Data Handling

12 October 2021

James Brenton and Mark Fernandes

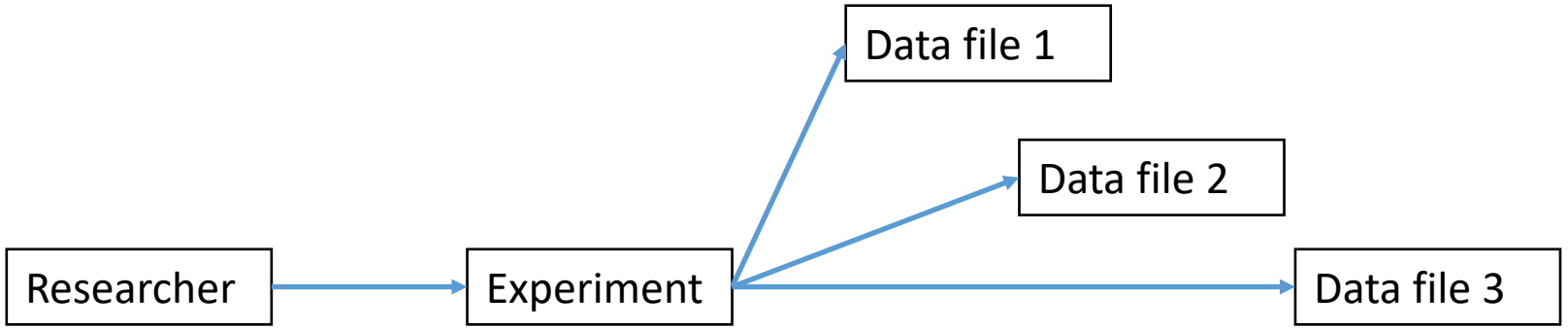
Goals

- Agreement that these are critical (!)
  - **Strong** data exploration skills
  - OpenRefine -> dplyr -> ggplot workflow
  - R markdown document for all lab results
  - Aspiration for high quality data carpentry (version control with git)

- Understanding
  - Limitations of excel
  - Good working practices upstream of R and other tools
  - Tidy data
  - Joins (introduction)

Develop **safe data practices** for handling your primary data

- keep primary data **unaltered**
- Use **consistent** identifiers (ISO date format e.g. 2020-10-12; integer identifiers e.g. JBLAB-XXXX)
- What are the **relationships** between your experiments and data?
- Use an electronic notebook
- Have a backup strategy
- Use a **database**



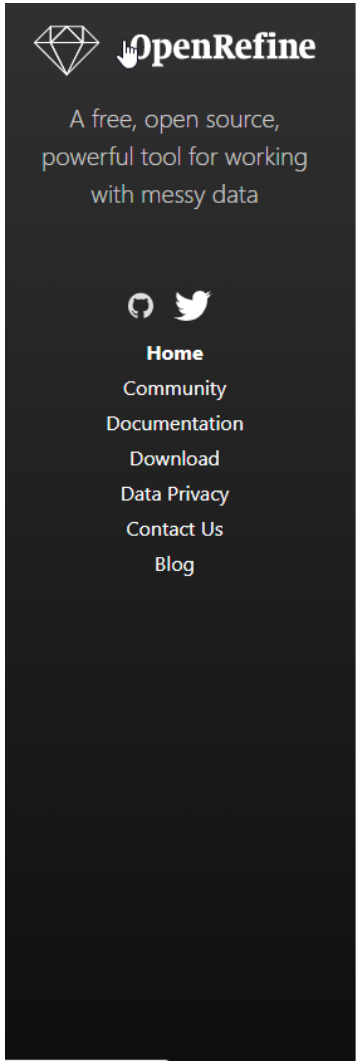
The 13 **critical** rules for using Excel

