



**HKU
Med**

LKS Faculty of Medicine
School of Biomedical Sciences
香港大學生物醫學學院

Research data management practices and challenges at a large comprehensive university: A cross-sectional mixed method study

Joshua W. K. Ho, PhD

Associate Professor | School of Biomedical Sciences
University of Hong Kong (HKU)

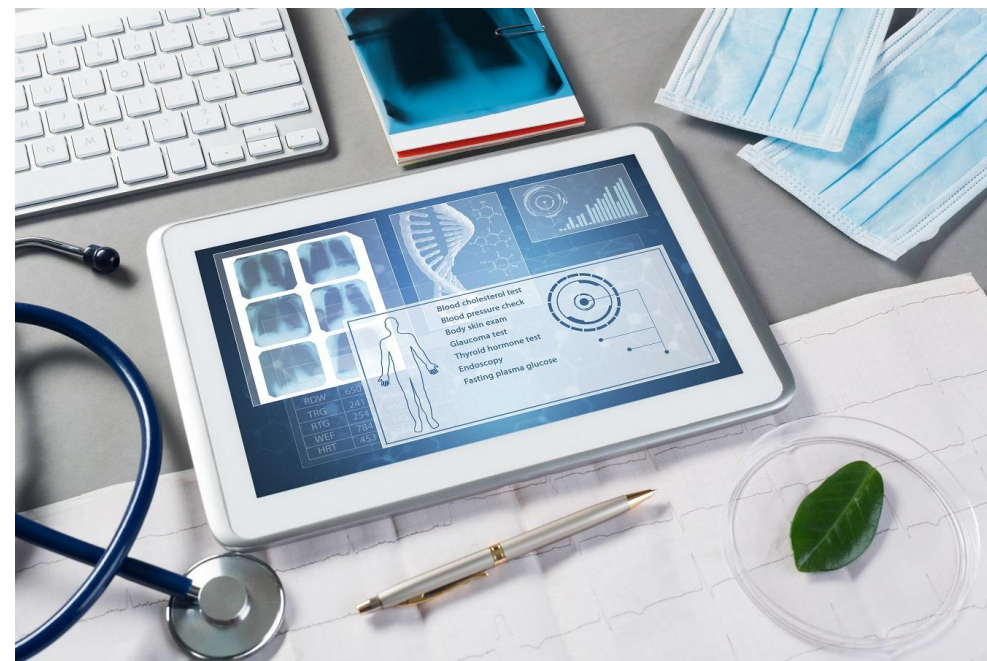
Co-director | BSc (Bioinformatics) Programme, HKU

