Data Science for the Humanities and Social Sciences

4:30-4:45 Appetizers, drinks, networking
4:45 Presentations begin
   Jonathan Halverson (Research Computing and PICSciE)
   Ryan Heuser (Center for Digital Humanities)
   Peter Ramadge (Center for Machine Learning)
   Sarah-Jane Leslie (Philosophy)
   Bobray Bordelon (Data and Statistical Services)
   Ofira Schwartz (Stokes Viz Hub)
   Boriana Pratt (Office of Population Research)
   Neggin Keshavarzian (Princeton Research Data Service)
   Irene Kopaliani (Research Computing)
   Musashi Jacobs-Harukawa (Data Driven Social Sciences)
5:20 Presentations end
5:20-5:30 More networking
Acknowledgments

Thanks to the Center for Digital Humanities for hosting.

Thanks to James Van Wyck, Gabby Xu (co-organizers) and the Graduate School for the food/drinks.

Why are we here today?

To inform students and researchers in the humanities and social sciences about resources for learning and doing data science.

Please share this information (slides will be sent out). And please give us your input.
Research Computing and PICSciE

Services and Support

- High-performance computing
- Data storage and networking
- Software engineering
- Visualization and GIS
- Debugging and troubleshooting
- Training
- 1-on-1 help

Researchers often come to Research Computing when their laptops prove to be insufficient for their work

- 90,000 CPU-cores
- 400+ GPUs
- 50 PB of storage
- 1600 users
- Free of charge

Free of charge
FALL 2022
MINI-COURSES AND WORKSHOPS

https://researchcomputing.princeton.edu/workshops

Scan the QR code to register and to learn more about the workshops.

Google Cloud Platforms (GCP)/Workshops
GCP Overview
October 20 | 10:00-11:30AM
Big Data & Machine Learning on GCP
October 20 | 12:30-2:00PM
Machine Learning Options on GCP
October 20 | 2:30-4:00PM

Michele Milano, Perceptive Architect and Scientist, Google
Hindy, Platform, Customer Experience, Google

Creating and Running Software Containers with Docker and Singularity
November 9 | 4:30-6:00PM
Jesse Nowak, Research Software & Computing Team Leader

Google Cloud Platform (GCP) Workshops

MACHINE LEARNING
Foundations of Deep Learning, Parts 1 & 2
August 31 | 1:00-5:00PM
September 1 | 1:00-5:00PM

Santhanam Ramchandran, NVIDIA

Getting Started with Transformers for Language Modeling
November 3 | 4:30-6:00PM
David Venat, Senior Research Software Engineer

Software Engineering
Continuous Integration and Automated Software Testing
September 31 | 10:00AM-12:00PM
Daniel Suss, Senior Linux Administration & Programming Analyst

Accelerate Your Coding with PyCharm and Visual Studio Code
October 6 | 4:30-6:00PM
Vincent Barsan, Senior Research Software Engineer

Introduction to Version Control Using Git, Parts 1 & 2
October 6 | 4:30-6:00PM
October 13 | 4:30-6:00PM
Dave Olson, Graduate Student

Everything You Didn’t Know You Needed: Research Computing Tips & Tricks
November 2 | 4:30-6:00PM
Mi Kim Coyle, Research Data Scientist

SUBSCRIBE TO RESEARCH COMPUTING NEWSLETTER /
EMAIL: research@princeton.edu

https://researchcomputing.princeton.edu/workshops
Useful Links

- Research Computing Training for the Humanities and Social Sciences
- Upcoming PICSciE/Research Computing Workshops and Live Training
- Getting Started with the Research Computing Systems
- Subscribe to the PICSciE-Research Computing mailing list

Contact

Jonathan D. Halverson, Ph.D.
Research Software and Computing Training Lead
Princeton Institute for Computational Science and Engineering
346 Lewis Library | 609-258-9575 | halverson@princeton.edu
‘Python for Poets’ and the Center for Digital Humanities

Wintersession 2023
Mixer
What is ‘digital humanities’?

