LSST: Informatics and Statistics Research Challenges

Kirk Borne
Dept of Computational & Data Sciences
George Mason University
kborne@gmu.edu , http://classweb.gmu.edu/kborne/

Outline

- Prelude
- Astroinformatics
- Example Application: The LSST Project
- Informatics & Statistics Challenge Problems
- Challenge Area: Distributed Data Mining
- Summary

Outline

- Prelude
- Astroinformatics
- Example Application: The LSST Project
- Informatics & Statistics Challenge Problems
- Challenge Area: Distributed Data Mining
- Summary

Prelude

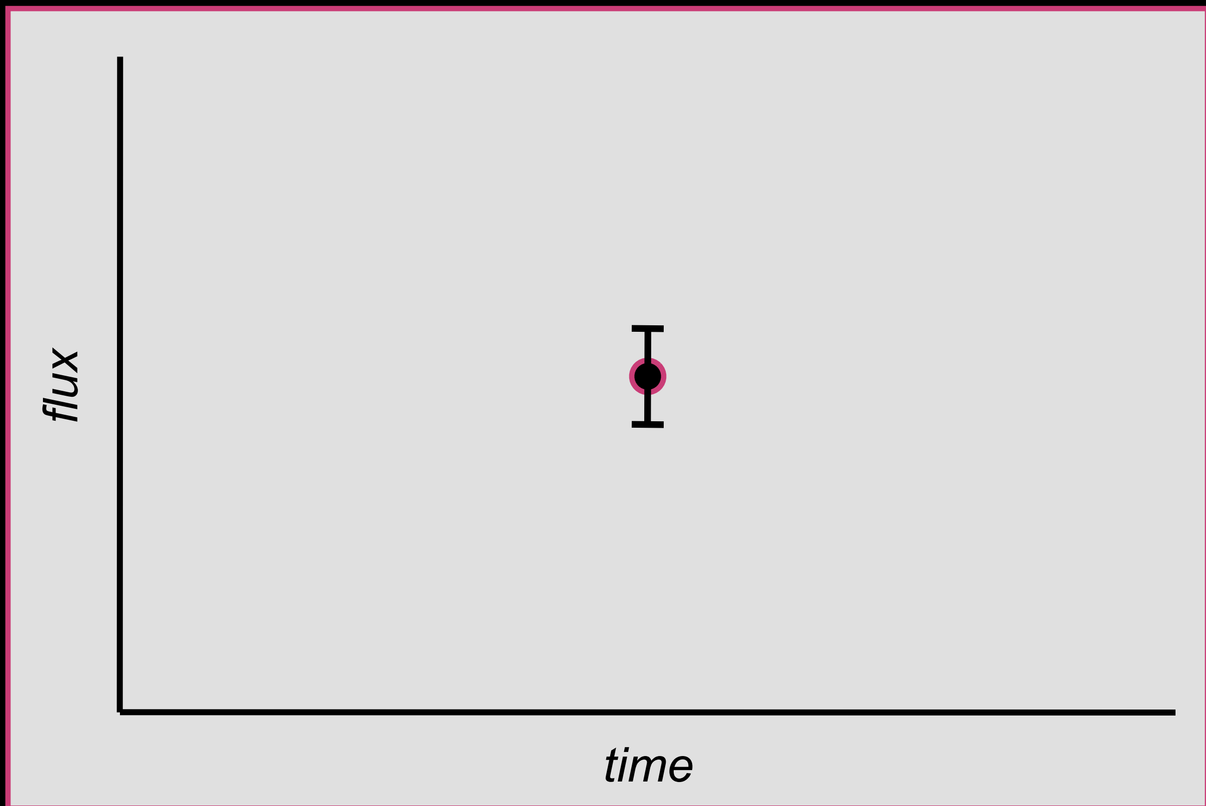
- LSST Challenge #1
- LSST Challenge #2
- The Data Flood
- Data-Enabled Science

LSST challenge #1

- Approximately 100,000 times each night for 10 years LSST will obtain the following data on a new sky event, and we will be challenged with classifying these data:

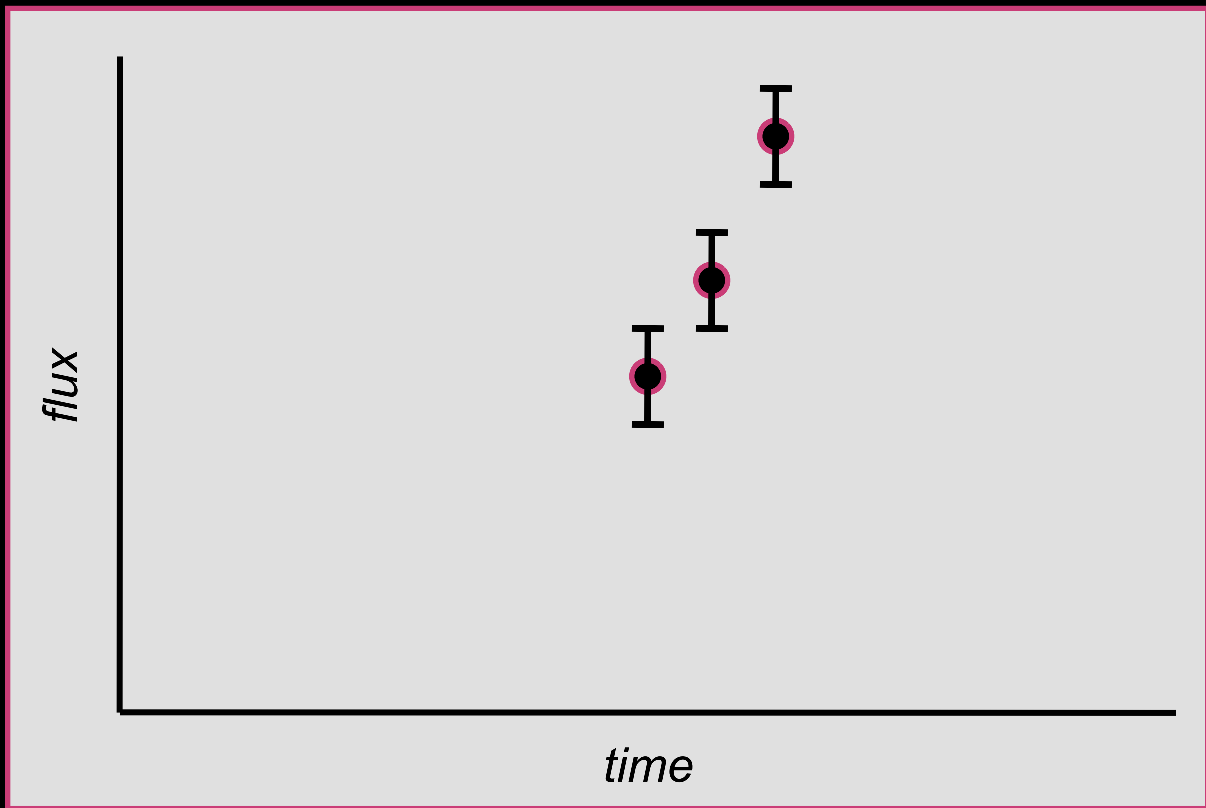
LSST challenge #1

- Approximately 100,000 times each night for 10 years LSST will obtain the following data on a new sky event, and we will be challenged with classifying these data:



LSST challenge #1

- Approximately 100,000 times each night for 10 years LSST will obtain the following data on a new sky event, and we will be challenged with classifying these data: *more data points help !*



LSST challenge #1

- Approximately 100,000 times each night for 10 years LSST will obtain the following data on a new sky event, and we will be challenged with classifying these data: *more data points help !*

Characterize first !
then Classify.

LSST challenge #2

- **Each night** for 10 years LSST will obtain the equivalent amount of data that was obtained by the entire Sloan Digital Sky Survey
- My grad students will be asked to mine these data (~20 TB each night ≈ 40,000 CDs filled with data):

LSST challenge #2

- **Each night** for 10 years LSST will obtain the equivalent amount of data that was obtained by the entire Sloan Digital Sky Survey
- My grad students will be asked to mine these data (~20 TB each night ≈ 40,000 CDs filled with data): *a sea of CDs*

Image: The CD Sea in Kilmington England (600,000 CDs)

LSST challenge #2

- **Each night** for 10 years LSST will obtain the equivalent amount of data that was obtained by the entire Sloan Digital Sky Survey
- My grad students will be asked to mine these data (~20 TB each night ≈ 40,000 CDs filled with data):
  - *A sea of CDs each and every day for 10 yrs*
  - *Cumulatively, a football stadium full of 200 million CDs after 10 yrs*

Responding to the Data Flood

- Big Data is a national challenge and a national priority … see August 9, 2010 announcement from OMB and OSTP @ http://www.aip.org/fyi (#87)
- More data is not just more data… ***more is different !***
- Several national study groups have issued reports on the urgency of establishing scientific and educational programs to face the data flood challenges.
- Each of these reports has issued a call to action in response to the data avalanche in science, engineering, and the global scholarly environment.

