

ODSC: Data Visualization with ggplot2

Part 1: Thinking with graphs

by Martin Frigaard

Written: February 08 2022

Updated: March 13 2022

Resources

Links:

- [Conference Website](#)
- [Website](#)
- [Part 1](#)
- [Part 2](#)

Materials:

- [RStudio.Cloud](#)
- [Github Repo](#)

Outline

Part 1

Exploratory data analysis

- *What is it, who does it, and why it's important*

A Bayesian mindset

- *Priors → new information → posteriors*

The grammar of graphics

- *Layers, aesthetics, and geoms*

Part 2

Build labels first

- *Set expectations*

Exercises & solutions

- *RStudio.Cloud*

Creating graphs

- *Building graphs layer-by-layer, global vs. local mapping, visual encodings*

Applying the grammar

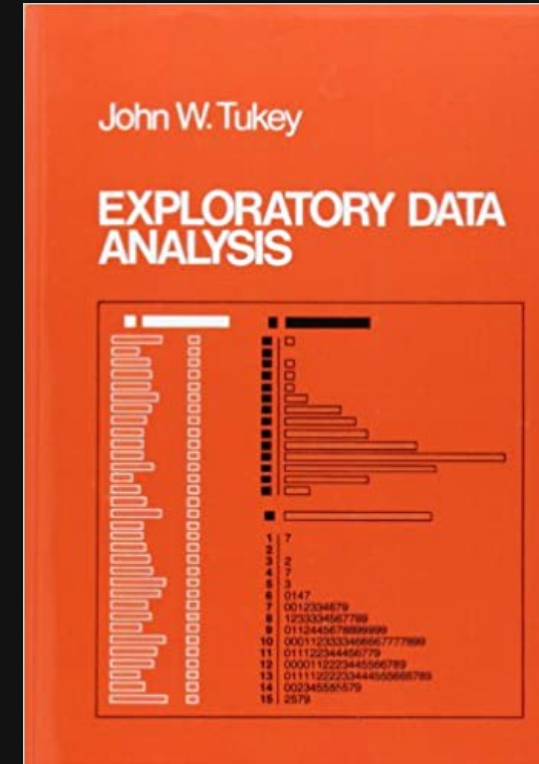
- *Mapping vs. setting aesthetics, combining layers, facets*

PART 1

Exploratory Data Analysis (EDA)

"EDA"

"Exploratory Data Visualization" first coined by American mathematician John Tukey in 1977



What is EDA?

John T. Behrens, Principles and Procedures of Exploratory Data Analysis:

Emphasis on substantive understanding of data

- i.e. "what is going on here?"

Iterative process with a focus on graphic representations of data

What is EDA?

John T. Behrens, Principles and Procedures of Exploratory Data Analysis:

- *Includes subset analyses, skepticism, and flexibility*
- *The role of the data analyst is to listen to the data in as many ways as possible until a plausible "story" of the data is apparent*

Who does EDA?

John Tukey, Exploratory Data Analysis:

A detective investigating a crime needs both tools and understanding.

If he has no fingerprint powder, he will fail to find fingerprints on most surfaces.

If he does not understand where the criminal is likely to have put his fingers, he will not look in the right places.

Equally, the analyst of data needs both tool and understanding.

EDA is a 'state of mind'

Hadley Wickham, [R for Data Science](#):

More than anything, EDA is a state of mind.

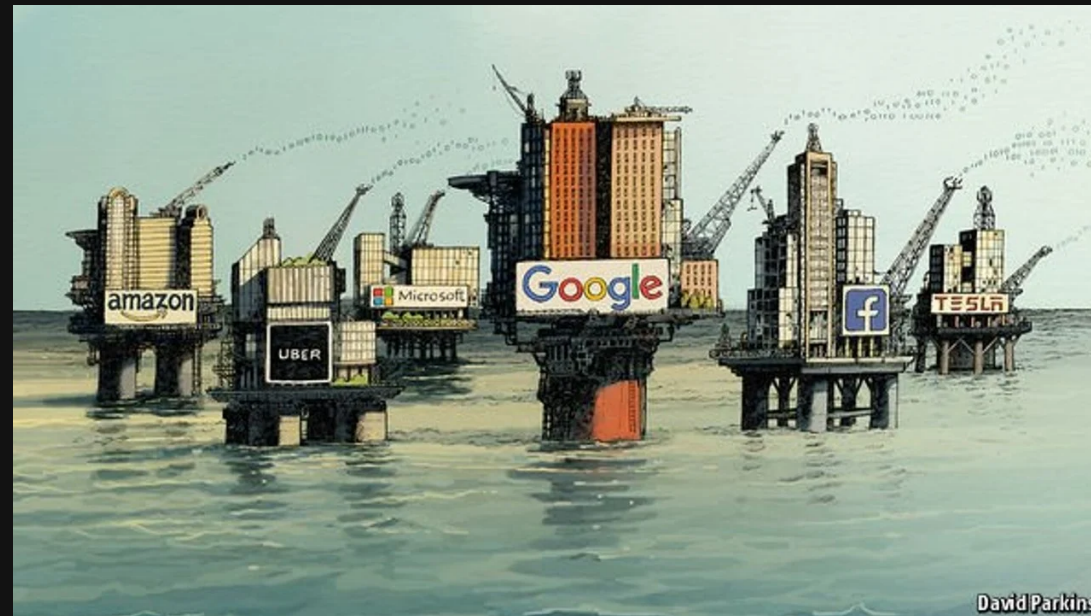
During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends.

As your exploration continues, you will home in on a few particularly productive areas that you'll eventually write up and communicate to others.

Why is EDA important?

