LIBRARY WORKSHOPS IN SUPPORT OF DATA-DRIVEN RESEARCH IN TOP NIH- AND NSF-FUNDED UNIVERSITIES

Tanja Bekhuis, PhD, MS, MLIS, AHIP (Distinguished)

EDDA Analytics Group™

Pittsburgh, Pennsylvania, USA

No. 00170205v2

PREFACE

In response to an urgent need for computationally-savvy researchers, leading university libraries are developing and offering participatory workshops to introduce their patrons to data-driven research methods and resources. In doing so, libraries help to improve the competitive advantage of their universities relative to other schools that compete for federal funds.

We conducted a qualitative study to explore the nature of library workshops offered in top NIH- and NSF-funded universities. To guide strategic planning, we present a catalog of the workshops offered in fall 2016, two indexes for resources and workshop content, and a thematic analysis.

A Brief Word about Methods

We used NIH Research Portfolio Online Reporting Tools and the NSF Budget Internet Information System to identify top-funded universities. From their corresponding websites, we extracted information on 99 workshops offered by five health sciences libraries and five main libraries in schools funded by NIH and NSF, respectively. We recorded workshop title, duration, and description by source of federal funding and library. Additionally, we used qualitative data analysis methods to identify themes in the dataset, and natural language processing to identify candidate terms for the resource and subject indexes. The results of the content analyses, as well as the indexes, yield insights regarding workshop coverage.

Keywords

Research libraries, data-driven research, library workshops, strategic planning, competitive advantage

Citation

Bekhuis T, EDDA Analytics Group™**.** Library workshops in support of data-driven research in top NIH- and NSF-funded universities [no. 00170205v2]. Pittsburgh, PA, USA: TCB Research & Indexing LLC. February 2017.

TABLE OF CONTENTS

NIH-FUNDED SCHOOLS AND LIBRARY WORKSHOPS

Johns Hopkins University Welch Medical Library

De-Identifying Human Subjects Data for Sharing
1 hour

With researchers increasingly encouraged or required to share their data, preparing to share datasets with confidential identifiers of people and organizations is particularly challenging. Join JHU Data Management Services for an overview of techniques for assessing disclosure risk and hiding personal identifiers and Protected Health Information in quantitative and qualitative data, in compliance with IRB and HIPAA guidance. We also discuss preparing consent forms that facilitate data sharing, and keeping identifier data secure during and after projects.

Introduction to ArcGIS
3 hours

This half-day course introduces students to ArcGIS, the most widely used geographic information systems (GIS) software. Learn the basics of ArcGIS, how to work with spatial data, and how to create maps. If your research calls for making maps or using a geographic information system to analyze data, then this class is for you.

Preparing Your Research Data for Re-use Workshop
1 hour

How can you organize, document and package up your research data so that you or another researcher can make easy use of your data in the future? Perhaps you are finishing a research project and want to make the associated data usable going forward. Or perhaps after using data from someone else's project, you vow not to let others be as frustrated when using your data!

In either case this workshop is for you. This workshop gives researchers practical steps for preparing research data for re-use by themselves and others. Data Management Services will take researchers through several data preparation steps (e.g., data selection, organization, documentation, preservation), and introduce a spreadsheet tool that facilitates and documents those steps. Attendees can apply these steps for their own research project!

This workshop is free to the JHU community and aimed at faculty and graduate students from all disciplines. Researchers at all stages of the research process are welcome to attend, whether they are just beginning or finishing a project, somewhere in between, or have a past project they would like to more thoroughly document, organize and archive.

Geocoding and Joining Data Using ArcGIS
3 hours

Prerequisite: Introduction to ArcGIS (Welch half-day workshop). Learn the steps required for successful joining of data and geocoding along with tips and suggestions for preparing data for use with ArcGIS. Common file formats will be discussed, (e.g. Excel, dBase, Access), along with text files and data with x, y coordinates. We'll explore street files available from the library along with optional interfaces for the geocoding process.

Finding and Importing GIS Data

3 hours

Prerequisite: Introduction to ArcGIS (Welch half day workshop). This 3-hour workshop will explore sources of maps and data. Participants will practice downloading data from internet sources such as the Census Bureau and the Geography Network and then learn how to import the data into ArcGIS. In addition, participants will learn how to identify and change map projections.

ArcGIS Online: The Key to Web Mapping

3 hours

Attendees will learn how to search, find, and share geographic content using Johns Hopkins ArcGIS Online Organization account. This class will provide the fundamental skills necessary to create, design, and share web maps, as well as use some of the various geoprocessing tools currently offered via the online application.

Data Management Plan Preparation

1 hour

Grant proposals for a growing number of funders require data management plans, including the National Science Foundation. Developing a competitive data management plan requires understanding and effectively addressing the many aspects of research data management that funders and reviewers emphasize (e.g., plans for research data security, sharing, and documentation). Join Data Management Services for a training session on preparing data management plans. During this one-hour workshop, we will cover the research data questions one should answer in creating an effective, competitive plan. Participants will have an opportunity to ask data management and planning questions specific to their research.

Data Sources for Health and Medical Research

1 hour

This one hour class introduces you to available secondary data sources for your research. Some of these resources are paid for such as several special physical and mental health archives (hosted by ICPSR), some are produced/hosted by various sub-agencies within the U.S. Department of Health and Human Services and institutes inside NIH. With the rise in data sharing mandates there are an increasing amount of clinical data repositories (biological specimens and patient-level data). Much of these data are restricted-use and require data use agreements and controlled access environments. This class includes explanations of what makes data restricted-use, and the typical components of a restricted-use data request.

University of California San Francisco Library

UCSC Genome Browser Workshop

4 hours

The UCSC Genome Browser integrates information from a wide variety of genomic resources, including gene predictions; disease associations from resources such as HGMD, OMIM and locus-specific databases; gene-expression data; copy-number variation; comparative genomics; SNPs; HapMap data; microarray mappings; and histone- and DNA- modification data.

Please join us for this UCSC Genome Browser workshop in which Robert Kuhn, PhD, Associate Director, UCSC Genome Browser, will highlight new and somewhat obscure but useful features. Saving and sharing sessions, for example, is one of the most powerful, yet not widely appreciated features of the Browser. You can save the state of the Browser for future reference or as a stable pointer to share with others. A *very* new feature now allows you to share something interesting with *everyone*. A variety of Custom Tracks data types will also be demonstrated.

Another new feature he will highlight is the multi-region exon-only mode that allows you to suppress display of introns and intergenic regions. This is especially useful for interpreting RNA-seq data and whole-exome sequencing. Another of the multi-region modes allows display of selected genes that do not have to be contiguous on the genome.

Python/R Work Session

2 hours

Learning Python or R? Need dedicated practice time where you can ask questions? This 2-hour work session is designed to give people a chance to work on their Python/R projects alongside other beginners, and receive support from other users and our in-house experts. Bring your laptop, questions, and anything you are working on. Brought to you by the UCSF Library Data Science Initiative and the Institute for Computational Health Sciences.

Open and Reproducible Research Workshop

3 hours

Please join us for a free, hands-on workshop co-hosted by the Center for Open Science and UCSF Library. This workshop will teach you reproducible research practices, including how to use the Open Science Framework - a free, open-source tool for connecting and managing your research workflow.

These workshops are open to anyone engaged in research including students, faculty, and staff, and does not require any specialized knowledge of programming.

MAKERS POP-UP: Assemble a 3D-Printed Prosthetic Hand

1 hour

Interested in learning more about medical applications for 3D printing? Curious about how 3D printed prosthetics work? Join us in the Makers Lab for a 60-minute pop-up about the eNABLE prosthetic hands project. Then stick around after the demo and try assembling a prosthetic hand yourself.

Pivot (Funding): Class
1 hour

Are you looking for funding to support your research, training, fellowship, postdoc or program/curriculum development? Pivot is a comprehensive database of funding opportunities for all disciplines from federal agencies and private foundations in the US and international sources.

<…> will help you navigate through Pivot to easily explore various avenues of funding and set up an alert for updated funding opportunities.

