Part 6 - Section 1: Introduction to data visualization

Dr. Nguyen Quang Huy

May 16, 2020

Introduction to data visualization

Data visualization is the graphical representation of data. By using visual charts, graphs, and maps, data visualization provides an accessible way to understand trends, outliers, and patterns in data.

- For most human brains, it is difficult to extract information from looing at the numbers, characters,...
- However, we can quickly identify red from blue, square from circle, ...

Introduction to data visualization

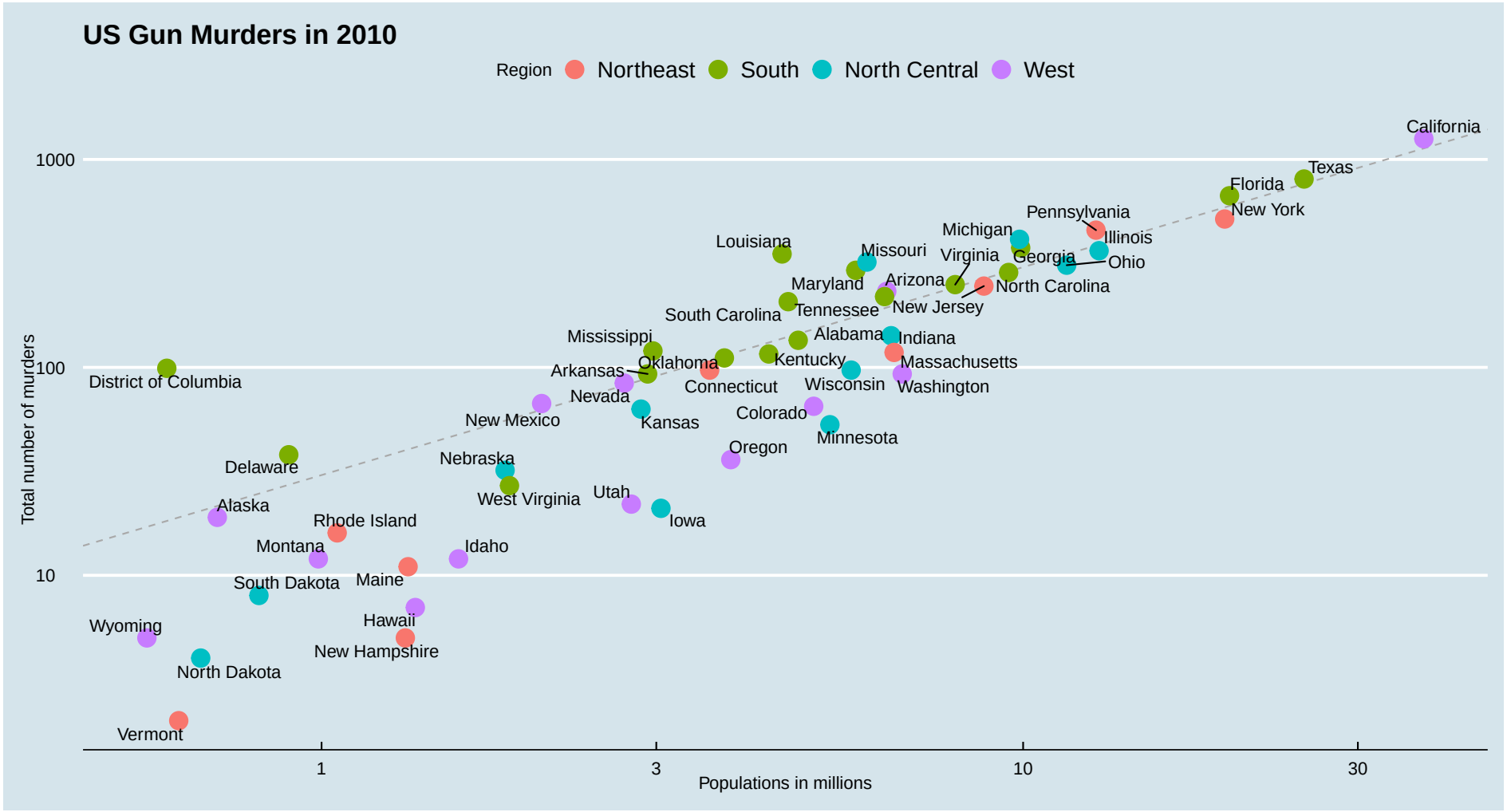
It is rarely useful when looking at the numbers, character strings from a dataset. How much information you get when look at **murders** dataset?

```
library(dslabs)
murders
```

```
##                   state abb        region population total
## 1               Alabama  AL         South    4779736   135
## 2                Alaska  AK          West     710231    19
## 3               Arizona  AZ          West    6392017   232
## 4              Arkansas  AR         South    2915918    93
## 5            California  CA          West   37253956  1257
## 6              Colorado  CO          West    5029196    65
## 7           Connecticut  CT     Northeast    3574097    97
## 8              Delaware  DE         South     897934    38
## 9  District of Columbia  DC         South     601723    99
## 10              Florida  FL         South   19687653   669
```

