Graphics: Critique & creation

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair

Department of Statistics / Rice University

September 2011

Exploratory graphics

Are for **you** (not others). Need to be able to create rapidly because your first attempt will never be the most revealing.

Iteration is crucial for developing the best display of your data.

Gives rise to two key questions:

What should I plot?

How can I plot it?

Two general tools

Plot critique toolkit:
"graphics are like pumpkin pie"

Theory behind ggplot2:
"A layered grammar of graphics"

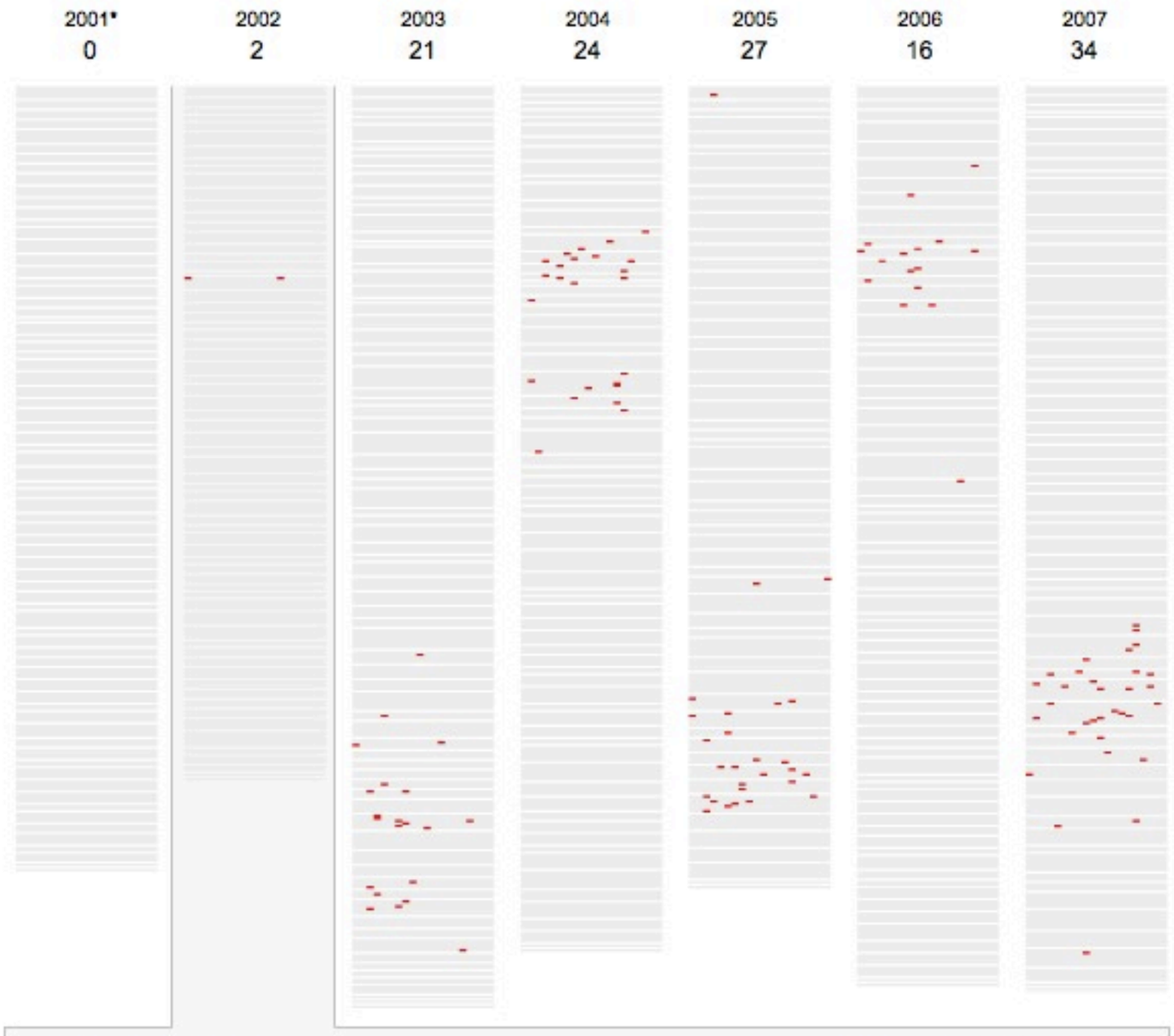
plus lots of practice...

