JPL-Caltech Virtual Summer School

Big Data Analytics

September 2 – 12, 2014

S. G. Djorgovski

Center for Data Driven Discovery (CD^3)

Caltech

Welcome and Introduction

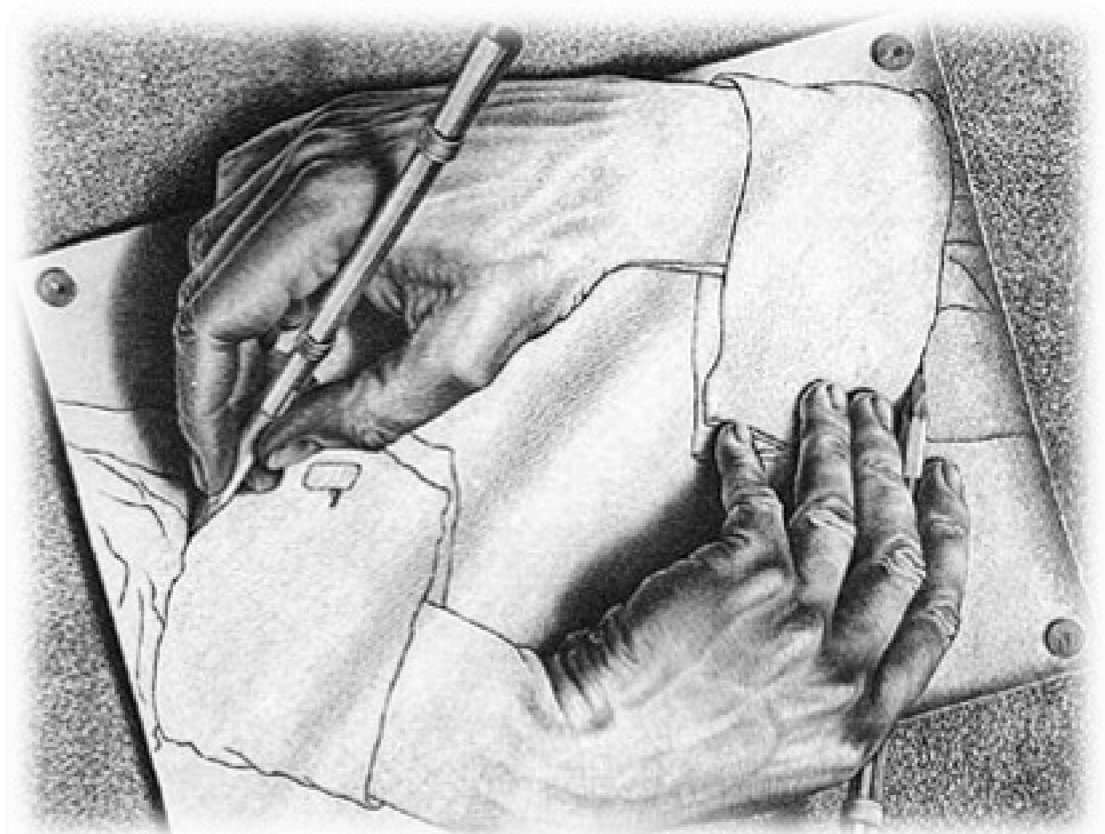
Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

What Is This School About?

- It is about ***applications*** of computer science tools and technologies and statistics to scientific data analysis
- A quick (!) introduction to a selected few topics, useful for data-intensive research
  - You should explore further
  - There are many topics that we do not cover (yet)
- It will evolve, and your feedback is welcome
- It is ***not*** about:
  - Computer science proper
  - High performance computing

