COMP 333 — Week 1 Nutshell

Data Analytics in a Nutshell

This lecture provides an overview of Data Analytics
to let you orient yourself for COMP 333
and see what is important and where to focus your efforts.
To quote the course outline:
"The aim of this course is to introduce students
to the Python programming language
and related tools for data analytics; and
to expose them to a broad range of data analysis problems across a range of disciplines."

Why?

Because

Data Analytics has permeated into every industry, government, and business function.
The future will need data-driven approaches for all fields of human endeavour.

What is Data Analytics?

The aim of data analytics is to add value to your data
so it becomes **actionable** data
which means it helps you and your organisation to make decisions.
You will see it termed as *"monetization of data"* in the business world.
The main steps of the data analytics are
- descriptive data analysis
- data wrangling
- exploratory data analysis
These steps fit into an overall data analytics process
where you combine an understanding of the data and the business
to come up with data-driven input into the decision-making of the organization.

THE IMPORTANT THINGS

This lecture is an overview.
It is an orientation of what is to come.
You are not meant to understand everything in this document today!
Each topic will be done in much more detail again later in the semester.

Descriptive Data Analysis (DDA)

DDA is a basic tool for understanding your data

DDA is used throughout all stages of data analytics.

Be aware of the type of data that you have:

- categorical versus continuous

– categorical: nominal versus ordinal

– continuous: interval versus ratio

- structured versus unstructured

and for numerical values, be aware of

- accuracy
- precision
- significant digits

Describe the data and the data distribution for each feature in the dataset

- central tendency: mean, median, mode
- variation: standard deviation, inter-quartile range (IQR)
- outlier values and extreme values
- skew
- kurtosis

You want descriptions that are *robust* to presence of outliers

Visualization as box-plots, violin plots, histograms, scatter plots

Data Wrangling

Data wrangling is extremely important because your data is typically “messy”
and remember Garbage-In-Garbage-Out (GIGO) rule for computation
so you need to tidy-up your data before doing “serious” work.

Data wrangling is generally 60%+ of the time and effort for data analytics!

For data wrangling, you need to look closely at your data, so DDA is a basic tool.

Steps in data wrangling:
Step 1: Discover
Step 2: Structure
Step 3: Cleanse
Step 4: Enrich
Step 5: Validate
Step 6: Publish

Issues for data cleaning:

- errors in data
- outliers and anomalies
- missing values and imputation of missing values
- unification and normalization so data is comparable
- entity recognition

Data wrangling is the traditional ETL (Extract-Load-Transform) process
from data warehouses and OLAP (online analytical processing).

The output of data wrangling is formatted as *“Tidy Data”*
which has three basic properties:

1. Each variable is saved in its own column
2. Each observation is saved in its own row
3. Each type of observation is stored in its own (single) table

Exploratory Data Analysis (EDA)

EDA grew out of the statistics community.

EDA is the heart of data analytics.

EDA involves data wrangling and descriptive data analysis.

EDA develops a data-driven solution to your problem

by exploring the data to find which features lead to a solution.

The steps of EDA

Step 1: Data wrangling: collect, load, enrich data

Step 2: Descriptive data analysis: check data types, check distributions

Step 3: Feature engineering

Step 4: Modeling

Step 5: Story-Telling

A checklist for EDA:

Q1. What question(s) are you trying to solve (or prove wrong)?

Q2. What kind of data do you have and how do you treat different types?

Q3. Whats missing from the data and how do you deal with it?

Q4. Where are the outliers and why should you care about them?

Q5. How can you add, change or remove features to get more out of your data?

The Data Analytics Process

The data analytics process is how the business community looks at data analytics.

Step 1: Business Understanding: What are the business goals and problems?

Step 2: Data Understanding: Explore and visualize the data.

Step 3: Data Preparation: Generate features

Step 4: Modeling: Create models

Step 5: Evaluating: Train models and evaluate effectiveness

Step 6: Deploying: Use this data-driven approach for the goal of the business on a regular basis.

This can be viewed as a highly iterative cycle:

- Define the Goal: What problem are you solving?
- Collect and Manage Data: What information do I need?
- Build the Model: Find patterns in the data that lead to solutions.
- Evaluate and Critique the Model: Does the model solve your problem?
- Present Results and Document: Establish that you can solve the problem, and how.
- Deploy Model: Deploy the model to solve the problem in the real world.

THE LESS IMPORTANT THINGS

Models and Machine Learning

Story-Telling and Visualization

Deployment and Big Data Infrastructure

THE LESS LESS IMPORTANT THINGS as these provide mainly context

Correlation, Causality, and Confounding Factors

Data Warehouses and Business Intelligence

Confirmatory Data Analysis

that is, the scientific method with

planned (not *exploratory*)

experimental design, data collection, and data analysis

Descriptive vs Predictive vs Prescriptive Data Analysis

Descriptive Data Analysis is describing your data from past activities, provides insight into the past and answer: *"What has happened?"*

Predictive Data Analysis provides results for unseen data for future activities, uses statistical models and forecasts to understand the future and answer: *"What could happen?"*

Prescriptive Data Analysis models viable solutions to a problem and the impact of considering a solution, uses optimization and simulation algorithms to advise on possible outcomes and answer: *"What should we do?"*