Data Mining and Analysis

Data Visualization: 2

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

Class Design Principles:

1. Don’t get too cute; focus on the data (focus on data variation, not design variation)

2. Have a good, explanatory, clear title

3. Data integrity; maintain scale, right axes (no assumptions: be clear with units, etc.)

4. Know your goal! Know your audience.

5. Balance high-level with detail

6. Consider and respect spatial logic

7. Don’t be overconfident

8. Color and design are important

Uh oh ...

The NSW Health system is…

* Nursing headcount figures at June includes non casual staff and 3rd schedule

NSW GOVERNMENT

NSW Ministry of Health March 2013

OPERATING REVENUES
$3,549,385
$4,520,362
$4,916,444
$6,814,503
$7,382,599
1970
1971
1972
1973
1974

NET INCOME (LOSS)
$(11,014)
$397,747
$521,943
$1,647,001
$1,435,102
1970
1971
1972
1973
1974

EXPLORATION & DEVELOPMENT EXPENDITURES
$351,341
$85,149
$75,243
$329,421
$1,226,007
1970
1971
1972
1973
1974

Commission Payments
To Travel Agents
In millions of dollars
$50
$57
$36
$70
$79
$50
$84
$102
$52
$100
$109
$64
'76
'77
First
Half
'78
DELTA
EASTERN
TWA
UNITED
AIRLINES

Comparative Annual Cost per Capita for care of Insane in Pittsburgh City Homes and Pennsylvania State Hospitals.

$147

South Mountain

$172

Pittsburgh

$198

Harrisburg

$213

Norristown

$214

Warren

Lie Factor

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

Los Angeles Times, A

Context is Essential

Graphics must not quote data out of context.

Nearly all the important questions are left unanswered by this display:

A few more data points add immensely to the account:

Tufte: Principles of Graphical Excellence

Substance, Statistics, Design

“Complex ideas communicated with clarity, precision, and efficiency”

Tufte’s principles for better viz?

- Above all else, show the data
- Maximize the data-ink ratio
  - Erase non-data-ink
  - Erase redundant data-ink
- Revise and edit

Maximize the Data-Ink Ratio

$$\text{Data-ink ratio} = \frac{\text{data-ink}}{\text{Total ink used to print graphic}}$$

= Proportion of a graphic's ink devoted to the non-redundant display of data-information.

= 1.0 – proportion of graphic that can be erased without the loss of information

Maximize the data-ink ratio

Within reason

In essence, you should be able to argue for every pixel

Starting point:

- erase non-data ink
- erase redundant data-ink

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

15%
10%
5%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

Show data variation, not design variation

Avoid using ink for non-data items

Avoid redundancy

Clear and detailed labeling should be used to defeat graphical distortion

Revise and Edit

Makeovers

Chart 2 - Total Expenditures on Health as a Percentage Share of GDP, by OECD Country, 2004

Source: OECD Health Data 2007.

Note: For the United States the 2004 data reported here do not match the 2004 data point for the United States in Chart 1 since the OECD uses a slightly different definition of "total expenditures on health" than that used in the National Health Expenditure Accounts.

Expenditures on Health as Percentage of GDP
for OECD Countries, 2004
United States
Switzerland
France
Germany
Austria
Belgium
Iceland
Portugal
Canada
Norway
Greece
Australia
Netherlands
Denmark
Sweden
Italy
New Zealand
United Kingdom
Spain
Hungary
Japan
Luxembourg
Turkey
Ireland
Finalnd
Czech Republic
Slovak Republic
Mexico
Poland
Korea
6
8
10
12
14
Share of GDP (%)

http://www.storytellingwithdata.com/gallery/

US Prison Population by Primary Sentencing Reason

As of January 25, 2014

Other includes Banking and Insurance, Counterfeit, Embezzlement, Courts or Corrections, Continuing Criminal Enterprise, National Security and other miscellanous sentencing reasons, accounting for <1% of total each.

Data source: US Sentencing Commission

US Prison Population by Primary Sentencing Reason

As of January 25, 2014

More than half of the current US prison population were incarcerated due to **drug offenses**.

Follow this up with additional context on why this is interesting or noteworthy, call to action, etc.

Other includes Banking and Insurance, Counterfeit, Embezzlement, Courts or Corrections, Continuing Criminal Enterprise, National Security and other miscellanous sentencing reasons, accounting for <1% of total each.

