

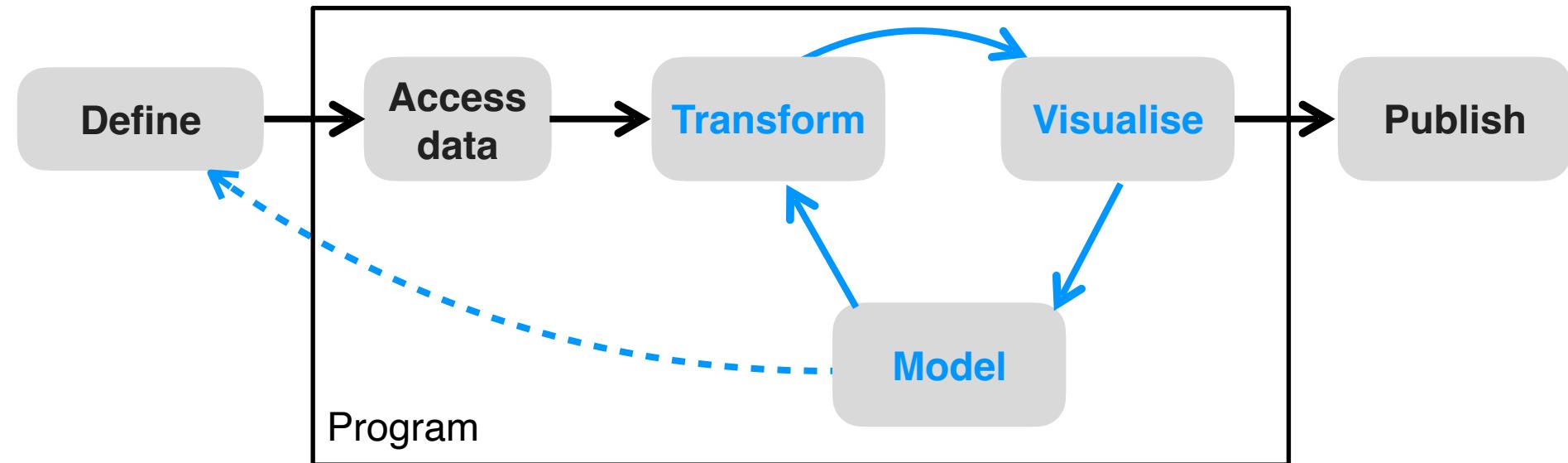


Applied Geodata Science I

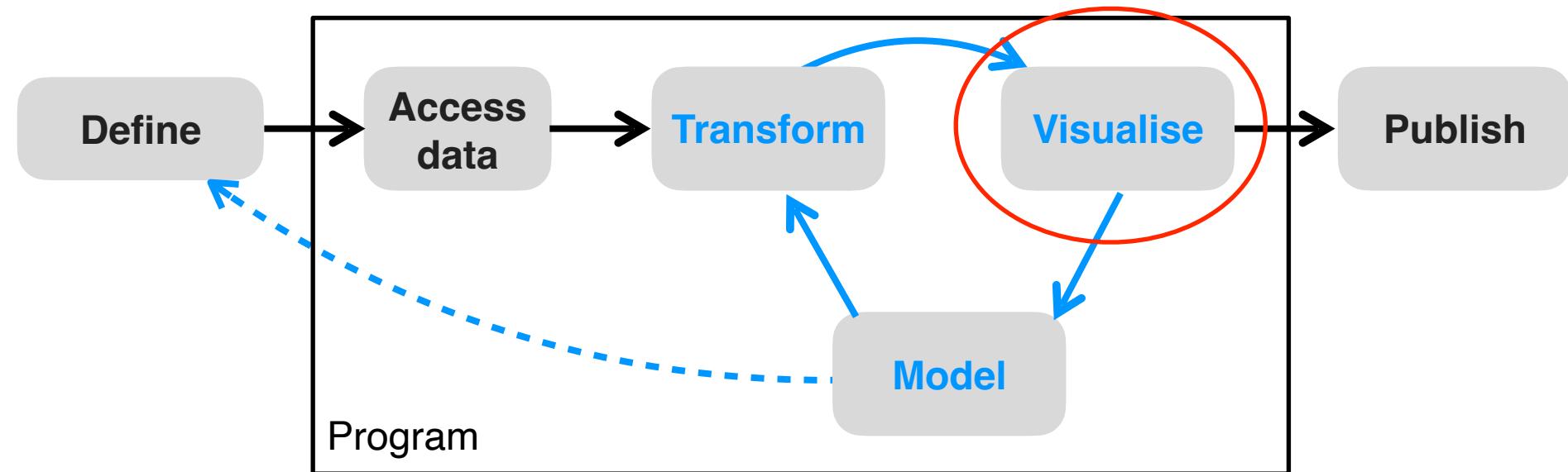
# Session 4

Prof. Dr. Benjamin Stocker  
Spring semester 2023

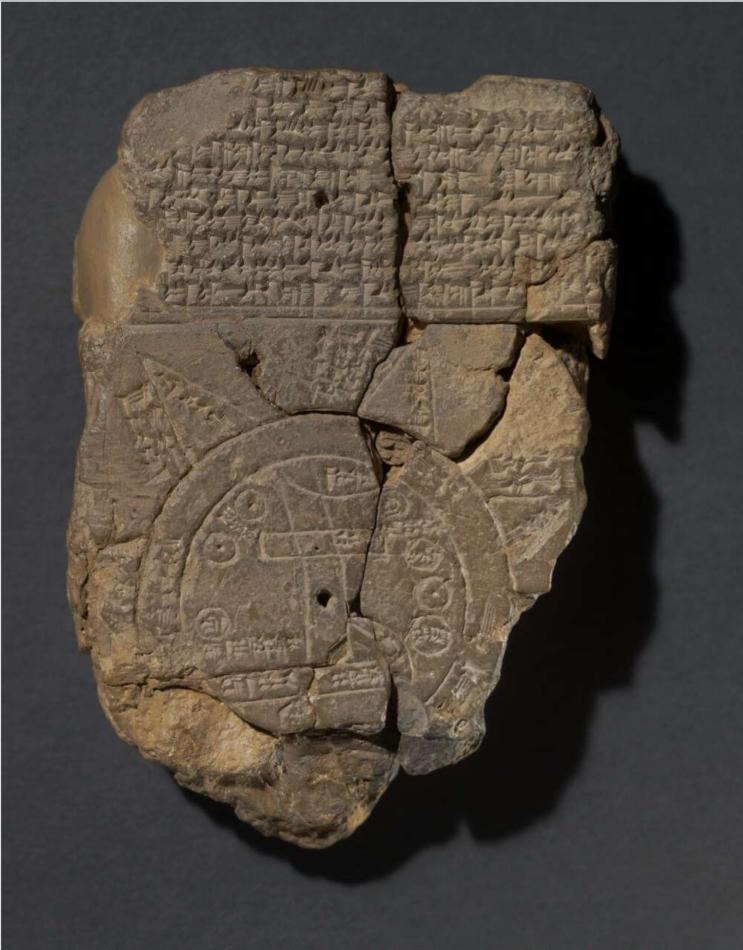
# The data science workflow



# The data science workflow



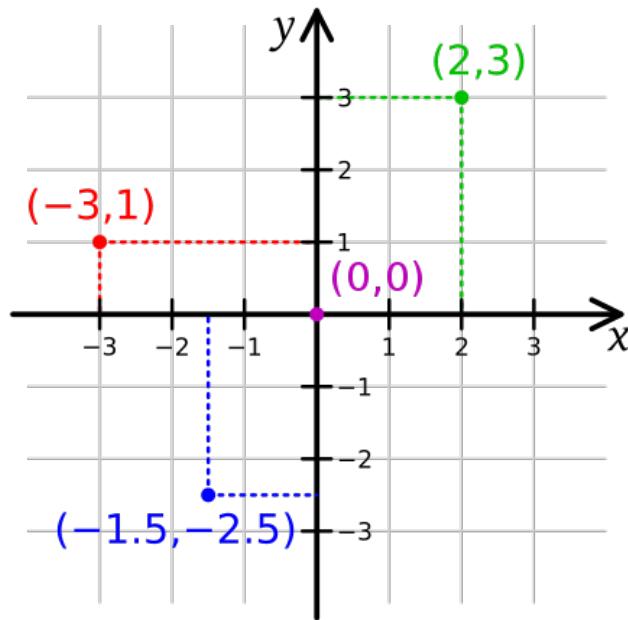
# Earliest data visualisation?



Babylon, ca. 500 BC

[https://www.britishmuseum.org/collection/object/W\\_1882-0714-509](https://www.britishmuseum.org/collection/object/W_1882-0714-509)

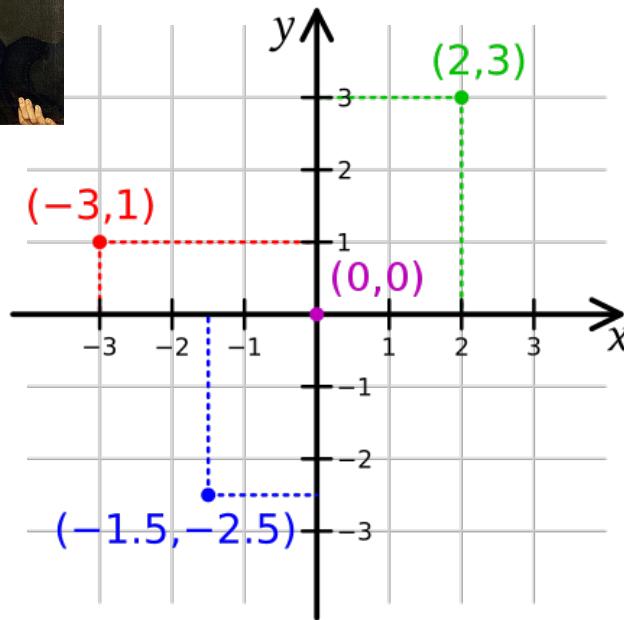
# Cartesian coordinate system



# Cartesian coordinate system



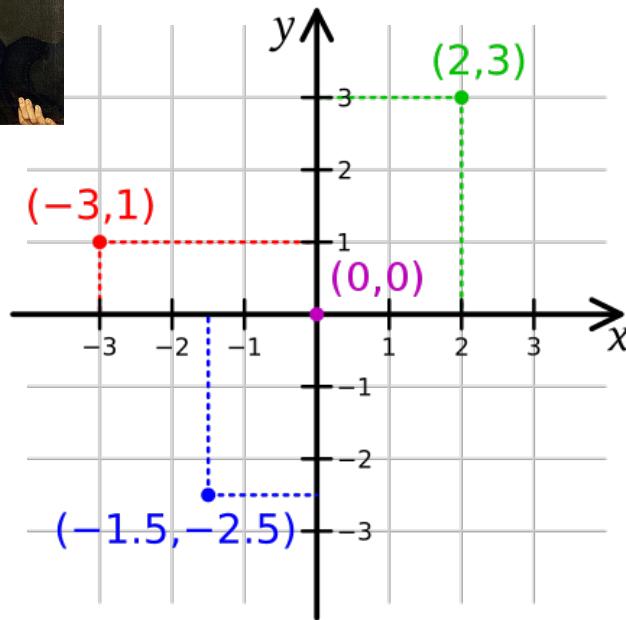
René Descartes  
1596-1650



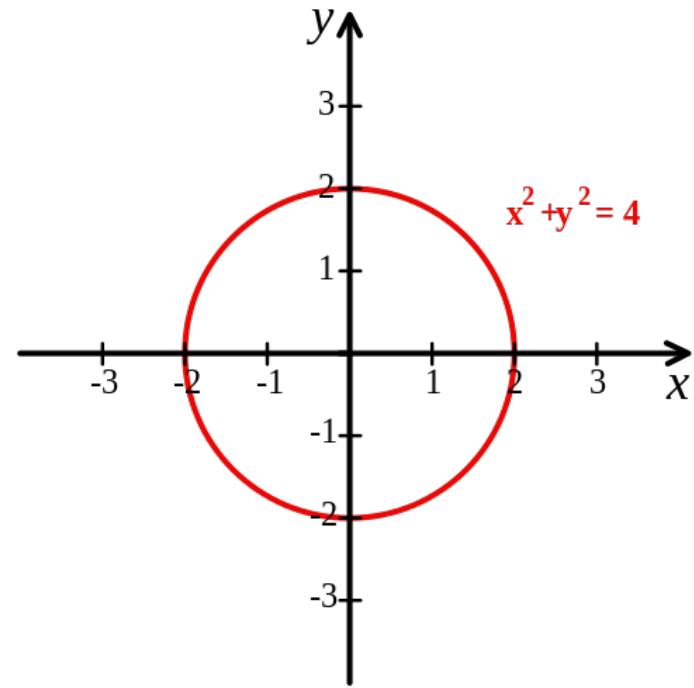
# Cartesian coordinate system



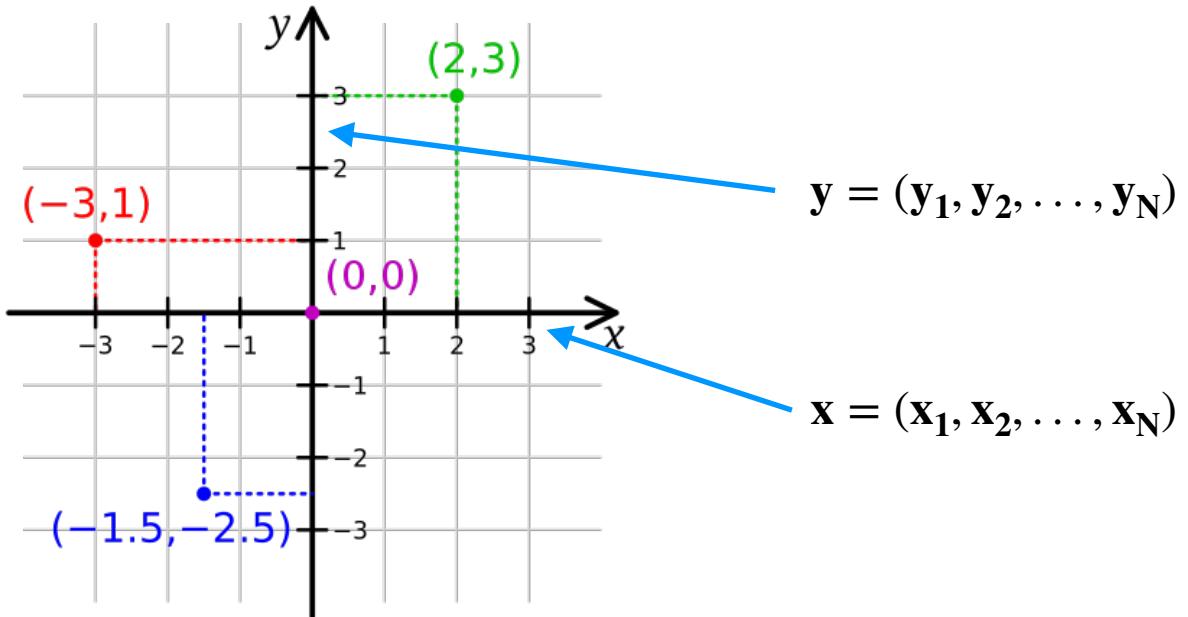
René Descartes  
1596-1650



Link between geometry and algebra!



# Cartesian coordinate system



# Data visualisation and the Grammar of Graphics



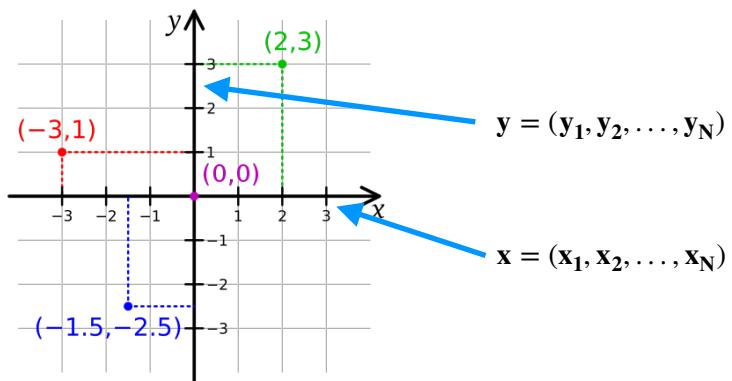
# Data visualisation and the Grammar of Graphics



- Visualising data is to convert values into visual elements that make up a graphic.

