ggplot2 theory

Hadley Wickham

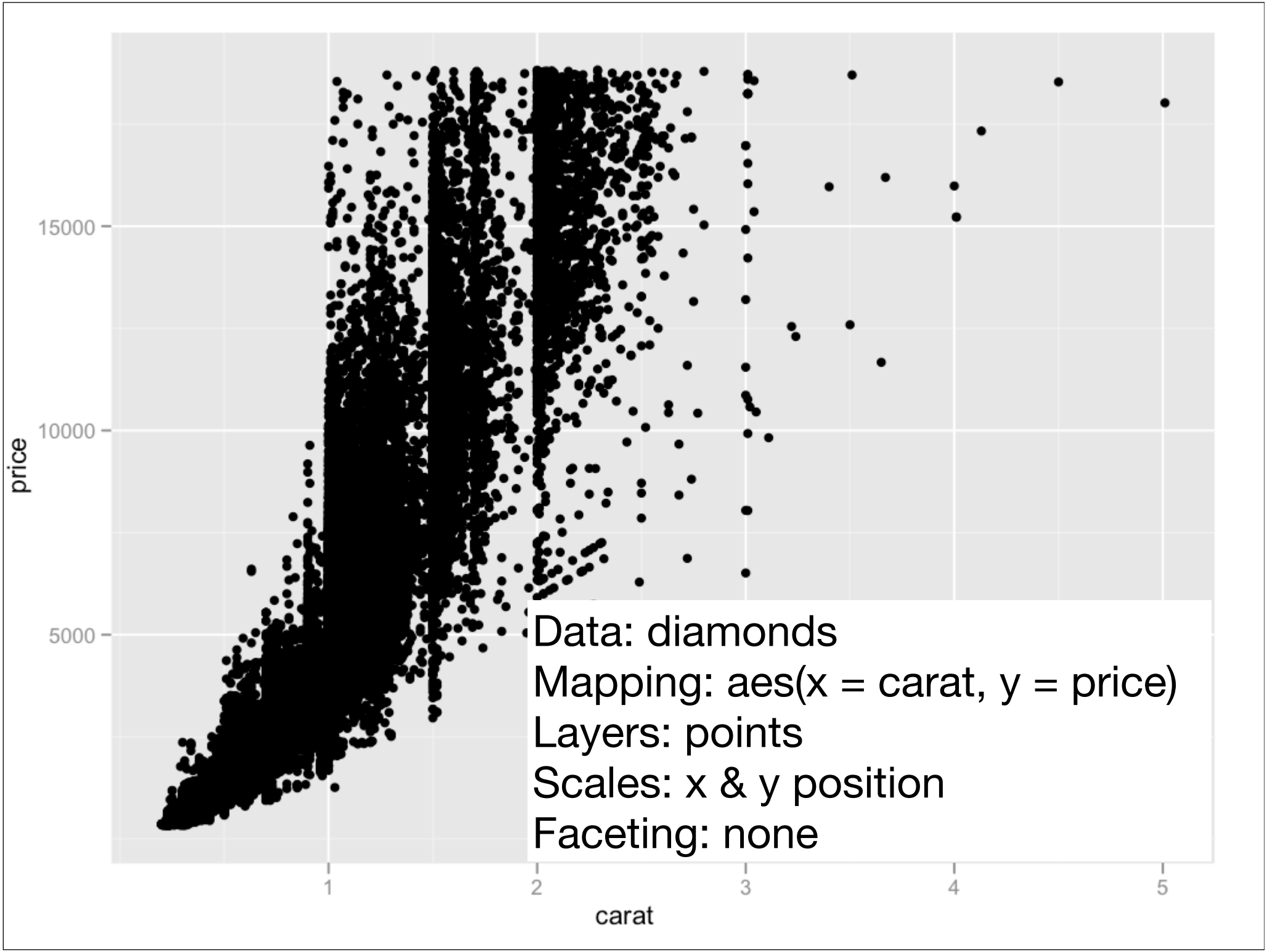
October 2009

1. What is a plot?
