ODSC: Data Visualization with ggplot2

Part 1: Thinking with graphs

https://bit.ly/odsc-ggplot2

by Martin Frigaard & Peter Spangler

Written: February 08 2022

Updated: April 21 2022

Resources

Links:

- Conference Website
- Website
- Part 1
- Part 2

Materials:

- RStudio.Cloud
- Github Repo

https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

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*

https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

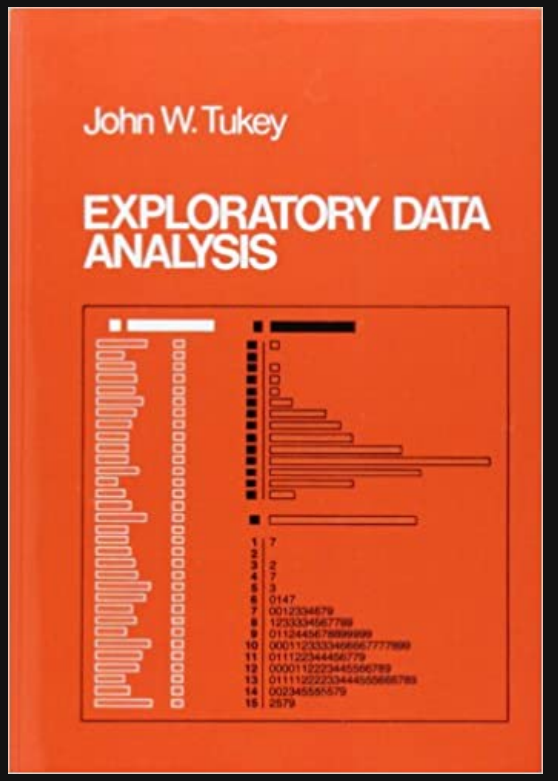
PART 1

https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

Exploratory Data Analysis (EDA)

"EDA"

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



https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

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:

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

https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

EDA is a means of visualizing complexity

ODSC
The Future of AI Gathers Here

- *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:*

Superior pattern processing is the essence of the evolved human brain - Frontiers in Neuroscience

What do you need?

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

Understanding = *...experience and feedback*

https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

A Bayesian Mindset

A Bayesian Mindset

The Future of AI Gathers Here

What we thought we knew (what we expect)

+

New information (what we see)

=

What we think now (what we've learned)

https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

A Bayesian Mindset

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

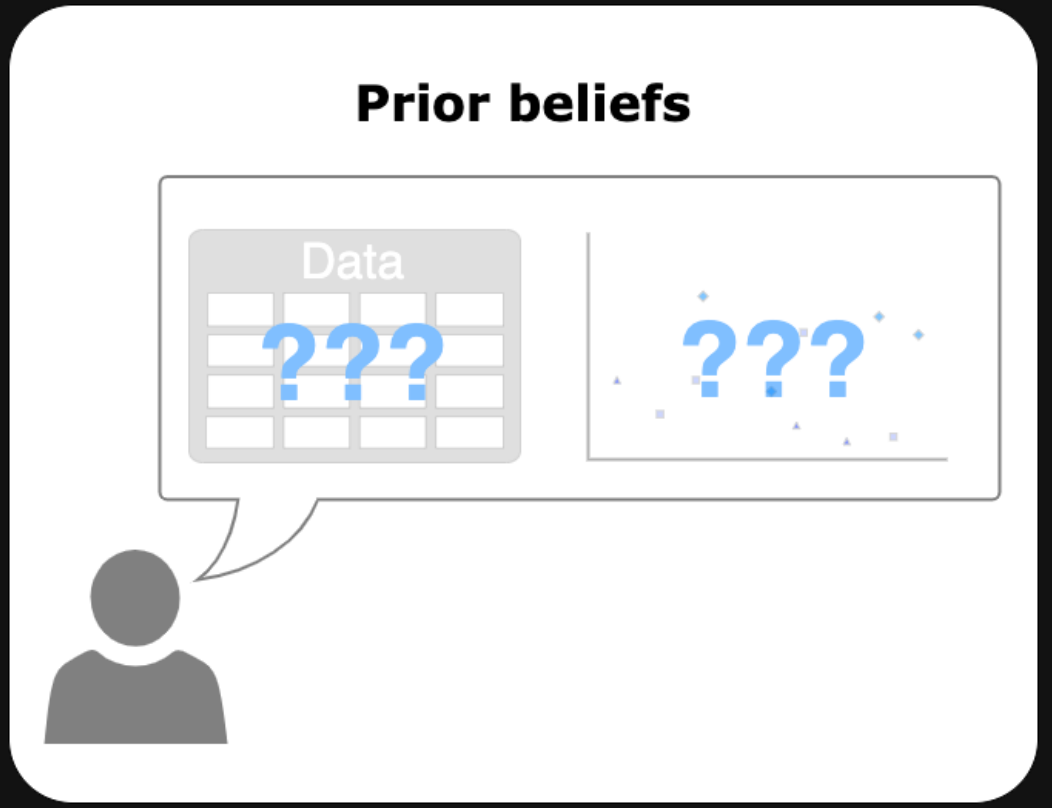
When we encounter new data or information, our *priors* get updated

