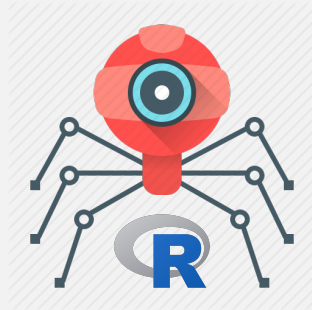


CS636 Data Analytics with R Program  
Prof. Zhi Wei



## CS636 Data Analytics with R Program

### Project 1

By  
Group 6

Duyen Ngyen | Bicheng Xiao | Shubham Gulia

Scope



### Scope of the project

Use R program to extract required information  
from specified website

Main page:

<https://bmcmmedgenet.biomedcentral.com>

Total articles: 1642

Required information:

10 fields:

DOI, Title, Authors, Author Affiliations,  
Corresponding Author, Corresponding Author's  
Email, Publication Date, Abstract, Keywords, Full  
Text (Textual format)

The screenshot shows the BMC Medical Genetics website. The header includes the BioMed Central logo and a search bar. The main navigation bar has links for HOME, ABOUT, ARTICLES, and SUBMISSION GUIDELINES. On the left, there are links for Sections, Supplements, and Receptor acknowledgements. The central area is titled 'Articles' and features a search bar with 'Search BMC Medical Genetics' and a 'Search' button. Below the search bar, there are options to 'Sort by' (Newest first) and 'Page 1 of 66'. A list of articles is displayed, with the first article titled 'The RS4939827 polymorphism in the SMAD7 GENE and its association with Mediterranean diet in colorectal carcinogenesis'. The article's abstract and authors are listed. On the right, there are links for 'Submit a manuscript', 'Editorial Board', 'Editor Profiles', and 'Sign up to article alerts'. A 'FOLLOW' button is also present.

## Contribution of group members

Bicheng Xiao	main program: (loadAllArticles.R) Functions(util.R): <ul style="list-style-type: none"> <li>• analysisArticle</li> <li>• extractAuthors</li> <li>• extractAffiliation</li> </ul>	Testing and debugging Project report(PDF file)
Duyen Ngyen	Function(util.R): <ul style="list-style-type: none"> <li>• loadArticleList</li> </ul>	Testing and debugging Readme.txt
Shubham Gulia	Function(util.R): <ul style="list-style-type: none"> <li>• extract</li> <li>• extracAttribute</li> </ul>	Testing and debugging ReadTheResult.R

## Challenges

unfamiliar with R packages, like XML	Check the online document and samples
extract author, corresponding author, corresponding author's email	<ul style="list-style-type: none"> <li>• Carefully check the HTML of the article page and find the xpath</li> <li>• Use a 2<sup>nd</sup> time extraction to get more detailed or related data</li> </ul>

CS636 Data Analytics with R Program  
Prof. Zhi Wei



**Thank you**