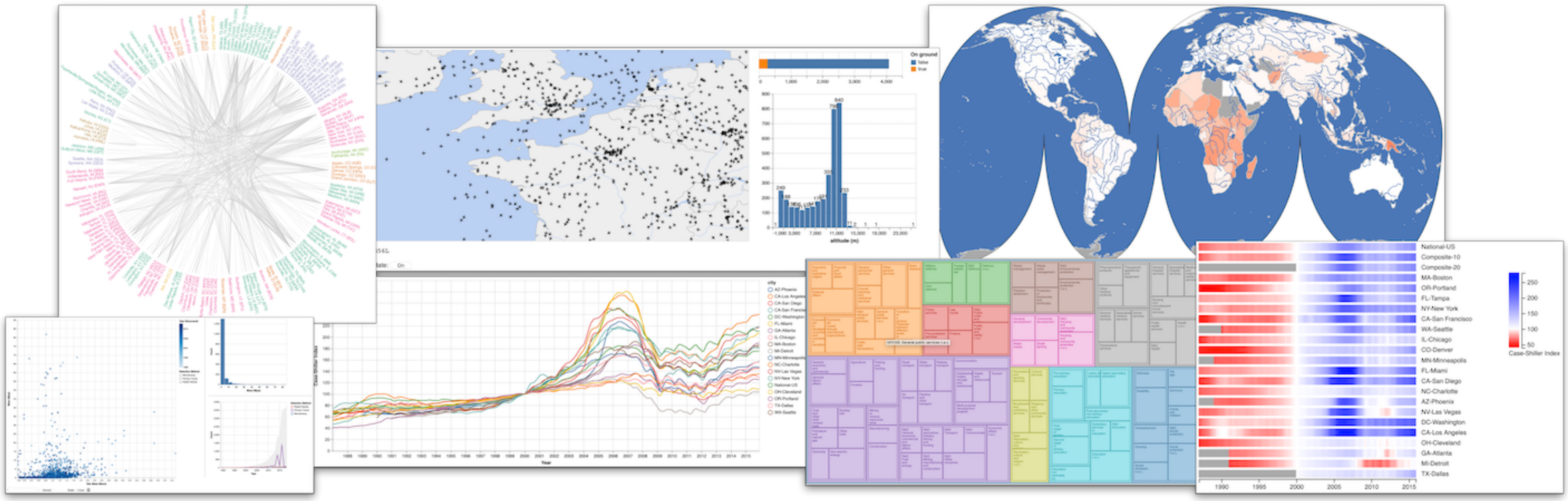
Data Visualization

INF552 (2023-2024)

Session 03
Multi-variate Data Visualisation (Part I)



INSTITUT POLYTECHNIQUE DE PARIS

Inria

emmanuel.pietriga@inria.fr

Flipped Classroom (s#04, s#05, s#07, s#08)

Oct-13 Oct-20 Nov-10 Nov-17

Organization

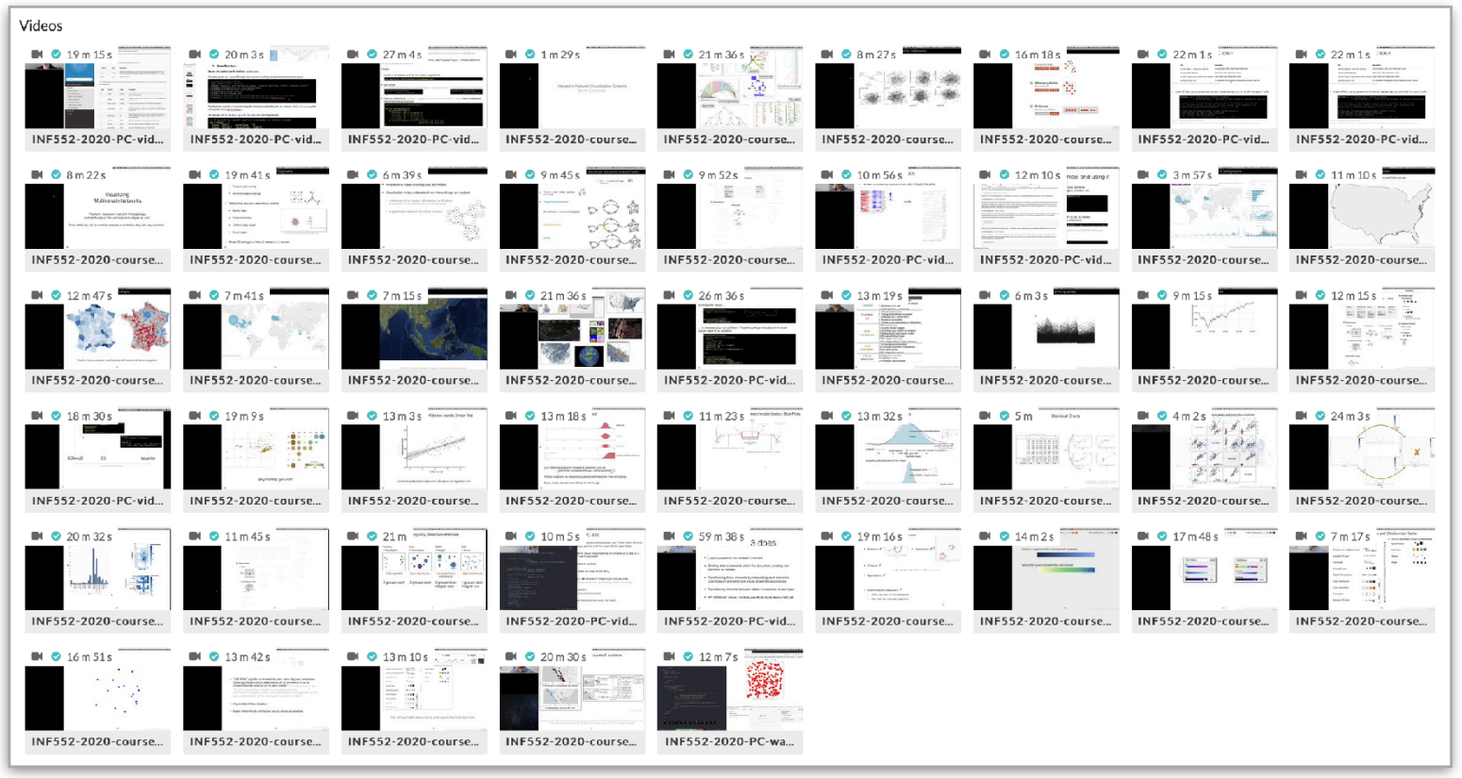
- Pre-recorded videos of the corresponding sessions are made available in advance.
  This includes sequences preparing for the following Petite Class session.

- Watch them before the actual class is scheduled to take place (Friday 8:30-10:30).
  Or at the very latest during that time slot.

- Petite Classe sessions remain unaffected.

- The course time slot is now free for the following activities:
  - one-on-one discussion about your individual project *(max 10mn)*;
  - ask questions about the course *(today's session or any previous one)*;
  - ask questions about what was done in previous Petite Classe sessions.

} Use the shared spreadsheet to book a slot *in advance*.



Flipped Classroom (s#04, s#05, s#07, s#08)

Oct-13 Oct-20 Nov-10 Nov-17

INF552-2020 FC

File Edit View Insert Format Data Tools Extensions Help Last edit was 2 minutes ago

100% $ % .0 .00 123 Default (Ari... 10

C12

	A	B	C	D
1				
2		**INF552 - 2022 Flipped Classroom**		
3				
4		Questions about your individual project	Questions about a course session (theory, design)	Questions about a petite classe session (software dev)
5		*<put your name below>*	*<put your name below>*	*<put your name below>*
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				

https://docs.google.com/spreadsheets/d/1EJ0Z4Juf4y_Yloozy9vUnp4CfaYardcARQ_QctZjLuM/edit?usp=sharing

Problem Statement (illustration)

THE WORLD BANK IBRD • IDA

Sign In

This page is in English Español Français عربي 中文

DataBank | World Development Indicators

Table | Chart | Map | Metadata

Download options

Variables | Layout | Styles | Save | Share | Embed

- Database — Available 70 | Selected 1
- Country — Available 217 | Selected 264
- Series — Available 3 | Selected 2

Search: fertility

A F W

- [ ] Adolescent fertility rate (births per 1,000 women ages 15-19)
- [x] Fertility rate, total (births per woman)
- [ ] Wanted fertility rate (births per woman)

Create Custom Indicator

Define Aggregation Rule

- Time — Available 20 | Selected 20

Preview

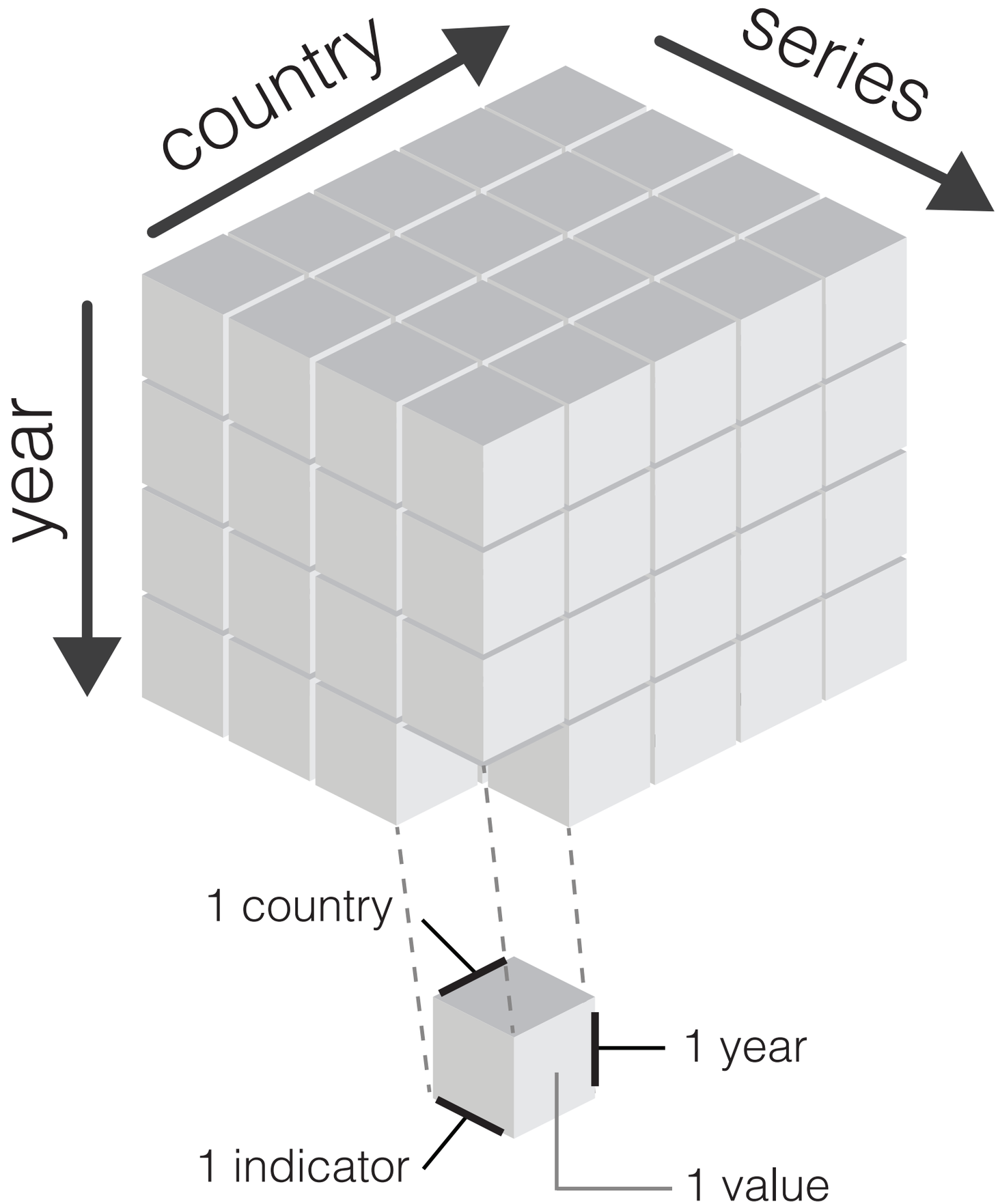
Clear Selection | Add Country (264) Add Series (2) Add Time (20)

France

	2015	2016	2017
Life expectancy at birth, total (years)	82.3	82.3	..
Fertility rate, total (births per woman)	2.0	2.0	..

Source: World Development Indicators. Click on a metadata icon for original source information to be used for citation.

- country / group (260+)
- series cc (1591)
- time (typically yearly)



Some general terminology:

- *Dimension*: attribute that groups, separates or filters data items.

- *Measure*: attribute that addresses the question of interest.

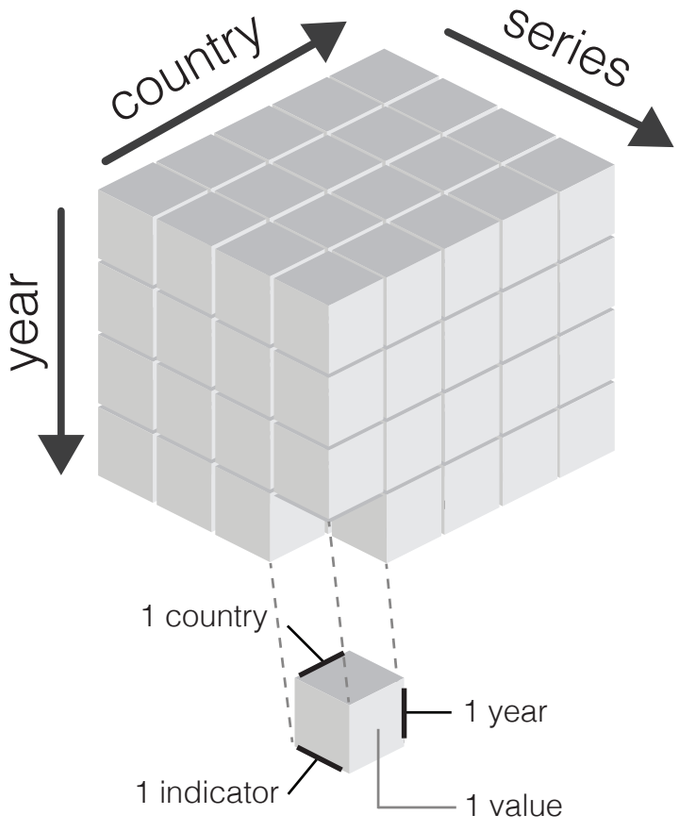
- *Independent variables*: those that the experimenter manipulates.

- *Dependent variables*: the outcomes of the experiment.

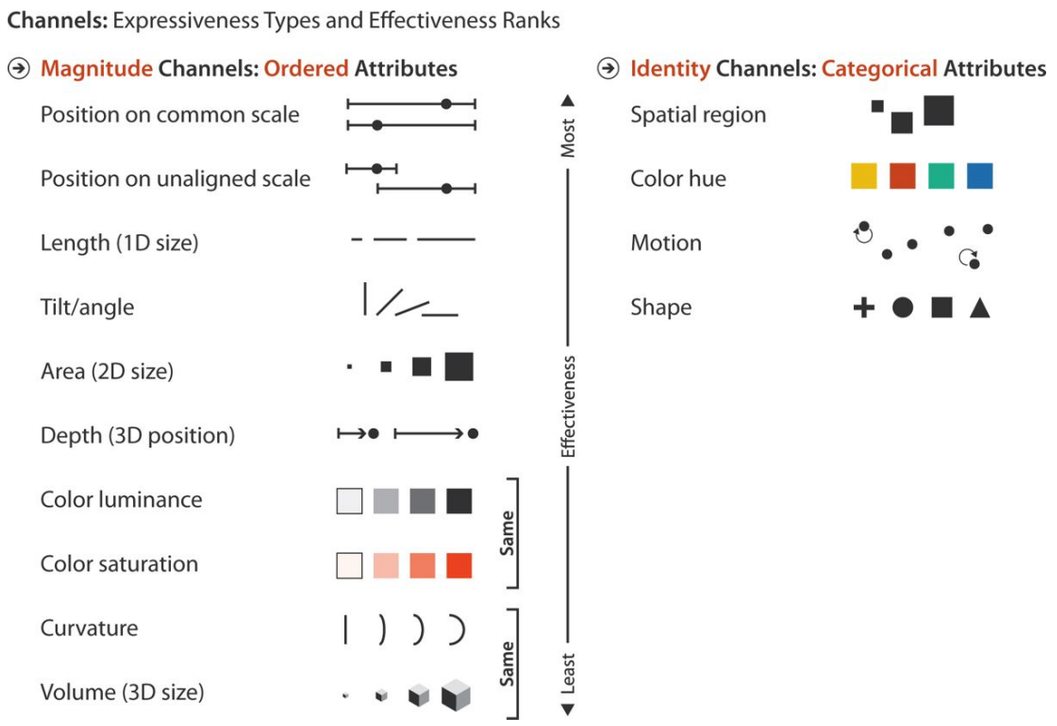
Problem Statement

Often many more dimensions than encoding channels.

Some channels are only good for categorical data.
Conversely, some others are only good for ordered data.



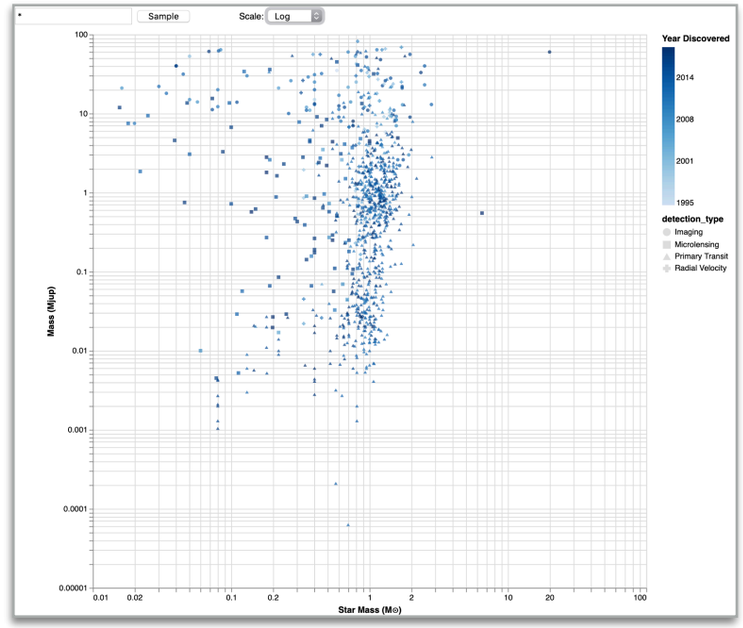
?



[Image source: Visualization Analysis and Design. Tamara Munzner, with illustrations by Eamonn Maguire. A K Peters Visualization Series, CRC Press, 2014]

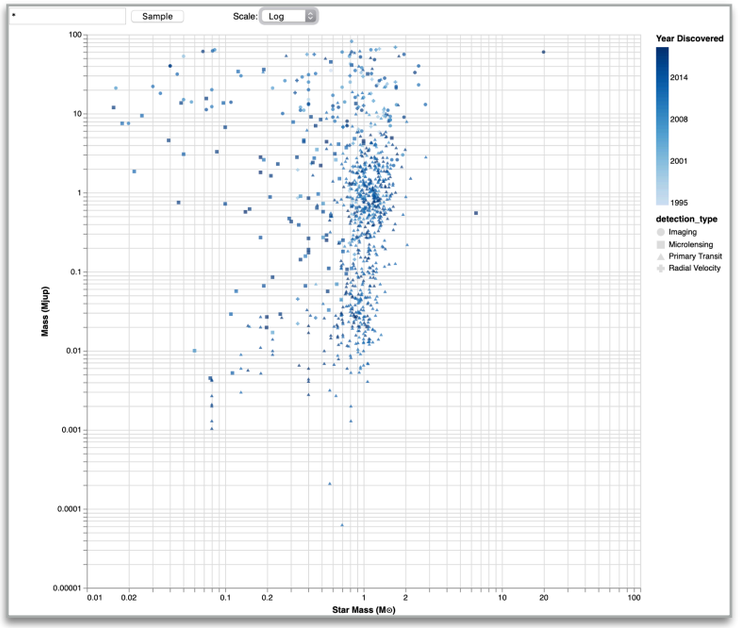
Main strategies

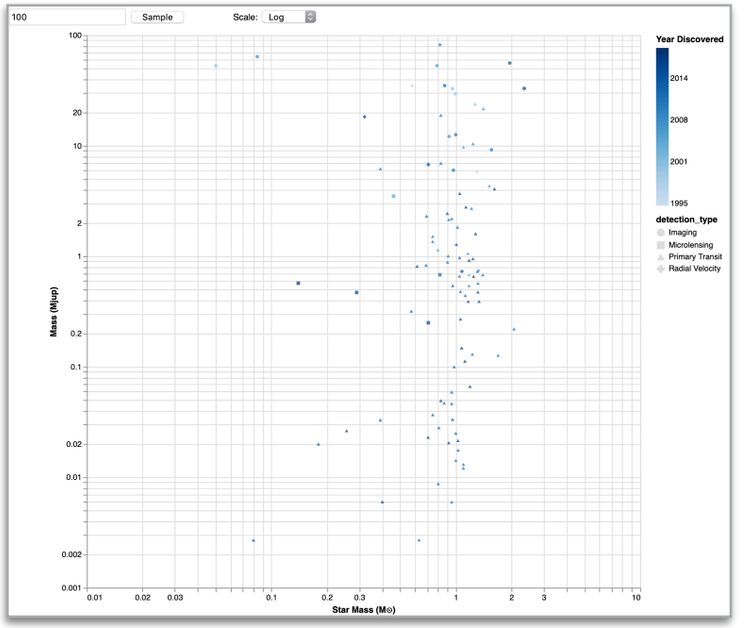
- *(try to)* visualize everything;
- visualize a subset of everything:
  - sample the data
  - filter the data
  - select a subset of dimensions and measures
  - use interaction to navigate between, or compare, different subsets
  - use dimensionality-reduction techniques (MDS, PCA, t-SNE, …)

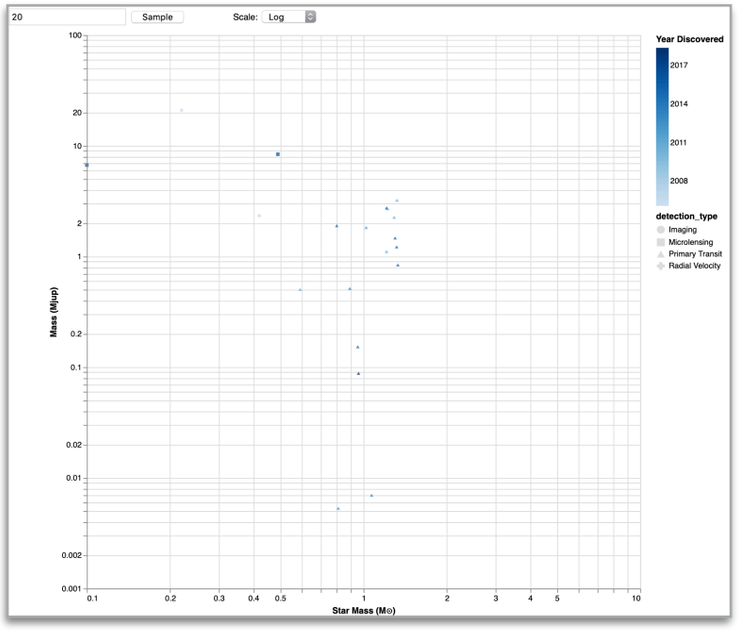


Main strategies

- *(try to)* visualize everything;
- visualize a subset of everything:
  - sample the data
  - filter the data
    - select a subset of dimensions and measures
    - use interaction to navigate between, or compare, different subsets
    - use dimensionality-reduction techniques (MDS, PCA, t-SNE, …)

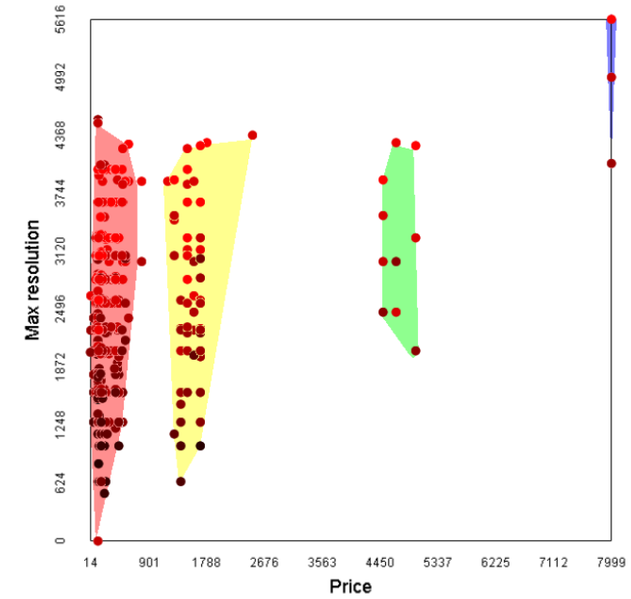


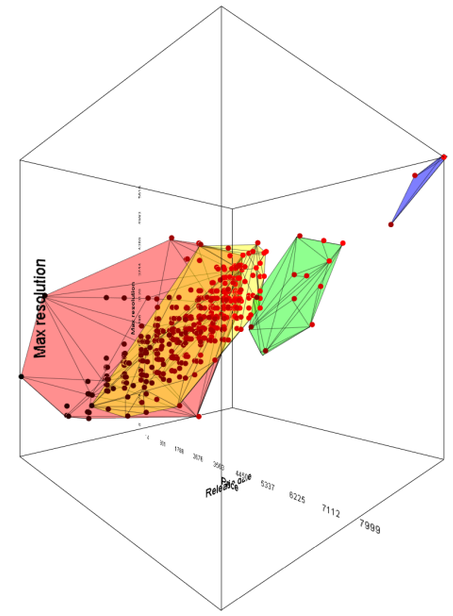


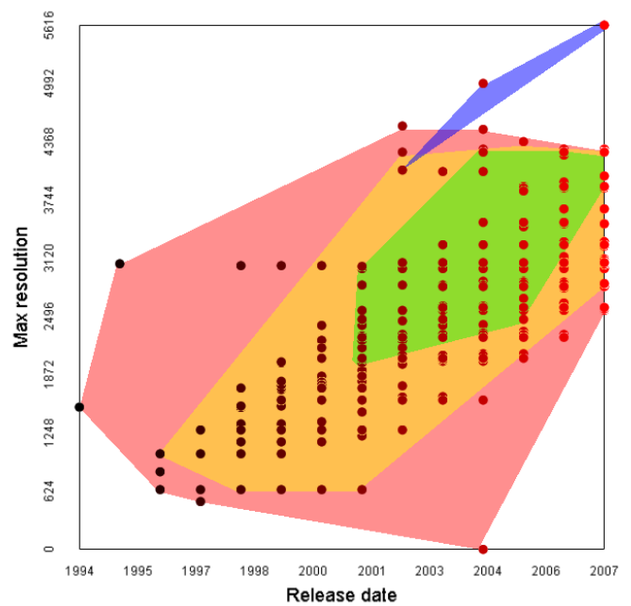


Main strategies

- *(try to)* visualize everything;
- visualize a subset of everything:
  - sample the data
  - filter the data
    - select a subset of dimensions and measures
    - use interaction to navigate between, or compare, different subsets
    - use dimensionality-reduction techniques (MDS, PCA, t-SNE, …)

