Data Scientist's Toolbox

Nicola Davide D'Avanzo

11/12/2014

This document is a generic overview about the data scientist's skills and attitudes.

What do data scientists do?

* Define the **questions of interest**
* Define the **ideal data set**
* Determine what **data you can access**
* **Obtain** the data
* **Clean** the data
* Exploratory **data analysis**
* **Statistical prediction/modeling**
* **Interpret** results
* **Challenge** results
* **Synthesize/write up** results
* Create **reproducible code**
* **Distribute** results to other people

Data Scientists tools:

* R (statistical programming environment)
* GitHub (Git repository web-based hosting service)
* Terminal Linux

R questions:

* What steps will reproduce the problem?
* What is the expected output?
* What do you see instead?
* What version of the product are you using?
* What operating system?

Data analysis questions:

* What is the question you are trying to answer?
* What steps/tools did you use to answer it?
* What did you expect to see?
* What do you see instead?
* What other solutions have you thought about?

Data analysis files:

* Data (Raw Data, Processed Data)
* Figures (Explorator figures, Final figures)
* R code (Raw scripts, Final scripts, R markdown files)
* Text (Readme files, Text of analysis)