THE UNIVERSITY OF
SYDNEY

CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE

ESSENTIALS OF DATA VISUALIZATION

THINKING ABOUT DRAWING DATA + COMMUNICATING SCIENCE

ENCODING

choosing a data representation

When you think of data visualization, the first ideas that come to mind are a scatter plot, or a bar chart, a box plot or a network diagram.

These are all data encodings—methods that relate data values to the positions, sizes and shapes of the lines or symbols that appear on the screen or in a figure.

There are many data encodings—which do you choose?

First, it must help answer relevant questions about the data that are difficult, or impossible, to answer by staring at the raw data itself. As such, the encoding may be of the data or some transformation of the data that addresses your questions. Just because you have a network doesn't mean you should automatically draw a force-directed hairball.

Second, it should accommodate the uncertainty of the data. How do you incorporate uncertainty in a scatter plot? Easy—error bars. How about a pie chart. Ah, well, umm. We'll come back to this later.

Third, it should be flexible enough to address questions that you haven't thought of yet. This sounds vague, I know. What I mean is that if the encoding warps the data or doesn't to at least try to limit occlusion (the phenomena where points overlap and hide behind each other), it's likely to be less useful.

DEPART | DEPARTURE | ABFAHRT

Trains au départ Departures trains Abfahrt der züge

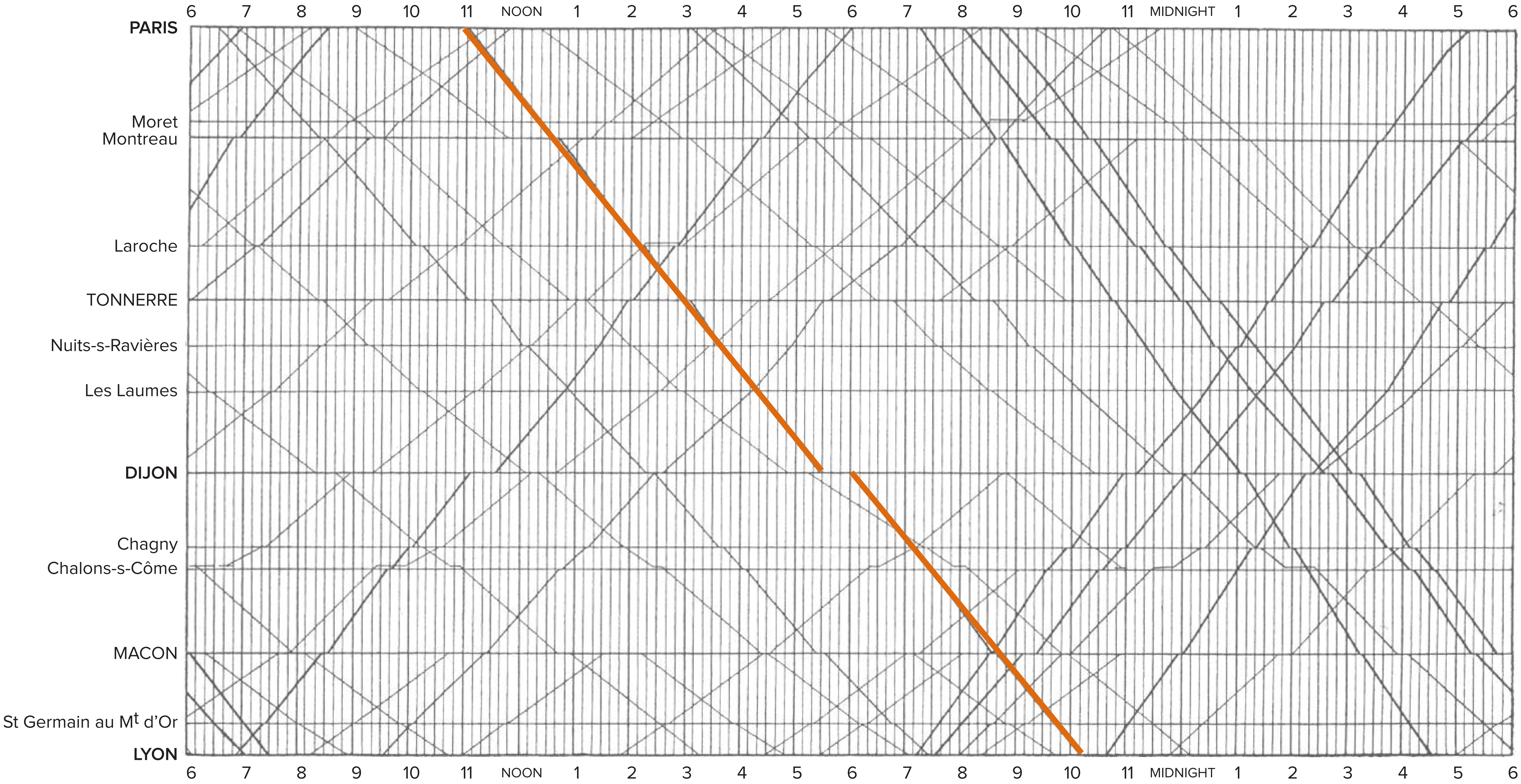
Zeit Time Heure	Nach Destination Destination	Hinweise Remarks Particularités		Zug nr Train nr Train nr	Gleis Platform Voie
11h58	DAMMARTIN CREPY SOISSONS ANIZY-PINON LAON		ter	849917	18
12h01	PERSAN CHAMBLY MERU ST-SULPICE BEAUVAIS		ter	847419	21
12h01	BRUXELLES-MIDI LIEGE AACHEN KOLN ESSEN	THALYS CONFORT 1 ET CONFORT 2		9437	9
12h07	CREIL COMPIEGNE CHAUNY TERGNIER ST-QUENTIN	1ère ET 2ème CLASSE		12309	11
12h10	ORRY-LA-VILLE CHANTILLY-GOUVIEUX CREIL		ter	847609	13
12h25	BRUXELLES ROTTERDAM SCHIPHOL AMSTERDAM	THALYS CONFORT 1 ET CONFORT 2		9339	
12h28	CREIL LONGUEAU AMIENS	1ère ET 2ème CLASSE		12011	
12h43	LONDON ST PANCRAS INT	HALL LONDRES	eurostar	9029	1erETAGE 1stFLOOR
12h46	LILLE FLANDRES	1ère ET 2ème CL avec RESERVATION	TGV	7043	
12h49	CREIL RIEUX PONT LONGUEIL COMPIEGNE		ter	847807	
12h52	ARRAS LENS BETHUNE HAZEBROUCK DUNKERQUE	1ère ET 2ème CL avec RESERVATION	TGV	7321	
12h52	ARRAS DOUAI VALENCIENNES	1ère ET 2ème CL avec RESERVATION	TGV	7121	
12h55	BRUXELLES-MIDI	THALYS CONFORT 1 ET CONFORT 2		9341	
13h01	PERSAN CHAMBLY MERU ST-SULPICE BEAUVAIS		ter	847421	
13h13	EBBSFLEET LONDON ST PANCRAS INT	HALL LONDRES	eurostar	9031	1erETAGE 1stFLOOR
13h16	LILLE FLANDRES	1ère ET 2ème CL avec RESERVATION	TGV	7045	
13h26	CREPY VILLERS SOISSONS ANIZY-PINON LAON		ter	849921	
13h59	LONGUEAU AMIENS ABBEVILLE	1ère ET 2ème CLASSE		2013	
14h01	PERSAN CHAMBLY MERU ST-SULPICE BEAUVAIS		ter	847423	
14h07	CREIL PONT COMPIEGNE TERGNIER ST-QUENTIN		ter	847905	

11:53

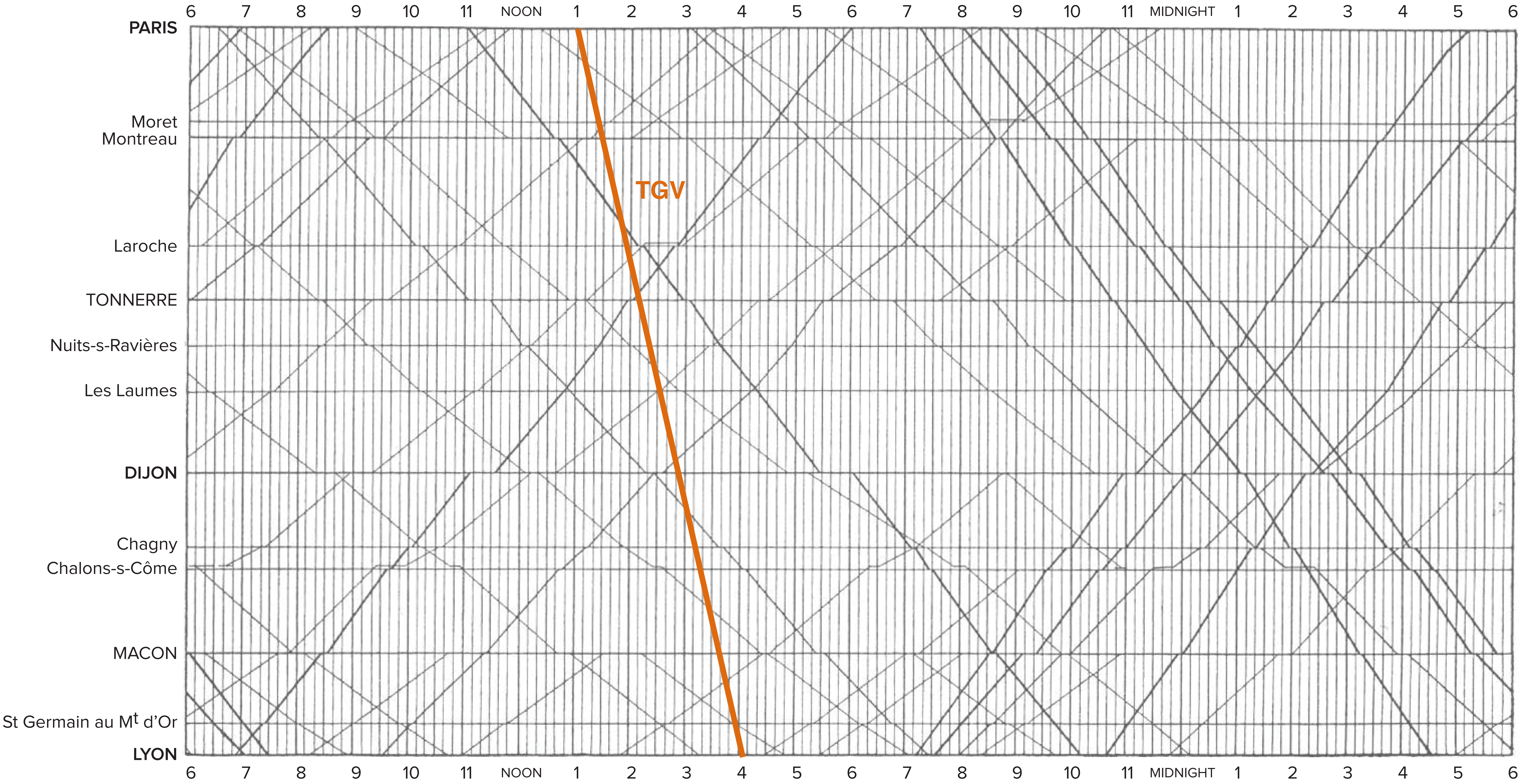
www.panoramio.com/photo/109233981

6
7
8
9
10
11
NOON
1
2
3
4
5
6
7
8
9
10
11
MIDNIGHT
1
2
3
4
5
6
PARIS
Moret
Montreau
Laroche
TONNERRE
Nuits-s-Ravières
Les Laumes
DIJON
Chagny
Chalons-s-Côme
MACON
St Germain au M^t d'Or
LYON

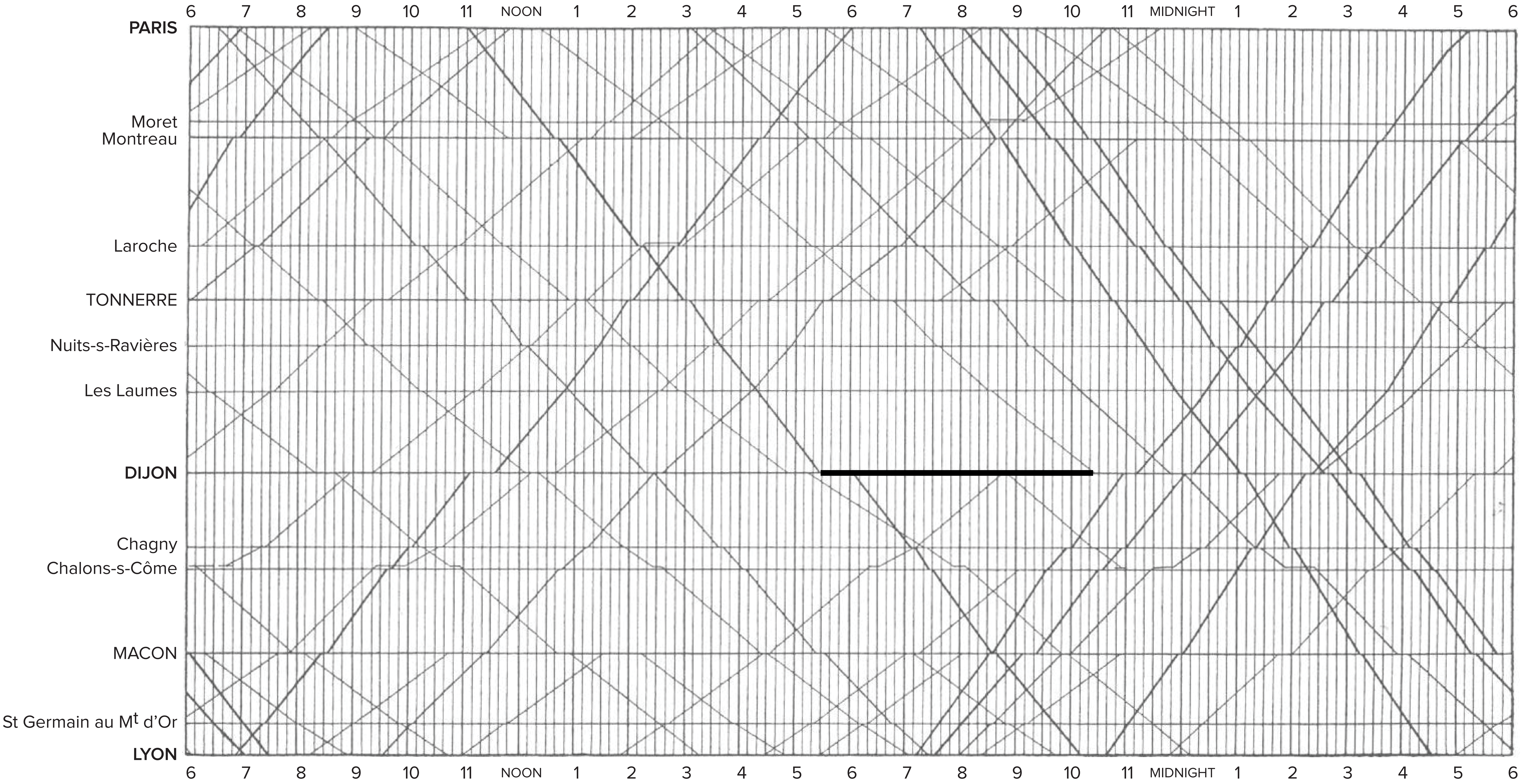
The visual display of quantitative information. R. Tufte (2001) 2nd ed. Text in original modified.



