CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD
COLLEGE

Materials by Sara Mathieson

Admin

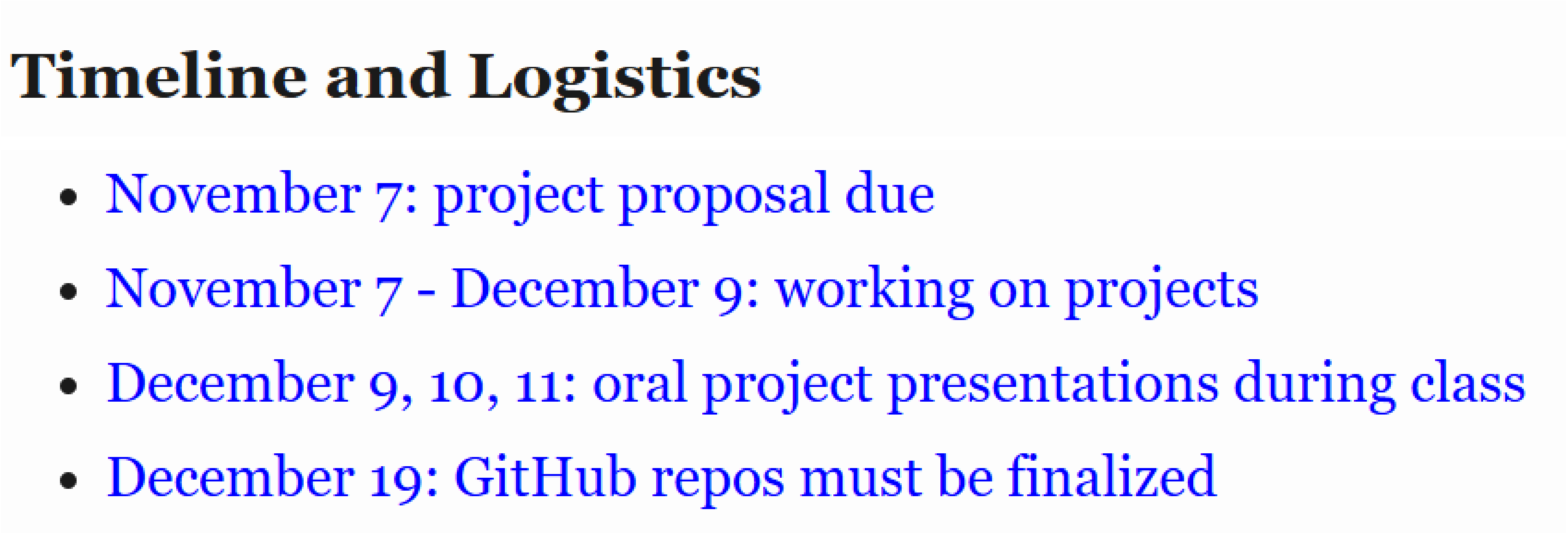
- **Lab 6** due tonight at midnight
- **Lab 7** posted (due next Tuesday Nov 4)
  - pair-programming optional
- **Final Project proposal** posted (due Nov 7)

Outline for today

- Discuss final project
- Review and practice logistic regression
- Introduction to visualization

Outline for today

- Discuss final project
- Review and practice logistic regression
- Introduction to visualization



Timeline and Logistics

- November 7: project proposal due
- November 7 - December 9: working on projects
- December 9, 10, 11: oral project presentations during class
- December 19: GitHub repos must be finalized

Outline for a typical project:

- Find a dataset (see project proposal writeup)
- Run an algorithm we’ve discussed on the dataset
- Try to do a comparison
  - run the algorithm in multiple ways
  - different data pre-processing
  - try a different algorithm
- Evaluate, interpret, and visualize the results

Project Proposal

- **Title** and **names** of both partners
  - Pair work is required!
- A **dataset** (what is n? what is p?)
- An **algorithm** or set of algorithms you will develop and/or apply to this dataset
- A **scientific question** you are trying to answer
  - "Will Naive Bayes or logistic regression perform better on my dataset?"
  - "How will pre-processing a dataset or subsampling features affect the results?"
- A way to **evaluate, interpret, and *visualize*** the results
- **References**

Project Group Options

- If you would like a random partner, please email me ASAP!

- If you *really* prefer to work individually or in a group of 3, email me ASAP!

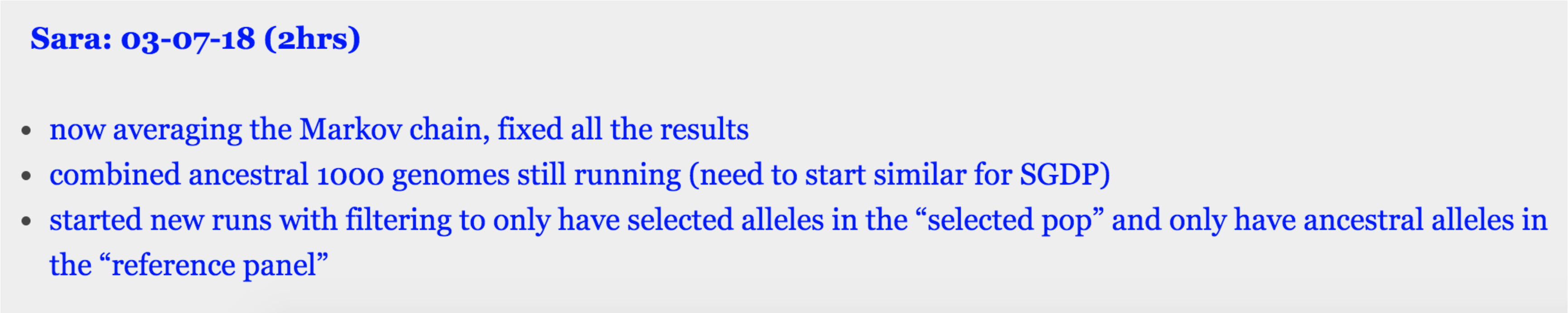
Final Project Deliverables

- Main deliverable: presentation
  - In class Dec 9, 10, 11 (last week of classes)
  - 8 min per group
  - 4 min for questions and peer feedback

- On GitHub (by Dec 19)
  - Project Code
  - Lab Notebook (in README.md)
  - Presentation Slides

Project Lab Notebook

- Running document
- Should say:
  - who was working (which partner)
  - date
  - how long
  - briefly describe what was accomplished



Sara: 03-07-18 (2hrs)

- now averaging the Markov chain, fixed all the results
- combined ancestral 1000 genomes still running (need to start similar for SGDP)
- started new runs with filtering to only have selected alleles in the "selected pop" and only have ancestral alleles in the "reference panel"

Outline for today

- Discuss final project

- Review and practice logistic regression

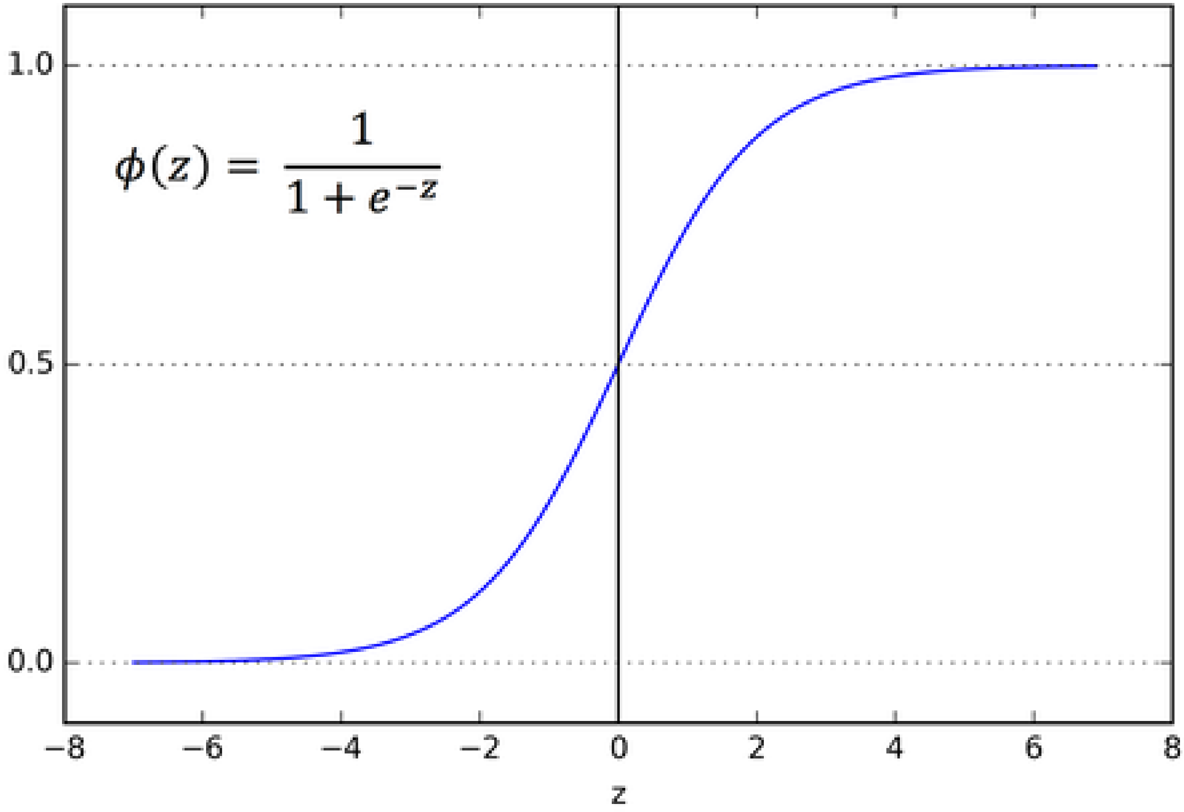
- Introduction to visualization

3 important pieces to SGD

- Hypothesis function (prediction)

$$\boxed{h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w} \cdot \boldsymbol{x}}}}$$

Logistic (sigmoid) function



3 important pieces to SGD

- Hypothesis function (prediction)

$$\boxed{h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}}$$

- Cost function (want to minimize)

$$\boxed{J(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log h_{\boldsymbol{w}}(\boldsymbol{x_i}) + (1 - y_i) \log(1 - h_{\boldsymbol{w}}(\boldsymbol{x_i}))}$$

3 important pieces to SGD

- Hypothesis function (prediction)

$$\boxed{h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}}$$

- Cost function (want to minimize)

$$\boxed{J(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log h_{\boldsymbol{w}}(\boldsymbol{x_i}) + (1 - y_i) \log(1 - h_{\boldsymbol{w}}(\boldsymbol{x_i}))}$$

- Gradient of cost wrt single data point x_i

$$\boxed{\nabla J_{\boldsymbol{x}_i}(\boldsymbol{w}) = (h_{\boldsymbol{w}}(\boldsymbol{x_i}) - y_i)\boldsymbol{x_i}}$$

Stochastic Gradient Descent for Logistic Regression (binary classification)

```
set $\vec{w} = \vec{0}$
while cost $J(\vec{w})$ is still changing:
    shuffle data points
    for i = 1,…,n:
        $\vec{w} \leftarrow \vec{w} - \alpha \underbrace{\nabla J_{\vec{x}_i}(\vec{w})}_{\text{derivative of } J(\vec{w}) \text{ wrt } x_i}$
    store $J(\vec{w})$
```

Cost function as Cross Entropy

probability distribution

$$J(\vec{w}) = -[y \log\left(\overline{h_{\vec{w}}(x)}\right) + (1 - y) \log\left(\overline{1 - h_{\vec{w}}(x)}\right)]$$

probability distribution

$$H(Y) = - \sum_{y \in vals(Y)} p(y)\, log\, p(y)$$

entropy

Cross entropy

$$H(p, q) = - \sum_{y \in vals(Y)} p(y)\, log\, q(y)$$

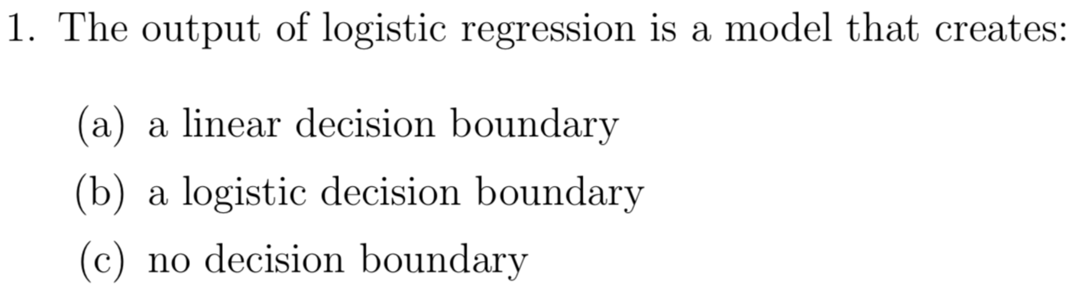
Cost function as Cross Entropy

- Example
  - – true: y=0, 1-y=1
  - – pred: h=0.4, 1-h=0.6

$$H(true, pred) = -(0\log(0.4) + 1\log(0.6)) = 0.5$$

Handout 14

Handout 14



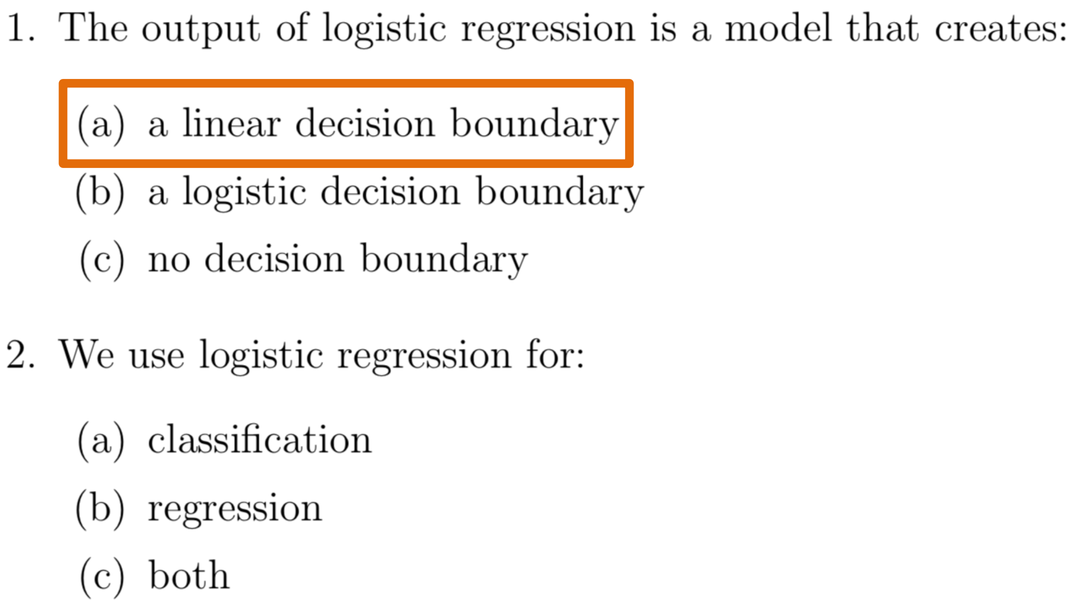
1. The output of logistic regression is a model that creates:

