



Big Data for Government Symposium

<http://www.ttcus.com>



@TECHTrain

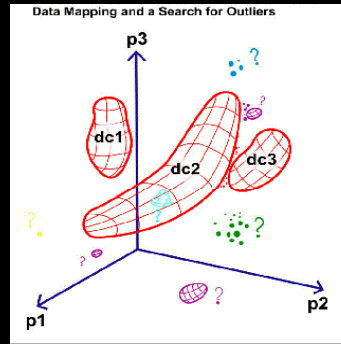
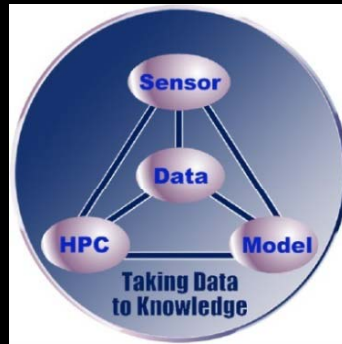


TTC™

Technology Training Corporation

Linkedin/Groups:
Technology Training
Corporation





Big Data in Space and Earth Sciences

Kirk Borne



[@KirkDBorne](https://twitter.com/@KirkDBorne)



School of Physics, Astronomy, & Computational Sciences
College of Science, George Mason University, Fairfax, VA

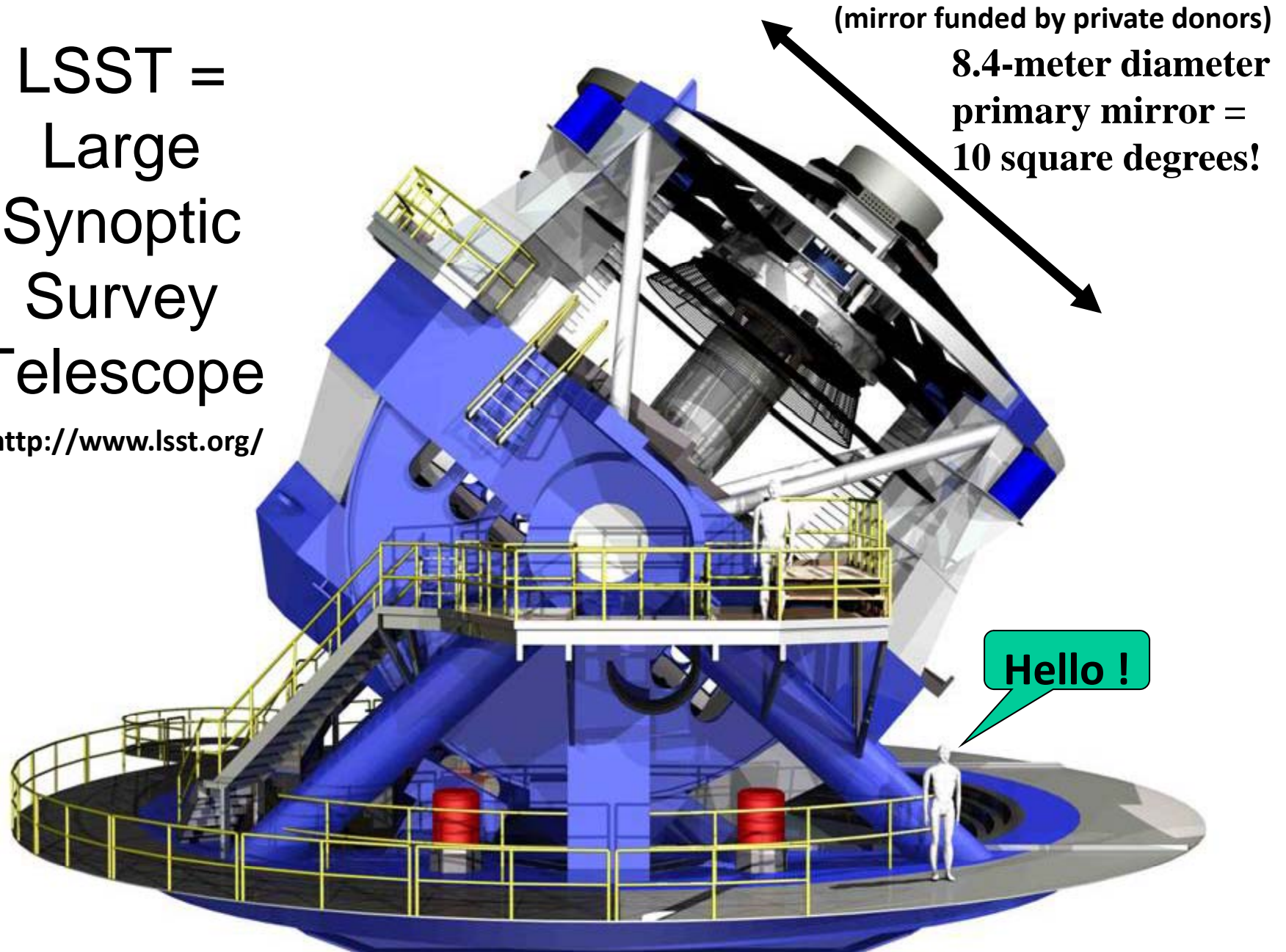
Astronomy Example

Let us begin with a space science example ...

The LSST (Large Synoptic Survey Telescope)

LSST = Large Synoptic Survey Telescope

<http://www.lsst.org/>



LSST =
Large
Synoptic
Survey
Telescope

<http://www.lsst.org/>

(mirror funded by private donors)

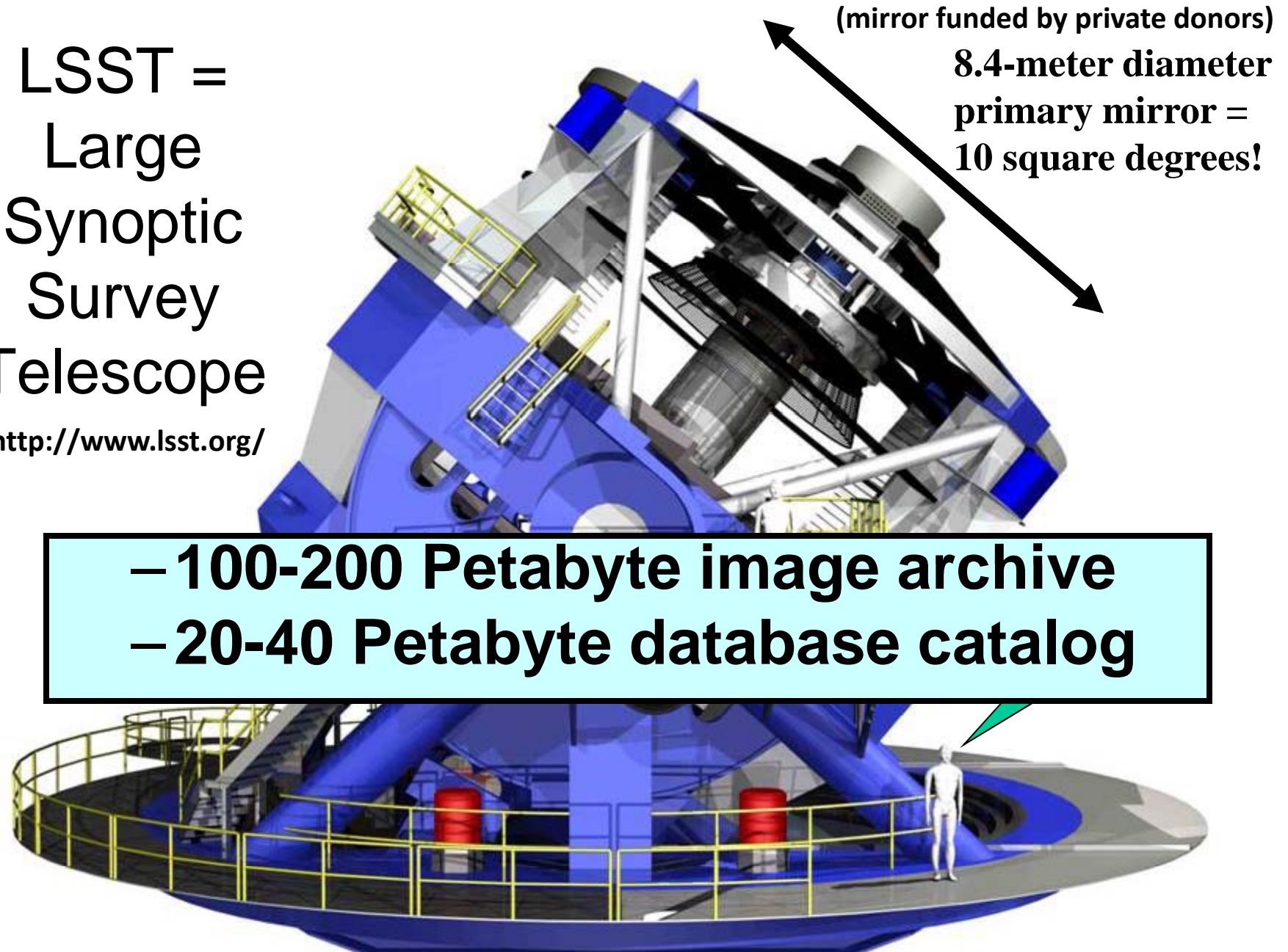
8.4-meter diameter
primary mirror =
10 square degrees!

Construction begins July 1, 2014

Hello !