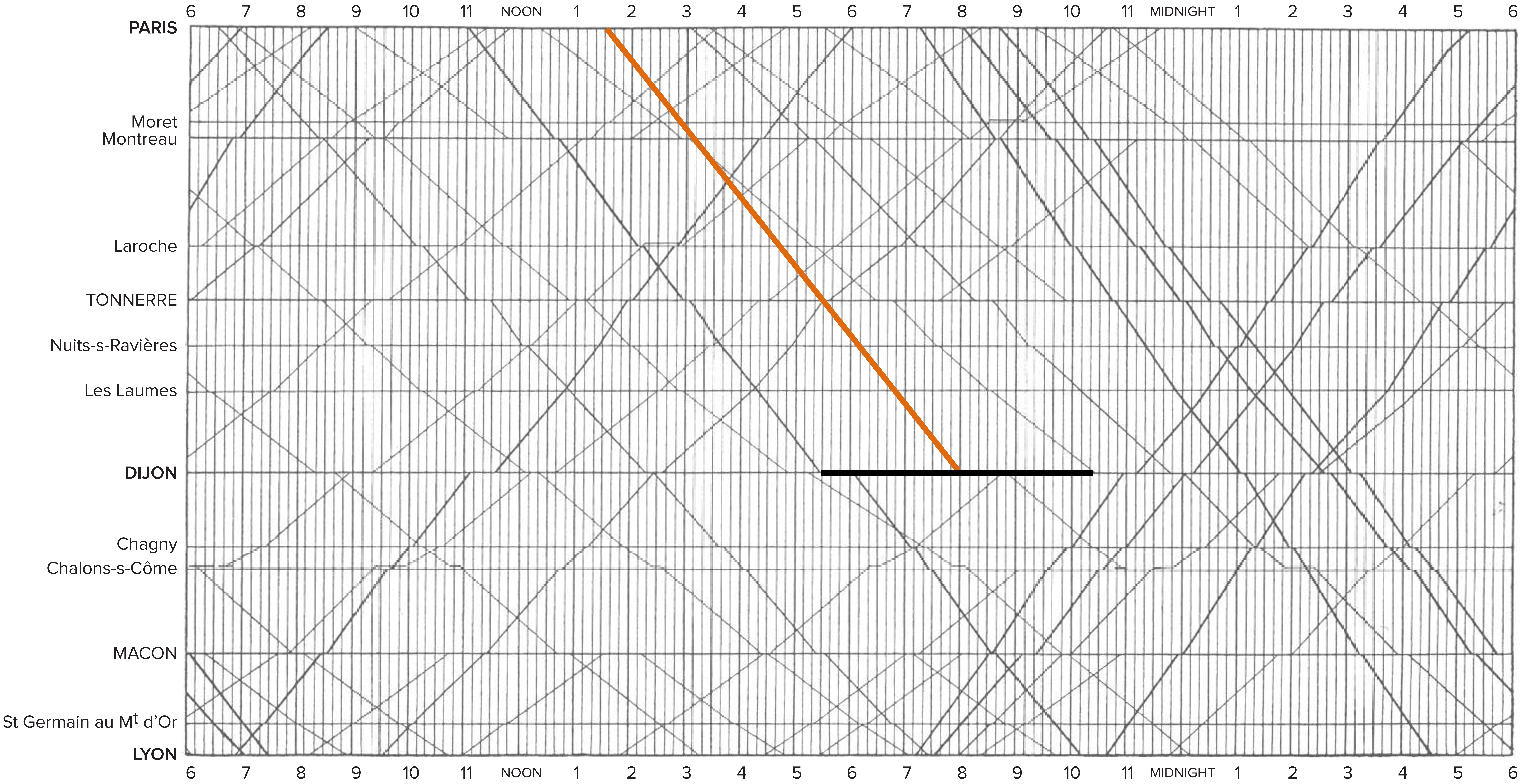
The visual display of quantitative information. R. Tufte (2001) 2nd ed. Text in original modified.



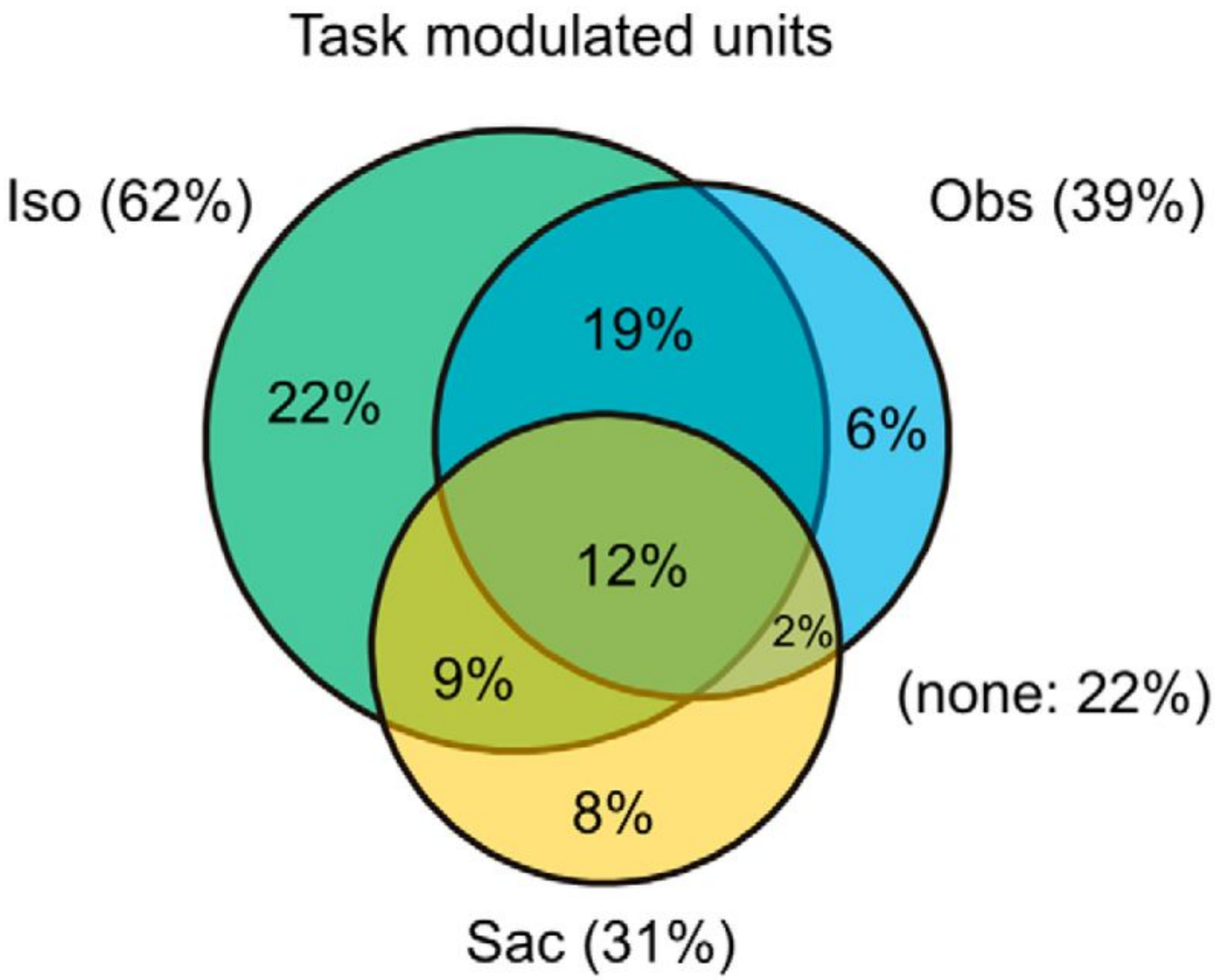
The visual display of quantitative information. R. Tufte (2001) 2nd ed. Text in original modified.

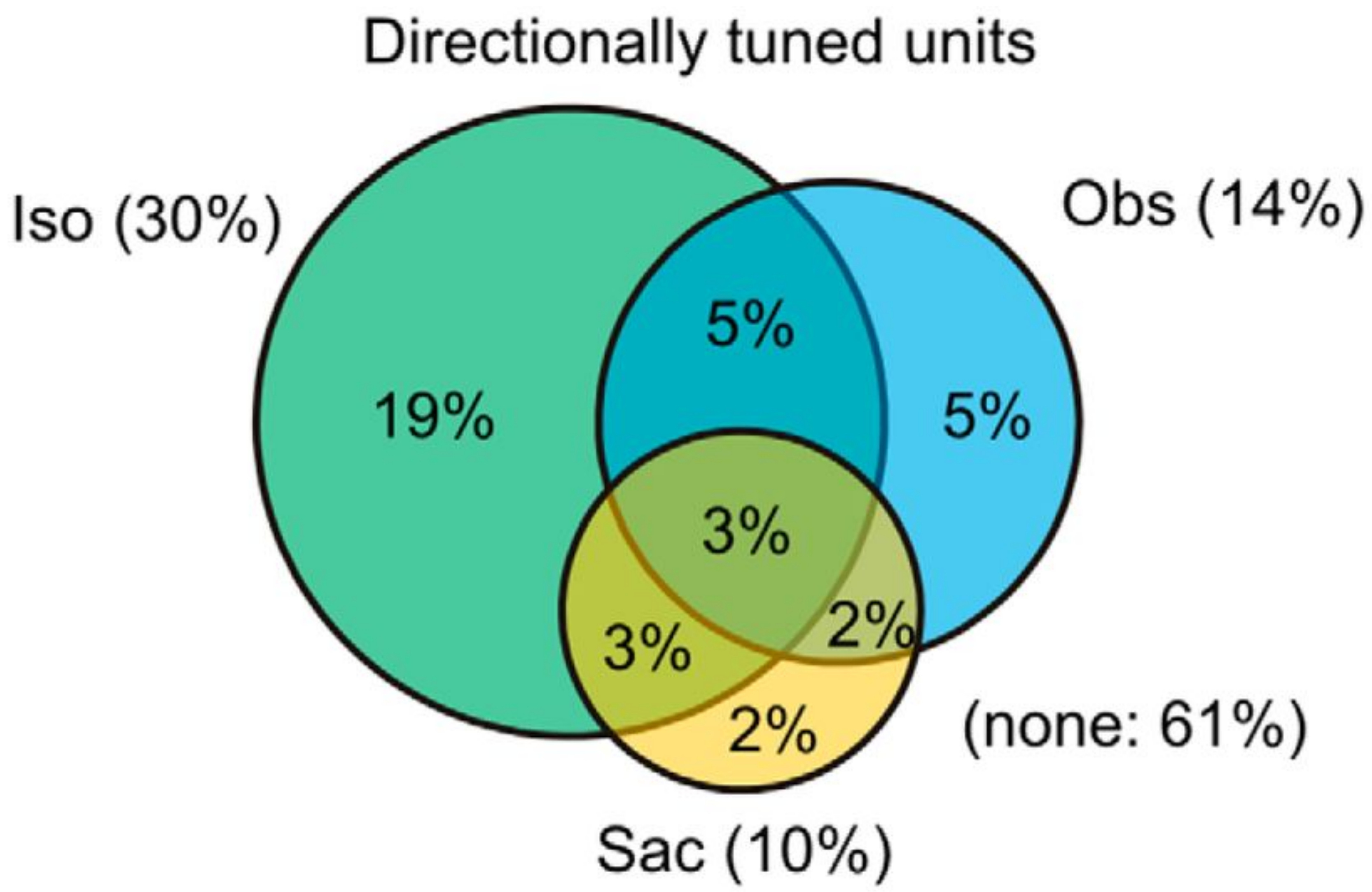


The visual display of quantitative information. R. Tufte (2001) 2nd ed. Text in original modified.



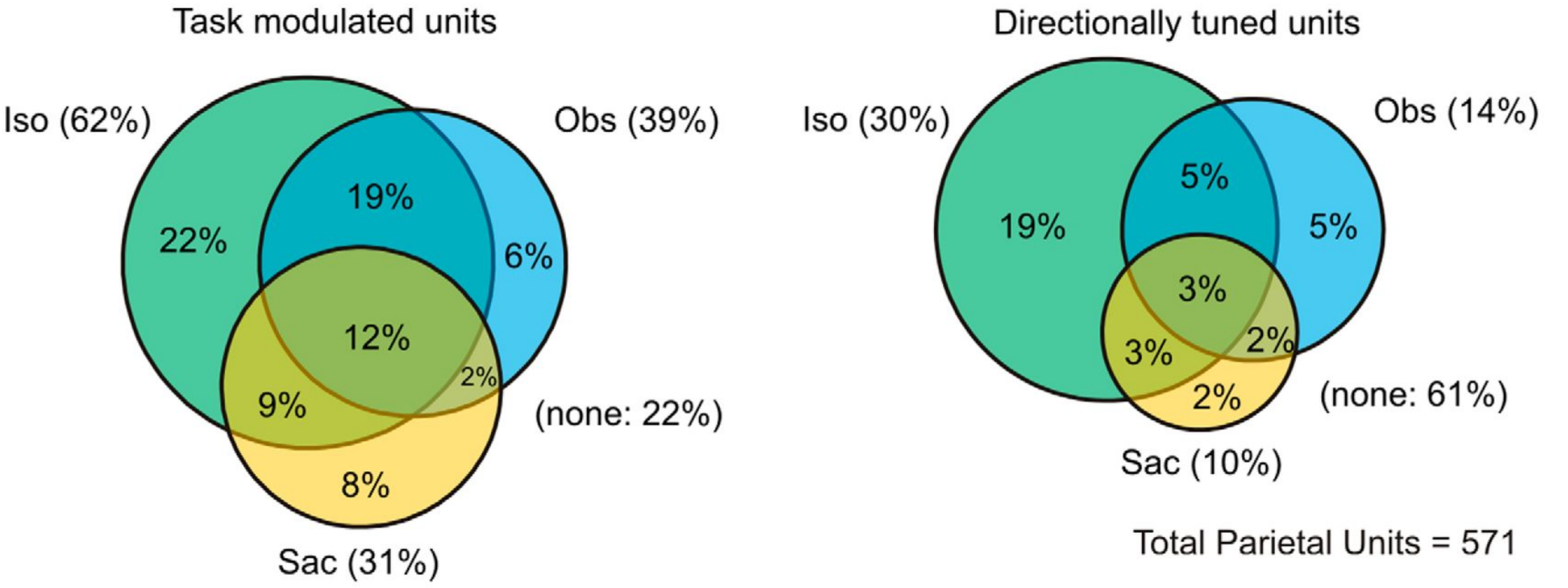
The visual display of quantitative information. R. Tufte (2001) 2nd ed. Text in original modified.

