



# Principles of Data Visualization II

Eamonn Maguire  
CERN School of Computing, Romania  
September 2019

# HOW

## We have to be careful when mapping data to the visual world

Some visual channels are more effective for some data types over others.

Some data has a **natural mapping** that our brains expect given certain types of data

There are many visual tricks that can be observed due to how the visual system works

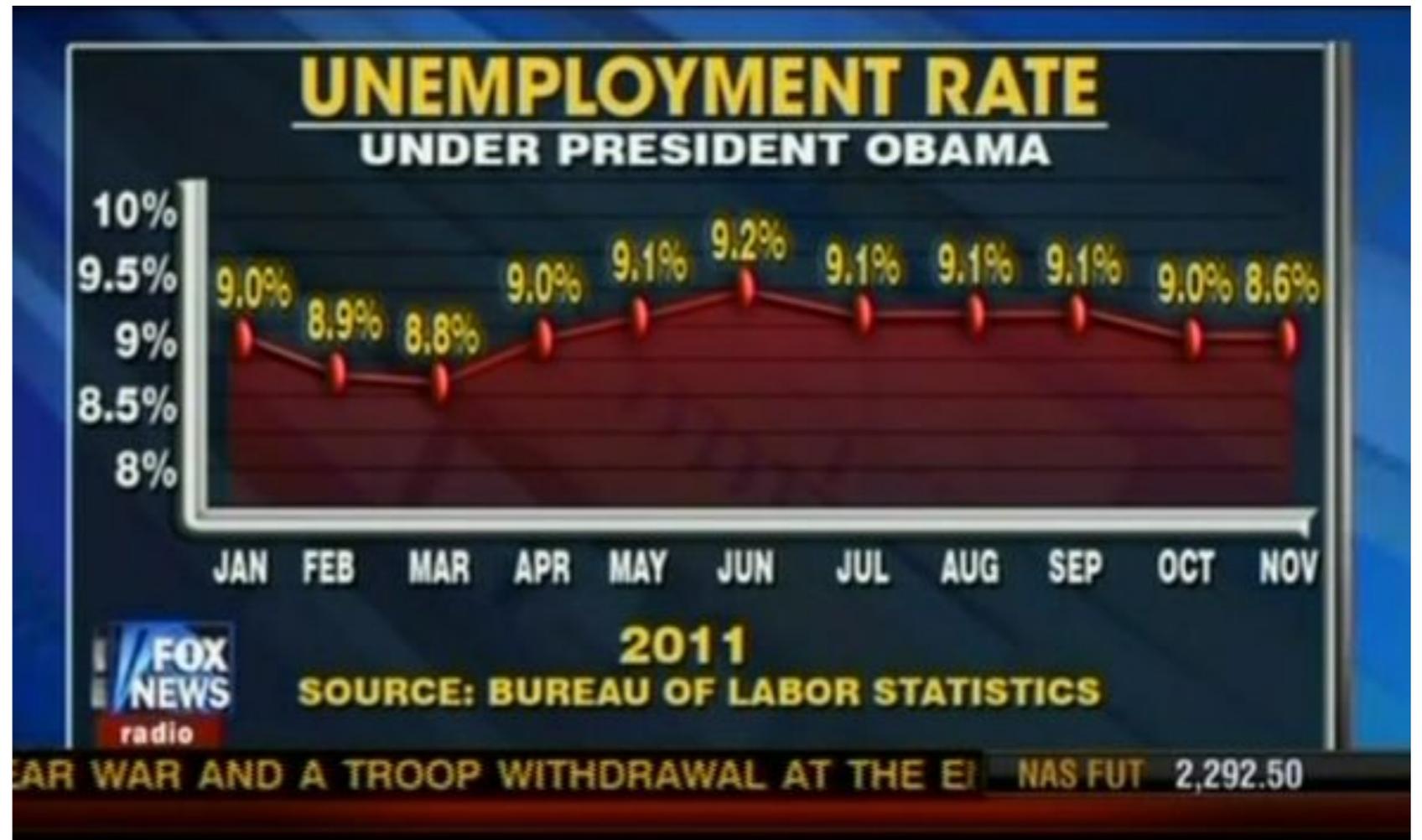
We don't see in 3D, and we have difficulties interpreting information on the Z-axis.

Colour

Scales

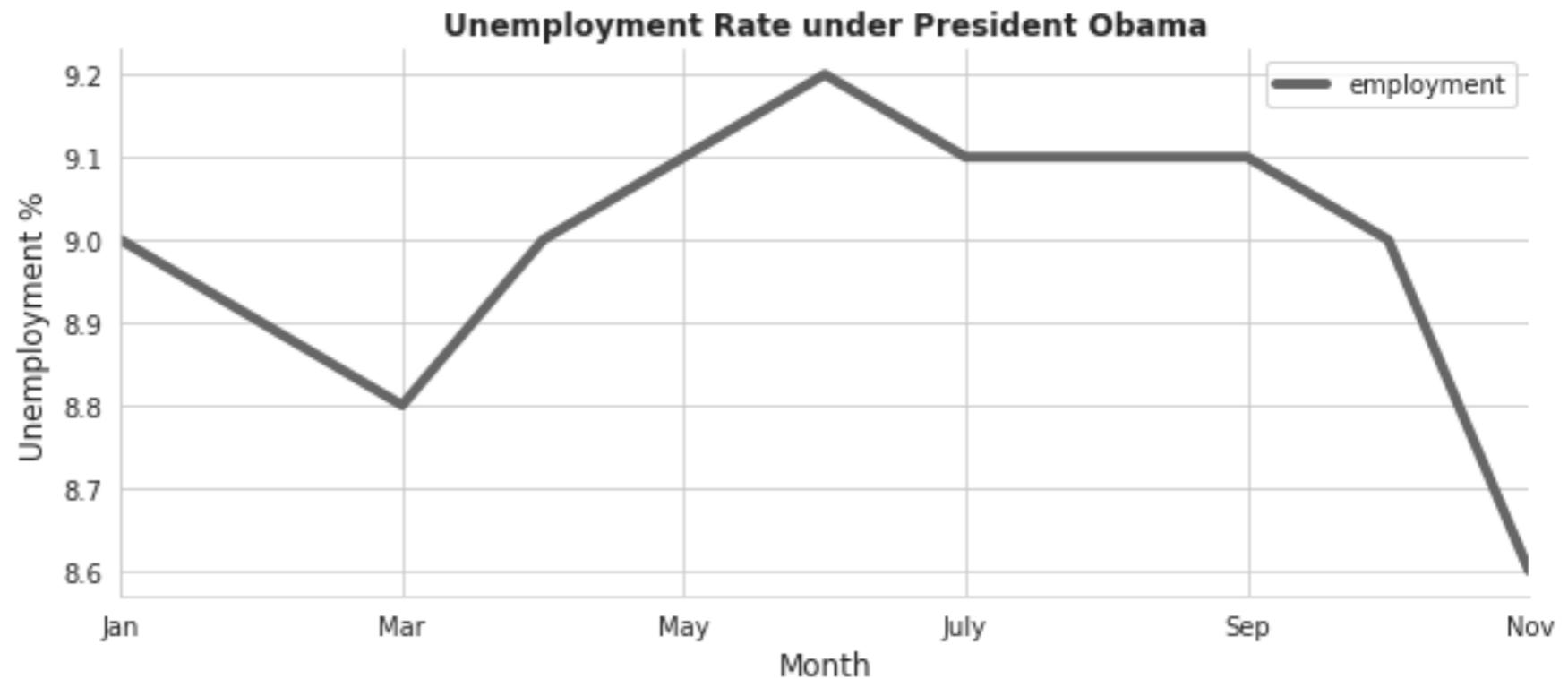
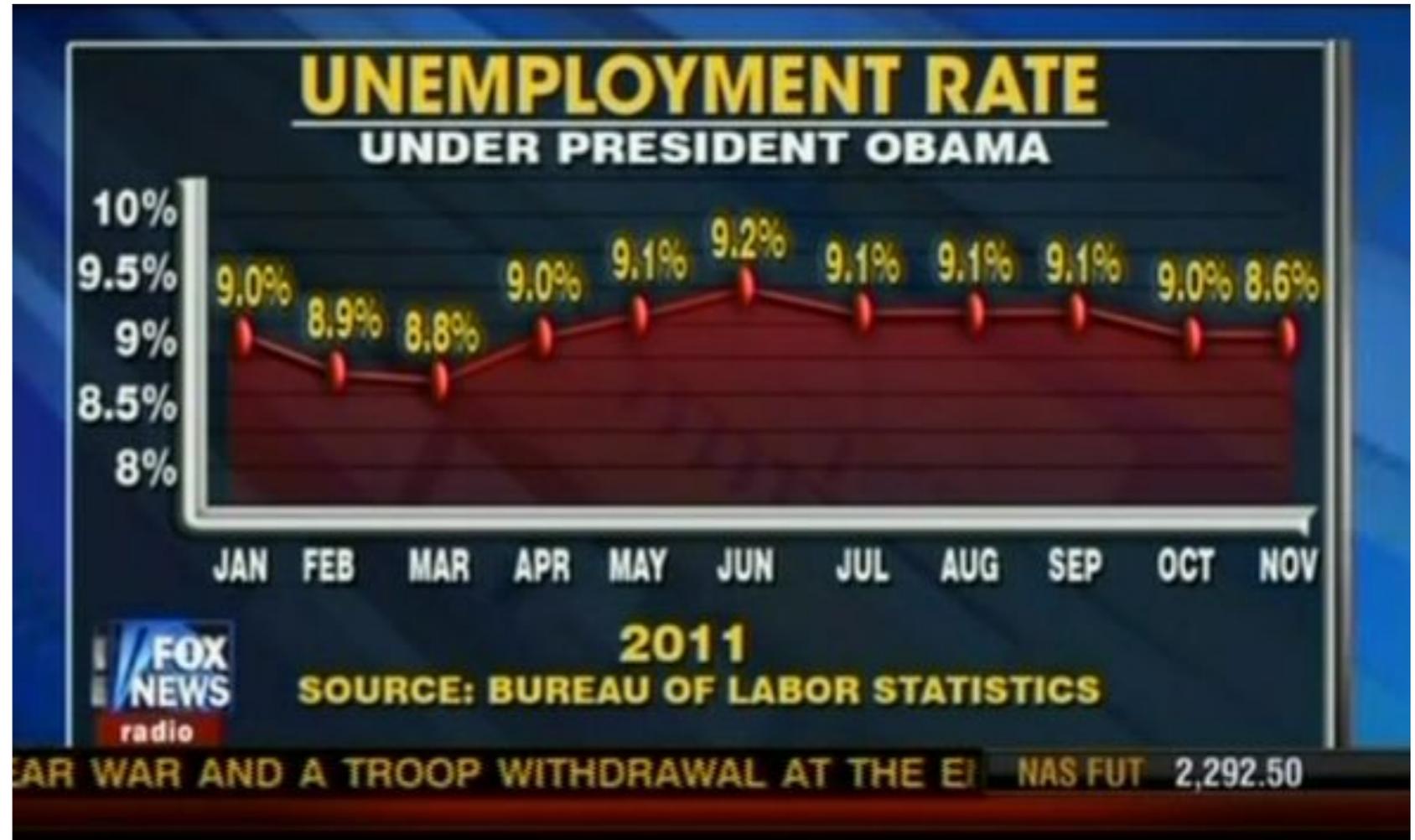
# Scales

Be aware of traps in visualizing data, when creating or reading. Especially with scales.



# Scales

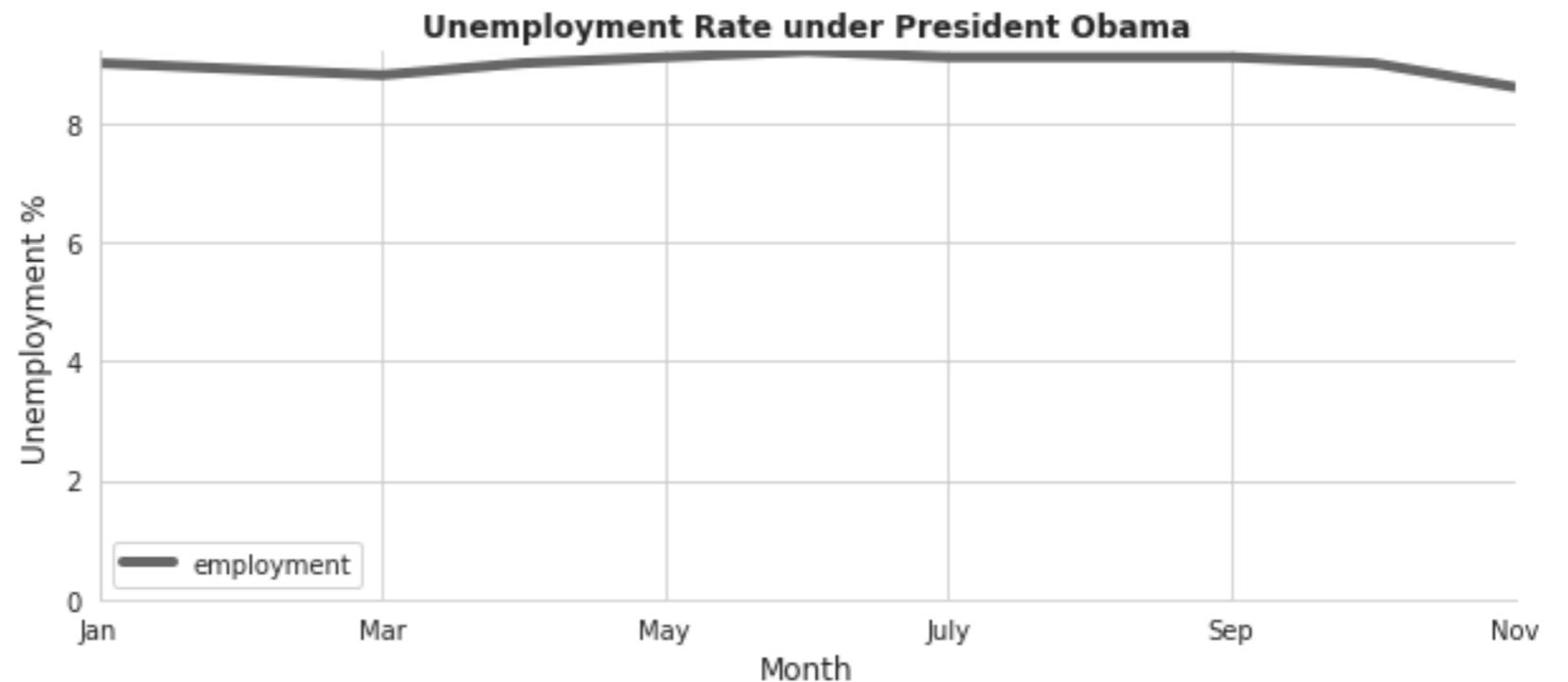
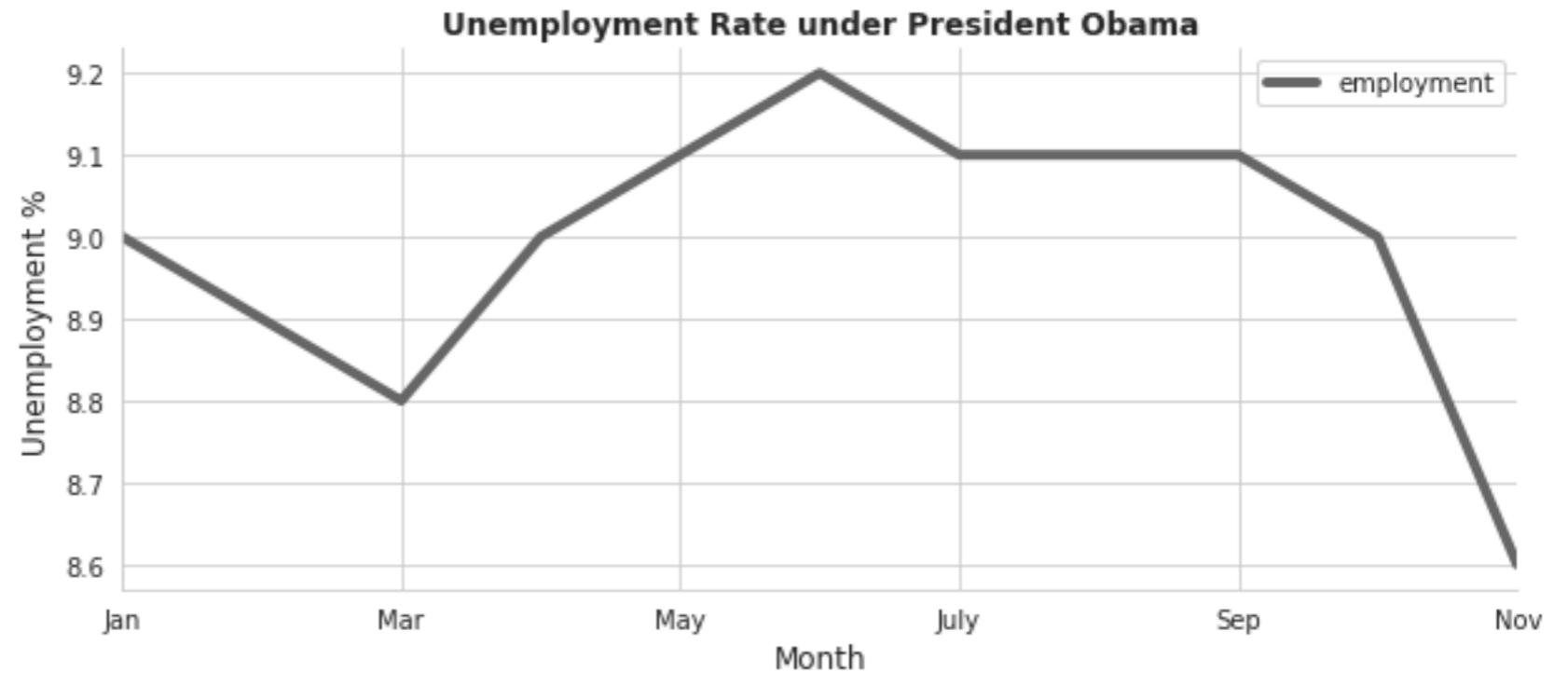
Be aware of traps in visualizing data, when creating or reading. Especially with scales.



# Scales

**You may have heard that charts should always have a zero baseline.**

But it's not always the case, especially if we never expect unemployment to go to zero!

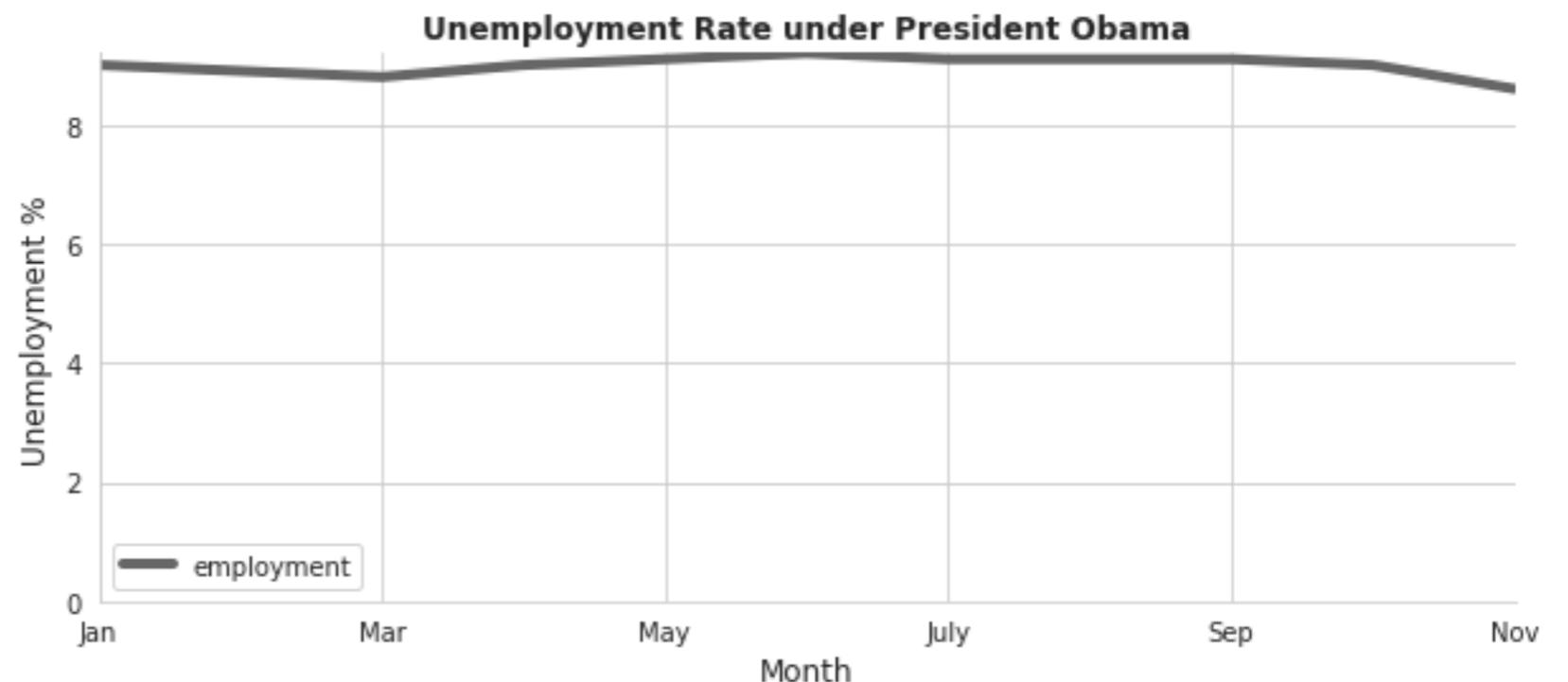
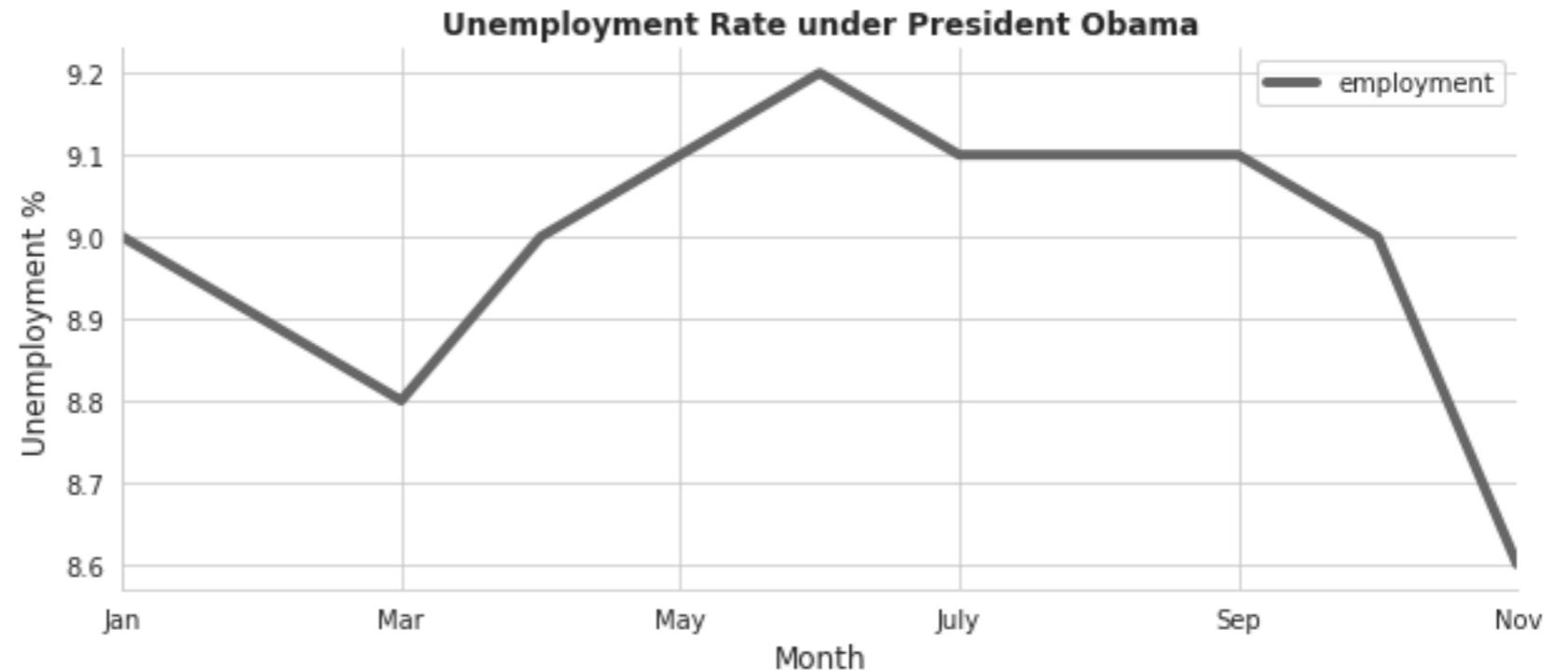


# Scales

**You may have heard that charts should always have a zero baseline.**

But it's not always the case, especially if we never expect unemployment to go to zero!

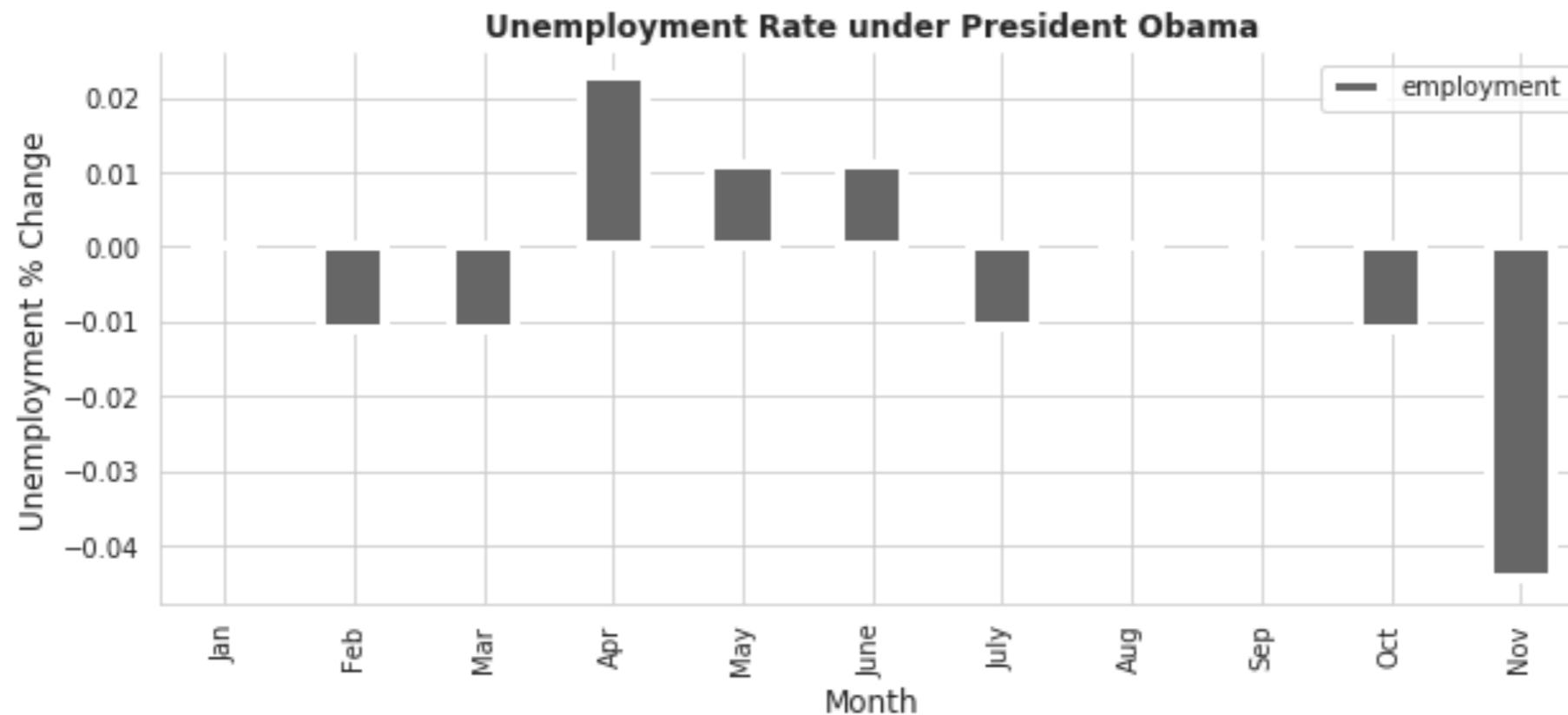
Zeroing the y axis here makes it difficult to see change....



# Scales

So, maybe we should think about other ways of showing change.

If our **task** is about **finding where there are intra-month changes**, then simply plotting the differences can be more informative.

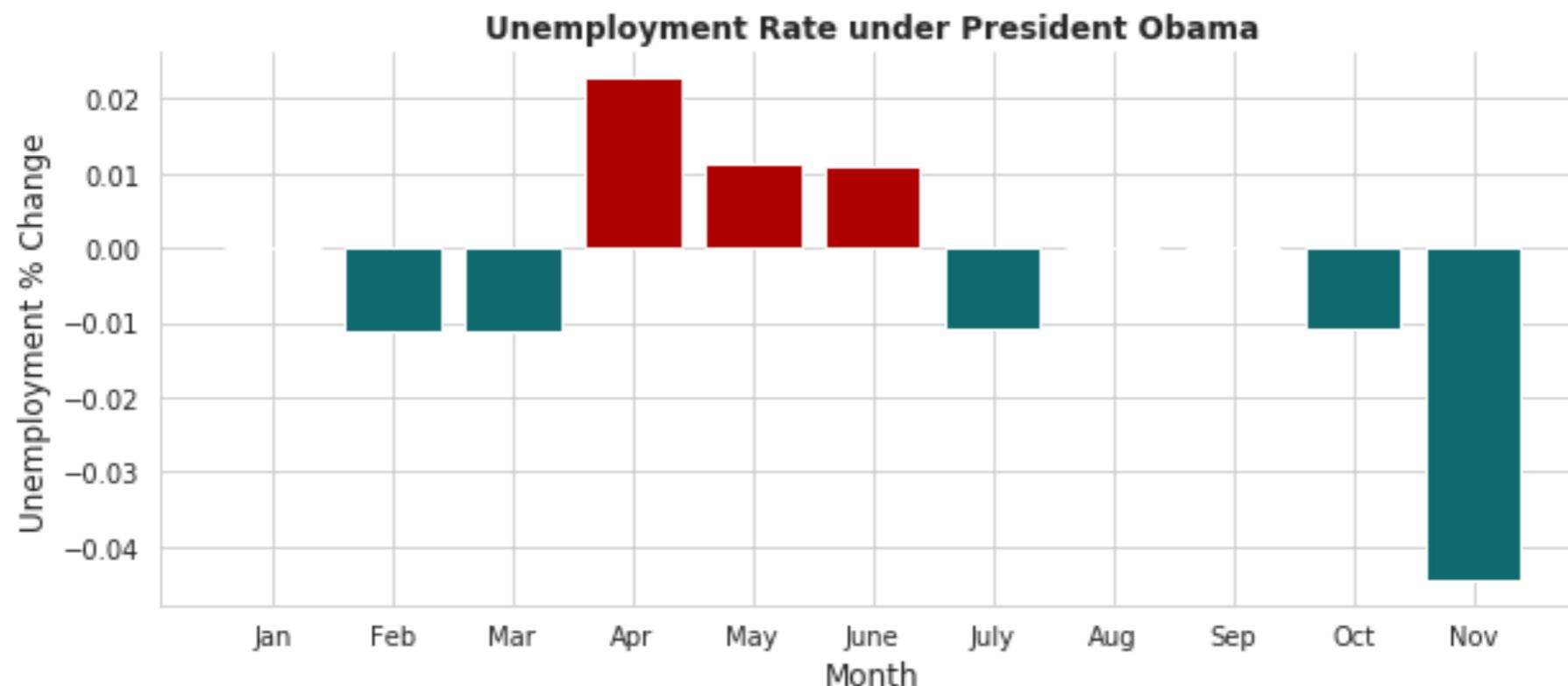


We can now see that the employment rate under Obama went down more than it went up, and that in November the drop was greatest...

