R Fundamentals and Best Practices for Mass Spectrometry Data Analysis

Sunday, November 8 (12:00-3:15pm Eastern)

Kylie Bemis, Northeastern University

Module #3: Basic data cisualization with ggplot2

Module #4: Data preparation for visualization & extended ggplot2

Data Visualization

Kylie A. Bemis

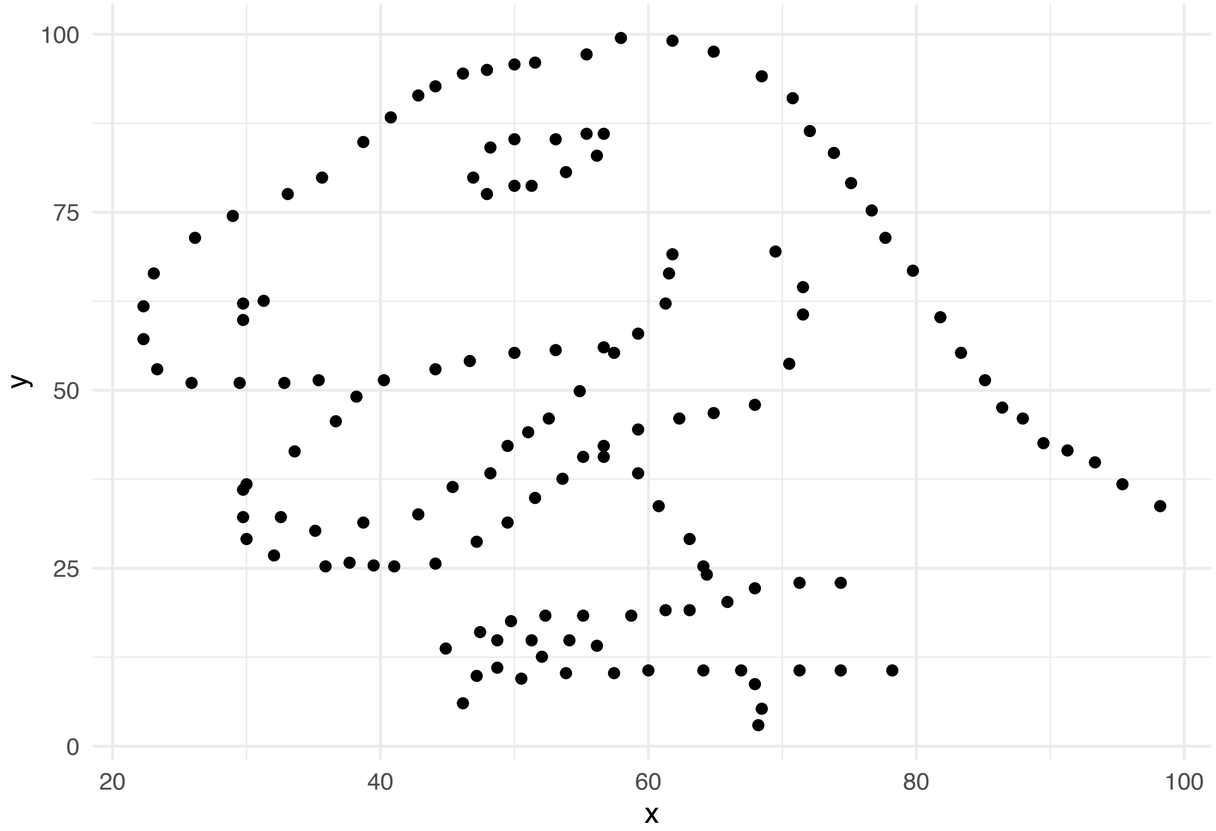
Northeastern University
Khoury College of Computer Sciences

Northeastern University

STATISTICAL GRAPHICS: WHY WE LOOK AT DATA

Why do we look at data?

Dataset #1:



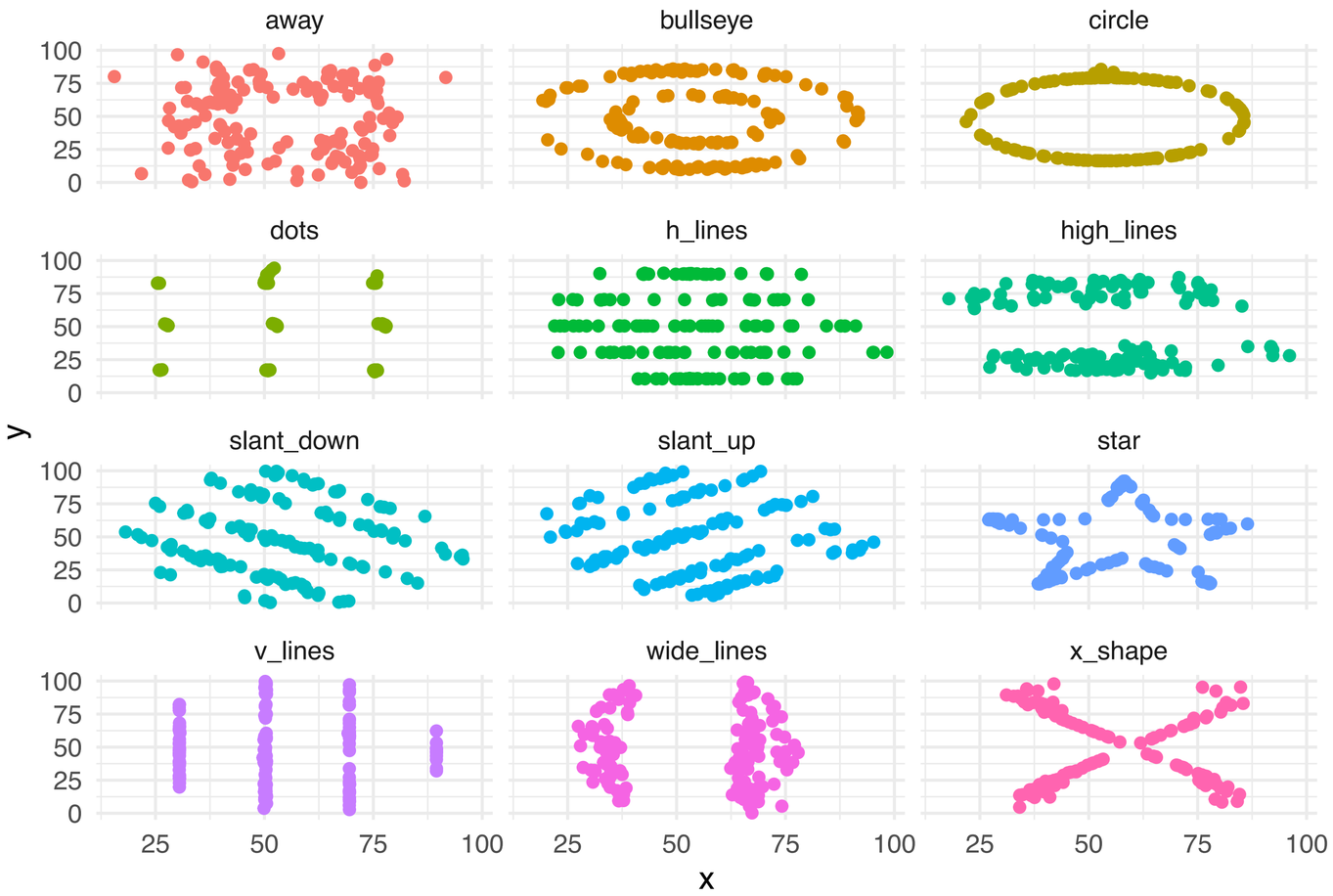
How similar are the other 12 datasets?

13 datasets with same statistics:

X mean	54.26
Y mean	47.83
X std. dev	16.76
Y std. dev	26.93
Correlation	-0.06

Why we look at data

Datasets #2-13:



X mean	54.26
Y mean	47.83
X std. dev	16.76
Y std. dev	26.93
Correlation	-0.06

"The Datasaurus Dozen": https://www.autodeskresearch.com/publications/samestats

Looking at data is important

- Summary statistics don't tell whole story
- Easily find patterns
- Identify potential outliers
- Check model assumptions
- Intuitively display results

What are some common ways of looking at data?

Some common statistical graphics

- Scatter plot
- Line plot
- Box-and-whisker plot
- Histogram
- Bar plot

Roles of statistical graphics

One variable

- Histogram
- Bar plot
- Box plot
- Pie chart

Two or more variables

- Scatter plot
- Line plot
- Box plot
- Faceting

Roles of statistical graphics

Distributions

- Histogram
- Bar plot
- Box plot
- Pie chart

Relationships

- Scatter plot
- Line plot
- Box plot
- Faceting

A GRAMMAR OF GRAPHICS

How do we plot data?

- By using the "name" of graphic?
  - Scatter plot
  - Box plot
  - Histogram
- Using "base" R (and similar approaches):
  - `plot()` - scatter plot
  - `boxplot()` - box plot
  - `hist()` - histogram

Is there a better way?

What are some common ingredients of statistical graphics?

