

A (very brief) introduction to data visualization



Stanford MEDICINE | Epidemiology and Population Health

Mathew Kiang, ScD

✉ mkiang@stanford.edu | [@mathewkiang](https://twitter.com/mathewkiang)



Agenda and Objectives

1. First half: Theory of visualizations
2. 15-minute break
3. Second half: Practice of visualizations (interactive)

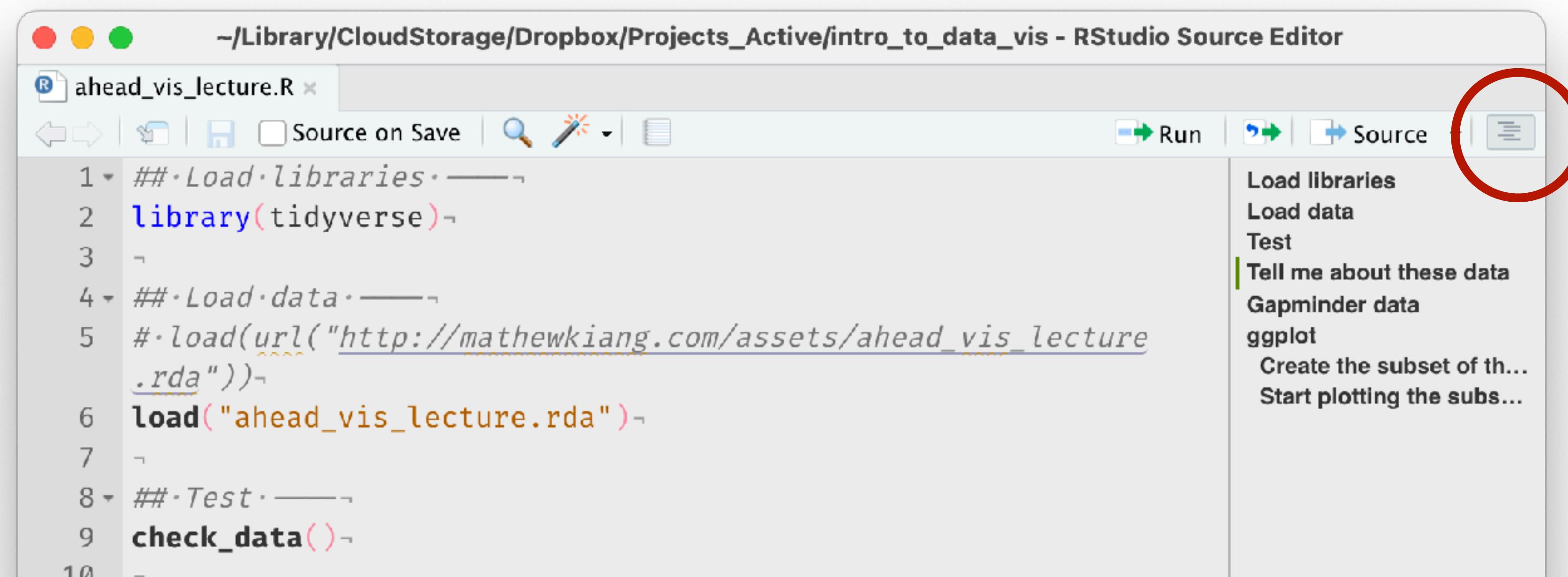
By the end of this class:

- You will have a framework to create, understand, critique, and reproduce statistical graphics
- You will learn how to create basic (but important) plots
- You will know enough code to expand your visualization toolkit on your own



Before we start

1. Open the ahead_vis_lecture.R file in RStudio
2. Enable “Show Document Outline”
3. Load the libraries and datasets (lines 2-6)
4. Run check_data() – if it doesn’t say “All datasets loaded.” let me know!

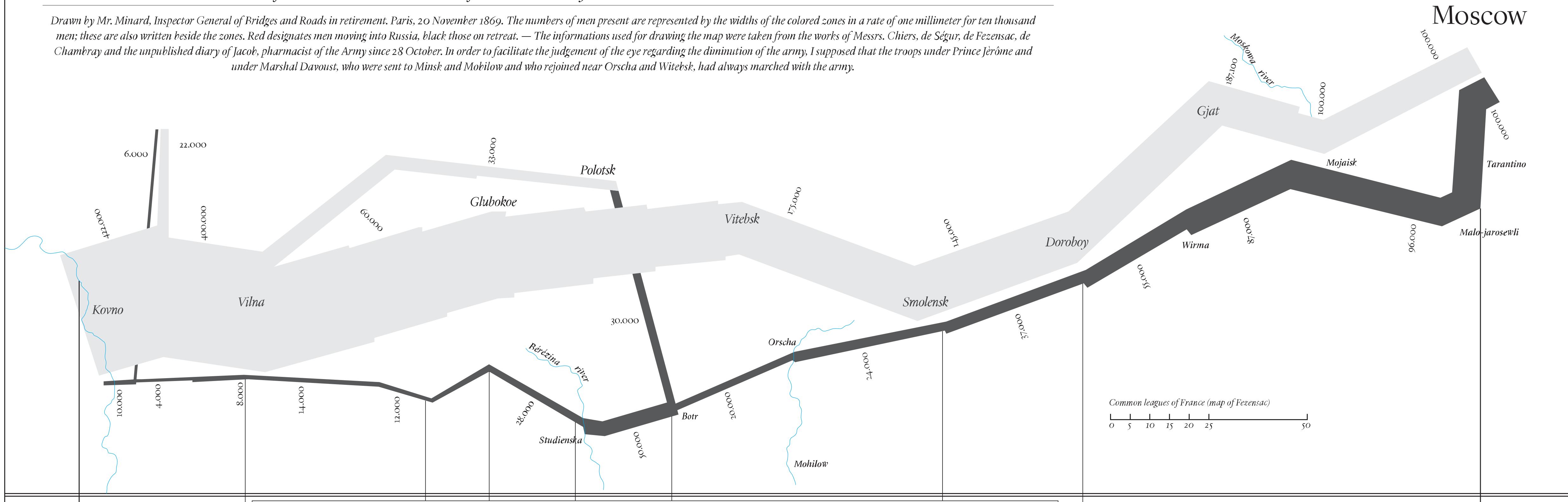


What is data visualization?

“The best statistical graphic ever drawn”

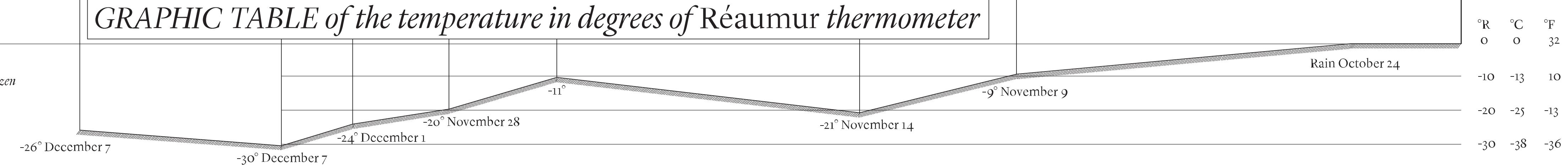
FIGURATIVE MAP of the successive losses in men of the French Army in the RUSSIAN CAMPAIGN OF 1812-1813

Drawn by Mr. Minard, Inspector General of Bridges and Roads in retirement. Paris, 20 November 1869. The numbers of men present are represented by the widths of the colored zones in a rate of one millimeter for ten thousand men; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The informations used for drawing the map were taken from the works of Messrs. Chiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October. In order to facilitate the judgement of the eye regarding the diminution of the army, I supposed that the troops under Prince Jérôme and under Marshal Davoust, who were sent to Minsk and Mohilow and who rejoined near Orscha and Vitebsk, had always marched with the army.



GRAPHIC TABLE of the temperature in degrees of Réaumur thermometer

The Cossacks pass the frozen Niemen at a gallop



Geographical information

FIGURATIVE MAP of the successive losses in men of the French army during its retreat from Moscow in 1812.

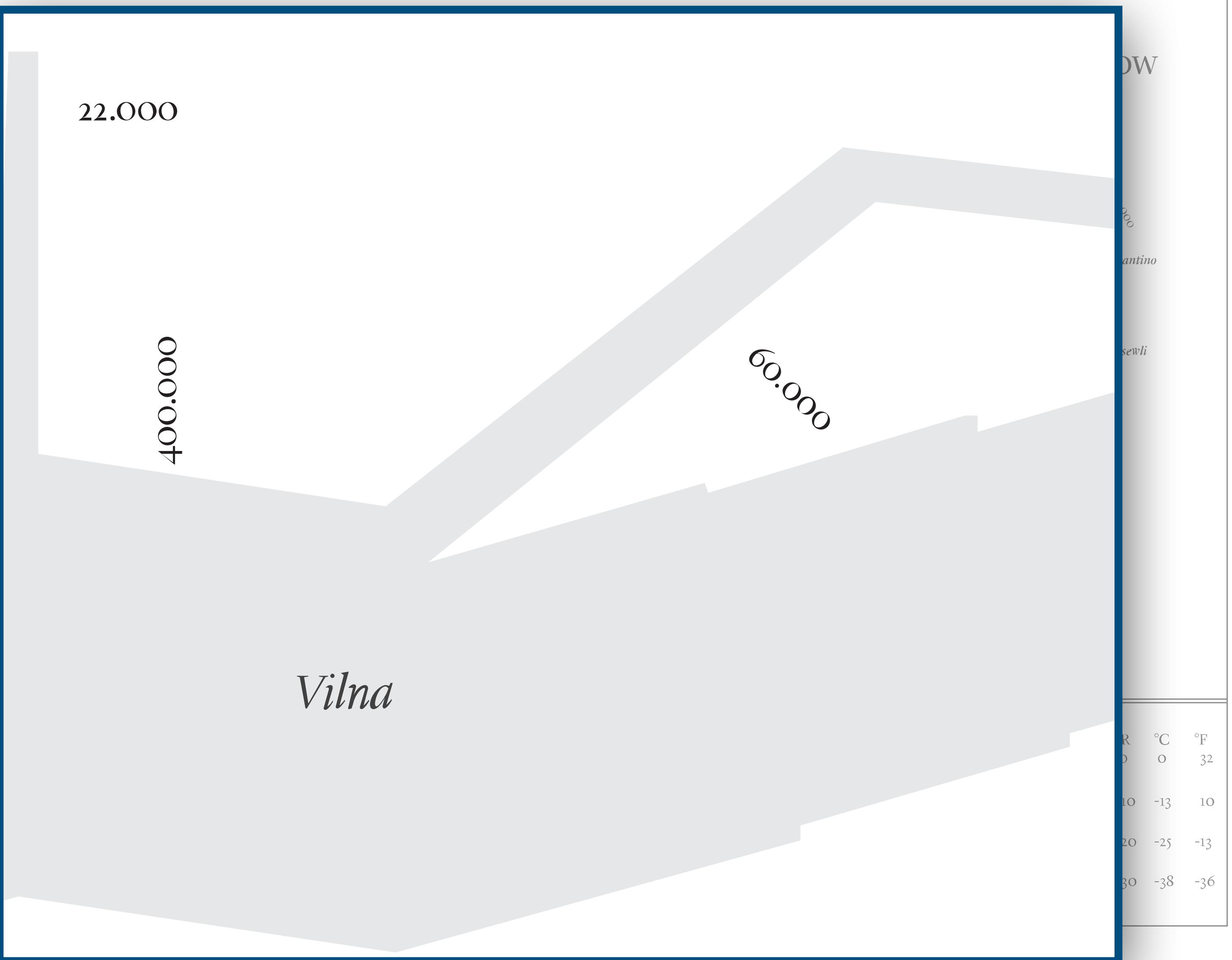
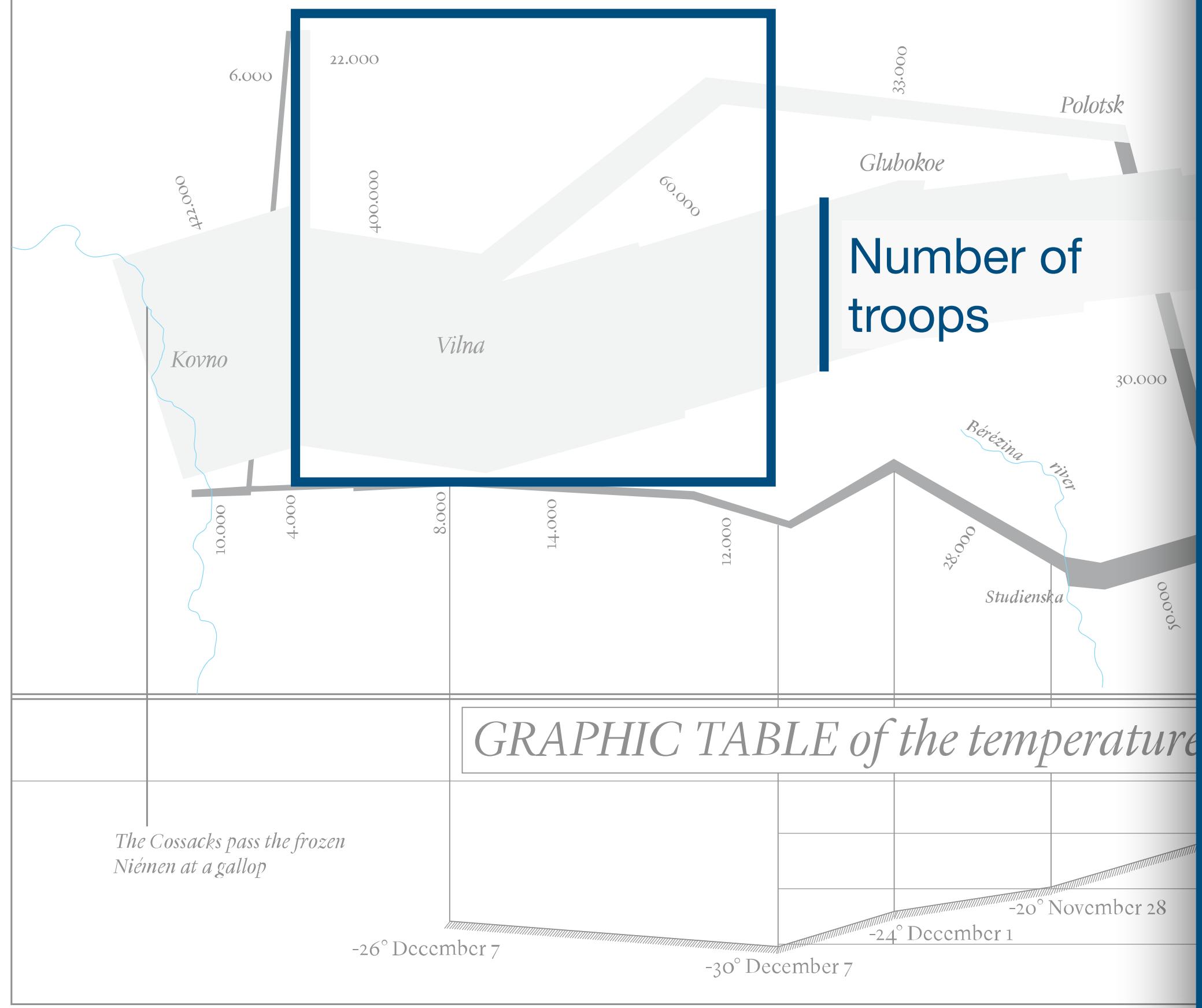
Drawn by Mr. Minard, Inspector General of Bridges and Roads in retirement. Paris, 20 November 1869. The numbers represent the number of men; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The map is based on the memoirs of Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October. In order to facilitate the comparison with the map of Fezensac, the distances are given in leagues of France (map of Fezensac).



