Data Munging and Exploratory Data Analysis

Tyrone Ryba
Fall 2021

Data munging and EDA

Prerequisites: { }

Course topics

1) A primer on R
2) Gathering, annotating, and structuring large datasets
3) Data tables and file connections
4) Preprocessing topics; aggregation, transformation, normalization, reduction, validation and consistency
5) Data structures and manipulation: UNIX shells, base R, reshape, dplyr
6) Exploratory data analysis and principles of data display
7) Exploratory graphics in base R, ggplot2, Python and alternatives
8) Interactive graphics and EDA applications
9) Reporting results in figures and tables
10) Group projects

Evaluation

- Three practical exams (week 4, 9, 14; 20% each)
- Two group projects (week 5, 11; 20% each)

Course topics: Data munging

- Originally described data processing steps that were impossible to reverse or reproduce – “mash until no good”

- Common early and intermediate stages in data analysis and data science projects

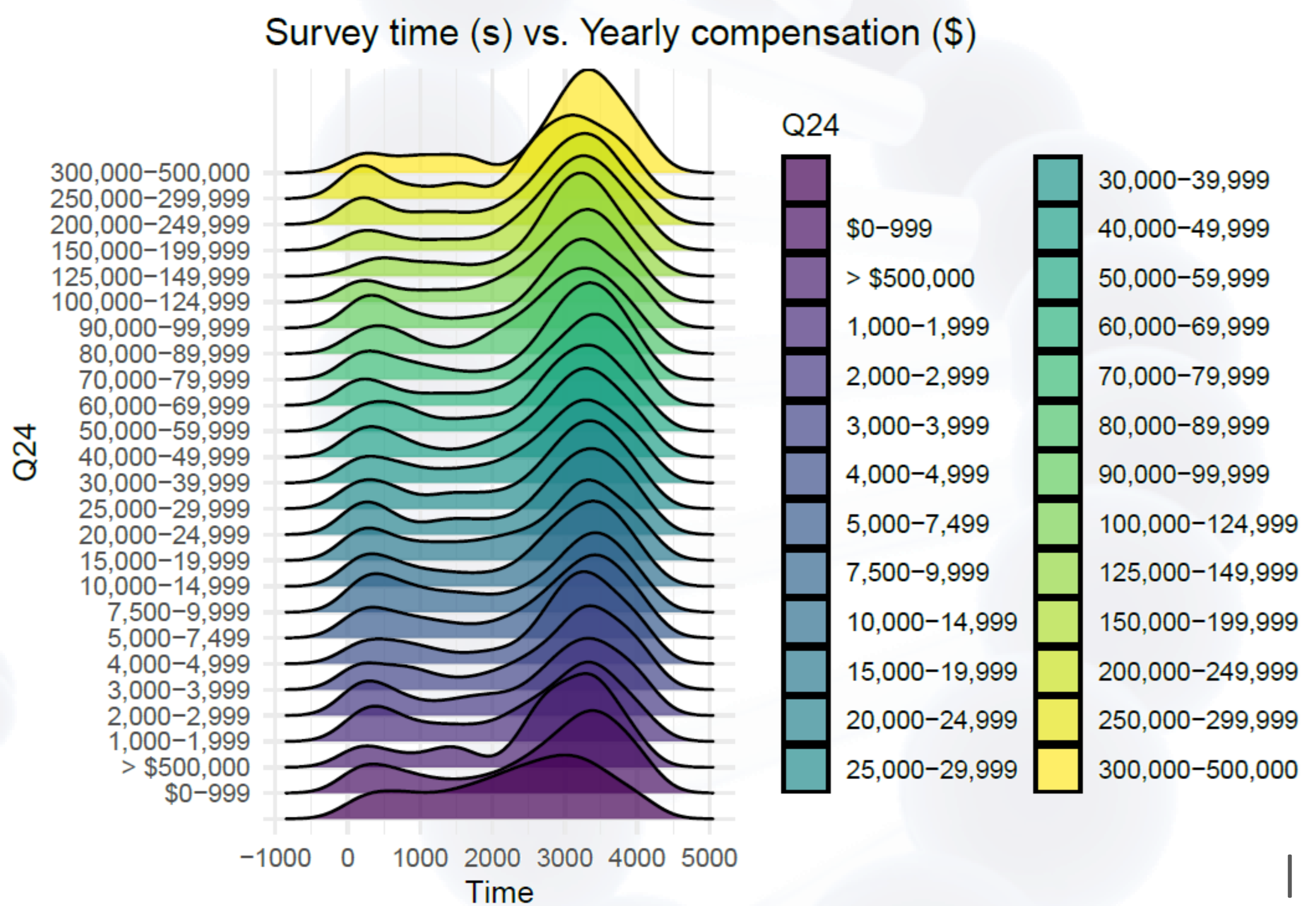
- Data provenance and reproducibility

Hadley Wickham: *“Tidy datasets are all alike but every messy dataset is messy in its own way.”*

