Informatics in Education and An Education in Informatics

Kirk Borne

Dept of Computational & Data Sciences

George Mason University

kborne@gmu.edu , http://classweb.gmu.edu/kborne/

General Themes in Informatics Research

- Information and knowledge processing, including natural language processing, information extraction, integration of data from heterogeneous sources or domains, event detection, feature recognition.
- Tools for analyzing and/or storing very large datasets, data supporting ongoing experiments, and other data used in scientific research.
- Knowledge representation, including vocabularies, ontologies, simulations, and virtual reality.
- Linkage of experimental and model results to benefit research.
- Innovative uses of information technology in science applications, including decision support, error reduction, outcomes analysis, and information at the point of end-use.
- Efficient management and utilization of information and data, including knowledge acquisition and management, process modeling, data mining, acquisition and dissemination, novel visual presentations, and stewardship of large-scale data repositories and archives.
- Human-machine interaction, including interface design, use and understanding of science discipline-specific information, information needs, and uses.
- High-performance computing and communications relating to scientific applications, including efficient machine-machine interfaces, transmission and storage, real-time decision support.
- Innovative uses of information technology to enhance learning, retention and understanding of science discipline-specific information.
- REFERENCE: http://grants.nih.gov/grants/guide/pa-files/PA-06-094.html

General Themes in Informatics Research

- Information and knowledge processing, including natural language processing, information extraction, integration of data from heterogeneous sources or domains, event detection, feature recognition.
- Tools for analyzing and/or storing very large datasets, data supporting ongoing experiments, and other data used in scientific research.
- Knowledge representation, including vocabularies, ontologies, simulations, and virtual reality.
- Linkage of experimental and model results to benefit research.
- Innovative uses of information technology in science applications, including decision support, error reduction, outcomes analysis, and information at the point of end-use.
- Efficient management and utilization of information and data, including knowledge acquisition and management, process modeling, data mining, acquisition and dissemination, novel visual presentations, and stewardship of large-scale data repositories and archives.
- **Human-machine interaction, including interface design, use and understanding of science discipline-specific information, information needs, and uses.**
- High-performance computing and communications relating to scientific applications, including efficient machine-machine interfaces, transmission and storage, real-time decision support.
- **Innovative uses of information technology to enhance learning, retention and understanding of science discipline-specific information.**
- REFERENCE: http://grants.nih.gov/grants/guide/pa-files/PA-06-094.html