Programmable Electronics Pop-Up
1 hour

Join us in the Makers Lab for a one-hour pop-up on programmable electronics, specifically using the Arduino. Haven't used an Arduino before? This is a great place to start! At its core, Arduino boards are able to read inputs (light on a sensor, a finger on a button, or a Twitter message) and turn it into an output (activating a motor, turning on an LED, publishing something online).

We will cover the basics of using the Arduino, discuss projects at UCSF currently using Arduinos, and we will even work our way through a few projects together using heat sensors, LED lights, fans, and motors.

We recommend bringing your Mac or PC laptop if possible to use with an Arduino during the pop-up, although it is not required. If you do bring your own laptop, you can download the Arduino software and Arduino codes ahead of time, and a Makers Lab staff member will be ready to help with the rest!

Remember, the Makers Lab is a work-in-progress and we welcome your suggestions for improving the space!

Tableau 101 - Visualize Your Data
2 hours

Are you interested in Tableau but don't know where to start? This class is designed to teach new Tableau users how to prepare their data, understand the Tableau interface, and create basic visualizations like bar charts and heat maps. We will be using the free version of Tableau, called Tableau Public, but the content will be applicable to users with a licensed version as well.

Programming and Pizza: Python/R
2 hours

Learning Python or R? Need dedicated practice time where you can ask questions? Like pizza? This 2-hour work session is designed to give people a chance to work on their Python/R projects alongside other beginners, and receive support from other users and our in-house experts.

Our Nov session will start with a short presentation from <…> on "Table work beyond Excel: Tools in Bash, Python Pandas, and R for biggish data." So if you have questions about working with data in tables (selecting rows, merging, grouping, sorting, calculating averages and variances) come and learn with us!

R/Python Programming Workshop - Software Carpentry
2 days

Are you looking for an introduction to programming in R or Python? Then this workshop is for you! This event, brought to you by the UCSF Library Data Science Initiative and the UCSF Institute for Computational Health Sciences, is a hands-on two-day workshop that will introduce you to the basics of programming in either R or Python as well as Git/GitHub and the Unix Shell.

This two-day workshop will run December 9-10, 9am-5pm each day. We will be running two concurrent bootcamps meaning both workshops are happening at the same time and you should register for either the R workshop or the Python workshop.

The schedule and content for the R workshop is here: https://michberr.github.io/2016-12-09-UCSF/

The schedule and content for the Python workshop is here: https://darencard.github.io/2016-12-09-ucsf_Python/

University of Michigan Taubman Health Sciences Library

Anatomage Table
By reservation

The Anatomage Table is the most technologically advanced anatomy visualization system for anatomy education and is being adopted by many of the world's leading medical schools and institutions. It has been featured in the TEDTalks Conference, PBS, Fuji TV, and numerous other journals for its innovative approach to anatomy presentation. The operating table form factor combined with Anatomage's renowned radiology software and clinical content separates the Anatomage Table from any other imaging system on the market.

The Anatomage Table was funded through the *Transforming Learning for a Third Century* Initiative, a collaboration between the U-M Library and Schools of Kinesiology, Dentistry, and Nursing.

Features

- An all-in-one, touch-interactive display system
- True human anatomy in a life-size scale
- Classroom and lab integration
- Clean, safe, reusable technology
- Radiological imaging workstation
- Available for use in the THL public space.

Navigating NCBI Molecular Data Using the Integrated Entrez System and BLAST: NCBI Discovery Workshop
3 hours

This workshop provides an introduction to the NCBI molecular databases and how to access the data using the Entrez text-based search system and BLAST sequence similarity search tool. You will learn the varied types of available molecular data, and how to find and display sequence, variation, genome information using organism sources (Taxonomy), data sources (Bioproject) and emphasizing the central role of the gene as an organizing concept to navigate across the integrated databases (Gene, Nucleotide, Protein, dbSNP and other resources).

A Practical Guide to NCBI BLAST: NCBI Discovery Workshop
3 hours

This workshop highlights important features and demonstrates the practical aspects of using the NCBI BLAST service, the most popular sequence similarity service in the world. You will learn about useful but under-used features of the service. These include access from the Entrez sequence databases; the new genome BLAST service quick finder; the integration and expansion of Align-2- Sequences; organism limits and other filters; re-organized databases; formatting options and downloading options; and TreeView displays. You will also learn how to use other important sequence analysis services associated with BLAST including Primer BLAST, an oligonucleotide primer designer and specificity checker; the multiple protein sequence alignment tool, COBALT; and MOLE-BLAST, a new tool for clustering and providing taxonomic context for targeted loci sequences (16S, ITS, 28S). These aspects of BLAST provide easier access and results that are more comprehensive and easier to interpret.

EDirect: Command Line Access to NCBI's Biomolecular Databases: NCBI Discovery Workshop
3 hours

The EDirect suite of programs allows easy command line access for searching and retrieving literature (PubMed) and accessing NCBI's biomolecular (Gene, Nucleotide, sequence databases, etc.) records. Its advantages include direct command-line access to NCBI's databases without writing Perl or Python scripts, construction of custom pipelines for processing data, built-in batch access, and the ability to generate highly flexible custom output reports. During the optional first hour of this workshop (9 – 10 AM), you will get a basic introduction to the Unix/Linux command line interface. The main workshop (10 AM – Noon) will cover how to use EDirect to set-up simple pipelines to retrieve and process data from PubMed, Gene, and the Nucleotide and Protein sequence databases. Access to EDirect installed in a Linux environment on a cloud service will be provided.

University of Pittsburgh Health Sciences Library System

Searching for Dollars: Grant Seeking to Support Research
1 hour

Want to know how to evaluate potential funders and keep up with their current grant priorities and opportunities? Then this is the class for you. This introduction to finding grant opportunities is especially appropriate for new investigators or those who want an overview of the research grant funding environment. The session covers both private (foundations, public charities) and government funders.

Crafting a Data Management Plan (webinar)
1 hour

What is a Data Management Plan? This session will answer that question, as well as describe the steps to creating a DMP, tools that can help with DMP development, and post-award management issues. University of Pittsburgh-specific guidelines and support resources will also be shared.

RNA-seq Analysis: CLC Genomics Workbench

3 hours

This hands-on workshop will provide an introduction to RNA-seq analysis using the library-licensed CLC Genomics Workbench. Participants will learn how to (1) align RNA-seq data to a reference genome, (2) calculate known genes and transcript expressions, (3) perform differential expression analysis, and (4) discover novel exons.

Please register for CLC Genomics Workbench.

Biomedical Genomics Workbench

2 hours

Biomedical Genomics Workbench is a comprehensive and accurate data analysis platform that enables you to find the signal in the noise in your cancer and hereditary disease NGS data. With its broad selection of end-to-end analysis workflows, tools, and visualization modules, it enables easy and accurate discovery, verification, and validation of novel disease biomarkers.

Discover novel insights with greater than 95% sensitivity and unsurpassed accuracy. Biomedical Genomics Workbench guides you through a complete analysis of your genome, exome, targeted amplicon, transcriptome, and epigenetic NGS sequencing data for results you can trust.

- Complete end-to-end and customizable analysis workflows for the comprehensive discovery, verification, and validation of novel biomarkers

- Specialized functionalities such as primer and primer-dimer removal for highly accurate targeted amplicon sequencing results

- High sensitivity detection of germline and low frequency variants from DNA-seq and RNA-seq data

- Unsurpassed accuracy for copy number detection in exome and targeted amplicon sequencing data

- Easy viewing of findings such as dynamic protein structures in 3D, and sequencing reads afford faster discovery

Please register for Biomedical Genomics Workbench.

QIAGEN Microbial Genomics

2.5 hours

QIAGEN Microbial Genomics Plug-ins expand upon CLC Genomics Workbench, the industry standard platform for bioinformatics computing. Plugins and modules add a layer of specialized tools and workflows to CLC Genomics Workbench, creating a comprehensive solution for microbial genomics and metagenomics data analysis, that include:

Microbiome profiling detects microbes and genes in metagenomic samples, and explores links between taxonomic or functional microbiome profiles and external factors like patient health or plant yield.