Number of troops

FIGURATIVE MAP of the successive losses in men of the French Army in the RU

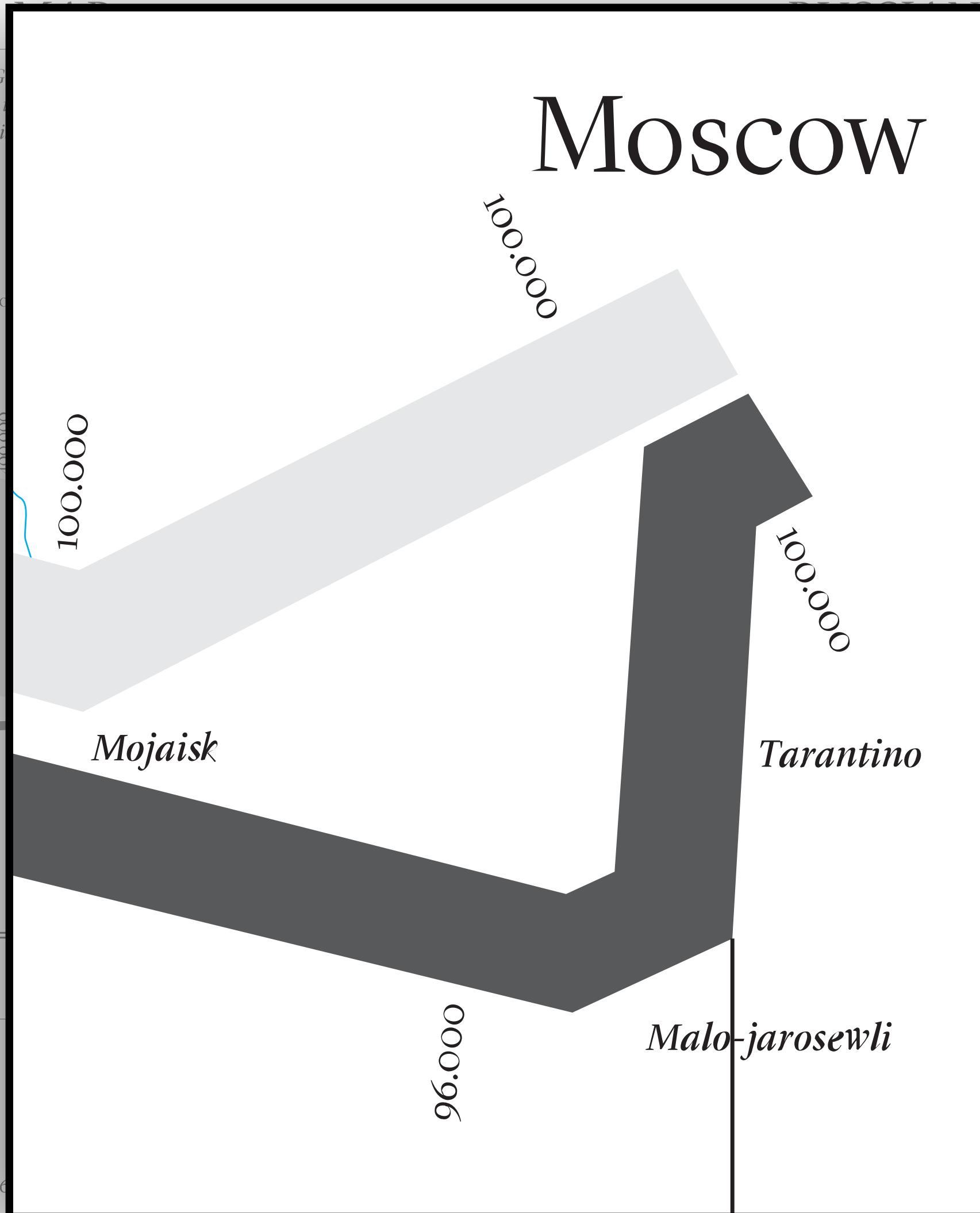
Drawn by Mr. Minard, Inspector General of Bridges and Roads in retirement. Paris, 20 November 1869. The numbers of men present are red; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The informations used for drawing were taken from the unpublished diary of M. Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October. In order to facilitate the judgement of the eye regarding the movements of the Russian army, the names of the principal cities are indicated. The names of the cities where the French army was quartered under Marshal Davoust, who were sent to Minsk and Mobilow and who rejoined near Orscha and Brest, are also indicated.



Progression of military movement

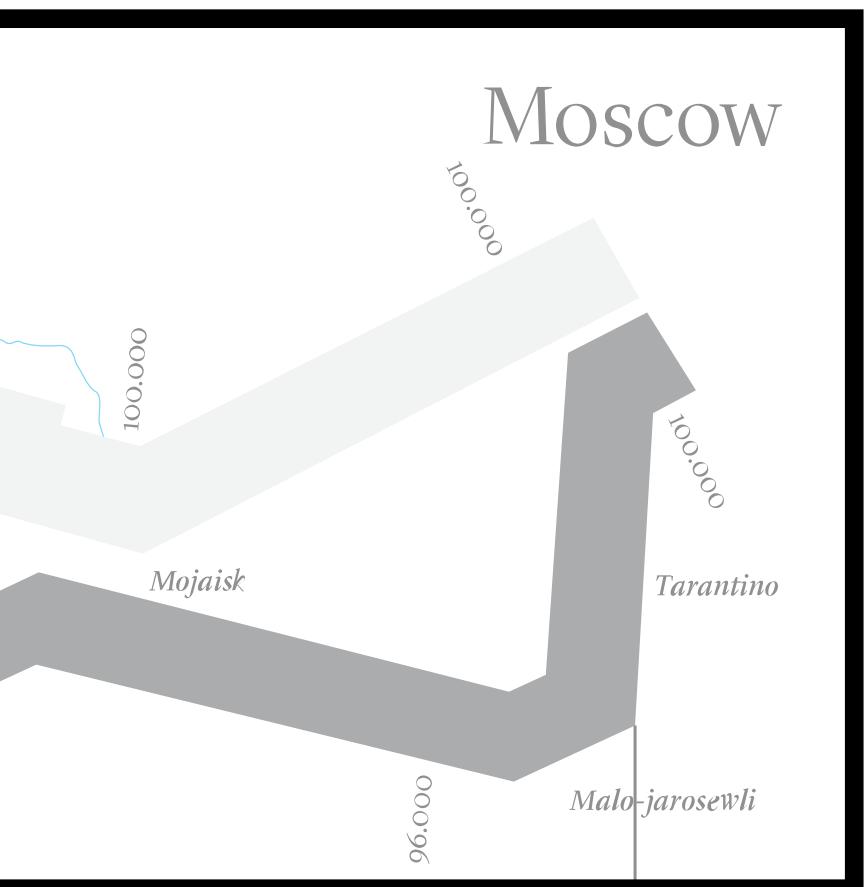
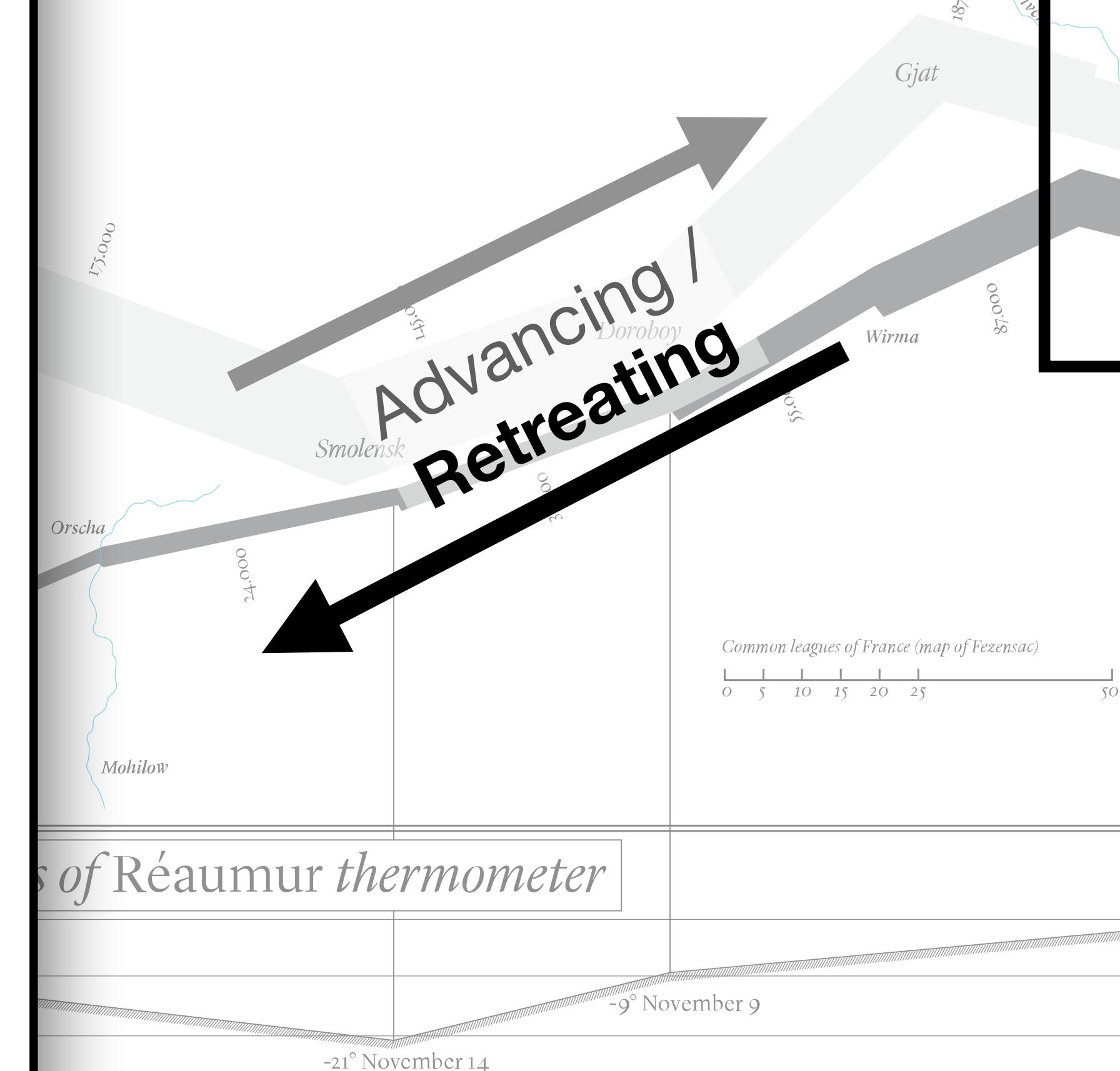
FIGURATIVE

Drawn by Mr. Minard, Inspector General of the Army; these are also written beside the maps of the campaign of 1812-1813.



CAMPAIGN OF 1812-1813

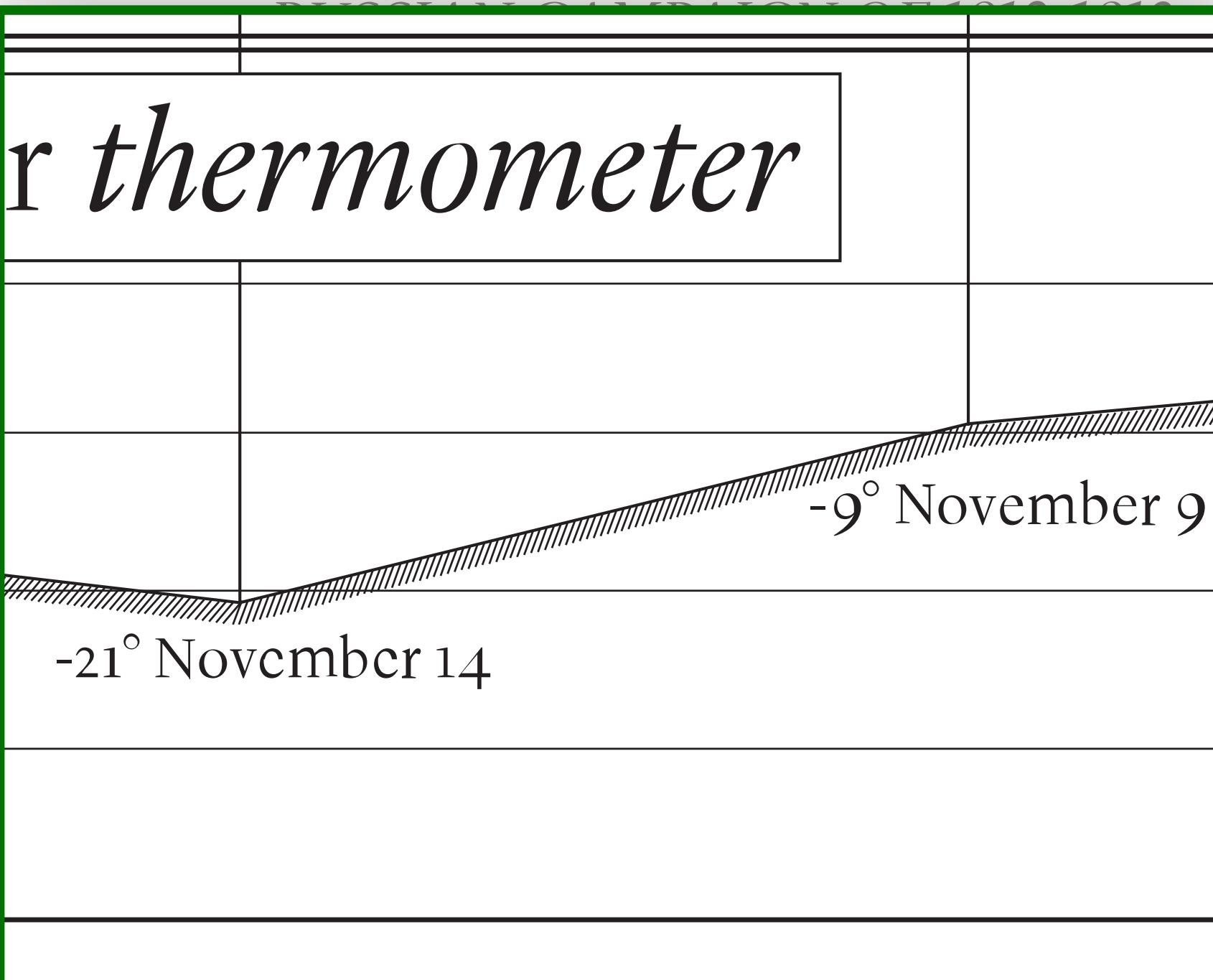
of the colored zones in a rate of one millimeter for ten thousand men from the works of Messrs. Chiers, de Ségur, de Fezensac, de la Motte, and others. I supposed that the troops under Prince Jérôme and his brother marched with the army.



Time and temperature (retreat only)

FIGURATIVE MAP of the successive losses in men of the French army during their retreat from Moscow in 1812.

Drawn by Mr. Minard, Inspector General of Bridges and Roads in retirement. Paris, 20 November 1869. This map shows the successive losses in men of the French army during their retreat from Moscow in 1812. Red designates men moving into Russia, black those on retreat. The map is based on Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October. In order to facilitate the reading of the map, it has been superimposed on a thermometer scale. The thermometer scale is also based on the same diary and on the unpublished memoirs of Marshal Davoust, who were sent to Minsk and Mohilow after the battle of Borodino.



Time and
Temperature

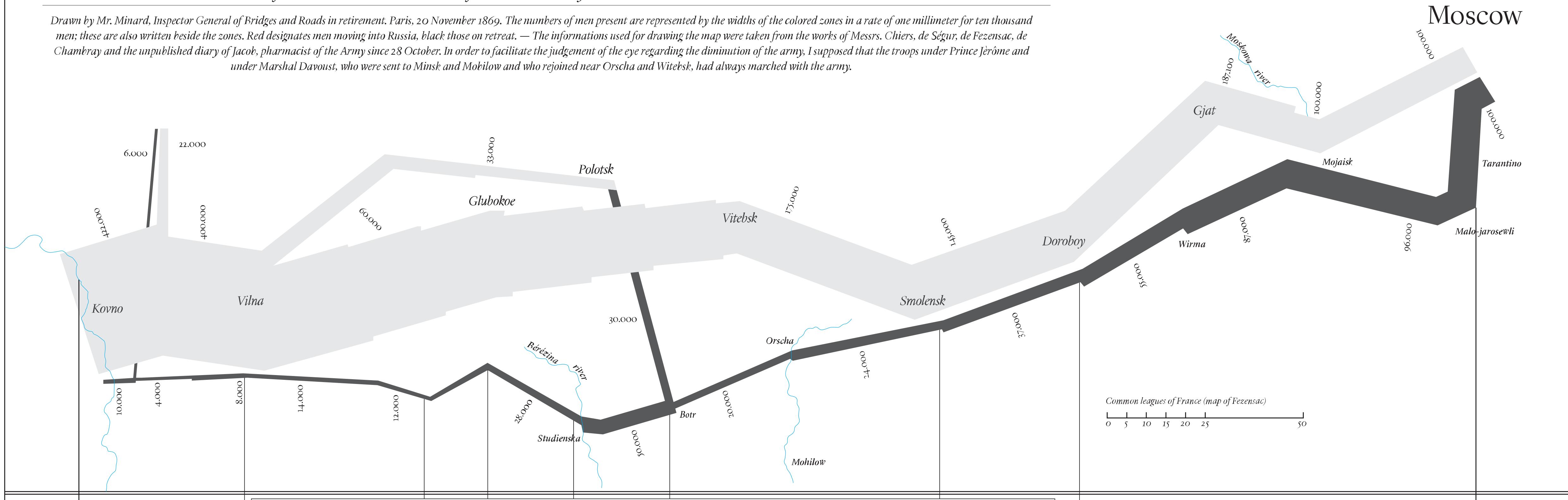


The Cossacks pass the frozen Niémen at a gallop

What is this plot showing us?

