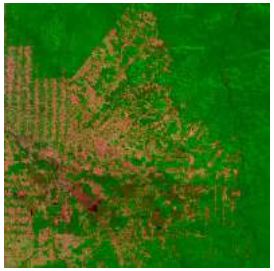
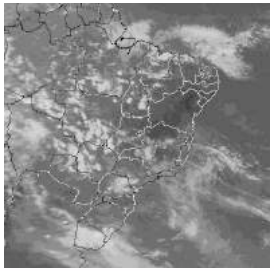




CAP-394
INTRODUCTION TO DATA SCIENCE

Rafael Santos – rafael.santos@inpe.br
Gilberto Ribeiro – gilberto.queiroz@inpe.br
www.lac.inpe.br/~rafael.santos/cap394.html

Updated in 2019

About this Lecture

Where are we?

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

Statistical Models

Create Data Product

Visualize and Report

Make Decisions

Based on ***Doing Data Science***,
Rachel Schutt and Cathy O'Neil, OReilly, 2014.

Introduction to Data Science

Raw Data and Tidy Data

Raw Data

- Data in databases, spreadsheets…
- Images, videos, audio…
- Time series...
- Logs, text, JSON files, XML files…

Based on Coursera's "Getting and Cleaning Data" course.

Tidy Data

- One table with all the data (or linked tables).
  - Each variable in its column.
  - Each observation in its row.
  - Variable names in the first row, with good, clear names.
- "Tidiness" is not an absolute feature!
  - Depends on what we have and what we want to do.

incidentLocation	location
4000 MT PLEASANT AV	(39.2910056,-76.5636502)
1000 W BALTIMORE ST	(39.2889640,-76.6339440)
1000 STAMFORD RD	(39.2963480,-76.7101510)

Based on Coursera's "Getting and Cleaning Data" course.

Raw Data to Tidy Data

- Create a Code Book that describes how we can get from Raw Data to Tidy Data.
- A simple (formatted) text file with:
  - Sources for the raw data.
  - More detail on the variables.
  - What was {selected, enhanced, preprocessed} and how.
  - Instruction on how the data was processed.
- Code Books essential to reproducibility!

Based on Coursera's "Getting and Cleaning Data" course.

Introduction to R

“R Programming”

