# Introduction to Data Visualization

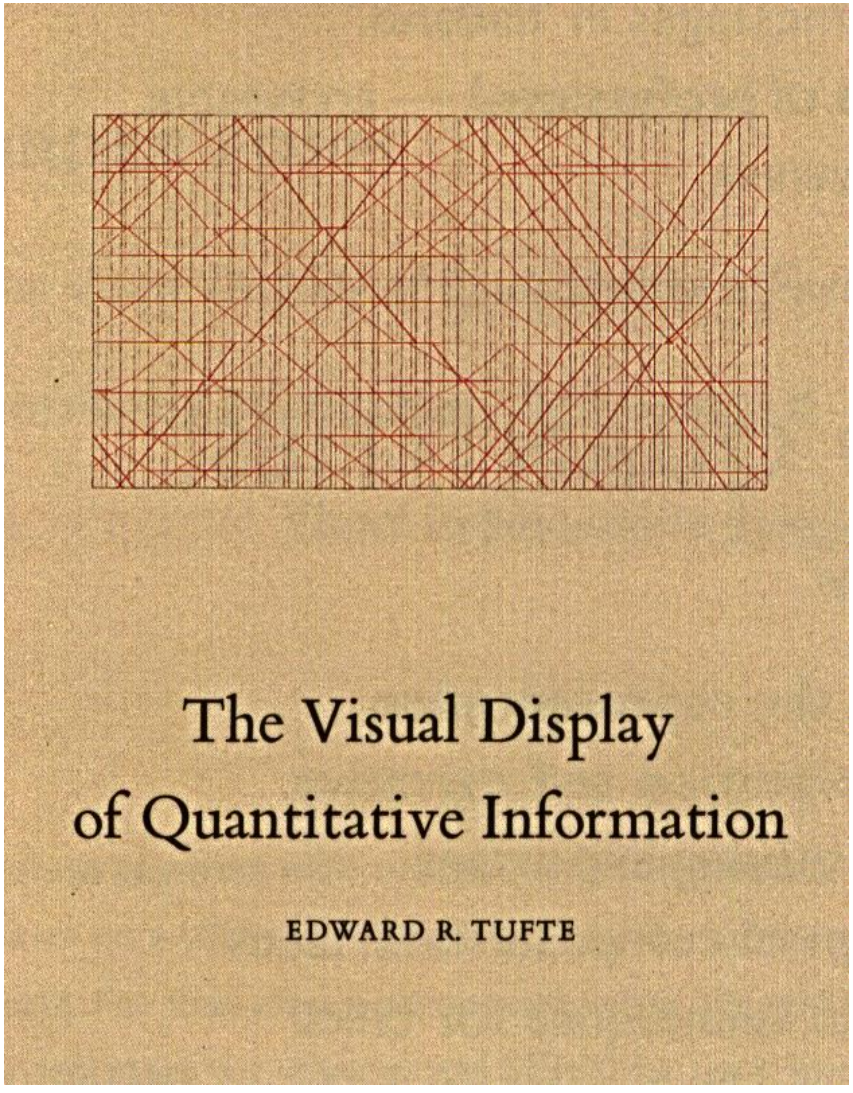
Author: Nicholas G Reich

*This material is part of the* **statsTeachR** *project*

# Visualization excellence

In Tufte's words:

- consists of complex ideas communicated with clarity, precision, and efficiency.
- is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
- is nearly always multivariate.
- requires telling the truth about the data.



# Components of data graphics

# Warm up

For each of the following graphics, work in pairs to

1. identify the variables displayed;
2. identify 2 features that you like and 2 that you don't;
3. sketch out the tidy data represented in the figure.