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

https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

A Bayesian Mindset

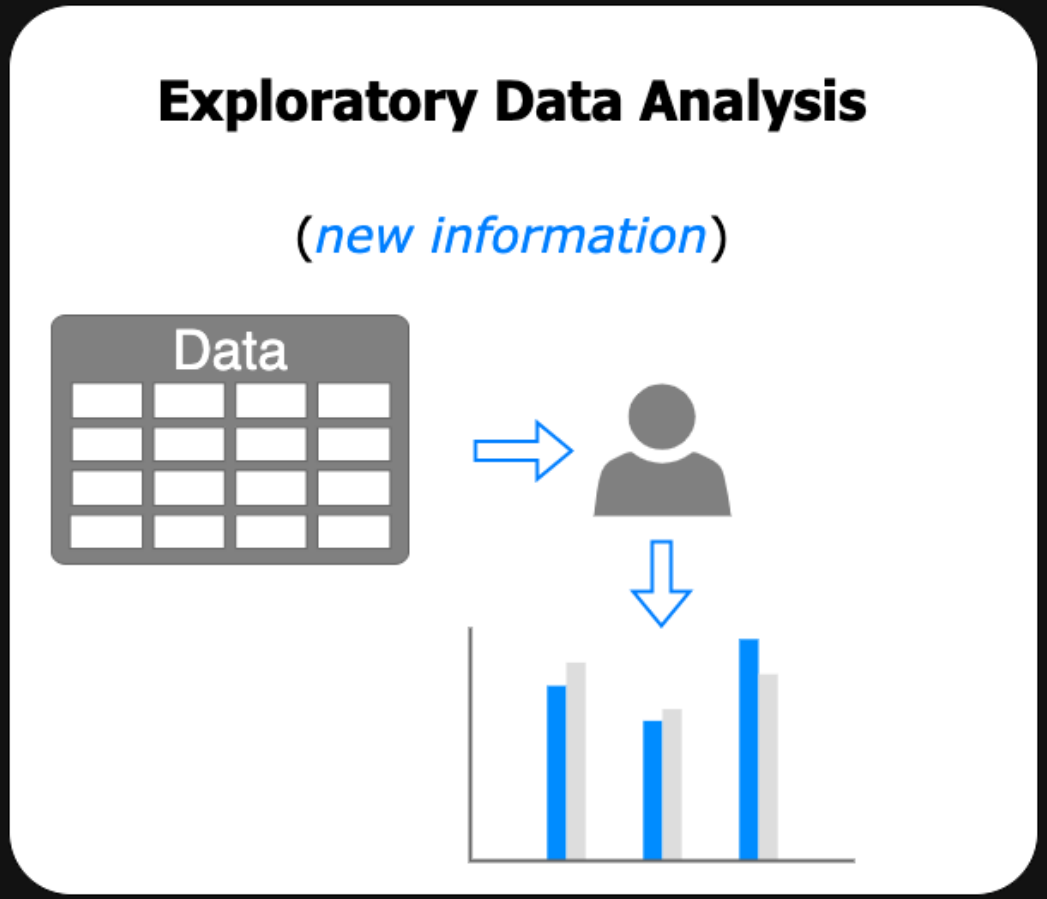
Before EDA, we start with expectations and/or assumptions about the data



https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

A Bayesian Mindset

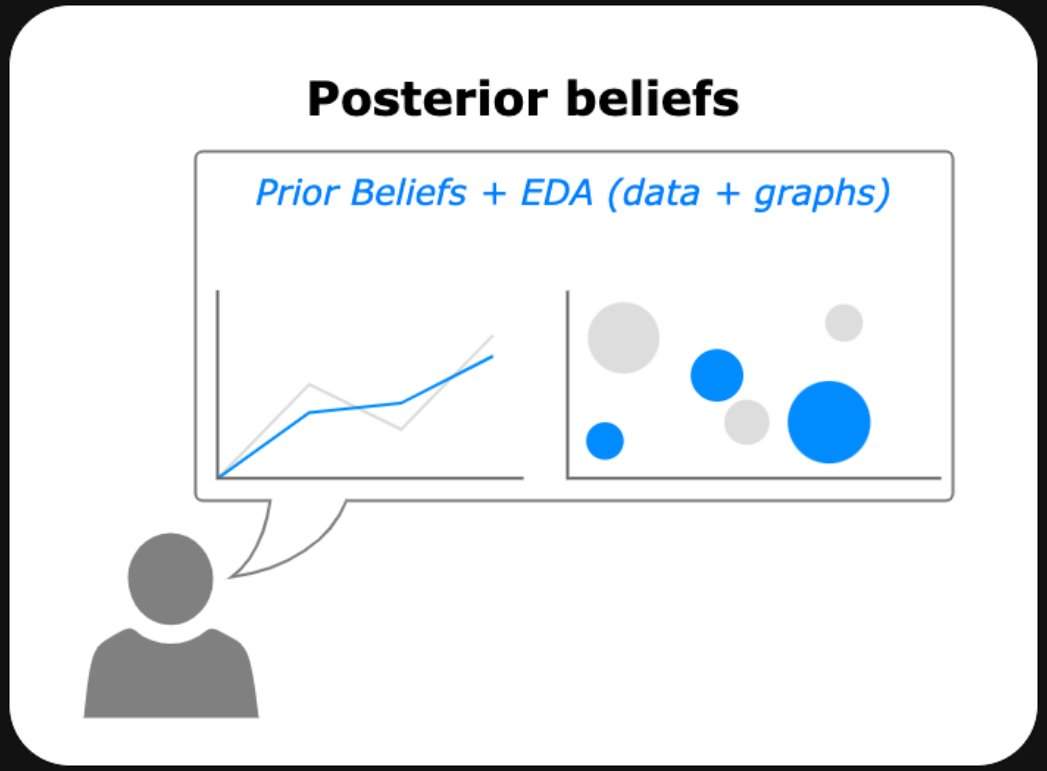
During EDA, we observe new information that either confirms or contradicts our prior beliefs



