# CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021

HAVERFORD COLLEGE

# Admin

- **Note-taker**: Trinity
- **Lab 6** due Thursday
  - Reminder to start early!
- Office hours TODAY: **3:30-5pm in H204**

# Outline for October 26

- Recap logistic regression and Handout 14
- Midterm 1: runtime (Part 5)
- Convert continuous features to discrete
- Introduction to visualization
- Begin: dimensionality reduction

# Outline for October 26

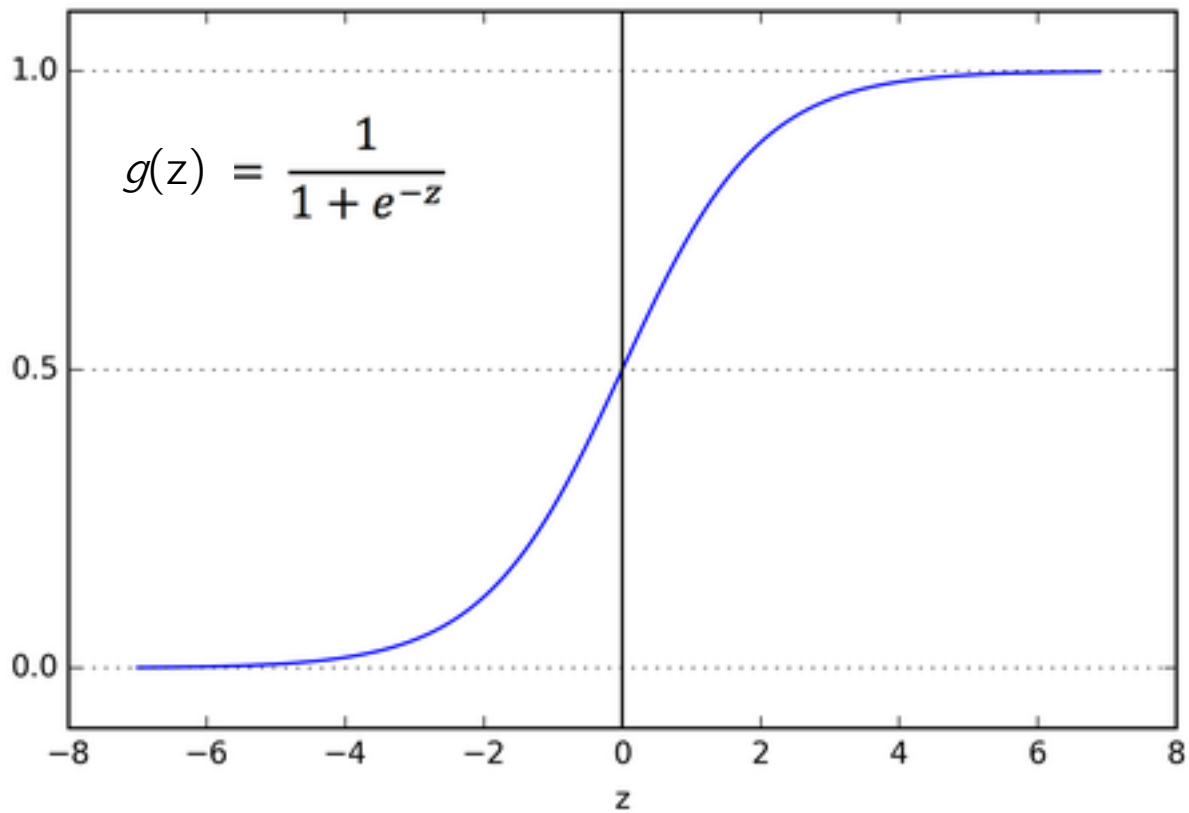
- Recap logistic regression and Handout 14
- Midterm 1: runtime (Part 5)
- Convert continuous features to discrete
- Introduction to visualization
- Begin: dimensionality reduction

# 3 important pieces to SGD

- Hypothesis function (prediction)

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

# Logistic (sigmoid) function



# 3 important pieces to SGD

- Hypothesis function (prediction)

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

- Cost function (want to minimize)

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

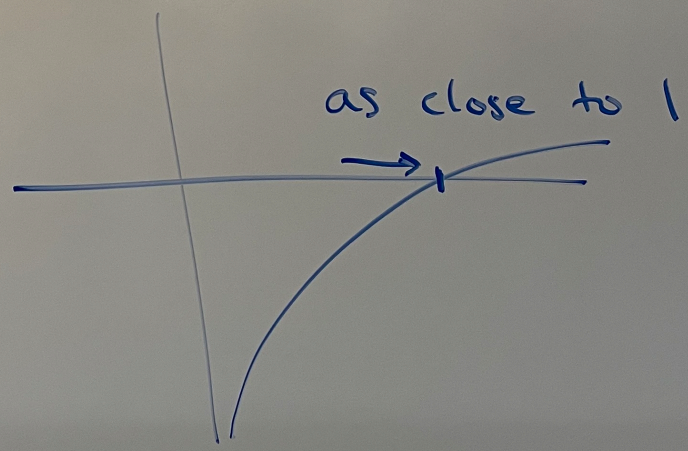
# True and predicted y-values as probability distributions


true

pred $y_i = 1$

$h_{\vec{w}}(\vec{x}_i) = \hat{y}_i = 0.6$

$1 - \hat{y}_i = 0.4$



# 3 important pieces to SGD

- Hypothesis function (prediction)

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

- Cost function (want to minimize)

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

- Gradient of cost wrt single data point $x_i$

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

# Stochastic Gradient Descent for Logistic Regression (binary classification)

```
set w = 0 vector
while cost J(w) still changing:
    shuffle data points
    for i = 1…n:
        w <- w – alpha(derivative of J(w) wrt x_i)
    store J(w)
```

Compare Handout 14 with a partner

# 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:

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

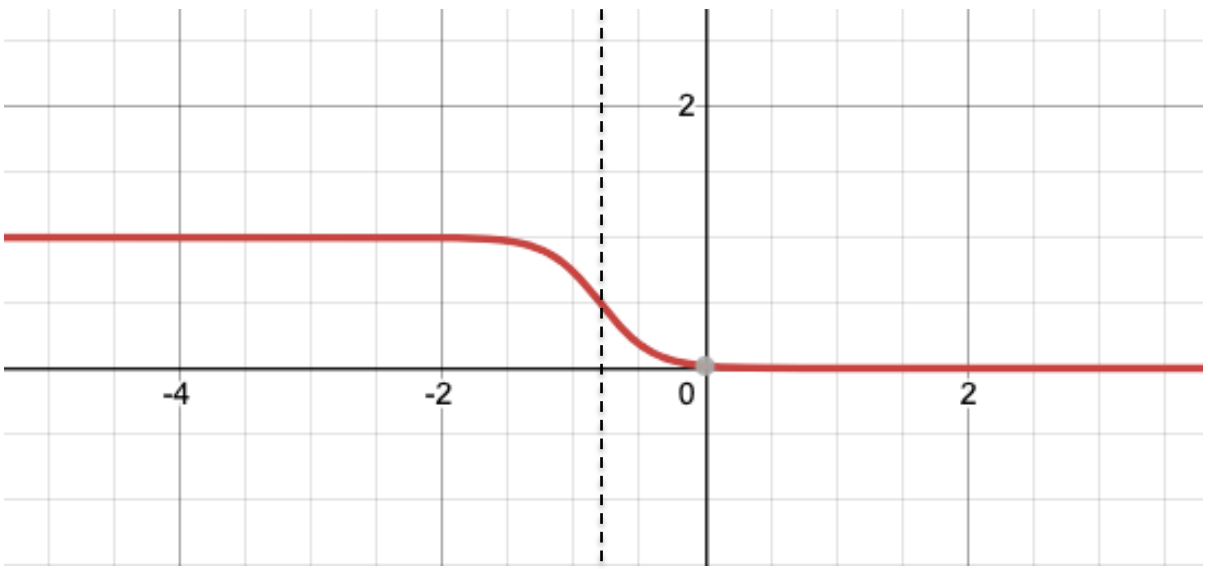
   If $\boldsymbol{w}$ is the zero vector (as it would be when starting SGD), what is the probability $y = 1$?

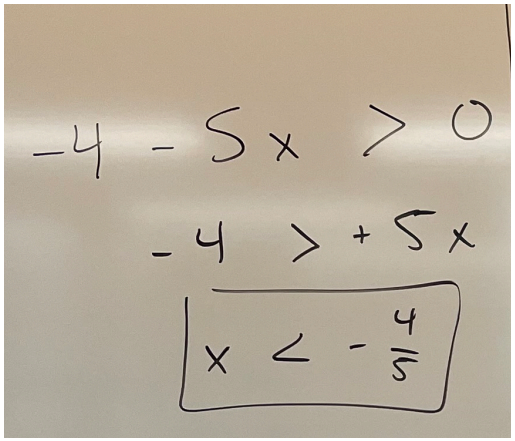
$\frac{1}{2}$

# 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 $x_{\text{test}} = -2$?

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

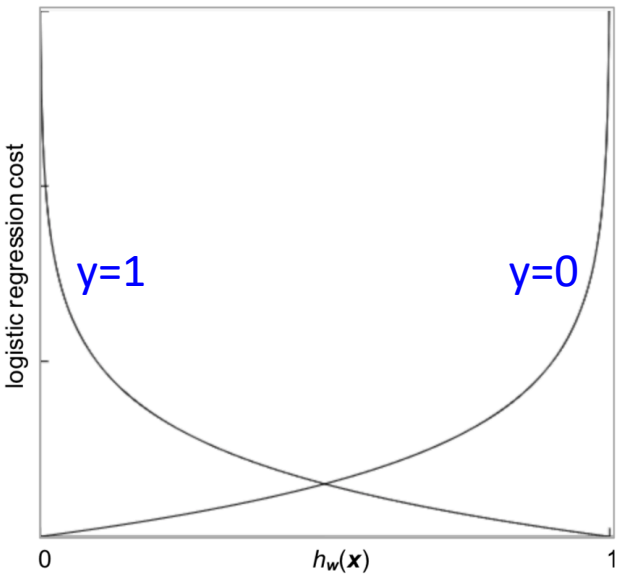



Decision boundary

# 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 $y = 1$?



# Outline for October 26

- Recap logistic regression and Handout 14
- Midterm 1: runtime (Part 5)

Not posted online

- Convert continuous features to discrete
- Introduction to visualization
- Begin: dimensionality reduction

# Outline for October 26

- Recap logistic regression and Handout 14
- Midterm 1: runtime (Part 5)
- Convert continuous features to discrete
- Introduction to visualization
- Begin: dimensionality reduction

# Continuous Features

(do this for the TRAIN only!)

