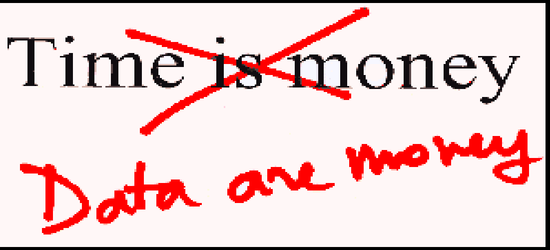
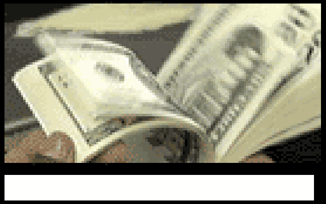
The Tonnabytes Big Data Challenge: Transforming Science and Education

Kirk Borne

George Mason University

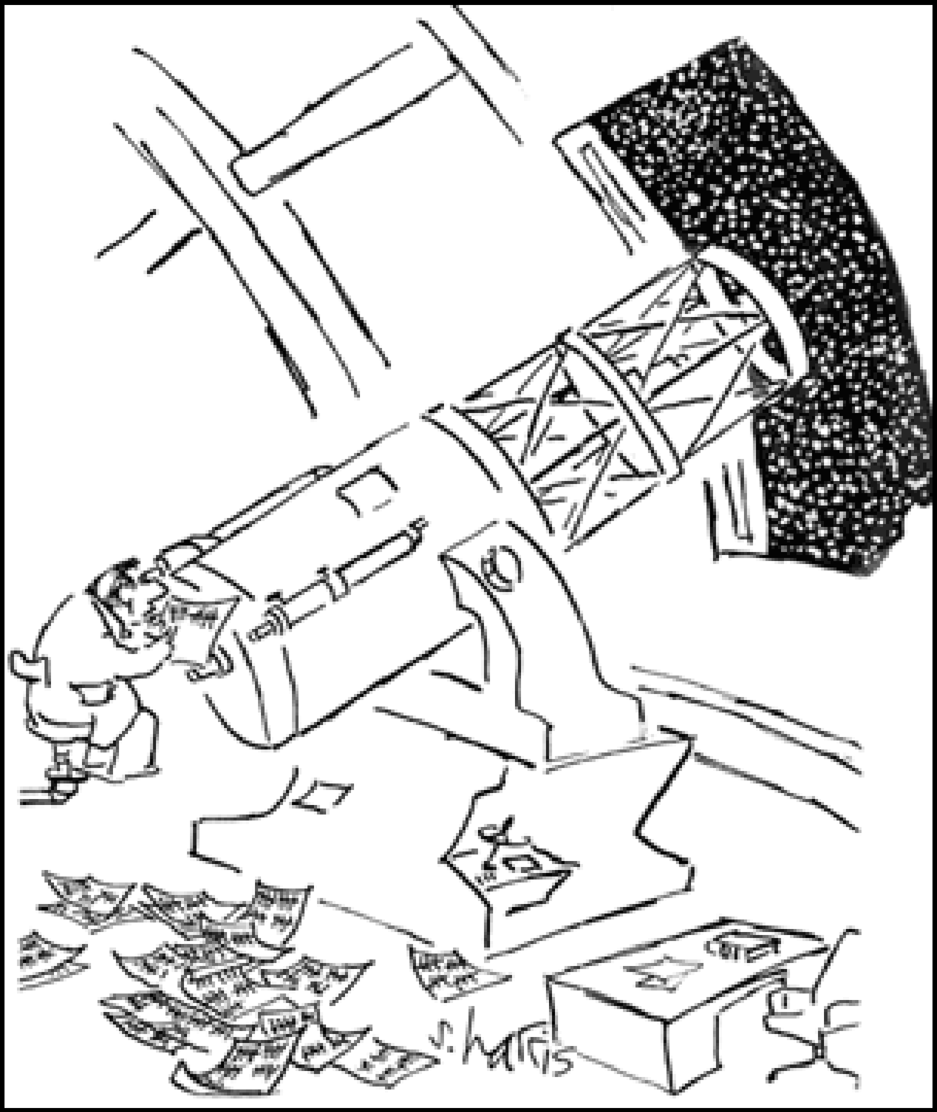




Ever since we first began to explore our world…

… humans have asked questions and …

… have collected evidence (data) to help answer those questions.



Astronomy: the world's second oldest profession !

Characteristics of Big Data

- **Big** quantities of data are acquired everywhere now. But…
- What do we mean by "**big**"?
  - Gigabytes? Terabytes? Petabytes? Exabytes?
  - The meaning of "big" is domain-specific and resource-dependent (data storage, I/O bandwidth, computation cycles, communication costs)
  - I say … we all are dealing with our own "**tonnabytes**"
- There are 4 dimensions to the Big Data challenge:
  1. **Volume** (*tonnabytes data challenge*)
  2. **Complexity** (*variety, curse of dimensionality*)
  3. **Rate of data and information flowing to us** *(velocity)*
  4. **Verification** (verifying inference-based models from data)
- Therefore, we need something better to cope with the data tsunami …

