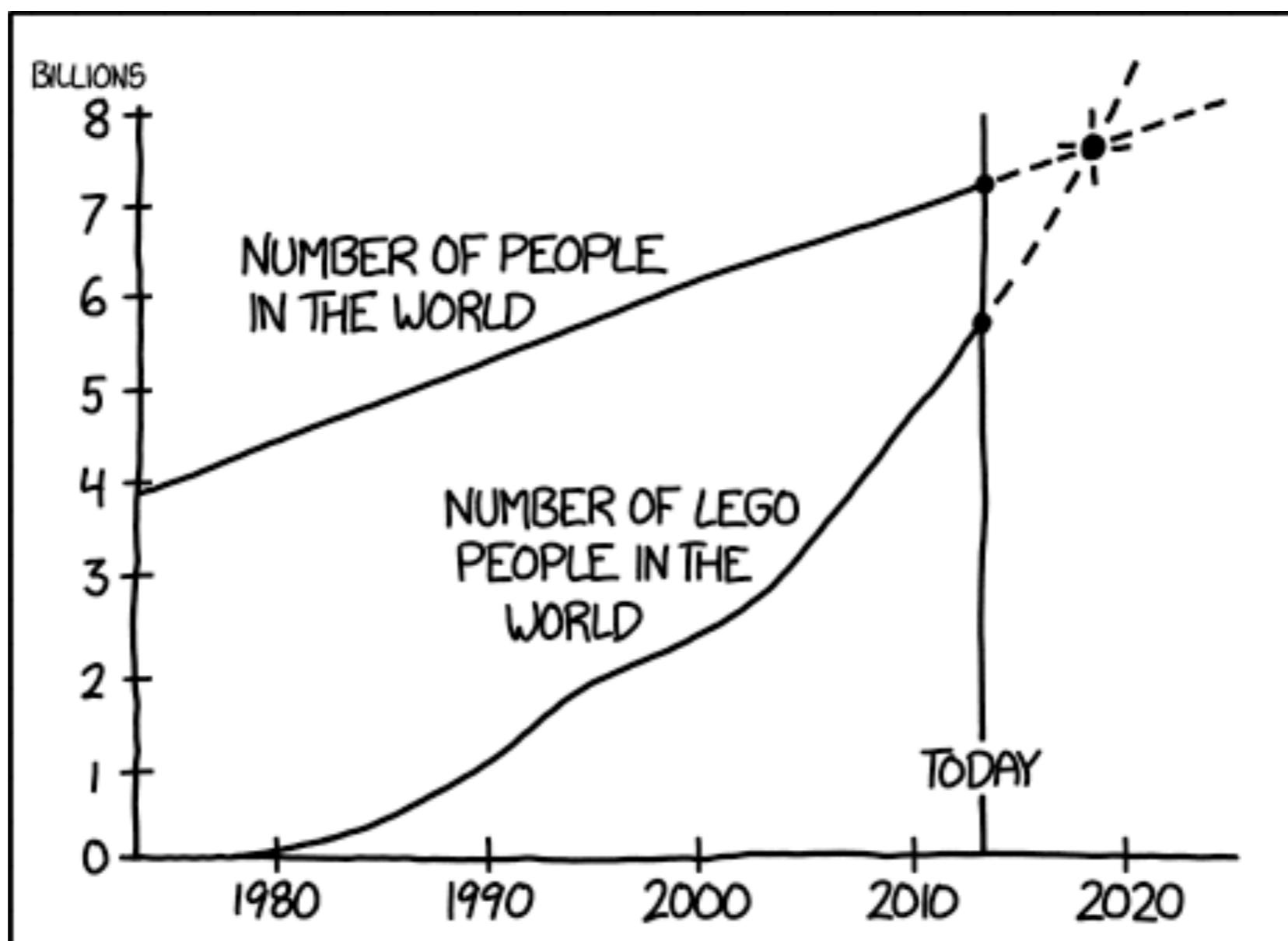


Data Visualisation for Beginners

Sara-Jayne Terp
2nd October 2014

Introduction



BY 2019, HUMANS WILL BE OUTNUMBERED.

(XKCD.com: data science humor)