Thinking humanistically about data and computation, and computationally about humanities research questions.
DH as a research method

• Primary sources as data
  o Transcribe
  o Organize
  o Visualize

  o Text Analysis
  o Mapping
  o Network Analysis
We partner with Princeton faculty, students, library colleagues and staff to design and develop world-class projects that span disciplines and technologies. In the process, the CDH serves as a space for critical discussions about data, technology and the human experience in the academy and in the public sphere. We raise awareness of the cultural context and implications of our increasingly data-driven scholarship and daily lives.
Recreating the world of the Lost Generation in interwar Paris.
Text Analysis

• Distant Reading
  – Instead of reading one book or a few books, reading 100 or 1000 books at once to identify larger patterns

Examples:

Mark Algee-Hewitt:  The Performance of Character

Micki Kaufman:  Quantifying Kissinger; Stack Graphs

New York Times:  Cliches of ESPN
<table>
<thead>
<tr>
<th>Artist</th>
<th># of Unique Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMX</td>
<td>2,675</td>
</tr>
<tr>
<td>21 Savage</td>
<td>2,675</td>
</tr>
<tr>
<td>A Boogie wit Da Hoodie</td>
<td>3,050-3,425</td>
</tr>
<tr>
<td>Lil Baby</td>
<td>3,425-3,800</td>
</tr>
<tr>
<td>Lil Durk</td>
<td>3,425-3,800</td>
</tr>
<tr>
<td>Wiz Khalifa</td>
<td>3,425-3,800</td>
</tr>
<tr>
<td>Lil Uzi Vert</td>
<td>3,800-4,175</td>
</tr>
<tr>
<td>NF</td>
<td>4,175-4,550</td>
</tr>
<tr>
<td>YoungBoy Never Broke Again</td>
<td>4,550-4,925</td>
</tr>
<tr>
<td>Common</td>
<td>15,000-19,000</td>
</tr>
<tr>
<td>The Roots</td>
<td>15,000-19,000</td>
</tr>
<tr>
<td>The Blackalicious</td>
<td>15,000-19,000</td>
</tr>
<tr>
<td>Canibus</td>
<td>15,000-19,000</td>
</tr>
<tr>
<td>Ghostface Killah</td>
<td>15,000-19,000</td>
</tr>
<tr>
<td>Immortal Technique</td>
<td>15,000-19,000</td>
</tr>
<tr>
<td>Wu-Tang Clan</td>
<td>15,000-19,000</td>
</tr>
<tr>
<td>Atmosphere</td>
<td>15,000-19,000</td>
</tr>
<tr>
<td>Del the Funky Homosapien</td>
<td>15,000-19,000</td>
</tr>
<tr>
<td>YG</td>
<td>2,675</td>
</tr>
<tr>
<td>YoungBoy Never Broke Again</td>
<td>2,675</td>
</tr>
</tbody>
</table>

Notes/sources:

(1) Since this analysis uses an artist's first 35,000 lyrics (prioritizing studio albums), an artist's era is determined by the years the albums were released. Some artists may be identified with a certain era (for example, Jay-Z with the 1990s, with Reasonable Doubt in 1996, In My Lifetime, Vol. 1 in 1997, etc.) yet continue to release music in the present day.

All lyrics are via Genius.
What is ‘Python for Poets’?

Thinking poetically about programming in Python, and learning to program through experiments on poetry and language.
Introduction to Cultural Analytics & Python

Designed by Melanie Walsh // Powered by Jupyter Book

This website hosts an online textbook, Introduction to Cultural Analytics & Python, which offers an introduction to the programming language Python that is specifically designed for people interested in the humanities and social sciences.

This book demonstrates how Python can be used to study cultural materials such as song lyrics, short stories, newspaper articles, tweets, Reddit posts, and film screenplays. It also introduces computational methods such as web scraping, APIs, topic modeling, Named Entity Recognition (NER), network analysis, and mapping.

These materials were originally created to support "Introduction to Cultural Analytics: Data, Computation & Culture," an undergraduate course taught at Cornell University and the University of Washington.

What is Cultural Analytics?
Cultural analytics is the study of culture with computational methods. Culture is a broad term that includes literature, history, politics, art, music, social media, and a lot more. Cultural analytics is a growing research area in fields like Digital Humanities and Information Science.

