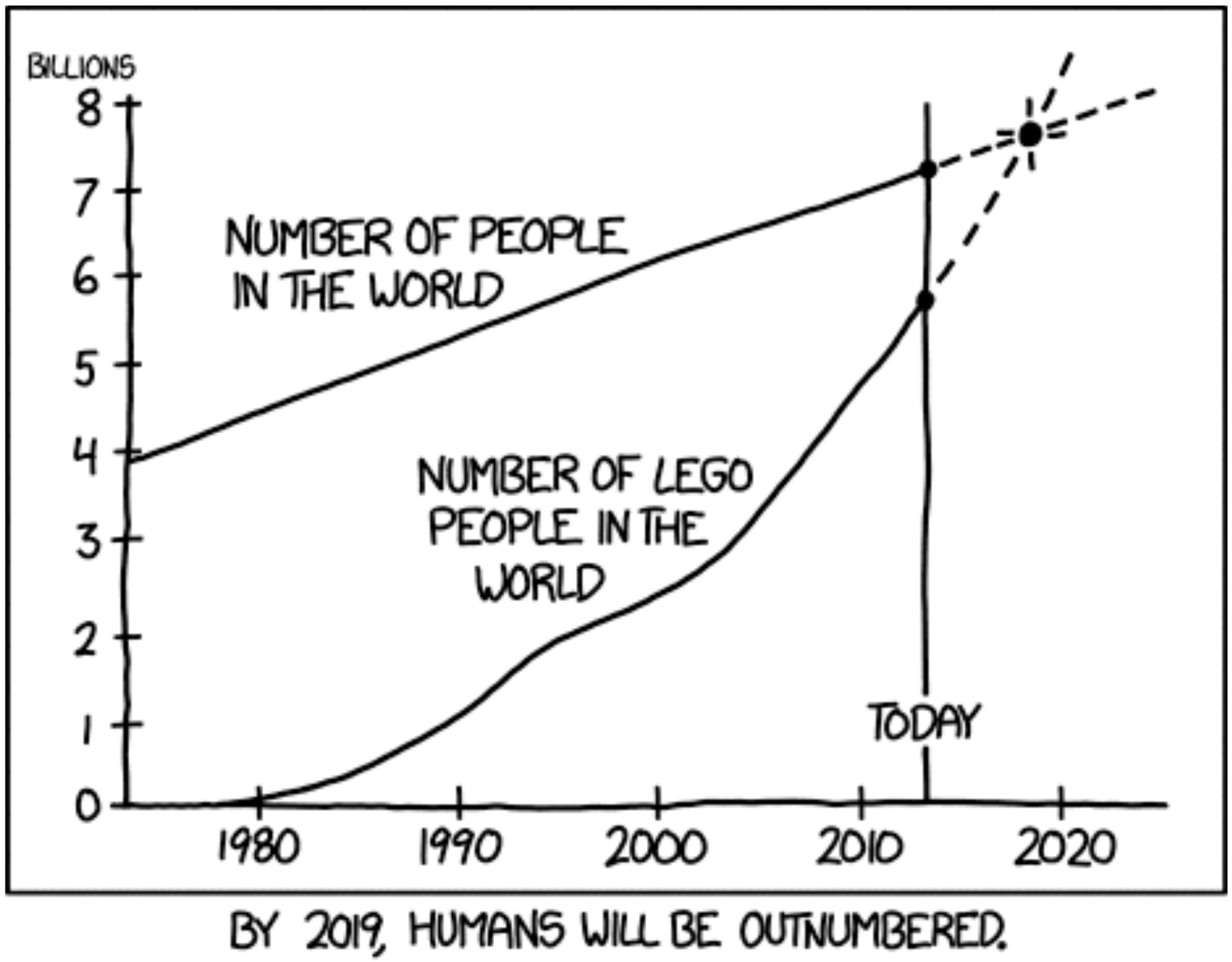
Data Visualisation for Beginners

Sara-Jayne Terp
2nd October 2014

Introduction



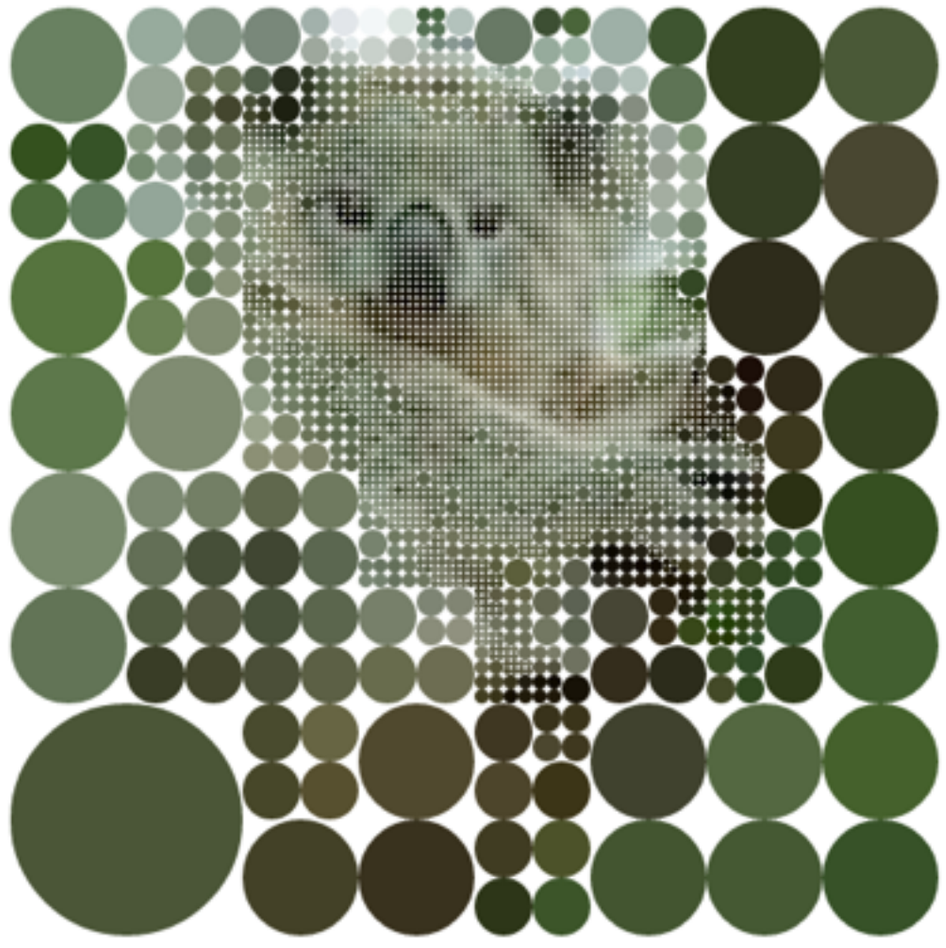
(XKCD.com: data science humor)

