The Grammar of Graphics & ggplot2

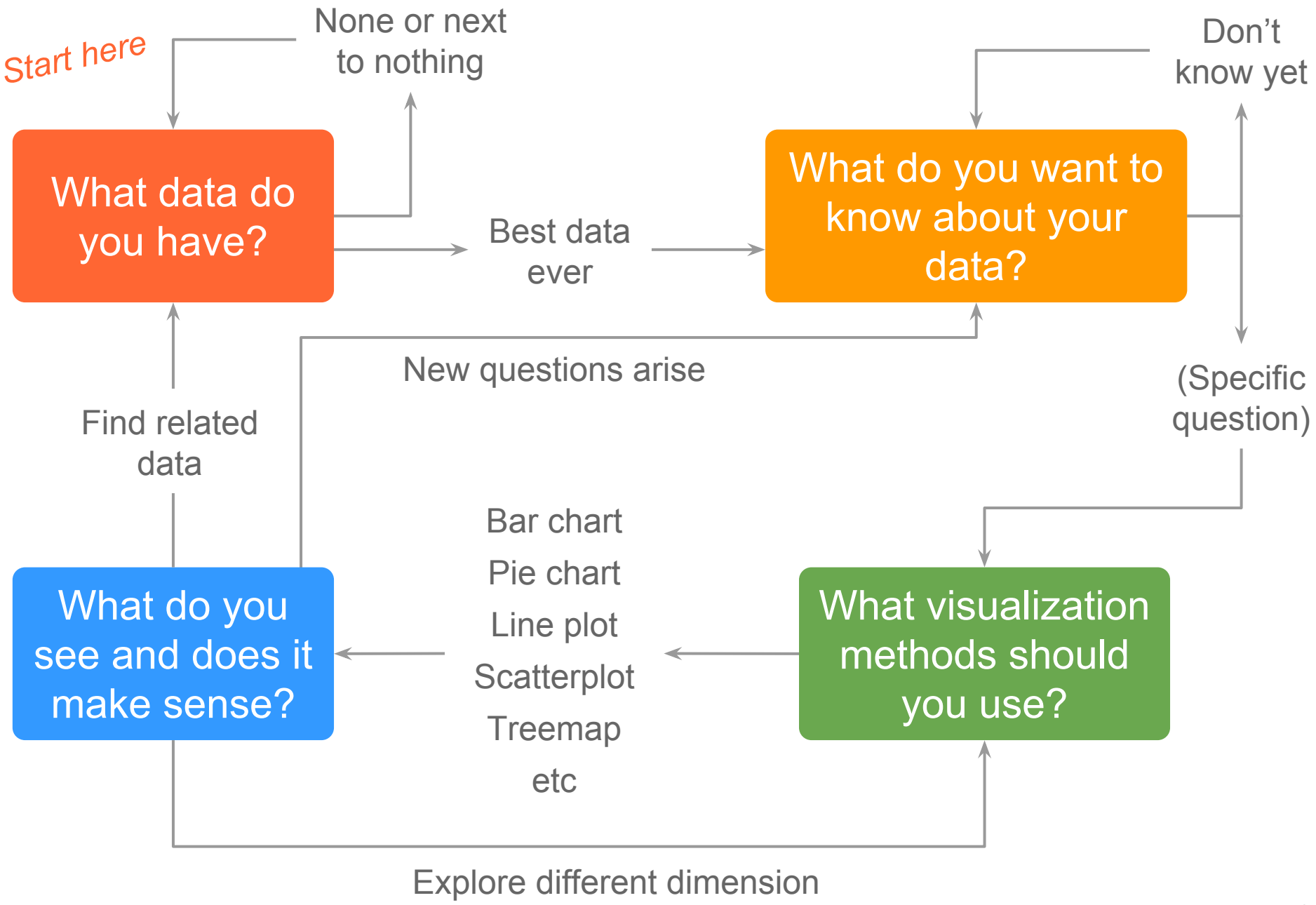
Stat 133 by Gaston Sanchez

So you want to make a plot?

The plotting steps vary by dataset and project

But you should consider four things:

1. What data do you have?
2. What do you want to know about the data?
3. What visualization methods should you use?
4. What do you see and does it make sense?



The iterative data exploration process by Nathan Yau

What data do you have?

How many variables?

1. One variable
2. Two variables
3. Three or more

What type of variables?

4. Quantitative, qualitative, time

What do you want to know about your data?

- Part-to-whole analysis
- Ranking analysis
- Deviation analysis
- Times series (trends in time)
- Distribution analysis
- Correlation analysis
- Multivariate analysis

What do you do see?

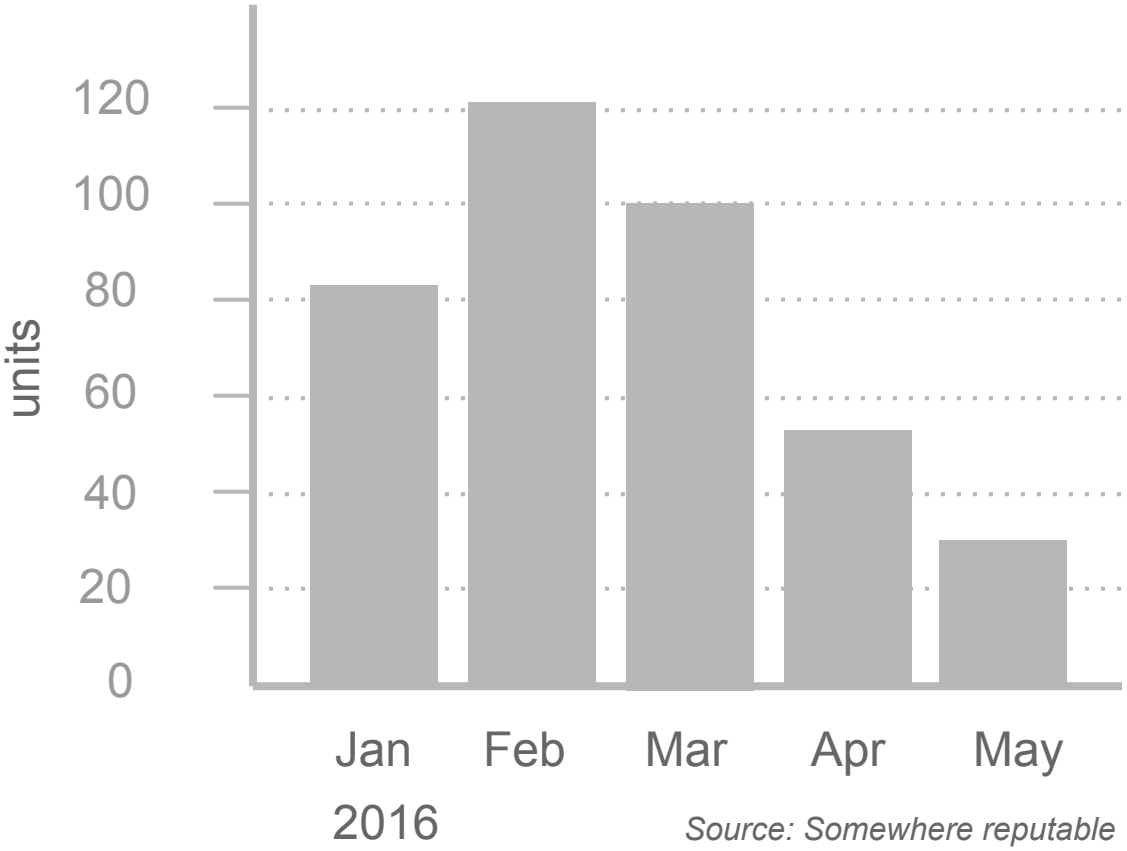
- Systematic variation
- Increasing patterns
- Decreasing patterns
- Atypical or outliers
- Noise?

Visualization is simply mapping data to geometry and color

Visualization is simply **mapping data** to **geometry** and **color**

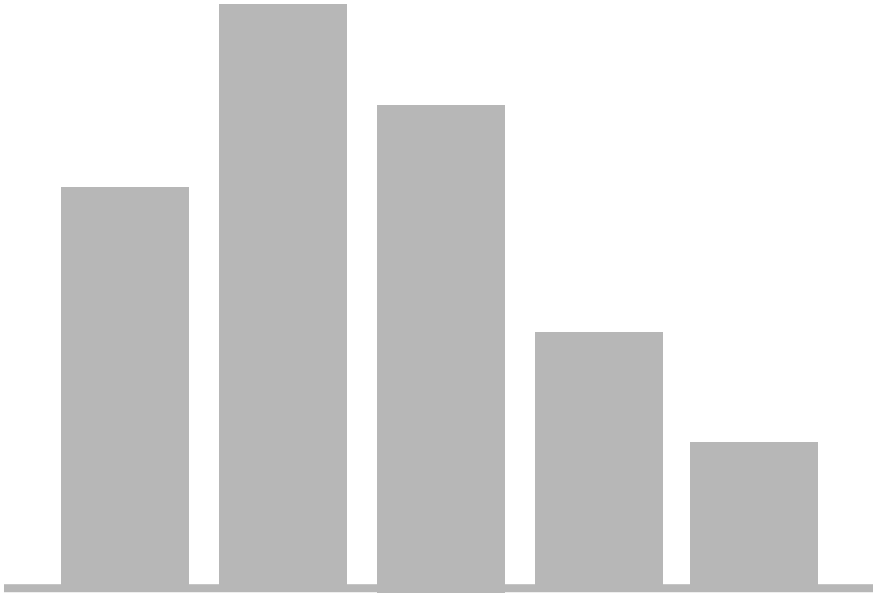
Title of this Graph

A description of the data or something worth highlighting to set the stage



Source: Somewhere reputable

Visual Cues
Encoding data with shapes, colors, and sizes. Which cues you choose depends on your data and your goals



Coordinate System
Mapping data requires a system of coordinates: cartesian, polar, etc

