Universal Provider Award

Data & Computational Science Series (DCS^2) Summary Report for 2018

Prepared for

The University of Cincinnati Office of the Provost

December 2018

Universal Provider Award

Data & Computational Science Series (DCS^2) Summary Report for 2018

Report prepared by

Amy Latessa, Ph.D. Candidate, Research Coordinator, UCIT Research and Development

Amy Koshoffer, Assistant Director for the Research & Data Services, UC Libraries

Suggested citation:

Koshoffer, A. & Latessa, A. (2018 December). *Data & Computational Science Series (DCS^2) Summary Report for 2018.* Cincinnati, Ohio: UC Information Technologies and UC Libraries, University of Cincinnati.

Universal Provider Award
Data & Computational Science Series (DCS^2)
Summary Report for 2018

Table of Contents

I. Proposal Background

The University of Cincinnati Information Technologies Department of Research and Development (UCIT R&D) and the University of Cincinnati Libraries (UCL) were funded $12,000 toward a collaborative symposium on Data Science and Computational Research. The proposals we each respectively submitted to the Office of the Provost Universal Provider Award initiative were combined and we called the symposium, the Data and Computational Science Series, DCS^2.

a. Team

The organizational team was composed of Ted Baldwin and Amy Koshoffer from UCL and Jane Combs and Amy Latessa from UCIT R&D. We have collaborated together on various initiatives including UC's Data Day and our job spheres often overlap in the fields of data management, data repositories, cyberinfrastructure and computational resources. We are the representatives at the UC new faculty orientation and feel that our collaborations together makes the UC research community more fortified.

b. Goals

Since the UCL and UCIT R&D proposals were similar and combined, we updated our goals to the following:

1) *Introduce free data & computational science resources and tools to the UC research community*
    a. Advanced research computing including high performance computing
    b. Qualitative Data tools
    c. Research Reproducibility
    d. Science Gateways
    e. Data Visualization
2) *Strengthen the UC computational research community*
    a. Bring multi-disciplinary researchers together to learn about domain-agnostic resources
    b. Introduce UC research faculty to each other, share data problems and solutions
    c. Introduce UC teaching & research faculty to resources and peers a regional universities – to encourage good stewardship of federal funding initiatives.

c. Acknowledgements

We would like to thank many other UC faculty and staff for their support to make the series a success. We could not have hosted all of the events without the help of people for room reservations, catering assistance and promotional materials and distribution. Thank you to Melissa Norris-Cox for designing our logo and committing to work with us for the 2019 series. And of course, special thanks to the Office of the Provost Universal Funding Award, that without, these workshops and collaborations were but a vision.

II. Events & Workshops

Table 1. DCS² 2018 Workshops & Events

Date	Workshop/Event Title
March 22, 2018	Introduction to Jetstream
March 23, 2018	Jetstream Atmosphere: Hands-on
May 2, 2018	Jetstream Atmosphere: Hands-on
May 2, 2018	Jetstream API
May 21, 2018	Jupyter Notebooks & Data Visualization & Analysis
May 31, 2018	Spatial Data: Sharing Your Research with Storymaps
June 4-7, 2018	XSEDE Monthly HPC Workshop: HPC Boot Camp
July 24, 2018	Qualitative Data: Screenscraping Using R
August 7, 2018	XSEDE Monthly HPC Workshop: OpenMP
September 5-6, 2018	XSEDE Monthly HPC Workshop: Big Data
October 3-4, 2018	Research Reproducibility: Code Ocean
October 30, 2018	Science Gateways and the Science Gateways Community Institute (SGCI)
October 31, 2018	Jetstream Atmosphere: Hands-on
October 31, 2018	DCS² Spooky Social
November 14, 2018	GIS Day
November 28, 2018	Data Visualization & Shiny

a. XSEDE[1] Monthly HPC Workshops

The University of Cincinnati was invited to be a satellite-site for these monthly high performance computing workshops hosted by XSEDE and the Pittsburgh Supercomputing Center after Jane Combs served on a NSF panel-review for NSF-funded Bridges.[2] These monthly workshops included:

1. 4-day HPC boot camp from beginner to advanced level with topics on MPI, OpenMP, OpenACC and accelerators
2. OpenMP to give C and Fortran programmers a hands-on introduction to MPI programming
3. Big Data with topics such as Hadoop and Spark

[1] eXtreme Science and Engineering Discovery Environment

[2] Bridges is a uniquely capable supercomputer designed to empower new research communities, bring desktop convenience to supercomputing, expand campus access and help researchers needing to tackle vast data to work more intuitively. Bridges consists of three tiered, memory-intensive resources to serve a wide variety of scientists, including those new to supercomputing and without specialized programming skills. Funded by a $9.65-million NSF grant, Bridges offers new computational capabilities to researchers working in diverse, data-intensive fields such as genomics, the social sciences and the humanities. The system was architected by Hewlett Packard Enterprise and features advanced technology from Intel® and NVIDIA®, including the first-ever installation of Intel® Omni-Path Architecture.

Photo of XSEDE Big Data Workshop on September 5, 2018

Instructor: John Urbanic, Parallel Computing Scientist, Pittsburgh Supercomputing Center						
Date	**Workshop (Format:** Wide-area classroom (WAC) training platform	**Attendees**				**Link**
		Faculty	**Graduate Students**	**Staff**	**Total**	
6/4-7	HPC Boot camp	2	9		11	https://www.youtube.com/channel/UCTVSqD_elHZtuRv0PKTLOEw/videos
8/7	OpenMP	0	5		5	https://www.youtube.com/channel/UCTVSqD_elHZtuRv0PKTLOEw/videos
9/5-6	Big Data	13	9		22	https://www.youtube.com/channel/UCTVSqD_elHZtuRv0PKTLOEw/videos

b. Cloud Computing: Jetstream

Jeremy Fischer, a senior technical advisor from Indiana University's Information Technology Services taught the workshops using Jetstream, describing Jetstream in greater detail, as well as how its unique combination of hardware, software, and user engagement support the "long tail of science." NSF-funded Jetstream Cloud (Indiana University and Texas Advanced Computing Center) is a user-friendly cloud environment designed to give researchers and students access to computing and data analysis resources on demand — from their tablets, laptops or desktop computers. People interact with the system through a menu of "virtual machines" designed to support research in many disciplines including biology, atmospheric science, earth science, economics, network science, observational astronomy and social sciences. https://jetstream-cloud.org/

The Infrastructure-as-a-Service platform comprised of two geographically isolated clusters, each supporting hundreds of virtual machines and data volumes. The two cloud systems are integrated via a user-friendly web application that provides a user interface for common cloud computing operations, authentication to XSEDE via Globus, and an expressive set of web service APIs. Jetstream enables on-demand access to interactive, user-configurable computing and analysis capability. It also seeks to democratize access to cloud capabilities and promote sharable, reproducible research.

Instructor: Jeremy Fischer, Senior technical advisor, Indiana University Information Technology							
Date	**Format**		**Attendees**				**Link**
			Faculty	**Graduate Students**	**Staff**	**Total**	
3/22	Introduction to Jetstream (1.5 hours)		2	3		5	https://www.youtube.com/watch?v=1dWwqkXWdRg&index=19&list=LLxq1lfQ71gslqLjZrxIzdpA&t=633s
3/23	Atmosphere Jetstream: Hands-on Workshop (3 hours)	Morning Session	1	4		5	
		Afternoon Session	3	14		17	
5/2	Introduction to Jetstream (1.5 hours)		6		1	7	https://www.youtube.com/watch?v=1dWwqkXWdRg&index=19&list=LLxq1lfQ71gslqLjZrxIzdpA&t=633s
5/2	Jetstream API Tutorial		2	2		4	https://www.youtube.com/watch?v=7bDFw41_PsM&t=1870s
10/31	Atmosphere Jetstream: Hands-on Workshop (3 hours)			2		2	

