Enabling Open Science and Data Sharing: Trust, Provenance, and Data Integrity

Michael Corn
Cybersecurity Advisor for Research Infrastructure
CORF
micorn@nsf.gov

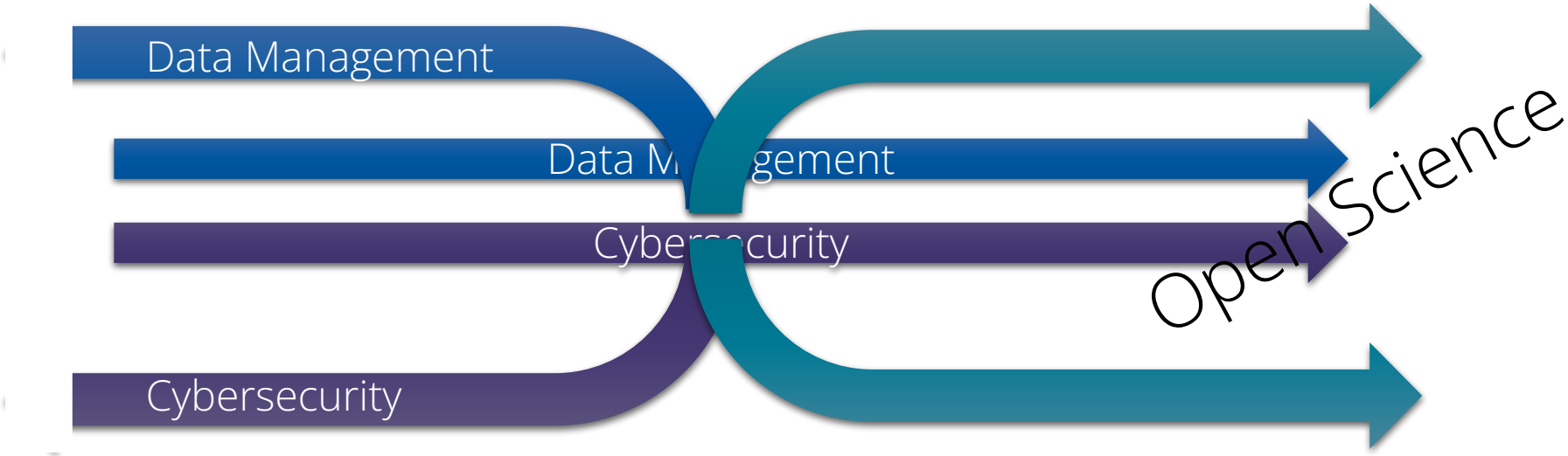
Abstract and Agenda

Open Science is driven by knowledge discovery and innovation and fueled by the wide dissemination of scholarly publications and data. Information Assurance (inclusive of cybersecurity, data protections (including privacy), cyber risk management, and resilience) provides tools that support the practical implementation of FAIR principles. This talk explores how FAIR, Information Assurance, and Research Security are related and why each domain needs to better recognize their shared concerns.

- Why are we here?
- Partnership not Ownership
- FAIR and Cybersecurity
- Forthcoming Research Infrastructure Guide Information Assurance Supplement
- Discussion

