CICompass

Funded by the National Science Foundation, Grant #2127548

Making the Major Facilities Data Lifecycle FAIR

To provide AI ready data

USC Viterbi School of Engineering Information Sciences Institute

INDIANA UNIVERSITY

UNIVERSITY OF NOTRE DAME

THE UNIVERSITY OF UTAH

TEXAS TECH UNIVERSITY

renci

Funded by the National Science Foundation, Grant #2127548

Agenda

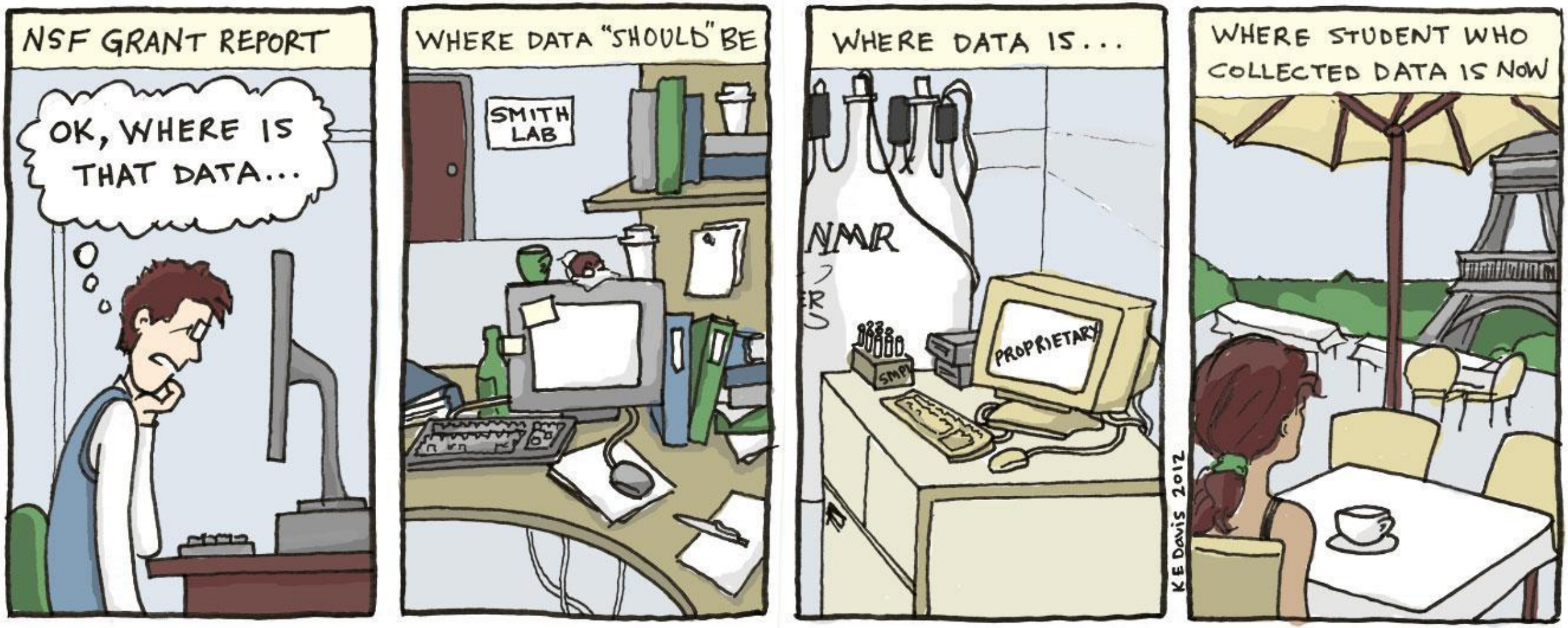
- Introduction (5 min) - Charles Vardeman
- FAIR Data and the MF Data Lifecycle (10 min) - Donald Brower
- Wikidata: A knowledge graph for the earth sciences (10 min) - Charles Vardeman
- ESIP Schema.org cluster: Utilizing the Schema.org vocabulary for FAIR Earth Science Data (15 min) - Adam Shepherd and Doug Fills
- NEON's first data releases: challenges and opportunities to provide FAIR data Draft Description (10 min) - Christine Laney
- Questions for first group of talks (5 min)
- AI strategies for improving data management components and preparing AI-ready data that meets FAIR and data provenance (15 min) - Giri Prakash
- Application of FAIR Principles in Astronomy (10 min) - Bruce Berriman