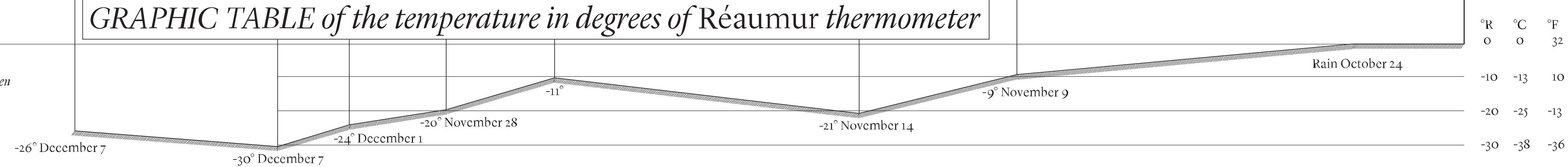
FIGURATIVE MAP of the successive losses in men of the French Army in the RUSSIAN CAMPAIGN OF 1812-1813

Drawn by Mr. Minard, Inspector General of Bridges and Roads in retirement. Paris, 20 November 1869. The numbers of men present are represented by the widths of the colored zones in a rate of one millimeter for ten thousand men; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The informations used for drawing the map were taken from the works of Messrs. Chiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October. In order to facilitate the judgement of the eye regarding the diminution of the army, I supposed that the troops under Prince Jérôme and under Marshal Davoust, who were sent to Minsk and Mohilow and who rejoined near Orscha and Vitebsk, had always marched with the army.



GRAPHIC TABLE of the temperature in degrees of Réaumur thermometer

The Cossacks pass the frozen Niemen at a gallop



So then, what *is*
data visualization?

The creation of
visual representations
of abstract data
to amplify cognition

The action

**The creation of
visual representations
of abstract data
to amplify cognition**

The product

The creation of
visual representations
of abstract data
to amplify cognition

The source material

The creation of
visual representations
of abstract data
to amplify cognition

The goal

The creation of
visual representations
of abstract data
to amplify cognition

Visualization has two parts

Encoding (i.e., creating the product)

1. Explore and manipulate abstract data
2. Encode that data into visual representations
3. Render the visual representation

Visualization has two parts

Encoding (i.e., creating the product)

1. Explore and manipulate abstract data
2. Encode that data into visual representations
3. Render the visual representation

Decoding (i.e., evaluating the product)

1. Perceive the visual representation
2. Interpret the visualization
3. Comprehend the visualization

Visualization has two parts

Encoding (i.e., creating the product)

1. Explore and manipulate abstract data
2. Encode that data into visual representations
3. Render the visual representation

Decoding (i.e., evaluating the product)

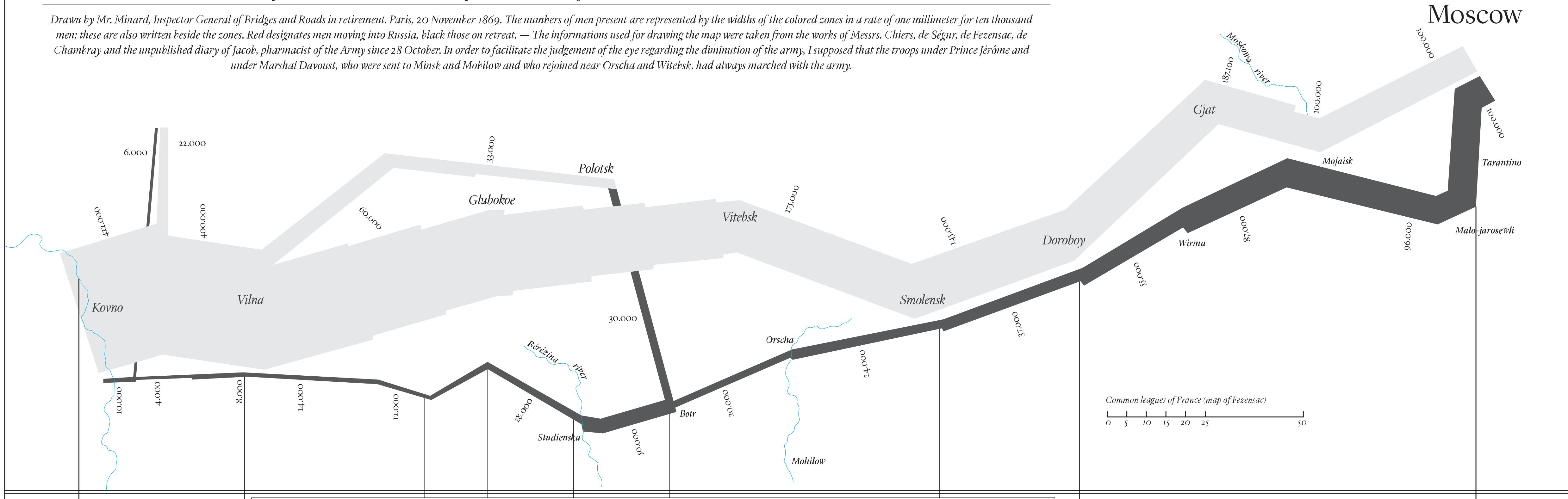
1. Perceive the visual representation
2. Interpret the visualization
3. Comprehend the visualization

Going to practice decoding first, then encoding.

How did we decode this figure?

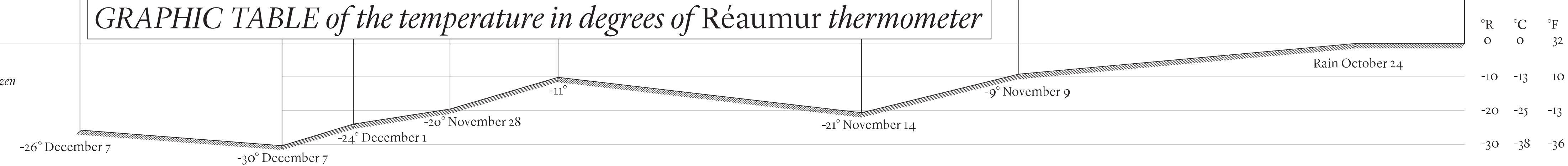
FIGURATIVE MAP of the successive losses in men of the French Army in the RUSSIAN CAMPAIGN OF 1812-1813

Drawn by Mr. Minard, Inspector General of Bridges and Roads in retirement. Paris, 20 November 1869. The numbers of men present are represented by the widths of the colored zones in a rate of one millimeter for ten thousand men; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The informations used for drawing the map were taken from the works of Messrs. Chiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October. In order to facilitate the judgement of the eye regarding the diminution of the army, I supposed that the troops under Prince Jérôme and under Marshal Davoust, who were sent to Minsk and Mohilow and who rejoined near Orscha and Vitebsk, had always marched with the army.



GRAPHIC TABLE of the temperature in degrees of Réaumur thermometer

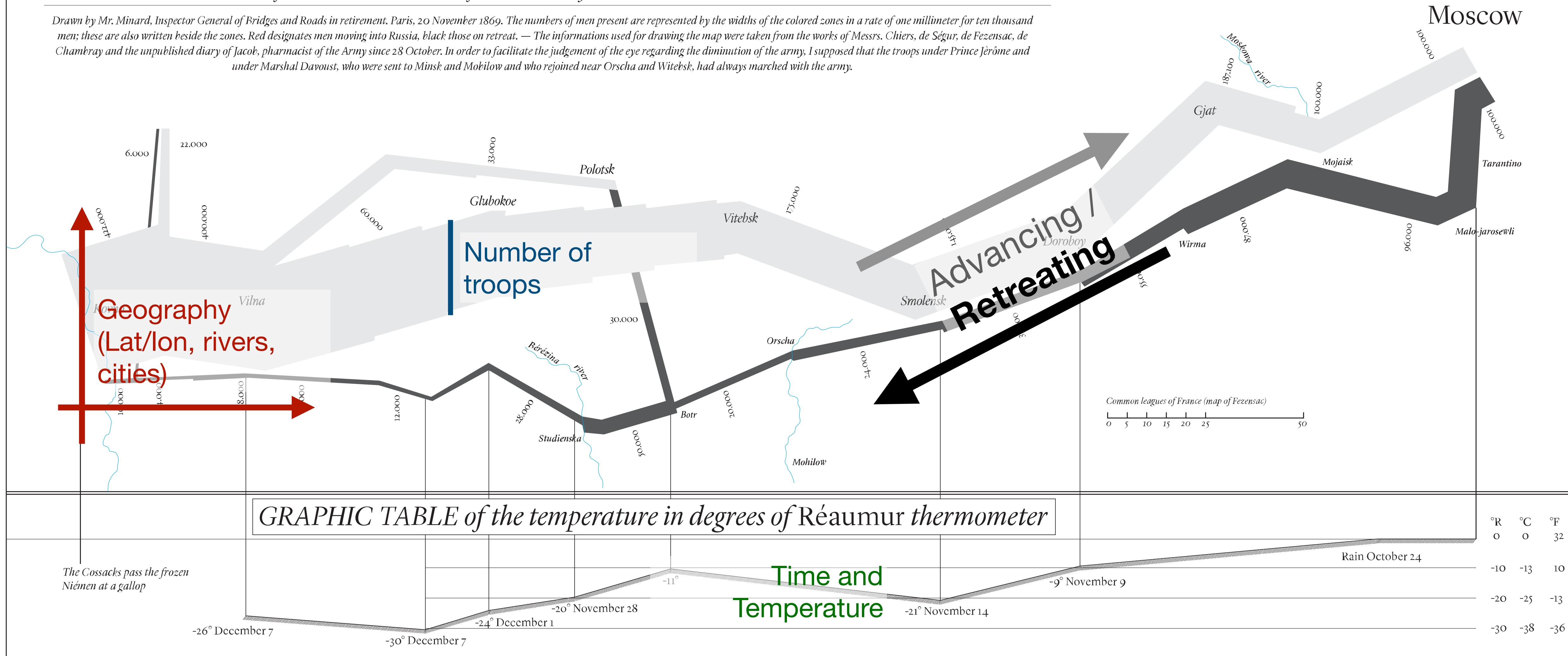
The Cossacks pass the frozen Niémen at a gallop



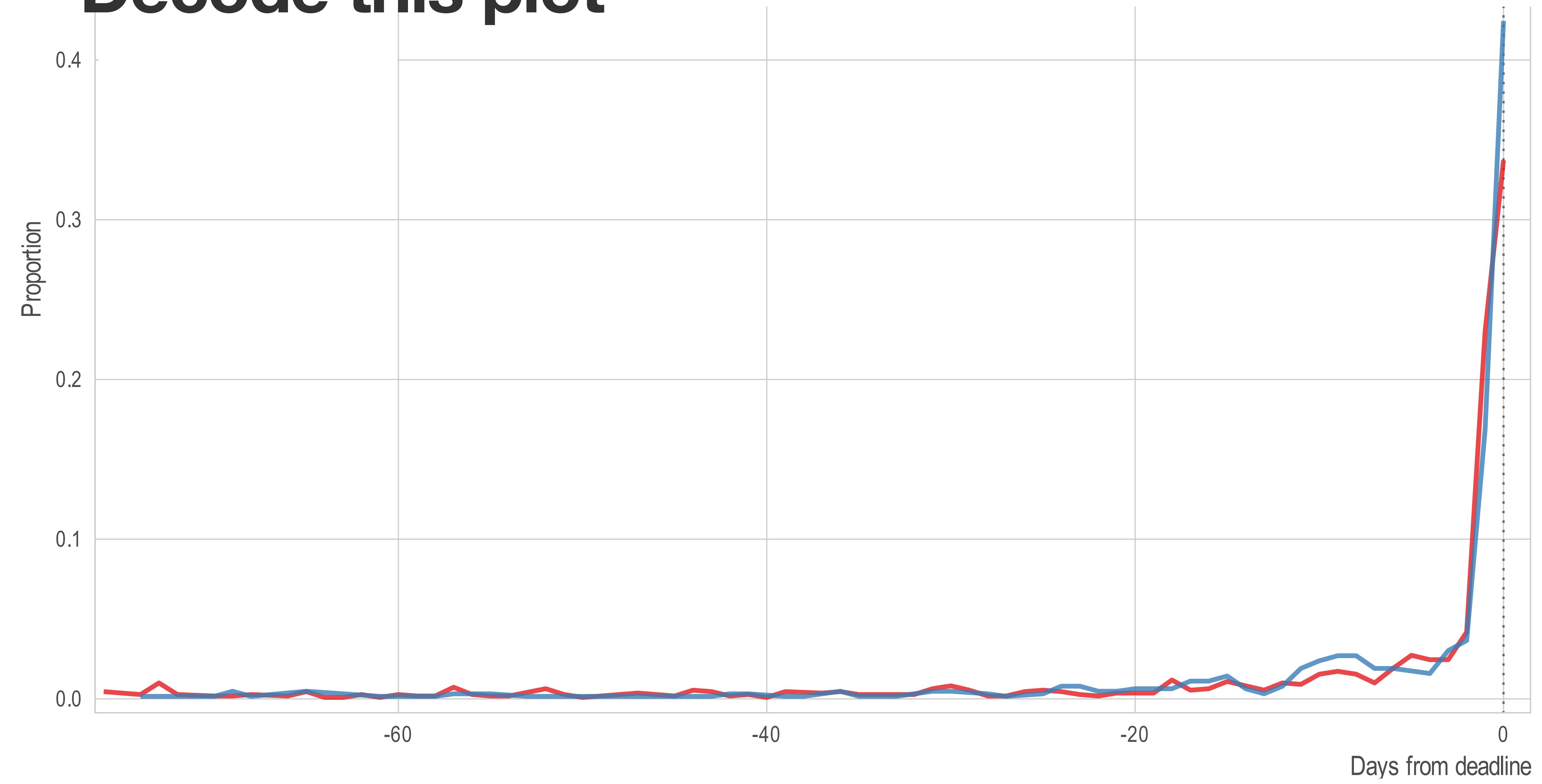
How did Minard encode this figure?

FIGURATIVE MAP of the successive losses in men of the French Army in the RUSSIAN CAMPAIGN OF 1812-1813

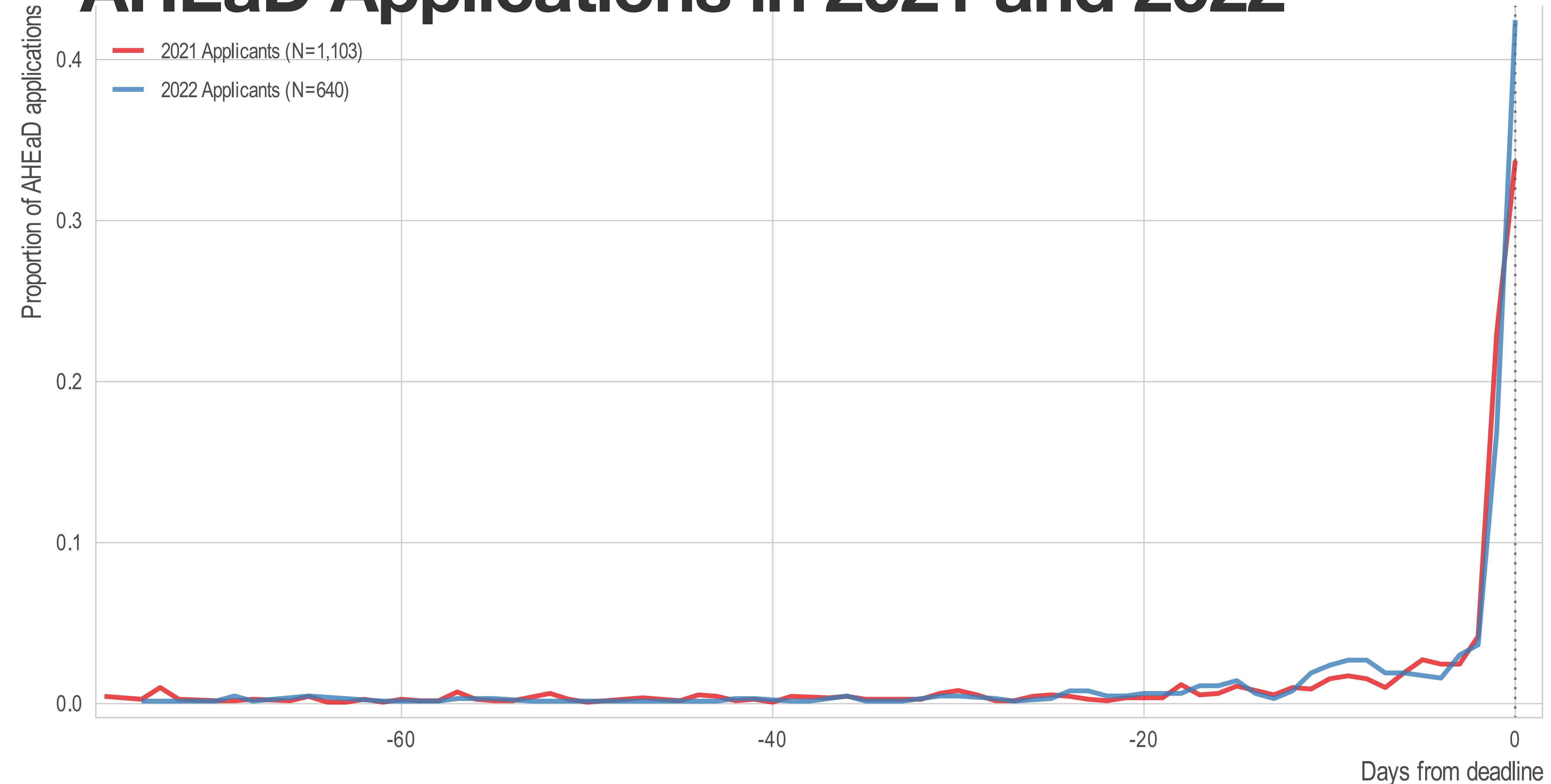
Drawn by Mr. Minard, Inspector General of Bridges and Roads in retirement. Paris, 20 November 1869. The numbers of men present are represented by the widths of the colored zones in a rate of one millimeter for ten thousand men; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The informations used for drawing the map were taken from the works of Messrs. Chiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October. In order to facilitate the judgement of the eye regarding the diminution of the army, I supposed that the troops under Prince Jérôme and under Marshal Davoust, who were sent to Minsk and Mohilow and who rejoined near Orscha and Vitebsk, had always marched with the army.