Data source: US Sentencing Commission

Tech
Chart of the Day
Online Hotel Revenue vs Number Of Travel Agents
Online hotel revenue
$ billion
Number of travel agents in the United States
Thousand
160
140
120
100
80
60
40
20
0
130
120
110
100
90
80
70
60
50
2000
01
02
03
04
05
06
07
08
09
10
11
12
13
2014
BUSINESS INSIDER
Source: McKinsey, Phocuswright, BLS

Hotel bookings shift

Increase in **online hotel bookings** is accompanied by decrease in **travel agents**

Online hotel revenue: massive growth in past 15 years

Global Competitive Summary

Global Competitive Summary (1/1/2015-12/31/2015)	
	Global
Total Projects	340
In Scope	264
Proposals	92
Wins	37

Timing & resource issues are top reasons we aren't submitting/winning proposals

Global project summary

January - December 2015

Why we **aren't submitting** proposals

Why we **aren't winning** proposals

1667
344
82
183

LACK OF DIVERSITY

Reports by Apple, Yahoo, Google, Facebook, LinkedIn, Twitter and eBay show that seven of the tech sector's largest companies are largely white and male. Apple has slightly more Hispanics and blacks, probably due to greater diversity at its retail stores.

1 APPLE
2 YAHOO
3 GOOGLE
4 FACEBOOK
5 LINKEDIN
6 TWITTER
7 EBAY

WORKFORCE BY GENDER AND ETHNICITY

Mostly male

August 2014 global staff

MALE 70%
FEMALE 30%

NOTE Data may not total 100% because of rounding

Mostly white

August 2014 U.S. staff

WHITE 55%
ASIAN 15%
HISPANIC 11%
BLACK 7%
TWO OR MORE RACES 2%
OTHER/NOT DISCLOSED 1%
UNDECLARED 9%

SOURCES Apple, Yahoo, Google, Facebook, LinkedIn, Twitter and eBay
Julie Snider USA TODAY

LACK OF DIVERSITY

Reports by Apple, Yahoo, Google, Facebook, LinkedIn, Twitter and eBay show that seven of the tech sector's largest companies are **largely white and male**. Apple has slightly more Hispanics and blacks, probably due to greater diversity at its retail stores.

BREAKDOWN BY **GENDER**

BREAKDOWN BY **ETHNICITY**

From critique to creation

...

Title of this Graph

A description of the data or something worth highlighting to set the stage.

Source: Somewhere reputable

Visual Cues

Visualization involves encoding data with shapes, colors, and sizes. Which cues you choose depends on your data and your goals.

Position
Where in space the data is

Length
How long the shapes are

Angle
Rotation between vectors

Direction
Slope of a vector in space

Shapes
Symbols as categories

Area
How much 2-D space

Volume
How much 3-D space

Color saturation
Intensity of a color hue

Color hue
Usually referred to as color

FIGURE 3-3 *Visual cues*

Coordinate System

You map data differently with a scatterplot than you do with a pie chart. It's x- and y-coordinates in one and angles with the other; it's cartesian versus polar.

Scale

Increments that make sense can increase readability, as well as shift focus.

Linear

Values are evenly spaced

Logarithmic

Focus on percent change

Categorical

Discrete placement in bins

Ordinal

Categories where order matters

Percent

Representing parts of a whole

Time

Units of months, days, or hours

Context

If your audience is unfamiliar with the data, it's your job to clarify what values represent and explain how people should read your visualization.

What _____ is the best and worst?	Maximums and minimums	
How has _____ changed over time?	Temporal patterns	
What _____ stands out from the rest?	Outliers	

What makes _____ different from _____?	Clustering	
How are _____ and _____ related to each other?	Correlation	
What's the breakdown for _____?	Distributions	

Storytelling

755

Steroids or Not, the Pursuit Is On

Barry Bonds is taking aim at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.

Lines are cumulative home runs.

Hank Aaron
755 homers
23 seasons

Babe Ruth
714 homers
22 seasons

Barry Bonds
708 homers
20 seasons

Bonds takes lead
Home runs after 16 seasons

Bonds	567
Aaron	554
Ruth	516

755
714
23 seasons
22 seasons
20 seasons

Bonds was injured last season. He played 14 games and hit 5 homers.

600
400
200
0
14th season
5 seasons
10
15

According to allegations in a book about Bonds, he began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.

Homer Pace After Age 34

If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.

PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS

Aaron
Actual homers slightly outpace projected homers for five seasons.

Ruth
Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.

Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.

Note: Ages as of July 1 of each season.

Others Taking Aim

Alex Rodriguez ▲
Is ahead of the pace set by all three home run leaders.

Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.

Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.

Differing Paths to the Top of the Charts

The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).