What should I plot?

Critique

- State of the union: http://nyti.ms/r8KdvU
- How different groups spend their day: http://nyti.ms/np29Yk
- CA primary results: http://nyti.ms/r8Sh8N (Click margin of victory)

Use of the phrase "Iraq" in past State of the Union Addresses



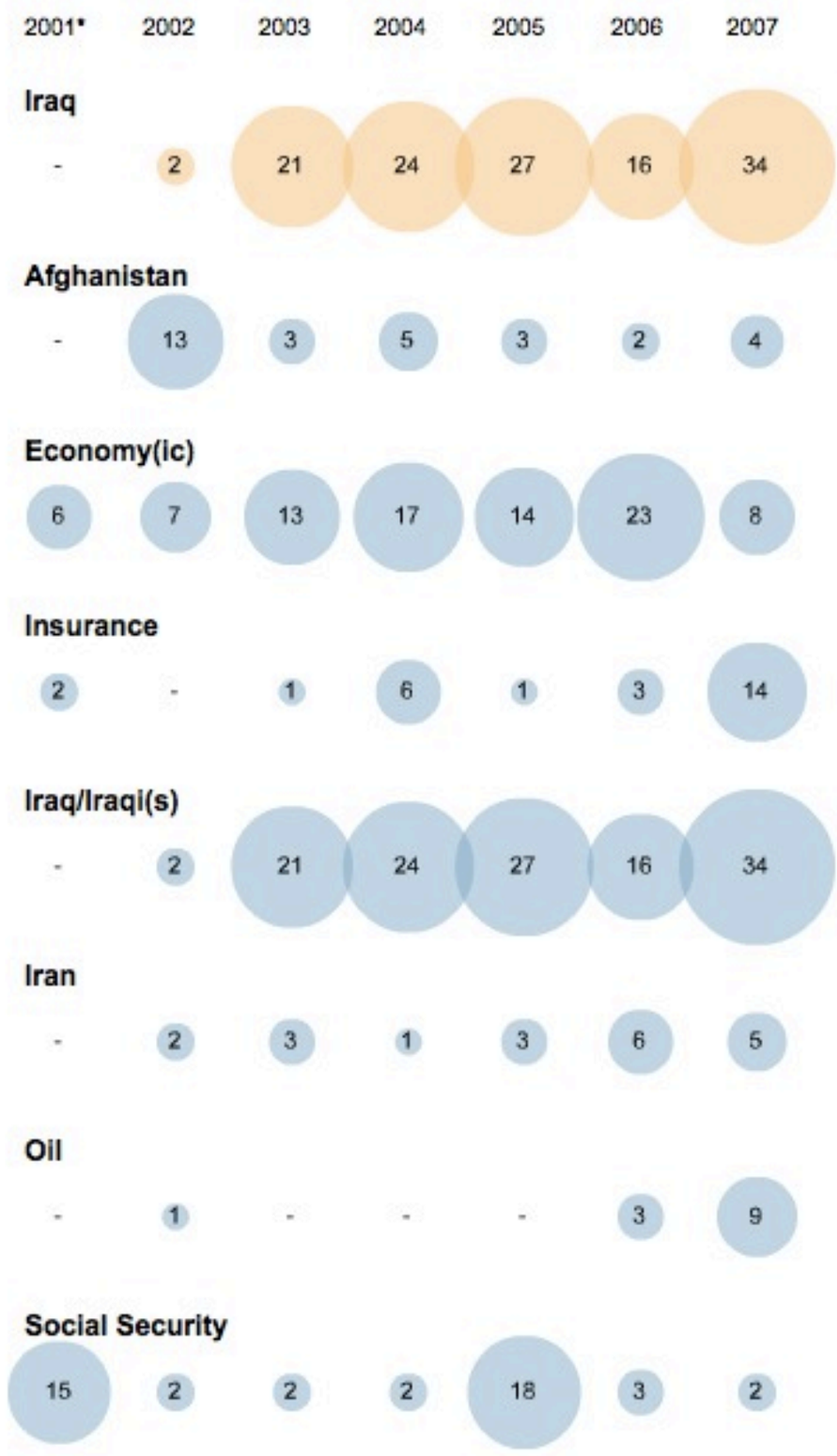
Next Instance of 'Iraq'

