COMP-5360 / MATH-4100 Introduction to Data Science Visualization

Alexander Lex
alex@sci.utah.edu

THE UNIVERSITY OF UTAH

[xkcd]

visualization *pictures*

The purpose of computing is insight, not numbers.

- Richard Wesley Hamming

- Card, Mackinlay, Shneiderman

Banana	***M. acuminata***
Date	***P. dactylifera***
Cress	***Arabidopsis thaliana***
Rice	***Oryza sativa***
Sorghum	***Sorghum bicolor***
Brome	***Brachypodium distachyon***

[D'Hont et al., Nature, 2012]

The Banana Chart Redesigned: UpSet

Visualization Definition

Visualization is the process that **transform**s
(abstract) **data** into
interactive graphical representations for the purpose of
exploration, confirmation, or presentation.

Good Data Visualization

... makes data accessible

... combines strengths of humans and computers

... enables insight

... communicates

Why Visualize?

To inform humans: Communication

How is ahead in the election polls?

When questions are not well defined: Exploration

What is the structure of a terrorist network?

Which drug can help patient X?

Purpose of Visualization

[Obama Administration]

Open Exploration ⟷ Confirmation ⟷ Communication

The Ability Matrix

Why not just use Statistics?

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.0	10	9.1	10	7.4	8	6.5
8	6.9	8	8.1	8	6.7	8	5.7
13	7.5	13	8.7	13	12.	8	7.7
9	8.8	9	8.7	9	7.1	8	8.8
11	8.3	11	9.2	11	7.8	8	8.4
14	9.9	14	8.1	14	8.8	8	7.0
6	7.2	6	6.1	6	6.0	8	5.2
4	4.2	4	3.1	4	5.3	19	12.
12	10.	12	9.1	12	8.1	8	5.5
7	4.8	7	7.2	7	6.4	8	7.9
5	5.						6.8

Mean x: 9 y: 7.50
Variance x: 11 y: 4.122
Correlation x – y: 0.816
Linear regression: y = 3.00 + 0.500x

Anscombe’s Quartett

Mean x: 9 y: 7.50
Variance x: 11 y: 4.122
Correlation x – y: 0.816
Linear regression: y = 3.00 + 0.500x

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, CHI 2017, Justin Matejka, George Fitzmaurice

Visualization =

Human Data Interaction

Visualization in the Data Science Process

Figure 2-2. The data science process

Why Humans?

Leveraging human capabilities

- **Pattern Discovery**: clusters, outliers, trends
- **Contextual Knowledge**: expectations for dataset, explanations for patterns
- **Action**: humans learn and take action

But: we also have to **design for Humans and their limitations**

Not everything that can be drawn can be read!

Limits of Cognition

Daniel J. Simons and Daniel T. Levin, Failure to detect changes to people during a real world interaction, 1998

History: John Snow

John Snow, 1854

E. Tufte, Visual Explanations, 1997

Minard: Napoleon's Russian Campaign

C.J. Minard, 1869

E. Tufte, Writings, Artworks, News

Communicate

Hans Rosling, TED 2006

Marks & Channels

How can I visually represent two numbers, e.g.,

4 and 8

Marks & Channels

Marks: represent **items** or **links**

Channels: change **appearance** based on **attribute**

Channel = Visual Variable

Marks for Items

Basic geometric elements

3D mark: Volume, but rarely used

Marks for Links

Containment

Connection

Channels (aka Visual Variables)

Control appearance proportional to or based on attributes

Using Marks and Channels

Mark: Line

Channel: Length/Position

1 quantitative attribute

1 categorical attribute

Mark: Point

Channel: Position

2 quantitative attr.

Adding Hue

+1 categorical attr.

Adding Size

+1 quantitative attr.

Good bar chart?

Rule: Use channel proportional to data!

Types of Channels

Magnitude Channels

How much?

Position

Length

Saturation …

Identity Channels

What? Where?

Shape

Color (hue)

Spatial region …

Ordinal & Quantitative Data

Categorical Data

Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes

Position on common scale

Position on unaligned scale

Length (1D size)

Tilt/angle

Area (2D size)

Depth (3D position)

Color luminance

Color saturation

Curvature

Volume (3D size)

Identity Channels: Categorical Attributes

Spatial region

Color hue

Motion

Shape

Position

Strongest visual variable

Suitable for all data types

Problems:

- Sometimes not available (spatial data)
- Cluttering

Example: Scatterplot

Position in 3D?