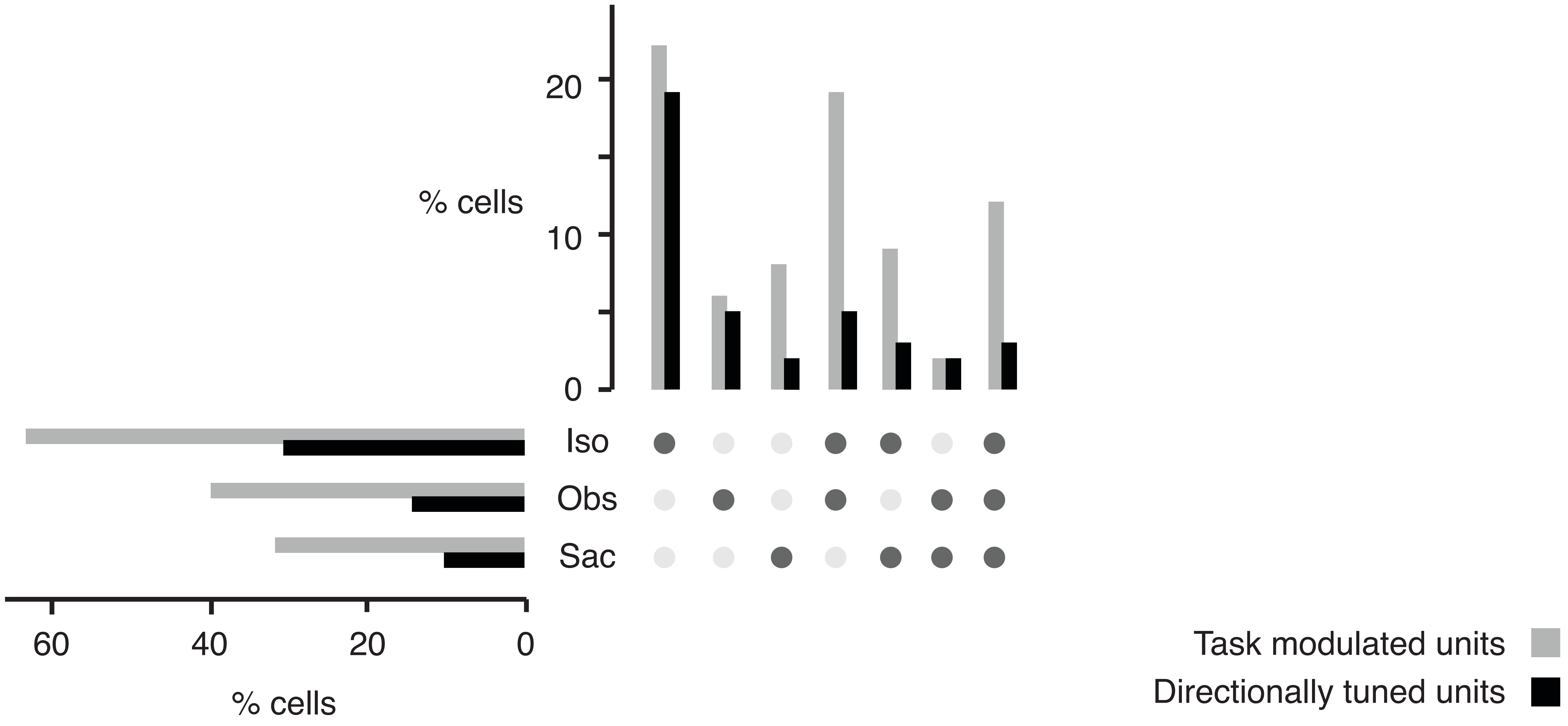




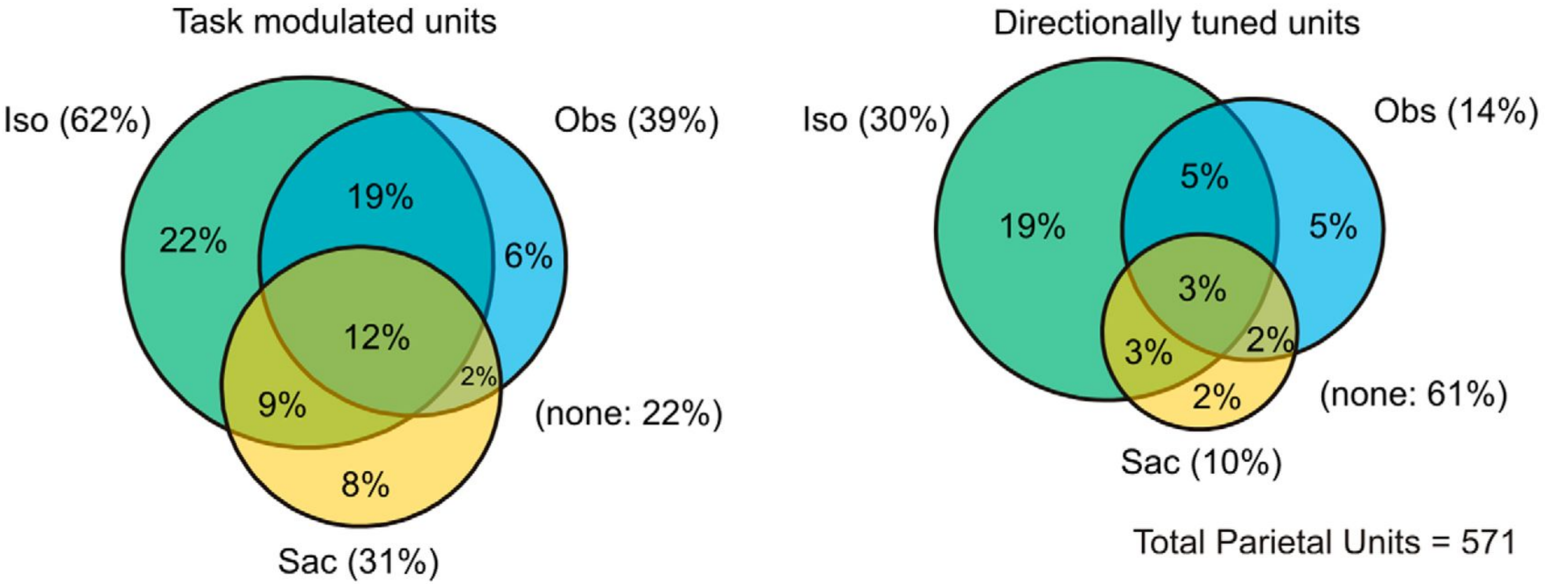
Total Parietal Units = 571

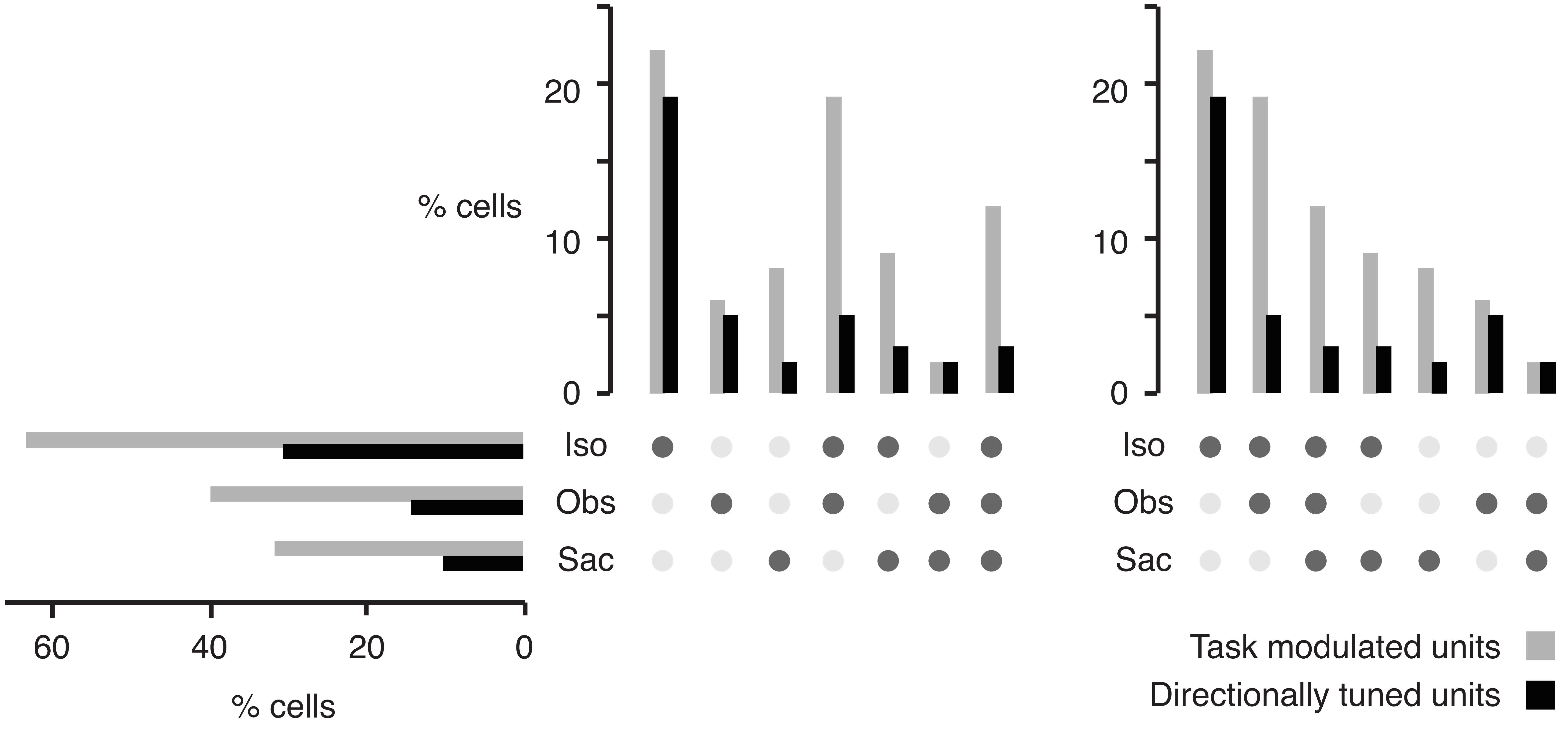
J Neurosci (2015) 35:10899-10910.



