Introduction to R and RStudio

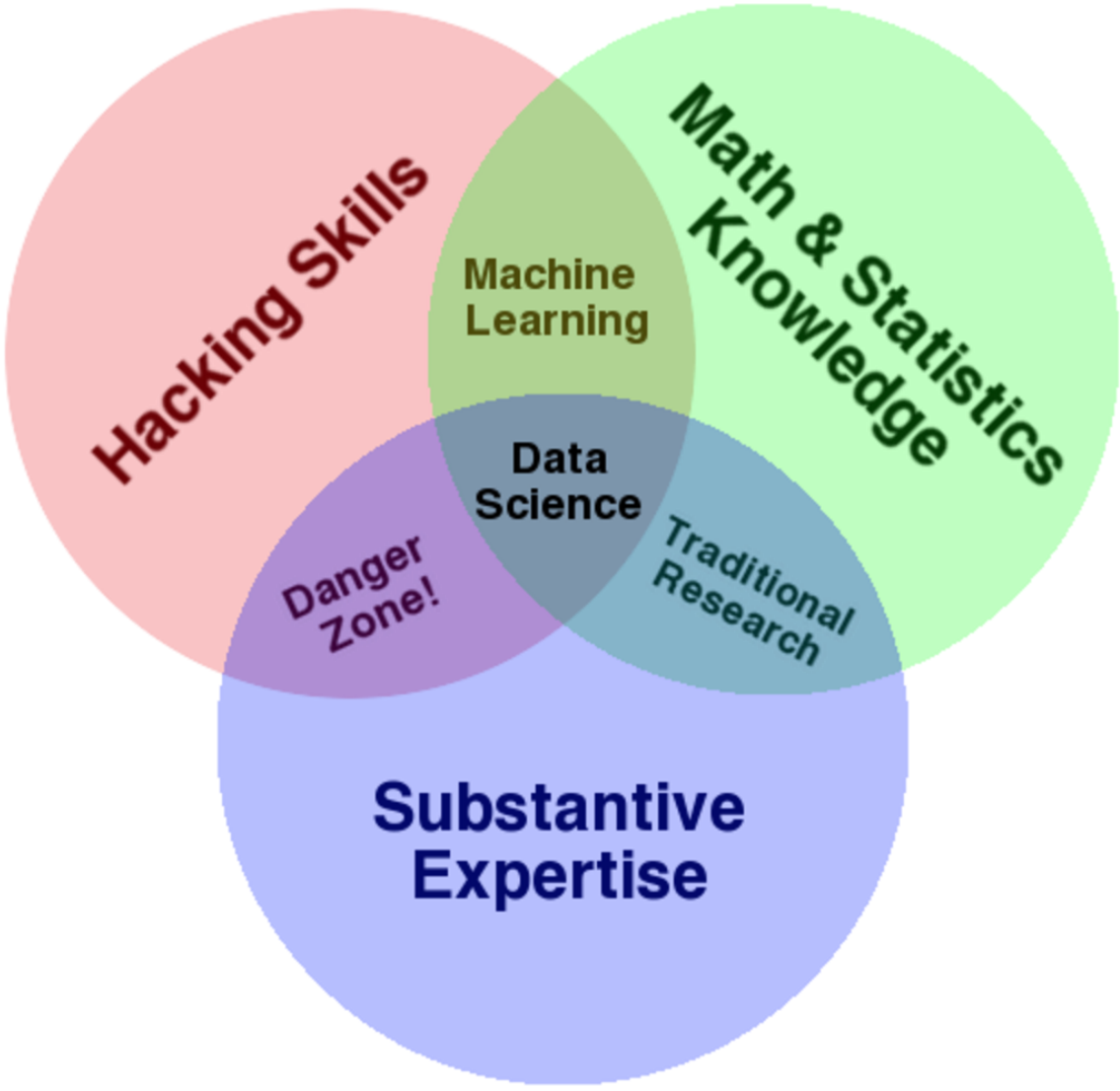
Chien-Liang Liu

August 22, 2016

Data Scientist: The Sexiest Job of the 21st Century

Source: `https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/`

Data Science Venn Diagram

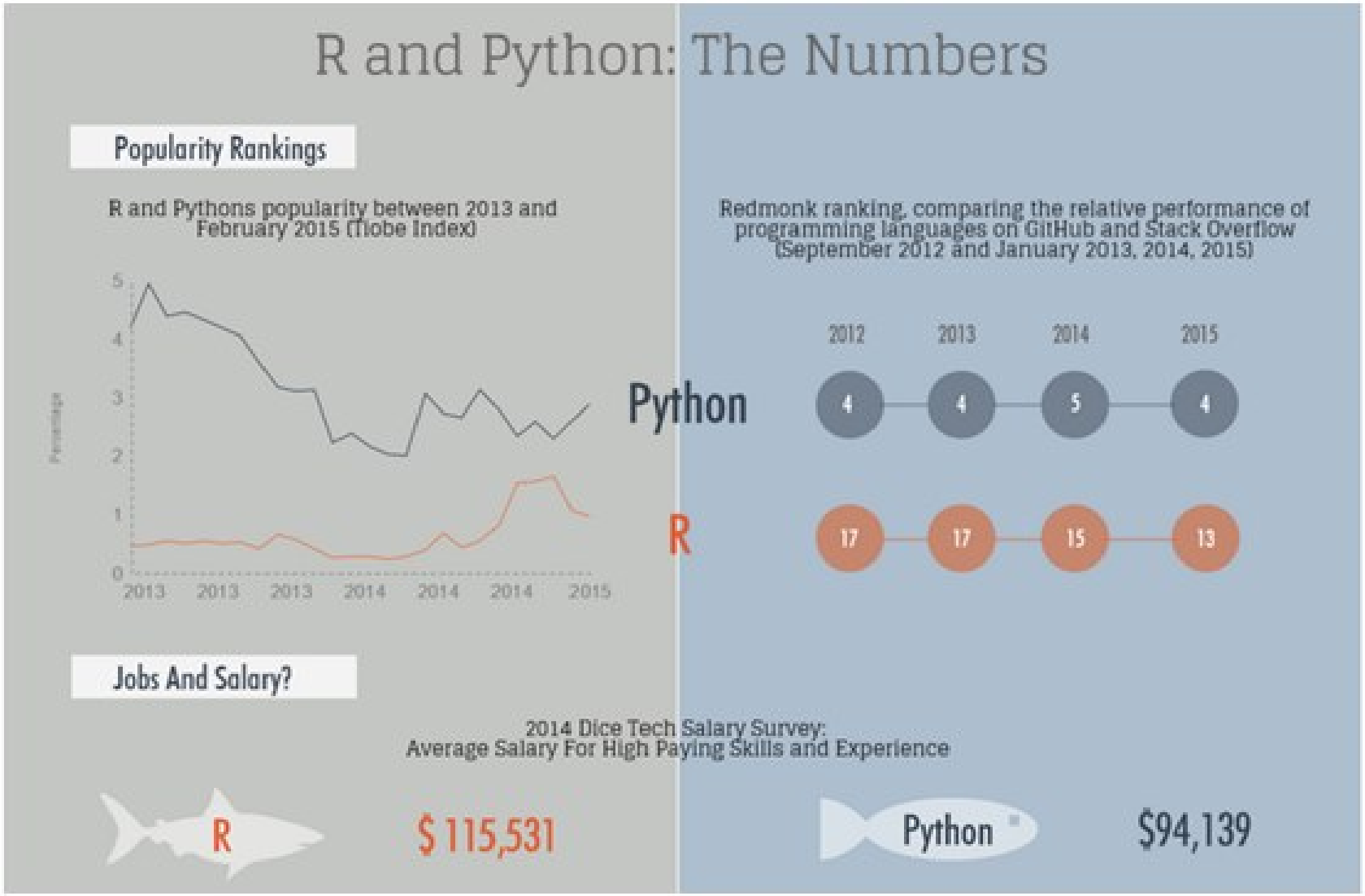


Source: `http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram`

Objective of this course

- This course focuses on hacking skill
- We use R as the programming language
- Learning by doing is the best approach
- R is simpler than C/C++, Java, and Python
- Besides R, we will also cover basic data analysis

R vs. Python



Source: `http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html`

When and how to use R?

- R is mainly used when the data analysis task requires standalone computing or analysis on individual servers
- It's great for exploratory work, and it's handy for almost any type of data analysis because of the huge number of packages and readily usable tests that often provide you with the necessary tools to get up and running quickly
- R can even be part of a big data solution.

IDE and packages for R

- When getting started with R, a good first step is to install the amazing RStudio IDE
- dplyr, plyr and data.table to easily manipulate packages
- stringr to manipulate strings
- zoo to work with regular and irregular time series,
- ggvis, lattice, and ggplot2 to visualize data, and
- caret for machine learning

When and how to use Python?

- You can use Python when your data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database
- Being a fully fledged programming language, it's a great tool to implement algorithms for production use

IDE and Packages for Python

- Python has no clear “winning” IDE. We recommend you to have a look at Spyder, IPython Notebook and Rodeo to see which one best fits your needs
- NumPy /SciPy (scientific computing)
- pandas (data manipulation) to make Python usable for data analysis
- matplotlib to make graphics
- scikit-learn for machine learning

Machine Learning

- Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence
- Machine learning explores the study and construction of algorithms that can learn from and make predictions on data
- Machine learning is closely related to and often overlaps with computational statistics; a discipline that also specializes in prediction-making. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field

Source: `https://en.wikipedia.org/wiki/Machine_learning`

Where is Data From?