| X | Y |
|---|---|
| 10 | Y |
| 7 | Y |
| 8 | N |
| 3 | Y |
| 7 | N |
| 12 | Y |
| 2 | Y |

1) Sort examples based on given feature

| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 3 | 7 | 7 | 8 | 10 | 12 |
| Y | Y | Y | N | N | Y | Y |

# Continuous Features

(do this for the TRAIN only!)

| X | Y |
|---|---|
| 10 | Y |
| 7 | Y |
| 8 | N |
| 3 | Y |
| 7 | N |
| 12 | Y |
| 2 | Y |

1) Sort examples based on given feature

| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 3 | 7 | 7 | 8 | 10 | 12 |
| Y | Y | Y | N | N | Y | Y |

2) Different label with same feature value, collapse to “None”

| | | | | | |
|---|---|---|---|---|---|
| 2 | 3 | 7 | 8 | 10 | 12 |
| Y | Y | None | N | Y | Y |

# Continuous Features

(do this for the TRAIN only!)

| X | Y |
|---|---|
| 10 | Y |
| 7 | Y |
| 8 | N |
| 3 | Y |
| 7 | N |
| 12 | Y |
| 2 | Y |

1) Sort examples based on given feature

| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 3 | 7 | 7 | 8 | 10 | 12 |
| Y | Y | Y | N | N | Y | Y |

2) Different label with same feature value, collapse to "None"

| | | | | | |
|---|---|---|---|---|---|
| 2 | 3 | 7 | 8 | 10 | 12 |
| Y | Y | None | N | Y | Y |

3) Whenever label changes, make a feature (use avg)

| | | | | | |
|---|---|---|---|---|---|
| 2 | 3 | 7 | 8 | 10 | 12 |
| Y | Y | None | N | Y | Y |

X <= 5

X <= 7.5

X <= 9

$\ldots \quad \boxed{x \geq 7.5}$

$\boxed{x \geq 5}$

$\vec{w} \cdot \vec{x}$

| | feature $x$ | label $y$ |
|---|---|---|
| T | 10 | Y |
| T | 7 | Y |
| T | 8 | N |
| F | 3 | Y |
| T | 7 | N |
| T | 12 | Y |
| F | 2 | Y |

$\Rightarrow \boxed{O(n)}$

① sort

| 2 | 3 | 7 | 7 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| Y | Y | Y | N | N | Y | Y |

② colapse features with more than one label

| 2 | 3 | 7 | 8 | 10 | 12 |
|---|---|---|---|---|---|
| Y | Y | None | N | Y | Y |

$x \geq 5$ $\quad$ $x \geq 7.5$ $\quad$ $x \geq 9$

③ change in label $\Rightarrow$ make a feature!

avg feature values

# Continuous Features (Handout 15)

(do this for the TRAIN only!)

| temp | Y |
|---|---|
| 80 | Y |
| 48 | Y |
| 60 | N |
| 48 | N |
| 40 | N |
| 48 | Y |
| 90 | Y |

1) Sort examples based on feature “temp”

2) Different label with same feature value, collapse to “None”

3) Whenever label changes, make a feature (use avg)

# Handout 15

| $x \geq 44$ | temp | label |
|---|---|---|
| T | ~~80~~ | Y |
| T | ~~48~~ | Y |
| T | ~~60~~ | N |
| T | ~~48~~ | N |
| F | ~~40~~ | N |
| T | ~~48~~ | Y |
| T | ~~90~~ | Y |

① 40 48 48 48 60 80 90

N Y N Y N Y Y

② 40 | 48 | 60 | 80 90

N | None | N | Y Y

$x \geq 44$ $x \geq 54$ $x \geq 70$

# Outline for October 26

- Recap logistic regression and Handout 14
- Midterm 1: runtime (Part 5)
- Convert continuous features to discrete
- Introduction to visualization
- Begin: dimensionality reduction