"Data are becoming the new raw material of business" - Craig Mundie, CEO at Microsoft

"Data is the oil of the digital era" - [The Economist](https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data)



Why is EDA important?

Data are complex:



```
...Handlers.RequestHandler", "method": "handle", "requestID": "b00...",
"URL": "/app/page/analyze", "webParams": "null", "class": "com.orgmanager.handlers.RequestHandler",
"requestID": "8249868e-afd8-46ac-9745-839146a20f09", "sessionID": "14402n620jm9trnd3s3n7wg0k", "deltaStartMillis": "0",
"durationMillis": "36"}{"timestamp": "2017-06-03T18:43:33.030", "class": "com.orgmanager.handlers.RequestHandler",
"requestID": "789d89cb-bfa8-4e7d-8047-498454af885d", "sessionID": "14402n620jm9trnd3s3n7wg0k", "deltaStartMillis": "0",
"durationMillis": "7"}{"timestamp": "2017-06-03T18:46:921.000", "class": "com.orgmanager.handlers.RequestHandler",
"requestID": "7ac6ce95-19e2-4a60-88d7-6ead86e273d1", "sessionID": "14402n620jm9trnd3s3n7wg0k", "deltaStartMillis": "0",
"durationMillis": "23"}{"timestamp": "2017-06-03T18:42:18.018", "class": "com.orgmanager.handlers.RequestHandler",
"requestID": "8249868e-afd8-46ac-9745-839146a20f09", "sessionID": "14402n620jm9trnd3s3n7wg0k", "deltaStartMillis": "0",
"durationMillis": "36"}{"timestamp": "2017-06-03T18:43:33.030", "class": "com.orgmanager.handlers.RequestHandler",
"requestID": "789d89cb-bfa8-4e7d-8047-498454af885d", "sessionID": "14402n620jm9trnd3s3n7wg0k", "deltaStartMillis": "0",
"durationMillis": "7"}{"timestamp": "2017-06-03T18:46:921.000", "class": "com.orgmanager.handlers.RequestHandler",
"requestID": "7ac6ce95-19e2-4a60-88d7-6ead86e273d1", "sessionID": "14402n620jm9trnd3s3n7wg0k", "deltaStartMillis": "0",
"durationMillis": "23"}{"timestamp": "2017-06-03T18:42:18.018", "class": "com.orgmanager.handlers.RequestHandler",
"requestID": "8249868e-afd8-46ac-9745-839146a20f09", "sessionID": "14402n620jm9trnd3s3n7wg0k", "deltaStartMillis": "0",
"durationMillis": "36"}{"timestamp": "2017-06-03T18:43:33.030", "class": "com.orgmanager.handlers.RequestHandler",
"requestID": "789d89cb-bfa8-4e7d-8047-498454af885d", "sessionID": "14402n620jm9trnd3s3n7wg0k", "deltaStartMillis": "0",
"durationMillis": "7"}{"timestamp": "2017-06-03T18:46:921.000", "class": "com.orgmanager.handlers.RequestHandler",
"requestID": "7ac6ce95-19e2-4a60-88d7-6ead86e273d1", "sessionID": "14402n620jm9trnd3s3n7wg0k", "deltaStartMillis": "0",
"durationMillis": "23"}...
```

It's hard to derive insight from data in it's raw form!

EDA is a means of visualizing complexity

- *It's hard to make sense of a dataset or database with millions of rows and thousands of columns*
- *Fortunately, humans are excellent at seeing patterns:*

**frontiers**
in Neuroscience

Superior pattern processing is the essence of the evolved human brain
REVIEW article
Front. Neurosci., 22 August 2014 | <https://doi.org/10.3389/fnins.2014.00265>

[Superior pattern processing is the essence of the evolved human brain](https://doi.org/10.3389/fnins.2014.00265) - Frontiers in Neuroscience

What do you need?

Tools = R, RStudio, Adobe, sketch pad, text editor (Atom, Sublime Text, Vim)

Understanding = ...*experience and feedback*

A Bayesian Mindset

A Bayesian Mindset

*What we thought we knew (**what we expect**)*

+

*New information (**what we see**)*

=

*What we think now (**what we've learned**)*

A Bayesian Mindset

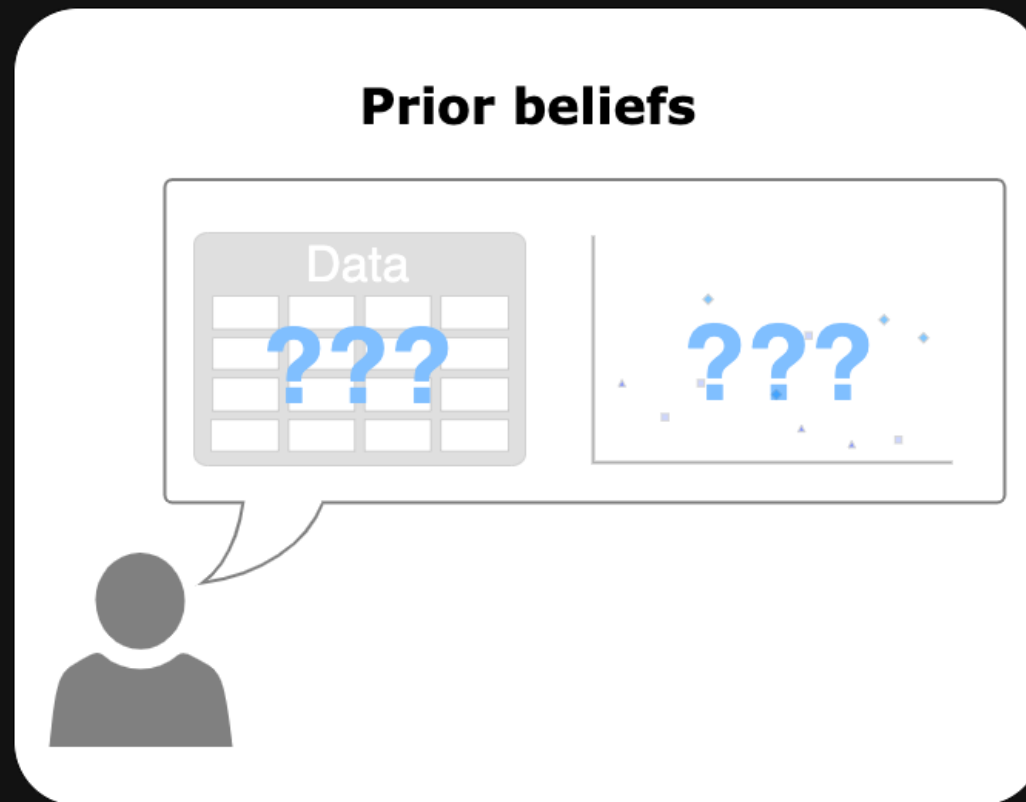
We all have implicit beliefs ('*priors*') about the world

