# Describing Data

## *A Visual Survival Kit*

IX International Physics Conference in honor of F.Pianca
December 22, 2021

**Jānis Lazovskis**, Riga Business School (Riga Technical University)
**Kristaps Bergfelds**, Institute of Numerical Modelling (University of Latvia)

Based on a talk given January 6, 2021

# Overview

1. Personal background & motivation
2. Tools: data, sources, software
3. Visual elements
4. Critisizing

*through examples

# Personal

Career path

- University of Waterloo
- University of Illinois at Chicago
- University of Aberdeen
- Riga Technical University

Outside of academia

- Family
- Sports
- Social service


**2009 - 2014**
BMath, MMath


**2014 - 2019**
PhD in Mathematics


**2019 - 2020**
Postdoctoral fellow


**2020 -**
Assistant Professor
BITL Program Academic Advisor

# Motivation

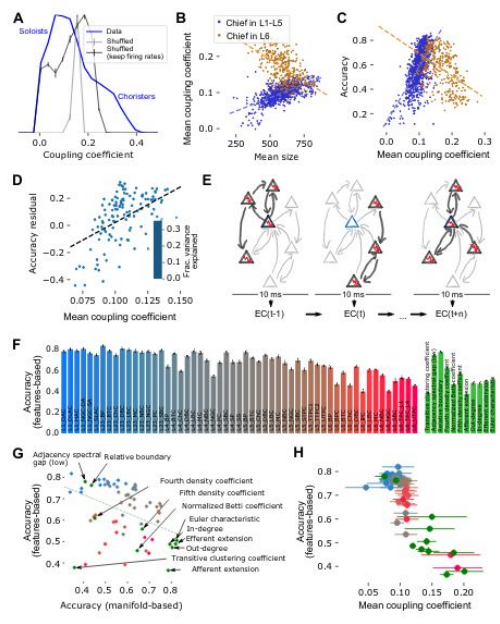
**The first look**

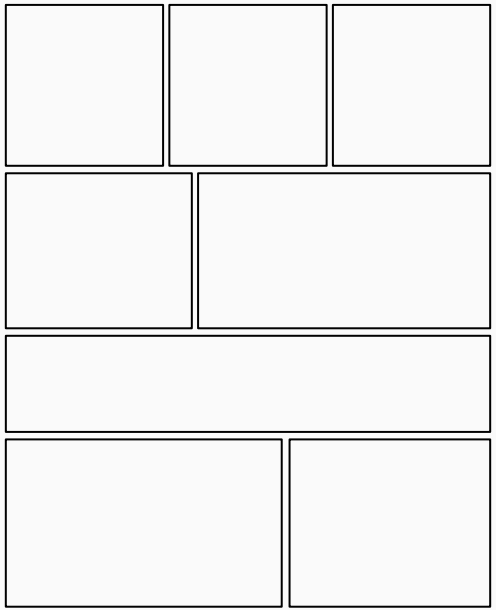
Seeing data:

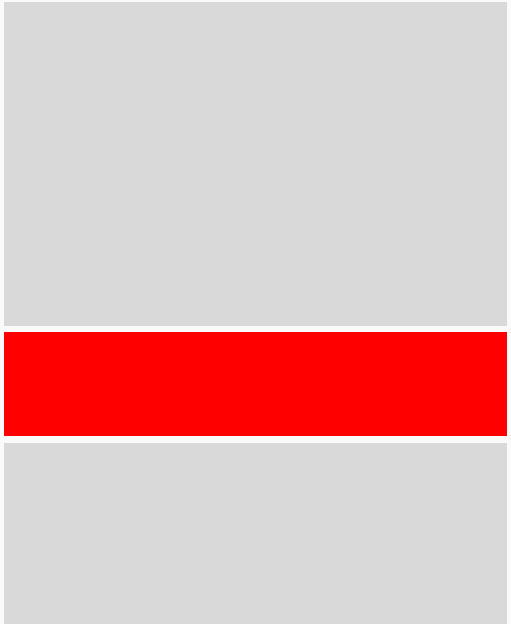
- What is going on?
- What should I be seeing?

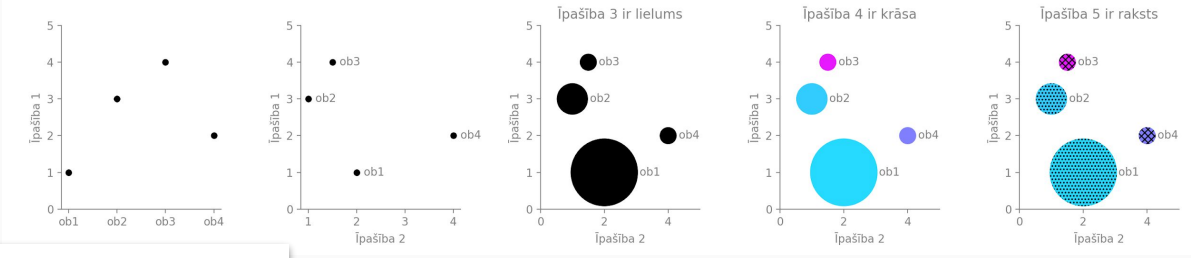
Presenting data:

- What do I want to say?
- What does the reader expect?
- What is unnecessary?