Outline

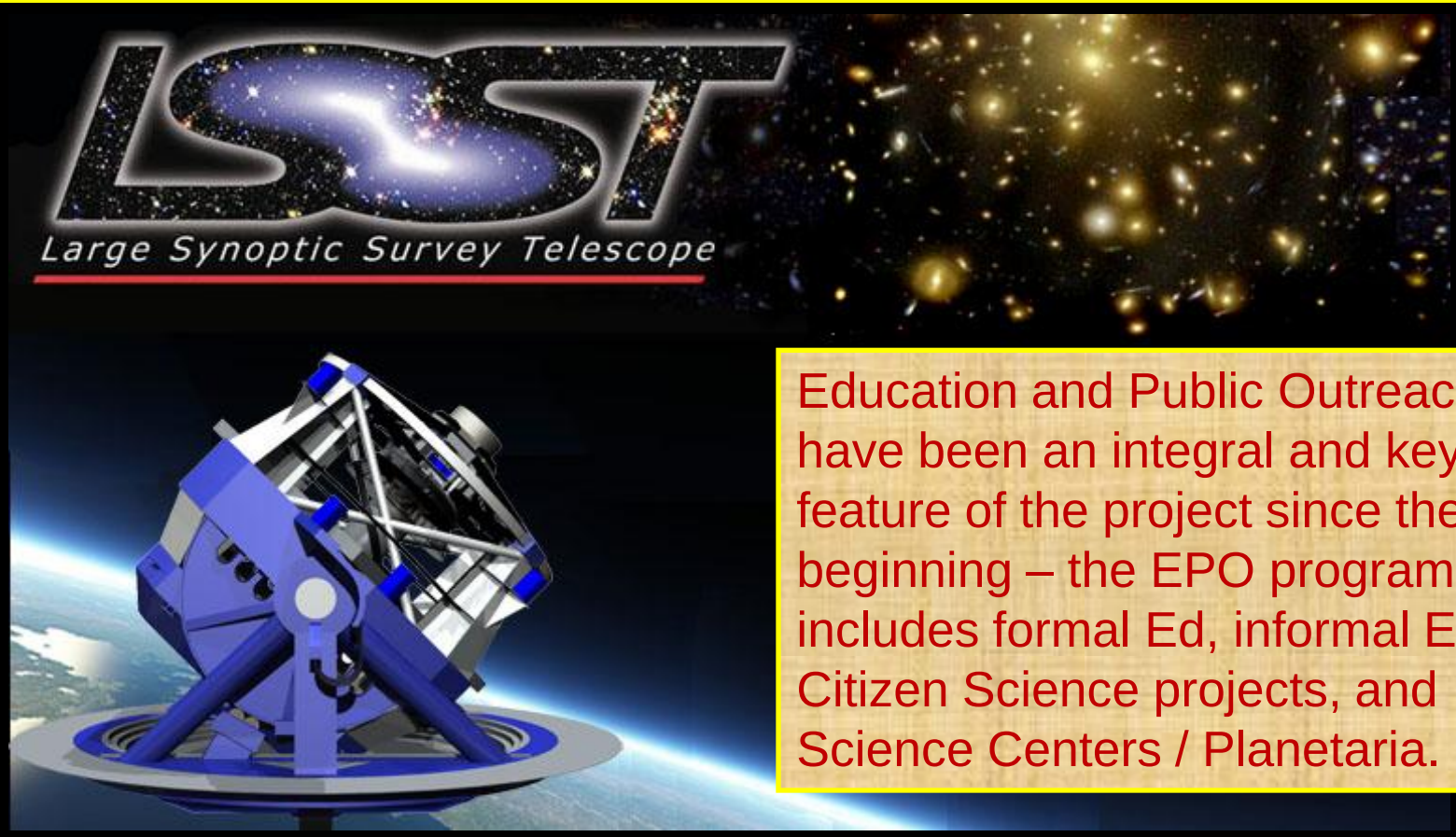
- Informal Education Example: Citizen Science
- Formal Education Example: Undergrad programs

Outline

- Informal Education Example: Citizen Science
- Formal Education Example: Undergrad programs

Observing Strategy: One pair of images every 40 seconds for each spot on the sky, then continue across the sky continuously every night for 10 years (2016-2026), with time domain sampling in log(time) intervals (to capture dynamic range of transients).

- **LSST (Large Synoptic Survey Telescope):**
  - Ten-year time series imaging of the night sky – mapping the Universe !
  - **100,000 events each night** – *anything that goes bump in the night !*
  - ***Cosmic Cinematography! The New Sky! @ http://www.lsst.org/***



Education and Public Outreach have been an integral and key feature of the project since the beginning – the EPO program includes formal Ed, informal Ed, Citizen Science projects, and Science Centers / Planetaria.

Citizen Science

- Exploits the cognitive abilities of **Human Computation**!
- Novel mode of data collection:
  - Citizen Science! = Volunteer Science = Participatory Science
  - *e.g.,* VGI = Volunteer Geographic Information (Goodchild '07)
  - *e.g.,* Galaxy Zoo @ http://www.galaxyzoo.org/
- Citizen science refers to the involvement of volunteer non-professionals in the research enterprise.
- The Citizen Science experience …
  - must be engaging,
  - must work with real scientific data/information,
  - must not be busy-work,
  - **must address authentic science research questions** that are beyond the capacity of science teams and enterprises, and
  - must involve the scientists.

