Data Visualization

Brice Randolph

University of California, Los Angeles

brandolph@ucla.edu

github.com/briceran

April 19, 2016

Agenda

- Social Experiment
- DataFest
- Introduction to data visualization
- Historic examples of data visualization
- Anomaly Detection
- 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
- Wear sunscreen! Especially on the back of your neck and face.

Statistician vs. Data Scientist

- 50 Years of Data Science by David Donoho
- Regardless of what you call yourself, you should start collecting you own data. Why?

Social Experiment

"Sometimes when you assume the world is sane, kind, and generally fair, it appears to be just that..." - Famous Dude

What would happen if you started shamelessly emailing famous people with a nice note?

Results?

I can't seem to get ahold of Calvin Harris...

However, Judea Pearl called me two nights ago, I heard back from David Mumford & Hadley Wickham, Terrance Tao replied last year, and I got an email from Noam Chomsky this morning...

Presentation Etiquette

- Never say "I" - always "We"
- Put your phones away when other groups present.
- Try to avoid distracting movements. (Unless it is your goal)

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]

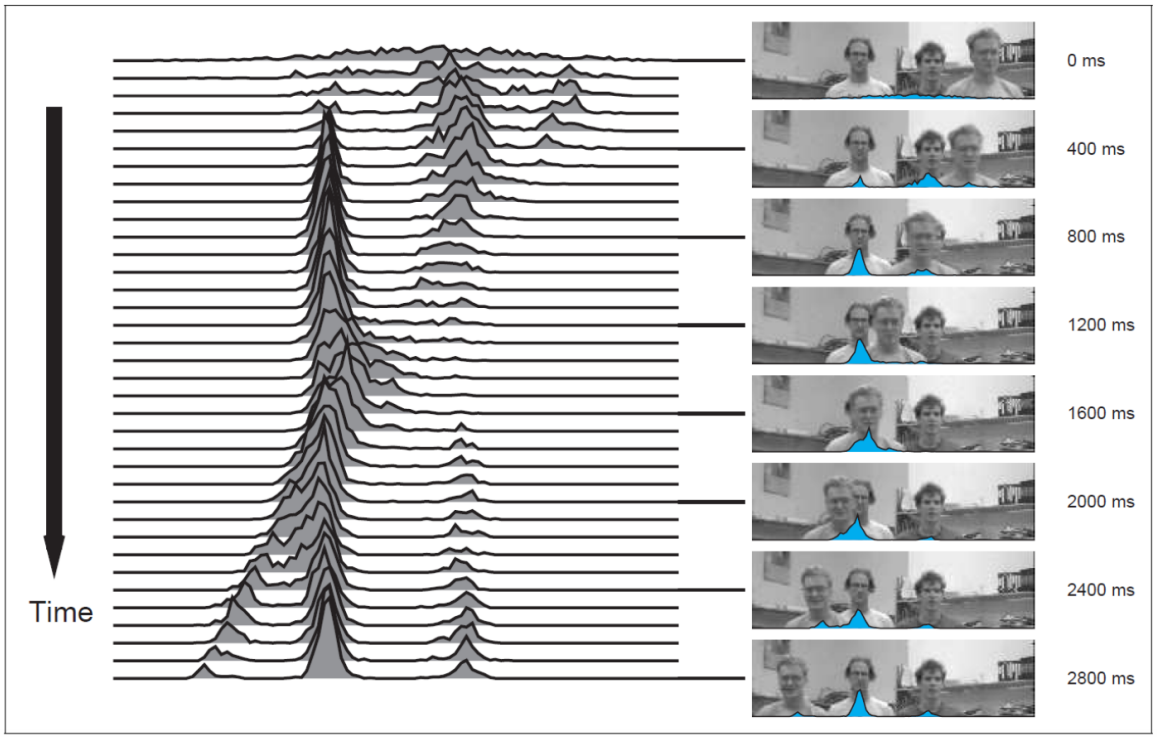
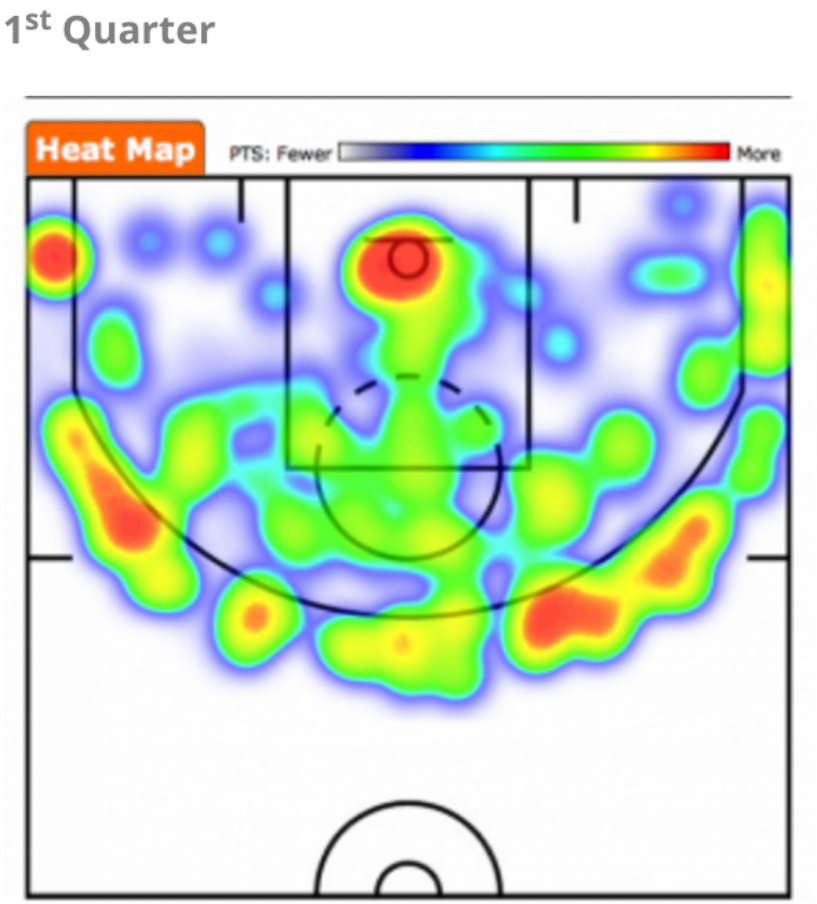