# Scales

So, maybe we should think about other ways of showing change.

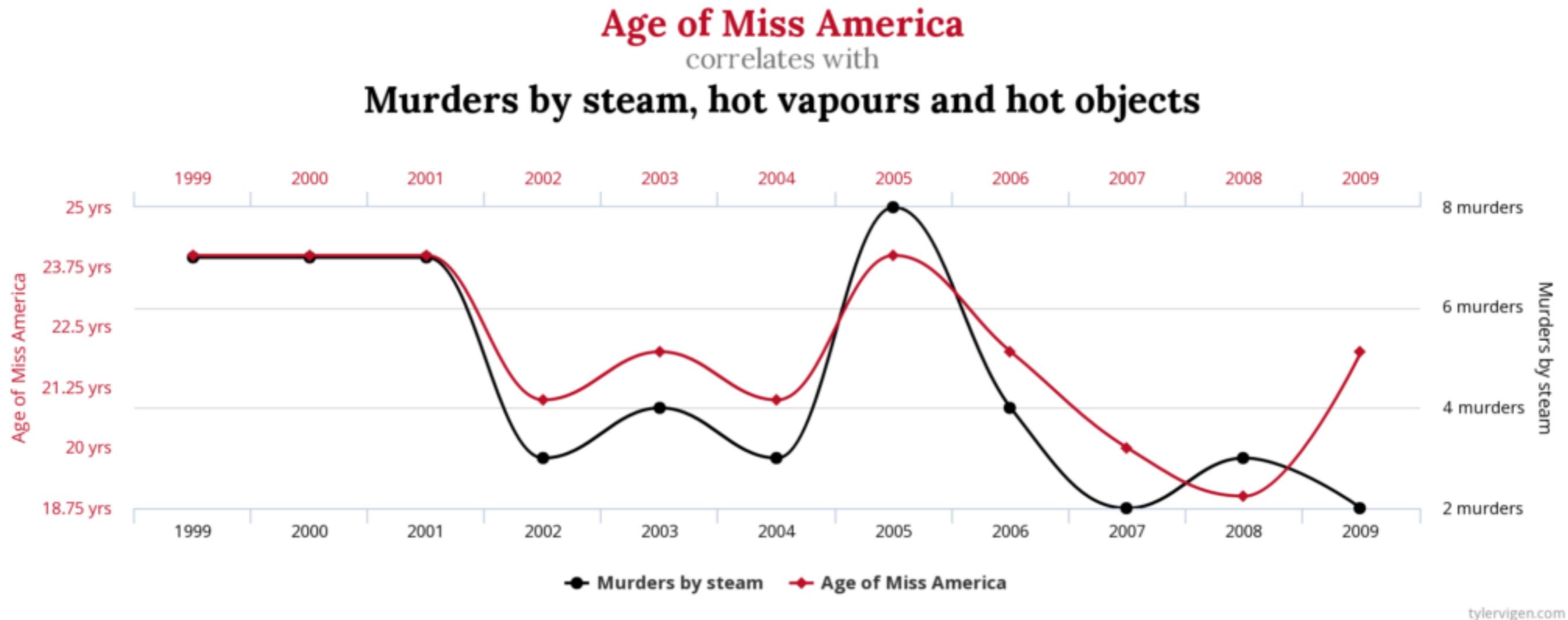
If our **task** is about **finding where there are intra-month changes**, then simply plotting the differences can be more informative.



We can now see that the employment rate under Obama went down more than it went up, and that in November the drop was greatest...

**Another problem is dual axes...**

# Scales

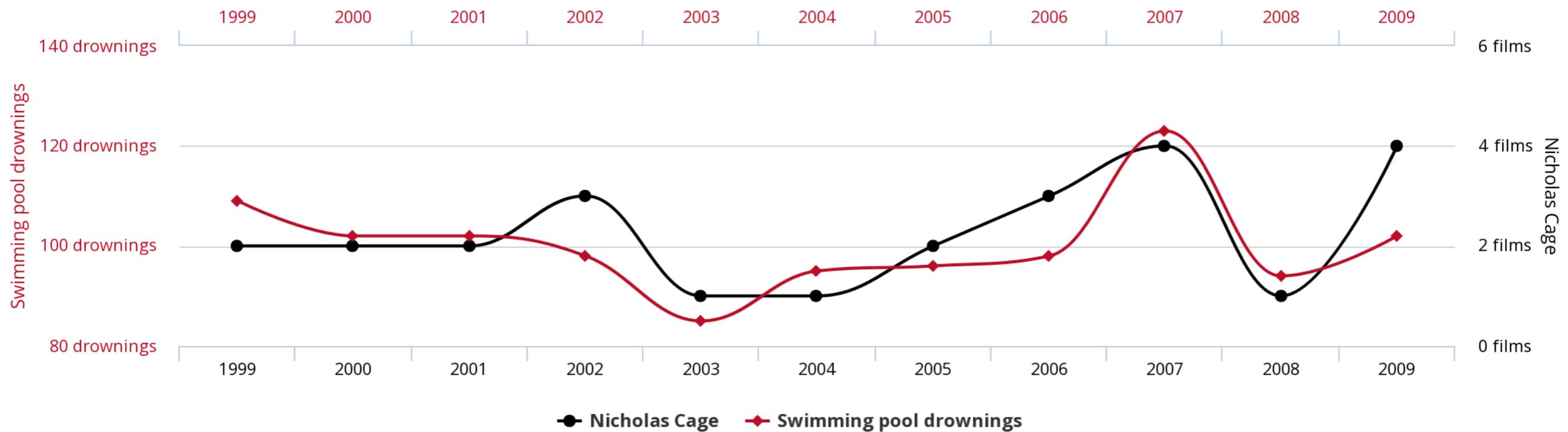


With dual axes, you can tell any story you want by changing one of the scales.

From <http://www.tylervigen.com/spurious-correlations>

# Scales

**Number of people who drowned by falling into a pool**  
correlates with  
**Films Nicolas Cage appeared in**

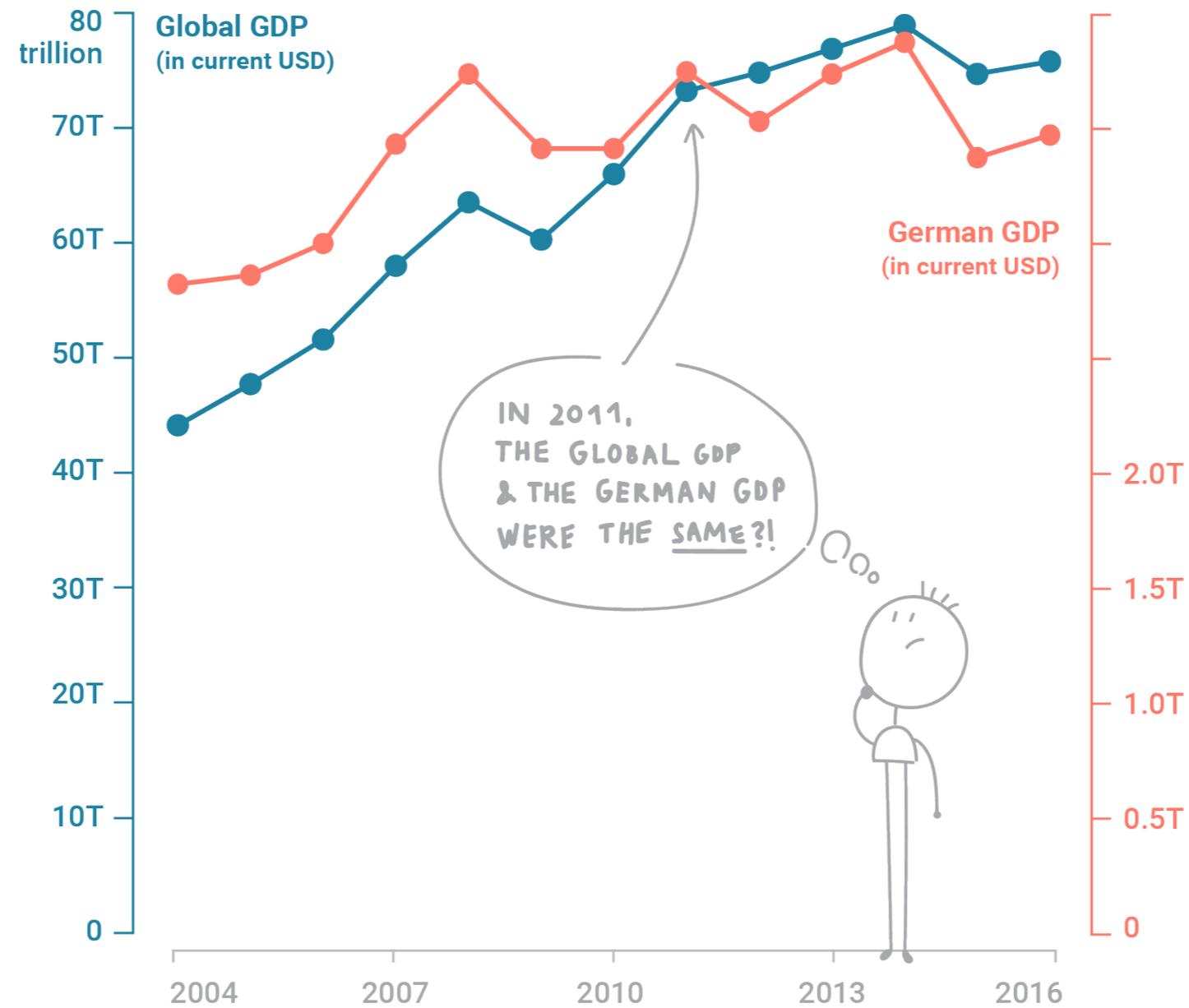


tylervigen.com

With dual axes, you can tell any story you want by changing one of the scales.

From <http://www.tylervigen.com/spurious-correlations>

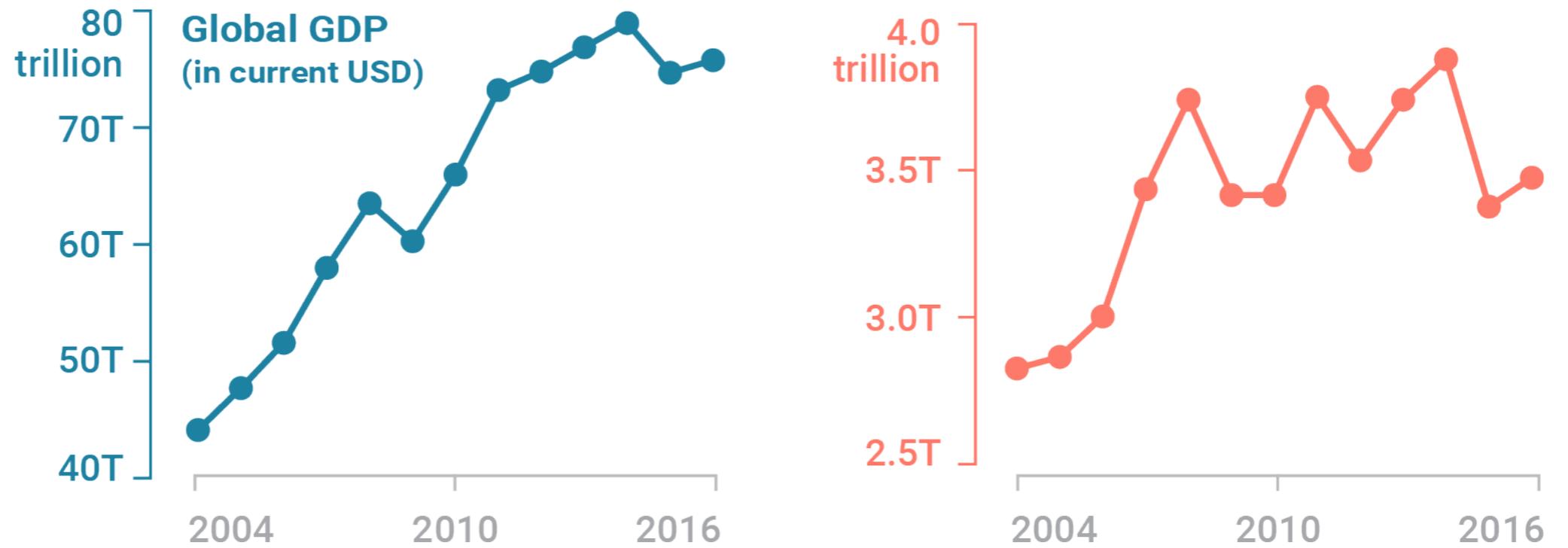
# Scales



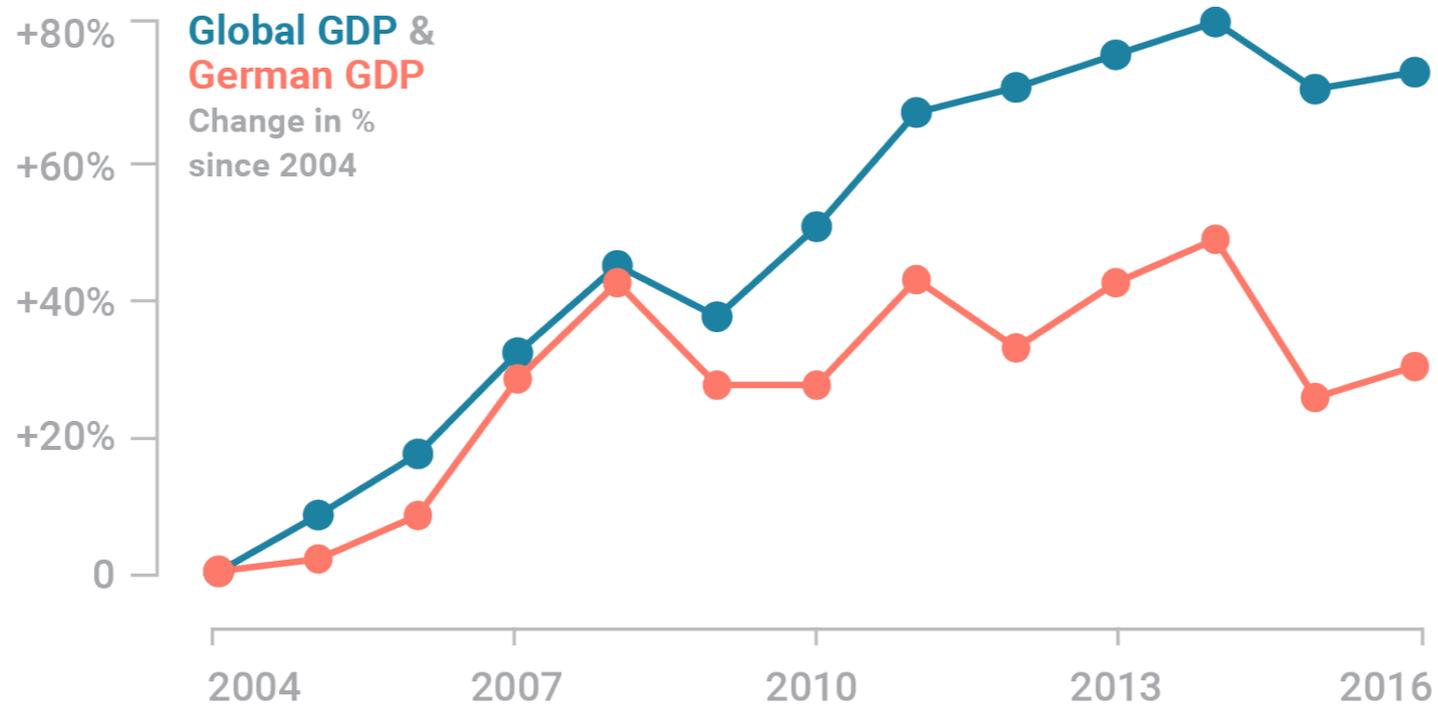
From <https://blog.datawrapper.de/dualaxis/>

# Scales

## Solution 1 Juxtapose



## Solution 2 Index



**Before stepping in to more complex multi-dimensional visualisations, let's look at an example...**

**Video Game Data Set From Kaggle**

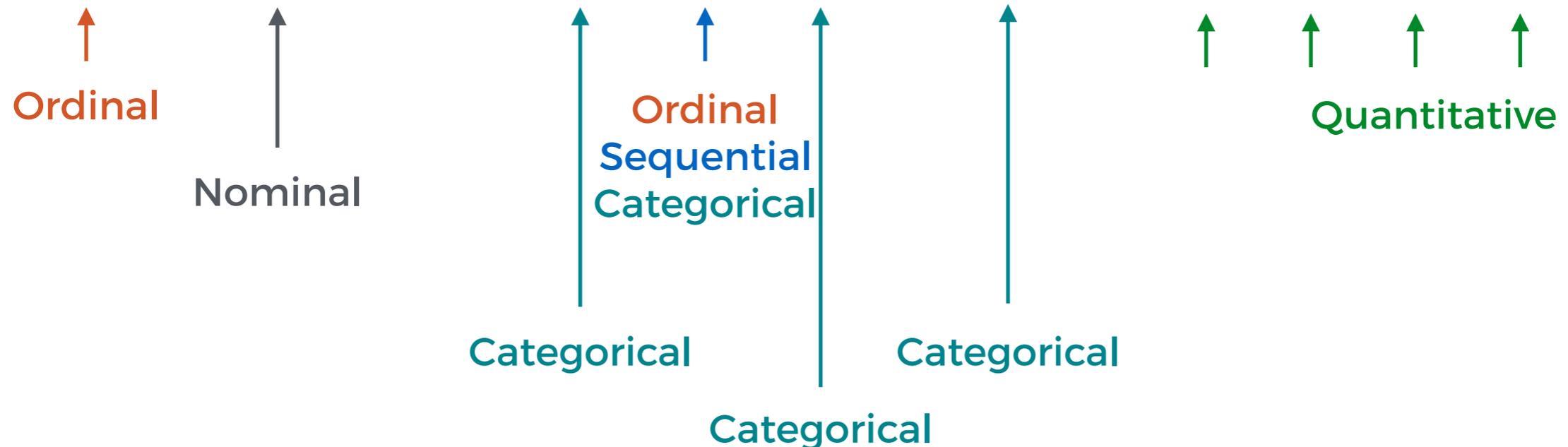
<https://www.kaggle.com/gregorut/videogamesales/version/2#>

# What are you visualising?

e.g. 16,000 rows of video game sales data (from Kaggle)

STATIC DATA | 2D Table | 11 features

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
5	6	Tetris	GB	1989.0	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26
6	7	New Super Mario Bros.	DS	2006.0	Platform	Nintendo	11.38	9.23	6.50	2.90	30.01
7	8	Wii Play	Wii	2006.0	Misc	Nintendo	14.03	9.20	2.93	2.85	29.02
8	9	New Super Mario Bros. Wii	Wii	2009.0	Platform	Nintendo	14.59	7.06	4.70	2.26	28.62
9	10	Duck Hunt	NES	1984.0	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
10	11	Nintendogs	DS	2005.0	Simulation	Nintendo	9.07	11.00	1.93	2.75	24.76
11	12	Mario Kart DS	DS	2005.0	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
12	13	Pokemon Gold/Pokemon Silver	GB	1999.0	Role-Playing	Nintendo	9.00	6.18	7.20	0.71	23.10



# Why are we visualising?

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82

...

## Task

I want to compare the general trends in Global Sales per Genre over time

We can break this task down in to

## Present



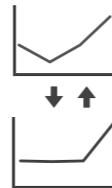
## Actions

→ Query

→ Identify



→ Compare



→ Summarise



## Targets

→ All Data

→ Trends



→ Outliers

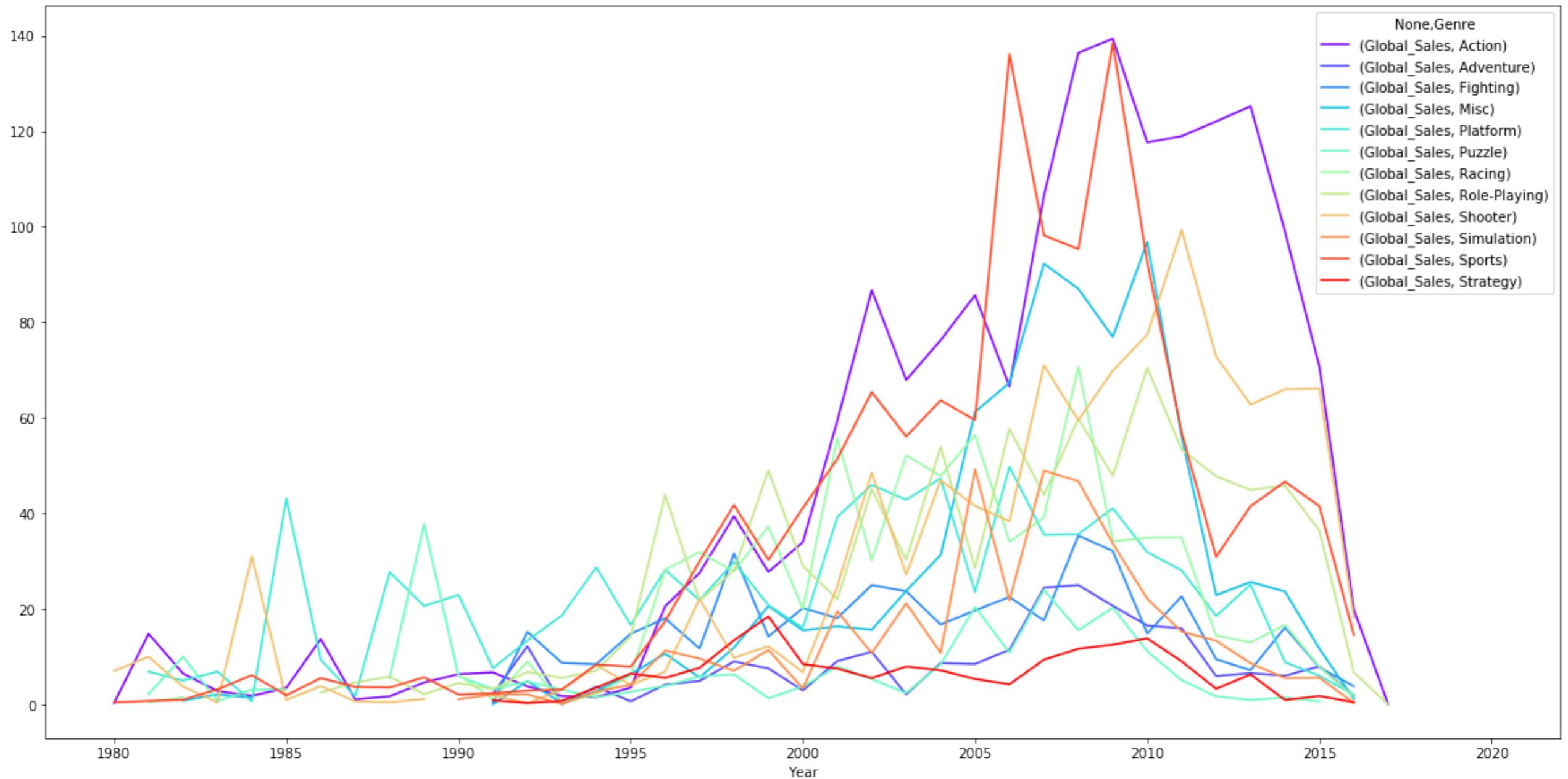


→ Features



We're presenting data, to enable **comparisons of trends.**

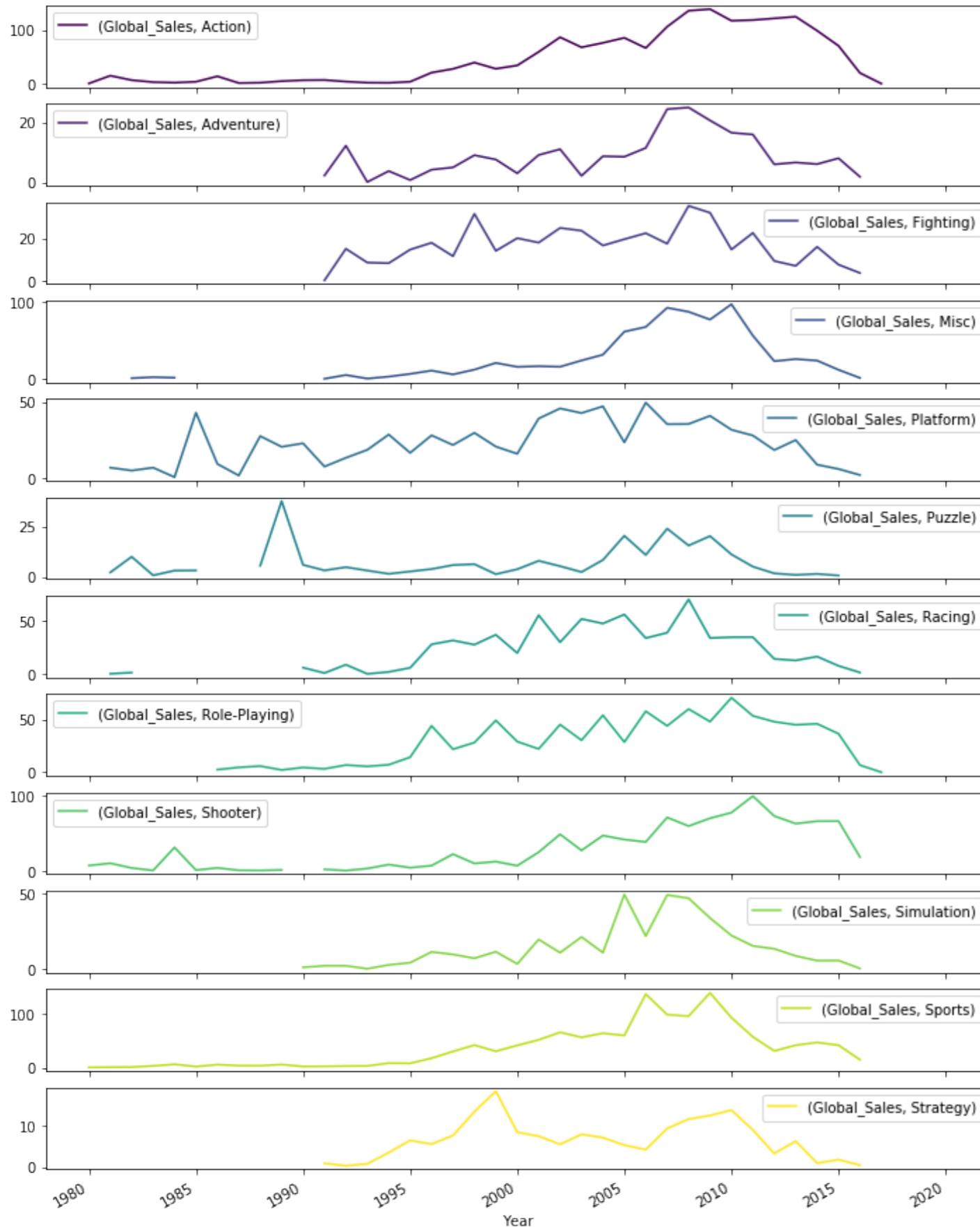
# How can you encode information optimally?



This is super hard to decode! So NOT a good visual encoding.

1. Too many colours (not all distinguishable).
2. Too many crossing lines (making it hard to see continuity)
3. Although less cognitively demanding than reading the whole spreadsheet, it's still pretty demanding to match the line to the series.

# How can you encode information optimally?

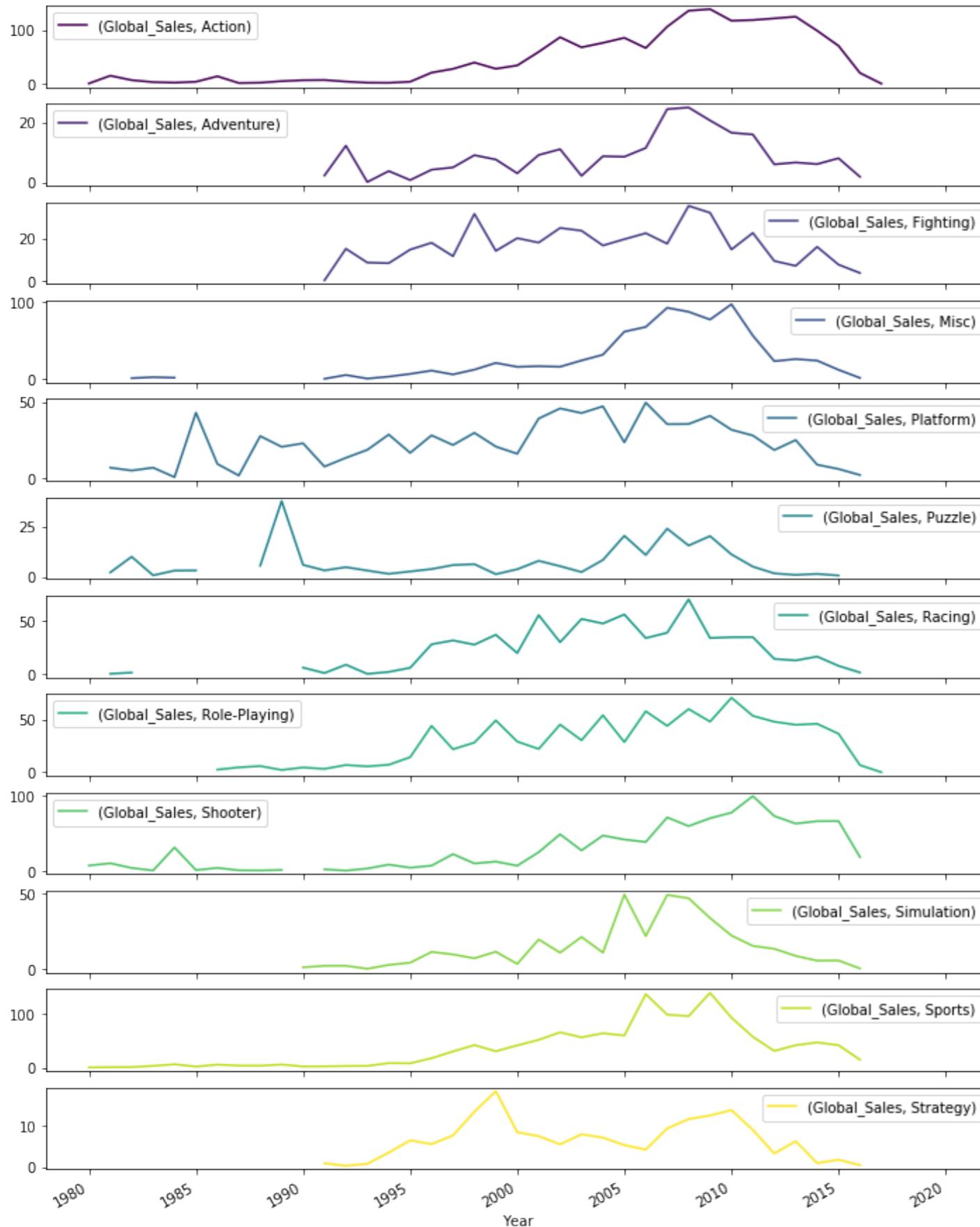


Much better.

Separating the series in to **small multiples** is generally good practice if you have many series to compare.

But can you see problems here?