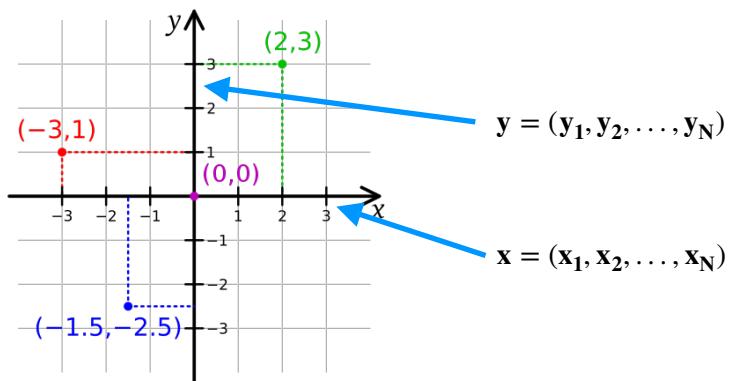
# Data visualisation and the Grammar of Graphics

- Visualising data is to convert values into visual elements that make up a graphic.
- We "map" data values onto quantifiable features of the resulting graphic - the aesthetics.

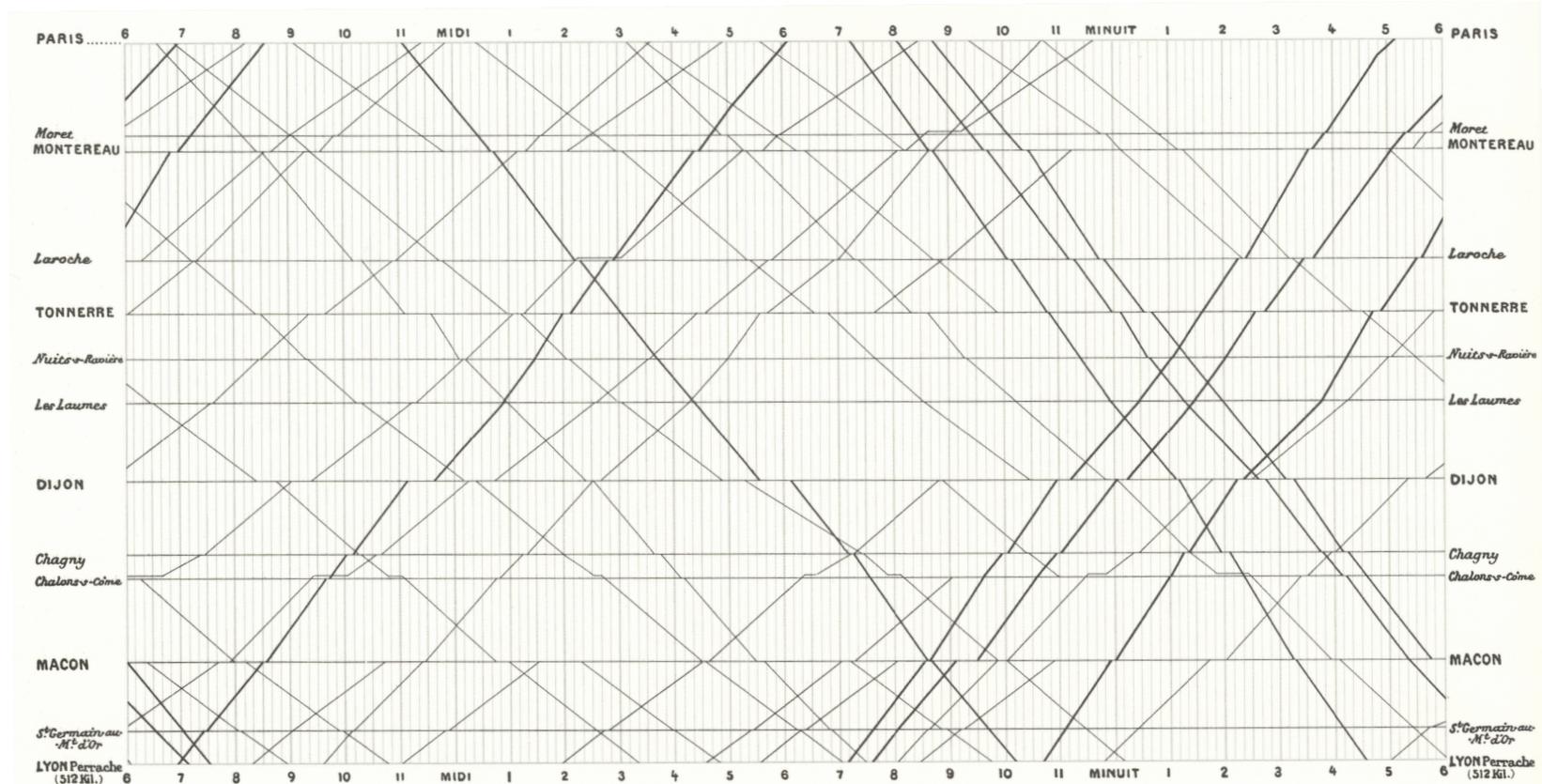


# Data visualisation and the Grammar of Graphics

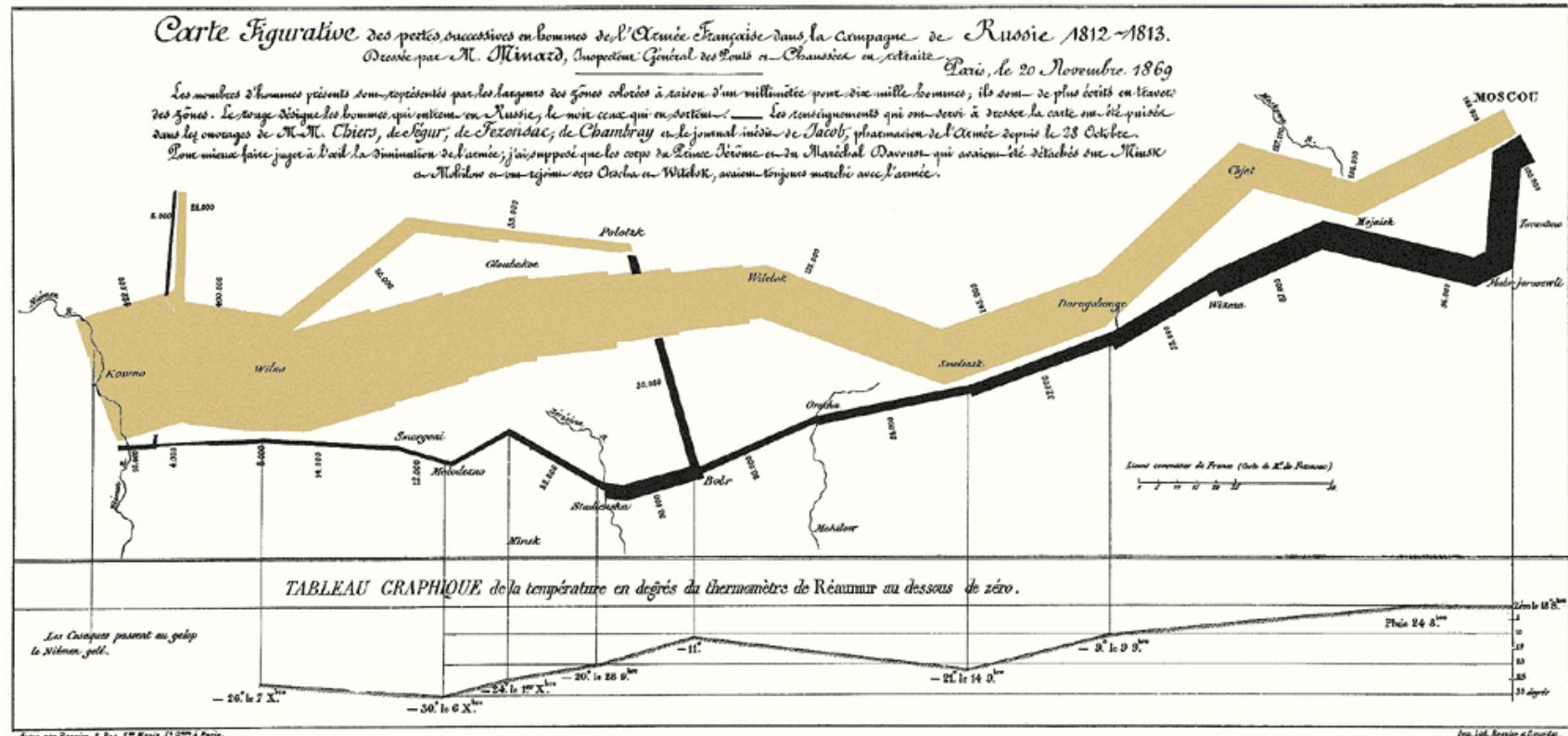
- Visualising data is to convert values into visual elements that make up a graphic.
- We "map" data values onto quantifiable features of the resulting graphic - the aesthetics.
- Variation along each "dimension" in the data is mapped onto one aesthetic.



