Python for Social Science

Ling 572: Python Scripting

Fall, 2017: No prerequisites

gawron@mail.sdsu.edu

TuTh 1100-1215 AH 2112

Jean Mark Gawron

San Diego State University, Department of Linguistics

2017-29-11

Overview

Introduction

Class Overview

Some background

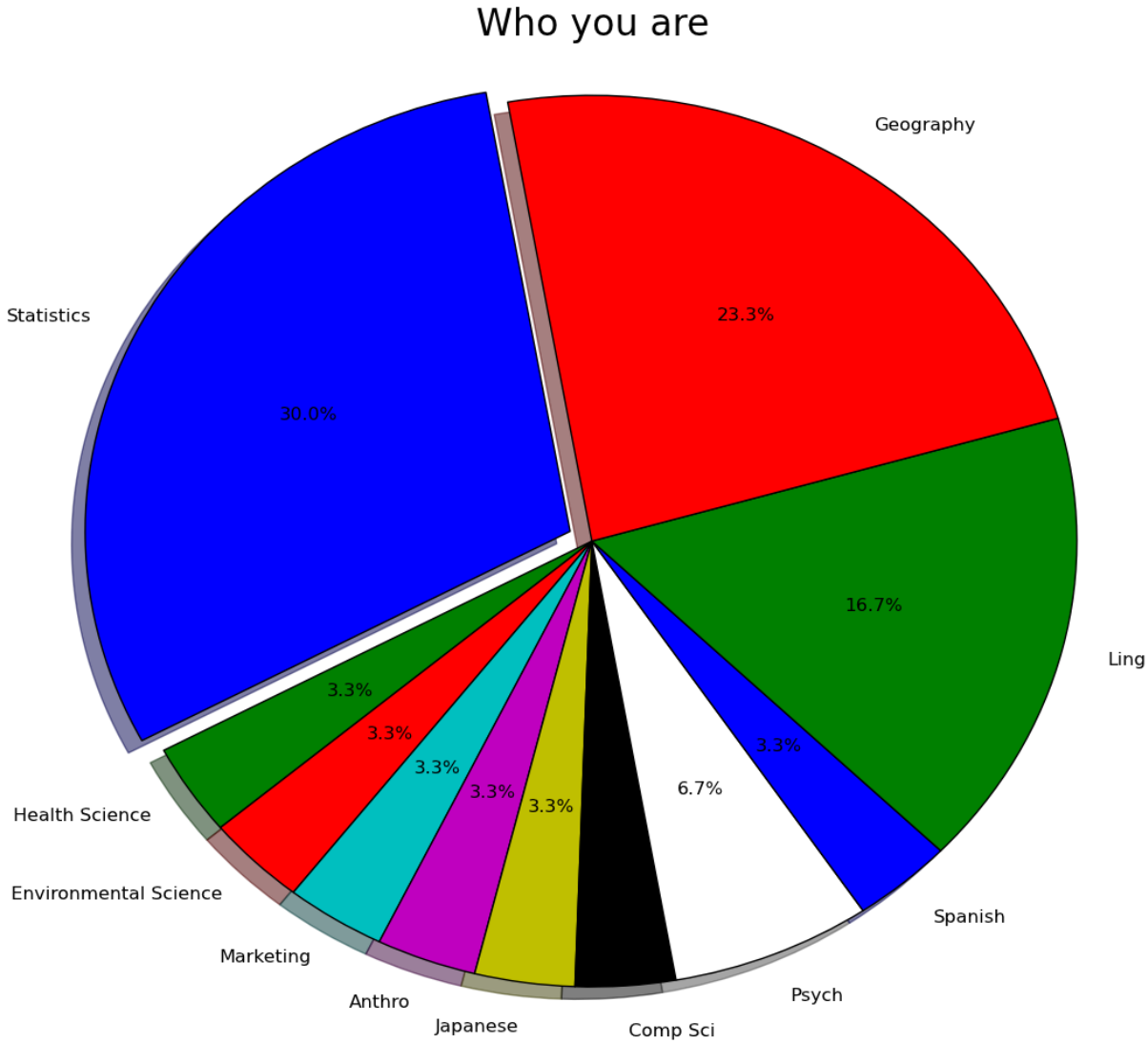
Assignments, quizzes, grading

Outline

Who am I?

