

# Module 3: Data Visualization

Ellen Bledsoe

2023-02-21

## A Visualization Primer

### Why Does Data Visualization Matter?

#### Small Groups

As a group, take 5 minutes to compare your notes and come to a group consensus on your top 3 reasons. Be sure to choose someone to report out!

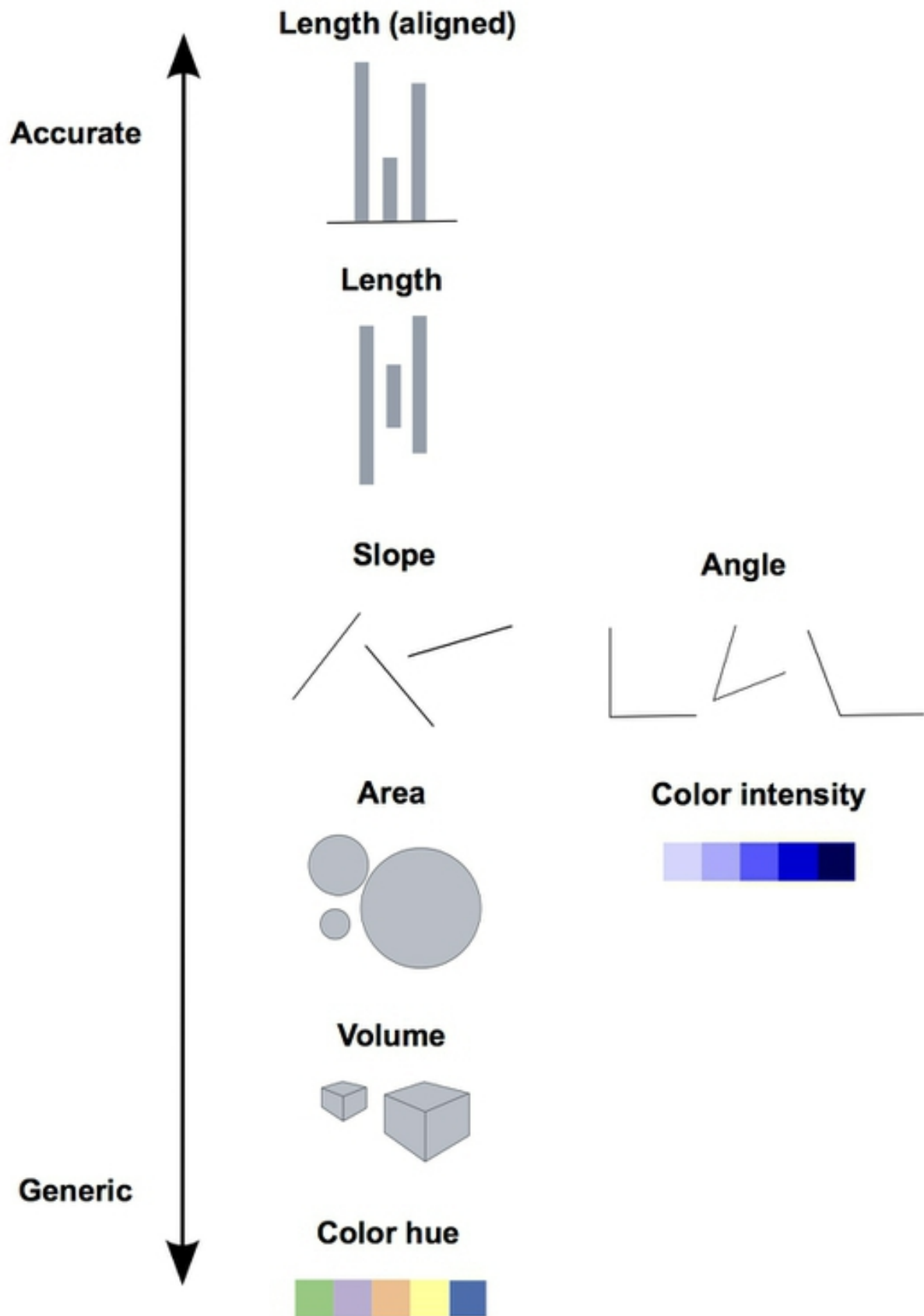
### Types of Visualization

#### Visual cues for communicating data

Here is a decent overview of some of the core concepts of data visualization.