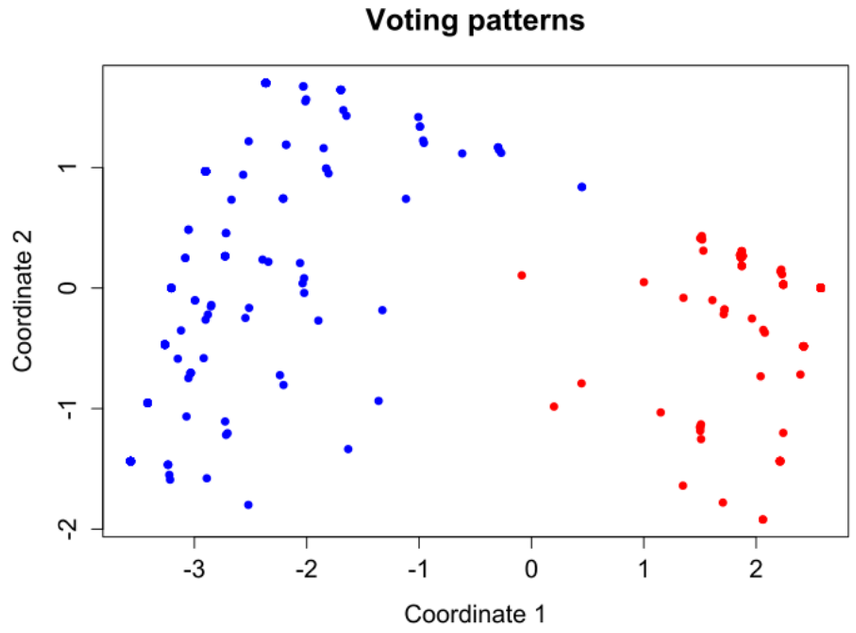


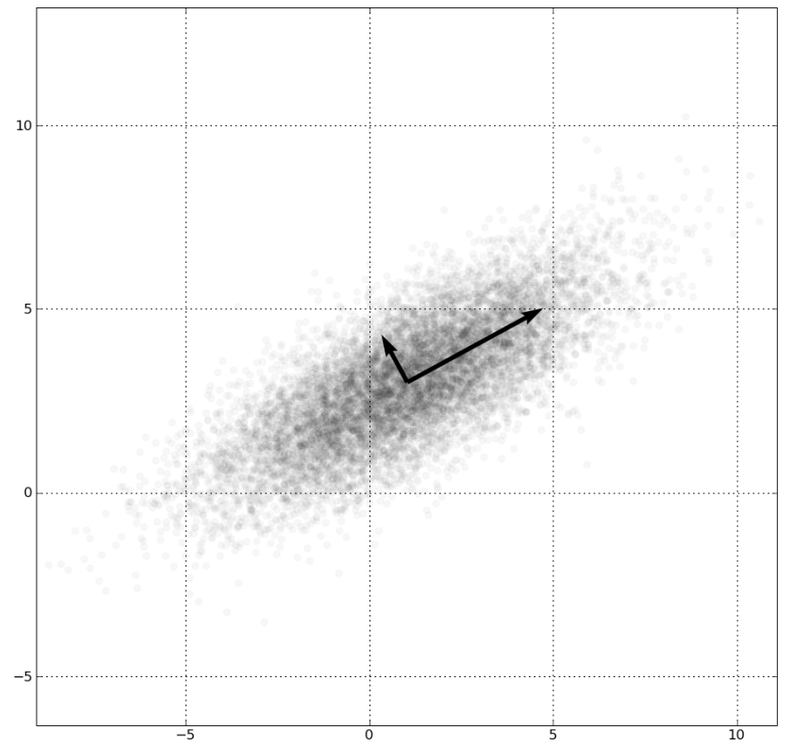


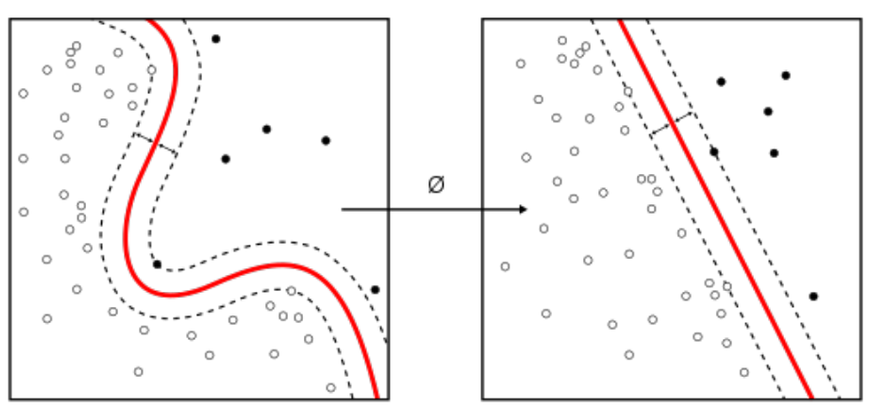


Main strategies

- *(try to)* visualize everything;
- visualize a subset of everything:
  - sample the data
  - filter the data
    - select a subset of dimensions and measures
    - use interaction to navigate between, or compare, different subsets
    - use dimensionality-reduction techniques (MDS, PCA, t-SNE, …)

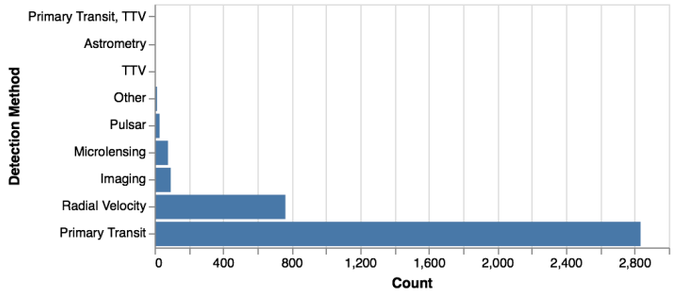


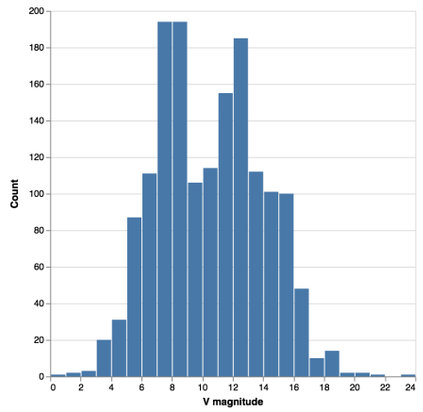


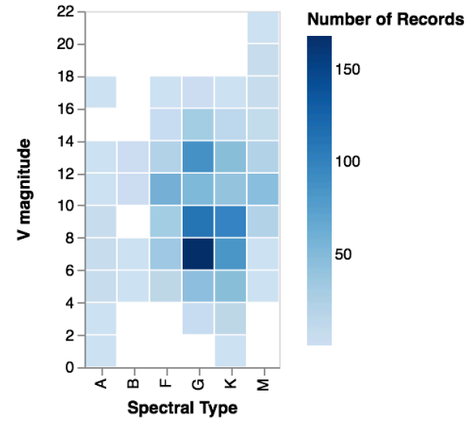


Choosing a chart type

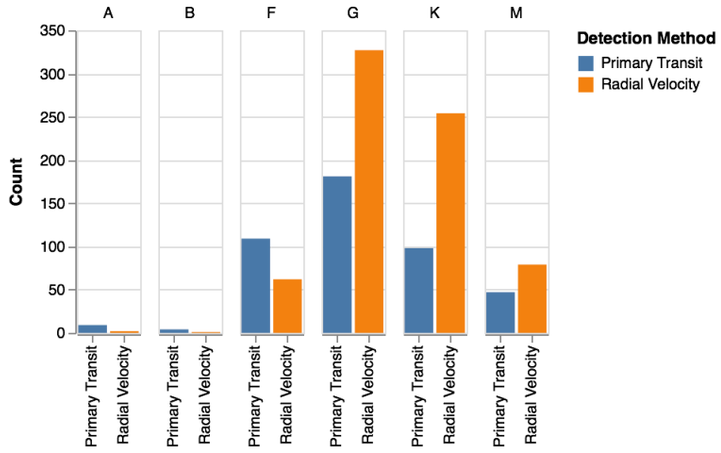
How is a measure distributed?





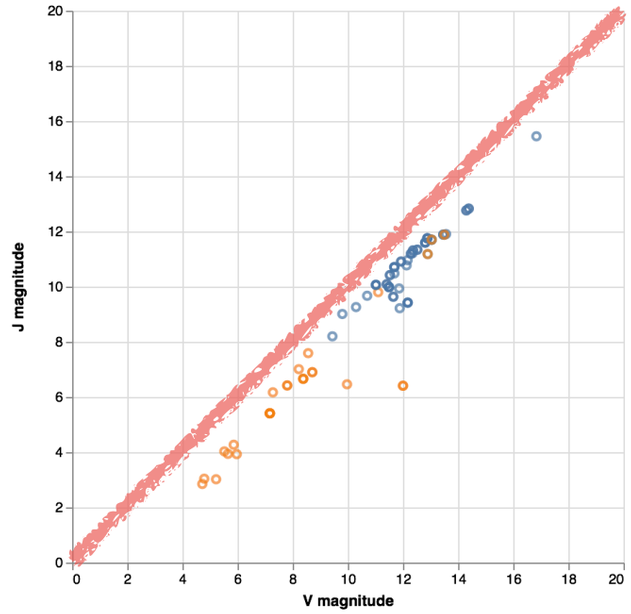


How do groups differ from each other?

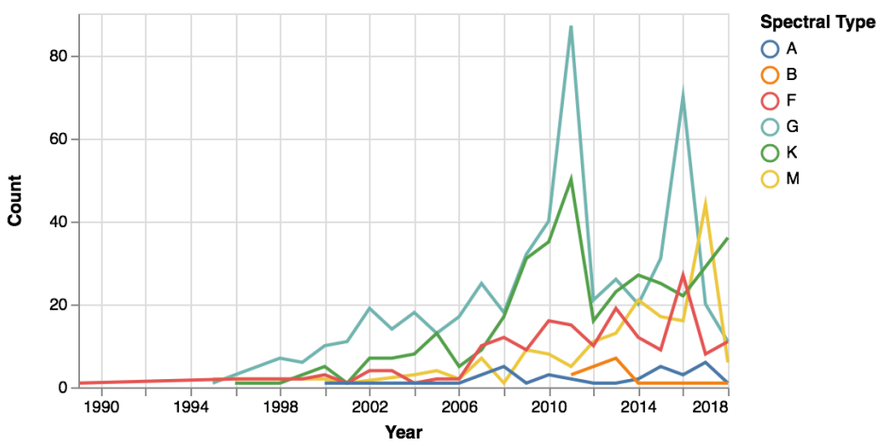


Do individual items fall into groups?

Is there a relationship between attributes of items?



How does an attribute vary continuously?

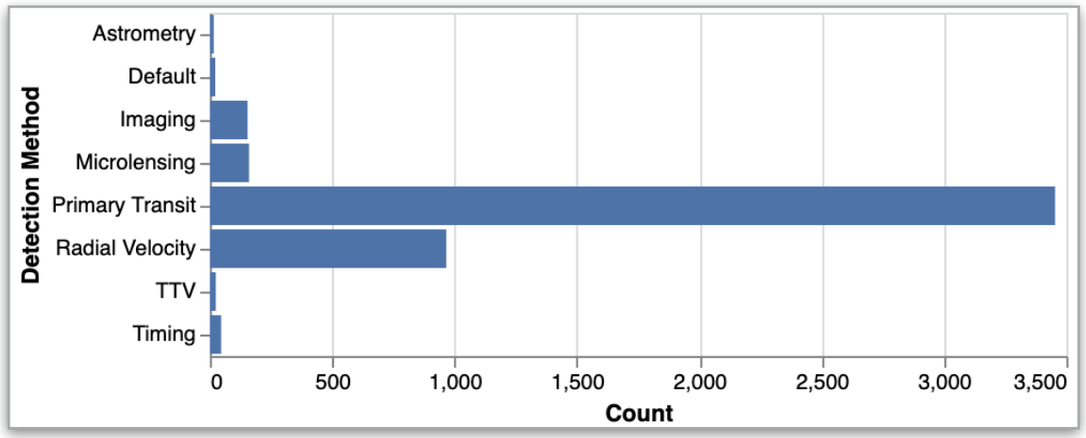


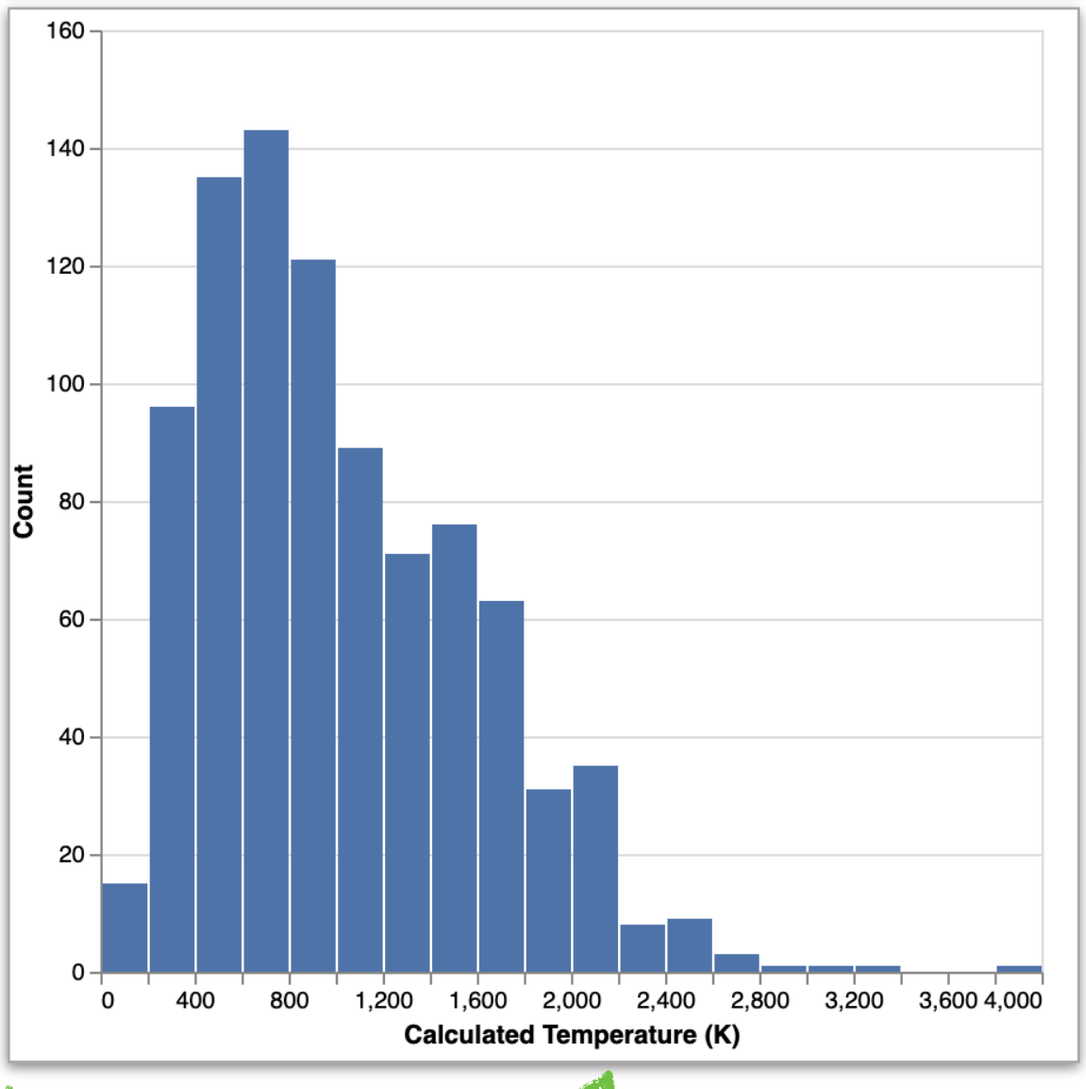
[Data source: exoplanets.eu]

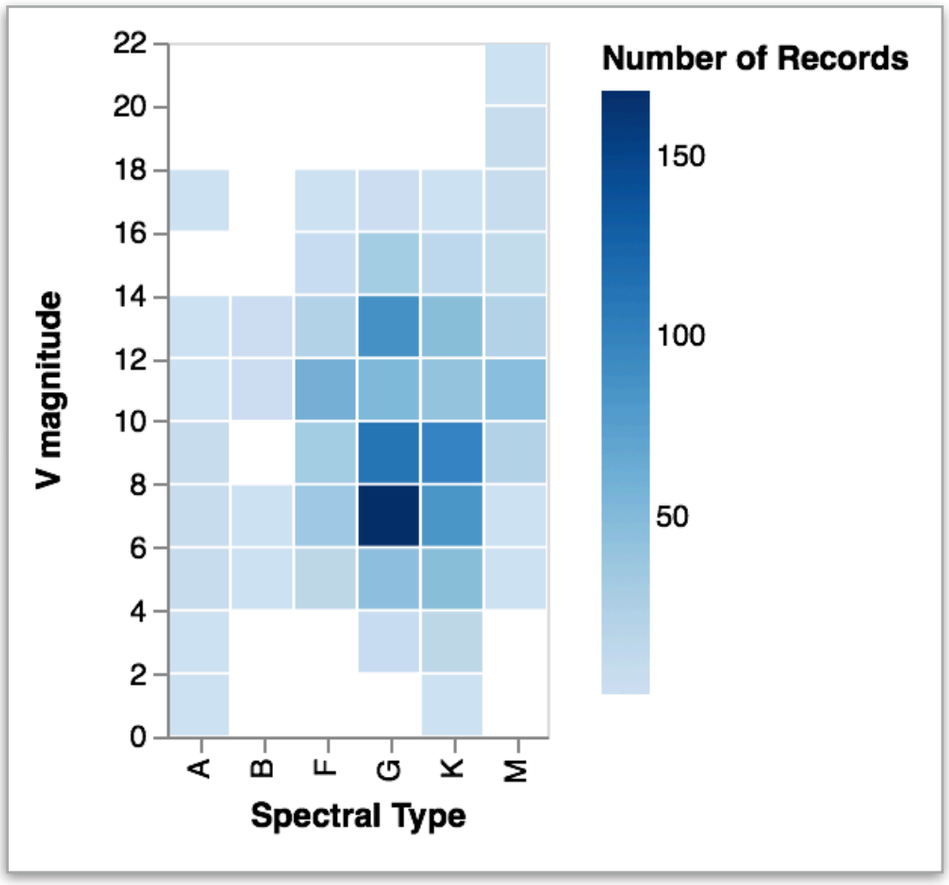
How is a measure distributed?

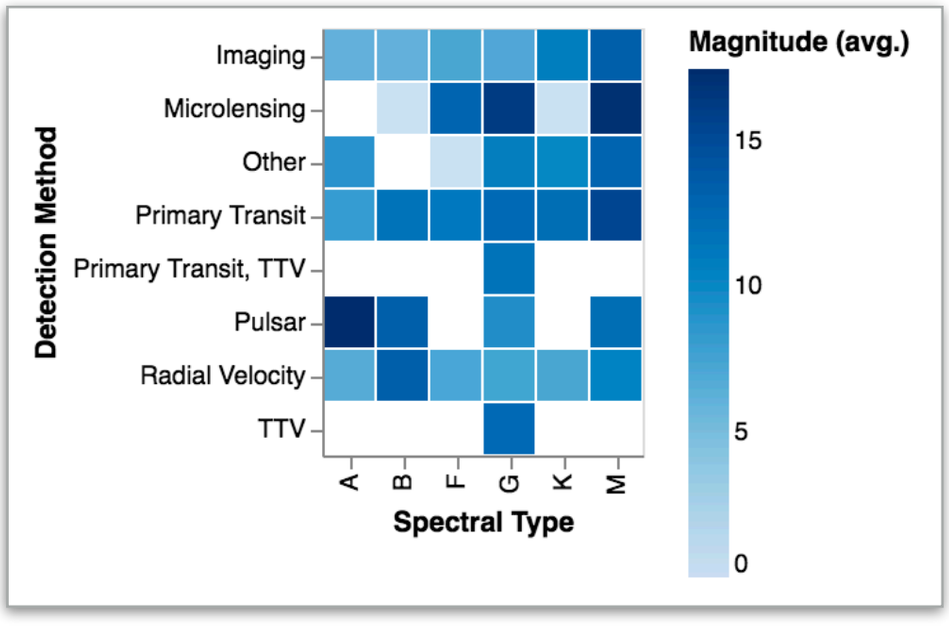
bar chart, histogram, 2D histogram, heatmap

categorical









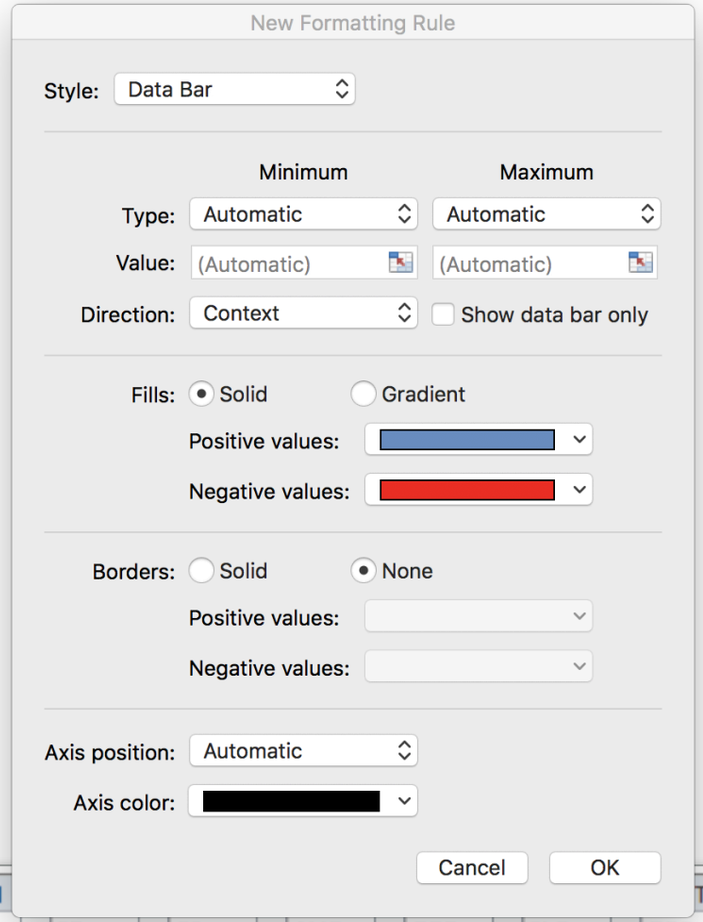
counting items
vs.
showing a measure

quantitative

[Data source: exoplanets.eu]

Visualizing amounts and distributions

Sometimes,
just using a spreadsheet program like MS-Excel,
you can turn your table into a heatmap
using conditional formatting



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	S	T	U	V	W	X
1	Rank	Player	Team	Pos	G	AB	R	H	2B	3B	HR	RBI	BB	SO	SB	CS	AVG	OBP	SLG	OPS
2	1	Posey, B	SF	C	148	530	78	178	39	1	24	103	69	96	1	1	0.336	0.408	0.549	0.957
3	2	Cabrera, M	DET	3B	161	622	109	205	40	0	44	139	66	98	4	1	0.33	0.393	0.606	0.999
4	3	McCutchen, A	PIT	CF	157	593	107	194	29	6	31	96	70	132	20	12	0.327	0.4	0.553	0.953
5	4	Trout, M	LAA	CF	139	559	129	182	27	8	30	83	67	139	49	5	0.326	0.399	0.564	0.963
6	5	Beltre, A	TEX	3B	156	604	95	194	33	2	36	102	36	82	1	0	0.321	0.359	0.561	0.921
7	6	Braun, R	MIL	LF	154	598	108	191	36	3	41	112	63	128	30	7	0.319	0.391	0.595	0.987
8	7	Mauer, J	MIN	C	147	545	81	174	31	4	10	85	90	88	8	4	0.319	0.416	0.446	0.861
9	8	Jeter, D	NYY	SS	159	683	99	216	32	0	15	58	45	90	9	4	0.316	0.362	0.429	0.791
10	9	Molina, Y	STL	C	138	505	65	159	28	0	22	76	45	55	12	3	0.315	0.373	0.501	0.874
11	10	Fielder, P	DET	1B	162	581	83	182	33	1	30	108	85	84	1	0	0.313	0.412	0.528	0.94
12	11	Hunter, T	LAA	CF	140	534	81	167	24	1	16	92	38	133	9	1	0.313	0.365	0.451	0.817
13	12	Butler, B	KC	1B	161	614	72	192	32	1	29	107	54	111	2	1	0.313	0.373	0.51	0.882
14	13	Cano, R	NYY	2B	161	627	105	196	48	1	33	94	61	96	3	2	0.313	0.379	0.55	0.929
15	14	Pacheco, J	COL	3B	132	475	51	147	32	3	5	54	22	61	7	2	0.309	0.341	0.421	0.762
16	15	Craig, A	STL	RF	119	469	76	144	35	0	22	92	37	89	2	1	0.307	0.354	0.522	0.876
17	16	Scutaro, M	SF	3B	156	620	87	190	32	4	7	74	40	49	9	4	0.306	0.348	0.405	0.753
18	17	Wright, D	NYM	3B	156	581	91	178	41	2	21	93	81	112	15	10	0.306	0.391	0.492	0.883
19	18	Jay, J	STL	CF	117	443	70	135	22	4	4	40	34	71	19	7	0.305	0.373	0.4	0.773
20	19	Murphy, D	TEX	LF	147	457	65	139	29	3	15	61	54	74	10	5	0.304	0.38	0.479	0.859
21	20	Rios, A	CWS	CF	157	605	93	184	37	8	25	91	26	92	23	6	0.304	0.334	0.516	0.85
22	21	Gonzalez, C	COL	LF	135	518	89	157	31	5	22	85	56	115	20	5	0.303	0.371	0.51	0.881

[Source (bottom image): dataremixed.com]

Visualizing proportions

DATA VISUALIZATION + ART

BEAUTIFUL SCIENCE DESERVES EFFECTIVE VISUALS

When you asphyxiate new knowledge with airless figures, you leave everyone breathless—in a bad way.

Let's look at a simple figure recently published in Nature Medicine, redesign it and take this opportunity to talk about fundamentals of organizing and communicating information.

The trouble isn't that we cannot find good ways to show complex data—it's that we fail at showing simple data.

PIE CHART IS THE UGLIEST DATA POEM

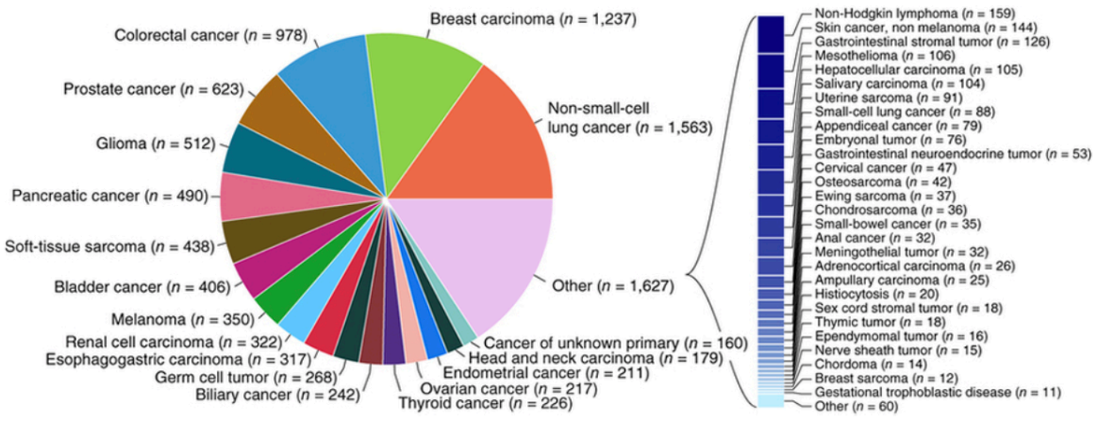
Consider the pie chart. Given the quote below, it can be considered a data poem—it attempts to hide even the most simple patterns.

> *In science one tries to tell people, in such a way as to be understood by everyone, something that no one ever knew before. But in poetry, it's the exact opposite.*
> —Paul Dirac, Mathematical Circles Adieu by H. Eves [quoted]

But the pie chart is a bad data poem. It never overcomes your initial disappointment and only rarely provides precise answers.

FIGURES THAT ARE VICTIMS OF THEIR OWN SUCCESS

This case study uses Figure 2b, shown below, from the recent report [1] of the sequencing of tumor and matched normal of over 10,000 patients.



▲ Overview of the MSK-IMPACT cohort. Distribution of tumor types among cases successfully sequenced from 10,336 patients. Cases represented 62 principal tumor types encapsulating 361 detailed tumor types. (Caption from Nature Medicine) (zoom, Ahmet Zehir et al. (2017) Nature Medicine doi:10.1038/nm.4333)

The reason why I chose to focus on this figure is not because it is a pie chart—there are countless pie charts out there that trigger only a mild reaction in me.

