Integrating data science across undergraduate STEM curriculum

Drs. Kim Dill-McFarland, Dave Oliver, Steven Hallam

May 25, 2018

CTLT Spring Institute

Outline

- Introduction to data science
- *As a student*: Introduction to command line module
- *As an instructor*: Assess module and outline your own
- Wrap-up

Data science

- **Interdisciplinary field** of scientific methods, processes, algorithms and systems to extract knowledge or insights from **data in various forms**, either structured or unstructured, similar to data mining

- Includes big data, statistics, mathematics, and computer science

Applications in STEM

- Next-generation DNA, RNA, protein, metabolite analyses
- Global systems monitoring
- Personalized medicine

Applications in your daily life

- E-commerce and advertising
- City planning and traffic patterns
- Netflix, Hulu, etc.
- FitBit, Garmin, etc.

Importance of teaching data science

- Data is everywhere
- Skills in one data science application are highly transferable to other areas
- Critical thinking, complex questions, and no "right" answer
- Communication and collaboration across disciplines

Keys to data science curriculum success

- Hands-on, experiential learning
- Connections to domain knowledge
- Open doors for underserved groups
- Collaborative experiences
- Foundations for independent learning

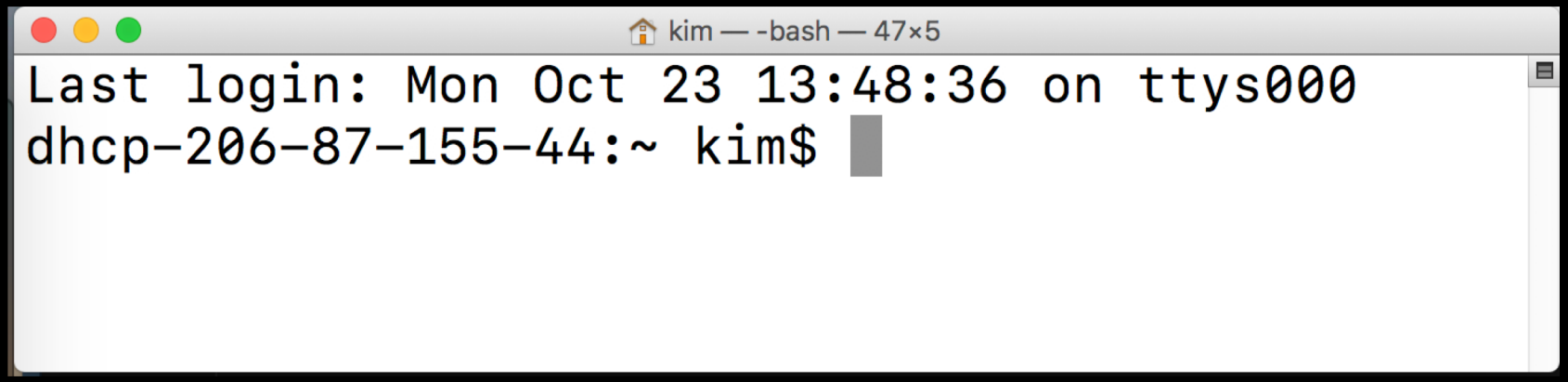
As a student:

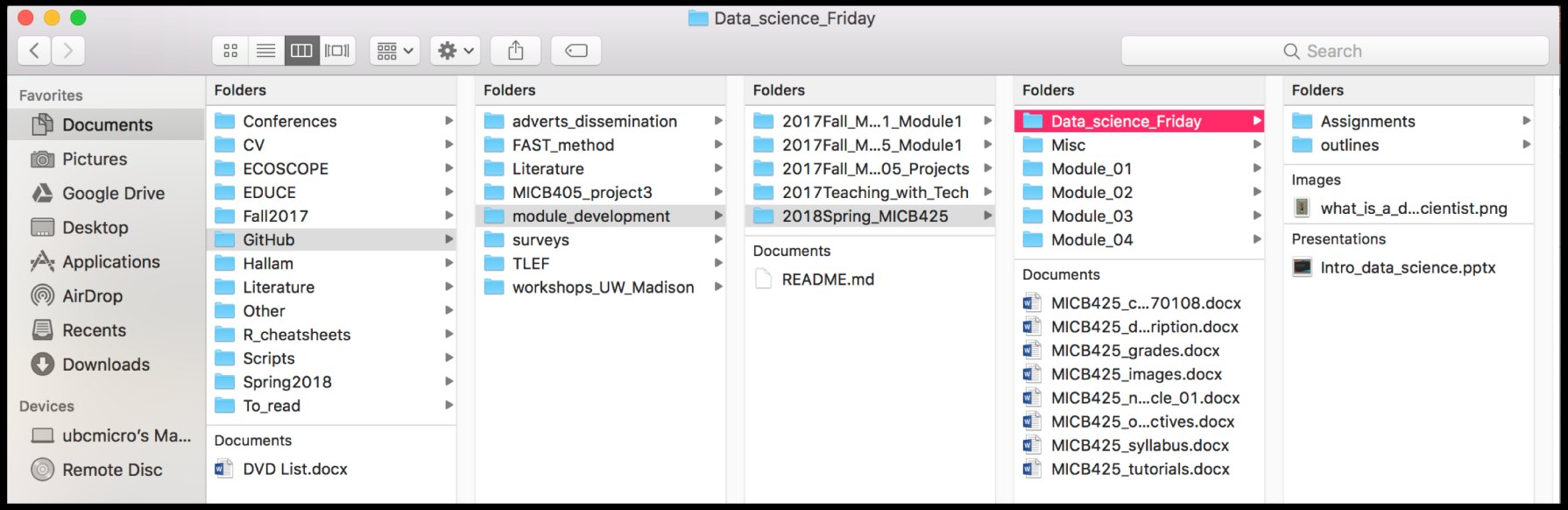
Introduction to command line module

Objectives

- Identify uses of command line inside and outside the classroom
- Apply basic syntax and functions in command line
- Use command line to navigate your computer's file structure

Command line terminal





Command line can…

- Do everything you do in Finder/File explorer
  - Make, move, copy, delete, rename files and folders
- Run specific programs
  - Text editors, R, sequence analysis tools
- Connect to and transfer files from remote servers like GitHub
- And more…

Why command line?

- Yes, everything we are doing in this class in command line could be done using the GUI (point and click).
- But…

Why command line?

- There are processes that are easier, faster, or can only be accomplished using the command line
  - Super computing resources can only be accessed via command line
- Command line allows you to provide reproducible code for your future self as well as supervisors and co-workers

Why command line?

- Commands are the same across different OS types and versions
- Command line knowledge makes you more competitive not only for careers in biology but also in other STEM areas, finance, etc…

Hands-on activity:

Follow along on your own computer

As an instructor:

Assess the module

What worked?

What could be improved?

Your own classroom?

- Run data faster on a remote server → Command line
- Statistics or Excel → Scripting in R/RStudio
- Repeated activities → Create a function in python
- Etc…

Data science at UBC

- Experiential Data science for Undergraduate Cross-disciplinary Education (EDUCE)
  http://ecoscope.ubc.ca/program-structure/educe/
  https://github.com/orgs/EDUCE-UBC/dashboard
- R workshops by ECOSCOPE
  http://ecoscope.ubc.ca/workshop-series/
- Masters in Data Science
  https://masterdatascience.science.ubc.ca/
- Skylight, CTLT, TLEFs https://tlef.ubc.ca/events/

Other data science resources

- Swirl http://swirlstats.com/
- Codecademy https://www.codecademy.com/catalog/subject/all
- edX https://www.edx.org/
- The Carpentries (data and software) https://carpentries.org/
- Compute Canada Bioinformatics Helpdesk https://bioinformatics.computecanada.ca/