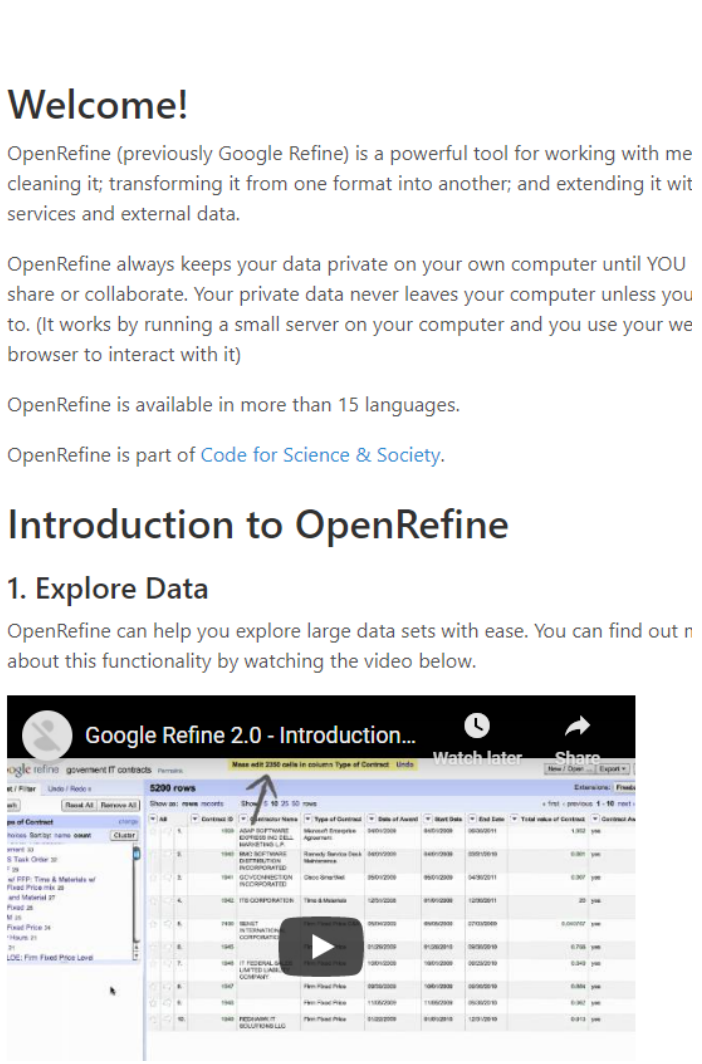
https://kbroman.org/dataorg/

Get good at data cleaning and exploratory data analysis

- **Don’t use Excel** for data cleaning
- Learn OpenRefine

https://openrefine.org/





Use R for your analyses

R for Data Science

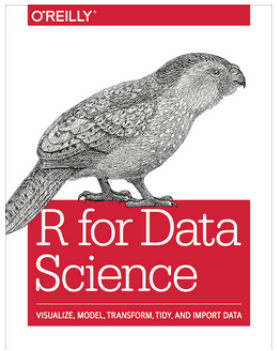
Garrett Grolemund

Hadley Wickham

Welcome

This is the website for **"R for Data Science"**. This book will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you'll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You'll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You'll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.

To be published by O'Reilly in late 2016. Pre-order from amazon.



http://r4ds.had.co.nz/

R Graphics Cookbook, 2nd edition

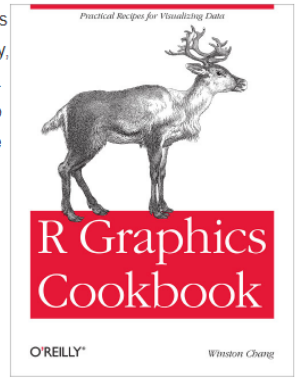
Winston Chang

2020-04-03

Welcome

Welcome to the **R Graphics Cookbook**, a practical guide that provides more than 150 recipes to help you generate high-quality graphs quickly, without having to comb through all the details of R's graphing systems. Each recipe tackles a specific problem with a solution you can apply to your own project, and includes a discussion of how and why the recipe works.

Read online here for free, or buy a physical copy on Amazon.

