Introduction to Data Wrangling & Visualization in R

Nelson Roque | 09/24/21

Slides & Code available on GitHub,

SCAN →

UCF

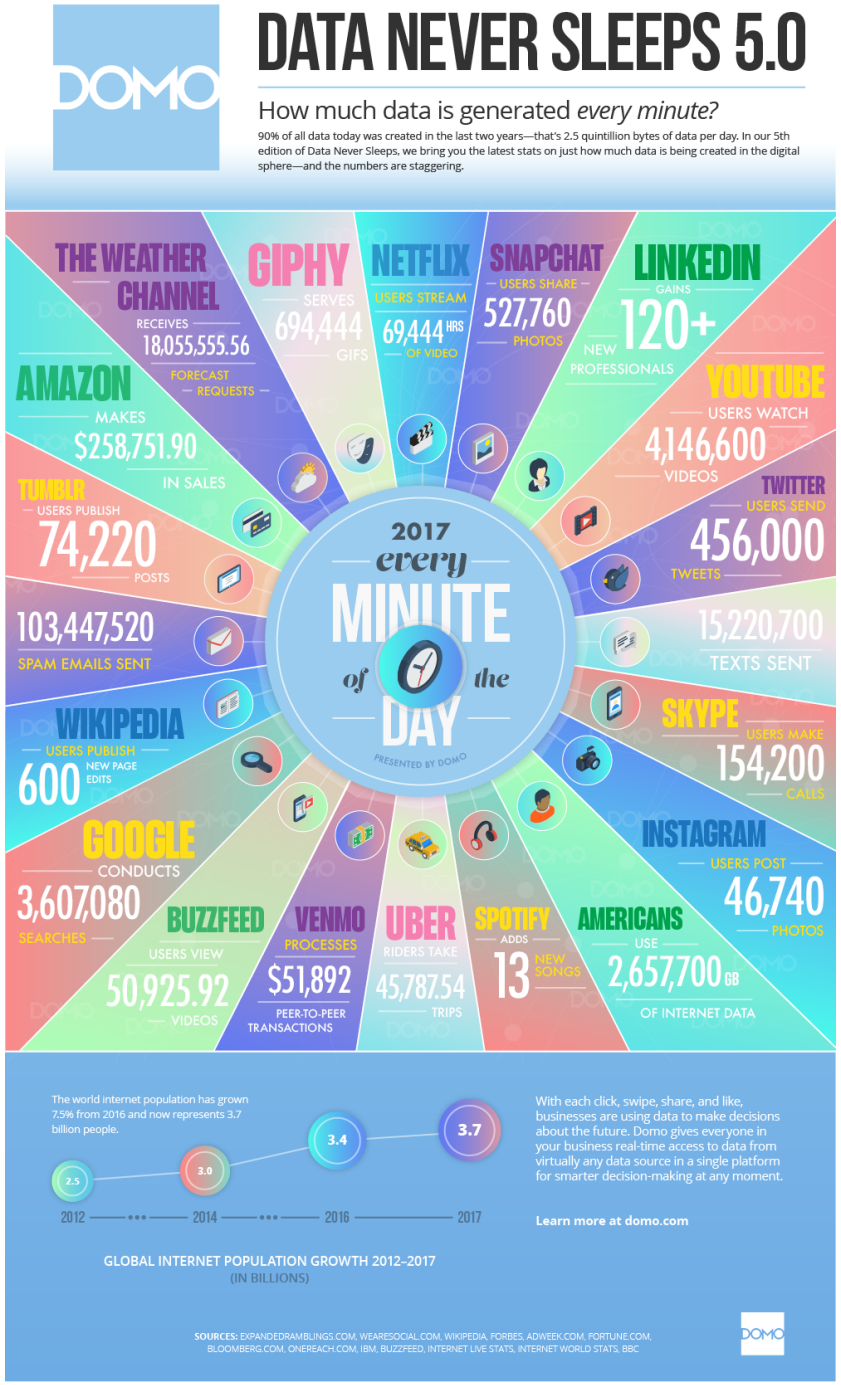
Goals for this Workshop

- Sharing motivation for a code-based approach to data processing
- Learn about data wrangling principles (e.g., tidy data)
- Import, pre-process & visualize data using R
- Share recommended readings, activities for continued learning

Background

- Humans are collectively outputting 2.5 quintillion bytes of data every day; by 2020, each person will generate ~ 1.7 MB every second (IBM Marketing Cloud, 2017).

