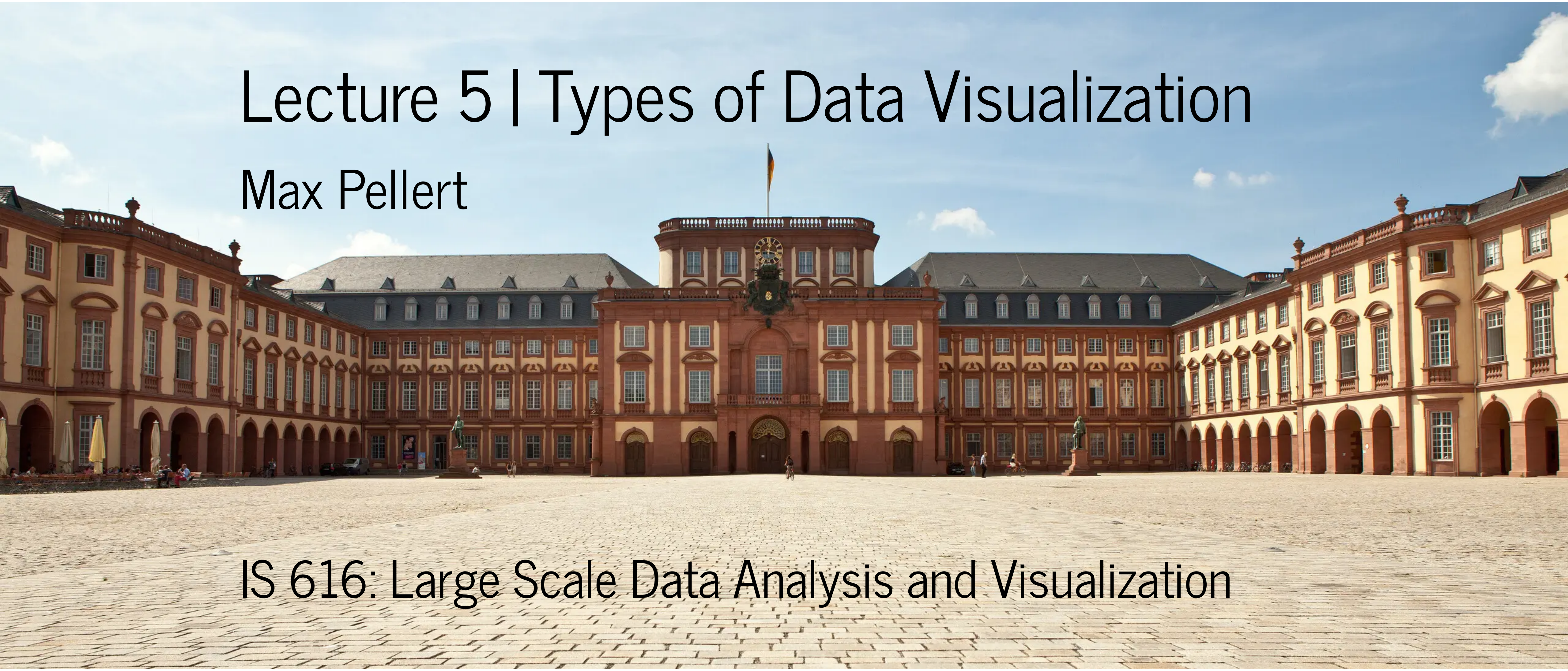


Lecture 5 | Types of Data Visualization

Max Pellert

IS 616: Large Scale Data Analysis and Visualization

UNIVERSITY
OF MANNHEIM

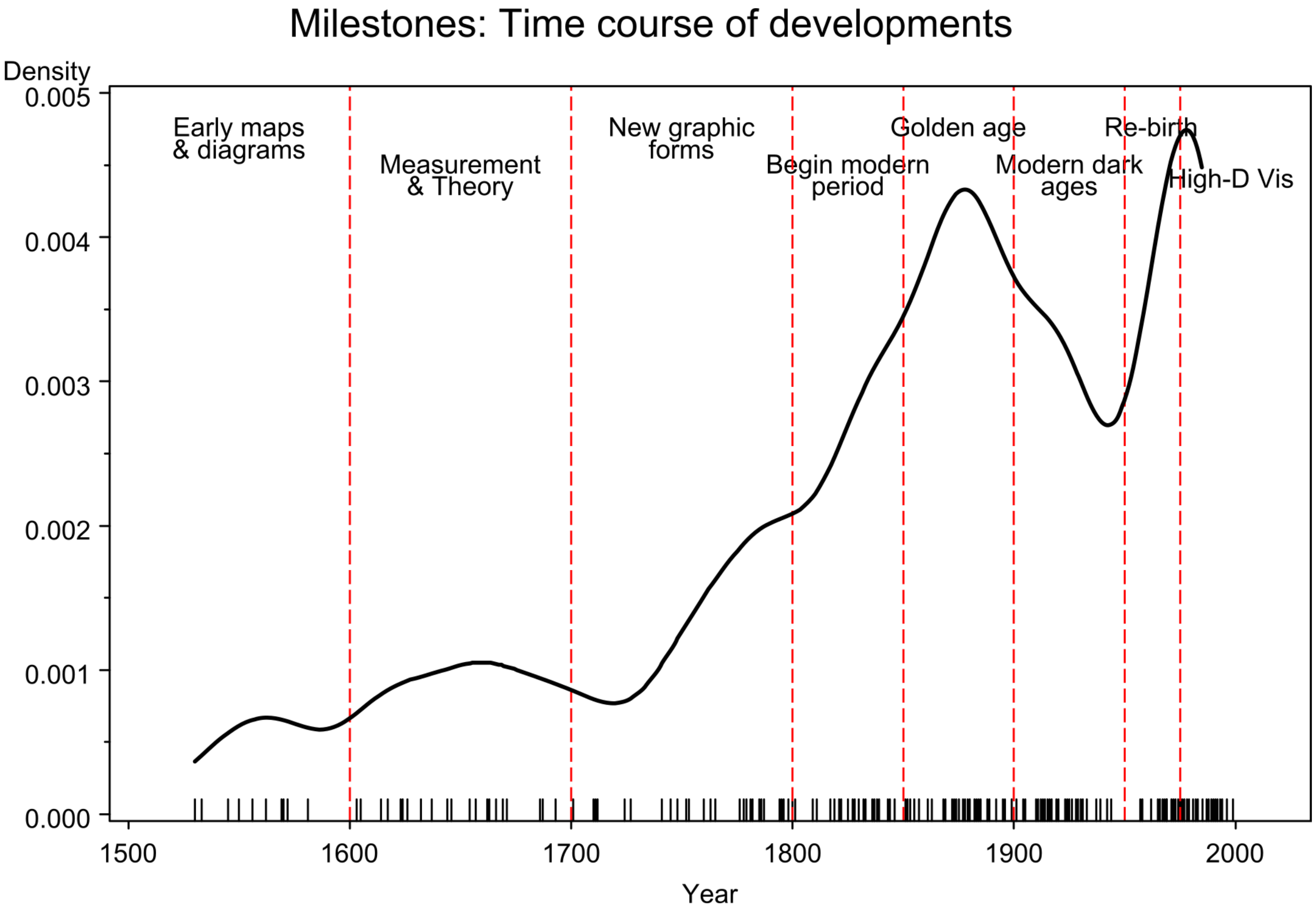


Figure 1.1. Time distribution of events considered milestones in the history of data visualization, shown by a rug plot and density estimate

The beginnings

“The earliest seeds of visualization arose in geometric diagrams, in tables of the positions of stars and other celestial bodies, and in the making of maps to aid in navigation and exploration.”

“The idea of coordinates was used by ancient Egyptian surveyors in laying out towns, earthly and heavenly positions were located by something akin to latitude and longitude by at least 200 B.C.,

The beginnings

and the map projection of a spherical earth into latitude and longitude by Claudius Ptolemy [c. 85–c. 165] in Alexandria would serve as reference standards until the 14th century."

Friendly, M. (2008). A Brief History of Data Visualization. In C. Chen, W. Härdle, & A. Unwin, Handbook of Data Visualization (pp. 15–56). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-33037-0_2

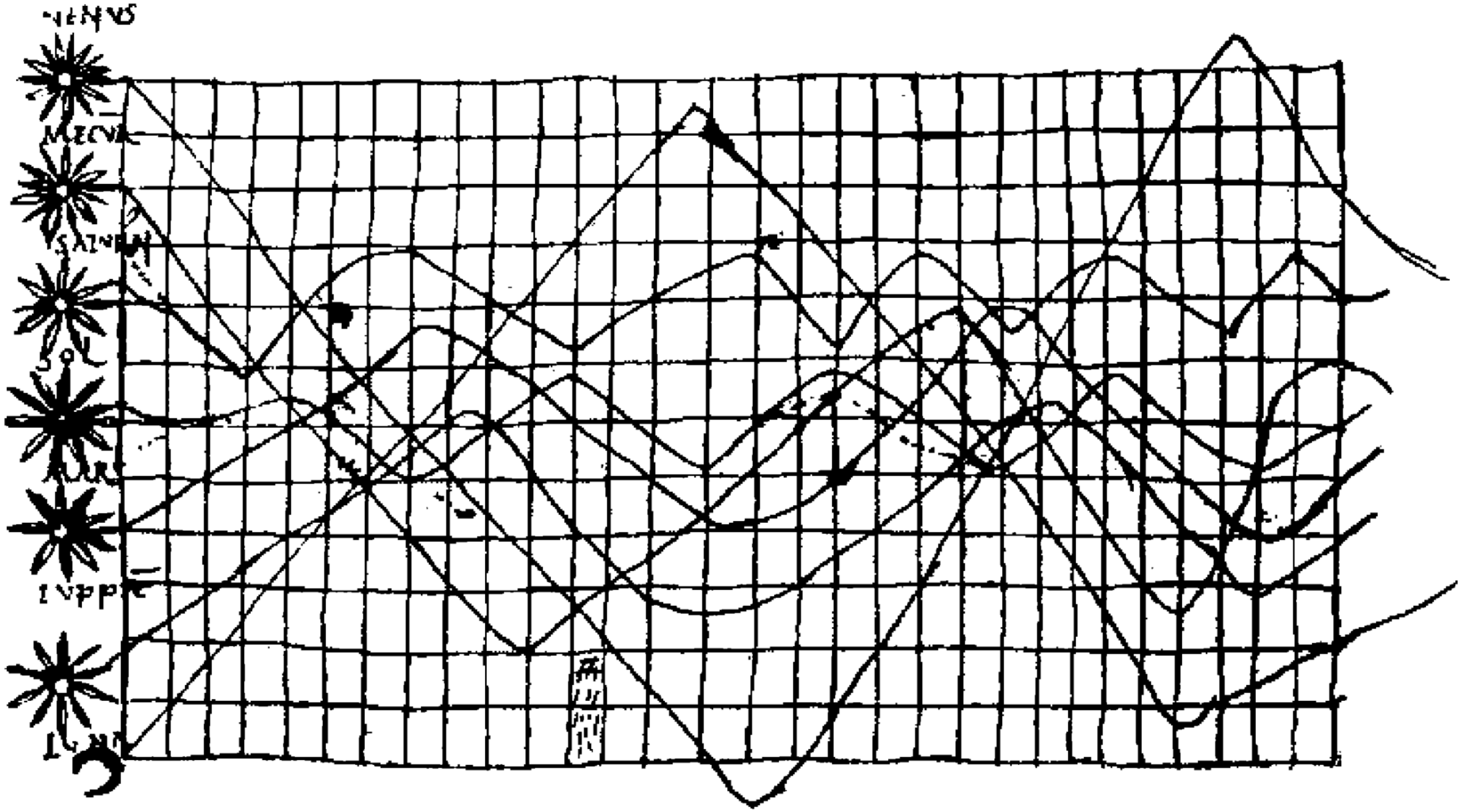


Figure 1.2. Planetary movements shown as cyclic inclinations over time, by an unknown astronomer, appearing in a 10th-century appendix to commentaries by A.T. Macrobius on Cicero's *In Somnium Sciponis*. *Source*: Funkhouser (1936, p. 261)

May 13, 2010
Volume 8, issue 5

PDF

A Tour through the Visualization Zoo

A survey of powerful visualization techniques, from the obvious to the obscure

Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky, Stanford University

Thanks to advances in sensing, networking, and data management, our society is producing digital information at an astonishing rate. According to one estimate, in 2010 alone we will generate 1,200 exabytes—60 million times the content of the Library of Congress. Within this deluge of data lies a wealth of valuable information on how we conduct our businesses, governments, and personal lives. To put the information to good use, we must find ways to explore, relate, and communicate the data meaningfully.