Exploratory data analysis

Methods to discover relationships between variables

- Correlations, associations
- Direct and indirect functional relationships
- Advocated by / largely attributed to John Tukey



Exploratory data analysis

Idea: Decisions about downstream processing steps should be supported by descriptive statistics and visualization

Goals:

- Make observations from multiple perspectives
- Check assumptions of statistical models
- Identify patterns for predictive models

John Snow’s map of the 1854 Cholera outbreak:

Data visualization and EDA

A major early success in epidemiology, and required only:

Ability to plot cases

Ability to notice patterns

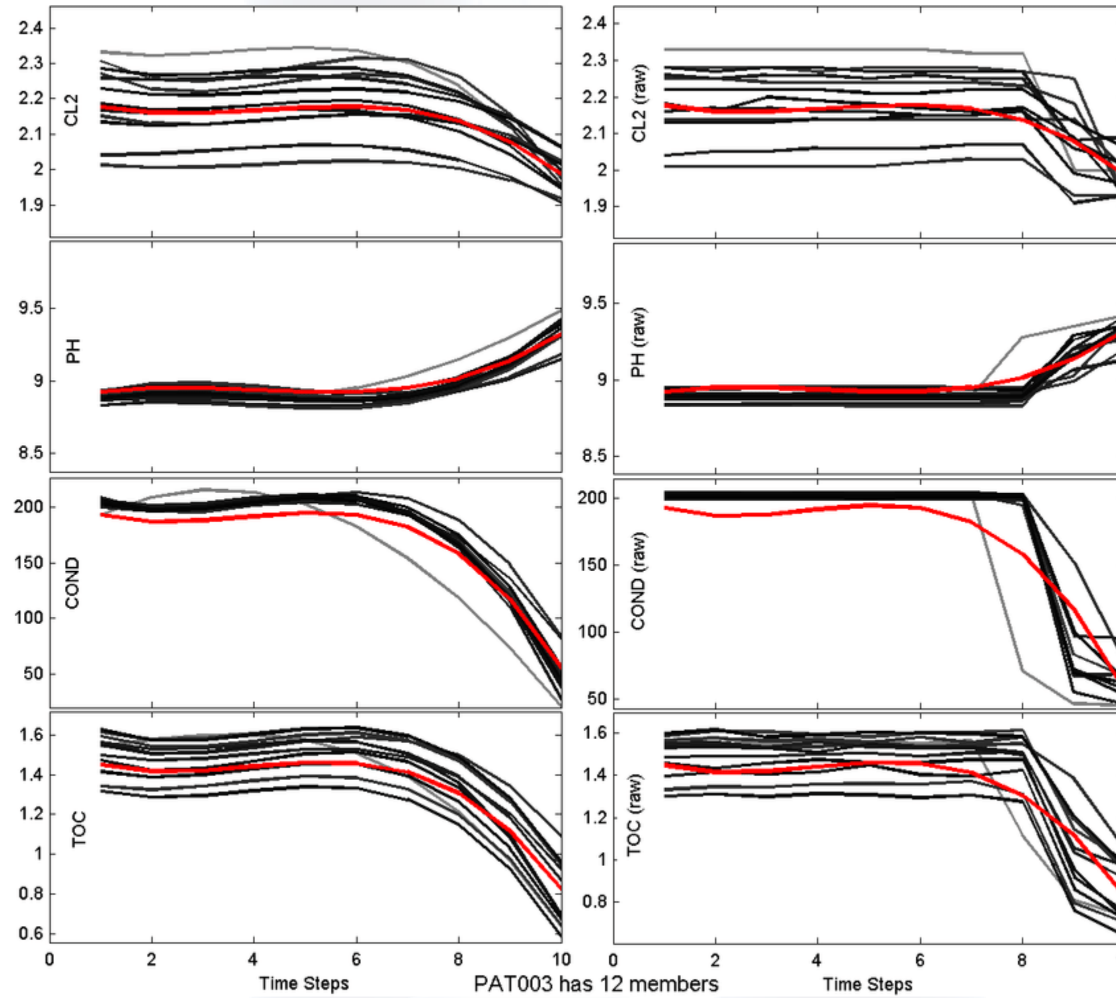
Data visualization and EDA

Some pattern finding examples:

Data visualization and EDA

Similar care will be required in interpreting exploratory graphics and models:

Changing interpretation with
- Color palettes
- Visualization method
- Preprocessing choices
- Color blindness, etc.



