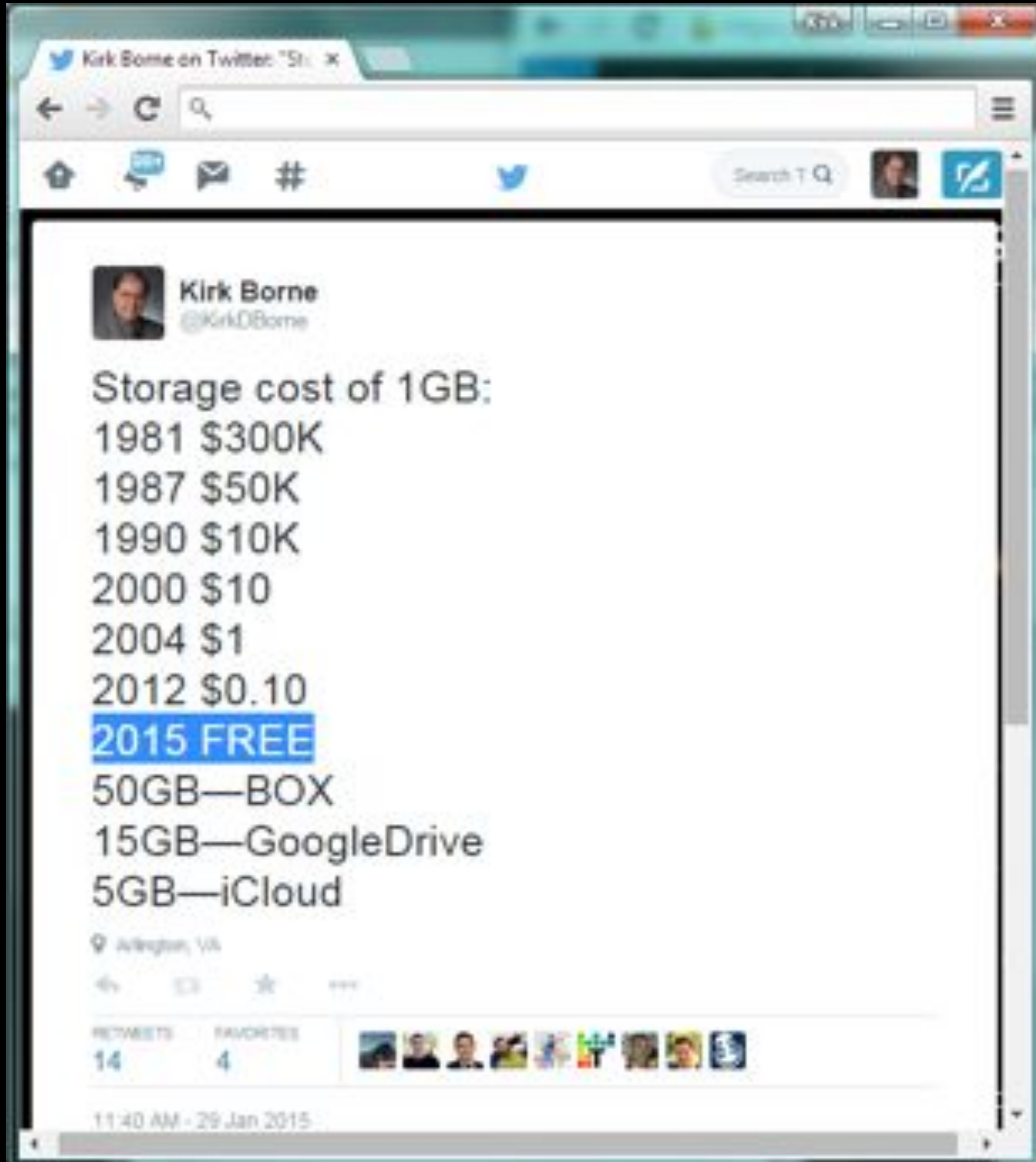
Data Literacy For All: Astrophysics and Beyond

(Astronomy is evidence-based forensic science, thus it is a data & information science)

Kirk Borne

George Mason University, Fairfax, VA • www.kirkborne.net @KirkDBorne

NSF Workshop: Towards Big Steps Enabled by “Big Data Science” January 29-30, 2015


Kirk Borne
@KirkDBorne
Storage cost of 1GB:
1981 $300K
1987 $50K
1990 $10K
2000 $10
2004 $1
2012 $0.10
2015 FREE
50GB—BOX
15GB—GoogleDrive
5GB—iCloud
14
4
11:40 AM - 29 Jan 2015

Data Science Programs at Mason

- http://spacs.gmu.edu/content/academic-programs
- **CSI = Computational Science & Informatics**
  - Graduate program at Mason since 1992 (centroid shifting to Data Science)
  - Over 200 PhD's graduated in past 20 years (~90 enrolled now)
- **CDS = Computational and Data Sciences**
  - Undergraduate program at Mason since 2007
  - Recently morphed from BS Major program into a Minor
  - Developed with the support of a grant (2006-2008) from the NSF DUE:

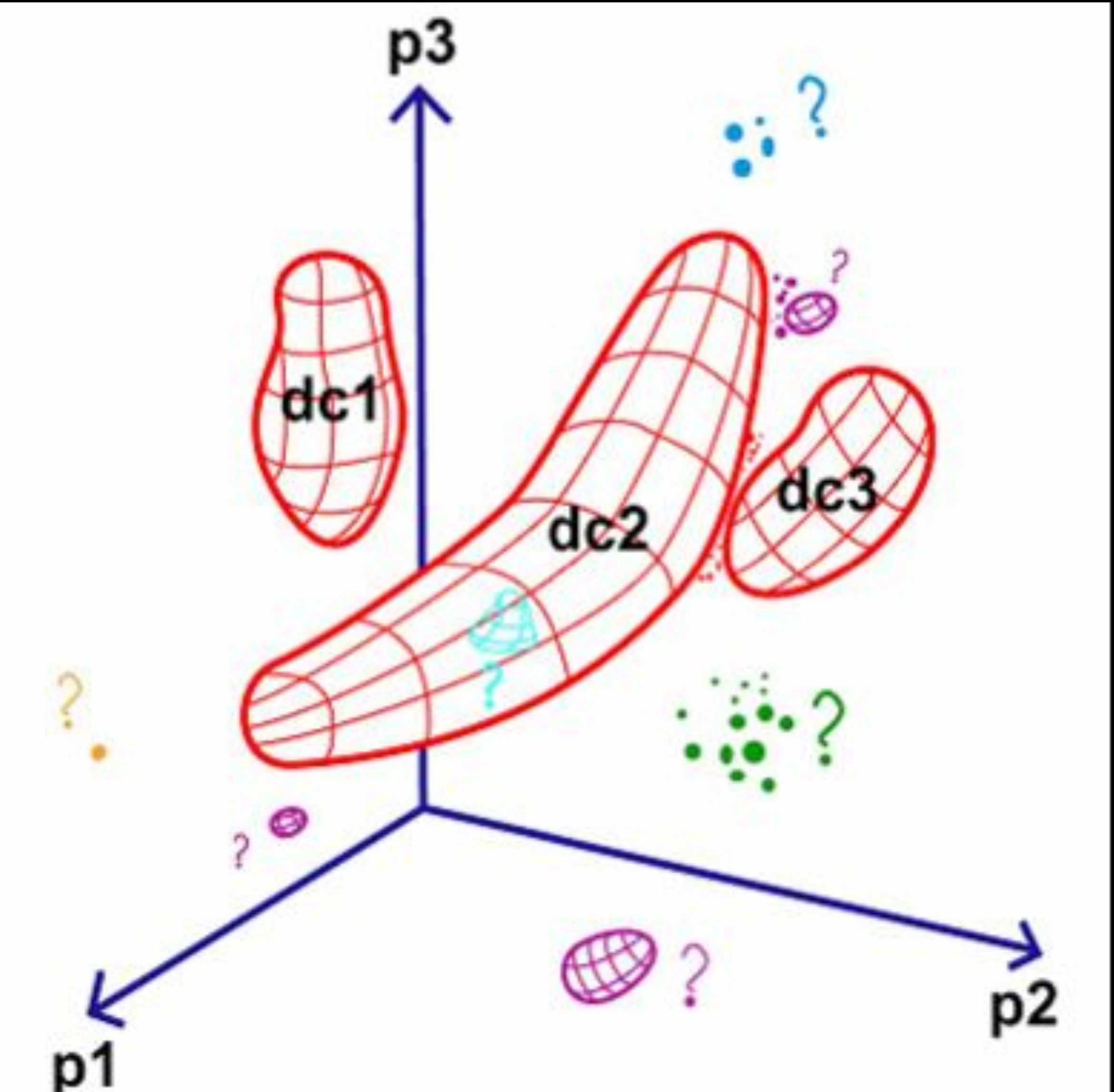
    CUPIDS = ***C****urriculum for an* ***U****ndergraduate* ***P****rogram* ***I****n* ***D****ata* ***S****ciences*
  - Primary Goals: *to increase students' skills in the use of data & increase their understanding of the role of data across the sciences and beyond.*
  - Objectives – students are trained:
  - … to access large distributed data repositories (with attention to Data Ethics)
  - … to conduct meaningful inquiries into the data (“ “ “ “ “)
  - … to mine, visualize, and analyze the data
  - … to make objective data-driven inferences, discoveries, and decisions

Data Science = 4 Types of Discovery

(Learning from Data)

1) Correlation Discovery
   - Finding patterns and dependencies, which reveal new natural laws or new scientific principles
2) Novelty Discovery
   - Finding new, rare, one-in-a-[million / billion / trillion] objects and events
3) Class Discovery
   - Finding new classes of objects and behaviors
   - Learning the rules that constrain class boundaries
4) Association Discovery
   - Finding unusual (improbable) co-occurring associations

This graph says it all …
3 Steps to Discovery – Learning from Data



Graphic provided by Professor S. G. Djorgovski, Caltech

- Unsupervised Learning : Cluster Analysis – partition the data items into clusters, without bias, ignoring any initially assigned categories = **Class Discovery !**

- Supervised Learning : Classification – for each new data item, assign it to a known class (*i.e.,* a known category or cluster) = **Predictive Power Discovery !**

- Semi-supervised Learning : Outlier/Novelty Detection – identify data items that are outside the bounds of the known classes of behavior = **Surprise Discovery !**

Goal of Data Science:

Take Data to Information to Knowledge to Insights (and Action!)

- ✓ From Sensors (Measurement & Data Collection)…
- ✓ … to Sentinels (Monitoring & Alerts) …
- ✓ … to Sense-making (Data Science) …
- ✓ … to Cents-making (Business ROI)

… *Actionizing and Productizing Big Data*

Astronomy Big Data Example

Large Synoptic Survey Telescope

The LSST (Large Synoptic Survey Telescope)

