Joining the Data Revolution:
Big Data in Education and Social Science Research
2021: June 1-11 | June 21-July 1 | July 26-August 5
Online Workshop

Course Description

We designed this workshop to help you acquire or expand data analysis skills to support your own efforts to develop and articulate an individual research question and to frame that question in a fashion suitable for you to pursue external funding to support your research. In addition to technical skills, participants will work with established investigators to develop and present ideas for projects that might eventually be submitted to NSF grant competitions. While examples will be drawn from UMETRICS data (and you are welcome to use those data to help develop your research questions), your topics and datasets may focus on other sources. Supported by the NSF as part of its *Harnessing the Data Revolution (HDR) Big Idea* (see https://www.nsf.gov/cise/harnessingdata/), the IRIS workshop is designed to help investigators from a wide range of backgrounds and disciplines acquire the tools and knowledge to secure grant funding for data-driven social science and education research. We are excited to present this workshop through the ICPSR Summer Program for the first time and wish to thank ICPSR for its support.

Course Goals & Guiding Principles

This workshop is funded by the Education and Human Resources (EHR) division of the National Science Foundation (NSF) under its Building Capacity in STEM Education Research (BCSER) program. It is designed to help researchers bring their expertise and experience in educational and social science (ESS) research to bear on questions pertaining to STEM education or related areas using data science tools and methods. We hope to help you increase your ability to define and develop projects that will result in competitive proposals that might be submitted to NSF or other funders. We also hope to help support a community of researchers who can share information, tools, and insights to help strengthen research and teaching capabilities involving large-scale data analysis in ESS fields.

Workshop Themes:

- Data exploration
- Data visualization
- Data linkage
- Basic data analysis
- Grant proposal writing

We believe that:

- Data science tools and methods are best learned through concrete, hands-on work to address real research questions with real data.
- You bring essential expertise and knowledge to this workshop. We have as much to learn from you as you do from us.
- Effective classroom support is necessary to help you over the technical challenges that inevitably accompany new tools.
- All research is, at one level or another, collaborative and is best accomplished within an engaged intellectual community.
- Data science teaching and learning, as well as the datasets and tools we develop and share, must be responsive to the needs and interests of diverse research communities.
- Data science skill consists primarily in figuring out how to make creative use of imperfect data that often comes from multiple sources to address important questions.

Course Structure

In this hands-on workshop, participants will work alone and in teams using large-scale datasets with the goal of achieving a better understanding of the research questions that can be answered with big data. During technical sessions, we will focus on working through well-documented examples of Python code in Jupyter Notebooks designed to address a shared research question using UMETRICS – Universities Measuring the EffecTs of Research on Innovation, Competitiveness, and Science – data.

Each day, we introduce the participants to the basics of data exploration and linkage techniques using Python and Structured Query Language (SQL) (e.g., select and join queries, pandas merges) along with tutorials on data visualizations using Python packages including Seaborn that enable you to create effective, attractive figures (such as histograms, scatter plots, box plots, etc.).

Workshop discussions, exercises, and pre-work materials will focus on examining whether and how different types of diversity in scientific research teams influence the amount, character, and impact of research produced by those teams. To this end, we have shared a set of orienting questions and associated readings to help you prepare for our collective work.

The guided, step-by-step analytic work we do together will help you develop skills in working with real administrative data in a privacy-protected research environment. More substantive discussions during introductory lectures, journal club, and other discussions will focus on literature and general approaches to questions that you will help define.

Class instructions and discussions will build on recommended pre-course materials, so participants are encouraged to read and watch all of them prior to the first day of class.

Workshop Activities

Journal Club

On about day 3 we will host a journal club. Journal clubs are a common practice in labs and research teams that share a mutual focus on a general topic, dataset, or type of analysis. The goal of journal club discussions is to dive deeply into a recent paper on a topic of shared interest with an eye toward jointly accomplishing three things: (1) clearly articulating what a paper finds and what it claims those findings mean; (2) closely examining the paper's data and methods to determine as precisely as possible how those findings were reached and justified; and (3) discussing how the data and tools we are working with might be used to challenge, expand, or replicate the findings presented in the paper we are considering. While critique is an important part of journal clubs, our primary goal will be to collectively work to understand the "what" and "how" of the paper and to frame our critiques in terms of empirical research we might do that builds on or contradicts the paper.