# How can you encode information optimally?



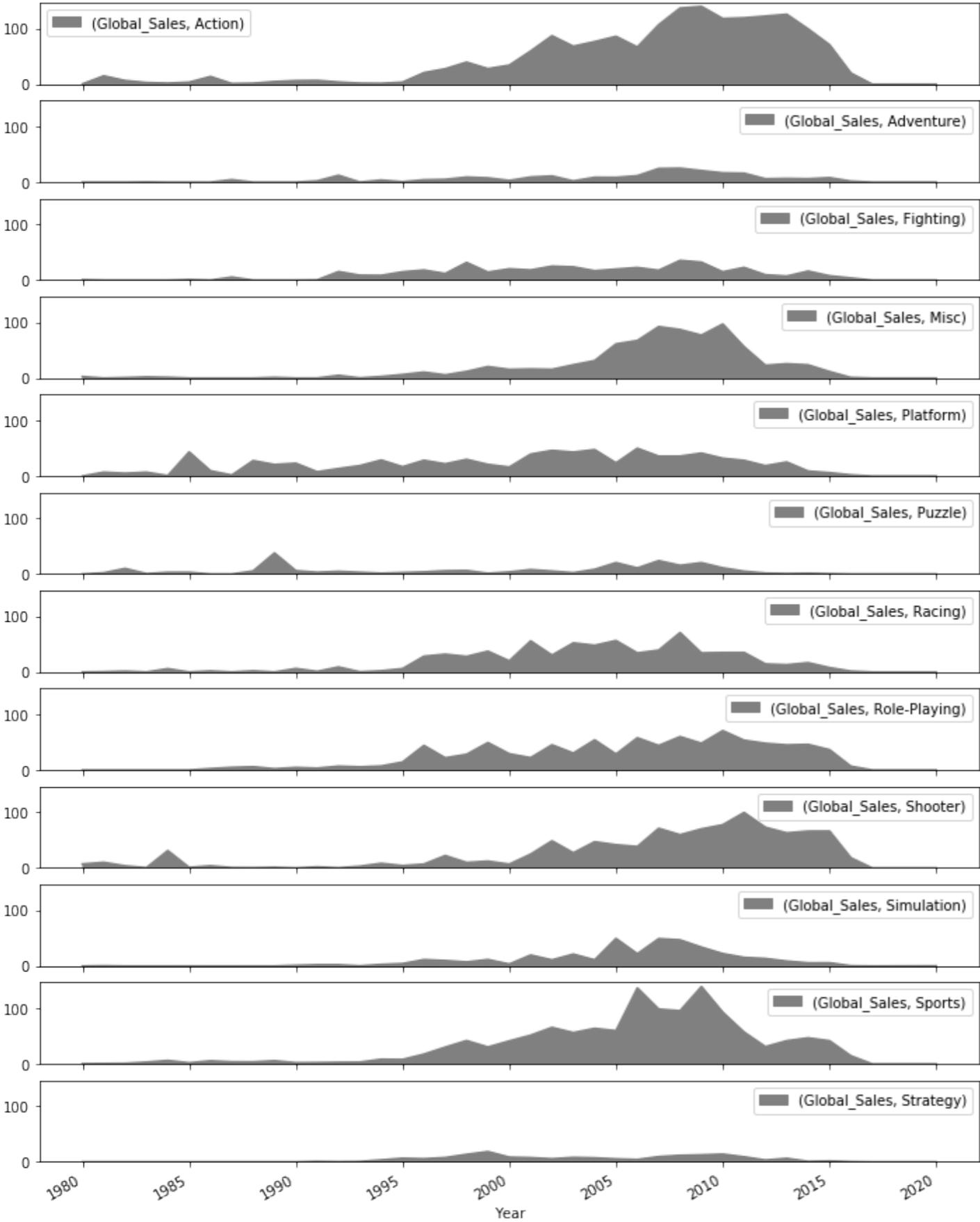
Much better.

Separating the series in to **small multiples** is generally good practice if you have many series to compare.

But can you see problems here?

**Axes are different per plot.**  
**Colour offers us nothing here.**

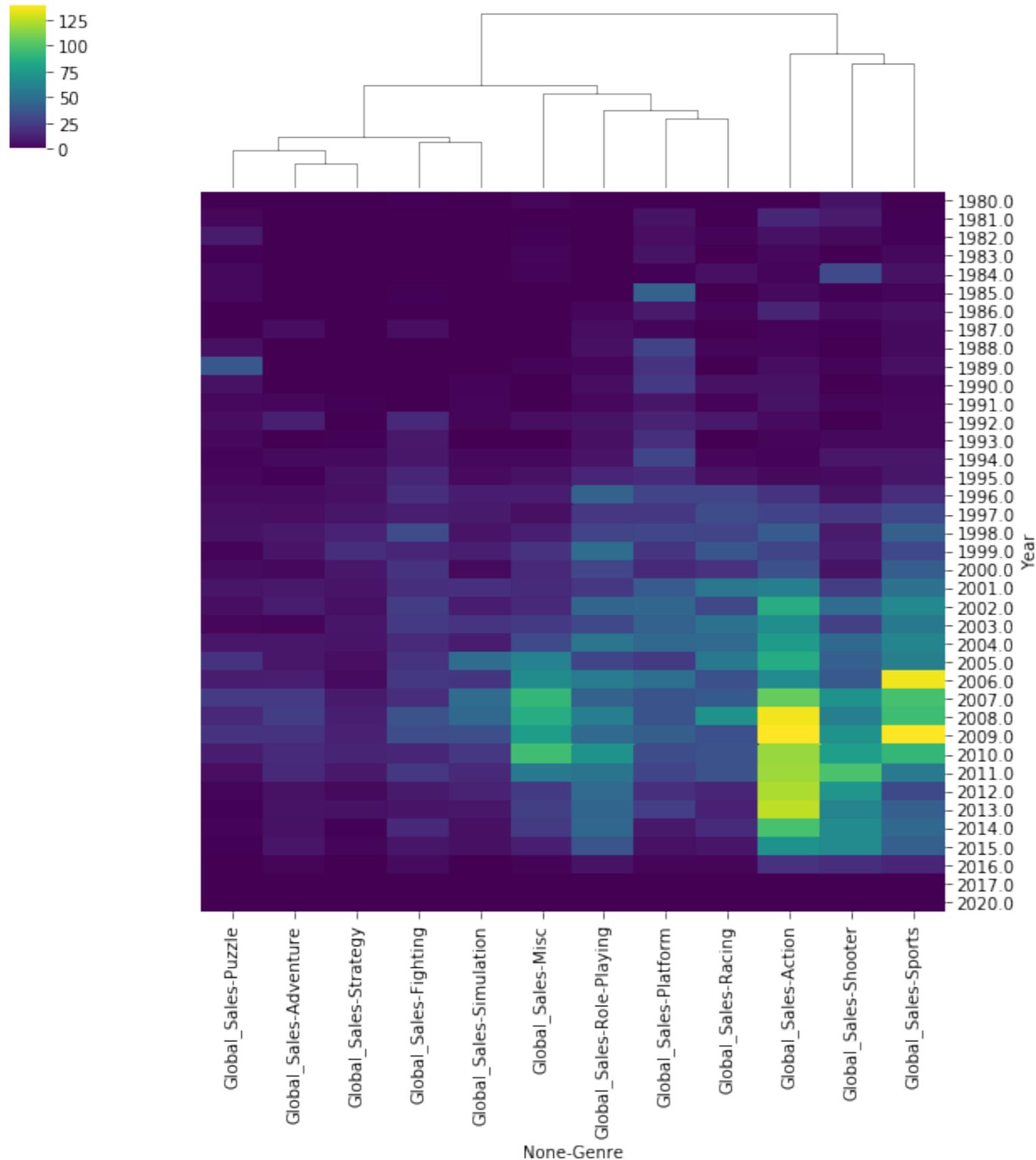
# How can you encode information optimally?



Easy to compare now between all plots.

But can we do better?

# How can you encode information optimally?



Comparing the trends is easier here since we can see all the data in one compact plot.

Here I've also clustered the genres to see which are most similar in terms of trend.

Although, it will be harder to map from the colour to an exact value. Here, we've given up some decoding power, i.e. the ability to go back to the original value.

# Why are we visualising?

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82

...

## Task

I want to compare the number of releases by genre per year

We can break this task down in to

## Present



## Actions

→ Query

→ Identify



→ Compare



→ Summarise



## Targets

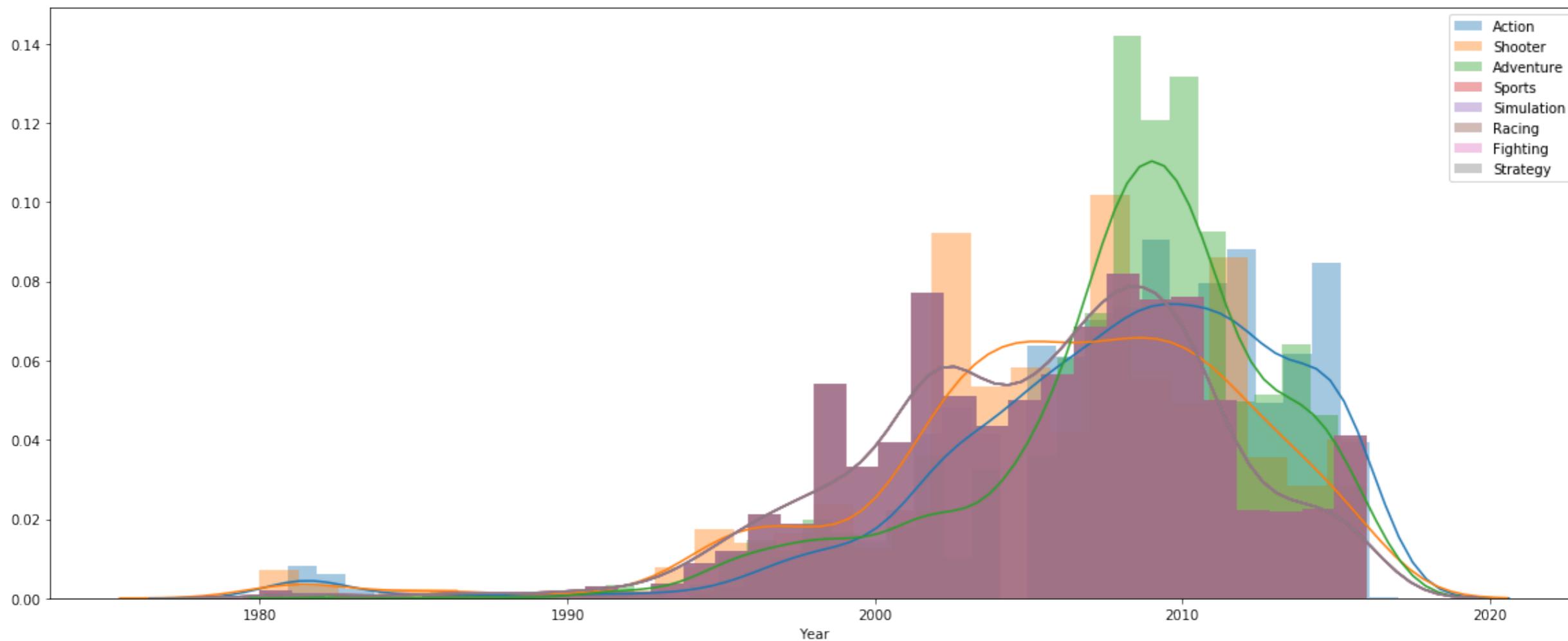
→ Distribution



We're presenting data, to enable comparisons of distributions.

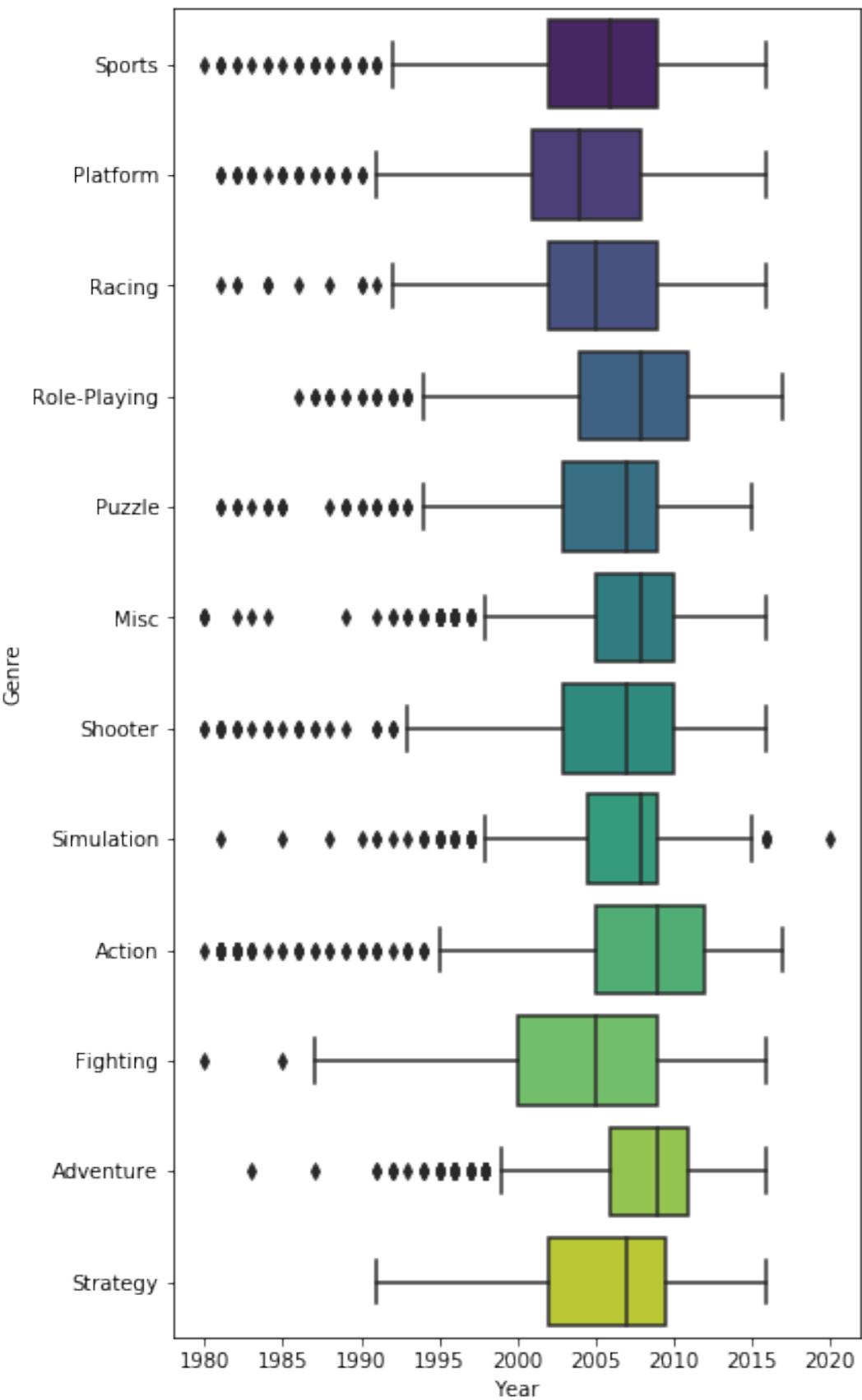


Naively, we would start by plotting the time distribution for each Genre, and overlay them on top of one another.



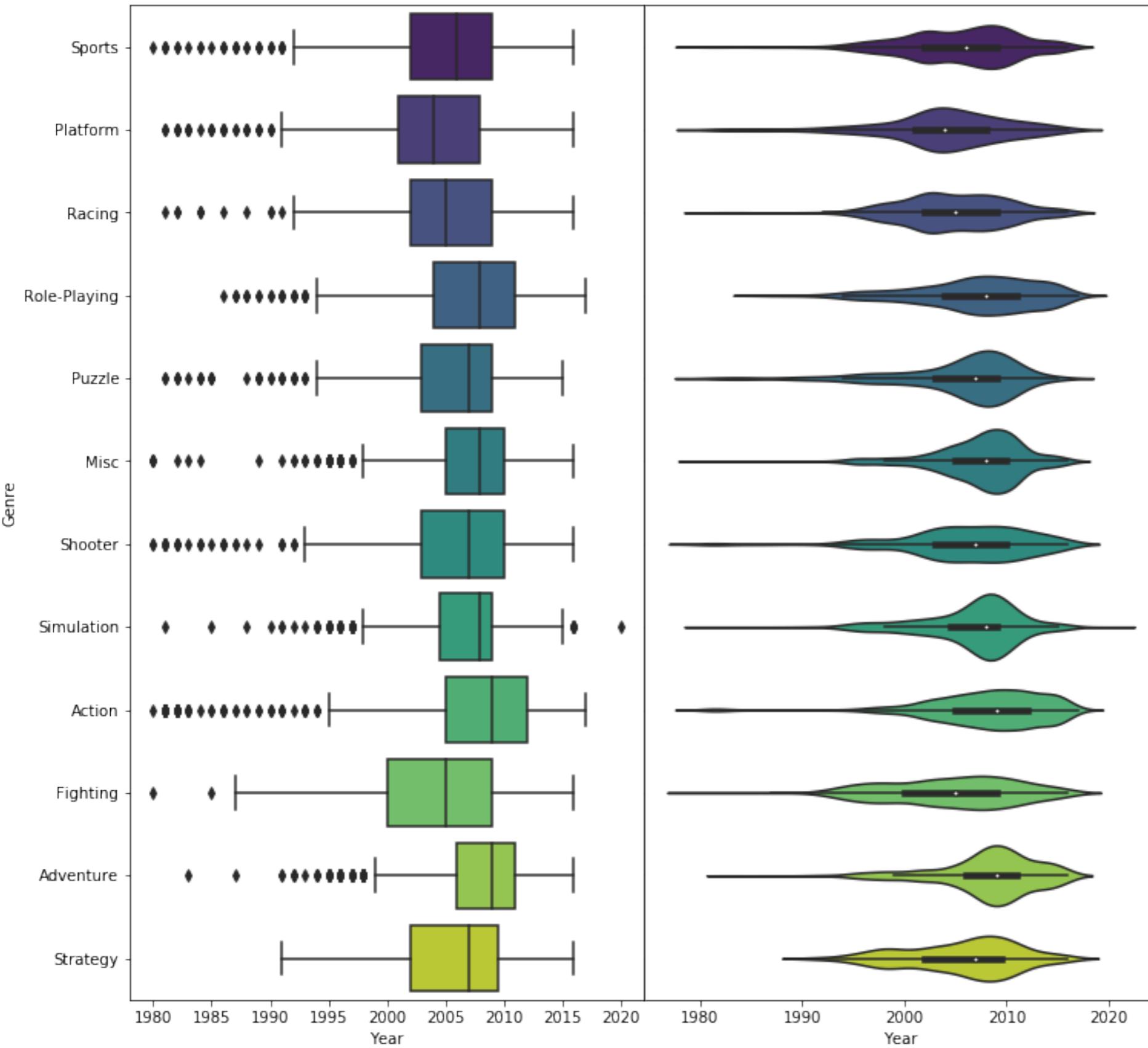
**Too many overlapping areas. It's a mess.**

# Box Plots



# Box Plots

# Violin Plots

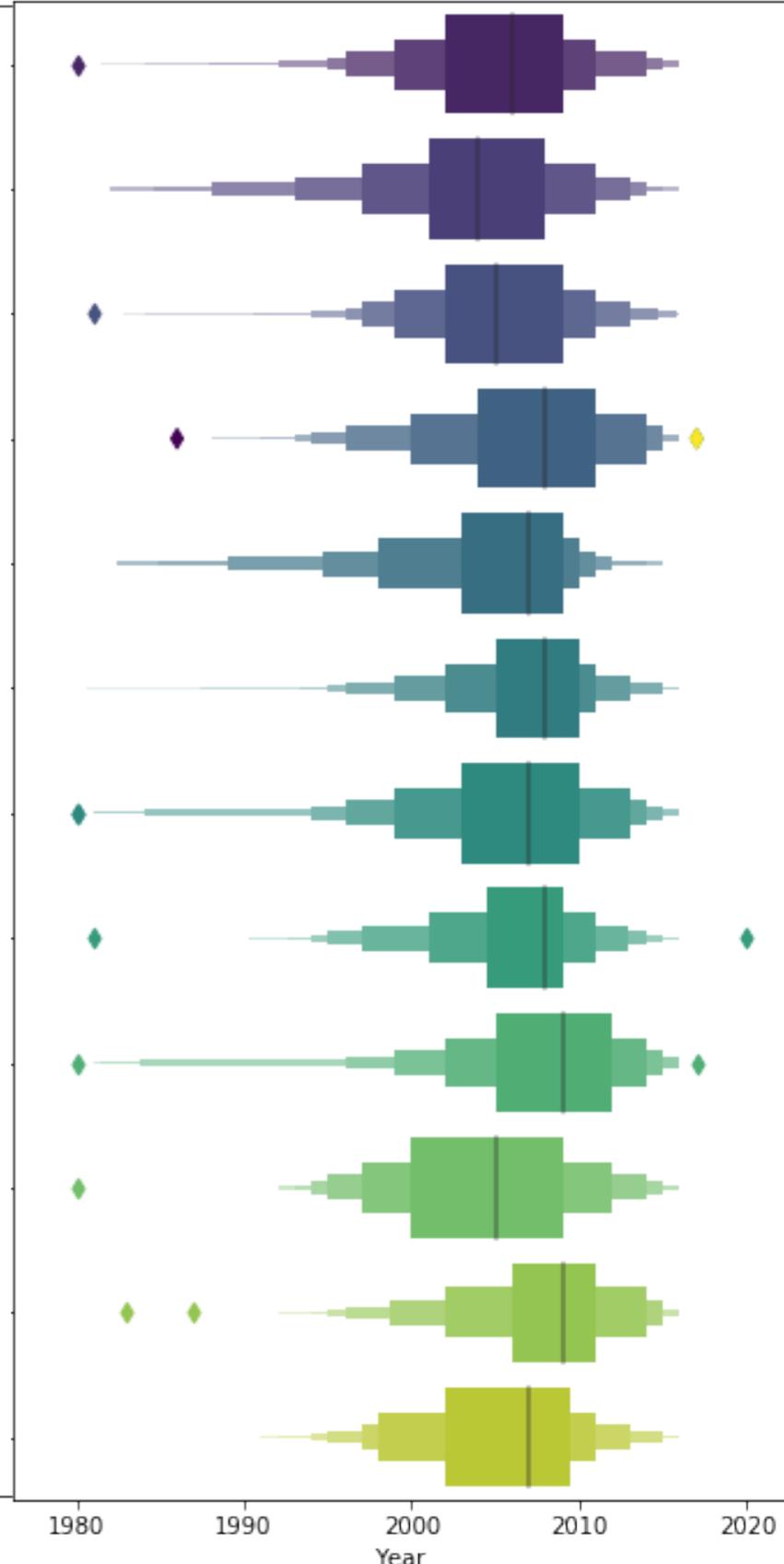
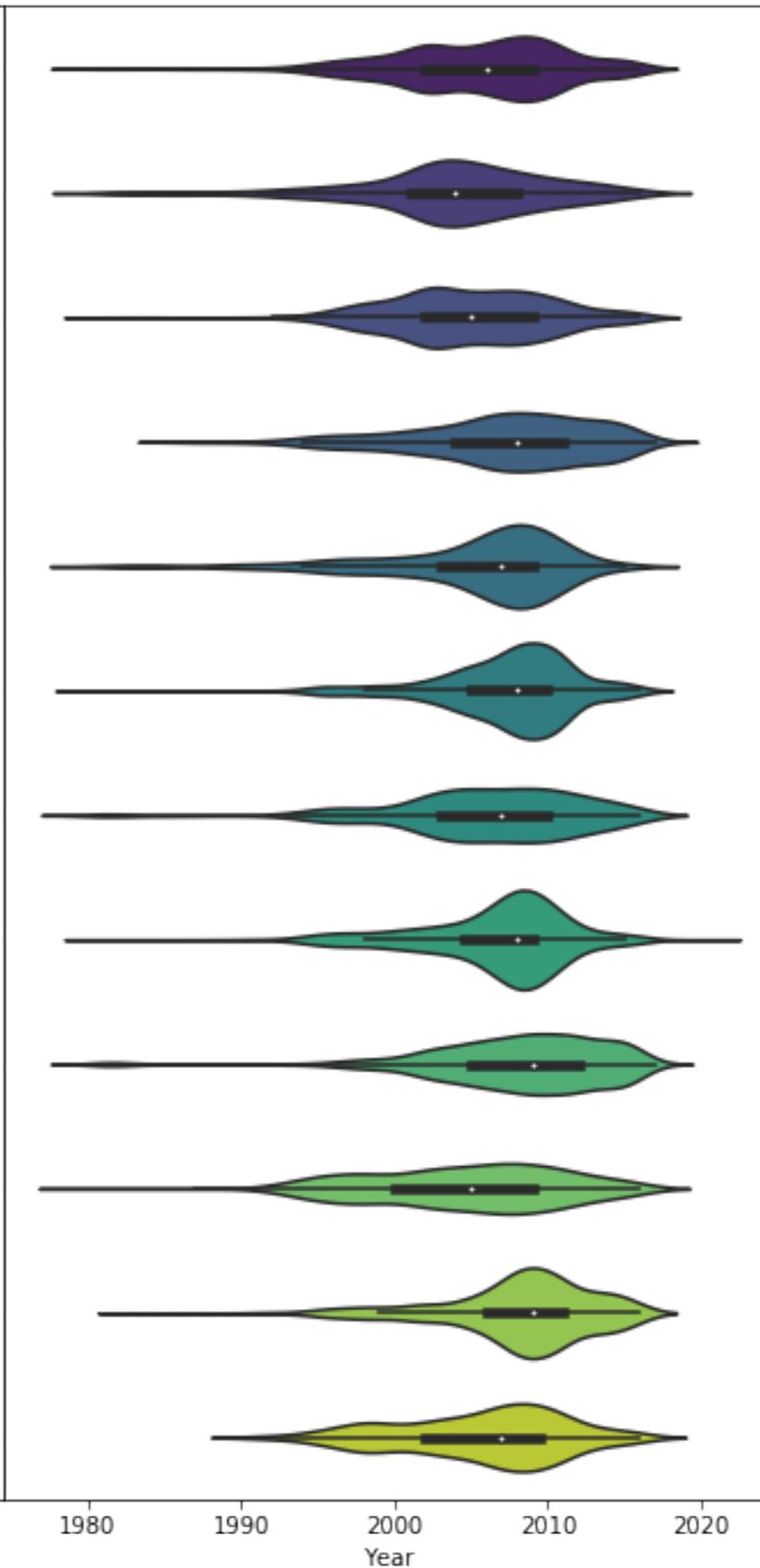
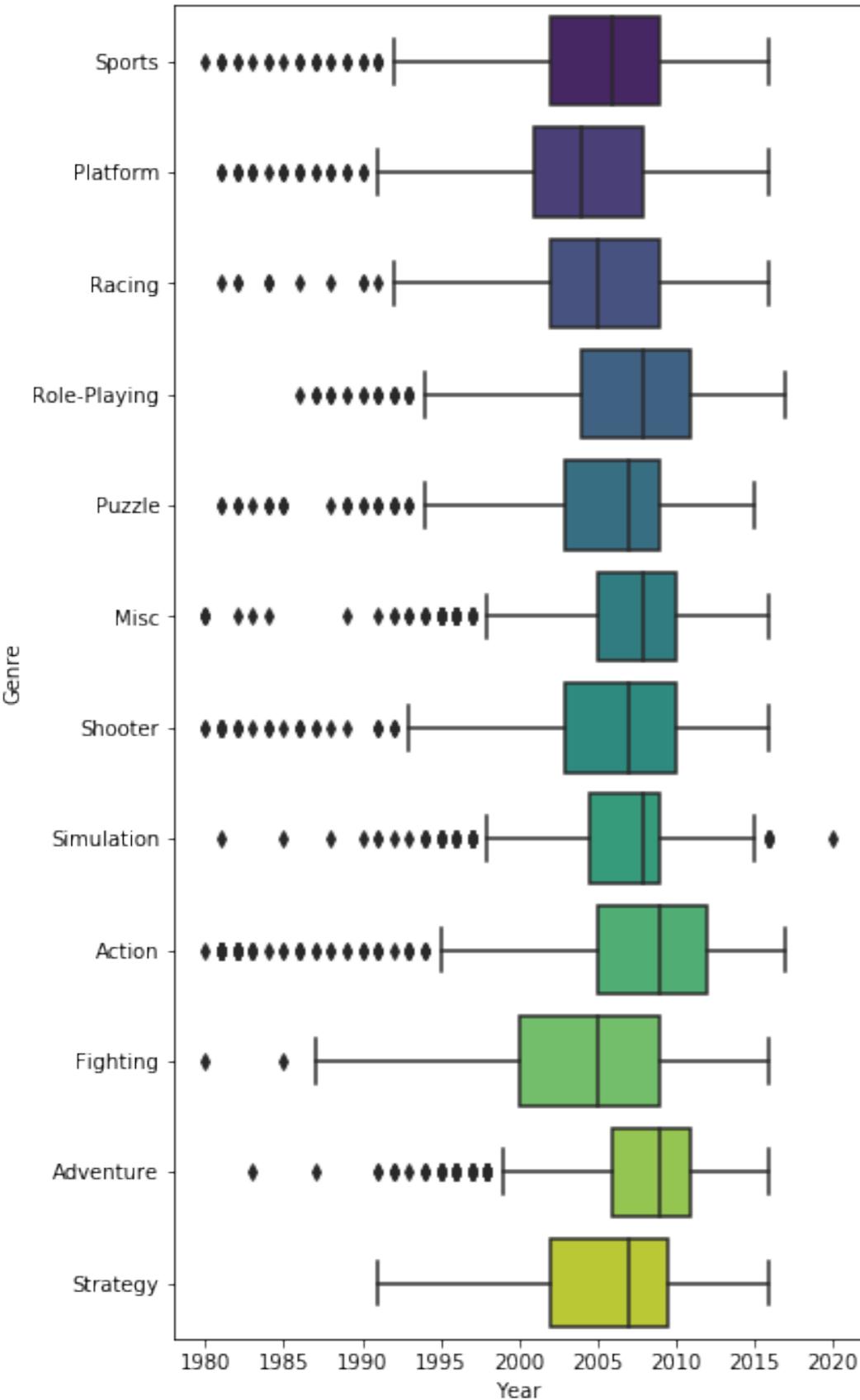


# Box Plots

# Violin Plots

# Boxenplots

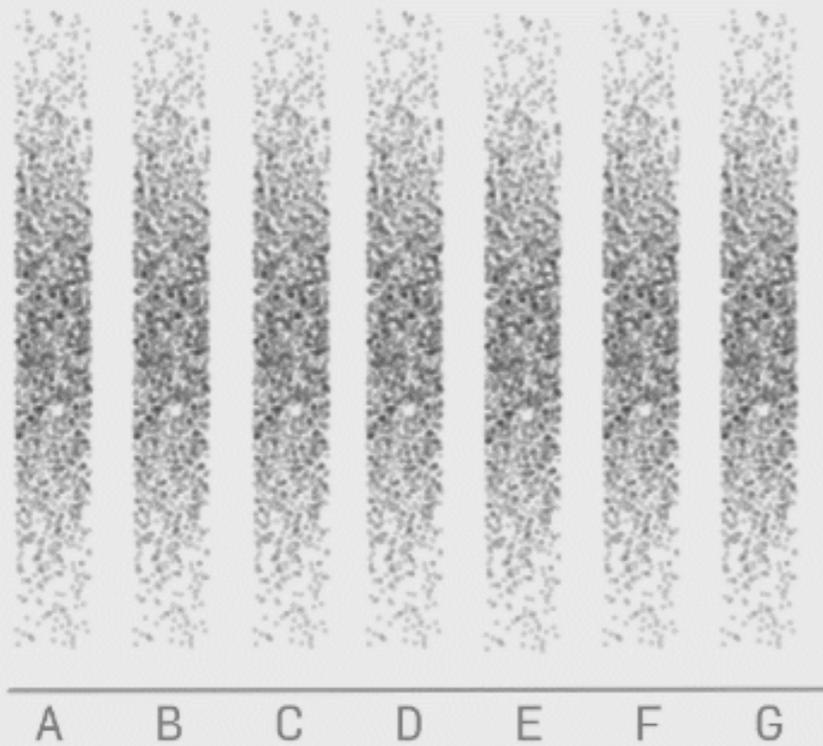
<https://vita.had.co.nz/papers/letter-value-plot.pdf>



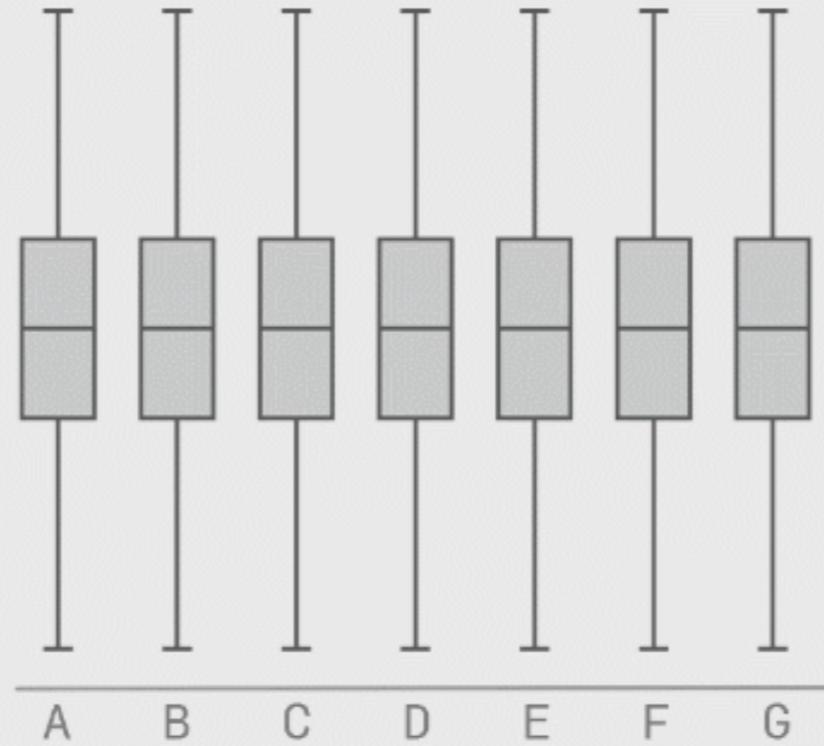
Why violin or boxenplots I hear you ask?