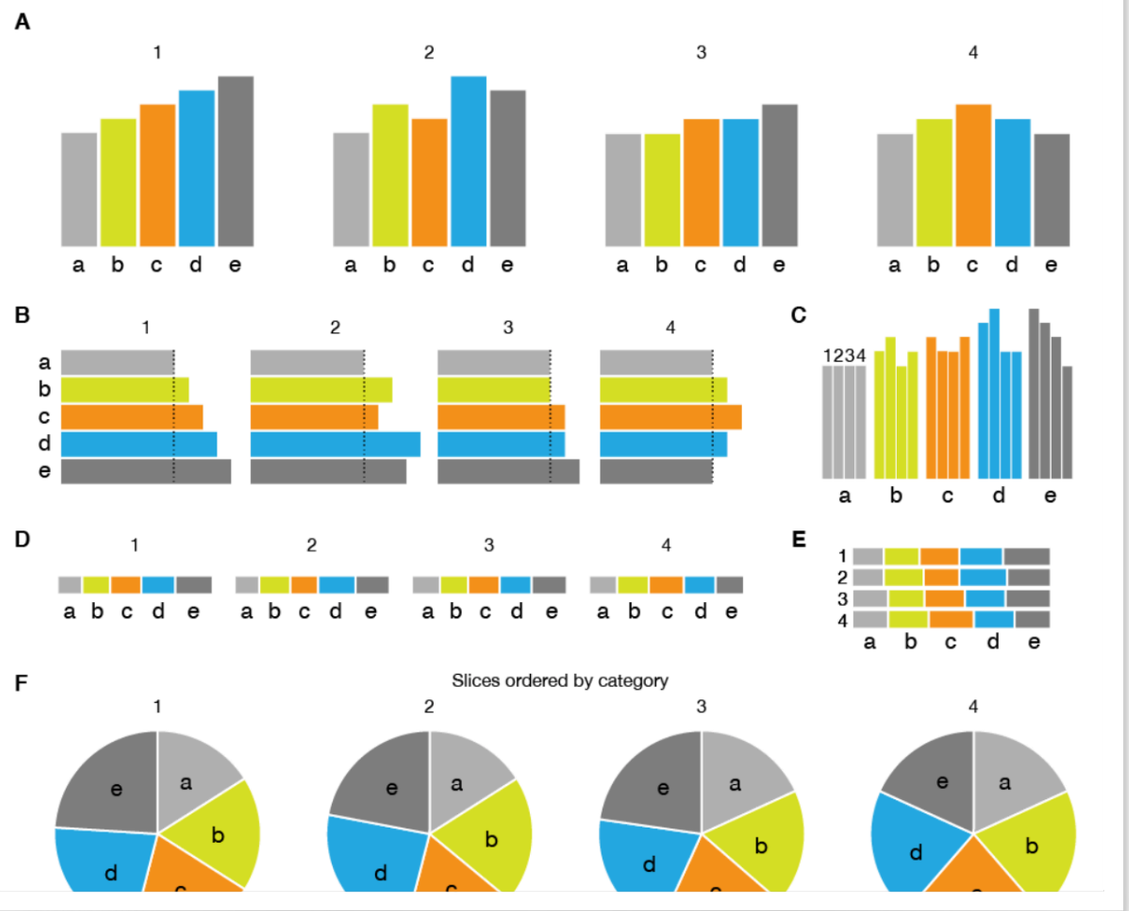
chart. This is a stunning blow to the pie chart—because of our lack of precision in assessing slice size, the pie chart cannot unambiguously communicate that two values are the same or that they are different by only a small amount.

And it gets worse.

THE PROBLEM WITH PIE CHARTS

My views here are hardly new. Much has been written about the shortcomings of pie charts and it may be that we do not even read pie charts by angle. However, in the context of the Nature Medicine figure, I want to focus on not merely data encoding and perception of patterns but also how pie charts fail at organizing information.

The figure below illustrates some (but not all) of the issues of pie charts: perception of proportions and label layout. It builds on the example above which demonstrated that a pie chart cannot tell us with precision that two values are the same.

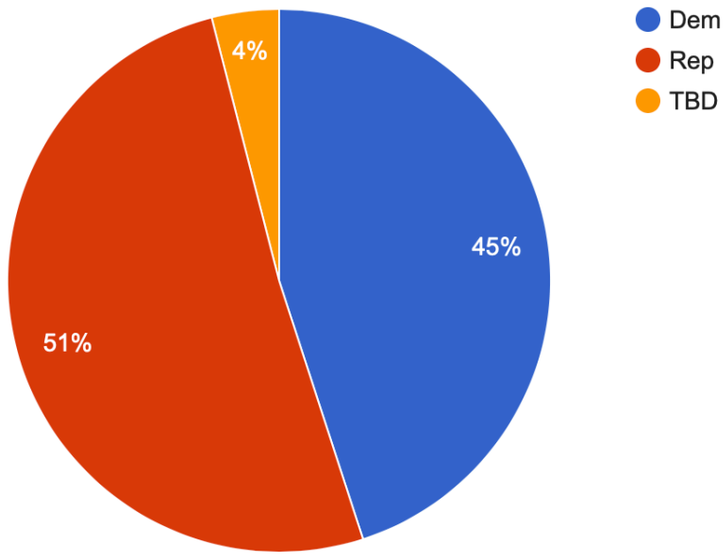


http://mkweb.bcgsc.ca/pie.charts/

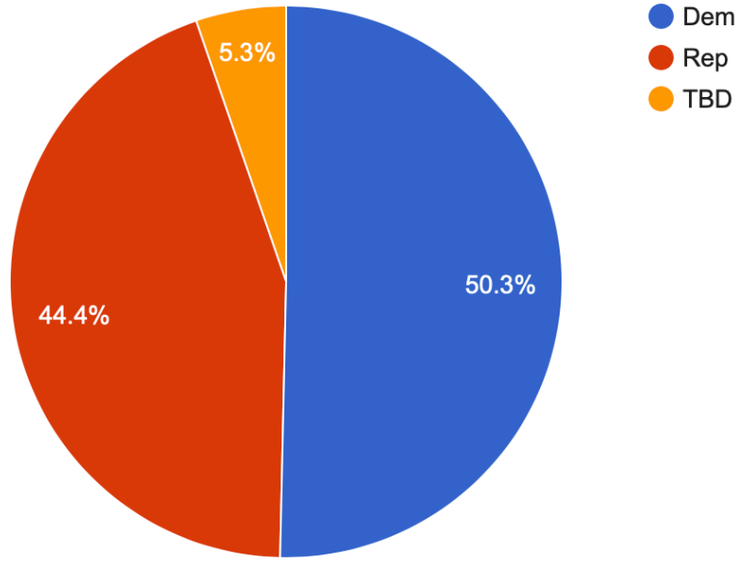
Visualizing proportions

Still, pie charts have some good properties:

- emphasize items as proportions of a whole
- highlight simple fractions: 1/2, 1/3, 1/4



Senate



House

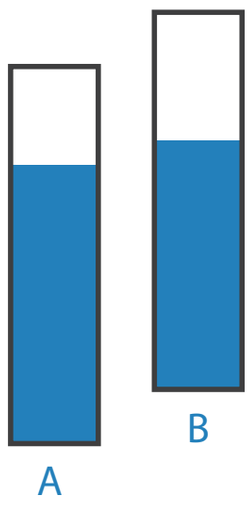
In a stacked bar chart, the relationship of each bar to the total is not visually obvious:

but...

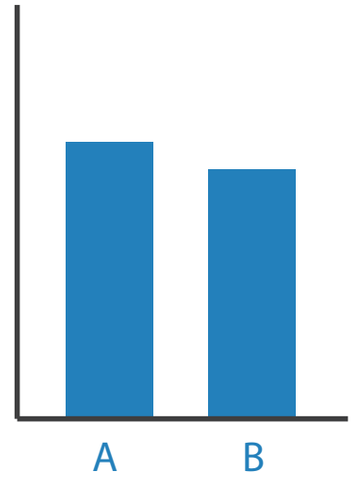
Relative *vs.* absolute judgements

- Our perceptual system mostly operates with relative judgements, not absolute ones.
- Accuracy increases with common frame/scale and alignment.

Length



Position along unaligned common scale

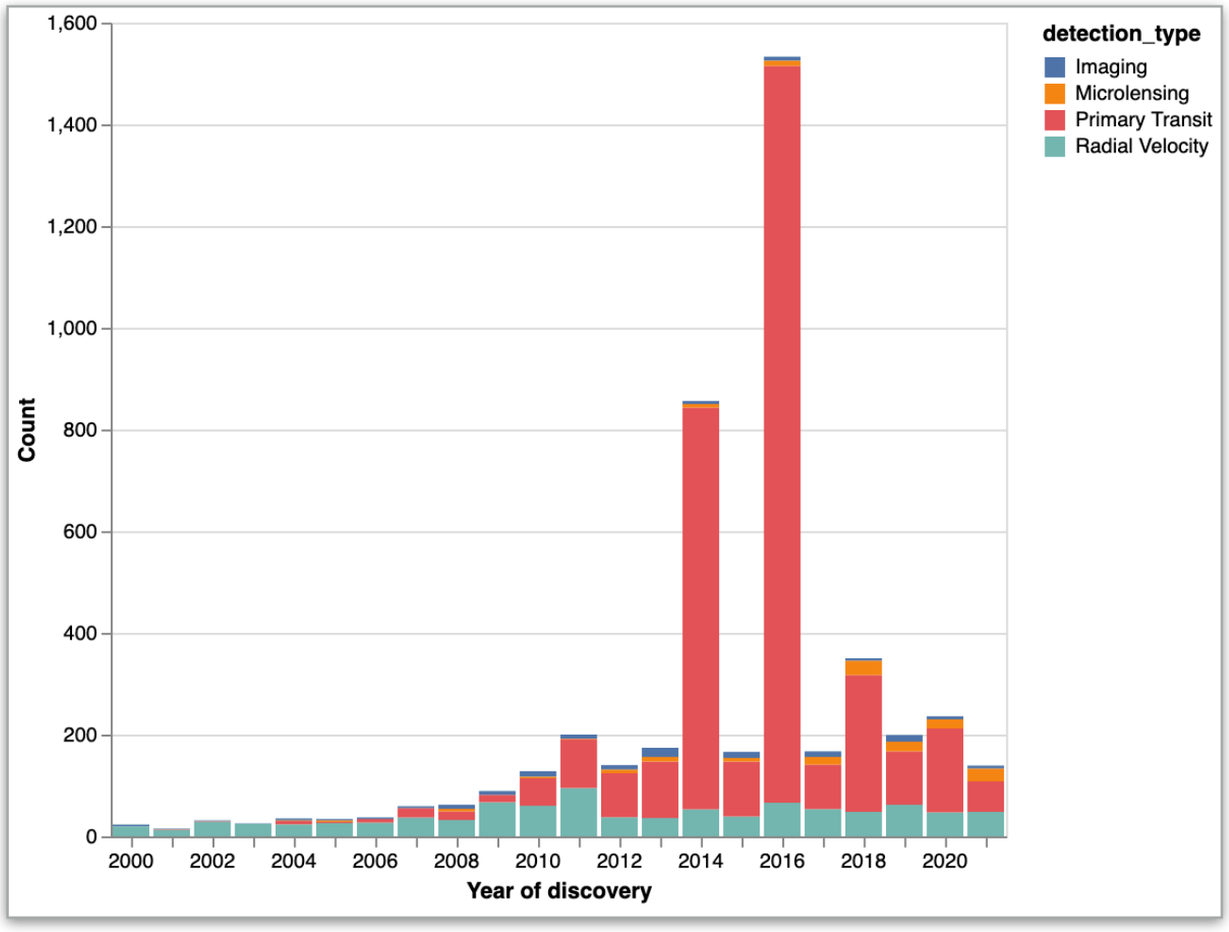


Position along aligned common scale

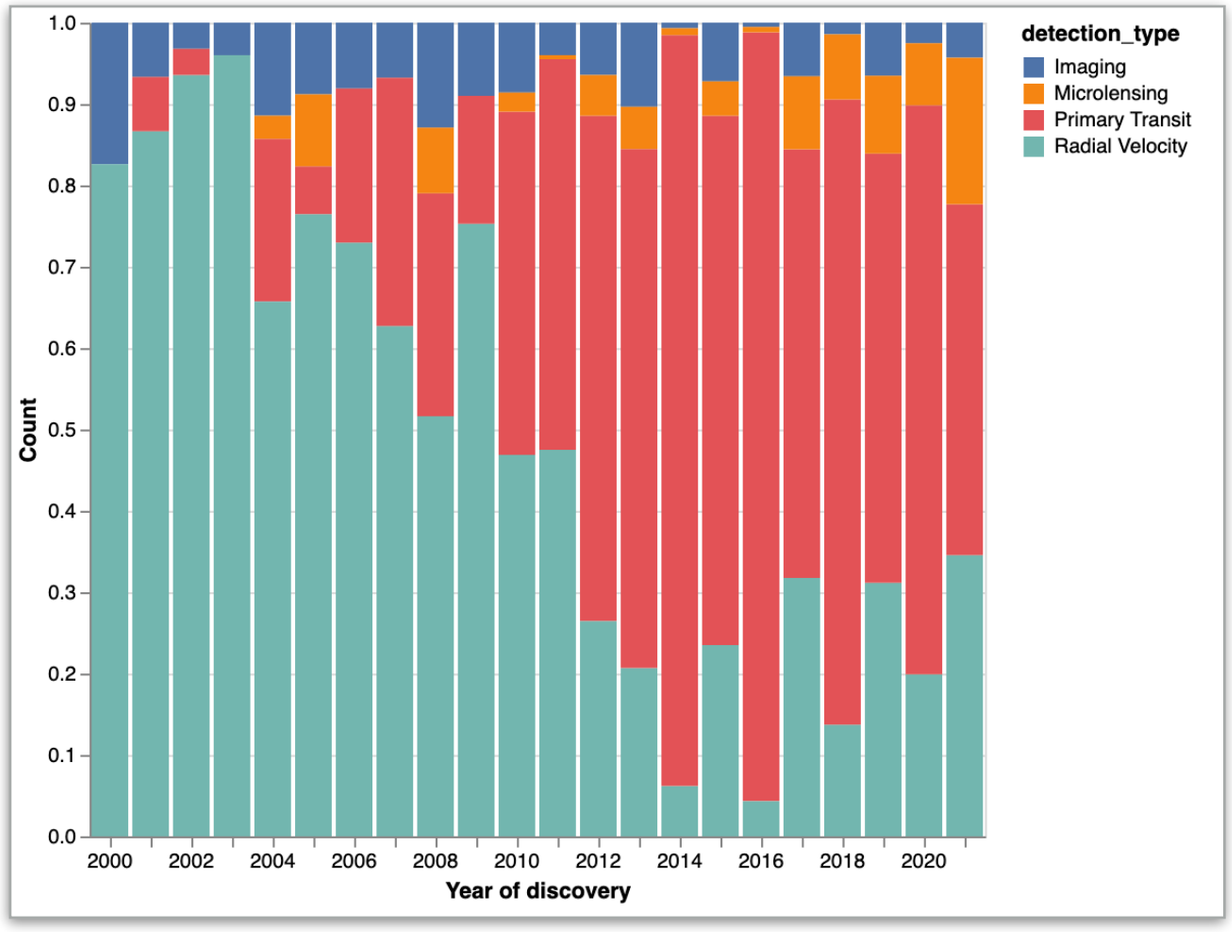
[Source: adapted from InfoVis course, T. Munzner, 2017]

Visualizing multiple distributions

... it does work well for showing counts and how they break down into parts:



absolute count



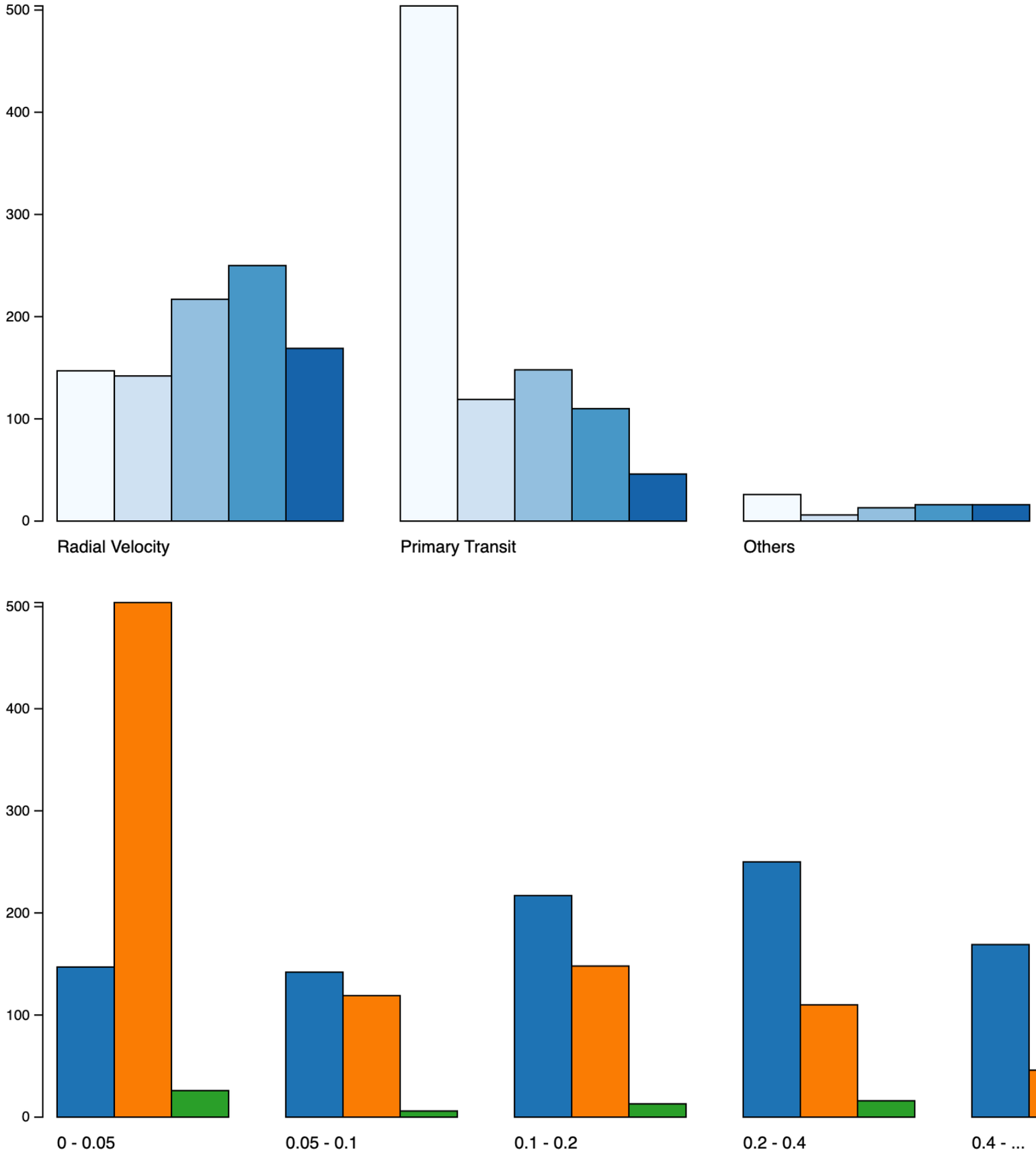
normalized to show proportions

[Data source: exoplanets.eu]

Visualizing multiple distributions

Grouping matters:

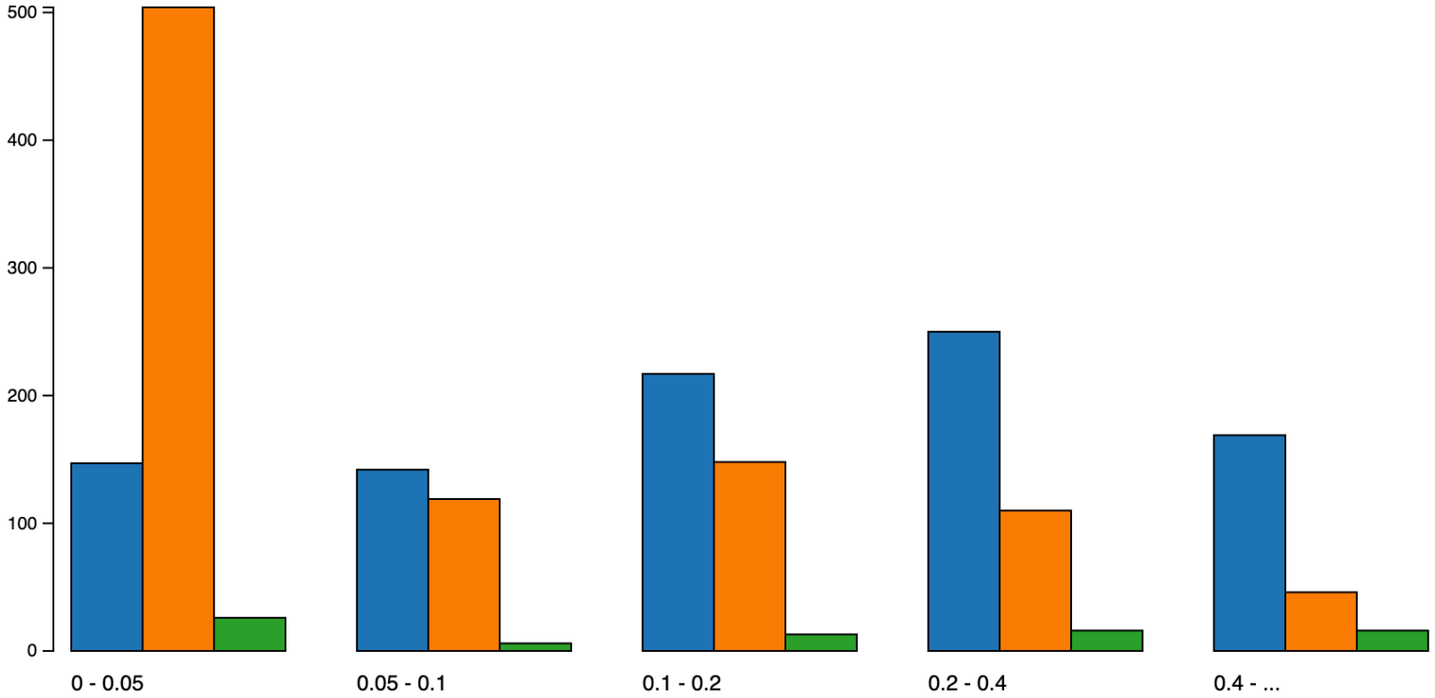
Easier comparison
within detection method

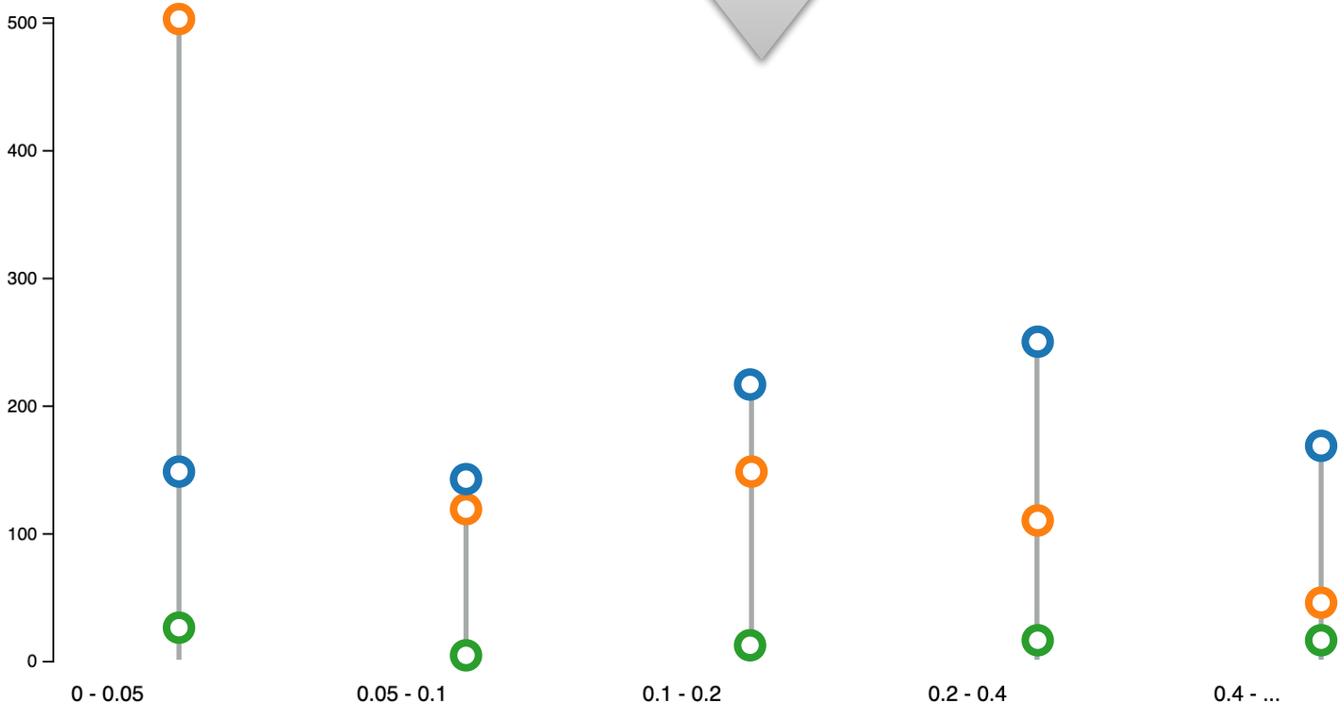


Easier comparison
within eccentricity bin

[Data source: exoplanets.eu]