Interactive Code 📊
This book’s pages contain Python code that can be opened and run in a web browser without any prior installation or configuration. Click the launch button 📊 in the top right corner of any page to open and play with the code. See How To Interact With This Book for more information.
Python for Poets (Part 1 of 2)
by Wintersession

Tue, Jan 24, 2023
2 PM – 5 PM EST (GMT-5)
Add to Calendar

Center for Digital Humanities Classroom
Princeton, NJ 08544, United States
View Map
Machine Learning: A Practical Introduction for Humanists and Social Scientists

Sarah-Jane Leslie

sjleslie@Princeton.edu

sarahjaneleslie.github.io
PHI 543/SML 543
Machine Learning: A Practical Introduction

- A regular, for-credit, transcripted graduate course (graded or audit)
- No prior coding experience or college-level math required
- Emphasis on developing practical skills
- “Zero-to-hero”
By the end of the course students will:

➢ be able to code and train a variety of basic deep learning models
➢ develop an appreciation of the range of humanities/social science research questions to which deep learning can be applied
➢ be fluent collaborators on research projects that involve machine learning experts
➢ gain an understanding than will inform theorizing about machine learning (e.g., for research in AI ethics, technology policy, etc)
PHI 543/SML 543
Machine Learning: A Practical Introduction

- For a tentative syllabus please visit:
  - sarahjaneleslie.github.io/PHI543-sp23
  - Or sarahjaneleslie.github.io and follow link
  - My email: sjleslie@princeton.edu
DATA AND STATISTICAL SERVICES
DSS SUMMARY

• Services began in 1964
• Collects, describes, organizes, & stores data (numeric/text)
  • Also linguistic corpora
• Helps discover & understand data
• Ensures contract conditions met
• Helps with quantitative statistical software & social science methodology
HOLDINGS

- ICPSR (world’s largest social science data archive)
- Roper (world’s largest public opinion archive)
- Sociometrics (primarily health data)
- Canadian microdata
- Purchased datasets from governmental, nonprofit, & corporate organizations
- Links & descriptions to hundreds of datasets, series, & archives
WHAT DOES DSS DO?

• **Subject Librarians**
  - Help researchers find appropriate data, understand methodology
  - Purchase data in their areas
  - Available to provide specialized sessions for classes

• **Statistical Consultants** (Professional and PhDs)
  - Provide assistance with statistical software
    - Primarily R and Stata; To a lesser extent depending on staffing SAS, SPSS, Matlab, Python
    - Helps subset, shape, create summary statistics, visualize
  - Provide assistance with choosing correct methods in the social sciences (primarily quantitative)
  - Apply models in the statistical software
  - Teach Stata and R workshops; Also available to provide specialized sessions for classes
DSS LAB (FIRESTONE A.12G)

- Statistics software
  - Stata/SE 9-17 (32/64 bit)
  - Stata/MP 17 (64 bit)
  - SPSS 28 (PASW)
  - SAS 9.4
  - R, RStudio
  - MATLAB
  - Python 2.7/3.6 (Anaconda suite)

- Hardware
  - 13 PCs with 28" monitors (32 GB RAM)
  - Super-mini computer - Dual Intel Xeon Platinum 8180M (28Core); 3Tb RAM
  - Emulation machine

- Conversion software
  - Stat/Transfer

- Compression software
  - 7-Zip (freeware)
Stoke Viz Hub

Ofira Schwartz
December 5, 2022
The Stokes Visualization Hub is a space and service that responds to the evolving digital research, data visualization, and qualitative analysis needs of the Princeton University community.
Stokes Viz Hub - Mission

Design and teach workshops focusing on data visualization, qualitative and quantitative data analysis

Support researchers’ qualitative and quantitative analysis, and data visualization needs

Work in collaboration with experts from other departments
# Workshops

**Stokes Viz Hub Workshops**  
Stokes Library, Wallace Hall, Room 070

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Graphics</th>
<th>Software</th>
<th>Coding</th>
</tr>
</thead>
</table>
| Intro to R for Policy: Reproducing Bertrand & Mullainathan (2004), racial discrimination in the labor market  
Sept. 20 at 1 p.m. to 2:30 p.m. | Intro to Qualitative Research Methods ...Qualitative v. quantitative, research, types of qualitative research  
Oct. 26 at 1 p.m. to 2:30 p.m. |  
| Reference Management with Zotero ...a single, searchable interface  
Sept. 23 at 1 p.m. to 2 p.m. | Intro to Data Analysis with Strata ...Learn data prep, descriptive stats, data visualization, linear regression  
Oct. 28 at 1 p.m. to 2:30 p.m. |  
| Intro to Data Analysis with R  
...Learn data prep, descriptive stats, data visualization, linear regression  
Sept. 30 at 1 p.m. to 2:30 p.m. | Visualizing Interconnections in the Social Sciences ...Library databases, VOSviewer  
Nov. 4 at 10 a.m. to 11:30 a.m. |  
| Intro to R for Policy: Reproducing Card and Krueger (1993), effect of minimum wage and employment  
Oct. 4 at 10 a.m. to 11:30 a.m. | Foundations of Data Visualization ...Best practices of data visualization  
Nov. 11 at 10:30 a.m. to 11:30 a.m. |  
| Intro to ArcGIS  
...Use software to explore geographic data, tables and maps  
Oct. 7 at 1:30 p.m. to 3:00 p.m. | Data Visualization with R: Introduction to ggplot...Visualizing data using R's ggplot package  
Nov. 18 at 1 p.m. to 2:30 p.m. |  
| Bibliometric Analysis Using R: Visualizing scholarly networks  
Oct. 13 at 2 p.m. to 3:30 p.m. | Questions?  
Please contact Ofra Schwartz-Solicher, Social Sciences Data & Sociology Librarian, oschwartz@princeton.edu |
Consultations

