2015

From Stars to Patients: Lessons from Space Science and Astrophysics for Health Care Informatics

S. George Djorgovski
California Institute of Technology

A. A. Mahabal
Caltech

D. J. Crichton
JPL

Basit Chaudhry
Tuple Health

From Stars to Patients: Lessons from Space Science and Astrophysics for the Health Care Informatics

S.G. Djorgovski, A.A. Mahabal (Caltech), D.J. Crichton (JPL), B. Chaudhry (TupleHealth)

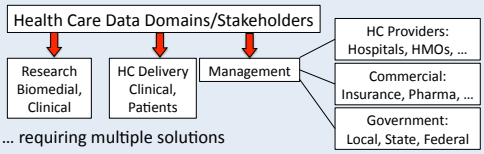
The motivation and the challenge:

Big Data are revolutionizing nearly every aspect of the modern society. One area where this can have a profound positive societal impact is the field of Health Care Informatics (HCI), which faces many challenges.

The key idea here is: can we use some of the experience and solutions from the fields that have successfully adapted to the Big Data era, namely astronomy and space science, to help accelerate the progress of HCI?

An Approach to HCI

It is a complex field with many constituencies and goals:



... requiring multiple solutions

We surveyed the HCI literature:

- Published studies cover: Bioinformatics; Neuroinformatics; Clinical Informatics; Public Health Informatics; Translational Bioinformatics
- Tools/analytics used include: Fuzzy trees, Random Forests, Area Under the Curve, Sensitivity, Specificity, Social Networks and Crowdsourcing
- The most frequently cited problems include:
  - Lack of interoperability, standards
  - Especially across data types: patients, disease, treatments, healthcare management
- There is great diversity in platforms/systems used, a likely hindrance to reproducibility: CareWeb, Caradigm Amalga, CER Hub, RedX, GLORE etc.

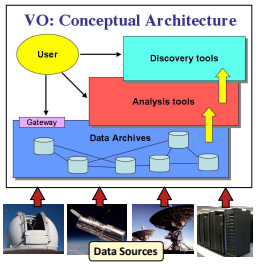
We surveyed the publicly available data sets:

- Most data are not publicly exposed due to: Proprietary nature; Monetary interests; Information blocking; Privacy issues
- Typical available data sets cover molecular, tissue, patient, population studies – but not in a connected fashion (diversity of formats and access mechanisms)
- Typical sizes of the publicly available data range from kB (highly derived data products) to GB-TB or even PB (raw instrument data)
- Most data sets do not have enough: metadata, standards, provenance
- Privacy protection limits hinder the attempts at reproducibility and possibilities of connecting multiple datasets easily
- There is growing emphasis on sustainability, rewards systems, usability

The Astronomy community's grassroots response to the challenges and opportunities of Big Data was

The Virtual Observatory (VO) Concept:

A complete, dynamical, distributed, open ***research environment for the new astronomy with massive and complex data sets***



VO provides and federates data and metadata from distributed archives, develops and provides data services, standards, and data analysis and exploration services

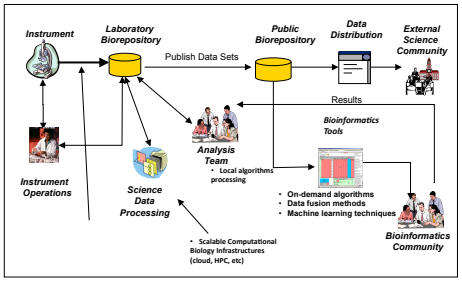
How Did the VO Succeed?