Academic Lead (Bioinformatics Core) | Centre for
PanorOmic Sciences (CPOS), HKU

Deputy Director of EdTech (AI) | Li Ka Shing Faculty of
Medicine, HKU

Lead Scientist | Laboratory of Data Discovery for Health

jwkho@hku.hk | <https://holab-hku.github.io/>



Big data challenges in research data management



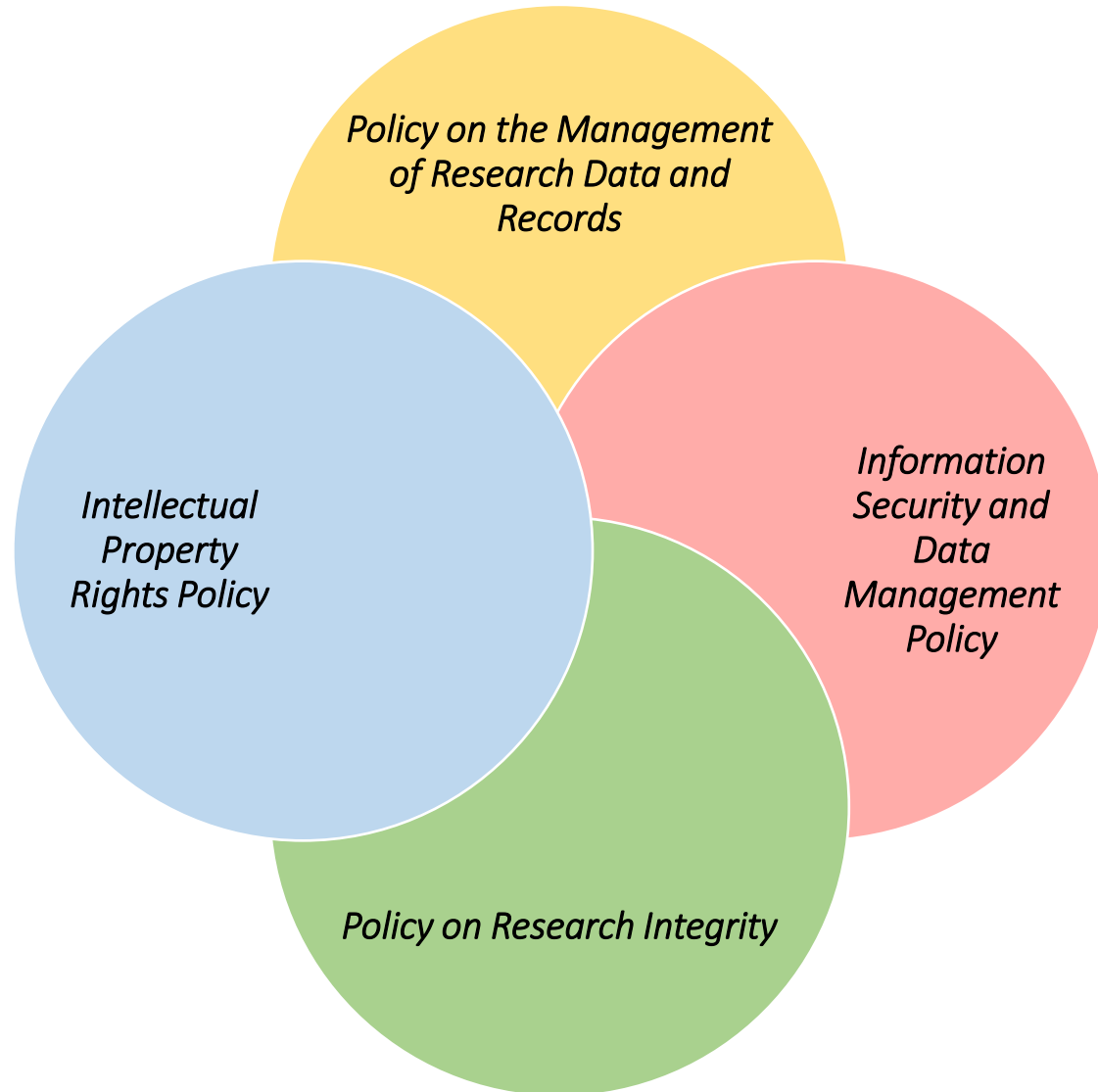
A professor had accumulated a large amount of human genomic data (>100TB) in their departmental computer server. Genomic sequencing and on-going data storage was continuously paid for by this professor via project grants. However, after the professor retired, the data was transferred to some external hard drives at this professor's department. In times, it was unclear if anyone knows how to retrieve all these valuable genomic data.

1. Who **owns** the data?
2. Where should these large data be stored in the **long-term**?
3. Who should **pay** for its long-term maintenance?
4. Who can decide whether these data **can be deleted**?
5. How and whether these data can be **re-used** by other people, given its use is compliant to the original IRB / informed consent?
6. Can de-identified genomic data be **re-identified**?