Data Sciences: A National Imperative

1. National Academies report: *Bits of Power: Issues in Global Access to Scientific Data,* (1997) downloaded from http://www.nap.edu/catalog.php?record_id=5504
2. NSF (National Science Foundation) report: *Knowledge Lost in Information: Research Directions for Digital Libraries,* (2003) downloaded from http://www.sis.pitt.edu/~dlwkshop/report.pdf
3. NSF report: *Cyberinfrastructure for Environmental Research and Education,* (2003) downloaded from http://www.ncar.ucar.edu/cyber/cyberreport.pdf
4. NSB (National Science Board) report: *Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century,* (2005) downloaded from http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf
5. NSF report with the Computing Research Association: *Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda,* (2005) downloaded from http://www.cra.org/reports/cyberinfrastructure.pdf
6. NSF Atkins Report: *Revolutionizing Science & Engineering Through Cyberinfrastructure: Report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure,* (2005) downloaded from http://www.nsf.gov/od/oci/reports/atkins.pdf
7. NSF report: *The Role of Academic Libraries in the Digital Data Universe,* (2006) downloaded from http://www.arl.org/bm~doc/digdatarpt.pdf
8. National Research Council, National Academies Press report: *Learning to Think Spatially,* (2006) downloaded from http://www.nap.edu/catalog.php?record_id=11019
9. NSF report: *Cyberinfrastructure Vision for 21st Century Discovery,* (2007) downloaded from http://www.nsf.gov/od/oci/ci_v5.pdf
10. JISC/NSF Workshop report on Data-Driven Science & Repositories, (2007) downloaded from http://www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf
11. DOE report: *Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale,* (2007) downloaded from http://www.sc.doe.gov/ascr/ProgramDocuments/Docs/DOE-Visualization-Report-2007.pdf
12. DOE report: *Mathematics for Analysis of Petascale Data Workshop Report,* (2008) downloaded from http://www.sc.doe.gov/ascr/ProgramDocuments/Docs/PetascaleDataWorkshopReport.pdf
13. NSTC Interagency Working Group on Digital Data report: *Harnessing the Power of Digital Data for Science and Society,* (2009) downloaded from http://www.nitrd.gov/about/Harnessing_Power_Web.pdf
14. National Academies report: *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age,* (2009) downloaded from http://www.nap.edu/catalog.php?record_id=12615

Recent (March 2010) NSF working group: Data-Enabled Science (DES)

- DES group prepared a white paper of related challenges and recommendations to inform the NSF MPSAC (Mathematical & Physical Sciences directorate Advisory Committee).
- MPSAC may use the white paper as a source of ideas and information to advise NSF in DES areas.
- DES committee: 2 scientists each from Astronomy, Physics, Chemistry, Mathematics, Statistics, and Materials Science.

Some of the members of the NSF DES Working Group

- Astronomy: Robert Hanisch, Kirk Borne
- Statistics: James Berger, Alan Karr
- Mathematical Sciences: David Keyes, …
- Physics: Patrick Brady (LIGO as eventful astronomy), Harrison Prosper (LHC)
- Chemistry: Brooks Pate (Interstellar chemistry), …
- Materials Science: Sharon Glotzer**, …

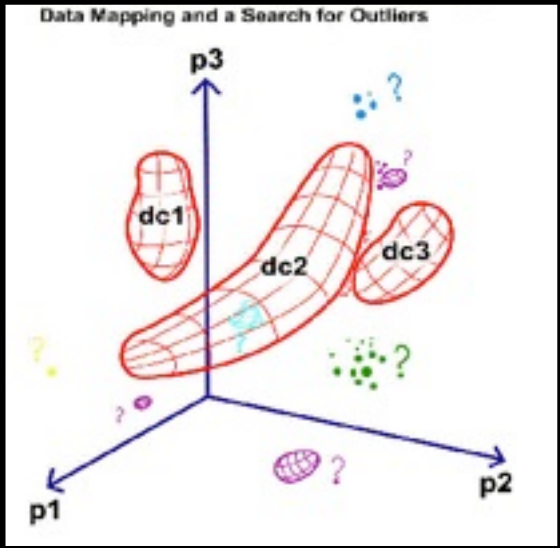
** Chair of the committee that authored the 400-page “Glotzer Report” (2009) on data-intensive science, specifically simulation-based science & engineering. Reference: http://www.wtec.org/sbes/

Examples of Recommendations: Scientific Inference with Massive or Complex Data

- Advances in fundamental mathematics and statistics are needed to provide the language, structure, and tools for many needed methodologies of data-enabled scientific inference.
  - Example 1: Exploitation of *sparsity* (e.g., out of a huge list of proteins, only an unknown few may be active in a particular metabolic process)
    - often hidden, discovered only with new mathematics involving harmonic analysis, approximation theory, numerical analysis and statistical theory;
    - led to compressed sensing.
  - Example 2: Machine learning in massive data sets
    - of late, an explosion in the utilization of nonparametric Bayes techniques
- Algorithmic advances in handling massive and complex data are crucial.
- Visualization (visual analytics) and citizen science (human computation or data processing) will play key roles.

Data-Enabled Science:
Scientific KDD (Knowledge Discovery from Data)

- Characterize the known (clustering, unsupervised learning)
- Assign the new (classification, supervised learning)
- Discover the unknown (outlier detection, semi-supervised learning)

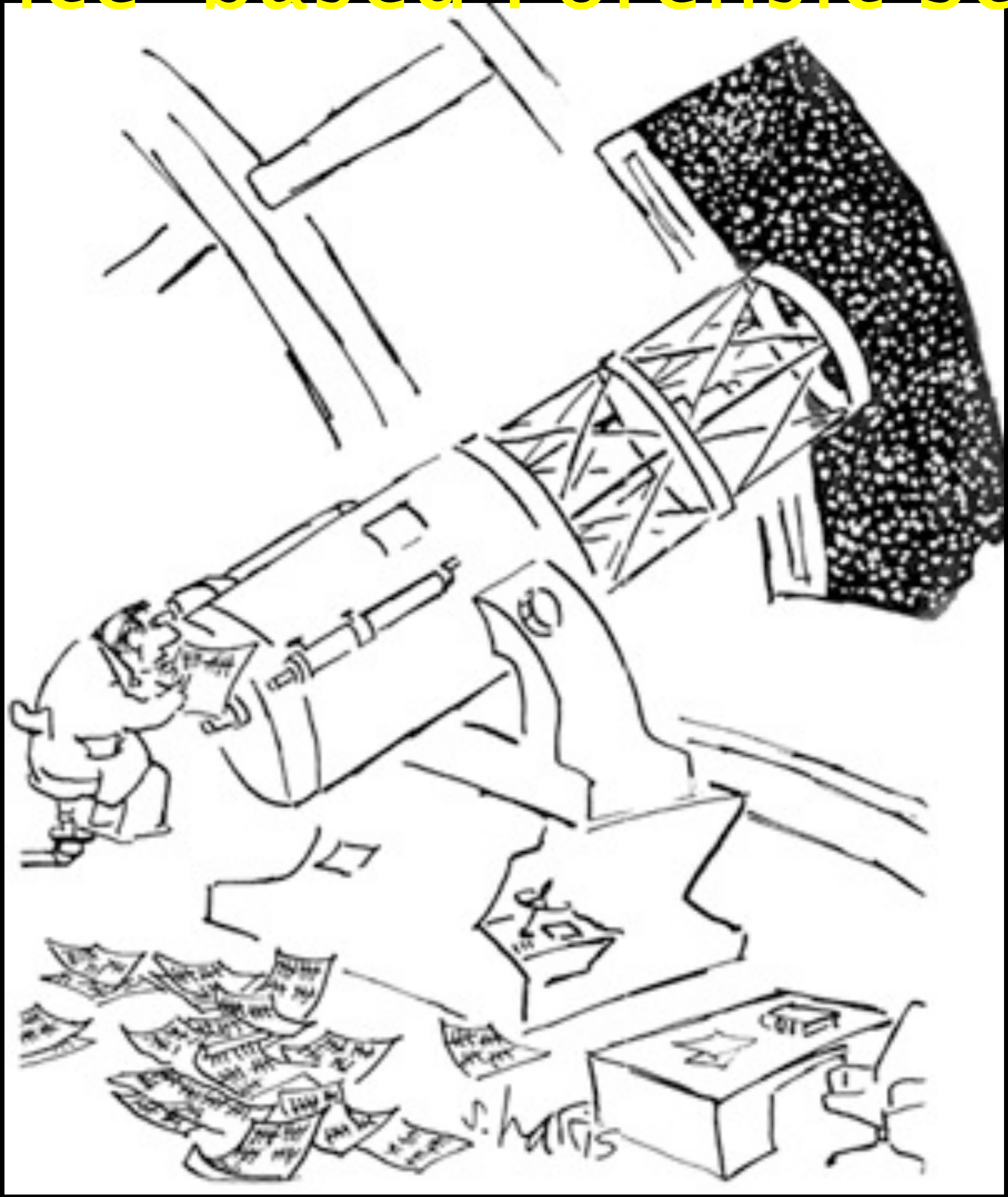


- **Benefits of very large datasets:**
  - **best statistical analysis of "typical" events**
  - **automated search for "rare" events**

Outline

- Prelude
- Astroinformatics
- Example Application: The LSST Project
- Informatics & Statistics Challenge Problems
- Challenge Area: Distributed Data Mining
- Summary

Astronomy: Data-Driven Science = Evidence-based Forensic Science



The Changing Landscape of Astronomical Research

- **Past:** 100's to 1000's of independent distributed heterogeneous data / metadata / information repositories.
- **Today:** Astronomical data are now accessible uniformly from federated distributed heterogeneous sources = **the Virtual Observatory**.
- **Future:** Astronomy is and will become even more data-intensive in the coming decade with the growth of massive data-producing sky surveys.
- **Challenge:** It will be prohibitively difficult to transport the data to the user application. Therefore … ***SHIP THE CODE TO THE DATA !***

From Data-Driven to Data-

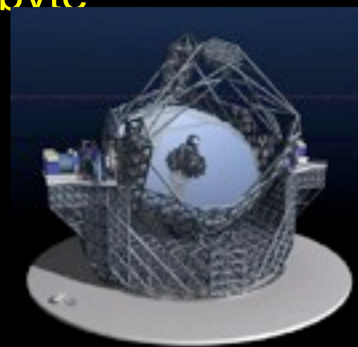
- Astronomy has always been a data-driven science
- It is now a data-intensive science: **Astroinformatics** !
  - Data-oriented Astronomical Research = "the 4th Paradigm"
  - Scientific KDD (Knowledge Discovery in Databases):
    - Characterize the known (clustering, unsupervised learning)
    - Assign the new (classification, supervised learning)
    - Discover the unknown (outlier detection, semi-supervised learning)
    - …

Scientific Knowledge !

