# Exploring Space in Cyberspace: Astronomy and Data Science

**Prof. S. George Djorgovski**

*Center for Data-Driven Discovery*
*And Astronomy Dept., Caltech*

Lecture 1
XXX Canary Islands Winter School
November 2018

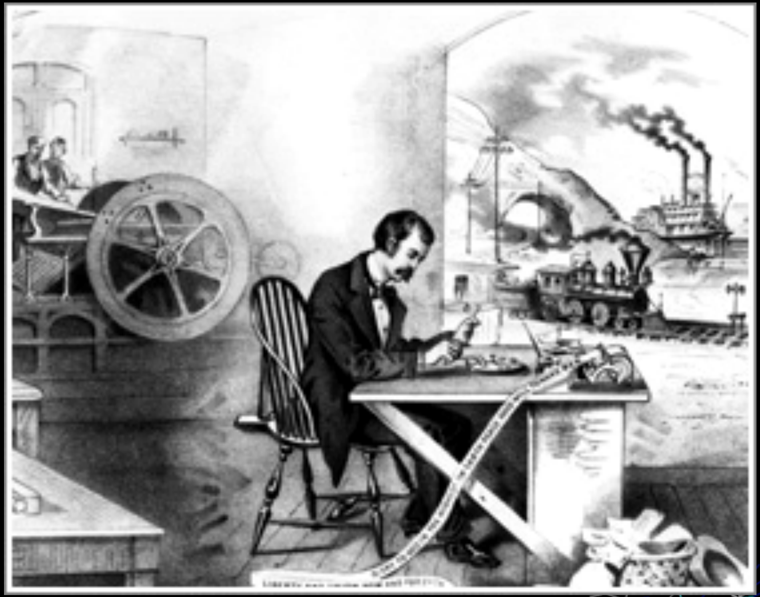
Caltech

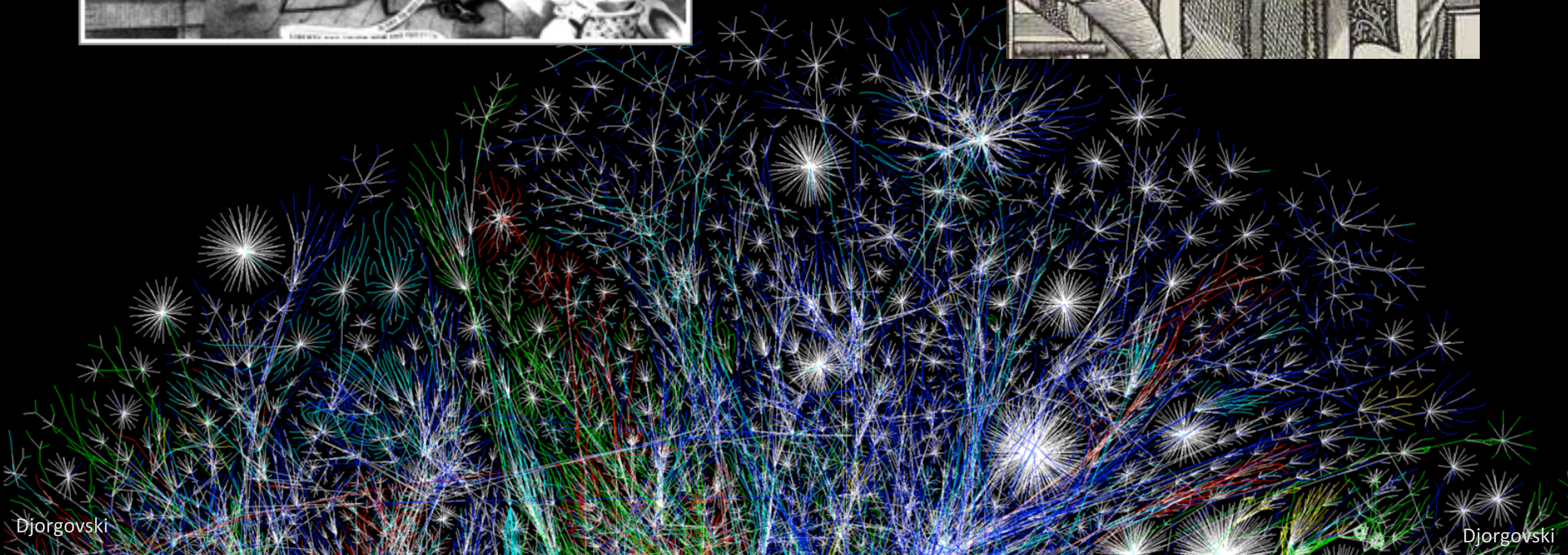
CENTER FOR DATA-DRIVEN DISCOVERY

# Overview

- Setting the stage: an ongoing transformation of science
- Astronomy in the era of an exponential data growth: from Virtual Observatory to Astroinformatics
- Exploration of parameter spaces and other outstanding challenges
- Science on the carbon-silicon interface: the rise of the machines
- Methodology transfer in action
- Concluding musings and comments

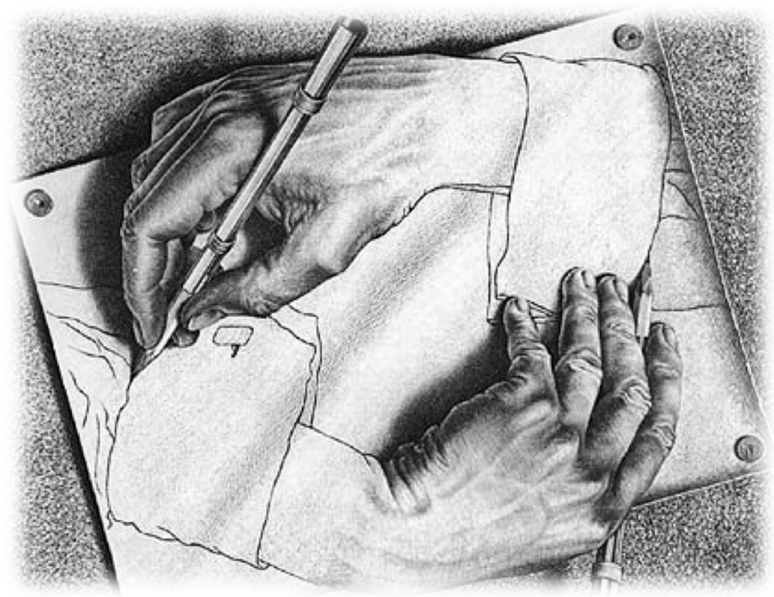
# These are Extraordinary Times



# Transformation and Synergy

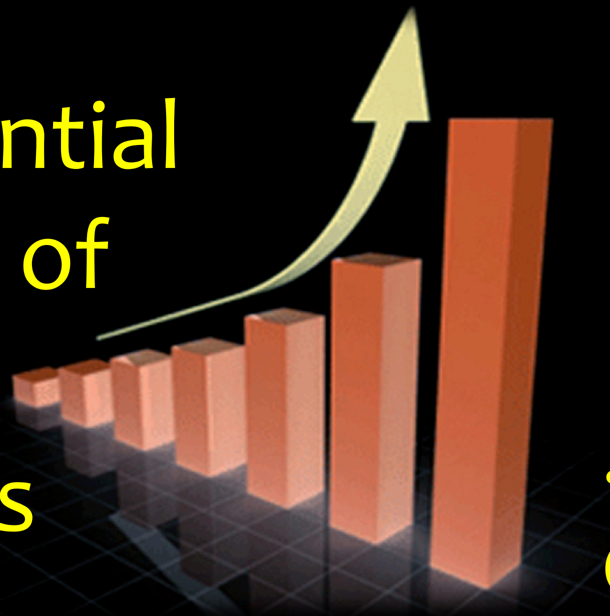
- ***All science*** in the 21st century is becoming cyber-science (aka e-Science) - and with this change comes the need for ***a new scientific methodology***
- The challenges we are tackling:
  - Management of large, complex, distributed data sets
  - Effective exploration of such data ➔ new knowledge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



***Cyberspace*** (today the Web, with all the information and tools it connects) is increasingly becoming the principal arena where humans interact with each other, with the world of information, where they work, learn, and play

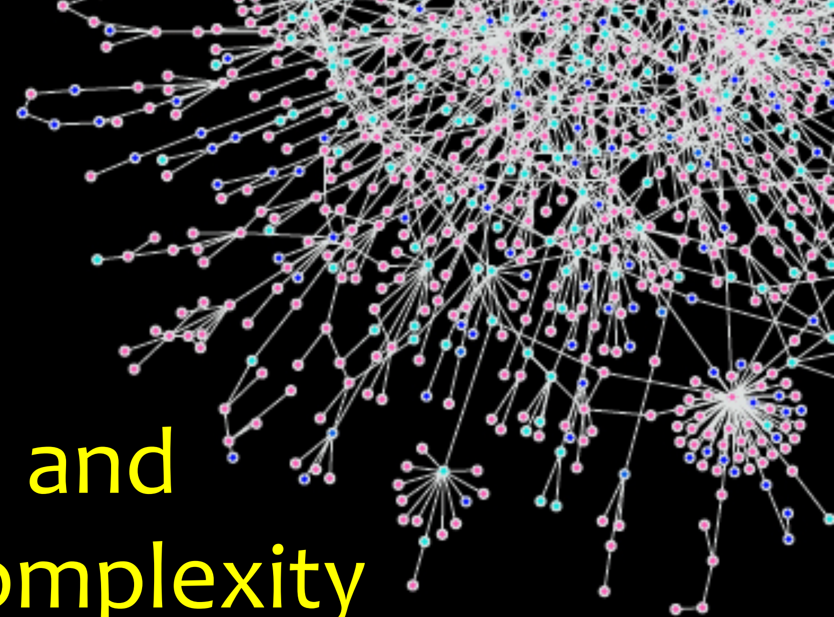
Essentially all aspects of the modern society are migrating to cyberspace, science and scholarship included, with their data, methods, publications, etc.

# 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

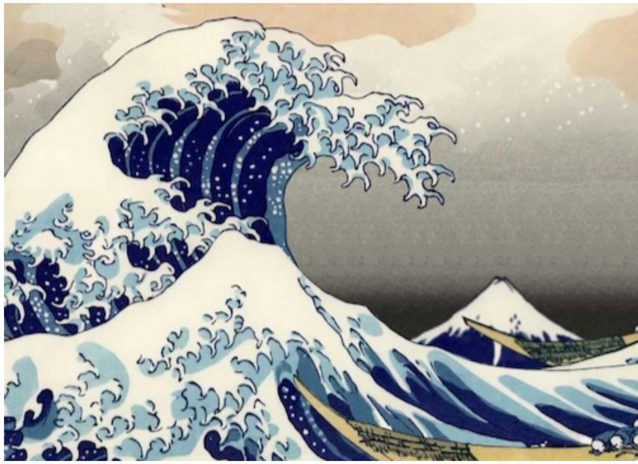
From centralized to distributed resources

From ownership of data to ownership of expertise

Djorgovski

# What is Fundamentally New Here?

- The ***information volumes and rates*** grow exponentially



➡ ***Most data will never be seen by humans***

- A great increase in the data ***information content***

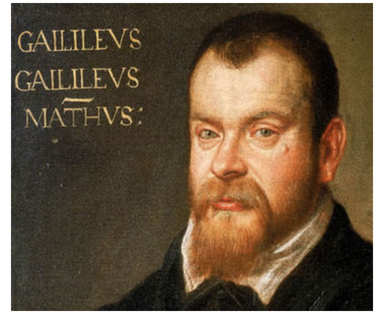
➡ ***Data driven vs. hypothesis driven science***

- A great increase in the ***information complexity***

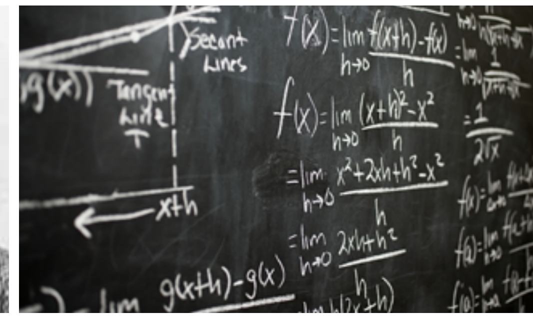
➡ ***There are patterns in the data that cannot be comprehended by humans directly***

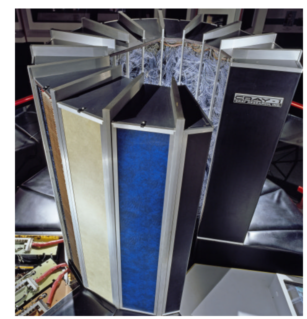
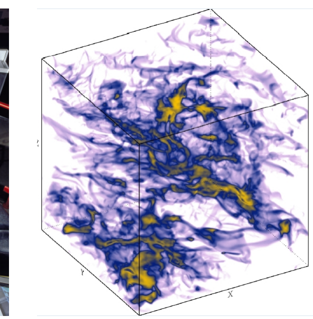
# The Evolving Paths to Knowledge

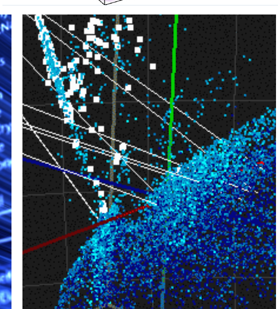
- The First Paradigm: Experiment/ Measurement
- The Second Paradigm: Analytical Theory
- The Third Paradigm: Numerical Simulations
- The Fourth Paradigm: Data-Driven Science

