

Statistics 101C Lecture 3: Data Visualization

Brice Randolph

University of California, Los Angeles

brandolph@ucla.edu

April 4, 2016

Agenda



What will we be doing in class today?

- Introduction to data visualization
- Historic examples of data visualization
- Examples using R
- Briefly talk about python + matplotlib

The Dude Recommends



- "It is easy to lie with statistics, but it is easier to lie without them."
-Fredrick Mosteller
- Start replicating good visualizations:
<http://r4stats.com/examples/graphics-ggplot2/>
- ggplot2 cheatsheet
- Read your first article:
<http://research.microsoft.com/pubs/66183/isard-blake-98.pdf>

Visualizations

Some things we might be trying to show with a visualization:

- Relationships between variables (clustering in a graph)
- Prediction outcome
- Complicated stuff in general [Isard & Blake '97]

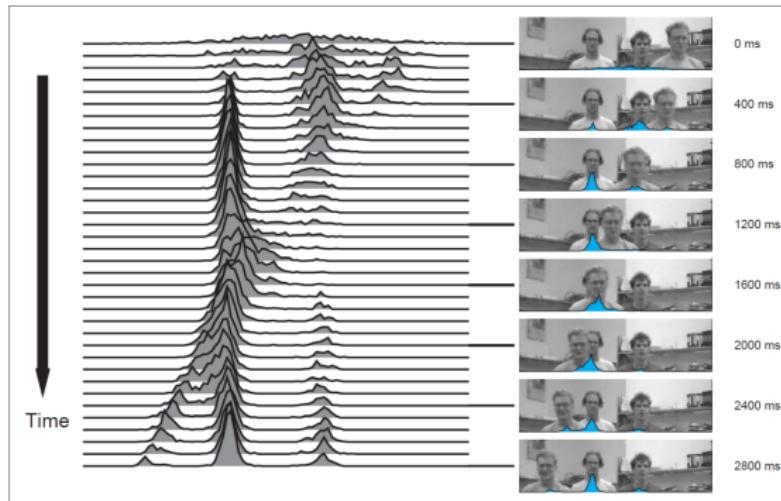
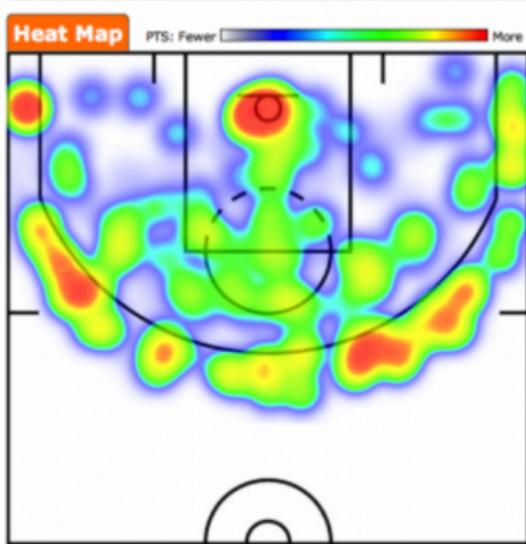


Figure 2.13: 1D projection of the state density across multiple frames of a video, from [100].

What do you think this visualization is showing?

Any idea what this is? Who?

1st Quarter



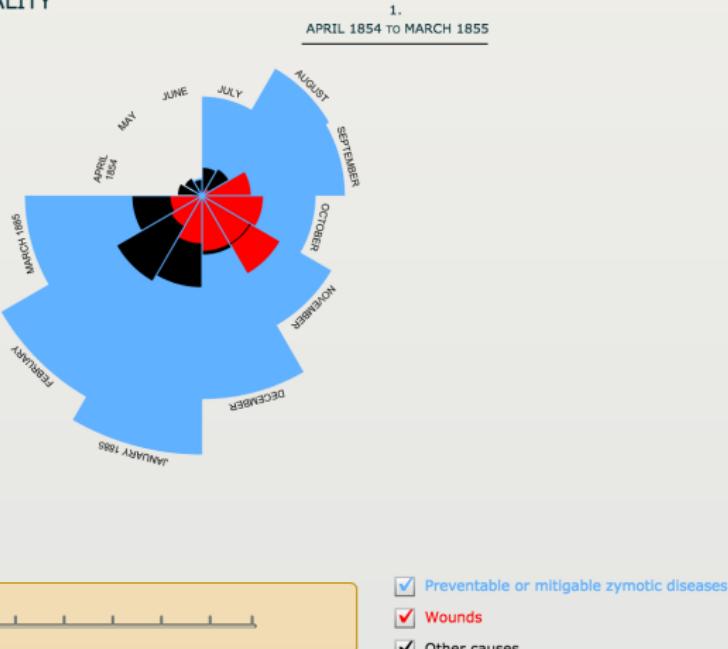
How might a team take advantage of this information?

