Early Detection Research Network

# EDRN Informatics Infrastructure

**Dan Crichton**
PI, JPL Informatics Center
Co-Director, JPL/Caltech Center for Data Science and Technology
NASA Jet Propulsion Laboratory, Caltech

With Heather Kincaid, NASA/JPL & Kristen Anton, Geisel School of Medicine, Dartmouth

*(in collaboration with NCI, Dartmouth, Fred Hutchinson Cancer Center and many others …)*

NASA JPL Jet Propulsion Laboratory California Institute of Technology

National Institutes of Health National Cancer Institute

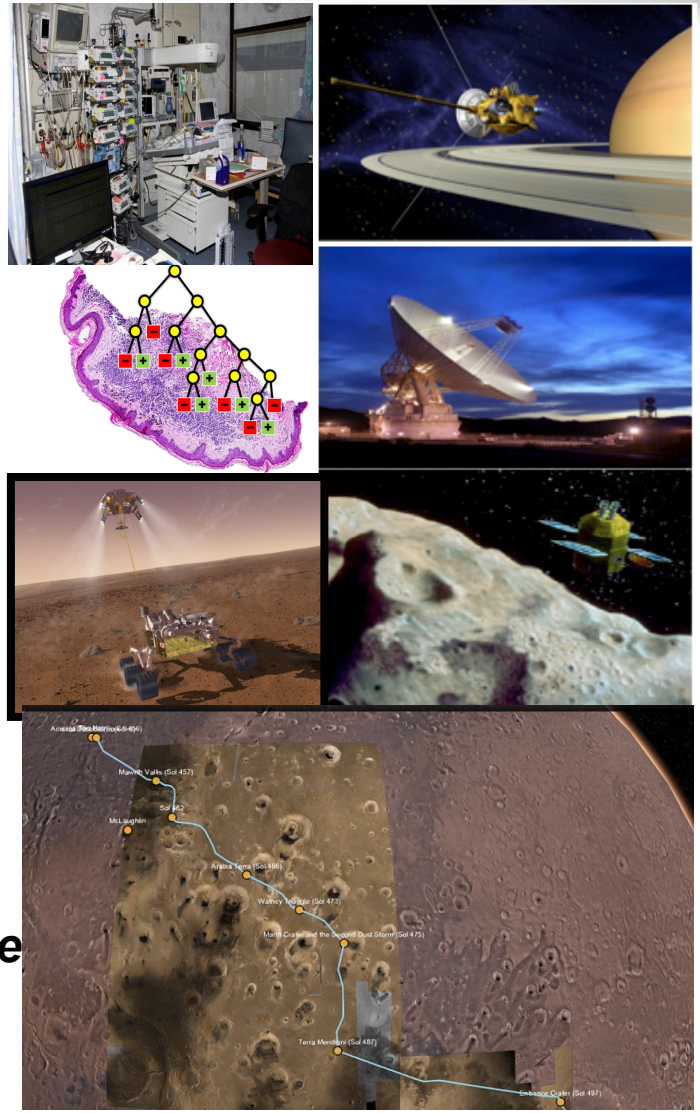
October 15, 2015

# NASA/JPL Informatics Center

Early Detection Research Network

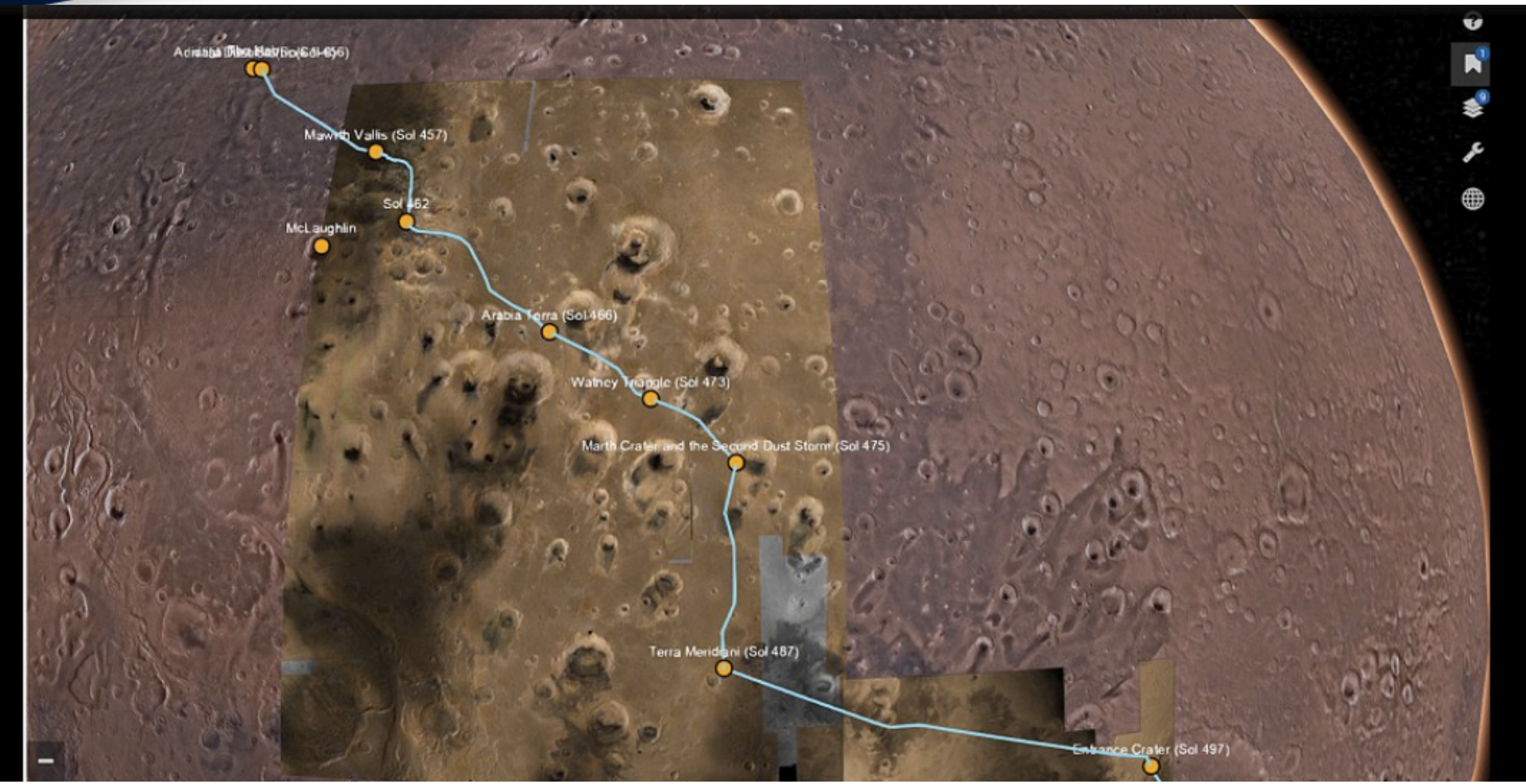
- **Development of an advanced Knowledge System for EDRN to *capture*, *share* and support *reproducible analysis* from the biomarker data results**