When we encounter new data or information, our *priors* get updated

These updated beliefs ('*posteriors*') depend on our implicit beliefs and our **perceptions** of the new information

A Bayesian Mindset

Before EDA, we start with expectations and/or assumptions about the data

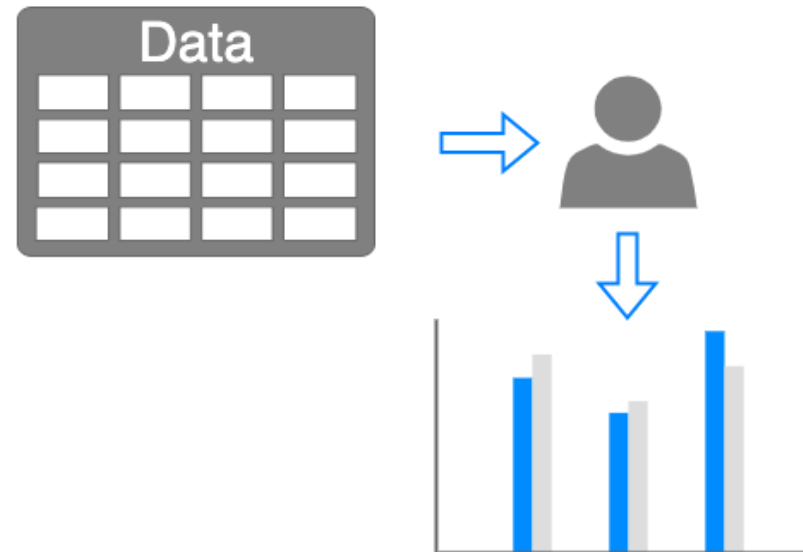


A Bayesian Mindset

During EDA, we observe new information that either confirms or contradicts our prior beliefs

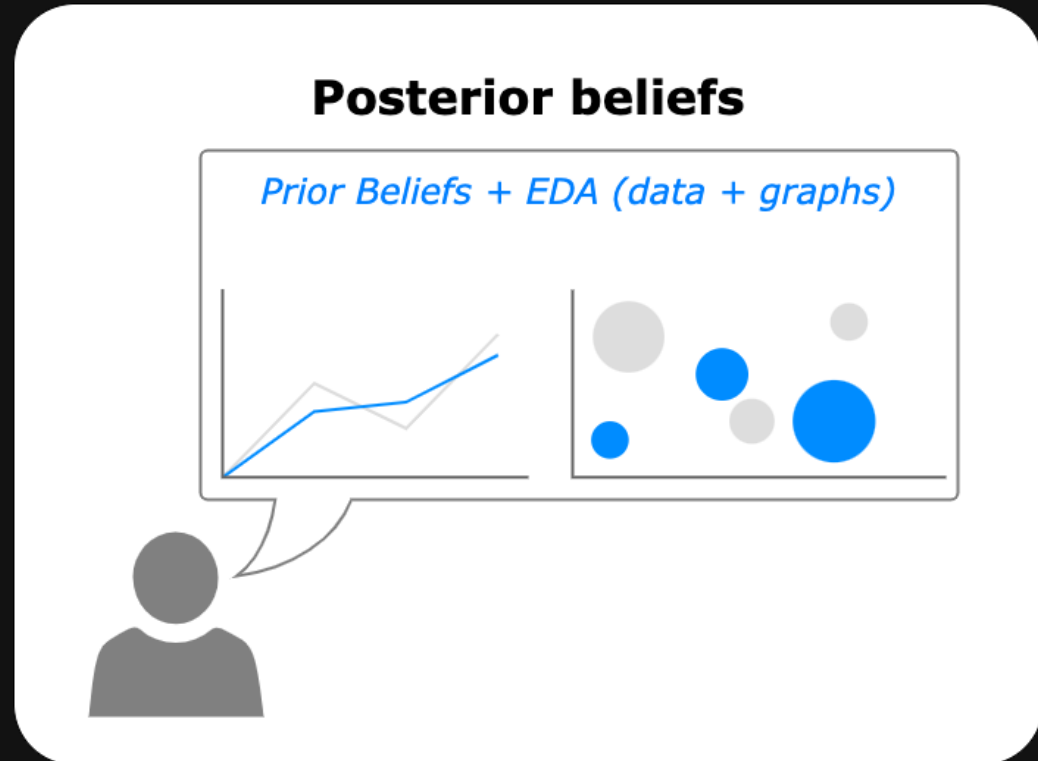
Exploratory Data Analysis

(*new information*)



