Data Visualization

Let's begin at the beginning...

Jackie Wirz, Alison Hill, & Steven Bedrick
CSE 629, 9/27/16

Our game plan for today:

1. Course logistics & overview
2. What are we really trying to do?
3. Let's talk about data...
4. Families of visualization
5. Data-Ink Ratios

Course Logistics:

Website:

http://cslu.ohsu.edu/~bedricks/TBD

Homework:

Generally low-key, will occur throughout course.

Final Project:

Begin thinking *now* about data sets;
We will be having you work on them in groups later in the course.

Course Logistics:

Format:

Tuesdays: lecture
Thursdays: lab.

Textbook (Optional but suggested)

Nathan Yau's "Visualize This!" and "Data Points"

Disclaimers!

1. This is a huge topic!

2. These are opinions!

3. There are no *absolute* rules!

4. Use yer’ noggin!

What does it mean to visualize data?

There exist many definitions, but we will start with:

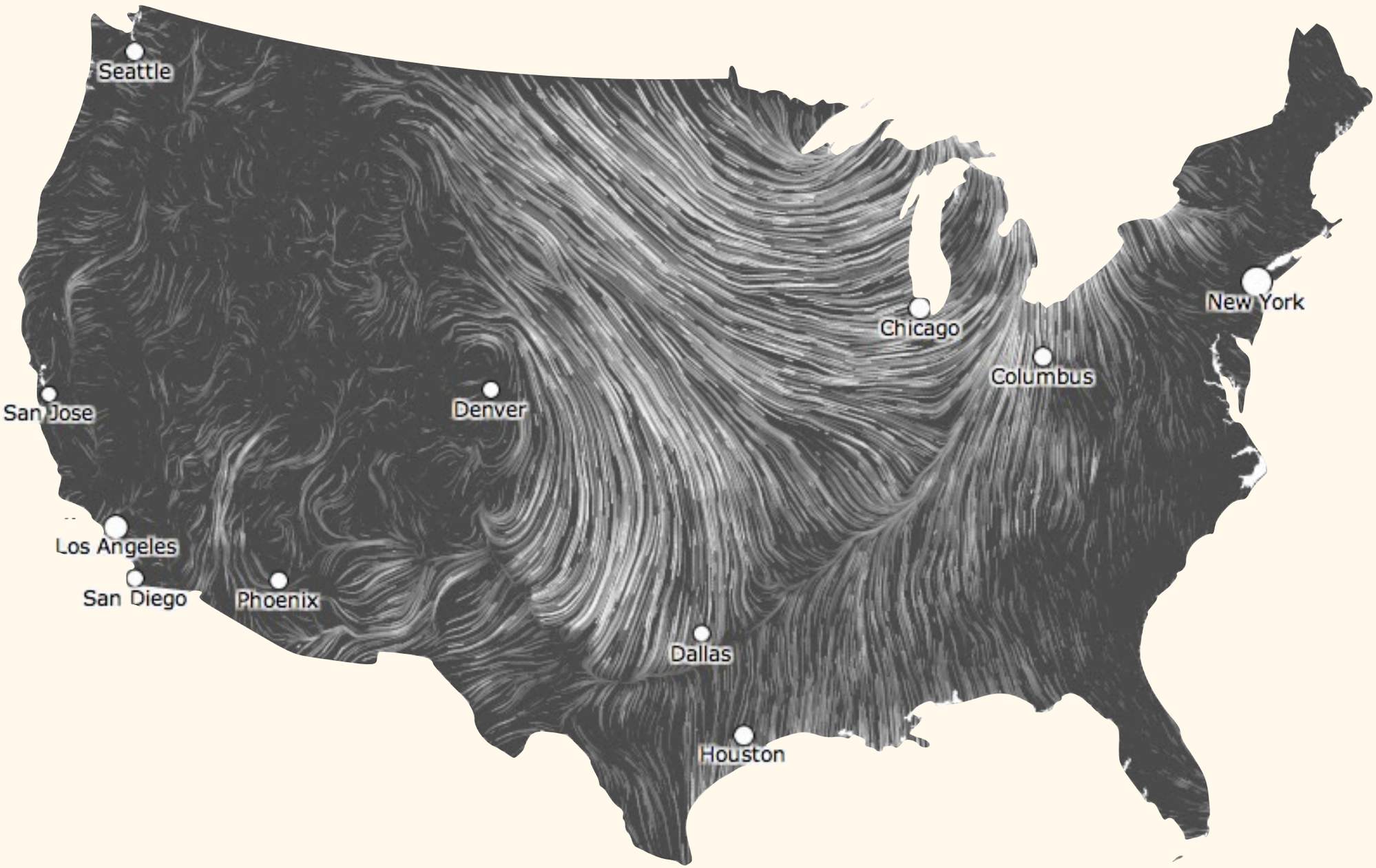
Creating a visual abstraction of data in order to make more easily understood.

Why do we need to do this?

No, seriously: why do we need to do this?

Why not just look at data directly?

Sometimes we have too much data for this to be practical:



http://hint.fm/wind/

What about when we don't have so much data?

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
7	4.82	7	7.26	7	6.42	8	6.89
5	5.68	5	4.74	5	5.73	19	12.5

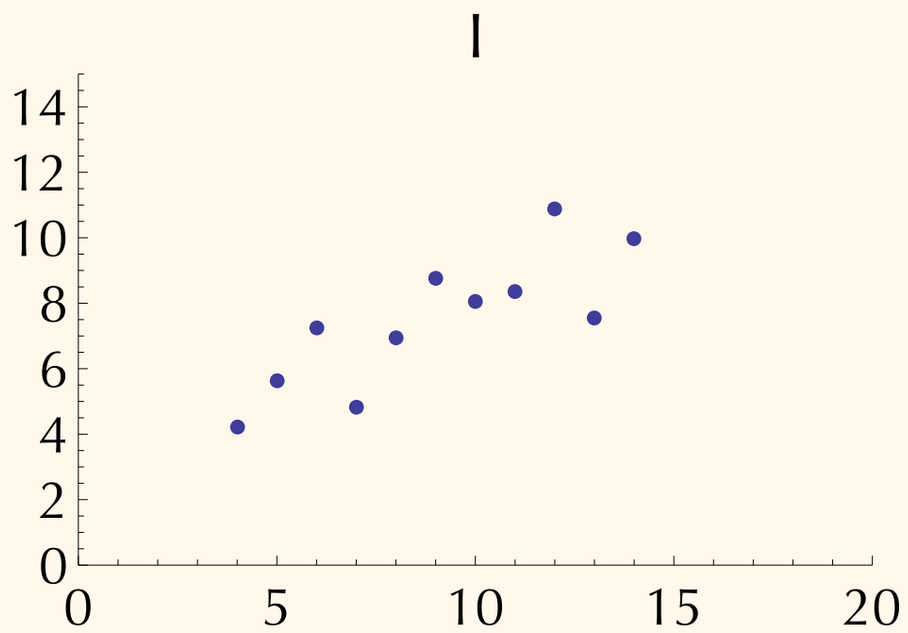
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
7	4.82	7	7.26	7	6.42	8	6.89
5	5.68	5	4.74	5	5.73	19	12.5

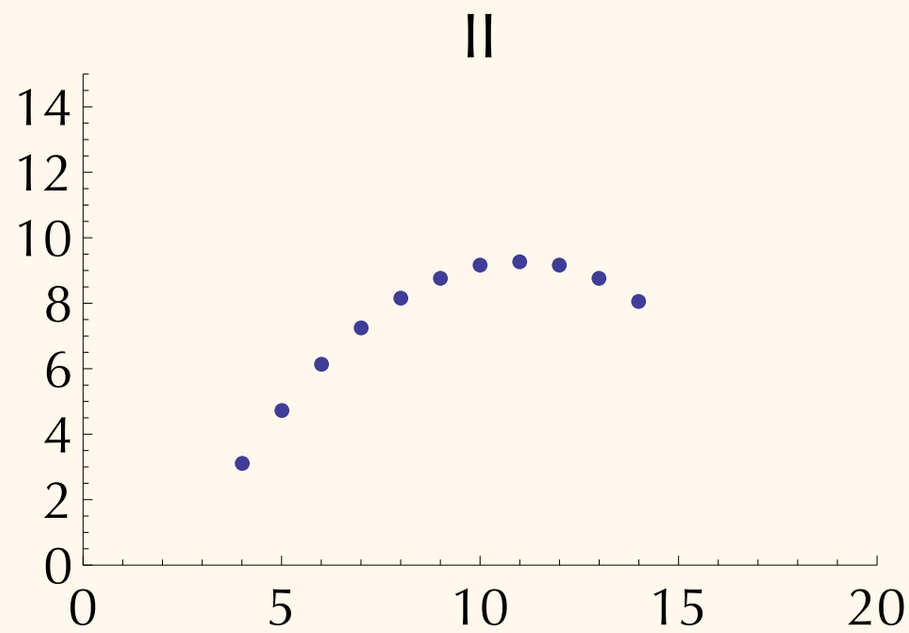
$$\bar{x} = 9$$

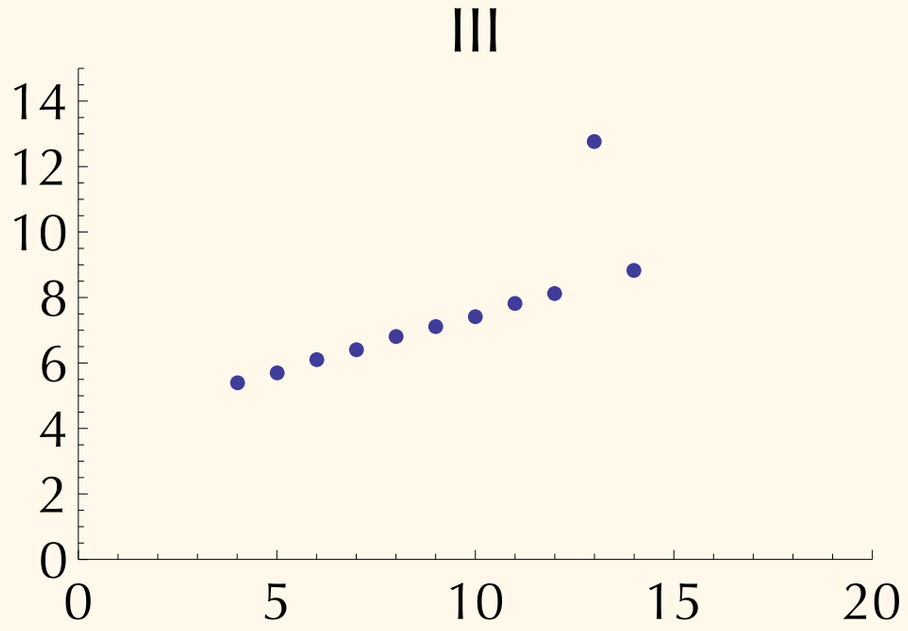
$$\bar{y} = 7.5$$

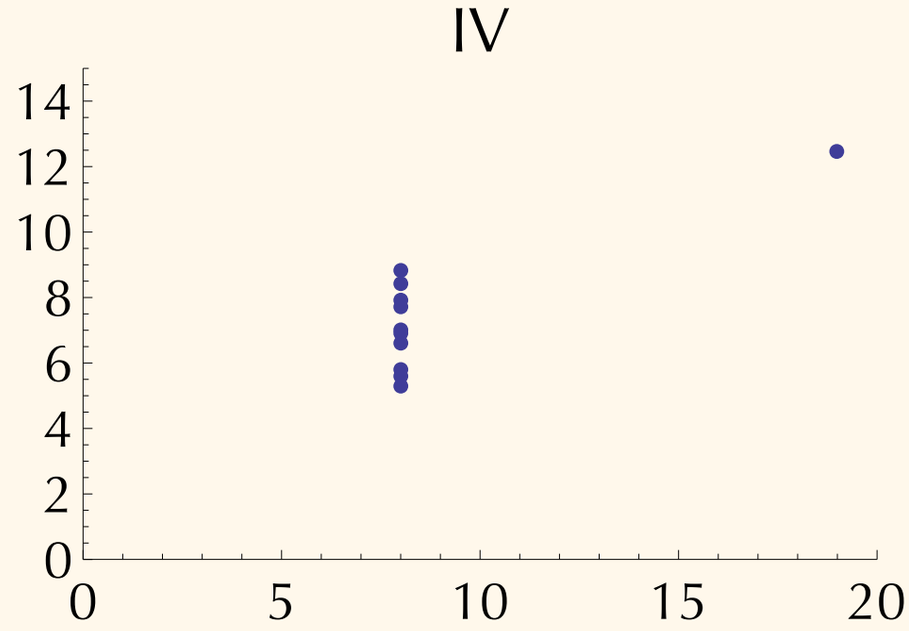
$$r^2 = 0.67$$

$$\hat{y} = 3 + 0.5x$$


I
14
12
10
8
6
4
2
0
0
5
10
15
20


II
14
12
10
8
6
4
2
0
0
5
10
15
20


III
14
12
10
8
6
4
2
0
0
5
10
15
20

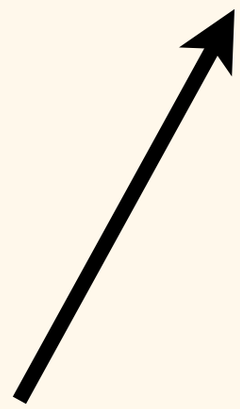

IV
14
12
10
8
6
4
2
0
0
5
10
15
20

Let’s try another definition/explanation:

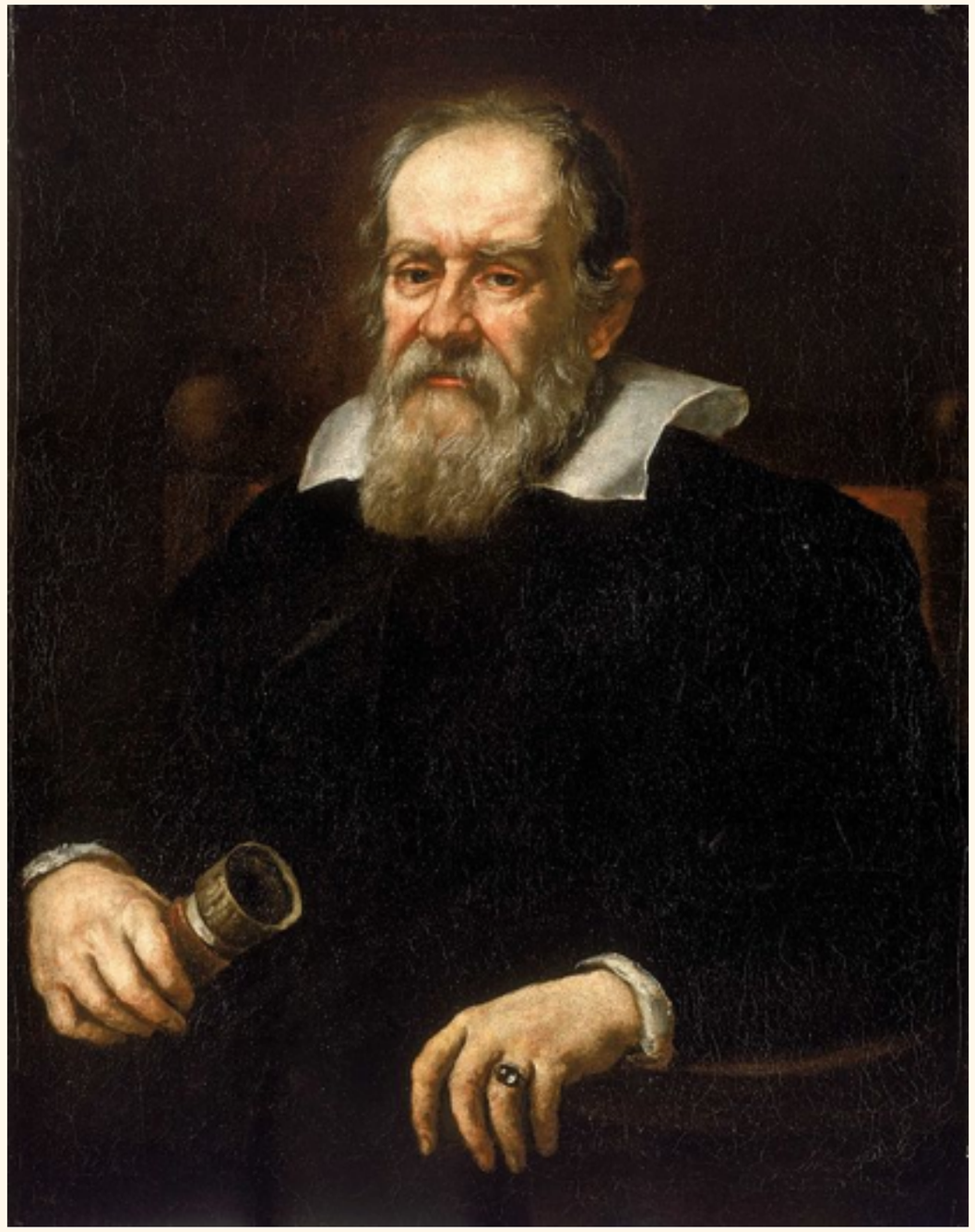
Visualization lets us use our human perceptual capabilities to gain insights into abstract data sets.

Let’s try another definition/explanation:

Visualization lets us use our human perceptual capabilities to gain insights into abstract data sets.