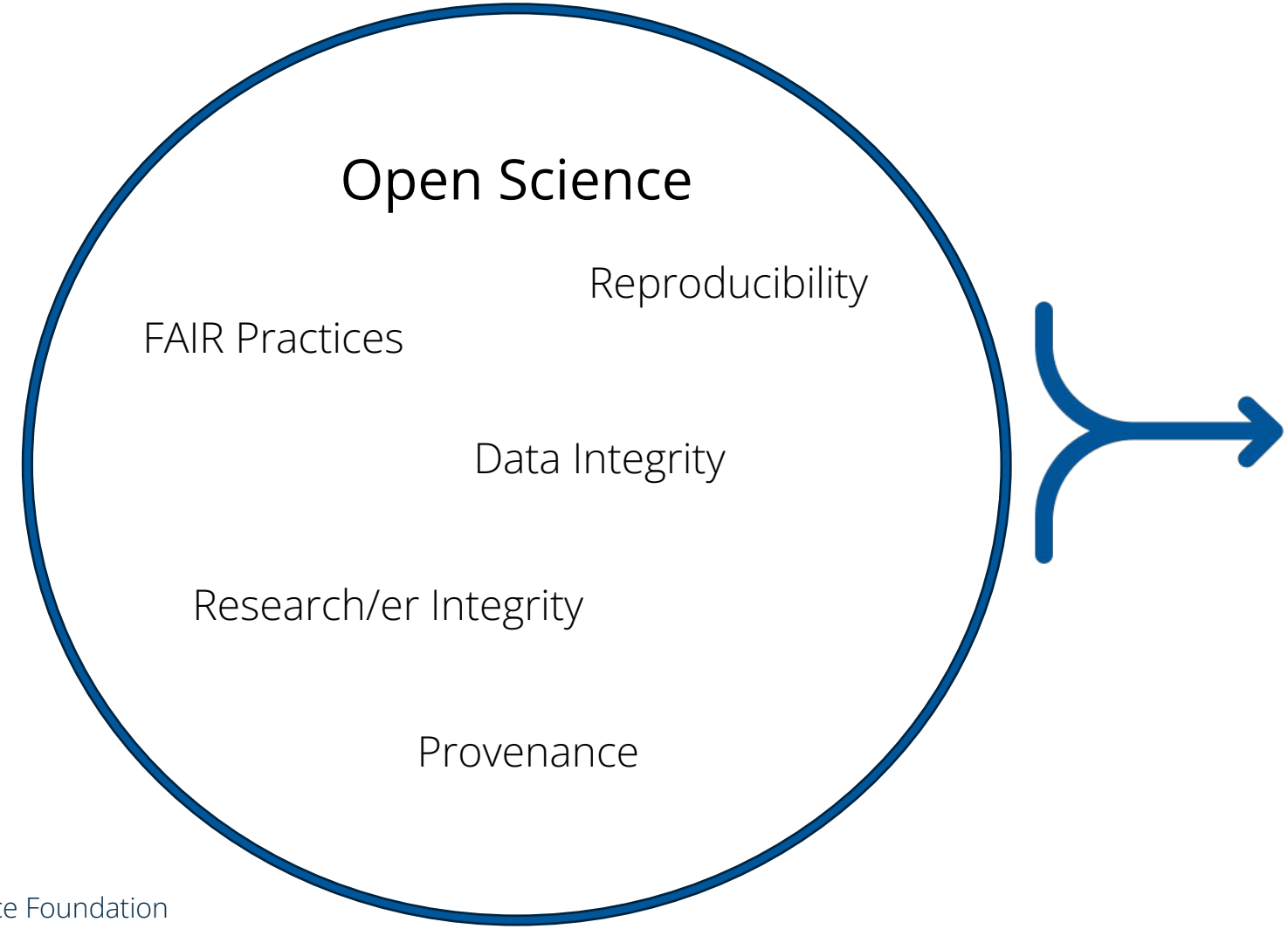
Cybersecurity and Data Management

Why am I here?



Not about Ownership but Partnership

Not about Ownership but Partnership



- Accelerated discovery
- Better science
- Trust in science
- National health & competitiveness

Same Concerns but Different Perspectives

	Cybersecurity	Data Management
Data Integrity	• Malicious Modification	• Validation and Reproducibility
Access Privileges	• Least Access	• Maximum Access
Time Horizons	• Short / immediate	• Long time horizons (metadata post-data)
Scope	• Metadata = Data	• Metadata (powers FAIR)
Provenance	• Forensic analysis	• Replicability
Research Integrity	• Insider threat	• FAIR

FAIR and Cybersecurity

Findable

Accessible

Interoperable

Reusable

Powered by metadata

Corruption / modification of metadata =>

- Data is no longer findable
- Metadata inconsistently or no longer accessible
- Interoperability fails
- Reuse becomes meaningless