LSST = Large Synoptic Survey Telescope

<http://www.lsst.org/>



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)



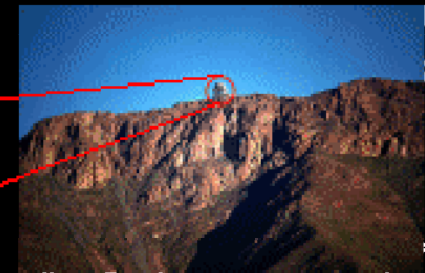
South America



Chile



Region de Coquimbo



LSST in time and space:

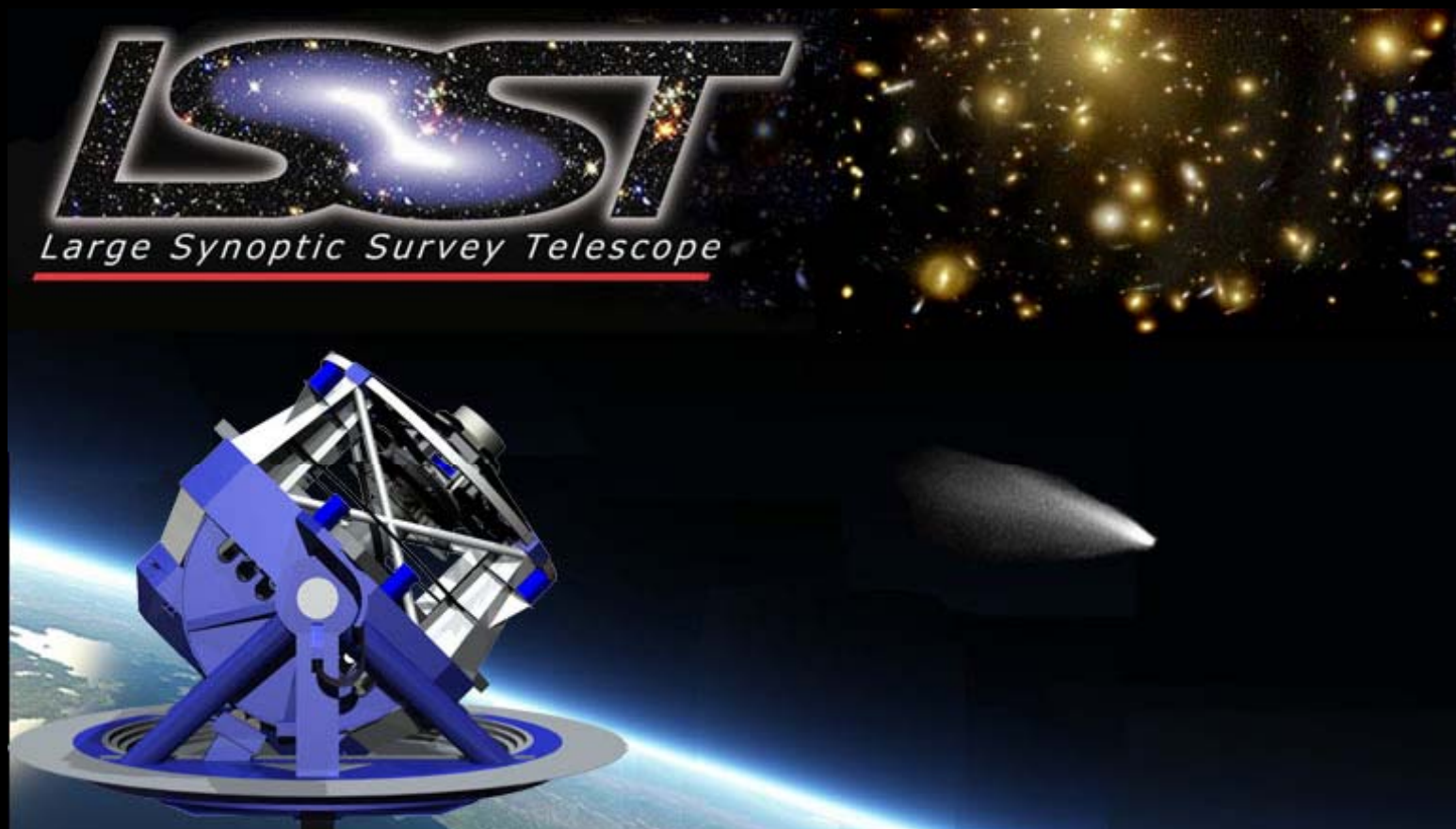
- When? ~2022-2032
- Where? Cerro Pachon, Chile

Architect's design
of LSST Observatory



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 (~2022-2032), with time domain sampling in log(time) intervals (to capture dynamic range of transients).

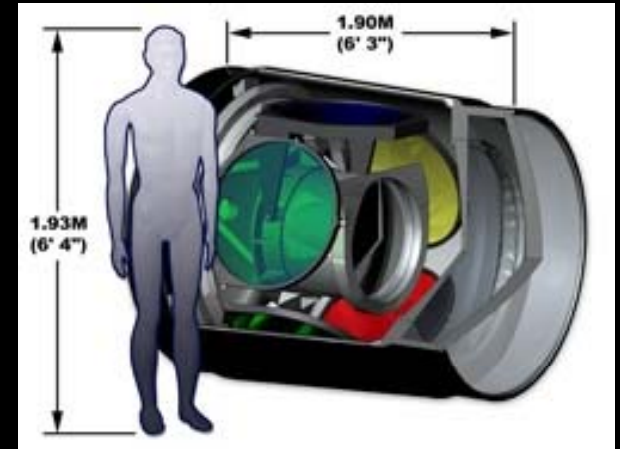
- **LSST (Large Synoptic Survey Telescope):**
 - Ten-year time series imaging of the night sky – mapping the Universe !
 - **~10,000,000 events each night** – *anything that goes bump in the night !*
 - **Cosmic Cinematography! The New Sky! @ <http://www.lsst.org/>**



LSST Summary

<http://www.lsst.org/>

- 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: ~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**



The LSST Big Data Challenges



10,000,000 events
every night

100 PB image
archive

50 billion object
database

20 PB science
catalog

What is Big Data Science good for?

- ✓ Discovery
- ✓ Data-driven Decision Support

Characteristics of Big Data Science

- The emergence of **Data Science** and **Data-Oriented Science** (the 4th paradigm of science).
- A complete data collection on any complex domain (*e.g.*, Earth, or the Universe, or the Human Body) has the potential to encode the knowledge of that domain, waiting to be mined and discovered.
- We call this “**X-Informatics**”: addressing the D2K (Data-to-Knowledge) Challenge in any discipline X using Data Science.
- Examples: **Astroinformatics**, Bioinformatics, Geoinformatics, Climate Informatics, Ecological Informatics, Biodiversity Informatics, Environmental Informatics, Health Informatics, Medical Informatics, Neuroinformatics, Crystal Informatics, Cheminformatics, Discovery Informatics, and more.

Rationale for Big Data Science

- 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, Big Data becomes the model for a domain X = we call this X -informatics.
- Anything we want to know about that domain is specified and encoded within the data.
- The goal of Big Data Science is to find those encodings, patterns, and knowledge nuggets.
- See article: [Big-Data Vision? Whole-population analytics](#)
 - From Reporting to Prediction
 - From hindsight & oversight to insight & foresight
 - From DATA to INFORMATION to KNOWLEDGE to UNDERSTANDING



Characterizing and Exposing the Big Data Hype: 3 V's or ?

<http://bit.ly/1hH6sB9>

- If the only distinguishing characteristic was that we have lots of data, we would call it **"Lots of Data"** (or a ***Tonnabytes!***)
- Big Data characteristics: **the 3+n V's** =
 1. **Volume** (*lots of data = "Tonnabytes"*)
 2. **Variety** (*complexity, curse of dimensionality, many formats*)
 3. **Velocity** (*high rate of data and information flow, real-time, incoming!*)
 4. **Veracity** (*necessary & sufficient data to test many hypotheses*)
 5. **Validity** (*data quality, governance, master data management*)
 6. **Value** (= *the all-important V!*)
 7. **Variability** (*dynamic, evolving, spatiotemporal data, time series*)
 8. **Venue** (*distributed, heterogeneous, multiple platforms/owners*)
 9. **Vocabulary** (*ontologies, semantics, schema, data models,...*)
 10. **Vagueness** (*confusion over the meaning of Big Data, tools, methods,...*)

