



Visualization: Grammar of Graphics

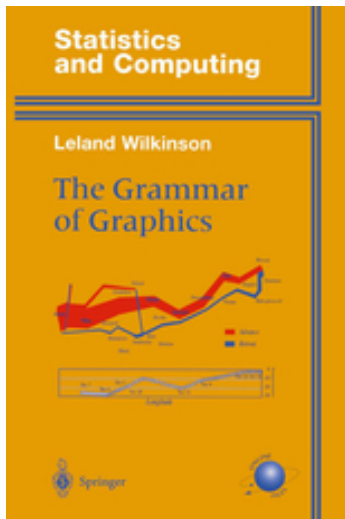
Data Science: Jordan Boyd-Graber
University of Maryland

FEBRUARY 11, 2018

Understanding Data

- After you've cleaned (wrangled) data
- Need to tell a story with the data
- Often a first step before you can build a model
- Necessary afterward to explain model works

Grammar of Graphics



- Don't focus on pixels
- Focus on data
- Easy combination / switches

Tiny Dataset

<i>x</i>	<i>y</i>	Shape
2	4	a
1	1	a
4	15	b
9	80	b

Data Tell a Story

x	y	Shape
25	11	circle
0	0	circle
75	53	square
200	300	square

What visualization helps you tell your story?

Components of a Plot

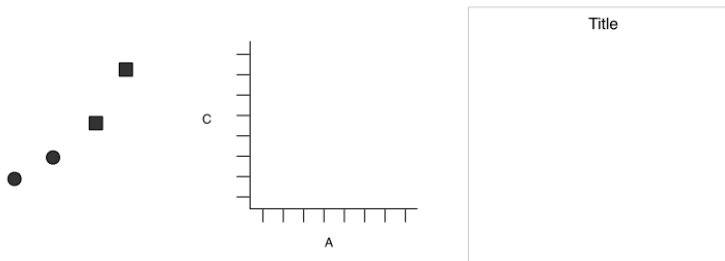
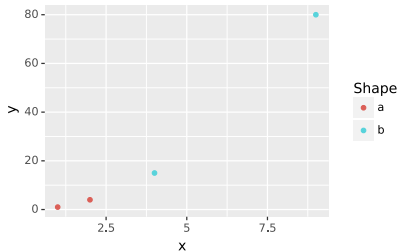


Figure 1. Graphics objects produced by (from left to right): geometric objects, scales and coordinate system, plot annotations.

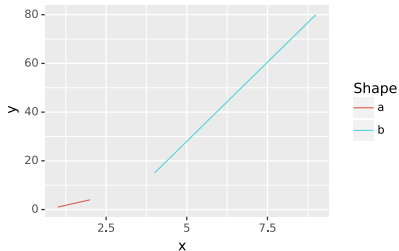
- Geom: How data turn into shapes
- Scales: Relative positioning
- Annotations: Text, explanations

Putting it Together



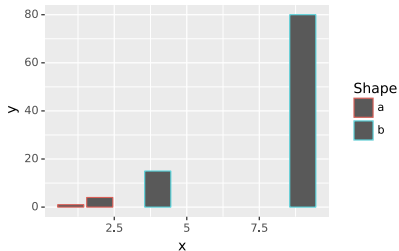
```
simple_point = (ggplot(demo,  
                      aes(color='Shape', y='y', x='x')) +  
               geom_point())  
simple_point.save("simple_point.pdf", scale=0.6,  
                 height=6, width=8)
```

Geometry Options: Line



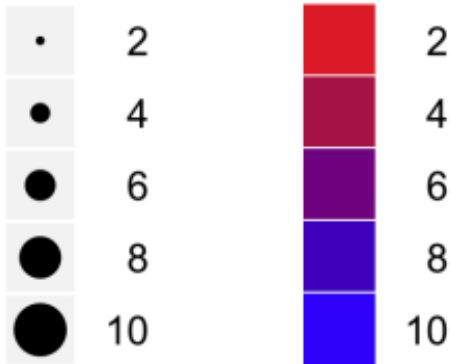
```
simple_point = (ggplot(demo,
                      aes(color='Shape', y='y', x='x')) +
               geom_line())
simple_point.save("simple_line.pdf", scale=0.6,
                 height=6, width=8)
```


