Informatics 143

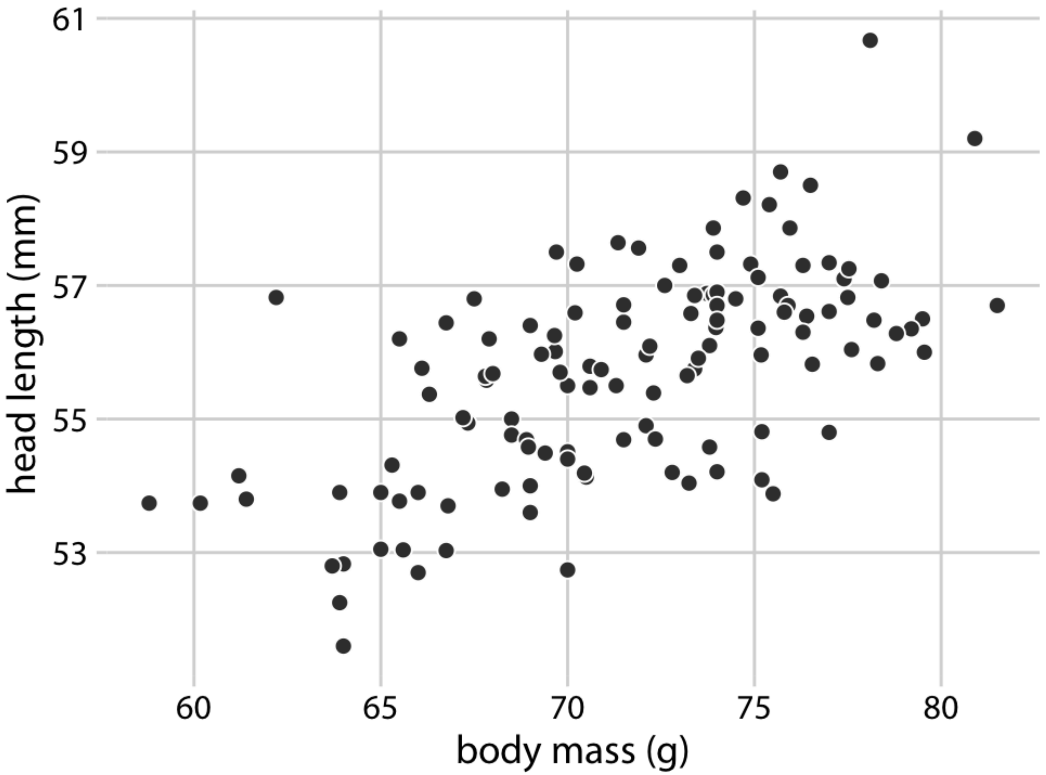
Information Visualization

Lecture 10

These course materials are based on books from Claus O. Wilke, Kieran Healy, Edward R. Tufte, Alberto Cairo, Colin Ware, Tamara Munzner, and others.
Powerpoint theme by Prof. André van der Hoek.

Visualization of relations between variables

- Useful to visualize how variables relate to each other in multidimensional datasets



Visualization of time series

- Useful to visualize how variables relate to each other in multidimensional datasets **and when one of the variables can be thought as time**

Visualization of time series

- Useful to visualize how variables relate to each other in multidimensional datasets **and when one of the variables can be thought as time**

- Data has:
  - At least two set of values and one of them can be thought as time

- *Some* standard geometrical mappings:
  - Standard relation mappings (e.g. scatterplots)
  - Line graphs
  - Connected scatterplots
  - Smoothed trendlines

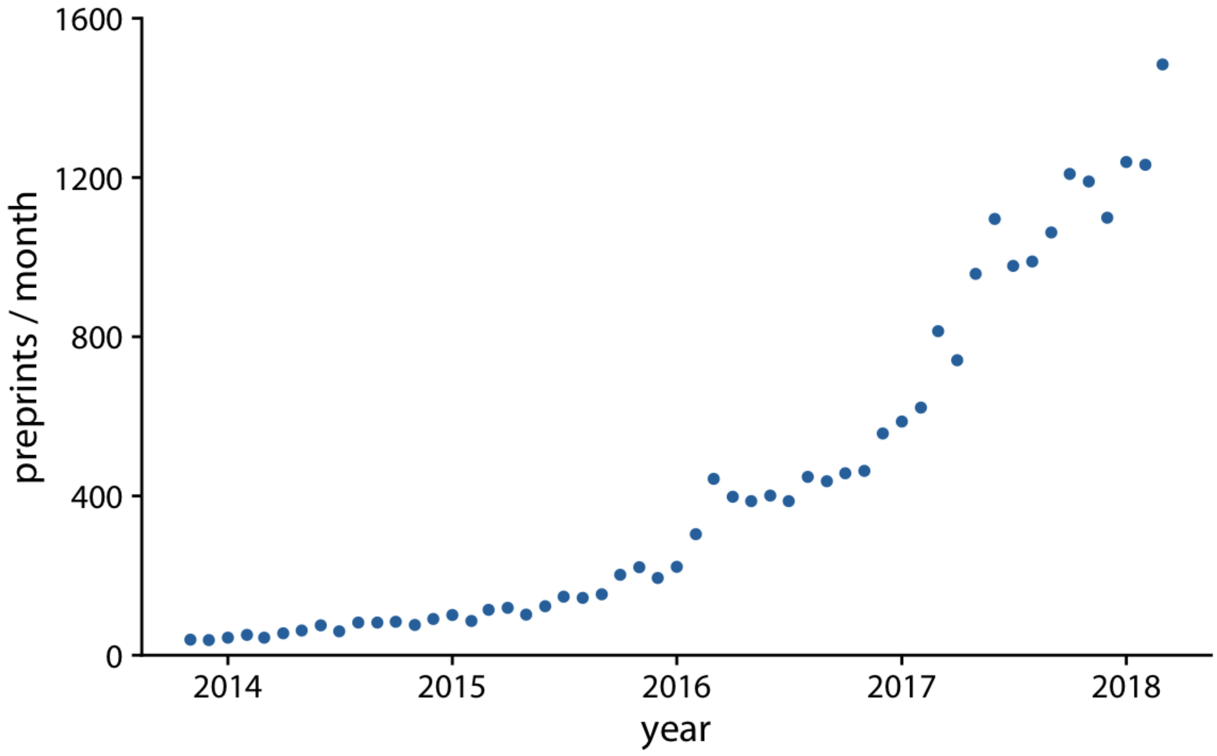
Visualization of time series

- Important notion:

Time imposes structure on the data: there is a natural order

Visualization of time series

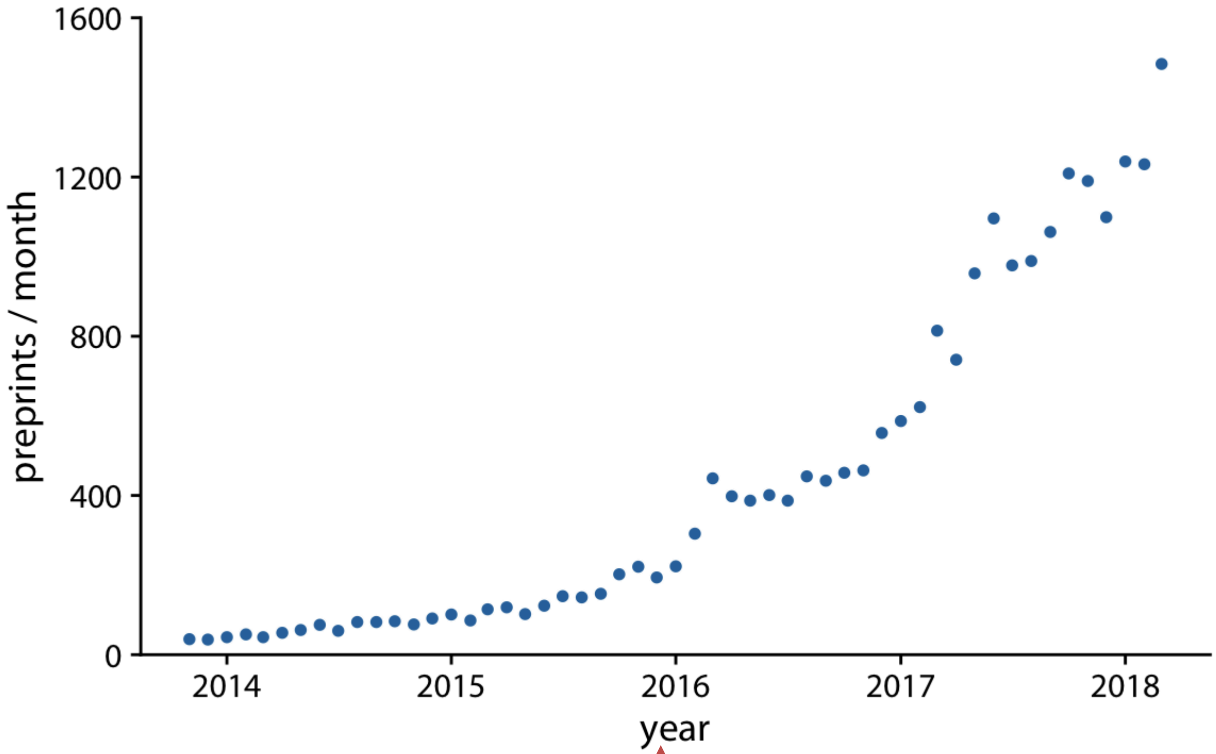
- Using scatterplots



Data source: Jordan Anaya, http://www.prepubmed.org/

Visualization of time series

- Using scatterplots

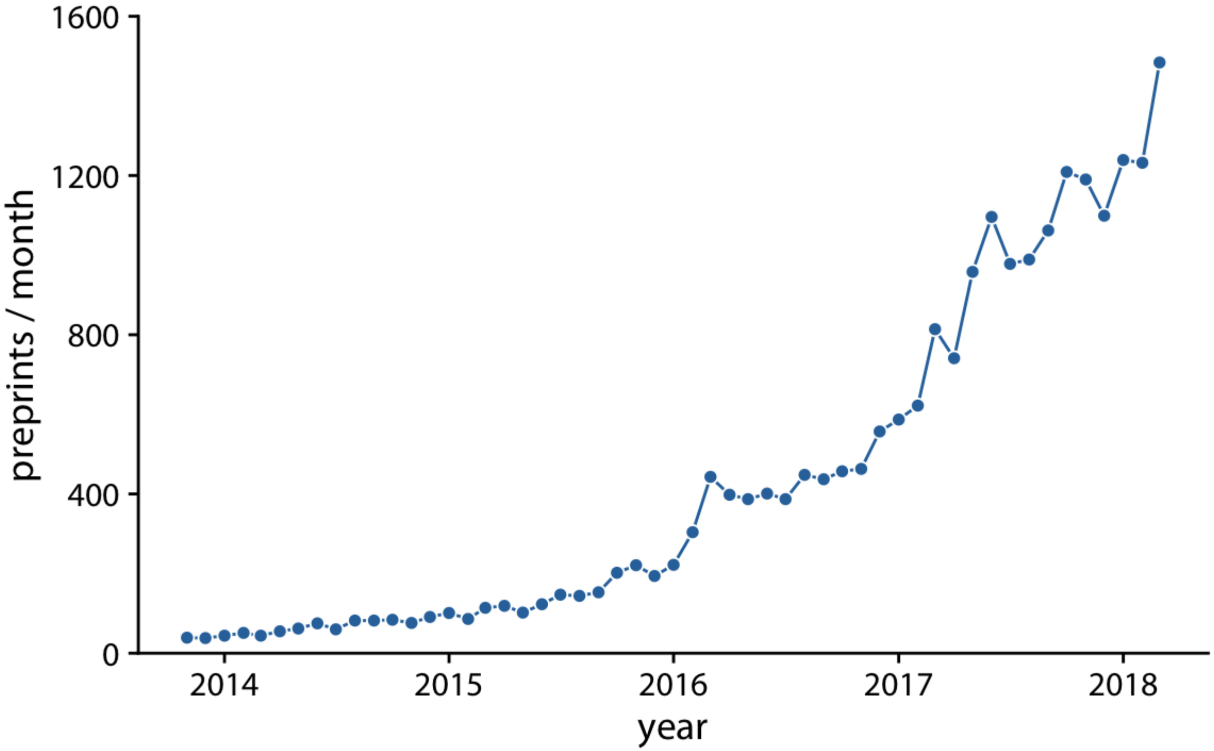


There is a natural and expected order due to the “time arrow”