Decode this plot



AHEaD Applications in 2021 and 2022



Why should you care
about visualization?

Roles of data visualization

1. Helps you **understand your data** and avoid errors
2. Accurately **interpret other results** from media or scientists
3. **Effectively communicate** your results

1. Understanding your data

Tell me about these data

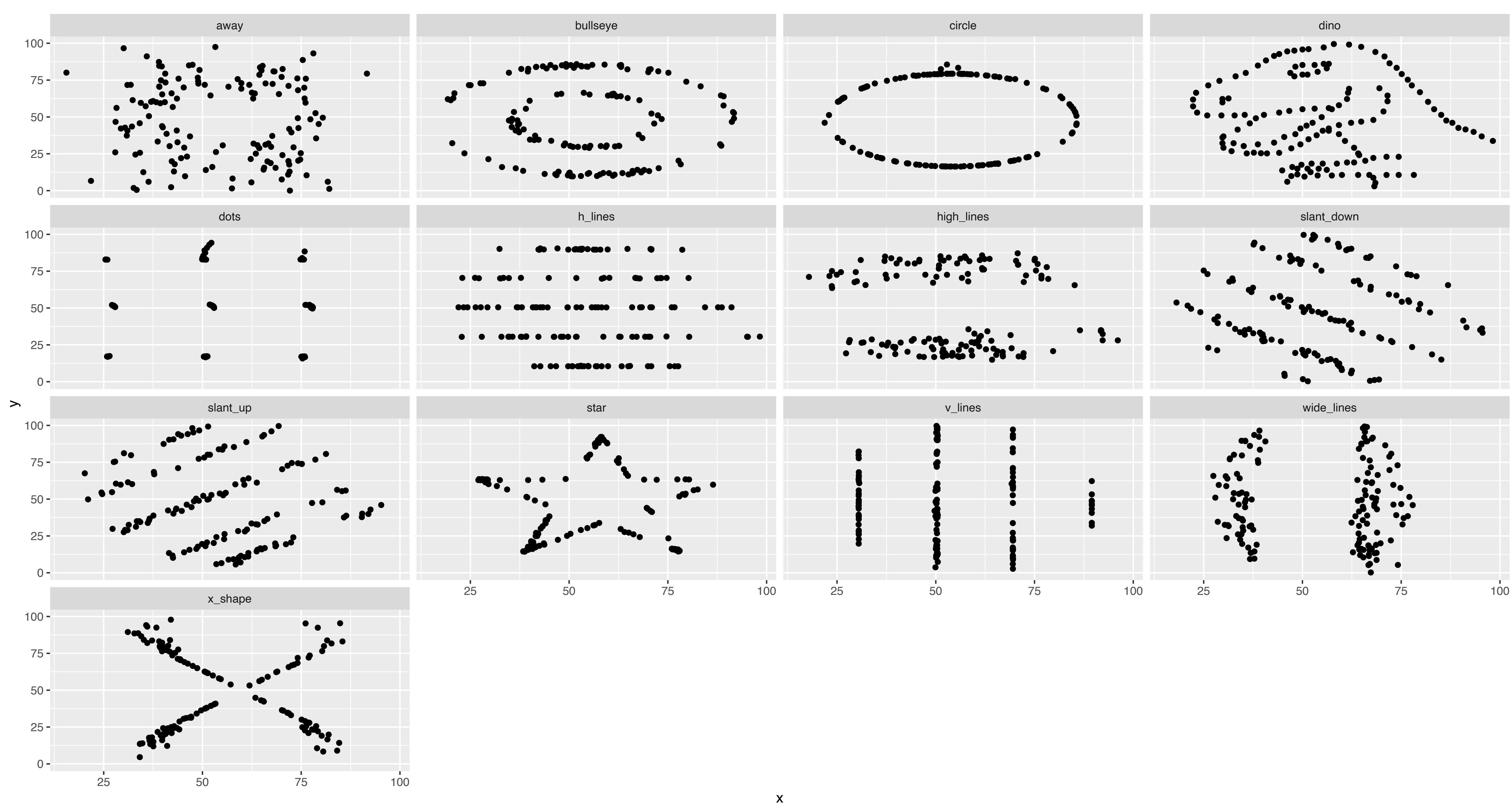
```
> dd
# A tibble: 1,846 × 3
  dataset      x      y
  <chr>    <dbl>  <dbl>
1 dino      55.4   97.2
2 dino      51.5   96.0
3 dino      46.2   94.5
4 dino      42.8   91.4
5 dino      40.8   88.3
6 dino      38.7   84.9
7 dino      35.6   79.9
8 dino      33.1   77.6
9 dino      29.0   74.5
10 dino     26.2   71.4
# ... with 1,836 more rows
```

Essentially identical data sets (?)

```
> dd %>%
+   group_by(dataset) %>%
+   summarize(
+     mean_x = mean(x),
+     sd_x = sd(x),
+     mean_y = mean(y),
+     sd_y = sd(y),
+     corr = cor(x, y)
+   )
# A tibble: 13 × 6
  dataset    mean_x    sd_x  mean_y    sd_y    corr
  <chr>      <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 away        54.3    16.8    47.8    26.9 -0.0641
2 bullseye    54.3    16.8    47.8    26.9 -0.0686
3 circle      54.3    16.8    47.8    26.9 -0.0683
4 dino         54.3    16.8    47.8    26.9 -0.0645
5 dots         54.3    16.8    47.8    26.9 -0.0603
6 h_lines      54.3    16.8    47.8    26.9 -0.0617
7 high_lines   54.3    16.8    47.8    26.9 -0.0685
8 slant_down   54.3    16.8    47.8    26.9 -0.0690
9 slant_up     54.3    16.8    47.8    26.9 -0.0686
10 star         54.3    16.8    47.8    26.9 -0.0630
11 v_lines      54.3    16.8    47.8    26.9 -0.0694
12 wide_lines   54.3    16.8    47.8    26.9 -0.0666
13 x_shape      54.3    16.8    47.8    26.9 -0.0656
```

Look at the data

```
ggplot(dd,  
       aes(x = x,  
            y = y)) +  
  geom_point() +  
  facet_wrap(~ dataset)
```