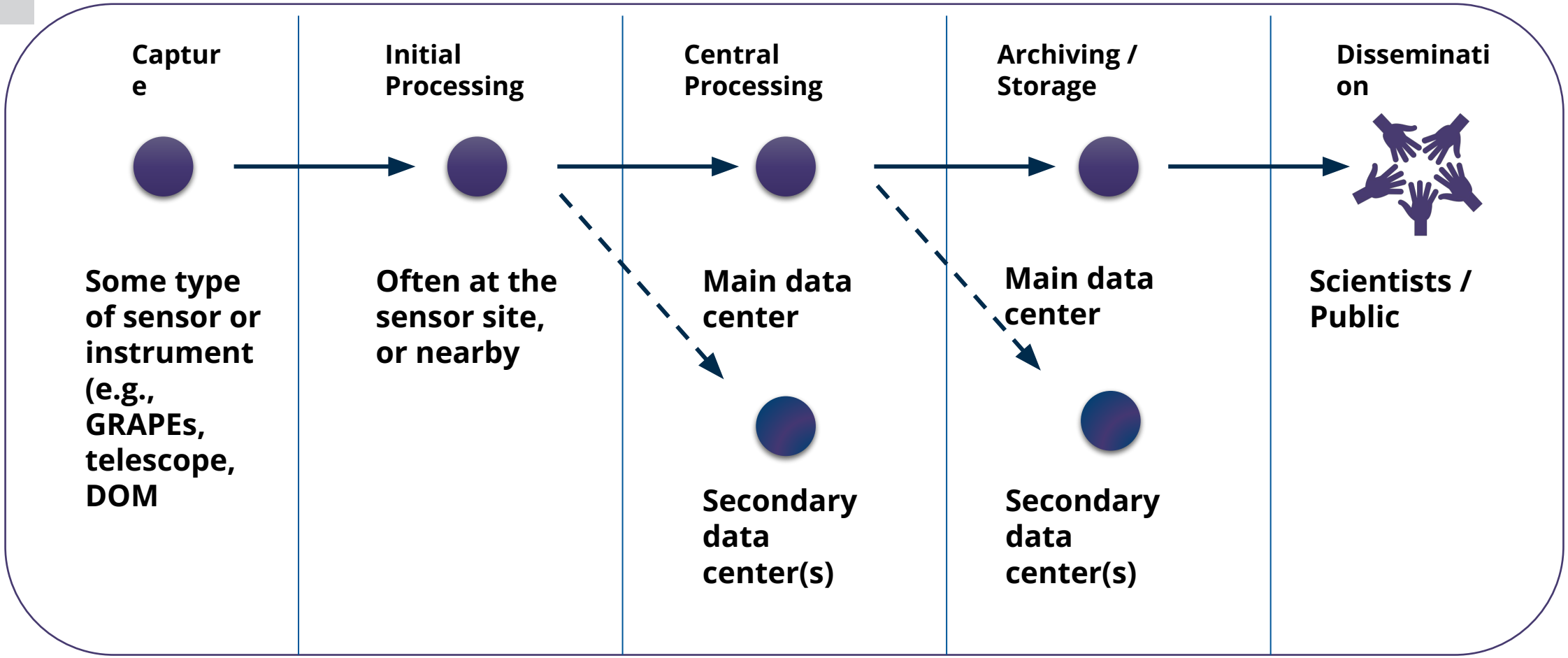


Diagram courtesy of CI Compass