2. What is a layer?
3. Minard’s march
4. Critiquing a graphic

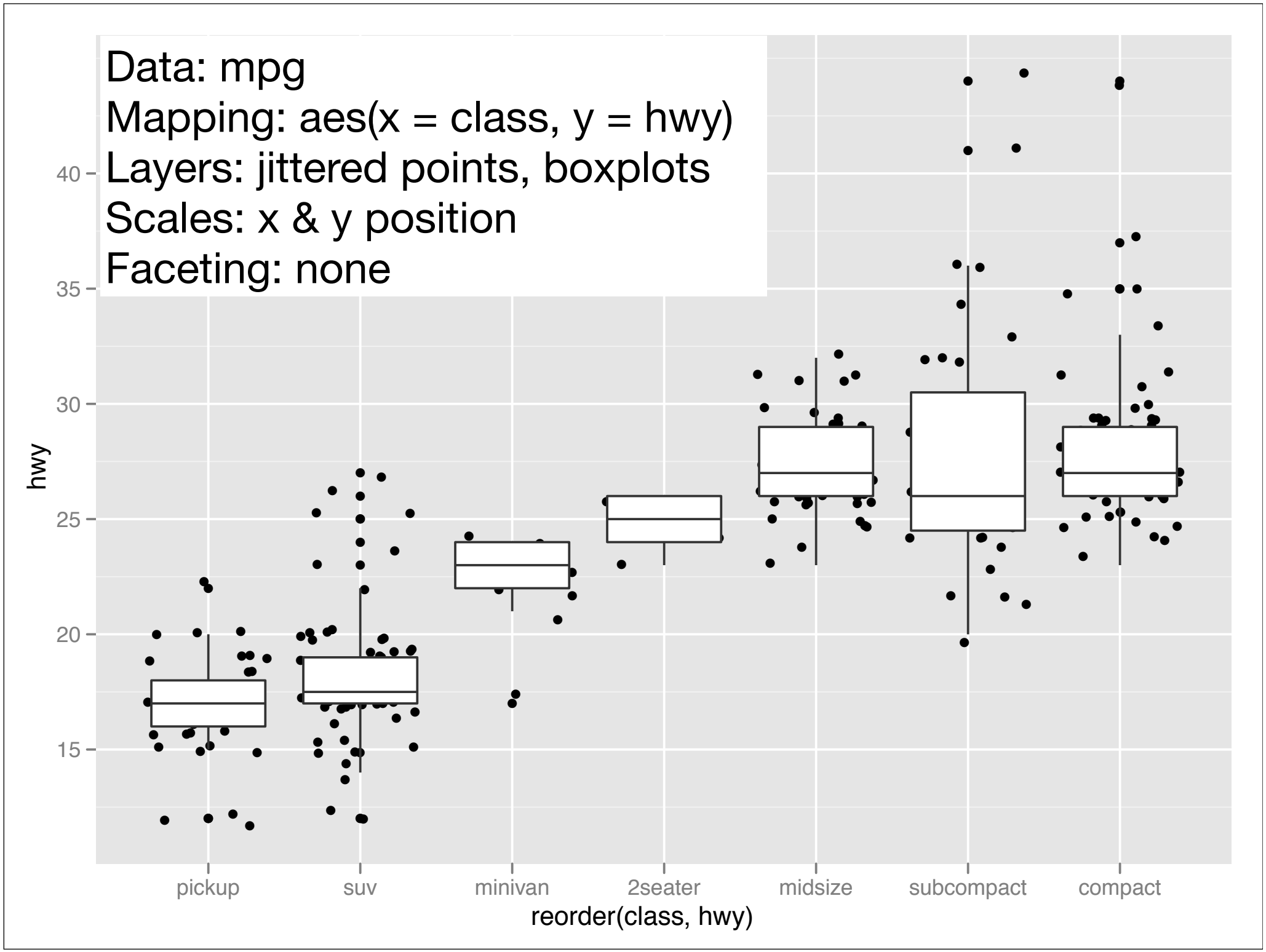
What is a plot?

