Nature of Data

Dr Sepi Chakaveh

31 January 2017
Most things start with a question...
Most people do not just use data for fun.

They have a story to tell or a problem to solve.
The purpose of this course is to start with a question and then explore a dataset with this question in mind.

Or we roam around and explore whether there is something interesting hidden in the data.
Two Ways Approach

**Top Down** - understand what kinds of problems the business has, and then look for specific solutions. (i.e. - pick a problem the business has - are the marketing pounds working as best as they can - and establishing which tools or techniques you can use to answer that question)

**Bottom Up** - understand what kinds of problems can be solved with Open Data first, and then look for specific applications in the business domain. (i.e. - before even looking at the business, learn about the techniques first)
Two Way Process
Relation Between Big Data, Open Data, Open Government

- **BIG DATA**
  1. Non-public data for marketing, business analysis, national security
  2. Large datasets from scientific research, social media, or other non-govt. sources
  3. Large public government datasets (e.g., weather, GPS, Census, SEC, health care)
  5. Business reporting (e.g., ESG data); other business data (e.g., consumer complaints)

- **OPEN DATA**
  4. Public data from state, local, federal govt. (e.g., budget data)

- **OPEN GOVT**
  2. Citizen engagement programs not based on data (e.g., petition websites)
What is an Open Data Process?

Open Government Platform (OGPL)

Data Management
- Data Catalogs
- Workflow
- Standard Metadata
- Validation
- Analytics

Data Portal
- Single Point Access
- Search & Discovery
- Open Formats
- Citizen Engagement
  Rate
  Social Media Connect

Feedback, Discussions, Ideas
Agriculture  Education  Communities
Energy
Health
But in order to consider any of these processes you should learn about

Data
What is Data?
You have 5 minutes to answer!

What is Data?
What is Data

• **Data** is a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things
Terminologies

1. We begin with an introduction to some of the basics of data

   - Open Data, Big Data
   - Qualitative, Quantitative,
   - Machine-readable, Discrete and Continuous data,
   - Static data & Dynamic data

2. For Data Wrangling & Manipulation

   - Data cleaning,
   - Data Visualisation,
   - Data Publishing,
   - static data & dynamic data
• **Data** is the fundamental building block of information, usually formatted in a special way.

• **Data** is information that has been translated into a form that is more convenient to move or process.

• Normally we use computer programs which are collections of instructions for manipulating *data*. 
Qualitative Data & Quantitative Data

- **Qualitative Data** is descriptive information

- **Quantitative Data**, is numerical information (numbers)
  
  - **Discrete Data** can only take certain values (like whole numbers) – **COUNTED**!
  
  - **Continuous Data** can take any value (within a range) - **MEASURED**!
DATA

Data

Qualitative

"It was great fun"

Quantitative

Discrete

Continuous

5

3.265...
Example: What do we know about Dexter the Dog

Qualitative:
• He is Golden brown
• He has long hair
• He has lots of energy

Quantitative:
• Discrete:
  • He has 4 legs
  • He has 2 brothers
• Continuous:
  • He weighs 25.5 kg
  • He is 565 mm tall
Data-Information-Knowledge-Wisdom
Link Data
Link Data

• Linked Data is about using the web to connect related data that wasn't previously linked.

• Or using the Web to lower the barriers to linking data currently linked using other methods.

• More specifically Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF."
Linked Data
Data for Computers
Data for Computers

• Computers are inherently different from humans.

• It can be exceptionally hard to make computers extract information from certain sources.

• Some tasks that humans find easy are still difficult to automate with computers.

• If you want your computer to process and analyse your data, it has to be able to read and process the data.

• This means it needs to be structured and in a machine-readable form
Data Usage

• **Data** is used at
  – **Computing (Private, Public)**
  – **Telecommunication (mostly private)**
  – **Publishing (e.g. Amazon)**
  – **Entertainment Industry (TV, Movie, Music, Games)**
  – **Business Sector**

  – Thus Data generated in all fields and systems in modern day life is **vast**.

• Hence we need to consider how we may **access, manipulate, analyse, consume** the continuously generated **data**.
Data Type for Computers

• **Data** is used as multiple forms of information, information content and information processing

• **Data** constitutes the use of (but not limited to) digital content to store and experience content.

• **Data** indicates that information or data being transferred over a network may be composed of one or more media types, including:
  
  • *Text*
  • *Graphics (2D, 3D)*
  • *Still images*
  • *Audio: files or real-time speech*
  • *Video: files or real-time broadcasting*
Data Types for Computers
Data Set Forms & Digital Object
• They are two data set forms or digital objects

• Static Data Sets
  – Records
  – Numbers
  – Values

• Dynamic Data Sets
  – Video
  – Audio
Data Set & Digital (Data) Object

- **Data set** is a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.

- **Data sets** normally deal with **Static Data**

- **Digital (Data) Object** is a Data set that the data are editable

- **Digital Objects** normally deal with **Dynamic Objects**
Digital Objects

- **Digital Objects** are composed of **Discreet Data Sets**
- Each object has a number of **attributes** associated with it, including shape, size in terms of border coordinates, colour of border, shadow and so on.
- **Editing** of a computer graph simply means *changing certain attributes*.
- To ensure that every computer **interprets** the commands correctly, **standard forms of representation or (FORMATS)** are used such as **GIF** (graphical interchange format) and **TIFF** (tagged image file format).
Data – Object Relationship

Data Set/Digital Objects
Metadata
Metadata

- **Metadata** (meta data, or sometimes *meta information*) is "data about other data", of any sort in any media.

- *Metadata* is information about data.

- *Metadata* is information about information.

- Metadata contains information about that data or other data.