Data Science – Informatics – Data Mining

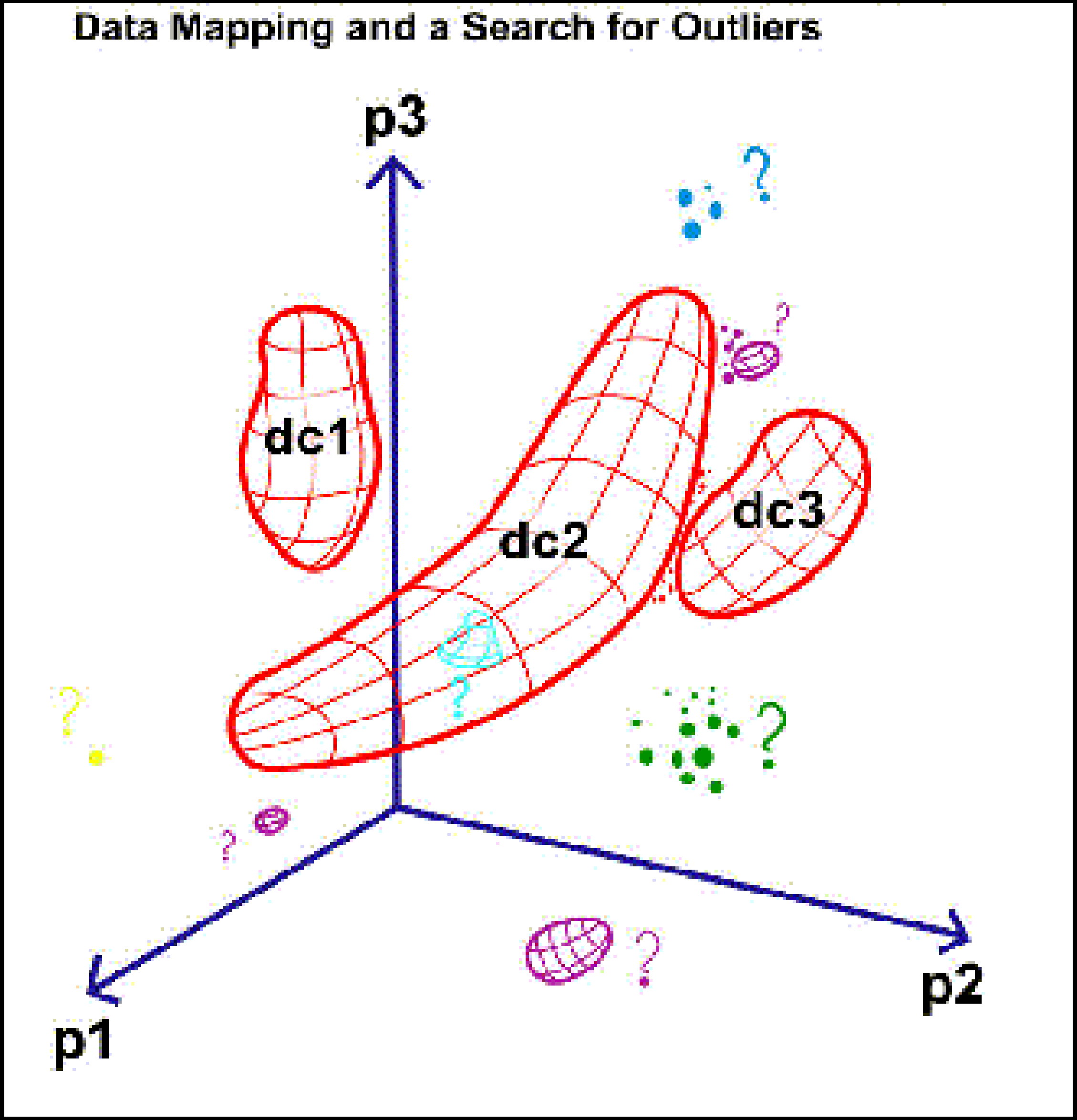
tagxedo.com

Examples of Recommendations**: Inference from 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 : Machine learning in massive data sets
- 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.

- ** From the NSF report: *Data-Enabled Science in the Mathematical and Physical Sciences*, (2010) http://www.cra.org/ccc/docs/reports/DES-report_final.pdf

This graphic says it all …

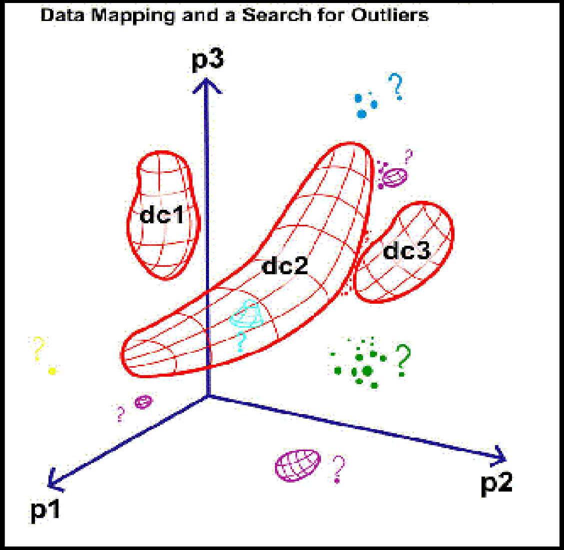


- Clustering – examine the data and find the data clusters (clouds), without considering what the items are = Characterization !
- Classification – for each new data item, try to place it within a known class (i.e., a known category or cluster) = Classify !
- Outlier Detection – identify those data items that don't fit into the known classes or clusters = Surprise !

Graphic provided by Professor S. G. Djorgovski, Caltech

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)