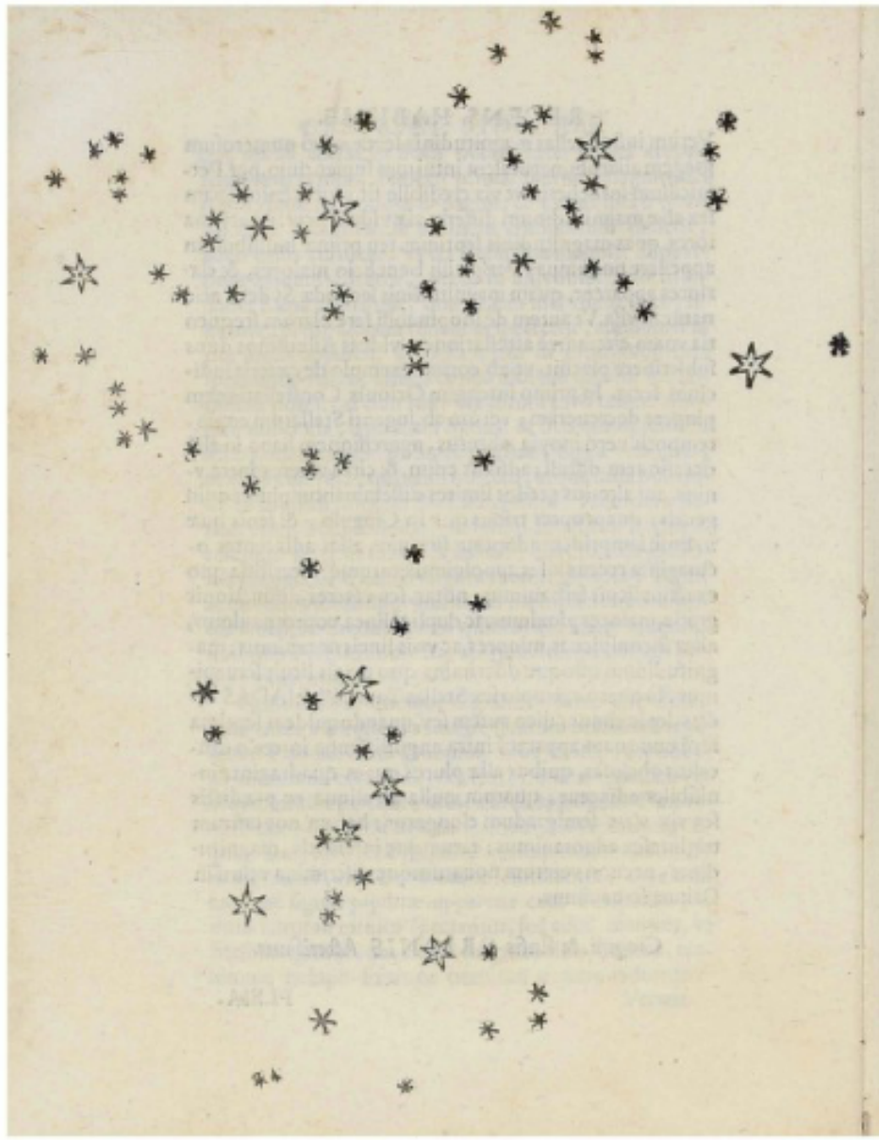


A good visualization makes us see things in our data that we couldn’t otherwise see!



Galileo Galilei
1564–1642

https://commons.wikimedia.org/wiki/File:Justus_Sustermans_-_Portrait_of_Galileo_Galilei,_1636.jpg#mediaviewer/File:Justus_Sustermans_-_Portrait_of_Galileo_Galilei,_1636.jpg



PLEIADVM CONSTELLATIO.

quòd tertio loco à nobis fuit obſeruatum, eſt ipſiuſmet LACTEI Circuli eſſentia, ſeu materies, quam Perſpicilli beneficio adeò ad ſenſum licet intueri, vt & altercationes omnes, quæ per tot ſæcula Philoſophos exercucia

OBSERVAT. SIDEREAE

Ori. ✱ ✱○ ✱ . Occ.

Stella occidentalioni maior, ambæ tamen valdè conſpicuæ, ac ſplendidæ: vtra quæ dikabat à Ioue ſcrupulis primis duobus; tertia quoque Stellula apparere cœpit hora tertia prius minimè conſpecta, quæ ex parte orientali Iouem ferè tangebat, eraique admodum exigua. Omnes ſeruat in eadem recta, & ſecundum Ecypticæ longitudinem coordinatæ.

Die decimatertia primum à me quatuor conſpectæ fuerunt Stellulæ in hac ad Iouem conſtitutione. Erant

13 January ✱ ○ ✱ ✱ ✱ ✱ ○ ✱ ✱ Occ.

Stellæ occidentales, & vna orientalis; lineam proximè

tenebant; media enim occidentalium paululum à recta Septentrionem versus deflectebat. Aberat orientalior à Ioue minuta duo: reliquarum, & Iouis intercapedines erant ſingulæ vnius tantum minuti. Stellæ omnes eandem præ ſe ferebant magnitudinem; ac licet exiguæ, laudiſſimæ tamen erant, ac fixis eiuſdem magnitudinis longè ſplendidiores.

Die decimaquarta nubibus fuit tempeſtas.

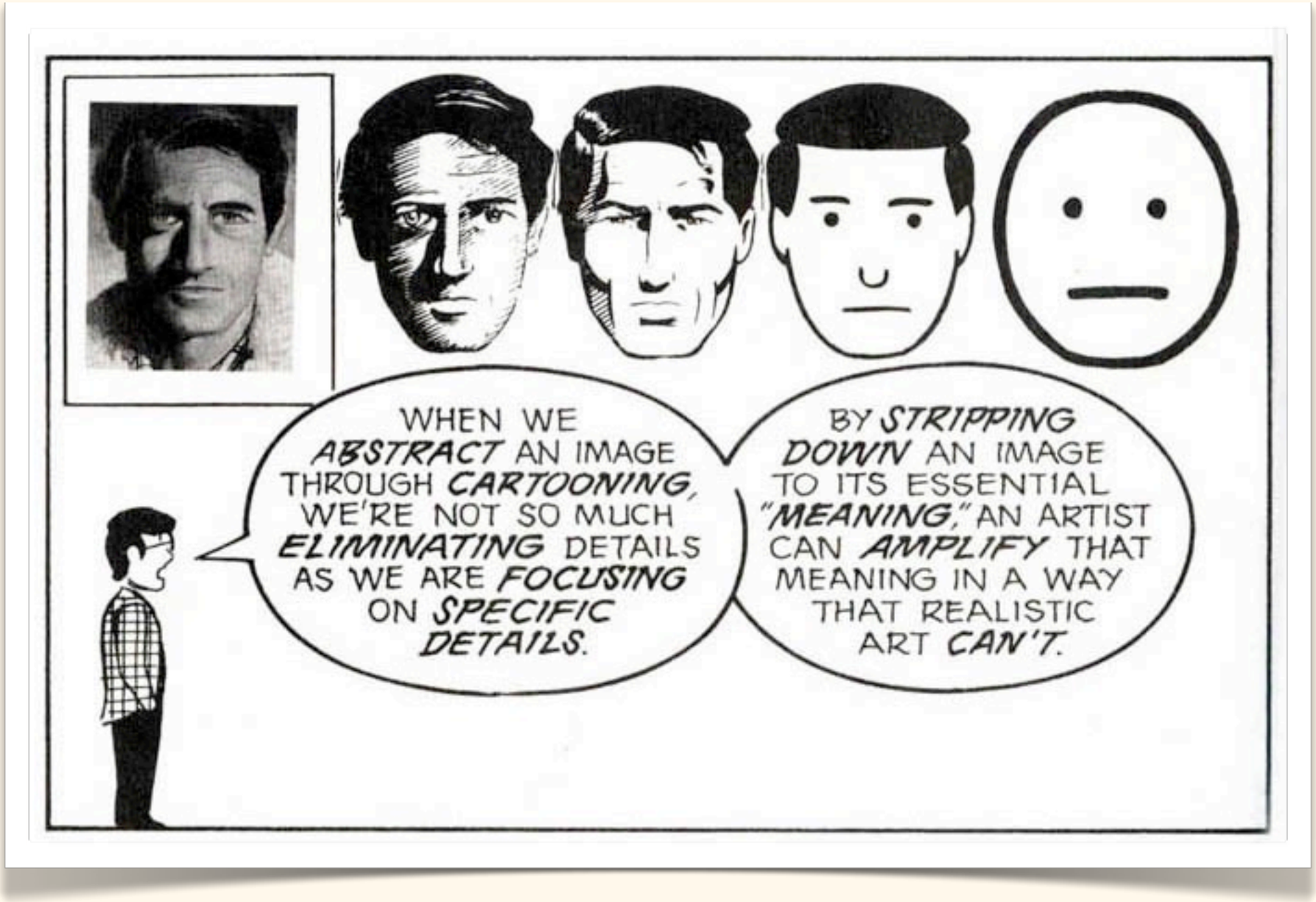
Die decimaquinta, hora noctis tertia in proximè

15 January ○ ✱ ✱ ✱ ✱ ○ ✱ ✱ ✱ Occ.

habitudine quatuor Stellæ ad Iouem;

ſunt omnes: ac in eadem proxim recta linea diſpoſitæ; quæ enim tertia à Ioue numerabatur paululum

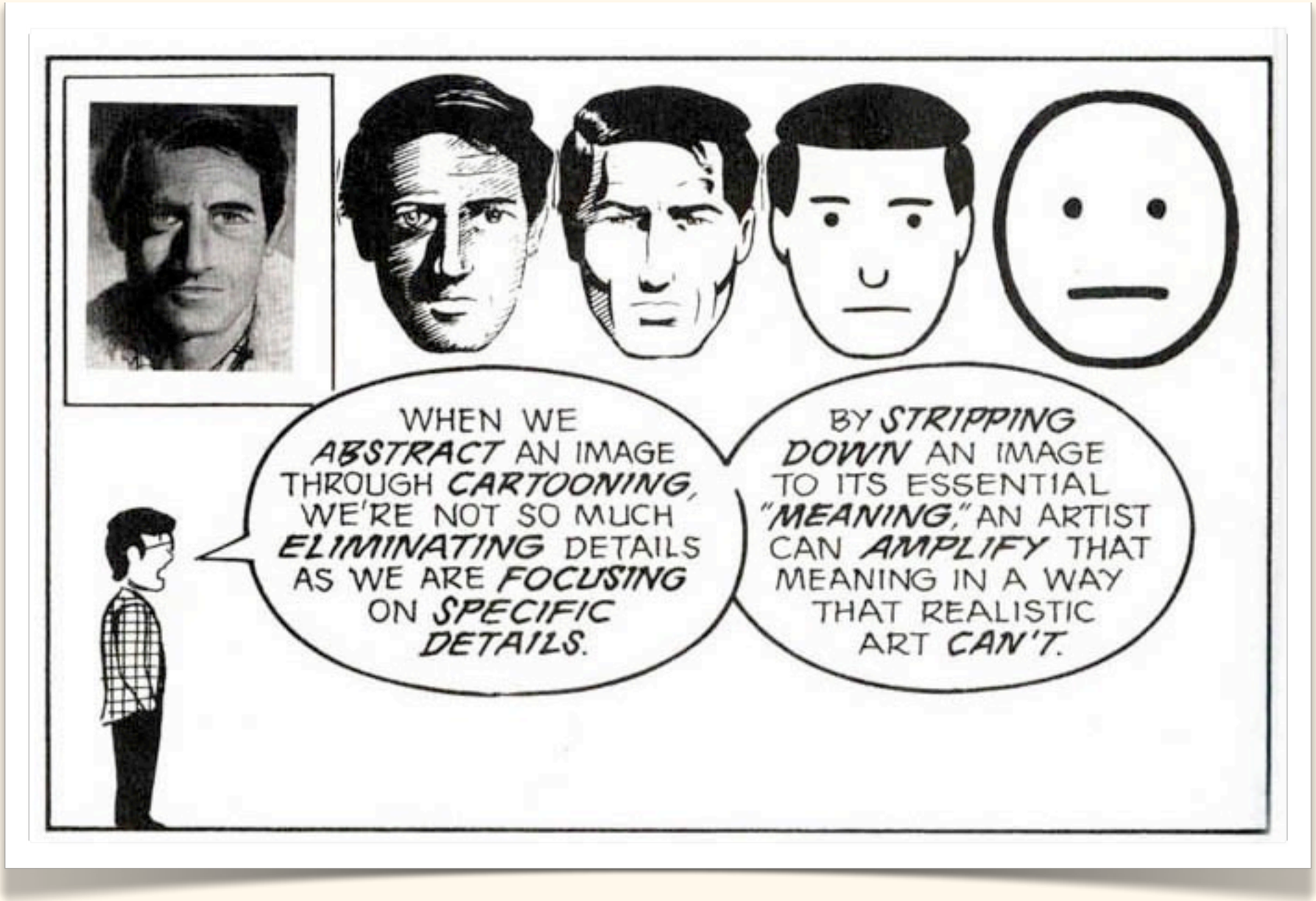
Another way to think about it:



Like comics, visualizations let us abstract away less-relevant details...

Scott McCloud, “Understanding Comics”, Chapter 2

Another way to think about it:

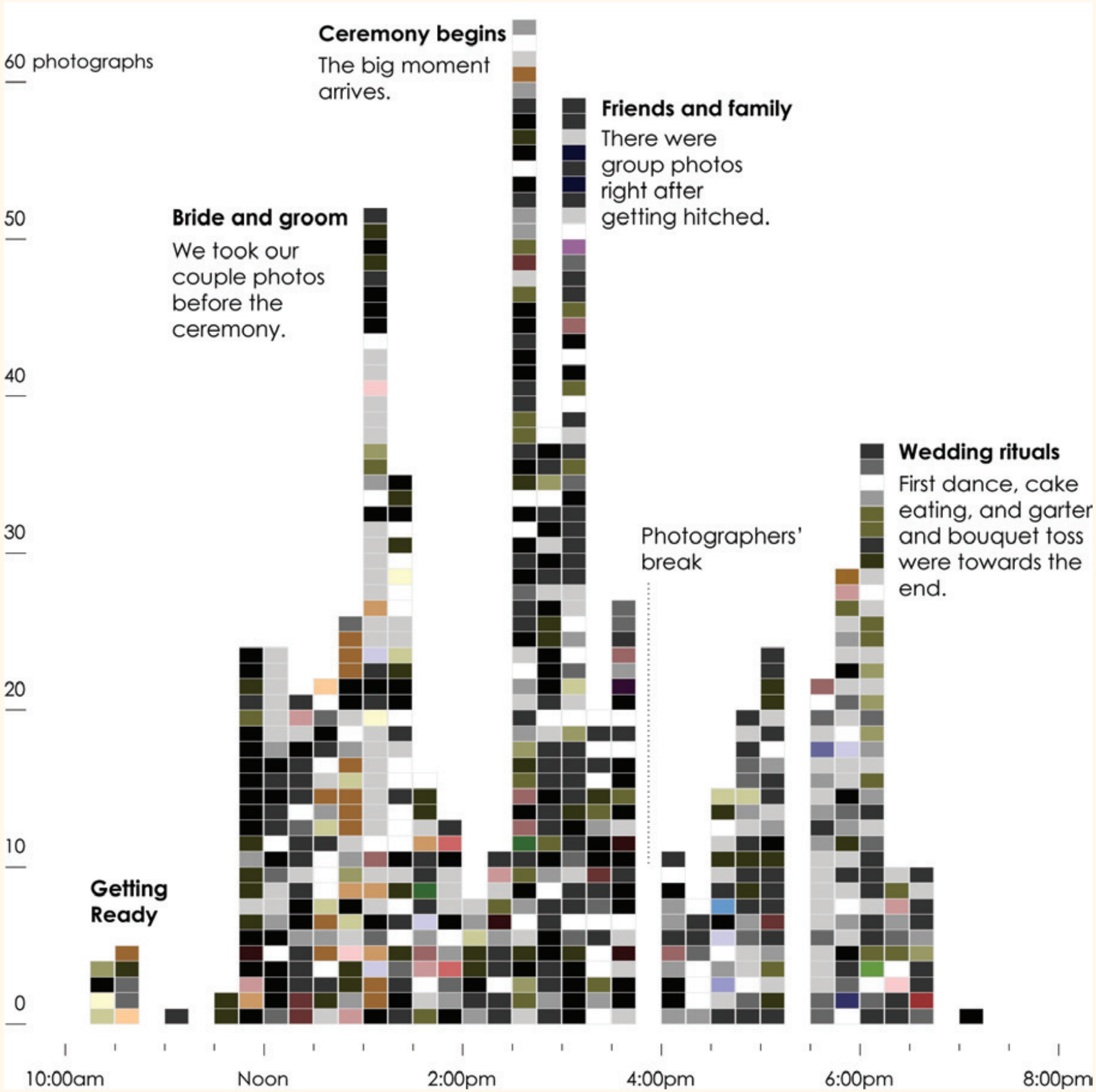


... and allow us to highlight the particular ideas and information we wish to convey.

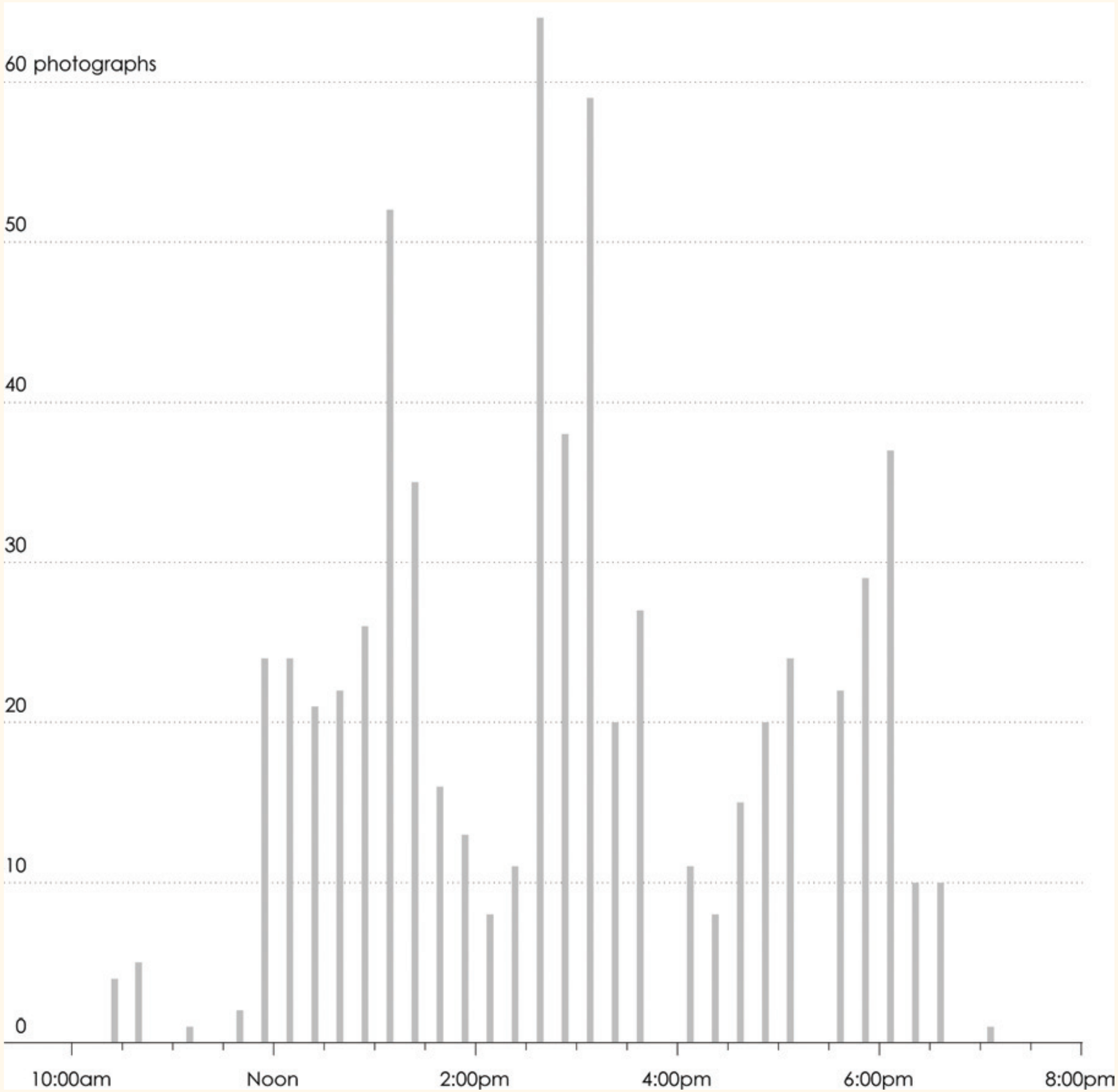
Scott McCloud, “Understanding Comics”, Chapter 2

An example:

Taken from Nathan Yau's "Data Points," chapter 1



Taken from Nathan Yau's "Data Points," chapter 1



Taken from Nathan Yau's "Data Points," chapter 1

Another way to think about it:

When we make a visualization from our data, we are telling a story.

