

# I ILLINOIS

## School of Information Sciences

### Research Impact and Bibliometrics

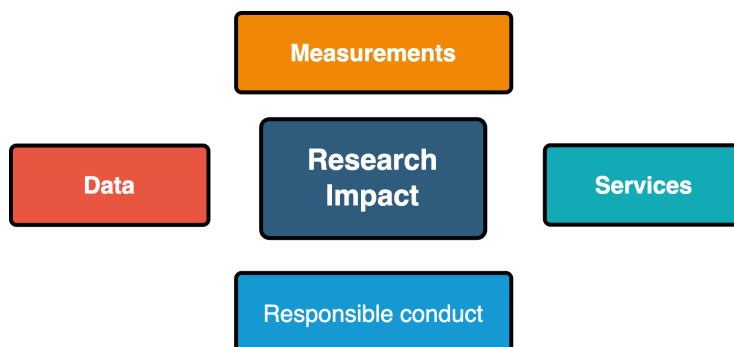
IS597RIO (CRN: 72551)

Spring 2023

Online (Canvas and Zoom)

Course meets 1 time per week for 1:55 minutes in Zoom, Tuesdays 2-3:55PM

Last updated 2023-07-24



### Instructors

**Jodi Schneider, PhD**

**Email:** [jodi@illinois.edu](mailto:jodi@illinois.edu) (preferred; please resend if I don't reply in 48 business hours)

**Office Hour:** After class and by appointment

**Phone (voicemail):** (217) 300-4328

**Tzu-Kun (Esther) Hsiao**

**Email:** [tkhsiao2@illinois.edu](mailto:tkhsiao2@illinois.edu) (preferred; please resend if I don't reply in 48 business hours)

**Office Hour:** Thursdays 1 PM and by appointment

### Course Description

Bibliometrics has long played a critical role in identifying developments and trends in science. Expertise in bibliometrics has gained visibility in aiding academic communities to measure and understand the research landscape and researchers', research groups', or institutions' positions. A bibliometric analysis requires expert knowledge in data collection, wrangling, visualization, and interpretation, along with understanding indicators and limitations. Outcomes of bibliometric analysis influence decision- and policy-making at various levels.

This course provides an intensive introduction to the fundamentals of bibliometrics from the research impact perspective. The course aims to acquaint students with the theories, concepts, methods, and applications of bibliometrics in research impact evaluation. The materials are broad in scope and cover topics including but not limited to research impact, research profiles, scientific



CC0 "No Rights Reserved"

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

collaboration, research opportunity identification, university ranking systems, and altmetrics. Students will acquire hands-on skills in implementing bibliometric analysis. Students will also learn the limitations of bibliometrics.

### **Pre- and Co-requisites**

None

Recommended experience: Data analysis, data cleaning, or programming experience. This could be from hands-on experience, or learned in a course.

*Example courses with relevant material include: IS 407 Introduction to Data Science, IS 417 Data Science in the Humanities, IS 430 Foundations of Information Processing, IS 457 Data Storytelling, IS 527 Network Analysis, IS 537 Theory and Practice of Data Cleaning, IS 575 Metadata in Theory and Practice, IS 577 Data Mining, IS 595 LD Linked Data Processing, IS 597 PY Python Standard Library, IS 597 DM Open Data Mashups, IS 597 PD - Practical Health Data Analytics*

### **# Credit hours**

4 graduate hours

### **Student Learning Objectives or Outcomes**

1. Understand fundamental concepts, theories, and methods of evaluating research impact.
2. Understand the data sources and the process of collecting, cleaning, and manipulating data.
3. Gain knowledge in measuring research impact, identifying research trends, and visualizing scientific collaborations.
4. Learn skills and workflows of conducting bibliometric analysis and interpreting results.
5. Design and execute a small-scale analysis project.

### **Course Context**

This course meets a number of learning outcomes connected to program objectives for the MSLIS, MSIM and PhD program, which in turn connect to larger iSchool and University of Illinois learning goals.

### **Program Learning Outcomes**

#### MSBio

1. Define and successfully address a tractable research question or real-world problem in health, medical, and bio-informatics using the appropriate scientific and/or research methods.
2. Accurately convey the implications of analytical results (in both oral and written modalities) to diverse stakeholders.



CC0 "No Rights Reserved"

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

MSIM

3. Manage information using best practices in management and policy; knowledge representation; and data analytics
4. Define and successfully address a tractable research question or real world problem in information management using the appropriate scientific and/or research methods.
5. Accurately convey the implications of analytical results (in both oral and written modalities) to diverse stakeholders.
6. Apply best practices for providing value, leadership, and team building.
7. Stay up-to-date by learning how to read, analyze, discuss, synthesize, and critique advances reported in the research literature.

MSLIS

1. Communicate capably with diverse stakeholders, promoting not just access to but also effective use of information services and systems in specific contexts.
2. Use evidence to help address information problems, meet information needs, and create relationships in their institutions, communities, profession, and the world.

PhD

1. Intellectual Reasoning and Knowledge
2. Creative Inquiry and Discovery
3. Social and Cultural Awareness and Understanding
4. Effective Leadership and Community Engagement

**iSchool Goal**

This course meets the following goal:

- Maintain global leadership in education for the information professions

**University of Illinois Campus-Wide Learning Goals**

This course meets the following goals):

1. Intellectual Reasoning and Knowledge
2. Creative Inquiry and Discovery
3. Effective Leadership and Community Engagement
4. Social Awareness and Cultural Understanding
5. Global Consciousness

**Course materials**

**Book** required purchase, \$17 from Illinois Bookstore; digital and used copies typically ~\$10; the [library ebook](#) does not allow printing more than 50 pages.

Sugimoto, Cassidy R., & Larivière, Vincent (2018). *Measuring Research: What Everyone Needs to Know*. Oxford University Press.

**Other readings are also available electronically.**



CC0 "No Rights Reserved"

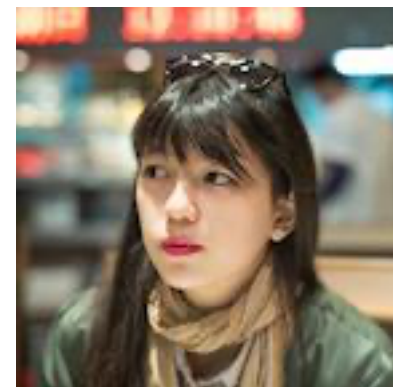
*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

### About the Instructors

**Jodi Schneider** is associate professor at the School of Information Sciences. Dr. Schneider studies science using argumentation, network science, knowledge graphs and metadata, and mixed methods. Her long-term research agenda analyzes controversies applying science to public policy; how knowledge brokers influence citizens; and whether controversies are sustained by citizens' disparate interpretations of scientific evidence and its quality. She earned degrees in informatics, mathematics, library & information sciences, and Great Books. Before becoming a researcher, she worked as an actuarial analyst for a Fortune 500 insurance company, as the gift buyer for a small independent bookstore, and in college and university libraries. She has held research positions across the U.S. as well as in Ireland, England, France, and Chile. She received an NSF CAREER award, and has held funding from NIH, Alfred P. Sloan Foundation, Institute for Museum and Library Services, the U.S. Office of Research Integrity, the European Union, and Science Foundation Ireland.