   (a) a linear decision boundary

   (b) a logistic decision boundary

   (c) no decision boundary

Handout 14



1. The output of logistic regression is a model that creates:

   [✓ (a) a linear decision boundary]

   (b) a logistic decision boundary

   (c) no decision boundary

2. We use logistic regression for:

   (a) classification

   (b) regression

   (c) both

Handout 14

1. The output of logistic regression is a model that creates:

   [✓ (a) a linear decision boundary]
   (b) a logistic decision boundary
   (c) no decision boundary

2. We use logistic regression for:

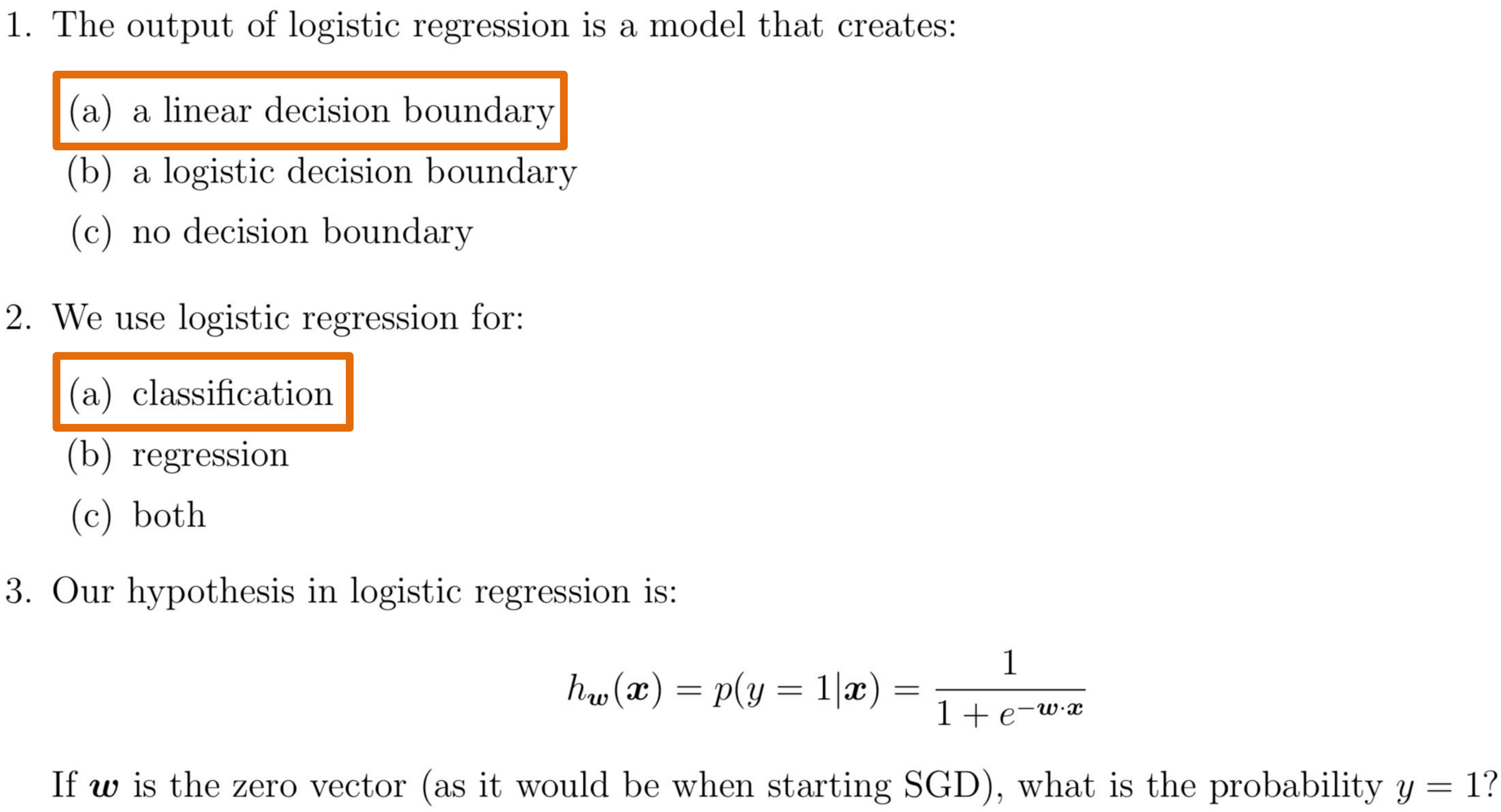
   [✓ (a) classification]
   (b) regression
   (c) both

3. Our hypothesis in logistic regression is:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

   If $\boldsymbol{w}$ is the zero vector (as it would be when starting SGD), what is the probability $y = 1$?

Handout 14



1. The output of logistic regression is a model that creates:

   [✓ (a) a linear decision boundary]

   (b) a logistic decision boundary

   (c) no decision boundary

2. We use logistic regression for:

   [✓ (a) classification]

   (b) regression

   (c) both

3. Our hypothesis in logistic regression is:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

If $\boldsymbol{w}$ is the zero vector (as it would be when starting SGD), what is the probability $y = 1$?

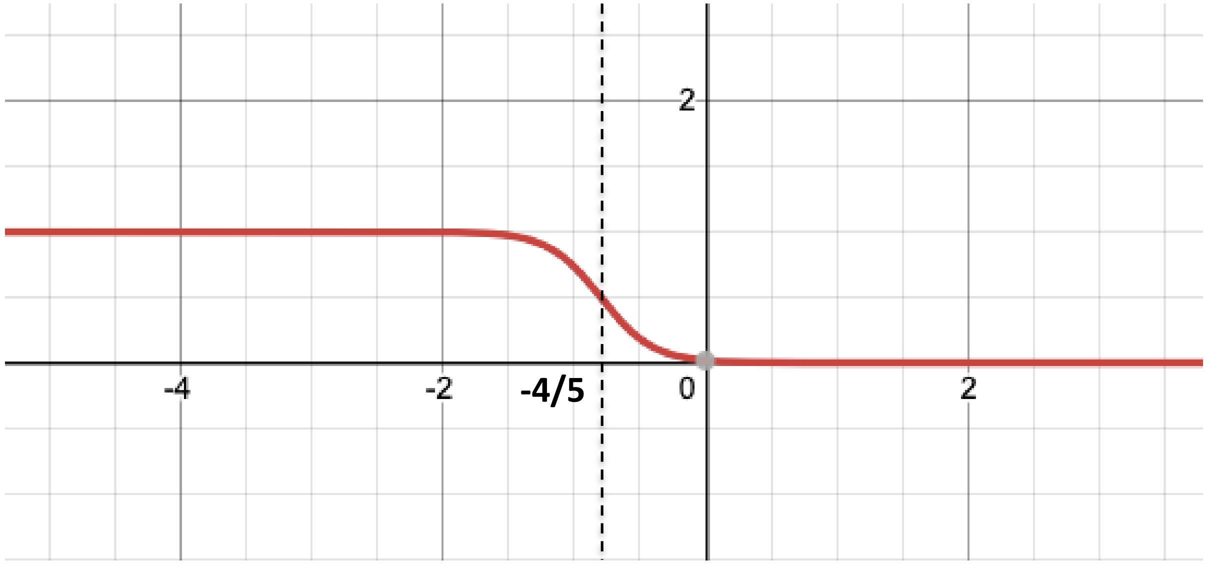
½

Handout 14

Say I train a binary logistic regression model (i.e. outcomes $\in \{0,1\}$) and end up with $\hat{\boldsymbol{w}} = [\hat{w}_0, \hat{w}_1]^T = [-4, -5]^T$. What is the decision boundary? Sketch a graph of this logistic model and label the decision boundary. How would you classify a new point $x_{\text{test}} = -2$?

predict y=1

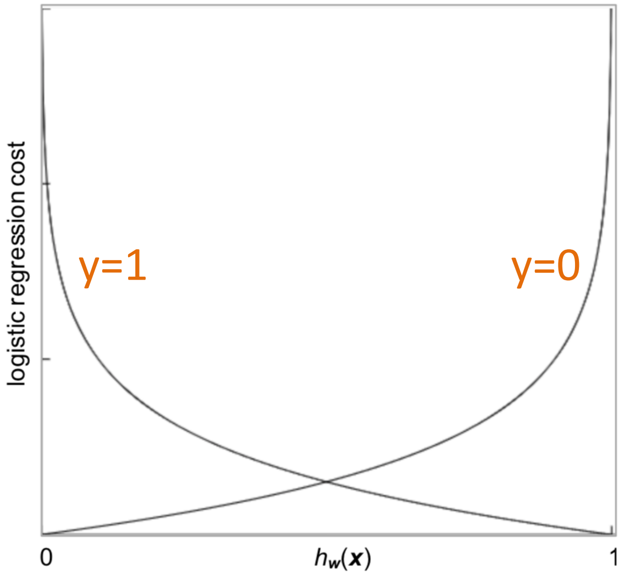
$$\frac{1}{\left(1 + e^{-(-4-5x)}\right)}$$



Decision boundary

Handout 14

5. The graph below shows the cost for logistic regression as a function of the hypothesis $h_{\boldsymbol{w}}(\boldsymbol{x})$, for one example $\boldsymbol{x}$. Which curve corresponds to the true label $y = 0$ and which corresponds to $y = 1$?



Single example $\vec{x}, y$

$$J(\vec{w}) = \begin{cases} -\log\left(h_{\vec{w}}(\vec{x})\right) \; if \; y = 1 \\ -\log\left(1 - h_{\vec{w}}(\vec{x})\right) \; if \; y = 0 \end{cases}$$

Outline for today

- Discuss final project
- Review and practice logistic regression
- Introduction to visualization

Ugly, bad, wrong visualizations

- **ugly**—A figure that has aesthetic problems but otherwise is clear and informative.

- **bad**—A figure that has problems related to perception; it may be unclear, confusing, overly complicated, or deceiving.

- **wrong**—A figure that has problems related to mathematics; it is objectively incorrect.