- Benefits of very large datasets:
  - best statistical analysis of "typical" events
  - automated search for "rare" events

Astronomy Data Environment : Sky Surveys

- To avoid biases caused by limited samples, astronomers now study the sky systematically = **Sky Surveys**
- Surveys are used to measure and collect data from all objects that are contained in large regions of the sky, in a systematic, controlled, repeatable fashion.
- These surveys include (... this is just a subset):
  - MACHO and related surveys for dark matter objects: ~ 1 Terabyte
  - Digitized Palomar Sky Survey: 3 Terabytes
  - 2MASS (2-Micron All-Sky Survey): 10 Terabytes
  - GALEX (ultraviolet all-sky survey): 30 Terabytes
  - Sloan Digital Sky Survey (1/4 of the sky): 40 Terabytes
  - and this one is just starting: Pan-STARRS: 40 **Peta**bytes!



- **Leading up to the big survey next decade:**
  - **LSST (Large Synoptic Survey Telescope):** 100 Petabytes!

Outline

- Prelude
- Astroinformatics
- Example Application: The LSST Project
- Informatics & Statistics Challenge Problems
- Challenge Area: Distributed Data Mining
- Summary

LSST

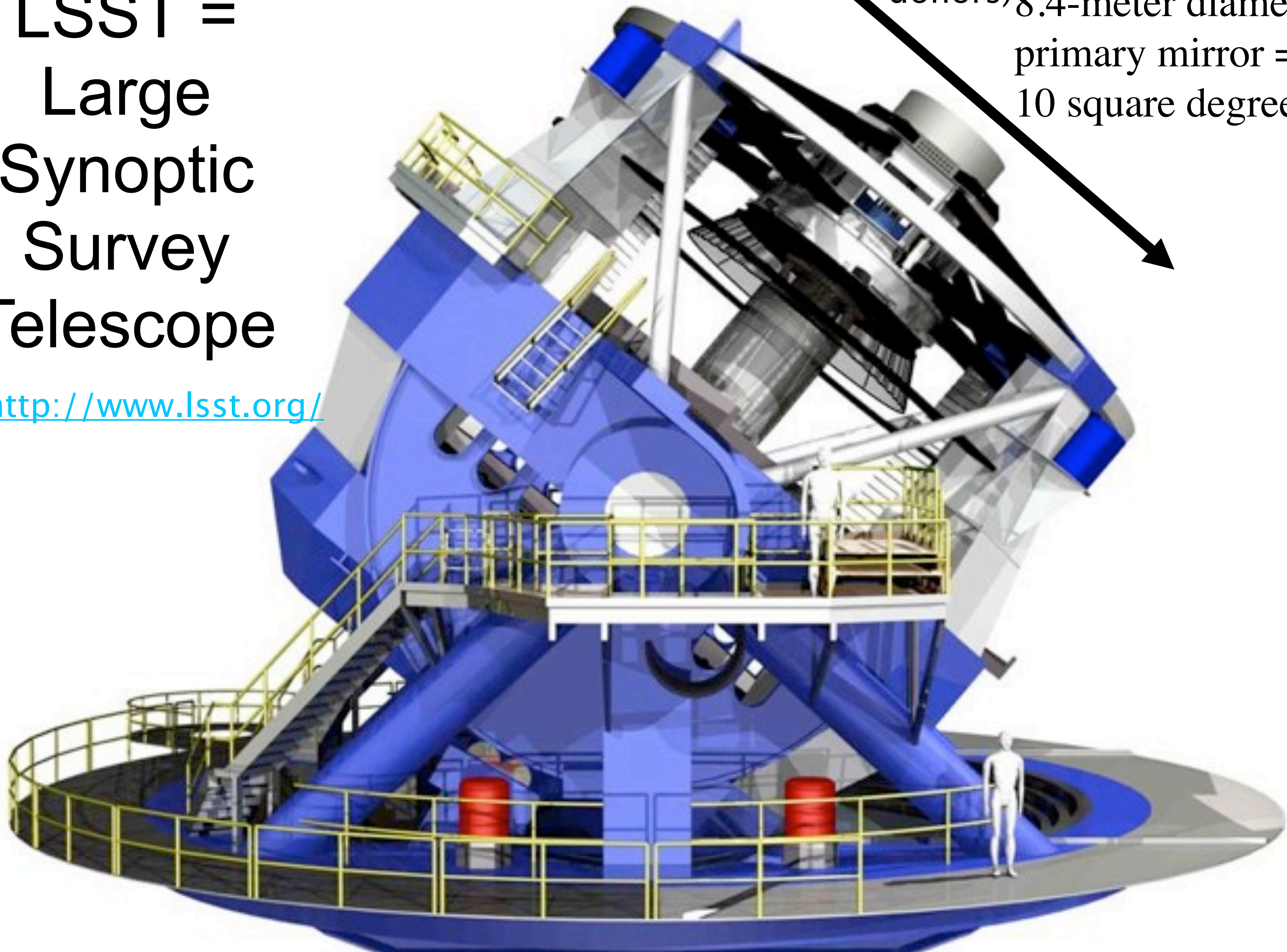
- The highest-ranked ground-based astronomy facility for the next decade in the Astro2010 Decadal Survey Report

LSST =
Large
Synoptic
Survey
Telescope

http://www.lsst.org/

(mirror funded by private donors)

8.4-meter diameter primary mirror = 10 square degrees!



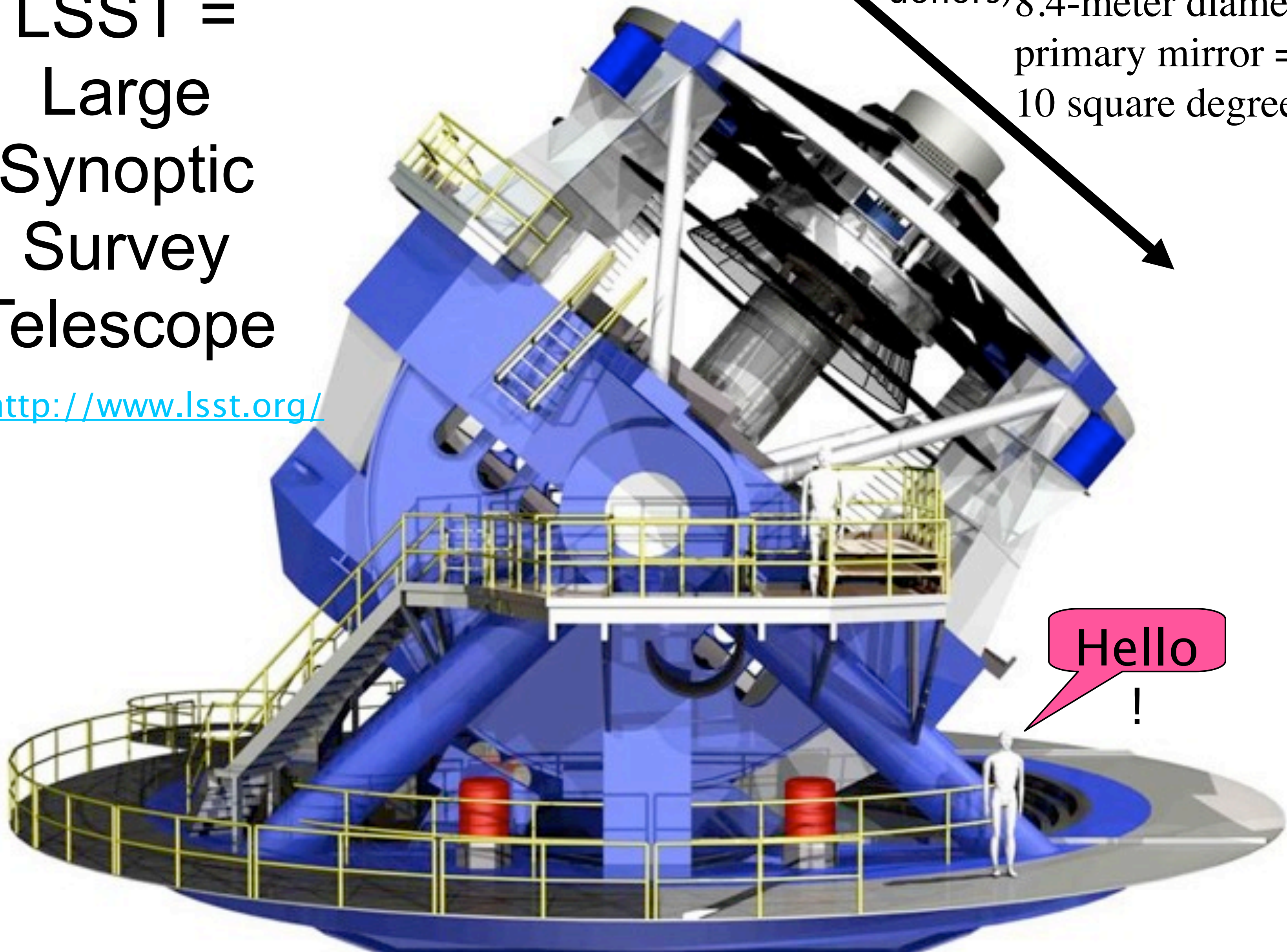
(design, construction, and operations of telescope, observatory, and data system: NSF)
(camera: DOE)

LSST = Large Synoptic Survey Telescope

http://www.lsst.org/

(mirror funded by private donors)

8.4-meter diameter primary mirror = 10 square degrees!



