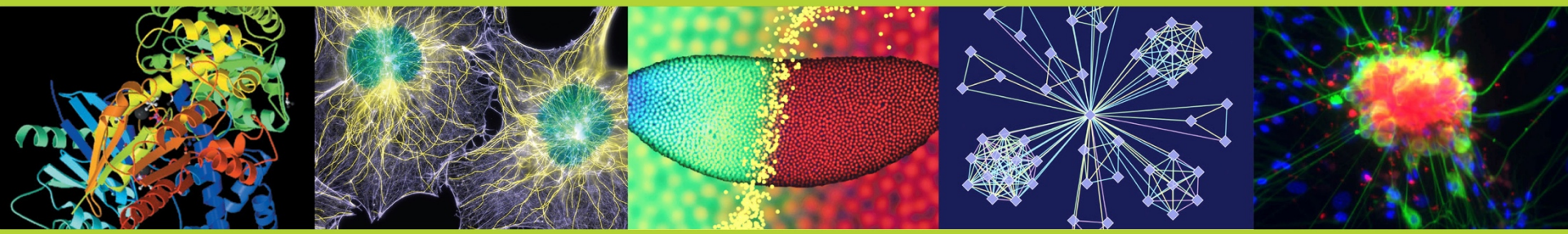
National Institute of General Medical Sciences

The NIH Strategic Plan for Data Science

Jon R. Lorsch, Ph.D., Director
National Institute of General Medical Sciences



Developing an NIH Strategic Plan for Data Science

- Requested by Congress
- The plan focuses on:
  - Modernizing the data resource ecosystem to increase its utility **for researchers and other stakeholders** and to optimize its efficiency of operation
  - Enhancing data sharing, access, interoperability
  - Improving ability to use EHR, clinical, and observational data for research while ensuring data confidentiality
  - Modernizing infrastructure, increasing capacity

A Couple of Definitions…

"Data science is an interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data."

FINDABLE

ACCESSIBLE

INTEROPERABLE

REUSABLE

Domains of Data Science	Description
Data Infrastructure	*Hardware, architecture, and platforms necessary to capture, organize, store, allow access to, and compute on data*
Data Resources	*Methods, practices, and associated features needed to increase the value and utility of data beyond its native state*
Advanced Management, Analytics, and Visualization Tools	*Algorithms, software, models, and tools necessary to extract knowledge and understanding from data*
Workforce Development	*Policies, practices, and programs to train and develop an outstanding data science workforce*
Policy, Stewardship, and Sustainability	*The policies and practices necessary for governance, financial management, and sustainable stewardship of the biomedical data science ecosystem*

Organization of the Strategic Plan

I. Overarching Goals

   i. Strategic Objectives

      1. Implementation Tactics

         a. Milestones and Performance Measures

Overarching Goal 1: Support Highly Efficient and Effective *Data Infrastructure for Biomedical Research*

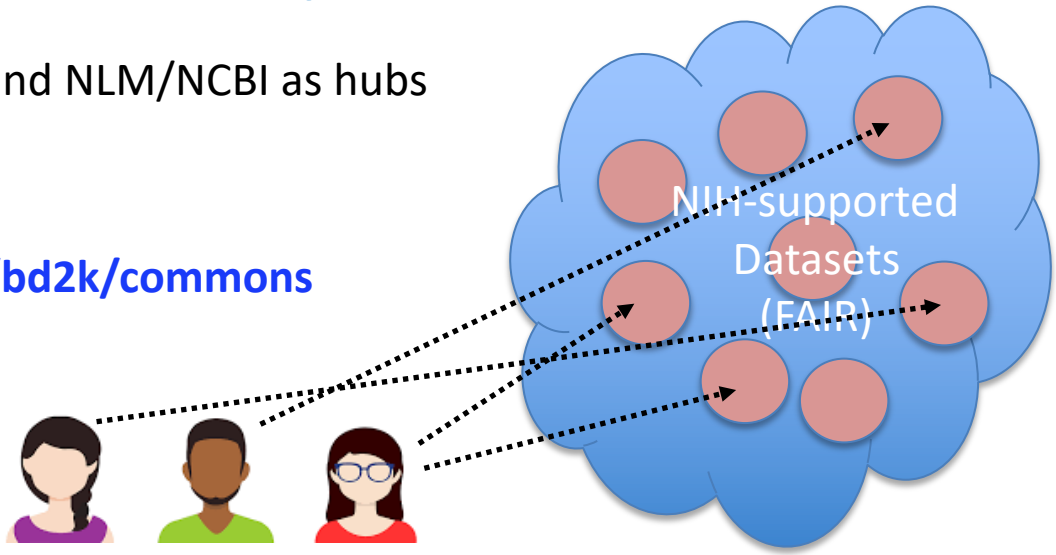
Strategic Objective 1-1: Optimize Data Storage, Access and Security

