

Reproducible and Attributable Materials Science Curation Practices: A Case Study

Ye Li*

yel@mit.edu

Sara L. Wilson

slwilson@mit.edu

Micah Altman

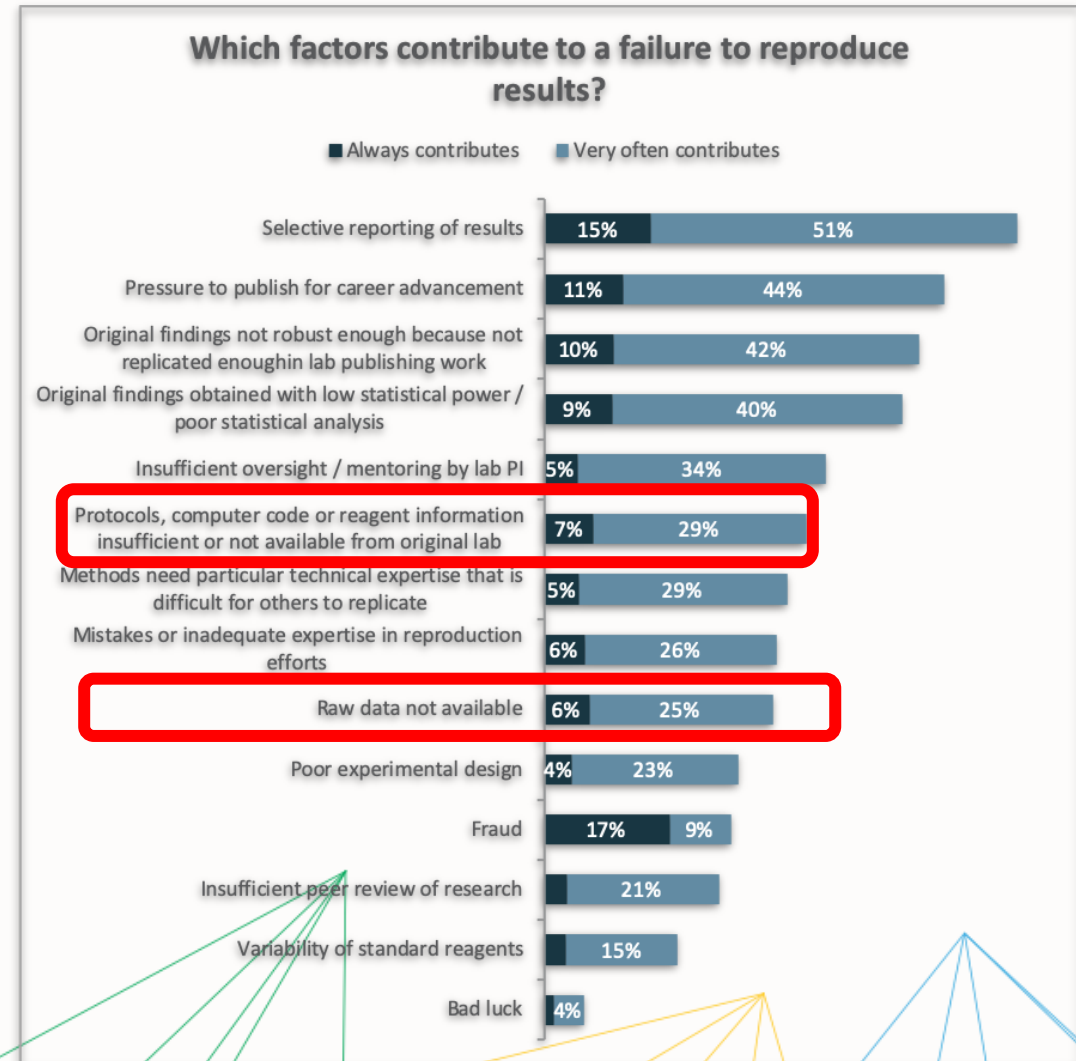
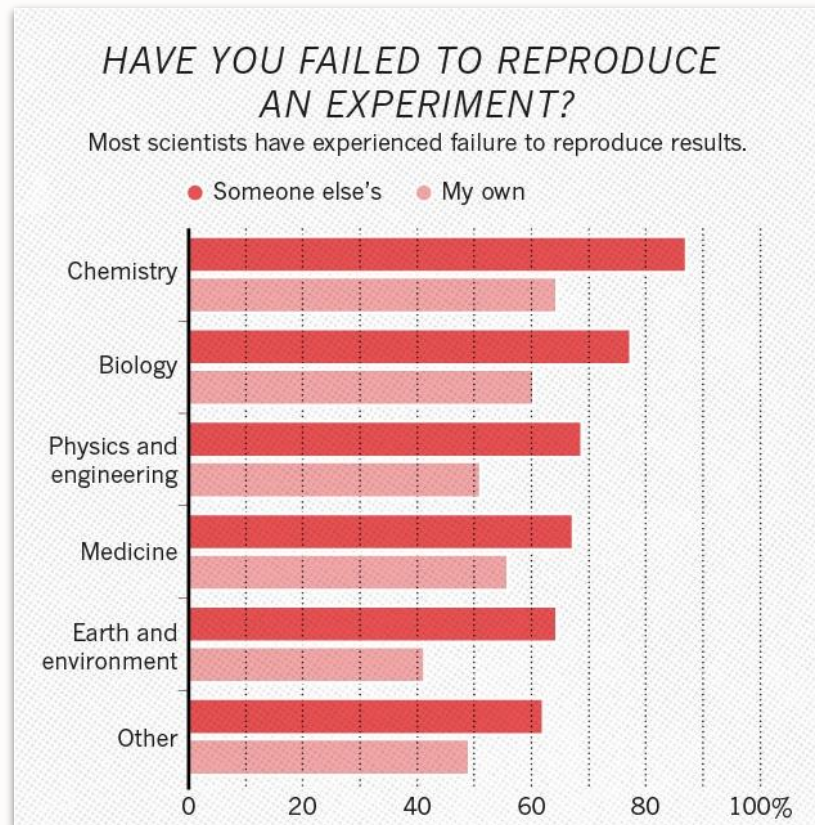
escience@mit.edu