Geometry Options: Bar



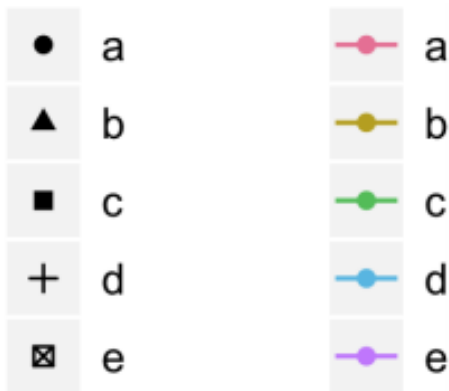
```
simple_point = (ggplot(demo,  
                      aes(color='Shape', y='y', x='x')) +  
               geom_bar(stat="identity"))  
simple_point.save("simple_bar.pdf", scale=0.6,  
                 height=6, width=8)
```

Aesthetic Options: Continuous Data



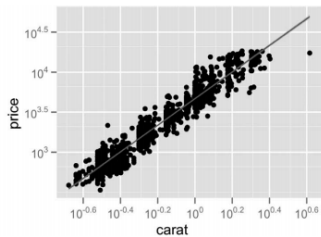
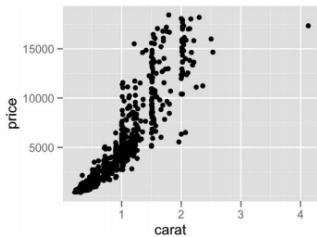
Size, color

Aesthetic Options: Discrete Data



Shape, color

Rescaling Data

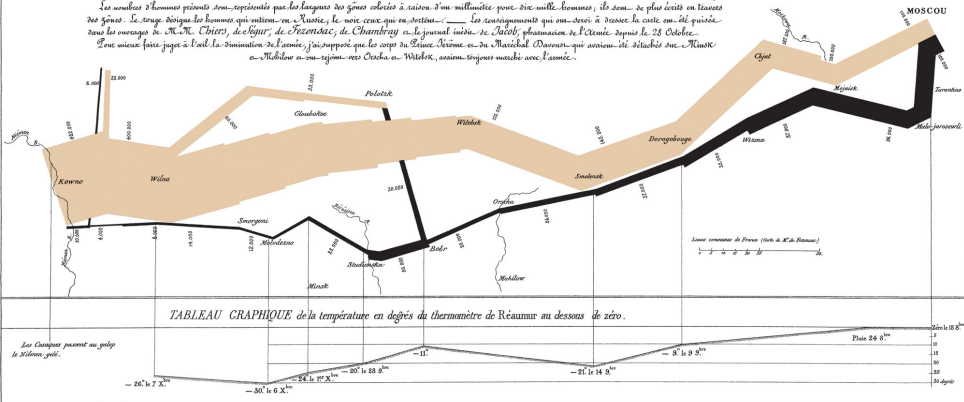


```
+ scale_x_log10()
```

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

dressée par M. Minard, Inspecteur Général des Ponts et Chaussées et extraite Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en lettres des zones. Le rouge désigne les hommes qui ont été en Russie; le noir ceux qui en sont restés. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Chéjuz, de Fessenet, de Chambray et le journal inédit de Sobol, pharmacien de l'Armée depuis le 28 Octobre. — Une seule fois j'ai vu à l'œil la diminution de l'armée; j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout qui avaient été détachés sur Minsk et Mielnois n'ont rejoint vers Oreska et Wilna, avaient toujours marché avec l'armée.



Edward Tufte on Charles Joseph Minard

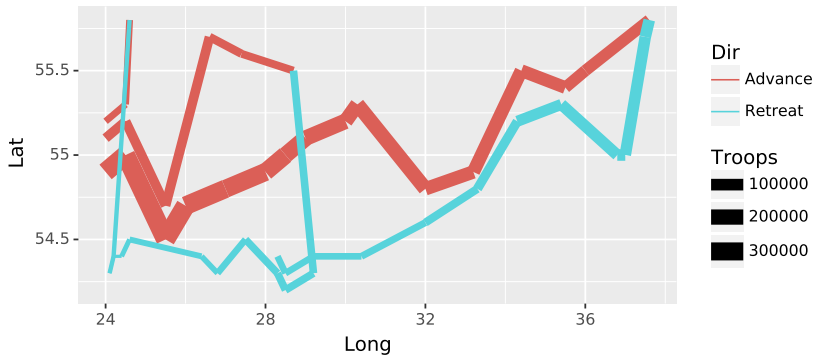
Data (Troops)

```
Long, Lat, Troops, Dir, Div
24.0, 54.9, 340000, Advance, 1
24.5, 55.0, 340000, Advance, 1
24.2, 54.4, 4000, Retreat, 2
24.1, 54.3, 4000, Retreat, 2
24.6, 55.8, 6000, Retreat, 3
24.2, 54.4, 6000, Retreat, 3
```