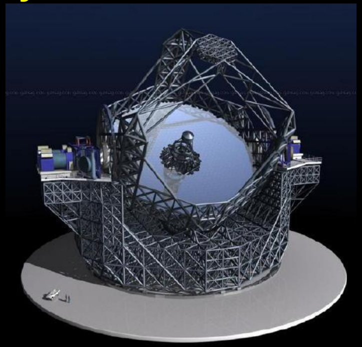


Graphic from S. G. Djorgovski

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

Astronomy Data Environment : Sky Surveys

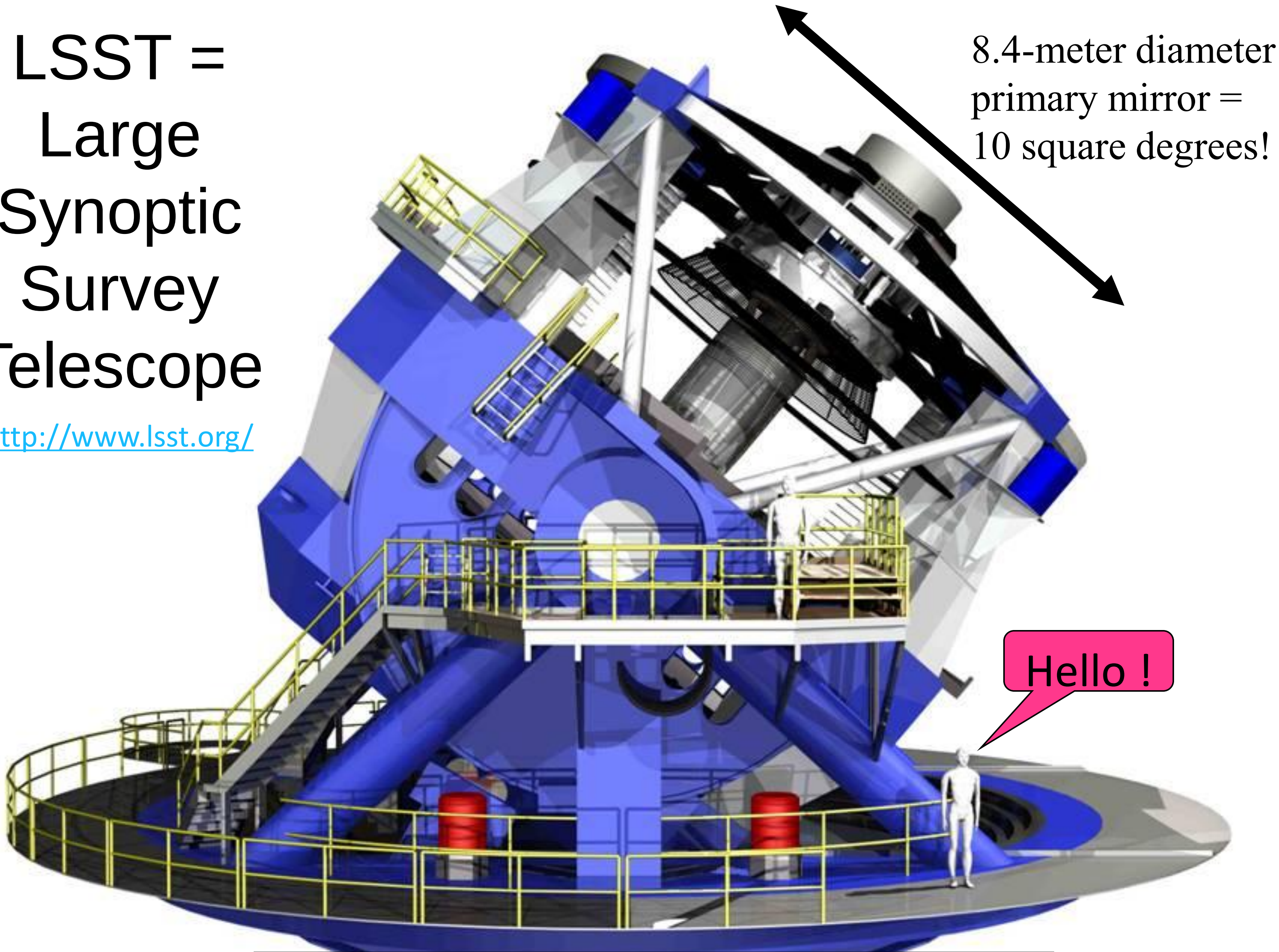
- To avoid biases caused by limited samples, astronomers now study the sky systematically = **Sky Surveys**
- Surveys are used to measure and collect data from all objects that are contained in large regions of the sky, in a systematic, controlled, repeatable fashion.
- These surveys include (... this 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!