Microbial typing of isolates at the level of genes and whole genomes is useful for characterizing pathogens, or to provide quality control for valuable strains.

Outbreak analysis leverages whole genome information for pathogen typing, source tracking, and epidemiological outbreak investigation.

Please register for CLC Genomics Workbench.

The BD2K Guide to the Fundamentals of Data Science Series

1 hour (5 hours total)

Join the HSLS Data Management Group to watch a weekly virtual lecture series on the data science underlying modern biomedical research. We will be available before and after each Webinar to answer any data management related questions you might have.

Sponsored by the Big Data to Knowledge (BD2K) Initiative, this webinar series will consist of weekly presentations from experts across the country covering the basics of data management, representation, computation, statistical inference, data modeling, and other topics relevant to “big data” in biomedicine. It will provide essential training suitable for individuals at an introductory overview level.

9/23 Finding and accessing datasets, Indexing and Identifiers <…>

9/30 Data curation and version control <…>

10/7 Ontologies <…>

10/14 Provenance <…>

10/21 Metadata standards …>

Literature Mining: InfoBoosters, Molecular Databases, & F1000Workspace

3 hours

This hands-on workshop will cover literature searching software, molecular databases, and tools to help quickly mine these resources for pertinent information. Participants will learn how to identify the most appropriate scientific articles as well as annotate them with results fetched from gene, protein, disease, and drug-related databases.

Genome Navigation: UCSC Genome Browser

3 hours

This hands-on workshop will focus on a variety of genome biology resources. Participants will learn to (1) identify and retrieve whole genome sequence information by searching databases (NCBI Genome, Integrated Microbial Genome), (2) navigate genome sequences and extract information from annotated genome data (UCSC Genome Browser), (3) construct complex queries and retrieve large-scale genome data (Table Browser), and (4) create custom genome browser tracks from users’ own uploaded data (UCSC Custom Track tool).

Turning Lemons into Lemonade: Making Negative Research Results Useful

1 hour

Have you unwittingly repeated research that was never reported because its results were negative? Human nature inclines us to report only positive outcomes. This results in duplication of effort, time, and money spent. Yet the growth of options for publishing and disseminating negative results demonstrates their importance. Learn what you can do to turn your own negative into a positive.

Infographics: Communicating Information Visually

1 hour

Make your raw data more appealing and consumable with infographics. In this class, you'll learn what infographics are and how they can make data easier to understand and share. Then we'll create one together using an online resource.

Data Visualization for Beginners

1 hour

You've collected your data. Now what? In this class we will discuss the basic principles of data visualization and demonstrate a variety of visualization tools that you can use to establish the significance of your data.

Locating and Citing Research Data

1 hour

Need to find a dataset to act as a control for your study? Or do you want to reuse open access data? This class will offer tips for locating and citing data and include hands-on exercises to explore directories of data repositories and data journals.

Gene Regulation: TRANSFAC, NextBio, ENCODE

3 hours

This hands-on workshop provides an overview of resources and search strategies on transcriptional regulation. Emphasis will be given to databases, including TRANSFAC (transcription factors), Proteome (promoter sequences) and RegulomeDB (SNPs with known and predicted regulatory elements). Software that will be covered: BIOBASE Match to locate transcription factor binding sites present in a query sequence, UCSC Genome Browser to visualize the regulatory regions present in the human genome generated by the Encyclopedia of DNA Elements (ENCODE) project, and the HSLS licensed tool – BaseSpace Correlation Engine (formerly NextBio) to mine microarray/RNA Seq data from the Gene Expression Omnibus (GEO) repository.

Please register for the following tools: TRANSFAC/Match, Proteome, and Correlation Engine.

Variant Detection & Analysis: CLC Genomics Workbench, dbSNP, COSMIC, & more

3 hours

This hands-on workshop focuses on human genetic variations and cancer mutations. It covers identification of variants using CLC Genomics Workbench software and introduces variation databases (dbSNP, ClinVar, OMIM, DGV, PheGenI, HGMD, BaseSpace Correlation Engine (formerly NextBio), COSMIC, Broad Tumor Portal, ExAc Browser, RegulomeDb). The workshop will also teach how to use bioinformatics tools for functional analysis of mutations (EBI Variant Effect Predictor).

Please register for the following tools: CLC Genomics Workbench and Correlation Engine.

Pathway Analysis: NIH DAVID & IPA

3 hours

This hands-on workshop introduces open access and commercial biological pathway informatics tools. First we will learn how to mine a list of differentially expressed genes associated with a disease of interest by searching the NCBI Gene Expression Omnibus (GEO) using the library-licensed tool BaseSpace Correlation Engine (formerly NextBio). Then we will focus on uncovering the biology hidden behind the extracted gene list by searching protein-protein interactions and literature-curated gene/protein knowledge bases using pathway informatics software, including Ingenuity Pathway Analysis (IPA) and NIH DAVID.

Please register for the following tools: IPA and Correlation Engine.

Emory University Woodruff Health Sciences Center Library

SPSS Basics (via Lynda.com)

1.5 hours

Target Audience: Health Sciences Faculty, Students, Staff

Topics covered: navigate IBM SPSS Statistics interface, upload dataset into SPSS, recode, create, combine

Objectives: After the session, the attendee will be able to:

- demonstrate how to upload a data file into SPSS
- modify data and variables in SPSS
- create various charts and tables within SPSS with multiple variables

MetaCore: Enabling Systems Biology Research Through Pathway Analysis (Part I)

3 hours

MetaCore™ is an integrated curated knowledge database and software suite for pathway analysis of experimental data and gene lists. The scope of data types includes microarray and sequence-based gene expression, SNPs and CGH arrays, RNAi screens, gene variants, proteomics, metabolomics, Co-IP pull-out and other custom interactions which can all by analyzed in tandem. MetaCore™ is based on a proprietary manually-curated database of human protein-protein, protein-DNA and protein-compound interactions, metabolic and signaling pathways for human, mouse and rat, supported by proprietary ontologies and controlled vocabulary. The analytical package includes easy-to-use, intuitive tools for searching and data visualization, enabling the identification of the most relevant biological pathways, networks, and processes in our “virtual lab.”

Benefits:

MetaCore’s best-in-class molecular database now includes over 1.6 million interactions including over 700,000 compounds and their targets.

Simplify the generation and prioritization of experimental hypotheses following ‘omics’ or NGS experiments.

Assess and validate potential therapeutic targets and disease biomarkers.

Analytical tools include easy-to-use workflows and reports.

Agenda for 9:00 am-12:00 pm:

Introduction to pathway analysis with MetaCore

1. Introduction to MetaCore

2. Knowledge mining using EZ search and advanced search

3. Enrichment analysis of a single data set

4. Comparing multiple datasets in a single workflow

MetaCore: Enabling Systems Biology Research Through Pathway Analysis (Part II)

3 hours

MetaCore™ is an integrated curated knowledge database and software suite for pathway analysis of experimental data and gene lists. The scope of data types includes microarray and sequence-based gene expression, SNPs and CGH arrays, RNAi screens, gene variants, proteomics, metabolomics, Co-IP pull-out and other custom interactions which can all by analyzed in tandem. MetaCore™ is based on a proprietary manually-curated database of human protein-protein, protein-DNA and protein-compound interactions, metabolic and signaling pathways for human, mouse and rat, supported by proprietary ontologies and controlled vocabulary. The analytical package includes easy-to-use, intuitive tools for searching and data visualization, enabling the identification of the most relevant biological pathways, networks, and processes in our "virtual lab."

Benefits:

MetaCore's best-in-class molecular database now includes over 1.6 million interactions including over 700,000 compounds and their targets.

Simplify the generation and prioritization of experimental hypotheses following 'omics' or NGS experiments.

Assess and validate potential therapeutic targets and disease biomarkers.

Analytical tools include easy-to-use workflows and reports.