A Tour through the Visualization Zoo: A survey of powerful visualization techniques, from the obvious to the obscure

Authors: Jeffrey Heer, Michael Bostock, Vadim Ogievetsky Authors Info & Claims

Queue, Volume 8, Issue 5 • May 2010 • pp 20–30 • https://doi.org/10.1145/1794514.1805128

Published: 01 May 2010 Publication History

Check for updates

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A Tour through the Visualization Zoo: A survey of powerful visualization techniques, from the obvious to the obscure. Queue, 8(5), 20–30. https://doi.org/10.1145/1794514.1805128

The goal of visualization is to aid our understanding of data by leveraging the human visual system's highly tuned ability to see patterns, spot trends, and identify outliers. Well-designed visual representations can replace cognitive calculations with simple perceptual inferences and improve comprehension, memory, and decision making. By making data more accessible and appealing, visual representations may also help engage more diverse audiences in exploration and analysis. The challenge is to create effective and engaging visualizations that are appropriate to the data.

"Creating a visualization requires a number of nuanced judgments."

"One must determine which questions to ask, identify the appropriate data, and select effective visual encodings to map data values to graphical features such as position, size, shape, and color."

"The challenge is that for any given data set the number of visual encodings—and thus the space of possible visualization designs—is extremely large."

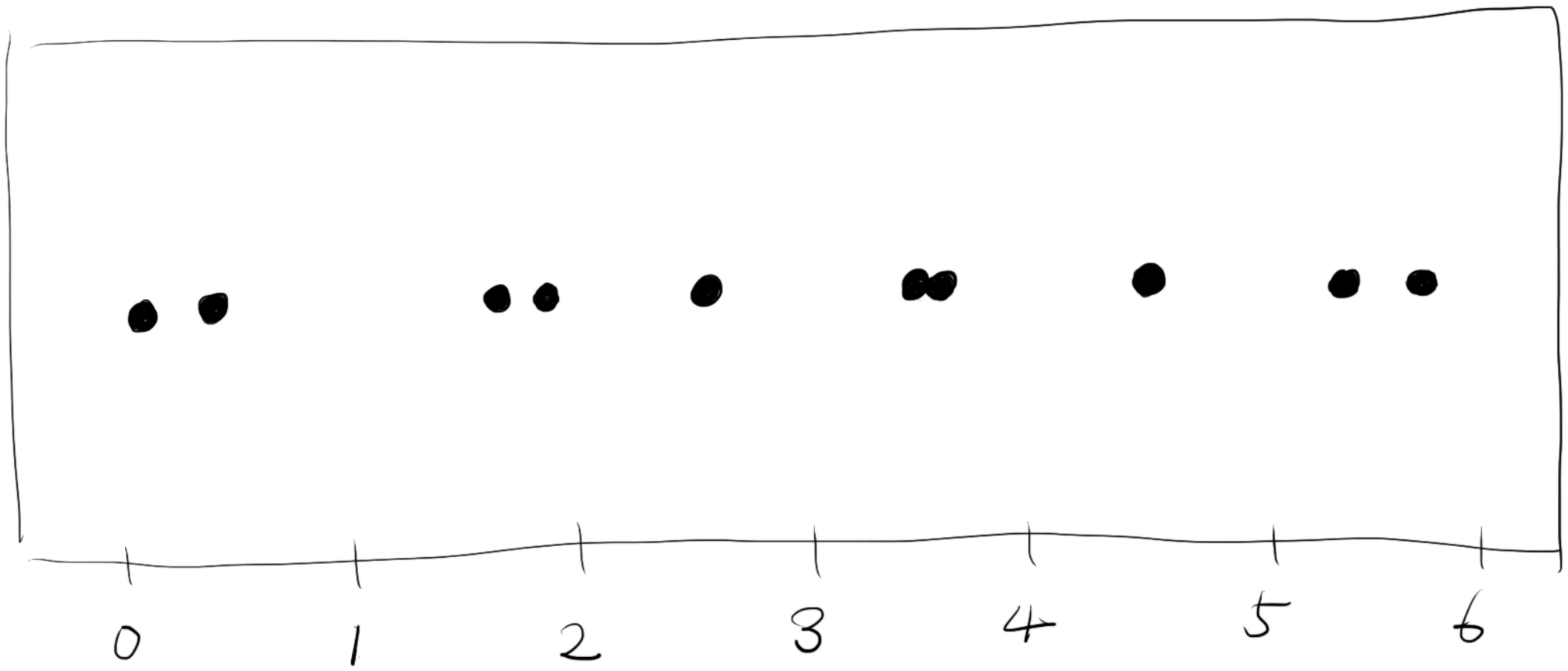
1D data

Person	Income
1	20.000 US$
2	150.000 US$
3	40.000 US$
4	55.000 US$
...	...

If you have for example 10 data points, what would be the most direct way to **visualize** this?*

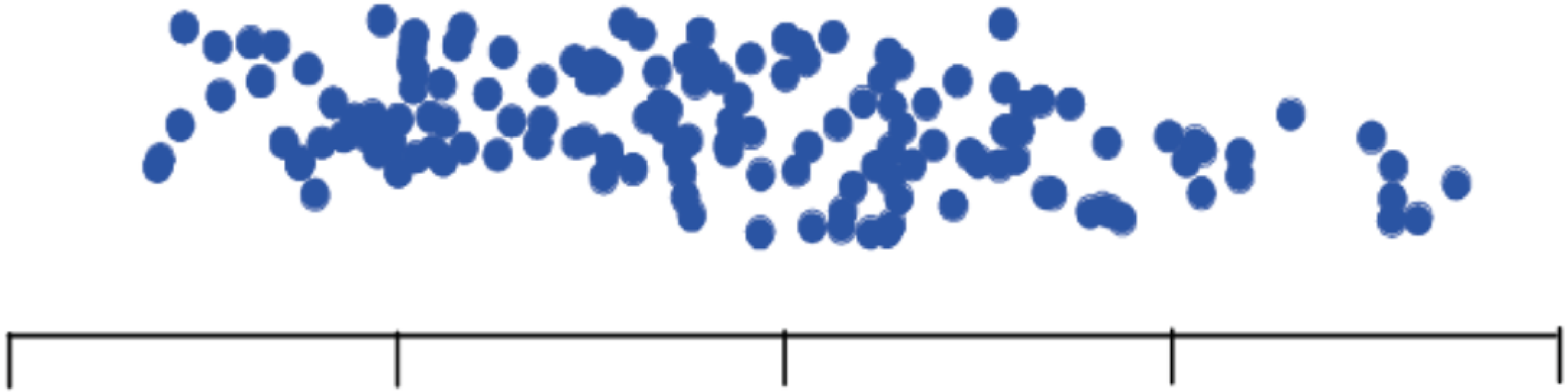
UNIVERSITY OF MANNHEIM

1D Scatterplot or “strip chart”

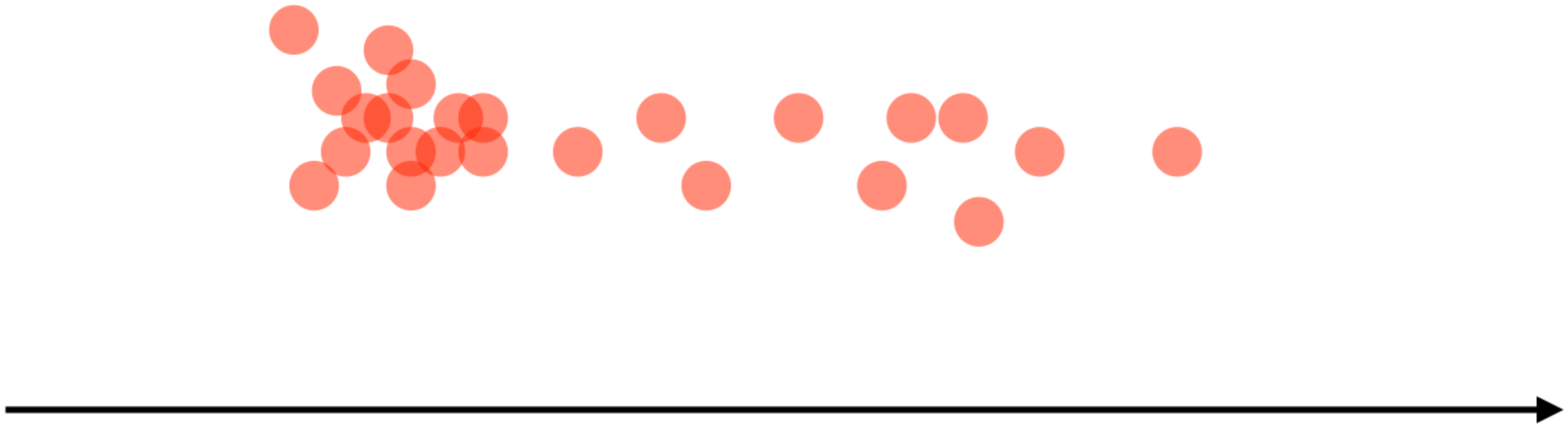


Problems?

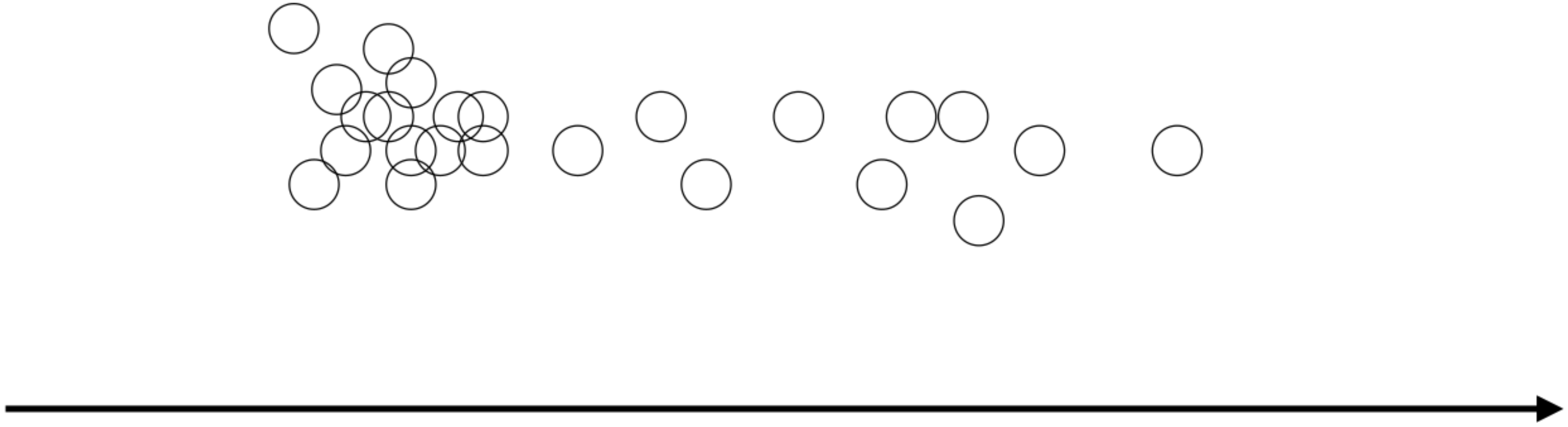
1D “Jittered” Scatterplot



Using transparency (“alpha”)



Using empty symbols such as rings



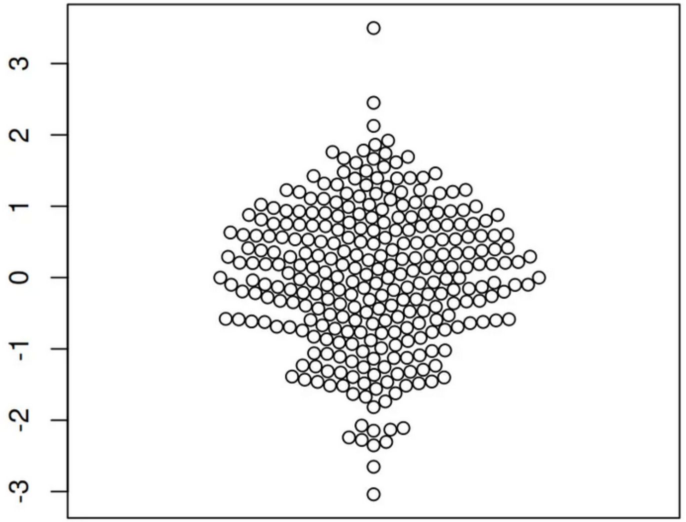
Basic beeswarm plot

The R `beeswarm` package contains a function of the same name that allows creating this type of plot. You need to input a numeric vector, a data frame or a list of numeric vectors.