c. Data Visualization and Analysis Tools: Jupyter

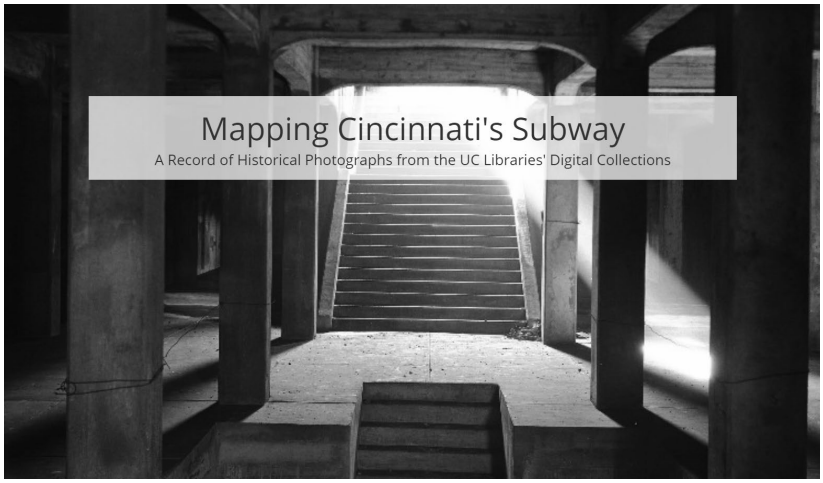
Jupyter Notebook (http://jupyter.org/) is an open-source web application where researchers can create and share documents that contain live code, equations, visualizations and narrative text. The use of this resource helps facilitate a hand- on experience for learning to use code. In the research environment, coding skills can help build an environment of research reproducibility. Jupyter notebooks are also a great resource for teaching data visualization best practices and how to use the open source coding language Python (https://www.python.org/) for data visualization.

Instructor: Bill Mcmillin, Digital Metadata Librarian, UC Libraries (now Data Analyst at 84.51˚)						
Date	**Format**	**Attendees**				**Link**
		Faculty	**Graduate Students**	**Staff**	**Total**	
5/21	Hands-on Workshop - (1 hour) Working with Jupyter Notebooks	4	8	10	22	No links available
5/21	Hand on Workshop - (2 hours) Jupyter Notebooks and Data Visualization in Python					

d. Spatial Data: Sharing Your Research with Storymaps

In the UC Libraries Data & GIS Collab, we are fielding more and more requests for assistance using web GIS resources such as Google Earth, ArcGIS online and other web mapping tools. Storymaps is an ESRI application available through ArcGIS online available to our UC community through our enterprise license. The presentation gave an introduction to the various tools available and how they can be used to great benefit in research. The hands-on workshop focused on the Storymaps tool and how to incorporate storytelling through text and images to enhance communication about research projects. Storymaps is also a great resource for faculty in teaching and helps students understand how to create

an effective digital scholarship project. The hands-on part of the workshop also highlighted the UC Libraries digital collections entitled Cincinnati Subway and Street Improvements, 1916-1955 - http://digital.libraries.uc.edu/collections/subway/.



Instructor: Joshua Sadvari, Research Commons Program Manager and Geographic Information Systems (GIS) Specialist, The Ohio State University						
Date	**Format**	**Attendees**				**Link**
		Faculty	**Graduate Students**	**Staff**	**Total**	
5/31	Introduction to Webmapping (1 hour)	5	11	4	30	https://youtu.be/qOPhOHTuUrg
5/31	2-hour Hands-on Workshop: Sharing Your Research with StoryMaps					