1:00pm-4:00pm

Advanced topics and workflows (choice of sessions)

Key Pathway Advisor – Hypothesizing key hubs using causal reasoning (~45 minutes)

Building networks with MetaCore (~60 minutes)

Constructing your own pathway maps and ontologies for enrichment (~60 minutes)

Using the Mircoarray repository for gene comparisons against public data (~45 minutes)

Analyzing and building networks with miRNA data (~45 minutes)

Analyzing multi-omics data (RNA-seq, proteomics, metabolomics, etc.) (~60 minutes)

MAXQDA 1-Day Boot Camp

6 hours

Please join us for our 1-day MAXQDA qualitative software bootcamp, led by <…>. Practice files will be provided to participants to use as they learn to create, navigate, code, manage, and analyze qualitative data.

Attendees will need to bring their personal laptops pre-loaded with the trial version of MAXQDA. Follow this link to sign up for the Demo Version:

http://www.maxqda.com/demo

An Introduction to Pathway Analysis Tools Available from the WHSC Library

1 hour

An overview of GeneGo's MetaCore, BioBase's Explain, and NextBio's pathway enrichment will be given. Objectives: Participants will be able to choose a Pathway analysis tool relevant to their research.

Finding Health Datasets for Secondary Analysis

1 hour

Topics Covered: locating health datasets for secondary analysis and formatting them for use in statistical software packages.

Objectives: After the session, the attendee will be able to:

- navigate the WHSCL Data Blog to locate Public Use Health datasets for analysis,
- analyze a dataset for relevancy and credibility of datasets by using the Secondary Data Analysis Worksheet

NSF-FUNDED SCHOOLS AND LIBRARY WORKSHOPS

University of Illinois Urbana-Champaign University Library

Intro to Crimson Hexagon

1 hour

Technology Services has secured access for campus to the social media listening and analytics service Crimson Hexagon. This service allows users to search social media via keyword queries and analyze the posts found through its built in tools or export the data to perform further analysis. This workshop will provide an introduction to the service and its capabilities.

http://www.crimsonhexagon.com/

Introduction to Text Mining with the HathiTrust Research Center Portal

1 hour

Students and researchers today have access to massive amounts of digitized text from the world's research libraries. Access to this growing digital record of human knowledge provides researchers with an unprecedented opportunity, but working with such material requires new tools to effectively analyze digitized text at so large a scale. This workshop will introduce cutting-edge software tools and cyberinfrastructure that are being developed at the Hathi Trust Research Center (HTRC) to meet these needs in the context of the digitized text collection of the Hathi Trust Digital Library, currently comprising more than 11 million digitized volumes.

https://analytics.hathitrust.org/

Data Rescue

1 hour

Have a mess of data floating around your computer or lab? Data management is an essential task for students and faculty but hard to get started. Go no further! In this workshop you will identify, group, and plan on how to reorganize your current and future data. We will discuss strategies for organization, folder structure, and create an organizational plan. An optional 30 minutes of extra lab time will be available for participants.

http://researchdataservice.illinois.edu/

Creating Data Documentation

1 hour

Writing project, code, and data documentation doesn't need to be the worst part of your day. This hands on workshop will give you experience using various types of documentation, discuss strategies for writing documentation, and get you started writing a template for your projects. Bring a dataset you'd like to work with but examples will be provided. An optional 30 minutes of extra lab time will be available for participants.

http://researchdataservice.illinois.edu/

GIS for Research I: Introduction to GIS Concepts, Software, and Data

2 hours

Not sure what GIS is or how it is used? This first workshop of the GIS for Research series will start you down the path to use geospatial technologies in your research by guiding you through the foundational concepts of GIS and how to think spatially. We will introduce different types of GIS software to discover geospatial data and learn about key concepts like vector vs. raster data, scale, and projections. We will also discuss resources available across campus and the web to help you utilize GIS for your research.

GIS experience needed: None

Analyzing and Interpreting Images

1 hour

Many of us agree with the veracity of the phrase “a picture is worth a thousand words,” yet it can often be difficult to analyze an image and discern its intended meaning. “Reading” images and visual media can be subjective, and multiple meanings can be transmitted through a visual work. How does one deconstruct various pictorial, graphic, technical, historical, cultural, and design components of a visual work in order to glean meaning? How does accompanying text or the source of a visual work affect one’s understanding of an image? Image interpretation and analysis are key components of the research and selection process, as an image must be understood accurately in order for it to be effective in a scholarly context. This workshop will provide an interpretative and analytical framework for evaluating and selecting visual media.

Finding and Using Census Data

1 hour

The Census Bureau provides socioeconomic and demographic data for large and small geographic areas in the U.S. Learn how to find everything from median income for your block to commuting time for every county in the U.S., as well as how to download data so that you can analyze it in Excel or statistical software.

Preparing for Data Sharing

1 hour

Making research data public is becoming a reality for many disciplines, but for many researchers and disciplines there is a complicated set of issues to consider before publication or sharing of data. This workshop will cover the basics steps of research data sharing & publication, from initial considerations to depositing. Participants will work through guidance to help them make decisions about when and how to publish or share data. An optional 30 minutes of extra lab time will be available for participants.

http://researchdataservice.illinois.edu/

GIS for Research II: GIS Research, Data Management, and Visualization

2 hours

The second GIS for Research workshop will help you build a solid foundation for framing your research to utilize GIS to its full potential. We will dive deeper into GIS software to uncover tools to help organize, manage, and visualize your data. This workshop will be mostly hands-on with GIS software.

GIS experience needed: None to Beginner

http://guides.library.illinois.edu/gis

Advanced Text Mining Techniques with Python and HathiTrust Data

1 hour

This workshop builds upon the Introduction to Text Mining with the HTRC Portal workshop: In this session, attendees will learn how to do moderately advanced text mining analysis approaches using the Extracted Features datasets generated from the HathiTrust Research Center and applying Python scripts to the data. We recommend that attendees be familiar with command line interfaces, though it is not required.

GIS for Research III: Geoprocessing, Analysis, and Web GIS

2 hours

Take your GIS skills to the next step! The final GIS for Research workshop will walk through different geoprocessing tools and analyses common in GIS for research. Additionally, an emphasis on sharing and visualizing GIS data on the web will challenge students to think differently about GIS data. This workshop will be mostly hands-on with GIS software.

GIS experience needed: Beginner to Intermediate

http://guides.library.illinois.edu/gis

Qualitative Data Analysis with ATLAS.ti

2 hours

Goals for the workshop:

Learn about the methodological principles behind ATLAS.ti.

Open ATLAS.ti software and import textual documents.

Understand the layout of the ATLAS.ti environment.

Learn the fundamental functions of ATLAS.ti for data description, exploration, analysis, and interpretation.

Learn how to print, save, and export output.

Smart and Simple Data Management

1.5 hours

This session aims to provide you with data management best practices and tools to increase your research efficiency and impact. We'll present a basic introduction to data management using a data management plan framework, hands on activities, and discuss how to find and vet resources for making data publicly accessible.

http://researchdataservice.illinois.edu/

Sharing Research with Story Maps

1 hour

Could a map enhance how you communicate your research to diverse audiences? Maps are becoming more accessible to the general public through the use of online mapping applications such as ArcGIS Online and Google Maps. This workshop will explore different examples of story maps and get you started creating your own. Prior to attending this workshop, students will need to sign up for an ArcGIS Online Public account at

GIS Experience needed: None

http://guides.library.illinois.edu/gis

Messy Data? Clean it up with OpenRefine!

1 hour

Join us for a workshop which introduces OpenRefine, a free, open source tool used to organize, clean up, and transform your data. We will provide an overview of the browser based application, as well as use cases that show the benefits of working with your data in OpenRefine, and demonstrate the basic functions to get you started on cleaning up your data.

http://openrefine.org/

Project Workflow Mapping

1 hour

Workflow mapping is a useful tool for teams of all sizes to understand how data, code, and other resources are being shared and passed around. Like retracing your steps after losing something, tracing a project through workflows identifies all the essential products and dependencies of the research process, and can be one of the most useful places to get started with data management. This workshop focuses specifically on large research processes, but can be adapted to most types of projects, including computational data project.

http://researchdataservice.illinois.edu/

Computational Data Workflow Mapping

1 hour

Workflow mapping is a useful tool for teams of all sizes to understand how data, code, software, and other resources are being shared and passed around. Like retracing your steps after losing something, tracing a project through workflows identifies all the essential products and dependencies of the research process, and can be one of the most useful places to get started with data management. This workshop focuses specifically on computational research processes, but can be adapted to most types of projects.

http://researchdataservice.illinois.edu/

Making Scanned Text Machine Readable through Optical Character Recognition

1 hour

Optical Character Recognition (OCR) is a process that converts scanned images and documents into editable, searchable formats. OCR helps your computer to recognize letter shapes in a scanned document and turn them into text you can copy and edit as needed. This allows you to extract information from documents quickly and easily. OCR also enables these texts to be used in key data and text mining projects. This workshop will give attendees a basic understanding of how to make use of optical character recognition software, including Adobe Acrobat Pro, ABBYY FineReader, and Tesseract, in their research, as well as give them a chance for hands-on experience with these programs in the Scholarly Commons.