- **NASA-NCI partnership, leveraging informatics and data science technologies from planetary and Earth science**
  - **Reproducible, Big Data Systems for exploring the universe**
  - **2011 NASA Group Award for "*innovative use of NASA software technologies to support cancer research*"**



# Example: Exploring Mars Imaging Data with Mars Trek

Early Detection Research Network




Derived from about 800 TBs of Imaging Data in the NASA Planetary Data System

# Building Bioinformatics Infrastructure

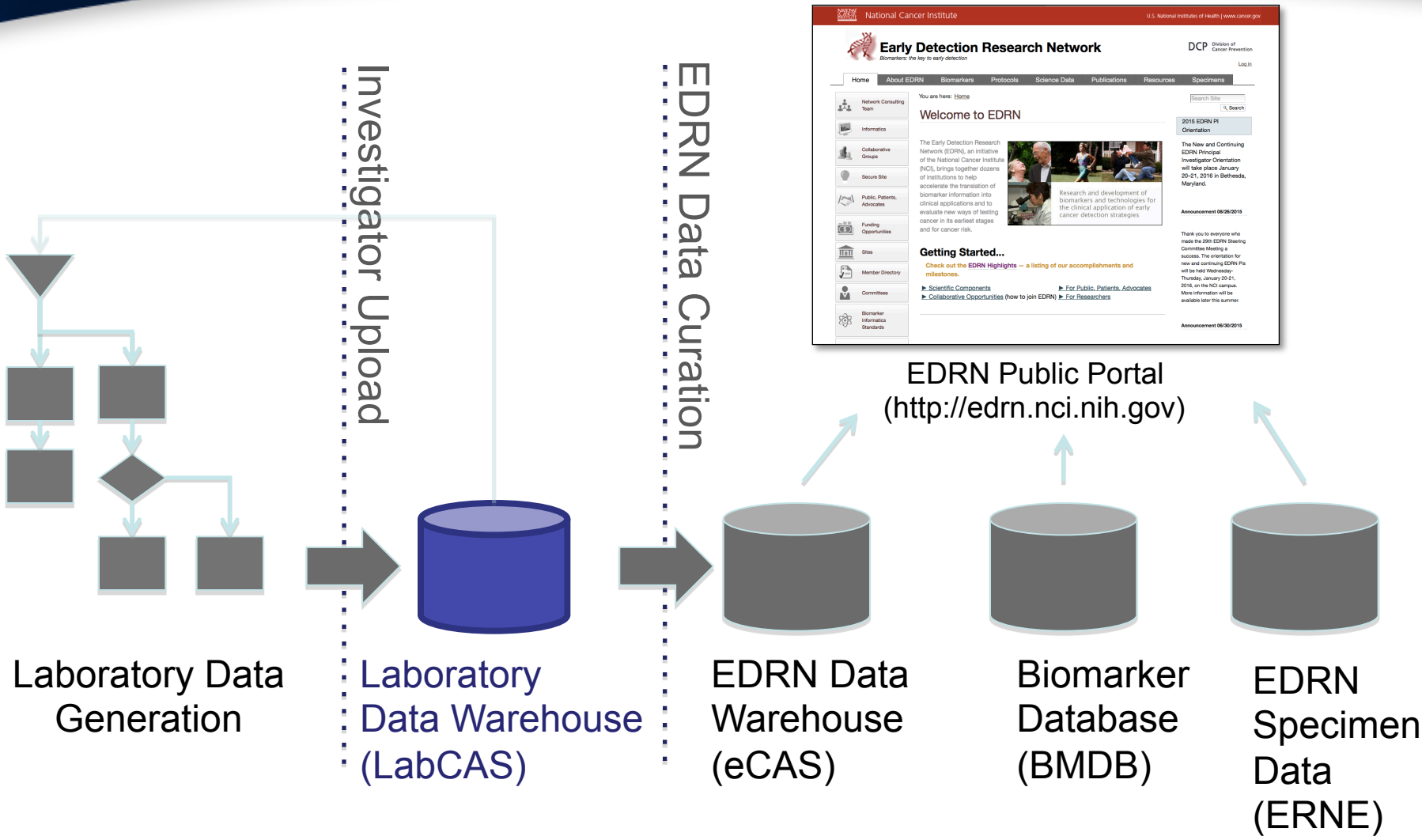
Early Detection Research Network

**In 2000, EDRN investigators identified the following needs:**

- Ability to develop informatics standards such as common data elements (CDEs) for biomarker research to allow for data to be linked together and analyzed
- Informatics to support multi-institutional validation studies
- Create tools to facilitate the use of the informatics standards for data collection (e.g. CDE Form Tool, EDRN Data Model, Public Portal)
- Ability to access specimens across EDRN Clinical validation Centers
- Ability to support secure data transfer, data analysis and communication
- Ability to support EDRN-wide data storage, curation and retrieval of multidimensional, multi-format data

# Capturing and Sharing Biomarker Data Results

Early Detection Research Network



# EDRN Informatics Projects

Early Detection Research Network

**EDRN has a number of focused projects that have been started in informatics to support the goals of capturing and sharing the state of biomarker research**

Virtual Specimen Repository (ERNE): Access to information on specimens across EDRN;