• A team of graduate students

• Expertise with the appropriate research methodology and software packages
  – R, Stata, Python, MAXQDA

• Consultations by appointment
E-classroom

Software packages include (and more):

R
Stata
SAS
Tableau
ArcGIS
NVivo
QDA Miner
SPSS Statistics 27
Thank You!

Image courtesy of: Ketut Subiyanto - pexels.com
R Data Wrangling: tidyverse packages tidyr and dplyr

Presented as part of Wintersession 2023

Boriana Pratt, Office of Population Research, Princeton University
R Data Wrangling: tidyverse package tidyr

- Tidy data vs messy data
- Reshape data
- Split and combine columns
R Data Wrangling: tidyverse package dplyr

- Subset observations
- Reorder observations
- Select variables
- Add new variables
- Group observations
- Summarize groups of observations
Research Data Management at Princeton

Neggin Keshavarzian
Research Data Management Specialist
Princeton Research Data Service
About PRDS

Part of the Office of Research Data and Open Scholarship, Princeton University Library

Joint initiative with:
Office of the Dean for Research
Office of Information Technology
Office of the Provost

Wind Cowles
Director, Office of Research Data and Open Scholarship

Matt Chandler & Neggin Keshavarzian
Research Data Management Specialists
Princeton Research Data Service

Rishi Joshi
Research Data Management and Storage Engineer,
Princeton Research Computing
PRDS Resources for Graduate Students

- Organizing Data
- Publishing Open Data
- Data Management Planning
- Open Research Practices
- Preserving Data
- Sharing Data

- Workshops & Programs
- Web Resources & Tools
- Consultations

researchdata.princeton.edu
PRDS Workshops & Programs

- **Modular Webinars**
  - Data File Inventory
  - Using Code Ocean
  - Data Wrangling

- **Recurring Workshop Series**
  - Humanities Data
  - Preparing Data for Publication
  - Data Spring Cleaning

- **Annual Research Data Stewardship Program** (Open registration now)
  - Comprehensive training geared toward grad students and postdocs
  - Spring Semester: Weekly workshops and hands-on interactive sessions
  - Certificate and alumni network

Find the latest and register on our [Events and Training](researchdata.princeton.edu) page.

View recordings on the [PRDS channel](researchdata.princeton.edu) in Media Central.
Web Resources & Tools

- **Research Lifecycle Guide**
  - Resources on all stages of the research lifecycle
  - Continuously expanding for different fields and use cases
  - Regularly updated with the latest policies and best practices

- **Ever-growing list of tools and resources**
## Task & Skill Clusters

<table>
<thead>
<tr>
<th>Organize</th>
<th>Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>File / folder naming</td>
<td>Planning (DMPs)</td>
</tr>
<tr>
<td>Versioning</td>
<td>Pre-registration</td>
</tr>
<tr>
<td>Open formats</td>
<td>Code commenting</td>
</tr>
<tr>
<td>Inventories &amp; backups</td>
<td>READMEs / codebooks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Collaborate</th>
<th>Publish &amp; Preserve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access and security</td>
<td>READMEs and metadata</td>
</tr>
<tr>
<td>Version control</td>
<td>Digital repositories</td>
</tr>
<tr>
<td>Containerization</td>
<td>Discovery and access</td>
</tr>
<tr>
<td>Project management</td>
<td>Reuse and citation</td>
</tr>
</tbody>
</table>
Welcome to the Princeton Research Data Service

Established in 2019, we provide Princeton’s diverse research community with expert services and infrastructure to store, manage, retain, and curate digital research data, and to make their digital research data available to the broader network of academic researchers, as well as the general public. We provide consultations, training, and data curation services to researchers throughout the life cycle of research projects, working with them to make the process of data management and storage as seamless as possible with their current research practices.
Thank You!

