THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Hands-on Exercise 1: Creating your own dataset

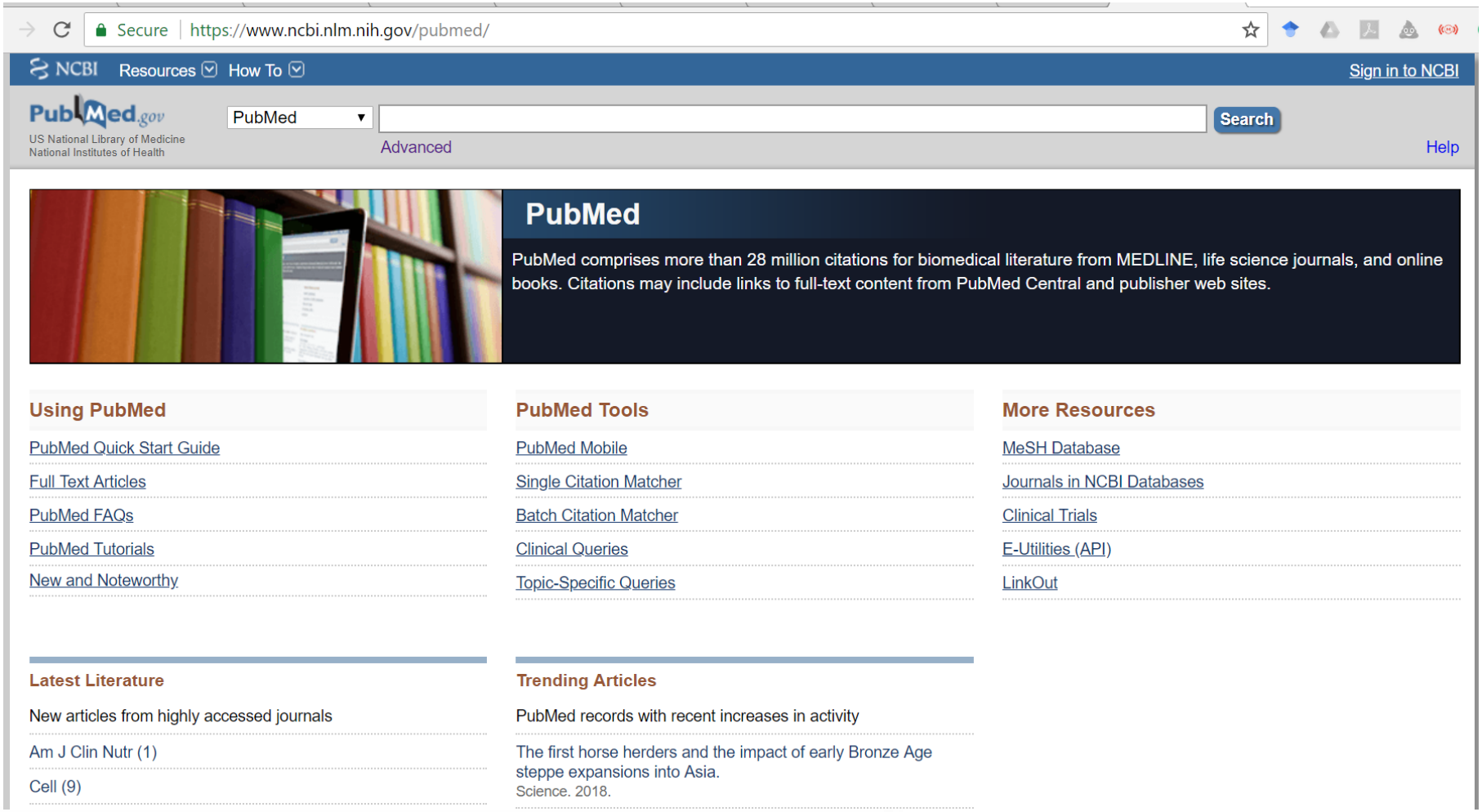
Heejun Kim

June 5, 2018

Objective of the Exercise

- Training/test data sets are already prepared for BRCA and Alzheimer's disease
- What if these topics do not interest you?
  - Learn how and where you can collect data
  - Experience with tools that help data pre-processing

PubMed



ENABLE

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

PubMed

- PubMed: a life science and biomedical literature search engine that is freely available by National Library of Medicine
  - Accessible to the MEDLINE database of references and abstracts
  - Manual access (https://www.ncbi.nlm.nih.gov/pubmed/) vs. programmatic access (https://www.ncbi.nlm.nih.gov/home/develop/api/)
  - Indexed by Medical Subject Heading (MeSH)

Medical Subject Heading (MeSH)

- “MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity.”
- “MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. ” (NLM)
- MeSH browser: https://meshb.nlm.nih.gov/search

