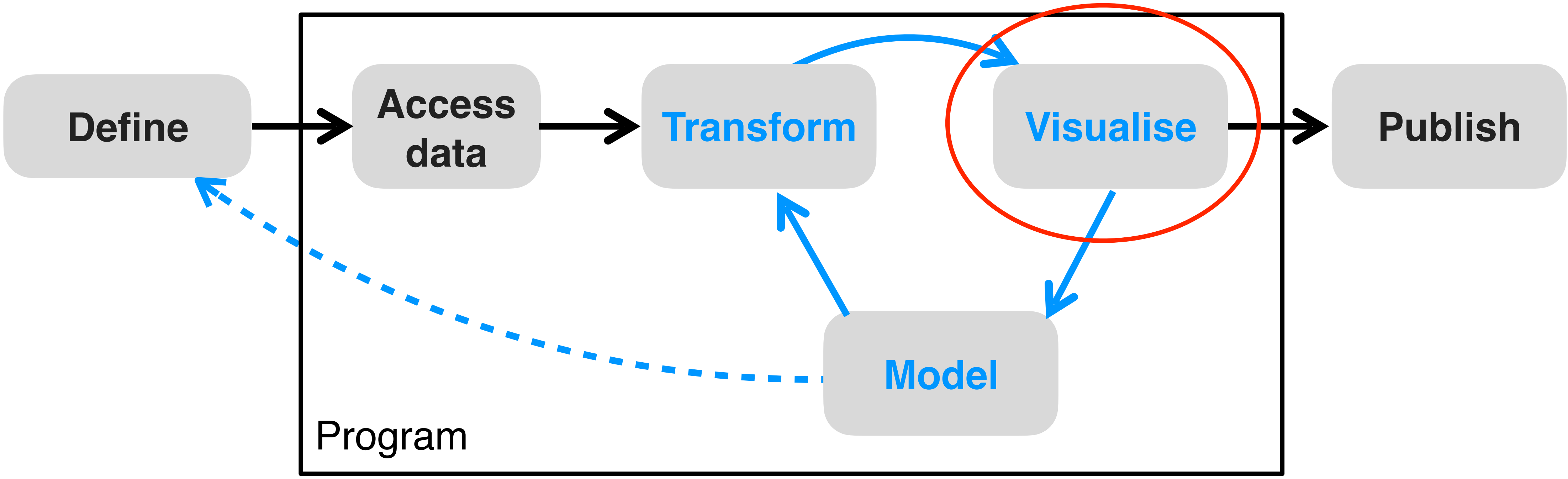
Applied Geodata Science I

Session 4

Prof. Dr. Benjamin Stocker

10.03.2025

The data science workflow



The course website: https://geco-bern.github.io/agds1_course/

Applied Geodata Science I

Introducing applications of machine learning in Geography and Environmental Sciences

AUTHOR
Benjamin Stocker

PUBLISHED
February 13, 2025

About this course

Course description

This course introduces essential methods of the data science workflow, supported by Geodata Science and demonstrates common applications of machine learning metho Environmental Sciences by worked examples.

Table of contents

Applied Geodata Science

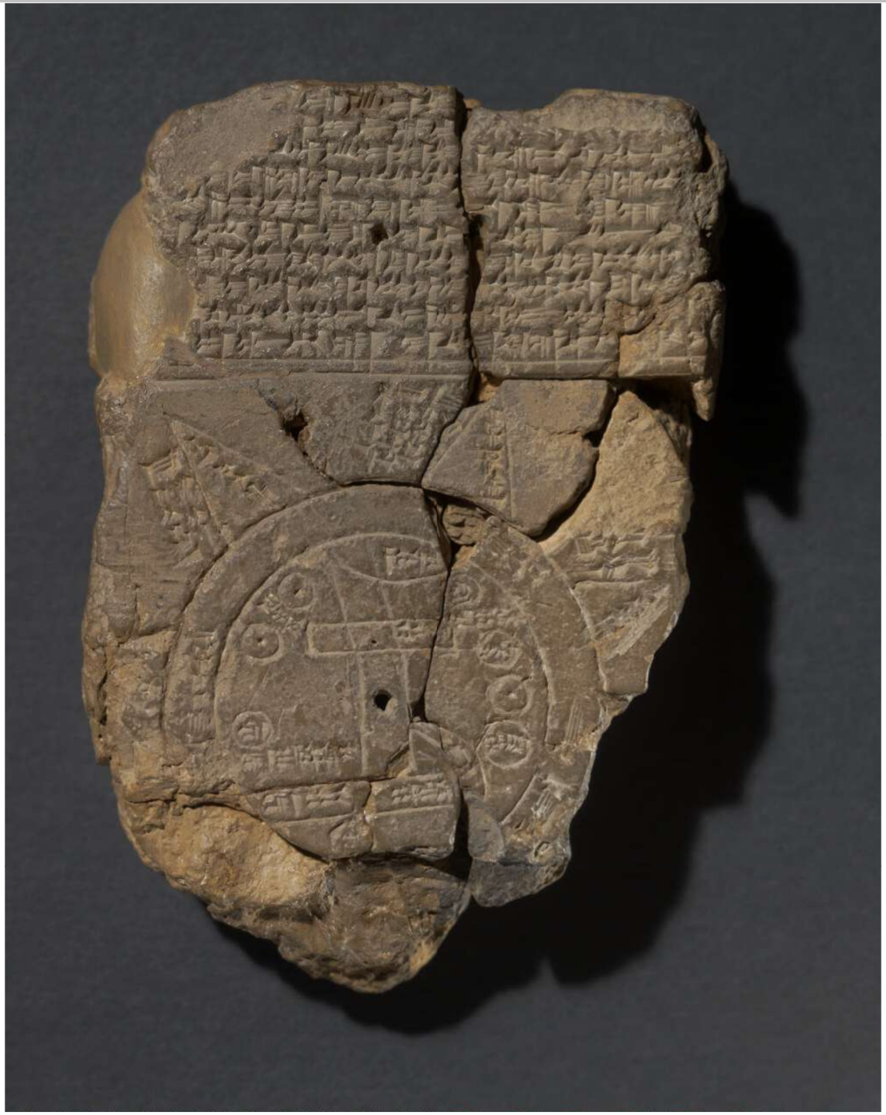
AUTHOR
Benjamin Stocker, Koen Hufkens, Pepa Arán, and Pascal Schneider

PUBLISHED
November 7, 2023

Table of contents

Preface

Earliest data visualisation?



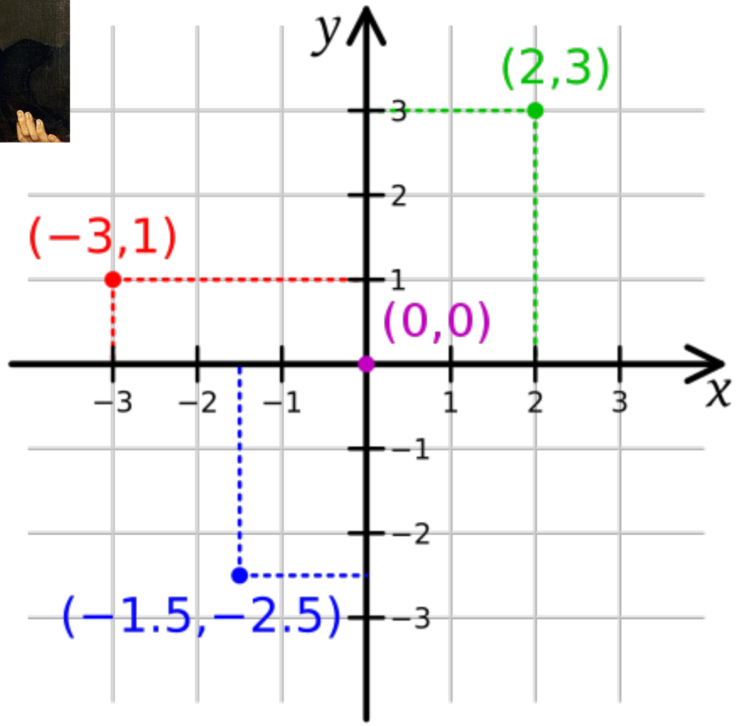
Babylon, ca. 500 BC

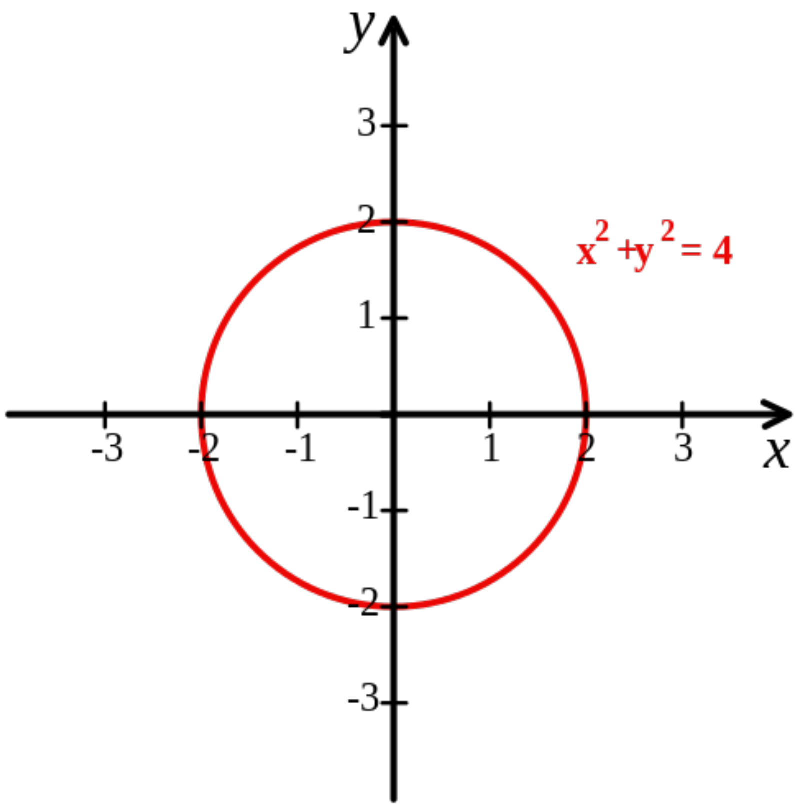
https://www.britishmuseum.org/collection/object/W_1882-0714-509

Cartesian coordinate system

René Descartes
1596-1650

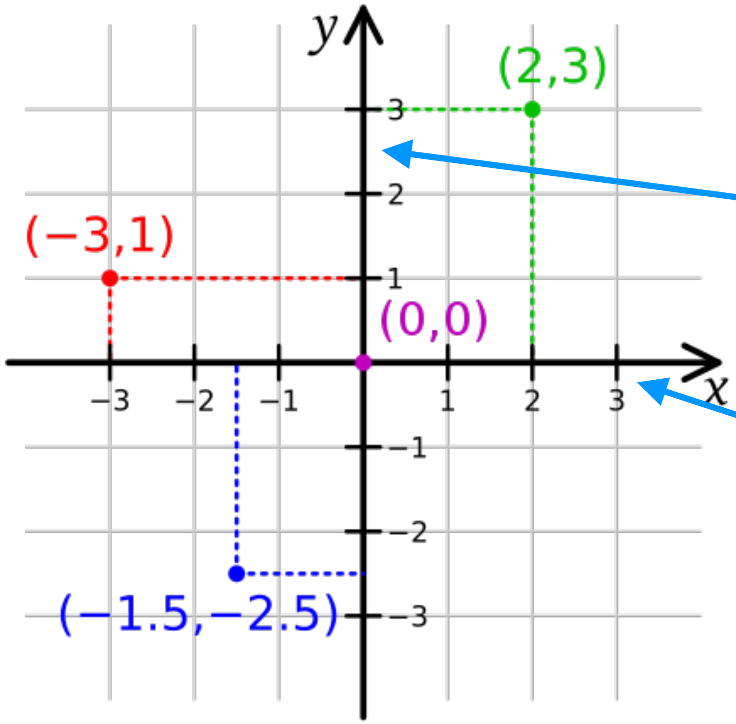
Link between geometry and algebra!