Another way to think about it:

There is information buried in our data, and we are finding the best way to make it accessible.

Important corollary:

“When we make a visualization from our data, we are telling a story.”

This means that a successful visualization has to have something to say or a question to answer...

Important corollary:

“When we make a visualization from our data, we are telling a story.”

... which means its creator (you!) needs to know what that story or question is!

Before you start, ask yourself:

What are you trying to say about your data?

What question are you trying to ask your data?

Once you know the answer to those questions, everything else follows naturally.

The first commandment of analysis:

Thou shalt know thy data.

There are many ways to think about different types of data...

I'm partial to Colin Ware's taxonomy:

<u>Entities</u>...

... exist in <u>Relationships</u> with one another ...

... which can have <u>Attributes</u> ...

... which can be comprised of multiple <u>Dimensions.</u>

Another dimension to consider: what scale(s) is your data built around?

Stevens' Taxonomy:

Nominal: *apples* and *oranges*

Ordinal: *always, sometimes, never*

Interval: 2011, 2012, 2013

Ratio: 35cm, 45cm, 65cm

Different visualization techniques work with different kinds of data!

Stevens, S. S. (1946). "On the Theory of Scales of Measurement"

Different kinds of questions make use of different dimensions of data...

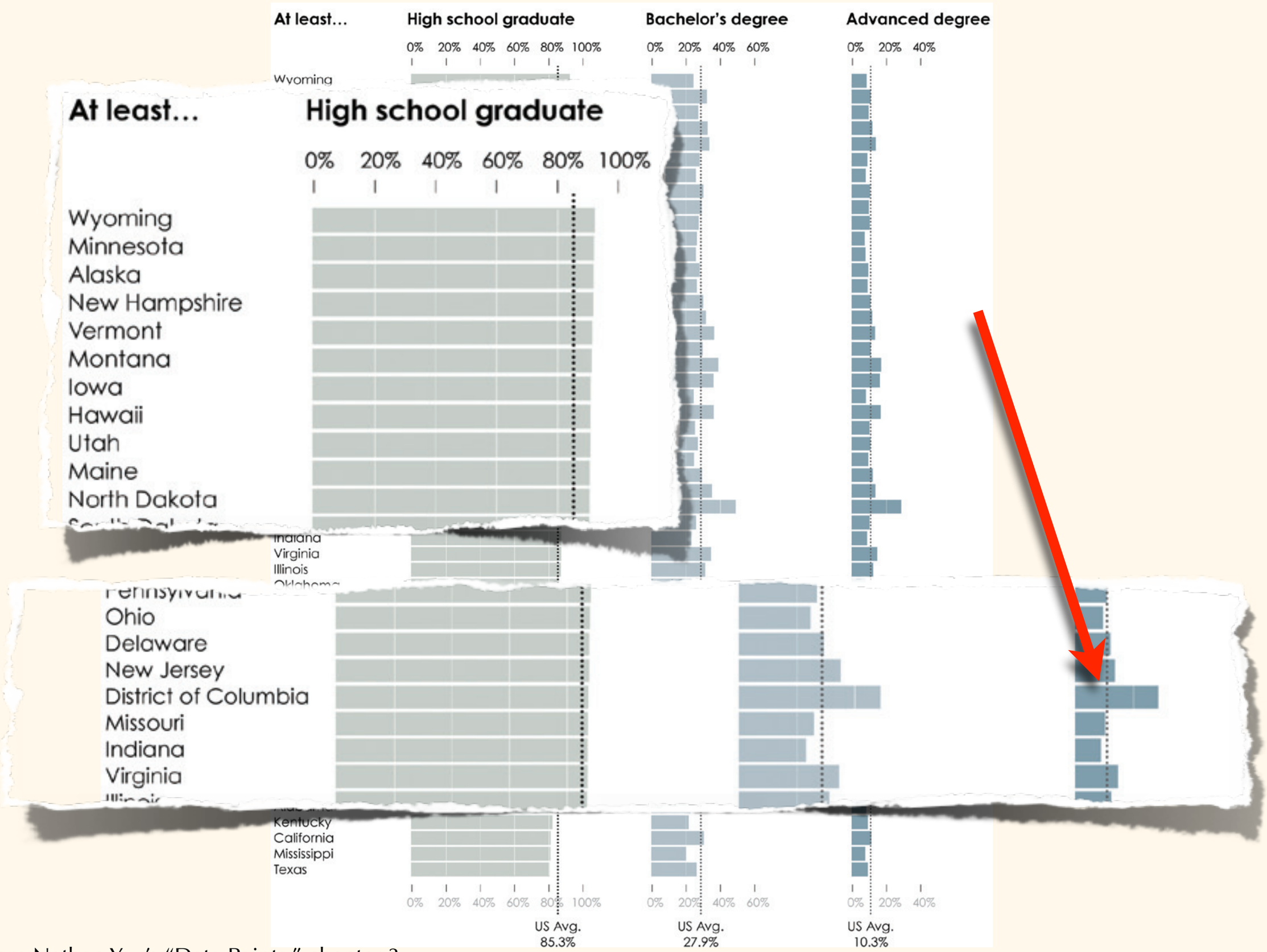
Another example:

Table 233. Educational Attainment by State: 1990 to 2009

[In percent. 1990 and 2000 as of April. 2009 represents annual averages for calendar year. For persons 25 years old and over. Based on the 1990 and 2000 Census of Population and the 2009 American Community Survey, which includes the household population and the population living in institutions, college dormitories, and other group quarters. See text, Section 1 and Appendix III. For margin of error data, see source]

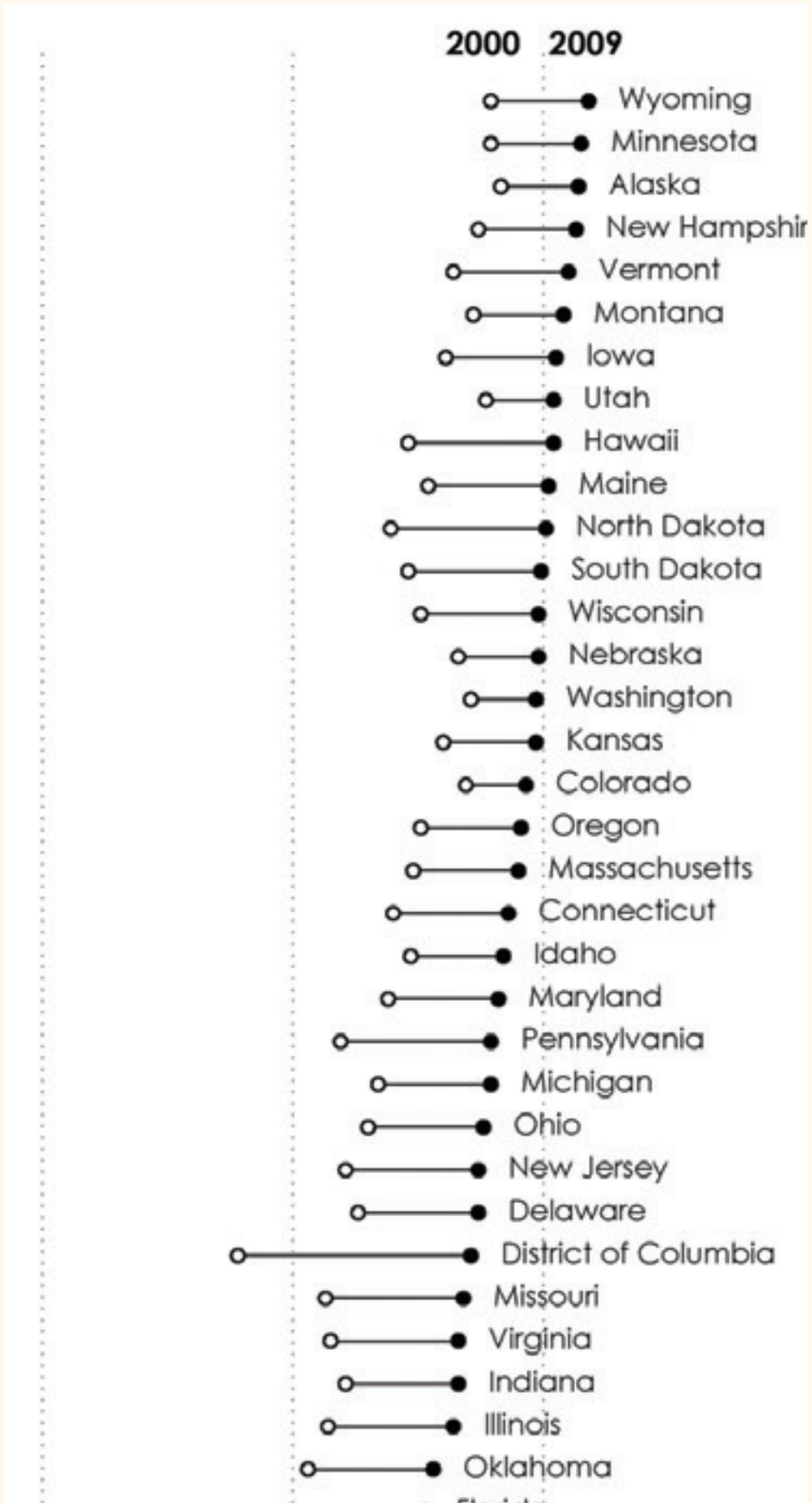
State	1990			2000			2009		
	High school graduate or more	Bachelor's degree or more	Advanced degree or more	High school graduate or more	Bachelor's degree or more	Advanced degree or more	High school graduate or more	Bachelor's degree or more	Advanced degree or more
United States	**75.2**	**20.3**	**7.2**	**80.4**	**24.4**	**8.9**	**85.3**	**27.9**	**10.3**
Alabama	66.9	15.7	5.5	75.3	19.0	6.9	82.1	22.0	7.7
Alaska	86.6	23.0	8.0	88.3	24.7	8.6	91.4	26.6	9.0
Arizona	78.7	20.3	7.0	81.0	23.5	8.4	84.2	25.6	9.3
Arkansas	66.3	13.3	4.5	75.3	16.7	5.7	82.4	18.9	6.1
California	76.2	23.4	8.1	76.8	26.6	9.5	80.6	29.9	10.7
Colorado	84.4	27.0	9.0	86.9	32.7	11.1	89.3	35.9	12.7
Connecticut	79.2	27.2	11.0	84.0	31.4	13.3	88.6	35.6	15.5
Delaware	77.5	21.4	7.7	82.6	25.0	9.4	87.4	28.7	11.4
District of Columbia	73.1	33.3	17.2	77.8	39.1	21.0	87.1	48.5	28.0
Florida	74.4	18.3	6.3	79.9	22.3	8.1	85.3	25.3	9.0
Georgia	70.9	19.3	6.4	78.6	24.3	8.3	83.9	27.5	9.9
Hawaii	80.1	22.9	7.1	84.6	26.2	8.4	90.4	29.6	9.9
Idaho	79.7	17.7	5.3	84.7	21.7	6.8	88.4	23.9	7.5
Illinois	76.2	21.0	7.5	81.4	26.1	9.5	86.4	30.6	11.7
Indiana	75.6	15.6	6.4	82.1	19.4	7.2	86.6	22.5	8.1
Iowa	80.1	16.9	5.2	86.1	21.2	6.5	90.5	25.1	7.4
Kansas	81.3	21.1	7.0	86.0	25.8	8.7	89.7	29.5	10.2
Kentucky	64.6	13.6	5.5	74.1	17.1	6.9	81.7	21.0	8.5
Louisiana	68.3	16.1	5.6	74.8	18.7	6.5	82.2	21.4	6.9
Maine	78.8	18.8	6.1	85.4	22.9	7.9	90.2	26.9	9.6
Maryland	78.4	26.5	10.9	83.8	31.4	13.4	88.2	35.7	16.0
Massachusetts	80.0	27.2	10.6	84.8	33.2	13.7	89.0	38.2	16.4
Michigan	76.8	17.4	6.4	83.4	21.8	8.1	87.9	24.6	9.4
Minnesota	82.4	21.8	6.3	87.9	27.4	8.3	91.5	31.5	10.3
Mississippi	64.3	14.7	5.1	72.9	16.9	5.8	80.4	19.6	7.1
Missouri	73.9	17.8	6.1	81.3	21.6	7.6	86.8	25.2	9.5
Montana	81.0	19.8	5.7	87.2	24.4	7.2	90.8	27.4	8.3

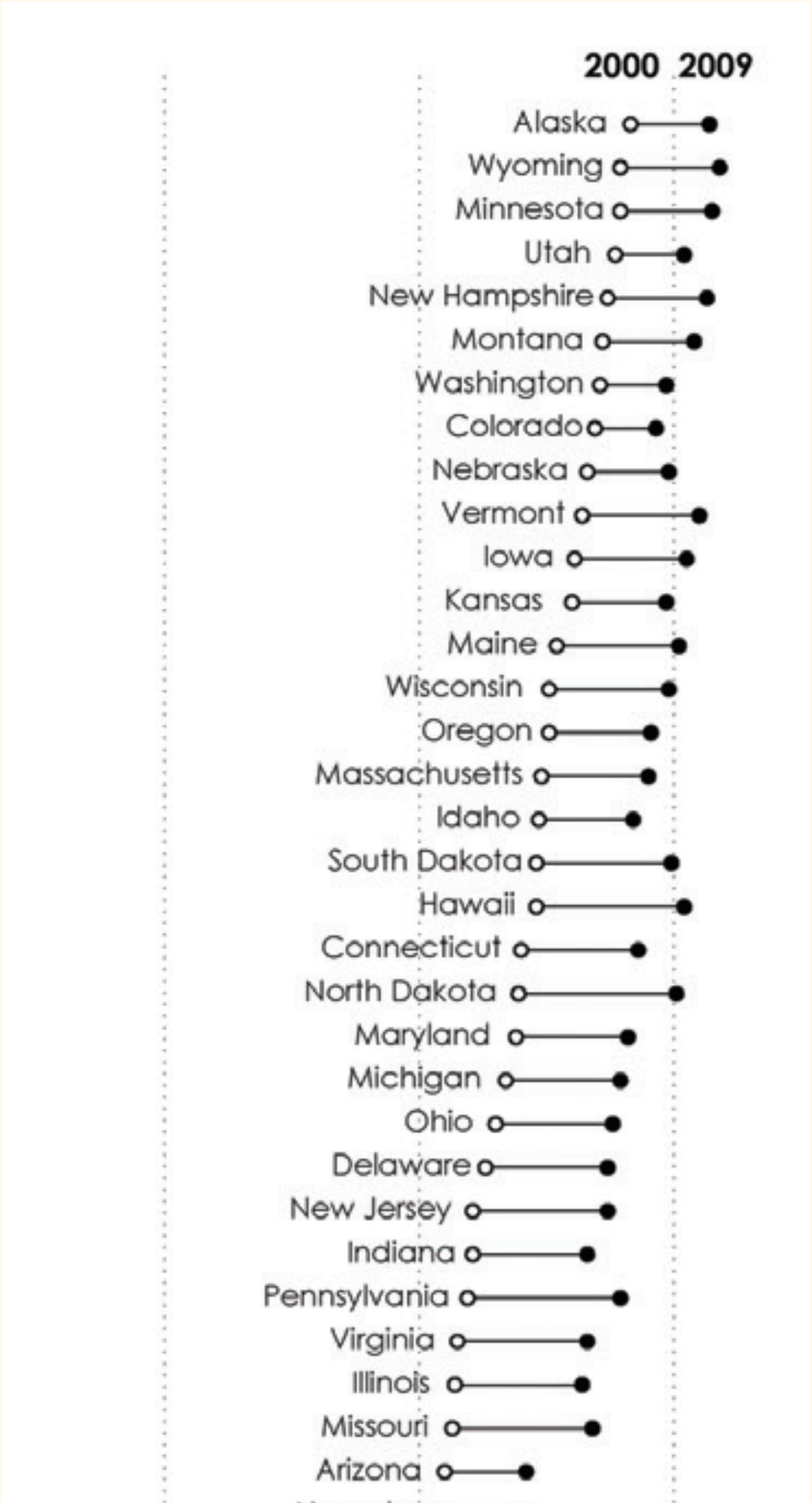
How do states compare to each other?



Taken from Nathan Yau's "Data Points," chapter 3

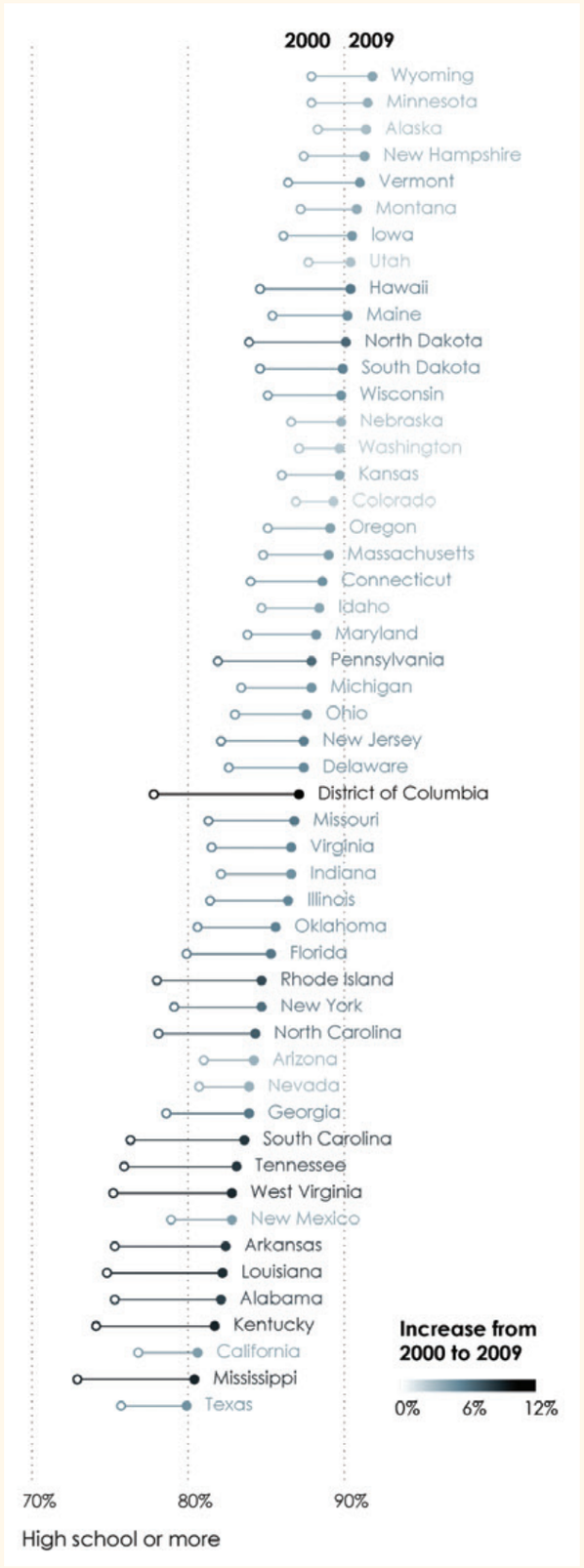
How have states changed over time?