Medical Subject Heading List: Breast Cancer

- Take the example of breast cancer in the context of MeSH. It has the following major terms:
    - Breast Neoplasms
    - Breast Cancer
    - Breast Carcinoma
    - Breast Tumors
    - Cancer of Breast
    - Malignant Neoplasm of Breast
    - Malignant Tumor of Breast
    - Mammary Neoplasm, Human

MeSH Definition: Breast Neoplasms

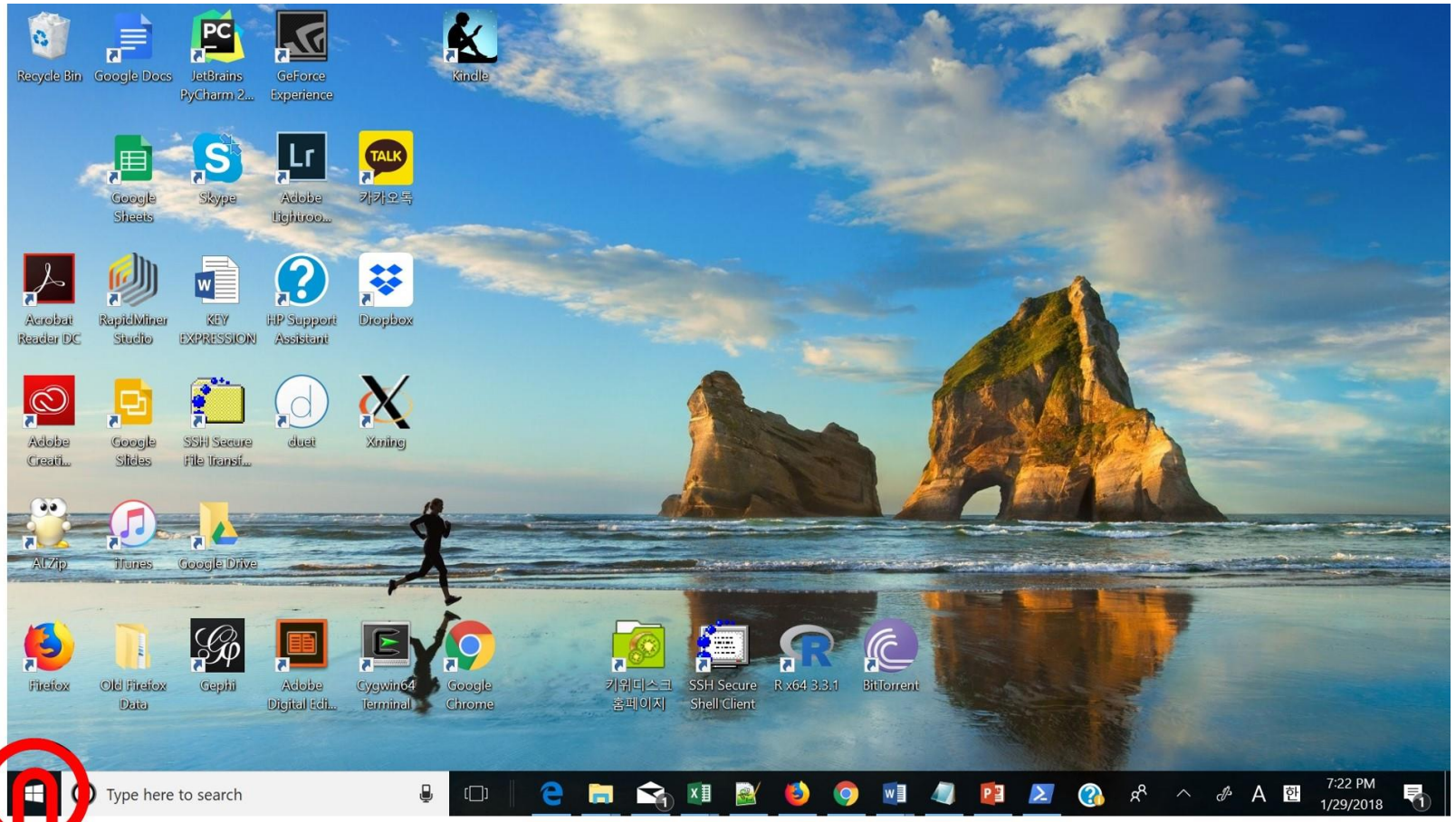
- Scope Note: Tumors or cancer of the human BREAST.
- Annotation: human only; BREAST NEOPLASMS, MALE is also available; for animal, index MAMMARY NEOPLASMS, ANIMAL or MAMMARY NEOPLASMS, EXPERIMENTAL; coordinate IM with histological type of neoplasm (IM)