Data source: Jordan Anaya, http://www.prepubmed.org/

Visualization of time series

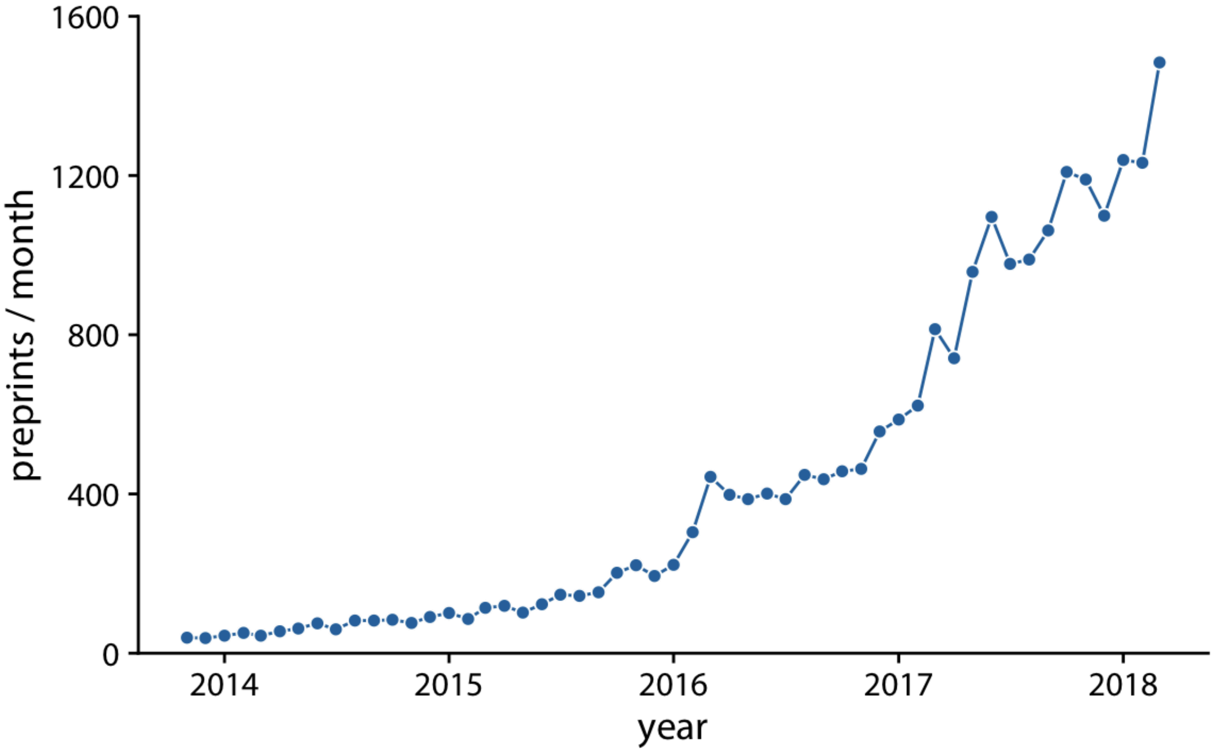
- Using a line graph



Data source: Jordan Anaya, http://www.prepubmed.org/

Visualization of time series

- Using a line graph

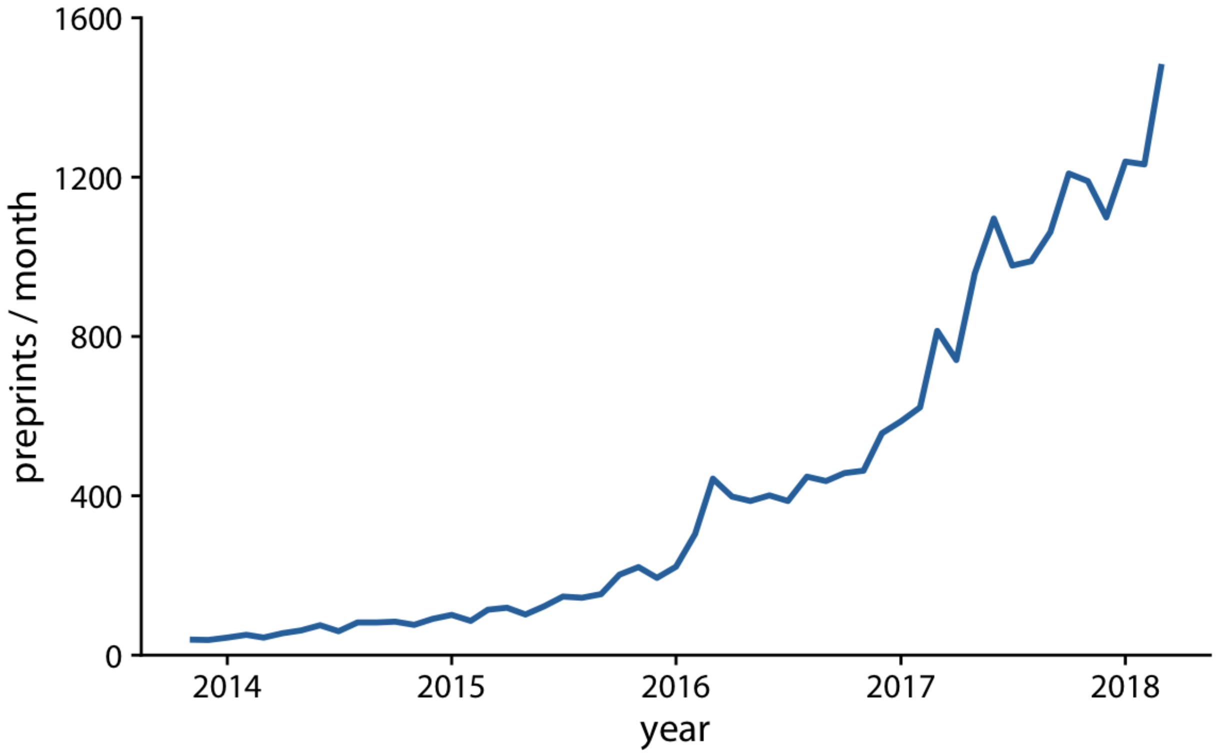


The lines between points do not represent observed data!
They are just a perception aid!

Data source: Jordan Anaya, http://www.prepubmed.org/

Visualization of time series

- Using a line graph

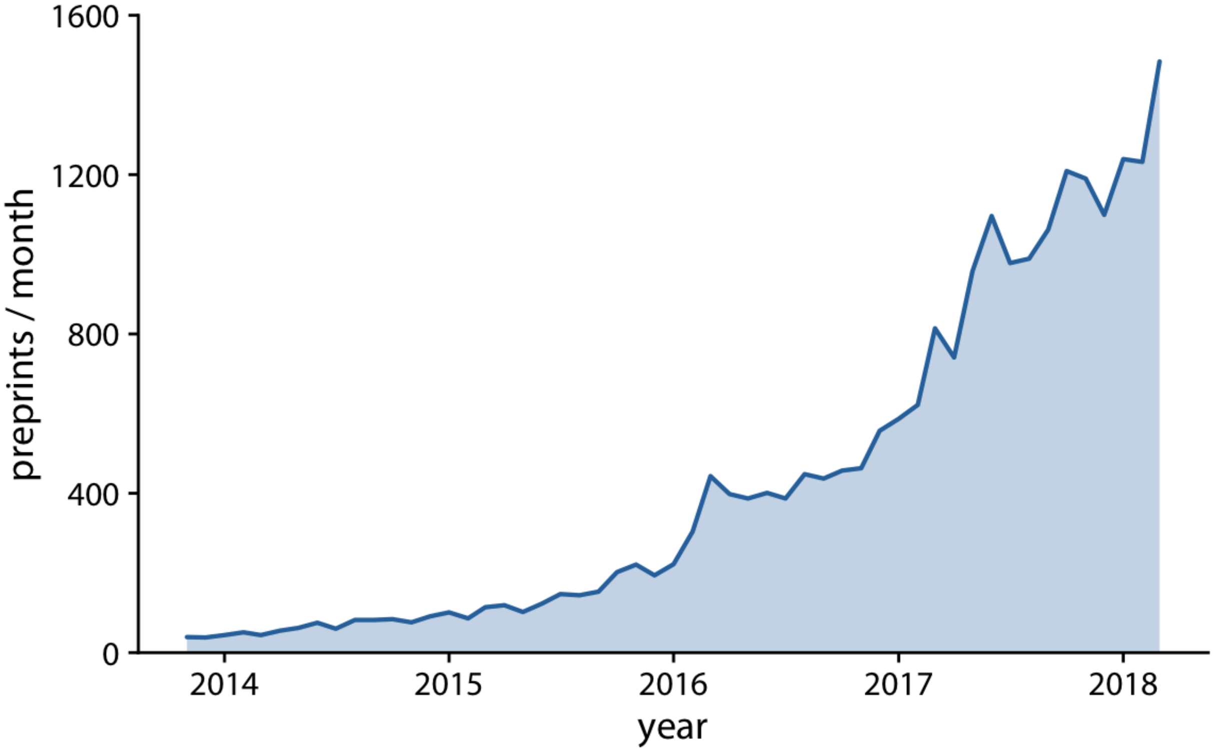


Sometimes points are hidden.
This is only acceptable if you are dealing with a large amount of data points!