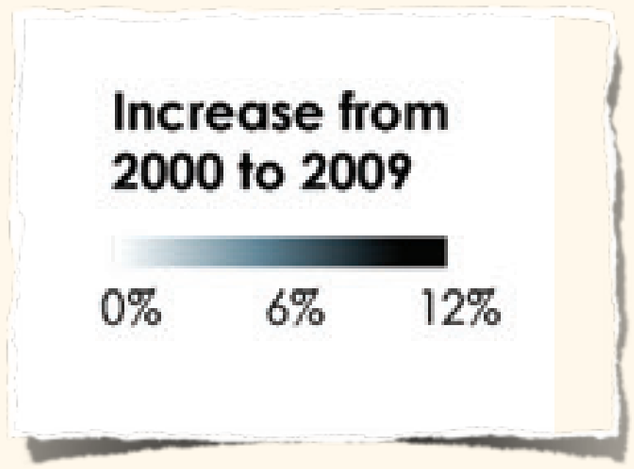


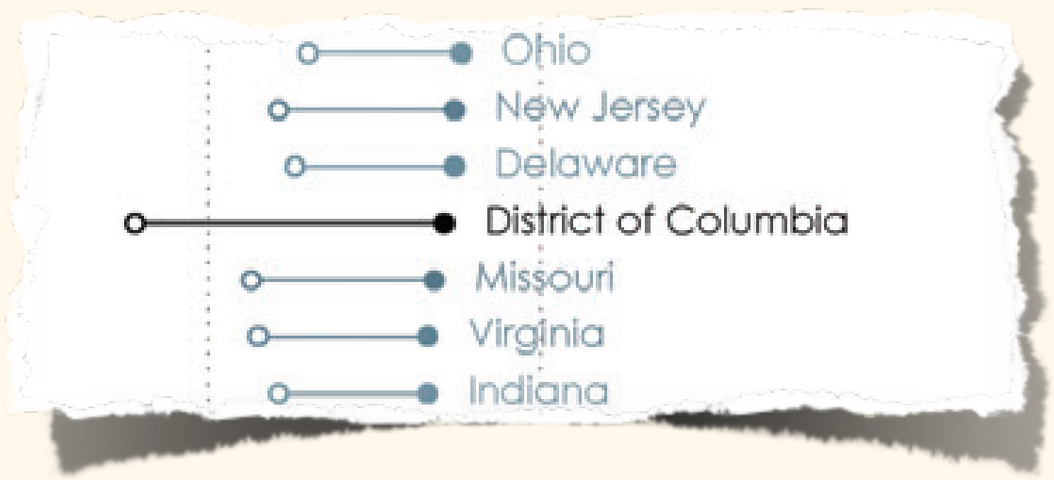


Taken from Nathan Yau's "Data Points," chapter 3

How have states changed over time?

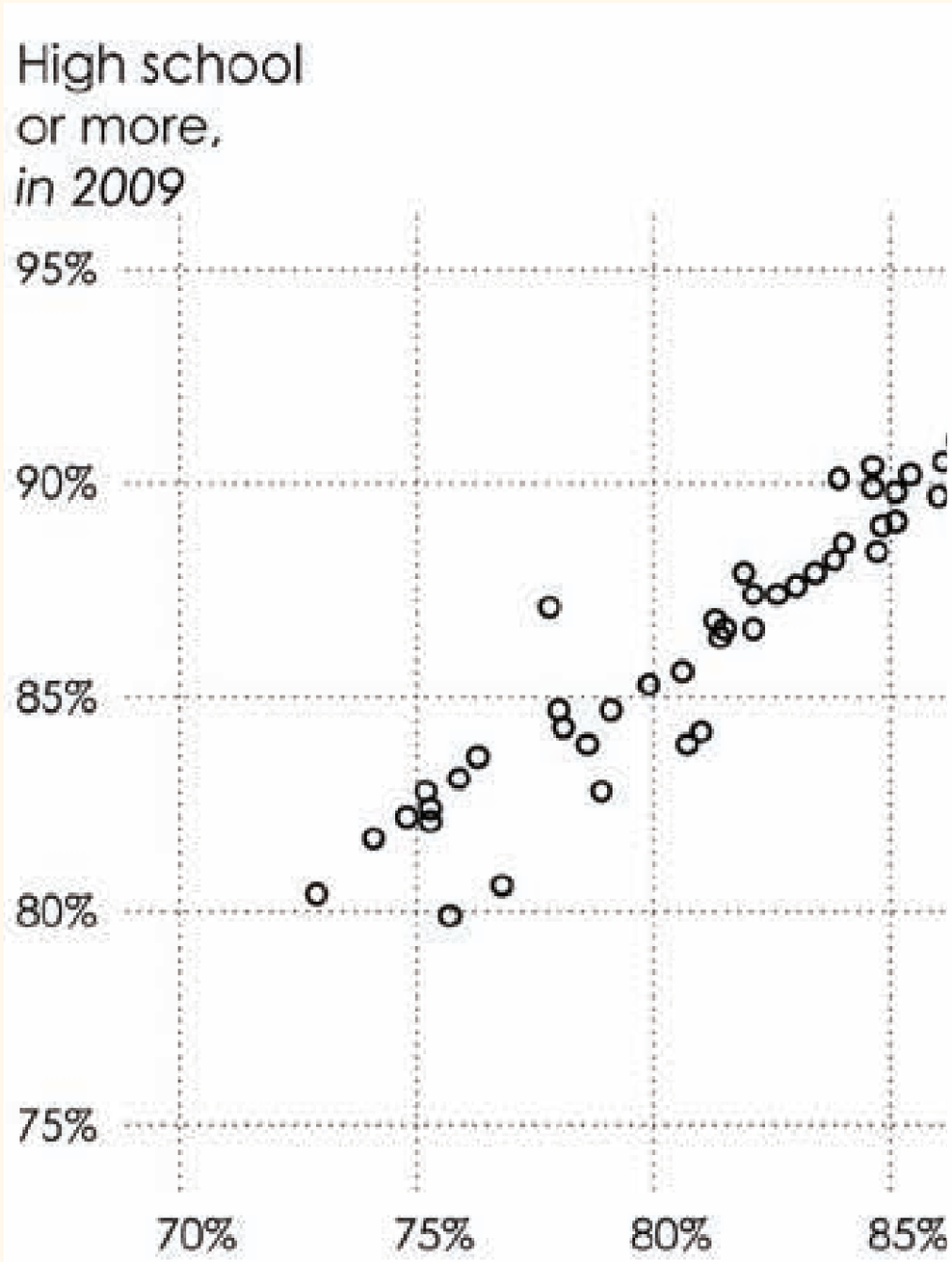


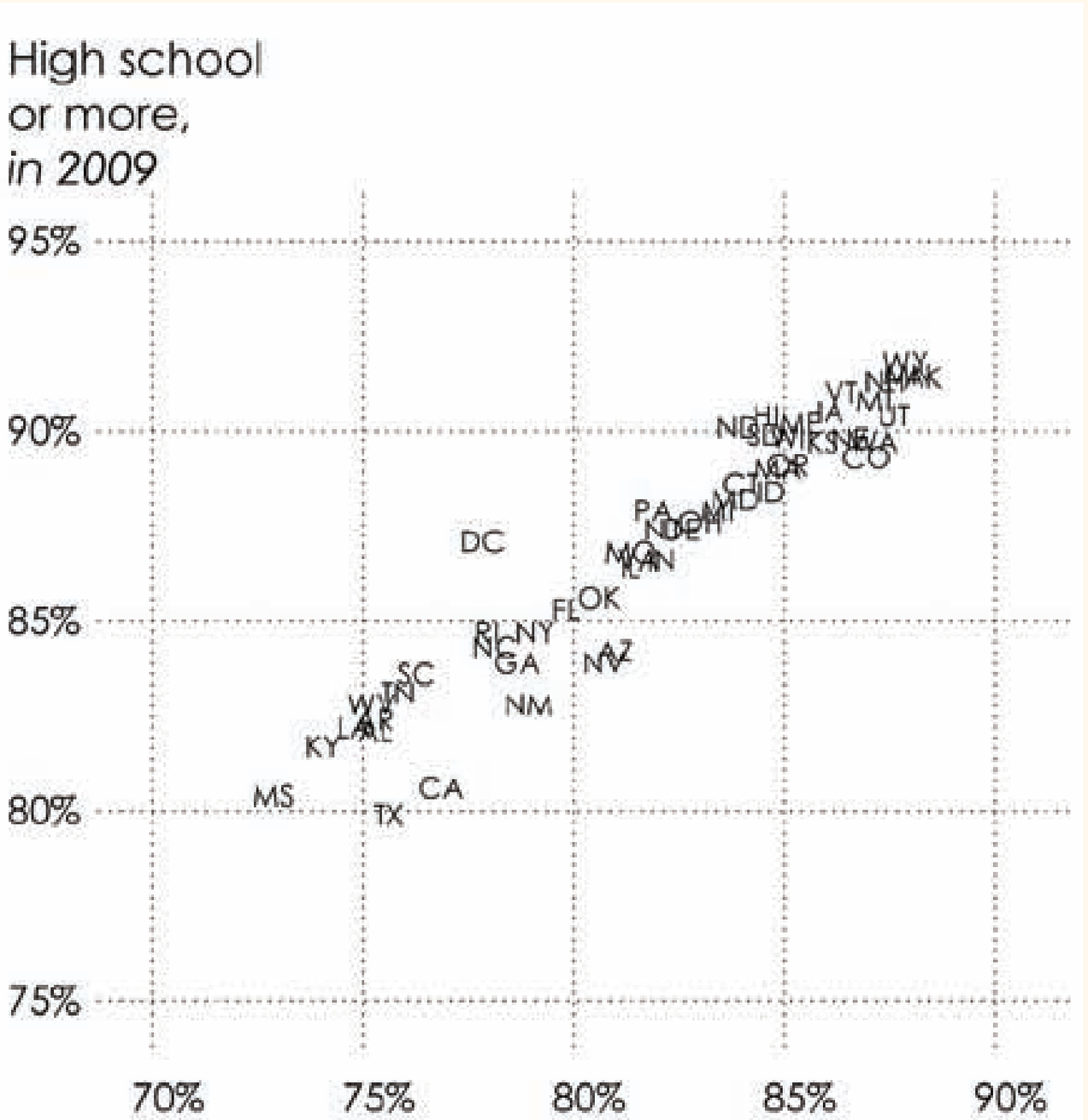




Taken from Nathan Yau's "Data Points," chapter 3

How have states changed over time?





Taken from Nathan Yau's "Data Points," chapter 3

Each of the preceding visualizations answered different questions...

... and used different components of the underlying data.

Neither one, on its own, told the entire story!

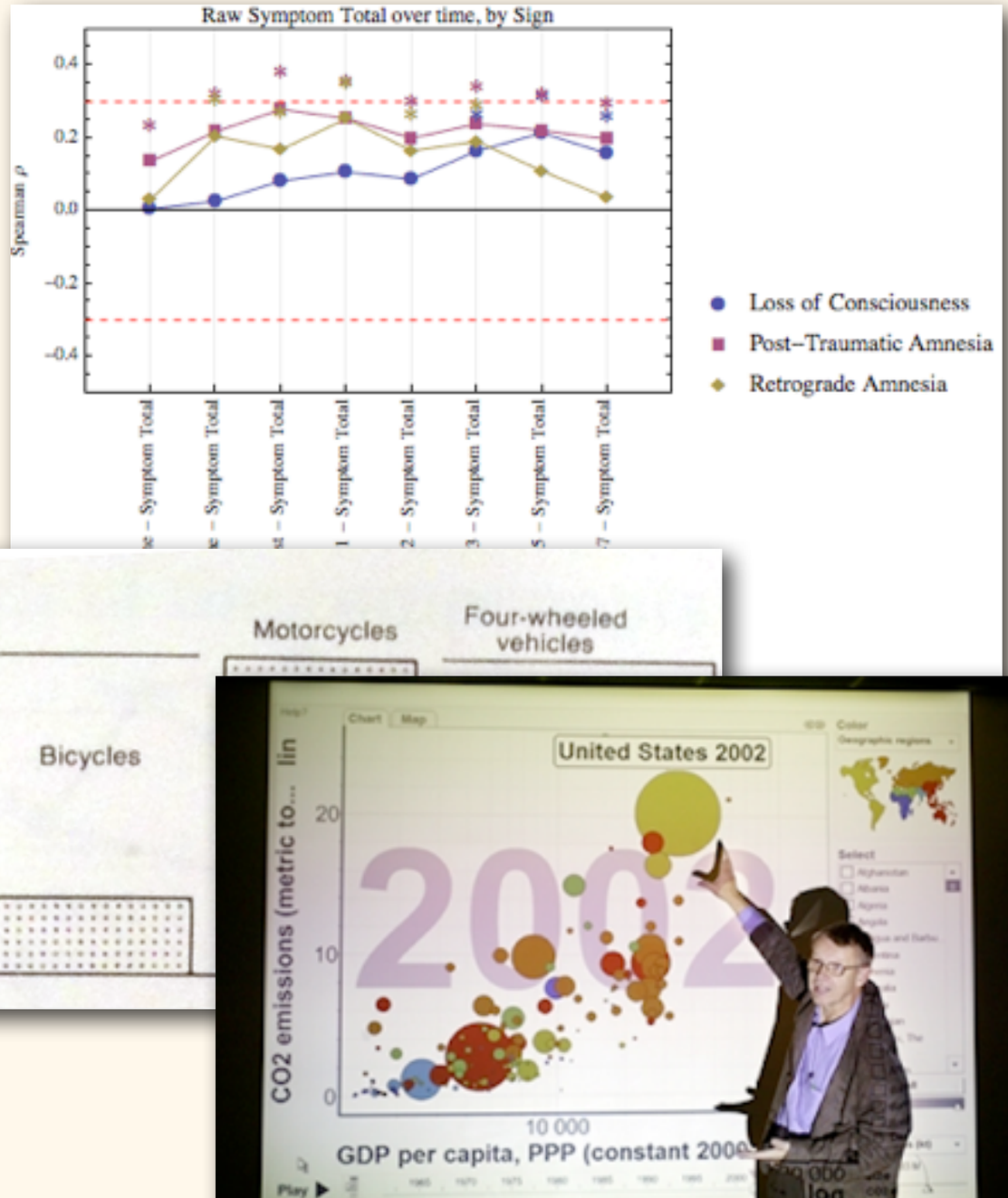
Our game plan for today:

1. Course logistics & overview
2. What are we really trying to do?
3. Let's talk about data...
4. Families of visualization
5. Data-Ink Ratios

There are 4.5 major families of information visualization:

1. Graphs
2. Charts
3. Maps
4. Diagrams
5. “Infographics”

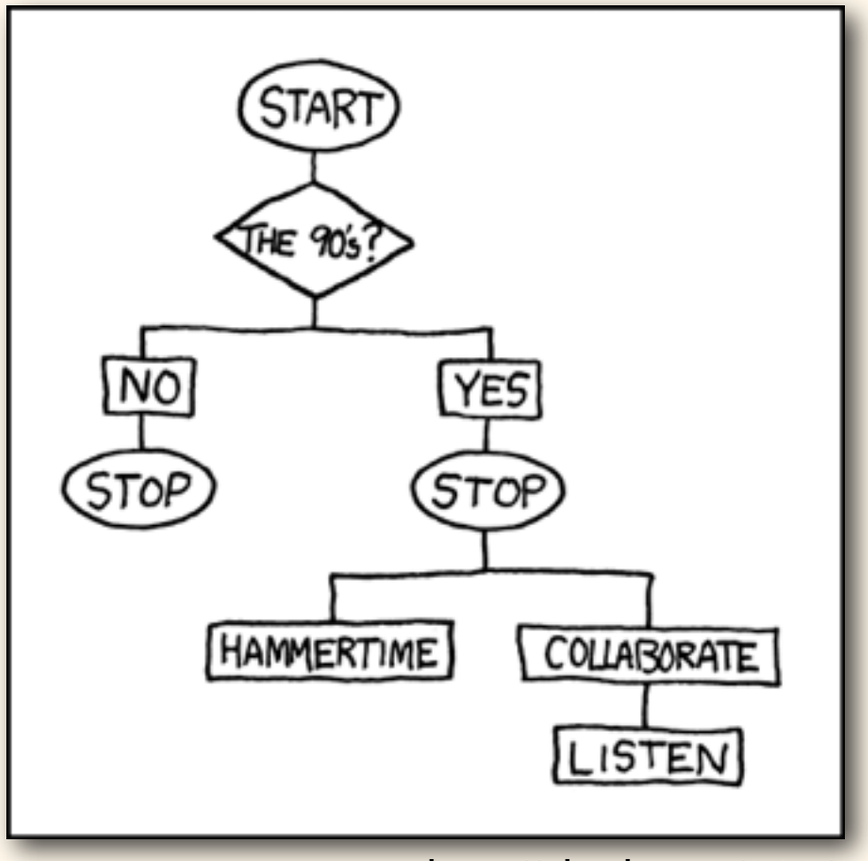
Graphs typically display quantitative information, and include ≥2 scales/axes.