# "Cities, traffic and CO2"[1]

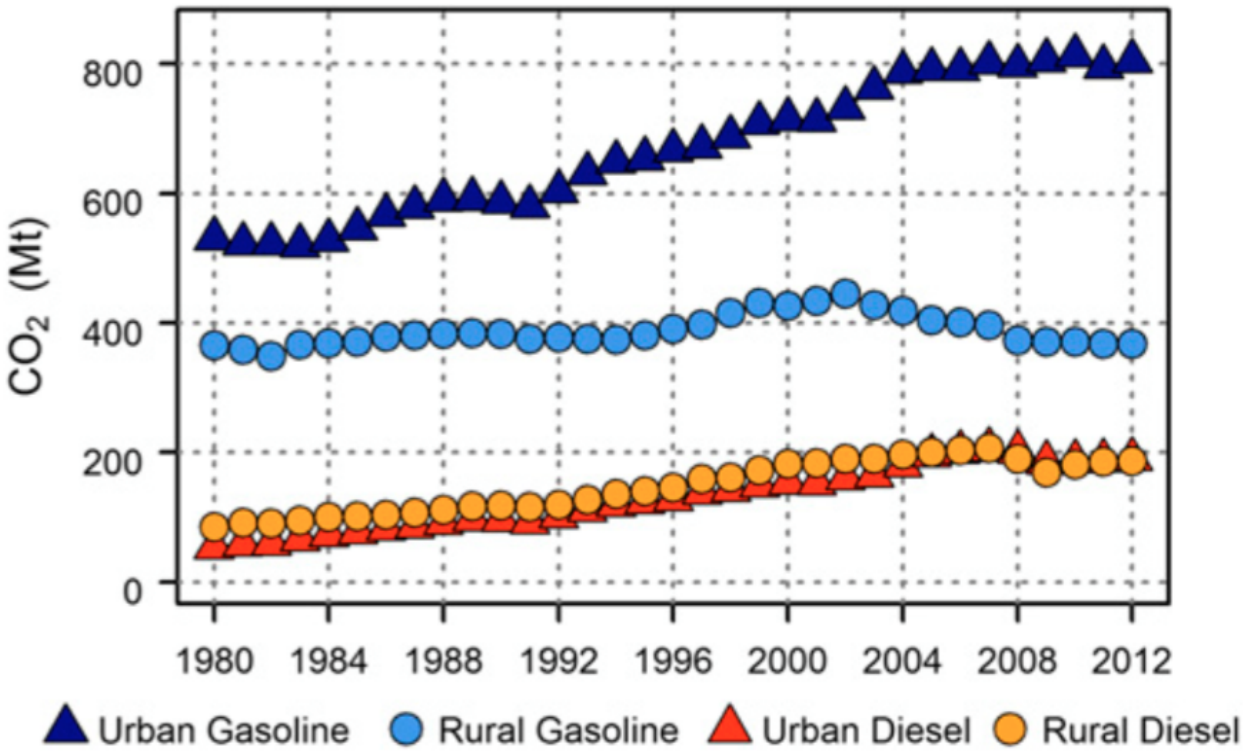
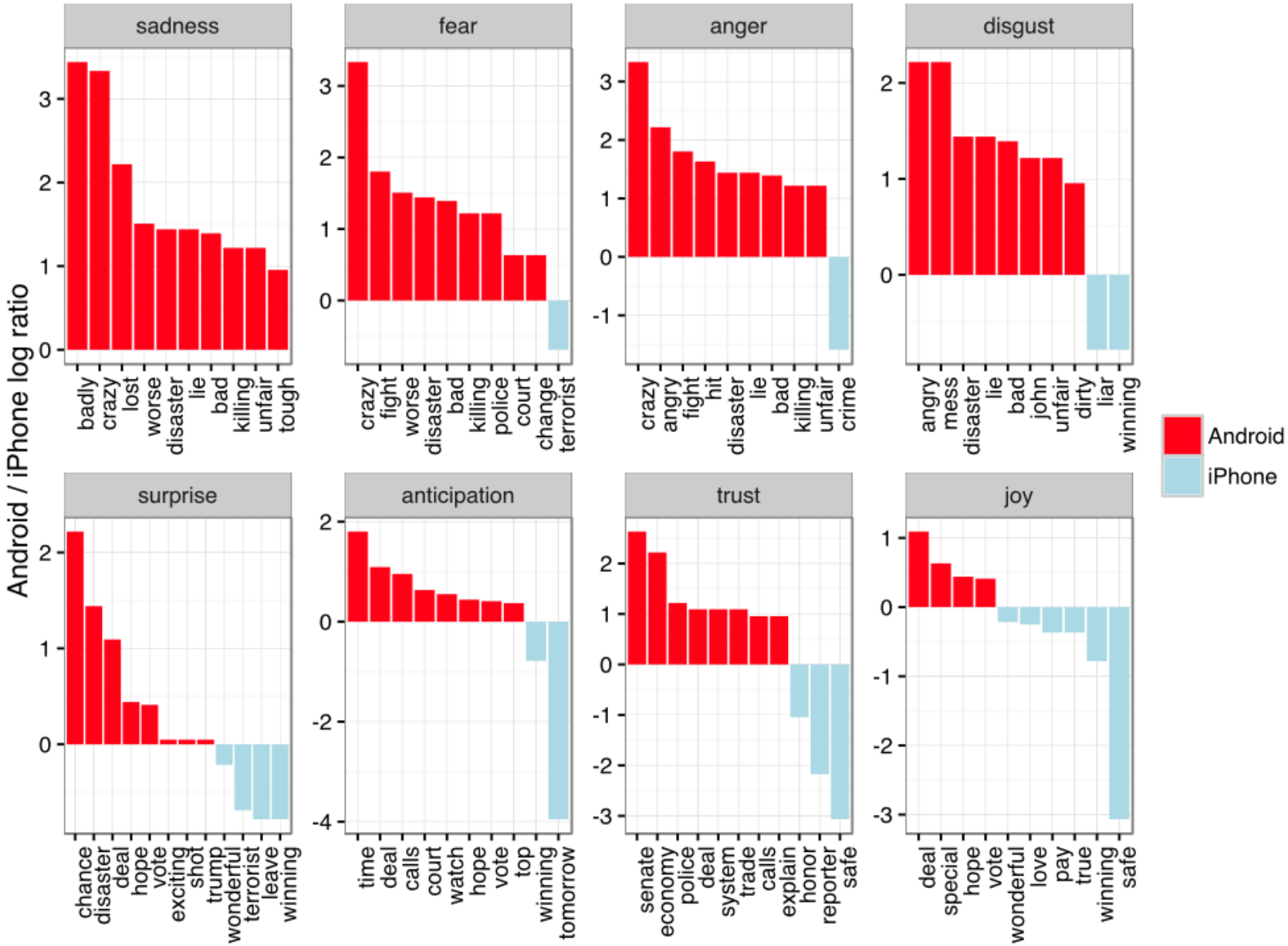