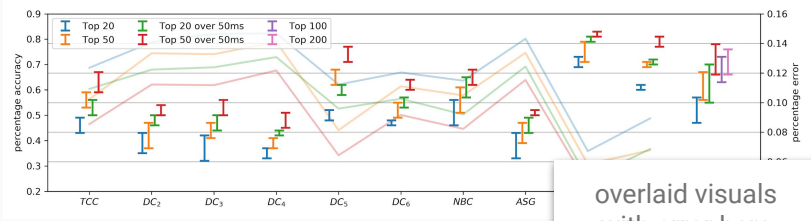


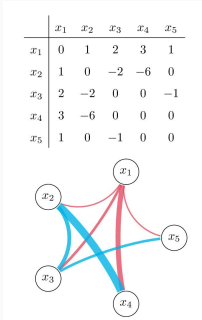




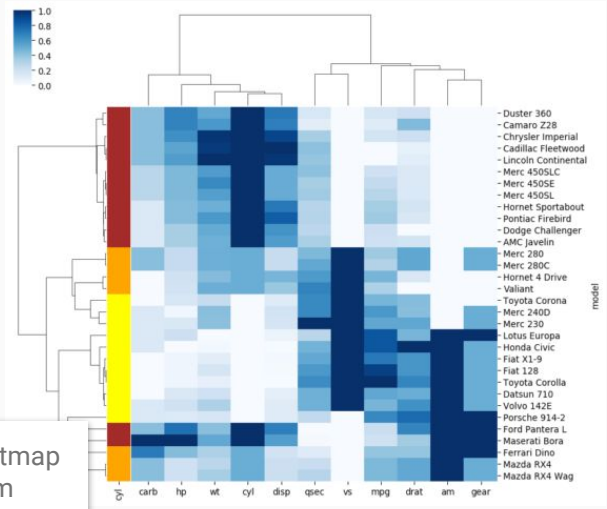

visualizing properties




overlaid visuals with error bars




relationships as a graph




properties as a heatmap and dendrogram

# Motivation

**Data relationships**

What is the first thing you notice?

- Color
- Contrast
- Text / whitespace

# Tools: data types

**Different requirements = different uses**

Objects and their properties

- Measurements, surveys
- parameters

Dynamically changing properties

- Repeated measurments
- Change of parameters

Relationships among objects

- Multiple dependencies
- Directional data

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | song_name | danceability | energy | key | loudness | mode |
| 2 | Mercury: Retrograde | 0.831 | 0.814 | 2 | -7.364 | 1 |
| 3 | Pathology | 0.719 | 0.493 | 8 | -7.23 | 1 |
| 4 | Symbiote | 0.85 | 0.893 | 5 | -4.783 | 1 |
| 5 | ProductOfDrugs (Prod. The Virus and Antidote) | 0.476 | 0.781 | 0 | -4.71 | 1 |
| 6 | Venom | 0.798 | 0.624 | 2 | -7.668 | 1 |
| 7 | Gatteka | 0.721 | 0.568 | 0 | -11.295 | 1 |
| 8 | kamikaze (+ pulse) | 0.718 | 0.668 | 8 | -4.162 | 1 |
| 9 | T.R.U. (Totally Rotten Underground) | 0.694 | 0.711 | 8 | -5.525 | 1 |
| 10 | I Put My Dick in Your Mental | 0.774 | 0.751 | 1 | -2.445 | 1 |
| 11 | Andromeda | 0.893 | 0.907 | 11 | -10.406 | 1 |
| 12 | BRAINFOOD | 0.864 | 0.365 | 8 | -10.219 | 1 |
| 13 | Troll Under the Bridge | 0.736 | 0.932 | 1 | -3.726 | 1 |
| 14 | 1000 Rounds | 0.825 | 0.761 | 8 | -5.389 | 1 |
| 15 | Sacrifice | 0.767 | 0.576 | 10 | -9.683 | 0 |
| 16 | Backpack | 0.765 | 0.726 | 5 | -5.58 | 1 |
| 17 | D(R)Own | 0.617 | 0.541 | 6 | -4.113 | 1 |
| 18 | Okay,ButThisIsTheLastTime | 0.755 | 0.298 | 1 | -15.032 | 1 |
| | | | 0.575 | 11 | -9.635 | 1 |

**table**
unordered objects with properties

**weighted adjacency matrix**
objects and pairwise properties

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | -4 | -2 | -2 | -1 | -3 | -2 | 1 |
| -4 | 0 | 1 | 0 | 0 | -2 | 1 | -1 |
| -2 | 1 | 0 | -1 | -3 | 1 | 1 | 0 |
| -2 | 0 | -1 | 0 | -2 | -3 | -1 | -2 |
| -1 | 0 | -3 | -2 | 0 | -3 | -1 | -2 |
| -3 | -2 | 1 | -3 | -3 | 0 | -3 | -1 |
| -2 | 1 | 1 | -1 | -1 | -3 | 0 | -3 |
| 1 | -1 | 0 | -2 | -2 | -1 | -3 | 0 |

```
{
  "particles": {
    "number": {
      "value": 80,
      "density": {
        "enable": true,
        "value_area": 800
      }
    },
    "color": {
      "value": "#01b6ed"
    },
    "shape": {
      "type": "circle",
      "stroke": {
        "width": 0,
        "color": "#01b6ed"
      },
      "polygon": {
        "nb_sides": 10
      },
      "image": {
        "src": "img/github.svg",
        "width": 100,
        "height": 100
      }
    },
    "opacity": {
      "value": 0.5,
      "random": false,
      "anim": {
        "enable": false,
```