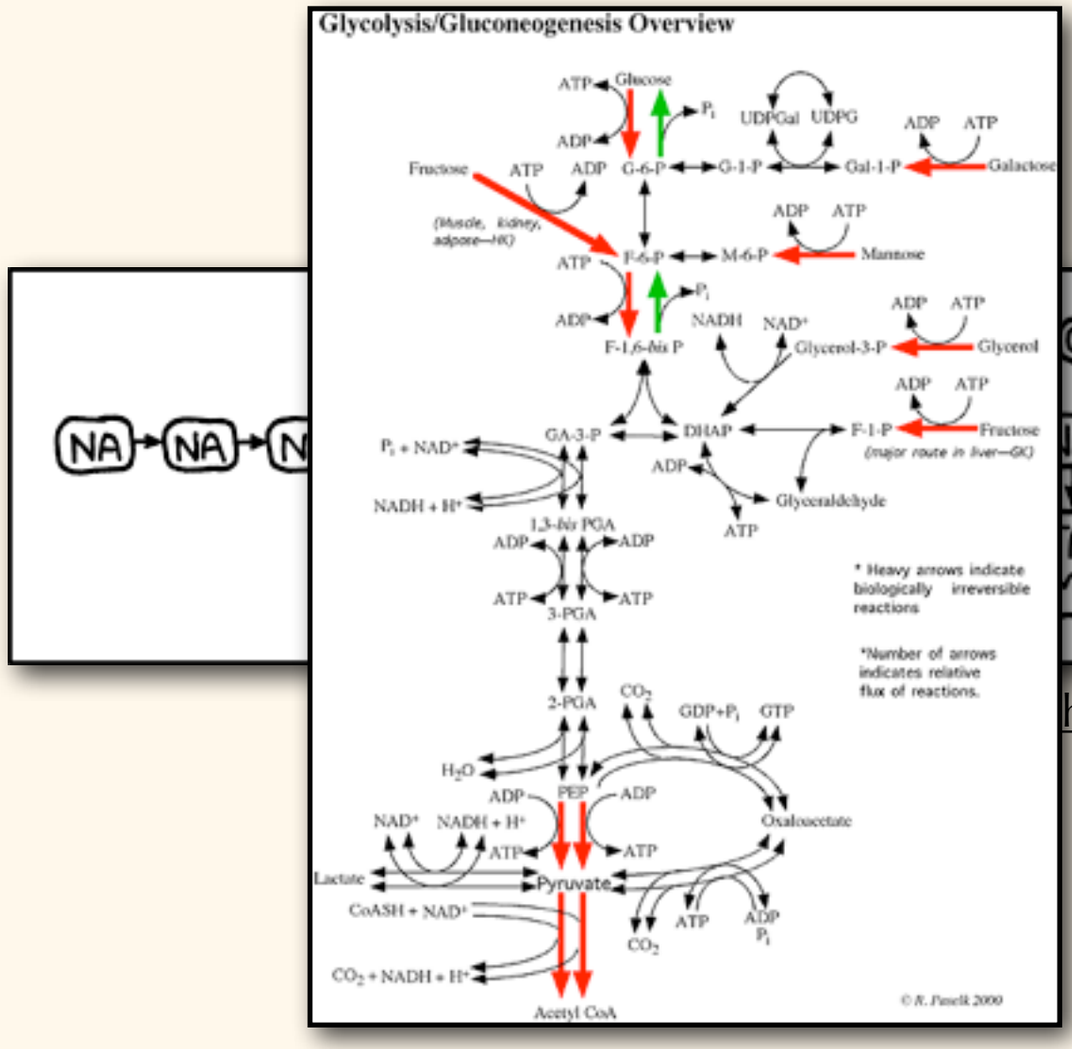
Descriptions based on slides by Marti Hearst

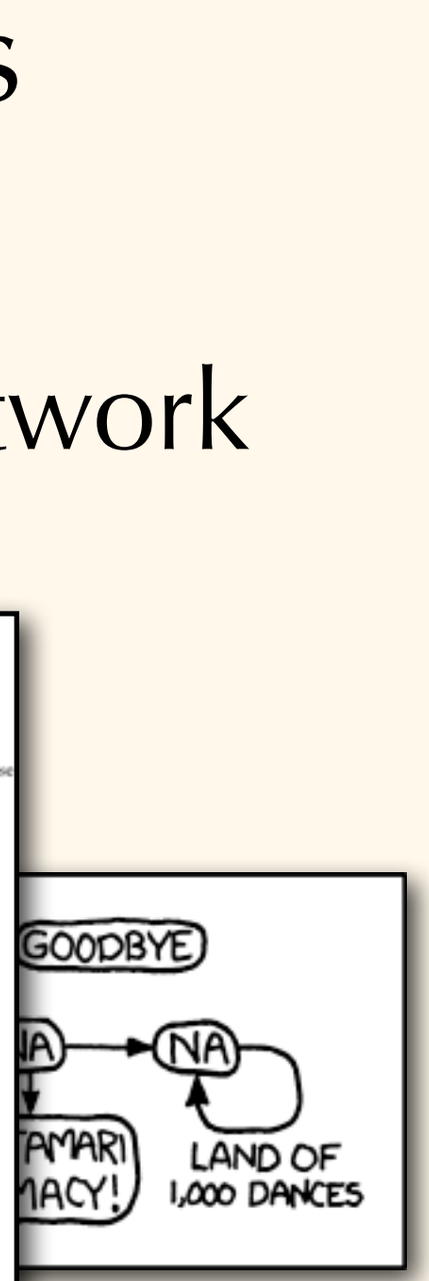
Charts display discrete relationships among discrete entities.

Flowcharts, family trees, (mis-named) network diagrams, etc.



http://xkcd.com/210/

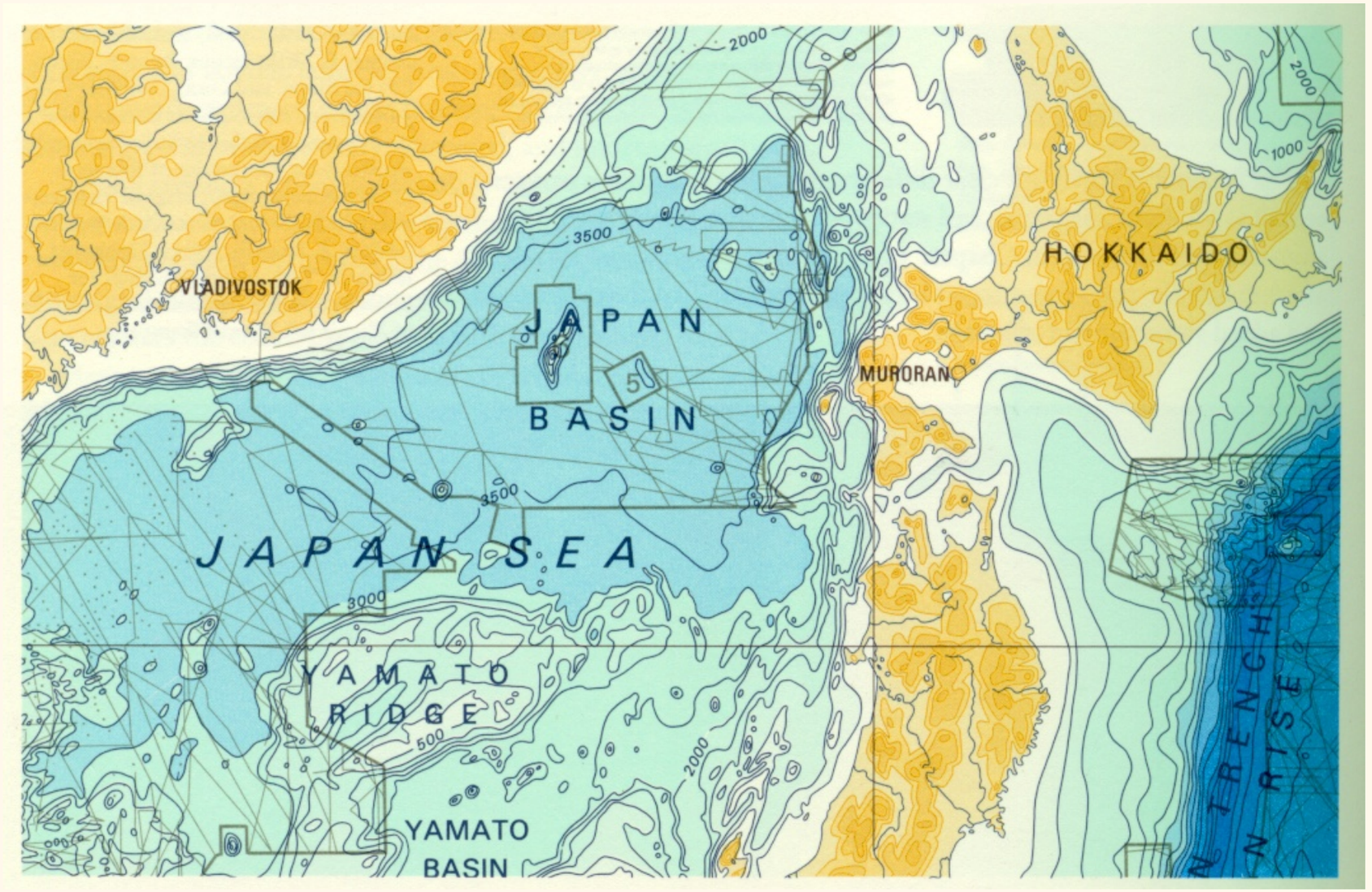




http://xkcd.com/851/

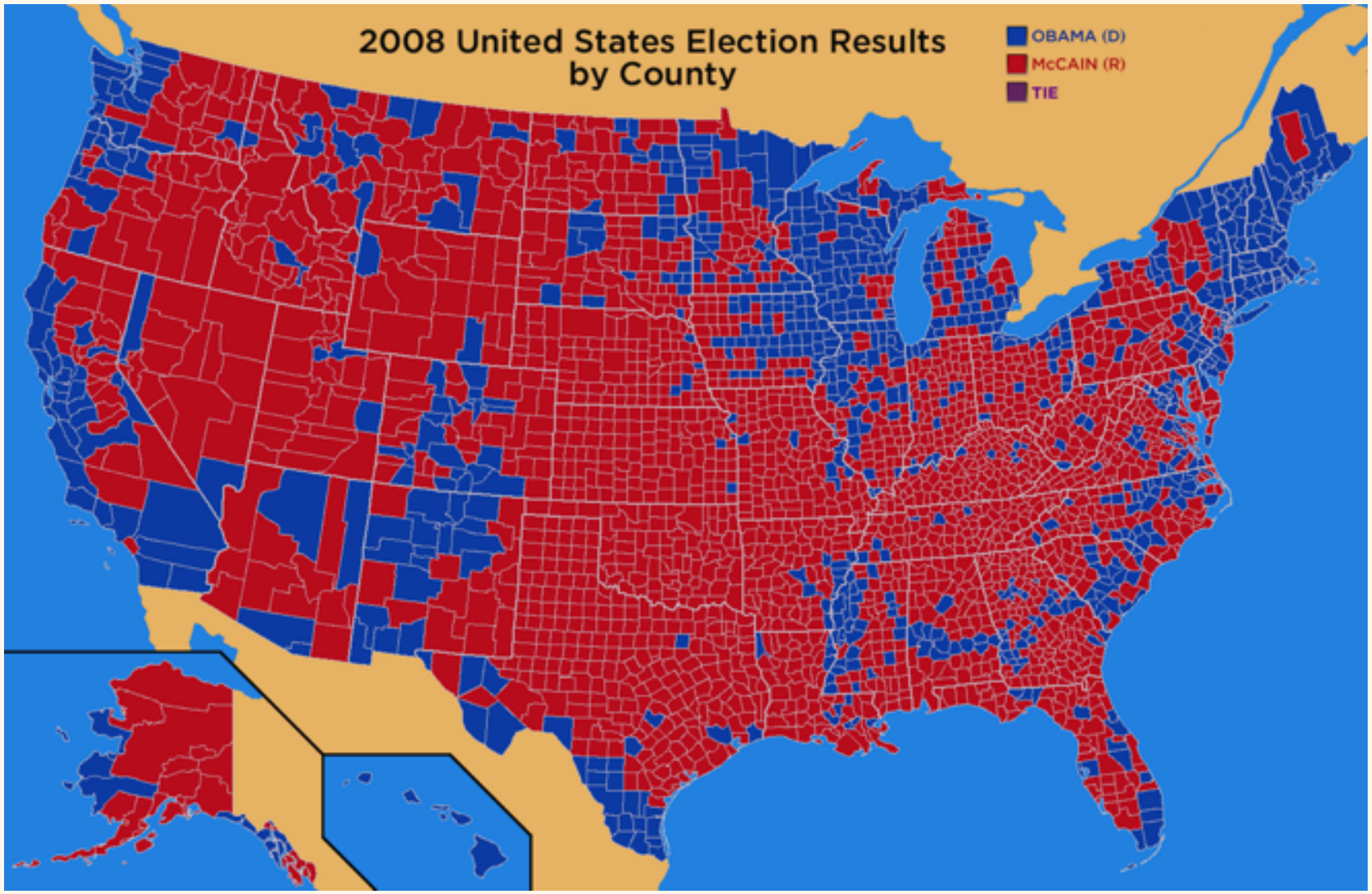
Descriptions based on slides by Marti Hearst http://users.humboldt.edu/rpaselk/C438.S10/C438Notes/C438nLec25_slide.htm

Maps display spatial information, possibly with labels and other information.



Tufte, VE 1997 p76

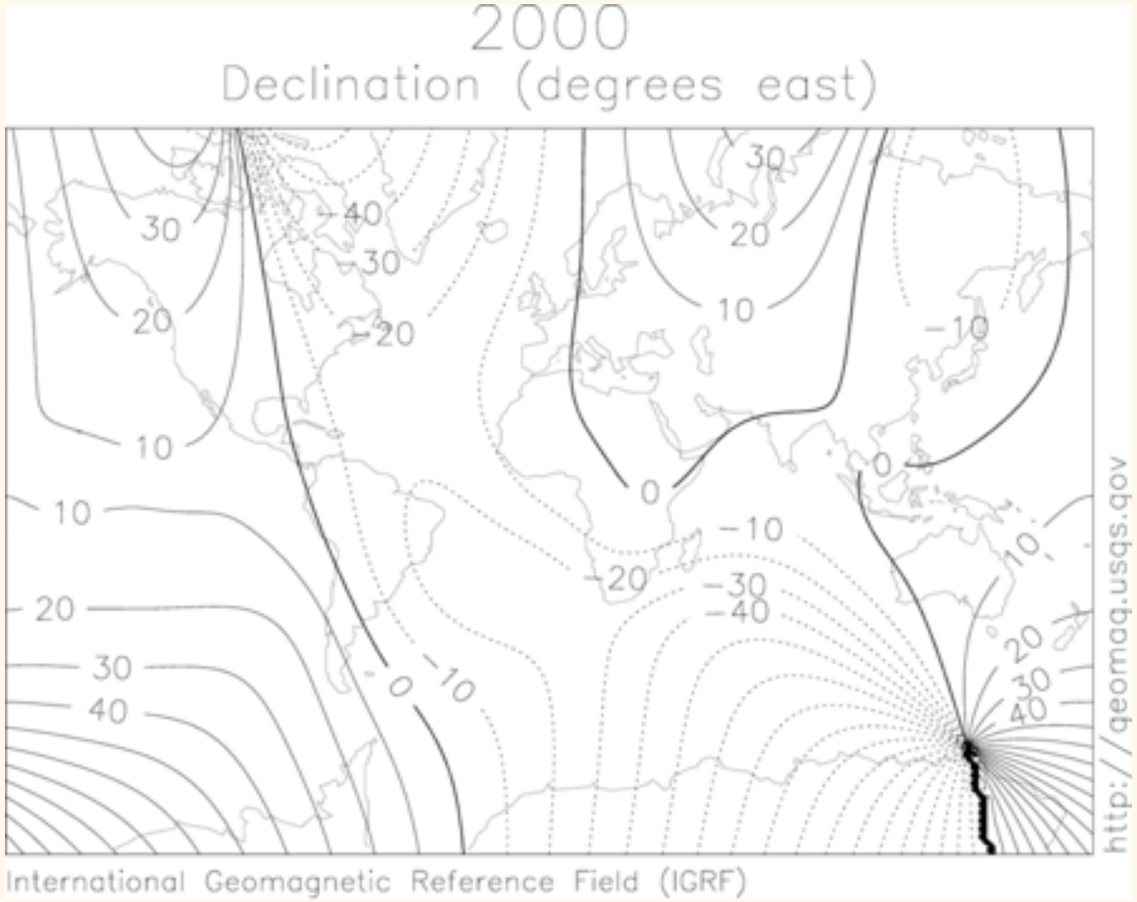
Maps display spatial information, possibly with labels and other information.

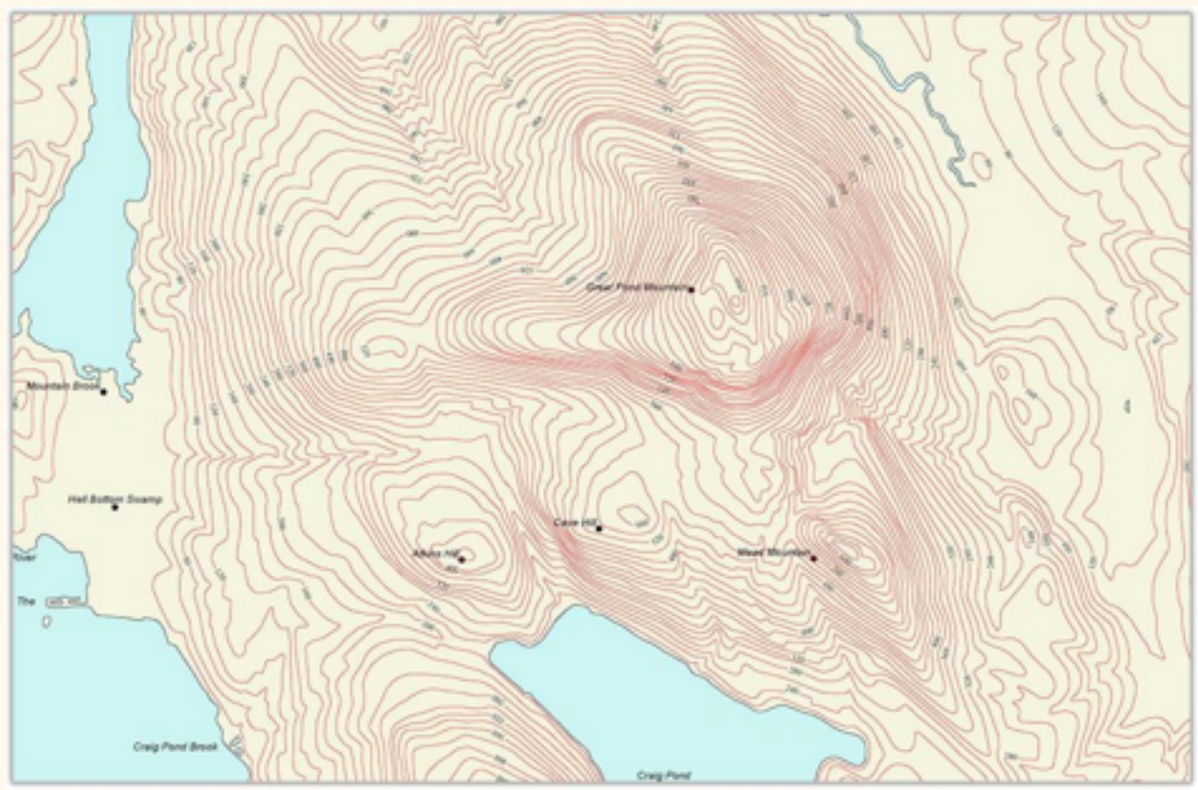


A chloropleth map displays categorical data...

http://upload.wikimedia.org/wikipedia/commons/4/4b/2008_General_Election_Results_by_County.PNG

Maps display spatial information, possibly with labels and other information.