https://en.wikipedia.org/wiki/Cartesian_coordinate_system

Cartesian coordinate system



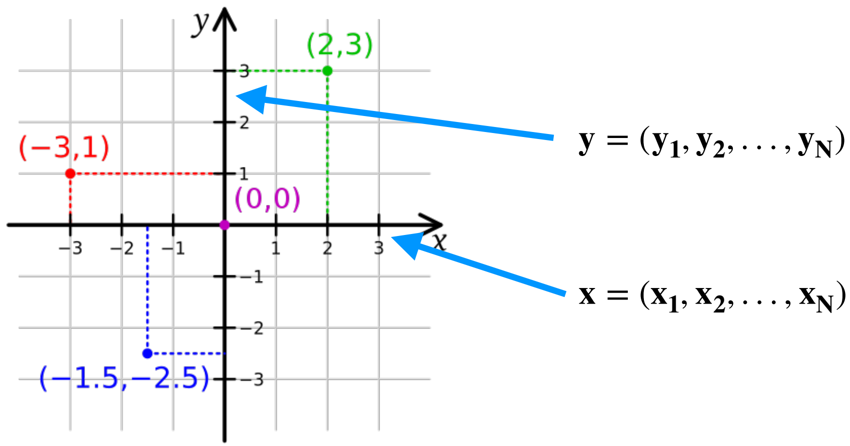
$$\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$$

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$$

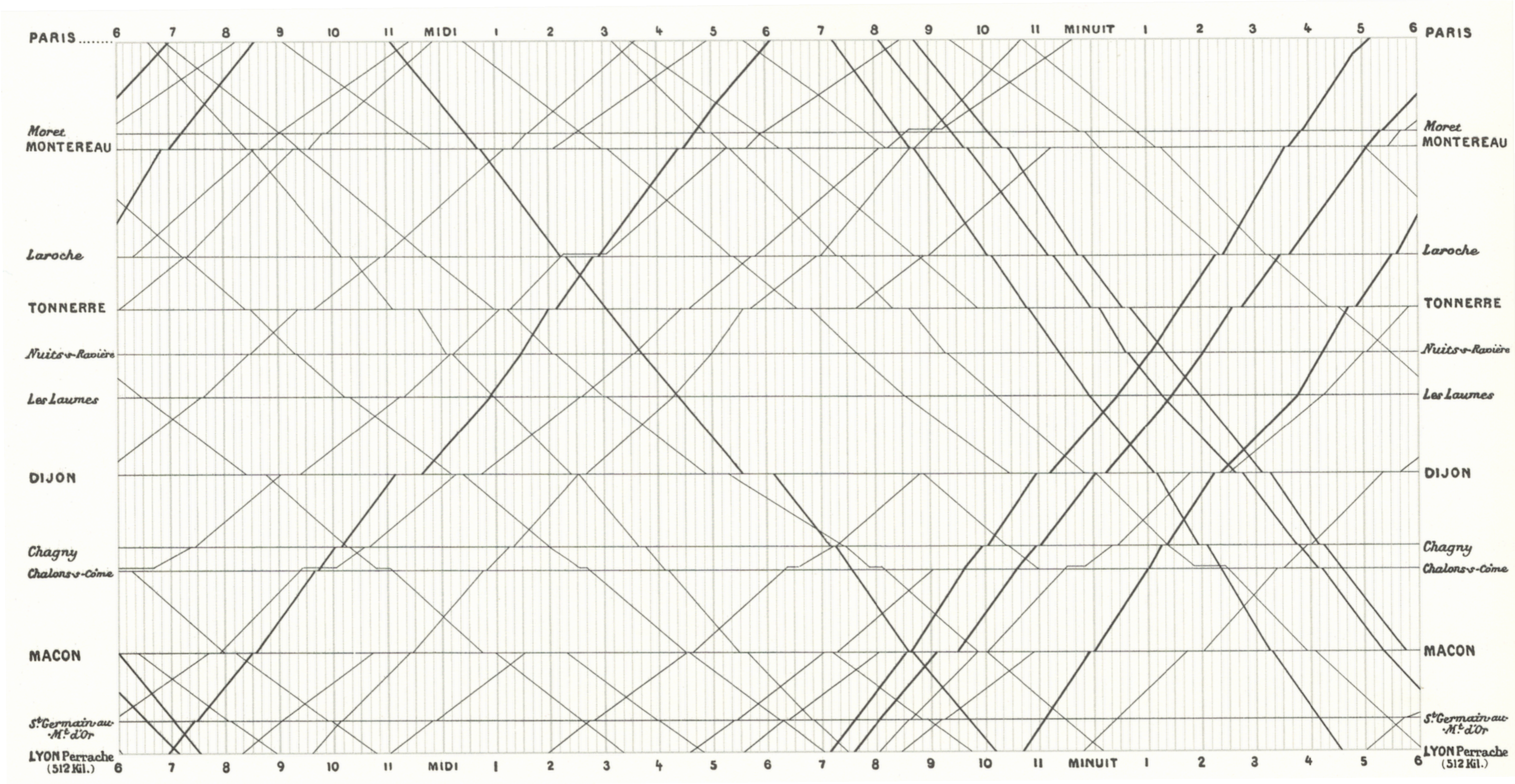
https://en.wikipedia.org/wiki/Cartesian_coordinate_system

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.

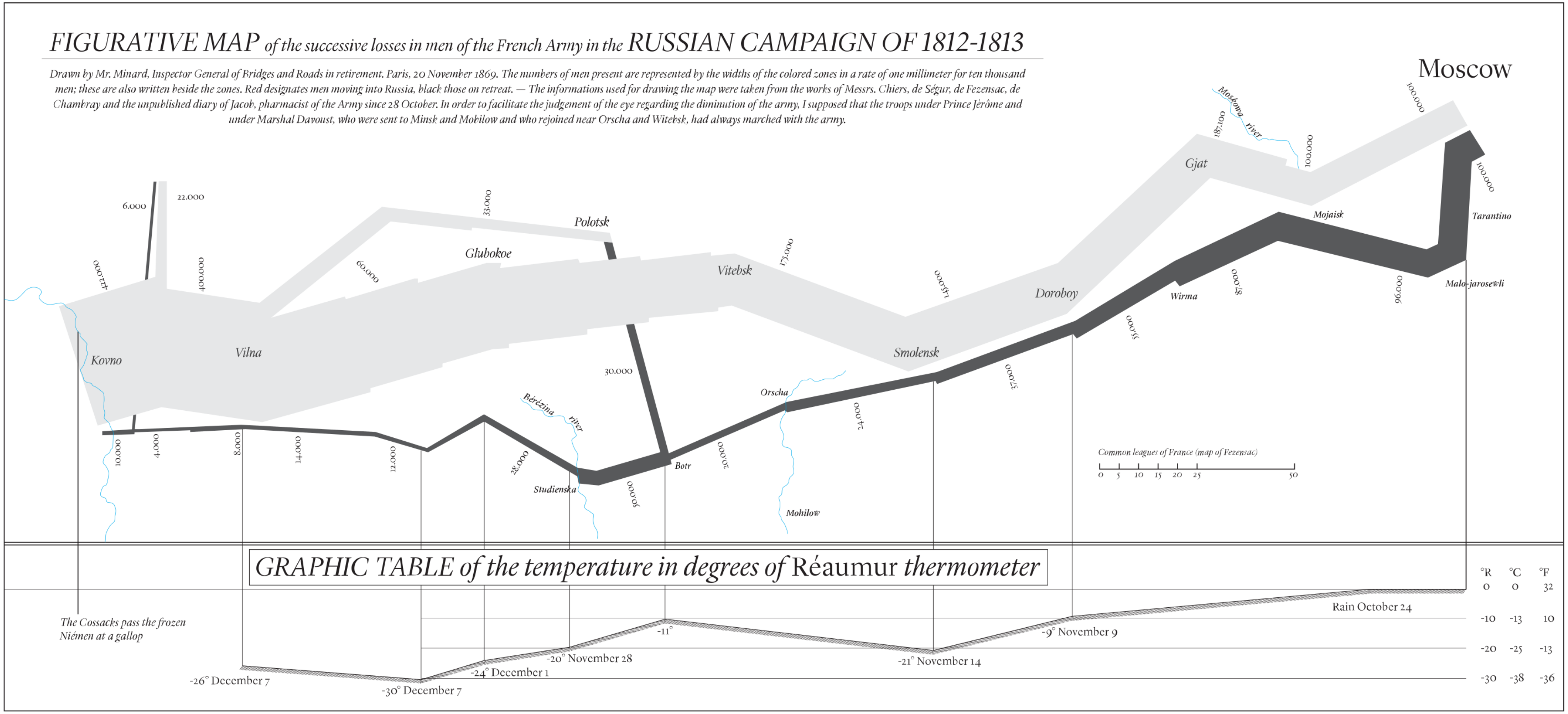


Visualising time



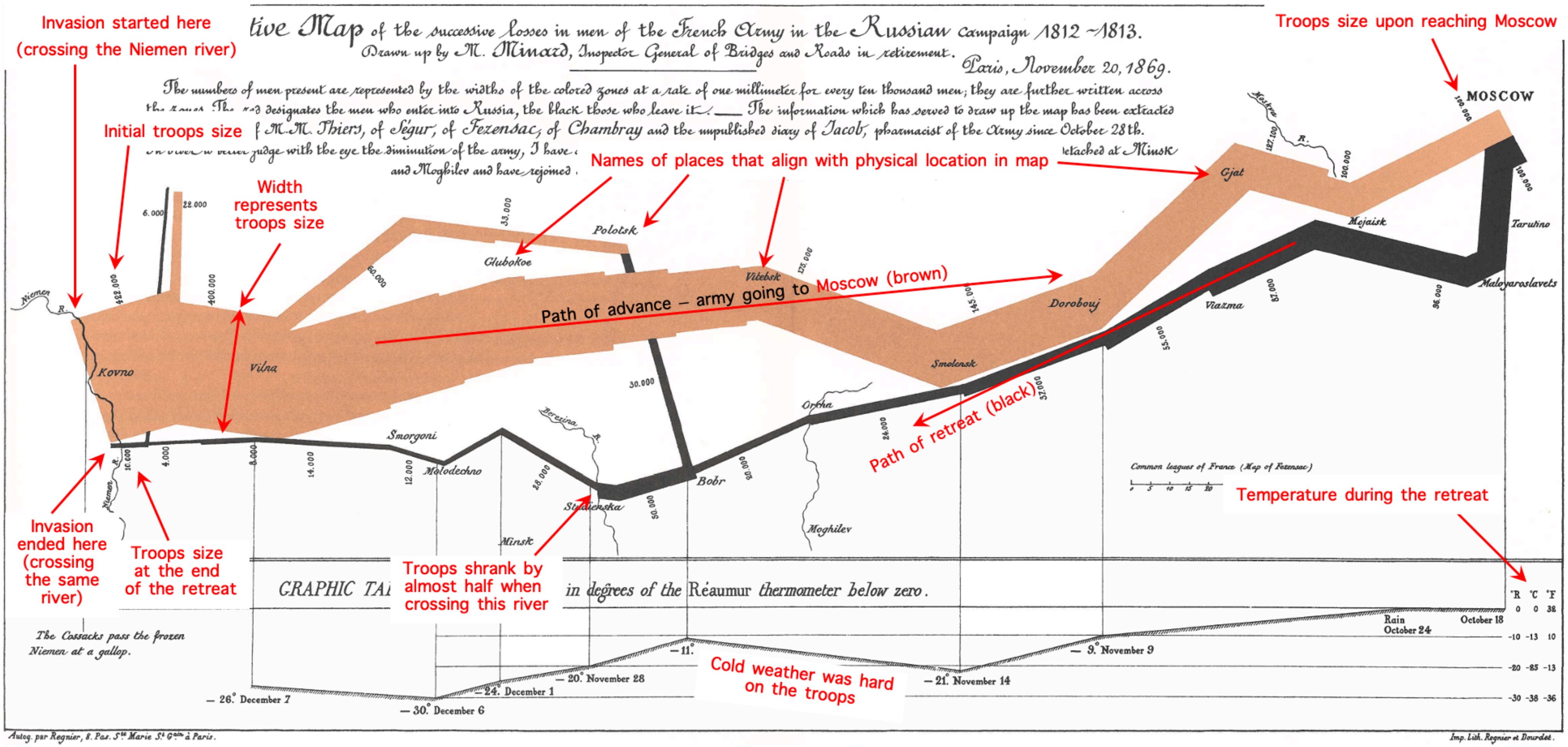
E. J. Marey, 1885
https://badriadhikari.github.io/data-viz-workshop-2021/minards/