- At this scale, intensive longitudinal data about humans' behavior facilitates new discovery about the patterning of thought and action and potentially better prediction and optimization of health and well-being.



A Digital Biomarker Approach

- Digital biomarkers are defined as **objective, quantifiable physiological and behavioral data** that are collected and measured by means of digital devices such as portables, wearables, implantables, or digestibles.

- The data collected are typically used to explain, influence, and/or predict health-related outcomes.

Source

A Digital Biomarker Approach: In the News

◆ WSJ NEWS EXCLUSIVE | TECH

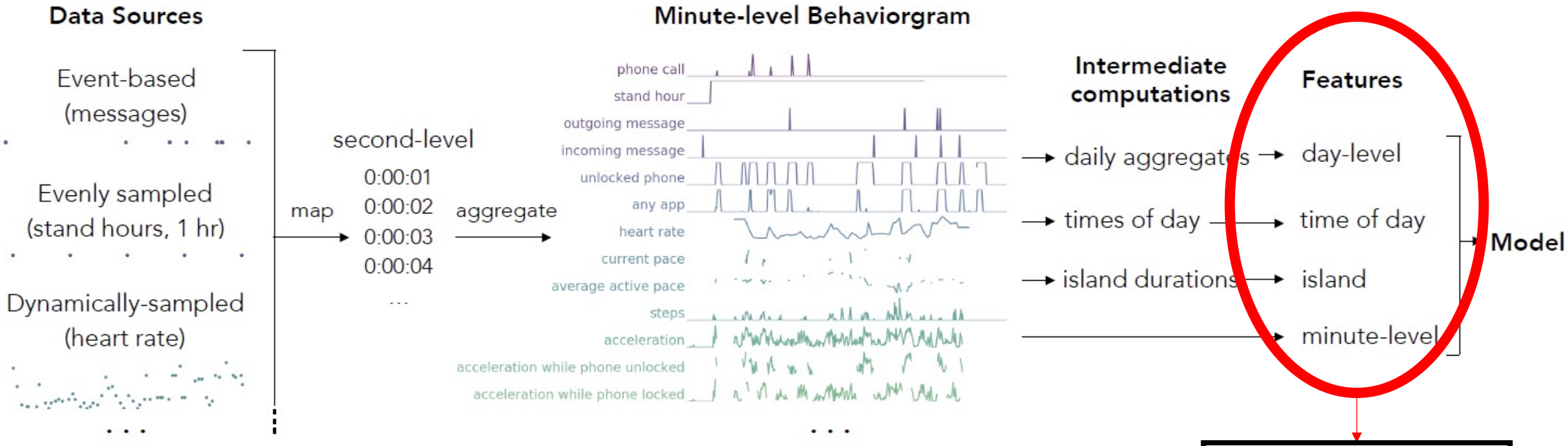
Apple Is Working on iPhone Features to Help Detect Depression, Cognitive Decline

Company is working with UCLA, Biogen to see if sensitive data like facial expressions, typing metrics could signal mental-health concerns