A Bayesian Mindset

After EDA, we have a new set of beliefs which account for the observed data



EDA is systematic, technical creativity

The 'exploration' stems from:

- 1) articulating our prior beliefs,
- 2) having clear ideas for what we expect to see, and
- 3) accurately describing our discoveries

A Grammar Of Graphics

ggplot2: grammar & syntax

Grammar: the system of rules for any given language

Syntax: the form, structure and order for constructing statements

ggplot2: the benefits of grammar & syntax

"**objects** are like the R language's nouns, and functions (**fn**) are like verbs"

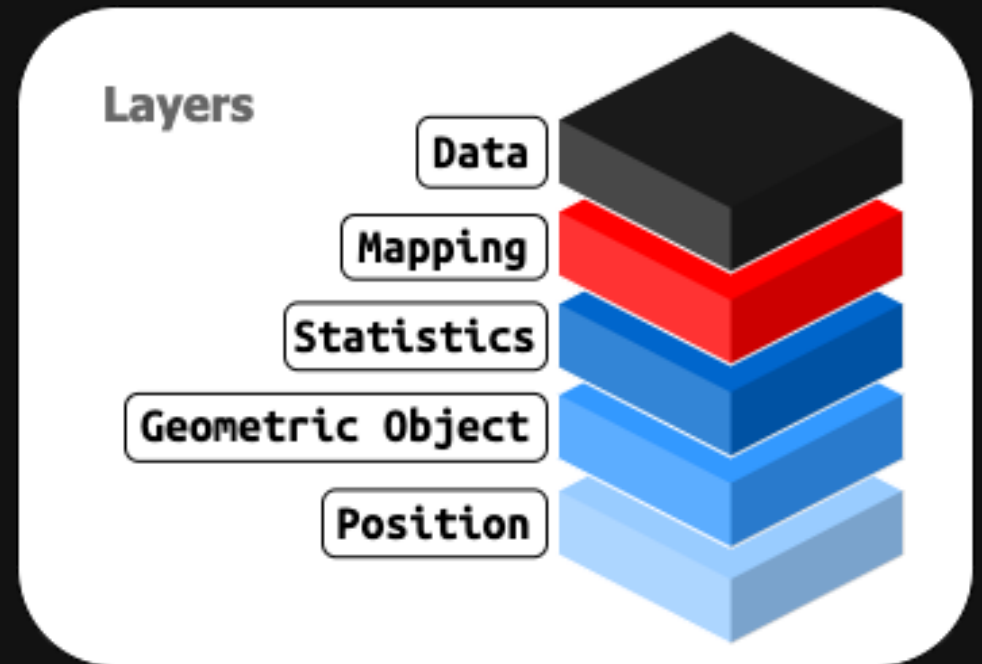


functions do things to objects

ggplot2: a layered language for graphs

ggplot2 is comprised of layers

- Data
- Mapping
- Statistics
- Geometric objects
- Position adjustments



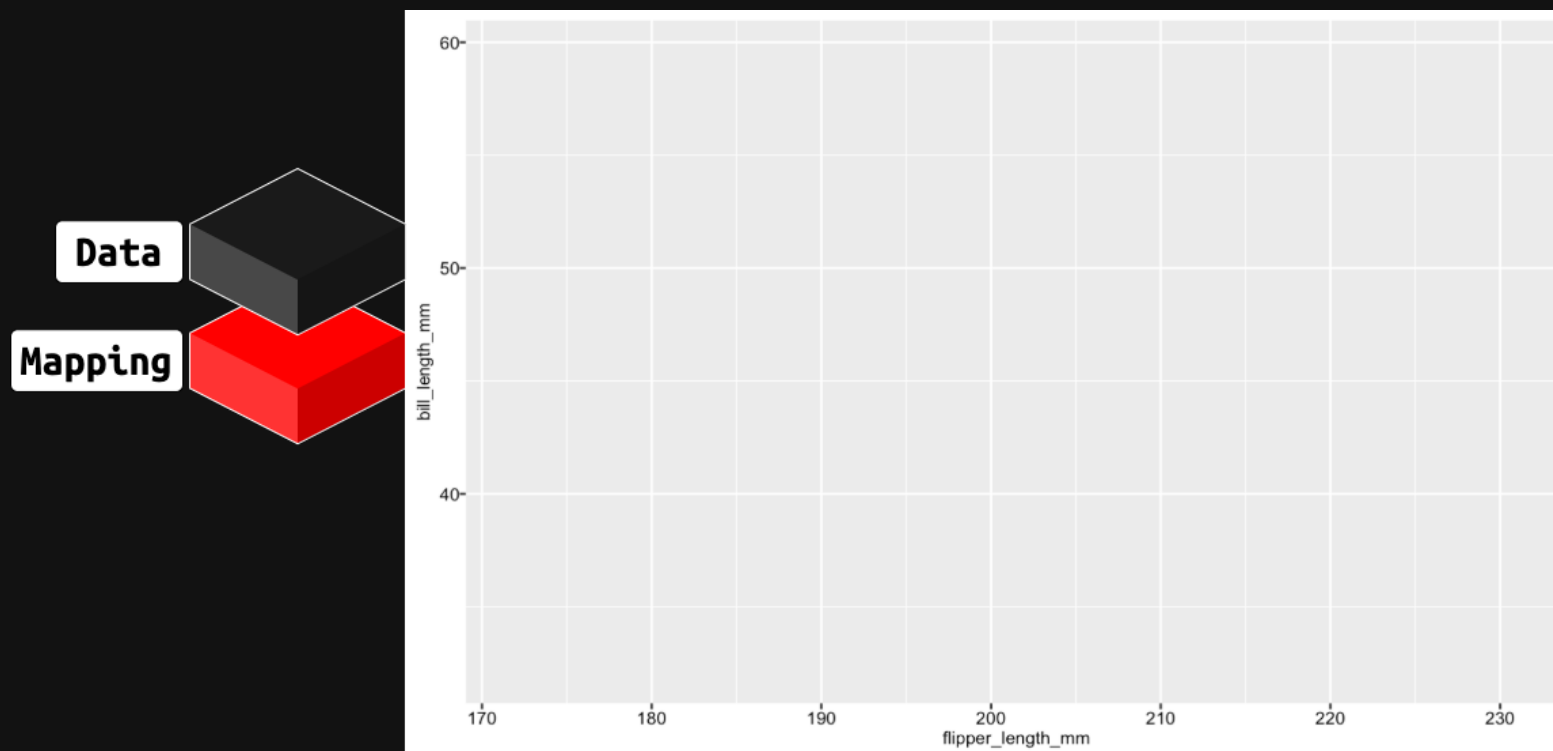
ggplot2: data

The data layer consists of a rectangular object (like a spreadsheet) with columns and rows



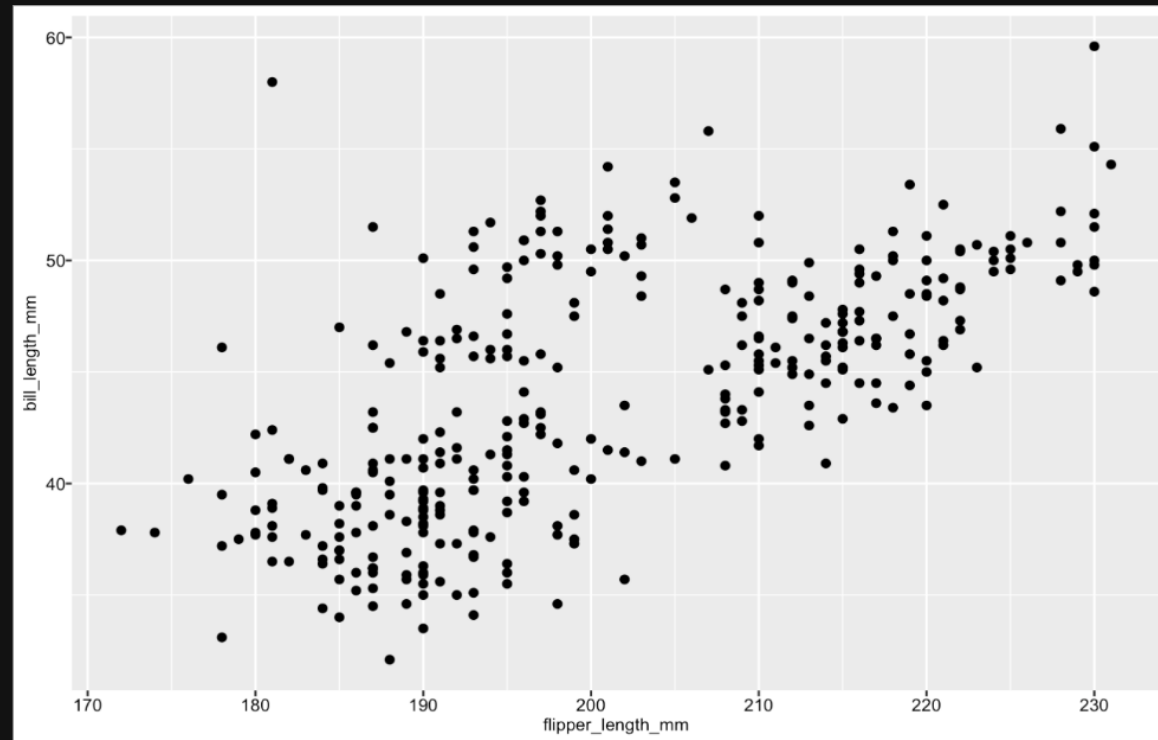
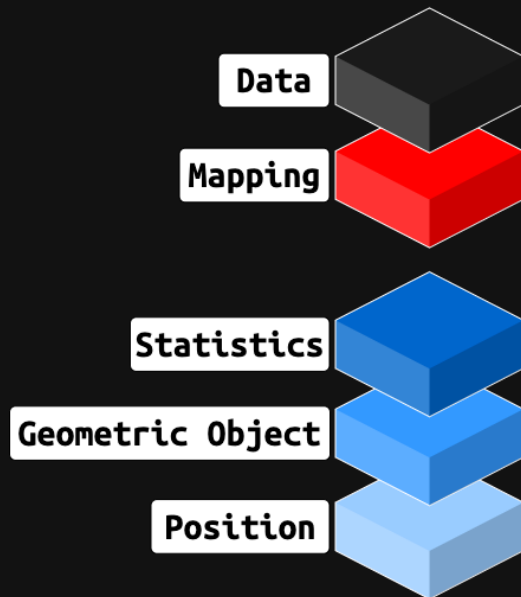
ggplot2: mapping

The mapping layer assigns columns (variables) from the data to a visual property (i.e. graph 'aesthetic')