Florence Nightingale

Audience: layperson & British parliament

Goal: To show that more soldiers were dying from preventable diseases than battle wounds

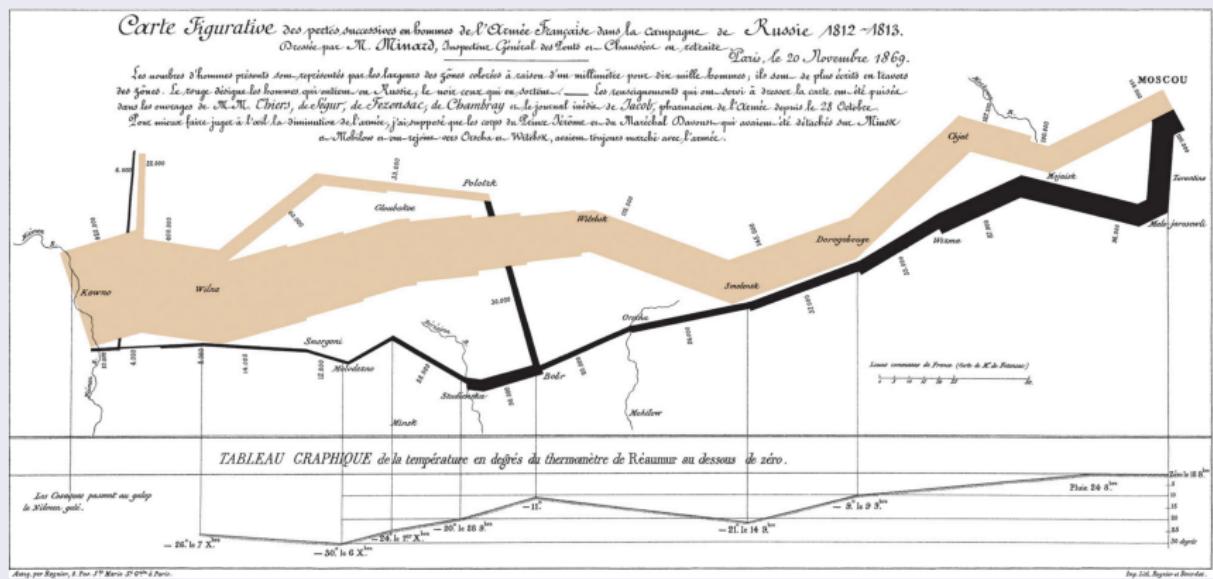
DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST



Napoleon's March 1812-1813

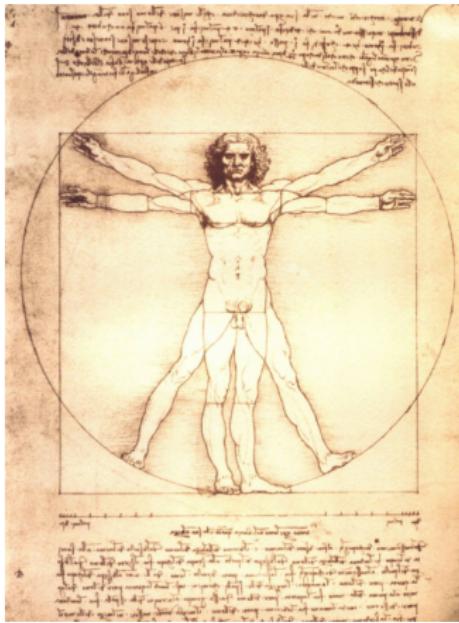
Six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates

Charles Joseph Minard



Leonardo Da Vinci

What does this graphic tell us?