From "Fundamentals of Data Visualization" by Claus Wilke

Ugly, bad, wrong visualizations

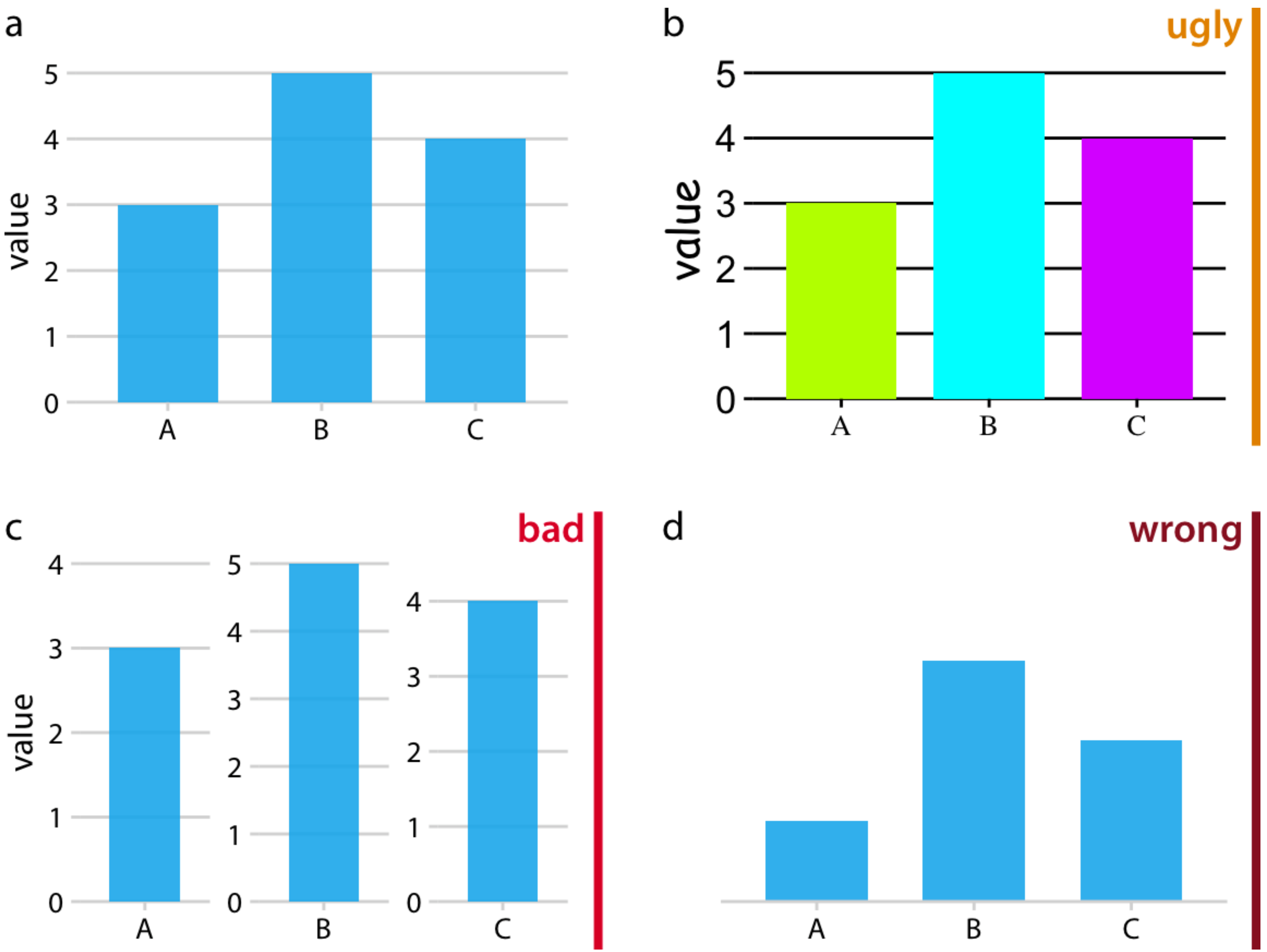


Fig 1.1 from “Fundamentals of Data Visualization” by Claus Wilke

Data Types: continuous vs. discrete

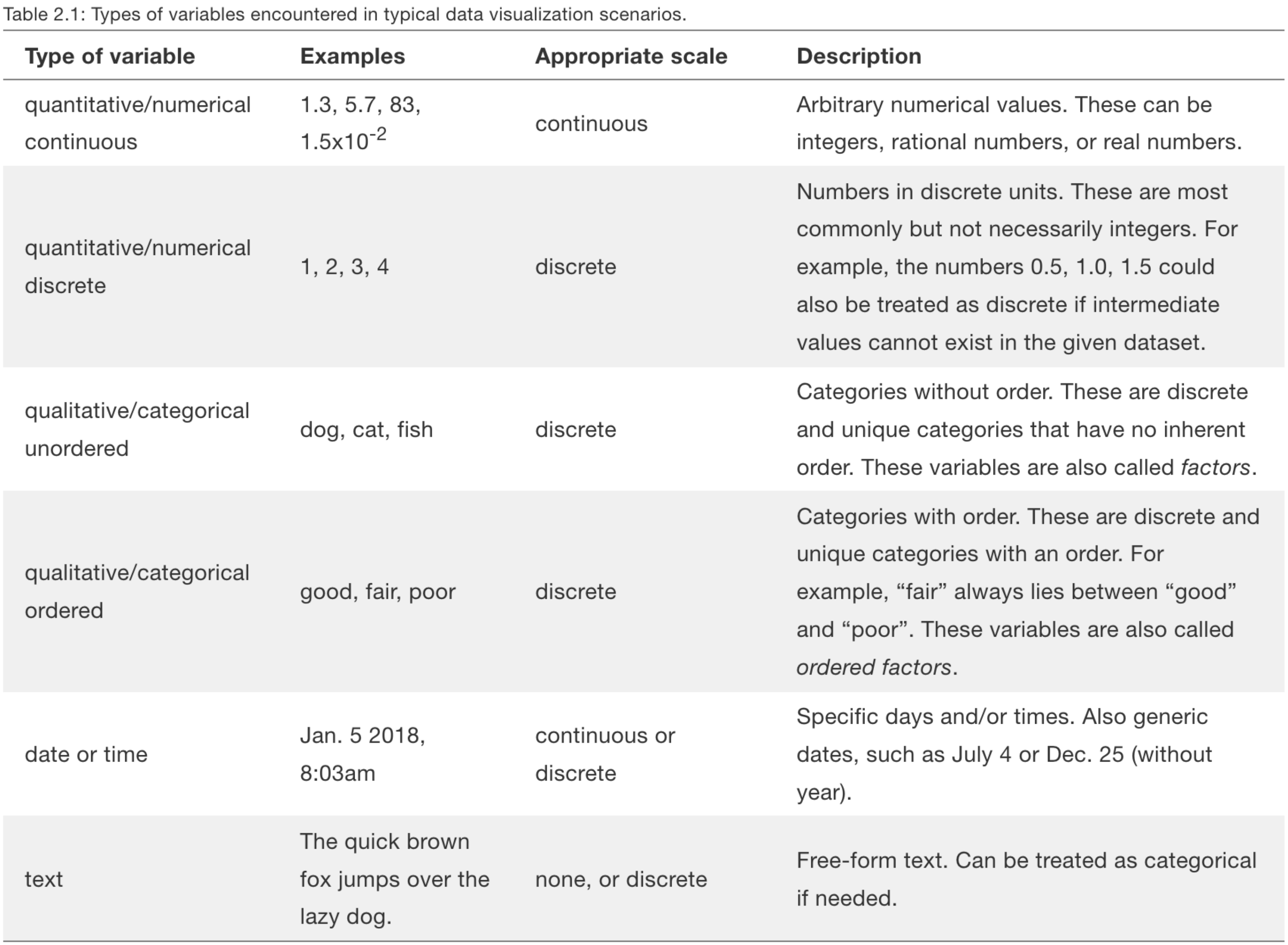


Table 2.1: Types of variables encountered in typical data visualization scenarios.

Type of variable	Examples	Appropriate scale	Description
quantitative/numerical continuous	1.3, 5.7, 83, 1.5x10^{-2}	continuous	Arbitrary numerical values. These can be integers, rational numbers, or real numbers.
quantitative/numerical discrete	1, 2, 3, 4	discrete	Numbers in discrete units. These are most commonly but not necessarily integers. For example, the numbers 0.5, 1.0, 1.5 could also be treated as discrete if intermediate values cannot exist in the given dataset.
qualitative/categorical unordered	dog, cat, fish	discrete	Categories without order. These are discrete and unique categories that have no inherent order. These variables are also called *factors*.
qualitative/categorical ordered	good, fair, poor	discrete	Categories with order. These are discrete and unique categories with an order. For example, "fair" always lies between "good" and "poor". These variables are also called *ordered factors*.
date or time	Jan. 5 2018, 8:03am	continuous or discrete	Specific days and/or times. Also generic dates, such as July 4 or Dec. 25 (without year).
text	The quick brown fox jumps over the lazy dog.	none, or discrete	Free-form text. Can be treated as categorical if needed.

Table 2.1 from "Fundamentals of Data Visualization" by Claus Wilke

Aesthetics in Data Visualization

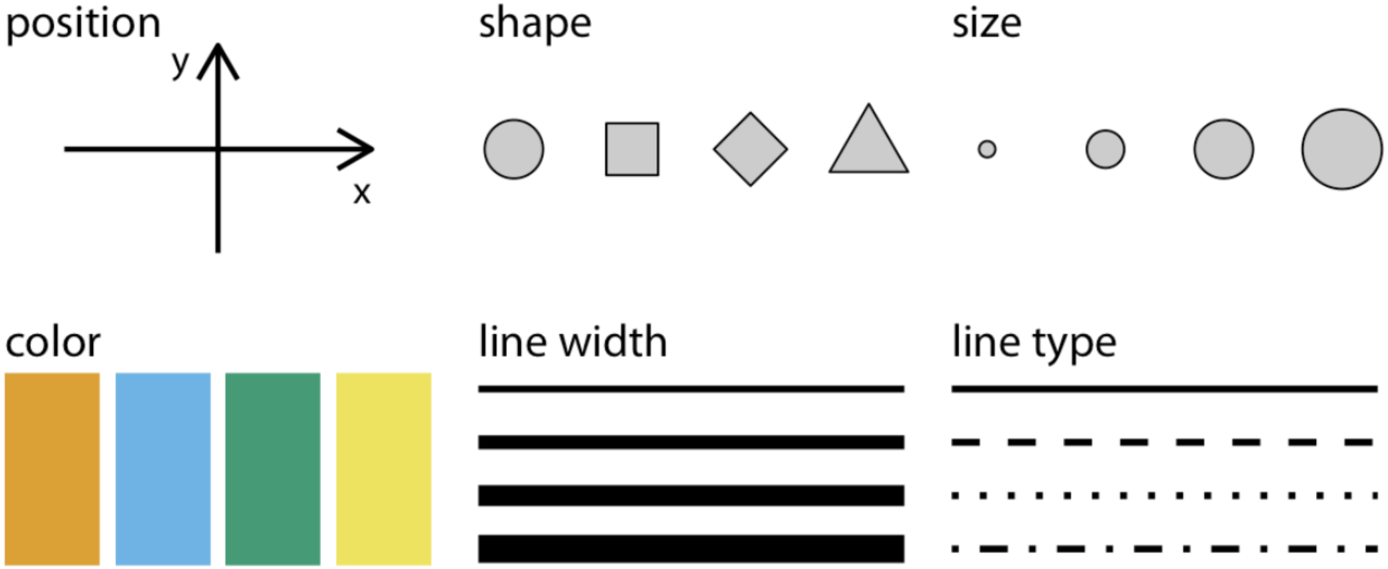
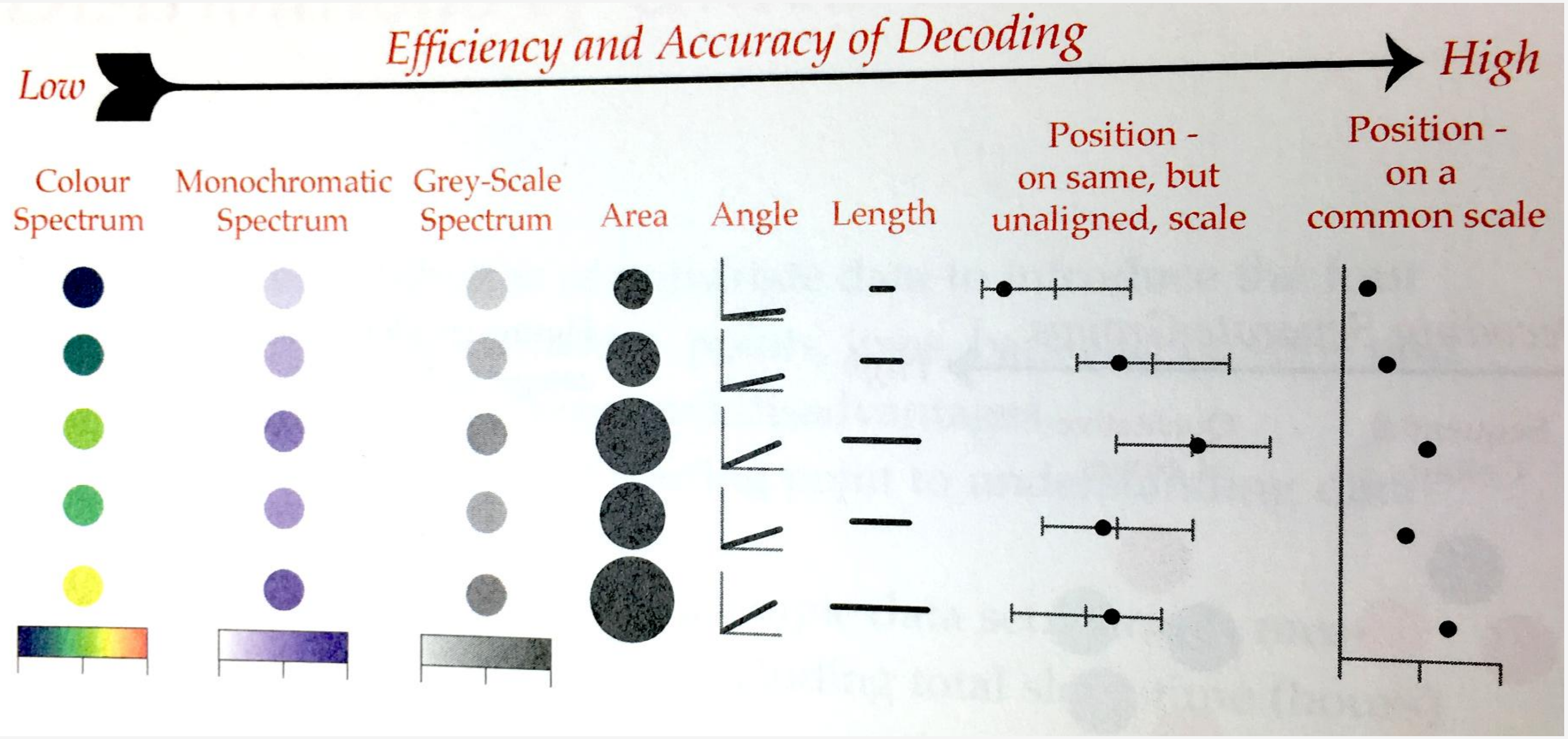


Figure 2.1: Commonly used aesthetics in data visualization: position, shape, size, color, line width, line type. Some of these aesthetics can represent both continuous and discrete data (position, size, line width, color) while others can usually only represent discrete data (shape, line type).