[Spotfire]

Length & Size

Good for 1D, OK for 2D, Bad for 3D

Easy to see whether one is bigger

Aligned bars use position redundantly

Example 2D Size: Bubbles

Value/Luminance/Saturation

OK for quantitative data when length & size are used.

Not very many shades recognizable

Example: Diverging Value-Scale

Color

Good for qualitative data (identity channel)

Limited number of classes/length (~7-10!)

Does not work for quantitative data!

Lots of pitfalls! Be careful!

My rule:

- minimize color use for encoding data
- use for brushing

What is a colormap?

[0,8] →

specifies a mapping between color and values

- categorical vs ordered
- sequential vs diverging
- segmented vs continuous
- univariate vs bivariate

expressiveness: match colormap to attribute characteristics!

http://colorbrewer2.org/

Qualitative Data Vis

Color labeling (nominal information coding)

recommended: about 6, no more than 10

22 colors, but only 8 distinguishable

Application: Labeling

Carefully designed color scheme for 12 colors [colorbrewer]

Quantitative Data Vis

use value

saturation works but not as good

don‘t use hue!

Danger: rainbow color map

[Rogowitz and Treinish, Why Should Engineers and Scientists Be Worried About Color? http://www.research.ibm.com/people/l/lloydt/color/color.HTM]

[Rogowitz and Treinish, How NOT to Lie with Visualization, www.research.ibm.com/dx/proceedings/pravda/truevis.htm]

Increment Threshold Experiment: Luminance

Increment Threshold Experiment: Rainbow Color Scale

Color Blindness

10% of males, 1% of females (probably due to X-chromosomal recessive inheritance)

Most common: red-green weakness / blindness

Reason: lack of medium or long wavelength receptors, or altered spectral sensitivity (most common: green shift)

Normal Color Perception | Deuteranopia (no green receptors) | Protanopia (no red receptors)

Color Blindness Tests

Color Blindness

Protanope	Deuteranope	Tritanope
No L cones	No M cones	No S cones
Red / green deficiencies		Blue / Yellow deficiency

Based on slide from Stone

NYT, Feb 12, 2012

Shape

Great to recognize many classes.

No grouping, ordering.

ALFA ROMEO
ACURA
ASTON MARTIN
ASTON MARTIN
BENTLEY
Audi
BMW
MITSUBISHI MOTORS
Ferrari
Photoshop plugins

Autobahn
Autobahn ends
Motor vehicles only
End of motor vehicles only
Autobahn exit
Direction to autobahn
Distance to next change of traffic
Auto repairs
Gas station also with lead free fuel
Autobahn snack machines
Autobahn restaurant
Autobahn hotel
Danger
Crossroads
Oncoming traffic
Falling stones
Danger streetcar
Tents and trailers camping area
Radio station traffic information number
First aid station
Telephone
Police
Recommended speed
Parking on curb permitted
Parking area
Parking disc clock
Park and ride area
Autobahn detour
Detour route sign
24 hour service off of the autobahn
FOR SAFETY
You must KNOW
German trafficsigns
USAREUR POSTER 190-34-12, Jan. 2001 (supersedes USAREUR Poster190-34, Nov95)
http://mwr.hqusareur.army.mil
Autobahn direction sign
Direction to autobahn
Change of traffic lanes
Direction sign
Bypass routing
Detour
Direction of travel
Directional sign
European highway
Federal highway number
Built-up area (front)
Built-up area (reverse) revision
Single curve
Double curve
Road narrows
Road narrows
No motorcycles
End of city limits
Direction of travel for bicycles
Point of interest
Name place
Streetcar/bus stop
Occasionally used with supplemental signs
Supplemental sign
Danger of unexpected ice
Orange traffic arrow
Recommended direction of travel in stau
Soft shoulder
Secondary route
Right of way changed
Gravel
Road damage
Quay or river bank
Low flying aircraft
Pedestrian crosswalk ahead
Pedestrian crosswalk
No vehicles carrying dangerous goods
Residents only
Speed and distance sign
Solid white line
Broken white line
Passing only from broken line side
Off limits markings
Arrow on pavement
Directional arrows
Pedestrian crosswalk
No parking
Traffic jam ahead
Traffic jam area ahead
Bicycle crossing
Dangerous downgrade
Dangerous upgrade
Rough road
No bicycles
Slippery road
Loose gravel
Children
Construction site
Drawbridge ahead
Wild animal crossing
Domestic animal crossing
Side wind
Smog
Motor vehicle traffic prohibited
Signal lights ahead
Stop
Yield right of way
Mandatory direction of travel
Mandatory direction of travel
Mandatory direction of travel
Mandatory direction of travel
Tourist office or information
Maximum height allowed
Maximum width allowed
Maximum weight allowed
No vehicles carrying more than 3000 liters of pollutants
Vehicles above a specific axle weight prohibited
Motor vehicles prohibited
Tractors and trucks with an authorized loaded weight of more than 3.5 tons prohibited
Minimum distance between vehicles of 3.5t
Maximum length allowed
No U turns
Prohibited for all vehicles
Entry prohibited
Mandatory direction of travel
Mandatory direction of travel
Beginning of a pedestrian priority area
End of pedestrian priority area
Traffic circle
Distance to guarded railroad crossing
Distance to unguarded railroad crossing
Railroad crossing
Railroad crossing
Guarded railroad crossing
Unguarded railroad crossing
Speed zone
End of speed limit restricted area
Speed limit
End of speed limit
Only when wet
Right on red green arrow sign
Pedestrian zone
End of pedestrian zone
Children playing
Traffic directed by school guard
Bus lane
Restricted no stopping
No stopping
No passing
End of no passing zone
No passing for trucks in excess of 3.5 tons authorized weight
End of no passing zone
Limited parking place clock card in windshield
End of limited parking zone
End of restriction
Marks streetlights that are not lighted parklights must be left on at night
Customs control
Oncoming traffic has right of way
Snow chains mandatory
Horse riders only
Pedestrians only
Bicycles only
Information panel at frontier crossings
Priority road
Priority road
End of Priority road
Right of way
Priority road ahead
Bridge carrying capacity for NATO vehicles
One way street in direction of arrow
Dead end
One-way traffic
Oncoming traffic must wait
Compulsory minimum speed limit
End of compulsory minimum zone
Separated bicycle and pedestrian pavements restricted speed area
Jointly used pavements for pedestrians and bicycles
Taxi parking only
Water protection area
Exit number