The journal club will be focused on a recent paper selected by the IRIS team related to our core topic of diversity in scientific teams. It will begin with a brief (no more than five minute) summary of the paper followed by guided group discussion to address each of the three goals of the session.

Small Group Work

The participants in this workshop all bring significant expertise and a wide range of experience with this kind of data analysis and with Python. We believe that one of the best ways to develop new skills is to actively apply them to answer a question you devise with support from your peers and from the instructional team. To accomplish this, we assign you to groups that will meet together on most workshop days to work on expanding or adapting the Jupyter

Notebooks we use to introduce key concepts to answer a new question pertaining to the workshop's orienting theme.

This does not need to be a complete and airtight analysis. Instead, we ask you to work together to define a question you think can be answered with the data we have produced for you and work together to develop one plausibly true finding (descriptive findings are just fine) and a means to effectively present it (such as a data visualization or figure).

Each group will be assigned a specific Teaching Assistant who can help you troubleshoot code and provide support as you learn (or expand your knowledge of) this new programming language.

Bump Meeting

As a means to help all participants learn from the work you do in your small groups, we borrow another meeting format common to large research groups. "Bump meetings" are designed to help individuals or groups develop, troubleshoot, or otherwise improve analyses by sharing and collectively discussing work in progress. On day 8 we will ask each group to bring forward something cool, unexpected, surprising, or difficult from their collective work. This could be a preliminary finding, a persistent question, a draft figure, a piece of code that works particularly well (or that you are having trouble getting to work) or anything else that you think merits broader discussion. Each group should nominate a participant to briefly (in 2-3 minutes) present their group's "bump in the road" which will then be discussed by the group as a whole to help with troubleshooting, honing, or otherwise improving the thing you wanted to discuss. This meeting format offers an opportunity for teams to become a bit more familiar with each other's questions, to see and react to other examples of analyses, and to share information and lessons learned across groups.

Open-Ended Discussion

We often find that learning new methods and using new tools creates more questions than it answers. In order to surface and do what we can to address those questions, we have two "open-ended" discussion sessions during which the teaching team has no particular plans or materials to share. The "Q&A with Jason" session on day 6 offers participants an opportunity to bring their questions about anything related to large-scale data analysis, workshop topics, grant-writing, or any other topic you find relevant to Jason for discussion with him and with your colleagues. The range of topics is meant to be broad and to be driven by your interests and concerns. On day 8 we offer a final opportunity for open-ended discussion of the workshop to help participants "wrap up" their experience of the workshop and to make sure that we do what we can to address new or lingering questions you might have.

Individual Work and Lightning Talks

One of the goals of this workshop (and of the NSF program that supports it) is to help prepare investigators to use new data and tools in projects that might be successfully submitted to relevant funders. The particular interest of our funder, EHR, is in STEM education research. Over the course of the workshop, we ask each participant to think about and work to develop a "lightning talk" presentation whose format is based on the key components of an NSF project summary. This work will largely occur outside of scheduled workshop time and will culminate in short (3 minute, 3 slide) presentations of your project idea to an expert panel who will offer you feedback. Your project idea may use UMETRICS data but does not have to. The goal is to have you think about and work through ways that the things you learn in this workshop might support your pursuit of funding to support your work. Lightning talks will take place on day 9. We will share a more detailed description of what those talks will entail early in the workshop. Day 8 will leave time for optional "open office hour" discussions of your work. An open office hour is one where other participants may be present to hear your questions and our answers, though unlike the open-ended sessions described above there is no expectation of group discussion.

Recommended Pre-Workshop Learning Resources

We're providing suggestions for a variety of Python and SQL learning resources for the 2021 Joining the Data Revolution workshop. Formats include text or e-books, videos, and interactive coding. We find that practice is ultimately what solidifies these skills, so keep that in mind if choosing a solely book or video-based learning strategy. Where possible we've provided rough estimates of time and whether or not additional cost is involved.

