MODULE - II

Handling Data

Gather Data- Topical Data- Data Scrapping- Formatting Data, tools- Formatting with code- Scientific Design Choices using Graphical Display Options - Adding Model with Overlaying (Statistical) Information - Higher-dimensional Displays and Special Structures with parallel coordinates- Scatter Plot Matrices and Mosaic Plot-Time Series and Maps.

Handling Data

Data Handling

Handling the data in such a way that it becomes easier for people to understand and comprehend the given information.

Hence,

The process of

Collecting⇨Recording⇨representing data

Is called Data handling.

Data handling also includes a broad range of activities like

Collection of raw data

- Ensuring its accuracy and integrity
- Processing it into a manageable form
- Analysing it statistically
- Presenting it in ways that are easy to understand

Gather Data

Data is the core of any visualization

There can be a lot of sources to find the data

Getting it from experts

Online applications or

Gathering it yourself

Different Data Collection Methods

Data Collection methods can be primarily categorized as

Primary methods &

secondary methods

Primary Data Collection:

Primary data collection involves the collection of original data directly from the source or through direct interaction

This method allows researchers to obtain firsthand information

various techniques for primary data collection

- Surveys and Questionnaires:
- Interviews:
- Observations:
- Experiments:
- Focus Groups:

Secondary Data Collection:

Secondary data collection involves using existing data collected by someone else for a purpose different from the current intent

Secondary data can be obtained from various sources

- Published Sources
- Online Databases
- Government and Institutional Records
- Publicly Available Data
- Past Research Studies

Managing the sourced data

When Data is Provided by Others

This route is common, someone else did all the data gathering work.

But need to be careful as lot of mistakes can happen in outsourced data.

The most common mistakes are

- Missing zeros
- Typo errors
- Missing digit
- Errors during conversion

Need to identify the context of the data

- Need to know where the original data came from
- How it was collected and
- What was the purpose

Example:

Poll survey data of 2000 could have almost no relevance to polls in 2024

Other aspects to mind while validating data

- Erroneous conclusions
- Data that might compromise public policy
- Incapacity to correctly respond to research inquiries
- Data that might Bringing harm to participants who are humans or animals
- Wilfully Deceiving other researchers
- The study's inability to be replicated and validated

Finding Sources

Here’s where you can start your search

Search Engines

How to find anything online, yes

- Google &
- Wolfram|Alpha, the computational search engine
- Among other online sources

Direct from the Source

If a direct query for “data” doesn’t provide anything

- Try searching for academics who specialize in the area/subject
- Sometimes data is on their personal sites
- Details/links can be found on their papers/ articles
- Can also try emailing relevant experts
- Spot sources in graphics published by news outlets
- Data sources are included in small print somewhere on the graphic

Universities

Most University libraries have boosted up their technology resources and may have some expansive data archives

Many of the statistics departments also maintain lots of data files

Suggested resources:

- Data and Story Library (DASL) (http://lib.stat.cmu.edu/DASL/)

  An online library of data files and stories that illustrate the use of basic statistics methods, from Carnegie Mellon
- Berkeley Data Lab (http://sunsite3.berkeley.edu/wikis/datalab/)

  Part of the University of California, Berkeley library system
- UCLA Statistics Data Sets (www.stat.ucla.edu/data/)

  Some of the data that the University of California, Los Angeles Department of Statistics uses in their labs and assignments

General Data Applications

There are many general data-supplying applications are available

Some applications provide large data files for download for free or for a fee

Some applications provide data accessible via Application Programming Interface (API) allowing use of data as a service

Few suggested resources

- Freebase (www.freebase.com)
  A community effort that mostly provides data on people, places, and things
- Infochimps (http://infochimps.org)
  A data marketplace with free and for-sale datasets
  (mostly government) data
- Aggdata (http://aggdata.com)
  Another repository of for-sale datasets
  Mostly focused on comprehensive lists of retail locations
- Amazon Public Data Sets (http://aws.amazon.com/publicdatasets)
  Host some large scientific datasets
- Wikipedia (http://wikipedia.org)
  A lot of smaller datasets in the form of HTML table

Topical Data

Apart from general data suppliers, there are also subject-specific sites offering loads of free data
Following are few site's available for the topic of choice

Geography
Plenty of geographic file types are available at.

