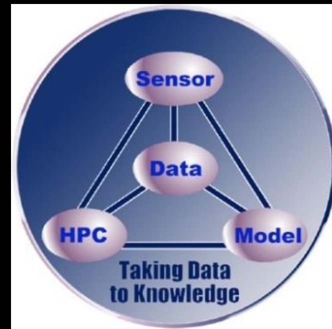
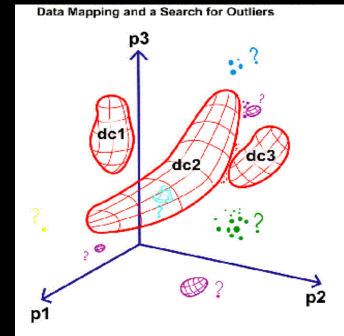
GEORGE MASON UNIVERSITY





What is Data Science and Why is it Needed?

Learning From Data, Big and Small

Kirk Borne

@KirkDBorne

School of Physics, Astronomy, & Computational Sciences

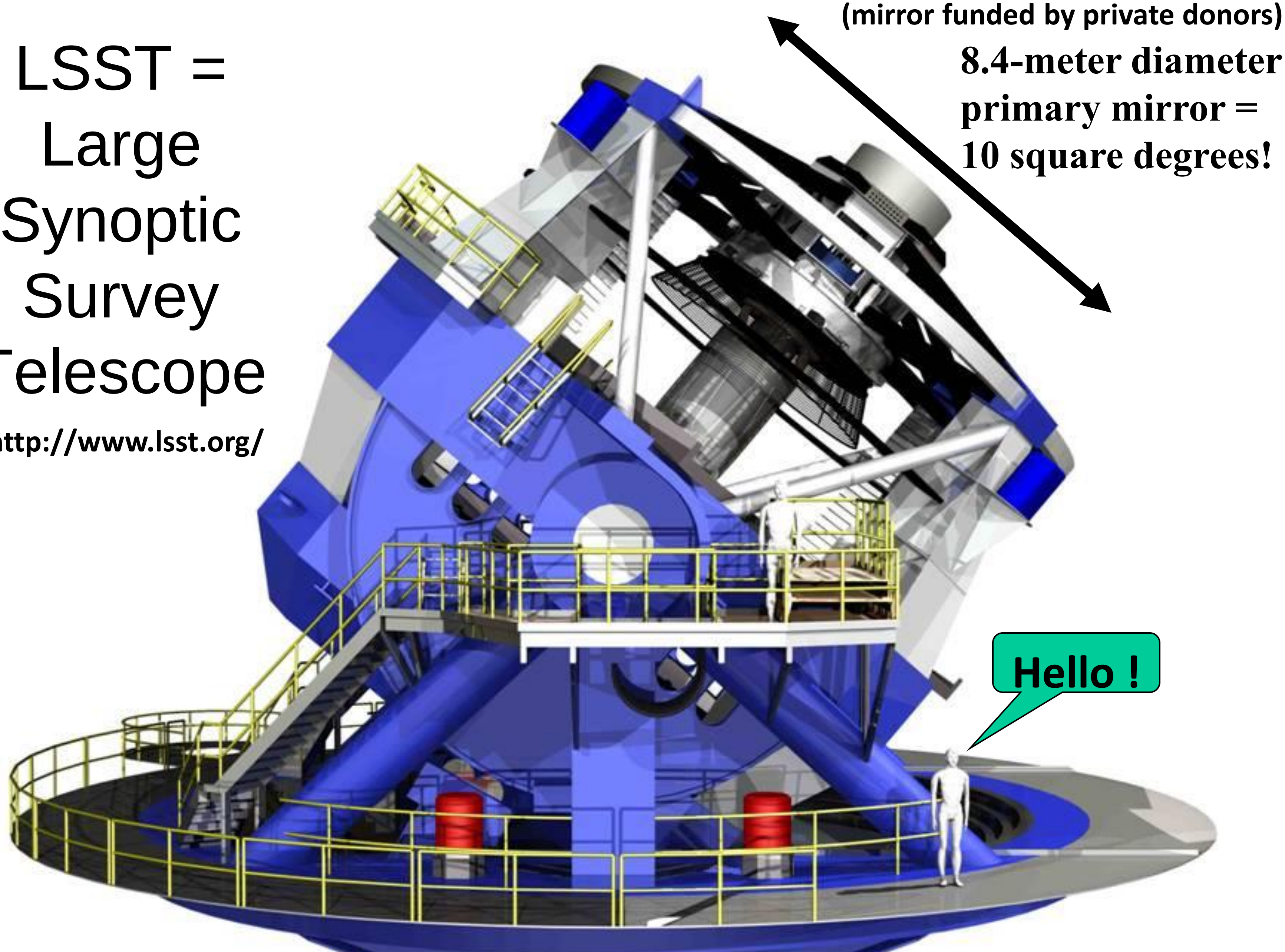
http://www.onalytica.com/blog/posts/onalytica-big-data-influencers-q4-13

Astronomy Example

- Before we look at Big Data and Data Science…
- ... Let us look at an astronomy example …
- The LSST (Large Synoptic Survey Telescope)
- … GMU is a partner institution and our scientists are involved with the science, data management, and education programs of the LSST

LSST = Large Synoptic Survey Telescope

http://www.lsst.org/

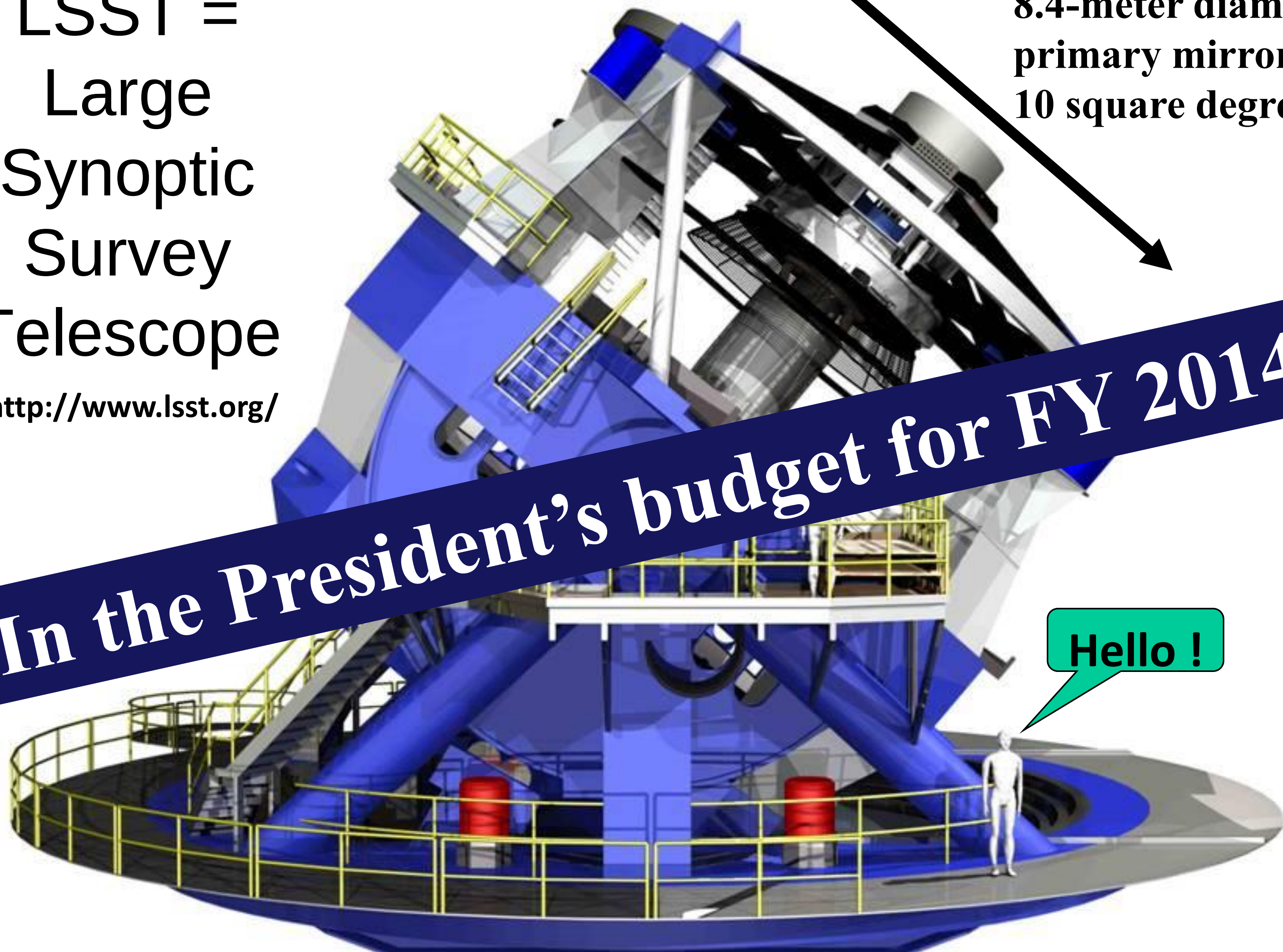


LSST = Large Synoptic Survey Telescope

http://www.lsst.org/

(mirror funded by private donors)

8.4-meter diameter primary mirror = 10 square degrees!