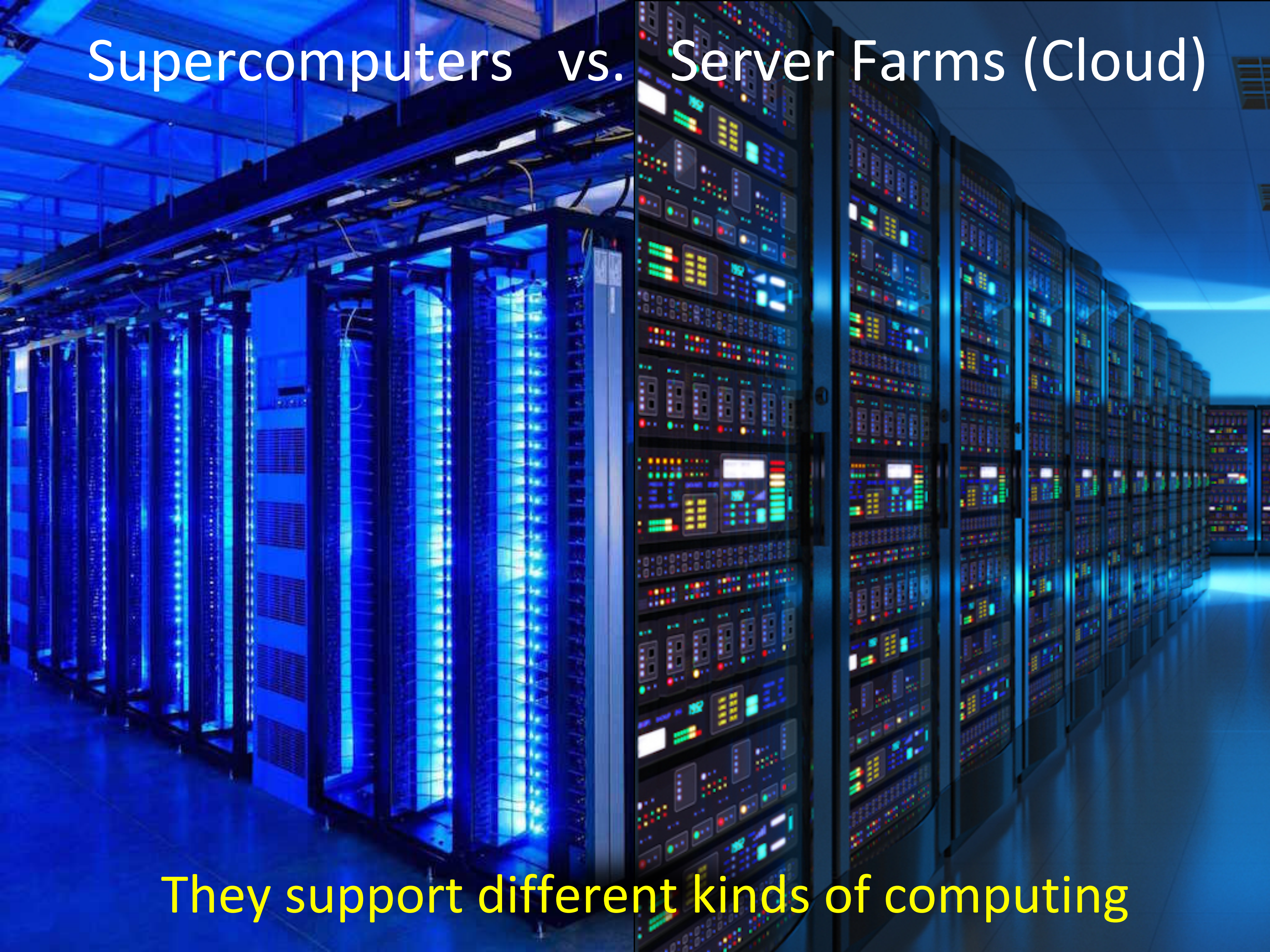




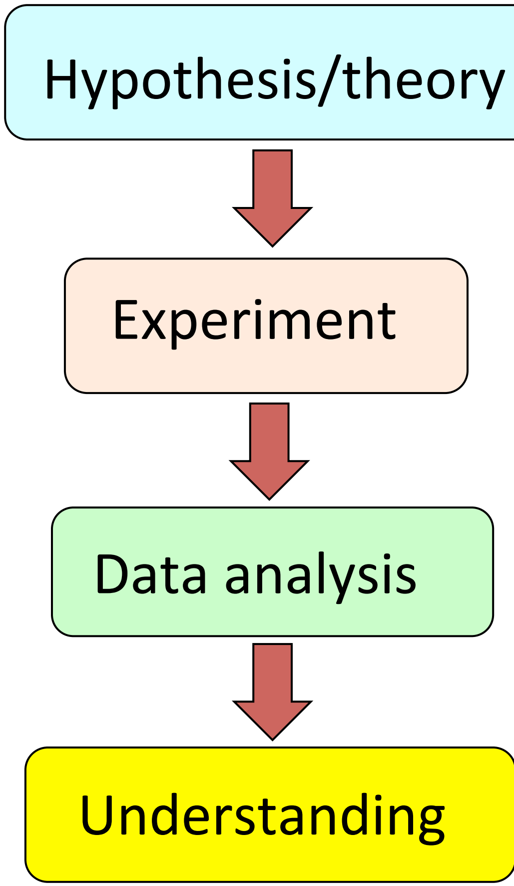


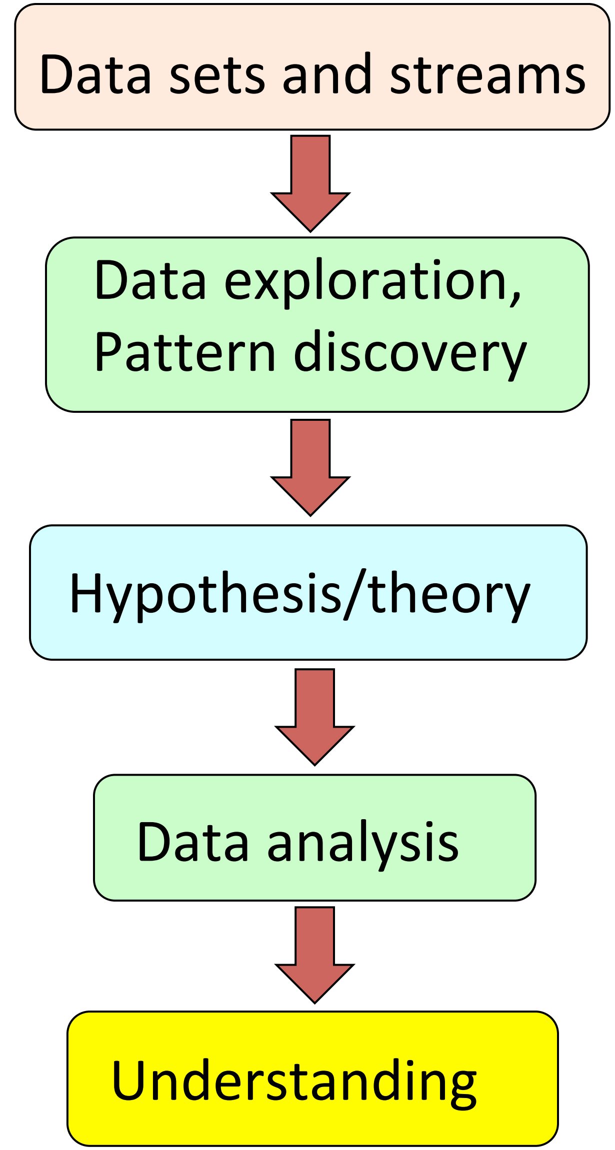




Supercomputers vs. Server Farms (Cloud)
They support different kinds of computing

## Hypothesis-driven science




## Data-driven science




The two approaches are complementary

Djorgovski

# A Modern Scientific Discovery Process

**Data Gathering** (finstruments, sensor networks, their pipelines…)

↳ **Data Farming:**

- Storage/Archiving
- Indexing, Searchability
- Data Fusion, Interoperability

} Databases, Data grids

↳ **Data Mining**

Key Technical Challenges ⇒

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

↳ **Data Understanding**

↳ **New Knowledge**

+feedback

Djorgovski

# Astronomy Has Become Very Data-Rich

- Typical digital sky surveys now generate ~ 1PB each, plus a comparable amount of derived data products
  - EB-scale data sets are on the horizon (e.g., SKA)
- Astronomy today has > 100 PB of archived data, and generates > 100 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 (?)

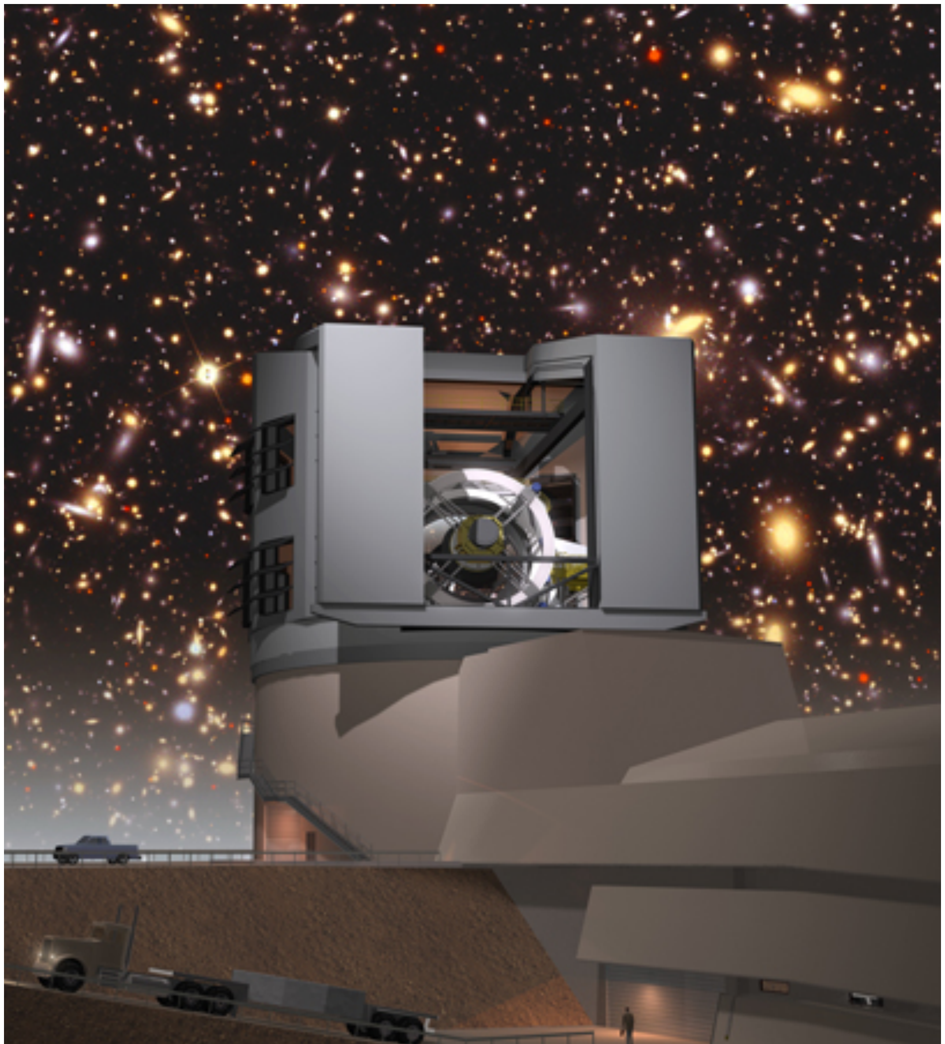
  1 TB ~ 2 million books

  Human Bandwidth ~ 1 TB / year (±)