Talking About

Source: wordle.net

Not Talking About

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

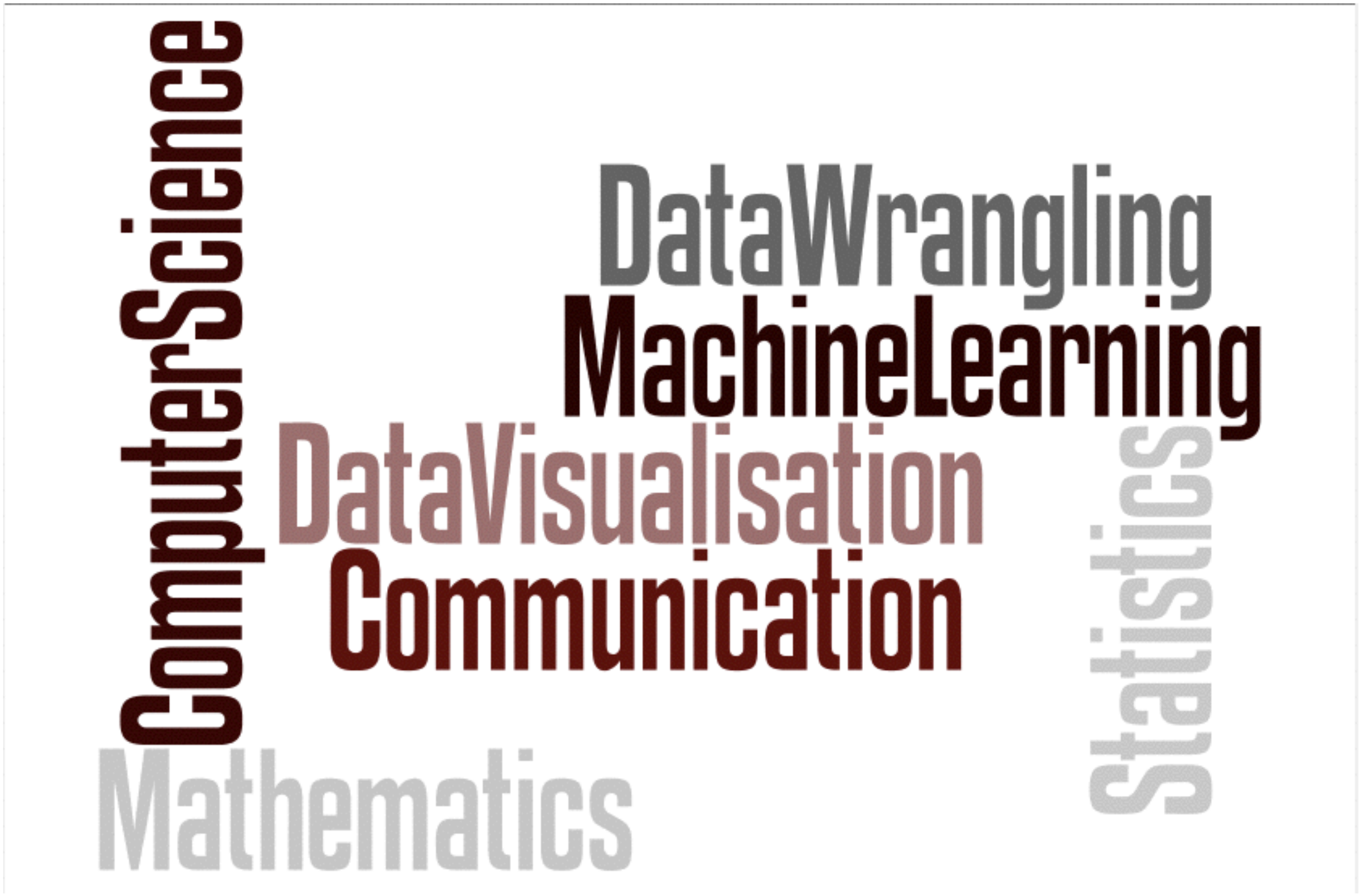


Cogs image from Worradmu on freedigitalphotos.net; Koala from koalastothemax.com

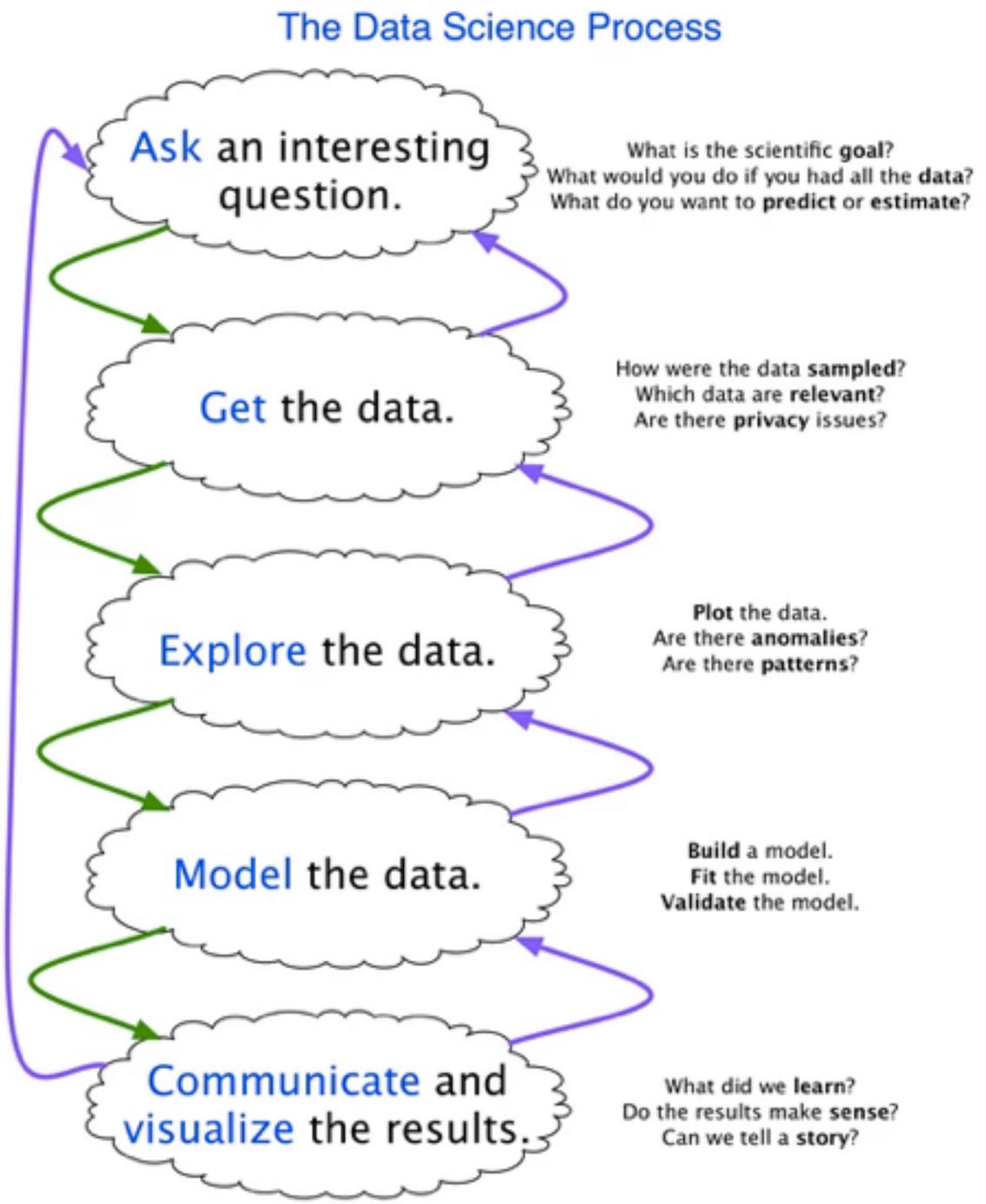
Data Visualisation

Images: Hans Rosling Ted Talk; erm.. something about India and women

Data Science



The Process


The Data Science Process
Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?
Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