A plot is composed of:

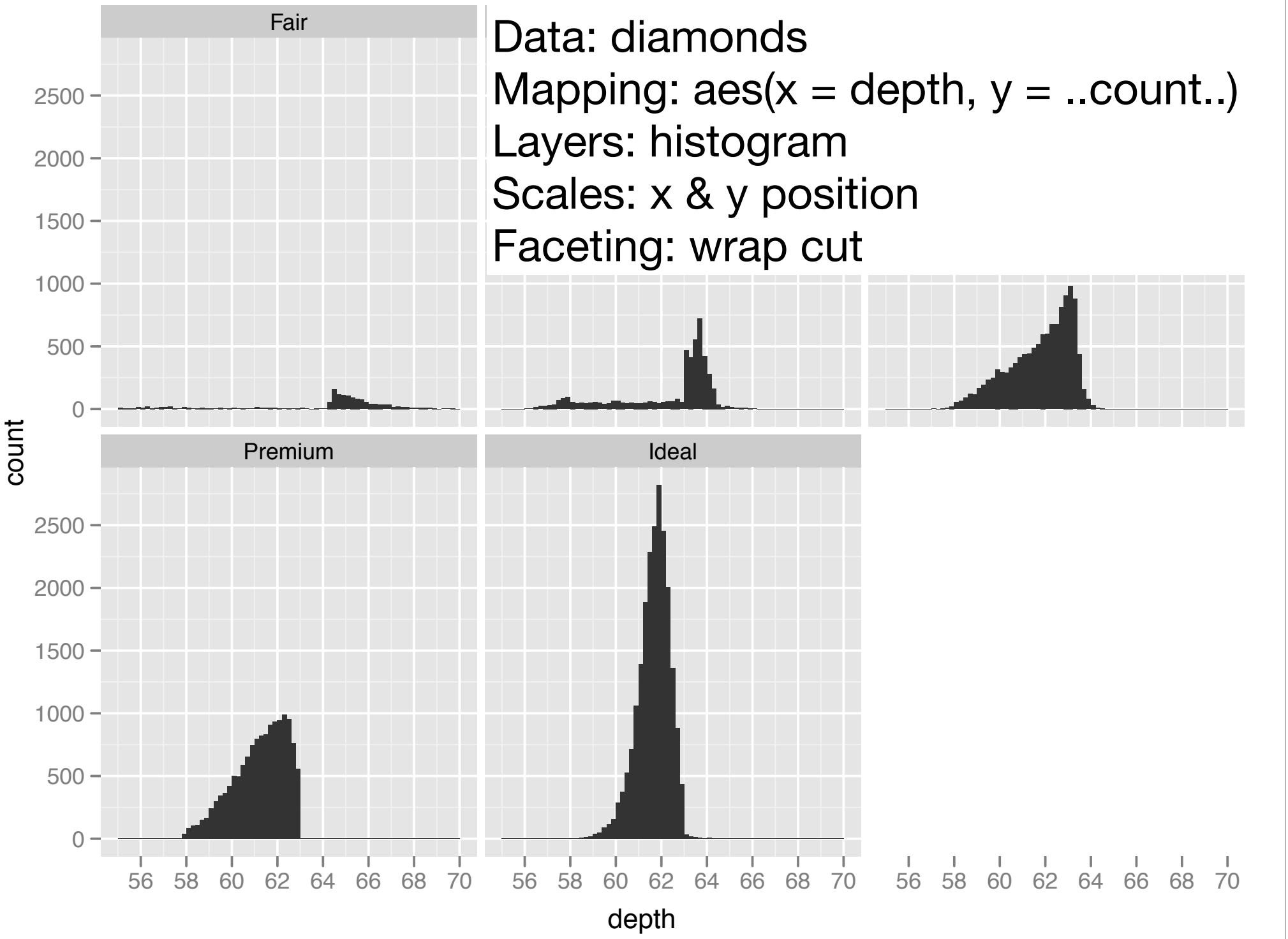
- A default dataset and set of aesthetic mappings
- Multiple layers
- A scale for each aesthetic
- A facetting specification
- A coordinate system


15000
10000
5000
price
1
2
3
4
5
carat
Data: diamonds
Mapping: aes(x = carat, y = price)
Layers: points
Scales: x & y position
Faceting: none