https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

A Bayesian Mindset

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



https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

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

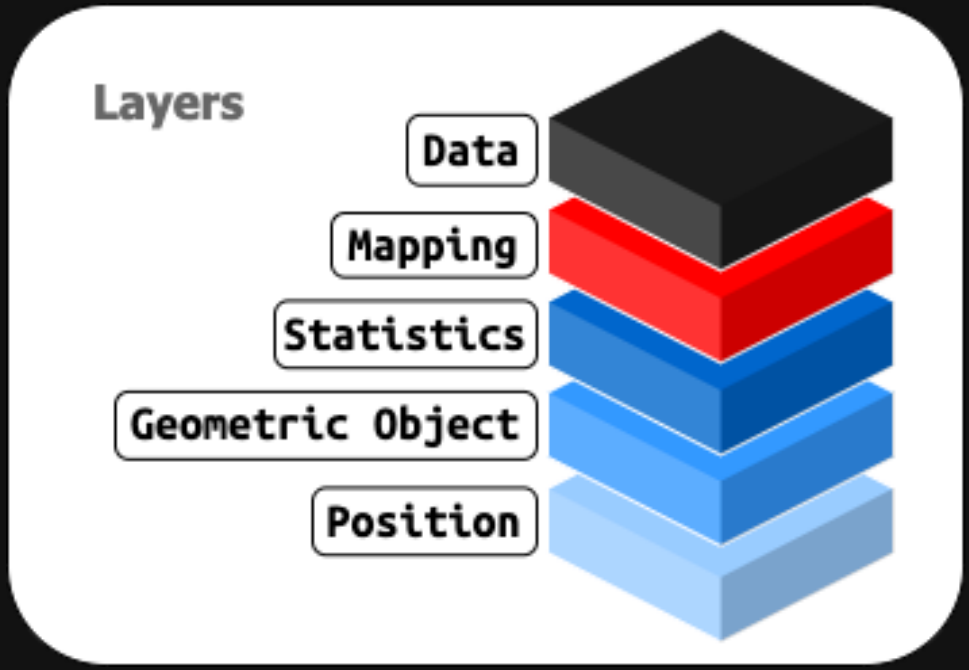
https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

ggplot2: a layered language for graphs

ODSC The Future of AI Gathers Here

ggplot2 is comprised of layers

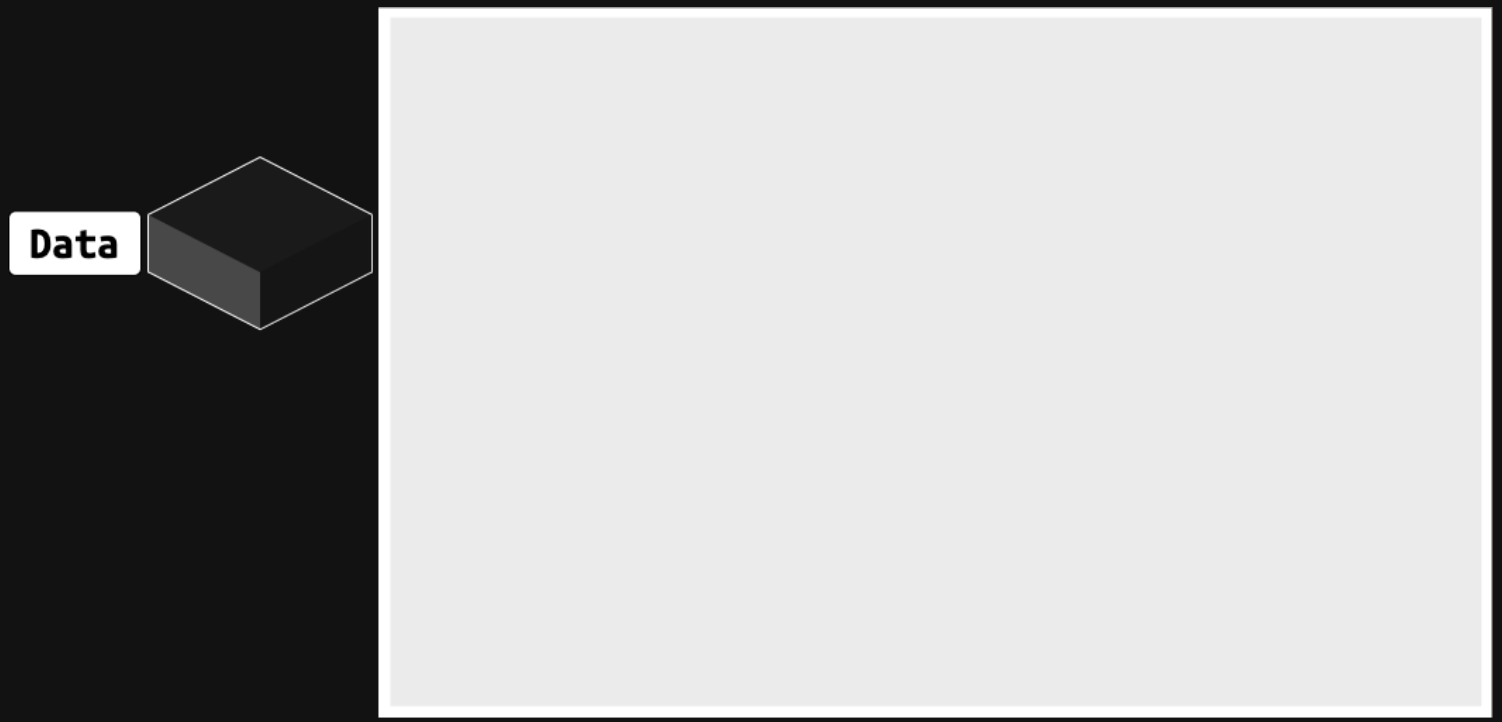
- Data
- Mapping
- Statistics
- Geometric objects
- Position adjustments