# 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

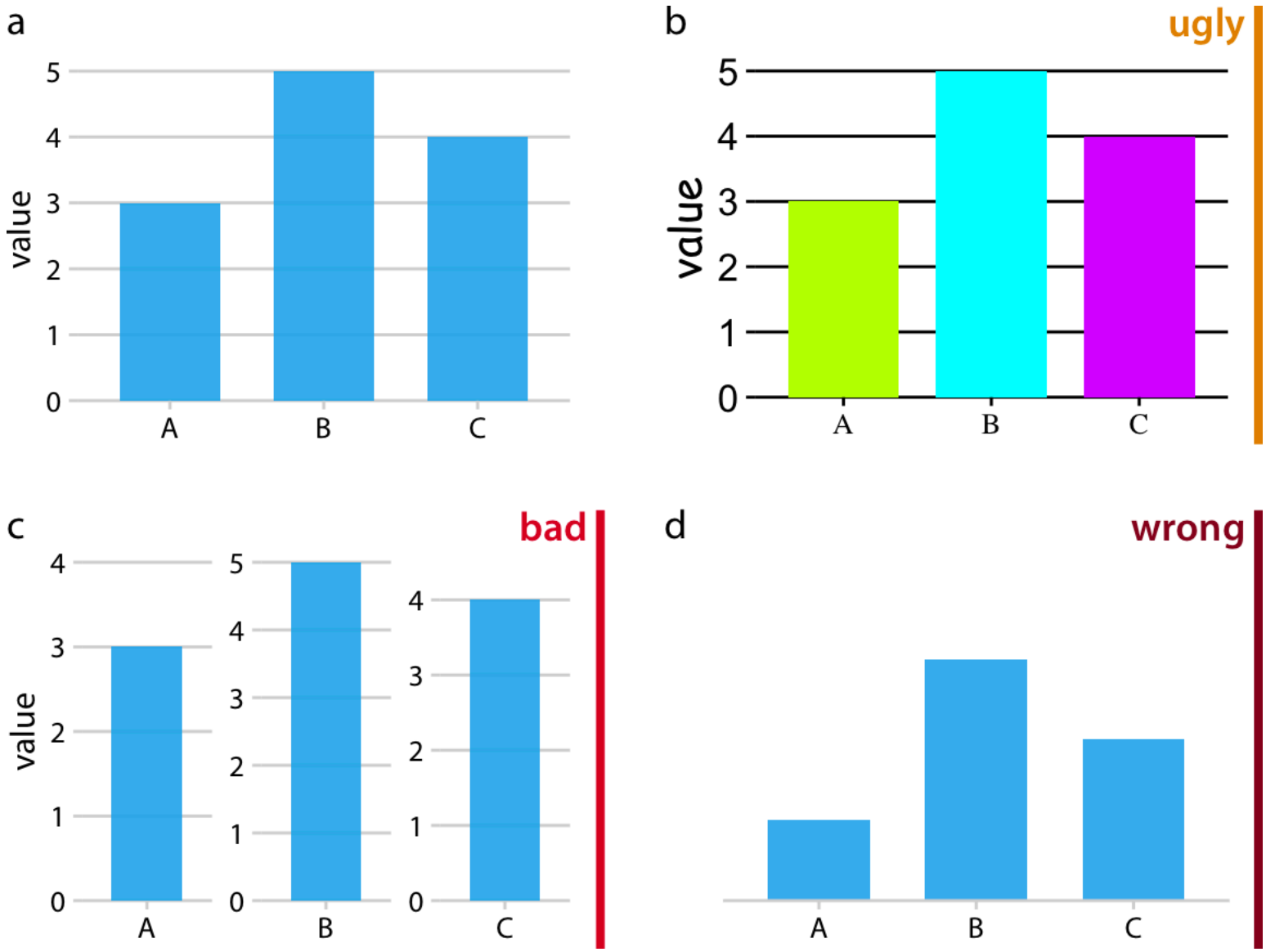



Fig 1.1 from “Fundamentals of Data Visualization” by Claus Wilke

Two different visualizations of the same data

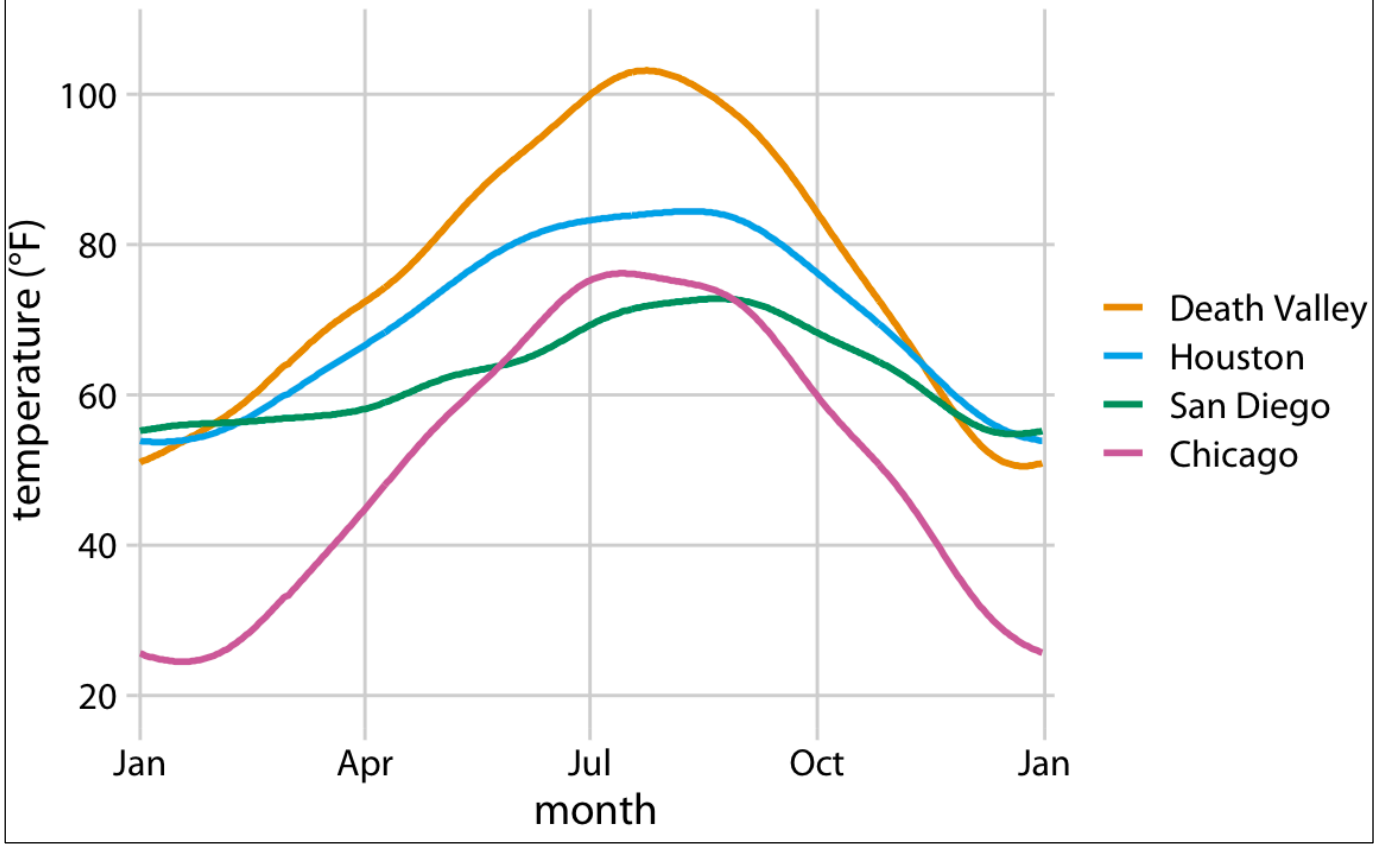
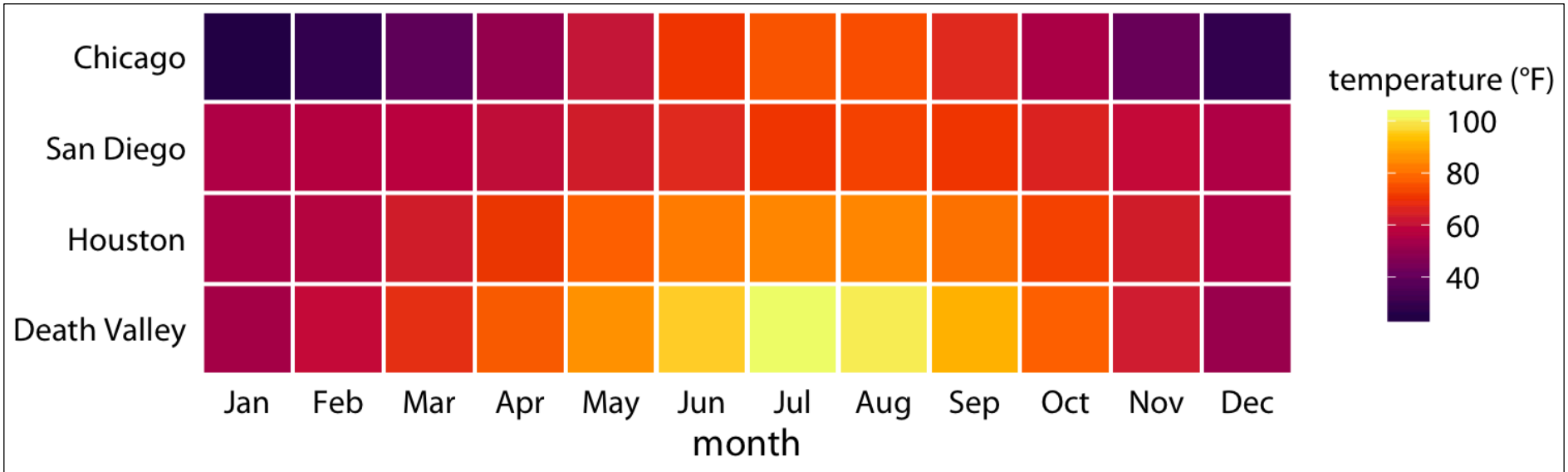



