Librarian-in-the-Loop Deep Learning To Curate Very Large Biomedical Image Datasets

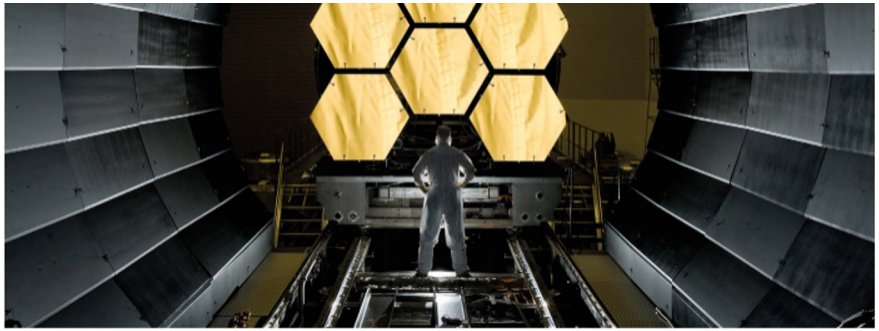
Zhiwu Xie, AUL for Research & Technology, UCR Library, University of California Riverside
Yinlin Chen, Assistant Director, Center for Digital Research & Scholarship, University Libraries, Virginia Tech

UC RIVERSIDE VIRGINIA TECH Yale

INSTITUTE of Museum and Library SERVICES

Librarian-in-the-Loop Deep Learning To Curate **Very Large Biomedical Image** Datasets

- In biology and medicine, structure often dictates the function.
- Enhanced Focused Ion Beam Scanning Electron Microscopy (FIB-SEM): a ‘quiet revolution’



The James Webb Space Telescope's 6.5-metre primary mirror (6 of 18 segments shown) can detect objects billions of light years away.

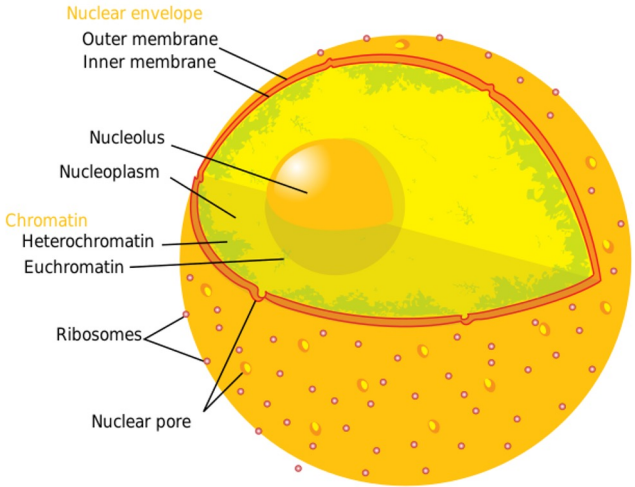
NASA/MSFC/DAVID HIGGINBOTHAM

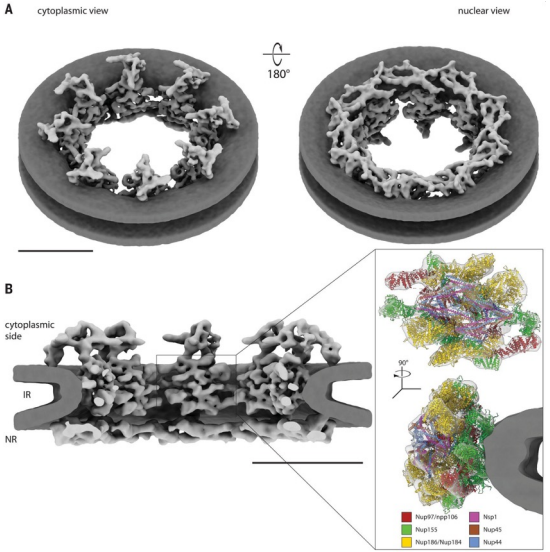
SEVEN TECHNOLOGIES TO WATCH IN 2023

Nature's pick of tools and techniques that are poised to have an outsized impact on science in the coming year. **By Michael Eisenstein**

