COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Lecture I: Course Overview, Intro to Data Science, and R

Data Science for Business Analytics

Thibault Vatter <thibault.vatter@unil.ch>

Department of Statistics, Columbia University and HEC Lausanne, UNIL

26.02.2018

Outline

1 Course overview

2 Intro to data science

3 R

A little about me

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

- 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:
  - Post-doctoral fellow at Columbia University
  - Live in New York city
- Hobbies:
  - Flying planes
  - Watching bay area teams (go 49ers and Warriors!)
  - Beers (formerly at Satellite, now in Brooklyn micro-breweries)

The basics

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

- Every two weeks: 02.26/03.12/03.26/04.09/04.30/05.14/05.28
- Lectures:
  - focus on introducing the concepts
  - 8:15-9:00am/9:15-10:00am + 1:15-3:00pm/3:15-4:00pm
  - classroom 237, Internef building
- Exercise sessions:
  - focus on the assignments and project
  - 10:15-11:00am/11:15-12:00pm + 3:15-4:00pm/4:15-5:00pm
  - lab room 143, Internef building
- TA: Natasha Tagasovska, `natasha.tagasovska@unil.ch`

Grading

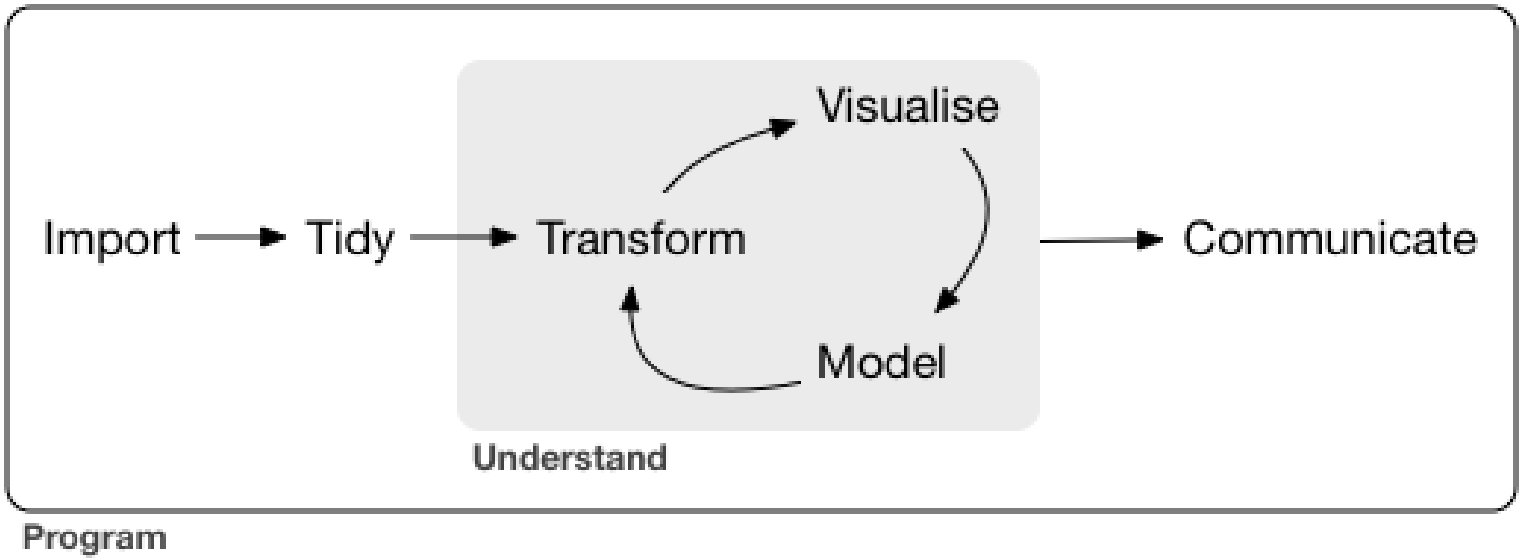
- 4 assignments (50%) and one project (50%)
  - Detailed reports for each assignment and final project
  - Presentation during last lecture for the project
- Final grade
  - According to

$$GRADE = \frac{\sum_{i=1}^{4} HW_i \cdot 12.5 + PR \cdot 50}{100}$$

  - HW_i for $i = \{1, 2, 3, 4\}$ and PR are from 0 to 100
  - $GRADE$ will then be adjusted from 1 to 6
- Groups of 1 or 2 members
  - Email to Natasha with the group members
  - One email per group is enough
  - Deadline for group registration is **March 19**
- Grades based on academic performance only!