Quantitative information in geographical space



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

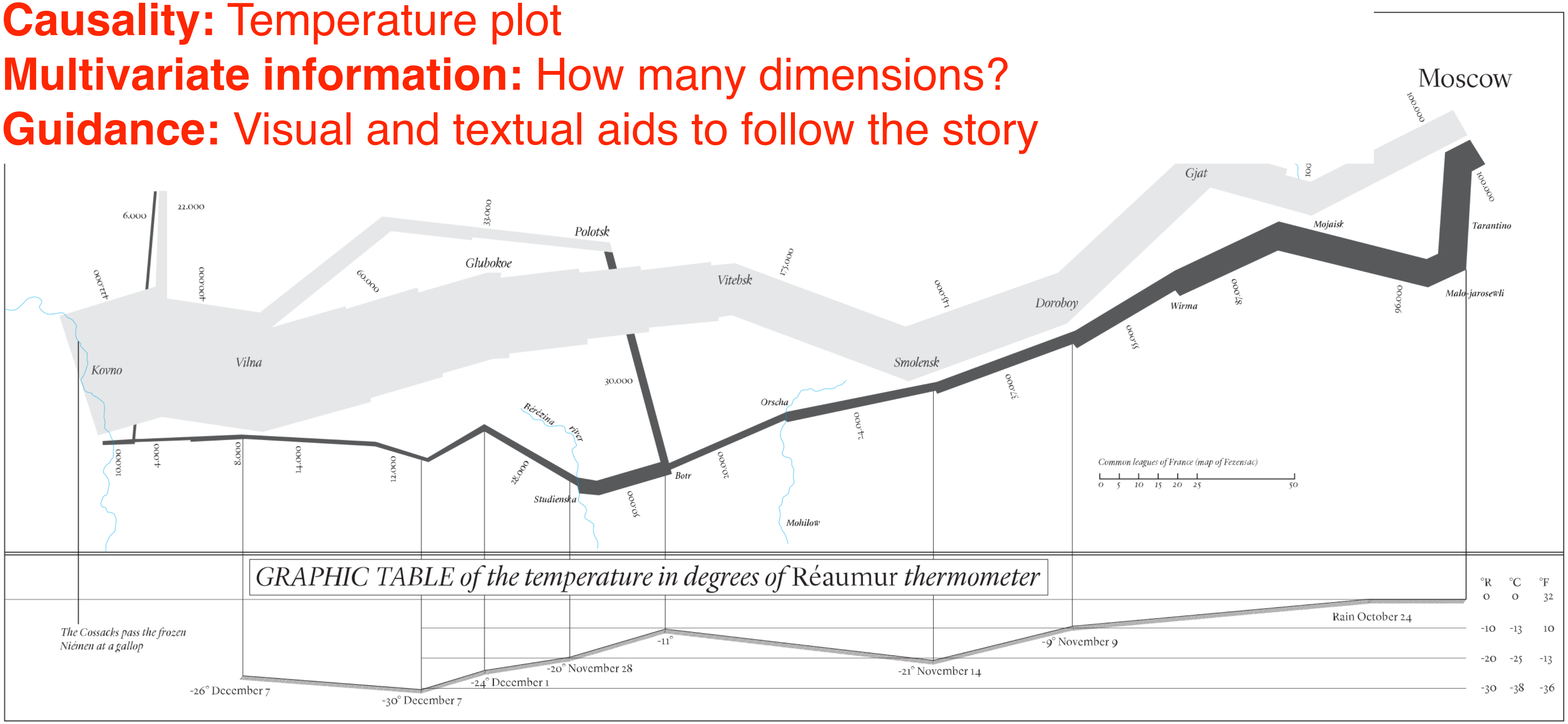
Quantitative information in geographical space



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



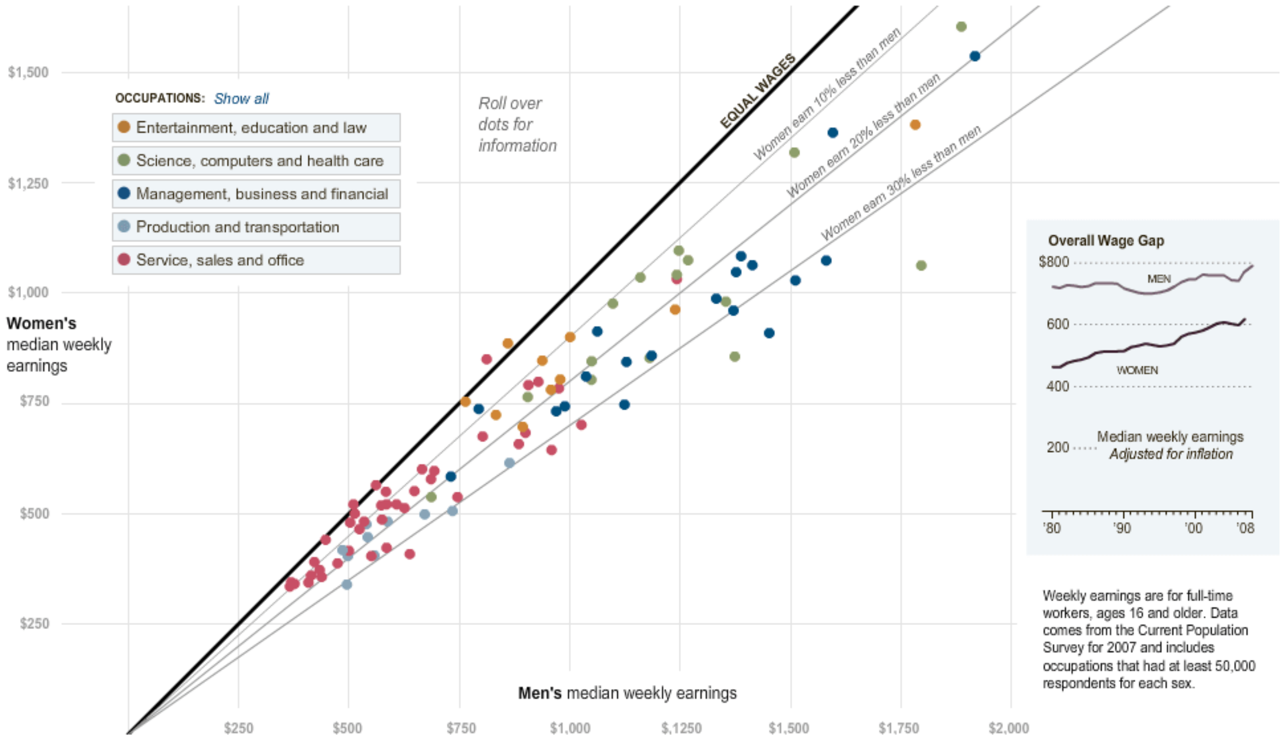
Charles-Joseph Minard, 1869
https://badriadhikari.github.io/data-viz-workshop-2021/minards/

Scatterplots

Published: May 18, 2010

Why Is Her Paycheck Smaller?

Nearly every occupation has the gap — the seemingly unbridgeable chasm between the size of the paycheck brought home by a woman and the larger one earned by a man doing the same job. Economists cite a few reasons: discrimination as well as personal choices within occupations are two major factors, and part of the gap can be attributed to men having more years of experience and logging more hours.



Hannah Fairfield and Graham Roberts/The New York Times | Send Feedback

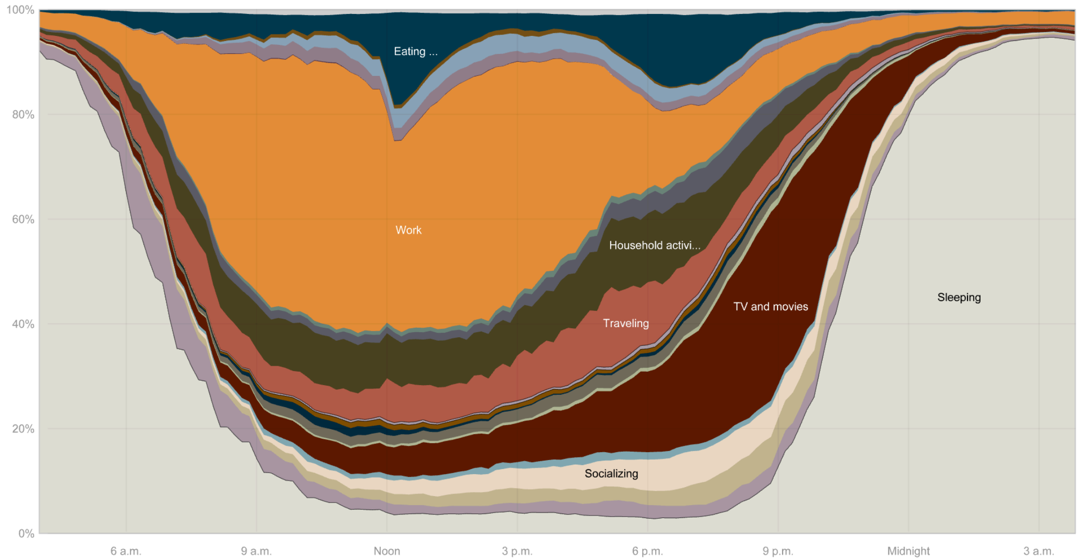
RECOMMEND TWITTER LINKEDIN SIGN IN TO E-MAIL SHARE

Sources: Bureau of Labor Statistics; Census Bureau

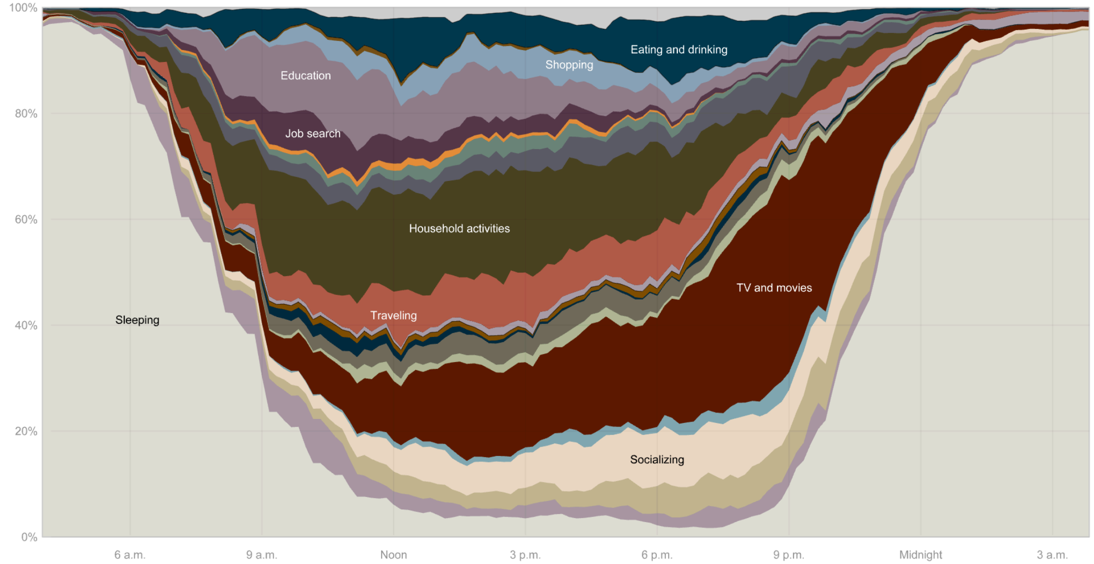
Hannah Fairfield and Graham Roberts
New York Times, 2010
https://eagereyes.org/journalism/the-explanatory-power-of-data-points