- Rely on private sector where possible

Strategic Objective 1-2: Connect NIH Data Systems

- Use NIH Data Commons and NLM/NCBI as hubs

https://commonfund.nih.gov/bd2k/commons



NIH Data Commons

Overarching Goal 2: Promote the Modernization of the *Data Resources* Ecosystem

Strategic Objective 2-1: Modernize the Data Repository Ecosystem

Implementation Tactics:

- Separate the support of databases and knowledgebases

Databases and Knowledgebases

- ***Databases***: data repositories that store, organize, validate, and make accessible the core data related to a particular system or systems
  - For example, core data might include genome, transcriptome, and protein sequences
  - Curation mostly focuses on QA/QC

- ***Knowledgebases****:* accumulate, organize, and link growing bodies of information related to core datasets
  - For example, information about expression patterns, splicing variants, localization, protein-protein interaction and pathway networks related to an organism or set of organisms; publication information
  - Traditionally require significant levels of human curation

Overarching Goal 2: Promote the Modernization of the *Data Resources* Ecosystem

Strategic Objective 2-1: Modernize the Data Repository Ecosystem

Implementation Tactics:

- Separate the support of databases and knowledgebases
- Use appropriate mechanism, review, and management for each type of repository
- Dynamically measure data use, utility, and modification
- Ensure privacy and security
- Create unified, efficient, and secure authorization of access to sensitive data
- Employ explicit evaluation, lifecycle, sustainability, and sunsetting expectations for data resources

Overarching Goal 2: Promote the Modernization of the *Data Resources* Ecosystem

Strategic Objective 2-2: Support the Storage and Sharing of Individual Datasets

Implementation Tactics:

- Link datasets to publications via PubMed Central and NLM/NCBI
- Longer-term: Expand NIH Data Commons to allow submission, open sharing, and indexing of individual, FAIR datasets

Overarching Goal 2: Promote the Modernization of the *Data Resources* Ecosystem

Strategic Objective 2-3: Leverage Ongoing Initiatives to Better Integrate Clinical and Observational Data into Biomedical Data Science

Implementation Tactics:

- Create efficient linkages among NIH data resources that contain clinical and observational information.
- Develop and implement universal credentialing protocols and user-authorization systems to enforce a broad range of access and patient-consent policies across NIH data resources and platforms.
- Promote use of the NIH Common Data Elements Repository.

All of Us RESEARCH PROGRAM

emerge network
ELECTRONIC MEDICAL RECORDS AND GENOMICS

CANCER MOONSHOT

ECHO
Environmental influences on Child Health Outcomes
A program supported by the NIH

TOPMed

Overarching Goal 3: Support the Development and Dissemination of *Advanced Data Management, Analytics, and Visualization Tools*

Strategic Objective 3-1: Support Useful, Generalizable, and Accessible Tools and Workflows

Implementation Tactics:

- Separate support for tools from support for databases and knowledgebases
- Use appropriate mechanism, review, and management for tool development
- Leverage commercial tools, software, workflows, and expertise
- Promote development of open source, openly shared and reusable tools, software, and workflows

Strategic Objective 3-2: Broaden Use of Specialized Tools

- Example: Algorithms from astronomy adapted for use in cellular imaging
- Support research for improving methods for using EHRs and other clinical data

Strategic Objective 3-3: Improve Discovery and Cataloging Resources

- Development and adoption of community standards for data indexing, citation and provenance will be key

Why Separate Support of Databases, Knowledgebases and Tool Development?

