Python for Data Science
Introduction: What is Data Science?

Heru Praptono

heru.pra@cs.ui.ac.id

Salemba, December 2018

Outline

- Introduction
  - Our Course Roadmap
  - Motivating Benefit: Examples
- Overview
  - What is it exactly?
  - Data Science vs Data Mining vs others
  - Data Mining: Problem Types
- Learning Tips

Our Course "Chestpass" Topics

- Day 1
  - Intro to Learning from Data
  - Data Collection & Preprocessing
  - Exploratory Data Analysis
- Day 2
  - Variables and Feature Analysis
  - Clustering
  - Regression & Classification
- Day 3
  - Text Mining
  - More: Selected Topics
  - Common Practices in Data Science

Our Course Hands On

- The course features **extensive hands-on sessions**
- The obtained knowledge **can be applied to real-world problems**
- The hands-on will be using **Python programming language**

... with no prior knowledge of Python is necessary as Python will be gently introduced throughout the whole sessions

Motivating Benefit: Examples

- Reduce time to market
  - **Bristol-Myers Squibb**, a global biopharmaceutical company, reduced the time it takes to run clinical trial simulations by 98%
- Optimize the workforce
  - Xerox used data science to reduce the turnover rate in its call centers by 20%.
    - had to understand what was causing the turnover, and determine ways to improve employee engagement
    - data allows them to select new hires that are a better fit for the company, reduce employee turnover, understand the skills and output of the existing workforce, and determine the talent the company needs moving forward

Motivating Benefit: Examples

- Improve financial performance
  - **The Weather Company and IBM**, will allow companies to better manage the impact of weather on business performance.
    - According to The Weather Company, weather has an economic impact of half a trillion dollars annually in the US.
    - The weather data is being collected from weather sensors and aircraft, as well as smartphones, buildings, and moving vehicles.
    - Financial benefits include: Retailers: adjust staffing and supply chain strategies, Energy companies: supply and demand forecasting, Insurance companies: warn policy holders of severe weather conditions

Motivating Benefit: Examples

- Sell intelligently. Slight modifications to sales and marketing strategies can have a profound effect on the bottom line.
  - **Kroger** personalises its marketing based on the shopping history of the individual customer
  - Also has loyalty card program that is rated No. 1 in the grocery industry. More than 90of its customers use its loyalty card when they purchase Kroger products!

... and many others..

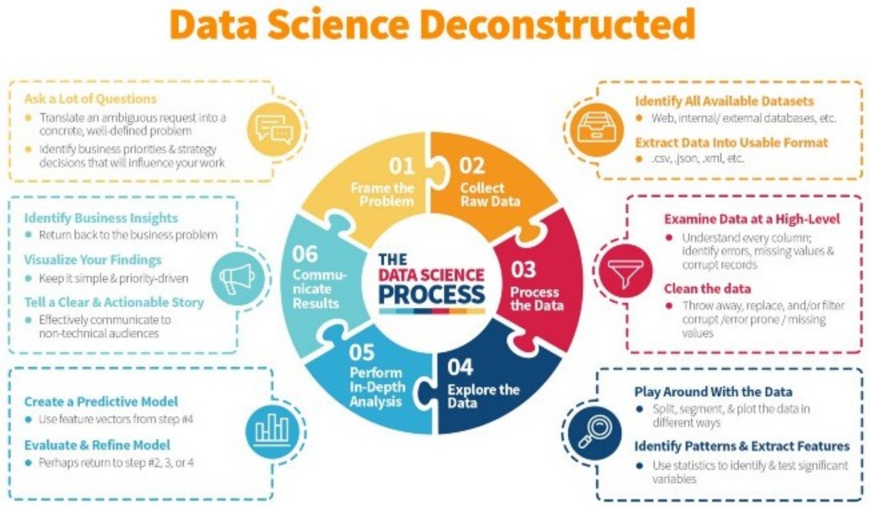
What is it exactly?[1]

The study of the computational principles, methods, and systems for extracting knowledge from data

IRDS Team, Edinburgh

[1] as it is difficult to find another term for this intersection

A DS Deconstruction[2]



Data Science Deconstructed

THE DATA SCIENCE PROCESS

01 Frame the Problem

Ask a Lot of Questions
- Translate an ambiguous request into a concrete, well-defined problem
- Identify business priorities & strategy decisions that will influence your work

02 Collect Raw Data

Identify All Available Datasets
- Web, internal/ external databases, etc.

Extract Data Into Usable Format
- .csv, .json, .xml, etc.

