support Activities
- Other Miscellaneous Manufacturing

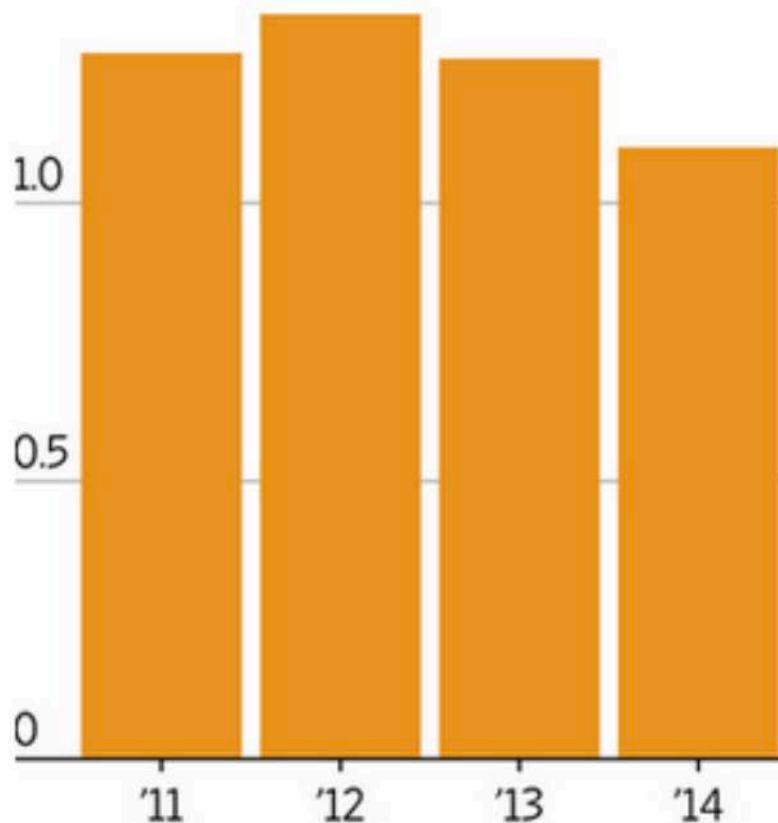
FUN FACTS

Musical Movements

Digital-music sales are falling as more people use streaming services like Spotify.

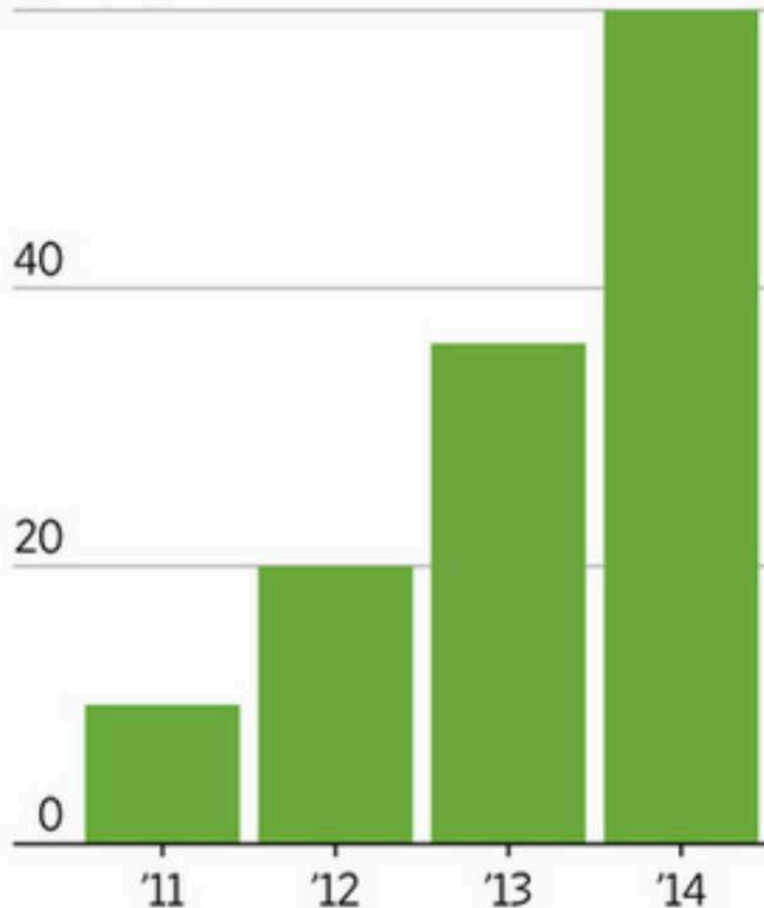
Digitally downloaded tracks, U.S.