Data source: Jordan Anaya, http://www.prepubmed.org/

Visualization of time series

- Using a line graph



Filling the area under the curve emphasizes the major trend

Data source: Jordan Anaya, http://www.prepubmed.org/

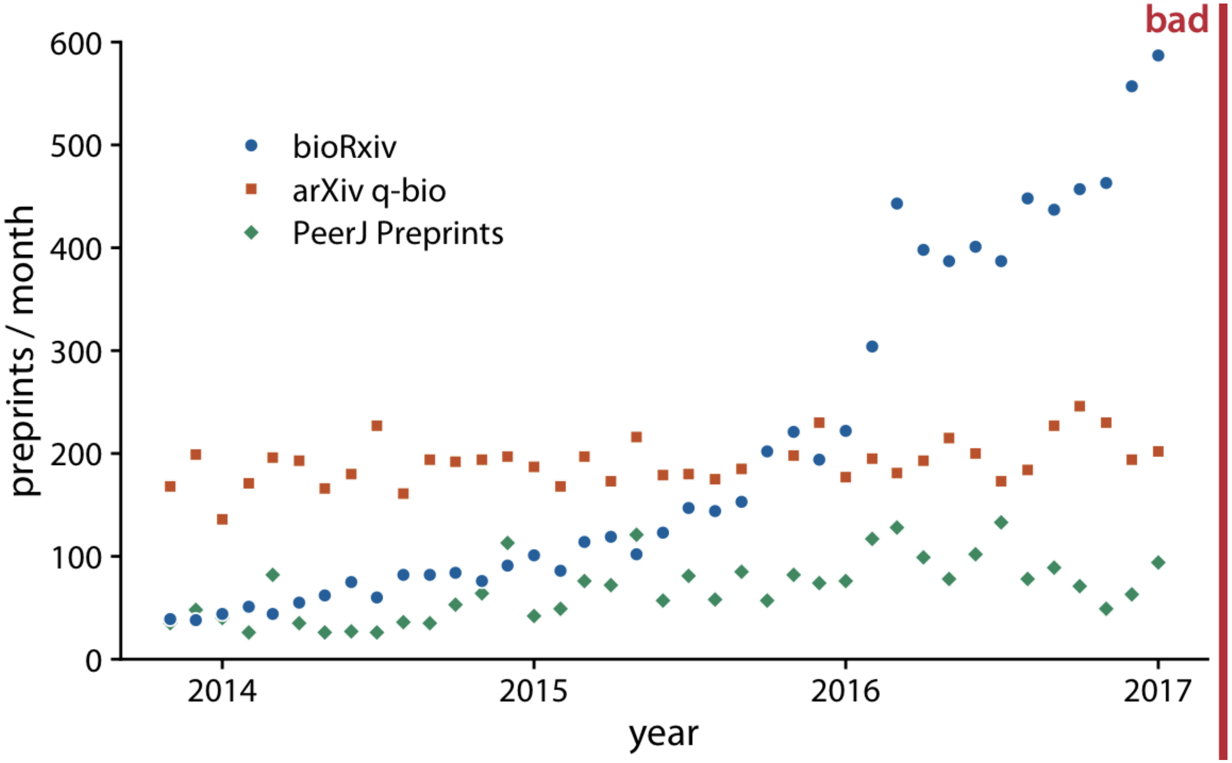
Visualization of multiple time series

- The need to visually represent the variation of multiple classes within the variable along the time is common

- This brings additional challenges
  – Direct adoption of scatterplots should be strongly avoided

Visualization of multiple time series

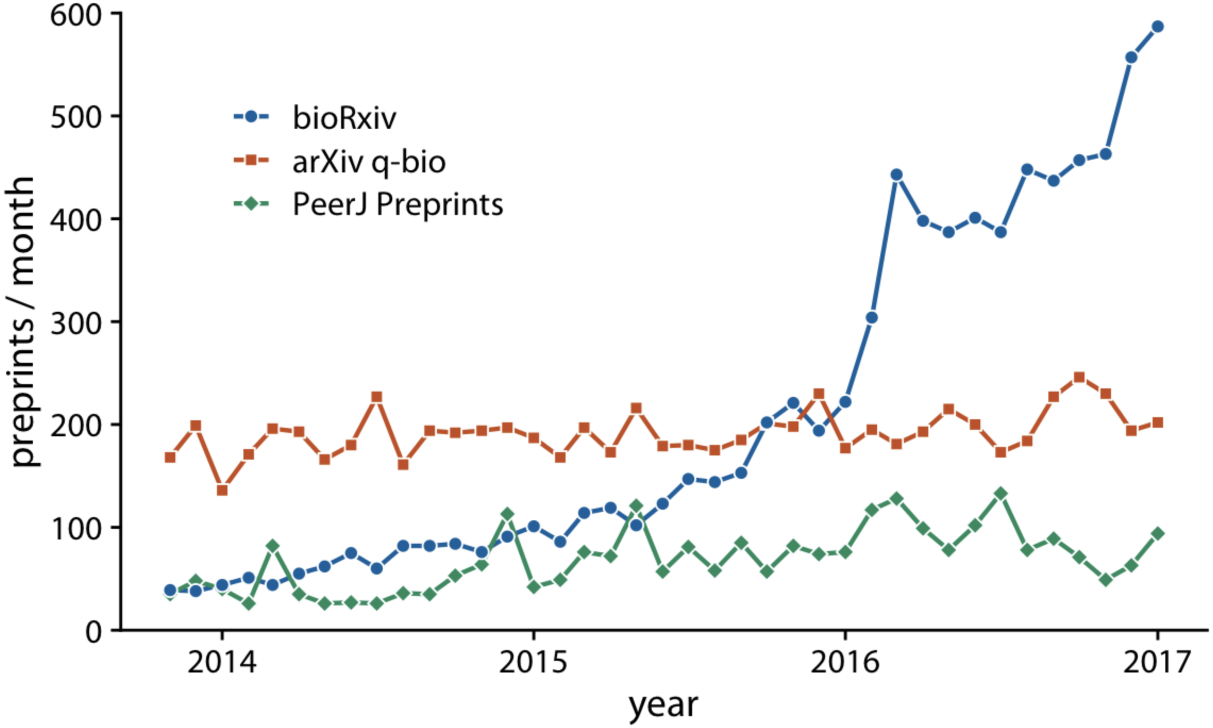
- This brings additional challenges
  - Direct adoption of scatterplots should be strongly avoided



Data source: Jordan Anaya, http://www.prepubmed.org/

Visualization of multiple time series

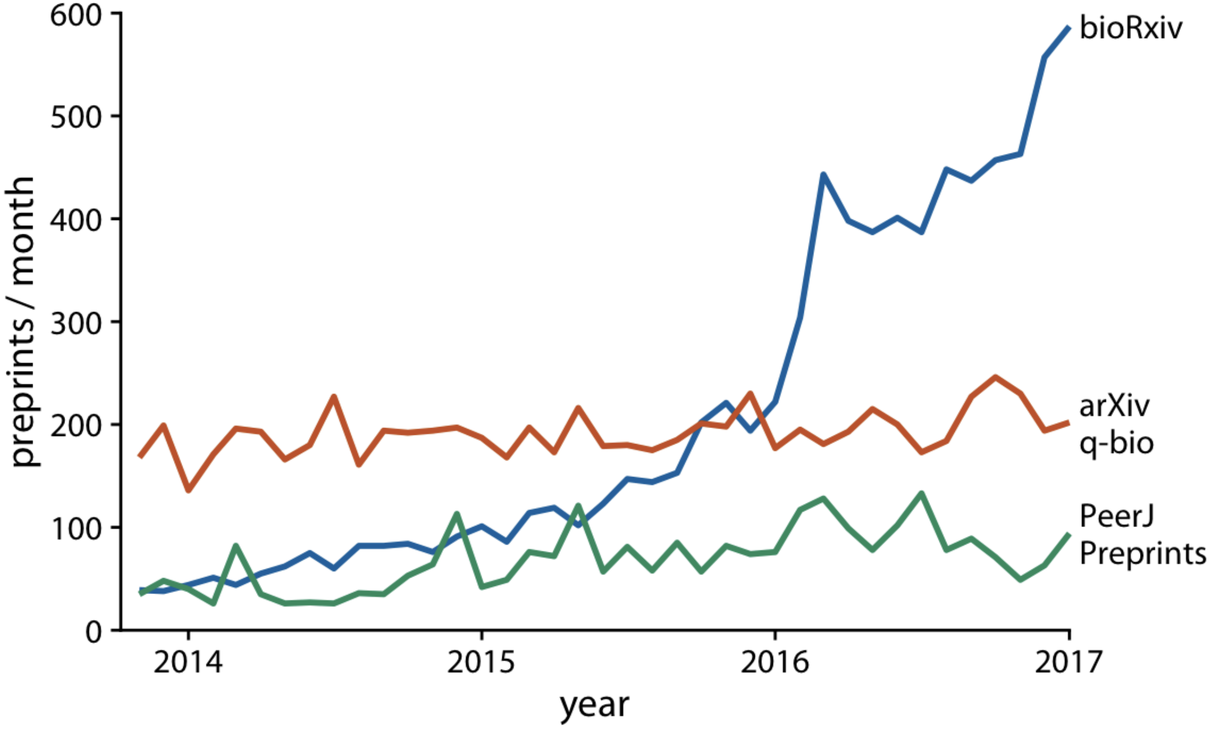
- This brings additional challenges
  - Direct adoption of scatterplots should be strongly avoided
  - But connecting the dots help to guide the eye



Data source: Jordan Anaya, http://www.prepubmed.org/

Visualization of multiple time series

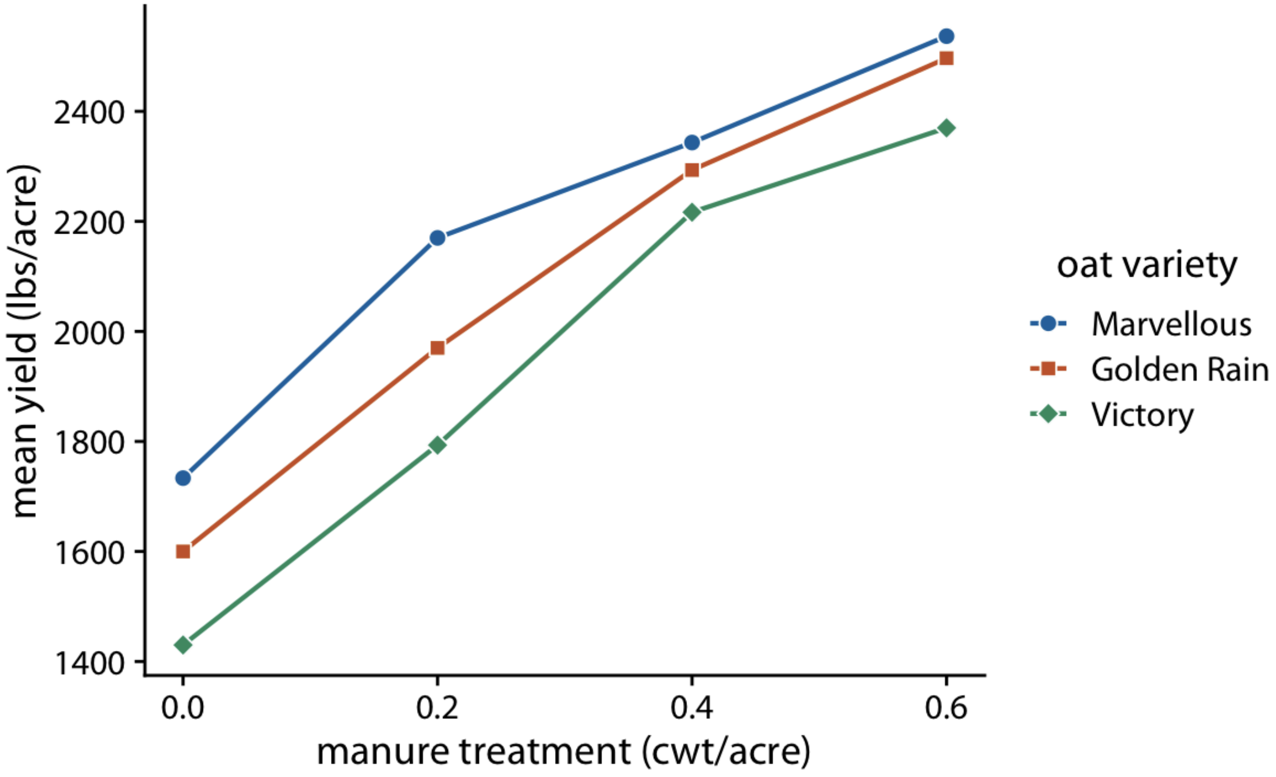
- If enought data is available, removing the dots and removing the separate legend helps to reduce cognitive load