Scale
Increments that make sense can increase readability as well as shift focus

Title of this Graph

A description of the data or something worth highlighting to set the stage

units

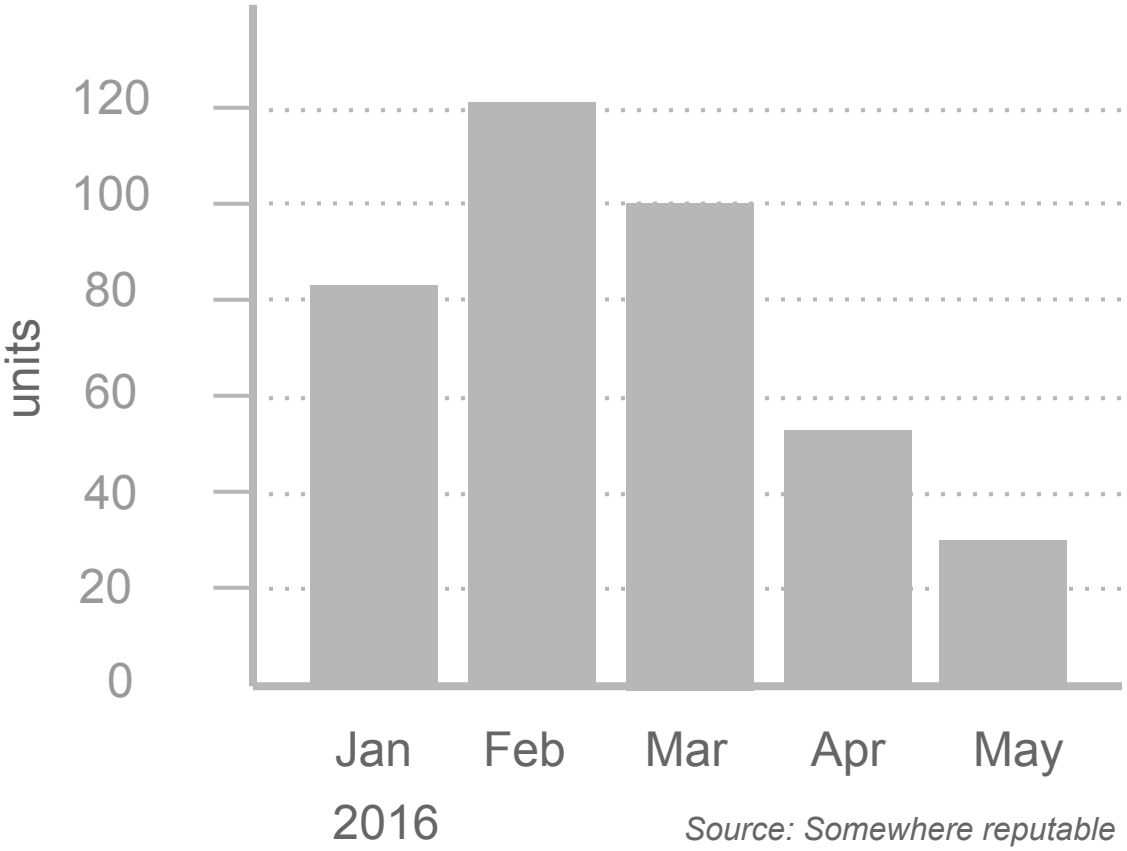
Context
If your audience is unfamiliar with the data, it's your job to clarify what values represent and explain how people should read your plot

2016

Source: Somewhere reputable

Title of this Graph

A description of the data or something worth highlighting to set the stage

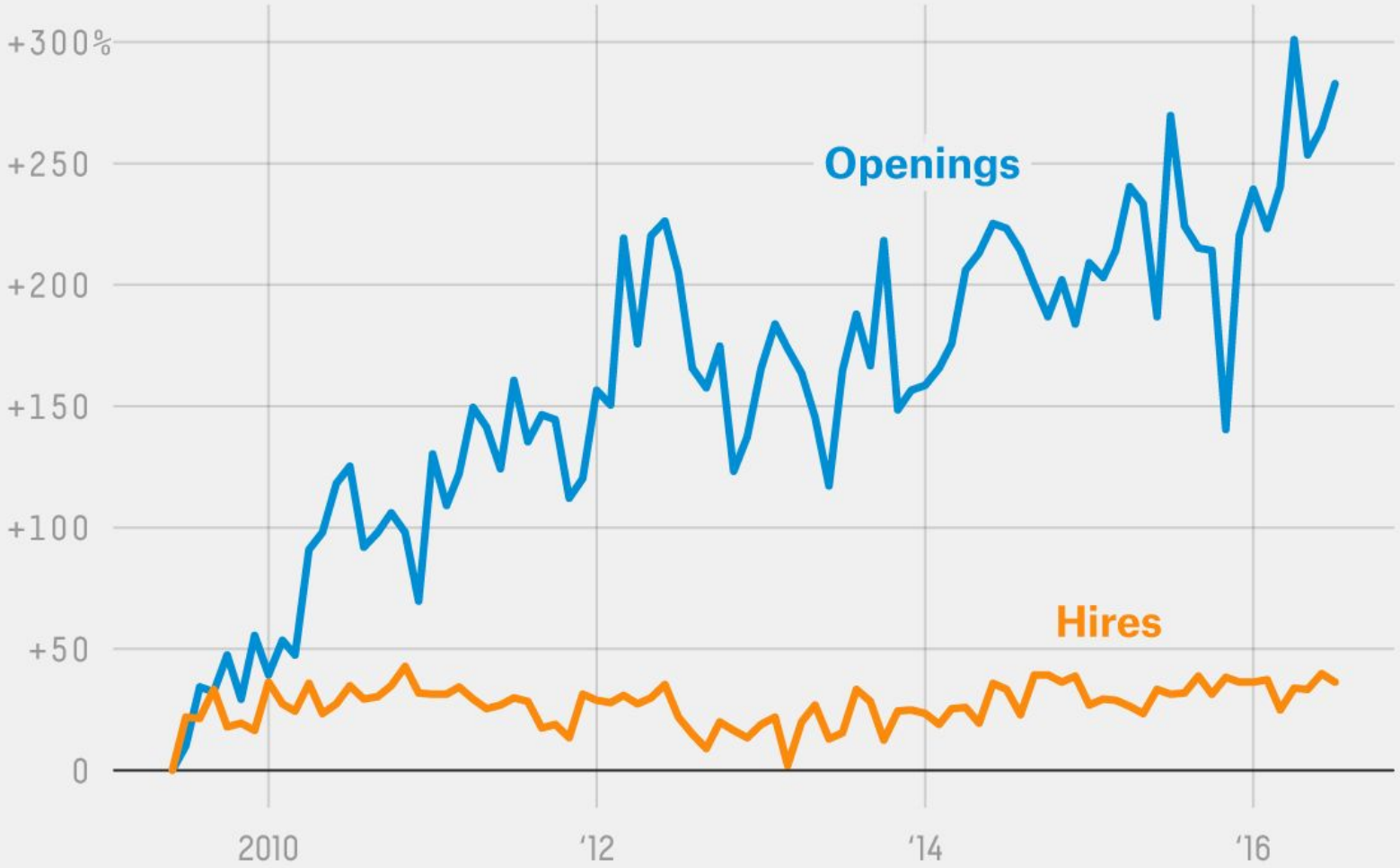


Source: Somewhere reputable

Chart Elements

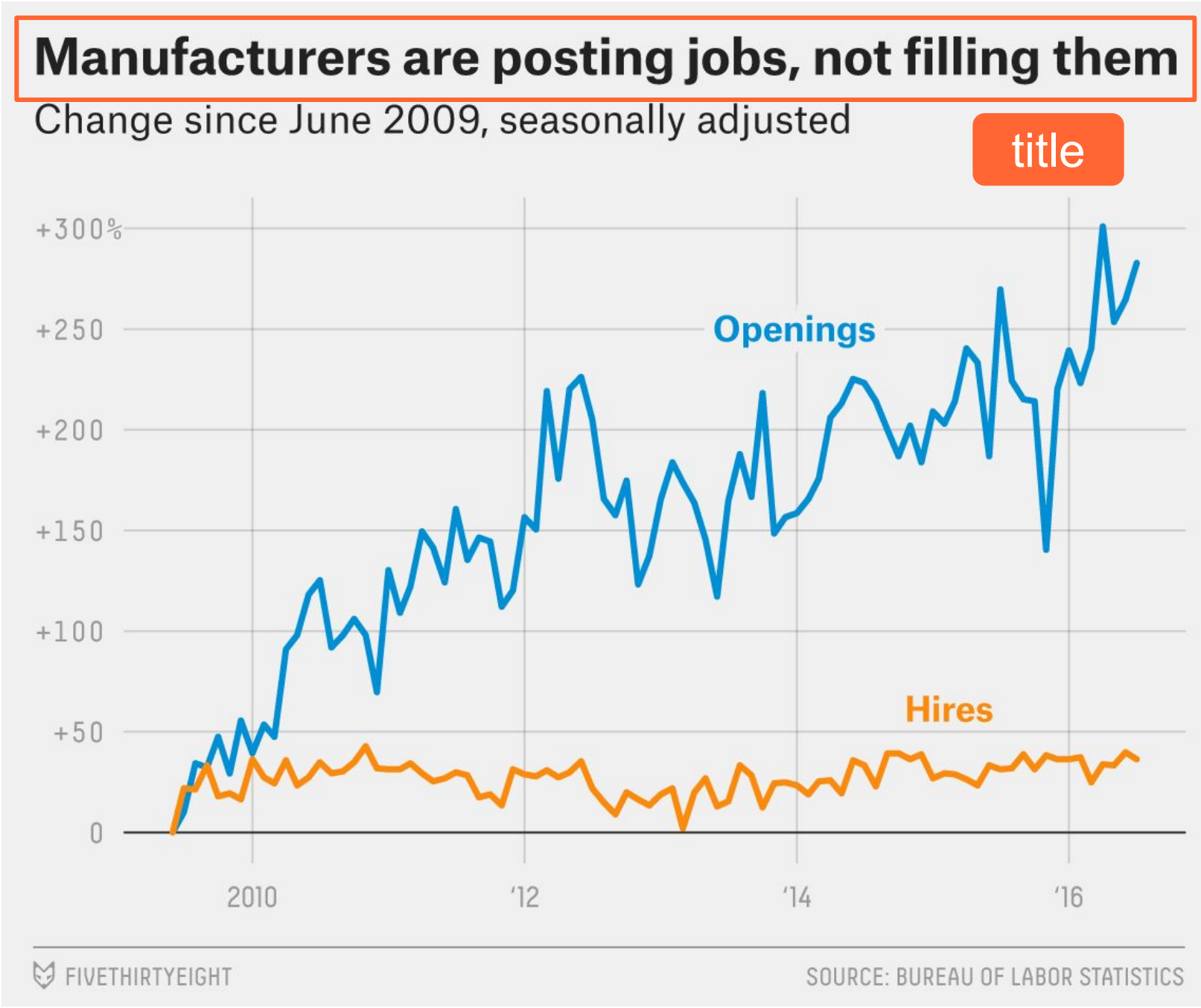
Manufacturers are posting jobs, not filling them