**Box plots can also be deceptive!**

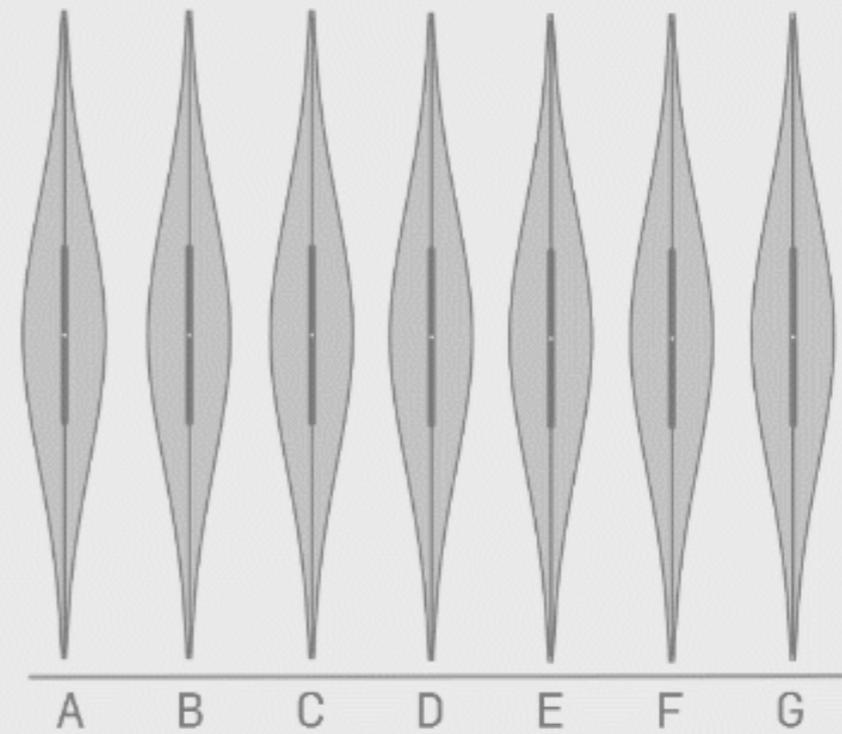
**Raw Data**



**Box-plot of the Data**



**Violin-plot of the Data**



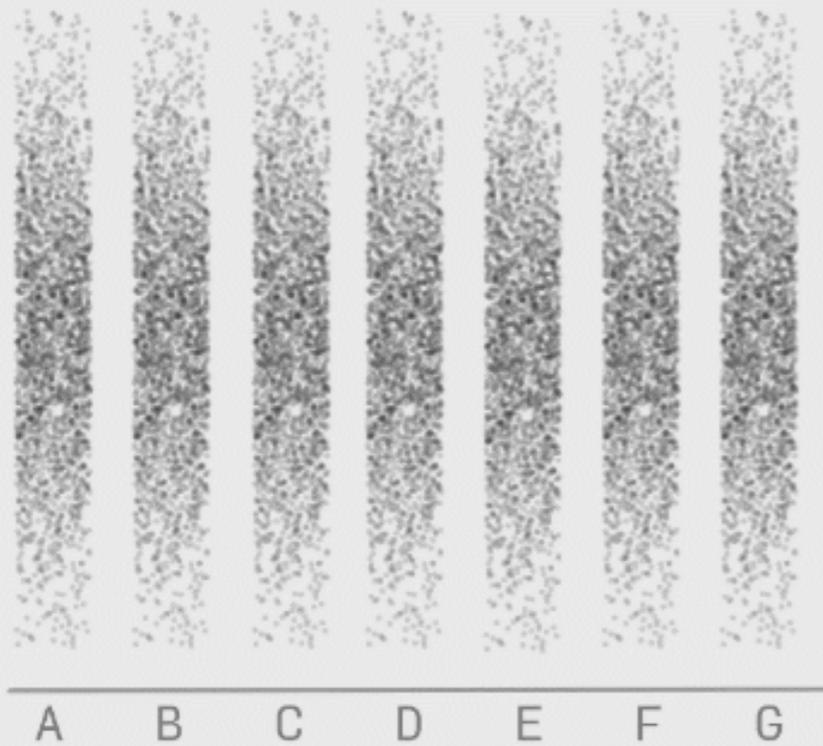
**But violin plots give a better representation of the data.**

<https://www.autodeskresearch.com/publications/samestats>

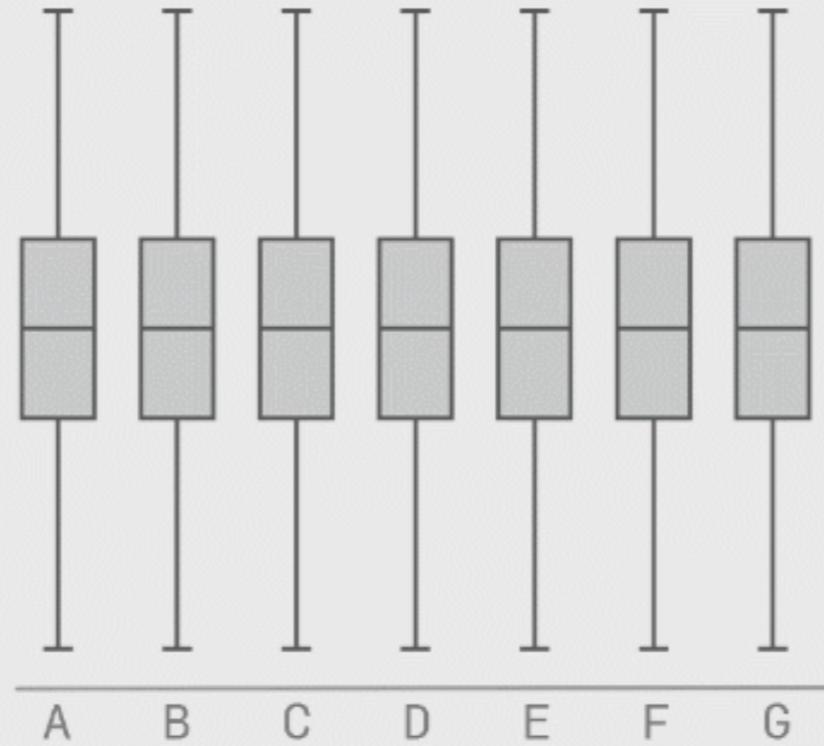
Why violin or boxenplots I hear you ask?

**Box plots can also be deceptive!**

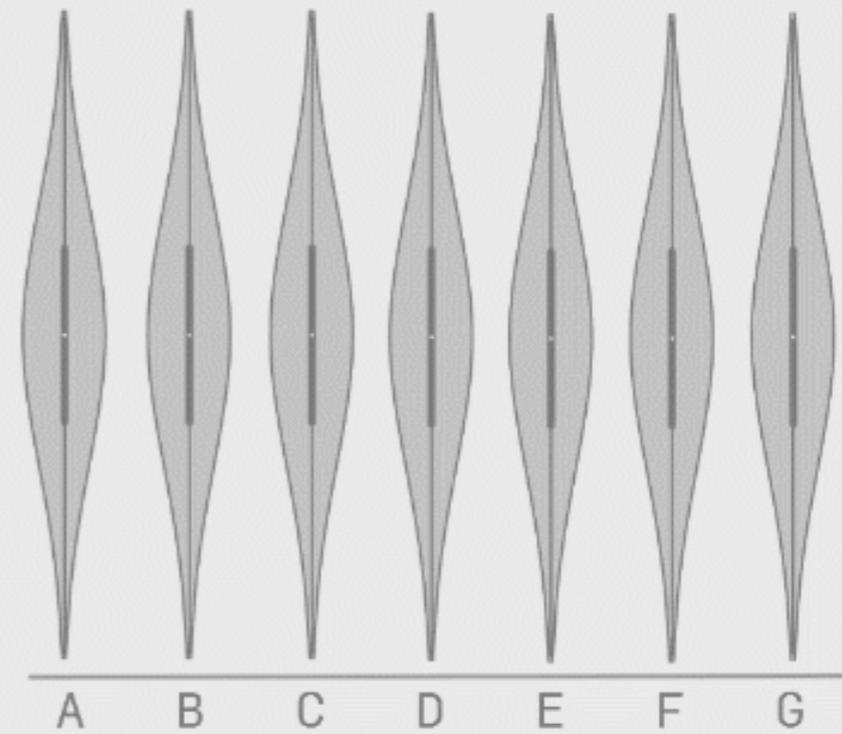
**Raw Data**



**Box-plot of the Data**

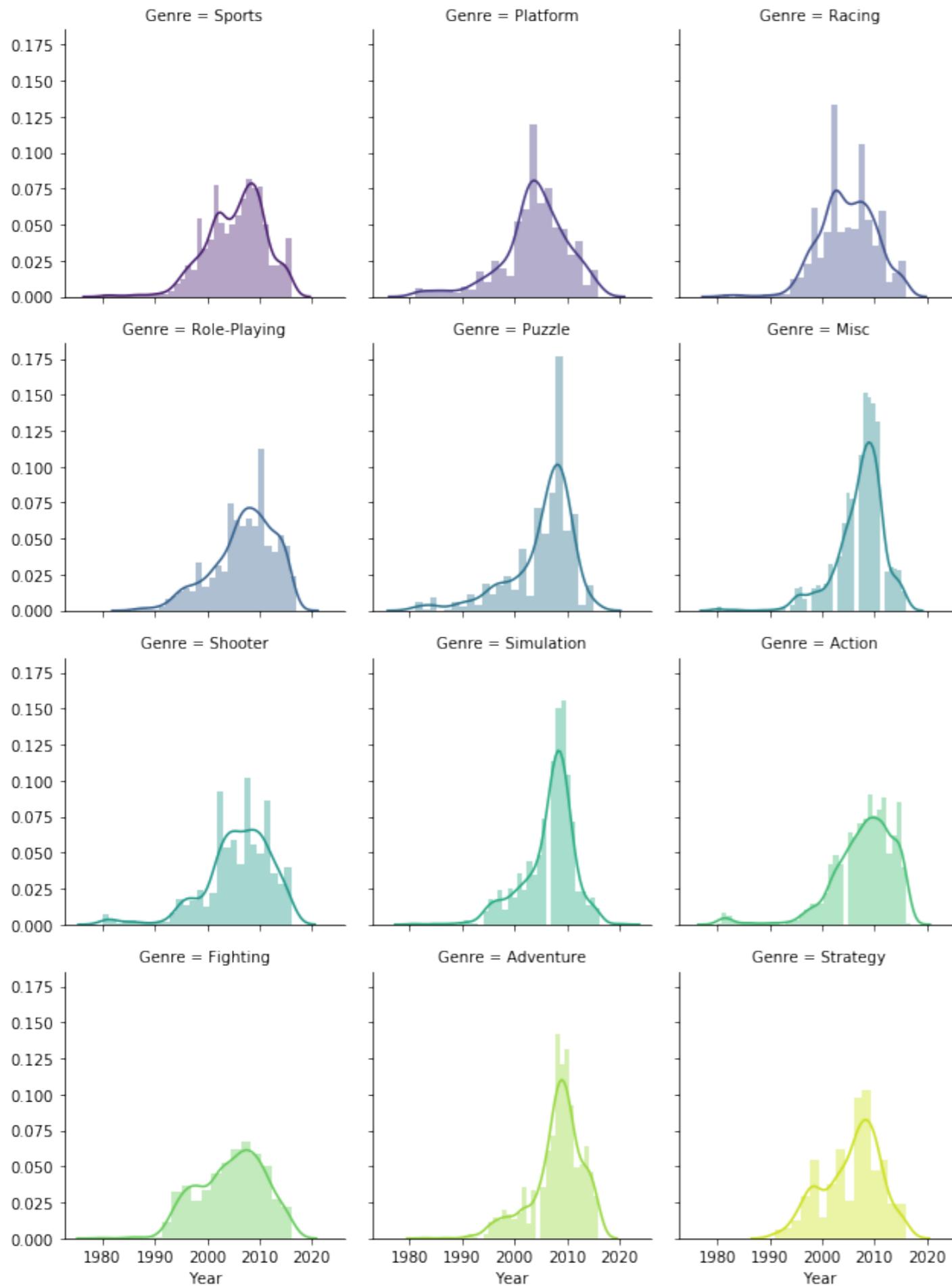


**Violin-plot of the Data**



**But violin plots give a better representation of the data.**

<https://www.autodeskresearch.com/publications/samestats>



## Facet Plots

Small multiples

While aesthetically nice, and this does provide a good detailed view of the data, it's hard to compare all the distributions.

**So far, we've only seen how to represent a low number of dimensions**

What happens when we have a high number of dimensions?

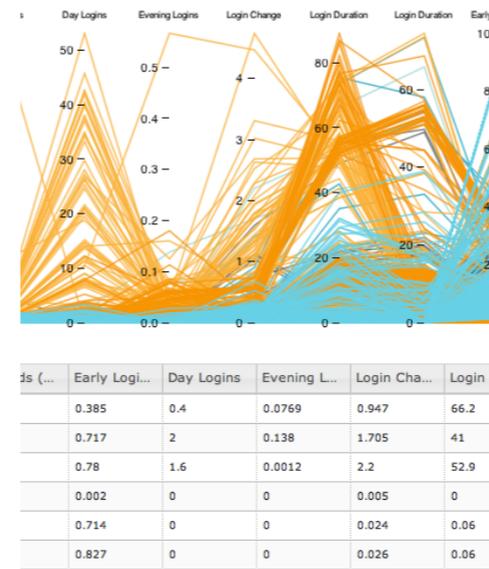
# Multidimensional Visualization



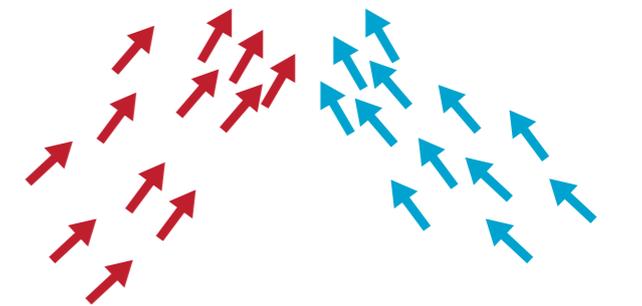
Scatter Plot  
Matrices



Dashboards



Parallel  
Coordinates



Temperature - Colour ■ ■  
 Wind direction - Orientation ↑ ↗ →  
 Wind Speed - Proximity  
 Location - Position

Glyphs

# Multidimensional Visualization

## Scatter Plot Matrices

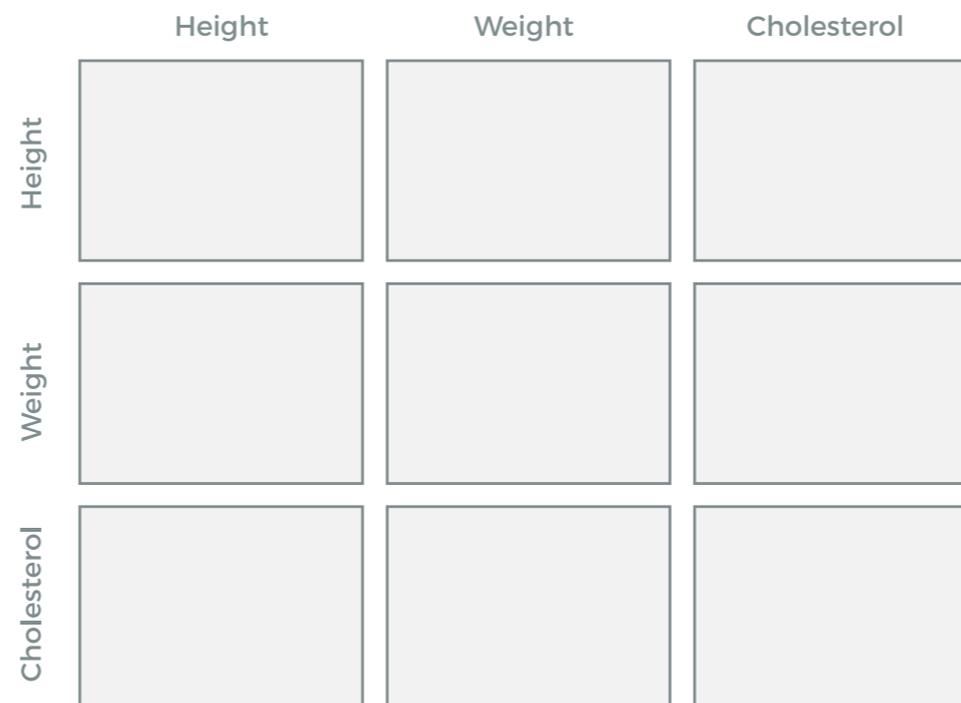
<b>Name</b>	<b>Height</b>	<b>Weight</b>	<b>Chol</b>
John	1.76	63	4.5
Mike	1.79	70	4.15
Jim	1.61	60	6.7
Francois	1.84	90	5.03

...

# Multidimensional Visualization

## Scatter Plot Matrices

Name	Height	Weight	Chol
John	1.76	63	4.5
Mike	1.79	70	4.15
Jim	1.61	60	6.7
Francois	1.84	90	5.03
	...		



# Multidimensional Visualization

## Scatter Plot Matrices

Name	Height	Weight	Chol
John	1.76	63	4.5
Mike	1.79	70	4.15
Jim	1.61	60	6.7
Francois	1.84	90	5.03
	...		



# Multidimensional Visualization

## Scatter Plot Matrices

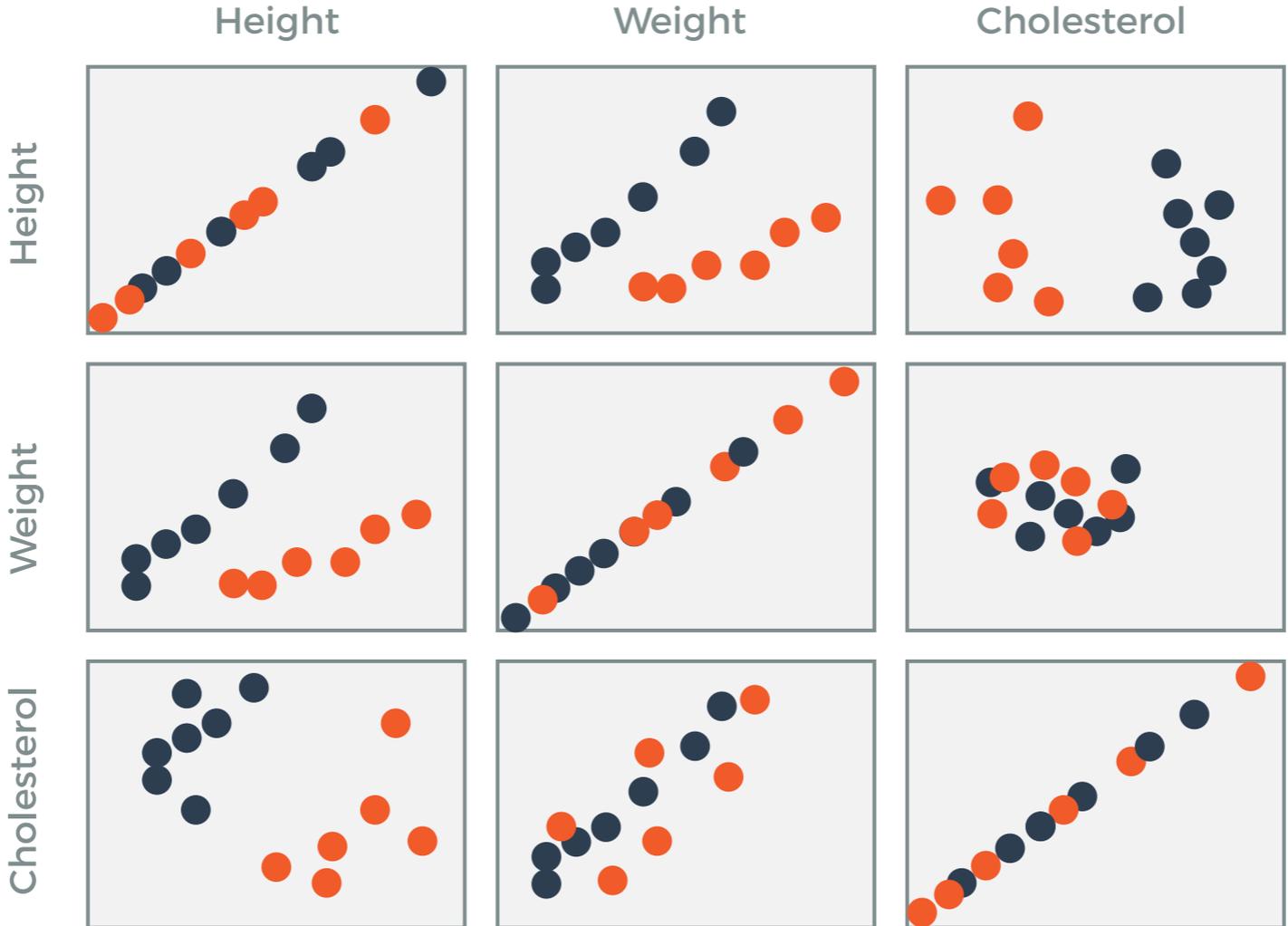
Name	Height	Weight	Chol
John	1.76	63	4.5
Mike	1.79	70	4.15
Jim	1.61	60	6.7
Francois	1.84	90	5.03
	...		

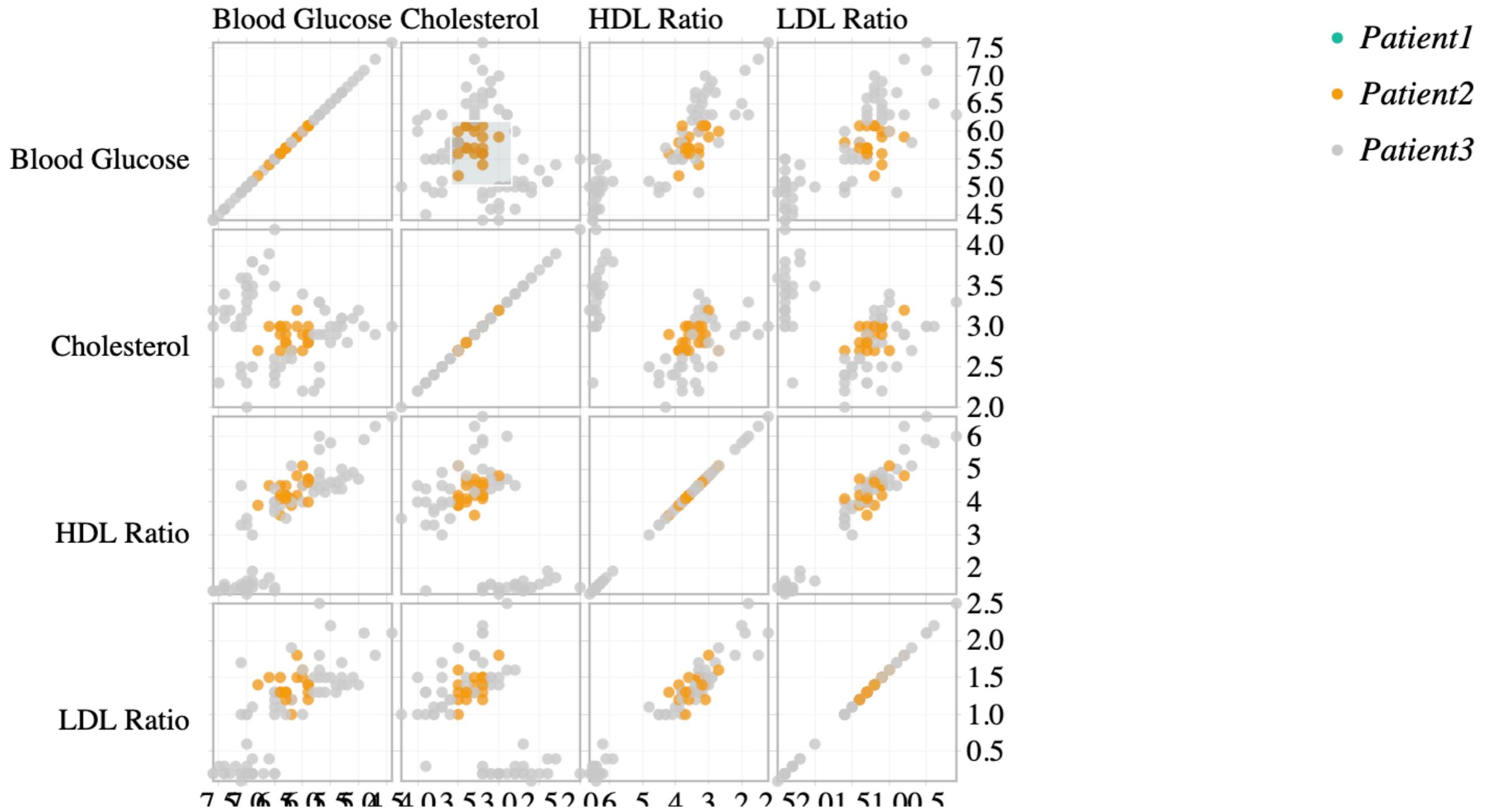


# Multidimensional Visualization

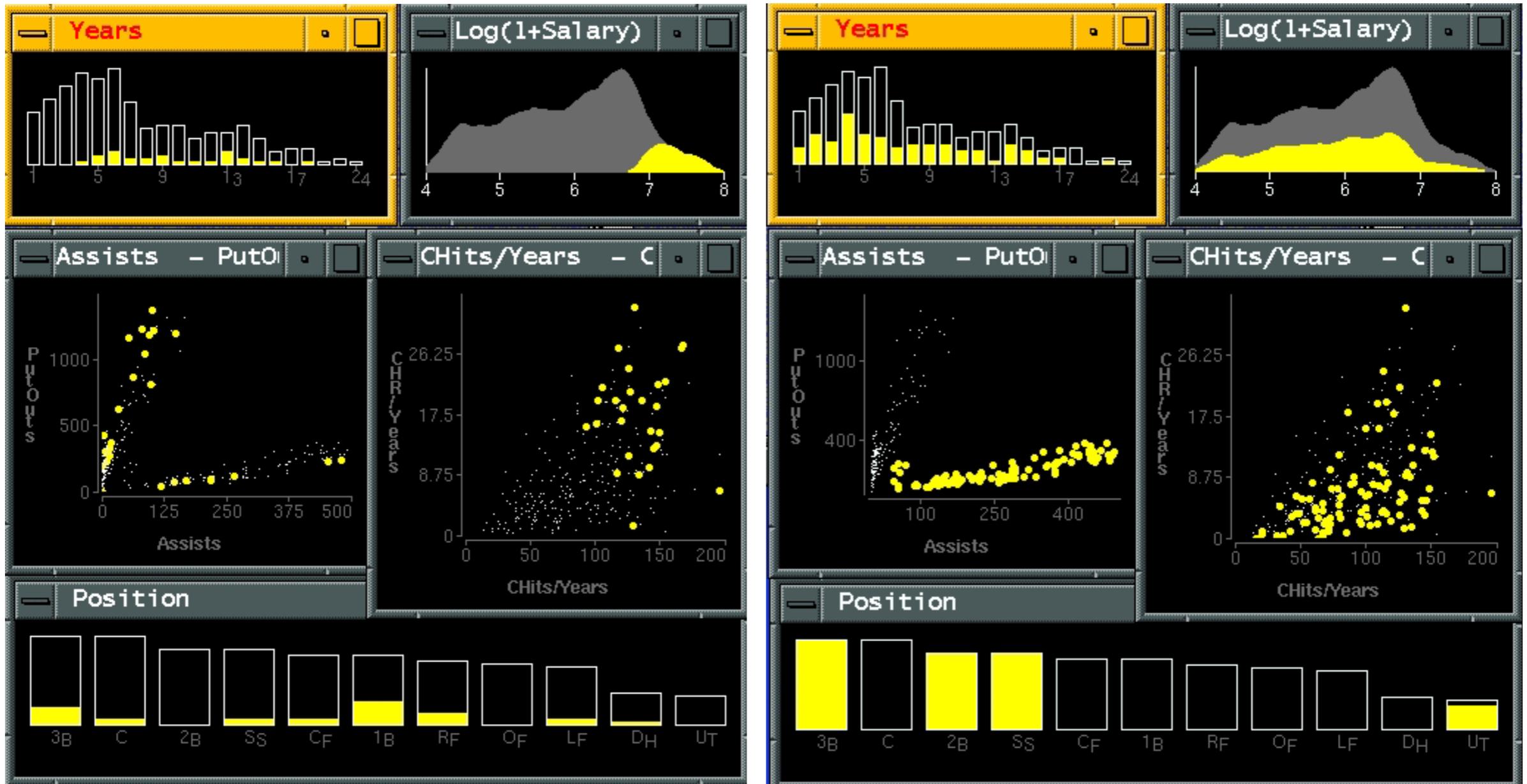
## Scatter Plot Matrices

Name	Height	Weight	Chol
John	1.76	63	4.5
Mike	1.79	70	4.15
Jim	1.61	60	6.7





# Multidimensional Visualization Dashboards



Visual Exploration of Large Structured Datasets. Wills. Proc. New Techniques and Trends in Statistics (NTTS), pp. 237-246. IOS Press, 1995.