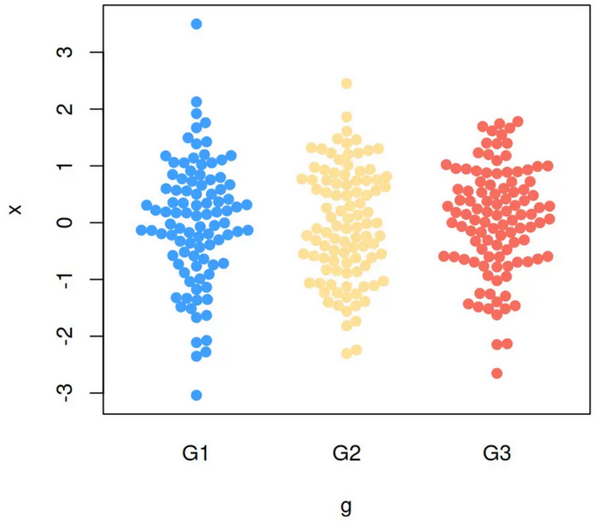
```r
# install.packages("beeswarm")
library(beeswarm)

# Data generation
set.seed(1995)
x <- rnorm(300)

# Bee swarm plot
beeswarm(x)
```



https://r-charts.com/distribution/beeswarm/




```r
# install.packages("beeswarm")
library(beeswarm)

# Data generation
set.seed(1995)
x <- rnorm(300)
g <- sample(c("G1", "G2", "G3"),
            size = 300, replace = TRUE)

# Bee swarm plot by group
beeswarm(x ~ g,
         pch = 19,
         col = c("#3FA0FF", "#FFE099", "#F7
```

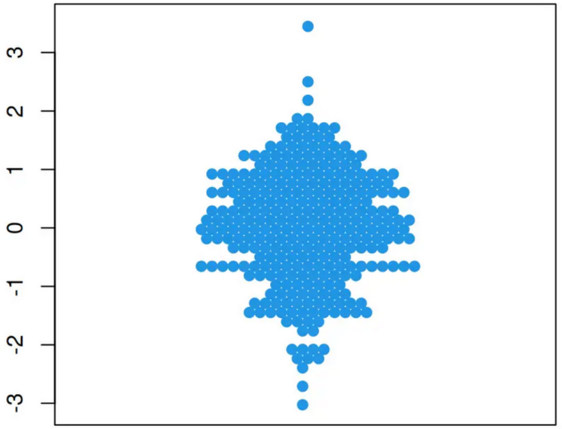
"hex" method

This method uses a hexagonal grid to place the data points.

```r
# install.packages("beeswarm")
library(beeswarm)

# Data generation
set.seed(1995)
x <- rnorm(300)

# hex method
beeswarm(x, col = 4, pch = 19,
         method = "hex")
```

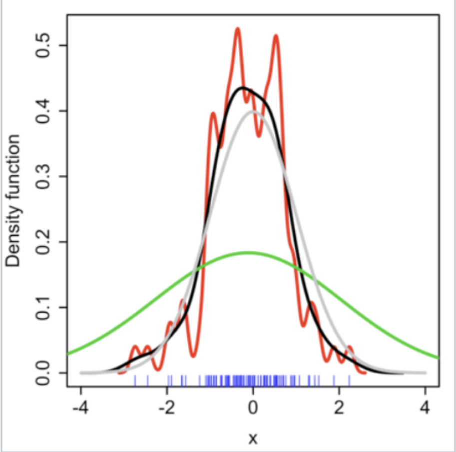


UNIVERSITY OF MANNHEIM

Not to be confused with Carpet plot.

A **rug plot** is a plot of data for a single quantitative variable, displayed as marks along an axis. It is used to visualise the distribution of the data. As such it is analogous to a histogram with zero-width bins, or a one-dimensional scatter plot.

Rug plots are often used in combination with two-dimensional scatter plots by placing a rug plot of the x values of the data along the x-axis, and similarly for the y values. This is the origin of the term "rug plot", as these rug plots with perpendicular markers look like tassels along the edges of the rectangular "rug" of the scatter plot.



A rug plot of 100 data points appears in blue along the x-axis. (The points are sampled from the normal distribution shown in gray. The other curves show various kernel density estimates of the data.)

External links [edit]

- Rug plots in R
- Rug plots in Matlab
- Rug plots in Python using the Seaborn library

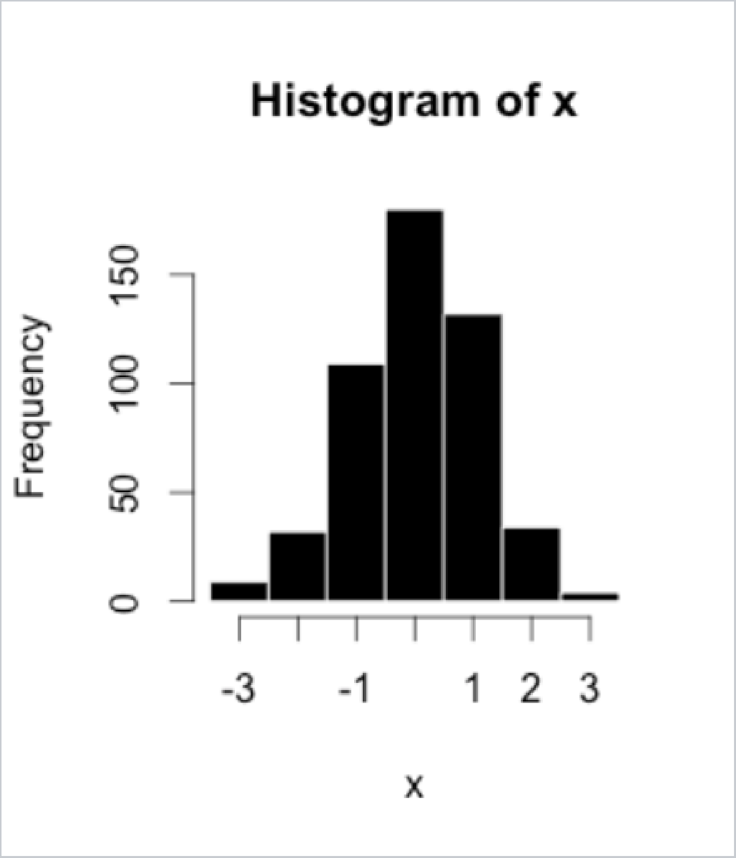
https://en.wikipedia.org/wiki/Rug_plot

With a lot of data, we may need to aggregate or summarize not to be overwhelmed by the mass of single data points

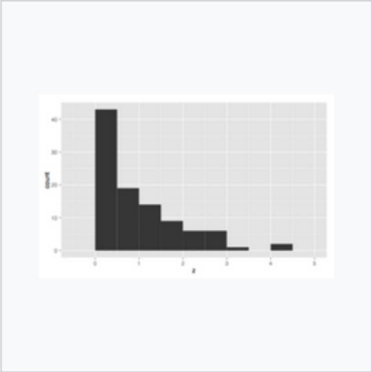
Histograms

show the prevalence of values grouped into bins

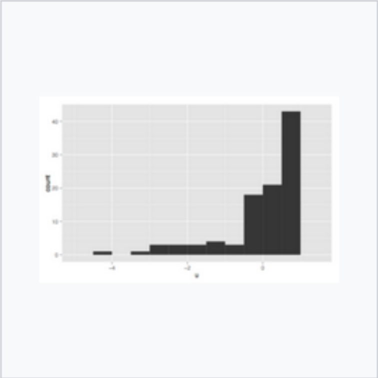
Bin/Interval	Count/Frequency
−3.5 to −2.51	9
−2.5 to −1.51	32
−1.5 to −0.51	109
−0.5 to 0.49	180
0.5 to 1.49	132
1.5 to 2.49	34
2.5 to 3.49	4



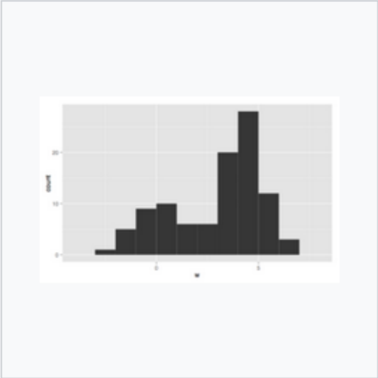
Symmetric, unimodal



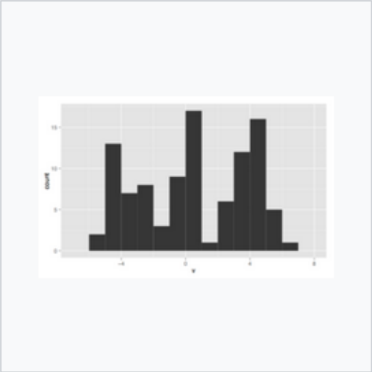
Skewed right



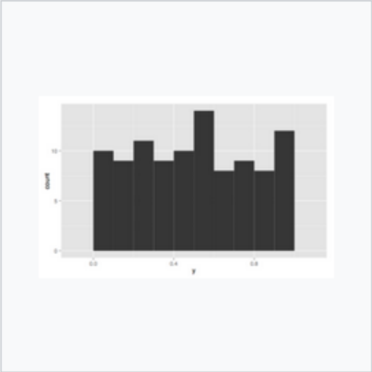
Skewed left



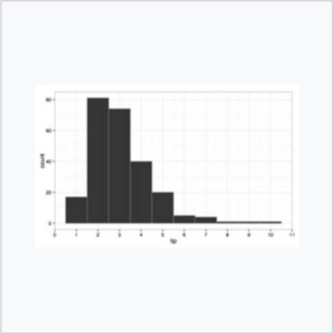
Bimodal



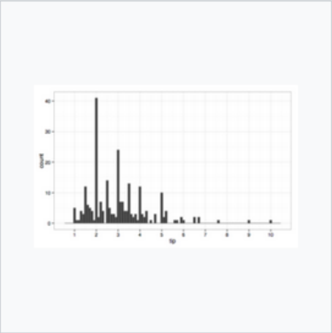
Multimodal



Symmetric

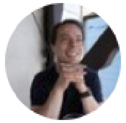


Tips using a $1 bin width, skewed right, unimodal



Tips using a 10c bin width, still skewed right, multimodal with modes at $ and 50c amounts, indicates rounding, also some outliers

Histograms can mislead



Nick Strayer
@NicholasStrayer

Histograms are fantastic, but make sure your bin-width/number is chosen well. This is the _exact_ same data, plotted with different bin-widths. Notice that the pattern doesn't necessarily get clearer as bin num increases. #dataviz

https://twitter.com/NicholasStrayer/status/1026893778404225024

http://nickstrayer.me/histogram_bins/

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


0:00

Boxplots

This refers to the box-and-whisker plot, which conveyes statistical features such as the mean, median, quartile boundaries or extreme outliers.

Wickham, H., & Stryjewski, L. (2012). 40 years of boxplots. had.co.nz.

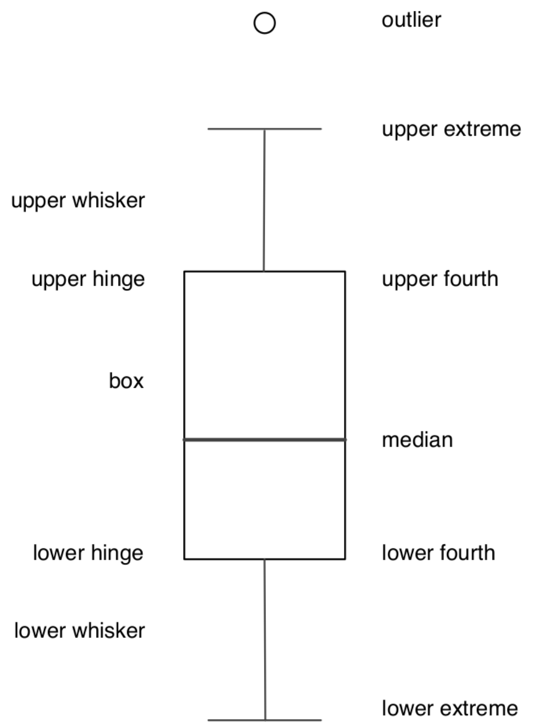


Figure 1: Construction of a boxplot. Labels on the left give names for graphic elements, labels on the right give the corresponding summary statistics.

Histogram vs. Boxplot

What are their strengths and weaknesses?

As a summarization method, a boxplot may be useful if you want to compare multiple (well-behaving) distributions. Boxplots will immediately and precisely show the median, the quartiles, and the rough range of the distribution.

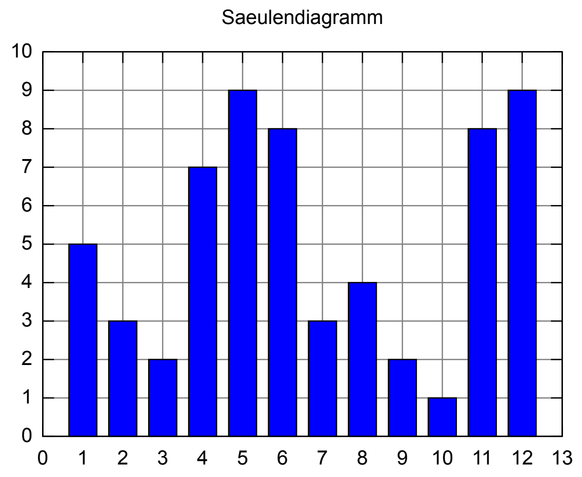
On the other hand, a boxplot may hide details in the distribution, particularly when the distribution is far from a normal distribution.