https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

ggplot2: data

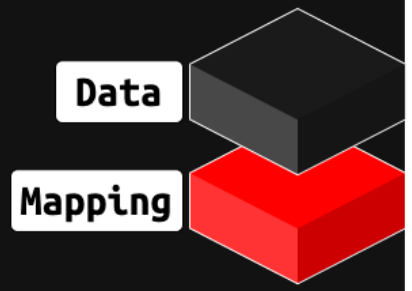
The data layer consists of a rectangular object (like a spreadsheet) with columns and rows

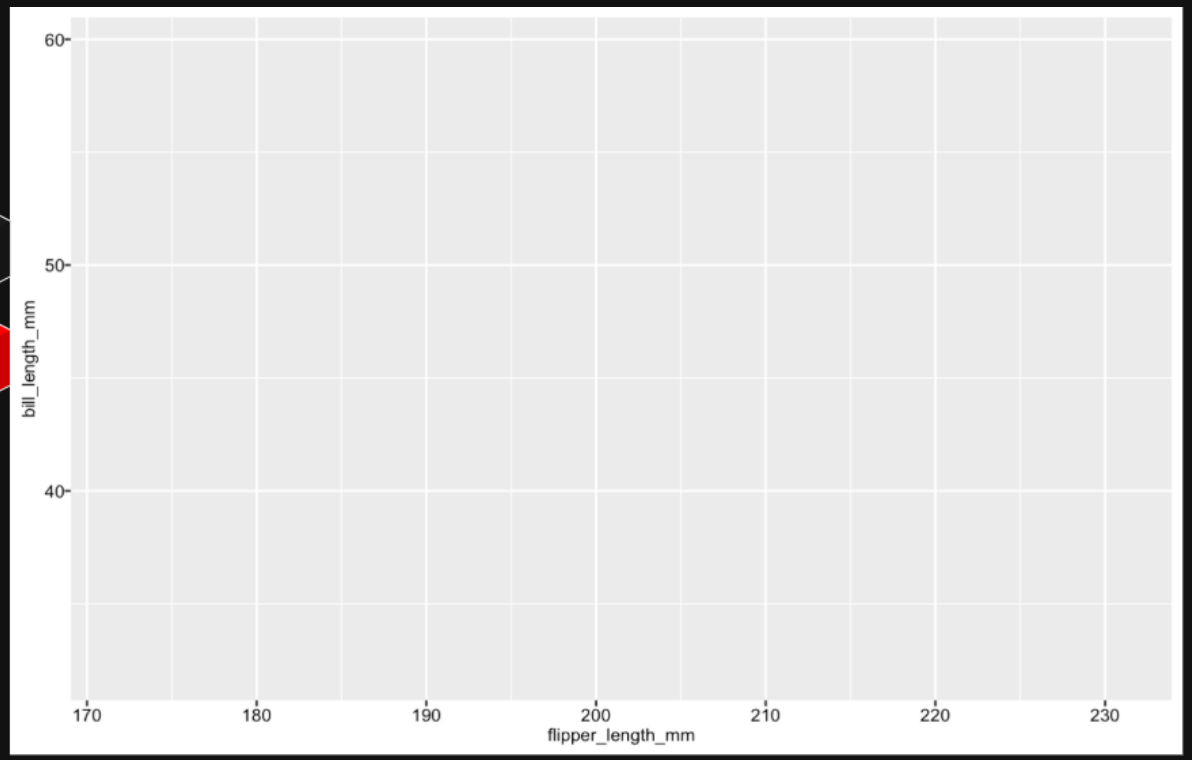