By Rolfe Winkler

Updated Sept. 21, 2021 1:07 pm ET

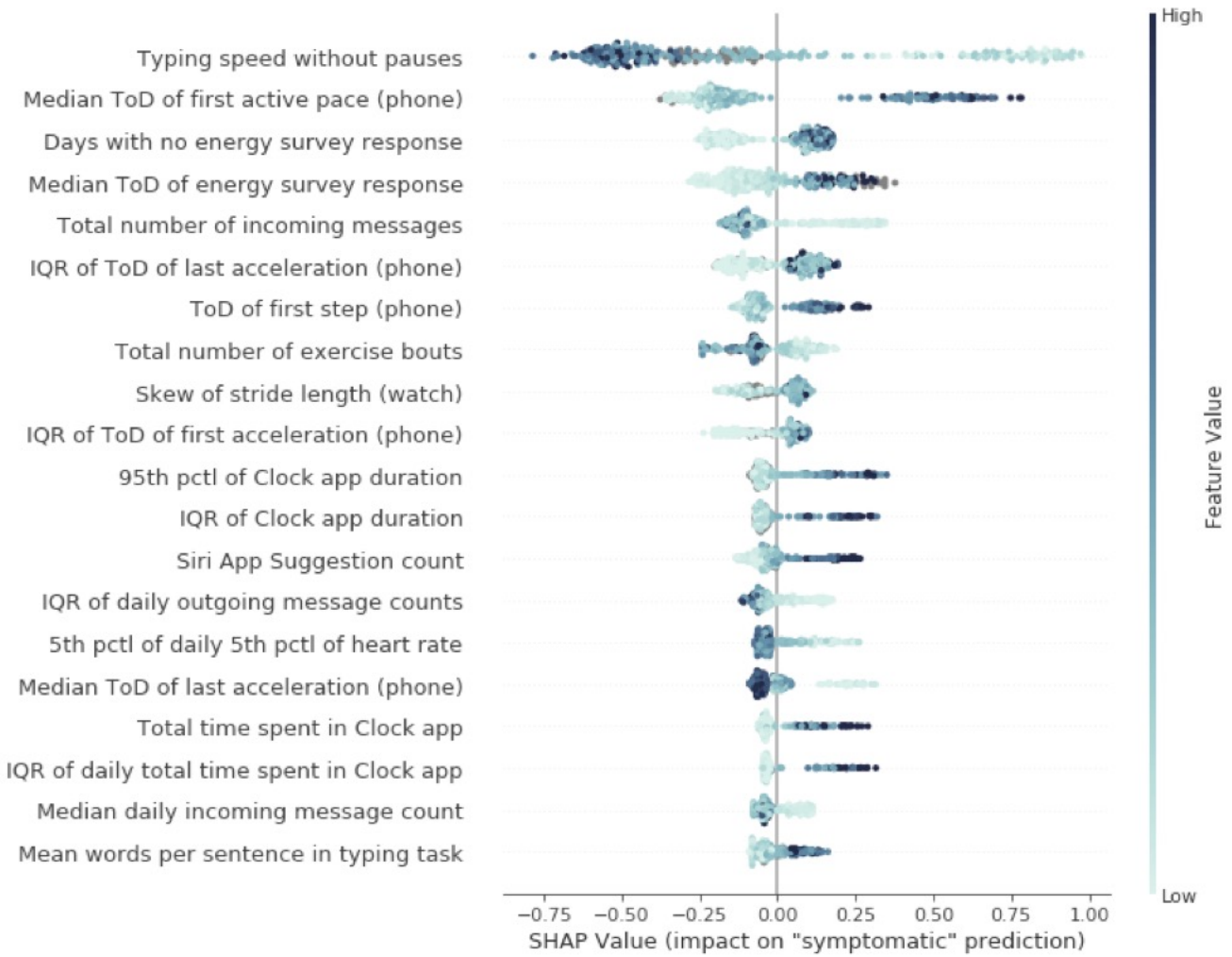
A Digital Biomarker Approach: Visualized



GitHub
tsfeaturex

Image source

Symptomatic participants tended to type slower, exhibit less routine behavior, receive fewer text messages, and spend more time using helper apps than healthy controls.



