max planck institut
informatik

Data Mining & Matrices:

Introduction to R

Saskia Metzler
3 May 2017

Agenda

Part 1: Why R?

Part 2: Learn basic tasks in R

Why R?

Why to learn another language?

Why to learn R?

R is ...

... good for statistical programming
and data analysis tasks

... used in companies like Google,
Bank of America, Shell

... for free

... available for Linux, Windows, OSX

R is ...

"R is really important to the point that it's hard to overvalue it," said Daryl Pregibon, a research scientist at Google, which uses the software widely.
"It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems."

New York Times, 2009
http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?_r=0

R is ...

... easy with reading and writing of data

... equipped with built in tools for statistics and plotting

... vectorized thinking

R in Comparison

Excel

Take the right tool
for each task.

C(++)
Java

Matlab

Python

R in Comparison

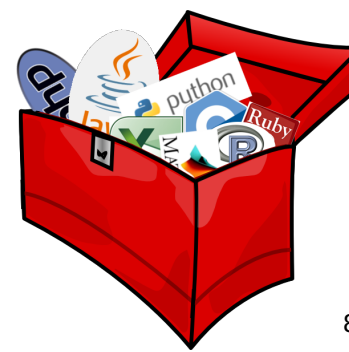
Excel

Take the right tool
for each task.

C(++)
Java

Matlab

Python



R vs. Excel

Importing data is easy, statistics and plotting are supported.

But did you ever try to do the same analysis on various datasets?

- on big data sets?
- with different sizes?
- on many of them?

Excel is designed for accounting.

R in Comparison

Excel

Take the right tool
for each task.

C(++)
Java

Matlab

Python

R vs. Java or C(++)

Java (or C, C++, ...) is capable but requires considerable programming overhead for

- reading data
- plotting data
- manipulating matrix/table data
- reformatting data to use different libraries

R in Comparison

Excel

Take the right tool
for each task.

C(++)
Java

Matlab

Python

R vs. Matlab

Conceptionally similar.

R is free, Matlab is not.
R is developed for statistics, Matlab for matrices.

R in Comparison

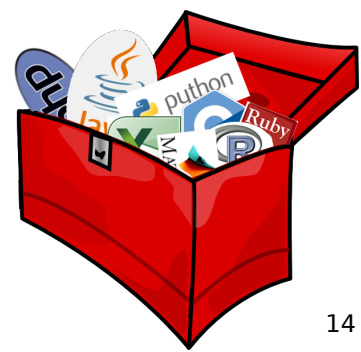
Excel

Take the right tool
for each task.

C(++)
Java

Matlab

Python



R vs. Python

Libraries make prototyping easy.
But these libraries are not inherent and might require differently formatted input.

Python doesn't come with the concept of vectors built in.

R in Comparison

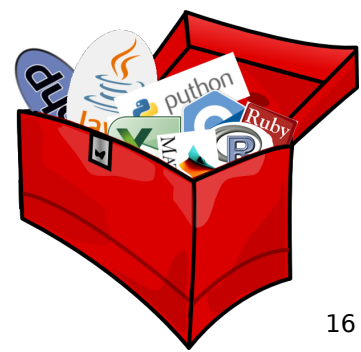
Take the right tool for each task.

Excel

C(++)
Java

Matlab

Python



Let's learn some R

Topics

1. Syntax basics & getting help
2. Vectors
3. Sequence generation
4. Matrices
5. Data frames
6. Reading & writing data
7. Plotting
8. Saving scripts
9. Function definitions

Topics

1. Syntax basics & getting help
2. Vectors
3. Sequence generation
4. Matrices
5. Data frames
6. Reading & writing data
7. Plotting
8. Saving scripts
9. Function definitions

Useful Links

What you get from help.start() too
https://stat.ethz.ch/R-manual/R-patched/doc/html/

"Quick R"
http://www.statmethods.net/

FAQ for very basic things
http://www.ats.ucla.edu/stat/r/faq/R_basics.htm

More tutorials
http://www.ats.ucla.edu/stat/r/