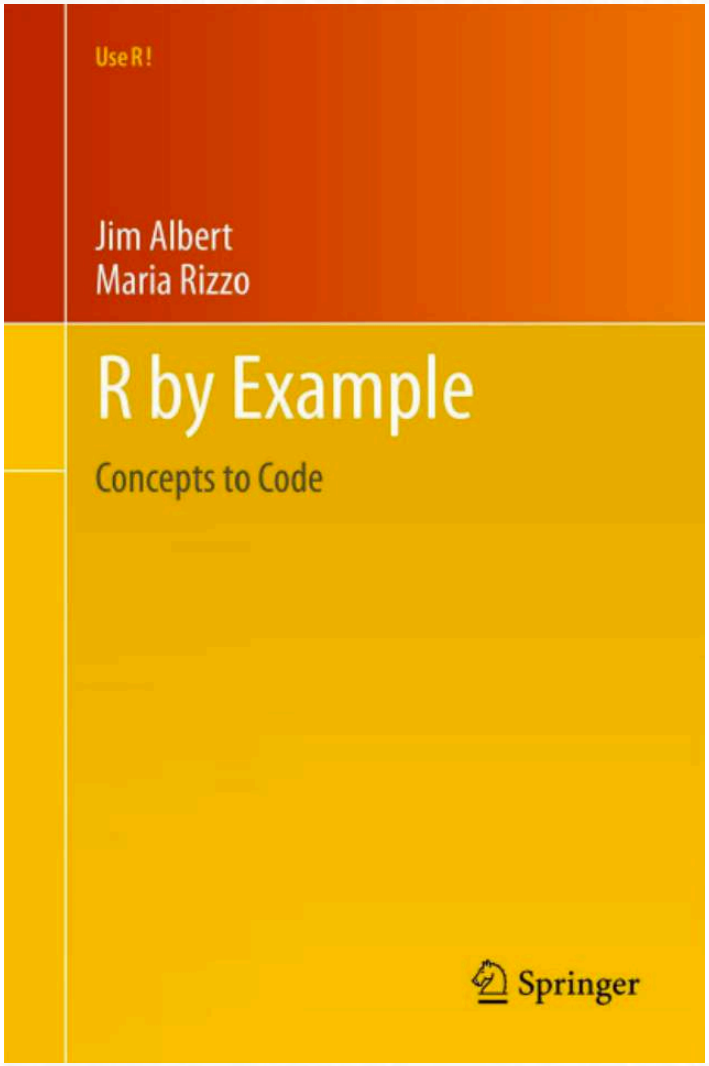
Introduction to R

- https://github.com/rafaeldcsantos/CAP-394

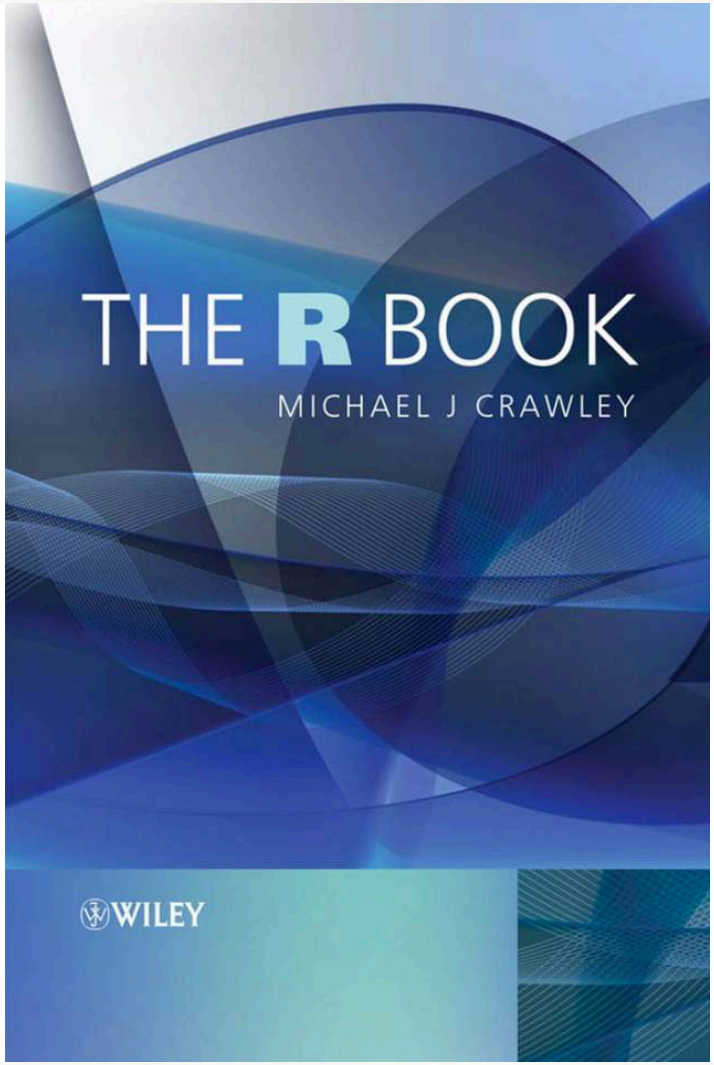
References

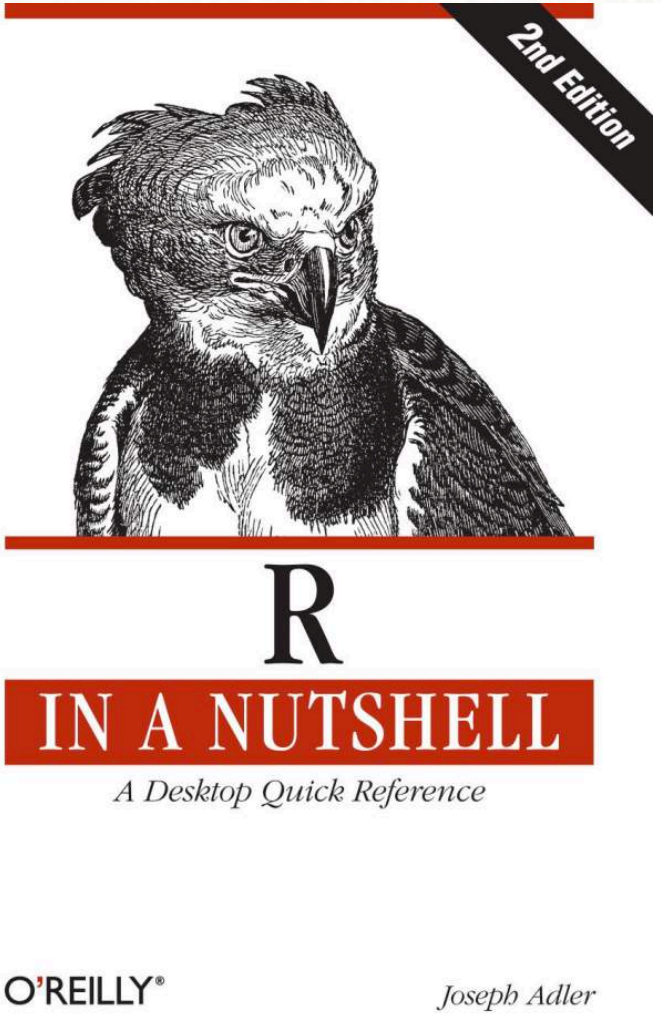
References


Proven Recipes for Data Analysis, Statistics, and Graphics
R
Cookbook
O'REILLY®
Paul Teetor

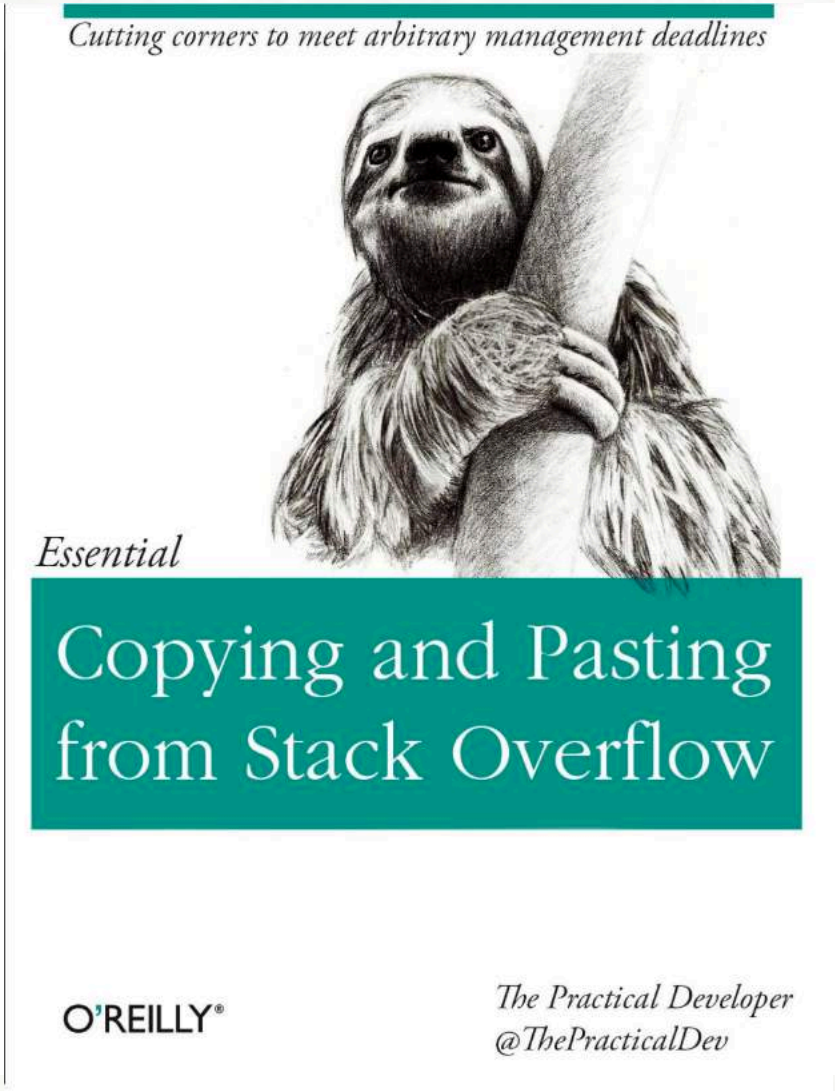

Use R!
Jim Albert
Maria Rizzo
R by Example
Concepts to Code
Springer

References


THE R BOOK
MICHAEL J CRAWLEY
WILEY


2nd Edition
R
IN A NUTSHELL
A Desktop Quick Reference
O'REILLY®
Joseph Adler

References


Cutting corners to meet arbitrary management deadlines
Essential
Copying and Pasting from Stack Overflow
O'REILLY®
The Practical Developer
@ThePracticalDev

Your Project

Research in Data Science

- Basic
  - New algorithms and variations.
  - Reference implementations.
  - Support tools (e.g. databases, data access, abstraction, automation).
- Applied
- Get your data, start doing:
  - Cleaning, munging.
  - EDA, visualization.
  - Basic model creation.
- …with real data, in a reproducible way.