Transformation and Synergy

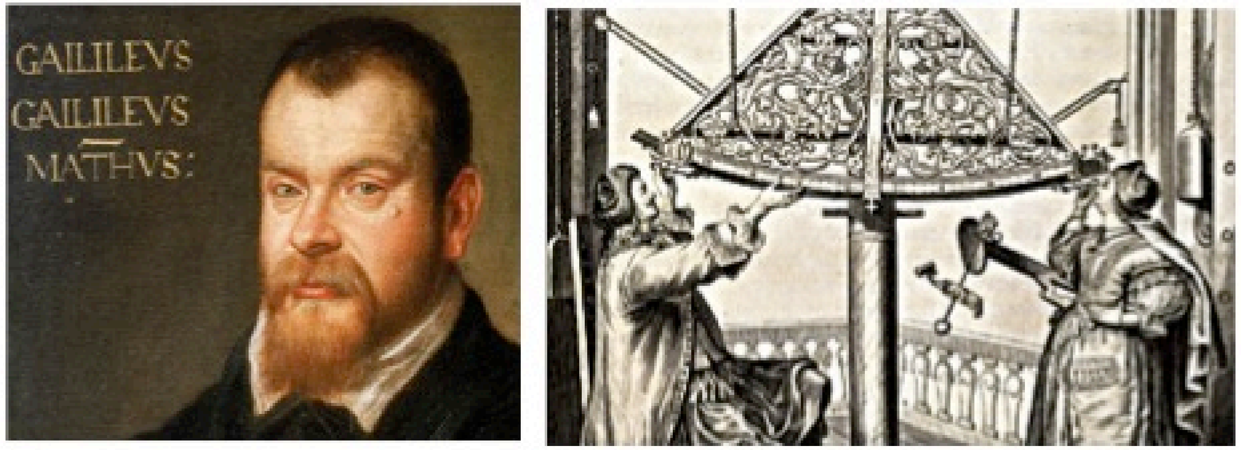
- ***All science*** in the 21st century is becoming cyber-science (aka e-Science) - and with this change comes the need for ***a new scientific methodology***
- The challenges we are tackling:
  - Management of large, complex, distributed data sets
  - Effective exploration of such data → new knowledge
  - **These challenges are universal**
- A great synergy of the computationally enabled science, and the science-driven technology