Strategic Objective 2-1: Modernize the Data Repository Ecosystem

Implementation Tactics:

- Separate the support of databases and knowledgebases
- Use appropriate mechanism, review, and management for each type of repository

Strategic Objective 3-1: Support Useful, Generalizable, and Accessible Tools and Workflows

Implementation Tactics:

- Separate support for tools from support for databases and knowledgebases
- Use appropriate mechanism, review, and management for tool development

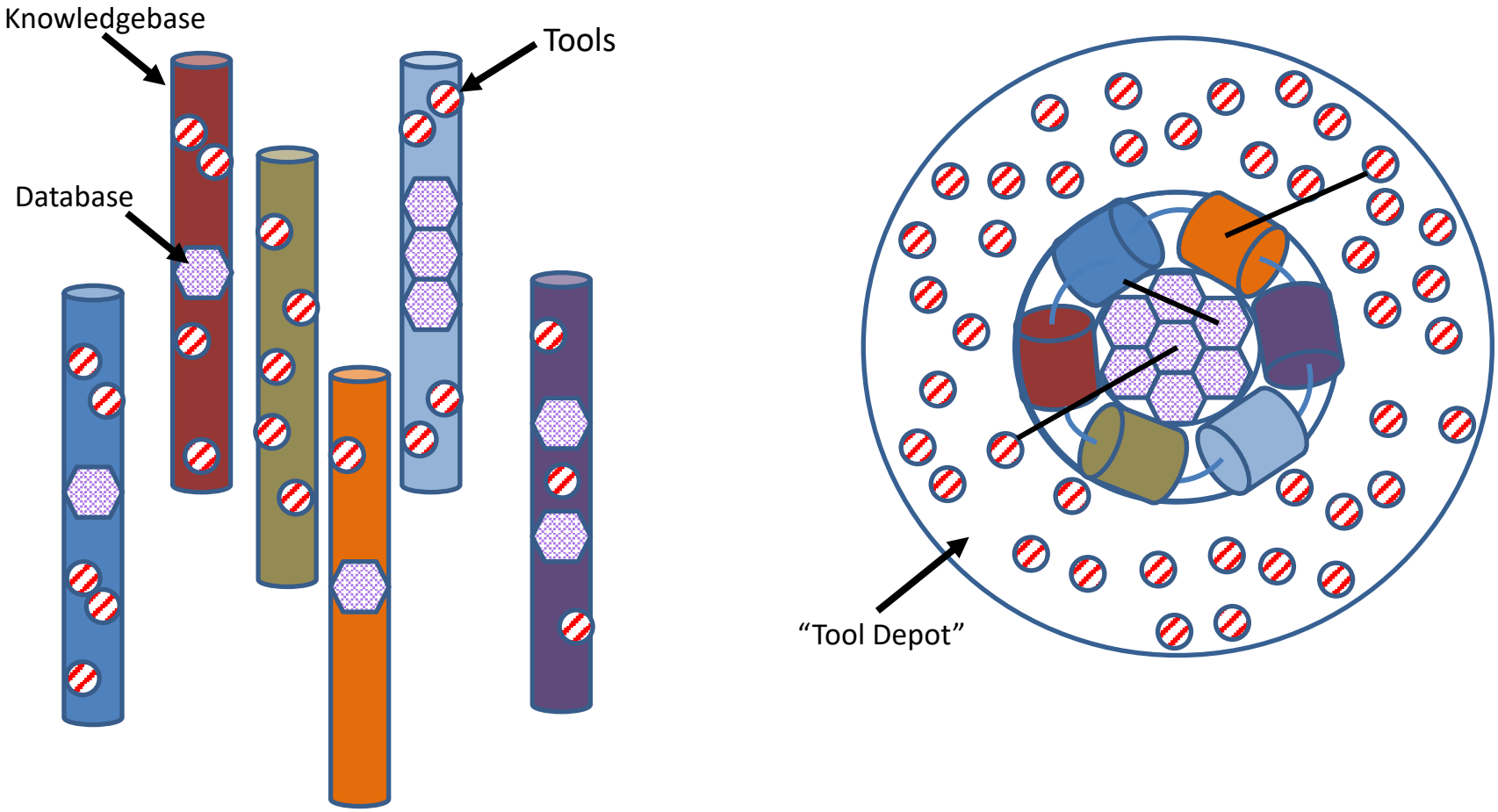
Why Separate Support of Databases, Knowledgebases and Tool Development?