HKU policies on research data management



**HKU
Med**



Research data needs to be managed well, and there are at least the following reasons:

Promote and safeguard research integrity (reproducibility and validity of research findings)

<http://www.rss.hku.hk/integrity/research-data-records-management>

<http://www.rss.hku.hk/integrity/rcr/policy>

Protect intellectual property rights (who invented what and what date/time)

<http://www.rss.hku.hk/contracts/ipr>

Enforce information security

<https://isd.hku.hk/>

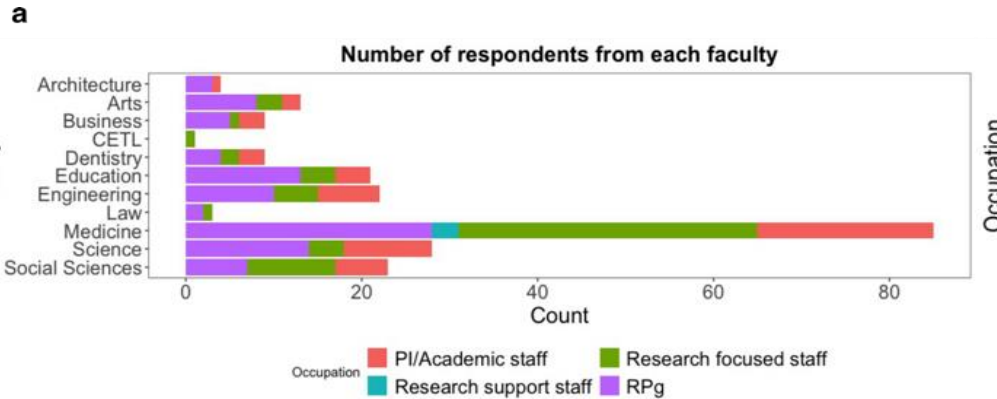
Goals of research data management



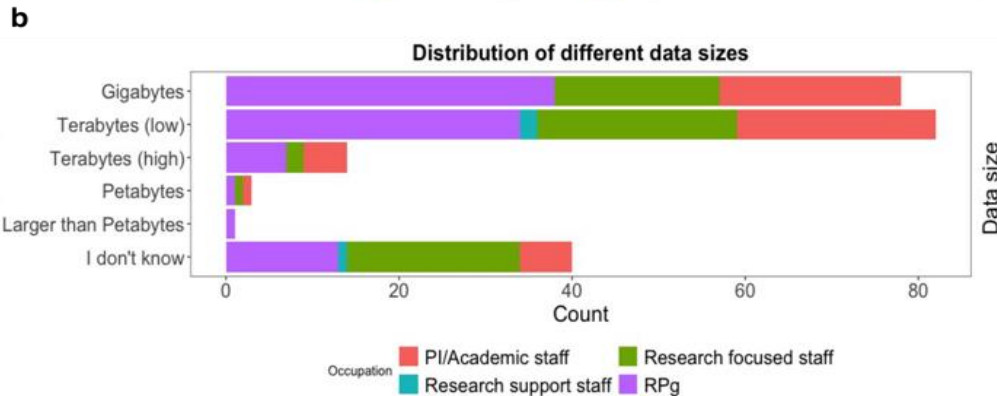
HKU
Med

- Extraction of value
 - Well organised data can be a valuable asset for researchers, institutions, the scientific community, and the general public
- Research integrity
 - Data validity and reproducibility is central to safeguard us from research misconduct, and promote sharing and transparency
- Data security
 - Safeguard personal identifiable or re-identifiable data of the research subjects
- Intellectual property protection
 - Retain information critical for determining priority, novelty, and inventorship of an intellectual property

