Slides for Day 2

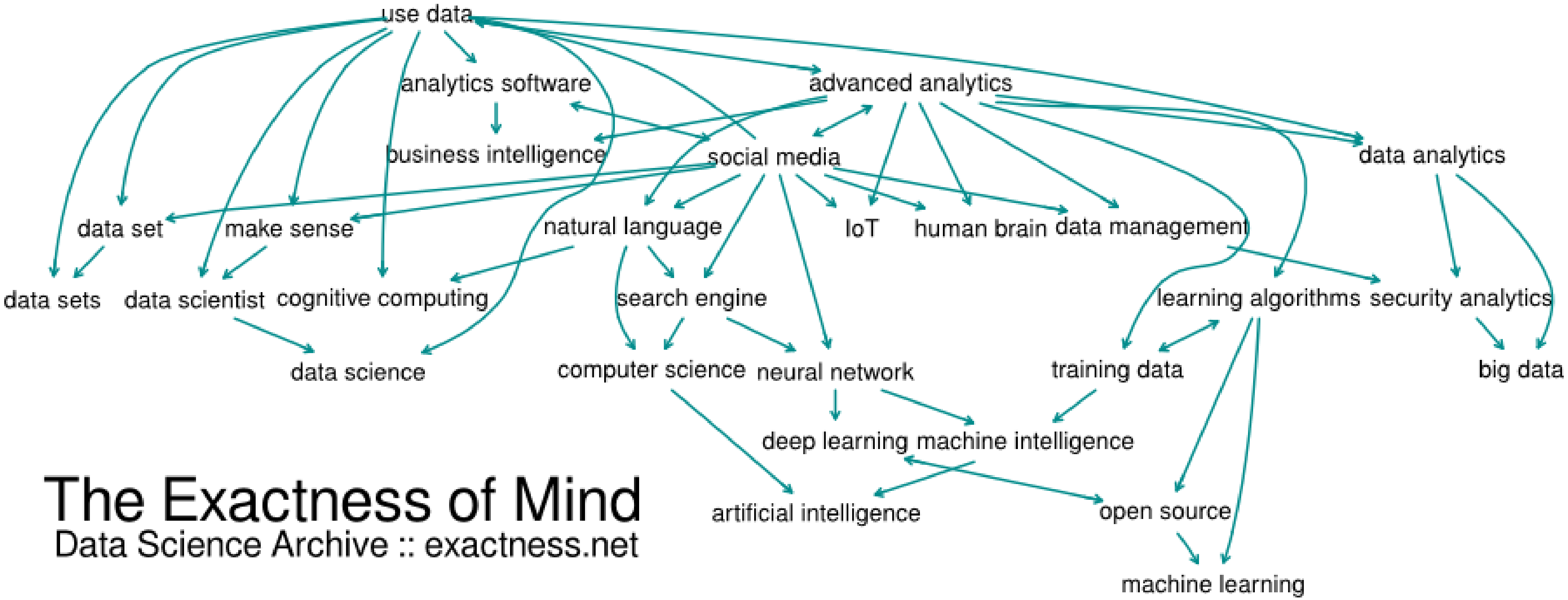
# Statistics with R

## A fast route to Data Science

Montserrat Guillen

Dept. Econometrics, Statistics and Applied Economics &
Riskcenter, UB

# What do we need?



The Exactness of Mind
Data Science Archive :: exactness.net

# What did we do?

- **RStudio projects, working directory, scripts and packages**
- **Data structures:** **vectors, matrices, data.frames, lists, objects**
- **Data wrangling:** **injestion+digestion**
- **R programming:** **if-the-else, loops, functions… a detour to graphics**
- **Rmarkdown:** **producing HTML, PDF, LateX, PPT**

# What will we do today?

- **A case study: Bank telemarketing**
- **Practice exploratory data analysis (EDA): what is in my data?**
- **Predictive modeling: Is there noise or is there something else?**
- **Other fancy models out there: decision tree model, random forests, support vector machine, Bayesian networks, neural networks,**
- **Prediction and cross-validation**
- **Build a package and deal with Spark**

# Today’s: Rmardown files

| | |
|---|---|
| Prog-07.Rmd | EDA bank data |
| Prog-08.Rmd | Logistic regression: bank data |
| Prog-09.Rmd | Further models: bank data |
| Prog-10.Rmd | Prediction and crossvalidation: bank data |

# Let’s create an R package

Doc-05.pdf

- Collect functions
- Create the package directory (easy if you install things before that or use RStudio)
- Document the functions
- Build process and install
- **Make the package a GitHub repository**

  **or Contribute to CRAN**
- **An example with our course**

R4DSUB.zip

# R and Spark

**Sparklyr is an R interface for Apache Spark, you can:**

- Connect to Spark from R. The sparklyr package provides a  complete dplyr backend.
- Filter and aggregate Spark datasets then bring them into R for analysis and visualization.
- Use Spark's distributed machine learning library from R.
- Create extensions that call the full Spark API and provide interfaces to Spark packages.

Once you have connected to Spark, then copying and interacting is super-fast and easy

Doc-06.pdf

# Python and/or R?

- Both can be used: There were a number of Python module choices to access R. They are: rpy2, pyRserve and PypeR.
- From R, Python can also be used:

rPython - an R package which allows the user to call Python from R

# References
# Statistics with R

- *http://rstudio.com/cheatsheets*

- **Introduction to R for Python Programmers**
  http://ramnathv.github.io/pycon2014-r/
- **The Art of R Programming** Norman Matloff
- **R in action**, Robert I. Kabacoff, Manning Publications
- **Introductory Statistics with R**, Peter Dalgaard, Springer
- **Data Analysis and Graphics using R** , John Maindonald &
  W. John Braun, Cambridge University Press
- **The R Book**, Michael J. Crawley, Ed. John Wiley & Sons
- **R for dummies**, Joris Meys, Andrie de Vries
- **Beginning R: An Introduction to Statistical Programming**, Larry Pace, Apress
- **Beginning R: The Statistical Programming Language**, Mark Gardener, Wrox

R YOU SURE?

- **If I want to upgrade my data analysis skills, which programming language should I learn?**

# Statistics with R

Some favorite quotes:

http://www.ub.edu/riskcenter/guillen
mguillen@ub.edu
@mguillen_estany

*“There are no routine statistical questions, only questionable statistical routines.”*
**Sir David Cox**

*"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."*
**John Tukey**

“All models are wrong, but some are useful. “
**George E. P. Box**