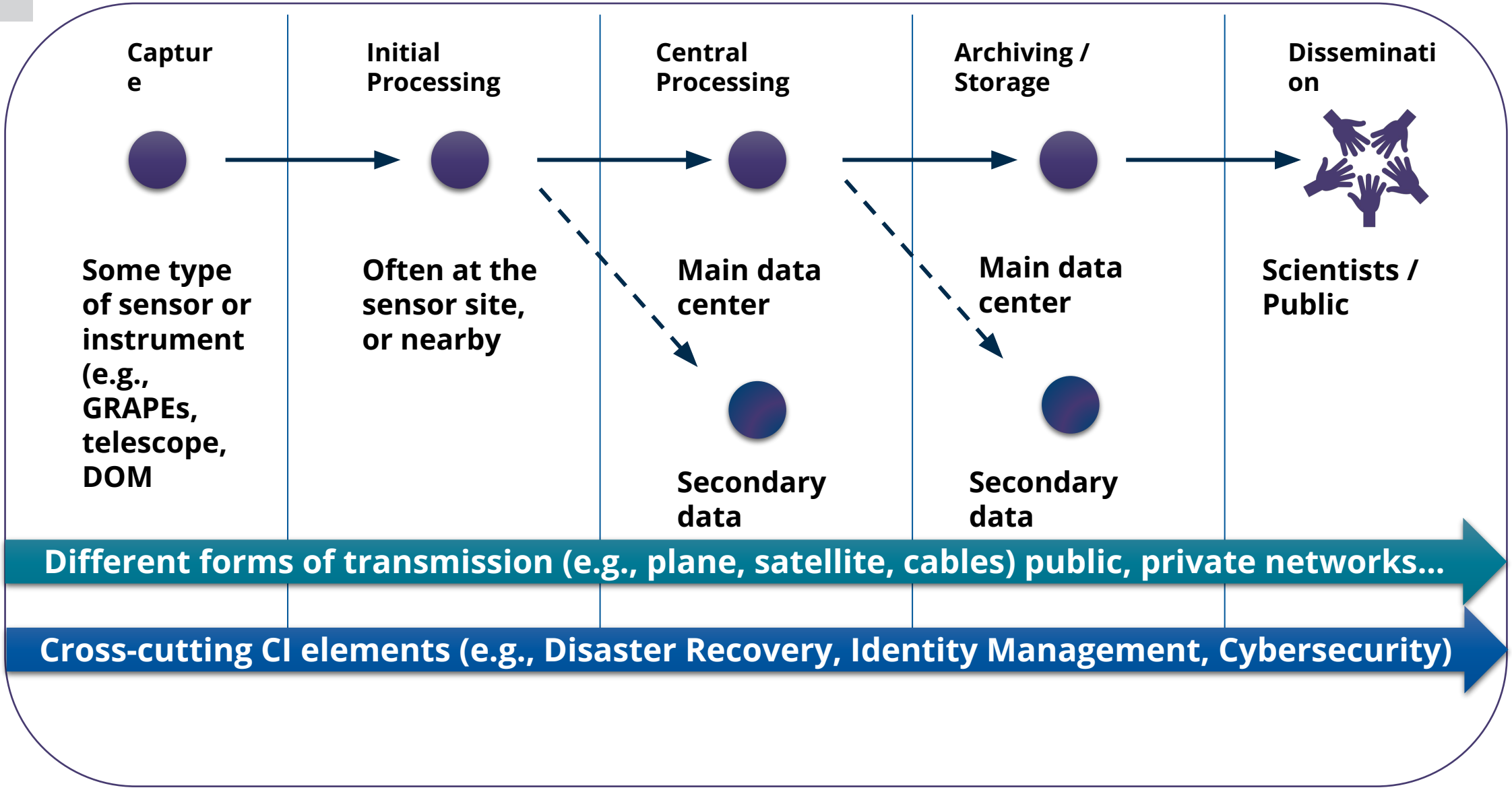