Nature 613, 794-797 (2023)
doi: https://doi.org/10.1038/d41586-023-00178-y

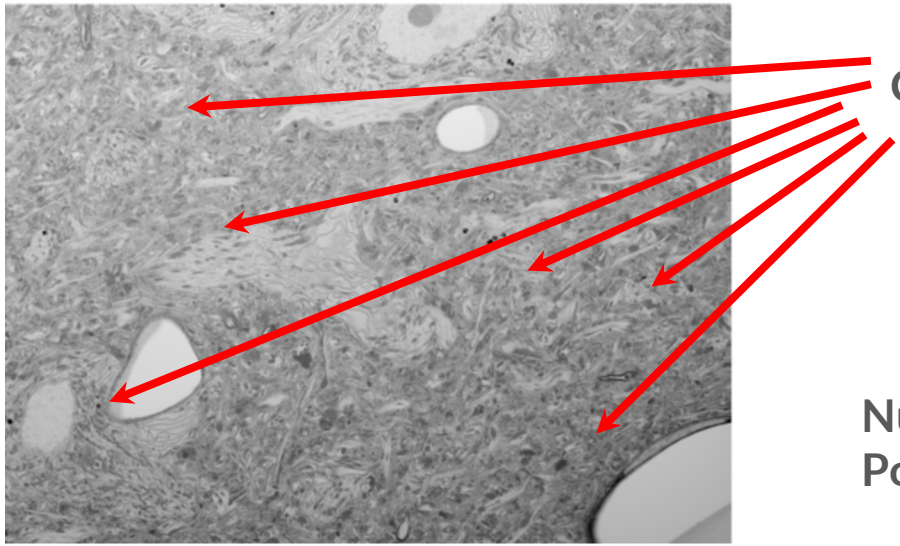
Cell Nucleus and Nuclear Pores



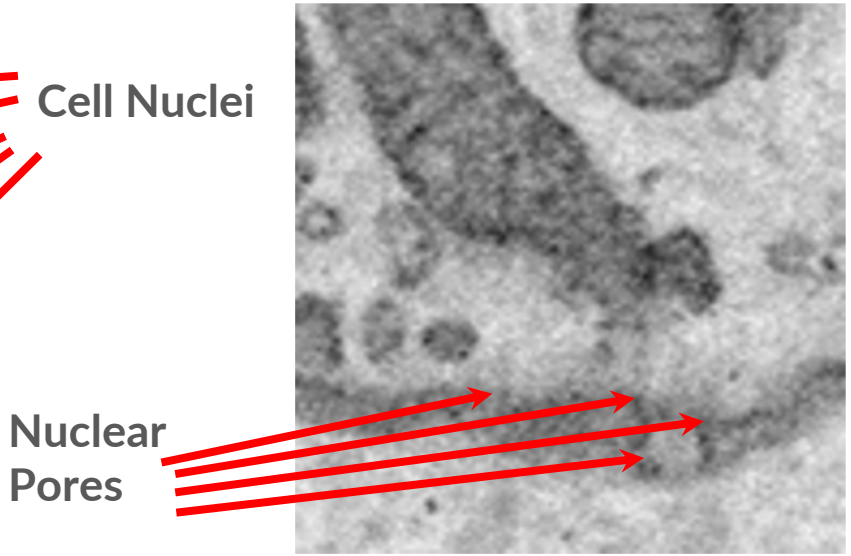


SCIENCE 11 Nov 2021 Vol 374, Issue 6573 DOI: 10.1126/science.abd9776

Mouse Neuron Nuclei @ 8nm Resolution



Zoomed out to 1/8000 of the original resolution, showing multiple cell nuclei

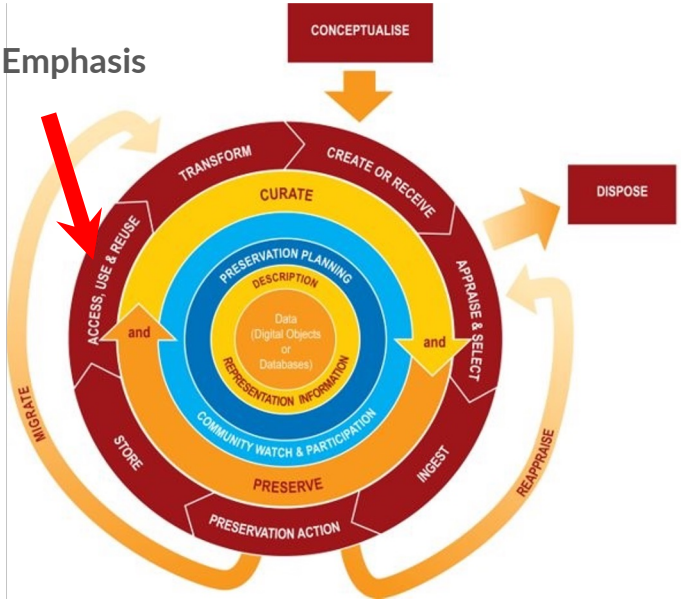


A tiny section of the same image in the original resolution, showing the pores

Librarian-in-the-Loop Deep Learning To **Curate** Very Large Biomedical Image **Datasets**

- Must data curation be fully separated from the science pipeline?
- Re-think data curation: move upstream to embed librarian work inside the science pipeline
- Data curation as the side effects of the science pipeline.