**Tzu-Kun (Esther) Hsiao** is a PhD student at the School of Information Sciences. Her research interests include knowledge diffusion, citation context, and scholarly communication. Her research combines natural language processing and machine learning techniques with citation analysis. She used citation links and citation contexts to study how patents were cited in academic articles and how retracted articles were cited after their retractions. She is currently investigating the reuse of text and facts in research articles.



### Methods of Assessment

#### *Summary of course assignments*

<b>Deliverable</b>	<b>Deadline(s)</b>	<b>Contribution to Course Grade</b>
Final Assignment (individual or group project)	Weeks 3-17	50%
Assignment 1 – Data wrangling and data manipulation	Week 4: February 7	15%
Assignment 2 – Building a research profile	Week 7: February 28	15%
Participation and Instructor Assessment	All semester	20%

All assignments are required for all students. Completing all assignments is not a guarantee of a passing grade. All work must be completed in order to pass this class. Contact the instructors if you need an extension, PRIOR to the due date. Late or incomplete assignments will not be given full credit unless the student has contacted the instructors prior to the due date of the assignment (or in the case of emergencies, as soon as practicable). Late assignments will include a penalty (half a letter grade for each day late), and no late assignments will be accepted after 5 days from



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

the date due. In the event that you are unsuccessful with an assignment, you may request to rewrite an assignment with instructor permission.

### **Attendance/ Participation Policy**

The iSchool expects students to attend all classes except in cases of emergency.

Student Code on Attendance: <http://studentcode.illinois.edu/article1/part5/1-501/>

You are also expected to participate in class. You are invited to initiate or engage in discussions.

### **Missing class should be rare.**

Attendance will be taken for each class meeting. When you need to miss class, we expect you to:

- Inform both instructors in advance, or, in case of emergency, as soon as possible after class.
- To get an excused absence:
  - For the first absence, we do not need an explanation. A short email “I will be absent on date \_\_\_\_\_.” is sufficient.
  - If you will be absent 2 or more times this semester, please contact me to make arrangements, so that you do not lose credit due to events outside your control.

Use the recording and office hours to catch up within the following week; prompt attention to the material is important because new material depends on the previous lectures.

### **Class conduct**

1. Students share some of the responsibility for fostering an inclusive classroom. Students are expected to be respectful of others' perspectives and lived experiences during class discussion.
2. Students are expected to demonstrate respect for the ideas and opinions of all other members of the class at all times. Failure to observe this course requirement can result in a failing course participation grade, and may result in a failing grade for the course.

### **Incomplete grades**

An exceptional request for an incomplete grade is most often granted to students encountering a medical emergency or other extraordinary circumstances beyond their control. Students must request an incomplete grade from the instructor. The instructor and student will agree on a due date for completion of coursework. The student must submit an Incomplete Form signed by the student, the instructor, and the student’s academic advisor to the front office:

<https://uofi.app.box.com/s/sx7arobhr0gfw12teaetmp1qq32ifdrd>

Please see the Student Code for full details: <http://studentcode.illinois.edu/article3/part1/3-104/>

### **Grading Scale**

94-100 = A

90-93 = A-

87-89 = B+

83-86 = B

80-82 = B-

77-79 = C+



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

73-76 = C  
 70-72 = C-  
 67-69 = D+  
 63-66 = D  
 60-62 = D-  
 59 and below = F

### **Academic Integrity**

Please review and reflect on the academic integrity policy of the University of Illinois, <https://studentcode.illinois.edu/article1/part4/1-401/> to which we subscribe. By turning in materials for review, you certify that all work presented is your own and has been done by you independently, or as a member of a designated group for group assignments. If, in the course of your writing, you use the words or ideas of another writer, proper acknowledgement must be given (using APA, Chicago, or MLA style). Not to do so is to commit plagiarism, a form of academic dishonesty. If you are not absolutely clear on what constitutes plagiarism and how to cite sources appropriately, now is the time to learn. Please ask! Please be aware that the consequences for plagiarism or other forms of academic dishonesty will be severe. Students who violate university standards of academic integrity are subject to disciplinary action, including a reduced grade, failure in the course, and suspension or dismissal from the University.

### **Statement of Inclusion**

<http://www.inclusiveillinois.illinois.edu/mission.html> As the state's premier public university, the University of Illinois at Urbana-Champaign's core mission is to serve the interests of the diverse people of the state of Illinois and beyond. The institution thus values inclusion and a pluralistic learning and research environment, one which we respect the varied perspectives and lived experiences of a diverse community and global workforce. We support diversity of worldviews, histories, and cultural knowledge across a range of social groups including race, ethnicity, gender identity, sexual orientation, abilities, economic class, religion, and their intersections.

### **Religious Observances**

In keeping with our Statement of Inclusion and Illinois law, the University is required to reasonably accommodate its students' religious beliefs, observances, and practices in regard to admissions, class attendance, and the scheduling of examinations and work requirements. Please contact the instructor regarding any concerns or accommodations.

### **Accessibility Statement**

To obtain accessibility-related academic adjustments and/or auxiliary aids, students with disabilities must contact the course instructor and the [Disability Resources and Educational Services](#) (DRES) as soon as possible. To contact DRES you may visit 1207 S. Oak St., Champaign, call 333-1970 (V/TTY), or e-mail a message to [disability@illinois.edu](mailto:disability@illinois.edu).



CC0 "No Rights Reserved"

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.



## Land acknowledgement Statement

Adopted by the University of Illinois in 2018

More information: [https://chancellor.illinois.edu/land\\_acknowledgement.html](https://chancellor.illinois.edu/land_acknowledgement.html)

*As a land-grant institution, the University of Illinois at Urbana-Champaign has a responsibility to acknowledge the historical context in which it exists. In order to remind ourselves and our community, we will begin this event with the following statement. We are currently on the lands of the Peoria, Kaskaskia, Peankashaw, Wea, Miami, Mascoutin, Odawa, Sauk, Mesquaki, Kickapoo, Potawatomi, Ojibwe, and Chickasaw Nations. It is necessary for us to acknowledge these Native Nations and for us to work with them as we move forward as an institution. Over the next 150 years, we will be a vibrant community inclusive of all our differences, with Native peoples at the core of our efforts.*

*Land Acknowledgement Statement Suggested by Native American House:*

*I/We would like to begin today by recognizing and acknowledging that we are on the lands of the Peoria, Kaskaskia, Piankashaw, Wea, Miami, Mascoutin, Odawa, Sauk, Mesquaki, Kickapoo, Potawatomi, Ojibwe, and Chickasaw Nations. These lands were the traditional territory of these Native Nations prior to their forced removal; these lands continue to carry the stories of these Nations and their struggles for survival and identity. As a land-grant institution, the University of Illinois has a particular responsibility to acknowledge the peoples of these lands, as well as the histories of dispossession that have allowed for the growth of this institution for the past 150 years. We are also obligated to reflect on and actively address these histories and the role that this university has played in shaping them. This acknowledgement and the centering of Native peoples is a start as we move forward for the next 150 years.*

## Mental Health Resources

Diminished mental health, including significant stress, mood changes, excessive worry, substance/alcohol abuse, or problems with eating and/or sleeping can interfere with optimal academic performance, social development, and emotional wellbeing. The University of Illinois offers a variety of confidential services including individual and group counseling, crisis intervention, psychiatric services, and specialized screenings at no additional cost. If you or someone you know experiences any of the above mental health concerns, it is strongly encouraged to contact or visit any of the University's resources provided below. Getting help is a smart and courageous thing to do – for yourself and for those who care about you.