The word in context

IRAQ continues to flaunt its hostility toward America and to support terror. The Iraqi regime has plotted to develop anthrax, and nerve gas, and nuclear weapons for over a decade. This is a regime that has already used poison gas to murder thousands of its own citizens -- leaving the bodies of mothers huddled over their dead children. This is a regime that agreed to international inspections -- then kicked out the inspectors. This is a regime that has something to hide from the civilized world.

-- 2002 (Paragraph 20 of 67)

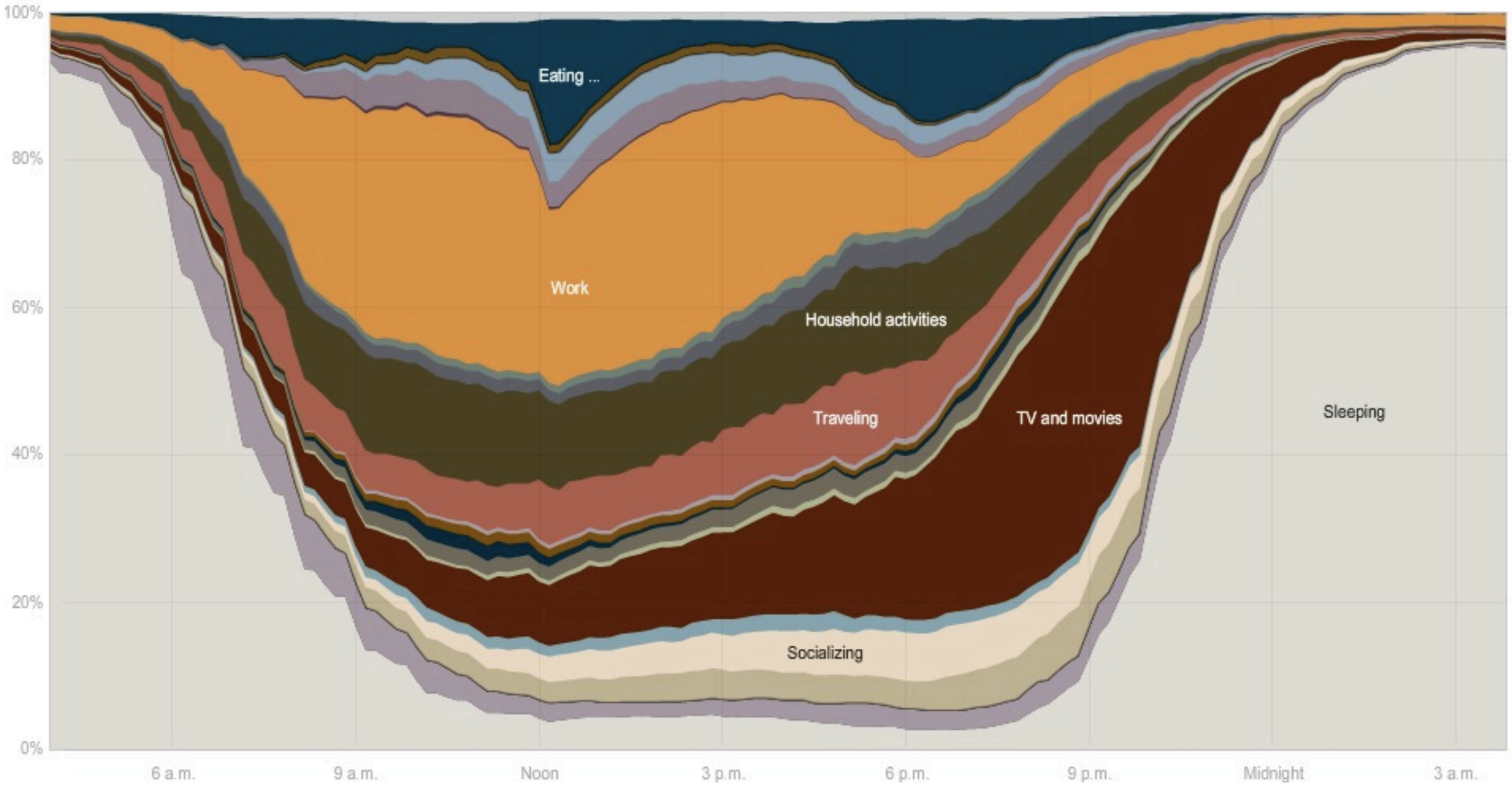
Compared with other words



Everyone

Sleeping, eating, working and watching television take up about two-thirds of the average day.

Everyone	Employed	White	Age 15-24	H.S. grads	No children
Men	Unemployed	Black	Age 25-64	Bachelor's	One child
Women	Not in lab...	Hispanic	Age 65+	Advanced	Two+ children

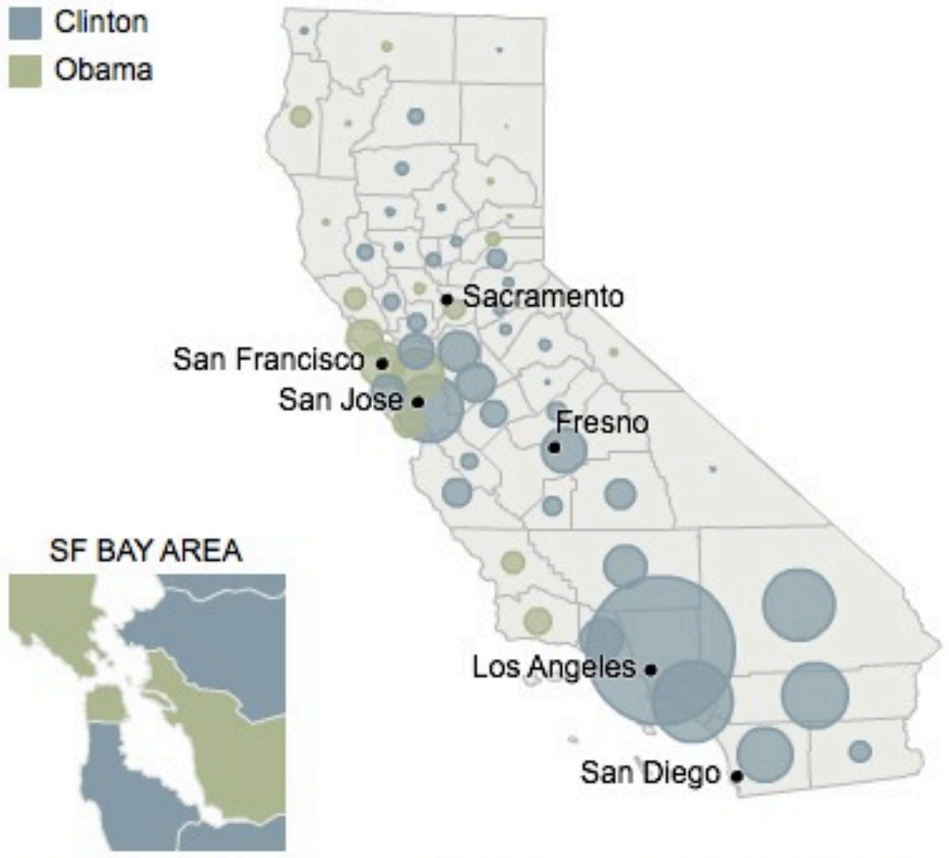


Results by County | **Margin of Victory**

Circles are proportional to the amount each county's leading candidate is ahead.