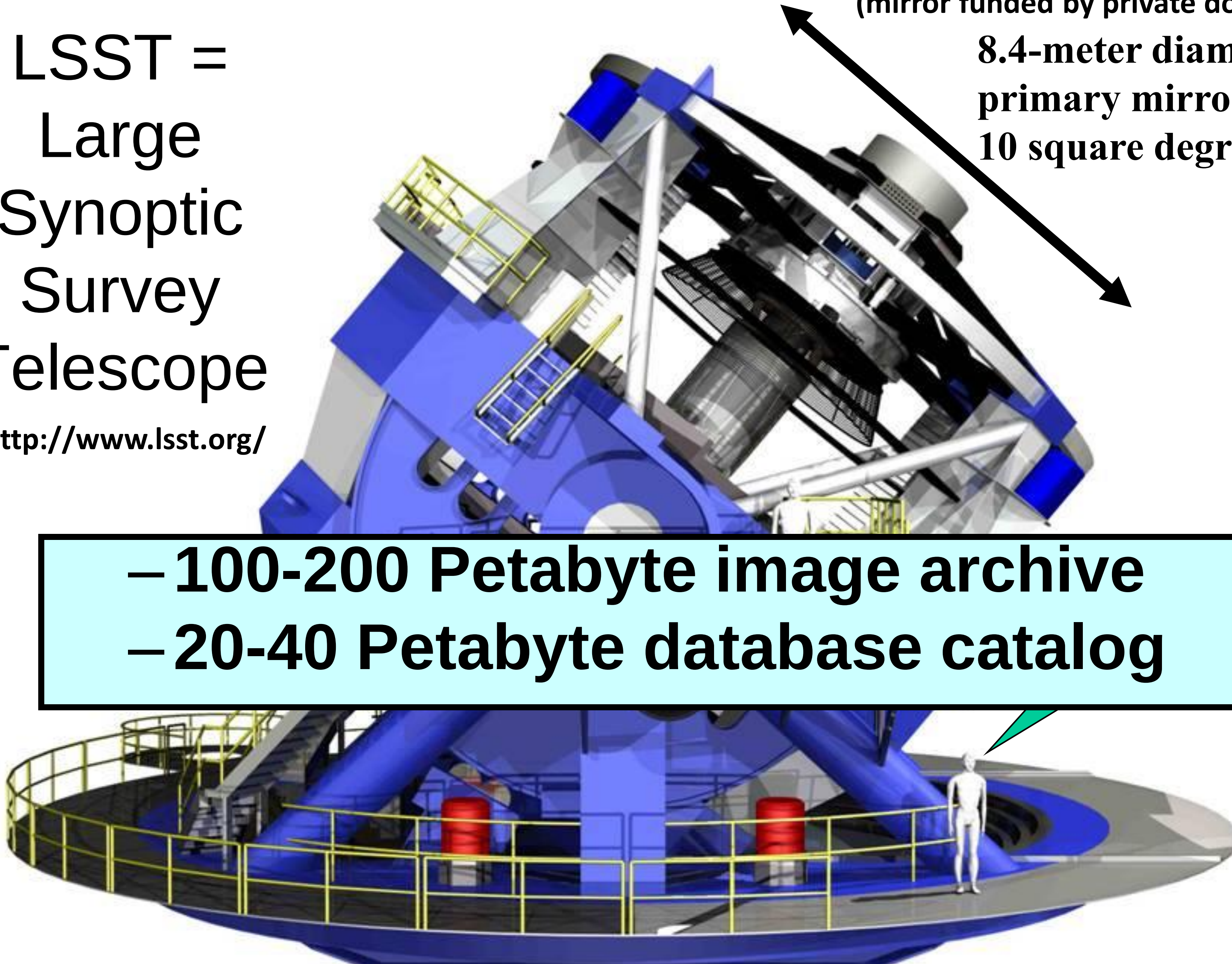
In the President's budget for FY 2014

LSST =
Large
Synoptic
Survey
Telescope

http://www.lsst.org/

(mirror funded by private donors)

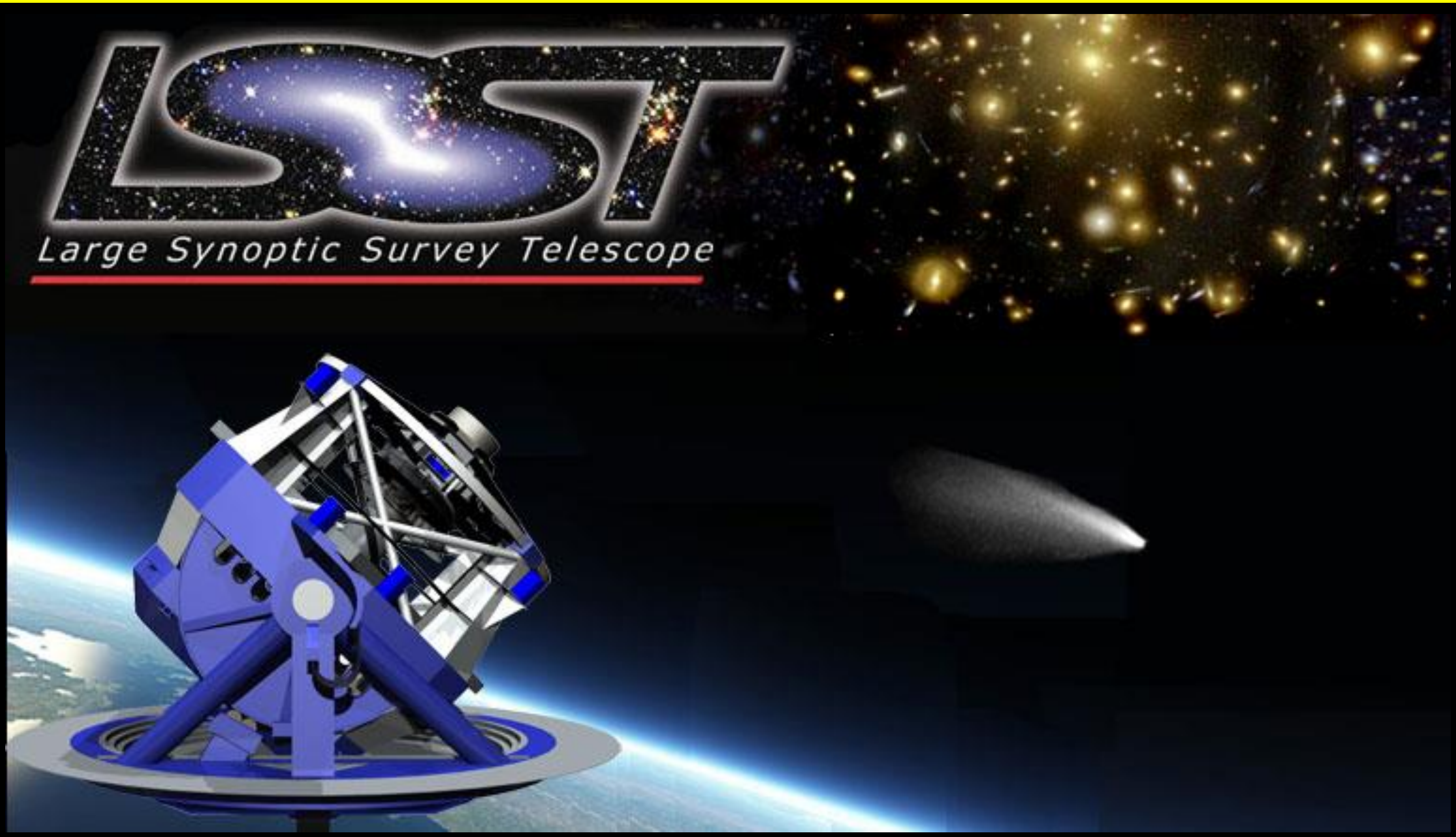
8.4-meter diameter primary mirror = 10 square degrees!



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

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 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

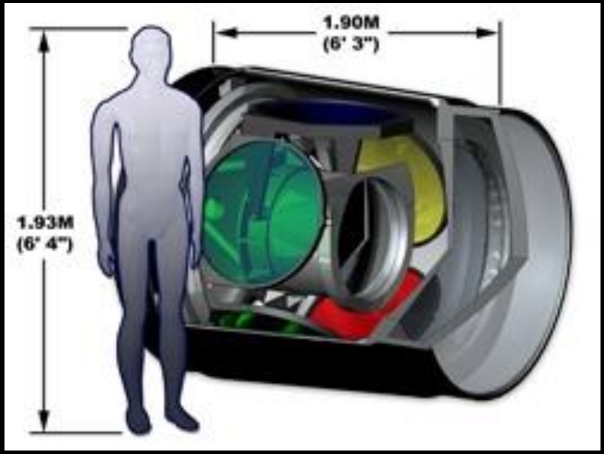
Chile

Region de Coquimbo

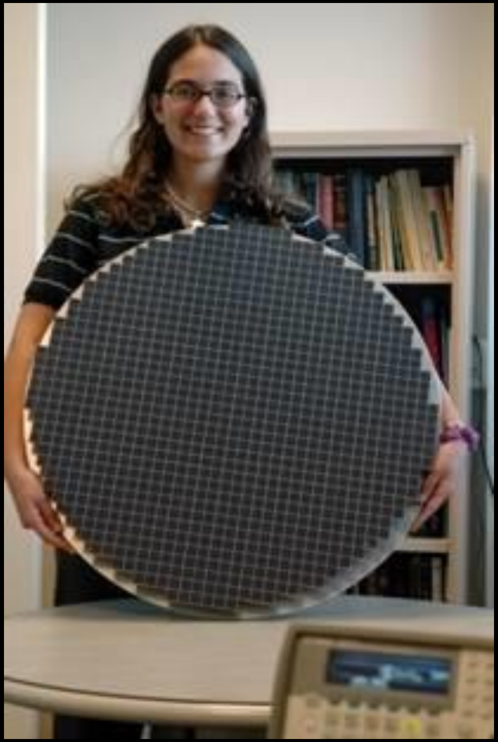
LSST in time and space:

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

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***



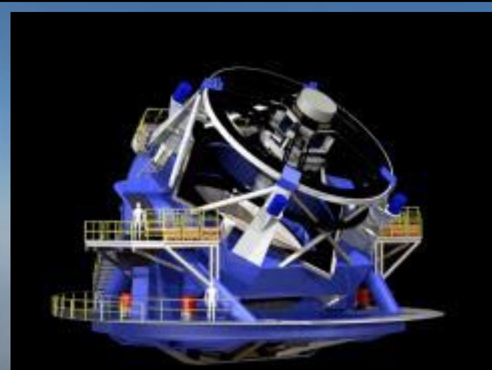
The LSST Data Challenges

Mason is an **LSST** member institution

Borne is chairman of the LSST Astroinformatics and Astrostatistics research team

@KirkDBorne

http://www.lsst.org/



Big Data Characteristics

Big Data is everywhere and growing

#1 priority for most businesses, social networks, and others…

http://www.v3.co.uk/v3-uk/news/2243045/onethird-of-businesses-planning-big-data-takeup-in-2013

Big Data is everywhere and growing

Examples:

http://bit.ly/1b2Qgci
http://bit.ly/154wYUq
http://bit.ly/19IjL4y

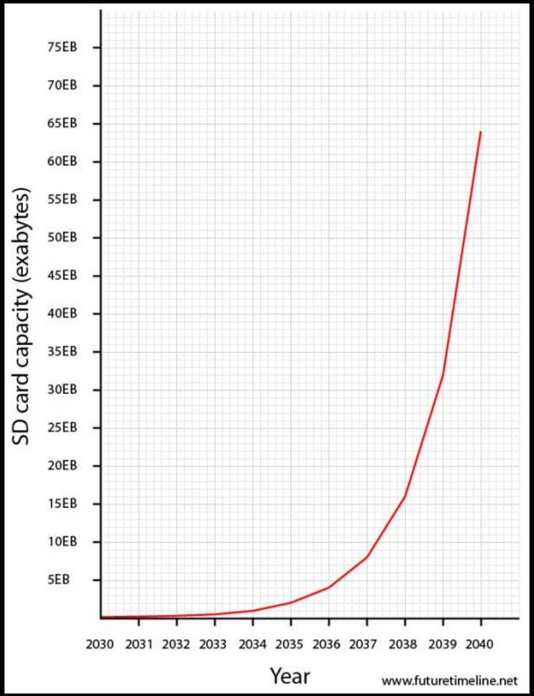
http://www.v3.co.uk/v3-uk/news/2243045/onethird-of-businesses-planning-big-data-takeup-in-2013