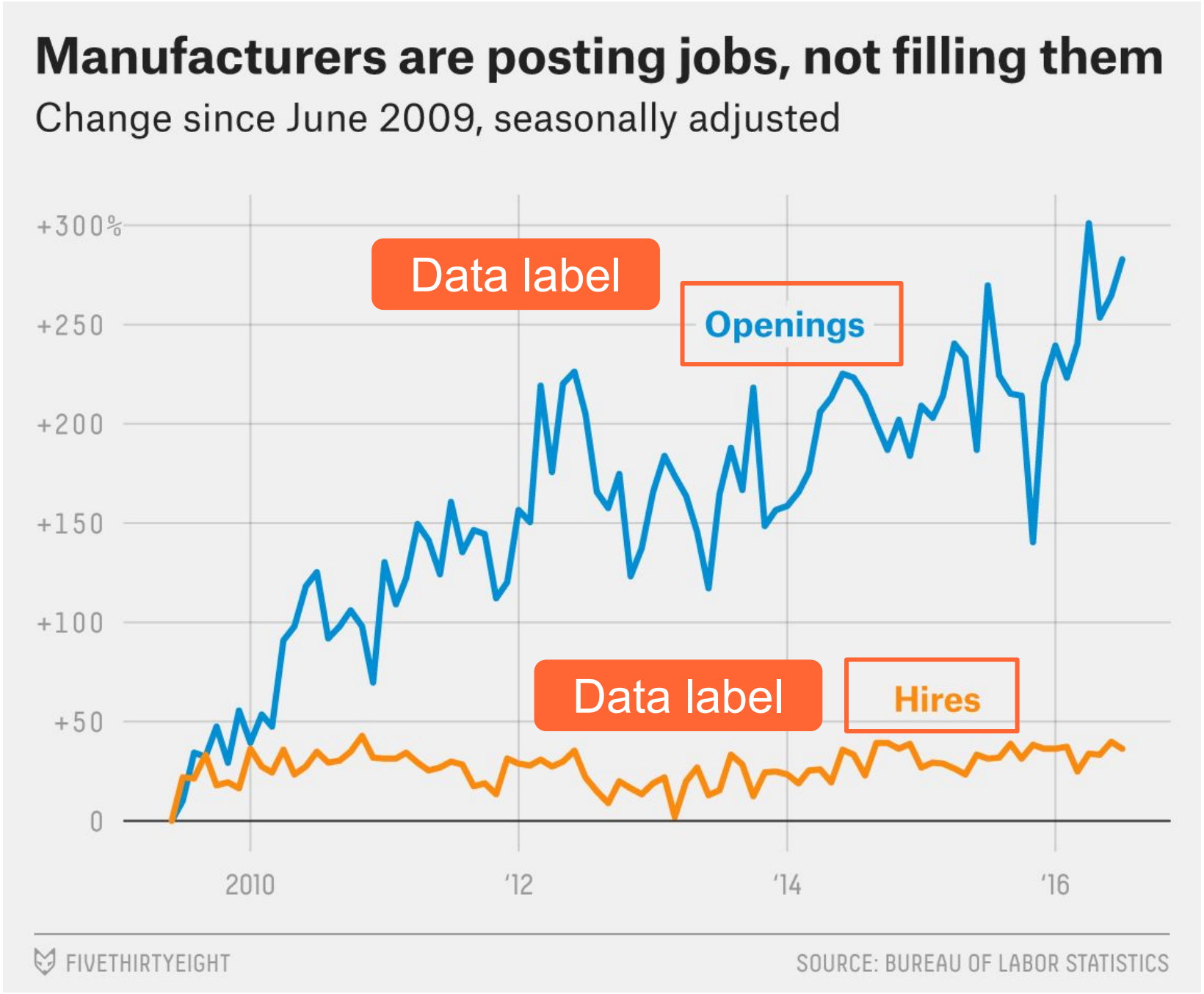
Change since June 2009, seasonally adjusted