Learning outcomes

- Manage and analyze data
- Develop data products
- Use data science in a business context



source: r4ds.had.co.nz

Lectures

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Date	Topic
02.26 (am)	Intro
02.26 (pm)	R workflow and RMarkdown
03.12 (am)	Wrangling (I)
03.12 (pm)	Visualization (I)
03.26 (am)	Wrangling (II)
03.26 (pm)	Visualization (II)
04.09 (am)	Modeling (I)
04.09 (pm)	Modeling (II)
04.30 (am)	Shiny
04.30 (pm)	Guest presentation
05.14 (am)	Big data
05.14 (pm)	Hadoop and Spark
05.28 (am)	Projects presentations
05.28 (pm)	Projects presentations

Lab sessions

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Date	Topic	HW#
02.26 (am)	R Refresher	/
02.26 (pm)	Workflow, RMarkdown, data w. & v. (I)	HW1
03.12 (am)	Project	/
03.12 (pm)	Workflow, RMarkdown, data w. & v. (I)	HW1
03.26 (am)	Project	/
03.26 (pm)	Data w. & v. (II), modeling (I and II)	HW2
04.09 (am)	Project	/
04.09 (pm)	Data w. & v. (II), modeling (I and II)	HW2
04.30 (am)	Project	/
04.30 (pm)	Shiny app	HW3
05.14 (am)	Project	/
05.14 (pm)	Spark	HW4

Milestones

Date	Assignment
03.26	HW1
04.09	Project proposal
04.30	HW2
05.14	HW3
05.14	Project update
05.28	HW4
05.28	Project report

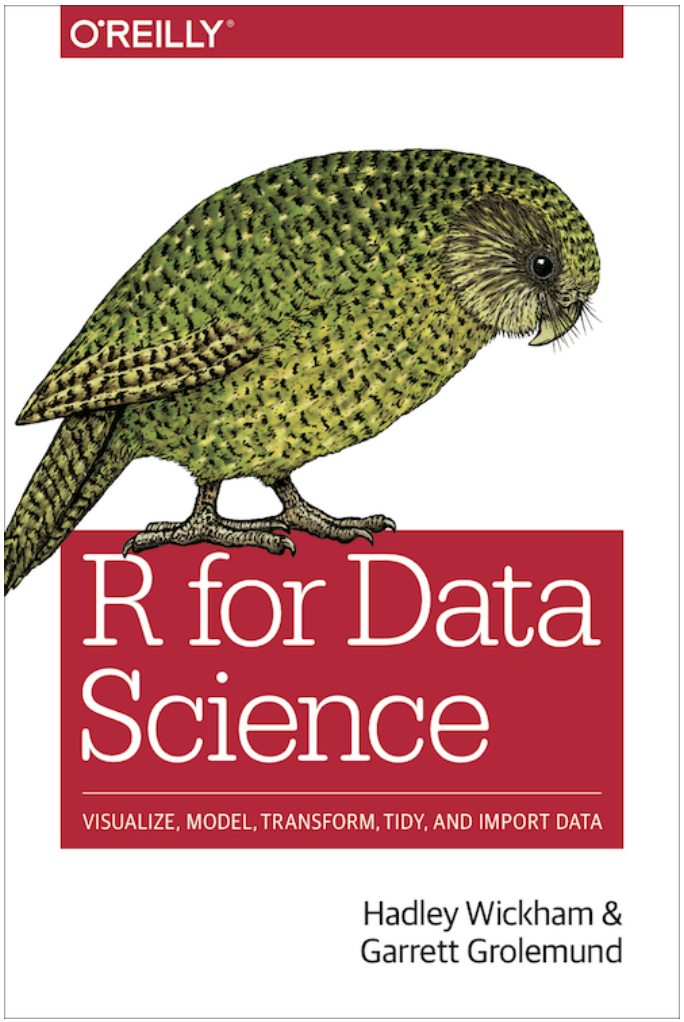
- To be submitted before midnight of the due date
- No late submission without medical certificate

Course website

All lecture notes, the syllabus, assignments, and additional resources are available at:

https://tvatter.github.io/dsfba_2018/

Resources



- R for data science
- The CRAN website
- Rstudio cheat sheets
- Much more in the resources section of the course website

Best place to look for answers?