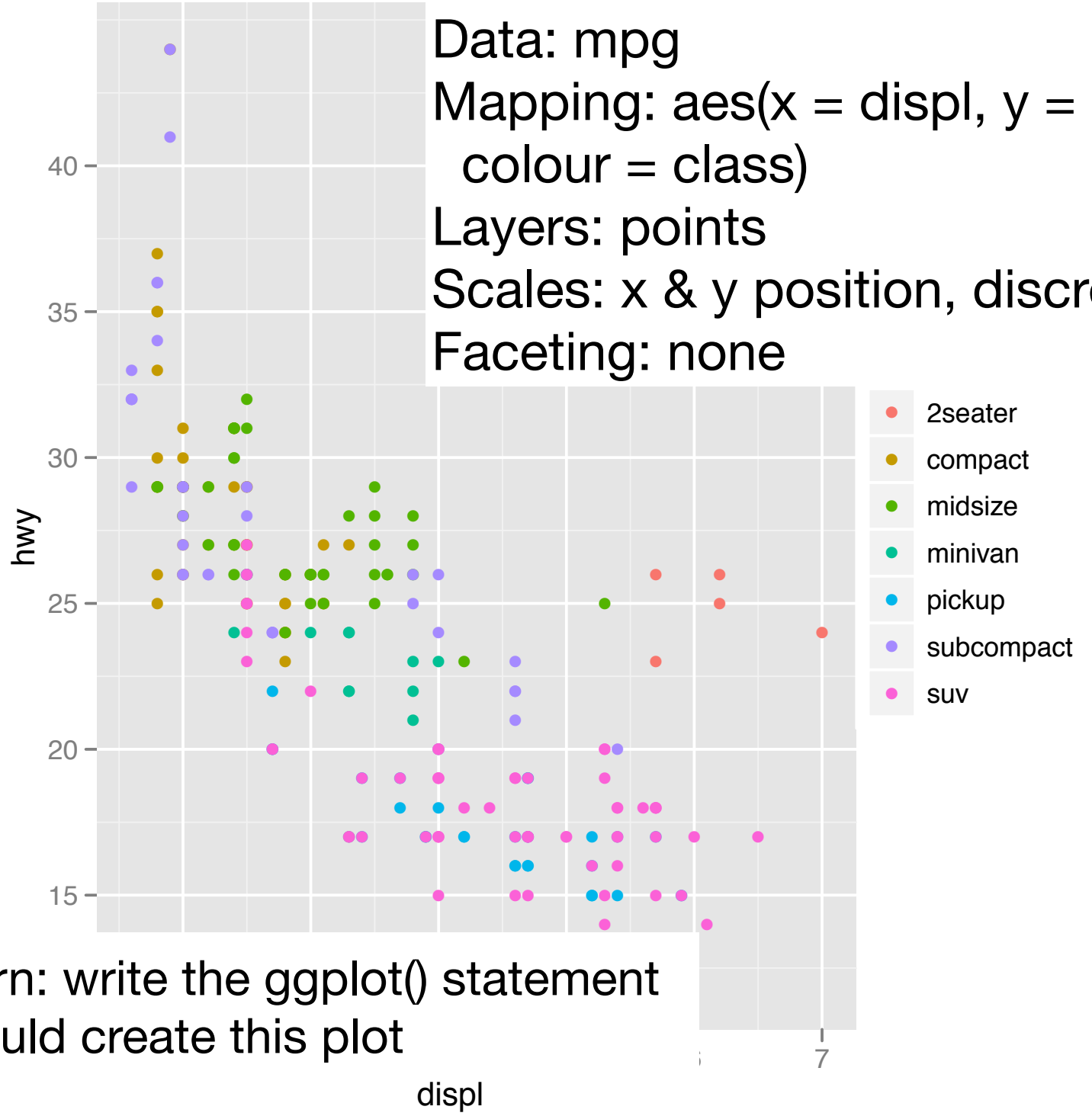
```
qplot(price, carat, data = diamonds)

ggplot(diamonds, aes(x = price, y = carat)) +
  geom_point() +
  scale_x_continuous() +
  scale_y_continuous()

# Scales added automatically, default aes params
ggplot(diamonds, aes(price, carat)) +
  geom_point()
```