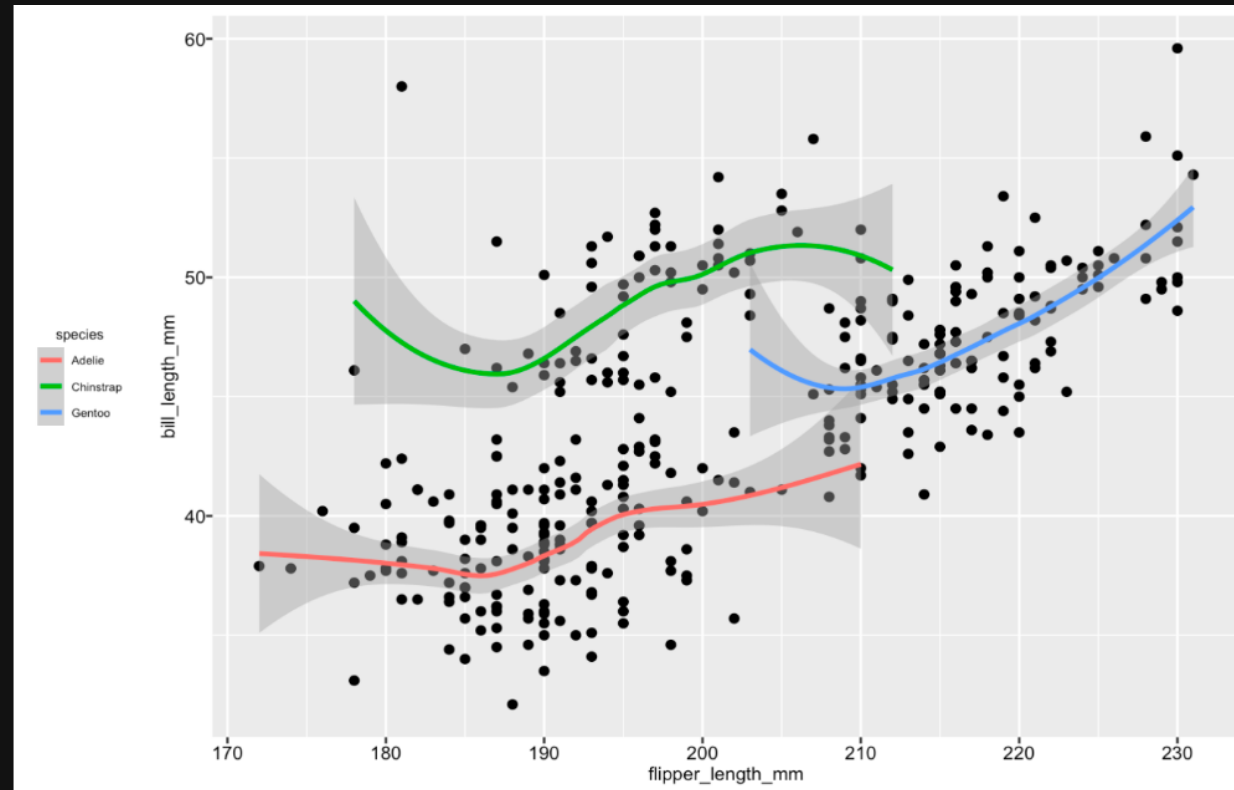
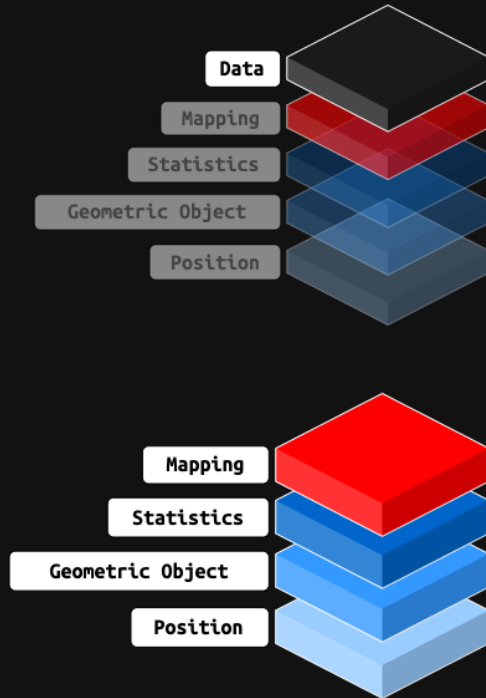
ggplot2: geoms

`geom_*()` functions include statistical transformations, shapes, and position adjustments for how to 'draw' the data on the graph



ggplot2: layers

We can have multiple layers (data, mappings, geoms) in a single graph



ggplot2: layers = infinitely extensible

Language is a system for

“making infinite use of finite means.” - [Wilhelm von Humboldt](#)

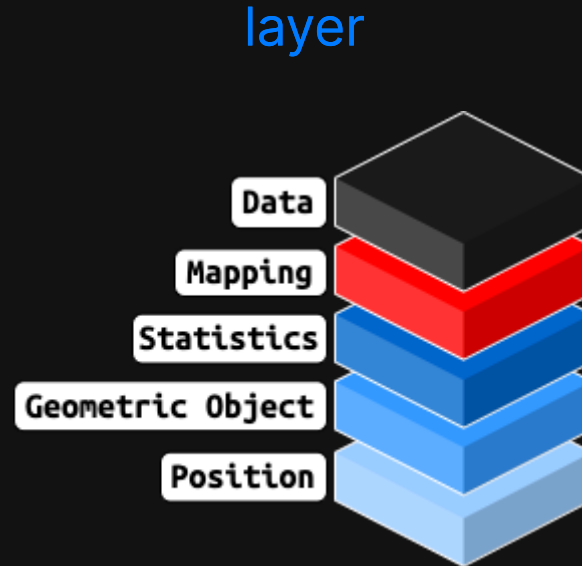
With a finite number of **objects** & **functions**, we can combine **ggplot2**s grammar and syntax to create an infinite number of graphs!