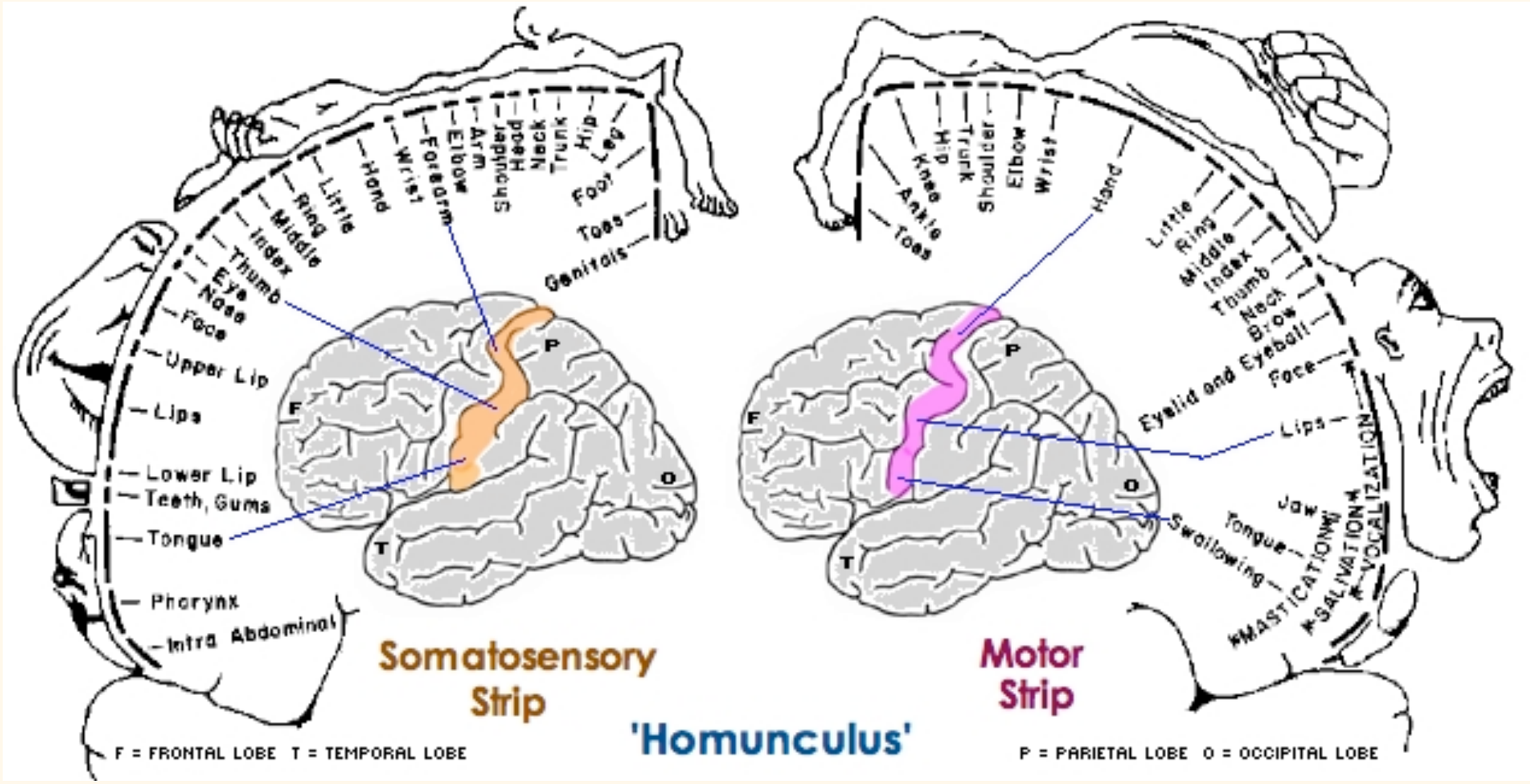




Contour maps show continuous data.

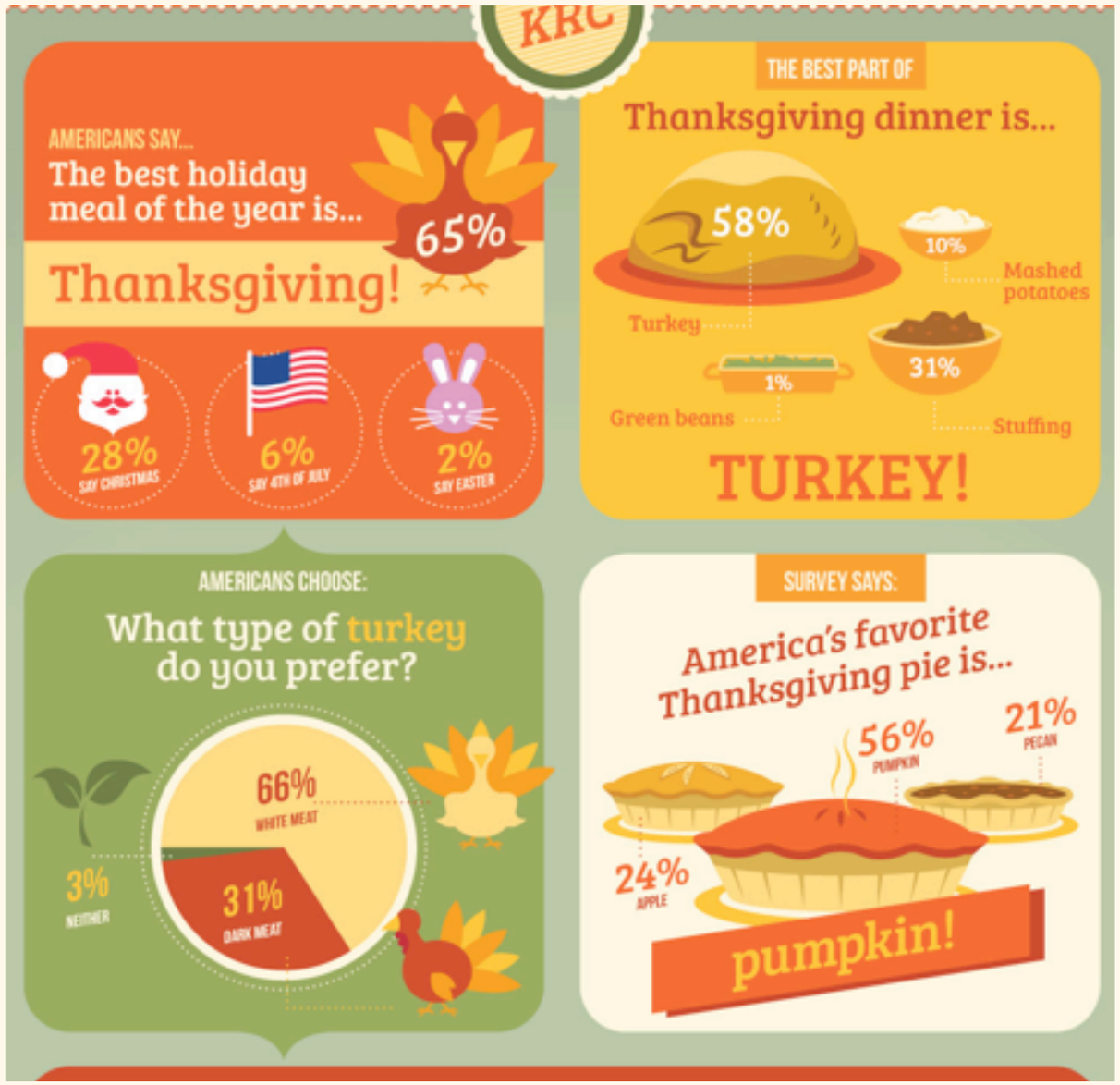
http://en.wikipedia.org/wiki/Contour_line#mediaviewer/File:IGRF_2000_magnetic_declination.gif
http://en.wikipedia.org/wiki/Contour_line#mediaviewer/File:Cntr-map-1.jpg

Diagrams are schematic pictures whose parts are symbolic (i.e., not photographic).



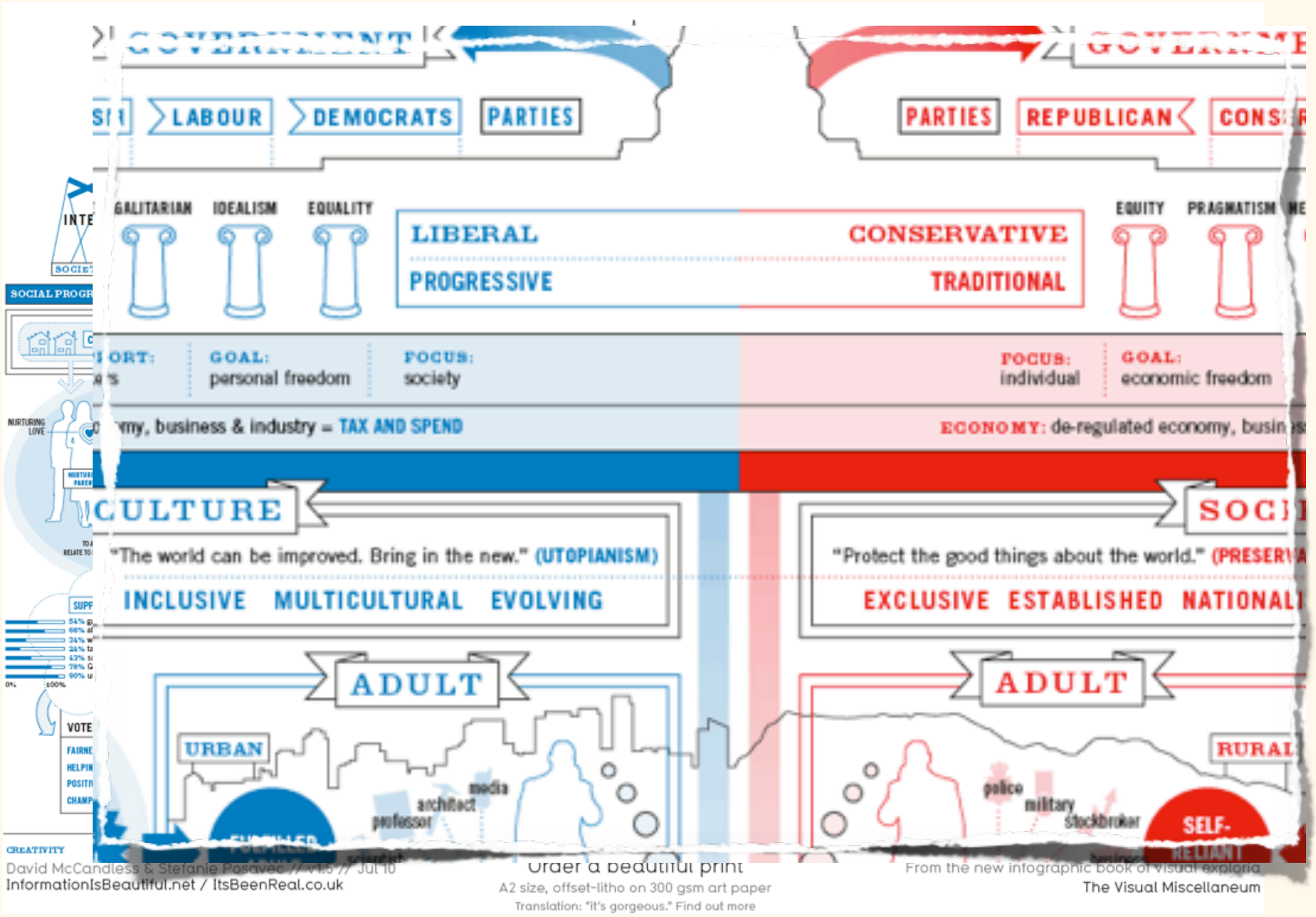
http://web.lemoyne.edu/~hevern/psy340_13F/lectures/psy340.04.3.research.meth.html

"Infographics" are a sort of hybrid of all of the above.

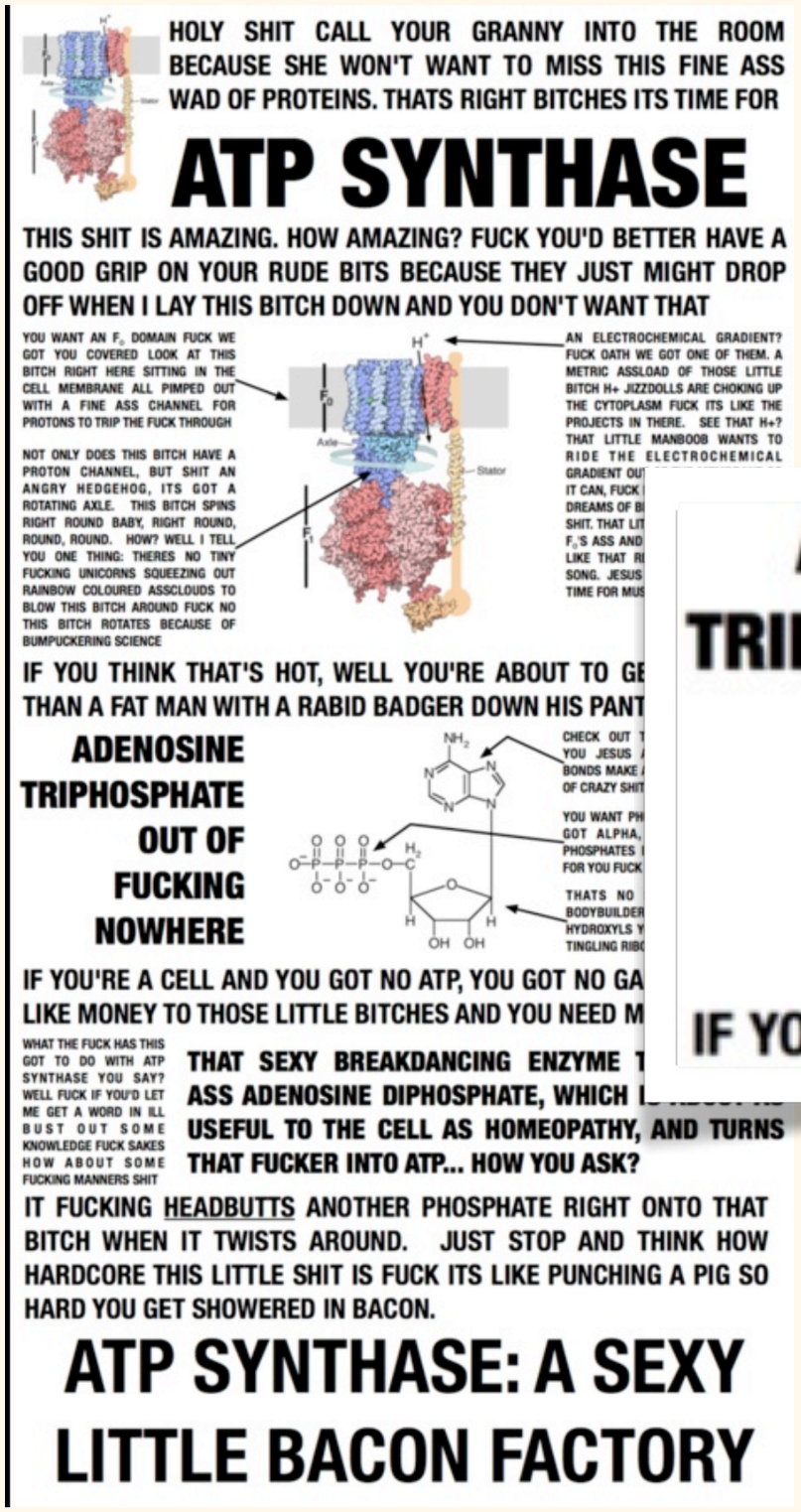


http://dailyinfographic.com/happy-thanksgiving-infographic/ooo
http://dailyinfographic.com/wp-content/uploads/2013/11/top-10-states-highest-computer-job-salaries.jpg

“Infographics” are a sort of hybrid of all of the above.



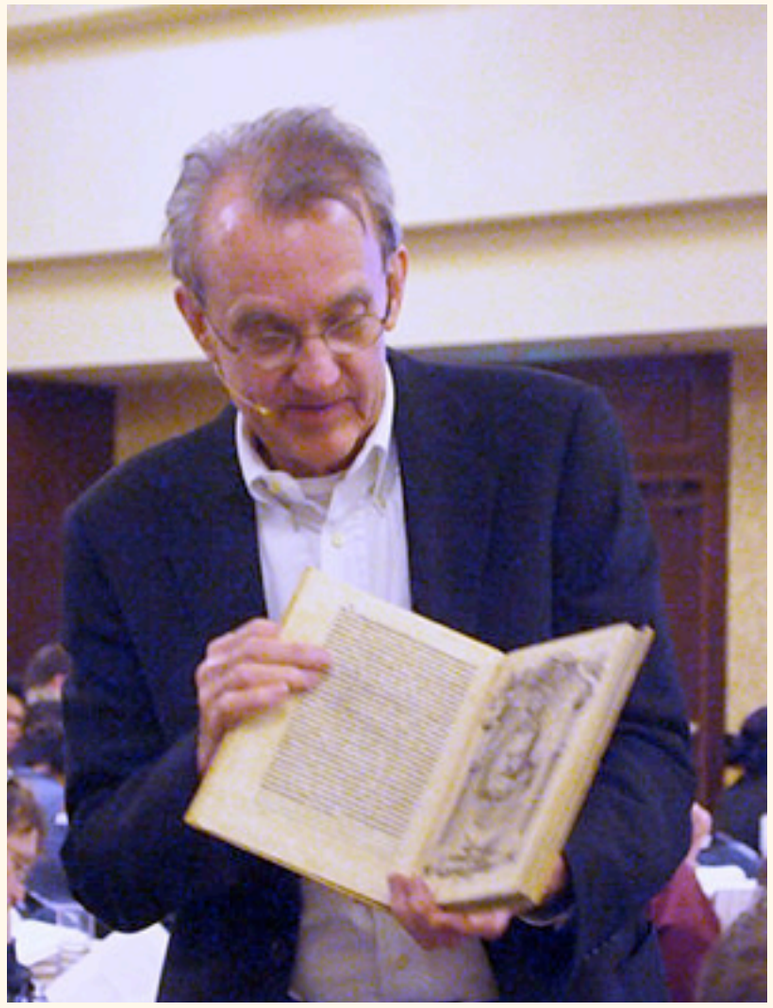
“Infographics” are a sort of hybrid of all of the above.