FINAL FINAL

POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become common-

the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Data visualization and EDA

Be aware of issues in measurement:

Systematic versus random errors (accuracy vs. precision)

Measurement directness (media H1N1 stories vs. search trends)

Algorithm may interact with measurement, as in suggested search terms

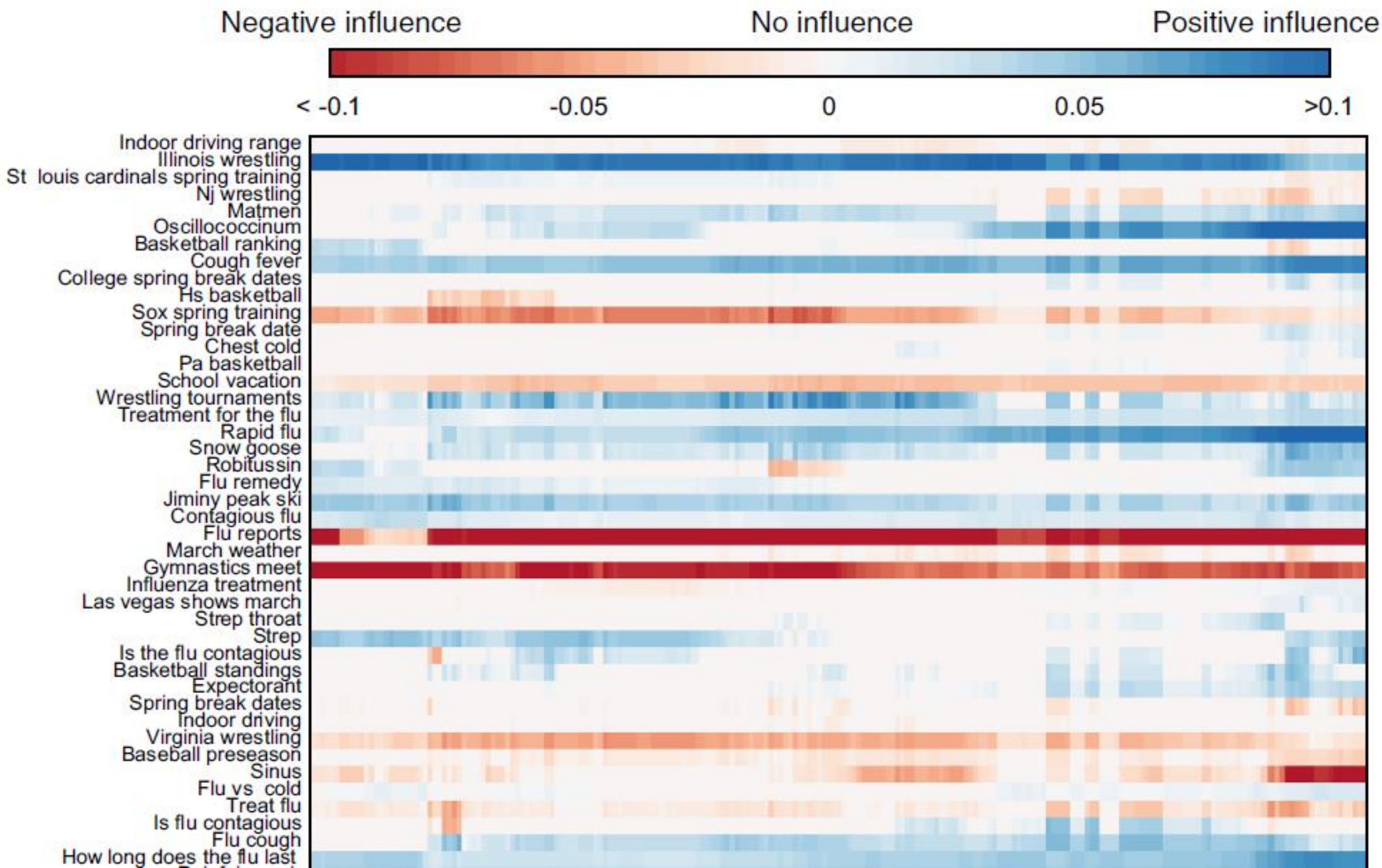
Authors dub "blue team dynamics" - endogenous feedback loops