Recipes for common statistical graphics

- Scatter plot
  - Maps variables to x- and y- axes
  - Uses points to represent observations
- Line plot
  - Maps variables to x- and y- axes
  - Uses lines to connect observations
- Box plot
  - Maps 5-number summary to x- or y-axis
  - Uses boxes and whispers to show this
- Histogram
  - Maps bins to x-axis and binned counts to y-axis
  - Uses bars+length to represent counts
- Bar plot
  - Maps categories to x-axis and counts to y-axis
  - Uses bars+length to represent counts
- Pie chart
  - Maps categories to color and proportion to angle
  - Uses wedges+area to represent proportions

Key ingredients for statistical graphics

- Some kind of data
- Encodings from data to aspects of the plot
  - Marks ("**geometric objects**", e.g., points and lines)
  - Channels ("**aesthetics**", e.g., color and size)
- Statistical transformations
- Coordinate system
- Scales and annotations

Building a plot

Consider a simple dataset:

A	*B*	*C*	*D*
2	3	4	a
1	2	1	a
4	5	15	b
9	10	80	b

http://vita.had.co.nz/papers/layered-grammar.pdf

How do we create a scatter plot of A versus C?

What if we want to include D?

Building a plot

We map the x-axis to **A**, the y-axis to **C**, and shape to **D**

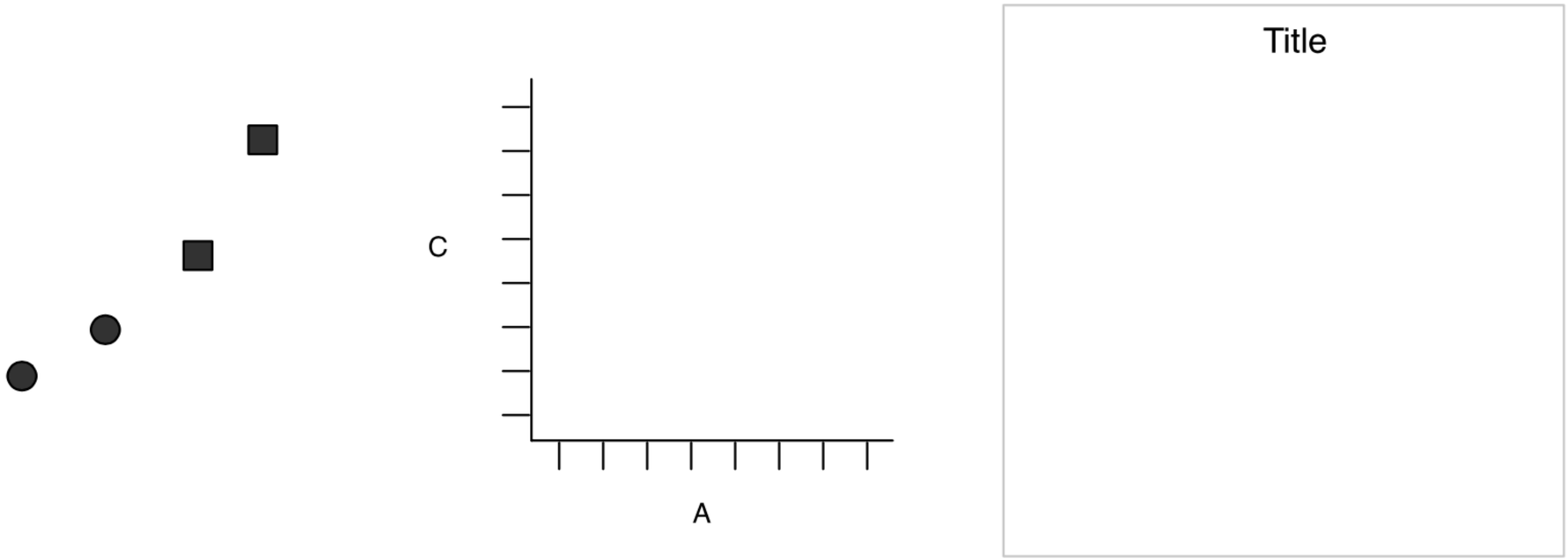
x	*y*	Shape
2	4	circle
1	1	circle
4	15	square
9	80	square

