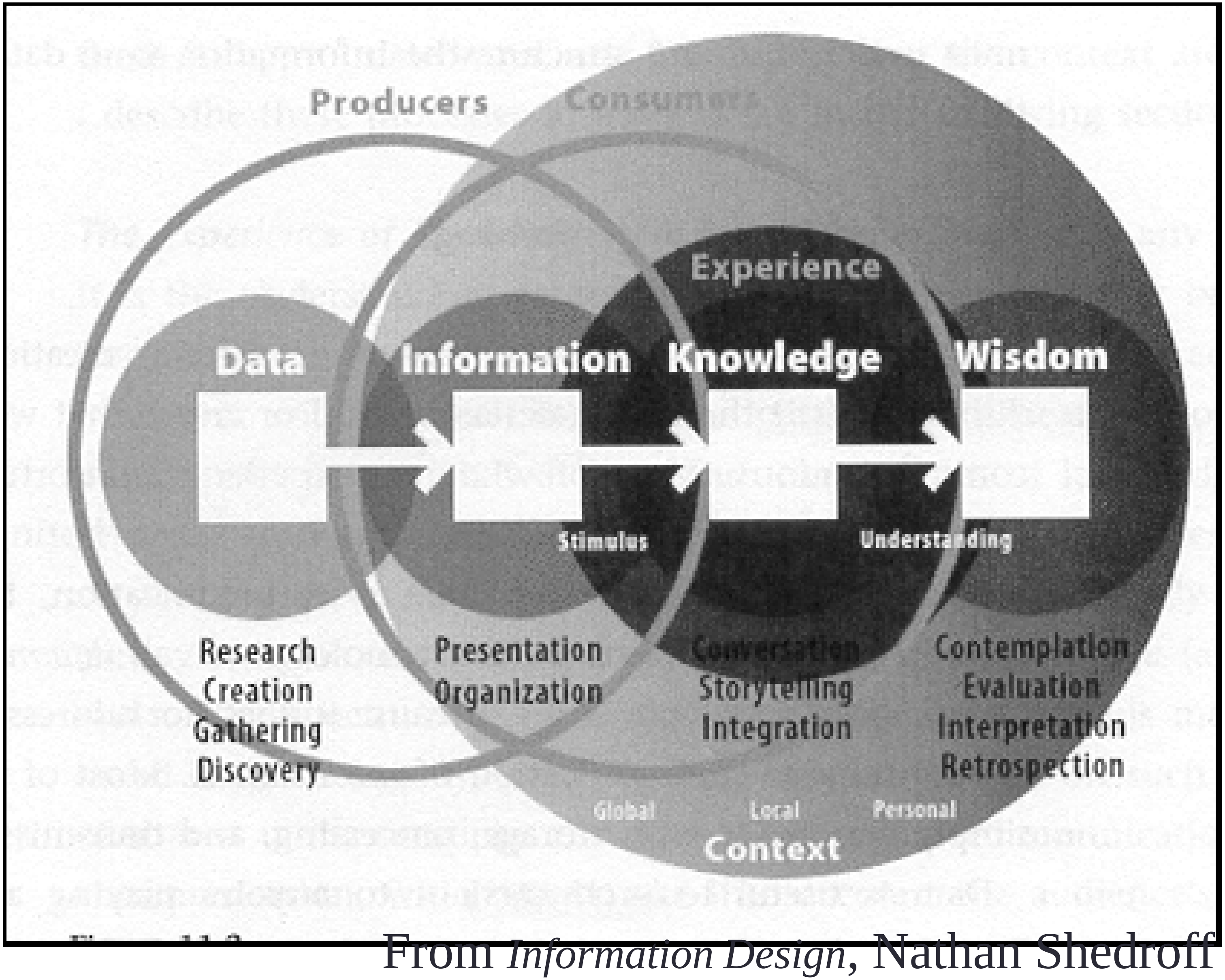
# THE ROLE OF BIOINFORMATICS IN STANDARDIZATION

Kristen Anton

Geisel School of Medicine at Dartmouth / NASA Jet Propulsion Laboratory

December 7, 2012



From *Information Design*, Nathan Shedroff

# Bioinformatics – A Definition

- *General*:  Use of information technology in facilitation of biology-related scientific tasks
- *NASA perspective*:  Focus on data processing pipelines, what you need to produce "standard" data products (Done routinely at NASA for massive data generated from remote sensing instruments)

