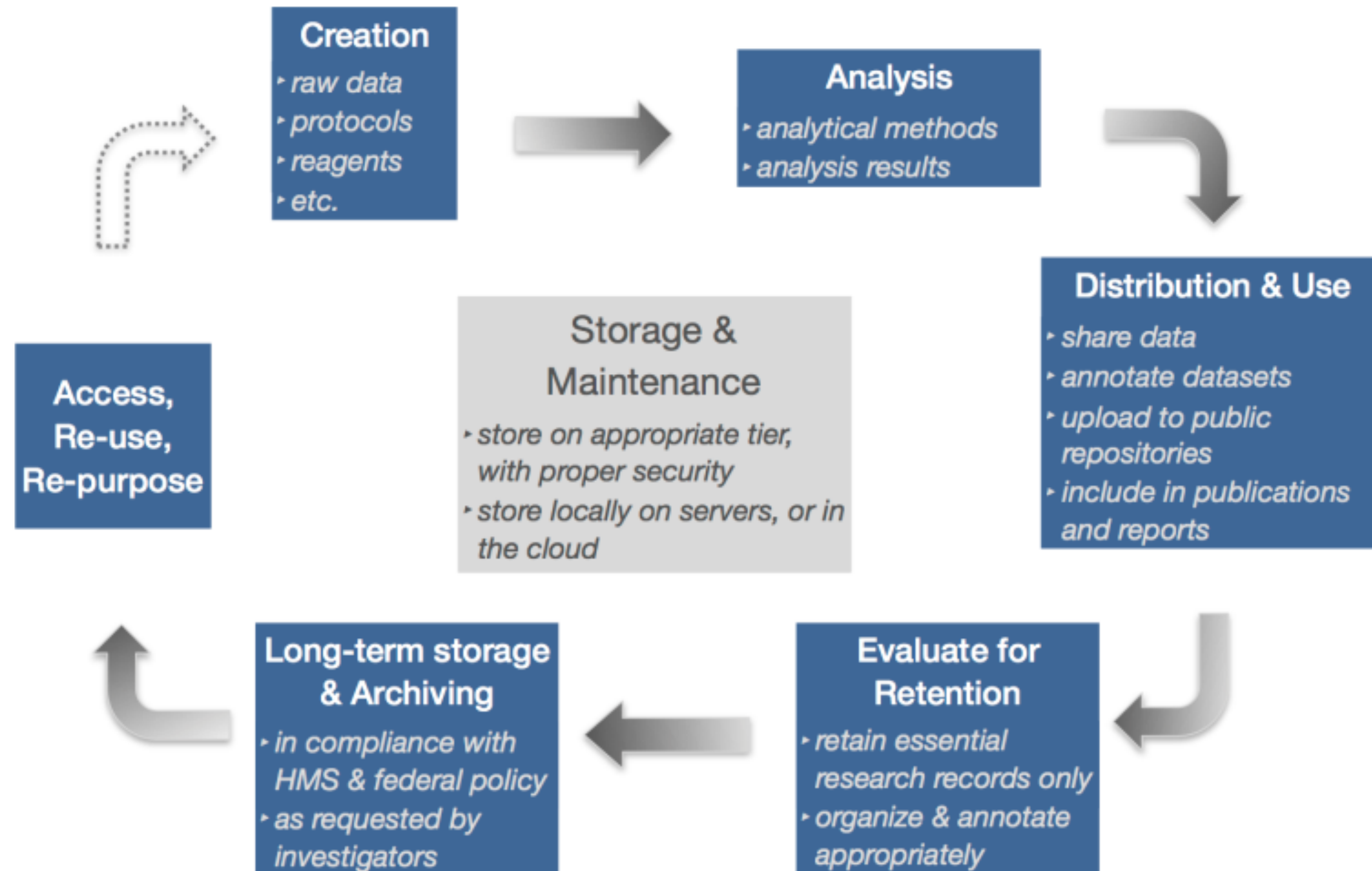


Data lifecycle for biomedical research



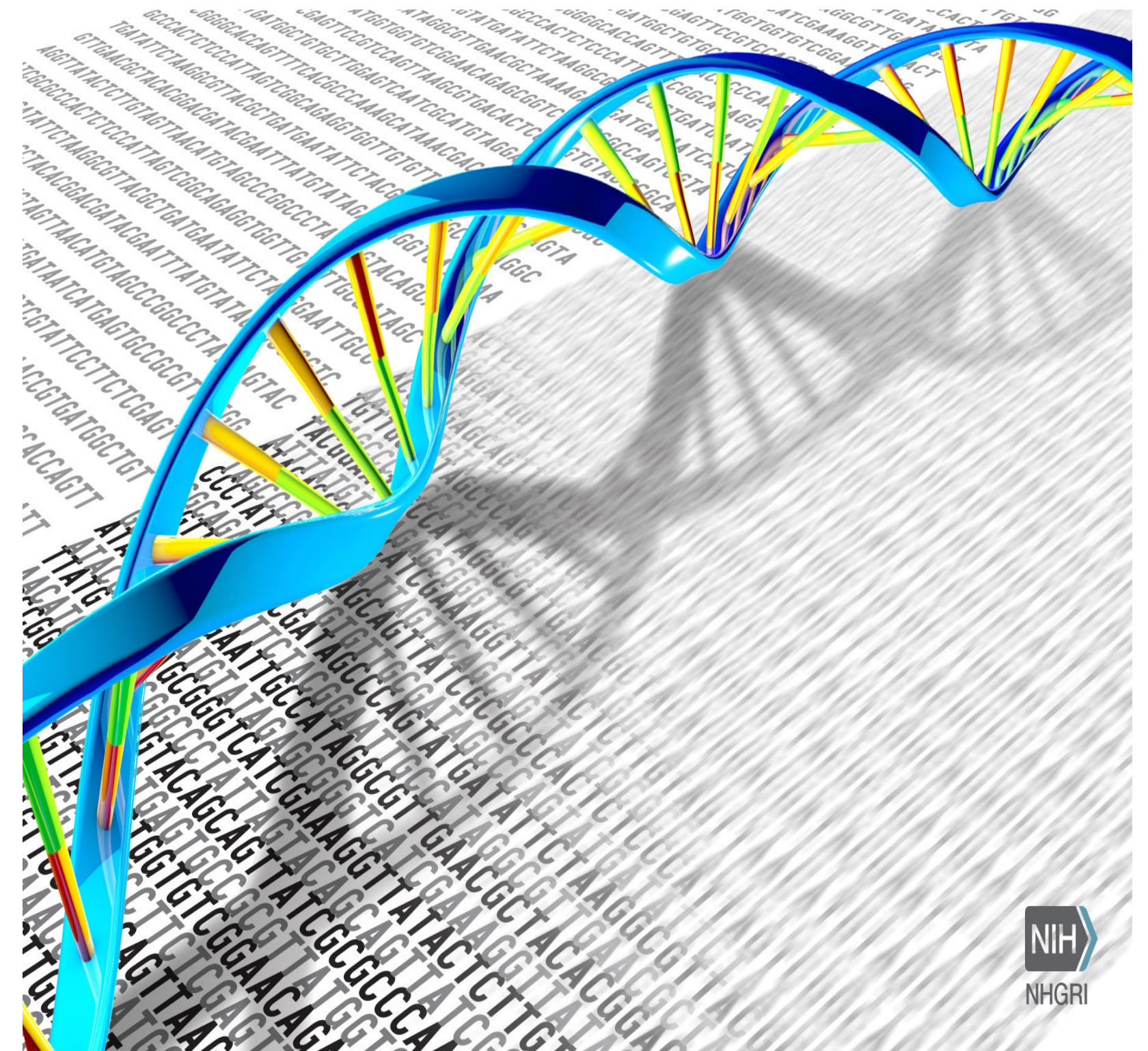
(HMS Data Management Working Group)

Data management planning

- ▶ Consider the **datatype and file sizes**. How much total raw data do you have?
- ▶ Consider the **type of analyses** you plan on performing. What types of intermediate data files will I generate?
- ▶ What is the best way to organize my **directory structure**?