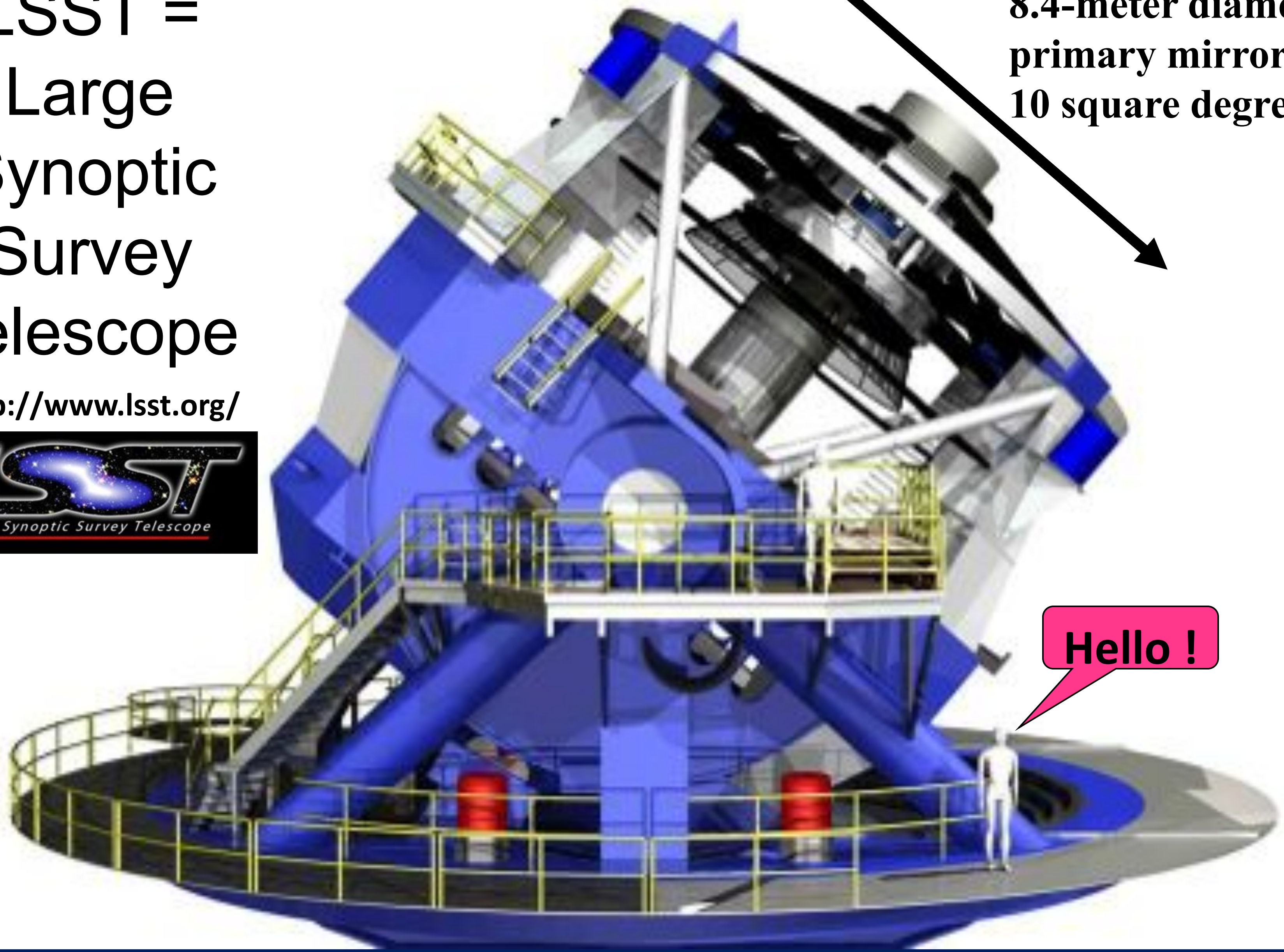
LSST = Large Synoptic Survey Telescope

http://www.lsst.org/

Large Synoptic Survey Telescope

(mirror funded by private donors)

8.4-meter diameter primary mirror = 10 square degrees!