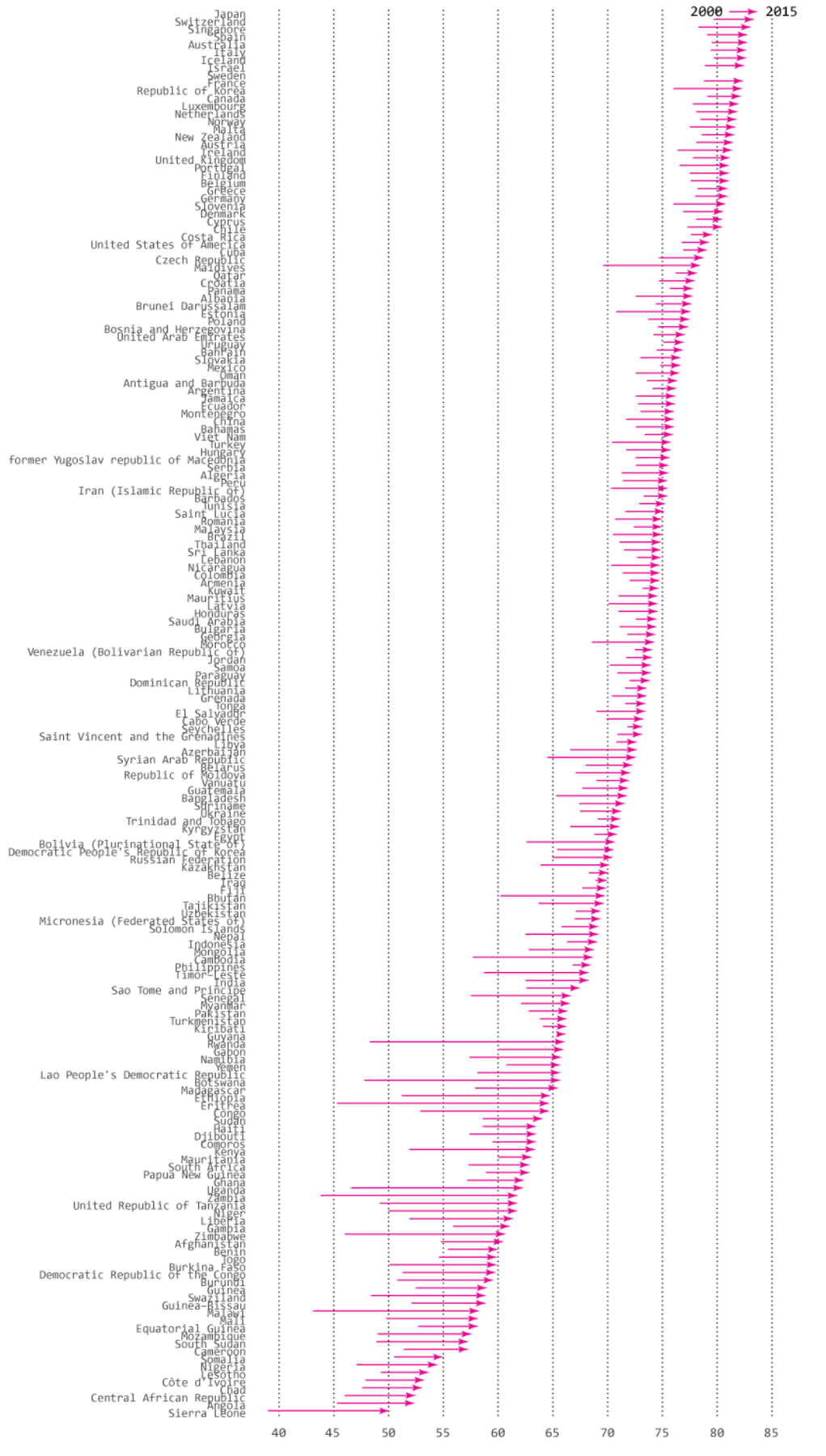
Lollipop Charts



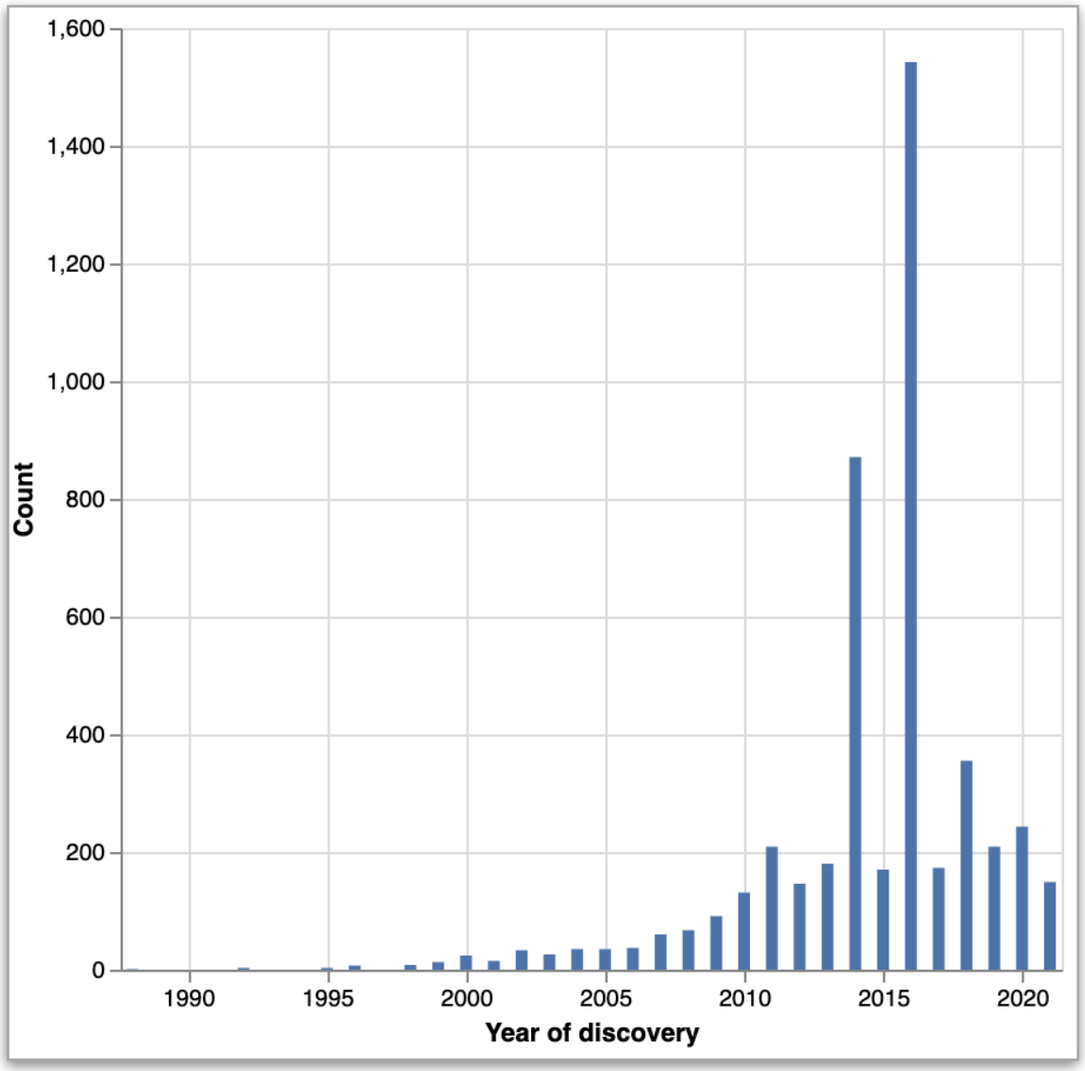


LIFE EXPECTANCY AT BIRTH
2000 vs. 2015

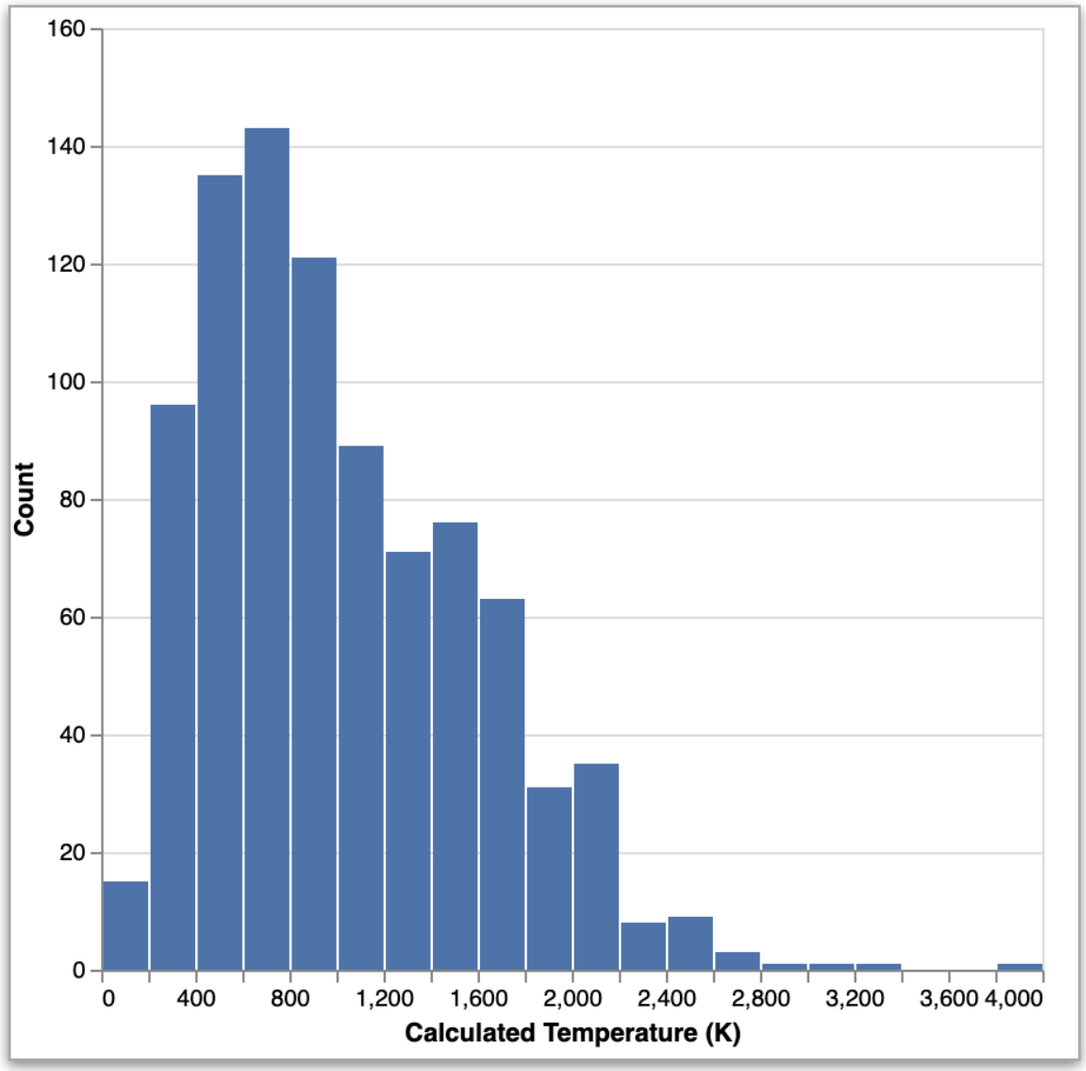
[Nathan Yau, One Dataset, Visualized 25 Ways
http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/]



Visualizing amounts and distributions



Values on a ranked scale:
no binning required



Values on a continuous scale:
binning required

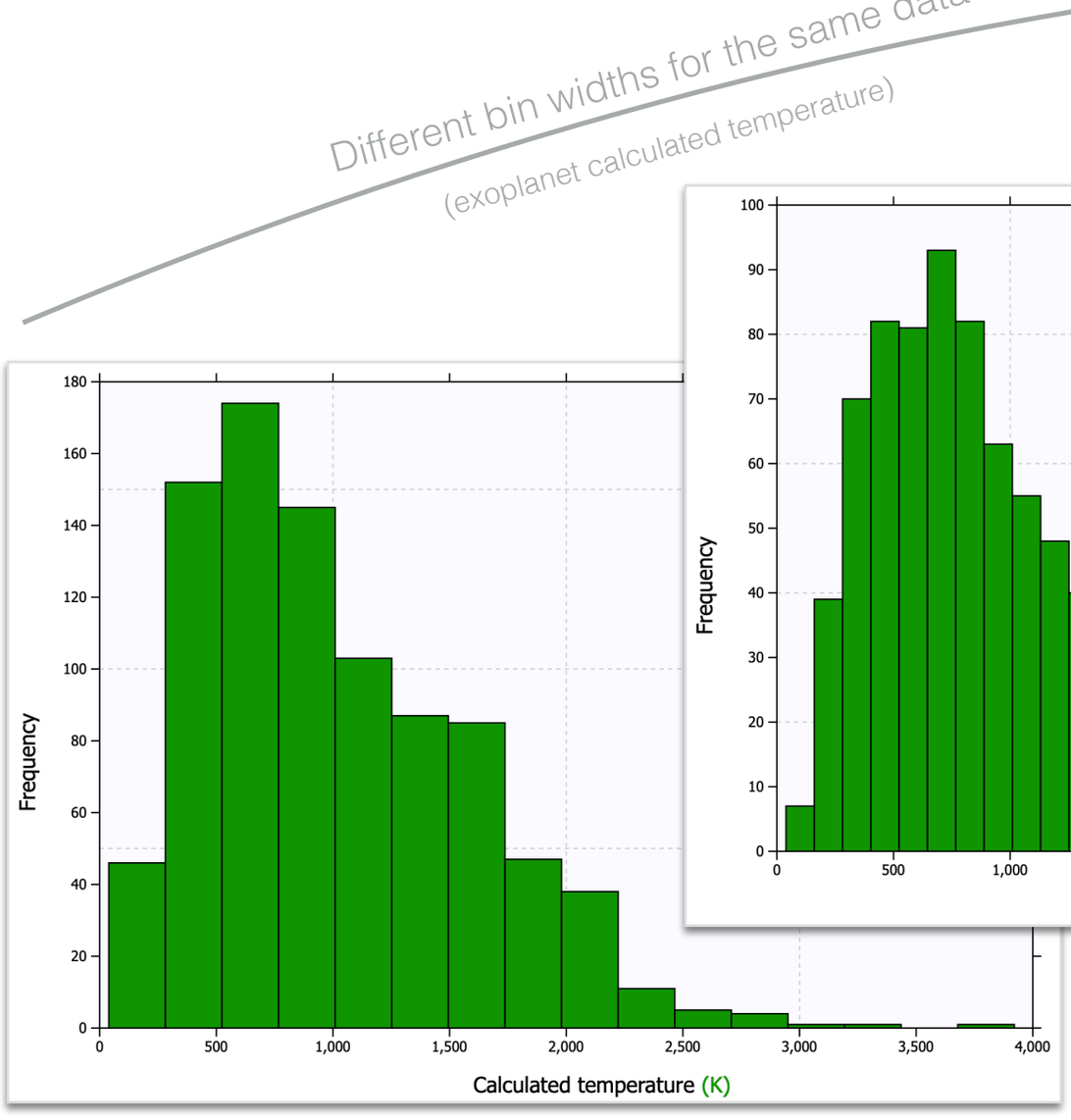
[Data source: exoplanets.eu]

Visualizing amounts and distributions

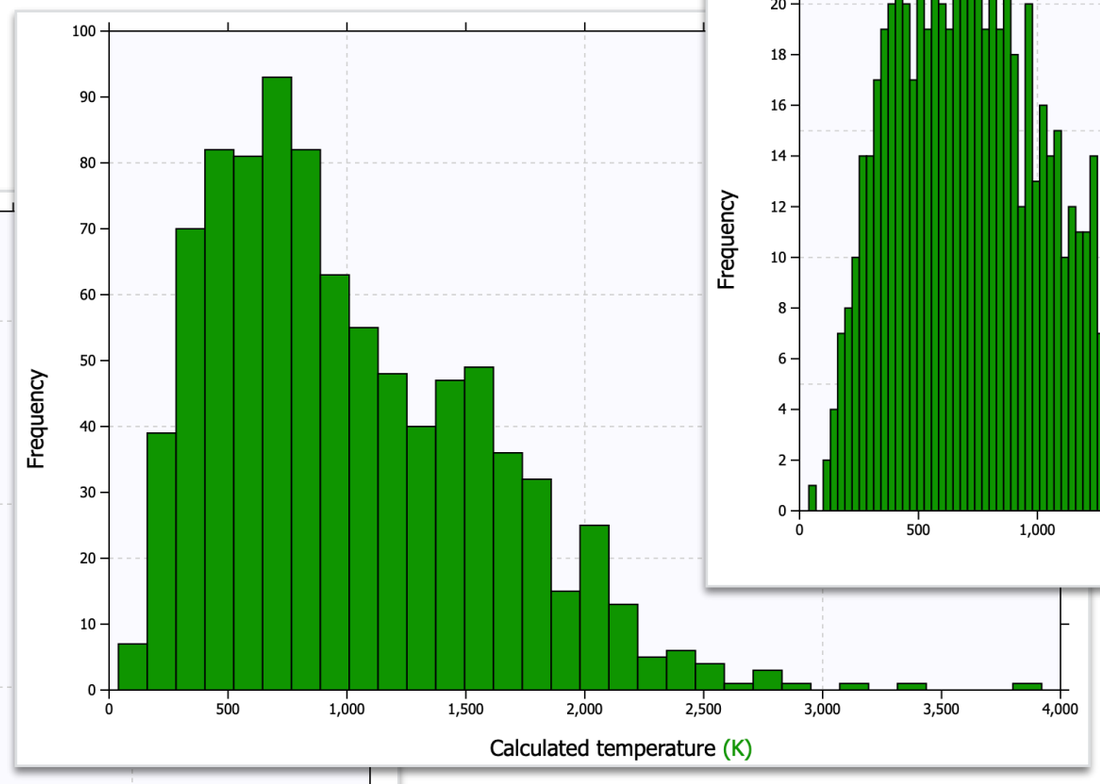
Binning settings matter:

Different bin widths for the same data

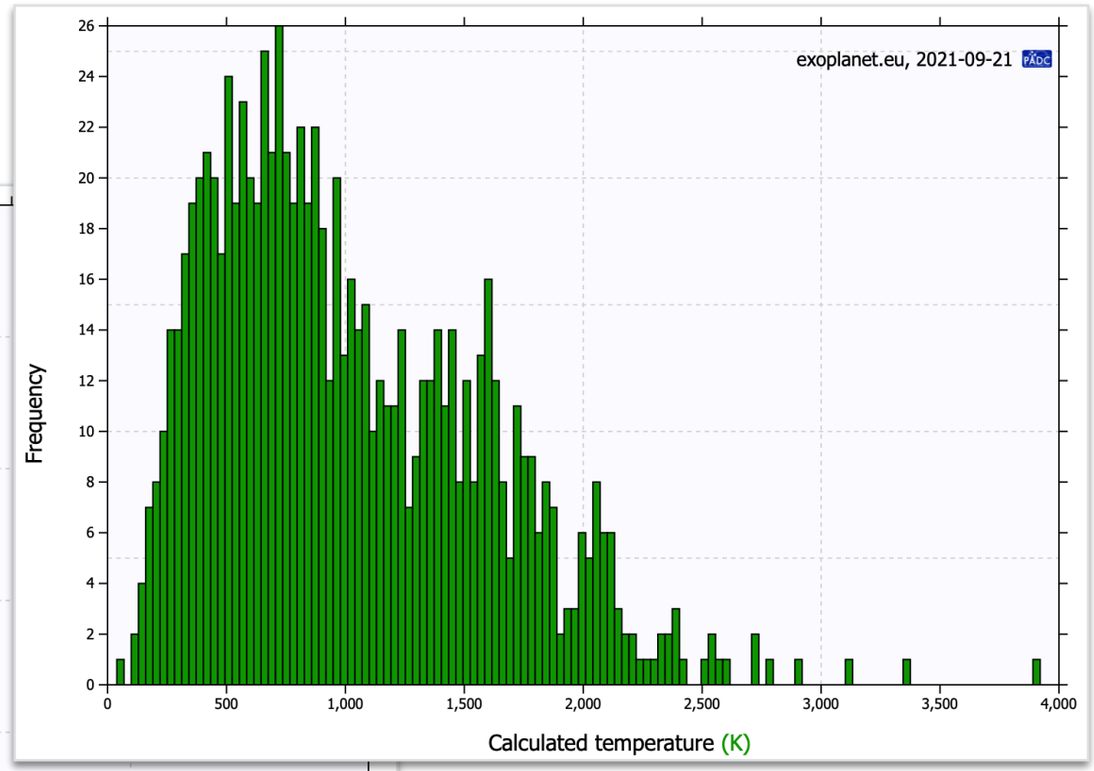
(exoplanet calculated temperature)



16 bins



32 bins



128 bins

A larger bin width can make smaller features disappear (too much smoothing)

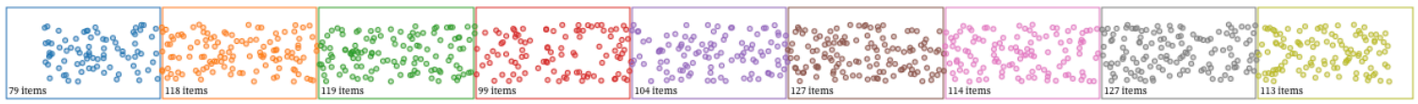
A small bin width can make the chart peaky and obscure the main trends

Optimal-bin-width auto-detect methods; interactive bin-width adjustment by the user

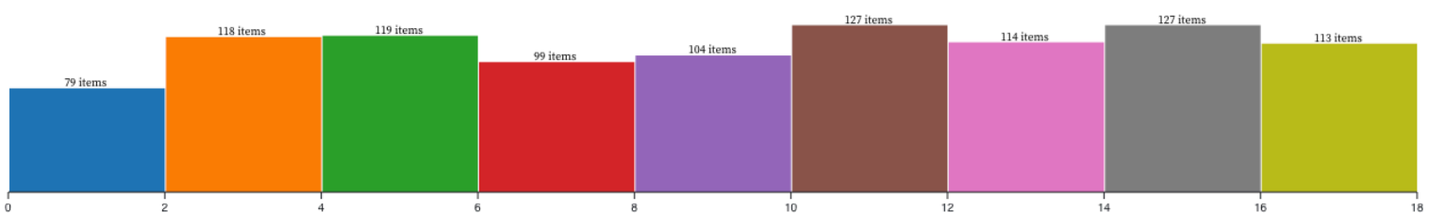
[Data source: exoplanets.eu]

```
buckets1 = ▸Array(9) [Array(79), Array(118), Array(119), Array(99), Array(104), Array(127), Array(114), Array(127), Arra
buckets1 = bin1(values1)
```

We can inspect this result visually, by drawing each bin as a rectangle enclosing the dots it contains. As we can see, all buckets have the same width; they do not necessarily contain the same number of elements:



Histograms allow to quickly grasp the shape of a distribution. The bins returned by d3.bin are customarily represented by a *histogram chart* (or *frequency bar chart*), in which each bar represents a bucket, and the height of each bar figures the number of elements it contains.

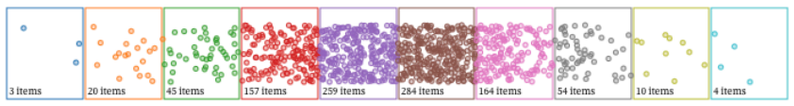


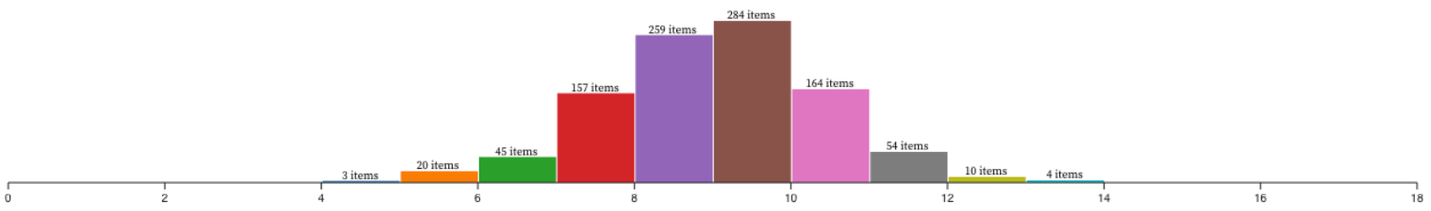
Since we're playing with generated data, let's try a different type of random distribution, and compute new buckets:

Click on the buttons below to try various shapes:

○ Uniform ◉ Normal ○ Two blobs ○ Skewed left ○ Skewed right

Distribution for values2






```
bin2 = ƒ(r)
bin2 = d3.bin()

buckets2 = ▸Array(10) [Array(3), Array(20), Array(45), Array(157), Array(259), Array(284), Array(164), Array(54), Array(
buckets2 = bin2(values2)
```

Parameters

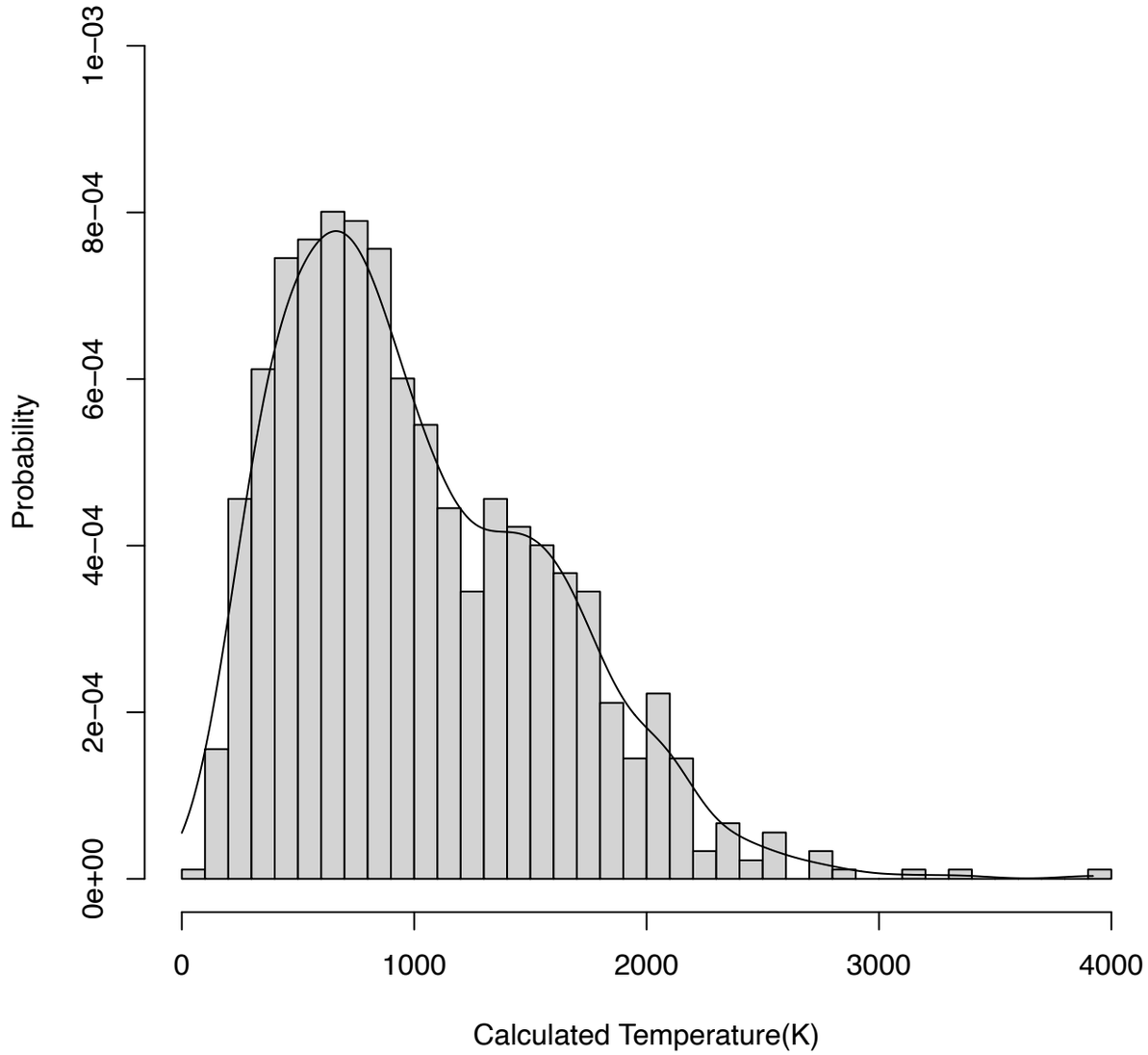
d3.bin doesn't generate the buckets directly: it returns a function that can be parametrized before it is applied to the data to generate the buckets.

