Reproducibility and Data Pipelines

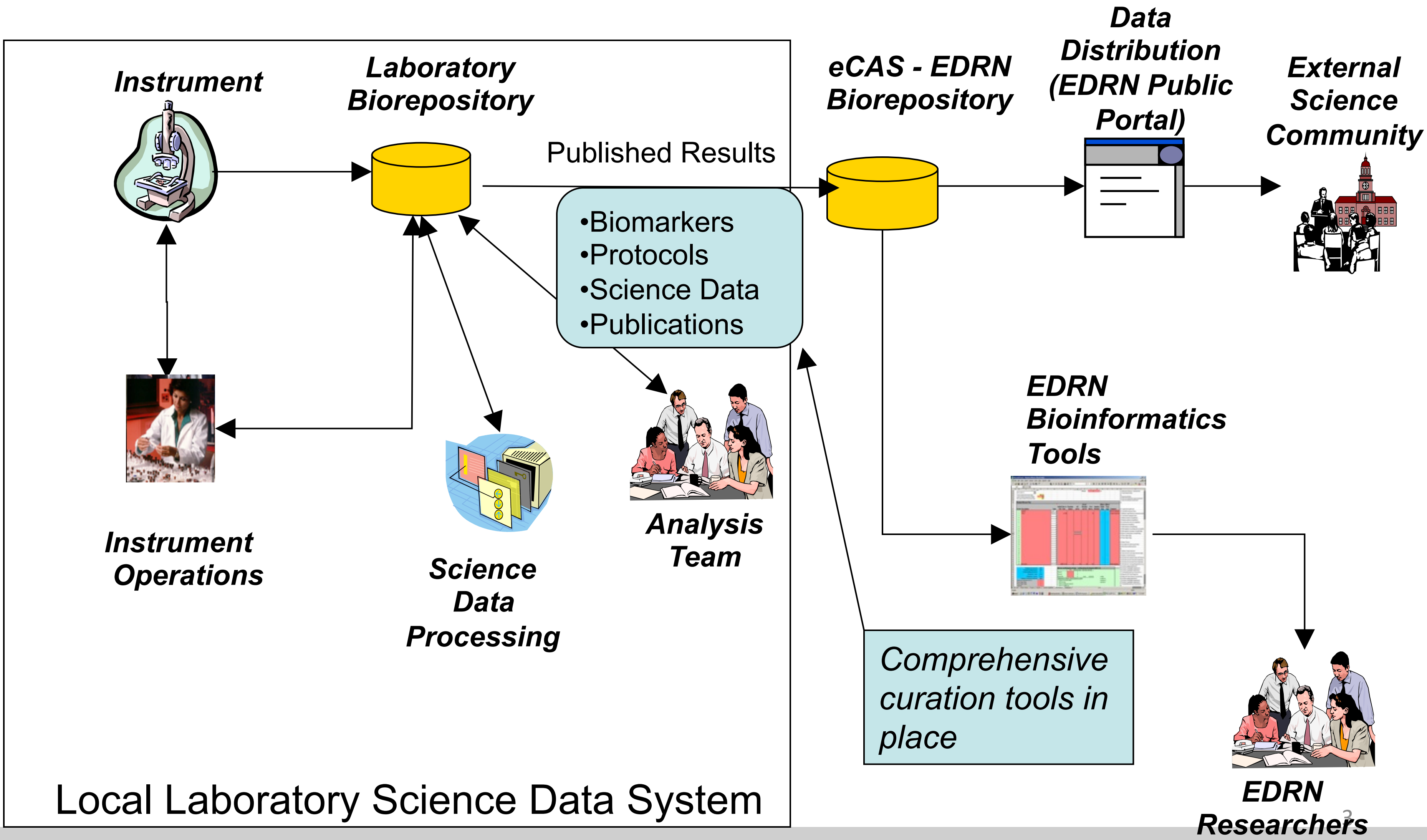
.

Agenda

- Introduction – D. Crichton
- Vanderbilt/JPL Collaboration on a Scalable Data Processing Pipeline for Proteomics – L. Cinquini
- Genomics Biomarker Discovery Pipeline – Marc Lenberg, Ashish Mahabal
- Reproducible Data Science in EDRN – Emplified with the PLCO Ovarian Phase III Validation Study – T. Fuchs

Capture of Public Science Data

An Integrated Repository of Public Data Sets



Instrument

Laboratory Biorepository

Published Results

eCAS - EDRN Biorepository

Data Distribution (EDRN Public Portal)