Neggin Keshavarzian
Research Data Management Specialist
neggink@princeton.edu
Data-Driven Social Science Initiative

Musashi Jacobs-Harukawa, DPhil
5 December 2022
About Me

- PoliSci/Data Science background
- Research on NLP, Interpretability:
  - Explanation and Validation for DL Language Models
  - Using LLMs for Corpus Description and Summary
  - How/when does domain-specific pre-training matter?

Website: muhark.github.io
About DDSS

- **Innovation** in data-/computationally-intensive research.
- **Support** faculty/students in changing technical landscape.
- **Community** building and production of public goods.
- **Position** Princeton as leader in quant soc sci.

Website: [ddss.princeton.edu](http://ddss.princeton.edu)
Workshop Preview (Next Semester)

1. Demystifying Deep Learning
2. (Tools for) Interpretation and Explanation
3. Computational Social Science: Mapping the State of the Field
‘Python for Poets’ and the Center for Digital Humanities

Wintersession 2023
Mixer
What is ‘digital humanities’?

Thinking humanistically about data and computation, and computationally about humanities research questions.
DH as a research method

• Primary sources as data
  o Transcribe
  o Organize
  o Visualize
  o Text Analysis
  o Mapping
  o Network Analysis
Projects
Exploring the intersections of the humanities and technology

We partner with Princeton faculty, students, library colleagues and staff to design and develop world-class projects that span disciplines and technologies. In the process, the CDH serves as a space for critical discussions about data, technology and the human experience in the academy and in the public sphere. We raise awareness of the cultural context and implications of our increasingly data-driven scholarship and daily lives.
Text Analysis

• Distant Reading
  – Instead of reading one book or a few books, reading 100 or 1000 books at once to identify larger patterns

Examples:

Mark Algee-Hewitt:  The Performance of Character

Micki Kaufman: Quantifying Kissinger; Stack Graphs

New York Times: Cliches of ESPN
<table>
<thead>
<tr>
<th>Artist</th>
<th>Words</th>
<th>1980s</th>
<th>1990s</th>
<th>2000s</th>
<th>2010s</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMX</td>
<td>2,675</td>
<td>3,050</td>
<td>3,425</td>
<td>3,800</td>
<td></td>
</tr>
<tr>
<td>A Boogie wit...</td>
<td>2,675</td>
<td>3,050</td>
<td>3,425</td>
<td>3,800</td>
<td></td>
</tr>
<tr>
<td>Lil Baby</td>
<td>3,050</td>
<td>3,425</td>
<td>3,800</td>
<td>4,175</td>
<td></td>
</tr>
<tr>
<td>Lil Durk</td>
<td>3,425</td>
<td>3,800</td>
<td>4,205</td>
<td>4,550</td>
<td></td>
</tr>
<tr>
<td>Wiz Khalifa</td>
<td>3,800</td>
<td>4,205</td>
<td>4,550</td>
<td>4,925</td>
<td></td>
</tr>
<tr>
<td>Lil Uzi Vert</td>
<td>4,175</td>
<td>4,550</td>
<td>4,925</td>
<td>5,300</td>
<td></td>
</tr>
<tr>
<td>YoungBoy.Navy</td>
<td>4,550</td>
<td>4,925</td>
<td>5,300</td>
<td>5,675</td>
<td></td>
</tr>
<tr>
<td>NF</td>
<td>4,925</td>
<td>5,300</td>
<td>5,675</td>
<td>6,050</td>
<td></td>
</tr>
<tr>
<td>Run-D.M.C.</td>
<td>5,300</td>
<td>5,675</td>
<td>6,050</td>
<td>6,425</td>
<td></td>
</tr>
<tr>
<td>2Pac</td>
<td>5,675</td>
<td>6,050</td>
<td>6,425</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Big L</td>
<td>6,050</td>
<td>6,425</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insane Clown...</td>
<td>6,425</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC Lyte</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scarface</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three 6 Mafia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UGK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dizzee Rascal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jadakiss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kano</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lil Kim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nelly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rick Ross</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T.I.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Chainz</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AZ &amp; Ferg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Big KRT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brockhampton</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cupcakke</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hopsin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jay Rock</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kendrick Lamar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mac Miller</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ScHoolboy Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyga</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vince Staples</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Das EFX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goodie Mob</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redman</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>straw</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alliance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bronson</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QAAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killah Priest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RZA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wu-Tang Clan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jedi Mind Tr...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MF DOOM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aesop Rock Band</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bueller</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes/sources:

1. Since this analysis uses an artist's first 35,000 lyrics (prioritizing studio albums), an artist's era is determined by the years the albums were released. Some artists may be identified with a certain era (for example, Jay-Z with the 1990s, with Reasonable Doubt in 1996, In My Lifetime, Vol. 1 in 1997, etc.) yet continue to release music in the present day.