Clinical trials, genomics and proteomics studies, specimen tracking initiatives *(etc.)*

benefit from study validation and visualization tools, tracking systems, uniform and validated data processing pipelines

# Bioinformatics – Challenges

*Bioinformatics systems:*

- Defining and managing views of bioinformatics models
- Providing model checking capabilities and validation
- Maintaining consistency among distributed information models
- Enabling real-time access to a variety of information that crosses institutional boundaries

*Integration of biomedical information systems:*

- Decentralized technology construction
- Decentralized control
- Unobtrusiveness
- Construction of a flexible system architecture
- Data model and data element variety
- Data ownership
- Training of users
- Security requirements (federal, institutional)

# Bioinformatics – Goals

*Supporting science-driven research needs:*
*Case Study - Early Detection Research Network (EDRN)*

- Research network of collaborating scientists from more than 40 institutions - international
- Focus on identifying and validating biomarkers of cancer at early stage/ preclinical

*Bioinformatics challenges in EDRN:*
*Developing computing infrastructure that is "biomarker-centric."*
*Improve research capability by enabling real-time access to a variety of information that crosses institutional boundaries.*

# Bioinformatics – Goals

*Supporting science-driven research needs*

- Coordinated discovery and validation of biomarkers across cancer research centers to increase accuracy of the results of studies
- Facilitation of analytics through data integration and single-point access
- Support workflows associated with various types of information

# Bioinformatics – Goals

*Supporting science-driven research needs*

- Linking highly diverse systems together to integrate and present data for analytics
- Defining a comprehensive information model for describing the problem space/ ontology
- Providing software interfaces for capture, discovery, and access of data resources
- Providing a secure transfer and distribution infrastructure
- Enabling all data sources to be heterogeneous and distributed
- Providing integrated portal for access to distributed data
- Providing bioinformatics tools/ pipelines for uniform data processing

# Bioinformatics
# EDRN Knowledge Environment

## *Functional architecture: Services*

- Data capture
- Data discovery
- Data access
- Data retrieval
- Data processing
- Data distribution

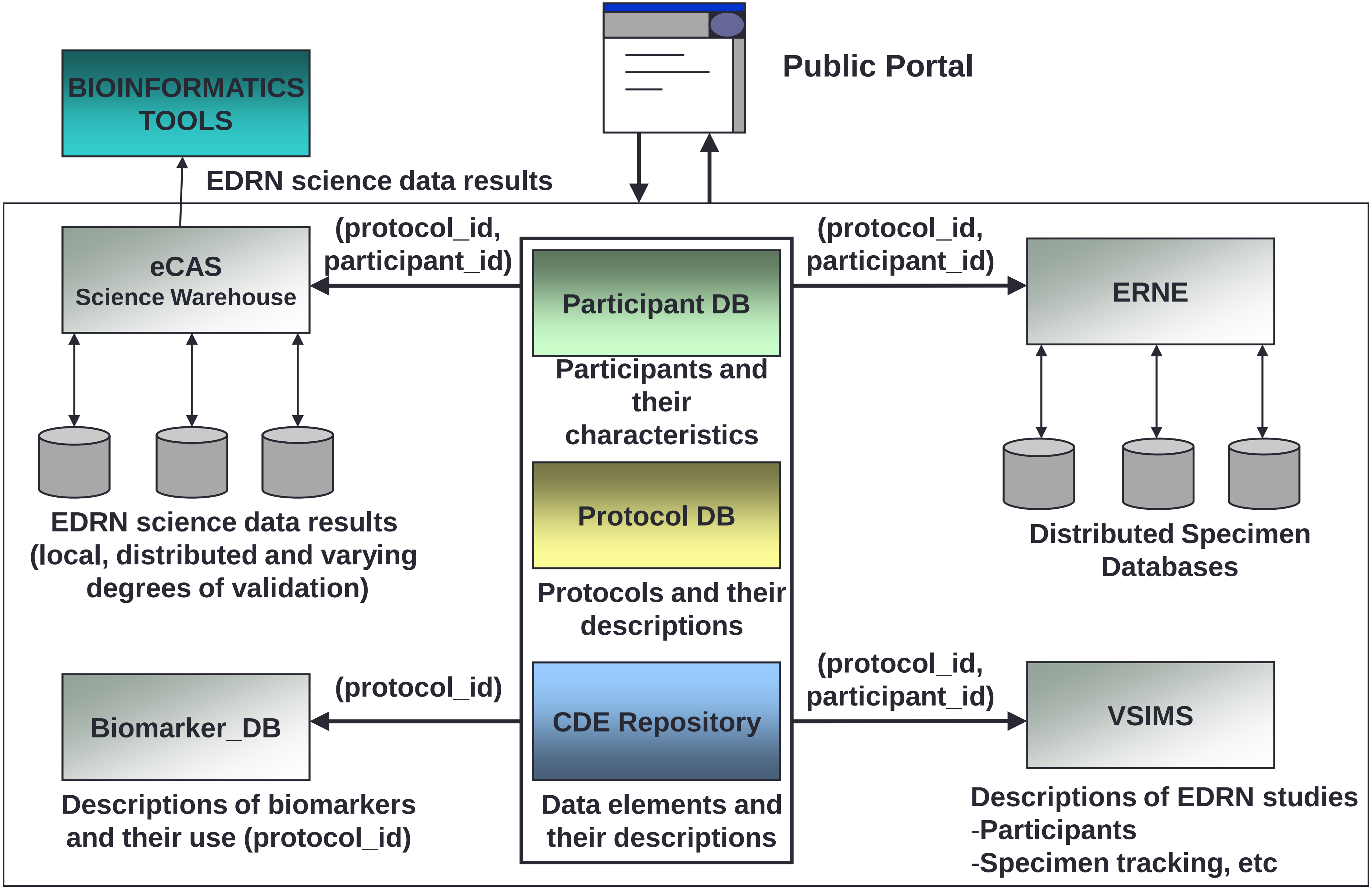
# Bioinformatics
# EDRN Knowledge Environment
## *Information architecture: Data Model*