External Science Community

- Biomarkers
- Protocols
- Science Data
- Publications

Instrument Operations

Science Data Processing

Analysis Team

EDRN Bioinformatics Tools

Comprehensive curation tools in place

EDRN Researchers

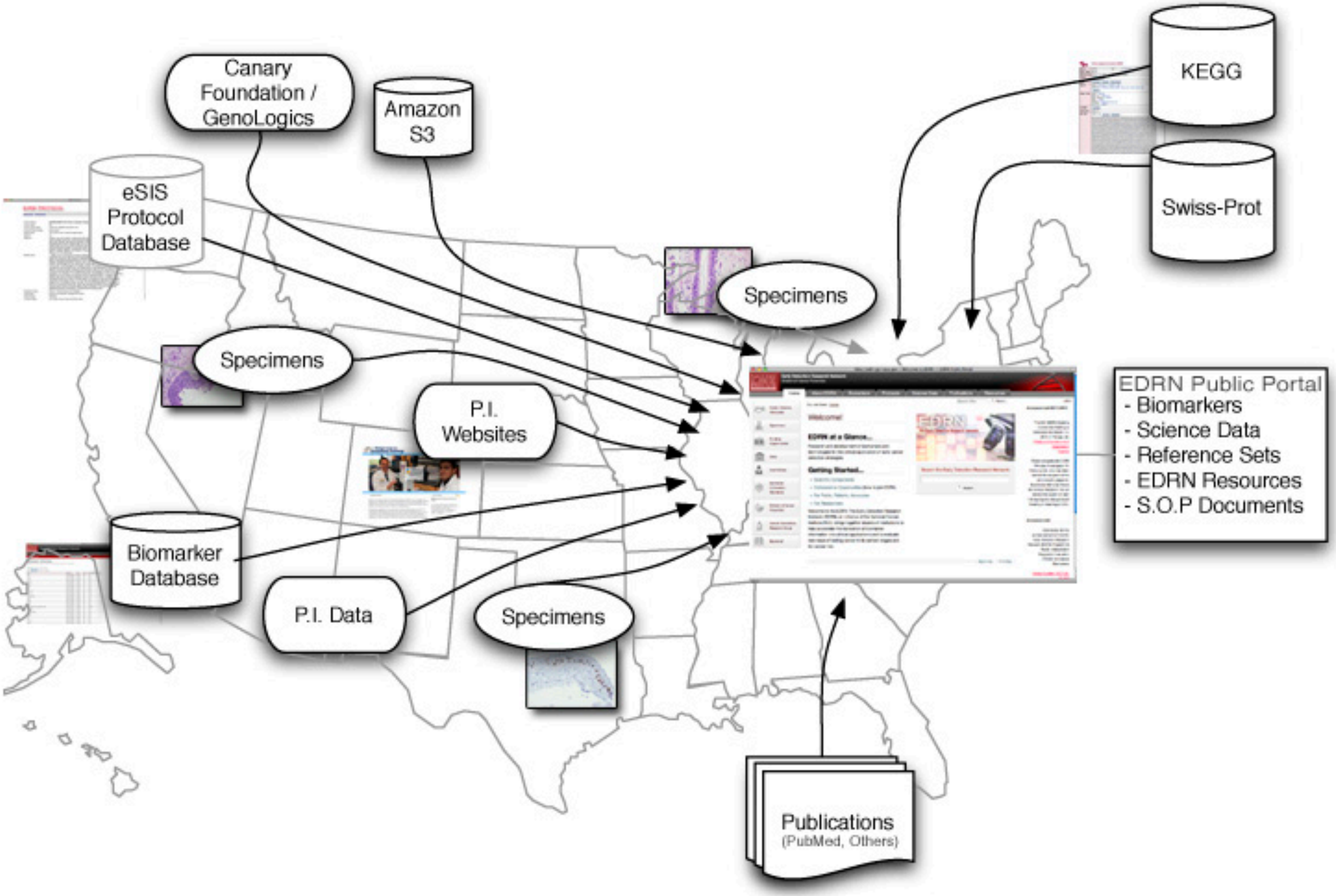
Local Laboratory Science Data System

Early Detection Research Network

A Virtual, National Integration Biomarkers Knowledge System

NASA

JPL
Jet Propulsion Laboratory
California Institute of Technology



EDRN Public Portal
- Biomarkers
- Science Data
- Reference Sets
- EDRN Resources
- S.O.P Documents