Counseling Center: 217-333-3704, 610 East John Street Champaign, IL 61820

McKinley Health Center: 217-333-2700, 1109 South Lincoln Avenue, Urbana, Illinois 61801.

The counseling center has resources for all students even if they are not located in Illinois.



CC0 "No Rights Reserved"

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

### **Other Resources**

Students experiencing economic hardships resulting in food insecurity, housing insecurity, homelessness, or other issues that may affect the quality of their work, are encouraged to reach out to iSchool Advisors.

### **Sexual Misconduct Reporting Obligation**

The University of Illinois is committed to combating sexual misconduct. Faculty and staff members are required to report any instances of sexual misconduct to the University's Title IX Office. In turn, an individual with the Title IX Office will provide information about rights and options, including accommodations, support services, the campus disciplinary process, and law enforcement options.

A list of the designated University employees who, as counselors, confidential advisors, and medical professionals, do not have this reporting responsibility and can maintain confidentiality, can be found here: <http://wecare.illinois.edu/resources/students/#confidential>

Other information about resources and reporting is available here: <http://wecare.illinois.edu>

### **Library Resources**

<https://www.library.illinois.edu/infosci/>

### **Writing Resources**

The iSchool Writing Resources is the in-house writing support team for graduate students at the iSchool. They are here to help you with your writing and help you feel more comfortable and confident in your skills. The writing consultants are not professors or evaluators. They simply know the struggles of graduate and undergraduate-level writing and want to help you learn how to succeed and improve your writing skills. The iSchool writing consultants can help you with every step of the writing process. For detailed information on our services please visit our website:

<https://publish.illinois.edu/ischoolwritingresources/>

While we primarily recommend the iSchool Writing Resources, the campus-wide Writers Workshop also provides free consultations. For more information see

<https://writersworkshop.illinois.edu/>



CC0 "No Rights Reserved"

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.



## Programming & Data Wrangling Resources

1. We recommend these Python tutorials.
  - a. Python - W3schools <https://www.w3schools.com/python/default.asp>
  - b. Python - Software Carpentry <http://swcarpentry.github.io/python-novice-gapminder/>
2. For Databases and SQL, we recommend this Software Carpentry resource: <http://swcarpentry.github.io/sql-novice-survey/>
3. For working with APIs, JSON, and XML, we recommend these resources:
  - a. Devlin, J. (2020, August 15). Python API Tutorial: Getting Started with APIs. *Dataquest*. <https://www.dataquest.io/blog/python-api-tutorial/>
  - b. W3Schools. (n.d.-a). *Python JSON*. Retrieved March 14, 2021, from [https://www.w3schools.com/python/python\\_json.asp](https://www.w3schools.com/python/python_json.asp)
  - c. W3Schools. (n.d.-b). *XML Introduction*. Retrieved March 14, 2021, from [https://www.w3schools.com/xml/xml\\_what\\_is.asp](https://www.w3schools.com/xml/xml_what_is.asp)
  - d. W3Schools. (n.d.-c). *XPath Tutorial*. Retrieved March 14, 2021, from [https://www.w3schools.com/xml/xpath\\_intro.asp](https://www.w3schools.com/xml/xpath_intro.asp)



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

## Week-by-Week Topic Schedule

All readings and assignments are subject to change; check the course Canvas site for the most up-to-date information. Readings are in suggested order.

Readings are due before class. Deliverables are due Tuesdays 11:59 PM unless otherwise noted.

### Week 1: Introduction to Bibliometrics and Research Impact (Tuesday January 17<sup>th</sup>)

*Icebreaker discussion post due*

Sugimoto, C. R., & Larivière, V. (2018). Table of Contents & Chapter 1: The basics. In *Measuring Research: What Everyone Needs to Know*. Oxford University Press.

### Week 2: Data and Data Sources (Tuesday January 24<sup>th</sup>)

*Lab Session in Class to prepare for Assignment 1*

Sugimoto, C. R., & Larivière, V. (2018). Chapter 2: The data. In *Measuring Research: What Everyone Needs to Know*. Oxford University Press.

Hutchins, B. I. (2021). A tipping point for open citation data. *Quantitative Science Studies*, 1–5. [https://doi.org/10.1162/qss\\_c\\_00138](https://doi.org/10.1162/qss_c_00138)

### One of the following four articles, which provide examples analyzing differences in database coverage:

Harzing, A.-W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, 120(1), 341–349. <https://doi.org/10.1007/s11192-019-03114-y>

Liu, W., Huang, M., & Wang, H. (2021). Same journal but different numbers of published records indexed in Scopus and Web of Science Core Collection: Causes, consequences, and solutions. *Scientometrics*, 126(5), 4541–4550. <https://doi.org/10.1007/s11192-021-03934-x>

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1), 871–906. <https://doi.org/10.1007/s11192-020-03690-4>

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 1–22. [https://doi.org/10.1162/qss\\_a\\_00112](https://doi.org/10.1162/qss_a_00112)



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

In class Lab session: Data Collection and Data Wrangling; Importance of Data Decisions; Data Cleaning (Preparation for Assignment 1)

- *Quote of the week:* “Make sure the data you collect and analyse is what you need; don’t use single measures when you can obtain several; always normalise your results against what is the average for that subject area, for that country, for that time period.”  
<https://thebibliomagician.wordpress.com/2016/06/20/event-report-theory-practice-of-bibliometric-analysis-by-charles-oppenheim/>

Week 3: Research Impact Evaluation (Tuesday January 31<sup>st</sup>)

*Final Assignment Milestone 1 due – Initial proposal*

Sugimoto, C. R., & Larivière, V. (2018). Chapter 3: The indicators. In *Measuring Research: What Everyone Needs to Know*. Oxford University Press.

Week 4: Altmetrics (Tuesday February 7<sup>th</sup>)

*Assignment 1 due – Data wrangling and data manipulation*

*Final Assignment Milestone 2 due for some – Office hours appointment/proposal feedback*

*Guest lecture – Professor Mike Thelwall, University of Wolverhampton, UK*

*altmetrics: A manifesto – altmetrics.org.* (n.d.). Retrieved March 21, 2021, from

<http://altmetrics.org/manifesto/>

Thelwall, M. (2020). The pros and cons of the use of altmetrics in research assessment. *Scholarly Assessment Reports*, 2(1), 2. <https://doi.org/10.29024/sar.10>

Week 5: Building Research Profiles (Tuesday February 14<sup>th</sup>)

*Final Assignment Milestone 2 due for some – Office hours appointment/proposal feedback*

*Lab Session in Class to prepare for Assignment 2*

Glänzel, W., & Abdulhayoğlu, M. A. (2018). Garfield number: On some characteristics of Eugene Garfield’s first and second order co-authorship networks. *Scientometrics*, 114(2), 533–544. <https://doi.org/10.1007/s11192-017-2623-4>

van Leeuwen, T. N., van Wijk, E., & Wouters, P. F. (2016). Bibliometric analysis of output and impact based on CRIS data: A case study on the registered output of a Dutch university. *Scientometrics*, 106(1), 1–16. <https://doi.org/10.1007/s11192-015-1788-y>

### **Disciplinary differences (read 1):**

Colavizza, G., & Romanello, M. (2019). Citation mining of humanities journals: The progress to date and the challenges ahead. *Journal of European Periodical Studies*, 4(1), 36–53.