1. A professor in the Department of Linguistics specializing in **Computational linguistics**
2. Machine translation, Speech recognition, Text classification, Topic identification
3. I have a lot of experience in introducing students without a lot of computational background to computational ideas

The code

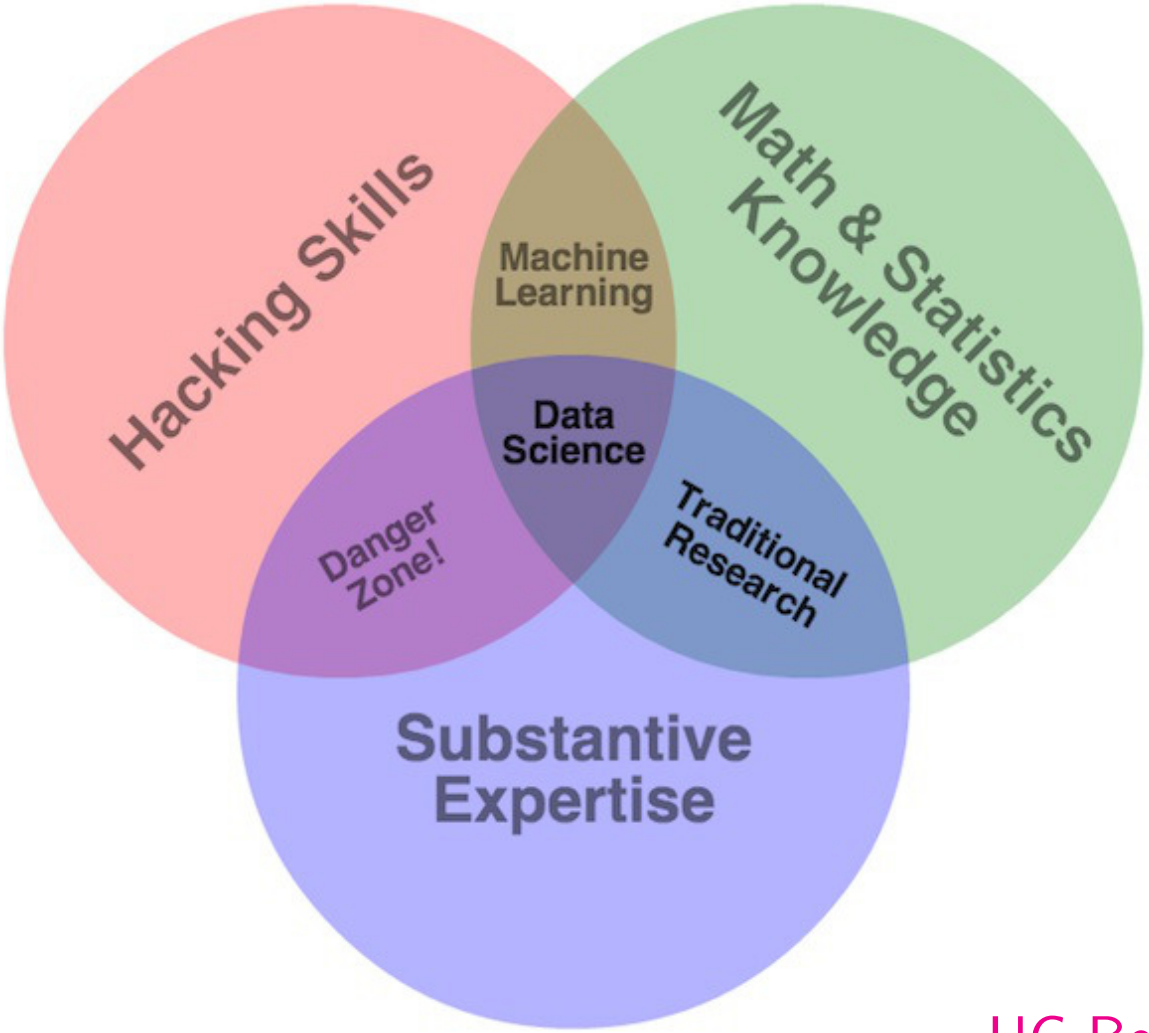

Who you are
Geography
23.3%
Statistics
30.0%
16.7%
Ling
3.3%
3.3%
3.3%
3.3%
3.3%
6.7%
3.3%
Health Science
Environmental Science
Marketing
Anthro
Japanese
Comp Sci
Psych
Spanish

Need

Big data and social science

Social scientists find themselves facing exponentially larger data sets without suitable tools to deal with them.

Where do you fit?



UC Berkeley Data Science

Master's

What is Python?