03 Process the Data

Examine Data at a High-Level
- Understand every column; identify errors, missing values & corrupt records

Clean the data
- Throw away, replace, and/or filter corrupt /error prone / missing values

04 Explore the Data

Play Around With the Data
- Split, segment, & plot the data in different ways

Identify Patterns & Extract Features
- Use statistics to identify & test significant variables

05 Perform In-Depth Analysis

Create a Predictive Model
- Use feature vectors from step #4

Evaluate & Refine Model
- Perhaps return to step #2, 3, or 4

06 Communicate Results

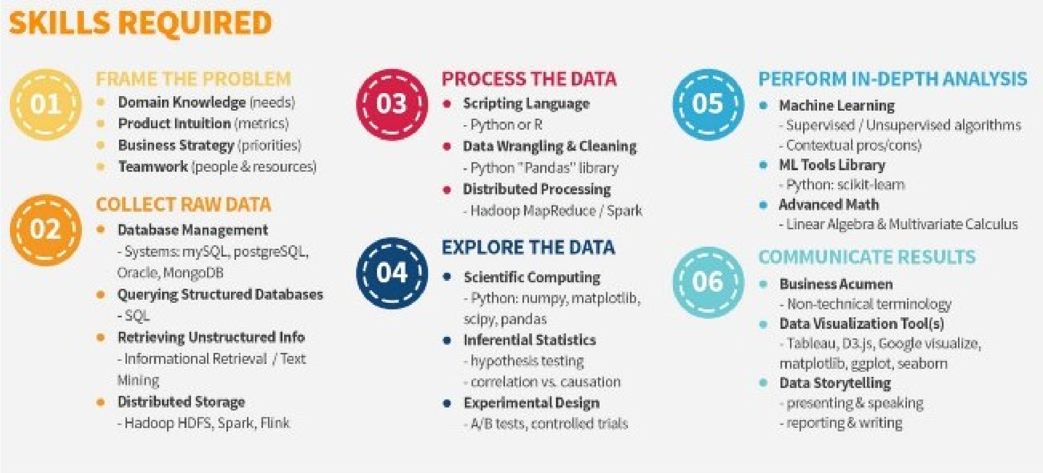
Identify Business Insights
- Return back to the business problem

Visualize Your Findings
- Keep it simple & priority-driven

Tell a Clear & Actionable Story
- Effectively communicate to non-technical audiences

[2] source: https://ajgoldstein.com/2017/11/12/deconstructing-data-science/

Skills



SKILLS REQUIRED

01 FRAME THE PROBLEM

- **Domain Knowledge** (needs)
- **Product Intuition** (metrics)
- **Business Strategy** (priorities)
- **Teamwork** (people & resources)

02 COLLECT RAW DATA

- **Database Management**
  - Systems: mySQL, postgreSQL, Oracle, MongoDB
- **Querying Structured Databases**
  - SQL
- **Retrieving Unstructured Info**
  - Informational Retrieval / Text Mining
- **Distributed Storage**
  - Hadoop HDFS, Spark, Flink

03 PROCESS THE DATA

- **Scripting Language**
  - Python or R
- **Data Wrangling & Cleaning**
  - Python "Pandas" library
- **Distributed Processing**
  - Hadoop MapReduce / Spark

04 EXPLORE THE DATA

- **Scientific Computing**
  - Python: numpy, matplotlib, scipy, pandas
- **Inferential Statistics**
  - hypothesis testing
  - correlation vs. causation
- **Experimental Design**
  - A/B tests, controlled trials

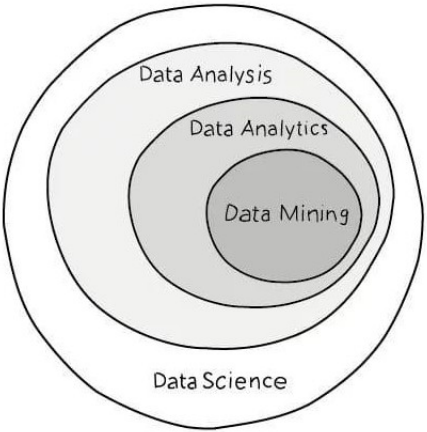
05 PERFORM IN-DEPTH ANALYSIS

- **Machine Learning**
  - Supervised / Unsupervised algorithms
  - Contextual pros/cons)
- **ML Tools Library**
  - Python: scikit-learn
- **Advanced Math**
  - Linear Algebra & Multivariate Calculus

06 COMMUNICATE RESULTS

- **Business Acumen**
  - Non-technical terminology
- **Data Visualization Tool(s)**
  - Tableau, D3.js, Google visualize, matplotlib, ggplot, seaborn
- **Data Storytelling**
  - presenting & speaking
  - reporting & writing

Data Science vs Data Mining vs others[3]



[3] source: https://www.quora.com/What-are-the-differences-between-Data-Science-and-Data-Mining-are-they-same

Data Mining

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner.

Hand, Mannila, Smyth, 2001

Data Mining: Problem Types

- Visualization
- Supervised: Prediction - Learn map $\mathbf{x} \longrightarrow y$
  - Classification: Predict categorical value
  - Regression: Predict a real value
  - Others:
    - Collaborative filtering
    - Learning to rank
    - Structured prediction
- Unsupervised: Description
  - Clustering
  - Dimensionality reduction
  - Density estimation
  - Finding patterns
    - Association rule mining
    - Detecting anomalies / outliers

Learning Tips

1. Learn statistics
2. Learn programming
3. Study, study, practice, practice, and practice even more
4. Stay hungry and curious
5. Put some structure to your learning
6. Join a community/meet up or get a mentor
7. Get on Kaggle/competitions as soon as you can
8. The best time is now
9. Keep going, never give up

Thank You