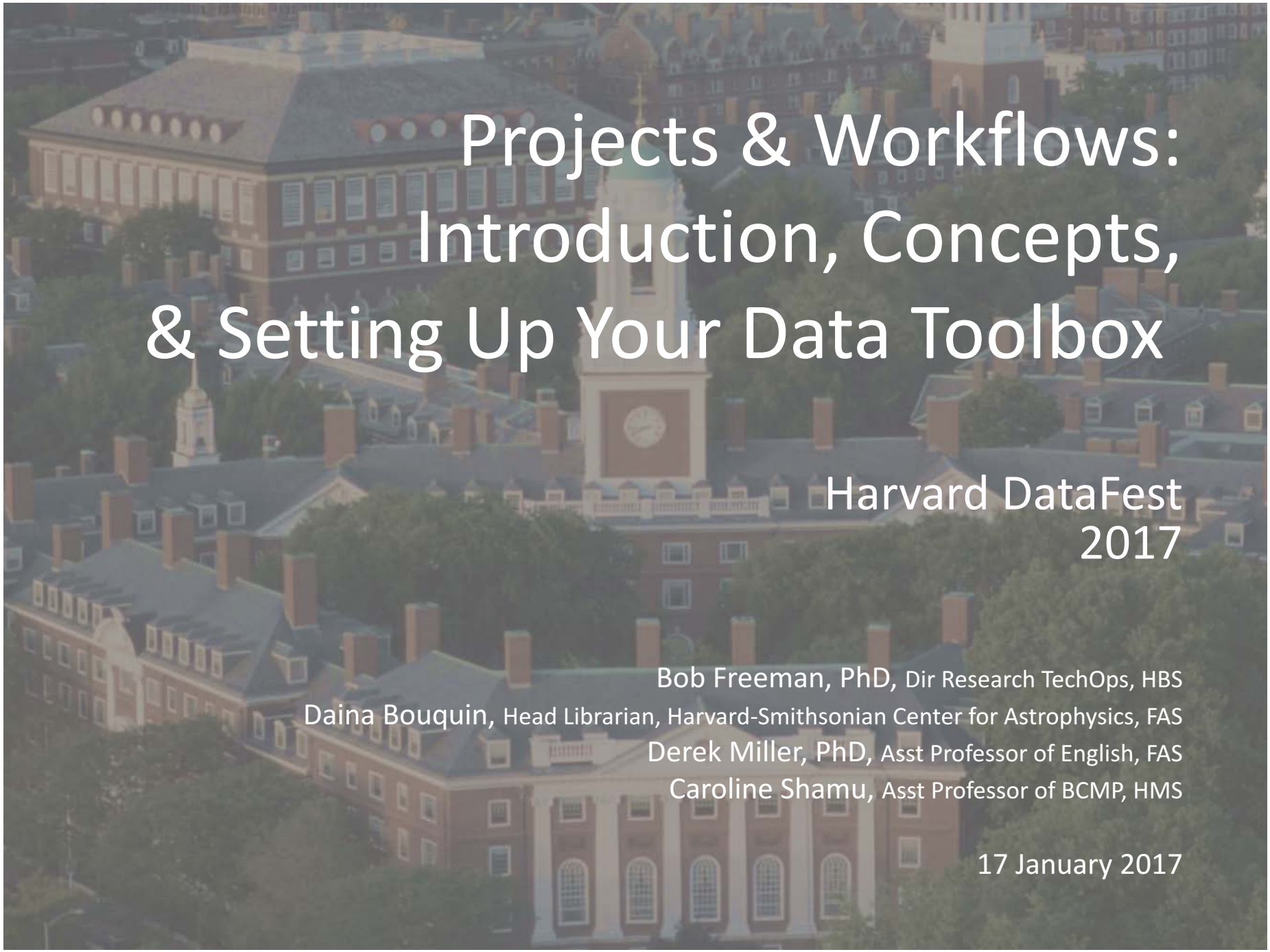


Projects & Workflows: Introduction, Concepts, & Setting Up Your Data Toolbox

Harvard DataFest 2017

Bob Freeman, PhD, Dir Research TechOps, HBS
Daina Bouquin, Head Librarian, Harvard-Smithsonian Center for Astrophysics, FAS
Derek Miller, PhD, Asst Professor of English, FAS
Caroline Shamu, Asst Professor of BCMP, HMS