Modes of Computing

- **Numerical Computation** *(in silico)*
  - Model-dependent, subjective, only as good as your best hypothesis
  - Fast, efficient
  - Processing power is rapidly increasing
- **Computational Intelligence**
  - Data-driven, objective (machine learning)
  - Often relies on human-generated training data
  - Often generated by a single investigator
  - Primitive algorithms
  - Not as good as humans on most tasks
- **Human Computation** *(Carbon-based Computing)*
  - Data-driven, objective (human cognition)
  - Creates training sets, Cross-checks machine results
  - Excellent at finding patterns, image classification
  - Capable of classifying anomalies that machines don't understand
  - Slow at numerical processing, low bandwidth, easily distracted

It takes a human to interpret a complex image

Examples of Citizen Science

- AAVSO (Amer. Assoc. of Variable Star Observers)
- Audubon Bird Counts
- Project Budburst
- Stardust@Home
- VGI (Volunteer Geographic Information)
- CoCoRaHS (Community Collaborative Rain, Hail and Snow network)
- Galaxy Zoo **(~20 refereed pubs so far…)**
- Light Curve Zoo (coming soon from the LSST project)
- Zooniverse (buffet of Zoos)

The Zooniverse

http://zooniverse.org/

- New funded NSF CDI grant (PI: L.Fortson, Adler Planetarium; J. Wallin, MTSU; K.Borne, GMU; & Chris Lintott, Oxford U)
- Building a framework for new Citizen Science projects, including user-based research tools
- Science domains:
  - Astronomy (Galaxy Merger Zoo)
  - The Moon (Lunar Reconnaissance Orbiter)
  - The Sun (STEREO dual spacecraft)
  - Egyptology (the Papyri Project)
  - and more (… accepting proposals from community)

Outline

- Informal Education Example: Citizen Science
- Formal Education Example: Undergrad programs

Dept of Computational & Data Sciences

@ GMU (George Mason University)

http://cds.gmu.edu/

Informatics-based Science Education

- Informatics enables transparent reuse and analysis of scientific data in inquiry-based classroom learning (http://serc.carleton.edu/usingdata/).
- Students are trained:
  - to access large distributed data repositories
  - to conduct meaningful scientific inquiries into the data
  - to mine and analyze the data
  - to make data-driven scientific discoveries
- The 21st century workforce demands training and skills in these areas, as all agencies, businesses, and disciplines are becoming flooded with data.
- Numerous Data Sciences programs now starting at several universities (GMU, Caltech, RPI, Michigan, Cornell, …).
- CODATA **ADMIRE** initiative: ***A****dvanced* ***D****ata* ***M****ethods and* ***I****nformation technologies for* ***R****esearch and* ***E****ducation*

Computational & Data Sciences at Mason: CUPIDS

- CUPIDS = *Curriculum for an Undergraduate Program in Data Sciences*
- NSF-funded program through CCLI (Course, Curriculum, and Laboratory Improvement)
- Starting year: 2008
- Primary Goal:
  - *to increase student's understanding of the role that data plays across the sciences as well as to increase the student's ability to use the technologies associated with data acquisition, mining, analysis, and visualization.*

http://cds.gmu.edu/

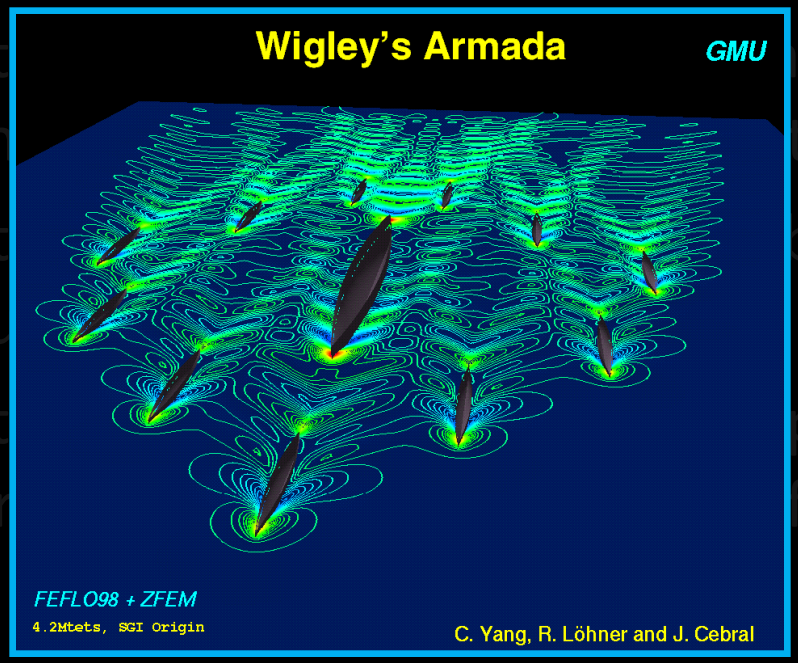
Objectives of Mason's CUPIDS project

1. To teach students what Data Science is and how it is changing the way science is being done across the disciplines
2. To change st… improve their confidence in … scientific data problem…
3. To increase s… alization for generating a… tions
4. To increase s… bases for scientific inquiry
5. To increase student's abilities to acquire, process, and explore experimental data with the use of a computer