Cities

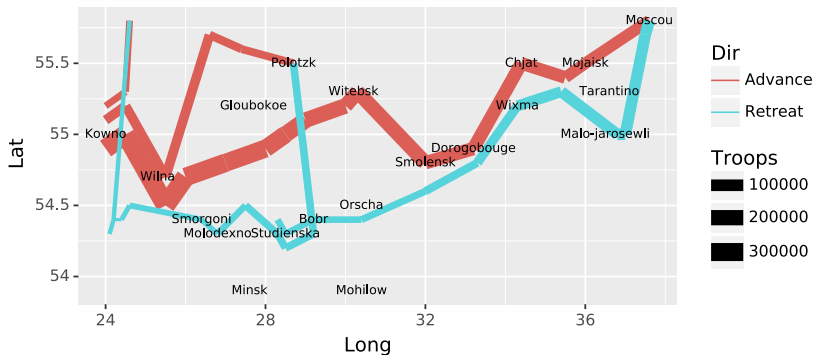
```
Long, Lat, City  
24.0, 55.0, Kowno  
25.3, 54.7, Wilna  
26.4, 54.4, Smorgoni  
26.8, 54.3, Molodexno  
27.7, 55.2, Gloubokoe  
27.6, 53.9, Minsk
```

Plot Troops



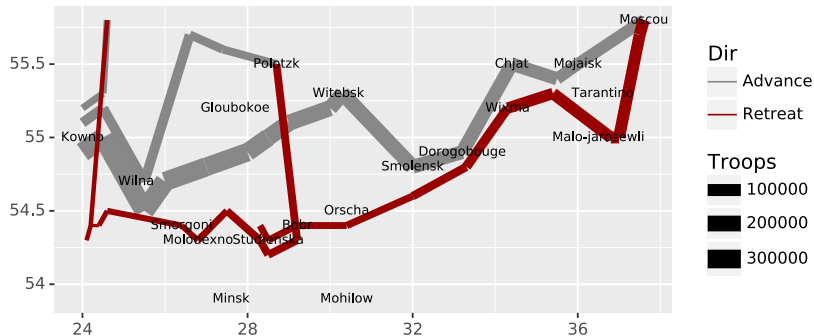
```
plot_troops = (ggplot(troops, aes('Long', 'Lat')) +  
  geom_path(aes(size = 'Troops',  
                color = 'Dir',  
                group = 'Div')))
```


Add Cities



```
both = plot_troops + geom_text(aes(label='City'),  
                               size=7, data=cities)
```

Make Prettier



```
polish = both + scale_color_manual(["#888888", "#990000"])
```

Practicalities

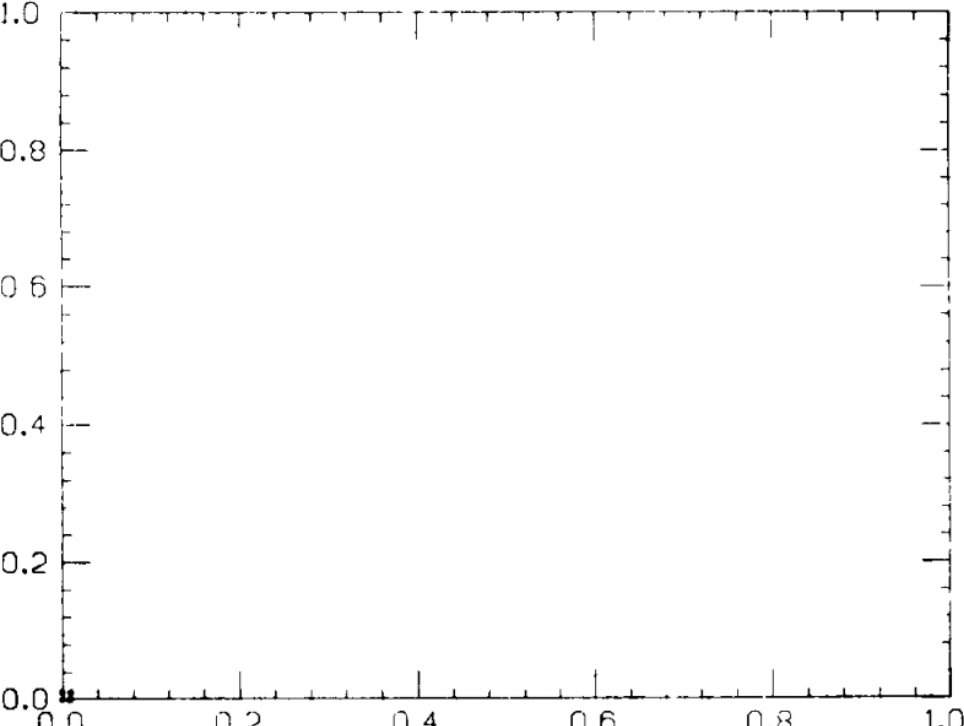
- Remember to have data in accessible directory
- Install plotnine (e.g., with pip)
- Keep track of how you generate every plot