DEMOCRATS

Clinton
Obama



REPUBLICANS

McCain
Romney

Note: Maps show election returns as reported by The Associated Press.

Graphics are like pumpkin pie

The four C's of critiquing a graphic

Content

Construction


Context


Consumption

Content

What data (variables) does the graph display?

What non-data is present?

What is **pumpkin** (essence of the graphic) vs what is **spice** (useful additional info)?

Your turn

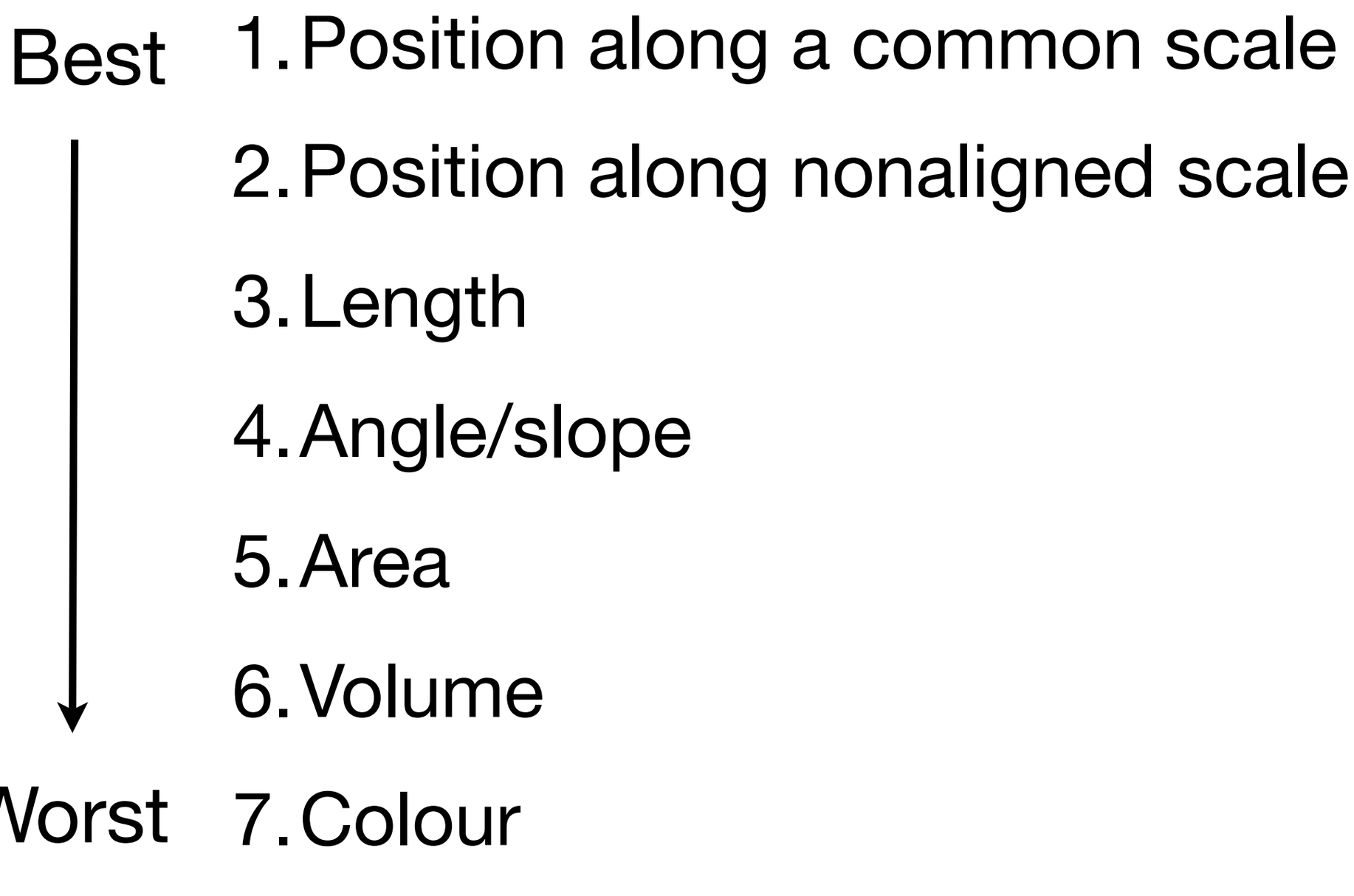
Pair up and identify the data and non-data in each of the three plots. Which features are the most important? Which are just useful background information?