Recommended: If you're brand new to Python we suggest focusing on several of the recommended resources, choosing content to complement your existing skill set. These choices include W3Schools web-based examples, Coleridge Initiative: Applied Data Analytics introductory/exploration/statistics videos, or DataCamp courses which provide opportunities for practicing code. Secondarily, some exposure to the Pandas library and Jupyter Notebooks would be highly beneficial. Additional reinforcement in Python and SQL can be obtained through any of the optional other learning resources.

- Open Source Resources (no cost)
    - W3Schools
        - These tutorials are self-paced and involve a mix of reading as well as opportunities to practice coding.
        - **Recommended:** Basic Python Tutorial
            - Stop after reaching *Python String Formatting*

- **Recommended:** Pandas Tutorial
  - Complete as much as possible
- Optional: Basic SQL Tutorial
  - Stop after reaching *SQL Operators*
- Time estimate: approximately 2 hours for each section above, depending on time spent completing exercises

- Coleridge Initiative: Applied Data Analytics
  - **Recommended:**
    - Introduction to Jupyter Notebooks
    - Data Exploration in Python Pandas
    - Datasets and Variables
    - Sampling
    - Descriptive Statistics and Graphs for Numerical Data
    - Sampling Distribution
    - Inference
    - Bootstrapping
    - Video format
    - Time estimate: about 30 minutes altogether
  - Optional:
    - ADA Course Jupyter Notebooks
    - Time estimate: 10 hours+
- Jupyter Notebook Tutorial
  - **Recommended:** Dataquest Jupyter Notebook Tutorial
  - Time estimate: 1.5 hours
- Python Tutorials (many options exist on various YouTube channels)
  - Optional: Python Tutorials for Beginners
  - These video tutorials are self-paced, but we don't suggest watching every video on the playlist. Instead, stop after the 9th video. Note that you can skip ahead to relevant timestamps and watch videos at higher speeds based on your comfort level. You would need to install Python and supporting software in order to follow along with the coding, but the first video walks you through this process.
  - Time estimate: 2-3 hours
- Python, SQL & Data Science MOOCs on Coursera, edX, Michigan Online, etc.
  - Optional: Python for Everybody (specialization consisting of 5 courses)
  - Format: mix of video lectures, readings, and programming assignments.

    - Time estimate: these courses typically span 4 weeks based on time commitment of 5-10 hours per week, though it often is possible to complete them more quickly at your own pace.

- Paid Resources (some cost required for full access or services)
    - DataCamp
        - Cost note: the first chapter of introductory/intermediate courses is free and you often get a promotional offer for a reduced subscription after completing it. You may be able to complete the full introductory course on a mobile device for free.
        - Each course is a mix of short video lectures and guided interactive coding practice.
        - **Recommended:**
            - Introduction to Python for Data Science
            - Intermediate Python
            - Pandas Foundations
        - Optional:
            - Dozens of other courses in Python, SQL, Data Science, etc.
        - Time estimate: 3-5 hours per course
        - Optional alternative: close equivalents to many of these courses are available at Dataquest.

Schedule

1-5 PM

Day	Topic	Speaker(s)
1 (Tues., June 1)	Introduction to data, guiding questions Logic of big, administrative data (1 - 2:30 p.m.) Hero's Journey (2:45 - 3:45 p.m.) Group work – logging in to & navigating the VDE (3:45 - 5 p.m.)	Jason Owen-Smith Nazha Gali Teaching Assistants
2 (Wed., June 2)	Day 1 Recap (1 - 1:10 p.m.) The rules of the road – IRIS policies & procedures for access, disclosure review (1:10 - 1:40 p.m.) Data exploration introduction, Jupyter notebook walk-through (1:50 - 3:40 p.m.) Small group work (3:50 - 5 p.m.)	Jason Owen-Smith Natsuko Nicholls Robert Truex Teaching Assistants
3 (Thurs., June 3)	Day 2 Recap (1 - 1:10 p.m.) Journal club (1:10 - 2:40 p.m.) Hofstra, B., Kulkarni, V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., McFarland, D. (2020). The diversity-innovation paradox in science. *Proceedings of the National Academy of Sciences*, *117*(17), 9284-9291. https://doi.org/10.1073/pnas.1915378117 Data exploration part 2 (2:50 - 3:50 p.m.) Small group work (4 - 5 p.m.)	Jason Owen-Smith Jason Owen-Smith, Teaching Assistants Robert Truex Teaching Assistants

