Columbia University
in the City of New York

DSFBA: Introduction

Data Science for Business Analytics

Thibault Vatter

Department of Statistics, Columbia University

09/16/2020

Outline

1 Introduction

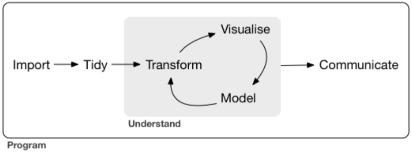
2 Organization

3 R

A little about me

- Born and raised in Geneva
- Education:
  - B.Sc. Physics (EPFL, '10)
  - M.Sc. Physics with minor in Financial Engineering (EPFL, '12)
  - Ph.D. Statistics (HEC Lausanne, '16)
- Worked a bit as a quant in finance
- Currently:
  - Assistant Professor in Statistics at Columbia University
  - Live in New York
- Hobbies:
  - Flying planes
  - Watching bay area teams (go 49ers and Warriors!)
  - Running
  - Beers (formerly at Satellite, now in Brooklyn micro-breweries)

What you will learn



- **Import** data from the web, a database, a stored file, etc.
- **Wrangle**:
  - **Tidy**: usually means that rows/columns are observations/variables.
  - **Transform**: narrowing in on observations of interests, creating new variables, calculating summary statistics.
- **Analyze**:
  - **Visualize**:
    - E.g., show unexpected things, or raise new questions.
    - Doesn't scale well as it requires human interpretation.
  - **Model**:
    - Sufficiently precise questions can be answered with a model.
    - Mathematical/computational tools generally scale well.
    - Even when it doesn't, computers are usually cheaper than brains!
- **Communicate** your results.
- Surrounding all these tools is **programming**.

Statistical computing & data science

- What's the difference between data science and statistics?

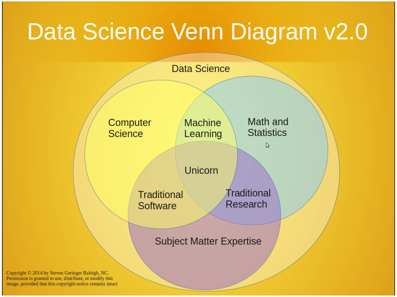
  "A data scientist is just a sexier word for statistician."
  — Nate Silver (outdated)

  "A data scientist is a better computer scientist than a statistician and is a better statistician than a computer scientist."
  — Unknown (still accurate)

- What does a data scientist do?
  - There is not one correct answer.
  - Transform data into valuable information!
  - A data scientist spends a significant portion of time processing data and less time modeling data.

What is Data Science?

- **Wikipedia:** "the extraction of knowledge from data"
- Precise definition a bit unclear and controversed. . .
- Practitioners "agree" on the components of data science:
  - database management
  - gathering and cleaning
  - exploratory analysis
  - predictive modeling
  - data summary and visualization



Applications

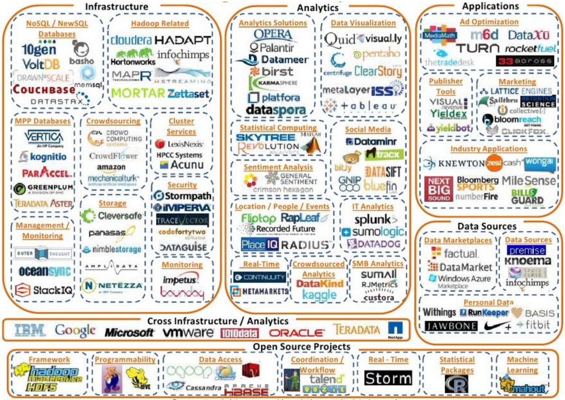
yelp
FLATIRON
foursquare
Mashable
Genentech
The New York Times
Capital One
Pfizer
1010DATA
Microsoft
ebay
JPMorganChase

Some of the hiring partners of *The Data Incubator*

- E-marketing
- Recommender systems
- Sport analytics
- Biotechnology
- Image or speech recognition
- Fraud and risk detection
- Social media

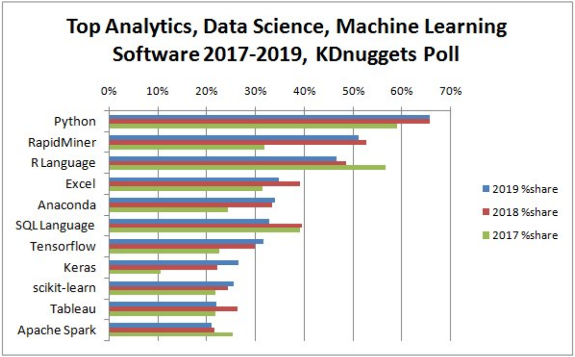
- Credit scoring
- E-commerce
- Government analysis
- Gaming
- Price comparisons
- Airline routes planing
- Delivery logistics