Storytelling

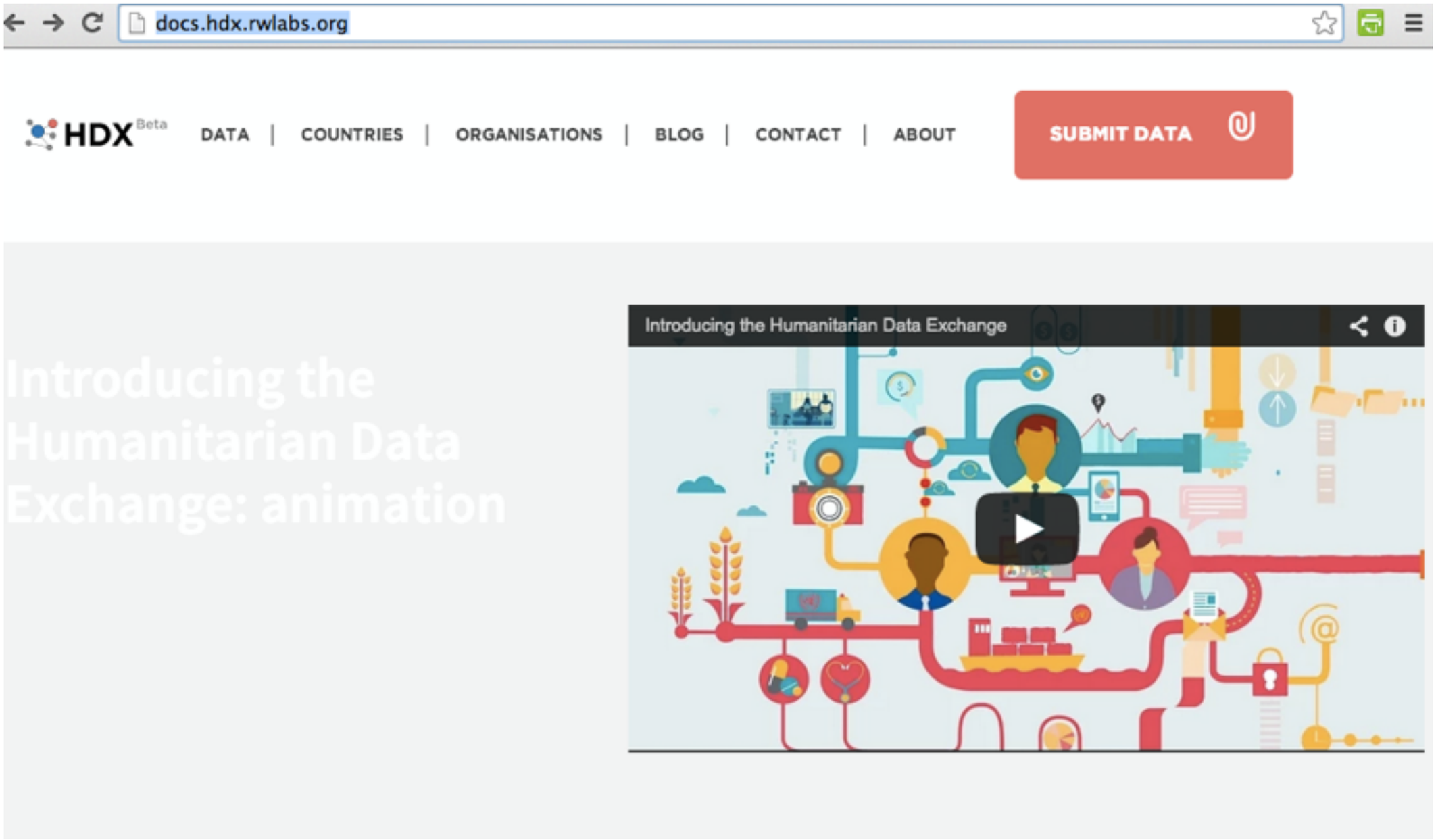
Ask Good Questions

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

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

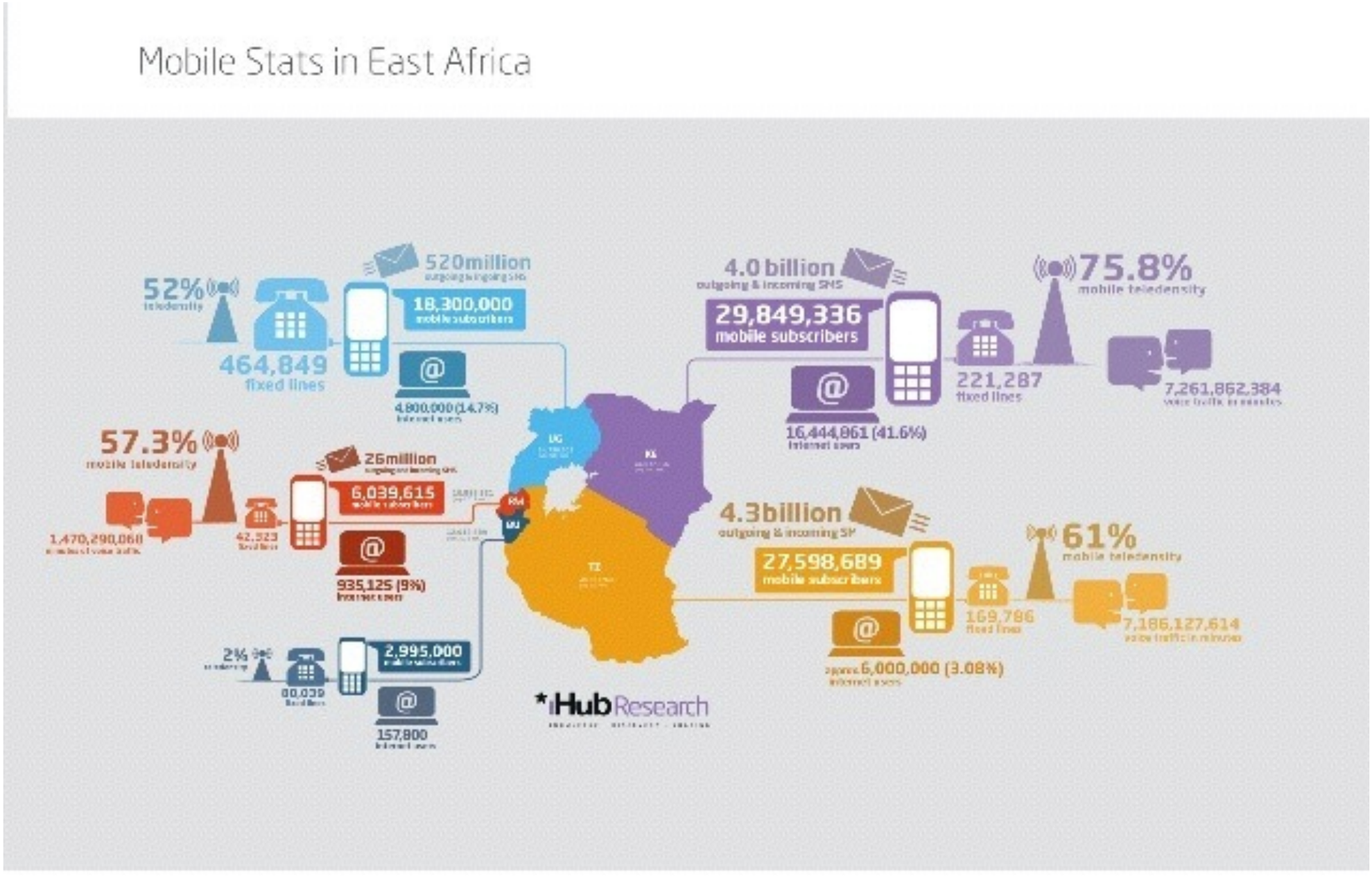
(Simple, Actionable,Incremental)

Get the data


docs.hdx.rwlabs.org
HDX Beta
DATA | COUNTRIES | ORGANISATIONS | BLOG | CONTACT | ABOUT
SUBMIT DATA
Introducing the Humanitarian Data Exchange: animation
Introducing the Humanitarian Data Exchange

An Aside: Big Data

Volume

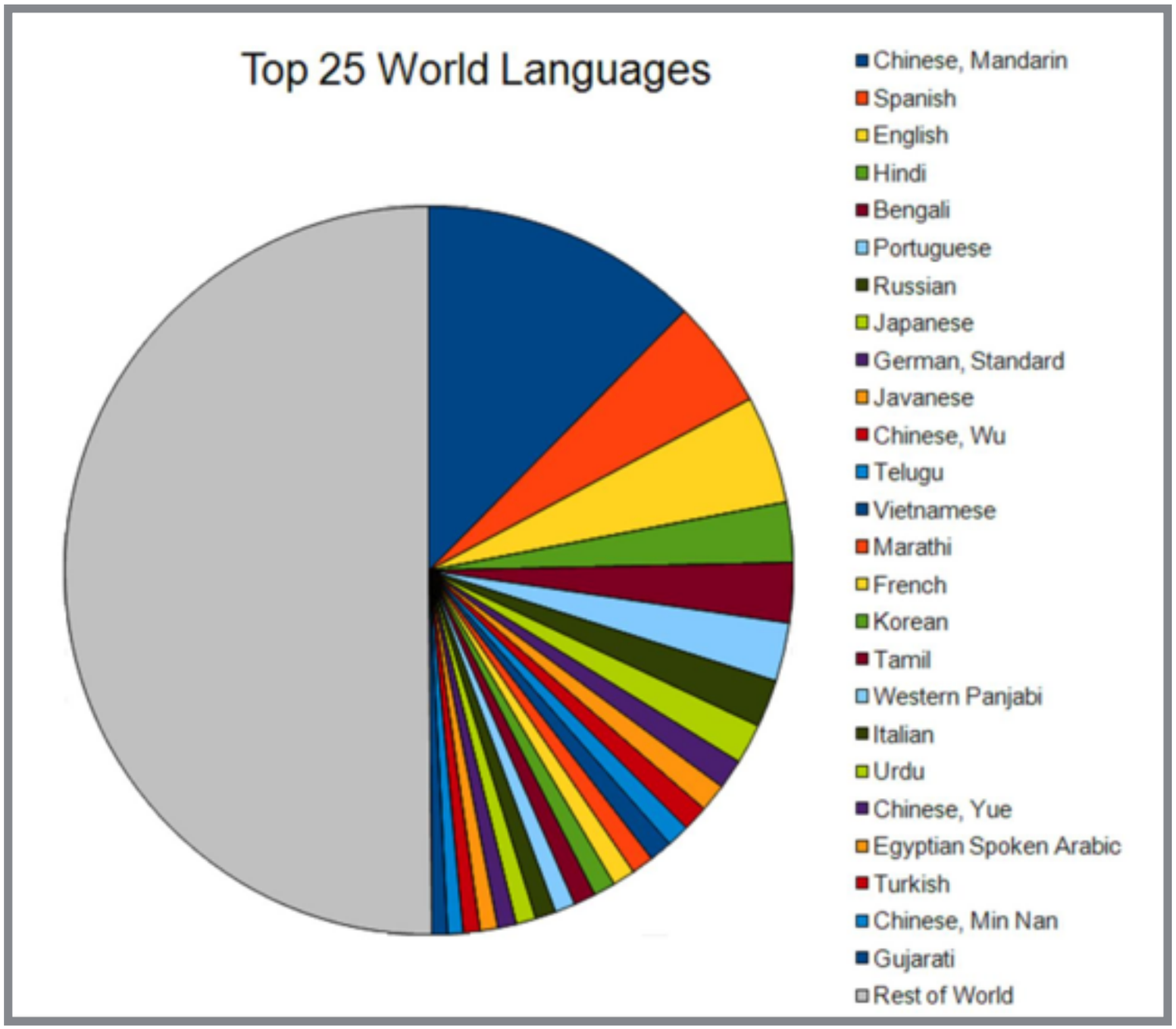


Sources:

Velocity


THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
571 NEW WEBSITES ARE CREATED.
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
EVERY MINUTE of the DAY
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKR USERS ADD 3,125 NEW PHOTOS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.

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

Tanzania_2012 census_Village_Statistics for population.pdf (page 1 of 505)

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

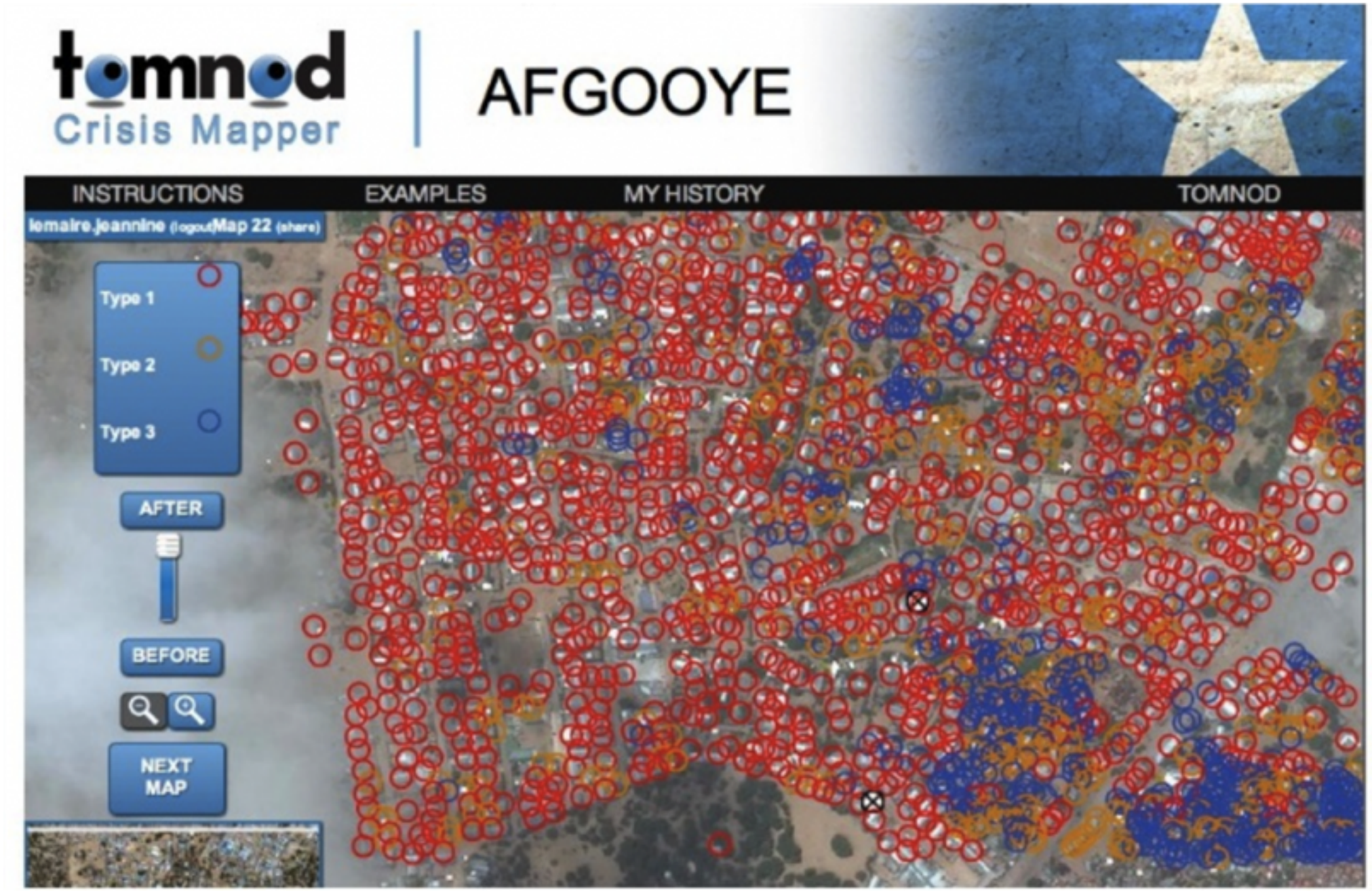
DR Congo in data.UN.org: “Congo, Democratic Republic of the”, “Congo Democratic”, “Democratic Republic of the Congo”, “Congo (Democratic Republic of the)”, “Congo, Dem. Rep.”, “Congo Dem. Rep.”, “Congo, Democratic Republic of”, “Dem. Rep. of Congo”, “Dem. Rep. of the Congo”