**dictionary**
hiereachical properties

| date | 2017-01-01 | 2017-01-02 | 2017-01-03 | 2017-01-04 | 2017-01-05 | 2017-01-06 | 2017-01-07 | 2017-01-08 | 2017-01-09 | 2017-01-10 | 2017-01-11 | 2017-01-12 | 2017-01-13 | 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| meantemp | 15.91 | 18.5 | 17.11 | 18.7 | 18.39 | 19.32 | 14.71 | 15.68 | 14.57 | 12.11 | 11 | 11.79 | 13.24 | |
| humidity | 85.87 | 77.22 | 81.89 | 70.05 | 74.94 | 79.32 | 95.83 | 83.53 | 80.81 | 71.94 | 72.11 | 74.58 | 67.06 | |
| wind_speed | 2.74 | 2.89 | 4.02 | 4.55 | 3.3 | 8.68 | 10.04 | 1.95 | 6.54 | 9.36 | 9.77 | 6.63 | 6.44 | |
| meanpressure | 59 | 1018.28 | 1018.33 | 1015.7 | 1014.33 | 1011.77 | 1011.38 | 1015.55 | 1015.95 | 1016.89 | 1016.78 | 1016.37 | 1017.53 | |

**time series**
unordered objects with ordered properties

# Tools: data sources

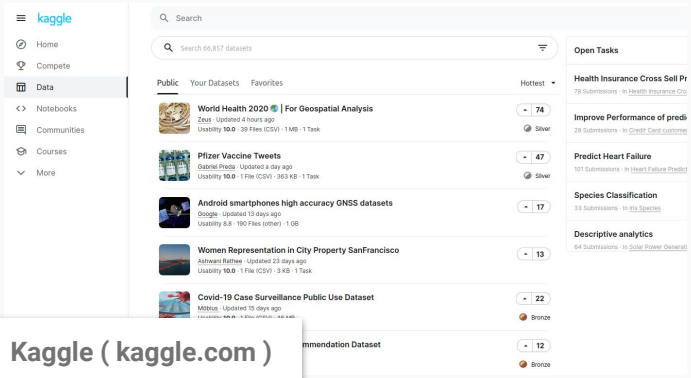
**Examples of data banks**

General

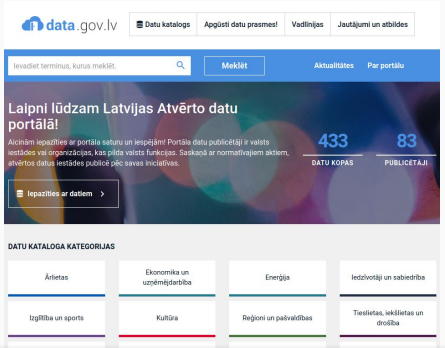
- Kaggle
- Atvērto Datu Portāls
- EU Open Data Portal

Specific

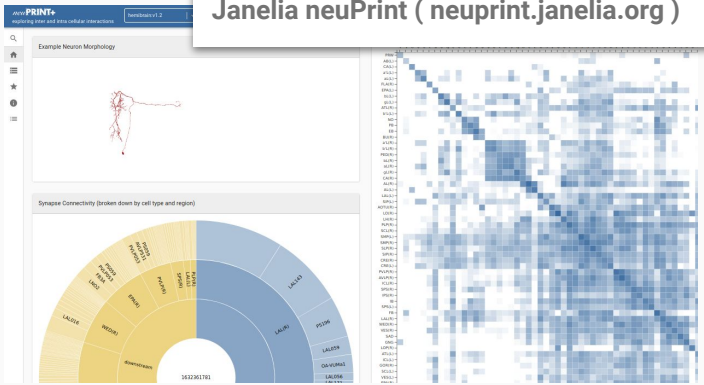
- Allen Brain Map
  portal.brain-map.org
- Blue Brain Nexus
  bluebrainnexus.io
- Janelia
  neuprint.janelia.org



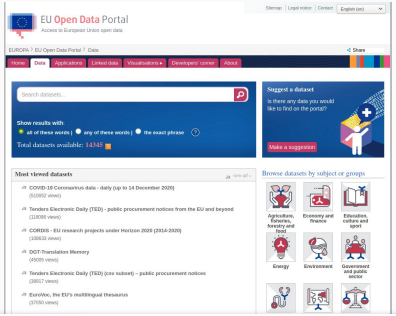

Kaggle ( kaggle.com )



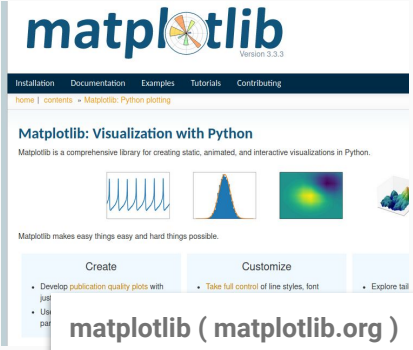

Atvērto Datu Portāls ( data.gov.lv )




Janelia neuPrint ( neuprint.janelia.org )



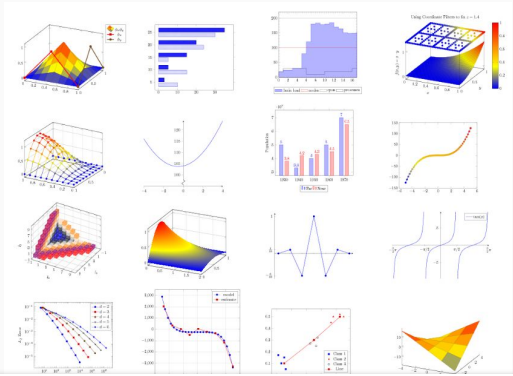

EU Open Data Portal ( data.europa.eu )




matplotlib ( matplotlib.org )


Plotly ( dash-gallery.plotly.host )


Python Graph Gallery
( python-graph-gallery.com )



PGFPlots gallery ( pgfplots.sourceforge.net )