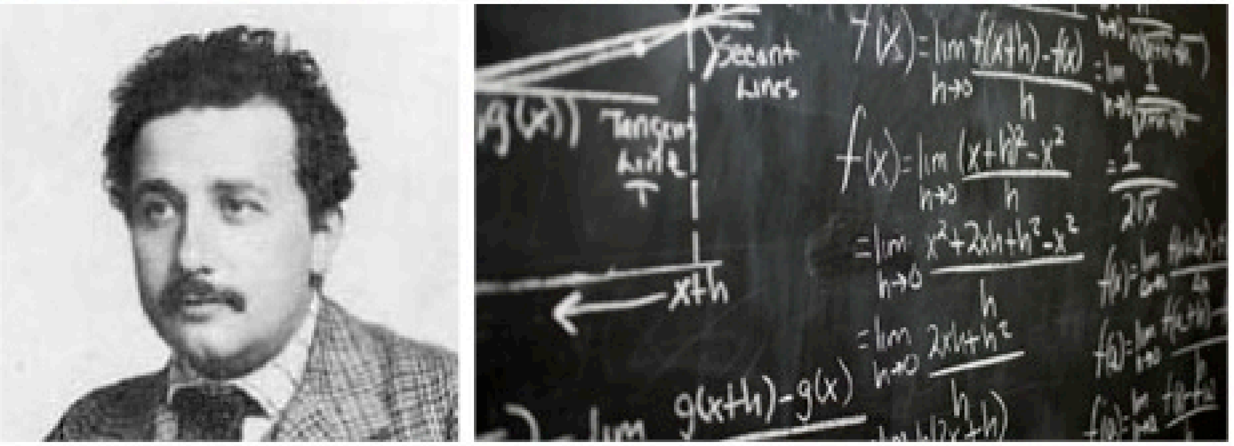
Djorgovski

The Evolving Paths to Knowledge

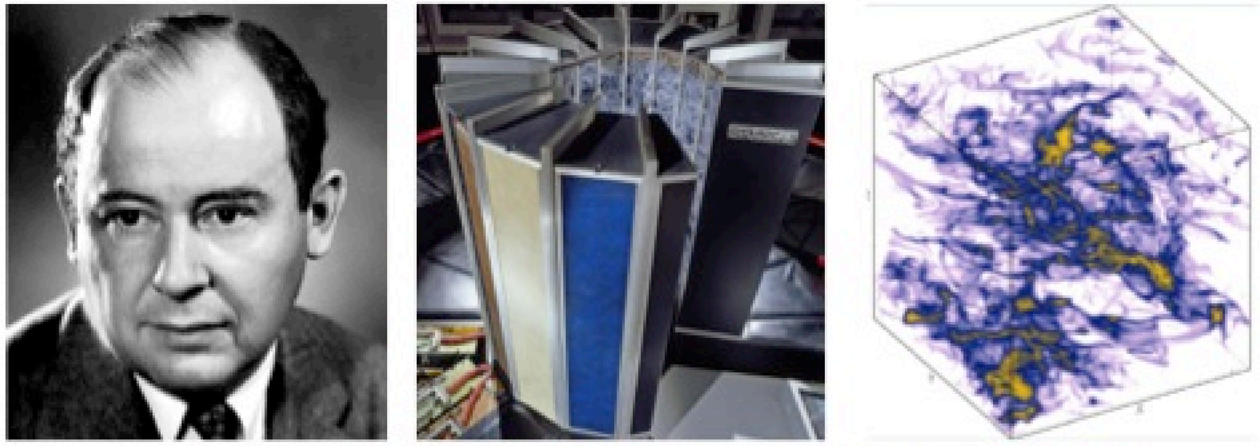
- The First Paradigm: Experiment/ Measurement



- The Second Paradigm: Analytical Theory



- The Third Paradigm: Numerical Simulations



- The Fourth Paradigm: Data-Driven Science

Djorgovski

Astronomy Has Become Very Data-Rich

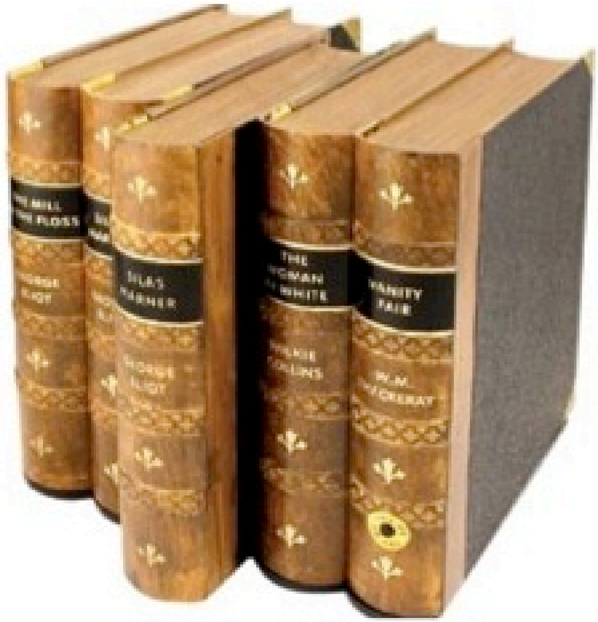
- Typical digital sky survey generate ~ 10 - 1000 TB each, plus a comparable amount of derived data products
  - Exabyte-scale data sets are on the horizon
- Astronomy today has ~ 10 PB of archived data, and generates ~ few ×10 TB/day
  - Both data volumes and data rates grow exponentially, with a ***doubling time ~ 1.5 years***
  - Even more important is the growth of ***data complexity***
- For comparison:

Human Genome < 1 GB

Human Memory < 1 GB (?)