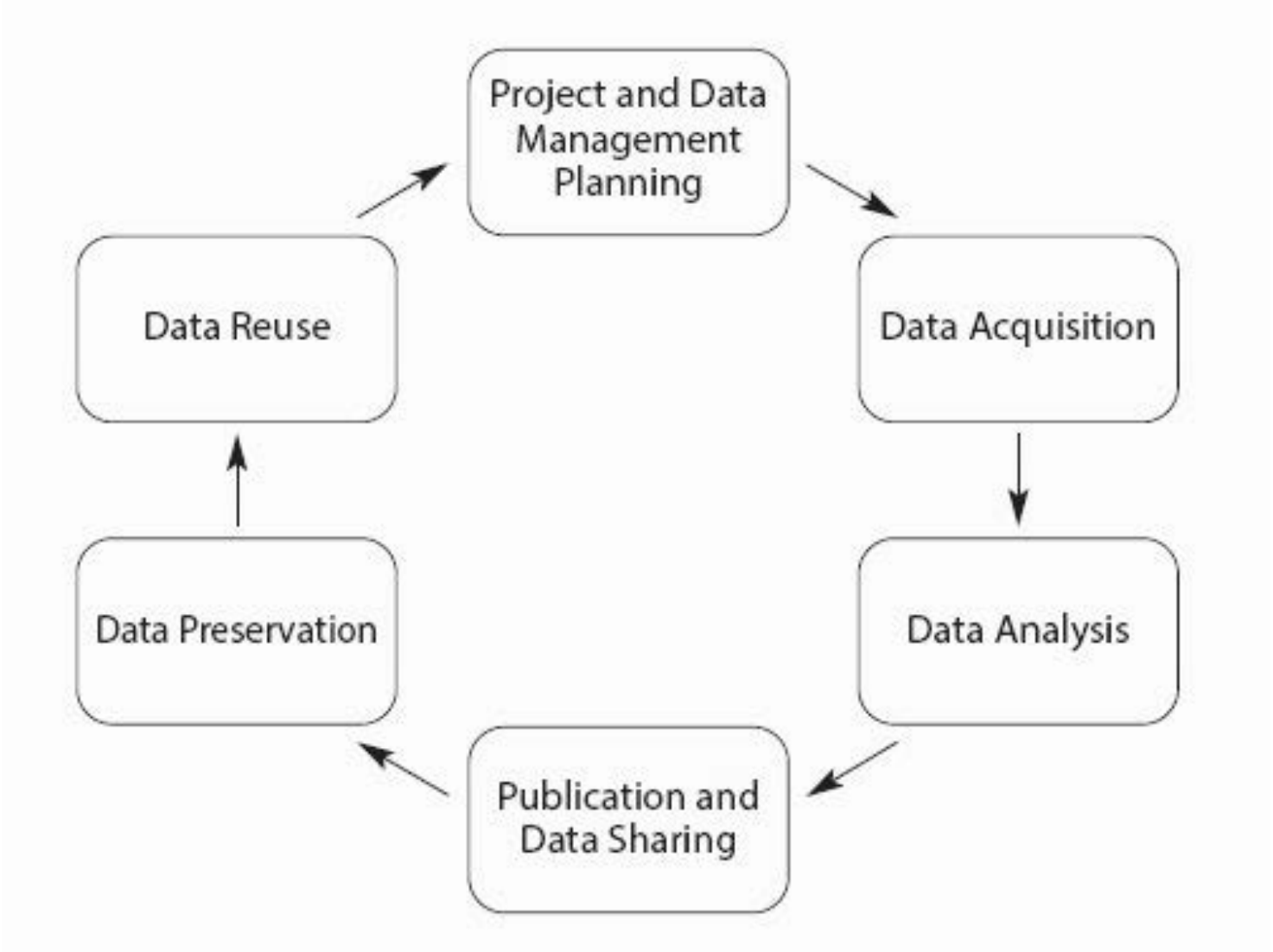
17 January 2017

Overview

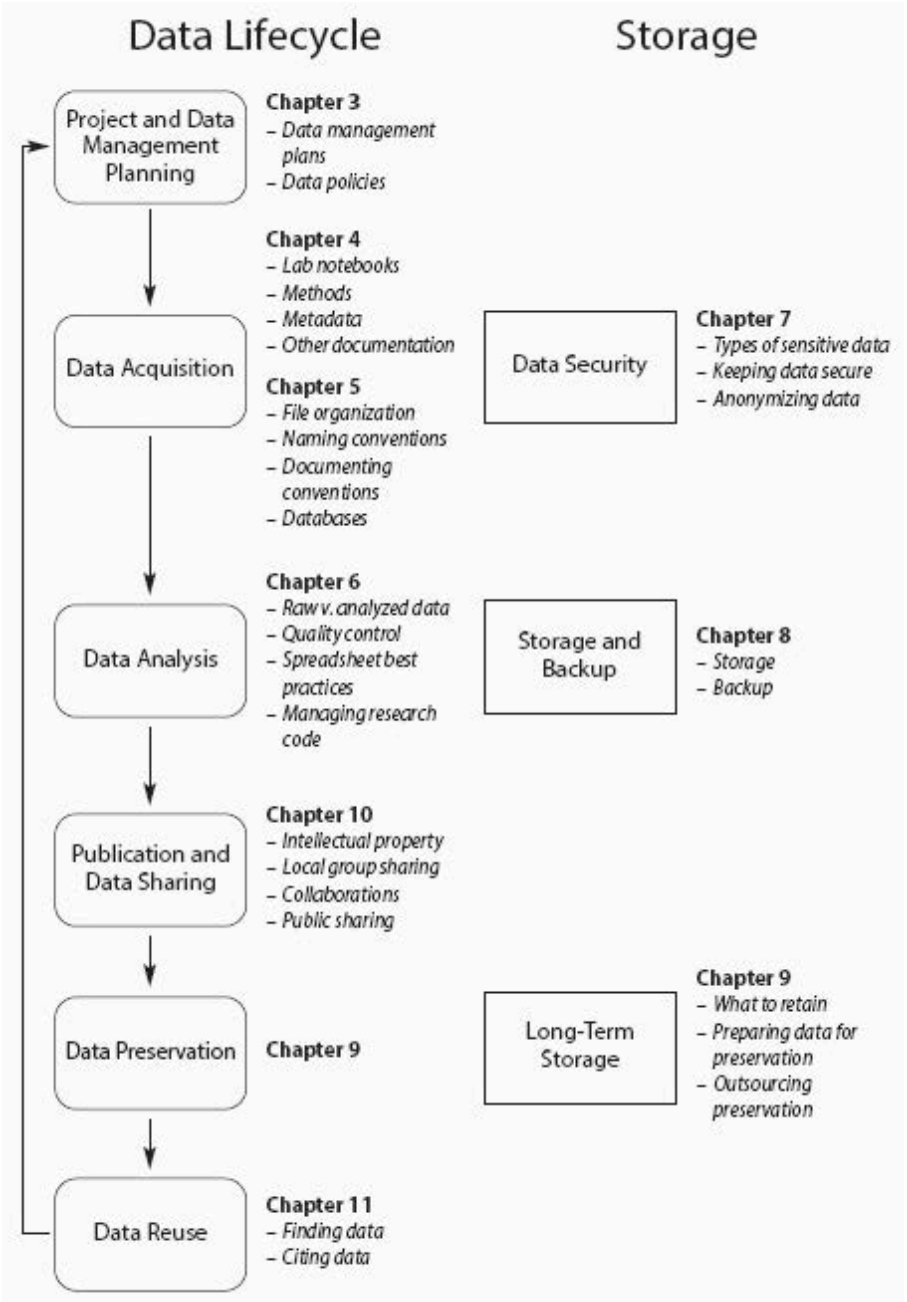
- Opening ideas for the afternoon session
- Why Plan?
- Why use frameworks and standards?
- What is in your Data Toolbox?