A programming language

1. Active and growing community of (data) scientists using it
2. Easy to learn
3. Easily constructed **scripts**: programs that construct pipelines combining the functionality of other programs
4. **Provides a formidable array of data collection, data manipulation, and data analysis tools**

Outline

Class goals

Who it's for

1. Graduate students and upper division undergraduates
2. Students with no knowledge of programming who want to get in on the data goldmine of the **Age of Information**
3. Students who have data that they need to drill into to reshape it or to extract specific kinds of information.
4. Students open to expanding their computational skills

Class prereqs

1. Some knowledge of what counts as interesting data in your particular discipline, and some experience working with it.
2. An interest in exploring some of the data opportunities provided by government websites, social networks, blogs, and the marketplace of ideas that is the Internet.

Topics

	1. Python Basics
Text Data	2. Searching for patterns in text and web data (regular expressions) 3. Classifying texts (machine learning) 4. Extracting information from big data sources (Government data)
Analysis/ visualization	5. Constructing social networks from data (visualizing social groups) 6. Connecting to your stat package (Python data frames) 7. Visualizing quantitative relationships on maps 8. Visualizing similarity relations

Data sources

1. PUMS (US Census)
2. Social Security Administration
3. Enron email data
4. Geocoding servers (Google) and geocoding DBs
5. USDA Food Database
6. Twitter
7. RSS news feeds

Outline

Good ideas

From Software carpenrty.org

- Documenting process for others
- Reproducibility of results
- Knowing how to test results
- Managing errors
- Posting to places like GitHub, BitBucket, and Figshare — The concept was more important than brandto make work sustainable even when students move on

Reproduceability

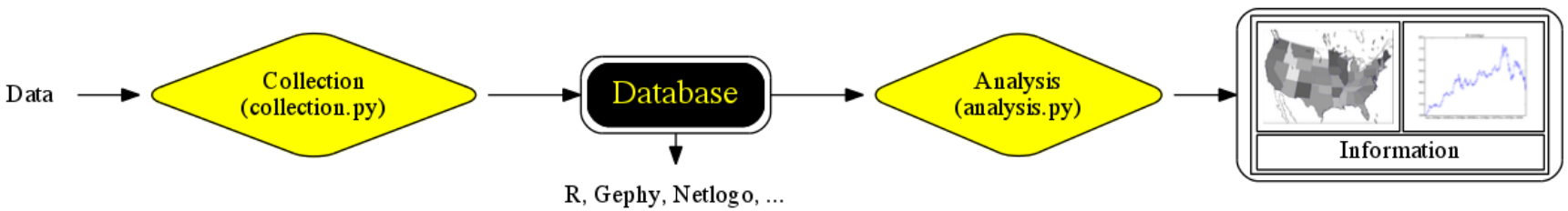
- Data management
- Documentation

Bedrock principle

Writing good programs to process our data is one of the best ways to keep track of exactly what we've done with our data.

The two tasks

Collection **Analysis**



Outline

Distribution

exercises 30%
quizzes 20%
midterm 25%
final project 25%

Assignments

1. Marked diagnostically (You generally get full credit if you turn in an assignment having made a sincere effort to complete it). As the term advances, the bar for what counts as sincere effort will get lifted a little higher.
2. Model answers will be posted online; we will go over these in class.
3. You will be responsible for assimilating the material in the model answers and learning from your mistakes.
4. The heart of the course: If you can successfully execute the assignments, you've learned a lot of Python.
5. Early assignments will be IPython (Jupyter) notebooks. Provides a web browser interface to Python. You can save your interaction sessions and turn them in. In later assignments, you will turn in python code files.
6. Most assignments will be eligible for group work. More on group work below.

Quizzes

1. Generally follow an assignment and test material covered on assignments
2. 1 week advance notice
3. Short answer: 20 minutes at the beginning of class (don't be late). Think of these as being like quizzes in a language class (French, Japanese, Russian, ...). You'll be asked how to say a few things, what a few things mean ...
   - 3.1 Using correct Python syntax, set the variable **L** to a list of the even integers, 20-30, inclusive.
   - 3.2 What are the Python types of the following objects?
     a. 'a': 'xx', 'b':'d'
     b. "Python"
     c. 1.7

Midterm

1. One inclass exam at the end of the **Basic Python** section of the class.
2. Union of the types of questions that have appeared on quizzes
3. Confirms that you have acquired basic elements of the language necessary for the data-oriented portion of the class.

Final project

1. An extended assignment. I will provide the data and the questions. [Maybe (group work allowed): Let's see how the group work is going!]
2. Optional for grad students: If you have a data collection/data analysis task that you think Python can help with (not too ambitious!), you can use this (in effect, I serve as a no-fee consultant).

Group assignments

1. Ideal size: (1-4)
2. Different backgrounds ok
3. Meet at least once outside of class for a working session
4. If you already know Python well, by all means, participate, but don't just give people the answers. Instead, help them figure out what's going wrong.