Funded by the National Science Foundation, Grant #2127548

Agenda

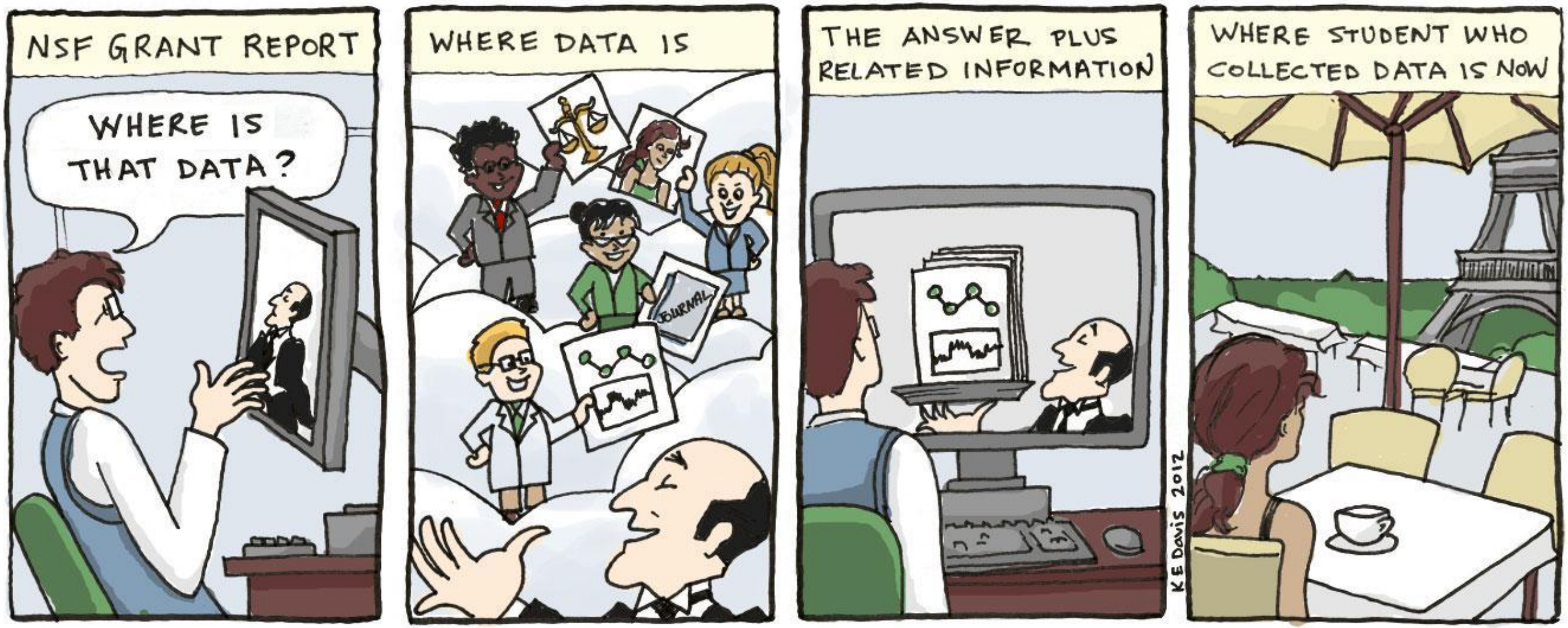
- Questions for second set of talks (5 min)
- Break - Close Poll at end of break
- Group Discussion over

Funded by the National Science Foundation, Grant #2127548



Credit: Kristina Furse and Charles Vardeman, University of Notre Dame, Center for Research Computing, NSF DASPOS: Data and Software Preservation for Open Science (2012)

Funded by the National Science Foundation, Grant #2127548



Credit: Kristina Furse and Charles Vardeman, University of Notre Dame, Center for Research Computing, NSF DASPOS: Data and Software Preservation for Open Science (2012)

Funded by the National Science Foundation, Grant #2127548

Wilkinson et al.
https://www.nature.com/articles/sdata201618

scientific data

Explore content | About the journal | Publish with us

nature > scientific data > comment > article

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, ... Barend Mons + Show authors

Scientific Data 3, Article number: 160018 (2016) | Cite this article

423k Accesses | 3576 Citations | 1934 Altmetric | Metrics

An Addendum to this article was published on 19 March 2019

Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Funded by the National Science Foundation, Grant #2127548

DeepMind › Blog › AlphaFold: a solution to a 50-year-old grand challenge in biology