Fig 2.1 from “Fundamentals of Data Visualization” by Claus Wilke

Data Types

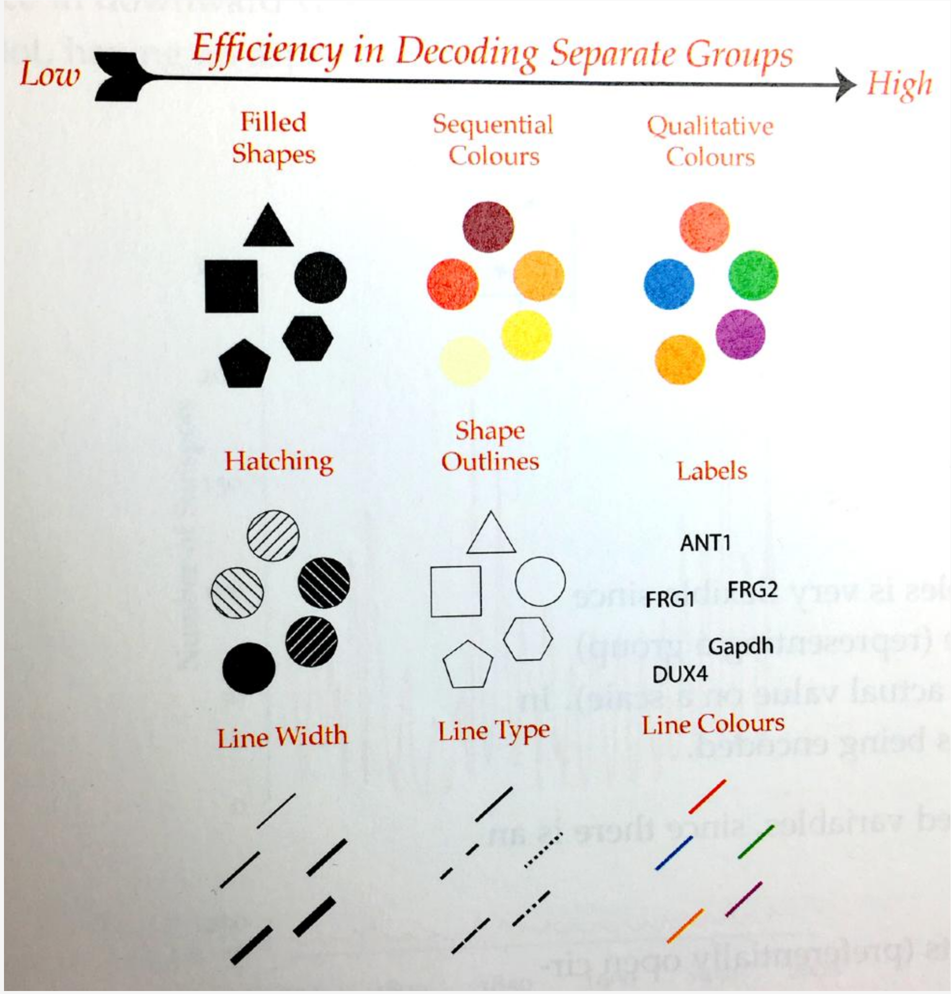
Continuous


Efficiency and Accuracy of Decoding
Low
High
Colour Spectrum
Monochromatic Spectrum
Grey-Scale Spectrum
Area
Angle
Length
Position - on same, but unaligned, scale
Position - on a common scale

Slide from: Eric Miller

Data Types

Discrete



Slide from: Eric Miller

Two different visualizations of the same data

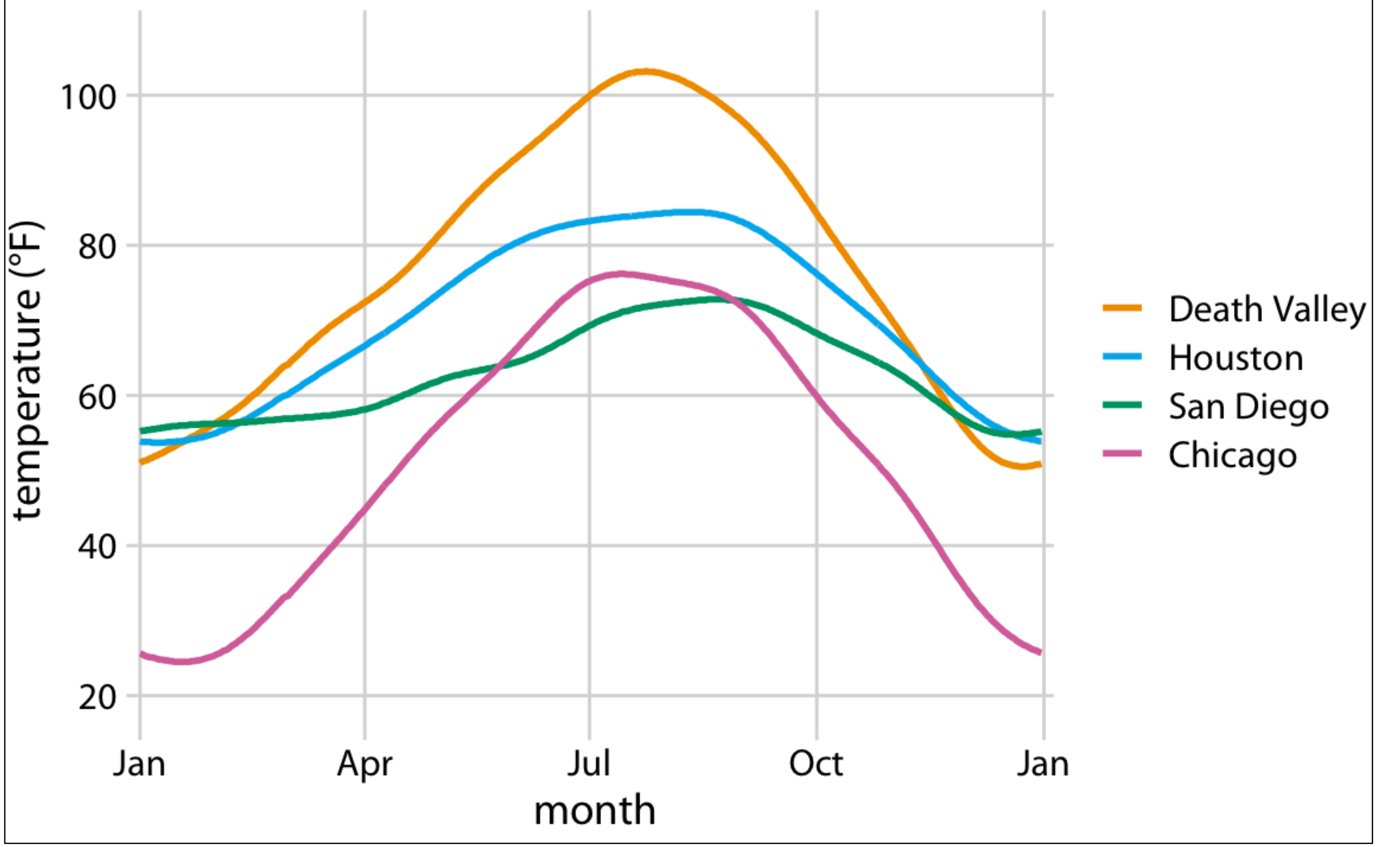
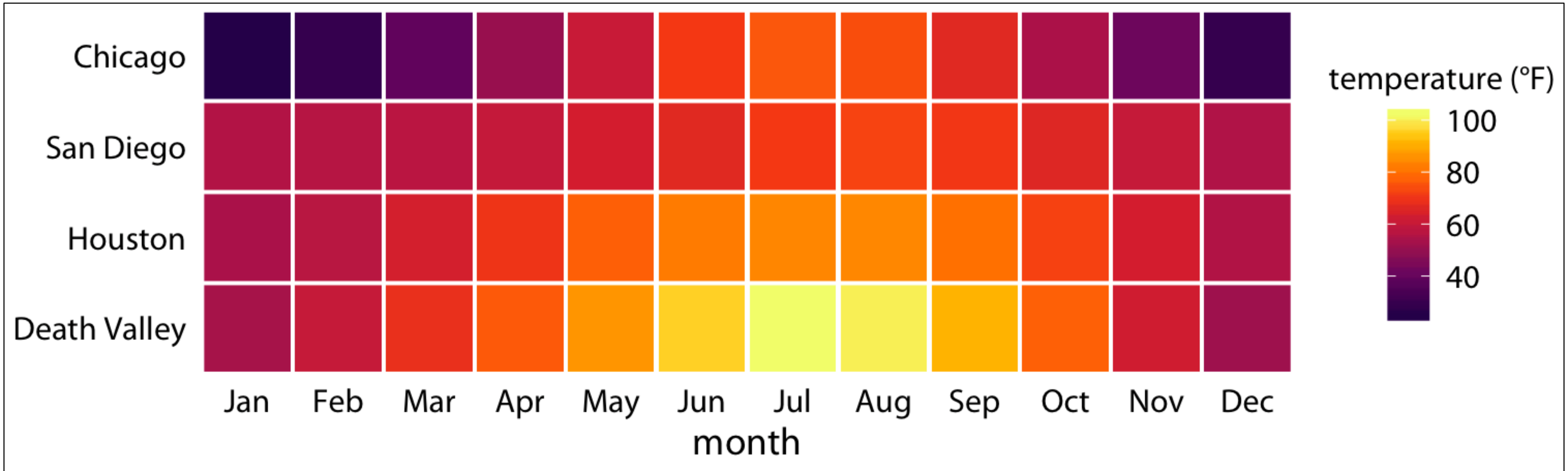


Fig 2.3/2.4 from "Fundamentals of Data Visualization" by Claus Wilke



All the same data – what do we want to convey?

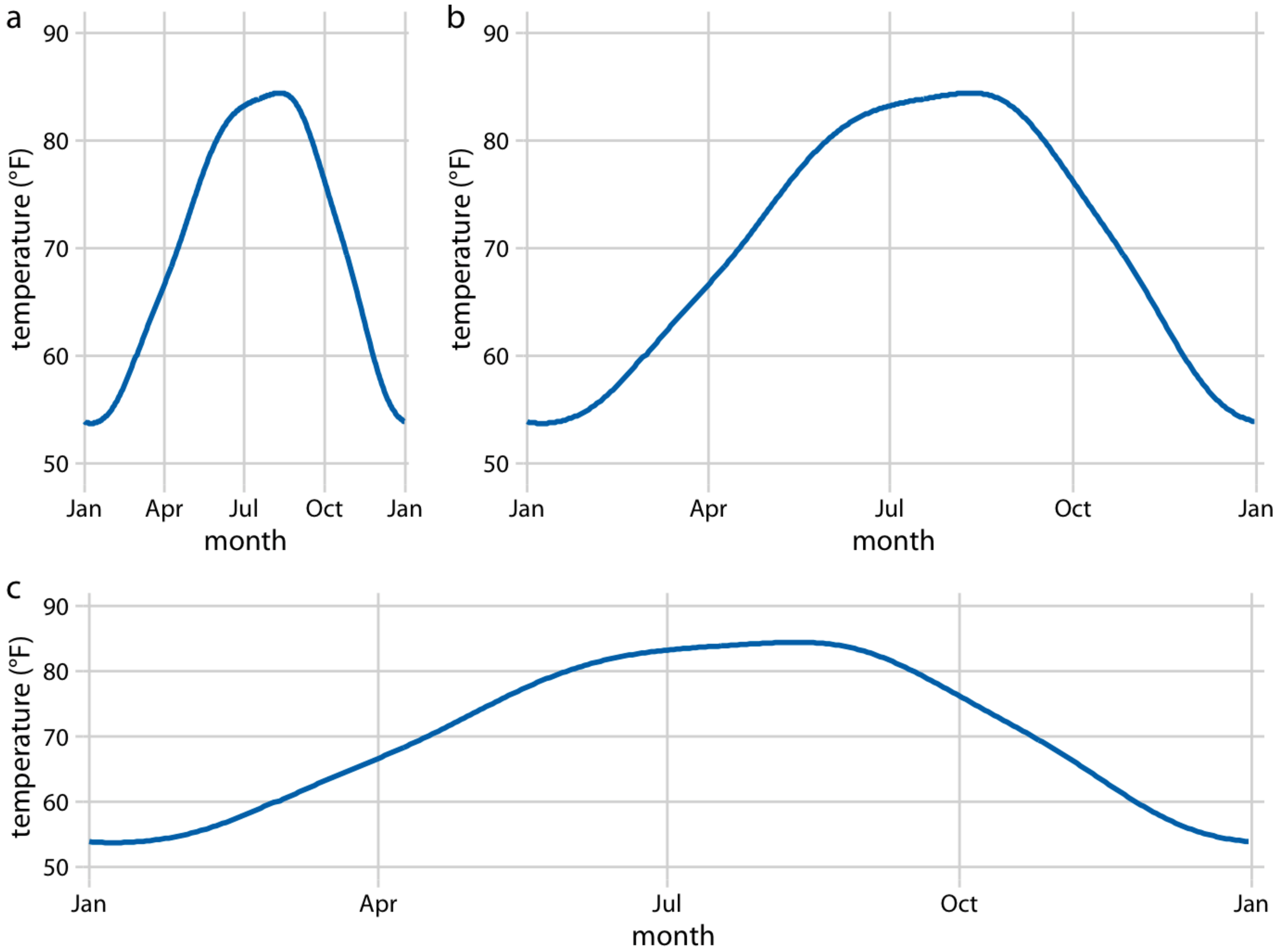


Fig 3.2 from "Fundamentals of Data Visualization" by Claus Wilke

Color

Why use color?

Which colors?