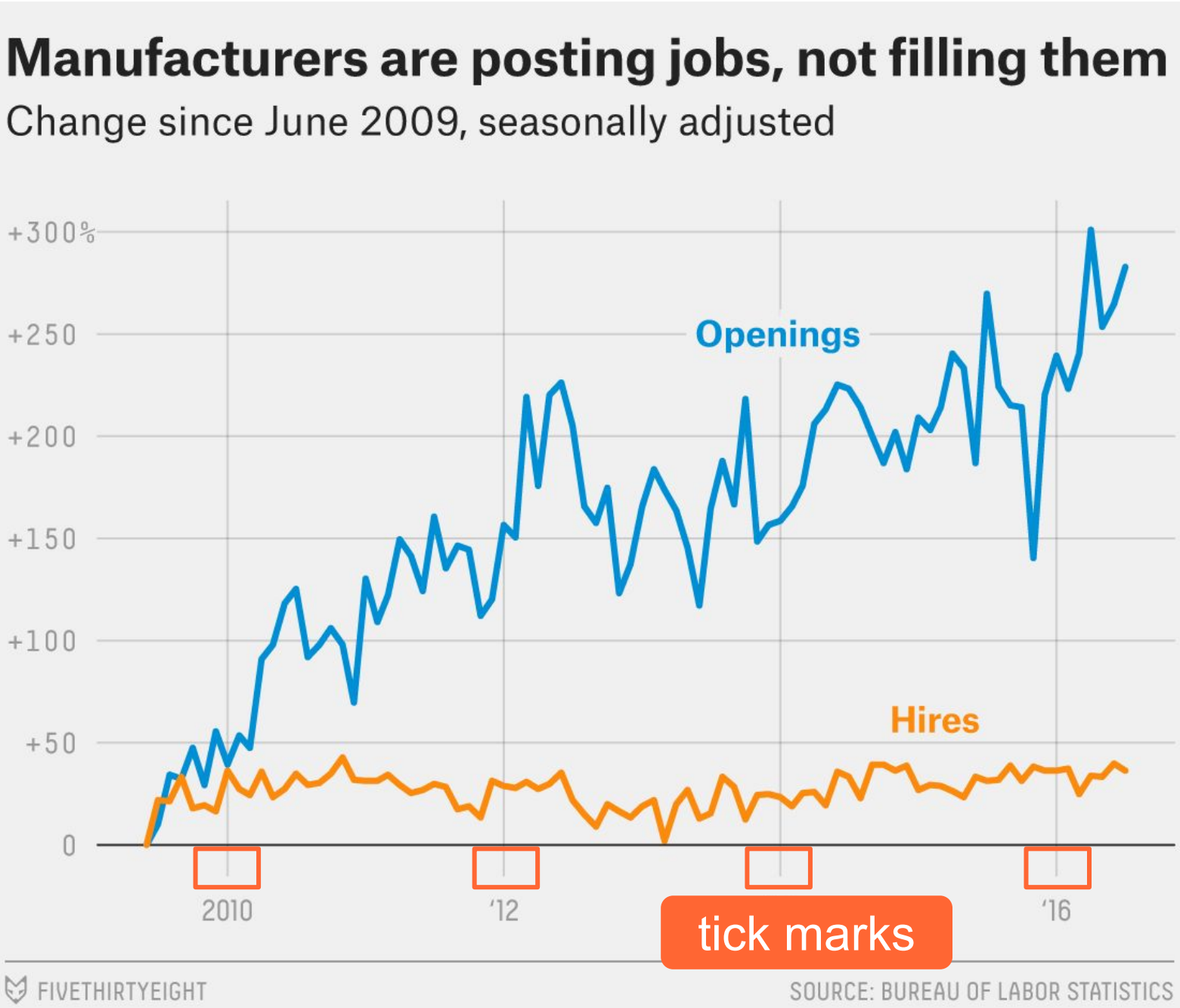


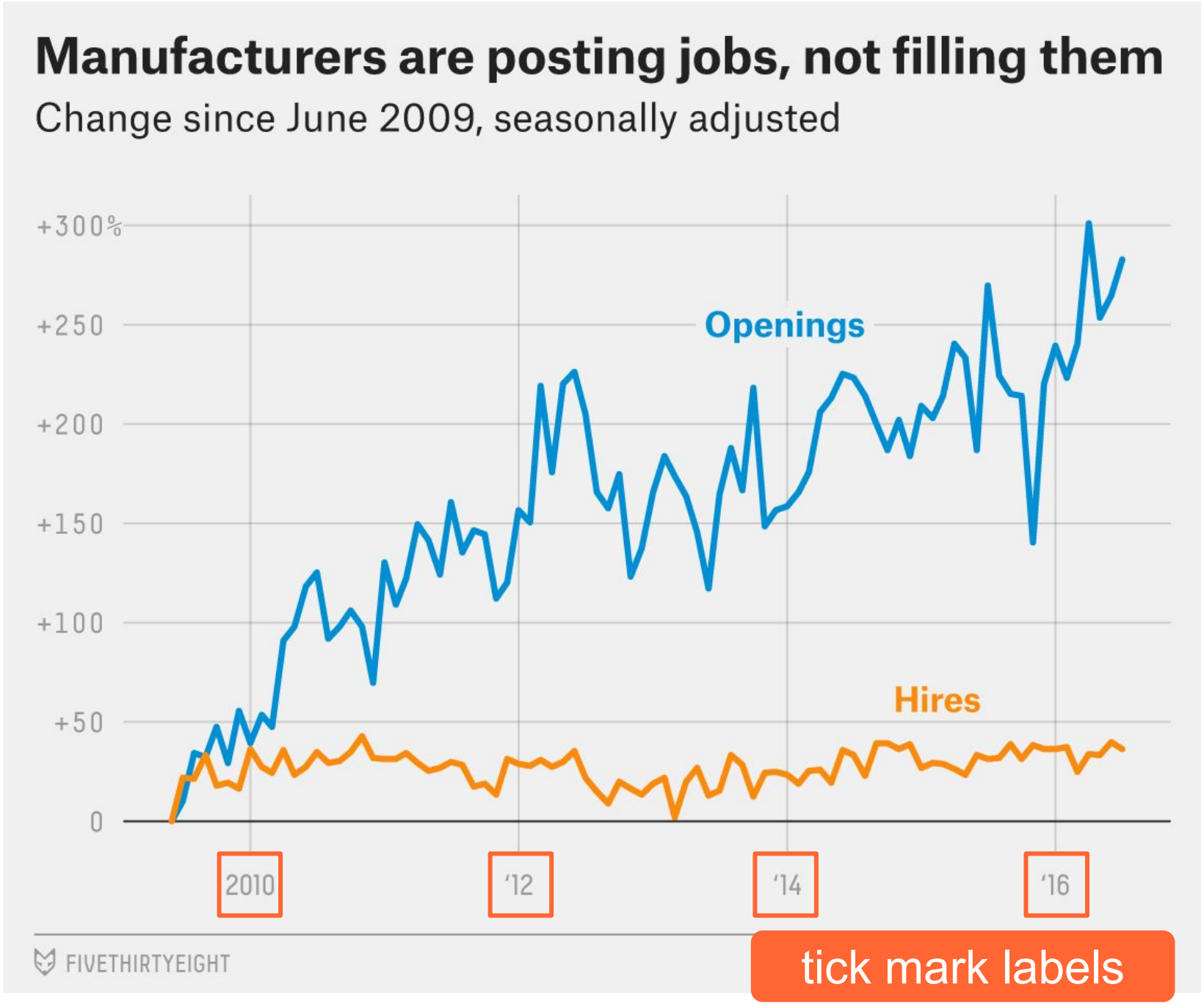
FIVETHIRTYEIGHT

SOURCE: BUREAU OF LABOR STATISTICS

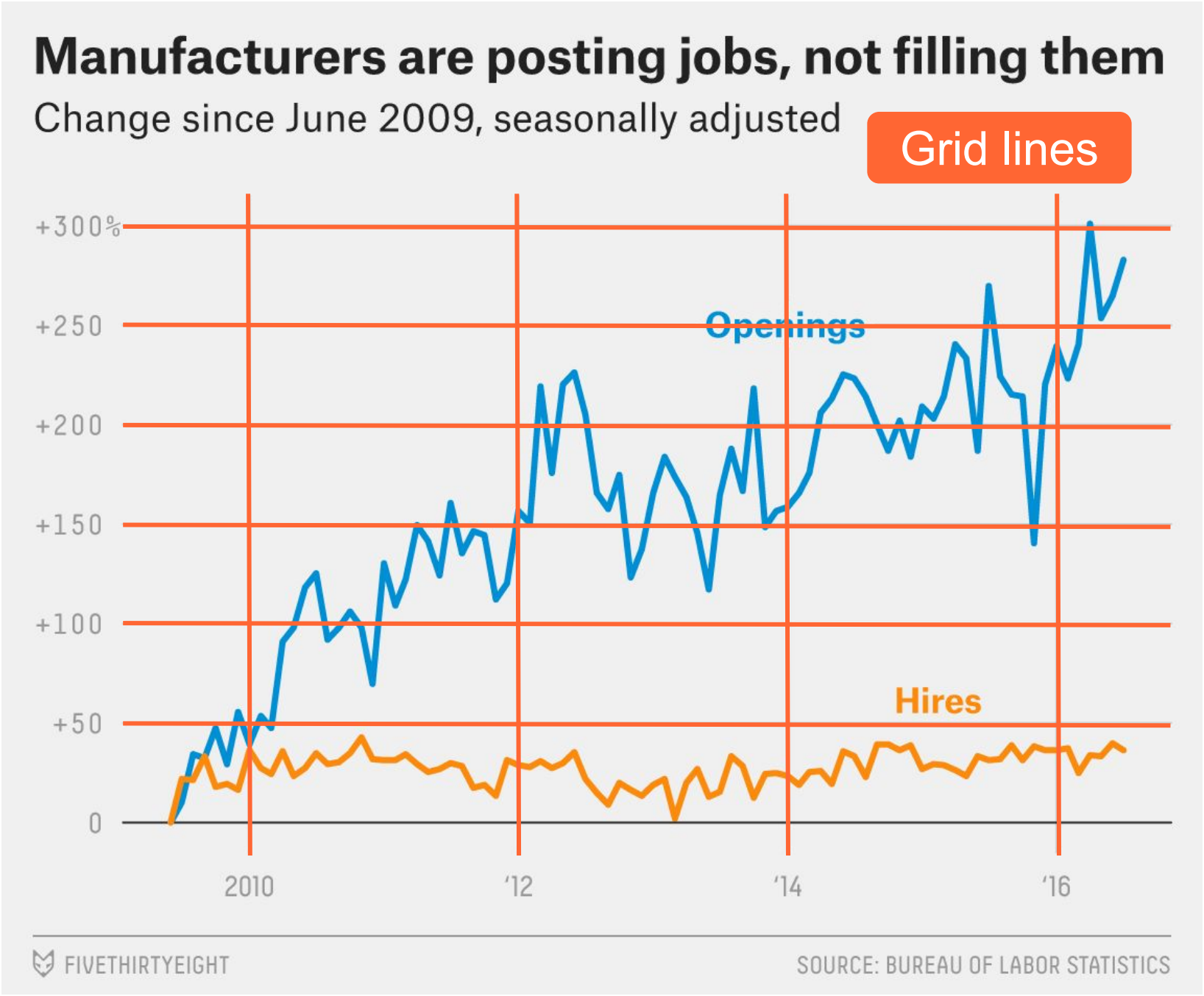

Manufacturers are posting jobs, not filling them
Change since June 2009, seasonally adjusted
title
+300%
+250
+200
+150
+100
+50
0
Openings
Hires
2010
'12
'14
'16
FIVETHIRTYEIGHT
SOURCE: BUREAU OF LABOR STATISTICS


Manufacturers are posting jobs, not filling them
Change since June 2009, seasonally adjusted
+300%
+250
+200
+150
+100
+50
0
2010
'12
'14
'16
Data label
Openings
Data label
Hires
FIVETHIRTYEIGHT
SOURCE: BUREAU OF LABOR STATISTICS


Manufacturers are posting jobs, not filling them
Change since June 2009, seasonally adjusted
+300%
+250
+200
+150
+100
+50
0
Openings
Hires
2010
'12
'16
tick marks
FIVETHIRTYEIGHT
SOURCE: BUREAU OF LABOR STATISTICS



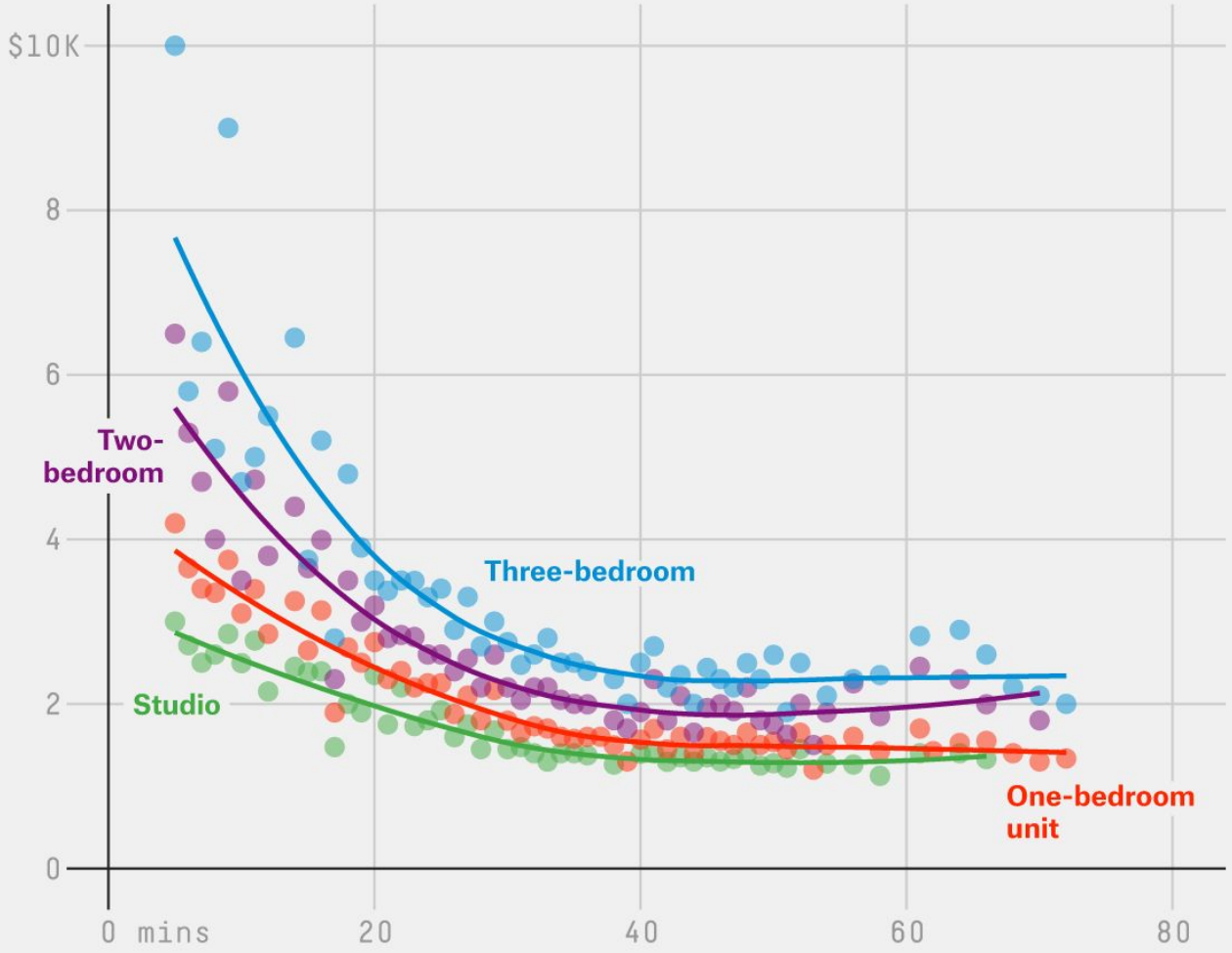
tick mark labels


Manufacturers are posting jobs, not filling them
Change since June 2009, seasonally adjusted
Grid lines
+300%
+250
+200
+150
+100
+50
0
Openings
Hires
2010
'12
'14
'16
FIVETHIRTYEIGHT
SOURCE: BUREAU OF LABOR STATISTICS