(design, construction, and operations of telescope, observatory, and data system: NSF)
(camera: DOE)

- Solar System Map (moving objects, NEOs, asteroids: census & tracking)
- Nature of Dark Energy (distant supernovae, weak lensing, cosmology)
- Optical transients (of all kinds, with alert notifications within 60 seconds)

dark matter)

South America

Chile

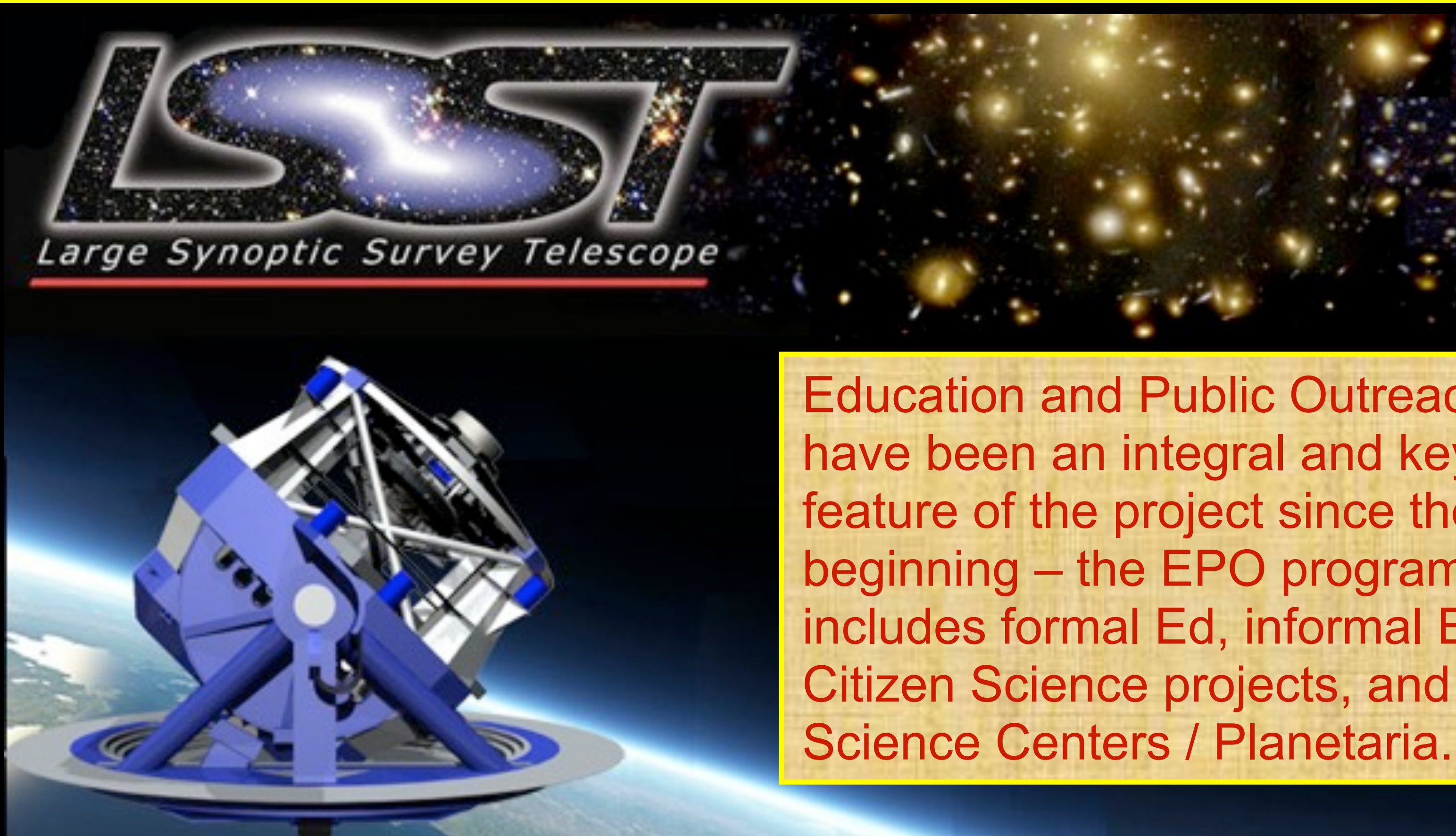
Region de Coquimbo

Summit of Cerro Pachon -

<u>LSST in time and space:</u>
- When? 2016-2026
- Where? Cerro Pachon, Chile

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/***



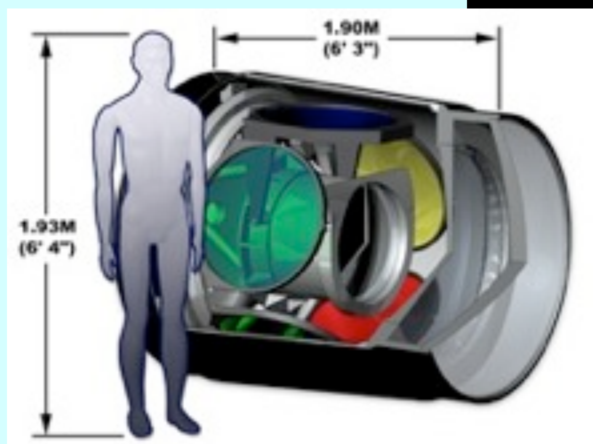
Education and Public Outreach have been an integral and key feature of the project since the beginning – the EPO program includes formal Ed, informal Ed, Citizen Science projects, and Science Centers / Planetaria.

The LSST focal plane array

Camera Specs: <u>(pending funding from the DOE)</u>
201 CCDs @ 4096x4096 pixels <u>each</u>!
= 3 Gigapixels = 6 GB per image, covering 10 sq.degrees
= ~3000 times the area of one Hubble Telescope image

LSST Data Challenges

- Obtain one 6-GB sky image in 15 seconds
- Process that image in 5 seconds
- Obtain & process another co-located image for science validation within 20^s (= 15-second exposure + 5-second processing & slew)
- Process the 100 million sources in each image pair, catalog all sources, and generate worldwide alerts within 60 seconds (e.g., incoming killer asteroid)
- Generate 100,000 alerts per night (VOEvent messages)
- Obtain 2000 images per night
- Produce ~30 Terabytes per night
- Move the data from South America to US daily
- Repeat this every day for 10 years (2016-2026)
- Provide rapid DB access to worldwide community:
  - **100-200 Petabyte image archive**
  - **20-40 Petabyte database catalog**



We proposed a new collaboration in 2009: Informatics and Statistical Sciences Collaboration (ISSC)

- We noted that **there is one significant research area that is not represented in the original 10 teams.**
  - That area is **Informatics and Statistics Research:**
    - **Astroinformatics**
    - **Astrostatistics**
- The Computer Science (data mining and machine learning) and Statistics research communities are becoming aware of and interested in LSST (astronomy data are abundant, interesting, and free).
  - The LSST data collection will be large and complex.

The new LSST ISSC research team

- In discussing the data-related research challenges posed by LSST, we identified several research areas:
  - Statistics
  - Data & Information Visualization
  - Data mining (machine learning)
  - Data-intensive computing & analysis
  - Large-scale scientific data management
- These areas represent Statistics and the science of Informatics (**Astroinformatics**) = Data-intensive Science = **the 4th Paradigm of Scientific Research**

The new LSST ISSC research team

- In discussing the data-related research challenges posed by LSST, we identified several research areas:
  - Statistics
  - Data & Information Visualization
  - Data mining (machine learning)
  - Data-intensive computing & analysis
  - Large-scale scientific data management

Informatics

- These areas represent Statistics and the science of Informatics (**Astroinformatics**) = Data-intensive Science = **the 4th Paradigm of Scientific Research**

The LSST ISSC Research Team

- Chairperson: K.Borne, GMU
- Core team: 3 astronomers + 2 =
  - K.Borne (scientific data mining in astronomy)
  - Eric Feigelsen, Tom Loredo (astrostatistics)
  - Jogesh Babu (statistics)
  - Alex Gray (computer science, data mining)
- Full team: ~30 scientists
  - ~60% astronomers
  - ~30% statisticians
  - ~10% data mining, machine learning computer scientists
- Original ISSC proposal: 50+ co-signers, only half were astronomers

Full list of team members:
http://www.lsstcorp.org/strausstest/StraussTest2.php

Some key astronomy problems that require informatics and statistical techniques … ***Astroinformatics & Astrostatistics!***

- Probabilistic Cross-Matching of objects from different catalogues
- The distance problem (e.g., Photometric Redshift estimators)
- Star-Galaxy separation ; QSO-Star separation
- Cosmic-Ray Detection in images
- Supernova Detection and Classification
- Morphological Classification (galaxies, AGN, gravitational lenses, ...)
- Class and Subclass Discovery (brown dwarfs, methane dwarfs, ...)
- Dimension Reduction = Correlation Discovery
- Learning Rules for improved classifiers
- Classification of massive data streams
- Real-time Classification of Astronomical Events
- Clustering of massive data collections
- Novelty, Anomaly, Outlier Detection in massive databases

ISSC "current topics"

- Advancing the field = Community-building:
  - Astroinformatics + Astrostatistics (several workshops this year!!)
  - Education, education, education! (Citizen Science, undergrad+grad ed…)
- LSST Event Characterization vs. Classification
- Sparse time series and the LSST observing cadence
- Challenge Problems, such as the Photo-z challenge and the Supernova Photometric Classification challenge
- Testing algorithms on the LSST simulations: images/catalogs PLUS observing cadence – can we recover known classes of variability?
- Generating and/or accumulating training samples of numerous classes (especially variables and transients)
- Proposing a mini-survey during the science verification year (Science Commissioning):
  - *e.g.,* high-density and evenly-spaced observations of extragalactic and Galactic test fields are obtained, to generate training sets for variability classification and assessment thereof
- Science Data Quality Assessment (SDQA): R&D efforts to support LSST Data Management team

LSST Level 3 Products from ISSC (TBD)

- Training sets for classification of various classes of variability or transient behavior
- Comparison samples of statistically robust classes of objects (non-transients), for use in evaluating the LSST object catalog
- Algorithms: e.g., for statistical analysis, time series analysis, photo-z estimation, star-galaxy separation, outlier (surprise) detection, data mining, …
- Results from precursor experiments on LSST event classification and characterization – training sets, results, algorithms, recommended cadences, etc.