Collecting Geospatial Data

2 hours

Collecting geospatial data in the field or with surveys is easier than ever before with new web and mobile tools. This workshop will provide an overview of the myriad of different devices and tools available to collecting geospatial data, included GPS devices, mobile devices, Collector for ArcGIS, and Survey123, among others. We will cover how to design a data collection workflow and will conclude with a hand-on geospatial data collection activity in the field (come prepared to be outside, weather permitting). Please bring your own iOS or Android device and install the Collector for ArcGIS and Survey123 apps prior to the workshop.

GIS experience needed: None to Advanced

http://guides.library.illinois.edu/gis

University of California Berkeley Library

NCBI Bioinformatics Tools: An Introduction

1 hour

A hands-on workshop introducing NCBI bioinformatics tools such as PubMed, Gene, Protein, Nucleotide, and BLAST.

The workshop will cover selecting the proper tools for your question, navigating through the interlinked NCBI databases, and saving your results.

The Bash Olympics: The Hacker Within

1.5 hours

[none]

Machine Learning for Kaggle Competitions with R: The Hacker Within

1.5 hours

Kaggle is a data science platform where data scientists from all over the world work together and compete in real-world machine learning challenges. These public data sets cover a wide array of interesting problems from diagnosing eye problems based on images of the retina to recommending coupons to users who visit a site. On Tuesday, we will explore the machine learning process in the context of competitions and how Kaggle is becoming a really good starting point for machine learning enthusiasts to collaborate and learn new things.

D3.js: The Hacker Within

1.5 hours

D3.js for building Exploratory Visualization Tools

Github Pages and Jekyll: The Hacker Within

1.5 hours

Github Pages is a free web hosting service by Github, which uses Jekyll to generate HTML files from files (themes, layouts, and data) in a special Github repository. Whenever you make a commit to a Github Pages repository, Github's servers run the Jekyll parser on the files in that repository, which generates a set of static HTML and CSS files on a special subdomain. The result can look nearly identical to traditional content management systems (like Wordpress or Drupal) that dynamically process requests from browsers using languages like PHP and querying live databases like MySQL.

Finding Health Statistics and Data: A Hands-on Workshop @D-Lab

1.5 hours

Participants in this workshop will learn about some of the issues surrounding the collection of health statistics, and will also learn about authoritative sources of health statistics and data. We will look at tools that let you create custom tables of vital statistics (birth, death, etc.), disease statistics, health behavior statistics, and more. The focus will be on U.S. statistics, but sources of non-U.S. statistics will be covered as well.

Whether you need a quick fact or a data set to analyze, this workshop will lead you to relevant data sources. Students will have a chance to explore some of these tools in class, so please bring your laptop.

Natural Language Processing for Python with NLTK: The Hacker Within

1.5 hours

NLTK. Text data requires a separate preprocessing stage often referred to as the 'NLP pipeline'. One popular library for its implementation is Python's NLTK (Natural Language Toolkit). This talk will cover how to clean text data, tag parts of speech (POS), identify named entities (NER), and quantify sentiment beyond dictionary look-up. While not explored in this talk, these preprocessing steps are often critical to developing more advanced, high-level models for document classifiers, topic modeling, and network models by providing targeted feature sets.

Data Visualization

1.25 hours

A well-designed figure can have a huge impact on the communication of research results. This workshop will introduce key principles and resources for visualizing data:

- Choosing when to use a visualization

- Selecting the best visualization type for your data

- Choosing design elements that increase clarity and impact

- Avoiding visualization issues that obscure or distort data

- Finding tools for generating visualizations

- Concepts and guidelines to follow when creating maps

Parallelization in Python: The Hacker Within

1.5 hours

[no description]

Digital Humanities for Tomorrow

2 hours

Join a conversation among digital humanities researchers and stakeholders about the future of your DH research. How are you currently preserving (or not) your own DH work? How do you maintain it? What will happen to your work after you move on? What does DH preservation look like? What is most important for you in ensuring your work is available to forthcoming scholars? How can the University Library and DH group support your goals?

The Python Olympics: The Hacker Within

1.5 hours

The fastest way to learn a programming language is to use it. So why not turn that into a game? All levels of experience welcome. We have Python puzzles for advanced coders and beginners alike. This will also be the world debut of a new kind of interactive IPython Notebook designed for group coding games. See you at the games!

Physical Computing: The Hacker Within

1 hour

[no description]

matplotlib: The Hacker Within

1.5 hours

matplotlib presentation through notebook demos.

Code to install (if you use Anaconda, use conda install instead of pip install):

```
pip install matplotlib

pip install Basemap

pip install mpld3

pip install folium

pip install bokeh
```

Introduction to matplotlib: Jupyter Notebook with example code

Using Basemap to plot geospatial data and other tricks/tools using matplotlib ("what used to bug me about using matplotlib, but doesn't anymore"): Jupyter Notebook with example code

Scrivener: Software for Writers Workshop

1 hour

Want a better way to tackle your long writing project? Scrivener can help! Scrivener is a software program that breaks down your writing into manageable "chunks" and keeps all of your research, brainstorming, and writing in a single conceptual workspace. Use Scrivener for your thesis, dissertation, book project, novel, or any longer writing project. Read more about Scrivener.

RStudio: The Hacker Within

1.5 hours

[no description]

Ensemble (Machine) Learning with Super Learner and H2O in R: The Hacker Within

1.5 hours

[no description]

Columbia University Libraries

R Open Lab

2 hours

Interested in exploring real-world data using R techniques? Join us in learning the capabilities of R, an open-source statistical programming language. We'll use data analytics in a process to understand the world better by using real world data.

Intro to Cartographic Design

1 hour

The DSSC Workshop series will offer Introduction to Cartographic Design a workshop led by <…>, GIS/Metadata Librarian. This workshop is geared towards students taking introductory GIS courses who are in the process of producing project presentations and papers as the end of the semester approaches. The two hour workshop will give an overview of some major cartographic concepts, including: - Communicating your message effectively - Creating clear, balanced layouts - Symbolization - Labeling - Use of color - Map elements Registration is not limited to current students in GIS courses. However, there is an expectation that attendees have had some level of exposure to GIS software.

Bloomberg Basics

1 hour

This workshop will provide an overview of the most popular and helpful tools available via the Bloomberg Terminal. It will cover Stock Overview, Historical Filling, Relative Valuation, Earning and Estimates, and more.

Python Open Labs: Programming Made Easy

2 hours

Join us in exploring the basics of programming via Python, one of the most widely used programming languages and indispensable tool in almost every field.

Map Club Workshop — D3.JS Part II

1.5 hours

This guided session explores the geovisualization capabilities of D3.js (https://d3js.org/), a JavaScript library for manipulating documents based on data. Map Club is a series of fast-paced hack sessions geared toward the rapid acquisition of skills in geospatial technology led by <…>, DSSC Spatial Research Intern. Each session provides an informal and fun opportunity for the exploration of a web-based library or framework. Sessions will be loosely divided into three phases: background and setup, self-paced making, and sharing.

HPC Open Lab

2 hours

Is computer processing time holding back your work? Are your projects running into the limits of your machine's capacities? High-Performance Computing (HPC)--sometimes called supercomputing--delivers more efficiency and accuracy than your typical workstation by aggregating computing power, allowing you to run more complex experiments and solve larger problems. Drop by to learn more. Staff from Research Computing Services are on call to help with projects, general questions, or just to get started!

