

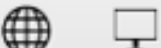





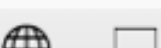
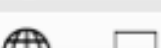


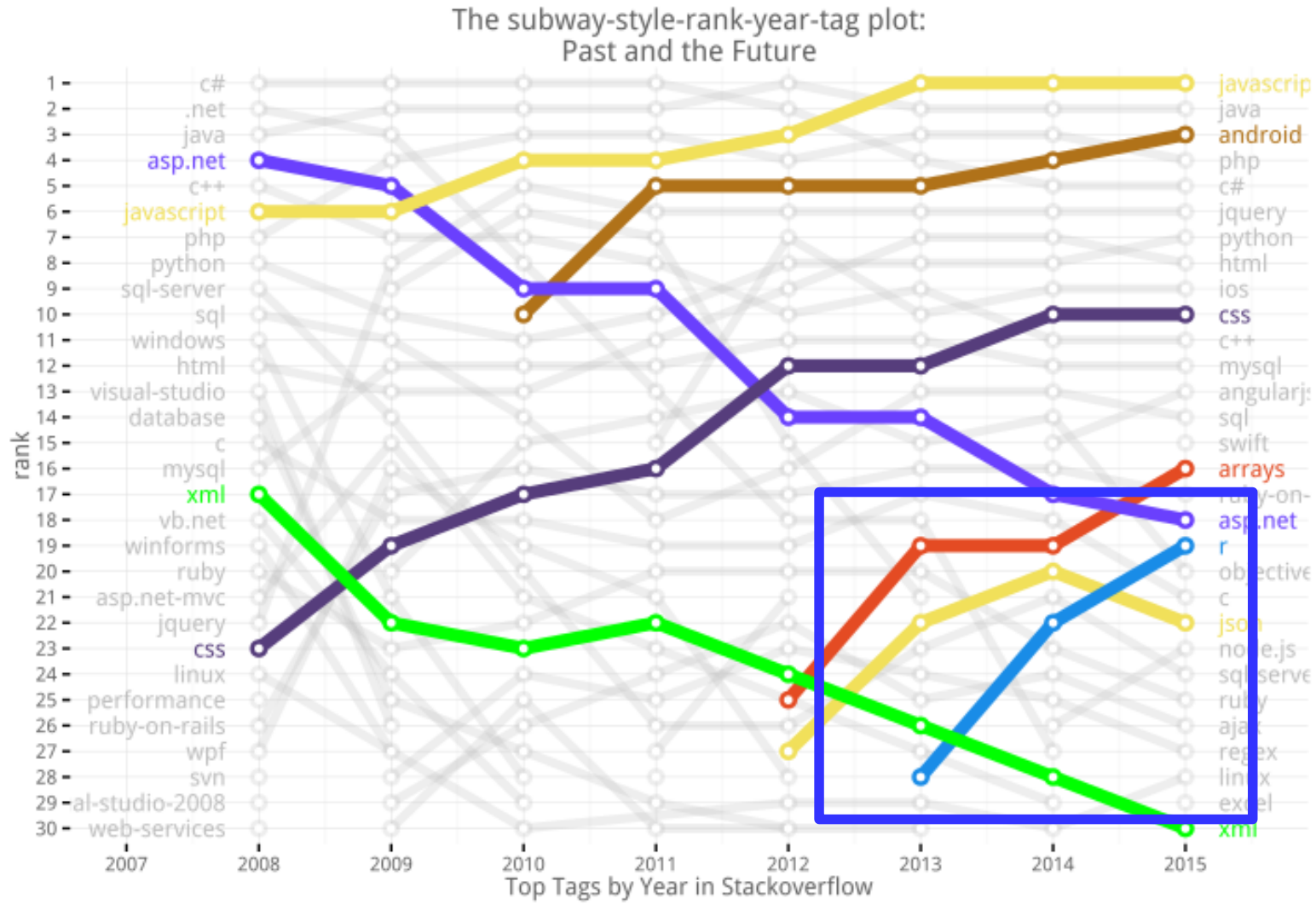
An intro to the R tidyverse: dplyr,ggplot2, and some Twitter conference data visualization