Planning & Organization

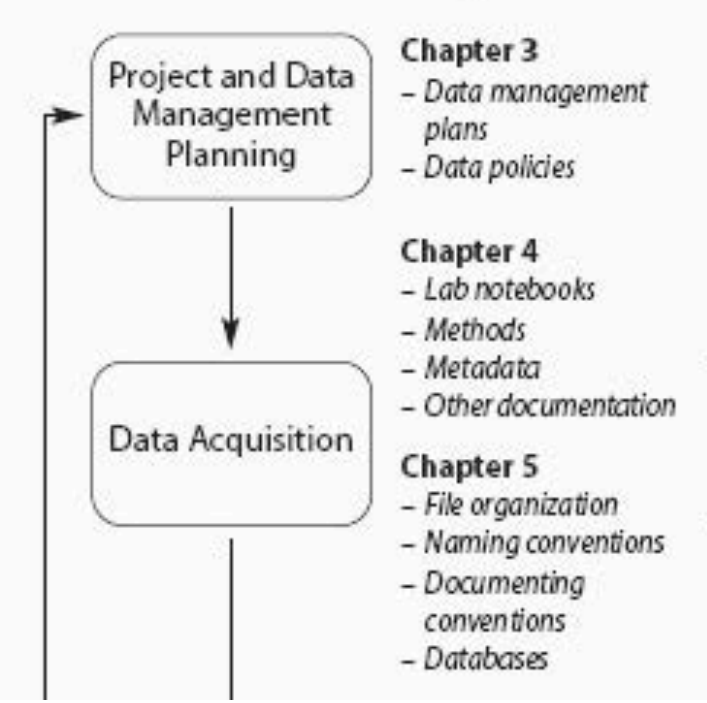
Modern Data LifeCycle



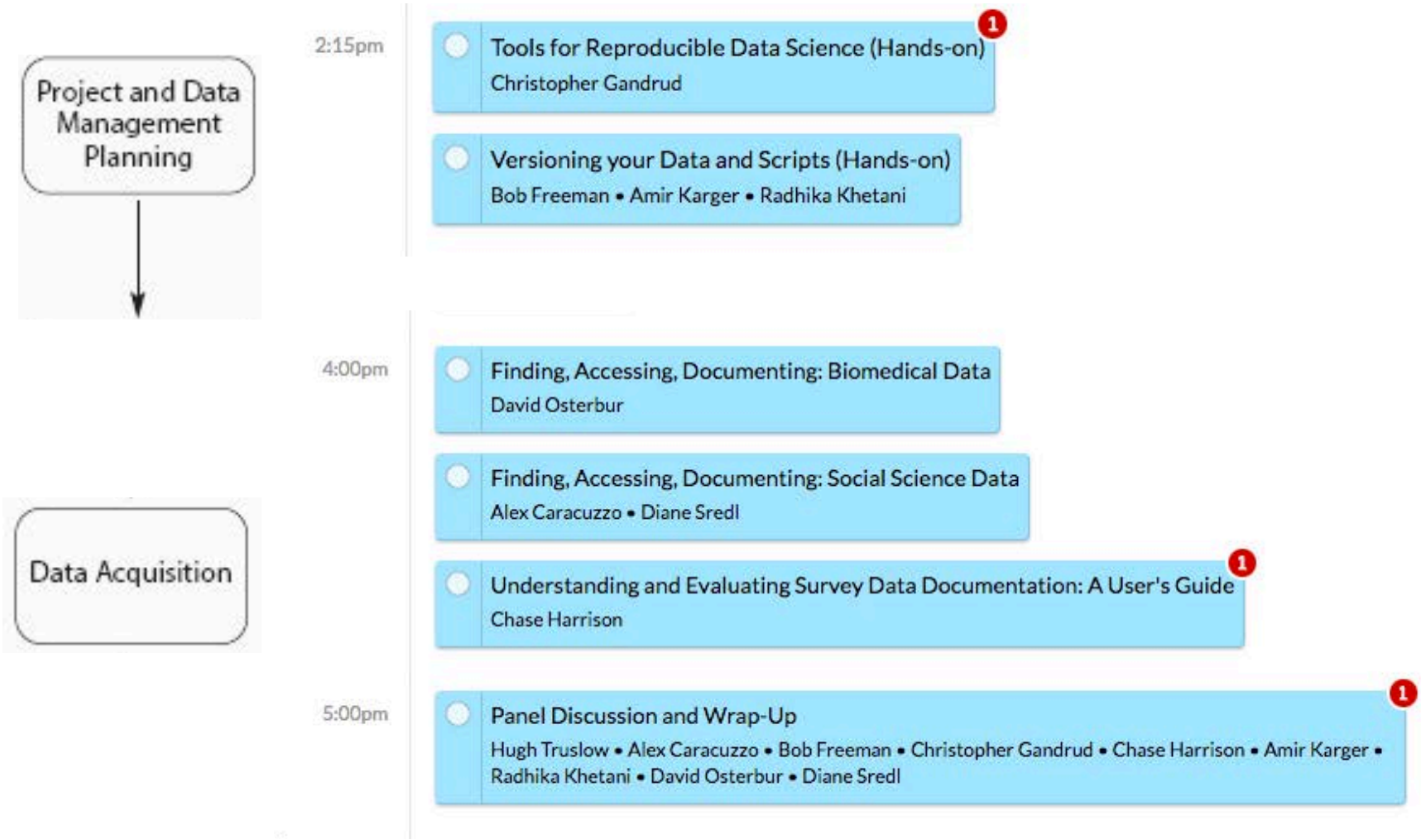
Excerpt From: Kristin Briney. "Data Management for Researchers." iBooks. https://itun.es/us/mnIf8.l