**MIT
Libraries**

February 20, 2024

IDCC'24 - Edinburgh, Scotland

Research data practice and workflow directly shape reproducibility, openness, and attribution of research outputs



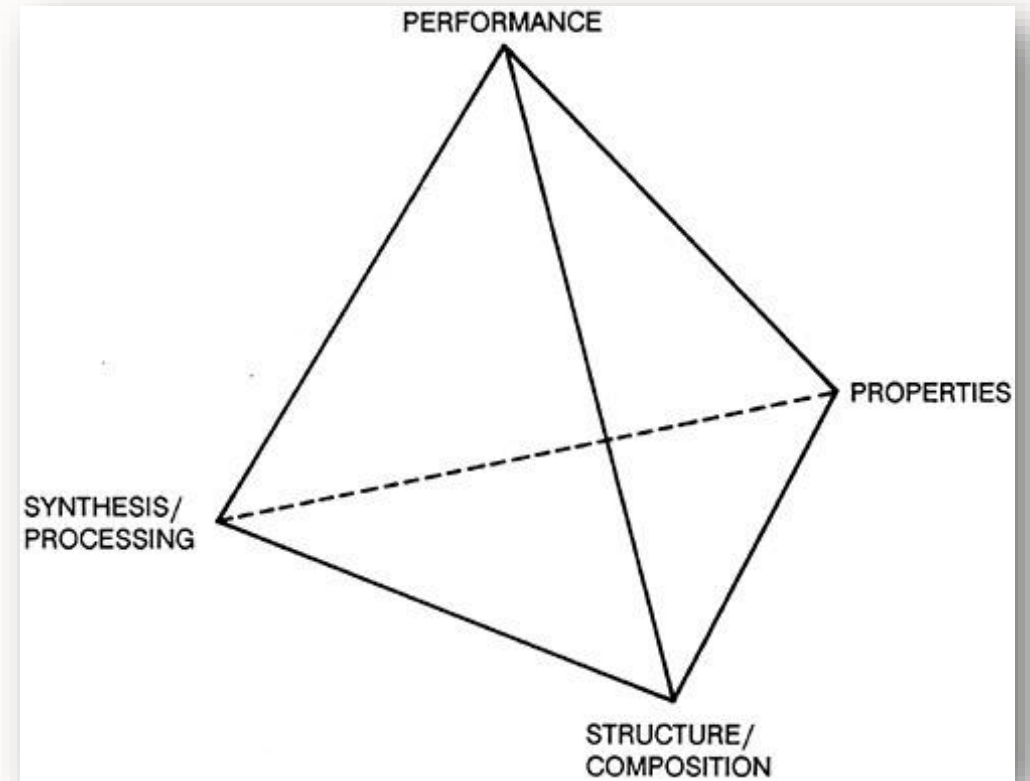
**MIT
Libraries**

Nature Reproducibility Survey 2017. **2018**. <https://doi.org/10.6084/m9.figshare.6139937.v4>.

Baker, M. 1,500 Scientists Lift the Lid on Reproducibility. *Nature* **2016**, 533 (7604), 452–454. <https://doi.org/10.1038/533452a>.

Why Material Science and Engineering (MSE) ?