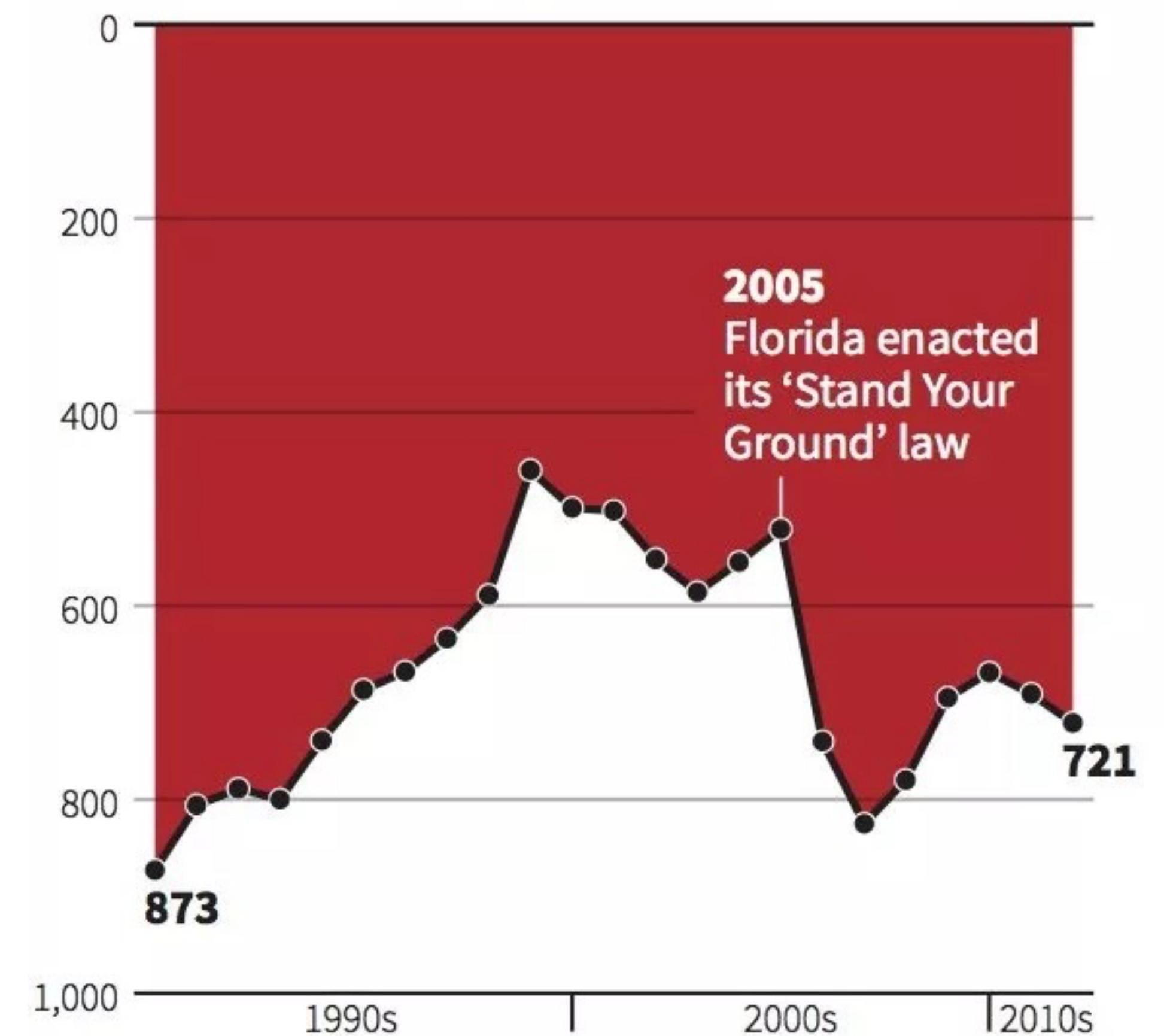


Source: <https://www.rdocumentation.org/packages/datasauRus/versions/0.1.6>

2. Interpret other plots

Gun deaths in Florida

Number of murders committed using firearms



Source: Florida Department of Law Enforcement

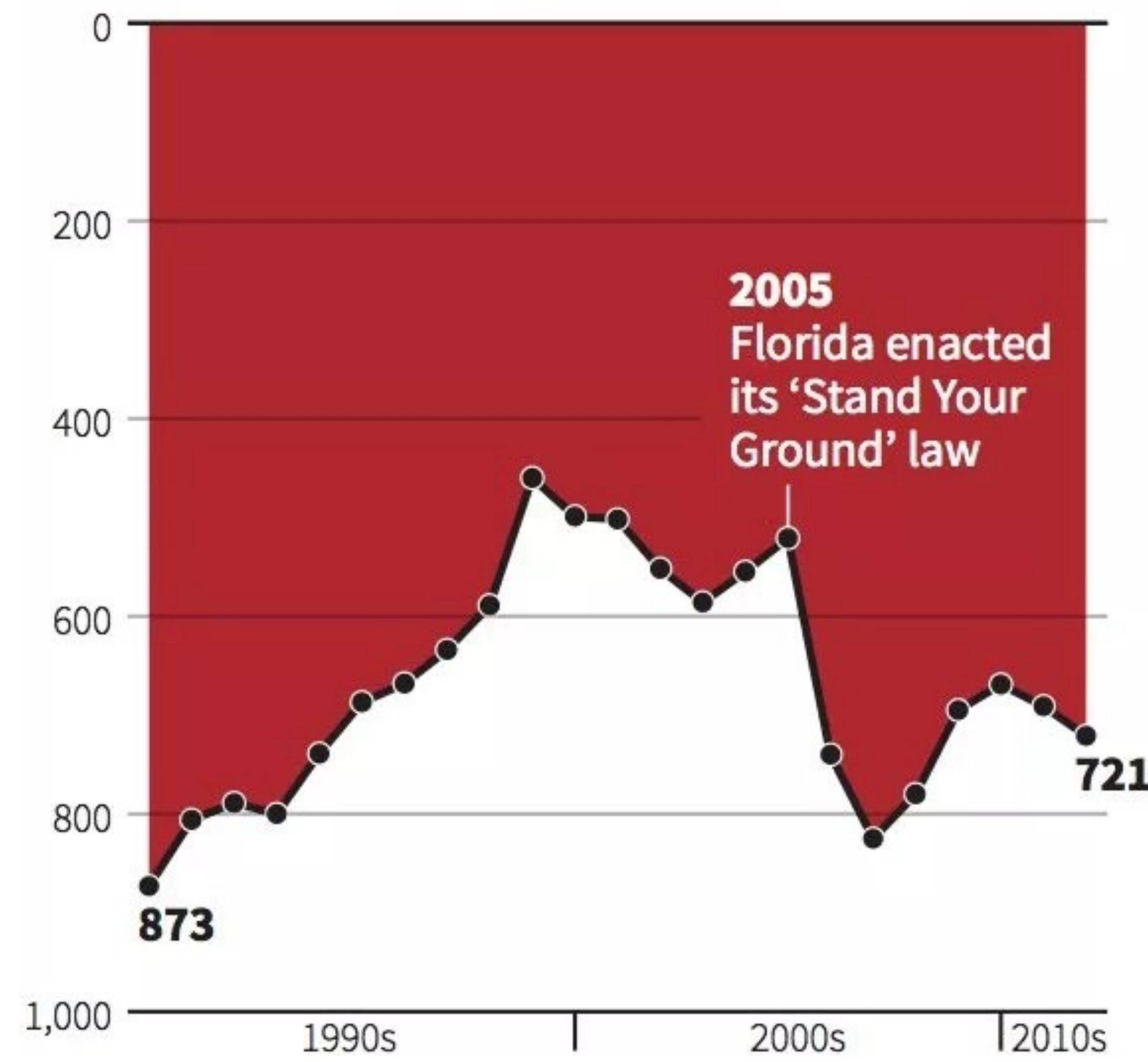
C. Chan 16/02/2014

REUTERS

Inverted y-axis

Gun deaths in Florida

Number of murders committed using firearms



Source: Florida Department of Law Enforcement

SOUTHWEST BORDER APPREHENSIONS

OCTOBER - APRIL



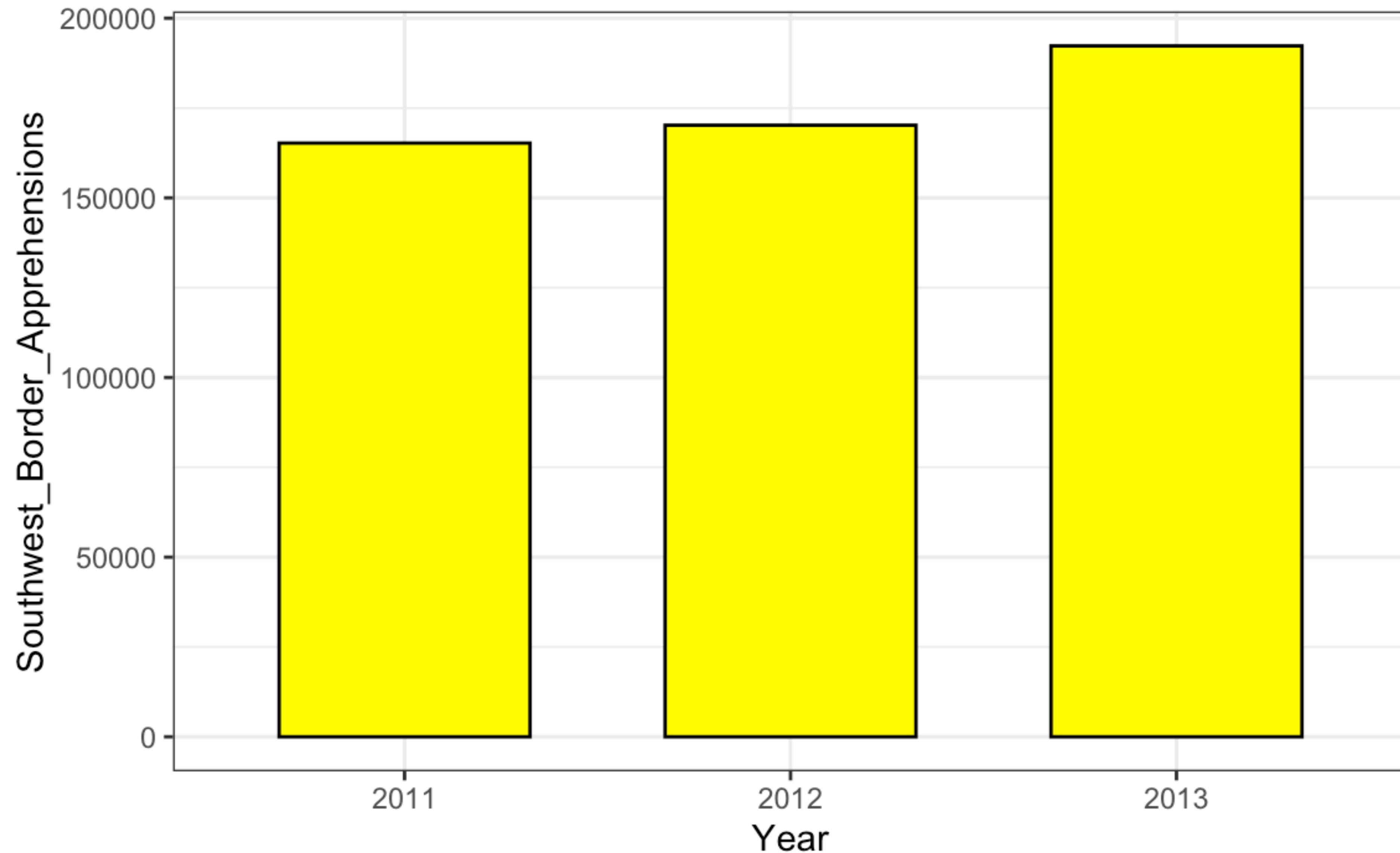
Source: U.S. Border Patrol

Happening
now

\$43 BIL IN FEB, DOWN 3.4% FROM JAN'S REVISED

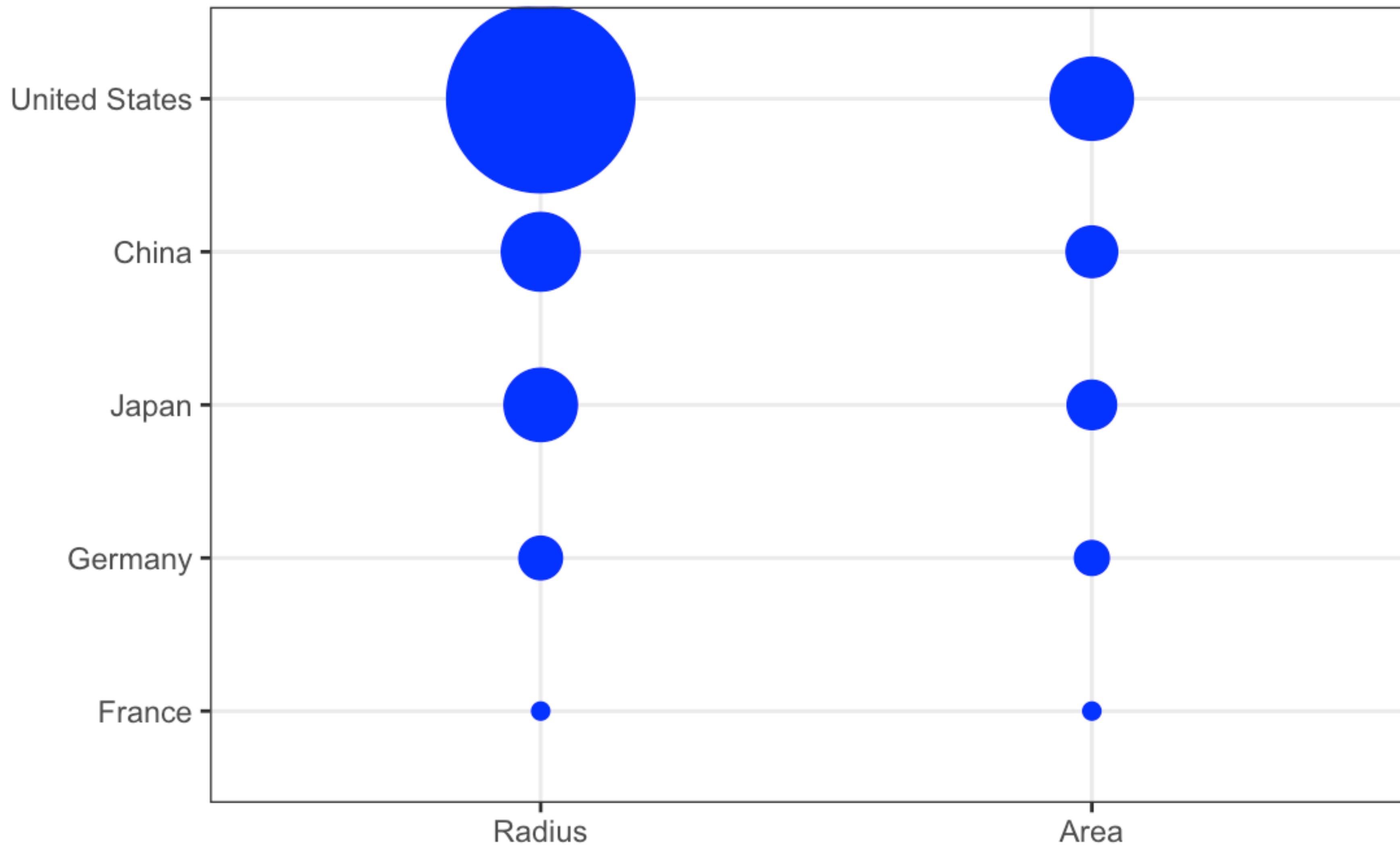
NAS ▼ 35.18

Not including 0 on the y-axis



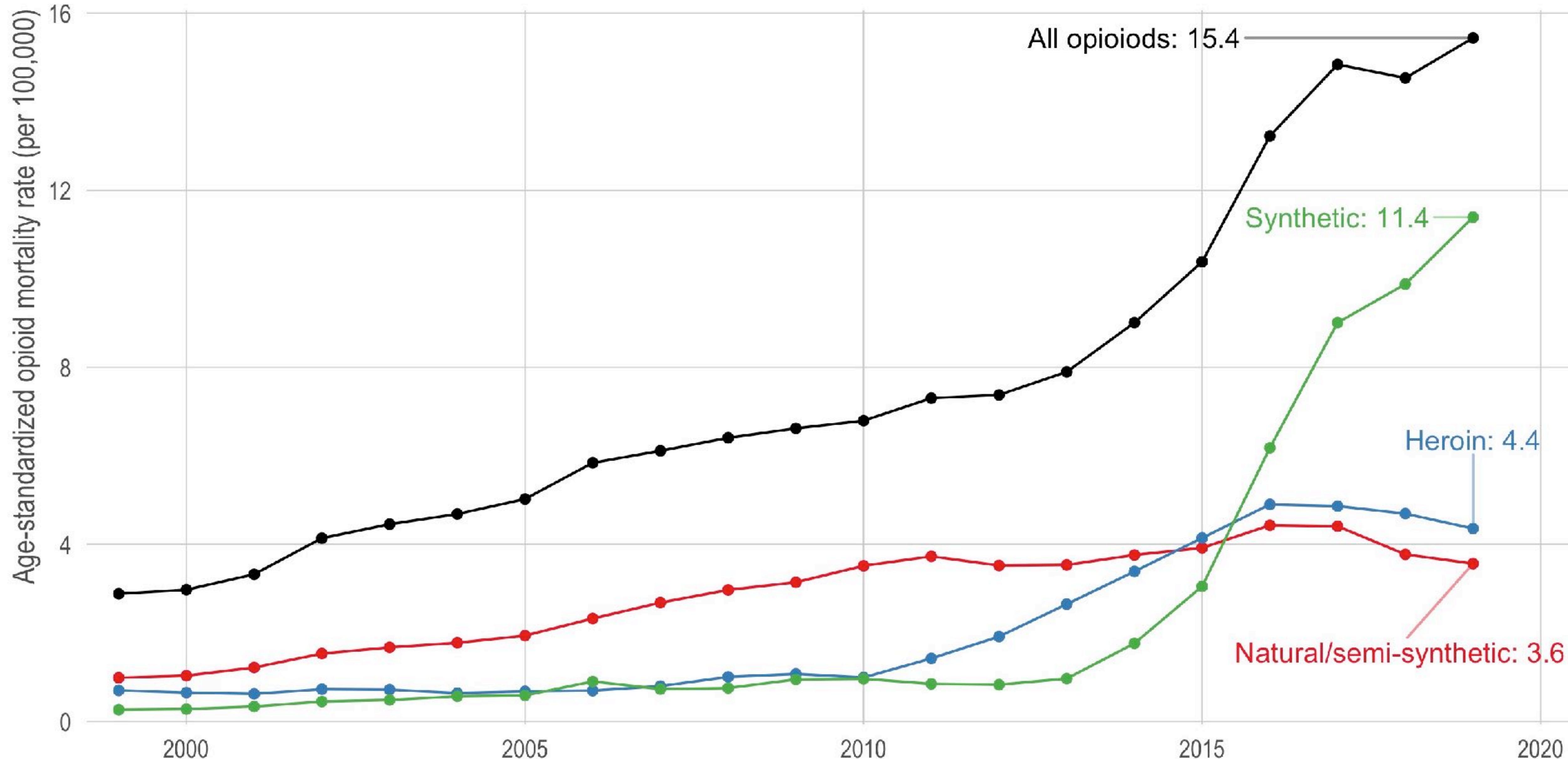


Using perceptually-deceptive scales

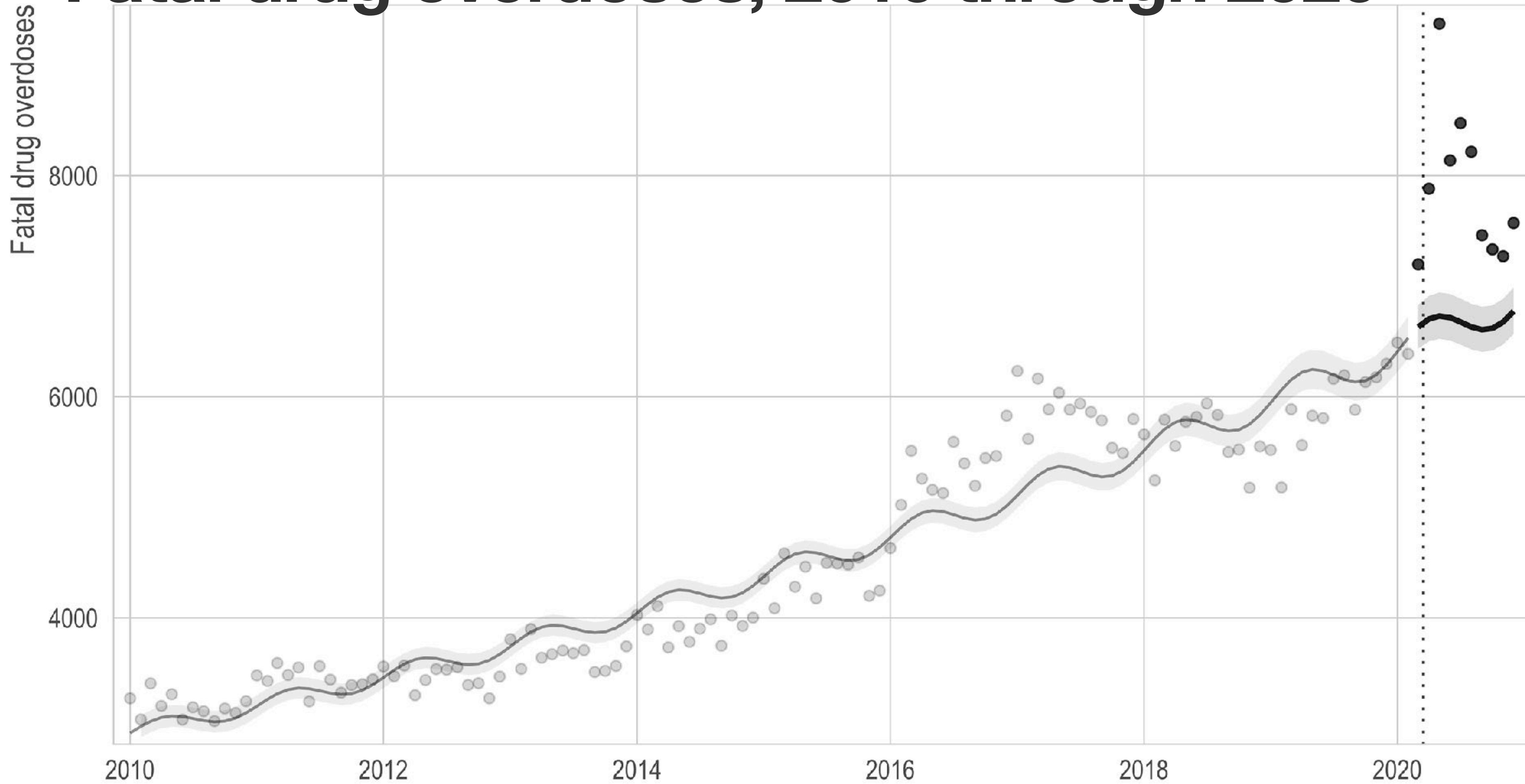


3. Communicate your results

The “Triple Wave” Opioid Epidemic



Fatal drug overdoses, 2010 through 2020



What do we mean by
“encoding”?