Data Lifecycle

Measurement → **Pre-processing** → **Analyses**

qualtrics XM

E-Prime 3

jspsych

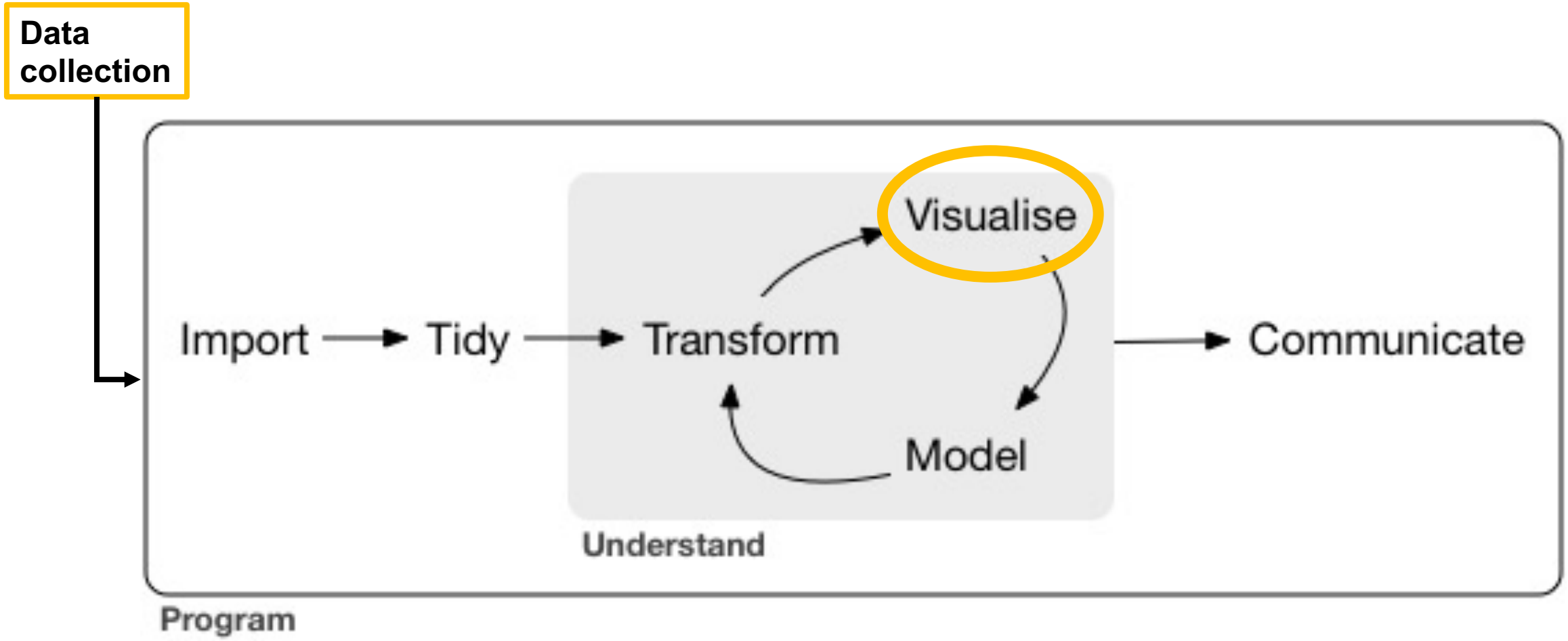
HTML CSS JS

python

sas

Mplus

Typical Data Science Workflow



Source: R for Data Science

What is tidy data?

1. Every column is a variable.
2. Every row is an observation.
3. Every cell is a single value.

participant	condition	avg_response_time	perc_accuracy
9991	control	506	90
9992	active	516	96
9993	control	526	99