- Representation of information associated with data objects managed within the knowledge system
- Models for:
  - Biomarkers
  - Studies
  - Participants
  - Organs
  - Data generated from instruments (e.g. mass spec, arrays)

# Bioinformatics
# EDRN Knowledge Environment
## *Information architecture: Data Model*

- Relationships between and among objects
- Standard set of metadata elements that can be used for annotating objects
- Multiple metadata schemata for machine usable explanations of the metadata descriptions
- Metadata descriptions describe the inception and composition of data
- Common language for describing data and associated attributes: Common Data Elements (CDEs)
- CDE has a Uniform Resource Identifier (URI) – URL form points to CDE definition page – used in XML standards

# EDRN Knowledge Environment



Public Portal

BIOINFORMATICS TOOLS

EDRN science data results

eCAS
Science Warehouse

(protocol_id, participant_id)

Participant DB

Participants and their characteristics

Protocol DB

Protocols and their descriptions

CDE Repository

Data elements and their descriptions

(protocol_id, participant_id)

ERNE

Distributed Specimen Databases

EDRN science data results (local, distributed and varying degrees of validation)

(protocol_id)

Biomarker_DB

Descriptions of biomarkers and their use (protocol_id)

(protocol_id, participant_id)

VSIMS

Descriptions of EDRN studies
-Participants
-Specimen tracking, etc

# EDRN Knowledge Environment

## *Success?*

- Biomarker Database holds 401 curated biomarkers, including panels/ signatures of biomarkers
- Biomarker Database modeled to reflect the data model: activity in multiple organs, protocols, data files – facilitate single-point data access
- eSIS contains 165 protocols
- eCAS holds 56 data sets, with many files in each set, and more added daily – standard metadata around each set and each product
- Two bioinformatics tools implemented: Proteomics "pipeline" (generating standardized biomarker identification files); REDCap (standardized data definition and capture at the project level) – eager to add more
- Common Data Elements (CDEs) contributed to the NCI repository
- CDE has a Uniform Resource Identifier (URI) – URL form points to CDE definition page – used in XML standards
- Portal facilitates authorized access to almost 200,000 specimens
- Publications and Resources

# EDRN Knowledge Environment

## *Technology*

- Iterative development
- Open Source philosophy and tools
- Apache OODT (Object Oriented Data Technology)

Software components developed independent of any data model:

*EDRN's computing infrastructure can be replicated*

# EDRN Informatics Team

*Acknowlegements*

- NASA Jet Propulsion Laboratory: Dan Crichton, Chris Mattmann, Heather Kincaid, Sean Kelly, Andrew Hart, Rishi Verma, Michael Joyce
- Dartmouth: Kristen Anton, Maureen Colbert, *BioInformatics Service Center*
- FHCRC: Ziding Feng, Mark Thornquist, Jackie Dahlgren, Suzanna Reid, Deanna Stelling
- NCI: Sudhir Srivastava, Christos Patriotis, *DCP*