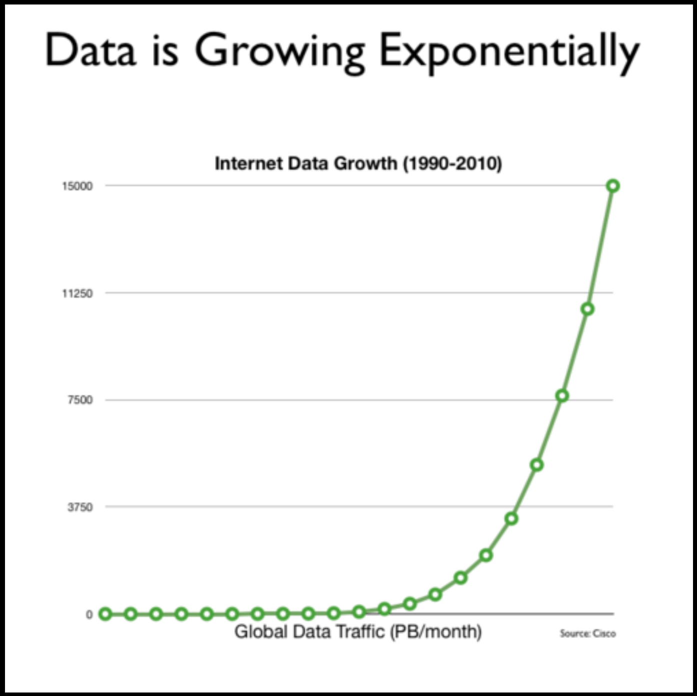
4 Characteristics of Big Data – #1234

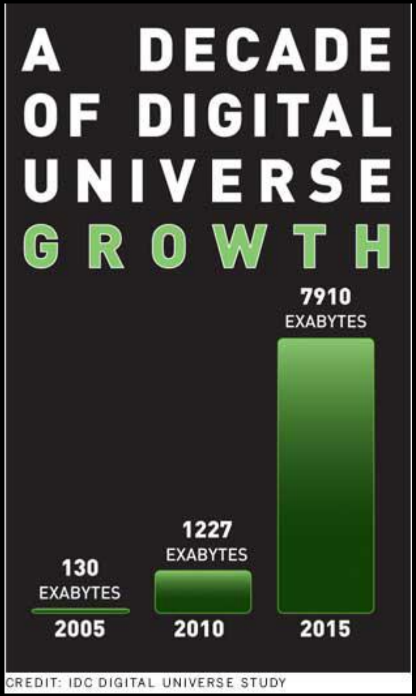
There are **huge** volumes of data in the world:

- From the beginning of recorded time until 2003, we created 5 billion gigabytes (exabytes) of data.
- In 2011 the same amount was created every two days.
- In 2013, the same amount is created every 10 minutes.

http://money.cnn.com/gallery/technology/2012/09/10/big-data.fortune/index.html



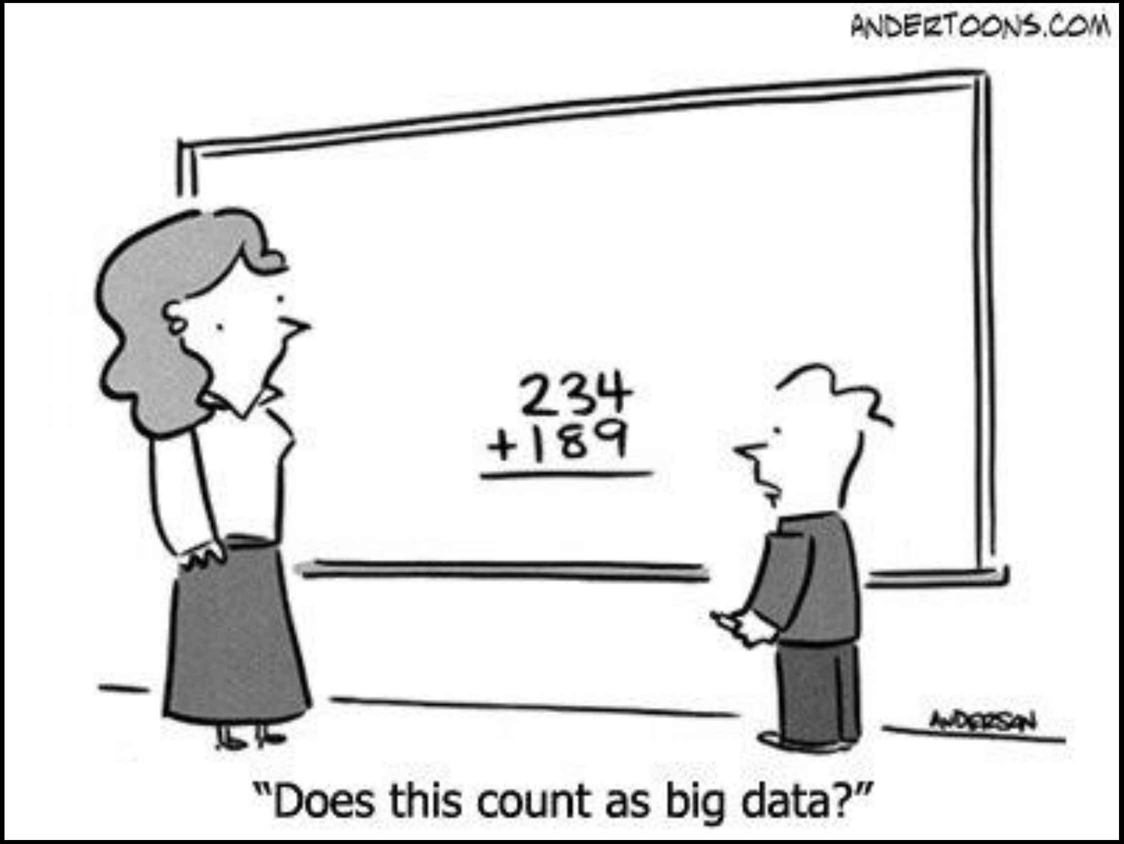




4 Characteristics of Big Data – #1234

Huge quantities of data are acquired everywhere:

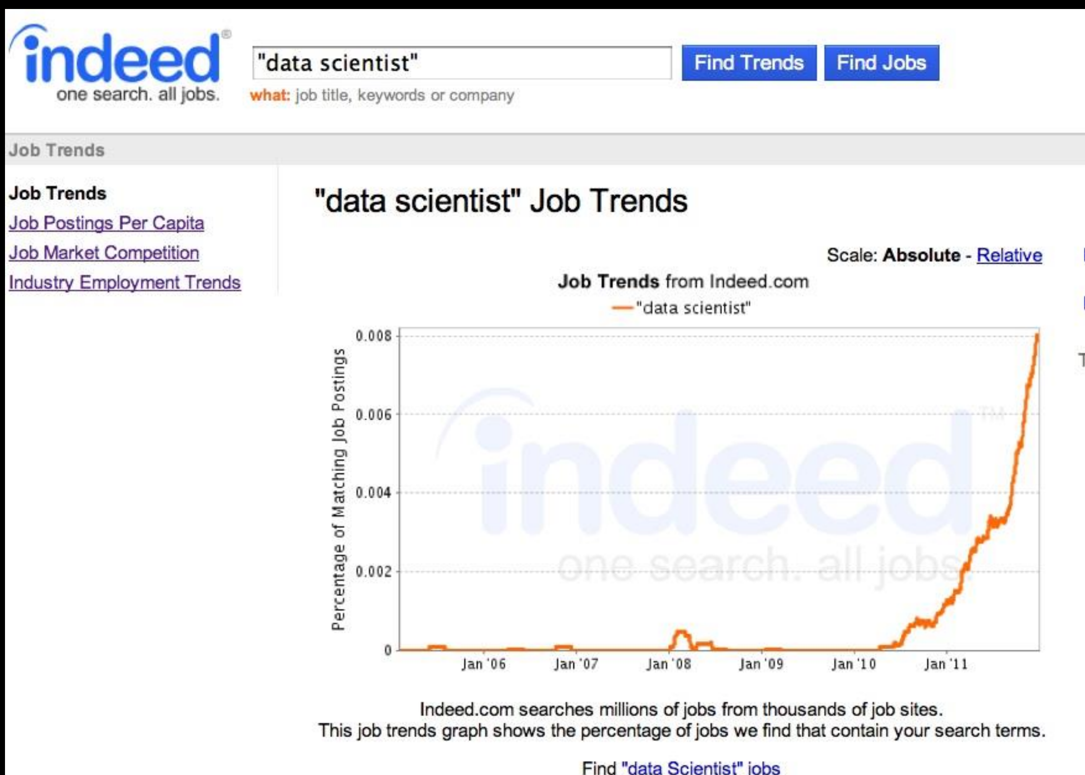
- **Big Data** is a big issue in all aspects of life: science, social networks, transportation, business, healthcare, government, national security, media, education, etc.

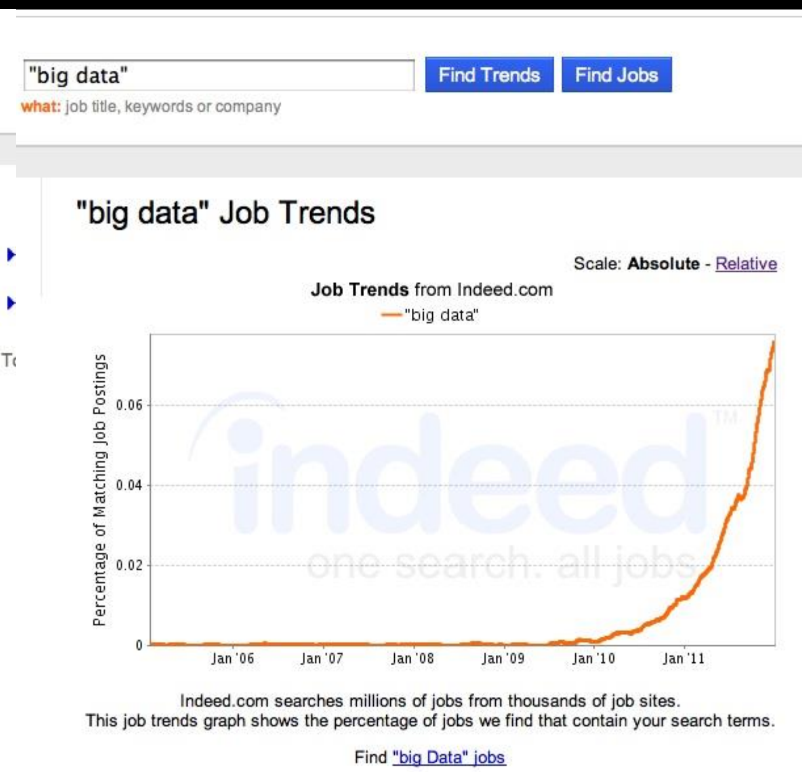


4 Characteristics of Big Data – #1234

Job opportunities are sky-rocketing:

- Extremely high demand for Big Data analysis skills
- Demand will continue to increase
- Old: “100 applicants per job”. New: “100 jobs per applicant”





4 Characteristics of Big Data – #1234

Job opportunities are sky-rocketing:

- Extremely high demand for Big Data analysis skills
- Demand will continue to increase
- Old: “100 applicants per job”. New: “100 jobs per applicant”

McKinsey Report (2011) :

- Big Data is the new “gold rush” , the “new oil”
- 1.5 million skilled data scientist shortage within 5 years
- Big Data investments will exceed $1 Trillion
- Machine-to-Machine Intelligence will be multi-Trillion $ industry
- Disruptive technologies will be $33 Trillion/year business
- http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation
- http://blogs.justonedatabase.com/2012/02/27/big-is-in-the-eye-of-the-beholder/
- http://bits.blogs.nytimes.com/2013/05/22/mckinsey-the-33-trillion-technology-payoff/

McKinsey Report (2013) :

- Big Data investments in Healthcare ~ $450 Billion
- http://www.thedoctorweighsin.com/the-value-of-big-data-in-health-care-450-billion/