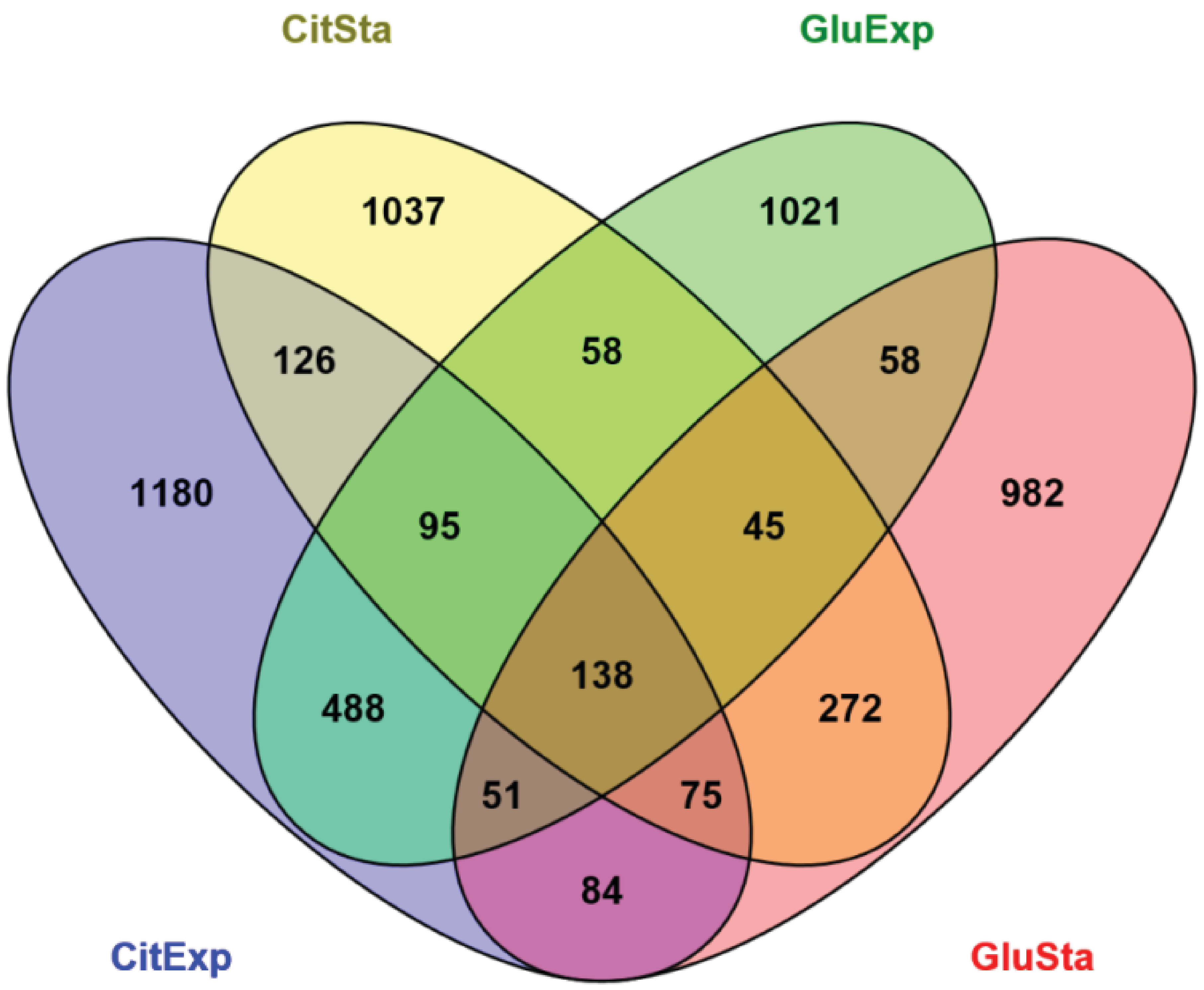


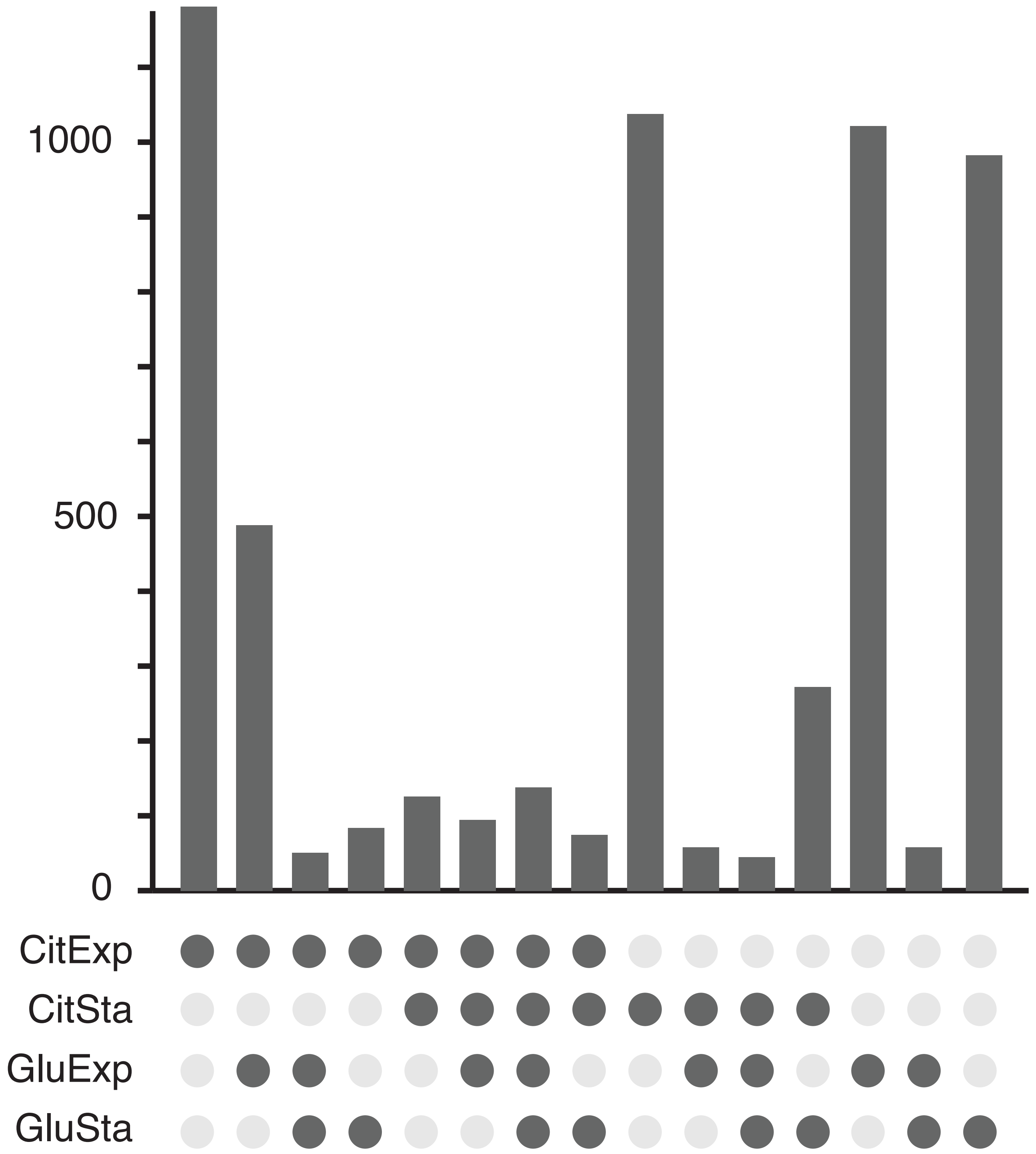
J Neurosci (2015) 35:10899-10910. Redesign using UpSet encoding (IEEE Trans Vis Comput Graph (2014) 20:1983–1992.)

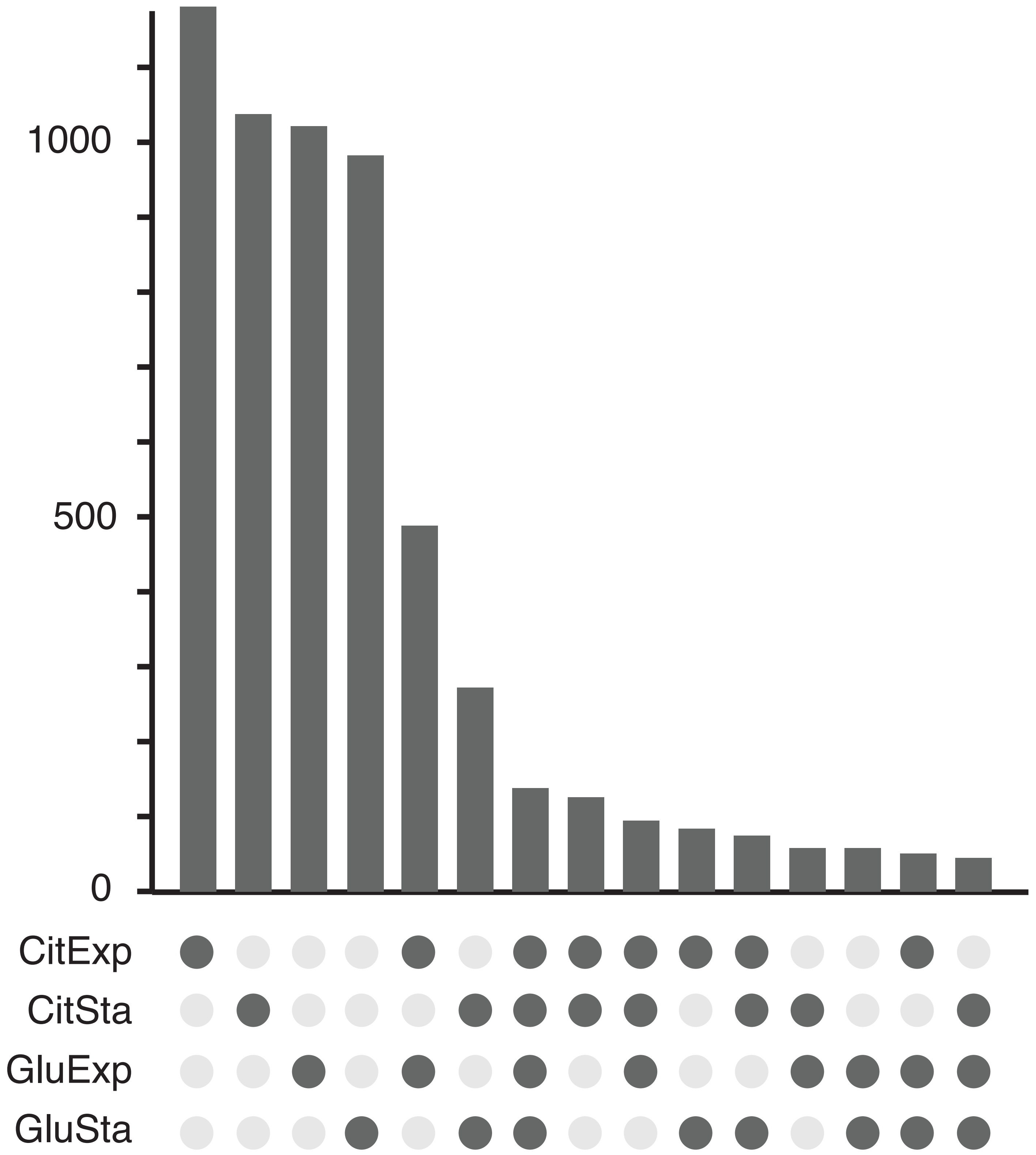




J Neurosci (2015) 35:10899-10910. Redesign using UpSet encoding (IEEE Trans Vis Comput Graph (2014) 20:1983–1992.)





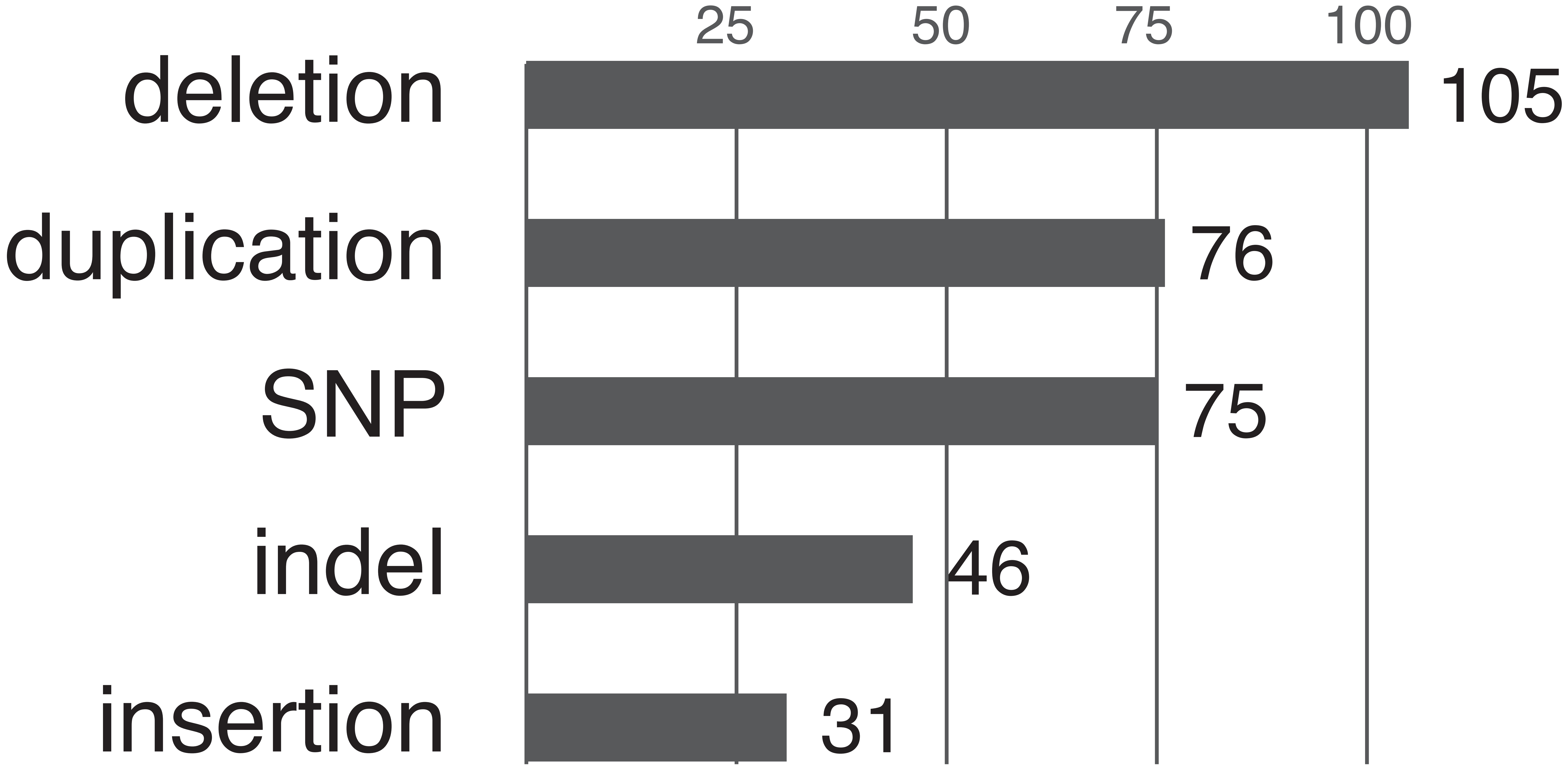


Submitted to Skills Beyond Science, 2015 Bactory Summer School, Helsingor, Denmark.
Redesign published in Environmental Microbiology, dx.doi.org/10.1111/1462-2920.13326

FREQUENCY OF VARIATION BY TYPE

deletion	105
duplication	76
SNP	75
indel	46
insertion	31

FREQUENCY OF VARIATION BY TYPE



One concept to always keep in mind is the so-called data-to-ink ratio.

Ask yourself: what ink on the page is directly related to data values and what ink is used for labels, grids, navigational components and design elements.

Then consider how to maximize ink used for data and minimize ink for everything else.