Day	Topic	Speaker(s)
4 (Fri., June 4)	Day 3 Recap (1 - 1:10 p.m.) Data visualization (1:10 - 2:40 p.m.) Small group work (2:50 - 4:30 p.m.) Week 1 wrap-up (4:40 - 5 p.m.)	Jason Owen-Smith Christopher Brown Teaching Assistants Jason Owen-Smith
5 (Mon., June 7)	Plan for the week (1 - 1:10 p.m.) Record linkage & disambiguation (1:10 - 2:10 p.m.) Data linkage (2:25 - 4:15 p.m.) Small group work (4:30 - 5 p.m.)	Jason Owen-Smith Jinseok Kim Raphael Ku Teaching Assistants
6 (Tues., June 8)	Day 5 Recap (1 - 1:10 p.m.) Model selection (1:10 - 2:40 p.m.) Data analysis (2:50 - 3:50 p.m.) Small group work (4 - 5 p.m.)	Jason Owen-Smith Jason Owen-Smith Matt VanEseltine Teaching Assistants
7 (Wed., June 9)	Day 6 Recap (1 - 1:10 p.m.) Weaving it all together – elements of a successful grant proposal (1:10 - 2:40 p.m.) Q&A with Jason (2:50 - 3:20 p.m.) Small group work (3:35 - 5 p.m.)	Jason Owen-Smith Jason Owen-Smith Jason Owen-Smith Teaching Assistants

Day	Topic	Speaker(s)
8 (Thurs., June 10)	Wrap-up: Questions, discussion of workshop topics (1 - 2 p.m.) "Bump" Meeting (2:15 - 3:45 p.m.) Lightning talk open office hours (4 - 5 p.m.)	Jason Owen-Smith Jason Owen-Smith, Teaching Assistants Teaching Assistants
9 (Fri., June 11)	Lightning presentations and feedback (1 - 5 p.m.)	Participants, Panel

Beginning on Day 3, Teaching Assistants will hold open office hours at 12:00-1:00 pm EDT

Teaching and Support Team

Lead Instructor/Project Director:

Jason Owen-Smith (IRIS Executive Director, Professor of Sociology)

Confirmed Speakers/Instructors:

Christopher Brown (IRIS Data Support Specialist)

Jinseok Kim (IRIS Assistant Research Professor)

Raphael Ku (IRIS Data Support Specialist)

Natsuko Nicholls

Robert Truex (IRIS Data Manager)

Matthew VanEseltine (IRIS Research Investigator)

Additional speakers TBD

Workshop Teaching Assistants:

Christopher Brown (IRIS Data Support Specialist)

Miles Butler (IRIS Teaching Assistant)

Raphael Ku (IRIS Data Support Specialist)

James Tang (IRIS Teaching Assistant)

Workshop Program Coordinators:

Nancy Calvin-Naylor (IRIS Managing Director)

Natsuko Nicholls (IRIS Research Manager)

Dan Meisler (IRIS Communication Specialist)

Research & Data Support:

Natsuko Nicholls (IRIS Research Manager)

Christopher Brown (IRIS Data Support Specialist)

Panel:

To be determined

Workshop Venue

We will meet each weekday using Zoom, beginning at 1 pm Eastern. Zoom login details will be forthcoming.

The total time commitment each day for presentations, team activities, and individual work should be approximately 4 hours. Course materials can be found on Canvas. Details about accessing the Canvas platform will be provided separately but note that you must use the "umich" email address provided by the ICPSR Summer Program to log in. The Canvas workshop site will be available for 2 weeks following the conclusion of the workshop.