Everywhere

Transaction Data

Government Data

- Traffic
- Real Estate
- ...

Scientific Data

- Genome
- Climate
- ...

UGC

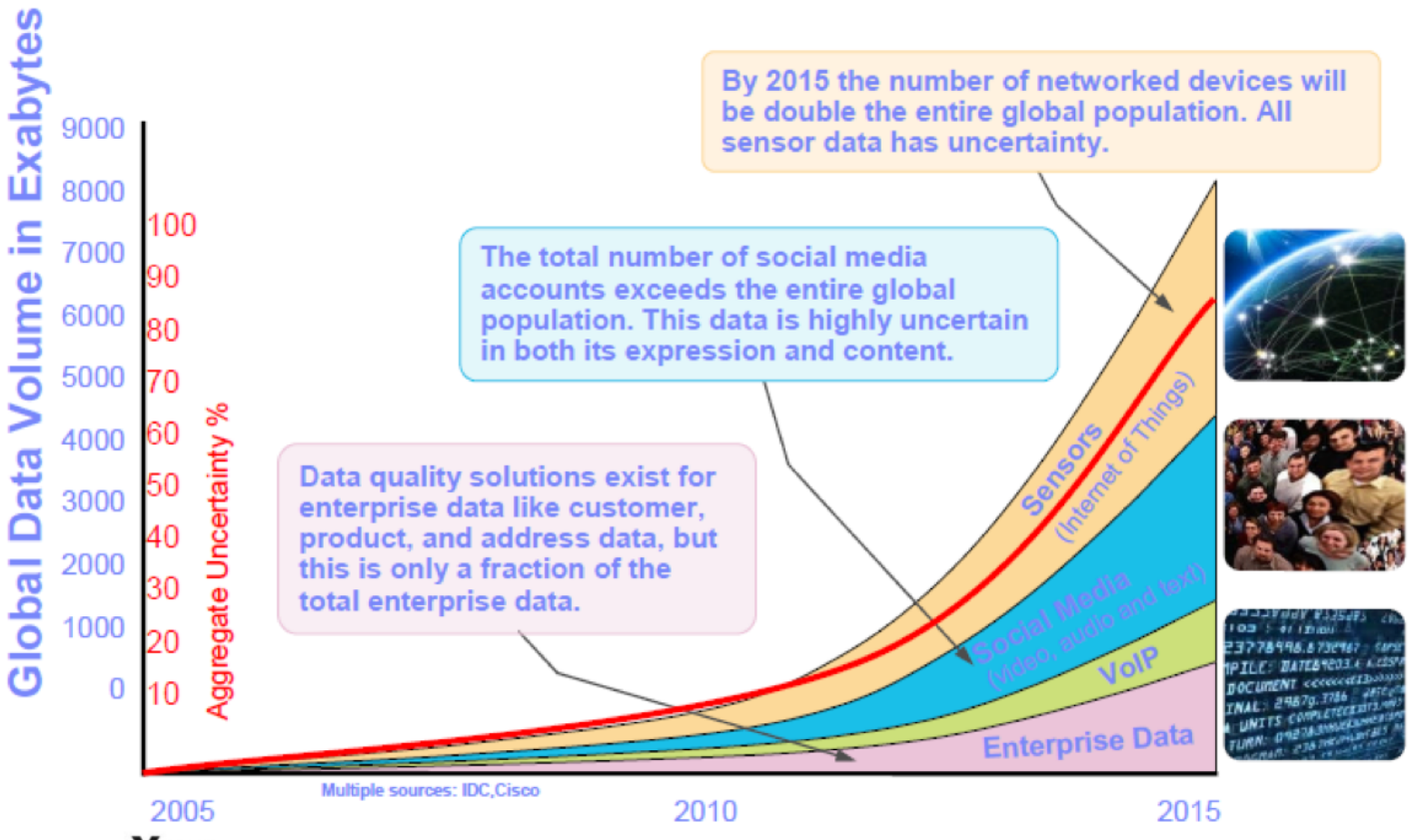
- Text
- Multimedia
  - Video
  - Image
  - Audio

Sensor Data

- GPS
- Gyroscope
- Accelerometer
- Thermometer
- ...