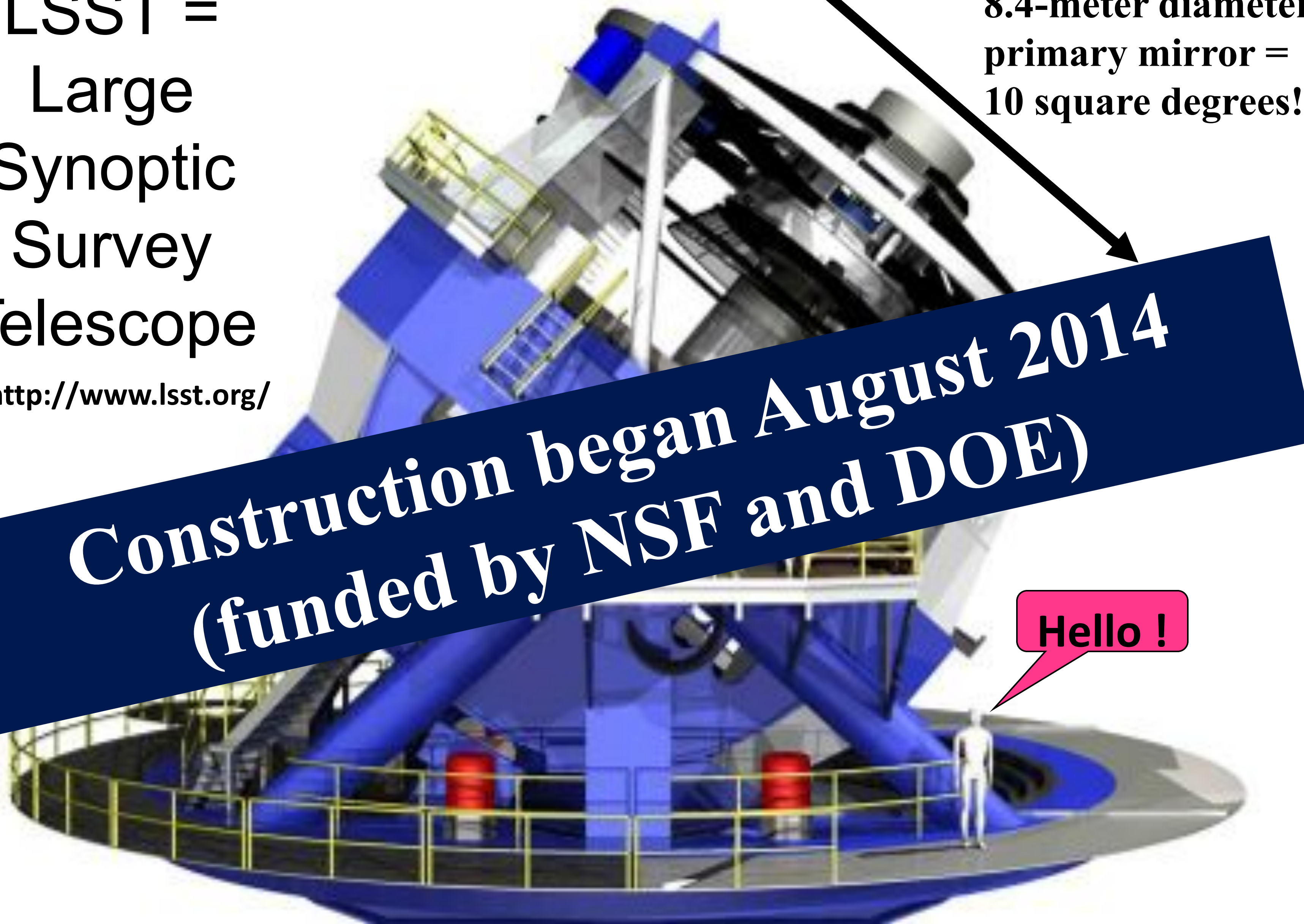


LSST =
Large
Synoptic
Survey
Telescope

http://www.lsst.org/

(mirror funded by private donors)

8.4-meter diameter primary mirror = 10 square degrees!



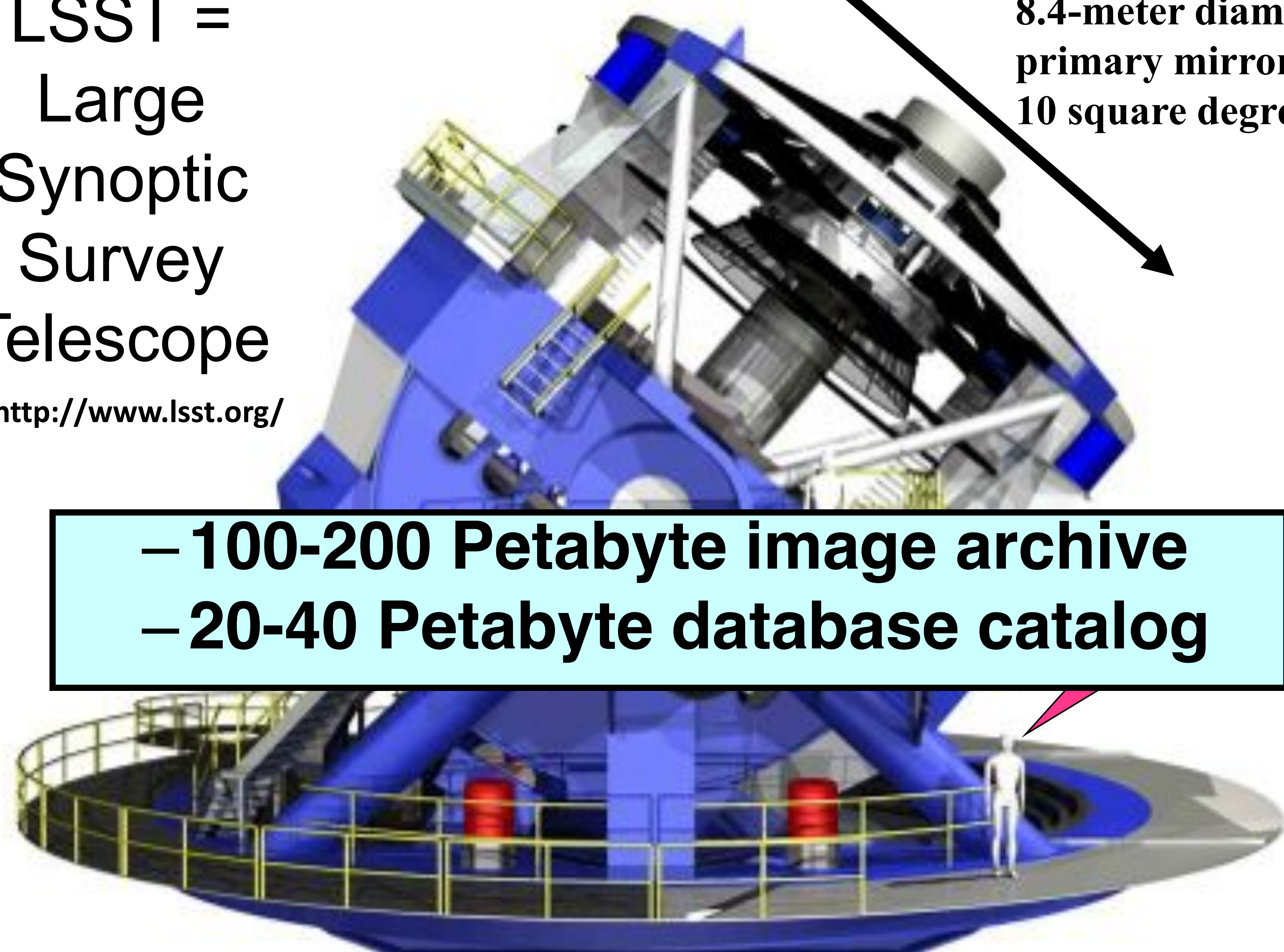
Construction began August 2014 (funded by NSF and DOE)

LSST =
Large
Synoptic
Survey
Telescope

http://www.lsst.org/

(mirror funded by private donors)

8.4-meter diameter primary mirror = 10 square degrees!



- **100-200 Petabyte image archive**
- **20-40 Petabyte database catalog**

LSST Key Science Drivers: Mapping the Dynamic Universe

- Complete inventory of the Solar System (Near-Earth Objects; killer asteroids???)
- Nature of Dark Energy (Cosmology; Supernovae at edge of the known Universe)
- Optical transients (10 million daily event notifications sent within 60 seconds)
- Digital Milky Way (Dark Matter; Locations and velocities of 20 billion stars!)