<https://doi.org/10.21825/jeps.v4i1.10120>

Melchiorson, P. M. (2018). Bibliometric differences – a case study in bibliometric evaluation across SSH and STEM. *Journal of Documentation*, 75(2), 366–378.

<https://doi.org/10.1108/JD-07-2018-0108>

In class Lab session: (Preparation for Assignment 2)



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

Week 6: Ethics and Responsible Conduct in Research Assessment (Tuesday February 21<sup>st</sup>)  
*Final Assignment Milestone 2 due for everyone – Office hours appointment/proposal feedback*

Sugimoto, C. R., & Larivière, V. (2018). Chapter 4: The big picture. In *Measuring Research: What Everyone Needs to Know*. Oxford University Press.

American Association of University Professors. 2016. “Statement on ‘Academic Analytics’ and Research Metrics.” March 22: [https://www.aaup.org/file/AcademicAnalytics\\_statement.pdf](https://www.aaup.org/file/AcademicAnalytics_statement.pdf)

Gadd, E. (2019). Influencing the changing world of research evaluation. *Insights*, 32(1), 6. <https://doi.org/10.1629/uksg.456>

Week 7: Research Portfolios and Policy Making (Tuesday February 28<sup>th</sup>)

*Assignment 2 due – Building research profiles*

*Start work on research fronts discussion post (due Tuesday March 21<sup>st</sup>)*

### Research portfolio analysis (read both)

Linton, J., & Vonortas, N. (2015). From research project to research portfolio: Meeting scale and complexity. *Форсайт*, 9(2), 38–43. <https://cyberleninka.ru/article/n/from-research-project-to-research-portfolio-meeting-scale-and-complexity>

Wallace, M. L., & Rafols, I. (2015). Research portfolio analysis in science policy: Moving from financial returns to societal benefits. *Minerva*, 53(2), 89–115. <https://doi.org/10.1007/s11024-015-9271-8>

### Read 1 of these papers on applications of research portfolio analysis:

den Houting, J., & Pellicano, E. (2019). A portfolio analysis of autism research funding in Australia, 2008–2017. *Journal of Autism and Developmental Disorders*, 49(11), 4400–4408. <https://doi.org/10.1007/s10803-019-04155-1>

Friedrich, D., & Luible, A. (2016). Supporting the development process for building products by the use of research portfolio analysis: A case study for wood plastics composite materials. *Case Studies in Construction Materials*, 4, 49–54. <https://doi.org/10.1016/j.cscm.2015.12.003>

Liggins, C., Pryor, L., & Bernard, M. A. (2010). Challenges and opportunities in advancing models of care for older adults: An assessment of the National Institute on Aging Research portfolio. *Journal of the American Geriatrics Society*, 58(12), 2345–2349. <https://doi.org/10.1111/j.1532-5415.2010.03157.x>

Kogler, D. F., Heimeriks, G., & Leydesdorff, L. (2018). Patent portfolio analysis of cities: Statistics and maps of technological inventiveness. *European Planning Studies*, 26(11), 2256–2278. <https://doi.org/10.1080/09654313.2018.1530147>

Week 8: Project Proposal Presentations and Feedback (Tuesday March 7<sup>th</sup>)

*Final Assignment Milestone 3 due in class – Group or individual presentation of your project proposal*

Week 9: Spring break (no class meeting)



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

Week 10: Careers, Competencies, and Institutional Structures for Research Impact Assessment  
(Tuesday March 21<sup>th</sup>)

*Final Assignment Milestone 5 due – Description of the initial approach, initial data visualization*  
*Guest panel discussion*

**No required reading this week - but here are some categorized suggestions!**

Research impact/bibliometrics librarianship:

*Day-in-the-life of a research impact/bibliometrics librarian*

(2017, June 5). In a day of a Bibliomagician (Shannon Gordon). *The Bibliomagician*.

<https://thebibliomagician.wordpress.com/2017/06/05/in-a-day-of-a-bibliomagician-shannon-gordon/>

*Competencies for bibliometrics in libraries*

Cox, A., Gadd, E., Petersohn, S., & Scaffi, L. (2019). Competencies for bibliometrics. *Journal of Librarianship and Information Science*, 51(3), 746–762.

<https://doi.org/10.1177/0961000617728111>

NIH Library. (2021, January 4). *BibSymp20 Panel: Roles for Librarians in Research Impact Services*. [https://www.youtube.com/watch?v=O0AQ1ejtkN4&ab\\_channel=NIHLibrary](https://www.youtube.com/watch?v=O0AQ1ejtkN4&ab_channel=NIHLibrary)

*Competency model for bibliometric work (version 1.1; July 2017)*. (n.d.). Retrieved April 17, 2021, from <https://thebibliomagician.files.wordpress.com/2017/10/competency-model-frame-work-design.pdf>

Petersohn, S. (2020). The competent bibliometrician – a guided tour through the scholarly and practitioner literature. In *Handbook Bibliometrics* (pp. 485–496). De Gruyter Saur. <https://doi.org/10.1515/9783110646610-044>

Example output from institutions:

Youtie, J., Porter, A. L., Shapira, P., & Newman, N. (2018). Lessons from 10 years of nanotechnology bibliometric analysis. In *Nanotechnology Environmental Health and Safety* (pp. 11–31). Elsevier. <https://doi.org/10.1016/B978-0-12-813588-4.00002-6>

Paige, S. B., Bourcier, E., Cahill, C., Hsu, C., & Kabel, C. (2012). Evaluating the Kaiser Permanente Community Fund’s social determinants of health portfolio. *The Foundation Review*, 4(1), 68–80. <https://doi.org/10.4087/FOUNDATIONREVIEW-D-11-00030>

Bibliometrics services for and at institutions (read 1) from any of the below subcategories:

*Service development - research evaluation & bibliometrics support services*

NIH Library: Belter, C. W. (2018). 4 - Providing meaningful information: Part B—Bibliometric analysis. In A. P. DeRosa (Ed.), *A Practical Guide for Informationists* (pp. 33–47). Chandos Publishing. <https://doi.org/10.1016/B978-0-08-102017-3.00004-8>

Survey of 140 libraries in Australia, New Zealand, Ireland, and the United Kingdom: Corral, S., Kennan, M. A., & Afzal, W. (2013). Bibliometrics and research data management services: Emerging trends in library support for research. *Library Trends*, 61(3), 636–674. <https://doi.org/10.1353/lib.2013.0005>



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). *Research Impact and Bibliometrics through Data Science [Syllabus]*. School of Information Sciences, University of Illinois Urbana-Champaign.

