BITS Pilani
Pilani Campus

Introduction to Data Science

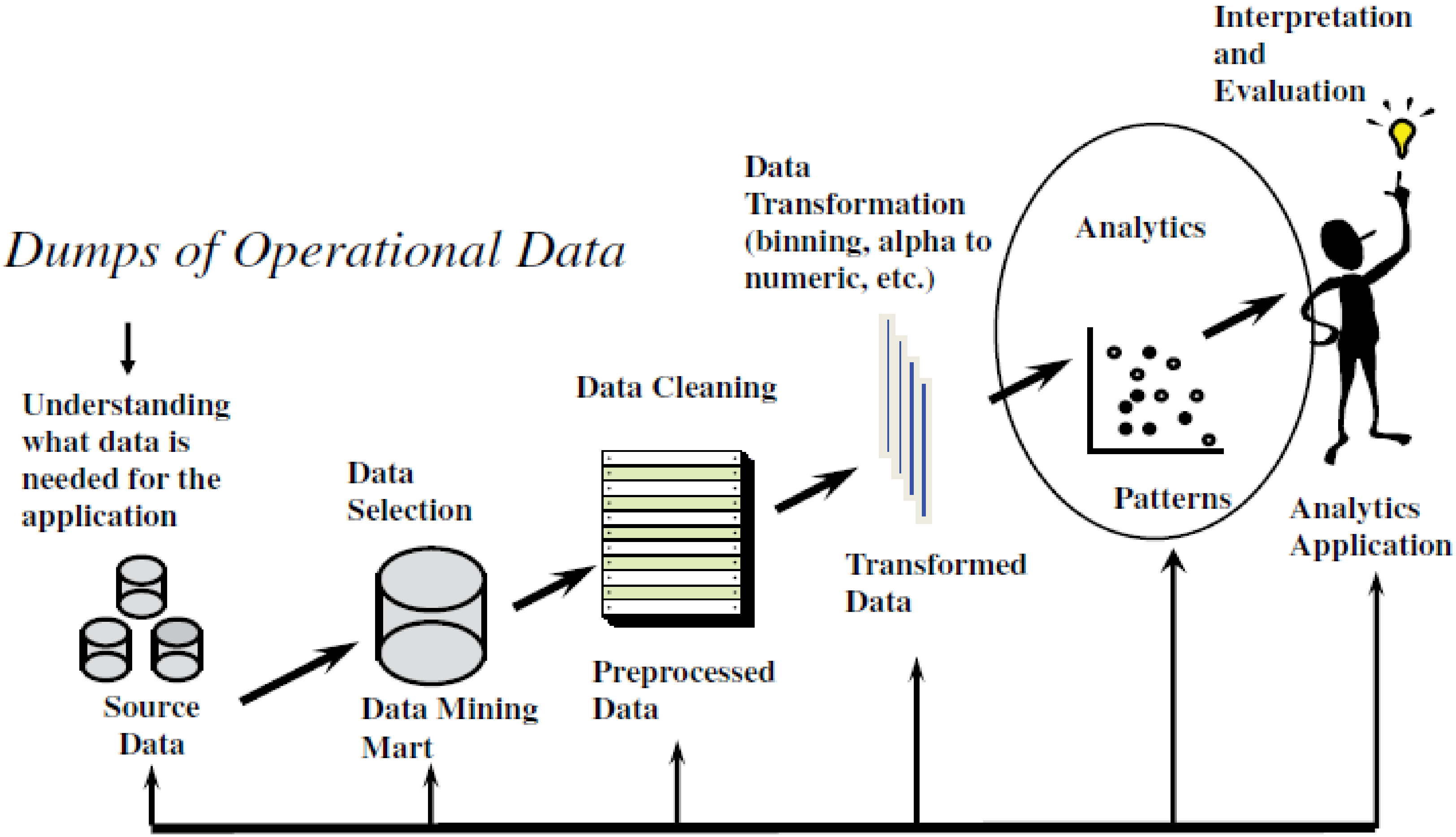
(BA ZG523 / CSI ZG523)

Tirtharaj Dash
Dept. of Comp. Sc. and APP Center for A.I. Research
BITS Pilani, K.K. Birla Goa Campus