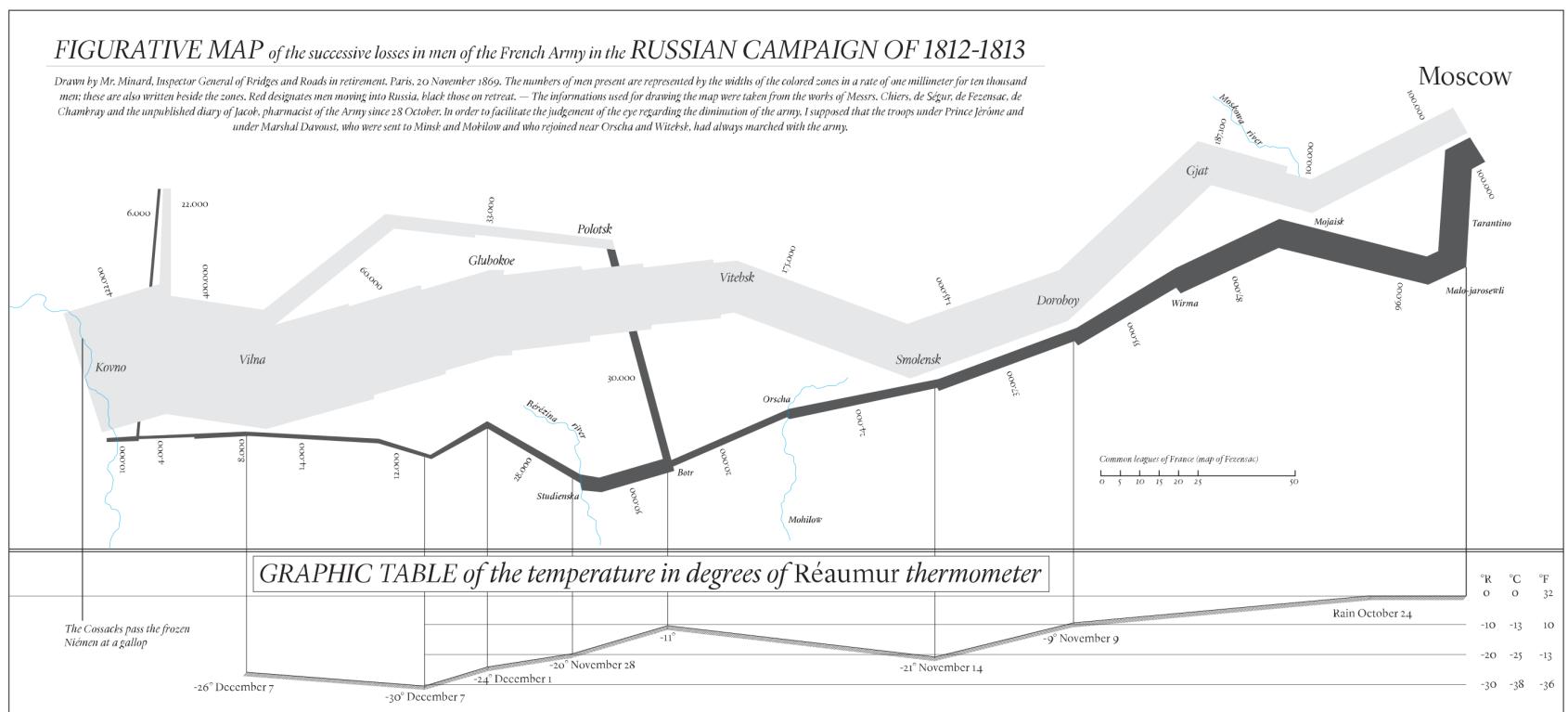
# Visualising time



# Quantitative information in geographical space

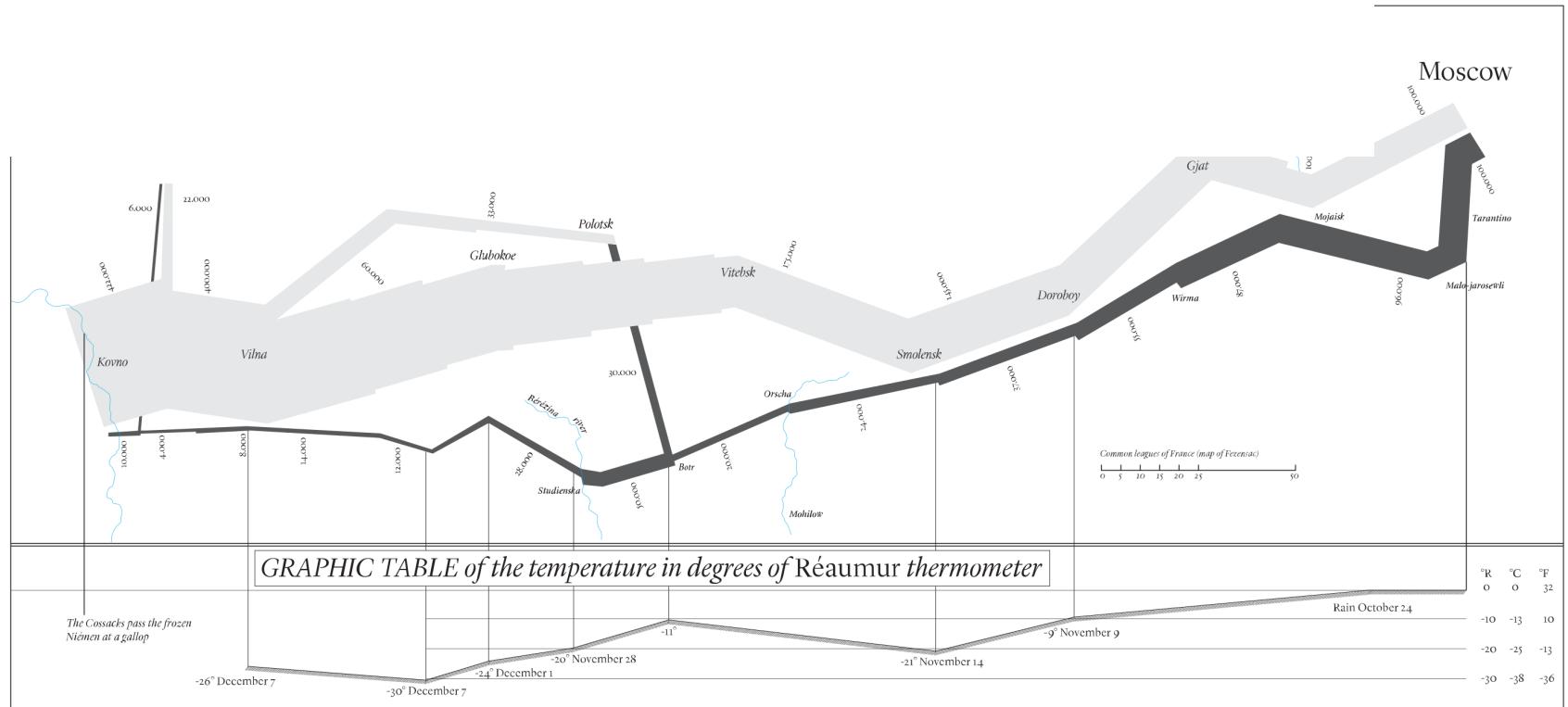


# Quantitative information in geographical space



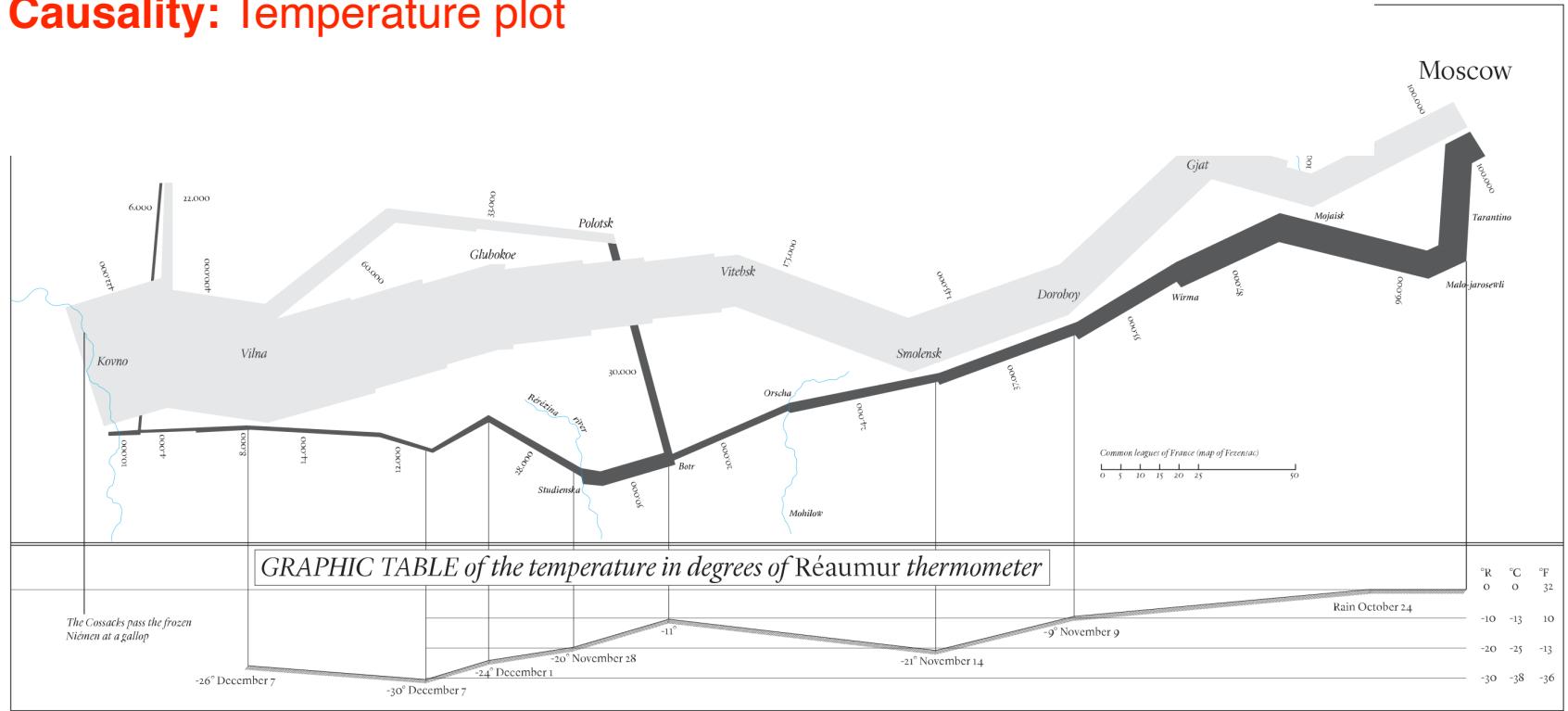
# Quantitative information in geographical space

- Contrast: Colors of advance and retreat. Width of lines back-to-back



# Quantitative information in geographical space

- **Contrast:** Colors of advance and retreat. Width of lines back-to-back
- **Causality:** Temperature plot

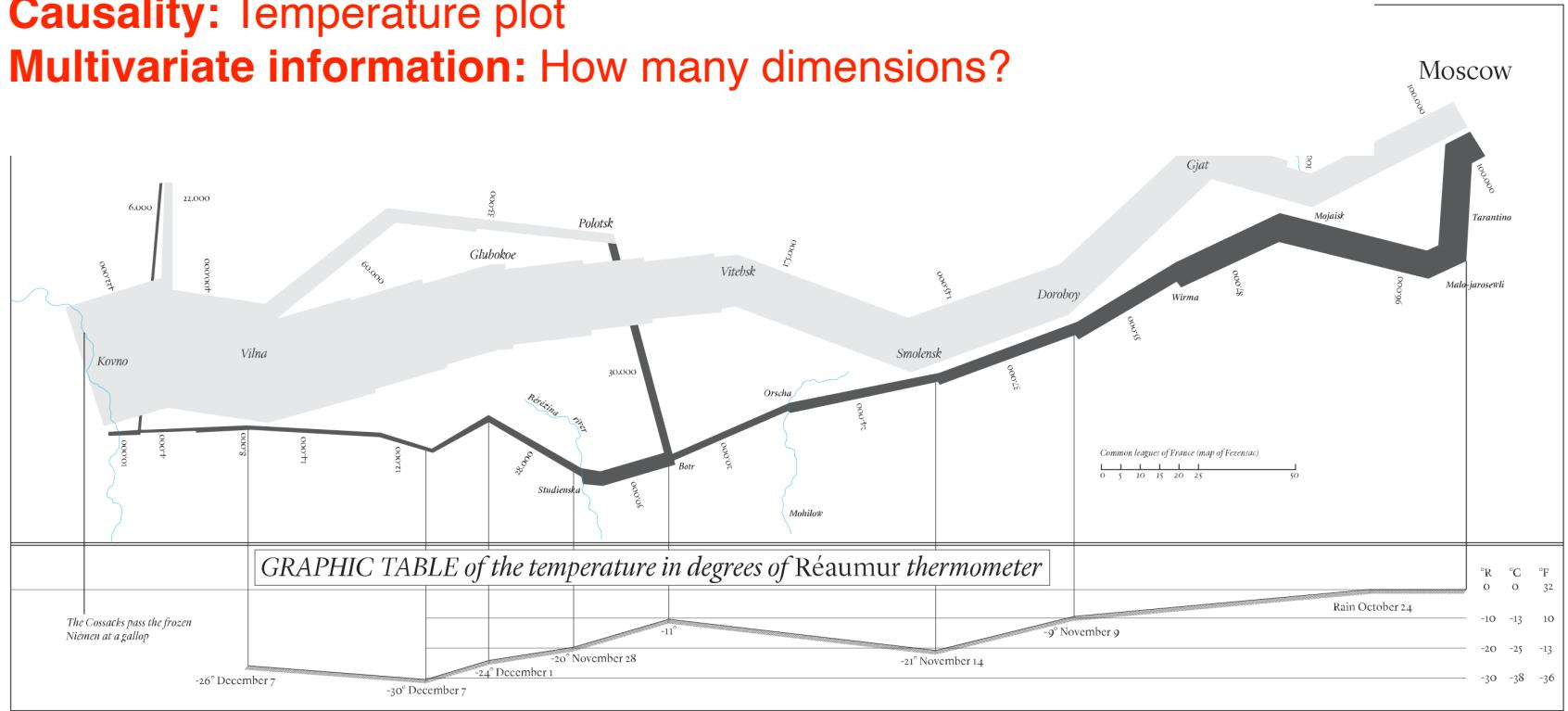


Charles-Joseph Minard, 1869

<https://badriadhikari.github.io/data-viz-workshop-2021/minards/>

# Quantitative information in geographical space

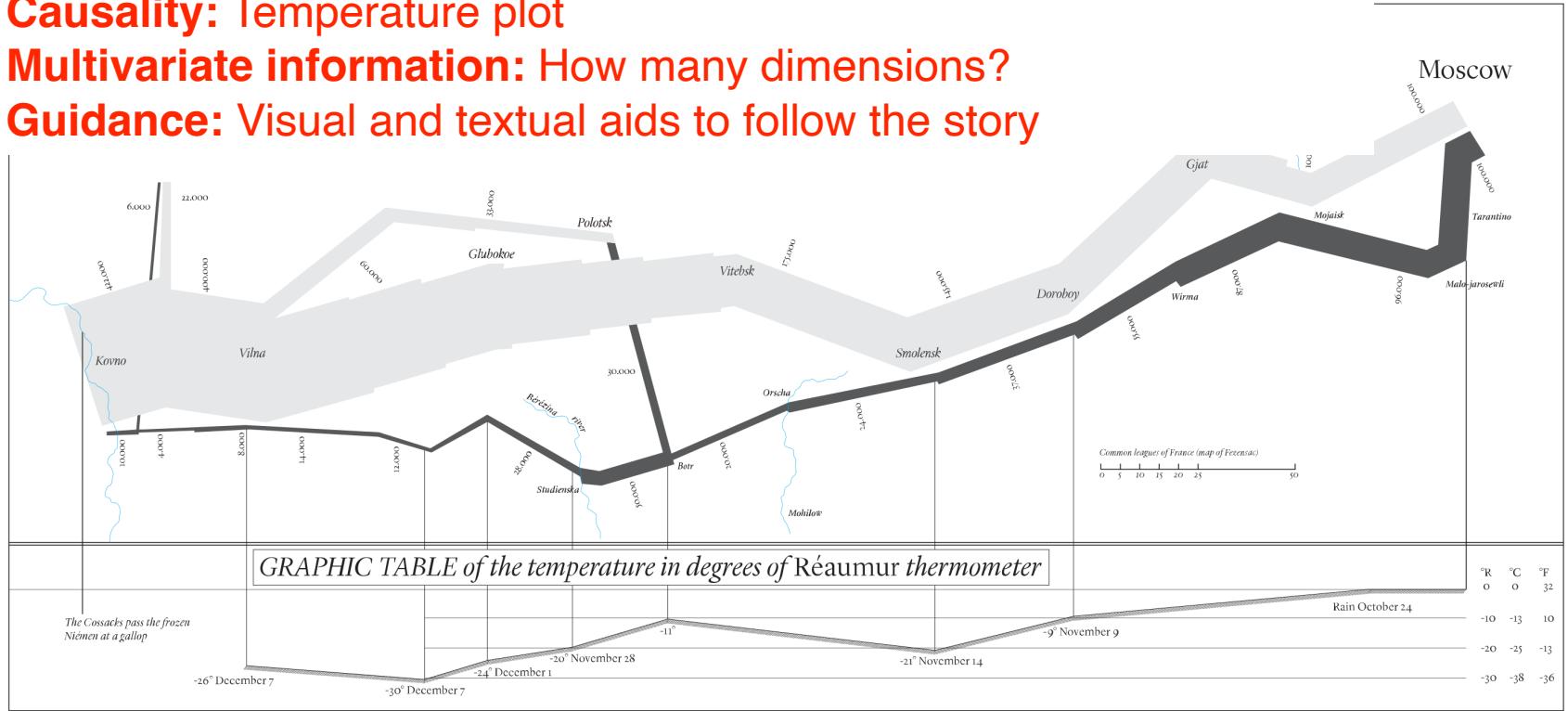
- **Contrast:** Colors of advance and retreat. Width of lines back-to-back
- **Causality:** Temperature plot
- **Multivariate information:** How many dimensions?



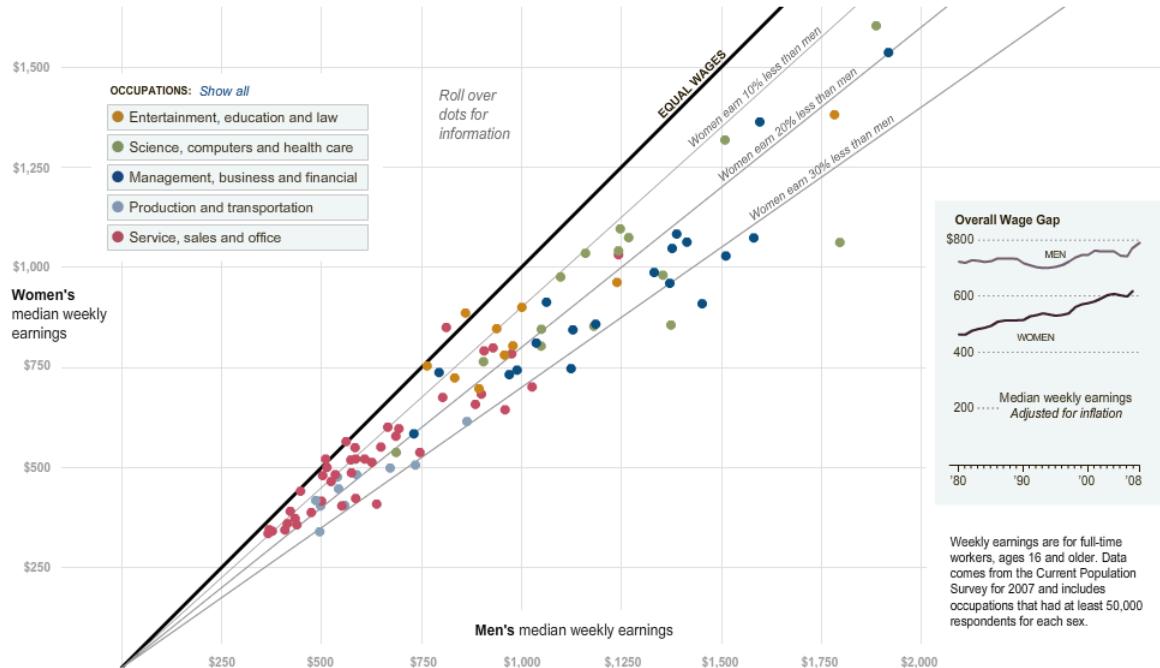
Charles-Joseph Minard, 1869  
<https://badriadhikari.github.io/data-viz-workshop-2021/minards/>

# Quantitative information in geographical space

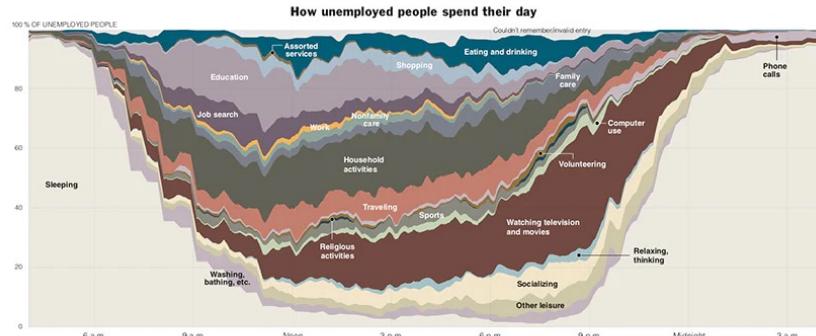
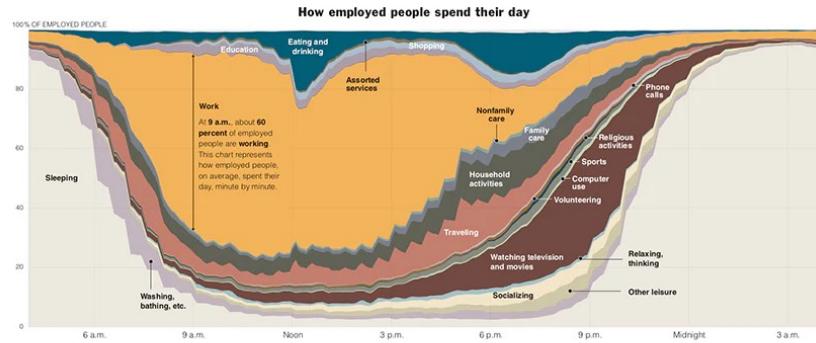
- **Contrast:** Colors of advance and retreat. Width of lines back-to-back
- **Causality:** Temperature plot
- **Multivariate information:** How many dimensions?
- **Guidance:** Visual and textual aids to follow the story



# Scatterplots



# Proportions over time



**Alarm clocks, or not**  
On weekdays at 6 a.m., more than 80 percent of unemployed people are sleeping, compared with nearly half of employed individuals. The unemployed sleep an hour more on weekdays than the employed.