ggplot2: layers = infinitely extensible

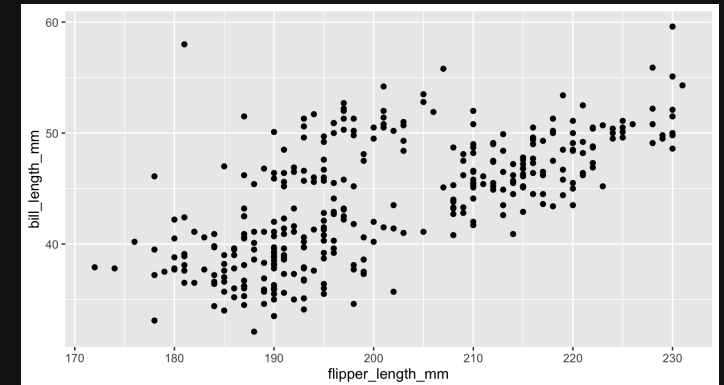
We can build graphs layer-by-layer

code

```
ggplot(data = penguins,  
  mapping = aes(x = flipper_length_mm,  
    y = bill_length_mm)) +  
  geom_point()
```



graph

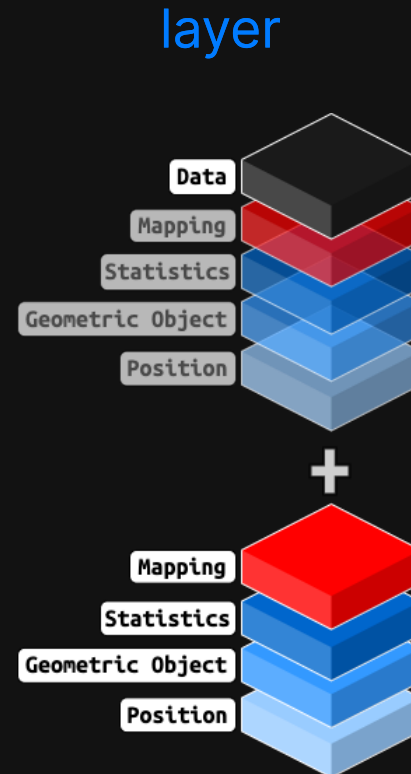


ggplot2: layers = infinitely extensible

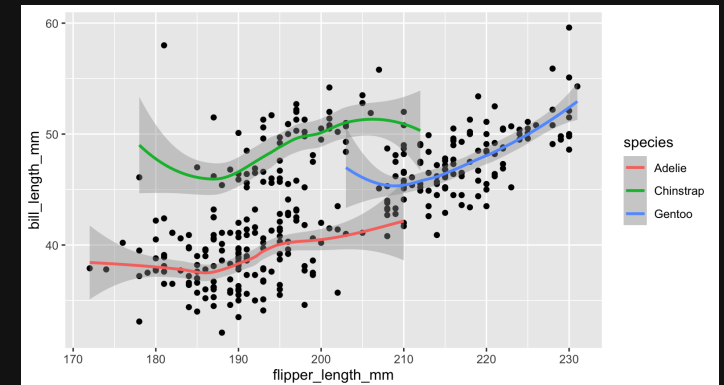
New layers can 'inherit' data from previous layers (or include their own data)

code

```
ggplot(data = penguins,  
       mapping = aes(x = flipper_length_mm,  
                     y = bill_length_mm)) +  
  geom_point() +  
  geom_smooth(  
    mapping = aes(x = flipper_length_mm,  
                  y = bill_length_mm,  
                  color = species))
```



graph



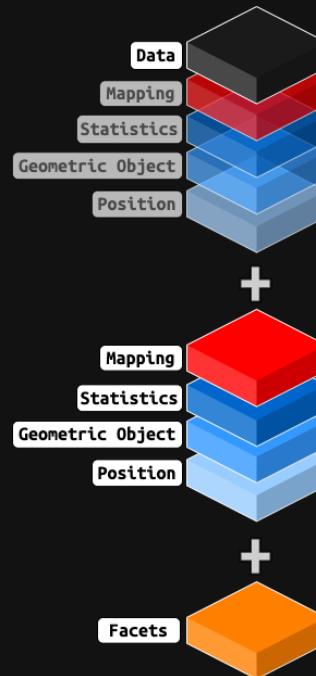
ggplot2: layers = infinitely extensible

Additional functions for facets, themes, etc.

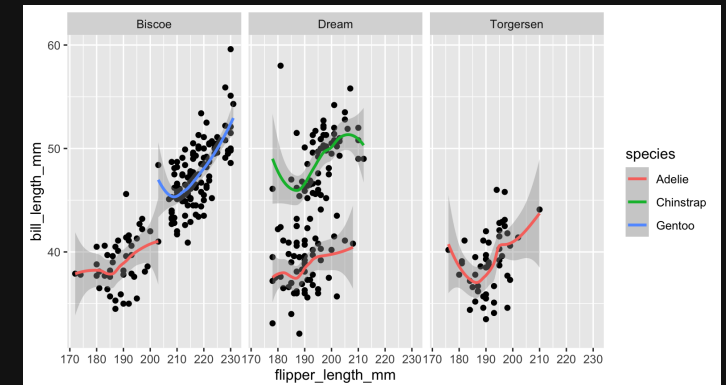
code

```
ggplot(data = penguins,  
       mapping = aes(x = flipper_length_mm,  
                     y = bill_length_mm)) +  
  geom_point() +  
  geom_smooth(  
    mapping = aes(x = flipper_length_mm,  
                  y = bill_length_mm,  
                  color = species)) +  
  facet_wrap(facets = . ~ island)
```