Data: mpg
Mapping: aes(x = class, y = hwy)
Layers: jittered points, boxplots
Scales: x & y position
Faceting: none
hwy
40
35
30
25
20
15
pickup
suv
minivan
2seater
midsize
subcompact
compact
reorder(class, hwy)

```
qplot(reorder(class, hwy), hwy, data = mpg,
  geom = c("jitter", "boxplot"))

ggplot(mpg, aes(reorder(class, hwy), hwy)) +
  geom_jitter() +
  geom_boxplot()
```


Fair
Premium
Ideal
Data: diamonds
Mapping: aes(x = depth, y = ..count..)
Layers: histogram
Scales: x & y position
Faceting: wrap cut
count
depth
0
500
1000
1500
2000
2500
56
58
60
62
64
66
68
70

```
qplot(depth, data = diamonds, binwidth = 0.2) +
   facet_wrap(~ cut)

ggplot(diamonds, aes(depth)) +
  geom_histogram(aes(y = ..count..), binwidth=0.2) +
  facet_wrap(~ cut) + xlim(50, 70)

ggplot(diamonds, aes(depth)) +
  geom_histogram(aes(y = ..density..), binwidth=0.2) +
  facet_wrap(~ cut) + xlim(50, 70)
```



Your turn: write the ggplot() statement that would create this plot

What is a layer?

- A dataset and aesthetic mappings (to override default)
- A geom
- A stat
- A position adjustment

	Geom	Stat
Scatterplot	point	identity
Histogram	bar	bin
Smooth	line + ribbon	smooth
2d histogram	rect	bin

Your turn

Convert some of the qplot() calls you've used this morning to ggplot(). How do you feel about the trade-off between verbosity and explicitness?

Overriding defaults

Layers can override the default data set and aesthetic mappings. This is useful if we want to put multiple datasets on the same plot.

Have seen two examples already.

```
ggplot(choropleth, aes(long, lat, group = group)) +
  geom_polygon(fill = "white", colour = "grey50") +
  geom_polygon(aes(fill = prop))

ggplot(bubble, aes(long, lat)) +
  geom_polygon(aes(group = group), data = states,
    fill = NA, colour = "grey50") +
  geom_point(aes(size = prop, colour = prop))
```

Plot default: `aes(mpg, wt)`

Add	`aes(colour = cyl)`	`aes(mpg, wt, colour = cyl)`
Override	`aes(y = disp)`	`aes(mpg, disp)`
Delete	`aes(y = NULL)`	`aes(mpg)`

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 en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