Massachusetts Institute of Technology Libraries

Get to Know Yewno: A Demo

1 hour

Curious about Yewno? Come to a demo and Q&A on this new kind of discovery tool that uses full text analysis and machine learning to create a visual, interactive map of connected concepts. With snacks!

Data Management Planning & the DMPTool

1 hour

Are you required to submit a data management plan (DMP) to a funder? Are you looking to create a data management plan and aren't sure where to start or what to include? This session will run through the components of a good data management plan and introduce the DMPTool, an online (and MIT-customized) tool for crafting funder-specific data management plans.

Look under the Hood of Yewno

1.5 hours

Join us for this interactive session presenting an in-depth look at Yewno's technology and how it works. Covering the technology underpinning the Yewno platform, including the use of machine learning and computational linguistics and how the results of this innovative approach can positively impact research outcomes. Includes lunch with registration!

Introduction to GIS

3 hours

We will introduce open source and proprietary Geographic Information System (GIS) software options and let attendees choose to work through exercises using ESRI ArcGIS (proprietary) and/or Quantum GIS (QGIS) (open source).

Learn to work with data from the MIT Geodata Repository, analyze the data, and create maps that can be used in reports and presentations.

GIS Level 2

3 hours

Expand your knowledge of GIS as you learn how to use several analysis tools.

Expand your experience with GIS software, and learn how to create and edit GIS files, re-project data, and use tools like Clip, Buffer, and Spatial Join. We will use both QGIS and ArcGIS.

Previous experience with GIS software is required, such as taking the Introduction to GIS workshop.

Patent Searching Fundamentals

1 hour

Demystify the patent literature and learn resources for finding patents.

This session will enable you to successfully find patent references from all over the world and obtain patent text and diagrams. This hands-on session will help de-mystify the patent literature and show key resources for finding patents.

Introduction to R

3 hours

Learn the basics of conducting statistical analyses in R.

Get an introduction to R, the open-source system for statistical computation and graphics. With hands-on exercises, learn how to import and manage datasets, create R objects, install and load R packages, conduct basic statistical analyses, and create common graphical displays. This workshop is appropriate for those with little or no prior experience with R.

Data Management: File Organization

1 hour

Don't struggle with organizing your research files any longer!

Do you struggle with organizing your research data? Wonder if there’s a better way to arrange and name your data files to optimize your work? This workshop will teach you practical techniques for organizing your data files. Topics will include: file and folder organizational structures and file naming. The session will include hands-on exercises to apply the concepts to your particular data project.

Basic R Programming for Data Analysis

3 hours

Expand your use of R by learning simple programming techniques.

This hands-on, intermediate course will guide you through a variety of programming functions in the open-source statistical software program R. It is intended for those already comfortable with using R for data analysis who wish to move on to writing their own functions.

To the extent possible, this workshop uses real-world examples. Concepts will be introduced as they are needed for a realistic analysis task. In the course of working through a realistic project, learn about interacting with web services, regular expressions, iteration, functions, control flow, and more.

Prerequisite: basic familiarity with R, such as acquired from an introductory R workshop.

Introduction to Python for GIS

3 hours

Learn enough Python to participate in the map creation workshop on Day 2.

This is the first day of a two-part workshop. You'll learn just enough Python scripting to work with it in ArcGIS and feel comfortable in Day 2 of the workshop, which focuses on using Python to automate map creation.

Python for Map Creation in ArcMap

3 hours

Automate map design in GIS.

This workshop will focus on using Python to automate map making. With the ArcPy mapping module you can easily create and update map layers and content to create customized maps that can be exported for presentations and reports. Whether you need to create 10 or 1,000 maps, you'll learn how to save time by using Python.

Prerequisite: Knowledge of basic Python commands.

University of Texas at Austin Perry-Castañeda Library

Introduction to R

2 hours

R is a free programming language designed originally to run statistical analyses and visualize data, but it can do much more. This introductory workshop will provide basic exposure to R for individuals with little to no previous experience with coding. The goal is to provide enough experience so that attendees are comfortable enough to continue learning independently after the workshop. We will cover: Coding in R with RStudio, basic programming concepts, interacting with data, and visualizing data with ggplot2. There will be time for questions and answers.

Web and Social Media Data Analysis

1.5 hours

Websites and Social Media Platforms are veritable treasure troves of textual data that can tell us so much about language and society. This workshop will provide you with an overview of a select range of tools you might use for scraping webpages and social media sites to collect and use that data. This workshop is intended for people new to web scraping and looking for how to get started. In the second half of the workshop, you will be
collecting data from Twitter.

Writing a Data Management Plan

1 hour

The rising tide of data sharing requirements from funding agencies, publishers, and institutions has created a new set of pressures for researchers who are already stretched for time and funds. While navigating these requirements can feel like yet another set of painful hurdles, in reality, the process can be a hugely useful exercise that can add value to your research data and increase the impact of your work. This workshop will present an overview of the current landscape of funder mandates, introduce you to the custom templates in the DMPTool, and give you some useful tips for writing a successful plan.

Managing Research Data (Part 1): A Guide to Good Practice.

1 hour

Dealing with the mountains of digital data that can accumulate in the course of a research project can seem like a daunting process, especially if your work is collaborative or stretches over several years. Adopting a few key good habits early on can save you huge amounts of time, money, and frustration searching for things and recovering lost files. In this one-hour workshop, we will introduce core data management concepts, offer top tips for things like backups and file formats, and share information about useful tools and resources available to UT faculty, staff, and students.

Introduction to OpenRefine

1.5 hours

OpenRefine is a powerful, free, and easy-to-use tool that's perfect for sifting through, cleaning up, and transforming datasets--tedious yet crucial tasks for data analysis and sharing. In this workshop, you will learn how to leverage OpenRefine's point-and-click interface and intuitive scripting language for basic data exploration and bulk transformations. No prior knowledge necessary.

CONTENT ANALYSIS

METHODS

To identify top-funded universities, we used NIH Research Portfolio Online Reporting Tools and the NSF Budget Internet Information System. From corresponding websites, we extracted information about workshops (n = 99) offered by health sciences libraries (n=5) and main libraries (n=5) in schools funded by NIH and NSF, respectively.

We included workshops if the title or description mentioned *data* AND (r*esearch* OR *analysis*). Lexical variants were allowed as indicators of relevance, such as *computational project, study,* or *analytics*. Furthermore, workshops had to have been offered in September, October, November, or December 2016.[1]

We excluded workshops about (a) searching databases, e.g., MEDLINE, CINAHL, or EMBASE; (b) using bibliographic software, e.g., EndNote or Zotero; (c) assessing one's research impact; (d) makers' spaces or labs, unless they dealt with medical or scientific applications; (e) literature review methods, including systematic reviews; (f) grant writing; (g) copyright issues; and (h) personal productivity.

Workshops were catalogued by library and source of federal funding for the university, along with title, duration, and description. We used NVivo 11 Pro (QSR International) for qualitative data analysis and TExtract® (TEXYZ) for semi-automated indexing of the catalog content. Themes were automatically identified in textual patterns and then were refined by an analyst. Thematic overlap was described across funding source. Additionally, we identified themes unique to each subset.

RESULTS

Main libraries in NSF-funded schools offered 36% more workshops than health sciences libraries in NIH-funded schools (57 vs 42) (Table 1).

Overall workshop duration ranged from 1 hour to 2 days; mode = 1 hour. Duration for health sciences library workshops was typically longer than for main libraries (NIH mode = 3 hours; NSF mode was greater than 1 hour and less than 3 hours) (Figure 1).

[1] If a list of workshops could not be found via a library website, the library of the next school in the NIH or NSF list was examined. Libraries might not offer workshops if other university organizations support faculty, researchers, and students interested in data-driven research. If so, patrons could be redirected to those organizations. In this study, we focused on workshops offered by libraries.