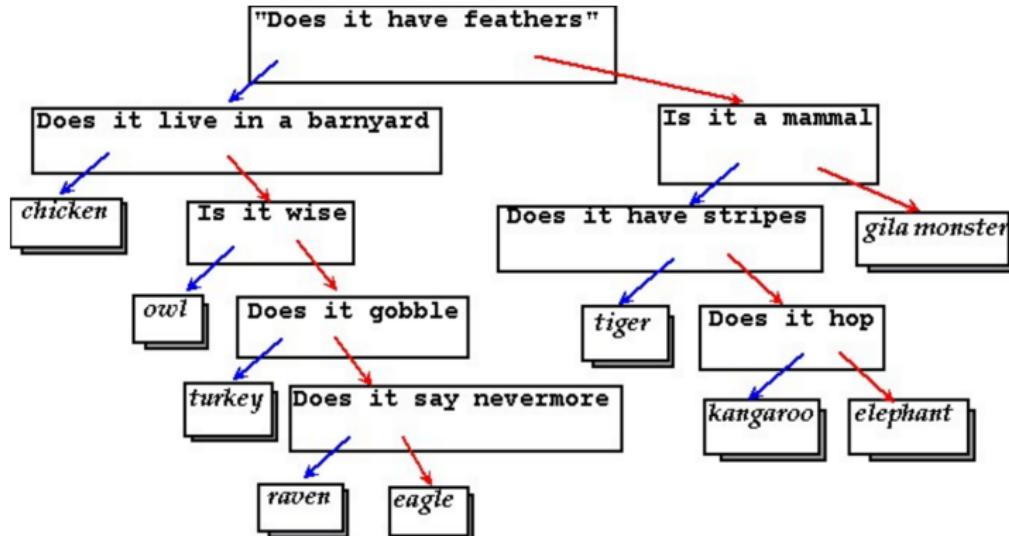
10 Questions

I'm thinking of a mathematical object.

You have 10 questions to try and guess what it is...

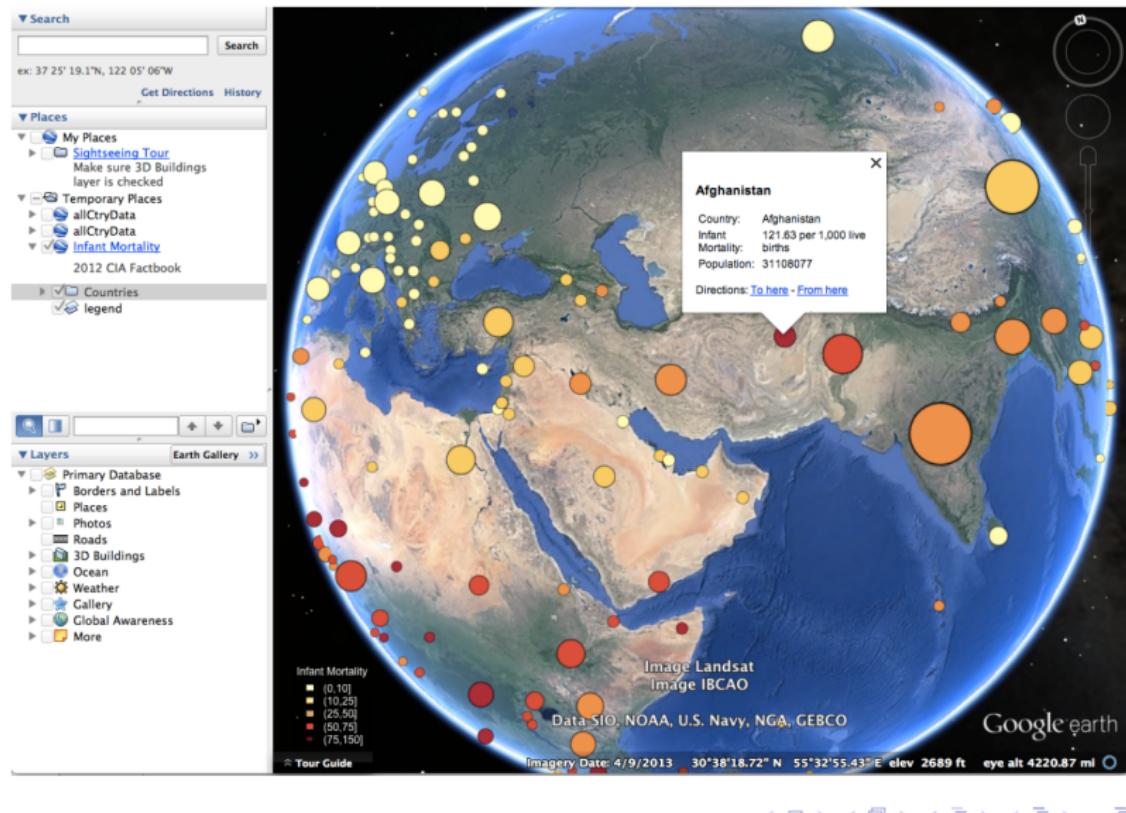
You can ask things like... Is it a complex number?