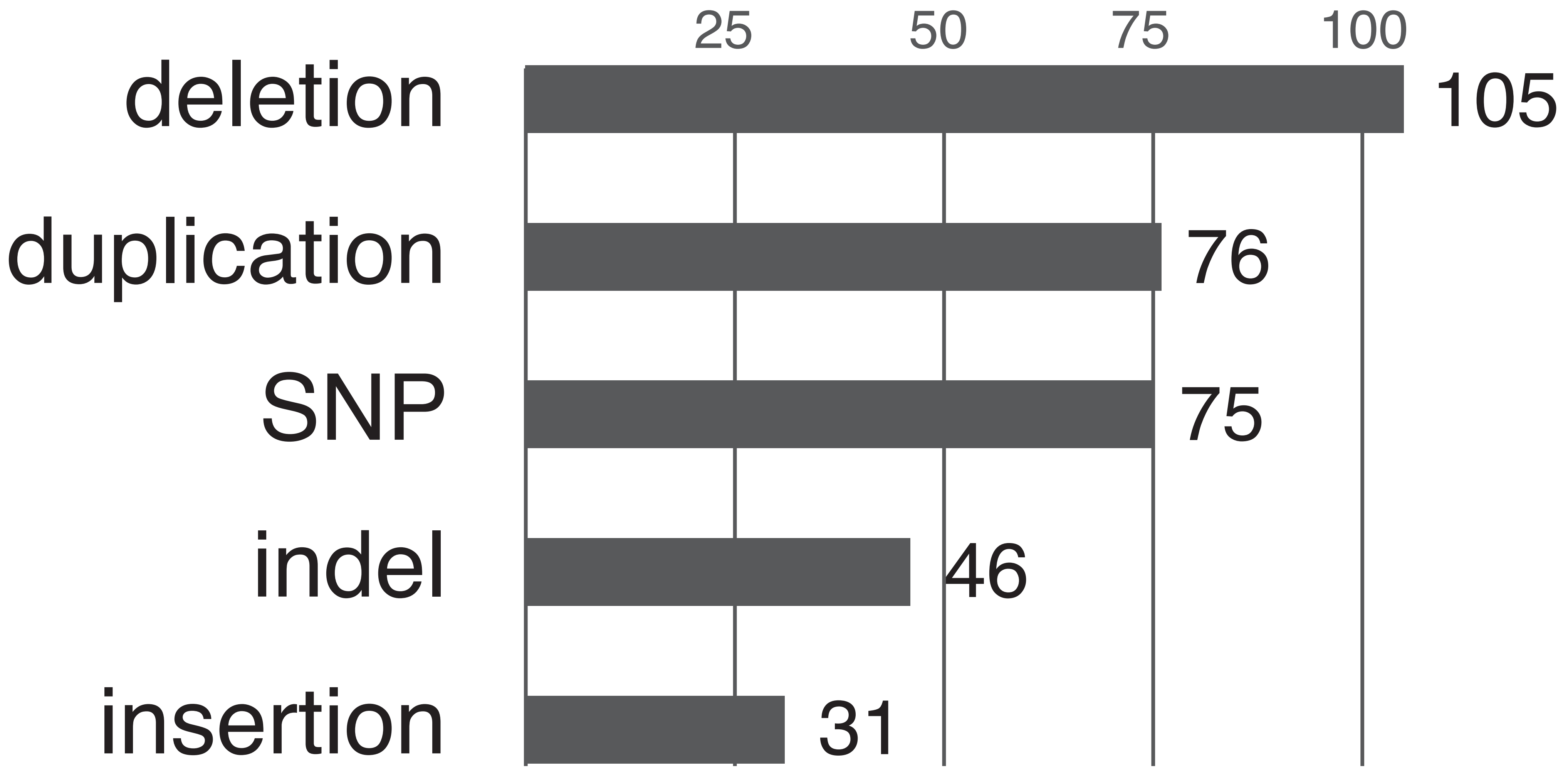
At the same time, always keep in mind ways to use less ink to tell the same story and show the same relationships without loss of accuracy and precision—and parsability!

Sometimes you need to use ink to clarify or avoid confusion. That's fine—do this. The speed at which your readers understand your message and the depth of this understanding is part of the model.

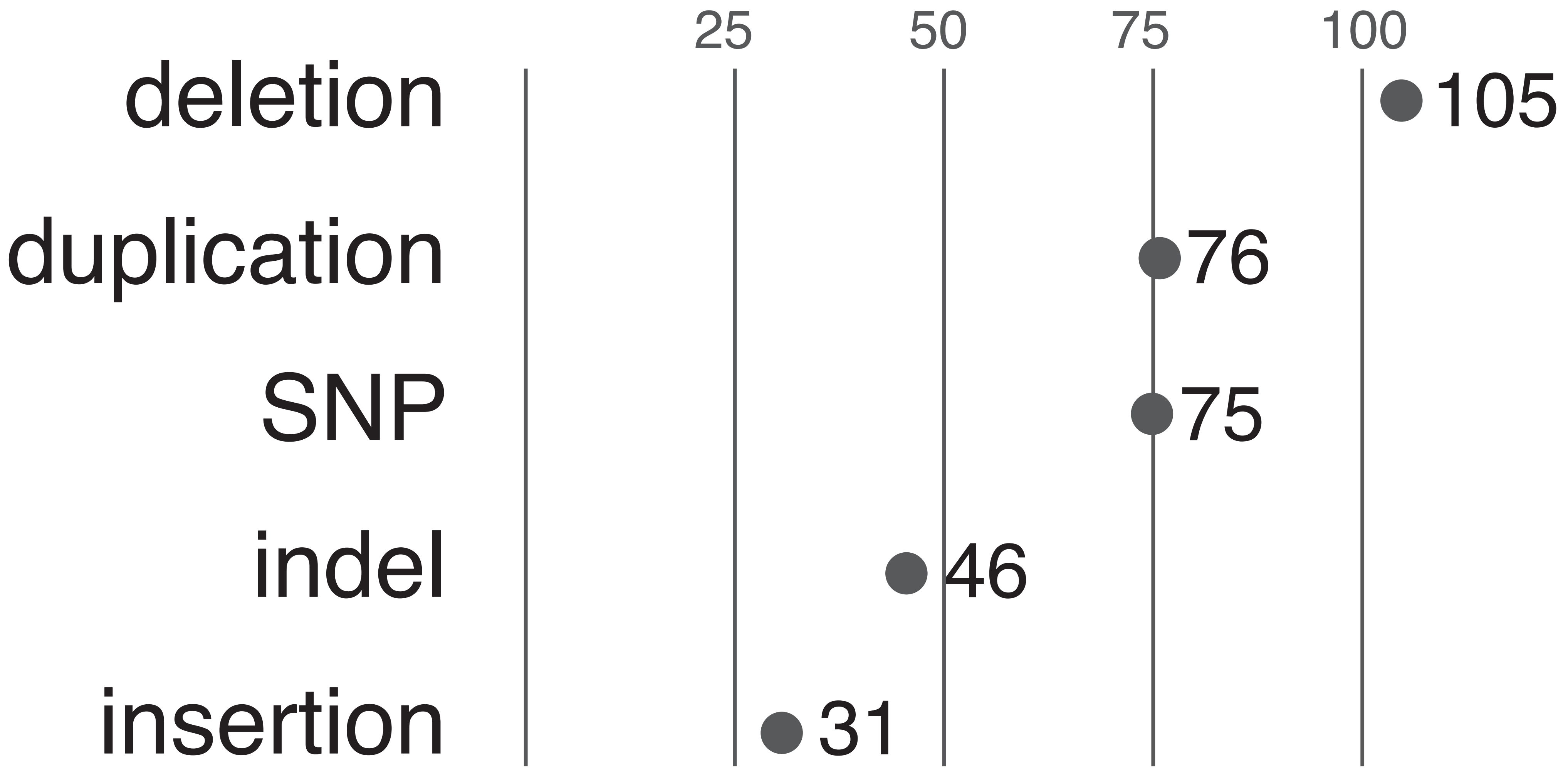
So, think "data-and-its-understanding-to-ink" ratio.

But not all the data is relevant. If you can figure out which is—that's the holy grail. In fact, if you knew this you might just compute on the data and bypass the visualization. But, I encourage you to think about "actionable-data-to-ink" ratio, too.