1 TB ~ 2 million books

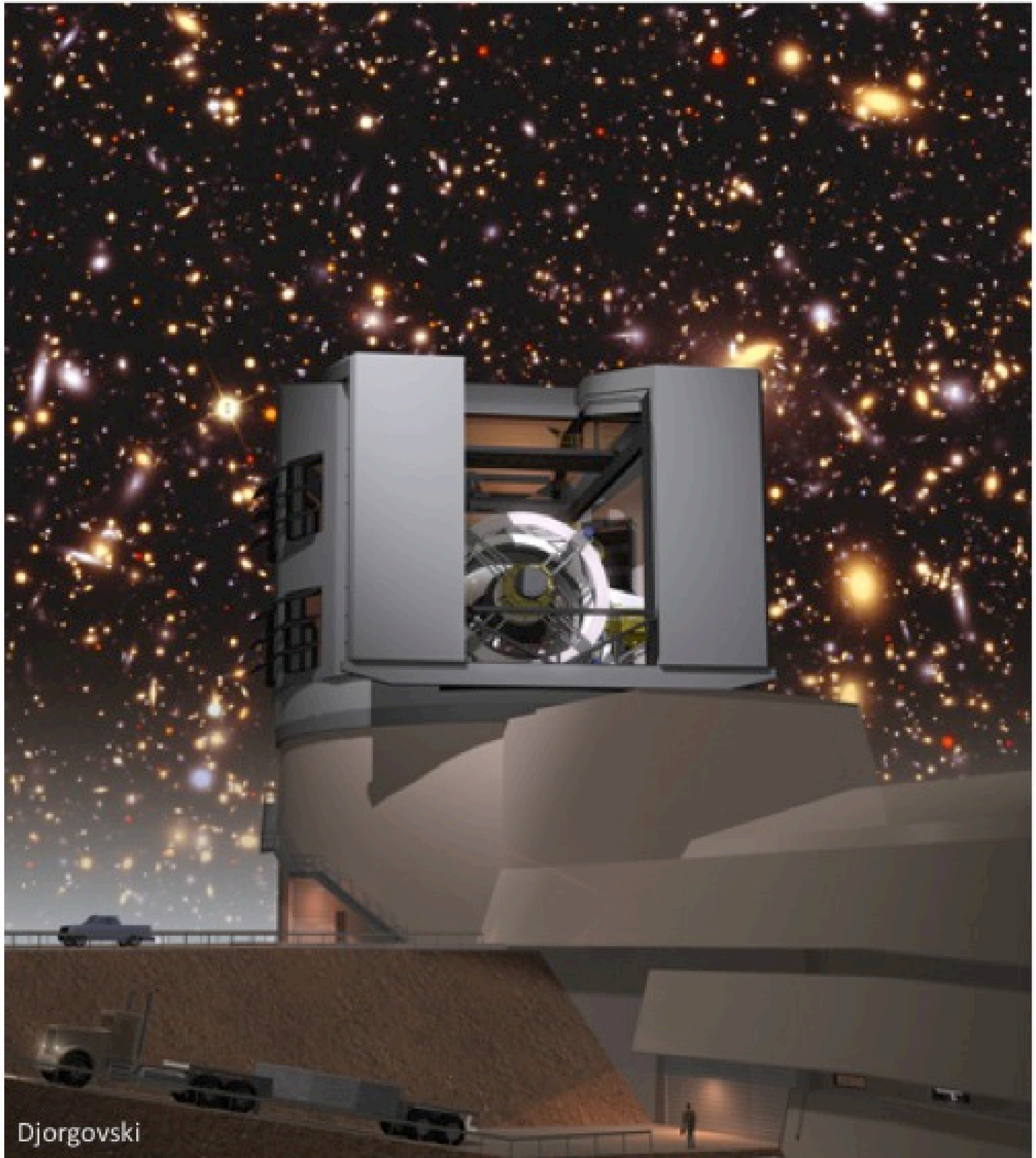
Human Bandwidth ~ 1 TB / year (±)