<u>LSST in time and space:</u>

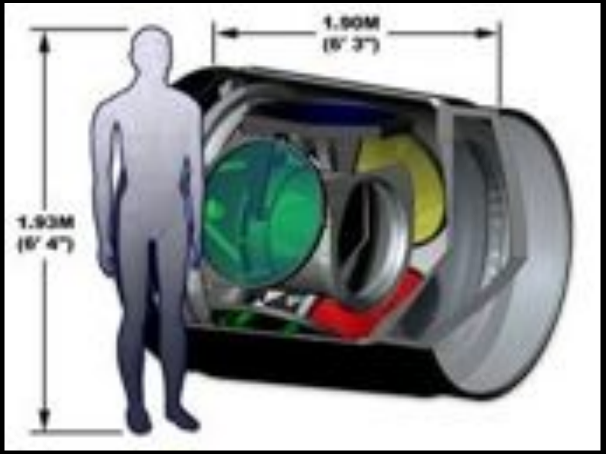
- When? ~2022-2032
- Where? Cerro Pachon, Chile


Architect's design of LSST Observatory

LSST Summary

http://www.lsst.org/

Large Synoptic Survey Telescope



- 3-Gigapixel camera
- One 6-Gigabyte image every 20 seconds
- 30 Terabytes every night for 10 years
- Repeat images of the entire night sky every 3 nights: ***Celestial Cinematography***
- 100-Petabyte final image data archive anticipated – **all data are public!!!**
- **20-Petabyte final database catalog anticipated**
- Real-Time Event Mining: ~10 million events per night, every night, for 10 years!
  - Follow-up observations required to classify these
  - Which ones should we follow up? …

… Decisions! Decisions! Data-to-Decisions!

Astronomy Data Science Organizations

- LSST Informatics & Statistics Science Collaboration
- AAS Working Group on Astroinformatics and Astrostatistics
- ASA interest group in Astrostatistics
- IAU Working Group on Astrostatistics and Astroinformatics
- International Astrostatistics and Astroinformatics (IAA) Professional Society (associated with ISI: International Statistical Institute)

http://arxiv.org/abs/1301.3069

Top Topics and Challenges for LSST ISSC

- Machine Learning and Statistics (Data Science) algorithm development
- Collaboration-building across disciplines
- Validated Training Sets needed for commissioning (e.g., for classifying multiple types of alerts)
- Inputs to LSST project on cadence and other
- Catalog-based data analysis tools (IDL, Python,…)
  - LSST is SDSS! (== that's a **FACTORIAL**!)
- Citizen Science (Crowdsourced Big Data Tasks)
- Finding Funding

The BIG Big Data Challenge:

Identifying, characterizing, & responding to millions of events in real-time streaming data

- Astronomy example:
  - ❖ Real-Time Event Mining: deciding which events (out of millions) need follow-up investigation & response (triage for maximum scientific return)
- Web Analytics example:
  - ❖ Web Behavior Modeling and Automated System Response (from online interactions & web browse patterns, personalization, user segmentation, 1-to-1 marketing, advanced analytics discovery,…)
- Many other examples:
  - ❖ Health alerts (from EHRs and national health systems)
  - ❖ Tsunami alerts (from geo sensors everywhere)
  - ❖ Cybersecurity alerts (from network logs)
  - ❖ Social event alerts or early warnings (from social media)
  - ❖ Preventive Fraud alerts (from financial applications)
  - ❖ Predictive Maintenance alerts (from machine / engine sensors)

Enter… Advanced Big Data Analytics!

- *Learning from Data (Data Science):*
  - Outlier / Anomaly / Novelty / Surprise detection
  - Clustering (= New Class discovery, Segmentation)
  - Correlation & Association discovery
  - Classification, Diagnosis, Prediction
- *… for the 3 D2D challenges:*
  - *Data-to-Discoveries*
  - *Data-to-Decisions*
  - *Data-to-Dividends*

    (big ROI = Return on Innovation)

The MIPS model

for **Dynamic Data-Driven Application Systems (DDDAS)**

http://dddas.org

- **MIPS** =
  - **M**easurement – **I**nference – **P**rediction – **S**teering
- This applies to any Network of Sensors:
  - Web user interactions & actions (web analytics data), Cyber network usage logs, Social network sentiment, Machine logs (of any kind), Manufacturing sensors, Health & Epidemic monitoring systems, Financial transactions, National Security, Utilities and Energy, Remote Sensing, Tsunami warnings, Weather/Climate events, Astronomical sky events, …
- Machine Learning enables the "IP" part of MIPS:
  - Autonomous (or semi-autonomous) Classification
  - Intelligent Data Understanding
  - Rule-based
  - Model-based
  - Neural Networks
  - Markov Models
  - Bayes Inference Engines

Alert & Response systems:
- LSST 10million events
- Automation of any data-driven operational system

The MIPS model

for **Dynamic Data-Driven Application Systems (DDDAS)**

http://dddas.org

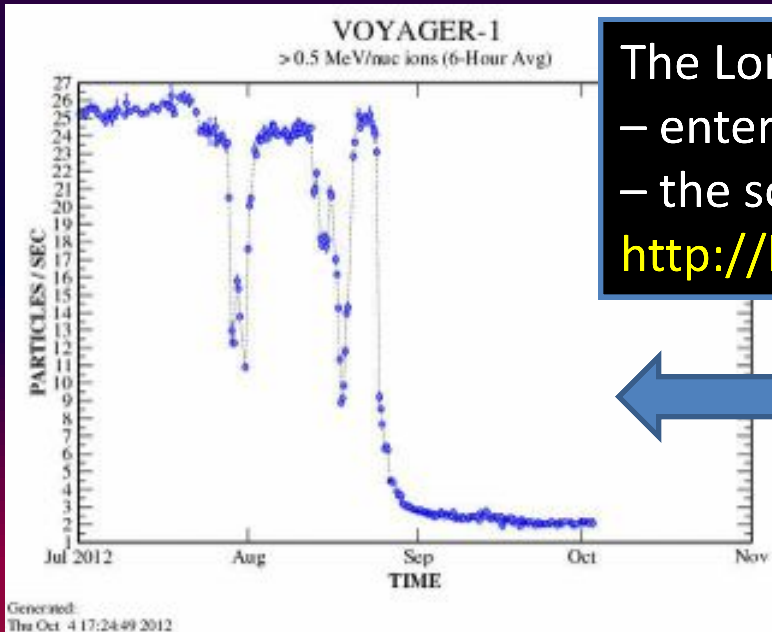
- **MIPS** =
  - **M**easurement – **I**nference – **P**rediction – **S**te
- This applies to any Network of Sen
  - Web user interactions & actions (web an network usage logs, Social network sentim (of any kind), Manufacturing sensors, He onitoring systems, Financial transactions, Nation and Energy, Remote Sensing, Tsunami war nate events, Astronomical sky events, …
- Mac g enables the “IP” part of MIPS:
  - r semi-autonomous) Classification
  - nt Data Understanding
  - ule-based
  - Model-based
  - Neural Networks
  - Markov Models
  - Bayes Inference Engines