With the default parameters, that function knows how to automatically adapt to the data: first it will look at the **value** of each data point; its **domain** will encompass the full extent of these values (from lowest to highest), slice it according to a certain number of equally-sized buckets separated by **thresholds**, and bin the data points into these buckets.

https://observablehq.com/@d3/d3-bin

Visualizing amounts and distributions

Density plots (visualizes the underlying probability distribution of the data)

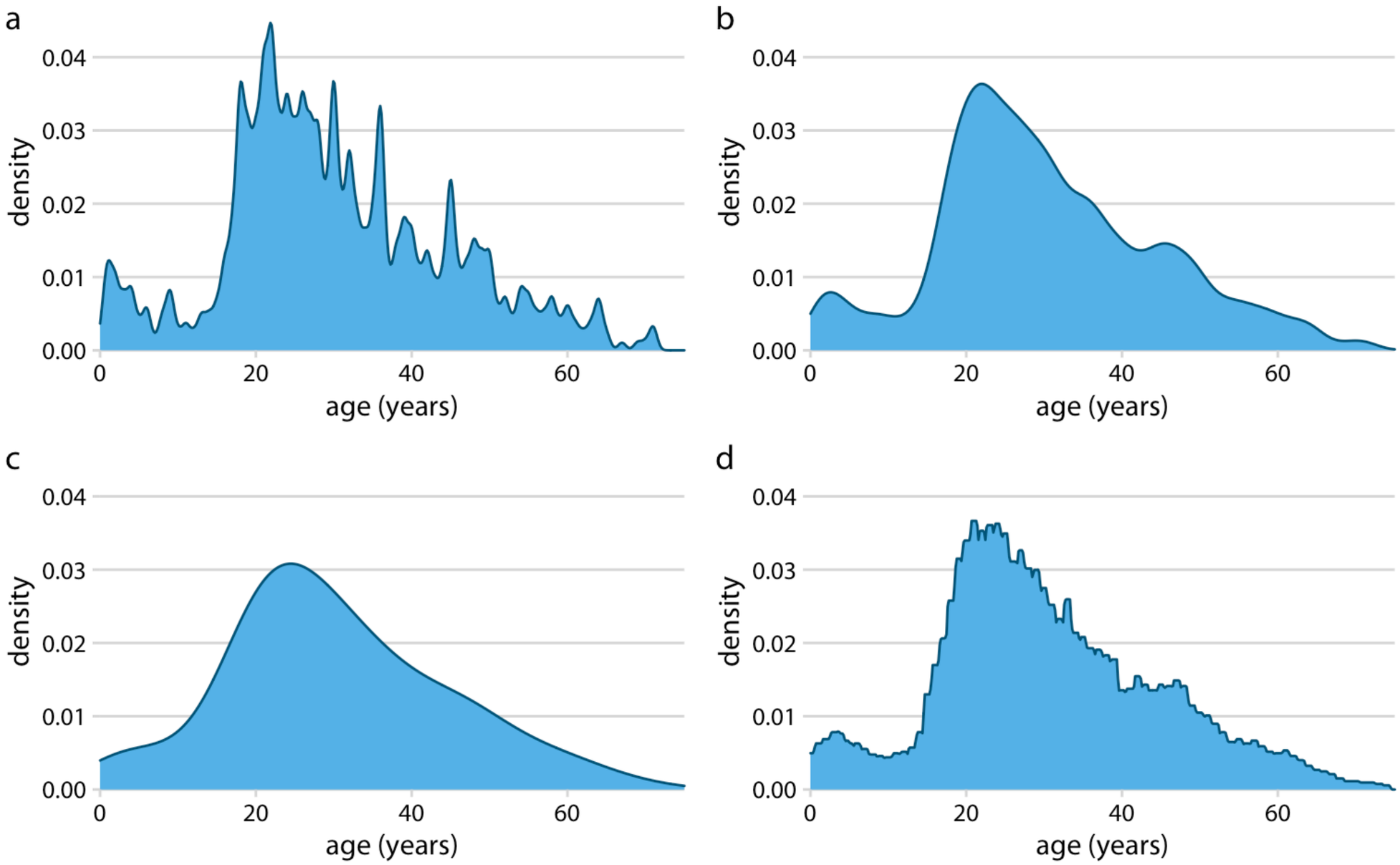


[Data source: exoplanets.eu]

Visualizing amounts and distributions

Density plots (visualizes the underlying probability distribution of the data)

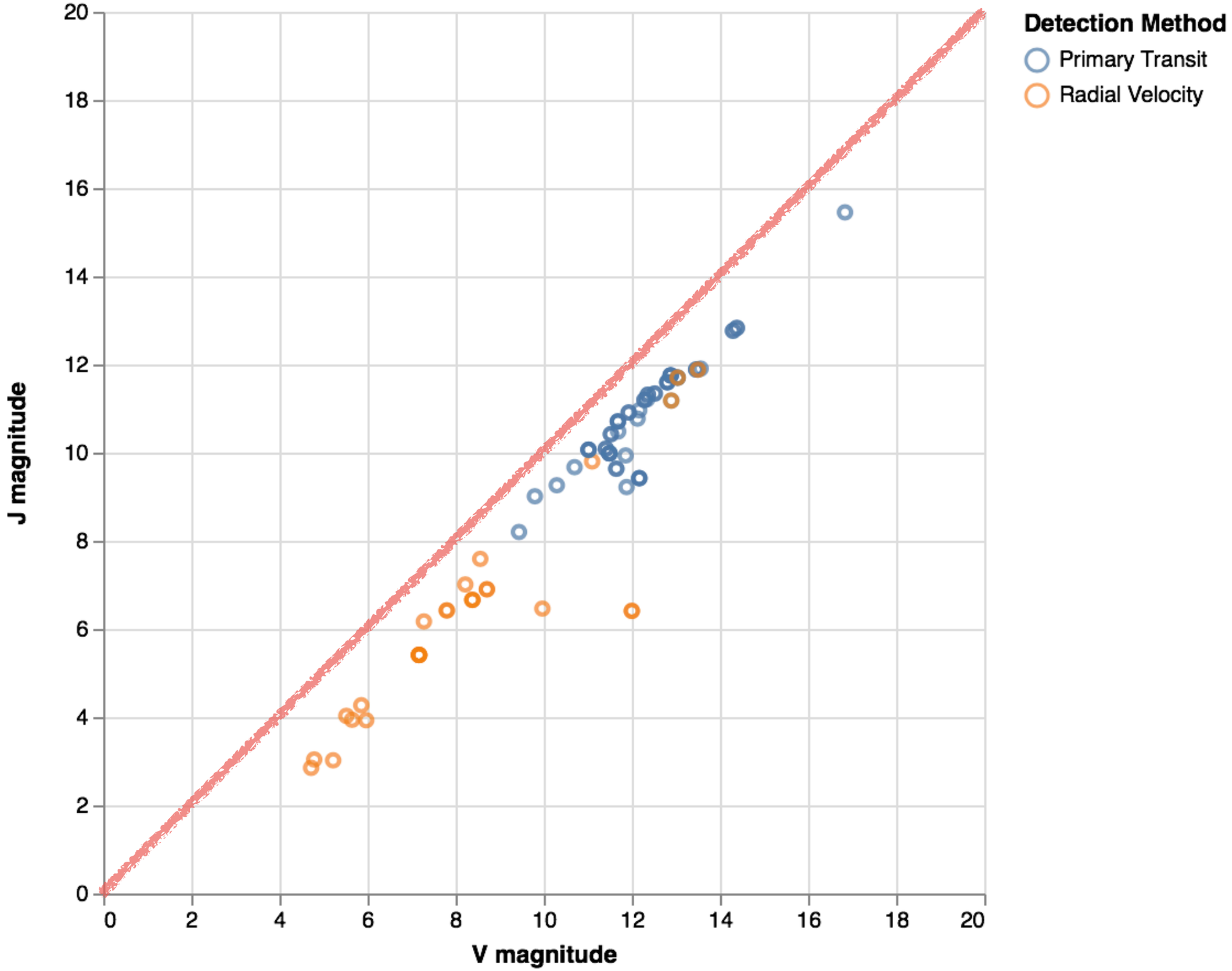
Kernel density estimator parameters (type & bandwidth) matter:



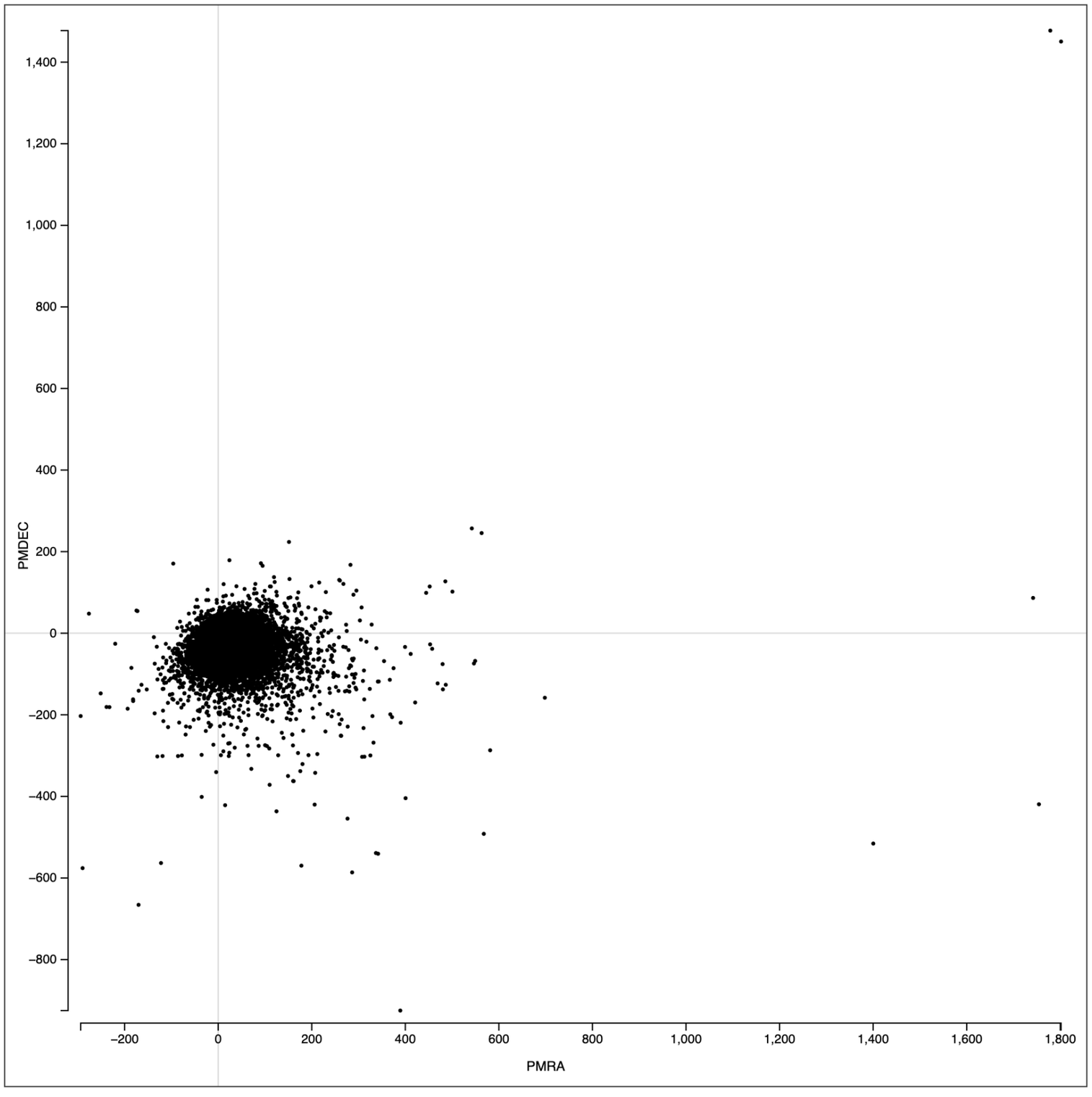
[Example source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Do individual items fall into groups?

Is there a relationship between attributes of items?



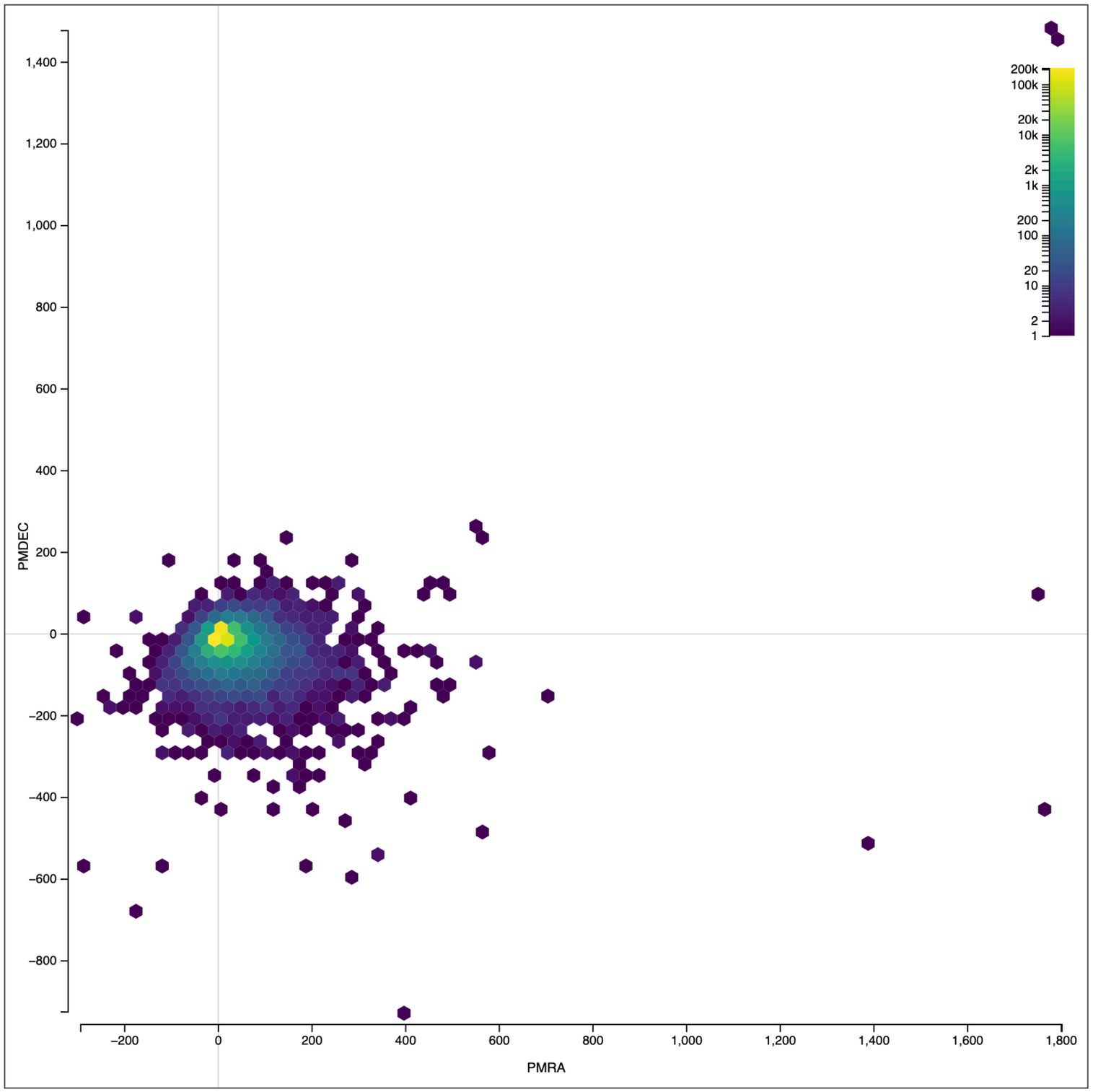
Deal with clutter using sampling, filtering, or binning:



Proper motion for 528,000+ stars

[Data Source: ESA, Gaia Early Data Release 3]

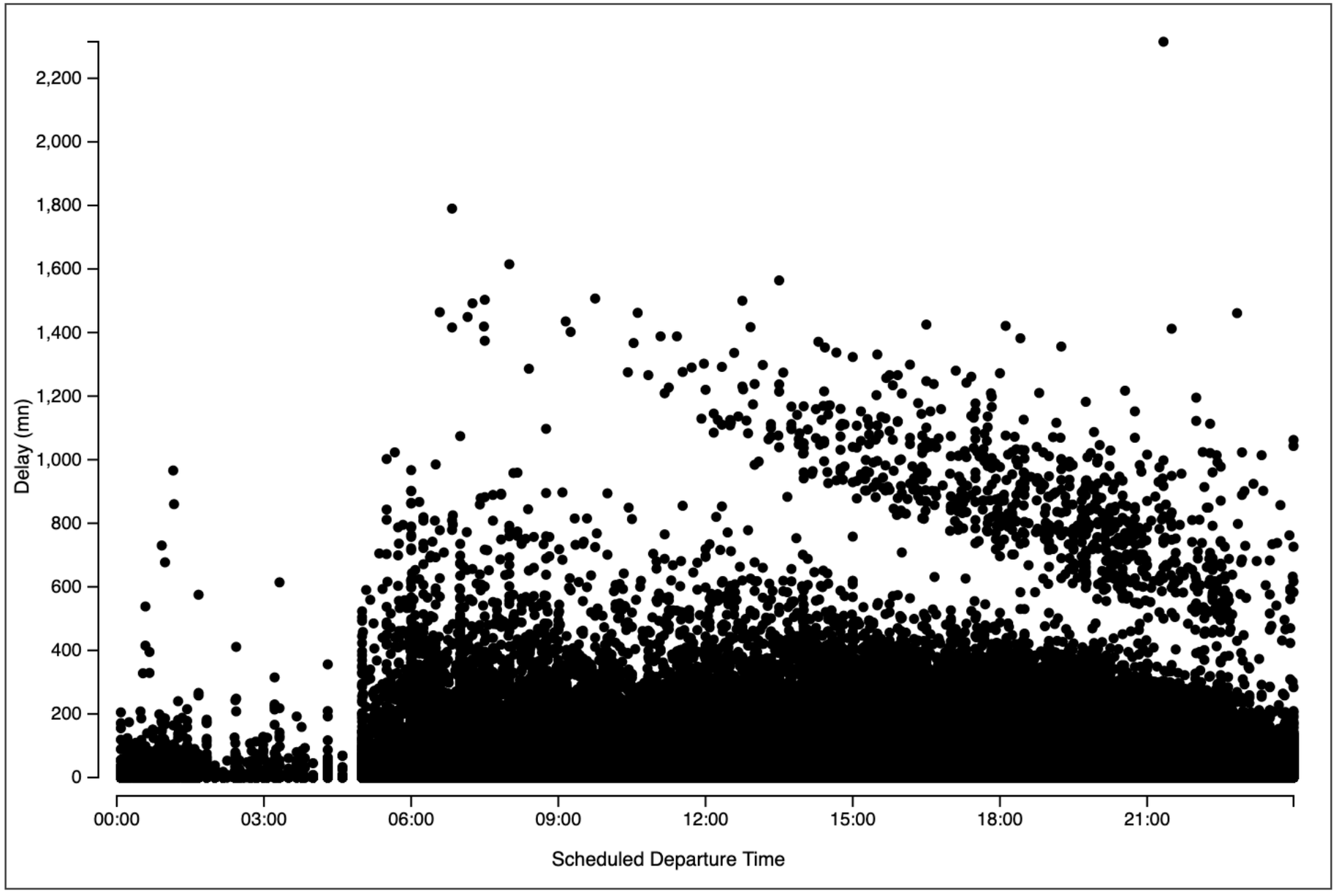
Deal with clutter using sampling, filtering, or binning:



Proper motion for 528,000+ stars

[Data Source: ESA, Gaia Early Data Release 3]

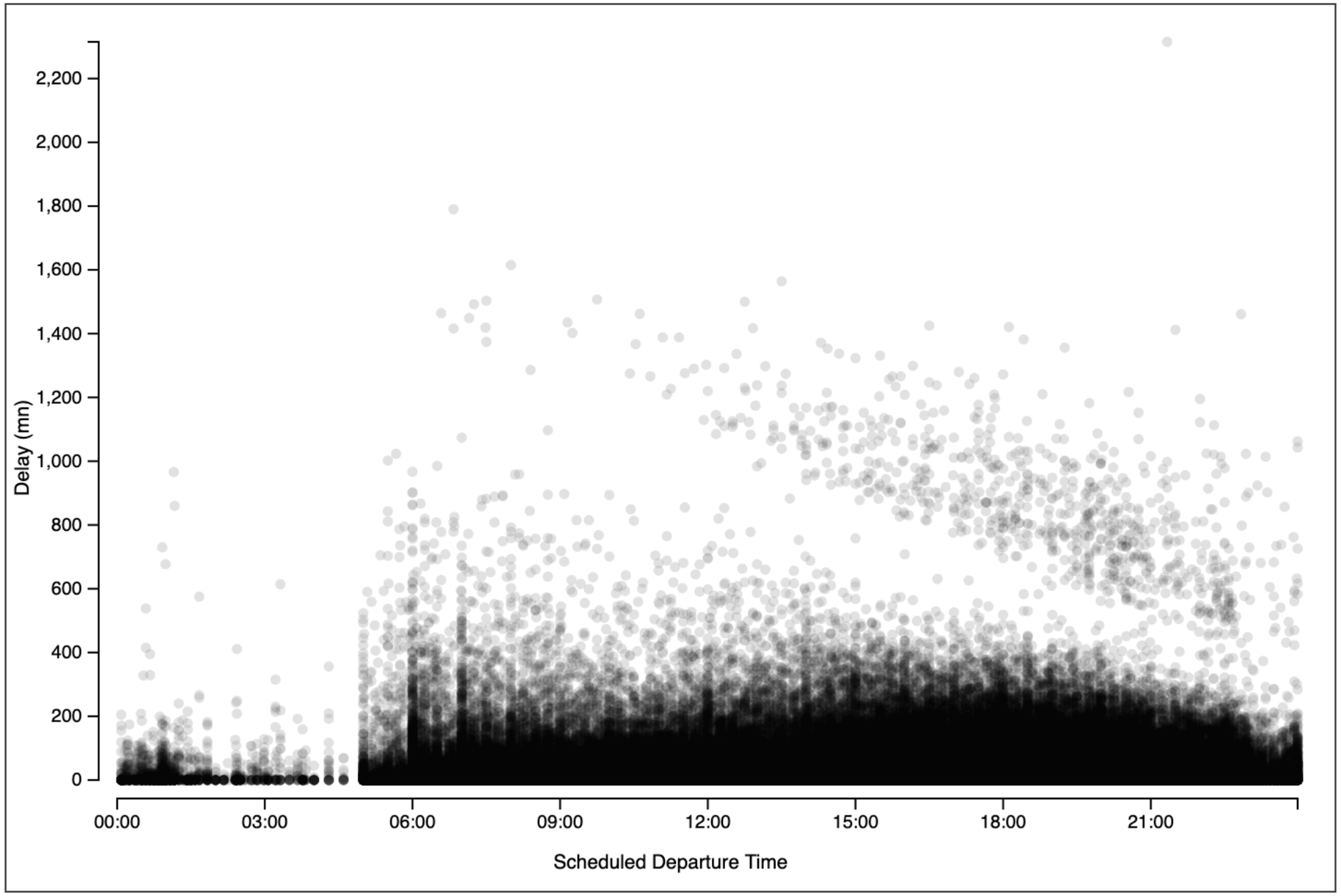
Deal with clutter using sampling, filtering, or binning:



Delay for 659,000+ flights (2015-01)

[Data Source: U.S. Bureau of Transportation Statistics, Airline On-Time Performance Data https://transtats.bts.gov/Tables.asp?QO_VQ=EFD&QO_anzr=Nv4yv0r%FDb0-gvzr%FDcr4s14zn0pr%FDQn6n&QO_fu146_anzr=b0-gvzr]

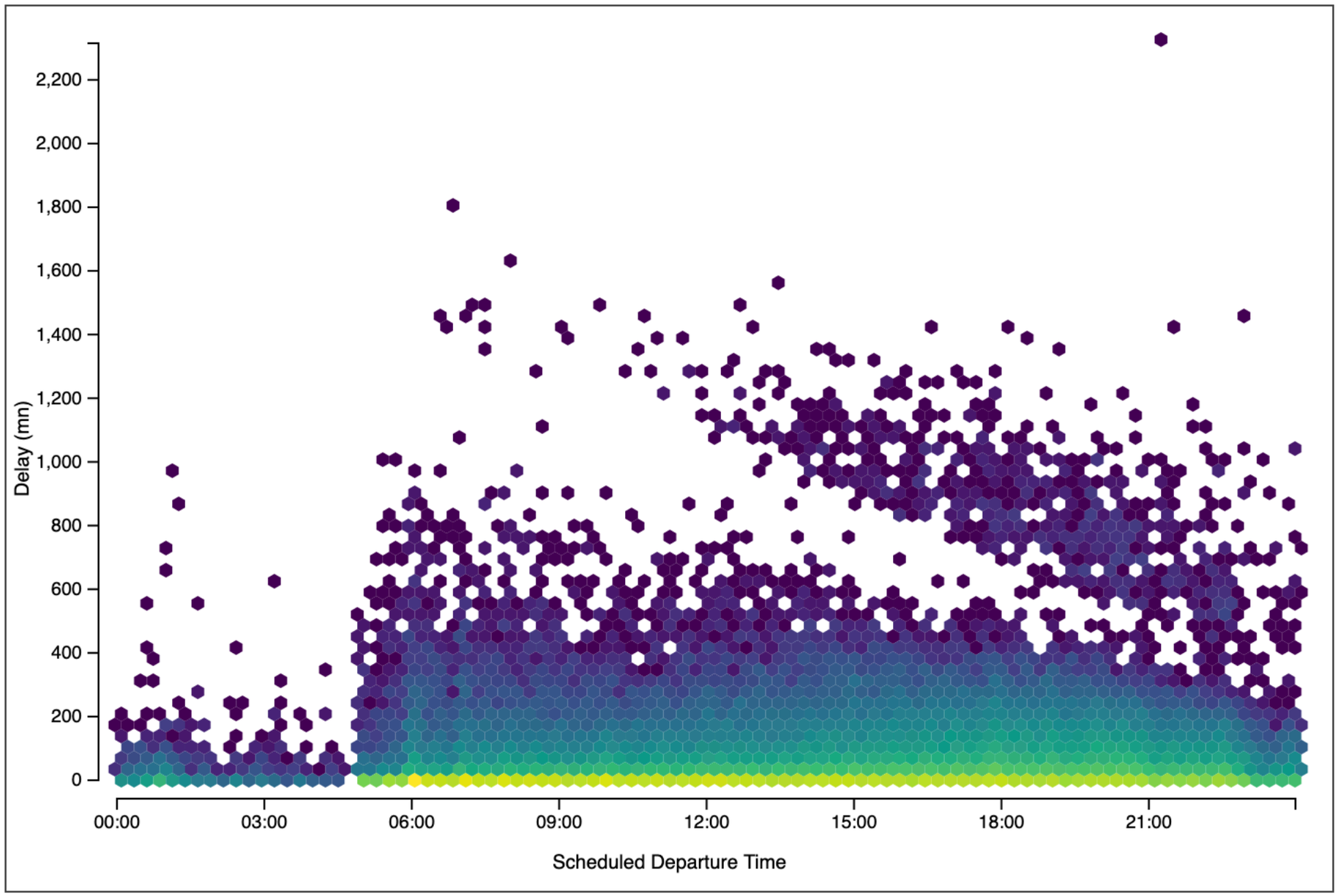
Deal with clutter using sampling, filtering, or binning:



Delay for 659,000+ flights (2015-01)

[Data Source: U.S. Bureau of Transportation Statistics, Airline On-Time Performance Data https://transtats.bts.gov/Tables.asp?QO_VQ=EFD&QO_anzr=Nv4yv0r%FDb0-gvzr%FDcr4s14zn0pr%FDQn6n&QO_fu146_anzr=b0-gvzr]

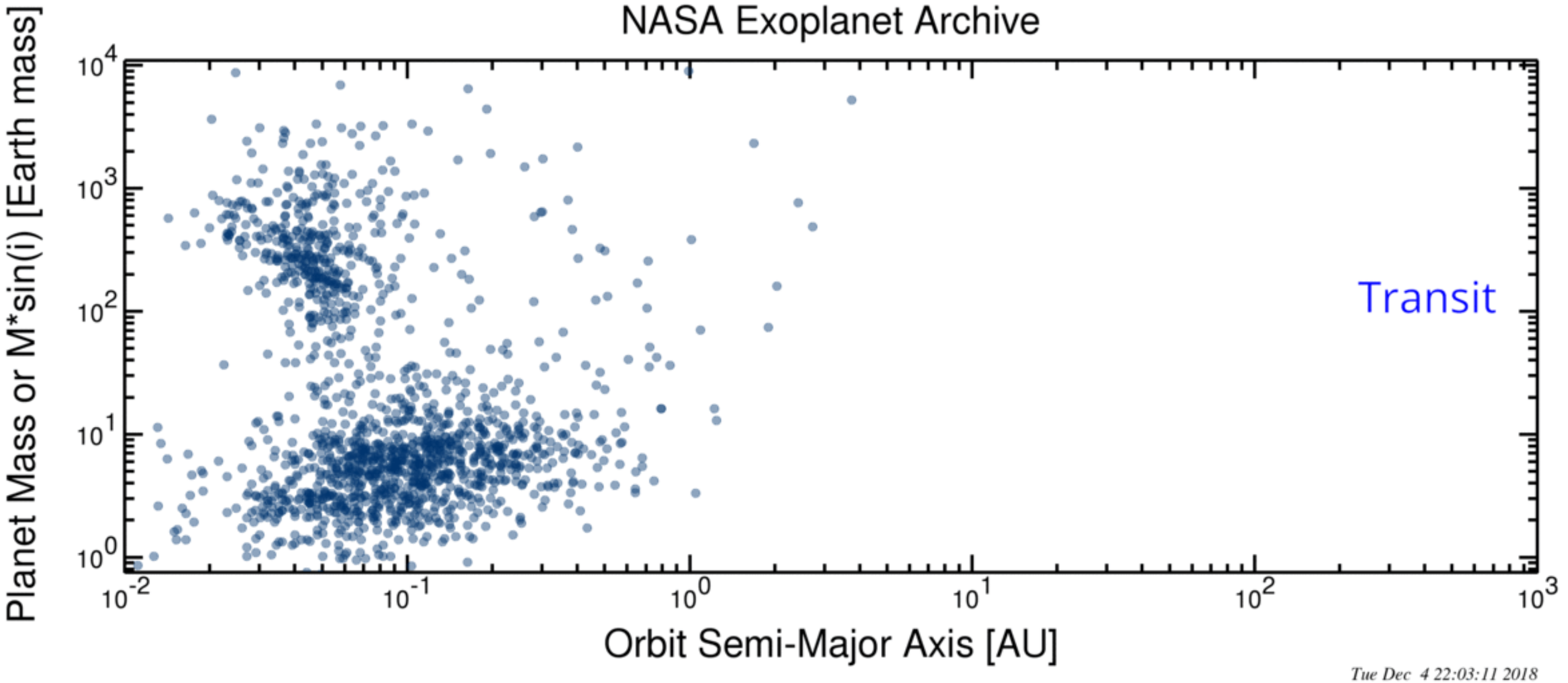
Deal with clutter using sampling, filtering, or binning:



Delay for 659,000+ flights (2015-01)

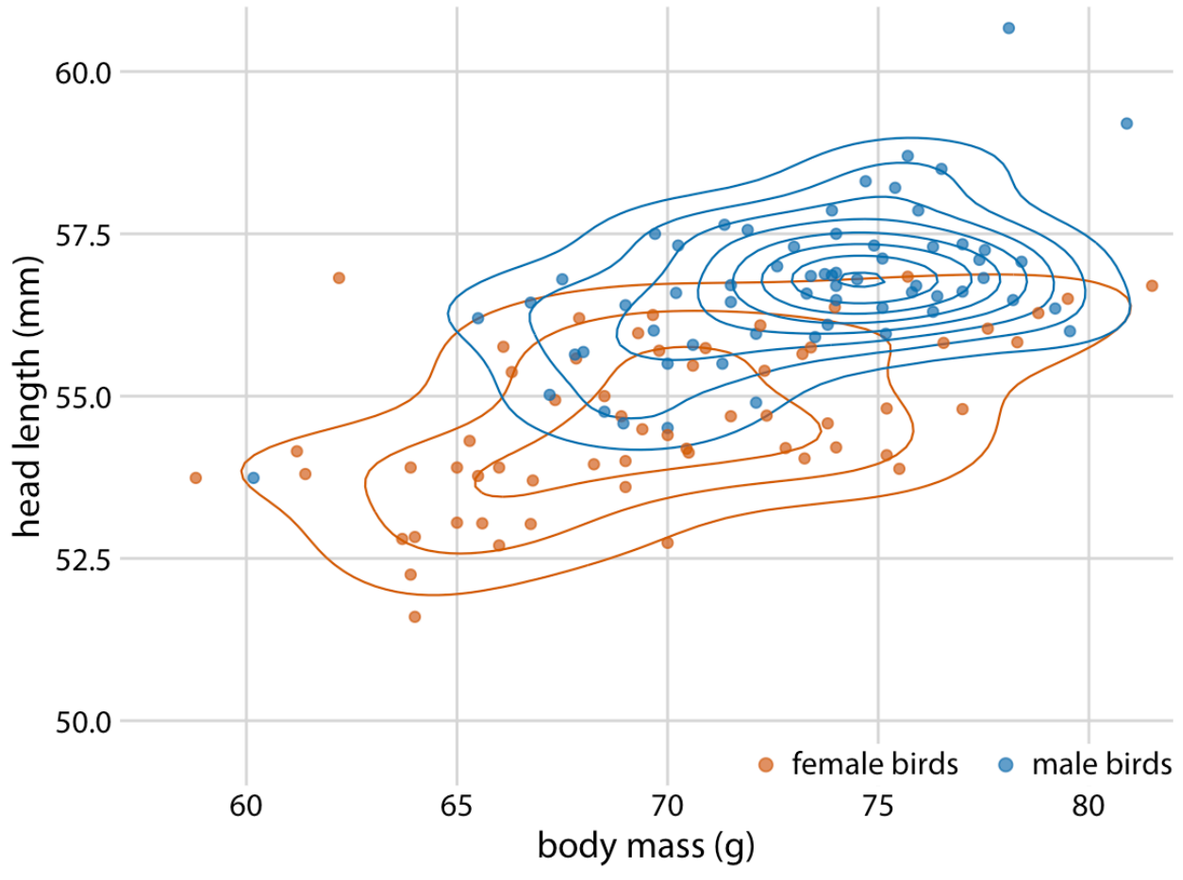
[Data Source: U.S. Bureau of Transportation Statistics, Airline On-Time Performance Data https://transtats.bts.gov/Tables.asp?QO_VQ=EFD&QO_anzr=Nv4yv0r%FDb0-gvzr%FDcr4s14zn0pr%FDQn6n&QO_fu146_anzr=b0-gvzr]

Deal with clutter using sampling, filtering, or binning,
or some more exotic method (dealing with multiple categories)…:



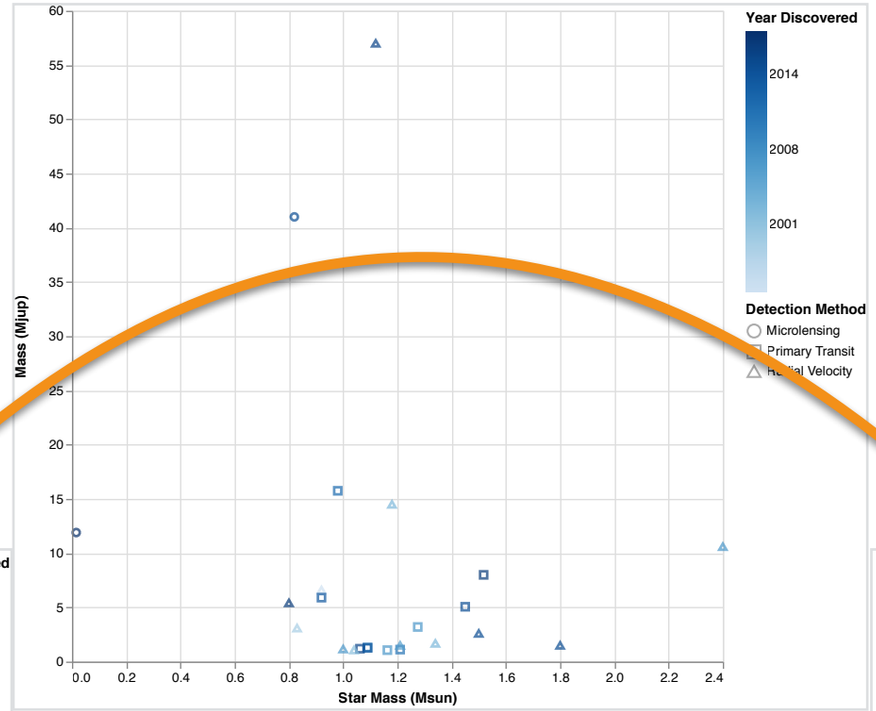
Deal with clutter using sampling, filtering, or binning,
or some more exotic method (dealing with multiple categories)…:

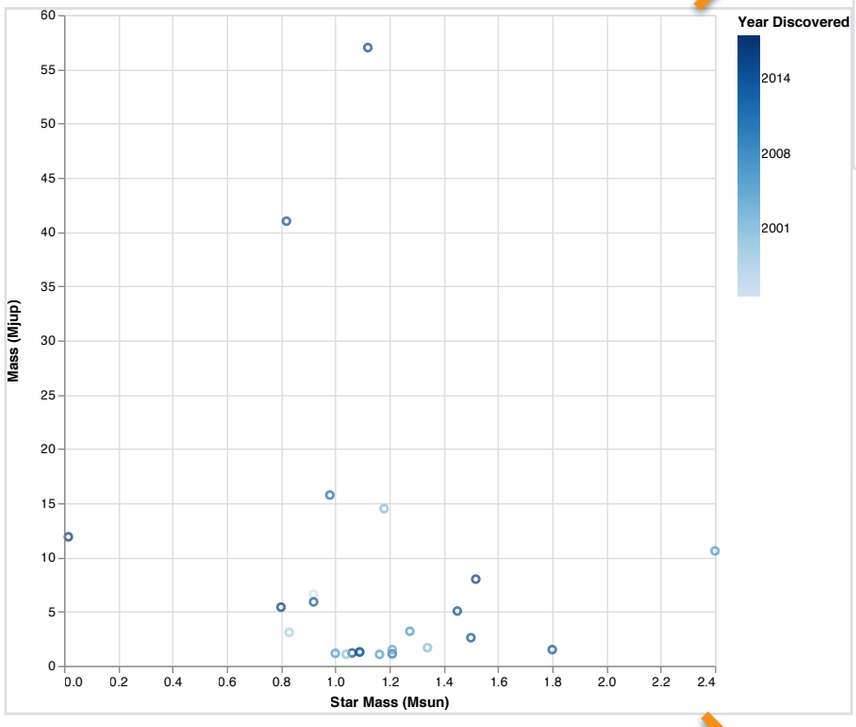
Contour lines delineate regions of similar point density.
Works well when the point density changes slowly along both x & y.
Works well for only a few categories, if the groups are spatially distinct.

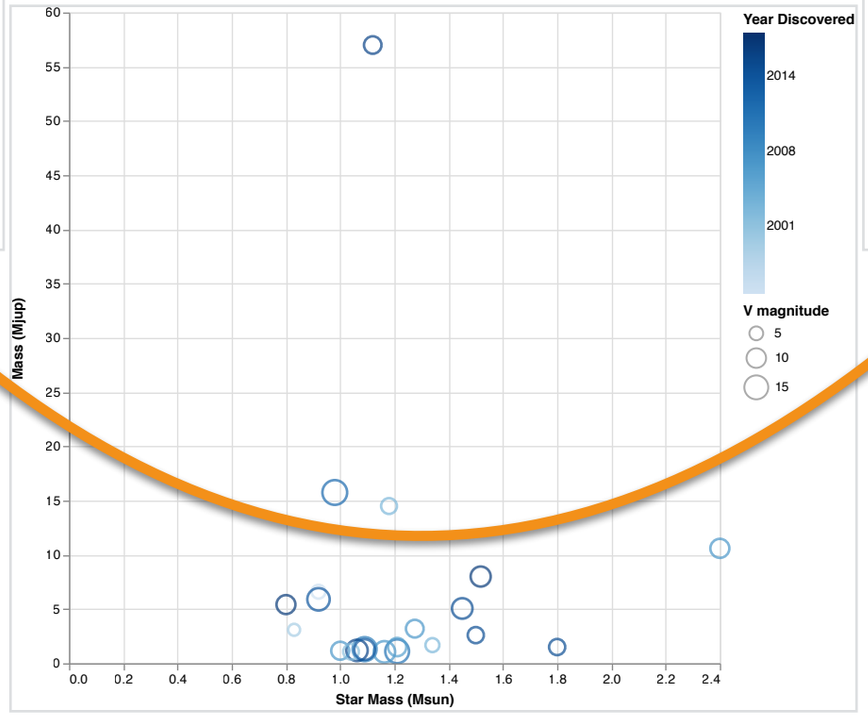


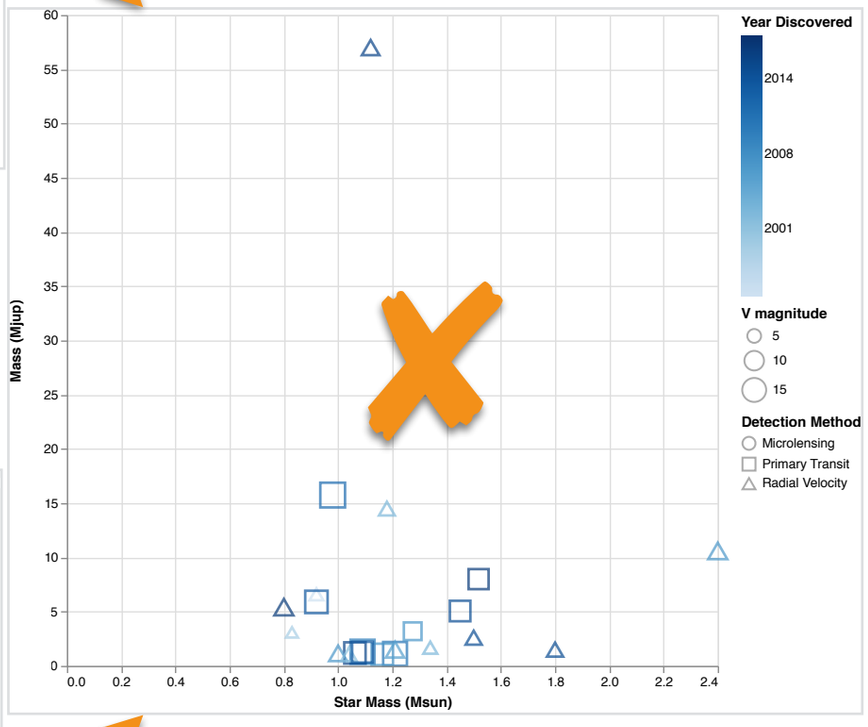
[Example source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

Adding more mappings…

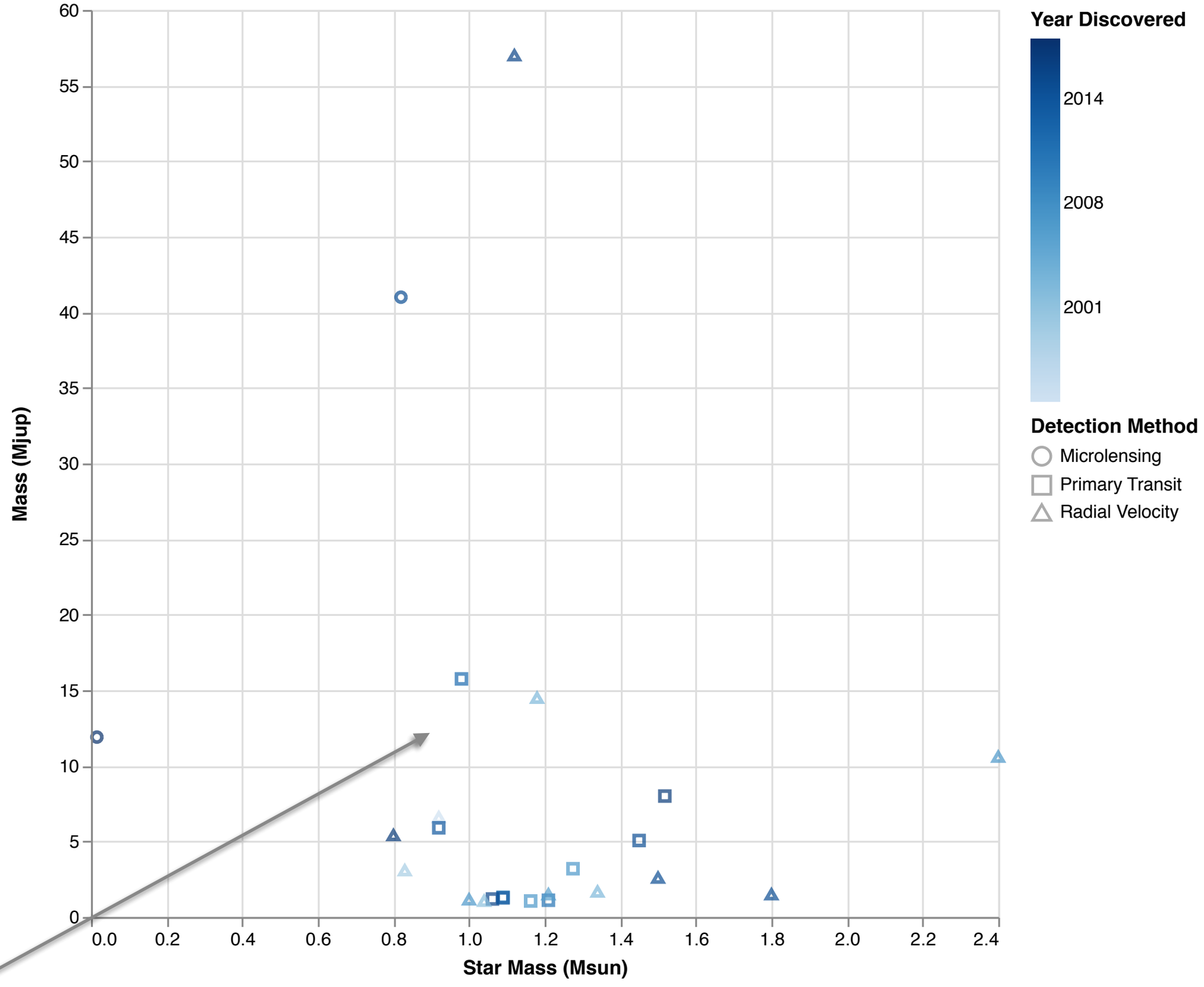






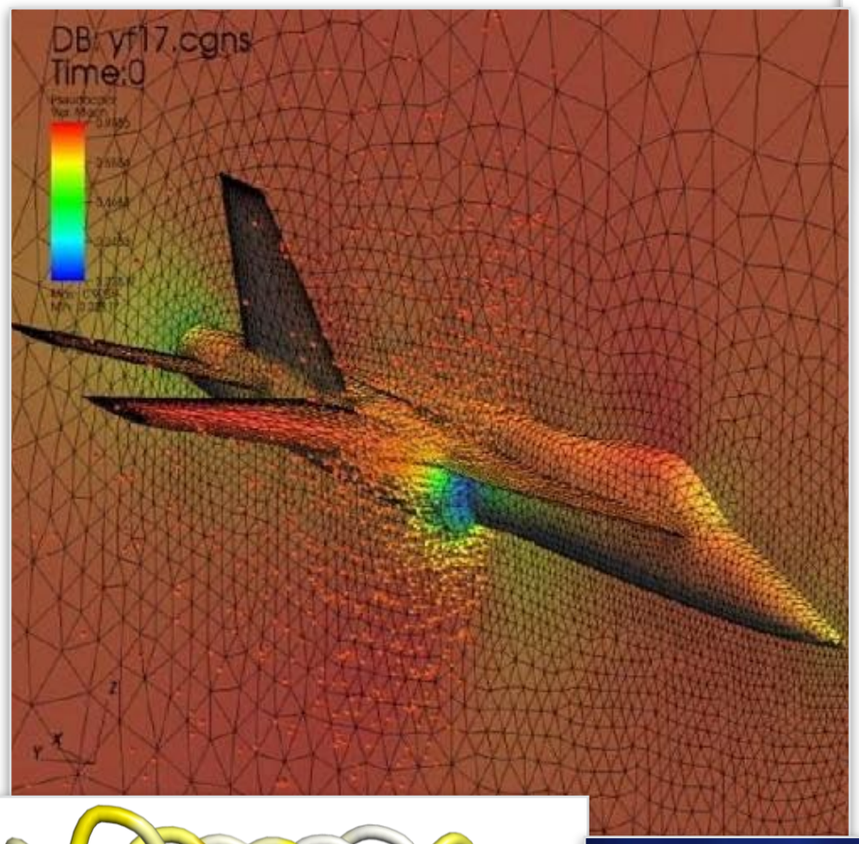


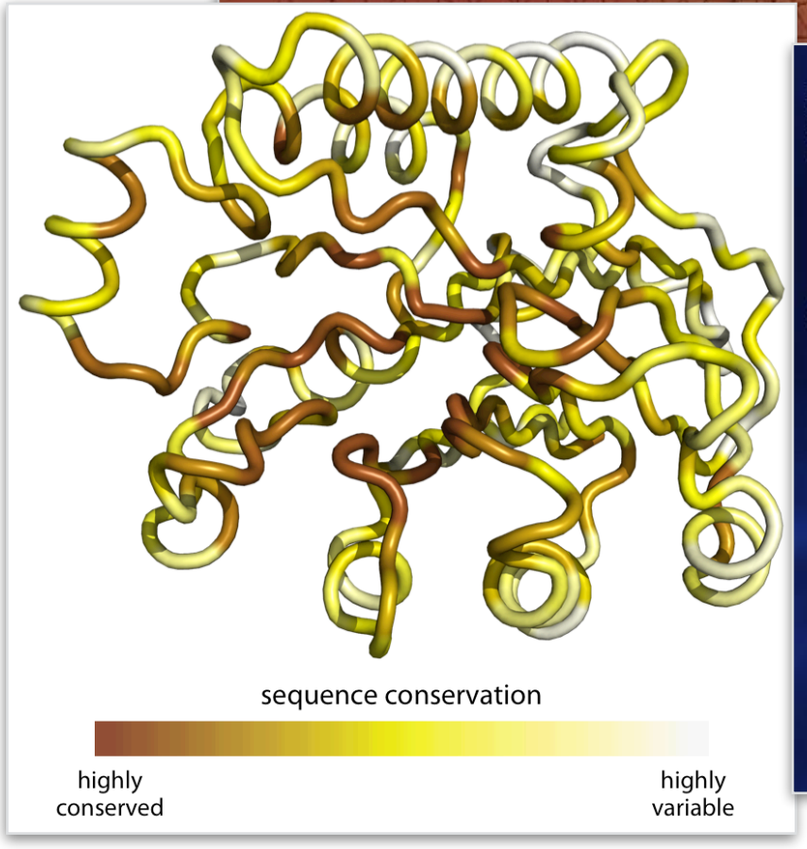
Adding more mappings…

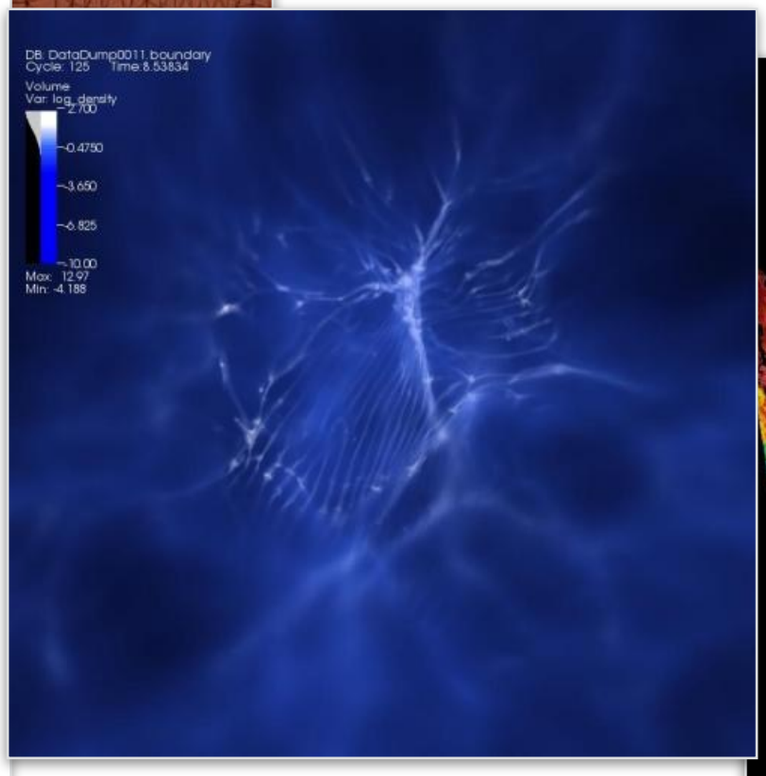


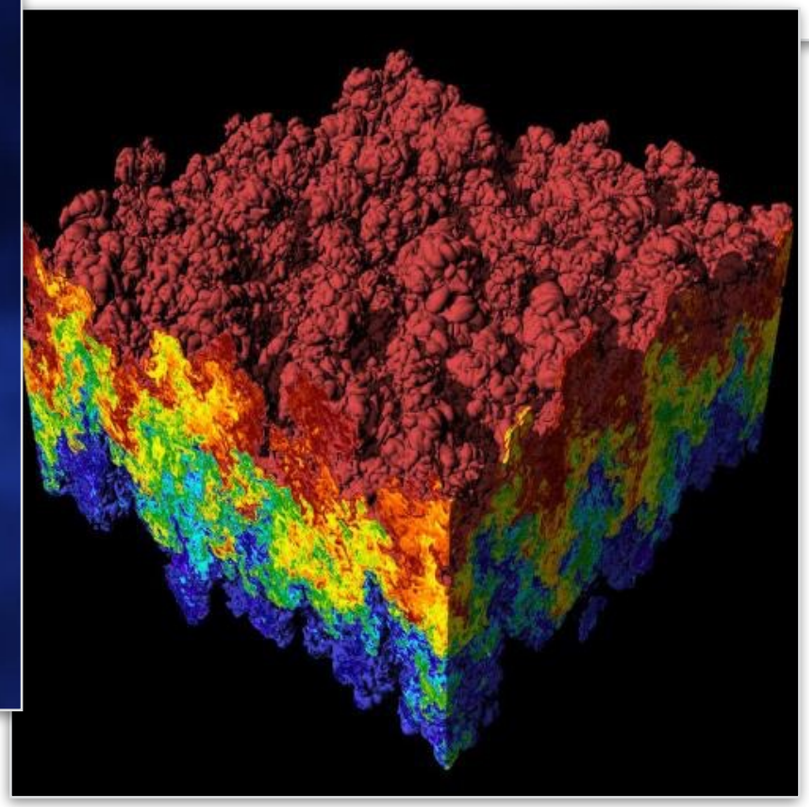
3rd axis ?