A histogram is sensitive to parameter choice as we have seen

Bar charts

Similar to histograms but the height of the bars must not necessarily be a count (or frequency) and the data can have "natural" categories not artificial bins

Rather, any (numeric) variable can be displayed



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

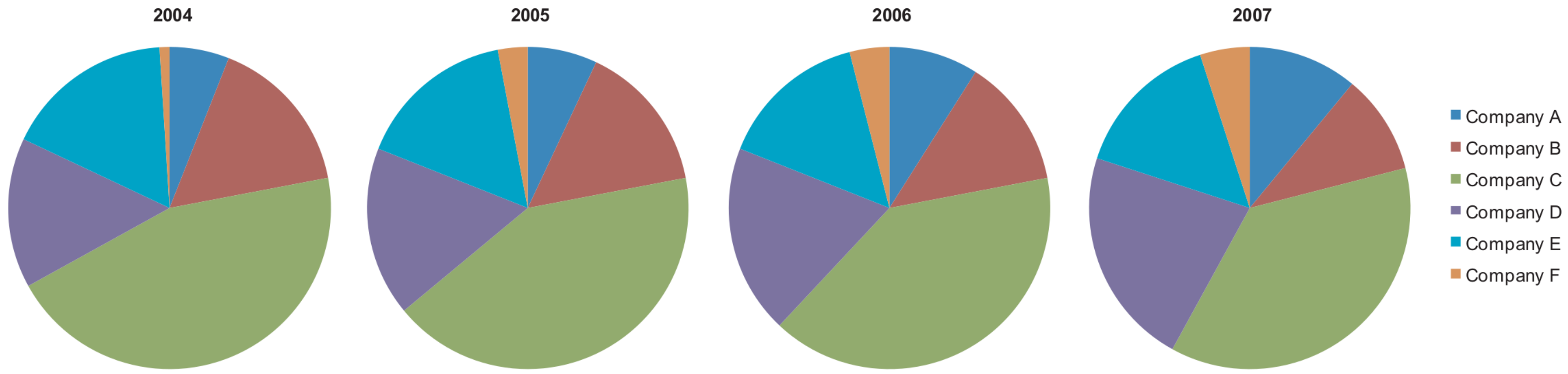
Pie charts

Similar to bar charts but the area is circle segments not bars

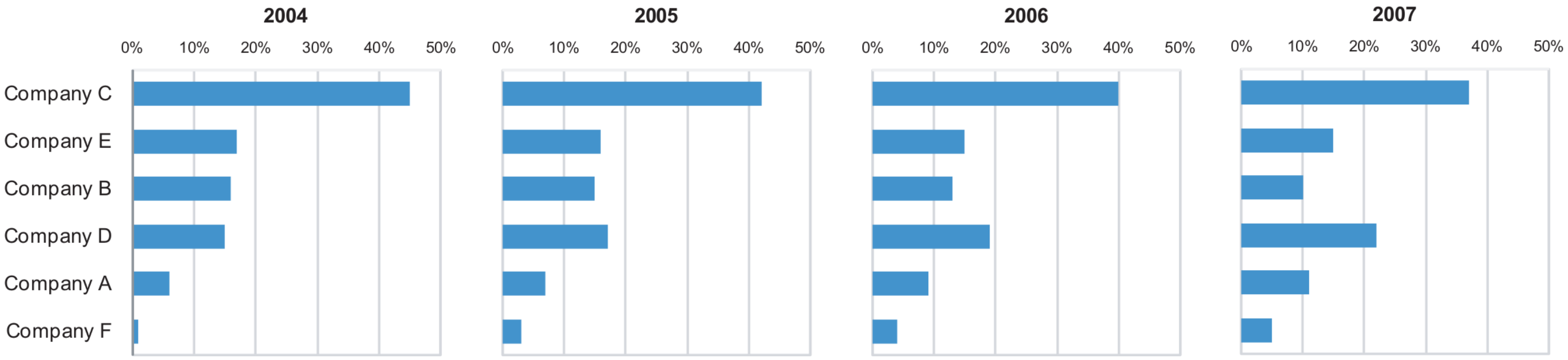
Have a very bad name, better not to use them not to trigger people (Few, S. (2007). Save the Pies for Dessert.)

Also good other reasons not to use them:

> Tables are preferable to graphics for many small data sets.[1] A table is nearly always better than a dumb pie chart; the only worse design than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies, as in this heavily encoded example from an atlas. Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used.[2]



Try to follow the changes of these various companies and how they compare to one another through time. It is nearly impossible. Notice how easily you can do it, however, using the following display:



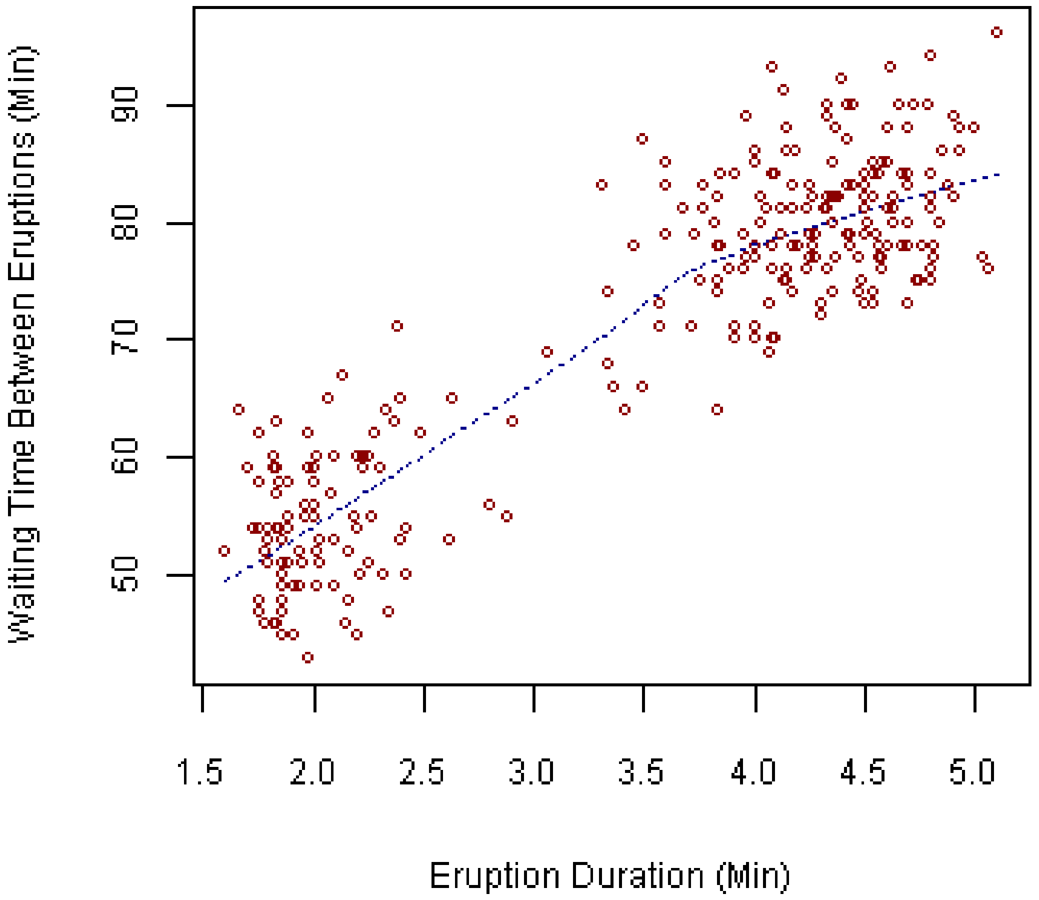
2D Scatter plots

So far, we were more or less only concerned with the x-axis

For example, the x-axis was set by the histogram bins or more general by groups or categories in the bar chart

If we relax this to plot arbitrary (numeric) variables on the x **and** y-axis, we get 2D scatter plots

Old Faithful Eruptions



2D Scatter plots

"Waiting time between eruptions and the duration of the eruption for the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA.

This chart suggests there are generally two types of eruptions: short-wait-short-duration, and long-wait-long-duration."

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

2D Scatter plots

Very common, often the first thing you plot usually by using points

That makes a lot of sense, as it is often an “honest” strategy that reveals a lot

It can prevent you from overlooking things, which may be embarassing later (see Anscombe’s quartett)

Same problems with a lot of data points as the 1D scatter plot, similar strategies to tackle that for example with alpha or rings

The “visualization zoo”

Time series

"Values changing over time"

Like a scatter plot, but the x-axis is a time dimension now

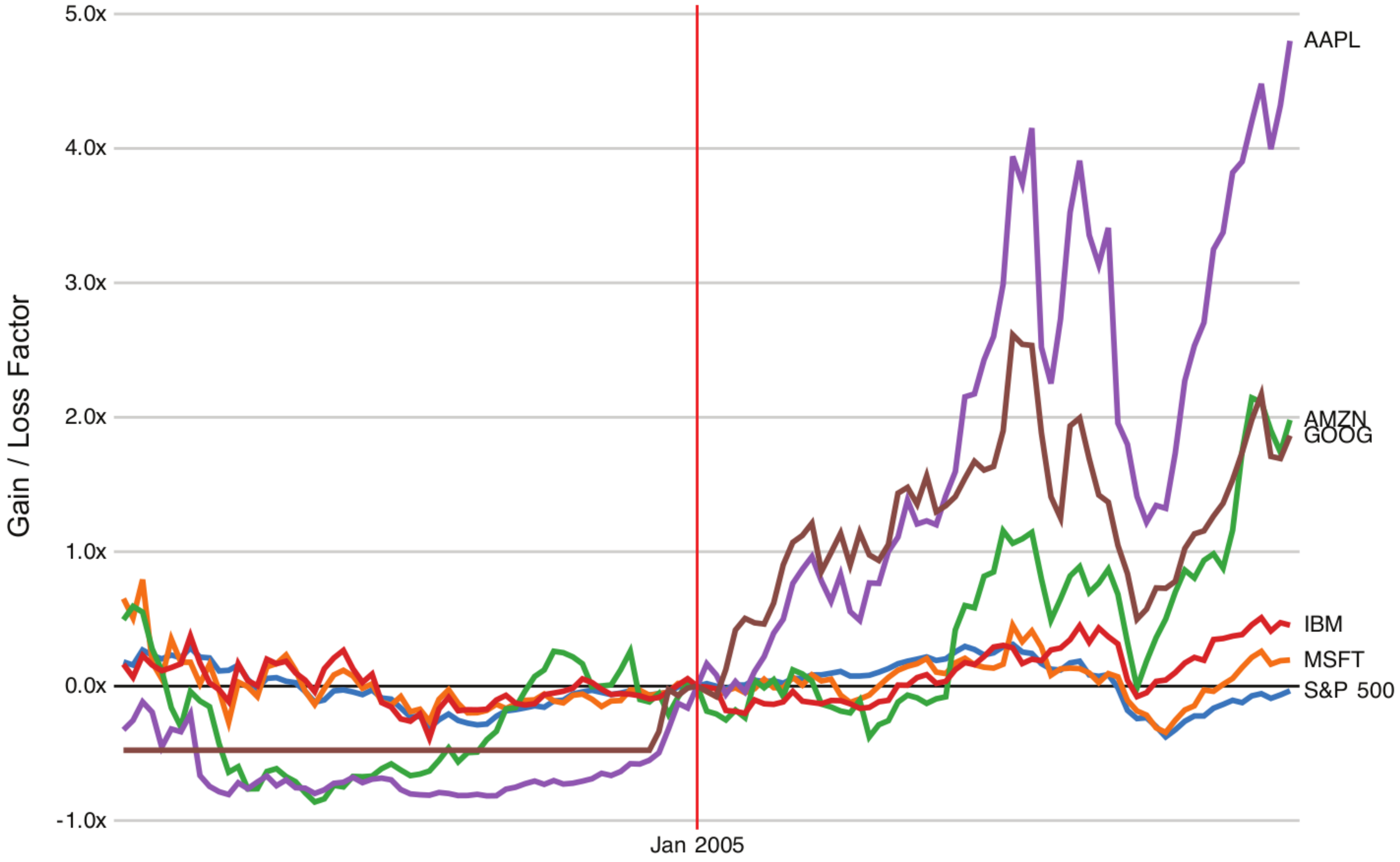
Often instead (or in addition) to points, lines are plotted

Raw values are often less important than relative changes

Mutltiple lines can often only meaningfully compared when they are normalized in some way