Proportions over time

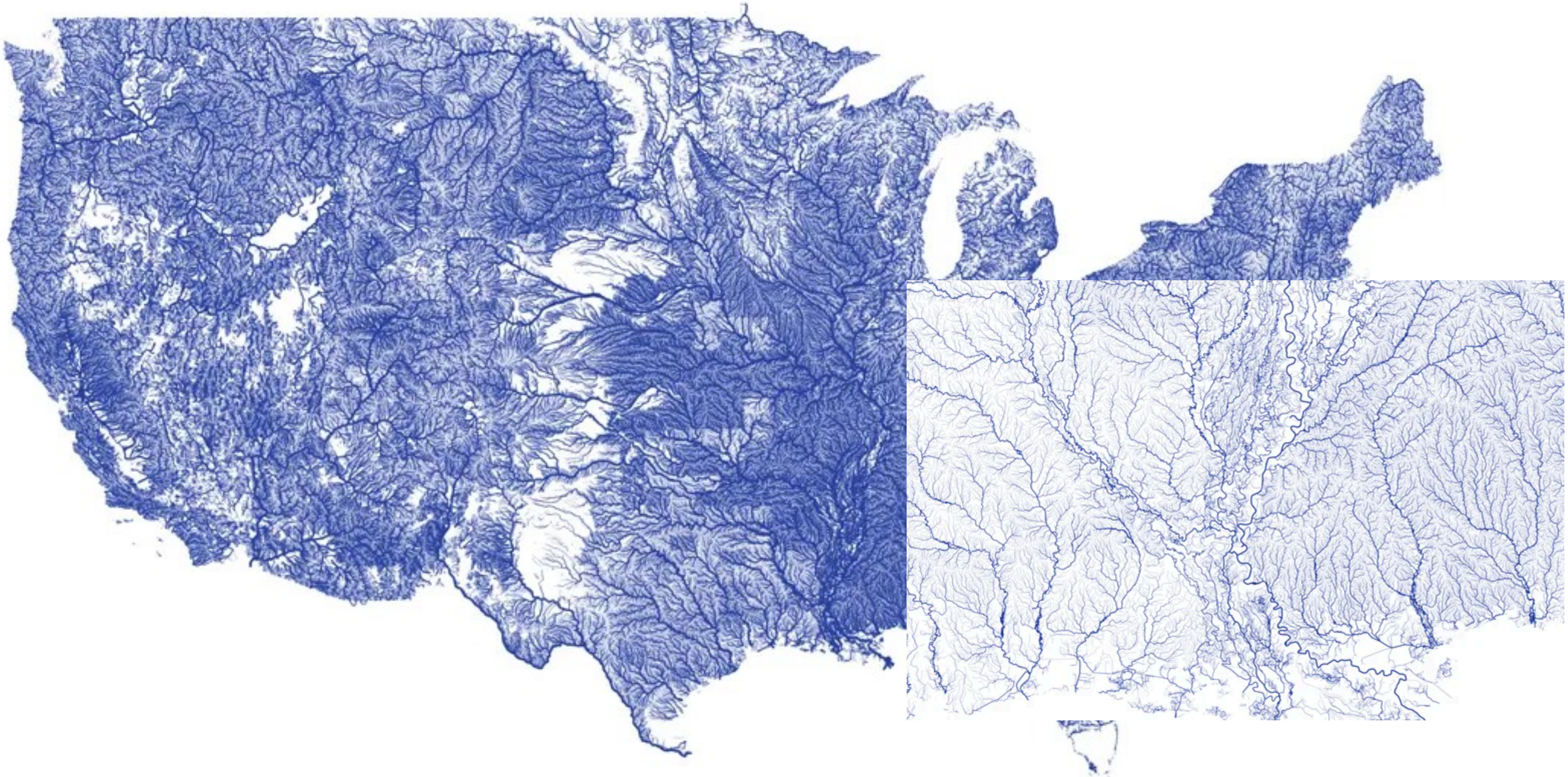
Employed



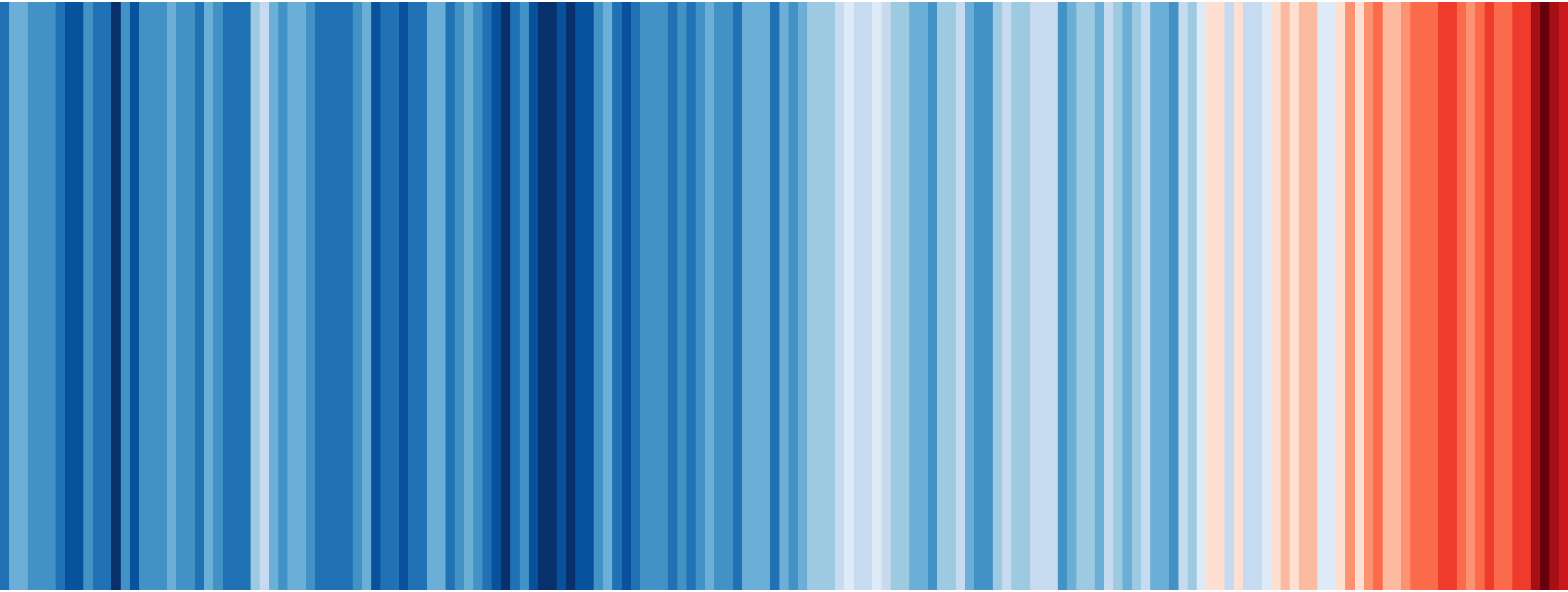
Unemployed



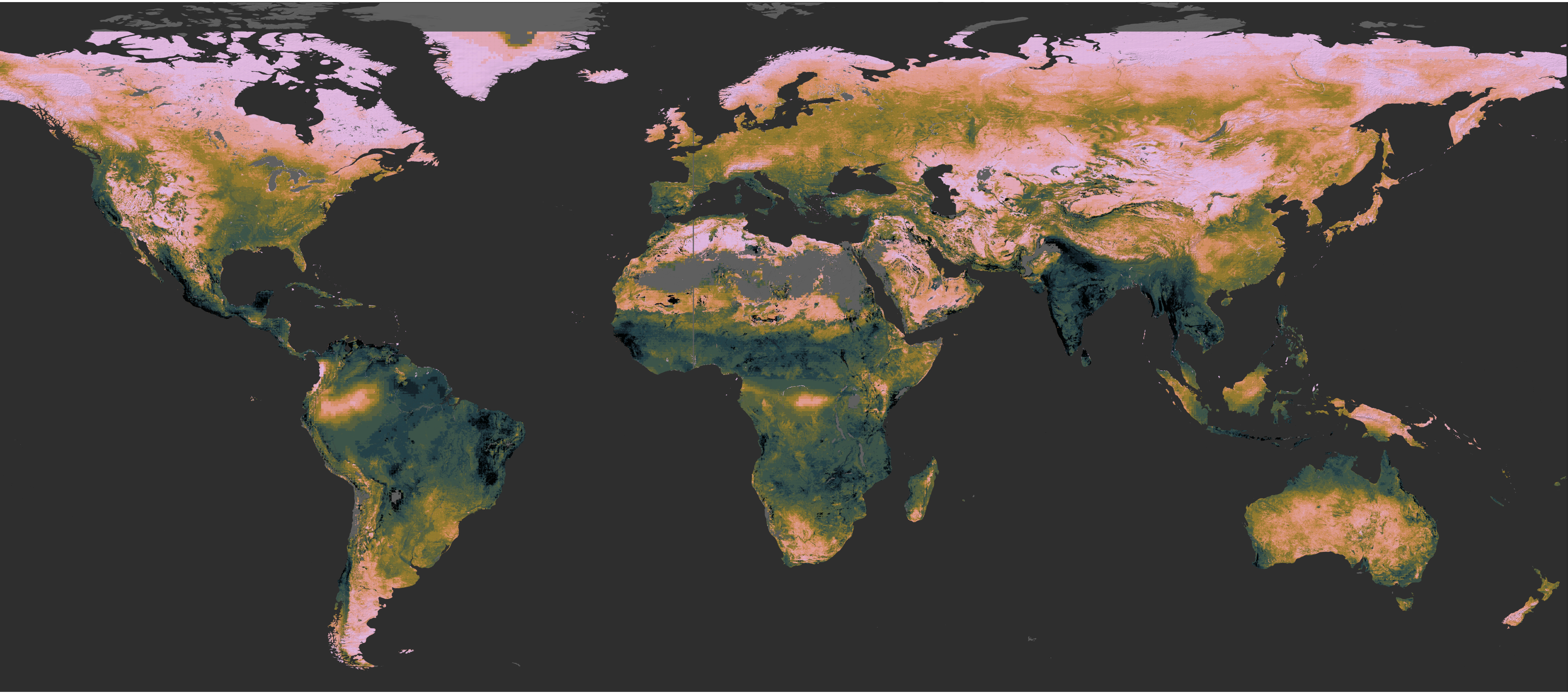
Amanda Cox, *New York Times*
https://gestalten.com/blogs/journal/visualizing-a-new-new-york-times