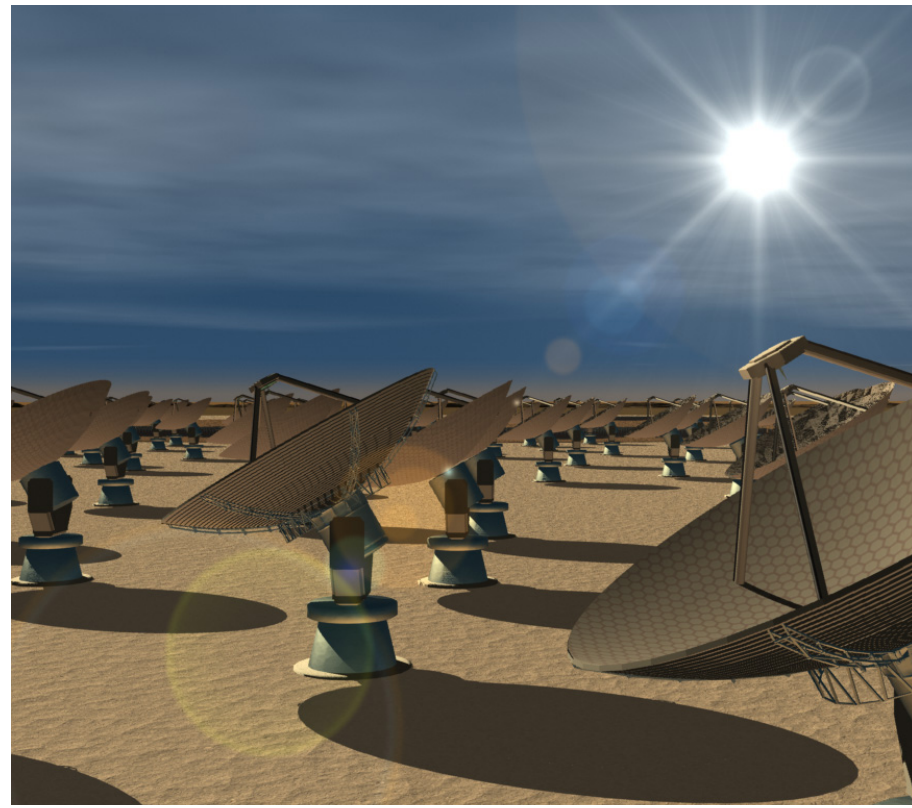

Djorgovski

# … And It Will Get Much More So

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



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

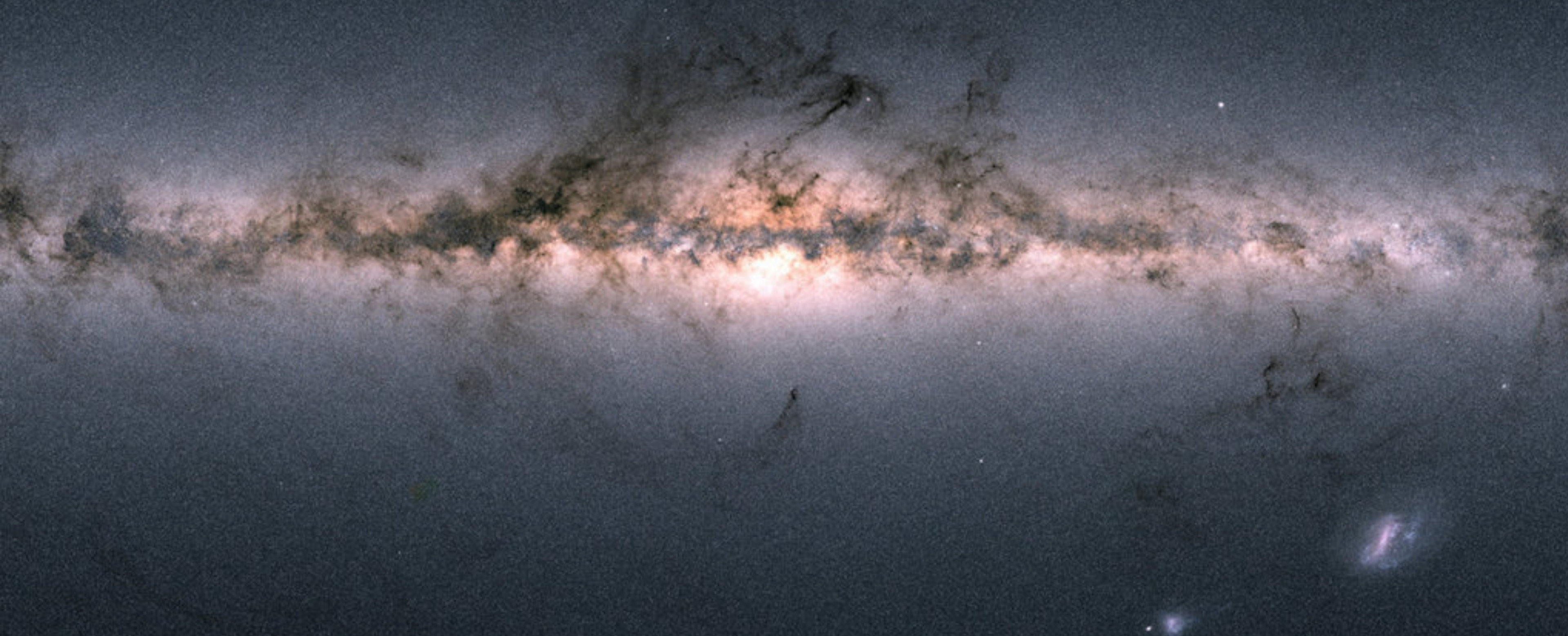


***Data triage*** becomes an issue

Djorgovski

# There Are *Lots* Of Stars In The Sky...

Modern sky surveys obtain ~ $10^{15}$ – $10^{16}$ bytes of images,
catalog ~ $10^{9}$ objects (stars, galaxies, etc.),
and measure ~ $10^{2}$ – $10^{3}$ numbers for each

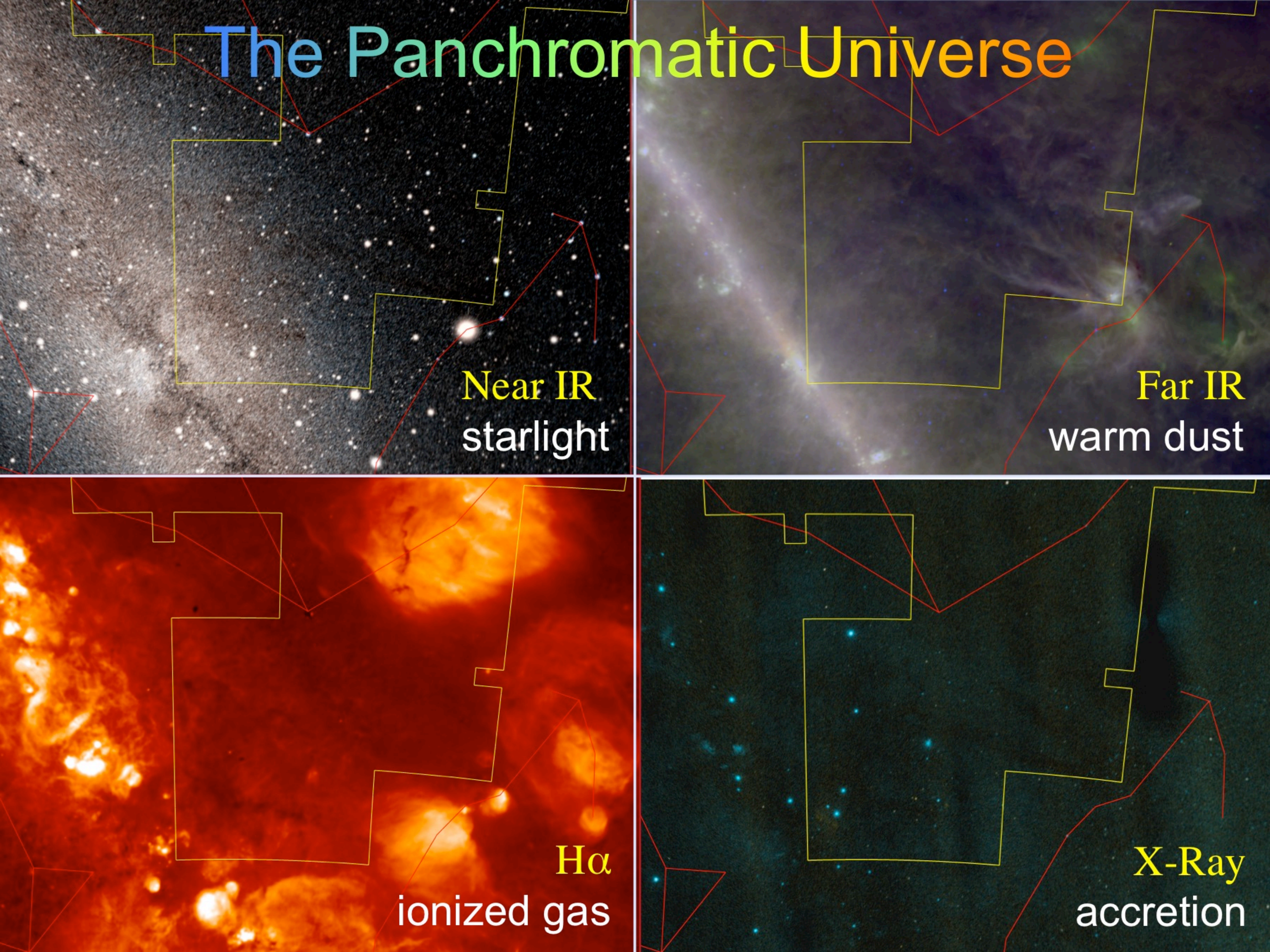


… and then do it again, and again, …

Gaia DR2 Catalog Image

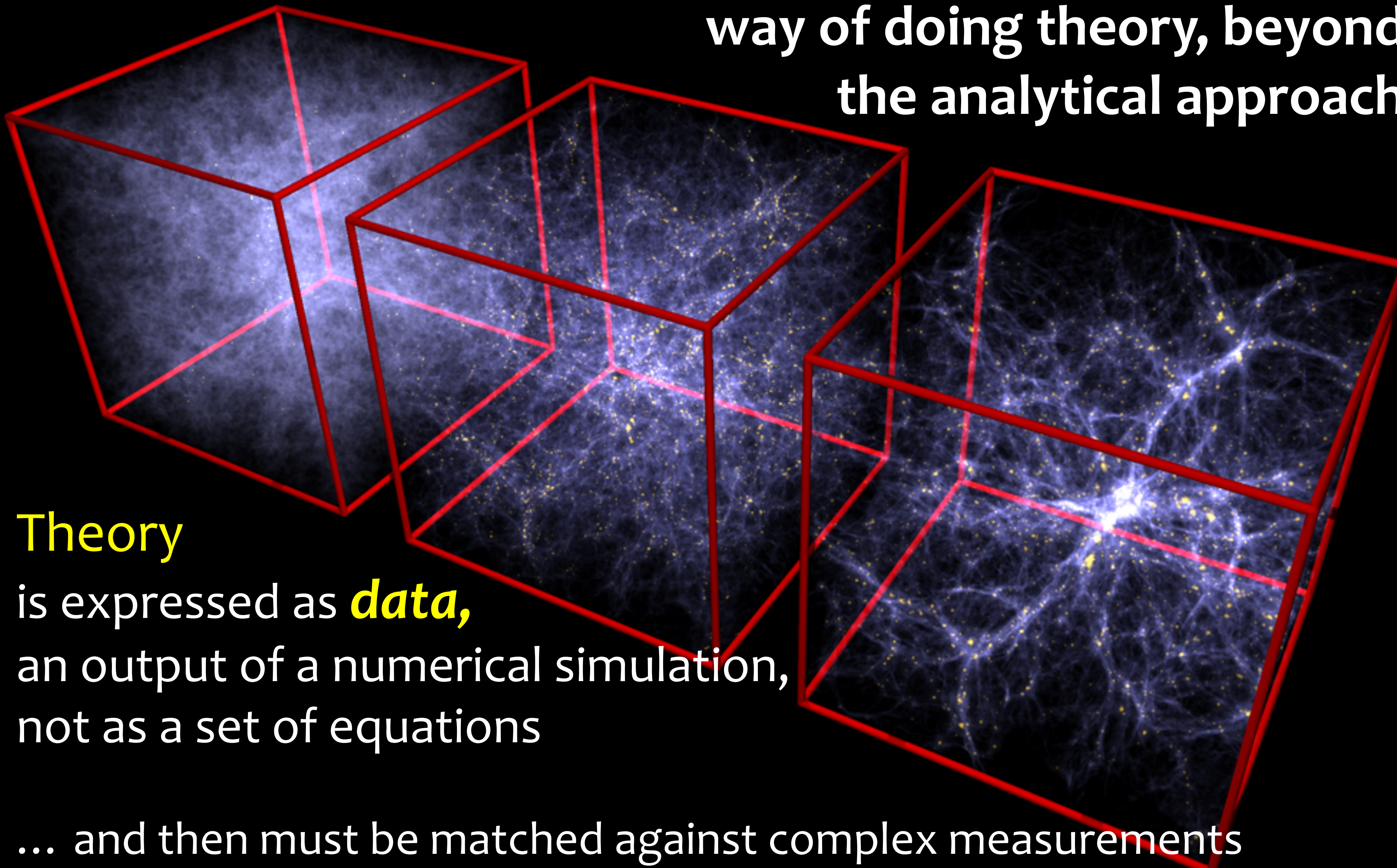
Djorgovski

# The Panchromatic Universe



# Numerical Simulations:

**A qualitatively different and necessary way of doing theory, beyond the analytical approach**



Theory

is expressed as ***data,***

an output of a numerical simulation,

not as a set of equations

… and then must be matched against complex measurements

Djorgovski

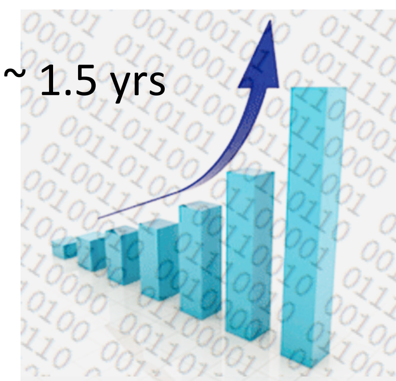
# The Evolving Data-Rich Astronomy

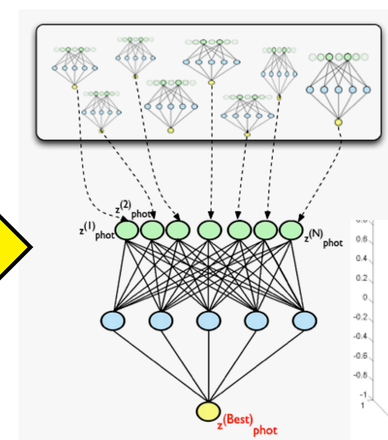
From “arts & crafts” to industry

From data subsistence to an exponential overabundance

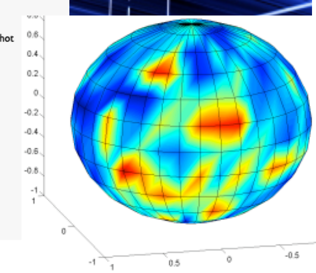

Astronomy is driven by the progress in information technology


$t_2$ ~ 1.5 yrs





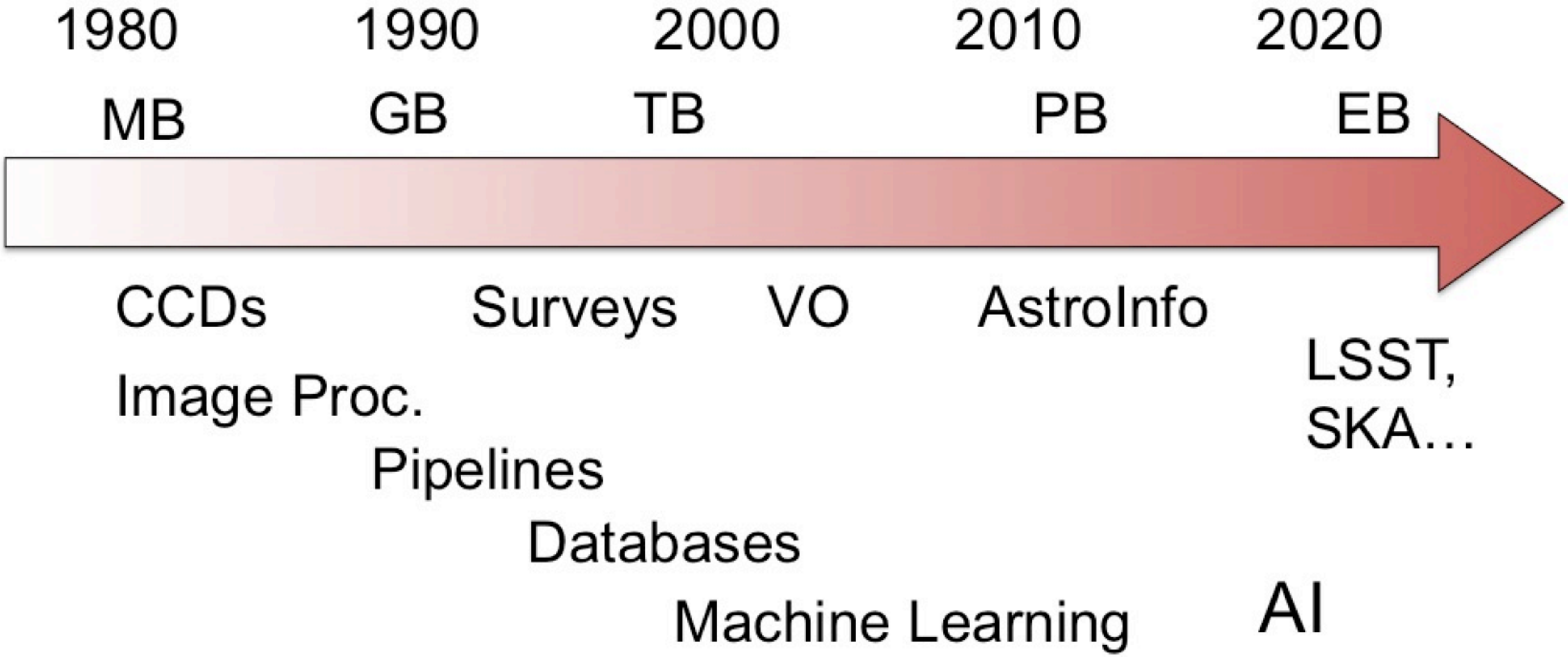




Telescope+instrument are “just” a front end to data systems, where the real action is