Construction

How many layers are on the plot?

What data does each layer display? What sort of geometric object does it use? Is it a summary of the raw data? How are variables mapped to aesthetics?

Perceptual mapping



Your turn

Answer the following questions for each of the three plots:

How many layers are on the plot?

What data does the layer display? How does it display it?

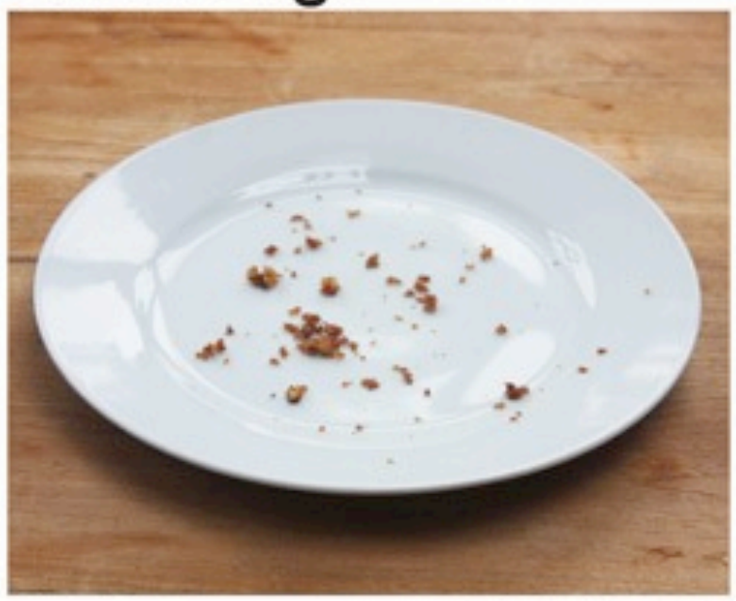
Another metaphor:

Data

Information

Presentation

Knowledge



http://epicgraphic.com/data-cake/

Can the explain composition of a graphic in words, but how do we create it?

How can I plot it?

"If any number of magnitudes are each the same multiple of the same number of other magnitudes, then the sum is that multiple of the sum."

Euclid, ~300 BC

"If any number of magnitudes are each the same multiple of the same number of other magnitudes, then the sum is that multiple of the sum."

Euclid, ~300 BC

$$m(\Sigma x) = \Sigma(mx)$$

The grammar of graphics

An abstraction which makes thinking about, reasoning about and communicating graphics easier.

Developed by Leland Wilkinson, particularly in “The Grammar of Graphics” 1999/2005

You’ve been using it in ggplot2 without knowing it! But to do more, you need to learn more about the theory.

What is a layer?