Outline

- Prelude
- Astroinformatics
- Example Application: The LSST Project
- Informatics & Statistics Challenge Problems
- Challenge Area: Distributed Data Mining
- Summary

Basic Astronomical Knowledge Problems – 1

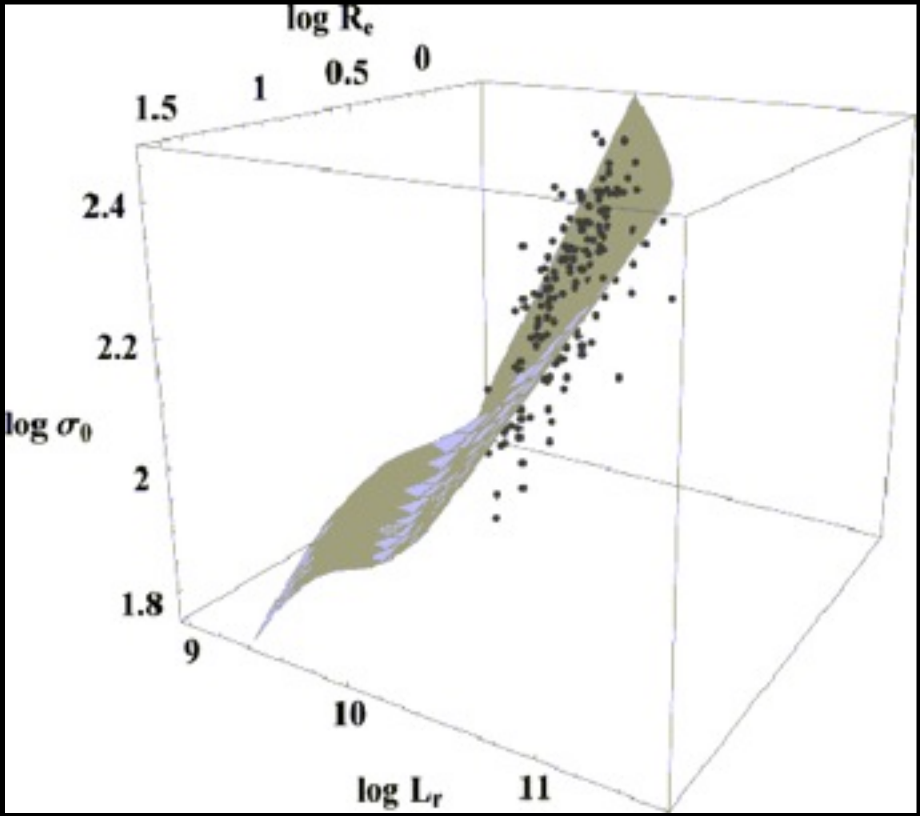
- **The clustering problem:**
  - Finding clusters of objects within a data set
  - What is the significance of the clusters (statistically and scientifically)?
  - What is the optimal algorithm for finding friends-of-friends or nearest neighbors?
    - N is $>10^{10}$, so what is the most efficient way to sort?
    - Number of dimensions ~ 1000 – therefore, we have an enormous subspace search problem
  - Are there pair-wise (2-point) or higher-order (N-way) correlations?
    - N is $>10^{10}$, so what is the most efficient way to do an N-point correlation?
      - algorithms that scale as $N^2 \log N$ won't get us there

Basic Astronomical Knowledge Problems – 2

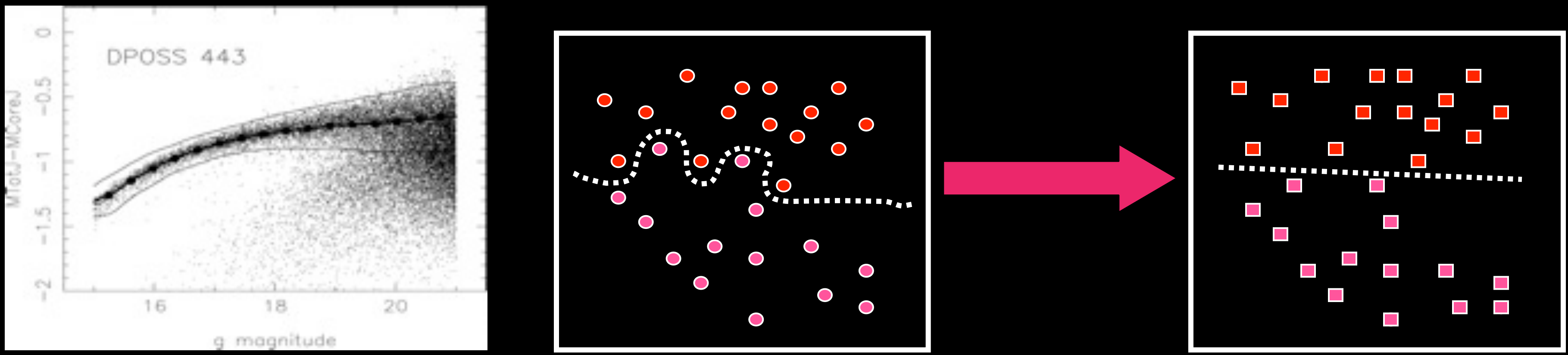
- **Outlier detection: (unknown unknowns)**
  - Finding the objects and events that are outside the bounds of our expectations (outside known clusters)
  - These may be real scientific discoveries or garbage
  - Outlier detection is therefore useful for:
    - Novelty Discovery – *is my Nobel prize waiting?*
    - Anomaly Detection – *is the detector system working?*
    - Data Quality Assurance – *is the data pipeline working?*
  - How does one optimally find outliers in 10^3-D parameter space? or in interesting subspaces (in lower dimensions)?
  - How do we measure their "interestingness"?

Basic Astronomical Knowledge Problems – 3

- **The dimension reduction problem:**
  - Finding correlations and “fundamental planes” of parameters
  - Number of attributes can be hundreds or thousands
    - **The Curse of High Dimensionality !**
  - Are there combinations (linear or non-linear functions) of observational parameters that correlate strongly with one another?
  - Are there eigenvectors or condensed representations (e.g., basis sets) that represent the full set of properties?



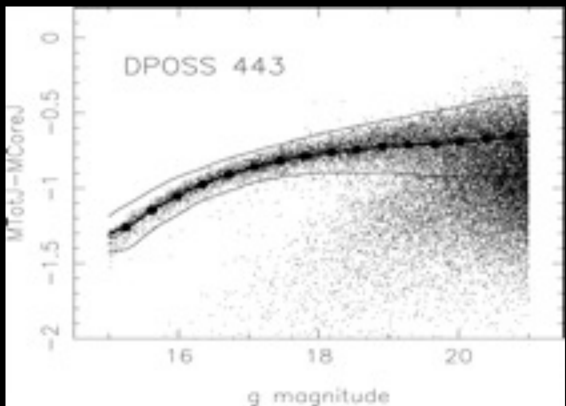
Basic Astronomical Knowledge Problems – 4

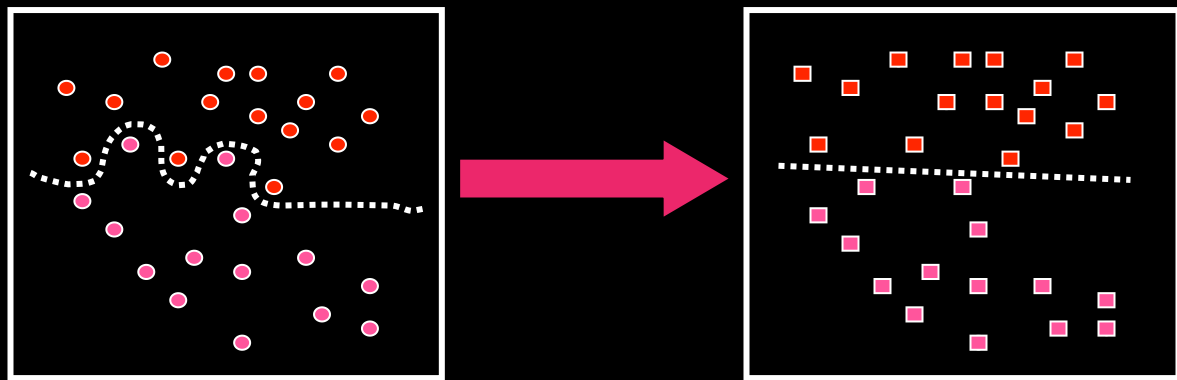


mercoledì 25 agosto 2010

Basic Astronomical Knowledge Problems – 4

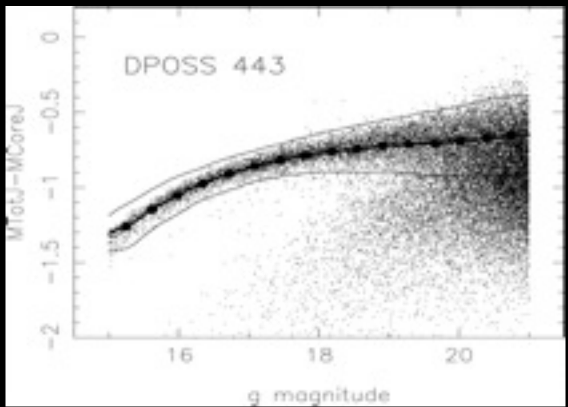
- **The superposition / decomposition problem:**
  - Finding distinct clusters (Classes of Object) among objects that overlap in parameter space