Nelson Minar,
https://www.wired.com/2013/06/infographic-this-detailed-map-shows-every-river-in-the-united-states/?viewall=true

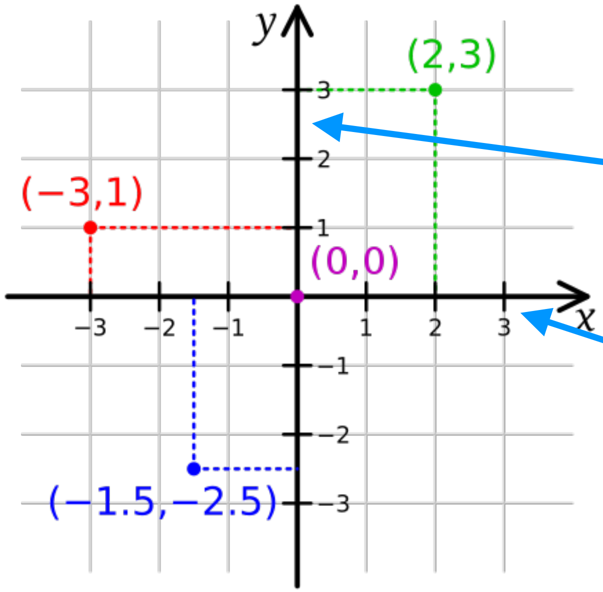


Ed Hawkins
https://www.climate-lab-book.ac.uk/2018/2018-visualisation-update/



Stocker et al., 2023 *Nature Geoscience*

Cartesian coordinate system



$$\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$$

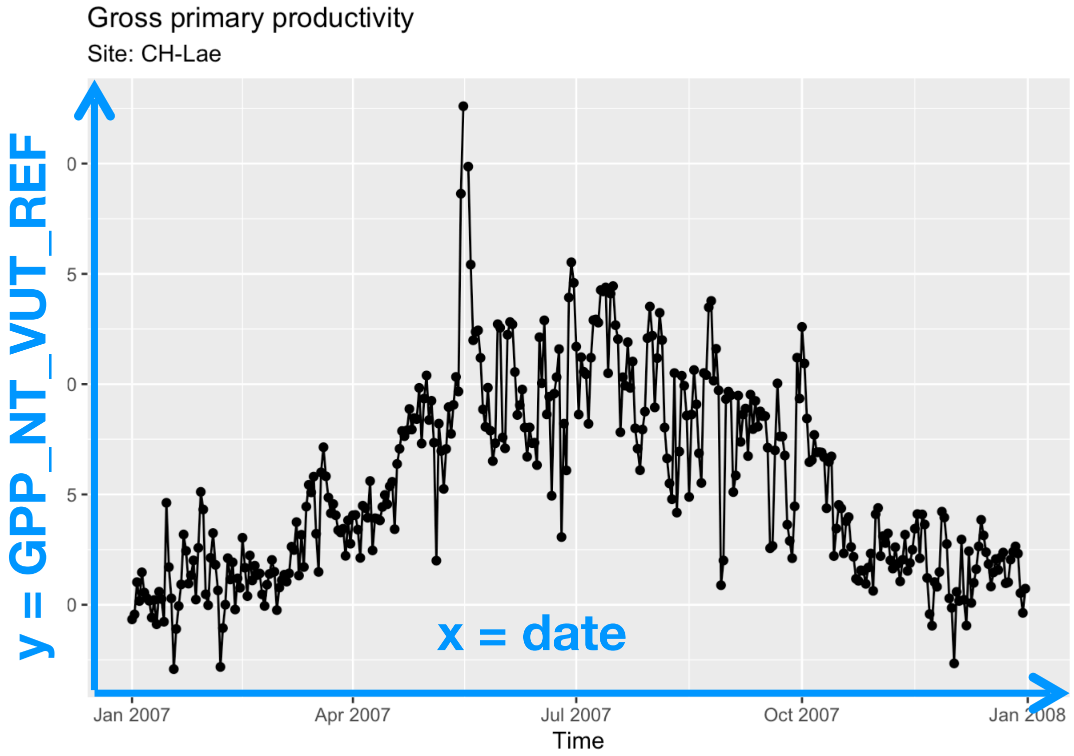
$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$$

- We "map" data values onto quantifiable features of the resulting graphic - the aesthetics.

https://en.wikipedia.org/wiki/Cartesian_coordinate_system

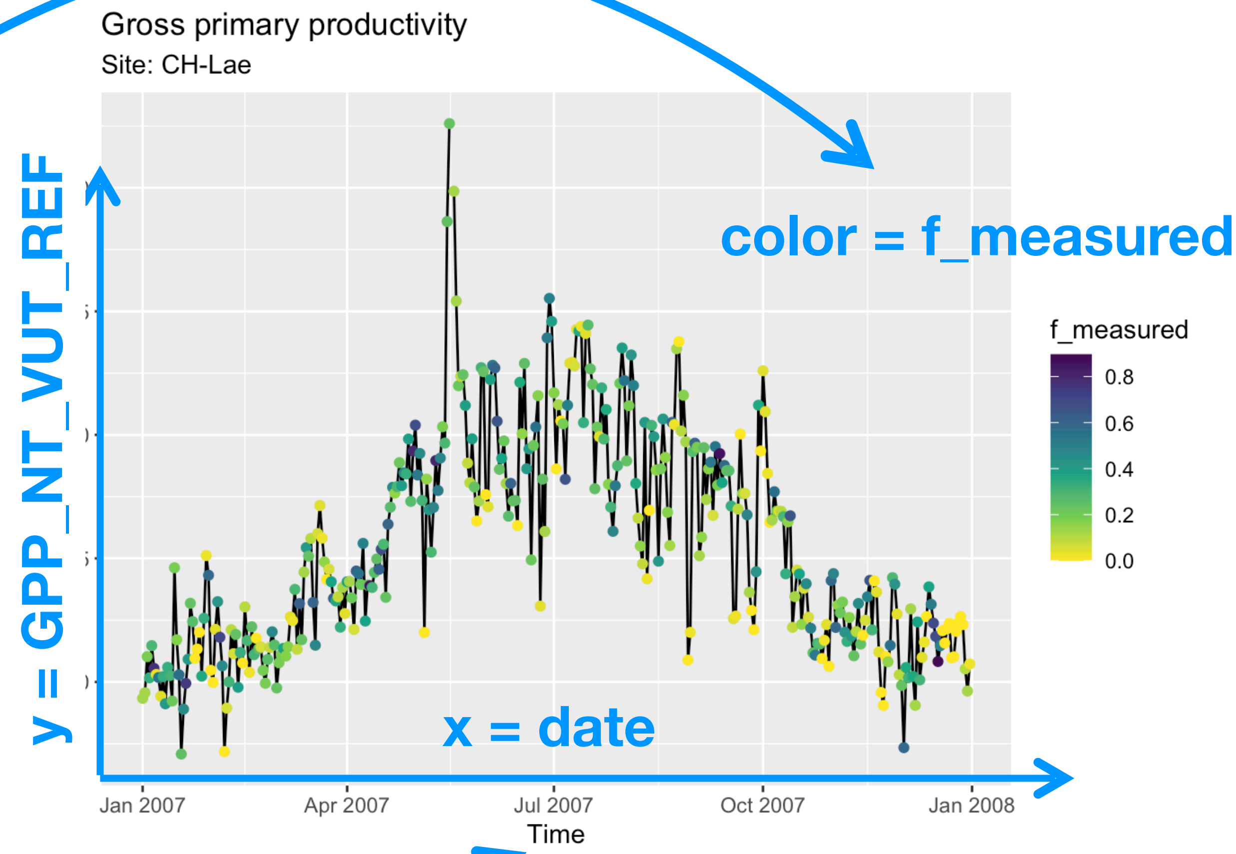
Mapping to two aesthetics

```
> 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 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      0.0417
 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
```



ggplot2

A data frame

Aesthetics mapping

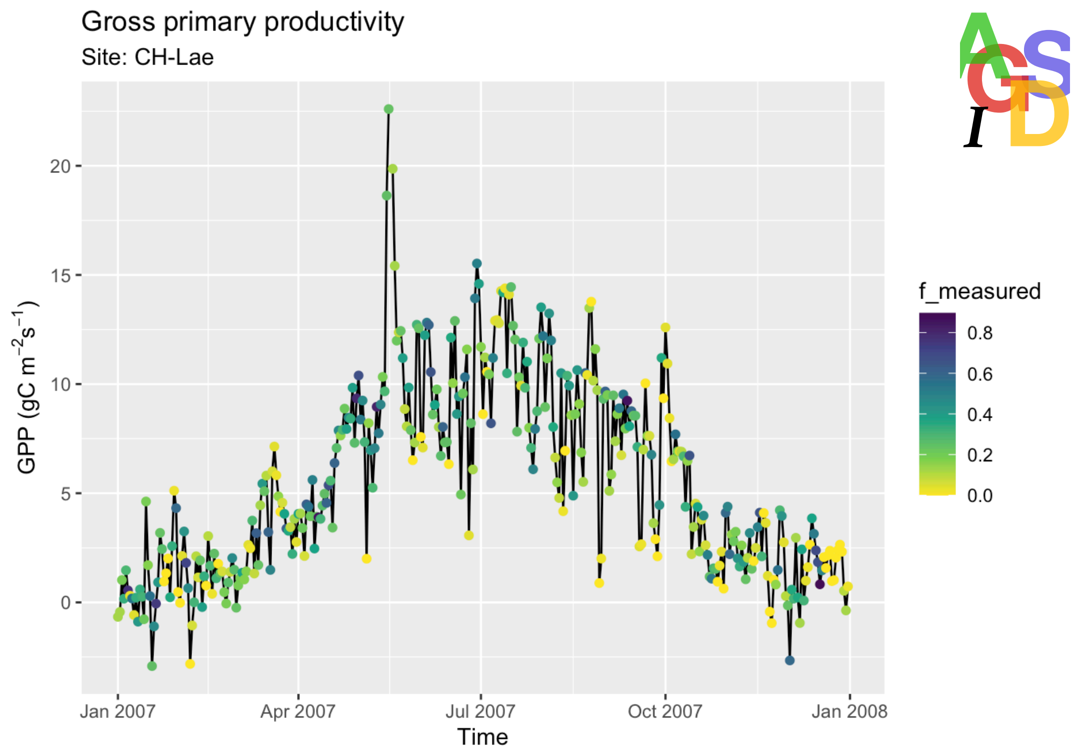
```
ddf %>%
  ggplot(aes(x = date, y = GPP_NT_VUT_REF)) +
  geom_line() +
  geom_point()
```

(A second visualisation element)

Visualisation element. Function interprets ‘x’ and ‘y’.