Technology ecosystem



source: rosebt.com

Most popular?



source: kdnuggets.com

Outline

1 Introduction

2 Organization

3 R

Course Description

Wednesday/Thursday 2:15pm-6:00pm

- **Registration:**
  - https://forms.gle/1oiSL9b2KQB5aAZG8 to register.
- **Zoom link:**
  - https://columbiauniversity.zoom.us/j/91365325993 for the lectures/exercise sessions.
- Instructor: Thibault Vatter
  - Email: thibault.vatter@unil.ch.
  - Office hours: by appointment.
- Teaching assistants
  - Aleksandr Shemendyuk
  - Maximilian Aigner
  - Hasini Gunawardena
  - Ilia Azizi
  - Office hours: by appointment.

Course website and Ed

- Two platforms:
  - Course website:
    - https://tvatter.github.io/dsfba_2020/
    - Syllabus/Schedule/PDFs.
  - Ed:
    - https://us.edstem.org/courses/2452
    - HWs1-2-3/Forum

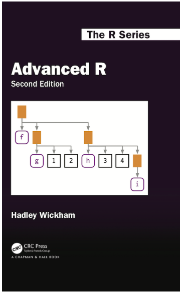
Grading

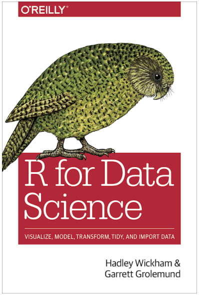
- 4 assignments (40%) and one project (60%)
  - First three assignments in Ed
  - Detailed reports for the last assignment and final project
  - Presentation during last lecture for the project
- Final grade
  - According to

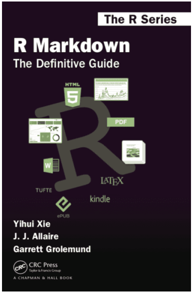
$$GRADE = \frac{\sum_{i=1}^{4} \frac{HW_i}{4} \cdot 40 + PR \cdot 60}{100}$$

  - HW_i for $i = \{1, 2, 3, 3, 4\}$ and PR are from 0 to 100
  - $GRADE$ will then be adjusted from 1 to 6
- For the project:
  - Groups of 1 or 2 members
  - More on that later
- Grades based on academic performance only!

Additional resources







- Books:
  - Advanced R
  - R for data science
  - R Markdown: The Definitive Guide
  - Most of the material in the slides taken from the first two.
- Additionally:
  - Rstudio cheat sheets
  - The CRAN website

Tentative outline

Date	Topic	Reading
09/16	Introduction	
09/23	Data Structures and Subsetting	ADVR 3+4
09/30	Data Structures and Subsetting	ADVR 3+4
10/07	Control Flows and Functions	ADVR 5+6
10/14	Control Flows and Functions	ADVR 5+6
10/21	Data Wrangling	R4DS 5, 9-16, 18
10/28	Data Wrangling	R4DS 5, 9-16, 18
11/04	Data Wrangling	R4DS 5, 9-16, 18
11/11	Visualization and Communication	R4DS 3+28, RMD 2
11/18	Visualization and Communication	R4DS 3+28, RMD 2
11/25	Presentations/Dashboards/Interactivity	RMD 4+5, htmlwidgets
12/02	Guest Lectures	
12/09	Projects Coaching	
12/16	Projects Presentations	

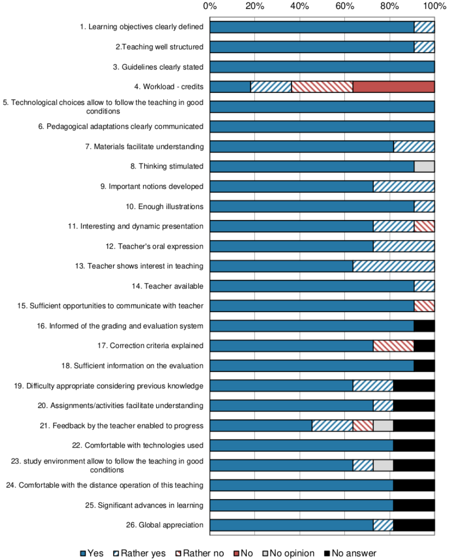
(numbers in the third column are book chapters)

Milestones

Date	Assignment
10/06	HW1
10/20	HW2
11/03	Project Proposal
11/10	HW3
12/01	HW4 (Project Update)
12/15	Project Report

Notice of caution!!!