Figure 2.13: 1D projection of the state density across multiple frames of a video, from [100].

What do you think this visualization is showing?

Any idea what this is? Who?

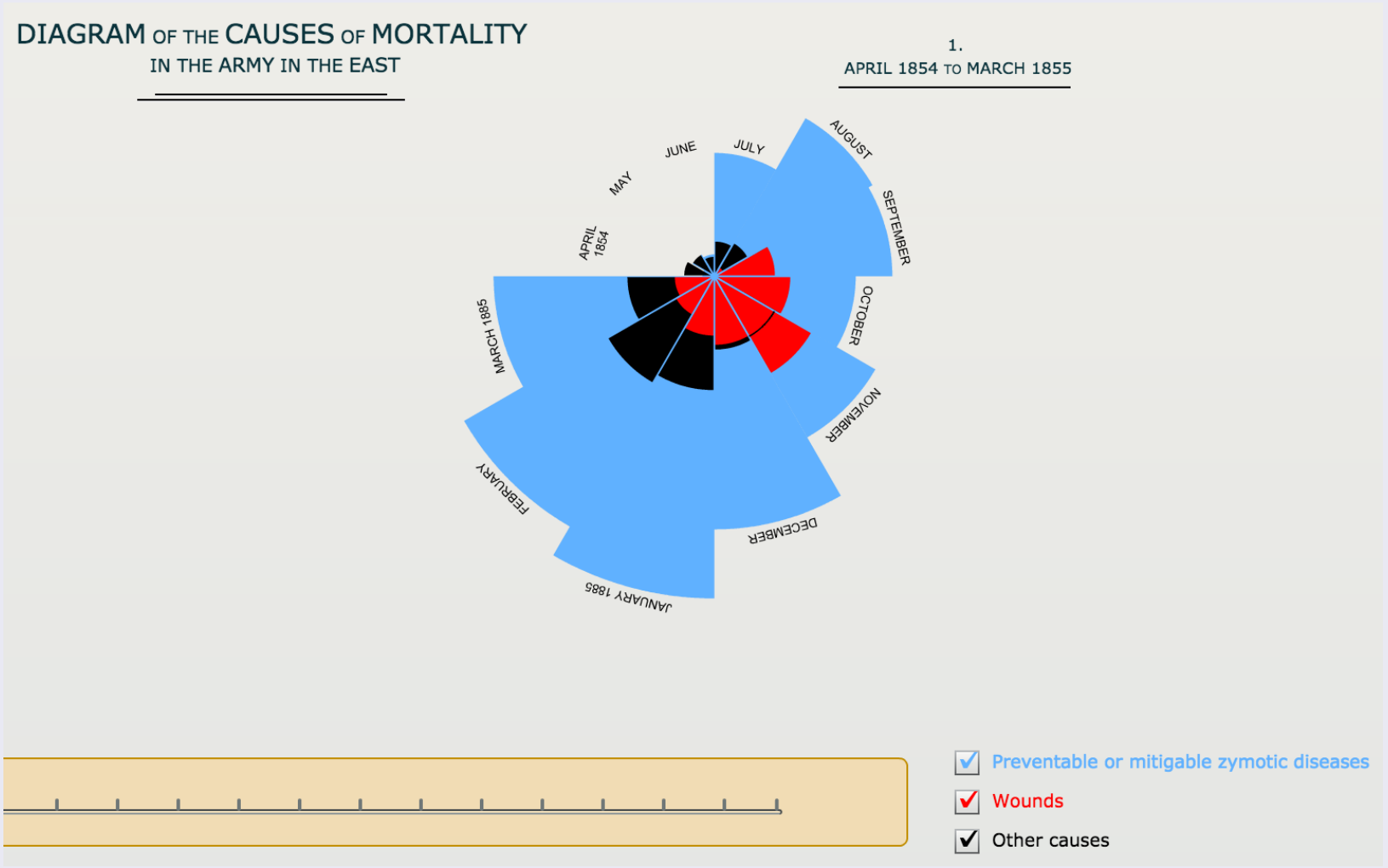


How might a team take advantage of this information?

