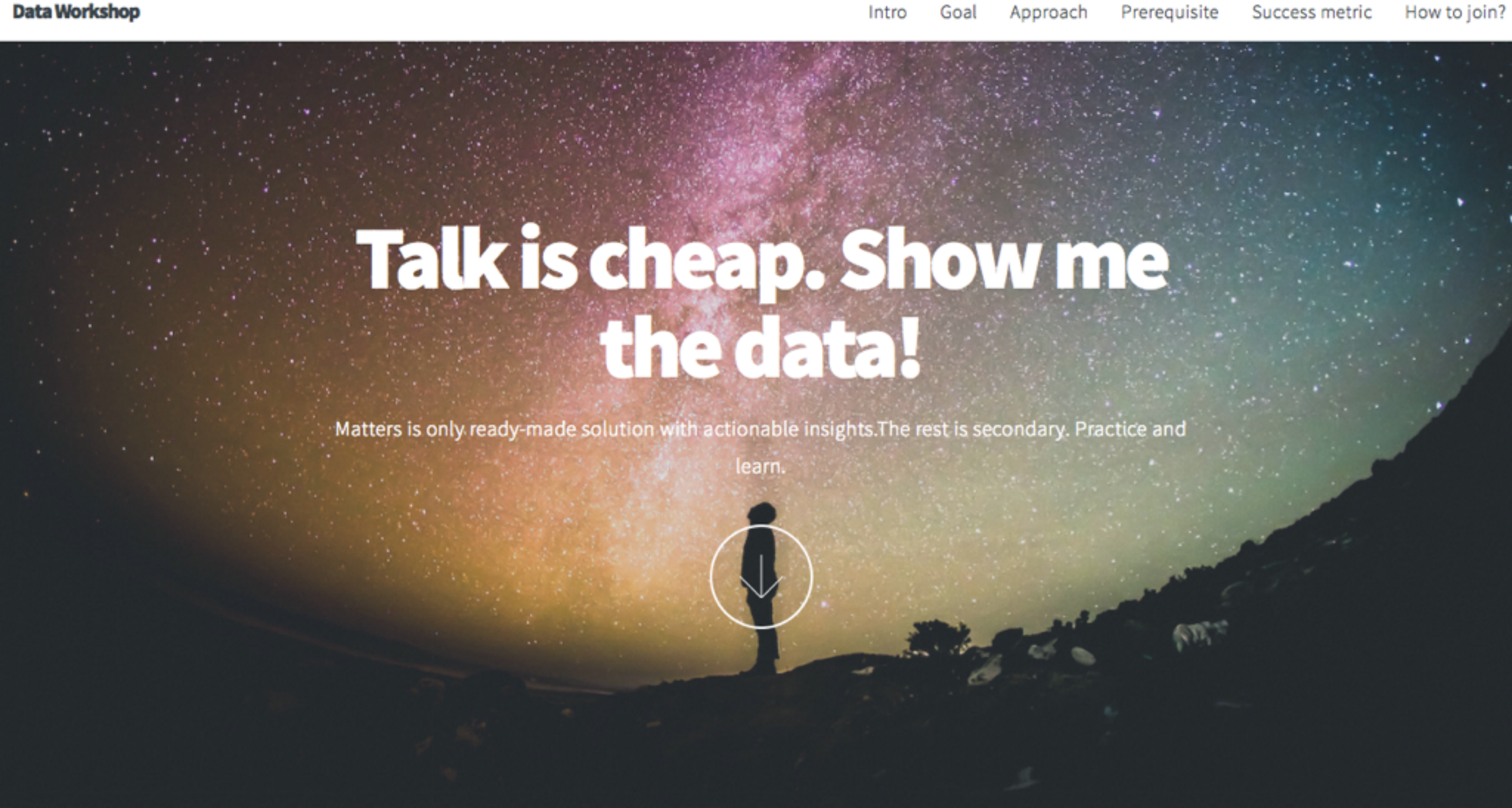
Data Workshop #5

Visualisation

dataworkshop.eu

Alekseichenko Vladimir

DataWorkshop.eu



About me

Vladimir Alekseichenko

Love analyze data

Architect

slon1024

slon1024

vova@vova.me

Disclaimer

Data Workshop *[all time]* focuses on the **intuition** and **practical** tips.

For a formal treatment, see something else.*

* papers or classical machine learning books

Environment

github.com/dataworkshop/**prerequisite**
github.com/dataworkshop/**environment**

github.com/dataworkshop/**visualization**

Packages

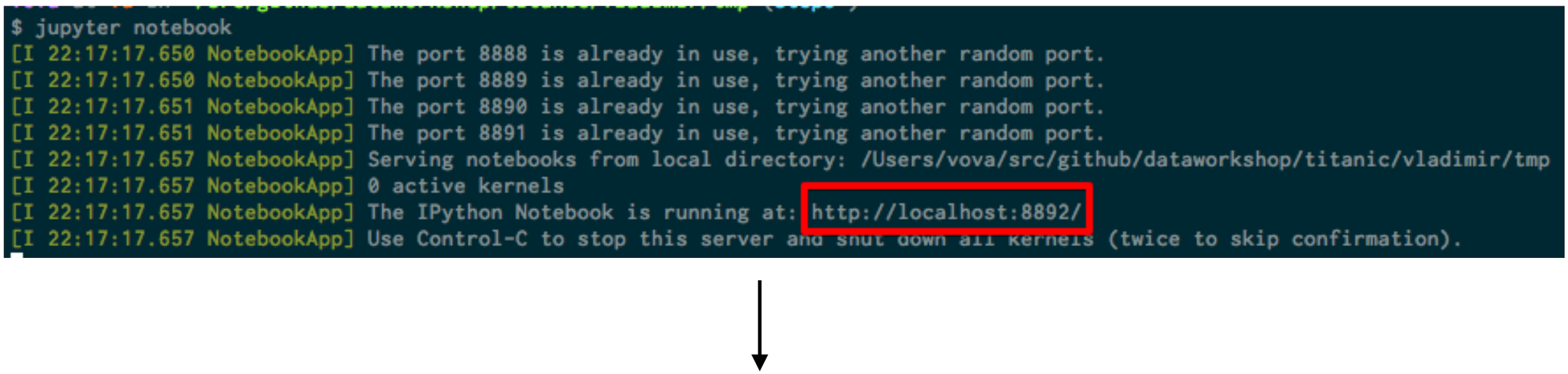
github.com/**dataworkshop/prerequisite**

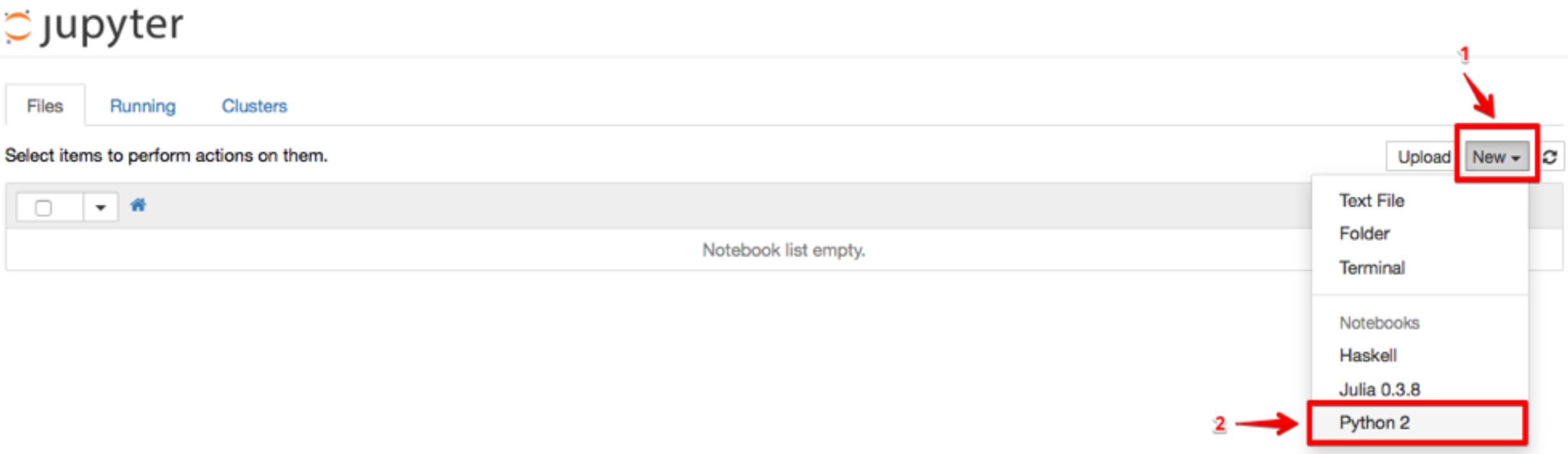
```
$ python run.py
seaborn-0.7.0 - OK
xgboost-0.4 - OK
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

==================================================
All right, you are ready to go on Data Workshop!
```

```
$ python run.py
seaborn-0.6 should be upgraded to seaborn-0.7
xgboost-0.4 - OK
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

==================================================
RECOMENDATION (without upgrade some needed features could be missing)
pip install --upgrade seaborn
```

```
$ python run.py
seaborn-0.7.0 - OK
xgboost - missing
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

==================================================
REQUIRED
Please install those packages before Data Workshop: xgboost
pip install xgboost
More info how to install xgboost: http://xgboost.readthedocs.org/en/latest/build.html
```

