**Track:** Methodologies and Analytics

**Session Format:** Extended Workshop (75 minutes)

**Title:** Python 4 ways: A beginner’s primer to practical uses of Python in data analytics

**Proposal Narrative\***

**Provide a proposal narrative describing the session objective, why it is important and/or innovative to the IR/IE field, and how it aligns with the conference theme. (1750 characters)**

Python is an incredibly popular programming language for data analytics. With its straightforward coding logic, accessible language, and extreme versatility, Python is a powerful tool for any IR professional to adapt into their work.

Our objective is to show how our team at UCOP has adapted Python into data management and analytics projects and provide the audience with a hands-on examples of codes for common data processing projects. In our workshop, we will share how we use Python in our own work, and include the audience using a mock dataset that highlights 1) Creating, Merging and Organizing large datasets; 2) Cleaning and Error Checking Data; 3) Querying a database; and 4) Automating lengthy processes.

Participants will walk away with Python codes and example datasets they can use in their work, ideas of how to use the code, and resources to guide them in their learning.

Our presentation is a primer for participants interested in using Python in their own work. Coding can be intimidating, so we will show why investing in Python training is beneficial to an IR professional. Continuously improving data processes and automation is important in the IR/IE field, as it increases productivity, efficiency, accuracy and reusability in data handling. Python also has an extensive library of packages that can be used for basic to complex analysis projects.

Our presentation focuses on strengthening skills that are important for the future of IR. As we move towards increased automation, AI, and other complex machine learning functions, it is important to develop skills that adapt to the work being done. Coding skills are important for the future of IR/IE, and this presentation will introduce those interested to easy ways to get started in Python.

**Session Abstract (for the conference program) (maximum 1000 characters)\***

Are you considering using Python for data management and analysis, but aren’t sure where to start?

Our team made the decision in 2022 to switch our data processes to Python, even though we had no prior experience with the coding language. Three years later, many of our data processing and analysis procedures are supported by or completely automated through Python.

In our workshop, we will show how we use Python in our own work at UCOP. We will then guide the audience through hands-on examples of Python code for 1) Creating, Merging and Organizing large datasets; 2) Cleaning and Error Checking Data; 3) Querying a database; and 4) Automating lengthy processes.

Our workshop is for IR professionals interested in using Python for their own data work. Python users at all levels are welcome, but the focus is for beginners looking to begin or expand their Python skills. All interactive examples will use Google Colab (a cloud implementation of Jupyter), with a free Google Account being the only necessary prerequisite.

**Learning Outcomes**

**What are the participant learning outcomes? (Please phrase as “Participants who attend this session will be able to…”)**

Participants who attend this session be able to learn four Python coding scenarios that focus on data parsing and processing methods for common IR/IE requests. This includes interactive coding practice throughout the presentation that attendees can follow along with on their own computers using Google Colab. These mini-projects are aimed towards IR professionals either thinking about or at the beginning of learning Python for their jobs.

Furthermore, Participants who attend this session will be able to access and take home the code examples used in the presentation, as well as a guide on how to apply what they have learned to their own data projects.

**Session Outline\***

**Provide an outline indicating how the time will be used during the session. Please note portions where there will be active participation with the participants specific to the type of session.**

This presentation includes both a PowerPoint lecture and hands-on examples of Python.

**0-10 Minutes:**

Introductions, how to use Google Colab on attendee computers to follow along

**10-25 Minutes:** **Section 1**

Discuss Creating, Merging and Organizing large datasets.

* *Power Point*: Provide examples from our UCOP team instances where we handle multiple large datasets
* *Active Participation*: Introduce mock-example, guide attendees through merging together multiple Excel Files from several data sources to create one file

**25-40 Minutes: Section 2**

* *Power Point***:** Provide examples from our UCOP team instances where we clean and error check our data against a template
* *Active Participation***:** Using the data file created in **Section 1,** check against a template file and scan the text for errors (blanks, incorrect ID formats, incorrect birthday formats, etc.) and create an error report

**40-55 minutes: Section 3**

* *Power Point***:** Provide examples from our UCOP team instances where need to query a database to include extra data fields in our dataset
* *Active Participation:* Using the cleaned and error checked data file created in **Section 2**, demonstrate how to query a SQL based data warehouse to access demographic information on students

**55-70 minutes: Section 4**

* *Power Point:* Provide examples from our UCOP team instances where we have automated lengthy data collecting, cleaning, organizing and analyzing processes for repeat use
* *Active Participation:* Using the dataset created in **Section 3**, demonstrate how to automate Python processes using variable expressions and generalized coding methods so that users can easily reuse the code for subsequent projects

**70-end:**

Wrapping up the project, go over resources that attendees have access to

**List 3-5 keywords that summarize your topic\***

Python

Data Automation

Data Cleaning