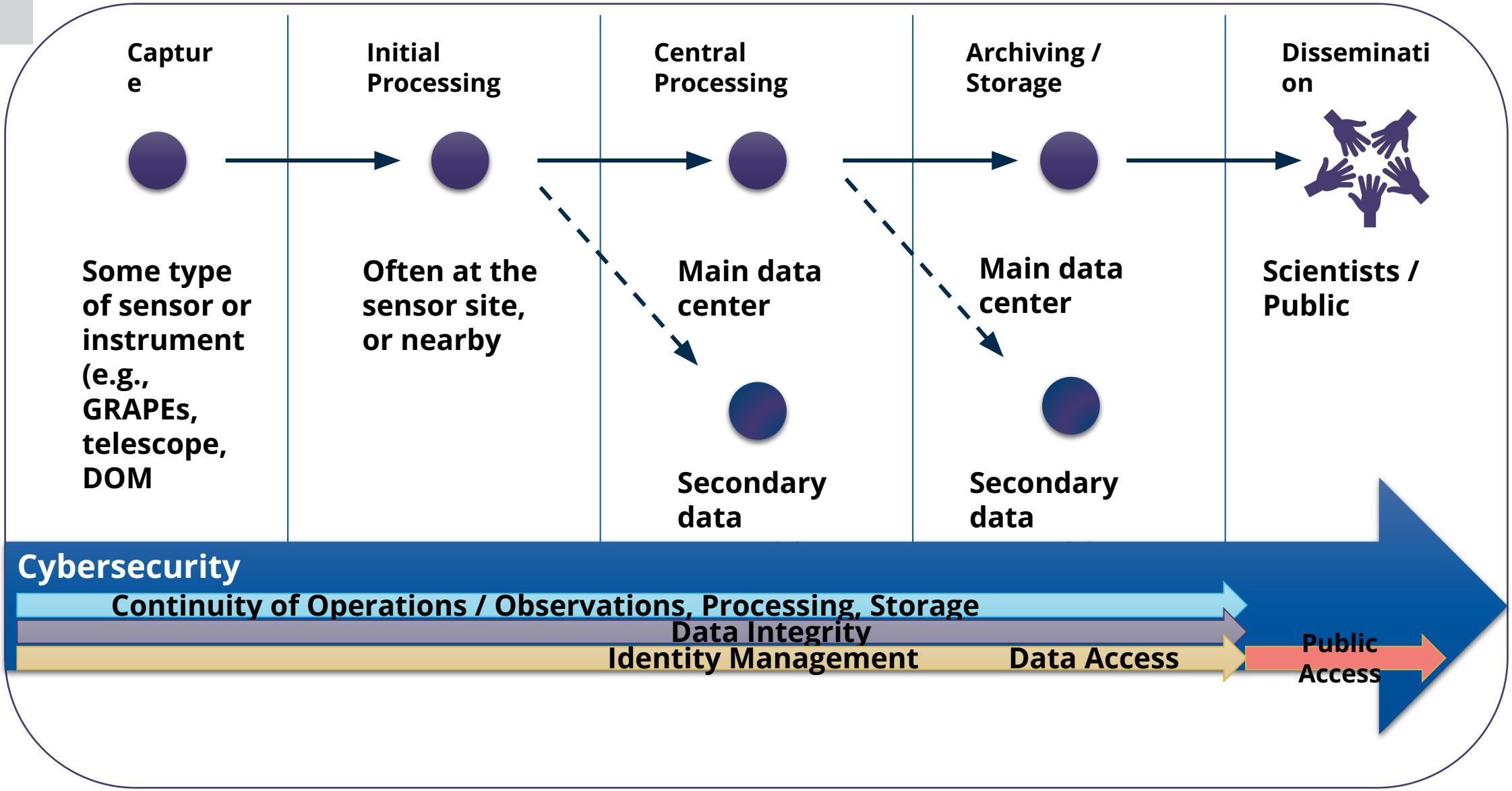
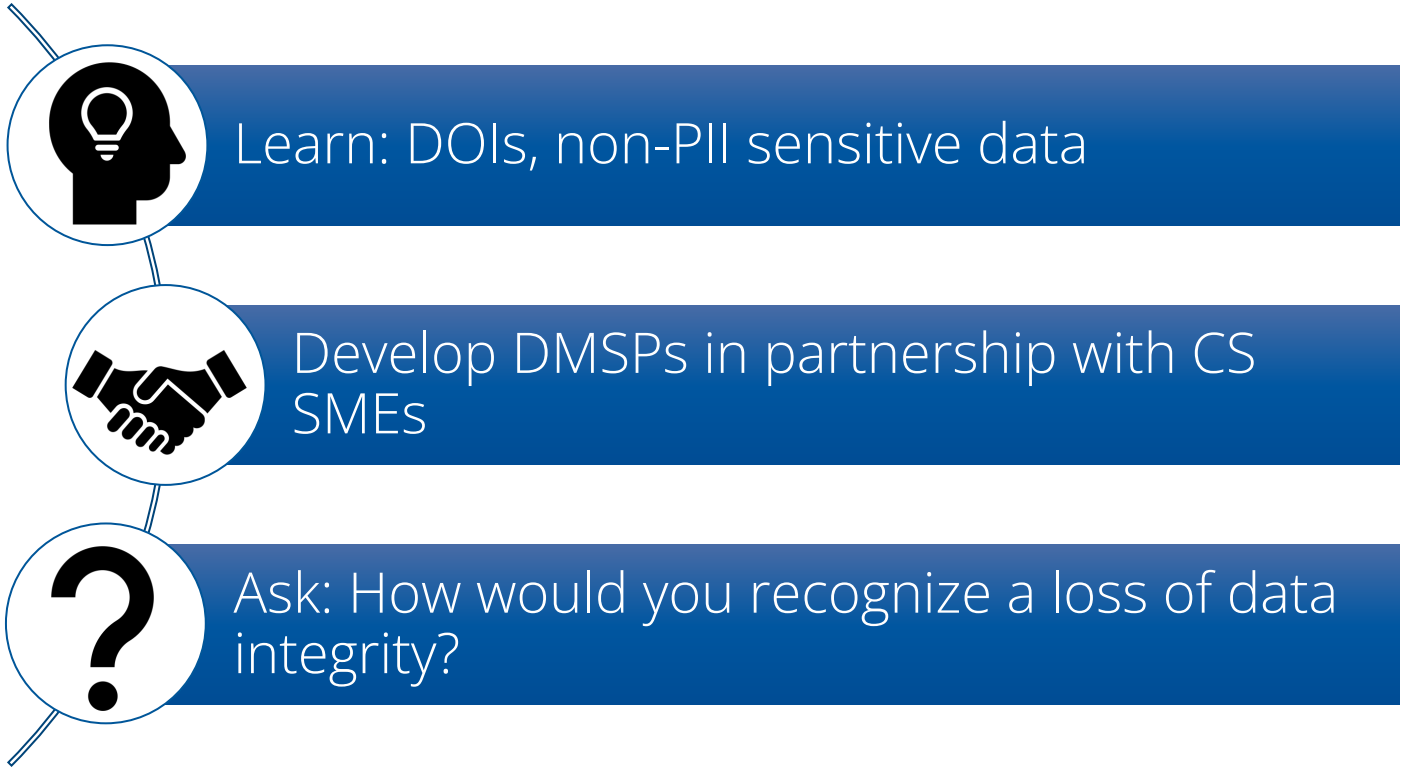


