COMP 333 — Week 4
MetaData, Self-Descriptive Data, Data Formats

Overview

We will look at

- Metadata, which is data describing data
- Self-descriptive data, which is data that is self-describing
- Data Formats common to datasets

Metadata

Metadata is data that provides information about other data

For example

- Means of creation of the data
- Purpose of the data
- Time and date of creation
- Creator or author of the data
- Location on a computer network where the data was created
- Standards used
- File size
- Data quality
- Source of the data
- Process used to create the data

The *provenance* of data is the origin and/or history of the data

Self-Descriptive Data

We have all struggled to understand a csv file of numbers
where there is no documentation.

Just what do the columns represent?
What units are meant to be assigned the a column headed *lth*?

And if someone gives you a binary file rather than a text file ...

The concept of *self-descriptive data*
means that it is human readable
and the human can make sense of it
as a stand-alone file.

Many self-descriptive data formats combine the metadata with the data.

Adding a header line to a csv file is a start towards self-descriptive data
but the best examples of self-descriptive data are

- ARFF
- HDF5
- XML

XML does a thorough job of specifying character sets and the meaning of tags.

Data Formats

There are several common, and not so common, formats used for data:

- Comma-separated values (csv)
- Tab-separated values (tsv)
- JSON
- Attribute-Relation File Format (ARFF)
- XML
- RDF triples
- Binary files (BLOBs)
- HDF5 (Hierarchical Data Format version 5)

Data Formats — ARFF — Weka

The Attribute-Relation File Format is the format used by Weka.

Weka is a popular open source Java package for Machine Learning.

The data examples for Weka are given as ARFF files.

ARFF is a self-descriptive data format.

An ARFF file is an ASCII file consisting of a header followed by data.

The Header consists of:

- the name of the relation,
- a list of the attributes (columns in data),
- their types

Here is an example:

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

The ARFF data looks like

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

HDF5

The Hierarchical Data Format is designed by `https://www.hdfgroup.org/solutions/hdf5/` to be

- self describing by containing metadata
- heterogeneous, allowing many types of data files
- hierarchical, essentially a directory structure

HDF5 is designed for large datasets from supercomputer problem domains.

> *"s an open source file format that supports large, complex, heterogeneous data. HDF5 uses a "file directory" like structure that allows you to organize data within the file in many different structured ways, as you might do with files on your computer. The HDF5 format also allows for embedding of metadata making it self-describing."*

We will not meet such datasets in this course.