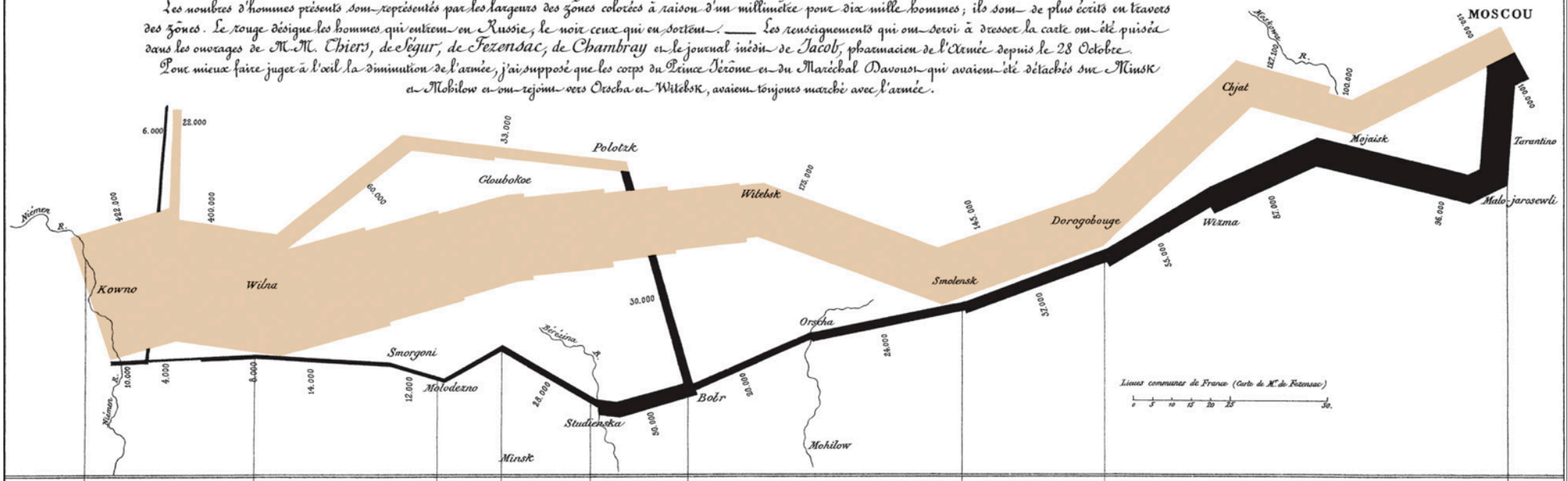
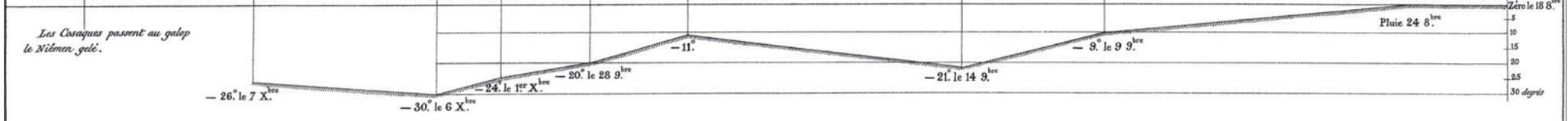
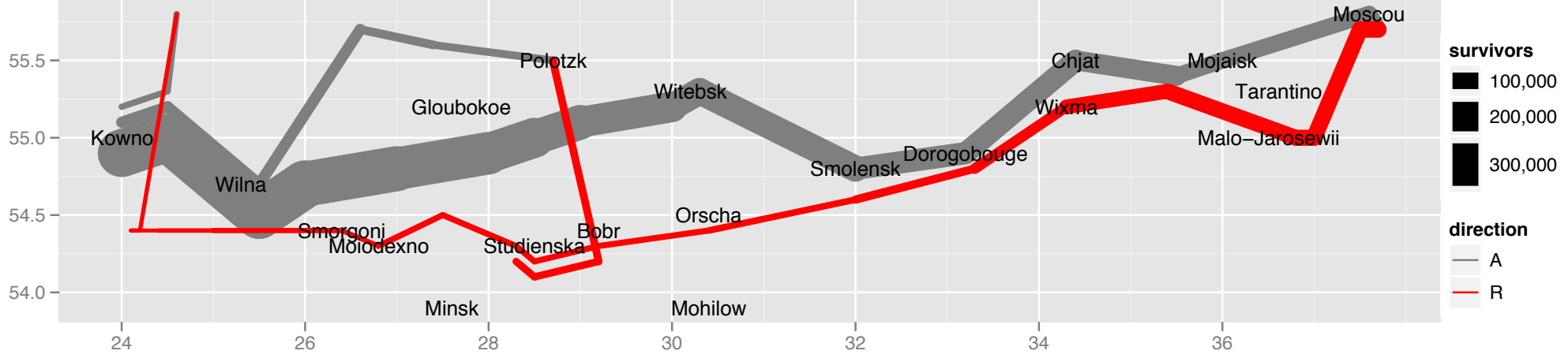


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.

Imp. Lith. Regnier et Dourdet.


Moscou
Chjat
Mojaisk
Polotzk
Witebsk
Tarantino
Gloubokoe
Wixma
Kowno
Malo−Jarosewii
Dorogobouge
Smolensk
Wilna
Orscha
Smorgonj
Bobr
Molodexno
Studienska
Minsk
Mohilow
55.5
55.0
54.5
54.0
24
26
28
30
32
34
36
survivors
100,000
200,000
300,000
direction
A
R

Your turn

How many layers are there on this graph?
What variables are plotted on each layer?
What datasets do we need to recreate this plot?

Layers

Bottom layer: Path of troops, coloured by direction, size proportion to number.

Top layer: Location of major cities

Your turn

Load the minard-troops and minard-cities dataset and attempt to recreate the plot. Make sure to specify the common features of each layer as plot defaults.