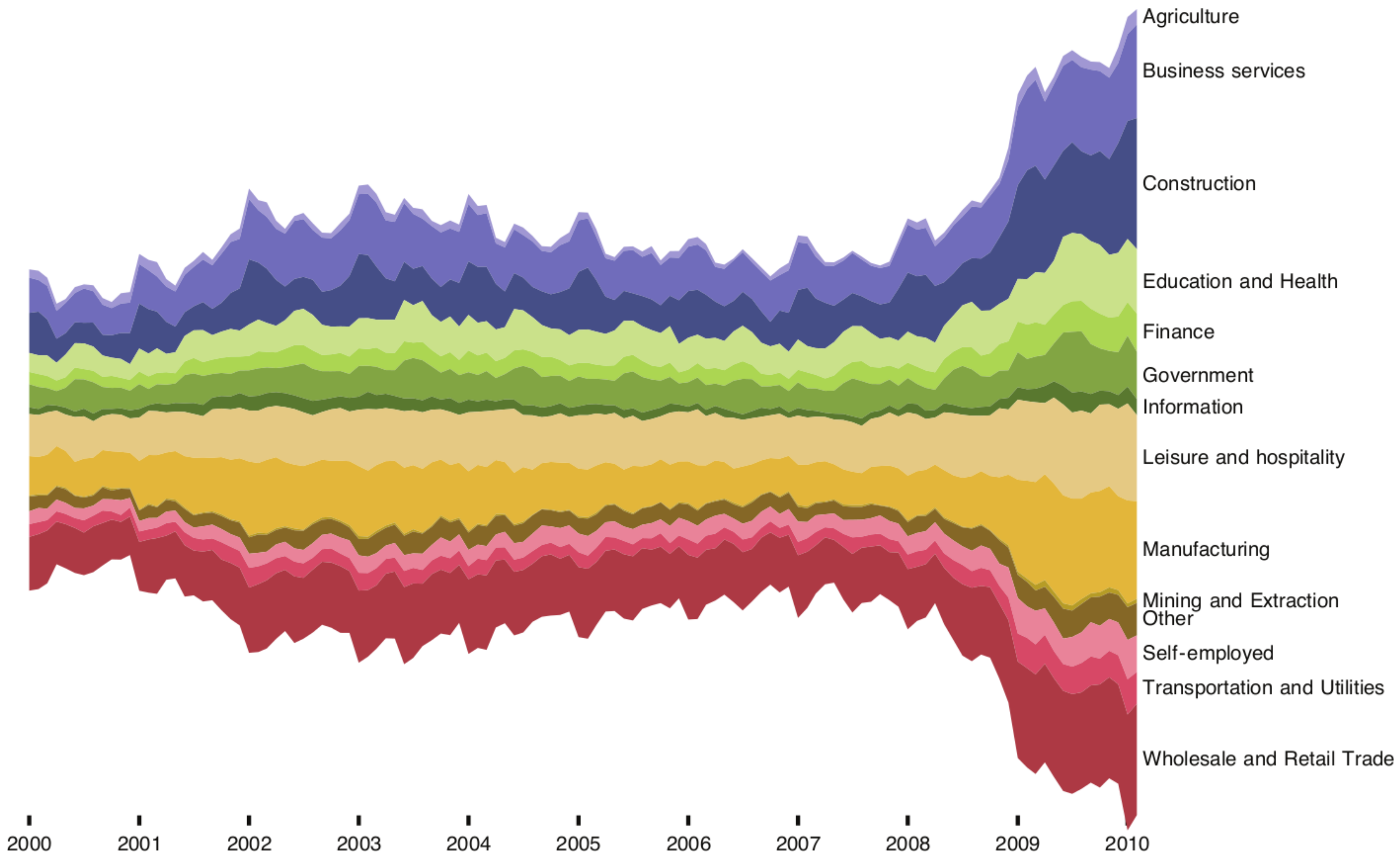
Multiple stocks may have totally different baseline prices for example

Index chart



5.0x
4.0x
3.0x
2.0x
1.0x
0.0x
-1.0x
Gain / Loss Factor
Jan 2005
AAPL
AMZN
GOOG
IBM
MSFT
S&P 500

Stacked graphs



Stacked graphs

Stacked Graph of Unemployed U.S. Workers by Industry, 2000-2010

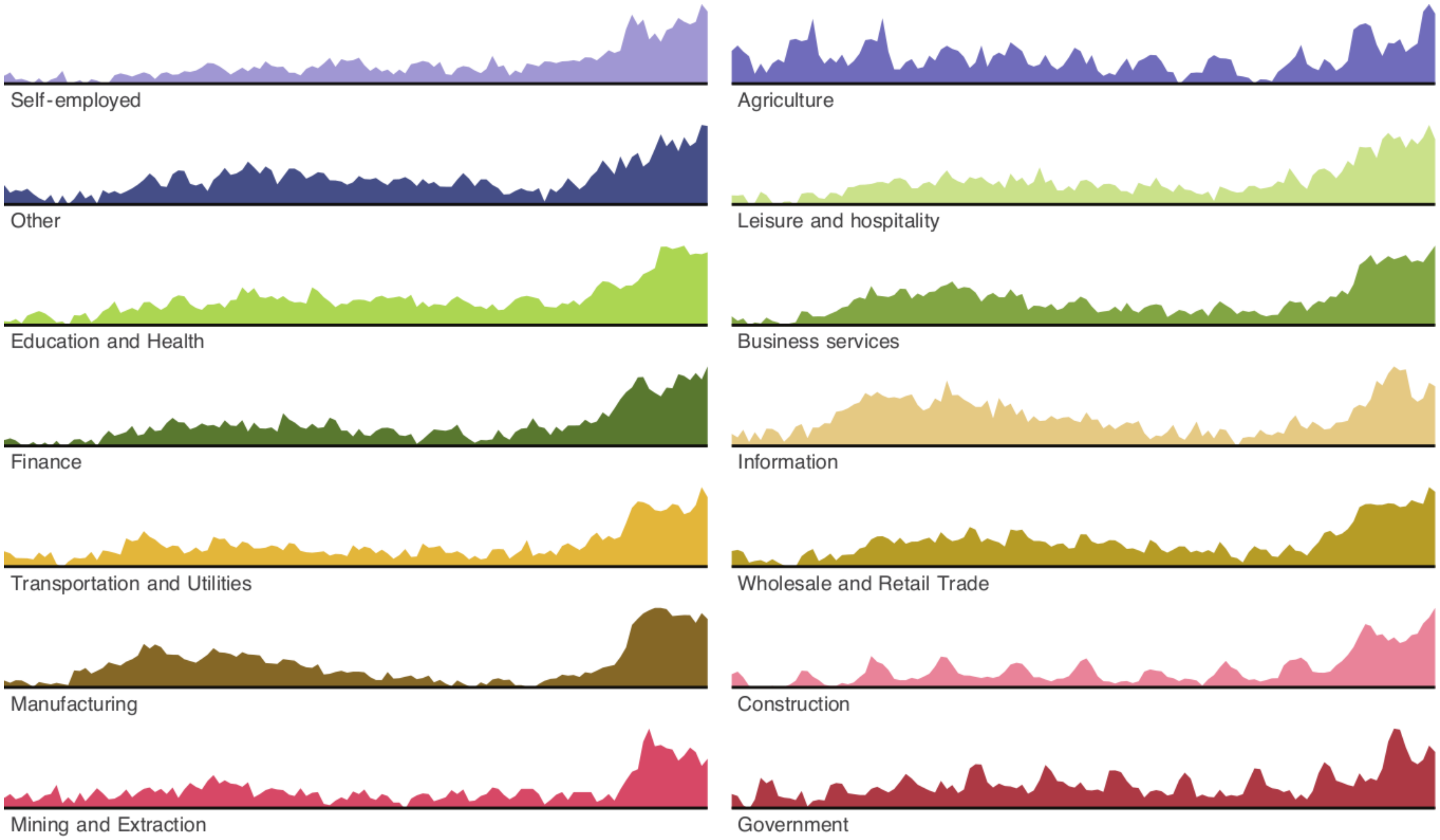
By stacking area charts on top of each other, we arrive at a visual summation of time-series values

Also called “stream graph”

Some limitations:

A stacked graph does not support negative numbers and is meaningless for data that should not be summed (temperatures, for example)

Small multiples instead



UNIVERSITY OF MANNHEIM

Horizon graph

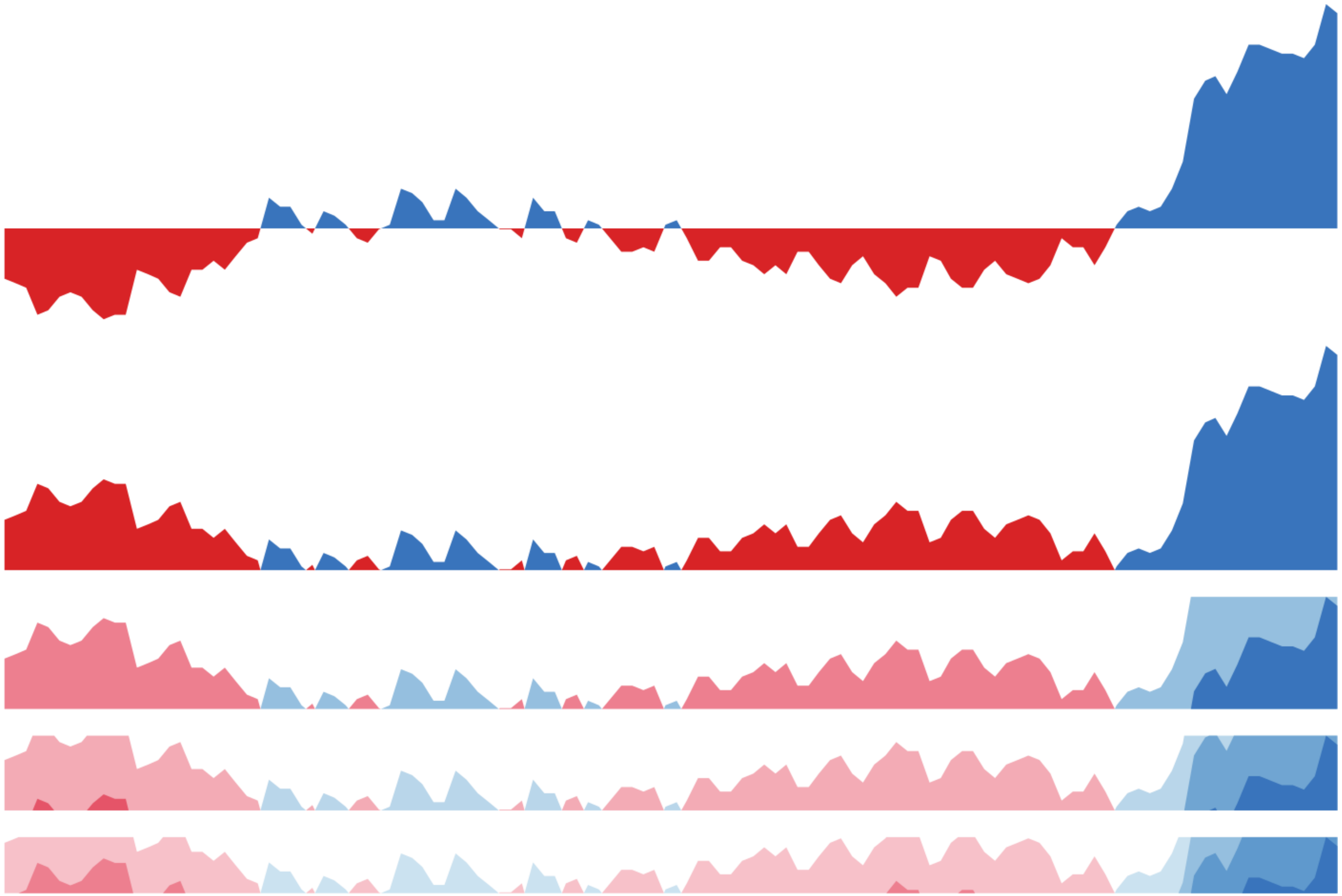
We start with standard area chart, with positive values colored blue and negative values colored red

“The horizon graph is a technique for increasing the **data density** of a time-series view while preserving resolution.”

We divide the graph into horizontal bands and layer them to create a nested form.

The result is a chart that preserves data resolution but uses only a quarter of the space.

Horizon graph



Statistical Distributions

Often, we want to do exploratory data analysis:

To gain insight into how data is distributed to inform data transformation and modeling decisions

We already covered the histogram and the boxplot, but there are many more techniques

Stem-and-leaf plots

0	1 1 1 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 5 6 7 8 8 8 8 8 8 9
1	0 0 0 0 1 1 1 1 2 2 3 3 3 3 4 4 4 4 5 5 6 7 7 8 9 9 9 9 9
2	0 0 1 1 1 5 7 8 9
3	0 0 1 2 3 3 3 4 6 6 8 8
4	0 0 1 1 1 1 3 3 4 5 5 5 6 7 8 9
5	0 2 3 5 6 7 7 7 9
6	1 2 6 7 8 9 9 9
7	0 0 0 1 6 7 9
8	0 0 1 2 3 4 4 4 4 4 4 4 5 6 7 7 7 9
9	1 3 3 5 7 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
10	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Stem-and-leaf plots

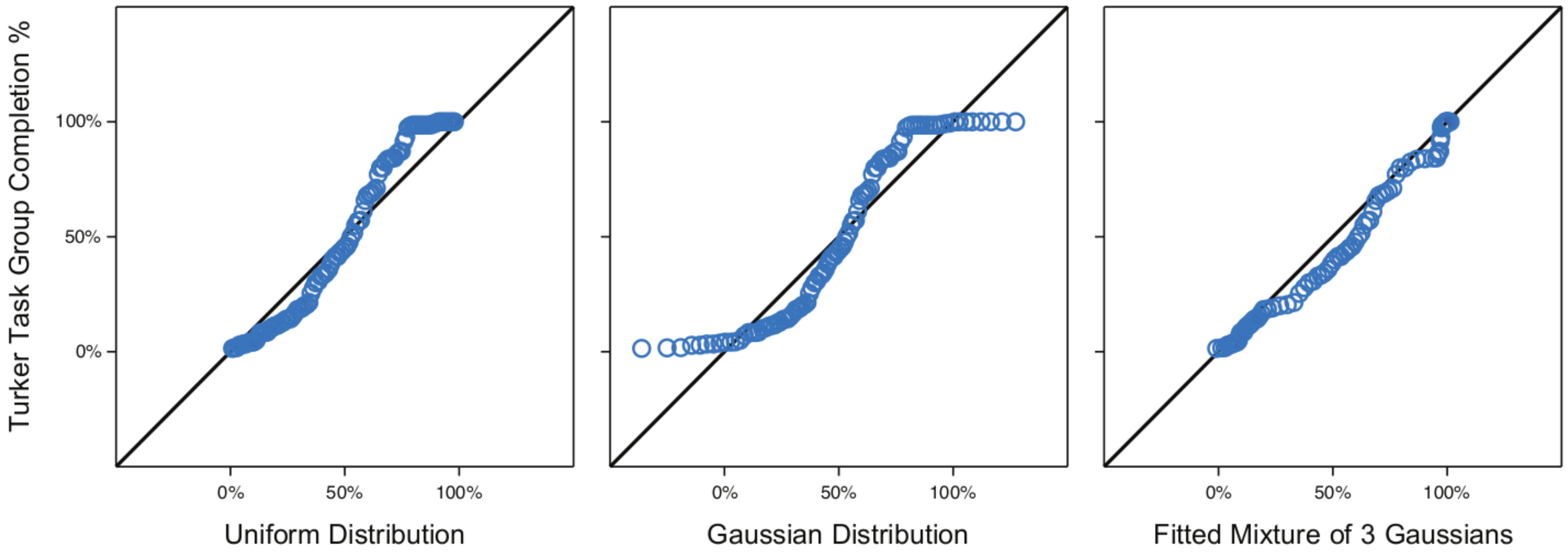
Stem-and-Leaf Plot of Mechanical Turk Participation Rates

It typically bins numbers according to the first significant digit and then stacks the values within each bin by the second significant digit.