Source: R for Data Science

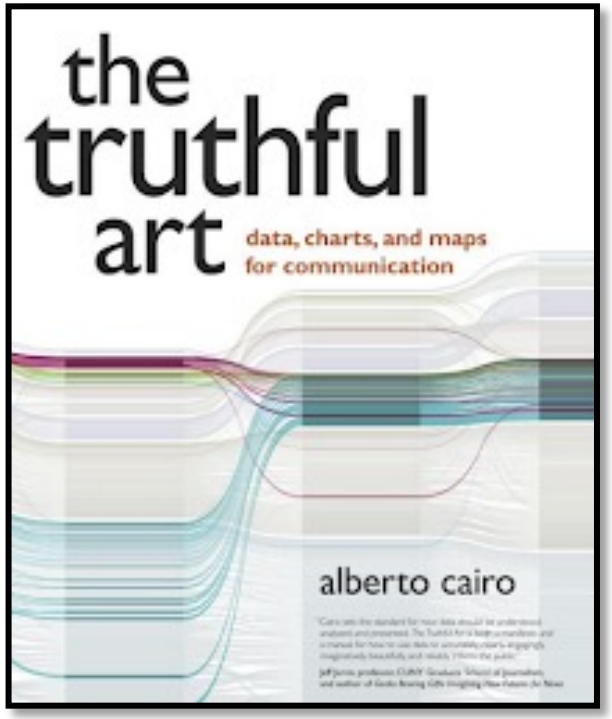
Uncleaned Data Problems

- In a perfect data-generating system, data are already tidy

- In practice, we must first determine
  - data structure issues, including:
    - Column headers are values, not variable names.
    - Multiple variables are stored in one column.
    - Variables are stored in both rows and columns.
    - Multiple types of observational units are stored in the same table.
    - A single observational unit is stored in multiple tables.
  - Data quality issues
    - Values out of range
    - Improperly/inconsistently coded response options
    - Inconsistent records counts

Five Qualities of Great Visualizations

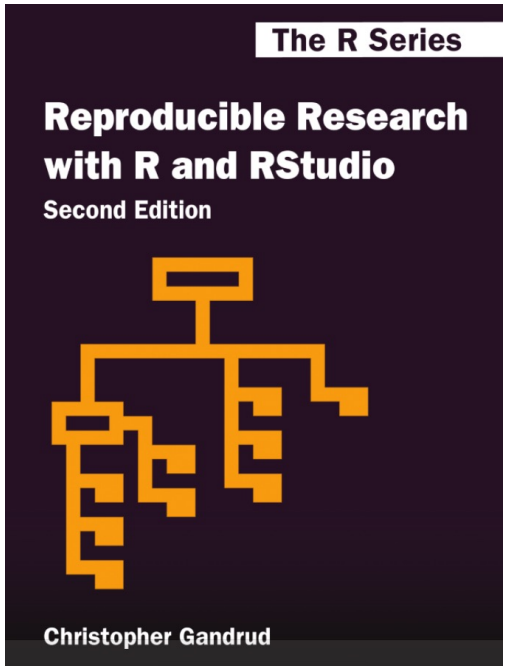
1. It is truthful
2. It is functional
3. It is beautiful
4. It is insightful
5. It is enlightening



Reproducible Workflows

- Documented, code-based approach to pre-processing and analytics
  - R (RMarkdown), Python (Jupyter notebooks)

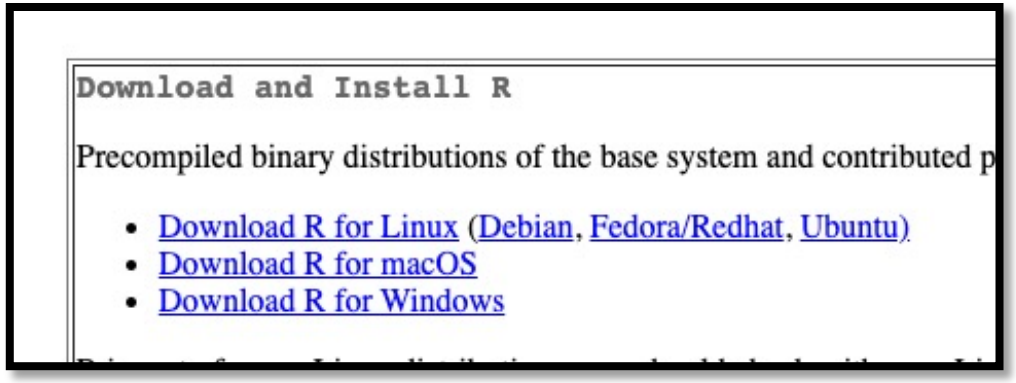
- File versioning with clear changelogs
  - Git and Github