http://cds.gmu.edu/

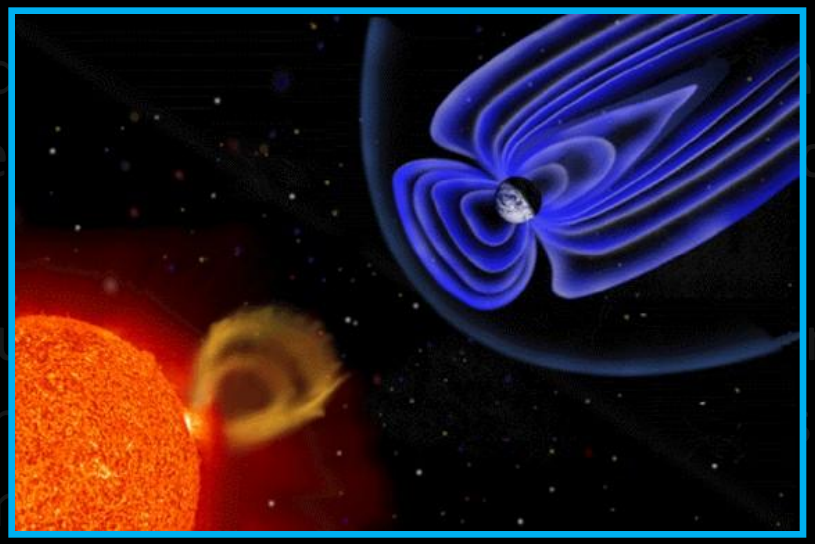
Objectives of Mason's CUPIDS project

1. To teach students what Data Science is and how it is changing the way science is being done across the disciplines
2. To change student's attitudes about and improve their confidence in using computers to address scientific data problems
3. To increase st lization for generating ar tions
4. To increase st bases for scientific inqu
5. To increase st process, and explore exper f a computer