LSST =
Large
Synoptic
Survey
Telescope

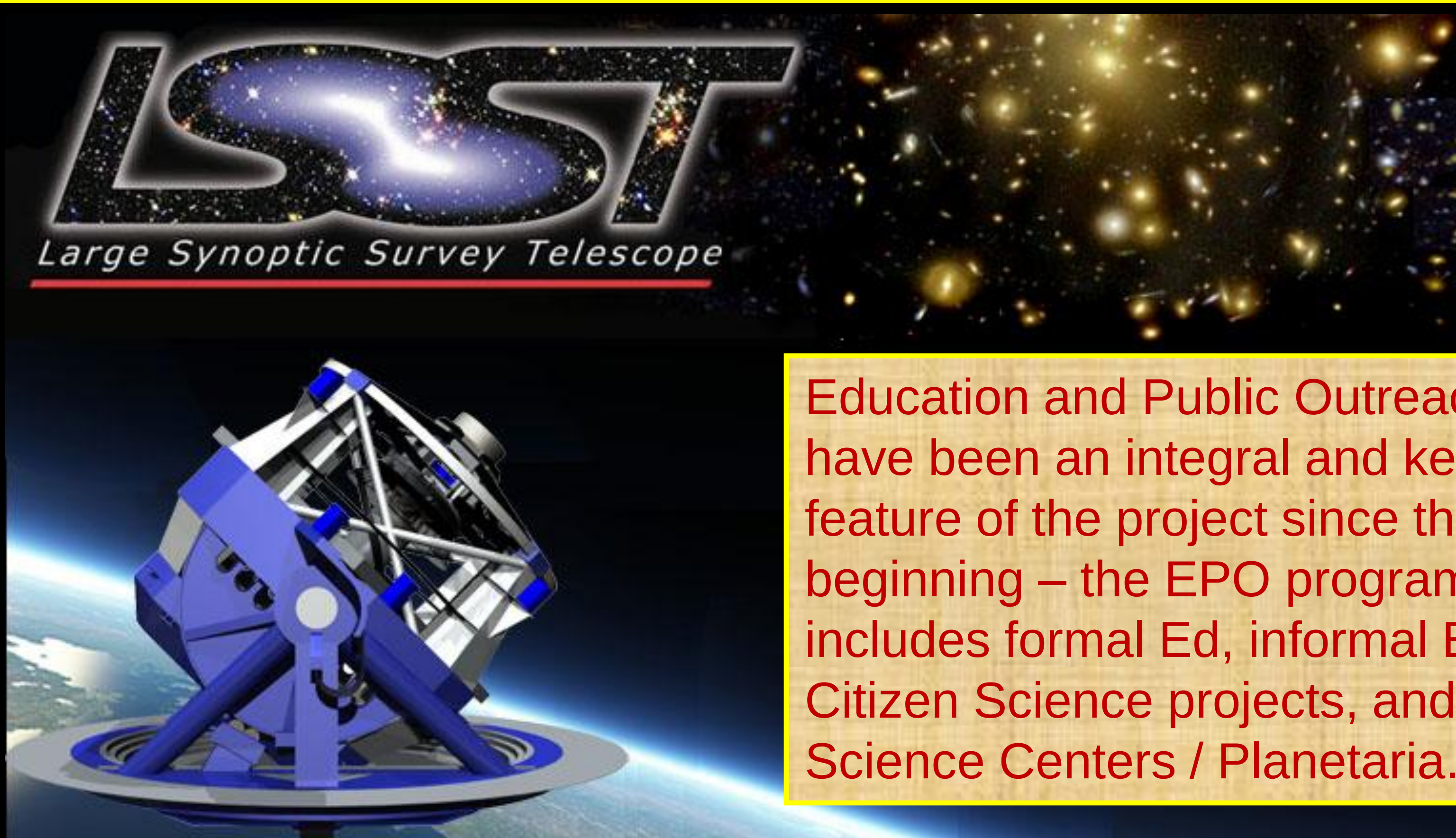
http://www.lsst.org/



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

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 (~2020-2030), 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 !
  - **~1,000,000 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.

LSST Key Science Drivers: Mapping the Dynamic Universe

– Solar System Inventory (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)
– Digital Milky Way (proper motions, parallaxes, star streams, dark matter)

South America

Chile

Region de Coquimbo

LSST in time and space:

– When? ~2020-2030
– Where? Cerro Pachon, Chile

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
- 100-Petabyte final image data archive anticipated – **all data are public!!!**
- **20-Petabyte final database catalog anticipated**
- Real-Time Event Mining: 1-10 million events per night, every night, for 10 yrs
  - Follow-up observations required to classify these
- Repeat images of the entire night sky every 3 nights: ***Celestial Cinematography***

LSST

Large Synoptic Survey Telescope

The LSST will represent a 10K-100K times increase in nightly rate of astronomical events.

This poses significant real-time characterization and classification demands on the event stream:

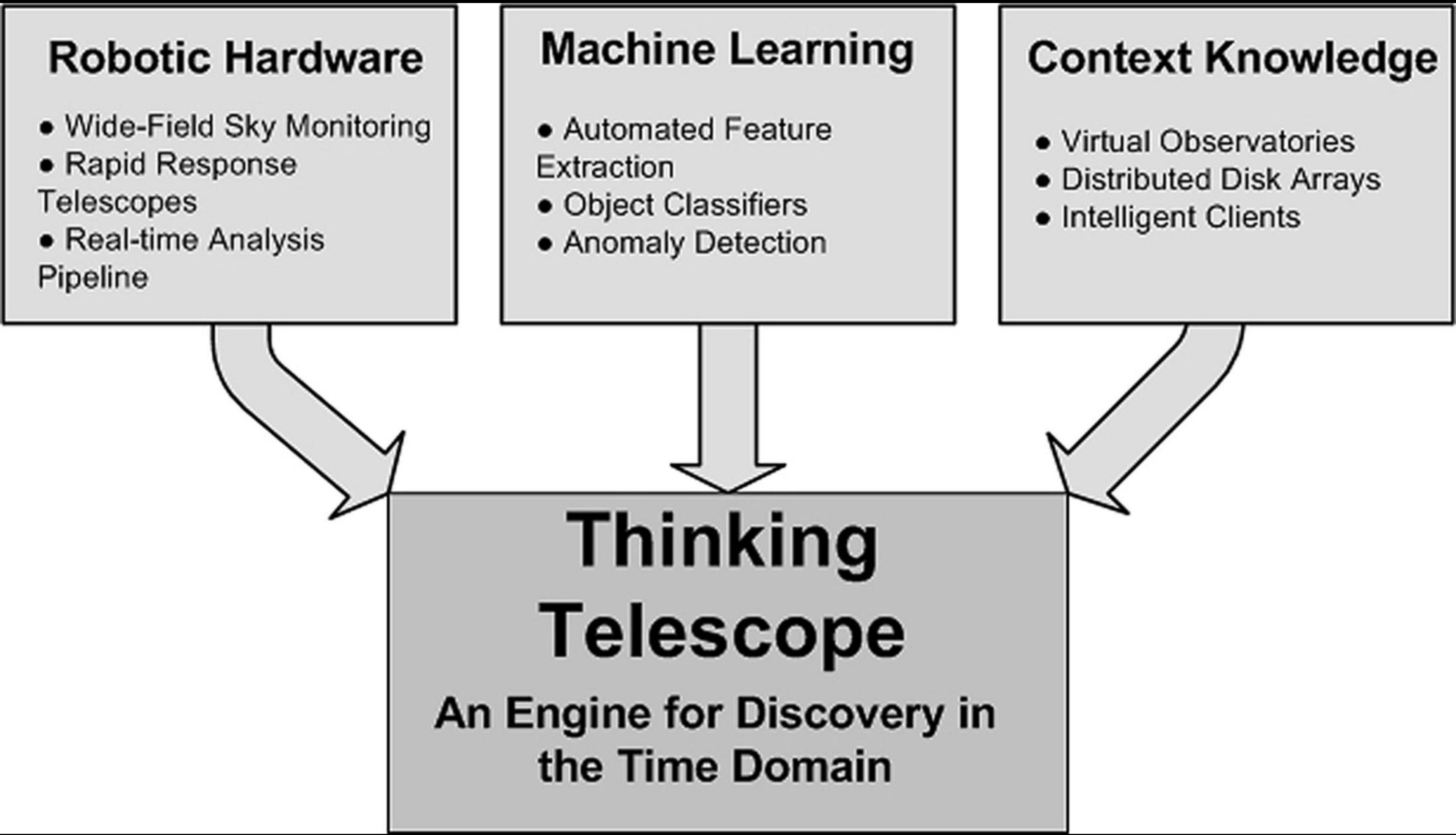
from data to knowledge!

from sensors to sense!