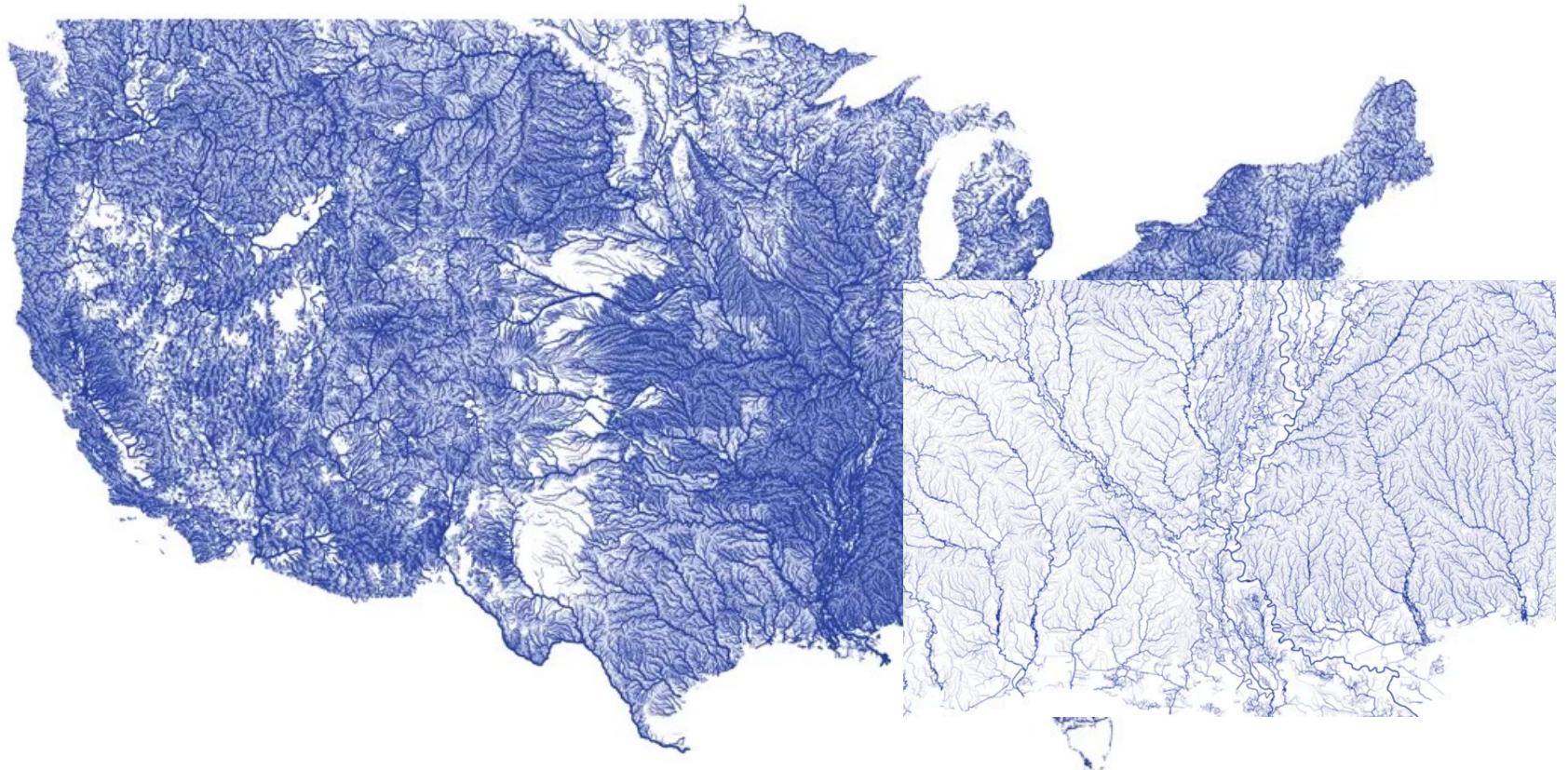
**Starting the daily grind**  
By 9 a.m., when more than half of the employed are working for pay, nearly half of the unemployed are still sleeping or looking for jobs. Only one in five are looking for work over the course of the day, but spend more than two hours doing so.

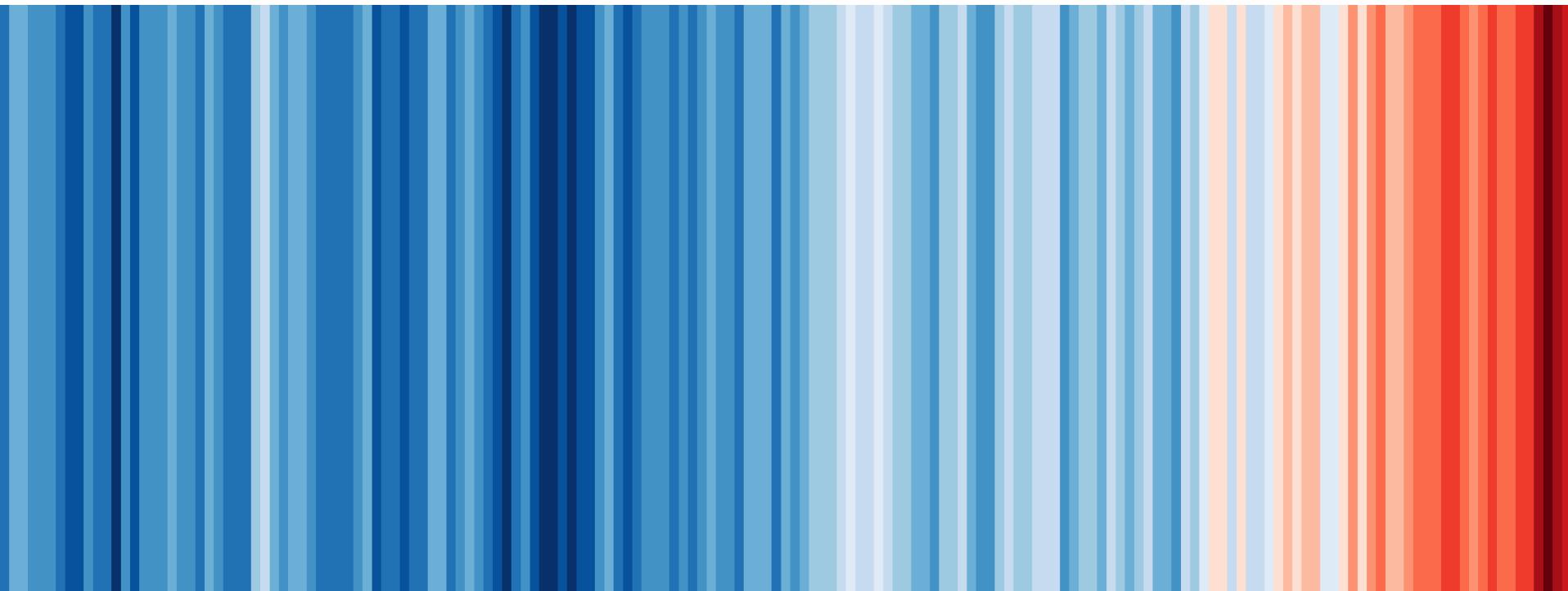
**The lunch break**  
The unemployed and the employed spend a comparable amount of time eating and drinking. But the meal schedule of the unemployed is less sharply defined. Workers are twice as likely to dine from noon to 1 p.m.

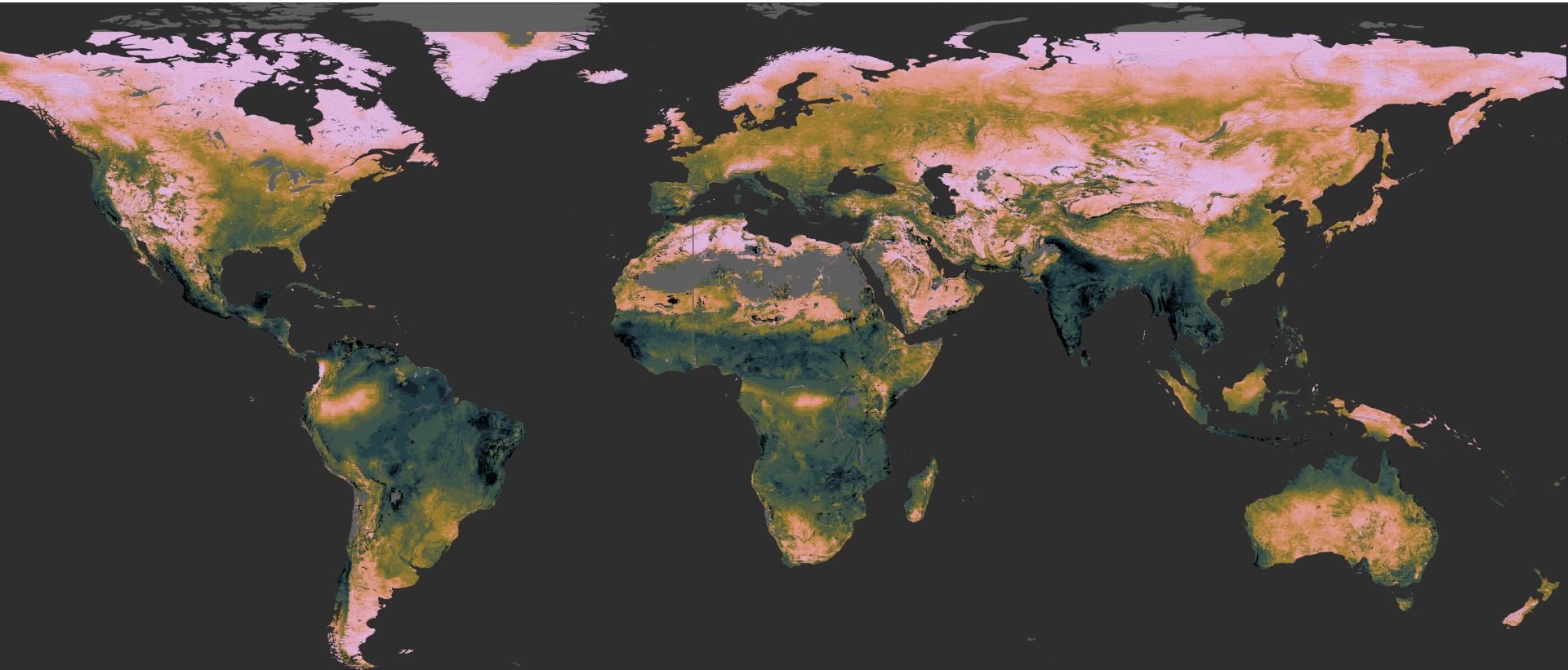
**On the road**  
The unemployed are just as mobile as the employed, as both spend over an hour traveling on average each day. But the peak for among the employed peaks at 5 p.m., while the unemployed are more likely to be traveling at times throughout the afternoon.

**TV time**  
At 9 p.m., about a third of all people surveyed are in front of the television. But at almost any hour of the day, a larger share of unemployed people are watching television or movies.

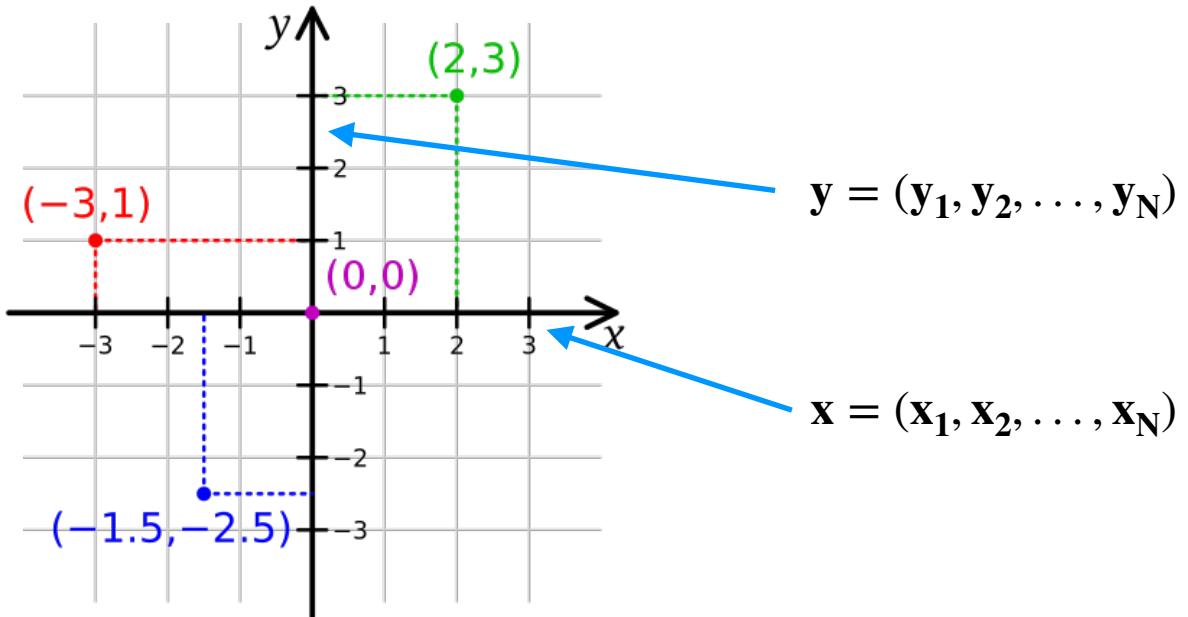
**Socializing before sleep**  
Socializing ramps up for both the employed and unemployed into the early evening, but the unemployed are more likely to continue these activities into the night. And they spend nearly three hours — twice as much as the employed — socializing and talking on the phone throughout the day.





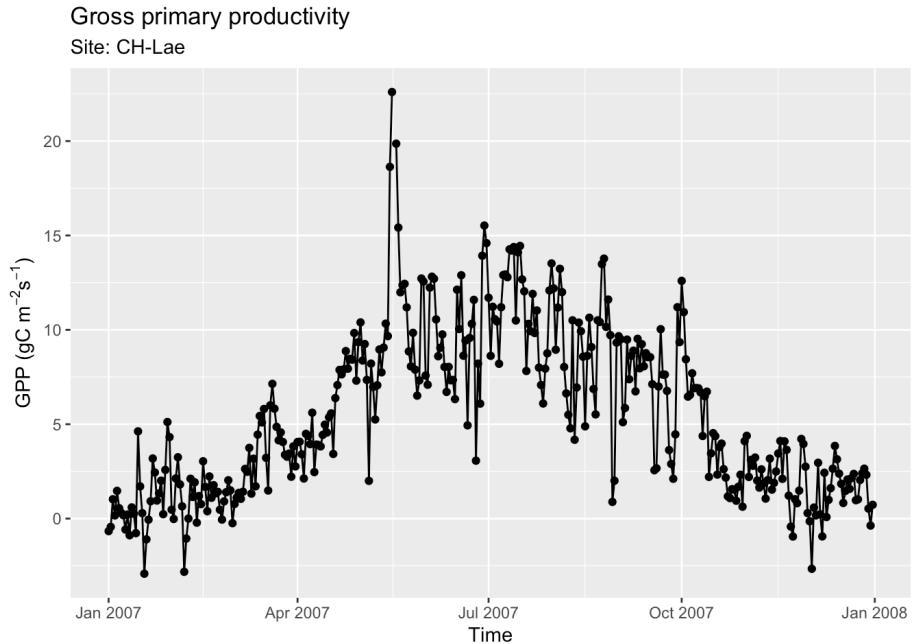


# Cartesian coordinate system



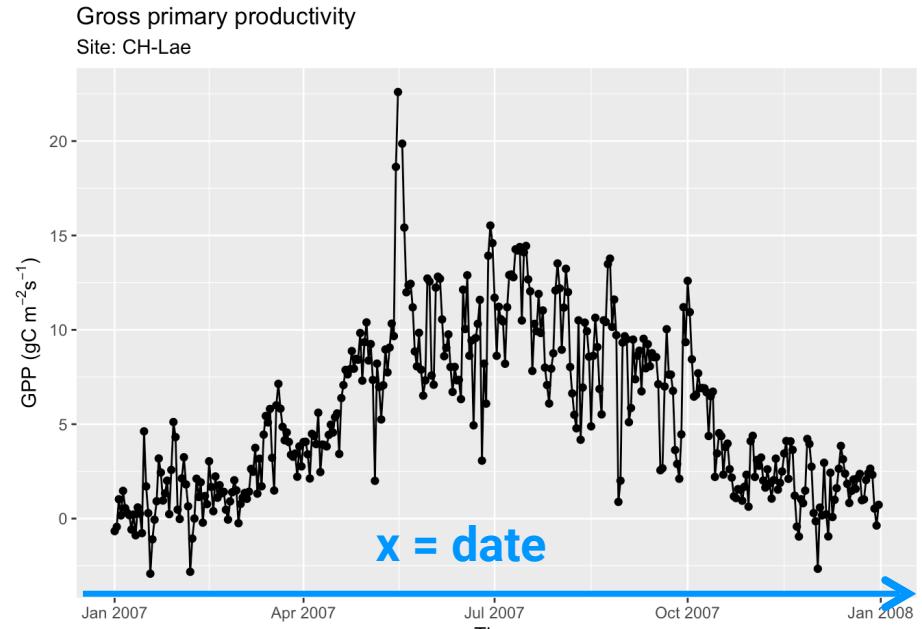
# Mapping to two aesthetics

```
> df
# A tibble: 3,912 x 3
  date      GPP_NT_VUT_REF f_measured
  <date>        <dbl>       <dbl>
1 2004-03-31     1.62     0.333
2 2004-04-01     5.03     0.396
3 2004-04-02     3.73     0.312
4 2004-04-03     3.87     0.5
5 2004-04-04     6.06     0.333
6 2004-04-05     4.46     0.396
7 2004-04-06    10.5     0.0417
8 2004-04-07     7.09     0.312
9 2004-04-08     6.90     0.208
10 2004-04-09    4.66     0.667
# ... with 3,902 more rows
```