Why are quantitative channels different?

Steven's Psychophysical Power Law: $S = I^N$

S = sensation
I = intensity

How much longer?

2x

How much longer?

4x

How much steeper?

~4x

How much larger?

5x

How much larger?

A B

2x diameter

4x area

area is proportional to diameter squared

How much larger (area)?

A B

3x

How much darker?

A B

2x

How much darker?

A B

3x

Other Factors Affecting Accuracy

Alignment

Distractors

Distance

Common scale

…

Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes

Position on common scale

Position on unaligned scale

Length (1D size)

Tilt/angle

Area (2D size)

Depth (3D position)

Color luminance

Color saturation

Curvature

Volume (3D size)

Same

Same

Most

Effectiveness

Least

Identity Channels: Categorical Attributes

Spatial region

Color hue

Motion

Shape

Design Guidelines

Rule #1: Use the Best Visual Channel Available for the Most Important Aspect of your Data

Effectiveness Principle

Rule #2: The visualization should show all of the data, and only the data

Expressiveness Principle

Tufte's Integrity Principles

Show **data variation**, not design variation

Clear, detailed, and thorough **labeling** and **appropriate scales**

Size of the **graphic effect** should be **directly proportional to the numerical quantities** ("lie factor")

The Lie Factor

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

Lie Factor - Graphical Integrity

Magnitude in data must correspond to magnitude of mark

Effect in Data: factor 1.14
Effect in Graphic: factor 5
Lie Factor: 5/1.14 = 4.38

Scale Distortions

Flowing Data

What’s wrong?

Viele Bezieher mit "ungeklärter Staatsbürgerschaft"

Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um

What’s wrong?

Viele Bezieher mit "ungeklärter Staatsbürgerschaft"

Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um

1
2
3

Asyl
Opf
frei
"Kr
Räd
geh
Zwe
Mit
Wol
Am
Stre
Mes
Abe

Viele Bezieher mit "ungeklärter Staatsbürgerschaft"

Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um

What’s wrong?

Grafik der Kronen-zeitung

Grafik in echt

OBAMACARE ENROLLMENT
7,000,000
7,100,000
ACTUAL ENROLLMENT
GOAL
FOX NEWS
LIVE

Start Scales at 0?

A. Kriebel, VizWiz

Scales at 0

Use a baseline that shows the data, not the zero-point.

Global Warming?

The Daily Mail, UK, Jan 2012

Global Warming?