Data source: Jordan Anaya, http://www.prepubmed.org/

Visualization of multiple time series

- Linegraphs
  - Widely used for time series
  - **But also for any data that has intrinsic ordering**



Data soure: Yates (1935)

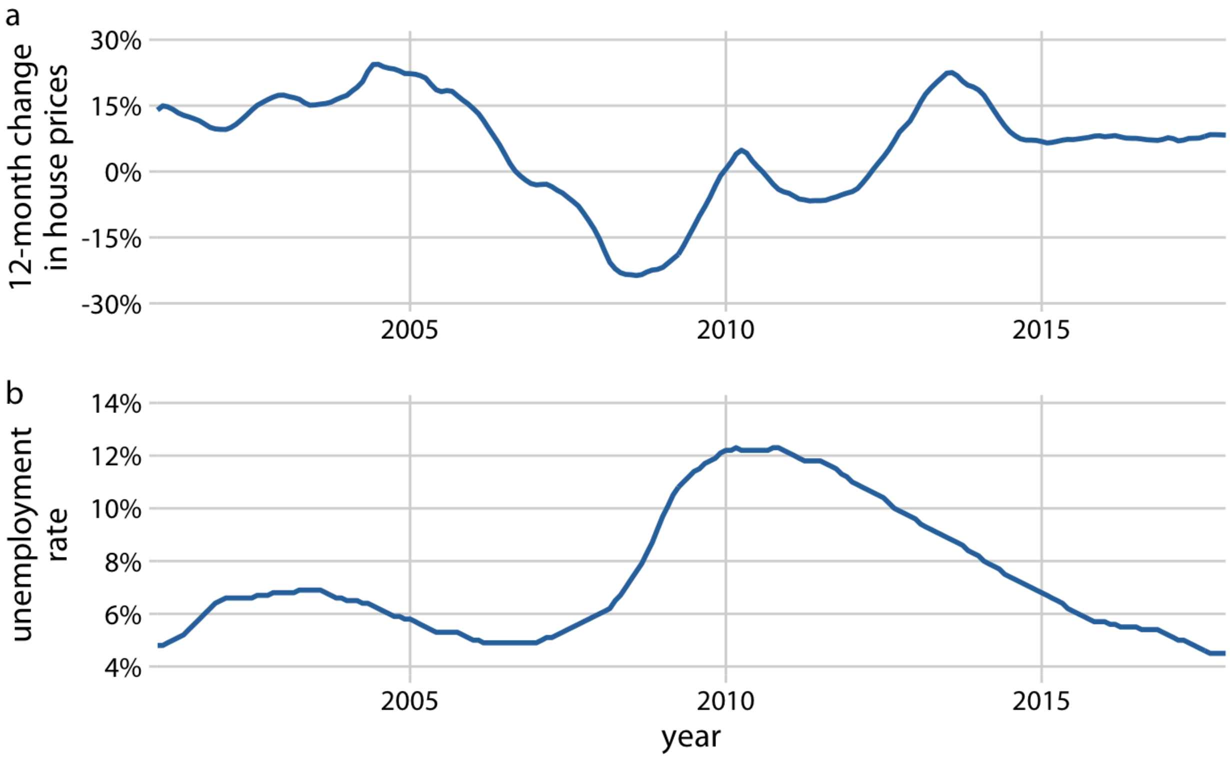
Visualization of multiple time series

- What to do when you are facing **multiple variables?**
  - E.g. the change in house prices **and** in unemployment rates in function of time

- A common visual representation is to adopt two (or more) separate line graphs **sharing the time axis**

Visualization of multiple time series

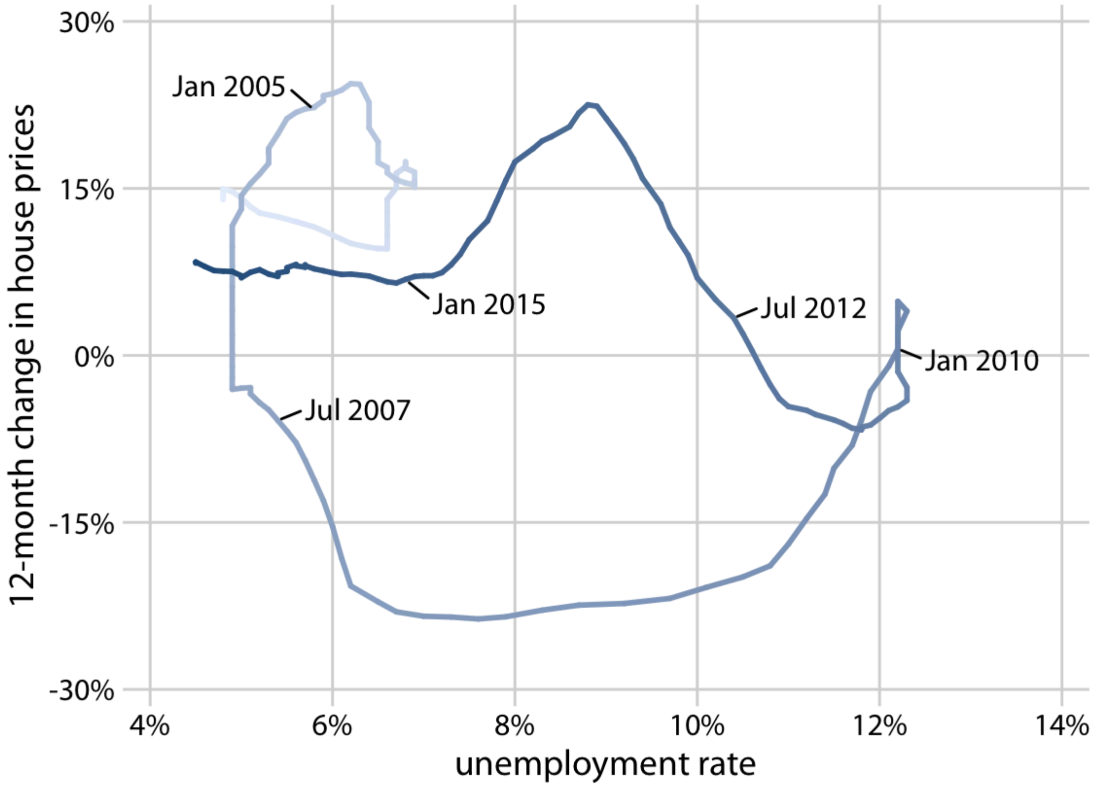
- A common visual representation is to adopt two (or more) separate line graphs sharing the time axis



Data sources: Freddie Mac House Prices Index, U.S. Bureau of Labor Statistics.

Visualization of multiple time series

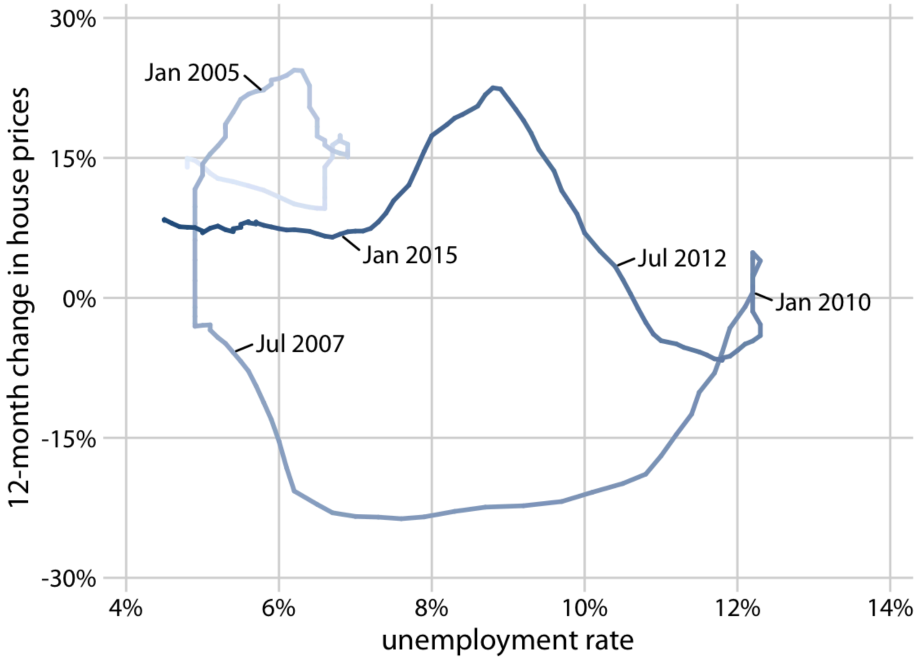
- **Alternative** representation *connected scatter plot*
  – Essentially a time annotated curve, in this case, also colored by date



Data sources: Freddie Mac House Prices Index, U.S. Bureau of Labor Statistics.

Visualization of multiple time series

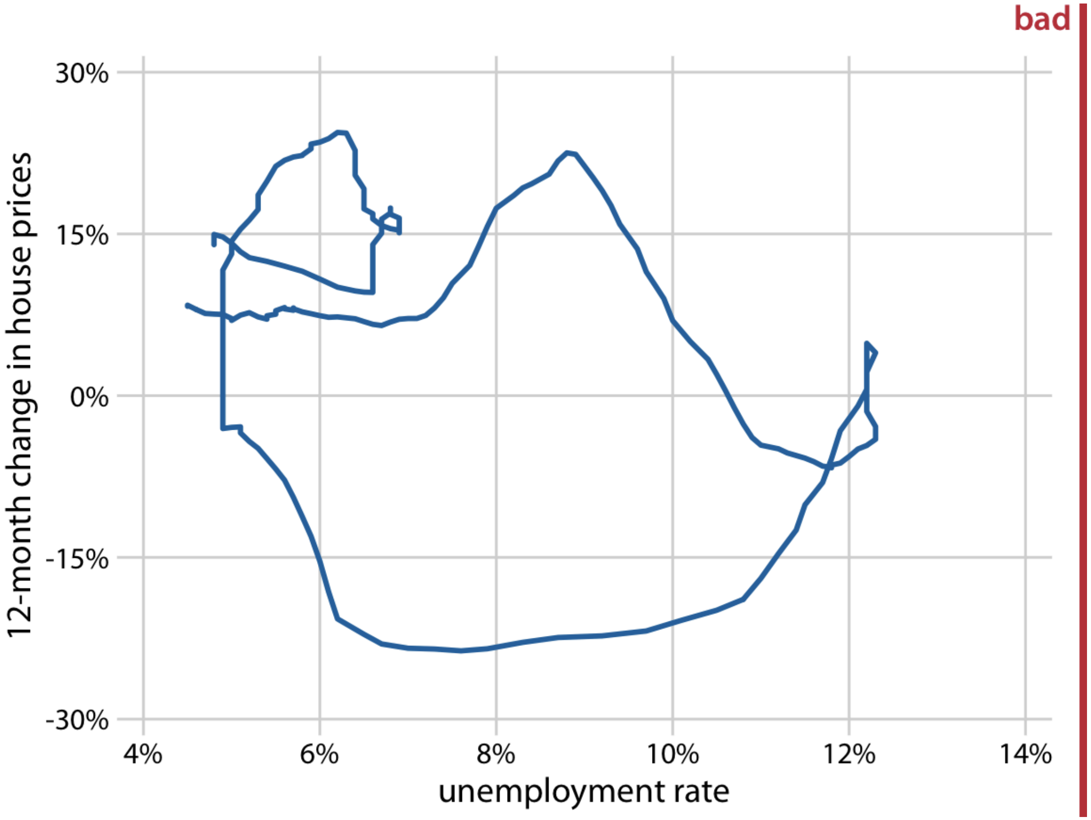
- Alternative representation *connected scatter plot*
  - Essentially a time annotated curve (in this case, also colored by date)
  - Lines from lower left to upper right indicate correlations
  - Lines from upper left to lower right represent anti-correlation



Data sources: Freddie Mac House Prices Index, U.S. Bureau of Labor Statistics.