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      0.0417
 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
```



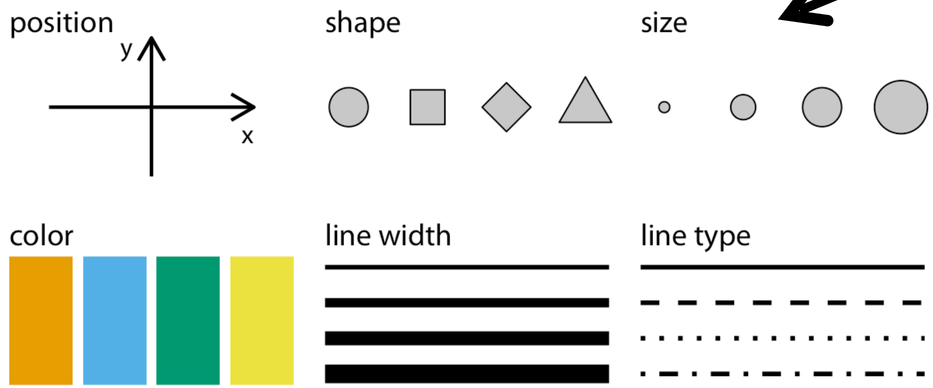

```
ddf %>%
  ggplot(aes(x = date, y = GPP_NT_VUT_REF, color = f_measured)) +
  geom_line() +
  geom_point() +
  scale_color_viridis_c()
```

Aesthetics for continuous and categorical data

Continuous but…

Categorical

Continuous



Continuous: color gradient
Categorical: color set

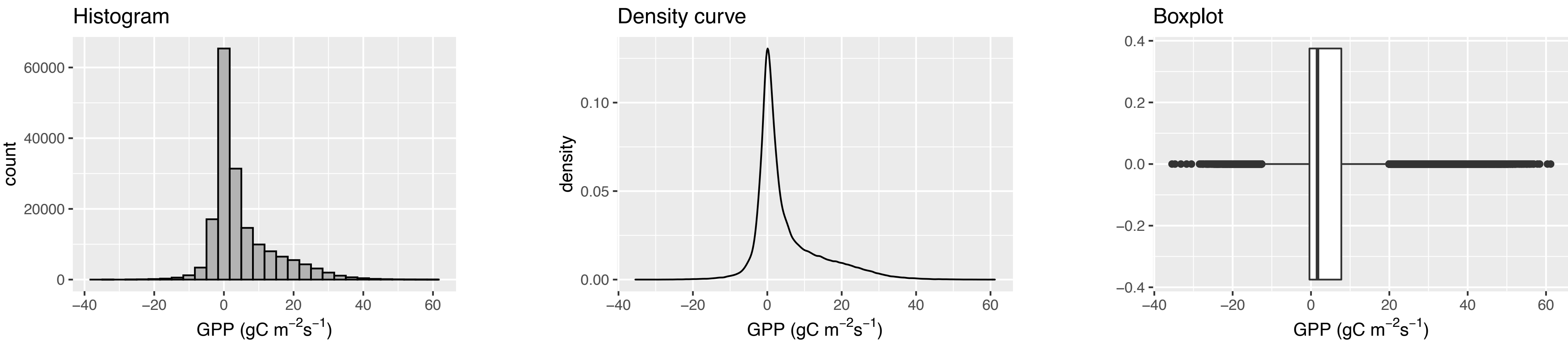
Continuous

Categorical

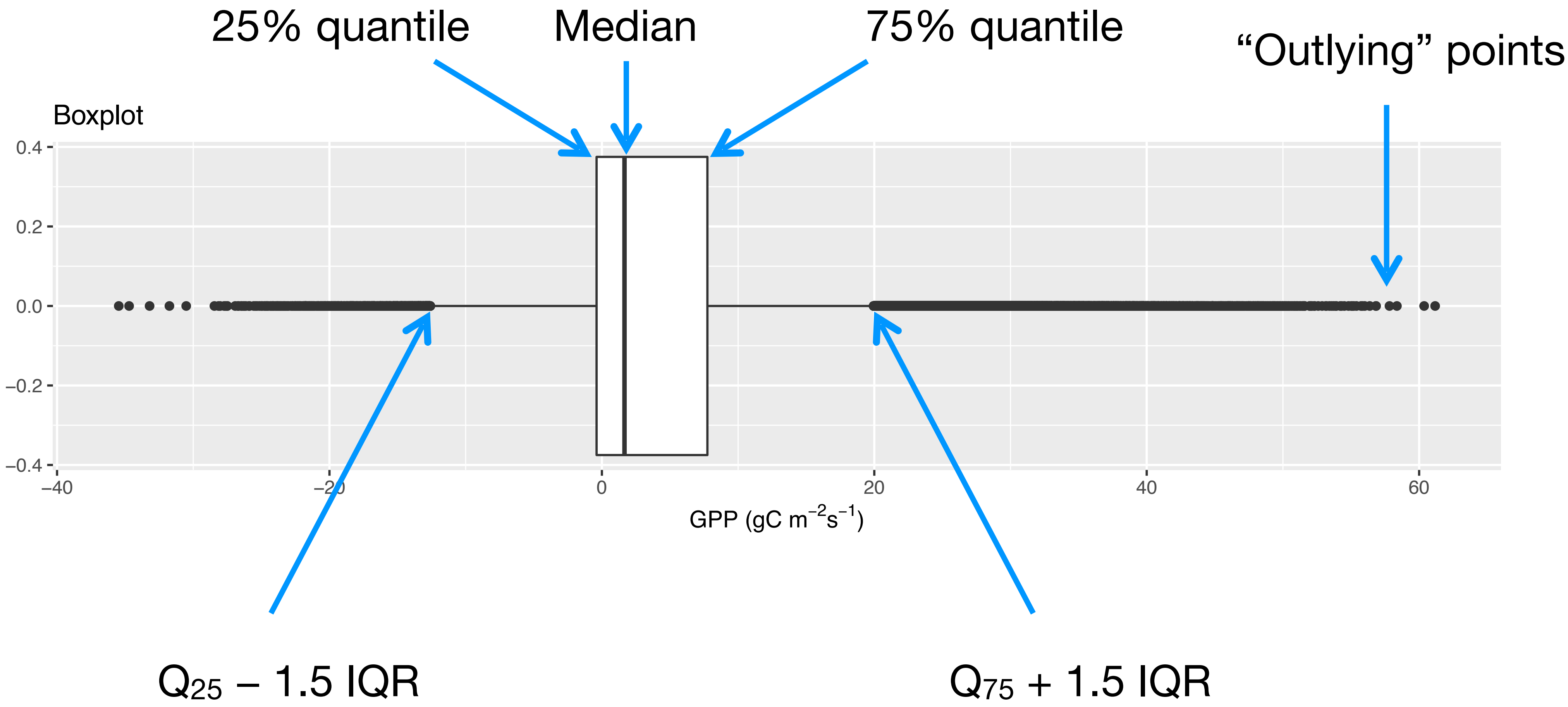
Figure from: Wilke, Data Visualization
https://clauswilke.com/dataviz/aesthetic_mapping_files/figure-html/common-aesthetics-1.png

Different “geoms” for different aspects of the data

Distributions

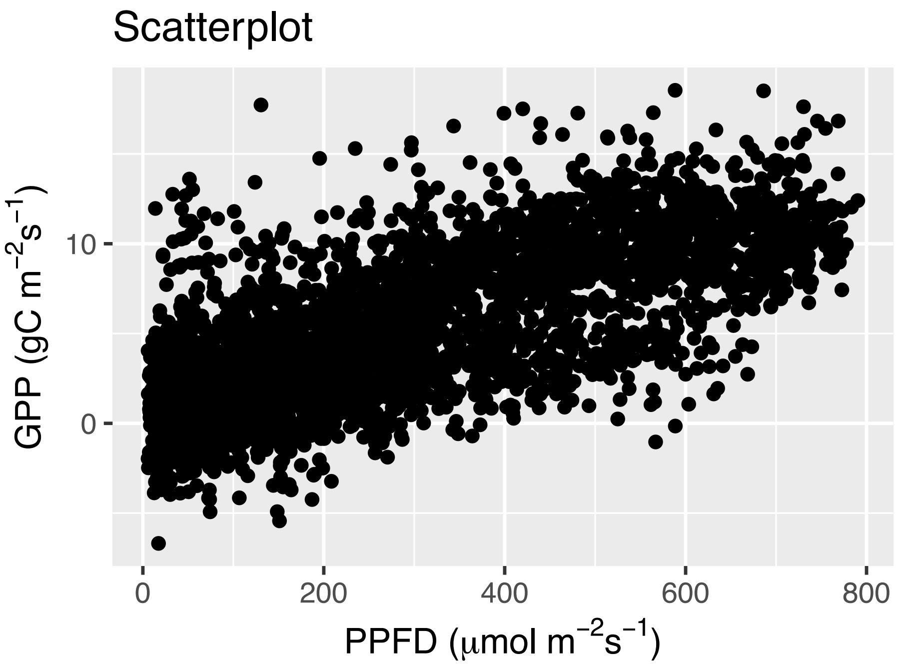


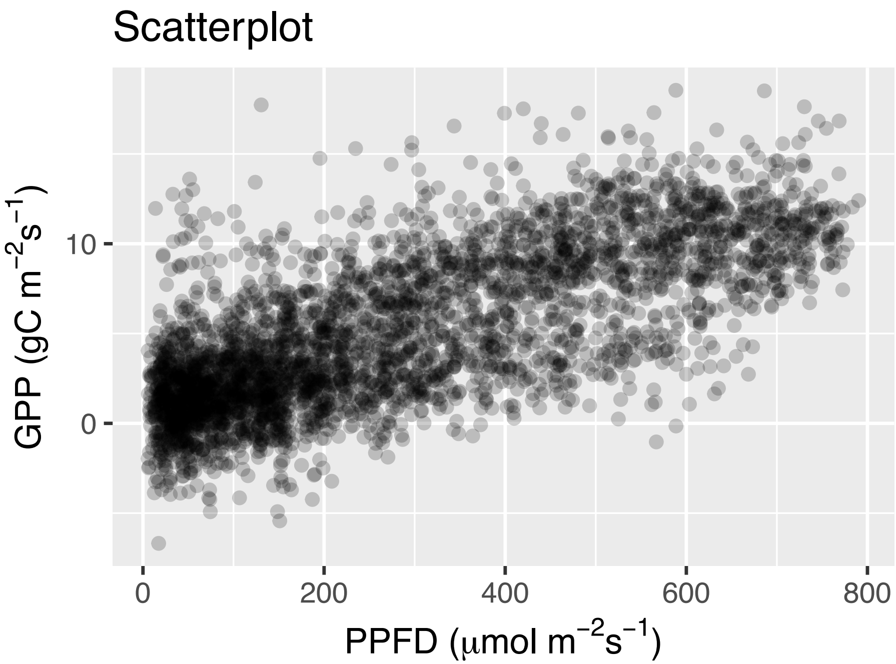