MIPS model for Event Follow-up

- MIPS =
  - **M**easurement – **I**nference – **P**rediction – **S**teering
- Heterogeneous Telescope Network = Global Network of Sensors (voeventnet.org, skyalert.org) :
  - Similar projects in NASA, NSF, DOE, NOAA, Homeland Security, DDDAS
- Machine Learning enables "IP" part of MIPS:
  - Autonomous (or semi-autonomous) Classification
  - Intelligent Data Understanding
  - Rule-based
  - Model-based
  - Neural Networks
  - Temporal Data Mining (Predictive Analytics)
  - Markov Models
  - Bayes Inference Engines

Example: The Los Alamos Thinking Telescope Project



Reference: http://en.wikipedia.org/wiki/Fenton_Hill_Observatory

From Sensors to Sense

Robotic Hardware

- Wide-Field Sky Monitoring
- Rapid Response Telescopes
- Real-time Analysis Pipeline

Machine Learning

- Automated Feature Extraction
- Object Classifiers
- Anomaly Detection

Context Knowledge

- Virtual Observatories
- Distributed Disk Arrays
- Intelligent Clients

From Data to Knowledge:
from sensors to sense (semantics)

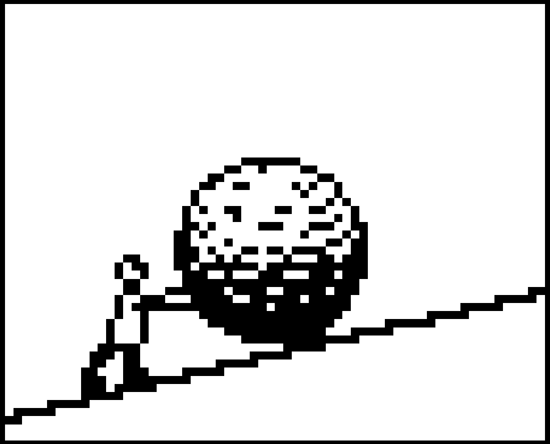
Thinking Telescope

An Engine for Discovery in the Time Domain

Data → Information → Knowledge

The LSST Data Mining Raison d'etre

- More data is not just more data … more is different!
- Discover the unknown unknowns.
- Massive Data-to-Knowledge challenge.



The LSST Data Mining Challenges

1. Massive data stream: ~2 Terabytes of image data per hour that must be mined in real time (for 10 years).

2. Massive 20-Petabyte database: more than 50 billion objects need to be classified, and most will be monitored for important variations in real time.

3. Massive event stream: knowledge extraction in real time for 1,000,000 events each night.

- Challenge #1 includes both the static data mining aspects of #2 and the dynamic data mining aspects of #3.

- Look at these in more detail ...

LSST challenges # 1, 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 (~30 TB each night ≈ 60,000 CDs filled with data):

LSST challenges # 1, 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 (~30 TB each night ≈ 60,000 CDs filled with data): *a sea of CDs*

LSST challenges # 1, 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 (~30 TB each night ≈ 60,000 CDs filled with data): *a sea of CDs*

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

LSST challenges # 1, 2

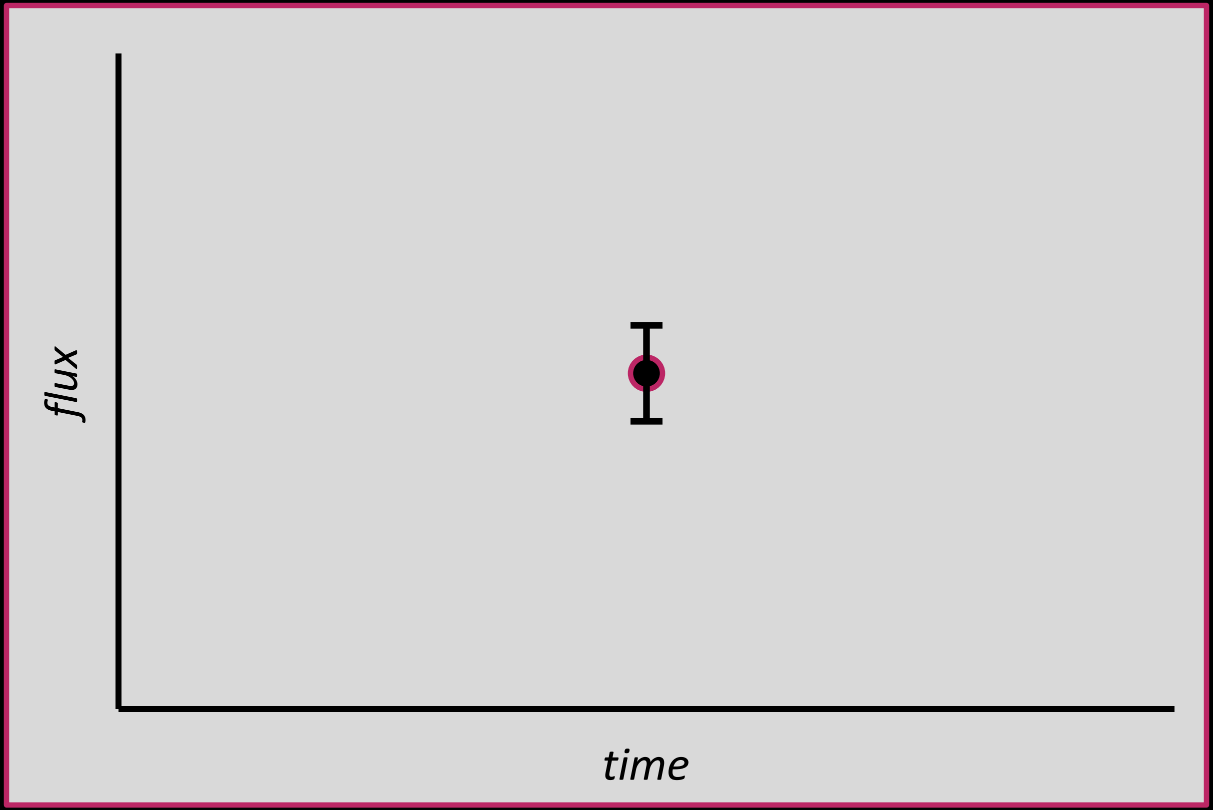
- **Each night** for 10 years LSST will obtain the equivalent amount of data that was obtained by the entire Sloan Digital Sky Survey
- My grad students will be asked to mine these data (~30 TB each night ≈ 60,000 CDs filled with data):
  - *A sea of CDs each and every day for 10 yrs*
  - *Cumulatively, a football stadium full of 200 million CDs after 10 yrs*
- The challenge is to find the new, the novel, the interesting, and the surprises (the unknown unknowns) within all of these data.
- *Yes, more is most definitely different !*