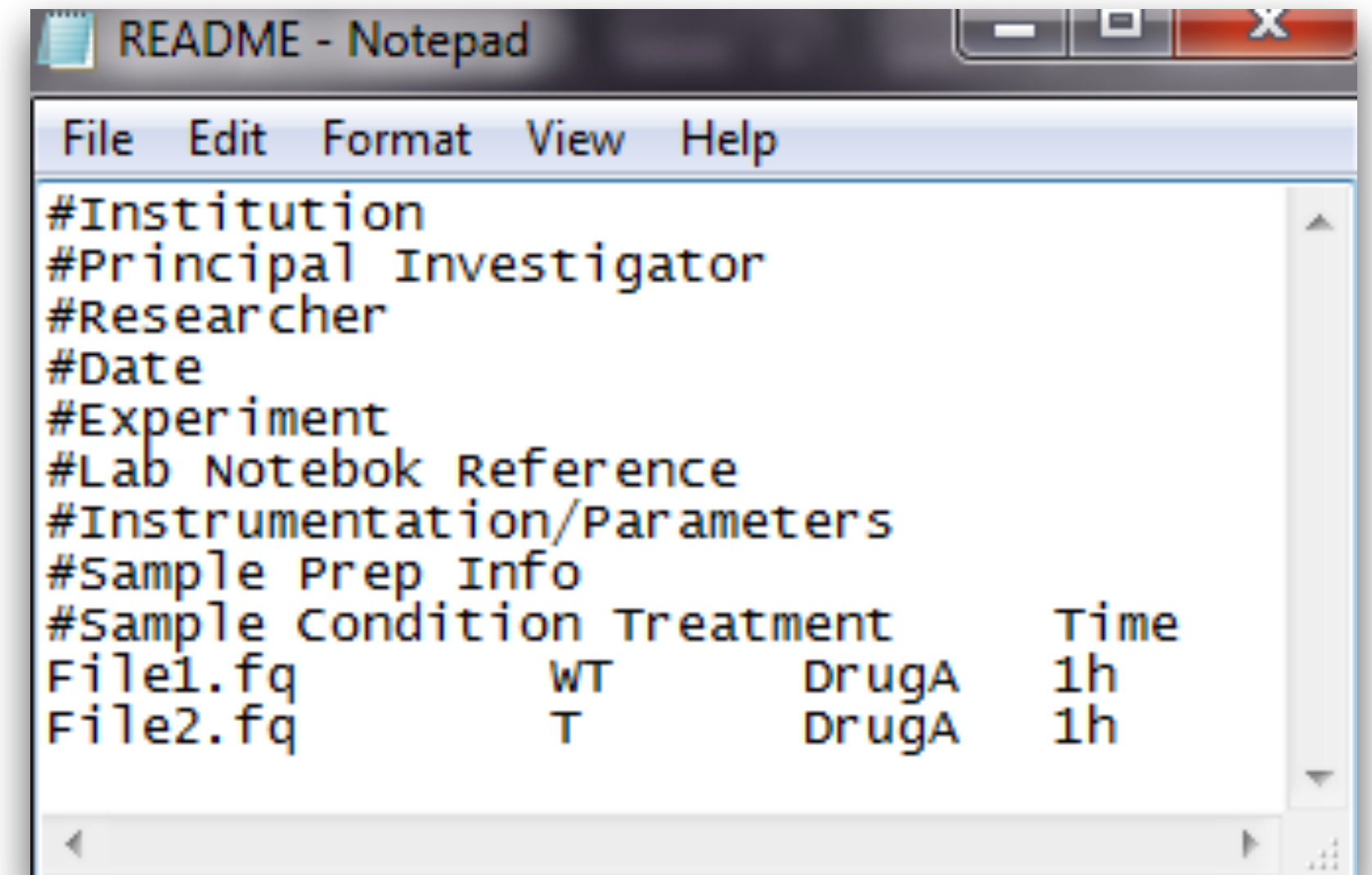
Data types: Raw data

- ▶ For NGS experiments, these are the FASTQ files you obtain from your sequencing facility (can be anything you start your analysis with)
- ▶ This data should never be directly modified (i.e. always keep a copy of this stored somewhere, untouched from its original state)
- ▶ Consider creating a read-only directory for this data



Metadata: README

- ▶ Create a plain text file (.txt) using a text editor (i.e Notepad, TextEdit, emacs, vi) and avoid proprietary formats (i.e Microsoft Word)
- ▶ Use this file to document information about the dataset including sample info, naming conventions, abbreviations, codes etc
- ▶ Use comments (hashtags) to describe information type
- ▶ Have a README file for each distance dataset.



```
File Edit Format View Help
#Institution
#Principal Investigator
#Researcher
#Date
#Experiment
#Lab Notebook Reference
#Instrumentation/Parameters
#Sample Prep Info
#Sample Condition Treatment Time
File1.fq WT DrugA 1h
File2.fq T DrugA 1h
```