This minimalistic representation uses the data itself to paint a frequency distribution,

replacing the “information-empty” bars of a traditional histogram bar chart and allowing one to assess both the overall distribution and the contents of each bin.

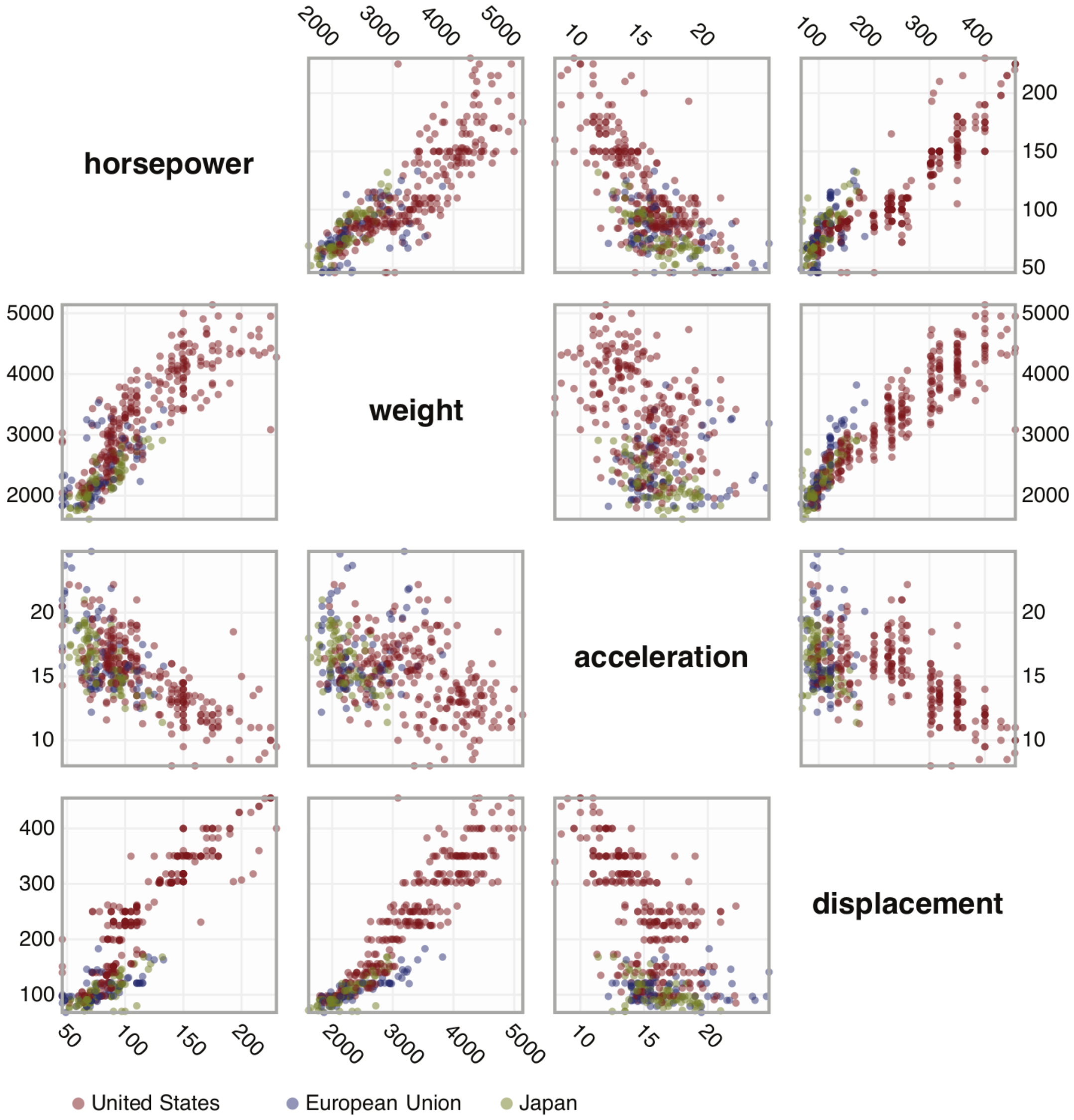
Q-Q plots



The Q-Q plot compares two probability distributions by graphing their quantiles

If the two are similar, the plotted values will lie roughly along the central diagonal

SPLOM (scatter plot matrix)



UNIVERSITY OF MANNHEIM

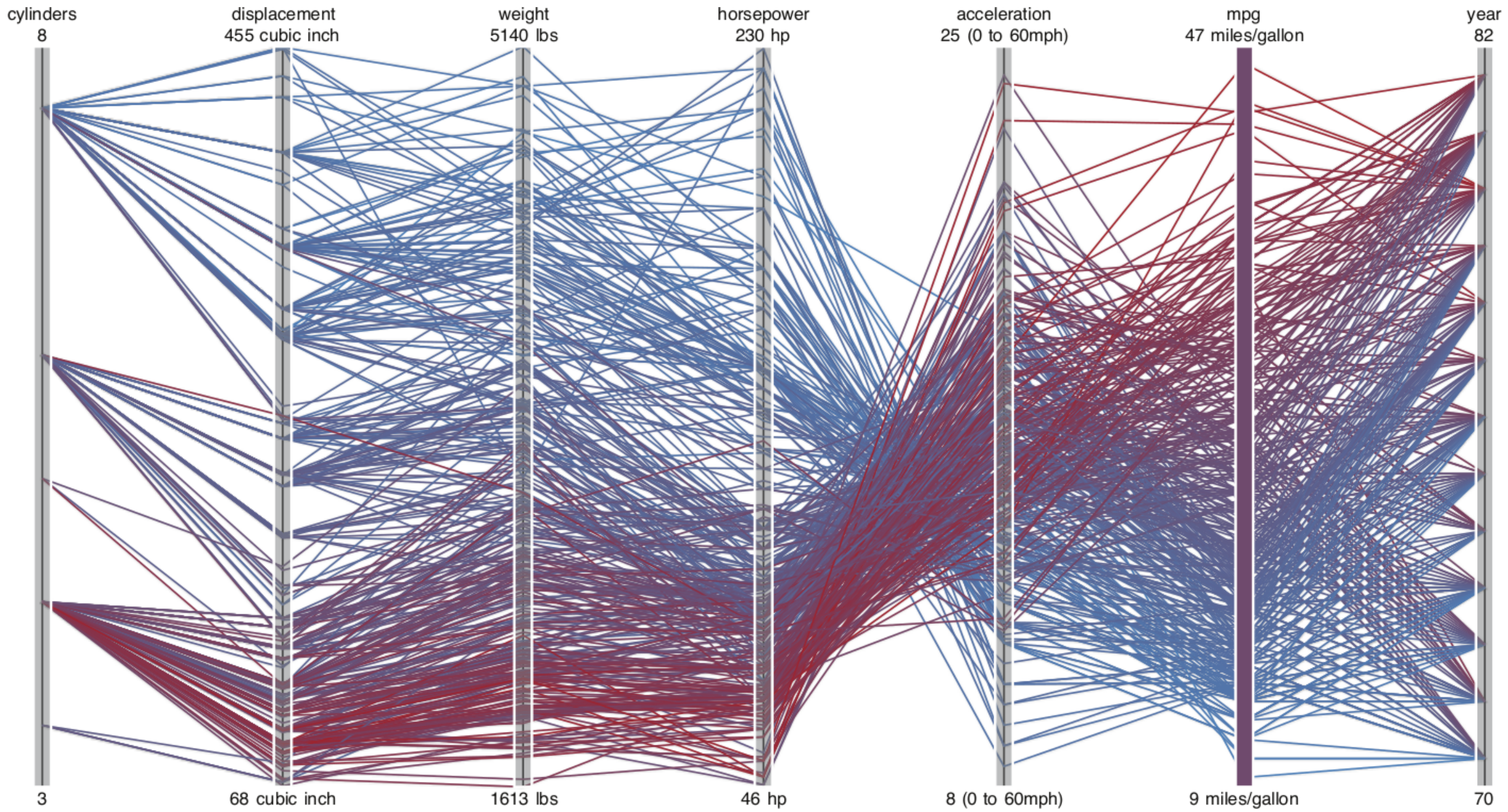
SPLOM (scatter plot matrix)

Scatter Plot Matrix of Automobile Data

Small multiples of scatter plots showing a set of pairwise relations among variables

A SPLOM enables visual inspection of correlations between any pair of variables.

Parallel coordinates



Parallel coordinates

Parallel coordinates take a different approach to visualizing multivariate data in a more compact way

Instead of graphing every pair of variables in two dimensions, we repeatedly plot the data on parallel axes and then connect the corresponding points with lines

Each line represents a single row in the database

Line crossings between dimensions often indicate inverse correlation

Reordering dimensions can aid pattern finding

*Do you need a graphic at all?

UNIVERSITY OF MANNHEIM

The conventional sentence is a poor way to show more than two numbers because it prevents comparisons within the data.

The linearly organized flow of words, folded over at arbitrary points (decided not by content but by the happenstance of column width), offers less than one effective dimension for organizing the data. Instead of:

> Nearly 53 percent of the type A group did something or other compared to 46 percent of B and slightly more than 57 percent of C.

Arrange the type to facilitate comparisons, as in this *text-table*:

The three groups differed in how
they did something or other:

Group A	53%
Group B	46%
Group C	57%

There are nearly always better sequences than alphabetical—for example, ordering by content or by data values:

Group B	46%
Group A	53%
Group C	57%

Tables also work well when the data presentation requires many localized comparisons. In this 410-number table that I designed for the *New York Times* to show how different people voted in presidential elections in the United States, comparisons between the elections of 1980 and 1976 are read across each line; within-election analysis is conducted by reading downward in the clusters of three to seven lines. The horizontal rules divide the data into topical paragraphs; the rows are ordered so as to tell an ordered story about the elections.

This type of elaborate table, a *supertable*, is likely to attract and intrigue readers through its organized, sequential detail and reference-like quality. One supertable is far better than a hundred little bar charts.

How Different Groups Voted for President

Based on 12,782 interviews with voters at their polling places. Shown is how each group divided its vote for President and, in parentheses, the percentage of the electorate belonging to each group.