# Tools: inspiration

**Collections of different examples**

Ready-made templates

- Plotly
- Chart.js, D3.js

Fine detail control

- The Python Graph Gallery
- matplotlib

Extreme detail control

- TeXample.net
- PGFPlots Gallery

# Tools: software

**Data management and exploration**

## Managing files

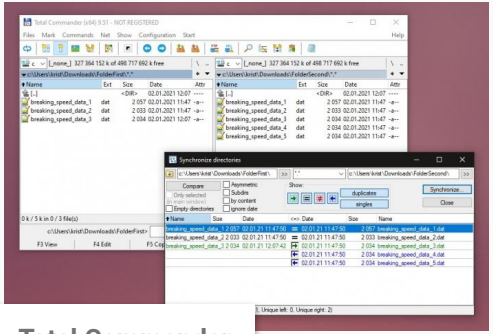
- Total Commander

## Reading text and numbers

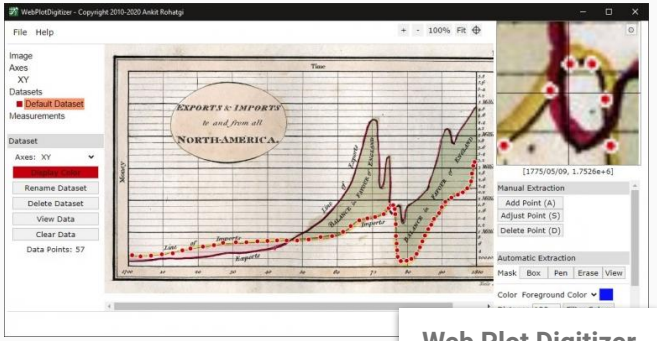
- Excel, Google Sheets
- Sublime Text, Notepad++

## Reading data from images

- Web Plot Digitizer
- GIMP (Gnu Image Manipulation Program)

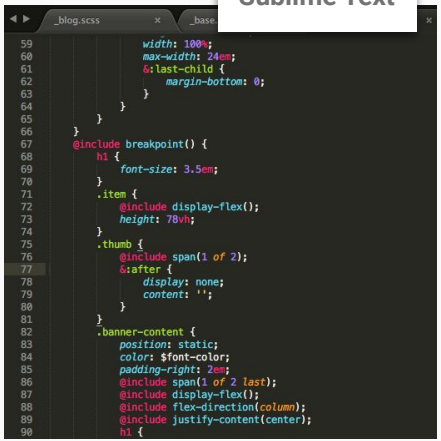


Total Commander

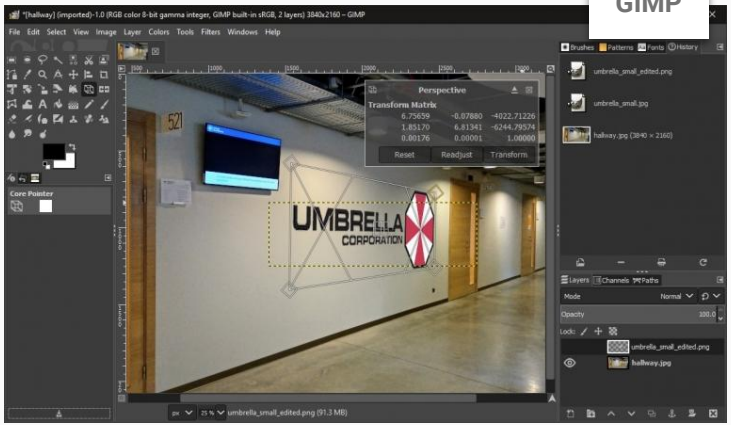


Web Plot Digitizer

Sublime Text



GIMP



# Tools: software

**Some coding required**

Plug and chug

- MS Excel, Google Sheets
- MS Word
- Canva
- Plotly, Chart.js

Unlimited control

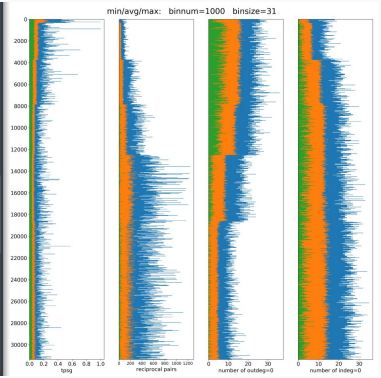
- Python: CSV, JSON, pandas
- LaTeX: TikZ, PGFPlots

```
>>> import pandas as pd
>>> df = pd.read_pickle('mc2_newparams2.pkl')
>>> df
      tpsg_strong_low  tpsg_strong_high  tpsg_low  tpsg_high  reciprocal_pairs  zero_outdeg  zero_indeg
0            0.456522          0.456522  0.266667   0.266667                 1           15           3
1            0.100756          0.100756  0.094926   0.094926                44           22          11
2            0.119700          0.119700  0.111111   0.111111                25           20           7
3            0.000000          0.000000  0.333333   0.333333                 0            9           1
4            0.133989          0.133989  0.114300   0.114300                20           23           3
...               ...               ...       ...        ...               ...          ...         ...
31341        0.068787          0.068787  0.068950   0.068950               162            3          16
31342        0.140519          0.140519  0.139227   0.139227                40            4           3
31343        0.030780          0.030780  0.030934   0.030934               803            3          16
31344        0.174490          0.174490  0.168207   0.168207                17            9          10
                                    ...  0.038262   0.038262               590            2          20
```

