Challenges in managing data in a multi-disciplinary research organization

Dr. Mark Greiner

Aviral Vaid

Research Data Management: Vision

Users **maintain their existing workflows**, while their **generated data is automatically digitized and categorized** for them, and is, subsequently, **available and easy to find** at a future time.

Research Data Management (RDM)

Experimental Workflow

- Sample lifecycle
- Lab book
- Data analysis tools

Laboratory Management

- Inventory
- Instruments
- Job scheduling, submitting

Data Lake

- Working storage
- Long-term backup and archiving

Project Management

- Proposals
- Research phases
- Presentations, Publications...

Integrations and Automated Workflows to Interface between Modules

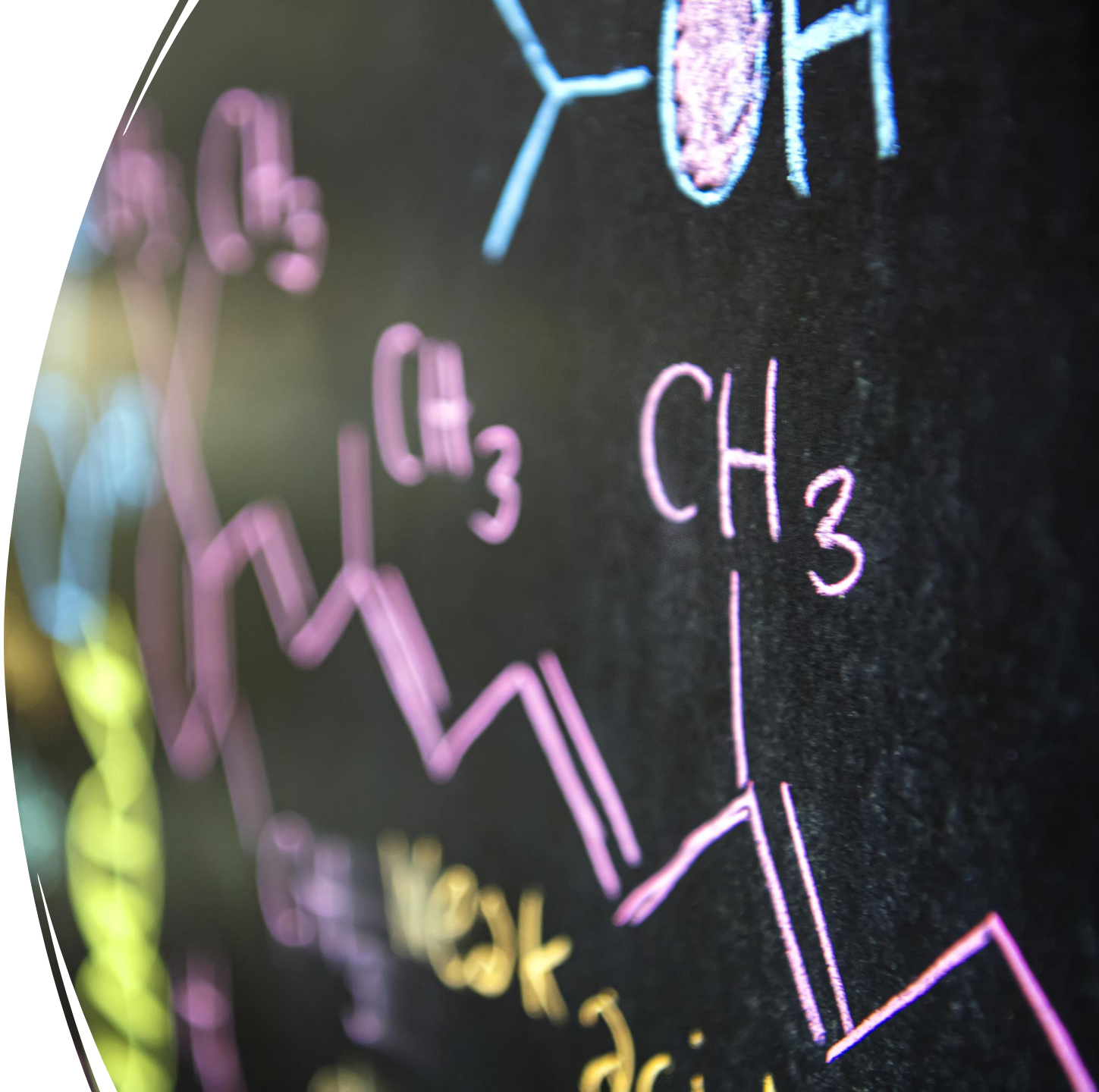
About MPI-CEC

Research Discipline

- Catalysis for chemical energy
- Water splitting, bio-catalysts, chemical production

Magnitude

- 218 Researchers
- 22 Research Groups
- 3 Departments



Diverse requirements

Synthesis Lab

- Focus on:
  - Creating new chemicals
- Workflow:
  - Plan, Synthesize, Characterize, Analyze results, Iterate, Test in some application
- Challenges:
  - Multi-disciplinary
  - Harmonize ELN with diverse data sources