jupyter notebook

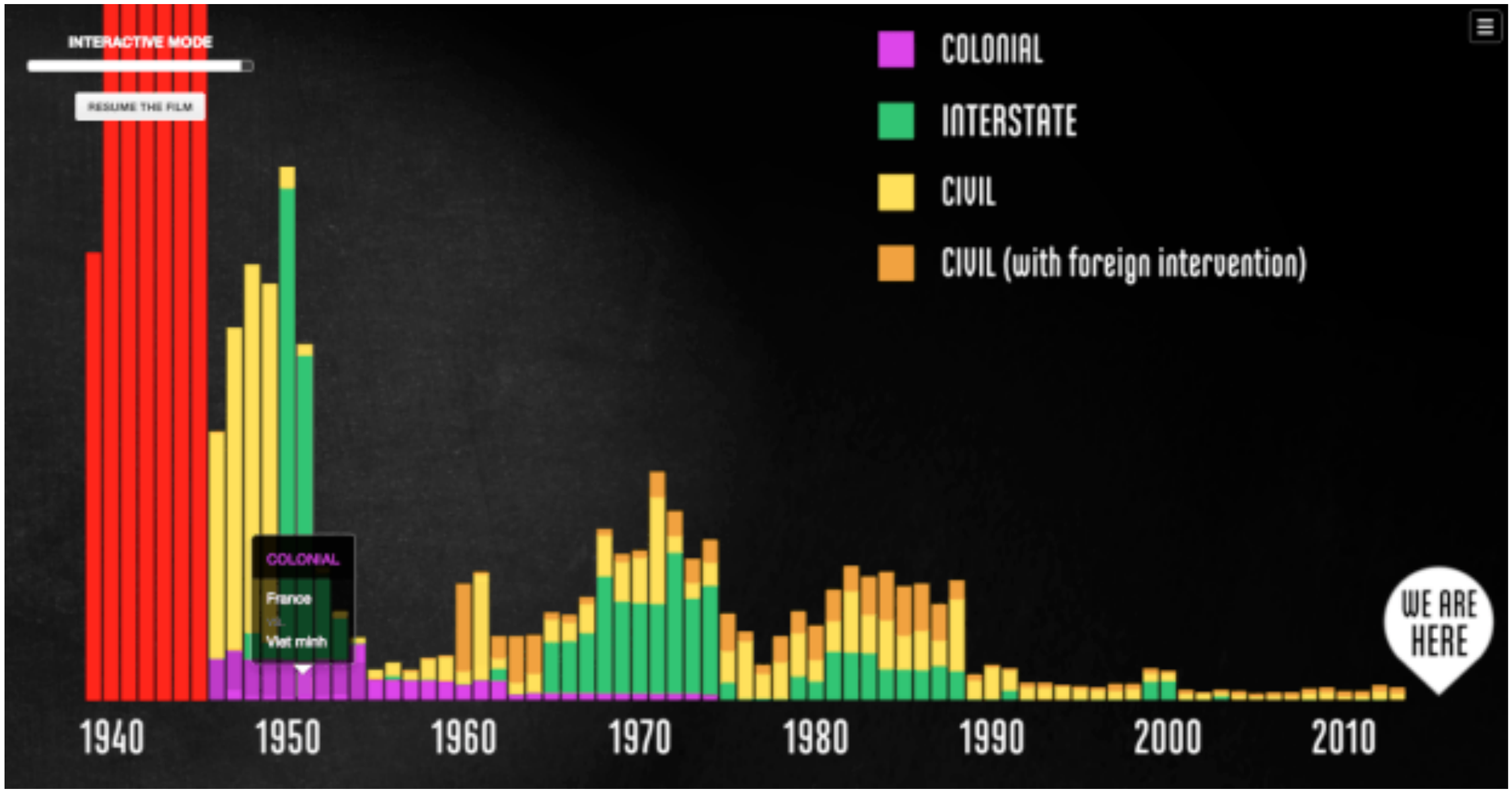




Motivation

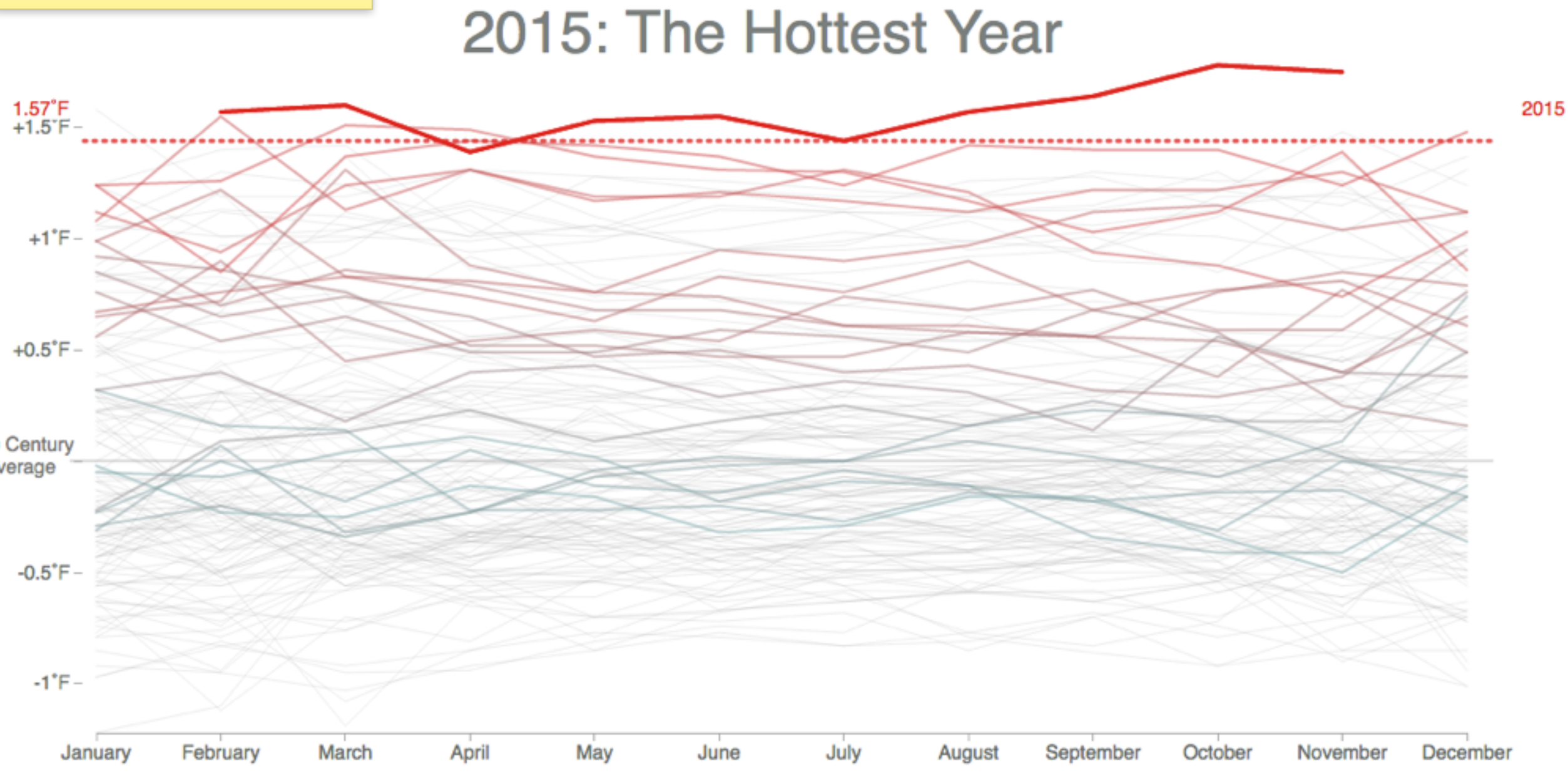
Millions of peopled died during World War II, but it's difficult to grasp what all the big numbers associated with the war mean. Neil Halloran explains in The Fallen of World War II, a hybrid between interactive visualization and documentary.

n of World War II



http://www.fallen.io/ww2/

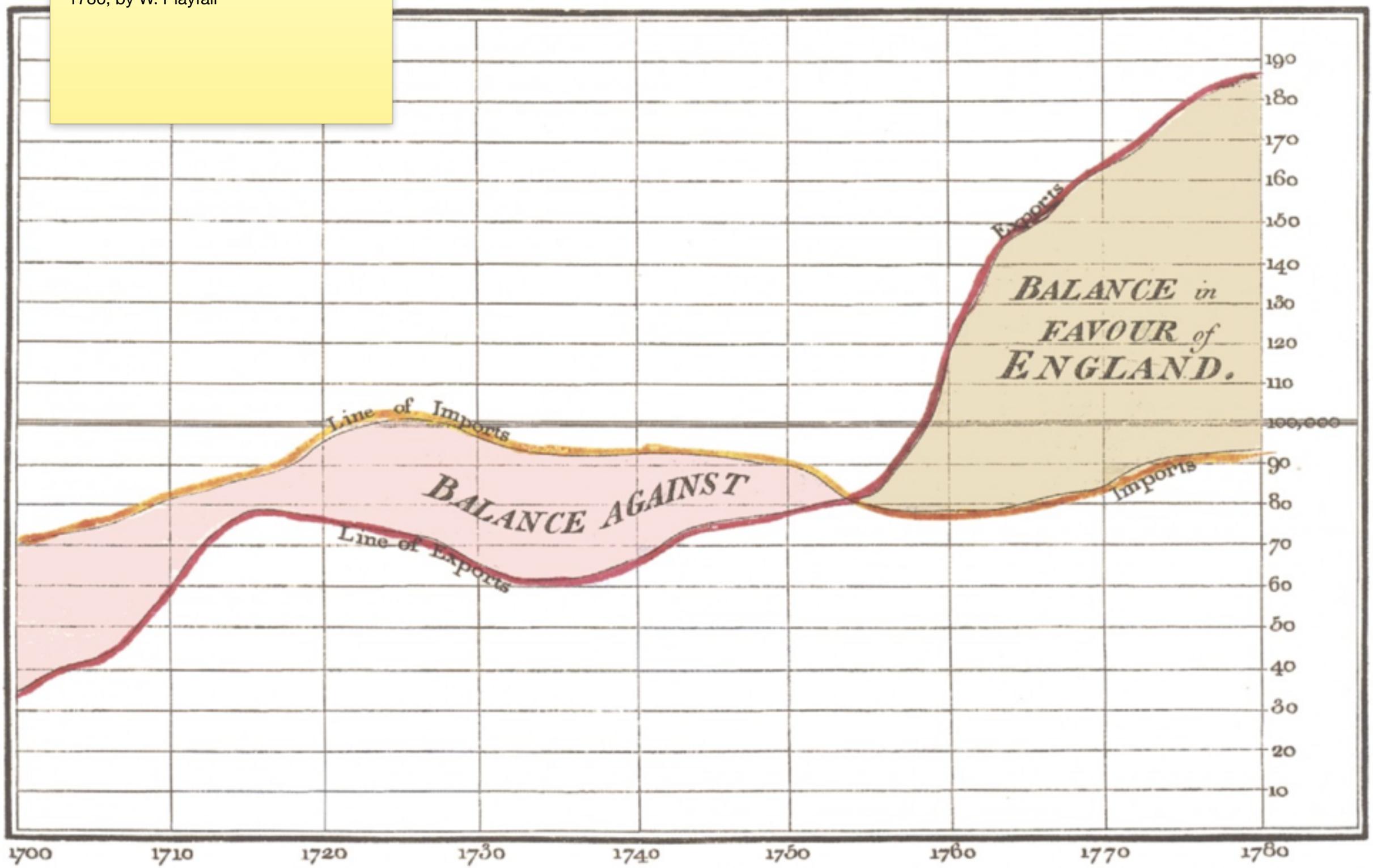
The animated chart by Tom Randall and Blacki Migliozzi for Bloomberg shows average monthly temperature. Each line represents a full calendar year, and as you get closer to present-day the line



http://www.bloomberg.com/graphics/2014-hottest-year-on-record/

Published as the Act directs, 14th May 1786, by W. Playfair

nports to and from DENMARK & NORWAY from 1700 to 1780.