20 Questions game visualization



But was this trained by hand? Or did it **learn** these classes?

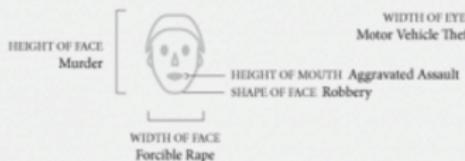
Infant Mortality by Country



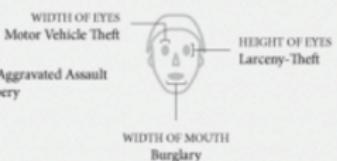
Chernoff Faces

The Face of Crime in the United States

Violent Crime

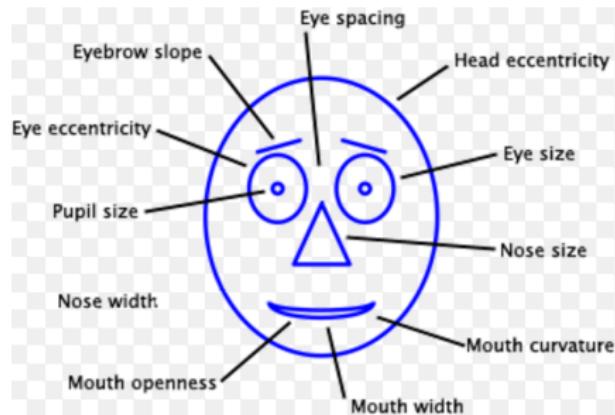


Property Crime



Chernoff Faces

Assuming we can read peoples faces in real life, can we leverage this on multivariate data?



Basics

- ① 5 number summary
- ② box & whiskers plot
- ③ density plots/ histograms
- ④ overlay predicted models on scatterplot

ggplot2

- ① data frames
- ② geometries
- ③ aesthetics
- ④ themes

ggplot2 Syntax

```
ggplot(data = <default data set>,
       aes(x = <default x axis variable>,
           y = <default y axis variable>,
           ... <other default aesthetic mappings>),
       ... <other plot defaults>) +
       geom_<geom type>(aes(size = <size variable for this geom>,
                             ... <other aesthetic mappings>),
                         data = <data for this point geom>,
                         stat = <statistic string or function>,
                         position = <position string or function>,
                         color = <"fixed color specification">,
                         <other arguments, possibly passed to the _stat_
                           function>) +
       scale_<aesthetic>_<type>(name = <"scale label">,
                                 breaks = <where to put tick marks>)
```

```
    labels = <labels for tick marks>,  
    ... <other options for the scale>) +  
  
theme(plot.background = element_rect(fill = "gray"),  
      ... <other theme elements>)
```

Constructing Plots

Surf the web and find a cool dataset

Try starting here: <http://archive.ics.uci.edu/ml/>

Think about questions you might want to ask & start plotting.

Please use ggplot2

Practice Interview Questions

Assume I don't know a thing about **shiny** (interactive web graphics in R).

Look at the following code and try to explain what the pieces do.

Feel free to tinker with the parameters:

<http://shiny.rstudio.com/gallery/kmeans-example.html>