Another example

New Yorkers pay up for a shorter commute

Median monthly NYC rent in 2015 vs. commute time by subway



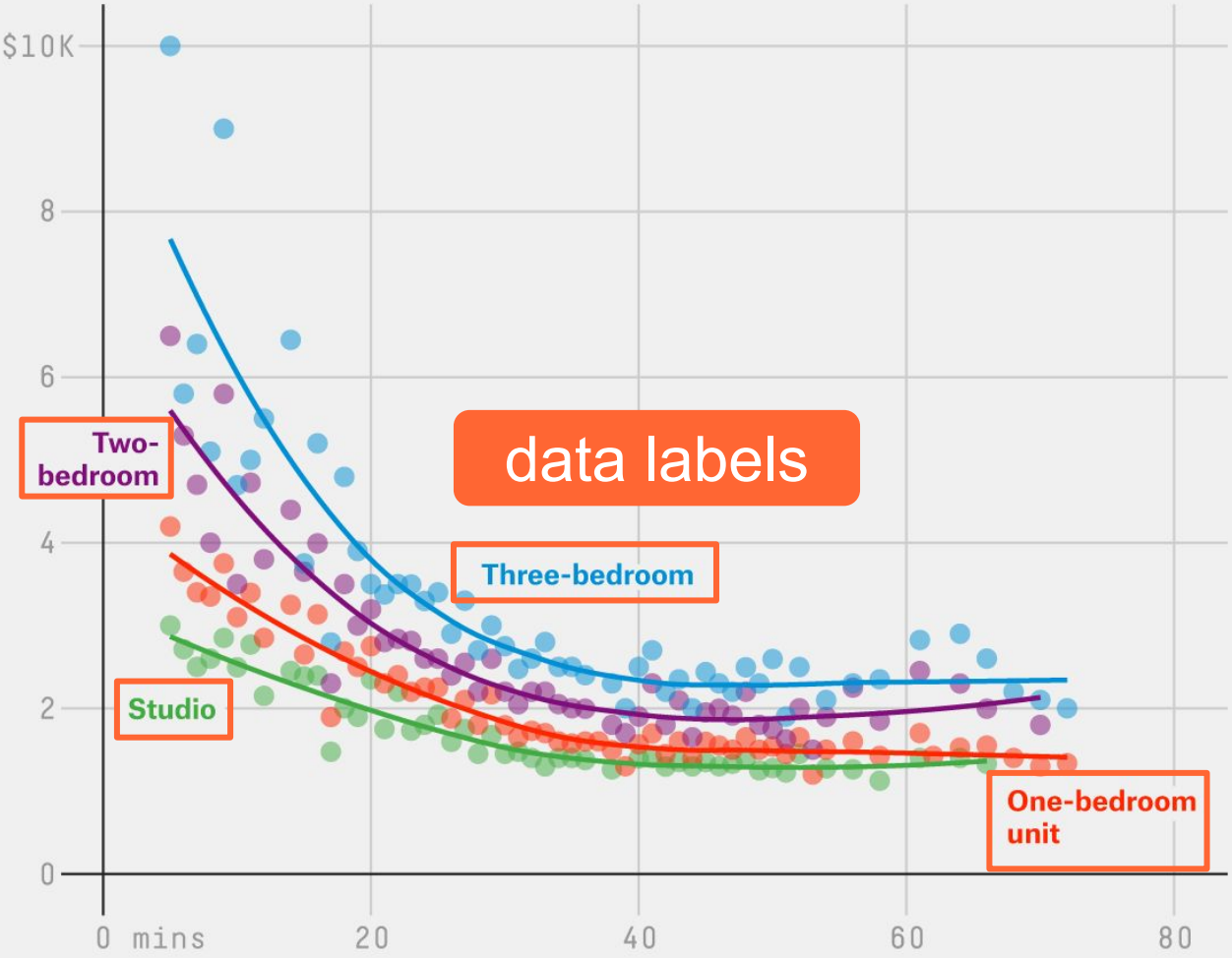
Rental prices are based on FiveThirtyEight's analysis of 2015 listings on StreetEasy. Commutes are calculated as the average time from the subway station nearest a home to the nearest 42nd Street and Chambers Street stations. Commutes without at least 10 listings are excluded.

FIVETHIRTYEIGHT

SOURCE: STREETEASY

New Yorkers pay up for a shorter commute

Median monthly NYC rent in 2015 vs. commute time by subway



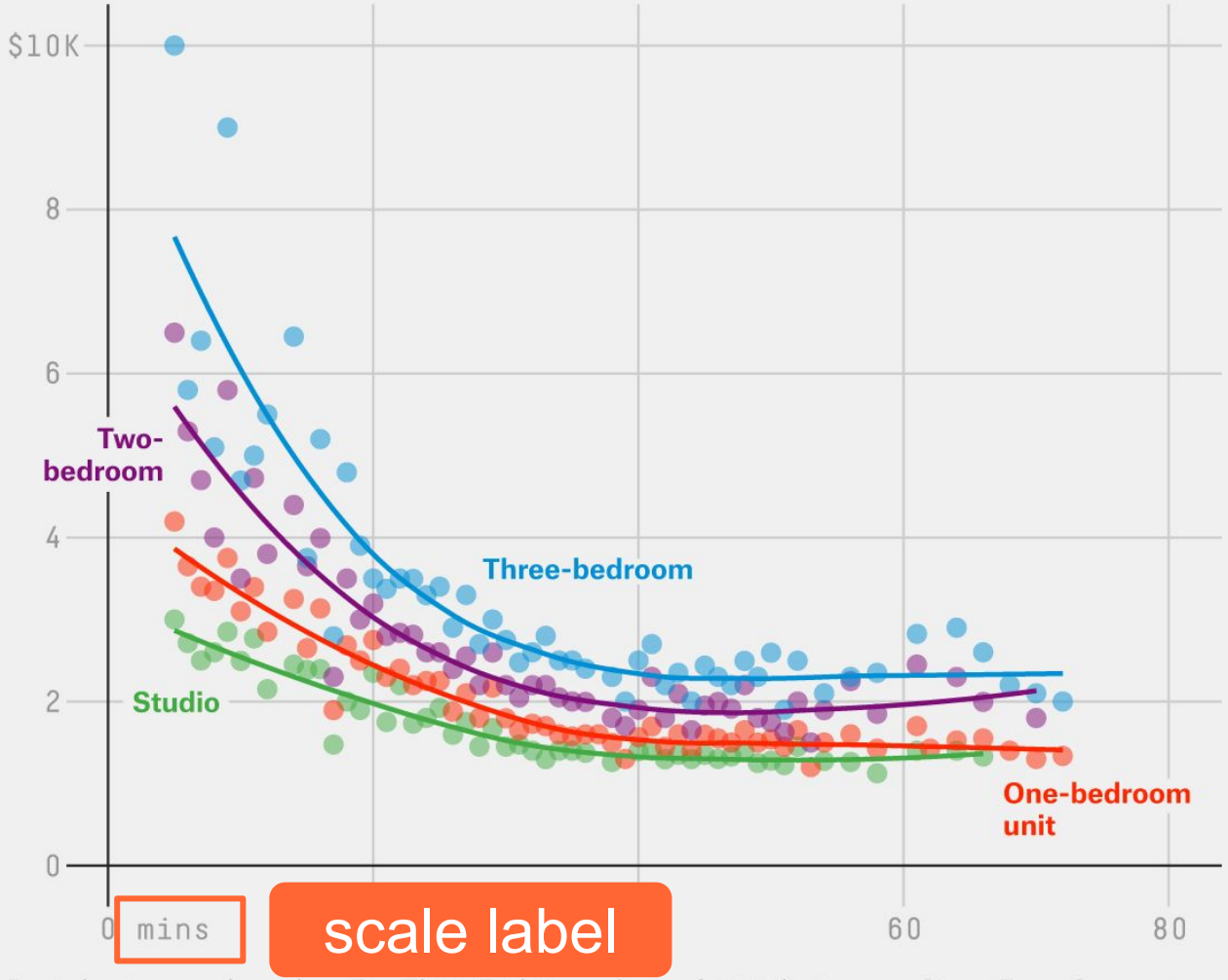
Rental prices are based on FiveThirtyEight's analysis of 2015 listings on StreetEasy. Commutes are calculated as the average time from the subway station nearest a home to the nearest 42nd Street and Chambers Street stations. Commutes without at least 10 listings are excluded.

FIVETHIRTYEIGHT

SOURCE: STREETEASY

New Yorkers pay up for a shorter commute

Median monthly NYC rent in 2015 vs. commute time by subway



Rental prices are based on FiveThirtyEight's analysis of 2015 listings on StreetEasy. Commutes are calculated as the average time from the subway station nearest a home to the nearest 42nd Street and Chambers Street stations. Commutes without at least 10 listings are excluded.