	CARTER	REAGAN	ANDERSON	CARTER-FORD in 1976
Democrats (43%)	66	26	6	77 - 22
Independents (23%)	30	54	12	43 - 54
Republicans (28%)	11	84	4	9 - 90
Liberals (17%)	57	27	11	70 - 26
Moderates (46%)	42	48	8	51 - 48
Conservatives (28%)	23	71	4	29 - 70
Liberal Democrats (9%)	70	14	13	86 - 12
Moderate Democrats (22%)	66	28	6	77 - 22
Conservative Democrats (8%)	53	41	4	64 - 35
Politically active Democrats (3%)	72	19	8	—
Democrats favoring Kennedy in primaries (13%)	66	24	8	—
Liberal Independents (4%)	50	29	15	64 - 29
Moderate Independents (12%)	31	53	13	45 - 53
Conservative Independents (7%)	22	69	6	26 - 72
Liberal Republicans (2%)	25	66	9	17 - 82
Moderate Republicans (11%)	13	81	5	11 - 88
Conservative Republicans (12%)	6	91	2	6 - 93
Politically active Republicans (2%)	5	89	6	—
East (32%)	43	47	8	51 - 47
South (27%)	44	51	3	54 - 45
Midwest (20%)	41	51	6	48 - 50
West (11%)	35	52	10	46 - 51
Blacks (10%)	82	14	3	82 - 16
Hispanics (2%)	54	36	7	75 - 24
Whites(88%)	36	55	8	47 - 52
Female (49%)	45	46	7	50 - 48
Male (51%)	37	54	7	50 - 48
Female, favors equal rights amendment (22%)	54	32	11	—
Female, opposes equal rights amendment (15%)	29	66	4	—
Catholic (25%)	40	51	7	54 - 44
Jewish (5%)	45	39	14	64 - 34
Protestant (46%)	37	56	6	44 - 55
Born-again white Protestant (17%)	34	61	4	—
18 - 21 years old (6%)	44	43	11	48 - 50
22 - 29 years old (17%)	43	43	11	51 - 46
30 - 44 years old (31%)	37	54	7	49 - 49
45 - 59 years old (23%)	39	55	6	47 - 52
60 years or older (18%)	40	54	4	47 - 52
Family income				
Less than $10,000 (13%)	50	41	6	58 - 40
$10,000 - $14,999 (14%)	47	42	8	55 - 43
$15,000 - $24,999 (30%)	38	53	7	48 - 50
$25,000 - $50,000 (24%)	32	58	8	36 - 62
Over $50,000 (5%)	25	65	8	—
Professional or manager (40%)	33	56	9	41 - 57
Clerical, sales or other white-collar (11%)	42	48	8	46 - 53
Blue-collar worker (17%)	46	47	5	57 - 41
Agriculture (3%)	29	66	3	—
Looking for work (3%)	55	35	7	65 - 34
Education				
High school or less (39%)	46	48	4	57 - 43
Some college (28%)	35	55	8	51 - 49
College graduate (27%)	35	51	11	45 - 55
Labor union household (26%)	47	44	7	59 - 39
No member of household in union (62%)	35	55	8	43 - 55
Family finances				
Better off than a year ago (16%)	53	37	8	30 - 70
Same (40%)	46	46	7	51 - 49
Worse off than a year ago (34%)	25	64	8	77 - 23
Family finances and political party				
Democrats, better off than a year ago (7%)	77	16	6	69 - 31
Democrats, worse off than a year ago (13%)	47	39	10	94 - 6
Independents, better off (3%)	45	36	12	—
Independents, worse off (9%)	21	65	11	—
Republicans, better off (4%)	18	77	5	3 - 97
Republicans, worse off (11%)	6	89	4	24 - 76
More important problem				
Unemployment (39%)	51	40	7	75 - 25
Inflation (44%)	30	60	9	35 - 65
Feel that U.S. should be more forceful in dealing with Soviet Union even if it would increase the risk of war (54%)	28	64	6	—
Disagree (31%)	56	32	10	—
Favor equal rights amendment (46%)	49	38	11	—
Oppose equal rights amendment (35%)	26	68	4	—
When decided about choice				
Knew all along (41%)	47	50	2	44 - 55
During the primaries (13%)	30	60	8	57 - 42
During conventions (8%)	36	55	7	51 - 48
Since Labor Day (8%)	30	54	13	49 - 49
In week before election (23%)	38	46	13	49 - 47

Source: 1976 and 1980 election day surveys by The New York Times/CBS News Poll and 1976 election day survey by NBC News.

Tables are clearly the best way to show exact numerical values, although the entries can also be arranged in semi-graphical form.

Some Winners and Losers in the Forecasting Game

About a year ago, eight forecasters were asked for their predictions on some key economic indicators. Here's how the forecasts stack up against the probable 1978 results (shown in the black panel).

Real G.N.P. Growth	Industrial Production Growth	Change in Consumer Prices	Corporate Profits Growth	Unemployment Rate
				Chase Econometrics: 7.4%
Council of Economic Advisers: +4.7%				Wharton Econometric Forecasting: 6.8%
Data Resources: +4.5%				Conference Board: 6.7%
Nat. Assoc. of Business Economists: +4.5%				Nat. Assoc. of Business Economists: 6.7%
Wharton Econometric Forecasting: +4.5%				I.B.M. Economics Department: 6.6%
Congressional Budget Office: +4.4%				Data Resources: 6.5%
Conference Board: +4.2%	Nat. Assoc. of Business Economists: +6.2%			Congressional Budget Office: 6.3%
I.B.M. Economics Department: +4.1%	I.B.M. Economics Department: +5.9%		Wharton Econometric Forecasting: +21%	Council of Economic Advisers: 6.3%
Real G.N.P. Growth: +3.8%	**Industrial Production Growth: +5.8%**	**Change in Consumer Prices: +7.7%**	**Corporate Profits Growth: +13.3%**	**Unemployment Rate: 6%**
Chase Econometrics: +2.8%	Conference Board: +5.5%	I.B.M. Economics Department: +6.6%	Data Resources: +10.5%	
	Data Resources: +5.2%	Nat. Assoc. of Business Economists: +6.5%	I.B.M. Economics Department: +10.4%	
	Wharton Econometric Forecasting: +4.8%	Conference Board: +6.2%	Chase Econometrics: +6.5%	
	Chase Econometrics: +1.9%	Data Resources: +6.2%		
		Chase Econometrics: +5.9%		
		Council of Economic Advisers: +5.9%		
		Wharton Econometric Forecasting: +5.4%		

Forecasters are not listed in categories for which they did not make a prediction.

*After taxes

Combined visualization types

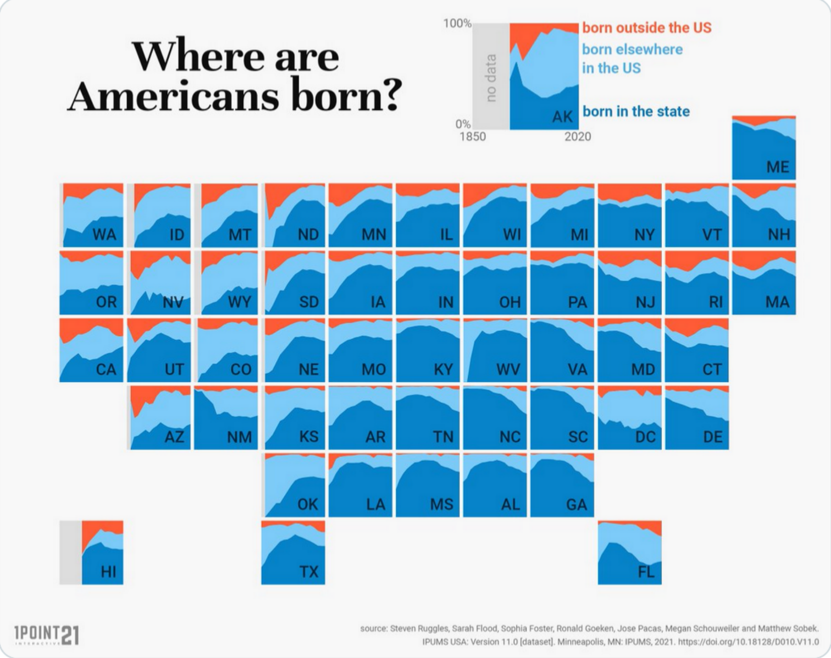
There can be interesting combinations of those types of graphs that we covered