The Bottom line is divided into Years, the Right hand line into £10,000 each.

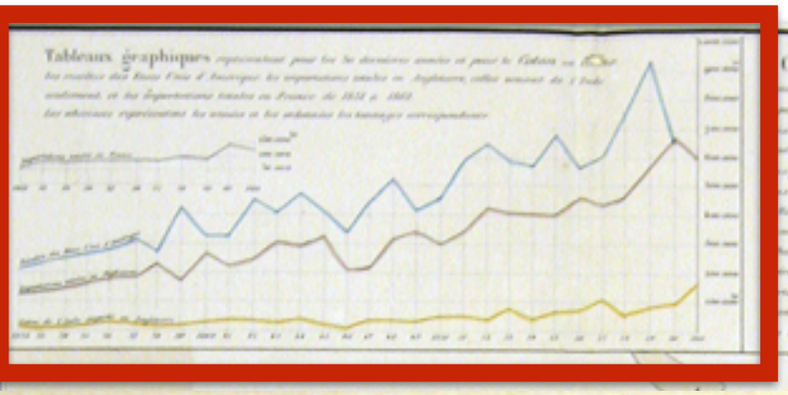
Published as the Act directs, 1st May 1786, by Wm Playfair *Neale sculpt 352, Strand, London.*

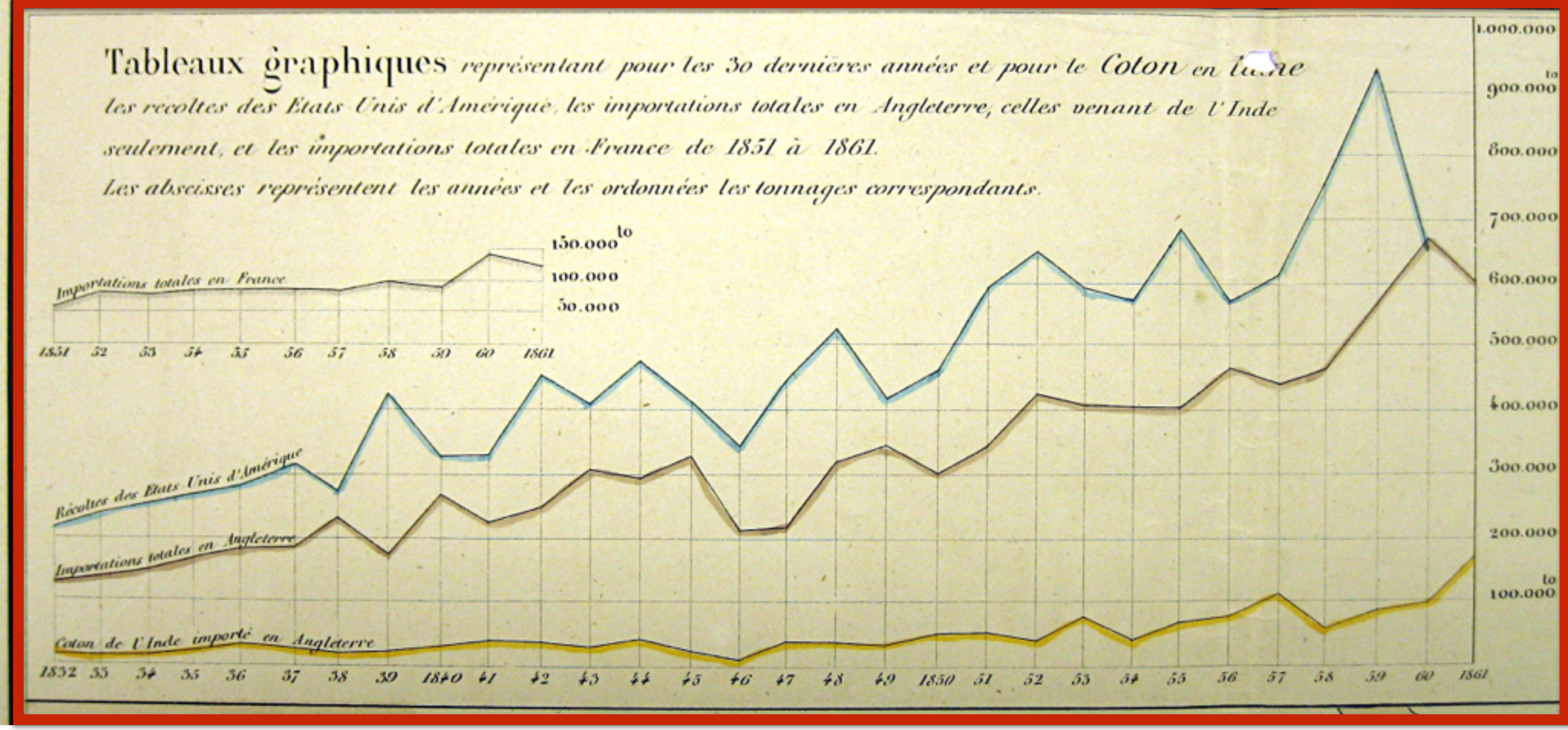
https://plot.ly/~MakotoRK/516/exports-and-imports-to-and-from-denmark-norway-from-1700-to-1780.embed

Tableaux graphiques *représentant pour les 30 dernières années et pour le Coton en laine*

les recoltes des Etats Unis d'Amérique, les importations totales en Angleterre, celles venant de l'Inde seulement, et les importations totales en France de 1851 à 1861.

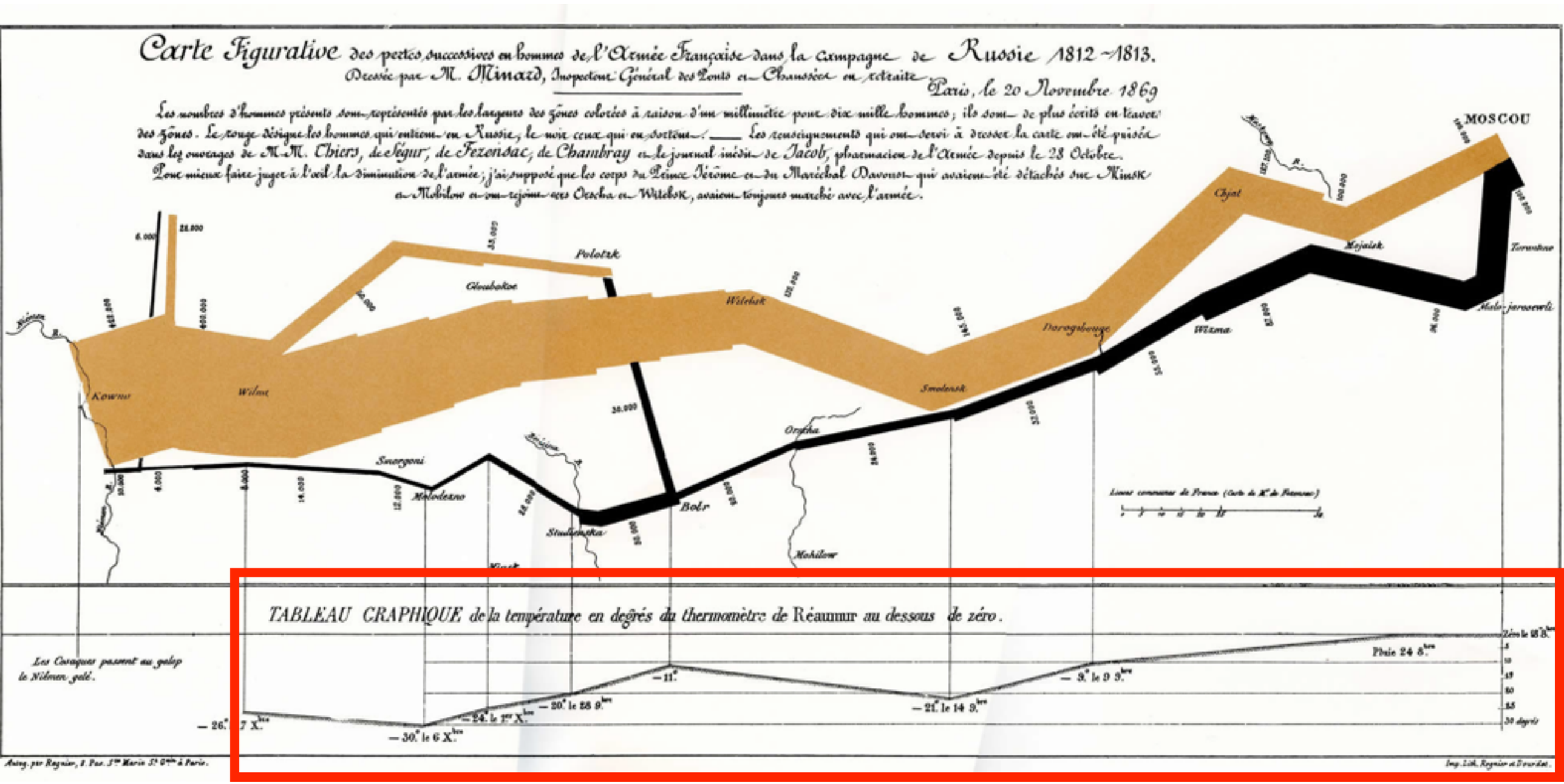
Les abscisses représentent les années et les ordonnées les tonnages correspondants.