Dr. Thomas E. Keller
@tek_keller

R, bolstered by interest in data-science, has grown

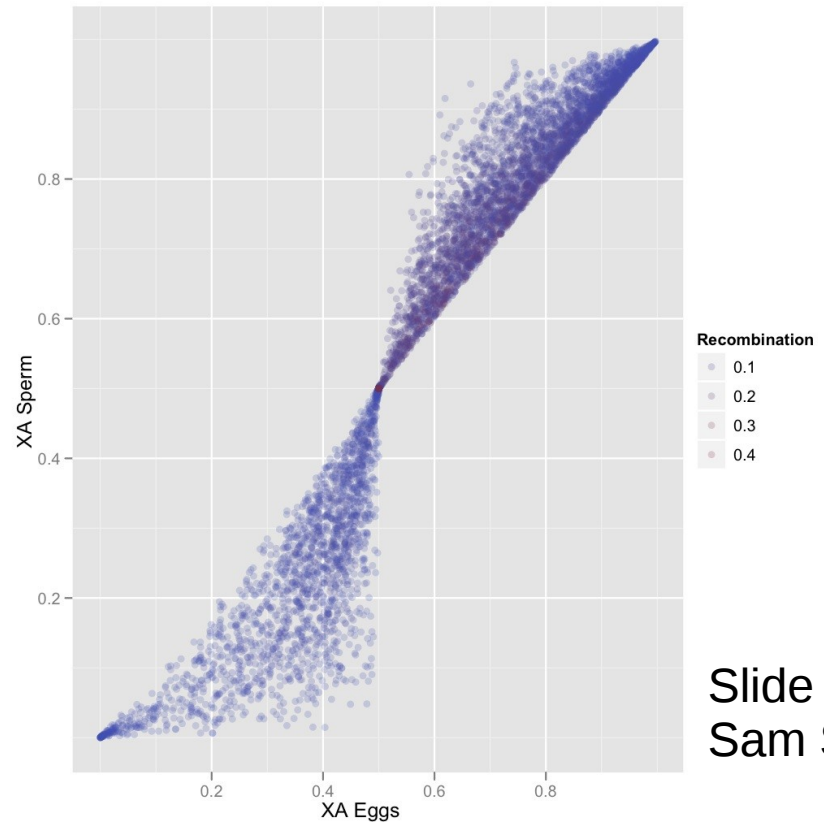
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

R gets on the subway



Why [R]?

- Advanced techniques
- A powerful research tool
- Repeatable code
- Price
- PRICE
- PRICE



Slide by
Sam Scarpino

R Cran – core binaries & host of the zillions of libraries for specialized stats and visualizations



CRAN

[Mirrors](#)

[What's new?](#)

[Task Views](#)

[Search](#)

About R

[R Homepage](#)

[The R Journal](#)

Software

[R Sources](#)

[R Binaries](#)

[Packages](#)

[Other](#)

Documentation

[Manuals](#)

[FAQs](#)

[Contributed](#)

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (Tuesday 2016-06-21, Bug in Your Hair) [R-3.3.1.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

