Data Collection and Modeling

Course 1 - Intro (Course & Concepts)

Why Would You Be Interested?

Stored data will reach 175 zettabytes in 2025

Data can be repurposed, reused, related, new insights can be generated

Useful in all domains

What Does DCM Mean?

It consists of two parts:

- Collection
- Modeling

Course/Lab Info

Course

- Theoretical concepts, case studies, demos
- 50% of final grade
- Exam (most likely quiz using Moodle)

Lab

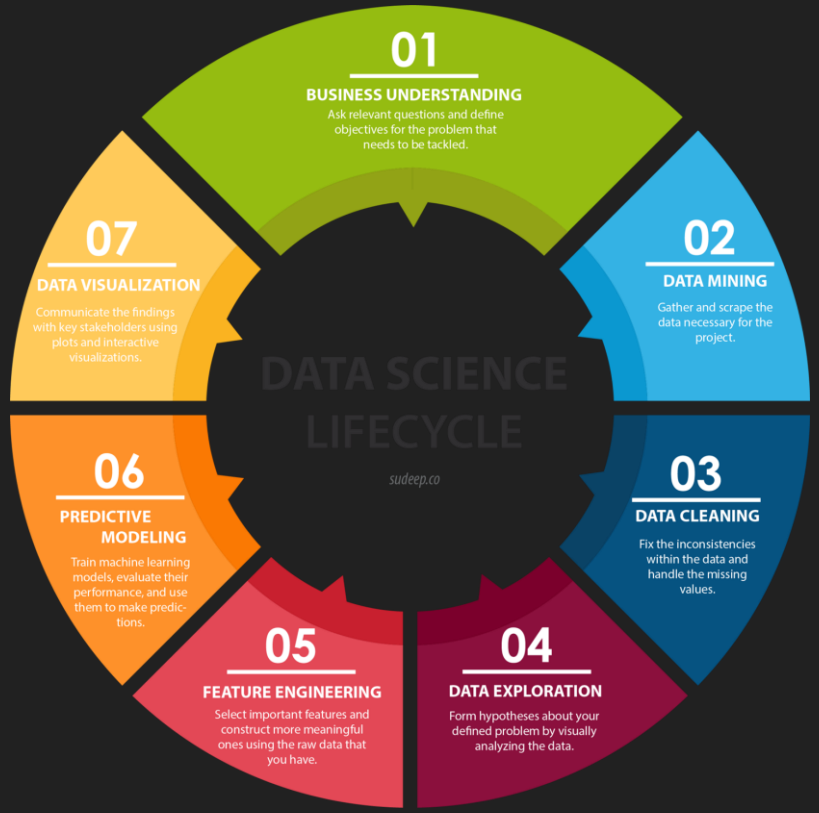
- Hands-on
- 50% of final grade
- Most likely project based

Topics

- Introduction
- File types; data sharing
- Local vs remote acquisition
- Data transport formats
- Data cleaning
- Web crawling
- APIs: REST vs SOAP
- SQL
- Data types and structures

Introduction

Data Science Lifecycle



Source: https://www.sudeep.co/data-science/2018/02/09/Understanding-the-Data-Science-Lifecycle.html

Data Collection

“Data collection is the process of gathering data for use in business decision-making, strategic planning, research and other purposes. It is an important part of data analytics applications and research projects”

(In Data Science) “Data collection is the process of accumulating data that's required to solve a problem statement”

Data Collection Steps

- Formulate the problem
- Determine data type to collect
- Identify data sources
- Set a time frame for data
- Collect data

Data Collection Types and Methods

- Primary Data
  - Surveys
  - Interviews/Focus Groups
  - Observations
- Secondary Data
  - Track Transactional Data
  - Social Media
  - Online Activity Tracking
  - Forms (Subscription, Registration)

Quantitative Data

Quantitative data relates to information that can be quantified (as numbers). It can be measured or counted, thus given a numerical value.

Quantitative data can be used to answer "how many", "how much" or "how often" type of questions.

Examples:

- How many students received offers after internship?
- How much will be the annual income for such a student?
- How often does a company organize internships?

Some of quantitative data collection methods are: surveys, experiments, polls.

Qualitative Data

Qualitative data has a descriptive nature, it can't be expressed as numerical values.

It is used to describe information that can't be measured or counted. It usualy refers to certain characteristics.

Quantitative data can be used to answer "why" or "how" type of questions.

Examples:

- The opinion of a student about its internship period (good, bad, worthy)

Some of qualitative data collection methods are in-depth interviews, focus groups, observations.

Qualitative vs Quantitative Data

Qualitative (Categorical) Data

- Nationality
- Gender
- Native Language
- High-School Specialization

Quantitative (Numerical) Data

- Age
- Height
- Weight
- Admission Grade

Data Collection Plan

2 main methods:

- diagram: visual, maps the flow of information

- plan table: analytical, list of variables by source

A workflow diagram starts with what data is being collected (quantitative/qualitative) and follows through from how it is collected to where it is stored and how it is shared(reporting presentation/online dashboard).

The plan outline consists in filling out a plan table. It helps organizing variables by source, method of acquisition, timeline, storage, and how it is analyzed and shared.

Information Workflow Diagram

Provides a graphic overview of the process

Uses symbols and shapes to depict the steps to complete from start to finish

Shows who is responsible for work at each step in the process

Each element of a workflow illustrates the flow between each step. Each step includes one of three parameters:

- Input: information required to complete the step
- Transformation: the changes that create the output
- Output: the result of the transformation

Data Collection Plan Outline

- What questions need to be answered?
- What data is available?
- How much data is needed?
- How to measure data?
- Who is performing data collection?
- Where is data collected from?
- Is a sample enough?
- What is the display format?

Sample Size

Not always is feasible to get results for/from all the instances. In this case a random sample should be taken to represent the population as a whole.

Sample size is important:

- Too small: outliers and/or anomalies distort the results
- Too big: complexity, increase cost, time consuming compared to accuracy gain

Sample size calculator

Data Quality

Criteria for data quality:

- Accuracy: data needs to be accurate
- Relevancy: should be appropriate for intended use
- Completeness: should not have missing values or records
- Timeliness: should be up to date
- Consistency: should be in the expected format
- Compliance: should comply with legal obligations