Florence Nightingale

Audience: layperson & British parliament

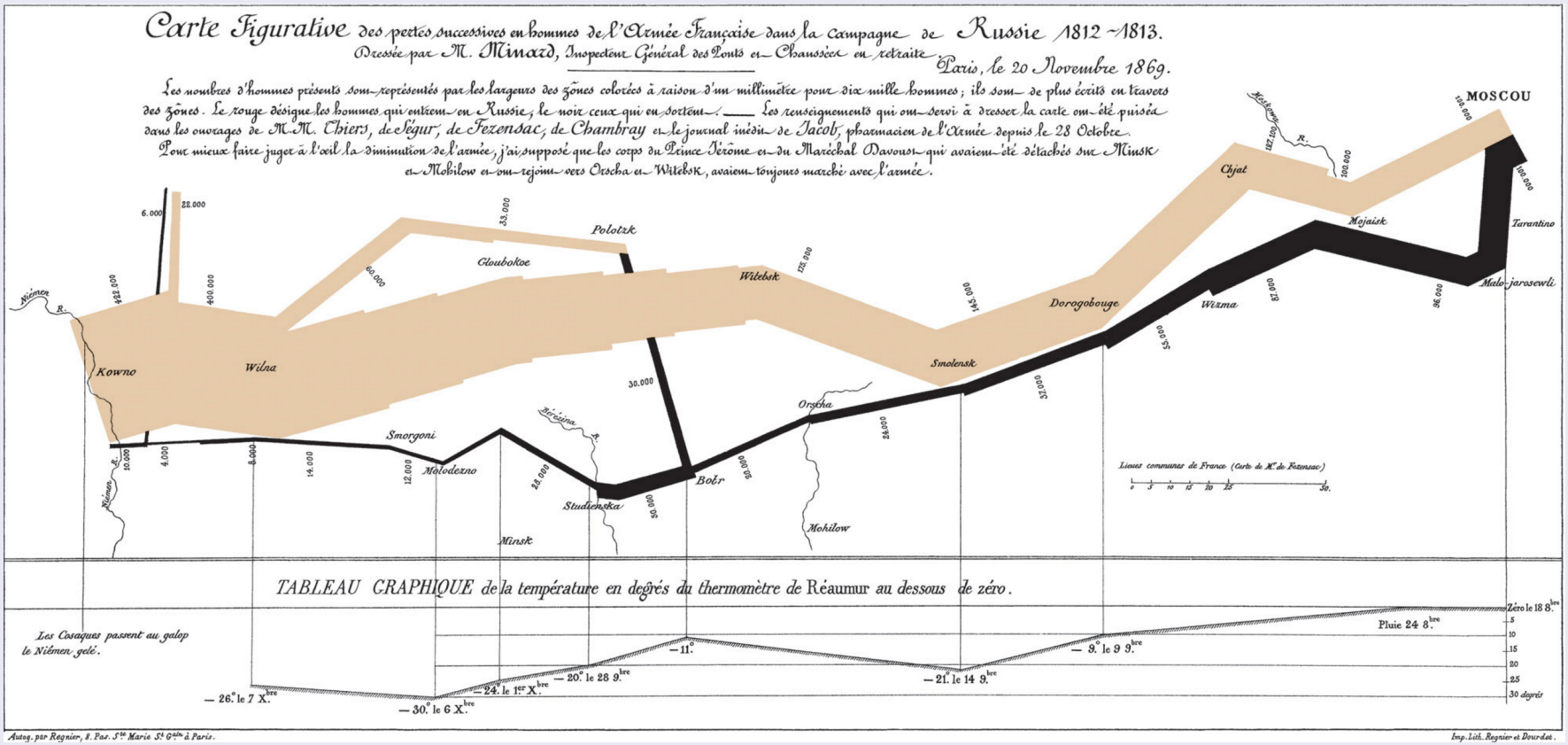
Goal: To show that more soldiers were dying from preventable diseases than battle wounds



Napoleon's March 1812-1813

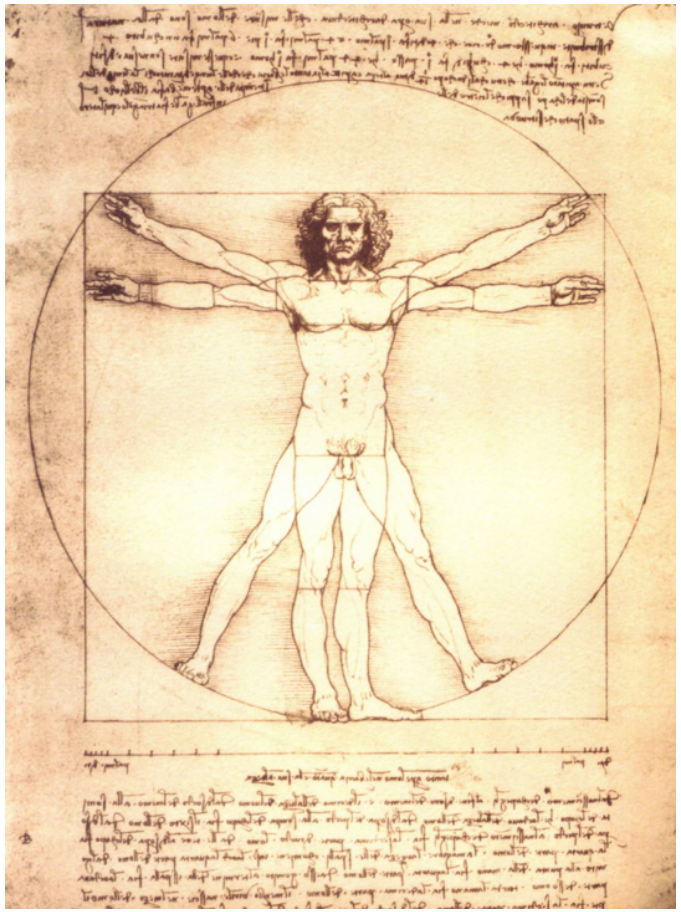
Six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates

Charles Joseph Minard

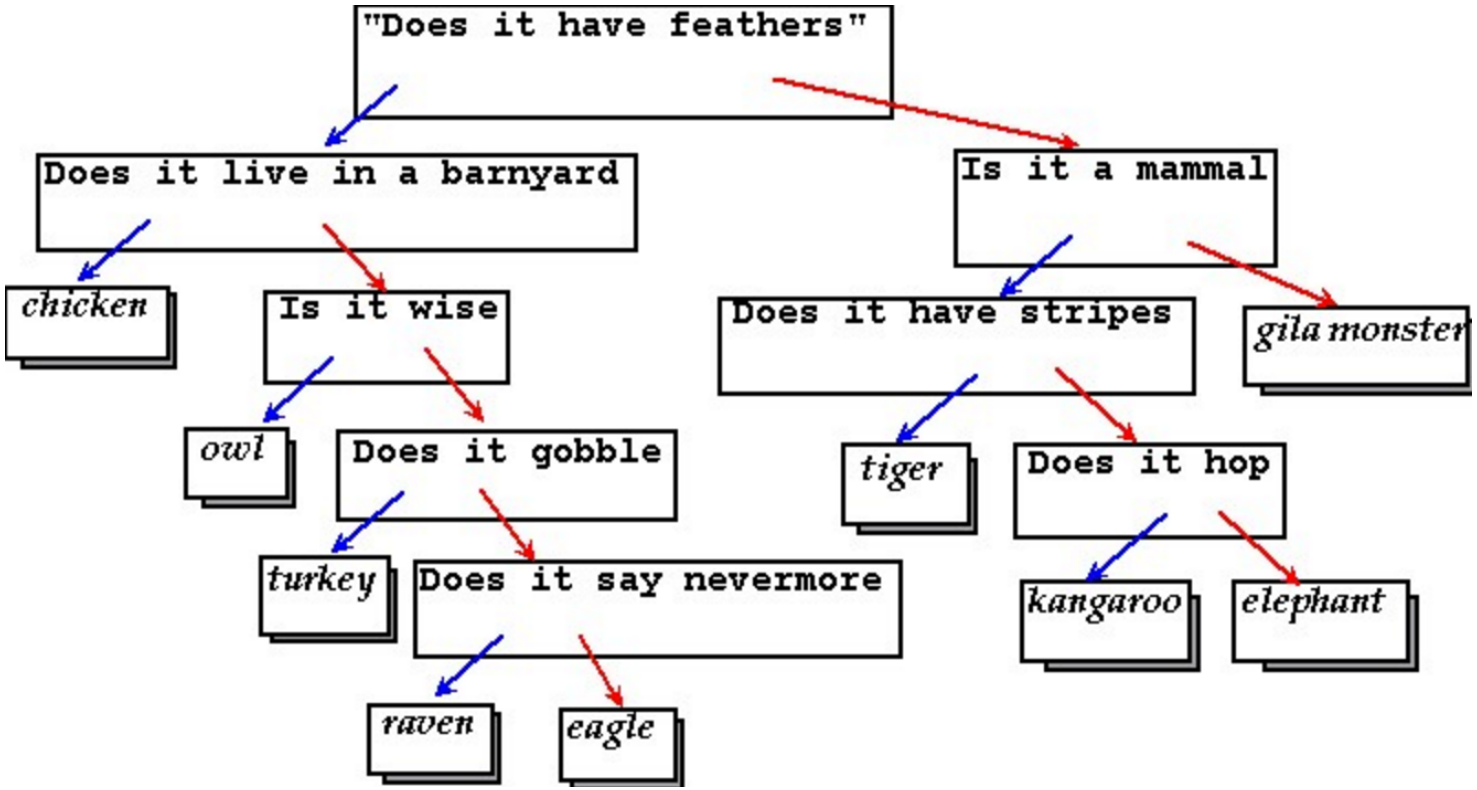


Leonardo Da Vinci

What does this graphic tell us?