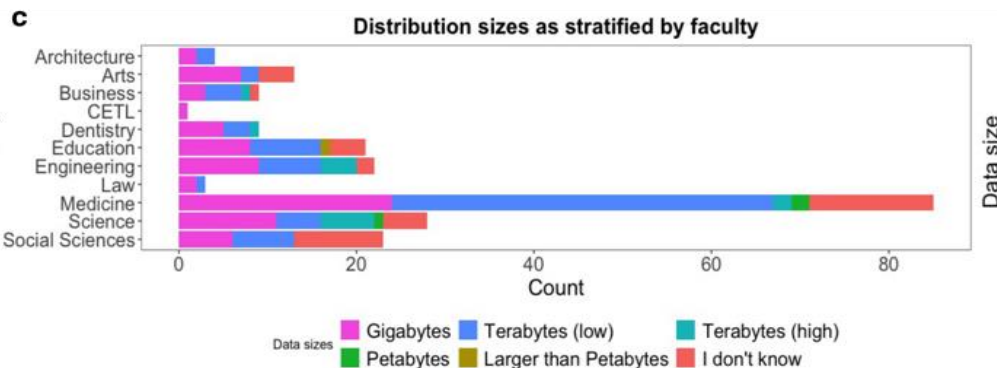
Big data users can be found everywhere



- Online questionnaire survey during May 2022.
- University wide recruitment (HKU)
- 218 completed and valid respondents



- Among the 218 respondents:
- 38% generate 1-500Tb of data
 - 6.4% generate 500-1000Tb of data
 - 1.3% generate petabytes of data
 - 18.3% did not know
 - Big data users can be found across most Faculties

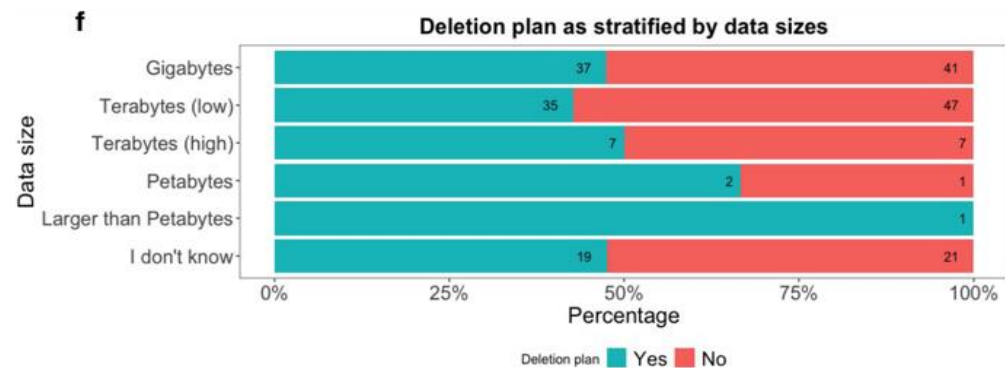
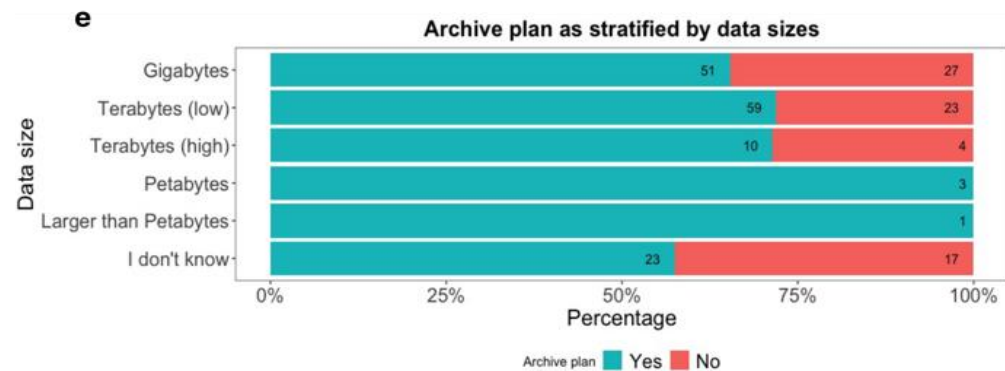
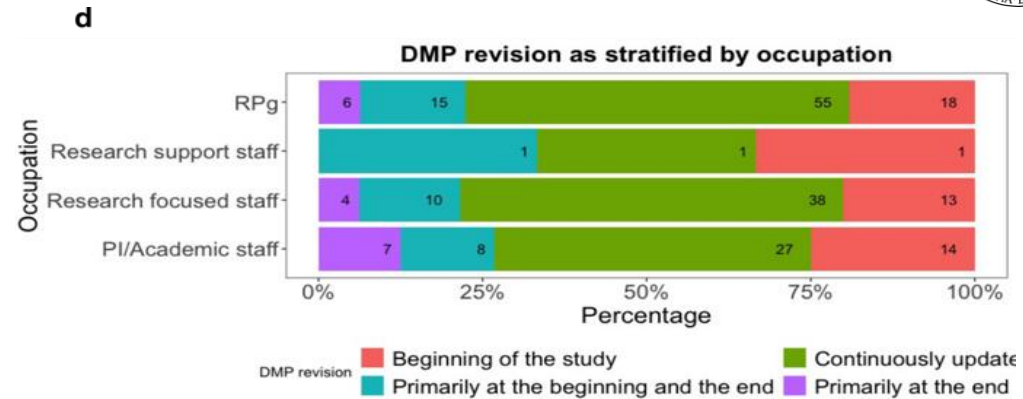