http://vita.had.co.nz/papers/layered-grammar.pdf

This is the "aesthetic mapping" of the plot

Building a plot

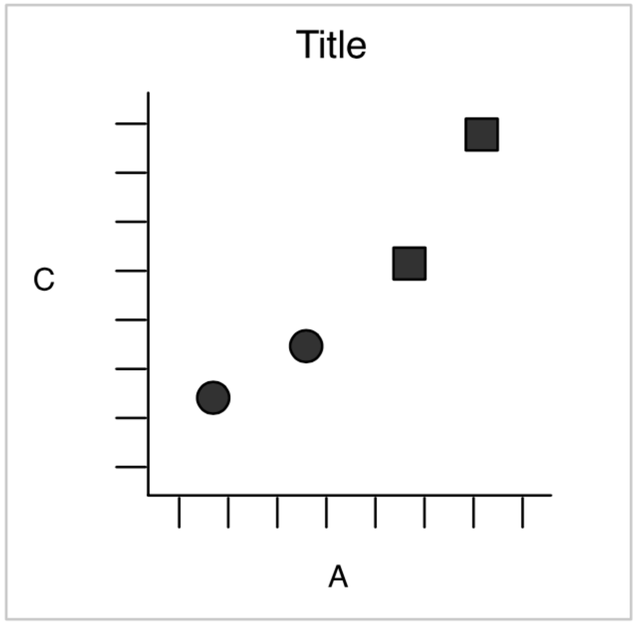
We have (1) marks or **geometric objects**, (2) **scales** and a **coordinate system**, and (3) plot **annotations**



http://vita.had.co.nz/papers/layered-grammar.pdf

Building a plot

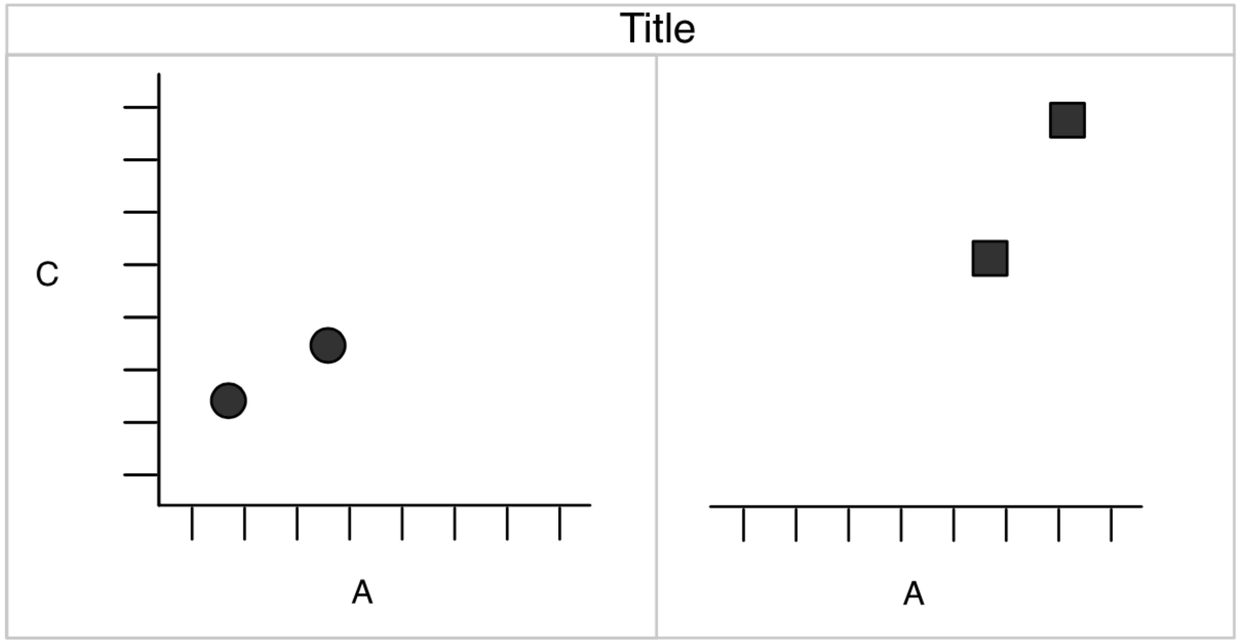
Putting the ingredients together, we have a plot:



http://vita.had.co.nz/papers/layered-grammar.pdf

Building a plot

If we want to compare the relationship between **A** and **C** for each level of **D**, we can facet on **D**



http://vita.had.co.nz/papers/layered-grammar.pdf

Faceting splits the data into subsets and creates sub-plots for each subset

Building a vocabulary

We can build more complicated plots by adding to our vocabulary:

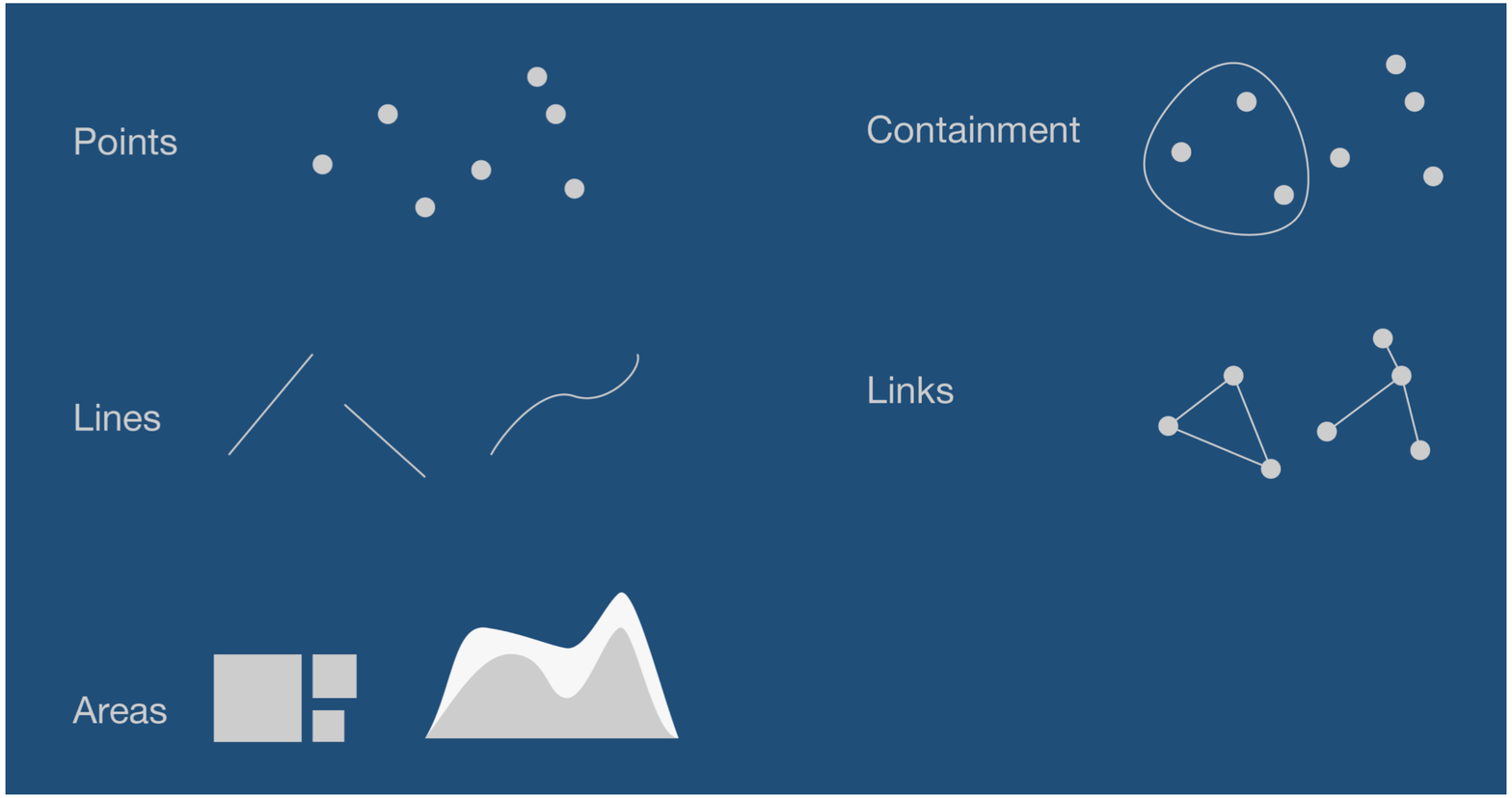
- Layers to overlay plots on top of each other
- Multiple datasets on the same plot
- Apply statistical transformations (e.g., binning)
- Apply position adjustments (e.g., jitter)
- *A way to build such plots programmatically*

A layered grammar of graphics

- Default dataset
- Default set of mappings from variables to graphical aesthetics
- One or more layers, each having:
  - Mark, or geometric object
  - Statistical transformation
  - Position adjustment
  - (Optionally) new dataset
  - (Optionally) new set of aesthetic mappings
- Scale for each mapped aesthetic
- Facet specification