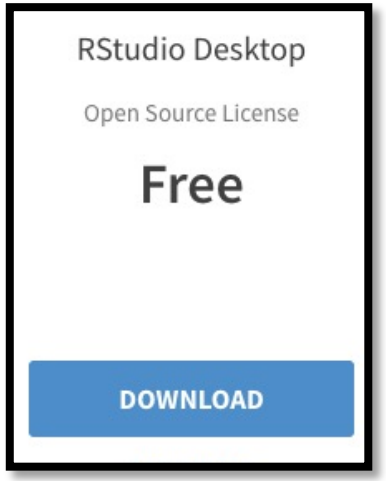
https://monashdatafluency.github.io/r-rep-res/

Software Required for Today's Workshop

- R - https://cran.r-project.org/



- RStudio - https://www.rstudio.com/products/rstudio/download/



The Data Source: Google Mobility

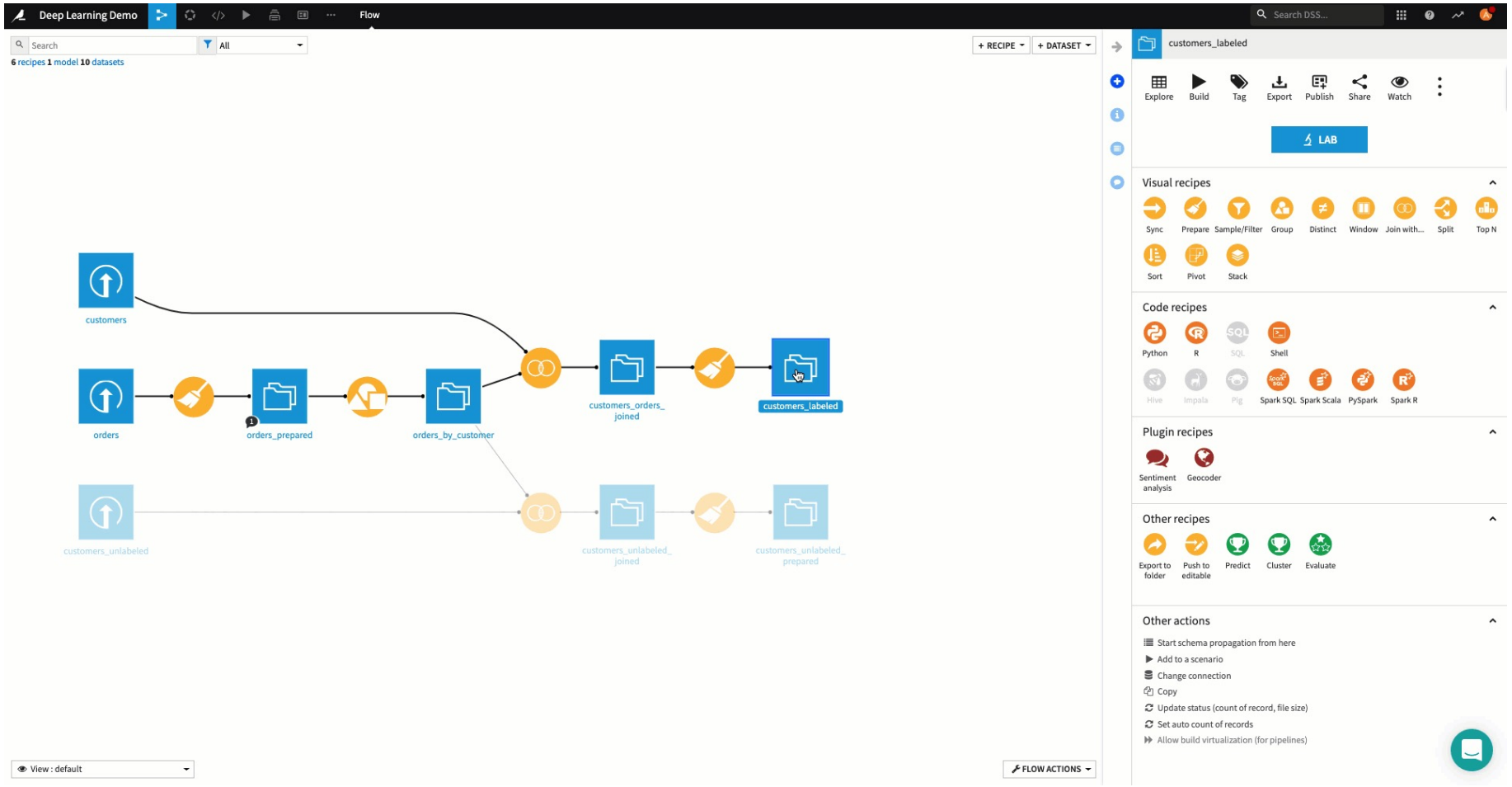
Learn more about this data
https://www.google.com/covid19/mobility/data_documentation.html?hl=en