Recommended Readings

Adil Yalcin, M., & Plaisant, C. (2016). Information visualization. In I. Foster, R. Ghani, R.S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big Data and Social Science* (Ch. 6). CRC Press. ISBN: 9781498751407. https://textbook.coleridgeinitiative.org/

Bender, S., Jarmin, R.S., Kreuter, F., & Lane, J. (2016). Privacy and confidentiality. I. Foster, R. Ghani, R.S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big Data and Social Science* (Ch. 12). CRC Press. ISBN: 9781498751407. https://textbook.coleridgeinitiative.org/

Biemer, P. (2016). Data quality and inference errors. In I. Foster, R. Ghani, R.S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big Data and Social Science* (Ch. 10). CRC Press. ISBN: 9781498751407. https://textbook.coleridgeinitiative.org/

Cook, L. & Chaleampong, K. (2010). The Idea Gap in Pink and Black. *NBER Working Paper No. 16331*. https://doi.org/10.3386/w16331

Healy, K., Moody, J. (2014). Data visualization in sociology. *Annual Review of Sociology*, *40*(1), 105-128. https://doi.org/10.1146/annurev-soc-071312-145551

Healy, K. (2020). *Principles of data visualization* [Video]. YouTube. https://www.youtube.com/watch?v=wHrzsO564uA

Hofstra, B., Kulkarni, V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., McFarland, D. (2020). The diversity-innovation paradox in science. *Proceedings of the National Academy of Sciences*, *117*(17), 9284-9291. https://doi.org/10.1073/pnas.1915378117

Lane, J., Bender, S., Nissenbaum, H., & Stodden, V. (Eds.). (2014). *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge (England): Cambridge University Press.

Lane, J. (2017). *Big data, privacy and the public good* [Video]. YouTube. https://youtu.be/XDT7FnXvM58

Leahey, E. (2016). From solo investigator to team scientist: Trends in the practice and study of research collaboration. *Annual Review of Sociology, 42*: 81-100. https://doi.org/10.1146/annurev-soc-081715-074219

Leahey, E., Beckman, C., & Stanko, T. (2017). Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Administrative Science Quarterly, 62*: 105-139. https://doi.org/10.1177/0001839216665364

McKinney, W. (October 2017). *Python for data analysis: Data wrangling with pandas, NumPy, and IPython* (2nd ed.). Sebastopol, CA: O'Reilly Media, Inc. ISBN: 9781491957660

Page, S. (2014). Where diversity comes from and why it matters? *European Journal of Social Psychology, 44*(4), 267-279. https://doi.org/10.1002/ejsp.2016

Page, S. (2017). *Why we need more diversity to solve complex problems* [Video]. YouTube. https://www.youtube.com/watch?v=2GYOx1PF3Bc

Reagans, R., & Zuckerman, E. (2001). Networks, diversity, and productivity: The social capital of corporate R&D teams. *Organization Science, 12*(4), 393-521. https://doi.org/10.1287/orsc.12.4.502.10637

Tokle, J., & Bender, S. (2016). Record linkage. In I. Foster, R. Ghani, R.S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big Data and Social Science* (Ch. 3). CRC Press. ISBN: 9781498751407. https://textbook.coleridgeinitiative.org/

Uzzi, B., Mukerjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*: 468-472. https://doi.org/10.1126/science.1240474

Wisnioski, M., et al. (2019). Does America need more innovators? *The MIT Press*. https://doi.org/10.7551/mitpress/11344.001.0001

Workshop Dataset and Underlying Source Data

The dataset we will be using defines a team as the complete set of individuals who are paid in a given year on grants for which a given faculty member is a principal investigator (PI). This definition follows standard practice in many science and engineering fields. Because UMETRICS data are based on payments made to employees from specific grants and because individuals can participate in multiple teams, we can begin to think about the ways in which individual locations in larger collaborative networks influence the behavior and outcomes of individual researchers and students. For the purposes of this workshop, we will treat scientific publications as key indicators of outcomes.

Team Diversity Data

The team diversity dataset features one row for each year of a team leader's awards and affiliated team members. Awards are those federally funded sponsored projects for which the team leader is a PI in a given year. Team members are the employees (if any) paid at the same university on those awards in the same year while the team leader is PI. Each row includes attributes of the team leader, members, and awards, including scientific publication outcomes. We selected two major federal sponsors, the National Institutes of Health (NIH) and the National Science Foundation (NSF), and linked PIs to UMETRICS employees in the data from 25 IRIS member universities. Several sources were integrated to construct this team dataset, including the UMETRICS employee transaction and award files, grant and PI information maintained by the NSF and NIH, and publication data provided by PubMed.