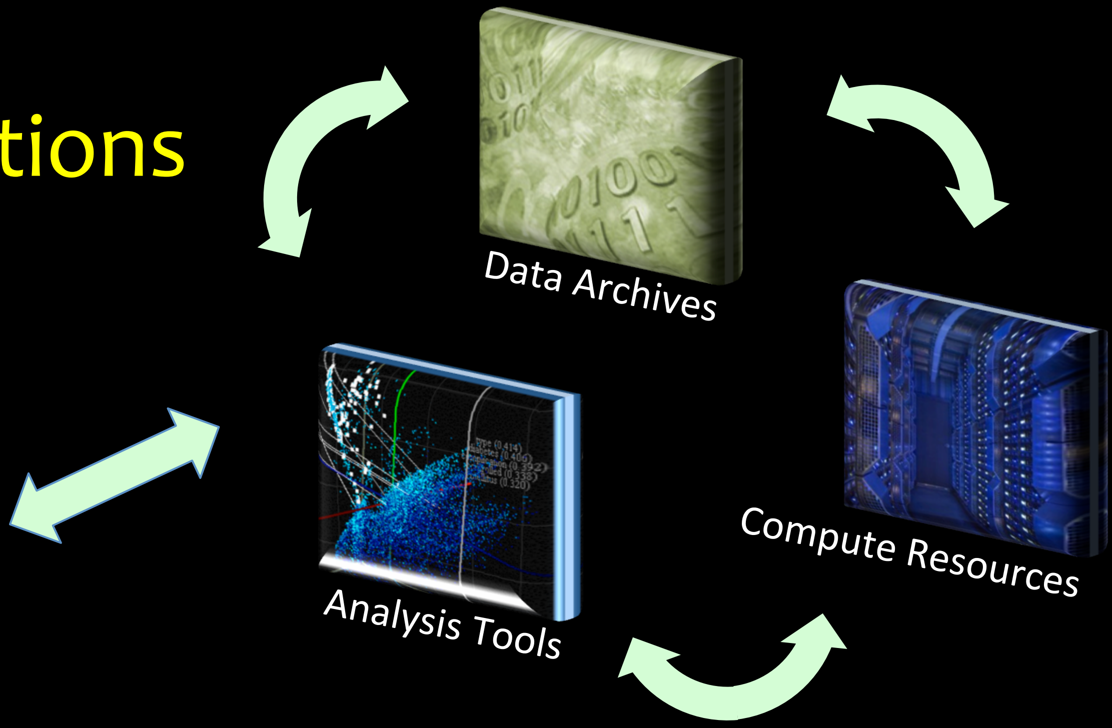
# The Evolving Data-Rich Astronomy

An example of a “Big Data” science driven by the advances in computing/information technology




***Key challenges: data heterogeneity and complexity***

# The Rise of Virtual Scientific Organizations




- A grassroots response to the challenges of the data glut
- A new type of scientific organizations:
  - ✧ Inherently geographically distributed (data, people, tools)
  - ✧ Discipline-based, not institution-based
  - ✧ Based on an exponentially changing technology and data
  - ✧ Crossing the traditional disciplinary boundaries

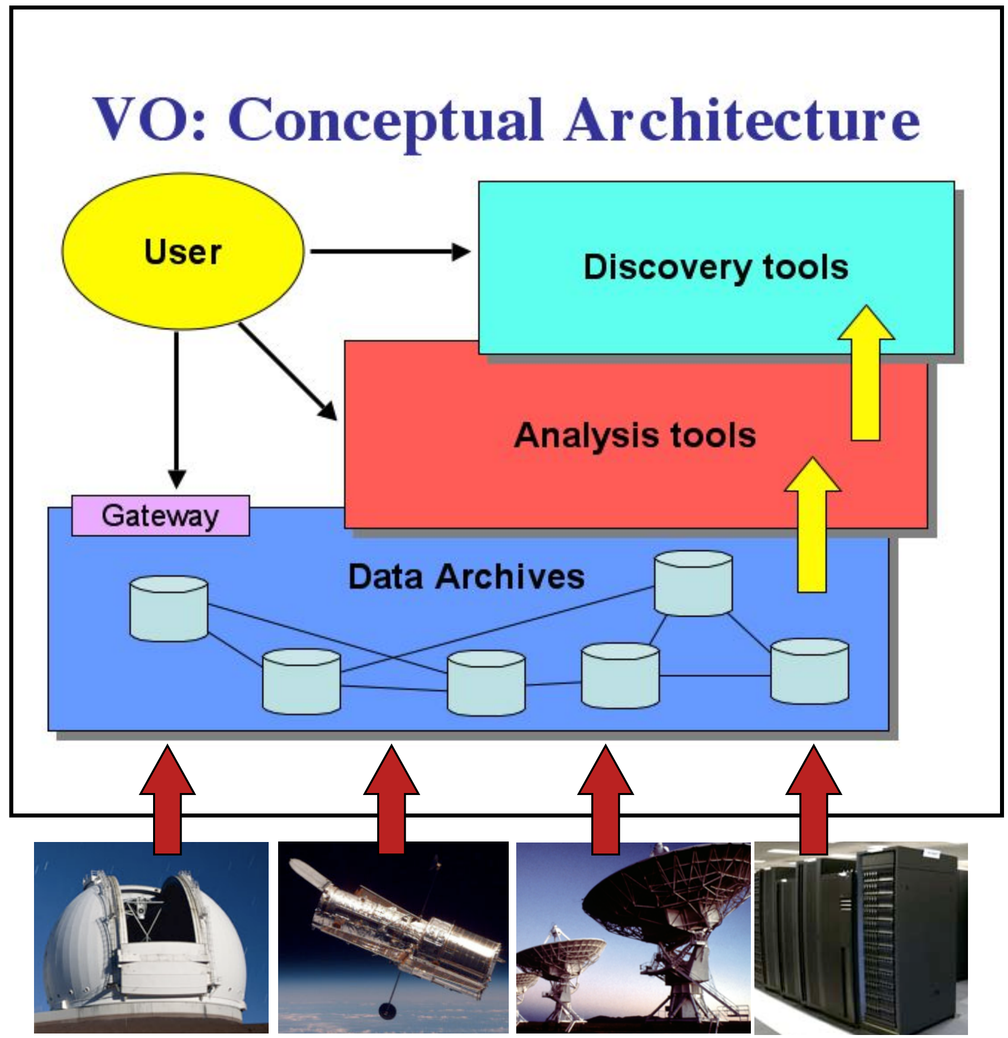
# The Virtual Observatory Concept

- A complete, dynamical, distributed, open ***research environment for the new astronomy with massive and complex data sets***

  – Provide and federate content (data, metadata) services, standards, and analysis/compute services

  – Develop and provide data exploration and discovery tools

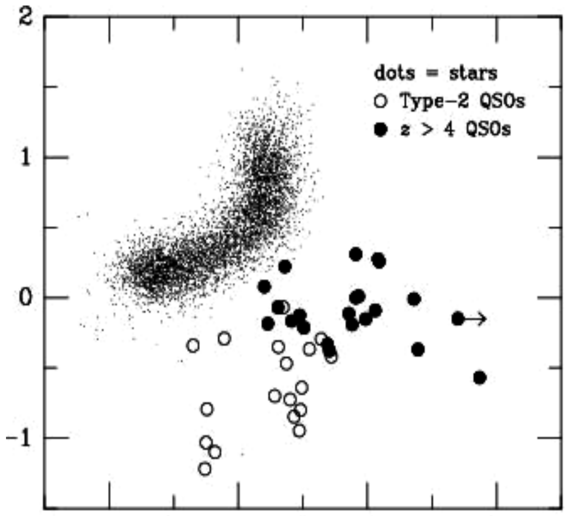
  – A successful example of an e-Science /Cyber-Infrastructure



# Virtual Observatory Science Examples

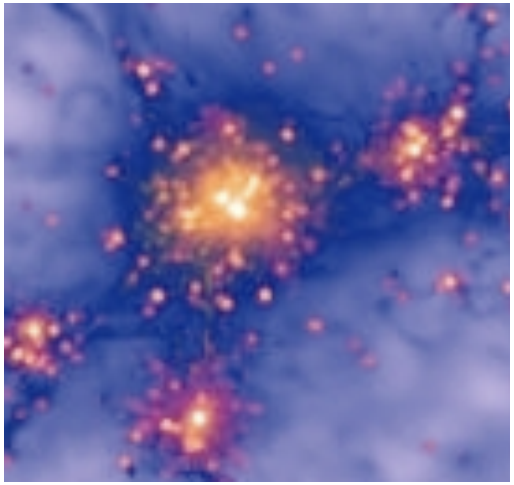
Combine the data from multi-TB, billion-object surveys in the optical, IR, radio, X-ray, etc.

- Large scale structure in the universe
- Structure of our Galaxy




Discover rare and unusual (one-in-a-million or one-in-a-billion) types of sources

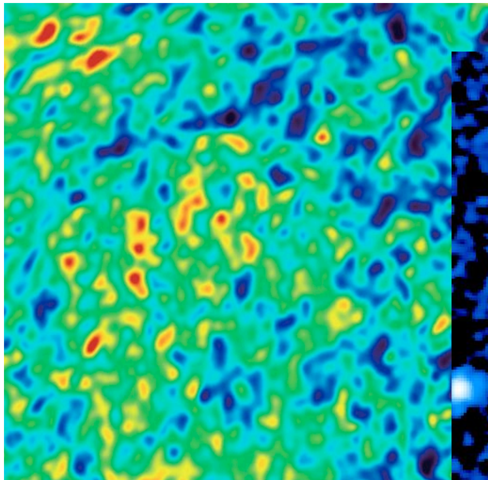
- E.g., extremely distant or unusual quasars, new types, etc.

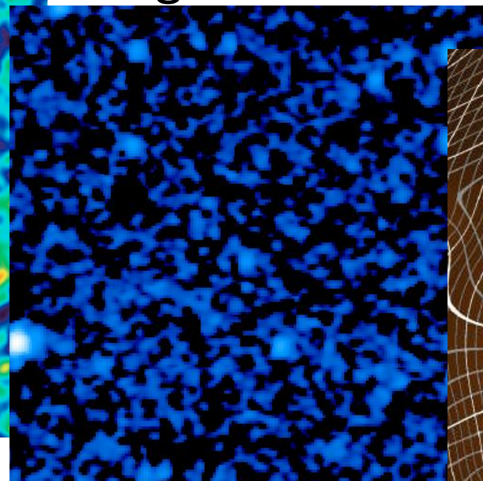


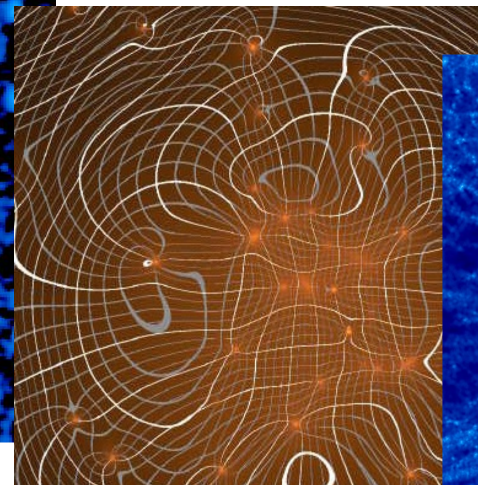
Match Peta-scale numerical simulations of star or galaxy formation with equally large and complex observations

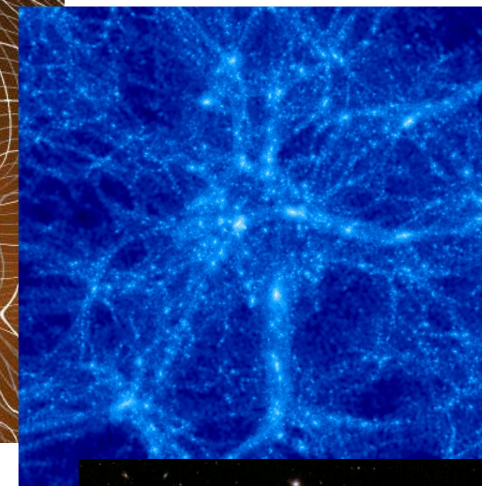
*... etc., etc.*

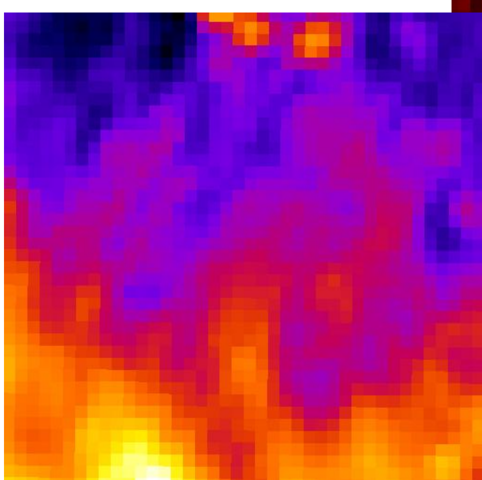
# Understanding the Cosmic Microwave Background and its Foregrounds