# Multidimensional Visualization Dashboards

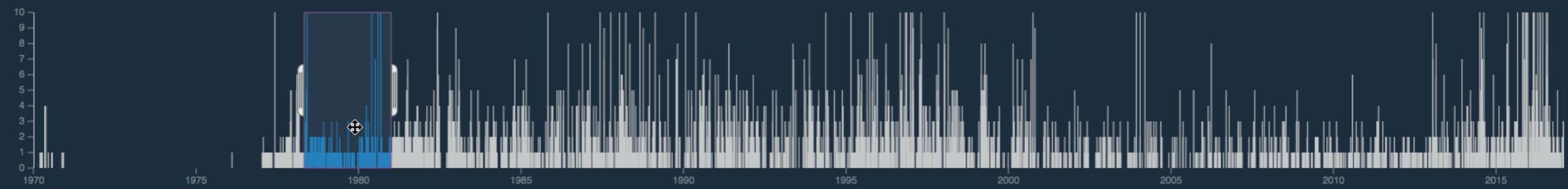


 **Eamonn Maguire**  
✉ eamonnmag@gmail.com

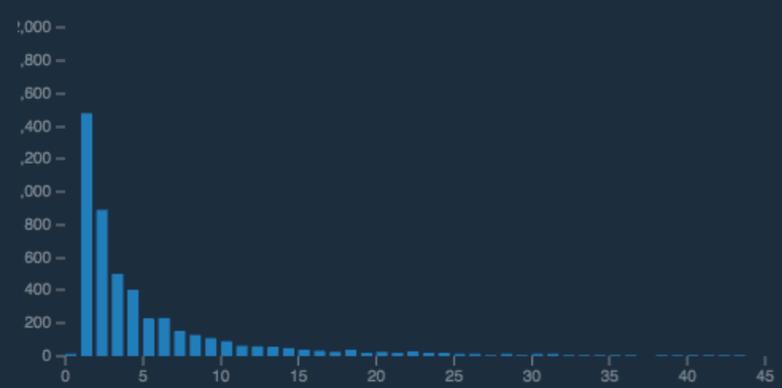
- Administration
- Submissions Overview
- Edit Profile

[Back to Dashboard](#)

Submissions By Last Update Time

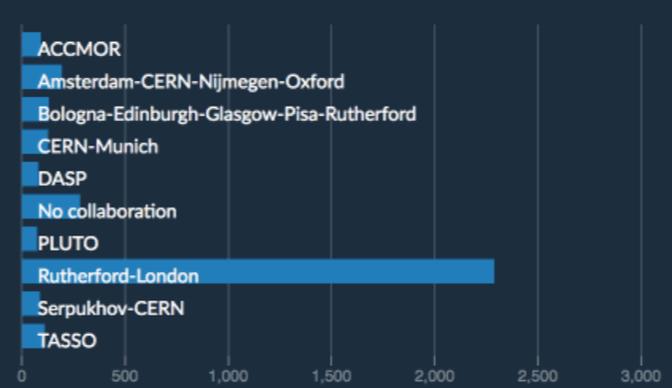


Data Records Per Submission



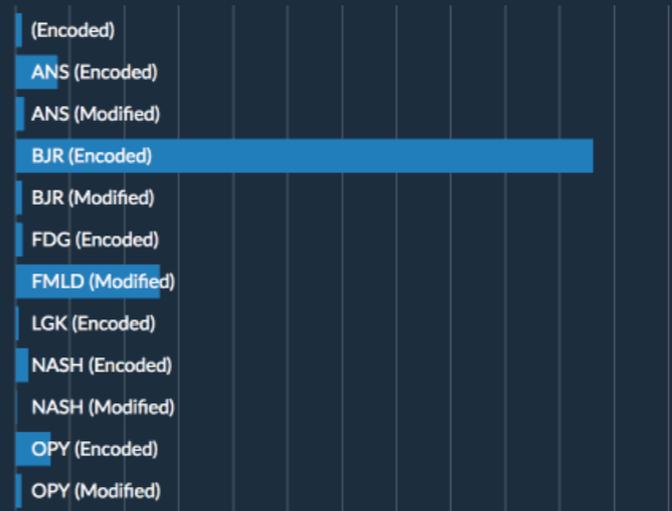
Submission Status

Collaboration TOP 10



Version

Submission Participants



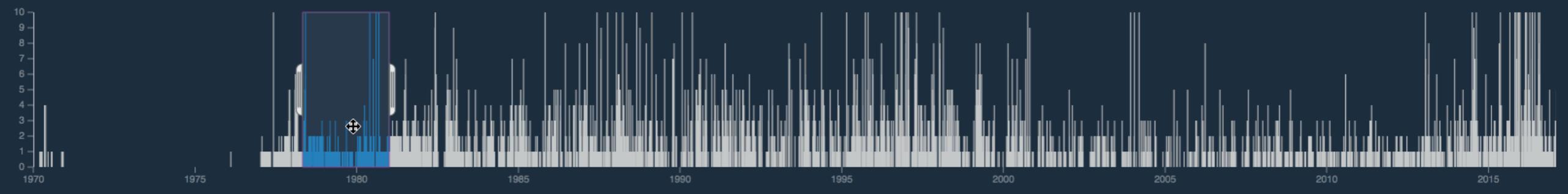
# Multidimensional Visualization Dashboards

**Eamonn Maguire**  
eamonnmag@gmail.com

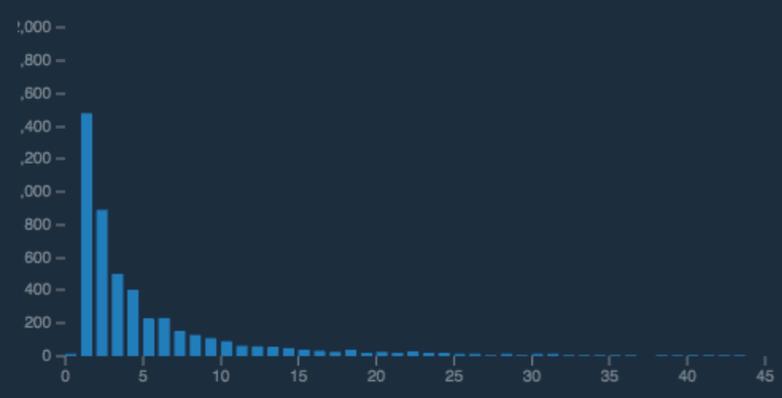
- Administration
- Submissions Overview
- Edit Profile

[Back to Dashboard](#)

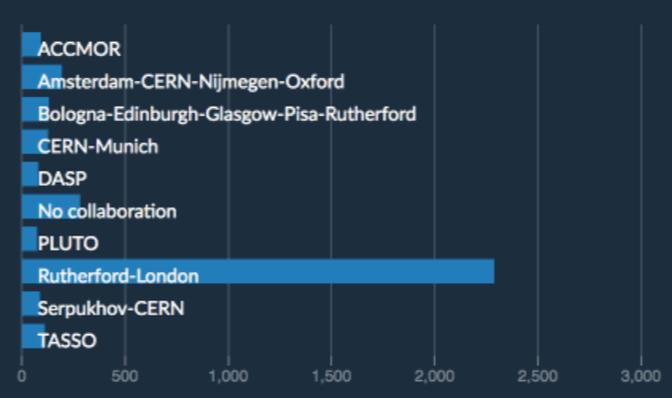
### Submissions By Last Update Time



### Data Records Per Submission



### Collaboration TOP 10



### Submission Participants



Submission Status

Version

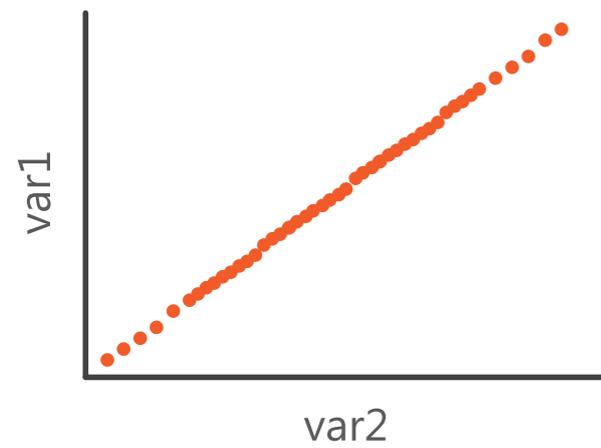
## My Tutorial on Creating Dashboard Visualizations

<https://thor-project.github.io/dashboard-tutorial/>

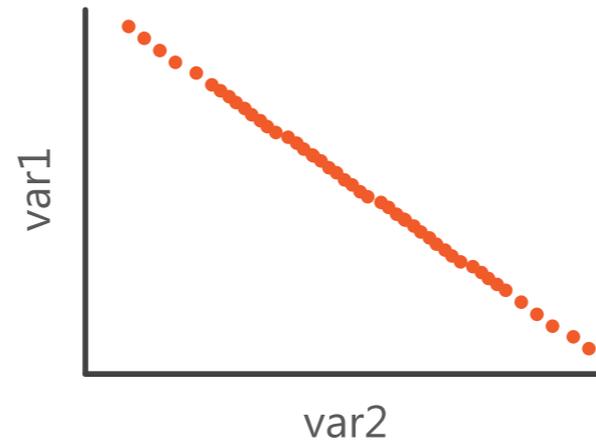
# Multidimensional Visualization

## Parallel Coordinate Plots

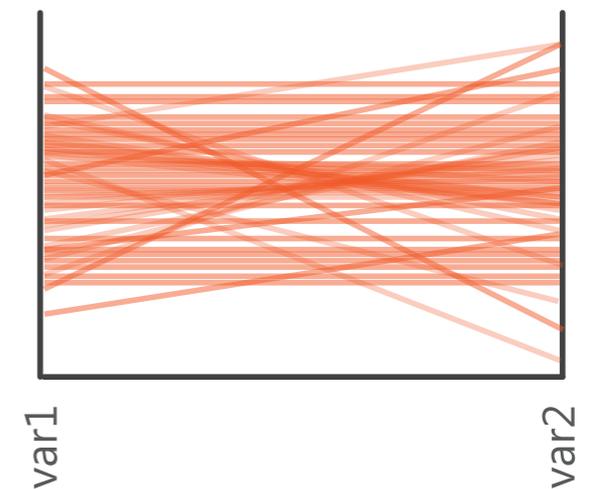
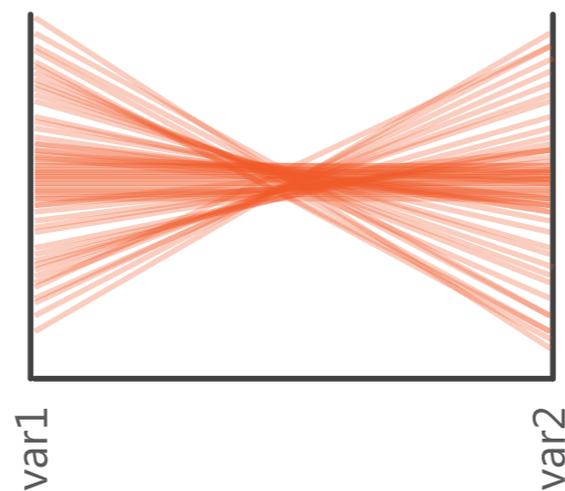
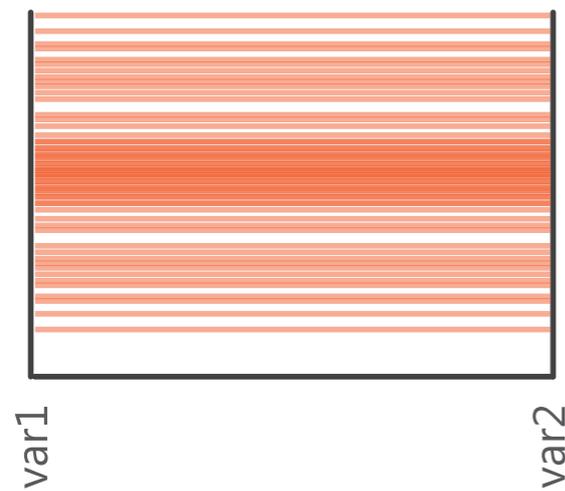
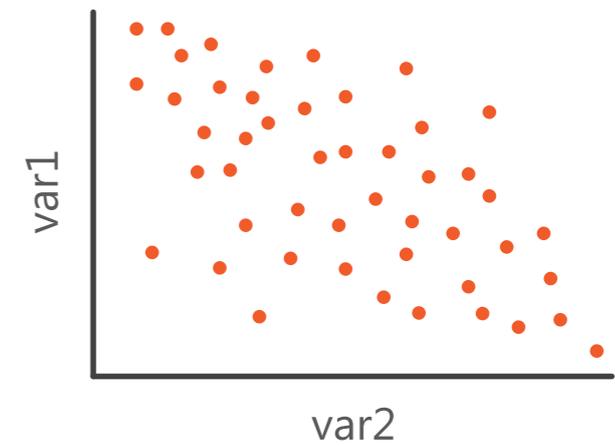
Positive Correlation



Negative (inverse) Correlation

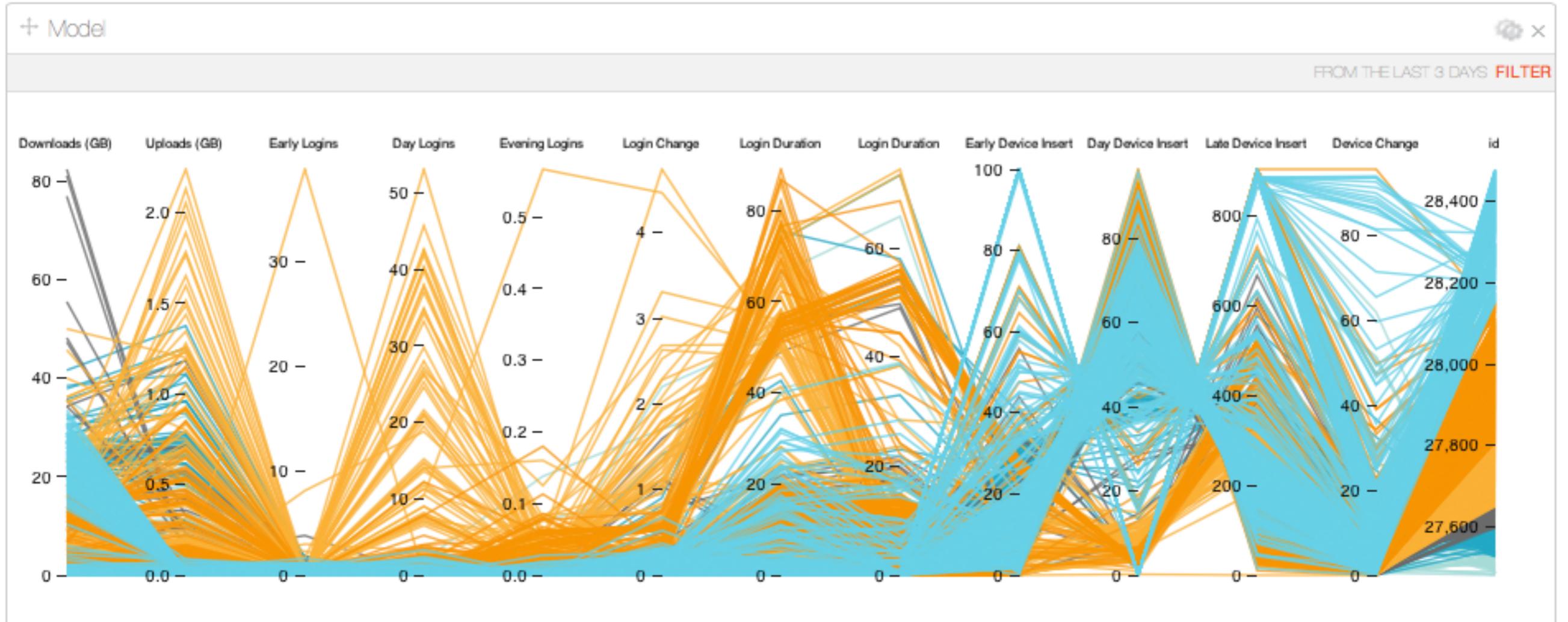


No Correlation



# Multidimensional Visualization

## Parallel Coordinate Plots



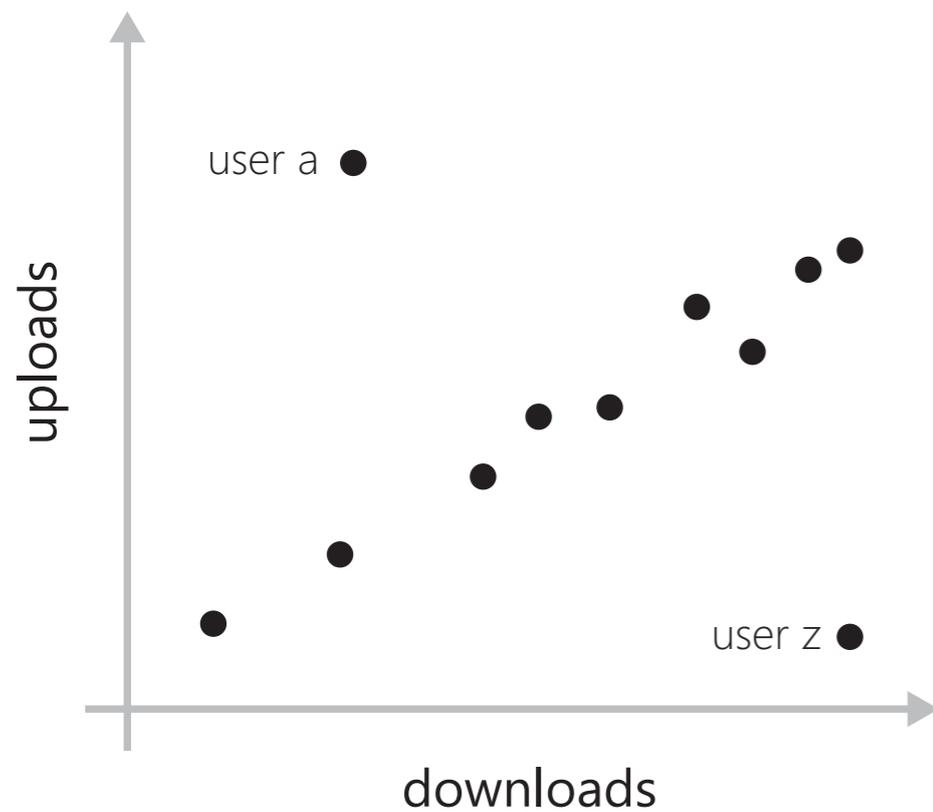
group	Downloads...	Uploads (...)	Early Logi...	Day Logins	Evening L...	Login Cha...	Login Dura...	Login Dura...	Early Devi...	Day Devic...	Late Devi...
Communicati...	19.9	0.285	0.385	0.4	0.0769	0.947	66.2	65.8	1.4	7.4	357
Communicati...	35.8	0.5	0.717	2	0.138	1.705	41	39	5.1	7.4	358
Communicati...	23.5	0.88	0.78	1.6	0.0012	2.2	52.9	51.3	2.6	12.8	329
Communicati...	0.28	0.004	0.002	0	0	0.005	0	0	99.48	0.24	876
Communicati...	0.85	0.024	0.714	0	0	0.024	0.06	0.06	81.11	15.87	717
Communicati...	0.85	0.024	0.827	0	0	0.026	0.06	0.06	81.11	15.87	717
Communicati...	0.85	0.024	0.011	0	0	0.024	0.06	0.06	81.11	17.94	717

# Multidimensional Visualization

## Parallel Coordinate Plots

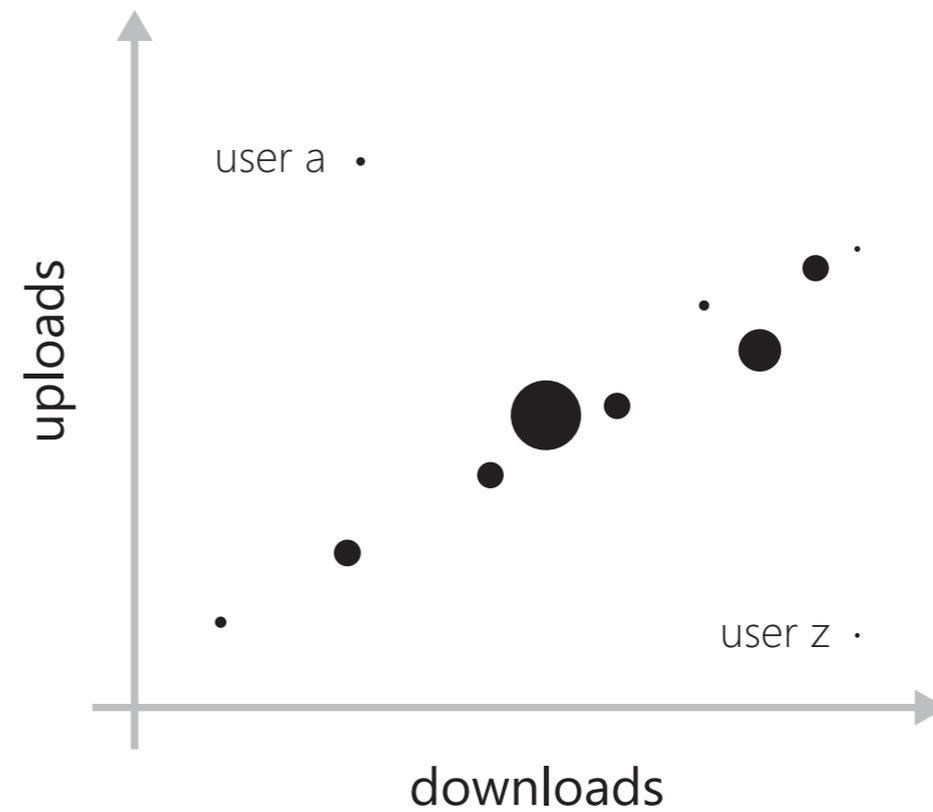
Lets take an example where we have many variables to display...  
Each user is represented by a circle

### 2 Dimensions



### 3 Dimensions

Size indicates number of logins per day

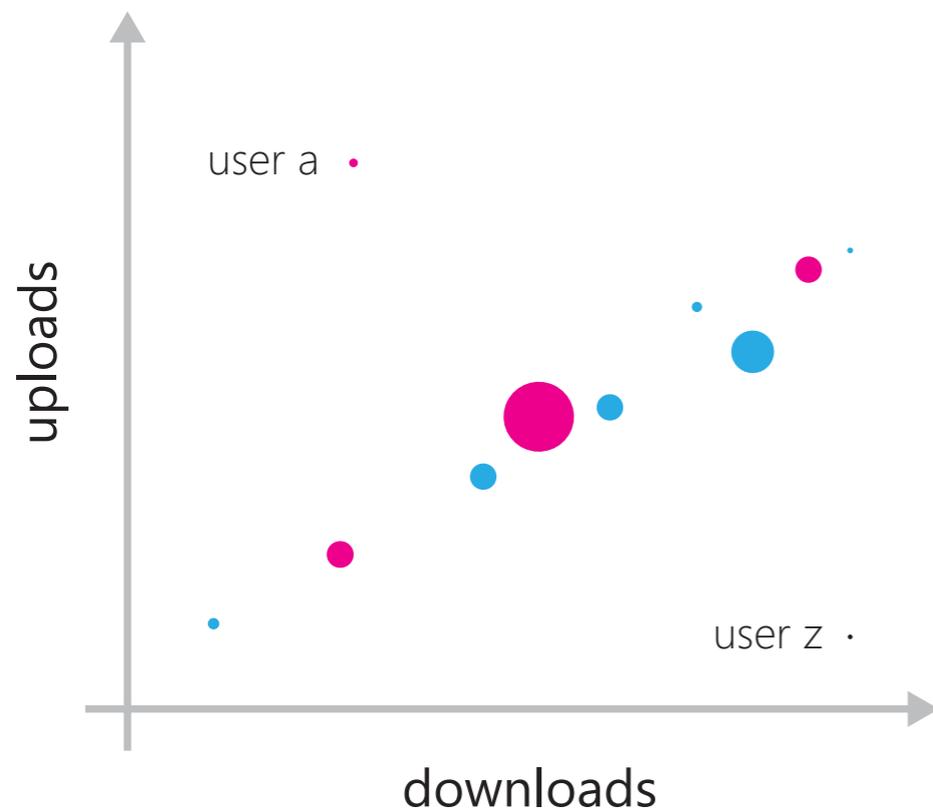


# Multidimensional Visualization

## Parallel Coordinate Plots

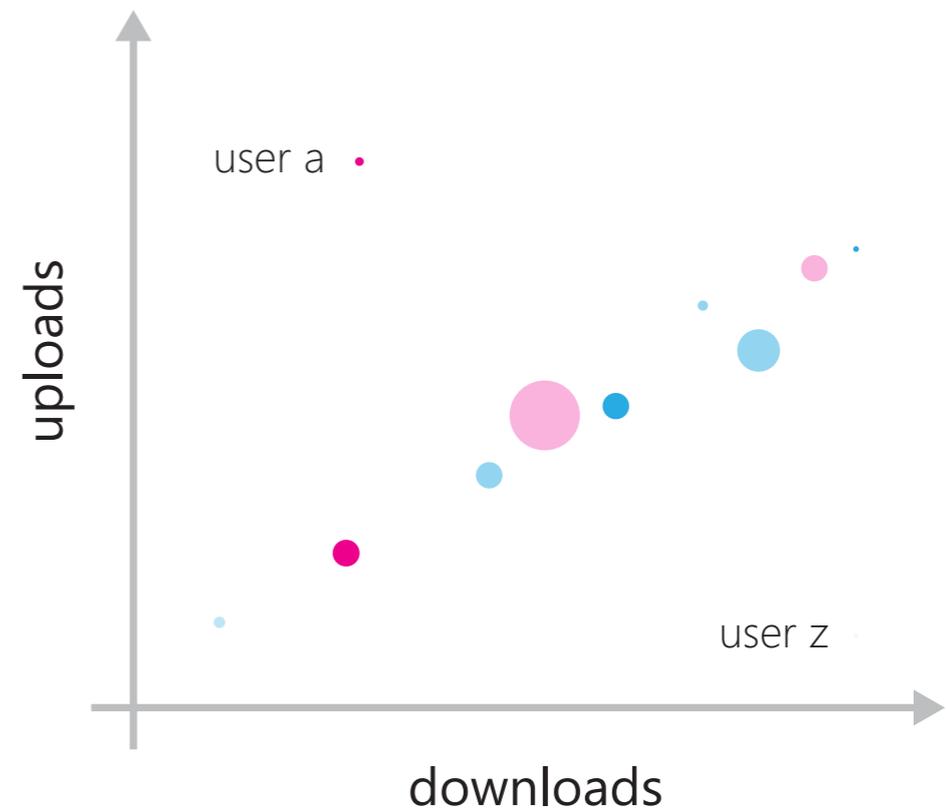
### 4 Dimensions

Color indicates users department



### 5 Dimensions

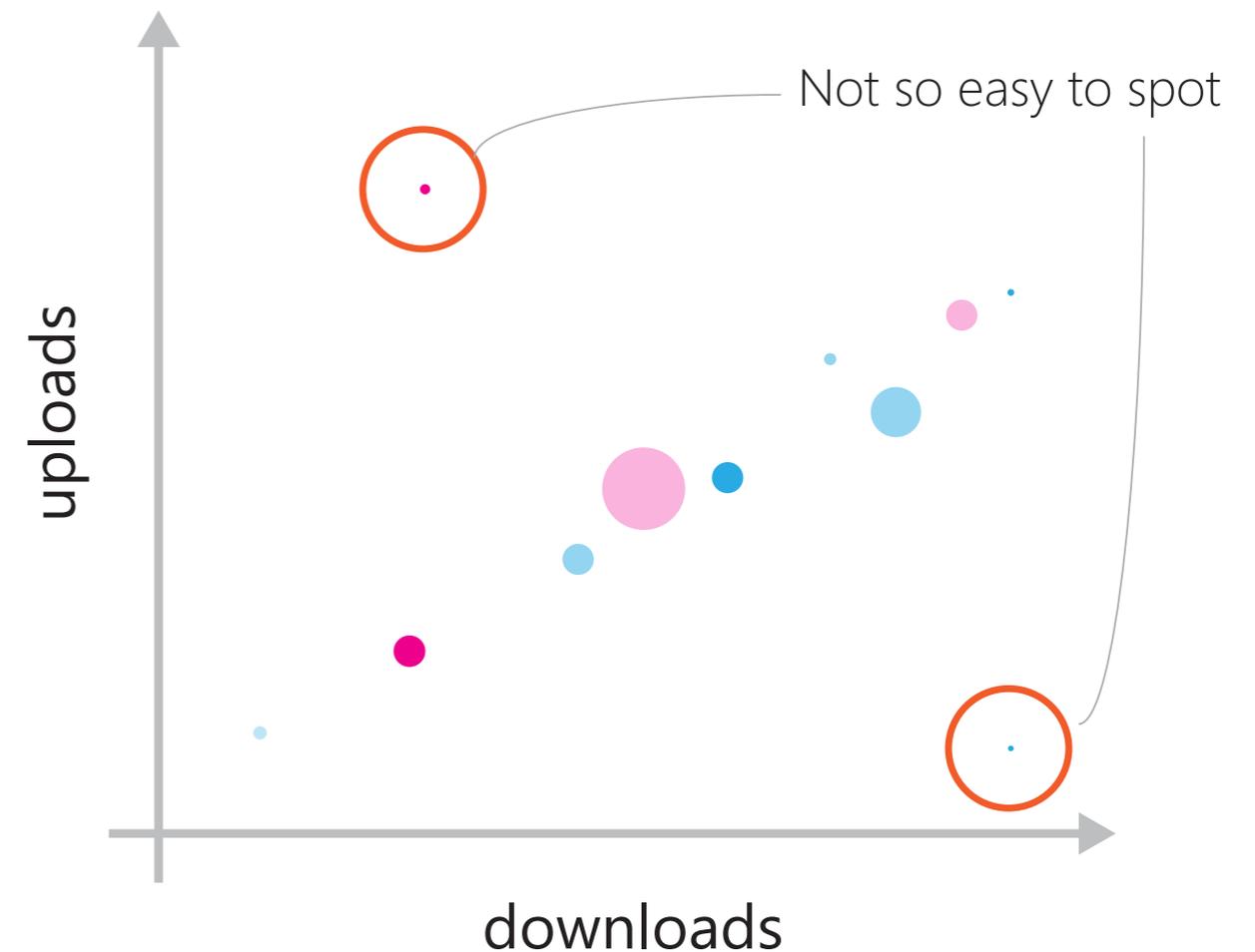
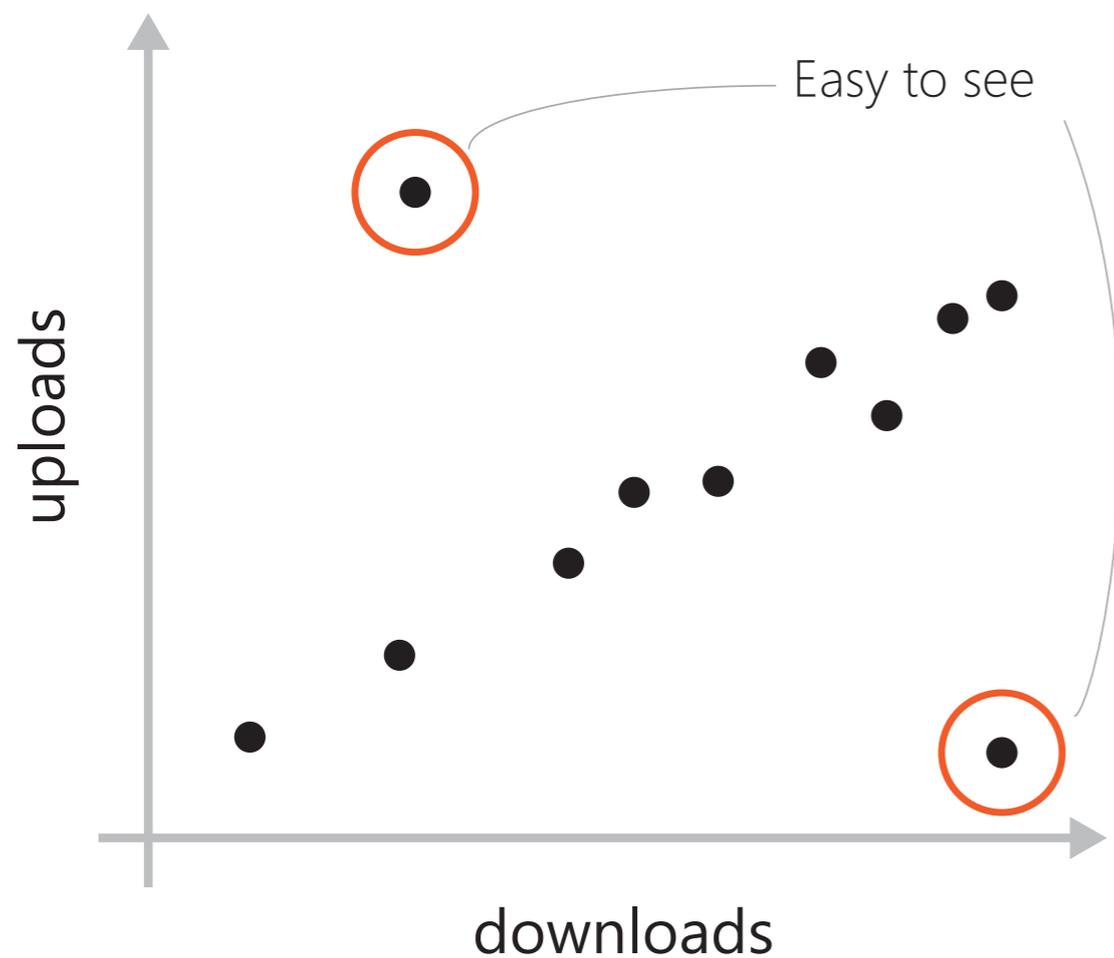
Transparency indicates consistency in logins



As we get to higher levels of dimensions, we'll have problems. Our choice of visual encoding will affect the visual availability of each dimension to the user.