Some of those advanced techniques will be covered in later course units

For example: geo-spatial placement of stacked time series

Erin
@erindataviz · Follow

Rather pleased with this map 🥳

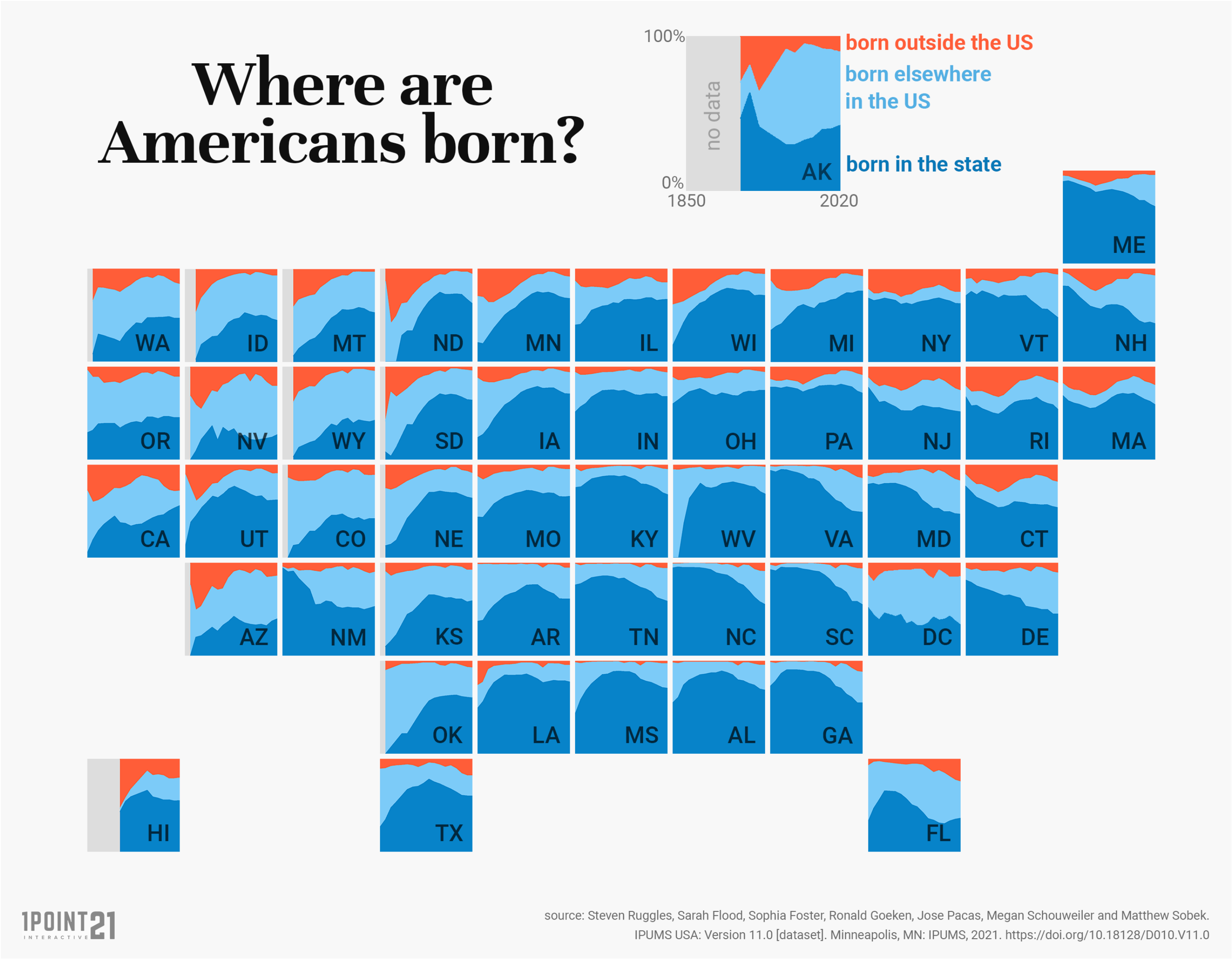


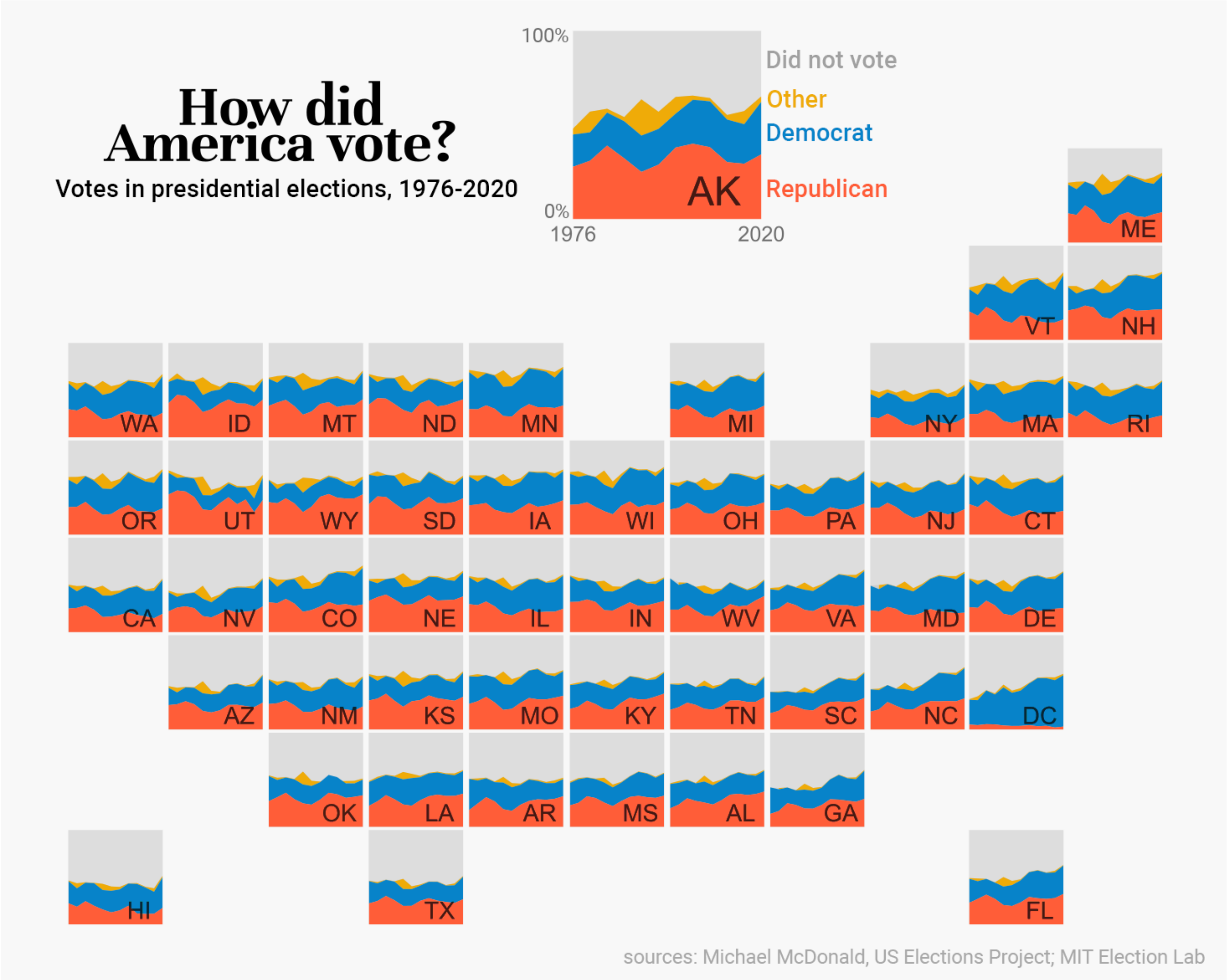
11:56 PM · Feb 2, 2022

36.7K Reply Copy link

Read 459 replies

https://erdavis.com/2022/02/09/how-i-made-the-viral-map/


Where are Americans born?
100%
0%
1850
2020
no data
AK
born outside the US
born elsewhere in the US
born in the state
ME
WA ID MT ND MN IL WI MI NY VT NH
OR NV WY SD IA IN OH PA NJ RI MA
CA UT CO NE MO KY WV VA MD CT
AZ NM KS AR TN NC SC DC DE
OK LA MS AL GA
HI TX FL
1POINT21
source: Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D010.V11.0



UNIVERSITY OF MANNHEIM

As so often, there are also examples that don’t serve so well as role models

US Dollar + Add to myFT

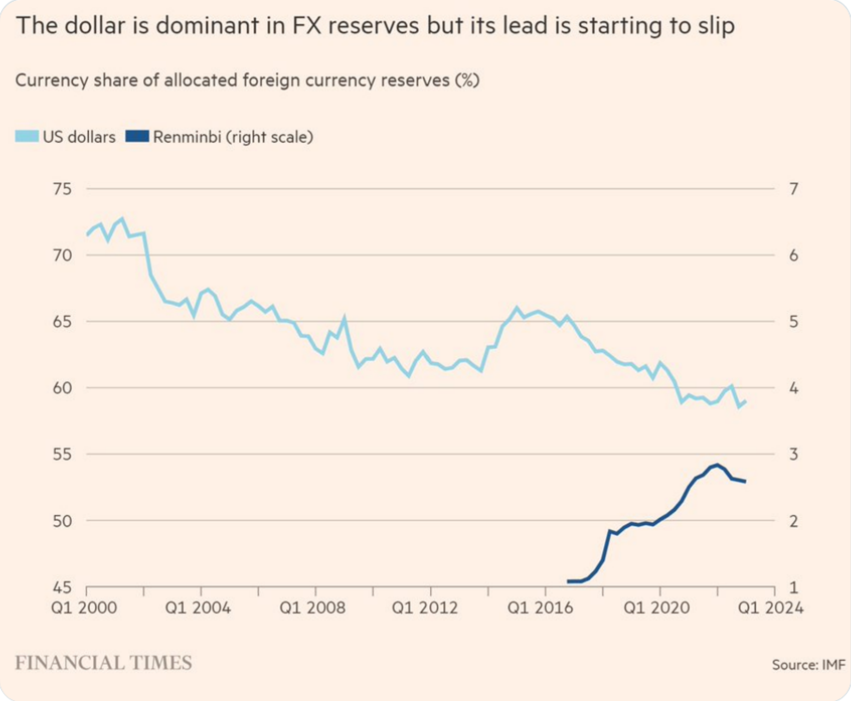
China capitalises on US sanctions in fight to dethrone dollar

Beijing uses developing world chagrin over Washington’s weaponisation of greenback to push global renminbi

https://www.ft.com/content/3888bdba-d0d6-49a1-9e78-4d07ce458f42

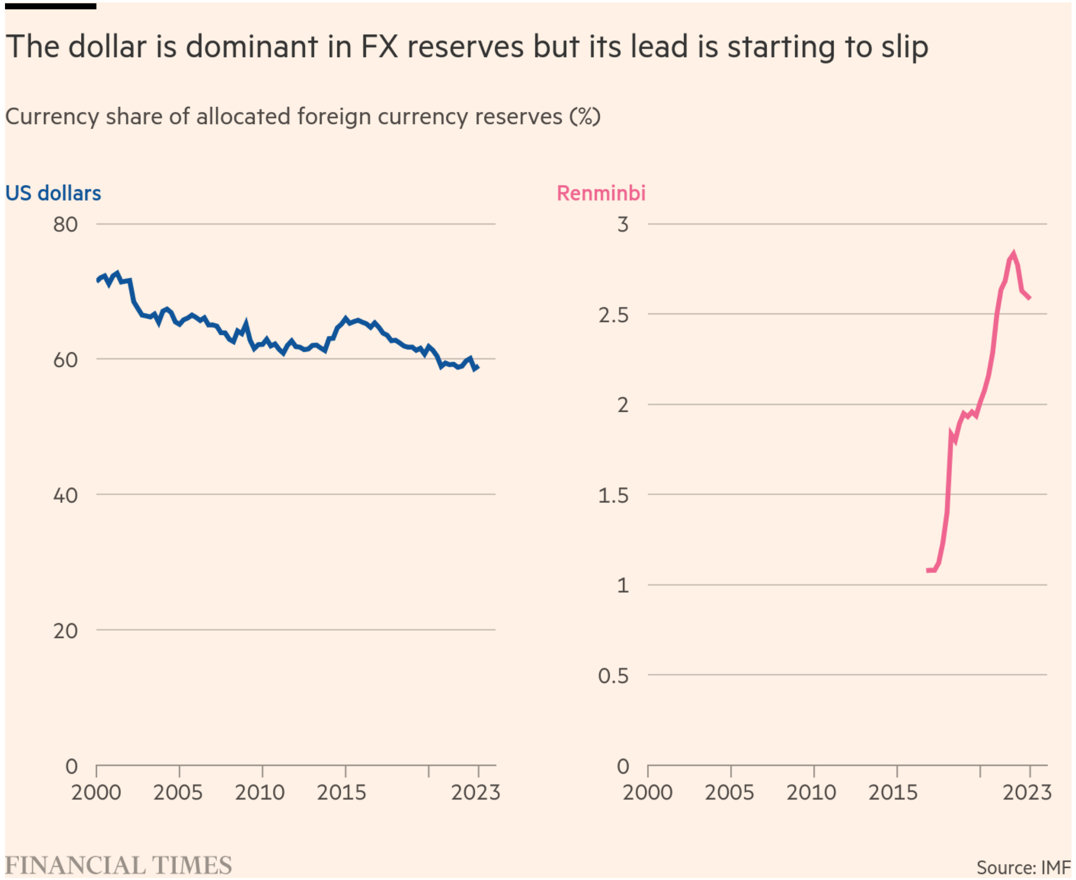
David Stillwell
@david_stillwell

In my big data class we cover data visualisation. I normally start by showing some terrible examples. Just saw this one in the FT of all places. I was very surprised to learn that the dollar is about to be surpassed by RMB as a % of foreign currency reserves... except it isn't:

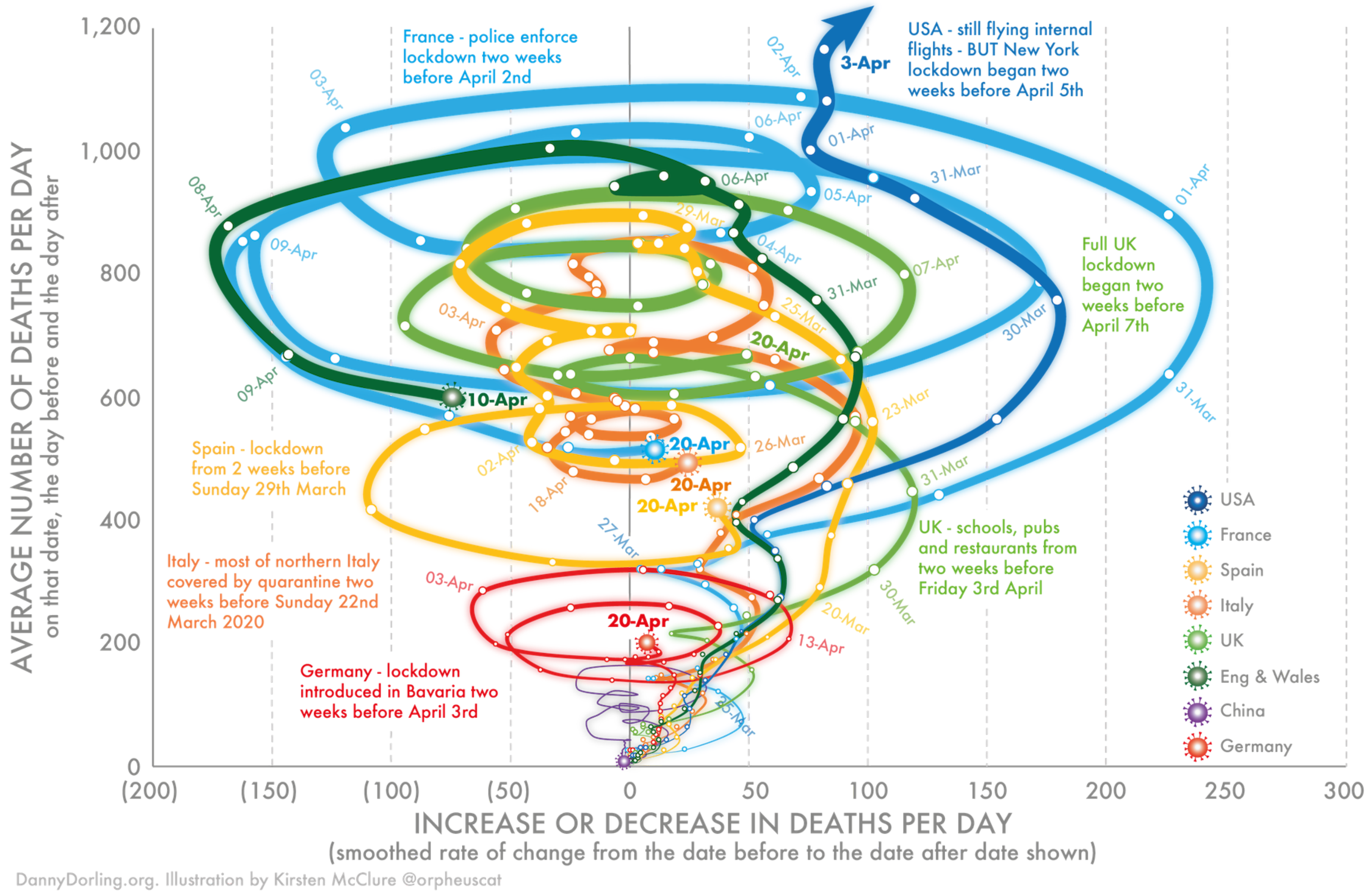


11:24 PM · Aug 24, 2023 · **1,457** Views

3 3 17 2



The chart accompanying this article has been amended to separate the dollar and the renminbi



UNIVERSITY OF MANNHEIM

THE CONVERSATION

Academic rigour, journalistic flair

Three charts that show where the coronavirus death rate is heading

Published: April 27, 2020 11.07am CEST • Updated: April 27, 2020 7.33pm CEST

https://theconversation.com/three-charts-that-show-where-the-coronavirus-death-rate-is-heading-137103

Design is choice. The theory of the visual display of quantitative information consists of principles that generate design options and that guide choices among options. The principles should not be applied rigidly or in a peevish spirit; they are not logically or mathematically certain; and it is better to violate any principle than to place graceless or inelegant marks on paper. Most principles of design should be greeted with some skepticism, for word authority can dominate our vision, and we may come to see only through the lenses of word authority rather than with our own eyes.

What is to be sought in designs for the display of information is the clear portrayal of complexity. Not the complication of the simple; rather the task of the designer is to give visual access to the subtle and the difficult—that is,

the revelation of the complex.

Graphical elegance is often found in simplicity of design and complexity of data.

Acknowledgements

https://yy.github.io/dviz-course/