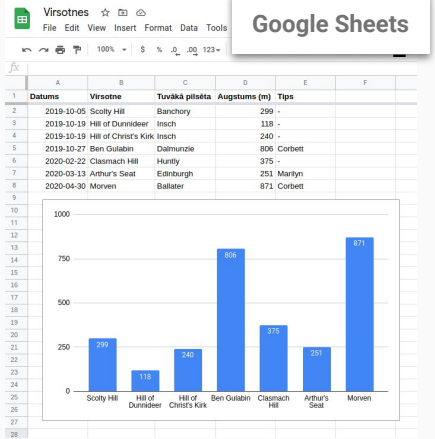
Python + pandas

```
def plot_sorted():
    # Load dataframe and initiate figure
    print('Loading dataframes',flush=True)
    df = pd.read_pickle(database_dir+'mc2_newparams.pkl')
    df_strong = pd.read_pickle(database_dir+'mc2_tpsgstrong.pkl')
    fig = plt.figure(figsize=(16,16)) # default is (8,6)

    # tpsg
    print('computing and plotting: tpsg',flush=True)
    ax1 = fig.add_subplot(4,1,1); plt.xscale('log');
    data_tpsg = df['tpsg_high'].values
    data_tpsg.sort()
    ax1.plot(data_tpsg[::-1],lw=3, label='tpsg', alpha=.8)
    data_tpsg_strong = df_strong['tpsg_strong_high'].values
    data_tpsg_strong.sort()
    ax1.plot(data_tpsg_strong[::-1],lw=3, label='tpsg_strong', alpha=.8)
    data_tt = [abs(df.loc[i,'tpsg_high']-df_strong.loc[i,'tpsg_strong_hi
    data_tt.sort()
    ax1.plot(data_tt[::-1],lw=3, label='difference', alpha=.8)
    ax1.set_ylabel('tpsg',fontsize=16); plt.yticks(fontsize=16)
    ax1.legend(fontsize=16)

    # reciprocal edges
    print('computing and plotting: reciprocal edges',flush=True)
    ax2 = fig.add_subplot(4,1,2); plt.xscale('log')
    data_re = df['reciprocal_pairs'].values
    data_re.sort()
    ax2.plot(data_re[::-1],lw=3)
    ax2.set_ylabel('reciprocal pairs',fontsize=16); plt.yticks(fontsize=

    # outdegree
    print('computing and plotting: out degree',flush=True)
    ax3 = fig.add_subplot(4,1,3); plt.xscale('log')
    data_od = df['zero_outdeg'].values
    data_od.sort()
    ax3.plot(data_od[::-1],lw=3)
    ax3.set_ylabel('number of outdeg=0',fontsize=16); plt.yticks(fontsiz
```

Python + matplotlib

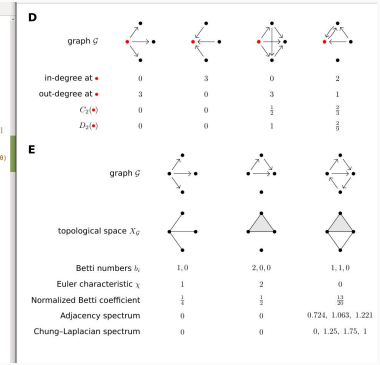



Google Sheets

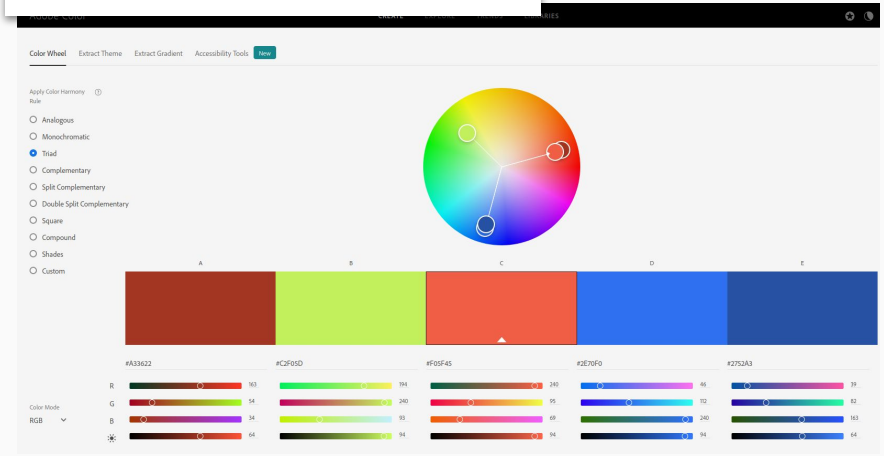



```
fig-topo.tex
\draw(edge) (\y) to [bend right=30] (\x);
}
\draw(edge) (c) to [bend right=45] (d);
\node[scale=\labelscale] at (.4,-.5) {strongly connected};
\end{scope}
\end{scope}
%%%
%%%
%%% D
%%%
%%%
\draw[\bboxcol] (-2.5,-1) rectangle (24,-6);
\node[anchor=north west] at (-2.5,-1) {\large\textbf{D}};
\begin{scope}[shift={(.4,-4.0)}]
% labels
\node[scale=\labelscale,anchor=east] at (0,2.6) {graph $\mathcal{G}$};
\node[scale=\labelscale,anchor=east] at (0,\ylev) {in-degree at \tikz[baseline=-3pt]{\fill[red]
  (0,0) circle (\rad);};
\node[scale=\labelscale,anchor=east] at (0,\ylev) {out-degree at
  \tikz[baseline=-3pt]{\fill[red] (0,0) circle (\rad);};
\node[scale=\labelscale,anchor=east] at (0,\yylev) {$C_2($\tikz[baseline=-3pt]{\fill[red] (0,0)
  circle (\rad);})$};
\node[scale=\labelscale,anchor=east] at (0,\yyylev) {$D_2($\tikz[baseline=-3pt]{\fill[red]
  (0,0) circle (\rad);})$};
% g0
\begin{scope}[shift={(1,2.6)}]
\foreach \x/\y/\c in {0/0/a/red, 1/0/b/black, .5/.7/c/black, .5/-.7/d/black}{
  \coordinate (\n) at (\x,\y);
  \fill[\c] (\n) circle (\rad);
}
\foreach \x/\y in {a/b,a/c,a/d}{
  \draw(edge) (\x)--(\y);
}
\node[scale=\labelscale] at (.5,\ylev-2.6) {$81$};
\node[scale=\labelscale] at (.5,\yylev-2.6) {$35$};
\node[scale=\labelscale] at (.5,\yyylev-2.6) {$95$};
\node[scale=\labelscale] at (.5,\yyyylev-2.6) {$85$};
\end{scope}
% g1
\begin{scope}[shift={(13.5,2.6)}]
\foreach \x/\y/\c in {0/0/a/red, 1/0/b/black, .5/.7/c/black, .5/-.7/d/black}{
  \coordinate (\n) at (\x,\y);
  \fill[\c] (\n) circle (\rad);
}
```