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

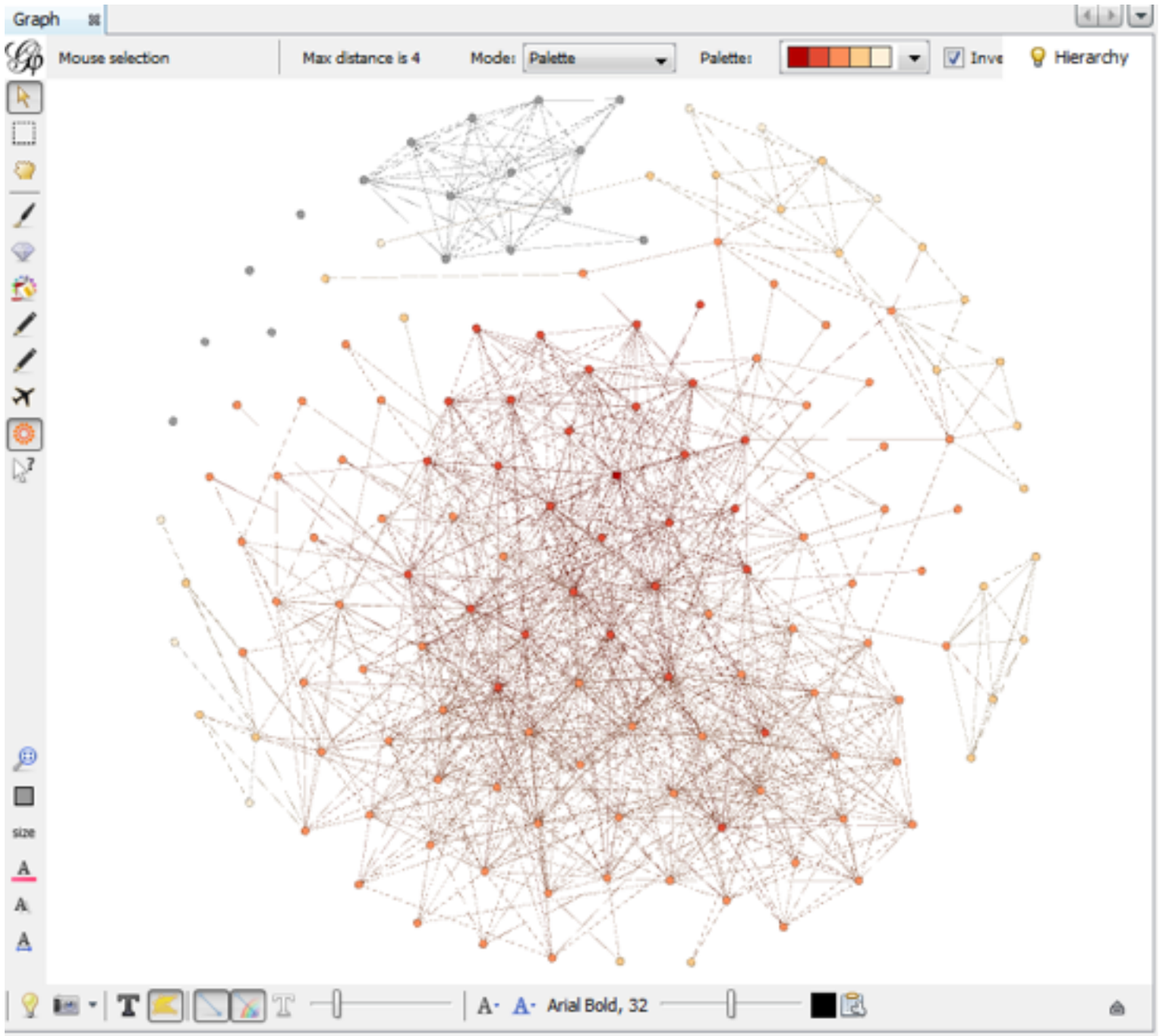
census	Arusha	Arusha	Daraja 2
shapefile	Arusha	Arusha Urban	Daraja Mbili
shapefile	Arusha	Ngorongoro	Endulen
census	Arusha	Ngorongoro	Enduleni
shapefile	Arusha	Ngorongoro	Engusero Sambu
census	Arusha	Ngorongoro	Enguserosambu
shapefile	Arusha	Longido	Gelai lumbwa
census	Arusha	Longido	Gelai Lumbwa
census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

Process your data



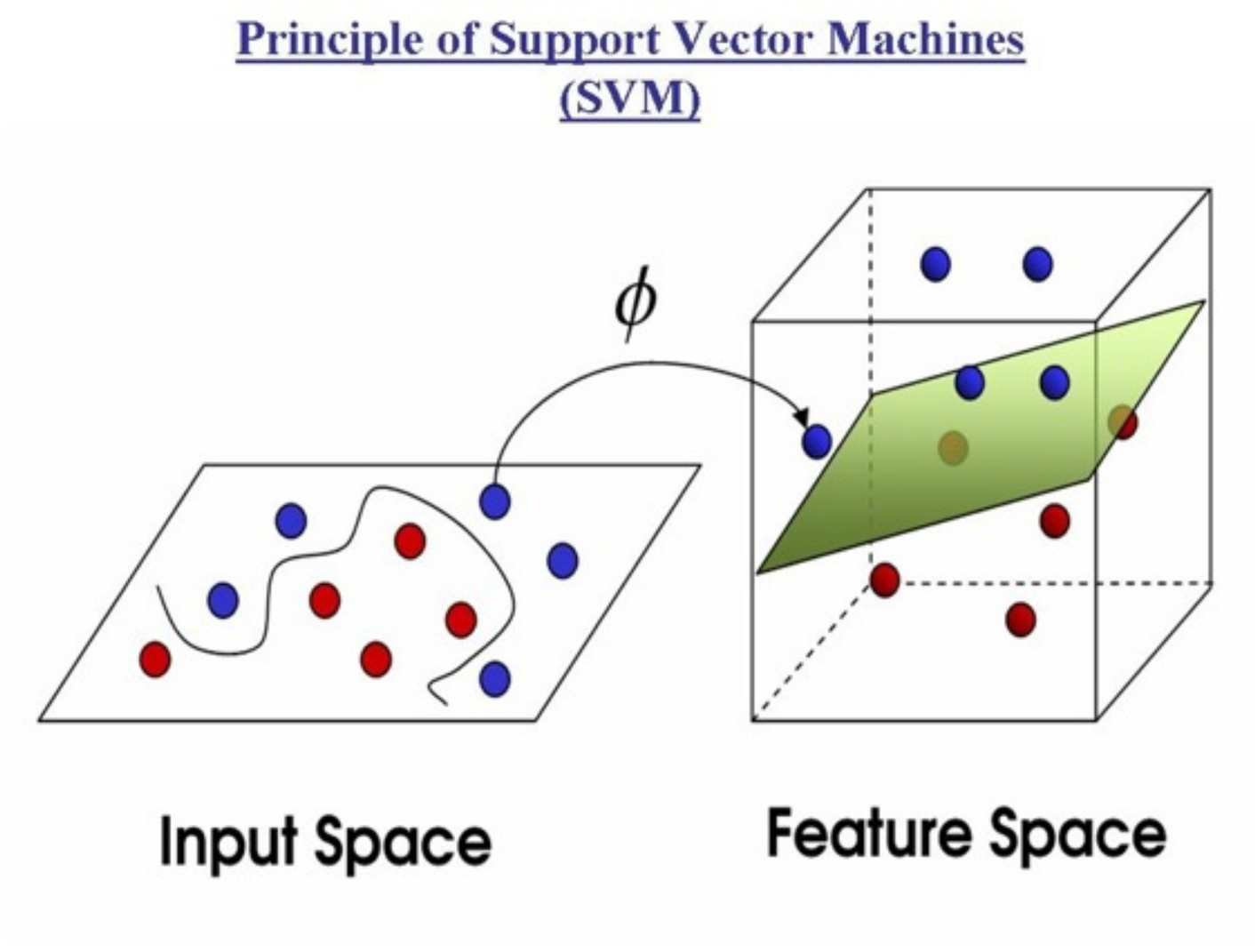
Everything is a dataset (if you look hard enough)