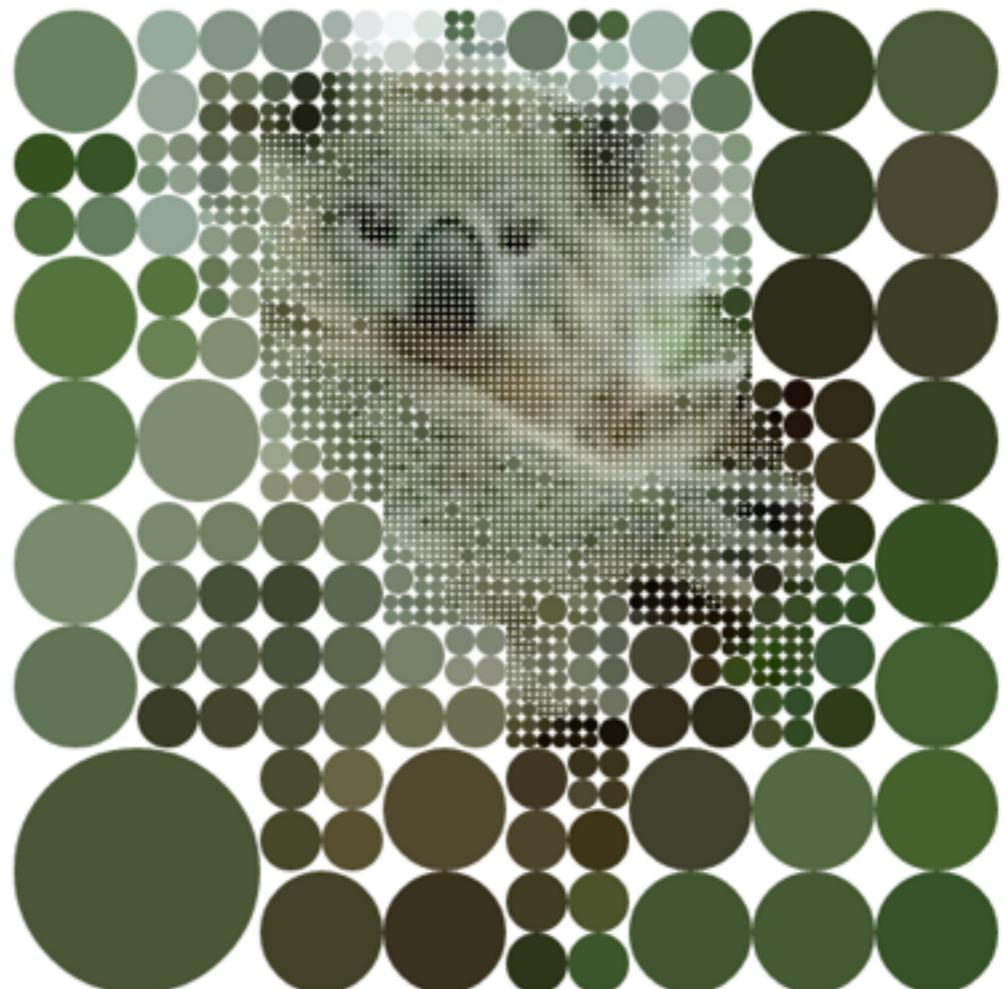
Talking About



Source: wordle.net

Not Talking About

- Algorithm details
- Machine learning
- Big data techniques
- Koalas



Data Visualisation



Data Science

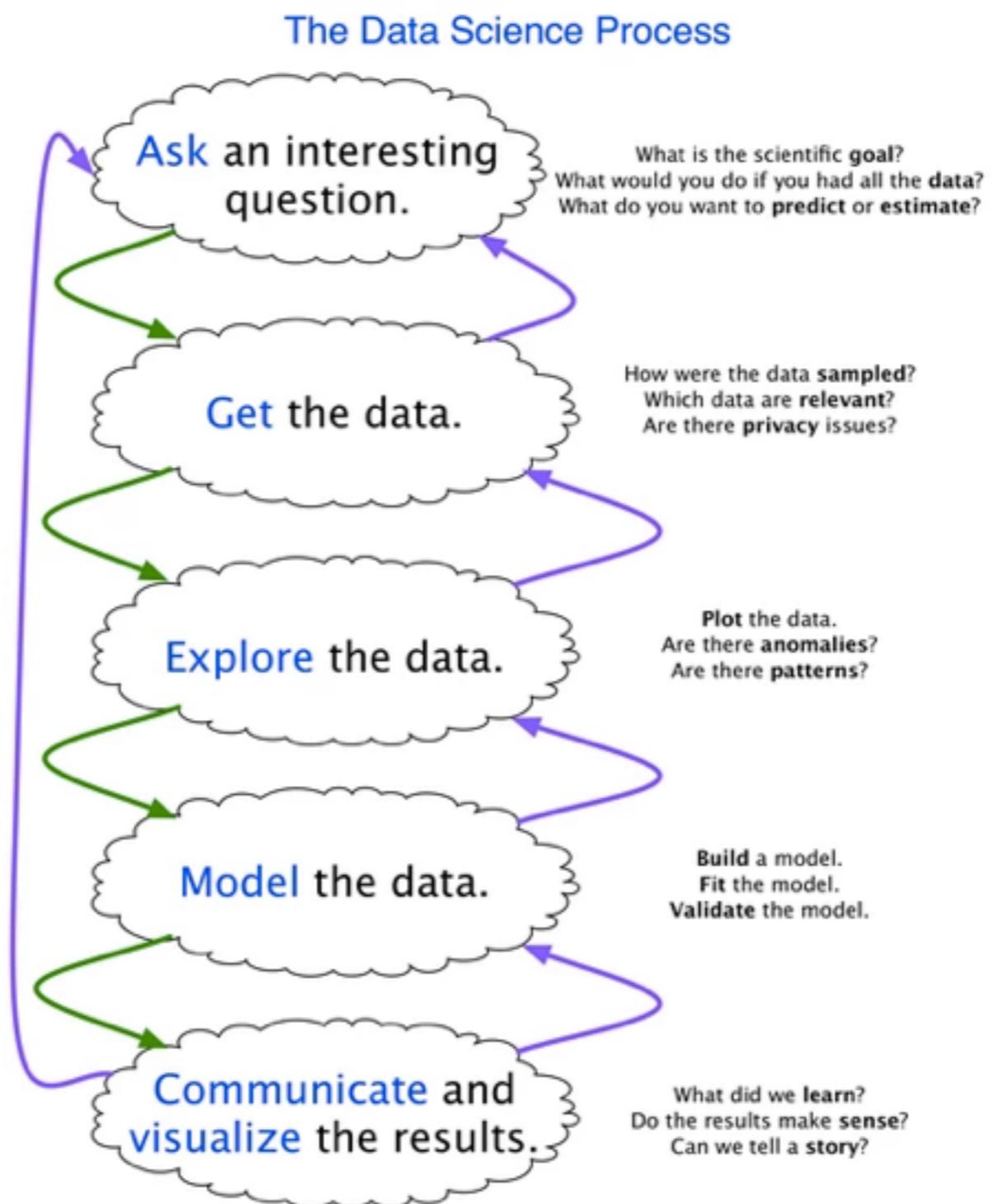
Computer Science
Mathematics

Data Wrangling
Machine Learning

Data Visualisation
Communication

Statistics

The Process



Storytelling

5 LOOM DATA Storytelling
THE ART AND SCIENCE OF SOCIAL MEDIA METRICS

PACKAGE Never force your data into a package--package your story from the data. The way you apply final data visualization and craft a modern data story needs to be flexible and agile like your strategy. Less data, more story. Less linear, more actionable. Less data showing-off and more zeroing in on opportunity. Modern social media data storytelling is all about adjustments--and a well-packaged data story will guide you to maximum results.

CONTEXT One of the most important and often forgotten layers of storytelling is context. Context isn't just about back-patting as you surpass your competitor in weekly virality. It's about answering the question "so what?" Every report should include competitive, industry and aspirational benchmarks but moreover, by applying context early in the process, you may uncover new stories from this layer of data.

CAPTURE Capture all--report on less. All social data is important data, but not just for reporting--for shaping and informing real-time campaign adjustments. Since social media is about directly communicating with your audiences--understanding what your users tell you via data is a critical part of serving them better.

ANALYZE Reams of unusable data need some visualization before they can be analyzed. Utilize multiple types of comparative charts, grids, and visuals to help give data a storytelling boost. Don't be afraid to uncover a story that requires you to drill back down and capture new data. The best analysis happens fluidly.

Ask Good Questions

“Data science is all about asking questions. You engage in it whenever you interactively and iteratively search for deep, hidden patterns.” - James Kobielski

- Do people have more phones than toilets?
- How is Ebola spreading?
- Is using wood fires sustainable here?
- Can we feed 9 billion people?

(Simple, Actionable, Incremental)

Get the data

← → C docs.hdx.rwlabs.org ⭐ ⌂ ⌂

HDX Beta DATA | COUNTRIES | ORGANISATIONS | BLOG | CONTACT | ABOUT SUBMIT DATA

Introducing the Humanitarian Data Exchange

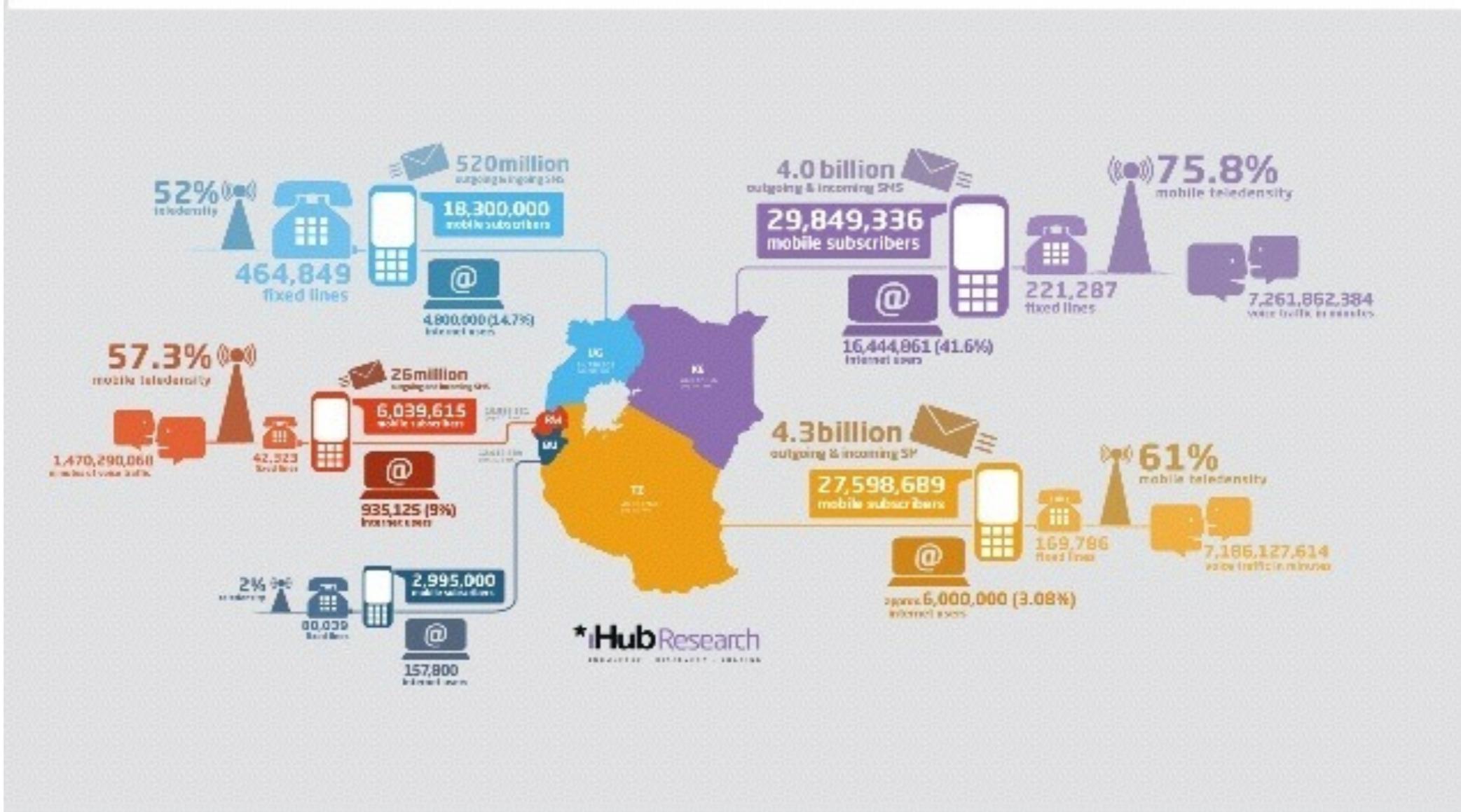
Introducing the Humanitarian Data Exchange: animation