File Naming conventions

- ▶ Should be **descriptive and provide contextual information**. Consider including some combination of the following:
 - ▶ Project/experiment name/acronym, Lab name/location, Date or date range, Type of data, Experimental conditions, Version number
- ▶ Consider **length** of filename
 - ▶ not too long (40-50 characters); limits differ by operating system
- ▶ Use naming conventions **consistently**

DO's


- ▶ Dates: YYYYMMDD e.g, 20161101
- ▶ Times: use 24-hour military time
- ▶ Sequential numbering: use leading zeros (e.g 001, 002, ... 010, 020, etc)
- ▶ Names: surname then given (i.e Smith_Bob)
- ▶ Versioning: use numbers to indicate updated versions (e.g. v1, v2)

DONT's

- ▶ Avoid special characters, these are sometimes used in specific tasks for some OS i.e. ~ ! @ # \$ % ^ & * () ` ; : < > ? . , [] { } ' " |
- ▶ Don't use carriage returns
- ▶ Don't use spaces. Instead try:
 - ▶ Underscores (e.g. file_name.xxx)
 - ▶ Dashes (e.g file-name.xxx)
 - ▶ No separation (e.g. filename.xxx)
 - ▶ Camel-case (e.g. fileName.xxx)

[COMMENT](#)[OPEN ACCESS](#)

Gene name errors are widespread in the scientific literature

[Mark Ziemann](#), [Yotam Eren](#) and [Assam El-Osta](#) 

Genome Biology 2016 17:177 | DOI: [10.1186/s13059-016-1044-7](https://doi.org/10.1186/s13059-016-1044-7) | © The Author(s). 2016

Published: 23 August 2016

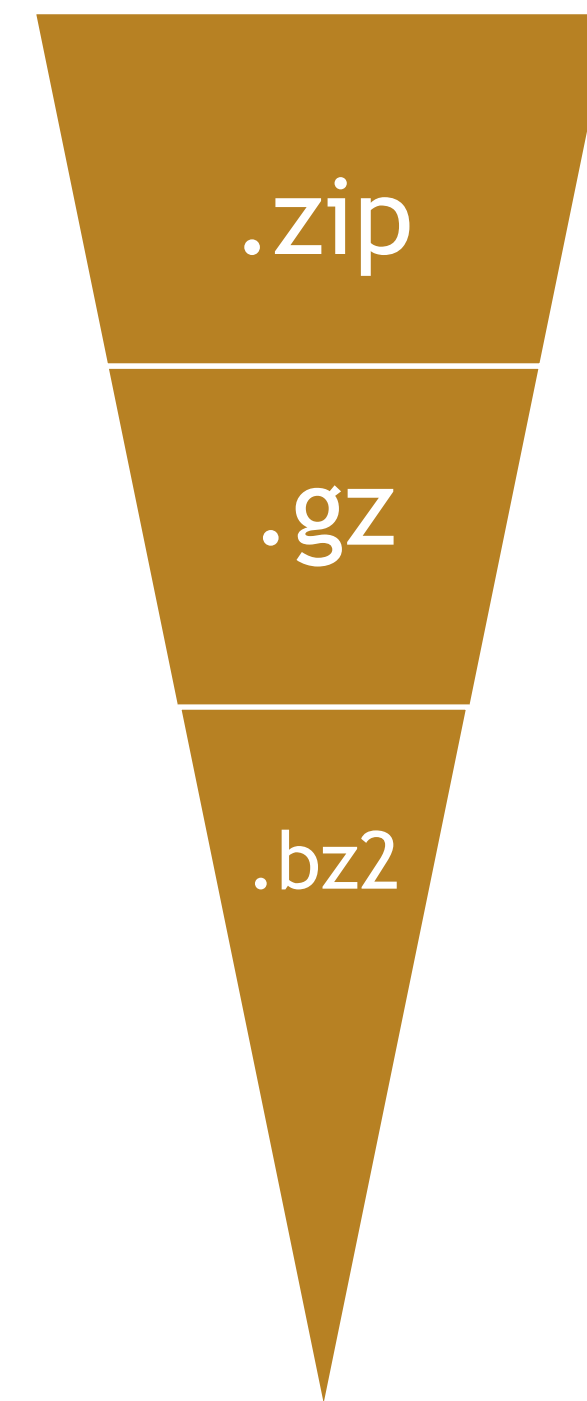
Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Be careful with Excel!