This dataset includes about 25,000 unique PIs who led their teams through over 60,000 sponsored projects funded by NIH or NSF between 2001 to 2019 (contingent on each university's available data). These awards together pay more than 200,000 unique team members, including other faculty members, post-doctoral scholars, graduate students, undergraduate students, and professional staff. The dataset consists of variables characterizing and measuring team composition, diversity, and performance. Each of these can serve as independent or dependent variables when studying team diversity and its impact on team performance and productivity. Team measures include age, gender, and racial/ethnic composition of each team as well as information on their occupational classifications (such as faculty, staff, or graduate student). In the course of the workshop we will also talk about the strengths and weaknesses of different measures and the limitations of the data.

IRIS UMETRICS Data

The team diversity dataset for this workshop was built on the UMETRICS data at the core of IRIS. The name refers to the original UMETRICS initiative (Universities Measuring the EffecTs of Research on Innovation, Competitiveness and Science), which collects and integrates large-scale data on university research. IRIS was founded in 2015 to continue this initiative by collecting, integrating, and expanding big data on university research.

UMETRICS data are centered on university financial and personnel administrative data pertaining to sponsored project expenditures submitted by IRIS member universities. Each IRIS member university contributes records from its sponsored projects, procurement, and human resources systems. Individual campus files are de-identified, cleaned, and aggregated by IRIS to produce the annual data release (available only via the IRIS Virtual Data Enclave). The 2020 release includes the transaction-level information on over 400,000 sponsored projects that represent approximately $100 billion in direct cost expenditures and employ more than 700,000 people at 33 universities, with coverage between 2001 and 2019 in a total of 60+ million rows (record counts) in relational tables.

These data are linked to information on scientific outcomes in order to better understand broad aspects of university research and its impact. Although this workshop does not cover how we manage record linkage through data manipulation and linkage techniques or applications, the dataset we have developed for this workshop is built on several different linkage analyses of research grant activities, university research, and scientific outcome (e.g., publication) at multiple levels of analysis.

Orienting Questions

Introduction

We believe that people develop new tools best in the context of trying to answer concrete, substantive questions. To that end, we will structure our work together around a common set of “starting point” questions. This will allow us to explore data and analytic techniques together using common objects of focus. While you may decide to proceed with further research around the questions we propose here, you need not. We will also spend some time introducing you more fully to the UMETRICS data that IRIS maintains, but all of our examples and teaching materials will be built around the common questions we outline below. We would be very pleased to see these questions be extended or amended according to your interests as the workshop progresses.

Overall question and some possible extensions

A long line of research suggests that more diverse teams are better able to reach innovative and effective solutions to challenging problems than more homogeneous teams (see Background section below). We have structured workshop materials around a dataset derived from UMETRICS that documents scientific teams at 25 universities. Our work together will focus on using these data to describe team diversity in science and engineering on multiple dimensions and relate measures of diversity to individual and group scientific outcomes.

These course materials are intended to prime the pump to address these questions based on recent work in the science of science. We are excited to learn more about how you will approach the following questions and what concepts, theories, or findings your work and your field bring to the discussion.

Overarching question: Are more diverse scientific teams more or differently productive than more homogenous scientific teams?

Brainstorming follow-on questions:

1. What are the dimensions along which team diversity can be measured?
2. Do different sorts of diversity have different effects on team-level outcomes?
3. What kinds of teams promote the strongest outcomes for graduate students or other trainees?
4. Does it matter whether the people who make diverse teams diverse are faculty?
5. Is there a tradeoff between scientific productivity and graduate or trainee outcomes at the team level? If so, is it stronger or weaker for more diverse teams?

Background

Research in all fields of inquiry has gotten more collaborative (Leahey, 2016). Larger teams and those that span institutions appear to produce more and higher impact research (Wuchty, Jones, & Uzzi, 2007), but there is also some evidence that smaller teams are better at generating more radical, disruptive innovations (Wu, Wang, & Evans, 2019). Across management, sociology, public policy, and the science of science, research is concerned with questions about how the composition, structure, and functioning of teams relate to what they produce and how their members do.