# Multidimensional Visualization

## Parallel Coordinate Plots



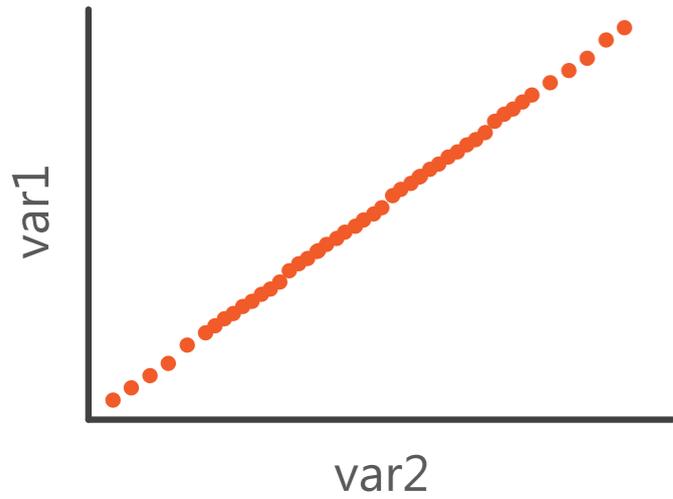
Parallel coordinates are a visualization technique employed when a large number of dimensions need to be displayed (often without a temporal element) and where each of those dimensions can be equally important in the decision making process.

In the scatter plots here, it's easy to see **correlation** between downloads and uploads, but with the other dimensions that's difficult.

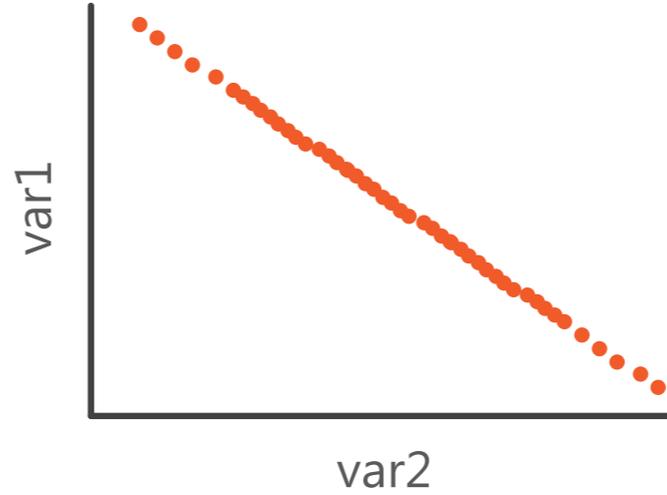
# Multidimensional Visualization

## Parallel Coordinate Plots

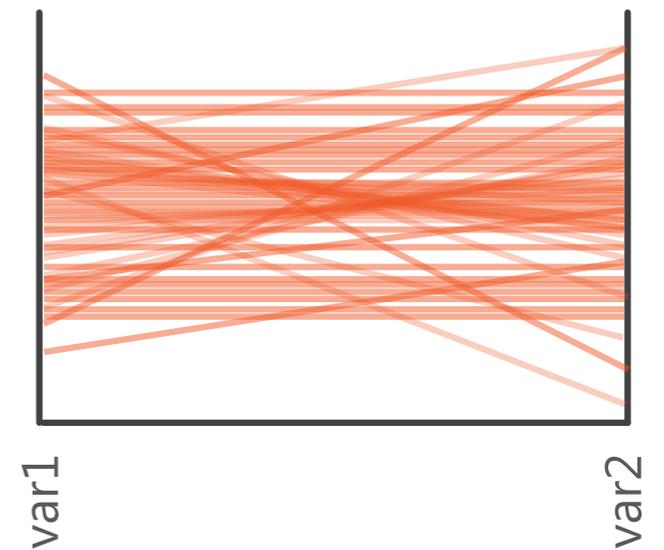
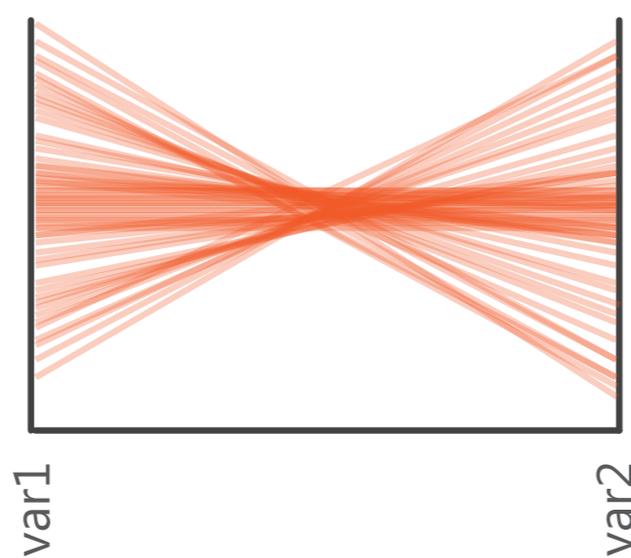
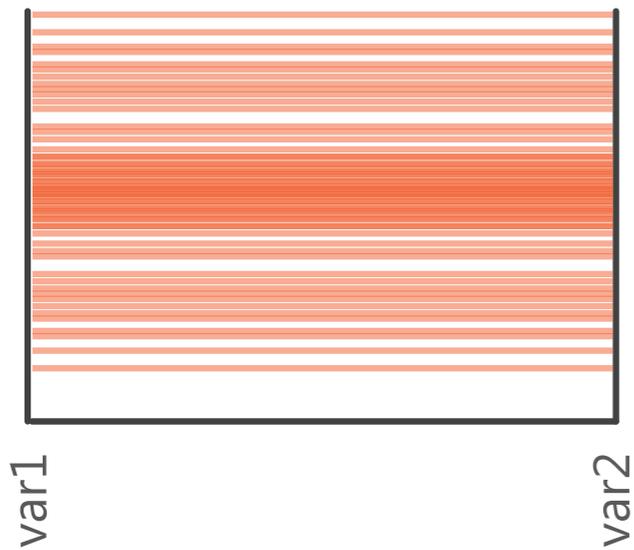
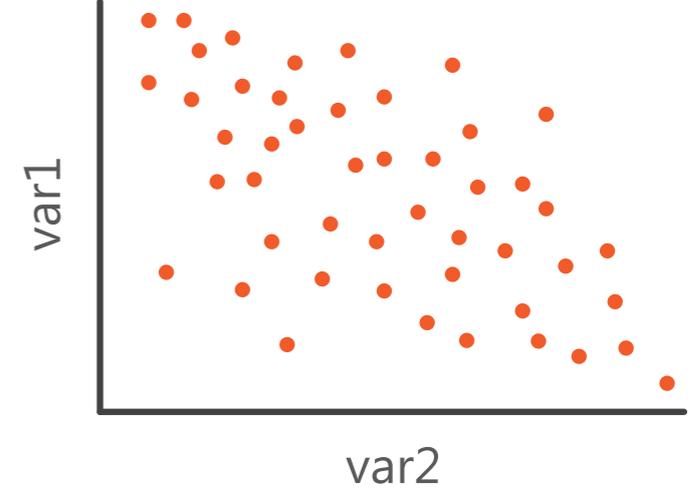
Positive Correlation



Negative (inverse) Correlation



No Correlation



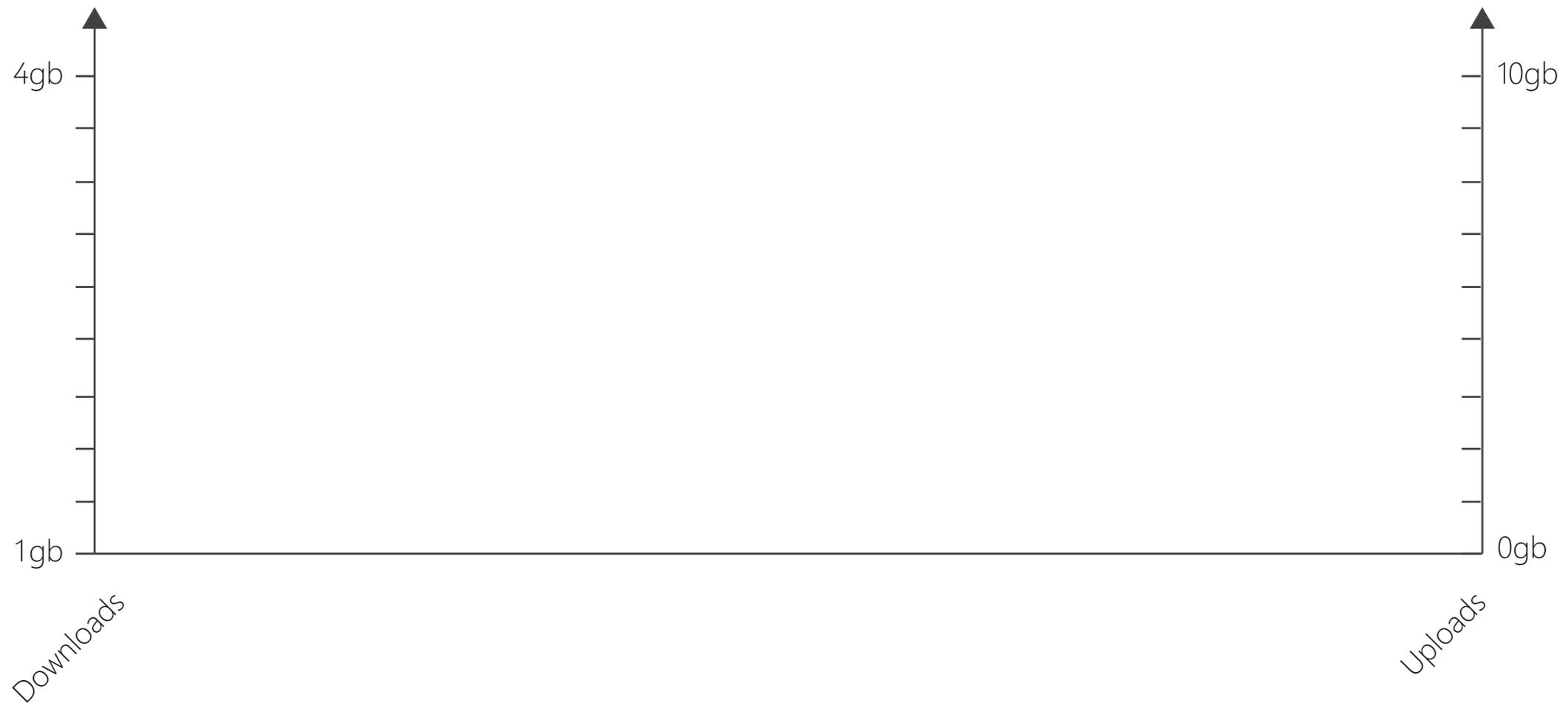
# Multidimensional Visualization

## Parallel Coordinate Plots

2 Dimensions

Uploads

Downloads



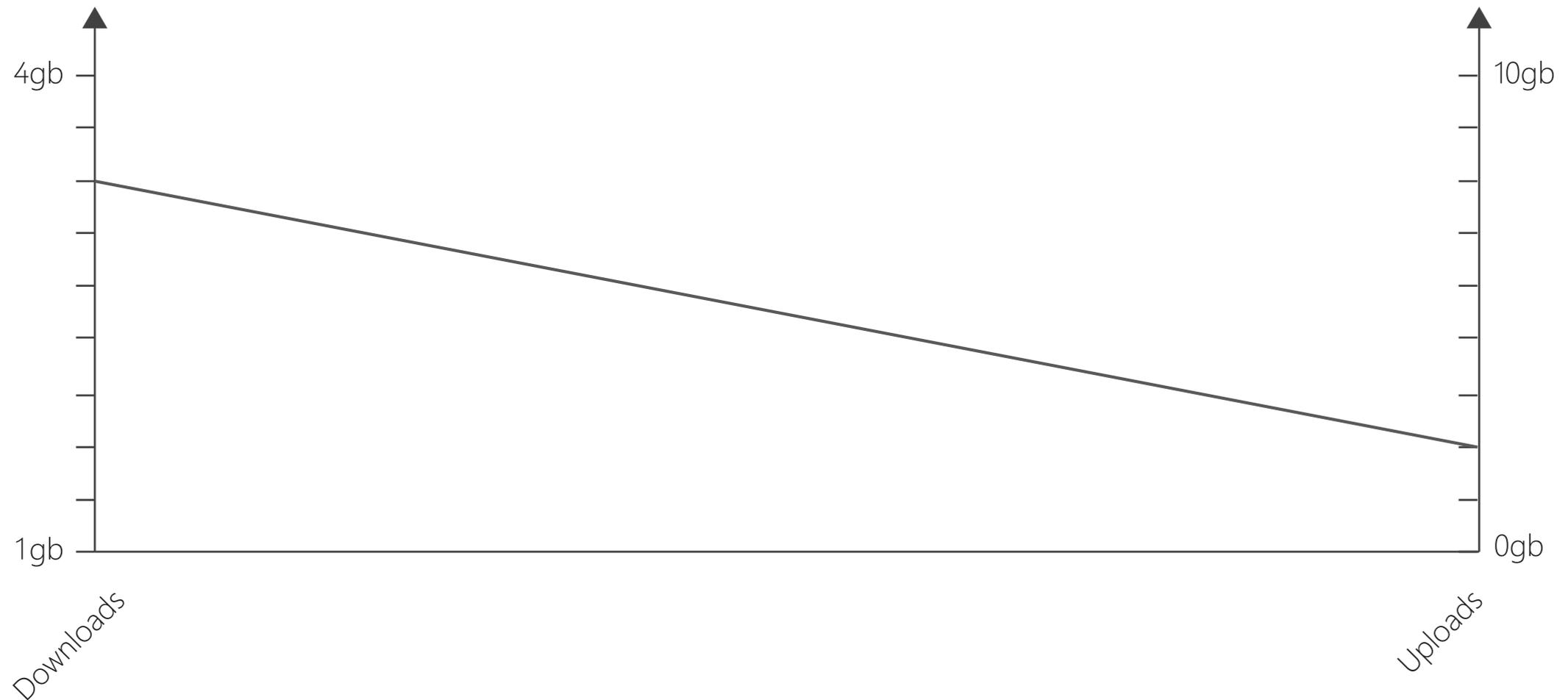
# Multidimensional Visualization

## Parallel Coordinate Plots

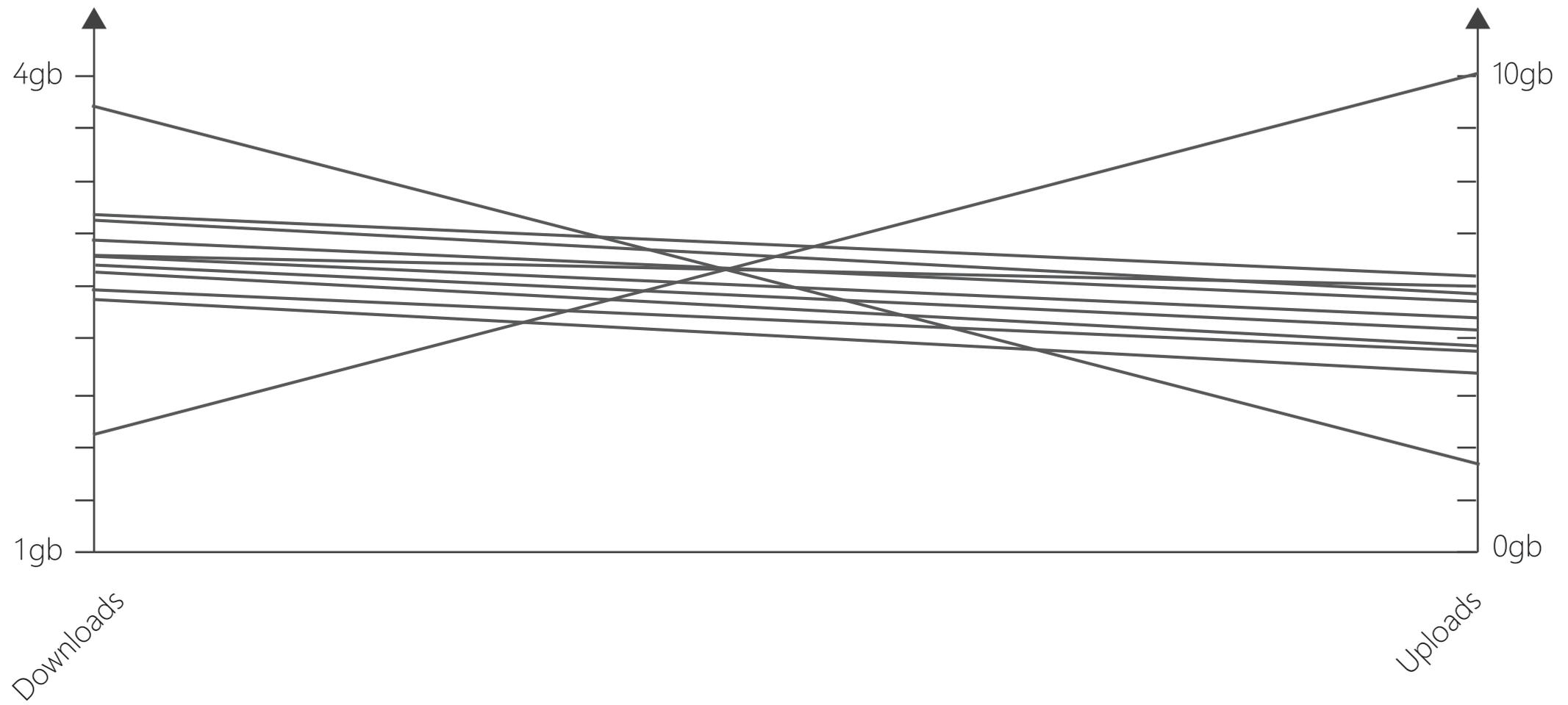
2 Dimensions

Uploads

Downloads

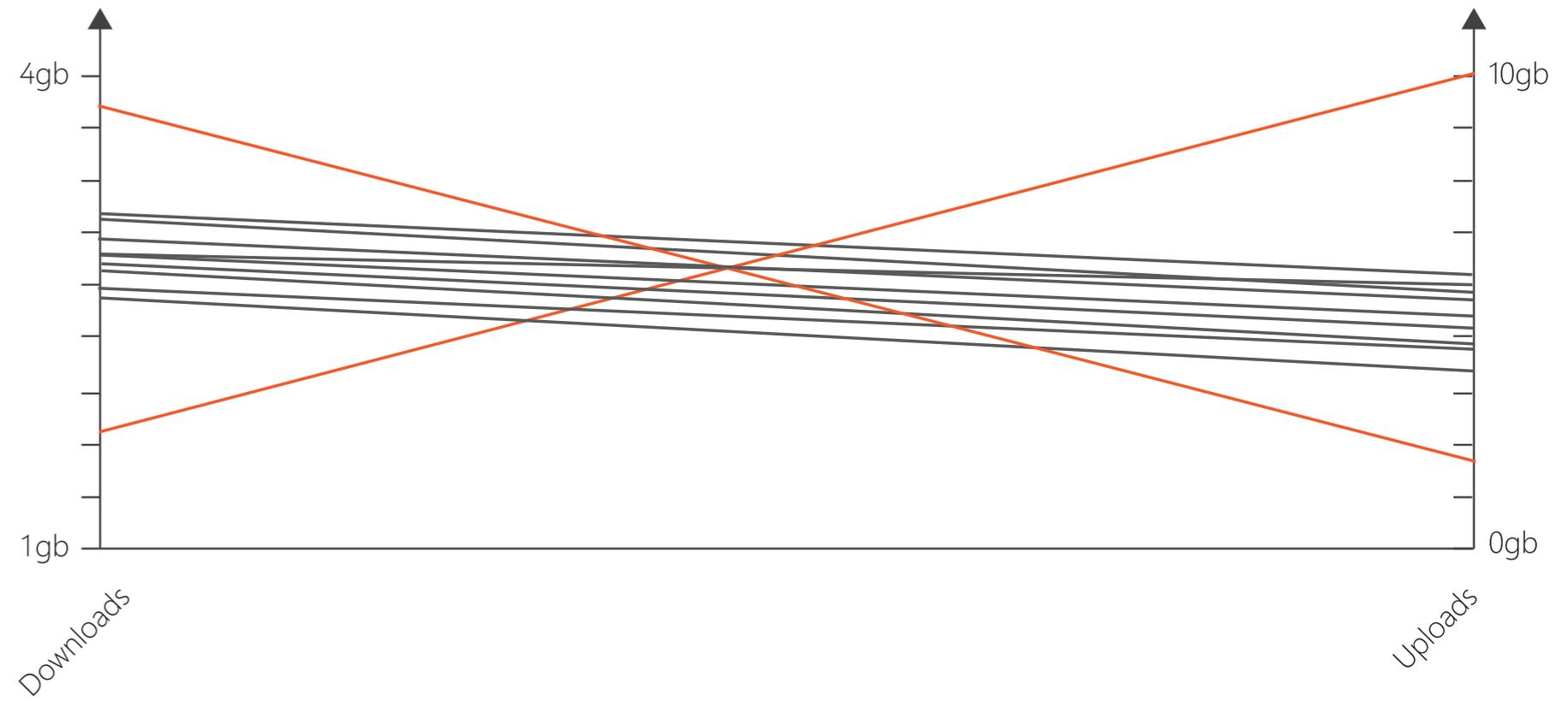
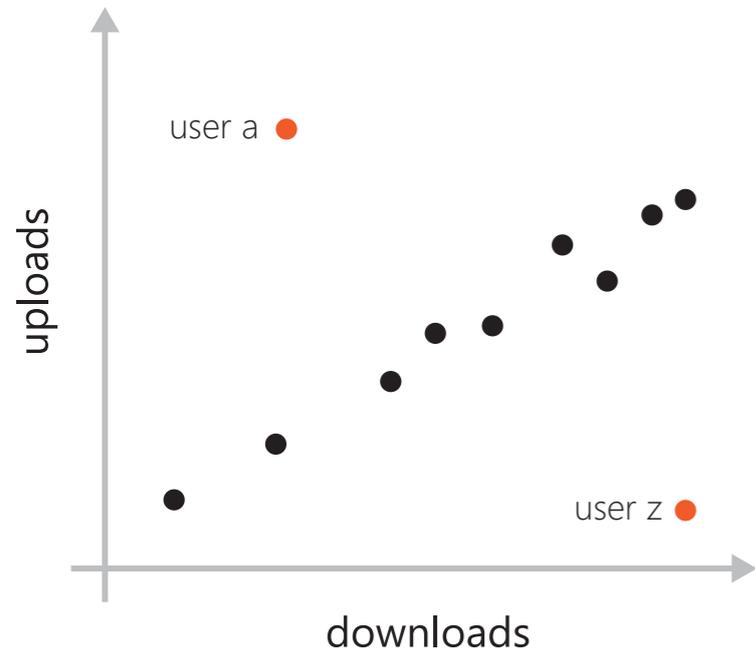


2 Dimensions  
Uploads  
Downloads



## 2 Dimensions

Uploads  
Downloads

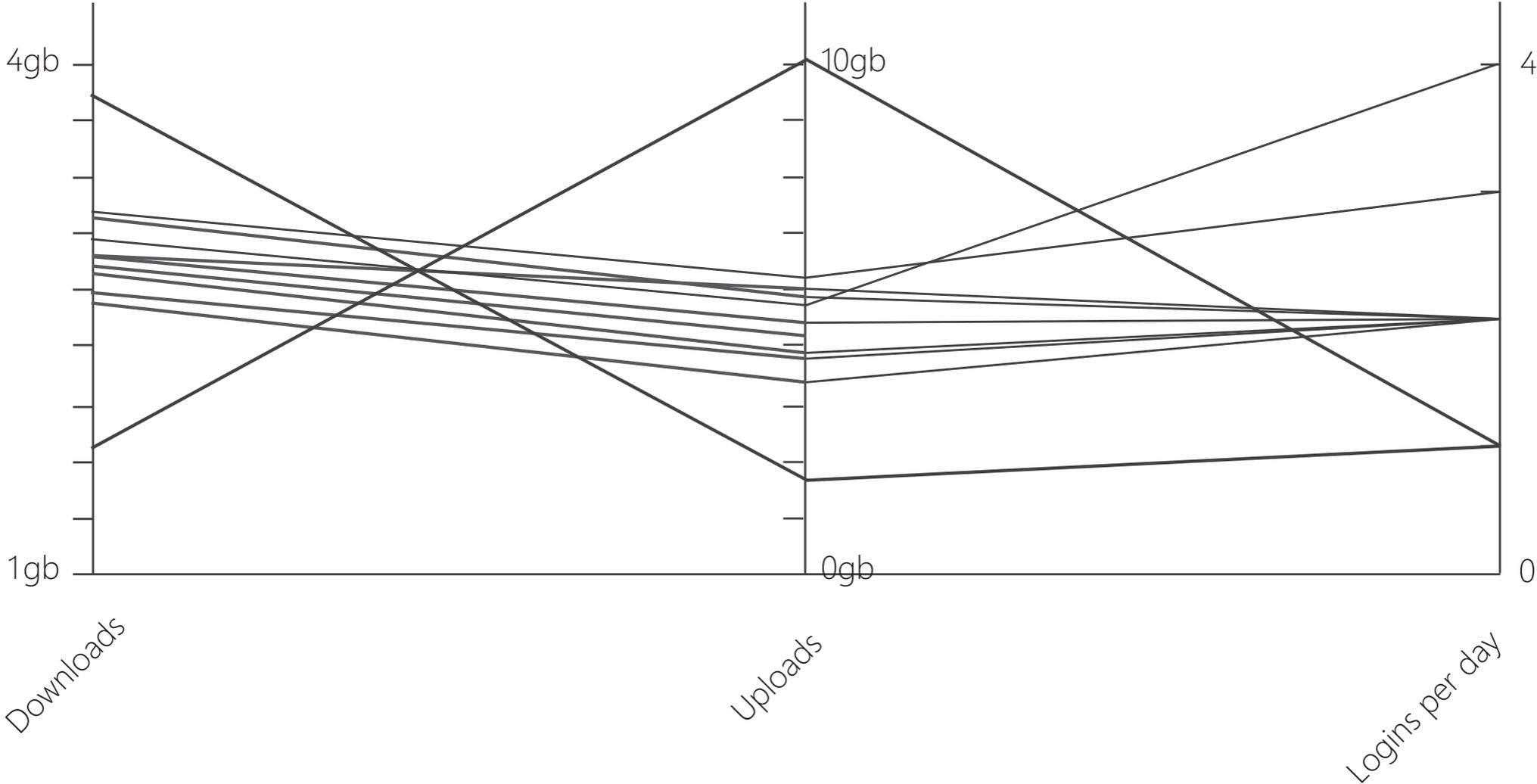


3 Dimensions

Uploads

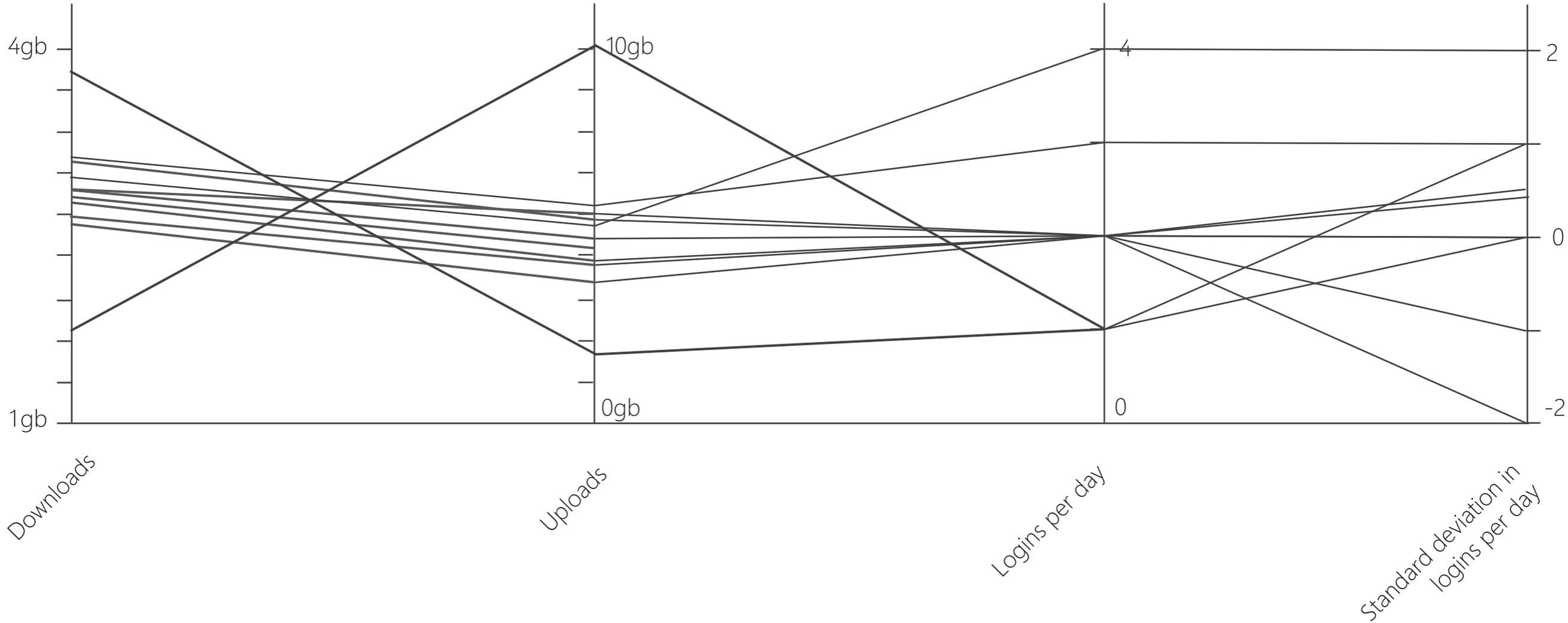
Downloads

Logins per day



4 Dimensions

- Uploads
- Downloads
- Logins per day
- Std. deviation in logins per day



5 Dimensions

Uploads

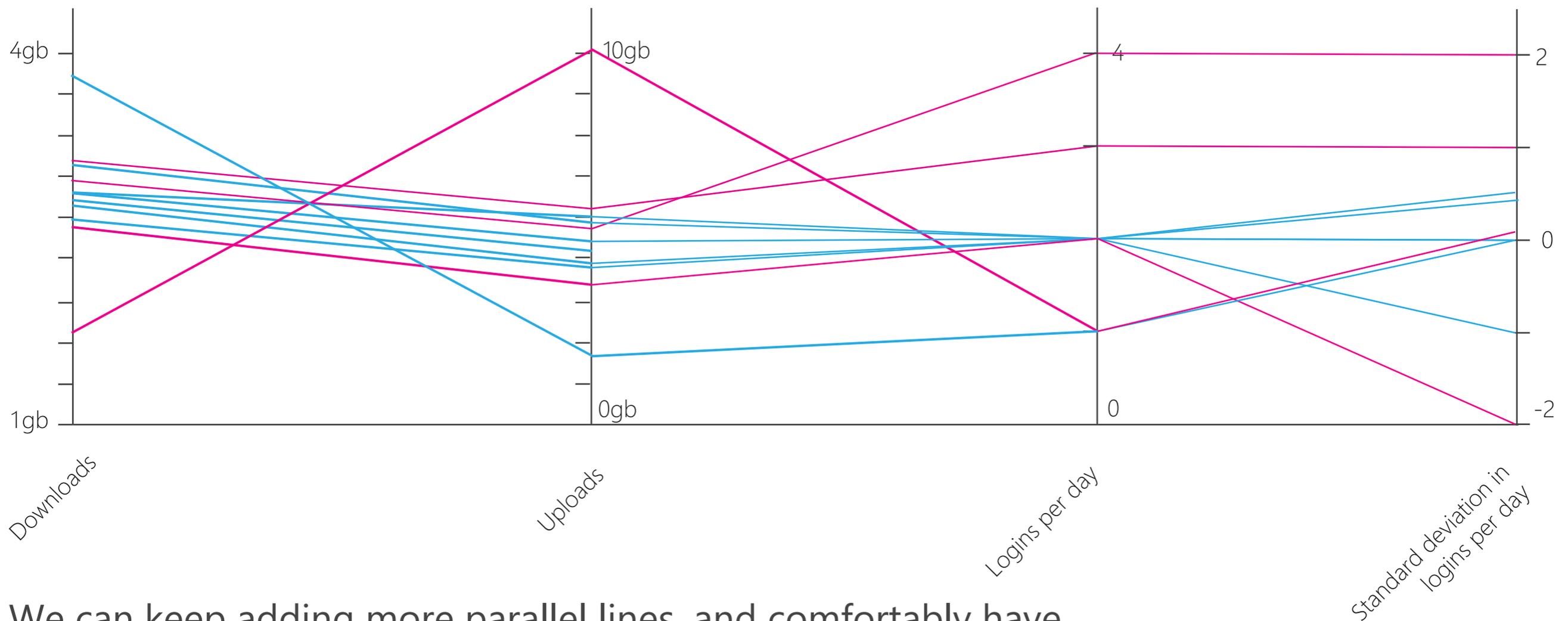
Downloads

Logins per day

Std. deviation in logins per day

Department

We use colour for department since it's categorical information.