- Course evaluations last semester:



Outline

1 Introduction

2 Organization

3 R

S and R

- S
  - A statistical programming language
  - First appeared in 1976
  - Developed by John Chambers and (in earlier versions) Rick Becker and Allan Wilks of Bell Labs
  - John Chambers, *[the aim is] to turn ideas into software, quickly and faithfully*
- R
  - Modern implementation of S
  - First appeared in 1993
  - Created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand
  - Currently developed by the *R Development Core Team*

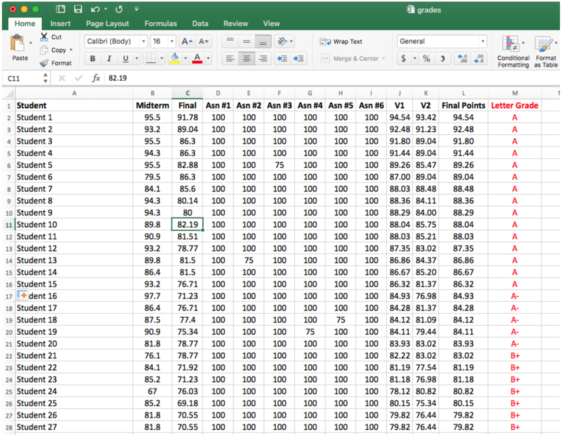
Some “technical” details about R

- **Part of the GNU free software project**
- Source code written primarily in C, Fortran, and R
- **Available for Windows, macOS, and Linux**
- Multi-paradigm: object-oriented, functional, procedural
- Dynamically typed
- Scripting language (interpreted)
- **Wide variety of statistical and graphical techniques**
- **Easily extensible through functions and packages**
- **Read/write from/to various data sources**

What about Excel?

source: fantasyfootballanalytics.net

Excel is great for certain things. . .



C11 f_x 82.19

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Student**	**Midterm**	**Final**	**Asn #1**	**Asn #2**	**Asn #3**	**Asn #4**	**Asn #5**	**Asn #6**	**V1**	**V2**	**Final Points**	**Letter Grade**
2	Student 1	95.5	91.78	100	100	100	100	100	100	94.54	93.42	94.54	A
3	Student 2	93.2	89.04	100	100	100	100	100	100	92.48	91.23	92.48	A
4	Student 3	95.5	86.3	100	100	100	100	100	100	91.80	89.04	91.80	A
5	Student 4	94.3	86.3	100	100	100	100	100	100	91.44	89.04	91.44	A
6	Student 5	95.5	82.88	100	100	75	100	100	100	89.26	85.47	89.26	A
7	Student 6	79.5	86.3	100	100	100	100	100	100	87.00	89.04	89.04	A
8	Student 7	84.1	85.6	100	100	100	100	100	100	88.03	88.48	88.48	A
9	Student 8	94.3	80.14	100	100	100	100	100	100	88.36	84.11	88.36	A
10	Student 9	94.3	80	100	100	100	100	100	100	88.29	84.00	88.29	A
11	Student 10	89.8	82.19	100	100	100	100	100	100	88.04	85.75	88.04	A
12	Student 11	90.9	81.51	100	100	100	100	100	100	88.03	85.21	88.03	A
13	Student 12	93.2	78.77	100	100	100	100	100	100	87.35	83.02	87.35	A
14	Student 13	89.8	81.5	100	75	100	100	100	100	86.86	84.37	86.86	A
15	Student 14	86.4	81.5	100	100	100	100	100	100	86.67	85.20	86.67	A
16	Student 15	93.2	76.71	100	100	100	100	100	100	86.32	81.37	86.32	A
17	dent 16	97.7	71.23	100	100	100	100	100	100	84.93	76.98	84.93	A-
18	Student 17	86.4	76.71	100	100	100	100	100	100	84.28	81.37	84.28	A-
19	Student 18	87.5	77.4	100	100	100	100	75	100	84.12	81.09	84.12	A-
20	Student 19	90.9	75.34	100	100	100	75	100	100	84.11	79.44	84.11	A-
21	Student 20	81.8	78.77	100	100	100	100	100	100	83.93	83.02	83.93	A-
22	Student 21	76.1	78.77	100	100	100	100	100	100	82.22	83.02	83.02	B+
23	Student 22	84.1	71.92	100	100	100	100	100	100	81.19	77.54	81.19	B+
24	Student 23	85.2	71.23	100	100	100	100	100	100	81.18	76.98	81.18	B+
25	Student 24	67	76.03	100	100	100	100	100	100	78.12	80.82	80.82	B+
26	Student 25	85.2	69.18	100	100	100	100	100	100	80.15	75.34	80.15	B+
27	Student 26	81.8	70.55	100	100	100	100	100	100	79.82	76.44	79.82	B+
28	Student 27	81.8	70.55	100	100	100	100	100	100	79.82	76.44	79.82	B+