https://cartographia.wordpress.com/category/charles-joseph-minard/

Sankey diagram

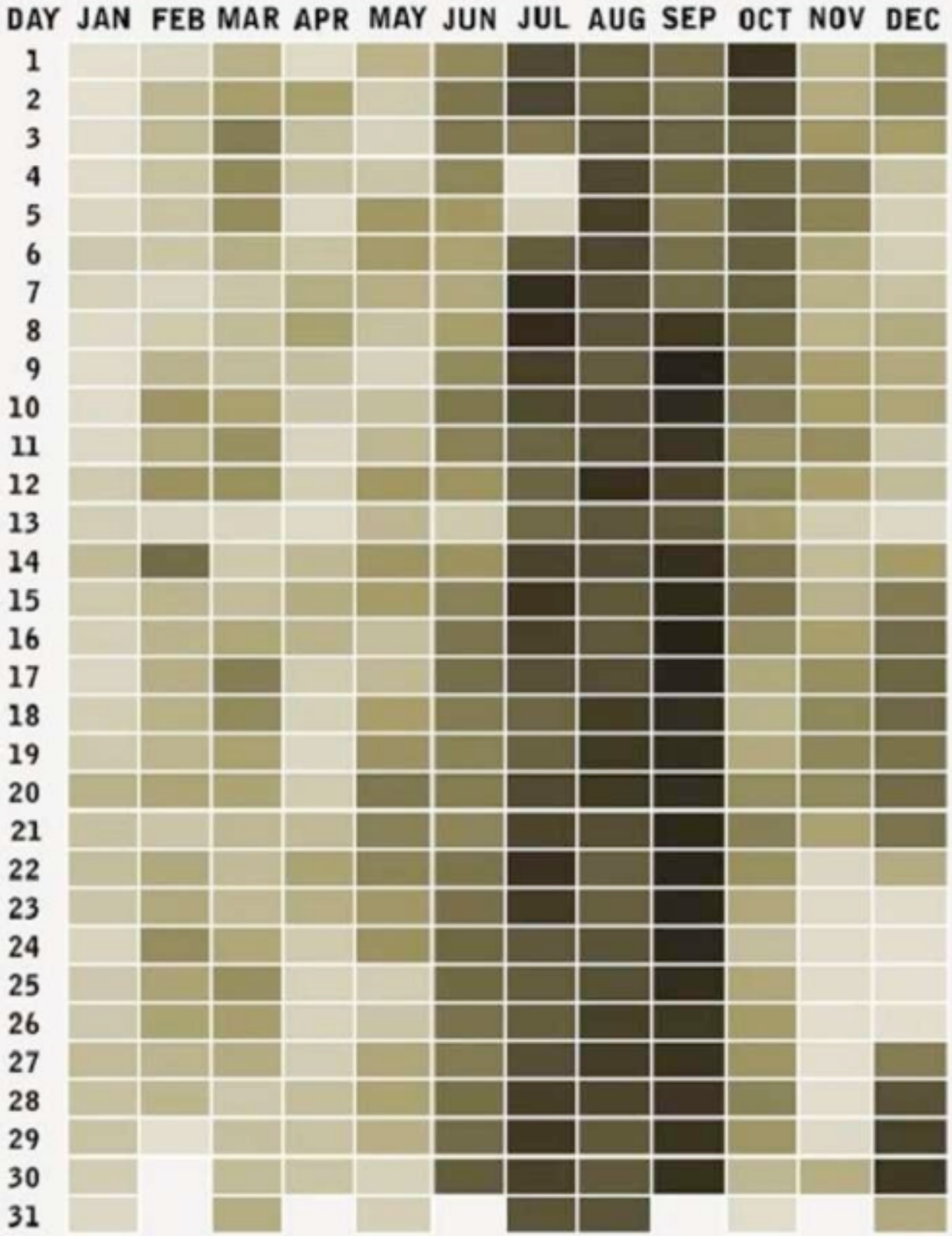


Temperature

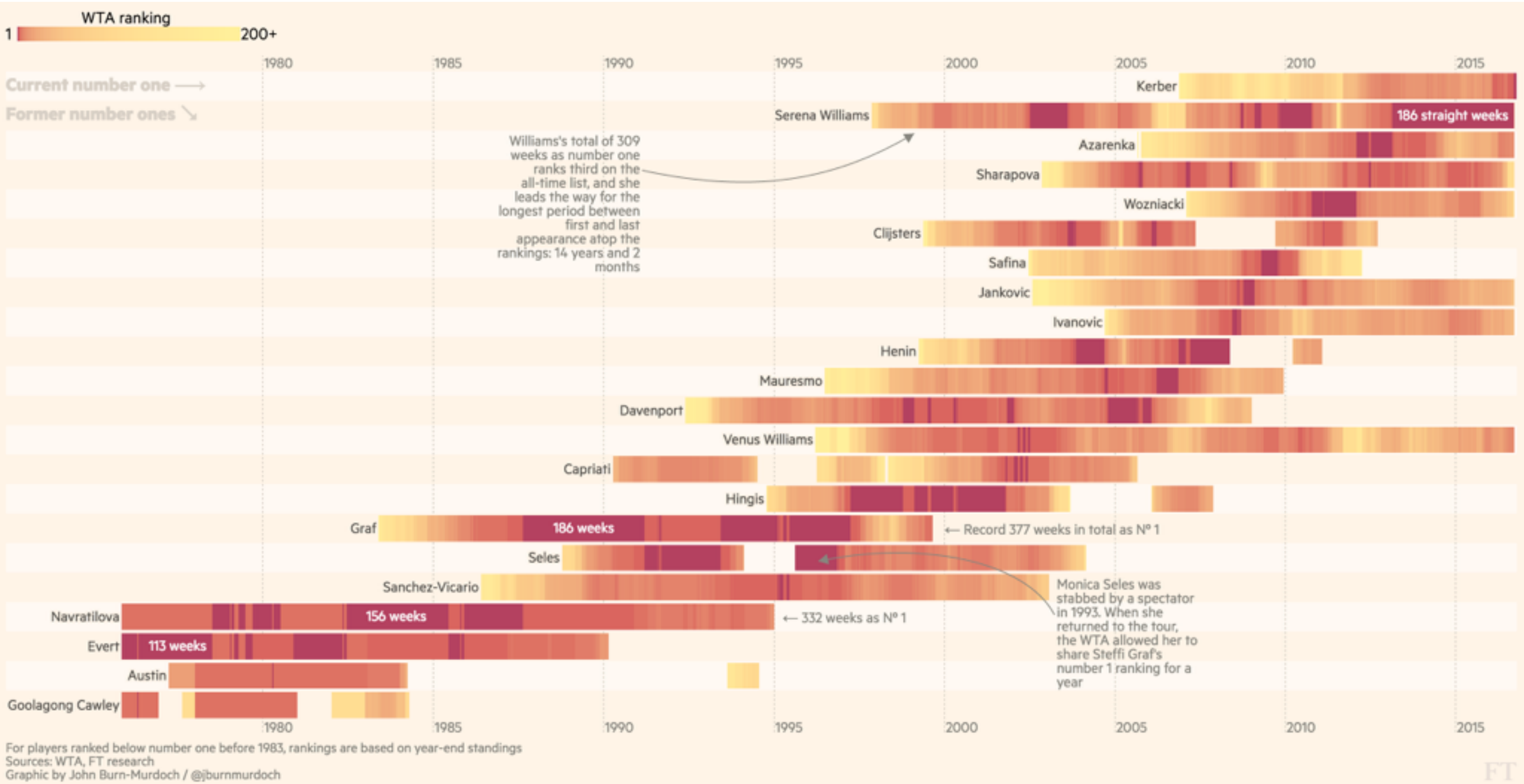
https://en.wikipedia.org/wiki/Sankey_diagram

Sankey diagram

Which Birth Dates Are Most Common?

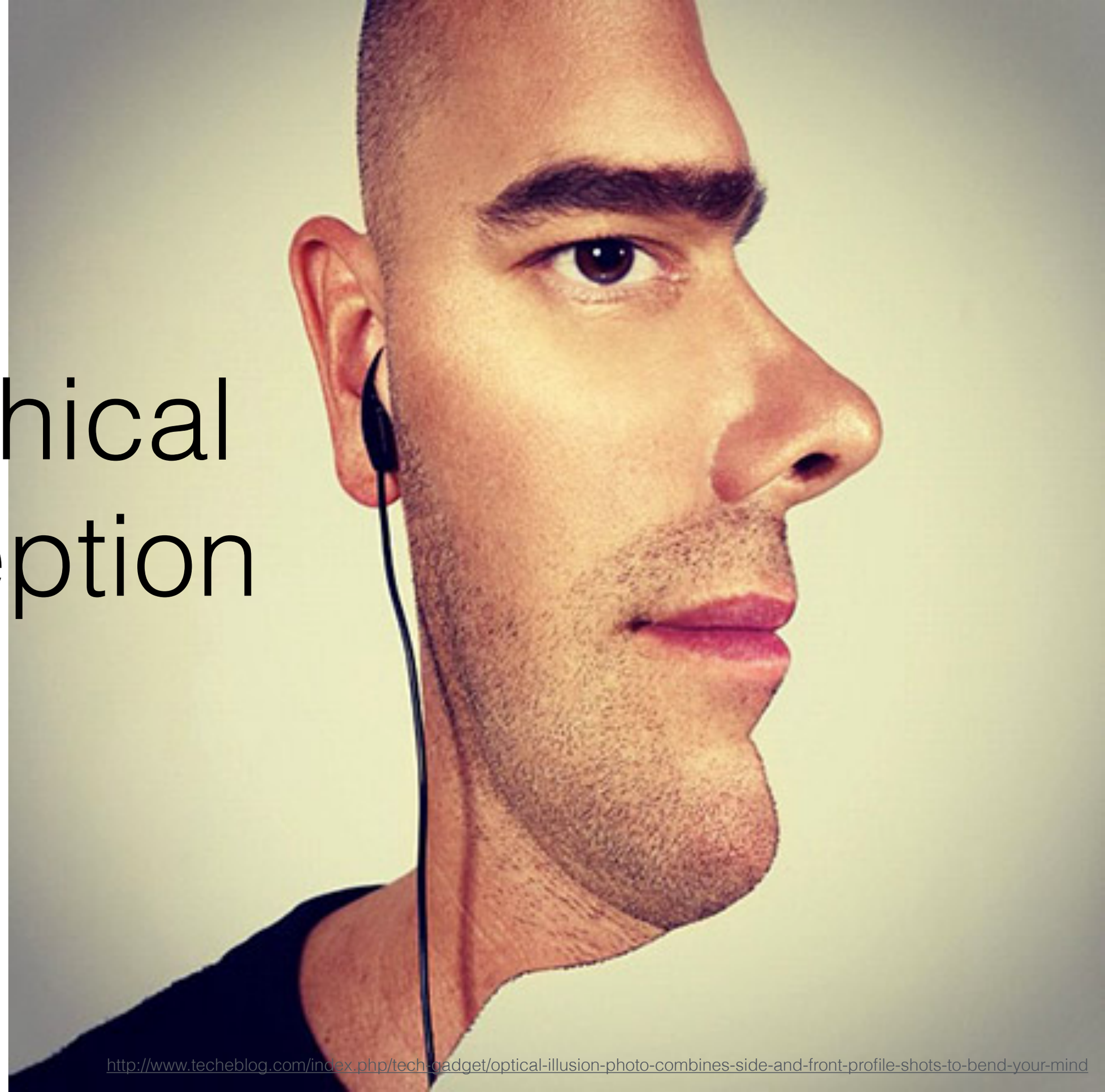


A visual history of women's tennis



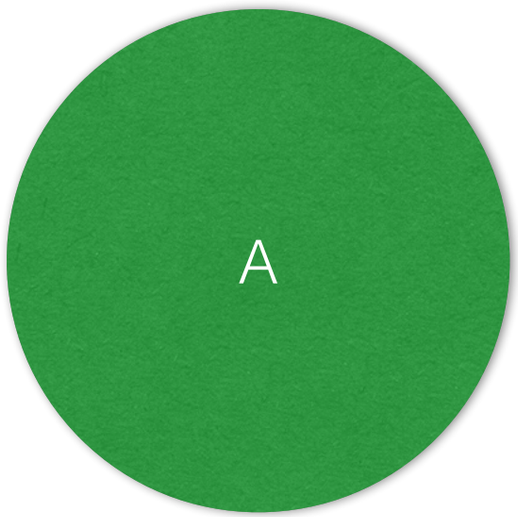
https://ig.ft.com/sites/visual-history-of-womens-tennis/