Introduction to Data Science

Lecture No. 02 (C.H.: 3, 4)

Analytics Process Model


Interpretation and Evaluation
Dumps of Operational Data
Data Transformation (binning, alpha to numeric, etc.)
Analytics
Understanding what data is needed for the application
Data Cleaning
Data Selection
Patterns
Analytics Application
Transformed Data
Preprocessed Data
Source Data
Data Mining Mart

Analytics Process Model (Source: T4, Chapter 1)

Analytics Process Model

- Source data: Data of interest
- Data mart: Gather data in a staging area or warehouse
- Cleaning: Remove inconsistencies (missing values, redundancies, outliers)
- Transformation: Represent data in a format that can be processed by a program at later stage, e.g. alpha to numbers, numeric encoding
- Analytics: Build an analytics model. Find patterns, sub-groups, market-basket analysis*, etc.
- Interpretation and Evaluation: Evaluate and interpret the results of the analytics model.

*Purchases that commonly happen together; e.g. brown bread and butter are bought together.

Roles and Responsibilities of a Data Scientist

A data scientist

- Works closely with business stakeholders.
- Understands business goals.
- Determines how data can be used to achieve these goals.
- Designs data modeling processes.
- Creates algorithms and predictive models to extract the data the business needs
- Help analyze the data and share insights with peers.

Specific activities they do:

1. Data gathering, preparation, and exploration
2. Data representation and transformation
3. Computing with data
4. Data modeling (using ML, AI, Statistics)
5. Data visualization
6. Presentation (to the business stakeholders)

Ethical Guidelines

There are **six** main ethical issues identified in [1]:

1. Autonomous machines: a threat to free will and responsibility?
    - Delegation of complex and critical decisions and tasks to machines *increases* the human capacity to act and poses a threat to human autonomy and free will and may water down responsibilities.
    - E.g. autonomous vehicles: (Q) Who is responsible in case of an accident? (Q) Will it be possible to overrule a machine's decision on lowest or allowable risk, i.e. in case of an emergency.

[1] Garzcarek, U., & Steuer, D. (2019). Approaching Ethical Guidelines for Data Scientists. In Applications in Statistical Computing (pp. 151-169). Springer, Cham. (https://arxiv.org/abs/1901.04824)

Cont.

2. Bias, discrimination and exclusion

- Algorithms and artificial intelligence can create biases, discrimination or even exclusion towards individuals and groups of people.
- Judgment and predictions are impacted by prior beliefs (and prejudices). These prior beliefs can sneak in to programs.
- Example 1. Bank loan application getting rejected even after satisfying eligibility requirements.
- Example 2. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) a software used in the US: … decisions by the judges whether defendants could get out on parole or had to go to jail were strongly influenced by the algorithm’s output and discriminated against black people. (Read more in [1])

Cont.

3. Algorithmic profiling

- Personalizing versus collective benefits: Individuals have gained a great deal from profiling and ever finer segmentation. This mindset of personalising can affect the key collective principles like democratic and cultural pluralism and risk-sharing in the realm of insurance.
- Example 1. Filter-bubble. An Internet user encounters only information and opinions that conform to and reinforce their own beliefs, caused by algorithms that personalize an individual’s online experience.
- Example 2. The scandal around Cambridge Analytica using Facebook data for micro-targeting a very specific subset of the public with the aim to influence the US elections in 2016.

Cont.

4. Preventing massive files while enhancing AI: seeking a new balance
   - AI by being based on advanced techniques of machine learning requires a significant amount of data. Still, data protection laws are rooted in the belief that individuals' rights regarding their personal data must be protected and thus prevent the creation of massive files.
   - Example. Studies in medical sciences have shown that true benefits for public health that can be generated from using large medical databases.

Cont.

5. Quality, quantity, relevance: the challenges of data curated for AI
    - The acceptance of the existence of potential bias in datasets curated to train algorithms is of paramount importance.
    - Example. Amazon built a system to help find the most qualified applicants in their huge stream of applications. But, the algorithm systematically downgraded applications of women. Probable causes: (C1) The training data contained mostly applications of men, so most of the successful applicants were men; (C2) Any appearance of the word woman reduced the chances of that applications.

Cont.