Log Data

Data is growing exponentially



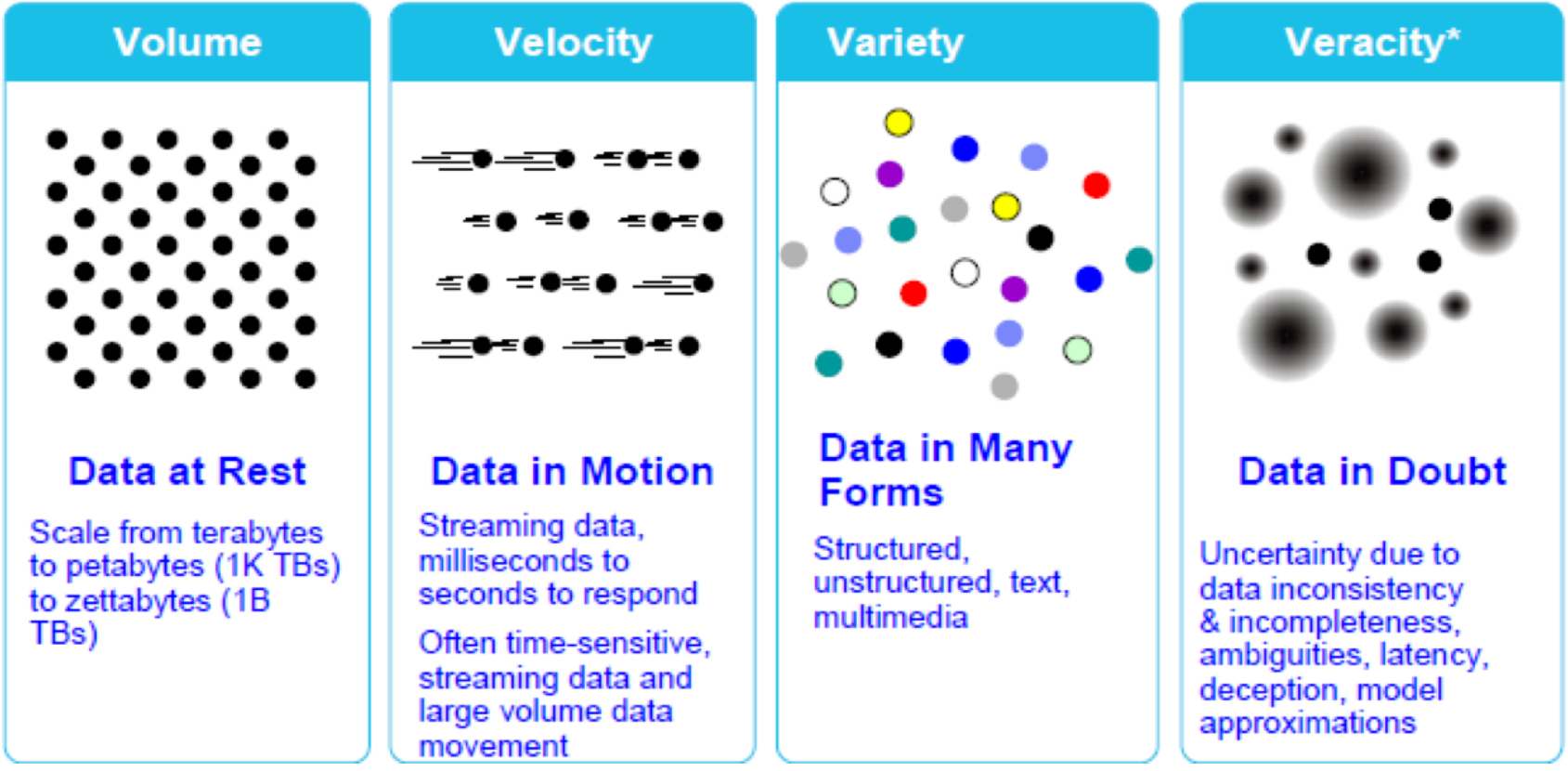
YouTube Video length 2006: <1hr/min
60hrs/min 2010
2005.02

tweets
2006.03 400k/Q 2007 | 100M/Q 2008 | 50M/Day 2010 | 140M/Day 2011

users
2004.02 100M 2008 | 300M 2009 | 500M 2010 | 800M 2011 | 1B 2012

What's Big Data

"Big Data" refers to data that grows so large that it is difficult to capture, store, manage, share, analyze and visualize with the typical hardware environments and database software tools



Source : Solutions Big Data IBM, 2012

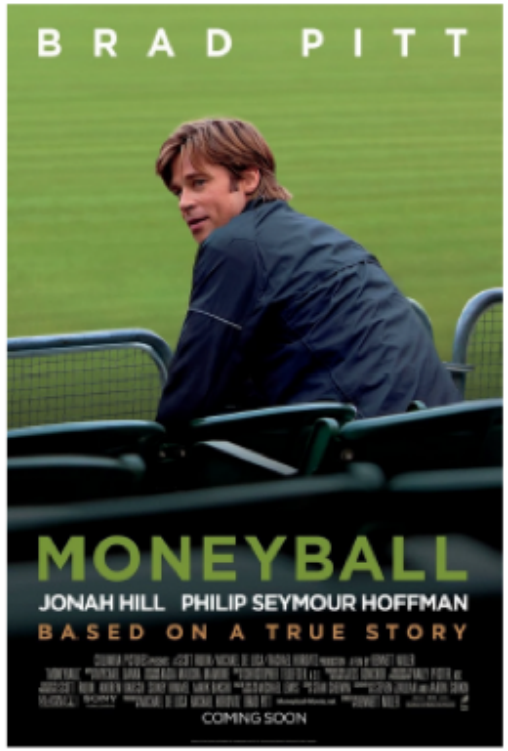
Data-Driven Decision Making

Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself `http://www.cra.org/ccc/docs/init/bigdatawhitepaper.pdf`

Corporations (baseball teams) need more data and advanced analytics to remain competitive.