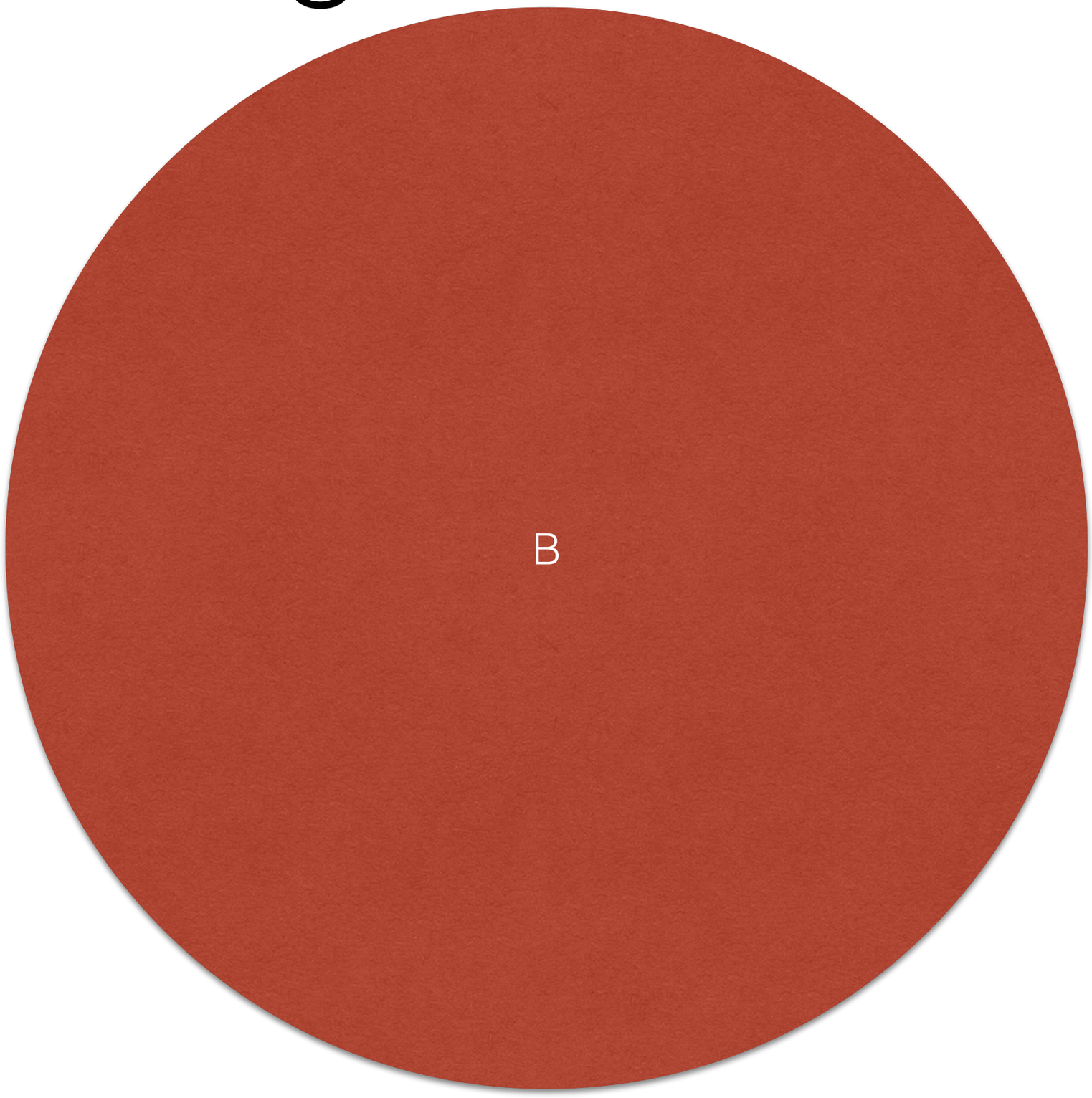
Graphical Perception



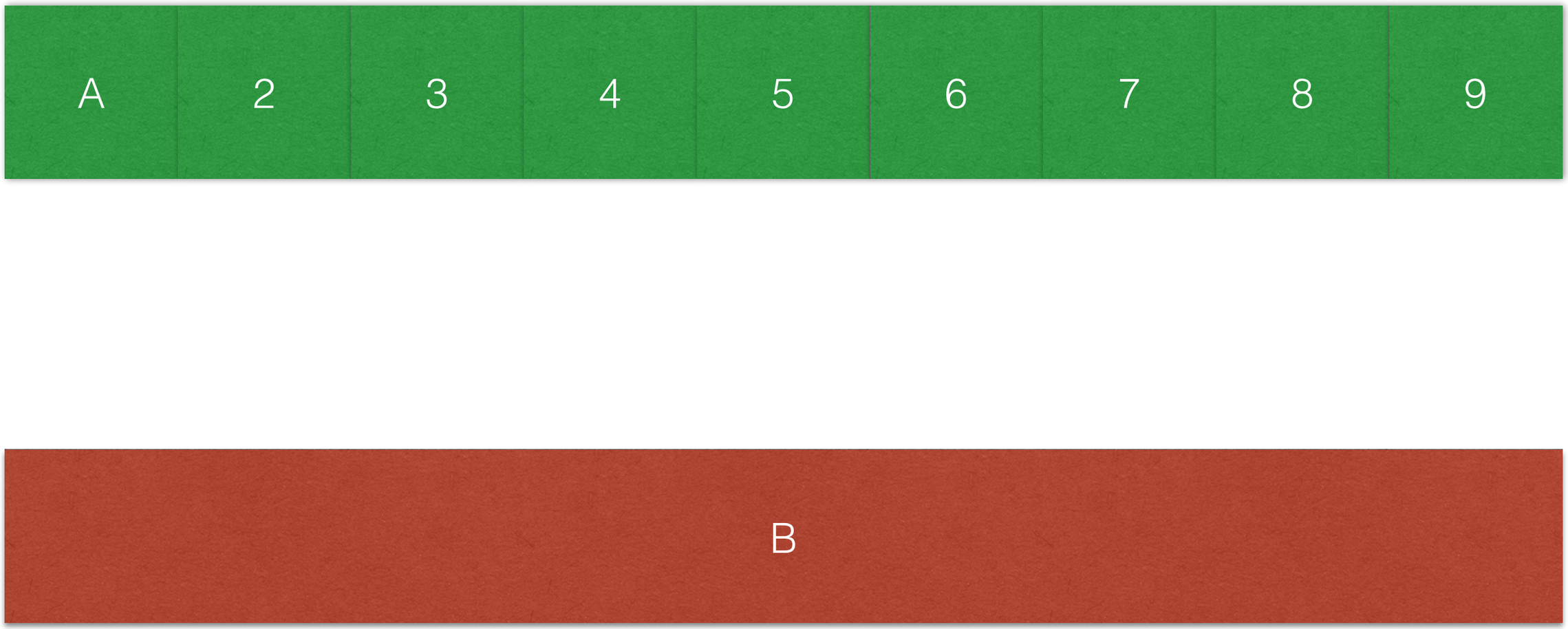
http://www.techeblog.com/index.php/tech-gadget/optical-illusion-photo-combines-side-and-front-profile-shots-to-bend-your-mind

How much larger is area?