Appropriate use of 3D visualization

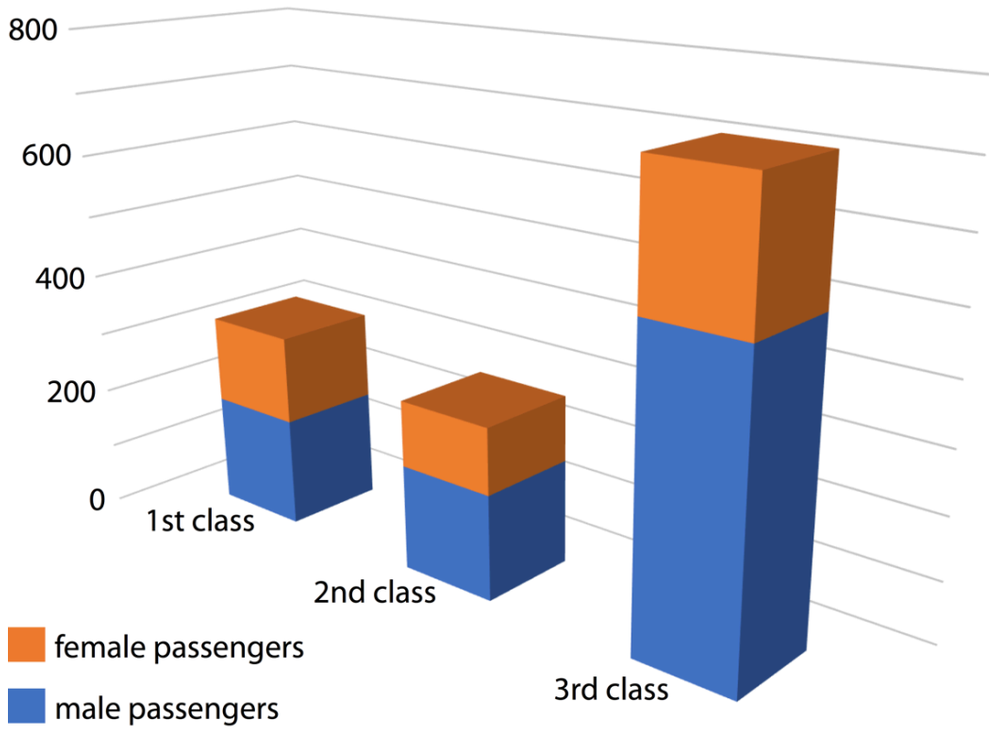


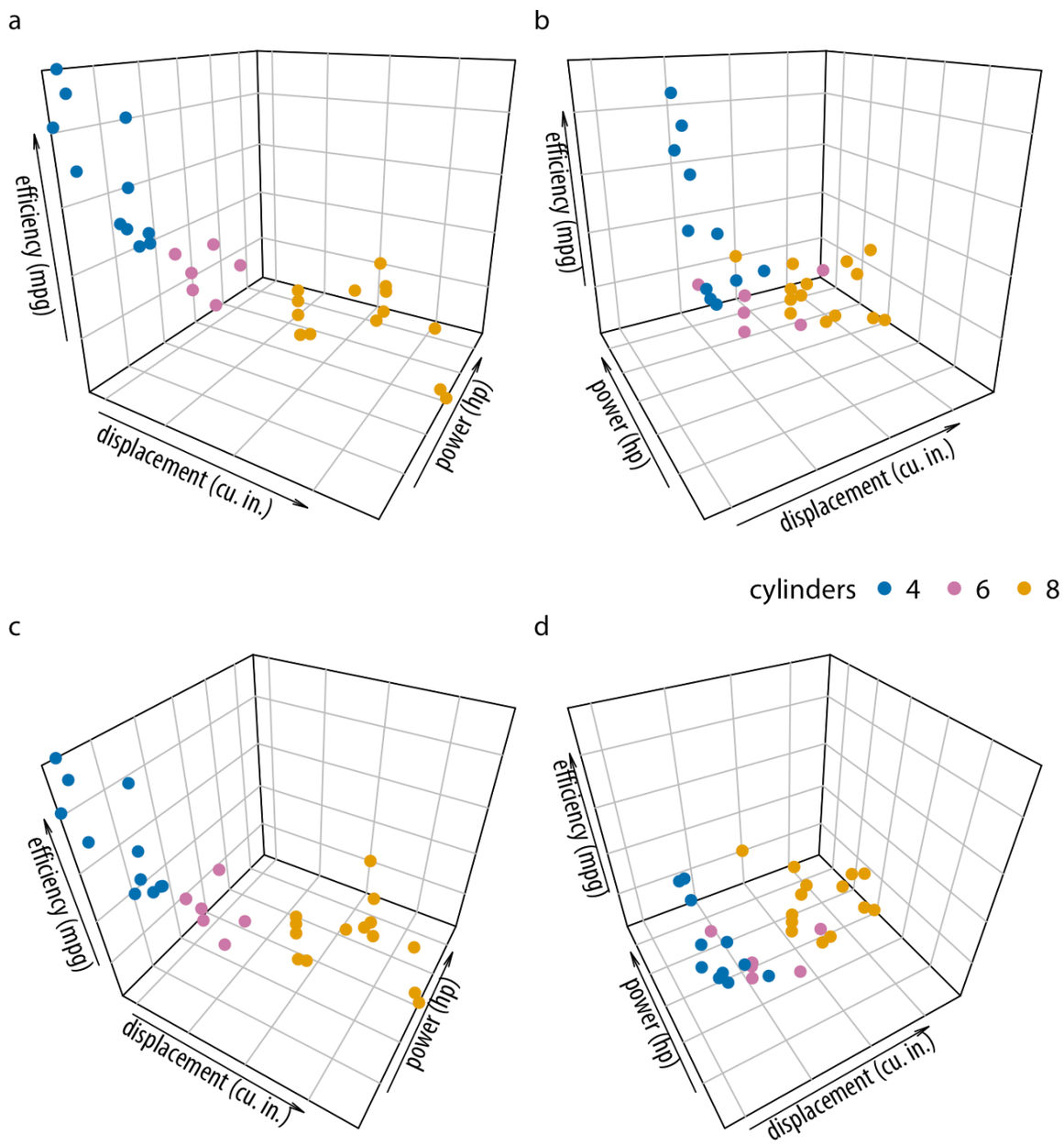


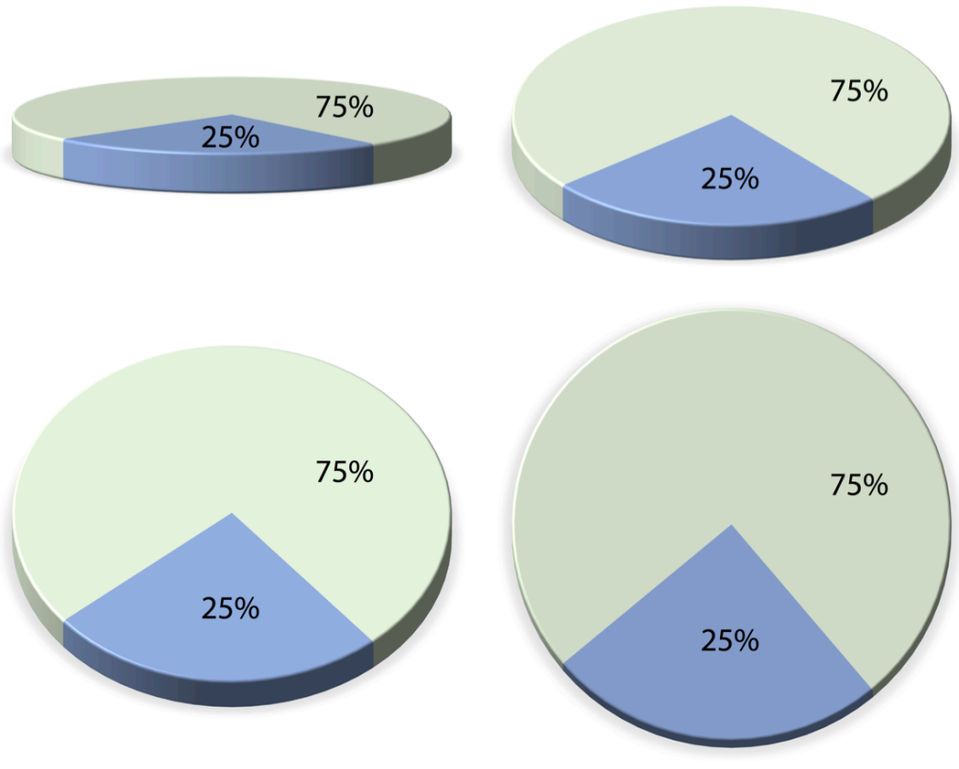




Gratuitous use of 3D visualization







[Example source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

RULE #1

Actual VR/AR immersion is not always necessary.

Analytics can still be situated, immersive, and ubiquitous even without 3D immersion.

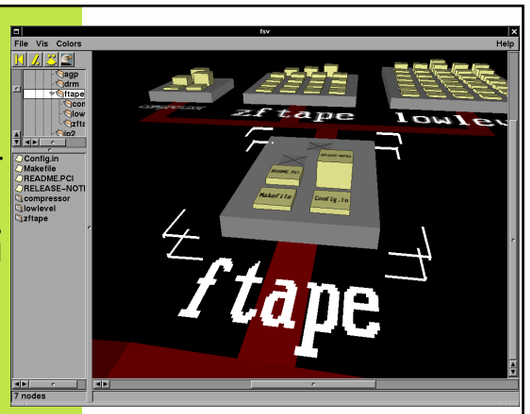
RULE #2

No unmotivated 3D.

3D ≠ 2D + 1.

Use 3D for 3D data, 2D for 2D data (and 1D for 1D data).

(And 1D or 2D for multidimensional…)

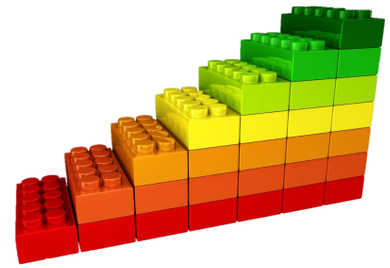


FSV (File System Visualizer)

RULE #3

3D perspective is not your friend.

Integrating visuals in the real world requires it, but be aware.



RULE #4

Occlusion is a major depth cue, but is also a major pain.

Find ways to mitigate this effect.

(Hint: Elmqvist 2006!)

RULE #5

3D navigation is hard!

Don't ask people to navigate in full 6DOF.

Physical navigation is nice (already experts).

RULE #6

Tilted 3D text is not legible.

Doesn't matter if it looks cool, you can't do it.

Make sure labels always face the user.

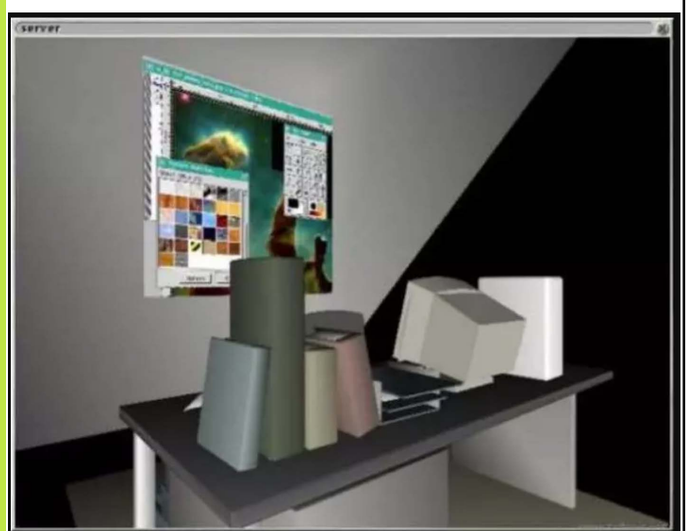
3Dwm (sorry!)

RULE #7

Don't try to replicate the real world…

…unless necessary.

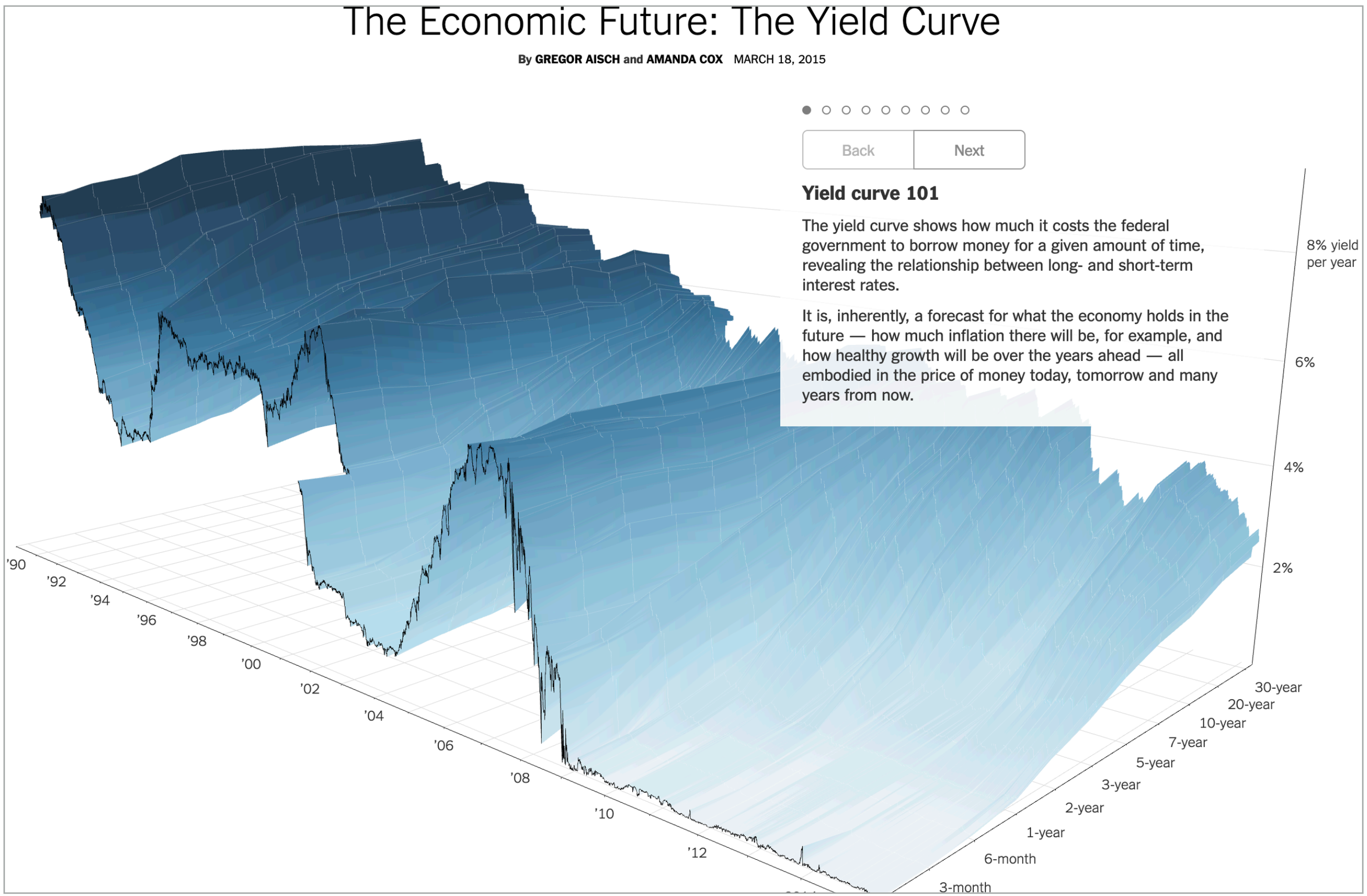
We use computers to augment our abilities, rebuild the real world and our limitations all over again?



3Dwm (sorry again!)

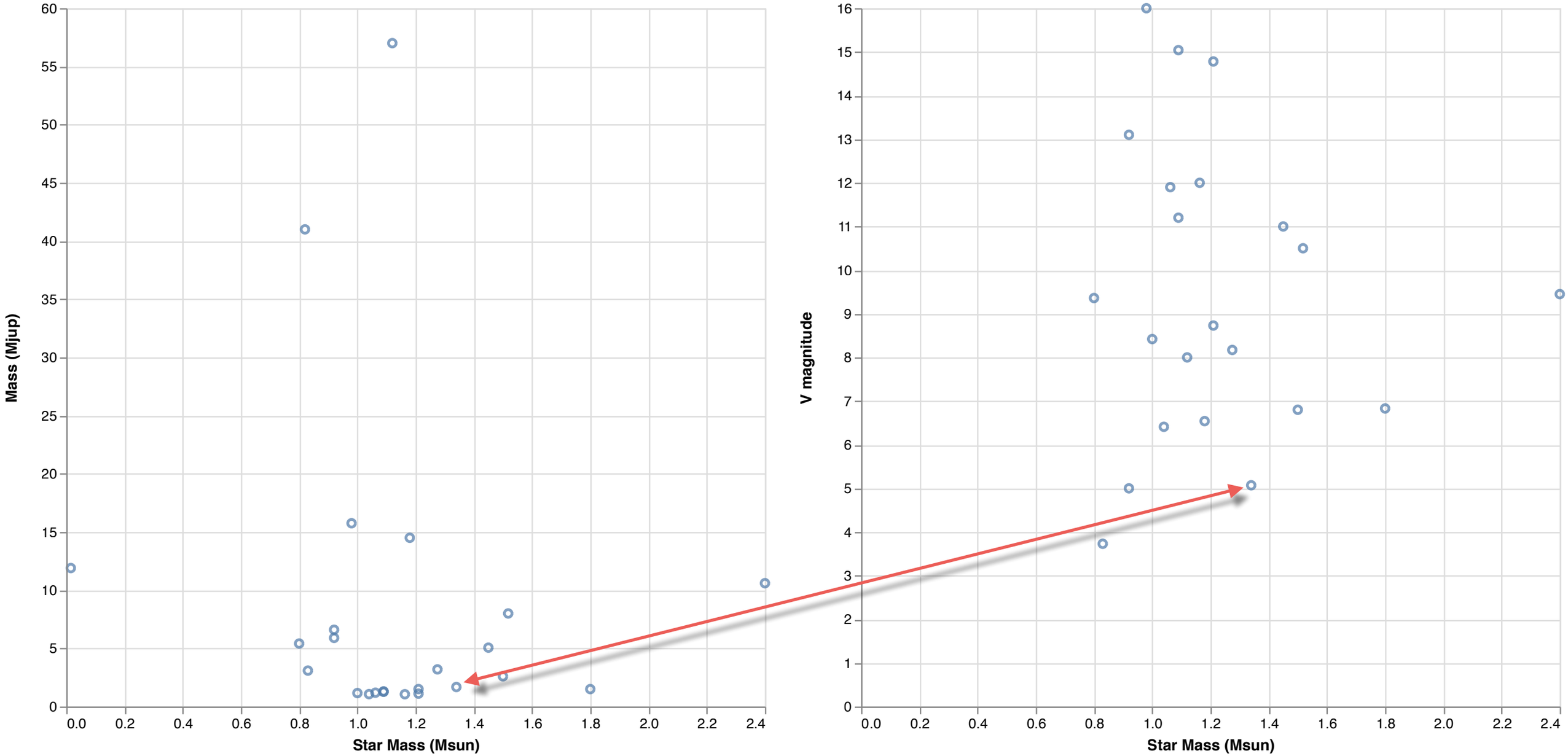
[Source: 3D Visualization for Nonspatial Data: Guidelines and Challenges, Niklas Elmquist, 2017]

3D plots can still be effective in some cases

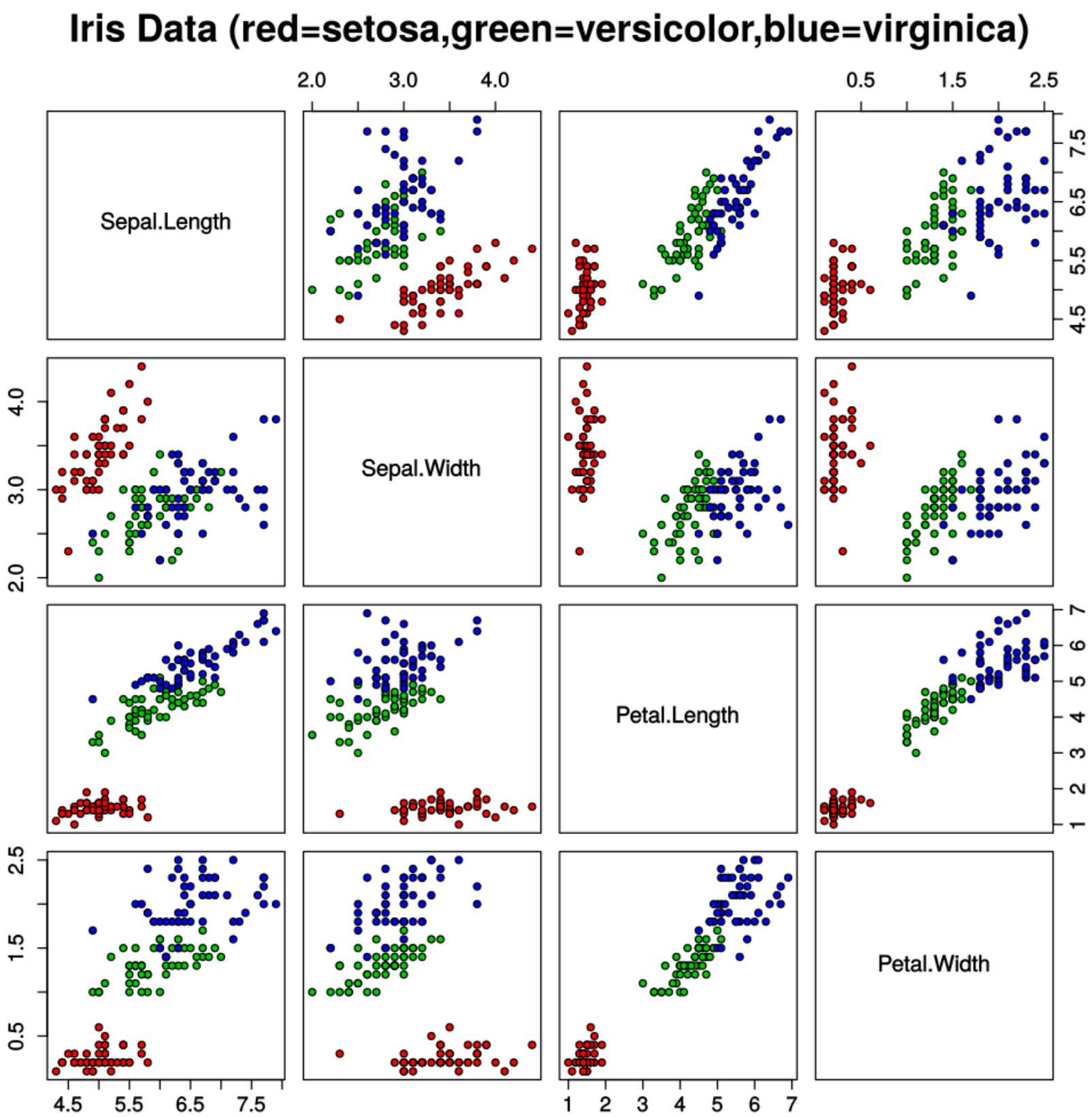


[https://www.nytimes.com/interactive/2015/03/19/upshot/3d-yield-curve-economic-growth.html]

Actually, we could add an axis…

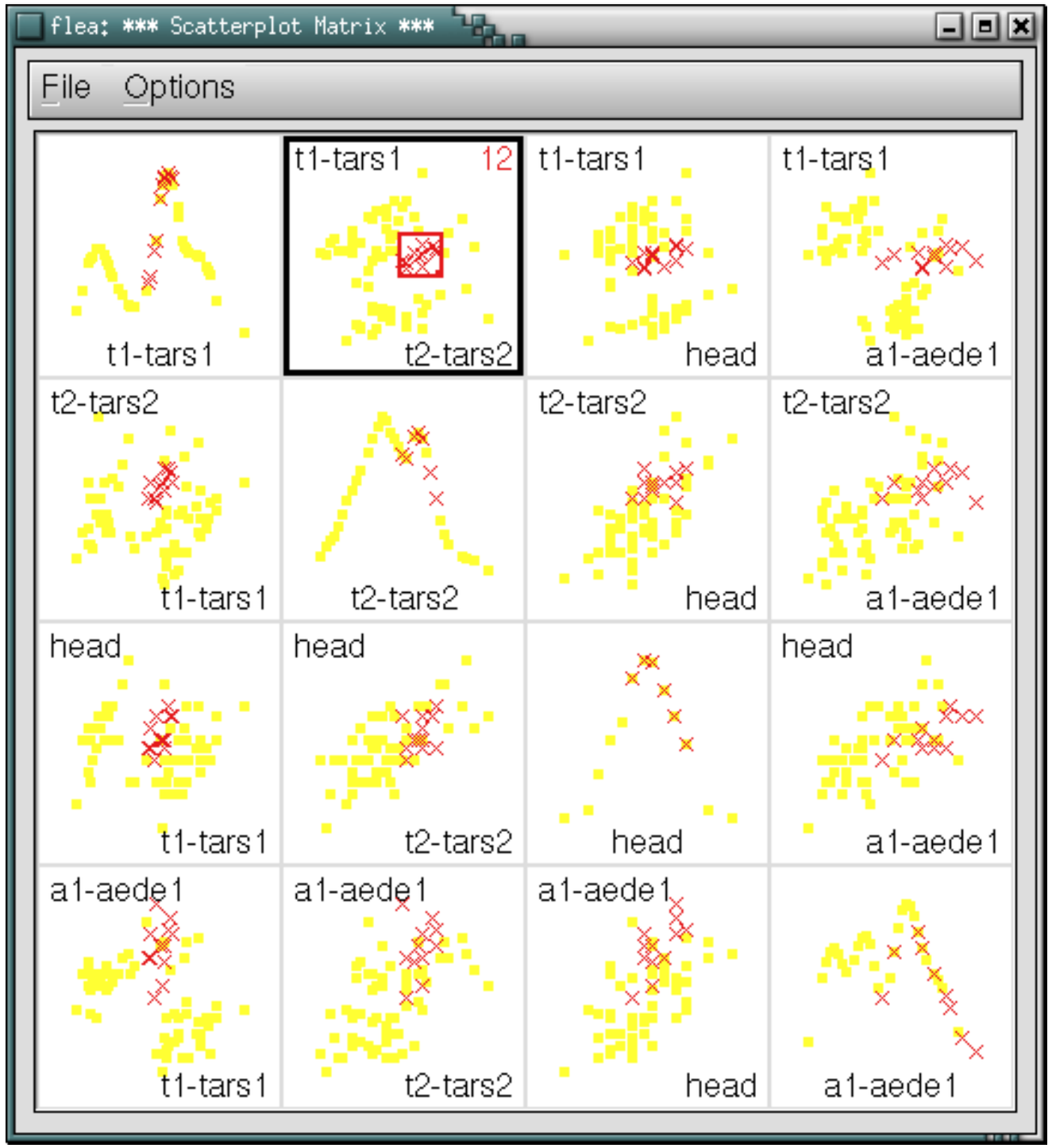


or even more than one, using a scatterplot matrix:



[Source: https://commons.wikimedia.org/wiki/File:Iris_dataset_scatterplot.svg]

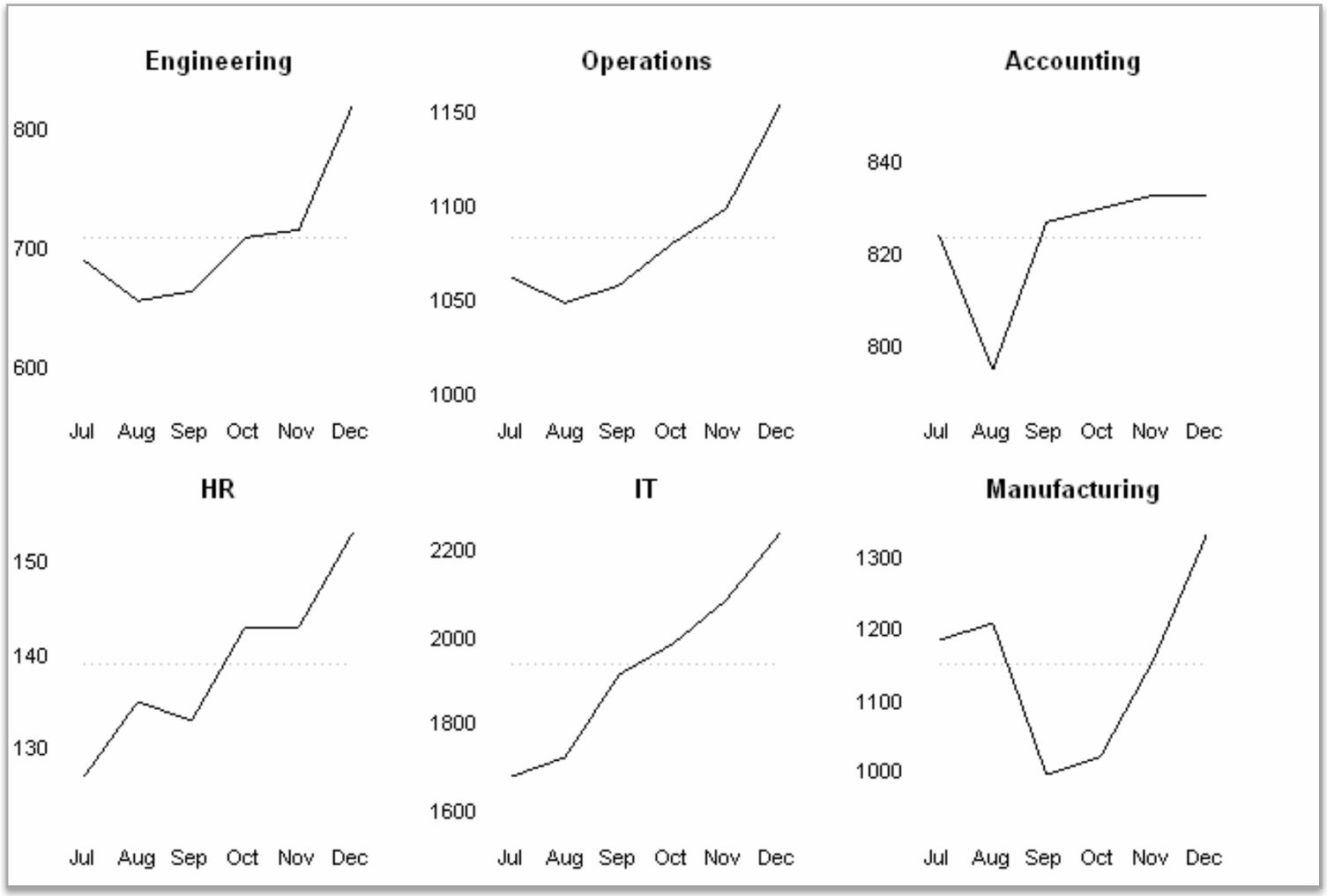
Brushing & linking



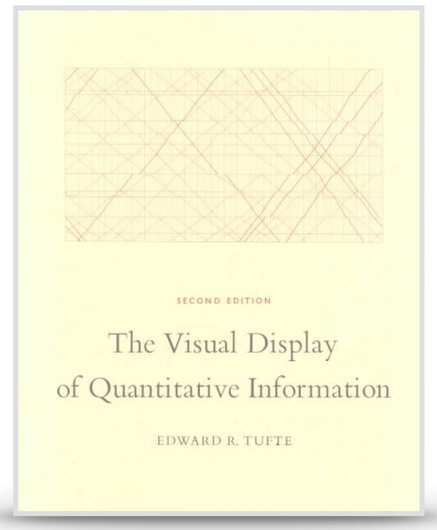
[Source: Robert Voigt, An Extended Scatterplot Matrix and Case Studies in Information Visualization, 2002]

which takes us to the higher-level concept of *Small Multiples*:

"a series of similar graphs or charts using the same scale and axes, allowing them to be easily compared"



example with lines charts instead of scatterplots



Term popularized by E. Tufte.