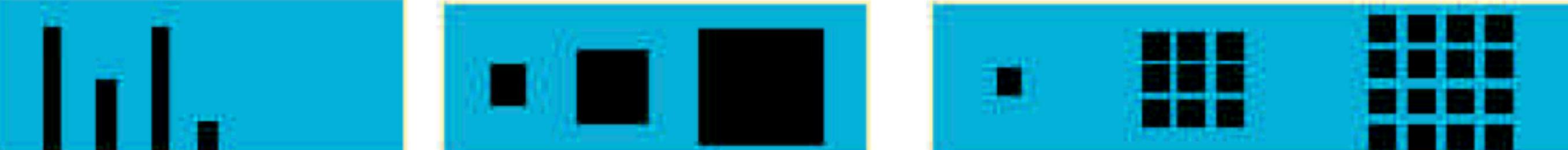
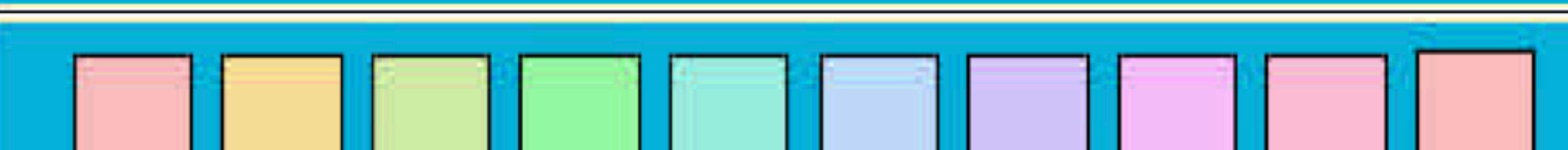
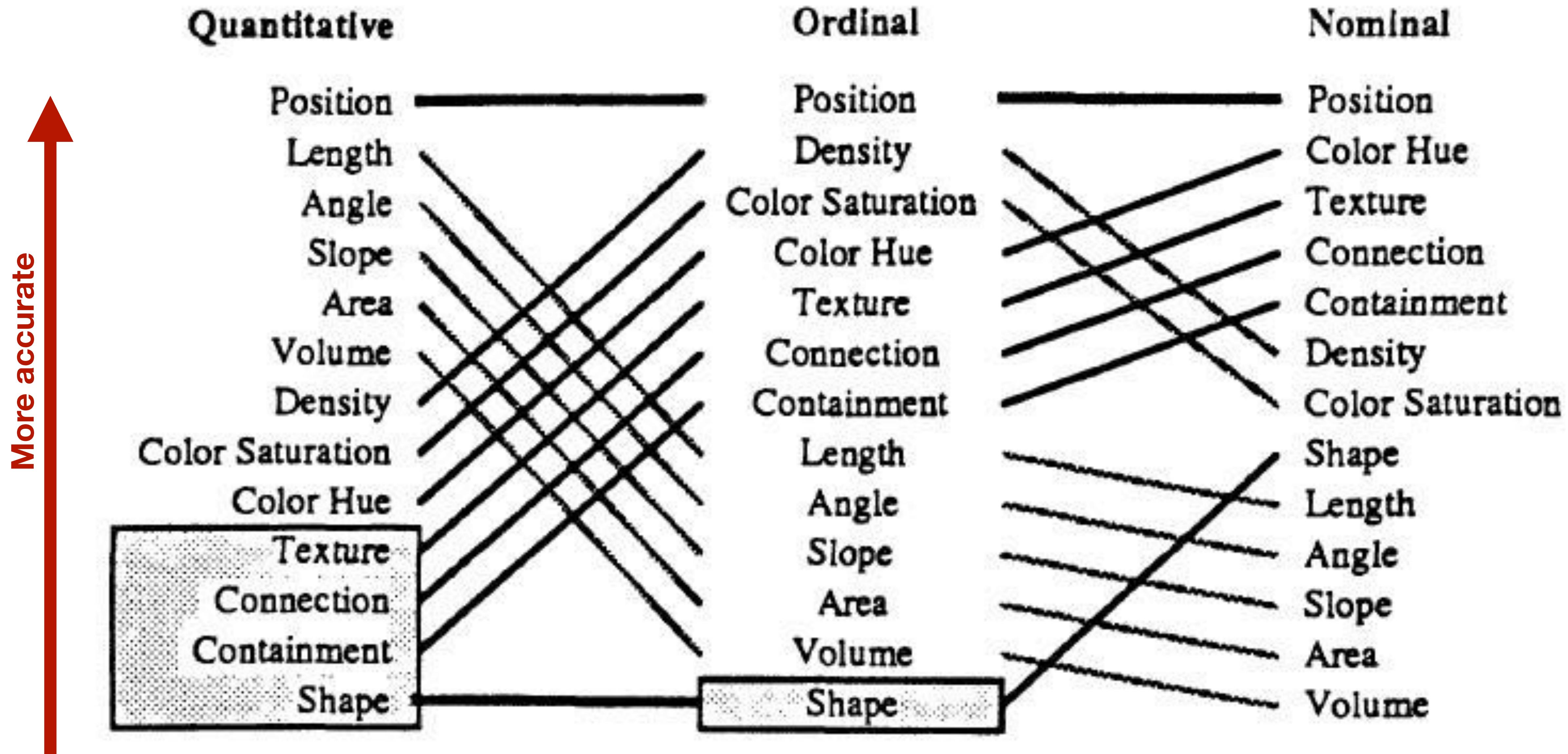
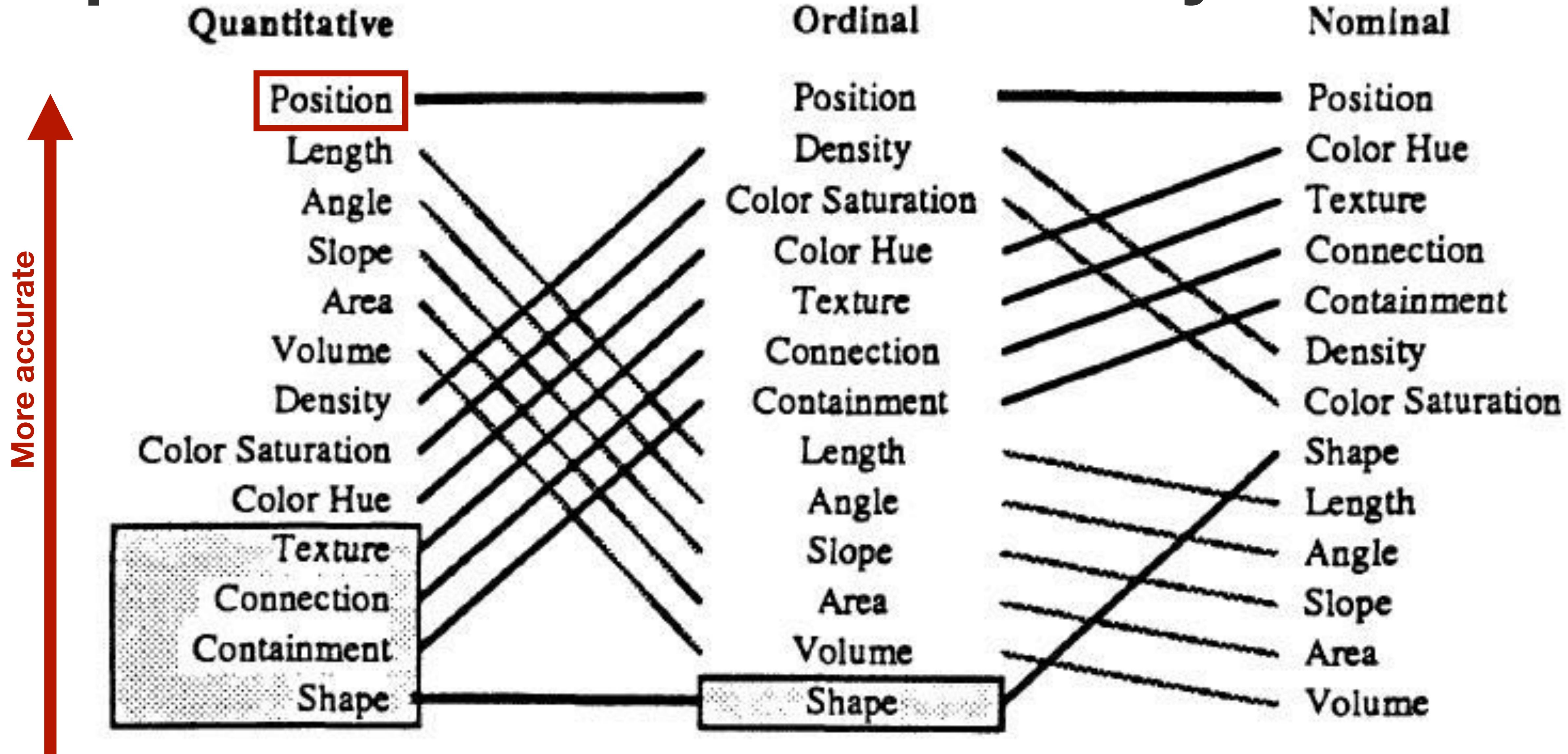
Bertin's Original Visual Variables	
Position changes in the x, y location	
Size change in length, area or repetition	
Shape infinite number of shapes	
Value changes from light to dark	
Colour changes in hue at a given value	
Orientation changes in alignment	
Texture variation in 'grain'	

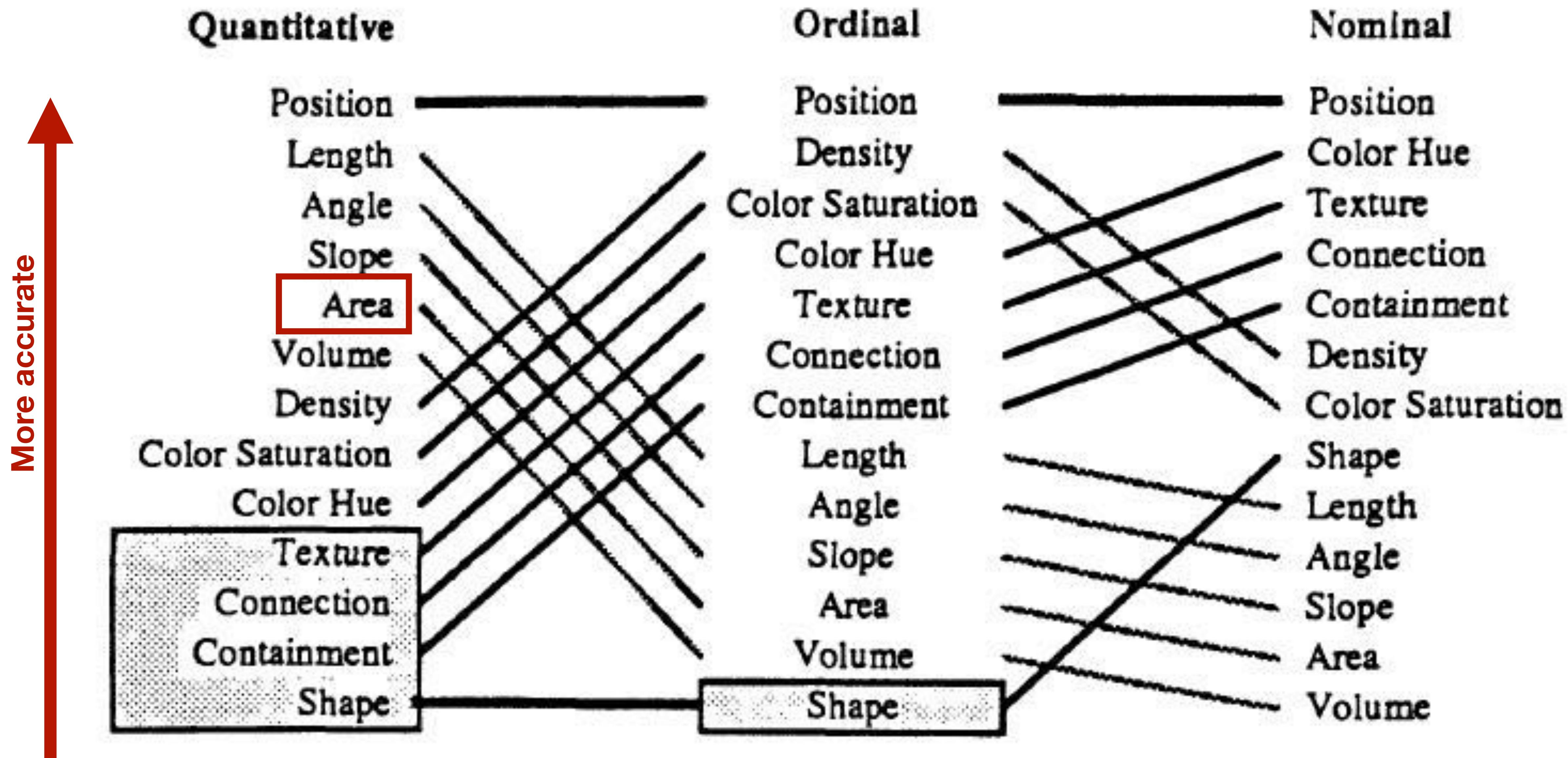
Table 1: These are Bertin's visual variables



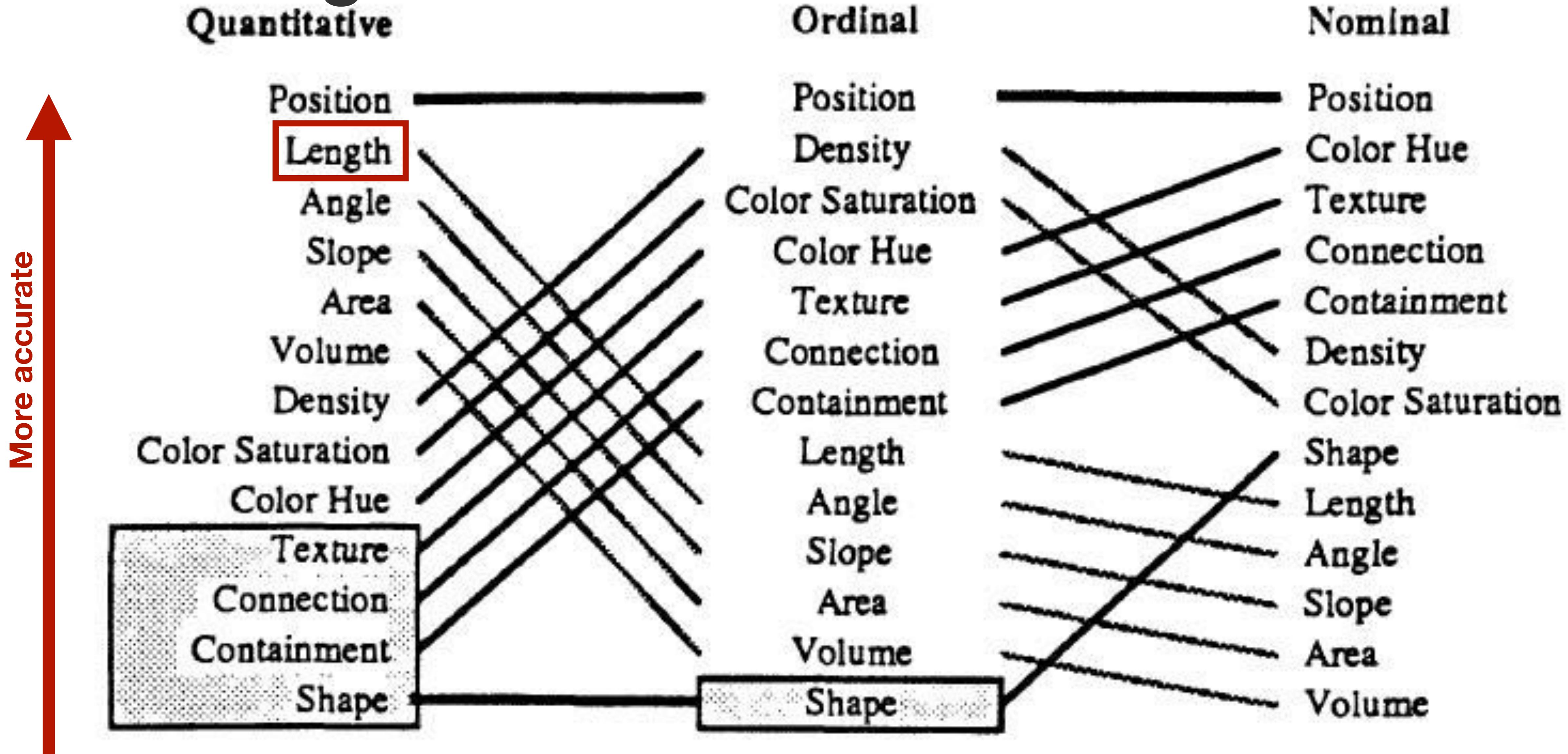
Important data should be on x- or y-axis



Pie charts are bad because area is hard



Truncating axes on bar charts is bad



How do we encode
visual information?

Visualization has two parts

Encoding (i.e., creating the product)

1. Explore and manipulate abstract data
2. Encode that data into visual representations
3. Render the visual representation

Decoding (i.e., evaluating the product)

1. Perceive the visual representation
2. Interpret the visualization
3. Comprehend the visualization

Going to practice decoding first, then encoding.