Fig 2.3/2.4 from "Fundamentals of Data Visualization" by Claus Wilke



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

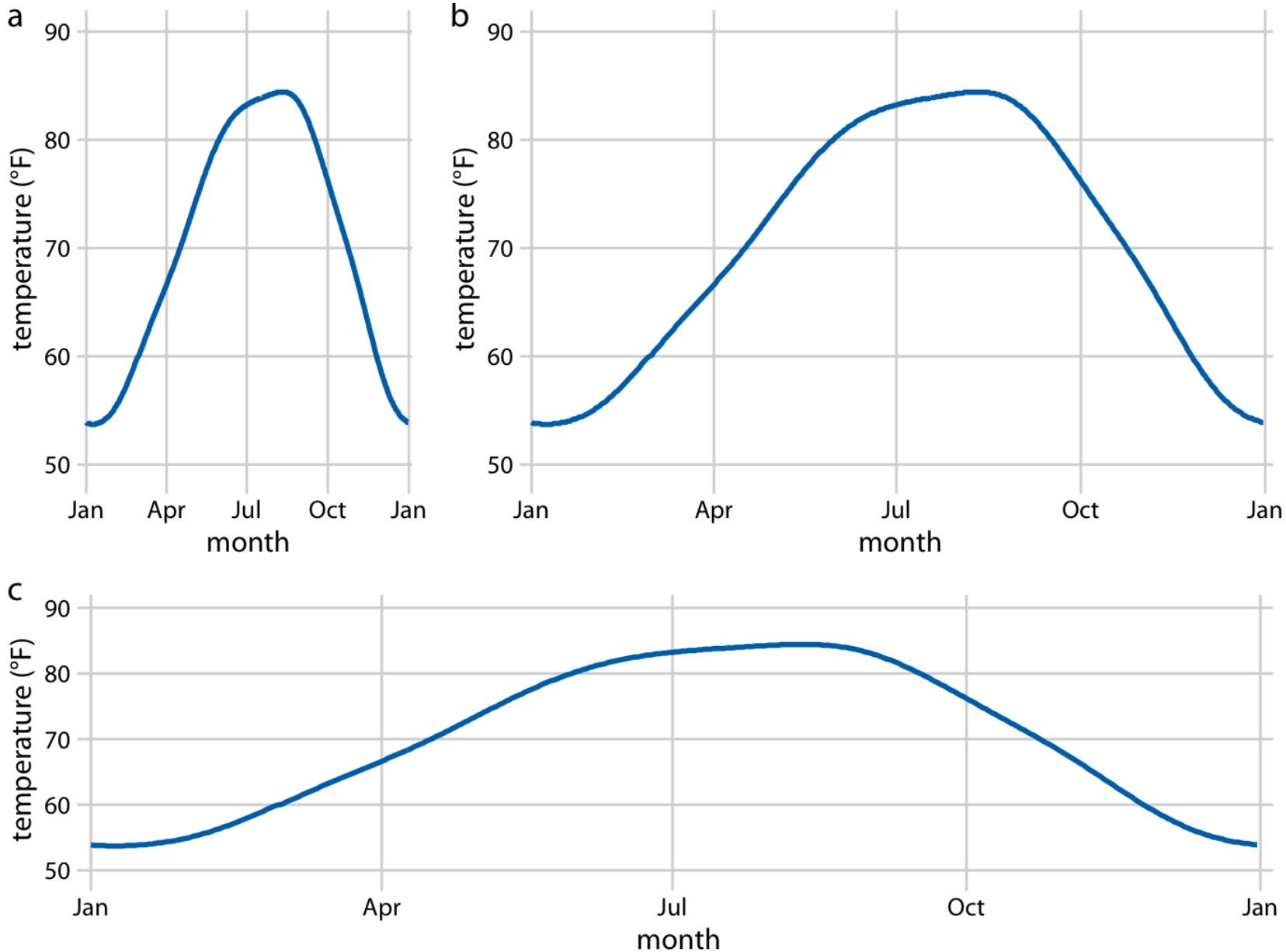



Fig 3.2 from “Fundamentals of Data Visualization” by Claus Wilke

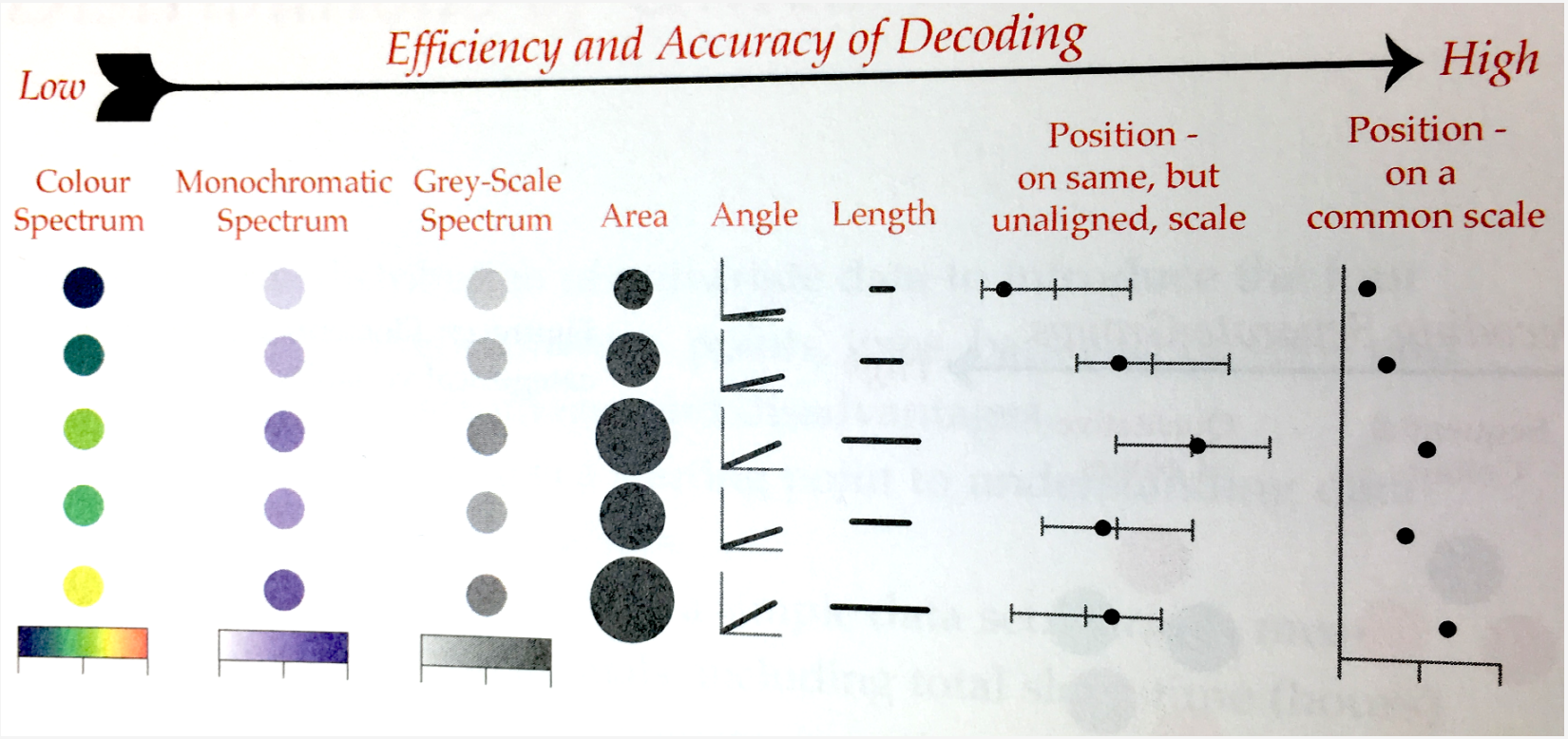
# Data Types

- Continuous vs. Discrete

Slide from: Eric Miller

# Data Types

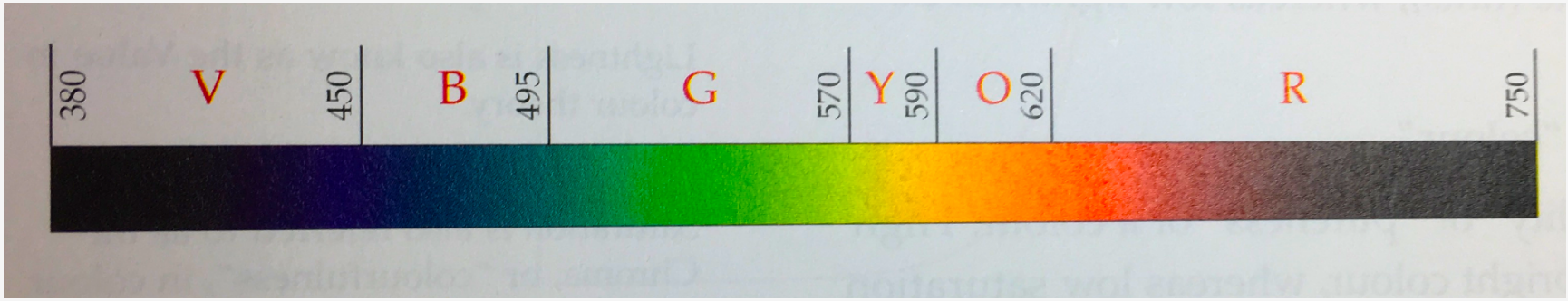
Continuous




Slide from: Eric Miller

# Data Types

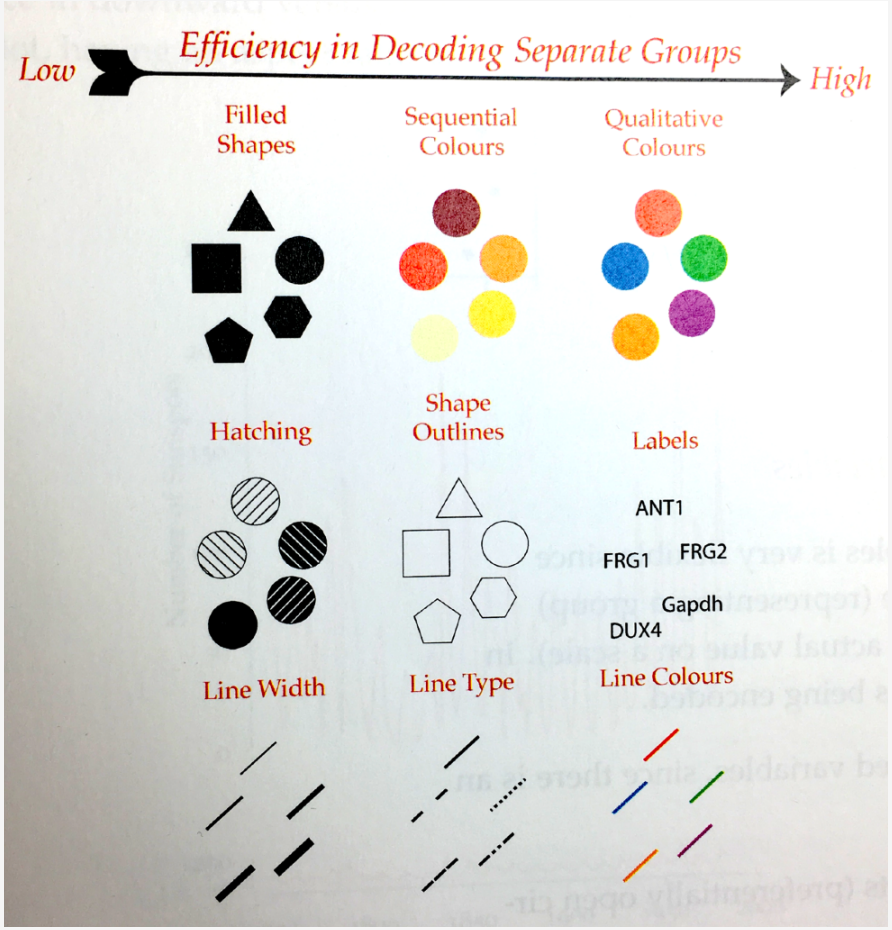
Continuous




Slide from: Eric Miller

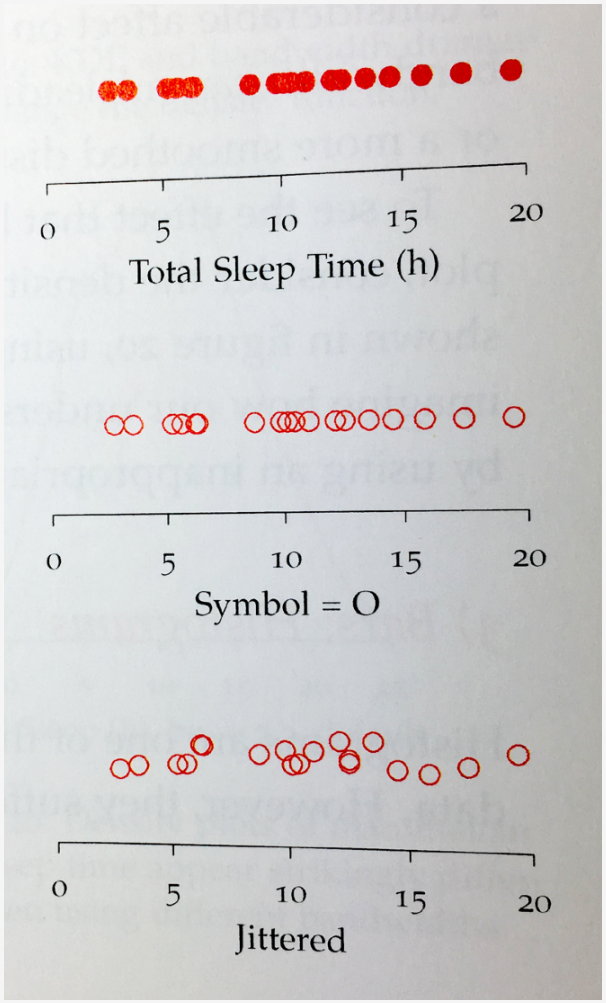
# Data Types

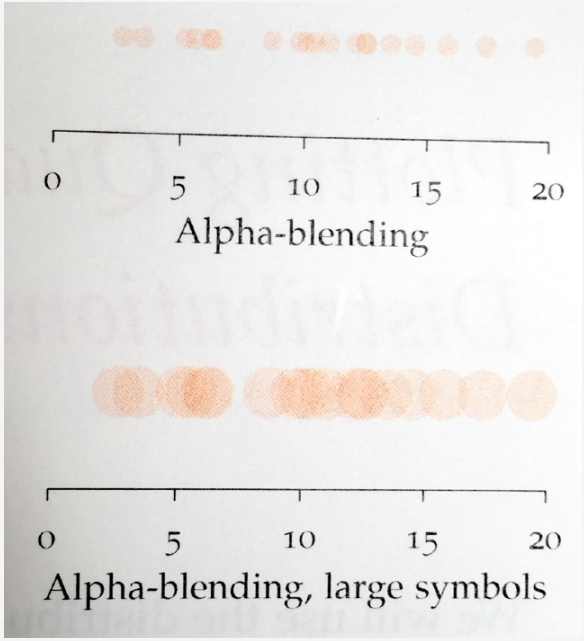
Discrete




Slide from: Eric Miller

# Overplotting




Slide from: Eric Miller

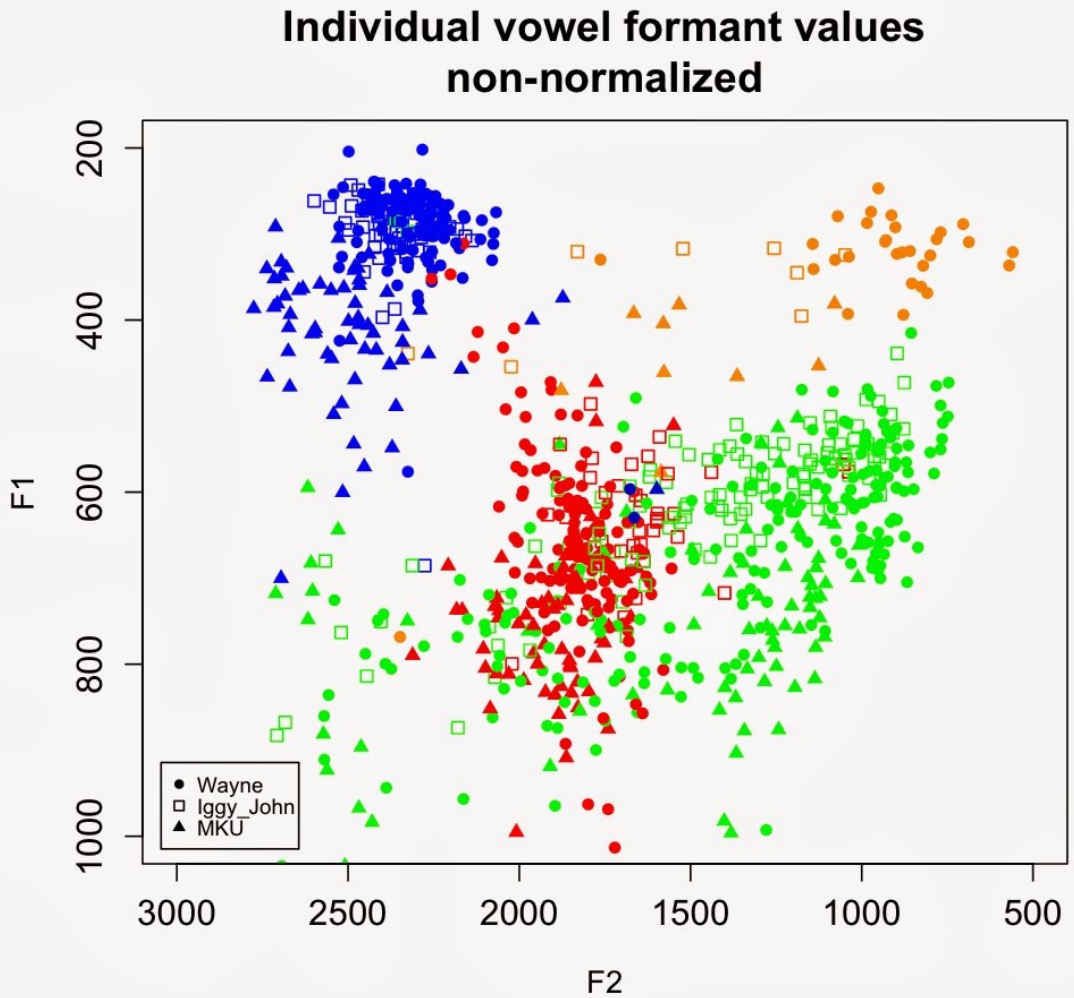
# Color

Why use color?

