DATA REUSE IN AGRICULTURAL GENOMICS RESEARCH:

Challenges, solutions, and a case study

Alenka Hafner

PhD candidate in Plant Biology, Penn State University

AgBioData Data Reuse WG

MOTIVATION FOR DATA REUSE WG

Sequence-based datasets + WWW + Open Science movement

□ a lot of data out there

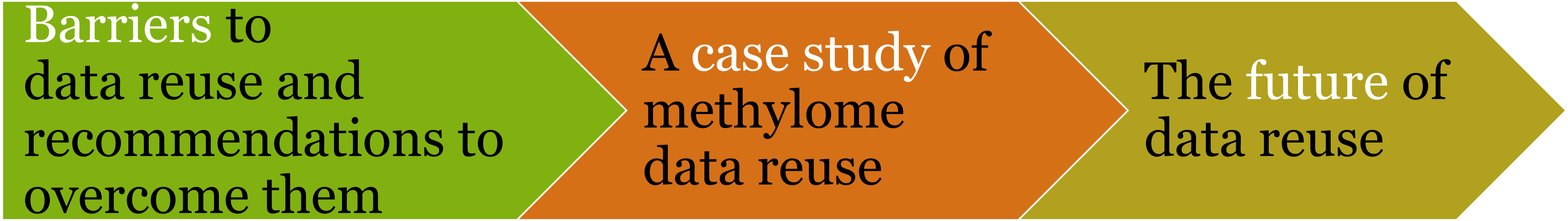
No dataset is perfect

Data are often underutilized due to insufficient metadata and other challenges related to the FAIR data standards

Objective for WG: to identify bottlenecks in data reuse and critical needs to propose solutions for agricultural genomics community

Personal objective: understand the lack of reusability of methylome data

AgBioData



Barriers to data reuse and recommendations to overcome them

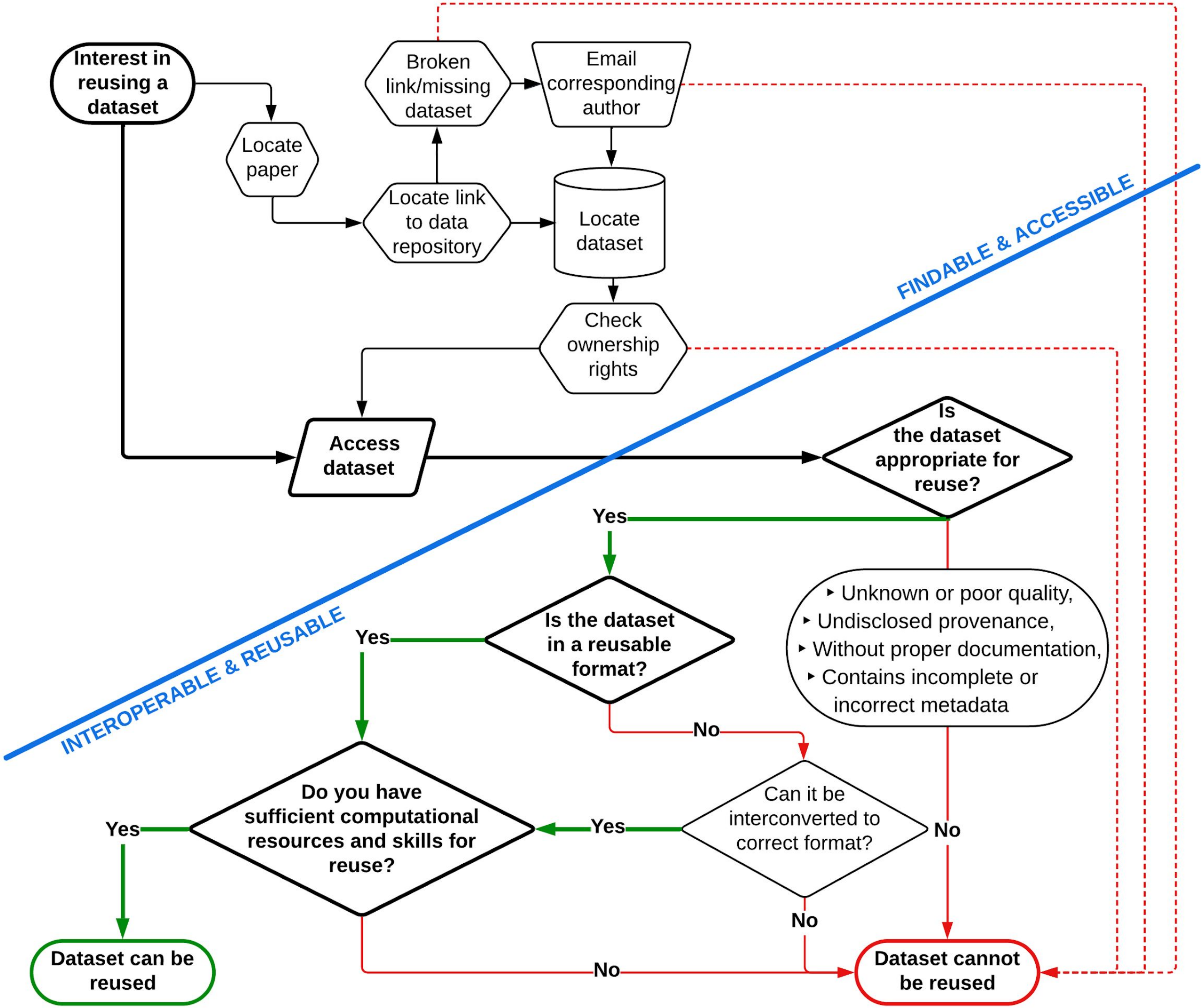
A case study of methylome data reuse

The future of data reuse

DATA REUSE =

the practice of utilizing existing data for

a novel scientific purpose beyond their original scope


Interest in reusing a dataset
Locate paper
Broken link/missing dataset
Email corresponding author
Locate link to data repository
Locate dataset
Check ownership rights
FINDABLE & ACCESSIBLE
Access dataset
Is the dataset appropriate for reuse?
Yes
Is the dataset in a reusable format?
Yes
INTEROPERABLE & REUSABLE
‣ Unknown or poor quality,
‣ Undisclosed provenance,
‣ Without proper documentation,
‣ Contains incomplete or incorrect metadata
No
Do you have sufficient computational resources and skills for reuse?
Yes
Can it be interconverted to correct format?
No
Yes
No
Dataset can be reused
No
Dataset cannot be reused

1. BARRIERS TO DATA REUSE & RECOMMENDATIONS TO OVERCOME THEM

Data quality

Metadata

Data availability

Interoperability

Data ownership

User skill and resources

DATA QUALITY: STANDARDS AS A SOLUTION

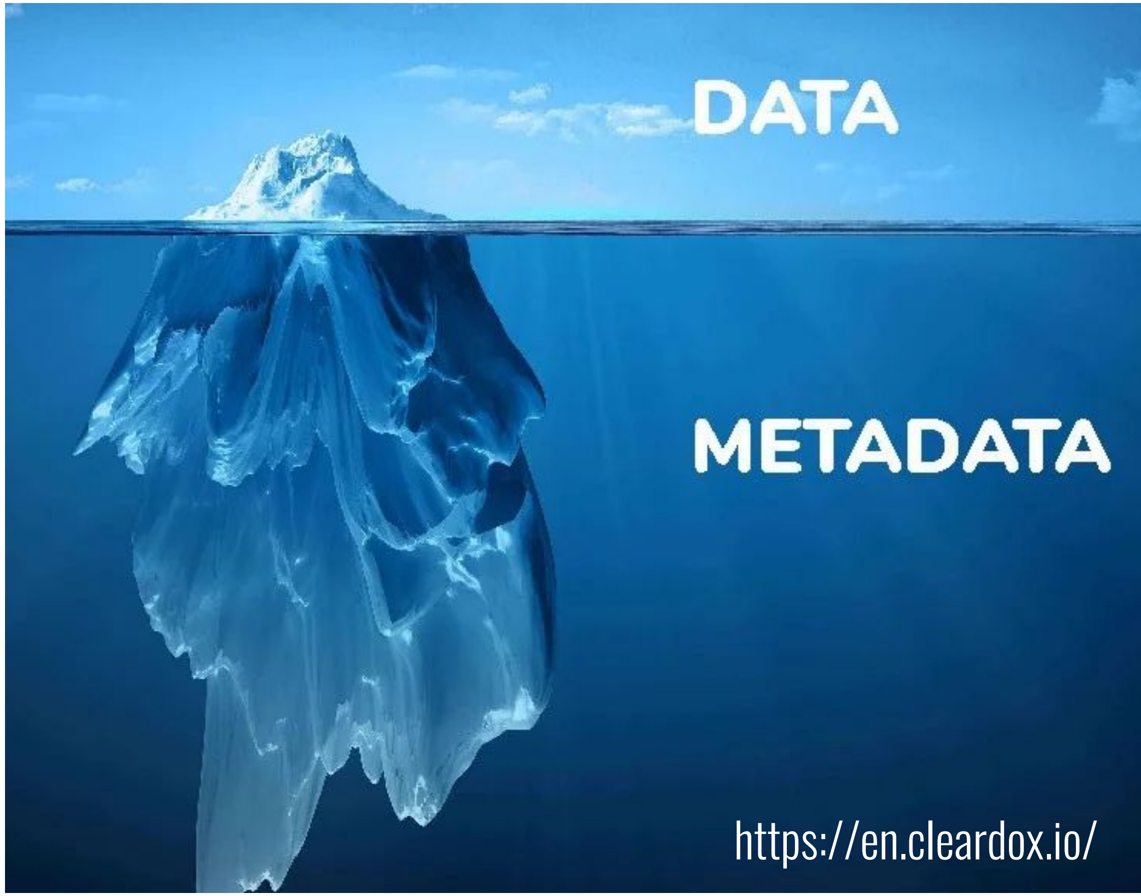
- Data made publicly available regardless of quality □ a (subjective) decision on suitability for reuse
- Factors to assess:
  - coverage,
  - depth,
  - technical and biological replication,
  - tissue type,
  - sample collection method,
  - extraction method and library preparation,
  - sequencing technology,
  - other dataset properties
- Limited scope of existing standards (even for common data types and organisms)

□ difficult to obtain experimental and computational protocols for informed reuse & meta-analyses

- **More protocol, pipeline, and statistical standards needed in agricultural genomics field**

ON THE ROAD TO COMPLETE METADATA: INCENTIVES

- Limited/incomplete/missing metadata submission templates
- Submission requires work □ Trade-off between collecting all some metadata via a lenient submission system and mandating comprehensive metadata
- **Incentives are needed! E.g., data citations…**

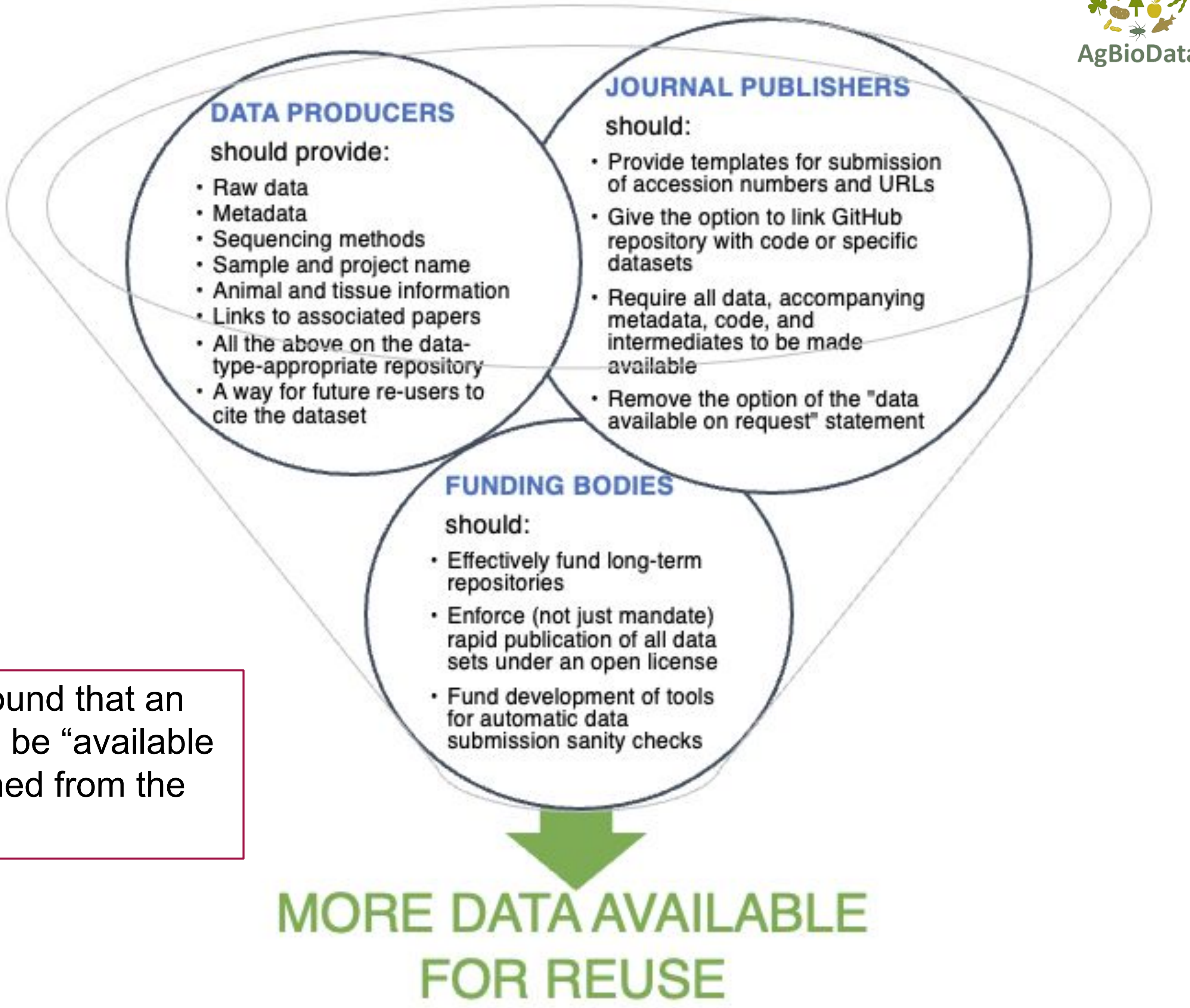


https://en.cleardox.io/

AgBioData

BRIDGING THE DATA AVAILABILITY GAP: A ROLE FOR ALL STAKEHOLDERS

Survey of *Science* and *Nature* in 2021 found that an alarming less than 50% of data stated to be “available upon request” could be effectively obtained from the original authors (Tedersoo et al., 2021)



DATA PRODUCERS

should provide:

- Raw data
- Metadata
- Sequencing methods
- Sample and project name
- Animal and tissue information
- Links to associated papers
- All the above on the data-type-appropriate repository
- A way for future re-users to cite the dataset

JOURNAL PUBLISHERS

should:

- Provide templates for submission of accession numbers and URLs
- Give the option to link GitHub repository with code or specific datasets
- Require all data, accompanying metadata, code, and intermediates to be made available
- Remove the option of the "data available on request" statement

FUNDING BODIES

should:

- Effectively fund long-term repositories
- Enforce (not just mandate) rapid publication of all data sets under an open license
- Fund development of tools for automatic data submission sanity checks

MORE DATA AVAILABLE FOR REUSE

TOWARDS INTEROPERABILITY: DATA FORMATTING

- Our community has converged on (meta)data standards for data file types:

  FASTQ, SAM/BAM, VCF, GTF, GFF3, BED, …

- Issue: ~~lack of standards~~ □ consistency of use
- Reference genome mapping can be an issue down the line
- “Backwards compatibility”: outdated lab and sequencing methodology can be combated through extensive metadata (https://www.protocols.io)

DATA OWNERSHIP & SHARING REQUIREMENTS

- Challenge: Having access to relevant, affordable study populations from breeding companies that can also be shared publicly as sequence or genotype data
- **Already many sharing requirements + 2026 mandate to make research funded by the USA government publicly available**

USER SKILL LEVEL & RESOURCE AVAILABILITY

- A recent study (LaFlamme et al., 2022) shows **that skill or perceived ability** was identified by many participants as a **major factor** influencing reuse behavior.
- 2017-2018 global survey: most scientists exhibited “**high and mediocre risk data practices**” (Tenopir et al., 2020).
- US-based institutions: computational resources likely not the limiting factor □ it’s skill level
- **Education programs** for awareness-raising and good practice training needed
- **Incentives (!):**DataWorks! Prize (https://www.herox.com/dataworks)

2. A CASE STUDY OF METHYLOME DATA REUSE

- Sample provenance
- Type of replicate

- Consistency of formats
- .pdf

- Computational resources
- Barrier of entry

Data quality | Metadata | Data availability | Interoperability | Data ownership | User skill and resources

- Depth of sequencing
- Experimental design
- Tissue type
- **REPLICATES!**

- Pipeline intermediates
- Code
- ”Available on request”
- .pdf

- Crops?