Excerpt From: Kristin Briney. "Data Management for Researchers." iBooks. https://itun.es/us/mnIf8.l

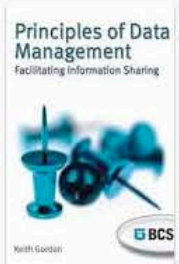


Excerpt From: Kristin Briney. "Data Management for Researchers." iBooks. https://itun.es/us/mnIf8.l

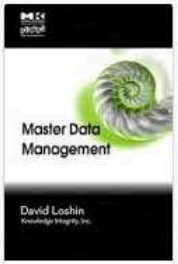


Excerpt From: Kristin Briney. "Data Management for Researchers." iBooks. https://itun.es/us/mnIf8.l

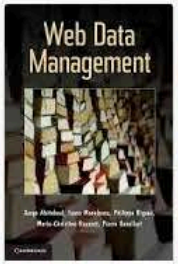
Google Search for “Data Management”



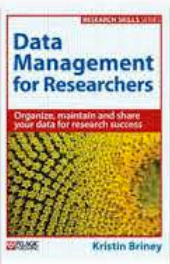
Principles of Data Manage...
Keith Gordon,...



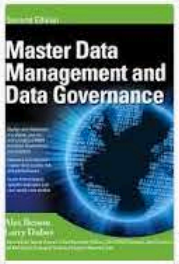
Master Data Management
David Loshin,...



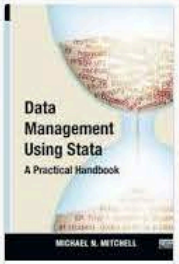
Web Data Management
2011



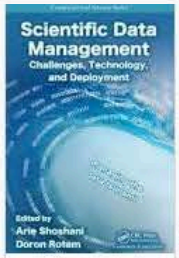
Data Management...
Kristin Briney,...



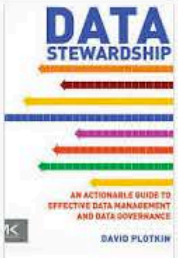
MASTER DATA MANAG..
2007



Data Management...
Michael N. Mi...

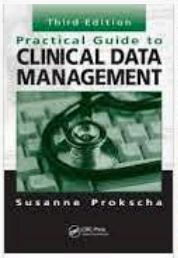


Scientific Data Management...
2009

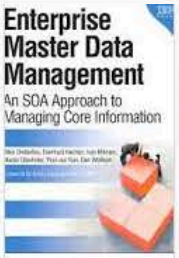


Data Stewardship: ...
David Plotkin,...

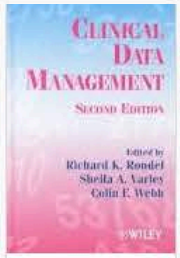
Research Data Manage...
2014



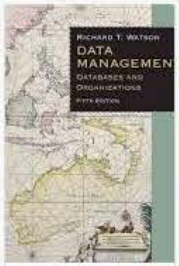
Practical Guide to Clini...
Susanne Prok...



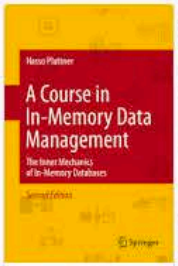
Enterprise Master Data ...
Allen Dreibelb...