Mother Jones

Global Warming - Frame the Data

Mother Jones

Death to Pie Charts

Share of coverage on TechCrunch

"I hate pie charts.
I mean, really hate them."

www.storytellingwithdata.com/2011/07/death-to-pie-charts.html

Cole Nussbaumer

Redesign

TechCrunch Coverage: 2005 - 2011
A slightly better pie?
Network/Hosting, 1%
Investor, 1%
PR, 1%
Cleantech, 1%
Music, 1%
News, 2%
Other, 2%
Enterprise, 2%
No Category, 3%
Video, 3%
Advertising, 3%
General Consumer Web, 23%
E-Commerce, 5%
Social Networks, 12%
Hardware, 6%
Entertainment, 6%
Search, 10%
Softward, 8%
Mobile, 9%

TechCrunch Coverage: 2005 - 2011
Bars are best!
General Consumer Web 23%
Social Networks 12%
Search 10%
Mobile 9%
Softward 8%
Entertainment 6%
Hardware 6%
E-Commerce 5%
Advertising 3%
Video 3%
No Category 3%
Enterprise 2%
Other 2%
News 2%
Music 1%
Network/Hosting 1%
Investor 1%
PR 1%
Cleantech 1%

Can you spot the differences?

Can you spot the differences?

My favorite pie chart

My second favorite pie chart

So, what to use instead?

imagine you just completed a pilot summer learning program on science aimed at improving perceptions of the field among 2nd and 3rd grade elementary children

http://www.storytellingwithdata.com/blog/2014/06/alternatives-to-pies

Alternative #1: Show the Number(s) Directly

After the pilot program,

68%

of kids expressed interest towards science,
compared to 44% going into the program.

Alternative #2: Simple Bar Graph

Alternative #3: 100% Stacked Horizontal Bar Graph

Alternative #4: Slopegraph

Visualization Design Principles

Maximize Data-Ink Ratio

Maximize Data-Ink Ratio

Eyes Beat Memory

Don’t make people memorize: Show them

http://www.randalolson.com/2015/08/23/small-multiples-vs-animated-gifs-for-showing-changes-in-fertility-rates-over-time/

Eyes Beat Memory: Small Multiples

A lot of charts

Do we need all of them?

Eyes Beat Memory: Small Multiples

Simplify!

Total fertility rate in USA and Japan, 1947 - 2010
Total fertility rate
5
4
3
2
1
1950
1960
1970
1980
1990
2000
2010
USA
Japan
Data source: Human Fertility Database (humanfertility.org)
Author: Randy Olson (randalolson.com / @randal_olson)

Average age when giving birth in USA and Japan, 1947 - 2010
Average age of women when giving birth
30
29
28
27
26
25
1950
1960
1970
1980
1990
2000
2010
Japan
USA
Data source: Human Fertility Database (humanfertility.org)
Author: Randy Olson (randalolson.com / @randal_olson)

Tabular Data

Histogram

Good #bins hard to predict

make interactive!

rules of thumb:

- #bins = sqrt(n)
- #bins = log2(n) + 1

10 Bins

20 Bins

Box Plots

aka Box-and-Whisker Plot

Show outliers as points!

Not so great for non-normal distributed data

Especially bad for bi- or multi-modal distributions

Wikipedia

One Boxplot, Four Distributions

Figure 1: Histograms and box plot: four samples each of size 100

http://stat.mq.edu.au/wp-content/uploads/2014/05/Can_the_Box_Plot_be_Improved.pdf

Box(and Whisker) Plots

http://xkcd.com/539/

Comparison

Streit & Gehlenborg, PoV, Nature Methods, 2014

Bar Charts vs Dot Plots

Data Source https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-10-67

https://twitter.com/robustgar/status/859318971920769024

Violin Plot

= Box Plot + Probability Density Function

http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

Showing Expected Values & Uncertainty

NOT a distribution!

Error Bars Considered Harmful:
Exploring Alternate Encodings for Mean and Error
Michael Correll, and Michael Gleicher

Heat Maps

binning of scatterplots

instead of drawing every point, calculate grid and intensities

Scatterplot Matrix & PCs

Scatterplot Matrix

Parallel Coordinates

Table

Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Pixel Based Displays

Each cell is a “pixel”, value encoded in color / value

Ordering critical for interpretation

If no ordering inherent, clustering is used

Scalable – 1 px per item

Good for homogeneous data

same scale & type

[Gehlenborg & Wong 2012]

Heterogeneous Data?

[Verhaak 2012]