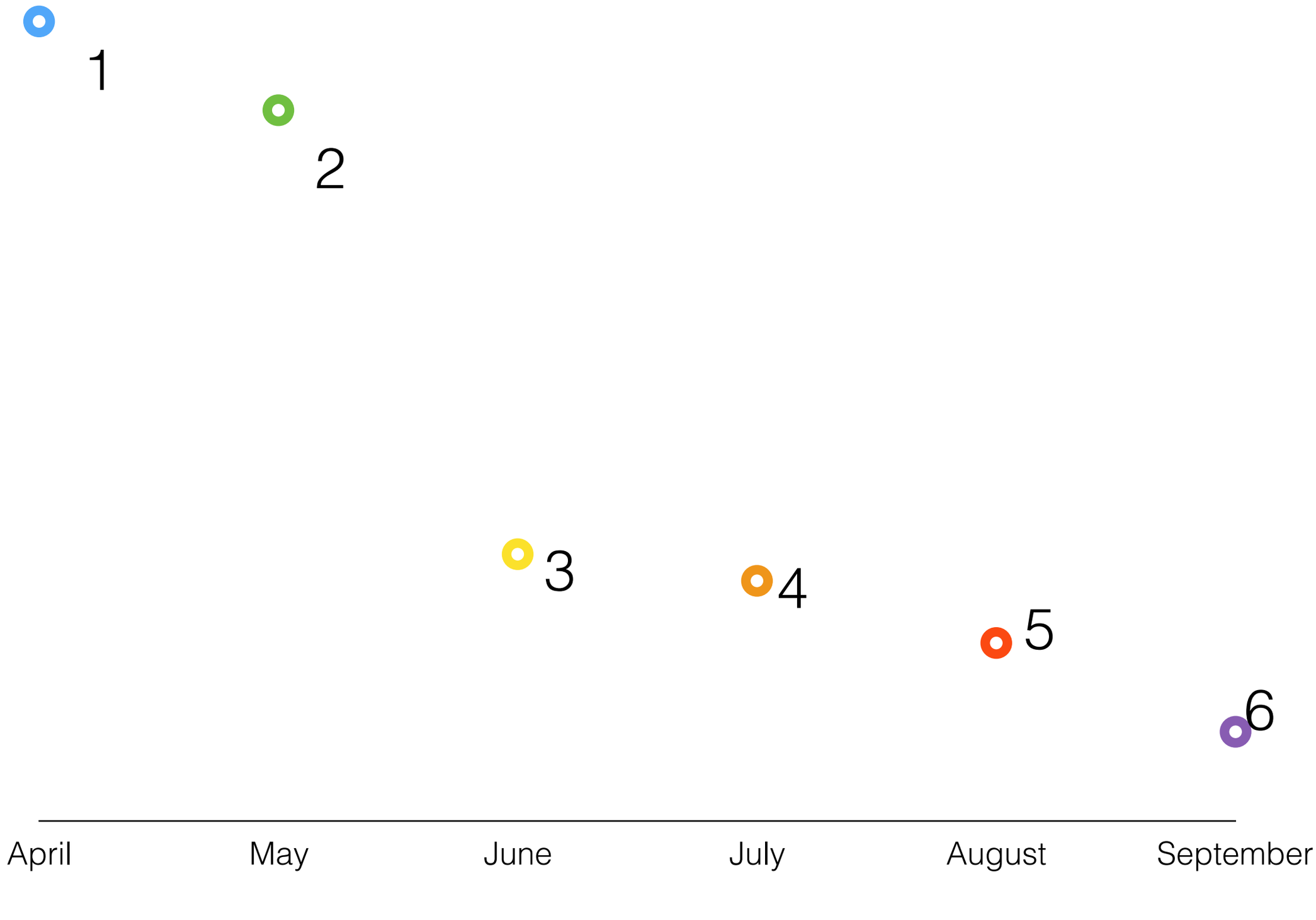


How much larger is area?