Outline

1 Course overview

2 Intro to data science

3 R

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

A brief (and opinionated) history

- 1960, Peter Naur publishes *Datalogy: the science of data and its place in education*
- 1962, John Tuckey, *The Future of Data Analysis*:

  > *. . . as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. . . I have come to feel that my central interest is in data analysis. . .*

- 1974, Peter Naur, *Concise Survey of Computer Methods*:

  > *The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.*

- 1977, John Tuckey publishes *Exploratory Data Analysis*

A brief (and opinionated) history

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

- 1989, first Knowledge Discovery in Databases (KDD) workshop (later maturing into the ACM SIGKDD Conference)
- 1994, *Database marketing*:

  > *Companies are collecting mountains of information about you, crunching it to predict how likely you are to buy a product, and using that knowledge to craft a marketing message precisely calibrated to get you to do so.*

- 1996, International Federation of Classification Societies (IFCS) meet in Tokyo (*Data science, classification, and related methods*)
- 1997, C.F. Jeff Wu inaugural's lecture *Statistics = Data Science?* for appointment to the H. C. Carver Professorship at the University of Michigan.

A brief (and opinionated) history

- 1997, the journal *Data Mining and Knowledge Discovery* is launched
- 2001, William Cleveland (Bell Labs) published *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics.*

  > *[a plan] to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called "data science."*

- 2001, Leo Breiman, *Statistical Modeling: The Two Cultures*

  > *There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.*

A brief (and opinionated) history

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

- 2002, creation of *Data Science Journal* (CS focus)
- 2003, creation of *Journal of Data Science* (stats focus)
- 2006, Hadoop
- 2008, DJ Patil (LinkedIn) and Jeff Hammerbacher (Facebook) coined the term *data scientist* to define their jobs
- 2009, reintroduction of the term NoSQL
- 2009, Hal Varian (chief economist at Google)

  the sexy job in the next 10 years will be statisticians

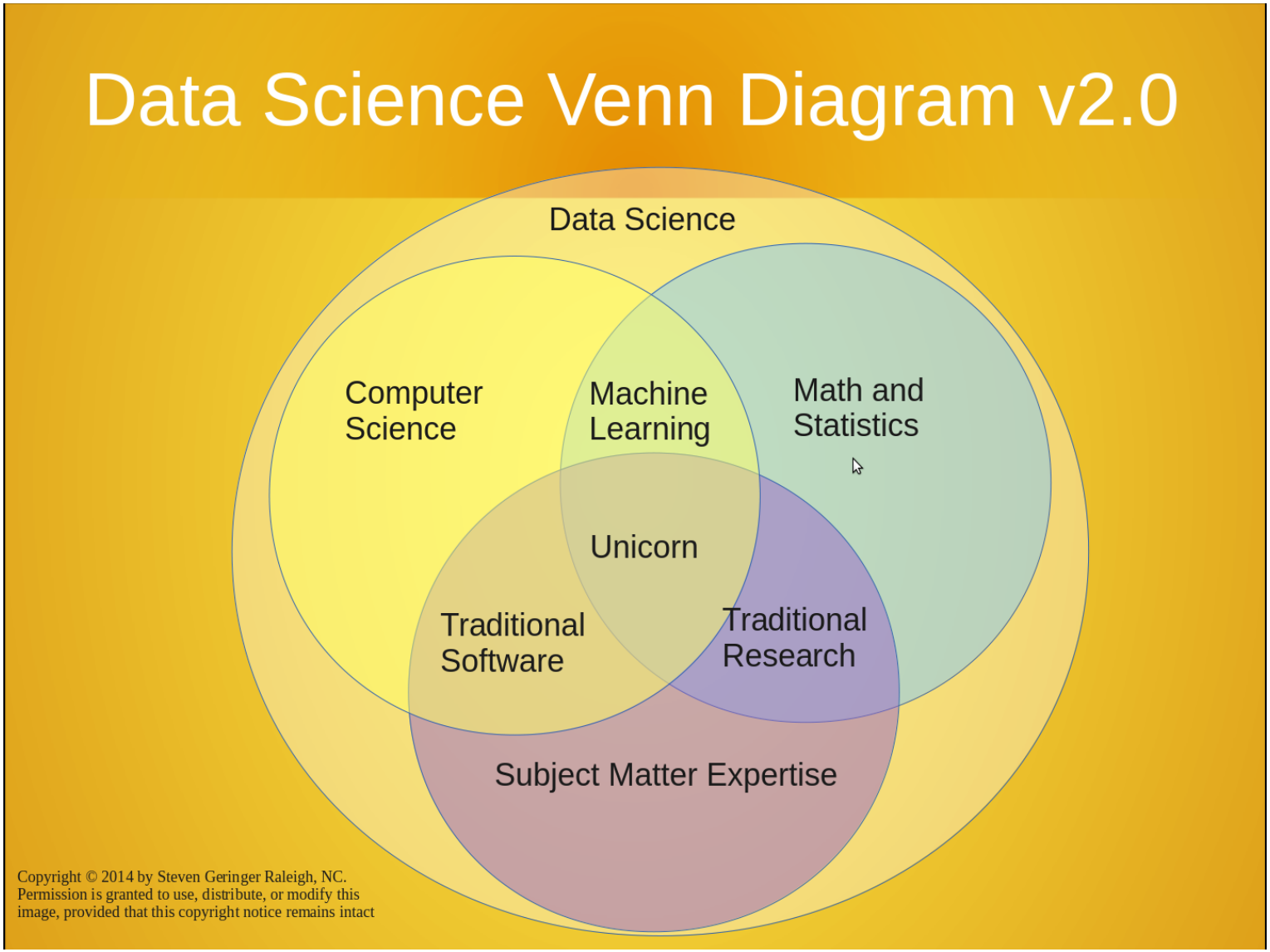
- 2012, Harvard Business Review publishes *Data Scientist: The Sexiest Job of the 21st Century*
- 2012, job listings for *Data Scientists* increased by 10,000%
- 2014, Bin Yu's *Let us own Data Science* speech
- 2015, DJ Patil as White House's first *Chief Data Scientist*