BLOG POST
RESEARCH

30 NOV 2020

AlphaFold: a solution to a 50-year-old grand challenge in biology

Article | Open Access | Published: 22 July 2021

Highly accurate protein structure prediction for the human proteome

Kathryn Tunyasuvunakool, Jonas Adler, ... Demis Hassabis + Show authors

Nature **596**, 590–596 (2021) | Cite this article

184k Accesses | **200** Citations | **1413** Altmetric | Metrics

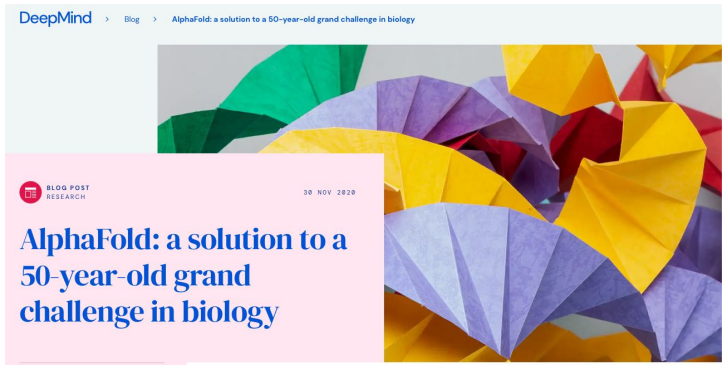
Abstract

Protein structures can provide invaluable information, both for reasoning about biological processes and for enabling interventions such as structure-based drug development or targeted mutagenesis. After decades of effort, 17% of the total residues in human protein sequences are covered by an experimentally determined structure[1]. Here we markedly expand the structural coverage of the proteome by applying the state-of-the-art machine learning method, AlphaFold[2], at a scale that covers almost the entire human proteome (98.5% of human proteins). The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence. We introduce several metrics developed by building on the AlphaFold model and use them to interpret the dataset, identifying strong multi-domain predictions as well as regions that are likely to be disordered. Finally, we provide some case studies to illustrate how high-quality predictions could be used to generate biological hypotheses. We are making our predictions freely available to the community and anticipate that routine large-scale and high-accuracy structure prediction will become an important tool that will allow new questions to be addressed from a structural perspective.

Funded by the National Science Foundation, Grant #2127548

Model confidence and added coverage

We predicted structures for the UniProt human reference proteome (one representative sequence per gene), with an upper length limit of 2,700 residues[6]. The final dataset covers 98.5% of human proteins with a full chain prediction.



The dataset adds high-quality structural models across a broad range of Gene Ontology (GO) terms[36,37], including pharmaceutically relevant classes such as enzymes and membrane proteins[38] (Fig. 1d). Membrane proteins, in particular, are generally underrepresented in the PDB because they have historically been challenging experimental targets. This shows that AlphaFold is able to produce confident predictions even for protein classes that are not abundant within its training set.

Funded by the National Science Foundation, Grant #2127548

The Question?

Where are we in terms of FAIR adoption in the MF community and where do we need/want to go with FAIR?

CICompass

Funded by the National Science Foundation, Grant #2127548

FAIR Data and the MF Data Lifecycle

Donald Brower

Funded by the National Science Foundation, Grant #2127548

FAIR

- Domain-independent, high-level principles that can be applied to a wide range of scholarly outputs
- Acronym for Findable, Accessible, Interoperable, Reusable.
- Proposed by Wilkinson et al. (2016)
- Not a standard or a specification
- May be applied in any combination and incrementally
- Recognizes the importance of "computational agents"

Funded by the National Science Foundation, Grant #2127548

F — Findable

- F1: (meta)data are assigned a globally unique and persistent identifier
- F2: data are described with rich metadata (as defined by R1)
- F3: metadata clearly and explicitly include the identifier of the data it describes
- F4: (meta)data are registered or indexed in a searchable source

Funded by the National Science Foundation, Grant #2127548

A — Accessible

- A1: (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1: the protocol is open, free, and universally implementable
  - A1.2: the protocol allows for an authentication and authorization procedure, where necessary
- A2: metadata are accessible, even when the data are no longer available

Funded by the National Science Foundation, Grant #2127548

I — Interoperable

- I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- I2: (meta)data use vocabularies that follow FAIR principles
- I3: (meta)data include qualified references to other (meta)data