Which colors?

Slide from: Eric Miller

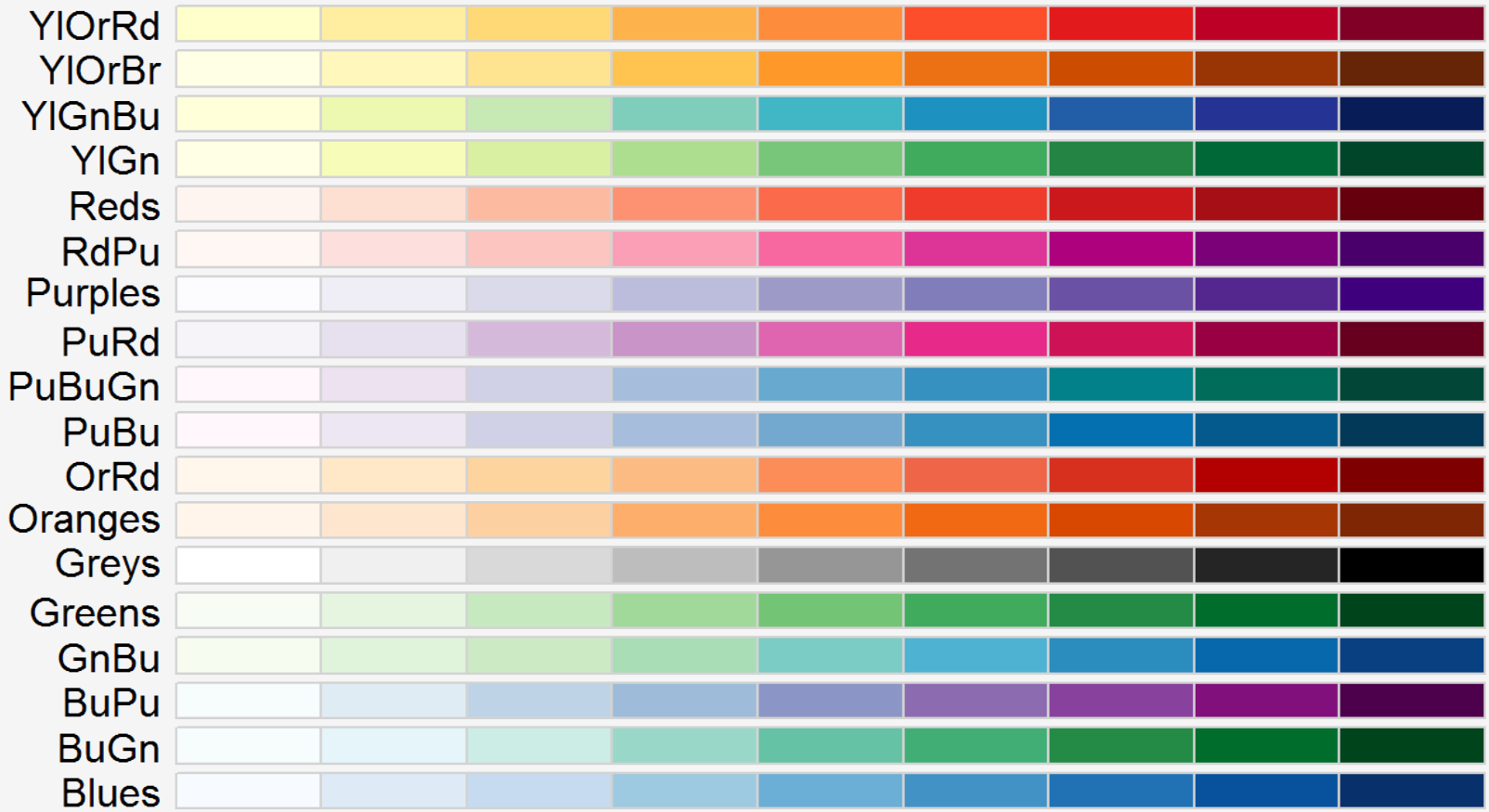
# Color




Slide from: Eric Miller

# Color

## Sequential




Slide from: Eric Miller

# Sequential example

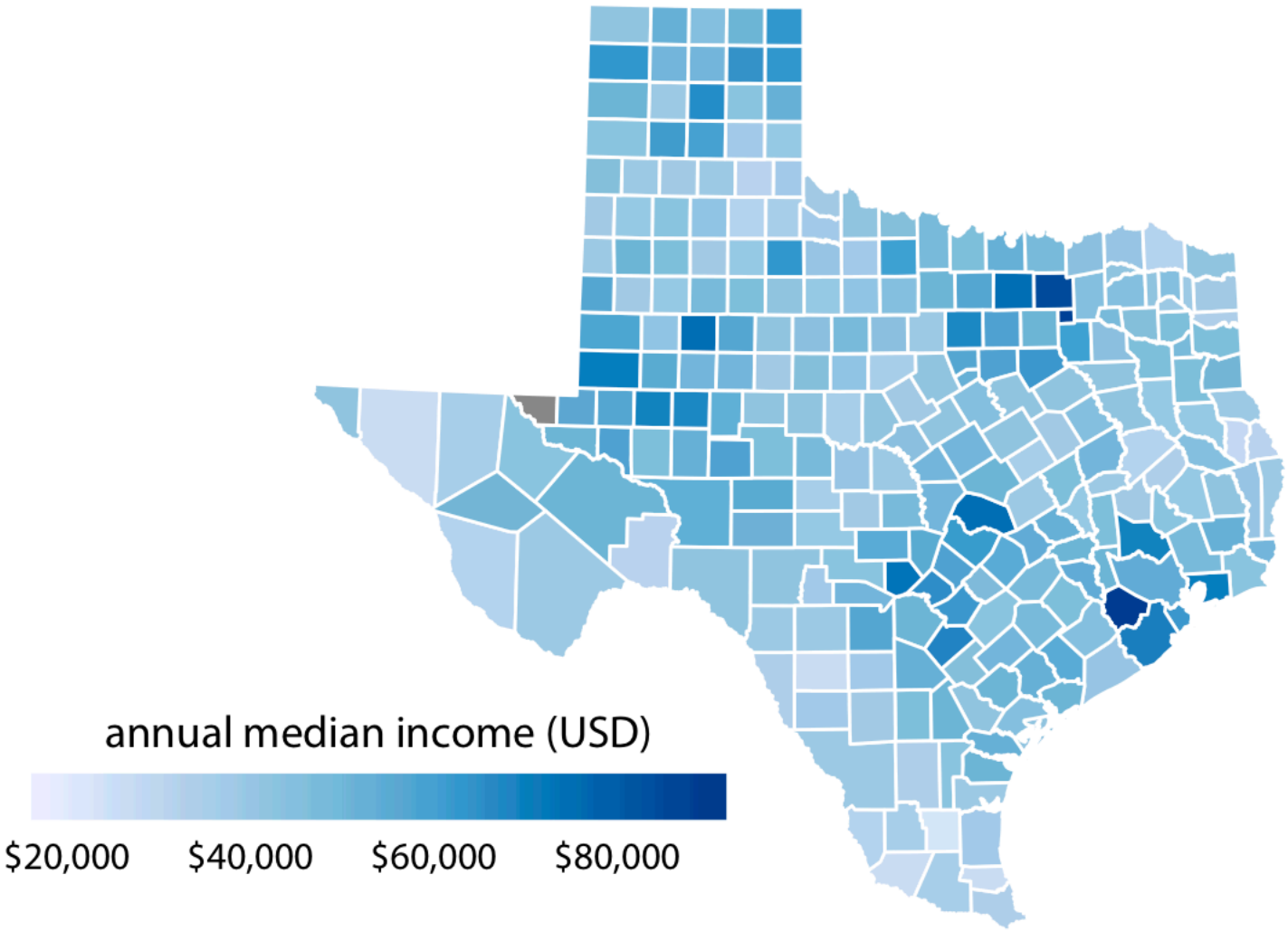
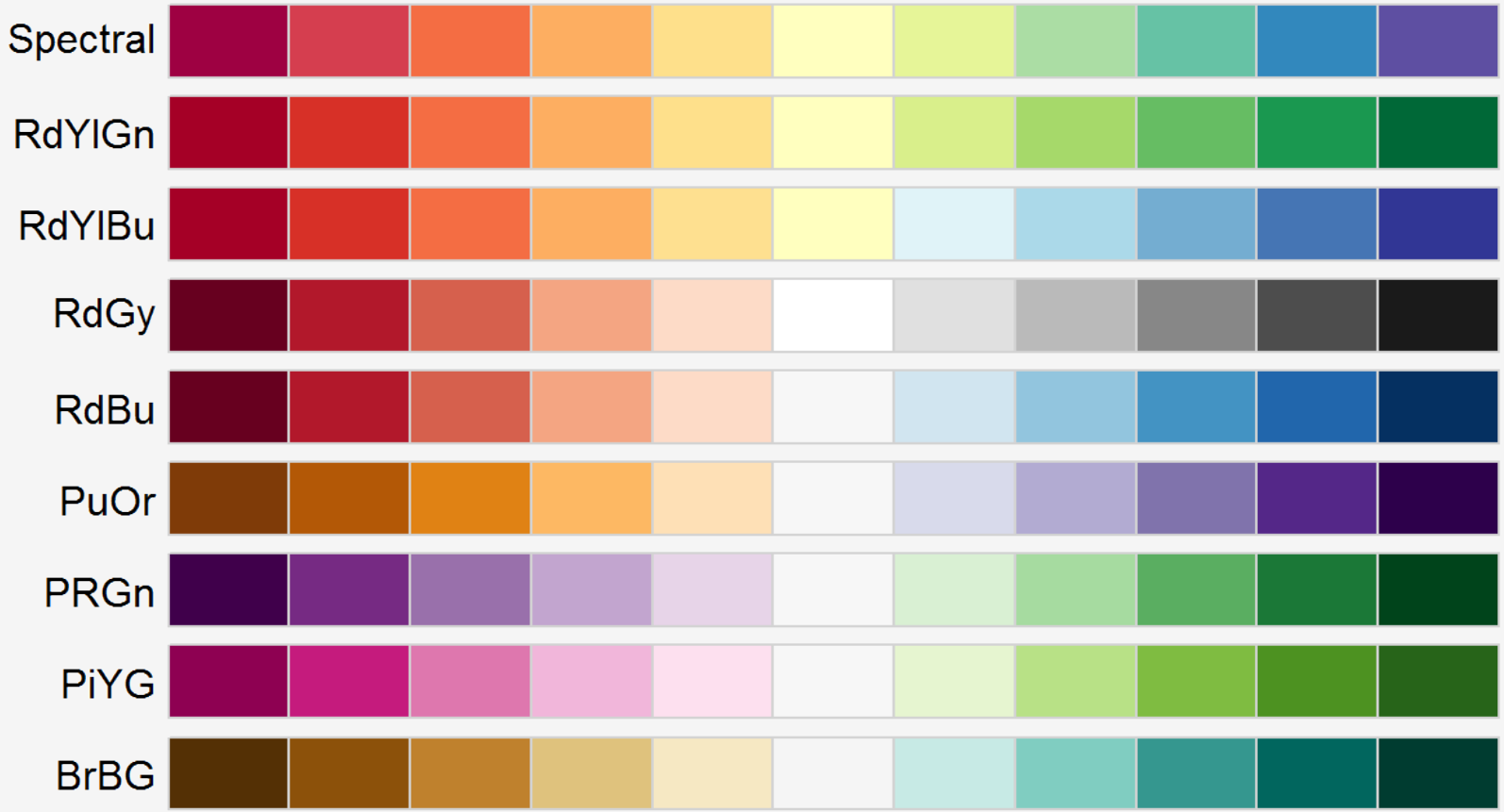