https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

ggplot2: mapping

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



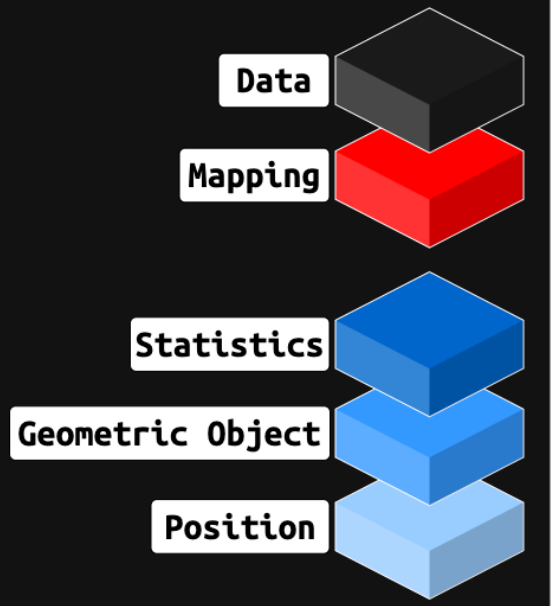


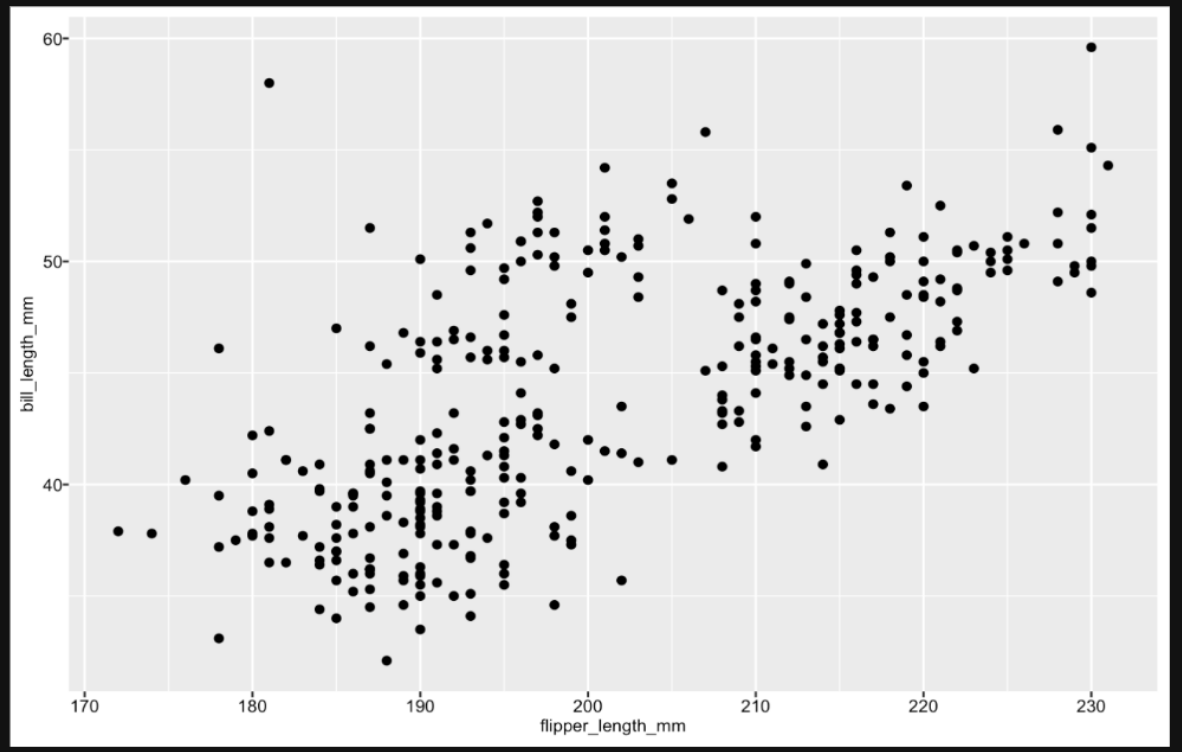
https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