- An item of metadata may describe
  - an individual datum,
  - content item,
  - collection of data including multiple content items and hierarchical levels, for example a database schema.
Metadata

- Metadata would document data about data elements or attributes,
- Name, Size, Data type,
- Data about records or data structures
- (length, fields, columns, etc) and
- Data about data (where it is located, how it is associated, ownership, etc.).
- Metadata may include descriptive information about the context, quality and condition, or characteristics of the data.
Metadata

• In data processing, metadata is definitional data that provides information about or documentation of other data managed within an application or environment.

• Metadata would document data about data elements or attributes, (name, size, data type, etc)

• Metadata would also document data about records or data structures (length, fields, columns, etc)

• Data about data (where it is located, how it is associated, ownership, etc.).

• Metadata may include descriptive information about the context, quality and condition, or characteristics of the data.
Data Granularity

- The **Data Granularity** of data refers to the fineness with which data fields are sub-divided.

- **Higher** granularity has **overheads** for data input and storage.

- Offers benefits in **flexibility** of data processing.
Granularity in Metadata

For example, a postal address can be recorded, with *low granularity*, as one field:

```
address = 200 2nd Ave. South #358, St. Petersburg, FL 33701-4313 USA
```

With *high granularity*, as many fields:

```
street address = 200 2nd Ave. South #358
city = St. Petersburg
postal code = FL 33701-4313
country = USA
or even higher granularity:
street number = 200
street = 2nd Ave. South #358
city = St. Petersburg
postal code state = FL
postal-code-first-part = 33701
postal-code-second-part = 4313
country = USA
```
Digital Asset
Digital Asset

• A **Digital Asset (DA)** is any form of **multimedia content** that have been formatted into a binary source which include the **right to use it**.

• A digital file without the right to use is not an asset.

• Digital assets are categorised in three major groups:
  - Textual content
  - Images (media assets)
  - Multimedia (media assets)
  - Art asset (video game & visual effects)
Binary Large Objects
Binary Large Object (BLOB)

- A Binary Large Object, also known as a BLOB, is a collection of binary data stored as a single entity in a database management system.

- Blobs are typically Media objects

- Sometimes binary executable code is stored as a blob.

- Database support for blobs is not universal.

- In open source culture, binary BLOB is a term for an object file loaded into the kernel of an open source operating system without publicly available source code.
Graph Theory
In mathematics and computer science, **graph theory** is the study of **graphs**.

In here mathematical structures used to model **pairwise relations** between objects from a certain collection or data-set.

A "**graph**" in this context refers to a collection of vertices or '**nodes**' and a collection of **edges** that connect pairs of vertices.

There are different ways to store graphs in a computer system.

The **data structure** used depends on both the **graph structure** and the **algorithm** used for **manipulating** the graph.
Tree Graph
Tree Graph

• **A tree** is a *graph* in which any two *vertices* are connected by *exactly one simple path*.

• In other words, any *connected* graph *without cycles* is a tree.

• A *forest* is a disjoint *union of trees*.

• The various kinds of trees used as data structures in computer science are not really trees in this sense, but rather, types of *ordered directed trees*.

• A **tree** is an undirected simple graph $G$ that satisfies any of the following equivalent conditions.
An example of a tree graph
Directed Acyclic Graph (DAG)
Directed Acyclic Graph (DAG)

- In mathematics, a **Directed Acyclic Graph (DAG)**, is a directed graph with **no directed cycles**.

- **DAG** is formed by a collection of vertices and directed edges.

- Each **edge** connecting one **vertex** to another.

- Mathematically **directed graph** or **digraph** is a pair which:
  - \( G = (V, A) \)
  - A set \( V \), whose **elements** are called **vertices** or **nodes**.
  - A set \( A \) of **ordered pairs**, i.e. a collection of objects having two coordinate of vertices, called **Arcs, directed edges**, or **Arrows**, (and sometimes simply **edges**).

- There is **no** way to start at some vertex \( V \) and follow a
An example of a directed acyclic graph
Scene Graph
Scene Graph

- A **scene graph** is a general **data structure** commonly used by **vector-based graphics** editing applications.

- The **scene graph** is a structure that arranges the **logical** and often (but not necessarily) **spatial** representation of a graphical scene.

- A **scene graph** is a collection of **nodes** in a graph or **tree** structure.

- A **node** may have many children but often only a **single parent**.

- Examples of such programs include AutoCAD, Adobe Illustrator, 3D Studio Max, Do, OpenSceneGraph, and others.
Parent & Children Relationship

P

C2

D1

D2

C1

E1
Scene Graph

- Root (Bicycle)
  - Frame
  - Translate (right)
    - Wheel
  - Translate (left)
    - Wheel
Scene Graph & Visualisations
Scene Graph

- Star
  - Rotation
    - Wobble!
      - Planet 1
        - Rotation
          - Moon A
          - Moon B
    - Rotation
      - Planet 2
        - Rotation
          - Moon C
          - Moon D
RDF
RDF

• The **Resource Description Framework (RDF)** is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model.

• It has come to be used as a general method for conceptual description or modelling of information that is implemented in web resources, using a variety of syntax notations and data serialization formats.

• It is also used in knowledge management applications.
RDF Ranking

• RDF Rank is an algorithm that identifies the more important or more popular entities in the repository by examining their interconnectedness.

• The popularity of entities can then be used to order the query results in a similar way to the internet search engines, the way Google orders search results using PageRank.

• The RDF Rank component computes a numerical weighting for all nodes in the entire RDF graph stored in the repository, including URIs, blank nodes and literals. The weights are floating point numbers with values between 0 and 1 that can be interpreted as a measure of a node’s relevance/popularity.