Boxplot


25% quantile
Median
75% quantile
"Outlying" points
Boxplot
0.4
0.2
0.0
−0.2
−0.4
−40
−20
0
20
40
60
GPP (gC m^{-2}s^{-1})
Q$_{25}$ – 1.5 IQR
Q$_{75}$ + 1.5 IQR

Different “geoms” for different aspects of the data

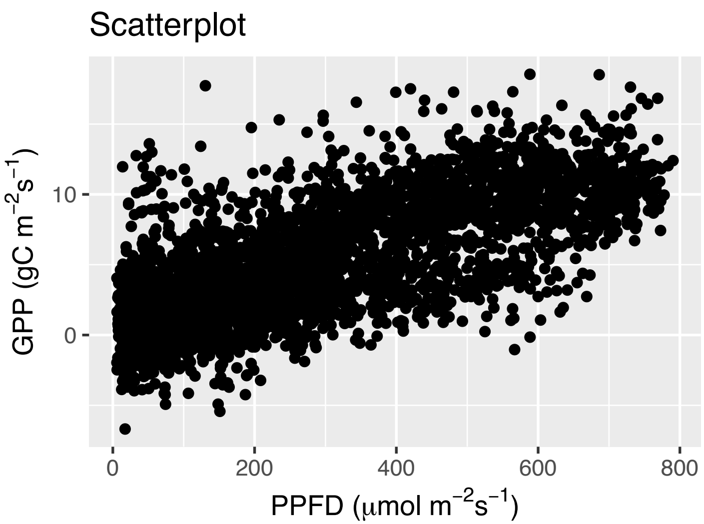
Relationships

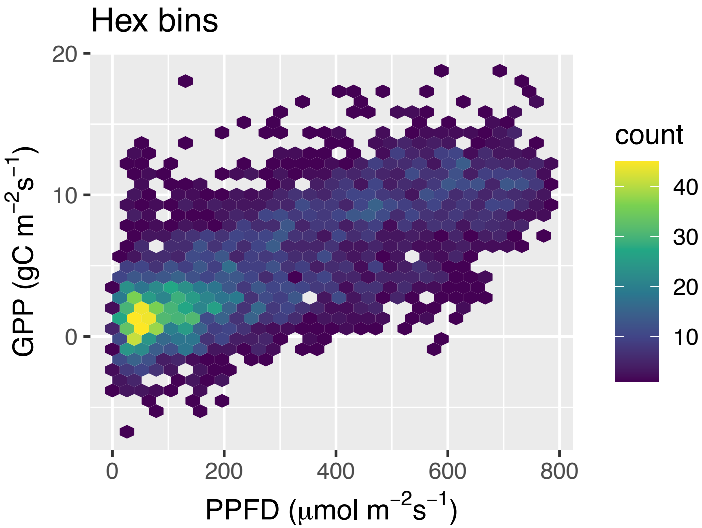


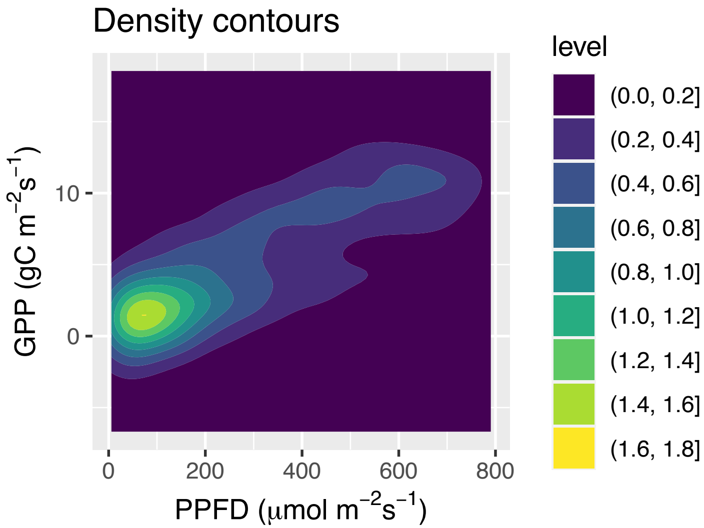


Different “geoms” for different aspects of the data

Relationships



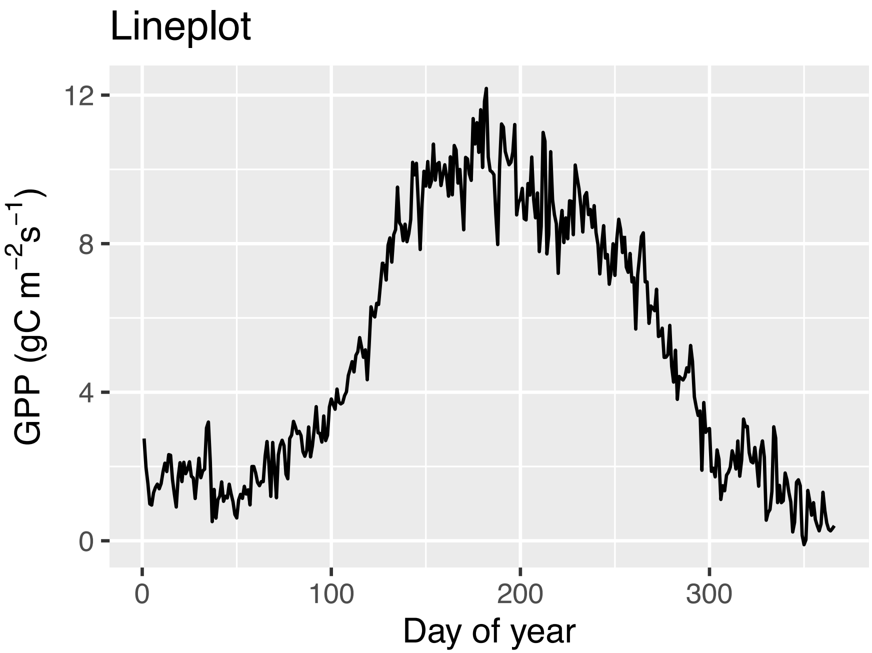


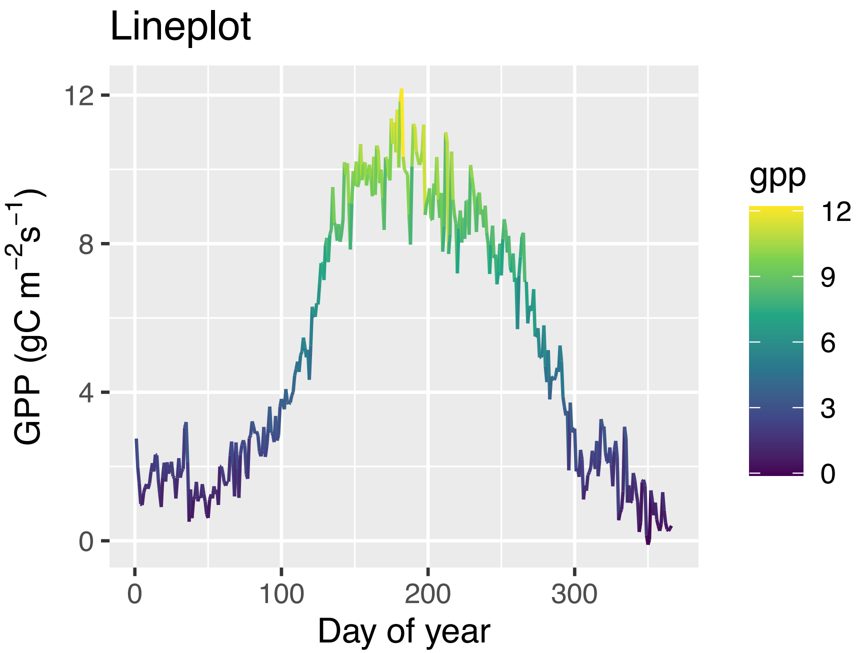


Different "geoms" for different aspects of the data

Time series

Avoid:
- Same dimension to multiple aesthetics

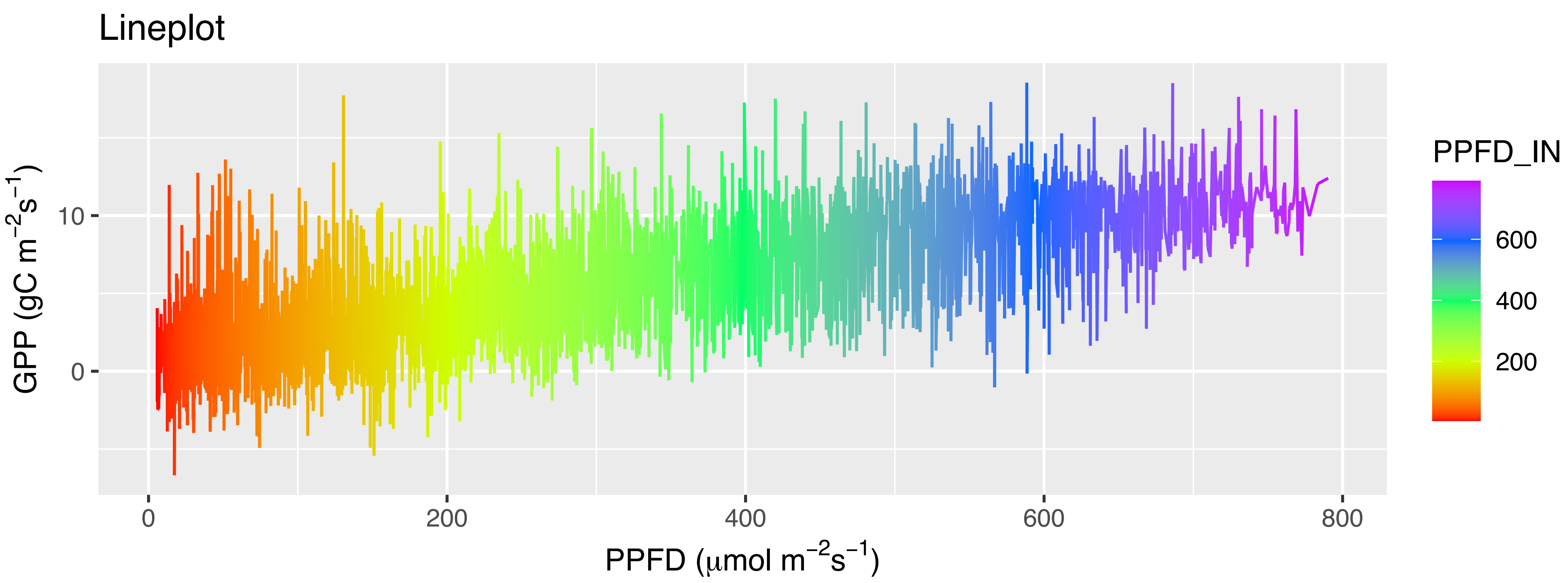




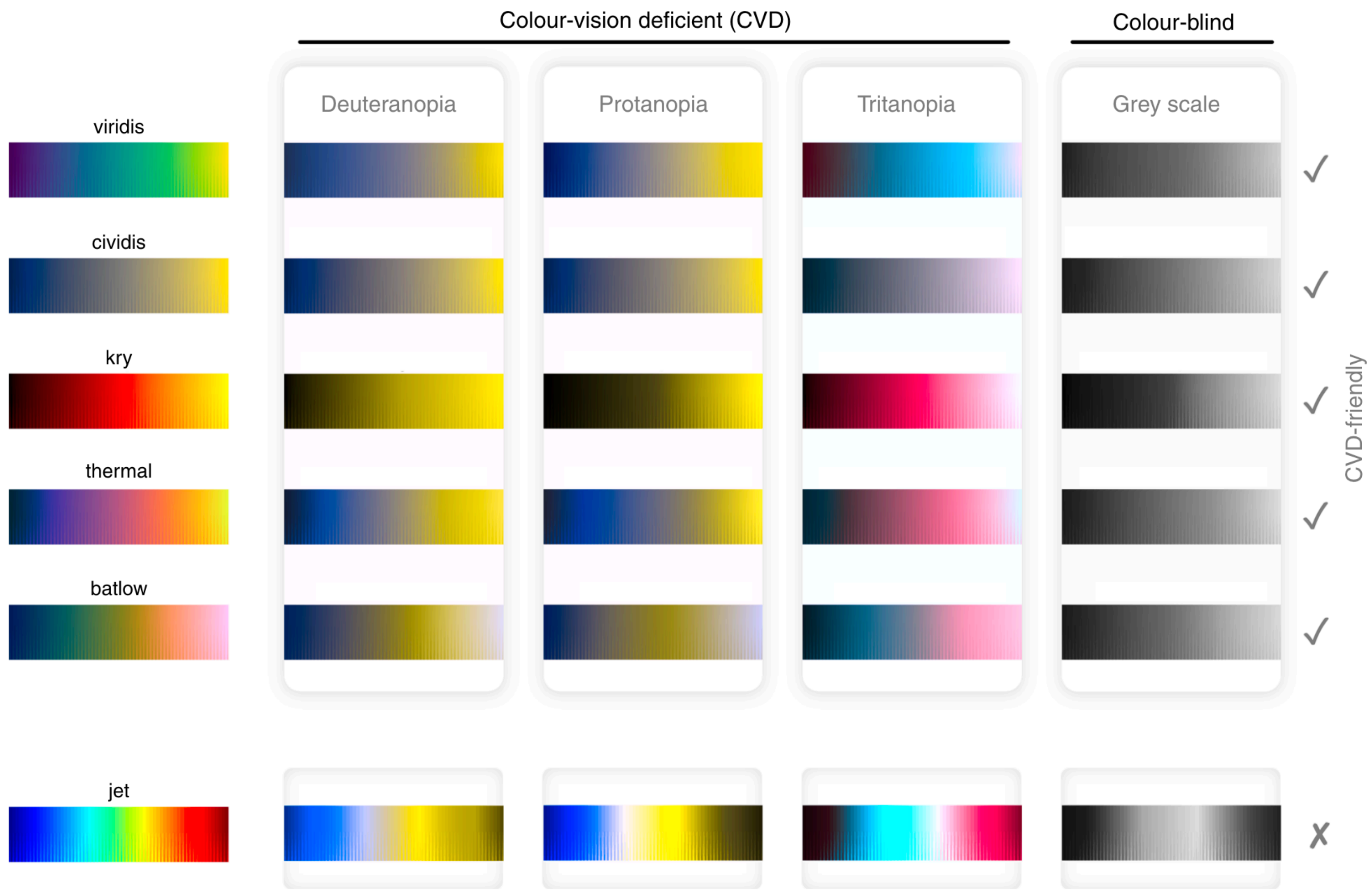
Different “geoms” for different aspects of the data

Avoid:

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