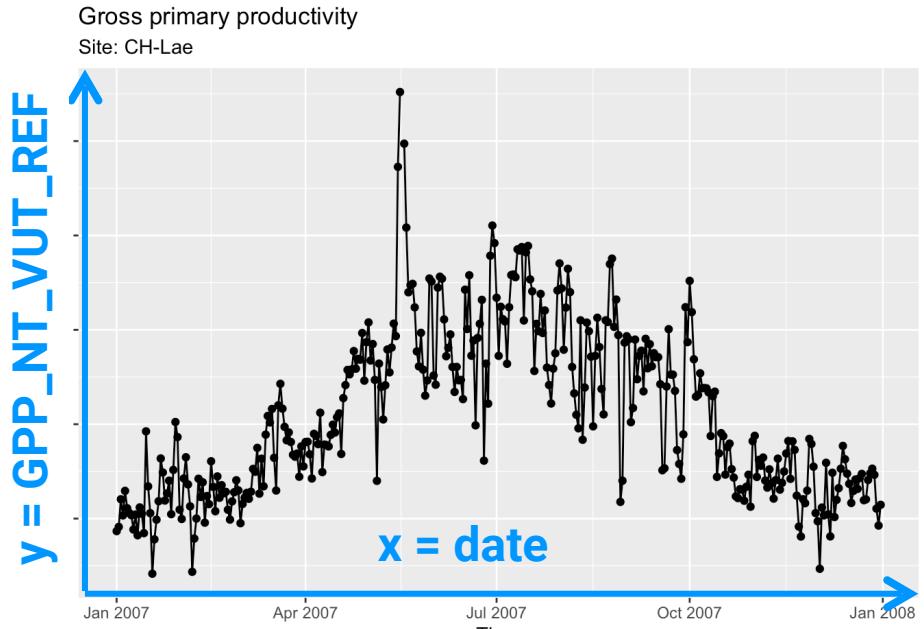
# Mapping to two aesthetics

```
> df
# A tibble: 3,912 x 3
  date      GPP_NT_VUT_REF f_measured
  <date>        <dbl>      <dbl>
1 2004-03-31     1.62     0.333
2 2004-04-01     5.03     0.396
3 2004-04-02     3.73     0.312
4 2004-04-03     3.87     0.5
5 2004-04-04     6.06     0.333
6 2004-04-05     4.46     0.396
7 2004-04-06    10.5     0.0412
8 2004-04-07     7.09     0.312
9 2004-04-08     6.90     0.208
10 2004-04-09    4.66     0.667
# ... with 3,902 more rows
```



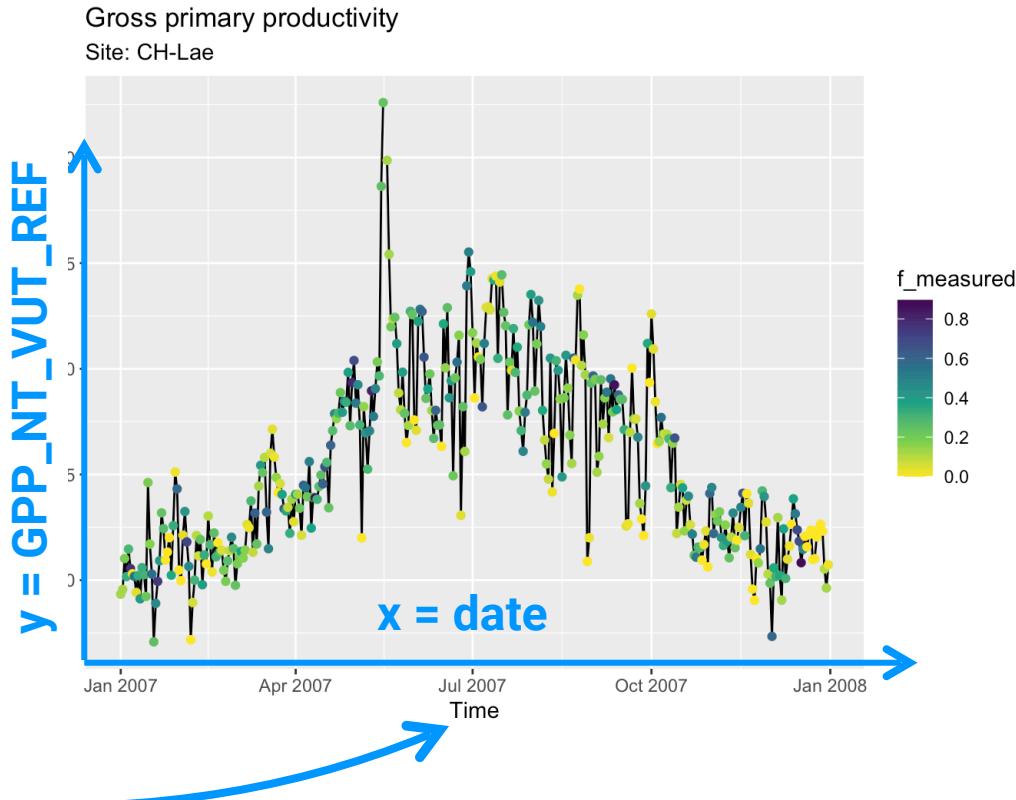
# Mapping to two aesthetics

```
> df
# A tibble: 3,912 x 3
  date      GPP_NT_VUT_REF f_measured
  <date>        <dbl>       <dbl>
1 2004-03-31     1.62     0.333
2 2004-04-01     5.03     0.396
3 2004-04-02     3.73     0.312
4 2004-04-03     3.87     0.5
5 2004-04-04     6.06     0.333
6 2004-04-05     4.46     0.396
7 2004-04-06    10.5     0.0412
8 2004-04-07     7.09     0.312
9 2004-04-08     6.90     0.208
10 2004-04-09    4.66     0.667
# ... with 3,902 more rows
```



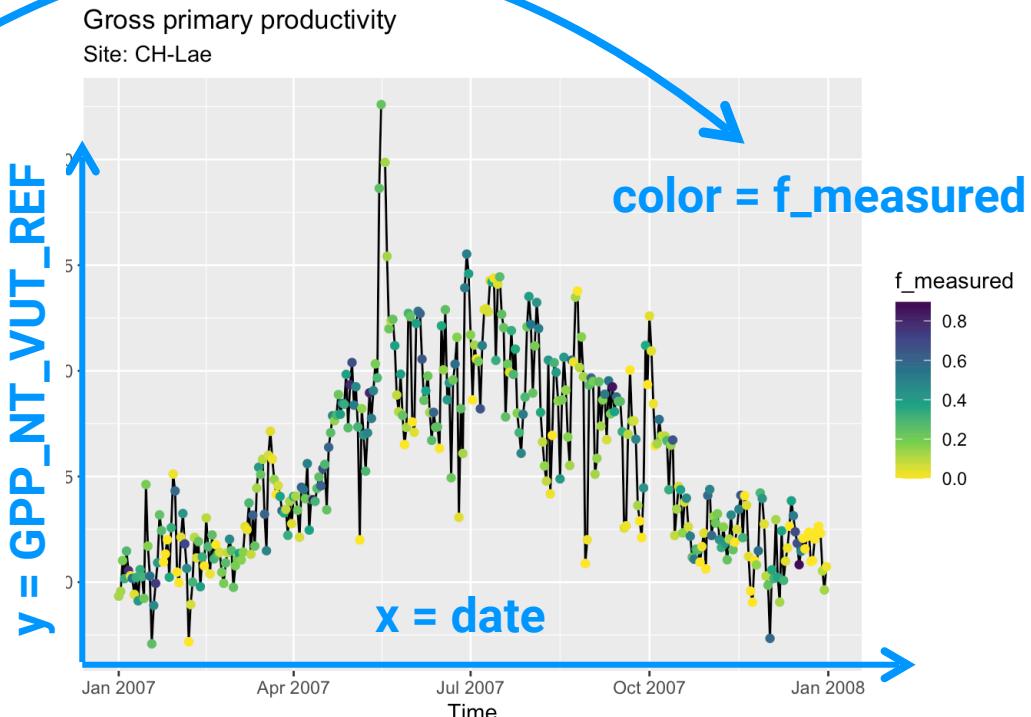
# Mapping to three aesthetics

```
> df
# A tibble: 3,912 x 3
  date      GPP_NT_VUT_REF f_measured
  <date>        <dbl>       <dbl>
1 2004-03-31     1.62     0.333
2 2004-04-01     5.03     0.396
3 2004-04-02     3.73     0.312
4 2004-04-03     3.87     0.5
5 2004-04-04     6.06     0.333
6 2004-04-05     4.46     0.396
7 2004-04-06    10.5     0.0412
8 2004-04-07     7.09     0.312
9 2004-04-08     6.90     0.208
10 2004-04-09    4.66     0.667
# ... with 3,902 more rows
```



# Mapping to three aesthetics

```
> df
# A tibble: 3,912 x 3
  date      GPP_NT_VUT_REF f_measured
  <date>        <dbl>       <dbl>
1 2004-03-31     1.62     0.333
2 2004-04-01     5.03     0.396
3 2004-04-02     3.73     0.312
4 2004-04-03     3.87     0.5
5 2004-04-04     6.06     0.333
6 2004-04-05     4.46     0.396
7 2004-04-06    10.5     0.0412
8 2004-04-07     7.09     0.312
9 2004-04-08     6.90     0.208
10 2004-04-09    4.66     0.667
# ... with 3,902 more rows
```



# ggplot2

```
ddf %>%
  ggplot(aes(x = date, y = GPP_NT_VUT_REF)) +
  geom_line() +
  geom_point()
```

# ggplot2

A data frame

```
ddf %>%
  ggplot(aes(x = date, y = GPP_NT_VUT_REF)) +
  geom_line() +
  geom_point()
```



# ggplot2

A data frame

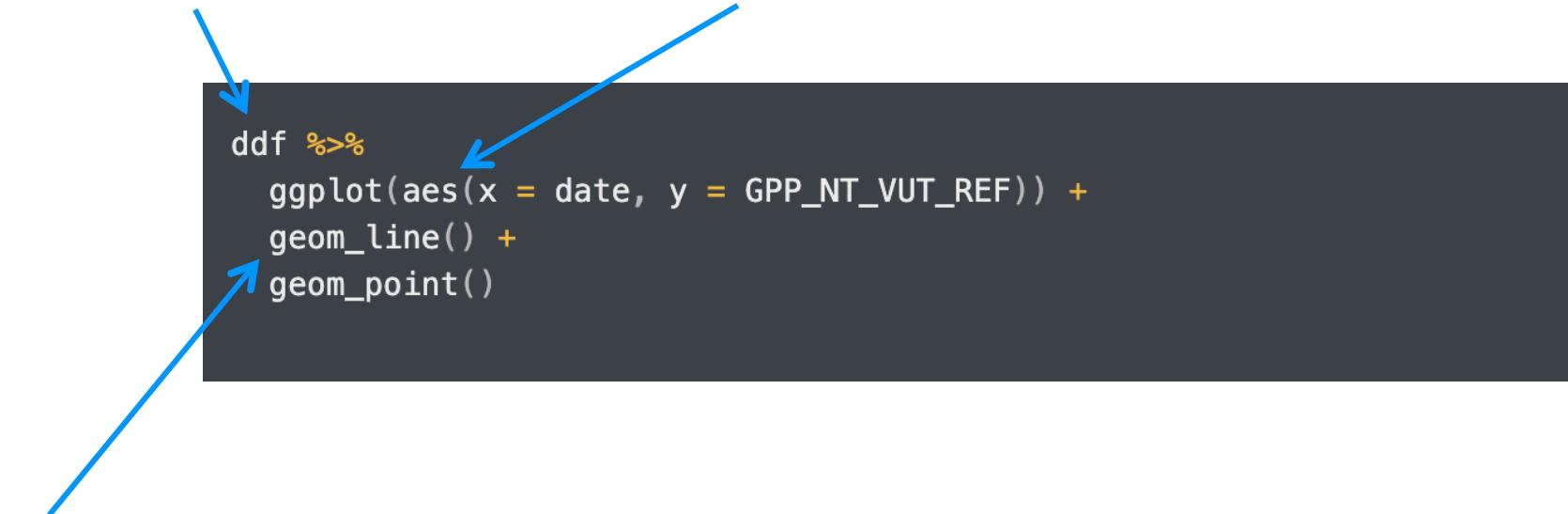
Aesthetics mapping

```
ddf %>%
  ggplot(aes(x = date, y = GPP_NT_VUT_REF)) +
  geom_line() +
  geom_point()
```

# ggplot2

A data frame

Aesthetics mapping



Visualisation element

# ggplot2

A data frame

Aesthetics mapping

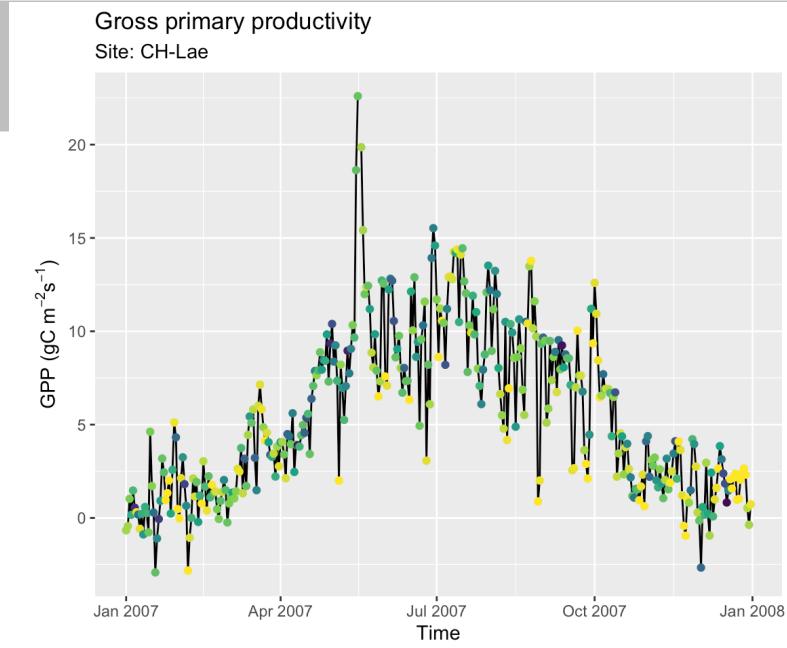
```
ddf %>%
  ggplot(aes(x = date, y = GPP_NT_VUT_REF)) +
  geom_line() +
  geom_point()
```

(A second visualisation element)

Visualisation element

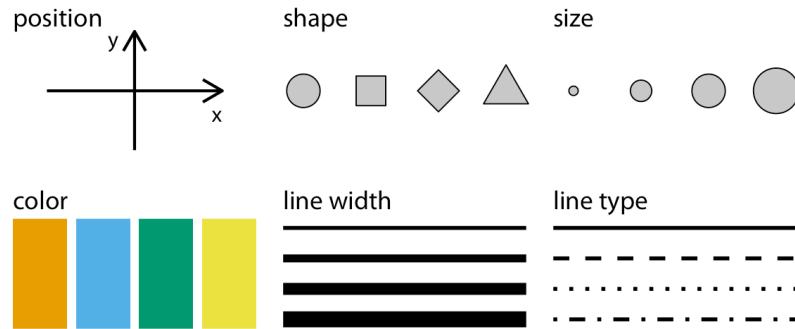
# ggplot2

```
> df
# A tibble: 3,912 x 3
  date      GPP_NT_VUT_REF f_measured
  <date>            <dbl>      <dbl>
1 2004-03-31        1.62     0.333
2 2004-04-01        5.03     0.396
3 2004-04-02        3.73     0.312
4 2004-04-03        3.87     0.5
5 2004-04-04        6.06     0.333
6 2004-04-05        4.46     0.396
7 2004-04-06       10.5     0.0412
8 2004-04-07        7.09     0.312
9 2004-04-08        6.90     0.208
10 2004-04-09       4.66     0.667
# ... with 3,902 more rows
```



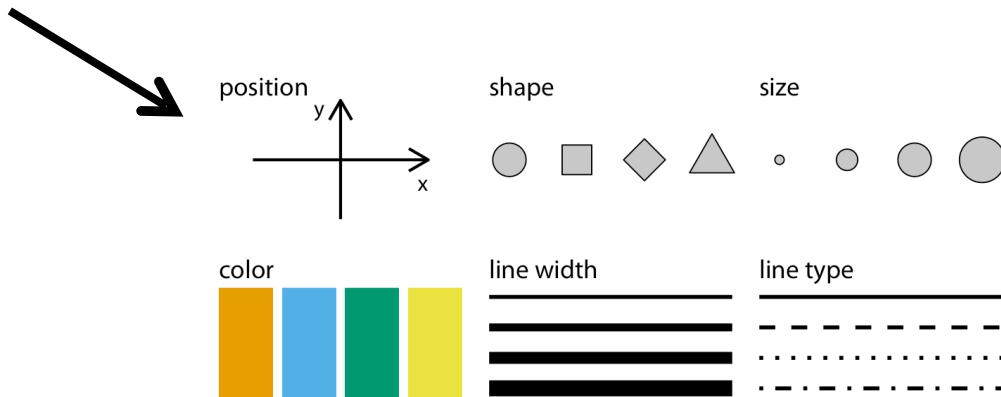
```
ddf %>%
  ggplot(aes(x = date, y = GPP_NT_VUT_REF, color = f_measured)) +
  geom_line() +
  geom_point() +
  scale_color_viridis_c()
```

# Aesthetics for continuous and categorical data



# Aesthetics for continuous and categorical data

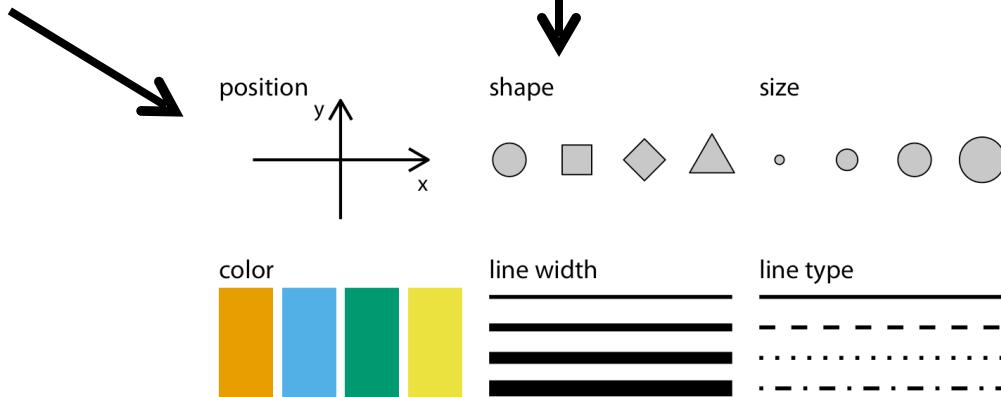
Continuous  
but...