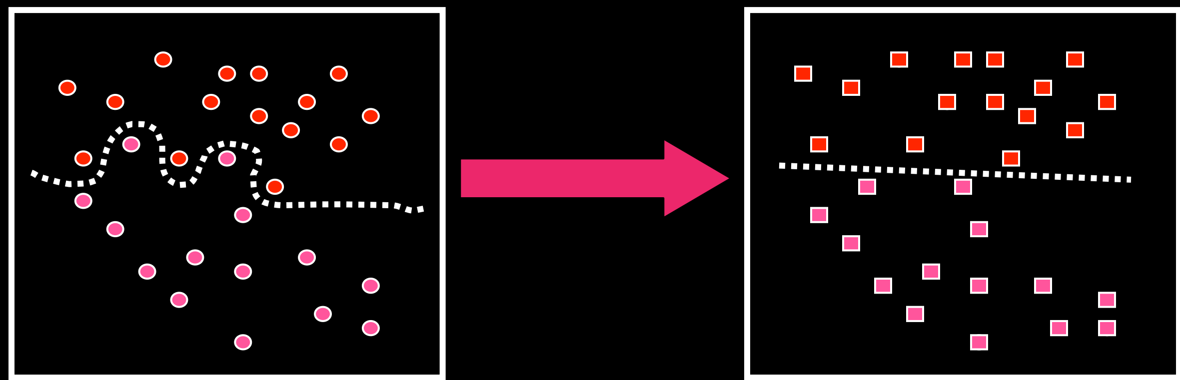




Basic Astronomical Knowledge Problems – 4

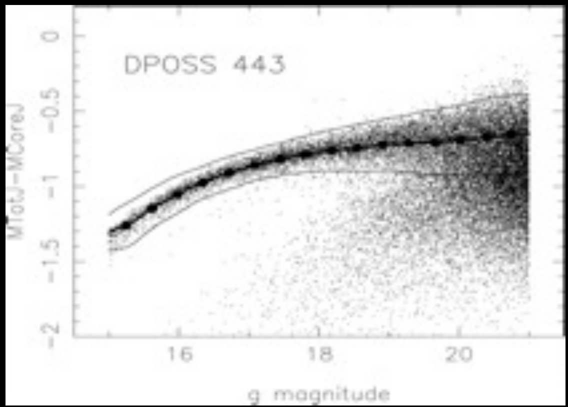
- **The superposition / decomposition problem:**
  - Finding distinct clusters (Classes of Object) among objects that overlap in parameter space

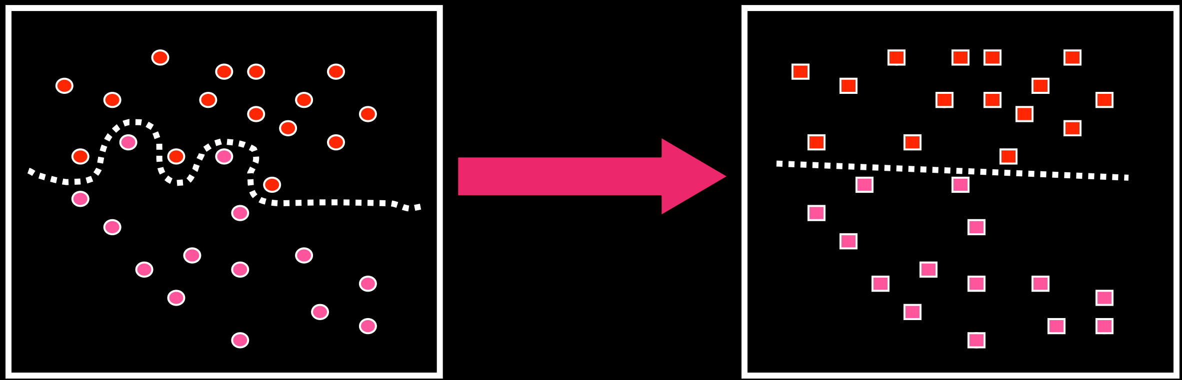




Basic Astronomical Knowledge Problems – 4

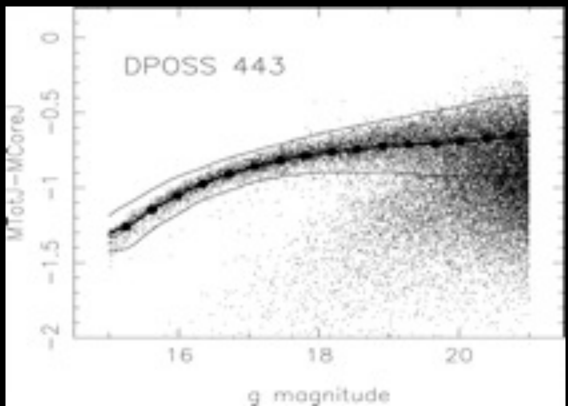
- **The superposition / decomposition problem:**
  - Finding distinct clusters (Classes of Object) among objects that overlap in parameter space

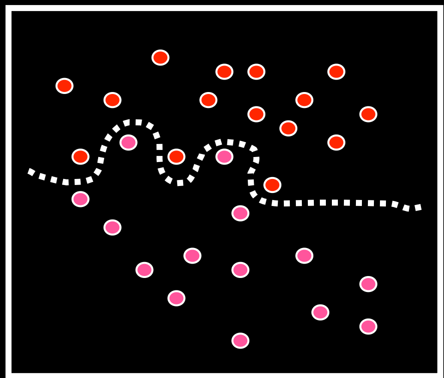


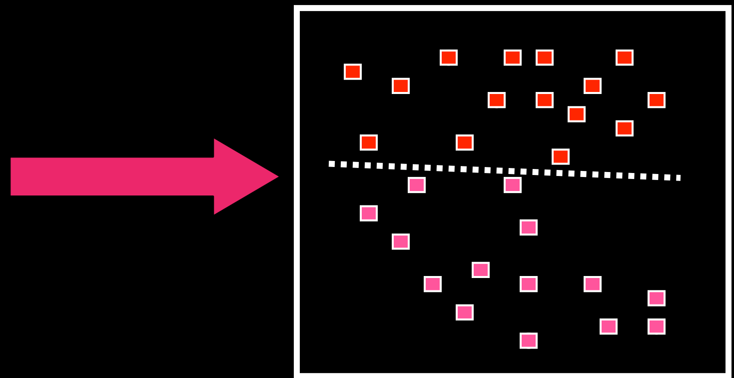


Basic Astronomical Knowledge Problems – 4

- **The superposition / decomposition problem:**
  - Finding distinct clusters (Classes of Object) among objects that overlap in parameter space

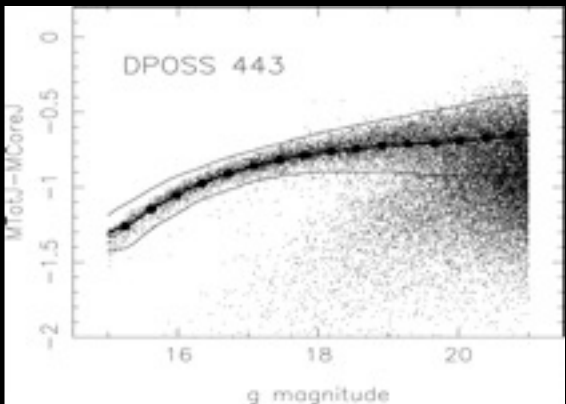


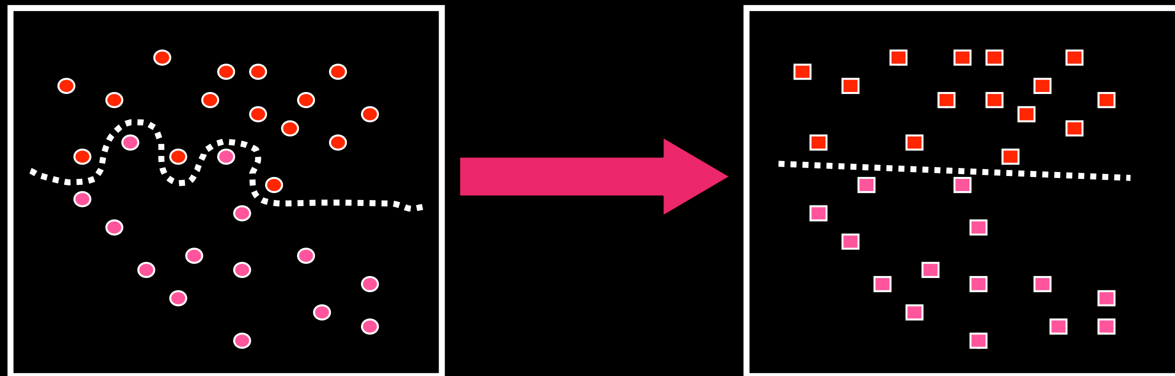




Basic Astronomical Knowledge Problems – 4

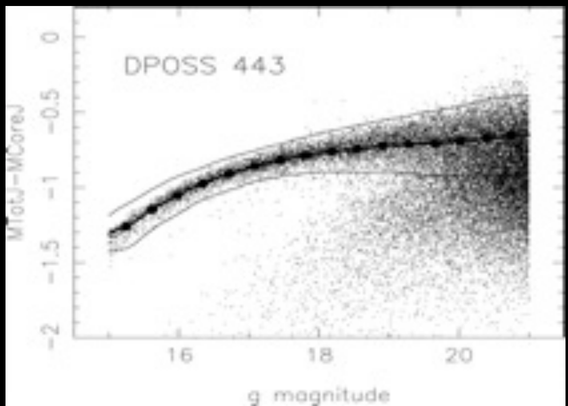
- **The superposition / decomposition problem:**
  - Finding distinct clusters (Classes of Object) among objects that overlap in parameter space