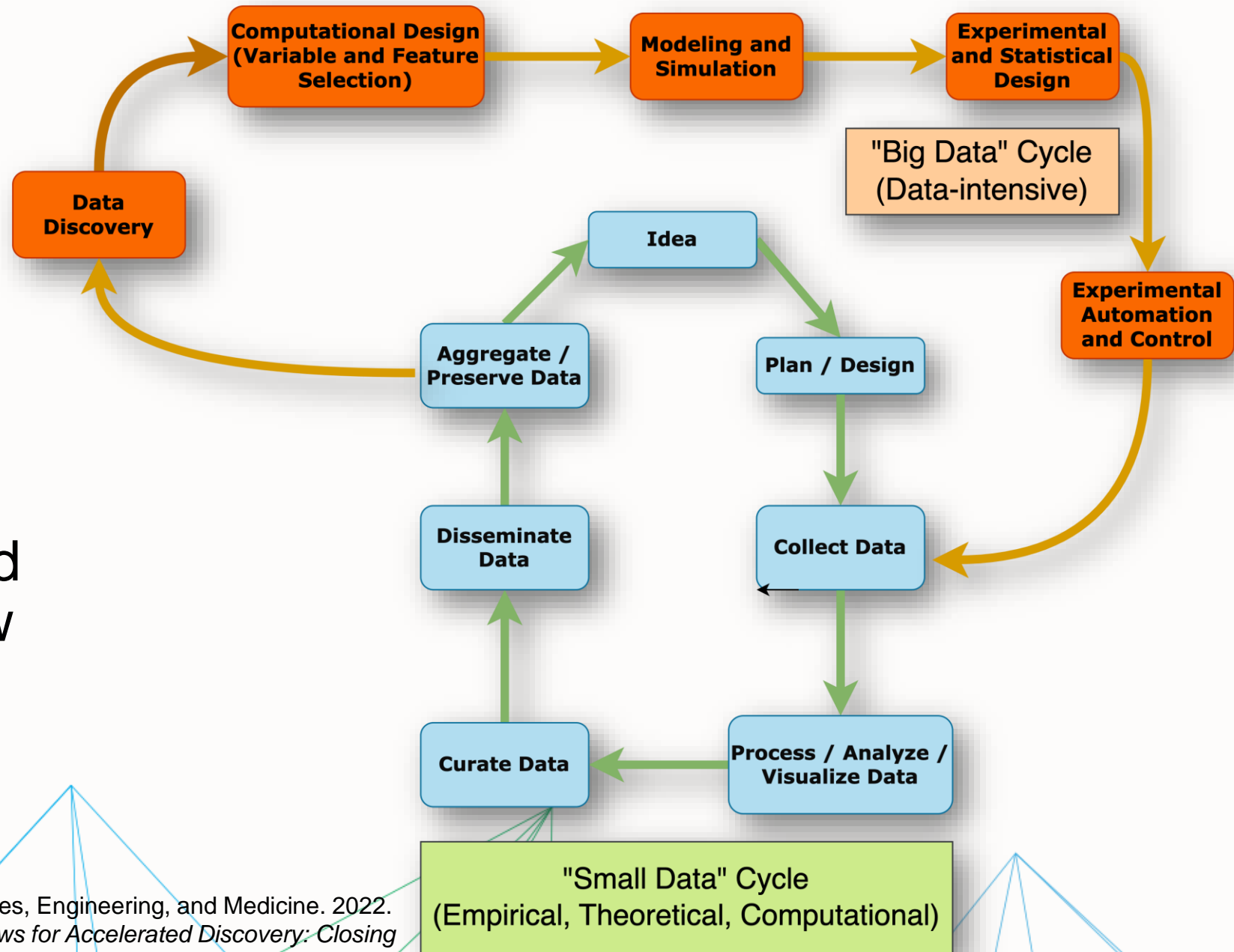
- Experimental measurement results in wealth of data, quantitative and qualitative
- Experimental process and data is often scattered in primary research articles from individual “small labs”
- Data sharing practices have not been broadly adopted



Four Elements of MSE Research

Why MSE?

- Closing the loop between “Big data” and “Small data” cycle for Automated Research Workflow

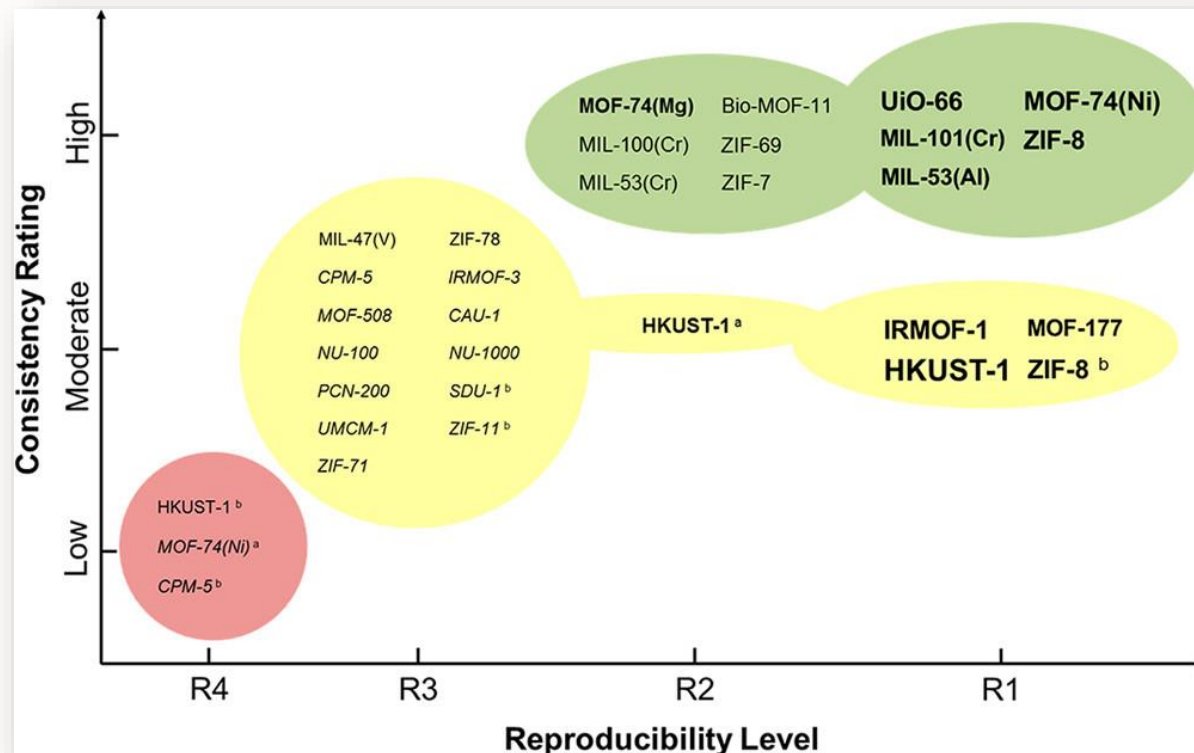


Why MSE?

Example of reproducibility issues in MSE

“...for the well-studied case of CO₂ adsorption there are only **15 of the thousands of known MOFs** for which enough experiments have been reported to allow strong conclusions to be drawn about the reproducibility of these measurements.”

“... In the examples in which **enough data exist to assess** the existence of outliers, approximately 20% of isotherms in the literature were classified as outliers.”