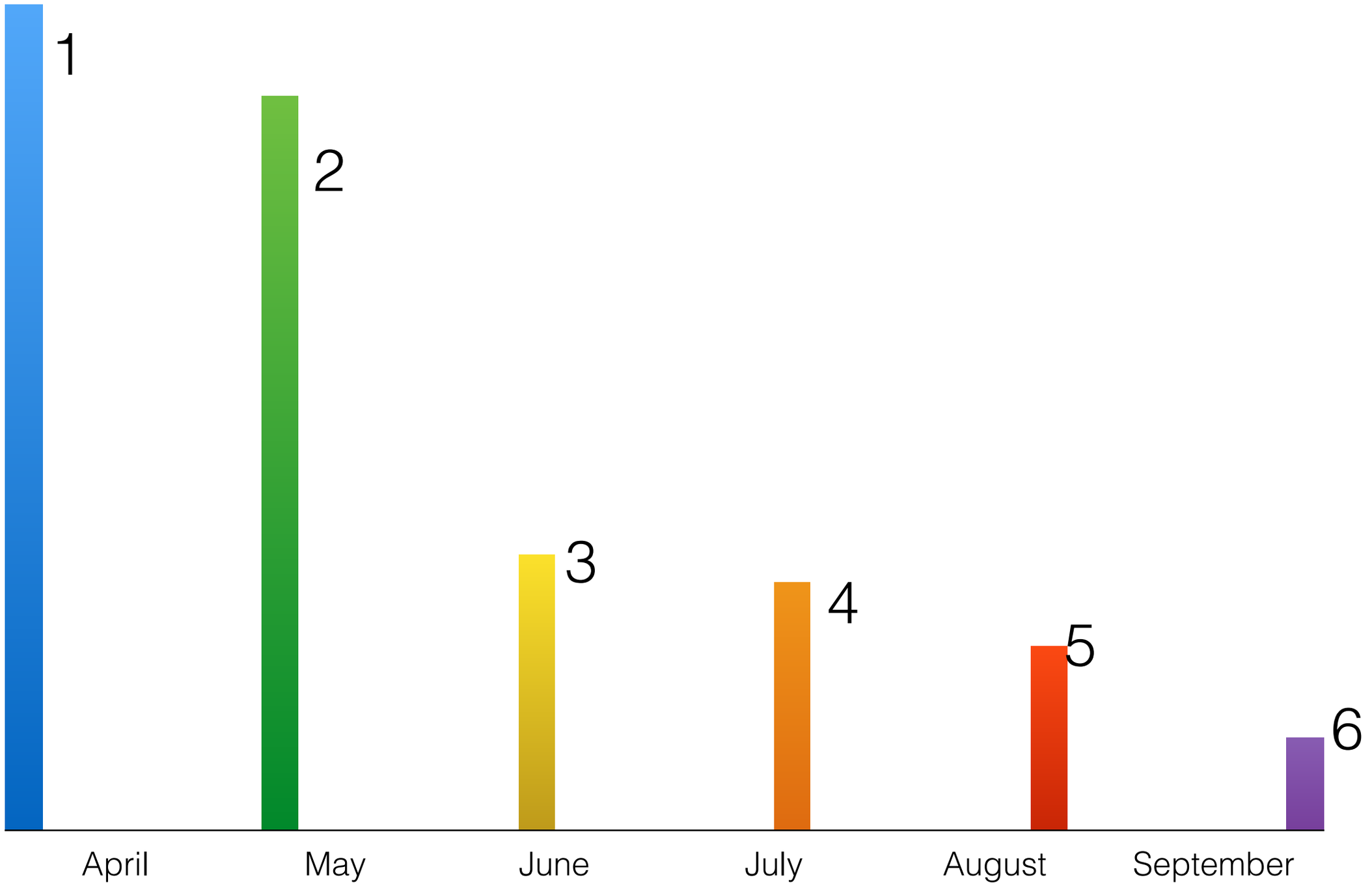


Pie Chart

Line Chart



Bar Chart

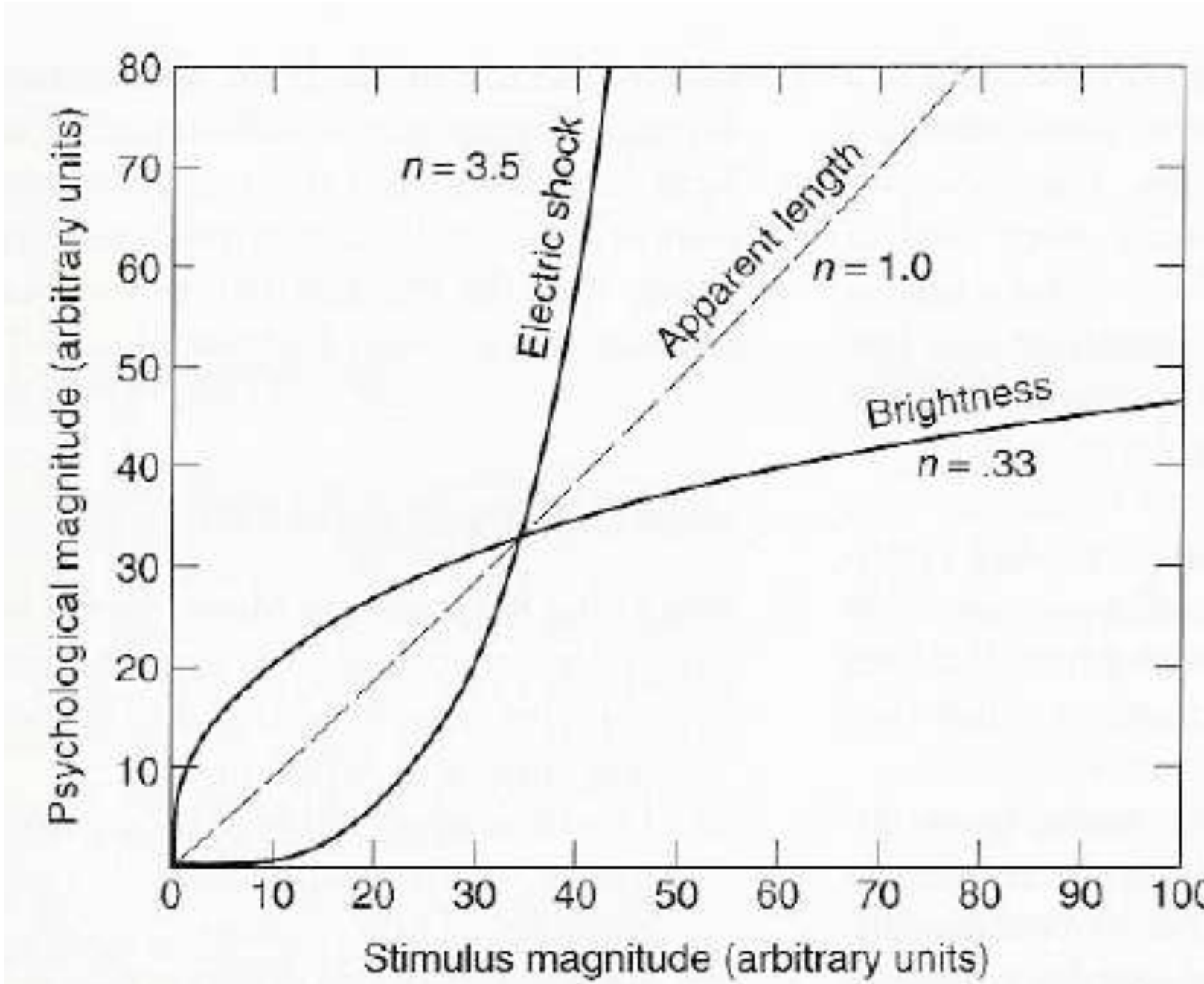


Steven's Power Law

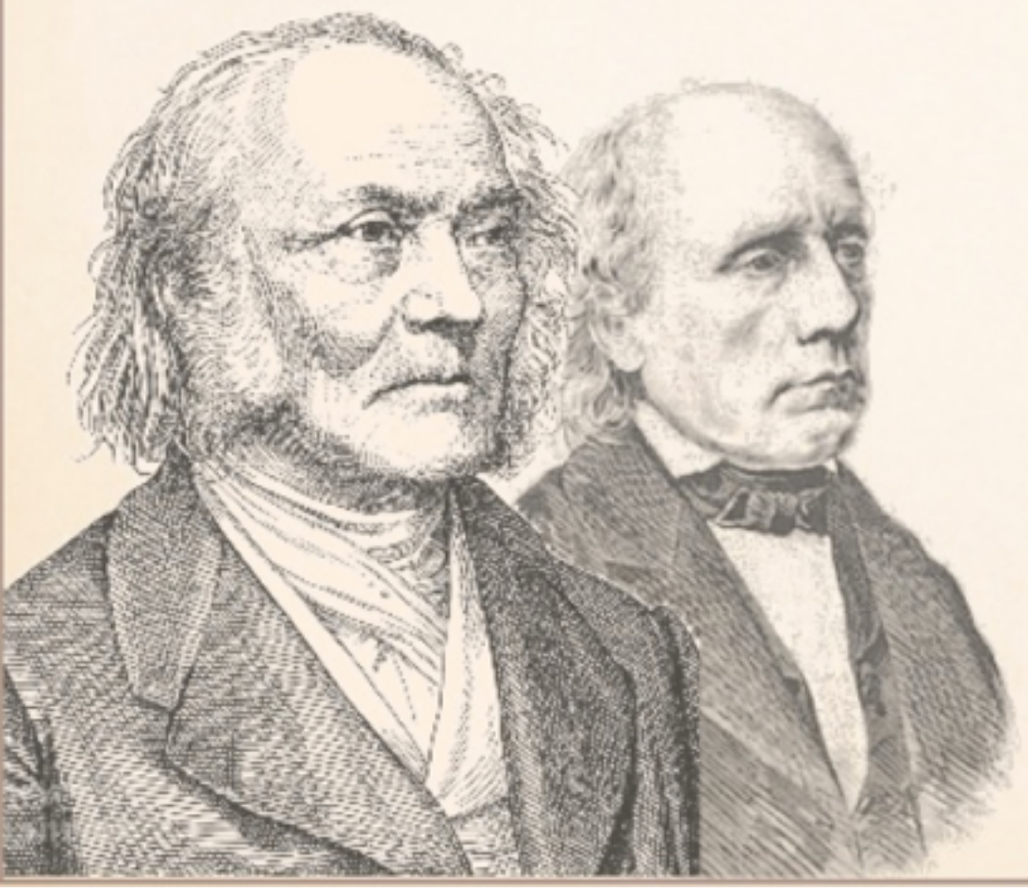
magnitude estimation

$$S = I^n$$

physical intensity



WEBER-FECHNER LAW



"The just-noticeable difference between two stimuli is proportional to the magnitude of the stimuli"

http://www.slideshare.net/FDConf/the-illusion-of-time-when-60-sec-is-not-1-minute

https://vivifychangecatalyst.wordpress.com/2014/06/28/weber-fechner-law-career-plateau-and-salary-hike/

Relative Magnitude Estimation

Most accurate

Least accurate

Position (common)
Position (non-aligned)

Length
Slope
Angle

Area
Volume
Color (hue,saturation,value)

Variables

Type of Variables

- Nominal (labels or categories)
- Ordered
- Quantative
  - Interval
  - Ratio

Type of Variables

Name	Operations	Example
Nominal	<>, ==	Apple, Orange, Banan Europe, Asia, America
Ordered	<>, ==, **>, <**	Low, Medium, High Child, Adult, Old
Quantitative-Interval	<>, ==, >, <, **-** ***measure distances***	*date:* 1 Jan 2016, 10 Jan 2016 *location:* Moscow, London
Quantitative-Ratio *(zero fixed)*	<>, ==, >, <, -, +, **%** ***measure ratios***	*physical measurements:*10cm, 30kg, 25C

Matplotlib

http://matplotlib.org

Type of Plots

- **plt.plot** – creates a line plot
- **plt.bar** – creates a bar chart
- **plt.boxplot** – makes a box and whisker plot
- **plt.hist** – makes a histogram
- **plt.scatter** – makes a scatter plot

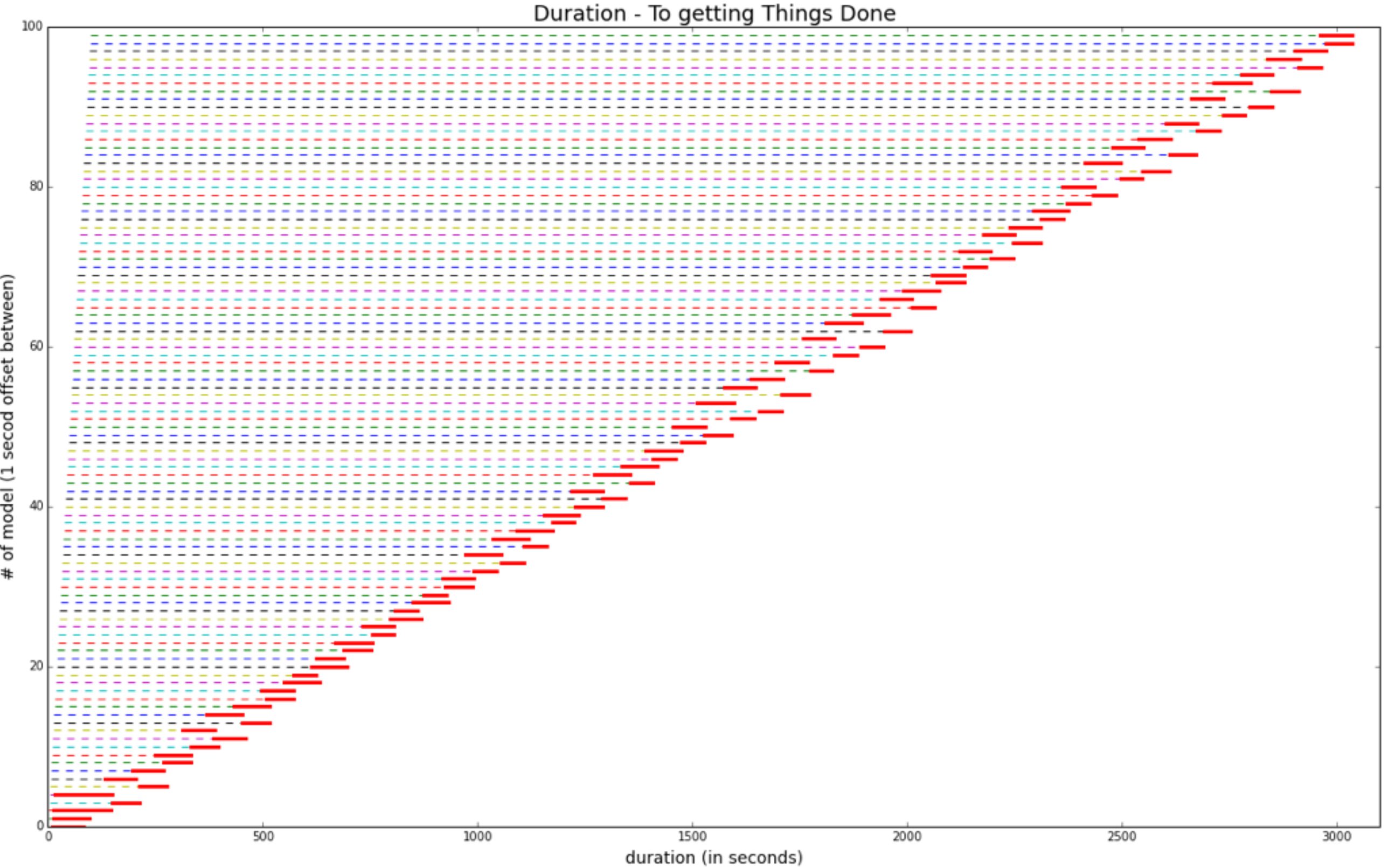
Title and Labels

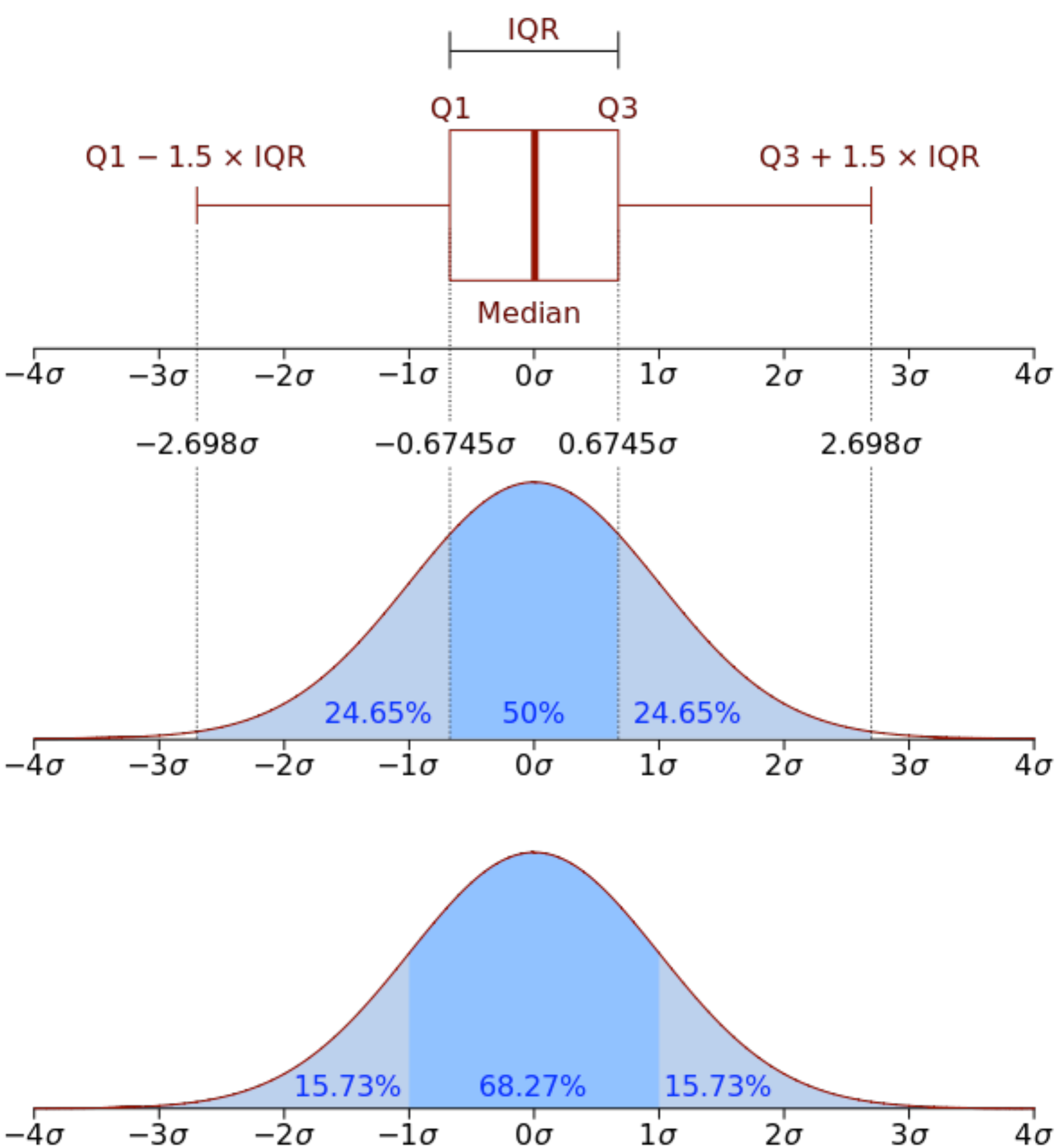
- **plt.title** – adds a title to the plot.
- **plt.xlabel** – adds an x-axis label.
- **plt.ylabel** – adds a y-axis label.