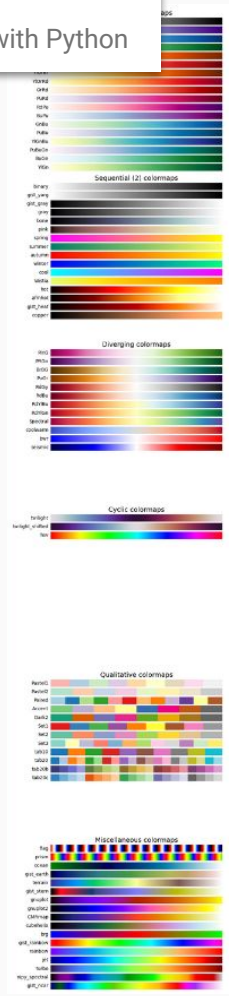
LaTeX + TikZ

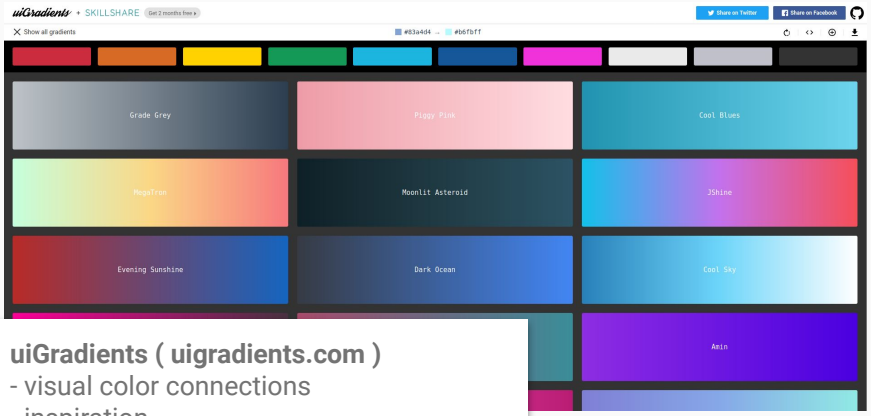


**Adobe Color Wheel ( color.adobe.com )**
- similar color choices
- accessibility tools for color impairment




**matplotlib colormap**
- color palettes for use with Python




**uiGradients ( uigradients.com )**
- visual color connections
- inspiration

# Elements

## Color choices

Emphasize

- Red, green
- Complementary colors

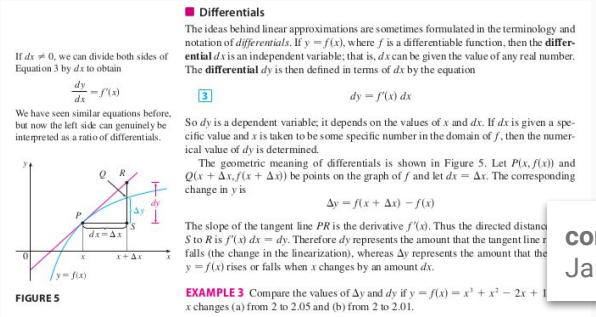
Reduce importance

- Light
- SImilar to background

Enrich

- Pattern overlay

** Organizations have style guides



■ **Differentials**

The ideas behind linear approximations are sometimes formulated in the terminology and notation of *differentials*. If $y = f(x)$, where $f$ is a differentiable function, then the **differential** $dx$ is an independent variable; that is, $dx$ can be given the value of any real number. The **differential** $dy$ is then defined in terms of $dx$ by the equation

$$dy = f'(x)\,dx \quad (3)$$

If $dx \neq 0$, we can divide both sides of Equation 3 by $dx$ to obtain

$$\frac{dy}{dx} = f'(x)$$

We have seen similar equations before, but now the left side can genuinely be interpreted as a ratio of differentials.

So $dy$ is a dependent variable; it depends on the values of $x$ and $dx$. If $dx$ is given a specific value and $x$ is taken to be some specific number in the domain of $f$, then the numerical value of $dy$ is determined.

The geometric meaning of differentials is shown in Figure 5. Let $P(x, f(x))$ and $Q(x + \Delta x, f(x + \Delta x))$ be points on the graph of $f$ and let $dx = \Delta x$. The corresponding change in $y$ is

$$\Delta y = f(x + \Delta x) - f(x)$$

The slope of the tangent line $PR$ is the derivative $f'(x)$. Thus the directed distance $S$ to $R$ is $f'(x)\,dx = dy$. Therefore $dy$ represents the amount that the tangent line rises or falls (the change in the linearization), whereas $\Delta y$ represents the amount that the curve $y = f(x)$ rises or falls when $x$ changes by an amount $dx$.

FIGURE 5

**EXAMPLE 3** Compare the values of $\Delta y$ and $dy$ if $y = f(x) = x^3 + x^2 - 2x + 1$ and $x$ changes (a) from 2 to 2.05 and (b) from 2 to 2.01.