EDRN-wide Portal: Access to EDRN-wide biomarker study results;

Science Data Warehouse (eCAS): Capture of biomarker data results from EDRN studies into a central repository. Security integrated.

Biomarker Database: Capture of biomarkers under study within the EDRN

Common Data Elements: Common sets of terms used to construct databases and forms

Laboratory Data Sharing (LabCAS): Enhance EDRN capabilities to automated the processing, capture and sharing of data from EDRN studies

# Key Informatics Accomplishments

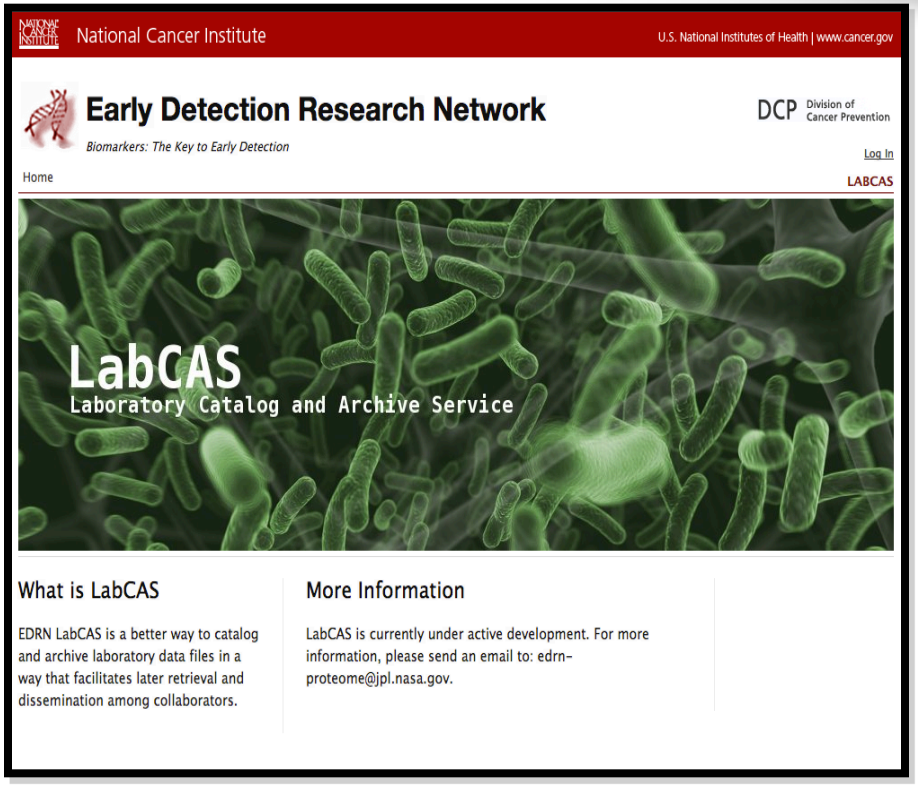
Early Detection Research Network

- Developed a **national, biomarker knowledge system** using advanced informatics technology
- Pioneered the concept providing **access to information** about **biospecimens** across EDRN at a national level (2001)
- Developed a **repository for capturing scientific data sets**; captured 90 data sets; integrated with the Canary Foundation infrastructure.
- Developed a **biomarker database** for capturing EDRN biomarkers; captured over 900 biomarkers
- Developed a **public portal** that provides **dissemination** of EDRN information as well as scientific data and results; over 2400 unique visitors a month
- Developing new tools for the Laboratory's to support the **processing, capture, curation and sharing of data** before publications
- Received NASA Award in 2011 for the "innovative use of NASA software technologies to support cancer research" due to **significant reuse of capability**.