An Aside: Big Data



Volume

Mobile Stats in East Africa

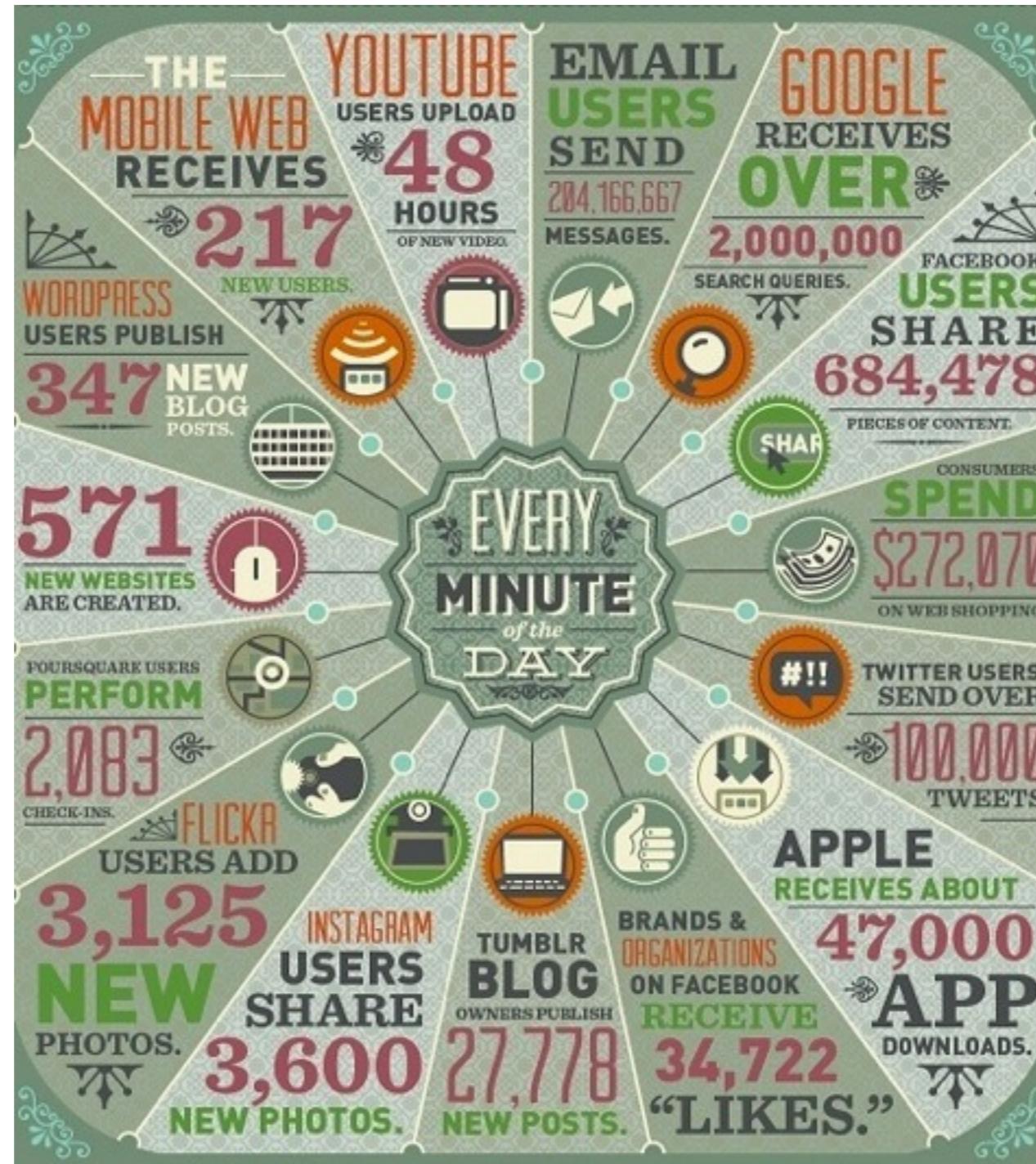


SOURCE

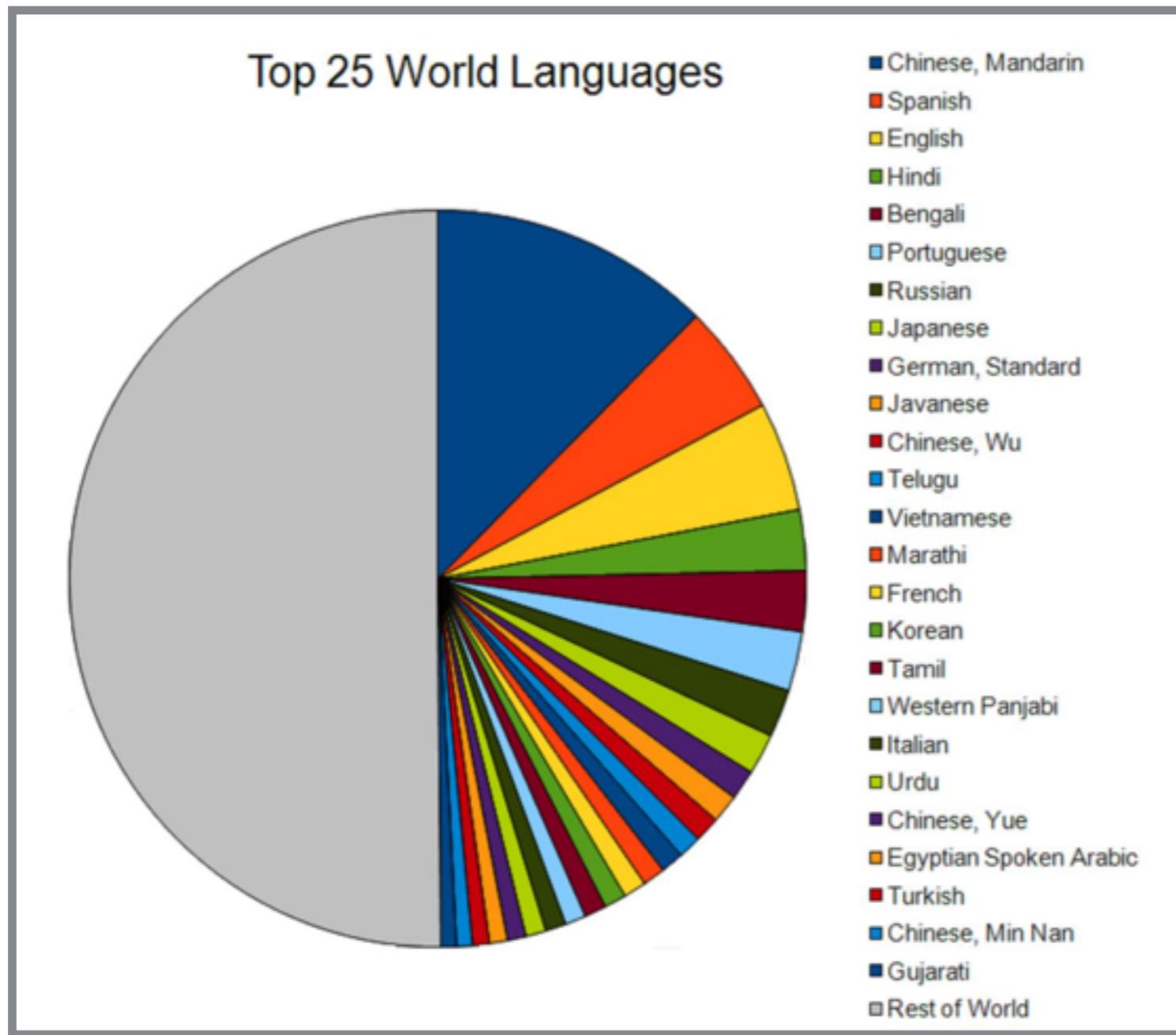
1. Data from <http://www.itu.int/ITU-D/ict/statistics/> dated 21.10.2013 & 2013 ITC World Telecommunication Indicators, including the latest available data for East African countries.
2. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Kenya.
3. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Uganda.
4. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Tanzania.
5. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Rwanda.
6. Data from <http://www.itu.int/ITU-D/ict/statistics/> dated 21.10.2013 & 2013 ITC World Telecommunication Indicators, including the latest available data for East African countries.
7. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Kenya.
8. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Uganda.
9. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Tanzania.
10. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Rwanda.

11. Data from <http://www.itu.int/ITU-D/ict/statistics/> dated 21.10.2013 & 2013 ITC World Telecommunication Indicators, including the latest available data for East African countries.
12. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Kenya.
13. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Uganda.
14. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Tanzania.
15. ITC Data Report - East Africa, December 2013, p. 273. It is estimated that there are 1.47 million mobile subscribers in Rwanda.

Velocity



Variety



CSV, json, xml,
excel, pdf,
text,
webpages, rss,
scanned pages,
images, videos,
audiofiles,
maps,
proprietary
formats etc.

Veracity



The “3 other Vs” are Viability, Validation, Verification.

Validation, checking that inputs are real, is a big deal for development data.

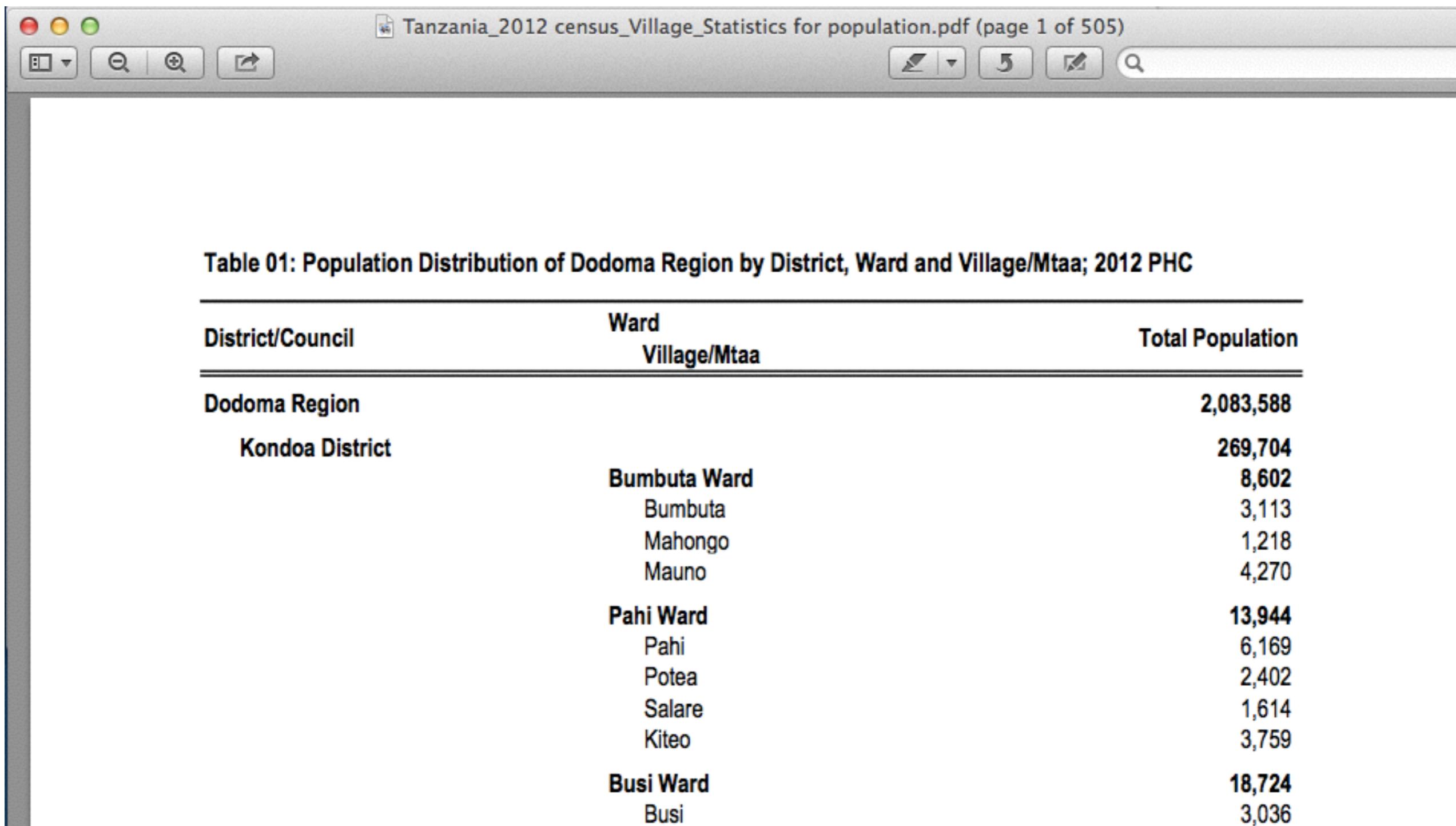
Getting more data



Explore your Data

- Spend time with your dataset:
 - Understand where it came from - can you live with the assumptions the data collectors made?
 - Look at it
 - Plot it
 - Where are there holes? Inconsistencies? Anomalies?
 - Clean your data, find better datasets, get more data

Data is Often Inconvenient



The screenshot shows a PDF document titled "Tanzania_2012 census_Village_Statistics for population.pdf (page 1 of 505)". The document contains a table titled "Table 01: Population Distribution of Dodoma Region by District, Ward and Village/Mtaa; 2012 PHC".