e. Qualitative Data: Screenscraping Using R

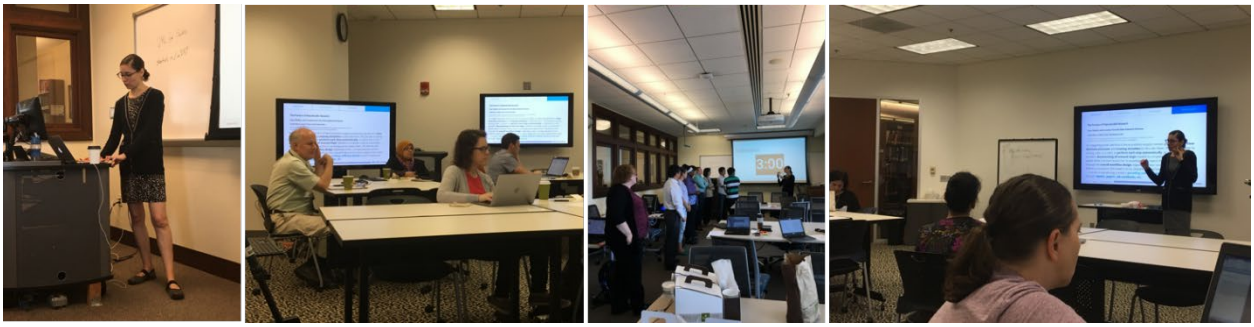
The Qualitative Data Repository (QDR) https://qdr.syr.edu/ is a NSF-funded repository dedicated to long term preservation of digital data generated or collected through qualitative and multi-method research in the social sciences. The presentation highlighted the unique nature of qualitative data collection and management and how the QDR is developing resources for research reproducibility, particularly the Annotation for Transparent Inquiry approach. Through this approach, researchers highlight the source text from which the data set within the repository derives and documents the pathway and methodology from source to data. Annotation allows the researcher to add in notes describing the analysis process further facilitating research reproducibility. Similar to the benefits of using Python, the programming language R (https://www.r-project.org/) is becoming an increasingly essential research tool as its use in data collection and analysis ensures greater research reproducibility. The hands on workshop led participants through a data set creation exercise using R to collect data from a website and create a qualitative data set.

Instructor: Sebastian Karcher, Associate Director of the Qualitative Data Repository						
Date	**Format**	**Attendees**				**Link**
		Faculty	**Graduate Students**	**Staff**	**Total**	
7/24	Annotation for Transparent Inquiry (ATI) (1 hour)	9	19	2	30	Presentation video: https://youtu.be/za7s54atuyY Slides: https://doi.org/10.6084/m9.figshare.6856826.v1
7/24	Hands-on Workshop: Sharing Your Research with StoryMaps (2 hours)	3	14	1	18	Workshop video: https://youtu.be/MTm_MiCXO7g Slides: https://adam3smith.github.io/web-scraping/

f. Research Reproducibility: Code Ocean

Code Ocean (https://codeocean.com/is a cloud-based computational reproducibility platform that provides researchers and developers an easy way to share, discover and run code published in academic journals and conferences. Their mission is to make the world's scientific code more reusable, executable and reproducible. They offer free accounts for academics and have working relationships with several publishers. These workshops offered a platform independent framework and approach focused on research reproducibility especially concerning code preservation. The focus of each specific workshop was targeted to the potential audience on each campus.

Instructor: April Clyburne-Sherin, Director of Scientific Outreach, Code Ocean						
Date	**Format**	**Attendees**				**Link**
		Faculty	**Graduate Students**	**Staff**	**Total**	
10/3	Hands-on Workshop (3 hour) Integrating reproducible best practices into biomedical & clinical research.	5	0	4	9	Workshop video: https://www.youtube.com/watch?v=V_VBmFNXWg0 Slides: http://bit.ly/2DToKHt
10/4	Hands-on Workshop (3 hour) Preparing your data and code for reproductive publication	3	7	4	14	Workshop video: https://youtu.be/TkQbtEYXuUA Slides: shorturl.at/iwDVY



g. Science Gateways and the Science Gateways Community Institute (SGCI)

Science Gateways - also called virtual research environments or virtual labs - allow science and engineering communities to access shared data, software, computing services, instruments and other resources specific to their disciplines and use them also in teaching environments. The U.S. Science Gateways Community Institute (SGCI) provides free resources, services, experts, and ideas for creating and sustaining science gateways. The challenges for creators of specific science gateways are manifold, and the expertise needed for well-designed science gateways is very diverse. The sustainability of science gateways is crucial to serve communities effectively, efficiently and reliably. One measure to achieve greater sustainability of science gateways is establishing on-campus teams. Researchers are served more efficiently since the support by experienced developers reduces individual project investments, and a team can bring the diversity of required expertise for a well-designed science gateway. The talk goes into detail about the challenges and the benefits of on-campus groups and of sharing resources across a campus. Four successful cases are discussed and the services of the Science Gateways Community Institute (SGCI) are described to support the process in building such groups, as well as the recommendation of strategies for using free campus resources.