Tell me about these data

1. How many observations?
2. How many columns?
3. How many countries?
4. How many years?
5. Characteristics of the columns (ordinal, numeric, categorical)?
6. What else?

```
> gap
# A tibble: 1,295 × 9
  country      year infant_mortality life_expectancy fertility population
  <fct>     <int>        <dbl>            <dbl>       <dbl>      <dbl>
1 Albania      1990        35.1             73.3        2.97    3281453
2 Algeria      1990        39.7             70.2        4.76    25912364
3 Angola       1990       134.              48.6        7.17    11127870
4 Antigua and ... 1990        24               73.5        2.06     61906
5 Argentina    1990        24.4             72.5        2.99    32729740
6 Armenia      1990        42.5             70.1        2.54    3544695
7 Aruba         1990        NA               73.5        2.25     62148
8 Australia     1990        7.6              77          1.9     17096869
9 Austria       1990         8               75.7        1.46    7706571
10 Azerbaijan   1990        75.5             65.6        2.97    7216503
# ... with 1,285 more rows, and 3 more variables: gdp <dbl>, continent <fct>,
#   region <fct>
```

1. Explore and manipulate abstract data

Create a subset of this file, called sub_gap, for 2010 and 1990 only.

```
> gap
# A tibble: 1,295 × 9
  country      year infant_mortality life_expectancy fertility population
  <fct>     <int>        <dbl>            <dbl>       <dbl>        <dbl>
1 Albania      1990        35.1             73.3        2.97    3281453
2 Algeria      1990        39.7             70.2        4.76    25912364
3 Angola       1990       134.              48.6        7.17    11127870
4 Antigua and ... 1990        24               73.5        2.06     61906
5 Argentina    1990        24.4             72.5        2.99    32729740
6 Armenia      1990        42.5             70.1        2.54    3544695
7 Aruba         1990        NA               73.5        2.25     62148
8 Australia     1990        7.6              77          1.9     17096869
9 Austria       1990         8               75.7        1.46    7706571
10 Azerbaijan   1990        75.5             65.6        2.97    7216503
# ... with 1,285 more rows, and 3 more variables: gdp <dbl>, continent <fct>,
#   region <fct>
```

1. Explore and manipulate abstract data

```
sub_gap <- gap %>%
  filter(year == 1990 | year == 2010)

> gap
# A tibble: 1,295 × 9
  country      year infant_mortality life_expectancy fertility population
  <fct>       <int>        <dbl>            <dbl>        <dbl>        <dbl>
  1 Albania     1990         35.1             73.3        2.97    3281453
  2 Algeria     1990         39.7             70.2        4.76    25912364
  3 Angola      1990        134.              48.6        7.17    11127870
  4 Antigua and ... 1990         24               73.5        2.06    61906
  5 Argentina   1990         24.4             72.5        2.99    32729740
  6 Armenia     1990         42.5             70.1        2.54    3544695
  7 Aruba       1990          NA              73.5        2.25    62148
  8 Australia   1990          7.6              77          1.9    17096869
  9 Austria     1990           8              75.7        1.46    7706571
 10 Azerbaijan  1990         75.5             65.6        2.97    7216503
# ... with 1,285 more rows, and 3 more variables: gdp <dbl>, continent <fct>,
#   region <fct>
```

2. Encode that data into visual representations

```
> sub_gap  
# A tibble: 185 × 9  
  country      year infant_mortality life_expectancy fertility population  
  <fct>     <int>           <dbl>            <dbl>        <dbl>       <dbl>  
1 Albania      2010            14.8             77.2        1.74      2901883  
2 Algeria      2010            23.5             76          2.82      36036159  
3 Angola       2010           110.              57.6        6.22      21219954  
4 Antigua and ... 2010            7.7             75.8        2.13      87233  
5 Argentina    2010            13               75.8        2.22      41222875  
6 Armenia      2010            16.1             73          1.55      2963496  
7 Aruba         2010            NA              75.1        1.7       101597  
8 Australia     2010            4.1              82          1.89      22162863  
9 Austria       2010            3.6              80.5        1.44      8391986  
10 Azerbaijan   2010            33.9             70.1        1.97      9099893  
# ... with 175 more rows, and 3 more variables: gdp <dbl>, continent <fct>,  
#   region <fct>
```

Rows are items

```
> sub_gap
# A tibble: 185 × 9
  country      year infant_mortality life_expectancy fertility population
  <fct>     <int>        <dbl>            <dbl>       <dbl>        <dbl>
1 Albania    2010         14.8             77.2       1.74      2901883
2 Algeria   2010         23.5             76          2.82      36036159
3 Angola     2010        110.              57.6       6.22      21219954
4 Antigua and ... 2010         7.7              75.8       2.13      87233
5 Argentina  2010         13               75.8       2.22      41222875
6 Armenia    2010        16.1              73          1.55      2963496
7 Aruba      2010         NA               75.1       1.7       101597
8 Australia   2010         4.1              82          1.89      22162863
9 Austria    2010         3.6              80.5       1.44      8391986
10 Azerbaijan 2010        33.9             70.1       1.97      9099893
# ... with 175 more rows, and 3 more variables: gdp <dbl>, continent <fct>,
#   region <fct>
```

(Data) Variables are (Aesthetic) Attributes

```
> sub_gap
# A tibble: 185 × 9
  country      year infant_mortality life_expectancy fertility population
  <fct>     <int>        <dbl>            <dbl>        <dbl>        <dbl>
1 Albania      2010         14.8            77.2        1.74    2901883
2 Algeria      2010         23.5            76          2.82    36036159
3 Angola       2010        110.             57.6        6.22    21219954
4 Antigua and ... 2010         7.7            75.8        2.13     87233
5 Argentina    2010         13              75.8        2.22    41222875
6 Armenia       2010        16.1            73          1.55    2963496
7 Aruba         2010         NA              75.1        1.7     101597
8 Australia     2010         4.1            82          1.89    22162863
9 Austria        2010         3.6            80.5        1.44    8391986
10 Azerbaijan   2010        33.9            70.1        1.97    9099893
# ... with 175 more rows, and 3 more variables: gdp <dbl>, continent <fct>,
#   region <fct>
```

Cells are values

```
> sub_gap  
# A tibble: 185 × 9  
  country      year infant_mortality life_expectancy fertility population  
  <fct>     <int>        <dbl>            <dbl>       <dbl>        <dbl>  
1 Albania      2010         14.8             77.2       1.74      2901883  
2 Algeria      2010         23.5             76          2.82      36036159  
3 Angola       2010        110.              57.6       6.22      21219954  
4 Antigua and ... 2010         7.7              75.8       2.13      87233  
5 Argentina    2010         13                75.8       2.22      41222875  
6 Armenia      2010        16.1              73          1.55      2963496  
7 Aruba        2010         NA                75.1       1.7       101597  
8 Australia    2010         4.1               82          1.89      22162863  
9 Austria       2010         3.6              80.5       1.44      8391986  
10 Azerbaijan   2010        33.9              70.1       1.97      9099893  
# ... with 175 more rows, and 3 more variables: gdp <dbl>, continent <fct>,  
#   region <fct>
```

What are we trying to encode?

```
> sub_gap  
# A tibble: 185 × 9  
  country      year infant_mortality life_expectancy fertility population  
  <fct>     <int>        <dbl>            <dbl>       <dbl>        <dbl>  
1 Albania    2010         14.8             77.2       1.74      2901883  
2 Algeria   2010         23.5              76          2.82      36036159  
3 Angola     2010        110.              57.6       6.22      21219954  
4 Antigua and ... 2010         7.7              75.8       2.13      87233  
5 Argentina  2010         13                75.8       2.22      41222875  
6 Armenia    2010        16.1              73          1.55      2963496  
7 Aruba      2010         NA                75.1       1.7       101597  
8 Australia  2010         4.1               82          1.89      22162863  
9 Austria    2010         3.6               80.5       1.44      8391986  
10 Azerbaijan 2010        33.9              70.1       1.97      9099893  
# ... with 175 more rows, and 3 more variables: gdp <dbl>, continent <fct>,  
#   region <fct>
```

How has infant mortality changed?

```
> sub_gap
# A tibble: 185 × 9
  country      year infant_mortality life_expectancy fertility population
  <fct>     <int>        <dbl>            <dbl>        <dbl>        <dbl>
1 Albania    2010         14.8             77.2        1.74     2901883
2 Algeria   2010         23.5             76          2.82     36036159
3 Angola     2010        110.              57.6        6.22     21219954
4 Antigua and ... 2010        7.7              75.8        2.13      87233
5 Argentina  2010         13               75.8        2.22     41222875
6 Armenia    2010        16.1              73          1.55     2963496
7 Aruba      2010         NA               75.1        1.7       101597
8 Australia  2010         4.1              82          1.89     22162863
9 Austria    2010         3.6              80.5        1.44     8391986
10 Azerbaijan 2010        33.9             70.1        1.97     9099893
# ... with 175 more rows, and 3 more variables: gdp <dbl>, continent <fct>,
#   region <fct>
```

What are we encoding? What is on the x/y?

```
> sub_gap  
# A tibble: 185 × 9  
  country      year infant_mortality life_expectancy fertility population  
  <fct>     <int>        <dbl>            <dbl>        <dbl>        <dbl>  
1 Albania    2010         14.8            77.2        1.74    2901883  
2 Algeria   2010         23.5            76          2.82    36036159  
3 Angola     2010        110.             57.6        6.22    21219954  
4 Antigua and ... 2010         7.7            75.8        2.13     87233  
5 Argentina  2010         13              75.8        2.22    41222875  
6 Armenia    2010        16.1             73          1.55    2963496  
7 Aruba      2010         NA              75.1        1.7      101597  
8 Australia  2010         4.1             82          1.89    22162863  
9 Austria    2010         3.6             80.5        1.44    8391986  
10 Azerbaijan 2010        33.9            70.1        1.97    9099893  
# ... with 175 more rows, and 3 more variables: gdp <dbl>, continent <fct>,  
#   region <fct>
```

Encoding using ggplot

?ggplot

In `ggplot()`, a plot is a *mapping* of data to the *aesthetic* attributes of geometric markings.

We make or *render* plots by layering on these mappings.

The screenshot shows the R documentation for the `ggplot` function. The title bar says "R: Create a new ggplot". Below it are three navigation icons: a left arrow, a right arrow, and a circular arrow. The main title is "ggplot {ggplot2}" and the subtitle is "Create a new ggplot". The "Description" section states that `ggplot()` initializes a ggplot object, used to declare input data frames and specify plot aesthetics. The "Usage" section shows the function signature: `ggplot(data = NULL, mapping = aes(), ..., envir = .GlobalEnv)`. The "Arguments" section includes two entries: "data" (the default dataset) and "mapping" (the aesthetic mappings). The "data" entry has a detailed description about its conversion to a data frame if not already one, and the "mapping" entry has a note about being a list of aesthetic mappings.

R: Create a new ggplot

ggplot {ggplot2} R Documentation

Create a new ggplot

Description

`ggplot()` initializes a ggplot object. It can be used to declare the input data frame for a graphic and to specify the set of plot aesthetics intended to be common throughout all subsequent layers unless specifically overridden.

Usage

```
ggplot(data = NULL, mapping = aes(), ..., envir = .GlobalEnv)
```

Arguments

<code>data</code>	Default dataset to use for plot. If not already a <code>data.frame</code> , will be converted to one by <code>fortify()</code> . If not specified, must be supplied in each layer added to the plot.
<code>mapping</code>	Default list of aesthetic mappings to use

3. Render the visual representation

Creating the base layer

```
ggplot(data = sub_gap,  
       mapping = aes(x = year, y = infant_mortality))
```

Adding markings (geometries)

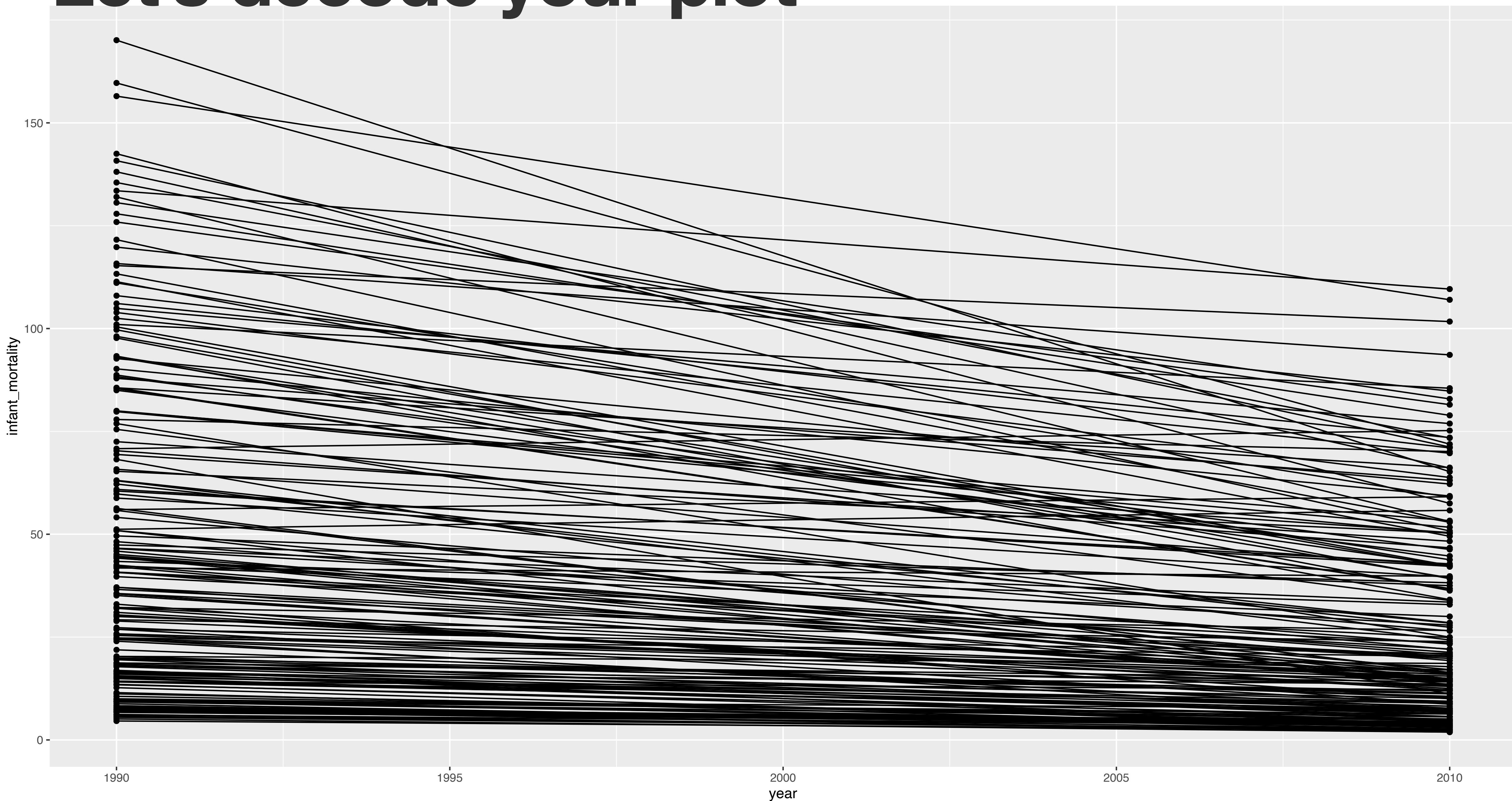
```
ggplot(data = sub_gap,  
       mapping = aes(x = year, y = infant_mortality)) +  
  geom_point()
```

Layering on more information

```
ggplot(data = sub_gap,  
       mapping = aes(x = year,  
                      y = infant_mortality,  
                      group = country)) +  
  geom_point() +  
  geom_line()
```

That was it!

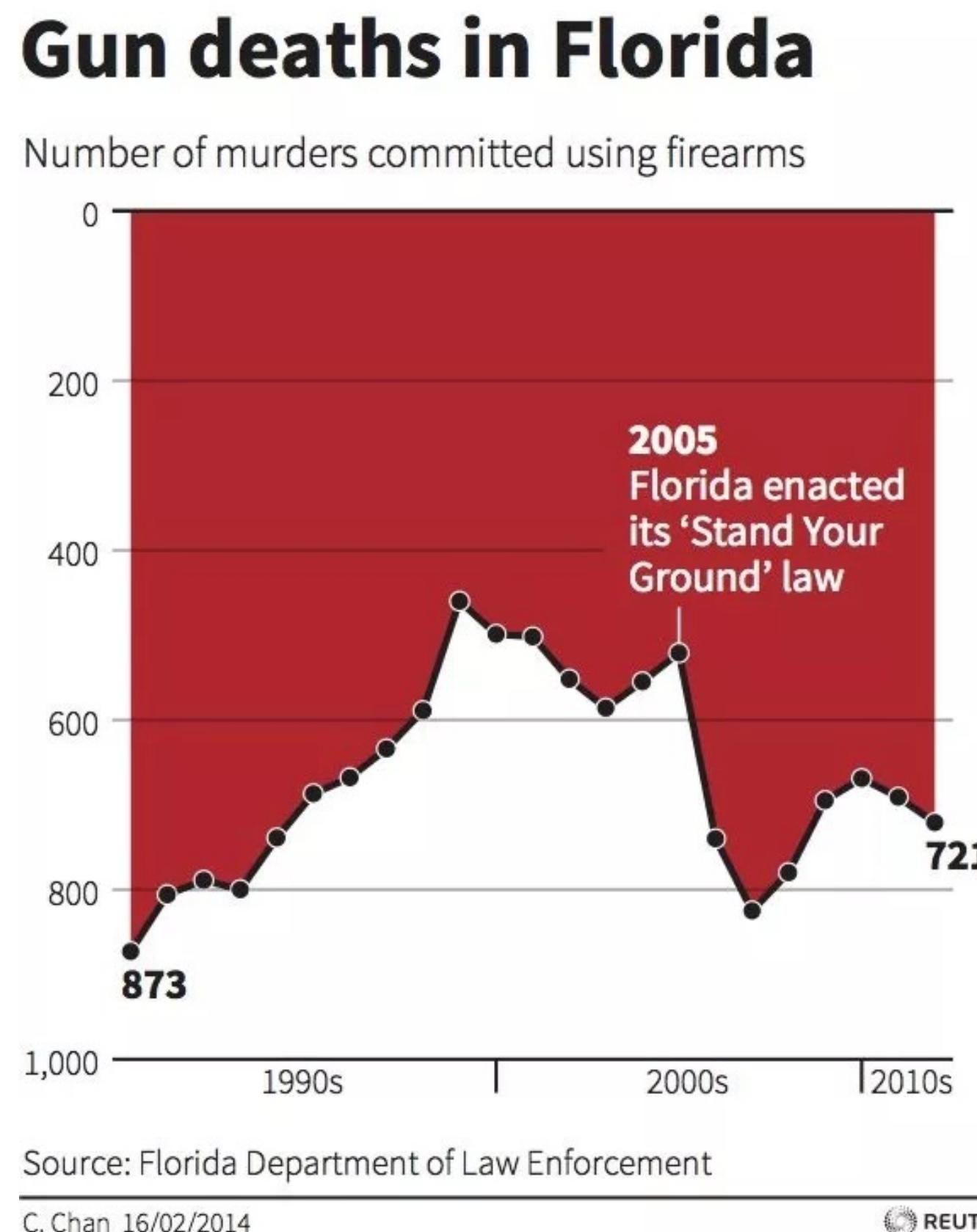
Let's decode your plot



Principles of good design

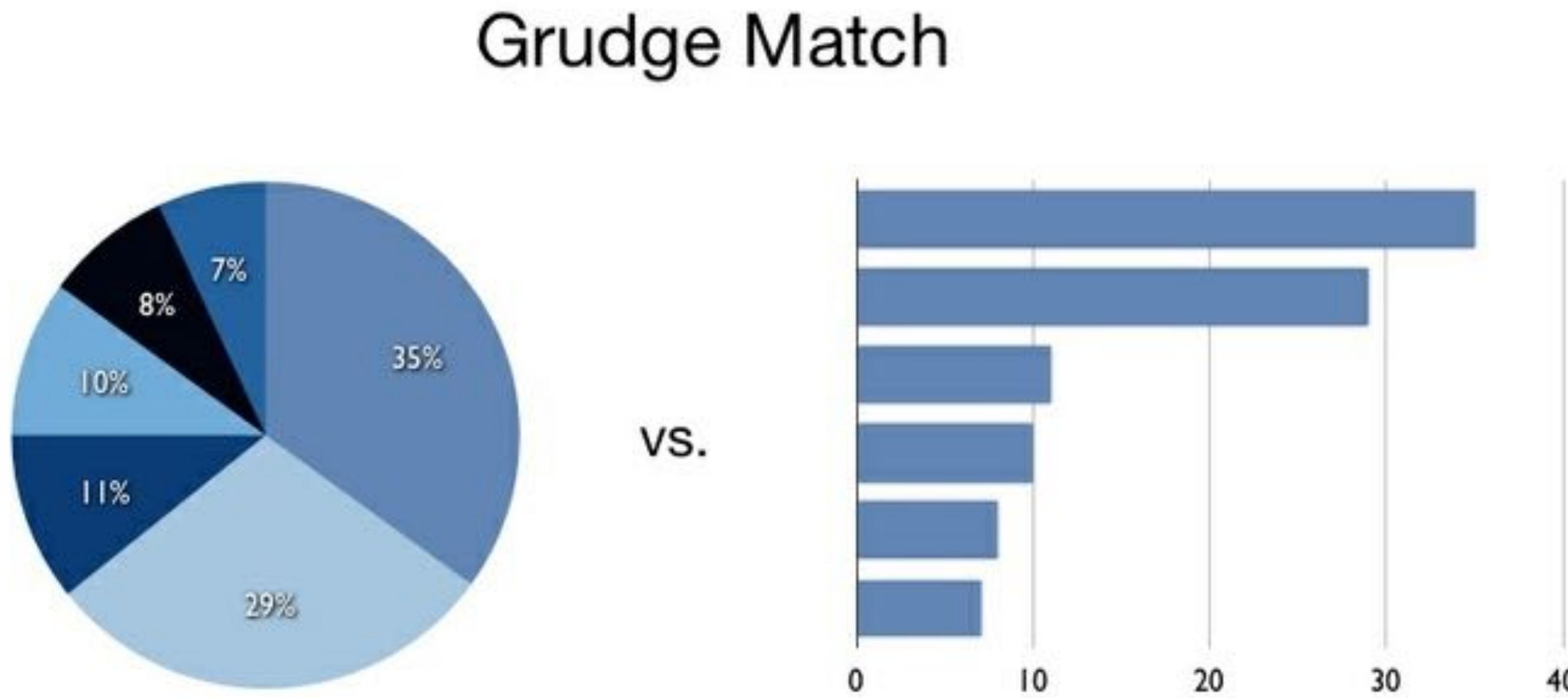
What makes a *bad* plot?

1. Adhere to common standards (e.g., do not flip your axes)



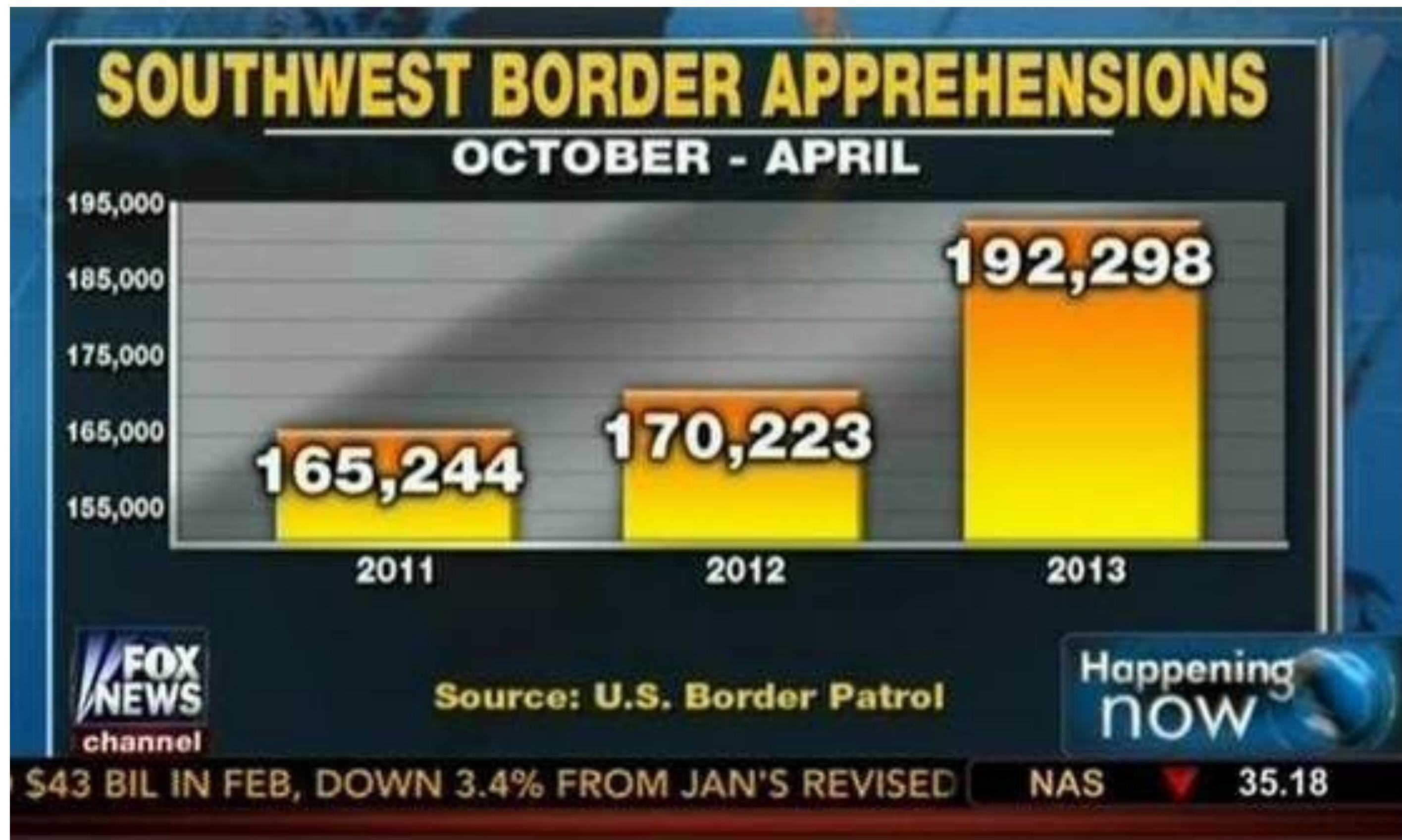
What makes a *bad* plot?

1. Adhere to common standards (e.g., do not flip your axes)
2. Don't use area to compare quantities



What makes a *bad* plot?

1. Adhere to common standards (e.g., do not flip your axes)
2. Don't use area to compare quantities
3. Don't truncate axes (especially on barcharts) – use points if appropriate



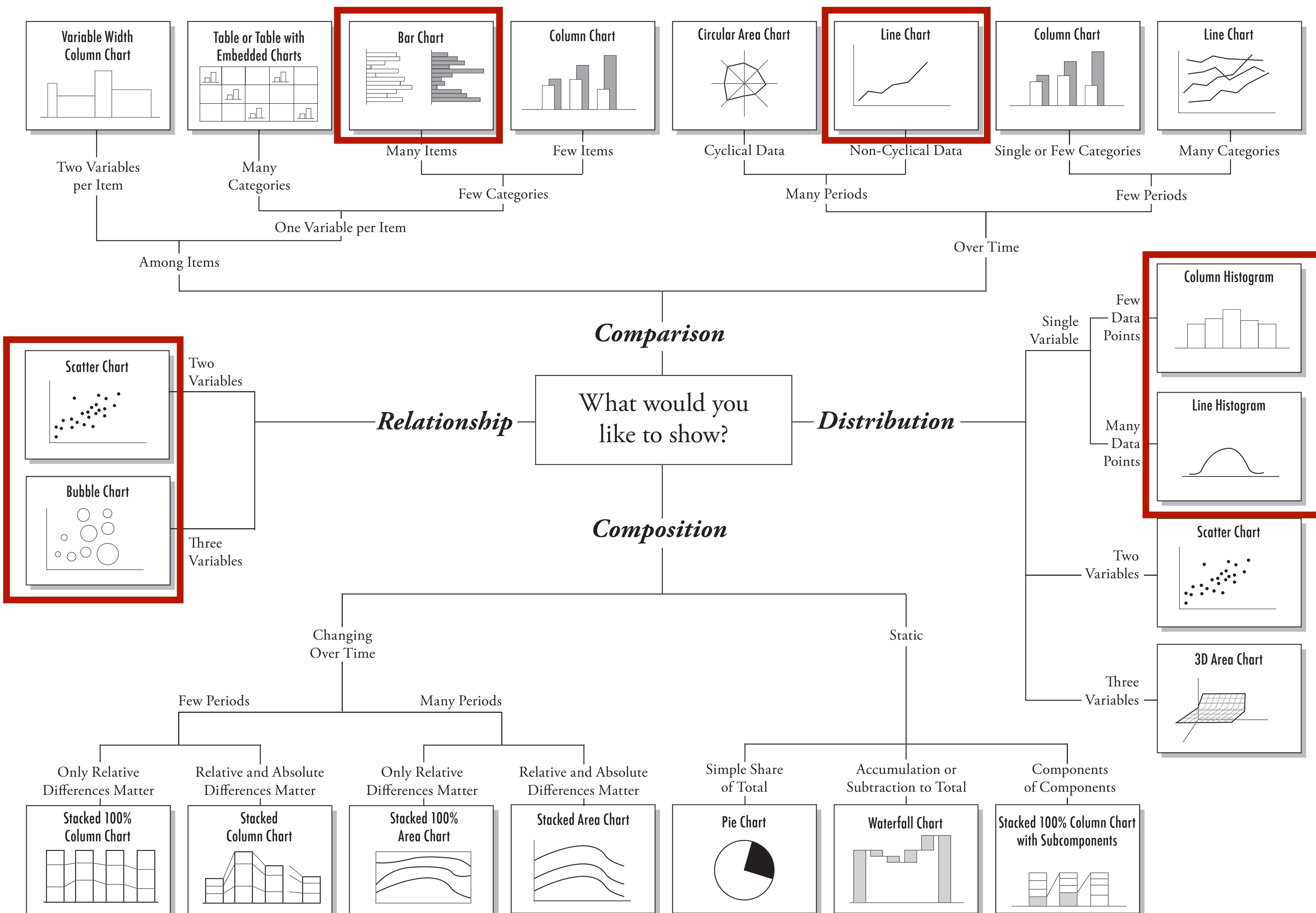
What makes a *good* plot?

1. Accurately conveys the intended message
2. Descriptive axis labels and title
3. Uses aesthetic mappings sparingly
(e.g., only use shapes if they serve a purpose)
4. Minimize clutter
5. Colors should be meaningful (and friendly to viewers with color vision deficiency)
6. Legends or annotations as appropriate
7. Plots should more or less be self-contained
(i.e., don't make the reader read the entire paper to understand it)
8. Rarely use 3D plots
9. Never use textures

(Many more but this will get you 90% of the way there.)

Live coding time
(Let's make the plot better)

Chart Suggestions—A Thought-Starter



Note: There are many (many many) more charts than this, but these are the most common.

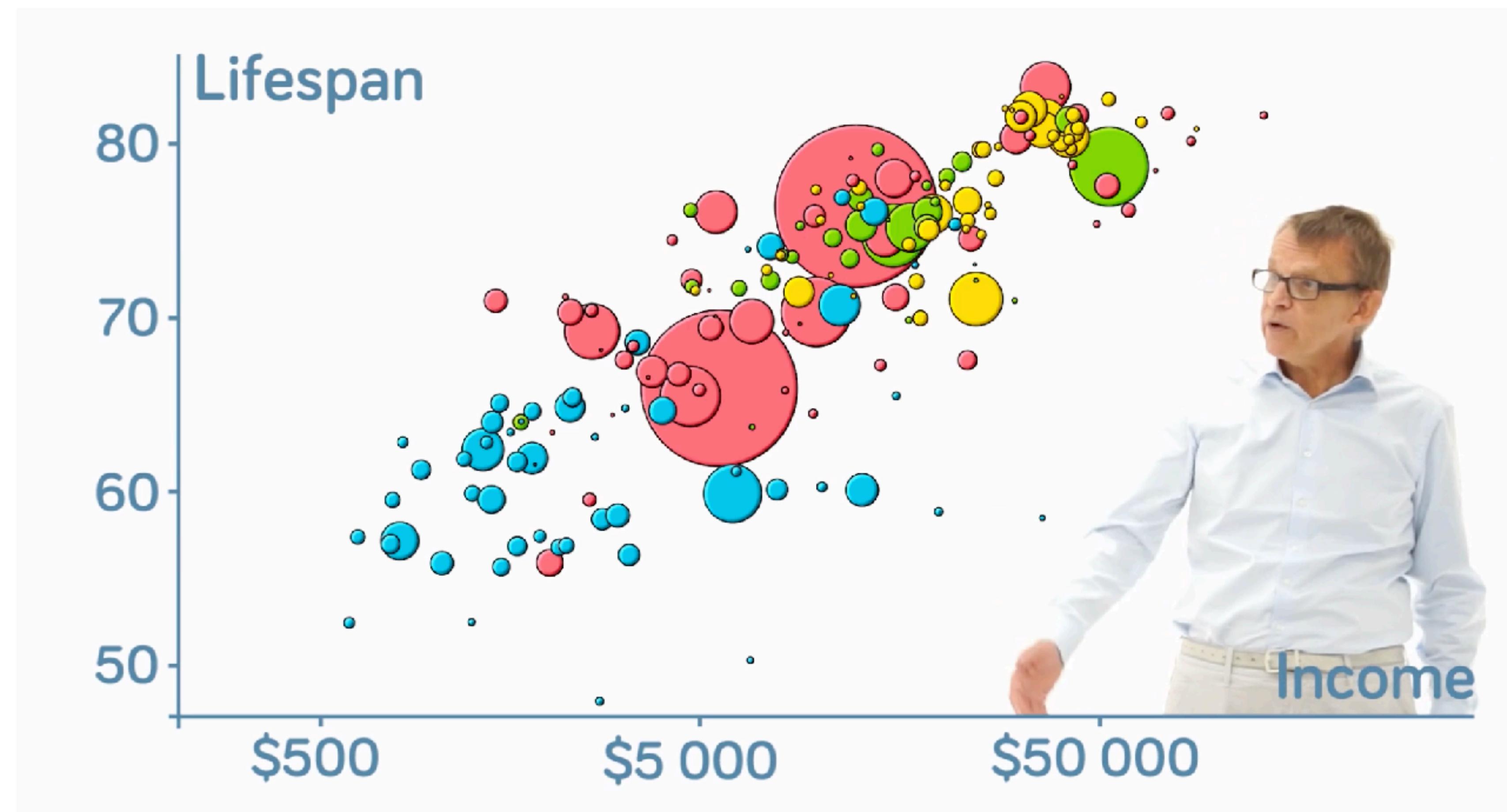
Exercises

1. What is the distribution? Density and histograms
2. What is the change (over time or space)? Line chart
3. How do things compare? Bar chart
4. How are things related? Scatterplot

Bonus: Let's replicate famous plots (or a couple of them)!

Can we replicate this plot?

How Does Income Relate to Life Expectancy?



Can we replicate this plot?

Measles