# Aesthetics for continuous and categorical data

Continuous  
but...

Categorical

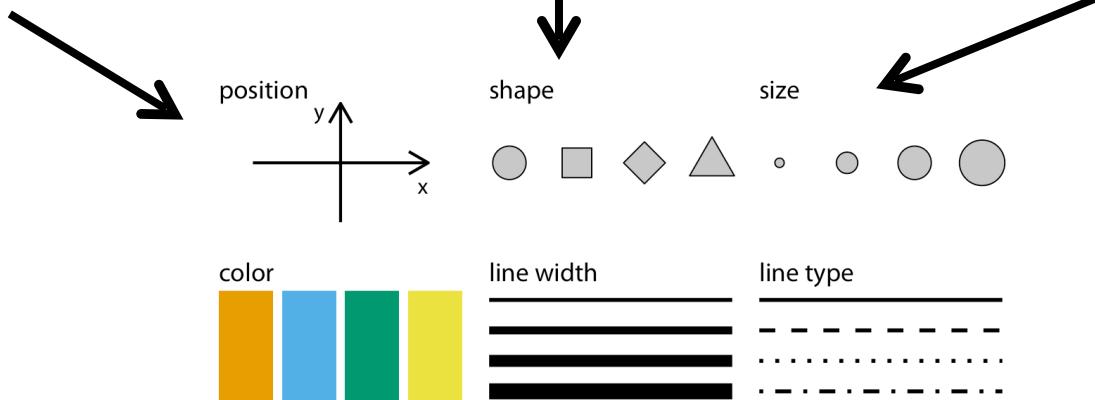


# Aesthetics for continuous and categorical data

Continuous  
but...

Categorical

Continuous

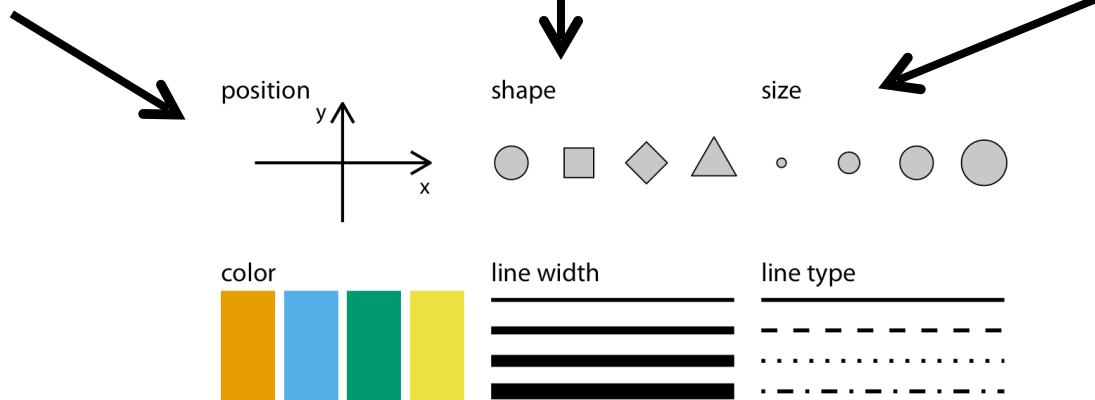


# Aesthetics for continuous and categorical data

Continuous  
but...

Categorical

Continuous

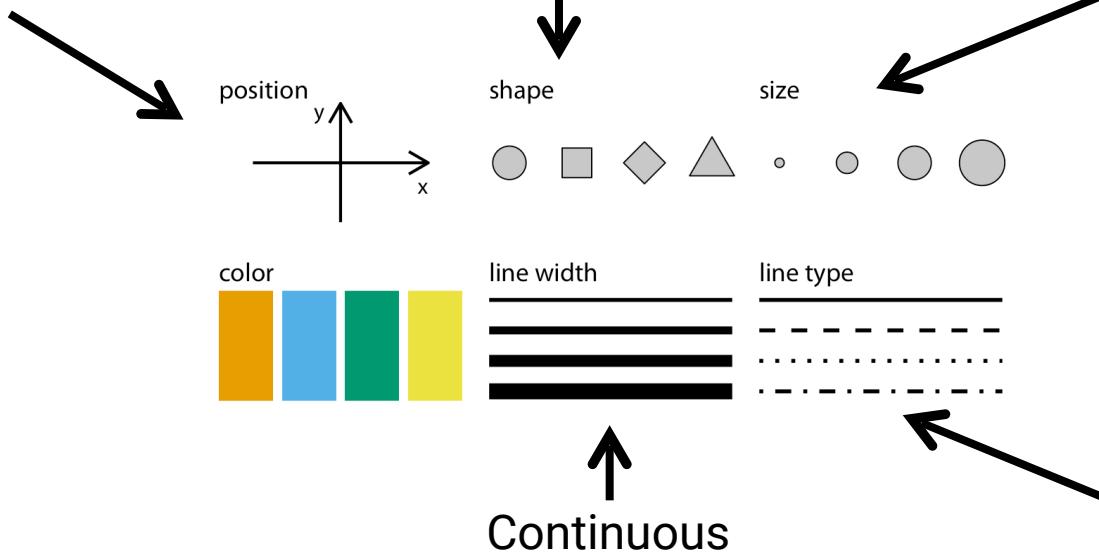


# Aesthetics for continuous and categorical data

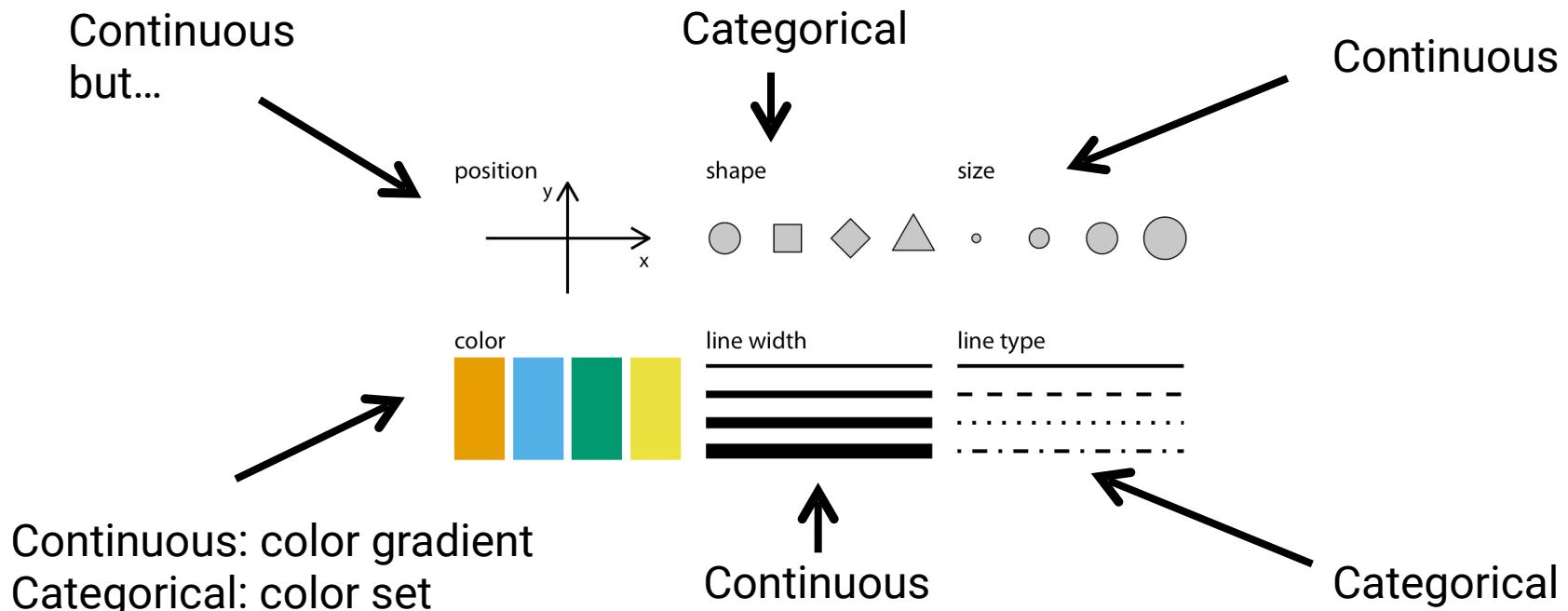
Continuous  
but...

Categorical

Continuous



# Aesthetics for continuous and categorical data



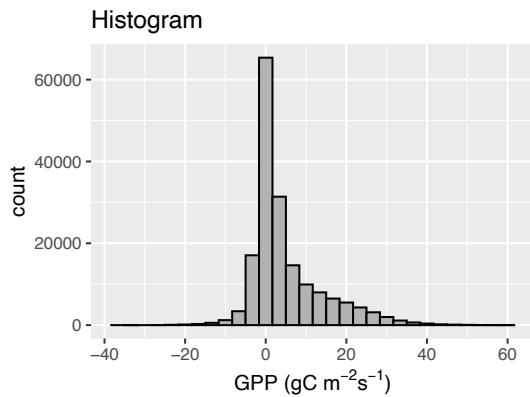
# Different “geoms” for different aspects of the data



## Distributions

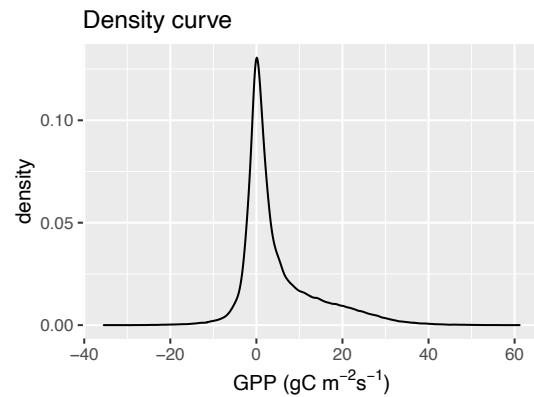
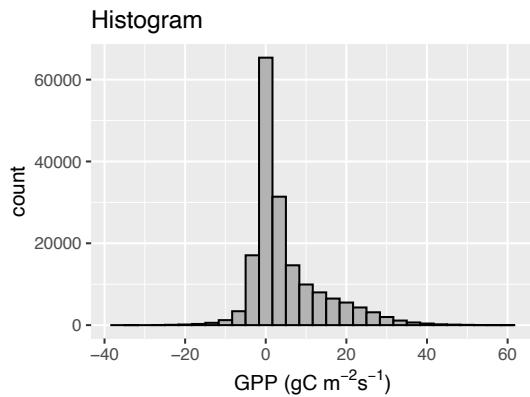
# Different “geoms” for different aspects of the data

## Distributions



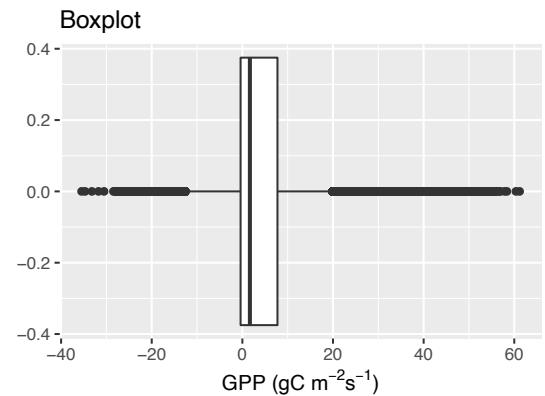
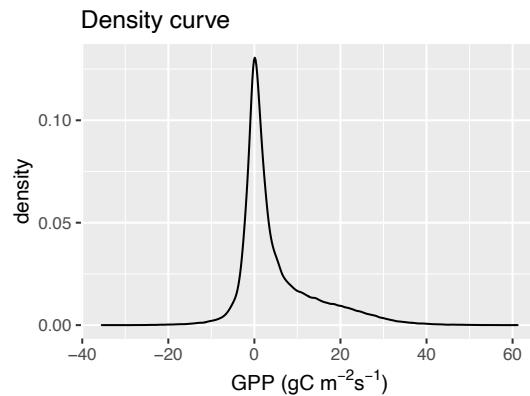
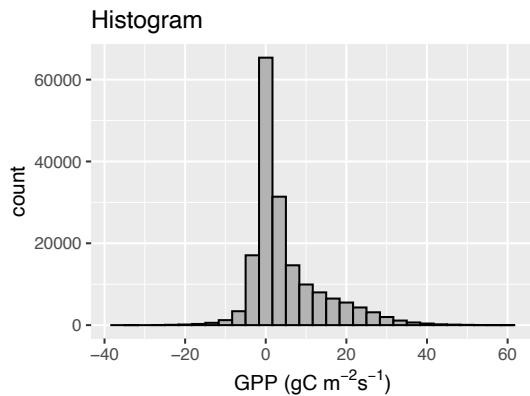
# Different “geoms” for different aspects of the data