All lyrics are via Genius.
What is ‘Python for Poets’?

Thinking poetically about programming in Python, and learning to program through experiments on poetry and language.
Introduction to Cultural Analytics & Python

Designed by Melanie Walsh // Powered by Jupyter Book

This website hosts an online textbook, *Introduction to Cultural Analytics & Python*, which offers an introduction to the programming language Python that is specifically designed for people interested in the humanities and social sciences.

This book demonstrates how Python can be used to study cultural materials such as song lyrics, short stories, newspaper articles, tweets, Reddit posts, and film screenplays. It also introduces computational methods such as web scraping, APIs, topic modeling, Named Entity Recognition (NER), network analysis, and mapping.

These materials were originally created to support "Introduction to Cultural Analytics: Data, Computation & Culture," an undergraduate course taught at Cornell University and the University of Washington.

**What is Cultural Analytics?**
Cultural analytics is the study of culture with computational methods. Culture is a broad term that includes literature, history, politics, art, music, social media, and a lot more. Cultural analytics is a growing research area in fields like Digital Humanities and Information Science.

**Interactive Code 🧪**
This book’s pages contain Python code that can be opened and run in a web browser without any prior installation or configuration. Click the launch button 🧪 in the top right corner of any page to open and play with the code. See How To Interact With This Book for more information.
Python for Poets (Part 1 of 2)

by Wintersession

Tue, Jan 24, 2023
2 PM – 5 PM EST (GMT-5)

Center for Digital Humanities Classroom
Princeton, NJ 08544, United States

Add to Calendar
View Map
Machine Learning: A Practical Introduction for Humanists and Social Scientists

Sarah-Jane Leslie

sjleslie@Princeton.edu

sarahjaneleslie.github.io
PHI 543/SML 543
Machine Learning: A Practical Introduction

- A regular, for-credit, transcripted graduate course (graded or audit)
- No prior coding experience or college-level math required
- Emphasis on developing practical skills
- “Zero-to-hero”
By the end of the course students will:

➢ be able to code and train a variety of basic deep learning models

➢ develop an appreciation of the range of humanities/social science research questions to which deep learning can be applied

➢ be fluent collaborators on research projects that involve machine learning experts

➢ gain an understanding than will inform theorizing about machine learning (e.g., for research in AI ethics, technology policy, etc)
PHI 543/SML 543
Machine Learning: A Practical Introduction

- For a tentative syllabus please visit:
  - sarahjaneleslie.github.io/PHI543-sp23
  - Or sarahjaneleslie.github.io and follow link
  - My email: sjleslie@princeton.edu
DATA AND STATISTICAL SERVICES
DSS SUMMARY

• Services began in 1964
• Collects, describes, organizes, & stores data (numeric/text)
  • Also linguistic corpora
• Helps discover & understand data
• Ensures contract conditions met
• Helps with quantitative statistical software & social science methodology
HOLDINGS

• ICPSR (world’s largest social science data archive)
• Roper (world’s largest public opinion archive)
• Sociometrics (primarily health data)
• Canadian microdata
• Purchased datasets from governmental, nonprofit, & corporate organizations
• Links & descriptions to hundreds of datasets, series, & archives
WHAT DOES DSS DO?