Visualization of multiple time series

- Alternative representation *connected scatter plot*

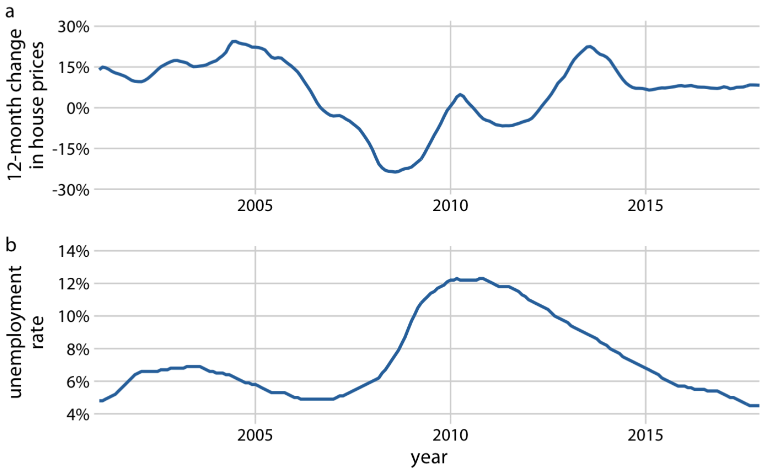


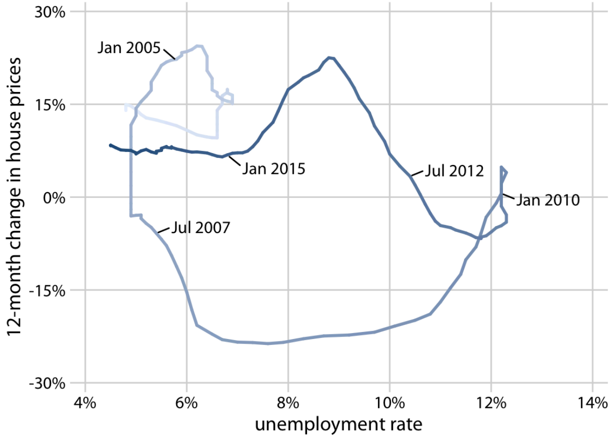
The time dimension must be explicit!

Data sources: Freddie Mac House Prices Index, U.S. Bureau of Labor Statistics.

Visualization of multiple time series

- Is one better than the other?

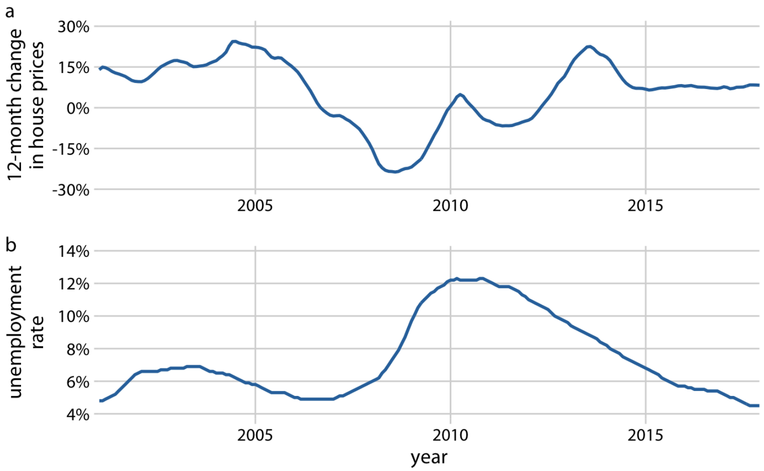


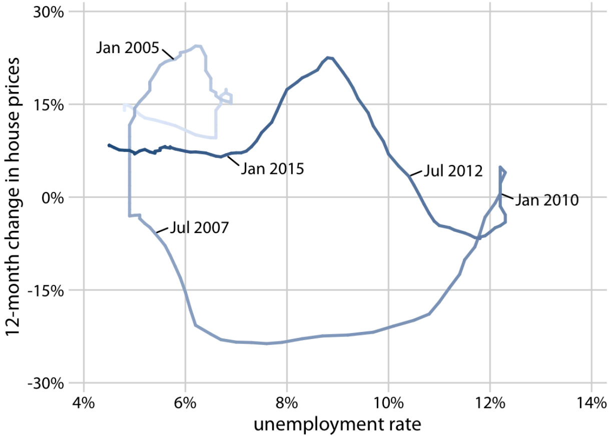


Data sources: Freddie Mac House Prices Index, U.S. Bureau of Labor Statistics.

Visualization of multiple time series

- Is one better than the other?



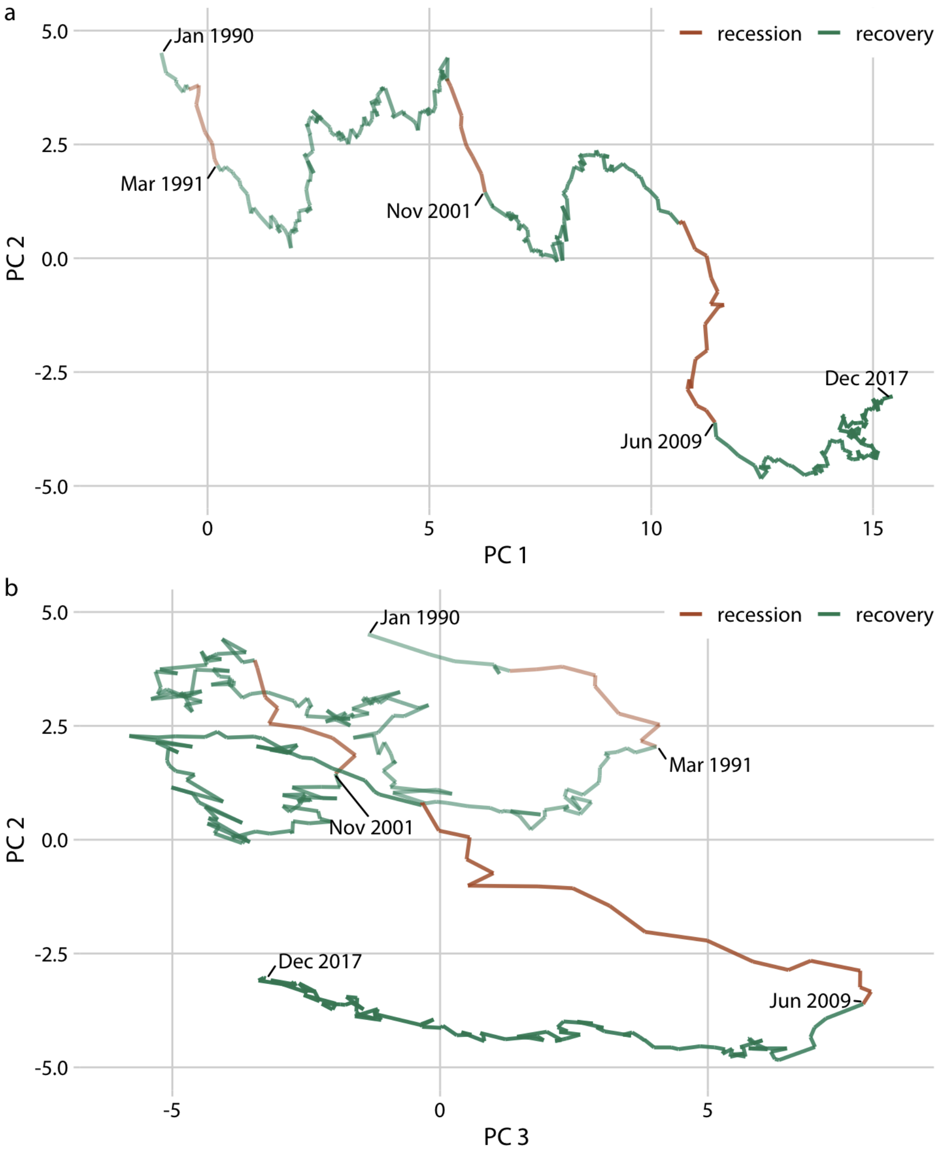


- No, it depends on the public and the story.

Data sources: Freddie Mac House Prices Index, U.S. Bureau of Labor Statistics.

Visualization of multiple time series

- Connected scatter plots can be effectively used to visualize high dimensional datasets by using PCA
  - This example shows the first PCs from the PCA of 100 macroeconomic indicators



Data source: M. W. McCracken, St. Louis Fed

Visualization of time series

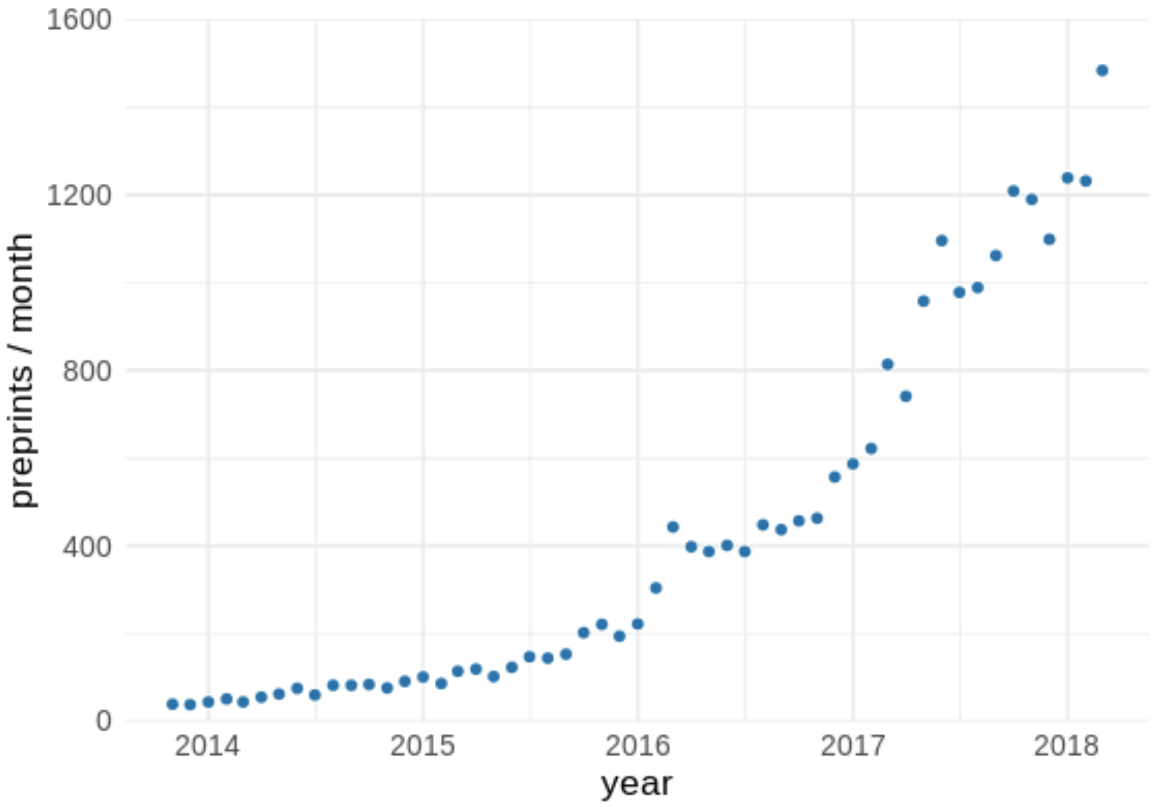
- How to build line graphs?
  - Use geom_line()
  - **And add geom_points() if you wish to keep the actual data points**

```
preprint_growth <-
read.csv("https://www.ics.uci.edu/~algol/teaching/informatics143w2021/preprint_growth.csv")
biorxiv_growth <- preprint_growth[which(preprint_growth$archive=="bioRxiv" &
                                        preprint_growth$count>0),]
```

```
ggplot(biorxiv_growth, aes(x=as.Date(date), y=count)) +
  geom_point(color = "white", fill = "#0072B2", shape = 21, size = 2) +
  scale_y_continuous(limits = c(0, 1600), expand = c(0, 0),
                     name = "preprints / month") +
  scale_x_date(name = "year") +
  theme_minimal() + theme(text = element_text(size=13))
```