Slide from: Eric Miller

Color: qualitative

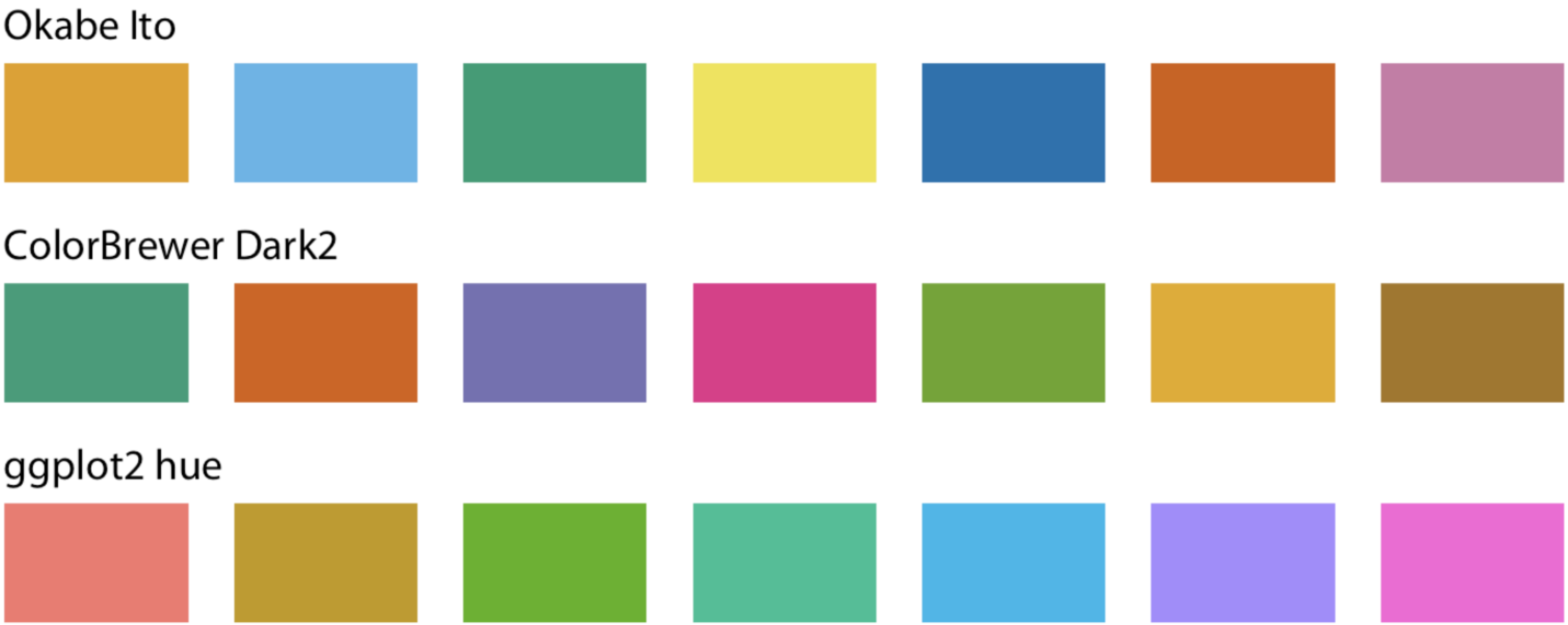


Fig 4.1 from “Fundamentals of Data Visualization” by Claus Wilke

Qualitative example

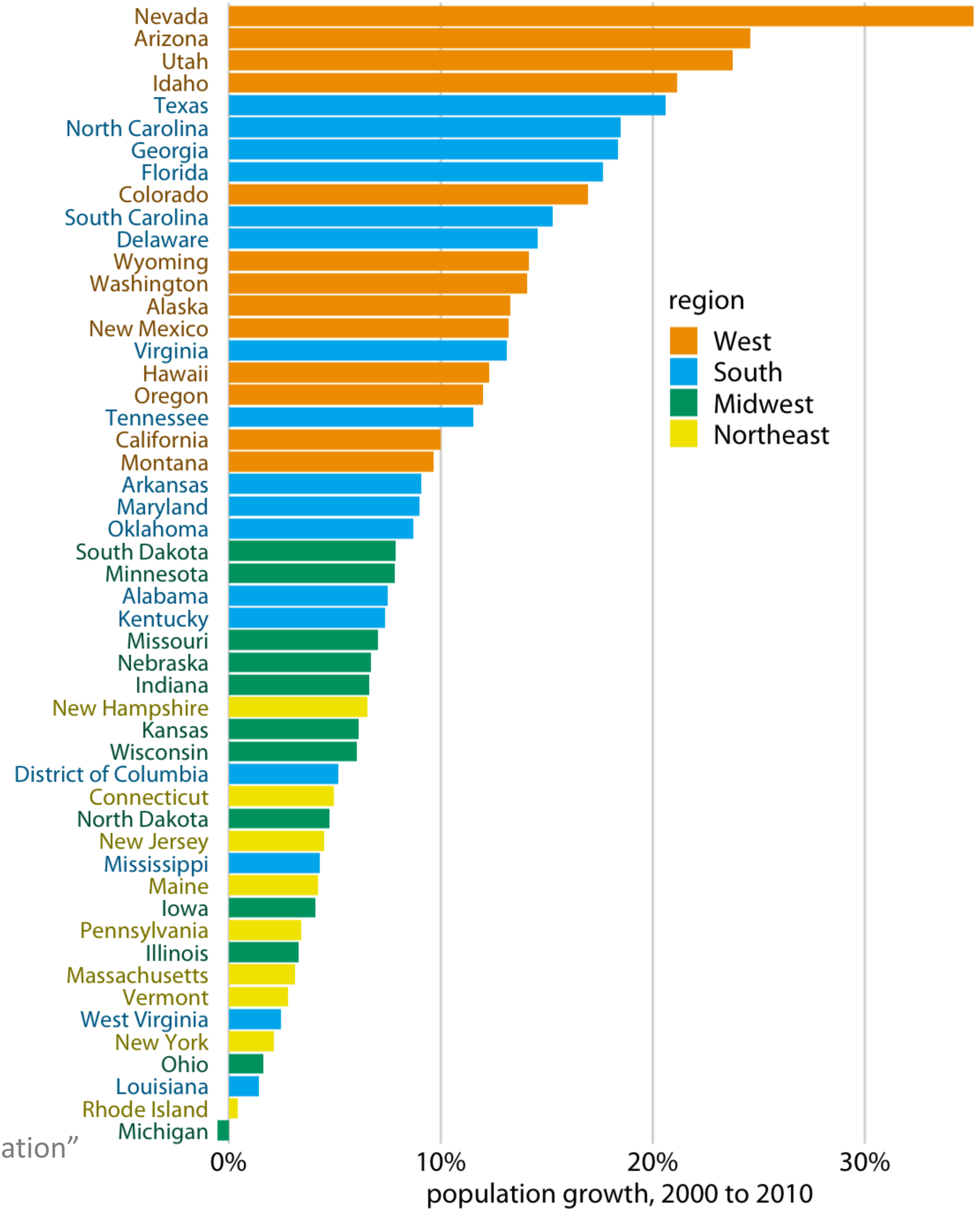


Fig 4.2 from “Fundamentals of Data Visualization”
by Claus Wilke

Color: sequential

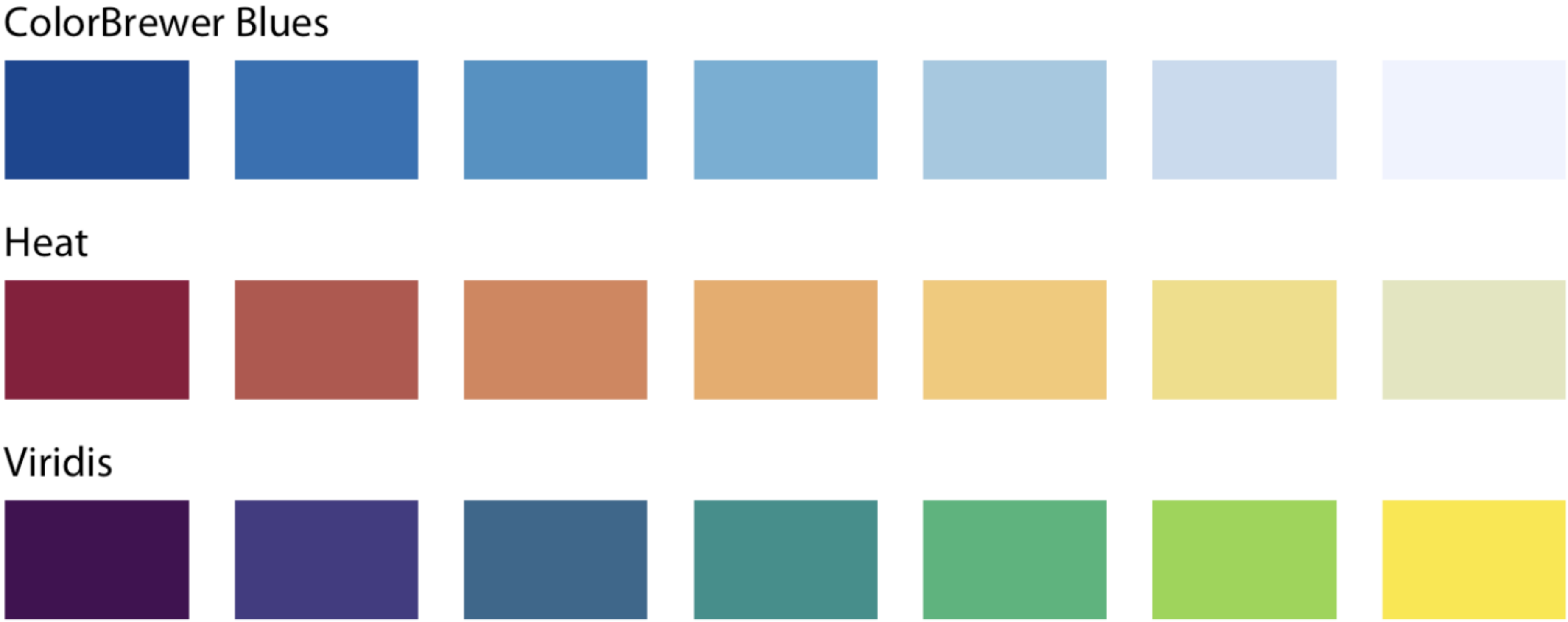


Fig 4.3 from “Fundamentals of Data Visualization” by Claus Wilke

