Plotting **One Variable Distributions** Cheat Sheet

MIT Biological Engineering Communication Lab

Understanding your data

- How many points do you have?
- What does your data look like?
- What is your message?

Example: "4 drug conditions, n = 3 mice per condition, each condition has small ranges"

▶ **Condition 1 results in longer survival.**

Choosing a representation

Bar graph
Discrete, n < 10
***must also show points**

Dot plot
Discrete, n < 30

Boxplot
Discrete/continuous, n > 10
****should also show points***

Violin plot
Continuous, n > 30
****can also show points***

Ridgeline plot (Joy plot)
Discrete/continuous, n > 30

Elements checklist

- ☐ Plot + data labels (label directly if possible)
- ☐ Appropriate axes scales and tick marks
- ☐ Gridlines if desired and needed
- ☐ Redundant (color + shape) markers
- ☐ Lines and points are thick and clear
- ☐ All text is clearly legible
- ☐ Units and annotations are directly on plot
- ☐ Plot is reproducible from clean workspace
- ☐ High-resolution output (DPI > 300)
- ☐ Statistical markers added if needed
- ☐ Peer/mentor proofread

Building a figure

"Biomechanics of predator–prey arms race in lion, zebra, cheetah and impala"

*Data was pulled directly from **Wilson et al. Nature (2018)** and describes peak power/ kg of muscle for pairs of predatory and prey.*

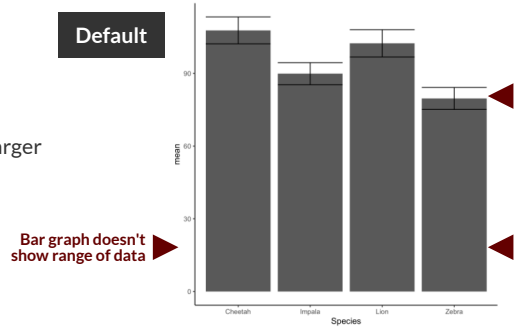
What does our data look like?
4 groups, n > 30 per group with a broad distribution per group

Intended message
Predator peak muscle power is larger prey muscle power, only slightly.

↳ **Boxplot or violin plot**

Follow along with the associated code notebooks in Python and R.

Default

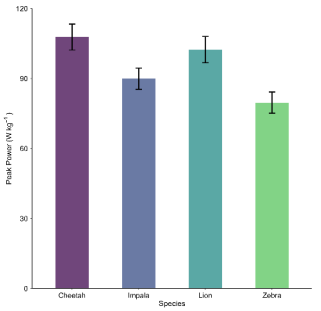


What do we need to fix?

- ☐ Text size is illegible
- ☐ Bars are wide and uniform
- ☐ Scales are off
- ☐ The distribution is hidden

Changing the axes limits and adjusting the bars eliminate

- ☑ ~~Bars are wide and uniform~~
- ☑ ~~Scales are off~~



Boxplot

> great for highlighting characteristics of the underlying distribution, but

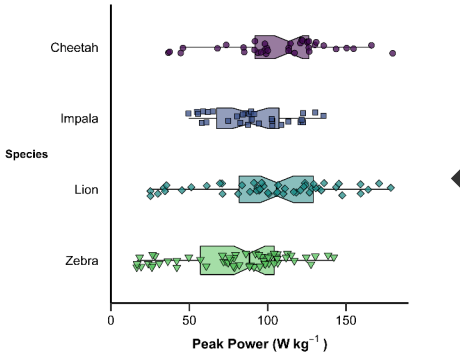
> can still be misleading if the points are not shown.

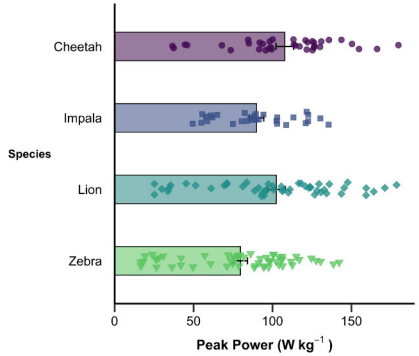
The notches represent the 95% confidence interval of the medians, providing

> the **characteristic meaningful to our message**

> information about the uncertainty of the median

> **ability for viewer to interpret** differences between species





Bar Graph*

Let's focus on

- ☐ Text size is illegible
- ☐ The distribution is hidden

1. Display all the points
2. Increase the font sizes

To make this even clearer,

- ☑ Colors and markers are redundant
- ☑ Horizontal plot makes labels easier to read

*With this data, we **would not** suggest a bar graph

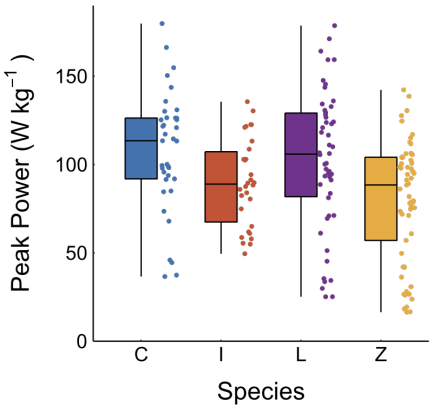
The boxplot was the authors' choice
Below is our reproduced version:

What are the strengths?

- ☑ Points are shown
- ☑ Points next to boxplots make it easier to understand both
- ☑ Large text, simple axes
- ☑ Median information make it easier to compare between species

What are the weaknesses?

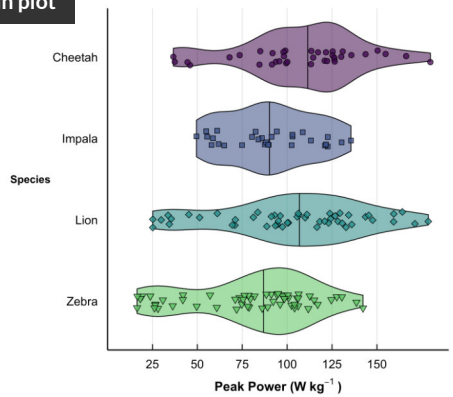
- ☐ Abbreviated labels, which are repeated throughout the paper, do not standalone
- ☐ The distribution of points is difficult to quantify
- ☐ Gathering exact quantities is difficult from few axis labels



Violin plot

We can address these limitations

- ☑ By adding **clear, horizontal species labels**
- ☑ **Using a violin plot** to display the underlying distribution
- ☑ **Light gridlines** and **expanded axis** allows to identify values and compare between species



All code and figures are stored at **https://github.com/MIT-BECL/awesome-becl-resources**

Learn more at **http://mitcommlab.mit.edu/be** – Updated August 2018