Intro to R for biologists Day 3

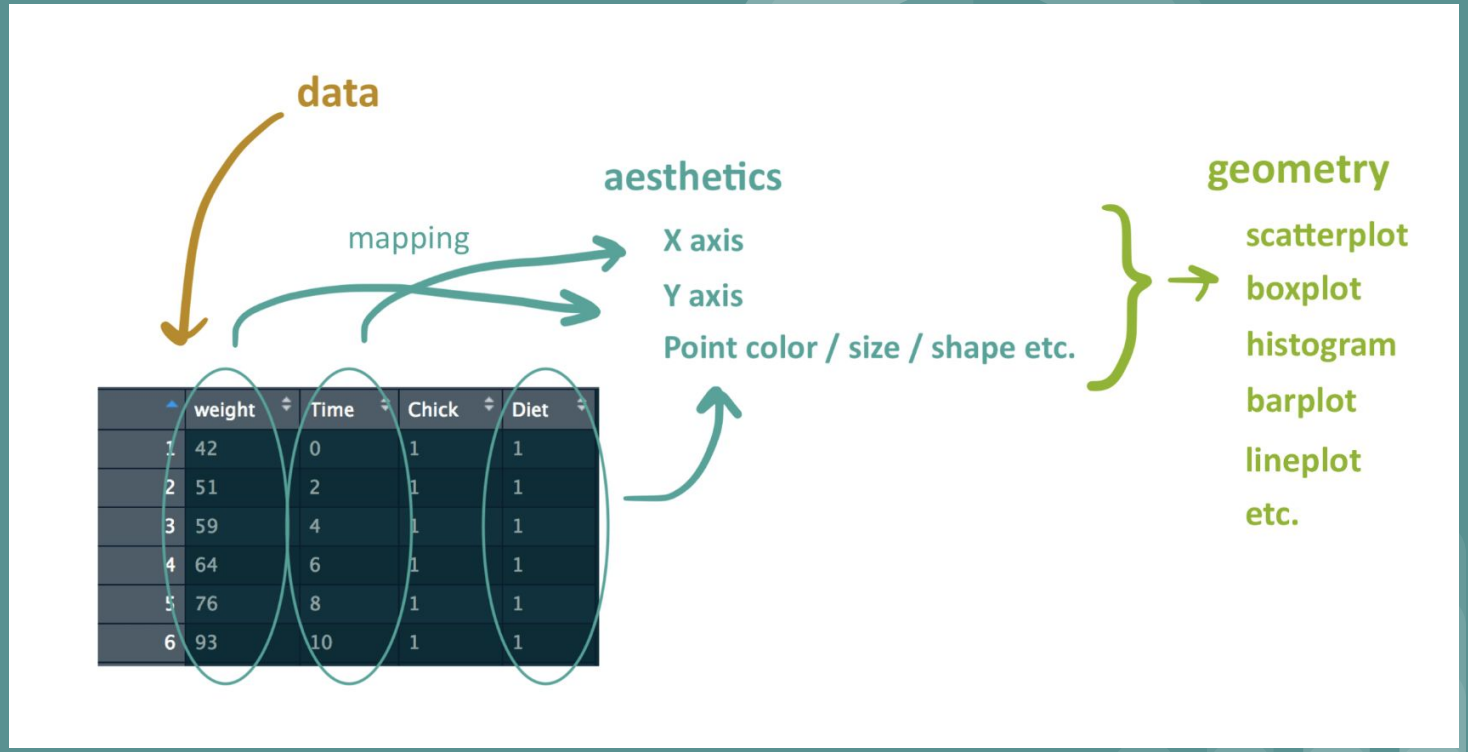
Data visualization with ggplot2
Analysis of your sequencing data

Brandon Ely, Doctoral Candidate
PhD program in Biology (Molecular, Cellular, Developmental)
CUNY Graduate Center

*Adapted from Dr. Weigang Qiu's "R Tutorials for biologists"

ggplot2 syntax

- Data
- Aesthetics
- Geometry



	weight	Time	Chick	Diet
1	42	0	1	1
2	51	2	1	1
3	59	4	1	1
4	64	6	1	1
5	76	8	1	1
6	93	10	1	1

Image credit:
https://www.rforecology.com/post/a-simple-introduction-to-ggplot2/

Building your plot

Specify your data and aesthetics:

```
ggplot(data = df, aes(x = Grazing, y = Root, color = Grazing))
```

Layer on the visuals (geometries):

**you can use just one or several at once*

```
+ geom_boxplot() #box plot
+ geom_point() # dot plot
+ geom_line() #line plot
+ geom_smooth() #add a trend line
+ geom_bar() #bar plot
```

The arrangement of data in your dataframe is very important when using it with ggplot!

Practice on the Iris datset

1. Make a box or violin plot for one of the numeric variables (include all species)
2. Make a scatter plot for 2 different numeric variables (include all species)
3. Make a panel of box plots for all 4 numeric variables (include all species)

Ecology statistics

From your sequencing results, we are going to answer the following questions using standard ecology methods and statistics:

1. What is the community composition of each sample?
    - Use relative abundance to look at composition of communities at the phylum level

2. Which sample is most diverse? Least diverse?
    - Alpha diversity using Shannon Index

3. Which samples are most similar/dissimilar to each other?
    - Beta diversity using Bray-Curtis dissimilarity metric