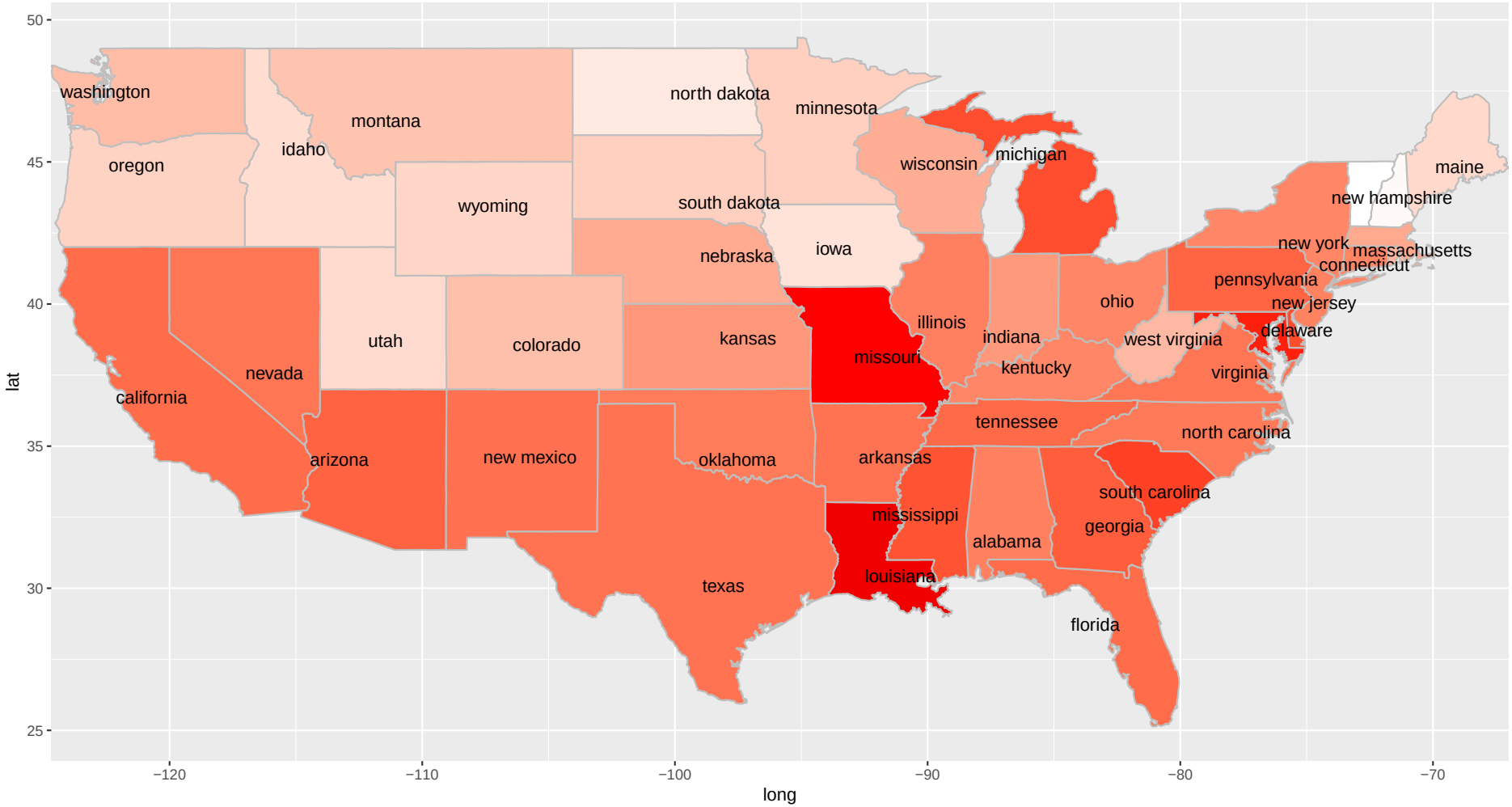
Introduction to data visualization

However, as people say, *"a picture is worth a thousand words"*. We have useful information from examining this plot



Introduction to data visualization

We can combine **murders** data with **map** data



Introduction to data visualization

Why data visualization is important?