- Biomedical libraries: Gutzman, K. E., Bales, M. E., Belter, C. W., Chambers, T., Chan, L., Holmes, K. L., Lu, Y.-L., Palmer, L. A., Reznik-Zellen, R. C., Sarli, C. C., Suiter, A. M., & Wheeler, T. R. (2018). Research evaluation support services in biomedical libraries. *Journal of the Medical Library Association*, 106(1), 1–14. <https://doi.org/10.5195/jmla.2018.205>
- Technical University of Munich: Leiss, C., & Gregory, K. (2016). *Visibility and impact of research: Bibliometric services for university management and academic staff*. Proceedings of the IATUL Conferences. Paper 3. <http://docs.lib.purdue.edu/iatul/2016/plenary/3>
- Sample Services - Bibliographic big data*
- Mabry, P. L., Yan, X., Pentchev, V., Van Rennes, R., McGavin, S. H., & Wittenberg, J. V. (2020). CADRE: A collaborative, cloud-based solution for big bibliographic data research in academic libraries. *Frontiers in Big Data*, 3. <https://doi.org/10.3389/fdata.2020.556282>
- Sample Services - Collection development*
- Belter, C. W., & Kaske, N. K. (2016). Using bibliometrics to demonstrate the value of library journal collections. *College and Research Libraries*, 77(4), 410. <https://doi.org/10.5860/crl.77.4.410>
- Martindale, T. (2020). More than collection development: Using local citation analysis to begin a career in business librarianship. *Collection Management*, 45(4), 321–334. <https://doi.org/10.1080/01462679.2020.1715315>
- Sample Services - Reference*
- Williams, B. (2020). Dimensions & VOSViewer bibliometrics in the reference interview. *The Code4Lib Journal*, 47. <https://journal.code4lib.org/articles/14964>
- Sample Services - Research Assessment*
- MacColl, J. (2010). Library roles in university research assessment. *LIBER Quarterly*, 20(2), 152–168. <https://doi.org/10.18352/lq.7984>

Week 11: Research Fronts (Tuesday March 28<sup>th</sup>)

*Final Assignment Milestone 5 due – Description of the initial approach, initial data visualization*

**Read one methodology paper (below):**

- Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12), 2389–2404. <https://doi.org/10.1002/asi.21419>
- Chen, C., & Song, M. (2017). Science mapping tools and applications. In C. Chen & M. Song, *Representing Scientific Knowledge* (pp. 57–137). Springer International Publishing. [https://doi.org/10.1007/978-3-319-62543-0\\_3](https://doi.org/10.1007/978-3-319-62543-0_3)
- Chen, C., & Song, M. (2019). Visualizing a field of research: A methodology of systematic scientometric reviews. *PLOS ONE*, 14(10), e0223994. <https://doi.org/10.1371/journal.pone.0223994>
- Rotolo, D., Rafols, I., Hopkins, M. M., & Leydesdorff, L. (2017). Strategic intelligence on emerging technologies: Scientometric overlay mapping. *Journal of the Association for Information Science and Technology*, 68(1), 214–233. <https://doi.org/10.1002/asi.23631>



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.



Wang, Z., Porter, A. L., Kwon, S., Youtie, J., Shapira, P., Carley, S. F., & Liu, X. (2019). Updating a search strategy to track emerging nanotechnologies. *Journal of Nanoparticle Research*, 21(9), 199. <https://doi.org/10.1007/s11051-019-4627-x>

Week 12: Topic Modeling (Tuesday April 4<sup>th</sup>)

*Final Assignment Milestone 6 due – Refined version of the initial approach*

*Lab Session in Class*

Chen, E. (2011). Introduction to Latent Dirichlet Allocation. Retrieved March 20, 2023, from <http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/>

The two images missing in the current web article are attached below (retrieved from web.archive.org).

*Trig Email image:*

Hello Governor Palin, Our **family** wanted to congratulate **you** and your **family** on the **birth** of your **son**, **Trig**. Our fourth **child**, Daniel, was **born** with **Down Syndrome**, and we can't imagine our **family** without him. Recently, I met a mom with a 34-year-old **daughter** with DS and she said it best: "Don't **you** feel like you've been chosen to be a member of a **very special** club?" **God** bless your **family**, what a **beautiful** example of **love** you are to all who see you! the Paul & Tricia Pietig **family**, Des Moines, Iowa

*Wildlife-Presidency Email image:*

We understand that **you** have been discussed as a possible choice for the **Vice Presidency**.

As **people** who **support** the democratic process and care about protecting our **wildlife** for future generations, we want **you** to know that we don't believe **people** in our states would vote for **you** for any office if they knew your record on these issues.

It is troubling that **you** are **now** working to deny more than 50,000 Alaskans a vote on **aerial** killing of **wolves** and **bears** with legislation now **being** considered in the Alaska legislature.

In class Lab session: Topic Modeling using jsLDA



CC0 "No Rights Reserved"

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

### Week 13: Peer Review (Tuesday April 11<sup>th</sup>)

*Final Assignment Milestone 7 due – Analysis/explanation of results*

Luo, J., Feliciani, T., Reinhart, M., Hartstein, J., Das, V., Alabi, O., & Shankar, K. (2021).

Analyzing sentiments in peer review reports: Evidence from two science funding agencies.

*Quantitative Science Studies*, 2(4), 1271–1295. [https://doi.org/10.1162/qss\\_a\\_00156](https://doi.org/10.1162/qss_a_00156)

Matsui, A., Chen, E., Wang, Y., & Ferrara, E. (2021). The impact of peer review on the contribution potential of scientific papers. *PeerJ*, 9, e11999.

<https://doi.org/10.7717/peerj.11999>

Bornmann, L. (2015). Interrater reliability and convergent validity of F1000Prime peer review.

*Journal of the Association for Information Science and Technology*, 66(12), 2415–2426.

<https://doi.org/10.1002/asi.23334>

### Week 14: Intermediate Project Presentations Part I (Tuesday April 18<sup>th</sup>)

*Final Assignment Milestone 8 due for all – Draft report and visualizations*

*Final Assignment Milestone 9 due **in class** this week or next– Project presentation & request for class feedback*

### Week 15: Intermediate Project Presentations Part II (Tuesday April 25<sup>th</sup>)

*Final Assignment Milestone 9 due **in class** for those who didn't present last week – Project presentation & request for class feedback*

### Week 16: Looking to the Future (Tuesday May 2<sup>nd</sup>)

(No additional readings this week; bring your favorite 1-2 readings to class for discussion)