In science and other knowledge-intensive areas, evidence about the importance of diversity for outcomes relies on an underlying theory of innovation that emphasizes the idea that solutions to problems at the frontiers of knowledge are driven by the recombination of ideas, skills, and knowledge (Owen-Smith, 2018; Schumpeter, 1947). In other words, new discoveries are taken to be the result of putting existing ideas and knowledge together in novel ways to either

accomplish something new or accomplish something known in a better way. Under this model, core questions have to do with how teams search for information and skills that are useful to their goals, and how more or less typical combinations of different types of knowledge relate to the character and impact of findings and innovations.

It is here that the literature tends to turn to questions about team diversity. The core idea is that teams which are "cognitively diverse" are better able to develop innovative solutions to problems because their members are familiar with and can work to integrate information drawn from lots of different fields, communities, or approaches (Page, 2007; Reagans & Zuckerman, 2001). So, this line of work suggests that markers of a team's cognitive diversity should be associated with more and more innovative discoveries.

But there are challenges to this view. First, researchers recognize that more diverse and dispersed teams face greater coordination challenges (Cummings & Kiesler, 2007; 2011). Second, teams with too much diversity may find it hard to collaborate at all, let alone to produce new findings (Rawlings et al., 2015). There is also some evidence that women and under-represented minority scholars are more likely to work in interdisciplinary research areas and that the coordination and socio-emotional work of making diverse teams "go" tends to fall on the shoulders of (particularly early career) women and minority researchers, with predictable and often negative career consequences (Leahey, Beckman, & Stanko, 2017). More generally, network researchers find that people who span and integrate very different communities face greater cognitive challenges and are more likely to face burnout (Sasovova et al., 2010). All of these findings suggest that while diverse teams may be better for team level productivity, improvement may come at the cost of individual outcomes for some researchers.

In other words, innovation in academic science and engineering may pose a particularly dangerous "diversity dilemma" where teams that are meaningfully diverse produce better and higher impact research, but the people who make those teams diverse face worse individual career outcomes. Understanding, explaining, and addressing this dilemma would represent an important step toward broader participation in STEM, and toward a more equitable, inclusive, and effective research ecosystem.

UMETRICS data are very well suited to research that explores these questions and tradeoffs. Integrating knowledge about diversity and its effects from education and other fields little represented in this brief summary also have great potential to inform out understanding of teams and their workings. For both these reasons, we will focus our collective work (both technical and substantive) around these general orienting questions using a dataset derived from IRIS UMETRICS that characterizes research teams in a fashion that allows for diversity to be measured on multiple dimensions and related to both team and individual level outcomes.

These course materials are intended to prime the pump to address these questions based on recent work in the science of science. We are excited to learn more about how you will approach the questions and what concepts, theories, or findings your work and your field bring to the discussion.

Background References

Cummings, J., & Kiesler, S. (2011). Organization theory and new ways of working in science. *2011 Atlanta Conference on Science and Innovation Policy*, 1-5.

Cummings, J., & Kiesler, S. (2007). Coordination costs and project outcomes in multi-university collaborations. *Research Policy, 36*(10): 1620-1634.

Leahey, E. (2016). From solo investigator to team scientist: Trends in the practice and study of research collaboration. *Annual Review of Sociology, 42*: 81-100. https://doi.org/10.1146/annurev-soc-081715-074219

Leahey, E., Beckman, C., & Stanko, T. (2017). Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Administrative Science Quarterly, 62*, 105-139. https://doi.org/10.1177/0001839216665364

Owen-Smith, J. (2018). *Research universities and the public good: Discovery for an uncertain future.* Redwood City, CA: Stanford University Press.

Page, S.E. (2007). *Difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Rawlings, C.M., McFarland, D.A., Dahlander, L., & Wang, D. (2015). Streams of thought: Knowledge flows and intellectual cohesion in a multidisciplinary era. *Social Forces, 93*(4), 1687-1722.

Sasovova, Z., Mehra, A., Borgatti, S.P., & Schippers, M.C. (2010). Network churn: The effects of self-monitoring personality on brokerage dynamics. *Administrative Science Quarterly, 55*(4), 639-670.

Schumpeter, J.A. (1947). The creative response in economic history. *The Journal of Economic History, 7*(2), 149-159. https://doi.org/10.1017/S0022050700054279

Wu, L., Wang, D., & Evans, J.A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*, 378-382. https://doi.org/10.1038/s41586-019-0941-9

Wuchty, S., Jones, B.F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5927), 1036-1039.