- **NIH-funded data *resources* historically have been evaluated and funded as *research* grants**
  - Misalignment of goals and review expectations: innovation, hypothesis generation/testing, etc. rather than user service, utility, efficiency of operation, usage
  - Has led to entanglement of tool development and resource management
  - Tool development and dissemination should be evaluated on their own merits rather than linked to the existence of important data resources
- **Database and knowledgebase functions, needs and uses are not the same**
  - Core data may be essential to the research community but (in some cases) the associated knowledgebase information less so
  - Cost of human curation is high and its utility and impact must be carefully evaluated, independent of the need for access to core data
- **Need for usage, utility, impact and efficiency metrics for each kind of resource and for tools**

Why Separate Support of Databases, Knowledgebases and Tool Development?

The same group could perform all three functions, but each function must be evaluated and funded separately to ensure that the research community is getting the best value for the taxpayers' money

Creating a More Modern and Coherent Data Resource Ecosystem



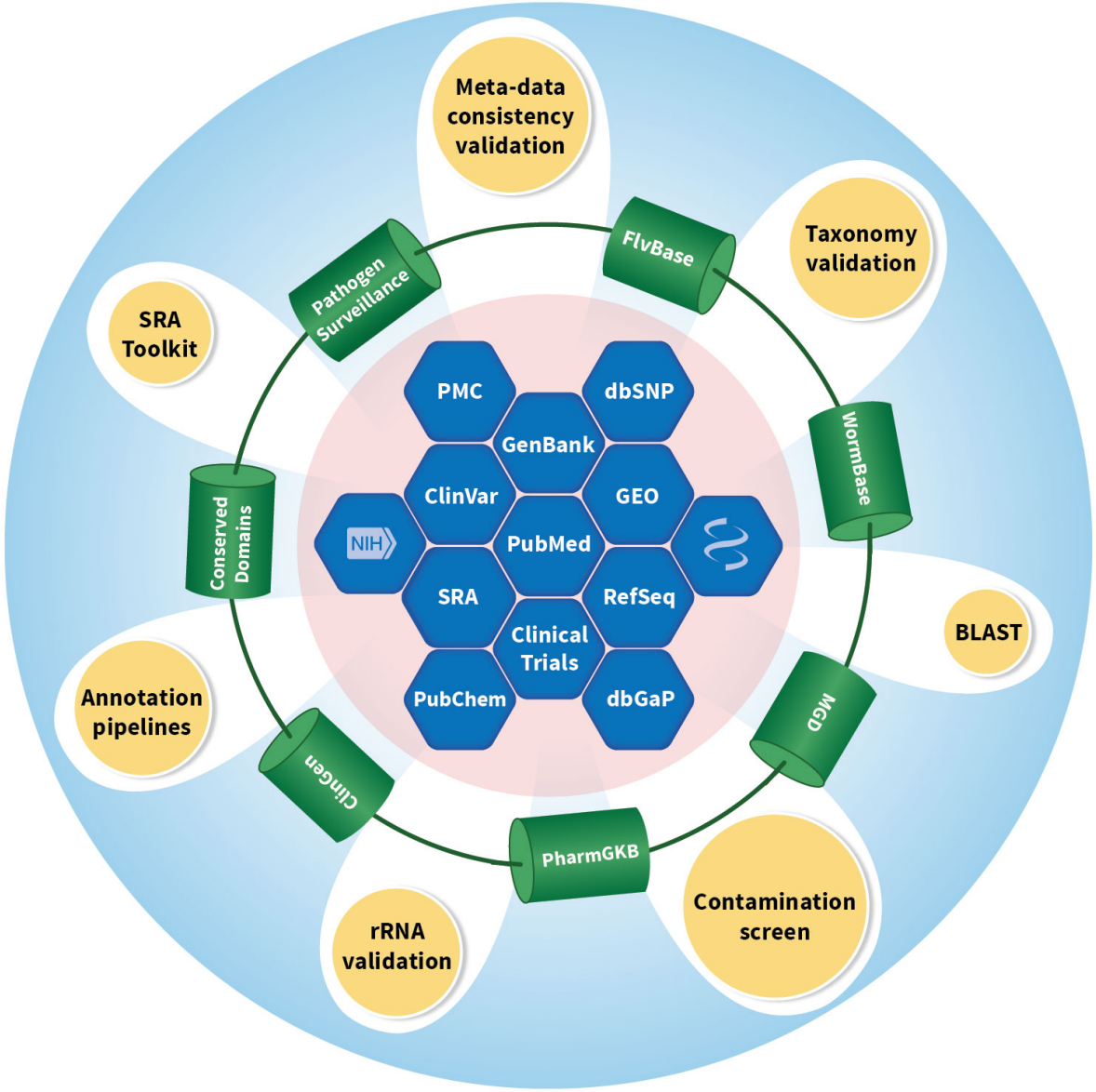


Figure Credit: Jim Ostell, NCBI/NLM

Overarching Goal 4: Enhance *Workforce Development* for Biomedical Data Science

Strategic Objective 4-1: Enhance the NIH Workforce

- E.g., data science training and education for NIH staff

Strategic Objective 4-2: Expand the National Research Workforce

- Enhance quantitative and computational training for graduate students and postdocs

Strategic Objective 4-3: Engage a Broader Community

- E.g., code-athons, bug-bounty programs, contests

Overarching Goal 5: Enact Appropriate *Policies* to Promote *Stewardship and Sustainability*

Strategic Objective 5-1: Develop Policies for a FAIR Data Ecosystem

- Policies must be achievable and minimize administrative burden

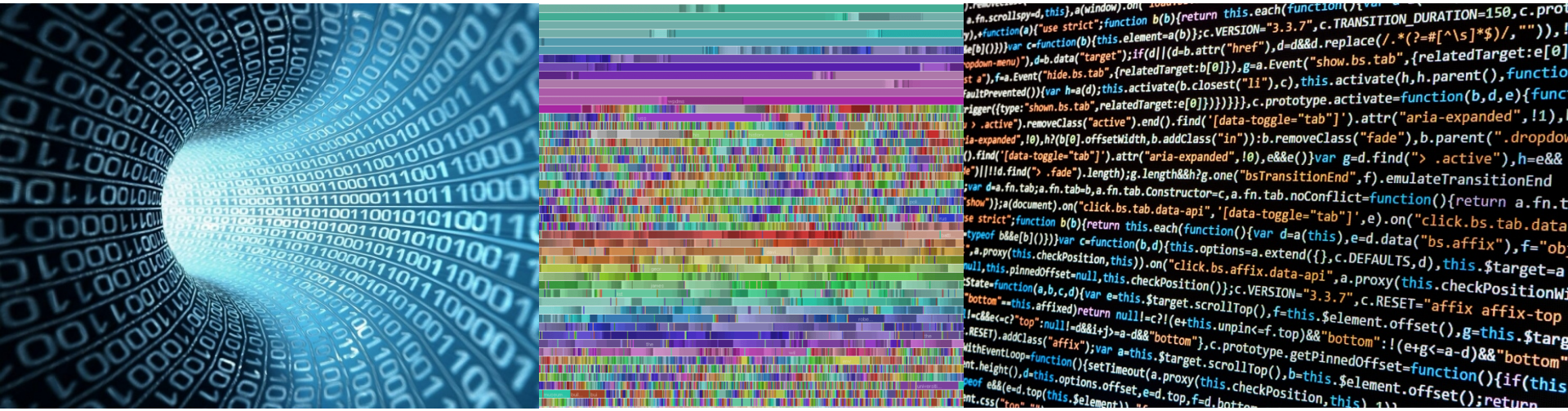
Strategic Objective 5-2: Enhance Stewardship

Implementation Tactics:

- Develop standard use and utility metrics and review expectations for data resources and tools
- Establish sustainability models for data resources

Next Steps

- The Strategic Plan was delivered to Congress in early May
- Posting of final plan is imminent
- The implementation phase has already started and will be ramping up fast
- Development of performance measures and milestones is key



Questions & Comments