Visual encodings: Marks

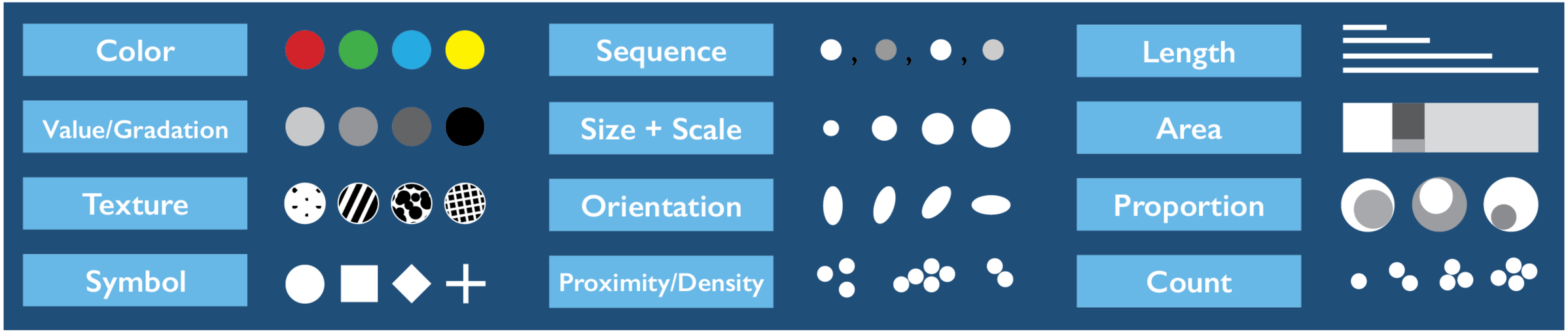
Marks, or **geometric objects**, draw the graphic



Courtesy of Steven Braun, CAMD Art + Design

Visual encodings: Channels

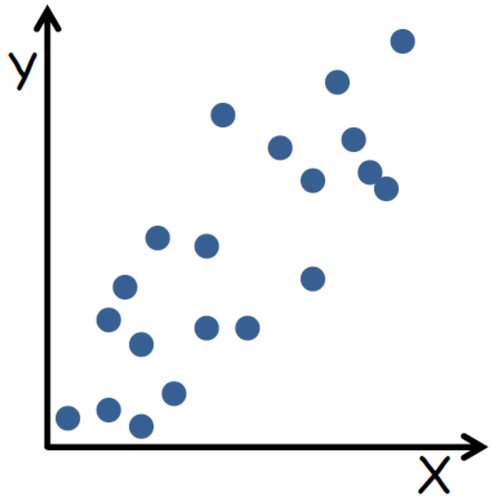
Channels, or **aesthetics** + **scales**, encode values



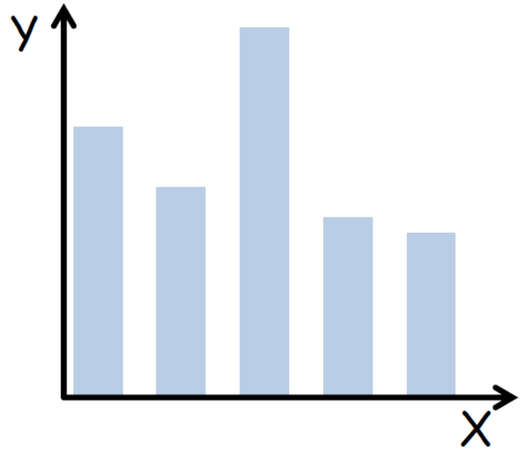
Courtesy of Steven Braun, CAMD Art + Design

Choosing visual encodings

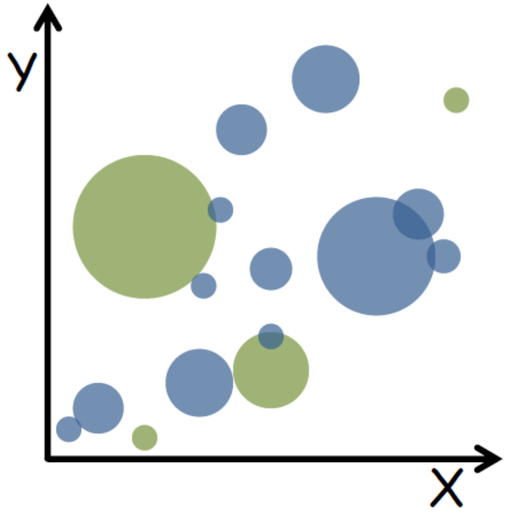
Your choice of **geometric objects,**
aesthetic mappings, and **scales** define the graphic