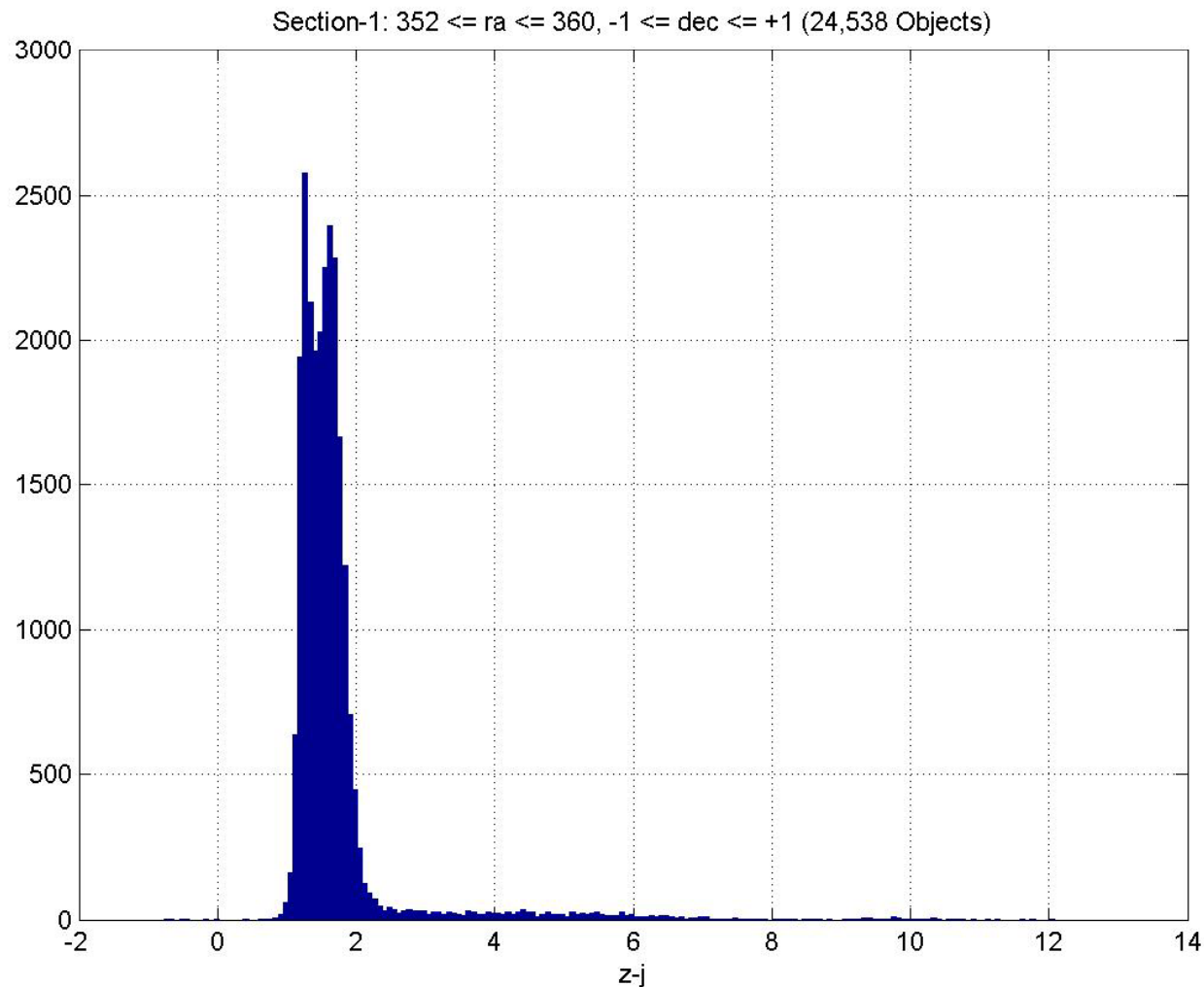


Characterizing and Exposing the Big Data Hype: 3 V's or ?

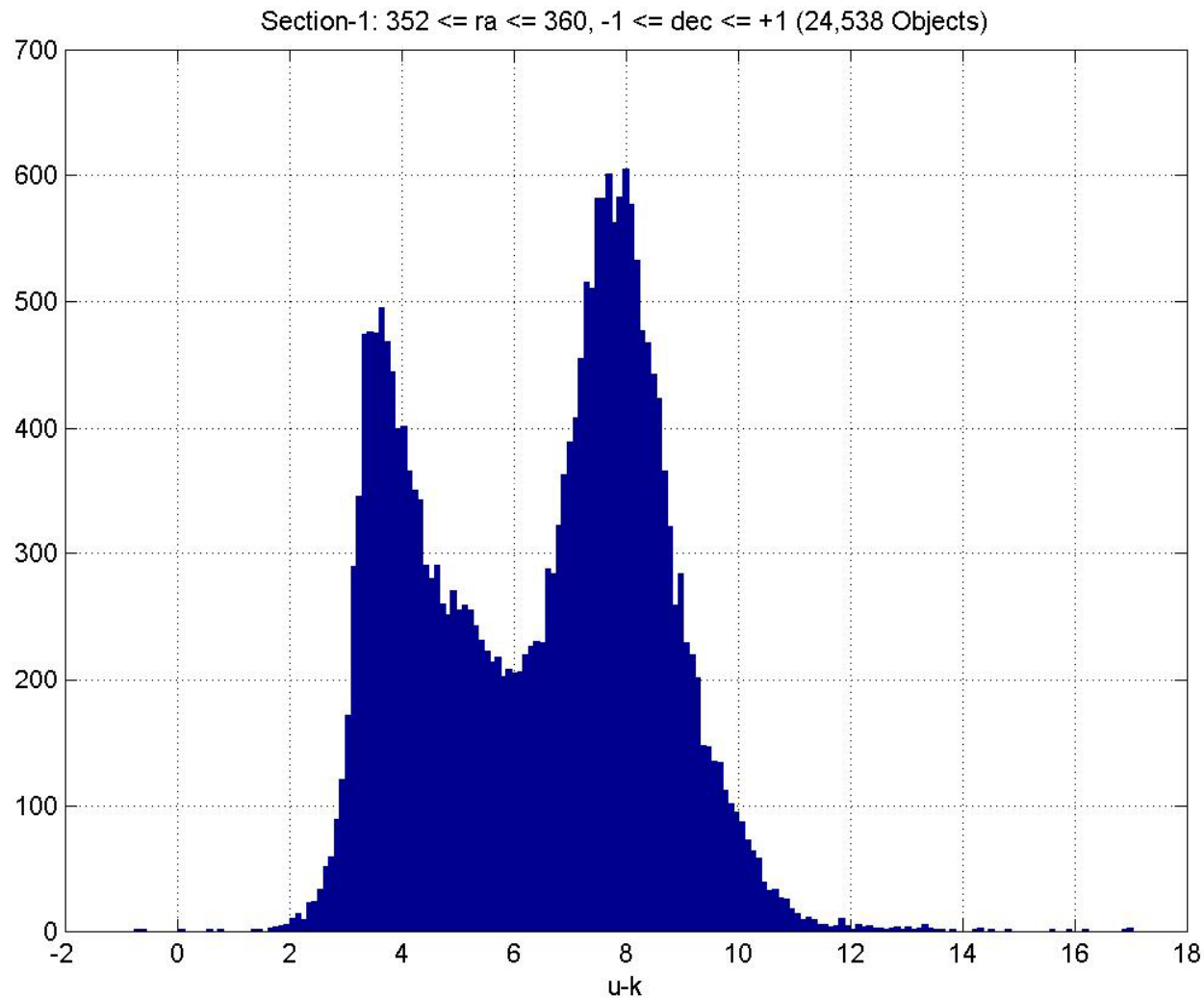
<http://bit.ly/1hH6sB9>

- If the only distinguishing characteristic was that we have lots of data, we would call it “**Lots of Data**” (or a ***Tonnabytes!***)
- Big Data characteristics: **the 3+n V's** =
 1. **Volume** (*lots of data = “Tonnabytes”*)
 2. **Variety** (*complexity, curse of dimensionality, many formats*)
 3. **Velocity** (*high rate of data and information flow, real-time, incoming!*)
 4. **Veracity** (*necessary & sufficient data to test many hypotheses*)
 5. **Validity** (*data quality, governance, master data management*)
 6. **Value** (*= the all-important V!*)
 7. **Variability** (*dynamic, evolving, spatiotemporal data, time series*)
 8. **Venue** (*distributed, heterogeneous, multiple platforms/owners*)
 9. **Vocabulary** (*ontologies, semantics, schema, data models,...*)
 10. **Vagueness** (*confusion over the meaning of Big Data, tools, methods,...*)