Sequential example

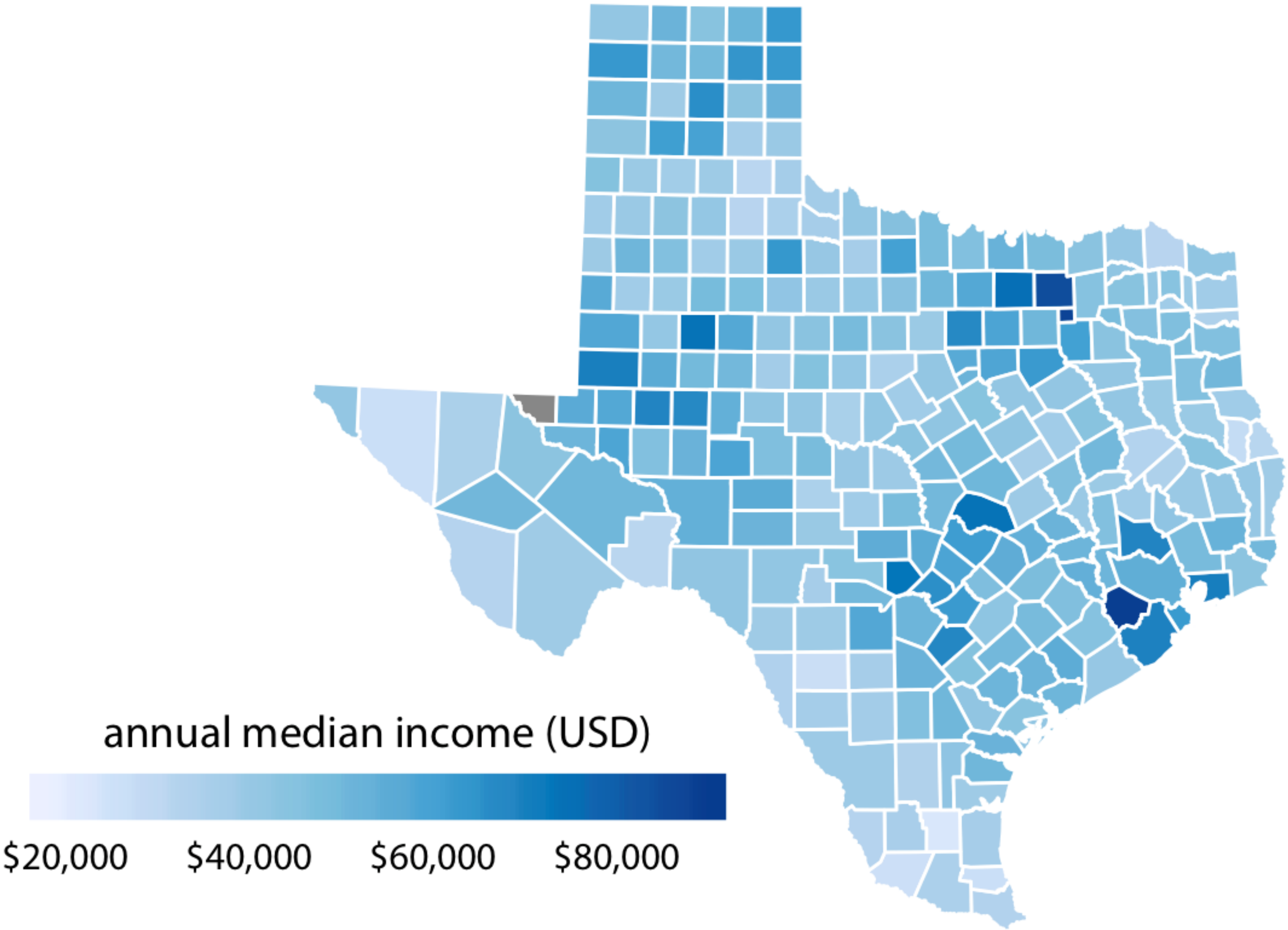


Fig 4.4 from “Fundamentals of Data Visualization” by Claus Wilke

Color: diverging

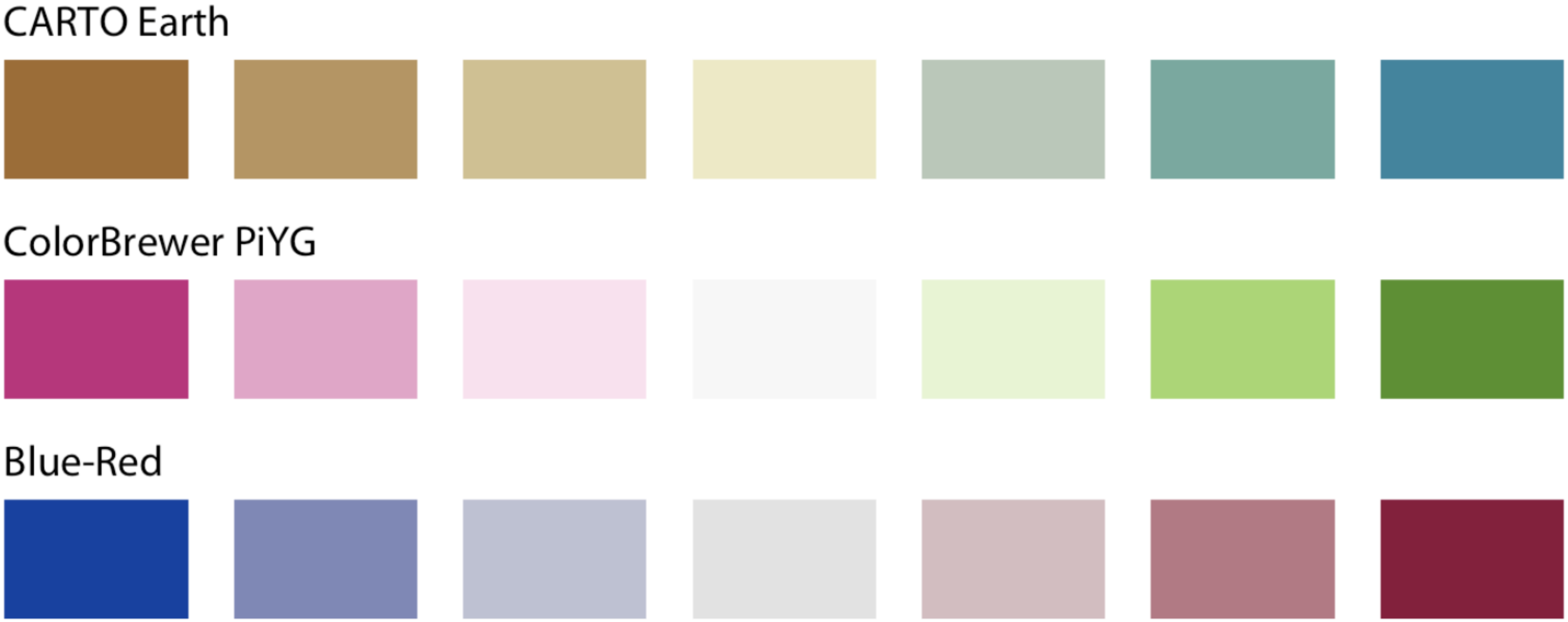


Fig 4.5 from “Fundamentals of Data Visualization” by Claus Wilke

Diverging example

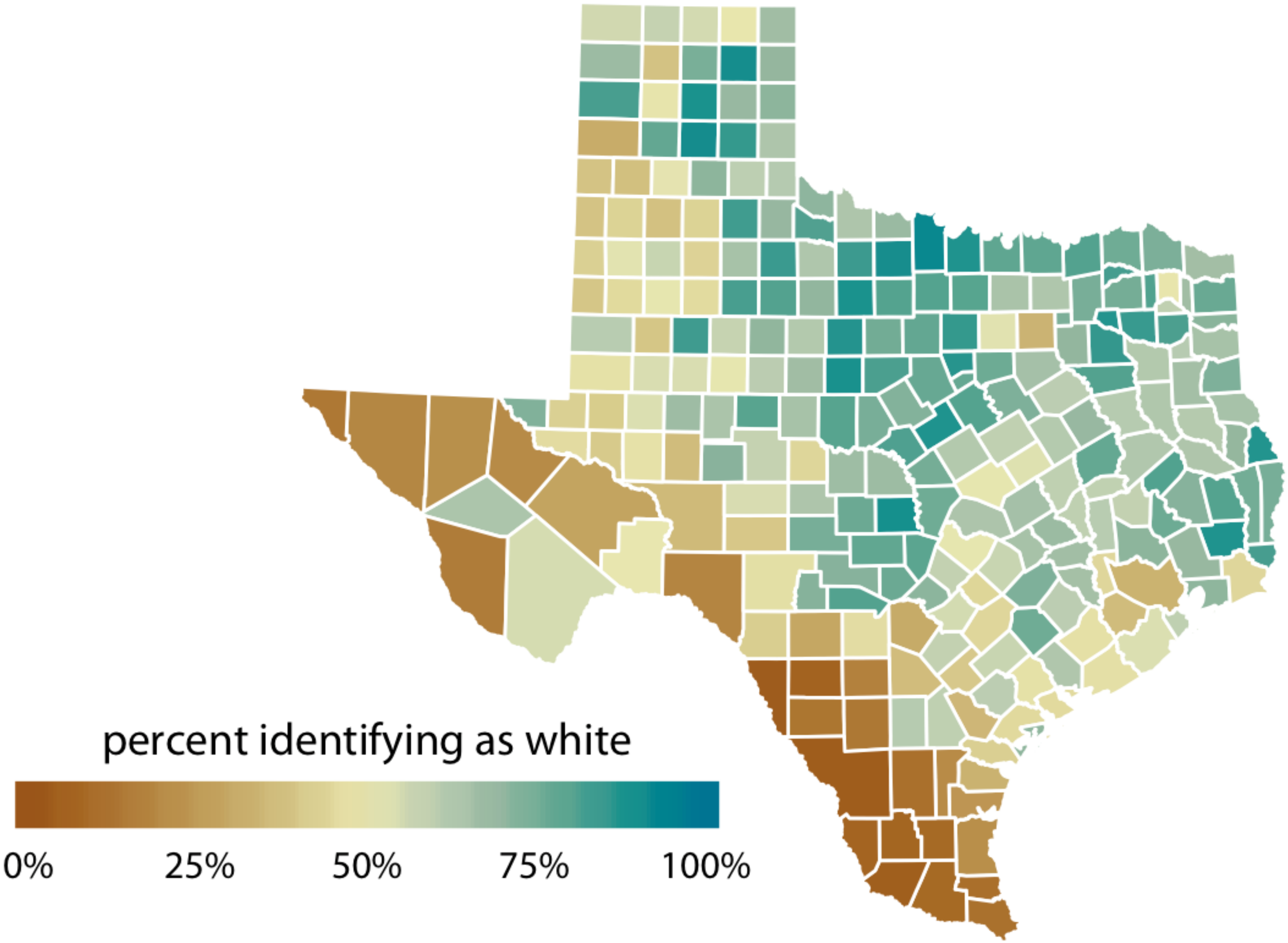
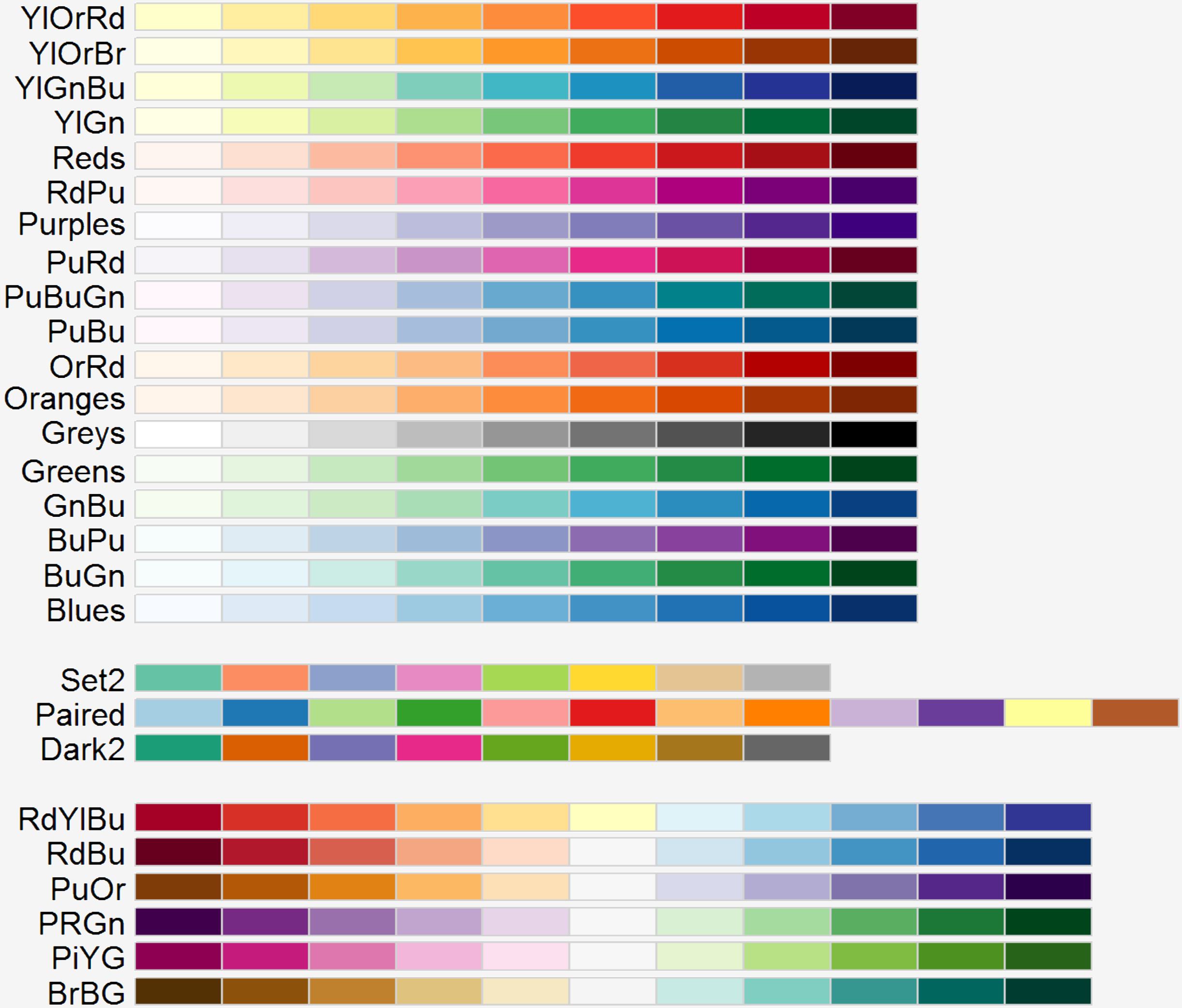