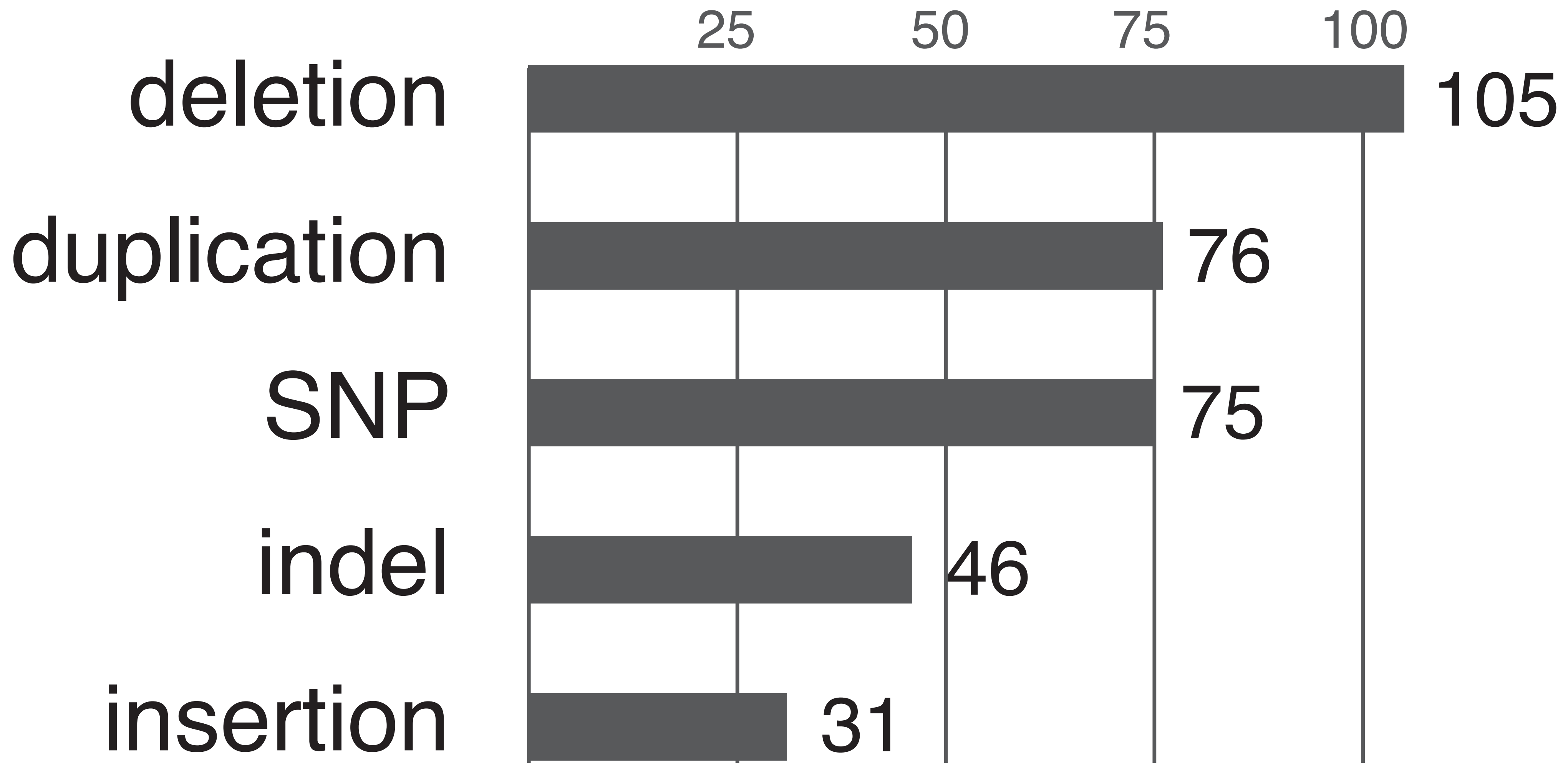
FREQUENCY OF VARIATION BY TYPE



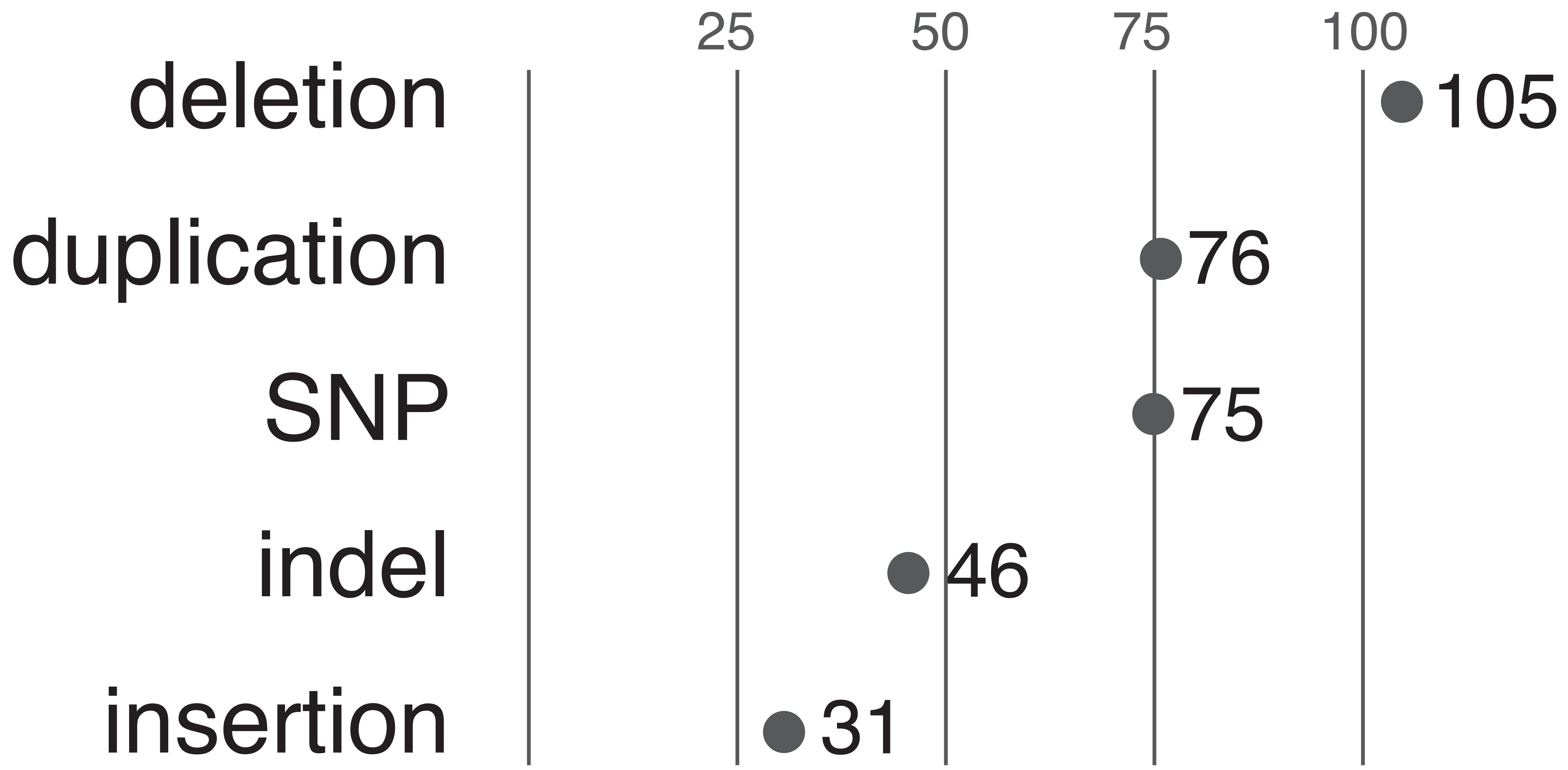
FREQUENCY OF VARIATION BY TYPE



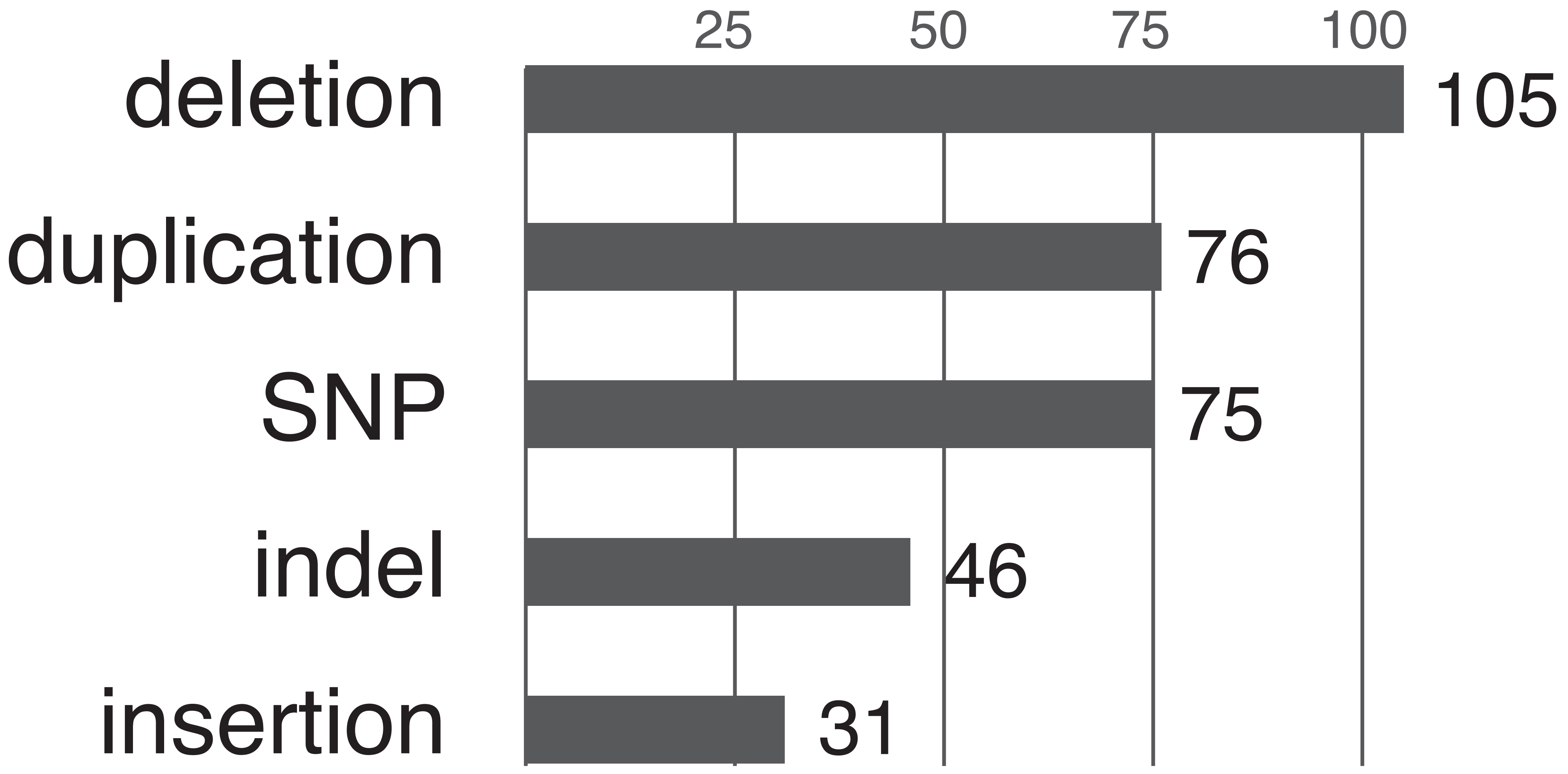
FREQUENCY OF VARIATION BY TYPE



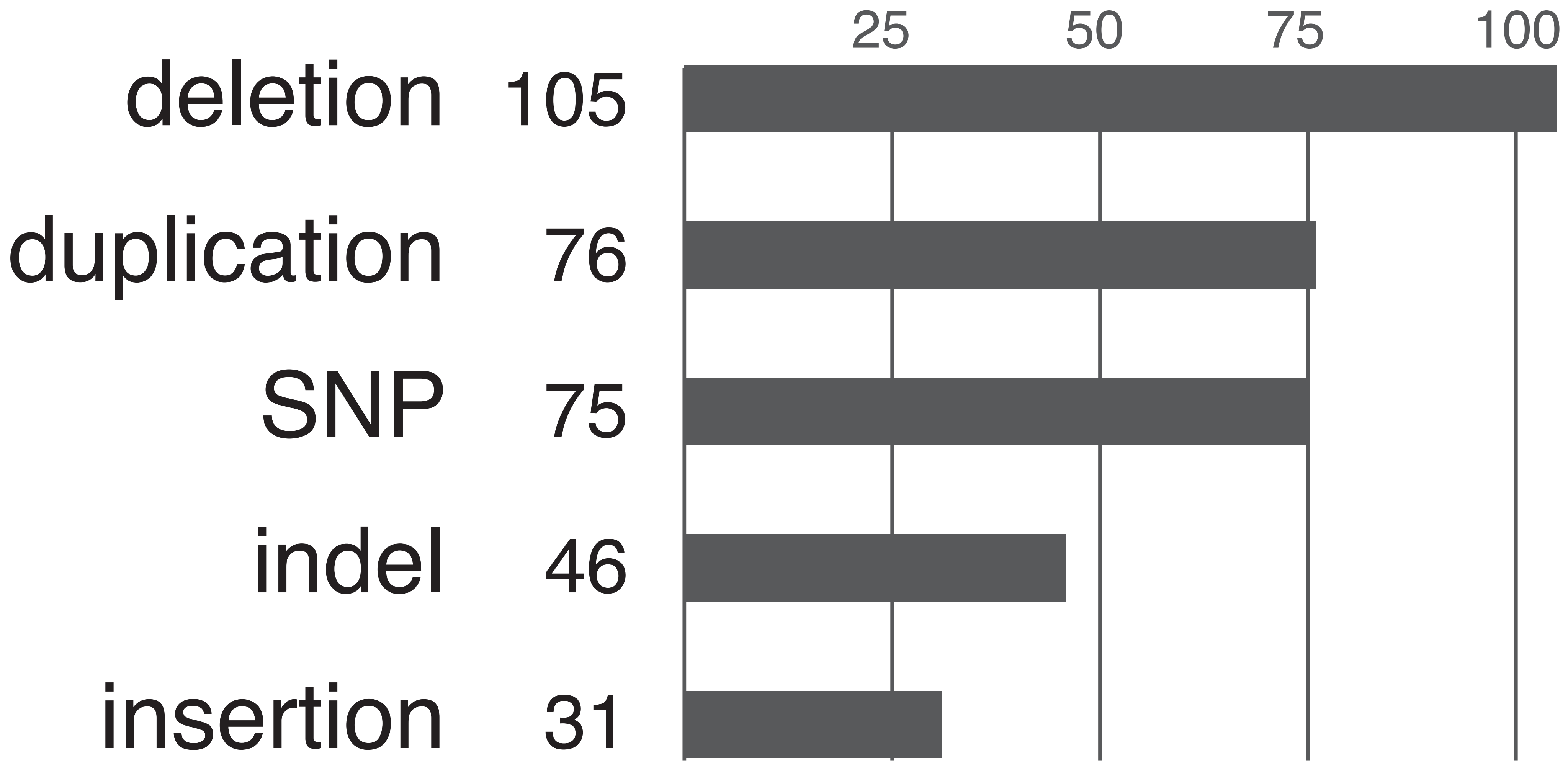
FREQUENCY OF VARIATION BY TYPE



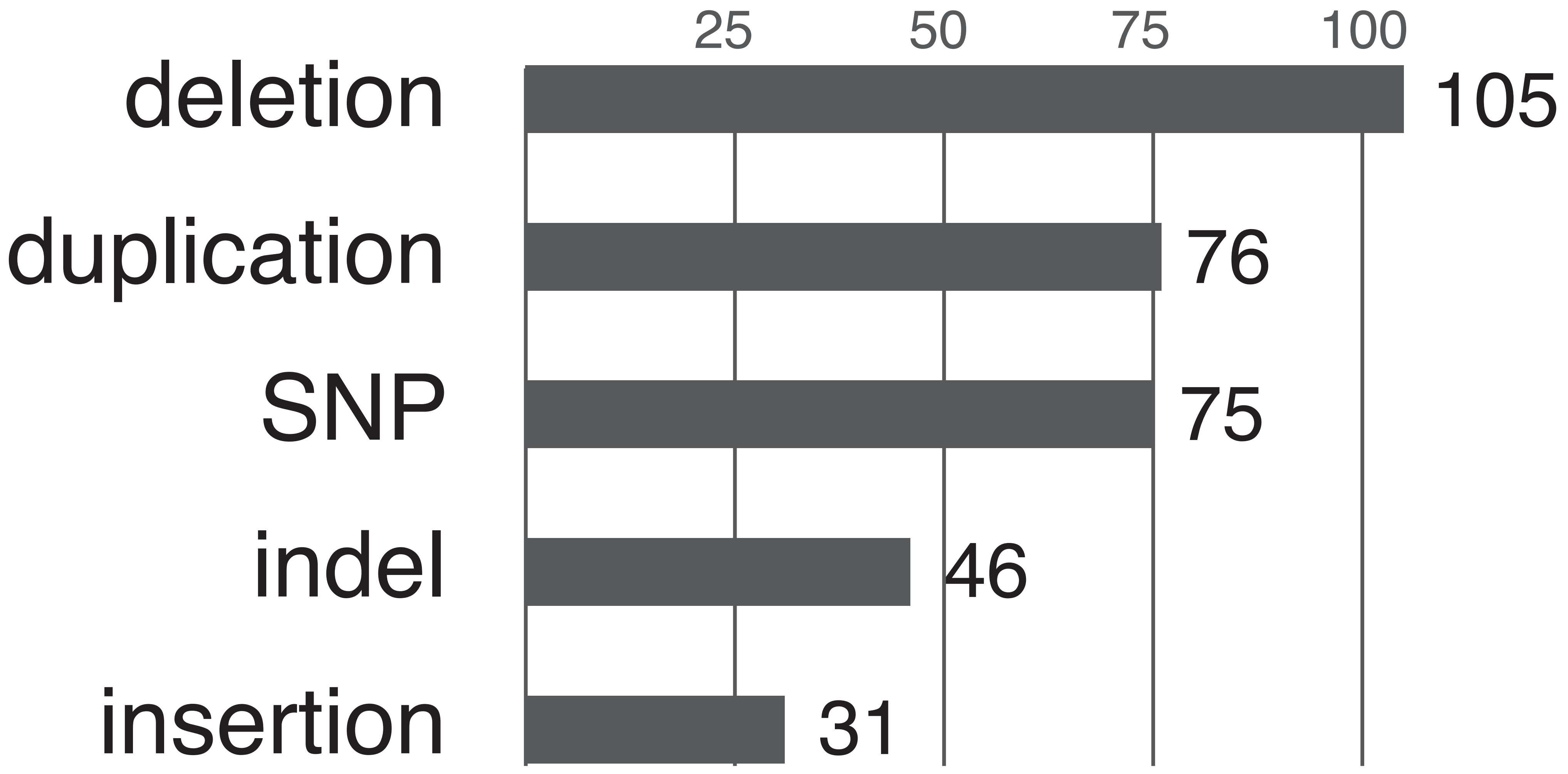
FREQUENCY OF VARIATION BY TYPE



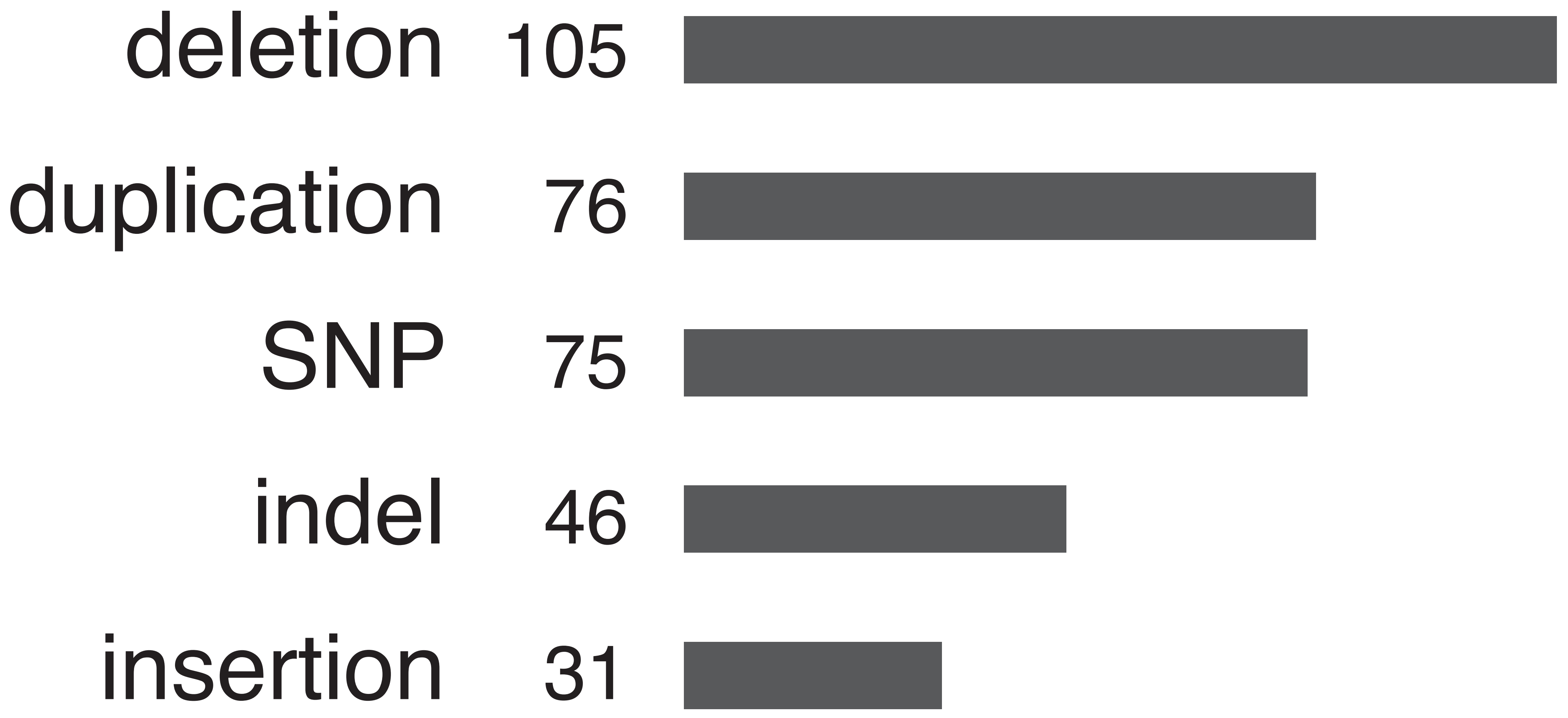
FREQUENCY OF VARIATION BY TYPE



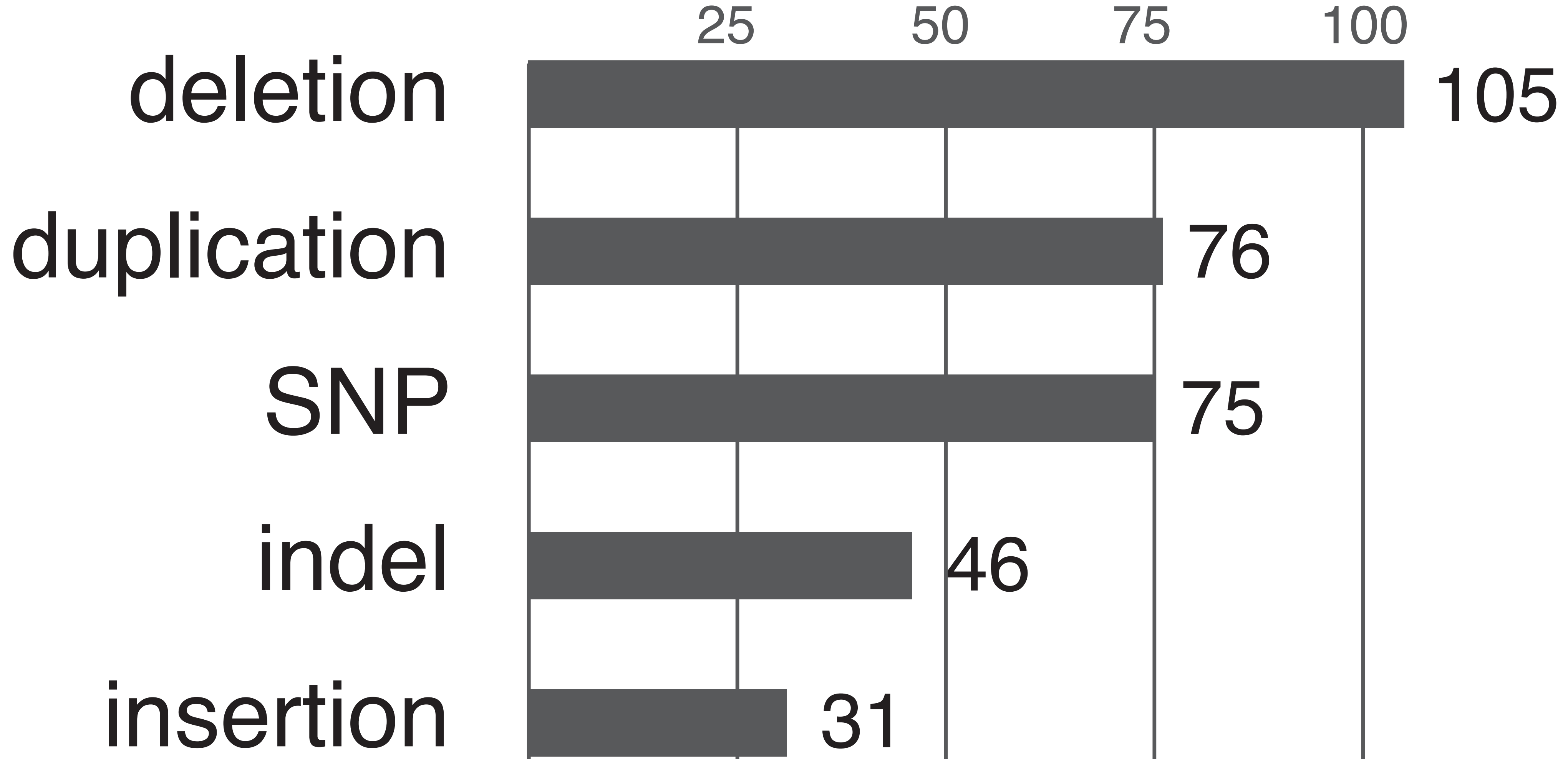