Process your data



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

Process your data



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

Model your data

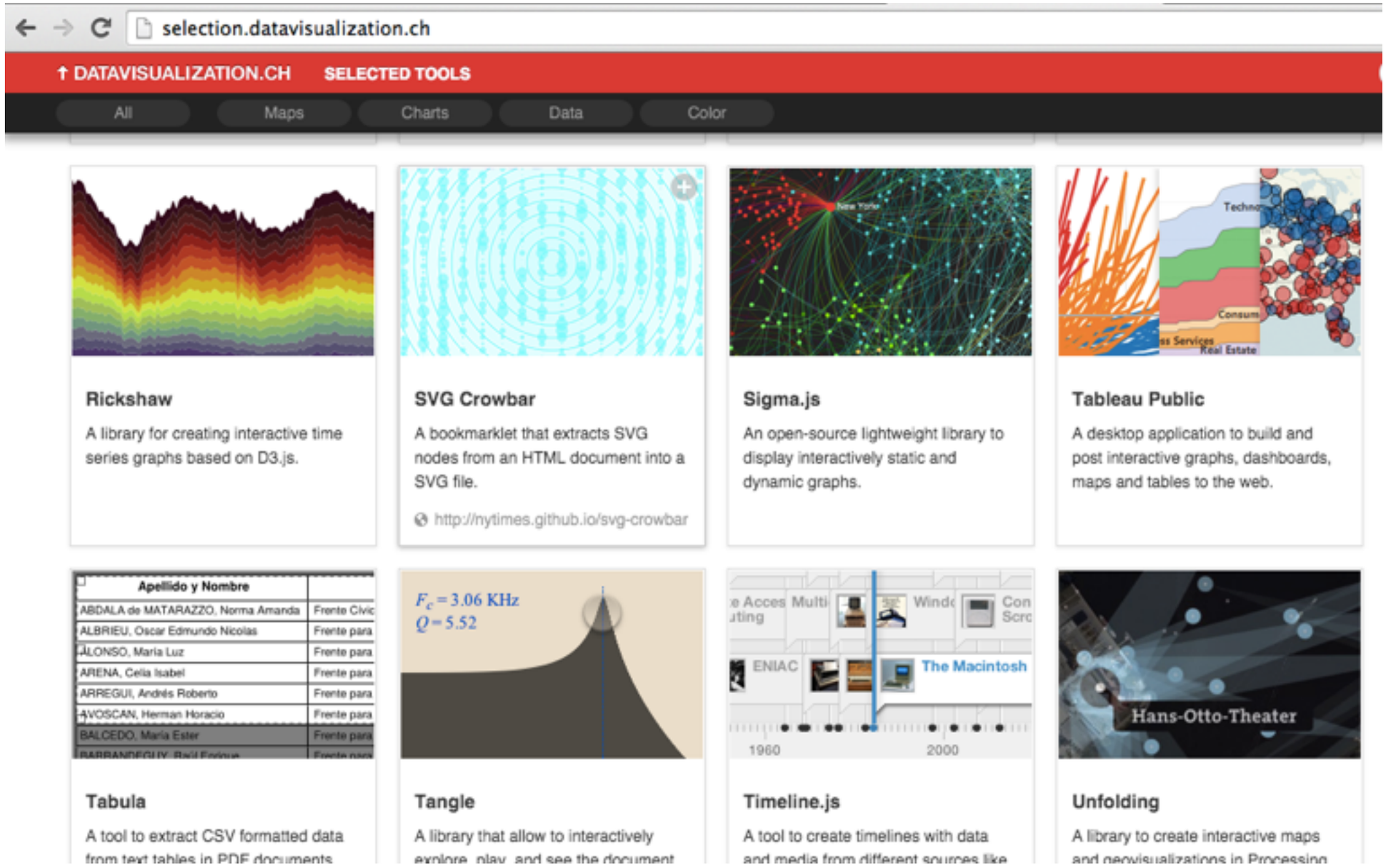
- You're persuading people with 'truths': do your best to make sure they're truthful
- Always cross-check
- Statistics is your friend

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.” *

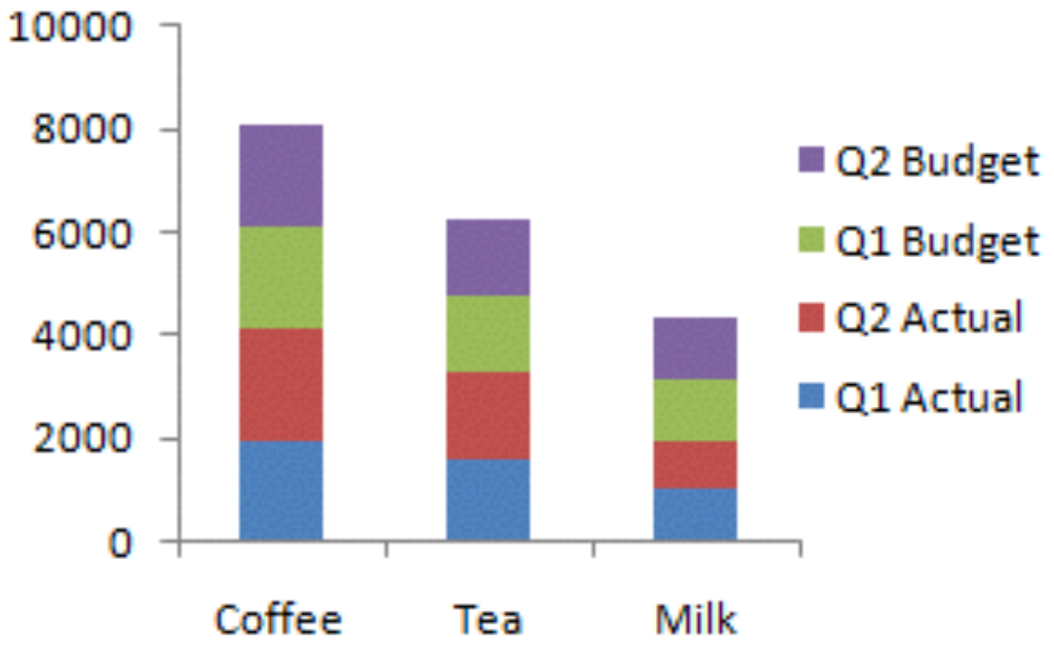
*: Pandey et al, The Persuasive Power of Data Visualisation