This website is pretty great and goes into a lot of detail about good practices in data visualization. If this is something that really piques your interest, I encourage you to check it out!

In this course, our main take-away from this website is the use of visual cues to communicate data and which ones are better than others.



## Small group activity

In groups, discuss as many all types of visualization that come to mind. It's okay if you don't know what they are called! Make a quick list or draw them out if you prefer.

Types of data visualizations:

This website is an amazing reference for data visualization methods and when to use what. It also has examples of each type plotted in `ggplot2`.

We will talk more about how to choose the right visualization for your data now and also in the rest of the module.

## Data matching activity

Still in your groups, open the PDF called “data\_viz\_matching”

Spend a few minutes seeing if you can match the data descriptions to the types of data visualizations.

Each “dataset” is pulled from the same overall dataset: measurements from 344 individual penguins from different species and on different islands over 3 years. Some datasets created multiple plots.

- A. Penguin body mass and flipper lengths
- B. Penguin flipper lengths
- C. Penguin species and flipper lengths
- D. Penguins body mass, species, and flipper lengths

## Matches

- 1.
- 2.
- 3.
- 4.
- 5.
- 6.

See the PowerPoint on D2L for full details!

## Figure Critique

First, read through this blog post on the “Dos and Don'ts of Data Visualization”

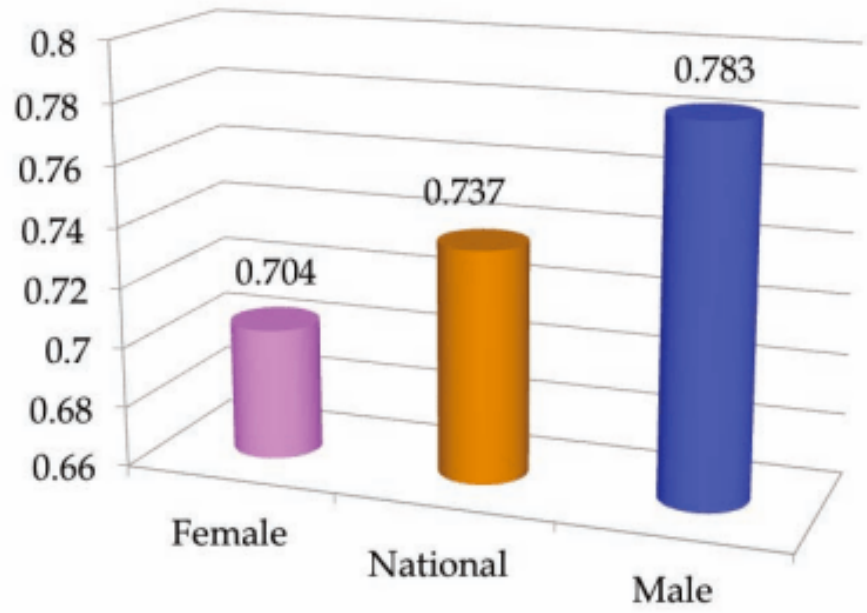
Now that you have some insights, let's critique these figures below. Yes, these are actual figures in the wild...

- what aspects don't work
- what aspects do work
- how how you present the data?