- TIGER (www.census.gov/geo/www/tiger/)
  From the Census Bureau, probably the most extensive detailed data about roads, railroads, rivers, and ZIP codes
- Openstreetmap (www.openstreetmap.org/)
  One of the best examples of data and community effort
- Geocommons (www.geocommons.com/)
  Both data and a mapmaker
- Flickr Shapefiles (www.flickr.com/services/api/)
  Geographic boundaries as defined by Flickr users

Sports
Sports data can be found on Sports Illustrated or team organizations' sites, and can also be found on sites dedicated to sports data.

- Basketball Reference (www.basketball-reference.com/)
  Provides data as specific as play-by-play for NBA games.
- Baseball Data Bank (http://baseball- databank.org/)
  Super basic site where you can download full datasets.
- Database Football (www.databasefootball.com/)
  Browse data for NFL games by team, player, and season.

World

Several noteworthy international organizations keep data about the world, mainly health and development indicators

- Global Health Facts (www.globalhealthfacts.org/)
  Health-related data about countries in the world.
- Undata (http://data.un.org/)
  Aggregator of world data from a variety of sources
- World Health Organization (www.who.int/research/en/)
  Again, a variety of health-related datasets such as mortality and life expectancy
- OECD Statistics (http://stats.oecd.org/)
  Major source for economic indicators
- World Bank (http://data.worldbank.org/)
  Data for hundreds of indicators and developer-friendly

Government and Politics

There has been a fresh emphasis on data and transparency in recent years, so many government organizations supply data

Data is provided by many nongovernmental sites that aim to make politicians more accountable.

- Census Bureau (www.census.gov/)
  Find extensive demographics here.
- Data.gov (http://data.gov/)
  Catalog for data supplied by government organizations.
- Data.gov.uk (http://data.gov.uk/)
  The Data.gov equivalent for the United Kingdom.
- Datasf (http://datasf.org/)
  Data specific to San Francisco.
- NYC datamine (http://nyc.gov/data/)
  Data specific to New York.
- Follow the Money (www.followthemoney.org/)
  Big set of tools and datasets to investigate money in state politics.
- Opensecrets (www.opensecrets.org/)
  Also provides details on government spending and lobbying.

Data Scrapping

Often the exact needed data is available but the data is spread across multiple HTML pages on multiple websites

If it concerns only a few pages
The **straightforward solution** can be employed
Visit every page and manually find the data and enter the data in a spreadsheet

If data is spread across thousands of pages
Straight forward solution of visiting all sites would be tedious and practically impossible
Such cases require data scraping method
Data scraping: writing code to visit all pages automatically fetch the data and store the same in database or a text file

Although coding is the most flexible way to scrape the data you need, you can also try tools such as Needlebase and Able2Extract PDF converter

Formatting Data

Different visualization tools use different data formats
Different structures require different data formats
Structure and tools required varies by the story you want to tell.
So

More flexible with structure and format of the data
=> **more possibilities/ options available**

Getting data in format and fit for specific needs

- Formatting applications
- Little bit of programming know-how
- Hire a programmer to format and parse the data

Data Formats

Most popular method for handling data is Excel.
Excel is good if everything
From
Processing => analyses => & visualization
Is done in the same Excel
If we have to rely on different applications for different stages, we would want data formats that are machine-readable and inter portable.
Other data formats

- Delimited text
- JavaScript Object Notation
- Extensible Markup Language.

Delimited Text

Delimited text most common and most familiar format

For a dataset in the context of rows and columns

A delimited text file splits columns by placing a delimiter between each column entry of a row.

Using a comma for delimiting the file is now a comma-delimited file.