Task

Analyze the result of load tests

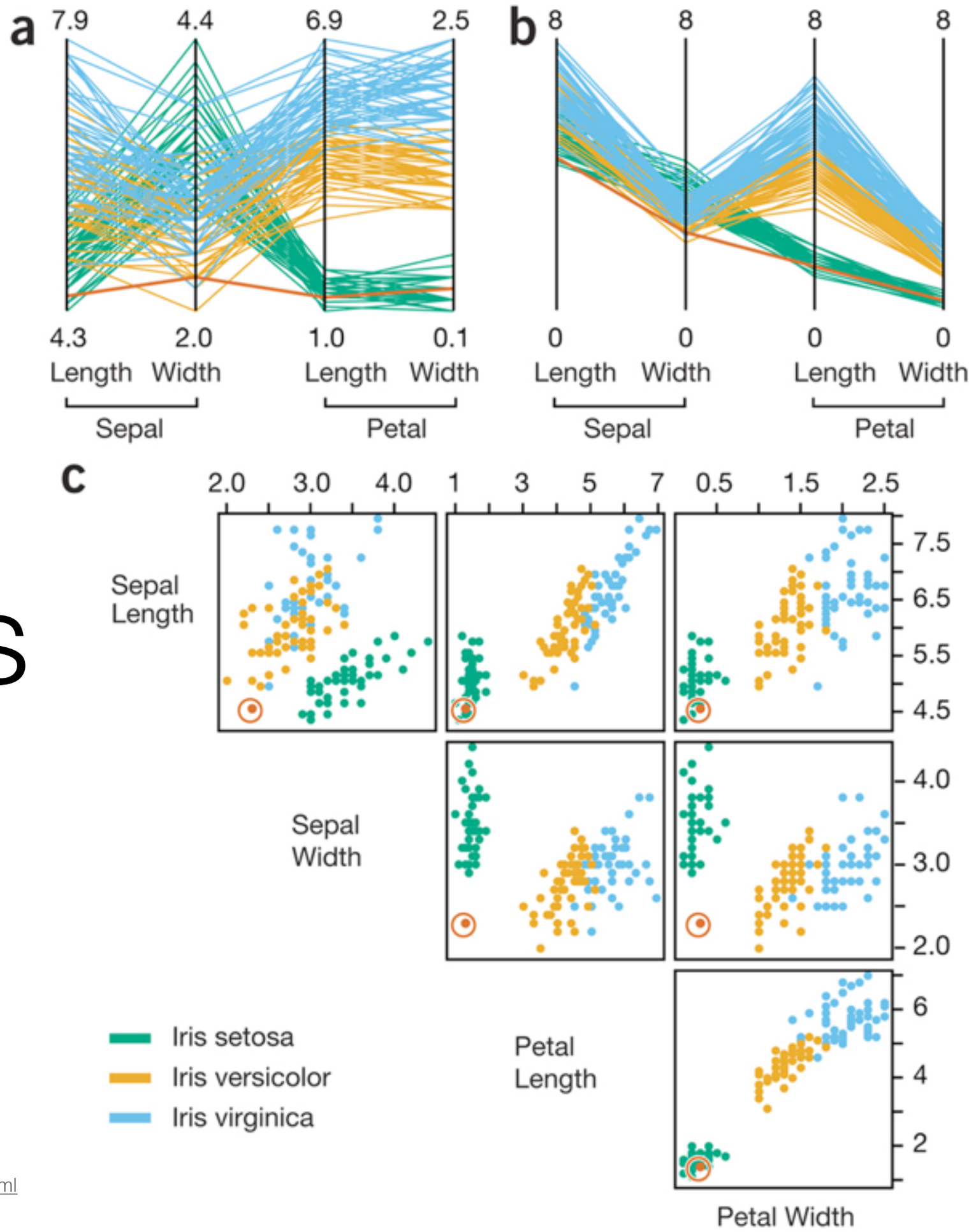
```
File Length : 1034494 bytes
File Creation Time : 10/6/2016 3:42:35 PM
File Modified Time : 8/17/2016 8:04:34 PM
File Accessed Time : 10/6/2016 3:42:35 PM

10/7/2016 2:08:39 PM : BEGIN COPYING FILES TO STAGING DIRECTORY
10/7/2016 2:08:39 PM : END COPYING FILES TO STAGING DIRECTORY

10/7/2016 2:08:39 PM : BEGIN PUBLISHING PDF

Server run id : df14a150-571a-46b5-988b-a4aaa62c2442

10/7/2016 2:08:50 PM : Server run status : RunAccepted

10/7/2016 2:10:13 PM : Server run status : RunInProgress

10/7/2016 2:10:23 PM : Server run status : RunQueued

10/7/2016 2:14:02 PM : Server run status : RunInProgress

10/7/2016 2:14:12 PM : Server run status : RunQueued

10/7/2016 2:14:33 PM : Server run status : RunInProgress

10/7/2016 2:15:57 PM : END PUBLISHING PDF

PDF for part.prt.1 created successfully

10/7/2016 2:15:57 PM : BEGIN COPY PDF TO LOCAL FOLDER
10/7/2016 2:15:57 PM : END COPY PDF TO LOCAL FOLDER

10/7/2016 2:15:57 PM : ==========END PROCESSING FILE==========
```


Duration - To getting Things Done
# of model (1 secod offset between)
duration (in seconds)
0
20
40
60
80
100
0
500
1000
1500
2000
2500
3000



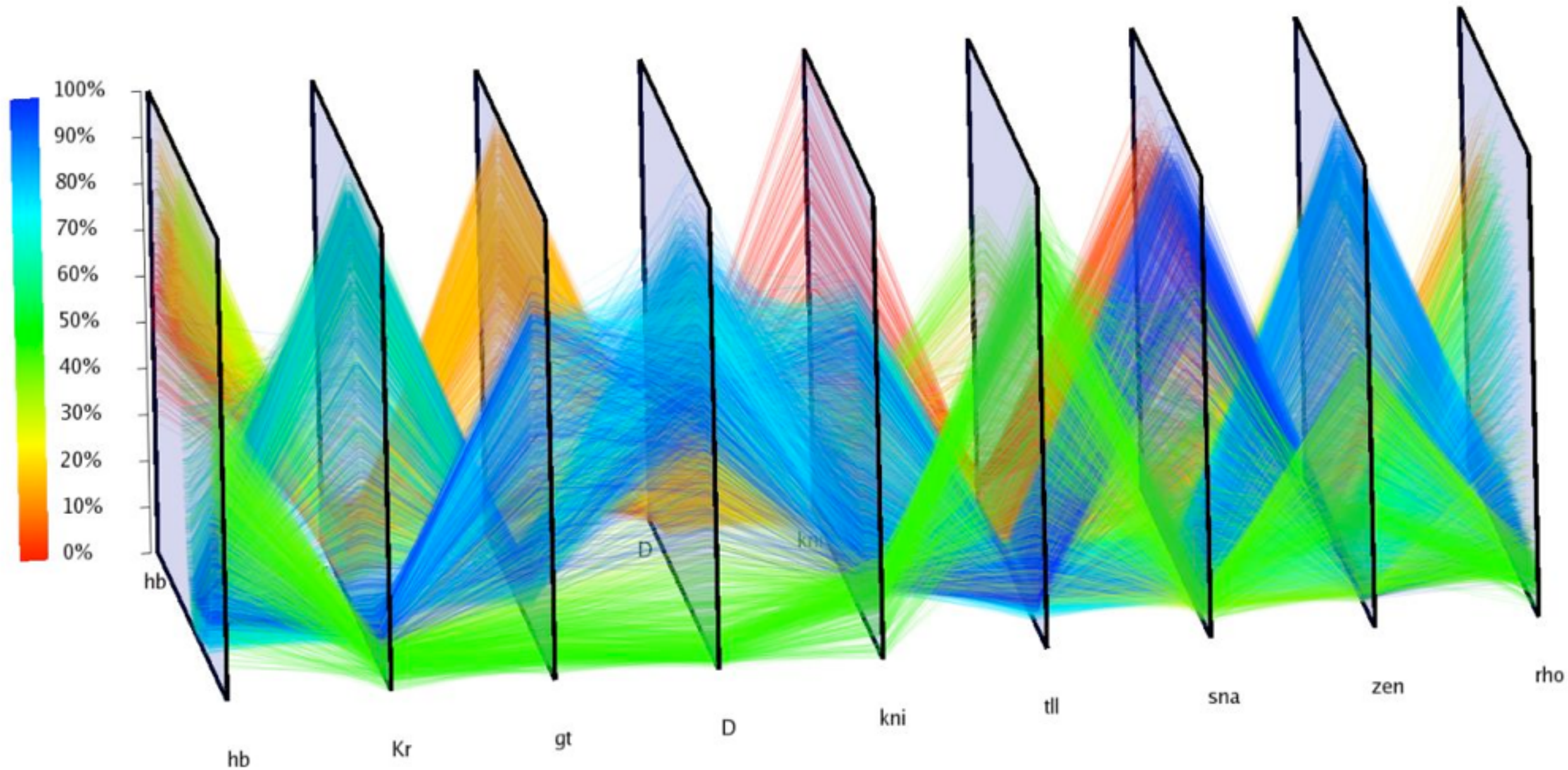
https://en.wikipedia.org/wiki/Box_plot

Parallel Coordinates Plot



http://www.nature.com/nmeth/journal/v9/n10/fig_tab/nmeth.2186_F1.html

3D Parallel Coordinates Plot



http://stackoverflow.com/questions/17156103/how-to-make-3d-parallel-coordinates-plot-in-r

Summary

Three things

if you can remember only three…

- Visualization help discover data
- Avoid pie charts (use histogram, bartplot …)
- Show more than 1 or 2 variables

Thank you