- Make data easier to understand and remember
- Discover unknown facts, outliers and trends
- Visualize relationships and patterns quickly
- Ask better questions and make better decisions

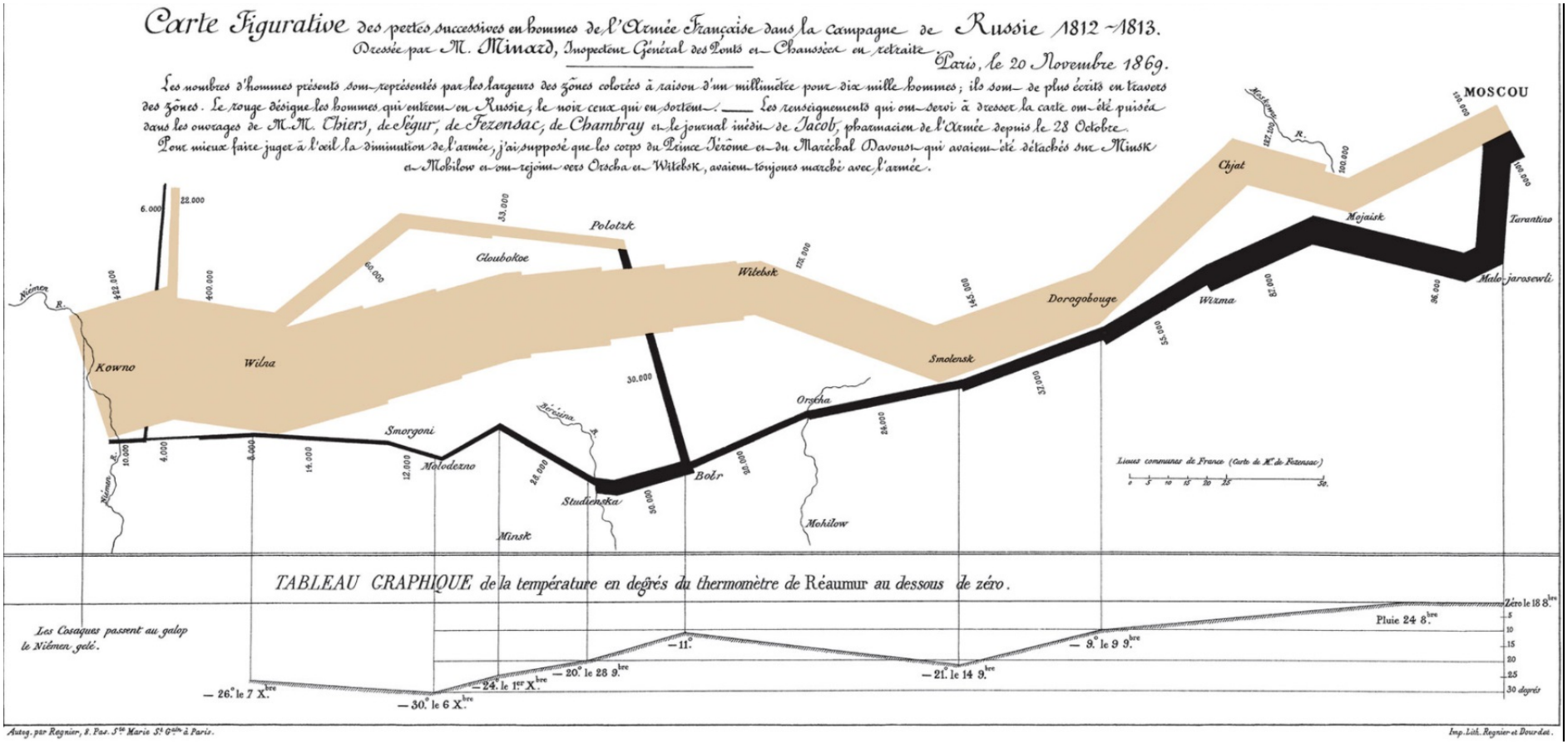
What makes a good data visualization?