Time for a Demo!

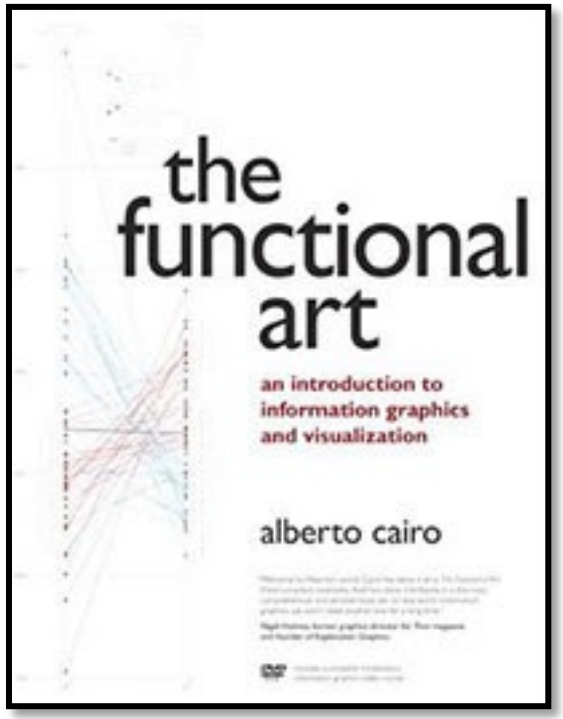
Scan the QR code to view the code we will work with