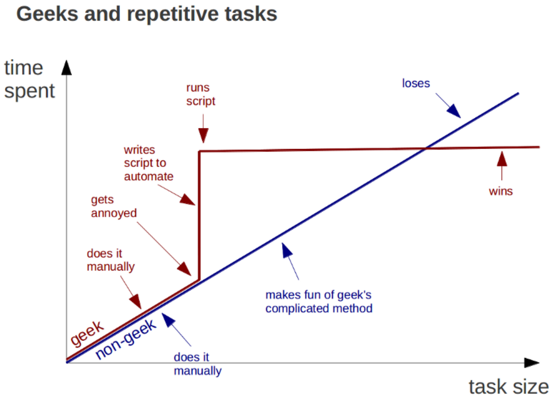
source: github.com/jdwilson4

. . . but not everything

R's advantages:

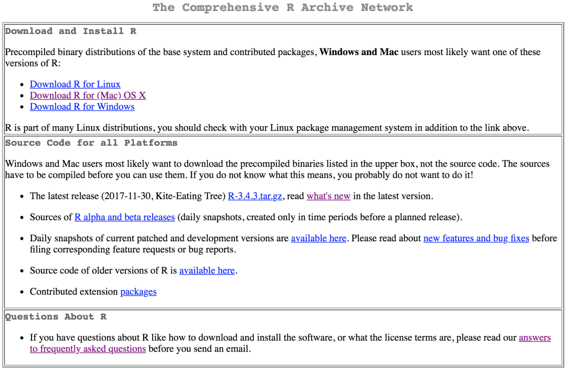
- **Easier automation**
- **Better reproducibility**
- Faster computation
- Supports larger data sets
- Reads any type of data
- More powerful data manipulation capabilities
- Easier project organization
- Easier to find and fix errors
- Free & open source
- Advanced statistics capabilities
- State-of-the-art graphics
- Runs on many platforms
- Anyone can contribute packages to improve its functionality

Automation and reproducibility



source: trendct.org

CRAN



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 (2017-11-30, Kite-Eating Tree) R-3.4.3.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

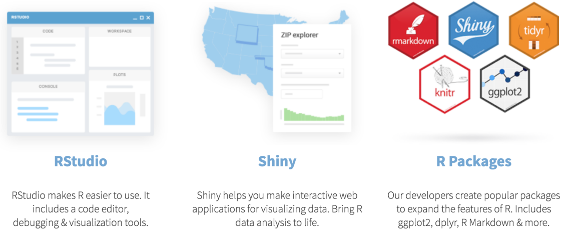
Questions About R

- 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.

source: cran.r-project.org

RStudio

- An open-source integrated development environment (IDE)
- RStudio Desktop available for Windows, macOS, and Linux



RStudio

RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.

Shiny

Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life.

R Packages

Our developers create popular packages to expand the features of R. Includes ggplot2, dplyr, R Markdown & more.

source: rstudio.com

Base R

- What is `Base R`?

  "The package named base is in a way the core of R and contains the basic functions of the language, particularly, for reading and manipulating data."

  — R for Beginners, Emmanuel Paradis

- `Base R` includes all default code for performing common data manipulation and statistical tasks.
- You might recognize some `Base R` functions:
  - `mean()`, `median()`, `lm()`, `summary()`, `sort()`
  - `data.frame()`, `read.csv()`, `cbind()`, `grep()`, `regexpr()`
  - Many many more...
- If you don't recognize any `Base R` functions, don't worry!

The tidyverse

- Common criticisms of **Base R**:
  - Function names/arguments are often inconsistent/confusing.
  - Functions often non type-stable objects.
  - Sometimes slow.
  - Other complaints exist. . .
- So what is the **tidyverse**? A collection of R packages
  - designed for data science,
  - sharing an underlying design philosophy, grammar, and data structures.
- Similar to **Base R**, but:
  - More descriptive function names and consistent inputs.
  - Type-stable.
  - Often faster than common **Base R** functions.

Core `tidyverse` packages

- `ggplot2`: declarative graphics, based on The Grammar of Graphics.
- `dplyr`: grammar of data manipulation.
- `tidyr`: functions that help you get to tidy data.
- `readdr`: reading in rectangular data.
- `purrr`: enhancing R's functional programming (FP).
- `tibble`: a `tibble`, or `tbl_df`, is a modern rethinking of the `data.frame`.
- `stringr`: functions designed to make working with strings as easy as possible.
- `forcats`: useful tools that solve common problems with factors.

More on the tidyverse website!