RStudio

- <http://rstudio.org/>

Slide by
Sam Scarpino

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for installing ggplot2, viewing the diamonds dataset, and creating a faceted scatter plot of Price vs. Carat, colored by clarity.
- Console:** Shows the output of the summary function for the diamonds dataset and the execution of the plotting commands.
- Workspace/History:** Lists the executed R commands and their timestamps.
- Plots Panel:** Displays a scatter plot titled "Diamond Pricing" with Carat on the y-axis and Price on the x-axis. Points are colored by clarity, with a legend on the right showing categories: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF.

```
install.packages("ggplot2")
library(ggplot2)

View(diamonds)
summary(diamonds)
summary(diamonds$price)

qplot(price, carat, data = diamonds)

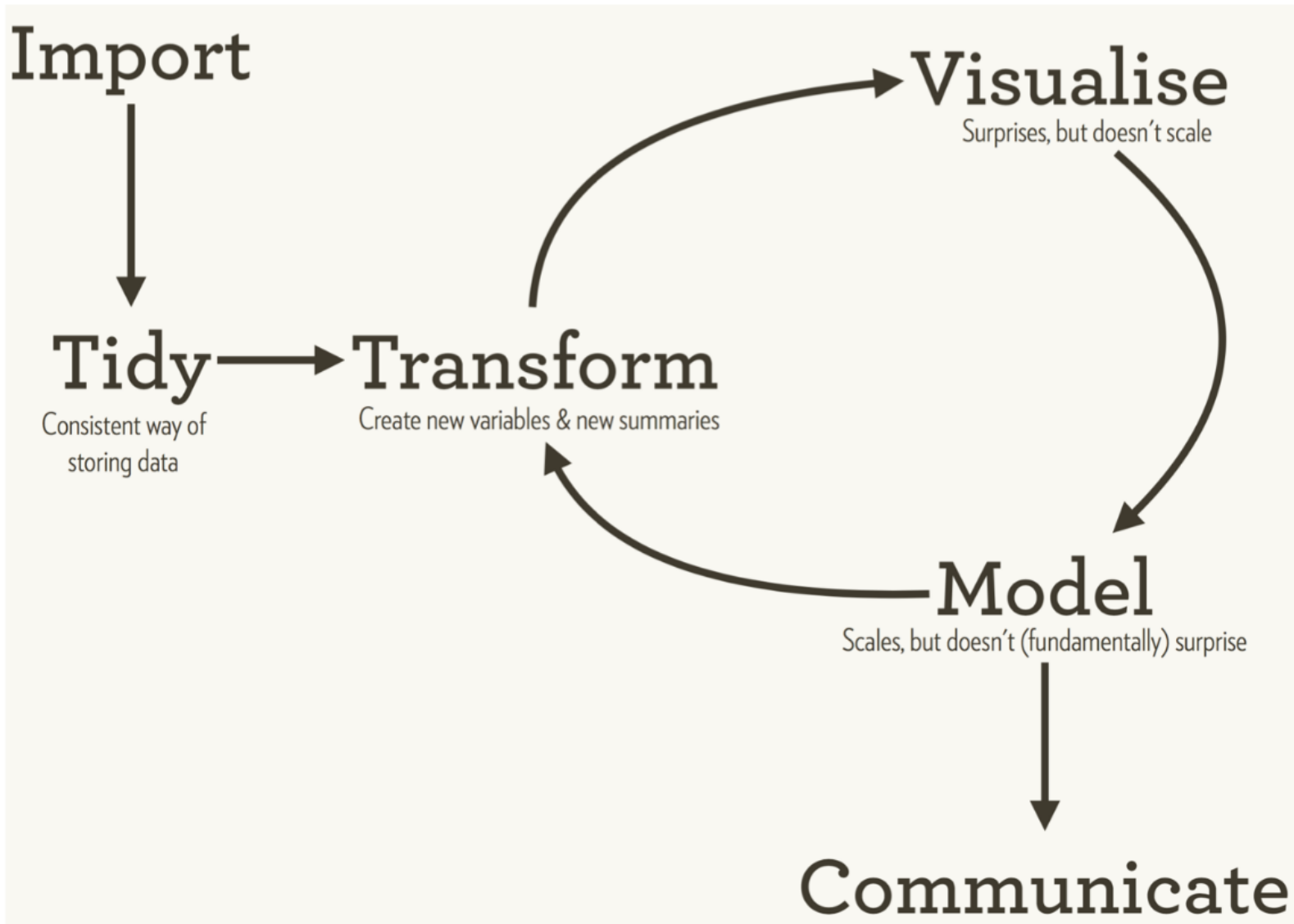
qplot(price, carat, data = diamonds, color=clarity,
      xlab = "Price", ylab = "Carat",
      main = "Diamond Pricing") +
  opts(plot.title = theme_text(size = 22))
```