Marks: points
Channels: position



Marks: lines
Channels: length, position



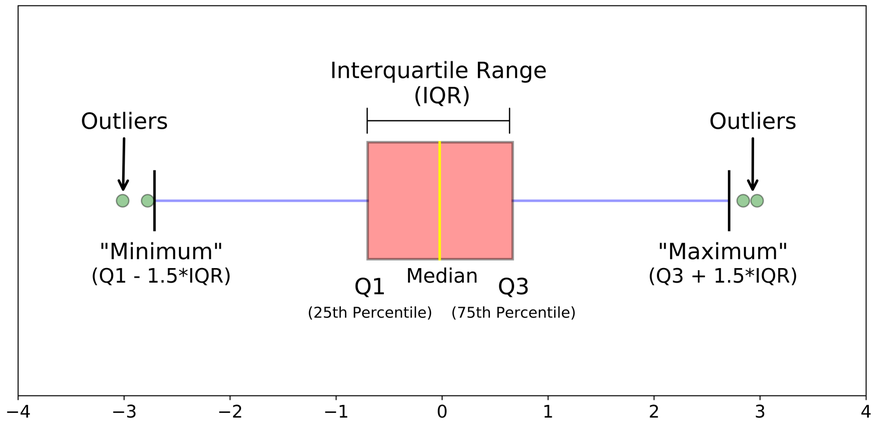
Marks: points
Channels: position, size, color

Courtesy of Steven Braun, CAMD Art + Design

Statistical transformations

Many statistical graphics utilize **statistical transformations**:

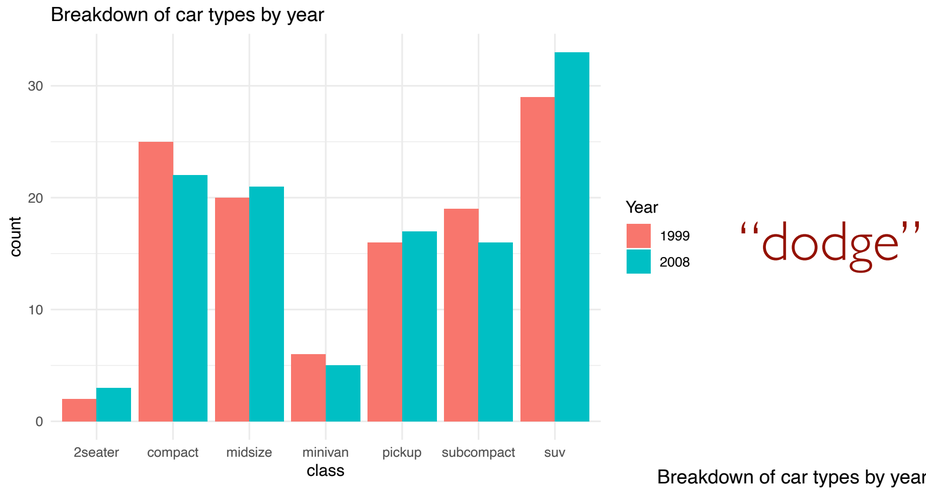
- Box plot
  - Five-number summary + outliers
- Histogram
  - Binning
- Bar plot
  - Counting



Position adjustments

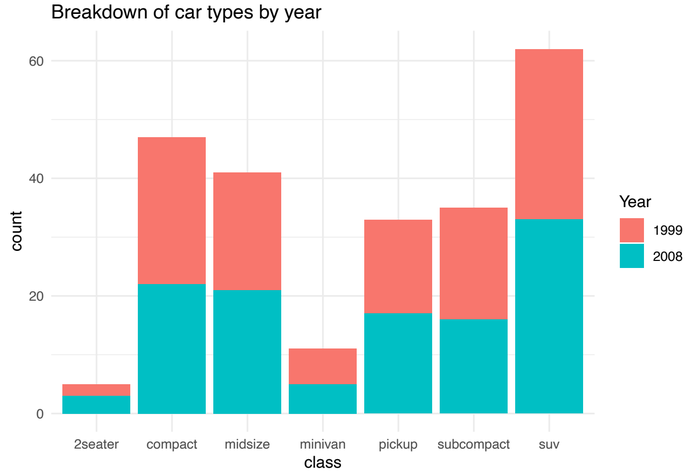
Many statistical graphics require **position adjustments**

- Scatter plot
  - Jitter
- Bar plot
  - Dodge
  - Stack



"dodge"