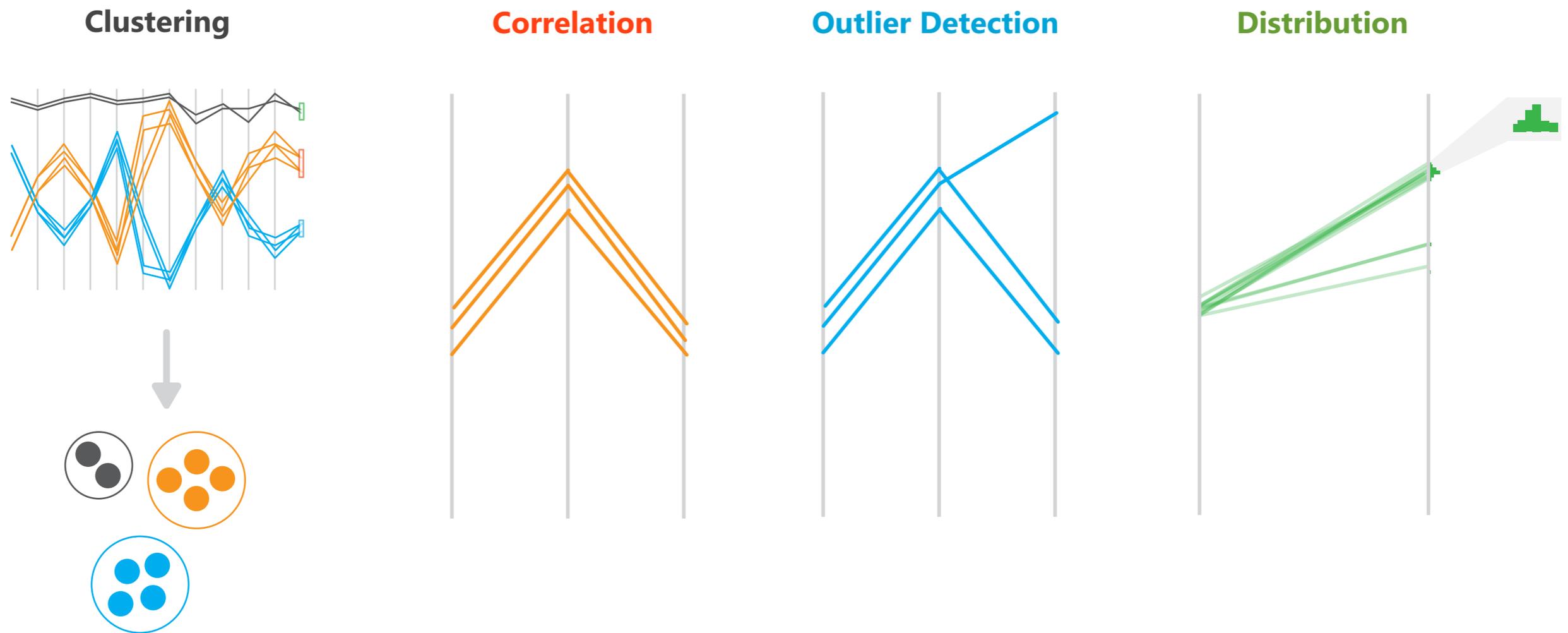


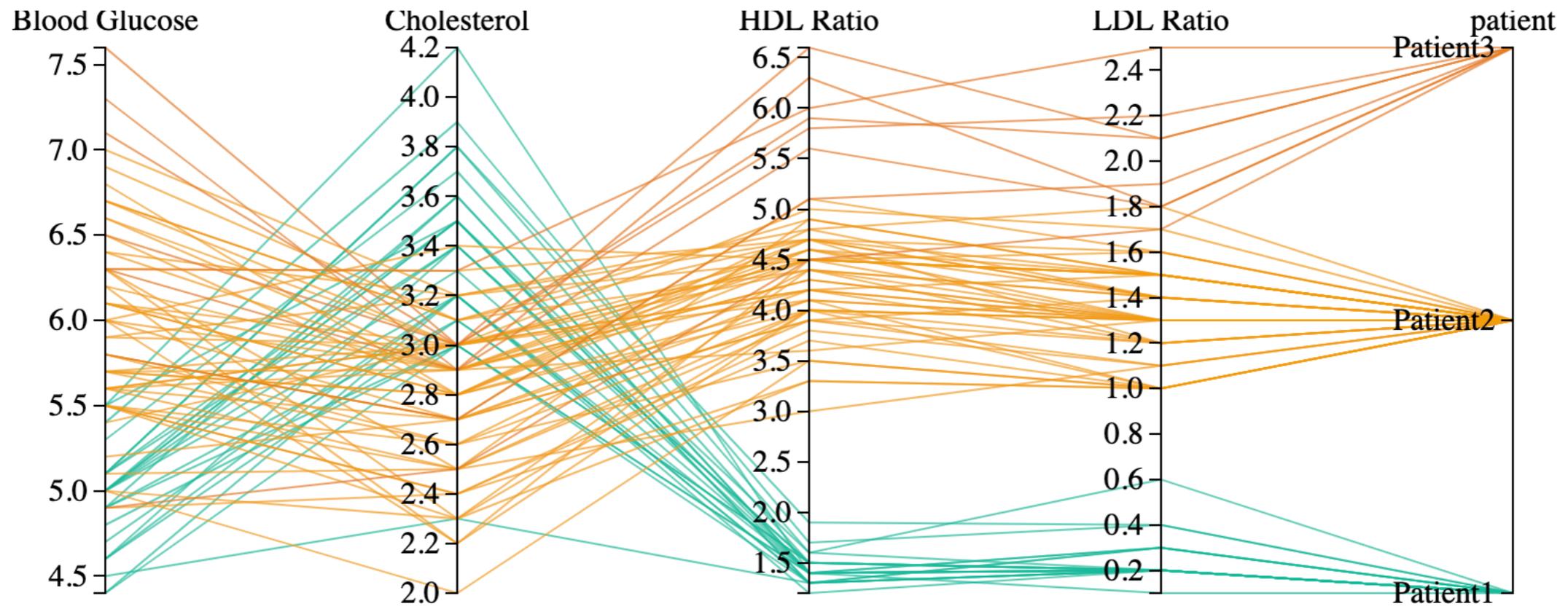
We can keep adding more parallel lines, and comfortably have around 20 dimensions for many users displayed at once.

# Multidimensional Visualization

## Parallel Coordinate Plots

Parallel coordinates provide an efficient way to visualize many variables, along with their associated **clusters**, **anomalies**, value **distributions** and **correlations**.

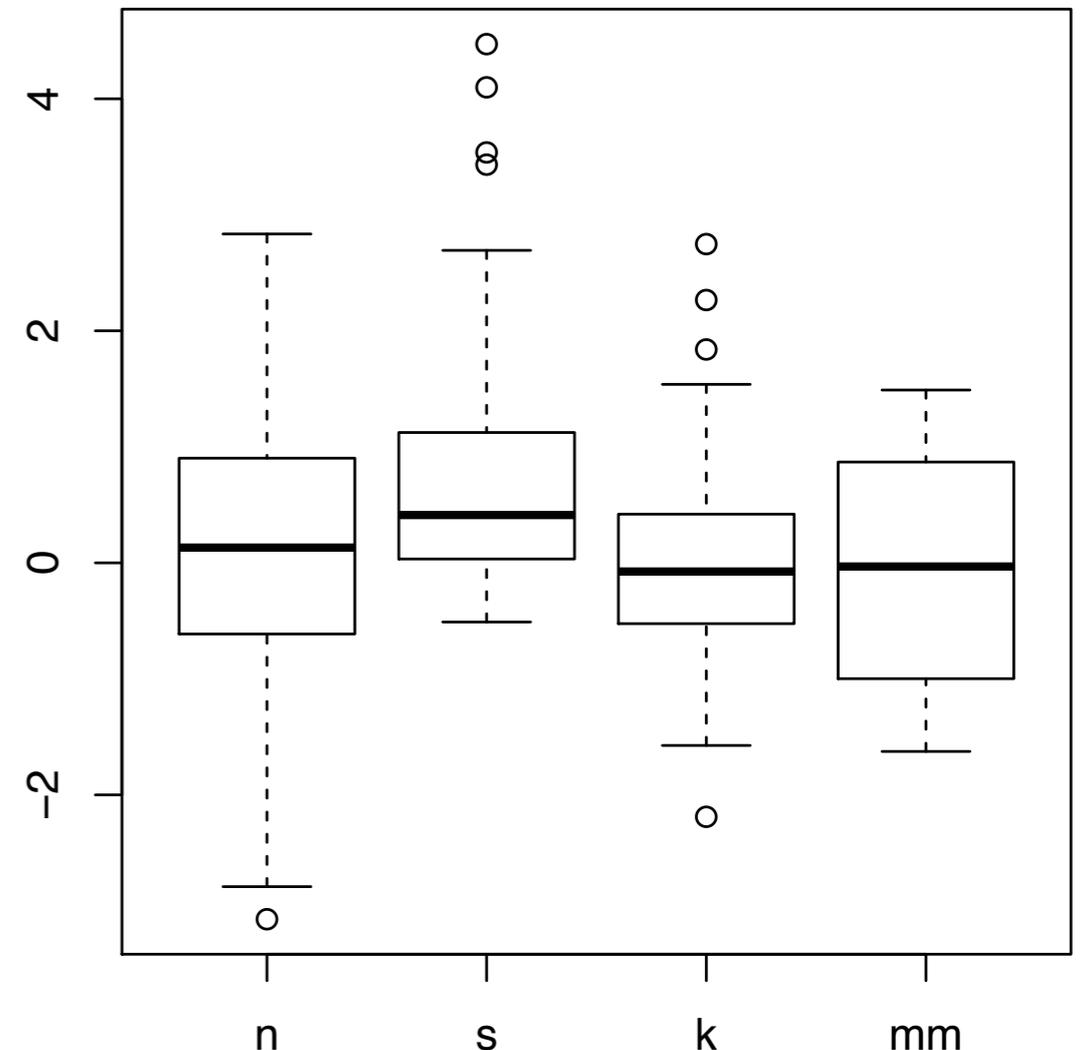




# Multidimensional Visualization

## Glyphs

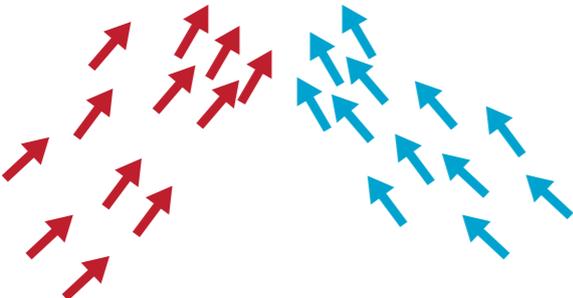
- static item aggregation
- **task:** find distribution
- **data:** table
- derived data
  - 4 quantitative attributes
    - median: central line
    - lower and upper quartile: boxes
    - lower upper fences: whiskers
  - outliers beyond fence cutoffs explicitly shown



# Multidimensional Visualization

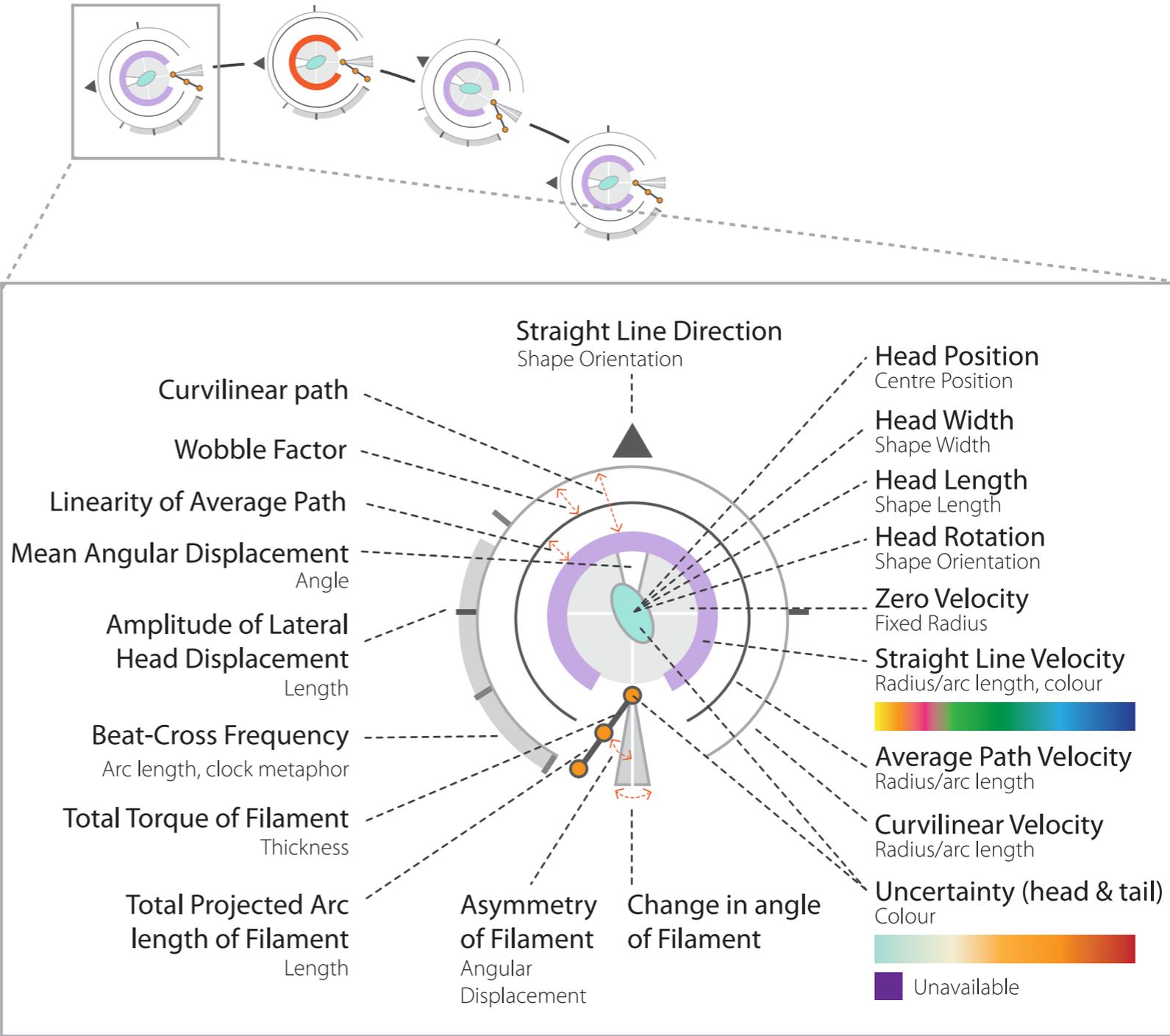
## Glyphs

### Simple Glyph



**Temperature - Colour** ■■■■  
**Wind direction - Orientation** ↑ ↗ ↘ →  
**Wind Speed - Proximity**  
**Location - Position**

### Complex Glyph



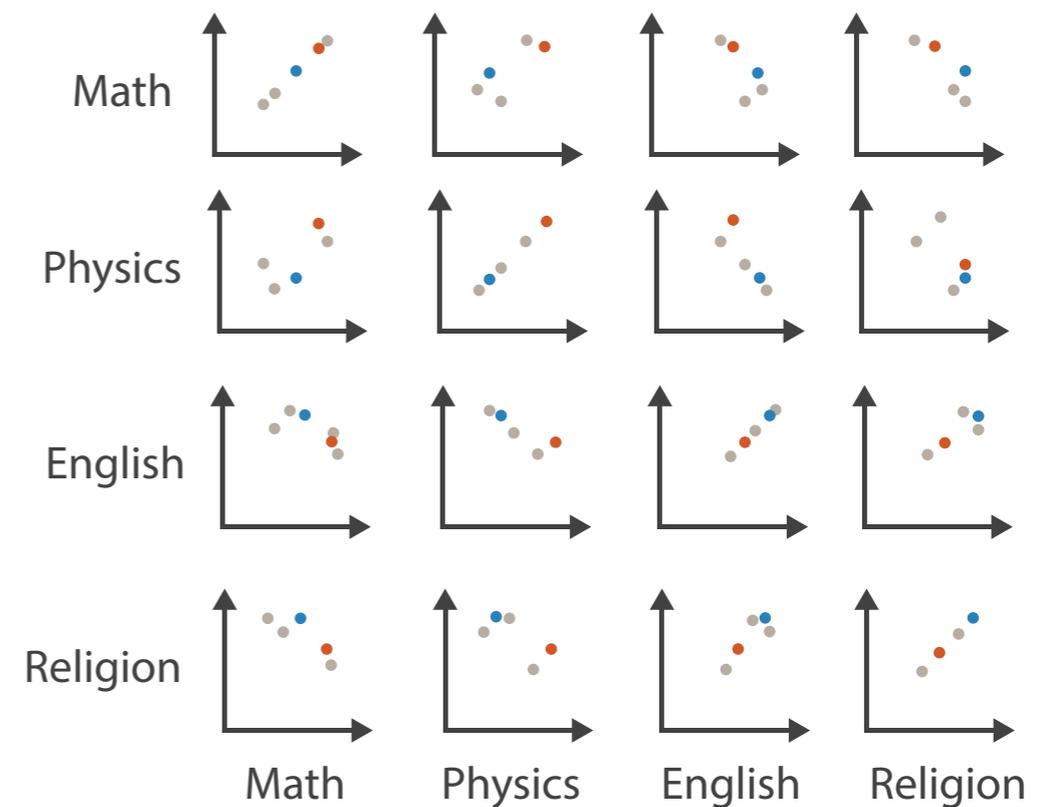
# Multidimensional Visualization

## A Simple Example | Student Test Results

Table

Math	Physics	English	Religion
85	95	71	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Scatter Plot Matrix



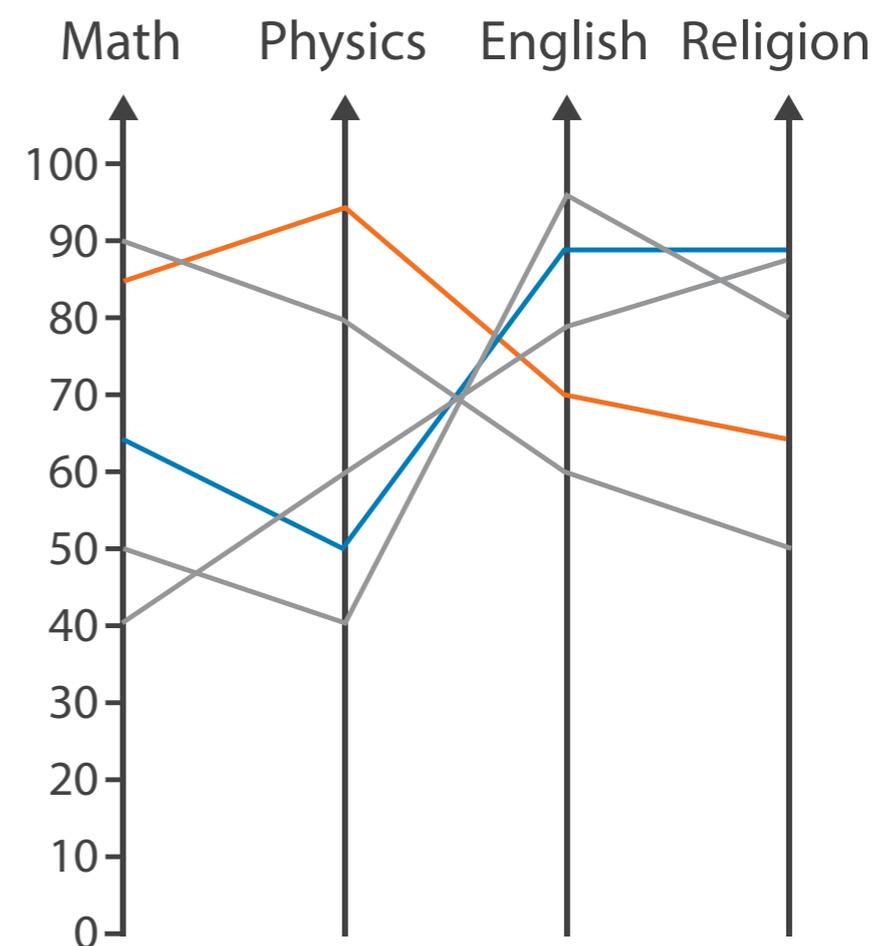
# Multidimensional Visualization

## A Simple Example | Student Test Results

Table

Math	Physics	English	Religion
85	95	71	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Parallel Coordinates



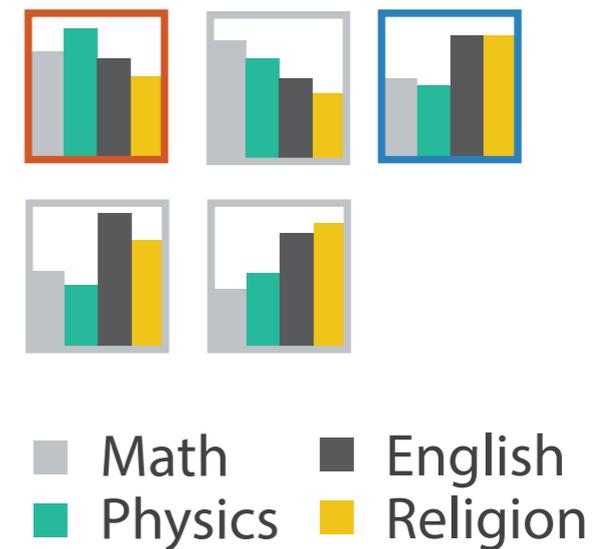
# Multidimensional Visualization

## A Simple Example | Student Test Results

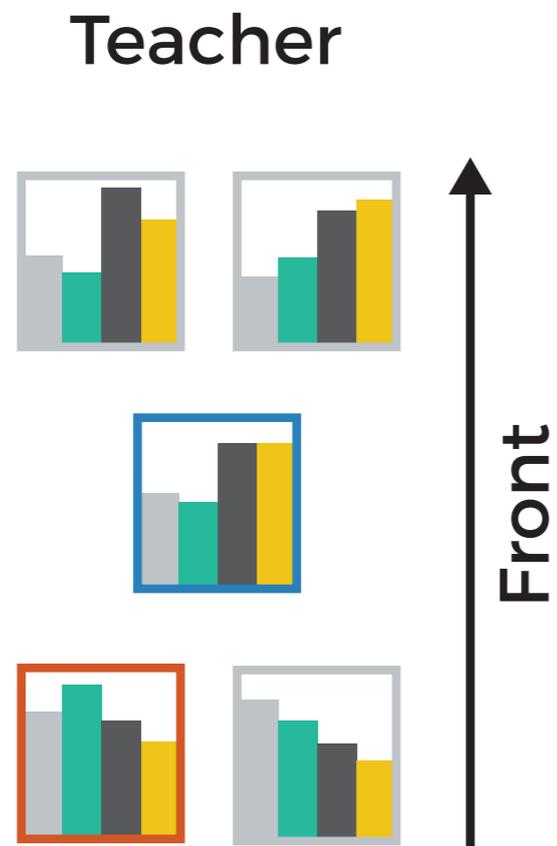
Table

Math	Physics	English	Religion
85	95	71	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Glyph

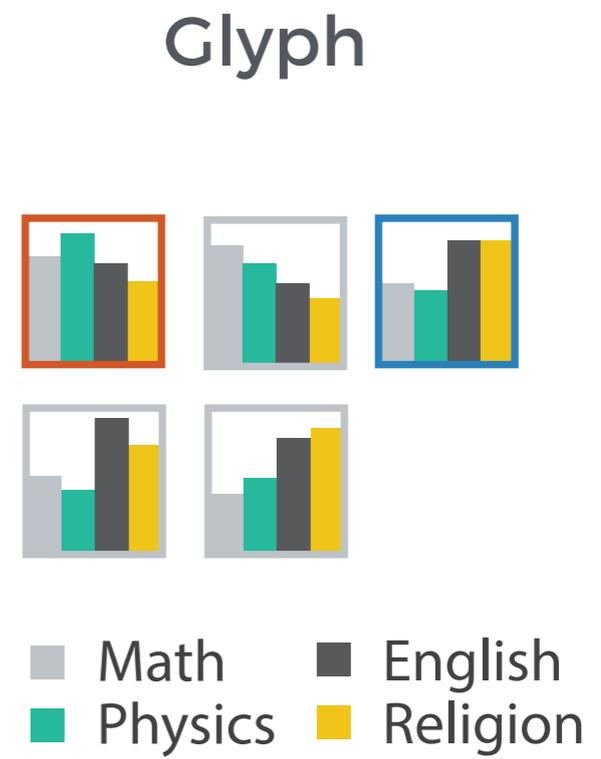


## Arrange Spatially



in

## it Test Results

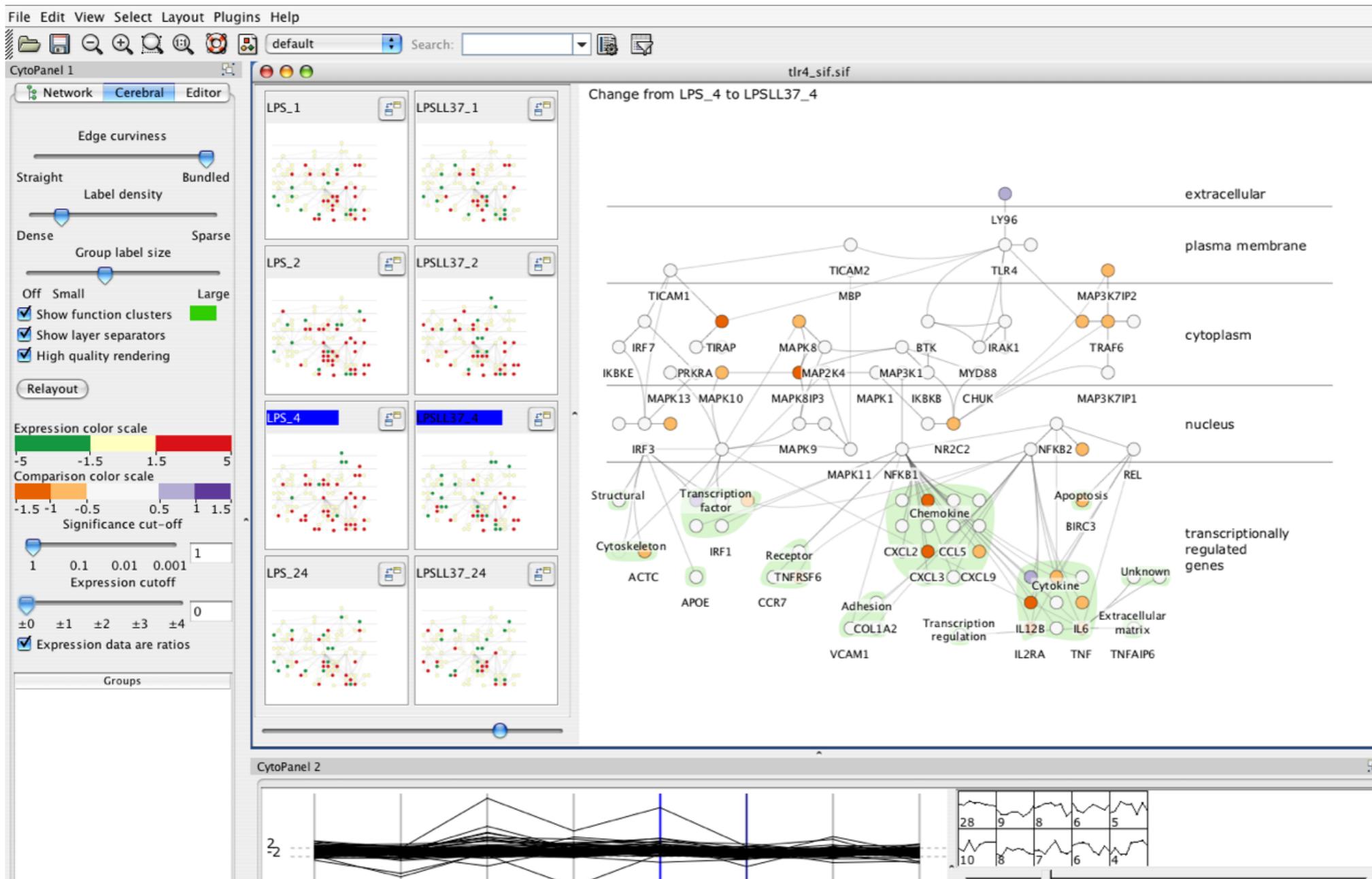


# What about topological data?

Representing trees and graphs...

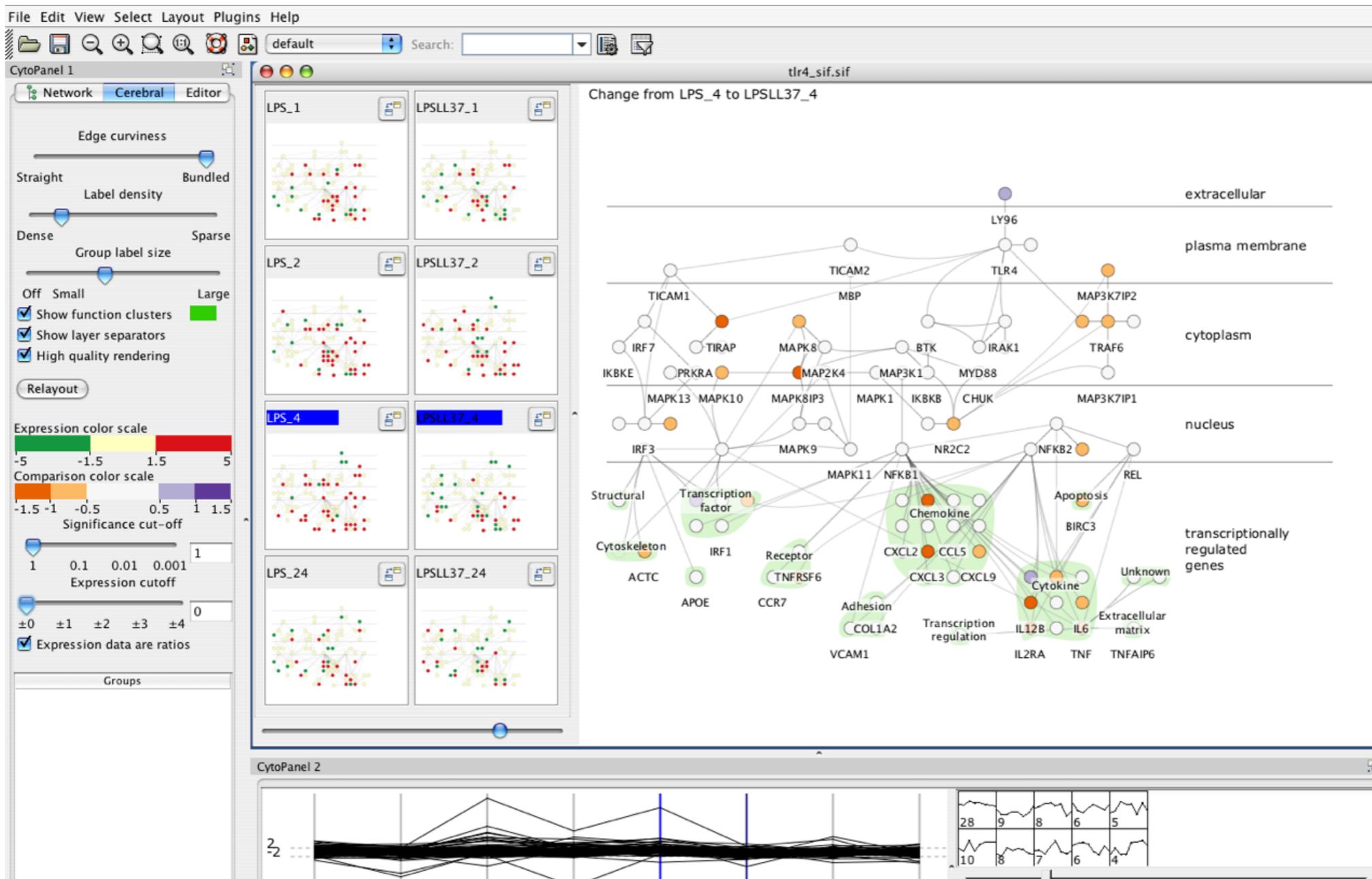
# Graphs/Networks

In this case, it's a semantic mapping to the underlying biological pathways.