Clinical Data Management
1993



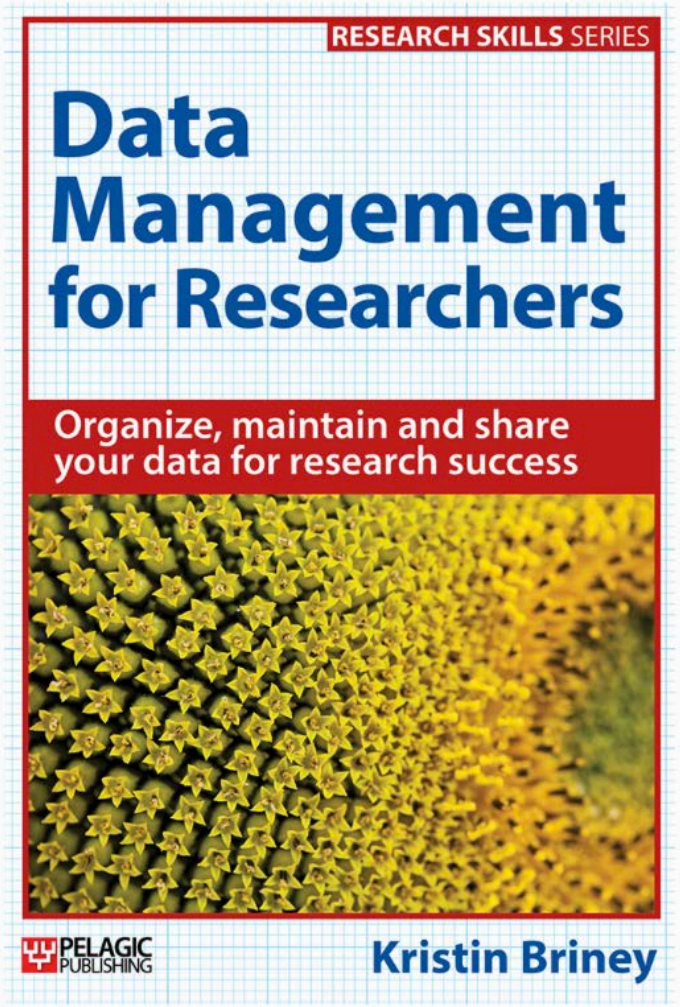
Data Management...
Tony J Watso...



A Course in In-Memory Da..
Hasso Plattn...

Big Data Management...
2013

Monetizing Data Manage...
Peter Aiken, 2...



Good Data Management…

Ask yourself…

- What types of data do I have? How much do I have?
- Do I use any third-party data?
- What data tools and technology are readily available to me?
- How much do I collaborate? Is this internal or external to my institution?
- How long must I keep my data?
- Will I share my data?
- Does my data have security concerns, such as personally identifiable information? What does my funder/institution/employer require?
- Is there anything particular in my research workflow that might affect my data management?
- What problems with my data do I often encounter

Excerpt From: Kristin Briney. "Data Management for Researchers." iBooks. https://itun.es/us/mnIf8.l

Good Data Management…

Your plan will cover…

- What data will you create?
- How will you document and organize your data?
- How will you store your data and, if necessary, keep it secure?
- How will you manage your data after the completion of the project?
- How will you make your data available for reuse, as necessary?

Good Data Management…

Your plan will cover…

- What data will you create?
- How will you document and organize your data?
- How will you store your data and, if necessary, keep it secure?
- How will you manage your data after the completion of the project?
- How will you make your data available for reuse, as necessary?

A New Hope®

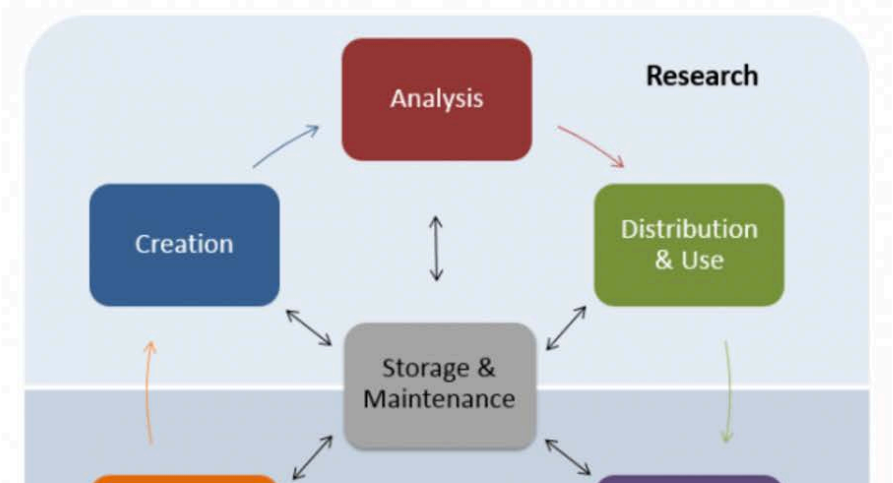
Harvard Biomedical Data Management

Best practices & support services for research data lifecycles

Data Management

Data Management is the process of providing the appropriate labeling, storage, and access for data at all stages of a research project. We recognize that best practices for each of these aspects of data management can and often do change over time, and are different for different stages in the data lifecycle.