Djorgovski

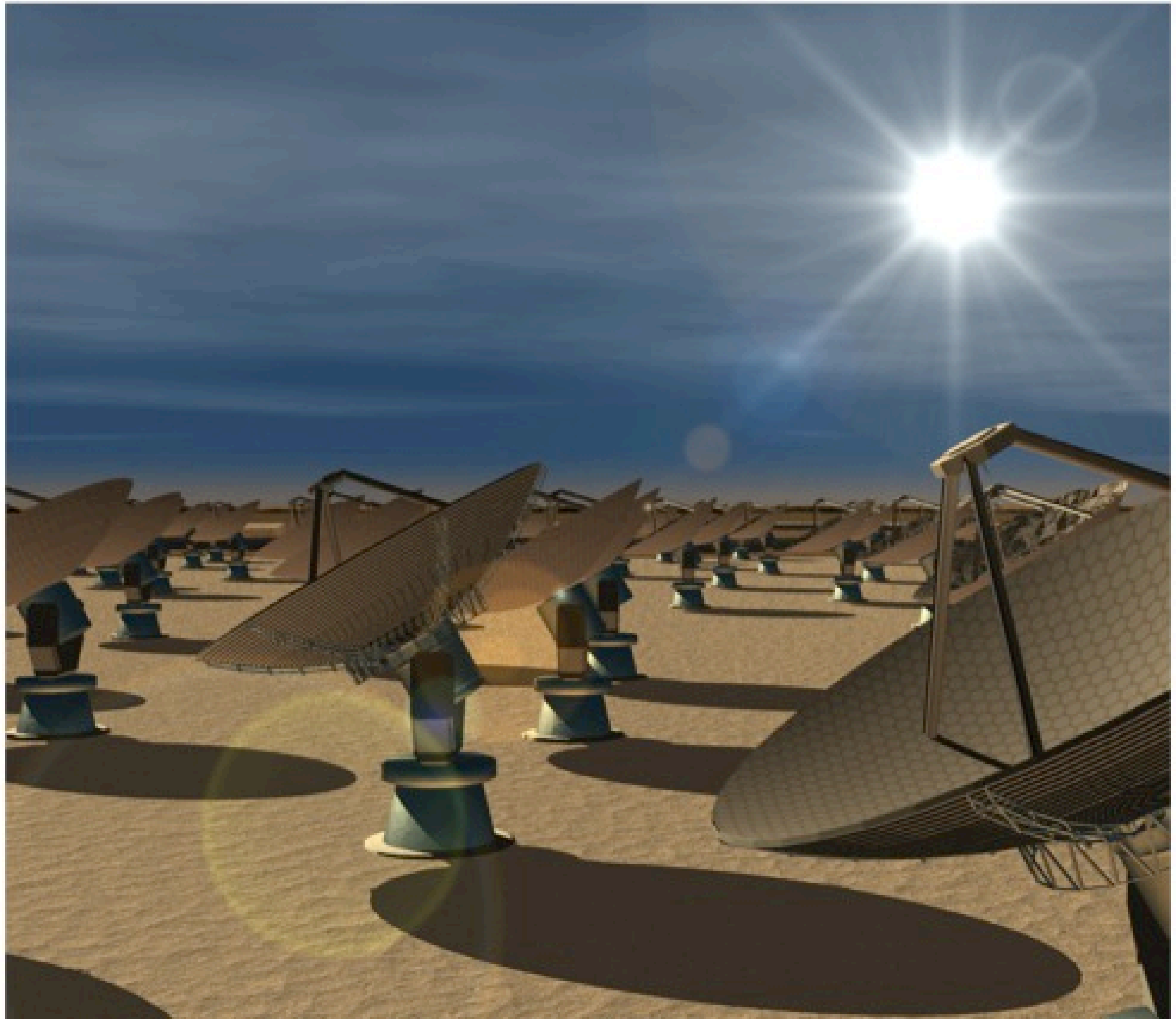
… And It Will Get Much More So

Large Synoptic Survey Telescope
(LSST) ~ 30 TB / night

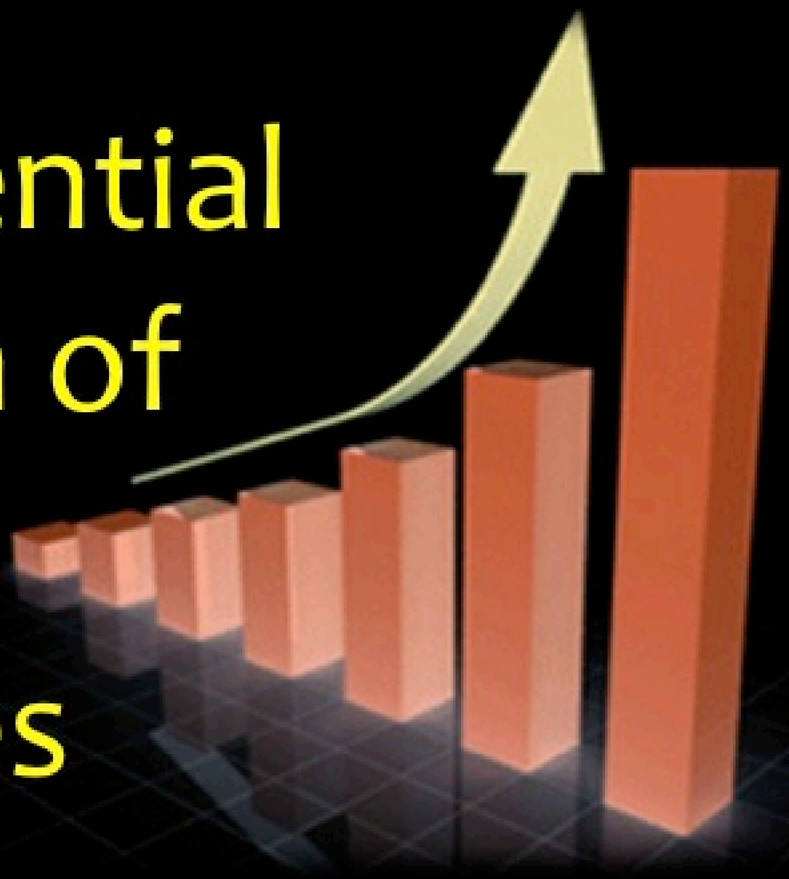


Djorgovski

Square Kilometer Array (SKA)
~ 1 EB / second (raw data)
(EB = 1,000,000 TB)



Exponential Growth of Data Volumes



on Moore’s law time scales

… and Complexity

Understanding of complex phenomena requires complex data!

From data poverty to data glut

From data sets to data streams

From static to dynamic, evolving data

From anytime to real-time analysis and discovery

From centralized to distributed resources

From ownership of data to ownership of expertise

Djorgovski

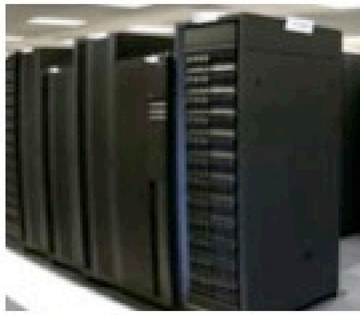
A Modern Scientific Discovery Process

Data Gathering (e.g., from sensor networks, telescopes…)

→ **Data Farming:**

Storage/Archiving
Indexing, Searchability
Data Fusion, Interoperability

} Database Technologies



→ **Data Mining** (or **K**nowledge **D**iscovery in **D**atabases):

Key Technical Challenges

Pattern or correlation search
Clustering analysis, classification
Outlier / anomaly searches
Hyperdimensional visualization

→ **Data Understanding**

Key Methodological Challenges

→ **New Knowledge**

+feedback

Djorgovski

Information Technology → New Science

- The information volume grows exponentially

  Most data will never be seen by humans!

⇨ The need for data storage, network, database-related technologies, standards, etc.

- Information complexity is also increasing greatly

  Most data (and data constructs) cannot be comprehended by humans directly!

⇨ The need for data mining and exploration, hyperdimensional visualization, AI/Machine-assisted discovery …

- We need to create ***a new scientific methodology*** for the computational science in the 21st century
- Important for practical applications beyond science – knowledge economy, etc.

Our goal is to help you start learning about the modern tools of scientific data analysis

JPL Keck Caltech

Enjoy!