Visualization of time series

- How to build line graphs?
  - Use geom_line()
  - **And add geom_points() if you wish to keep the actual data points**



Visualization of time series

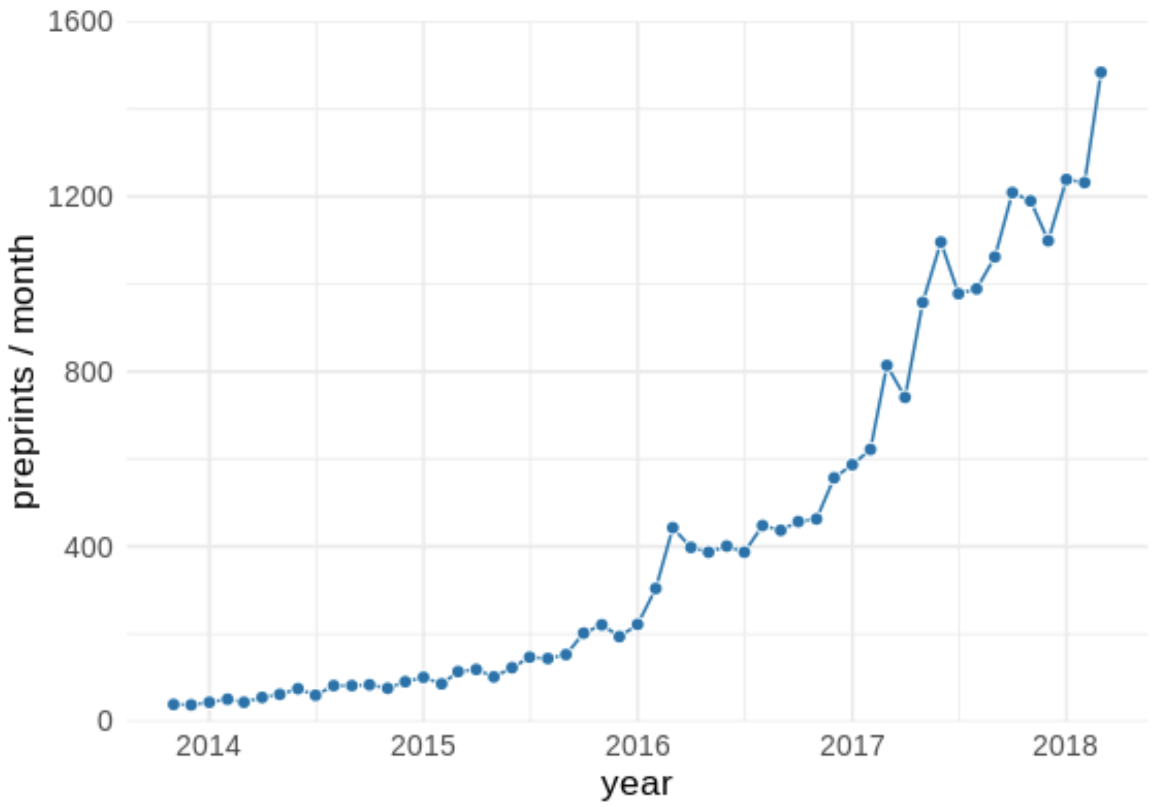
- How to build line graphs?
  - Use geom_line()
  - And add geom_points() if you wish to keep the actual data points

```
preprint_growth <-
read.csv("https://www.ics.uci.edu/~algol/teaching/informatics143w2021/preprint_growth.csv")
biorxiv_growth <- preprint_growth[which(preprint_growth$archive=="bioRxiv" &
                                        preprint_growth$count>0),]
```

```
ggplot(biorxiv_growth, aes(x=as.Date(date), y=count)) +
  geom_line(size = 0.5, color = "#0072B2") +
  geom_point(color = "white", fill = "#0072B2", shape = 21, size = 2) +
  scale_y_continuous(limits = c(0, 1600), expand = c(0, 0),
                     name = "preprints / month") +
  scale_x_date(name = "year") +
  theme_minimal() + theme(text = element_text(size=13))
```

Visualization of time series

- How to build line graphs?
  - **Use geom_line()**
  - And add geom_points() if you wish to keep the actual data points



Visualization of time series

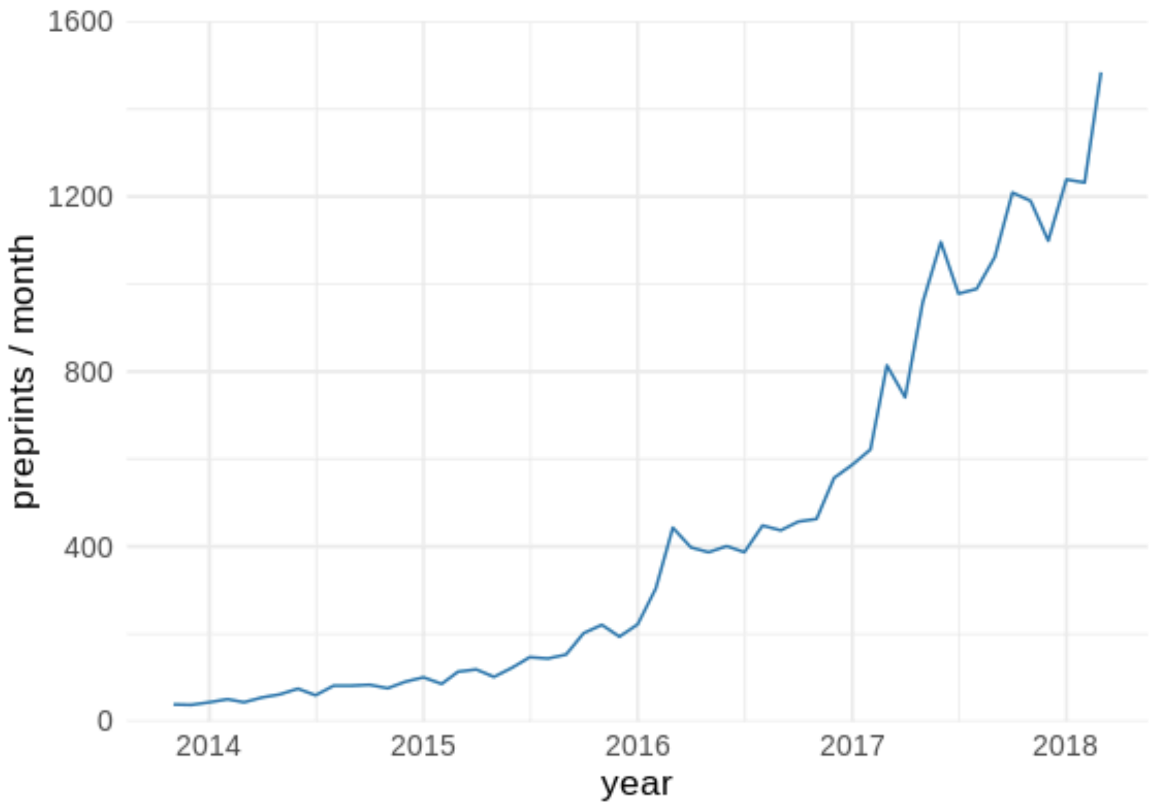
- How to build line graphs?
  - – You can remove the points **if you have enough data**...

```
preprint_growth <-
read.csv("https://www.ics.uci.edu/~algol/teaching/informatics143w2021/preprint_growth.csv")
biorxiv_growth <- preprint_growth[which(preprint_growth$archive=="bioRxiv" &
                                        preprint_growth$count>0),]
```

```
ggplot(biorxiv_growth, aes(x=as.Date(date), y=count)) +
  geom_line(size = 0.5, color = "#0072B2") +
  scale_y_continuous(limits = c(0, 1600), expand = c(0, 0),
                     name = "preprints / month") +
  scale_x_date(name = "year") +
  theme_minimal() + theme(text = element_text(size=13))
```

Visualization of time series

- How to build line graphs?
  – You can remove the points **if you have enough data**...



Visualization of time series

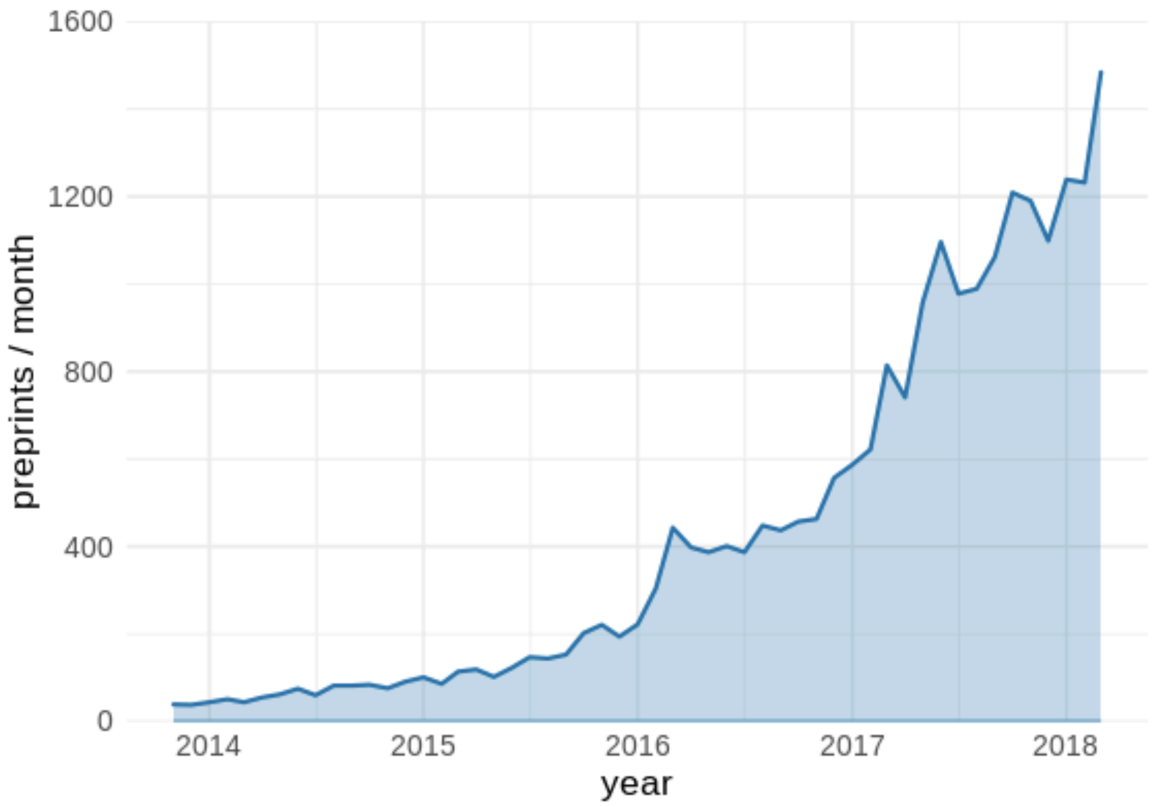
- How to build line graphs?
  - To create the filled region, you can use geom_ridgeline() from the ggridges package. Remember to set the height aesthetics and set y to 0.

```
preprint_growth <-
read.csv("https://www.ics.uci.edu/~algol/teaching/informatics143w2021/preprint_growth.csv")
biorxiv_growth <- preprint_growth[which(preprint_growth$archive=="bioRxiv" &
                                        preprint_growth$count>0),]
require(ggridges)
```

```
ggplot(biorxiv_growth, aes(x=as.Date(date), height=count, y=0)) +
  geom_ridgeline(color = "#0072B2", fill = "#0072B240", size = 0.75) +
  scale_y_continuous(limits = c(0, 1600), expand = c(0, 0),
                     name = "preprints / month") +
  scale_x_date(name = "year") +
  theme_minimal() + theme(text = element_text(size=13))
```