Insufficient Variety: stars & galaxies are not separated in this parameter

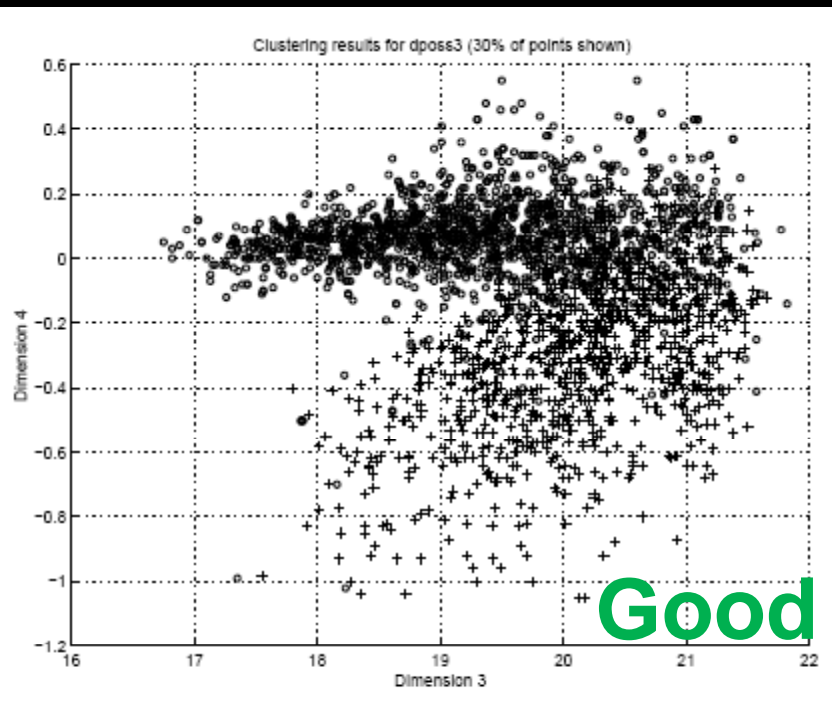
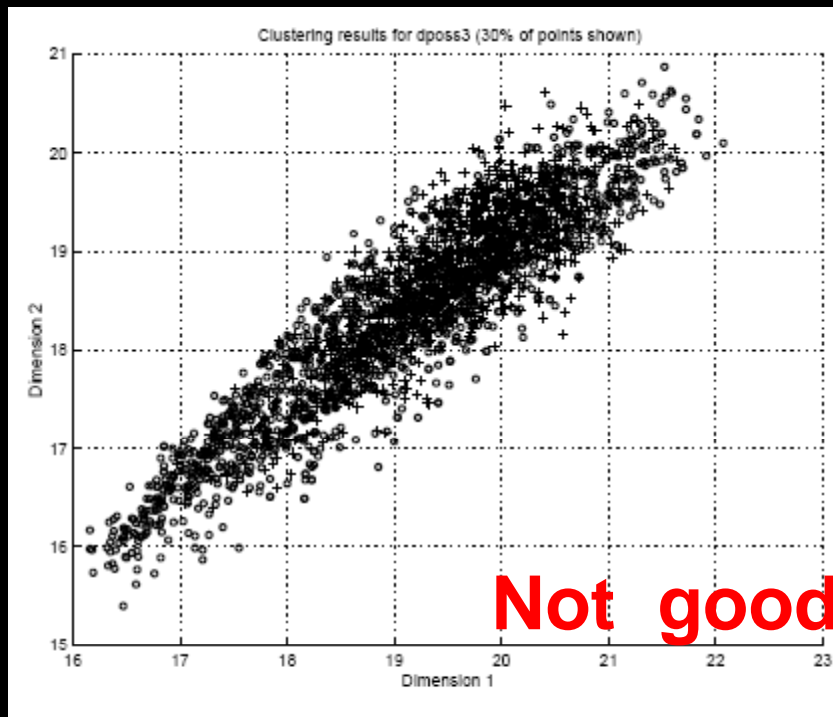
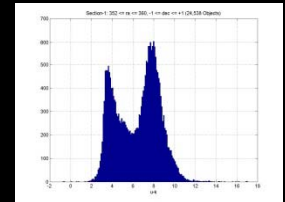
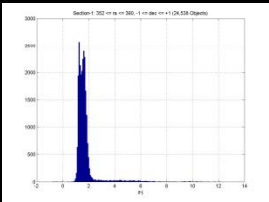


Sufficient Variety: stars & galaxies are separated in this parameter



The 3 important D's of Big Data Variety: feature Disambiguation, Discrimination between multiple classes, and Discovery of new classes.

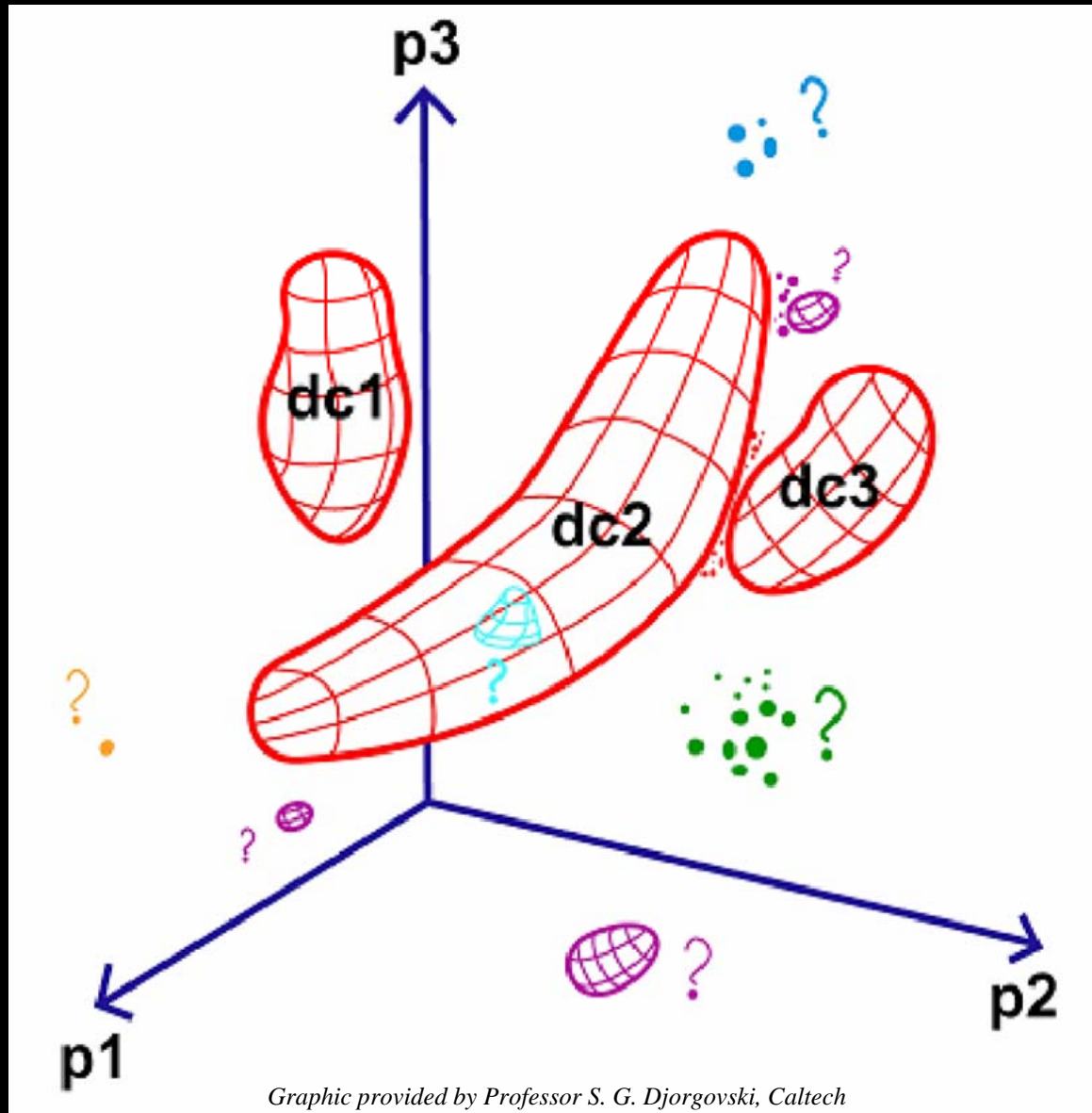
The separation and discovery of classes improves when a sufficient number of "correct" features are available for exploration and testing, as in the following two-class discrimination test:



<http://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/BrunnerDPS.pdf>

This graph says it all ...

3 Steps to Discovery – Data Mining your Big Data



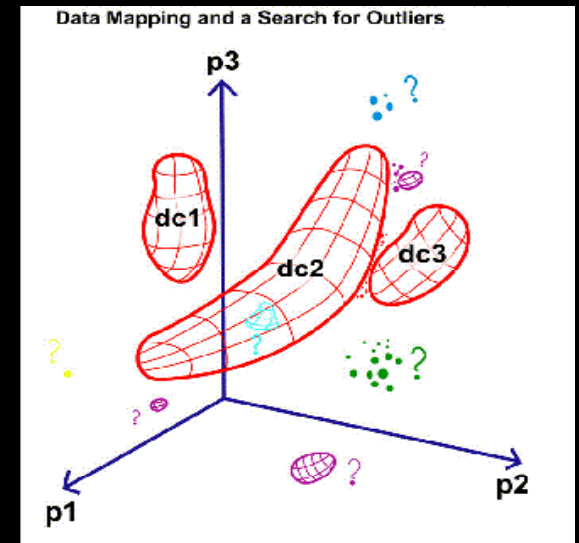
Graphic provided by Professor S. G. Djorgovski, Caltech

- **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 !**

Data-Driven Discovery:

(KDD: Knowledge Discovery from Data)

1. Class Discovery
2. Correlation Discovery
3. Novelty Discovery
4. Association Discovery
(Finding unusual (improbable) co-occurring associations)