**MIT
Libraries**

Park, Jongwoo, *et al.* 2017. How Reproducible Are Isotherm Measurements in Metal–Organic Frameworks? *Chemistry of Materials* 29 (24): 10487–95.

<https://doi.org/10.1021/acs.chemmater.7b04287>

Reproducibility map for a comprehensive summary of reproducibility, consistency, and outlier levels for CO₂ isotherms in MOFs

Why MSE?

- Jaramillo Lab at MIT
 - A “small lab” in MSE
 - Dedicated to reproducible research and open science



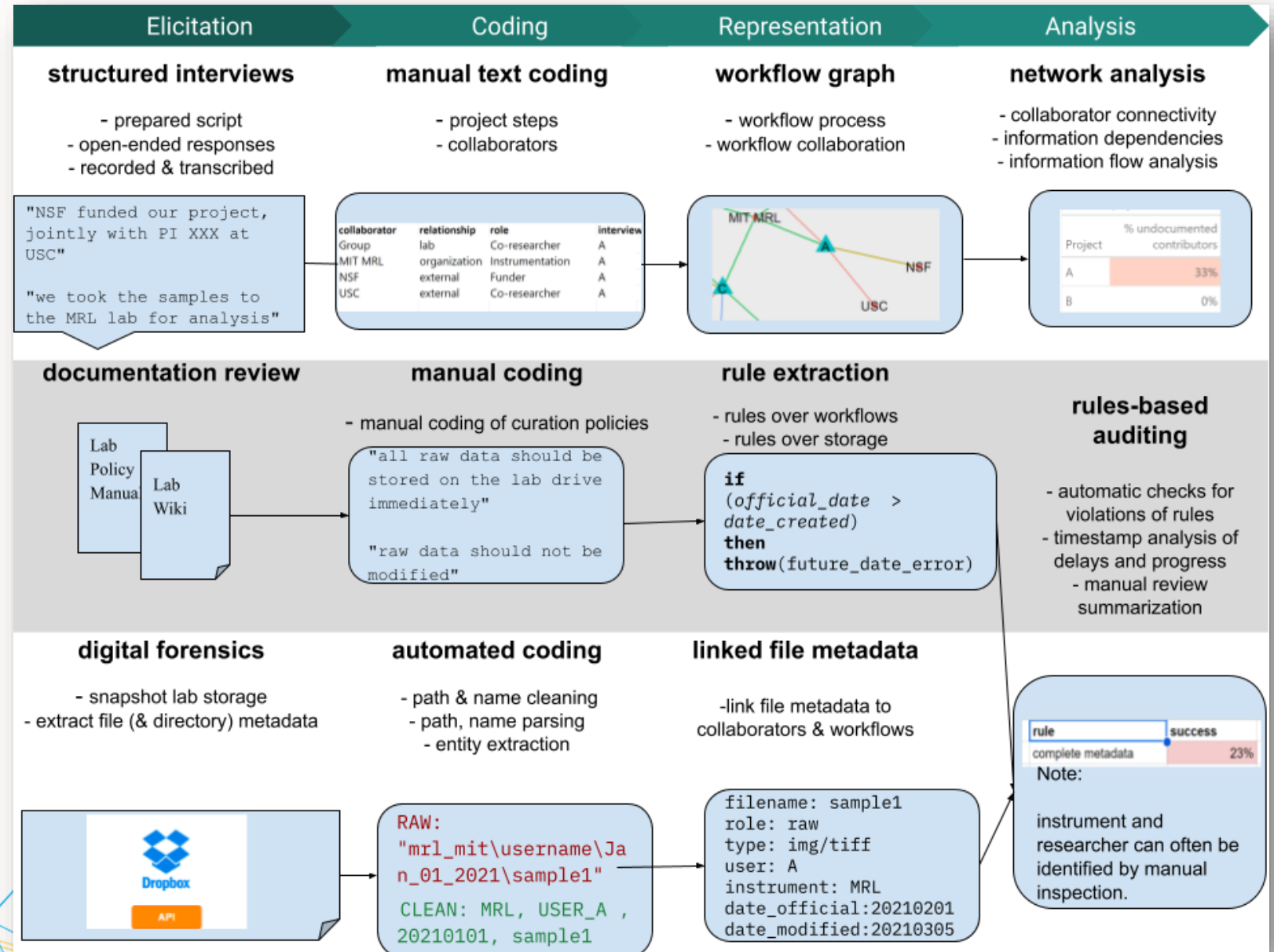
Research Questions

- To what extent does research **depend on manual processes** for information management?
- Explicit **processes**:
 - What processes concerning data and research workflow management are **documented**?
 - To what extent are documented processes **consistent with practice**?
- To what extent are documentation processes **complete** enough to support another person's **replication** of a result within the lab (without further communication with the original researcher)?
- To what extent are data management processes **robust** enough to survive the **departure** of a project member or the **loss** of an individual's personal computer or storage?
- To what extent are workflow data, outputs, and documentation **sufficient** to describe **responsibility** (or support attribution) for published results?