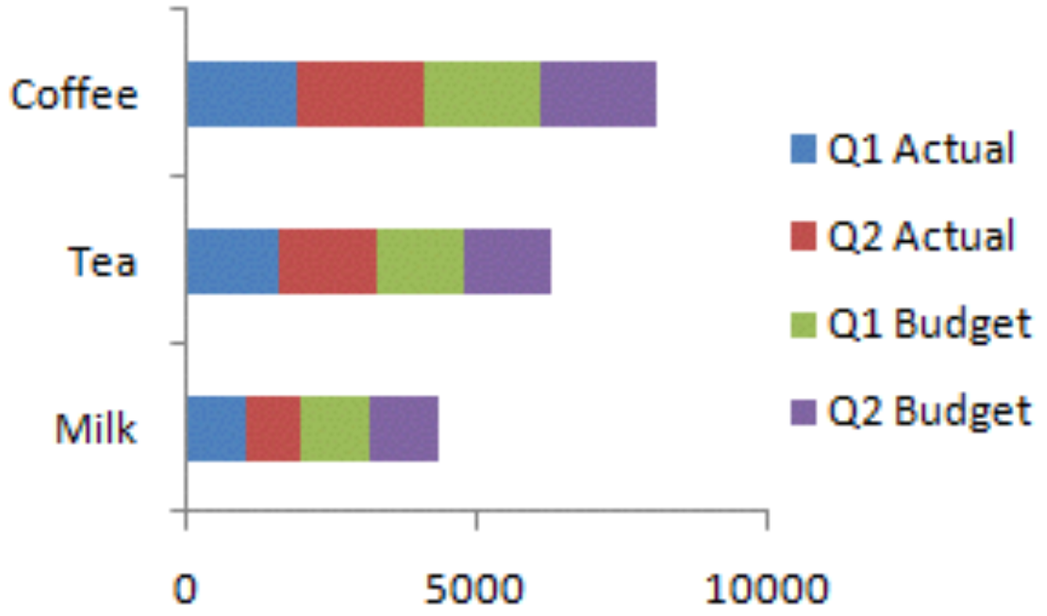
Tools



selection.datavisualization.ch

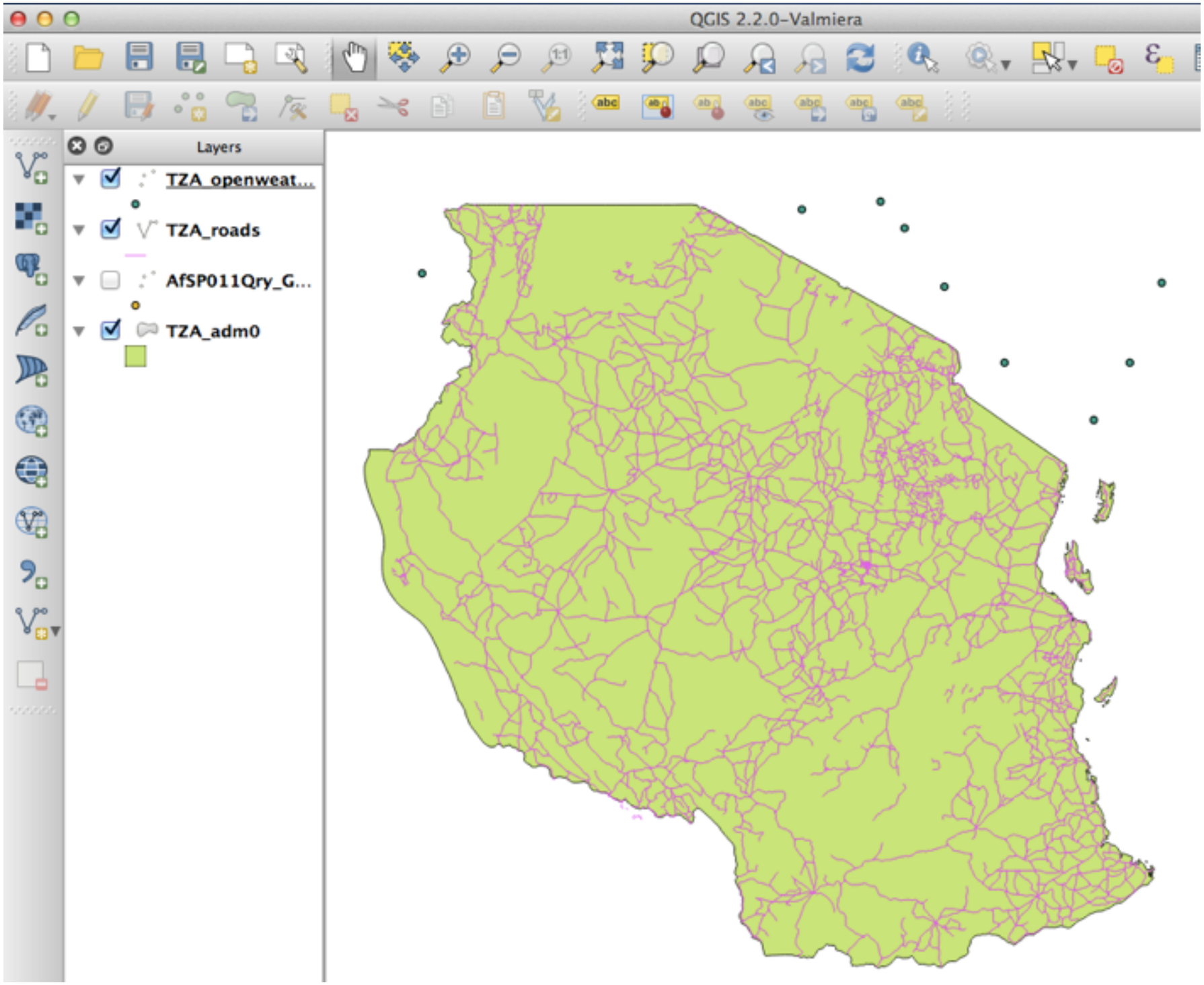
Tools: Excel



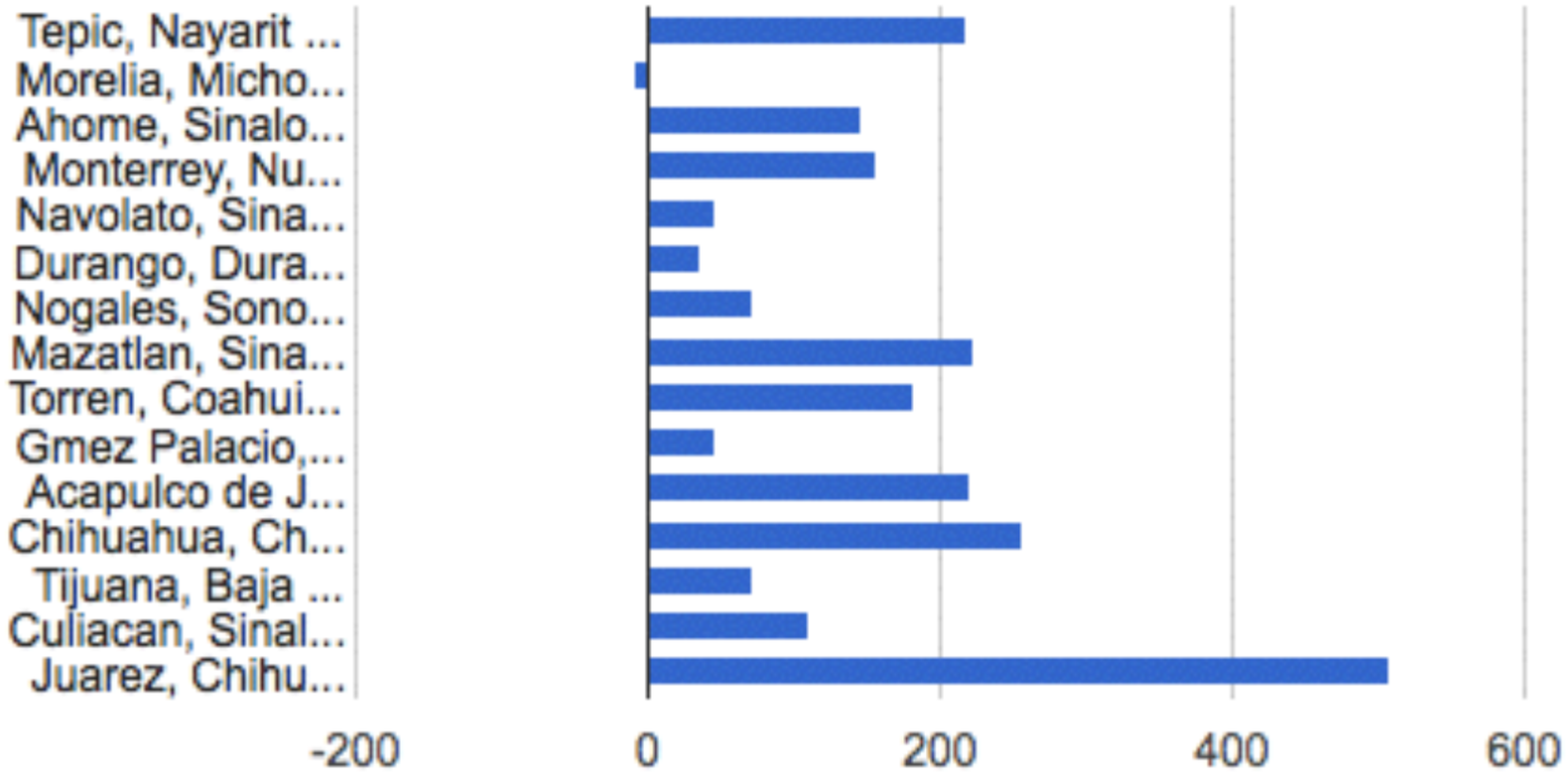


http://peltiertech.com/clustered-stacked-column-bar-charts/

Tools: QGIS

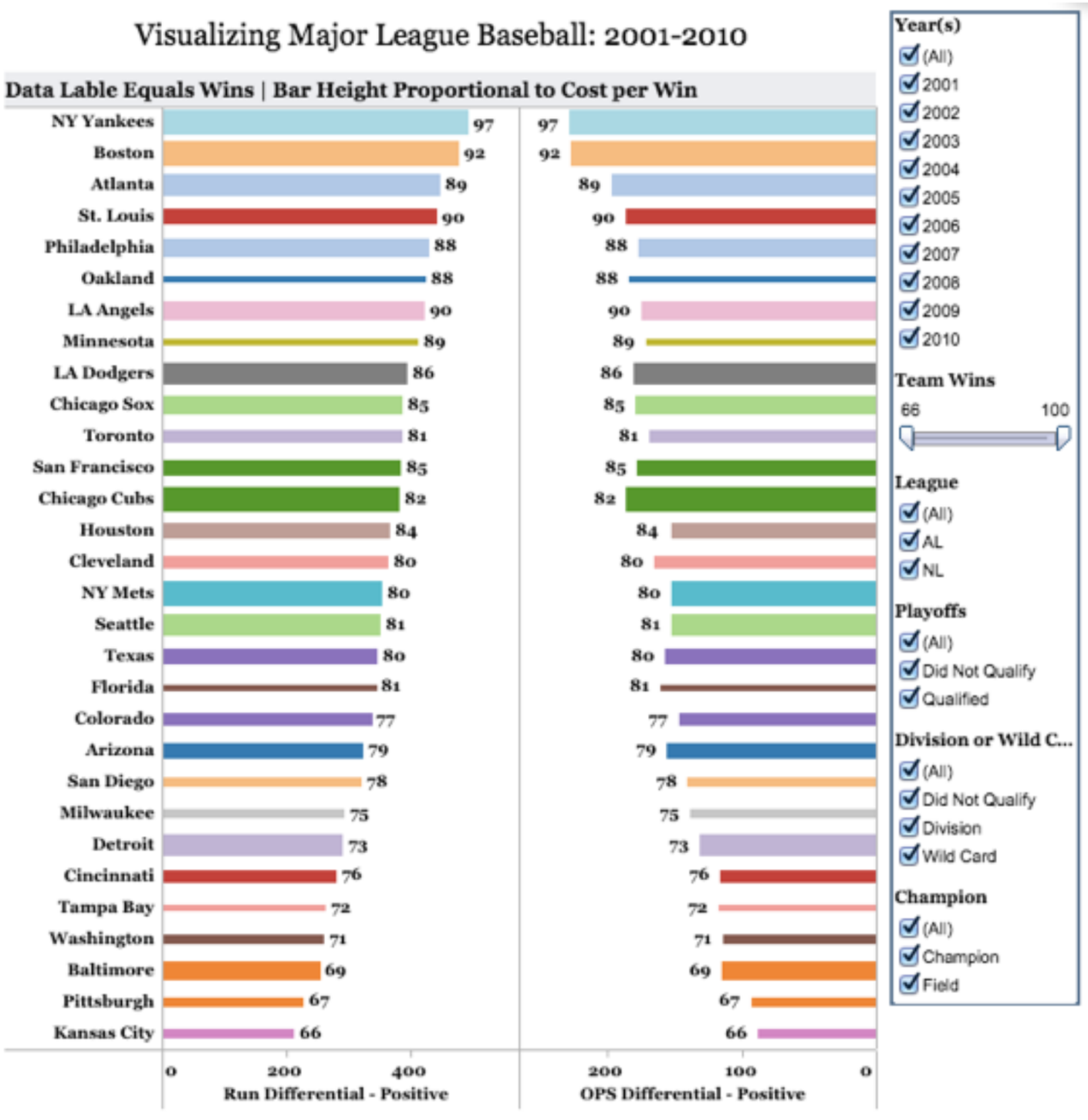

QGIS 2.2.0-Valmiera
Layers
TZA_openweat...
TZA_roads