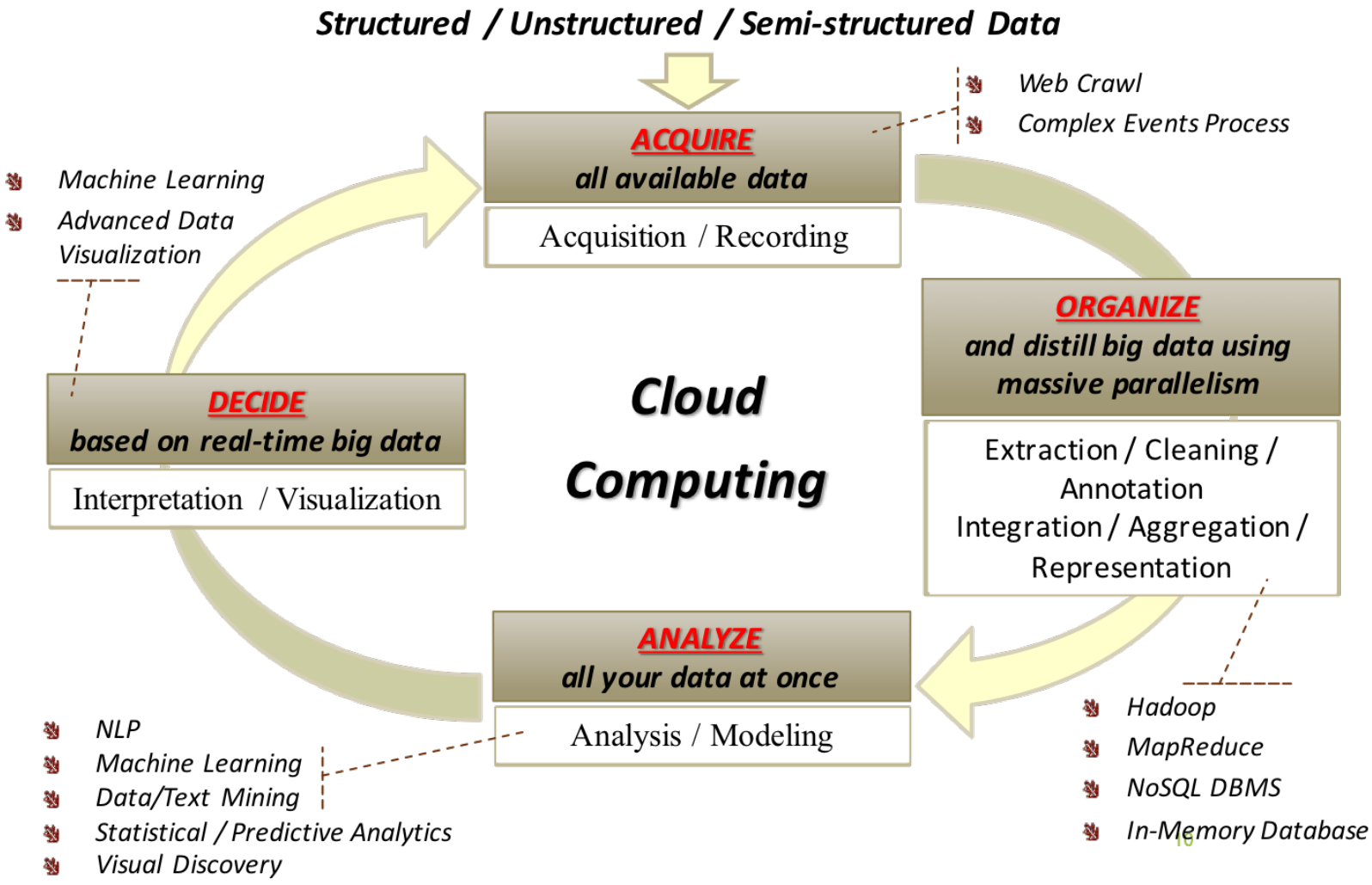
E.R.A evolved into ... Skill-Interactive Earned Run Average

$$SIERA = 6.145 - 16.986*(SO/PA) + 11.434*(BB/PA) - 1.858*((GB-FB-PU)/PA) + 7.653*((SO/PA)^2) +/- 6.664*(((GB-FB-PU)/PA)^2) + 10.130*(SO/PA)*((GB-FB-PU)/PA) - 5.195*(BB/PA)*((GB-FB-PU)/PA)$$



Source : New York Times

Data Processing Flow



R & RStudio

What's R

- R is a scripting language for statistical data manipulation and analysis
- It was inspired by, and is mostly compatible with, the statistical language S developed by AT&T
- R has become more popular than S/S-Plus, both because it's free and because more people are contributing to it

Why R?