Data Sources and Methods

Purposive case study

- Structured interviews
- Documentation review
- Digital forensics



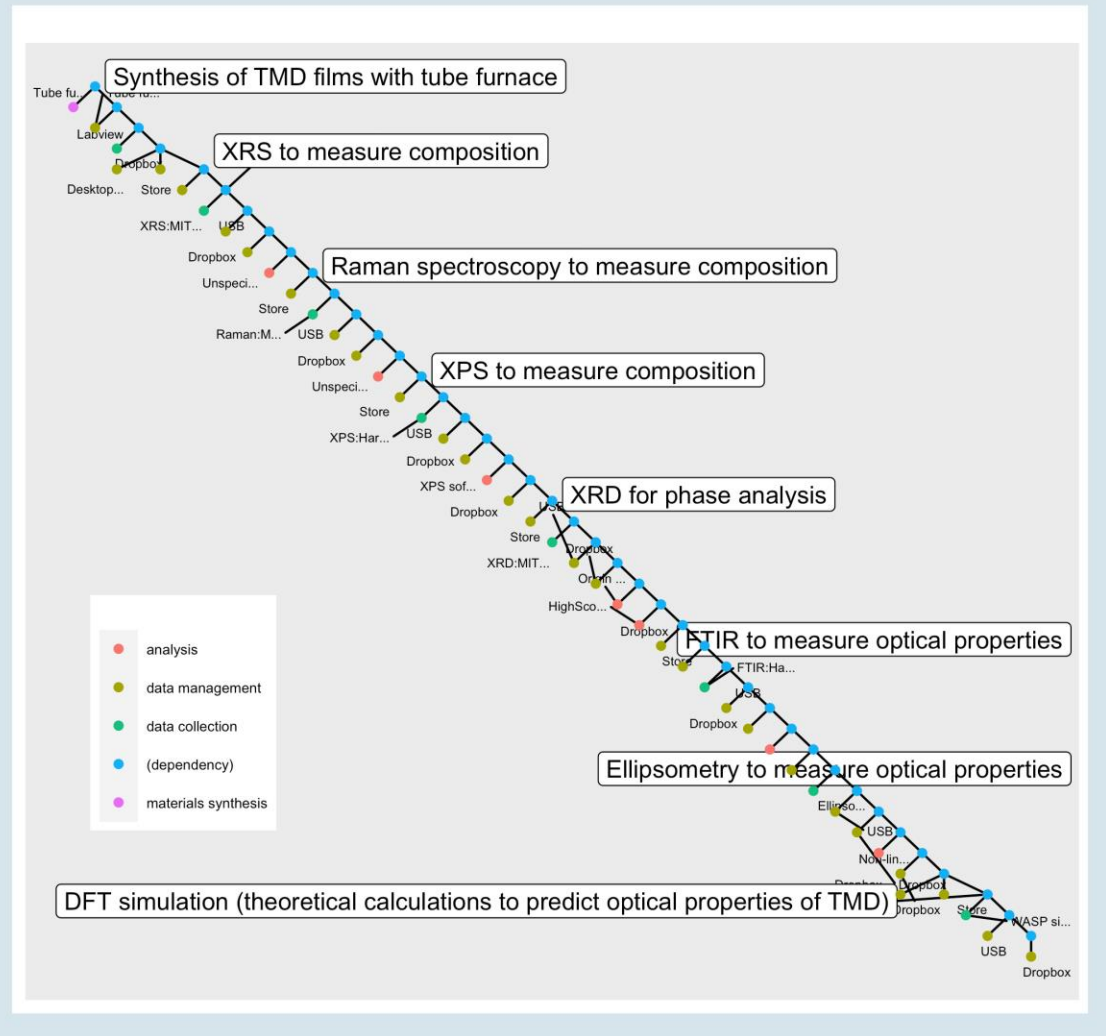
Research workflow graphs

- Four interviewees' workflows encoded and visualized in graphs
 - Hierarchical for each project
 - How the sample, data and metadata flowing through the research process
 - No connecting branches across projects

MIT
Libraries

Figure 2D: workflow overview -- project D

Workflow steps by phases, description, and type.



