科目ナンバリング	U-LAS30 20033 SE11						
授業科目名<英訳>	Processing and Analyzing Data I-E2 : Shell-based data processing fundamentals Processing and analyzing data I-E2 :Shell-based data processing fundamentals			担当者所属職名・氏名	医学研究科　助教　VEALE, Richard Edmund		
群	情報学科目群	分野(分類)	(各論)		使用言語	英語	
旧群		単位数	2単位	週コマ数	1コマ	授業形態	演習（対面授業科目）
開講年度・開講期	2024・後期	曜時限	金3	配当学年	全回生	対象学生	全学向

[授業の概要・目的]

As the world and the sciences become increasingly computerized, it is increasingly important to understand how to search, process, and analyse large bodies of digital data. This course is designed for all students of all disciplines. The purpose is to learn the the basic concepts and methods for systematic processing of data encountered in any field.
Lectures will focus on learning basic command line tools for automatic processing of data, including sorting, filtering, summarizing, searching, and other related programming.

[到達目標]

At the end of the course, students should be able to operate a computer to automatically:
(1) search for specific entries in large collections of data
(2) search for pattern-like entries in large collections of data
(3) filter desired content from large collection of data
(4) perform basic summary and counting statistics on data
(5) assemble small processing pipelines from the various tools they will study

[授業計画と内容]

(1) What is a computer, what is an operating system?
Remove microsoft/apple preconceptions.
Using Command Line Interfaces (CLI) to interact with computers: Shell.
Logging in to a remote machine (SSH, public/private keys, etc.)

(2) Using remote and local machines.
Basic Networking: TCP, FTP/HTTP, IP.
Managing data: Disk management, file systems, file system structure (tree), file permissions.
Moving data between machines: SCP, RSYNC.
Installing software: package managers (RPM, APT).
Security: Super User (su, sudo), users, groups.
Diagnostic tools: PS, HTOP, DF, etc.

(3) Complex commands for string manipulation and search.
Moving data between programs: standard in/out/error streams, piping, redirecting.
String manipulation: Regular Expressions, wildcards, AWK, SED
Loops: for/while loops, loop conditions.
Finding information: Stack Overflow, MAN pages.

(4) Shell Scripts and programming languages.

Processing and Analyzing Data I-E2 :Shell-based data processing fundamentals(2)へ続く↓↓↓

What is a "program"? Libraries, functions, paths, environmental variables.
Programming languages: interpreted versus compiled, lazy versus strict evaluation, data types. Python, R, Perl, Fortran, C/C++, Java.

(5) Data Formats
Binary versus Textual (CSV etc.). HDF5 (computer independent representation).
Statistics: Summary statistics on data. Good/bad ways of thinking.

(6) Data representation/presentation
Simple plotting/graphing (matlab, matplotlib, R, ggplot, gnuplot).
Why excel is bad (limitations).
Formats: PDF, vector versus raster.

(7) Representation of large data sets.
(Relational) Databases, SQL, "queries", subsets.

(8) Keeping track of your work (Version Control).
Version Control: CVS, SVN, GIT, mercurial. Remote versus local repositories.
Backing up: Version Control is not back-up. Backing up practices (tape, disks, etc.).

(9) Data processing THEORY
Best practices: concepts to reproduce reusibility.
Basic parallelization (GNU parallel).

(10) "Big Data" processing.
Parallelizing: MapReduce, Hadoop, Spark, MPI.
Big filesystems: HDFS, lustre, NFS.
Clusters, Supercomputers.
Scheduling computer time and resources (scheduler): TORQUE

(11) Modeling, optimization, parameter search
Gradient descent methods, neural networks
Parameter estimation: markov chain monte-carlo, evolutionary algorithms.
Random seeds: pseudorandom issues on large machines

(12) Project

(13) Project

(14) Project (presentations)

(15) Feedback

[履修要件]

No prior knowledge of computer programming or data processing is necessary

Processing and Analyzing Data I-E2 :Shell-based data processing fundamentals(3)へ続く↓↓↓

Processing and Analyzing Data I-E2 :Shell-based data processing fundamentals(3)

[成績評価の方法・観点]

Class attendance and participation (10%), Quizzes (40%), Final Project/Report (50%)

[教科書]

No textbook used, lecture materials will be provided in class and online via PANDA.
Documentation about processing tools (e.g. manpages) will be introduced in class.

[参考書等]

（参考書）

Introduced during class

[授業外学修（予習・復習）等]

Students are strongly recommended to practice class materials and on their own data outside of class to deepen their understanding.

[その他（オフィスアワー等）]

A personal computer is strongly recommended and makes the course significantly more accessible. While Windows-based, macOS-based and GNU/Linux systems are all acceptable, the majority of the course will focus on UNIX-based tools.