**Lecture-4**

1. Ways to find data:
   1. Datasets are available for direct download from different sources like IMDB, Movielens, DOE, last fm, medical databases and much more.
   2. Data can be derived from WWW using a technique called web scraping.
   3. API as provided by Twitter, Facebook, Fourquare etc

Note: JSON is one format for data exchange used on web (Java script object notation). JSON is an [open standard](http://en.wikipedia.org/wiki/Open_standard) format that uses [human-readable](http://en.wikipedia.org/wiki/Human-readable) text to transmit data objects consisting of [attribute–value pairs](http://en.wikipedia.org/wiki/Attribute%E2%80%93value_pair)

1. Data munging: processing and generating data.

Note: HTTP is a kind of hand shaking protocol which helps in communication between server and client to have data exchange between them. This can be in form of GET and POST requests. GET request are essentially queries for requesting data. I.e. Google query: <https://www.google.co.in/#q=manoj+gulati>

GET request has limited number of characters in GET request (1000 max). Even security is an issue as GET requests are not encrypted.

POST request are encrypted. It does not send the query in URL rather it sends query in HTTP header.

Note: XML is another type of mark-up language. It is base behind semantic web. It allows descriptors with in the language.

HTML: It is similar to XML but it is simplest of all mark-up languages. In web scraping you are actually extracting data from HTML.

CSS: In html or any markup language you make classes and divisions in order to facilitate dedicated styling schemes using CSS. Cascading style sheets are essentially used for styling purposes.

DOM: Document object model which defines the hierarchy of a webpage. Tree like structure of a webpage is called DOM. Elements of HTML webpage forms the elements of a DOM.

1. Web Scraping: Is done while following rules
   1. Data is public
   2. Non sensitive
   3. Anonymized
   4. Citations should be added
   5. You should check with robots.txt which specifies what you are allowed to scrap
2. “Import requests” is a python library for scraping data from webpages.

Another one on same line is “pattern” and “BeautifulSoup”