1.5 billion



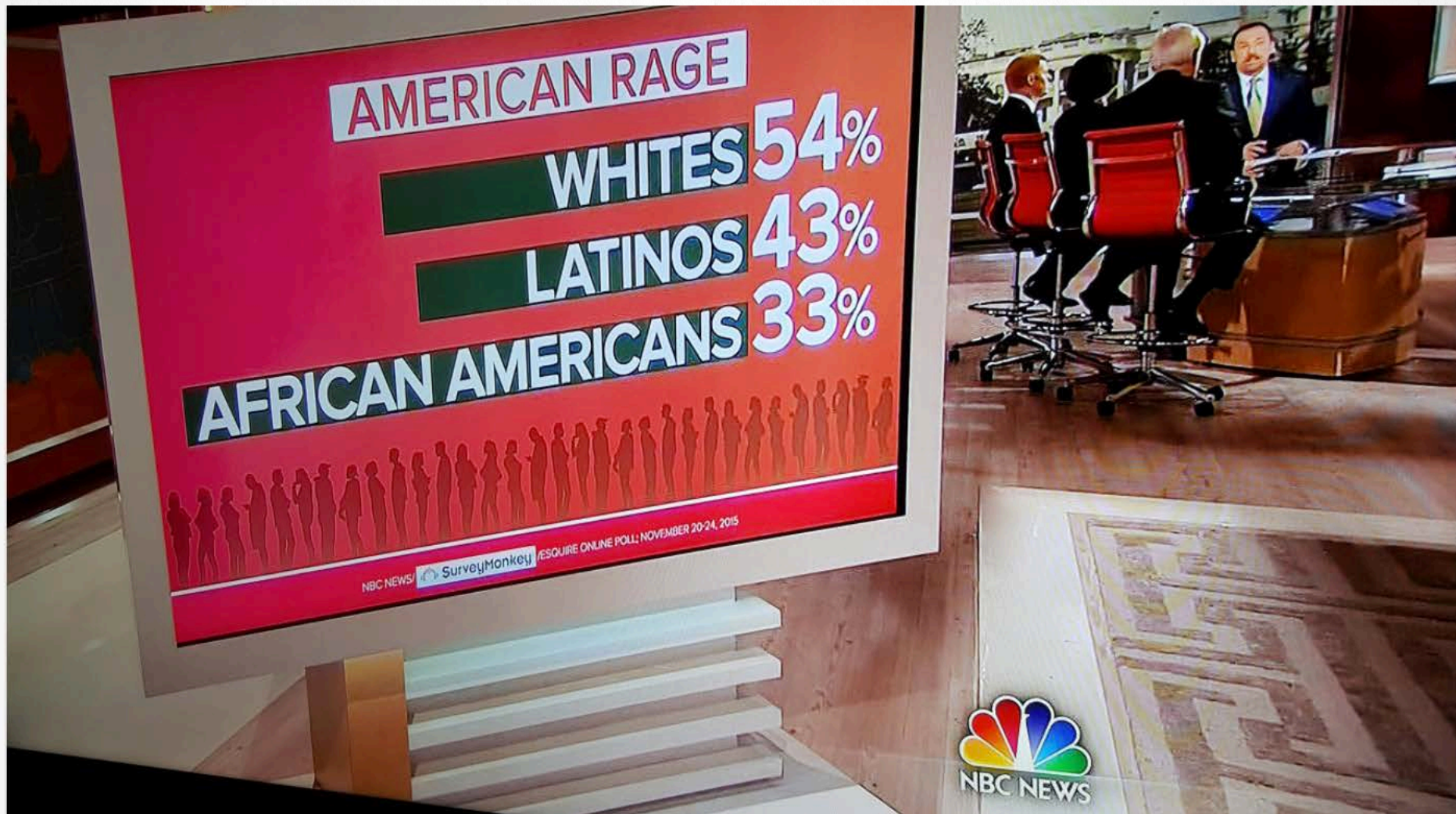
Monthly active Spotify users, global

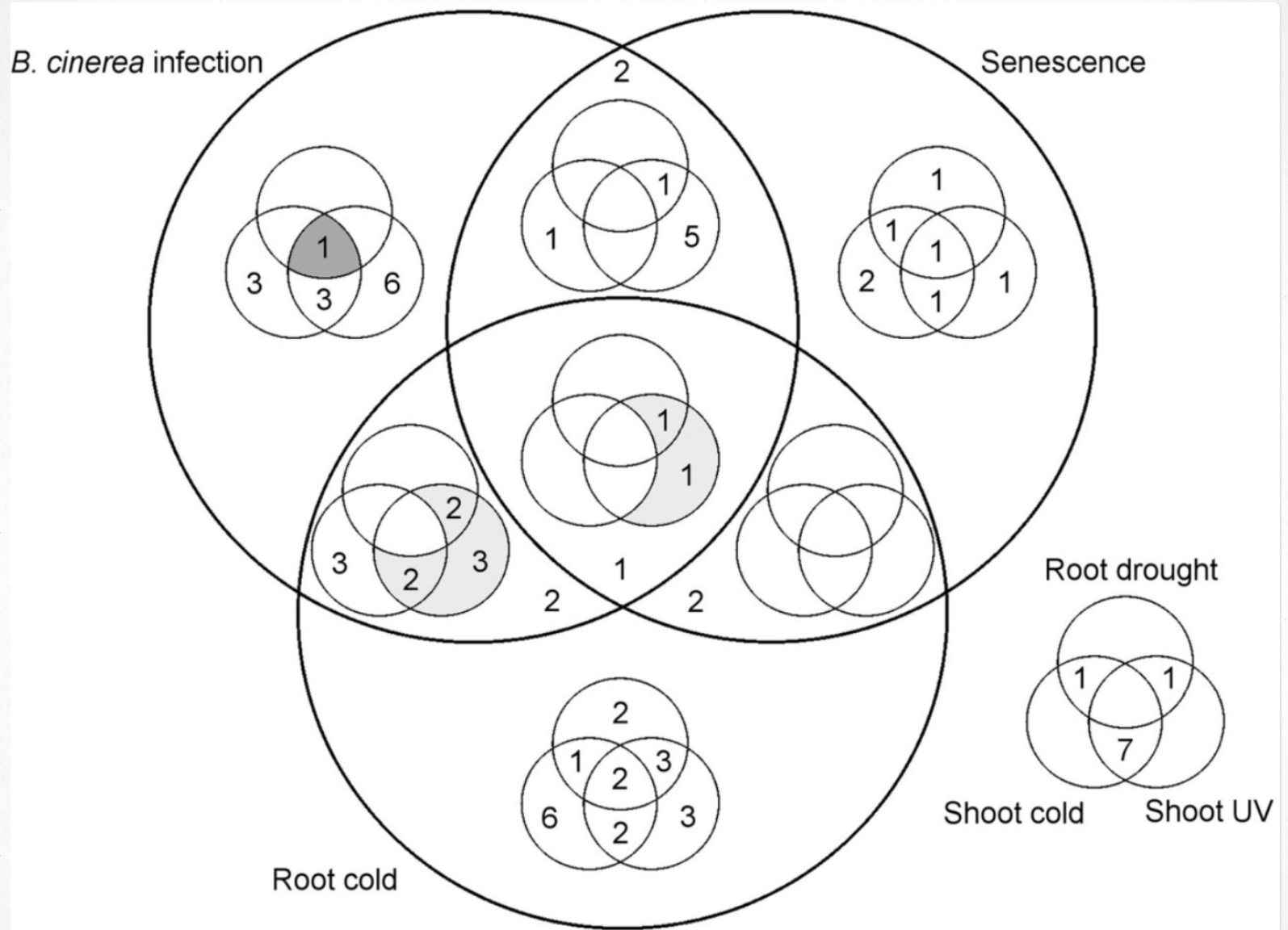
60 million



Note: Spotify usership data released at unequal intervals
Sources: Nielsen Music (digital downloads); Spotify (users)

THE WALL STREET JOURNAL.