Measurements themselves may be manipulated

Red team dynamics", as in efforts to top Google searches or Twitter trends

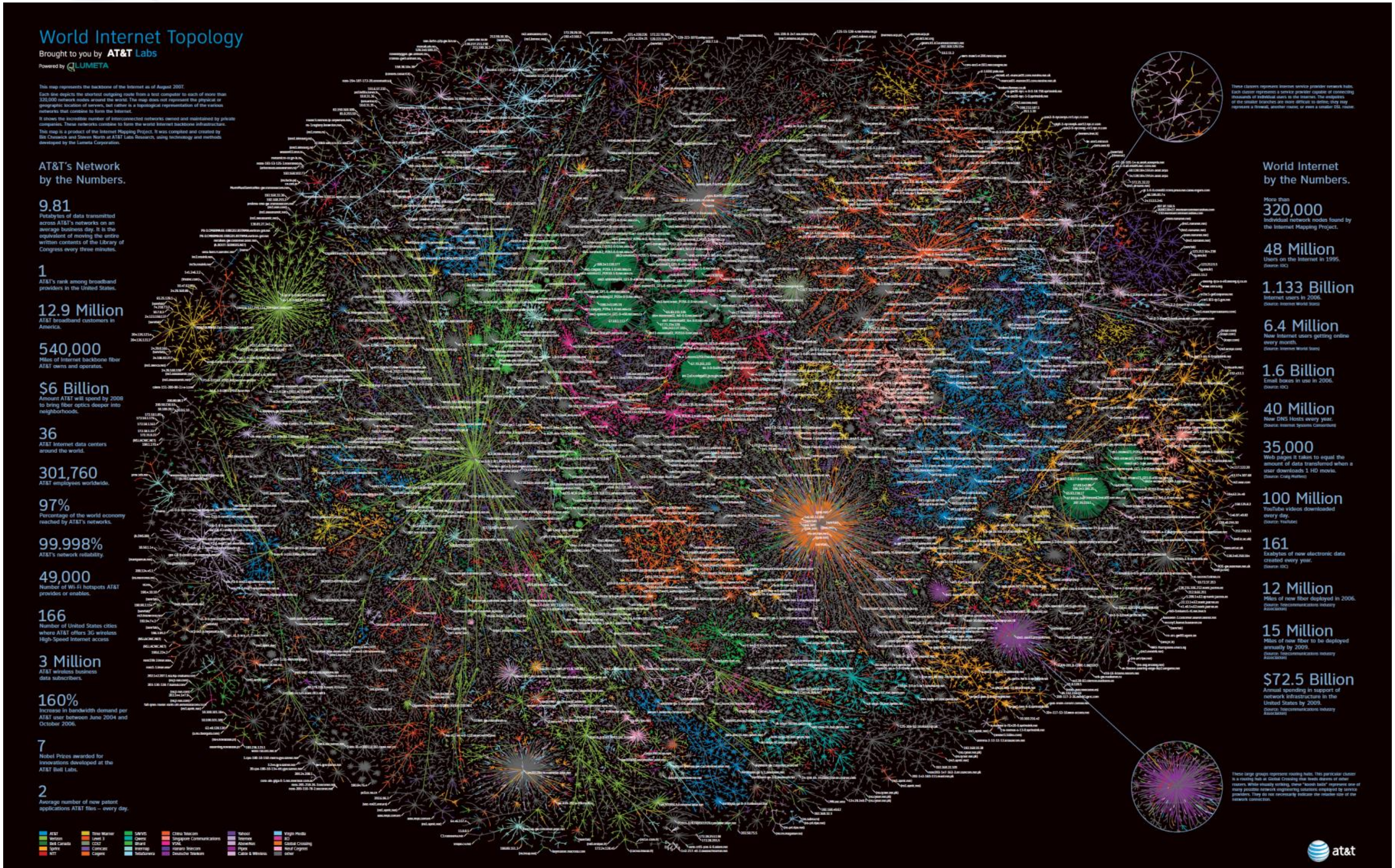
Data visualization and EDA

Most models are not static, with parameters that must be adjusted over time:

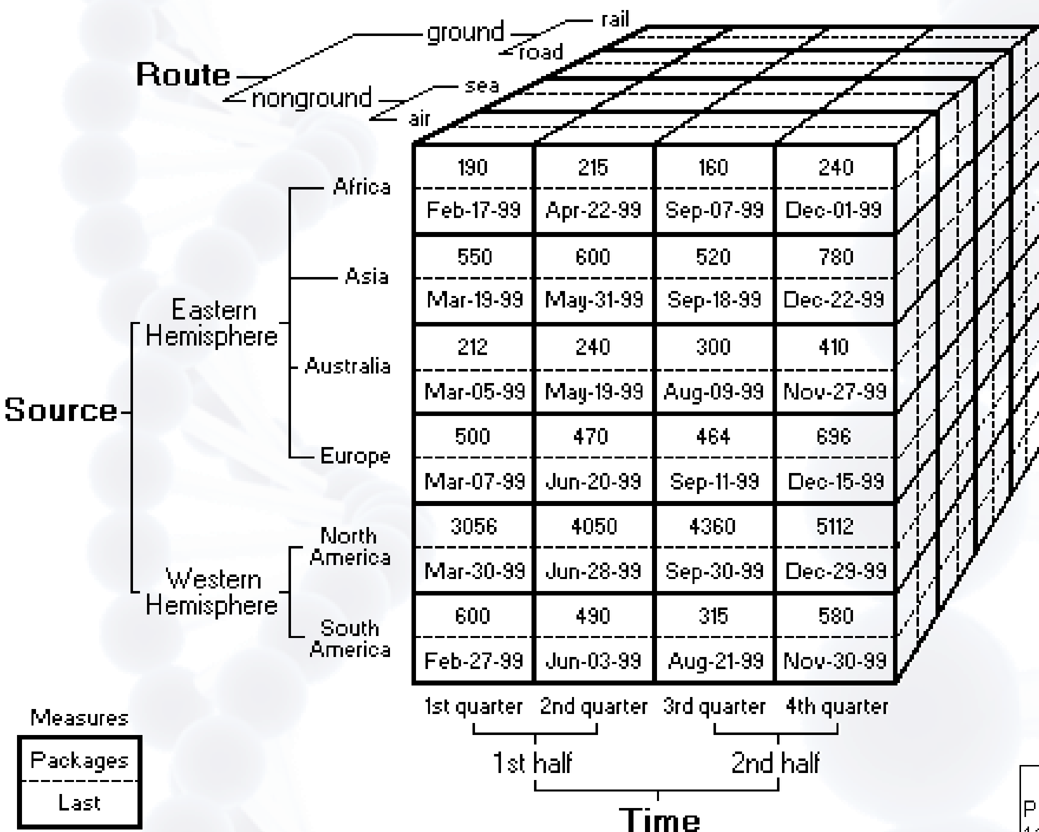


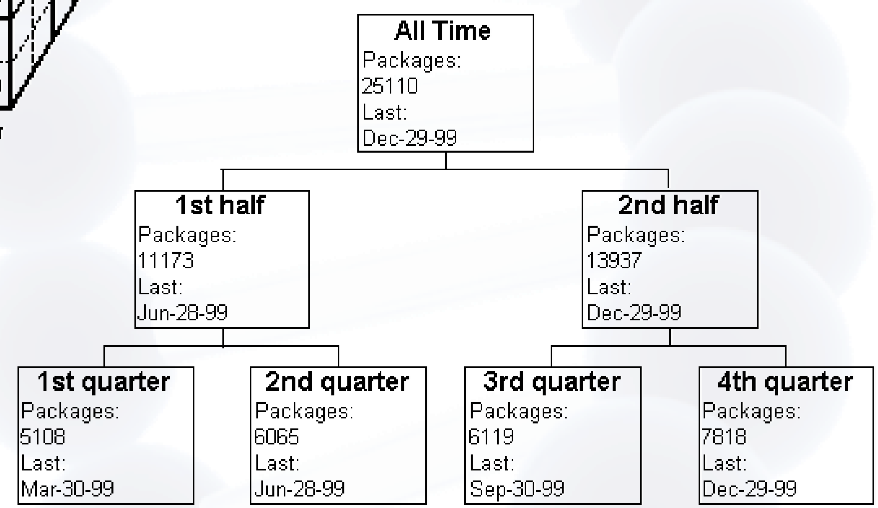
Data visualization and EDA

AT&T Map of Internet structure, circa 2007:



Course topics: Data munging





https://technet.microsoft.com/en-us/library/aa216365

Practical tools

R / RStudio (most focus in the course)