Fig 4.6 from “Fundamentals of Data Visualization” by Claus Wilke

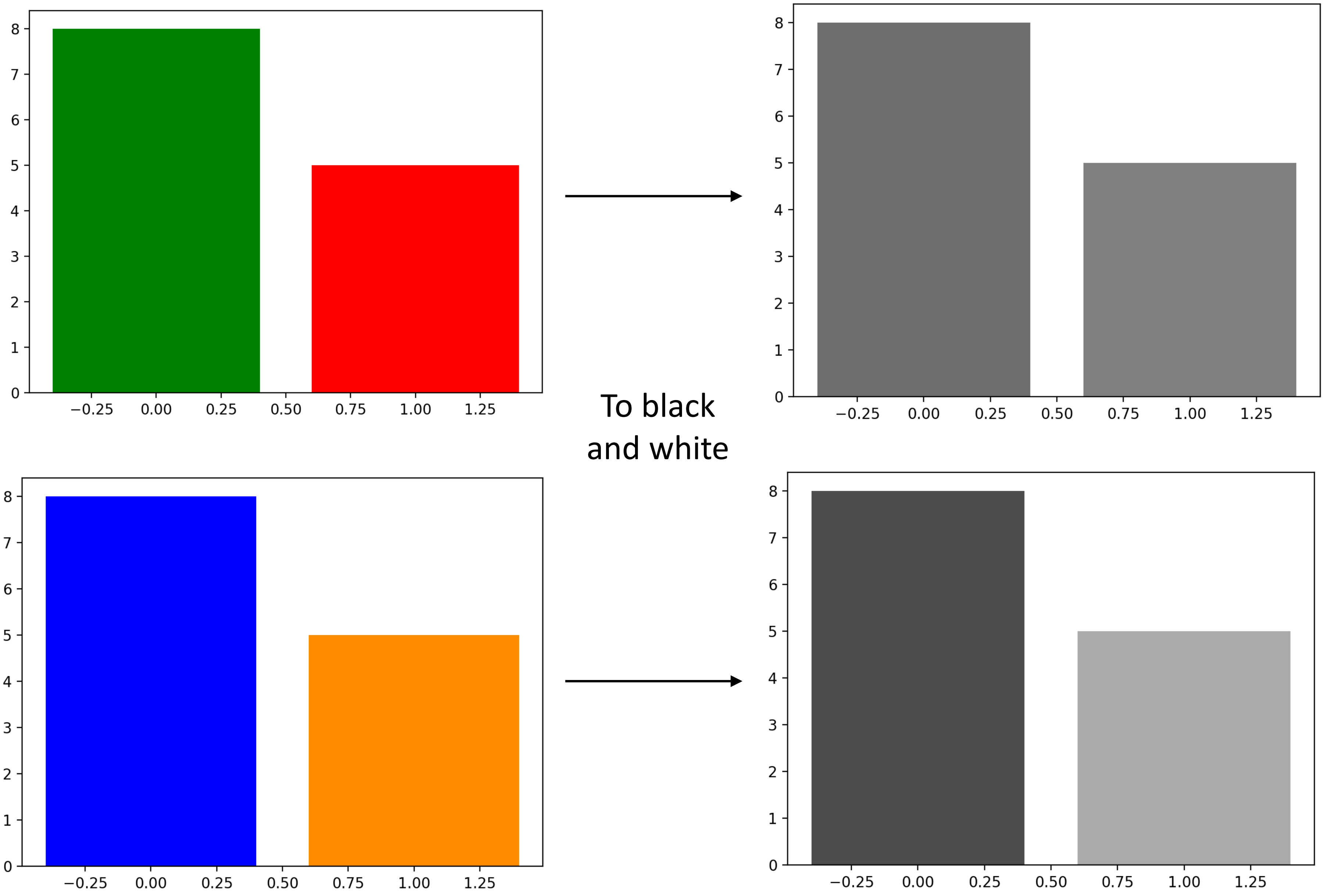
Color

Color-blind friendly



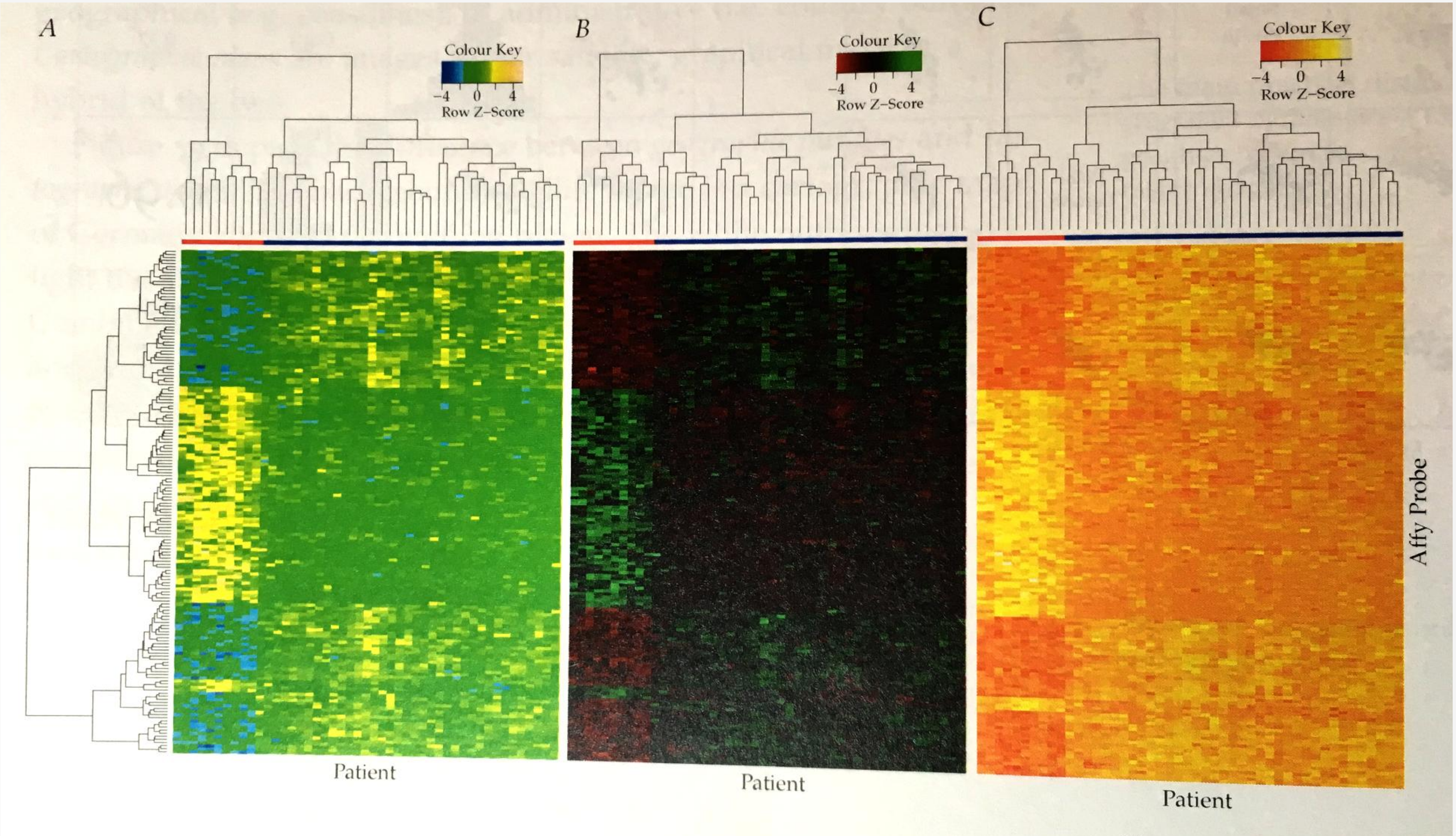
Slide from: Eric Miller

Red/green vs. blue/orange



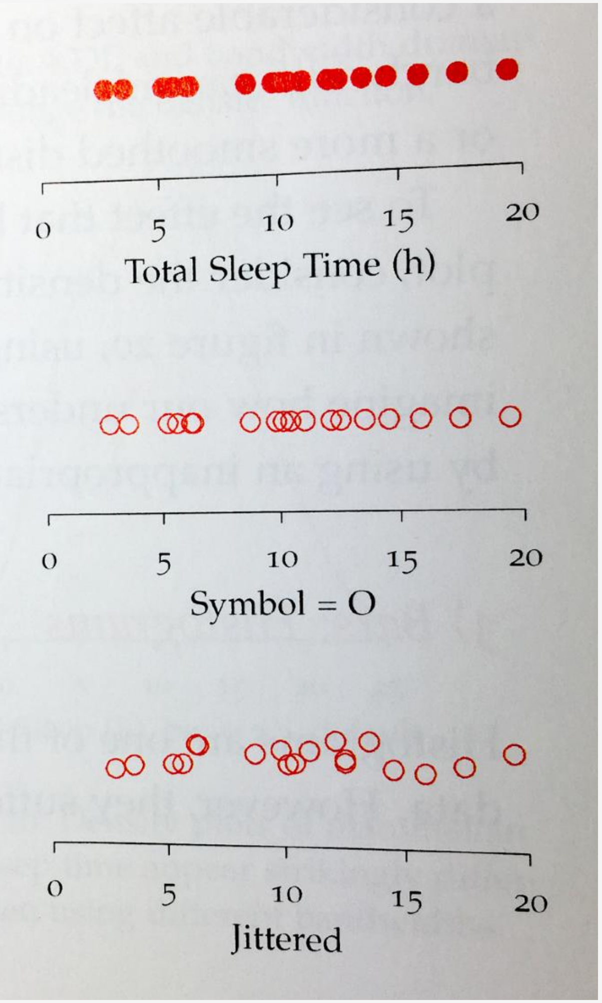
To black and white

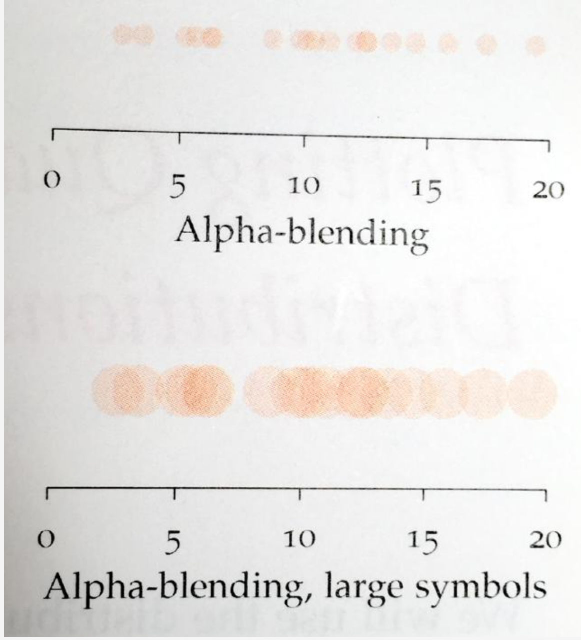
Color



Slide from: Eric Miller

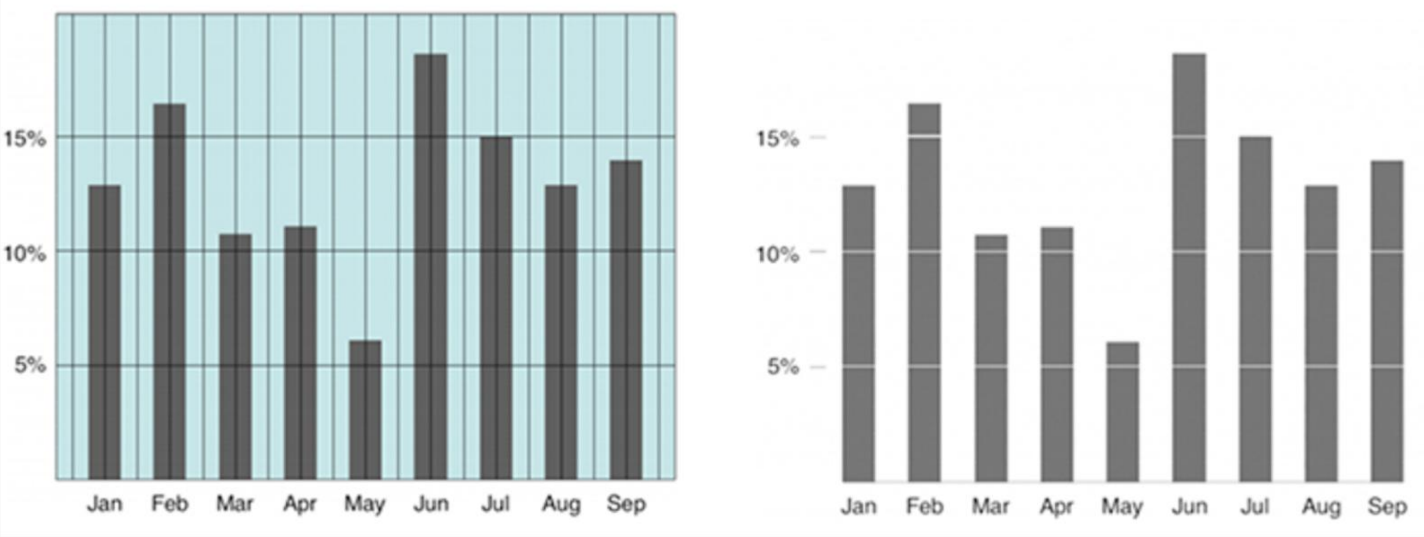
Overplotting





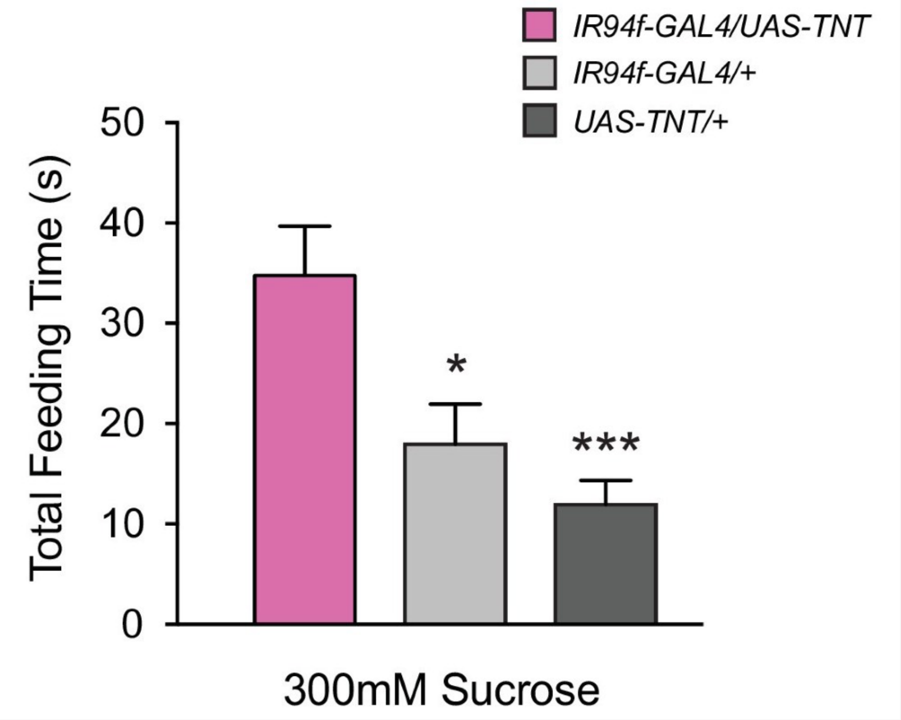
Slide from: Eric Miller

Data::Ink



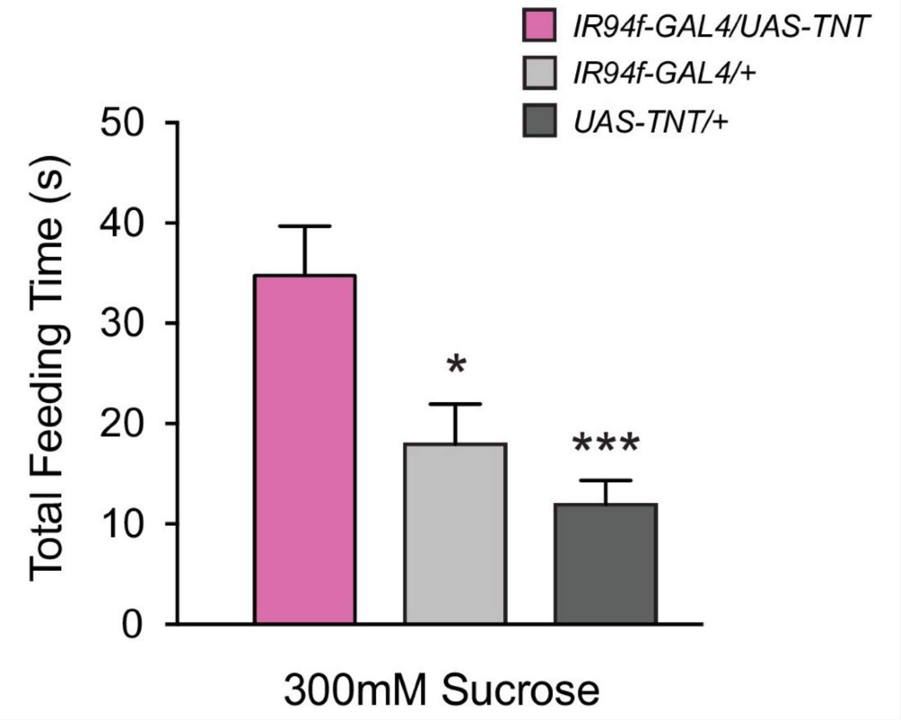
Slide from: Eric Miller

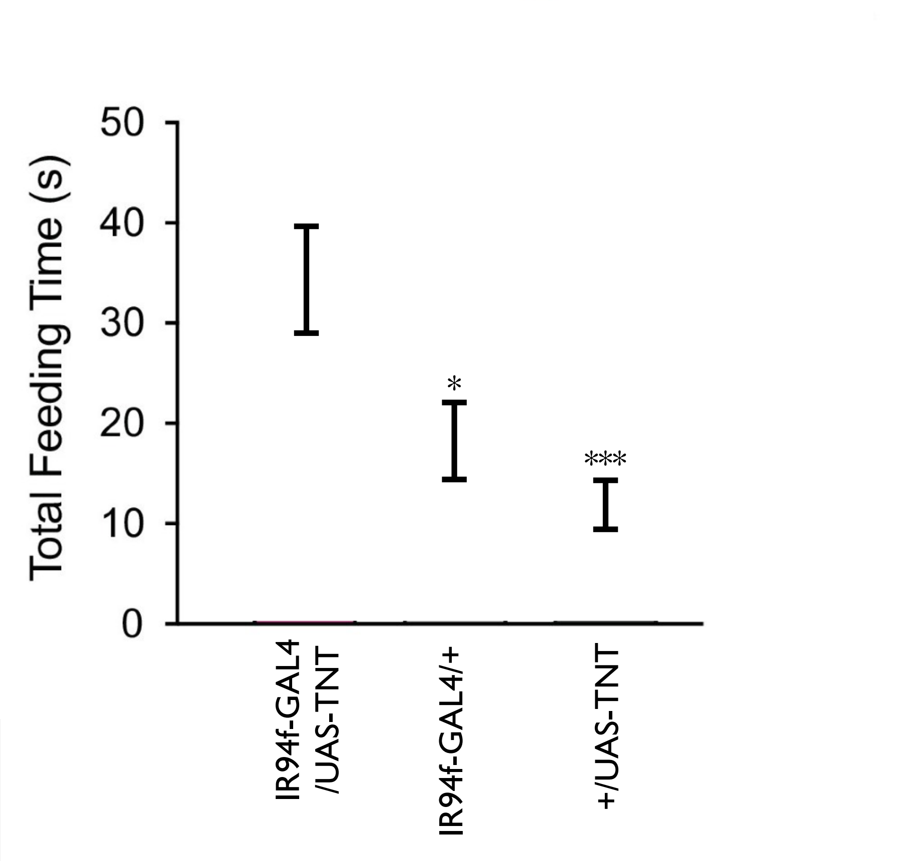
Data::Ink



Slide from: Eric Miller

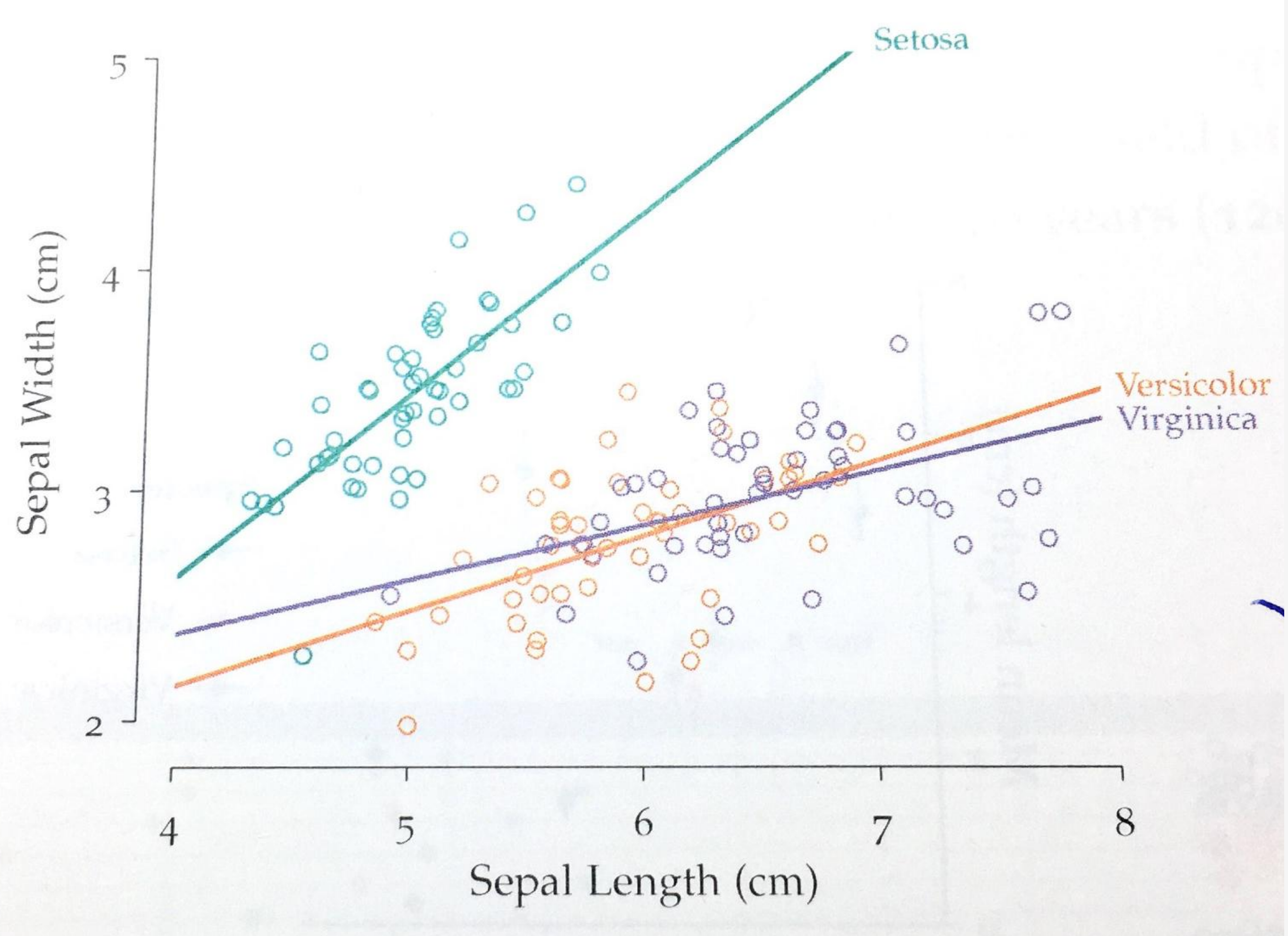