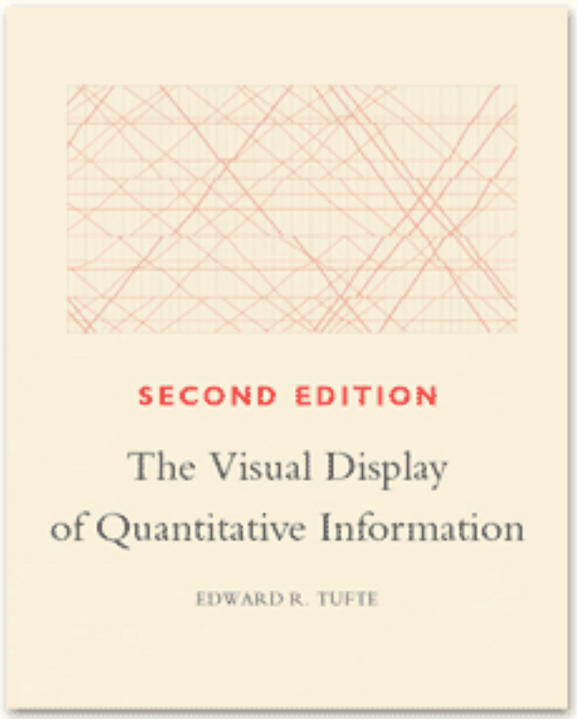


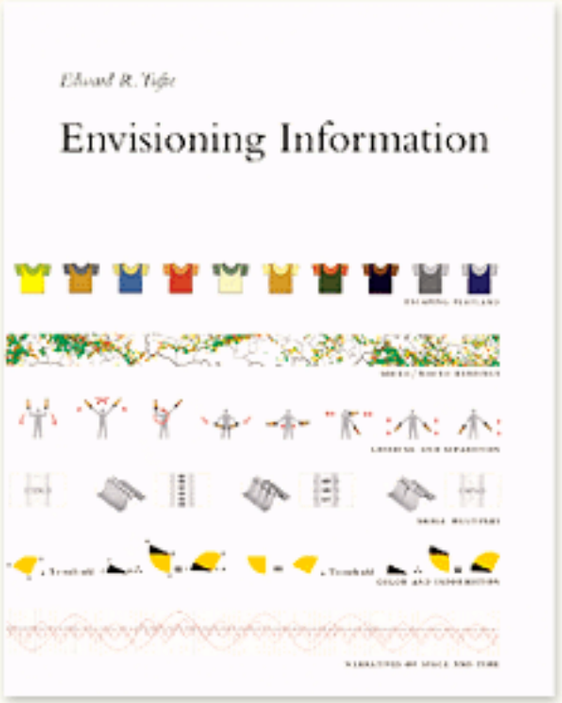
Our game plan for today:

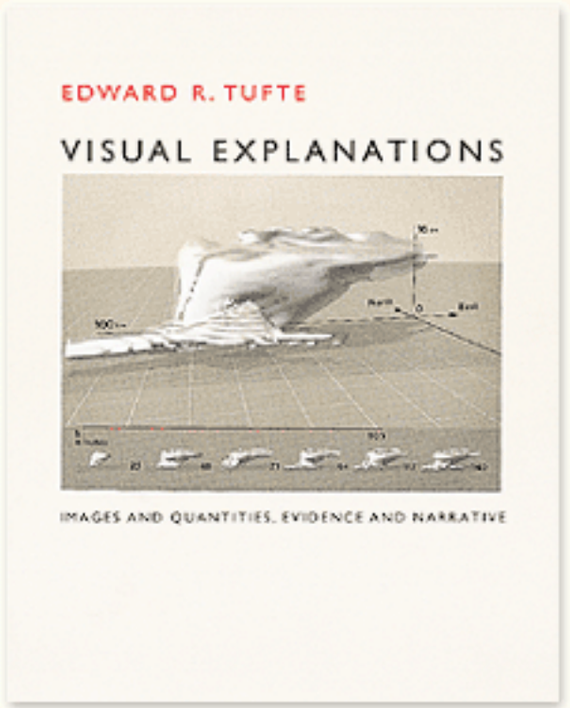
1. Course logistics & overview
2. What are we really trying to do?
3. Let's talk about data...
4. Families of visualization
5. Data-Ink Ratios

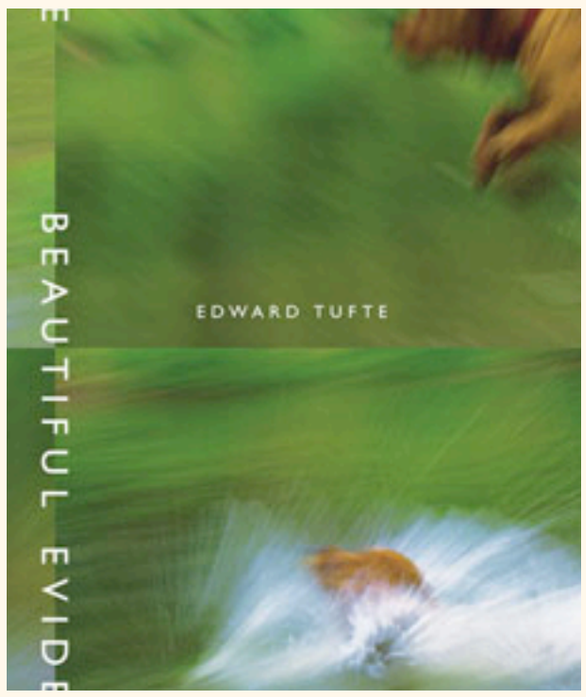
Let’s talk Tufte:











In VDQI, Tufte lays out five characteristics of "Graphical Excellence":

1. "... the well-designed presentation of interesting data- a matter of substance, statistics, and design."
2. Complex ideas communicated with clarity, precision, and efficiency.
3. That which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
4. Nearly always multivariate.
5. Requires telling the truth about the data.

E. Tufte, The Visual Display of Quantitative Information, p. 51


Yeah, well, you know...
That's just , like...your opinion, man

In VDQI, Tufte lays out five characteristics of "Graphical Excellence":

1. "... the well-designed presentation of interesting data- a matter of substance, statistics, and design."
2. Complex ideas communicated with clarity, precision, and efficiency.
3. That which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
4. Nearly always multivariate.
5. Requires telling the truth about the data.

E. Tufte, The Visual Display of Quantitative Information, p. 51

In VDQI, Tufte lays out five characteristics of "Graphical Excellence":

1. "... the well-designed presentation of interesting data- a matter of substance, statistics, and design."
2. Complex ideas communicated with clarity, precision, and efficiency.
3. That which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
4. Nearly always multivariate.
5. Requires telling the truth about the data.

E. Tufte, The Visual Display of Quantitative Information, p. 51

Maximizing the data-to-ink ratio:

$$\text{d-i ratio} = \frac{\text{data ink}}{\text{total ink}}$$

The d-i ratio captures the proportion of the image spent on non-redundant display of information.

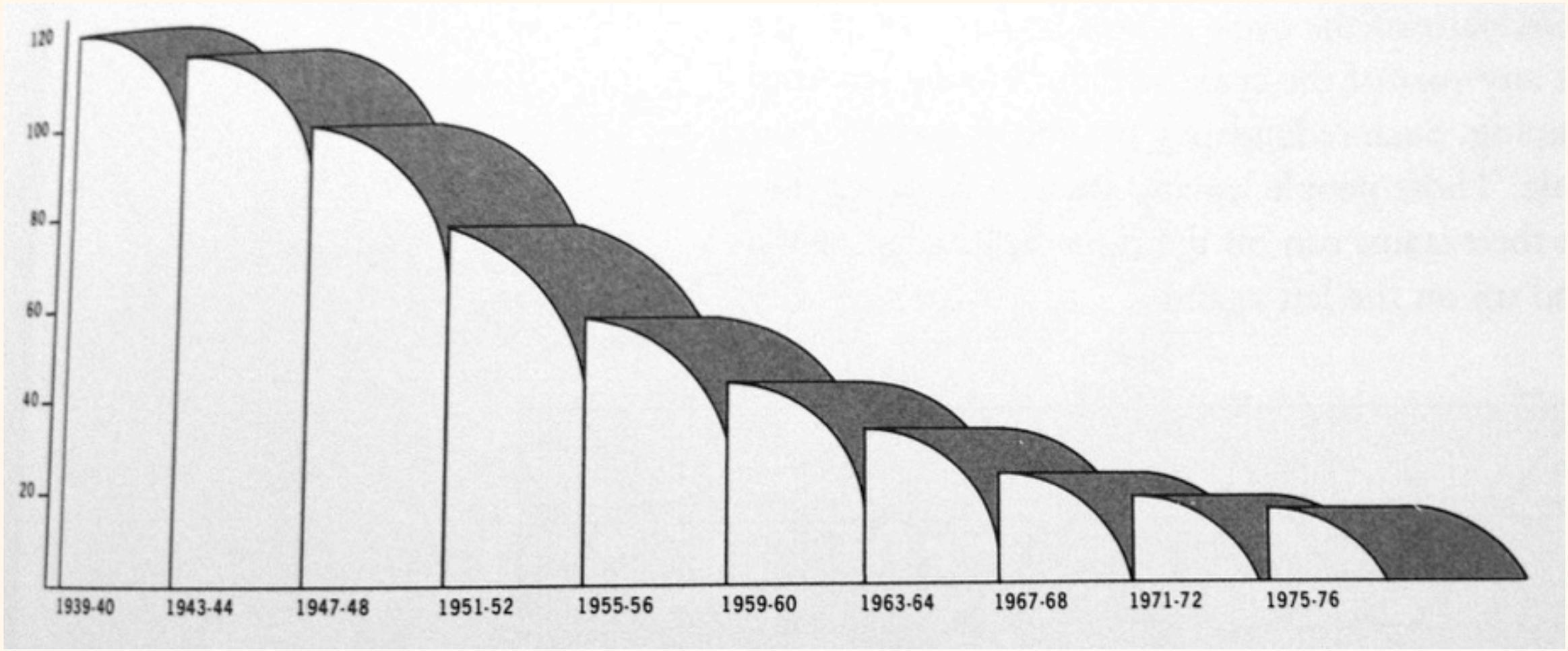
The intuitive explanation:

Everything on the page has to go through the visual processing system...

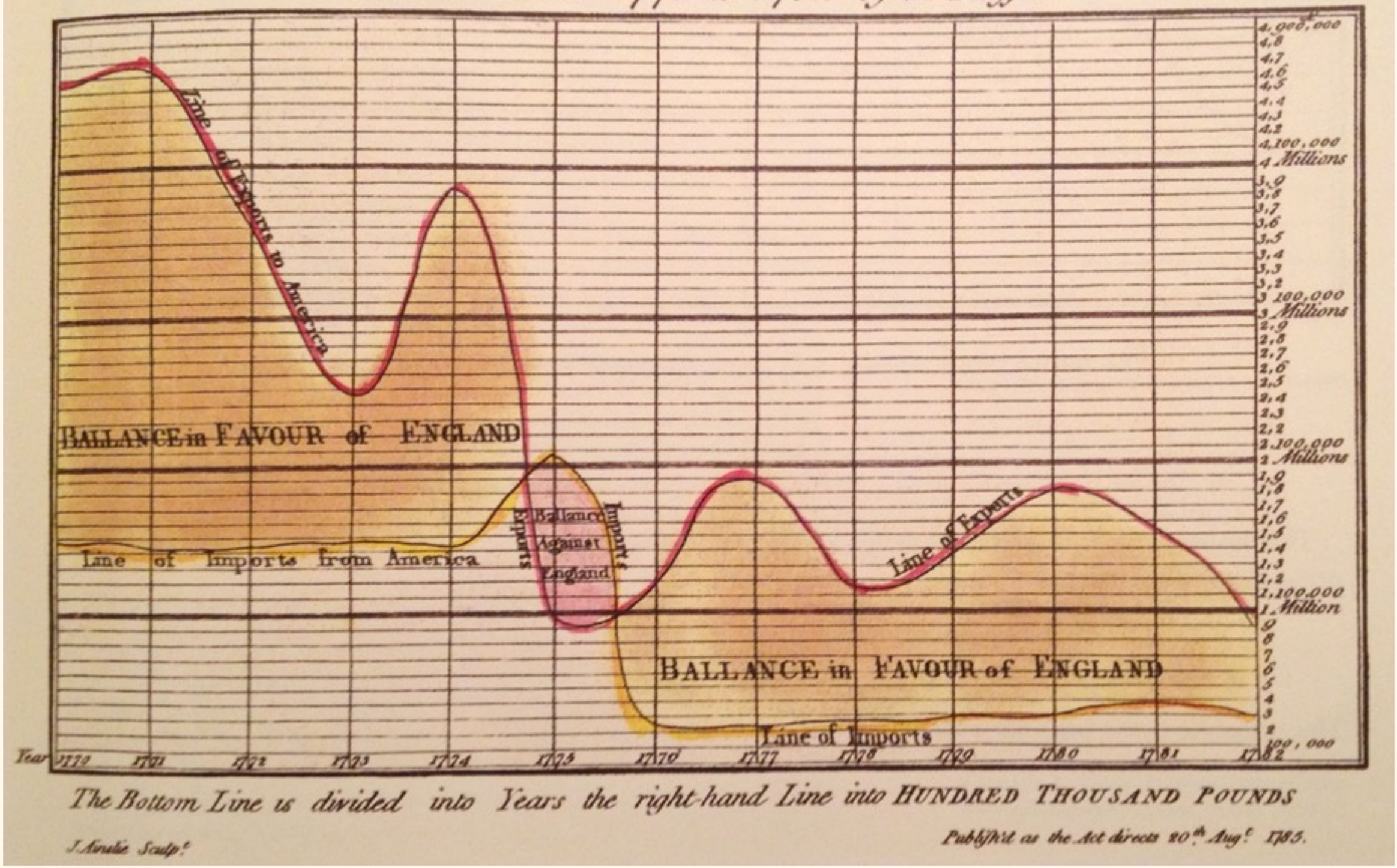
... so we want as much of that processing time as possible being spent on thinking about the data-

- *not* trying to sort out which part of the graph is the data and which is the frame.

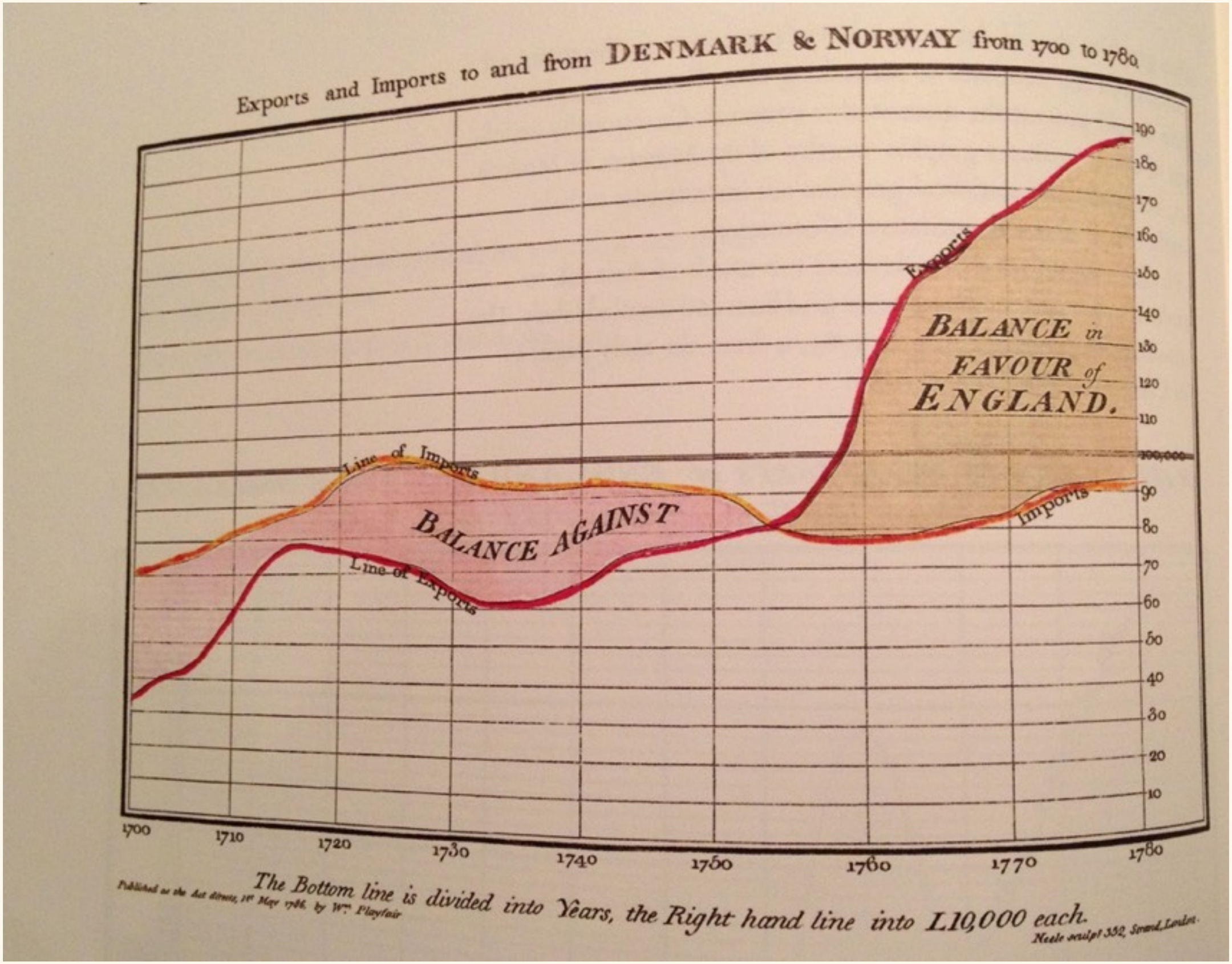
Adding unnecessary content makes it harder to interpret our data.