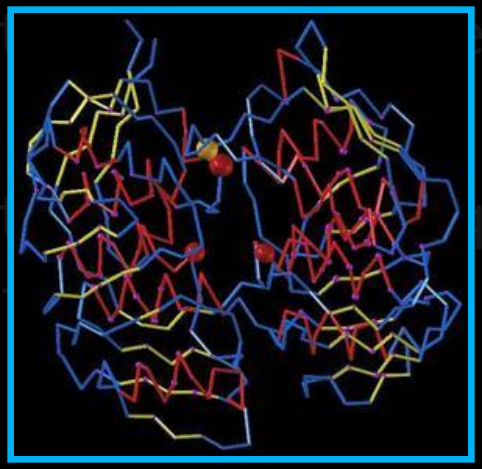


http://cds.gmu.edu/

Objectives of Mason's CUPIDS project

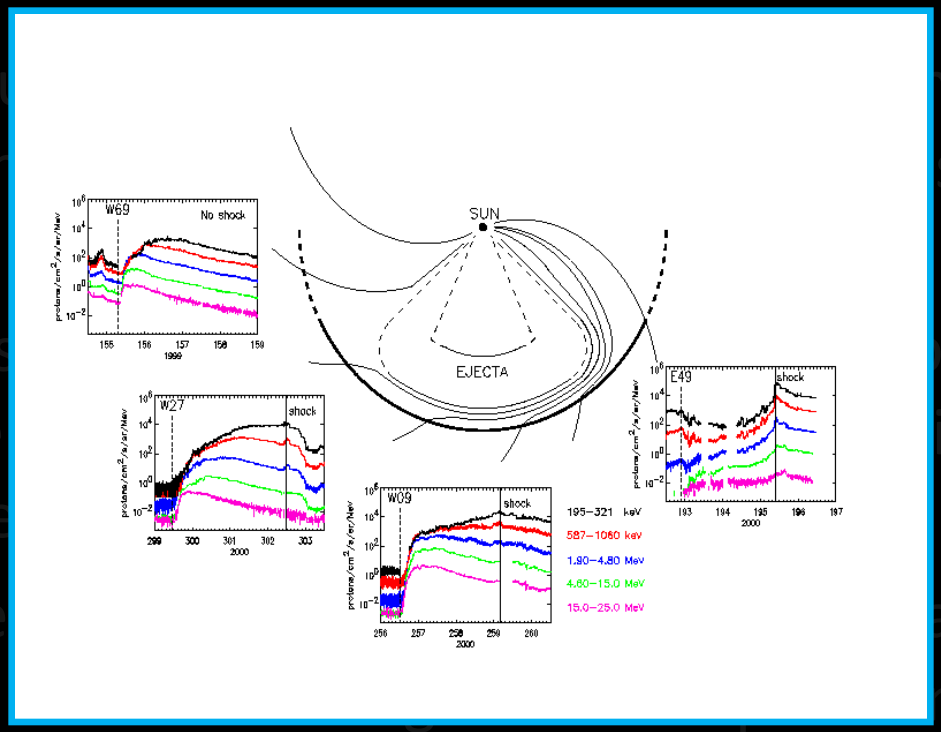


1. To teach stud d how it is changing the cross the disciplines
2. To change stu mprove their confidence in scientific data problem
3. To increase student's abilities to use visualization for generating and addressing scientific questions
4. To increase studen databases for scientific inquiry
5. To increase studen quire, process, and explore experimen use of a computer



http://cds.gmu.edu/

Objectives of Mason's CUPIDS project

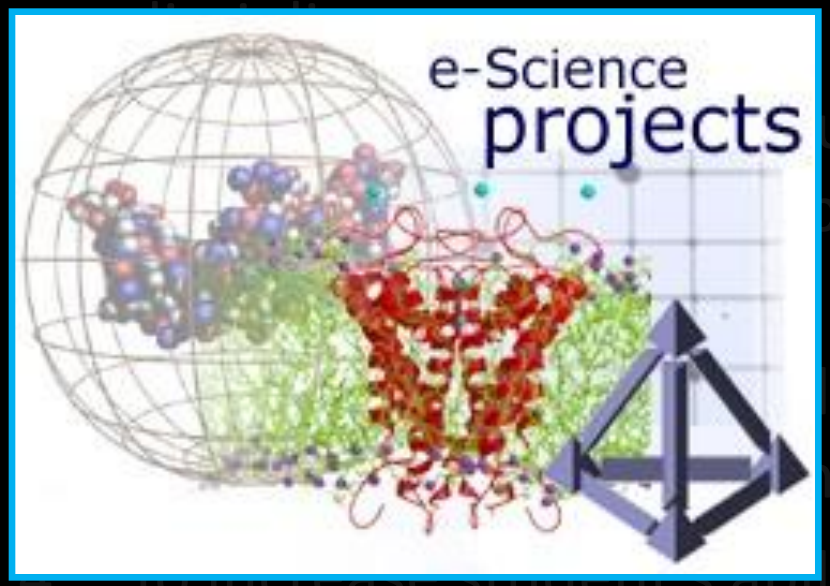


4. To increase student's abilities to use databases for scientific inquiry

http://cds.gmu.edu/

Objectives of Mason's CUPIDS project

1. To teach students what Data Science is and how it is changing the way science is being done across the



scientific inquiry

5. To increase student's abilities to acquire, process, and explore experimental data with the use of a computer

http://cds.gmu.edu/

What is the CDS Degree Program?

http://cds.gmu.edu/

- This is a *SCIENCE* degree (**not** a Computer Science degree) with a strong computational and data-oriented approach to science:
  - The degree is Computational and Data Sciences
- Started in Spring 2008 with CDS major (for a B.S. degree in the Mason College of Science), and then added the CDS minor
- Each student must declare a concentration:
  - Physics, Chemistry, or Biology
  - Other choices may be added in the future: astronomy, geosciences, materials science