Diverse requirements

Specialized Characterization

- Focus on:
  - in-depth analysis
- Workflow:
  - Job request, Plan with user, Sample management, Scheduling, Data analysis, Data to user
- Challenges:
  - Data analysis
  - Harmonize with user ELN

Diverse requirements

Self-service facilities

- Focus on:
  - routine measurements
- Can be performed with minimal training
- Workflow:
  - Schedule, Measure, Retrieve data, Analyze data,
- Challenge:
  - Harmonize users' ELN with instrument
  - Associate data with sample

Diverse requirements

Testing facilities

- Focus on:
  - Behavior in applications
  - Testing parameters
- Workflow:
  - Plan, Schedule, Measure, Retrieve data, Analyze data
- Challenges:
  - Analysis
  - Linking data-sample-conditions

Diverse requirements

Large facilities

- Focus on:
  - Characterizing
- Workflow:
  - Plan, Schedule, Measure, Retrieve data, Analyze data
- Major challenges:
  - Data sizes
  - Integrating with home ELN

Workflow depends on Research

Examples:

- Synthesis centered
  - Main challenge is sample management
  - Which data belongs to which sample, after which treatment
- Analysis centered
  - Main challenges are data conversion, analysis versioning

Problem Statements

- Data is spread out across different machines and unstructured. Thus, knowledge cannot be automatically extracted from it, and tools cannot be (economically) built to automate regular processes.

Problem Statements

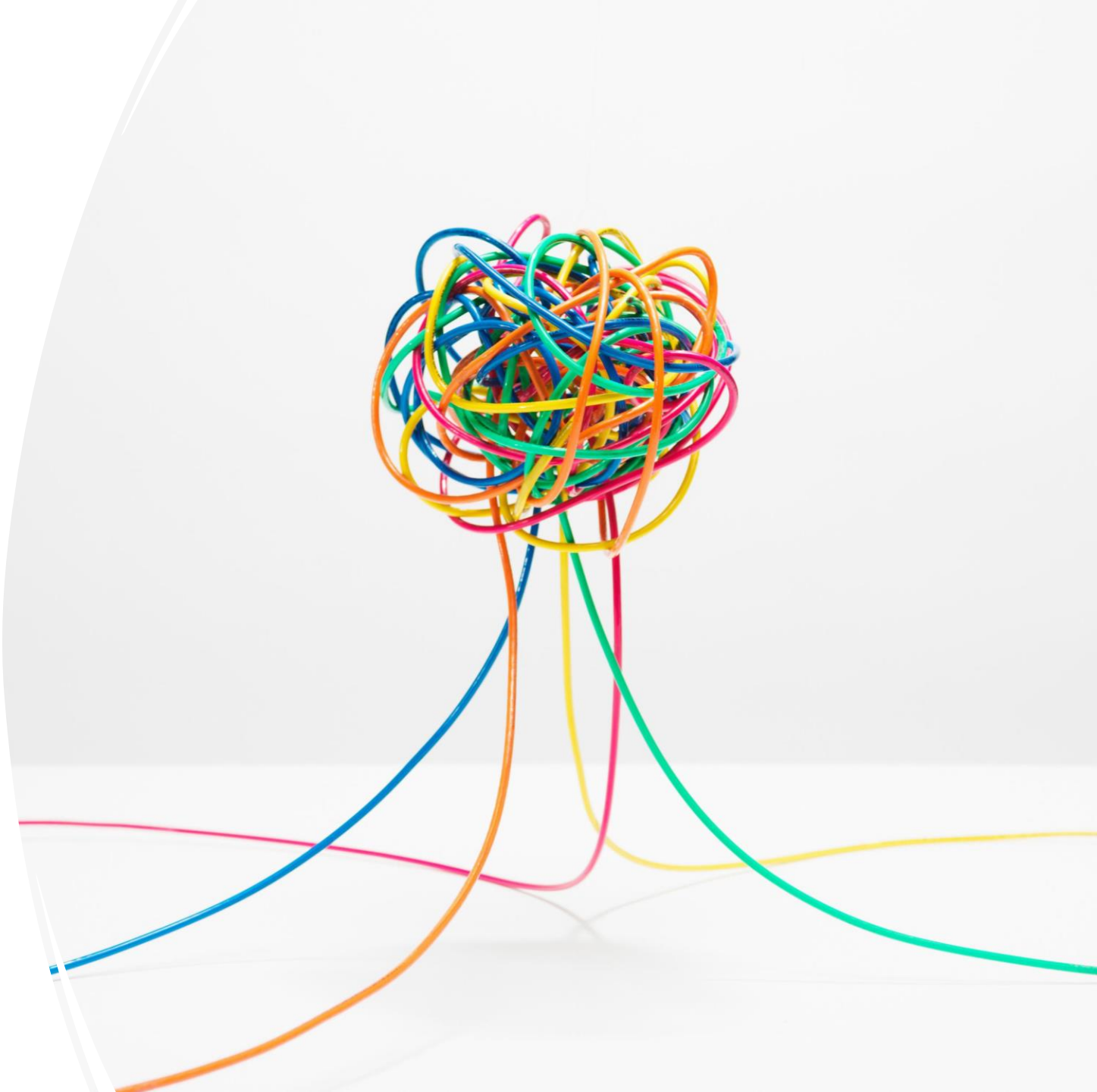
- Researchers spend too much time on repetitive manual work, related to generating data, processing data and publishing data.
    - Solving this would: save time, reduce errors, make better use of resources by minimizing re-work