ggplot2: geoms

The Future of AI Gathers Here

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

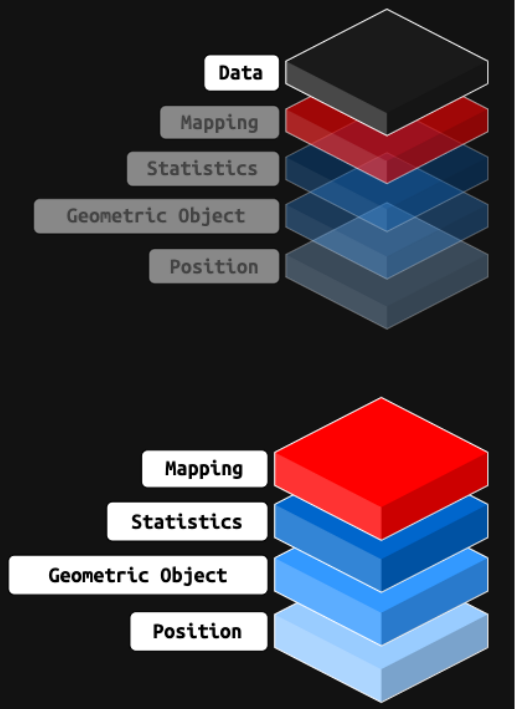


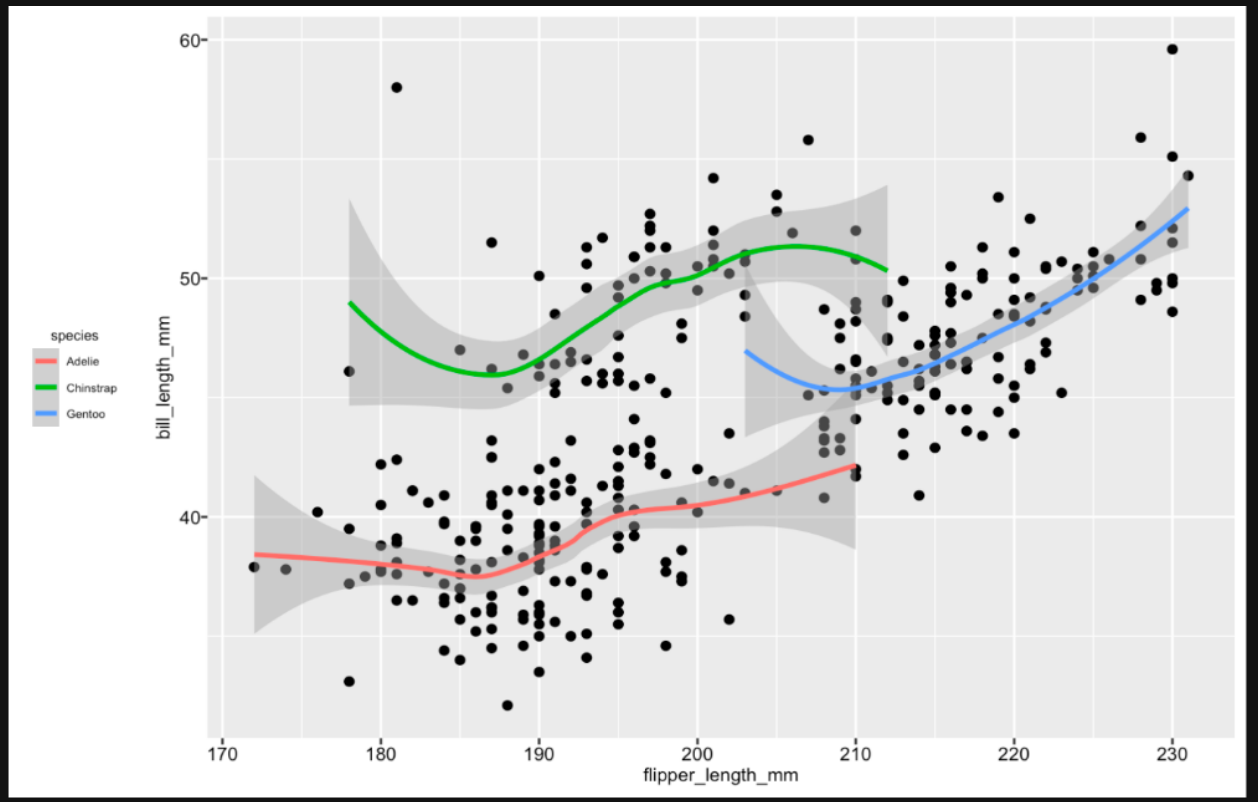


https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

ggplot2: layers

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





https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

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 ggplot2s grammar and syntax to create an infinite number of graphs!