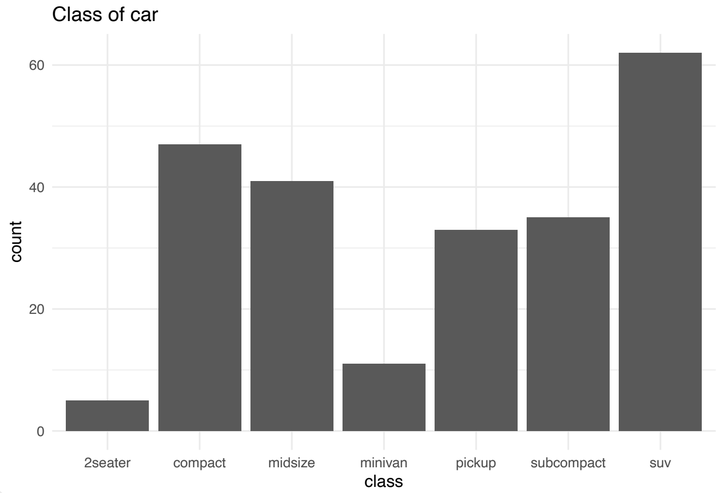
"stack"

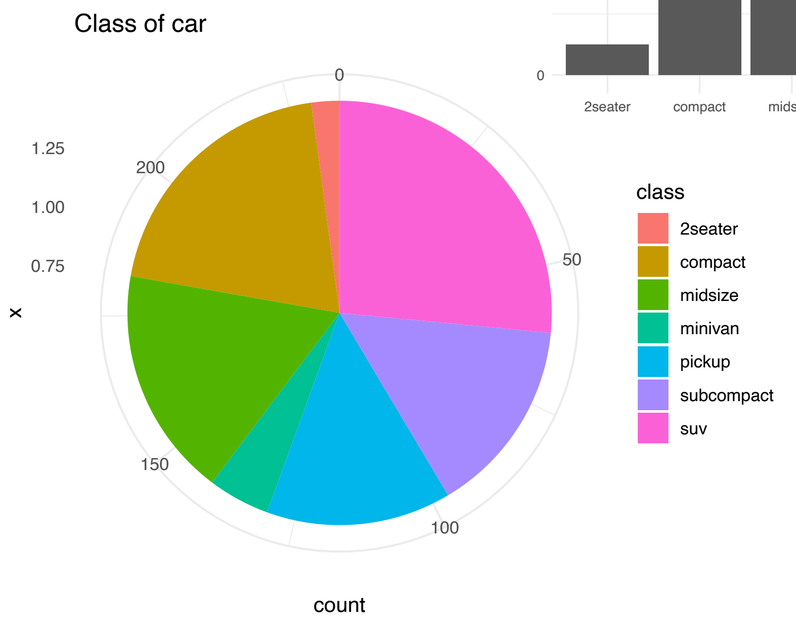


Coordinate systems

Some graphics may require different **coordinate systems**

- Cartesian
- Polar
- Map





Implementing a grammar of graphics

A version of the "layered grammar of graphics" is implemented in **ggplot2**:


```
ggplot(data = <DATASET>,
       mapping = aes(<MAPPINGS>)) +
  layer(geom = <GEOM>,
        stat = <STAT>,
        position = <POSITION>) +
  <SCALE_FUNCTION>() +
  <COORDINATE_FUNCTION>() +
  <FACET_FUNCTION>()
```

or, more simply:

```
ggplot(data = <DATASET>,
       mapping = aes(<MAPPINGS>)) +
  <GEOM_FUNCTION>()
```

Recipes for common statistical graphics

- Scatter plot
  - Geom = "point"
  - Stat = "identity"
- Line plot
  - Geom = "line"
  - Stat = "identity"
- Box plot
  - Geom = "boxplot"
  - Stat = "boxplot"
- Histogram
  - Geom = "bar"
  - Stat = "bin"
- Bar plot
  - Geom = "bar"
  - Stat = "count"
- Pie chart
  - Geom = "bar"
  - Stat = "count"

+ coords = "polar"

GGPLOT2

Thank you to the instructors and to the teaching assistants!

Ryan Benz
Meena Choi
Niyati Chopra
Miguel Cosenza
Matthias Fahrner
Amanda Figueroa-Navedo
Melanie Foell
Omkar Reddy Gojala
Dan Guo
Shubhanshu Gupta
Ting Huang
Maanasa Kaza
Smit Anish Kiri
Devon Kohler
Sai Srikanth Lakkimsetty
Danielle LaMay

Ajeya Makanahalli Kempegowda
Yogesh Nizzer
Harish Ramani
Ruthvik Ravindra
Abdul Rehman
Sai Divya Sangeetha Bhagavatula
Siddarth Sathyanarayanan
Gopalika Shama
Rishabh Rajesh Shanbhag
Sagar Singh
Mateusz Staniak
Sara Taheri
Anuska Tak
Derrie Susan Varghese
Amrutha Vempati

Video of the presentation: https://www.youtube.com/channel/UCnbUMFlIRLaY7fwfSintWuQ/