# LabCAS: Laboratory Catalog and Archive Service

Early Detection Research Network

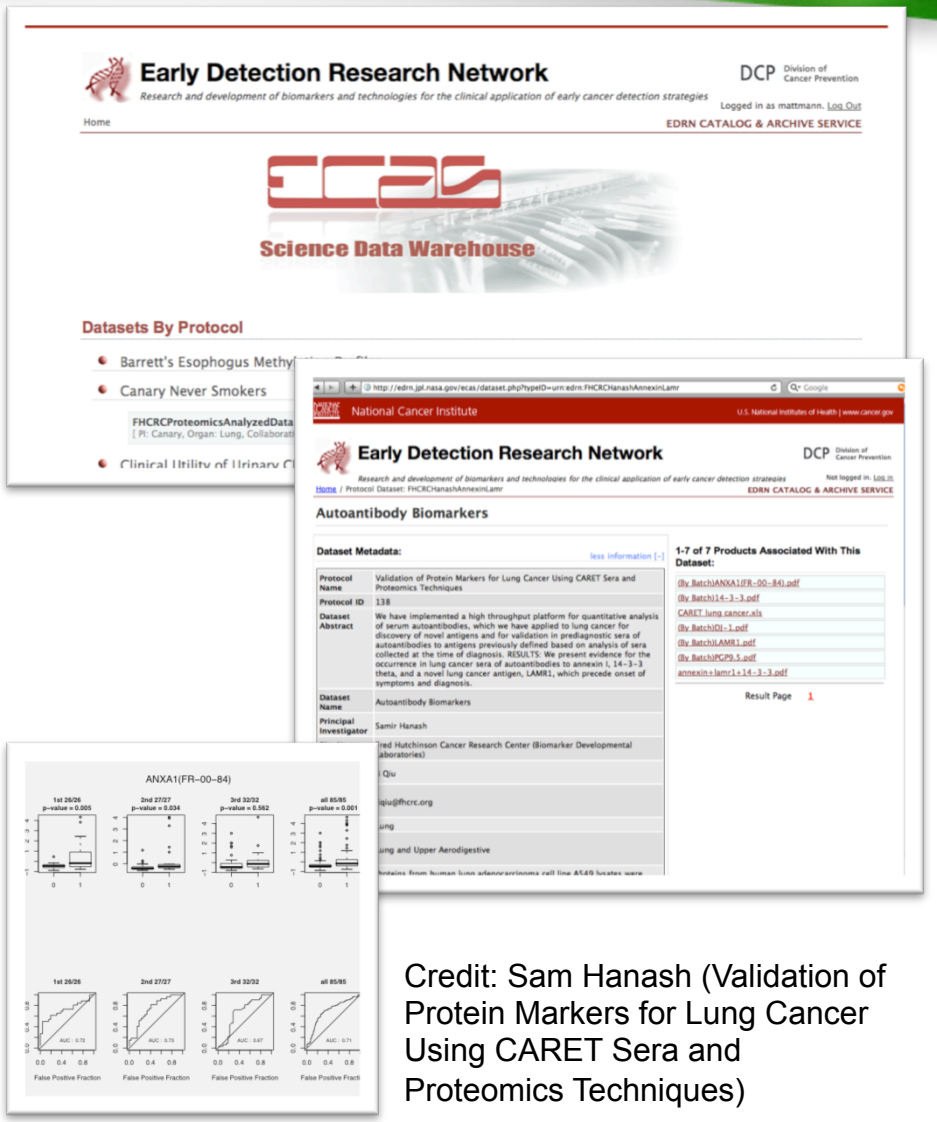
- LabCAS is a new capability under development to provide investigators with a secure, reliable means to capture their **pre-publication** research datasets
- LabCAS also provides integrated data processing
- Enable investigators and collaborative groups/projects to share data in a secure manner as early as possible
- Scale to support data intensive projects
  - Repeatable data processing pipelines



# eCAS: Capture and Sharing of Public Data Sets

Early Detection Research Network

- **EDRN has a warehouse of public biomarker data for use today!**
  - Uses the EDRN CDEs to populate a catalog describing the data sets
  - Supports public release/access to the data
  - Supports peer review of the data by collaborative groups prior to public release
  - Integrated with the rest of EDRN systems
- **Provides a long term and central capture of EDRN study results for the broad community**




Credit: Sam Hanash (Validation of Protein Markers for Lung Cancer Using CARET Sera and Proteomics Techniques)

# Biomarker Database: Capture of EDRN Biomarkers Under Research

Early Detection Research Network

- **A database of annotated biomarkers that are either under development or reported in publications**
  - Over 900 biomarkers captured
  - Based on EDRN research
- **A national biomarker registry to track EDRN progress**
  - Similar to gene and protein type registries
  - Essential to tracking progress and supporting national collaboration of biomarker discovery and validation
- **Share results with the broader research community**
- **Integrate with existing databases (e.g., genomic, publication, etc. databases)**

# Biomarker Database Content

Early Detection Research Network

928 individual biomarkers

Ten organs represented

- Bladder (2)
- Breast (193)
- Colon (13)
- Esophagus (11)
- Head and Neck (8)
- Liver (9)
- Lung (182)
- Ovary (205)
- Pancreas (7)
- Prostate (387)

- 27 biomarker “panels” or “signatures” included
- 79 biomarkers associated with multiple organs (e.g. information detailing *p16* activity in Breast, Esophagus, Lung and Prostate)
- Expert review completed in each Collaborative Group
- Markers from all four Collaborative Groups “Accepted” and released to public view

*Research, Communicate, Curate, Annotate*

# Ability to locate specimens across EDRN Clinical Centers: ERNE

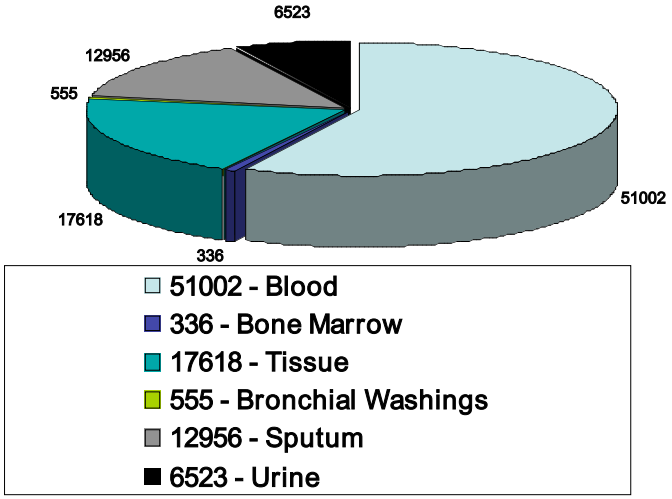
Early Detection Research Network

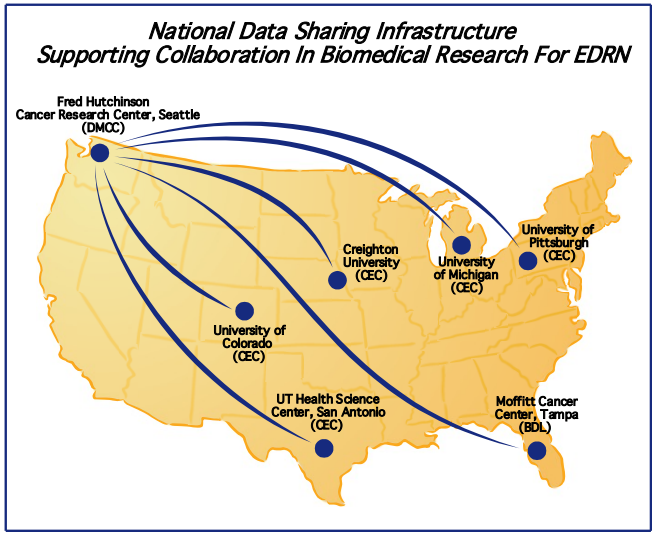
- **Specimen Locator System, dubbed ERNE, EDRN Resource Network Exchange was developed to query data across EDRN's Clinical Validation Centers (CVC)**
- **The system is based on NASA JPL's Object-oriented Data module which can be easily tailored to the CVC's institutional informatics system**
  - Same software module used to share earth and planetary science data
- **ERNE allows the user to query the availability of specimens in real-time**
- **This is a first-ever system developed to query specimen on disparately distributed specimens across the country**
  - "ERNE" has been a model studied by many groups

# Participants in Virtual Specimen Bank

Early Detection Research Network

- H. Lee Moffitt Cancer Center
- University of Texas, San Antonio
- Creighton University
- University of Colorado
- University of Pittsburgh
- University of Michigan/Dartmouth University (Great Lakes New England Consortium)
- Brigham and Womens
- MD Anderson
- New York University
- UCSD
- Center for Disease Control
- Johns Hopkins
- Duke University
- Fred Hutchinson Cancer Research Center
- Fox Chase Cancer Research Center



# Public Portal

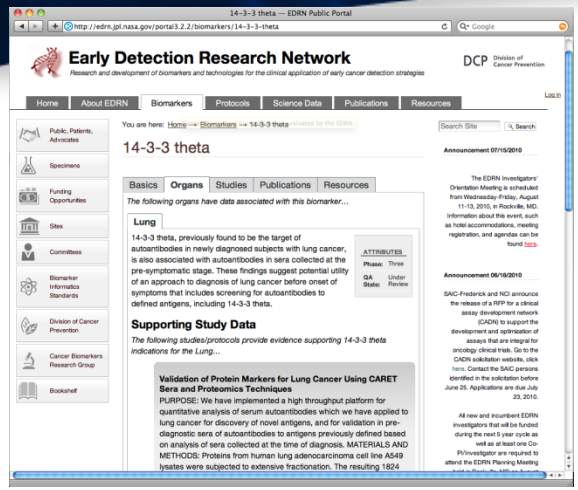
- **A one stop shop to access EDRN programmatic and science information**
  - Operated by NCI on the cancer.gov network
  - Used by NCI and EDRN for disseminating program information
  - Information spans from biomarker data all the way to member information
  - Integrates information from the DMCC
  - **Google-like search feature**

- **We have approximately 900 registered users and we get about 2400 unique visits a month**
  - A lot of ad hoc requests for data from outside EDRN

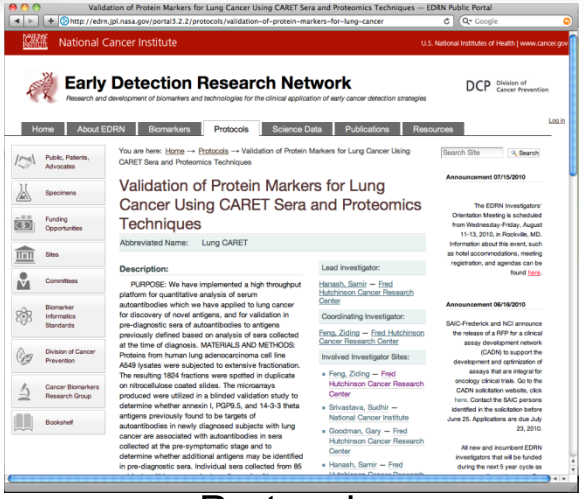

EDRN Public Portal
(http://edrn.nci.nih.gov)