Are we going the Basic way?

- There is ***nothing*** basic to it!

1. Develop a new (or derived) algorithm.
   - Justify why!
   - Package and document it extensively!
   - Add reproducible test cases!
   - Publish your code!
2. Develop a data access or analysis tool.
   - Better if it is thematic.
   - Document, add test cases, etc.

Are we going the Applied way?

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

Statistical Models

Create Data Product

Visualize and Report

Make Decisions

Based on ***Doing Data Science***,
Rachel Schutt and Cathy O'Neil, OReilly, 2014.

A Comic Book guide to the Applied Data Science way

1. Think about your data.
   - Where is it? How can you access it?
   - How is it stored, formatted?

2. Which are interesting questions about it?
   - Can we answer those questions with the data or do we need more data?
   - Can we get more data?

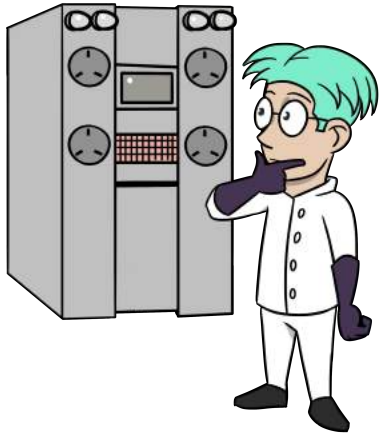
A Comic Book guide to the Applied Data Science way

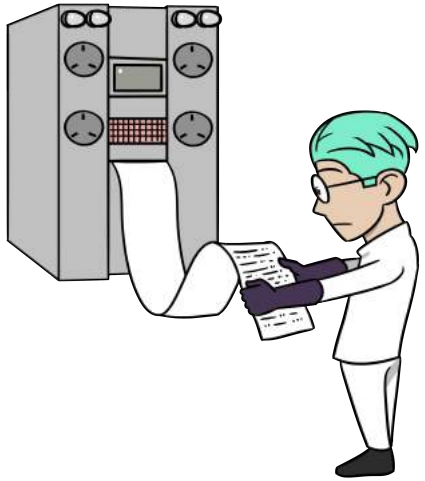
3. Create code to explore it.

   - Explore its structure, completeness, features.
   - Do some basic statistics, EDA, visualization.
   - Don’t worry about failures in the code:
     worry about failures in the data!

A Comic Book guide to the Applied Data Science way

4. Consider hypotheses about your data.
   - Make sure it makes (at lease some) sense!
   - Check the data again!
   - Learn how to create models.
   - Apply and evaluate models on your data.

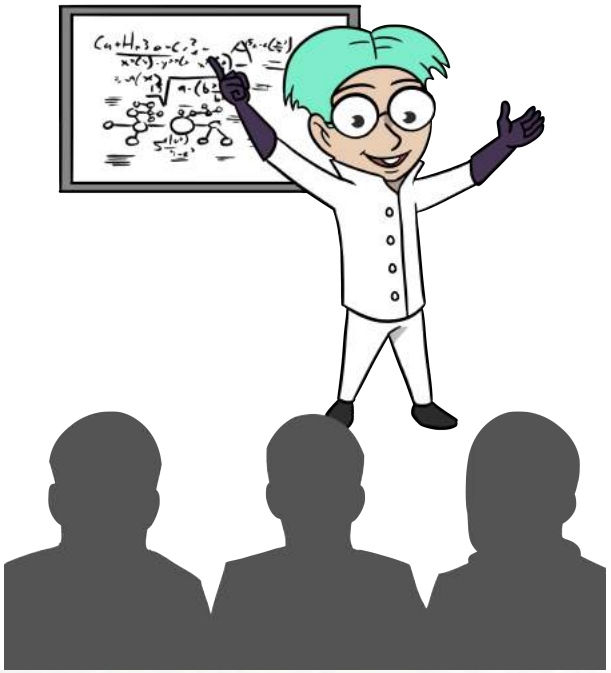




A Comic Book guide to the Applied Data Science way

5. Communicate and document your results.

- Intermediary results if they help to tell a story about the data.
- Even bad results if they can teach us something!
- Have you been using notebooks?
- Can you create a new data product?



A Comic Book guide to the Applied Data Science way

- A *very common* alternative approach: