(RESEARCH) DATA MANAGEMENT IN THE MODERN ERA

Jeffrey Thompson, PhD
Chief Research Informatics Officer
Associate Vice Chancellor of Research Data and Analytics
University of Kansas Medical Center


DATA MANAGEMENT IS A VAST TOPIC

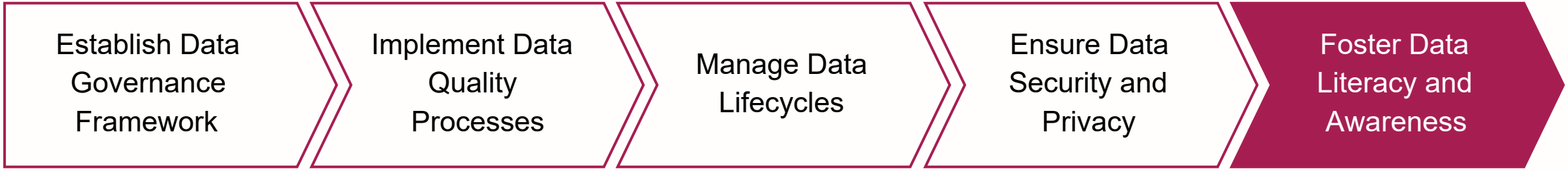
DATA MANAGEMENT IS FUNDAMENTAL TO RESEARCH BUT IS OFTEN OVERLOOKED

1. Data Management Trends
2. Playing FAIR
3. Being Reliable
4. Using Your (Artificial) Intelligence

Section 1

DATA MANAGEMENT TRENDS

DATA GOVERNANCE STRATEGIES



Define clear roles, responsibilities, and decision-making processes for data management and oversight. Identify data owners, stewards, and custodians to ensure accountability.

Develop and enforce policies and procedures to maintain data accuracy, completeness, and consistency. Implement data validation checks, data cleansing routines, and data profiling techniques.

Implement a comprehensive data lifecycle management strategy, including data creation, storage, retention, archiving, and secure disposal. Ensure alignment with regulatory and compliance requirements.

Establish robust data security measures, such as access controls, encryption, and data masking, to protect sensitive information. Comply with relevant data privacy regulations and guidelines.

Provide training and resources to promote data literacy and awareness among employees. Empower users to understand and effectively utilize data for decision-making.

DATA MANAGEMENT TRENDS

- **Unification of Data Sources**

  Consolidating disparate data sources into a centralized, integrated platform to enable a more holistic view of data assets.

- **Democratization of Data Access**

  Empowering users with self-service tools and intuitive interfaces to access, analyze, and derive insights from data without relying solely on IT support.

- **Adoption of Data Governance Frameworks**

  Implementing structured policies, processes, and controls to ensure data quality, security, and compliance across the organization.

- **Leveraging Artificial Intelligence and Machine Learning**

  Applying advanced analytics techniques to automate data processing, generate predictive insights, and uncover hidden patterns in large datasets.

- **Emphasis on Data Lineage and Provenance**

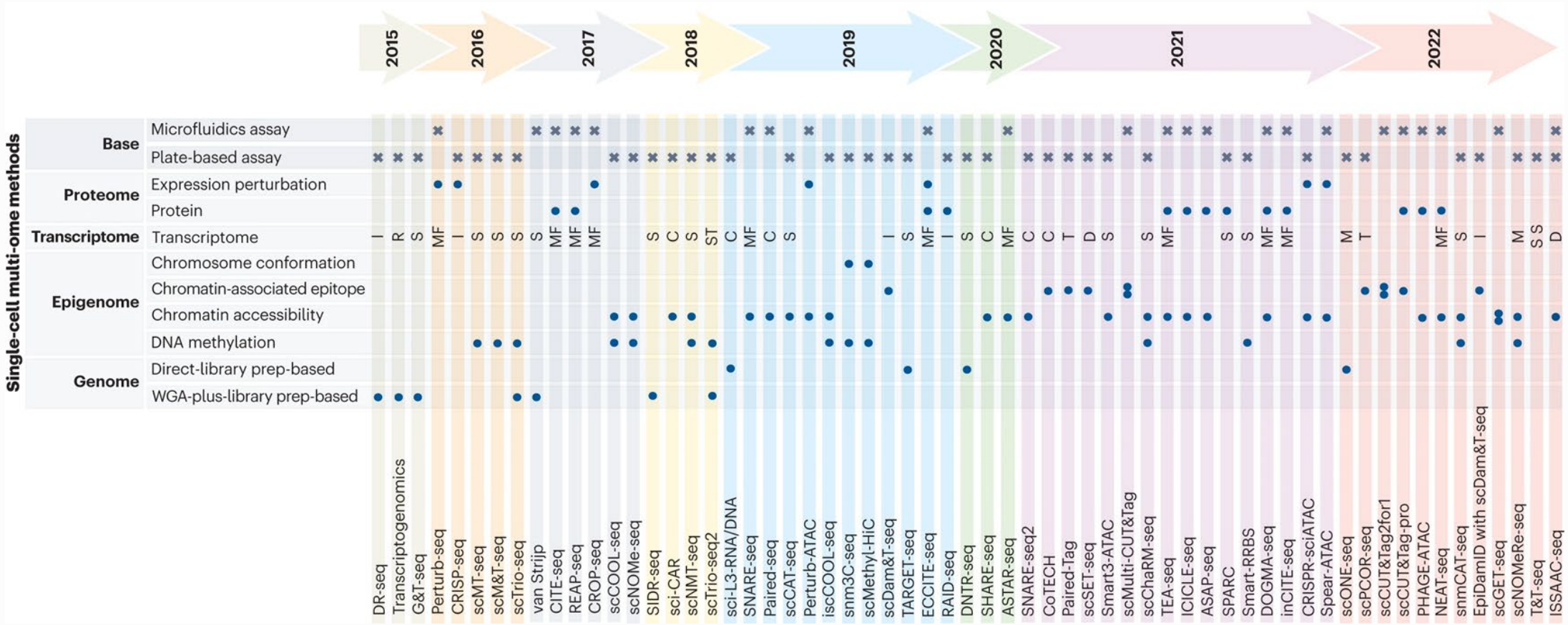
  Capturing and maintaining detailed information about the origin, transformation, and usage of data to enhance transparency and enable better data-driven decision-making.

IT'S GETTING MORE CHALLENGING TO MANAGE RESEARCH DATA

Although the challenges are not really new, the trends continue.

- There is a greater **variety** of data than ever
- The data are of greater **volume** than ever
- The data are coming at higher **velocity** than ever

THERE HAS BEEN PHENOMENAL GROWTH JUST IN THE AREA OF SINGLE-CELL AND SPATIAL OMICS



Vandereyken, K., Sifrim, A., Thienpont, B. et al. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet 24, 494–515 (2023). https://doi.org/10.1038/s41576-023-00580-2

INCREASINGLY, DATA INTEGRATION IS REQUIRED

- As a result, the legacy approach to studies is starting to break down
- KUMC Research Informatics might be asked to provide linked data from medical records, genetic testing, tumor registry, ECG, imaging, or more
- For research, it is very powerful to be able to do so, but it requires a great deal of work to efficiently manage such a system
- That’s why it helps if data are FAIR

“DOESN'T MATTER HOW MUCH DATA YOU HAVE, IT'S WHETHER YOU USE IT SUCCESSFULLY THAT COUNTS”

BERNARD MARR

Section 2

PLAYING FAIR

DATA MANAGEMENT IS NOT A LUXURY

- For many, it is a requirement.
- For example, under the NIH 2023 Data Management and Sharing Policy, researchers are expected to maximize appropriate sharing of data.
- There are more stringent regulations for some.
- But regardless of regulatory requirements, you should do more than pay lip service to data management.

HOW LONG DOES THE NIH REQUIRE A RESEARCHER TO RETAIN DATA?

- 3 years following the *closeout* of a grant
- For a five-year research project, you might need to retain data for 8 years!

THE VALUE OF DATA GO FAR BEYOND COMPLIANCE

You

University

Society

A TRUE STORY PART I

- A researcher collected some sequencing data
- The results were great and resulted in a paper
- Five years later they did proteomics using the same samples
- They thought it would be great to integrate the new data with the old

A TRUE STORY PART II

- The researcher had their bioinformatics collaborator download the raw data from the lab
- The lab used different labels for the samples than what were recorded in the lab notebook
- The researcher couldn't find the spreadsheet that had both sets of labels
- The proteomics data didn't have a data dictionary, making it unclear what column contained what data

THE MORAL OF THE STORY

- Your data are extremely valuable
- Organizations benefit from well -developed data management practices
- Data management is complex, and you care more about your data than anyone else
- Do you know where your data are? You should!

FAIR DATA

- Findable
- Accessible
- Interoperabe
- Reusable

KUMC FAIR DATA EXAMPLES

- Findable – KUMC Research Informatics has created a Digital Research Platform that includes a centralized catalog that can be used to track research data, furthermore, we provide tools such as C3OD and HERON
- Accessible – Data from the Digital Research Platform can be retrieved using Delta Sharing (an open standard)
- Interoperable – KUMC provides data to the All of Us Research Program using the OMOP Common Data Model
- Reusable – KUMC's research catalog can be used to enhance data with meta -data, track licensing, and more

META DATA ARE CRUCIAL

- Meta data describe your data
- They are key to knowing what you have and finding it again later
- They are extremely helpful for collaboration, so that the person analyzing the data knows what they are looking at

Section 3

BEING RELIABLE

DATA PRIVACY AND SECURITY

Introduction to Data Privacy and Security

In today's digital world, where vast amounts of personal and sensitive information are collected, stored, and shared, the importance of data privacy and security has become paramount.

Threats to Data Privacy

Cybercriminals, data breaches, unauthorized data sharing, and lack of transparency in data collection practices pose significant threats to data privacy. Organizations must take proactive measures to protect people's personal information.

Principles of Data Security

Data security involves the implementation of technological and organizational measures to protect data from unauthorized access, modification, or destruction. Key principles include confidentiality, integrity, and availability of data.

Regulatory Frameworks and Compliance

Governments and organizations have introduced various regulations, such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA), to ensure the protection of personal and sensitive data.

CAN YOU TRUST YOUR DATA?

Good data are the foundation of meaningful results, but many things can affect their quality and reliability

- Data entry errors are common
- System are upgraded
- High velocity data from sensors and other sources can have unexpected gaps or spurious signals
- People feel pressure to produce results
- Disgruntled staff may strike at data

THE THREATS ARE REAL

KUMC has seen examples of all of these

- Manual transfer of data to spreadsheets leads to errors, like labeling a mouse as being in a different experimental group than it was
- Sensor data from CGMs or other continuous data can have gaps or spurious signals when sensors are dislodged, machines are connected/disconnected, etc.
- Academic misconduct.
- A fired employee once tried to leave KUMC with all of a project's data.

DATA MANAGEMENT SOLUTIONS

Some examples

- Important data should have an audit trail (e.g. CTMS, REDCap, or Lab Archives, but custom solutions are available for any data)
- Data monitoring can be used to ensure incoming data are reliable
- Back up data to the cloud, where different levels of storage are available
- Use data integrations to automatically enter data whenever possible (such as FHIR or REDCap API).

Section 4

USING YOUR (ARTIFICIAL) INTELLIGENCE

DATA CHALLENGES WITH AI

- AI training often relies on large amounts of data, so we need systems capable of pumping those data to the optimizer without choking.
- Data now often come from systems that use AI. For example, AI might be used to identify suspicious lesions in an image. How can we ensure that they are accurate?
- Generative AI systems can ‘hallucinate’. This is a byproduct of the way they work. This means, even if they are asked to summarize something, they may insert items that were not present in the data. It is becoming popular to use an ambient AI scribe for medical documentation, so how can we be sure it is faithful?

OTHER AI DATA CHALLENGES

- The use of AI to interrogate data, while promising, can hide the process of data science from us, which might cause us to miss data issues that bias the results.
- In order for an AI model to be accurate, the data it was trained on must reflect the current state of the world it works in. This means collecting data to continuously assess model validity.
- Currently, things like decisions about equipment purchasing don't typically consider AI. For example, an institution might find the best overall deal for ECG equipment. But what if there are downstream AI algorithms that take data from those machines? They were trained on input that looks a certain way.

WHY ARE THESE DATA MANAGEMENT CHALLENGES?

Data management sits at the intersection of where the data are coming from and how they are being used.

DATA MANAGEMENT SOLUTIONS

- Use distributed computing and a hyper -scaler to adjust to meet the needs of a project
- Centralize data management for research to provide a consistent approach
- Use meta-data to 'tag' data elements that come from AI
- Have humans make decisions and track whether a human has reviewed generative AI elements for accuracy

MONITOR FOR ACCURACY OF OUTPUT AND QUALITY OF DATA

- Continuously monitor AI output for accuracy, collecting metrics. If metrics drift past a certain threshold, automatically alert someone.
- Automatically monitor data needed by AI to detect data quality issues that might come from upstream (Have units changed? Are data missing? Is the volume different?).

Section 5

SUMMING THINGS UP

DATA MANAGEMENT IN THE MODERN ERA – PLAYING FAIR

- Whenever possible, budget for data management support. Yes, it's an extra cost that's hard to squeeze in, but it will pay dividends in the future.
- Work with your research data team, whoever they are, to make sure your data are findable, accessible, interoperable, and reusable. For you, the most important thing you can do is give your data rich meta -data.
- Know where your data are.

DATA MANAGEMENT IN THE MODERN ERA – BEING RELIABLE

- For any data that have value to you, store them in a system that is compliant with regulations, but also has audit trail capabilities.
- Use zero trust. Grant minimum privileges.
- Ask your data management team to configure alerts when data are modified unexpectedly.

DATA MANAGEMENT IN THE MODERN ERA – USING YOUR ARTIFICIAL INTELLIGENCE

- Tag data coming from AI sources (preferably, have the data management team tag them in a centralized resource)
- Monitor data to detect when sources change.
- Monitor the quality of AI – don't assume it will always keep working like it did.
- If possible, do this all in one place.