Introduction to Data Science

- Point of data science is answering a question—the science is more important than the data
- Key challenges faced when doing data science:
    - Don't have enough data
    - Have too much data
- Statistics: science of learning from data; useful whenever there is uncertainty
- Data science skills = hacking skills + math and stats knowledge + substantive expertise
- What do data scientists do?
    1. Define question
    2. Define ideal data set
    3. Determine what data is accessible
    4. Obtain data
    5. Clean data
    6. Exploratory data analysis (look for patterns)
    7. Statistical prediction / modeling
    8. Interpret results
    9. Challenge results
    10. Synthesize results / write-up results
    11. Create distributable code
    12. Distribute results
- Key characteristics of hackers:
    - Willing to independently find answers, and know how to do so
    - Unintimidated by new data types or packages
    - Unafraid to say they don't know
    - Polite but relentless in seeking the answer

About R

R is main workhorse of data science

Benefits of R

- Most common language for data science (complemented by Python)
- Wide range of packages for all steps in data science process
- Free
- Amazing IDE (RStudio)
- Amazing ecosystem of developers
- Packages easy to install, and play nicely with each other

R scripts are textfiles ending in .R

R markdown documents are used to document research; .Rmd files are "knit" into HTML

Getting Help

There are several important R functions to get help. Let's say need information on the function rnorm, and want to access help from the console.

- ?rnorm will return some information, but requires the exact name.
- help.search("rnorm") will search the docs, and doesn't require the exact name
- args("rnorm") will display the arguments
- just typing the function name will spit out some basic info

Asking Questions about R

To ask a question about R, should provide answers to the following questions:

1. What steps reproduce the problem?
2. What is the expected output?
3. What do you see instead?
4. What version of product is being used (R + packages)?
5. What operating system is being used?

Where to look for answers:

- Archive of class forums
- Read manual
- Search web
- Ask skilled friend
- Post to class forums, R mailing list, or http://stackoverflow.com

Asking Questions about Data Analysis

To ask a question about data analysis, should provide answers to the following questions:

1. What is the question trying to answer?
2. What steps/tools are being used to answer it?
3. What am I expecting to see?
4. What am I seeing instead?
5. What other solutions have I tried / thought about?

The same resources for seeking answers about R are useful when seeking answers about data analysis, but another good resource specific to data analysis questions is http://crossvalidated.com.

When googling questions about data analysis, try searching for *[data type]data analysis* or *[data type] R package*

Try to identify what the data analysis field is called for the data type, and use that term when looking for answers

- Biostatistics for medical data
- Data science for web analytics data
- Machine learning for computer science / computer vision data
- Natural language processing for data from texts
- Signal processing for data from electrical signals
- Business analytics for customer data
- Econometrics for economic data
- Statistical process control for data about industrial processes.

Etiquette of Asking for Help

Etiquette when asking for help:

- Use specific titles
- Describe goal
- Be explicit
- Provide the minimum information necessary
- Be polite
- Follow up and post solutions if solve elsewhere
- Don’t use personal email