PSS718 - Data Mining

Lecture 1

Asst.Prof.Dr. Burkay Genç

Hacettepe University, IPS, PSS

October 3, 2016

Data Science is Art

Science is knowledge which we understand so well that we can teach it to a computer.

Donald Knuth

Data Science is Art

- We have many tools for doing data science
- But, we have to know how to use them to solve a problem

Steps of Analysis

Data analysis is an iterative and interactive process.
There is no recipe that cures them all.

Steps of Analysis

- State and refine a question
- Explore the data you have
- Build formal statistical models
- Interpret the results
- Communicate the results

Data Mining

Data mining is the art and science of intelligent data analysis.

The aim is to discover meaningful insights and knowledge from data.

Anywhere we collect data, data mining is being applied and feeding new knowledge into human endeavour.

Data->Info->Knowledge

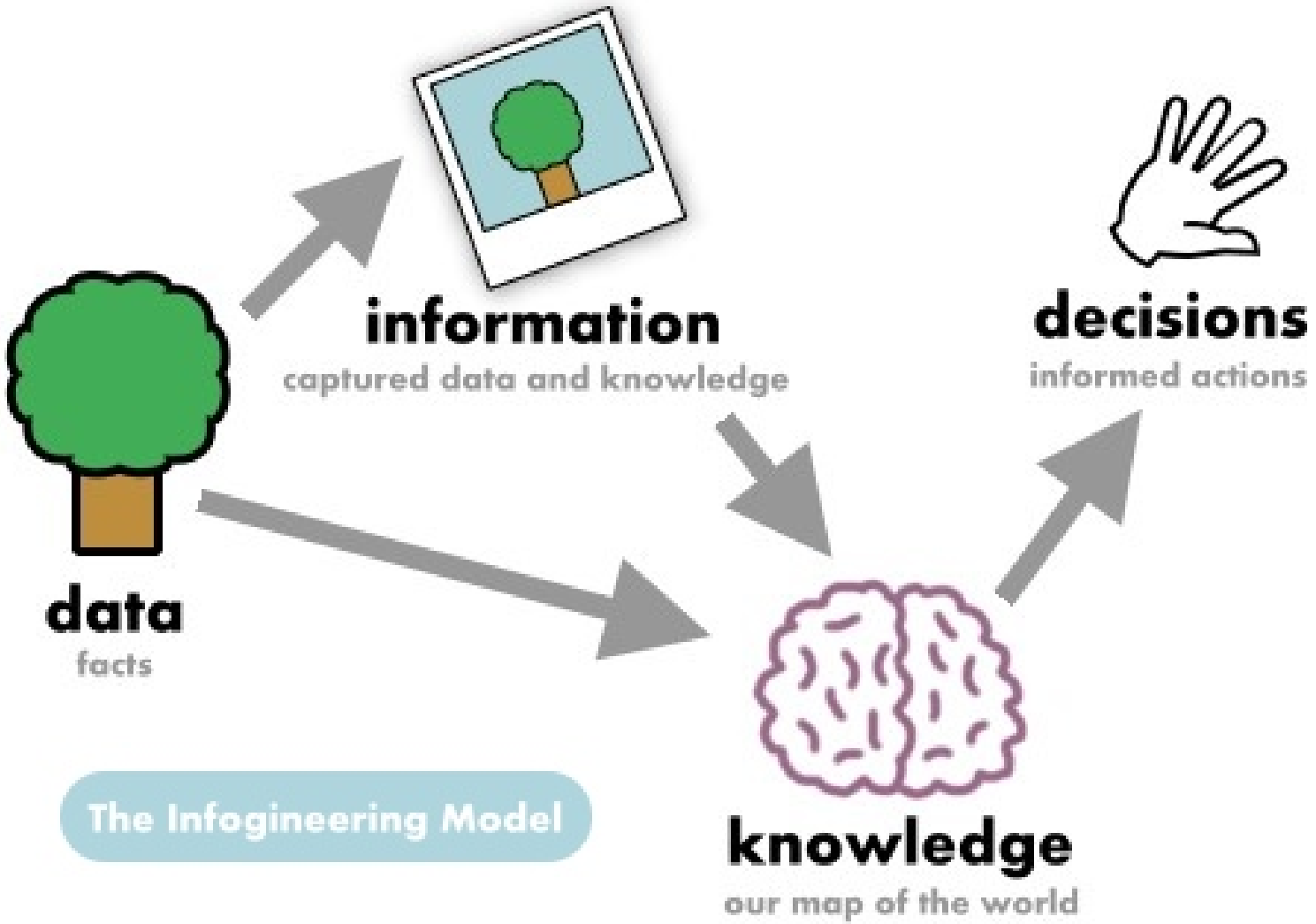

Figure: Data -> Information -> Knowledge

Steps of Data Mining

According to CRISP-DM (Cross Industry Process for Data Mining):

- Problem Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployement

Business vs Science

- Data Miners do not have deeper knowledge of the business
- Business Owners have very little knowledge of data mining
- You need to give enough time to both for understanding each other

Documentation

- Documentation is crucial
- Always document each step!
- Document for you
- Document for team members
- Document for business owner
- Document for public

Script vs Interactive

- Prefer scripts over interactive mode
- Scripts can be saved and later reused
- Scripts also work as a documentation

Why R?

- Open source doesn't mean cheap!
- Open source means peer reviewed

Why R?

- R is the most comprehensive software available
- Graphical capabilities of R are outstanding
- Has over 4800 packages
- R is cross platform
- R is friendly to almost all other tools
- R has extensive documentation on the web

Downsides

- R is difficult to learn
- Most of the time you need to know statistics to understand what is going on
- Some packages may not be of top quality
- R may be a bit too hungry for memory

Load Rattle

```
> library(rattle)
Rattle: A free graphical interface for data mining with R.
Version 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
> rattle()
```

Example

- On the Data tab, choose Library option
- Choose weather:rattle dataset
- Click on Execute: This will load the data into R
- Go to the Model tab and click on Execute
- You have made your first analysis on Rattle!

Loading a Dataset

- Click on "New" to start a new project
- Click "Execute" on Data tab to automatically load the weather data
- Data Type tells you the automatically assigned type
- Inputs are your independent variables
- Target is your dependent variable
- Ident is a variable that uniquely identifies each row
- Ignore is ignored because it doesn't carry any information