District/Council	Ward Village/Mtaa	Total Population
Dodoma Region		2,083,588
Kondoa District		269,704
	Bumbuta Ward	8,602
	Bumbuta	3,113
	Mahongo	1,218
	Mauno	4,270
	Pahi Ward	13,944
	Pahi	6,169
	Potea	2,402
	Salare	1,614
	Kiteo	3,759
	Busi Ward	18,724
	Busi	3,036

(plus often sitting on someone's laptop)

Data is (almost) Always Dirty

DR Congo in [data.UN.org](#): “Congo, Democratic Republic of the”, “Congo Democratic”, “Democratic Republic of the Congo”, “Congo (Democratic Republic of the)”, “Congo, Dem. Rep.”, “Congo Dem. Rep.”, “Congo, Democratic Republic of”, “Dem. Rep. of Congo”, “Dem. Rep. of the Congo”

DR Congo in common standards: “Democratic Republic of the Congo” (UN Stats), “Congo, The Democratic Republic of the” (ISO3166), “Congo, Democratic Republic of the” (FIPS10, Stanag), “180” (UN Stats), “COD” (ISO3166, Stanag), “CG” (FIPS10)

Use multiple datasets

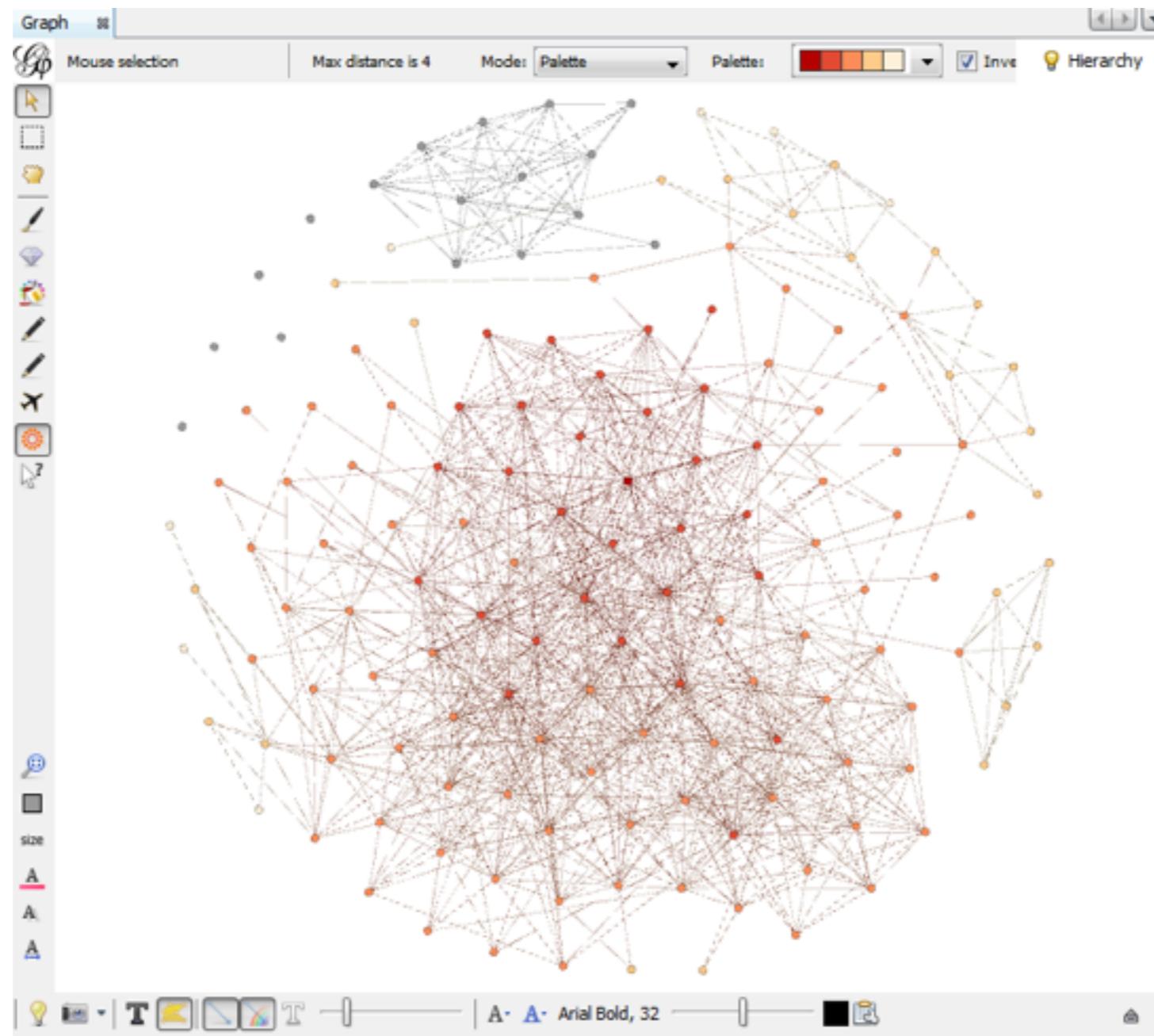
census	Arusha	Arusha	Daraja 2
shapefile	Arusha	Arusha Urban	Daraja Mbili
shapefile	Arusha	Ngorongoro	Endulen
census	Arusha	Ngorongoro	Enduleni
shapefile	Arusha	Ngorongoro	Engusero Sambu
census	Arusha	Ngorongoro	Enguserosambu
shapefile	Arusha	Longido	Gelai lumbwa
census	Arusha	Longido	Gelai Lumbwa
census	Arusha	Longido	Ketumbeine
shapefile	Arusha	Longido	Kitumbeine
census	Arusha	Arusha	Levolos
shapefile	Arusha	Arusha Urban	Levolosi

Process your data



Everything is a dataset (if you look hard enough)

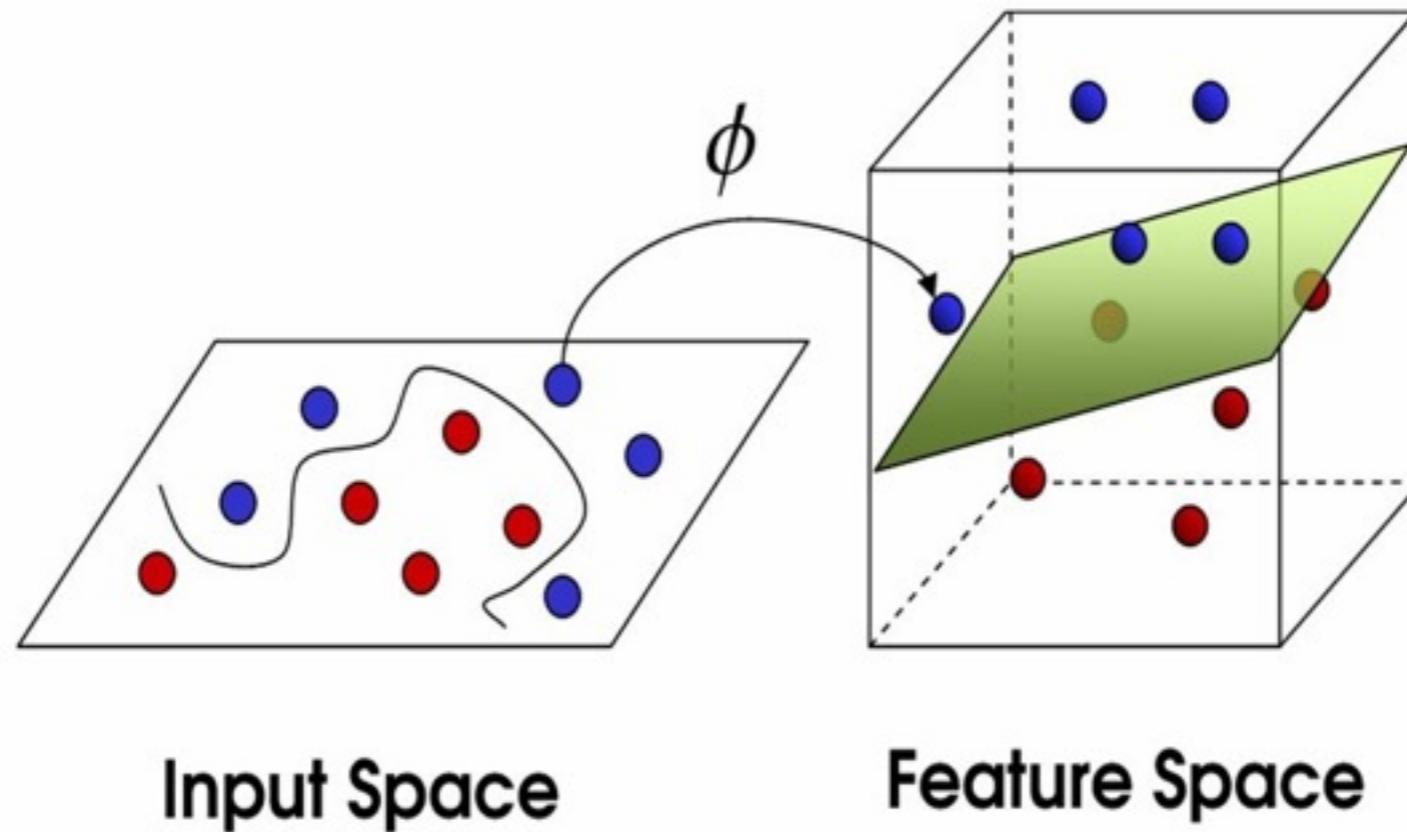
Process your data



(The relationships between things are interesting
- these are my Facebook 'friends', on Gephi)

Process your data

Principle of Support Vector Machines (SVM)



(machine learning can be useful too,
e.g. if you're working in a language with no stopword lists)

Model your data

- You're persuading people with 'truths': do your best to make sure they're truthful
- Always cross-check
- Statistics is your friend

Explain your results

- You're trying to persuade people to change:
 - Their opinions
 - Their actions
- Visuals are (often) more persuasive:
 - "I already knew that increased incarceration didn't lower crime, but I wasn't sure of the statistics. To see it on the graphs is really eye opening." *

*: Pandey et al, The Persuasive Power of Data Visualisation

Tools

selection.datavisualization.ch

DATAVISUALIZATION.CH SELECTED TOOLS

All Maps Charts Data Color

Rickshaw
A library for creating interactive time series graphs based on D3.js.