FREQUENCY OF VARIATION BY TYPE



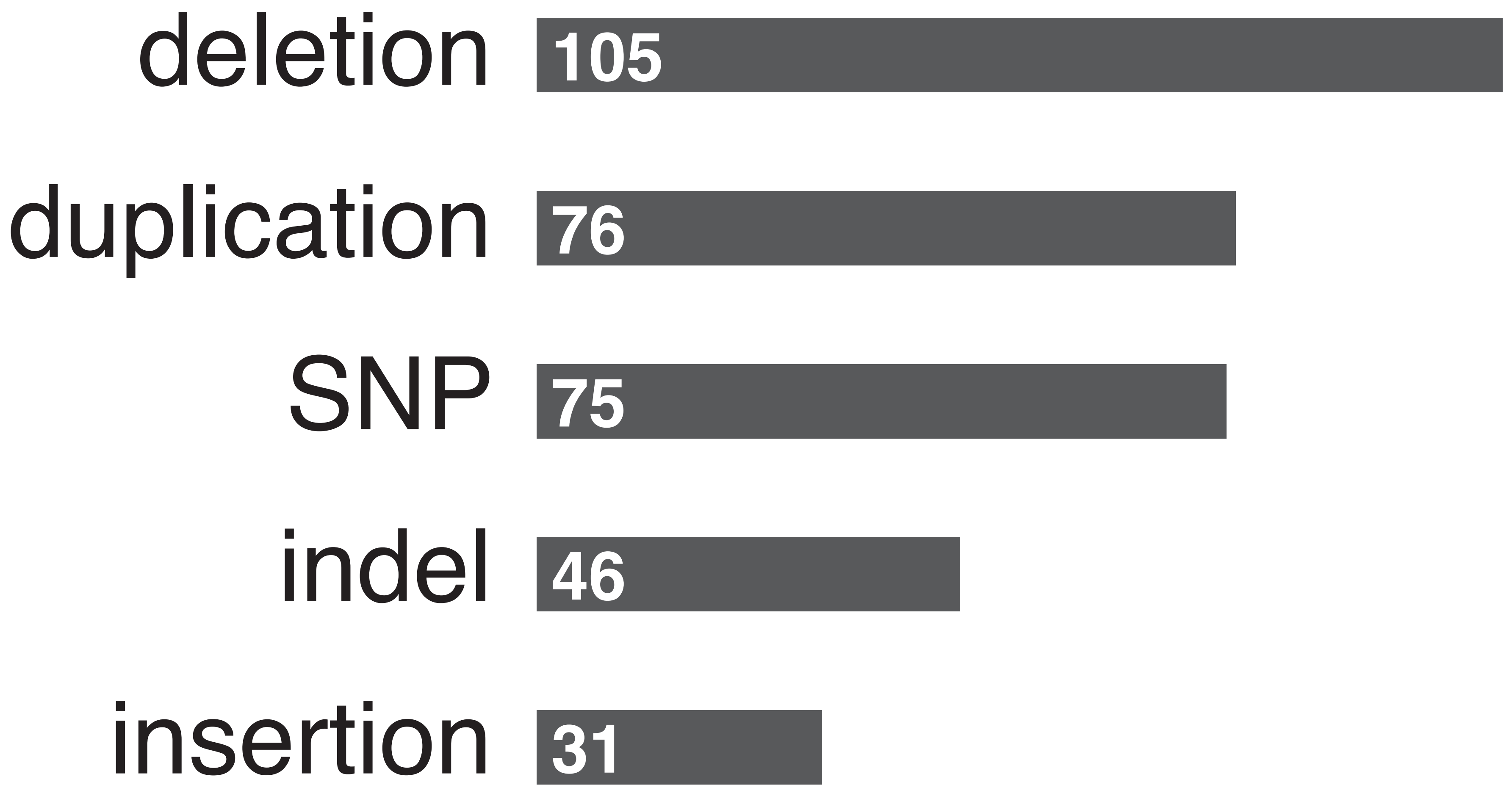
FREQUENCY OF VARIATION BY TYPE



FREQUENCY OF VARIATION BY TYPE



FREQUENCY OF VARIATION BY TYPE



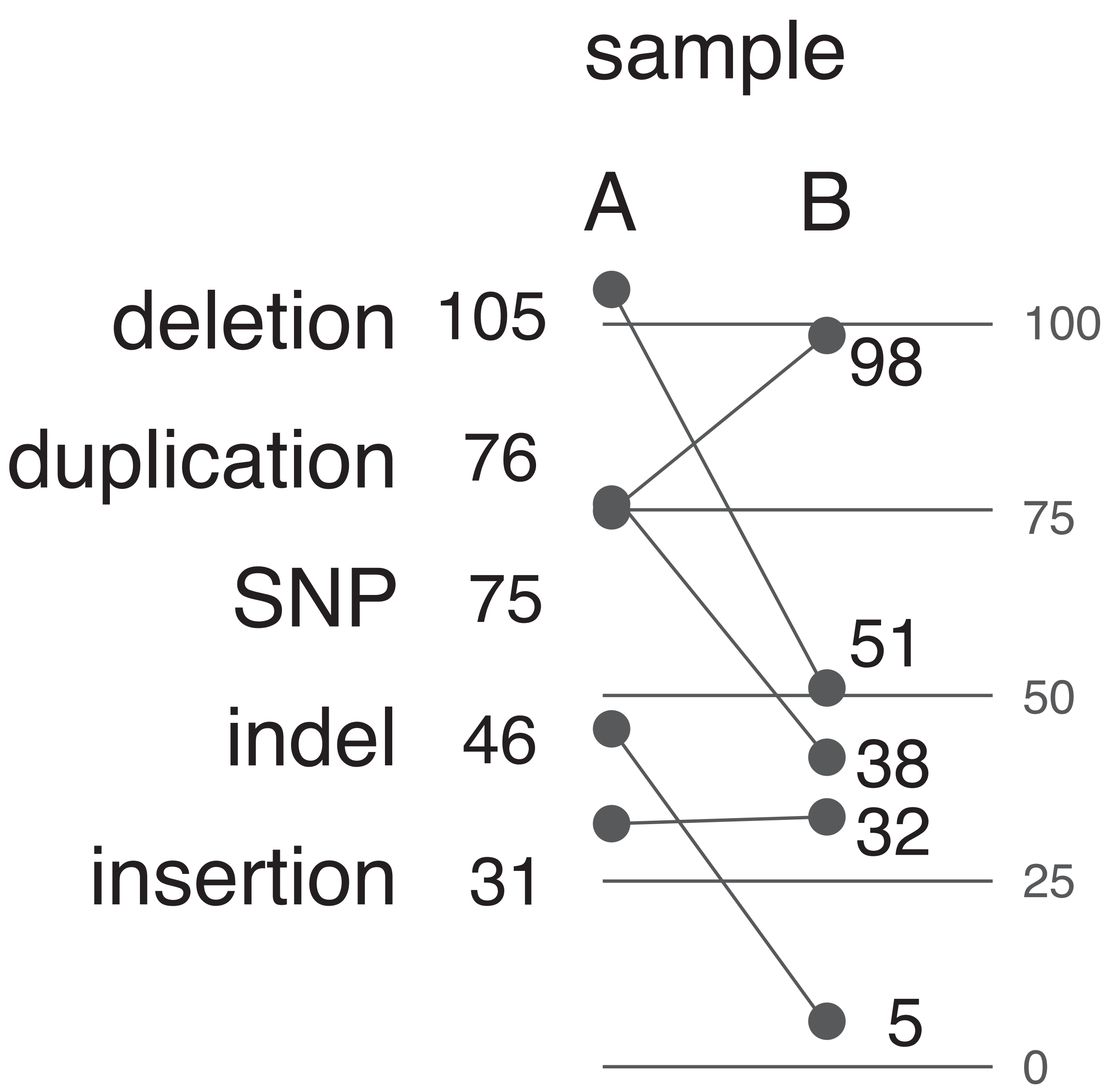
FREQUENCY OF VARIATION BY TYPE

	sample	
	A	B
deletion	105	51
duplication	76	38
SNP	75	98
indel	46	5
insertion	31	32

FREQUENCY OF VARIATION BY TYPE

	sample	
	A	B
deletion	105	51
duplication	76	38
SNP	75	98
indel	46	5
insertion	31	32