Tips

- Avoid clutter
- Organize logically, not arbitrarily
- Encourage / enable comparisons
- Don't overload with variables
- Overplotting

Tips

- Avoid clutter
- Organize logically, not arbitrarily
- Encourage / enable comparisons
- Don't overload with variables
- Overplotting
- Examples of what not to do



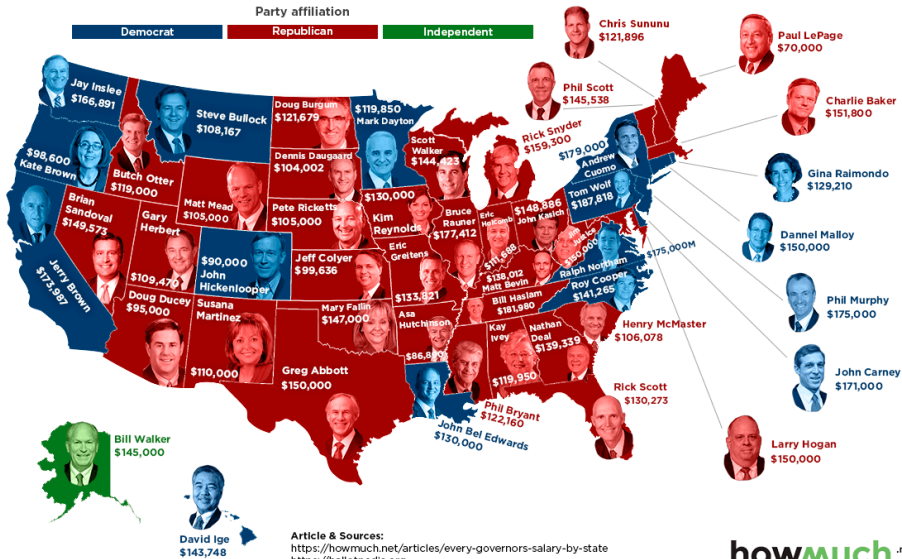
Every Governor's Salary by State

Party affiliation

Democrat

Republican

Independent



Article & Sources:
<https://howmuch.net/articles/every-governors-salary-by-state>
<https://ballotpedia.org>

howmuch.net

Reflect data

U.S. trade with China and Taiwan

(in millions of U S dollars)

3,000

2,000

1,000

U S exports
to China

U S imports
from China

1972 1974 1976 1978 1980

(in millions of U S dollars)

6,000

4,000

2,000

U S imports
from Taiwan

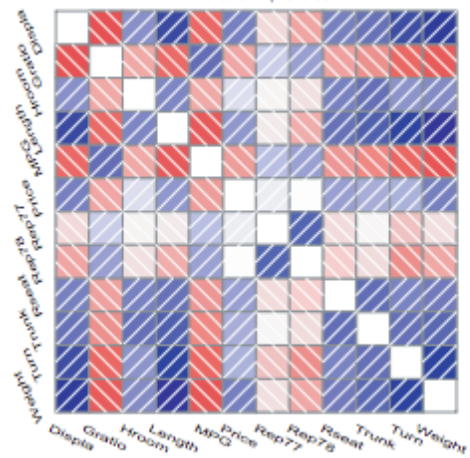
U S exports
to Taiwan

1970 1972 1974 1976 1978 1980

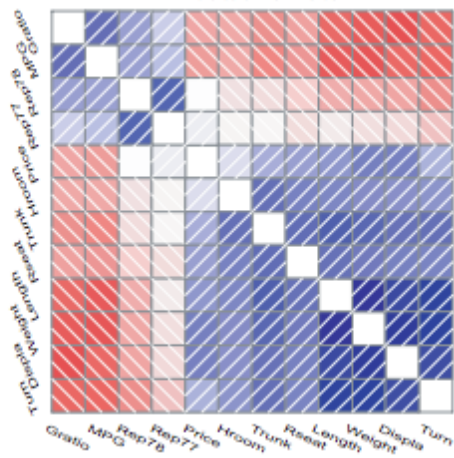
Source Department of Commerce

Enable comparisons

Auto data: Alpha order



Auto data: PC2/1 order



Order sensibly

Wrap Up

- Cleanup
- Explore
- Model
- Copy
- Explain