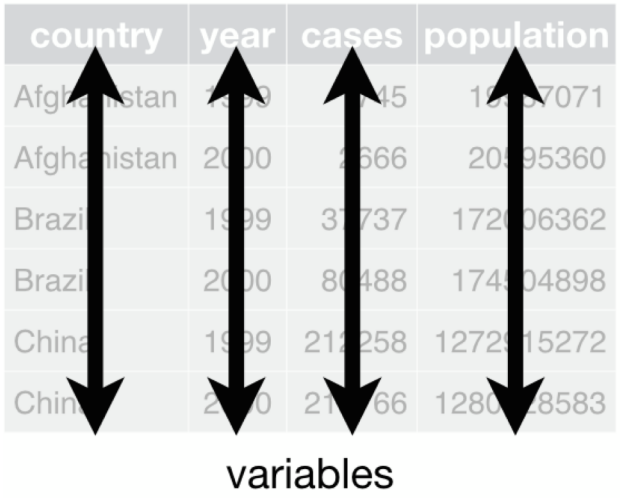


https://r-graphics.org/

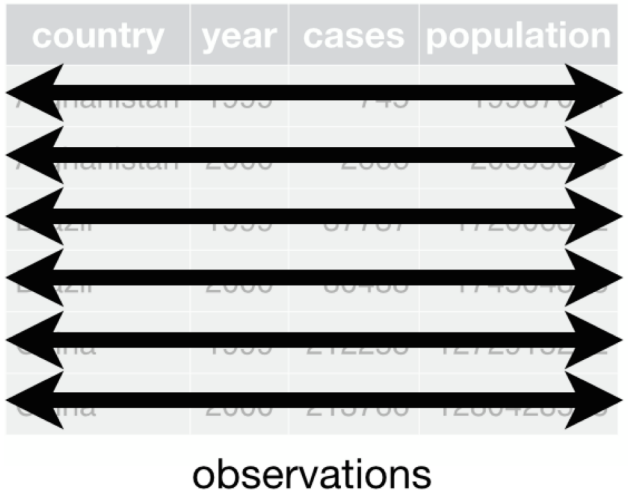
Make your data tidy (this saves you time!)

“Happy families are all alike; every unhappy family is unhappy in its own way” — Leo Tolstoy
“Tidy datasets are all alike, but every messy dataset is messy in its own way” — Hadley Wickham

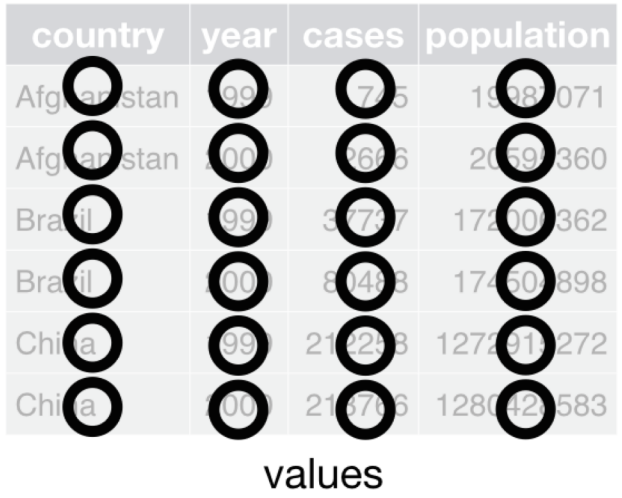
Each variable must have its own column



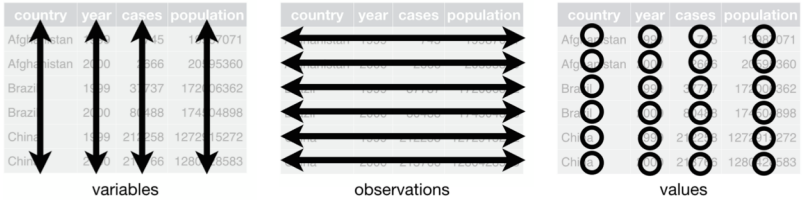
Each observation must have its own row



Each value must have its own cell



Why use tidy data?



1. Provides a consistent way of storing data for your analyses

2. Consistent data structures make it much easier to learn the tools that work with it because they have an underlying **uniformity**

3. Placing variables in columns allows R (and other tools) to work efficiently. Most built-in R functions, work with vectors of values.

4. You want to make transforming (tidy) data feel natural and easy

Use a database for your data

Databases and SQL

In the late 1920s and early 1930s, William Dyer, Frank Pabodie, and Valentina Roerich led expeditions to the Pole of Inaccessibility in the South Pacific, and then onward to Antarctica. Two years ago, their expeditions were found in a storage locker at Miskatonic University. We have scanned and OCR the data they contain, and we now want to store that information in a way that will make search and analysis easy.