# Accessing EDRN Data: Data may come from any of these sources

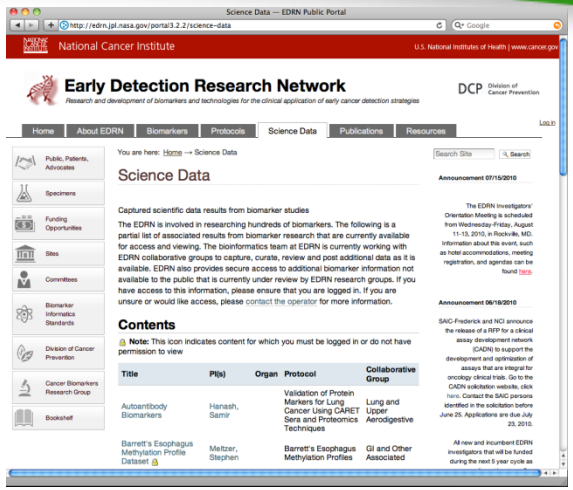
Early Detection Research Network



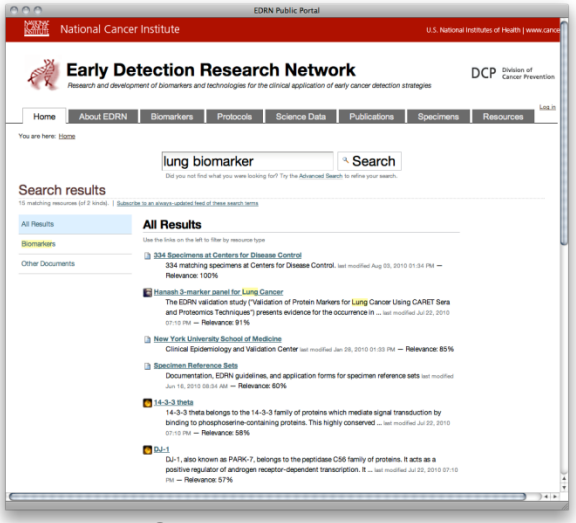

Biomarker Annotations




Protocols

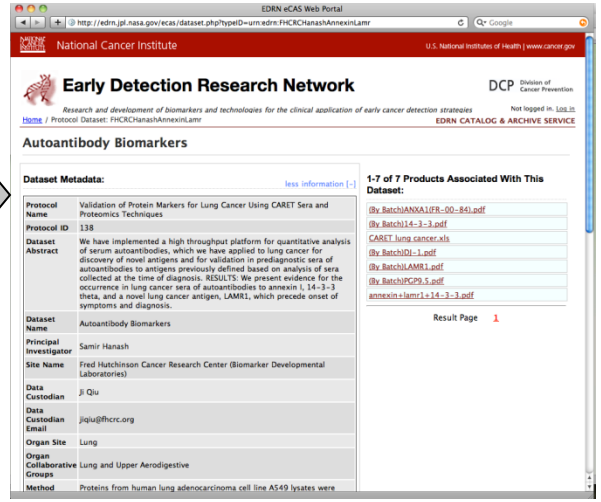



Biomarker Data Results



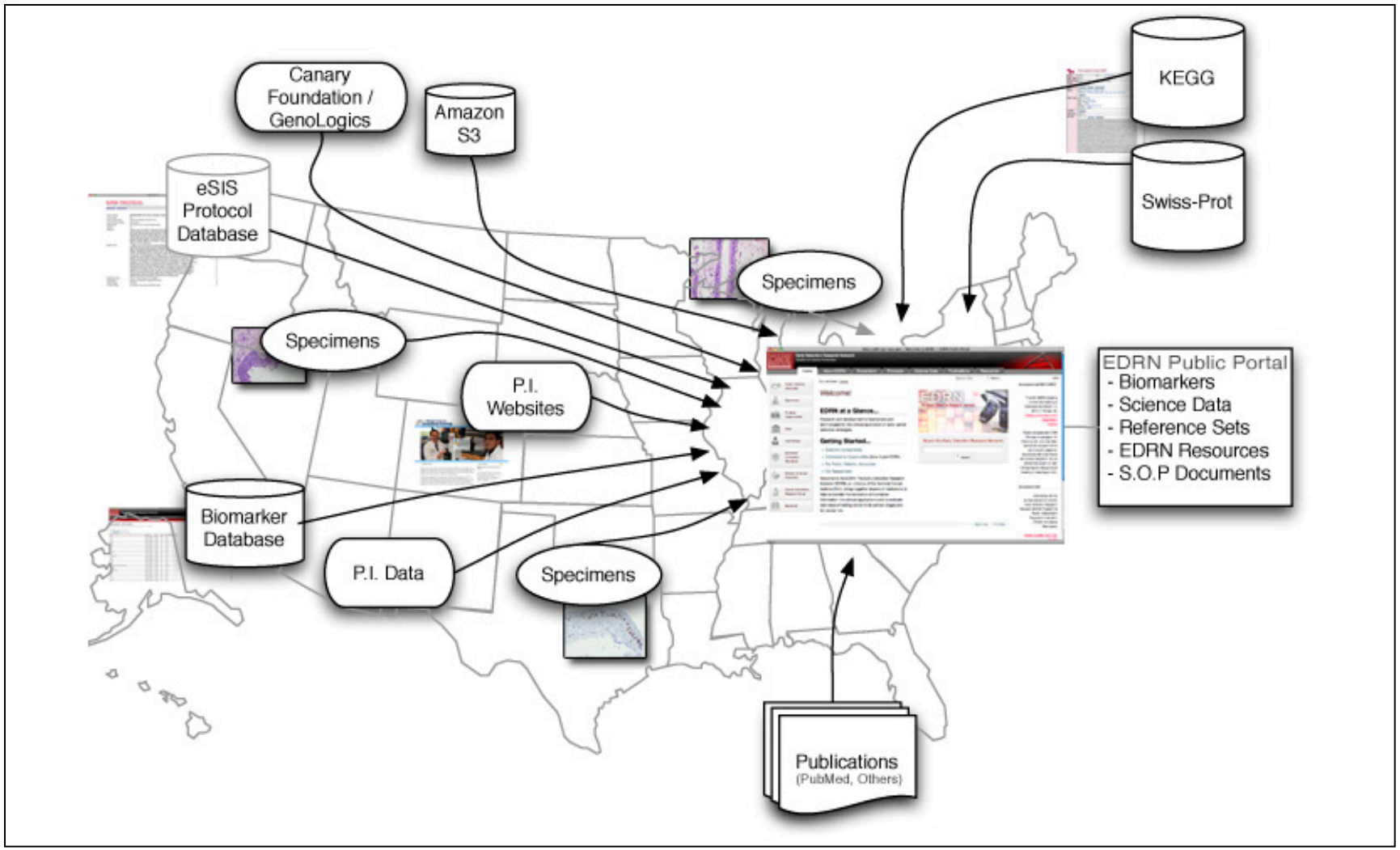

Specimens


Linked through Public Portal




Access to download data

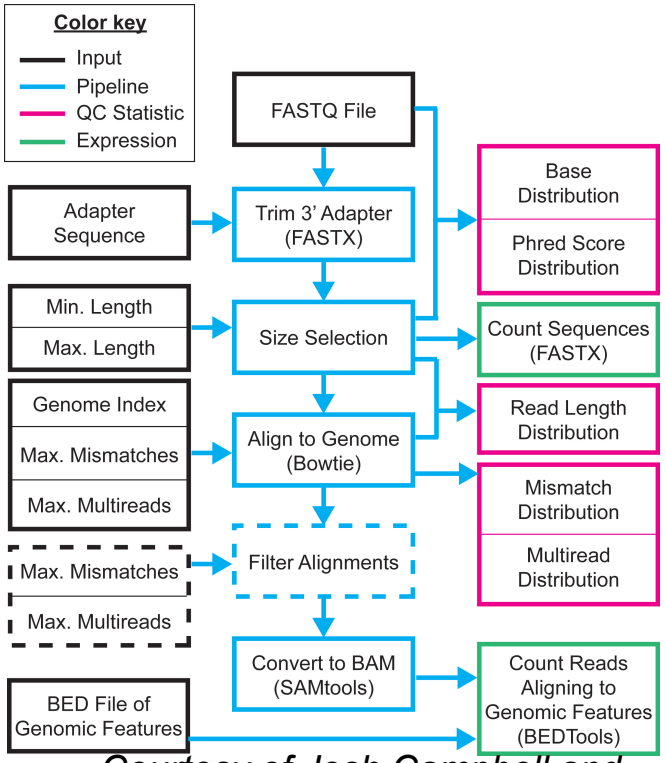