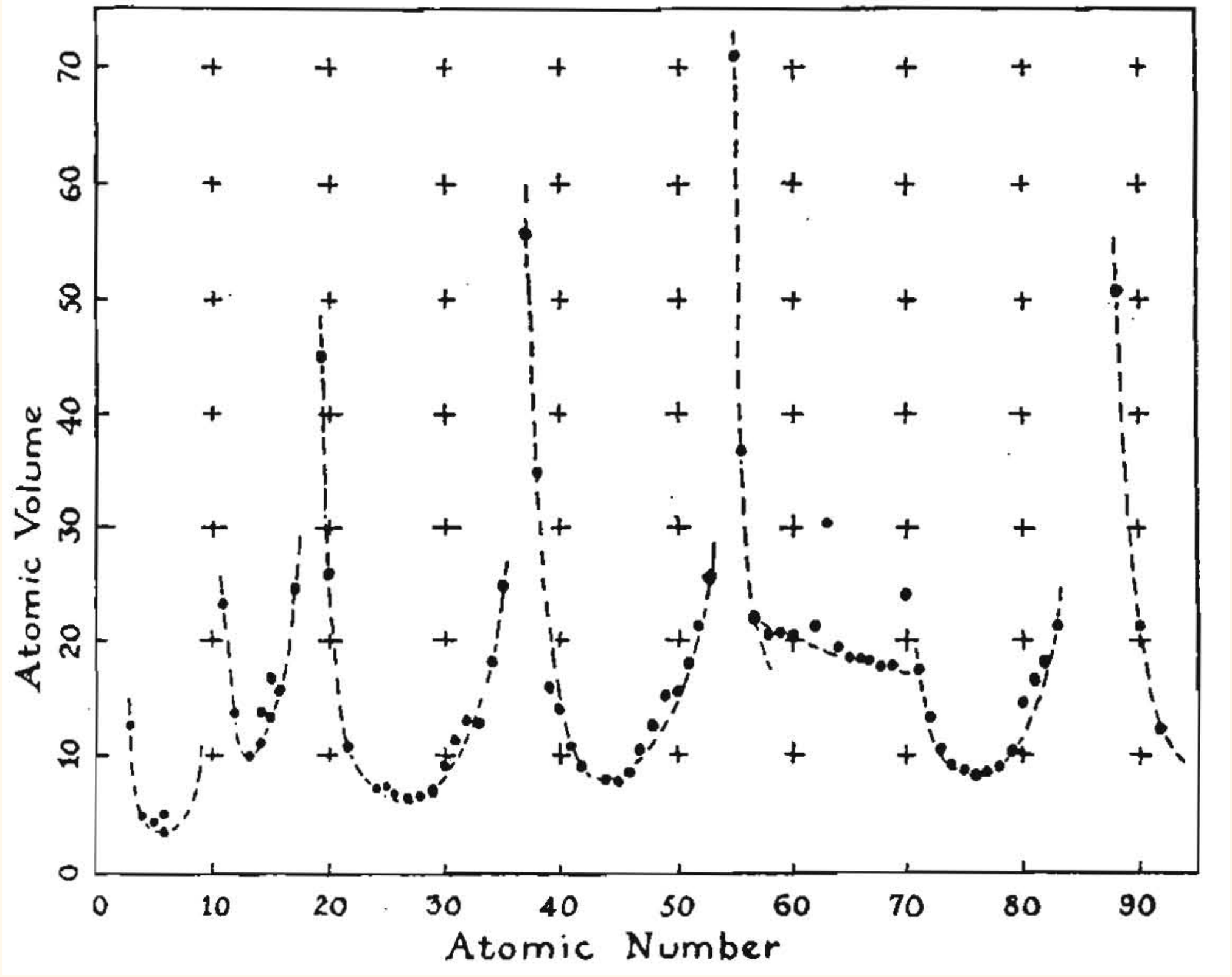
E. Tufte, The Visual Display of Quantitative Information



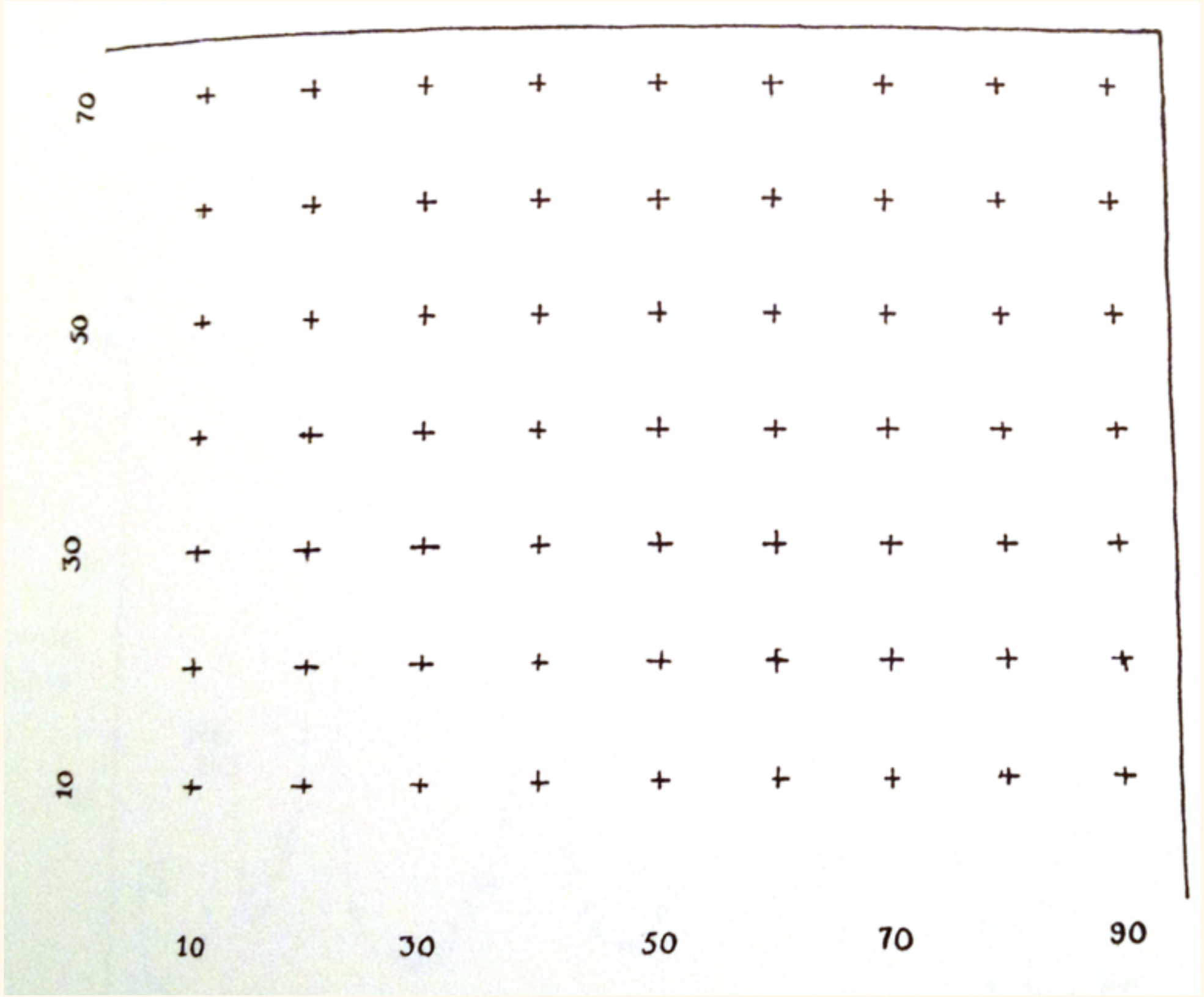
E. Tufte, The Visual Display of Quantitative Information



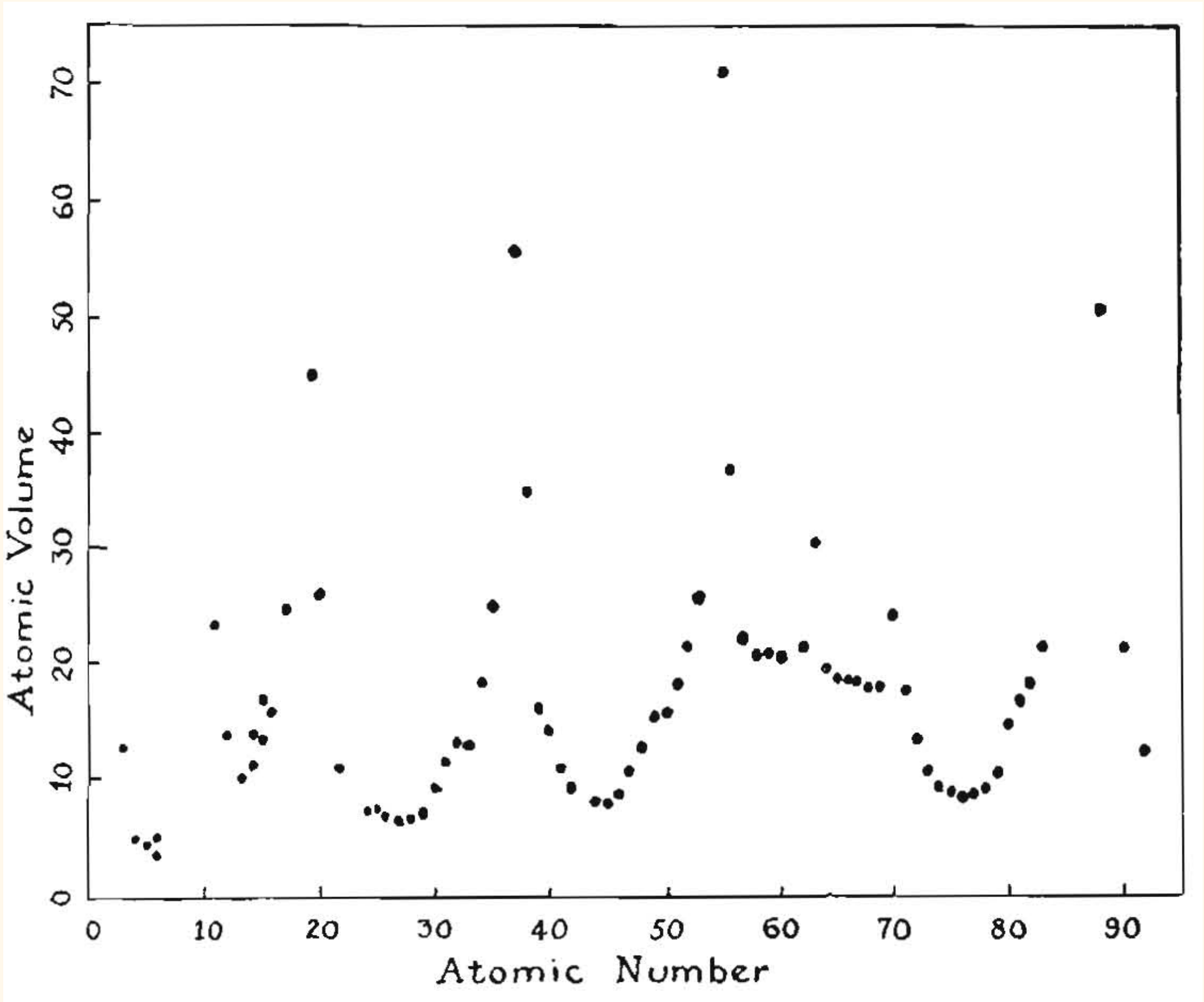
E. Tufte, The Visual Display of Quantitative Information



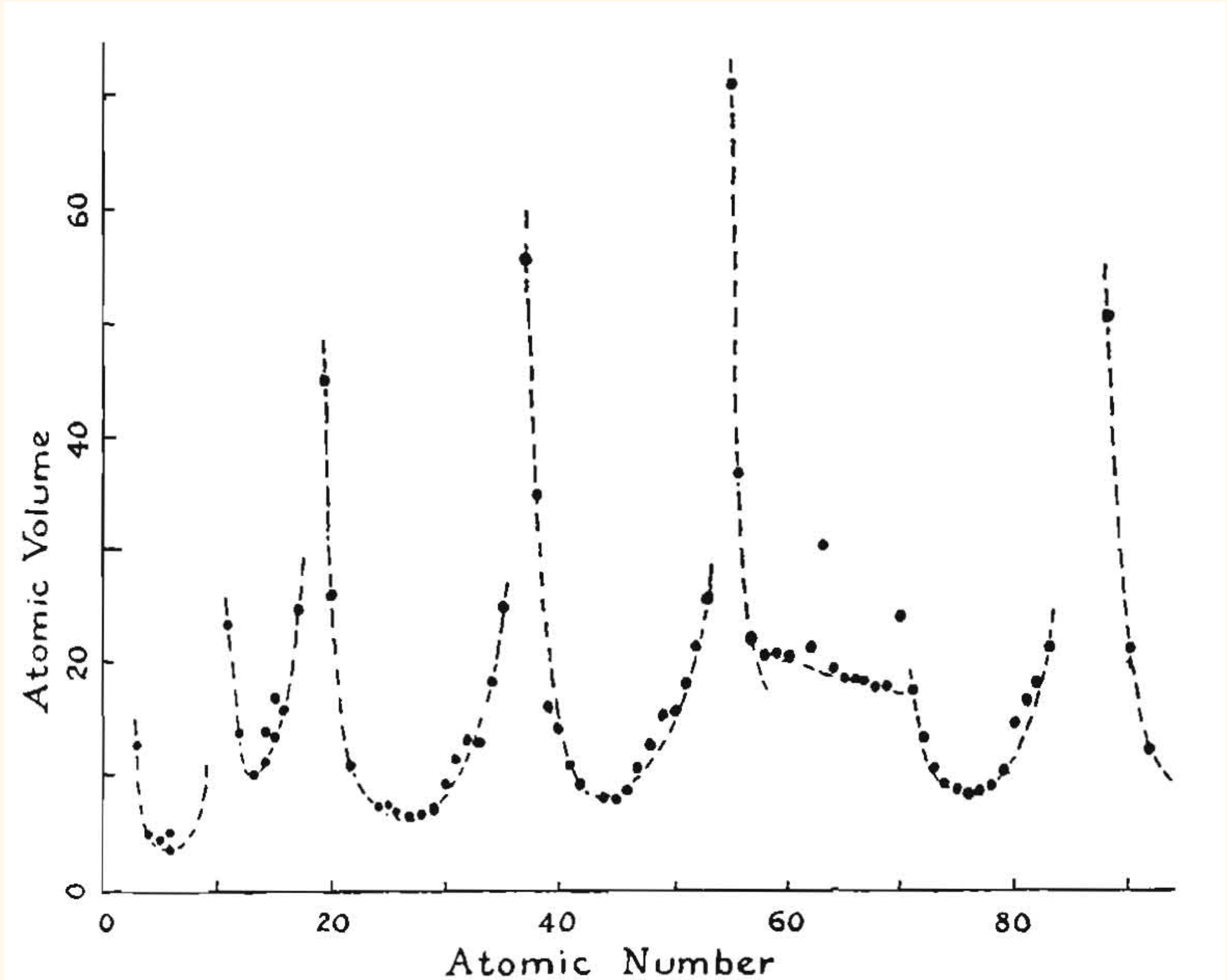
Reproduced from Tufte, The *Visual Display of Quantitative Information*, using a graphic modified from an original by Roger Hayward, published in Pauling's *General Chemistry* (1947). Oy.



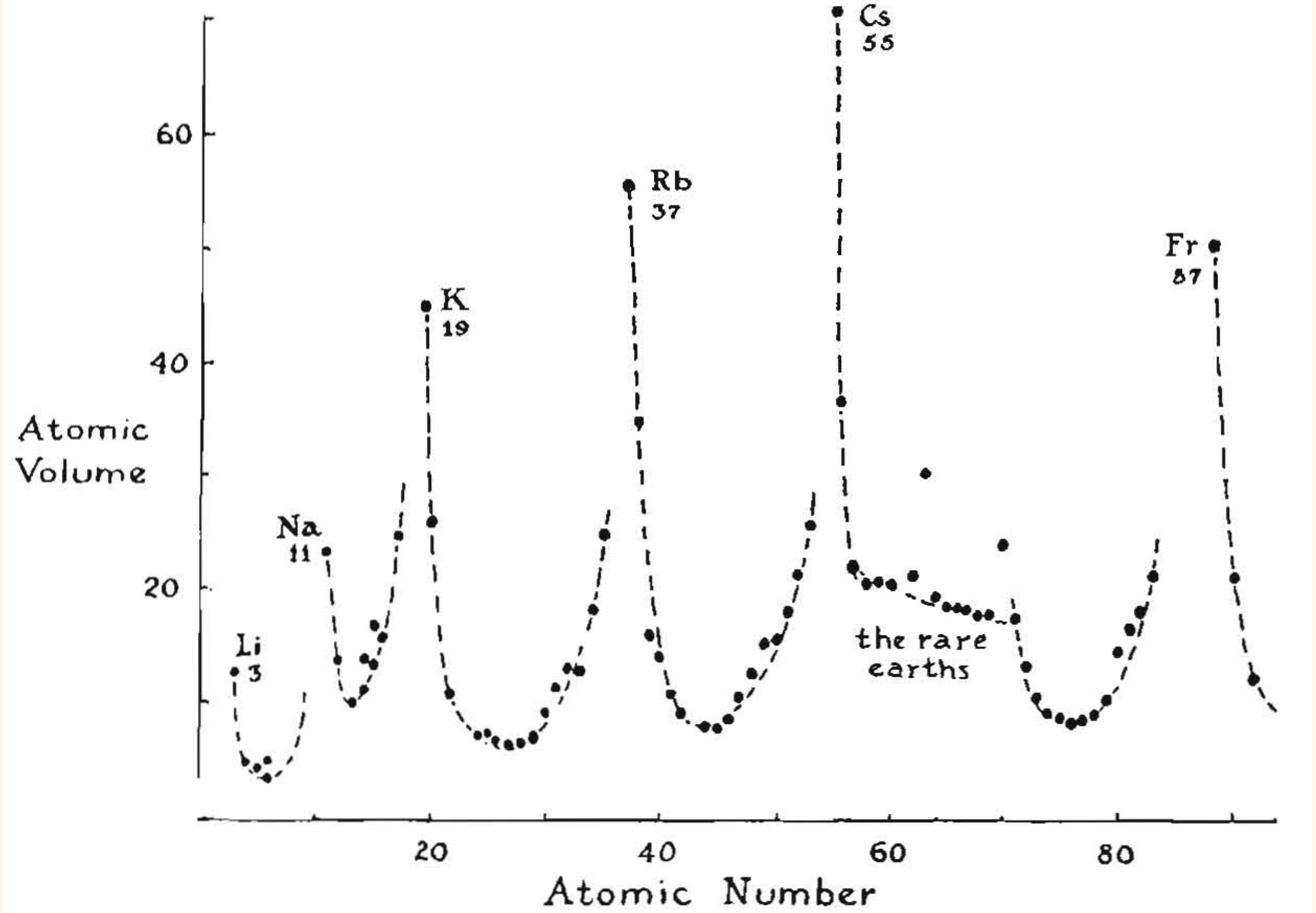
Reproduced from Tufte, The *Visual Display of Quantitative Information*, using a graphic modified from an original by Roger Hayward, published in Pauling's *General Chemistry* (1947). Oy.



Reproduced from Tufte, The *Visual Display of Quantitative Information*, using a graphic modified from an original by Roger Hayward, published in Pauling's *General Chemistry* (1947). Oy.



Reproduced from Tufte, The *Visual Display of Quantitative Information*, using a graphic modified from an original by Roger Hayward, published in Pauling's *General Chemistry* (1947). Oy.



Reproduced from Tufte, The *Visual Display of Quantitative Information*, using a graphic modified from an original by Roger Hayward, published in Pauling's *General Chemistry* (1947). Oy.

That's it for today...

Your homework for next time:

Find two examples of a visualization:

One "good", one "bad"...

... within the context of your field!

Also: Install LaTeX as prep for Thursday's lab!