From Sensors to Sentinels to Sense

Alert & Response systems:
- LSST 10million events
- Automation of any data-driven operational system

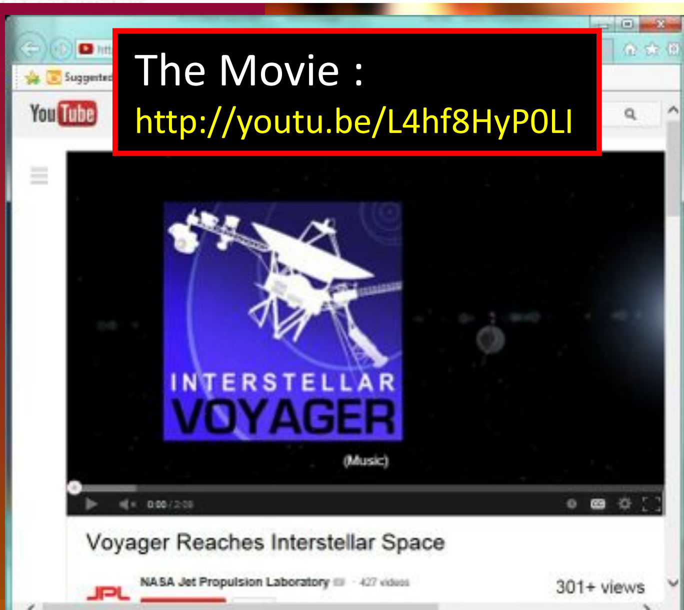
Voyager 1 becomes first human-made object to leave solar system

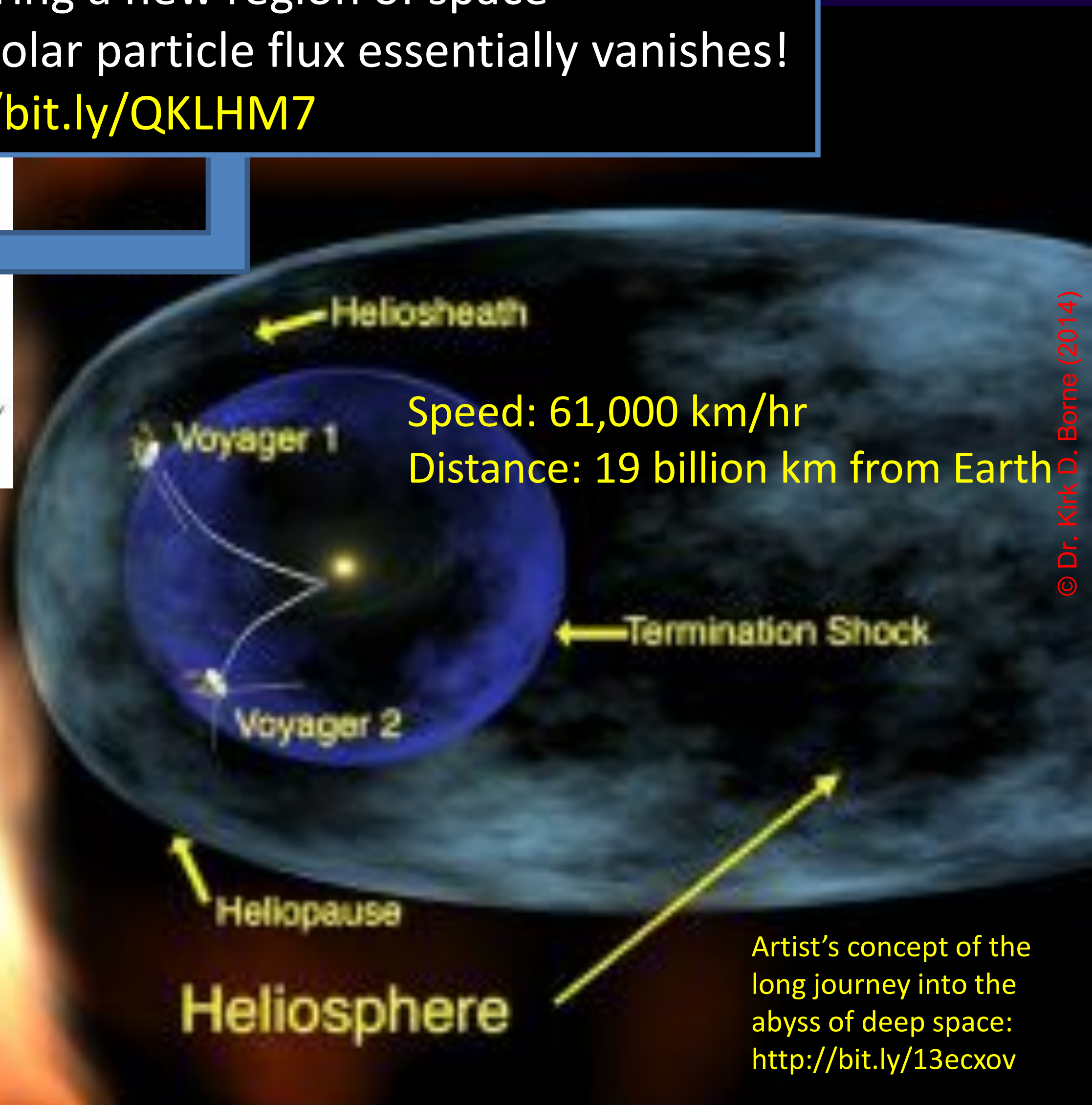
http://www.nasa.gov/mission_pages/voyager/