- **Subject Librarians**
  - Help researchers find appropriate data, understand methodology
  - Purchase data in their areas
  - Available to provide specialized sessions for classes
- **Statistical Consultants** (Professional and PhDs)
  - Provide assistance with statistical software
    - Primarily R and Stata; To a lesser extent depending on staffing SAS, SPSS, Matlab, Python
    - Helps subset, shape, create summary statistics, visualize
  - Provide assistance with choosing correct methods in the social sciences (primarily quantitative)
  - Apply models in the statistical software
  - Teach Stata and R workshops; Also available to provide specialized sessions for classes
DSS LAB (FIRESTONE A.12G)

Statistics software
- Stata/SE 9-17 (32/64 bit)
- Stata/MP 17 (64 bit)
- SPSS 28 (PASW)
- SAS 9.4
- R, RStudio
- MATLAB
- Python 2.7/3.6 (Anaconda suite)

Hardware
- 13 PCs with 28" monitors (32 GB RAM)
- Super-mini computer - Dual Intel Xeon Platinum 8180M (28Core); 3Tb RAM
- Emulation machine

Conversion software
- Stat/Transfer

Compression software
- 7-Zip (freeware)
Stoke Viz Hub

Ofira Schwartz
December 5, 2022
Stokes Viz Hub - Vision

The Stokes Visualization Hub is a space and service that responds to the evolving digital research, data visualization, and qualitative analysis needs of the Princeton University community.
Stokes Viz Hub - Mission

Design and teach workshops focusing on data visualization, qualitative and quantitative data analysis

Support researchers’ qualitative and quantitative analysis, and data visualization needs

Work in collaboration with experts from other departments
# Workshops

**Stokes Viz Hub Workshops**

Stokes Library, Wallace Hall, Room 070

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Graphics</th>
<th>Software</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intro to R for Policy:</strong> Reproducing Bertrand &amp; Mullainathan (2004), racial discrimination in the labor market</td>
<td><strong>Intro to Qualitative Research Methods:</strong> Qualitative v. quantitative research, types of qualitative research</td>
<td><strong>Intro to Data Analysis with Strata:</strong> Learn data prep, descriptive stats, data visualization, linear regression</td>
<td><strong>Visualizing Interconnections in the Social Sciences:</strong> Library databases, VOSviewer</td>
</tr>
<tr>
<td>Sept. 20 at 1 p.m. to 2:30 p.m.</td>
<td>Oct. 26 at 1 p.m. to 2:30 p.m.</td>
<td>Oct. 28 at 1 p.m. to 2:30 p.m.</td>
<td>Nov. 4 at 10 a.m. to 11:30 a.m.</td>
</tr>
<tr>
<td><strong>Reference Management with Zotero:</strong> a single, searchable interface</td>
<td><strong>Intro to Data Analysis with R:</strong> Learn data prep, descriptive stats, data visualization, linear regression</td>
<td><strong>Foundations of Data Visualization:</strong> Best practices of data visualization</td>
<td><strong>Data Visualization with R:</strong> Introduction to ggplot...Visualizing data using R’s ggplot package</td>
</tr>
<tr>
<td>Sept. 23 at 1 p.m. to 2 p.m.</td>
<td>Sept. 30 at 1 p.m. to 2:30 p.m.</td>
<td>Nov. 11 at 10:30 a.m. to 11:30 a.m.</td>
<td>Nov. 18 at 1 p.m. to 2:30 p.m.</td>
</tr>
<tr>
<td><strong>Intro to R for Policy:</strong> Reproducing Card and Krueger (1993), effect of minimum wage and employment</td>
<td><strong>Bibliometric Analysis Using R:</strong> Visualizing scholarly networks</td>
<td><strong>Questions?</strong> Please contact Ofra Schwartz-Solicher, Social Sciences Data &amp; Sociology Librarian, <a href="mailto:oschwart@princeton.edu">oschwart@princeton.edu</a></td>
<td></td>
</tr>
<tr>
<td>Oct. 4 at 10 a.m. to 11:30 a.m.</td>
<td>Oct. 13 at 2 p.m. to 3:30 p.m.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Registration required. Learn more at [bit.ly/stokes_vizhub_2022](bit.ly/stokes_vizhub_2022)
Consultations

• A team of graduate students

• Expertise with the appropriate research methodology and software packages
  – R, Stata, Python, MAXQDA

• Consultations by appointment
E-classroom

Software packages include (and more):

- R
- STATA
- SAS
- Tableau
- ArcGIS
- NVIVO
- QDA Miner
- SPSS

STATISTICS 27
Thank You!