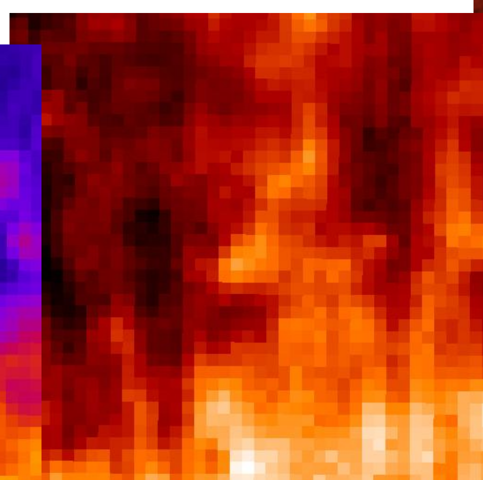

CMB Signal

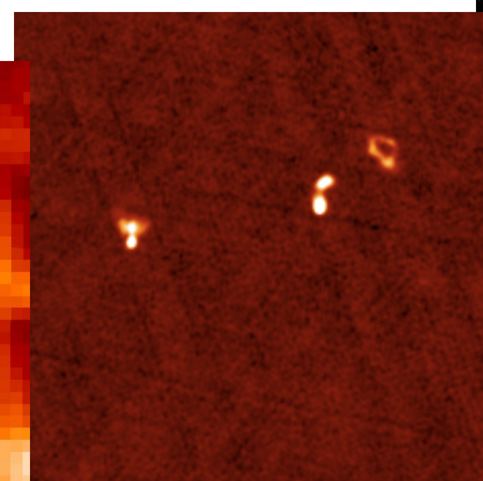
Integrated SZ


Grav. Lensing


Sachs-Wolfe


Gal. Nonthermal


Galactic Thermal



Radio Sources


Galaxies (SF)

Djorgovski

# IVOA: The Virtual Observatory Reified

- Formed in 2002 to facilitate the international collaborative effort needed to enable integrated access to astronomical archives
- 21 international members
- Working Groups and Interest Groups overseen by Technical Coordination Group reporting to Executive Committee:
  - Applications
  - Data Access Layer
  - Data Models
  - Grid and Web Services
  - Registry
  - Semantics
  - Data Curation and Preservation
  - Knowledge Discovery in Databases
  - Education
  - Operations
  - Solar System
  - Theory
  - Time Domain
- Committee for Science Priorities
- Engage with big projects

IVOA.net

# Resources at http://ivoa.net

INTERNATIONAL VIRTUAL
OBSERVATORY ALLIANCE

Home | Astronomers | Deployers | Members | About

## VO Applications for Astronomers

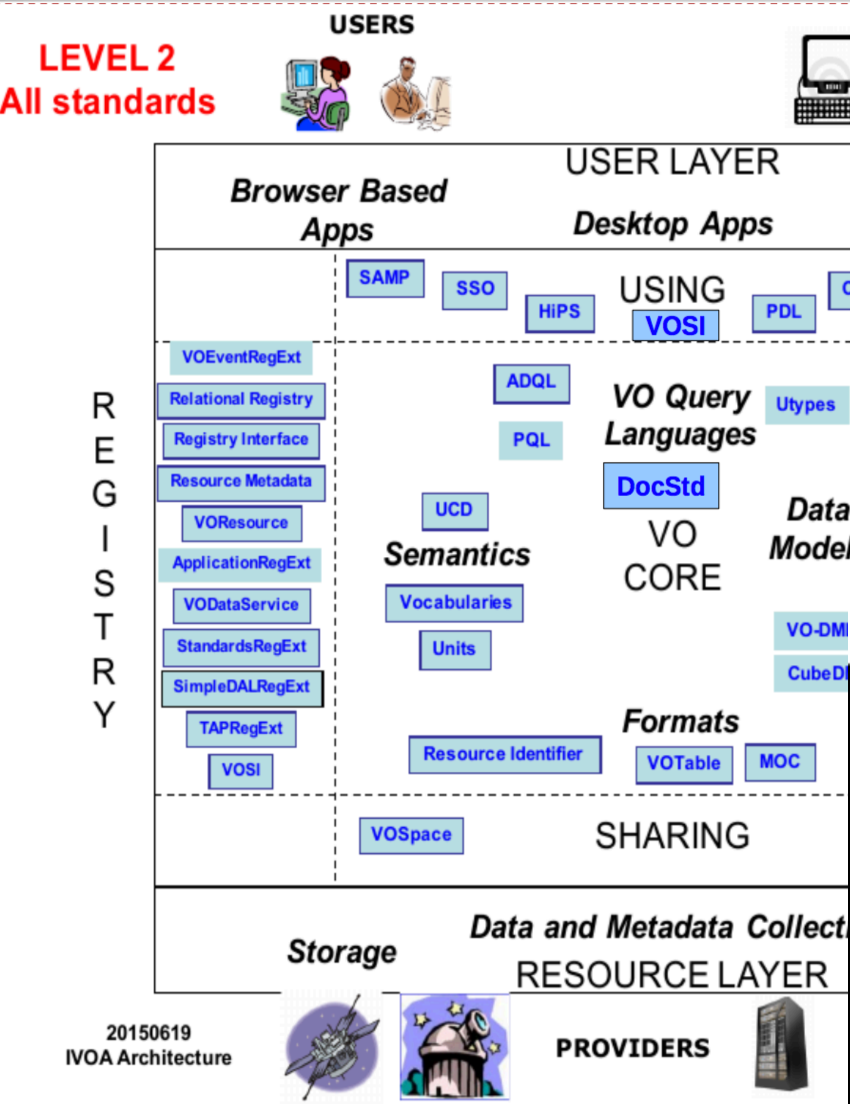
In this section, scientists can find available VO-compatible applications for their immediate use to do science. The level of maturity of the applications depends on a high degree on the level of maturity of the corresponding IVOA protocols and standards.. As a consequence of the flexibility of the standards, several of the applications might overlap in functionality. **The IVOA does not manage or guarantee these services/tools.**

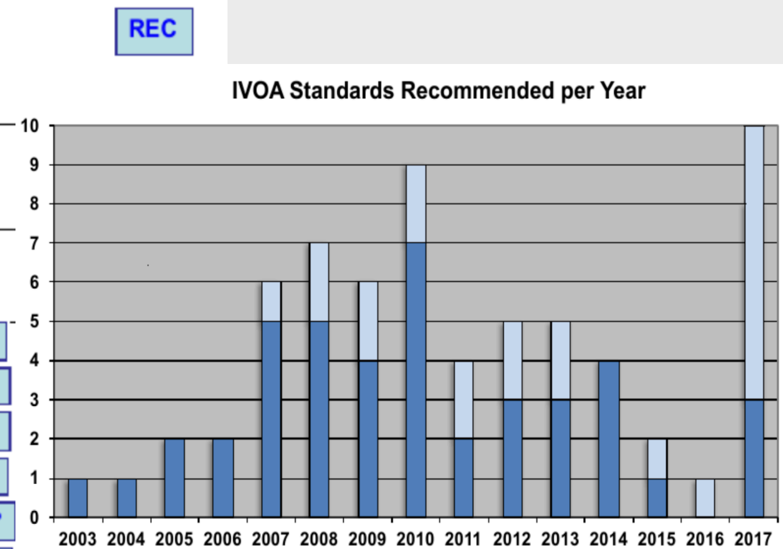
| Applications (in alphabetical order) | Functionality | VO-compliant Tools & Services |
|---|---|---|
| Aladin<br>AppLauncher<br>CASSIS<br>CDS Xmatch Service<br>Data Discovery Tool<br>Filter Profile Service<br>Iris<br>Montage<br>Octet<br>SkyView<br>Specview<br>SPLAT<br>TAPHandle<br>… | **Search for Images:**<br>Aladin, Datascope, SkyView, VODesktop, Data Discovery Tool<br>**Search for Spectra:**<br>Aladin, CASSIS, Datascope, SPLAT, Specview, VOServices, VOSpec, Data Discovery Tool<br>**Search for Catalogues:**<br>Aladin, Datascope, TOPCAT, VODesktop, Data Discovery Tool<br>**Search for Time Series**<br>… | DS9: Image visualiasation<br>GOSSIP: SED fitting<br>VirGO: Search for Images and Spectra<br>IRAF: Image Reduction & Analysis<br>World Wide Telescope<br>Gaia - Graphical Astronomy and Image Analysis<br>SIMBAD<br>TESELA<br>VizieR<br>… |

A compilation of tools and services

IVOA is now mainly a standards coordination body

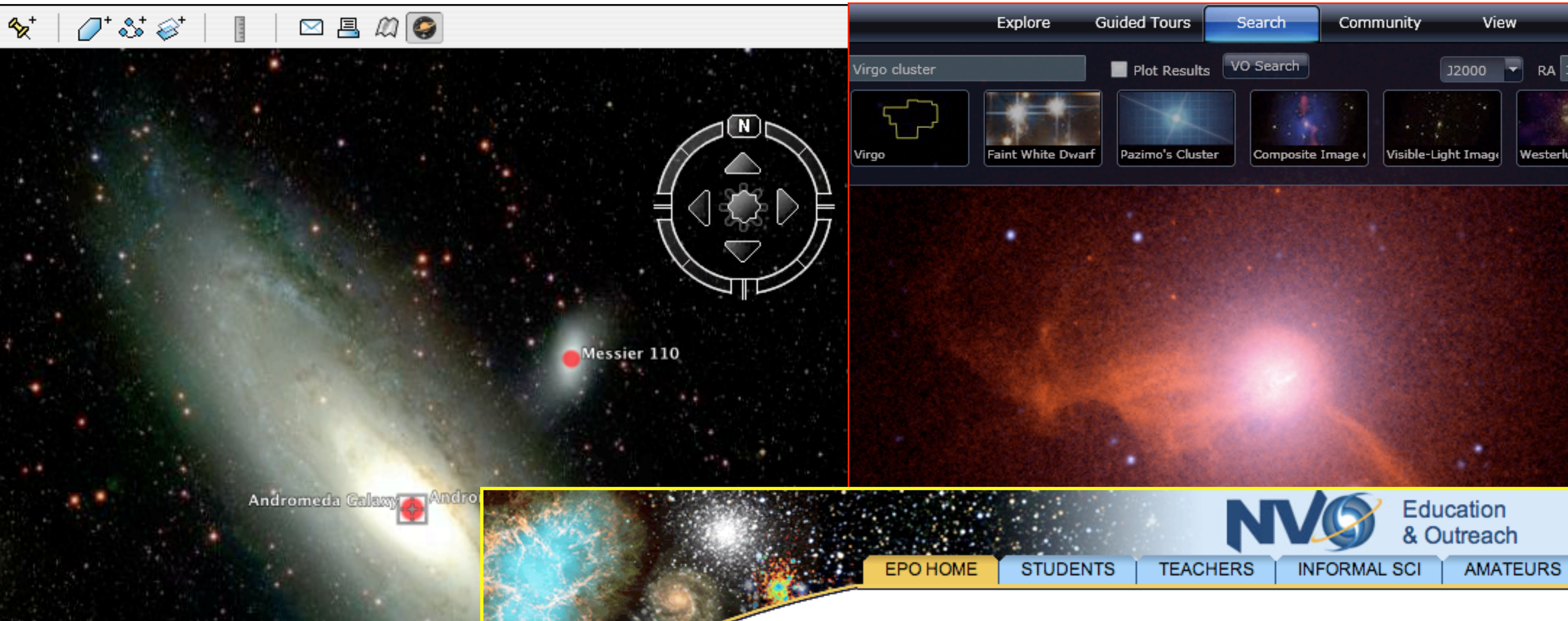
# What has the IVOA achieved?




# VO Education and Public Outreach

## *"Weapons of Mass Instruction"*

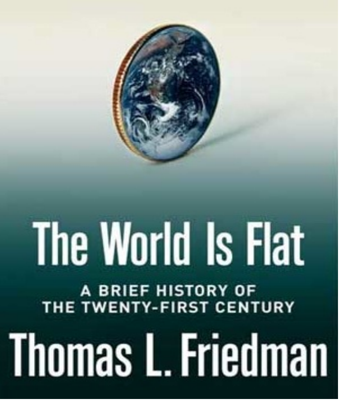



- Unprecedented opportunities in terms of the content, broad geographical and societal range, at all levels
- Astronomy as a gateway to learning about physical science in general, as well as applied CS and IT

***Galaxy M81 seen by a visible-light telescope***

0h42m44.30s Streaming

**This site:** This site is a gateway to education and public




# The Cyberworld Is Also Flat

***Possibly the most important aspect of the IT revolution***

- **Professional Empowerment:** Scientists and students anywhere with an internet connection should be able to do a first-rate science (access to data *and* tools)
  - A broadening of the talent pool democratization of science
  - They can also be substantial contributors, not only consumers of scientific content
- Riding the exponential growth of the IT is far more cost effective than building expensive hardware facilities

… and computational science magnifies their impact

# How Did the VO Succeed?

- All data collected in a digital form
- Computer- and data-savvy community
- Some standard formats in place
- Large data collections in funded, agency mandated archives
- Established culture of data sharing
- Community initiative driven by the needs of an exponential data growth
- Federal agency support/funding
- Data have no commercial value or privacy issues

# VO: Some Lessons Learned

- **Educate your community**. People will share out of an enlightened self-interest. Enlighten them.
- **The uptake is slow**, because:
  A. Cultural inertia: transition from a data poverty to a data glut
  B. Scientists respond to two stimuli:
     1. Resources ⇨ Need agency support, mandates
     2. Results ⇨ ***Need knowledge discovery tools***

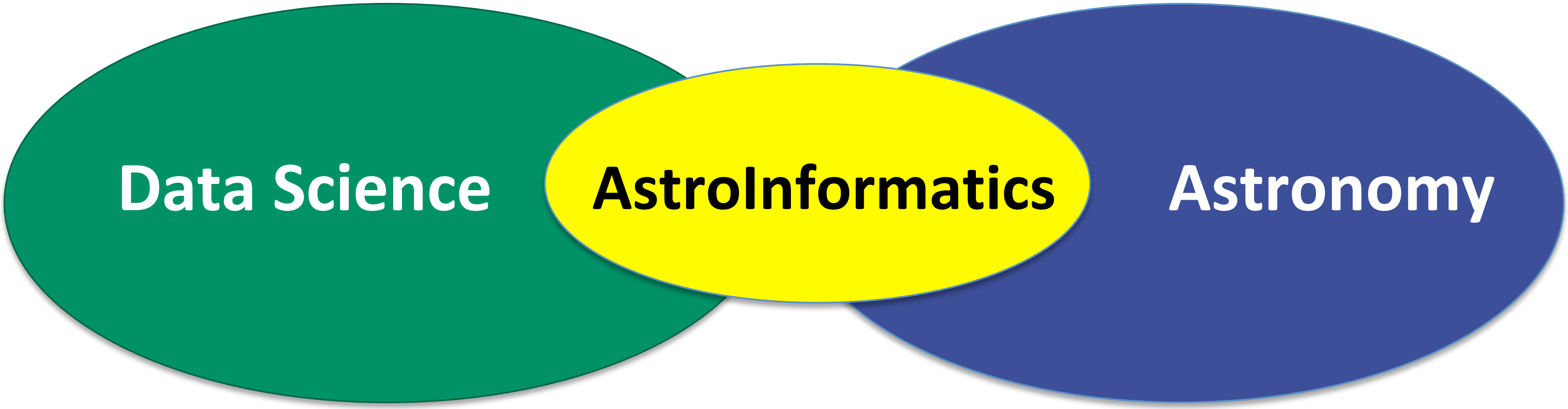
And because of that…

- Don't let the archives people take over! Data commons are essential, but ***only*** because they enable science.

VO ***failed*** at the last bullet. Thus: **Astroinformatics**

# AstroInformatics

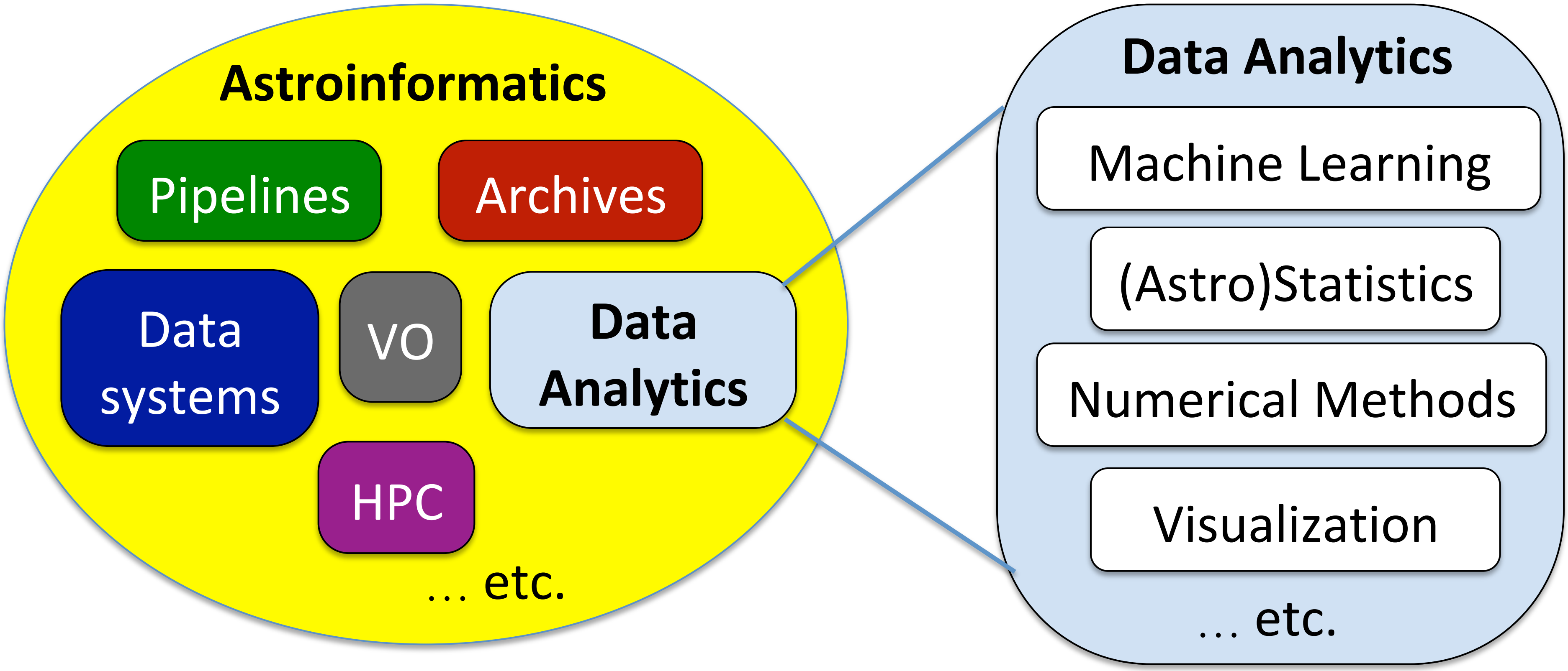
is essentially astronomical applications of Data Science




- While VO became a global data grid of astronomy, astroinformatics focuses of the **knowledge discovery tools**
- It includes a growing community of scientists, both as contributors and as users
- Like other X-Informatics (X = bio, geo, … ) it is a bridge between astronomy and data science, and for the methodology sharing with other fields.

# AstroInformatics

It contains all of the components of Data Science, in their astronomical applications

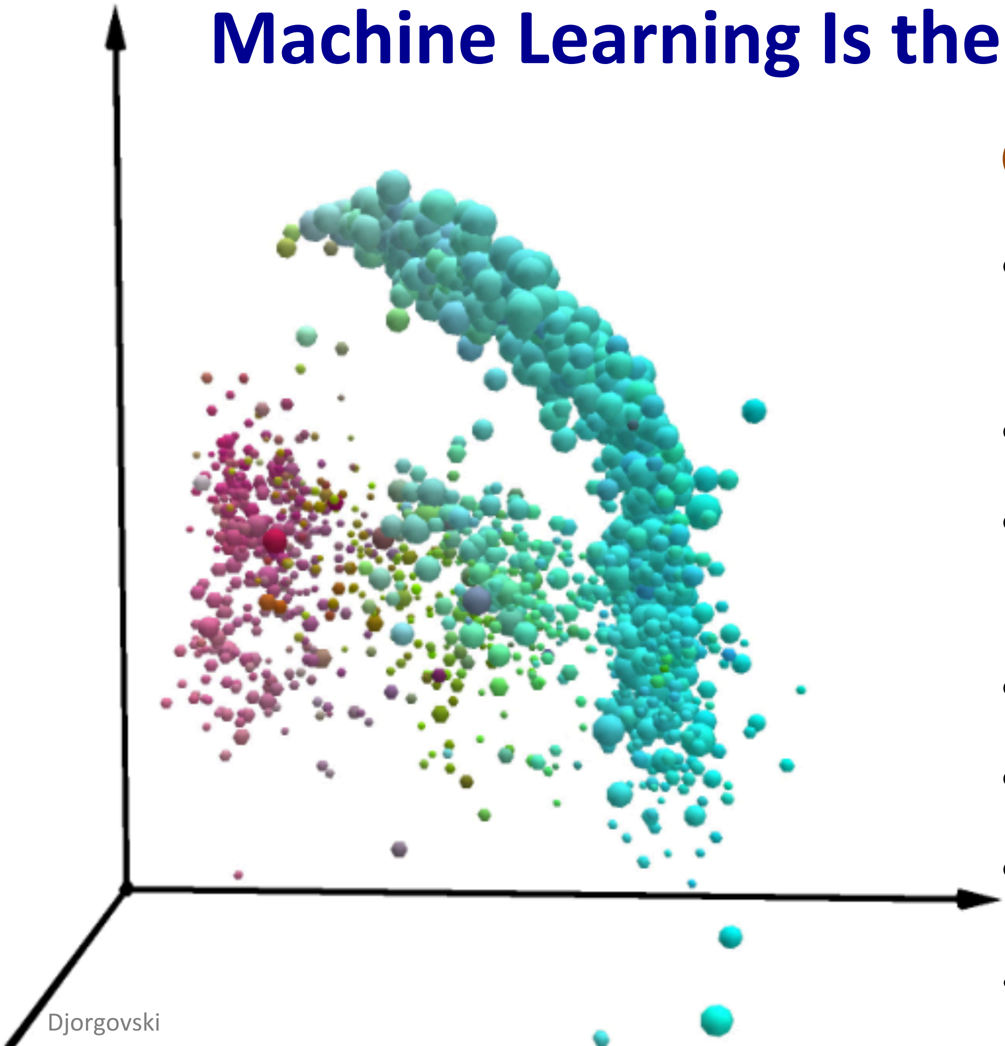



… and their interconnections

# Exploration of Parameter Spaces is a Central Problem of Data Science

Clustering, classification, correlation and outlier searches, …

## Machine Learning Is the Key Methodology



**Challenges:**

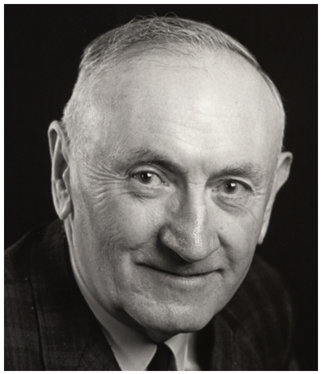
- Algorithm and data model choices
- Data incompleteness
- Feature selection and dimensionality reduction
- Uncertainty estimation
- Scalability
- Visualization

… etc.

} Especially with the data dimensionality

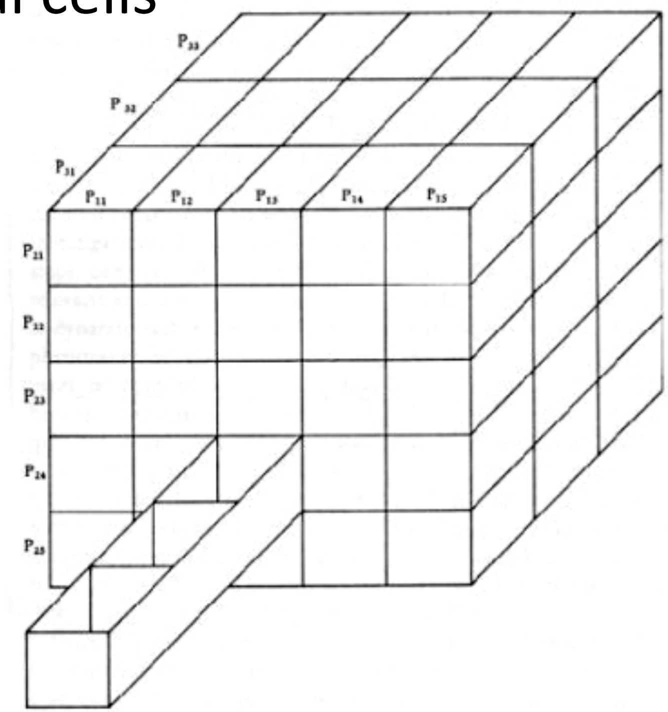
Djorgovski

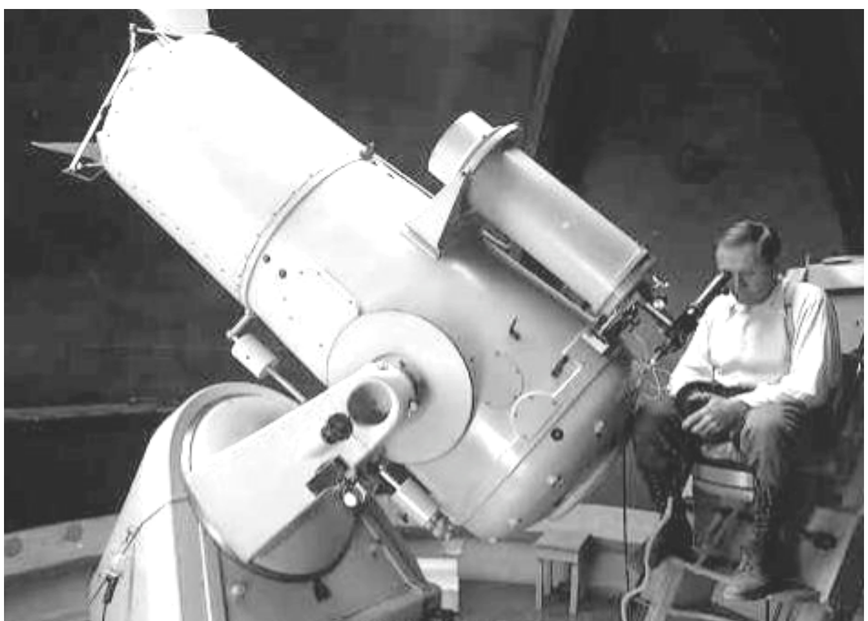
# From "Morphological Box" to the Observable Parameter Spaces



Fritz Zwicky

Zwicky's concept: explore all possible combinations of the relevant parameters in a given problem; these correspond to the individual cells in a "Morphological Box"