The Long Tail of the Data tells the story –
– entering a new region of space –
– the solar particle flux essentially vanishes!
http://bit.ly/QKLHM7

The Movie :
http://youtu.be/L4hf8HyP0LI



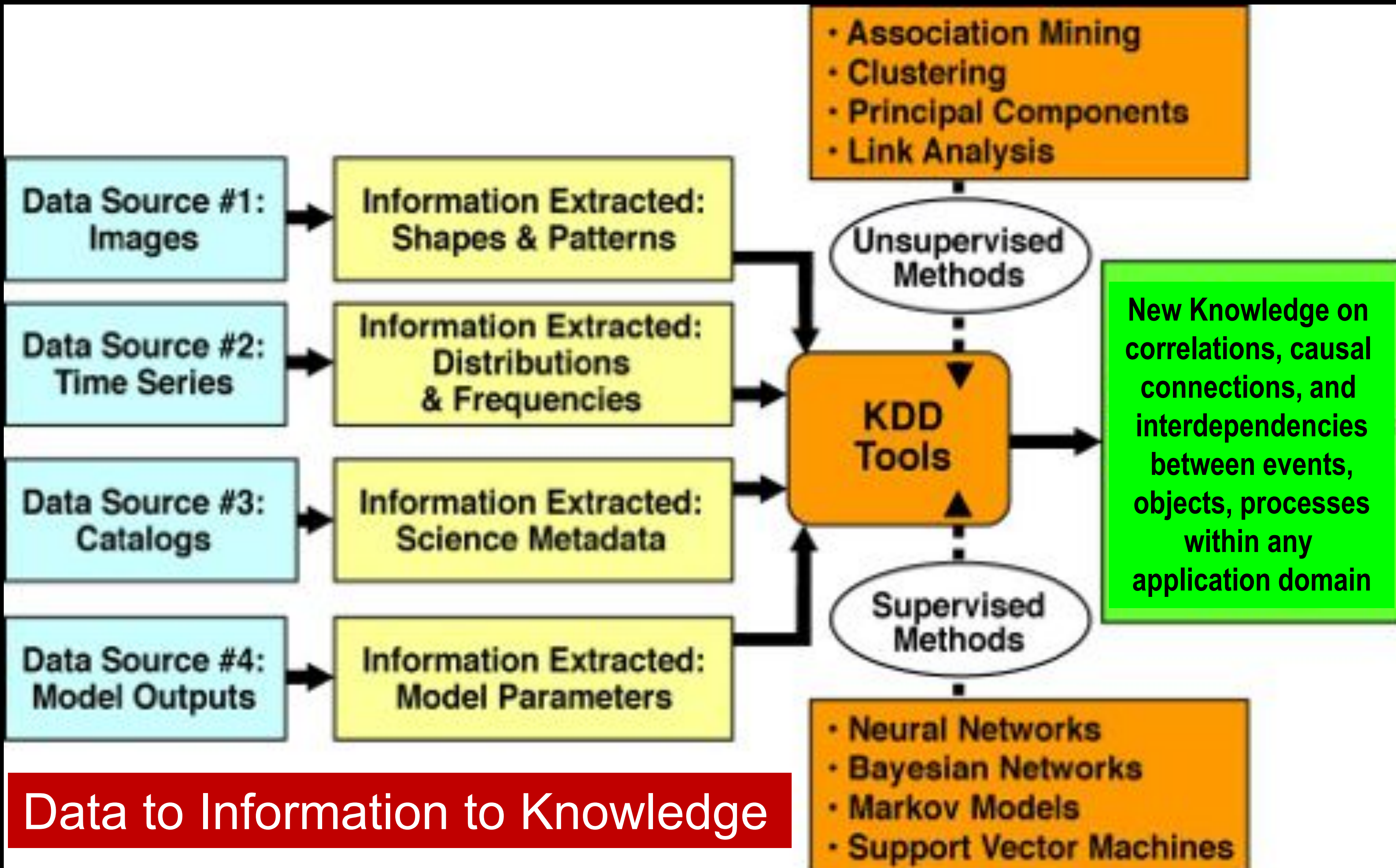


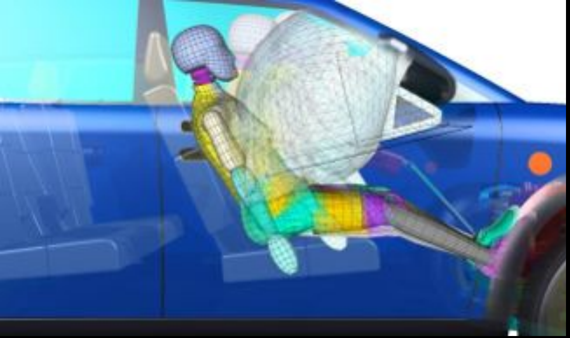
Speed: 61,000 km/hr
Distance: 19 billion km from Earth

Artist's concept of the long journey into the abyss of deep space: http://bit.ly/13ecxov

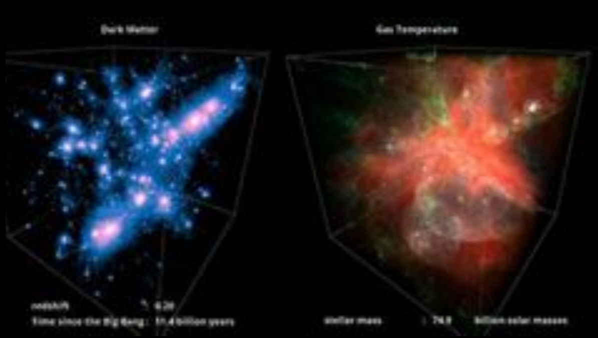
Knowledge Discovery for multi-source Data:

Heterogeneous data collections are the new normal





Big Data Science meets HPC's Big Simulations



- Cosmology (colliding galaxies: crash science)
- Fusion Science
- Climate Science
- Vehicle Safety (colliding cars: crash science)
- Digital Manufacturing
- Aircraft, Ship, and Automotive Design
- Multiphysics, Turbulence, Energy systems, etc. …

Characterize, measure, and track massive data outputs for: deviations, anomalies, emergent behavior & patterns, "events", signals of changes in system stationarity,…

- Enabling Discovery and Data-Driven Decision-making

Meeting the 3 D2D Challenges**

1. Characterize and Contextualize first.
2. Collect and Curate each entity's features.

…then Come to the data-driven decision!