Visualization of time series

- How to build line graphs?
  - To create the filled region, you can use geom_ridgeline() from the ggridges package. Remember to set the height aesthetics and set y to 0.



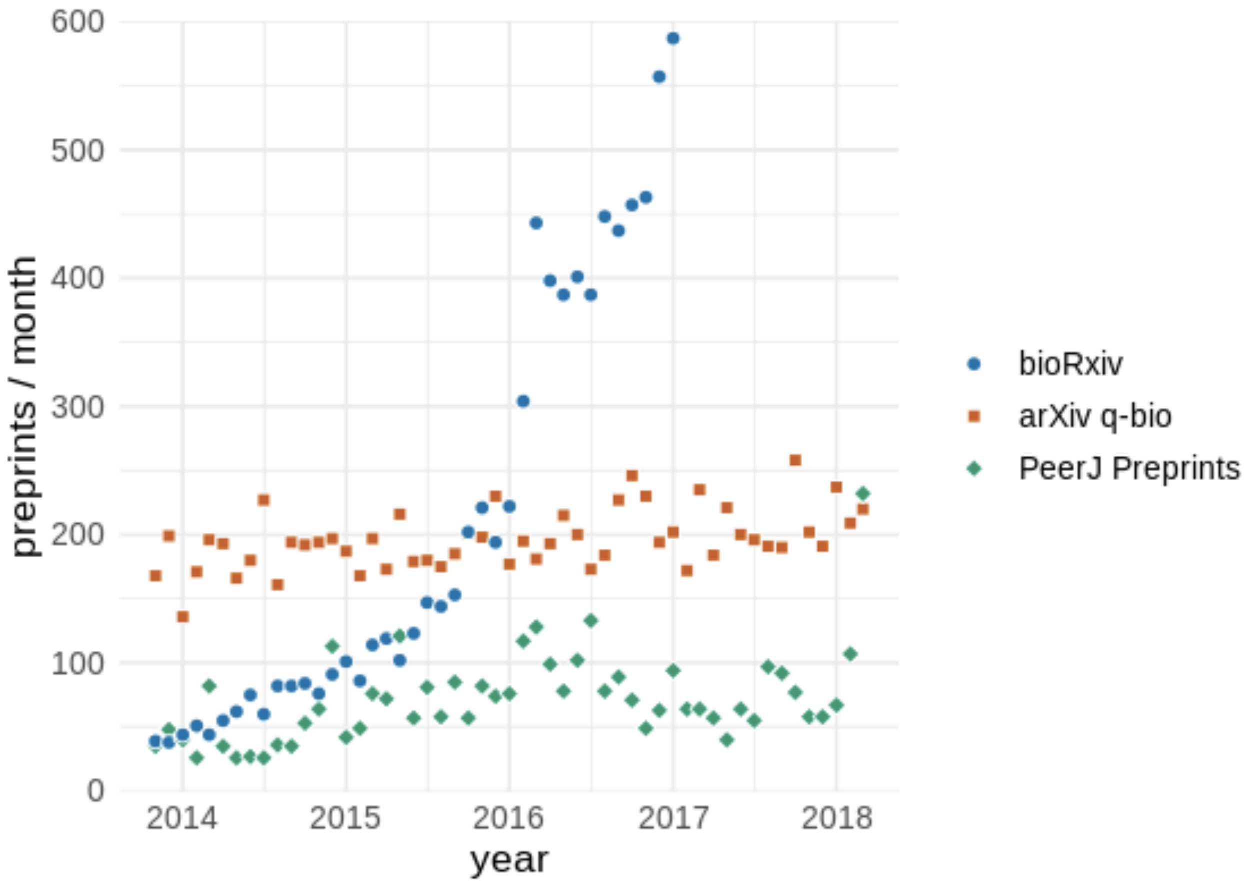
Visualization of time series

- How to build line graphs with multiple series?
  - **Just map the different classes (as factors) into aesthetics like color / shape / etc.**

```
preprint_growth <-
read.csv("https://www.ics.uci.edu/~algol/teaching/informatics143w2021/preprint_growth.csv")
biorxiv_growth <- preprint_growth[which(preprint_growth$archive=="bioRxiv" &
                                        preprint_growth$count>0),]
require(ggridges)

preprints <- preprint_growth[which(
         preprint_growth$archive %in% c("bioRxiv", "arXiv q-bio", "PeerJ Preprints")
         & preprint_growth$count>0),]
preprints$archive <- factor(preprints$archive,
                            levels = c("bioRxiv", "arXiv q-bio", "PeerJ Preprints"))
preprints$date <- as.Date(preprints$date)

ggplot(preprints, aes(x=date, y=count, color = archive, fill = archive,
                      shape = archive)) +
  geom_point(color = "white", size = 2) +
  scale_shape_manual(values = c(21, 22, 23), name = NULL) +
  scale_y_continuous(limits = c(0, 600), expand = c(0, 0),
                     name = "preprints / month") +
  scale_x_date(name = "year", limits =
                   range(preprints$date[which(preprints$archive=="bioRxiv")])) +
  scale_color_manual(values = c("#0072b2", "#D55E00", "#009e73"), name = NULL) +
  scale_fill_manual(values = c("#0072b2", "#D55E00", "#009e73"), name = NULL) +
  theme_minimal() + theme(text = element_text(size=13))
```

Visualization of multiple time series

- How to build line graphs with multiple series?
  - **Just map the different classes (as factors) into aesthetics like color / shape / etc.**



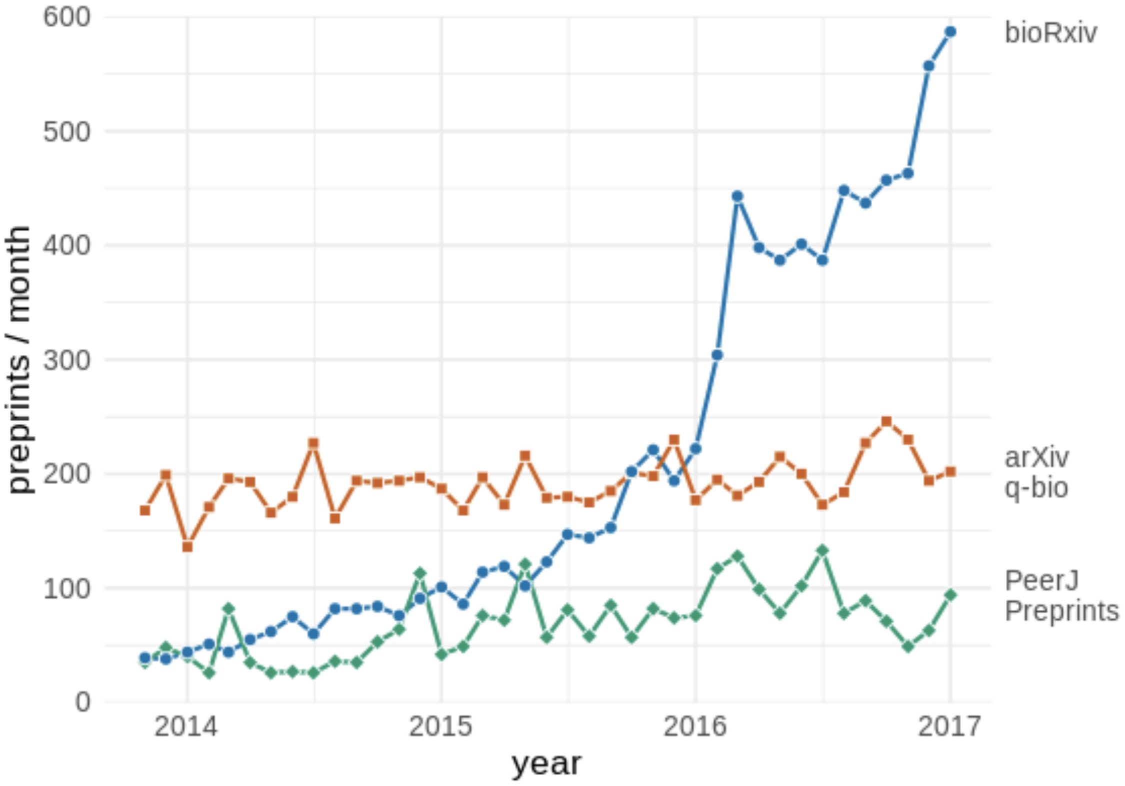
Visualization of multiple time series

- How to build line graphs with multiple series?
  - **To add the lines just add the geom_line() call**
  - **And to set the legents at the last point positions you need to specify a secondary y axis and set the breaks and labels at the correct positions. Also, you need to erase the legend by setting the legend.position = “none” in the theme**

```
ggplot(preprints, aes(x=date, y=count, color = archive, fill = archive,
                      shape = archive)) +
  geom_line(size = 0.75) + geom_point(color = "white", size = 2) +
  scale_shape_manual(values = c(21, 22, 23), name = NULL) +
  scale_y_continuous(limits = c(0, 600), expand = c(0, 0),
                     name = "preprints / month",
                     sec.axis = dup_axis(
      breaks = preprints$count[which(preprints$date==as.Date("2017-01-01"))],
      labels = c("arXiv\nq-bio", "PeerJ\nPreprints", "bioRxiv"),
      name = NULL)) +
  scale_x_date(name = "year",
      limits = c(min(preprints$date[which(preprints$archive=="bioRxiv")]),
                     as.Date("2017-01-01"))) +
  scale_color_manual(values = c("#0072b2","#D55E00","#009e73"), name = NULL) +
  scale_fill_manual(values = c("#0072b2","#D55E00","#009e73"), name = NULL) +
  theme_minimal() + theme(text = element_text(size=13),
                        legend.position = "none")
```