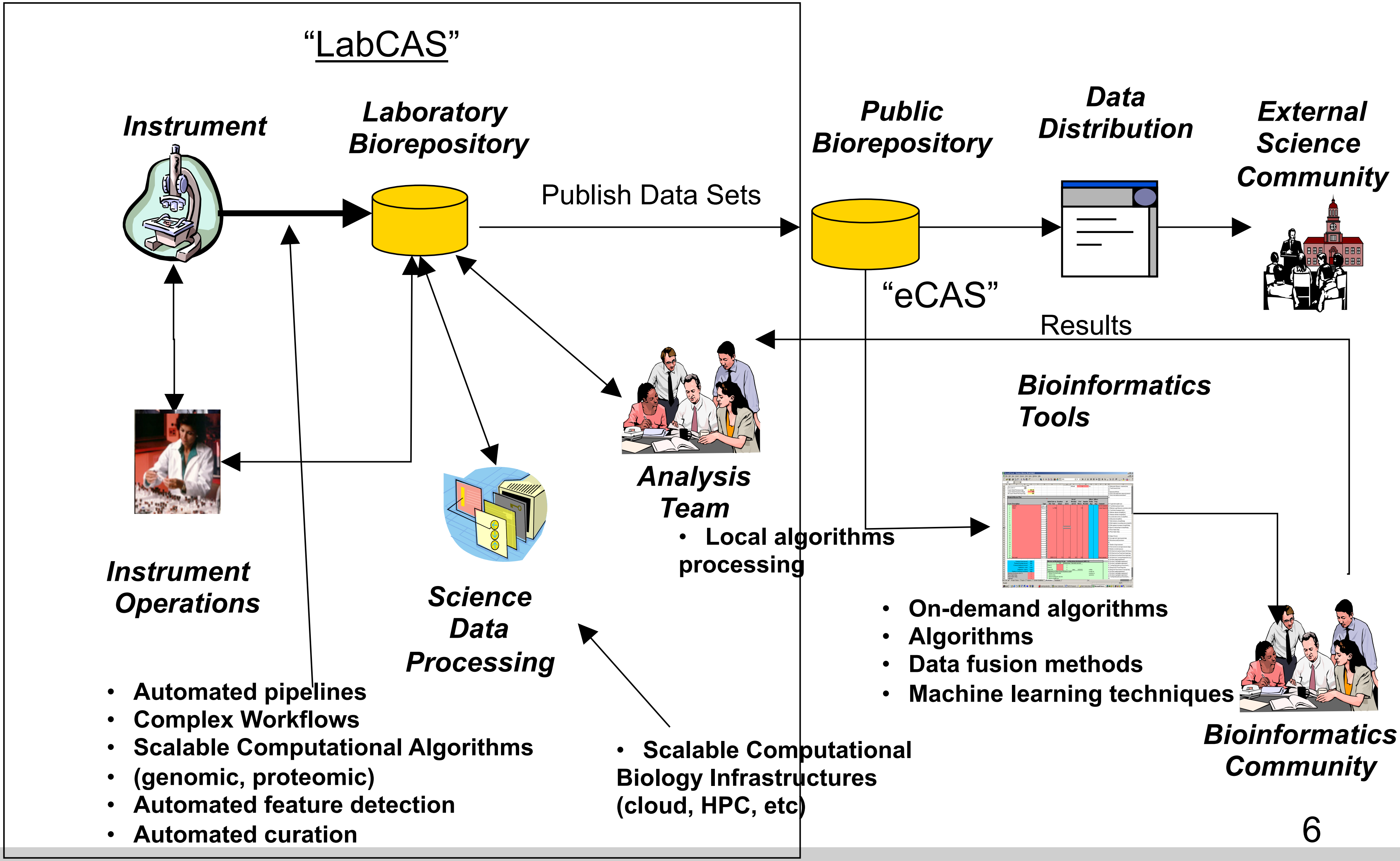
Cancer Biomarker Bioinformatics Workshop

- The EDRN and NASA Jet Propulsion Laboratory held a workshop in May 2013 at Caltech to address informatics and data-driven research in cancer biomarkers
  - http://edrn.nci.nih.gov/cancer-bioinformatics-workshop/cancer-biomarker-bioinformatics-workshop-report-may-2013
  - A major outcome focused on data usability, reproducibility of results, methods and algorithms to systematize data analysis, and scalable computing infrastructures.
- Key Recommendations
  - **Systematic approaches to the generation, capture, management of data to enable reproducibility.**
  - Increased emphasis on data curation to promote data reuse
  - **Automation of data processing/analytics software pipelines**
  - Data integration and fusion of data from multiple platforms, studies
  - Scalable data infrastructures and repositories
  - **Use of big data tools and bioinformatics techniques to scale data analysis**
  - Increased training of scientists in the use of computational tools/methods

Early Detection Research Network

Moving towards data-driven science for cancer biomarkers

NASA JPL Jet Propulsion Laboratory California Institute of Technology



"LabCAS"

Instrument

Laboratory Biorepository

Publish Data Sets

Public Biorepository

Data Distribution

External Science Community

"eCAS"

Results

Bioinformatics Tools

Analysis Team

- Local algorithms processing

Instrument Operations

Science Data Processing

- On-demand algorithms
- Algorithms
- Data fusion methods
- Machine learning techniques

Bioinformatics Community

- Automated pipelines
- Complex Workflows
- Scalable Computational Algorithms
- (genomic, proteomic)
- Automated feature detection
- Automated curation

- Scalable Computational Biology Infrastructures (cloud, HPC, etc)

Agenda

- Introduction – D. Crichton
- Vanderbilt/JPL Collaboration on a Scalable Data Processing Pipeline for Proteomics – L. Cinquini
- Genomics Biomarker Discovery Pipeline – Marc Lenberg, Ashish Mahabal
- Reproducible Data Science in EDRN – Exemplified with the PLCO Ovarian Phase III Validation Study – T. Fuchs

Backup