**Fig. 2.** Time series of US on-road $CO_2$ emissions. Urban roads accounted for 80% of total emissions growth since 1980. Rural road emissions have been declining since 2002.

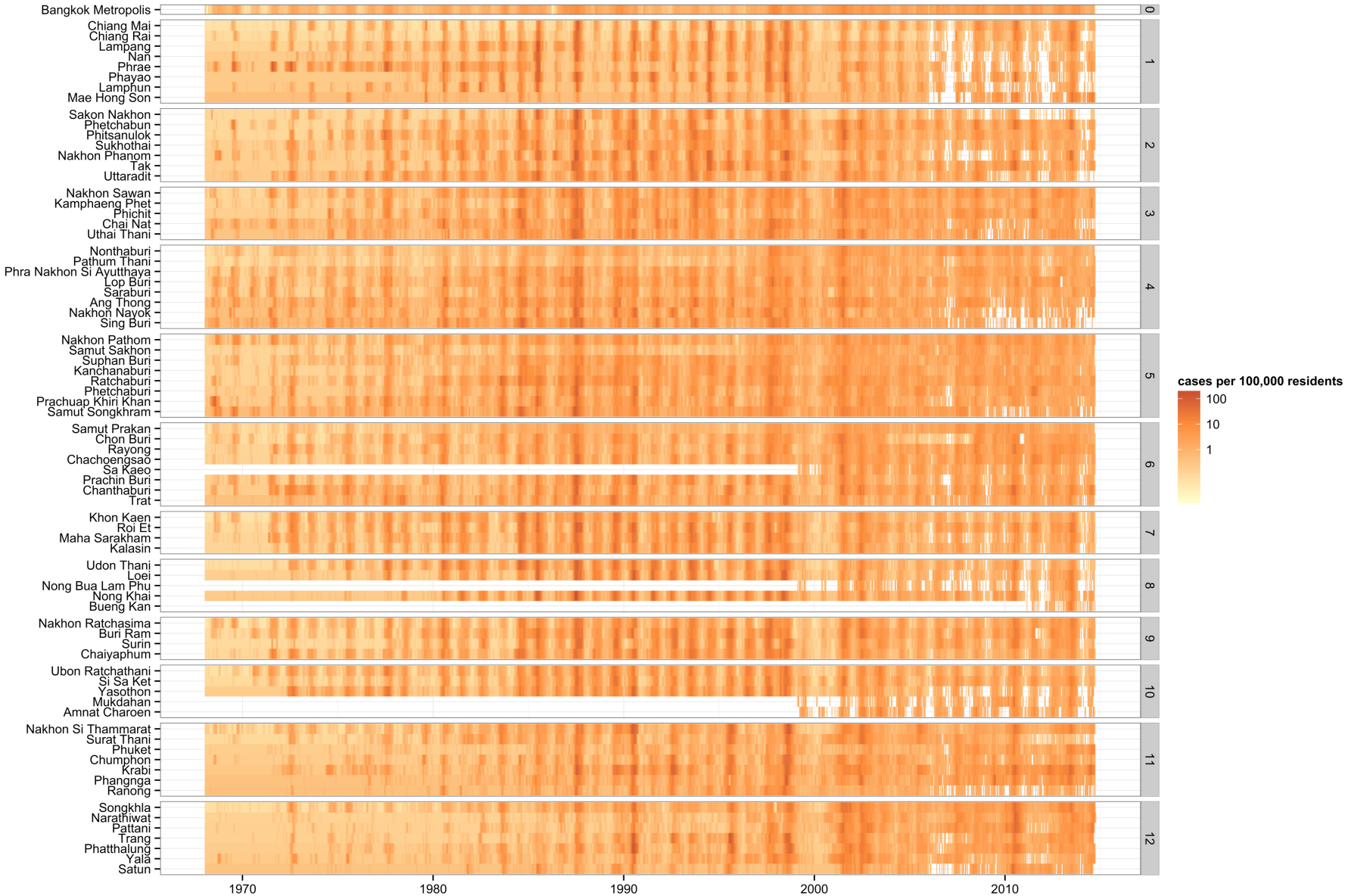
[1] from "Cities, traffic, and CO2: A multidecadal assessment of trends, drivers, and scaling relationships", Gately et al, PNAS, 2015.