## Distributions

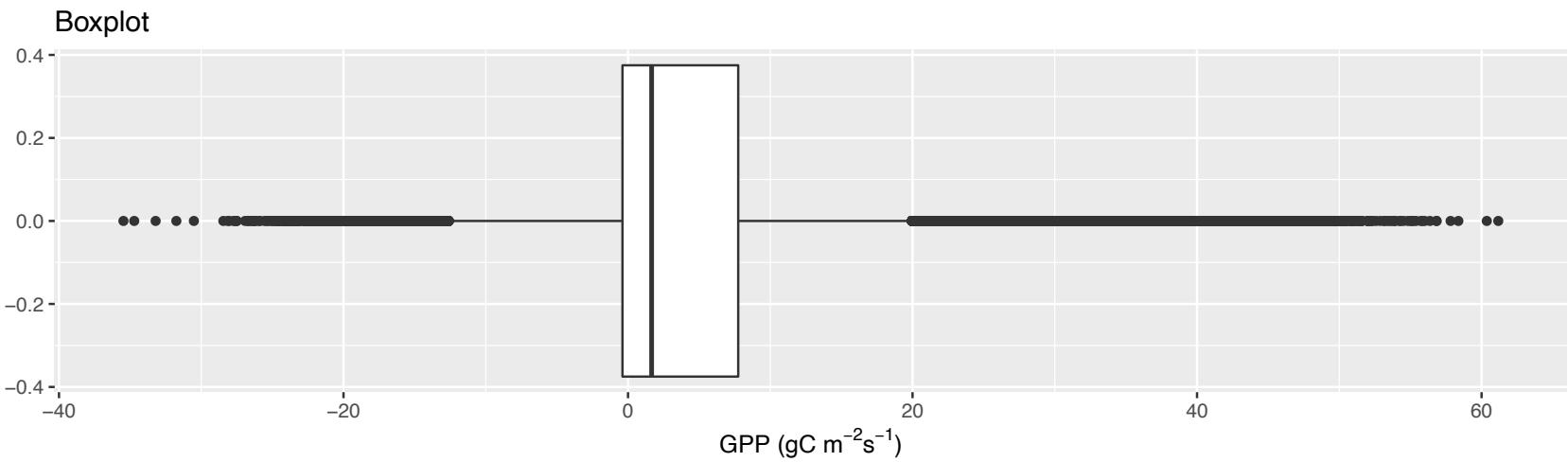


# Different “geoms” for different aspects of the data

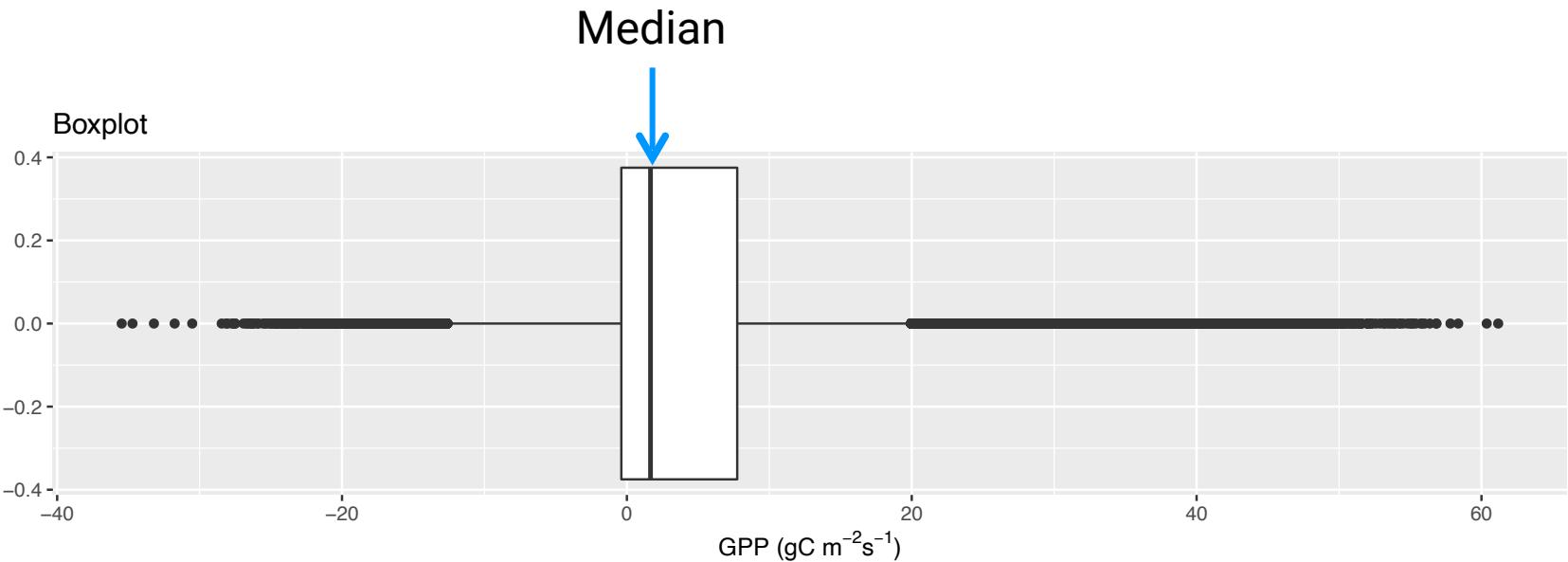
## Distributions



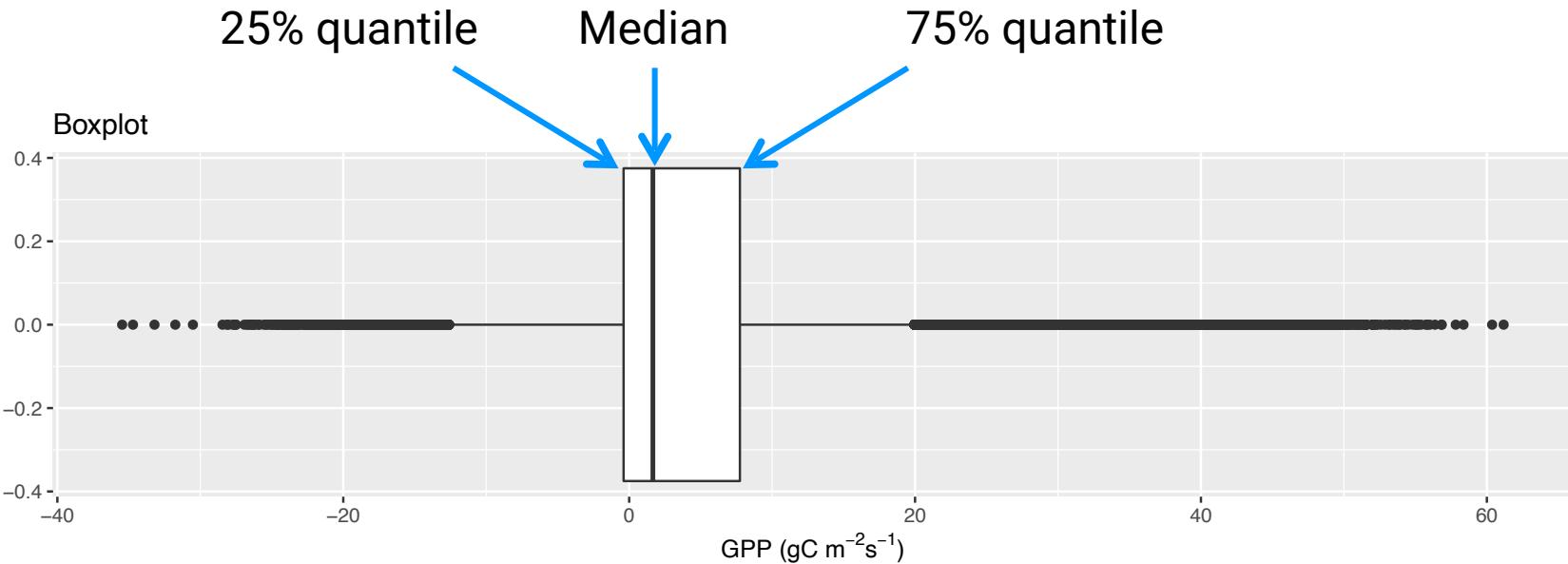
# Boxplot



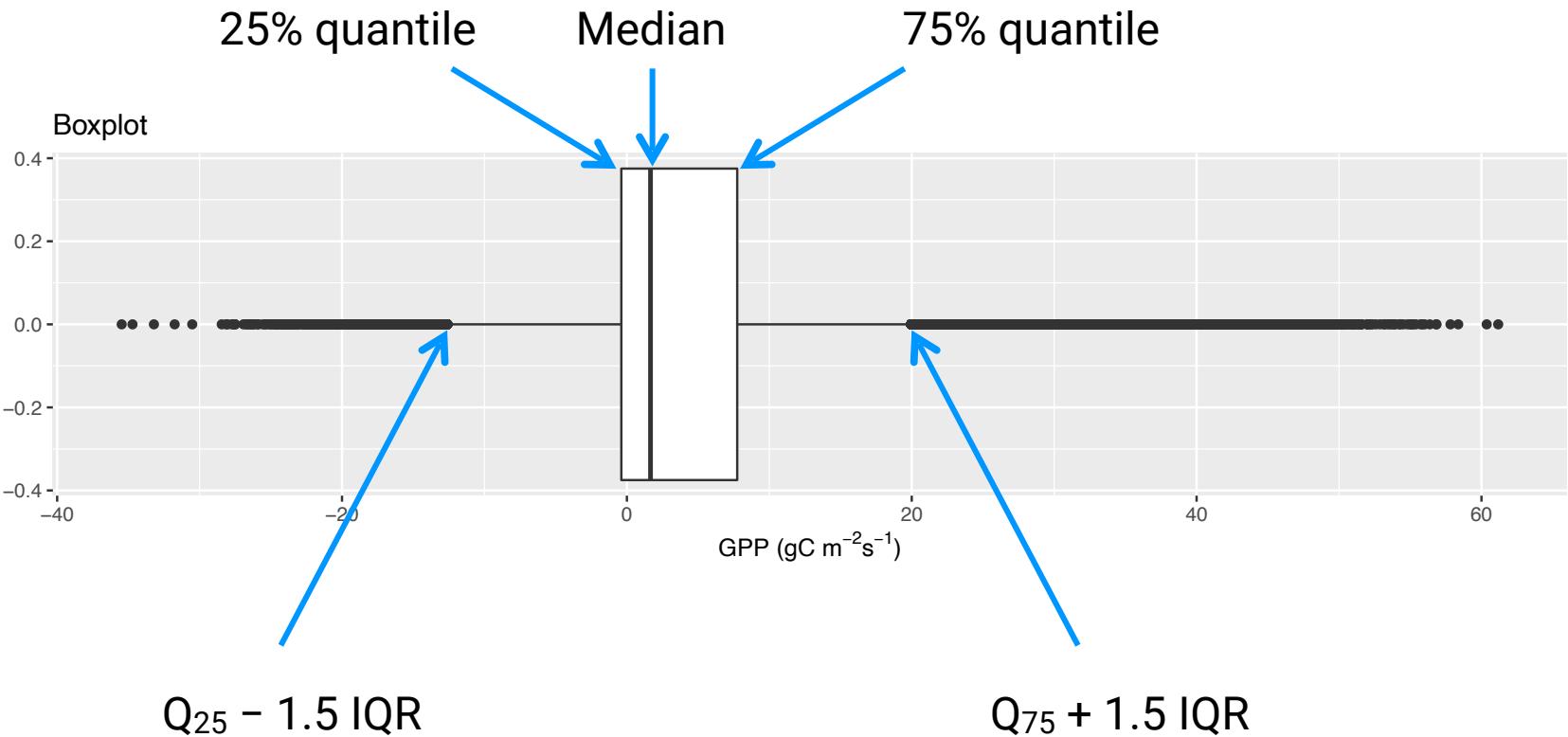
# Boxplot



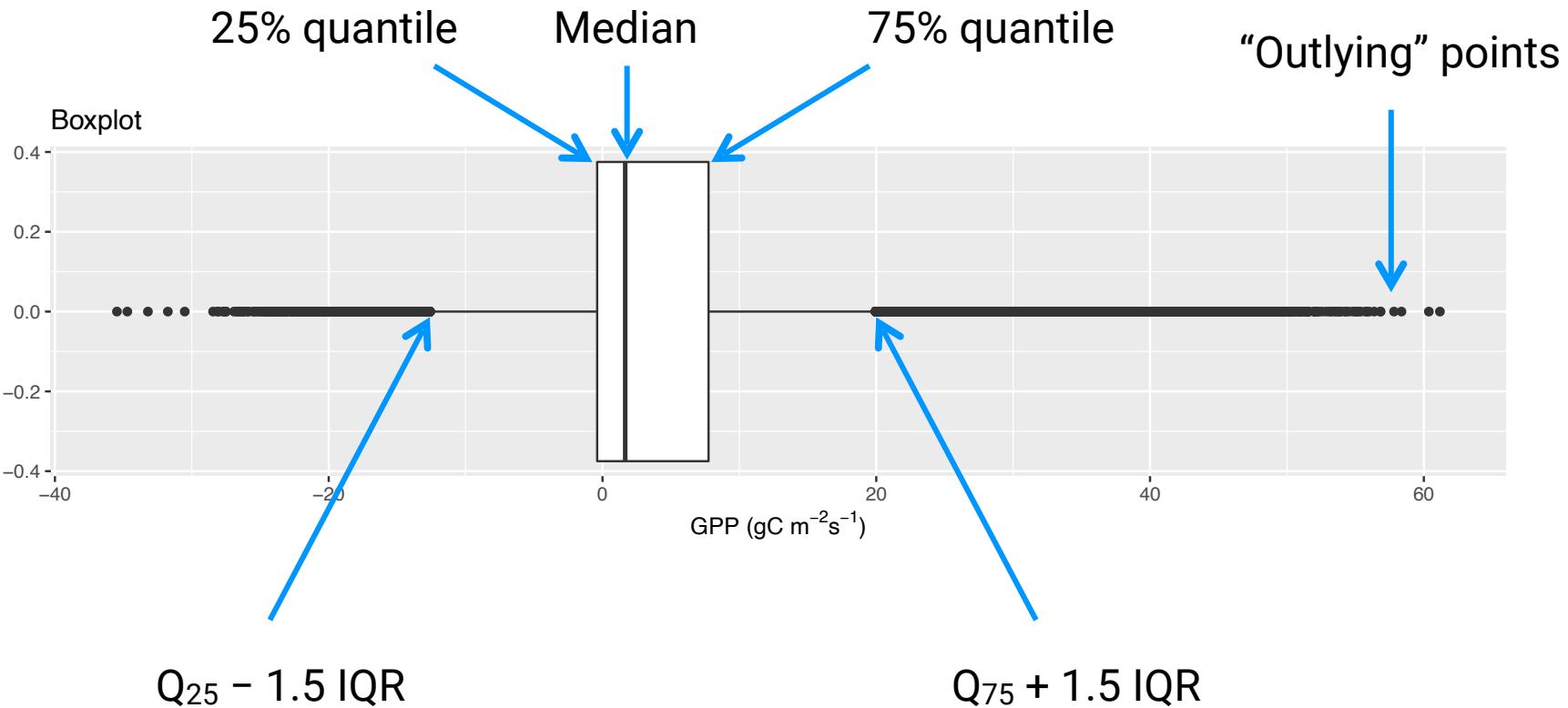
# Boxplot



# Boxplot



# Boxplot



# Different “geoms” for different aspects of the data

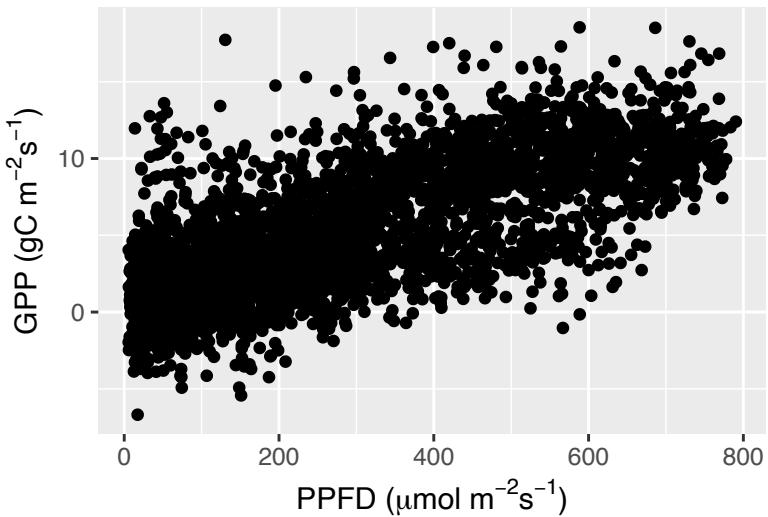


## Relationships

# Different “geoms” for different aspects of the data

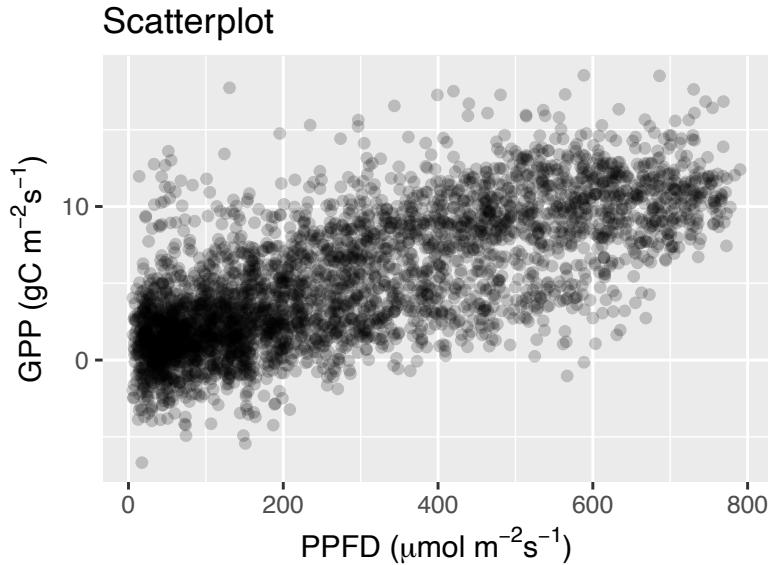
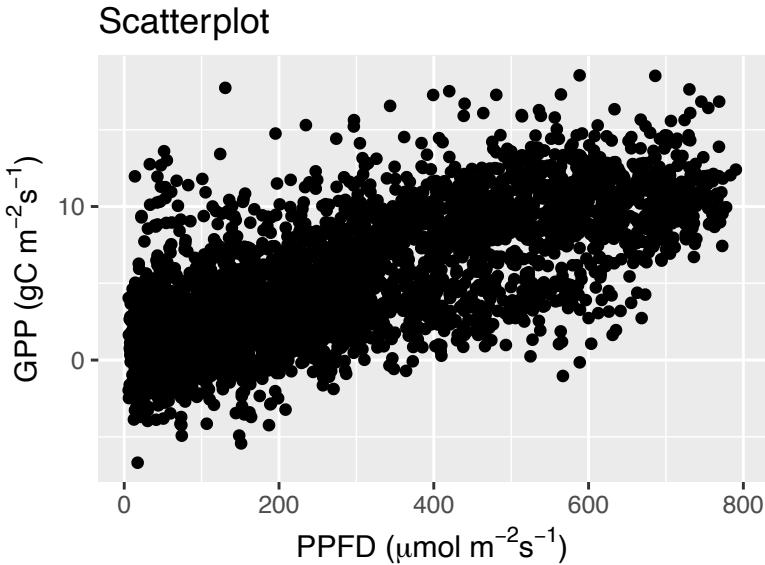
## Relationships