May 3, 2008

SIGN IN TO E-MAIL OR SAVE THIS | FEEDBACK

All of Inflation's Little Parts

Each month, the Bureau of Labor Statistics gathers 84,000 prices in about 200 categories – like gasoline, bananas, dresses and garbage collection – to form the Consumer Price Index, one measure of inflation.

It's among the statistics that the Federal Reserve considered when it cut interest rates on Wednesday. The categories are weighted according to an estimate of what the average American spends, as shown below.

An Average Consumer's Spending

Each shape below represents how much the average American spends in different categories. Larger shapes make up a larger part of spending.

Color shows change in prices from March 2007 to March 2008



ZOOM IN

ZOOM OUT

Food and beverages 15%

The high price of oil is a factor that has made food prices rise quickly.

Miscellaneous 3%

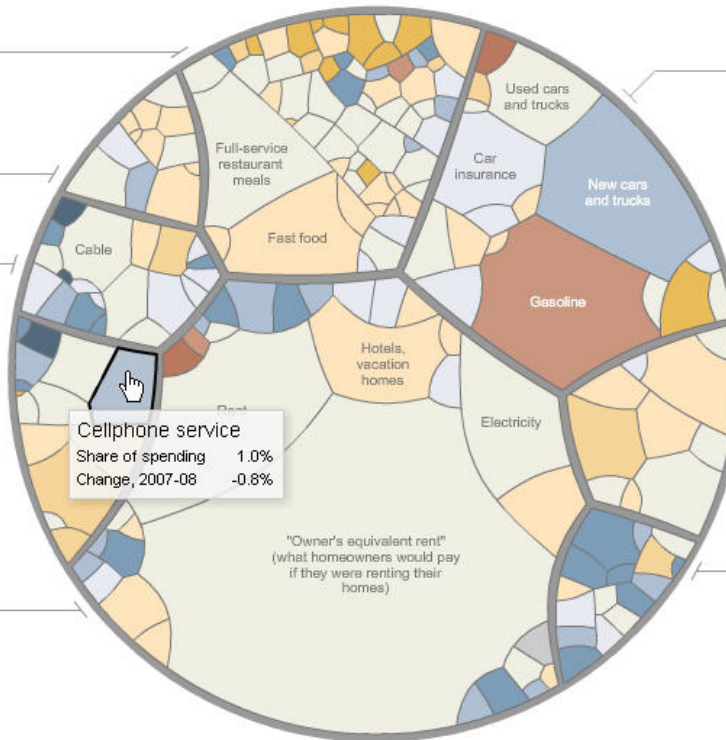
Recreation 6%

Education/Communication 6%

Cellphones were added to the index in 1997. Because the Consumer Price Index can be slow to add new goods, which are often cheaper, it may overstate parts of inflation.

Housing 42%

In the C.P.I., home ownership costs track rent prices more closely than housing prices. This means inflation may have been understated when home prices were rising faster than rents.



Transportation 18%

Gas is 5.2 percent of spending nationwide, but only 3.8 percent in the New York area.

Health care 6%

As a group, the elderly spend about twice as much of their budget on medical care.

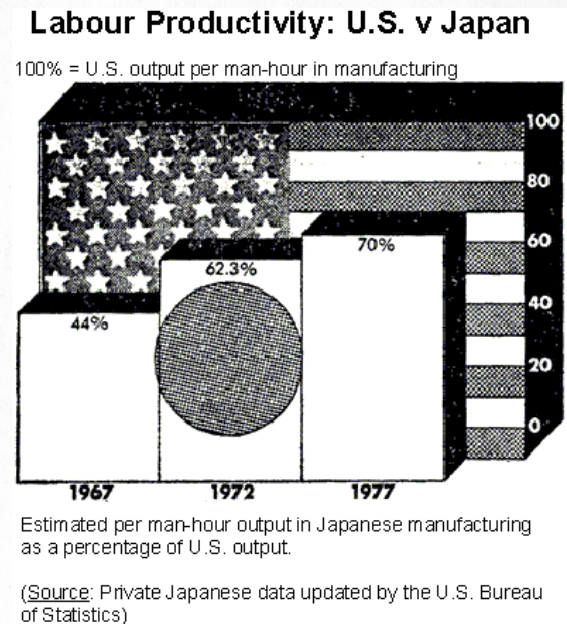
Apparel 4%

The ratio of spending on women's clothes to that on men's clothes is about 2 to 1.