SVG Crowbar
A bookmarklet that extracts SVG nodes from an HTML document into a SVG file.
<http://nytimes.github.io/svg-crowbar>

Sigma.js
An open-source lightweight library to display interactively static and dynamic graphs.

Tableau Public
A desktop application to build and post interactive graphs, dashboards, maps and tables to the web.

Tabula
A tool to extract CSV formatted data from text tables in PDF documents.

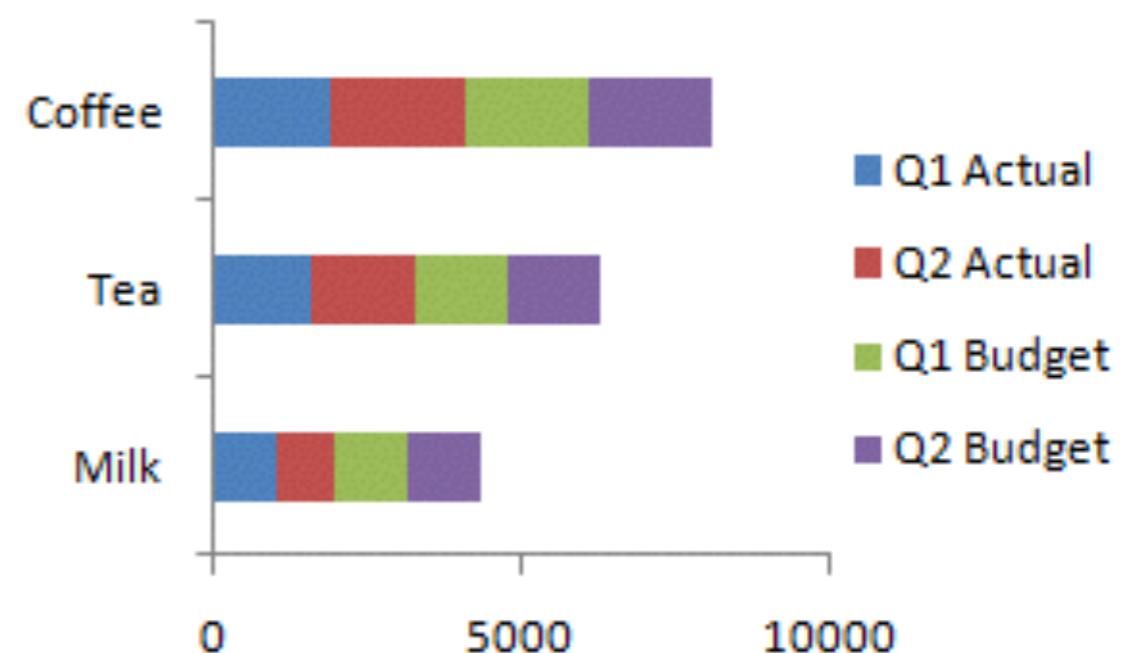
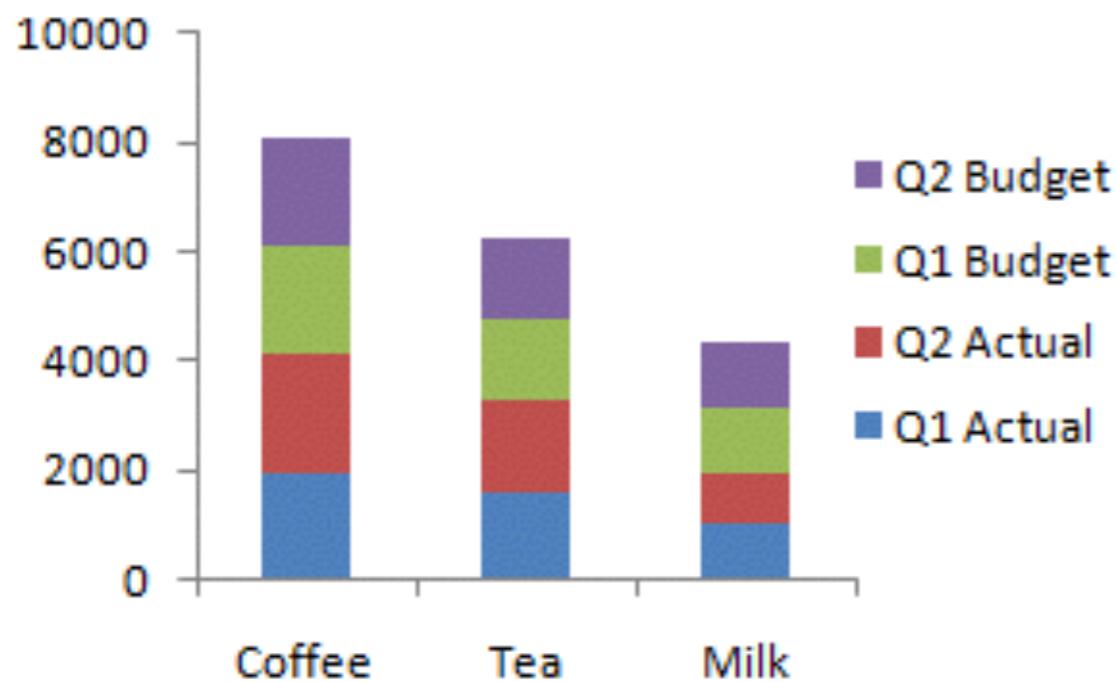
Tangle
A library that allows to interactively explore, play and see the document.

Timeline.js
A tool to create timelines with data and media from different sources like ENIAC, The Macintosh, etc.

Unfolding
A library to create interactive maps and mosaics in Processing.