Characteristics of Big Data – #1234

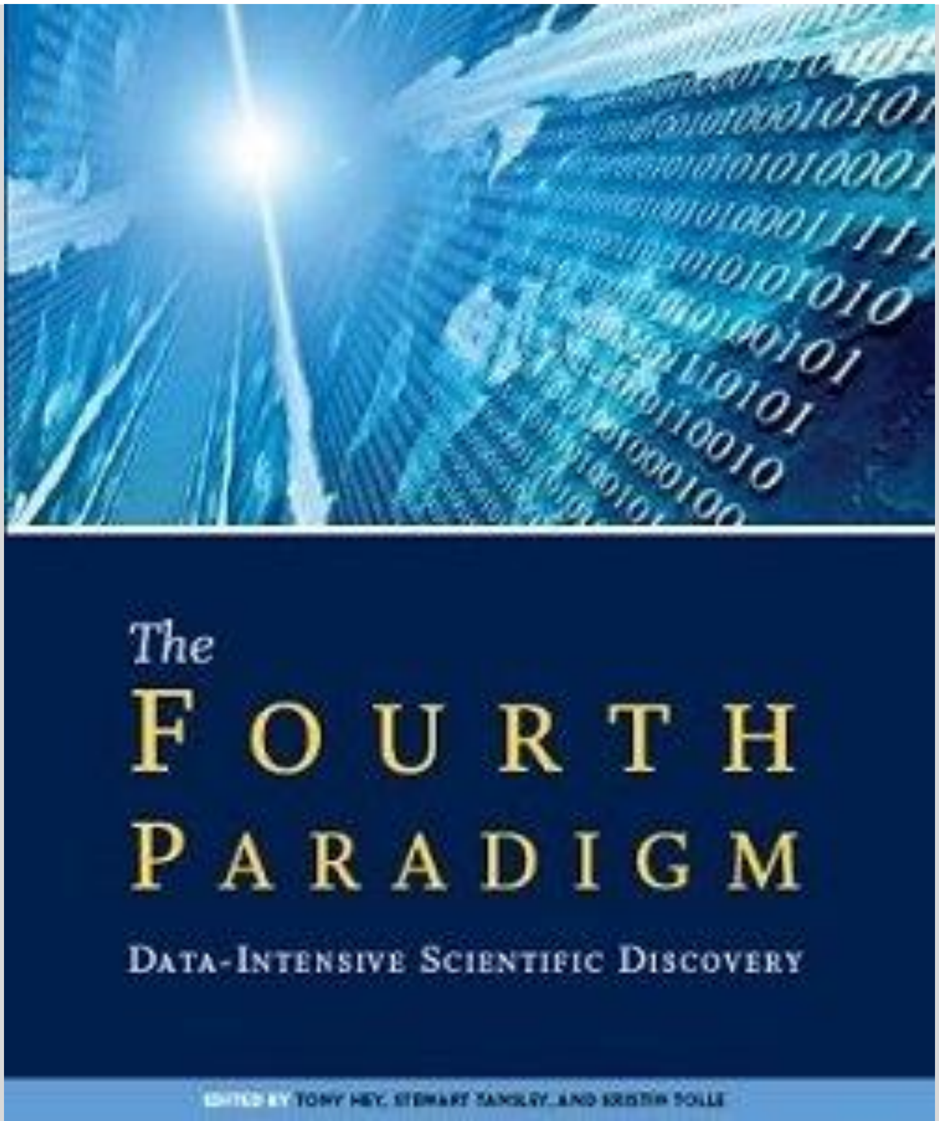
- The emergence of **Data Science** and **Data-Oriented Science** (the [4th paradigm of science]).
  - *"Computational literacy and data literacy are critical for all."* - Kirk Borne

Data Science: A National Imperative

1. National Academies report: *Bits of Power: Issues in Global Access to Scientific Data,* (1997) 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://archive.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/storage/documents/publications/digital-data-report-2006.pdf
8. NSF report: *Cyberinfrastructure Vision for 21st Century Discovery,* (2007) downloaded from http://www.nsf.gov/od/oci/ci_v5.pdf
9. JISC/NSF Workshop report on Data-Driven Science & Repositories, (2007) downloaded from http://www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf
10. 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.sci.utah.edu/vaw2007/DOE-Visualization-Report-2007.pdf
11. DOE report: *Mathematics for Analysis of Petascale Data Workshop Report,* (2008) downloaded from http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Peta_scaled_at_a_workshop_report.pdf
12. 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
13. 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
14. NSF report: *Data-Enabled Science in the Mathematical and Physical Sciences*, (2010) downloaded from https://www.nsf.gov/mps/dms/documents/Data-EnabledScience.pdf
15. National Big Data Research and Development Initiative, (2012) downloaded from http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf
16. National Academies report: *Frontiers in Massive Data Analysis*, (2013) downloaded from http://www.nap.edu/catalog.php?record_id=18374

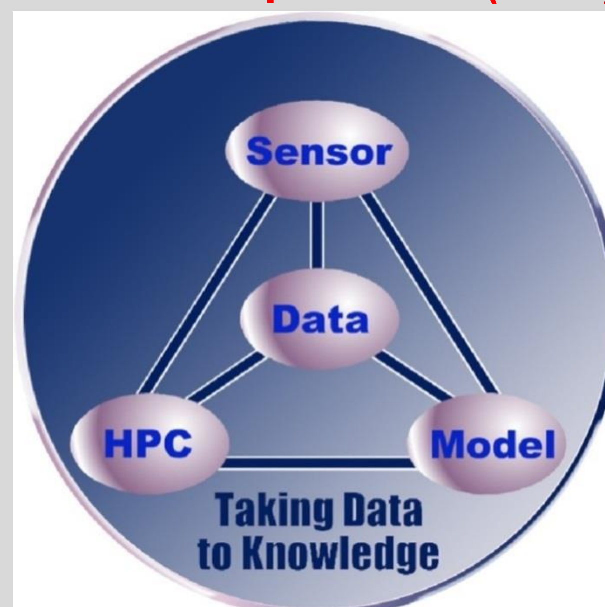
The Fourth Paradigm: Data-Intensive Scientific Discovery

http://research.microsoft.com/en-us/collaboration/fourthparadigm/



The 4 Scientific Paradigms:

1. Experiment (sensors)
2. Theory (modeling)
3. Simulation (HPC)
4. **Data Exploration (KDD)**



Characteristics of Big Data – #1234

- The emergence of **Data Science** and **Data-Oriented Science** (the 4th paradigm of science).
  - *"Computational literacy and data literacy are critical for all."* - Kirk Borne
- 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.
  - *"Somewhere, something incredible is waiting to be known."* - Carl Sagan

Characteristics of Big Data – #1234

- The emergence of **Data Science** and **Data-Oriented Science** (the 4th paradigm of science).
  - *"Computational literacy and data literacy are critical for all."* - Kirk Borne
- 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.
  - *"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: **Astroinformatics**, Bioinformatics, Geoinformatics, Climate Informatics, Ecological Informatics, Biodiversity Informatics, Environmental Informatics, Health Informatics, Medical Informatics, Neuroinformatics, Crystal Informatics, Cheminformatics, Discovery Informatics, and more.

News #1 - Scary:
Big Data is taking us to a Tipping Point

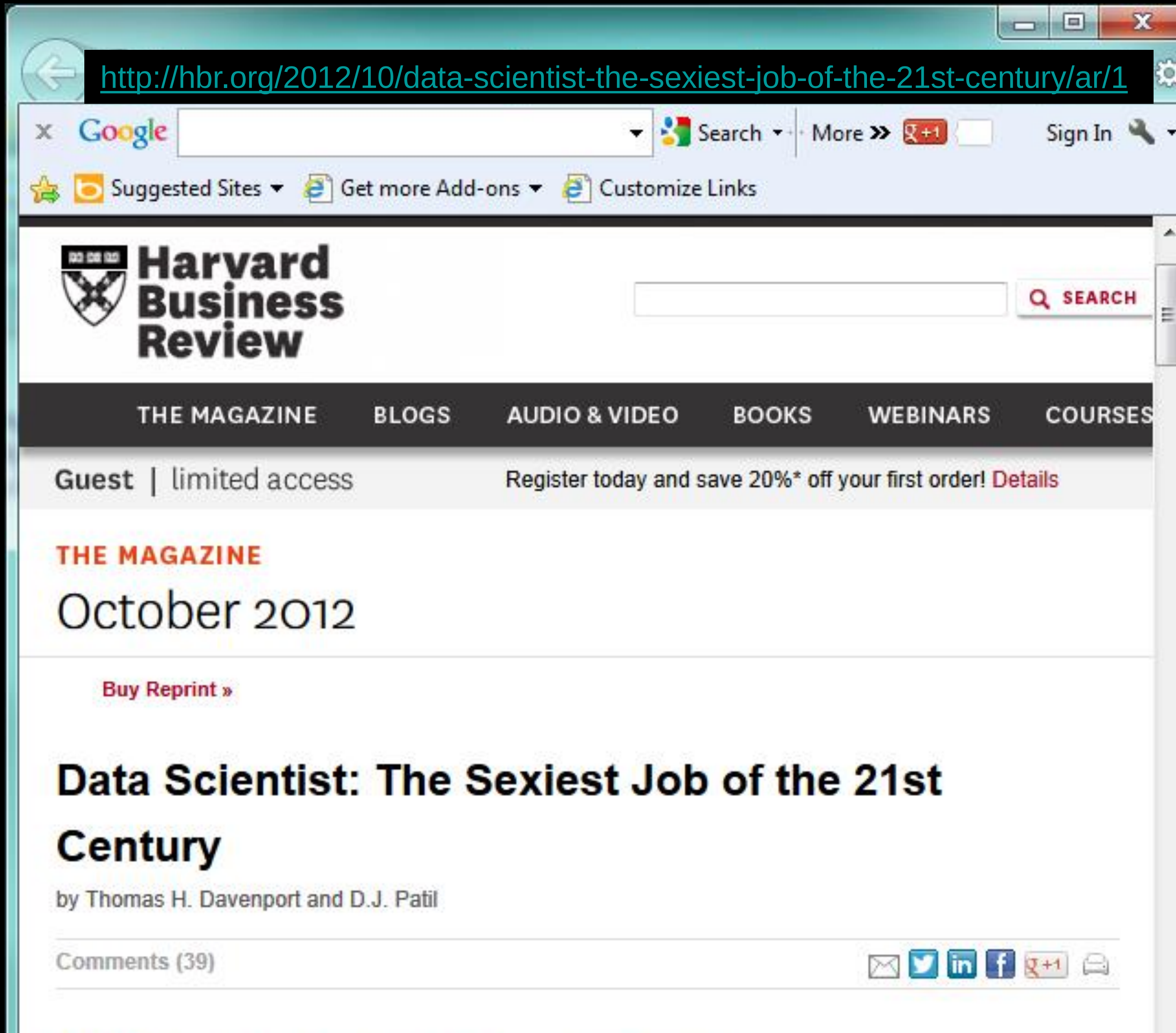
http://bit.ly/HUqmu5

http://goo.gl/Aj30t

News #2 - Promising: Big Data leads to Big Insights and New Discoveries

http://news.nationalgeographic.com/news/2010/11/photogalleries/101103-nasa-space-shuttle-discovery-firsts-pictures/

News #3 - Good: Big Data is Sexy



Big Data headlines in the news

- Smart cities market worth $1 trillion by 2016
- Schreiner University Adopts Predictive Analytics Suite
- Big Data Drives Big IT Spending ($34B in 2013; $232B thru 2016)
- Big Data Tackles Patients Who Don't Take Meds (cost healthcare system $317B/yr)
- Creating a better world with data
- Big Data: The Management Revolution
- Can Big Data Revitalize Public Transit in Los Angeles?
- Social networks can predict the spread of infectious disease
- Big Data Analytics Today Lets Businesses Play Moneyball
- How Big Data Can Make Us Happier & Healthier
- Facial Analytics: From Big Data to Law Enforcement
- Predictive Policing: prediction and probability in crime patterns
- Foot Locker Deploys Big Data Visual Analytics System
- Can data analytics prevent the next offshore oil spill?
- San Francisco bars: Buy a drink, become profiled by cameras
- Big Data Astrophysics is out!