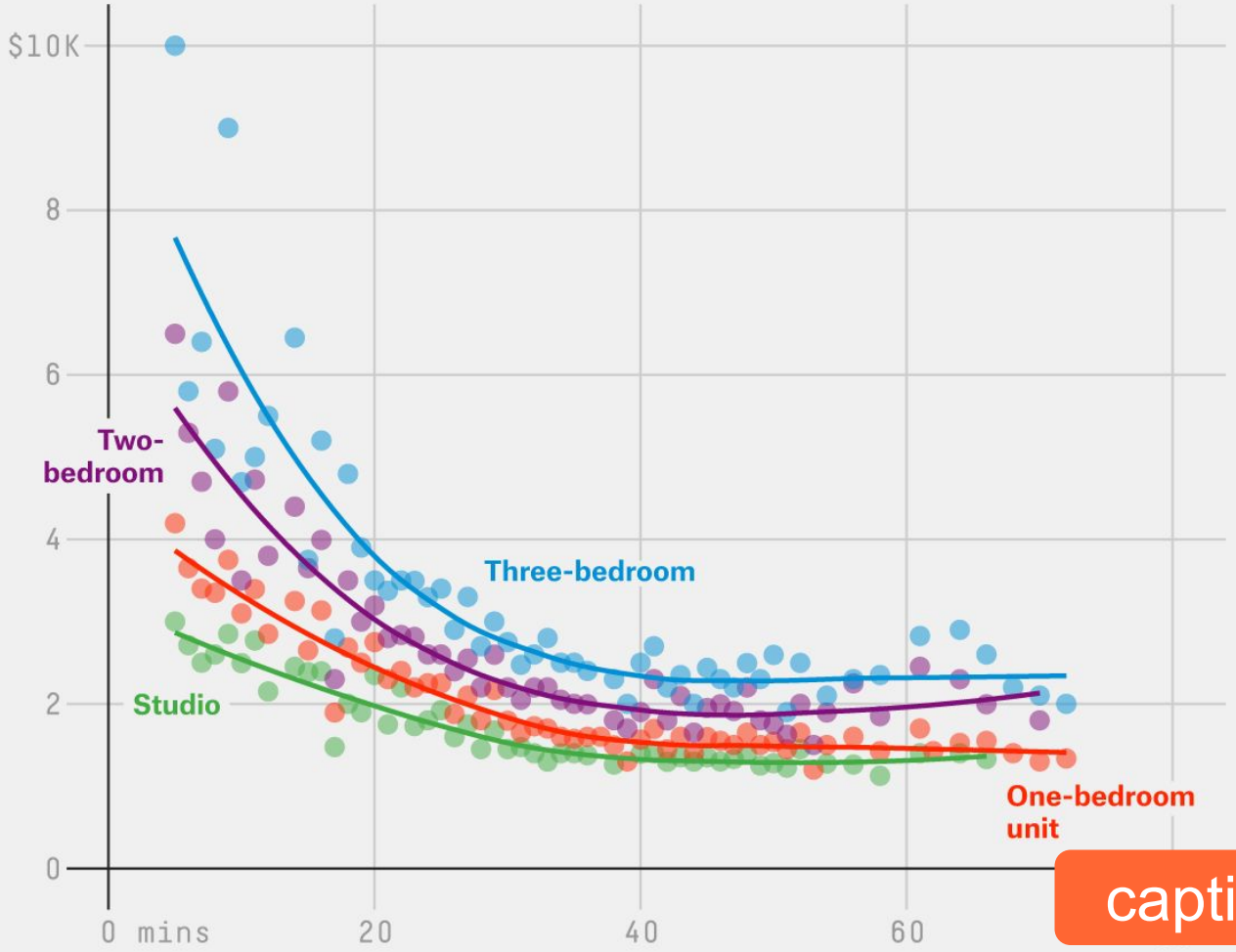
FIVETHIRTYEIGHT

SOURCE: STREETEASY


New Yorkers pay up for a shorter commute
Median monthly NYC rent in 2015 vs. commute time by subway
$10K
scale label
8
6
4
2
0
Two-bedroom
Three-bedroom
Studio
One-bedroom unit
0 mins
20
40
60
80
Rental prices are based on FiveThirtyEight's analysis of 2015 listings on StreetEasy. Commutes are calculated as the average time from the subway station nearest a home to the nearest 42nd Street and Chambers Street stations. Commutes without at least 10 listings are excluded.
FIVETHIRTYEIGHT
SOURCE: STREETEASY

New Yorkers pay up for a shorter commute

Median monthly NYC rent in 2015 vs. commute time by subway



caption

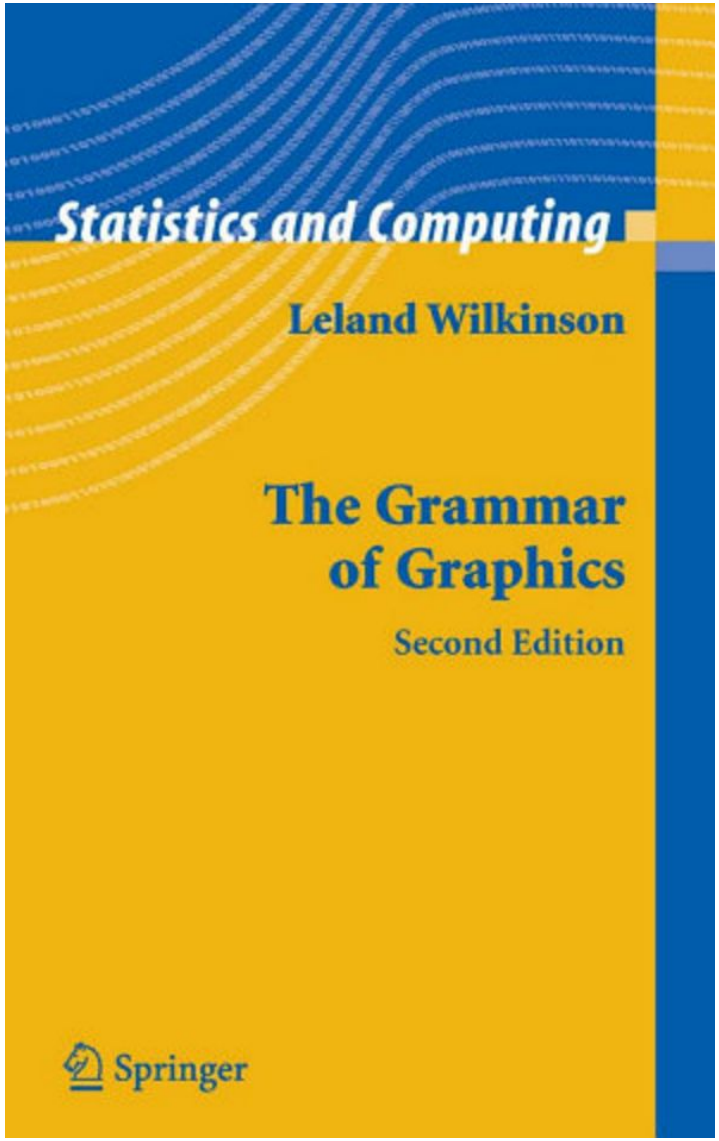
Rental prices are based on FiveThirtyEight's analysis of 2015 listings on StreetEasy. Commutes are calculated as the average time from the subway station nearest a home to the nearest 42nd Street and Chambers Street stations. Commutes without at least 10 listings are excluded.

FIVETHIRTYEIGHT

SOURCE: STREETEASY

Grammar of Graphics with “ggplot2”

The Grammar of Graphics



About the grammar of graphics

The Grammar of Graphics is Wilkinson's attempt to define a theoretical framework for graphics

Grammar: formal system of rules for generating graphics:

- Some rules are mathematic
- Some rules are aesthetic

Aesthetics ≠ Beauty

Aesthetics (GG): attributes of the geometric objects

Meaning of aesthetic in the Grammar of Graphics

Aesthetics: pertaining to sense perception

Aisthesthai = perceive

GG aesthetic attributes: visual properties that affect the way observations are displayed