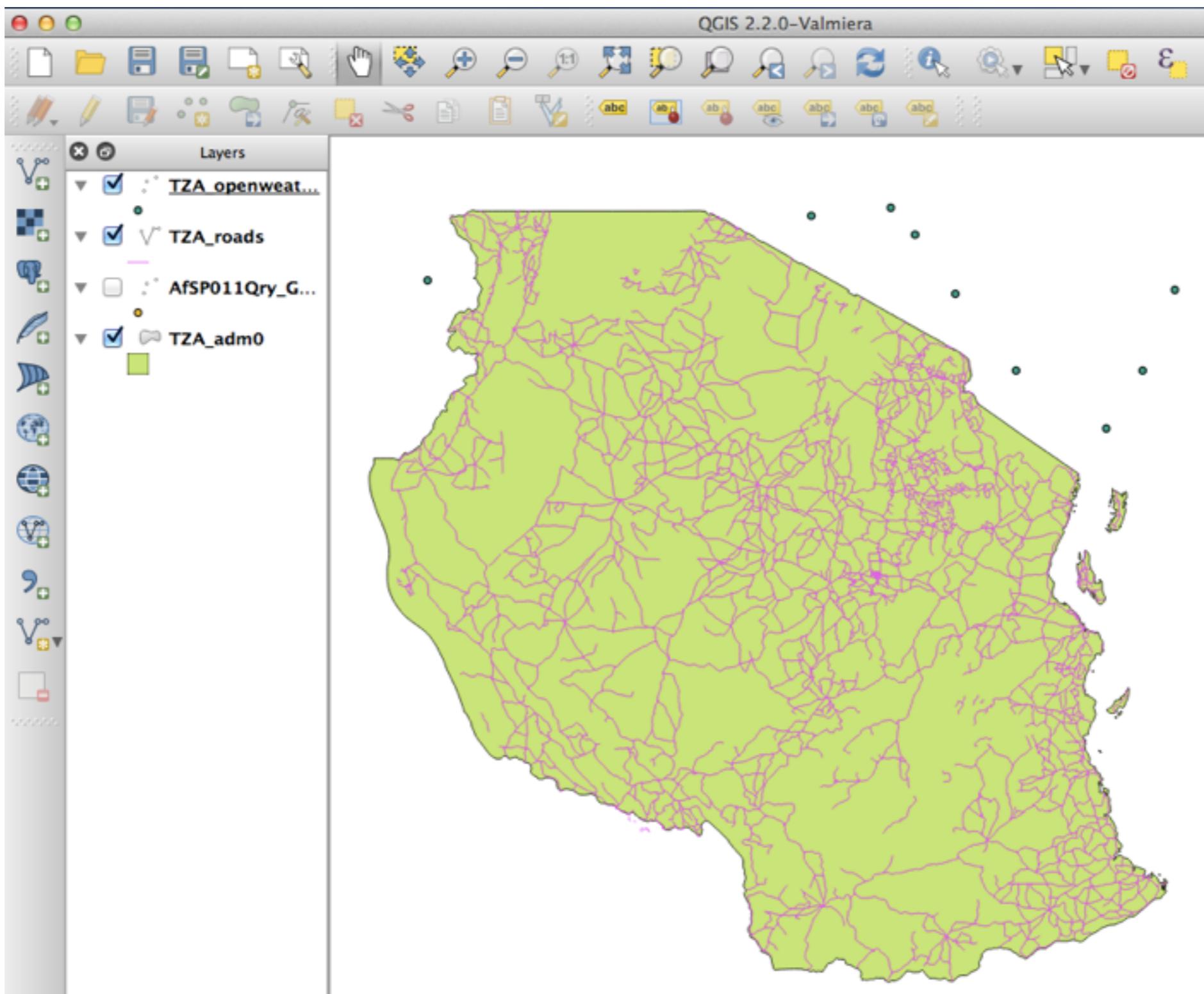
selection.datavisualization.ch

Tools: Excel

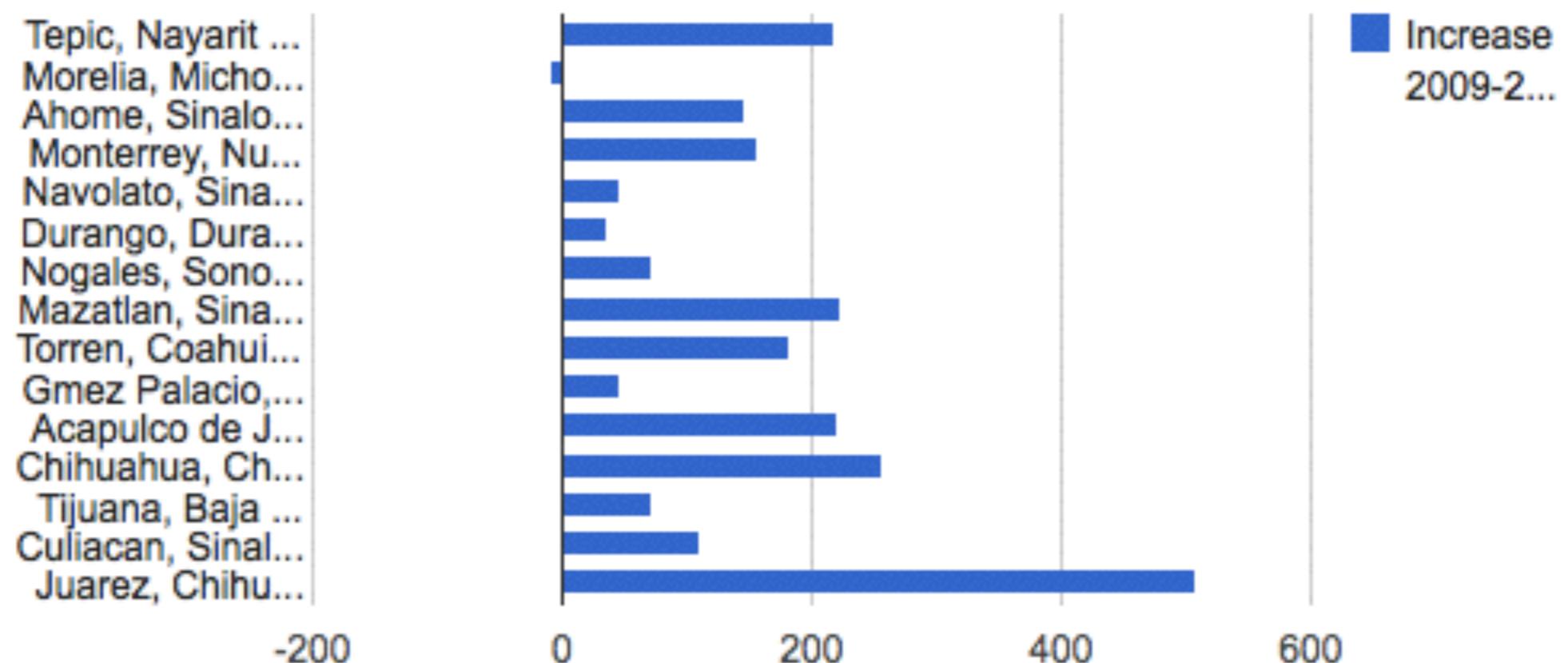


<http://peltiertech.com/clustered-stacked-column-bar-charts/>

Tools: QGIS

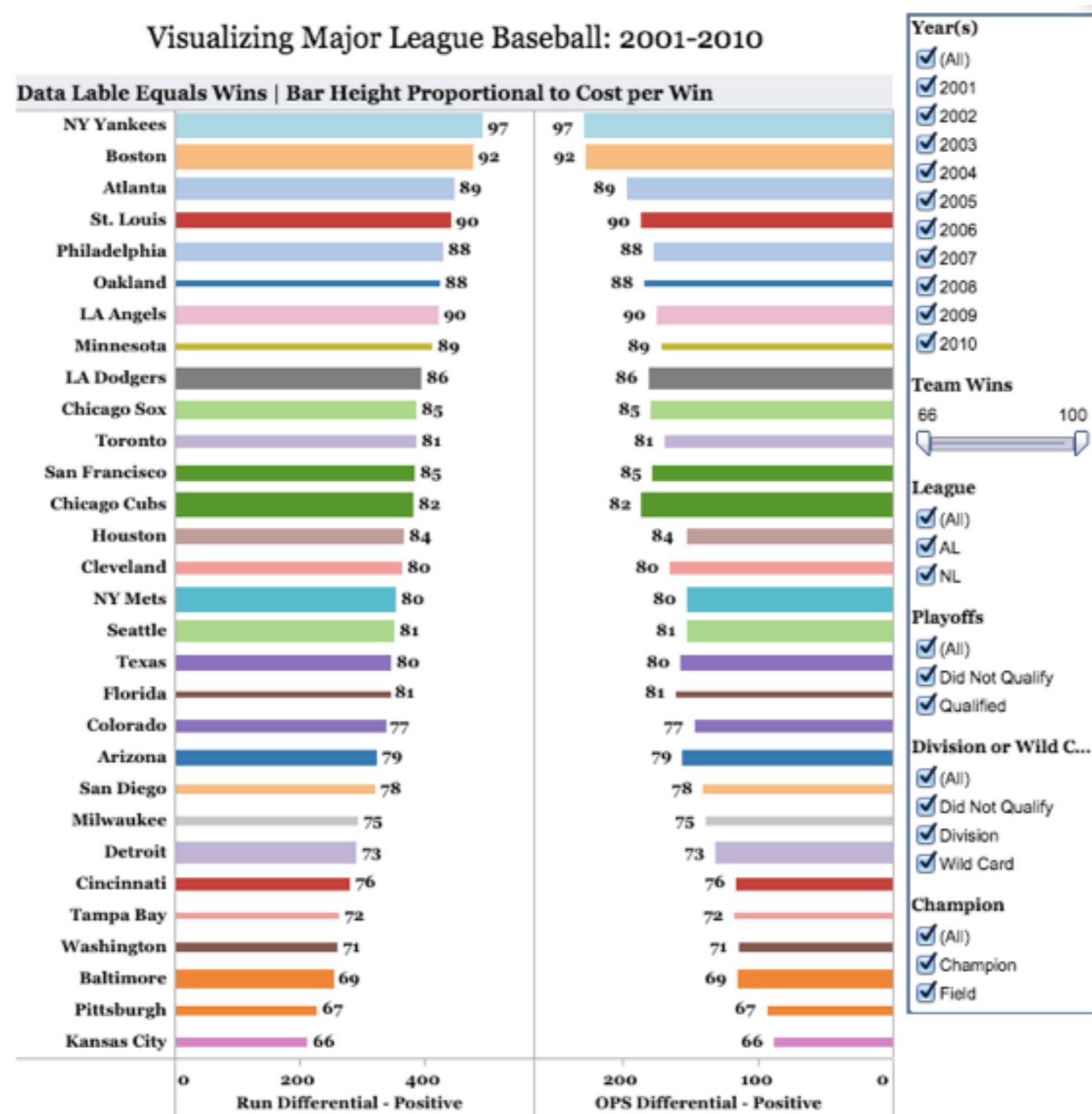


Tools: Google Fusion Tables

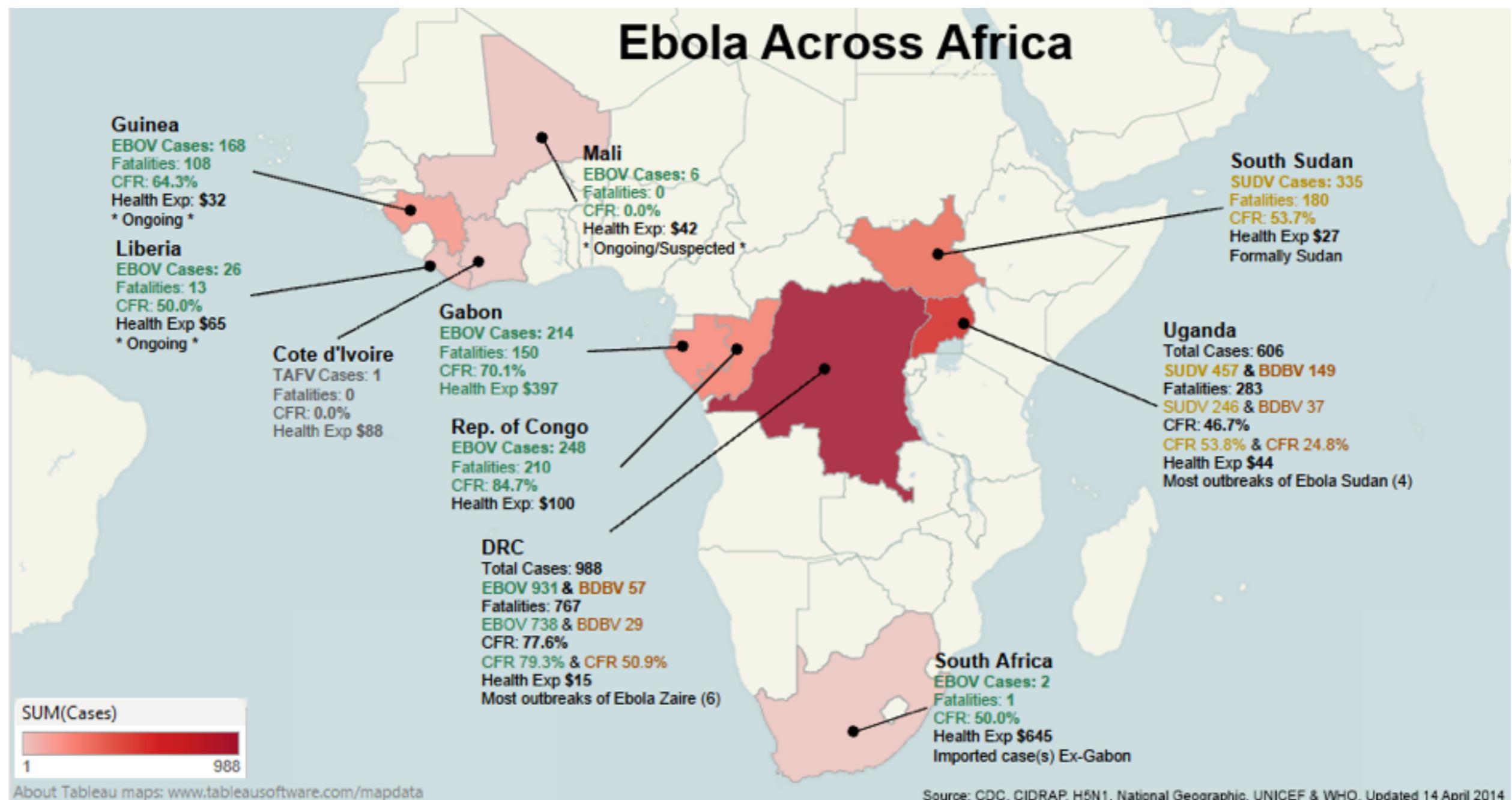


(link to google spreadsheets)

Tools: Tableau



Tools: Tableau



(this is a choropleth)

Tools: Python, R

worst	-3
worth	2
worthless	-2
worthy	2
wow	4
wowow	4
wowww	4
wrathful	-3
wreck	-2
wrong	-2
wronged	-2
wtf	-4
yeah	1
yearning	1
yeees	2
yes	1
youthful	2
yucky	-2
yummy	3
zealot	-2
zealots	-2
zealous	2

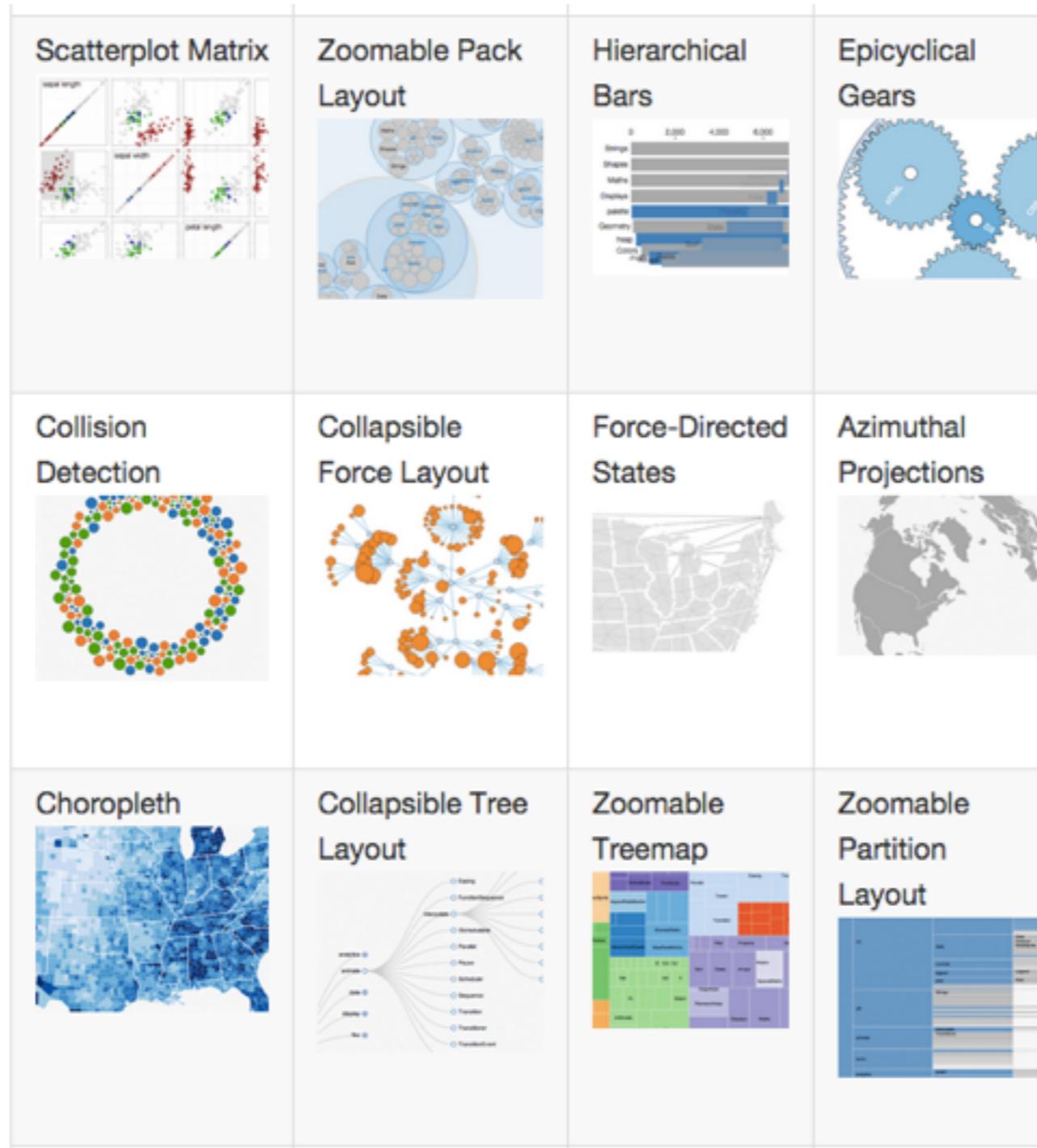
```
def getsentiments(fpSENT, fPTweet):
    #convert sentiment file to dictionary