Bad Data Viz

- Not informative
- Data is obscured (Tufte's "Chart junk"*)
- Pie charts (3d!!)
- Issues of scale

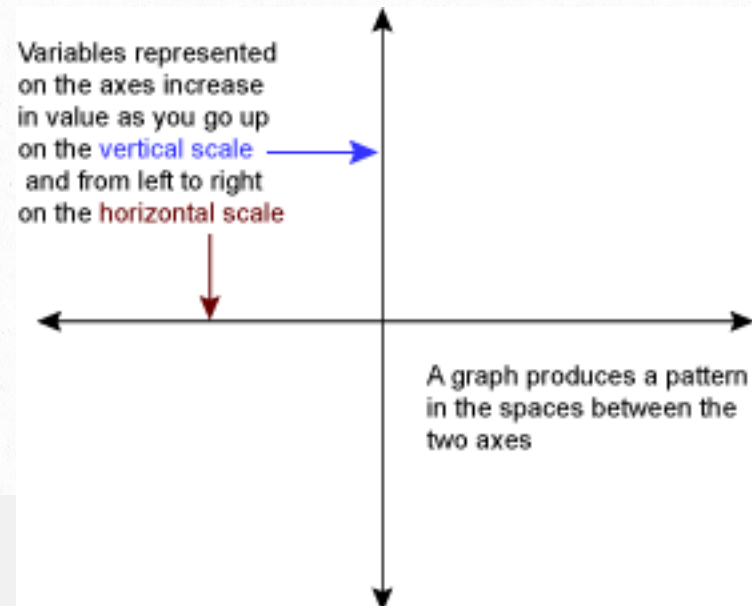


*Tufte, E. R. *The visual display of quantitative information*



Graphical Proficiency

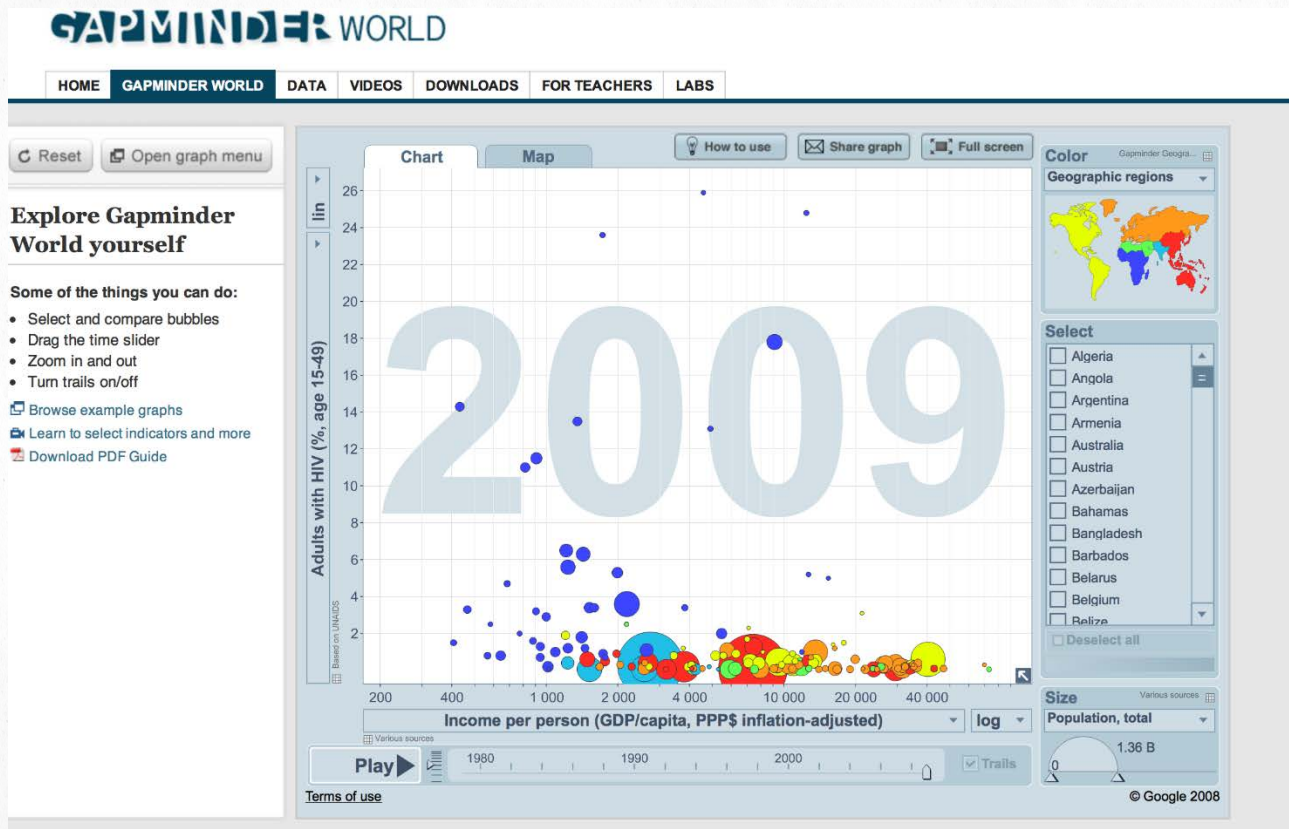
- WHAT IS THE STORY?
- WHAT DO YOU NEED TO KNOW TO INTERPRET IT?



