# EE3801 Cheatsheet

Intro to Data Engineering

github.com/securespider

## 01. Intro

### Data science vs engineering

- **Science** - Learn, optimise, analytics, aggregate and labelling
- **Engineering** - Cleaning, data storage, logging, sensors, pipelines

### Data structure

#### Unstructured data

- Chaotic no order to data

#### Structured data

- Data stored access in the same format

#### Semi structured data

- Can contain both forms of data
- Some structure but not all data points follow same format

### Big data

**Volume, Variety, Variability**

**Velocity** High rate of data generation

- Must create a robust and scalable pipeline

### Raw Data

- Tend to have gaps

### Data wrangling

Used to understand raw data

**Discovery** Understand what is in your data

**Structure**

**Cleaning** Dealing with gaps (nulls), outliers, formatting bugs

**Enrichment** Derive other data from other information/ additional data augmentation (feature selection)

**Validation** Verify data quality, sources

**Publishing** Give data scientist

### Process

**Extraction** Retrieve raw data from unstructured pool and migrate to temp repo

**Transformation** Structure enrich and convert raw data

**Loading** Loading structured data into data warehouse