# What's Emerged…
# The EDRN Knowledge System



# Numerous Collaborations

Early Detection Research Network

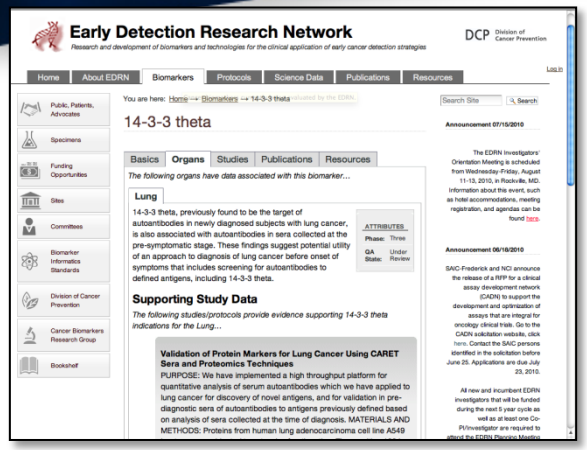
- **Reproducibility experiments**
  - EDRN PLCO
- **Analytic Data Pipelines**
  - Boston University
  - Vanderbilt
  - PNNL
  - NIST/Stanford
- **Bioinformatics Tool Integration**
  - Secretome
- **Data citation in eCAS**
- **Biomarker Database/Results**