    sents = getscores(fpSENT)

    #Readlines returns a list of strings...
    tweettext = []
    for line in fPTweet.readlines():
        jline = json.loads(line)
        if jline.has_key("text"):
            #tweettext.append(jline["text"])
            words = jline["text"].split()
            score = 0
            for word in words:
                if sents.has_key(word):
                    score += sents[word]
            print(float(score))

    return(tweettext)
```

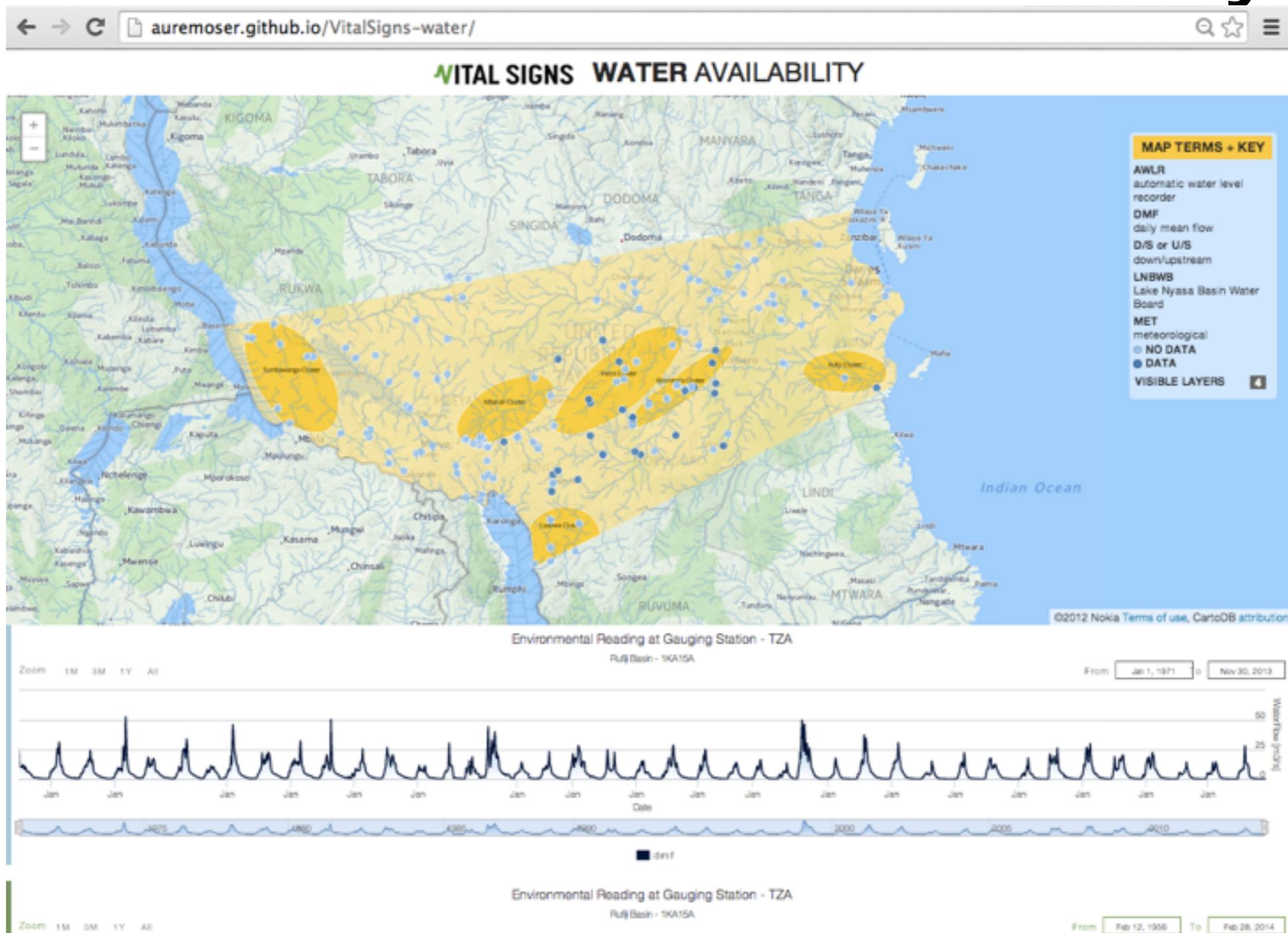
(AFINN sentiment analysis: sometimes you have to code)

Tools: D3, Javascript



D3 gallery: <https://github.com/mbostock/d3/wiki/Gallery>

D3: Interactive Play



auremoser.github.io/VitalSigns-water

What's that visualisation?