Applications

Some of the hiring partners of *The Data Incubator*

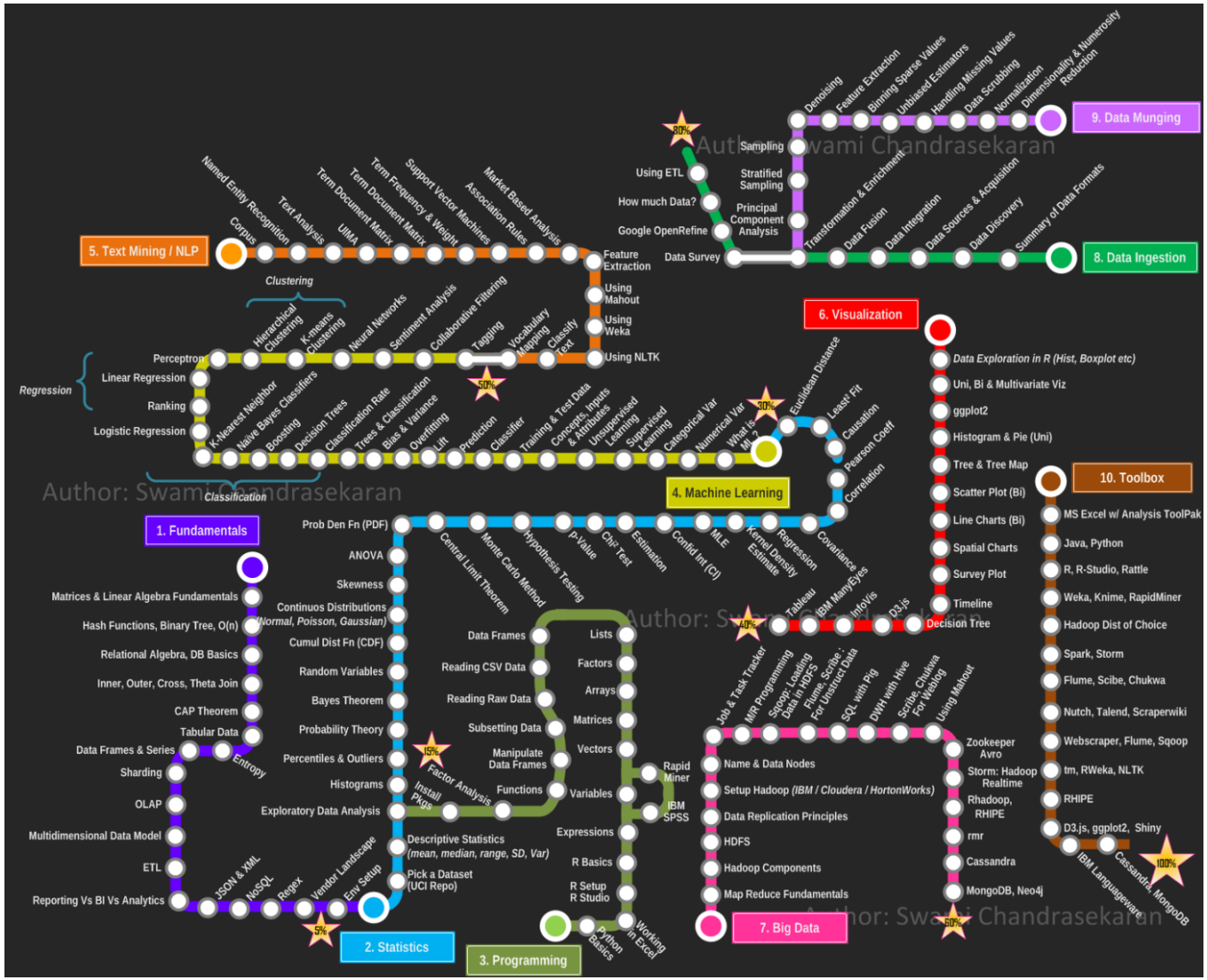
- E-marketing
- Recommender systems
- Sport analytics
- Biotechnology
- Image or speech recognition
- Fraud and risk detection
- Social media
- Credit scoring
- E-commerce
- Government analysis
- Gaming
- Price comparisons
- Airline routes planing
- Delivery logistics