The delimiter can be

- A tab
- Spaces
- Semicolons
- Colons
- Slashes etc...

Comma and tab are the most common.

Delimited text is widely used because

- It can be read into most spreadsheet programs such as Excel or Google Documents
- Can be exported from spreadsheets
- Good for sharing data with others
- Not depend on any particular program/ application.

JavaScript Object Notation (JSON)

Common format offered by web APIs

Designed to be both machine- and human-readable

Based on JavaScript notation

But not dependent on the language.

JSON works with keywords and values,

Treats items like an objects.

Comparing JSON data as regular database=> each object is a row

A number of applications, languages, and libraries accept JSON as input

JASON is verry useful while designing data graphics for the web

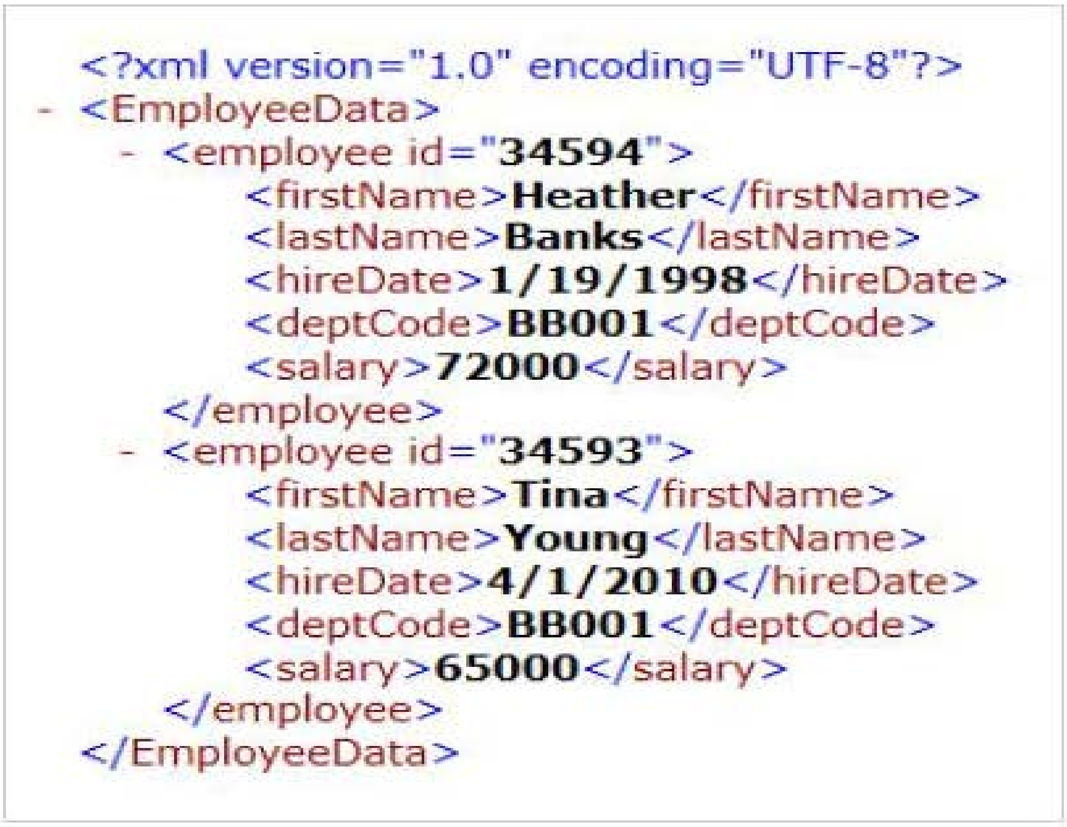
Extensible Markup Language (XML)

XML is another popular format on the web

Often used to transfer data via APIs

It is a text document with values enclosed by tags

Example, an XML file



```
  <?xml version="1.0" encoding="UTF-8"?>
- <EmployeeData>
  - <employee id="34594">
        <firstName>Heather</firstName>
        <lastName>Banks</lastName>
        <hireDate>1/19/1998</hireDate>
        <deptCode>BB001</deptCode>
        <salary>72000</salary>
    </employee>
  - <employee id="34593">
        <firstName>Tina</firstName>
        <lastName>Young</lastName>
        <hireDate>4/1/2010</hireDate>
        <deptCode>BB001</deptCode>
        <salary>65000</salary>
    </employee>
  </EmployeeData>
```

Formatting Tools

Previously, converting data from one format to another definitely require lots of coding/ programming

Requiring lots of time and efforts.