Problem Statements

- Users to not have easy access to all their data assets.
  - Solving this would: Anable users to access their data from anywhere, anytime (improved accessibility)

Problem Statement

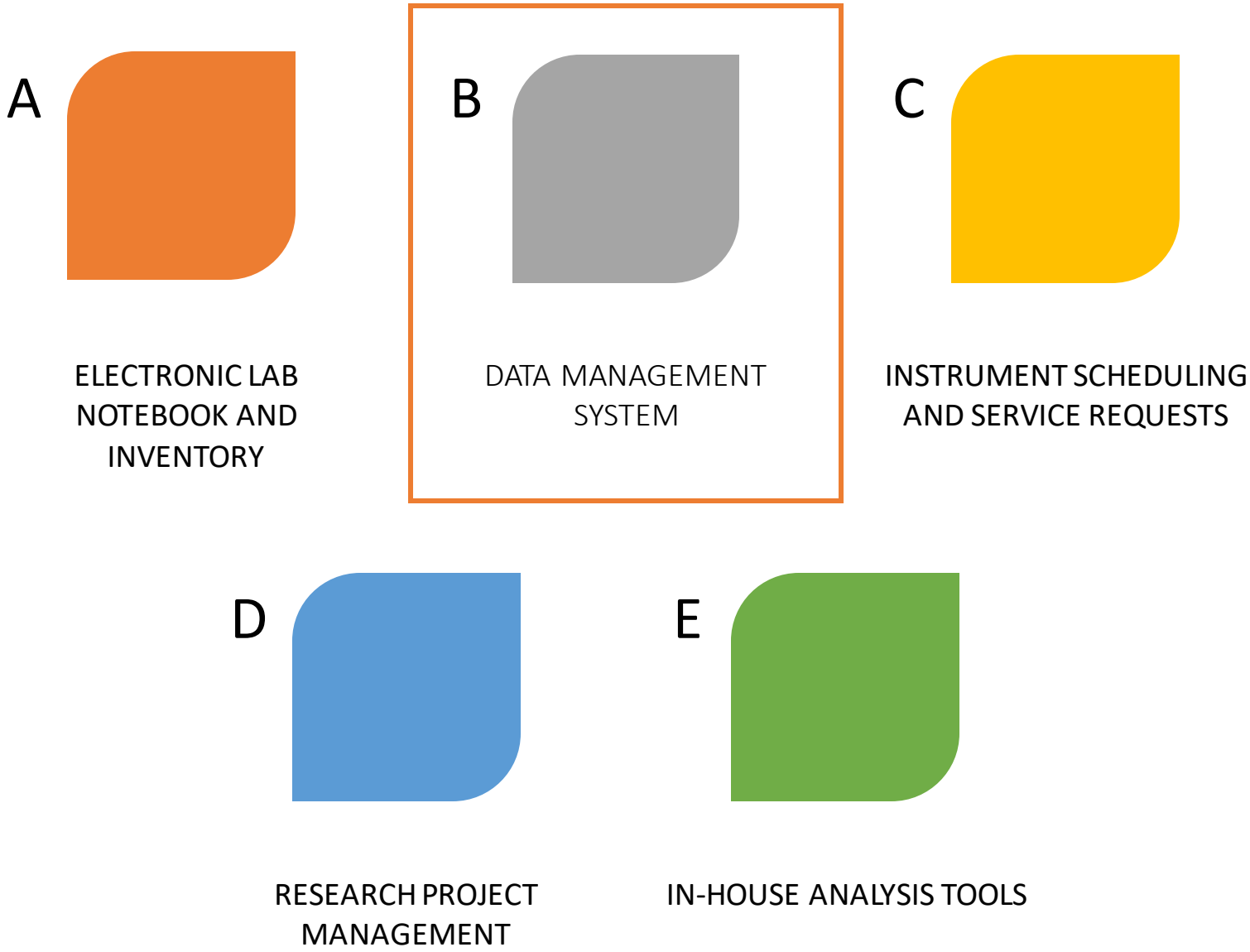
- It is difficult, sometimes impossible, to trace back the origin of a research result.
  - Solving this would: Allow a user and the organization complete traceability and reproducibility over all data artifacts (improving management, scientific audits)



Problem Statement

- The structure of the organization's data assets are not suitable for large-scale analysis algorithms.
  - Solving this would: Future-proof an organization's assets, and open up new research opportunities (e.g. artificial intelligence)

Phases of Research Data Management



Thank You!