Big Data headlines in the news

- Smart cities market worth $1 trillion by 2016
- Schreiner University Adopts Predictive Analytics Suite
- Big Data Drives Big IT Spending ($34B in 2013; $232B th
- Big Data Tackles Patients Who Don't Take M n $317B/yr)
- Creating a better world wit
- Big Data: Th
- Can
- Socia ous disease
- Big Da Businesses Play Moneyball
- How B an Make Us Happier & Healthier
- Facial Analytics: From Big Data to Law Enforcement
- Predictive Policing: prediction and probability in crime patterns
- Foot Locker Deploys Big Data Visual Analytics System
- Can data analytics prevent the next offshore oil spill?
- San Francisco bars: Buy a drink, become profiled by cameras
- Big Data Astrophysics is out!

For up-to-date synopses of Big Data and Data Science in the news, follow @KirkDBorne on Twitter…

http://www.onalytica.com/blog/posts/onalytica-big-data-influencers-q4-13

Characterizing the Big Data Hype

- If the only distinguishing characteristic was that we have lots of data, we would call it **"Lots of Data"**.

Characterizing the Big Data Hype

- If the <u>only</u> distinguishing characteristic was that we have lots of data, we would call it **"Lots of Data"**.
- Big Data characteristics: the 3+n V's =
  1. **Volume** *(lots of data = "Tonnabytes")*
  2. **Variety** *(complexity, curse of dimensionality)*
  3. **Velocity** *(rate of data and information flow)*
  4. **V**
  5. **V**
  6. **V**
  7. **V**
  8. **V**

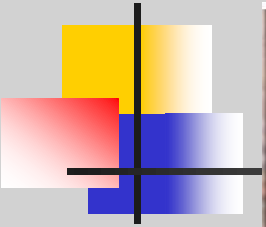
Characterizing the Big Data Hype

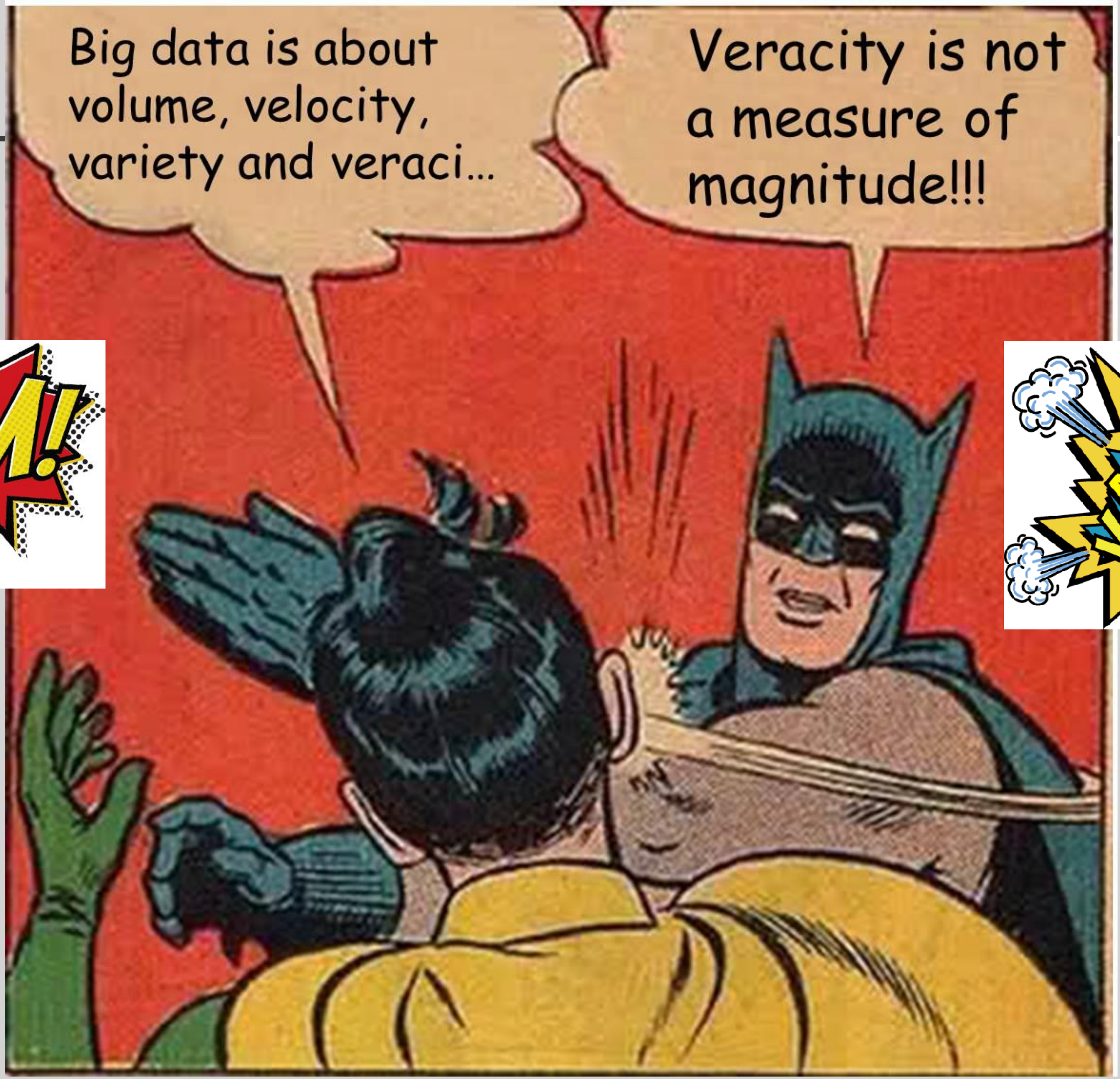
- If the only distinguishing characteristic was that we have lots of data, we would call it **"Lots of Data"**.
- Big Data characteristics: the 3+n V's =

1. **Volume** *(lots of data = "Tonnabytes")*
2. **Variety** *(complexity, curse of dimensionality)*
3. **Velocity** *(rate of data and information flow)*
4. **Veracity** *(data to verify many hypotheses)*
5. **Variability**
6. **Venue**
7. **Vocabulary**
8. **Value**

We will return to these Later.

http://whatsthebigdata.com/2013/11/14/batman-rejects-big-data-v-inflation/





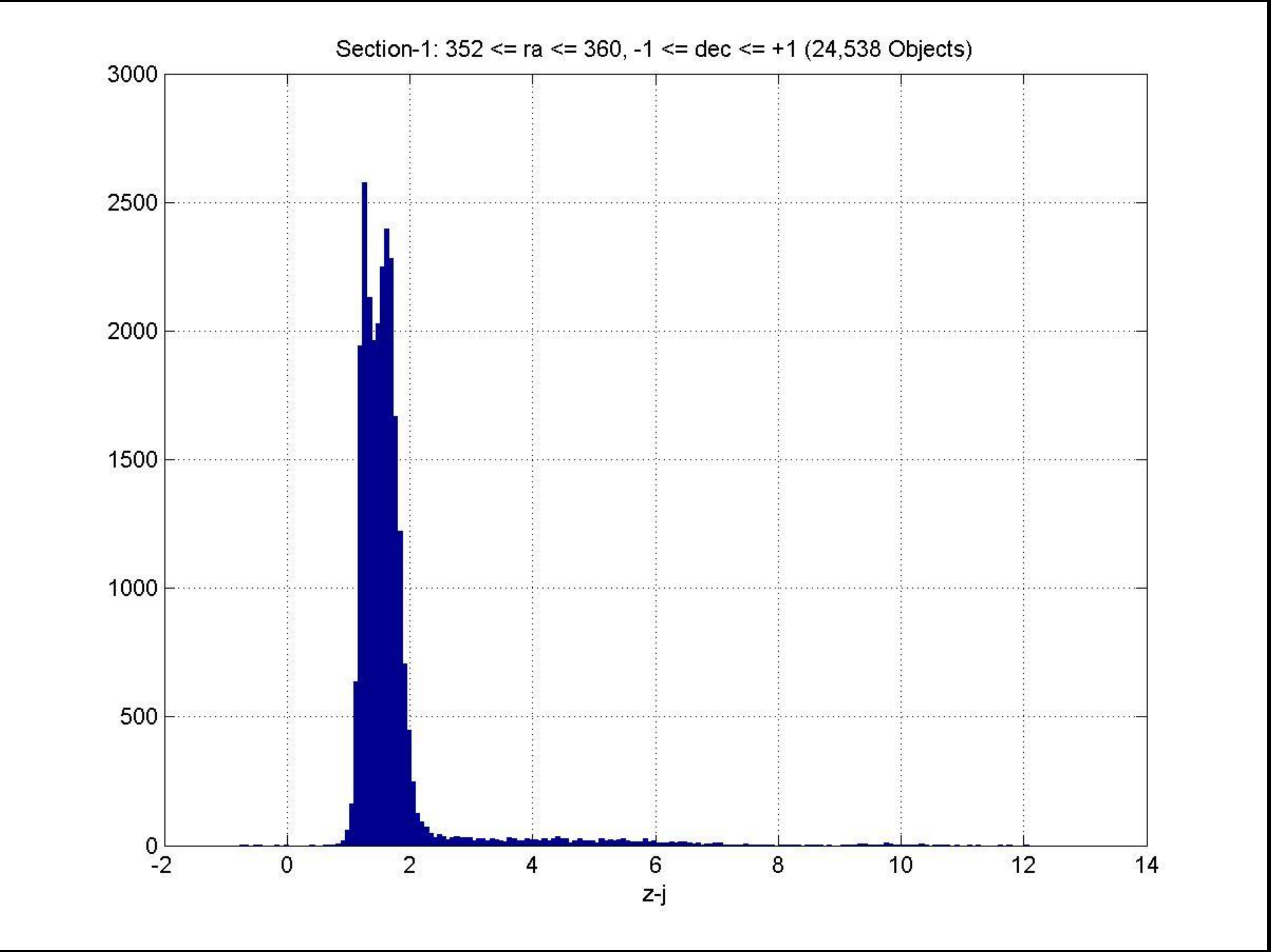
Characterizing the Big Data Hype

- If the <u>only</u> distinguishing characteristic was that we have lots of data, we would call it **"Lots of Data"**.
- Big Data characteristics: the 3+n V's =