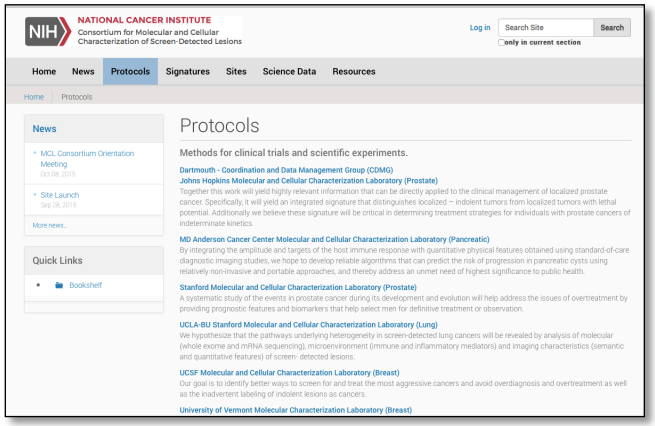

*Courtesy of Josh Campbell and Teresa Wang*

# MCL Knowledge Environment

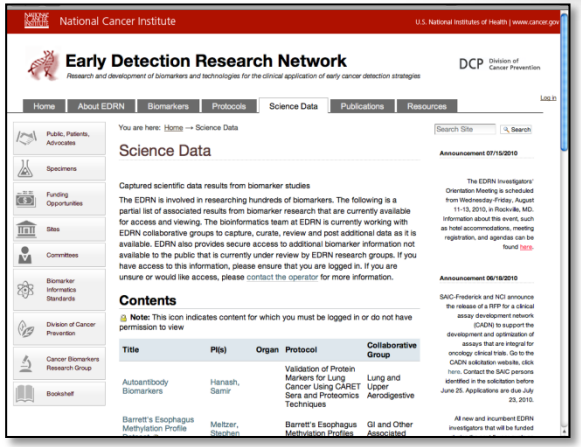
Early Detection Research Network



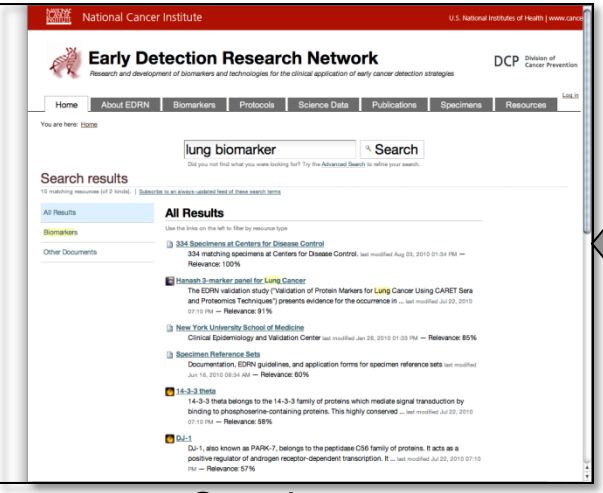

Signature Annotations (Biomarker Database)



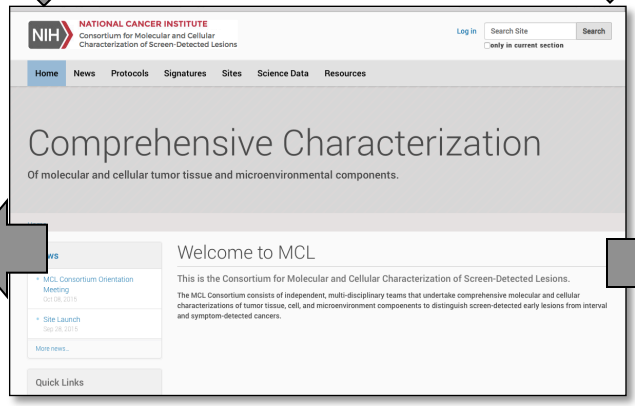

Protocols



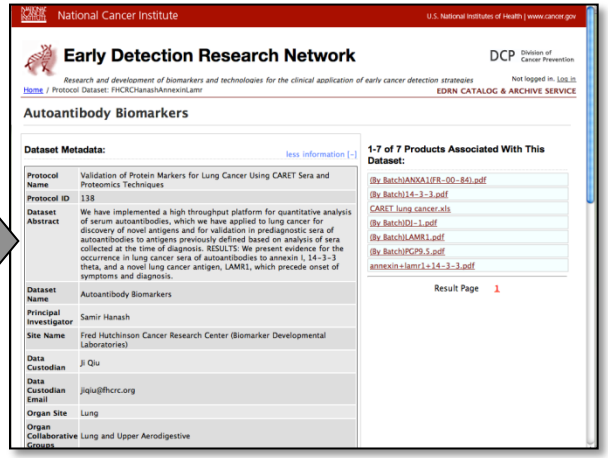

Signature Data Results (eCAS)




Specimens (ERNE)



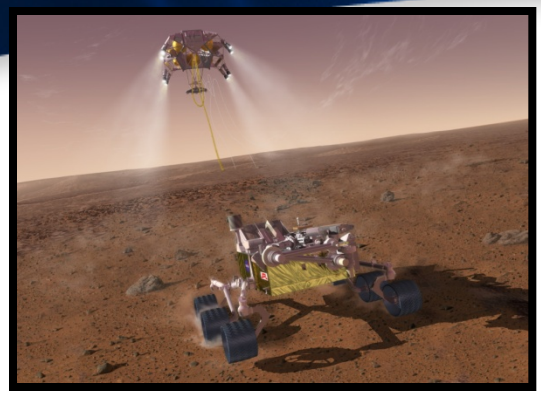

Linked through Public Portal




Access to download data (eCAS)

# Conclusion

Early Detection Research Network





Reproduced from *Wired* magazine

**EDRN has a focused, leading-edge, informatics platform that can be leveraged today for sharing data.**

http://twitter.com/edrn_ic

http://www.facebook.com/group.php?gid=56938589930