About the grammar of graphics

Three stages of graphic creation

Specification: link data to graphic objects

Assembly: put everything together

Display: render of a graphic

R package `ggplot2`

Resources

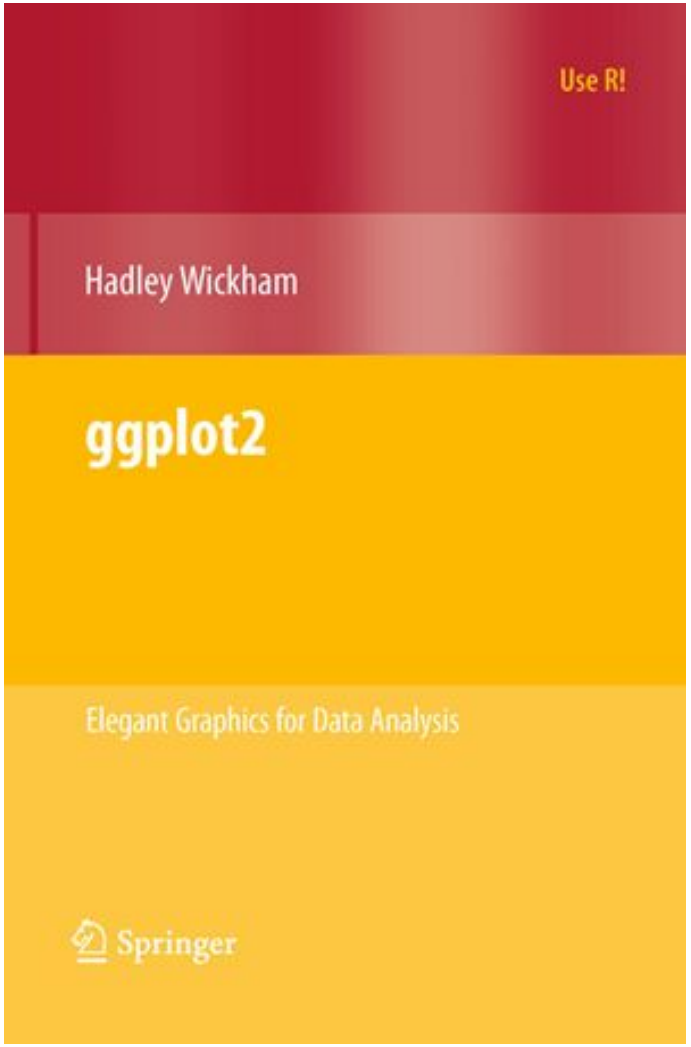
Documentation: http://docs.ggplot2.org

Book: **ggplot2: Elegant Graphics for Data Analysis** by Hadley Wickham

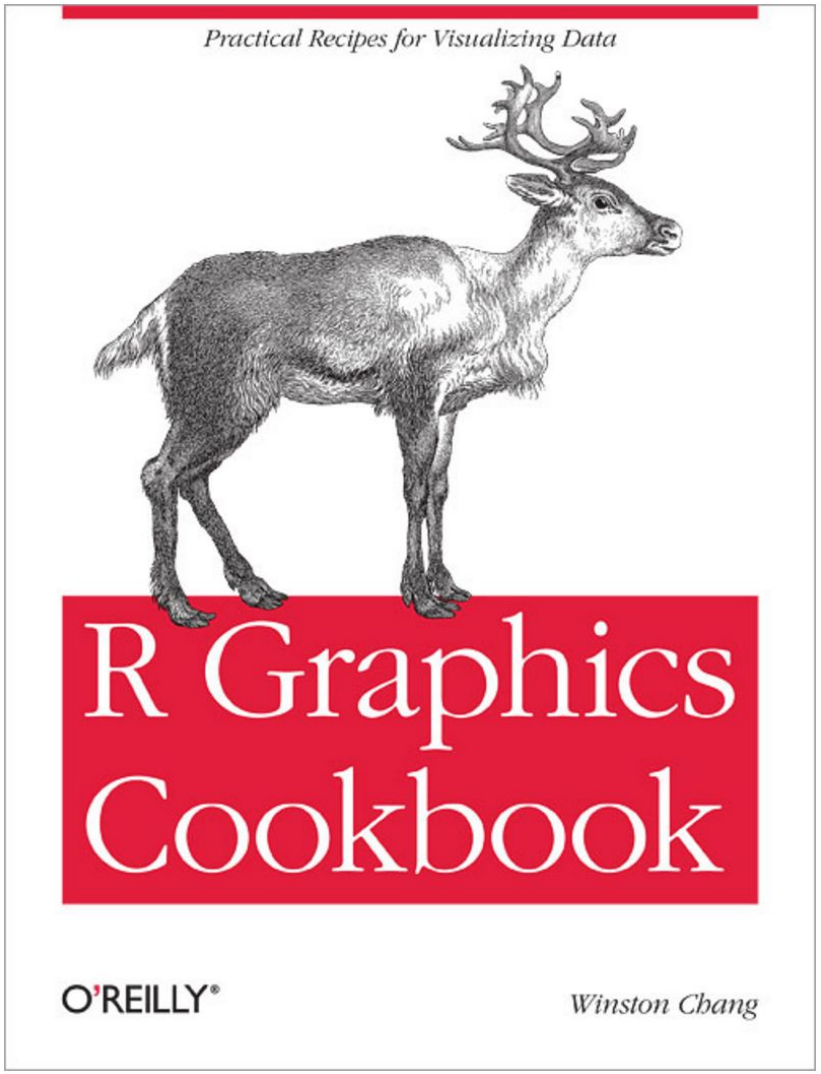
Book: **R Graphics Cookbook** by Winston Chang

RStudio ggplot2 cheat sheet
https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

ggplot2 book



ggplot2 book



R package “ggplot2”

Remember to install ggplot2 (just once)

```
install.packages("ggplot2")

library(ggplot2)

?ggplot
```

About ggplot2

"ggplot2" is an R package for producing statistical graphics.

It provides a framework based on Leland Wilkinson's **Grammar of Graphics**.

"ggplot2" provides beautiful plots while taking care of fiddly details like legends, axes, colors.

About ggplot2

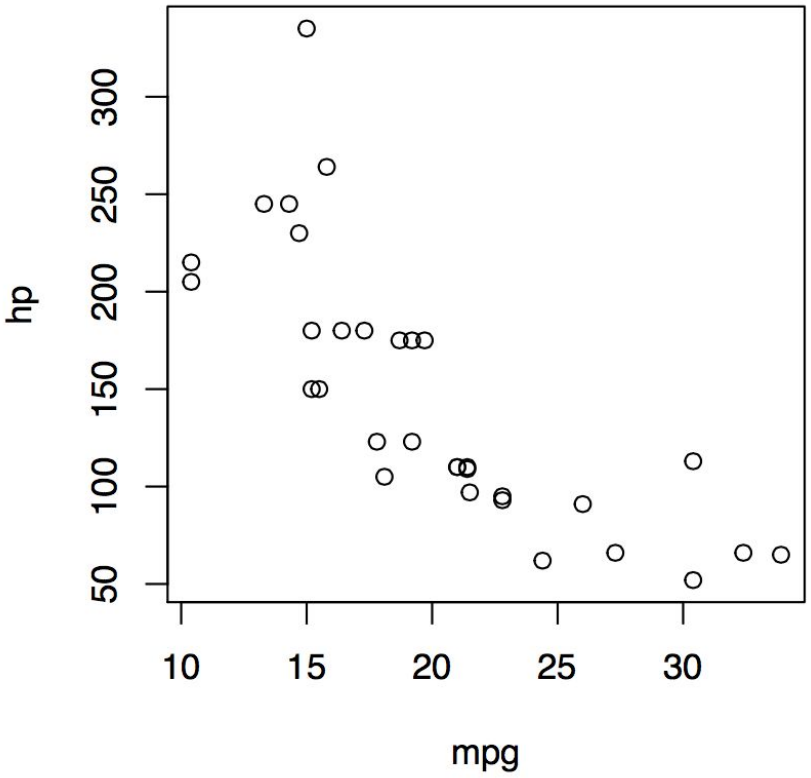
Default appearance of plots carefully chosen

Designed with visual perception in mind

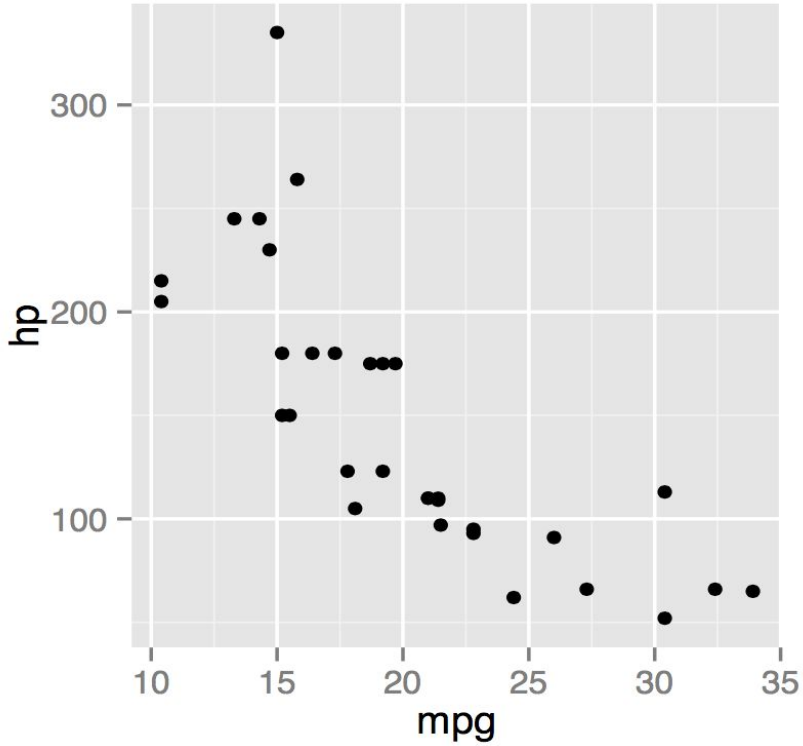
Inclusion of some components, like legends, are automated

Great flexibility for annotating, editing, and embedding output

base graphics



ggplot2



About ggplot2

"ggplot2" is the name of the package (don't forget the 2)

The *gg* in ggplot2 stands for Grammar of Graphics

Inspired in the Grammar of Graphics by Lee Wilkinson

`ggplot()` is the main function in "ggplot2"

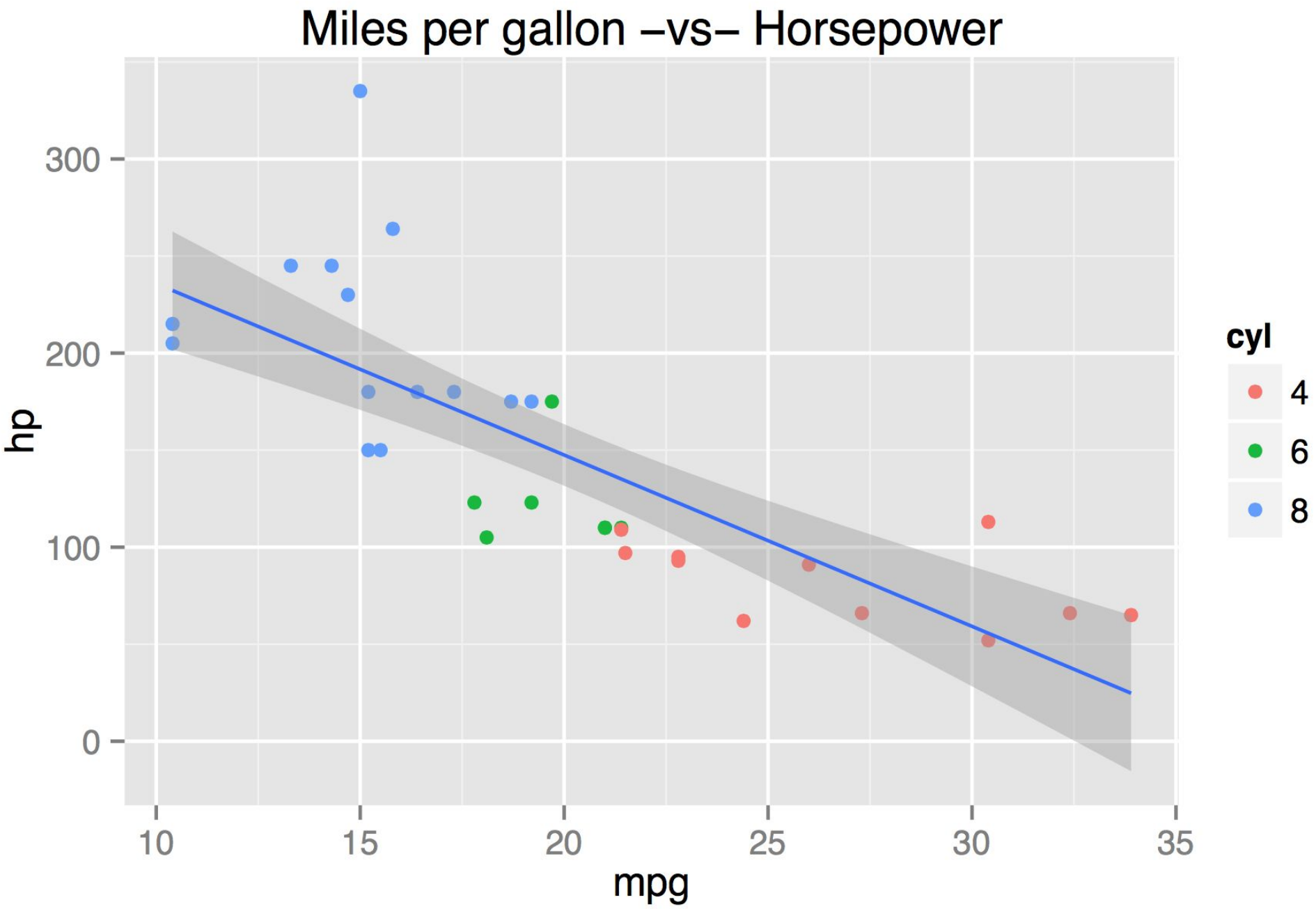
ggplot2 philosophy:

Describe a wide range of graphics with a compact syntax and independent components

What is a Statistical Graphic?