Fig 4.4 from “Fundamentals of Data Visualization” by Claus Wilke

# Color

## Diverging




Slide from: Eric Miller

# Diverging example

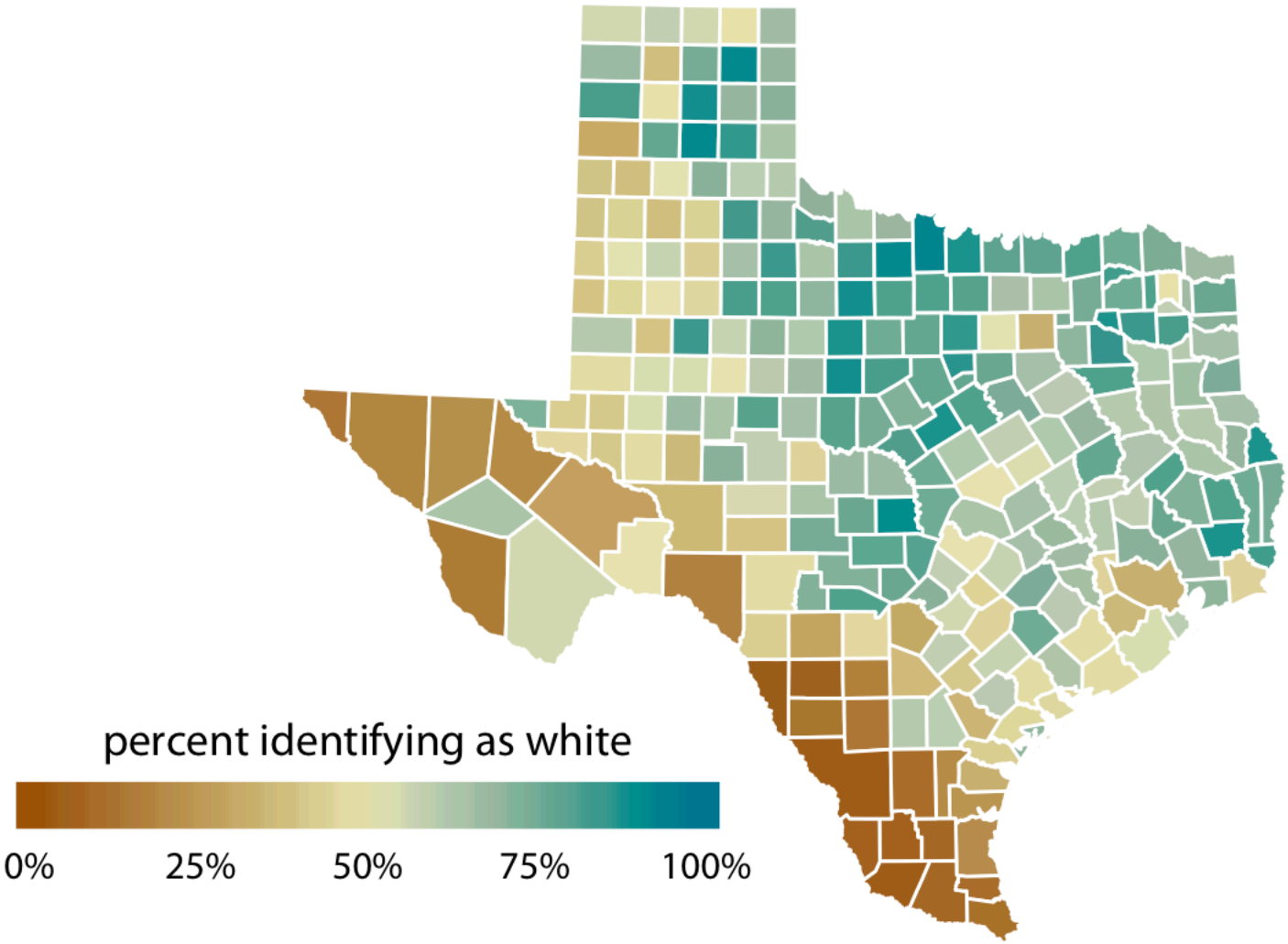
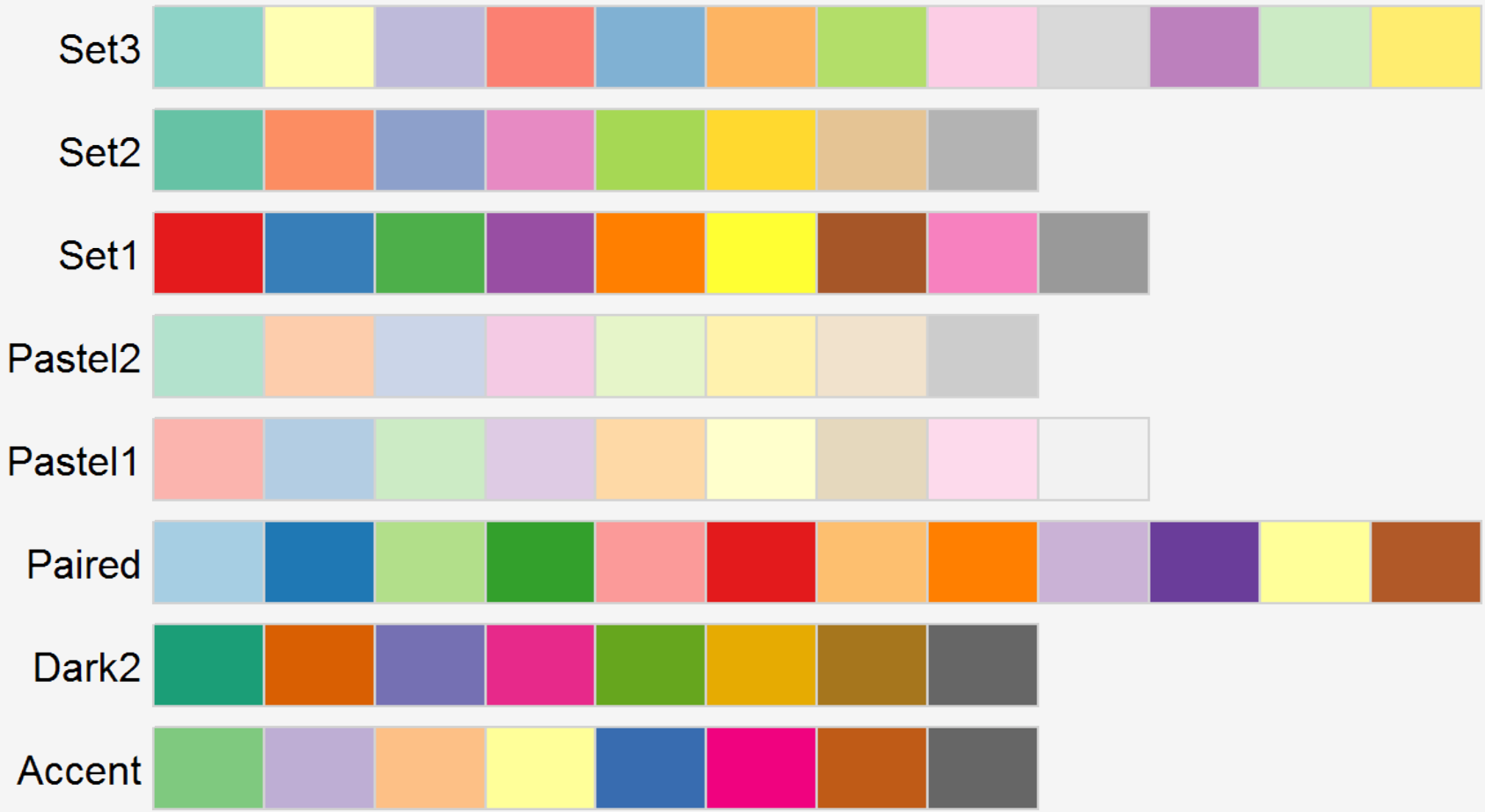