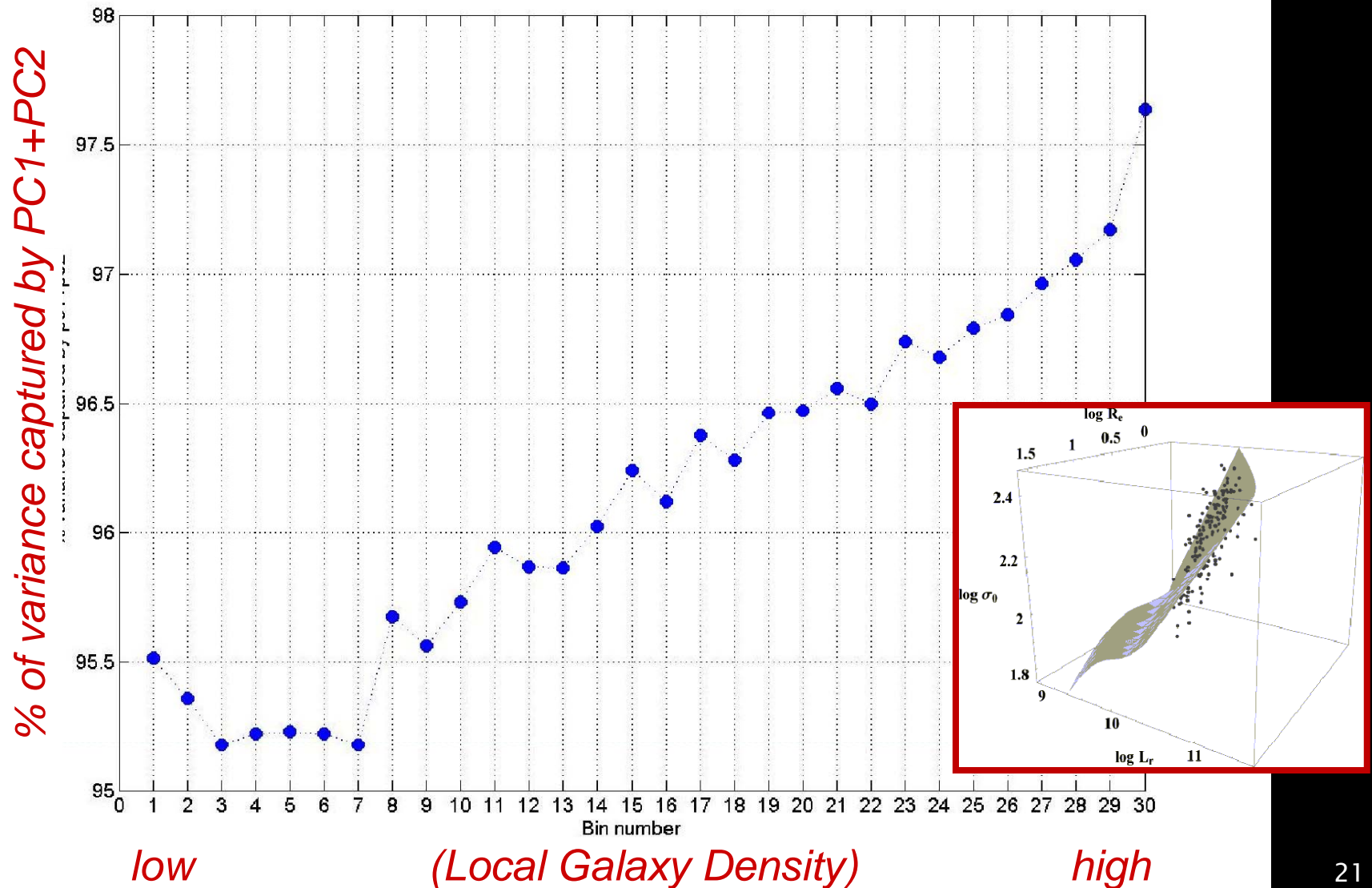


Graphic from S. G. Djorgovski

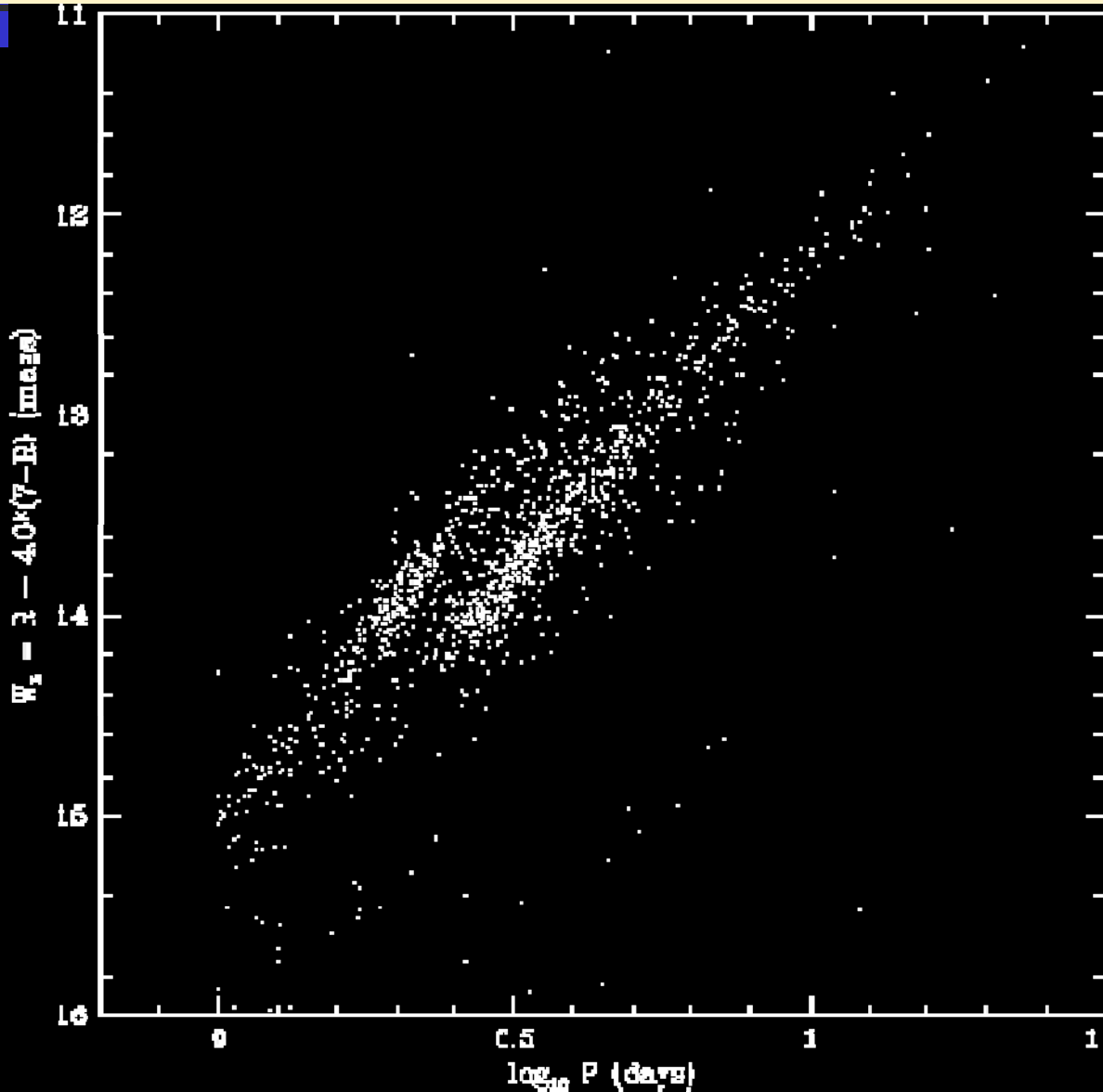
- Benefits of very large datasets:
 - best statistical analysis of “typical” events
 - automated search for “rare” events

Correlation Discovery: Fundamental Plane for 156,000 cross-matched Sloan+2MASS Elliptical Galaxies: plot shows variance captured by first two Principal Components as a function of local galaxy density.

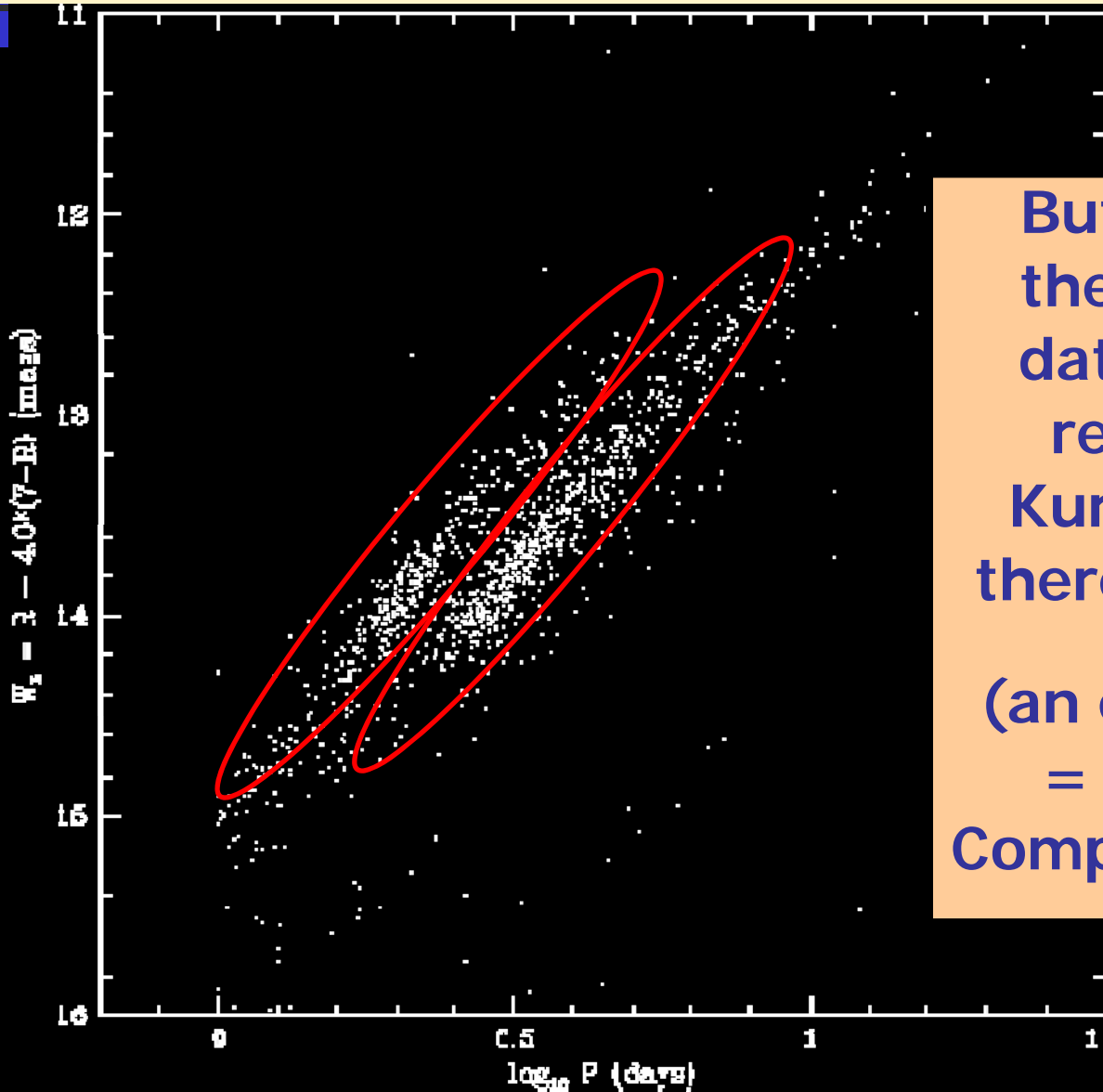
Reference: Borne, Dutta, Giannella, Kargupta, & Griffin 2008