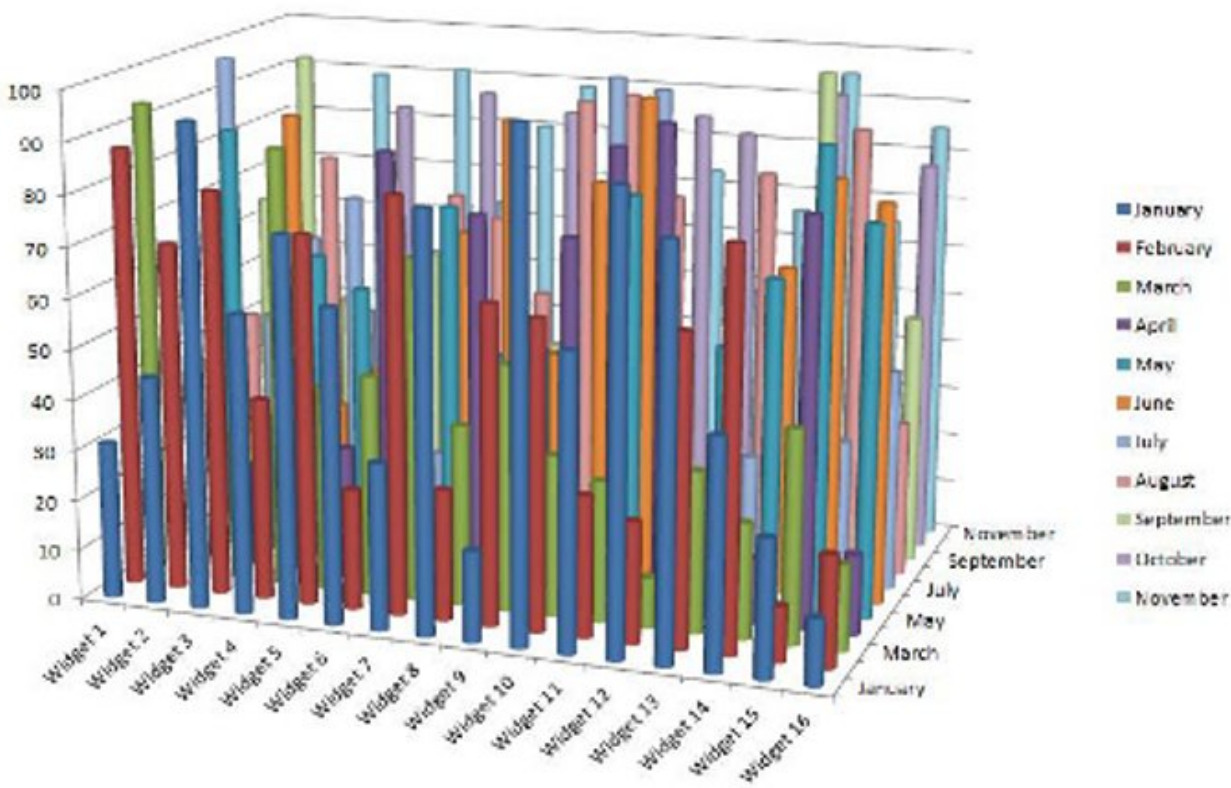
Figures:

- 1.

Figure 11: GNH index by gender



2.



3.

## Distribution of All TFBS Regions

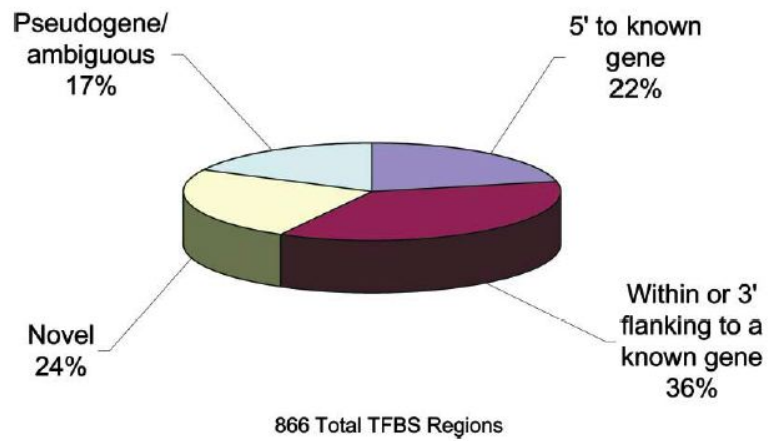


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5' most exon of a gene, within 5 kb of the 3' terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

4.