Fig 4.6 from “Fundamentals of Data Visualization” by Claus Wilke

# Color

## Qualitative




Slide from: Eric Miller

# Qualitative example

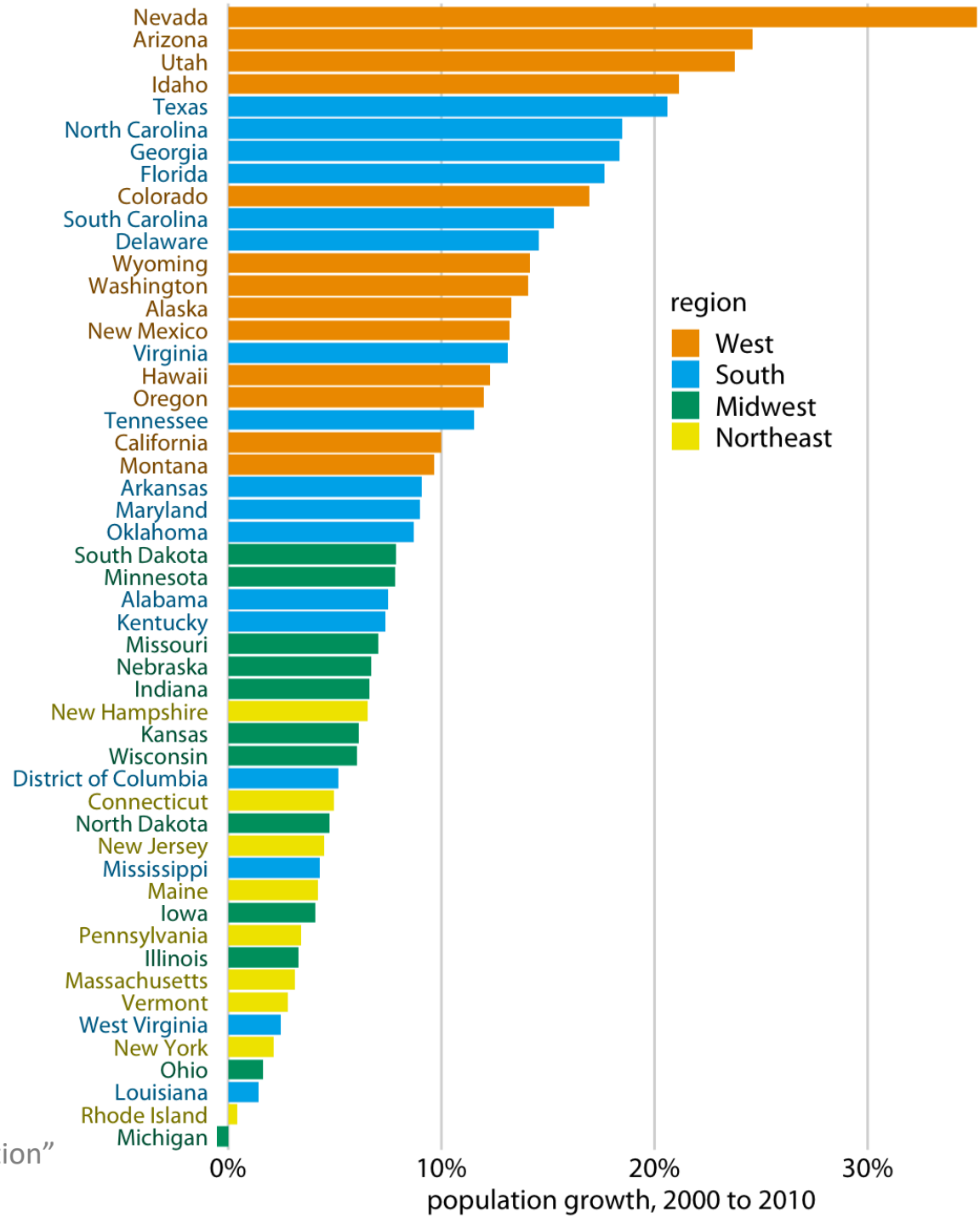
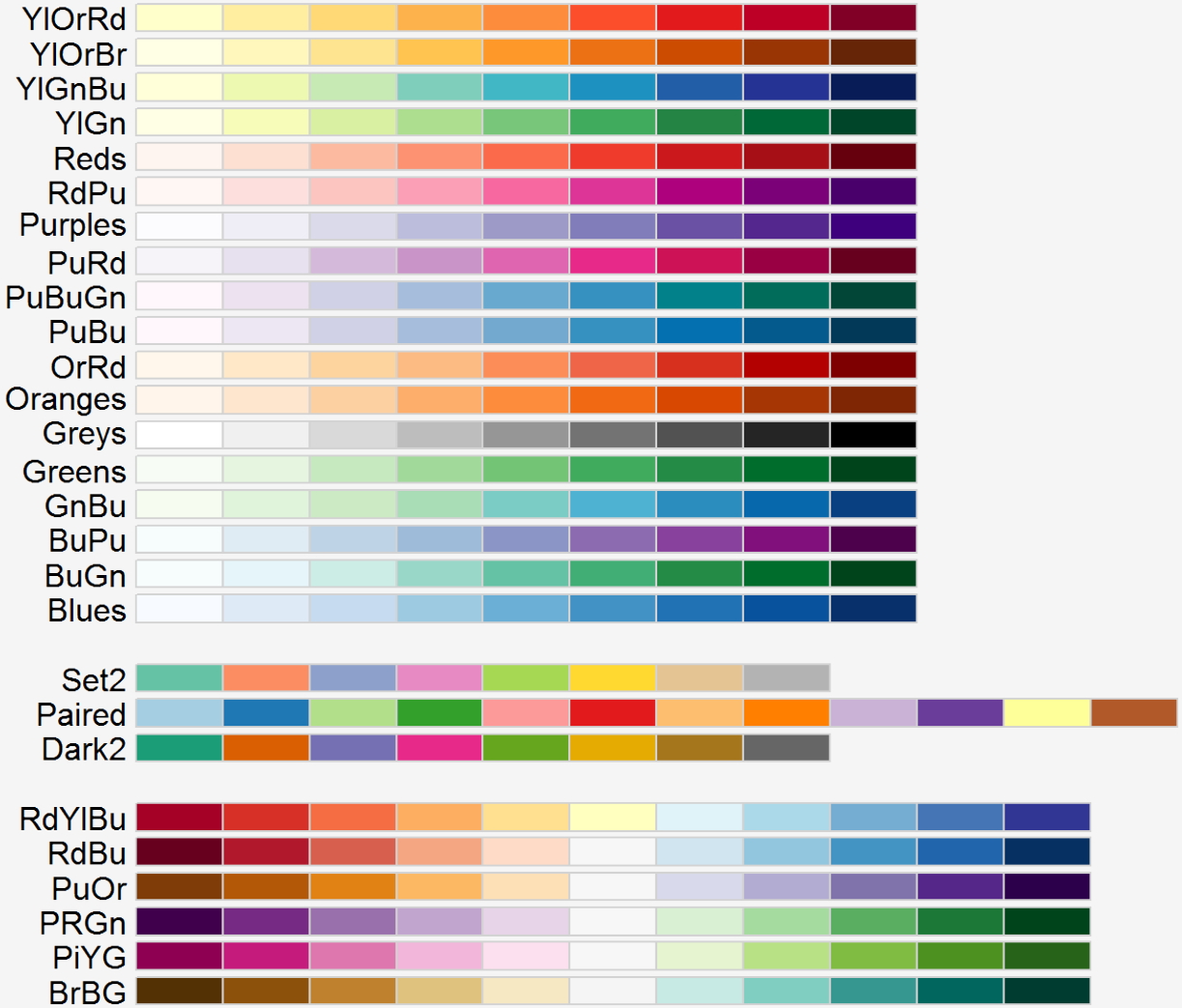



