Data Science and Open Source Software

Iraklis Varlamis
Assistant Professor
Harokopio University of Athens
varlamis@hua.gr
What is data science?
Why data science is important?

More data (volume, variety,...) ⇒ better analysis ⇒ better decision making

New data (trends, events) ⇒ faster analysis ⇒ new opportunities
The skills of a data scientist *

Non-technical skills

● **Education**: Computer Science, Mathematics, Statistics
● **Domain Knowledge**
● **Intellectual Curiosity**
● **Communication Skills**

● Balance between technical, communication and presentation skills

Technical skills

● **Programming**: R, Python
● **Machine Learning/Data Mining**: scikit-learn, KNIME, TensorFlow, Theano, PyTorch
● **Big Data Processing**: Hadoop, Spark, Flink or Storm
● **Data management**: SQL, noSQL
● **Data Visualization**: D3.js, Chart.js, Tableau, ggplot or lattice in R, matplotlib, plotly or seaborn in Python

*According to Matthew Mayo @ KD Nuggets: https://www.kdnuggets.com/2016/05/10-must-have-skills-data-scientist.html
There is no “one ring to rule them all”

Statisticians choose R:

- Many statistical tests and models
- Better in ad hoc analysis
- New statistical models can be written in a few lines
- Comprehensive R Archive Network (12576 packages)
  - https://cran.r-project.org/

Programmers choose Python:

- Lots of code in stackoverflow
- Better data manipulation
- Supports scripting website or other applications
- 140,784 projects in Python Package Index (PyPI)
  - https://pypi.org/

Run Python from R
https://github.com/rstudio/reticulate

Run R from Python
https://pypi.org/project/rpy2/
Crunching Big Data and Analytics

**Store**
- Document databases (e.g. MongoDB, CouchDB): free-form JSON documents
- Key-value stores (e.g. Redis, Memcached): from integers or strings to JSON documents
- Wide column stores (e.g. Cassandra, HBase): data in columns instead of rows
- Graph databases (e.g. Neo4j): networks or graphs of related entities
- Search engines (e.g. Elasticsearch, Solr): allow full-text or faceted search, search by synonyms etc.

**Process**
- Distributed processing and cluster-computing
- Hadoop Common, Hadoop Distributed File System, Hadoop YARN (job scheduling and cluster resource management), Hadoop MapReduce (parallel processing)
- Apache Spark
  - With native bindings for Java, Scala, Python and R
  - Supports SQL, streaming data, machine learning, and graph processing
Crunching Big Data and Analytics

Analyze and Visualize

- **ETL**
- **Explore**: Examine data sets to gain insights
- **Learn and predict**: Find trends, predict future activity, Predictive analytics
- **Visualize**

Process on the fly - Stream analytics

- **Orchestrate** (messaging): Kafka
- **Stream data** (or in mini-batches): Flink, Storm or Spark streaming
- **Process and predict**: Spark MLLib
- **Store results**: MongoDB, Cassandra
- **Visualize**: D3.js, Plotly
When it comes to data streams (e.g. IoT)

Source: InfoQ.com
When it comes to social networks

Networks are graphs, and social network analysis requires graph operations at large scale!

Distributed storage and processing is the only solution.

ETL as graph traversal

Storage
Cassandra, HBase, ...

Indexing
Elasticsearch, Solr
When it comes to document streams (e.g. tweets)

The ELK stack

Twitter Streaming → logstash → elasticsearch → kibana

Text Mining NLP

When it comes to Machine Learning and AI
“Open Source Is The New Normal...

...In Data and Analytics”, Scott Gnau, CTO at Hortonworks, @ Forbes, July 2017

● And how Open Source makes money?
  ○ Customization of Code
  ○ Sell Value-Added Enhancements
  ○ Provide Support
  ○ Sell Documentation
  ○ Sell Binaries
  ○ Offer OSS

● Open Source S...
Open Source Services

- Accelerated product development by eliminating product dependencies
- Agile, Innovative and cost-effective development
- Cloud services (Amazon, Microsoft, Google, IBM, Salesforce)
Thank you and questions