Once you get that done, use your plot polishing skills to try and make the plot look as close to the original as possible.

```
ggplot(cities, aes(long, lat)) +
  geom_path(aes(size = survivors, colour = direction,
    group = interaction(group, direction)), data = troops) +
  geom_text(aes(label = city), hjust = 0, vjust = 1, size = 4)

# Polish appearance
last_plot() +
  scale_x_continuous("", limits = c(24, 39)) +
  scale_y_continuous("") +
  scale_colour_manual(values = c("grey50","red")) +
  scale_size(to = c(1, 10))
```

Graphics are like pumpkin pie

Or, the four C's of critiquing a graphic

Content

Construction


Context

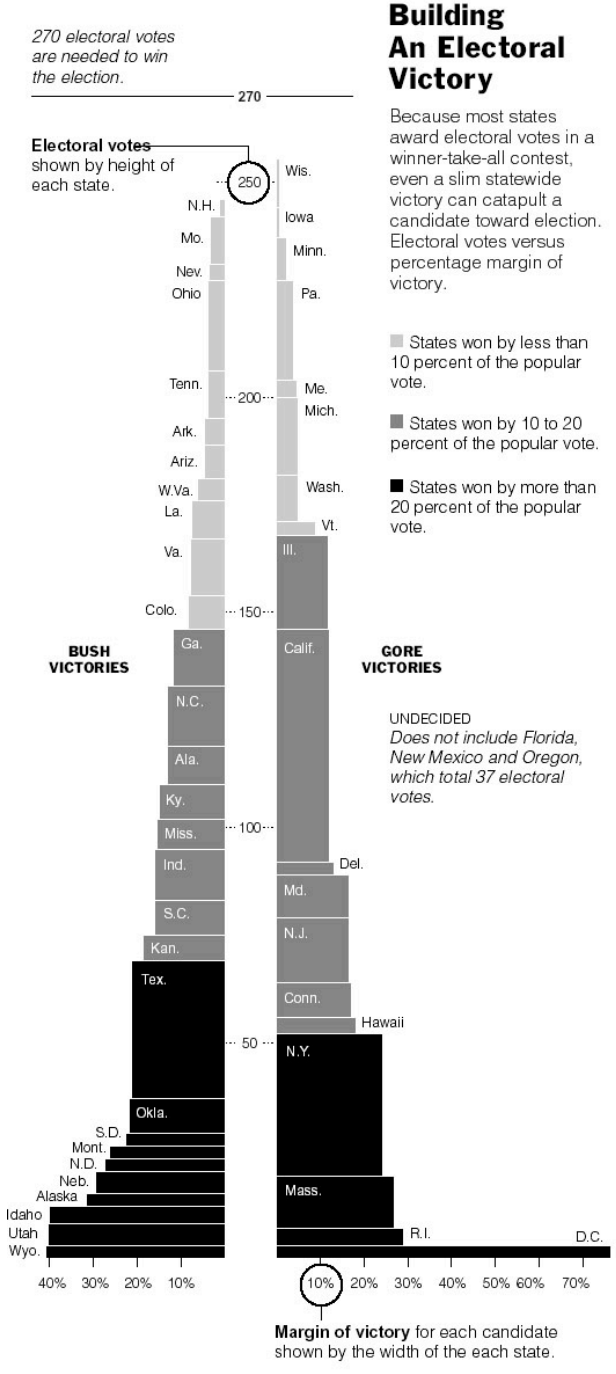

Consumption

Content

- What data (variables) does the graph display?
- What non-data is present?

Building An Electoral Victory

Because most states award electoral votes in a winner-take-all contest, even a slim statewide victory can catapult a candidate toward election. Electoral votes versus percentage margin of victory.



States won by less than 10 percent of the popular vote.

States won by 10 to 20 percent of the popular vote.

States won by more than 20 percent of the popular vote.

UNDECIDED

Does not include Florida, New Mexico and Oregon, which total 37 electoral votes.

The New York Times

United States Geological Survey

The Las Vegas metropolitan area in the 21st century.

Magnet City

1905 Las Vegas is founded as a railroad town, after completion of the main railway linking Southern California with Salt Lake City.

1931 Nevada legalizes gambling and adopts "quickie" divorce laws. Building begins on Hoover Dam, which brings in workers and lifts the region's economy during the Great Depression.