Three common options for storage are text files, spreadsheets, and databases. Text files are easiest to create, and work well with version control, but then we would have to build search and analysis tools ourselves. Spreadsheets are good for doing simple analyses, but they don't handle large or complex data sets well. Databases, however, include powerful tools for search and analysis, and can handle large, complex data sets. These lessons will show how to use a database to explore the expeditions' data.

Prerequisites

- This lesson requires the Unix shell, plus SQLite3 or DB Browser for SQLite.
- Please download the database we will use: survey.db

Schedule

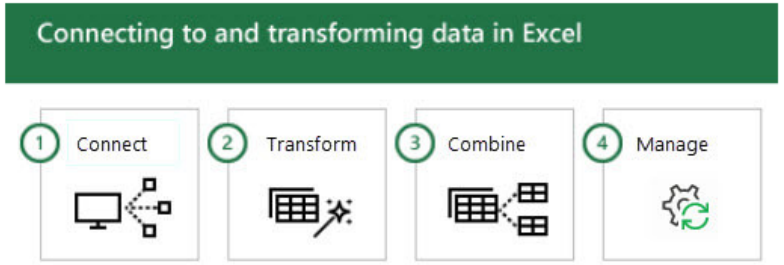
	Setup	Download files required for the lesson
00:00	1. Selecting Data	How can I get data from a database?
00:15	2. Sorting and Removing Duplicates	How can I sort a query's results? How can I remove duplicate values from a query's results?
00:35	3. Filtering	How can I select subsets of data?
00:55	4. Calculating New Values	How can I calculate new values on the fly?
01:05	5. Missing Data	How do databases represent missing information? What special handling does missing information require?
01:35	6. Aggregation	How can I calculate sums, averages, and other summary values?
01:55	7. Combining Data	How can I combine data from multiple tables?
02:35	8. Data Hygiene	How should I format data in a database, and why?
03:05	9. Creating and Modifying Data	How can I create, modify, and delete tables and data?
03:30	10. Programming with Databases - Python	How can I access databases from programs written in Python?

Excel has good features

Getting Started with Get & Transform in Excel

Excel for Microsoft 365, Excel 2019, Excel 2016

With **Get & Transform** in Excel, you can search for data sources, make connections, and then *shape* that data (for example remove a column, change a data type, or merge tables) in ways that meet your needs. Once you've shaped your data, you can share your findings or use your query to create reports.