Data set mtcars

	mpg	hp	cyl
Mazda RX4	21.0	110	6
Mazda RX4 Wag	21.0	110	6
Datsun 710	22.8	93	4
Hornet 4 Drive	21.4	110	6
Hornet Sportabout	18.7	175	8
Valiant	18.1	105	6
Duster 360	14.3	245	8
Merc 240D	24.4	62	4
Merc 230	22.8	95	4
Merc 280	19.2	123	6


Miles per gallon –vs– Horsepower
300
200
100
0
hp
10
15
20
25
30
35
mpg
cyl
4
6
8

Elements to draw the chart “manually”

Coordinate system

x and y axes

Axis tick marks

Axis labels, and title

Points (with colors)

Regression line (and ribbon)

Legend

A statistical graphic is ...

A mapping from data to aesthetic attributes (color, shape, size) of geometric objects (points, lines, bars)

A plot may also contain statistical transformations of the data

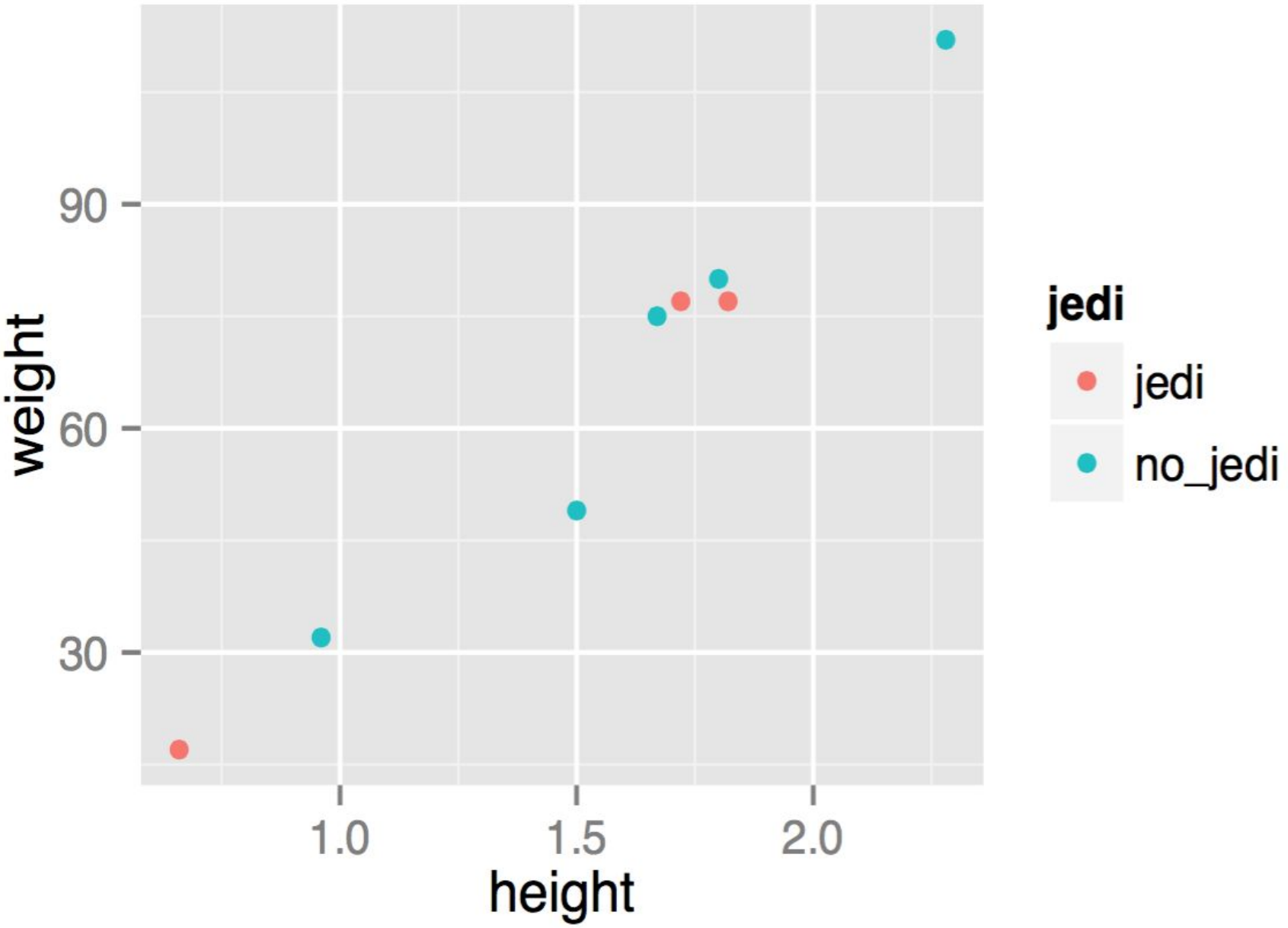
A plot is drawn on a specific coordinate system

Sometime faceting can be used to get the same plot for different subsets of the dataset

Example

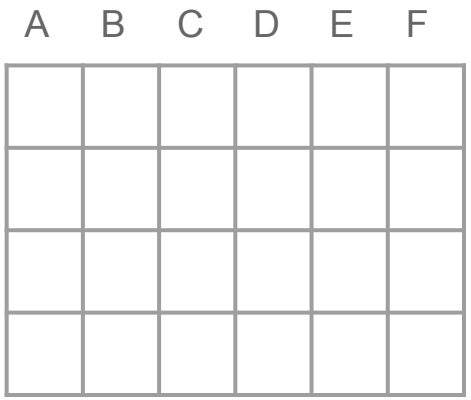
name	gender	height	weight	jedi	species	weapon
Luke Skywalker	male	1.72	77	jedi	human	lightsaber
Leia Skywalker	female	1.5	49	no_jedi	human	blaster
Obi-Wan Kenobi	male	1.82	77	jedi	human	lightsaber
Han Solo	male	1.8	80	no_jedi	human	blaster
R2-D2	male	0.96	32	no_jedi	droid	unarmed
C-3PO	male	1.67	75	no_jedi	droid	unarmed
Yoda	male	0.66	17	jedi	yoda	lightsaber
Chewbacca	male	2.28	112	no_jedi	wookiee	bowcaster

Let's use these variables to make a scatterplot

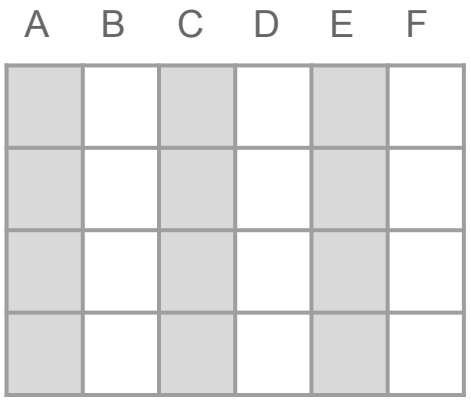

weight
90
60
30
1.0
1.5
2.0
height
jedi
jedi
no_jedi

How does it work?

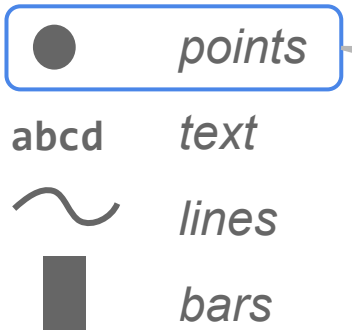
1 Dataset



2 Which variables



3 Which Geometric objects



4 Which Aesthetic attributes

x = A
y = B
color = C
size = *default*
shape = *default*

Building a scatterplot

Dataset: starwars

Variables: height, weight, jedi

Geoms: points

Aesthetic (perceptive attributes):

- X-axis: height
- Y-axis: weight
- Color: jedi

Scatterplot with ggplot2

```
ggplot(data = starwars) +
    geom_point(aes(x = height, y = weight, color = jedi)
```

`ggplot()` initializes a "ggplot" object

You specify the data set (data frame) with **`data`**

`geom_point()` indicates the type of geometric object

You use **`aes()`** to map aesthetic attributes to variables:

- X-position: height
- Y-position: weight
- Color: jedi

Automated things in ggplot2

- Axis labels
- Legends (positions, labels, symbols)
- Choice of colors for points
- Background color (i.e. gray)
- Grid lines (major and minor)
- Axis tick marks

You can always override the default settings (this is the tricky part in ggplot2)

Mapping

data values

height	weight	jedi
1.72	77	jedi
1.50	49	no_jedi
1.82	77	jedi
1.80	80	no_jedi
0.96	32	no_jedi
1.67	75	no_jedi
0.66	17	jedi
2.28	112	no_jedi

mapping

aesthetic attributes

x	y	color
x_1	y_1	#F8766D
x_2	y_2	#00BFC4
x_3	y_3	#F8766D
x_4	y_4	#00BFC4
x_5	y_5	#00BFC4
x_6	y_6	#00BFC4
x_7	y_7	#F8766D
x_8	y_8	#00BFC4

These values are meaningful to us, but not to the computer

They need to be converted from data units to physical units that the computer can display

Main elements

A graphic is a mapping from data to aesthetic attributes (color, shape, size) of geometric objects (points, lines, bars, etc)

```
ggplot(data, ...)
```

```
aes()
```

```
geom_objects()
```

How does ggplot2 work?

Plots are created piece-by-piece

Plot components added with + operator

Aesthetic attributes mapped to data values

Computation of scales for aesthetic attributes

The data MUST BE in a data frame!

Always ask

What is the data set of interest?

What variables (columns) will be used to make the plot?

What graphic shapes (geoms) will be used to display the data?

What features of the shapes will be used to represent the data values?

Warning

ggplot2 comes with the function `qplot()` (i.e. quick plot)

Avoid using it!

As Karthik Ram says: “you’ll end up unlearning and relearning a good bit”