15 times hit 30 or more (M.L. most).

Hit only 20 over first five seasons.

Averaged 52 from 2000 to 2004.

No one hit more from 1950-69.

Three 60-homer seasons is record.

Triple Crown in '66 (49, 122, .316).

First to hit 70 in a season.

Only McGwire had more in the 90's.

Youngest to reach 400 homers.

Second most ever in first five seasons.

Amanda Cox and Joe Ward/The New York Times

Storytelling Analysis

How did you engage with the story?

Walk us through your initial reaction to follow-on details

What do you think is the author's intent? What is the purpose?

Can you think of better (or worse) ways to communicate the same story?

So What is a Good Story?

Introduce interesting characters

Challenge is believable

There are hurdles to overcome

Outcome or prognosis is clear

Who is your audience?

What questions are you answering?

Why should the audience care?

What are your major insights and surprises?

What change do you want to affect?

Don’t bury the lede!

At a glance, does your audience understand your main thesis?

http://www.nytimes.com/interactive/2016/04/16/upshot/stephen-curry-golden-state-warriors-3-pointers.html

But can repeated and more in-depth looks add to our understanding?

Stephen Curry's 3-Point Record in Context: Off the Charts

By GREGOR AISCH and KEVIN QUEALY APRIL 16, 2016

This chart contains 752 lines — one for each N.B.A. player who finished in the top 20 in 3-point attempts made in each season since 1980. Sitting atop it is the Golden State Warriors' Stephen Curry, who finished the regular season with a record 402 3-pointers.

The record is an outlier that defies most comparisons, but here is one: It is the equivalent of hitting 103 home runs in a Major League Baseball season.

The colors show a clear progression toward more 3-pointers. In the 1979-80 N.B.A. season, the first to feature the 3-pointer, making just 21 was good enough to put a player among the league's top 20. On Feb. 27, Curry made 12 3-pointers in a single game.

How can we best put the gap between Curry and the best

Guide your audience

Don’t waste your audience’s cognitive effort!

Lead your audience through the major steps of your story

Point out interesting key facts and insights using captions and annotations

http://www.nytimes.com/interactive/2015/05/03/upshot/the-best-and-worst-places-to-grow-up-how-your-area-compares.html

Brazos County is very bad for income mobility for children in poor families. It is better than only about 12 percent of counties.

Consider framing effects

Tells the audience: here is the right way to think about my problem

Catch the audience's attention and shapes the story

Hopefully, your insights will seem obvious given your framing

Can be persuasive

http://www.computerworld.com/article/3088166/data-analytics/wh
how-to-lie-with-statistics-did-us-a-disservice.html

Iraq's bloody toll
Monthly fatalities
2003
2004
2005
2006
2007
2008
2009
2010
2011
Coalition deaths
4,802
Civilian deaths
113,728
May
US President George W. Bush declares an end to combat
December
Saddam Hussein captured
Jan 2004
US hands sovereignty to interim Iraqi government
December, 2005
Iraqis elect full-term government and parliament
August, 2005
950 killed in a single incident when fears of an insurgent attack caused a huge crowd of Shiite pilgrims to stampede on a bridge
January, 2009
US forces come under Iraqi mandate. Control of Green Zone handed to Iraq
January, 2007
Bush announces a surge of extra troops will be deployed
December, 2006
Saddam Hussein executed
March, 2003
US-led forces invade
The United States officially marked the end of almost nine years of bloody military engagement in Iraq on Thursday. Over 4,800 coalition soldiers and tens of thousands of Iraqis lost their lives in a war that defined a decade.
US ground troops in Iraq
President Bush approves 21,500 extra troops
Coalition deaths
United States
4,484
Total
4,802
Others
139
Britain
179
Iraqi security forces
10,125
Coalition fatalities by area
Baghdad and the vast desert province of Anbar saw some of the fiercest fighting, the latter being the scene of sectarian tensions and a bloody Sunni insurgency
0-99
100-499
500+
Dohuk
Nineweh
Arbil
Sulaymaniyah
Kirkuk
Salahuddin
Diyala
Baghdad
Anbar
Kut
Babil
Wasit
Qadisiyah
Maysan
Najaf
Dhi Qar
Basra
Muthanna
The biggest killers
How civilians died, according to the data available*
Incidents to March 19, 2008
Air strikes
Roadside bombs
Suicide bombs
Vehicle bombs
Mortar fire
Executions
Small arms gunfire

Iraq: Deaths on the decline

Think about how your audience engages with your story

Draw your line on the chart below