# Data Science - Project Stage 2

## Web Data Extraction

**Team Members:**

- Karan Dharni (dharni@wisc.edu)
- Sudarshan Maru (maru@wisc.edu)
- Amogh Joshi (asjoshi4@wisc.edu)

---

### Description of Data Sources-

The two sources of data we selected for our data extraction task are:

- amazon.com
- booksamillion.com

We have extracted data about books from the following three categories: Space Opera, Dark Fantasy and Psychological Thrillers. Moreover, we have extracted data about recently released and paperback books only.

### Description of extracted Entity-

The entity type we have extracted from the above mentioned sources is : **Books**. We have built two tables (CSV format), one for each source. There are 2 CSV files are named books_amazon_output.csv and books_millions_output.csv.

The schema followed by both the tables is:
*{Name, Category, Author, Price, Series, Pages, Publisher, Date, Language, ISBN-10, ISBN-13, Dimensions, Weight}*

The number of tuples in the table extracted from **amazon.com :- 3000**
The number of tuples in the table extracted from **booksamillion.com :- 9000**

Here's a brief description of the attributes extracted :-
Name: Title of the book
Category: Category of the book (Ex: In these tables, we have extracted data for books limited to categories Space Opera, Dark Fantasy and Psychological Thrillers)
Author: Author(s) of the book
Price: Cost of the book
Series: Series of the book
Pages: Number of pages in the book

Publisher: Publication house of the book
Date: Date on which this book was (or will be) released
Language: Language in which this book is written
ISBN-10 and ISBN-13: The 2 ISBNs of a book
Dimensions: Size of the book
Weight: Shipping weight of the book

Some attributes are sparsely populated. For example- In the file books_millions_output.csv, Dimensions and Weight are sparsely populated, as compared to books_amazon_output.csv.

## Description of methodology-

The process of extracting the data from both the sources was identical. In the first step, after applying appropriate filters on the data-source websites, hundreds of pages are displayed; and each page contains entries for multiple books. We grab the embedded URLs of the books on each page, store them in a file and then move to the next page. Thus, at the end of this step, we have two files, which contain more than 9000 (booksamillion) and 3000 (amazon) URLs of books.

In the next step, we read the URLs from the files created in previous step, and crawl each of these URLs individually. We extract the required attributes such as Name, Author, Price, Series, Pages, Publisher, Date, Language, ISBN-10, ISBN-13, Dimensions and Weight for each book. A tuple consisting of these attributes is then stored in a CSV file.

So finally, after successful completion of both these steps, we obtain two CSV format files, one for each data source; containing the extracted data in tabular form. Each tuple of a table describes the data about one particular book.

## Description of Tools Used-

We used the tools '*Beautiful Soup*' and '*Selenium*' tools of the python framework in our project. Selenium is used to get the HTML page source; whereas Beautiful Soup is used to extract the attributes from this source.