AfSP011Qry_G...
TZA_adm0

Tools: Google Fusion Tables



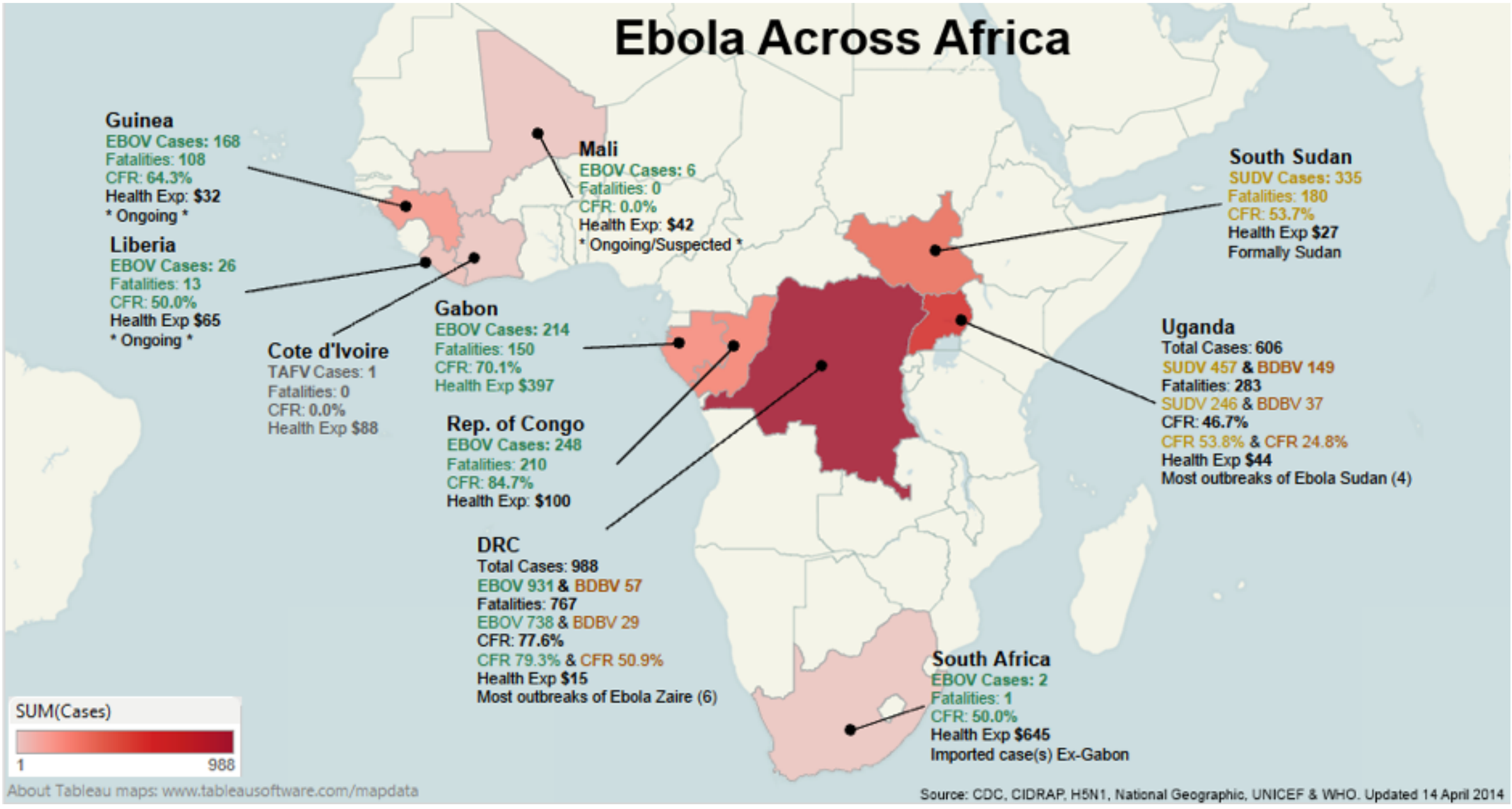
(link to google spreadsheets)

Tools: Tableau



http://www.tableausoftware.com/public/gallery/major-league-baseball

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
```

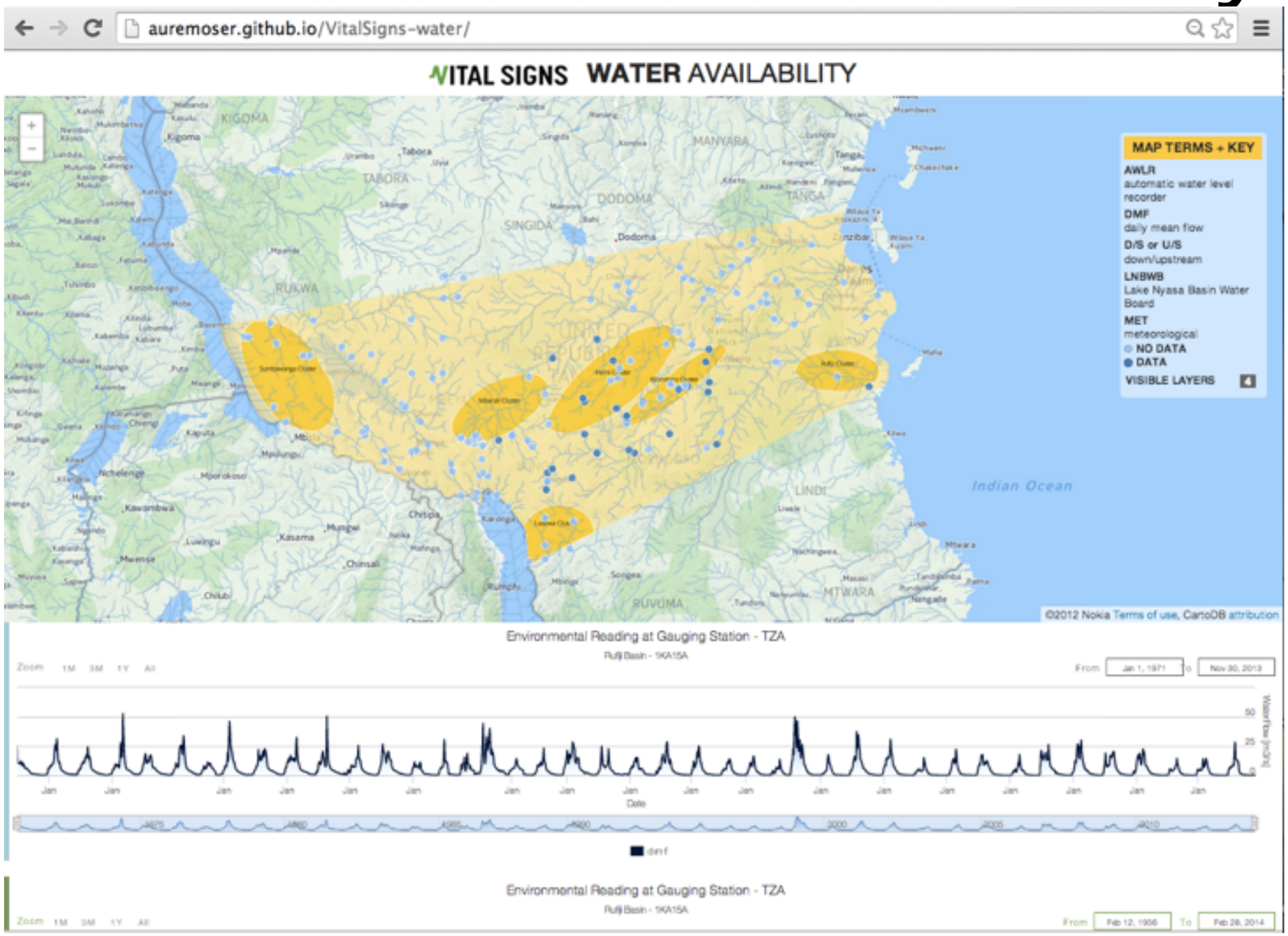
```
def getsentiments(fpsent, fptweet):