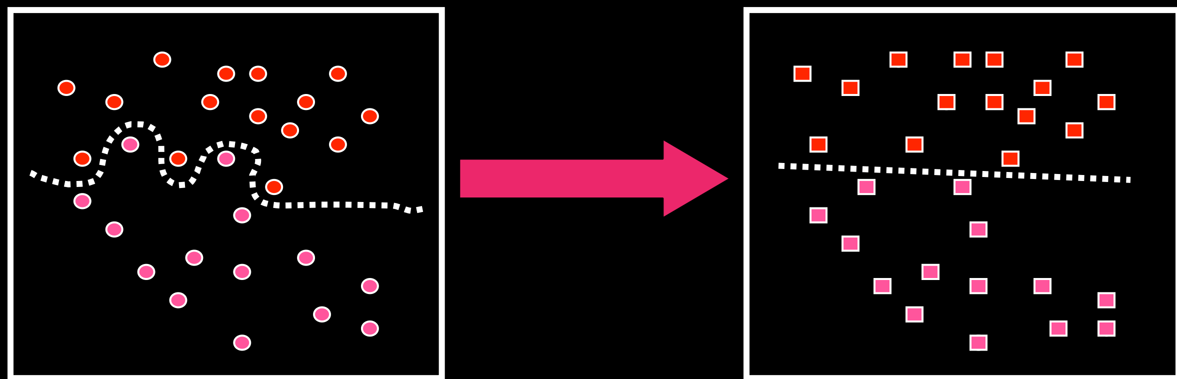




Basic Astronomical Knowledge Problems – 4

- **The superposition / decomposition problem:**
  - Finding distinct clusters (Classes of Object) among objects that overlap in parameter space

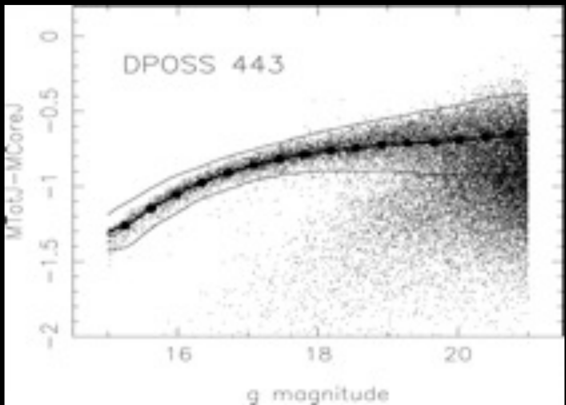


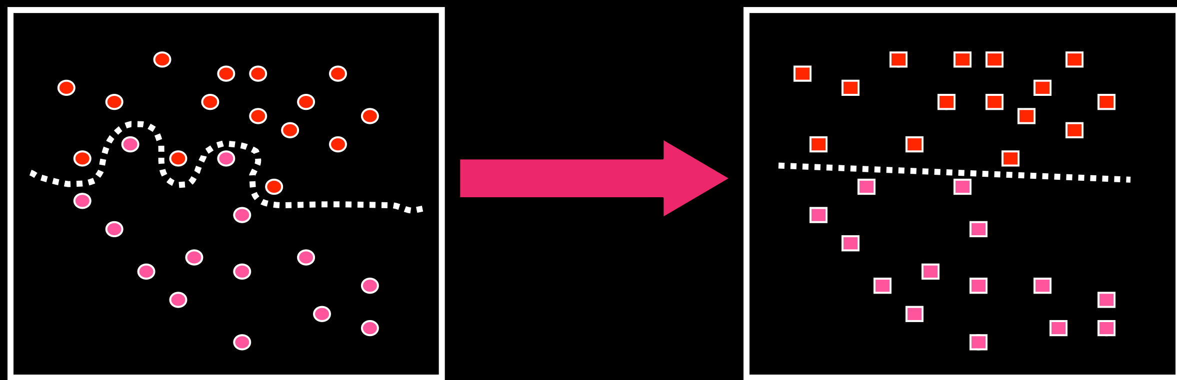


  - What if there are 10^{10} objects that overlap in a 10^{3}-D parameter space?

Basic Astronomical Knowledge Problems – 4

- **The superposition / decomposition problem:**
  - Finding distinct clusters (Classes of Object) among objects that overlap in parameter space

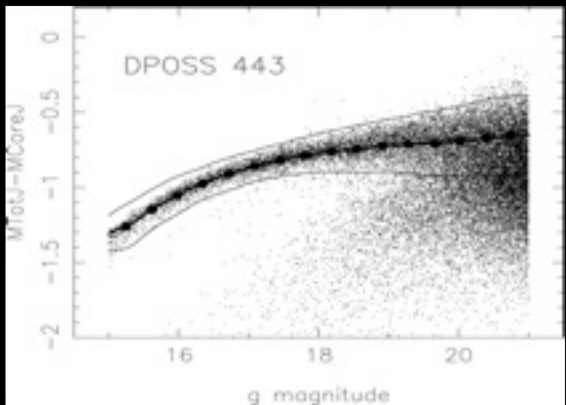


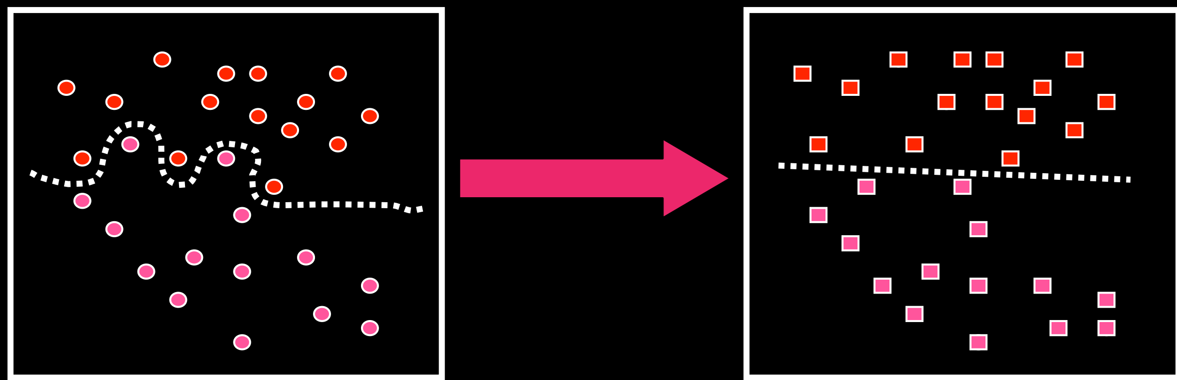


  - What if there are 10^{10} objects that overlap in a 10^3-D parameter space?
  - What is the optimal way to separate and extract the different unique classes of objects?

Basic Astronomical Knowledge Problems – 4

- **The superposition / decomposition problem:**
  - Finding distinct clusters (Classes of Object) among objects that overlap in parameter space