# Trump tweets[2]




[2] http://varianceexplained.org/r/trump-tweets/

# Dengue cases in Thailand[3]




[3] adapted from Reich et al, 2016.

# Why do we visualize data?

# Exploratory graphics

- The most valuable graphics are often the simple ones you make for yourself.
- Exploratory graphics can introduce you to a dataset.
- Key goal: understand the variation.
- What do you want to know about these data?

```
data(airquality)
head(airquality)

##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

# Exploratory summaries: airquality data

Some quick text-based/tabular summaries

```
nrow(airquality)

summary(airquality)

table(airquality$Month)

with(airquality, table(Month, Day))
```

# Univariate graphics: airquality data

```
library(ggplot2)

p <- ggplot(airquality)

## better or worse than the table?
p + geom_bar(aes(x=factor(Month)))

## which of these do you prefer and why?
p + geom_density(aes(Ozone))
p + geom_histogram(aes(x=Ozone))
```

# Multivariate graphics: airquality data

```
p + geom_boxplot(aes(x=factor(Month), y=Ozone))

p2 <- ggplot(airquality, aes(x=Temp, y=Ozone))
p2 + geom_point()
p2 + geom_point() + geom_smooth()
p2 + geom_point() + geom_smooth(se=FALSE)

p3 <- ggplot(airquality,
             aes(x=Temp, y=Ozone, color=factor(Month)))
p3 + geom_point() + geom_smooth(se=FALSE)
```

# Multivariate graphics: pairs plots!

Pairs plots are sweet, but can take some time to render (especially for big-datasets).

```
library(GGally)
ggpairs(airquality)
```

# Your turn!

Try visualizing some of the NHANES data

```
library(NHANES)
data(NHANES)
?NHANES
```

# ggplot2

# Choices for R graphics

You have three central choices for making graphics in R:

- "Base graphics"
- ggplot2
- lattice

# Understanding the “grammar” of ggplot2

The grammar ...

- geom
- aesthetics (‘aes‘)
- scales
- facets
- data
- ... and more here: http://docs.ggplot2.org/current/

# What is a "geom"?

From Hadley:

- Geoms define the basic "shape" of the elements on the plot
- Basics: point, line, bar, text, hline, vline
- Statistics: histogram, smooth, density
- Others: boxplot, pointrange, linerange, ribbon

For more info check out the documentation:
http://docs.ggplot2.org/current

# What are "aesthetics"?

Aesthetics define a mapping between data and the display.[4]

| length | width | depth | trt |
|---|---|---|---|
| 2 | 3 | 4 | a |
| 1 | 2 | 1 | a |
| 4 | 5 | 15 | b |
| 9 | 10 | 80 | b |

| x | y | colour |
|---|---|---|
| 2 | 3 | a |
| 1 | 2 | a |
| 4 | 5 | b |
| 9 | 10 | b |

[4] Figure credits: Hadley Wickham

# geom_point

Each geom has a different set of aesthetics.
What aesthetics do we need for `geom_point`?

# geom_point

Each geom has a different set of aesthetics.
What aesthetics do we need for geom_point?

- x (required)
- y (required)
- alpha
- color
- fill
- shape
- size

# geom_line

What aesthetics do we need for geom_line?

# geom_line

What aesthetics do we need for geom_line?

- x (required)
- y (required)
- alpha
- color
- linetype
- size

# Try mplot for learning ggplot2 syntax

```
library(mosaic)
## downsample the dataset to make it smaller
NHANES_samp <- sample(NHANES, size = 1000)
mplot(NHANES_samp)
```

# Summary: Key principles of data graphics

- "**Show** the data"
- "Encourage the eye to **compare** different pieces of data"
- **Simplify** by maximizing the "data-ink ratio."
- Leverage color, shapes, facets to highlight multivariate data.
- Annotate your figures with context.