ggplot2: layers = infinitely extensible

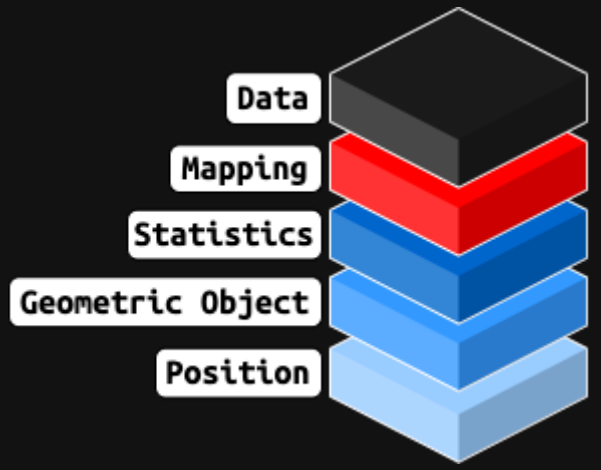
The Future of AI Gathers Here

We can build graphs layer-by-layer

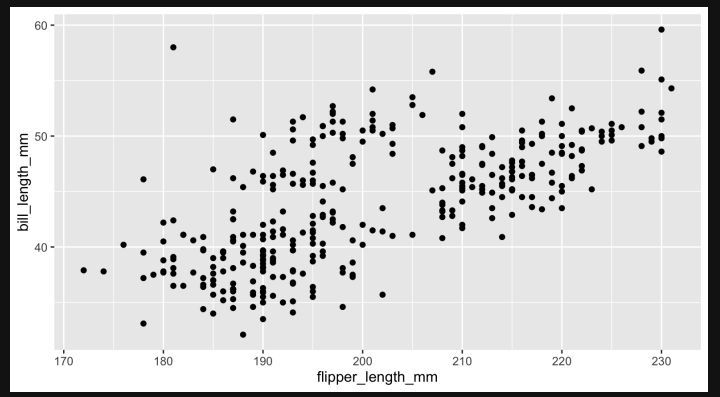
code

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

layer



graph



https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

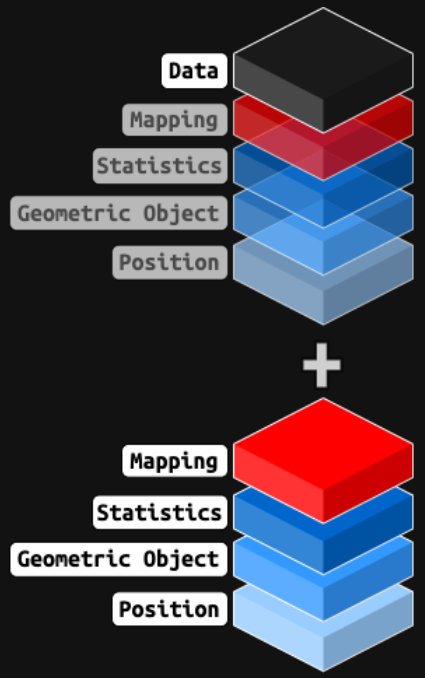
ggplot2: layers = infinitely extensible

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

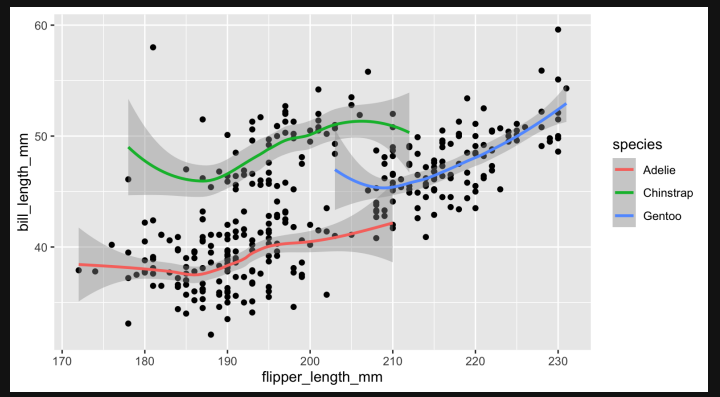
code

```
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))
```

layer



graph



https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

ggplot2: layers = infinitely extensible

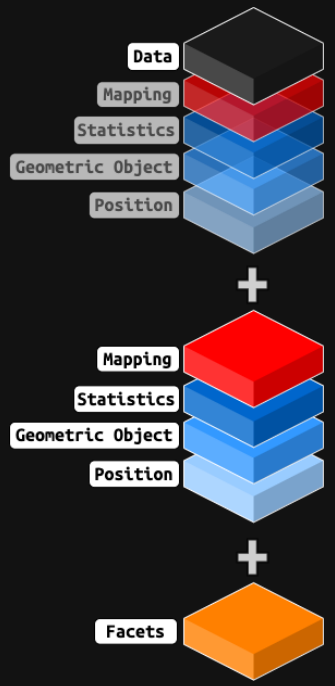
The Future of AI Gathers Here

Additional functions for facets, themes, etc.