Initial impression is that the data are extended in only one direction (principal component)



Initial impression is that the data are extended in only one direction (principal component)



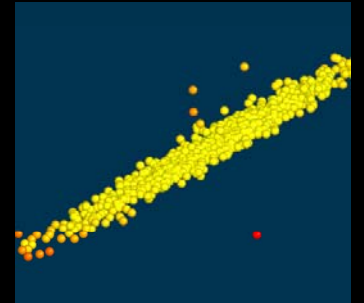
But, a cut across the width of the data distribution reveals a high Kurtosis ... hence there are 2 Classes!

(an example of ICA = Independent Component Analysis)

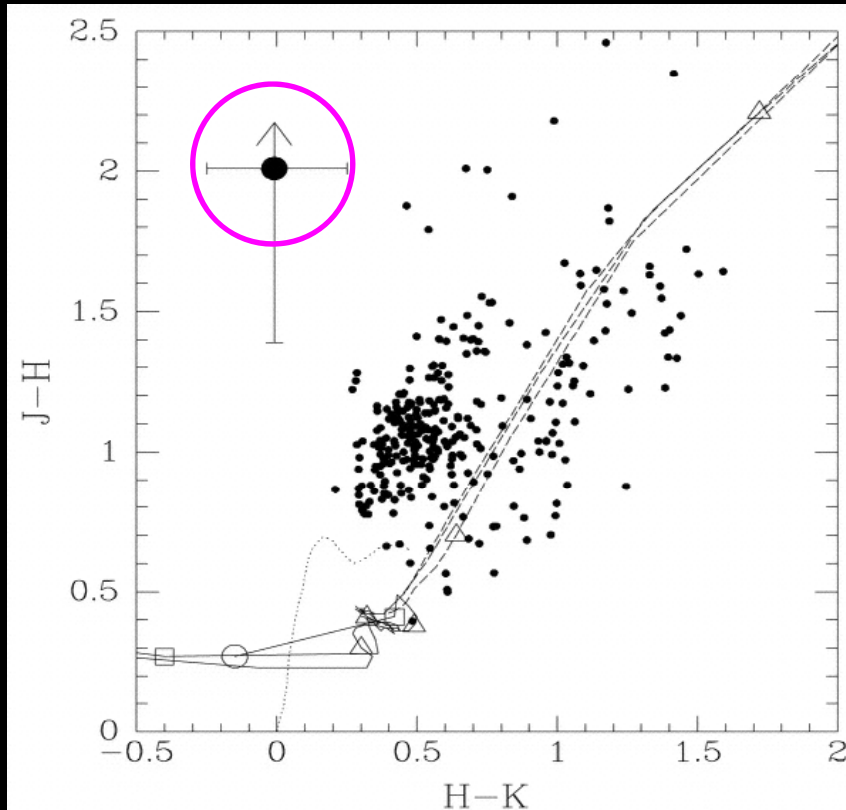
Novelty Detection = Surprise Discovery!

- **Outlier detection: (unknown unknowns)**

- Anomaly Detection, Deviation Detection, **Surprise Discovery**, Novelty Discovery: Finding objects and events that are outside the bounds of our expectations (outside known clusters)
- Finding new, rare, one-in-a-million(billion)(trillion) objects and events
- These may be real scientific discoveries or garbage
- Outlier detection is therefore useful for:
 - Anomaly Detection – *is the detector system working?*
 - Data Quality Assurance – *is the data pipeline working?*
 - Novelty Discovery – *is my Nobel prize waiting?*
- 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”?



Novelty Detection = improved discovery of rare objects across multiple databases



2MASSW J1217-03

A methane (T-type) dwarf in the constellation Virgo

The near-infrared view

2MASS Composite JHK_s Atlas Image

The optical view

Palomar Digitized Sky Survey

A.J. Burgasser (Caltech), J.D. Kirkpatrick (IPAC/Caltech), M.F. Brown (Caltech),
 L.N. Reid (U. Penn.), J.L. Gizis (U. Mass.), C.C. Dahn & D.G. Monet (USNO, Flagstaff),
 C.A. Beichman (JPL), J.L. Liebert (Arizona), R.M. Cutri (IPAC/Caltech), M.F. Skrutskie (U. Mass.)

The 2MASS Project is a collaboration between the University of Massachusetts and IPAC



Association Discovery = finding interesting co-occurring associations

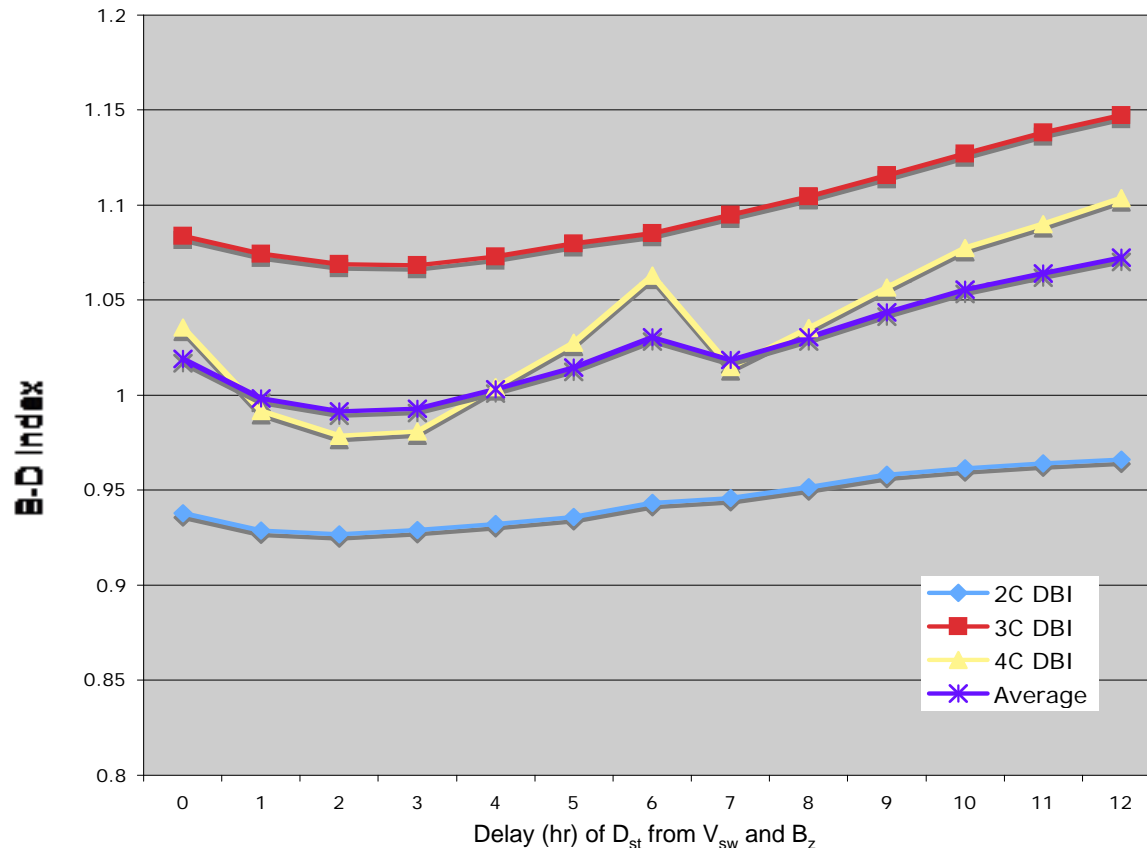
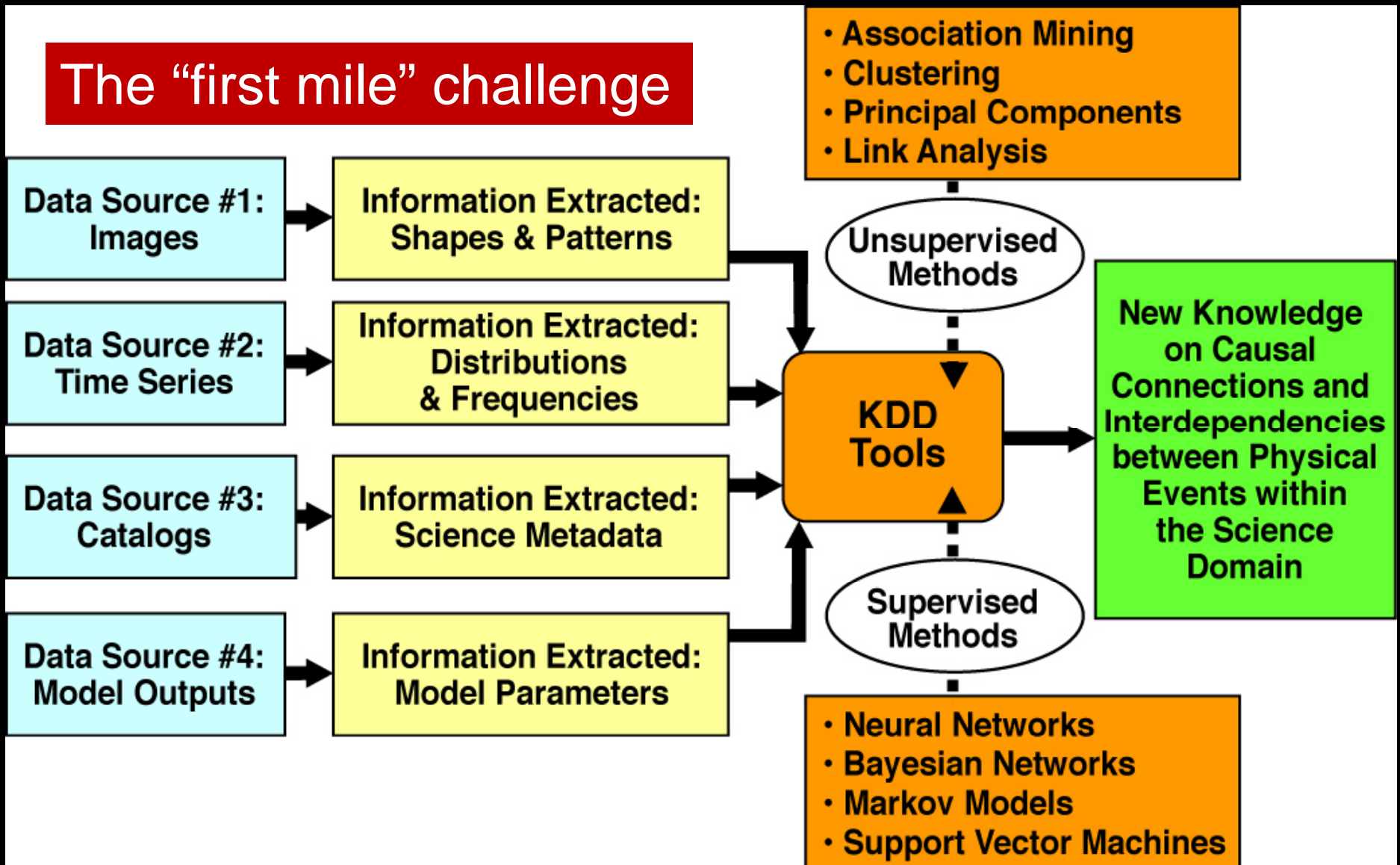