CDS Core Courses (required)

- *CDS 101 – Introduction to Computational & Data Sciences*
- *CDS 301 – Scientific Information and Data Visualization*
- *CDS 302 – Scientific Data and Databases*
- *CDS 401 – Scientific Data Mining*
- *CDS 410 – Modeling and Simulation I*
- *CDS 411 – Modeling and Simulation II*

New Courses (not required)

- *CDS 130 – Computing for Scientists*
- *CDS 151 – Data Ethics in an Information Society*
- *CDS 351 – Introduction to Scientific Programming*

http://cds.gmu.edu/

Similar programs elsewhere

- Cornell data-driven science DISCOVER program
- RPI Web Science program
- Caltech CACR and e-Science programs
- U. Washington e-Science Institute
- U. Michigan astronomy+medicine+CS program
- POCA (Partnership in Observational and Computational Astronomy) at SCSU & Clemson
- Purdue's Discovery Informatics program
- College of Charleston Discovery Informatics program
- Vanderbilt Initiative in Data-intensive Astrophysics (VIDA)
- University of Sydney Astroinformatics

Data Science Education paper available !

State of the Profession position paper, submitted to the Astro2010 Decadal Survey
3/15/2009

http://mason.gmu.edu/~kborne/Borne_data_sciences_education_CDH_EPO.pdf
http://www8.nationalacademies.org/astro2010/publicview.aspx

The Revolution in Astronomy Education

Data Science for the Masses

Authorship: This Position Paper was prepared and endorsed by the following team of astronomers, educators, and information scientists. The lead authors are Kirk D. Borne (Dept. of Computational and Data Sciences, George Mason University, kborne@gmu.edu) and Suzanne Jacoby (LSST Education and Public Outreach, sjacoby@lsst.org).

Astroinformatics 2010 Conference

- http://www.astro.caltech.edu/ai10/
- Participants discussed the following question (in an informatics context**):
  - *What are the top three things you think we need to teach the next generation of scientists?***
- It was then decided that a community of interest (COI) should form to discuss and establish an education plan, curriculum changes, courses, best practices, etc. – perhaps as a virtual institute for space and earth science informatics education (VISESIE).

Astroinformatics 2010 Conference

- http://www.astro.caltech.edu/ai10/
- Participants discussed the following question (in an informatics context**):
  - *What are the top three things you think we need to teach the next generation of scientists?***
- It was then decided that a community of interest (COI) should form to discuss and establish an education plan, curriculum changes, courses, best practices, etc. – perhaps **the CODATA ADMIRE initiative (Advanced Data Methods and Information technologies for Research and Education).**

Who wants to be involved?

- Kirk Borne
- Your name here

Related References

- Borne (2009): *"**U-Science**",* http://essi.gsfc.nasa.gov/pdf/Borne2.pdf
- Borne, Jacoby, et al. (2009): *"**The Revolution in Astronomy Education: Data Science for the Masses**",* http://arxiv.org/abs/0909.3895
- Borne (2009): *"**Astroinformatics: A 21st Century Approach to Astronomy**",* http://arxiv.org/abs/0909.3892
- Borne (2010): *"**Astroinformatics: Data-Oriented Astronomy Research and Education**", Journal of Earth Science Informatics, vol. 3, pp. 5-17.*
- M. F. Goodchild (2007): *"**Citizens as Sensors: the World of Volunteered Geography**",* GeoJournal, 69, pp. 211-221.
- Lintott et al. (2009): *"**Galaxy Zoo: 'Hanny's Voorwerp', a quasar light echo?**",* http://arxiv.org/abs/0906.5304
- Raddick et al. (2009): *"**Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers**",* http://arxiv.org/abs/0909.2925
- Raddick, Bracey, Carney, Gyuk, Borne, Wallin, & Jacoby (2009): *"**Citizen Science: Status and Research Directions for the Coming Decade**",* http://www8.nationalacademies.org/astro2010/DetailFileDisplay.aspx?id=454
- Baehr, Vedachalam, Borne, & Sponseller (2010): *"Data Mining the Galaxy Zoo Mergers",* (NASA CIDU 2010 conference)