6. Human identity before the challenge of artificial intelligence

   - Hybridization between humans and machines challenges the notion of our human uniqueness. How should we view the new class of objects, humanoid robots, which are likely to arouse emotional responses and attachment in humans?
   - However, this is not an ethical issue where data scientists have a special responsibility due to their expertise.

Intermission
(10m)

Last homework:

- I assume:
  - You have a computer system (with Linux or Windows or Mac OS)
  - You know how to type using computer keyboard
  - You are already enjoying this course

- Task 0: Install Python latest (3.x) on your machine. You will find it here:

  https://www.python.org/downloads/

- Task 1: Practice (Chapter 1 and 2) from:

  https://www.py4e.com/

  (this is an engaging book on Python)

Has anybody done it?

Let’s play a bit, shall we?

- Which of the following is/are not part(s) of a computer system?

A. Input device (e.g. keyboard)
B. Output device (e.g. display monitor)
C. Main memory (also called Random Access Memory)
D. Secondary memory (e.g. hard disk)
E. Web browser (e.g. Chrome, Firefox)

Answer: E

(Web browser is a subsidiary application that is not required to operate a computer system.)

- I am watching a movie on my computer. The movie is half-way. Suddenly, due to some issues, my computer got shutdown. After restarting, I find that:

  A. The movie is intact. I can start the movie from where I had left.

  B. I have to reopen/reload the application that was playing the movie.

  Answer: B

  (Computer does not remember which applications were open)

- Why did you have to restart the application?

  A. Because, the application had some issues.

  B. The main memory did not contain the application anymore.

  Answer: B

  (Information stored in main memory vanishes when computer is turned off)

- We want to ask a computer: Find out the average of the following 5 numbers: 10, 4, 22, 67, 11. What are the two important things you need (assume that you have a fully working computer system that can take instructions)?

  A. A language in which to talk, and a set of instructions in that language

  B. An interface, a language and a set of instructions in that language

  C. A set of instructions in English language

  D. It is my computer. It knows what I want.

Answer: B

('A' is not completely correct. You also need an interface + 'A')

Example: Python is a (programming language). Terminal/Command Prompt or any other IDE is an interface. The program that you write for computing average is a set of instructions.

- You opened a Terminal window (in Linux) or Command Prompt (in Windows) or "whatever!" (in Mac). Then, you typed the command: `python<enter>`. Which of the following does it show and what does this mean?

A. $ and it means that command is successful.

B. > and it means that command is successful.

C. >>> and it means that Python interpreter is waiting for next command.

D. >>> and it means that Python interpreter is not working.

Answer: C

- How to check version of `Python` you are working on?
  A. On Terminal type: `python --version <enter>`
  B. On Terminal, if you type `python <enter>`, it automatically shows the version.
  C. Right click on python icon and click ‘version’
  D. I just think about python and a version number appears on my screen.

Answer: A and B.

- Computer cannot understand any other language than a machine language (some binary executable code; human-unreadable). The conversion from a written program to machine instruction is done either by a compiler or an interpreter. Python is

  A. a compiler

  B. an interpreter

  C. a machine language. It doesn’t need to be converted.

Answer: B

(An interpreter interprets instructions on the fly)

(A compiler takes full program –a file– and converts it to machine language)

- When we write `x = 123`, where does this variable x gets stored?
  A. CPU
  B. Main memory
  C. Hard disk
  D. Mouse memory

Answer: B

(All variables are stored in main memory so that they can be used during program execution.)

(Because, the program itself is loaded in the main memory)

- What will be the output of the following program?

```
x = 20
x = x + 1
print(‘x’)
```

A. 21
B. 20
C. x + 1
D. x

Answer: D

(Because, the argument inside the print function call is of string type)

(If you intended to print the updated value of x, you should remove the quotes.)

Python Demo

- We will now play with some Python syntaxes and semantics.
- At the end of this demo, we should be able to solve simple problems using Python.

- For my convenience, I will be using jupyter-notebook. If you also want to use jupyter-notebook for your work, you can. The information for this is provided at the end of this lecture slide.
- All the programs that I am playing around here will be made available via my GitHub repository page.

Course Repo

- I have created a dedicated GitHub repository for this course. Please find it here:

  https://github.com/tirtharajdash/IntroductionToDataScience
- README.md file will give more information on this.