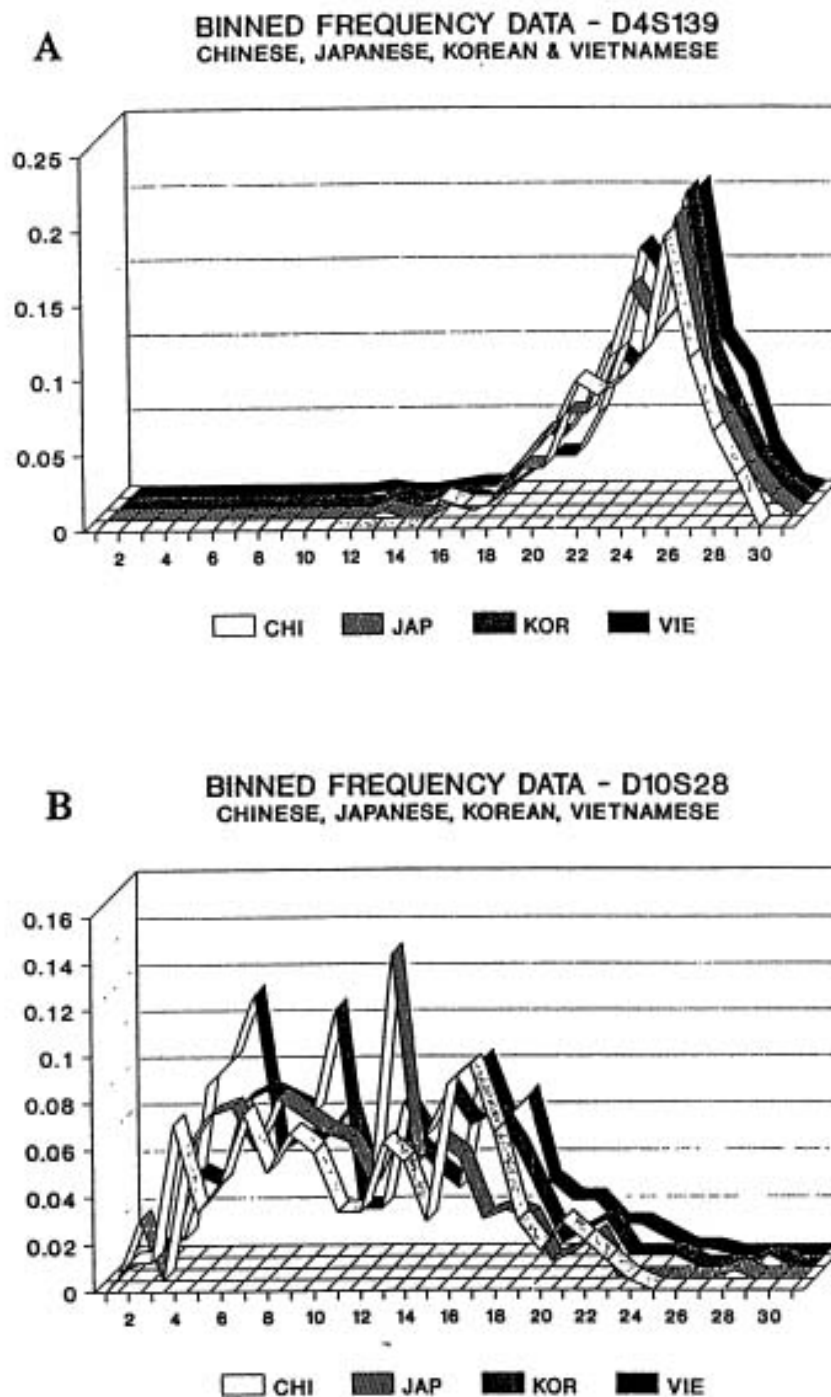


FIG. 4. Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are 103, 125, 93 and 215 for D4S139 and 120, 137, 100 and 193 for D10S28. The horizontal axis is the bin number; bins are not of equal length.