**

- Data-to-Discoveries
- Data-to-Decisions
- Data-to-Dividends

Characterization

Feature & Context Detection and Extraction:

- Identify and characterize features in the data:
  - Machine-generated
  - Human-generated
  - Crowdsourced? (**Citizen Science** = **Tap the Power of Human Cognition to find patterns and anomalies in massive data!**)
- Extract the context of the data: the source, the channel, the data user, the use cases, the value, the re-uses … where, when, who, how, what, why = *Metadata!*
- Curate these features for search, re-use, and **D2D!**
  - Include other parameters and features from other data sources and databases – integrate all information to help characterize & contextualize (and ultimately make decision regarding) each new event.

Contextualization

Feature & Context Detection and Extraction:

- Identify and characterize features in the data:
  - Machine-generated
  - Human-generated
  - Crowdsourced? (Citizen Science = **Tap the Power of Human Cognition to find patterns and anomalies in massive data!**)
- Extract the **context** of the data: the source, the channel, the data user, the use cases, the value, the re-uses … where, when, who, how, what, why = *Metadata!*
- Curate these features for search, re-use, and **D2D!**
  - Include other parameters and features from other data sources and databases – integrate all information to help characterize & contextualize (and ultimately make decision regarding) each new event.

Collection & Curation

Feature & Context Detection and Extraction:

- Identify and characterize features in the data:
  - Machine-generated
  - Human-generated
  - Crowdsourced? (Citizen Science = **Tap the Power of Human Cognition to find patterns and anomalies in massive data!**)
- Extract the context of the data: the source, the channel, the data user, the use cases, the value, the re-uses … where, when, who, how, what, why = *Metadata!*
- Curate these features for search, re-use, and **D2D!**
  - Include other parameters and features from other data sources and databases – integrate all information to help characterize & contextualize (and ultimately make decision regarding) each new event.

Key Feature of Zooniverse:

Data Mining the volunteer-contributed characterizations

- Train the automated pipeline classifiers with:
  - Improved classification algorithms
  - Better identification of anomalies
  - Fewer classification errors
- Millions of training examples
- Hundreds of millions of class labels (tags)

ZOONIVERSE
REAL SCIENCE ONLINE

Advancing Science through User-Guided Learning in Massive Data Streams

Tags produce a new data flood

- Tagging enables semantic data fusion and integration, for knowledge acquisition / representation / sharing
- User-contributed content adds **more** data to the data flood.
- Tagging is applicable to any data source, including document repositories – adding lightweight semantics to the data repository (taxonomies, folksonomies, annotations)
- Tagging improves data discovery, search and retrieval, and knowledge management

Data Science – putting it all together:
(the whole is greater than sum of the parts)

Data Science is TRANSDISCIPLINARY Science!

It is the collection of mathematical, computational, scientific, and domain-specific methods, tools, and algorithms that are applied to Big Data for discovery, decision support, and data-to-knowledge transformation:

- Advanced Database / Data Management & Data Structures
- Data Mining (Machine Learning) & Analytics (KDD)
- Statistics and Statistical Programming
- Data & Information Visualization
- Semantics (Natural Language Processing, Ontologies)
- Everything is a graph (Network Analysis and Graph Mining)
- Data-intensive Computing (e.g., Hadoop, Cloud, …)
- Modeling & Simulation (computational data science)
- Metadata for Indexing, Search, & Retrieval
- Domain-Specific Data Analysis Tools

Profile of a Big-Data-Enabled Specialist

generated by "Oceans 11" panel of experts convened by the Oceans of Data Institute (August 2014)

http://oceansofdata.org/our-work/profile-big-data-enabled-specialist

Panel

Kirk Borne
Professor of Astrophysics and Computational Science
George Mason University
Fairfax, Virginia

Randy Bucciarelli
Programmer/Analyst
Scripps Institution of Oceanography
UC San Diego
La Jolla, California

Tim Chadwick
Principal Engineer
Dynamic Network Services, Inc.
Manchester, New Hampshire

Benjamin Davison
Quantitative User Experience Researcher
Google
Boston, Massachusetts

Lucy Drotning
Associate Provost of Planning and Institutional Research
Columbia University
New York, New York

Ryan Kapaun
Law Enforcement Analyst
Eden Prairie Police Department
Eden Prairie, Minnesota

Juan Miguel Lavista Ferres
Principal Data Scientist
Bing/Microsoft
Seattle, Washington

Shannon McWeeney
Head of Division of Bioinformatics and Computational Biology
Oregon Health & Science University
Portland, Oregon

Jay Parker
Earth Scientist
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Steve Ross
Consultant on Data Quality Control
Corporate Editor
Broadband Communities Magazine
Revere, Massachusetts

Kartik Shah
Principal Consultant
Strategic Solutions
Toronto, Canada

Oceans of Data Institute

Ruth Krumhansl
Director

Profile Facilitators

Joseph Ippolito

Joyce Malyn-Smith

Suggested Citation:
Oceans of Data Institute. (2014). *Profile of a big-data-enabled specialist.* Waltham, MA: Education Development Center, Inc.

Profile of a Big-Data-Enabled Specialist

EDC OCEANS of DATA INSTITUTE

http://oceansofdata.org | Email: oceansofdata@edc.org