Our Emphasis

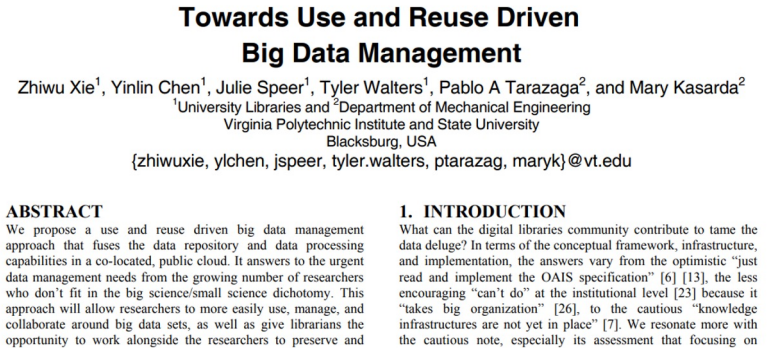


Sarah Higgins. The DCC Curation Lifecycle Model
https://doi.org/10.2218/ijdc.v3i1.48

Librarian-in-the-Loop Deep Learning To **Curate** Very Large Biomedical Image **Datasets**

- Use and Reuse Driven Big Data Management
  - Scientists collect data to answer science questions, not for data curation
- Librarians expanding into knowledge creation must be capable of helping answer science questions and embed in the science pipeline
- Data use & reuse →Data analysis
  - How many pores on each nucleus? Pore density? Size distribution?

Z Xie et al. Towards Use And Reuse Driven Big Data Management. JCDL' 15. https://doi.org/10.2218/ijdc.v3i1.48



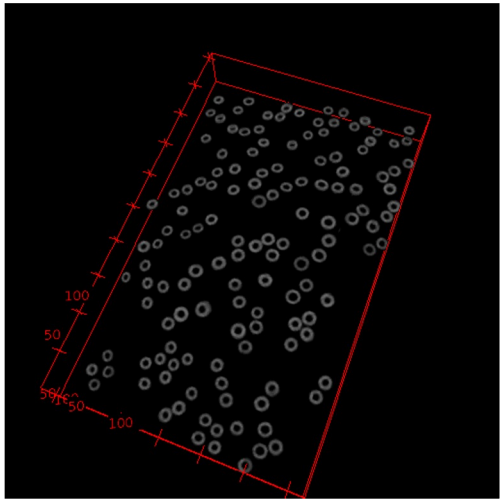
Towards Use and Reuse Driven Big Data Management