- Data
- Mappings from variables to aesthetics (**aes**)
- A geometric object (**geom**)
- A statistical transformation (**stat**)
- A position adjustment (**position**)

```
layer(geom, stat, position, data, mapping, ...)

layer(
  data = mpg,
  mapping = aes(x = displ, y = hwy),
  geom = "point",
  stat = "identity",
  position = "identity"
)

layer(
  data = diamonds,
  mapping = aes(x = carat),
  geom = "bar",
  stat = "bin",
  position = "stack"
)
```

```
# A lot of typing!

layer(
  data = mpg,
  mapping = aes(x = displ, y = hwy),
  geom = "point",
  stat = "identity",
  position = "identity"
)

# Every geom has an associated default statistic
# (and vice versa), and position adjustment.

geom_point(aes(displ, hwy), data = mpg)
geom_histogram(aes(carat), data = diamonds)
```

```r
# To actually create the plot
ggplot() +
  geom_point(aes(displ, hwy), data = mpg)
```

```r
ggplot() +
  geom_histogram(aes(displ), data = mpg)
```

```
# Multiple layers
ggplot() +
  geom_point(data = mpg, aes(displ, hwy)) +
  geom_smooth(data = mpg, aes(displ, hwy))

# Avoid redundancy:
ggplot(aes(displ, hwy), data = mpg) +
  geom_point() +
  geom_smooth()
```

```
# Different layers can have different aesthetics
ggplot(mpg, aes(displ, hwy)) +
  geom_point(aes(colour = class)) +
  geom_smooth()

ggplot(mpg, aes(displ, hwy, colour = class)) +
  geom_point() +
  geom_smooth(method = "lm", se = F)
```

```
ggplot(mpg, aes(displ, hwy, group = class)) +
  geom_point(aes(colour = class)) +
  geom_smooth(method = "lm", se = F)

ggplot(mpg, aes(displ, hwy)) +
  geom_point(aes(colour = class)) +
  geom_line(aes(group = class), stat = "smooth",
    method = "lm", se = F)
```

	stat	geom
histogram	bin	bar
smooth	smooth	ribbon
boxplot	boxplot	boxplot
density	density	line
freqpoly	bin	line

Your turn

For each of the following plots created with qplot, recreate the equivalent ggplot code.

```
qplot(carat, price, data = diamonds)
```

```
qplot(hwy, cty, data = mpg, geom = "jitter")

qplot(reorder(class, hwy), hwy, data = mpg,
  geom = c("jitter", "boxplot"))

qplot(log10(carat), log10(price),
data = diamonds, colour = color) +
geom_smooth(method = "lm")
```

```
ggplot(diamonds, aes(carat, price)) +
  geom_point()

ggplot(mpg, aes(hwy, cty)) +
  geom_jitter()

ggplot(mpg, aes(reorder(class, hwy), hwy)) +
  geom_jitter() +
  geom_boxplot()

ggplot(diamonds, aes(log10(carat), log10(price),
  colour = color)) +
  geom_point() +
  geom_smooth(method = "lm")
```

More geoms & stats

See http://had.co.nz/ggplot2 for complete list with helpful icons:

Geoms: (0d) point, (1d) line, **path**, (2d) boxplot, bar, **tile**, **text**, polygon, linerange.