**consistency across all visuals**
Jame Stewart, *"Calculus"*

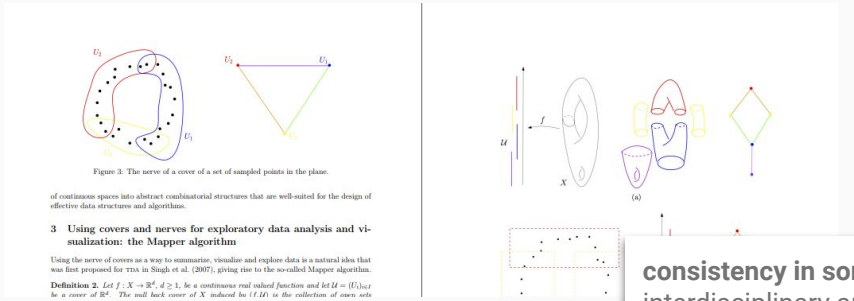


Figure 3: The nerve of a cover of a set of sampled points in the plane.

of continuous spaces into abstract combinatorial structures that are well-suited for the design of effective data structures and algorithms.

**3 Using covers and nerves for exploratory data analysis and visualization: the Mapper algorithm**

Using the nerve of covers as a way to summarize, visualize and explore data is a natural idea that was first proposed for TDA in Singh et al. (2007), giving rise to the so-called Mapper algorithm.

**Definition 2.** *Let $f : X \to \mathbb{R}^d$, $d \geq 1$, be a continuous real valued function and let $\mathcal{U} = (U_i)_{i \in I}$ be a cover of $\mathbb{R}^d$. The pull back cover of $X$ induced by $(f, \mathcal{U})$ is the collection of open sets*



(a)

**consistency in some visual types**
interdisciplinary articles

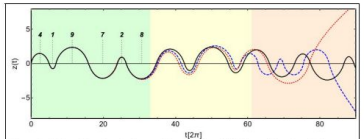


**Figure 2:** Motion of the planet above and beyond the orbital plane of the stars in the Sitnikov problem. x-axis shows time in units of revolution of the primaries. y-axis gives distance from planet above and beyond the orbital plane. The plot shows motion of the planet for three different initial conditions. In the beginning, the motion is very similar, but after about 33 revolutions of the primaries, first differences appear and after 60 revolutions the three initial conditions develop into totally different types of motion. The red line shows a planet moving very far away from the stars in one direction; the blue line shows the same behaviour with a planet moving away in the other direction. In both cases a very long and possible endless winter would begin. The black line shows a planet that continues to have a more or less regular sequence of seasons.

**3.2. Non-mathematical explanation**

For those who - like the maesters of Westeros - refuse to indulge in the wonders of celestial mechanics, let us explain the main point in simpler words. Let's define the time the two stars need for one revolution around the barycenter as one "year". We can then define any infinitely long and arbitrary sequence of whole numbers. The theorem of Mosers stated

**Figure 3:** Visualization of how a Sitnikov system could look like. The planet oscillates between a star and a black hole.
(Created with Universe Sandbox² http://universesandbox.com/)

**4. A universal mechanism for every s**

Close to the orbital plane of the stars th
large amount of stellar energy. If the pla
flux will be reduced. In other words: the
the warmer it will be. A planet that oscil
regular seasons like on Earth. But as th
oscillate with any sequence of duration
longer the cold season will be. To reproduce a seemingly chaotic sequence of summers and

**inconsistency**
most technical articles

# Elements

**Scheme consistency**

Often overlooked

- *Done well:* adds to the story being told
- *Done poorly:* distracts from the main ideas

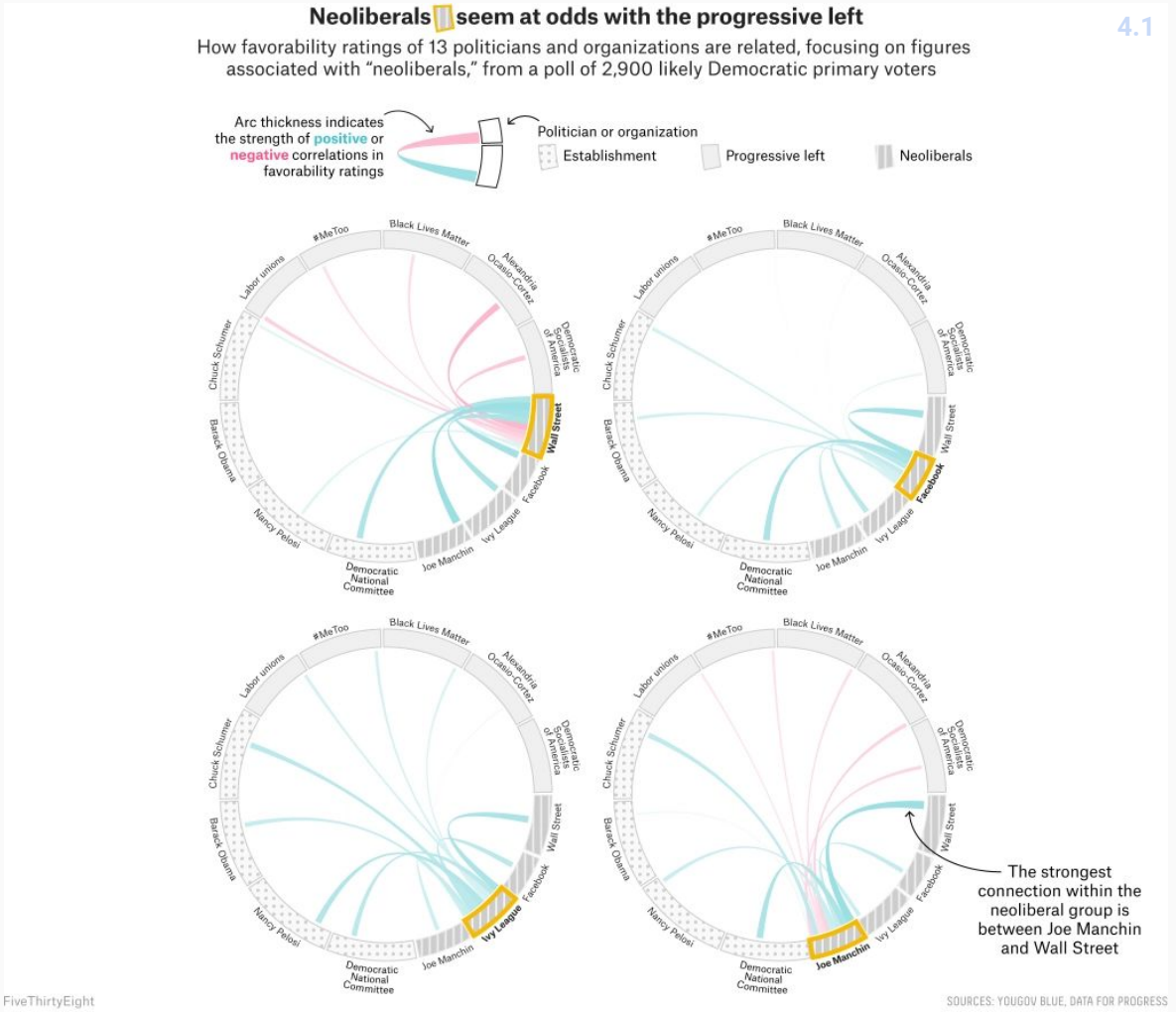
** The data **does not** speak for itself