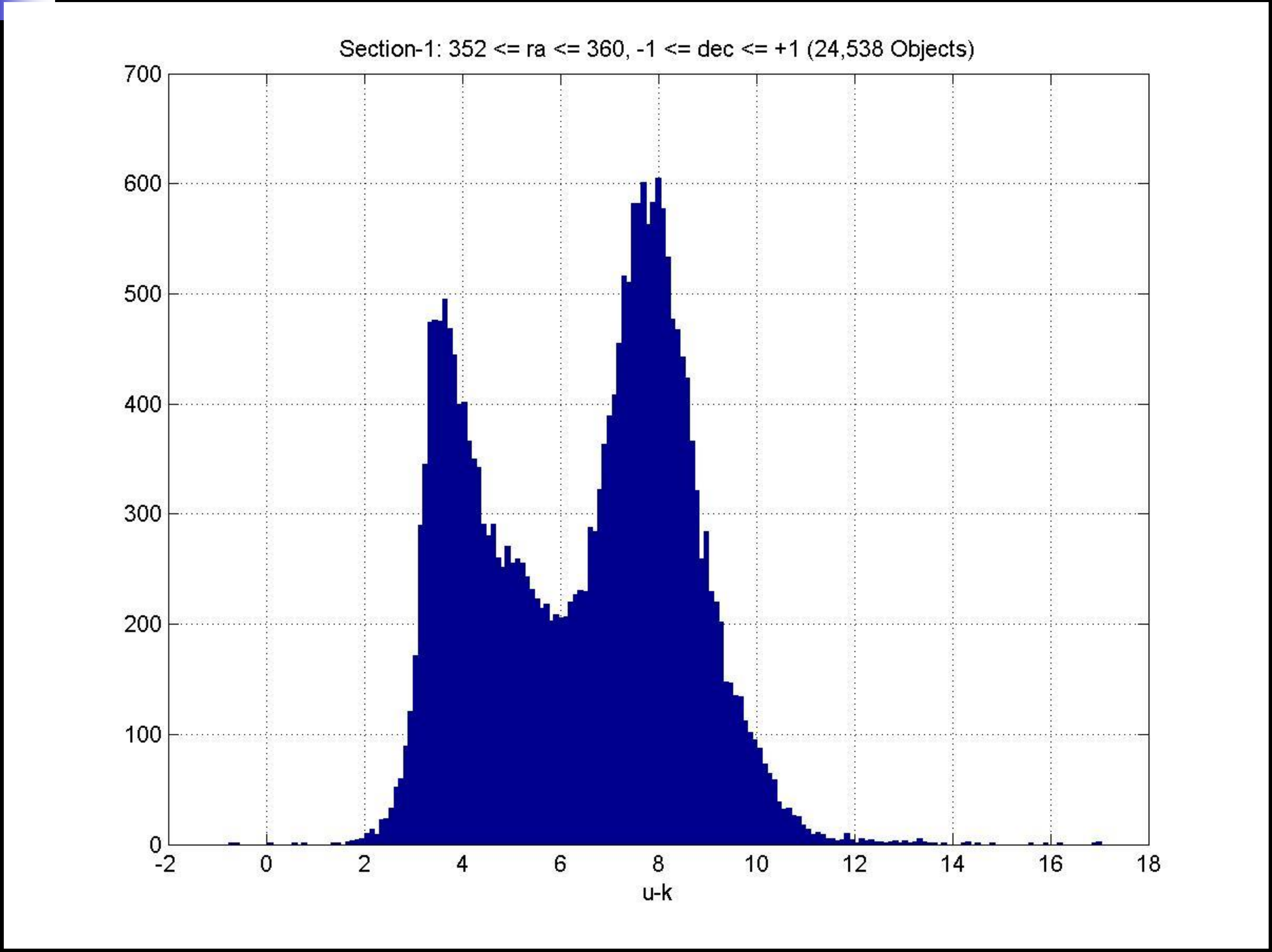
Big Data Example :

2. **Variety : this one helps us to discriminate subtle new classes (= Class Discovery)**
3. **Velocity**
4. **Veracity**
5. **Variability**
6. **Venue**
7. **Vocabulary**
8. **Value**

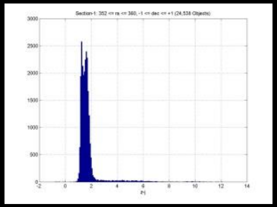
Insufficient Variety: stars & galaxies are not separated in this parameter



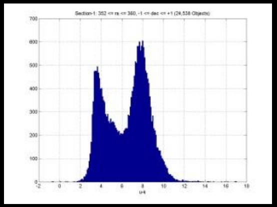
Sufficient Variety: stars & galaxies are separated in this parameter

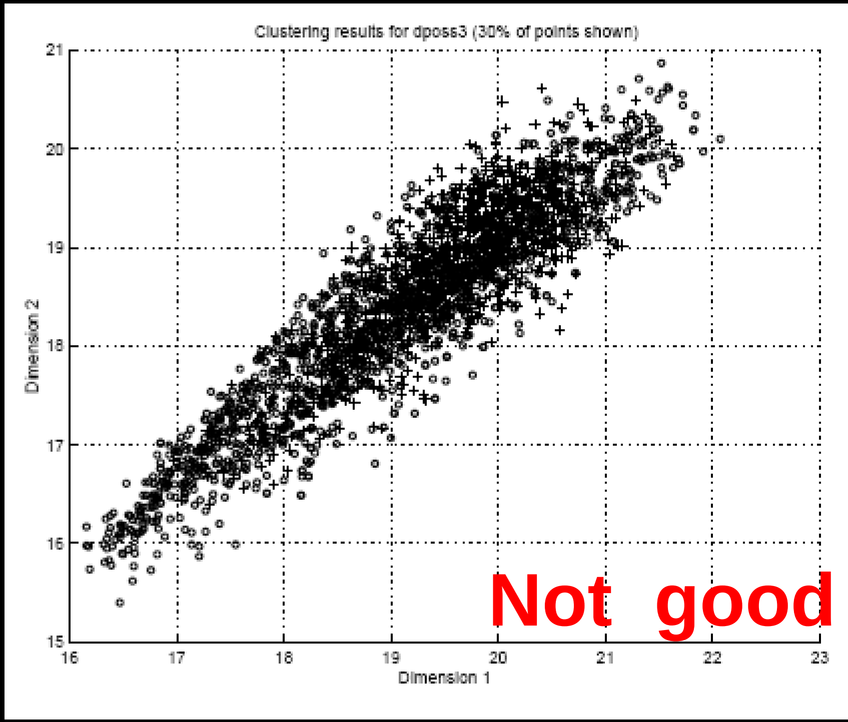


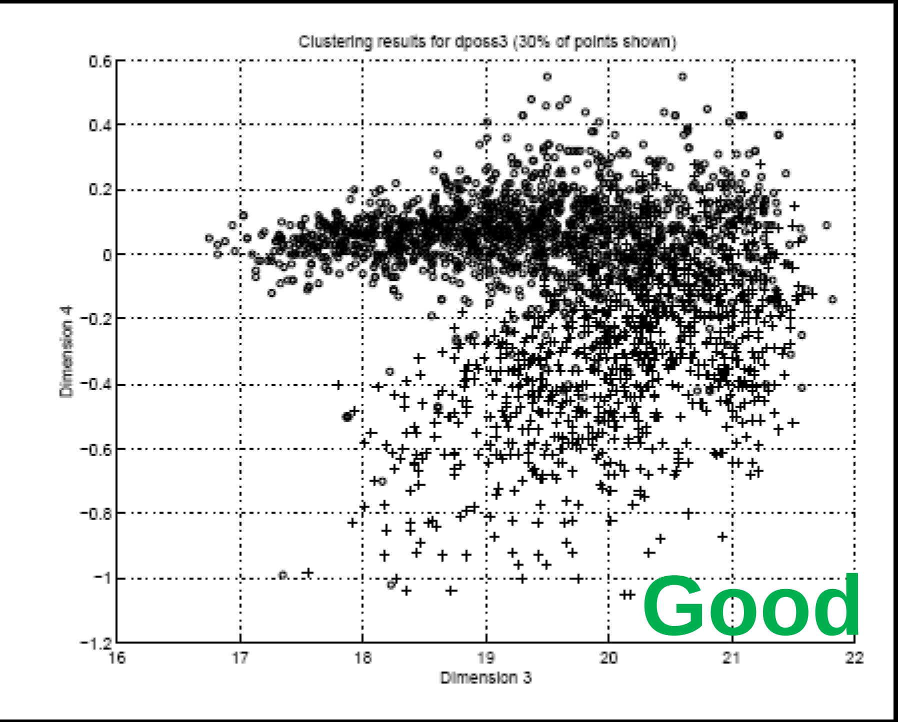
The 3 important D's of Big Data Variety: feature Disambiguation, Discrimination between multiple classes, and Discovery of new classes.



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:

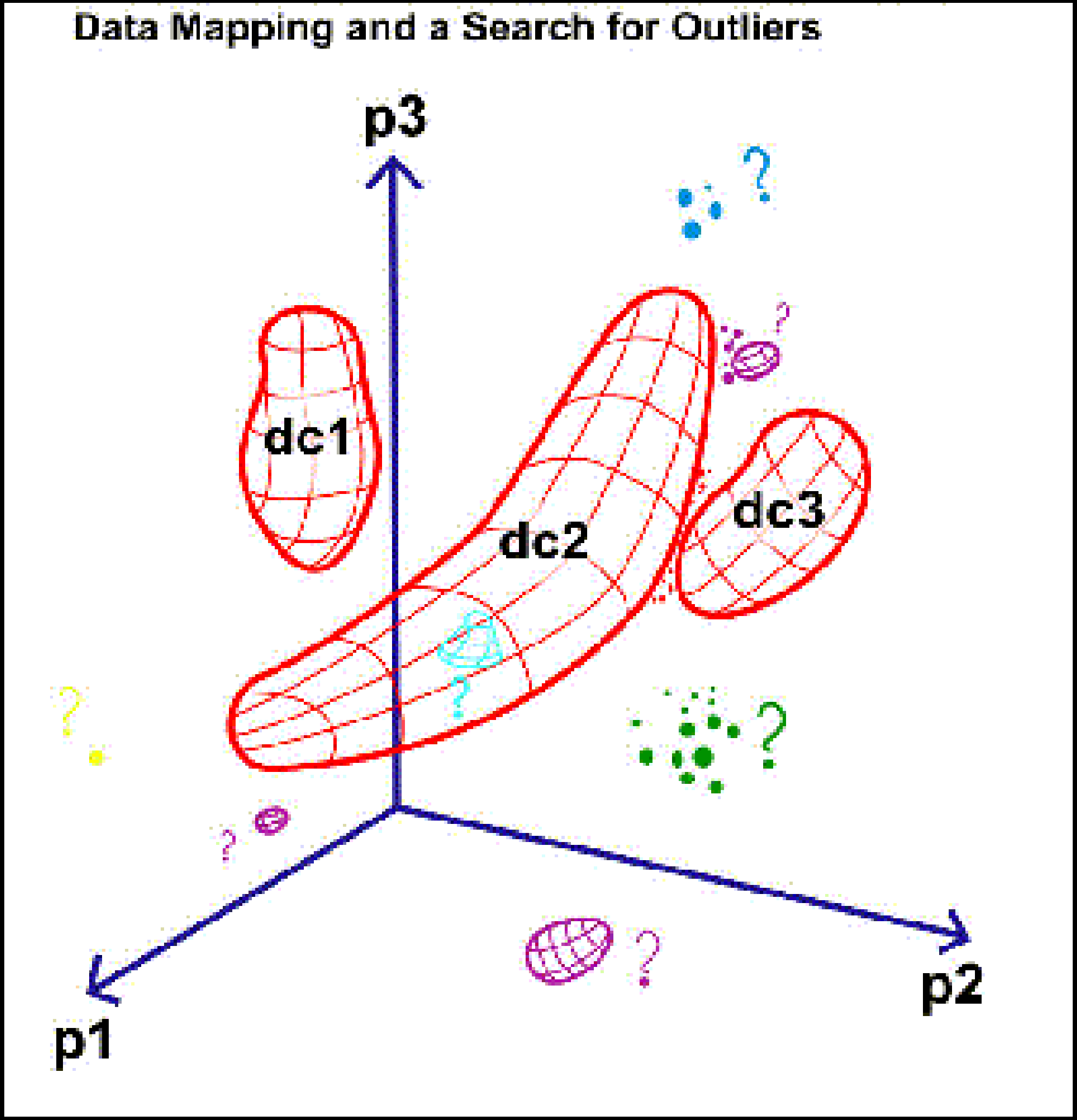






http://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/BrunnerDPS.pdf

This graphic says it all …

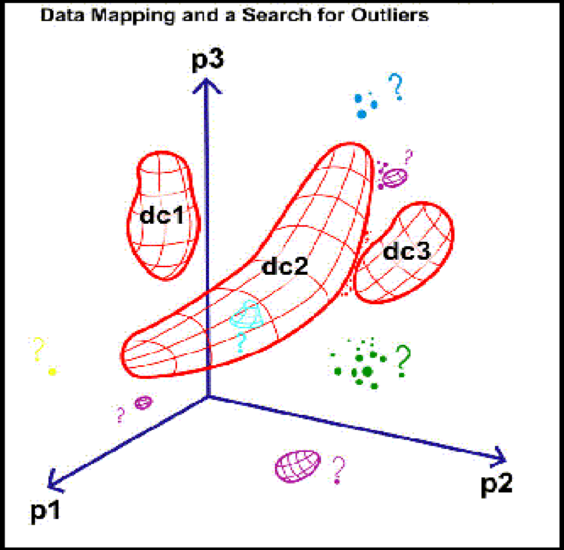


- Clustering – examine the data and find the data clusters (clouds), without considering what the items are = Characterization !
- Classification – for each new data item, try to place it within a known class (i.e., a known category or cluster) = Classify !
- Outlier Detection – identify those data items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

Graphic provided by Professor S. G. Djorgovski, Caltech

Data-Driven Discovery:
(KDD: Knowledge Discovery from Data)

1. Correlation Discovery
2. Novelty Discovery
3. Class Discovery
4. Association Discovery



Graphic from S. G. Djorgovski

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

4 Categories of Data Science Analytics: Knowledge Discovery from Big 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

What is Association Discovery?

- Identifying connections between different things (people or events)
- Finding unusual (improbable) co-occurring combinations of things (for example: in your shopping cart)
- Finding things that have much fewer than "six degrees of separation"

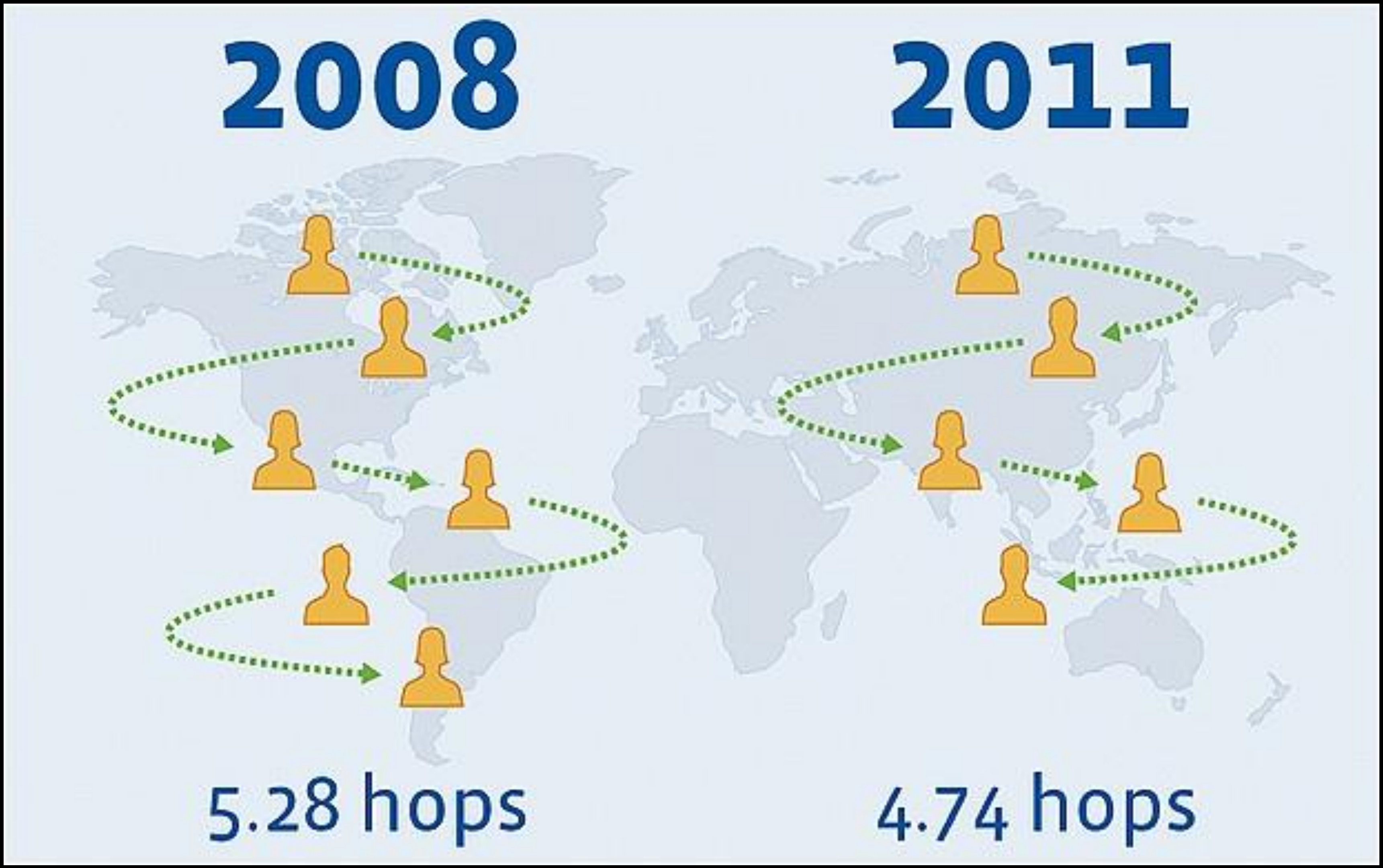
6 Degrees of Separation:

Everyone is on average approximately 6 steps away from any other person on Earth (through their relationships with each other).

http://info.logicmanager.com/bid/86132/ERM-and-the-Six-Degrees-of-Separation-Theory

Less than 6 Degrees of Separation: due to Social Networks!

http://www.telegraph.co.uk/technology/facebook/8906693/Facebook-cuts-six-degrees-of-separation-to-four.html



6 Degrees of Separation works with people, and it works with any type of network, including the network of things that you like or purchase linked to things that other people buy or like.

http://nosql.mypopescu.com/post/46508012660/graph-based-recommendation-systems-at-ebay

4 Examples: how Big Data is shrinking your world – Small World Connections through Associations

amazon.com

NETFLIX

Walmart

<u>Example #1</u>: how Big Data is shrinking your world – Small World Connections through Associations

- **Classic Textbook Example of Data Mining** (Legend?): Data mining of grocery store logs indicated that men who buy diapers also tend to buy beer at the same time.

Example #2: how Big Data is shrinking your world – Small World Connections through Associations

- **Amazon.com** mines its customers' purchase logs to recommend books to you: *"People who bought this book also bought this other one."*

Example #3: how Big Data is shrinking your world – Small World Connections through Associations

- **Netflix** mines its video rental history database to recommend rentals to you based upon other customers who rented similar movies as you.

Example #4: how Big Data is shrinking your world – Small World Connections through Associations

- **Wal-Mart** studied product sales in their Florida stores in 2004 when several hurricanes passed through Florida.

- Wal-Mart found that, before the hurricanes arrived, people purchased 7 times as many of ***{one particular product}*** compared to everything else.

Example #4: how Big Data is shrinking your world – Small World Connections through Associations

- **Wal-Mart** studied product sales in their Florida stores in 2004 when several hurricanes passed through Florida.

- Wal-Mart found that, before the hurricanes arrived, people purchased 7 times as many **strawberry pop tarts** compared to everything else.

Strawberry pop tarts???

http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html
http://www.hurricaneville.com/pop_tarts.html

Definitions of Big Data

From Wikipedia:

- Big Data refers to any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

Definitions of Big Data

<u>From Wikipedia</u>:

- ~~Big Data refers to any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.~~
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

<u>My suggestion</u>:

- **Big Data refers to "Everything, Quantified and Tracked!"**

- According to the standard (Wikipedia) definition, even the Ancient Romans had Big Data! That's ridiculous!
  - See my article ***"Today's Big Data is Not Yesterday's Big Data"*** at: http://bit.ly/1aXb7hD

Definitions of Big Data

<u>From Wikipedia</u>:

- ~~Big Data refers to any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.~~
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

<u>My suggestion</u>:

- **Big Data refers to "Everything, Quantified and Tracked!"**

- The challenges **do not** change – but their scale, scope, scariness, discovery potential **do** change!
- Examples:
  - Big Data Science Projects
  - Social Networks
  - IoT = Internet of Things
  - M2M = Machine-to-Machine

Exponential Growth

- Data growth is coupled to the growth in computer processing power (Moore's Law) = the ability to generate data from computational processes!
- Exponential function has this property:

$$df/dx \sim f$$

Exponential Growth

- Data growth is coupled to the growth in computer processing power (Moore's Law) = the ability to generate data from computational processes!
- Exponential function has this property:

$df/dx \sim f$ … therefore, $d^2f/dx^2 \sim f$, etc.

All derivatives of e^x are also exponential.