LSST data mining challenge # 3

- Approximately 1,000,000 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:

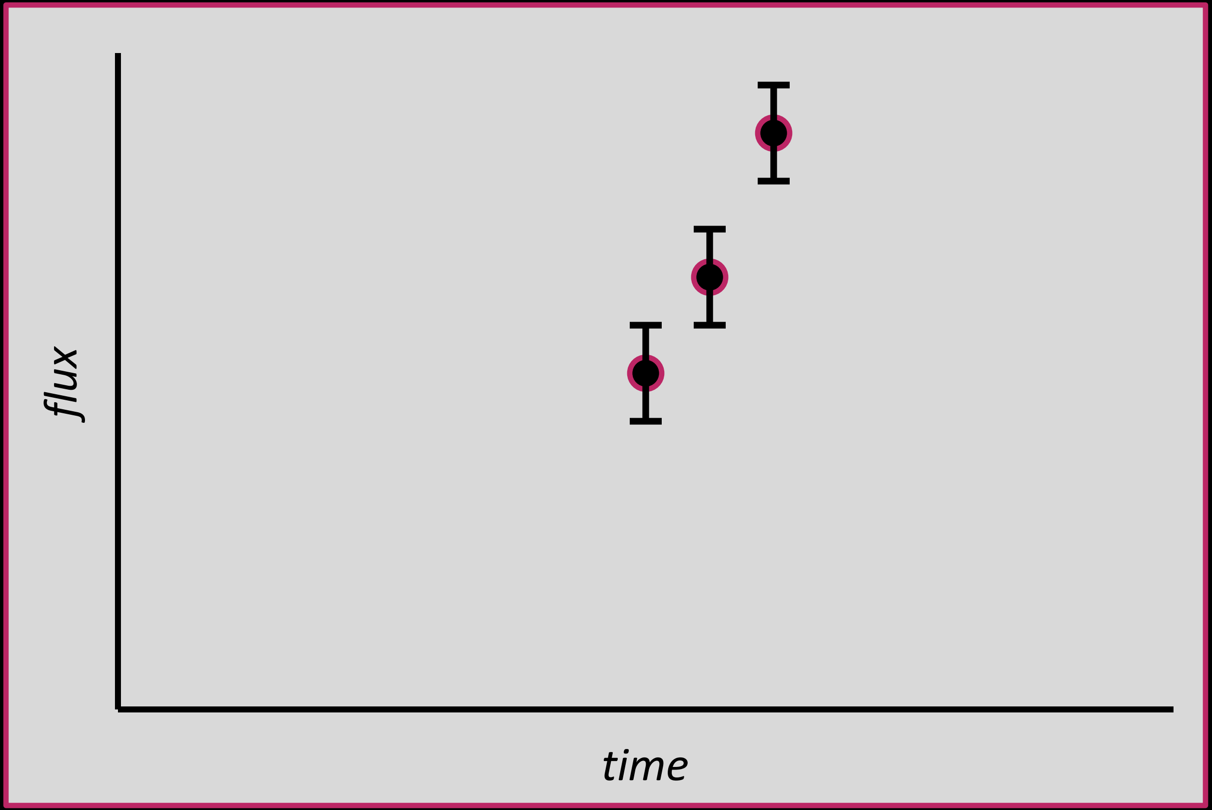
LSST data mining challenge # 3

- Approximately 1,000,000 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 data mining challenge # 3

- Approximately 1,000,000 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 data mining challenge # 3

- Approximately 1,000,000 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 !
(Unsupervised Learning)

Classify later.

Characterization includes …

- Feature Detection and Extraction:
  - Identifying and describing features in the data
  - Extracting feature descriptors from the data
  - Curating these features for search & re-use
  - Finding other parameters and features from other archives, other databases, other information sources – and using those to help characterize (ultimately classify) each new event.
  - … hence, coping with a highly multivariate parameter space

- Interesting question: can we standardize these steps?

Data-driven Discovery (Unsupervised Learning)

- **Class Discovery – Clustering**
  - Distinguish different classes of behavior or different types of objects
  - Find new classes of behavior or new types of objects
  - Describe a large data collection by a small number of condensed representations
- **Principal Component Analysis – Dimension Reduction**
  - Find the dominant features among all of the data attributes
  - Generate low-dimensional descriptions of events and behaviors, while revealing correlations and dependencies among parameters
  - Addresses the Curse of Dimensionality
- **Outlier Detection – Surprise / Anomaly / Deviation / Novelty Discovery**
  - Find the unknown unknowns (the rare one-in-a-billion or one-in-a-trillion event)
  - Find objects and events that are outside the bounds of our expectations
  - These could be garbage (erroneous measurements) or true discoveries
  - Used for data quality assurance and/or for discovery of new / rare / interesting data items
- **Link Analysis – Association Analysis – Network Analysis**
  - Identify connections between different events (or objects)
  - Find unusual (improbable) co-occurring combinations of data attribute values
  - Find data items that have much fewer than "6 degrees of separation"

Why do all of this?