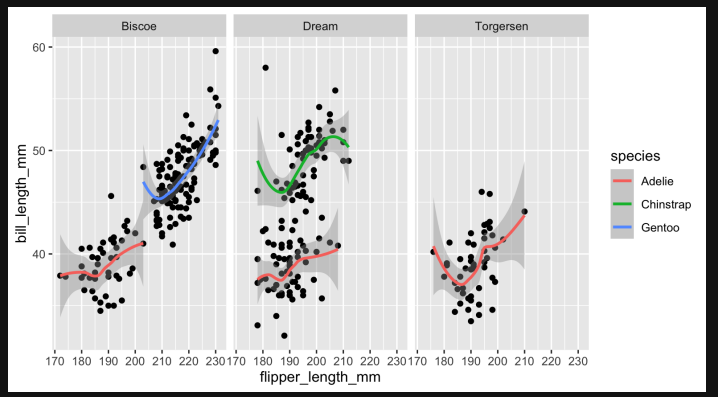
code

```
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



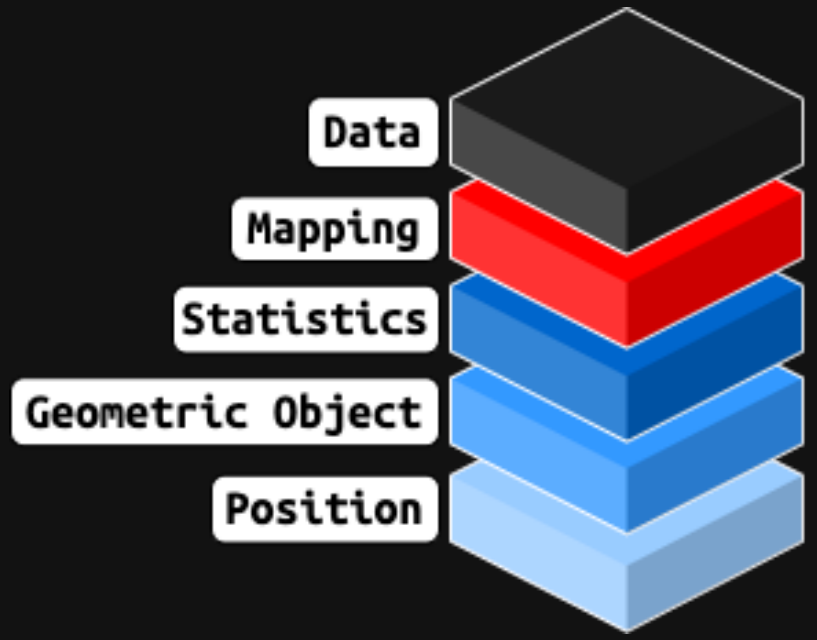
https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

ggplot2: templates

The Future of AI Gathers Here

Basic Template: Data, aesthetic mappings, geom

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



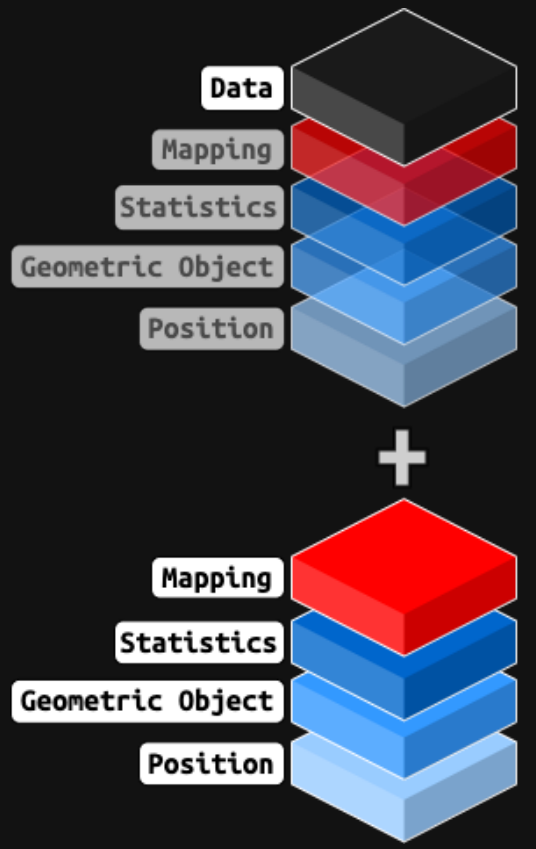
https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

ggplot2: templates

The Future of AI Gathers Here

Template + 1 Layer: More geoms and aesthetic mappings

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



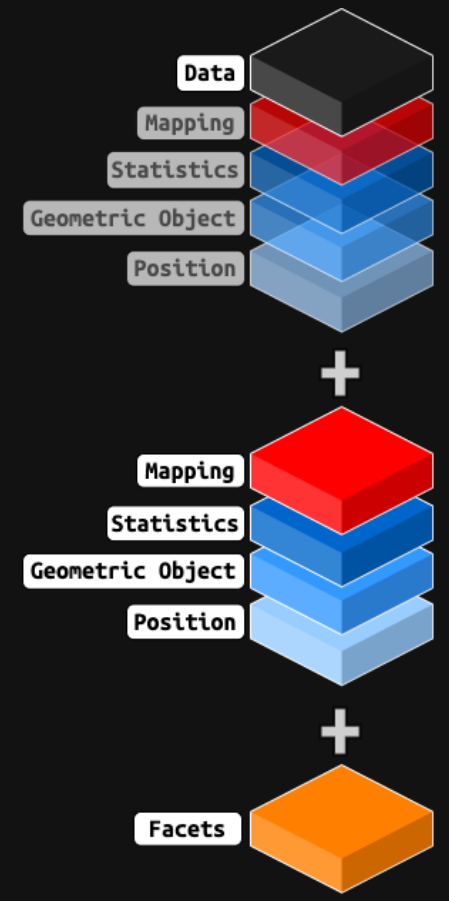
https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

ggplot2: templates

The Future of AI Gathers Here

Template + 2 Layers: Faceting

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



https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

templates = infinitely extensible!

The Future of AI Gathers Here

Themes

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

Don't forget labels!

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


https://github.com/mjfrigaard/talks/tree/main/odsc-eda-2022-04

ODSC
The Future of AI Gathers Here

Next up: Part 2!

@mjfrigaard
@mjfrigaard
mjfrigaard@pm.e
What does "λέξις" mean?