We identified 15 main themes overall: coverage was maximal for *data visualization* and *statistical programming*, and minimal for *writing* and *natural language processing (NLP)*. Thematic distributions varied with funding source. For example, coverage of *bioinformatics* was greater in the NIH-funded subset; coverage of *statistical programming* was greater in the NSF subset. Main themes unique to the NIH subset included *reproducible research, makers' labs,* and *finding funds for research.* Themes unique to the NSF subset included *writing*, *natural language processing (NLP),* and *machine learning* (Figure 2)*.*

A hierarchical block chart displays the relative coverage of themes and subthemes (Figure 3). For example, *statistical programming* has the largest block with 11 subthemes, mainly of tools and software (see upper left corner). The smallest blocks have no subthemes, such as the blocks for *open science* and *big data* (see lower right corner)*.*

The 20 most informative candidate terms for the Subject Index were identified after sorting and discretizing into 7 quantiles by source of university funding (Table 2). For example, top terms from the NIH subset included *data visualization*, *pathway analysis of experimental data*, and *data management.* Top terms for the NSF subset included *data analysis*, *data management plan (DMP)*, and *analysis tools*. Six terms occurred in both subsets (indicated by bold italics in Table 2). The percentage of term overlap was 18% (6/34), where % overlap is defined as the number of indexing terms occurring in the intersection divided by the number occurring in the union x 100.

Table 1. Library by source of university funding sorted by number of workshops

Library in top NIH-funded university	N workshops
University of Michigan Taubman Health Sciences Library [a]	4
Emory University Woodruff Health Sciences Center Library	6
Johns Hopkins University Welch Medical Library	8
University of California San Francisco Library	9
University of Pittsburgh Health Sciences Library System	15
Subtotal	*42*
Library in top NSF-funded university	
University of Texas at Austin Perry-Castañeda Library	5
Columbia University Libraries	6
Massachusetts Institute of Technology Libraries	11
University of California Berkeley Library	16
University of Illinois Urbana-Champaign University Library	19
Subtotal	*57*
Total	*99*

[a] The count for the University of Michigan library may be an underestimate because the online calendar disallowed retrospective searching.

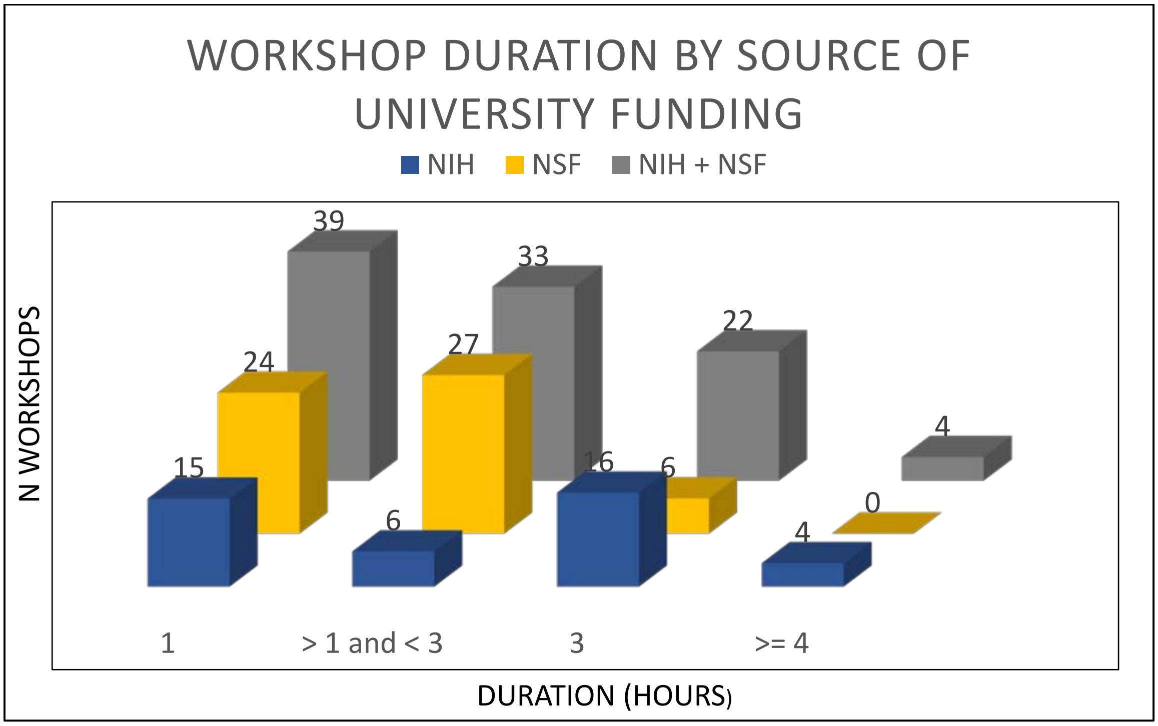

Figure 1. Workshop duration by source of university funding. The distribution of duration varies across top NIH- and NSF-funded universities (minimum = 1 hour; maximum = 2 days).

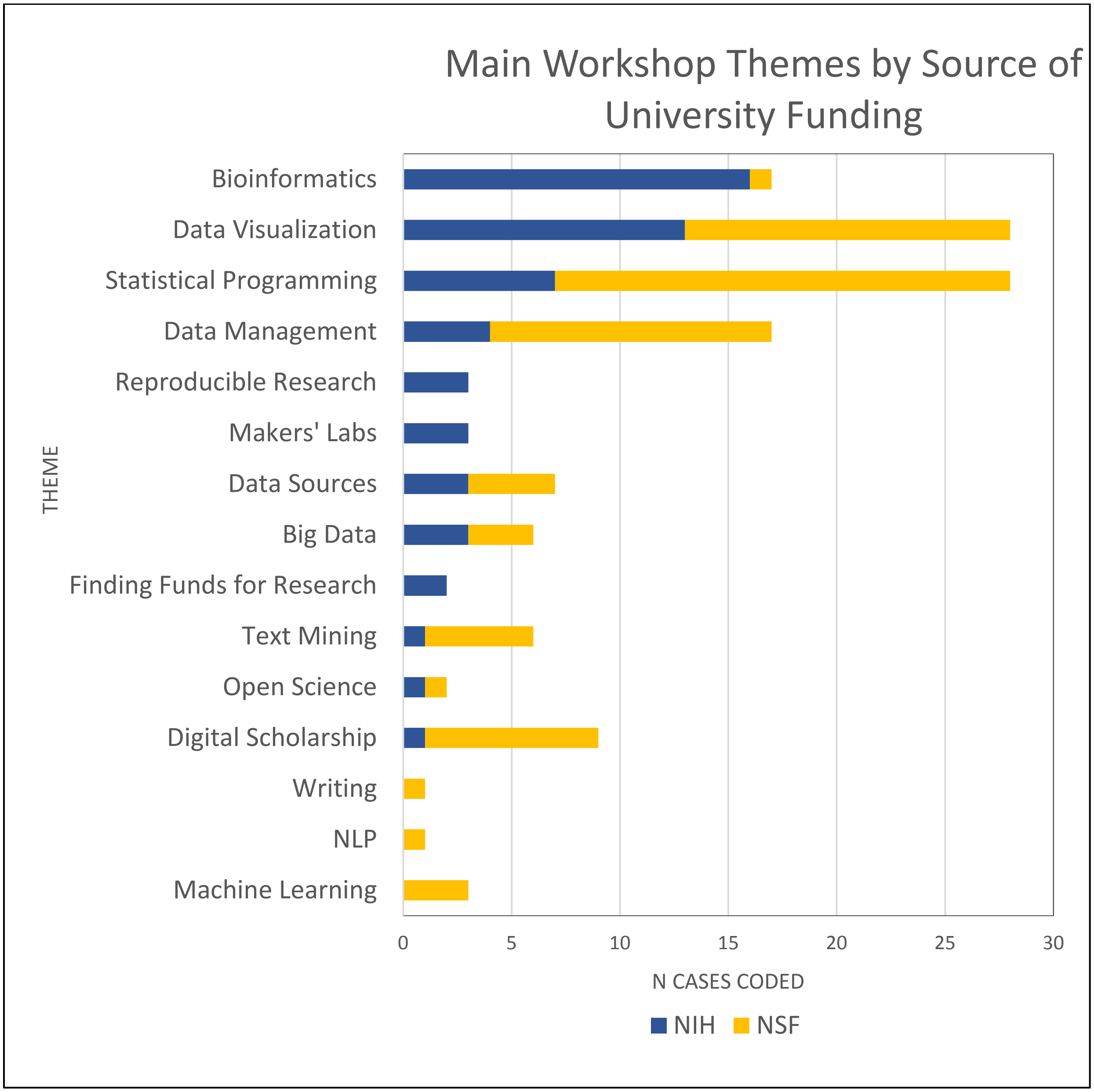

Figure 2. Main workshop themes vary with source of university funding. Coverage of *data visualization, data sources,* and *big data* is about the same across funding subsets, whereas coverage for the remaining themes varies quite a bit.