MeSH Ontology Structure

- Neoplasms [C04]Neoplasms by Site [C04.588]
  - Abdominal Neoplasms [C04.588.033]
  - Anal Gland Neoplasms [C04.588.083]
  - Bone Neoplasms [C04.588.149]
  - **Breast Neoplasms [C04.588.180]**
    - Breast Carcinoma In Situ [C04.588.180.130]
    - Breast Neoplasms, Male [C04.588.180.260]
    - Carcinoma, Ductal, Breast [C04.588.180.390]
    - Carcinoma, Lobular [C04.588.180.437]
    - Hereditary Breast and Ovarian Cancer Syndrome [C04.588.180.483]
    - Inflammatory Breast Neoplasms [C04.588.180.576]
    - Unilateral Breast Neoplasms [C04.588.180.682]
    - Triple Negative Breast Neoplasms [C04.588.180.788]
  - Digestive System Neoplasms [C04.588.274]
  - Endocrine Gland Neoplasms [C04.588.322]
  - Eye Neoplasms [C04.588.364]

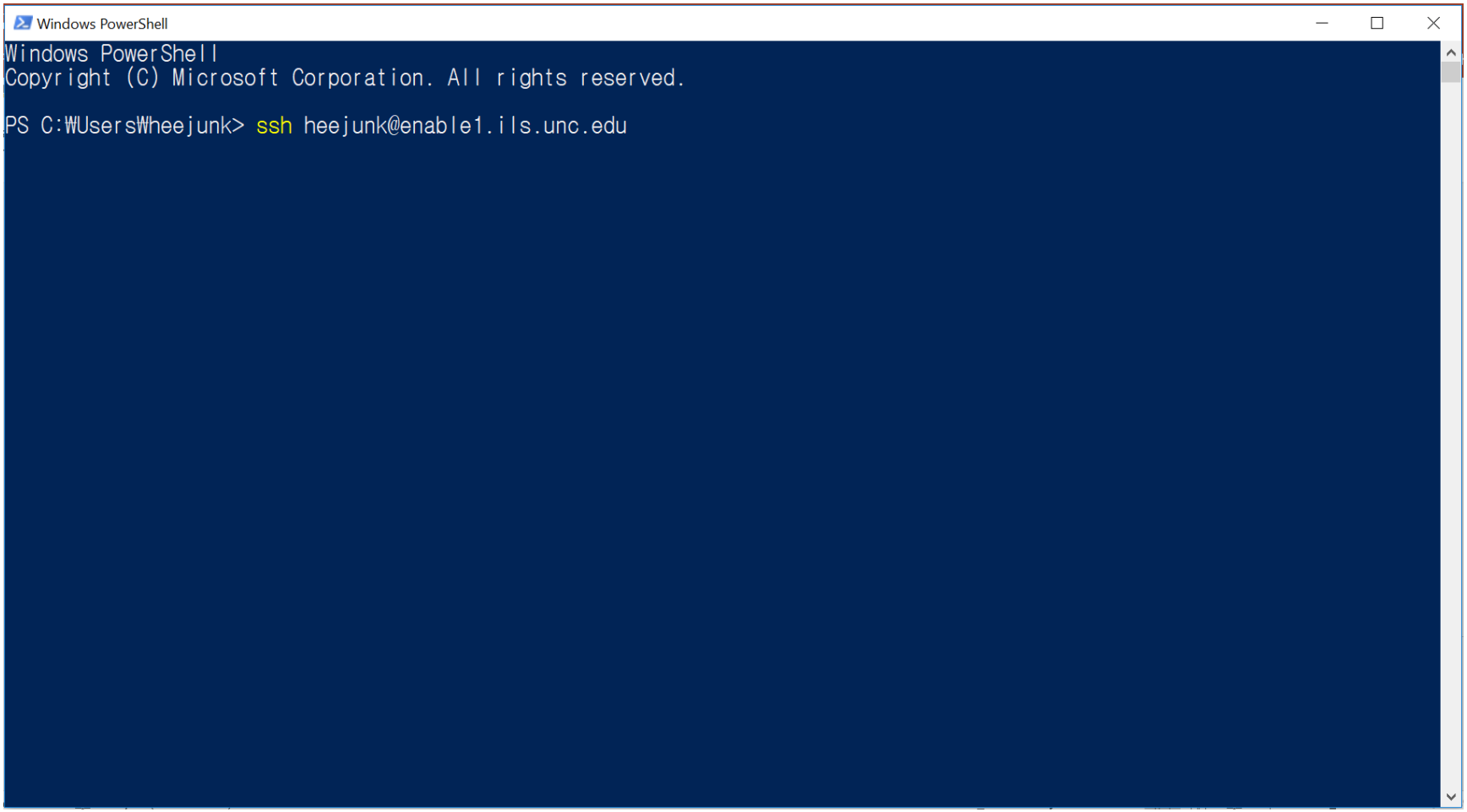