[Source: https://en.wikipedia.org/wiki/Small_multiple]

Another example, with maps:

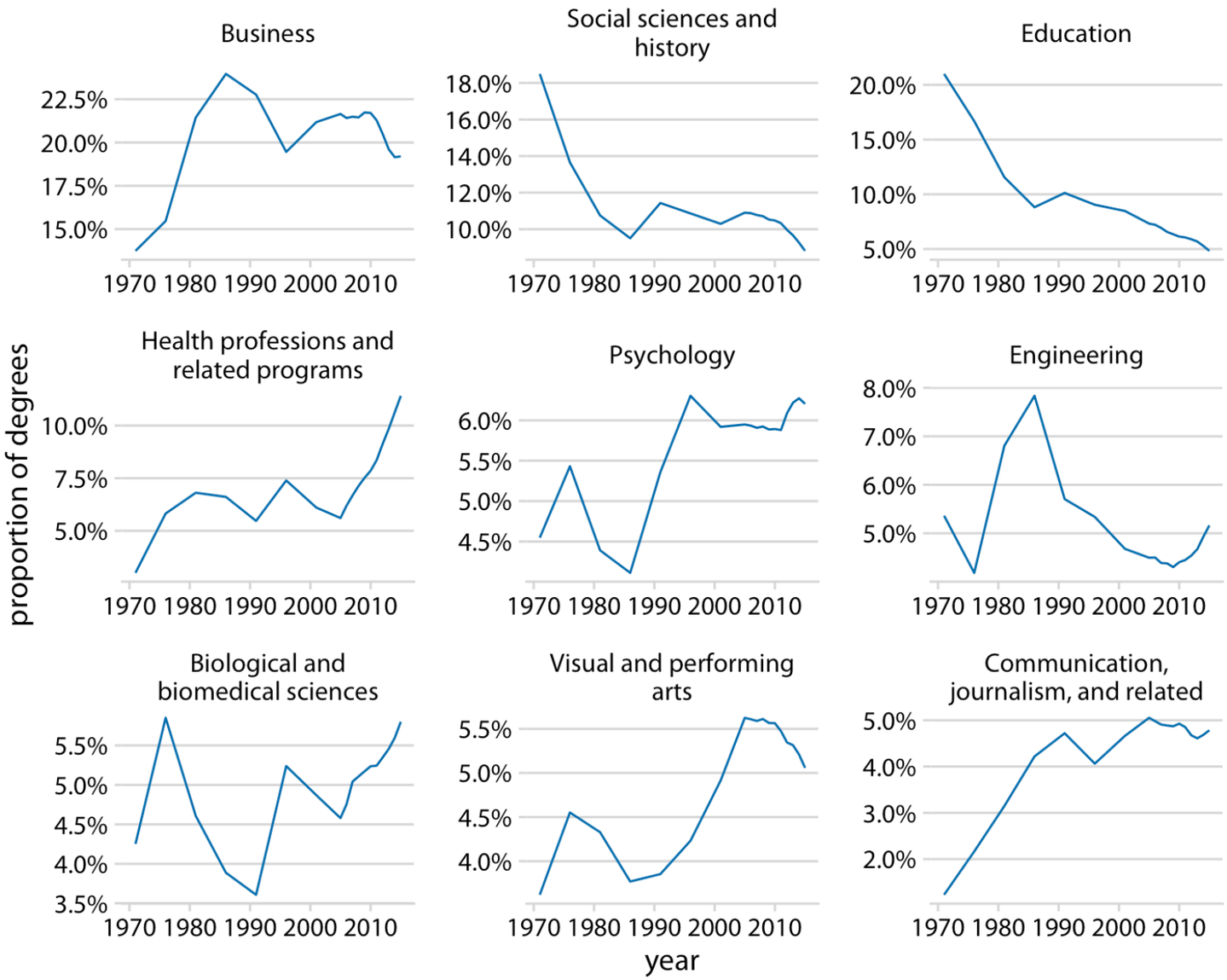
2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support

Orange and green colors correspond to states where support for vouchers was greater or less than the national average.
The seven ethnic/religious cagetories are mutually exclusive. "Evangelicals" includes Mormons as well as born-again Protestants.
Where a category represents less than 1% of the voters of a state, the state is left blank.

[Source: https://andrewgelman.com/2009/07/15/hard_sell_for_b/]

Small multiples - design considerations

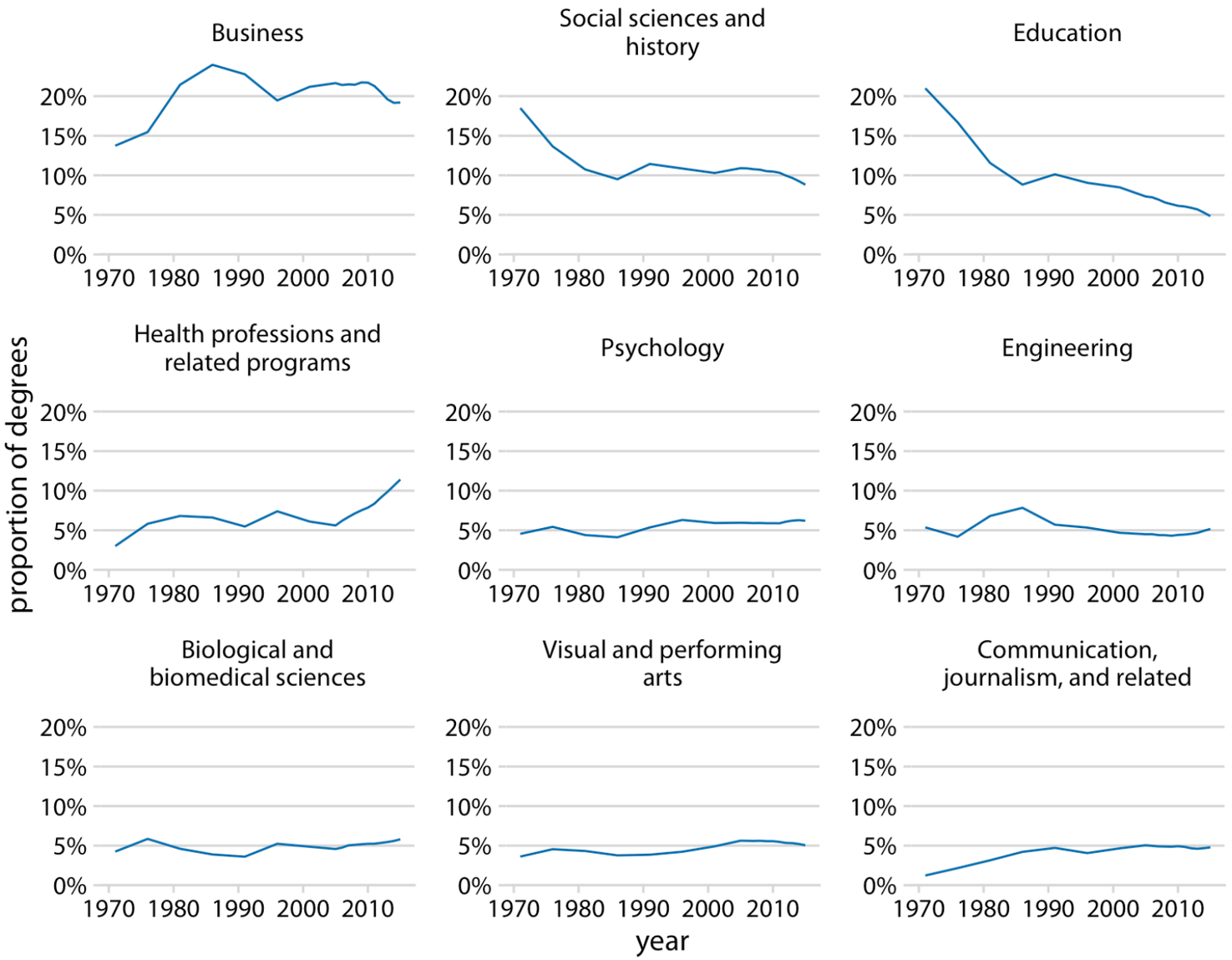
- logical ordering
- comparable encodings & scales



[Example source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

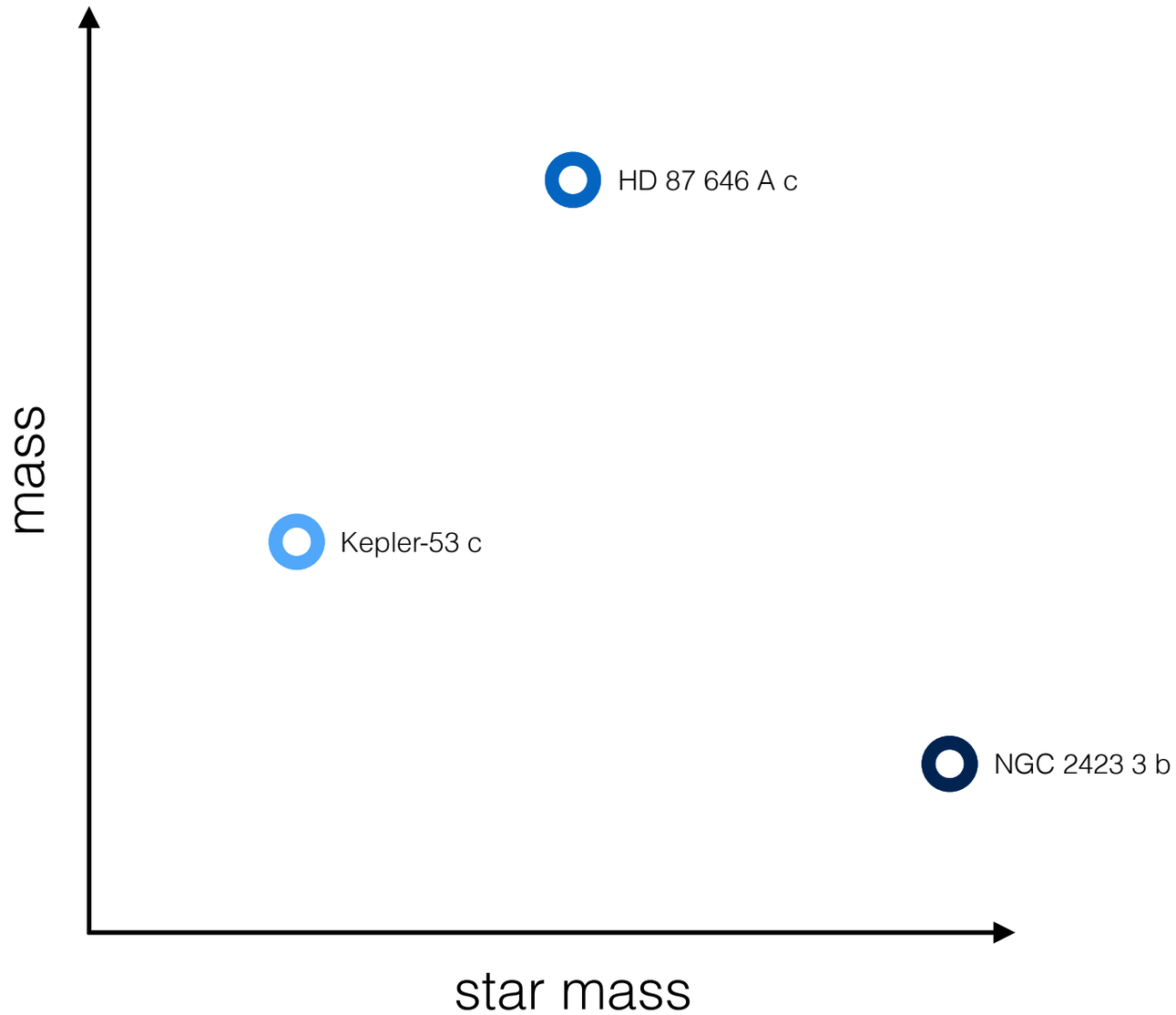
Small multiples - design considerations

- logical ordering
- comparable encodings & scales

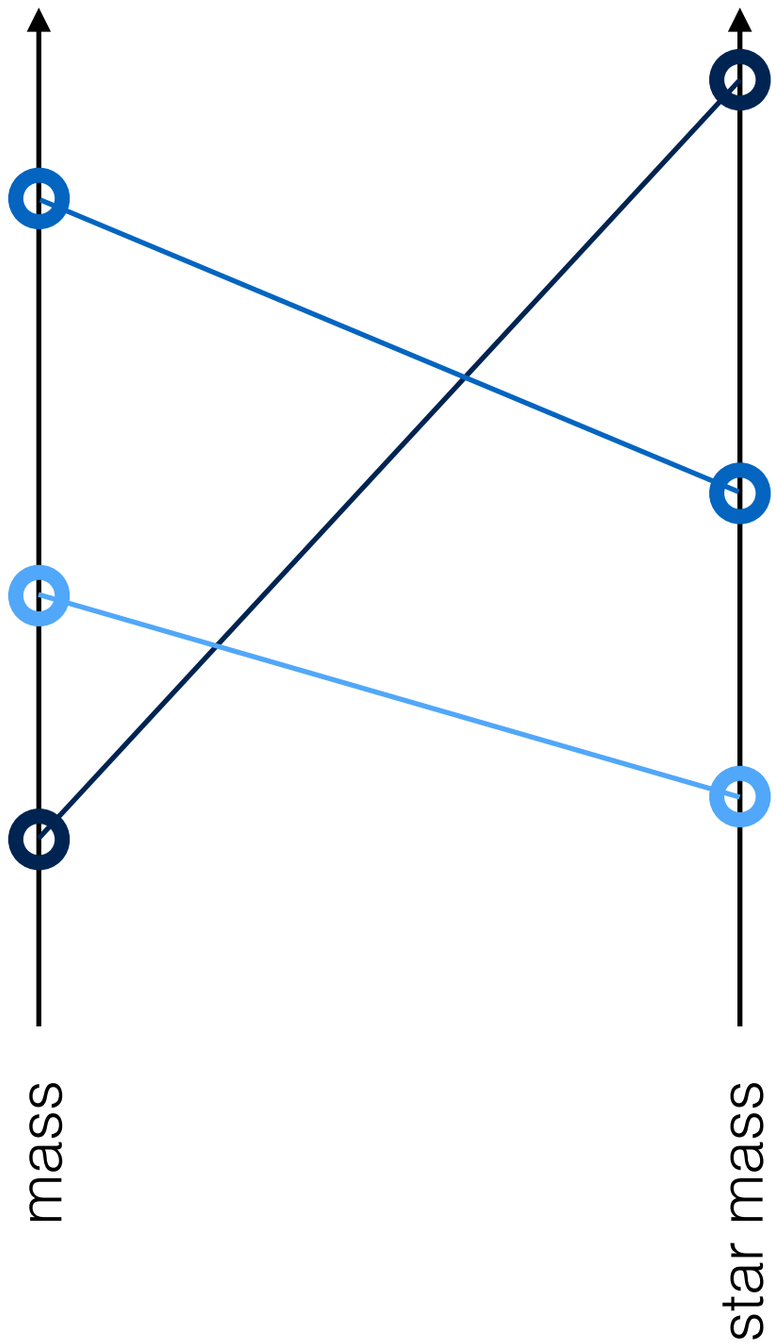


[Example source: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly, 2019]

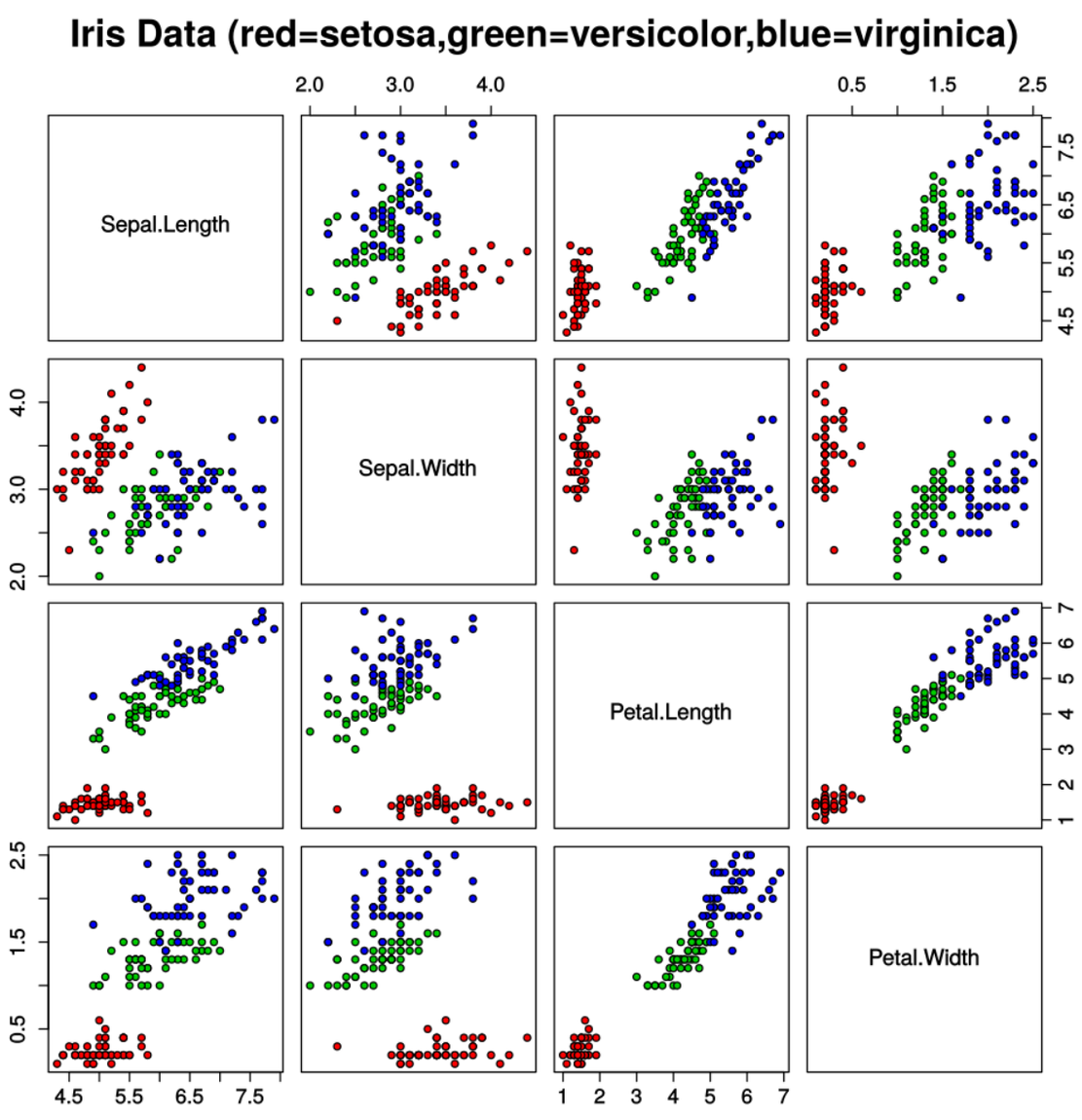
Parallel Coordinates

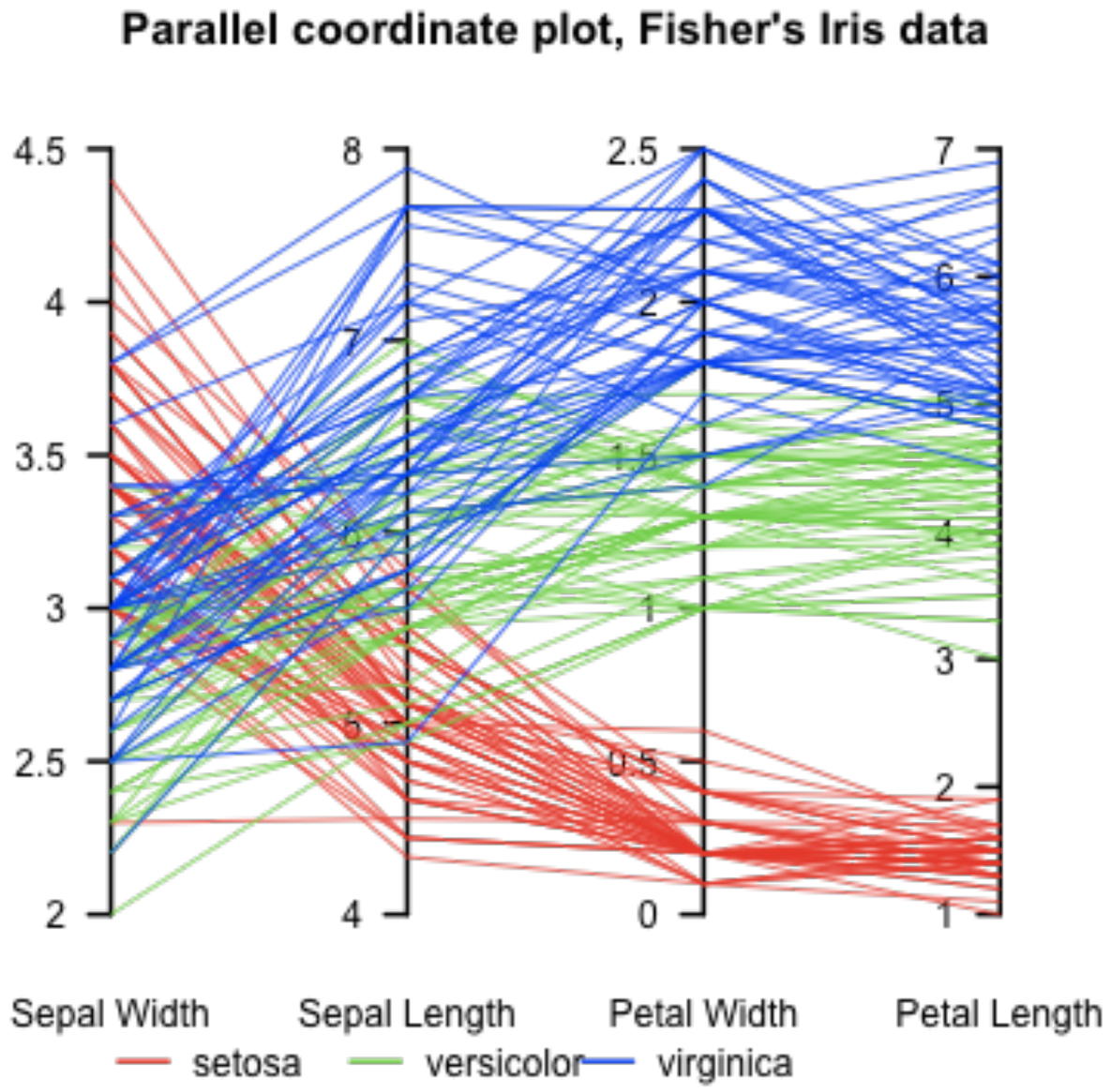


Parallel Coordinates



Parallel Coordinates





UPDATED February 12, 2012

LINKEDIN SHARE

Four Ways to Slice Obama's 2013 Budget Proposal

Explore every nook and cranny of President Obama's federal budget proposal.

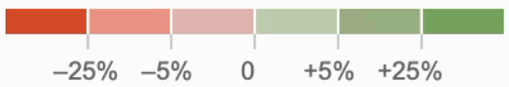
All Spending | Types of Spending | Changes | Department Totals

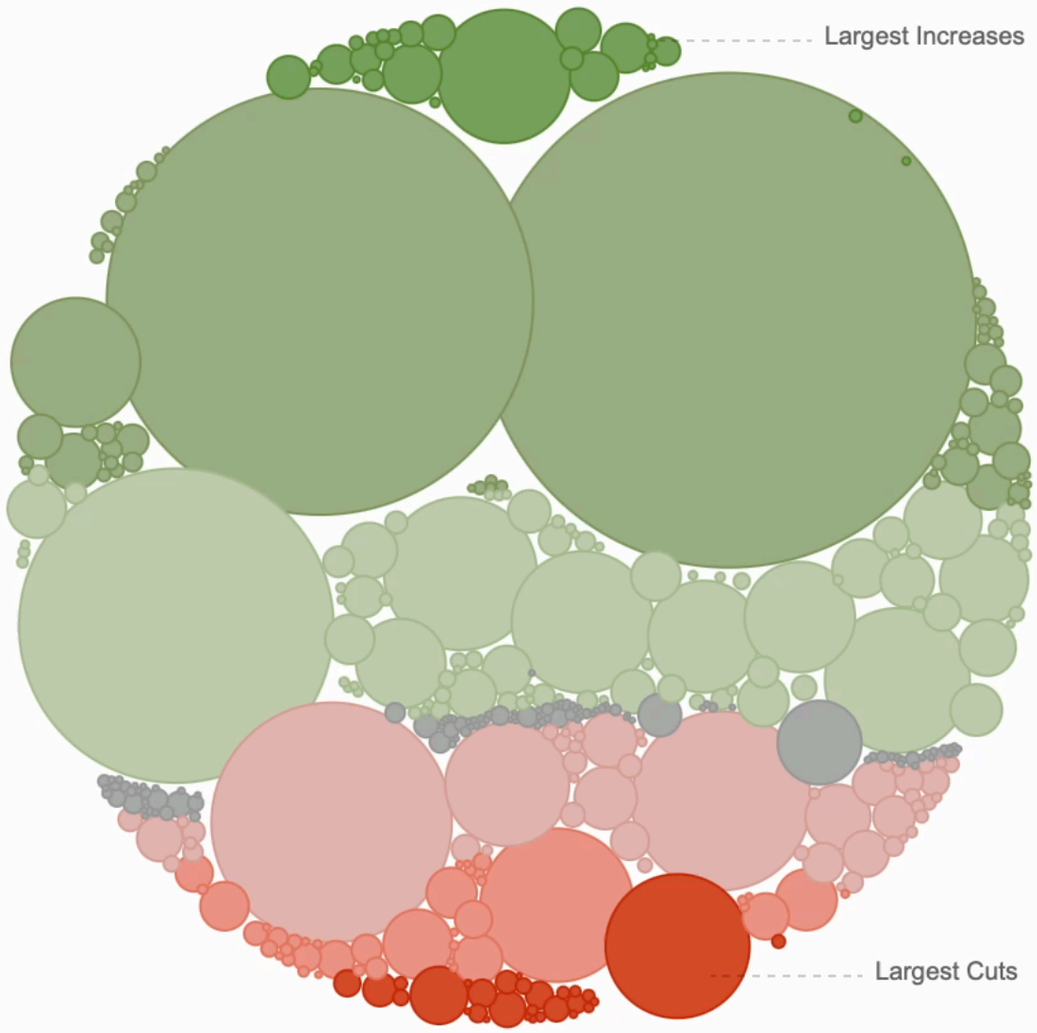
How $3.7 Trillion Is Spent

Mr. Obama's budget proposal includes $3.7 trillion in spending in 2013, and forecasts a $901 billion deficit.

Circles are sized according to the proposed spending.

Color shows amount of cut or increase from 2012.





The proposal forecasts a $901 billion deficit.

Chart shows $3.7 trillion authorized to be spent in 2013. (Total spending is estimated to be $3.8 trillion, including funds authorized in other years). Negative budget authority, which results from fees or other collections, is shown only on the department totals tab, but is included in other totals.

By SHAN CARTER | Send Feedback