Data compression methods

- ▶ zip: DEFLATE coding
- ▶ gzip: Lempel-Ziv coding
- ▶ bzip2: Burrows-Wheeler block sorting text and Huffman coding
- ▶ tar: archival utility, preserving hierarchy and permissions, often used with gzip and bzip2



Data Deposition

- ▶ Certain NIH-funded research requires deposition of data into public repositories
- ▶ SRA, dbGAP, GEO, NDAR
- ▶ Harvard DataVerse: Harvard Libraries, HUIT, IQSS
- ▶ Structural Biology Data Grid

Data Retention and LTS

- ▶ **Adhere to your lab's standard practices for data management and organization.**
 - ▶ If you do not have standards, make them, write them down and follow them.
- ▶ **Keep your data for at least seven (7) years.**
 - ▶ IP (Intellectual Property), human subjects information and other factors can influence (extend) retention timelines, so before you delete data, check with your lab and sponsor guidelines first. If you have IP, talk to the Office of Technology Department (OTD).
- ▶ **Store your data on University premises and/or systems.**
 - ▶ HMS IT offers a Long-Term Storage service for storing large quantities of **infrequently** accessed data that do not need to be retrieved immediately
 - ▶ For more information about the service, please visit: <http://rits.hms.harvard.edu/dm/lts>
 - ▶ If you are not sure what options are available to you, contact HUIT.

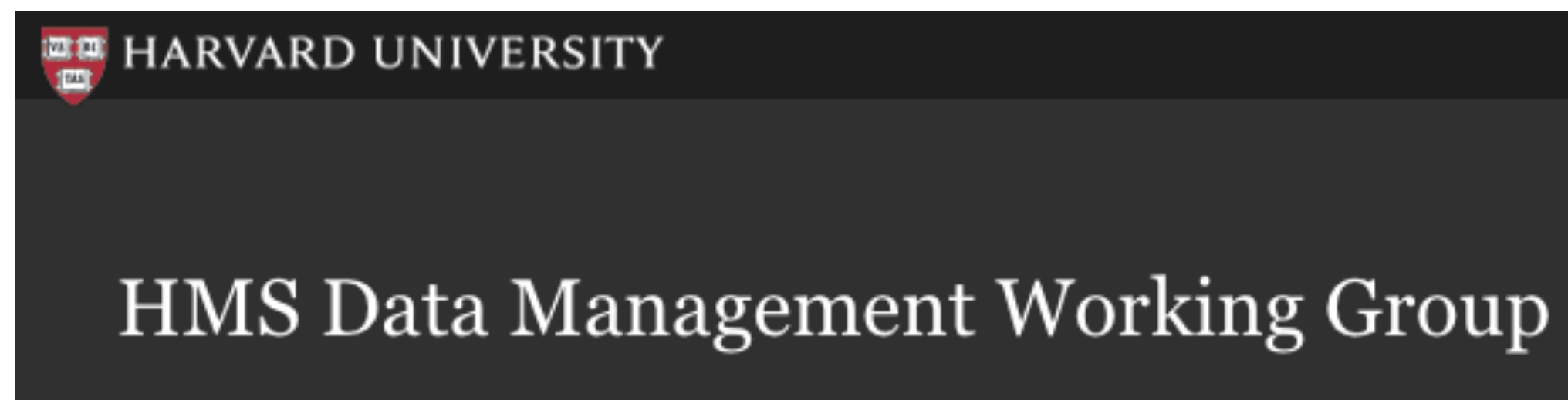
Scripting: Version Tracking

- ▶ Record changes (additions/deletions/replacements) in your code
- ▶ Revert back to an older (working) version with ease
- ▶ Collaboration: facilitate many people to work on a file at once
- ▶ Public or private repository options



Credits

- ▶ These materials were adapted from existing materials created by Radhika Khetani, Jessica Pearce from the HMS Data Management Working Group and Kris Holton from HMS Research Computing



HARVARD
MEDICAL SCHOOL

Information Technology

RESEARCH COMPUTING
<https://rc.hms.harvard.edu/>

These materials have been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