Now with growing volumes of data

There are verry reliable tools developed to handle data conversions

Most commonly used tools

- Google Refine
- Mr. Data Converter
- Mr. People
- Spreadsheet Software

Google Refine

Google Refine is the evolution of Freebase Gridworks.

Google Refine (Gridworks 2.0) has an easier-to-use interface and with more features.

It runs from the browser but can work on the system data

Refine is also open source, => if required it can be downloaded and or used in any other applications

Refine, provides

- A familiar spreadsheet interface with rows and columns.
- Allows
  - Easy sort by field
  - Search for values
  - Identify inconsistencies in data
  - Consolidating data.

Mr. Data Converter

When there is a need to convert Data to another format

Excel/ Spread Sheets can be useful when exporting as CSV(Delimited Text) but for any other formats there is a need for specific data conversion application -Data Converter

Mr. Data Converter

- A simple and free tool
- Simple steps to convert the data
  - Copy and paste data from excel in the input section
  - Then select what output format needed in the bottom half
  - Can Choose from variants of XML, JSON, and a number of other formats.
- The source code to mr. Data converter is freely available hence, allowing personal modifications if needed

Mr. People

Mr. People is specifically for parsing names

Mr. People enable

- Copying and paste data into a text field
- The tool parses and extracts the names from text

Identifies

- The first name last names, and middle names
- Initial
- Prefix, and
- Suffix

O/p is provided in Table format

Mr. People is also available as open-source and open for

- Inclusion into other applications
- custom modifications

Spreadsheet Software

Spread sheet software's (Excel/Google sheets)
Are verry fast and efficient on
Small/ simple tasks
Very limited dataset
For
Managing large datasets
Performing complex tasks
Should prefer
Specific software's or
Custom coding

Formatting with Code

Working with readymade applications for managing data might be useful and simple, it might have their own limitations,

- Might not handle large datasets
- Might overload and crash
- Might not do the precise manipulation needed

in some cases, they might be of no help

in such cases it would be wise to start writing code to perform custom/tailored operations on data
example on how to change formats of data
XML to CSV
XML data file

```
<weather_data>
<observation>
<date>20090101</date>
<max_temperature>26</max_temperature>
</observation>
<observation>
<date>20090102</date>
<max_temperature>34</max_temperature>
</observation>
<observation>
<date>20090103</date>
<max_temperature>27</max_temperature>
</observation>
<observation>
<date>20090104</date>
<max_temperature>34</max_temperature>
</observation>
...
</weather_data>
```

Each day’s temperature is enclosed in <observation> tags with a <date> and the <max_temperature> tags enclosing date and temperature within observation.

Python Code to convert XNL to csv

```
import csv
reader = csv.reader(open(‘wunder- data.txt’, ‘r’), delimiter=",")
print "{ observations: [" rows_so_far = 0
for row in reader:
rows_so_far += 1 print ‘{‘
print ‘"date": ‘ + ‘"‘ + row[0] + ‘", ‘ print ‘"temperature": ‘
+ row[1]
```

output in CSV format:

```
20090101,26
20090102,34
20090103,27
20090104,34
```

Module II b
Scientific Design Choices using Graphical Display Options
Adding Model with Overlaying (Statistical) Information
Higher-dimensional Displays and Special Structures with parallel coordinates
Scatter Plot Matrices and Mosaic Plot
Time Series and Maps.