- Step 1. Clean data (is ready to visualize)
- Step 2. Pick the right chart
- Step 3. Design and customize your visualization
- Step 4. Publish, share and communicate

Remember, simplicity is the key.

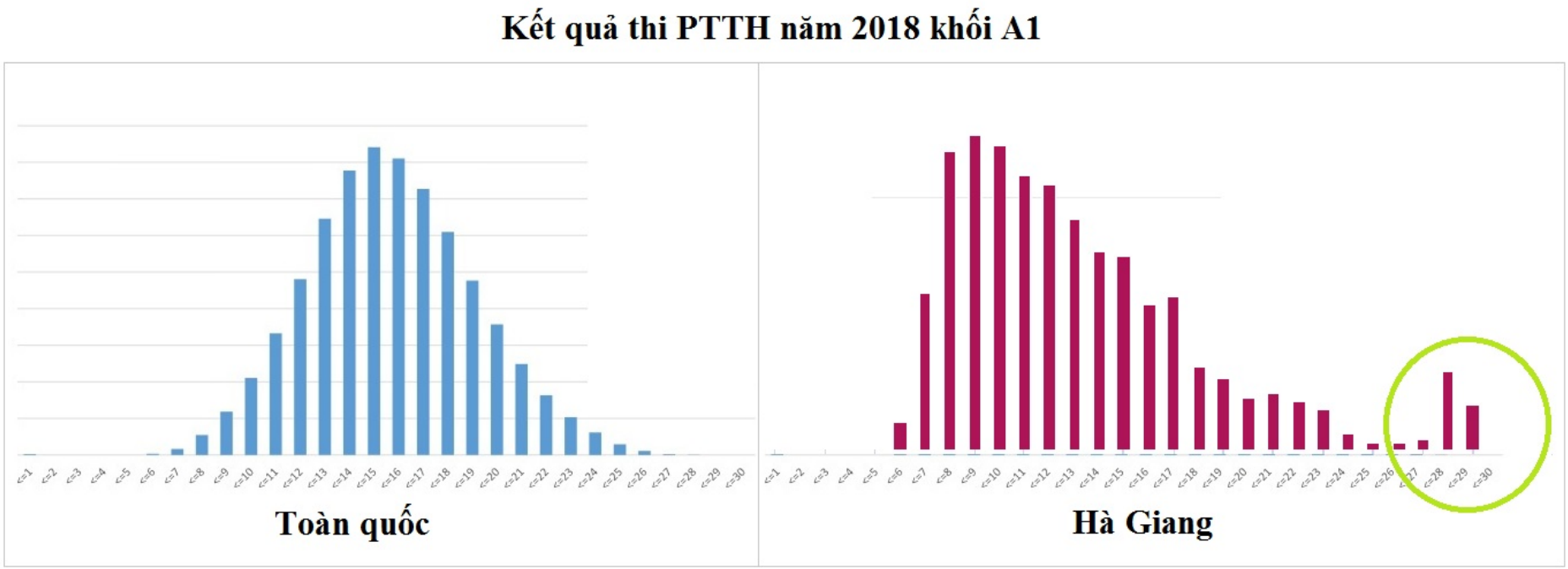
Introduction to data visualization

Storytelling in the work of Charles Joseph Minard (1780-1871) about Napoleon's Russian campaign of 1812