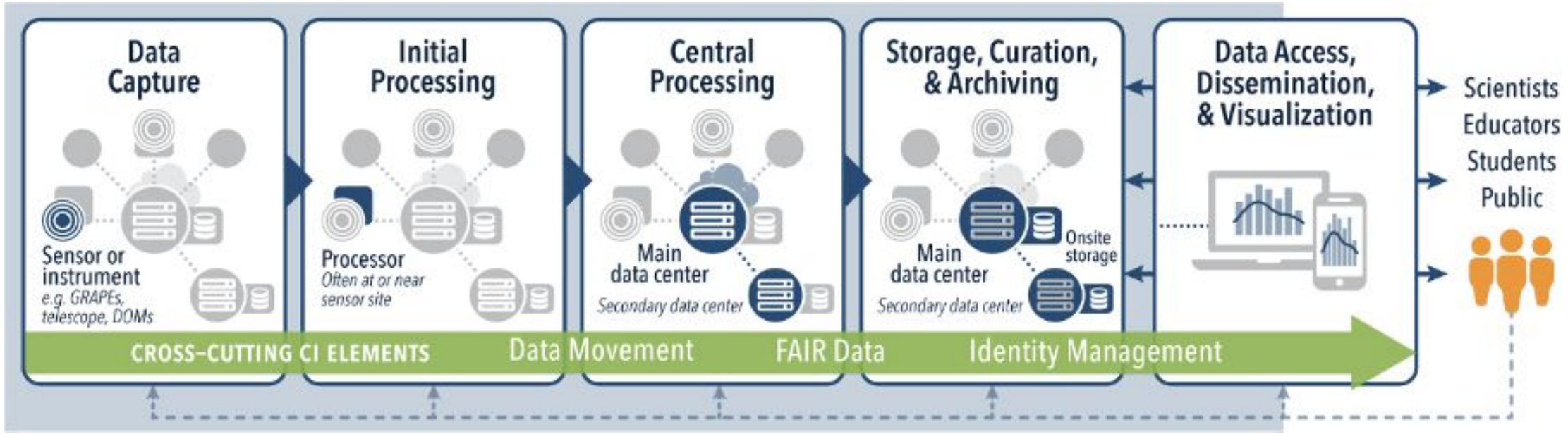
Funded by the National Science Foundation, Grant #2127548

R — Reusable

- R1: meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1: (meta)data are released with a clear and accessible data usage license
  - R1.2: (meta)data are associated with detailed provenance
  - R1.3: (meta)data meet domain-relevant community standards

Funded by the National Science Foundation, Grant #2127548

Data Lifecycle



Funded by the National Science Foundation, Grant #2127548

- Identifiers are essential at each step of lifecycle to identify and describe data
- Need ability to add metadata and provenance at each step
- Metadata/provenance helps with further reuse, eg. AI-ready data
- Some domain specific efforts to "FAIR"-ify data https://www.go-fair.org/

Funded by the National Science Foundation, Grant #2127548

- "good data management and stewardship is not a goal in itself, but rather a pre-condition supporting knowledge discovery and innovation" (Wilkinson 2016)
- Data management is a cross-cutting concern and not limited to the archiving and dissemination phases of the data lifecycle

Funded by the National Science Foundation, Grant #2127548

References

- Wilkinson, Mark D., et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Scientific Data 3, no. 1 (December 2016): 160018. https://doi.org/10.1038/sdata.2016.18.
- McMurry, Julie A, et al. “Identifiers for the 21st Century: How to Design, Provision, and Reuse Persistent Identifiers to Maximize Utility and Impact of Life Science Data”. https://doi.org/10.1101/117812.

Funded by the National Science Foundation, Grant #2127548

Extra: Persistent Identifiers

- URI schemes. Difference from URLs.
- Opaque identifiers
- DOI, ORCID, ARK

CICompass

Funded by the National Science Foundation, Grant #2127548

Wikidata: A knowledge graph for the earth sciences

Charles F Vardeman II

CICompass

Funded by the National Science Foundation, Grant #2127548

ESIP Schema.org cluster: Utilizing the Schema.org vocabulary for FAIR Earth Science Data

Adam Shepherd and Doug Fills

Funded by the National Science Foundation, Grant #2127548

[Link to slides]

CICompass

Funded by the National Science Foundation, Grant #2127548

NEON's first data releases: challenges and opportunities to provide FAIR data

Christine Laney

Funded by the National Science Foundation, Grant #2127548

AI strategies for improving data management components and preparing AI-ready data that meets FAIR and data provenance

Giri Prakash

CICompass

Funded by the National Science Foundation, Grant #2127548

Application of FAIR Principles in Astronomy

Bruce Berriman

Funded by the National Science Foundation, Grant #2127548

Discussion…

Funded by the National Science Foundation, Grant #2127548

Explain

1. Which FAIR principle are you farther along with regarding implementation, and how would you evaluate your progress towards being FAIR? (check all that apply)
    a. Findable
    b. Accessible
    c. Interoperable
    d. Reusable
    e. Very good progress
    f. Some progress
    g. A little progress
    h. No progress