… for 4 very simple reasons:

- (1) Any real data table may consist of thousands, or millions, or billions of rows of numbers.
- (2) Any real data table will probably have many more (perhaps hundreds more) attributes (features), not just two.
- (3) Humans can make mistakes when staring for hours at long lists of numbers, especially in a dynamic data stream.
- (4) The use of a data-driven model provides an objective, scientific, rational, and justifiable test of a hypothesis.

Why do all of this?

… for 4 very simple reasons:

- (1) Any real data table may consist of
  Volume nds, or millions, or billions of rows of numbers.
- (2) Any real data table will probably have
  Variety nore (perhaps hundreds more) attributes (features), not just two.
- (3) Humans can make mistakes when
  Velocity for hours at long lists of numbers, especially in a dynamic data stream.
- (4) The use of a data-driven model provides
  Veracity ective, scientific, rational, and justifiable test of a hypothesis.

Why do all of this?

… for 4 very simple reasons:

- (1) Any real data table may consist of … billions of rows of numbers.

Volume — It is too much !

- (2) Any real data table will probably have … ds more) attributes (features), not just two.

Variety — It is too complex !

- (3) Humans can make mistakes when … of numbers, especially in a dynamic data stream.

Velocity — It keeps on coming !

- (4) The use of a data-driven model provides … justifiable test of a hypothesis.

Veracity — Can you prove your results ?

Rationale for BIG DATA – 1

- Consequently, if we collect a thorough set of parameters (high-dimensional data) for a complete set of items within our domain of study, then we would have a "perfect" statistical model for that domain.
- In other words, the data becomes the model.
- Anything we want to know about that domain is specified and encoded within the data.
- The goal of Data Science and Data Mining is to find those encodings, patterns, and knowledge nuggets.
- Recall what we said before …

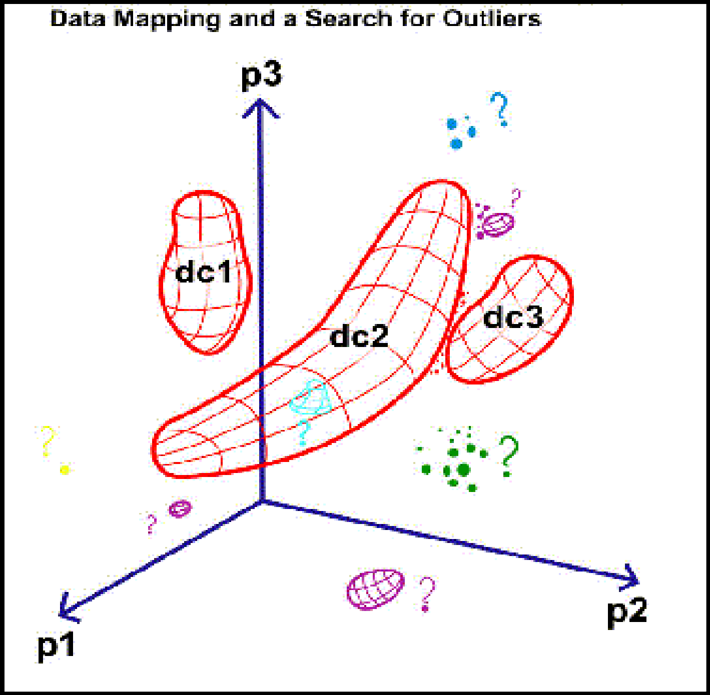
Rationale for BIG DATA – 2

- … **one of the two major benefits of BIG DATA is to provide the best statistical analysis ever(!) for the domain of study.**

Remember this :

Benefits of very large datasets:

1. best statistical analysis of “typical” events
2. automated search for “rare” events