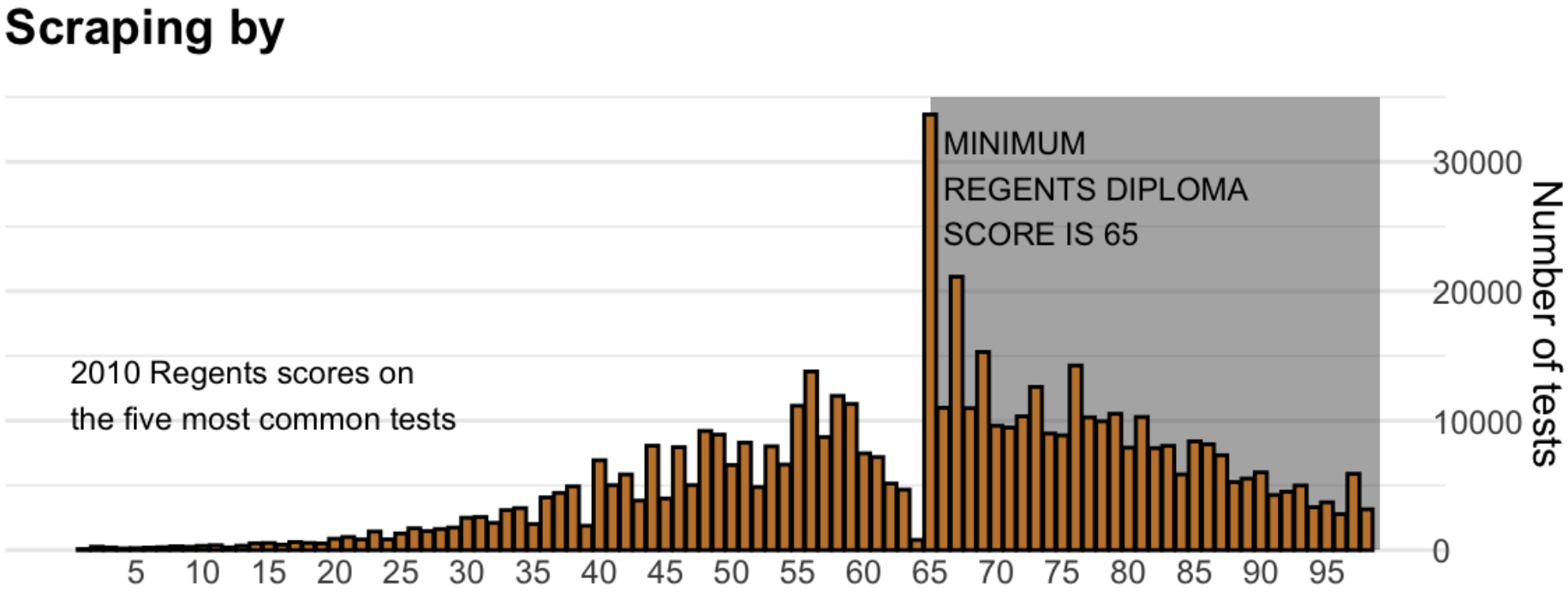
Introduction to data visualization

"The greatest value of a picture is when it forces us to notice what we never expected to see"



Introduction to data visualization

What is the problem with the NYC Regents Exam in 2010 where you need a score of 65 to pass?



Introduction to ggplot2

- *ggplot2* is an R package for producing statistical graphics with a deep underlying grammar.
- The grammar, based on the grammar of graphics (Wilkinson, 2005), is made up of a set of independent components that can be composed in many different ways.
- The grammar of graphic tells us that a statistical graphic is a mapping from data to aesthetic attributes (color, shape, size, ...) of geometric objects (points, lines, bars, ...)
- *ggplot2* is powerful because users are not limited to a set of pre-specified graphics, but they can create new graphics that are appropriate for their problem.