    #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

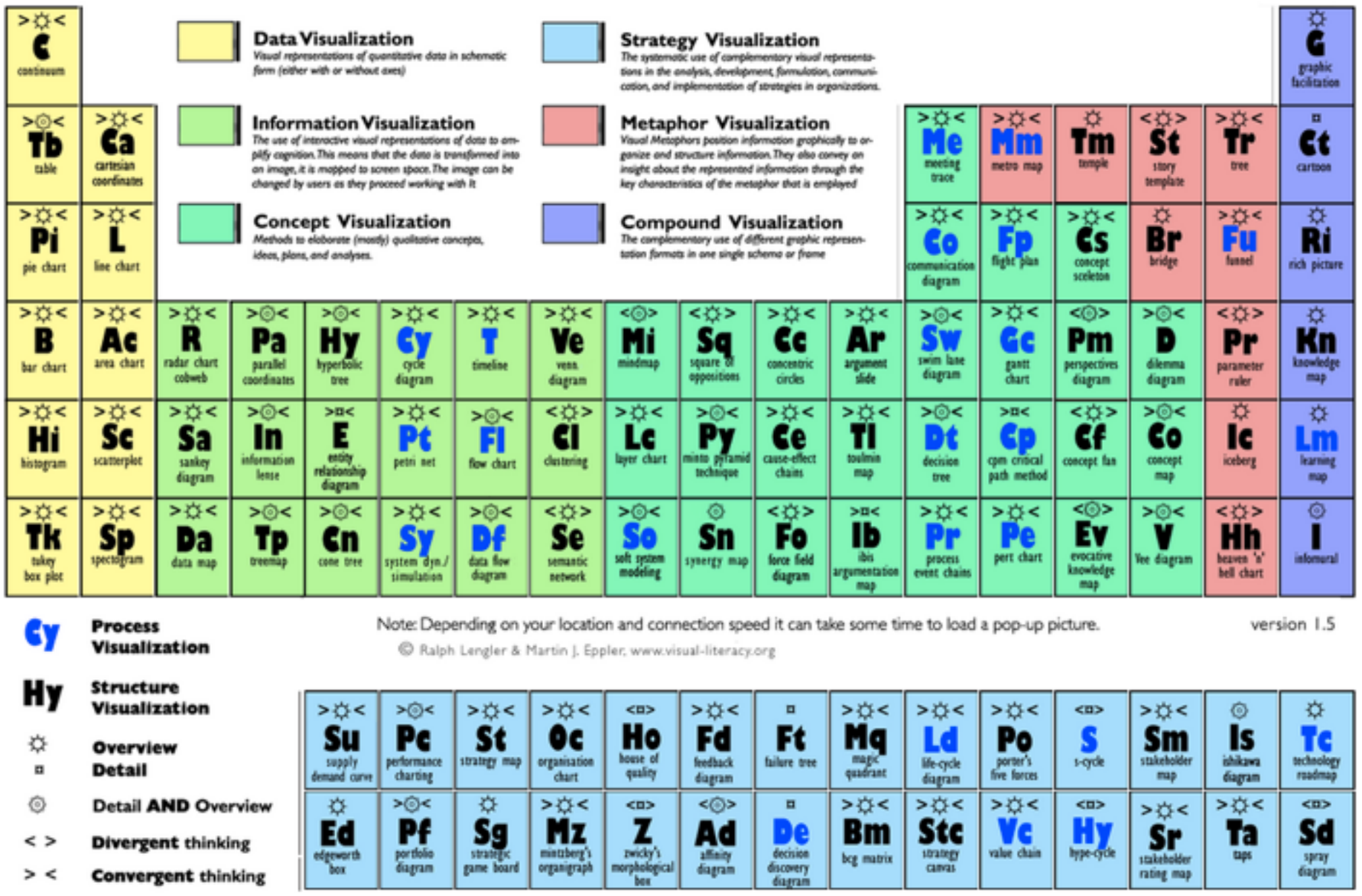
D3 gallery: https://github.com/mbostock/d3/wiki/Gallery

D3: Interactive Play



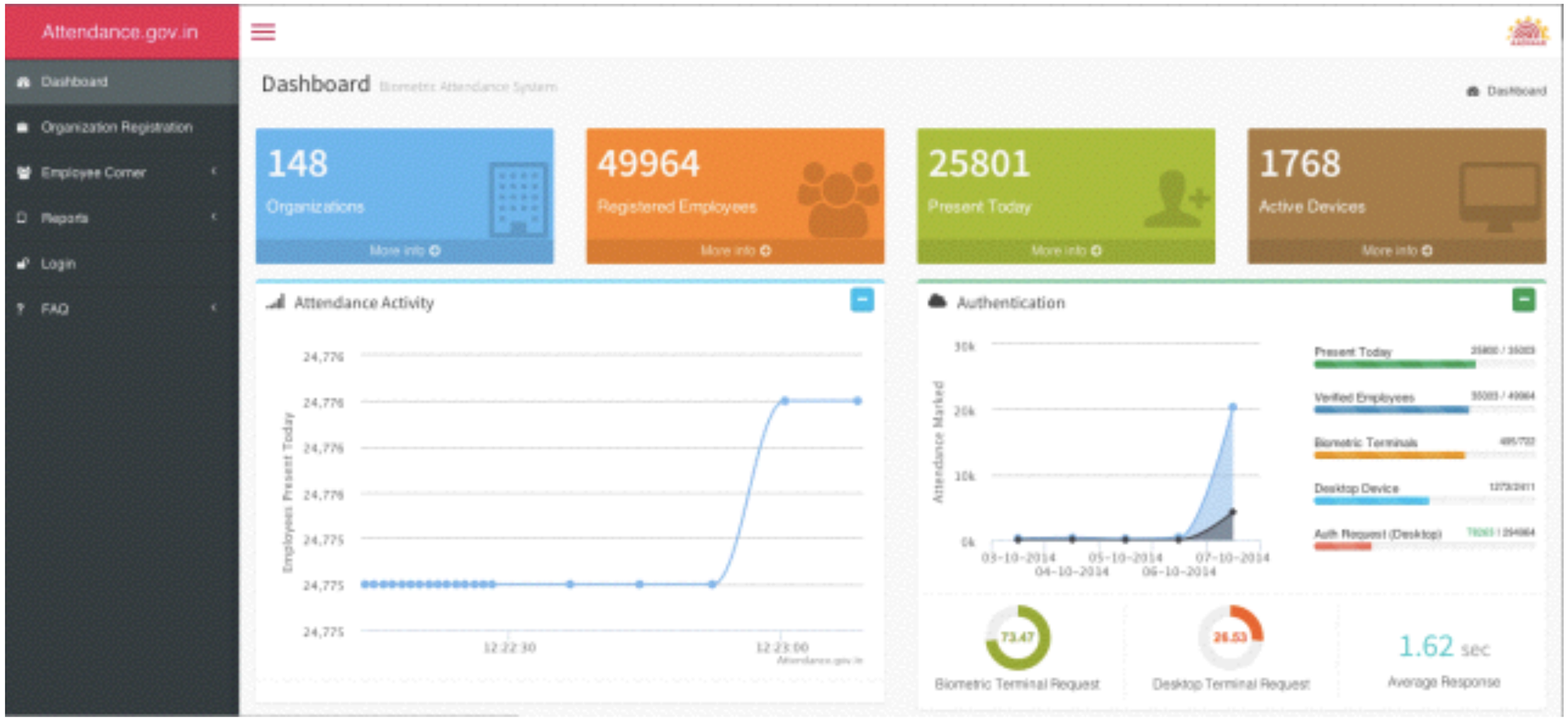
auremoser.github.io/VitalSigns-water

What's that visualisation?



http://www.visual-literacy.org/periodic_table/periodic_table.html

What’s a dashboard?



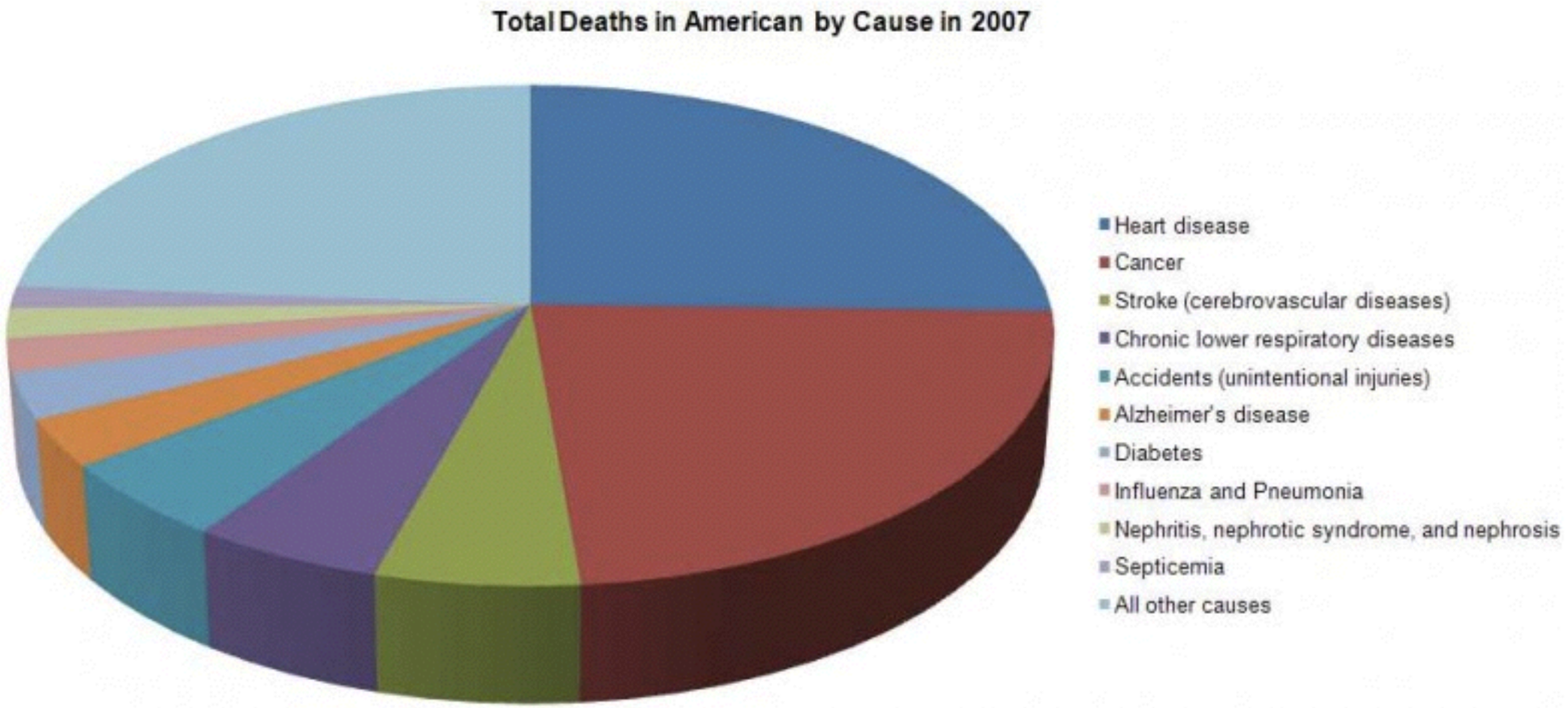
(India’s government employee attendance system)

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

http://www.creativebloq.com/how-design-better-data-visualisations-8134175.

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