Stats: bin, density, summary, sum

Advanced layering

Layering

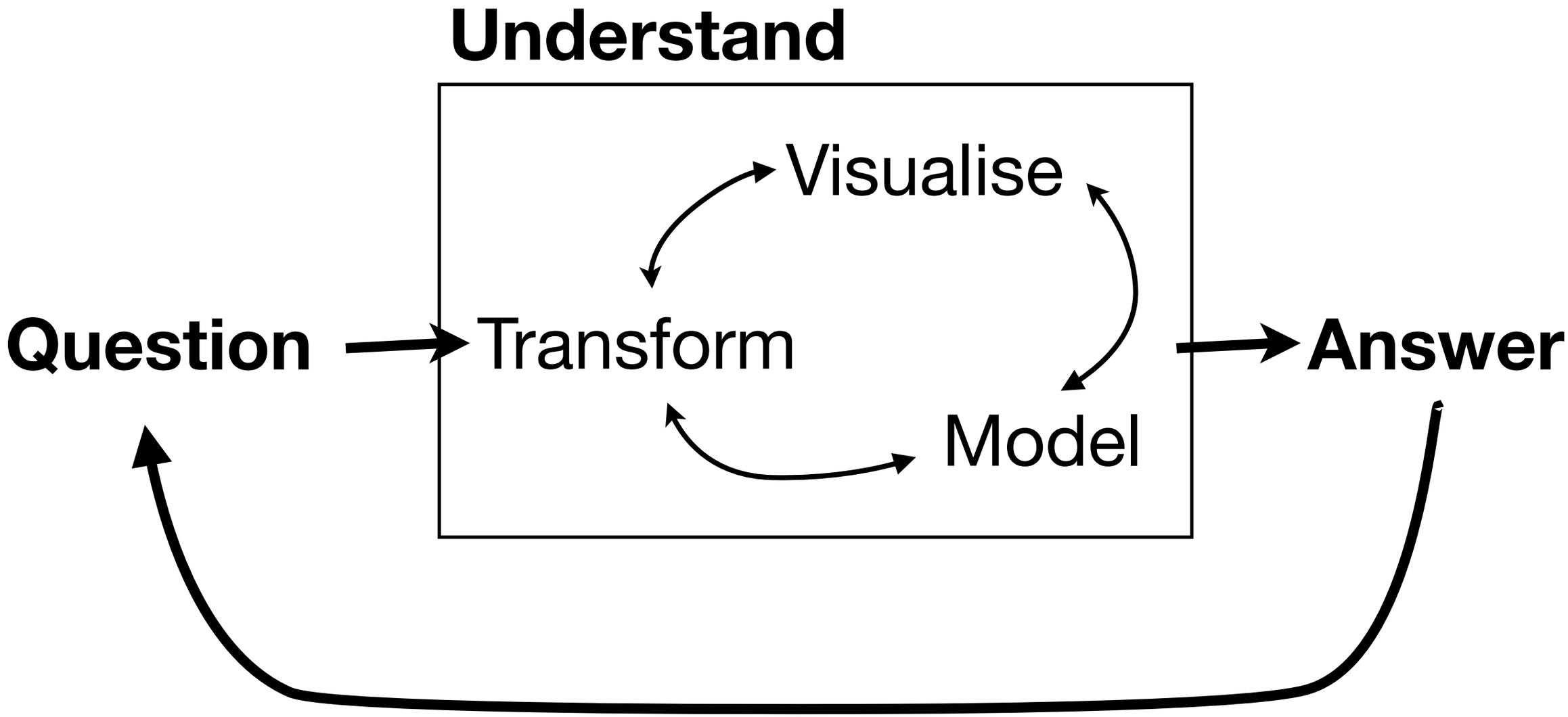
Key to rich graphics is taking advantage of layering.

Three types of layers: context, raw data, and summarised data

Each can come from a different dataset.

Iteration

- First plot is never the best. Have to keep iterating to understand what's going on.
- Don't try and do too much in one plot.
- Best data analyses tell a story, with a natural flow from beginning to end.


Understand
Visualise
Question
Transform
Answer
Model

```
qplot(x, y, data = diamonds)
diamonds$x[diamonds$x == 0] <- NA
diamonds$y[diamonds$y == 0] <- NA
diamonds$y[diamonds$y > 20] <- NA
```

```
diamonds <- mutate(diamonds,
  area = x * y,
  lratio = log10(x / y))

qplot(area, lratio, data = diamonds)
diamonds$lratio[abs(diamonds$lratio) > 0.02] <- NA
```

```
ggplot(diamonds, aes(area, lratio)) +
  geom_point()

ggplot(diamonds, aes(area, lratio)) +
  geom_hline(yintercept = 0, size = 2, colour = "white") +
  geom_point() +
  geom_smooth(method = lm, se = F, size = 2)

ggplot(diamonds, aes(area, abs(lratio))) +
  geom_hline(yintercept = 0, size = 2, colour = "white") +
  geom_point() +
  geom_smooth(se = F, size = 2)
```

```
ggplot(diamonds, aes(area, abs(lratio))) +
  geom_hline(yintercept = 0, size = 2, colour = "white") +
  geom_boxplot(aes(group = round_any(area, 5))) +
  geom_smooth(se = F, size = 2)

ggplot(diamonds, aes(area, abs(lratio))) +
  geom_hline(yintercept = 0, size = 2, colour = "white") +
  geom_boxplot(aes(group = round_any(area, 5)))

ggplot(diamonds, aes(area, lratio)) +
  geom_hline(yintercept = 0, size = 2, colour = "white") +
  geom_boxplot(aes(group = interaction(sign(lratio),
   round_any(area, 5))), position = "identity")
```