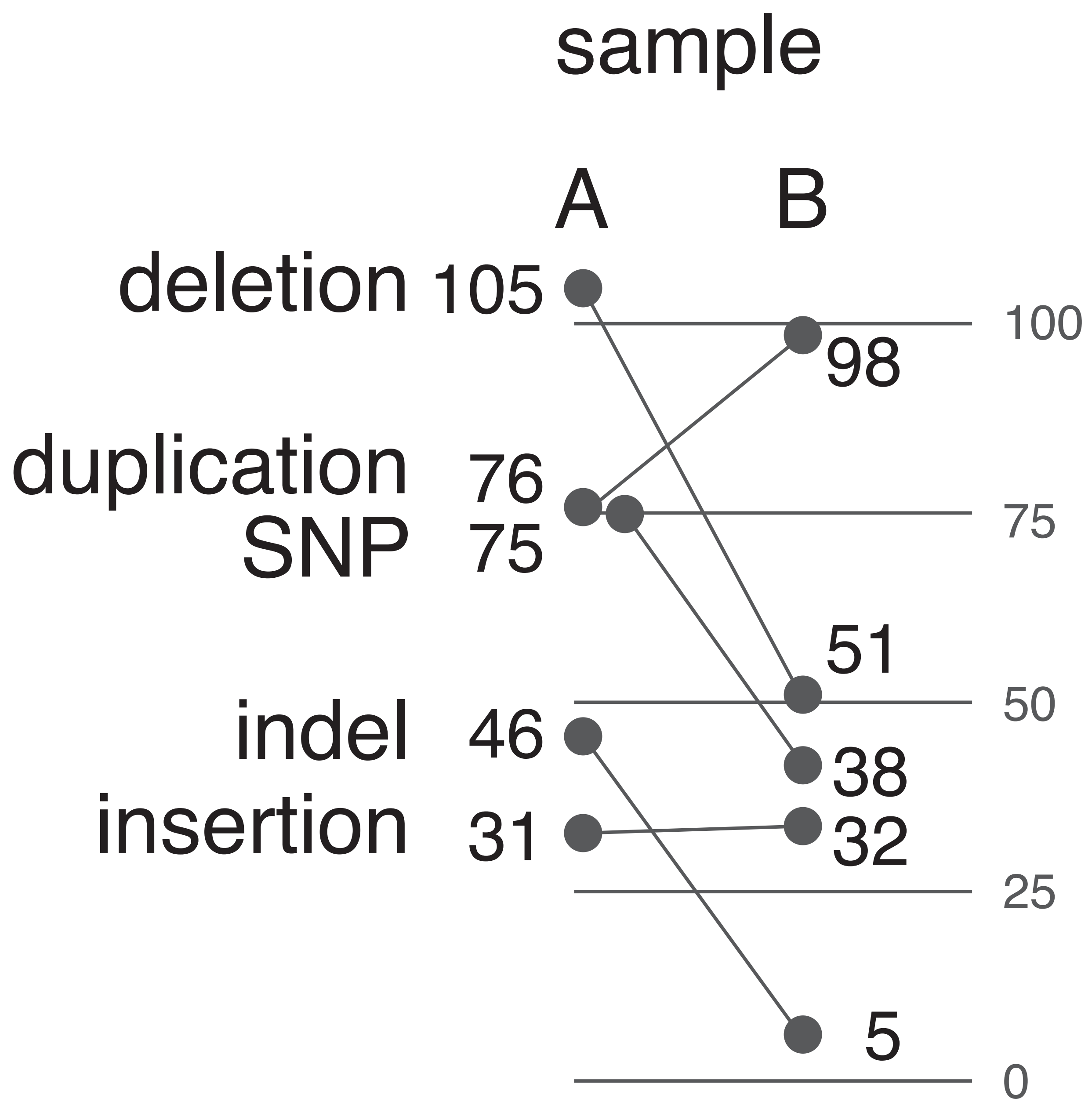
FREQUENCY OF VARIATION BY TYPE



FREQUENCY OF VARIATION BY TYPE

	sample	
	A	B
deletion	105	51
duplication	76	38
SNP	75	98
indel	46	5
insertion	31	32

FREQUENCY OF VARIATION BY TYPE

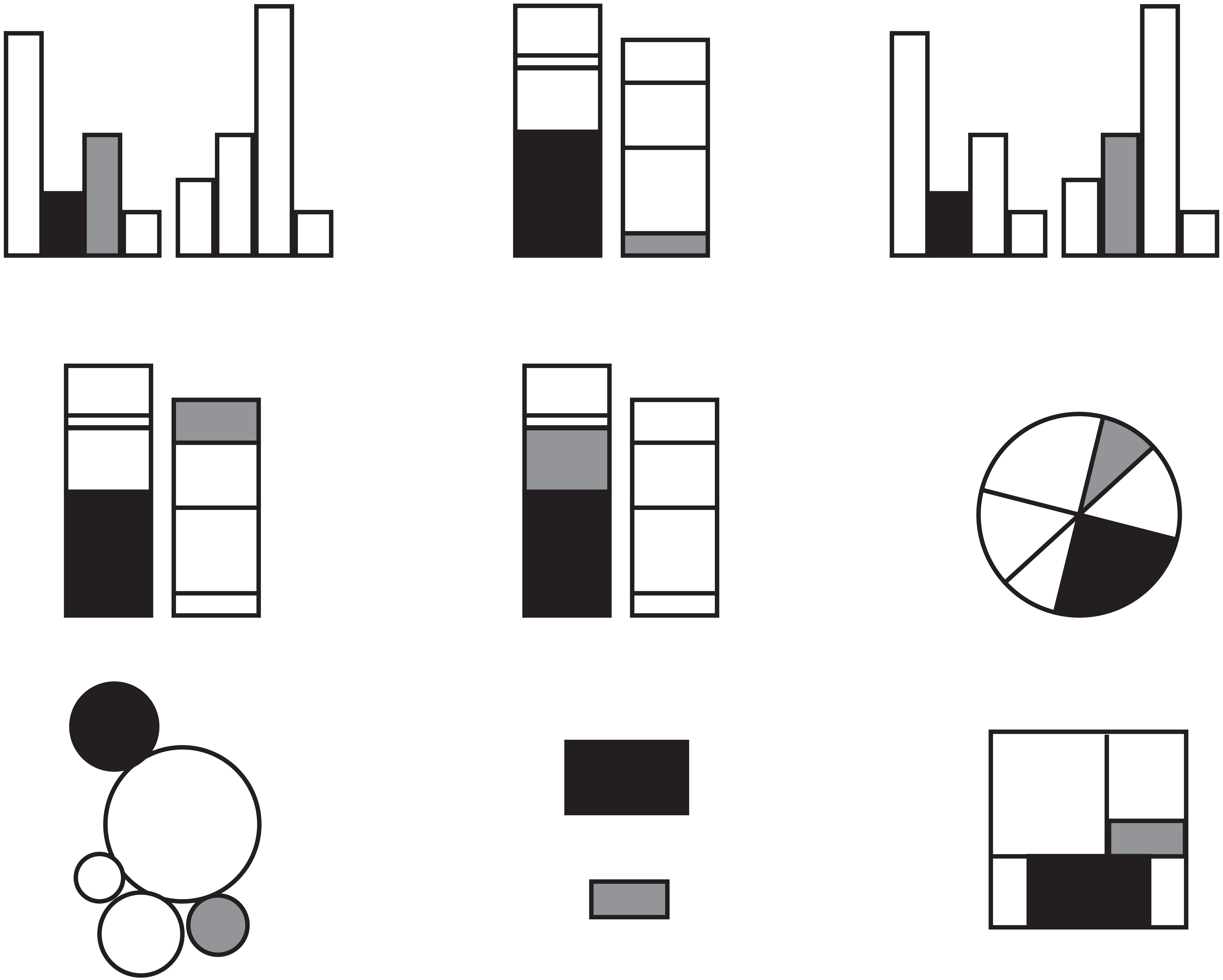


$\log_2(\text{error})$

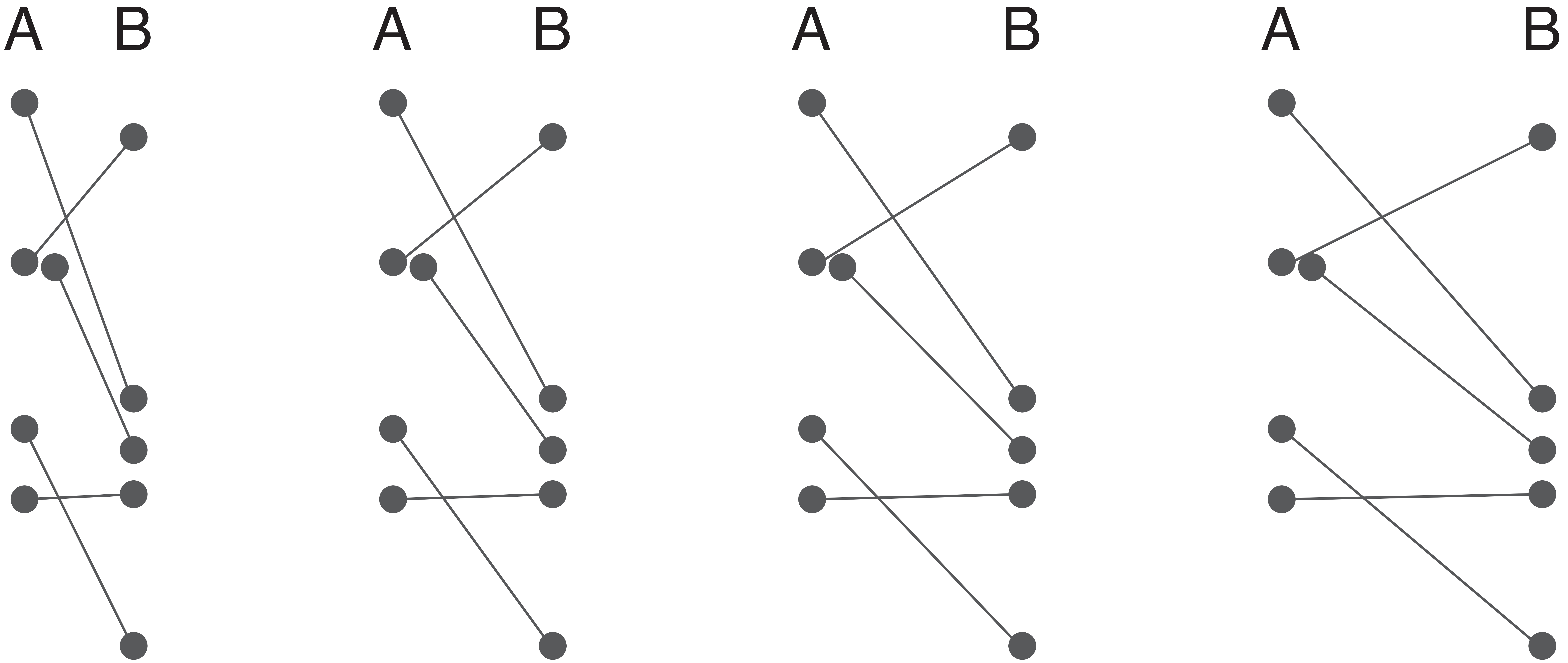
1.5

2

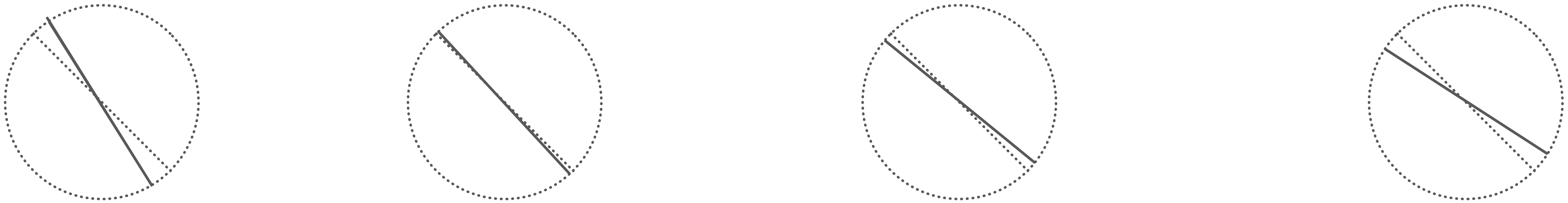
2.5



Proceedings of the 28th international conference on Human factors in computing systems (2010) ACM: Atlanta, Georgia, USA. 203–212.



average line angle



We’ve started with some simple examples. There’s a good reason for this.

Always appreciate and use to your advantage the fundamental principles of data visualization. Simple examples embody these.

At no point, do you ever eject these principles. I don’t care how big your data set is. Sure, they might be more nuanced or interrelated, if you’re showing a lot of stuff on the page, but fundamentally you’re still thinking about the same things:

Am I using ink responsibly?

Am I drawing shapes whose position and size can be accurately judged?

Am I making comparisons easy to make?

Am I visually emphasizing the things that are relevant?

Once your data size grows, a popular technique in visualization is called small multiples. Here, you break your data down into a large number of smaller sets and represent each set with the same primitive encoding. So, you might have a matrix of scatter plots or bar plots. Each of those individual small plots must be handled with care.

created by

Martin Krzywinski, Kim Bell-Anderson & Philip Poronnik

written and designed by

Martin Krzywinski

production

One Ski Digital Media Productions

with financial support by

University of Sydney

filmed at

University of Sydney, Australia

EXERCISE 1

Represent the "frequency of variation by type" two-column table as a bar plot.

Pay careful attention to bar width and the distance within groups (e.g. the deletion bars for A and B) and between groups (the deletion and duplication groups). Start with a ratio of

bar width : within : between = 1 : 0.2 : 1.5

How does this look to you? Incorporate the golden ratio ($\varphi = 1.62$) and its inverse ($1/\varphi = \varphi - 1 = 0.62$) and use ratio of

bar width : within : between = $1 : 1/\varphi^2 : \varphi$

Does this look better?

Reading vertical text is difficult—draw both vertical and horizontal versions of your bar plot. Which one requires you to angle text more?

FREQUENCY OF VARIATION BY TYPE

	sample	
	A	B
deletion	105	51
duplication	76	38
SNP	75	98
indel	46	5
insertion	31	32

EXERCISE 2

Differences in data are important—sometimes more important than the data themselves. Think of the differences as the data.

For example, in the "frequency of variation by type" two-column table, the number of deletions drops by 54 from 105 in A to 51 in B.

Try to add this value to your bar plot. Is a relative or absolute difference meaningful here? Suggest reasons to show one or the other.

FREQUENCY OF VARIATION BY TYPE

	sample	
	A	B
deletion	105	51
duplication	76	38
SNP	75	98
indel	46	5
insertion	31	32

EXERCISE 3

Draw a scatter plot using the data in the "frequency of variation by type" two-column table. Use the number of variations in A on the X axis and the relative change on the Y axis.

Does this representation make some questions easier to answer?

Address the challenge of labeling the points—this can be difficult!

What has been gained?

FREQUENCY OF VARIATION BY TYPE

	sample	
	A	B
deletion	105	51
duplication	76	38
SNP	75	98
indel	46	5
insertion	31	32

EXERCISE 4

You are a terrific visual calculator and comparer.

How much faster can you compare lengths than numbers? Let’s check.

Print this slide. Now, with pen and paper, time yourself to see how quickly you can identify the larger of each of the two numbers in each pair. Next, time yourself to see how quickly you can identify the longer of the two lengths in each pair.

Are you surprised at the timing difference? Which task was more tiring?

Notice that as the numbers get larger the longer the comparison takes—your eye has to travel further. Can you think of a way to align the number pairs differently to speed this up?

6	9
3	6
1	5
3	5
9	4
9	1
1	5
5	4
3	1
1	9
93	12
12	99
79	53
17	67
57	27
99	45
45	46
54	98
14	41
65	85
424	978
227	617
379	437
394	388
436	755
237	263
774	973
819	687
517	717
655	289