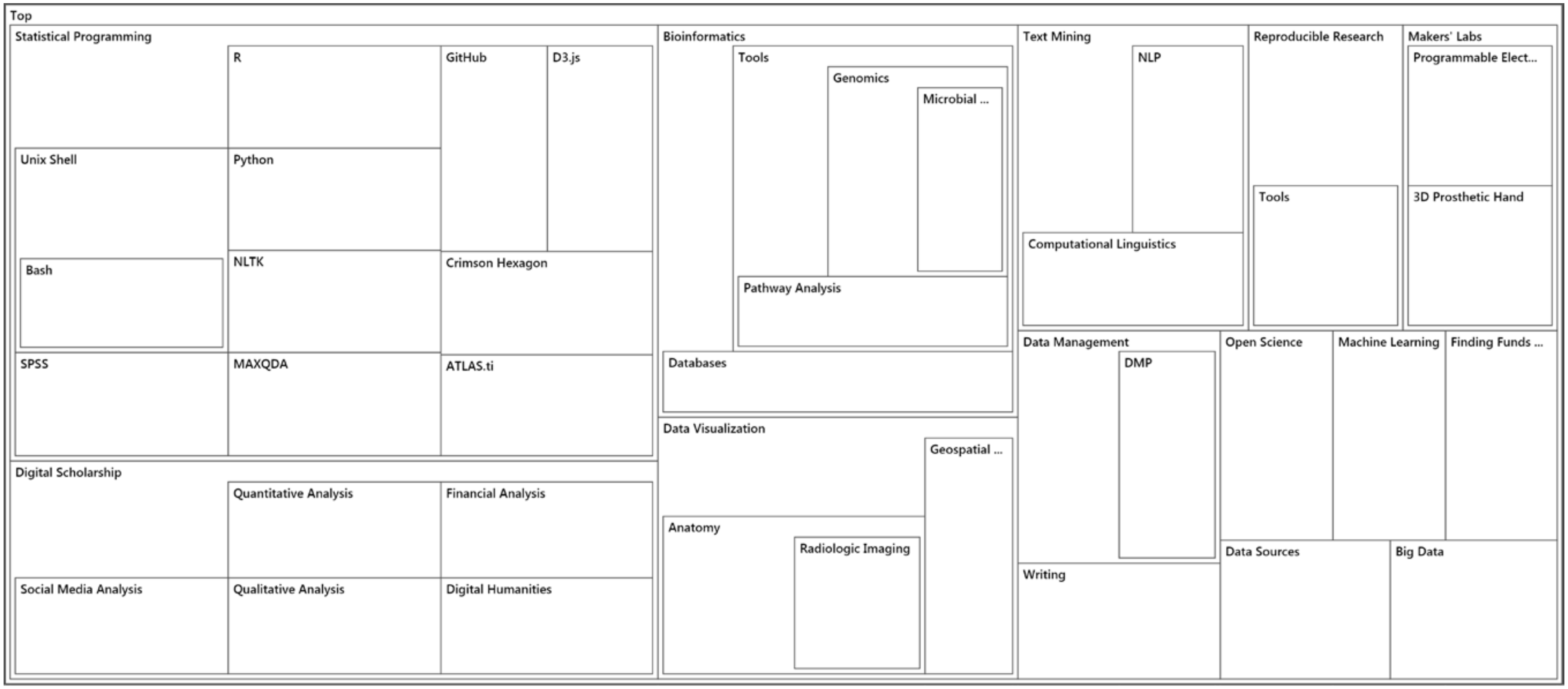

Figure 3. Hierarchical block chart comparing relative coverage of workshop themes and subthemes. Smaller blocks nested in larger blocks display subthemes.

Table 2. Most informative indexing terms by source of university funding

NIH (n = 20 terms)	NSF (n = 20 terms)
data visualization	data analysis
pathway analysis of experimental data	***data management plan (DMP)***
data management	analysis tools
datasets	data collection workflow
locating health datasets for secondary analysis	data files
molecular databases	***data management***
RNA seq data	data project, computational
comprehensive database of funding	finding tools for generating visualizations
data sources	***geographic Information system (GIS)***
enabling systems biology research	geoprocessing tools and analyses
finding health datasets	hacker
funding	***projects***
genes	Python puzzles for advanced coders and beginners
makers lab	social media data analysis
navigating NCBI molecular data	code
participants	***data visualization***
projects	demographic data
searching databases	geospatial data, plot
data management plan (DMP)	optical character recognition (OCR)
geographic Information system (GIS)	***participants***

Note: Informative candidate terms for the Resource Index are omitted. Shared candidate terms for the Subject Index are in ***bold italics***.

DISCUSSION

In 2012, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) Initiative (https://datascience.nih.gov/bd2k/about). This trans-NIH effort recognizes, in part, the need for an educated workforce to cope with the onslaught of data in its many forms.

Although the stream of 'big data' presents many challenges for basic, clinical, and translational research, less-than-massive datasets also present challenges—mainly because data gleaned from our digital ecosystem may be partly structured or unstructured, as well as heterogeneous (e.g., numeric, textual, audio, and visual data).

In addition to knowledge of how best to design and analyze computational projects, investigators need to be conversant with methods of other disciplines. This is because federally-funded research is carried out by distinctly multidisciplinary teams of scientists. For example, teams might consist of computer scientists, engineers, informaticians, biologists, and clinicians. To further promote multidisciplinary research, the National Science Foundation (NSF) is partnering with NIH and many other federal agencies and organizations (https://www.nsf.gov/about/partners/fedagencies.jsp).

In response to this urgent need for computationally-savvy researchers, leading university libraries are developing and offering participatory workshops to introduce their patrons to data-driven research methods and resources. In doing so, libraries help to improve the competitive advantage of their universities relative to other schools that compete for federal funds.

Developing a set of relevant and useful workshops requires strategic planning. We offer the following suggestions:

Before you develop new workshops, **conduct an environmental scan** of your university to identify participatory training opportunities in data-driven research other than credit-bearing courses, which are too long or expensive to jump-start new research endeavors. The value of a library's roster of workshops is in offering gateway opportunities to their patrons. This is especially true for researchers and faculty who are unlikely to take a course, but who need to acquire new skills. If other groups on campus cover main themes identified in this report, consider collaborating to reduce wasted effort.

Identify your gaps in coverage of data science topics. First compare your roster of workshops to those offered by other schools in the subset you most closely resemble. Initially, pay particular attention to the themes well covered by competitor schools in this study. If these themes are well covered by your library, consider developing workshops poorly covered in your subset, i.e., covered by just a few libraries.

Next, based on informative indexing terms occurring in both subsets, **consider setting as a priority the following themes:** *data visualization (including GIS)* and *data management (including data management plans or DMPs)*.

To **differentiate your school from your competitors** and to prepare your patrons for the extreme multidisciplinarity of team science, consider themes covered in the other subset. To the extent that team science and data-driven research are responsive to collaborative funding opportunities across NIH and NSF, themes in the other subset should not be ignored.

Finally, regularly evaluate whether your workshops cover emerging topics in the digital ecosystem. Figure out how to **demonstrate the value your library adds** to the university's research enterprise. Widely disseminate evidence of your value across the university.

RESOURCE INDEX

Resources mentioned in the catalog of workshops are indexed.

SUBJECT INDEX

Subjects mentioned in the catalog of workshops are indexed.