A Hands-on Practice

Starting Windows Powershell



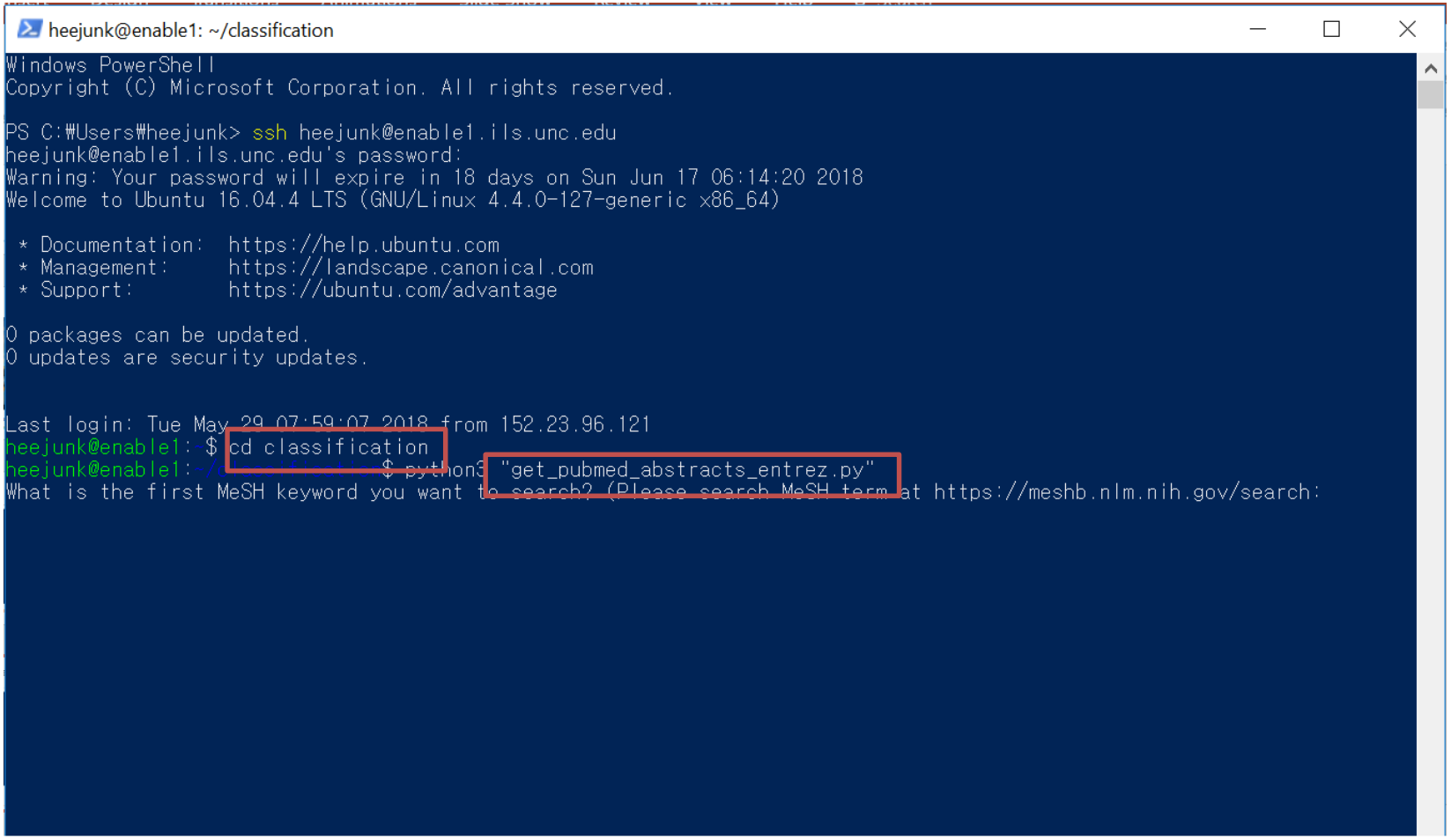
Right-click on the Windows icon at the bottom left of Windows screen and select "Windows Powershell."

Connecting to a Server



Replace your own ID with *heejunk* and type the Onyen password.

Running a Code for Collecting dataset



Move into a “classification” directory, run “get_pubmed_abstracts_entrez.py” and select MeSH.

Data Collected

- The Python script will prepare two data files: training.txt and testing.txt
- Each of these files contains 2,000 data instances and labels (MeSH)

Let's Modify the Python Code

- It's time to do your first programming
- It should be fun
- But, you should be very instructive and kind to computers…

How should the command to the computer be clear?

Link

Any Questions?