- Consequently, the rate of growth is growing exponentially, and the rate of growth of the rate of growth (acceleration) is growing exponentially, etc.
- This rapidly becomes "out of control" = which we call a tipping point, or unstable equilibrium, …

Exponential Growth

- Data growth is coupled to the growth in computer processing power (Moore's Law) = the ability to generate data from computational processes!
- Observed Fact : Volume of data doubles every year (roughly) = 100% growth rate.
- Consider – Compound Interest at 100% APR:
  - Invest $1 in your 401(k) at age 20
  - Total invested = $1
  - Value of your 401(k) fund at age 65 = **$???**
    - See my article ***"Big Data: Compound Interest Growth on Steroids"*** at: http://bit.ly/19fR2lI
    - Related article: ***Simple Math Formula is Basically Responsible for All of Modern Civilization***

Exponential Growth

- Data growth is coupled to the growth in computer processing power (Moore's Law) = the ability to generate data from computational processes!
- Observed Fact : Volume of data doubles every year (roughly) = 100% growth rate.
- Consider – Compound Interest at 100% APR:
  - Invest $1 in your 401(k) at age 20
  - Total invested = $1
  - Value of your 401(k) fund at age 65 = **$35 Trillion!**
    - See my article ***"Big Data: Compound Interest Growth on Steroids"*** at: http://bit.ly/19fR2lI
    - Related article: ***Simple Math Formula is Basically Responsible for All of Modern Civilization***

It was the best of times, it was the worst of times…

- January 1986 – Shuttle Challenger disaster!!!
- August 1986 – Hubble Space Telescope (HST) was scheduled for launch, but postponed until April 1990.
- 1986-1990: Time of reflection, re-tooling, improvements, and …

It was the best of times, it was the worst of times…

- 1986-1990: new look at Scientific Data Management!
  - Initially, NASA managers decided that HST didn't need a data archive, just a "Data Management Facility" (*e.g.,* that wooden crate holding the Ark of the Covenant at the end of Indiana Jones movie "Raiders of the Lost Ark")

  - After some "lobbying" by HST science managers, the concept of a Hubble Science Data Archive was born! (and Borne! – who eventually became HST Data Archive Project Scientist!)

Science Data "Management" in the HST and Sloan Era

- The Hubble Data Archive became a widely used research tool for scientists, who conducted "secondary" investigations on the data that were initially collected for some PI's primary research program.
- The Sloan Digital Sky Survey carried out an imaging (and spectroscopic) survey of ¼ of the sky ("Pi in the Sky"). The Sloan project scientists had their own primary science programs, but the real value came in the community re-use of the data:
  - Over 5000 refereed papers thus far!

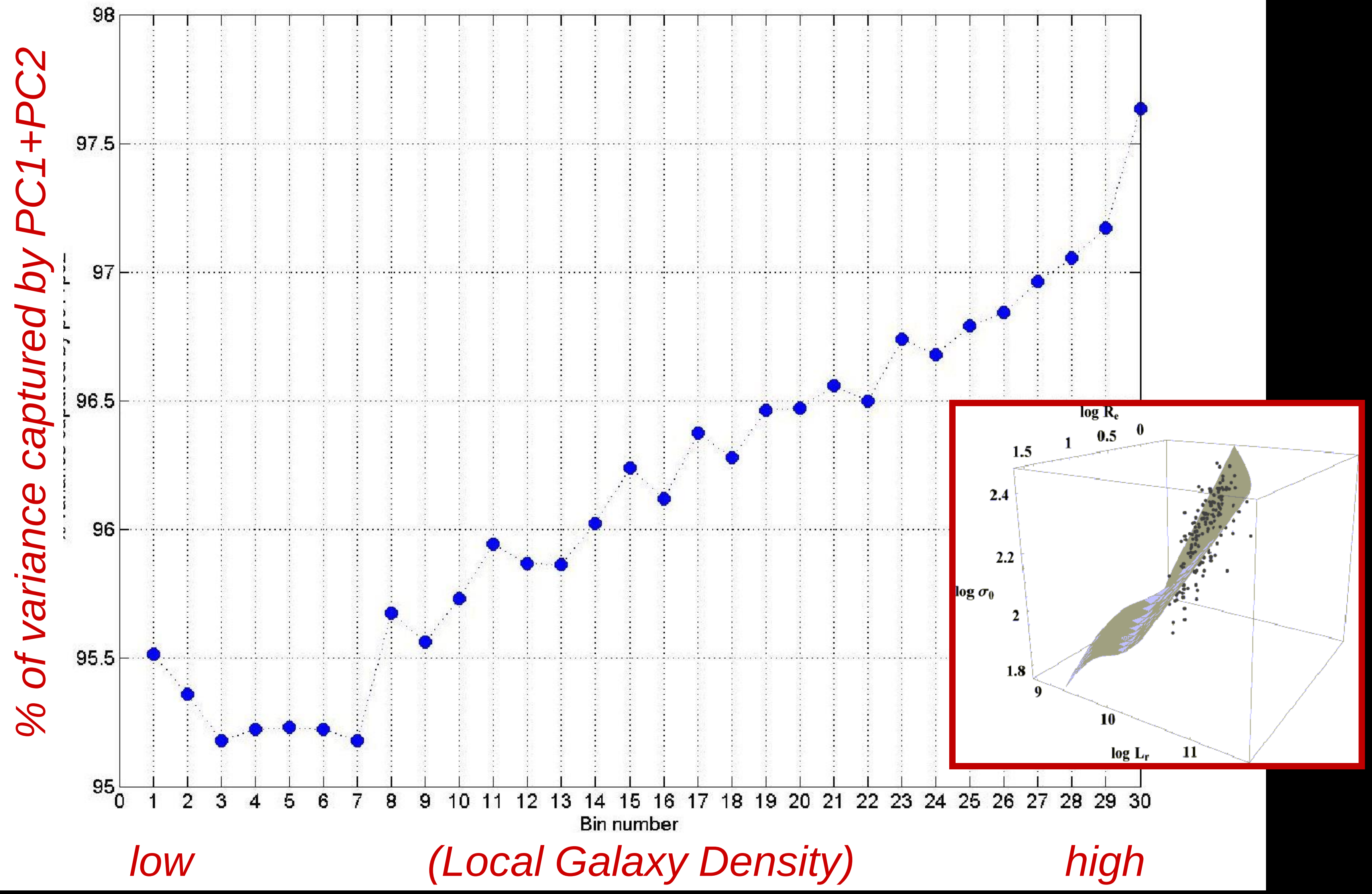
    http://blog.sdss3.org/2013/03/26/sdss-has-now-been-used-by-over-5000-refereed-papers/
- Now, the number of refereed papers for HST science is larger for archival research than for primary observation programs.
- Science Data = now focused on Discovery, not Management!

Data-Oriented Discovery

- Scientific experiments can now be run against the data collection.
- Hypotheses are inferred, questions are posed, experiments are designed & run, results are analyzed, hypotheses are tested & refined!
- This is the 4th Paradigm of Science
- This is especially (and correctly) true if the data collection is the “full” data set for a given domain:
  - astronomical sky surveys, human genome (the 1000 Genomes Project), social networks, large-scale simulations, earth observing system, ocean observatories initiative, banking, retail, national security, cybersecurity, … and the list goes on and on …

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



Other examples

- Earth Science – pattern detection (fire, cyclone, typhoon)
- Education – personalized learning / interventions
- Social Networks – targeted ads / recommendations
- Law Enforcement – predictive policing
- Healthcare – personalized medicine; medical discovery
- …

Data Science & The 4th Paradigm

- The tools of the 4th Paradigm are the tools of **Data Science**:
  - data mining (machine learning algorithms), visualization, data structures and indexing schemes, statistics, applied math, semantics (ontologies, taxonomies), data-intensive computational methods (Hadoop/MapReduce; data-parallelism vs. task-parallelism; shared-nothing,…)
- Similarly, the tools of 3rd Paradigm are the tools of **Computational Science**:
  - parallel computing methods, applied math algorithms, data structures, grid methods, high-performance computing, memory allocation techniques, modeling & simulation methods (Monte Carlo, CFD grid-based, N-body point-based, Agent-based modeling,…)

What is Data Science?

- It is a collection of mathematical, computational, scientific, and domain-specific methods, tools, and algorithms to be applied to Big Data for discovery, decision support, and data-to-knowledge transformation…
  - Statistics
  - Data Mining (Machine Learning) & Analytics (KDD)
  - Data & Information Visualization
  - Semantics (Natural Language Processing, Ontologies)
  - Data-intensive Computing (e.g., Hadoop, Cloud, …)
  - Modeling & Simulation
  - Metadata for Indexing, Search, & Retrieval
  - Advanced Data Management & Data Structures
  - Domain-Specific Data Analysis Tools

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, intelligent agents, 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**

Why do all of this?

… for 4 very simple reasons:

- (1) Any real data collection may consist of millions, or billions, or trillions of sampled data points.

Why do all of this?

… for 4 very simple reasons:

- (1) Any real data collection may consist of millions, or billions, or trillions of sampled data points.
- (2) Any real data set will probably have many hundreds (or thousands) of measured attributes (features, dimensions).

Why do all of this?

… for 4 very simple reasons:

- (1) Any real data collection may consist of millions, or billions, or trillions of sampled data points.
- (2) Any real data set will probably have many hundreds (or thousands) of measured attributes (features, dimensions).
- (3) Humans can make mistakes when staring for hours at long lists of numbers, especially in a dynamic data stream.

Why do all of this?

… for 4 very simple reasons:

- (1) Any real data collection may consist of millions, or billions, or trillions of sampled data points.
- (2) Any real data set will probably have many hundreds (or thousands) of measured attributes (features, dimensions).
- (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 collection may consist of **Volume** s, or billions, or trillions of sampled data points.
- (2) Any real data set will probably have **Variety** undreds (or thousands) of measured attributes (features, dimensions).
- (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 able test of a hypothesis.

Why do all of this?

… for 4 very simple reasons:

- (1) Any real data collection may consist of s, llions of sampled data points.

Volume It is too much !

- (2) Any real data set will probably have ds) of measured attributes (features, dimensions).

Variety It is too complex !

- (3) Humans can make mistakes when f 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 Data Science – 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, 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

Rationale for Data Science – 1

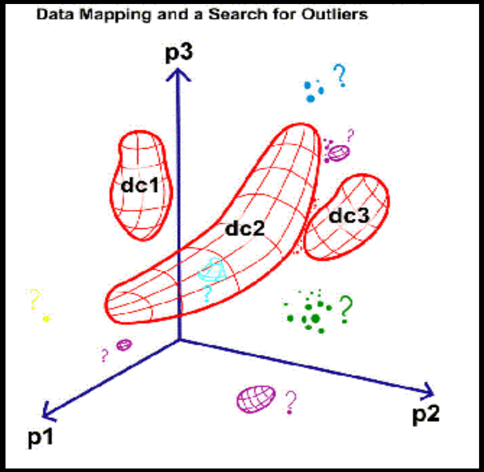
- Consequently, if we collect a thorough set of parameters (high-dimensional data) for a complete set of items within our do… study, then we would ha… al model for tha…
- … er for … matics.
- … to know about that domain is s… ed and encoded within the data.
- The goal of Big Data Science is to find those encodings, patterns, and knowledge nuggets.
- **See article:** [Big-Data Vision? Whole-population analytics]()

This is the difference between hypothesis-driven modeling & analysis and **data-driven modeling & analysis.**

Rationale for Data Science – 2

Data Science helps us to achieve Data-Driven Discovery from Big Data

1. Correlation Discovery
2. Novelty Discovery
3. Class Discovery
4. Association Discovery



Graphic from S. G. Djorgovski

- The 2 Big Benefits of Big Data:
  - best statistical analysis of "typical" events
  - automated search for "rare" events

Rationale for Data Science – 3

Data Science in and for Education

- 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")
- 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
- Big Data & Data Science programs emerging at "every" university! (RPI, Georgetown, UC Berkeley, U. Washington, NCSU, U. Illinois, …)
- **Informatics as a new Gen Ed requirement ? … why not ?**

Data Science addresses Big (and Small) Data's Data-to-Knowledge Challenges:

- Finding order in Data
- Learning from Data
- Finding the unknown unknowns
- Data Literacy for all !

because …

- Everything is Quantified and Tracked!