- a public-domain implementation of the widely-regarded S statistical language; R/S is the de facto standard among professional statisticians
- comparable, and often superior, in power to commercial products in most senses
- available for Windows, Macs, Linux
- in addition to enabling statistical operations, it's a general programming language, so that you can automate your analyses and create new functions
- object-oriented and functional programming structure
- open-software nature means it's easy to get help from the user community, and lots of new functions get contributed by users, many of which are prominent statisticians

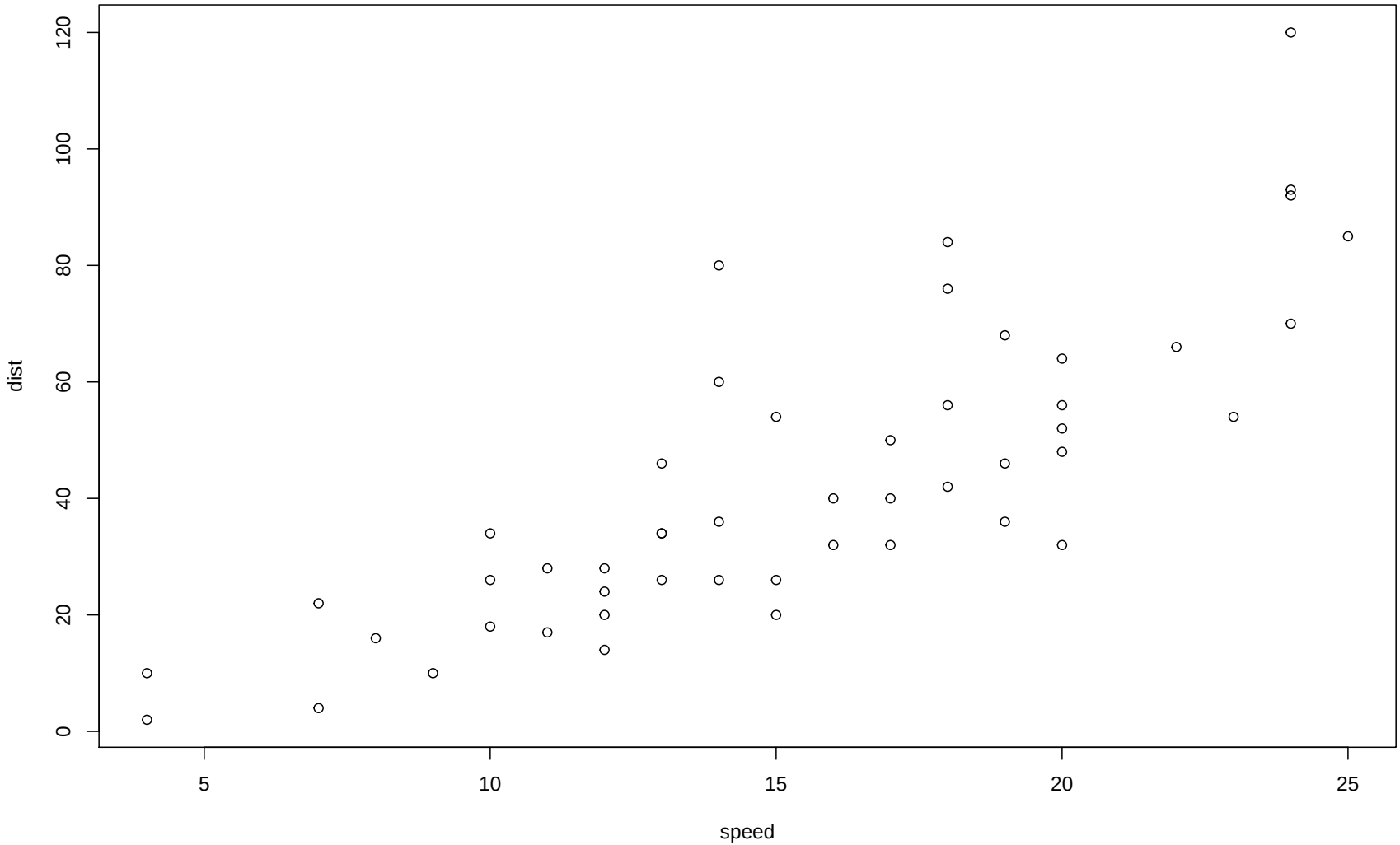
R and functional programming

R has many functional programming features. Roughly speaking, these allow one to apply the same function to all elements of a vector, or all rows or columns of a matrix or data frame, in a single operation. The advantages are important:

- Clearer, more compact code
- Potentially much faster execution speed
- Less debugging (since you write less code)
- Easier transition to parallel programming

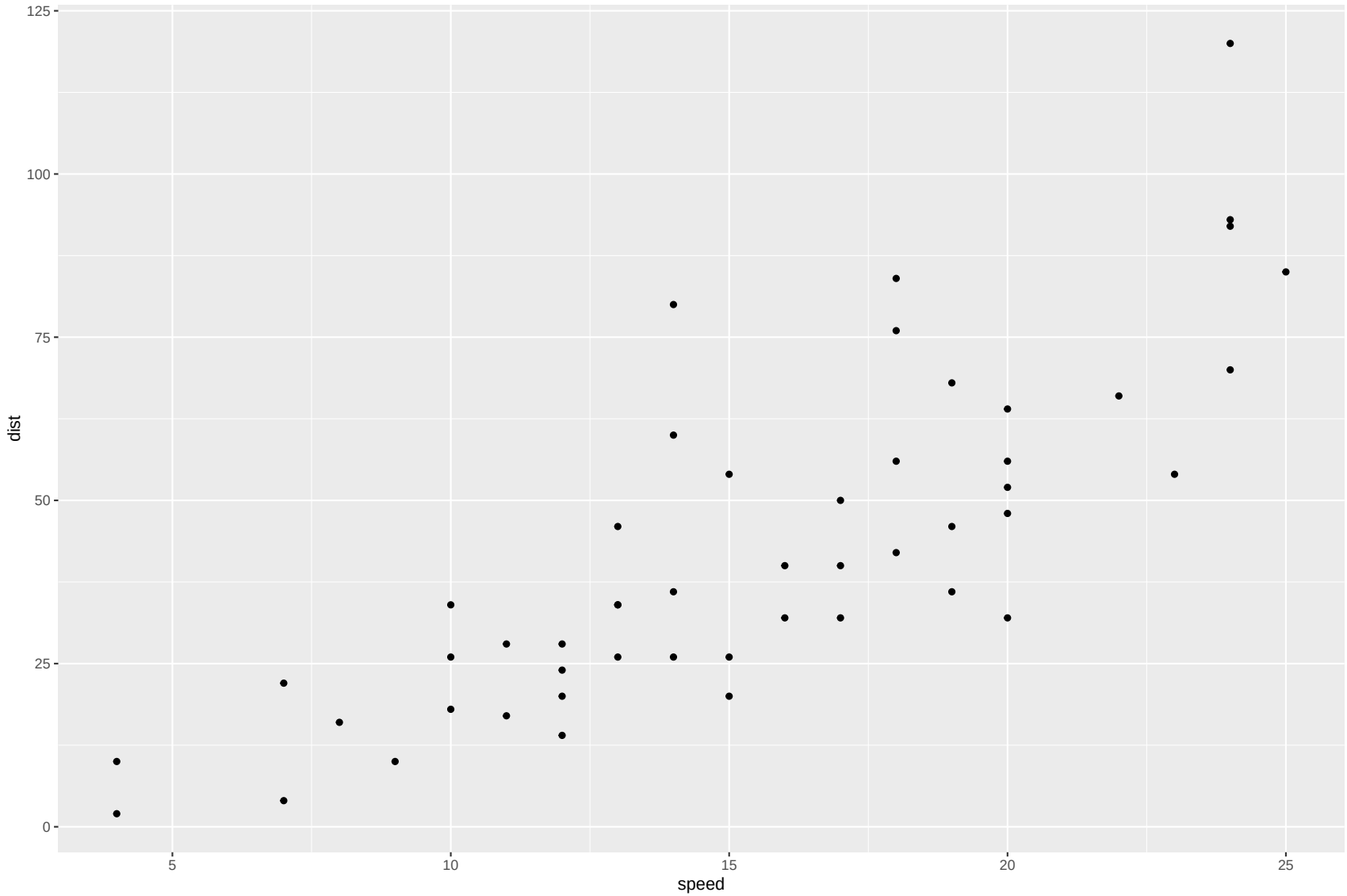
Visualization

R has provided a basic plotting envirorment for you.



ggplot2

But, ggplot2 is a more powerful visualization package in R.



R Installation

You can download R from `https://cran.r-project.org/` according to your OS.

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

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.