1940's Clark County's population nearly triples, in large part because of the construction of Nellis Air Force Base and the Basic Magnesium Plant, which produces munitions for World War II.

1946 A New York mobster, Benjamin (Bugsy) Siegel, opens the Flamingo Hotel. After the war, tourism becomes the region's largest employer as lavish casinos and resorts are built. The 1950's and 60's are the heyday of Mafia rule.

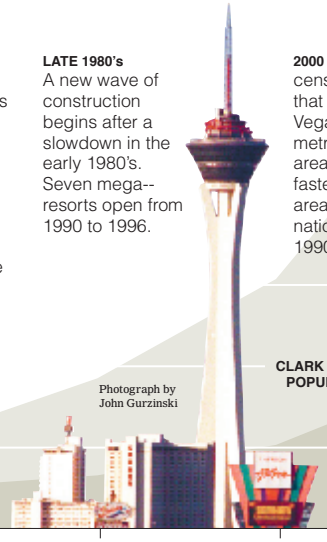
1955 The nine-story Riviera becomes the first high-rise hotel on the Las Vegas Strip.

1966 The reclusive billionaire Howard Hughes moves into the Desert Inn, which he later acquires as part of a hotel and casino buying spree.

1967 After the Nevada Legislature allows publicly traded companies to obtain gambling licenses, corporations move in and begin their domination of the casino industry.

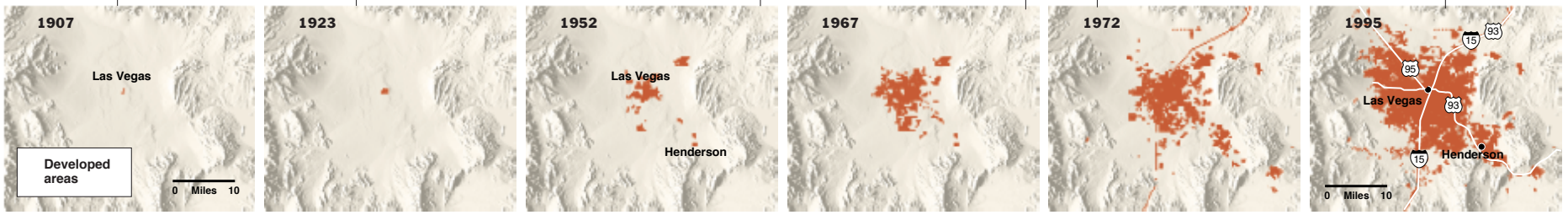
LATE 1980's A new wave of construction begins after a slowdown in the early 1980's. Seven mega--resorts open from 1990 to 1996.

2000 The census finds that the Las Vegas metropolitan area was the fastest-growing area in the nation in the 1990's.



Photograph by John Gurzinski

	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
CLARK COUNTY	3,371	4,859	8,532	16,414	48,589	127,016	273,288	463,087	768,203	1,428,690
INCREASE	+214%	+44%	+76%	+92%	+196%	+161%	+115%	+69%	+66%	+86%
U.S. POPULATION	+21%	+15%	+16%	+7%	+14%	+19%	+13%	+11%	+10%	+13%



Sources: City of Las Vegas; Las Vegas Chamber of Commerce; Nevada State Demographer; U.S. Census Bureau; U.S. Geological Survey

The New York Times

More questions

- What is missing?
- What other data would be useful?
- Where did the data come from?

Construction

- What type of plot is it?
- How are the variables mapped to objects in the graphic?
- How is non-data used?

Perceptual mapping

Best

Worst

1. Position along a common scale
2. Position along nonaligned scale
3. Length, direction, angle
4. Area
5. Volume, curvature
6. Shading, colour saturation

More questions

- Are conventions followed?
- Good perceptual mappings?
- High "data-ink" ratio?
- Low "lie-factor"?
- Appropriate aspect ratio?
- Is it aesthetically pleasing?

Context

- Why was the graphic created?
- What story is it telling?

- What is the graphic trying to make you think?
- Is there a hidden agenda?

Consumption

- How does the graph make you feel?
- Are you still hungry for more?

More questions

- What were your initial impressions?
- How did you feel after studying it longer?