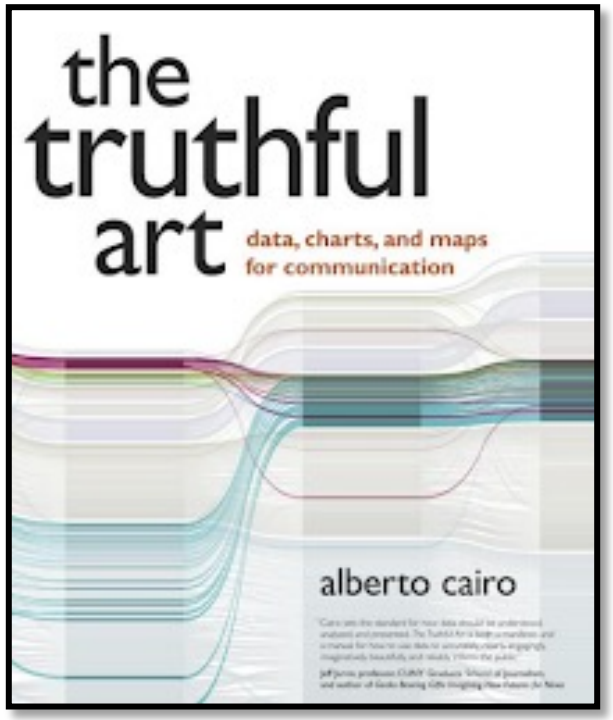
The Future of Data Workflows

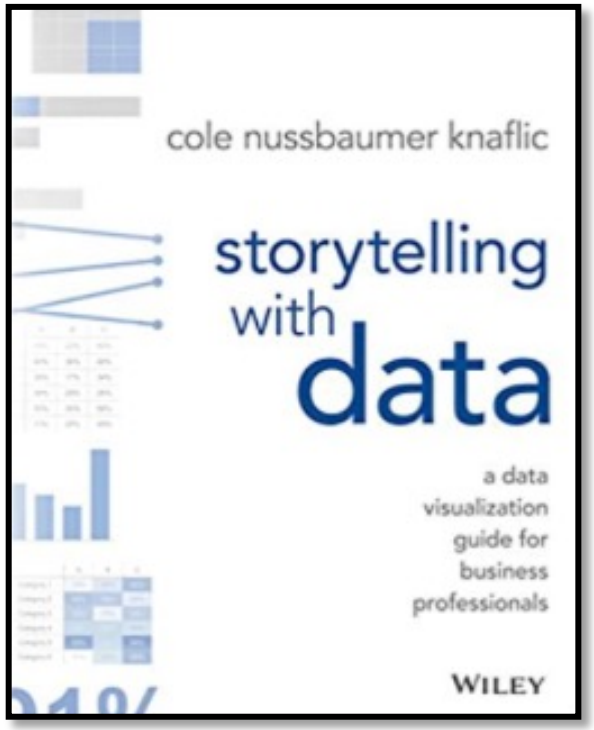
Seamless integration platforms, for example:



Reading Recommendations

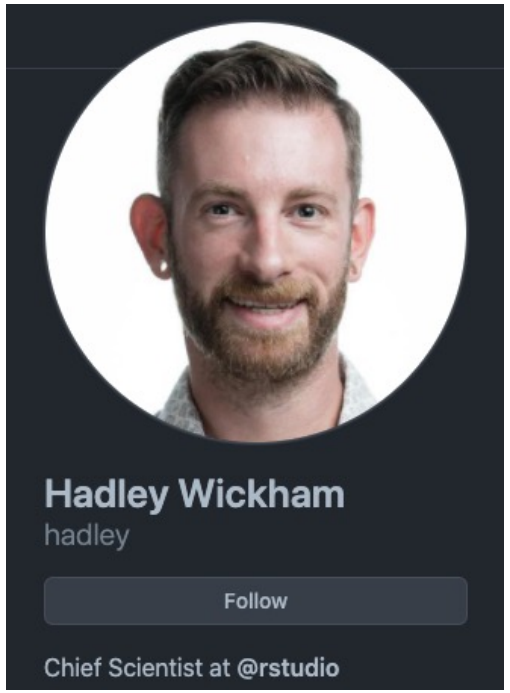






Tutorial Recommendations

- Anything by Hadley Wickham



- https://www.r-bloggers.com/
- https://www.r-statistics.com/
- https://blog.revolutionanalytics.com/
- https://r-charts.com/
- http://www.cookbook-r.com/Graphs/
- https://plotly.com/r/

Download Cheatsheets

- Data Import
  - https://github.com/rstudio/cheatsheets/raw/master/data-import.pdf)
- Data Wrangling Cheatsheet
  - https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf)
- Data Transformation with dplyr
  - https://github.com/rstudio/cheatsheets/raw/master/data-visualization.pdf)
- String Manipulation
  - https://github.com/rstudio/cheatsheets/raw/master/strings.pdf)
- Work with dates/times
  - https://github.com/rstudio/cheatsheets/raw/master/lubridate.pdf)
- R Markdown
  - https://github.com/rstudio/cheatsheets/raw/master/rmarkdown.pdf)
- More cheatsheets
  - https://www.rstudio.com/resources/cheatsheets/)

Download Open Data

- Kaggle
  - https://www.kaggle.com/datasets
- Data.gov
  - https://www.data.gov/

thank you

nelson.roque@ucf.edu

nelsonroque.com

https://github.com/nelsonroque