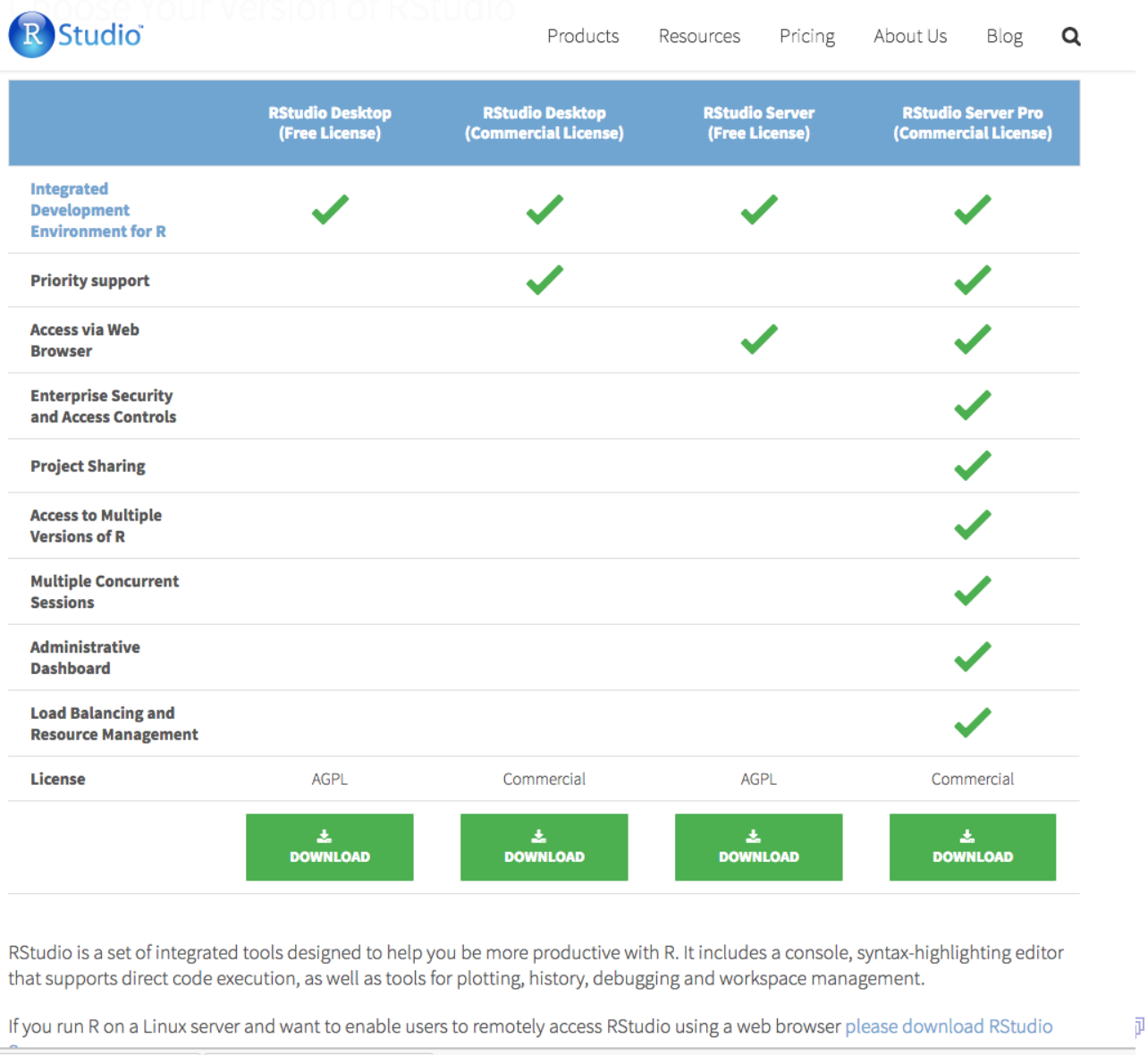
Source Code for all Platforms

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 (Tuesday 2016-06-21, Bug in Your Hair) R-3.3.1.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.

RStudio

You can download RStudio from `https://www.rstudio.com/products/rstudio/download2/`.



RStudio Products Resources Pricing About Us Blog

	RStudio Desktop (Free License)	RStudio Desktop (Commercial License)	RStudio Server (Free License)	RStudio Server Pro (Commercial License)
Integrated Development Environment for R	✓	✓	✓	✓
Priority support		✓		✓
Access via Web Browser			✓	✓
Enterprise Security and Access Controls				✓
Project Sharing				✓
Access to Multiple Versions of R				✓
Multiple Concurrent Sessions				✓
Administrative Dashboard				✓
Load Balancing and Resource Management				✓
License	AGPL	Commercial	AGPL	Commercial
	DOWNLOAD	DOWNLOAD	DOWNLOAD	DOWNLOAD

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

If you run R on a Linux server and want to enable users to remotely access RStudio using a web browser please download RStudio

Customize your RStudio

You can customize RStudio according to your own requirements. For example, you can change the code appearance, panel layout, and background color.

Howework 0

- Install R & RStudio
- Customize your own RStudio environment and try to be familiar with it
- Start coding