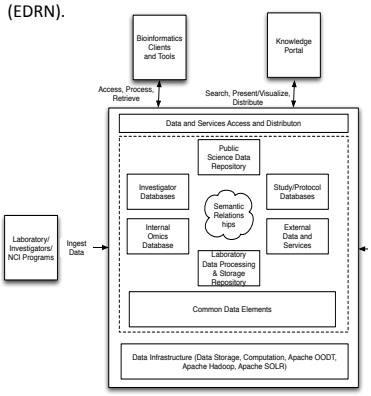
- All data collected in a digital form
- Computer- and data-savvy community
- Some standard formats in place
- Large data collections in funded, agency mandated archives
- Established culture of data sharing
- Community initiative driven by the needs of an exponential data growth
- Federal agency support/funding
- Data have no commercial value or privacy issues

Today, VO is the global data grid of astronomy, and is regarded as one of the success stories of the virtual scientific organizations and cyberinfrastructure

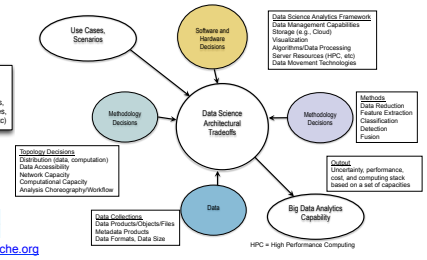
An Example of a Successful Methodology Transfer From Space Science to Medicine:

Using a state-of-the-art informatics infrastructure developed at JPL and leveraging successful efforts to build similar open source systems in space science open-source Object Oriented Data Technology (OODT) to design and an effective national integrated Knowledge System for the National Cancer Institute's Early Detection Research Network (EDRN).





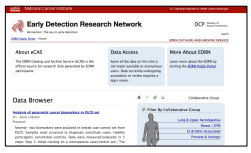
Designing a scalable architecture:



APACHE OODT
http://oodt.appache.org

LabCAS - Laboratory Catalog and Archive Service to Integrate the processing and data management capabilities to support the automated capture of the data directly from the labs into the knowledge environment.

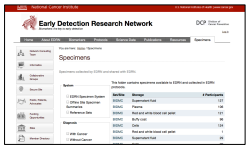
eCAS - EDRN Catalog and Archive Service to capture both structured and unstructured data using EDRN CDEs to catalog the metadata. Provenance information is also captured (approx. 40 data set and analysis information).



Biomarker Database- to capture and share EDRN biomarker annotations (approx. 900 encompass 11 organ and tissue sites. EDRN researchers have published over 230 papers on these biomarkers, with many more papers in preparation).

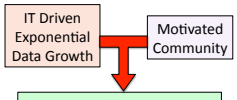
EDRN Virtual Specimen Repository – access to EDRN distributed specimen repository (approx. 200,000) .

EDRN Protocols – to share scientific protocols (approximately 200).



The EDRN Public Portal: A resource for collaborative science: http://cancer.gov/edrn

Conclusions and Recommendations:



Solution(s) Including:

- Data access / standards
- Legal and privacy issues
- Interoperable data grid
- Data analytics tools
- Trained user community

As the example above shows, data methodology transfer from astronomy and space science into HCI is clearly possible, at least within a given domain. However, the greater complexity and sociological issues will likely delay progress.

The Key Outstanding Challenges:

- The HC community does not yet have a well developed data culture and technical skills; acquiring them will take some time and resources
- The data and metadata must be understood, relevant, repeatable, with standard formats, properly curated, with interoperability protocols
- Diverse stakeholders may have colliding interests, yet find common goals
- Adequate, but not stifling privacy protection mechanisms are needed

A Recommended Path Forward:

- Promote multidisciplinary collaborations between communities of practice that have developed large scale open data environments (e.g., astronomy, space science, physics) and HCI
- Organize knowledge transfer activities through which best practices and lessons learned can be shared between research communities to help accelerate progress in HCI
- Work with funding agencies to facilitate multidisciplinary collaboration around large scale data analysis and infrastructure development
- Facilitate development of training programs in HCI that draw on expertise from other domains

This work was supported in part by a grant from the AcademyHealth Electronic Data Methods Forum, by the Center for Data-Driven Discovery at Caltech, and by the Center for Data Science and technology at JPL

Caltech CDD CENTER FOR DATA-DRIVEN DISCOVERY JPL CDST AcademyHealth