Unix shell (glue between scripts)

Python (alternatives and stable programs)

Packages useful for reshaping data:

From Hadley Wickham:

Tidyr

Plyr/Dplyr

Reshape/Reshape2

From Matt Dowle:

Data.table

Long vs. wide format

month	day	variable	value
5	1	ozone	41
5	2	ozone	36
5	3	ozone	12
5	4	ozone	18
5	5	ozone	NA
5	6	ozone	28

month	day	ozone	solar.r	wind	temp
5	1	41	190	7.4	67
5	2	36	118	8.0	72
5	3	12	149	12.6	74
5	4	18	313	11.5	62
5	5	NA	NA	14.3	56
5	6	28	NA	14.9	66

http://seananderson.ca/

Packages useful for reshaping data:

From Hadley Wickham:

Tidyr

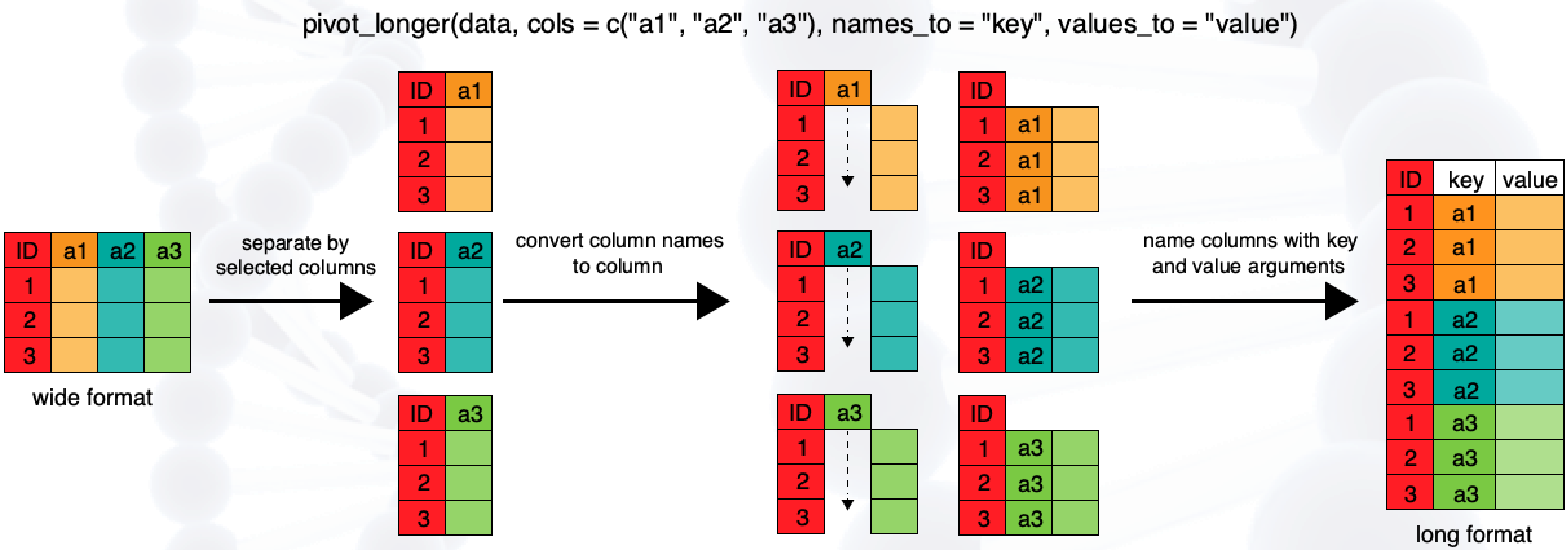
Plyr/Dplyr

Reshape/Reshape2

From Matt Dowle:

Data.table

Long vs. wide format



Resources

Supplementary articles, links, videos on Canvas

Resources

Supplementary texts (most from https://www.bookdown.org/):

1. *R for Data Science*, by Hadley Wickham and Garrett Grolemund
2. *Advanced R*, by Hadley Wickham.
3. Hands-On Programming with R, by Garrett Grolemund
4. *R in Action: Data Analysis and Graphics with R*, by R. Kabacoff.
5. *Practical Data Science with R*, by J. Mount.
6. *Data Science at the Command Line*, by Jeroen Janssens
7. *An Introduction to Statistical Learning,* by G. James *et al.*

(Available at: https://faculty.marshall.usc.edu/gareth-james/ISL/)

Other specialized texts under the "Books" tab on https://bookdown.org.