Visualization of multiple time series

- How to build line graphs with multiple series?
  - To add the lines just add the geom_lin() call
  - And to set the legents at the last point positions you need to specify a secondary y axis and set the breaks and labels at the correct positions. Also, you need to erase the legend by setting the legend.position = “none” in the theme



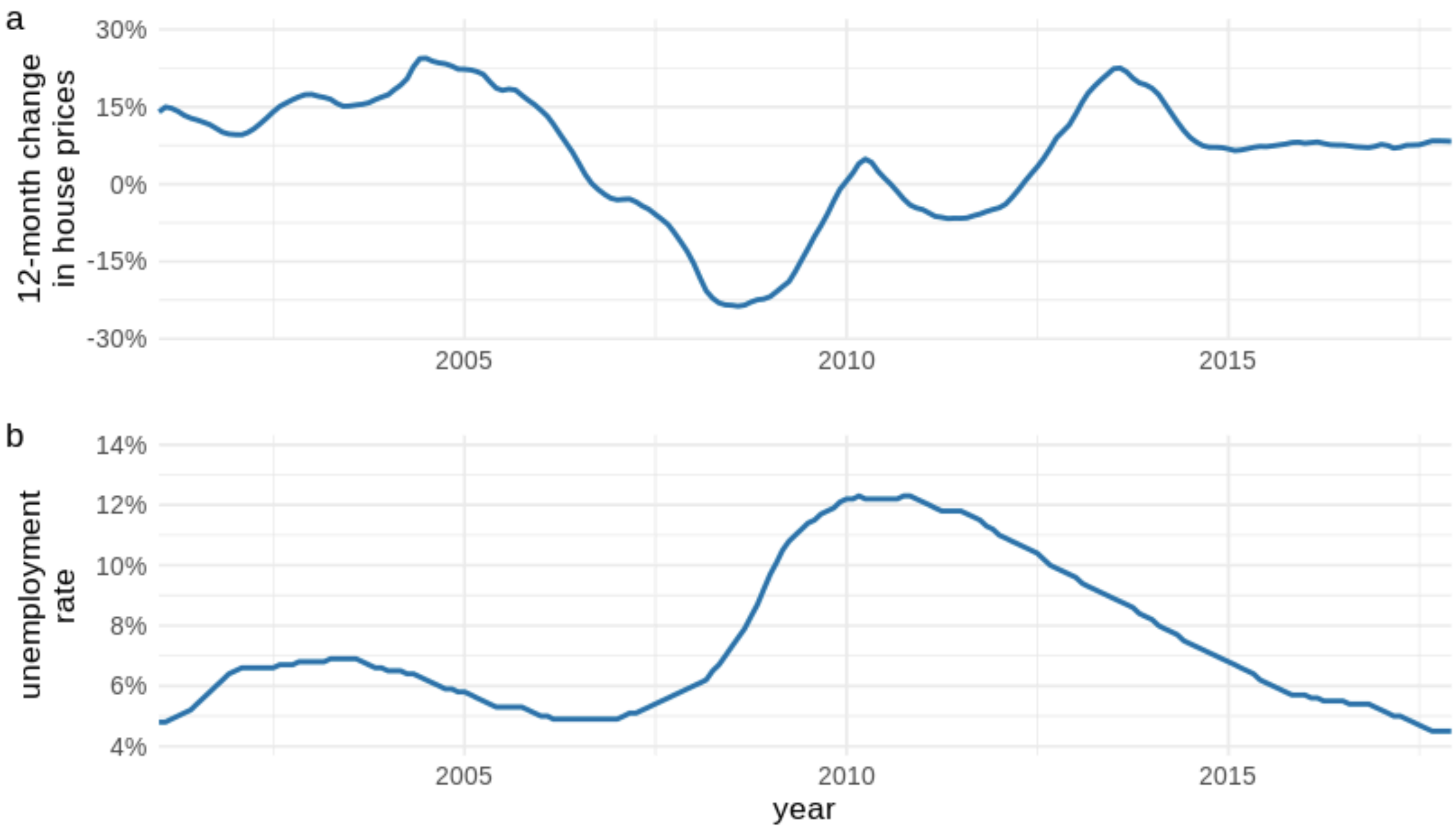
Visualization of multiple time series

- How to build line graphs with multiple series?
  - **Create two ggplots objects and add them to a same visualization**

```
CA_house_prices <-read.csv(
  "https://www.ics.uci.edu/~algol/teaching/informatics143w2021/CA_house_prices.csv")
CA_house_prices$date <- as.Date(CA_house_prices$date)

p1 <- ggplot(CA_house_prices, aes(date, house_price_perc)) +
  geom_line(size = 1, color = "#0072b2") +
  scale_y_continuous(limits = c(-0.3, .32), expand = c(0, 0),
    breaks = c(-.3, -.15, 0, .15, .3),
    name = "12-month change\nin house prices",
    labels = scales::percent_format(accuracy = 1)) +
  scale_x_date(name = "", expand = c(0, 0)) +
  coord_cartesian(clip = "off") +
  theme_minimal() + theme(text = element_text(size=13))

p2 <- ggplot(CA_house_prices, aes(date, unemploy_perc/100)) +
  geom_line(size = 1, color = "#0072b2") +
  scale_y_continuous(limits = c(0.037, 0.143),
    name = "unemployment\nrate",
    labels = scales::percent_format(accuracy = 1),
    expand = c(0, 0)) +
  scale_x_date(name = "year", expand = c(0, 0)) +
  theme_minimal() + theme(text = element_text(size=13))

cowplot::plot_grid(p1, p2, ncol = 1, align = 'v’,
                   labels = 'auto', label_fontface = "plain", hjust = 0, vjust = 1)
```

Visualization of multiple time series

- How to build line graphs with multiple series?
  - **Create two ggplots objects and add them to a same visualization**



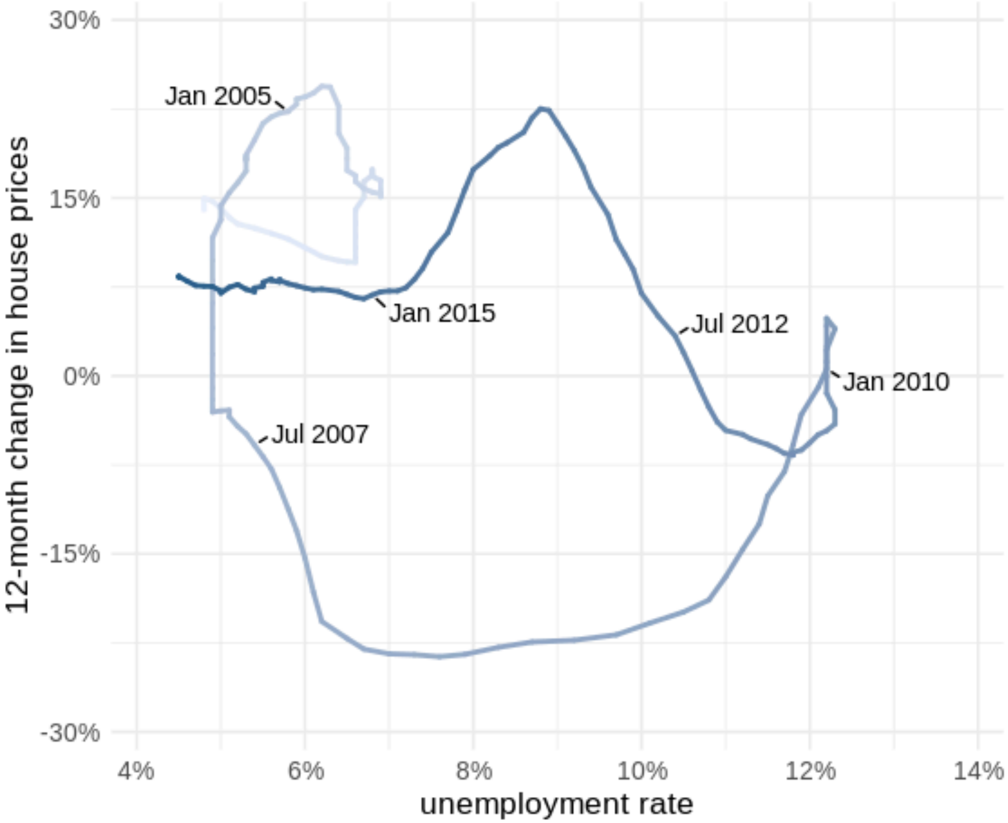
Visualization of multiple time series

- How to build a connected scatterplot?
  - Use geom_path() and set the correct variables (X,Y) and use the date in the color
  - Use ggrepel’s geom_text_repel() to make text repel and to write the dates

```
ggplot(CA_house_prices) +
  aes(unemploy_perc/100, house_price_perc, colour = as.numeric(date)) +
  geom_path(size = 1, lineend = "round") +
  geom_text_repel(
    aes(label = label), point.padding = .2, color = "black",
    min.segment.length = 0, size = 11/.pt,
    hjust = CA_house_prices$hjust,
    nudge_x = CA_house_prices$nudge_x,
    nudge_y = CA_house_prices$nudge_y,
    direction = "y") +
  scale_x_continuous(
    limits = c(0.037, 0.143),
    name = "unemployment rate", labels = scales::percent_format(accuracy = 1),
    expand = c(0, 0)) +
  scale_y_continuous(
    limits = c(-0.315, .315), expand = c(0, 0),
    breaks = c(-.3, -.15, 0, .15, .3),
    name = "12-month change in house prices",
    labels = scales::percent_format(accuracy = 1)) +
  scale_colour_gradient(low = "#E7F0FF", high = "#035B8F") + #"#0072b2") +
  guides(colour = FALSE) +
  coord_cartesian(clip = "off") +
  theme_minimal() + theme(text = element_text(size=13))
```

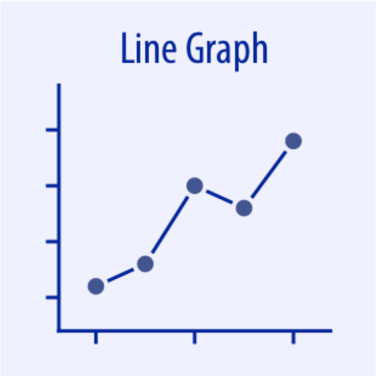
Visualization of multiple time series

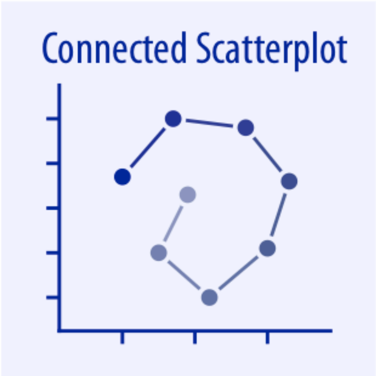
- How to build a connected scatterplot?
  - Use geom_path() and set the correct variables (X,Y) and map the date to color
  - Use ggrepel's geom_text_repel() to make text repel and to write the dates

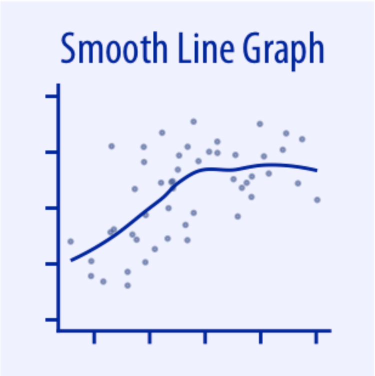


Visualization of relations between variables

- Helping interpretation by adding trends




Connected Scatterplot


Smooth Line Graph