### Week 17: Finals Week (no class meeting) - Final Report due Tuesday May 9<sup>th</sup> 11:59 PM



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

## Week-by-Week Reading Schedule

Week	# items to read	What to read
Week 1: Introduction to Bibliometrics and Research Impact	1	1. Sugimoto, C. R., & Larivière, V. (2018). Table of Contents & Chapter 1
Week 2: Data and Data Sources	3	1. Sugimoto, C. R., & Larivière, V. (2018). Chapter 2 2. Hutchins, B. I. (2021) 3. One of the articles analyzing <b>differences in database coverage</b>
Week 3: Research Impact Evaluation	1	1. Sugimoto, C. R., & Larivière, V. (2018). Chapter 3
Week 4: Altmetrics	2	1. altmetrics: A manifesto 2. Thelwall, M. (2020)
Week 5: Building Research Profiles	3	1. Glänzel, W., & Abdulhayoğlu, M. A. (2018) 2. van Leeuwen, T. N., van Wijk, E., & Wouters, P. F. (2016) 3. One of the articles in the <b>disciplinary differences</b>
Week 6: Ethics and Responsible Conduct in Research Assessment	3	1. Sugimoto, C. R., & Larivière, V. (2018). Chapter 4 2. American Association of University Professors. 2016 3. Gadd, E. (2019)
Week 7: Research Portfolios and Policy Making	3	1. Linton, J., & Vonortas, N. (2015) 2. Wallace, M. L., & Rafols, I. (2015) 3. One of the articles in the <b>applications of research portfolio analysis</b>
Week 8: Project Proposal Presentations and Feedback	0	No readings
Week 9: Spring break	0	No class and no readings
Week 10: Careers, Competencies, and Institutional Structures for Research Impact Assessment	0	No required reading this week - but feel free to pick something from the following categories: 1. <b>Research impact/bibliometrics librarianship</b> category 2. <b>Example output from institutions</b> category 3. <b>Bibliometrics services for and at institutions</b> category
Week 11: Research Fronts	1	1. One of the <b>methodology papers</b>
Week 12: Topic Modeling	1	1. Chen, E. (2011)
Week 13: Peer Review	3	1. Luo, J., Feliciani, T., Reinhart, M., Hartstein, J., Das, V., Alabi, O., & Shankar, K. (2021) 2. Matsui, A., Chen, E., Wang, Y., & Ferrara, E. (2021) 3. Bornmann, L. (2015)
Week 14: Intermediate Project Presentations Part I	0	No readings
Week 15: Intermediate Project Presentations Part II	0	No readings
Week 16: Looking to the Future	0	No readings - but bring your favorite 1-2 readings to class



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

### *Assignment Details*

**Reminder: Contact the instructors if you need an extension, PRIOR to the due date.**

<b>Deliverable</b>	<b>Deadline(s)</b>	<b>Contribution to Course Grade</b>
Final Assignment (individual or group project)	Weeks 3-17	50%
Assignment 1 – Data wrangling and data manipulation	Week 4: February 7	15%
Assignment 2 – Building research profiles	Week 7: February 28	15%
Participation and Instructor Assessment	All semester	20%

**Final Assignment – Individual or Group Project (*Milestones due Weeks 3-17 – 50% of grade*) – due Tuesdays at 11:59 PM unless otherwise noted**

Milestone 1, due Jan 31<sup>st</sup> (Week 3) : Initial proposal for an individual or group project of your choice, relating to research impact assessment, bibliometrics, or other science of science related topics.

Milestone 2 January 31<sup>st</sup> to Feb 21<sup>st</sup> (Weeks 3-5): Office hours appointment/proposal feedback

Milestone 3, due in class March 7<sup>th</sup> (Week 8): Proposal presentation in class

*Week 9: Spring break*

Milestone 4, due March 21<sup>st</sup>(Week 10): Data source

Milestone 5, due March 28<sup>th</sup> (Week 11): Initial approach/visualization

Milestone 6, due March 28<sup>th</sup> (Week 12): Refine initial approach

Milestone 7, due April 4<sup>th</sup> (Week 13): Analysis/explanation of results

Milestone 8, due April 18<sup>th</sup> (Week 14): Draft report and visualizations

Milestone 9, due in class April 18<sup>th</sup> or 25<sup>th</sup> (Week 14 or 15): Intermediate presentation in class

*Week 16: Last day of class*

Milestone 10, due May 9<sup>th</sup> (Week 17, Finals Week): Final Report



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

***Participation and Instructor Assessment (All Semester – 20% of grade) – all semester***

This grade is a combination of your attendance, participation, and overall progress throughout the semester; all assessed by your instructor.

Participation involves the following activities:

- Engaging with the material in readings, labs, and lectures
- Asking questions and raising discussion points during class
- Asking questions in the Open Discussion in Canvas, or answering classmates' questions
- Communicating with the instructors in office hours

Ungraded activities fall into this category, including items with specific deadlines:

- Readings due before class most weeks
- Posting in the icebreaker activity discussion (Week 1 – due Tuesday January 17th)
- Take the student survey to help instructors understand your background (Week 1 – due Tuesday January 17th)
- Posting in the investigating a research front discussion (Week 10 – due Tuesday March 21st)

The following rubric will be used to assign a score:

**10**-Student has been an active participant in class discussion, bringing to the class insights from their interpretations of readings and lived experiences and is demonstrating an increasing grasp of the key concepts covered in class.

**8**-Student has been an active participant in some of the class discussion and is demonstrating some gains in grasping key concepts covered in class.

**6**-Student is occasionally active in class and is demonstrating some learning, but it is clear they are not performing to their full capabilities.

**4**-Student has missed several classes and/or is not always active when attending class.

**0**-Student has consistently missed class during the rated period.



CC0 "No Rights Reserved"

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

## Assignment 1—Data wrangling and data manipulation

**Due Tuesday February 7<sup>th</sup> at 11:59 PM (Week 4)**

Points: 100 (15% of course grade)

### Goals

1. Practice filtering, sorting, and grouping data.
2. Practice extracting information from the data.
3. Practice merging data files.
4. Learn or practice data wrangling and manipulation skills, using a Jupyter notebook with the Python Pandas package (encouraged, not mandatory).

### Instructions

This assignment has three parts. Each part contains a set of questions and tasks.

The assignment has two tracks: non-programming and programming. You can choose either track, but you cannot switch between them. In other words, if you choose to go with the non-programming track, you need to complete all the questions and tasks described in the prompts for the non-programming track (and vice versa for choosing the programming track). Extra credit (+10 points) will be awarded to the students who choose the programming track.

### What to submit

Choose one of the following tracks

- Non-programming track: Submit *authority.xlsx* and *icite.xlsx*, with your answers in them.
- Programming track: Submit *assignment1.ipynb* with your answers in it.

### Information provided

1. *authority.xlsx*: This file contains author name instances for 5,000 articles. Each row is an author instance in an article. This file has three columns: *author ID*, *PMID*, and *author name position*. *Author ID* is the unique ID of an author. *PMID* is the unique ID of an article. *Author name position* indicates the author's position on the paper. For example, 1 indicates the author is the first author. Please see data source 1 for further details. **Do not delete or modify any columns, rows, or cell values in the 'authority' tab of this Excel file.**
2. *icite.xlsx*: This file contains data for 5,000 articles. Each row in the file contains the data for an article. Please see data source 3 for column descriptions. **Do not delete or modify any columns, rows, or cell values in the 'icite' tab of this Excel file.**



CC0 "No Rights Reserved"

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.