Rationale for BIG DATA – 3

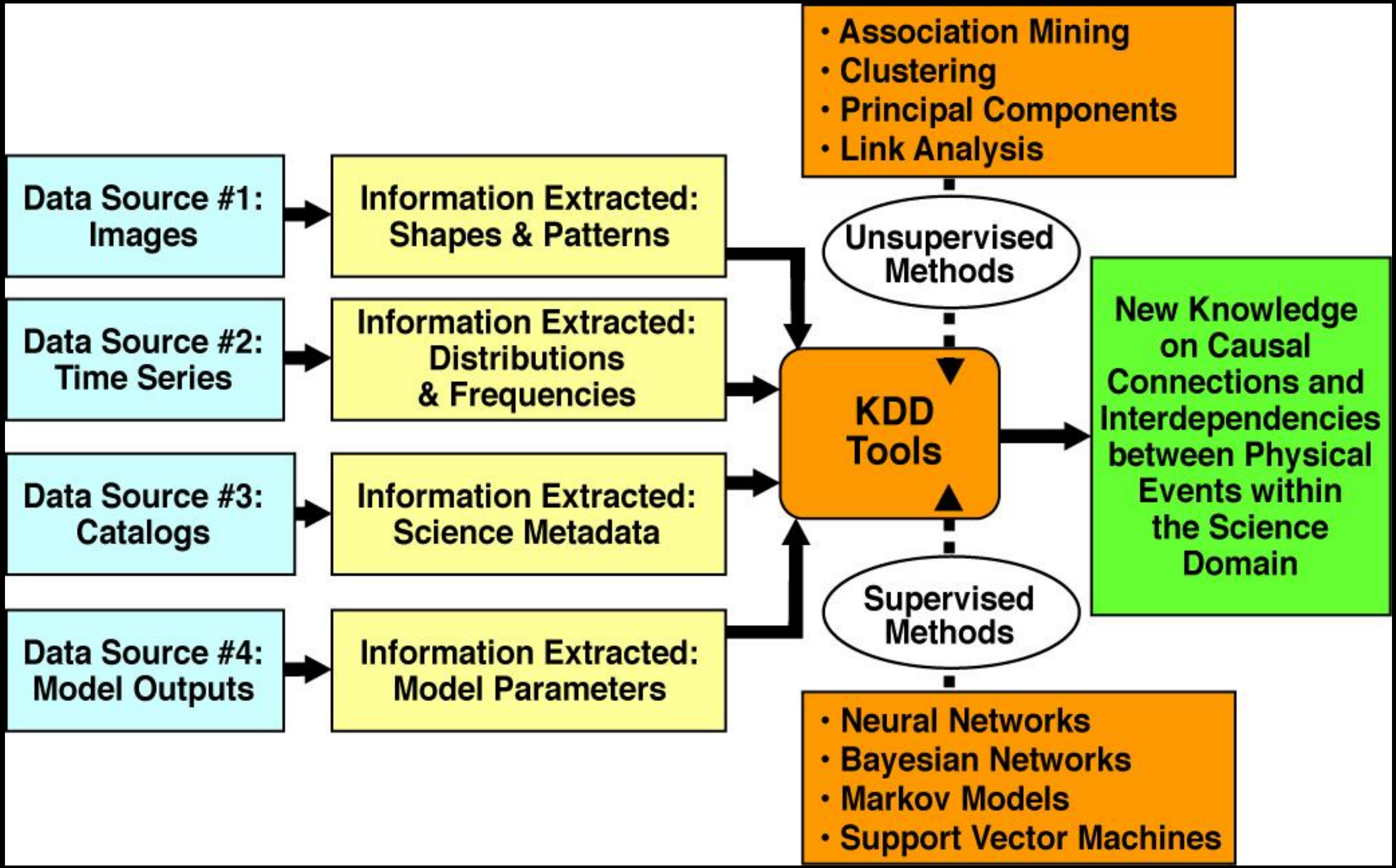
- Therefore, the 4th paradigm of science (which is the emerging data-oriented approach to any discipline X) is **different** from Experiment, Theory, and Computational Modeling.
  - *"Computational literacy and data literacy are critical for all."* – Kirk Borne
- A complete data collection on a domain (*e.g.,* the Earth, or the Universe, or the Human Body) encodes the knowledge of that domain, waiting to be mined and discovered.
  - *"Somewhere, something incredible is waiting to be known."* – Carl Sagan
- We call this "**X-Informatics**": addressing the D2K (Data-to-Knowledge) Challenge in any discipline X using Data Science.

- Examples: Bioinformatics, Geoinformatics, Astroinformatics, Climate Informatics, Ecological Informatics, Biodiversity Informatics, Environmental Informatics, Health Informatics, Medical Informatics, Neuroinformatics, Crystal Informatics, Cheminformatics, Discovery Informatics, and more …

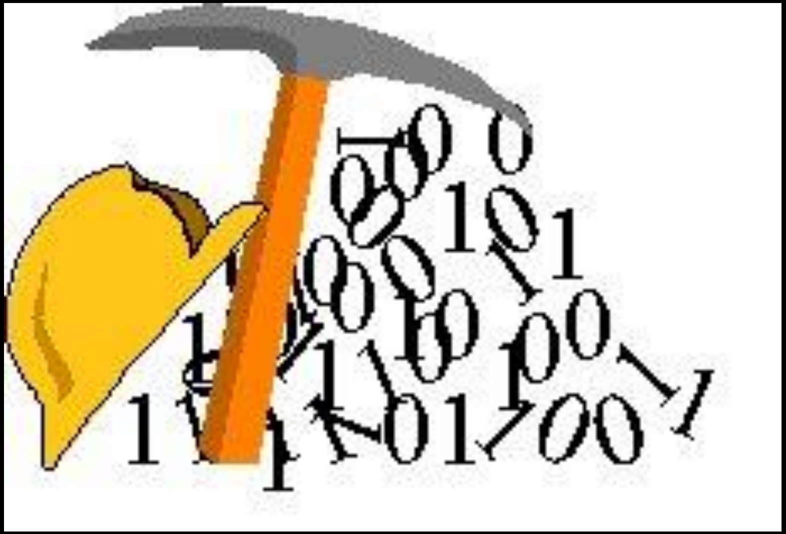
Addressing the D2K (Data-to-Knowledge) Challenge

Complete end-to-end application of informatics:

- Data management, metadata management, data search, information extraction, data mining, knowledge discovery
- All steps are necessary – skilled workforce needed to take data to knowledge
- Applies to any discipline (not just science)



Informatics in Education
and
An Education in Informatics



Data Science Education: Two Perspectives

- Informatics in Education – working with data in all learning settings
  - Informatics (Data Science) enables transparent reuse and analysis of data in inquiry-based classroom learning.
  - Learning is enhanced when students work with real data and information (especially online data) that are related to the topic (any topic) being studied.
  - http://serc.carleton.edu/usingdata/ ("Using Data in the Classroom")
  - Example: CSI The Cosmos
- An Education in Informatics – students are specifically trained:
  - … to access large distributed data repositories
  - … to conduct meaningful inquiries into the data
  - … to mine, visualize, and analyze the data
  - … to make objective data-driven inferences, discoveries, and decisions
- Numerous Data Science programs now exist at several universities (GMU, Caltech, RPI, Michigan, Cornell, U. Illinois, and more)
  - http://cds.gmu.edu/ (Computational & Data Sciences @ GMU)

Summary

- All enterprises are being inundated with data.
- The knowledge discovery potential from these data is enormous.
- Now is the time to implement data-oriented methodologies (Informatics) into the enterprise, to address the 4 Big Data Challenges from our "Tonnabytes" data collections: Volume, Variety, Velocity, and Veracity.
- This is especially important in training and degree programs – training the next-generation workers and practitioners to use data for knowledge discovery and decision support.
- We have before us a grand opportunity to establish dialogue and information-sharing across diverse data-intensive research and application communities.