Data scientists?



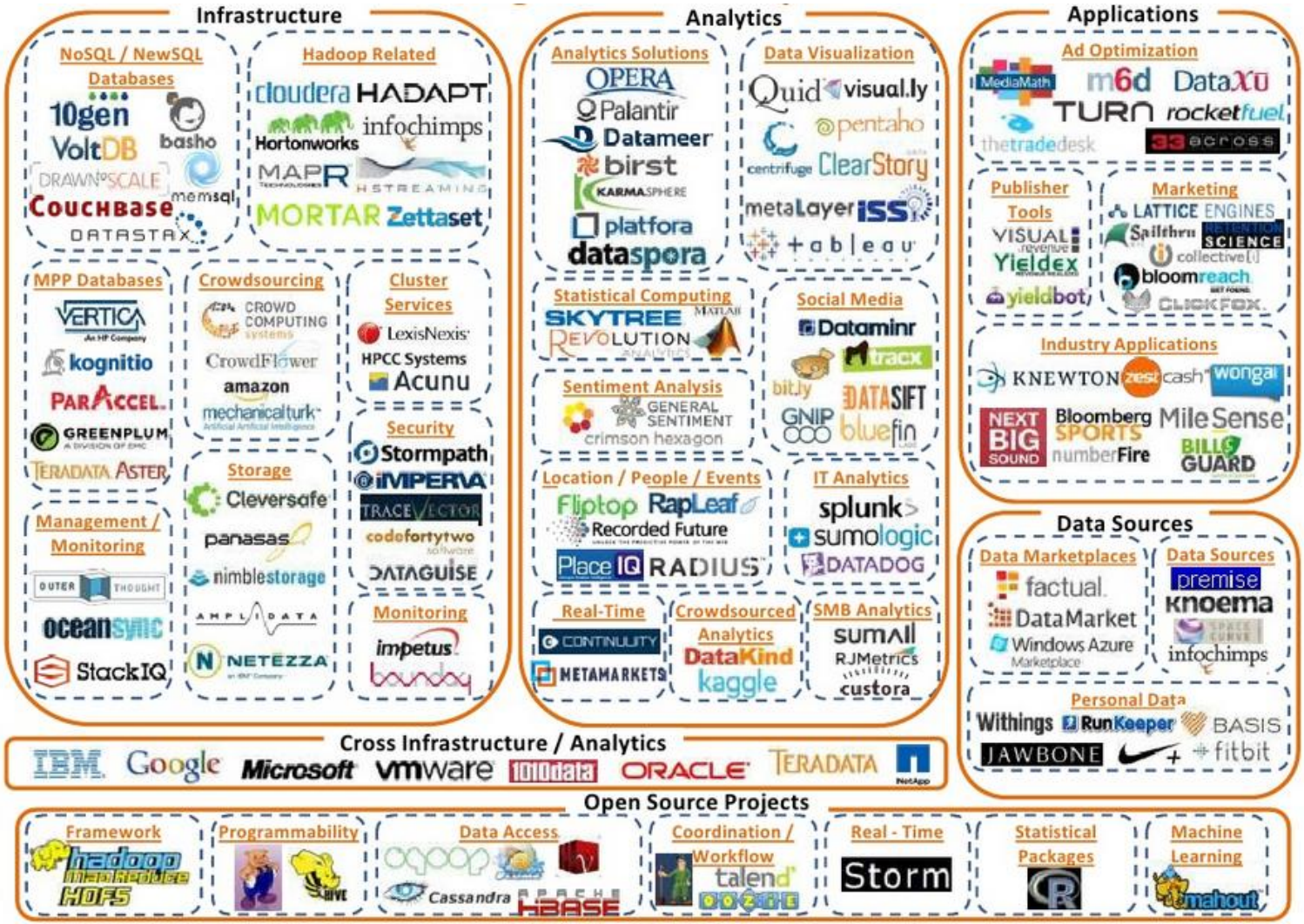
The data science toolbox

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK



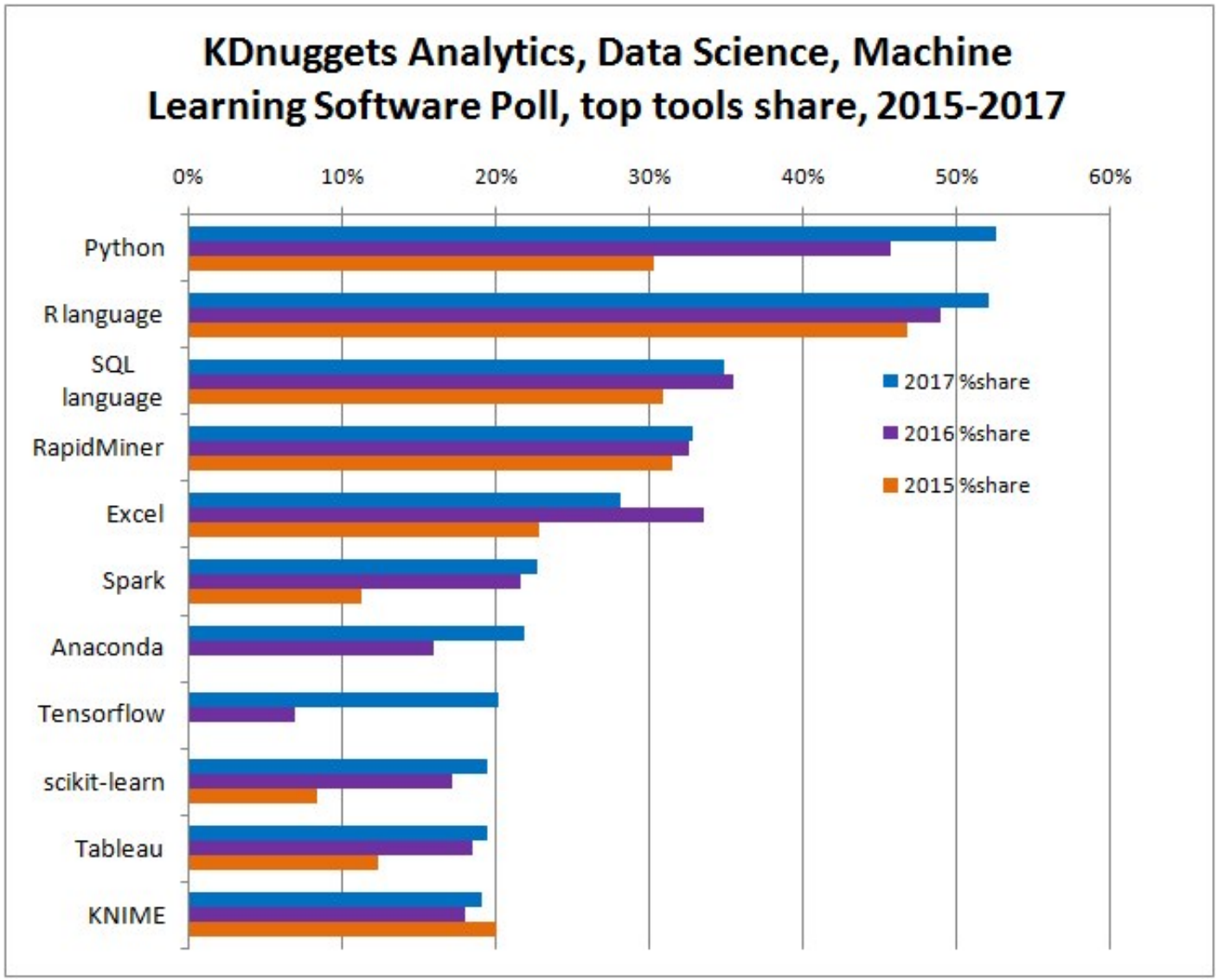
source: datasciencecentral.com

Technology ecosystem



source: rosebt.com

Most popular?