Insufficient planning for data archival and removal



Among the 218 respondents:

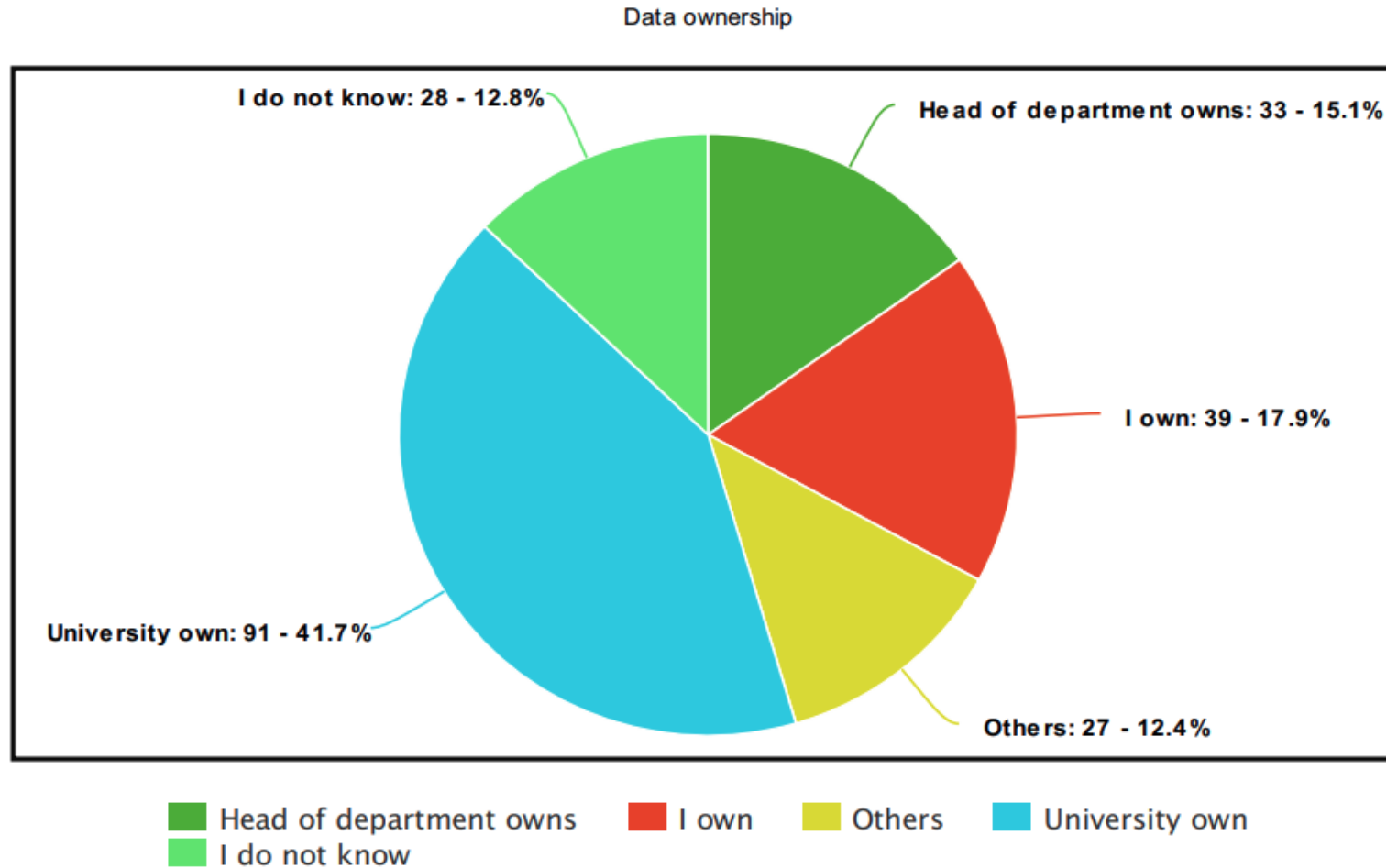
- Only 67% had a concrete plan for archival of research data
- Only 46% had a concrete plan to delete their archived data
- Similar proportions among small or big data generators