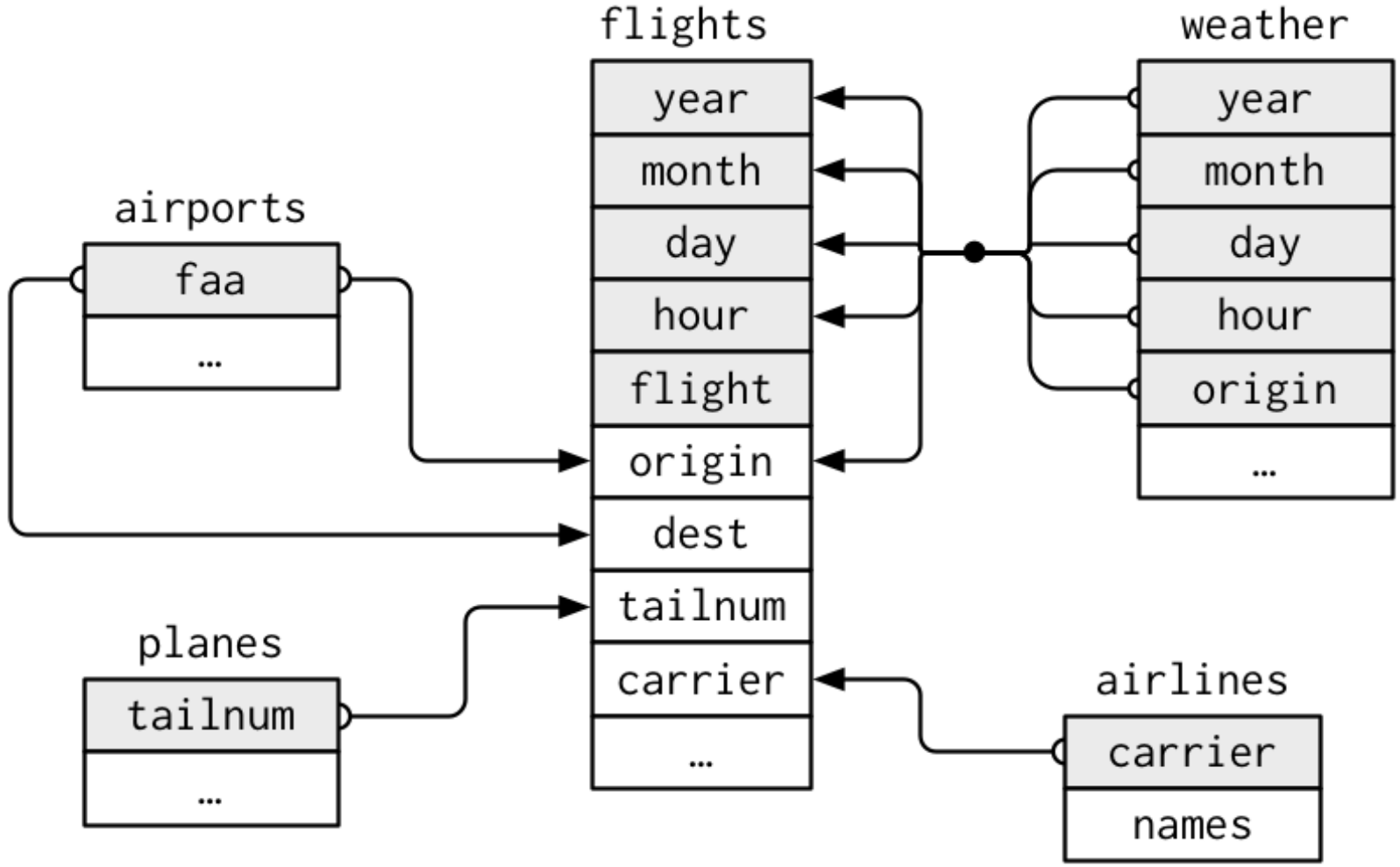


Looking at those steps in order, they often occur like this:

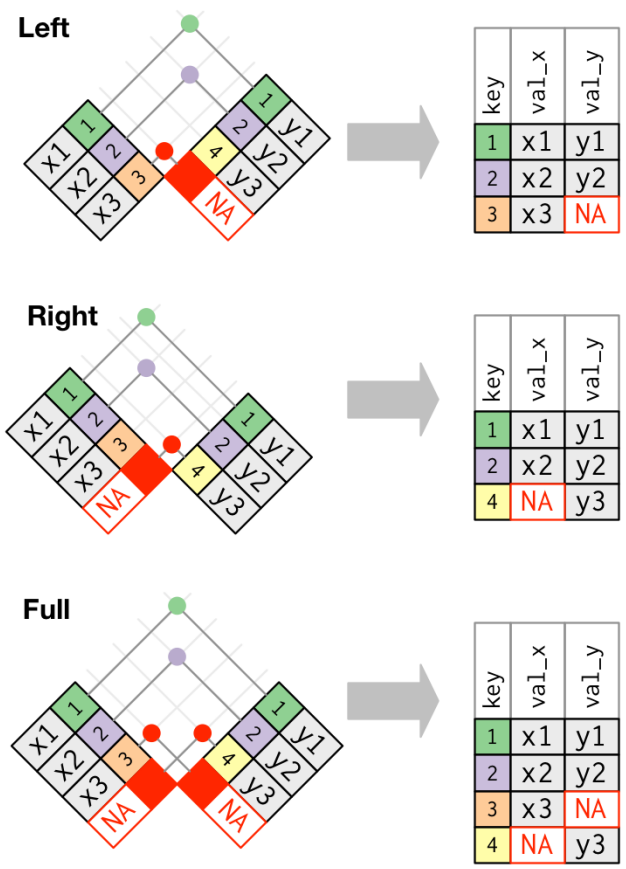
- **Connect** – make connections to data sitting in the cloud, in a service, or locally
- **Transform** – shape the data to meet your needs; the original source remains unchanged
- **Combine** – create a data model from multiple data sources, and get a unique view into the data
- **Manage** – once your query is complete you can save it, copy it, or use it for reports