x	y	z
Min. : 0.000	Min. : 0.000	Min. : 0.000
1st Qu.: 4.710	1st Qu.: 4.720	1st Qu.: 2.910
Median : 5.700	Median : 5.710	Median : 3.530
Mean : 5.731	Mean : 5.735	Mean : 3.539
3rd Qu.: 6.540	3rd Qu.: 6.540	3rd Qu.: 4.040
Max. : 10.740	Max. : 58.900	Max. : 31.800

```
> summary(diamonds$price)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  326    950    2401    3933    5324   18820

> qplot(price, carat, data = diamonds)
> qplot(price, carat, data = diamonds, color = clarity, xlab =
"Price", ylab = "Carat", main = "Diamond Pricing") +
  opts(plot.title = theme_text(size = 22))
>
```

Statistical analysis cycle (Wickham)



Tidy that data (one observ. Per row)

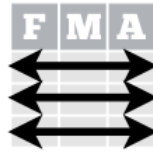
Tidy Data - A foundation for wrangling in R

In a tidy data set:



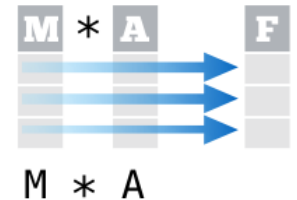
Each **variable** is saved in its own **column**

&

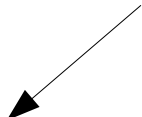


Each **observation** is saved in its own **row**

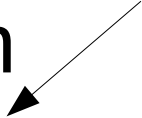
Tidy data complements R's **vectorized operations**. R will automatically preserve observations as you manipulate variables. No other format works as intuitively with R.



Untidy (No, Bad)

- #> country year type count
 - #> 1 Afghanistan 1999 cases 745
 - #> 2 Afghanistan 1999 population 19987071
 - #> 3 Afghanistan 2000 cases 2666
 - #> 4 Afghanistan 2000 population 20595360
 - #> 5 Brazil 1999 cases 37737
 - #> 6 Brazil 1999 population 172006362
- One Observation
- 

Tidy!

- #> country year cases population  One Observation
- #> 1 Afghanistan 1999 745 19987071
- #> 2 Afghanistan 2000 2666 20595360
- #> 3 Brazil 1999 37737 172006362
- #> 4 Brazil 2000 80488 174504898
- #> 5 China 1999 212258 1272915272
- #> 6 China 2000 213766 1280428583

The pipe (think linux pipe)

dplyr::%>%

Passes object on left hand side as first argument (or . argument) of function on righthand side.

x %>% f(y) *is the same as* **f(x, y)**

y %>% f(x, ., z) *is the same as* **f(x, y, z)**

"Piping" with %>% makes code more readable, e.g.

```
iris %>%  
  group_by(Species) %>%  
  summarise(avg = mean(Sepal.Width)) %>%  
  arrange(avg)
```

Twitter Mining with R

- Key libraries
 - `twitteR`, `streamR` to download from twitter
 - `ROAuth` to actually get authorized to use your account
 - `twitteR` is fine for short term things, but for large, sustained analyses use `streamR` (downloading 100k stream of `#rio2016`, etc)
 - `TwitteR` is limited to 1 week window of twitter search API
 - `Tidytext`, `wordcloud`, `lubridate`, `ggraph`
 - Breaking down tweets into words and doing sentiment analysis, in a “tidy” fashion
 - `Lubridate` for date parsing (it is a nightmare in R)
 - `Ggraph` -plot networks with `ggplot`

Resources

- <http://r4ds.had.co.nz/>
- <http://docs.ggplot2.org/current/>
- The ggplot2 book (v2 is out)!
- <https://github.com/hadley/ggplot2-book>
- #rstats on twitter is great, not huge egos, helpful
- Stack overflow