Ambiguity about ownership of research data



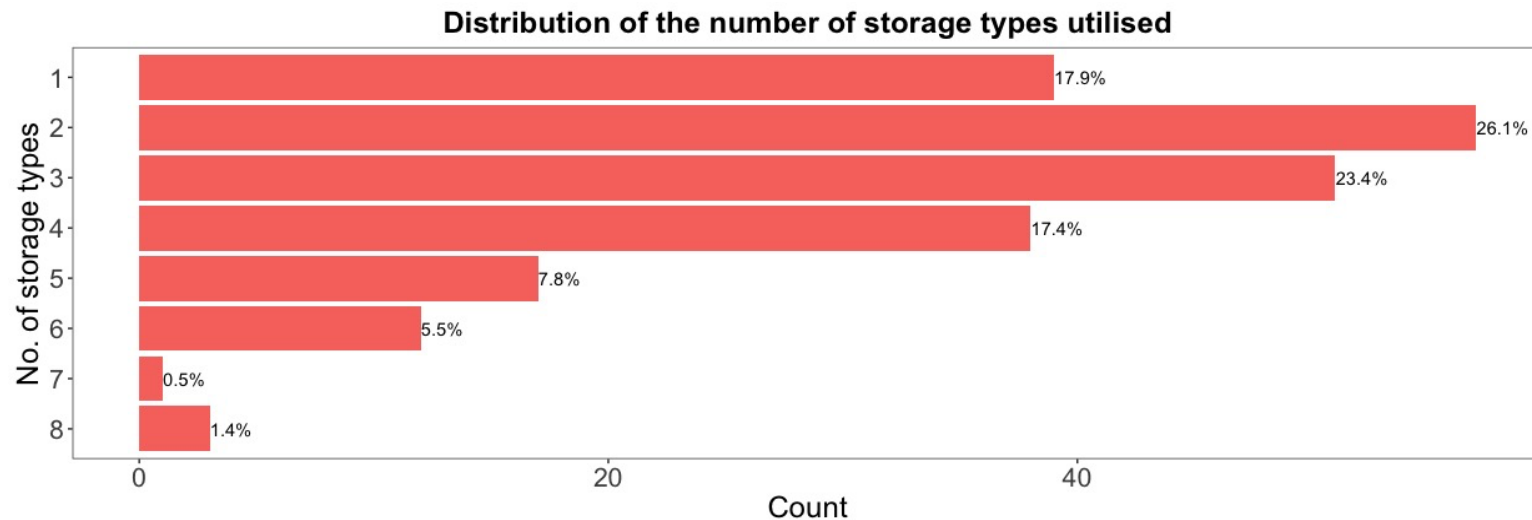
There is a diversity of understanding about ownership of research data.