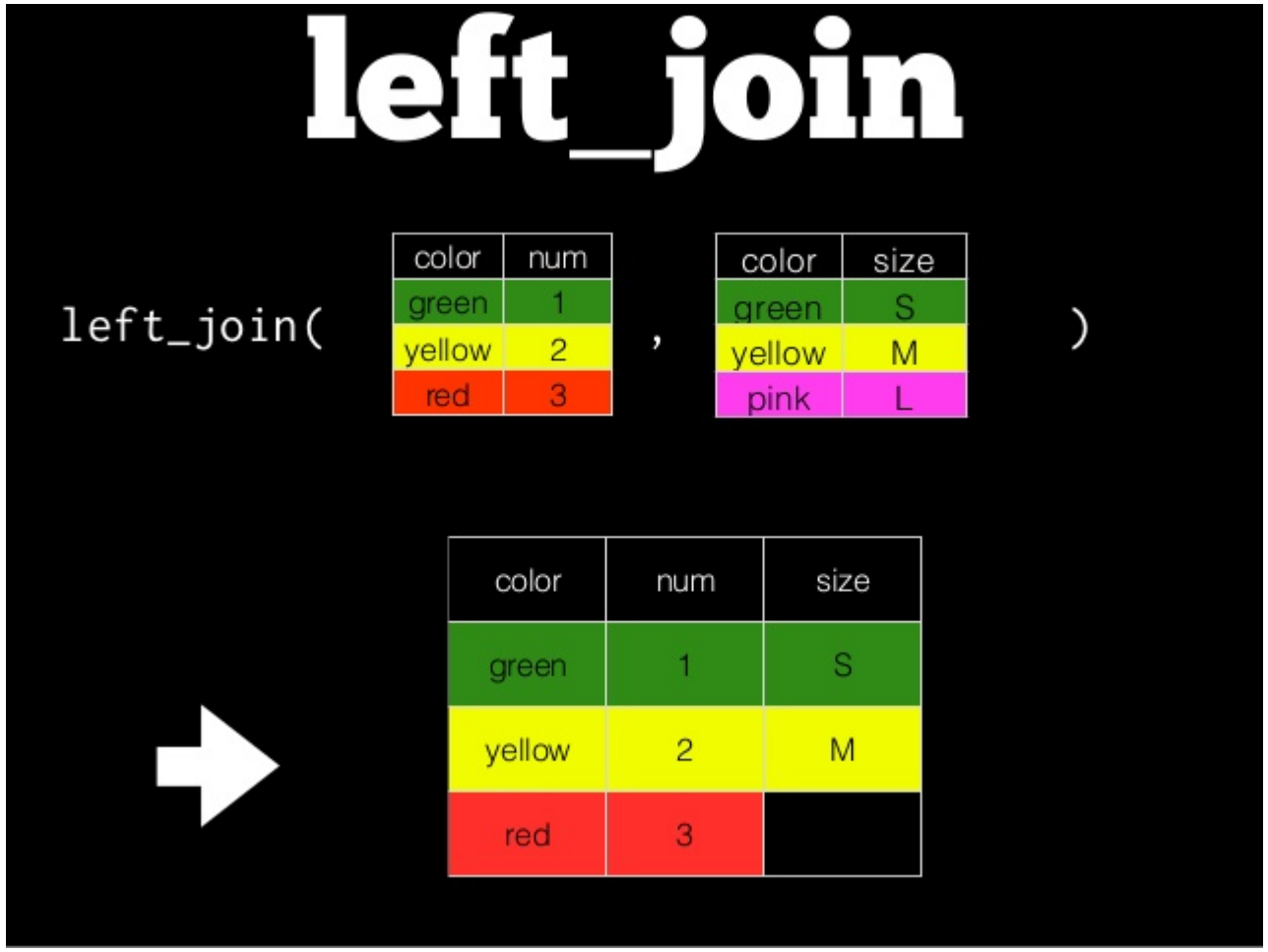
https://support.microsoft.com/en-us/office/getting-started-with-get-transform-in-excel-a8310388-2a12-438c-9d29-c6d29cb8df6a

Think relationally! Joining data is a common task



Dplyr has powerful joins





Summary

1. Know limitations (and strengths) of Excel http://kbroman.org/dataorg/
2. Use RStudio and make a R markdown document for every analysis you do
3. Focus on using dplyr, tidyr and ggplot for everything (or at least mostly everything) and learn all the functionality in these powerful libraries
4. Develop reliable data carpentry skills (version control, joins, programmatic analysis)