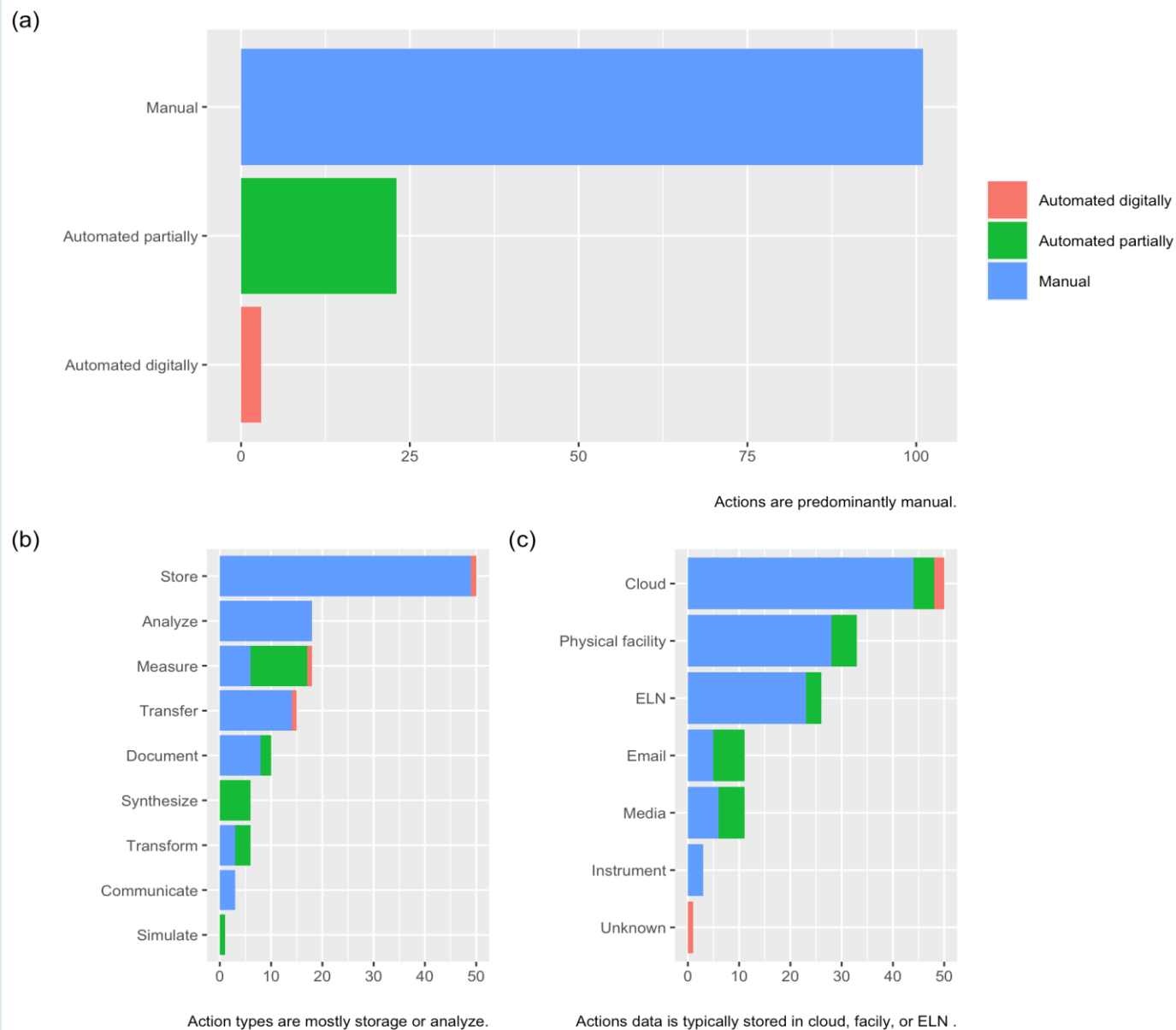
Notable Workflow Characteristics

- Dominated by manual activities
 - Especially for data transmission
- Action types are mostly storage and analyze
- Email and portable media being used for data storage substantially

MIT
Libraries

Figure 3: Selected characteristics of the workflow steps

Summarizes type of action described in each step, proximate data source, and level of automation.



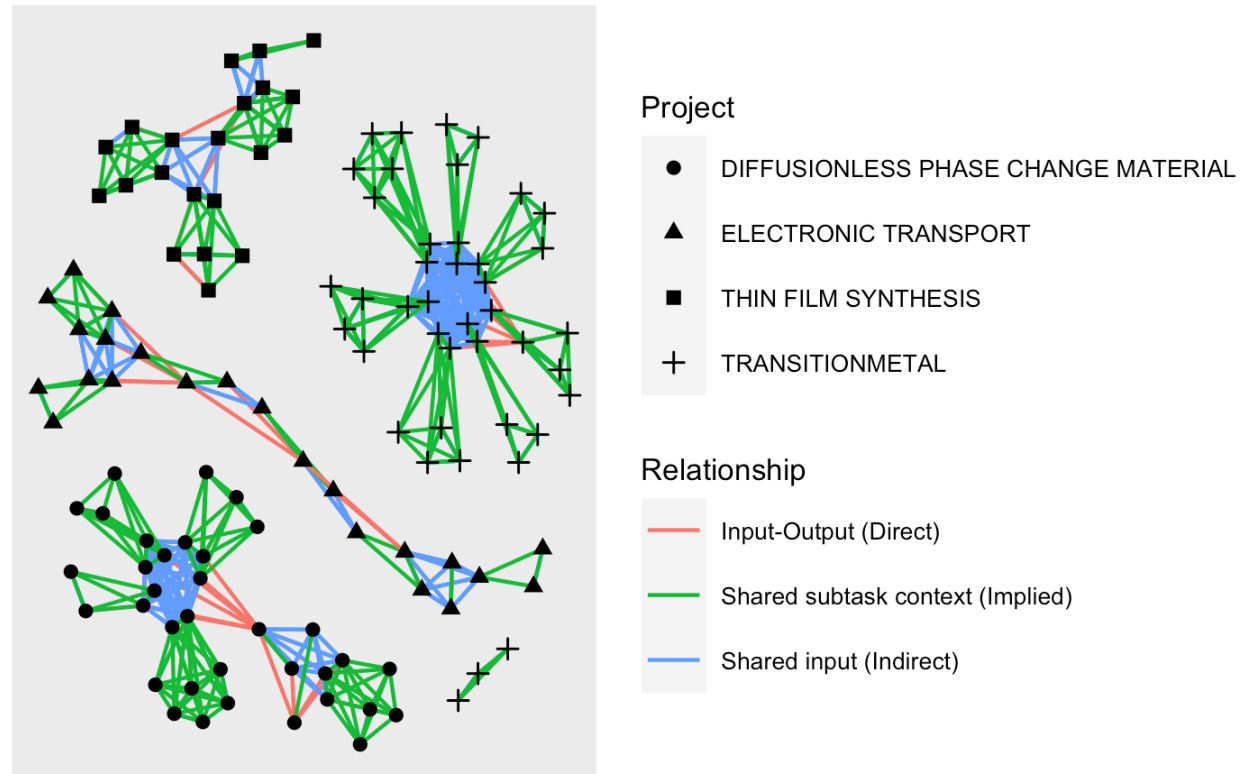
(N =127 workflow steps)

Project Information Exchange

- Implicit and indirect data/information exchange occurs frequently within projects, but does not connect projects

Figure 4: Project Information Exchange.