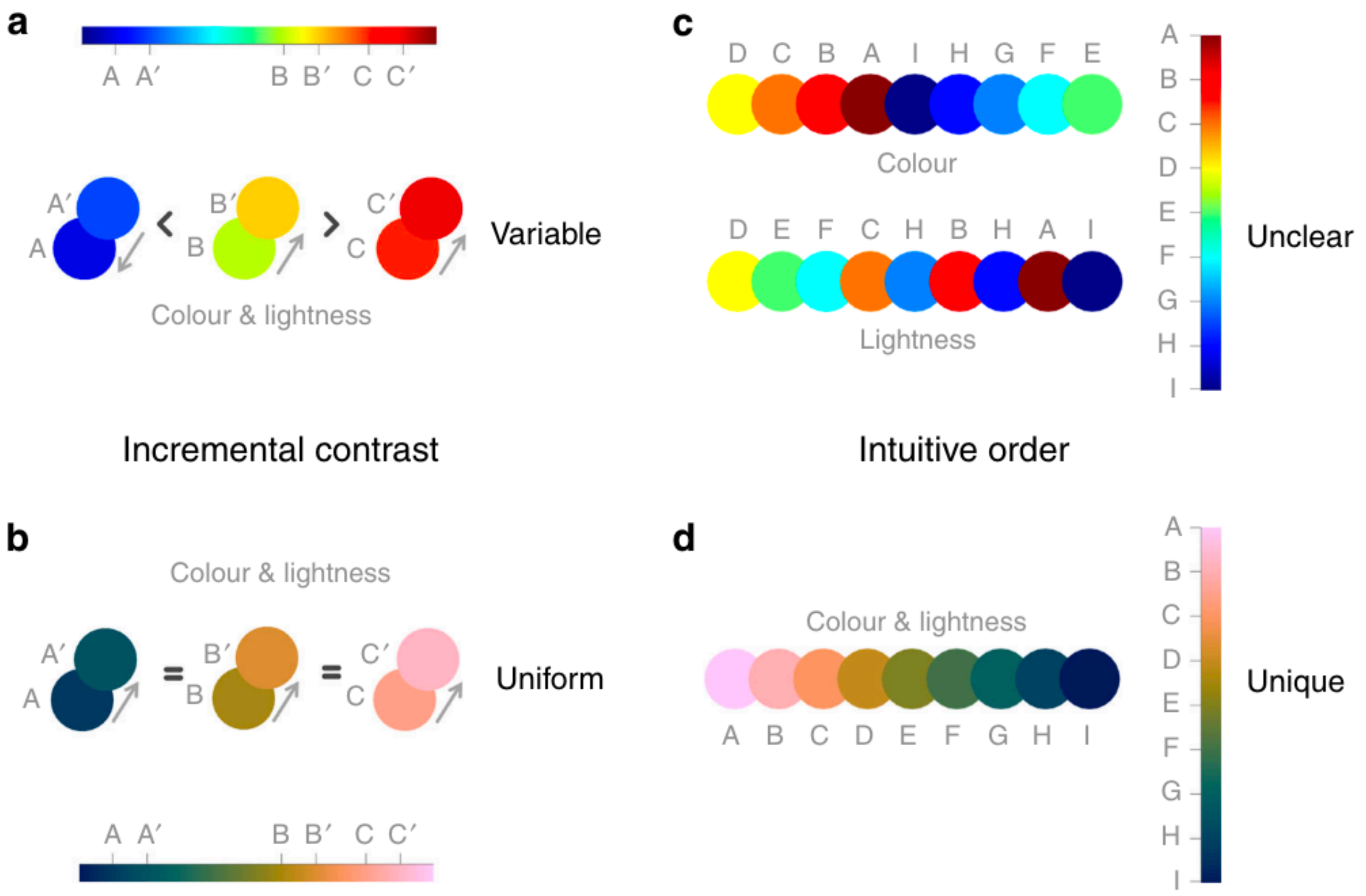


Color scales and color vision deficiencies



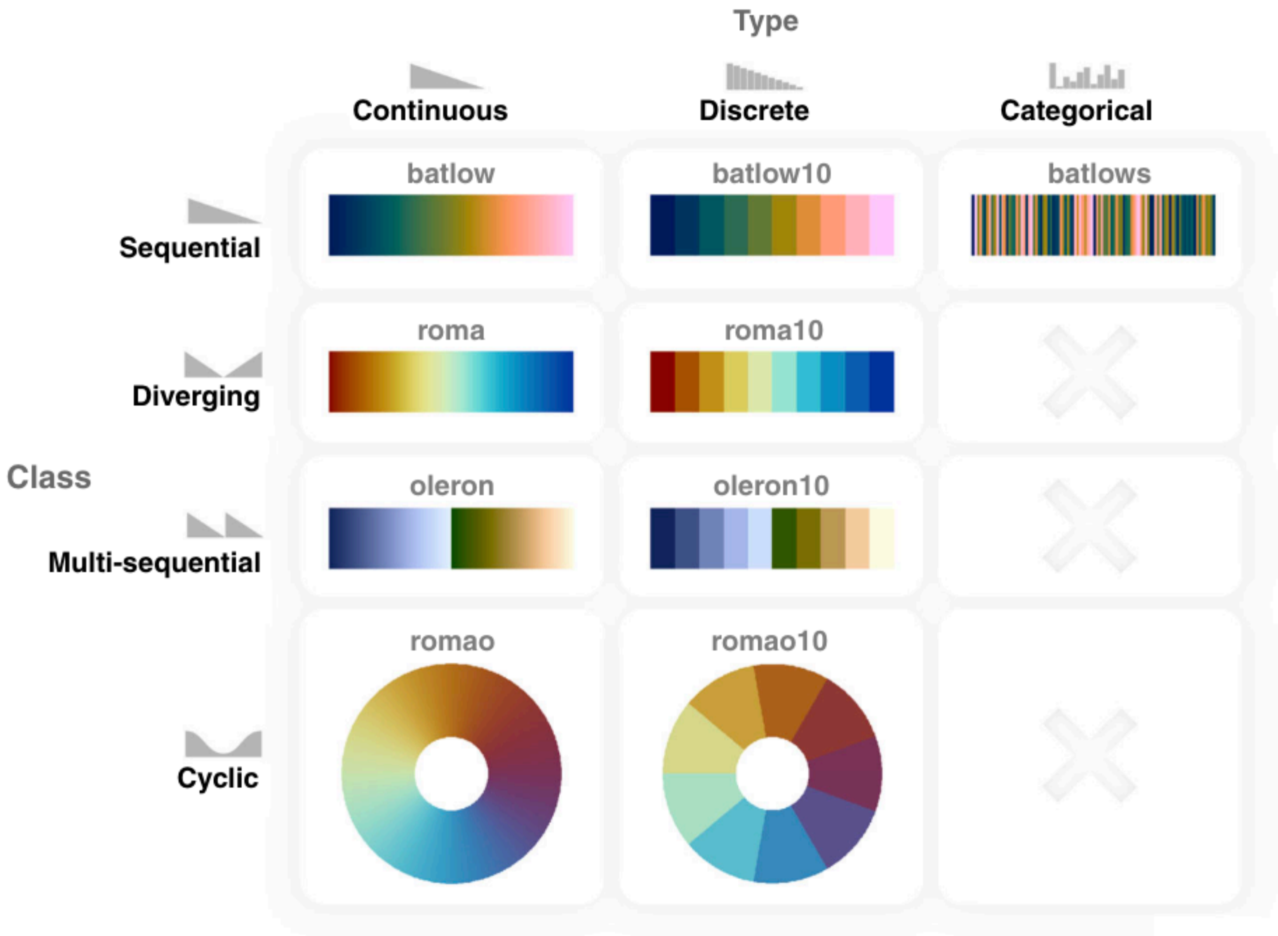
Crameri et al., 2021 *Nature Communications*

Color scales and lightness gradients



Crameri et al., 2021 *Nature Communications*

Color scales for different purposes



Crameri et al., 2021 *Nature Communications*

Color scales in R

khroma 1.16.0 Reference Articles Changelog

khroma

CRAN 1.16.0 CRAN OK downloads 1710/month

repo status Active

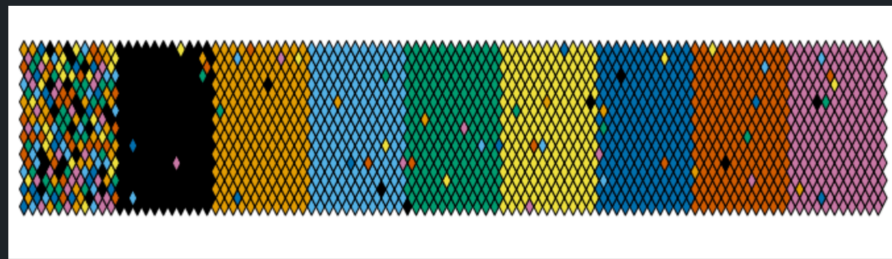
DOI 10.5281/zenodo.1472077

Diagnostic tools

Test how well the colors are identifiable

```
## Okabe & Ito's color scheme
okabe <- color("okabe ito")

set.seed(12345)
plot_map(okabe(8))
```



Simulate color-blindness

```
plot_scheme_colorblind(okabe(8))
```

https://packages.tesselle.org/khroma/