Hafner, A., Mackenzie, S. Re-analysis of publicly available methylomes using signal detection yields new information. *Sci Rep* **13**, 3307 (2023). https://doi.org/10.1038/s41598-023-30422-4

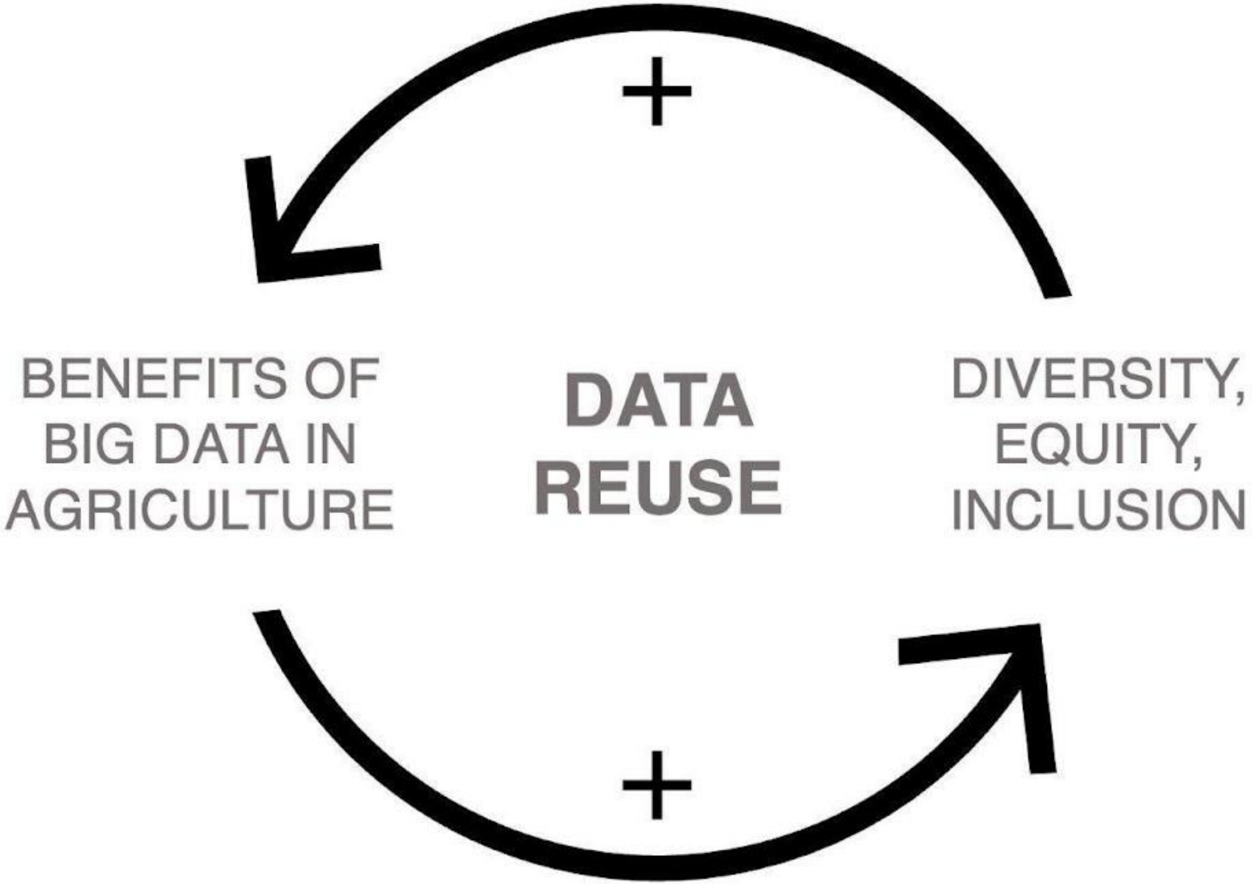
3. THE FUTURE OF DATA REUSE

The importance and benefits of equity and inclusion

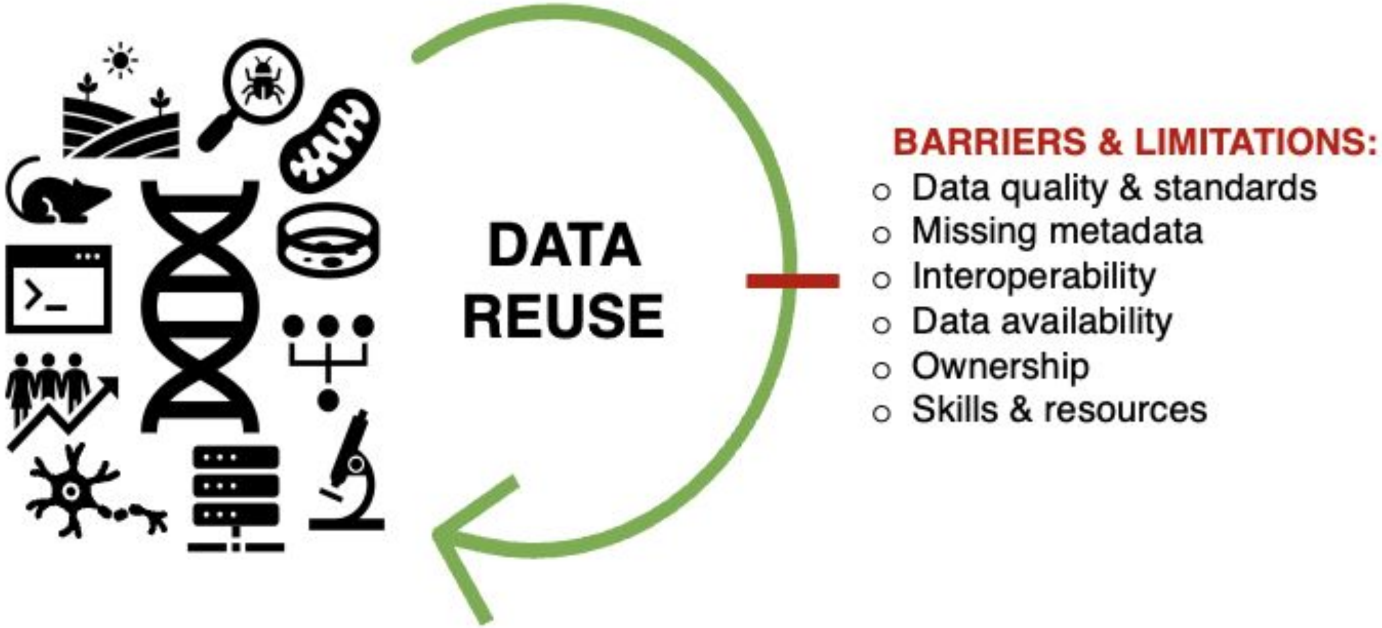
Take-aways and looking ahead

THE IMPORTANCE AND BENEFITS OF EQUITY AND INCLUSION

- ***Diversity breeds innovation***
- Reuse requires computational capacity, internet access, digital literacy, and proficiency in dominant languages
- Data sovereignty: https://localcontexts.org


+
BENEFITS OF BIG DATA IN AGRICULTURE
DATA REUSE
DIVERSITY, EQUITY, INCLUSION
+

TAKE-AWAYS AND LOOKING AHEAD



DATA REUSE

BARRIERS & LIMITATIONS:

- Data quality & standards
- Missing metadata
- Interoperability
- Data availability
- Ownership
- Skills & resources

The future of data reuse is bright and exciting!

- Integration of datasets
- Emerging data types:

  phenomes, metabolomes, proteomes, interactomes, enviromes, microbiomes, lipidomes, and glycomes
- AI and ML

WG’s white paper

James Koltes, Iowa State University (WG chair)

Alenka Hafner, Penn State University (WG co-chair)

Victoria DeLeo - Bowery Farming

Cecilia Deng- The New Zealand Institute for Plant and Food Research Limited

Christine G. Elsik - University of Missouri

Damarius Fleming, USDA

Peter W. Harrison, European Bioinformatics Institute

Ted Kalbfleisch, University of Kentucky

Bruna Petry, Iowa State University

Boas Pucker, TU Braunschweig

Elsa H Quezada-Rodríguez, Universidad Autónoma Metropolitana -Xochimilco; Universidad Nacional Autónoma de México

Christopher K. Tuggle, Iowa State University

https://doi.org/10.20944/preprints202401.0780.v1