Zhiwu Xie[1], Yinlin Chen[1], Julie Speer[1], Tyler Walters[1], Pablo A Tarazaga[2], and Mary Kasarda[2]
[1]University Libraries and [2]Department of Mechanical Engineering
Virginia Polytechnic Institute and State University
Blacksburg, USA
{zhiwuxie, ylchen, jspeer, tyler.walters, ptarazag, maryk}@vt.edu

ABSTRACT
We propose a use and reuse driven big data management approach that fuses the data repository and data processing capabilities in a co-located, public cloud. It answers to the urgent data management needs from the growing number of researchers who don't fit in the big science/small science dichotomy. This approach will allow researchers to more easily use, manage, and collaborate around big data sets, as well as give librarians the opportunity to work alongside the researchers to preserve and

1. INTRODUCTION
What can the digital libraries community contribute to tame the data deluge? In terms of the conceptual framework, infrastructure, and implementation, the answers vary from the optimistic "just read and implement the OAIS specification" [6] [13], the less encouraging "can't do" at the institutional level [23] because it "takes big organization" [26], to the cautious "knowledge infrastructures are not yet in place" [7]. We resonate more with the cautious note, especially its assessment that focusing on

Librarian-in-the-Loop **Deep Learning** To Curate Very Large Biomedical Image Datasets



Manually label a few tiny slices taken from the whole cell image

Train A Deep Neural Network Using *3D UNet* To Recognize Pores From FIB-SEM Images

If predictions are not good enough, add more ground truths and do more training

Predict pores on the whole cell

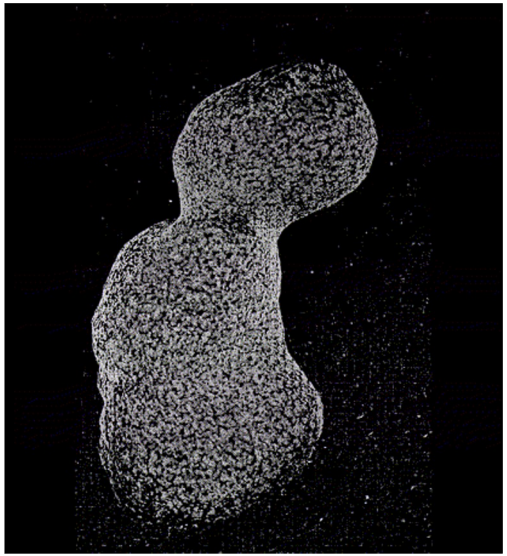
Librarian-in-the-Loop Deep Learning To Curate Very Large Biomedical Image Datasets

- A special case of human-in-the-loop machine learning.
- Embed librarians in the science loop. But why?
  - We can
    - We are professional informationists.
      - Digging into information and data has always been part of our job.
      - Labeling (categorizing) image pixels are not fundamentally different from cataloging or creating metadata
    - More librarians are gaining data and AI skills as used in this project
    - We are eager to learn more
  - We should
    - These skills are much needed, while we are eager to become campus research partners
    - We are service oriented and do not need to fixate on our own research agenda
    - Abundant opportunities exist for those who are well prepared
    - Data curation as the side effects of the science pipeline.

Librarian-in-the-Loop Deep Learning To Curate Very Large Biomedical Image Datasets

Label 2 additional tiny slices as ground truth

Both false positives and false negatives are significantly reduced



Key Takeaways

- Librarianship is whatever we make it to be, not defined by a fixed set of doctrines.
- To qualify as competent research partners, librarians need to embed ourselves deeper in the science pipeline to help answer critical research questions.
- Focus on our partner's research agenda, achieve our own agenda through the side effects of our collaborations.
- Practical machine learning/AI work is not all about programming and technical skills. It's a combination of hard and soft skills including manual labor, human intuition, and trial-and-error.

Questions? Comments? Please Get In Touch

Zhiwu Xie
Assistant University Librarian for
Research & Technology, UCR Library
zhiwux@ucr.edu

Yinlin Chen
Assistant Director, Center for Digital Research
& Scholarship, Virginia Tech Libraries
ylchen@vt.edu