layer



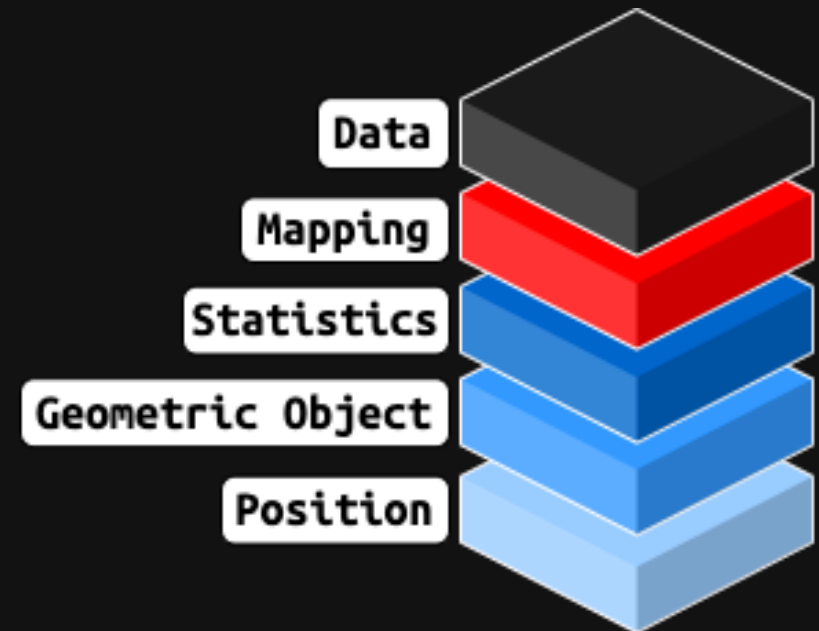
graph



ggplot2: templates

Basic Template: Data, aesthetic mappings, geom

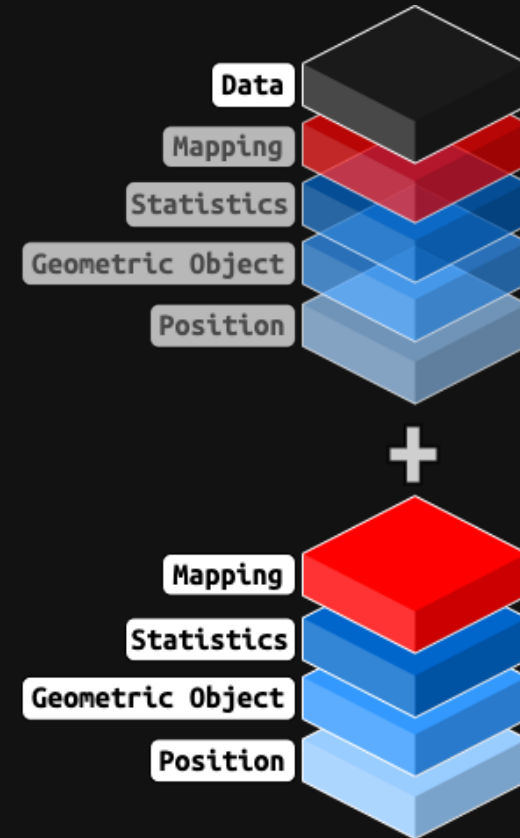
```
ggplot(data = <DATA>) +  
  geom_*(mapping = aes(<AESTHETIC MAPPINGS>))
```



ggplot2: templates

Template + 1 Layer: More geoms and aesthetic mappings

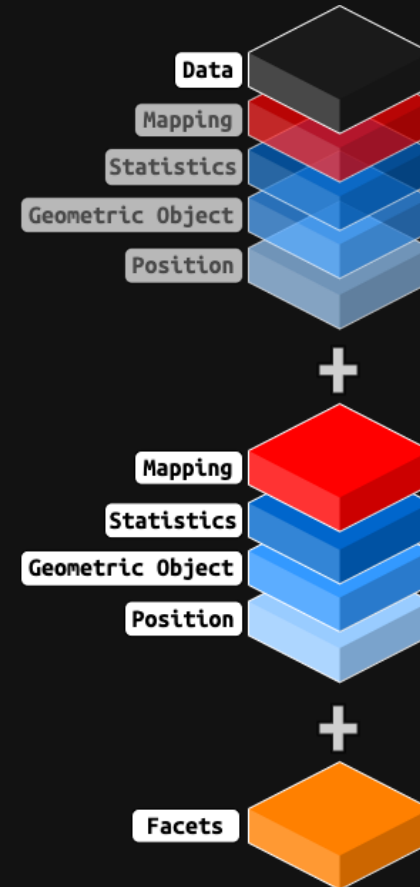
```
ggplot(data = <DATA>) +  
  geom_(mapping = aes(<AESTHETIC MAPPINGS>)) +  
  geom_(mapping = aes(<AESTHETIC MAPPINGS>))
```



ggplot2: templates

Template + 2 Layers: Faceting

```
ggplot(data = <DATA>) +  
  geom_(mapping = aes(<AESTHETIC MAPPINGS>)) +  
  geom_(mapping = aes(<AESTHETIC MAPPINGS>)) +  
  facet_*
```



templates = infinitely extensible!

Themes

Don't forget labels!

```
ggplot(data = <DATA>) +  
  geom_(mapping = aes(<AESTHETIC MAPPINGS>)) +  
  geom_(mapping = aes(<AESTHETIC MAPPINGS>)) +  
  facet_* +  
  theme_*
```

```
ggplot(data = <DATA>) +  
  geom_(mapping = aes(<AESTHETIC MAPPINGS>)) +  
  geom_(mapping = aes(<AESTHETIC MAPPINGS>)) +  
  facet_* +  
  theme_* +  
  <LABELS>
```

Next up: Part 2!

[@mjfrigaard](#) 

[@mjfrigaard](#) 

mjfrigaard@pm.e 

[What does "λέξις" mean?](#)