3. assignment1.ipynb: A Jupyter notebook contains the assignment prompts. **Do not delete or modify any columns, rows, or cell values in the *authority* and *icite* dataframes. Do not delete code and comments in the notebook cells.**
4. Sample code from Lab 1, lab1.ipynb

#### Data sources (for future reference only)

1. Author-ity 2018 - PubMed author name disambiguated dataset  
<https://databank.illinois.edu/datasets/IDB-2273402>
2. iCite API <https://icite.od.nih.gov/api>
3. iCite snapshot Version 34 (2022-09-11)  
[https://nih.figshare.com/collections/iCite Database Snapshots NIH Open Citation Collection /4586573/34](https://nih.figshare.com/collections/iCite_Database_Snapshots_NIH_Open_Citation_Collection_/4586573/34)

#### Step-by-step instructions

##### Part 1: authority.xlsx (30 points)

0. Read the file
  - Non-programming track: Open the *authority.xlsx* file in Excel.
  - Programming track: Read the *authority.xlsx* file as instructed in cell 2 in assignment1.ipynb
1. How many unique author IDs are in the file?
  - Non-programming track:  
Step 1: Copy and paste the author ID values in the *authority* sheet to the *Author ID-deduplicated* column in the *part1-q1* sheet.  
Step 2: De-duplicate the values in the *Author ID-deduplicated* column.  
Step 3: Calculate the number of unique author IDs. Enter the answer into cell B2 under *number\_of\_unique\_authors*.
  - Programming track:  
In cell 3, calculate and display the number of unique author IDs.
2. How many authors are there in each paper?
  - Non-programming track:  
Step 1: Go to the sheet *part1-q2*. There should be two columns (*Author ID* and *PMID*) in the sheet. These columns were copied and pasted from the *authority* sheet.



CC0 “No Rights Reserved”

*Recommended citation*: Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

Step 2: Calculate the number of authors in each paper. This can be achieved by creating a pivot table using the *Author ID* and *PMID* columns like Figure 1. The pivot table should have two columns. The first column will contain the row labels (i.e., the PMIDs and grand total). The second column will contain the values associated with the labels (i.e., the number of authors in the paper). Save this pivot table next to the *PMID* column in the *part1-q2* sheet.

Row Labels	Count of Author ID
8952	3
15616	3
19645	6
32697	3
43783	2
70999	3
79317	3
81362	5
81654	5

Figure 1. First five rows in the pivot table



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

- Programming track:

**Please answer this question in cell 4.**

Step 1: Calculate the number of authors in each paper. This can be achieved by grouping the *authority* dataframe by the *PMID* column and count the number of author IDs associated with each PMID (hint: `groupby()`). Name this dataframe as *author\_counts\_1*.

Step 2: The `groupby()` function sets the column where the grouping is performed (i.e., the *PMID* column) as a dataframe index. Set the PMIDs back to a column (Hint: `reset_index()`).

Step 3: Rename the *Author ID* column in the *author\_counts\_1* dataframe as *author\_counts\_authority*.

Step 4: Sort the *author\_counts\_1* dataframe by the *author\_counts\_authority* values. The sorting should be in descending order.

Step 5: Display the *author\_counts\_1* dataframe.

	PMID	author_counts_authority
<b>2986</b>	17501563	163
<b>3718</b>	21983444	31
<b>4815</b>	29442540	29
<b>2721</b>	15994562	27
<b>3186</b>	18786894	23

Figure 2. First five rows in the *author\_counts\_1* dataframe



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

3. How many papers have more than five authors?

- Non-programming track:

Step 1: Copy the pivot table that you just created in the *part1-q2* sheet.

Step 2: Use "Paste Values" to paste the copied pivot table into the *part1-q3* sheet.

Step 3: Rename the column headers as *PMID* and *author\_counts\_authority*.

Step 4: Delete the (*blank*) and the *Grand Total* rows in the pasted table.

Step 5: Sort the table by the *author\_counts\_authority* values in descending order.

Step 6: Calculate the number of papers having more than five authors. Enter the answer in cell C2 (i.e., the cell under *number\_of\_authors\_per\_paper*).

- Programming track:

In cell 5, calculate the number of papers having more than five authors. This can be achieved by filtering the *author\_counts\_1* dataframe by the condition that the value in the *author\_counts\_authority* column is greater than five. Display the shape of the filtered dataframe.



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

## Part 2: icite.xlsx (50 points)

## 0. Read the file

- Non-programming track: Open the *icite.xlsx* file in Excel.
- Programming track: Read the *icite.xlsx* file as instructed in cell 6 in *assignment1.ipynb*

## 1. How many papers were published between 2000 and 2010?

- Non-programming track:  
Step 1: Go to the *part2-q1* sheet.  
Step 2: Calculate the number of papers published between 2000 and 2010 (i.e.,  $2000 \leq \text{year} \leq 2010$ ) by the *year* column in the *icite* sheet. Enter the answer in cell A2 (i.e., the cell under *number\_of\_papers\_published\_between\_2000\_and\_2010*).
- Programming track:  
**Please answer this question in cell 7.**  
Step 1: Filter the *icite* dataframe by the condition that the *year* values are between 2000 and 2010 (i.e.,  $2000 \leq \text{year} \leq 2010$ ). Name this filtered dataframe as *in\_year\_range*.  
Step 2: Display the shape of the *in\_year\_range* dataframe.

## 2. How many papers are research articles? How many are not?

- Non-programming track:  
Step 1: Copy and paste the *is\_research\_article* column in the *icite* sheet to the *part2-q2* sheet.  
Step 2: Create a pivot table using the *is\_research\_article* column in the *part2-q2* sheet. The pivot table should have two columns. The first column contains the row labels (i.e., Yes, No, and Grand Total). The second column contains the values associated with the labels (i.e., the numbers of articles). Save this pivot table next to the *is\_research\_article* column in the *part1-q2* sheet.

Row Labels	Count of is_research_article
No	714
Yes	4286
(blank)	
<b>Grand Total</b>	<b>5000</b>



CC0 “No Rights Reserved”

*Recommended citation*: Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

Figure 3. Expected result

- Programming track:  
In cell 8, group the *icite* dataframe by the *is\_research\_article* column and count the PMIDs. Display the grouping result.

		PMID
<b>is_research_article</b>		
<b>No</b>		714
<b>Yes</b>		4286

Figure 4. Expected result

### 3. How many papers have a DOI?

- Non-programming track:  
Step 1: Go to the *part2-q3* sheet.  
Step 2: Calculate the number of papers having a DOI using the *doi* column in the *icite* sheet. Note that a DOI prefix always starts with “10.”, and the *doi* column contains bad values. To avoid counting bad values, count the number of papers having a DOI by counting the number of cells in the *doi* column that have values starting with “10.”. Enter the answer in cell A2 (i.e., the cell under *number\_of\_papers\_having\_DOI*).
- Programming track:  
**Please answer this question in cell 9.**  
Step 1: Calculate the number of papers having a DOI using the *doi* column in the *icite* dataframe. Note that a DOI prefix always starts with “10.”, and the *doi* column contains bad values. Filter the *icite* dataframe by the conditions that DOI values are starting with **10.** and are not NaNs. Name this filtered dataframe as *papers\_with\_doi*.  
Step 2: Display the shape of the *papers\_with\_doi* dataframe.



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.