Fig 4.2 from "Fundamentals of Data Visualization" by Claus Wilke

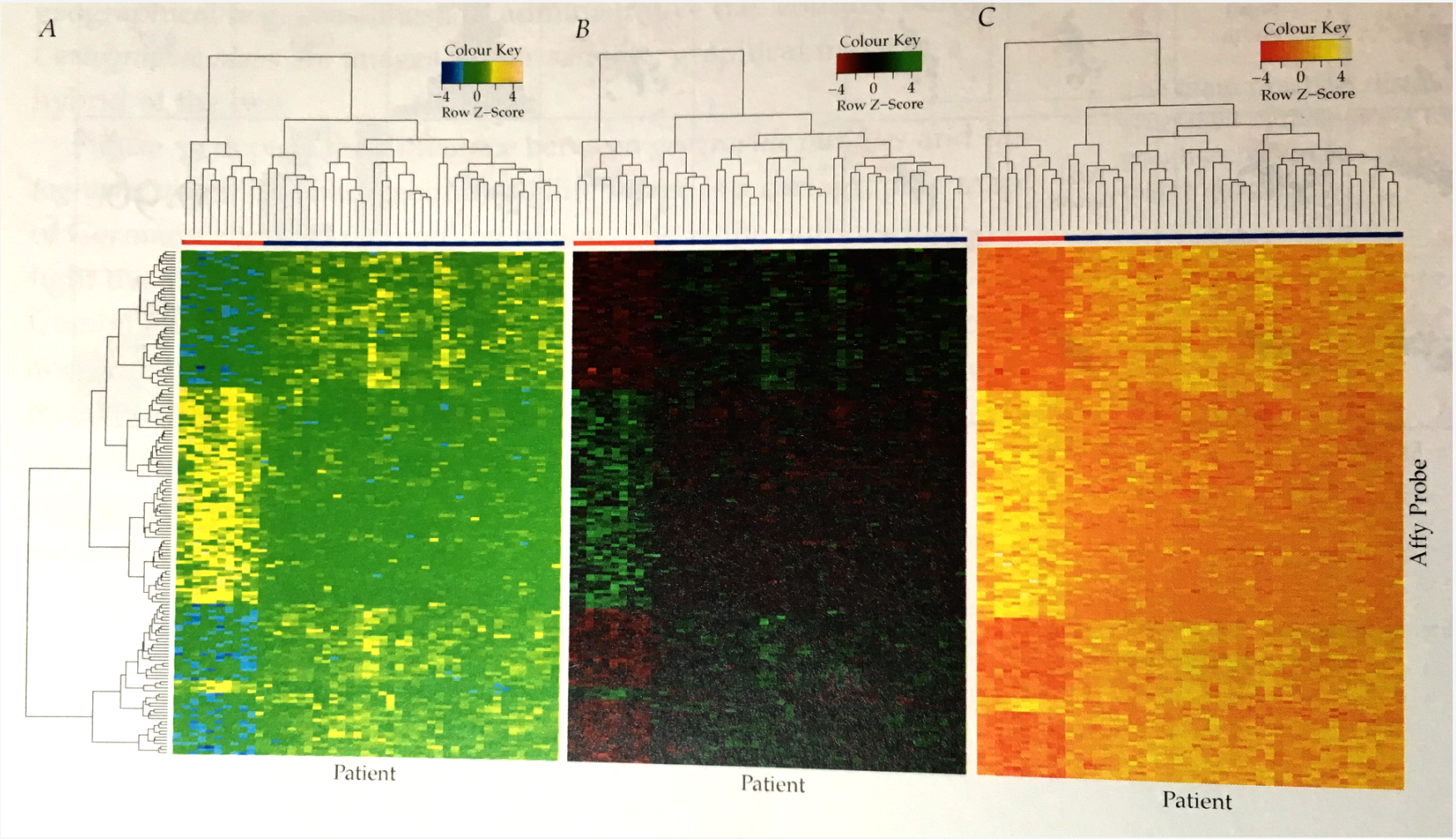
# Color

Color-blind friendly



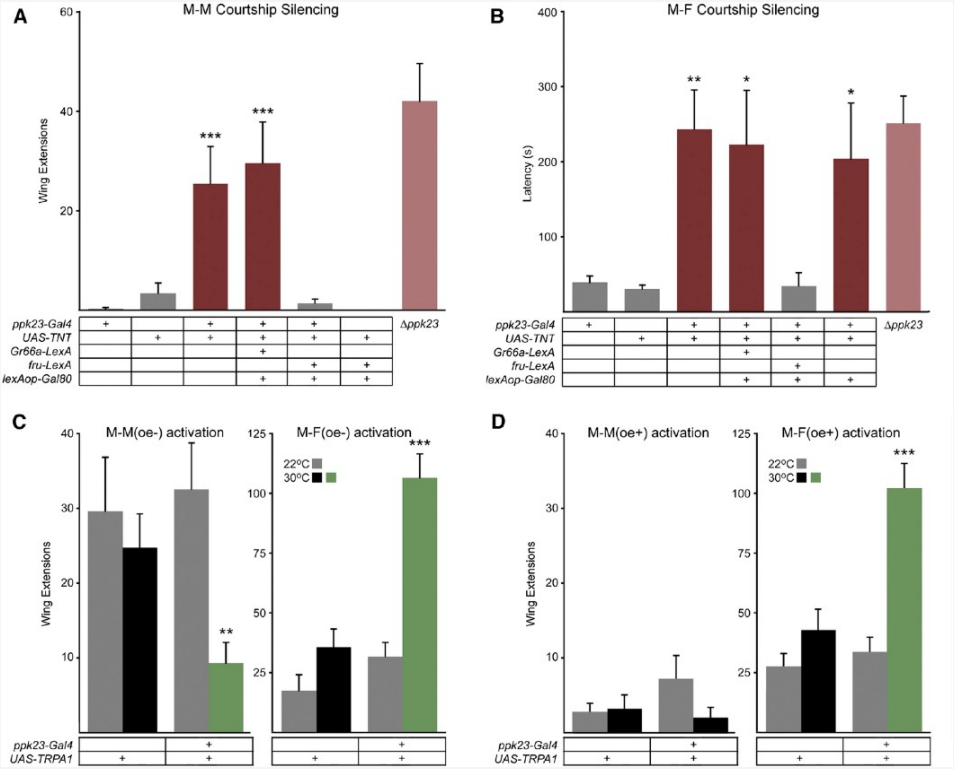

Slide from: Eric Miller

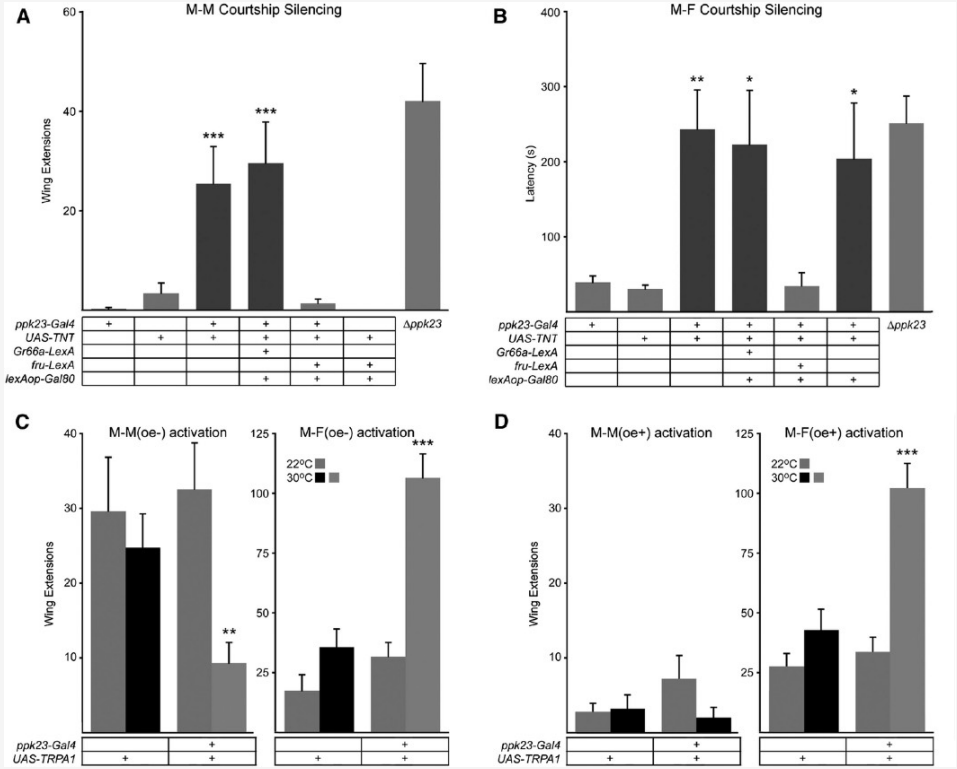
# Color




Slide from: Eric Miller

# Color

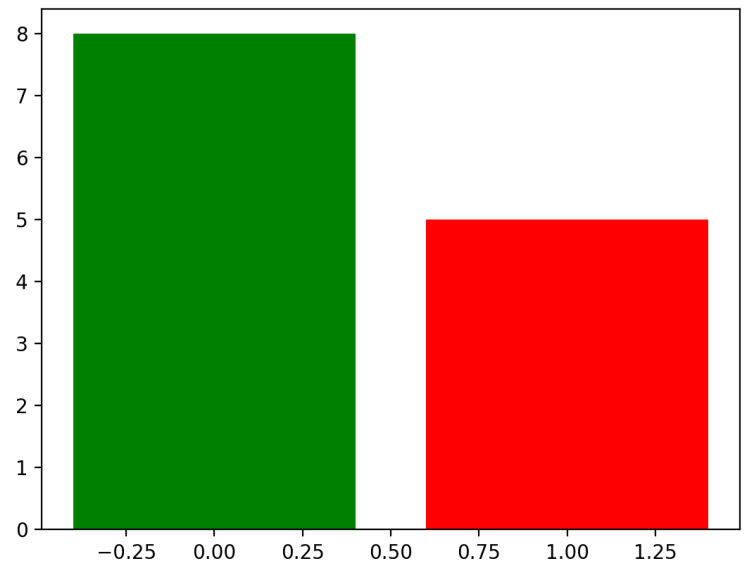


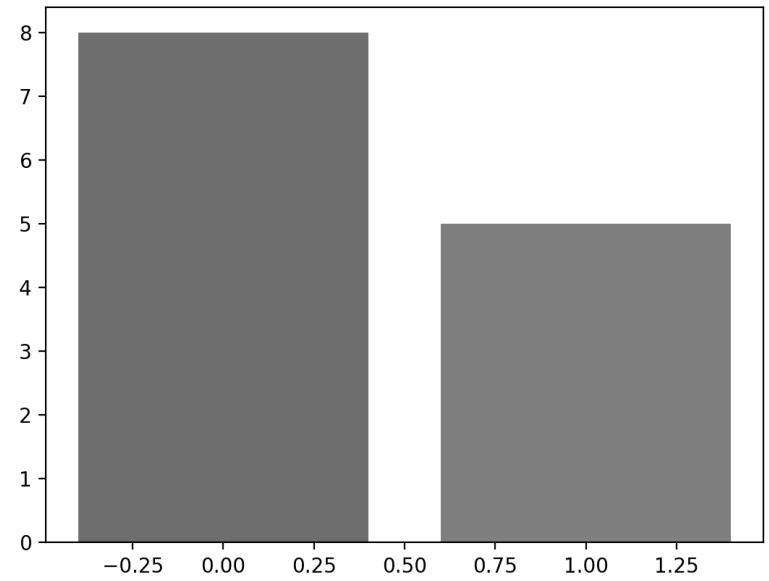




A)Red bars denote responses different than ppk23-Gal4 control.
B)Red bars denote responses different than ppk23-Gal4
C)???
D)???

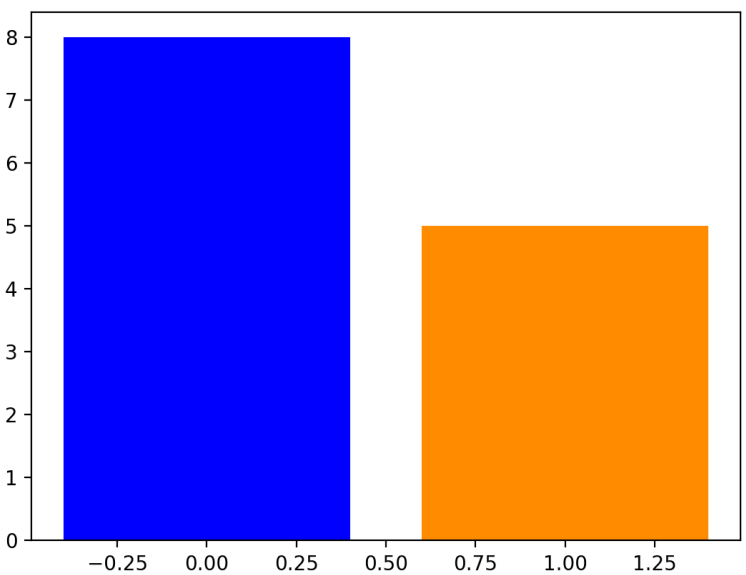
Slide from: Eric Miller

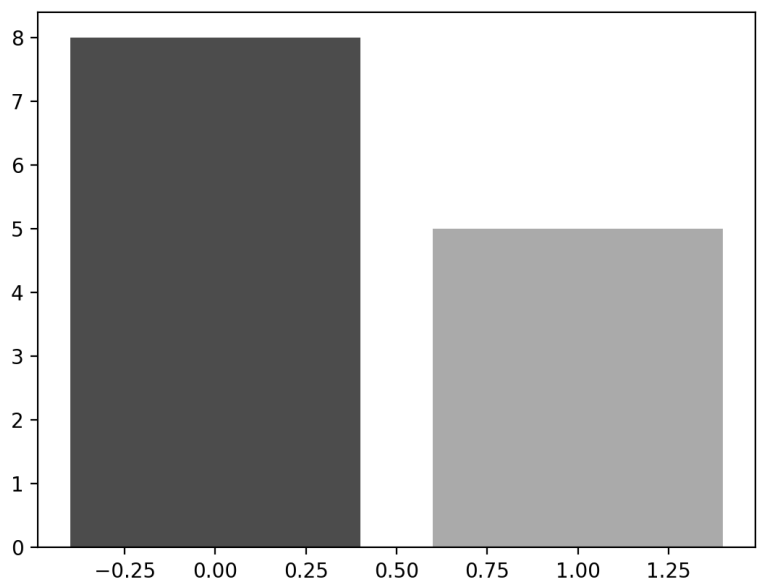
# Red/green vs. blue/orange




To black and white



# Color

- Avoid harsh, non-printable colors
- Use color with purpose: highlight data, or make it equally visible
- Do not encode continuous variables with color; categorical variables are best
- Avoid red/green; think about black/white printing
- Consistently use a color scheme throughout your figures

Slide from: Eric Miller