Programming for Data Science

Data science using R language

Marco Beccuti

Università degli Studi di Torino

Dipartimento di Informatica

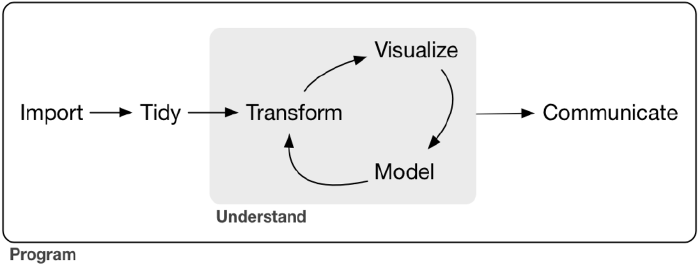
STUDII TAURINENSIS UNIVERSITAS 1404

Data science: an introduction

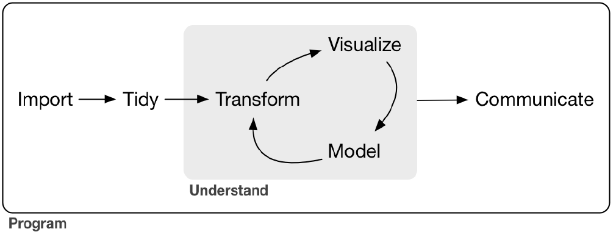
- Data science allows you to turn raw data into *understanding, insight, and knowledge*;

- In this course you will learn the most important tools in R to do data science.

- A typical data science project can be sketched as follows:

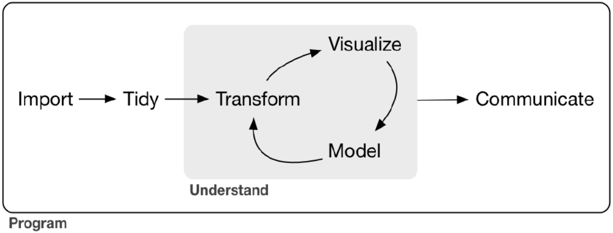


Data science: an introduction



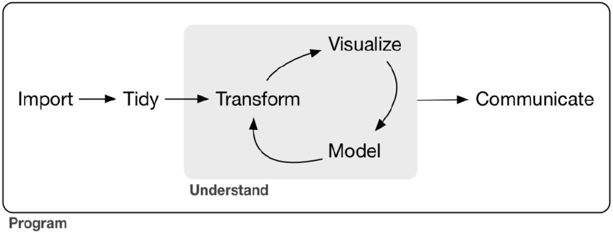
- *Import* your data stored in a file, database, web API into R;
- *Tidying* your data in a consistent dataset form:
  - each column is a variable;
  - each row is an observation.

Data science: an introduction



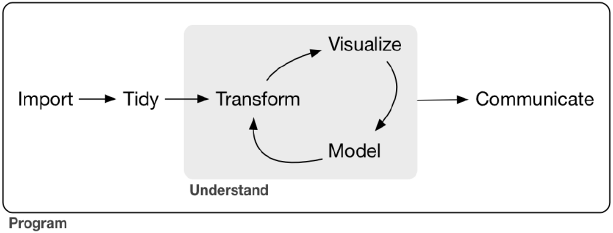
- *Transformation:*
  - to include the dataset narrowing in on observations of interest;
  - to create new variables that are functions of existing ones (e.g. speed from acceleration and distance);
  - to calculate a set of summary statistics (e.g. means,. . .)

Data science: an introduction



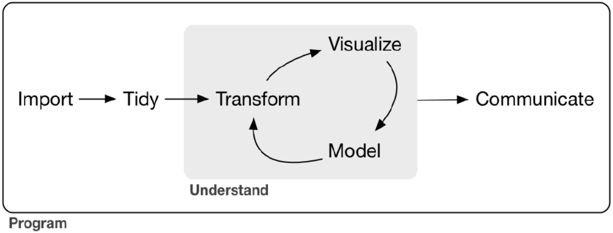
- Visualization and modeling are the two main tools for knowledge generation;
- They have complementary strengths and weaknesses;
- A real analysis will iterate between them many times.

Data science: an introduction



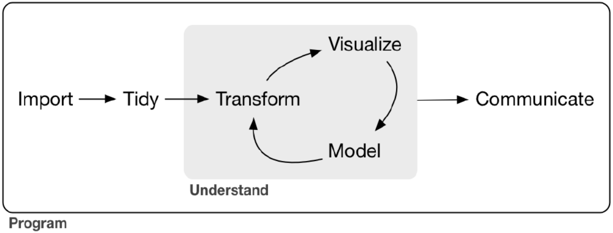
- *Visualization:*
  - it is a fundamentally human activity;
  - it could show you things that you did not expect or raise new questions;
  - it does not scale particularly well because it requires a human to be interpreted.

Data science: an introduction



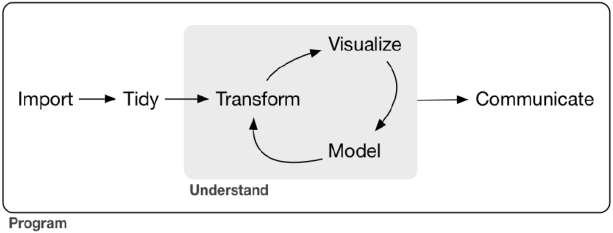
- *Models:*
  - they are a complementary tools to visualization;
  - Once you have made your questions sufficiently precise, you can use a model to answer them;
  - They are a fundamentally mathematical or computational tool, so they generally scale well.

Data science: an introduction



- *Communication:*
  - it is critical part of any data analysis project;
  - it does not matter how well your models and visualization have led you to understand the data unless you can also communicate your results to others.

Data science: an introduction



- These steps are typically carried out using a *mix of languages* (e.g. R, Python, Julia, . . .)
- It is important to master one tool at time;
- R is a great place to start: it is not just a programming language, but it is also an interactive environment for doing data science.

Rectangular Data

- Rectangular data are a collection of values that are each associated with a variable and observation;
- In this course we focus exclusively on rectangular data;
- There are datasets that do not fit on this paradigm (e.g. images, sound, . . .)

Hypothesis generation vs Hypothesis confirmation

- Hypothesis generation or data exploration generates many interesting hypotheses to help explain why the data behaves the way it does;
- Hypothesis confirmation studies if a hypothesis is confirmed or not;
- Commonly modeling is considered a tool for hypothesis confirmation, and visualization a tool for hypothesis generation;
- This is false dichotomy: models are often used for exploration, and with a little care visualization can be exploited for confirmation.