Image courtesy of: Ketut Subiyanto – pexels.com
R Data Wrangling: tidyverse package tidy\(r\)

- Tidy data vs messy data
- Reshape data
- Split and combine columns
R Data Wrangling: tidyverse package dplyr

- Subset observations
- Reorder observations
- Select variables
- Add new variables
- Group observations
- Summarize groups of observations
Research Data Management at Princeton

Neggin Keshavarzian
Research Data Management Specialist
Princeton Research Data Service
About PRDS

Part of the Office of Research Data and Open Scholarship, Princeton University Library

Joint initiative with:
Office of the Dean for Research
Office of Information Technology
Office of the Provost

researchdata.princeton.edu

Wind Cowles
Director, Office of Research Data and Open Scholarship

Matt Chandler & Neggin Keshavarzian
Research Data Management Specialists
Princeton Research Data Service

Rishi Joshi
Research Data Management and Storage Engineer,
Princeton Research Computing
PRDS Workshops & Programs

● Modular Webinars
  ○ Data File Inventory
  ○ Using Code Ocean
  ○ Data Wrangling

● Recurring Workshop Series
  ○ Humanities Data
  ○ Preparing Data for Publication
  ○ Data Spring Cleaning

● Annual Research Data Stewardship Program (Open registration now)
  ○ Comprehensive training geared toward grad students and postdocs
  ○ Spring Semester: Weekly workshops and hands-on interactive sessions
  ○ Certificate and alumni network

Find the latest and register on our Events and Training page.

View recordings on the PRDS channel in Media Central.
Web Resources & Tools

- **Research Lifecycle Guide**
  - Resources on all stages of the research lifecycle
  - Continuously expanding for different fields and use cases
  - Regularly updated with the latest policies and best practices

- **Ever-growing list of tools and resources**
<table>
<thead>
<tr>
<th>Task &amp; Skill Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organize</strong></td>
</tr>
<tr>
<td>File / folder naming</td>
</tr>
<tr>
<td>Versioning</td>
</tr>
<tr>
<td>Open formats</td>
</tr>
<tr>
<td>Inventories &amp; backups</td>
</tr>
<tr>
<td><strong>Document</strong></td>
</tr>
<tr>
<td>Planning (DMPs)</td>
</tr>
<tr>
<td>Pre-registration</td>
</tr>
<tr>
<td>Code commenting</td>
</tr>
<tr>
<td>READMEs / codebooks</td>
</tr>
<tr>
<td><strong>Collaborate</strong></td>
</tr>
<tr>
<td>Access and security</td>
</tr>
<tr>
<td>Version control</td>
</tr>
<tr>
<td>Containerization</td>
</tr>
<tr>
<td>Project management</td>
</tr>
<tr>
<td><strong>Publish &amp; Preserve</strong></td>
</tr>
<tr>
<td>READMEs and metadata</td>
</tr>
<tr>
<td>Digital repositories</td>
</tr>
<tr>
<td>Discovery and access</td>
</tr>
<tr>
<td>Reuse and citation</td>
</tr>
</tbody>
</table>

researchdata.princeton.edu
Welcome to the Princeton Research Data Service

Established in 2019, we provide Princeton's diverse research community with expert services and infrastructure to store, manage, retain, and curate digital research data, and to make their digital research data available to the broader network of academic researchers, as well as the general public. We provide consultations, training, and data curation services to researchers throughout the life cycle of research projects, working with them to make the process of data management and storage as seamless as possible with their current research practices.
Thank You!

Neggin Keshavarzian
Research Data Management Specialist
neggink@princeton.edu

researchdata.princeton.edu
prds@princeton.edu
Cloud Usage at Princeton

• Collection of available Cloud resources
  https://researchcomputing.princeton.edu/systems/cloud-computing

• To get started
  Cloud Computing - Princeton Service Portal (service-now.com)

• Things to keep in mind
  • Shared Security
  • Data Transfer Fees
  • Help on Campus

Irene Kopaliani
Cloud Architect
ik8@princeton.edu
Data-Driven Social Science Initiative

Musashi Jacobs-Harukawa, DPhil
5 December 2022
About Me

- PoliSci/Data Science background
- Research on NLP, Interpretability:
  - Explanation and Validation for DL Language Models
  - Using LLMs for Corpus Description and Summary
  - How/when does domain-specific pre-training matter?

Website: [muhark.github.io](https://muhark.github.io)
About DDSS

- **Innovation** in data-/computationally-intensive research.
- **Support** faculty/students in changing technical landscape.
- **Community** building and production of public goods.
- **Position** Princeton as leader in quant soc sci.

Website: [ddss.princeton.edu](http://ddss.princeton.edu)
Workshop Preview (Next Semester)

1. Demystifying Deep Learning
2. (Tools for) Interpretation and Explanation
3. Computational Social Science: Mapping the State of the Field