4. What are the minimum, maximum, mean, and median of the citation counts?
  - Non-programming track:  
 Step 1: Go to the *part2-q4* sheet.  
 Step 2: Count the minimum, maximum, mean, and median of the citation counts. Enter the answers in the cells next to the *minimum*, *maximum*, *mean*, and *median* cells.
  - Programming track:  
 In cell 10, count the minimum, maximum, mean, and median of the citation counts by using `describe()` to get the results. Display the results.
5. How many authors are there in each paper? (Hint: Author names in the *authors* column are separated by “|”.)
  - Non-programming track:  
 Step 1: Go to the *part2-q5* sheet. There should be three columns (*pmid*, *authors*, and *author\_counts\_icite*) in the sheet. The first two columns were copied and pasted from the *icite* sheet. The third column is blank.  
 Step 2: Calculate the number of authors by the author names in the *authors* column. Fill the values in the *author\_counts\_icite* column with the calculated numbers.
  - Programming track:  
 In cell 11, generate a new column, *author\_counts\_icite*, in the *icite* dataframe. The *author\_counts\_icite* should contain the numbers of authors calculated from the *authors* column.



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

## Part 3: Merging data from different sources (20 points)

## 0. Prepare a table for merging

- Non-programming track:

Step 1: In the *icite.xlsx* file, copy the *pmid* and *author\_counts\_icite* columns from the *part2-q5* sheet to the *part3* sheet.

- Programming track:

In cell 12, copy the *PMID* and *author\_counts\_icite* columns in the *icite* dataframe as a new dataframe. Name this new dataframe *icite\_selected*.

1. Merge the prepared table with another table (non-programming track: the table in the *part1-q3* sheet; programming track: the *author\_counts\_1* dataframe) by the following steps:

- Non-programming track:

Step 1: Create a new column, *author\_counts\_authority\_lookup*, in the *part3* sheet.

Step 2: For each row (i.e., for each PMID) in the *author\_counts\_authority\_lookup* column, fill in the value by using VLOOKUP to look up the PMID's *author\_counts\_authority* value in the *part1-q3* sheet.

- Programming track:

**Please answer this question in cell 13.**

Step 1: Merge the *icite\_selected* dataframe with the *author\_counts\_1* dataframe. Name the merged dataframe as *merged\_df*.

Step 2: Rename the *author\_counts\_authority* column as *author\_counts\_authority\_lookup*.

Step 3: Print the shape of the *merged\_df* dataframe. The shape should be (5000, 3).

Step 4: Display the *merged\_df* dataframe.

	PMID	author_counts_icite	author_counts_authority_lookup
0	10009783	1	1
1	10009800	3	3
2	1005199	2	2
3	10066046	3	3
4	10068902	7	7

Figure 5. First five rows in the *merged\_df*



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

2. Subtract *author\_counts\_authority\_lookup* from *author\_counts\_icite*, and put the results into a new column, *author\_counts\_diff*. How many values in the *author\_counts\_diff* column are not equal to zero?

- Non-programming track:

Step 1: Create the *author\_counts\_diff* column in the *part3* sheet.

Step 2: Enter the answer to the question in the cell under *diff\_count*.

- Programming track:

**Please answer this question in cell 14.**

Step 1: Create the *author\_counts\_diff* column in the *merged\_df* dataframe.

Step 2: Display your answer to the question in cell 14



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

## Assignment 2 – Building a research profile

**Due Tuesday February 28<sup>th</sup> at 11:59 PM (Week 7)**

Points: 100 (15% of course grade)

### Goals

1. Build a research profile.
2. Identify possible limitations of the profile and the data source(s).

### Instructions

In this assignment, you will practice building a research profile for an author. In particular, you will (1) collect data for the author, (2) analyze the data for building the profile, and (3) discuss the possible limitations of the profile and the data source(s). You may select any author for the profile. We recommend selecting an author you are familiar with since knowing the author may help you identify potential limitations.

Building a research profile is a complex task. Numerous items that can be included in the analysis, and each item reflects a portion of the author's research activities and/or impact. Here are a few examples:

- The number of publications partially reflects the quantity of research outputs.
- Citations are often used to measure the impact of research, but we know that measuring research impact should not solely depend on citation counts.
- Affiliated institutions are tied to an author's career path. For example, a professor affiliated with an institution is likely to have had a different affiliation when he or she was a Ph.D. student.
- Collaborations (e.g., co-authorship networks) reflect research partnerships.
- Keywords provide information about an author's research topics and fields. Publication venues and types are also field-related. For instance, conferences are popular venues for computer science researchers.
- Grants have an impact on research output and quality.

It is also worth noting that the items listed above may be related to one another (e.g., publication and citation counts) and to *time*. A tenured professor, for example, is likely to have more publications than a young researcher because the tenured professor has worked in the field longer.



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

The profile should be written in organized paragraphs with relevant statistics, tables, or figures. Given the time constraints and complexity of building a profile, the profile does not have to be exhaustive. However, the profile should have at least analyses of publications, citations, research topics, and collaborators. In the profile, you should describe what items (publication count, article distribution by research topics, etc.) are included as well as why they are included (i.e., what information you want to show and why).

**Recommended Data Sources:** WoS and/or Scopus, or another data source of your choice with permission of the instructors.

**Recommended Tools:** R, Python, or end-user tools such as VOSViewer, Publish or Perish, etc.

**What to submit:** Submit a Word or a PDF file.

**Additional Resources (See Canvas for the PDF files)**

- Sample\_AIR\_2018b.pdf
- Sample\_RIR\_2019\_web.pdf

These documents contain examples of analyzing publications, citations, collaborations, and research topics. You are not required to repeat their analysis in your research profile. However, the examples may give you some useful ideas for building a research profile. When you read these resources, we encourage you to think about the following:

- What was evaluated? For each of the evaluated items, how was the result produced?
- What kinds of information (citations, publications, author names, etc.) were collected for the analysis? Where can you get the data if you would like to perform the same analysis? What kinds of data preprocessing may be needed?
- What did the results tell you about the evaluated article (Sample\_AIR\_2018b.pdf) and the researcher (Sample\_RIR\_2019\_web.pdf)? What was not covered in the results?



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.

[Assignment 2 – Building a research profile](#)

### Part 1: Data collection (20 points)

This part should address the following information (in a paragraph or paragraphs):

1. The name of the author
2. The selected data source(s)
3. The date of data collection
4. How the data was collected and prepared for analysis

### Part 2: The research profile (50 points)

The profile should contain the following:

1. The included items (publication count, article distribution by research topics, etc.)
2. The reasons for including these items in the profile
3. Your interpretations of the results. For example, what does article distribution by research topics tell you about the author?

### Part 3: Possible limitations (30 points)

Address the possible limitation(s) of the profile and the data source(s) through answering the following questions:

1. How well do you think the profile represents the author? What kinds of impact cannot be measured from or may be underrepresented in the profile?
2. Do you think the data source(s) of your choice has good coverages of the publication and citation records for your author? Why or why not?
3. What kinds of impact cannot be measured from or may be underrepresented in the publication and citation records?



CC0 “No Rights Reserved”

*Recommended citation:* Tzu-Kun Hsiao and Jodi Schneider. (2023 Spring). Research Impact and Bibliometrics through Data Science [Syllabus]. School of Information Sciences, University of Illinois Urbana-Champaign.