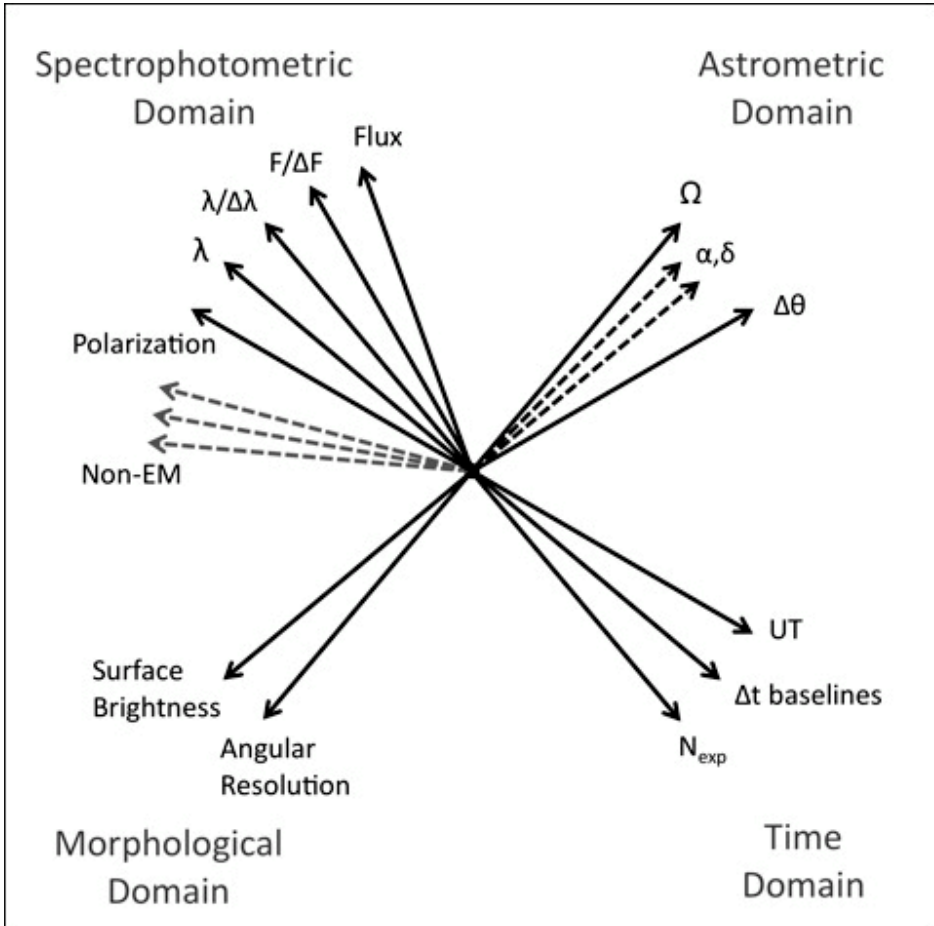


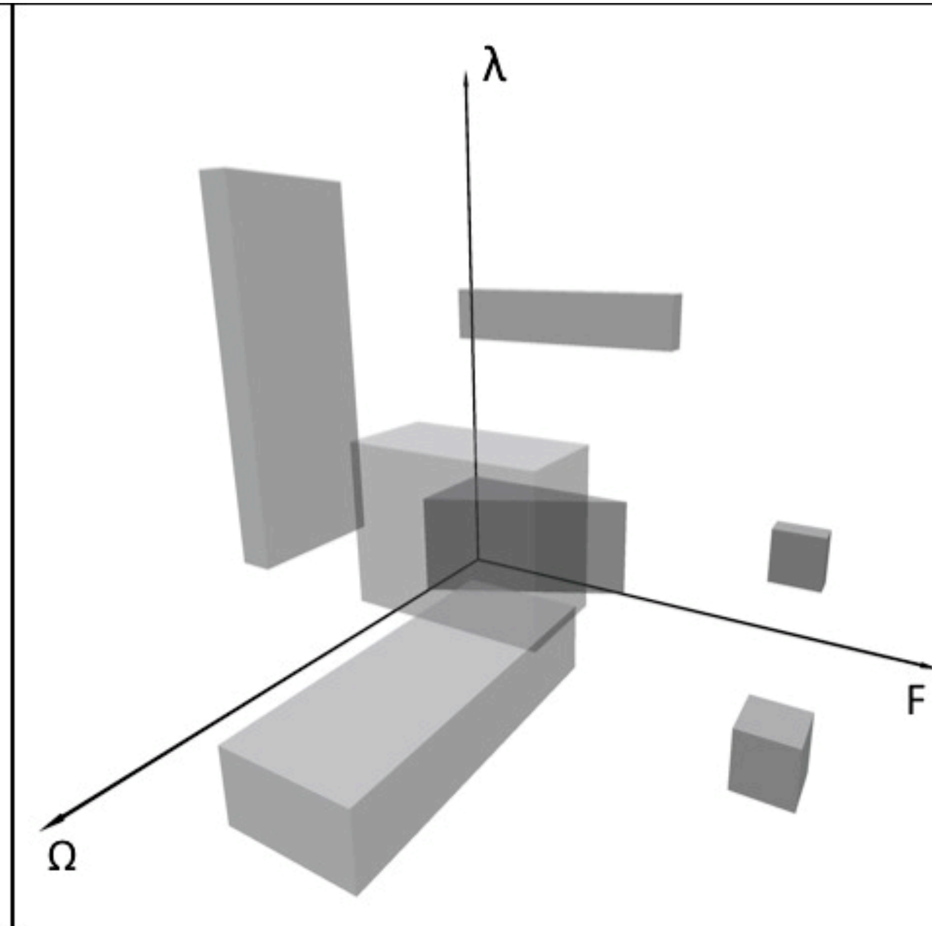
Example: Zwicky's discovery of the compact star-forming dwarfs

# Systematic Exploration of the Observable Parameter Spaces (OPS)

Its axes are defined by the observable quantities

Every observation, surveys included, carves out a hypervolume in the OPS

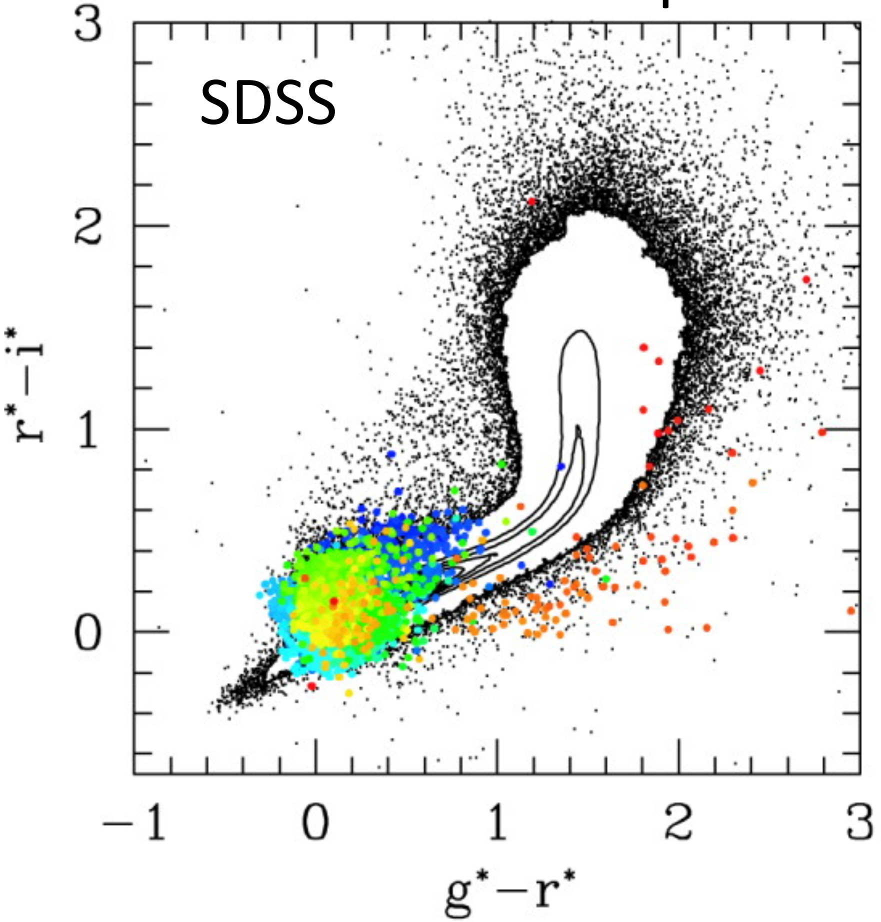



Technology opens new domains of the OPS → New discoveries

## Measurements Parameter Space

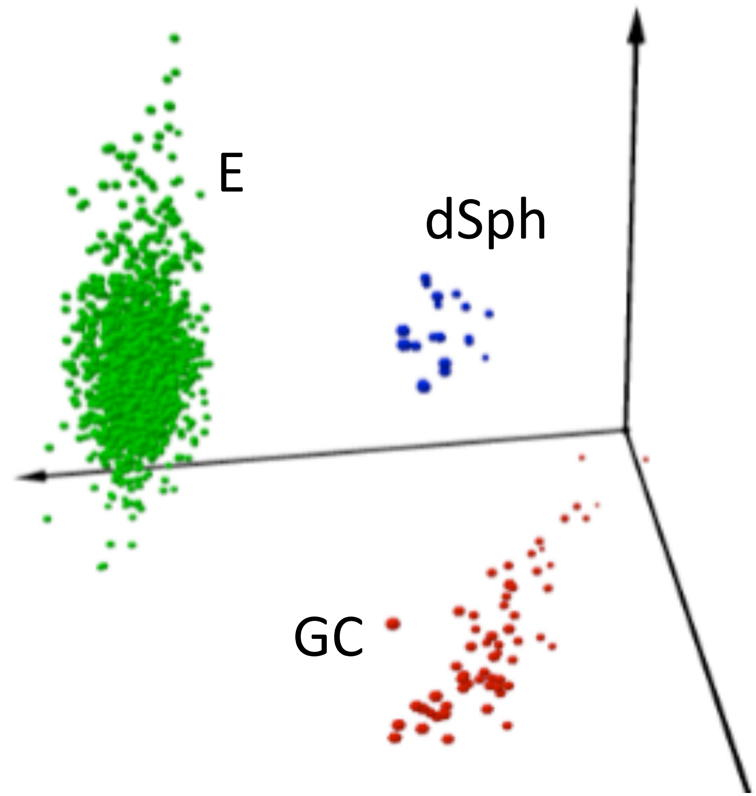
Colors of stars and quasars




Dimensionality ≤ the number of observed quantities

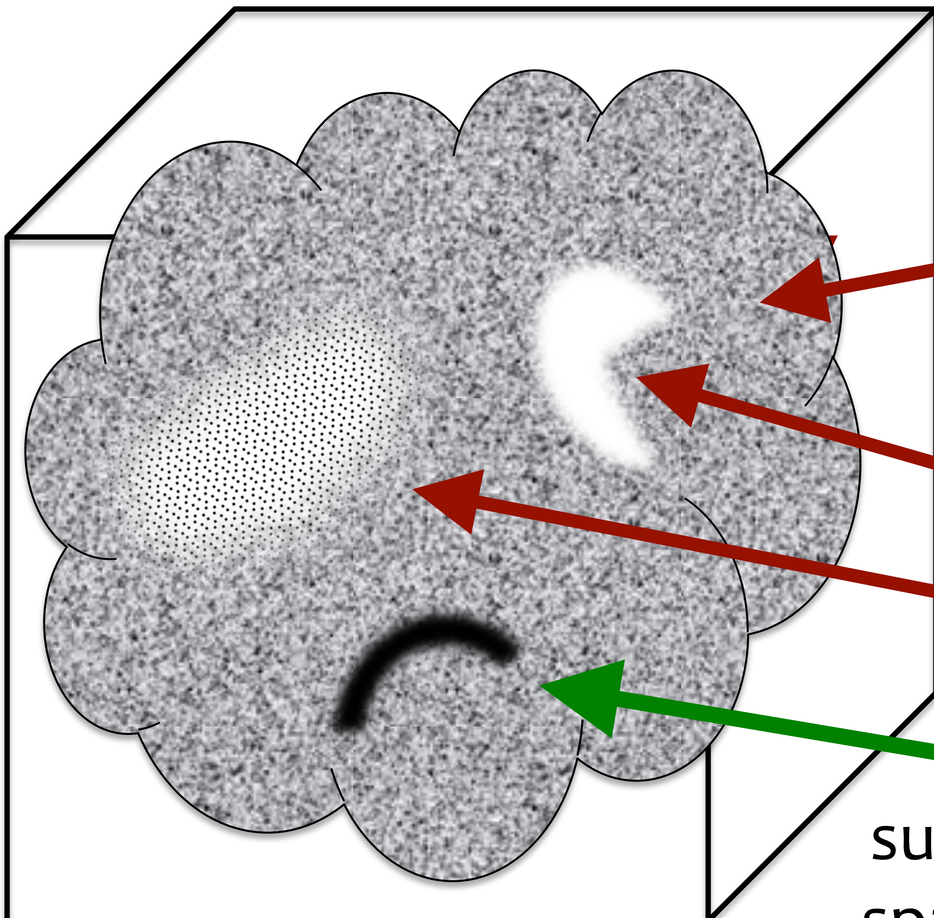
## Physical Parameter Space

Fundamental Plane of hot stellar systems




Both are populated by objects or events

# Pattern or structure (Correlations, Clustering, Outliers, etc.) Discovery in High-Dimensional Parameter Spaces



D >> 3 parameter/feature space hypercube

High-D data cloud: mostly noise, with an arbitrary PDF distribution

Missing data

Data heterogeneity

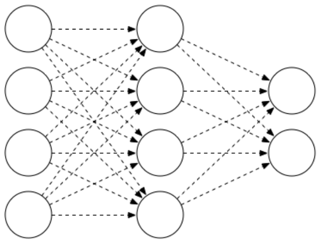
But in some corner of some subset of dimensions of this data space, there is ***something ≠ noise***, i.e., a statistically significant structure with an unknown form

**Mapping the Entropy of Large Data Spaces?**

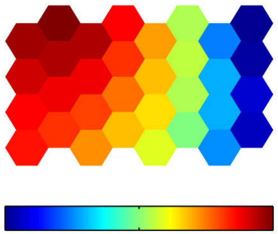
Djorgovski

# Classification, Clustering, and Outliers

- **Supervised learning (classification):** use a known set of objects to train a classifier
  - Hard to find previously unknown things



- **Unsupervised learning (clustering):** let the data tell you how many different kinds of things are there
  - Could find previously unknown types as outliers



**Supervised Algorithms**
Neural Networks (MLP)
Boltzmann Machines
RBM
Decision Trees
Nearest Neighbor
Naive Bayes Classifiers
Bayesian Networks
Gaussian Processes
Regression
…

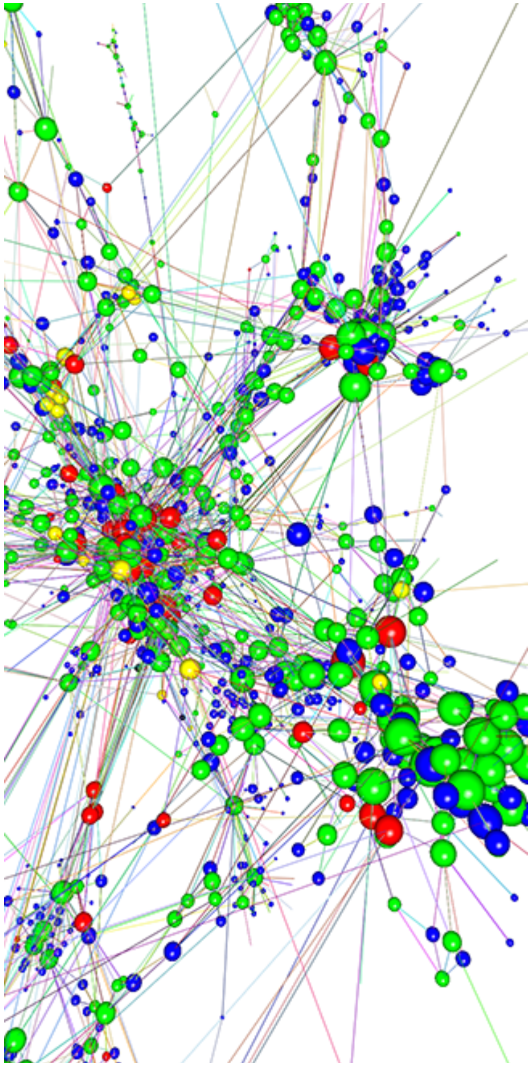
There is **no** "one size fits all": different choices for different problems

**Unsupervised Algorithms**
K-Means
Self-Organizing Maps
RDF
Fuzzy Clustering
CURE
ROCK
Vector Quantization
Probabilistic Principal Surfaces
…

# The principal challenges of knowledge discovery do not come from the data size, but from the data complexity

- How do we recognize highly complex patterns that involve interactions of many variables in many dimensions?
- How do we visualize data spaces with 10's, 100's or 1000's of dimensions?
- How do we decide what algorithms to use in a given situation?
- How do we interpret and explain the results?