Early and attentive management at each step of the data lifecycle will ensure the discoverability and longevity of your research.



Ask us your biomedical research data management questions!

HARVARD MEDICAL SCHOOL | The Francis A. Countway Library of Medicine

Receive Data Management Updates

Organization

OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Education

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble[1,2]*

1 Department of Genome Sciences, School of Medici
Engineering, University of Washington, Seattle, Was

Introduction

Most bioinformatics coursework focuses on algorithms, with perhaps some components devoted to learning programming skills and learning how to use existing bioinformatics software. Unfortunately, for students who are preparing for a research career, this type of curriculum fails to address many of the day-to-day organizational challenges associated with performing computational experiments. In practice, the principles behind organizing and documenting computational experiments are often learned on the fly, and this learning is strongly influenced by personal predilections as well as by chance interactions with collaborators or colleagues.

The purpose of this article is to describe one good strategy for carrying out computational experiments. I will not describe profound issues such as how to formulate hypotheses, design experiments, or draw conclusions. Rather, I will focus on relatively mundane issues such as organizing files and directories and documenting progress. These issues are important because poor organizational choices can lead to significantly slower research progress. I do not claim that the strategies I outline here are optimal. These are simply the principles and practices that I have developed over 12 years of bioinformatics research, augmented with various suggestions from other researchers with whom I have discussed these issues.

Principles

The core guiding principle is simple: Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why. This "someone" could be any of a variety of people: someone who read your published article and wants to try to reproduce your work, a collaborator who wants to understand the details of your experiments, a future student working in your lab who wants to extend your work after you have moved on to a new job, your research advisor, who may be interested in

OPEN ACCESS Freely available online

PLOS BIOLOGY

Community Page

Best Practices for Scientific Computing

Greg Wilson[1]*, **D. A. Aruliah**[2], **C. Titus Brown**[3], **Neil P. Chue Hong**[4], **Matt Davis**[5], **Richard T. Guy**[6¤], **Steven H. D. Haddock**[7], **Kathryn D. Huff**[8], **Ian M. Mitchell**[9], **Mark D. Plumbley**[10], **Ben Waugh**[11], **Ethan P. White**[12],

1 Mozilla Foundation, Toron
Michigan, United States of A
States of America, **6** Univers
8 University of California Ber
University of London, Londo
13 University of Wisconsin, M

Introduction

Scientists spend an
using software. Howev
do this efficiently. As
practices that would
maintainable code wit
practices for scientific
foundations in resear
scientists' productivity

Software is as imp
telescopes and test tub
computational proble
scientists, more and mo
around developing nev
large amounts of dat
projects, combining dis
and other computation

Scientists typically de
because doing so requi
As a result, recent studi
30% or more of their
90% or more of them a
lack exposure to basic
writing maintainable
trackers, code reviews,

We believe that soft
apparatus [3] and shou
as any physical appar

PLOS BIOLOGY

COMMUNITY PAGE

Computing Workflows for Biologists: A Roadmap

Ashley Shade[1,2]*, **Tracy K. Teal**[2,3]

CrossMark
click for updates

Code and Data for the Social Sciences: A Practitioner's Guide

Matthew Gentzkow Jesse M. Shapiro[1]

Chicago Booth and NBER

March 10, 2014

Planning & Metadata Frameworks

Metadata Frameworks & Standards

Daina Bouquin

Head Librarian, Harvard-Smithsonian Center for Astrophysics

Derek Miller

Assistant Professor of English, Harvard University

Caroline Shamu

Assistant Professor, Harvard Medical School


Data Toolbox

Data Toolbox

- Will post more information on the GitHub repository
- Invite others to send me their recommendations

- Important to know your (system) limitations…
  - Transitioning to cloud and cluster (HPC/HTC) systems doesn't have to be difficult!