source: kdnuggets.com

Outline

1 Course overview

2 Intro to data science

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] o 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

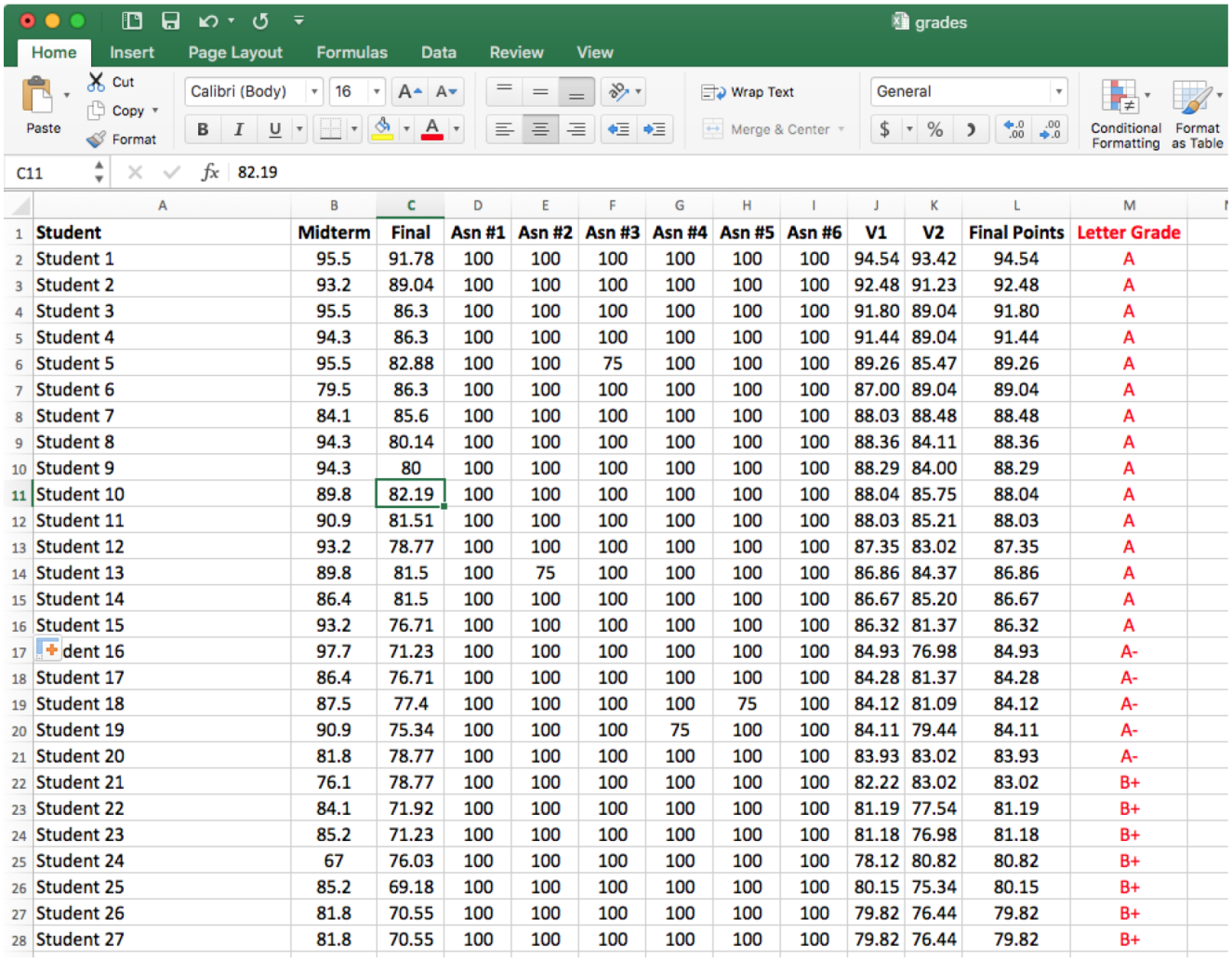
- **Part of the GNU free software project**
- Source code written primarily in C, Fortran, and R
- **Available for Windows, macOS, and Linux**
- Multi-paradigm: object-oriented, functional, procedural
- Dynamically typed
- Scripting language (interpreted)
- **Wide variety of statistical and graphical 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...