Introduction to ggplot2

Advantages of *ggplot2*

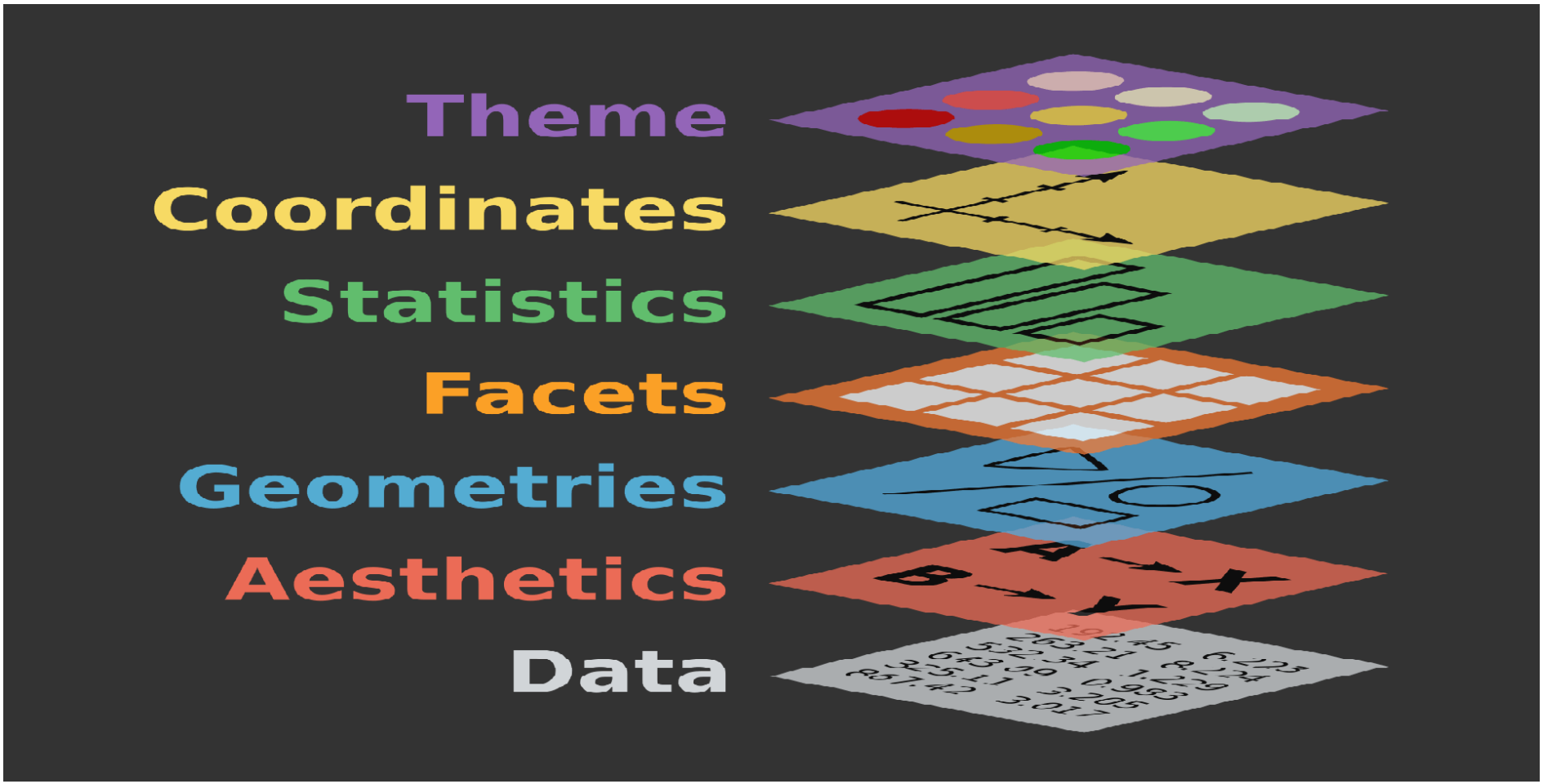
- Users are not limited to a set of pre-specified graphics. You can build graphics that precisely tells your story.

Disadvantages of *ggplot2*

- *ggplot2* is useful only when users have some basic knowledge in R.
- *ggplot2* doesn't suggest what graphics you should use to answer the questions you are interested in.
- *ggplot2* is not designed to create dynamic and interactive graphics i.e. *ggplot2* is suitable with static data.

Grammar of graphics

The grammar of graphics describes the deep features that underlie all statistical graphics:

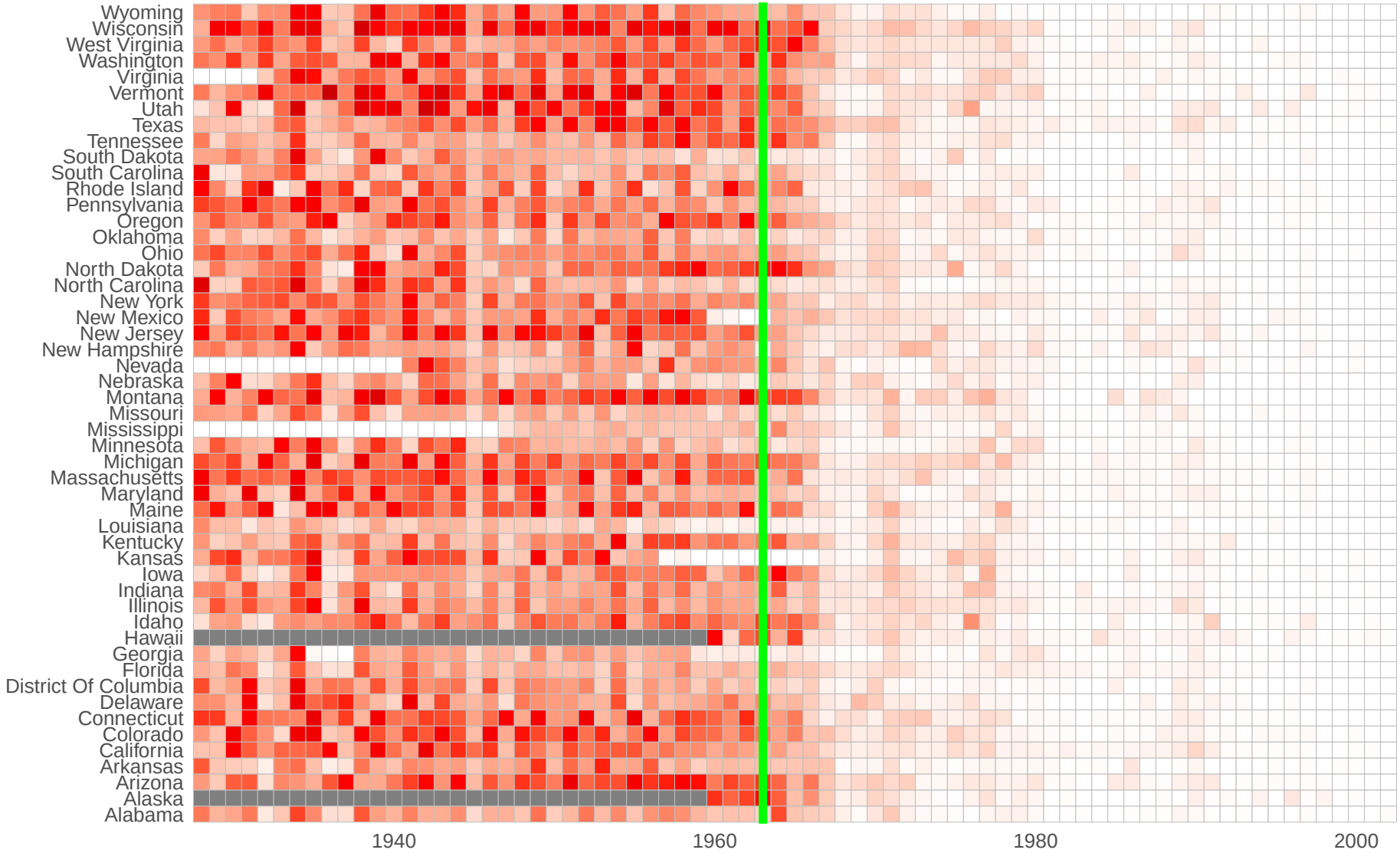


Grammar of graphics

- 1. **Data** that you want to visualise.
- 2. **Aesthetic mappings (aes)** describing how variables in the data are mapped to aesthetic attributes.
- 3. **Geometric objects (geoms)** represent what you actually see on the plot: points, lines, polygons, etc.
- 4. **A faceting** describes how to break up the data into subsets.
- 5. **Statistical transformations (stats)** summarise data in many useful ways.
- 6. **A coordinate system** describes how data coordinates are mapped to the plane of the graphic.
- 7. **A theme** controls the finer points of display, like the font size and background colour.

Introduction to data visualization

Measles vacinne was licensed in 1963 in the United States



Introduction to data visualization

```
p1<-us_contagious_diseases%>%filter(disease=="Measles")%>%
  mutate(rate=count*1000/population)%>%
  ggplot(aes(year,state,fill=rate))+geom_tile(color="grey")
p1+scale_fill_gradientn(colors = c(rgb(1,1,1),rgb(1,0,0),
                        rgb(0.8,0,0)),trans = "sqrt")+
  geom_vline(xintercept=1963, col = "green", size=2) +
  scale_x_continuous(expand=c(0,0))+
  theme_minimal()+
  theme(panel.grid = element_blank(),
        legend.position="bottom",
        text = element_text(size = 12))+
  xlab(label="")+
  ylab(label="")
```

End of Section 1