Instructor: Sandra Gesing, research assistant professor at the Department of Computer Science and Engineering and a computational scientist at the Center for Research Computing at the University of Notre Dame						
Date	**Format**	**Attendees**				**Link**
		Faculty	**Graduate Students**	**Staff**	**Total**	
10/30	The Sustainability of Science Gateways via On-Campus Groups	2	1	8	11	
10/30	General Introduction to Science Gateways & the Science Gateways Community Institute (SGCI) and Science Gateways Presentation On Usability With Hands-On Portion	3	3	2	8	

h. DCS² Spooky Social

Beyond speakers and hands-on trainings, we organized a *DCS²* Social on October 31, 2018. Feedback from UC faculty attendees indicated that they seek a space for conversation and informal networking with other research-oriented faculty. We provided soda, water and pizza and had a good turnout

considering the rainy night. Bringing people together in these ways encourages interdisciplinary thinking and results and reaches beyond simple attendance numbers. We believe that the community-building component of our collective efforts is essential component for long-term faculty retention.

Jane Combs with Dr. Prashant Khare (CEAS)

Location: UC Catskeller						
Date	**Format**	**Attendees**				**Link**
		Faculty	**Graduate Students**	**Staff**	**Total**	
10/31	Open social	6	3	5	14	

i. GIS Day

This community building event brings to together researchers and practitioners in the field of Geographic Information Systems (GIS) for a celebration of the application and power of GIS. In addition to the Provost Office contribution through the UP grant, the event is sponsored by the Geography Graduate Student Association, Graduate Student Governance Association, Department of Geography & GIS, and Joint Center for GIS and Spatial Analysis.

Instructor: Dr. Daniel Sui, Vice President for Research University of Arkansas; Former Division Director National Science Foundation							
Date	**Format**	**Attendees**					**Link**
		Faculty	**Graduate Students**	**Undergrad**	**Staff**	**Total**	
11/14	GIS and Emergence of Convergence Research	5	14	6	10	35	Presentation video and slides: https://scholar.uc.edu/show/bn999796n

j. Data Visualization with Shiny (R Studio)

If a picture is worth of a thousand words, data visualization is worth millions: Toward a framework for actionable visual insights. Current shift in scientific landscape toward cross-disciplinary teams, evolving cyberinfrastructure and complex data requires a new kind of data analysis and visualization tools. This talk introduced a visualization framework developed at Cyberinfrastructure for Network Science Center (CNS) at Indiana University, founded and directed by Professor Katy Börner, Victor H. Yngve Distinguished Professor of Engineering and Information Science (http://cns.iu.edu). The talk was followed by a two hour hands-on workshop that stepped through the process of building, visualizing, deploying, and sharing Shiny web applications. Learning this workflow will enable you to build your own interactive tools that can be used for research and teaching.

Images of the 2-hour Shiny workshop in Langsam Library room 475 given by Dr. Olga Scrivner

Instructor: Dr. Olga Scrivner, research scientist at Cyberinfrastructure for Network Science Center (CNS) in the School of Informatics, Computing, and Engineering at Indiana University, faculty fellow at the Center of Excellence for Women in Technology, and corporate faculty in Data Analytics at Harrisburg University of Science and Technology

Date	Format	Attendees				Link
		Faculty	Graduate Students	Staff	Total	
11/28	Lunch & Learn: The Importance of Data Visualization; 2-hour hands-on workshop: Building Interactive Web Applications: Data Visualization with Shiny	13	15	5	33	

III. General Summary

- Organized, hosted and planned 9 speakers/seminars and 23 workshops and hands-on training sessions.
- Trained over 300 people, including 87 faculty (29%). Faculty attendees came from 8 UC colleges, representing 14 departments/disciplines: A&S (math, chemistry, biology), Lindner College of Business (OBAIS), College of Medicine (environmental health, biomedical informatics), CEAS (computer science, mechanical & materials engineering, engineering

education), CECH (education, criminal justice), College of Nursing, College of Allied Health Sciences, and University of Cincinnati Libraries.