Figure 10. Davies-Bouldin index for various time delays of D_{st} from V_{sw} and B_z for cases of 2 (blue), 3 (red), 4 (yellow) clusters, and the overall average (purple), indicating an optimal delay of ~2-3 hours for D_{st} .

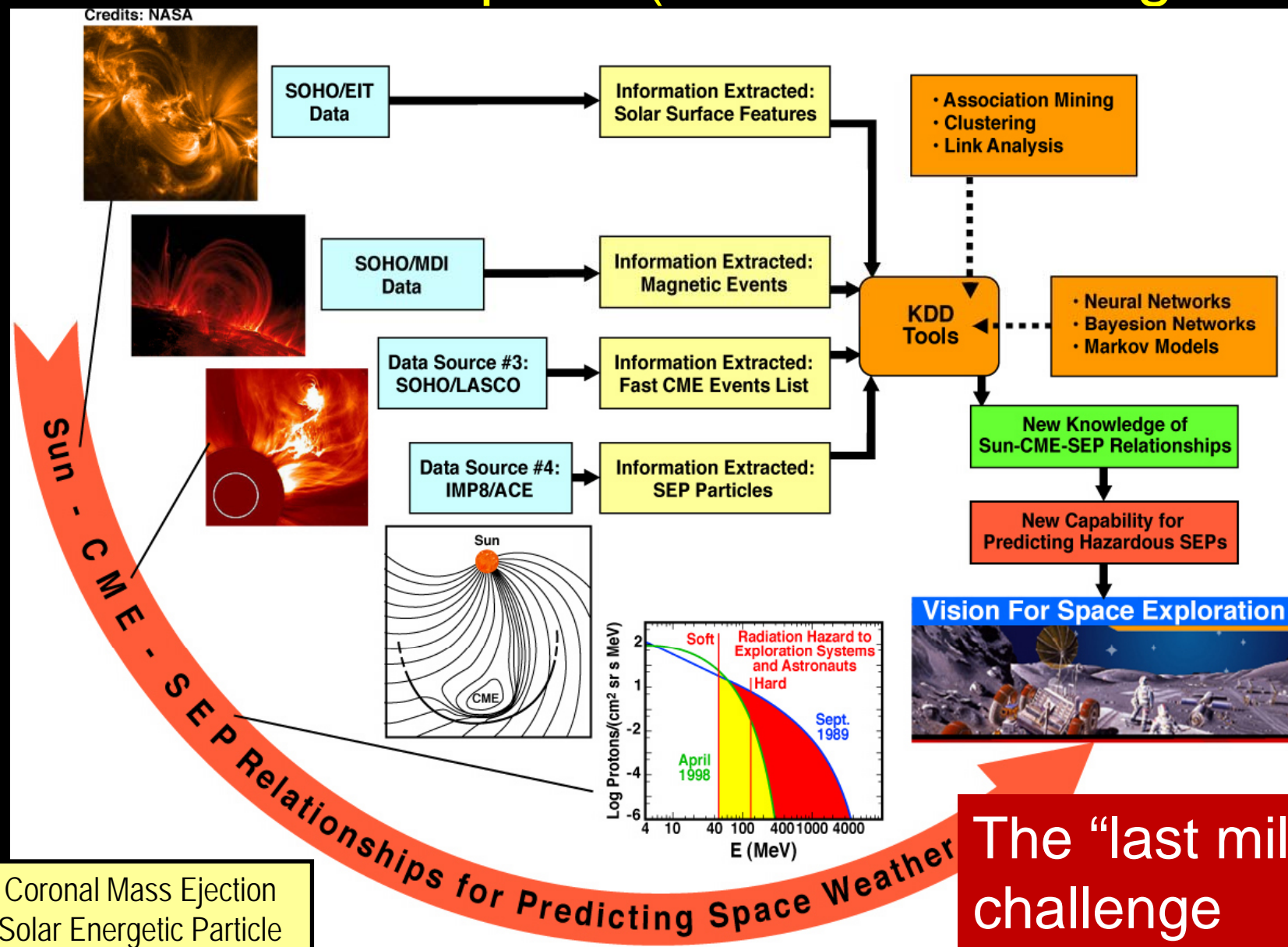
Solar wind properties have the strongest association with space plasma responses within the Earth's magnetosphere about 2-4 hours after a major plasma outburst occurs on the Sun.

Knowledge Discovery for Multi-source Data: Heterogeneous data collections are the new normal.

