COMP 333 — Week 4 Structured and Unstructured Data

Overview

In Week 3 and Week 4 (this lecture) we will cover Descriptive Data Analysis which is a basic technique of Data Analytics.

It is very important to understand your data.

Here we will cover

- Structured Data
- Unstructured Data
- The Five V's of Big Data

Structured Data

Structured data is the data commonly found in business databases.

The data in relational databases, csv files, and spreadsheets is typically

structured

as tables organized into rows and columns with certan properties

- each entry is an atomic (indivisible) value such as a number, id, char, or string
- each column has one type of data
- the organization of the table is described by a data schema
- each entity is specified by an identifier, such as a primary key or a foreign key

This structure makes it easier to manipulate, query, and analyse the data.

This course mainly focuses on structured data.

Unstructured Data

Everything that is not structured is *unstructured*.

The bulk of the unstructured data consists of text communications, images, and videos.

To this, social media adds a layer of network connectivity data
using friends, likes, etc
and mobile devices add geospatial and temporal data though GPS and timestamps.

Network data can be extremely useful in adding value to an organisation
depending on the domain and the problem.

This is a rich source of data that is the focus of much data analysis.

Unstructured data complicates data analysis, especially data wrangling, through

entity recognition to determine the identitifer for an entity's text reference

and *schema matching* to align features across different data sources.

Unstructured data is an important topic that is further covered in COMP 479.

Typical human-generated unstructured data includes:

- Text files: Word processing, spreadsheets, presentations, email, logs.
- Email: Text, but some internal structure thanks to its metadata.
- Social Media: Data from Facebook, Twitter, LinkedIn.
- Website: YouTube, Instagram, photo sharing sites.
- Mobile data: Text messages, locations.
- Communications: Chat, IM, phone recordings, collaboration software.
- Media: MP3, digital photos, audio and video files.
- Business applications: MS Office documents, productivity applications.

Typical machine-generated unstructured data includes:

- Satellite imagery: Weather data, land forms, military movements.
- Scientific data: Oil and gas exploration, space exploration, seismic imagery, atmospheric data.
- Digital surveillance: Surveillance photos and video.
- Sensor data: Traffic, weather, oceanographic sensors.

What Data is Gathered?

- GPS in your watch and car
  - Your location at all times
- Your web browser
  - Which sites you visit
  - How long you stay
  - What search queries you ask
  - When, from where, how often
- Amazon and Shops
  - What you look at
  - What you buy
  - Are you influenced by recommendations
- Email
  - What you say
  - Who you know
- Social Media
  - What you say
  - Who you know
  - What you like
- Your Fitness Tracker
  - Your health
  - Your exercise regime
- Credit Card and Bank
  - What you buy
  - When, from where, how often

Big Data — The V's

Big Data is an issue because of the **V** properties

- **Volume**: amount of data
- **Variety**: different types of data
- **Velocity**: rate at which data is generated
- **Veracity**: trustworthiness, level of noise
- **Value**: usefulness of data to a business

that create data management, computation, and interpretation problems.

The dirvers behind the generation of Big Data are

- **Transactions** with companies for commerce
- **Mobile** generating communication via text
- **Social Media** full of people interacting
- **Internet of Things** full of devices generating data

MGI Report McKinsey Global Institute, *Big data: The next frontier for innovation, competition, and productivity*, May 2011.