EXECUTIVE SUMMARY

With almost every aspect of transportation research and practice driven to utilize complex computer software and innovative data sources (University Transportation Research Center, 2014), researchers and professionals increasingly face the computing challenges that have plagued other science and engineering disciplines (Merali, 2010). Most students of science, engineering, and planning are never taught to build, use, validate, and share software well (Merali, 2010). To help students and professionals in transportation research cope with these challenges, this course will apply lessons from similar programs in other disciplines (for example, Wilson et al., 2012) and aims to equip them with proper scientific computing skills.

The course plan and materials are developed with funding from TREC to support training transportation students and professionals in basic data science. The course materials are open-source under the Creative Commons (CC) license and publicly available online on GitHub at https://cities.github.com/datascience and the project report is avaialble at http://ppms.trec.pdx.edu/media/project_files/NITC_854_Introduction_to_Data_Science.pdf.

The project identifies and absorbs the best practices in data science in academic and professional literature (for example, Wilson et al., 2012), as well as successful courses and workshops for similar purpose, such as the Software Carpentry lessons, and develops the following course topics:

- Best practices in data science,
- Coding and scripting basics,
- Version control using Git,
- The import-tidy-transform-visualize-model-communicate workflow,
- Communication and reproducible research, and
- How to find help.

INTRODUCTION TO DATA SCIENCE

A Summer Course at TREC

Did you ever feel you are “drinking from a hose” with the amount of data you are attempting to analyze? Have you been frustrated with the tedious steps in your data processing and analysis process and thinking, “There’s gotta be a better way to do things”? Are you curious what the buzz of data science is about? If any of your answers are yes, then this course is for you.

Although computing is now integral to every aspect of science and engineering, transportation research included, most students of science, engineering, and planning are never taught how to build, use, validate, and share software well. As a result, many spend hours or days doing things badly that could be done well in just a few minutes and in a repeatable and self-documented way. The goal of this course is to empower students to spend less time wrestling with software and more time doing useful research/work.

Building on successful data science training programs such as the Software Carpentry (http://www.software-carpentry.org/) and Data Carpentry, and recent developments in related software and research, this course exposes transportation students and professional to the best practices in data science and scientific computing through lectures, discussion, and hands-on lab sessions and aims to help them tackle the challenge of “drinking from a hose” when dealing with an overwhelming amount of data.

1. DATE AND LOCATION

Date:

Part I: 8/6 – 8/7, 2018
Part II: 8/8 – 8/10, 2018

Location: Portland State University
Professional credits available

2. FORMAT AND SOFTWARE

Classes will all be hands-on sessions with lecture, discussions and labs. A major component of the class is the class project, in which students go through data retrieval, processing and analysis, and develop a report/article while learning the best practices of data science.

This course will use free R statistical software, with RStudio (https://www.rstudio.com/) as the main interface to R. The lecture and lab instructions will use R. It is possible (and encouraged) for existing Python users (and potentially users of other software, such as, Stata, Matlab, etc) to keep using the software they already know well. Students are encouraged to bring their own laptops. The instructor and TA will help the students set laptops up to run all examples and exercises in lectures and labs, and to re-run them later for review and their own project.

3. PREREQUISITES

Basic knowledge and experience of working with quantitative data; experiences and skills in (or keenness to learn) a programming language (e.g. Python) and/or data processing and statistical software (e.g. Python, R, Matlab, Stata).

4. TEXTBOOK AND READINGS

The course will use the following textbook:

- Wickham, H., Grolemund, G., 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data, 1 edition. ed. O'Reilly Media. (R4DS).

An electronic version of the book is available without charge on Hadley Wickham's website.

For Python users, Wes McKinney's book is recommended:

- McKinney, W., 2012. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 1 edition. ed. O'Reilly Media.

Additional articles and online resources are used as supplements to the textbook.

5. TOPICS

Topics are tentative and subject to change according to students' needs, but will be centered on these topical areas:

1. Part I:
    a. Coding and scripting basics
    b. Best practices in data science
    c. Version control using Git
    d. Workflow and RStudio, and
    e. How to find help
2. Part II:
    a. The import-tidy-transform-visualize-model-communicate workflow
    b. tidyverse suite of packages (dplyr, tidyr, ggplot2, and purrr)
    c. R Markdown

6. LICENSE

7. ACKNOWLEDGEMENTS

Development of the course has been supported by NITC grant # 854. Parts of the course materials have been adapted from the following sources:

- R for Data Science by Hadley Wickham

- Software Carpentry workshop lessons
- UBC Stat 545 by Professor Jenny Bryan at University of British Columbia
- NEU 5110 Introduction to Data Science by Professor Jan Vitek at Northeastern University

8. REFERENCES

Merali, Z., 2010. Computational science: ...Error. Nat. News 467, 775–777. doi:10.1038/467775a

University Transportation Research Center, 2014. Ground Transportation Technology Symposium: Big Data and Innovative Solutions for Safe, Efficient and Sustainable Mobility. New York, NY.

Wilson, G., Aruliah, D.A., Brown, C.T., Hong, N.P.C., Davis, M., Guy, R.T., Haddock, S.H.D., Huff, K., Mitchell, I.M., Plumbley, M., Waugh, B., White, E.P., Wilson, P., 2012. Best Practices for Scientific Computing. ArXiv12100530 Cs.

This course is hosted by the Transportation Research and Education Center (TREC) at Portland State University. The curriculum was developed under an education grant from the National Institute for Transportation and Communities (NITC), a program of TREC.

TREC, the Transportation Research and Education Center for Portland State University, houses the National Institute for Transportation and Communities (NITC), and the archives of its predecessor grant program, the Oregon Transportation Research and Education Consortium (OTREC). TREC also administers the Initiative for Bicycle and Pedestrian Innovation (IBPI), and other transportation grants and programs. We produce timely, practical research useful to transportation decision makers and support the education of future transportation professionals through curriculum development and student participation in research.

The National Institute for Transportation and Communities (NITC), one of five U.S. Department of Transportation national university transportation centers, is a program of the Transportation Research and Education Center (TREC) at Portland State University. The NITC program is a Portland State-led partnership with the University of Oregon, Oregon Institute of Technology, University of Utah and new partners University of Arizona and University of Texas at Arlington. We pursue our theme — improving mobility of people and goods to build strong communities — through research, education and technology transfer.