The “first mile” challenge



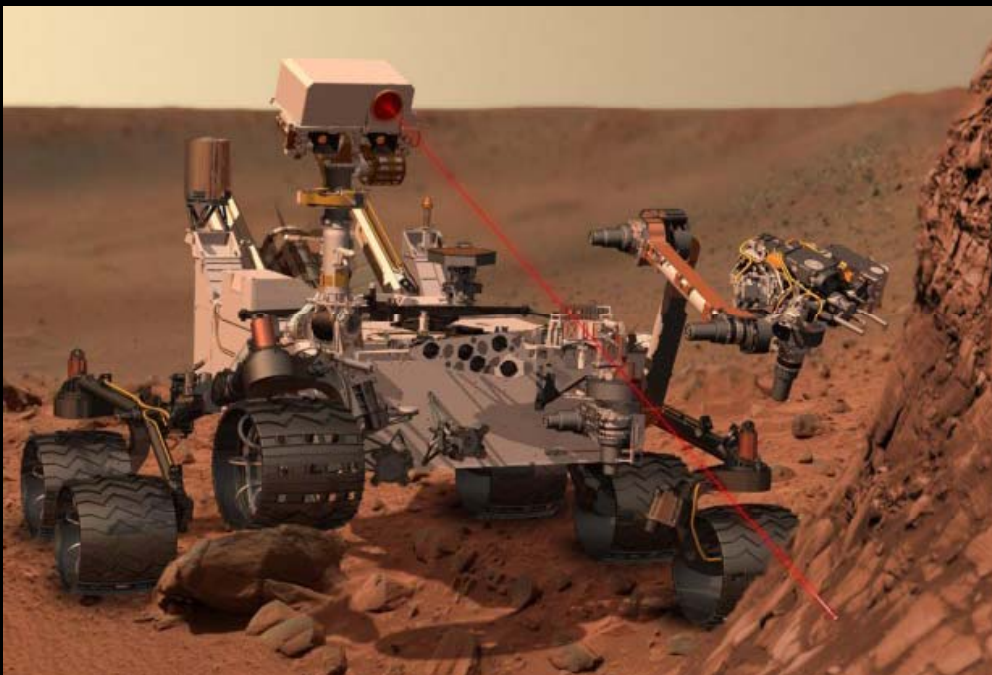
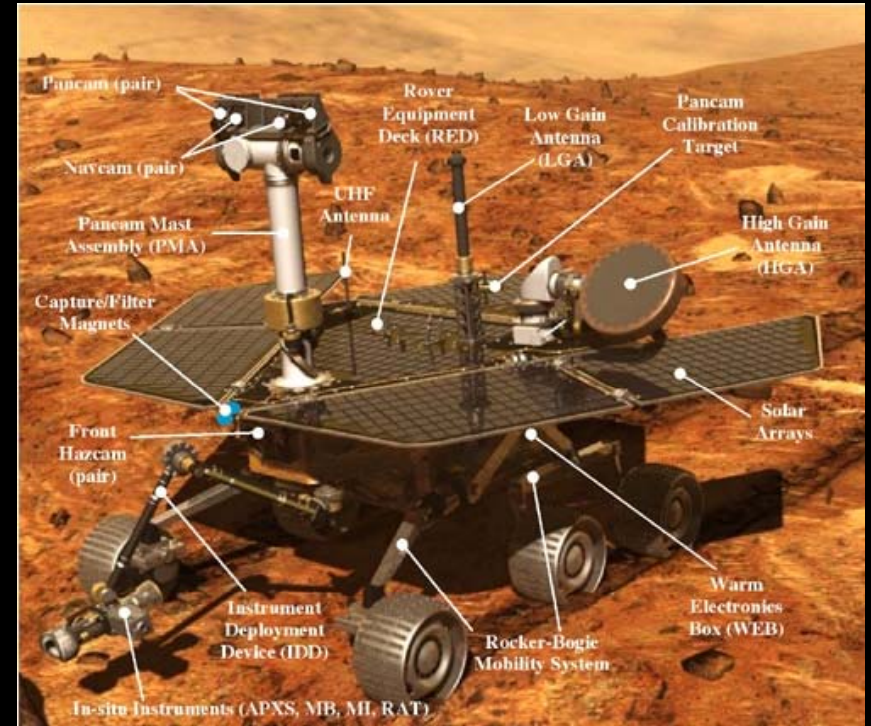
Space Weather Example: Early Warning System for Astronauts in Space (actionable intelligence!)



What is Big Data Science good for?

- ✓ Discovery
- ✓ Data-driven Decision Support

Case Study - Mars Rovers



Mars Rover: intelligent data-gatherer, mobile data mining agent, and autonomous science decision-support system

Rove around the surface of Mars and take samples of rocks (experimental technique: mass spectroscopy = data histogram)

Intelligent Data Operations in Action:

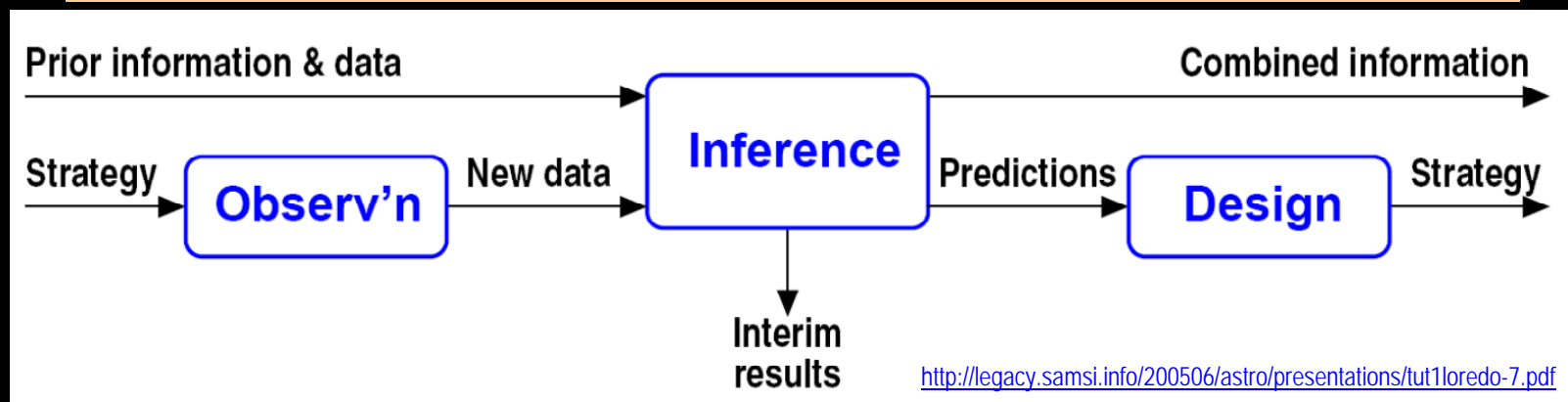
- **Supervised Learning** (search for rocks with known compositions)
- **Unsupervised Learning** (discover what types of rocks are present, without preconceived biases)
- **Association Mining** (find unusual associations)
- **Clustering** (find the set of unique classes of rocks)
- **Classification** (assign rocks to known classes)
- **Deviation/Outlier Detection** (one-of-kind; interesting?)
- On-board Intelligent Data Understanding & Decision Support Systems (**Fuzzy Logic** & **Decision Trees** & **Cased-Based Reasoning**) =
= **Science Goal Monitoring** :
 - *“stay here and do more”* ; or else *“move on to another rock”*
 - *“send results to Earth immediately”* ; or *“send results later”*

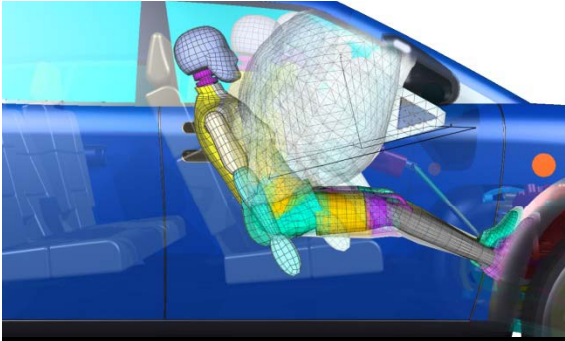
Mars Rover = Decision Science Engine

(enabling data-to-decisions = DTD)

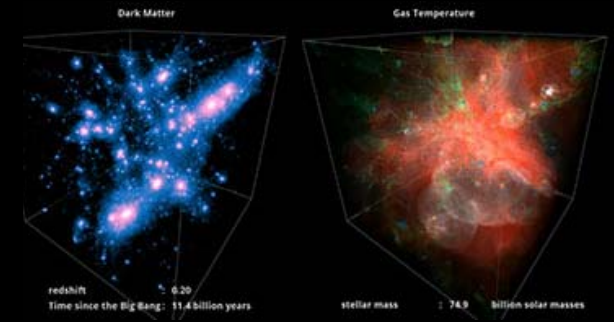


- Decisions are based on data mined, prior experience, new knowledge, and the set of learned rules.
- Rover acts autonomously, without human intervention, in Deep Space environment.
- Actions are driven by mining actionable data from all sensors.





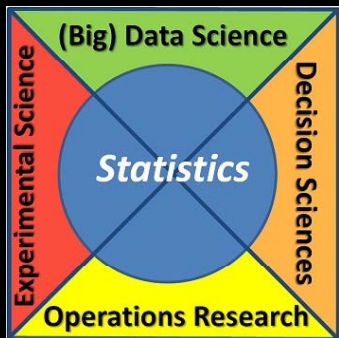
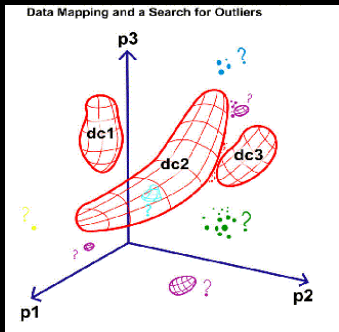
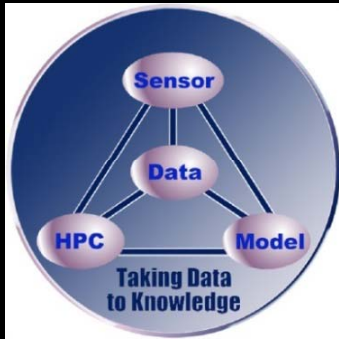
Big Data Science meets Big Science Simulations



- Cosmology (colliding galaxies: crash science)
- Climate Science
- Fusion 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



The Big Picture of Big Data in Space and Earth Sciences

✓ Knowledge Discovery

- BD2K = Big Data-to-Knowledge <http://bd2k.nih.gov>
- Class discovery: predictive power discovery
- Association discovery
- Correlation discovery
- Novelty discovery: *surprise!*

✓ Data-driven Decision Support

- Data-to-Decisions (DTD)
- Predictive & Prescriptive Analytics
- The Last Mile Challenge (Actionable Intelligence)
- Decision Science-as-a-Service™ <http://www.syntasa.com/>

✓ Big ROI (Return On Innovation) !!!