# Graphs/Networks

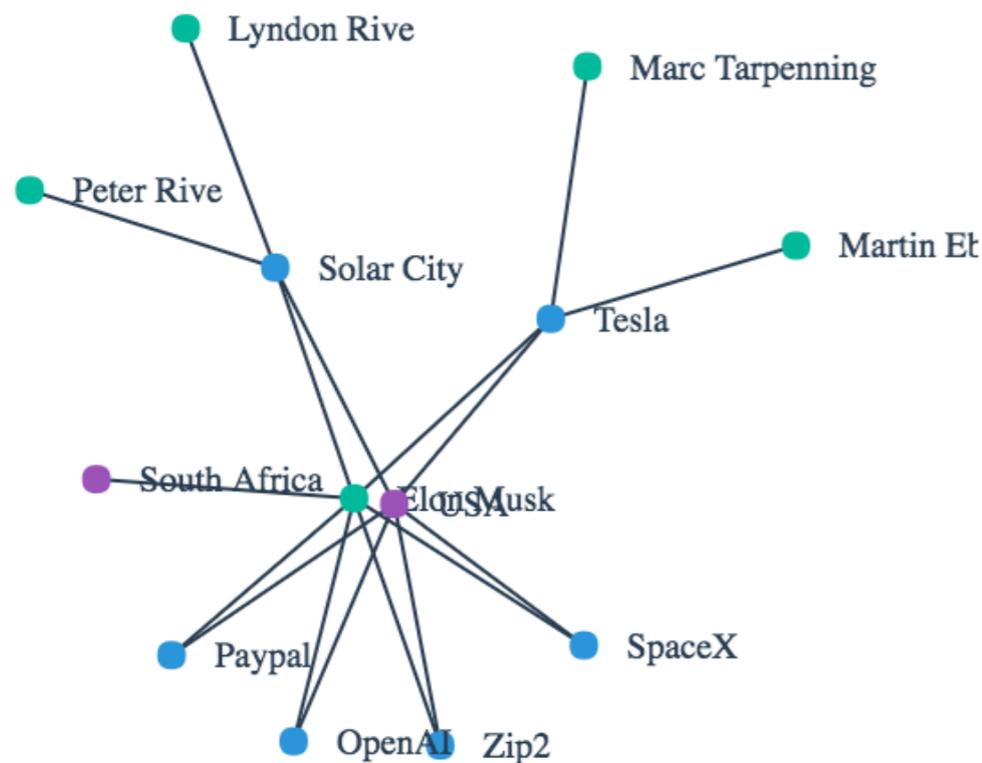
In this case, it's a semantic mapping to the underlying biological pathways.



*Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context. Barsky, Munzner, Gardy, and Kincaid. IEEE TVCG (Proc. InfoVis) 14(6):1253-1260, 2008.]*

# Graphs/Networks

## Force Directed Graphs



<http://jsfiddle.net/7a7b5dwp/>

The most used of all graphical layouts on the web.

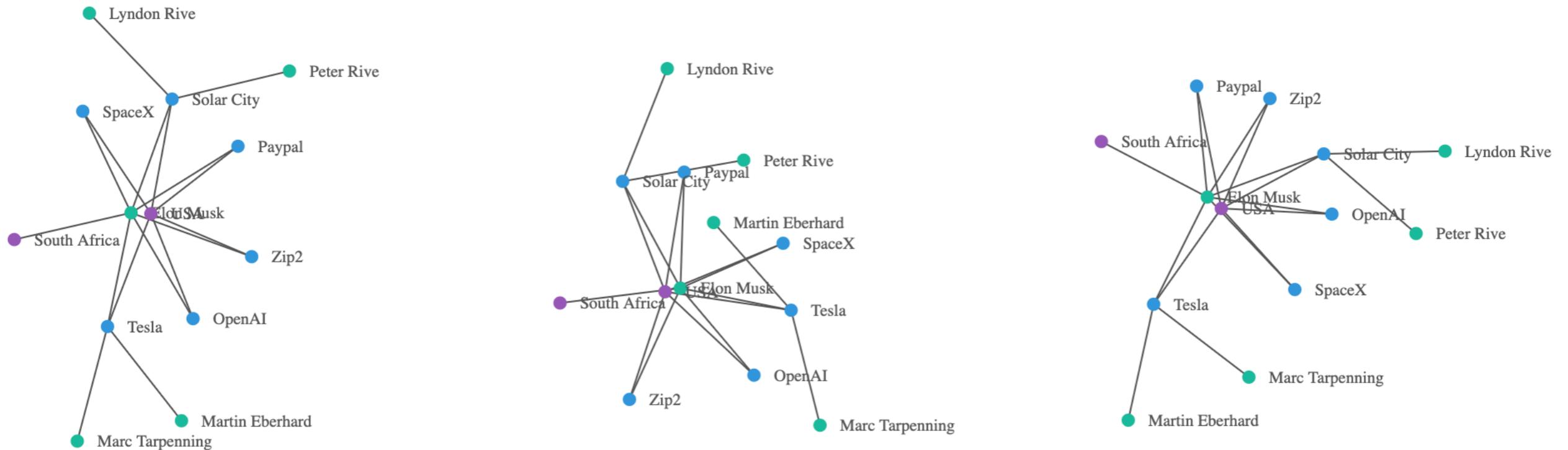
But beware. As we saw earlier, Gestalt laws tell us that items that are close together are seen as more similar than those that are not.

Unfortunately, completely unrelated nodes can be perceived as being more similar due to the layout algorithm in force directed graphs.

# Graphs/Networks

## Force Directed Graphs

A greater problem is that the network can change every time you run the code, and adding or removing nodes can drastically affect



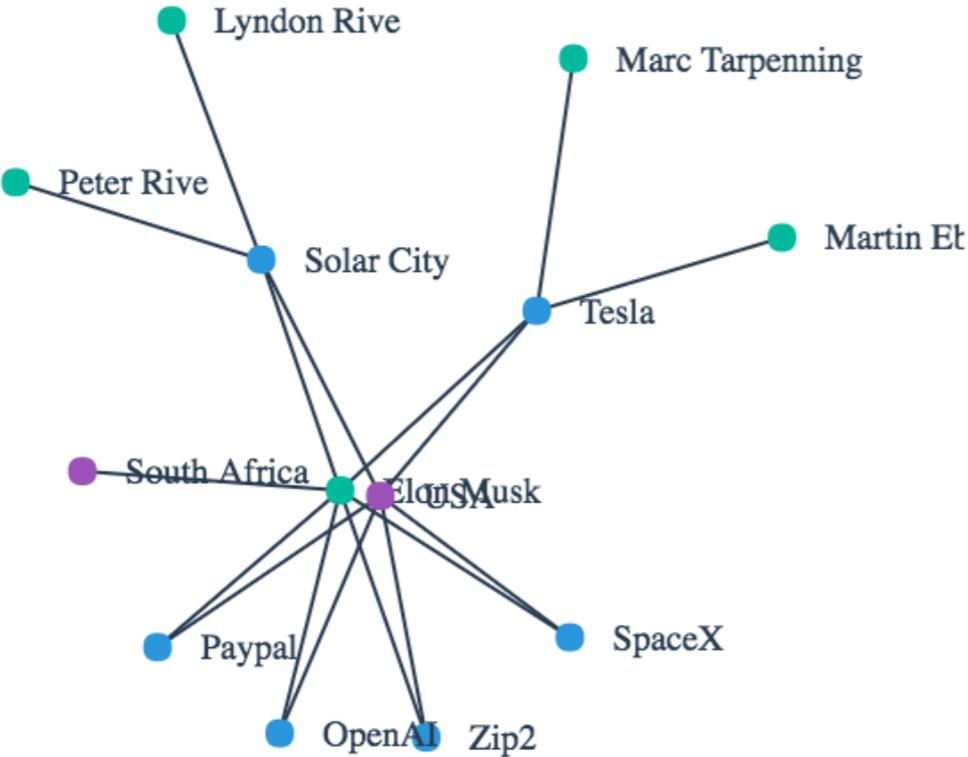
<http://jsfiddle.net/7a7b5dwp/>

**Same graph, different layouts. Not easy to compare.**

**Are there solutions to this?**

# Graphs/Networks

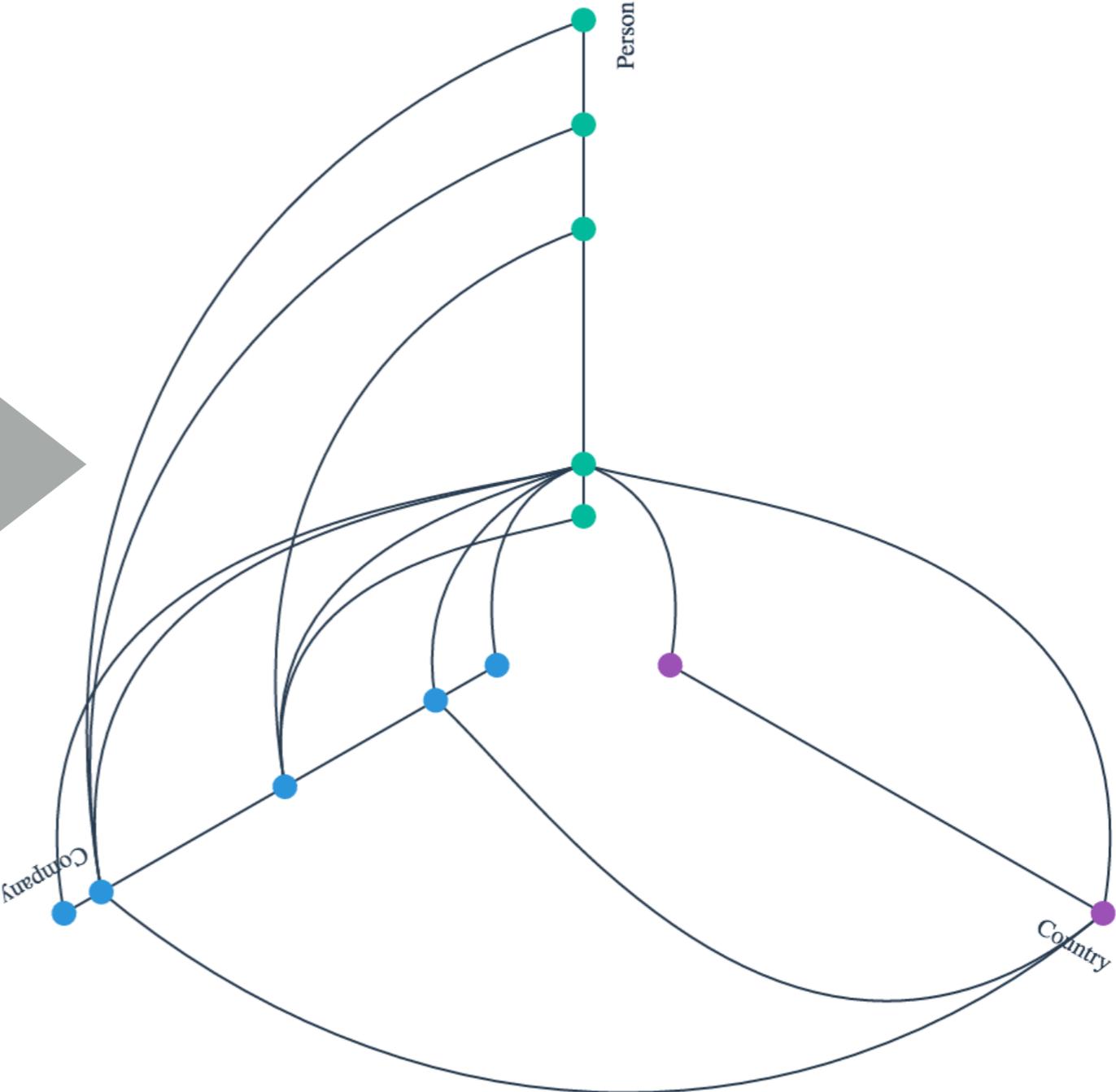
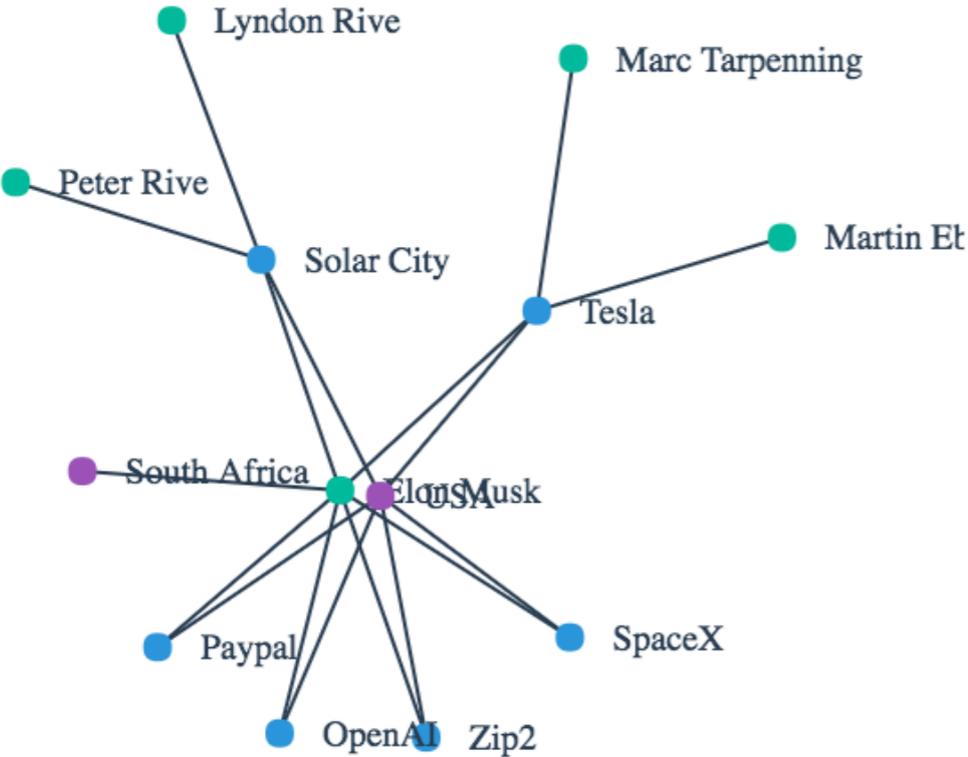
## Hive Plots



<http://jsfiddle.net/7a7b5dwp/>

# Graphs/Networks

## Hive Plots



<http://jsfiddle.net/7a7b5dwp/>

<http://jsfiddle.net/eamonnmag/vso70qnr/>

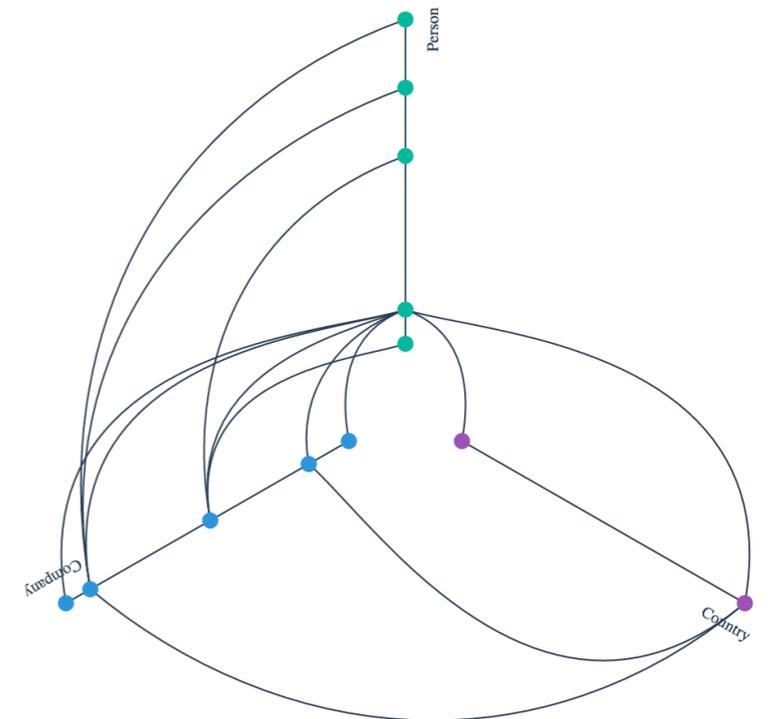
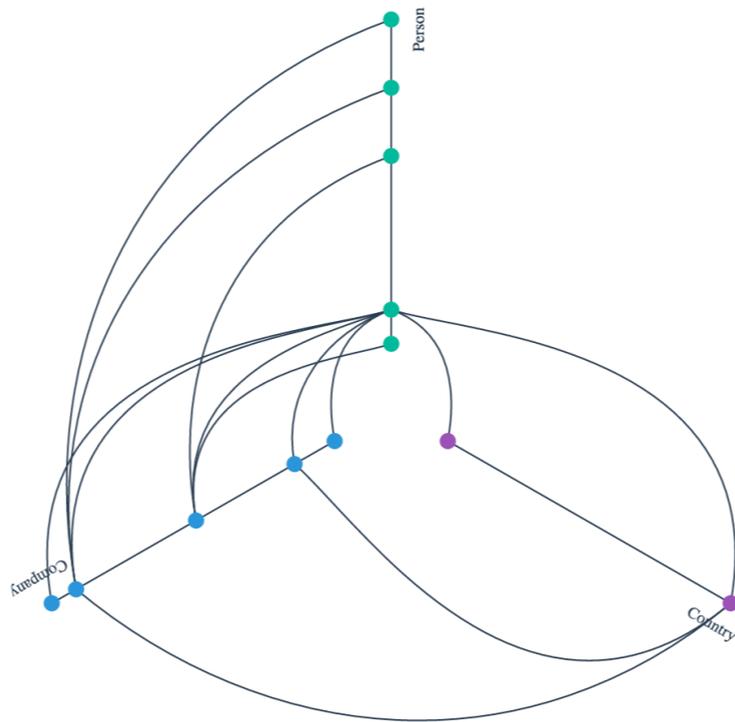
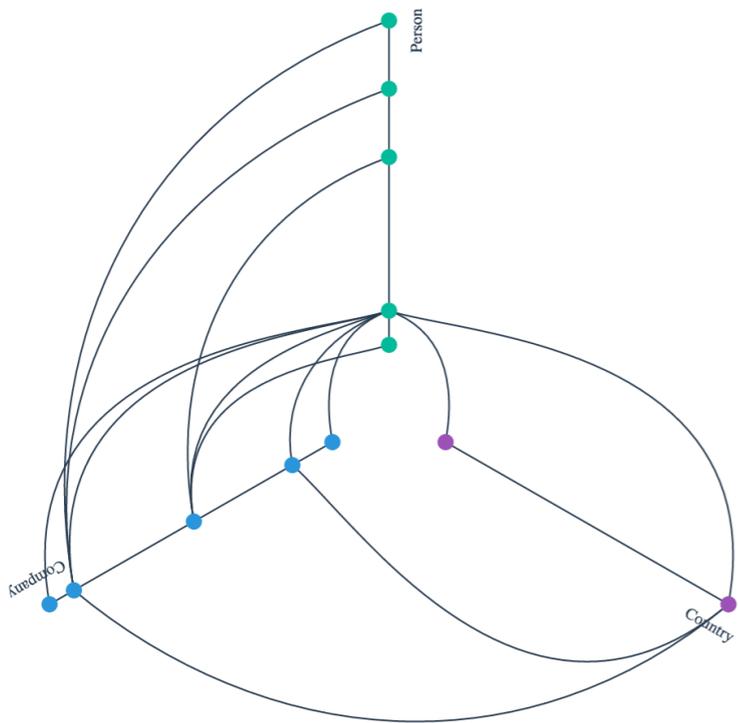
# Graphs/Networks

Hive Plots

**Same graph, same layouts.**

# Graphs/Networks

## Hive Plots

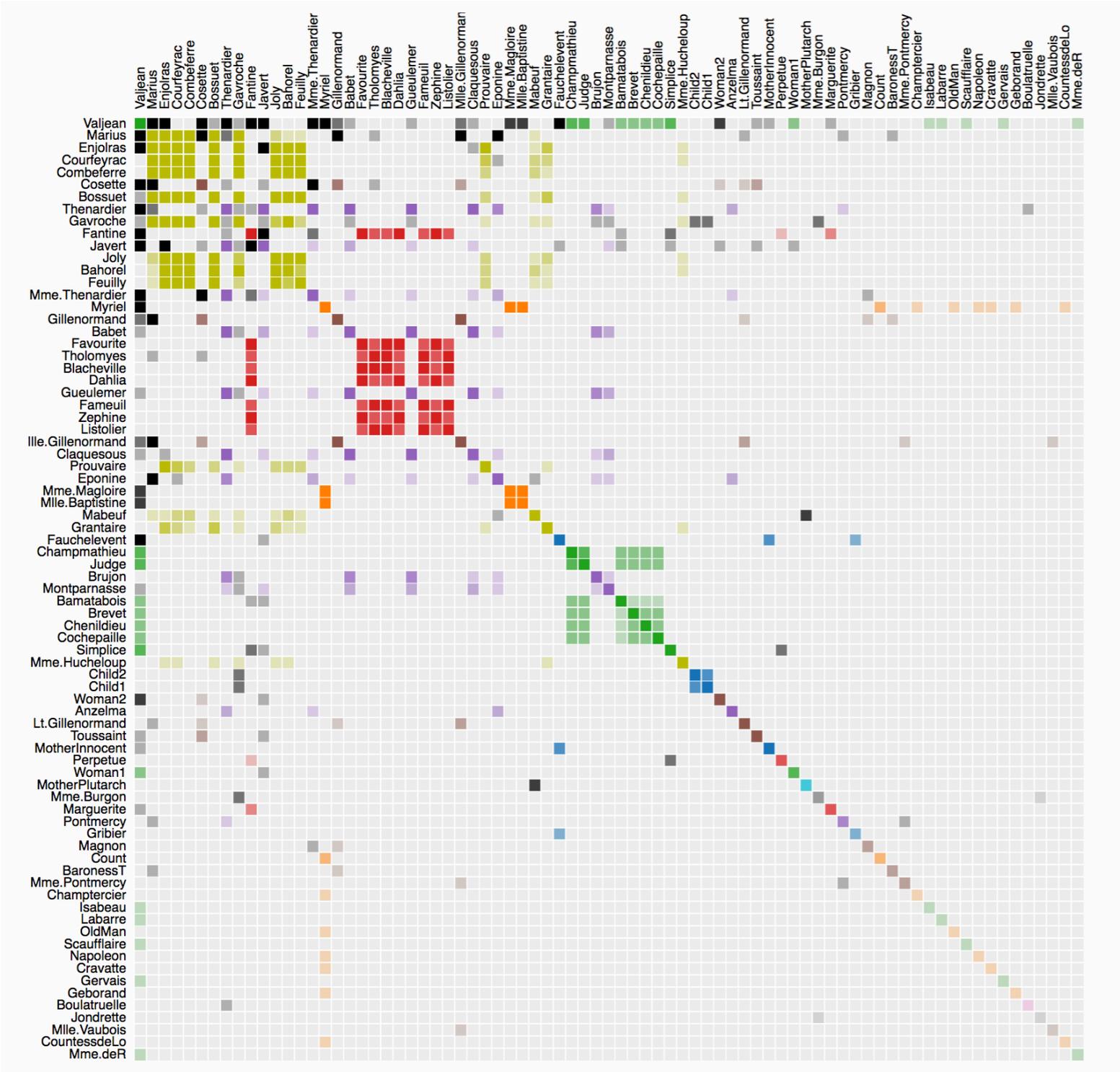


<http://jsfiddle.net/eamonnmag/vso70qnr/>

**Same graph, same layouts.**

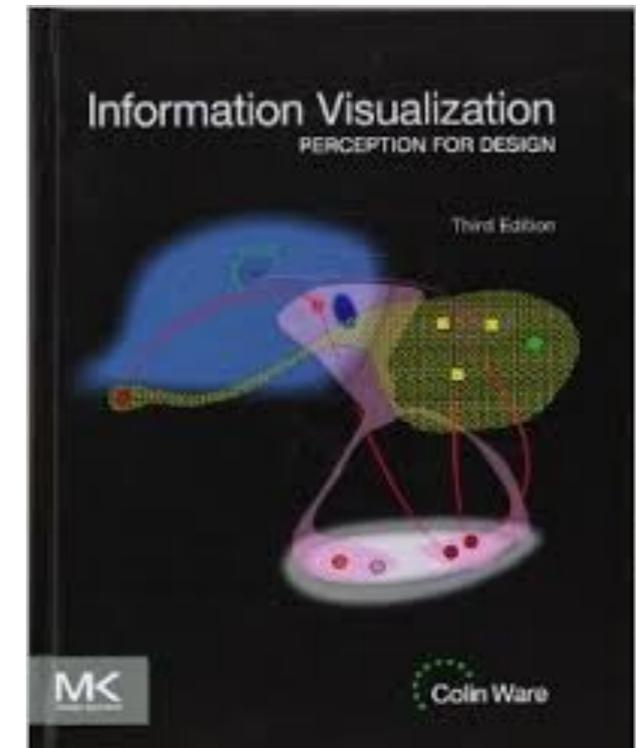
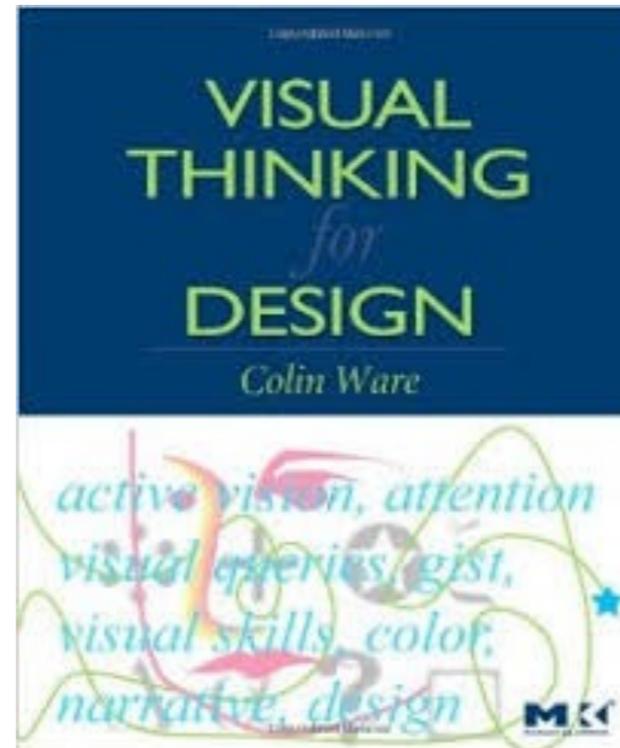
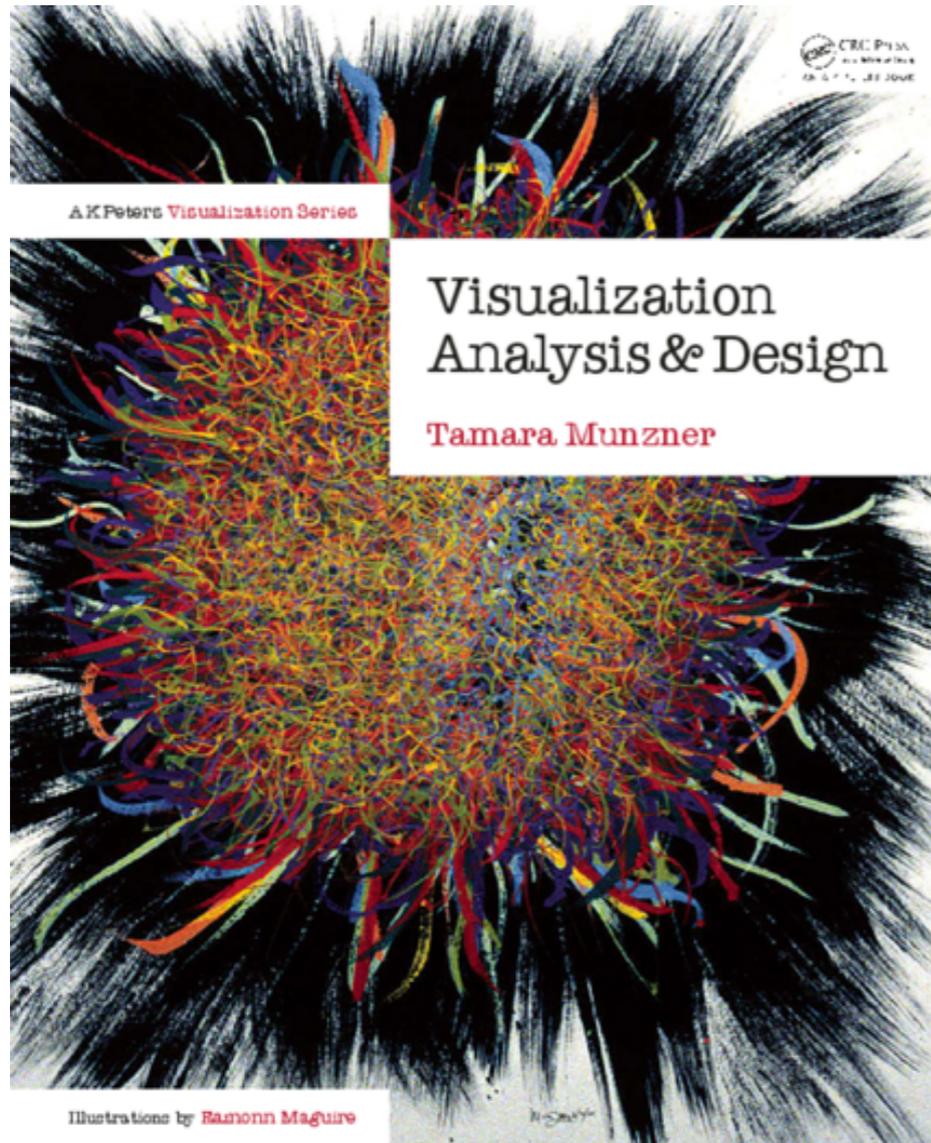
# Graphs/Networks

## Matrix Representations



<https://bost.ocks.org/mike/miserables/>

# More??



Visualization Analysis and Design.  
Munzner. A K Peters Visualization Series, CRC Press, Visualization Series, 2014.

# Further Links

## Tutorials

**D3** <http://antarctic-design.co.uk/biovis-workshop15/>

**Dashboards** <https://thor-project.github.io/dashboard-tutorial/>

## D3 Examples

<https://blockbuilder.org/search>

## Visualization Sites

Set Visualization - <http://www.cvast.tuwien.ac.at/SetViz>

Time Series Visualization - <http://survey.timeviz.net/>

<http://flowingdata.com/>

[Data Vis Catalogue](#)

## Python Data Vis Tools

[Pandas Data Vis](#)

Matplotlib

Seaborn

Altair



# Questions

[@antarcticdesign](#)  
[eamonnmag@gmail.com](mailto:eamonnmag@gmail.com)