A PERIODIC TABLE OF VISUALIZATION METHODS

 Process
Visualization

Hy Structure Visualization

 Overview
 Detail

Detail AND Overview

< > Divergent thinking

> < **Convergent thinking**

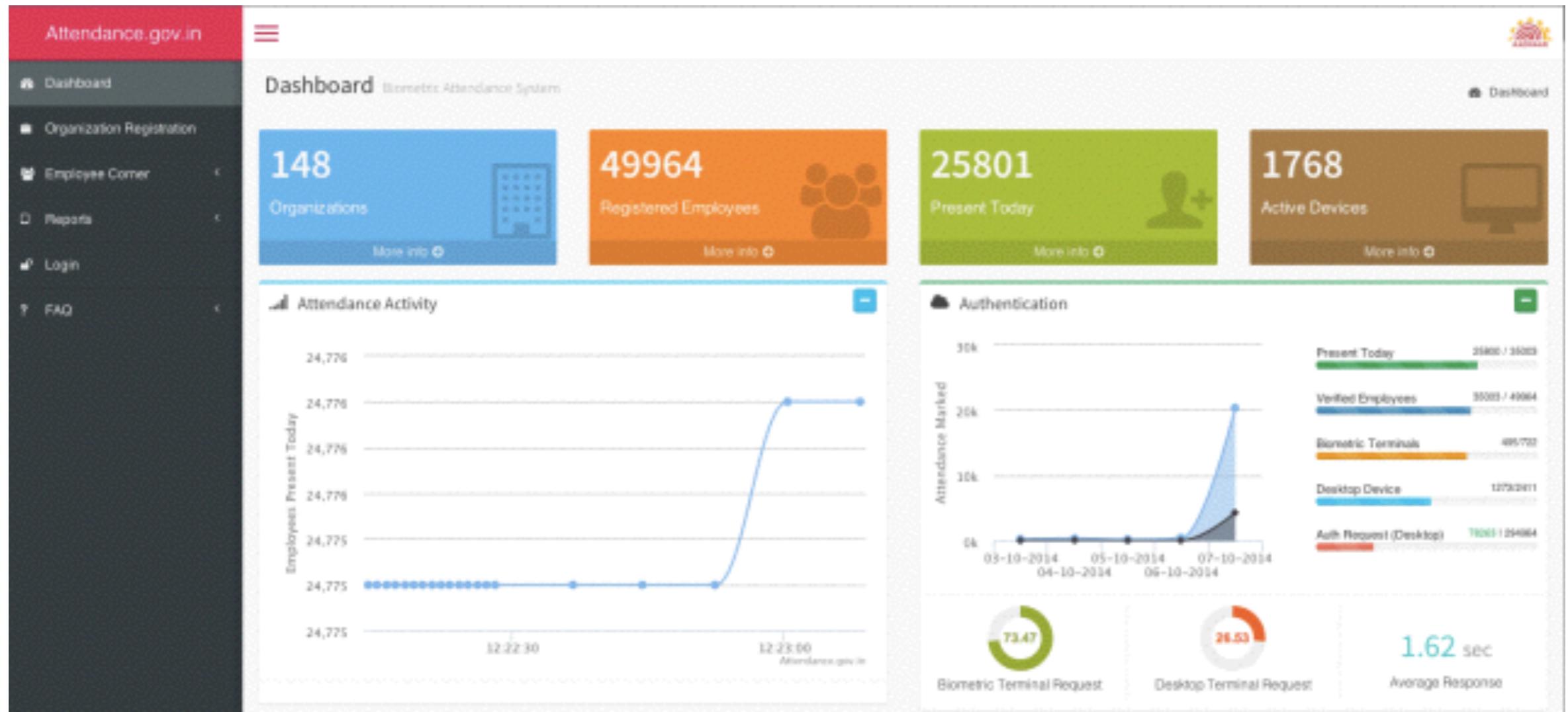
Note: Depending on your location and connection speed it can take some time to load a pop-up picture.

© Ralph Lengler & Martin J. Eppler: www.visual-literacy.org

version 1.5

> < Su supply demand curve	> < Pc performance charting	> < St strategy map	> < Oc organisation chart	< < Ho house of quality	> < Fd feedback diagram	□ Ft failure tree	> < Mq magic quadrant	> < Ld life-cycle diagram	> < Po porter's five forces	< < S s-cycle	> < Sm stakeholder map	○ Is ishikawa diagram	● Tc technology roadmap
 Ed edgeworth box	> < Pf portfolio diagram	 Sg strategic game board	> < Mz mintzberg's organograph	< < Z zwickly's morphological box	< < Ad affinity diagram	□ De decision discovery diagram	> < Bm big matrix	> < Stc strategy canvas	> < Vc value chain	< < Hy hyper-cycle	> < Sr stakeholder rating map	> < Ta taps	< < Sd spray diagram

What's a dashboard?

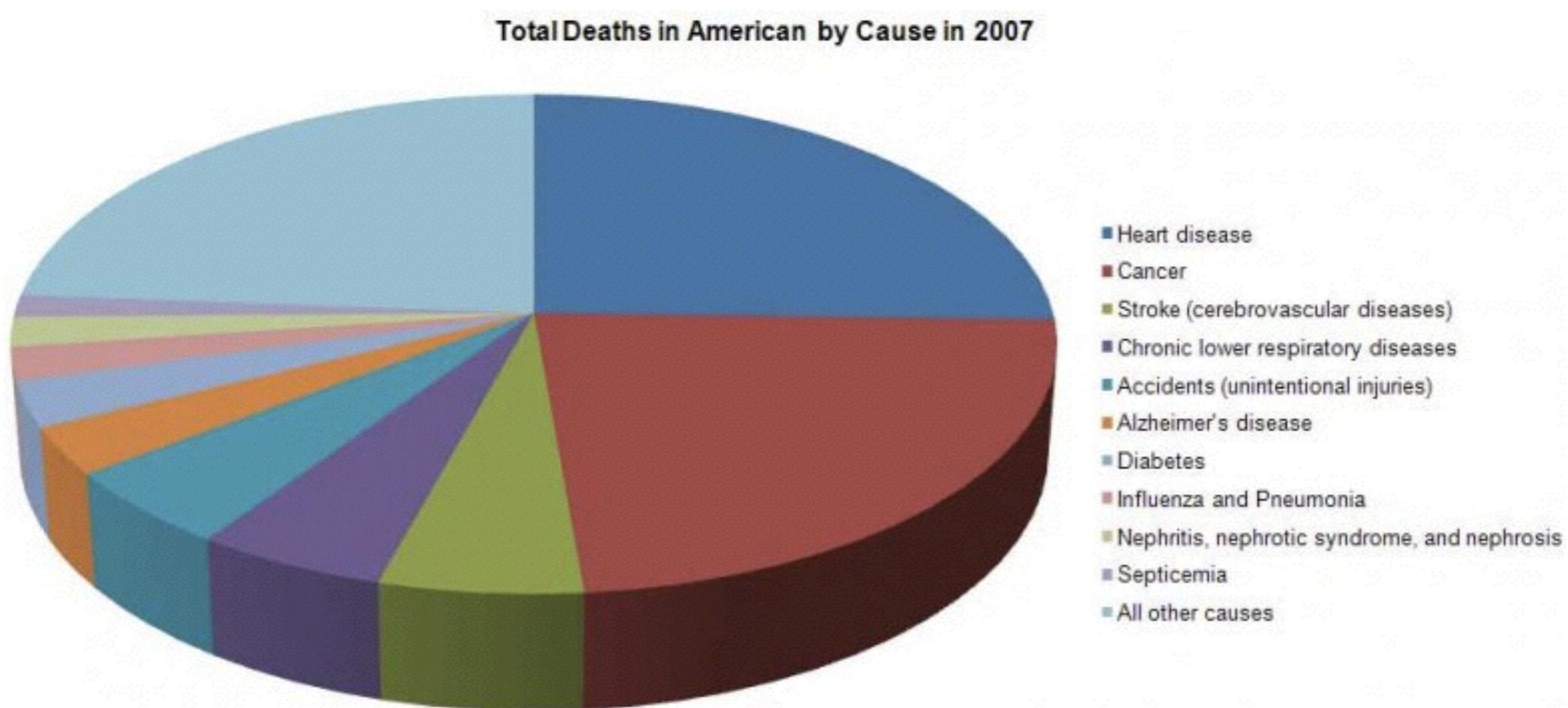


(India's government employee attendance system)

Psychology checklist

- What's your important message (and what are you trying to hide?)
- Medium matters (laptop, phone, sms?)
- Colours matter
- As do angles and relative lengths.
- And think about your audience, e.g. what local effects do you need to be aware of, how do you compensate for colorblindness etc etc.
- New visualization type? Check the Gestalt principles

Please don't do this...



Tools checklist

- Who's your audience?
- What's the medium: paper, static webpage, tablet, phone?
- Which languages do you need to display?
 - And are they right-to-left?
- Is this a one-off visualisation or will you need to update it as new data comes in?
- Are your audience viewing this online or offline?
- What resources do you have for updating the visualisation?

Where to go from here

- Websites, e.g. Information is Beautiful, DataScience Central, flowing data, ILoveCharts, Chart Porn, junk charts, visual.ly blog, fivethirtyeight.com
- Meetups and events, e.g. DataKind, NYC Data Skeptics
- Books e.g. Nathan Yao “Visualise this!”, anything by Tufte
- Spring course on data science

Ask good questions;
Tell good stories