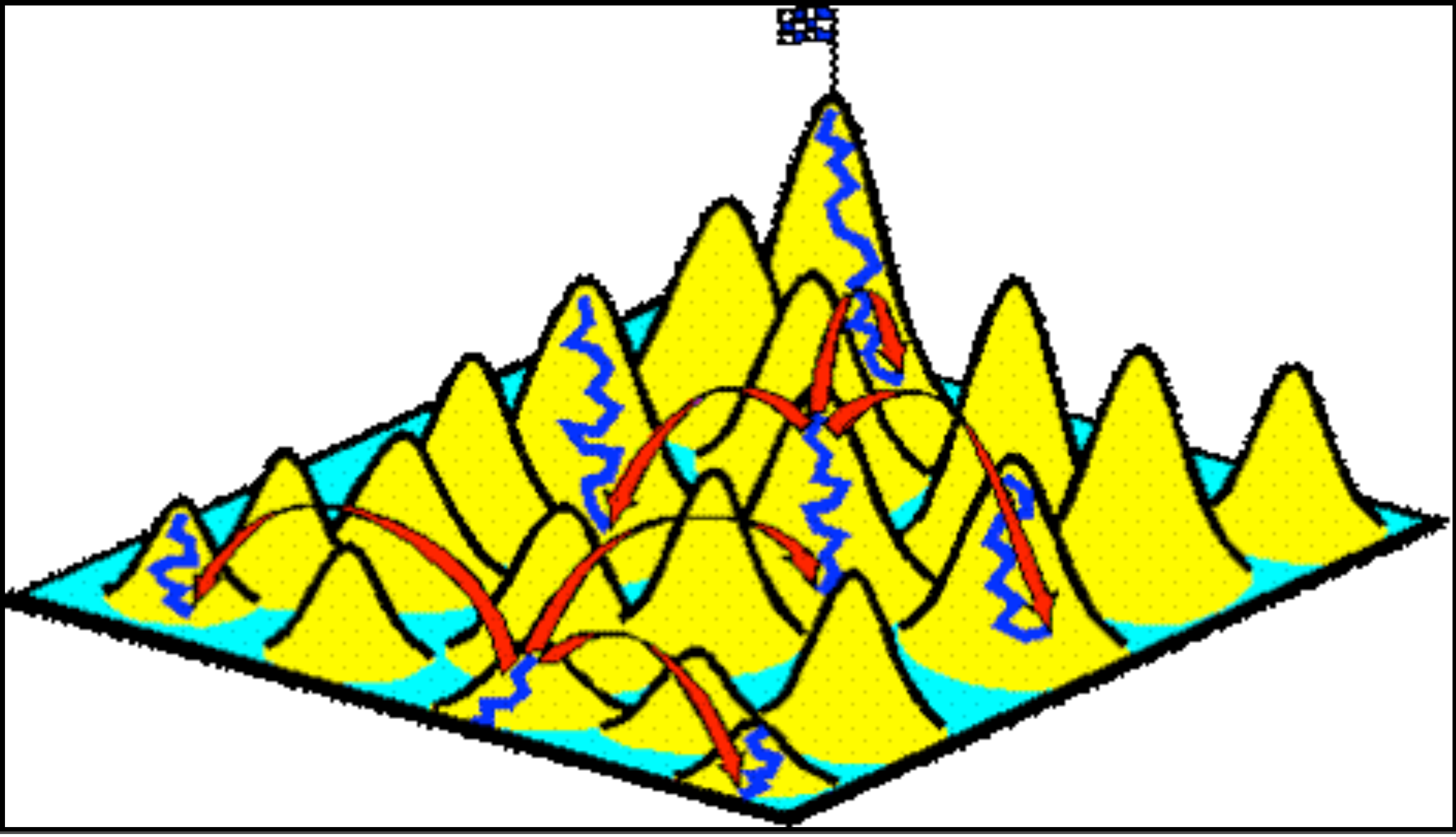




  - What if there are 10^{10} objects that overlap in a 10^3-D parameter space?
  - What is the optimal way to separate and extract the different unique classes of objects?
  - How are constraints applied?

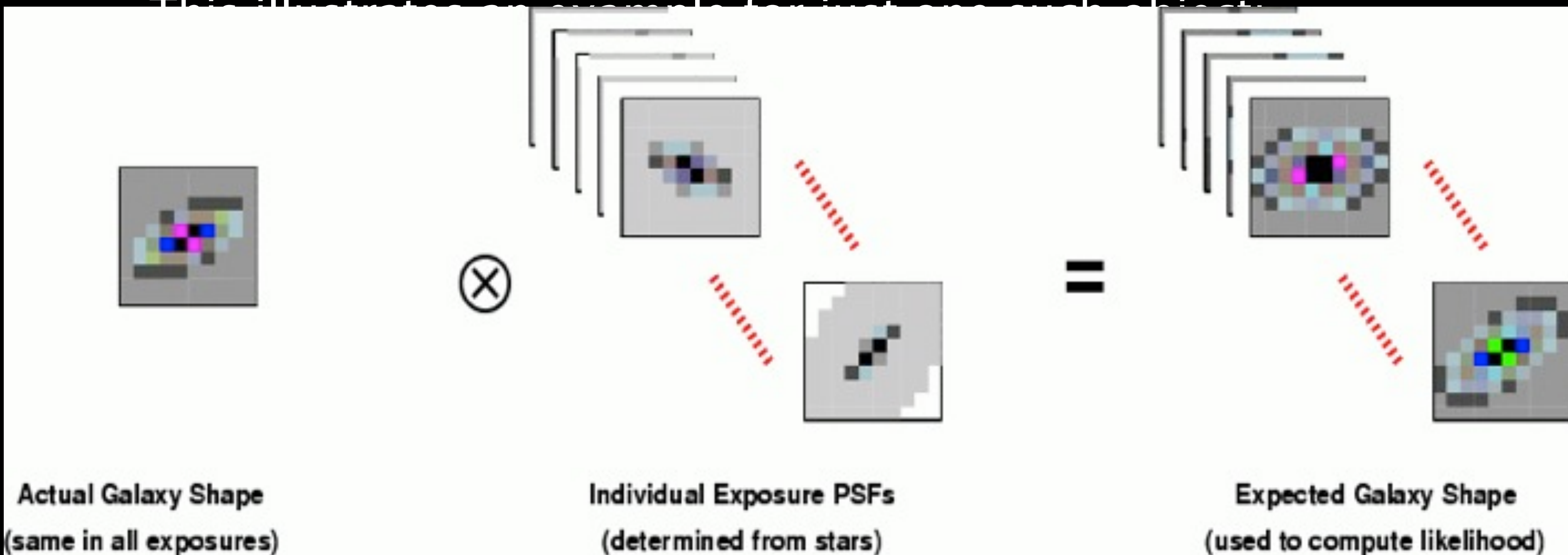
Basic Astronomical Knowledge Problems – 5

- **The optimization problem:**
  - Finding the optimal (best-fit, global maximum likelihood) solution to complex multivariate functions over very high-dimensional spaces



Example: Beyond Exascale Computational & Data Science

- Find the optimal <u>simultaneous</u> solution for 20,000,000,000 objects' shapes across 2000 image planes, each of which has 201x4096x4096 pixels ... 10^{23} floating-point operations!
  - This illustrates an example for just one such object:



References:
http://universe.ucdavis.edu/docs/MultiFit-ADASS.pdf

Outline

- Prelude
- Astroinformatics
- Example Application: The LSST Project
- Informatics & Statistics Challenge Problems
- Challenge Area: Distributed Data Mining
- Summary

Why Distributed Data Mining (DDM)?

Because …

… many great scientific discoveries have come from inter-comparisons of diverse data sources:

- Quasars
- Gamma-ray bursts
- Ultraluminous IR galaxies
- X-ray black-hole binaries
- Radio galaxies
- . . .

Why Distributed Data Mining

"Just checking."

“Just Checkin g”

Because ...

... many great scientific discoveries have come from inter-comparisons of diverse data sources:

- Quasars
- Gamma-ray bursts
Ultraluminous IR
laxies
-ray black-hole
inaries
- Radio galaxies
- . . .

Distributed Data

- Distributed data are the norm (across people, institutions, projects, agencies, nations, ...)
- Data are usually heterogeneous (e.g., databases, images, catalogs, file systems, web interfaces, document libraries, binary, text, structured, unstructured, ...)
- Scientists want to query and to mine these data (= 2 different user scenarios)
- Virtual Observatory lementations enable data disco d integration, but do not yet fac arge-scale

Distributed Data Mining (DDM)

- DDM comes in 2 types:
  1. **Distributed Mining** of Data
  2. Mining of **Distributed Data**
- Type 1 requires sophisticated algorithms that operate with data in situ …
  - Ship the Code to the Data
- Type 2 takes many forms, with data being centralized (in whole or in partitions) or data remaining in place at distributed sites
- References: http://www.cs.umbc.edu/~hillol/DDMBIB/
  - C. Giannella, H. Dutta, K. Borne, R. Wolff, H. Kargupta. (2006). Distributed Data Mining for Astronomy Catalogs. Proceedings of 9th Workshop on Mining Scientific and Engineering Datasets, as part of the SIAM International Conference on Data Mining (SDM), 2006. [http://www.cs.umbc.edu/~hillol/PUBS/Papers/Astro.pdf]
  - H. Dutta, C. Giannella, K. Borne and H. Kargupta. (2007). Distributed Top-K Outlier Detection from Astronomy Catalogs using the DEMAC System. Proceedings of the SIAM International Conference on Data Mining, Minneapolis, USA, April 2007. [http://www.cs.umbc.edu/~hillol/PUBS/Papers/sdm07.pdf]

Outline

- Prelude
- Astroinformatics
- Example Application: The LSST Project
- Informatics & Statistics Challenge Problems
- Challenge Area: Distributed Data Mining
- Summary

Data Science Challenge Areas in Astronomy over the next 10 years – addressable by Astroinformatics

- Scalability of statistical, computational, & data mining algorithms to peta- and exa- scales
- Algorithms for optimization of simultaneous multi-point fitting across massive multi-dimensional data cubes
- Multi-resolution, multi-pole, fractal, hierarchical methods and structures for exploration of condensed representations of petascale databases
- Petascale analytics for visual exploratory data analysis of massive databases (including feature detection, pattern & interestingness discovery, correlation mining, clustering, class discovery, eigen-monitoring, dimension reduction)
- Indexing and associative memory techniques (trees, graphs, networks) for highly-dimensional petabyte databases
- Rapid query and search algorithms for petabyte databases

Astroinformatics Research paper available !

Addresses the data science challenges, research agenda, application areas, use cases, and recommendations for the new science of *Astroinformatics*.

Borne (2010): “***Astroinformatics: Data-Oriented Astronomy Research and Education***”, *Journal of Earth Science Informatics, vol. 3, pp. 5-17.*

See also http://arxiv.org/abs/0909.3892

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

Astroinformatics: A 21st Century Approach to Astronomy

Authorship: This Position Paper was prepared and endorsed by the following team of 91 astronomers and information scientists (listed separately). The lead author is Kirk D. Borne (Dept. of Computational and Data Sciences, George Mason University, kborne@gmu.edu). The team maintains a web site that hosts information about the authors (including email addresses and links to web sites) and supporting information for this document: http://inference.astro.cornell.edu/Astro2010/ .