Strategic Data Science Series: Introduction to Data Science

Summer 2018
7/2/2018 - 7/4/2018
9:00 to 12:00 noon, 13:00 – 16:00

Instructor: Dr. Karl Ho, University of Texas at Dallas; kyho@utdallas.edu
Course github: **https://github.com/kho7/sds/introduction**

Overview: This is an introductory course covering the scope and methods of data science. It aims at providing a comprehensive framework in understanding data science, and the training roadmap for strategic data scientists, who host critical positions in national security, military strategies and defense and decision-making team members in key government agencies. The new knowledge of data science is not only intellectually important for this new genre of data scientists but also strategically necessary. This training will present the big picture to strategic data scientists covering from the basics of data science to such advanced topics as machine learning and spatial modeling. The course is designed to familiarize strategic data scientists with data concepts, tools and best practices. Topics on new developments and tools of big data will also be covered.

Learning Objectives:

At the completion of the course, students will be able to:

1. Understand scope and methods of data science training
2. Identify the roadmap for training and areas for specialization
3. Develop skills for prospective enrichment

Required Text:

Foster, Ian, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane (editors). 2016. Big Data And Social Science: A Practical Guide to Methods and Tools, Chapman and Hall/CRC Press.

Saltz, J.S. and Stanton, J.M., 2017. *An Introduction to Data Science*. SAGE Publications.

Schedule (This schedule is subject to change according to class progress):

Pre-reading:
Lazer, D., Kennedy, R., King, G. and Vespignani, A., 2014. The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), pp.1203-1205.

Day 1 outline:

1. What is data?
2. What is Big data?
3. Data Science and Data Analytics
4. The Story of Google Flu Trend
5. A Theory of Data: Understanding Data Generating Process
6. Statistical Modeling: The Two Cultures
7. Data scientist showcase:
   - **a.** Hans Rosling
   - **b.** Hal Varian
8. Data Science Roadmap
9. What is Data Literacy?

Prepare for Day 2:

1. Bring your own device
   a. Laptop computer on either Windows or MacOS operating system (OS). Tablets are not recommended.
   b. Run the latest update on OS since the computer is your most important companion in this class.
2. Software
   a. R (https://cran.r-project.org)
   b. RStudio (https://www.rstudio.com)
   c. Program/text editor at your choice. The following is only recommended but not required:
      i. Notepad++
      ii. WinEdt
      iii. Sublime Text (Mac)
      iv. Atom (Mac/Windows)
3. Reading
   a. Read recommended articles
4. Data
   a. Identify a literature of academic topic and collect data (secondary data will work)
   b. Collect charts from studies
5. Class discussion:
   a. What is data (in your understanding)?
   b. What is the most challenging data in your research area?
   c. Explain.

Prepare for Day 3

1. Reading:
   Wickham, H., 2010. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, *19*(1), pp.3-28.
2. Watch the following videos and write a few notes for discussion:
   1. Druckrey, Inge. 2012. *Teaching to See* (http://teachingtosee.org/film/TeachingToSee.html?gclid=EAIaIQobChMI38GMg8eH2QIVFLbACh0vIwqcEAAYASAAEgKzp_D_BwE)
   2. "The Big Deal about Big Data" with Dr. Gary King (https://www.youtube.com/watch?v=5h6MTLybccs)
   3. Software
      1. Tabula (https://tabula.technology/)
      2. QGIS (https://www.qgis.org/en/site/)

3. Class discussion:
   1. Data Analytics in application
   2. Design and operation
   3. Future of Data Science

Day 3 Outline:

1. Data Visualization
   a. Tableau
   b. R
   c. Web
   d. Animation
2. Spatial Models and Methods
   a. R/Leaflet
3. Data Analytics Methods
   a. Machine Learning
   b. Sentiment Analysis

Prepare Day 4

a. Population your own website with new charts and programs
b. Reading:
    a. Wickham, Hadley. 2007. ggplot2: past, present and future ((http://ggplot2.org/resources/2007-past-present-future.pdf)
c. Watch the following video and write a one-page note and review:
    a. Tufte, Edward. 2016. *The Future of Data Analysis* (https://channel9.msdn.com/Events/Machine-Learning-and-Data-Sciences-Conference/Data-Science-Summit-2016/MSDSS11)
d. Prepare for presentation schedule 8/1
    a. Identify the research question
    b. Data
    c. Visuals
    d. Anticipated contributions

Day 4 outline

1. Data Visualization
    a. Cognitive Science
    b. Color Science
    c. R packages
2. Workshop
    a. R data manipulation
    b. Presentation techniques

Prepare for Day 5

1. Watch:
    a. McGhee, Geoff. 2011. *Journalism in the Age of Data*, available at Vimeo (https://vimeo.com/14777910
2. Software
    a. Google Refine/Open Refine
        i. Watch Google Refine on Youtube: https://www.youtube.com/watch?v=B70J_H_zAWM
    b. Shiny
        i. Shiny tutorial: https://shiny.rstudio.com/tutorial/

Day 5 outline

1. Text mining/analytics
2. Web scraping
3. Sentiment analysis
4. Workshop
    a. R/Leaflet
    b. Shiny
    c. Python
    d. IDE

Prepare for Day 6

1. Learn about Github (https://guides.github.com)
2. Presentation

Day 6 outline

1. D3 library
2. Presentation
3. Discussion:

a. Future of Data Science in Social Sciences