Data::Ink





Slide from: Eric Miller

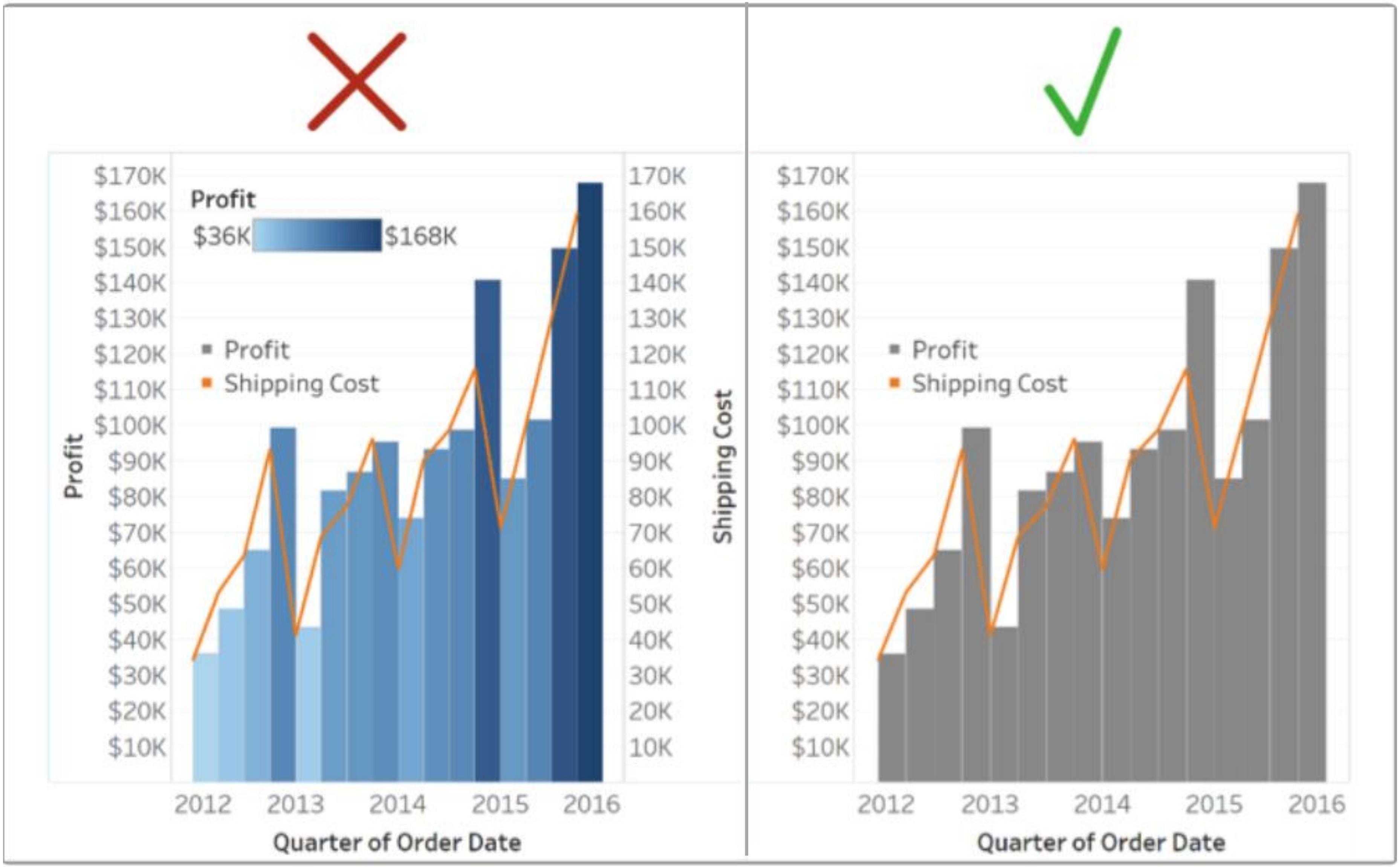
Data::Ink



Where is the legend?

Slide from: Eric Miller

Double encoding



Double encoding

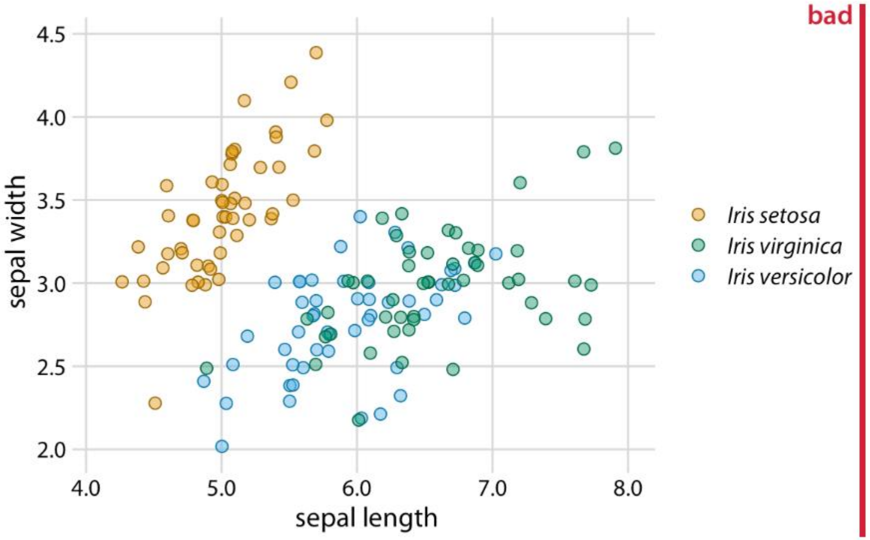
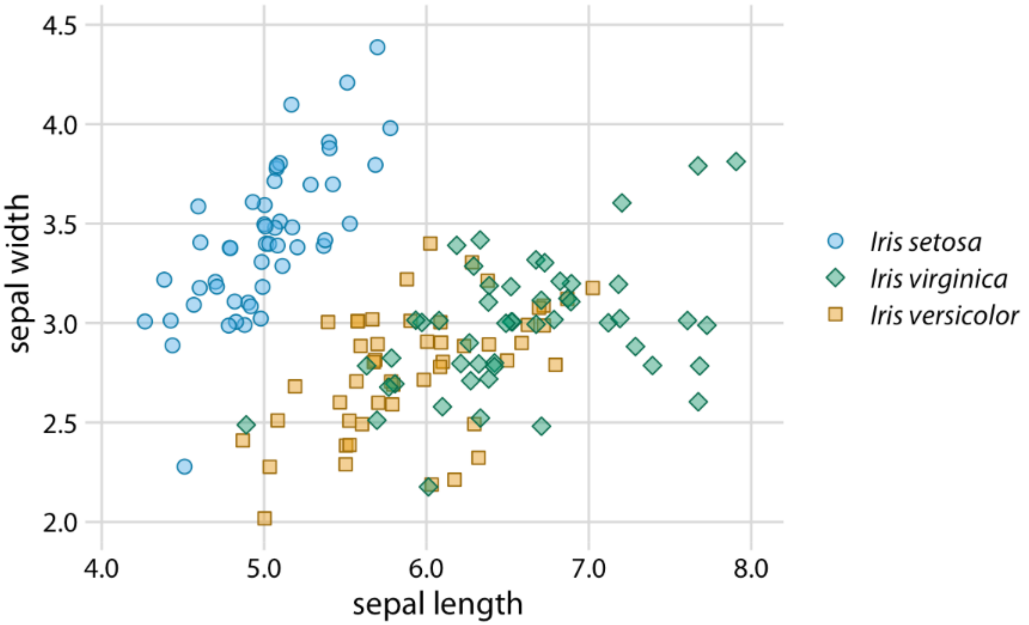
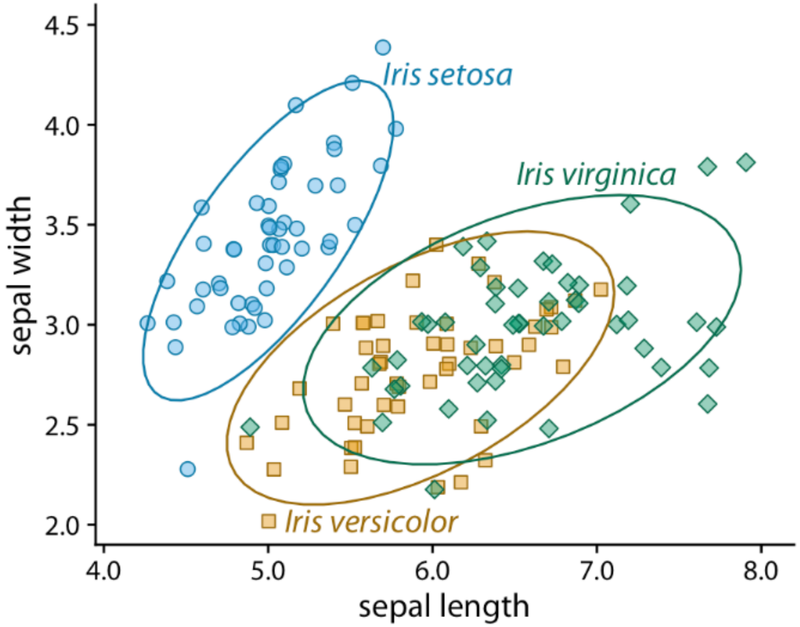


Fig 20.1/20.3/20.9 from “Fundamentals of Data Visualization” by Claus Wilke





Data::Ink

- Remove excess ink
- Show distributions, instead of bars
- Can you remove the legend?
- Remove double encodings when appropriate
- Is a log scale appropriate? https://www.lrs.org/2020/06/17/visualizing-data-the-logarithmic-scale/
- What do the ‘error bars’ represent?

Slide from: Eric Miller