Interactive Visualization

Examples & Tools you can use

INTERACTIVE GRAPHICS



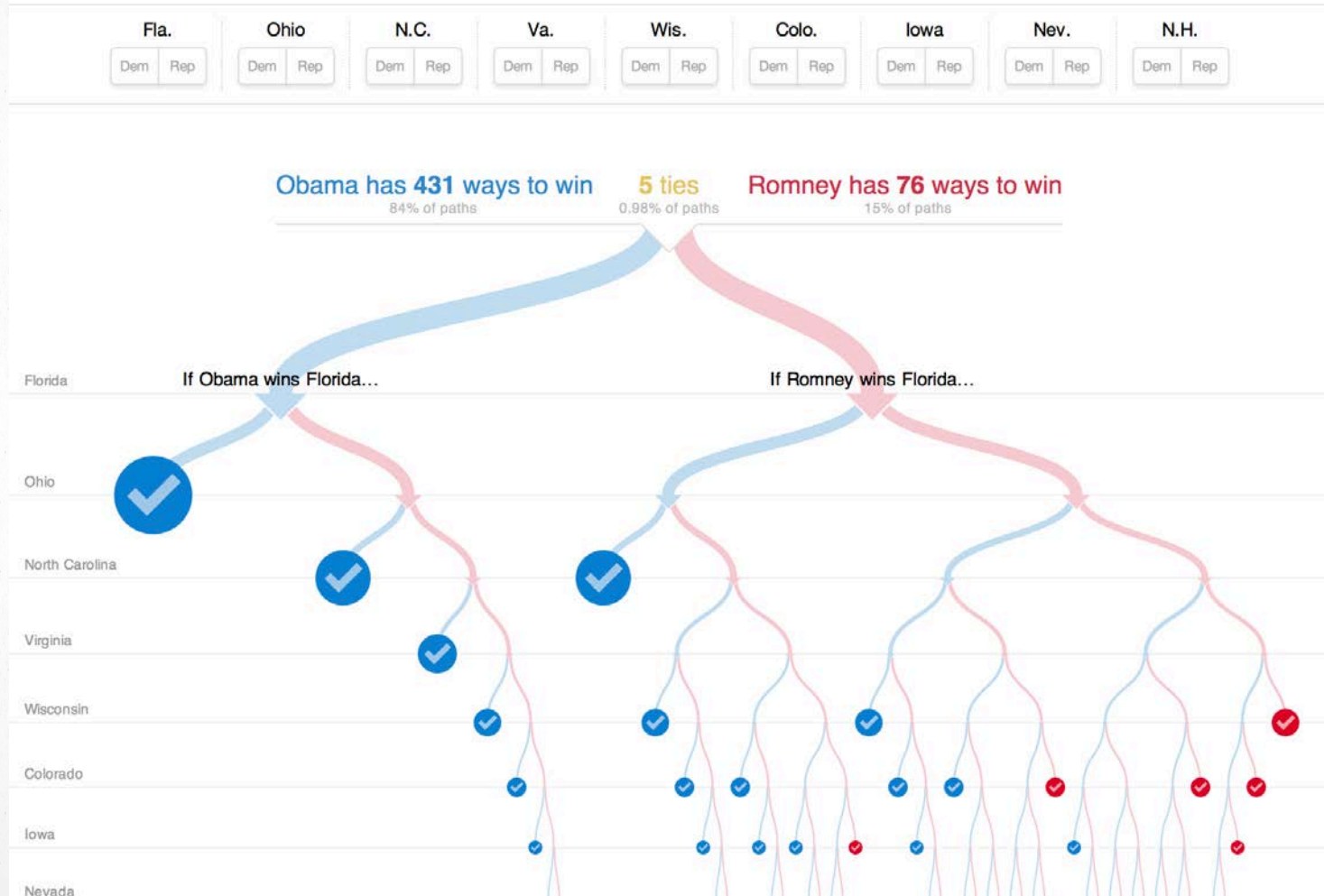
GAPMINDER.ORG




Interactive Data: Path Models


512 Paths to the White House


Select a winner in the most competitive states below to see all the paths to victory available for either candidate.



Google Charts

 Google Developers

Charts  Search

shannon.mcweeney@gmail.com
Sign out 

Products > Charts

Charts

Interactive charts for browsers and mobile devices.