# Analysis: color

**Ryan Best, Shom Mazumder, FiveThirtyEight**

- Main idea in title
- Soft color scheme

- Small value = light color
- Extra information = patterns



# Analysis: overload

**Graphic Detail blog, The Economist**

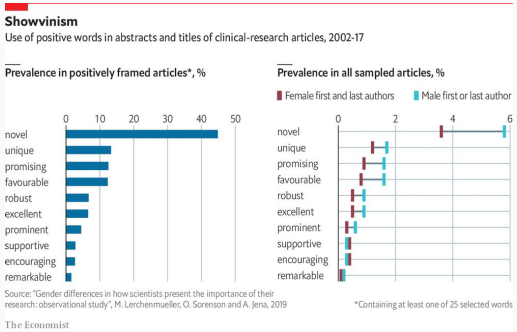
High volume of information

- Is it worth showing?
- What does it add to message?
- What are the common ideas?
- How to best combine it?

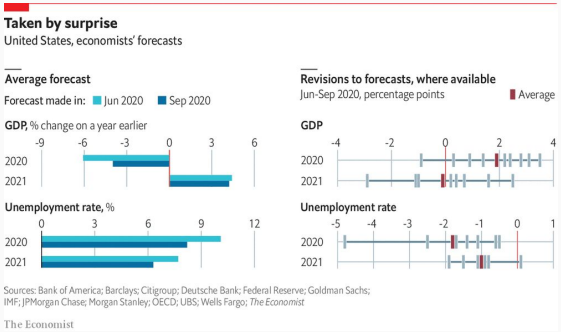
Interactive information

- Plotly
- D3.js

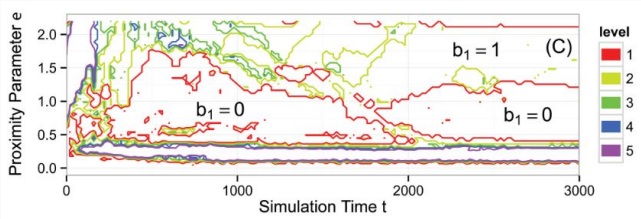
separate visuals: good

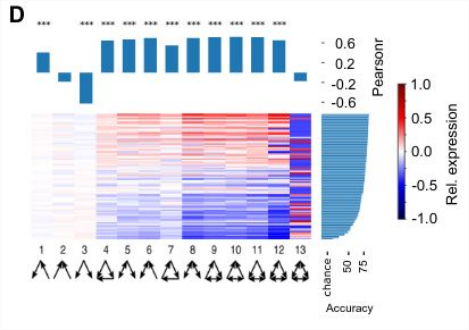



separate visuals: bad

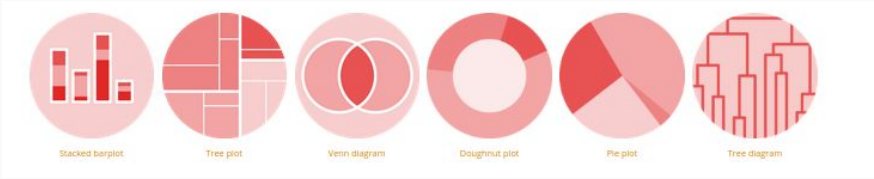



academic visuals: need knowledge to evaluate



1. Background / Motivation
2. Tools / Elements / Analysis




Sources:

- The Economist Graphic Detail blog economist.com/graphic-detail
- FiveThirtyEight fivethirtyeight.com
- Adobe Color Wheel color.adobe.com/create/color-wheel
- The Python Graph Gallery python-graph-gallery.com
- PGFPlots Gallery pgfplots.sourceforge.net/gallery.html
- TikZ and PGF examples texample.net/tikz/examples
- Lisa Charlotte Rost github.com/lisacharlotterost/talk-slides
- James Gleick “The Information: A History, A Theory, A Flood”
- Topaz et al “Topological Data Analysis of Biological Aggregation Models”
- Reimann et al “Topology of synaptic connectivity constrains neuronal stimulus representation”