Today

- Visualization
  - grammar of graphics theory
  - demo
    - gog with any plotting tools
    - gog with ggplot
- Model checking
  - demo: ants case study

Visualization

- Main goal is to tell the story (data + model)
- Is the story clear?
- Are the data clear?
- Asking "How can I make the story or data clear?" is more useful than asking "What type of plot should I make?"
- The elements of the grammar help with this perspective

Visualization

- All data plots are built from marks and channels
- Different combinations of marks and channels are more effective for visual perception than others
- Best practice: evidence based

Marks

Points

0D

Lines

1D

Polygons
Areas

2D

Munzner T (2015) Visualization Analysis and Design

Channels

Position

→ Horizontal → Vertical → Both

Color

Shape

Tilt

Size

→ Length → Area → Volume

Munzner T (2015) Visualization Analysis and Design

Combining

mark:
line
channel 1
quantitative:
vertical
position
channel 2
category:
horizontal
position

mark:
point
channel 1
quantitative:
vertical
position
channel 2
quantitative:
horizontal
position

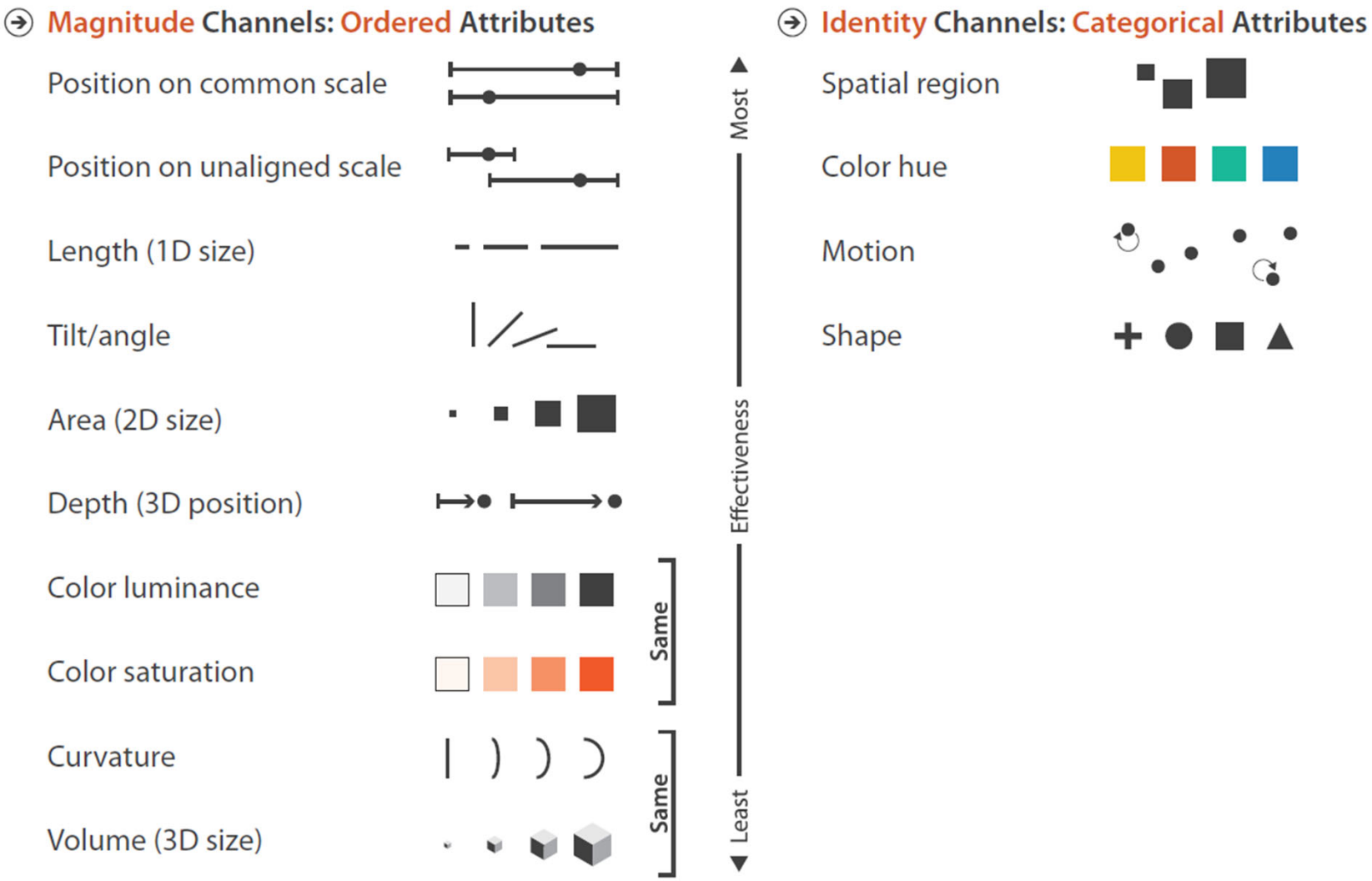
+
channel 3
category:
color

+
channel 4
quantitative:
size

Fundamental principles

- Expressiveness: visual encoding should express all, and only, info in the variable
  - ordered data: magnitude channels
  - unordered data: identity channels (should not imply ordering)
- Effectiveness: code the most important data features with the most effective channels

Effectiveness of channels



Munzner T (2015) Visualization Analysis and Design

Summary Wickham paper

- Marks: geom (geometric object): points, lines, polygons, ...
- Channels: aes (aesthetic) x, y, shape, color
- Mapping: map data to channels of marks
- Scales: one per aes
- Statistics (summarizing function)
- Coordinate system (one per plot)
- Annotations
- Facets

Layer

ggplot template

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>( mapping = aes(<MAPPINGS>),
                   <AES_SETTINGS>,
                   stat=<STAT>,
                   position=<POSITION> ) +
  <GEOM_FUNCTION>(...) +
  ...
  <SCALE_MODIFIERS> +
  <COORDINATE_SYSTEM> +
  <FACETS> +
  <THEME_MODIFIERS>
```

Layers

Axes

Legends

Labels

Color scales

stat & geom are intertwined: each has a reciprocal default

Fonts, background, margins, tick styles, borders, etc

Cheatsheet: https://www.rstudio.com/resources/cheatsheets (visualization)

ggplot inheritance

```
ggplot() +
  geom_point(data=mpg, mapping=aes(x=displ, y=hwy)) +
  geom_smooth(data=mpg, mapping=aes(x=displ, y=hwy))
```

```
ggplot(data=mpg, mapping=aes(x=displ, y=hwy)) +
  geom_point() +
  geom_smooth()
```

layers inherit

Diagnostics: model checking

- Systematic departures of the process (biological) model from the data
- Poor error distribution
- Mistakes in data
- Outliers
- Influential data points

Tools

- Plot the fitted model with the data
- Residuals vs fitted values
- QQ plot, histogram of residuals
- Leave one out (LOO) influence algorithm
- What should these diagnostics look like (on average and variation)?
  - make plots of them from simulated data of the fitted model