- Connected faculty with external experts, resources and tools. Introduced UC faculty to free regional and national research and computing resources, technical and disciplinary experts, and enhanced research practices to increase quality of research outputs. Numerous faculty have established accounts to utilize these new resources, including (1) Indiana University's cloud-computing resource, *Jetstream*; (2) the University of Pittsburgh's supercomputing resource *Bridges*, and (3) the National Science Foundation's Extreme Science and Engineering Discovery Environment (XSEDE), a single virtual system that scientists can use to interactively share computing resources, data and expertise. Faculty across disciplines have consulted with technical experts and instructors, including Sebastian Karcher (Assoc Director, Qualitative Data Repository, Syracuse University) and Josh Sadvari (Program Manager and GIS Specialist, Ohio State University).
- In support of enhanced pedagogy, a College of Business faculty member is adopting the Jetstream platform as an educational tool to teach a Business Intelligence course (50 students in fall semester, 180 in spring semester).

IV. Marketing & promotional campaign & post-workshop feedback

Melissa Cox-Norris, Directory of Library Communication, University of Cincinnati Libraries (UCL) designed a flyer and banner template for us to maintain a brand uniformity of the series (see Appendix A). We were able to use this template for all of our events and we created a communication's distribution list to get the word out. Most of the events' registration were through the Faculty Development OneStop,[3] which allowed us to set workshop size, share completion of workshop certificates with attendees, and have a repository of our events. There was no way to limit registration to faculty only, which we were initially concerned about, but as our registration numbers show, it turned out to not be a problem, as we had mixed faculty, student and staff at each event. Each event was posted on the UC Liblog and the UCIT R&D website.

After each workshop we shared a 6 questions Qualtrics survey to assess attendee satisfaction. Although the complete rate of the survey is low, we did gather some insights (Appendix C).

Q1 - How did you hear about the workshop?
Q2 - How well did the workshop meet the description?
Q3 - How well did this class meet your personal objectives?
Q4 - Overall, how effective was the training session?
Q5 - What did you hope to learn in this workshop?
Q6 - Please list suggestions for other courses, questions and/or comments

V. Recommendations for DCSS 2.0

a. Further streamline our efforts:

[3] The XSEDE monthly HPC workshops maintained their own registration since they needed to provide users with free log-on accounts.

i. Amy K and Amy L met with Melissa Cox-Norris to set a marketing campaign for the 2019 series.
    1. We hope to set up a WordPress site so that we can promote the entire series – one of the workshop feedbacks was "Is there any list for all the seminars to review?"
ii. The team will meet in mid-December to sketch the series; we were awarded the entire amount for our 2019 proposal, so the proposed speakers and workshops are already outlined.
iii. Richard Johansen, UC's Data Visualization specialist will join our team to promote and utilized the new data visualization wall in the Geology-Math - Physics library

b. The DCS2 social has more potential and we plan to host two in the 2019 series
    i. Incorporate a poster session or 3-minute thesis/research "open-mic"
c. Gather more pre- and post- faculty input
d. Work with the Faculty Enrichment Center
e. Provide parking passes to faculty so we can host the events on both East and West campus

VI. DCSS 2019 Proposal

Our team proposed enhancements and expansion of the *Data and Computational Science Series (DCS²)* expert speaker and hands-on training events during calendar year 2019; the sponsorship of two *DCS²* networking socials (including flash talks and poster sessions); enhancement of UCL and UCIT R&D's existing Research Data Services directed to faculty (e.g., additional research software available for all faculty in Research Data Services' lab spaces). In total, these educational offerings, networking opportunities, and enhanced service capabilities will bring us closer to providing a more comprehensive toolkit of resources that enable the faculty to succeed and prosper. The *DCS²* 2019 will include both in-house and external experts in a wide range of computational and data science domains, with the aim to enrich the quality of faculty teaching, scholarship and research. Planned areas of focus:

1. **Data Visualization**
2. **Advanced research computing**
3. **Machine Learning/AI/Big Data**
4. **DCS² Socials**

We were awarded the full requested amount of $24,966 and we look forward to a bigger and better 2019 series.

VII. Conclusion

The team had some learning curves; organizing faculty-wide events is not for the faint of heart. However, our mixed levels of expertise and knowledge about how to coordinate and operate visiting guests and workshops at UC proved a good mix. The UCL and UCIT R&D representatives are in direct contact with UC research faculty, supporting their data and research infrastructure needs, thus this extra funding only enhances the services and resources that we can introduce and provide to increase the success of UC faculty research, teaching, funding, publications, community-building and retention.