HOME GUIDES REFERENCE SUPPORT

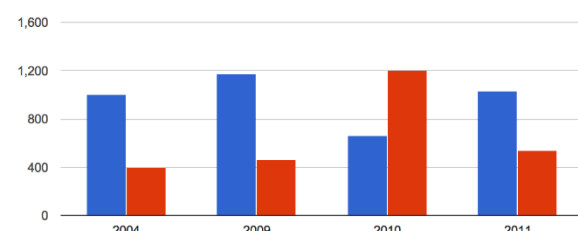
Display live data on your site

About Google chart tools







Google chart tools are powerful, simple to use, and free. Try out our rich gallery of interactive charts and data tools.

[GET STARTED](#) [CHART GALLERY](#)

Column Chart - [view source](#)



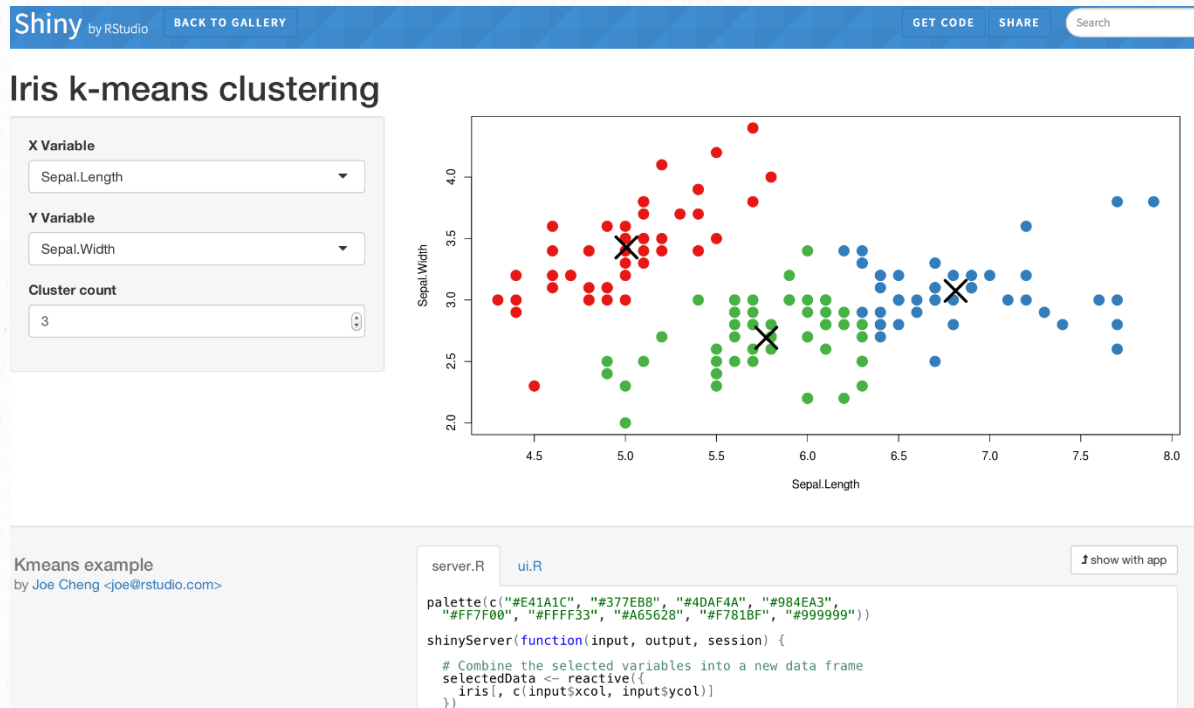
Year	Blue Series	Red Series
2004	1,000	400
2009	1,200	450
2010	650	1,200
2011	1,000	550