Diagram courtesy of CI Compass


Captur
e
Initial
Processing
Central
Processing
Archiving /
Storage
Disseminati
on
Some type
of sensor or
instrument
(e.g.,
GRAPEs,
telescope,
DOM
Often at the
sensor site,
or nearby
Main data
center
Secondary
data
Main data
center
Secondary
data
Scientists /
Public
Cybersecurity
Continuity of Operations / Observations, Processing, Storage
Data Integrity
Identity Management
Data Access
Public
Access

How do we get there?



- Provenance (DOIs for instrumentation, algorithms, software, storage, data artifacts)
- Protection of sensitive information (PII, endangered species, geolocation data)

- Develop DMSP in partnership with cybersecurity SMEs
- Ask cybersecurity SMEs to review entire data workflow

- How would you know if you've been hacked?
- How would you know if your data's been changed? i.e., how would you recognize a loss of data integrity?

Data Management and the RIG

The Research Infrastructure Guide aka RIG, comprises NSF's guidance on meeting the letter of and spirit of requirements in cooperative agreements for major facilities and midscale research infrastructure.

The draft under revision significantly expands on previous versions regarding Information Assurance. This includes a strong recommendation to partner with cybersecurity SMEs in the review of data workflow and pipeline development and management.

§ Data Management and Curation

- Address cybersecurity implications in DMSP
- Identify and engage institutional data management expertise
- Monitor the 2022 OSTP Public Access Memorandum (and address budgetary impact)
- Review cybersecurity of interfaces to data repositories
- Identify data sets with extraordinary data integrity requirements
- Discuss "hidden" sensitive data even w/in non-human data sets (e.g., geolocation for endangered species)

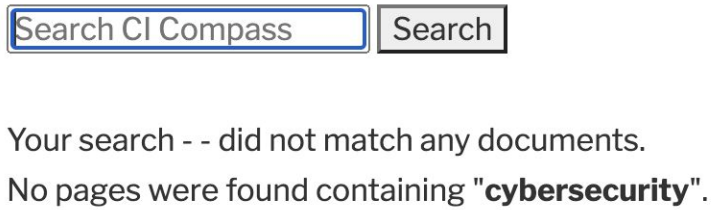
Facility Resilience

With the sophistication of modern, state sponsored or facilitated attacks, breaches of accounts and systems are inevitable. Resilience for a facility consists of,

- **Minimize the likelihood of successful attacks in general, and unsophisticated, opportunistic attacks specifically.**
- **Minimize the impact of even sophisticated attacks by constraining the 'blast radius' or ability to spread throughout a facility.**
- **Minimize the period of the disruption of scientific operations.**
- **Ensure the integrity of scientific data and artifacts despite the occurrence of a cybersecurity incident.**

Perfect cybersecurity is impossible: Resilient facilities is our goal

Resources



- Reproducibility and Replicability in Science https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science
- TDWG Survey Report https://doi.org/10.5281/zenodo.3906865
- https://www.trustedci.org/search?q=data
- Recommendations for Improving the Security of a Science Gateway https://scholarworks.iu.edu/dspace/handle/2022/26780
- Securing Science Gateways https://opensky.ucar.edu/islandora/object/articles%3A19076
- Custos https://dl.acm.org/doi/10.1145/3311790.3396635
- FAIRS (FAIR + Security) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9790701/

Slide Bank