Scatterplot



# Different “geoms” for different aspects of the data

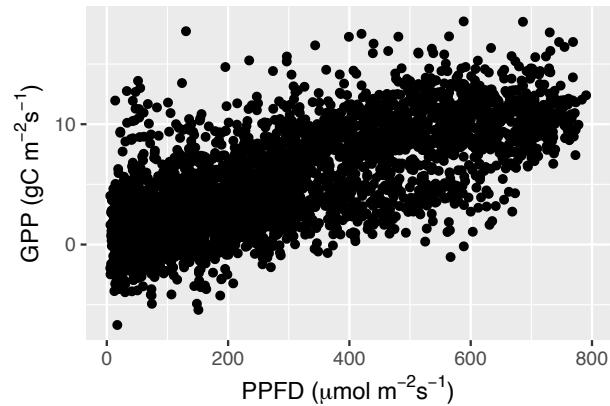
## Relationships



# Different “geoms” for different aspects of the data

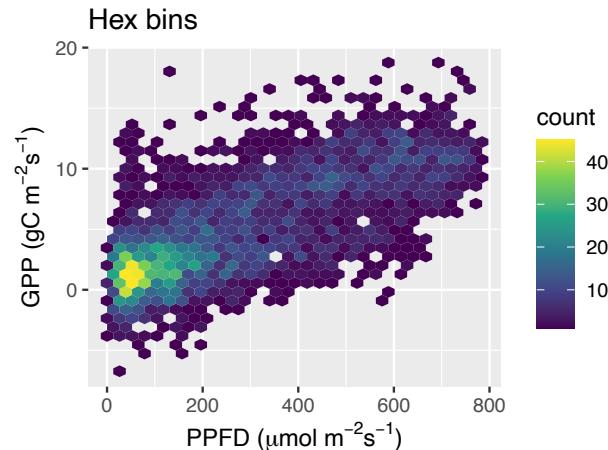
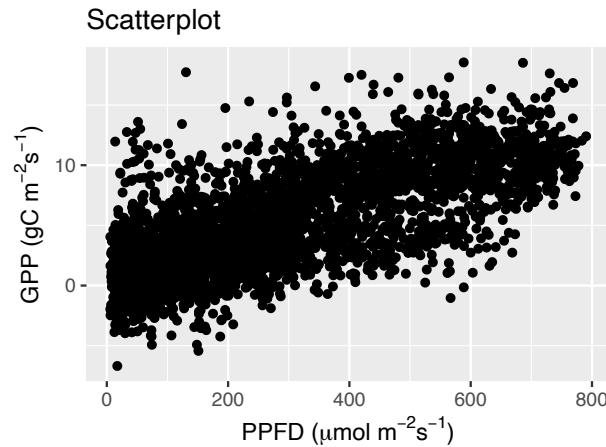
## Relationships

Scatterplot



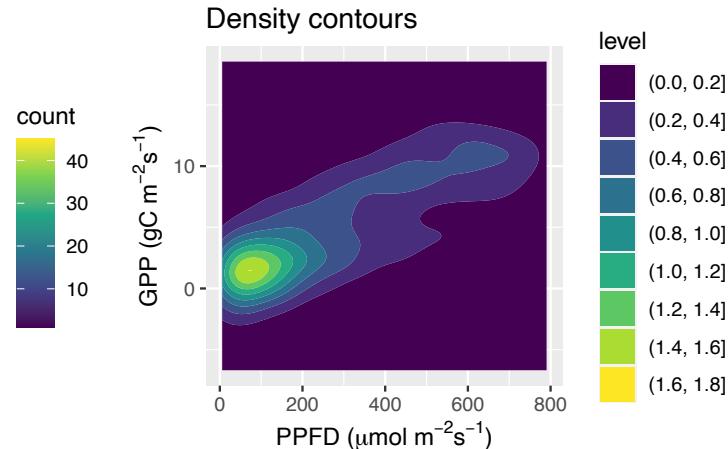
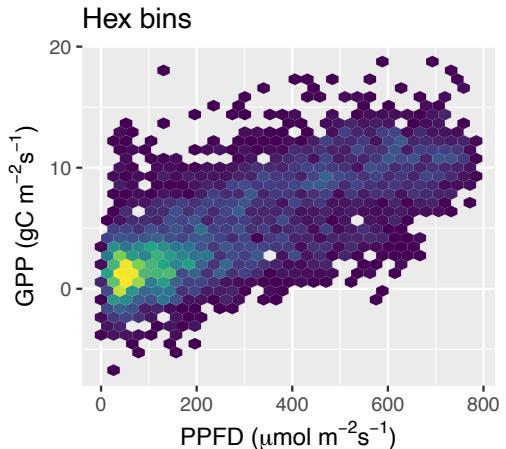
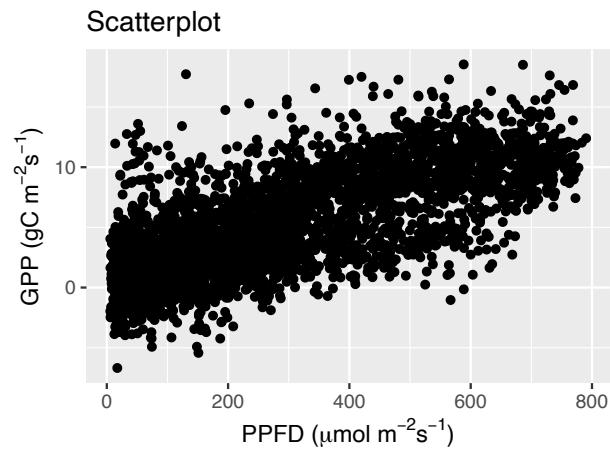
# Different “geoms” for different aspects of the data

## Relationships



# Different “geoms” for different aspects of the data

## Relationships



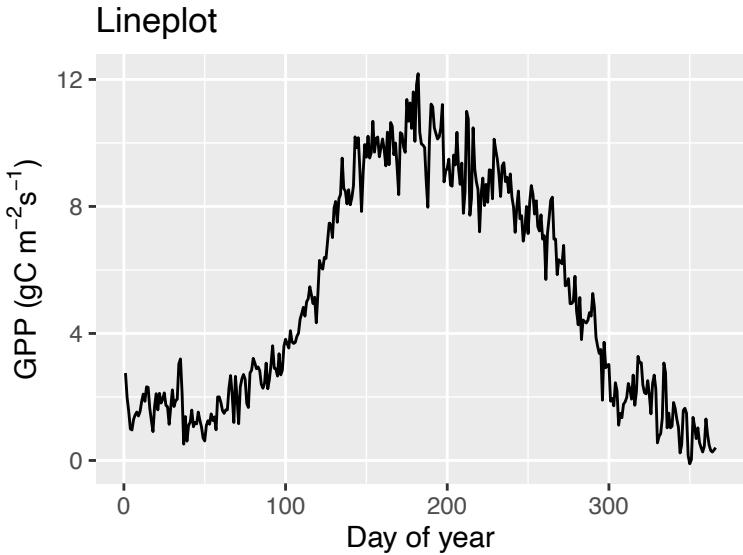
# Different “geoms” for different aspects of the data



## Time series

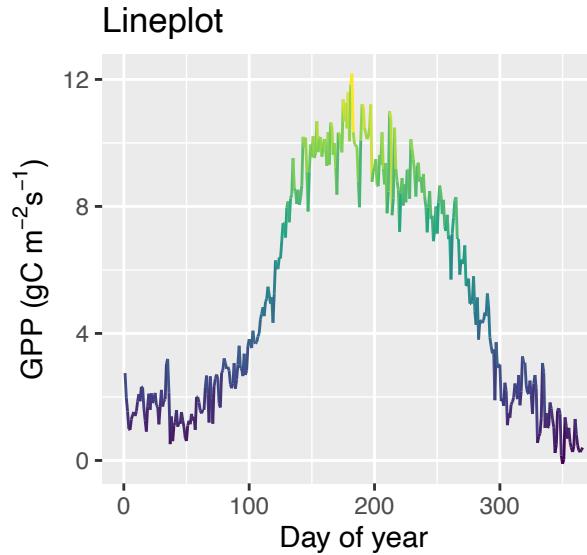
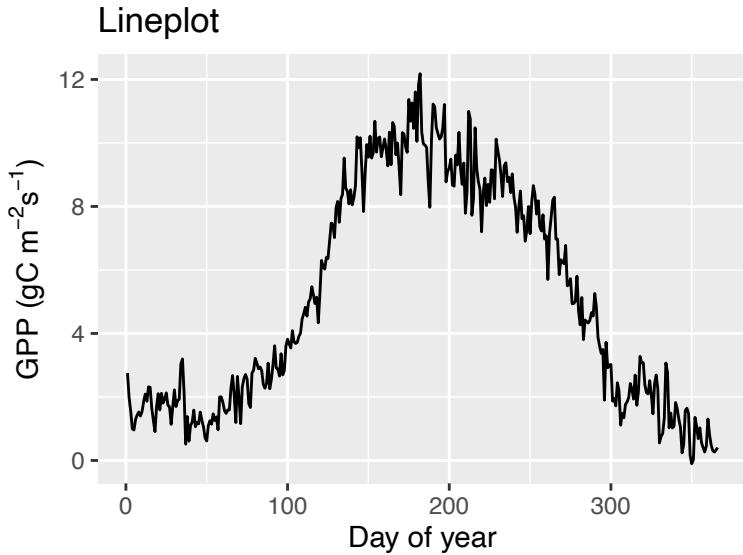
# Different “geoms” for different aspects of the data

## Time series



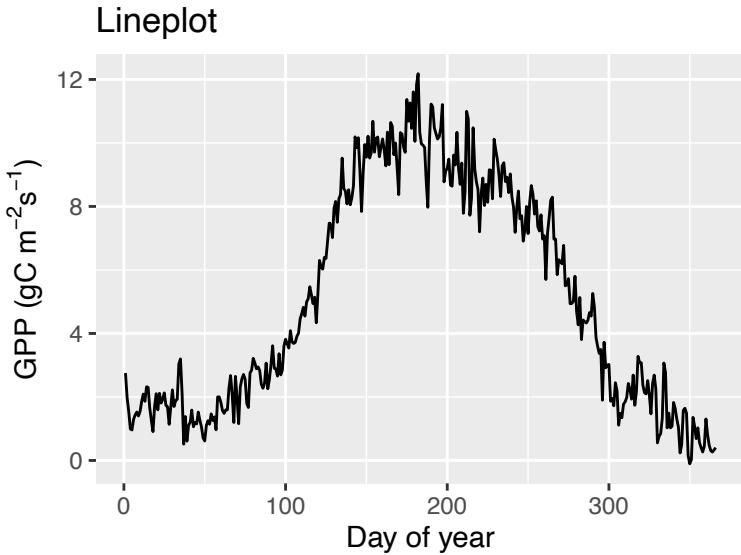
# Different “geoms” for different aspects of the data

## Time series



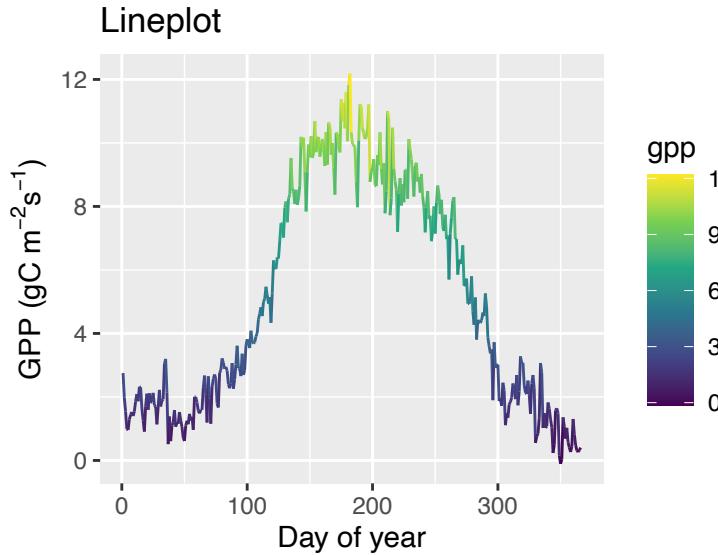
# Different “geoms” for different aspects of the data

## Time series



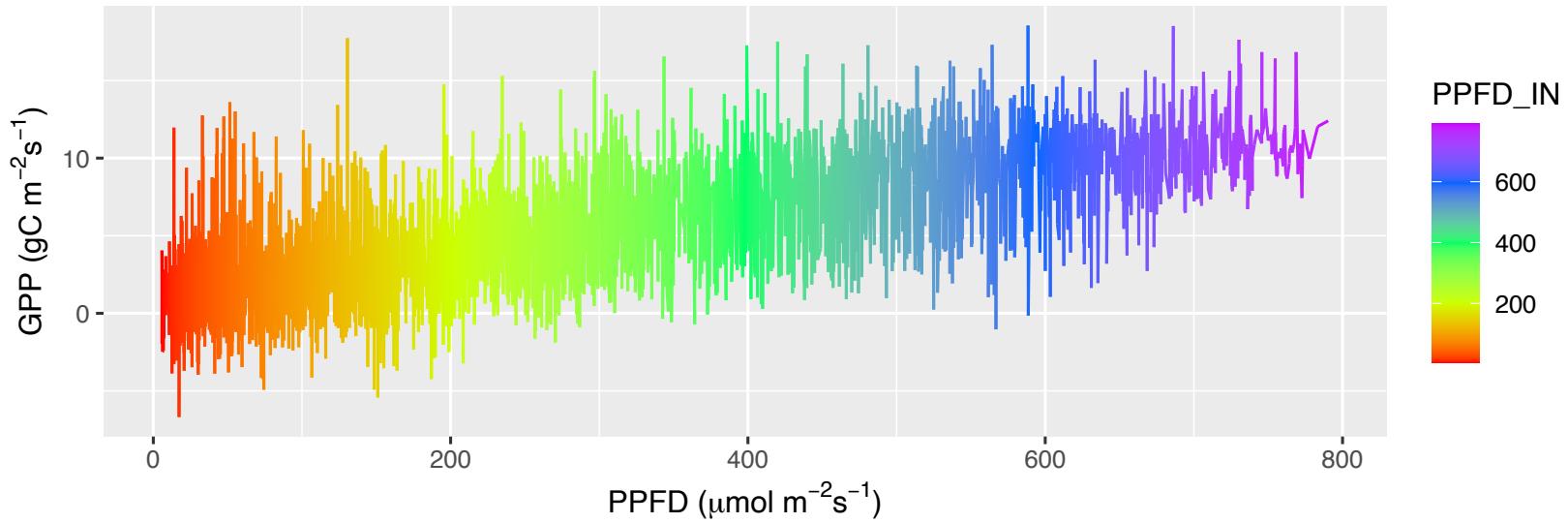
Avoid:

- Same dimension to multiple aesthetics



# Different “geoms” for different aspects of the data

Lineplot



# Different “geoms” for different aspects of the data

## Avoid:

- Non-monotonic color scale for numeric values
- Color scale not adjusted to color vision deficiency

Lineplot