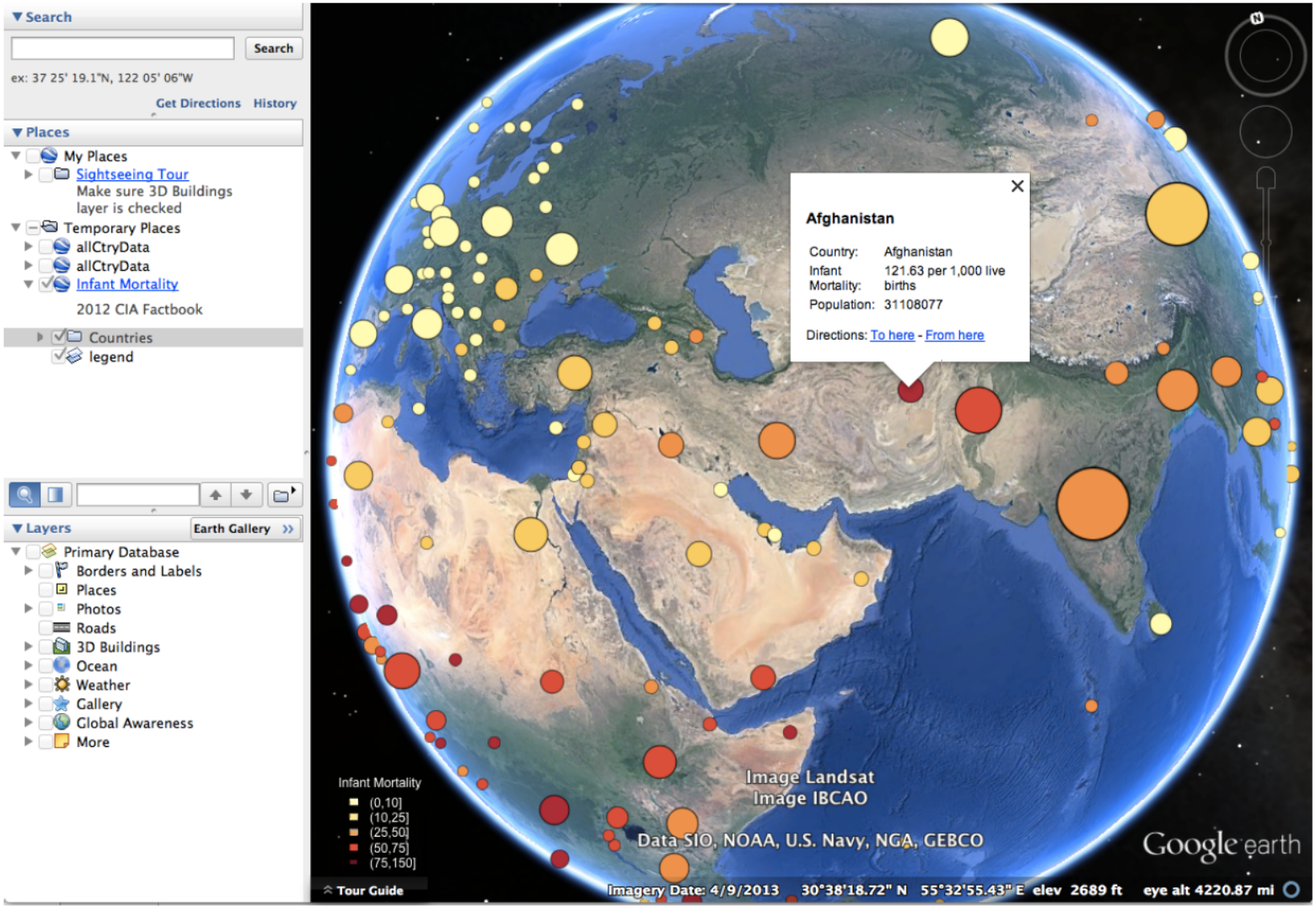


20 Questions game visualization



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

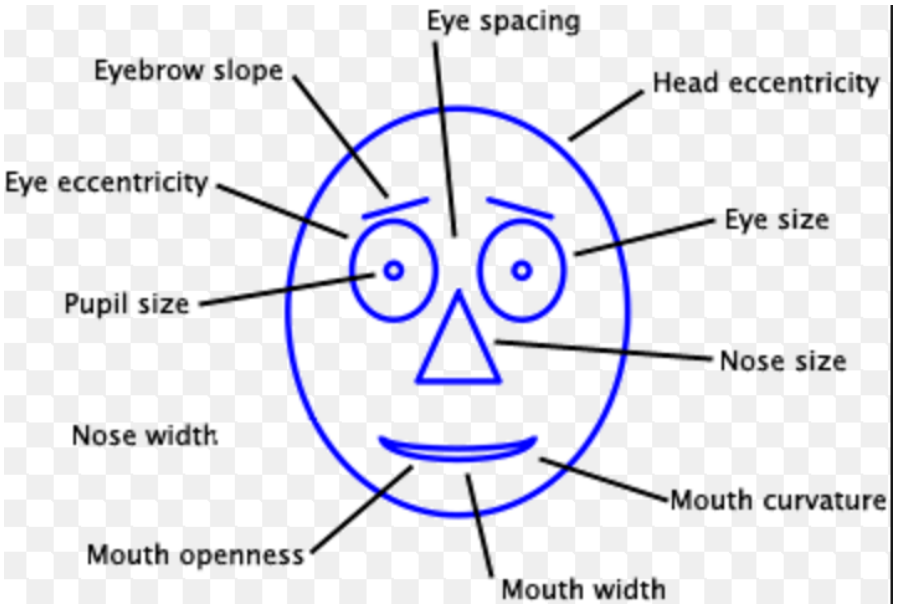
Infant Mortality by Country



Chernoff Faces

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

Anomaly Detection

Where's Waldo search. But we don't know what Waldo looks like, what he's wearing, and sometimes we think we see Waldo, but we don't. Oh yeah, and it's often on the scale of Manhattan.

Let's look at two examples of anomaly detection:

1. Is this guy cheating in Halo?
2. Credit Card Fraud

Cheating in Halo

What are some ways you could cheat in the game of Halo?

- Auto-headshot?
- Flying around map?

Imagine you are Bungie.net and have all the data...

- How do you catch someone cheating?
- What data do you look at?

Credit Card Fraud

Visa changed their fraud system in mid 90s and decreased fraud by 2/3rds.

They used to only analyze 2% with average fraud rates per category .
Warning signs?

In certain categories with transactions of $200 or more, prepaid cards were used 85% of the cases of fraud.

What information does this give us?
Do billing and shipping addresses differ?

Instagram Likes

- Fraud target (create and sell false followers and likes)
- How might we detect something like this?

What's trending?

How do you deduce what is trending on google/twitter?

Twitter trends are automatically generated by a program looking at the terms mentioned in tweets.
March 2013 400,000,000 per day.

Dec 2010 95,000,000 per day.

What is trending depends on how we define anomaly. This depends on what you are looking for.

In 2010 twitter changed what they called a trend.

Now: what is "most breaking and immediately popular". before this: twitter trends were dominated by what was popular- anything new was pushed out by things like Justin Bieber.

What worked at one time, didn't work later

Dynamic models! Note that twitter changed it's definition. Data is always changing

"The new algorithm identifies topics that are immediately popular, rather than topics that have been popular for a while or on a daily basis, to help people discover the 'most breaking' breaking news from across the world."

-Twitter As time goes on, the anomalies trail increases, and our tools get better.

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>)
```

Please load up Rstudio

and issue a library(ggplot2) command.

Practice Before DataFest

Surf the web and find a cool dataset

Try starting here: http://archive.ics.uci.edu/ml/ or kaggle

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