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK



	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

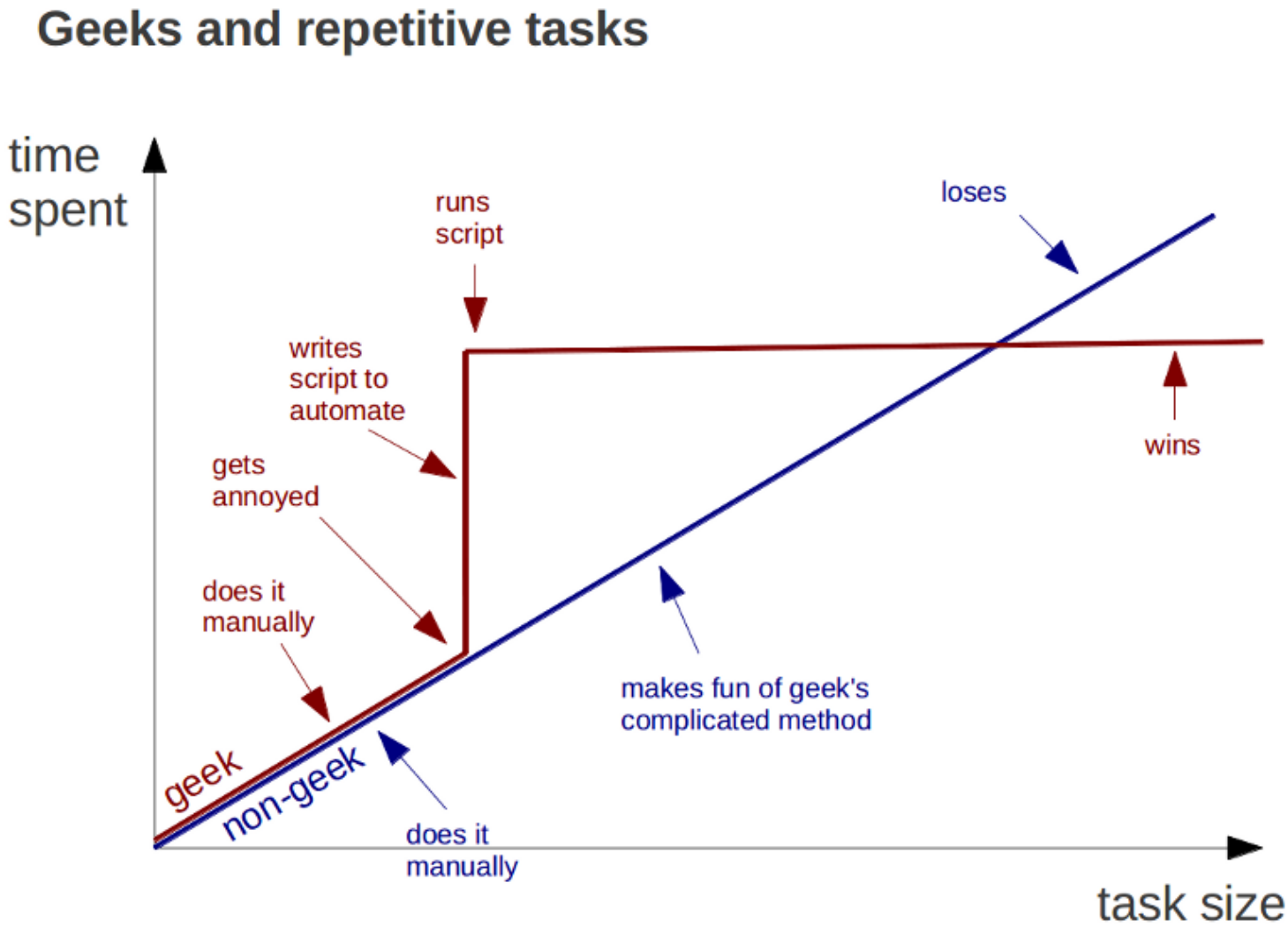
...but not everything

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

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK



source: trendct.org

How about Python?

source: python.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.

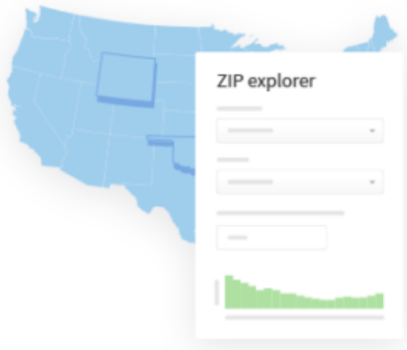
source: cran.r-project.org

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