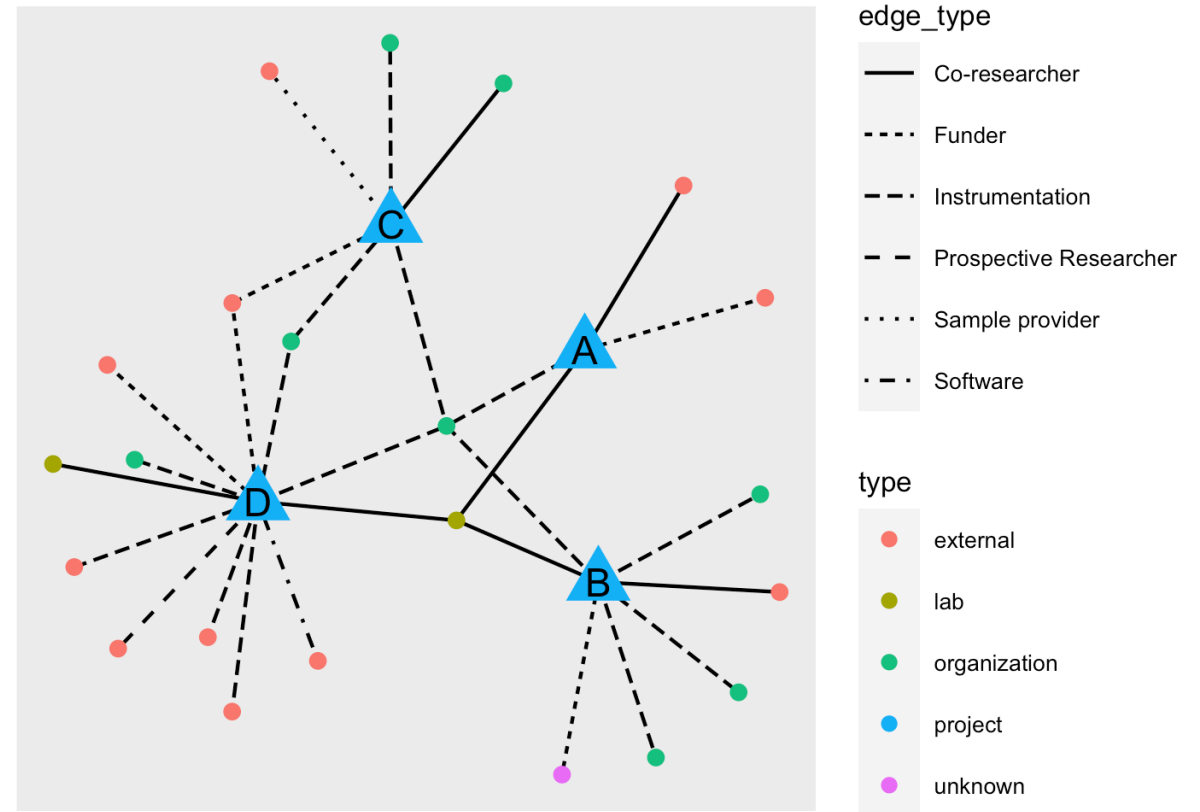
Implicit and indirect information exchange occurs frequently within projects, but does not connect projects.



Collaboration Networks

- Collaboration networks are also partitioned by project

Figure 5: Project Collaboration.



Comparison of Documented Group Practice with Recalled Individual Practices

Documented Procedures	Inconsistencies with Practice
<i>Information sharing</i>	Practices are predominantly consistent with documentation, although occasional lapses occur.
<i>Information security</i>	Practices are consistent with documentation.
<i>Information organization</i>	Practices are frequently inconsistent with documentation, however the instrument, username, data, and sample can often be identified by human inspection of the file and directory name.

Workflow graph

Workflow graph +
follow-up interview

Digital forensic of
shared drive

E.g. Of the 31929 deposited over two years of proximity, less than a quarter (23%) provided could be readily assigned a collection date, researcher, and instrument.

Internal Replicability

Sufficiency of Documentation Process

Documentation
integrated with
data

Data manually
documented in
a separate file

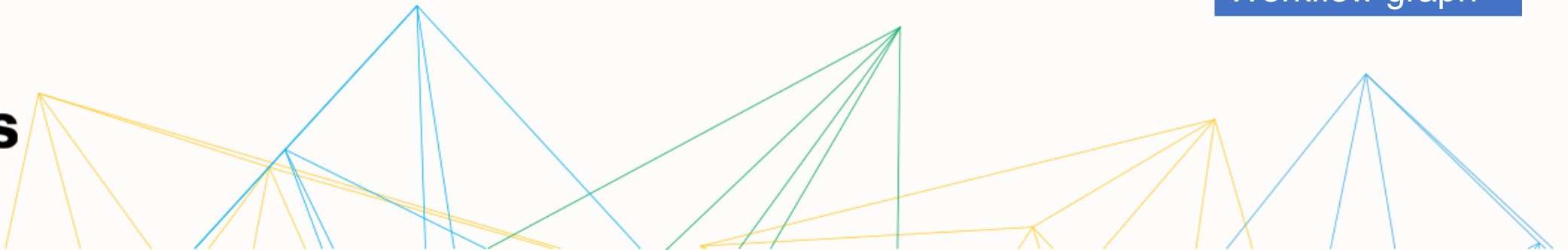
Documentation
derived from
process

Documentation
missing

	integrated	manual	implicit	missing
processed data	0 (0%)	8 (88.89%)	1 (11.11%)	0 (0%)
analysis	0 (0%)	8 (44.44%)	0 (0%)	10 (55.56%)
raw data	7 (50.00%)	5 (35.71%)	2 (14.29%)	0 (0%)

Workflow graph

MIT
Libraries



Robustness of Storage Practices

Proportion of output in group managed storage, by type.