      [more](#)

<https://developers.google.com/chart/?csw=1>



Shiny (Rstudio)



<http://shiny.rstudio.com/>

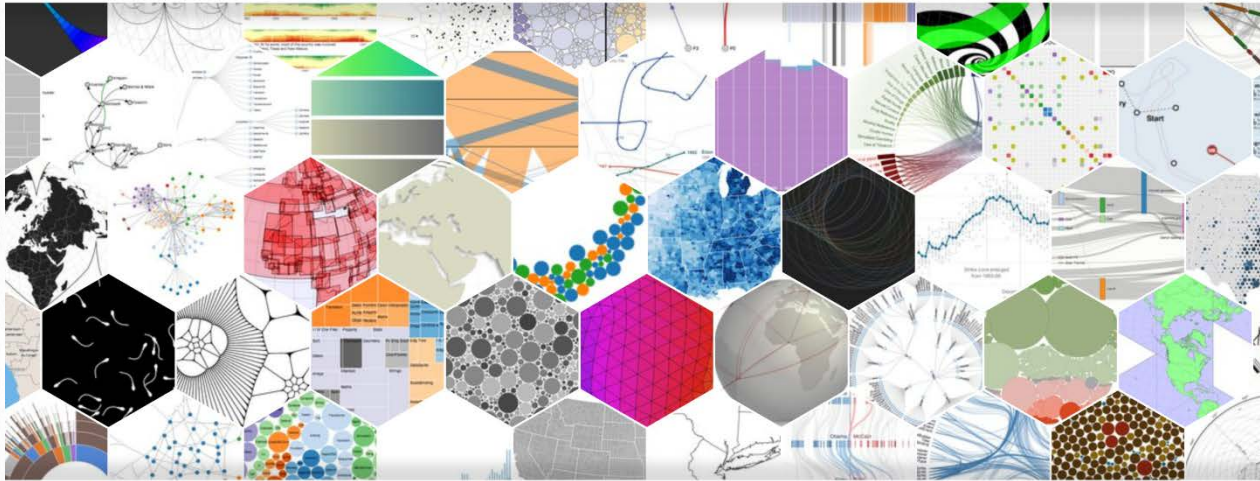


Data Driven (D3JS)

[Overview](#) [Examples](#) [Documentation](#) [Source](#)



Fork me on GitHub



D3.js is a JavaScript library for manipulating documents based on data. **D3** helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of

[See more examples.](#)

<http://d3js.org>



“If you don’t think you have a quality problem with your data, you haven’t looked at it”

Every data set has quirks.

5 Stages of Data Grief

- Denial
- Anger
- Bargaining
- Depression
- Acceptance (+ Hope!)



Visual **Points** to remember

- Software shouldn't dictate the Visual
- Tell a story
- Follow best practices (be mindful)