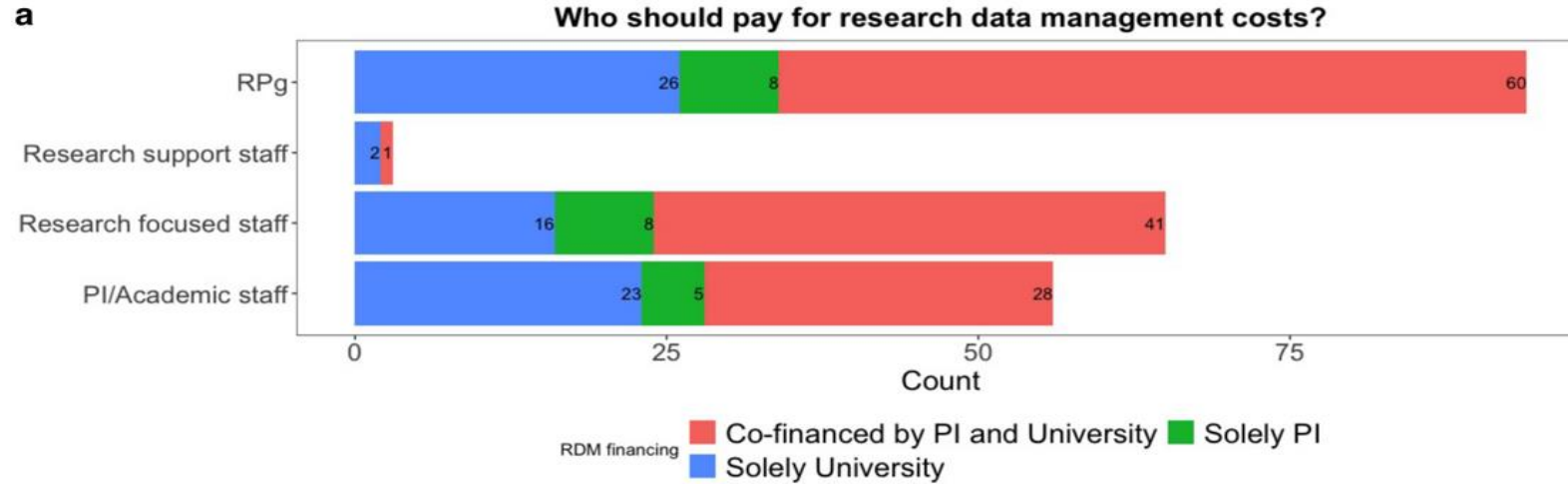
A variety of data storage media is used



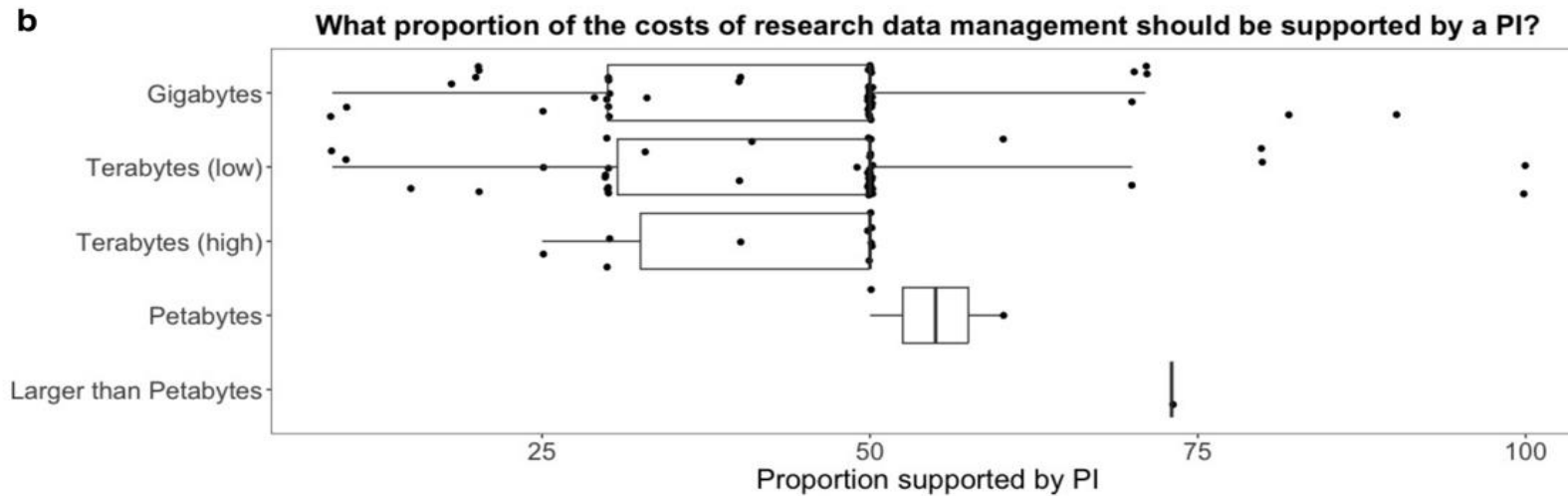
Storage types	No. of times each storage method was used (percentage)
Cloud based storage	139 (63.8%)
Personal computer	137 (62.8%)
USB / Hard disk	100 (45.9%)
Lab computer	92 (42%)
Laboratory server	63 (29%)
Online data repositories	44 (20%)
Physical Laboratory Notebook	35 (16%)
Electronic Laboratory Notebook	31 (14%)
Network Attached Systems (NAS)	1 (0.5%)
Physical Storage	1 (0.5%)



Who should fund RDM?



>50% of respondents agree that RDM cost should be supported by both the PI and the university.



Perceived lack of central RDM infrastructure is a major challenge



HKU
Med

Problems self-reported by researchers	No. of times selected (percentages)
Unsure of current infrastructure provided by HKU	49 (22%)
Lack of systems / infrastructure to store data	42 (19%)
Unsure of infrastructure available which are suitable to my research needs.	40 (18%)
Collaborators and HKU do not use the same system	36 (17%)
Hard to share data seamlessly	35 (16%)
Unsure on which data sets should be kept and which data sets should be deleted	34 (16%)
Costly external data storage	33 (15%)
Data size is too big to use university provided infrastructure	22 (10%)
Unsure where to publicise my datasets to share my data	18 (8.3%)
No problems	2 (0.9%)

Thematic analysis on written responses



1. Lack of clarity in the practical steps in RDM
 - “I don't understand to what extent the research data has to be managed. Is it raw data? or processed data? and do I need to write a procedure for how to convert raw data to processed data? and how to properly describe a research data so that others will know what this is for?”
2. Require centralised and customisable data storage infrastructure
 - “When planning data protection policy, it is important to allow for a flexible approach and recognise that scholars (particularly those in Social Sciences/Arts) work with types of data (such as handwritten field notes) generated in a wide range of social settings. Sharing of datasets is not always possible in all instances. “

In summary



HKU
Med

Our research identified some challenges related to research data management in a university setting

1. Lack of planning for long terms data archival or removal
2. Data ownership can be unclear
3. Perceived or actual lack of RDM infrastructure is a major challenge

Acknowledgements



HKU
Med

This research project was supported in part by the University Research Committee (URC) of the University of Hong Kong (HKU).

Joint work with

- Nikkie Stables
- Ian Ooi
- Lydia Ma

Support/advice:

- Danny Chan