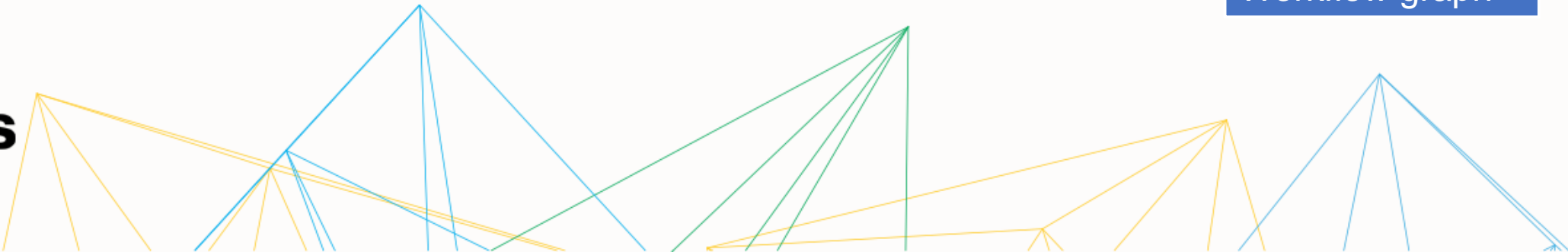
Type of research outputs	Percentage
metadata	44%
analysis	44%
raw data	79%
processed data	100%

At risk!

Note: processed data includes derived, linked and cleaned data; metadata includes configuration files, output logs, and manual documentation

MIT
Libraries

Workflow graph



Attribution Robustness

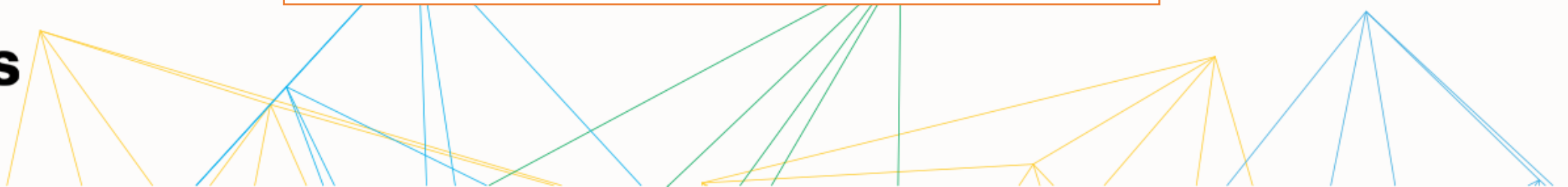
Undocumented Collaborators

Project	Undocumented Types	% undocumented contributors
A	Co-researcher	33%
B	[None]	0%
C	Co-researcher, Sample provider	40%
D	Instrumentation, Prospective Researcher, Co-researcher, Software	40%

Compared collaborators named by the interviewee and the list of collaborators detected through the workflow outputs and documentation

Workflow graph

**MIT
Libraries**

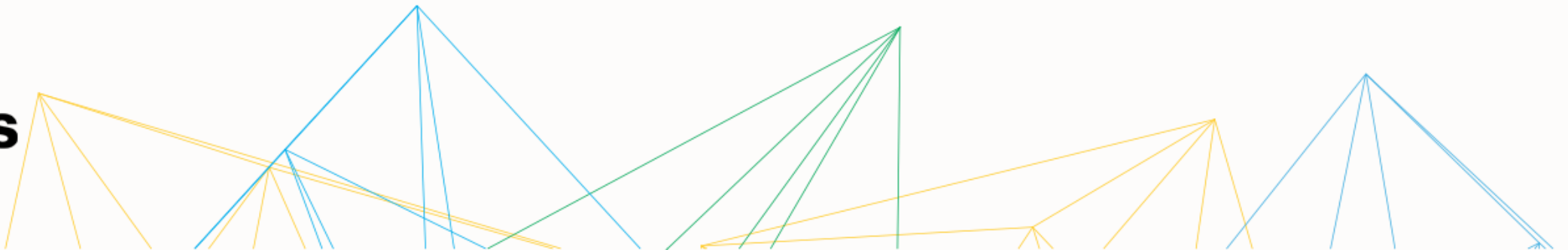


Recommendation: Auditing

- Automated audits
 - E.g. implementing scripts to
 - verify naming conventions being followed
 - verify backups
- Automated validation
 - Detect system failure
 - Flag unusual activities for further investigation
 - Correlate timing of lab notebook updates with the timing of data deposits into group storage system

Recommendation: Upgrading Infrastructure

- Reduce manual data transfer and operations
 - E.g. adopting USB-compatible mobile storage devices, including built-in wireless networking and data synchronization capabilities
- Eliminate the use of multiple storage locations
- Monitor target folders for automated synchronization via tools



Recommendation: Refining Practices

- Develop explicit practices around collaboration attribution systematically
 - Enhancing project documentation
 - Explicitly saving external contributors' work in the group storage
 - Defining contributors' roles according to taxonomy (e.g. CRediT)
- Develop explicit practices around reproducibility beyond the stage of raw data
 - Standardizing practices at commonly used external equipment
 - Establishing a group-shared location for metadata
 - Encouraging analyses to be conducted in a framework that builds reproducibility

Future Research

- How effective practices can be aligned with incentives, training, institutional coordination, and infrastructure improvement
 - Recognizing the value of FAIR data sharing
 - Realization of AI-powered automated research workflow
 - Embedded data curators with disciplinary expertise
 - Improvement of human-computer interaction, accessibility and security of cloud-based system

Acknowledgement

- Professor Rafael Jaramillo at MIT for his commentary and enabling access to lab records
- Members of the Jaramillo research group for participation in interviews
- MIT Libraries for the special fund and support to the project