⇨ **The key challenges stem from the high dimensionality of data**



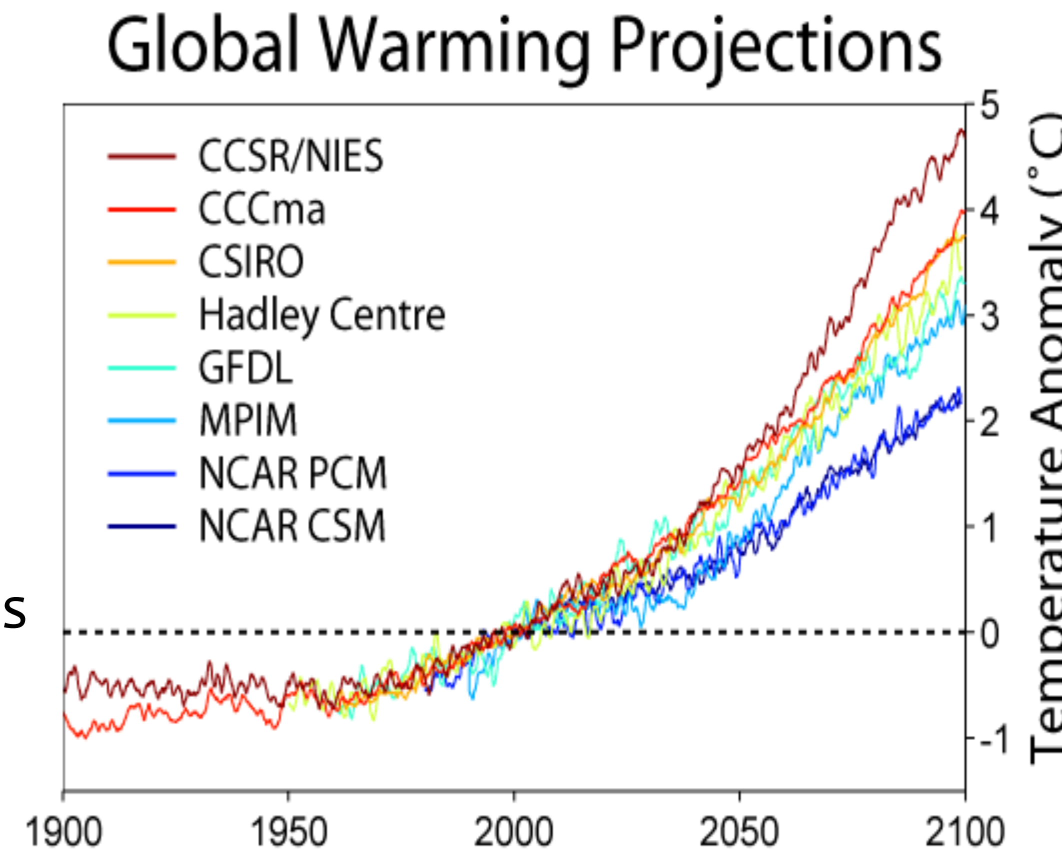
# Quantifying Model Uncertainty

… Whether the data come from measurements or from the output of numerical models and simulations

The sources of uncertainty:

- Measurement errors
- Numerical errors
- Sample sizes
- Processing algorithms
- Data representation
- Data mining choices and their implementations

… etc. etc.

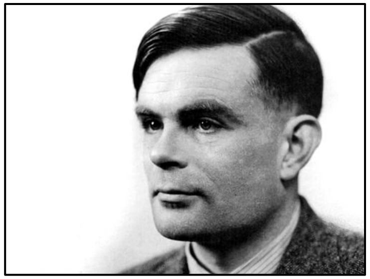


**The key role of data analysis is to replace the raw complexity seen in the data with a reduced set of patterns, regularities, and correlations, leading to their theoretical understanding**

**However, the complexity of data sets and interesting, meaningful constructs in them is *starting to exceed the cognitive capacity of the human brain***

Djorgovski

# A Brief History of AI

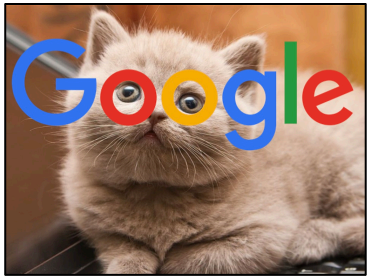


**1950:** A. Turing publishes “Computing Machinery and Intelligence”

***The field of AI/ML starts***

**1960:** J. C. R. Licklider* publishes “Man-Computer Symbiosis” (*You can thank him for the Internet)

**Early 1990’s:** Astronomers start using ML tools



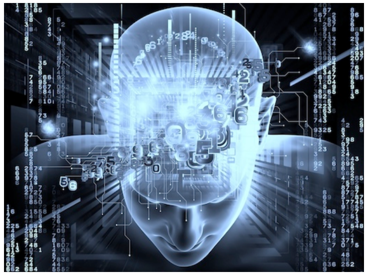

**~1998:** Google starts − common AI use

**1998:** Computer becomes the world chess champion

**2011-2015:** AI talks (Siri, Cortana, Alexa)


**2012:** Google AI learns to recognize pictures of cats

**2016:** Computer becomes the world *Go* champion



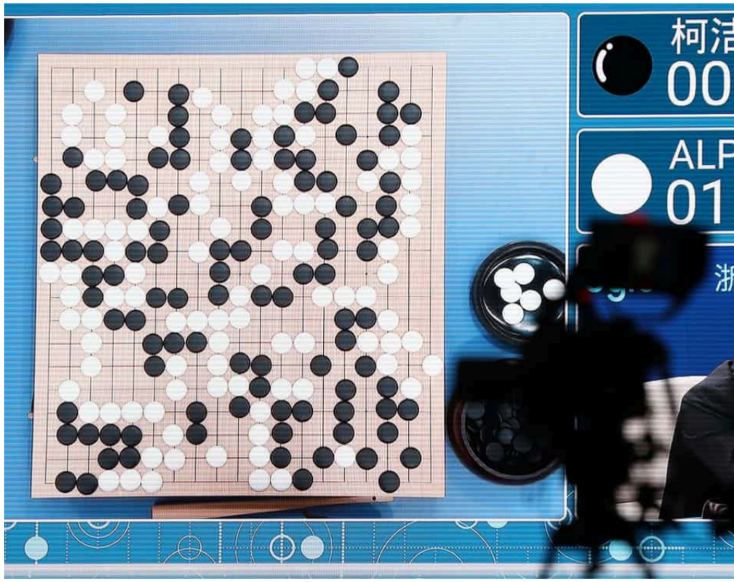
**2017:** A *self-taught AI* beats the previous AI *Go* champion

**Soon?** Collaborative human-computer discovery

# The Rise of the Machines

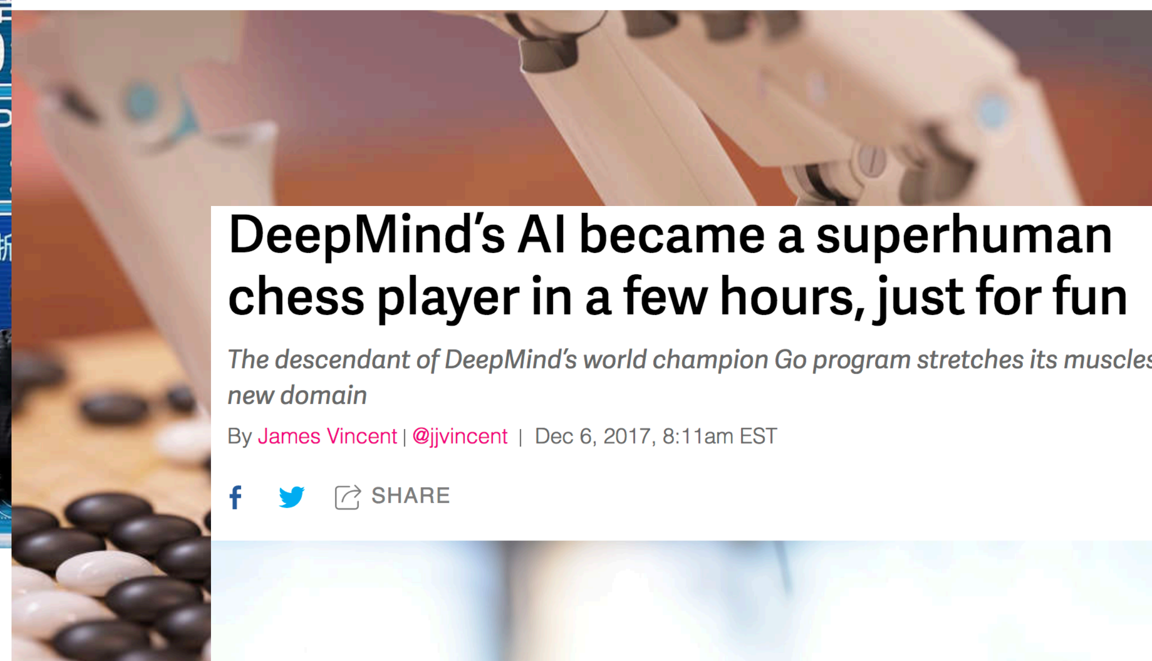
## World's best Go player flummoxed by Google's 'godlike' AlphaGo AI

**Ke Jie, who once boasted he would never be beaten by a co[…] at the ancient Chinese game, said he had 'horrible experie[…]**



## Google's "AlphaGo Zero" AI Taught Itself To Become World Champion In Just Three Days

58 SHARES | Share on Facebook | Share on Twitter | +

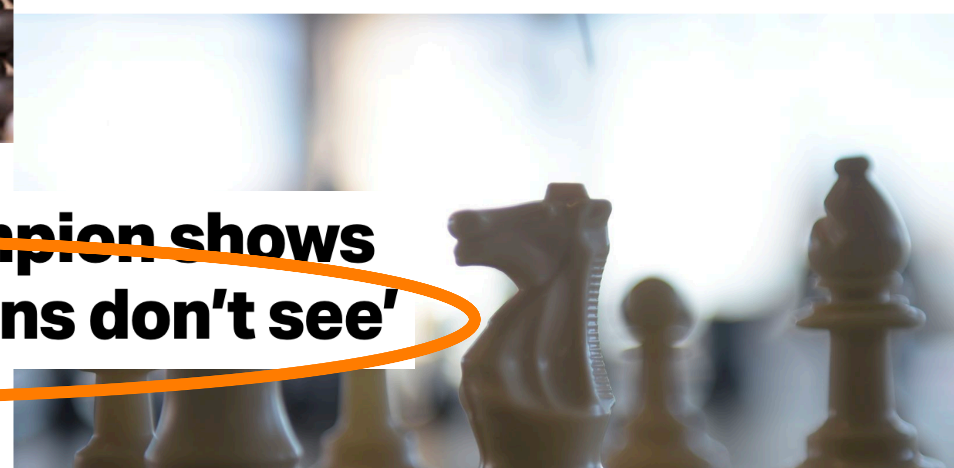


## DeepMind's AI became a superhuman chess player in a few hours, just for fun

*The descendant of DeepMind's world champion Go program stretches its muscles i[…] new domain*

By James Vincent | @jjvincent | Dec 6, 2017, 8:11am EST

SHARE



## Google: Defeating Go champion shows AI can 'find solutions humans don't see'

# What Can Possibly Go Wrong?



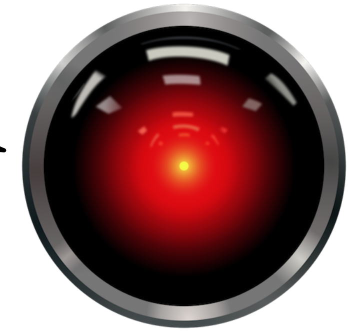


From which we can conclude:

1. Hollywood has no imagination
2. We anthropomorphize *everything*


We are at the start of the AI Era
We have created an Alien Intelligence
and it is not going away
How do we interact/collaborate with it?
(and achieve a symbiotic relationship)

# **Everything is going *extremely well*, George**



The goal is not to replace the humans but to ***amplify our capabilities,*** and it was always thus, from the opposable thumbs to grasp tools, to the modern day:

- ✧ Transportation (cars, airplanes, submarines, spacecraft… )
- ✧ Medicine: enhancing the immune system, replacing organs…
- ✧ Telecommunications over the large distances
- ✧ From print to Google: augmenting our memory
- ✧ Computing, cognition tech, neuro tech… **enhance our minds**

**We create technology, and the technology changes us**

And so it will be with the machine intelligence

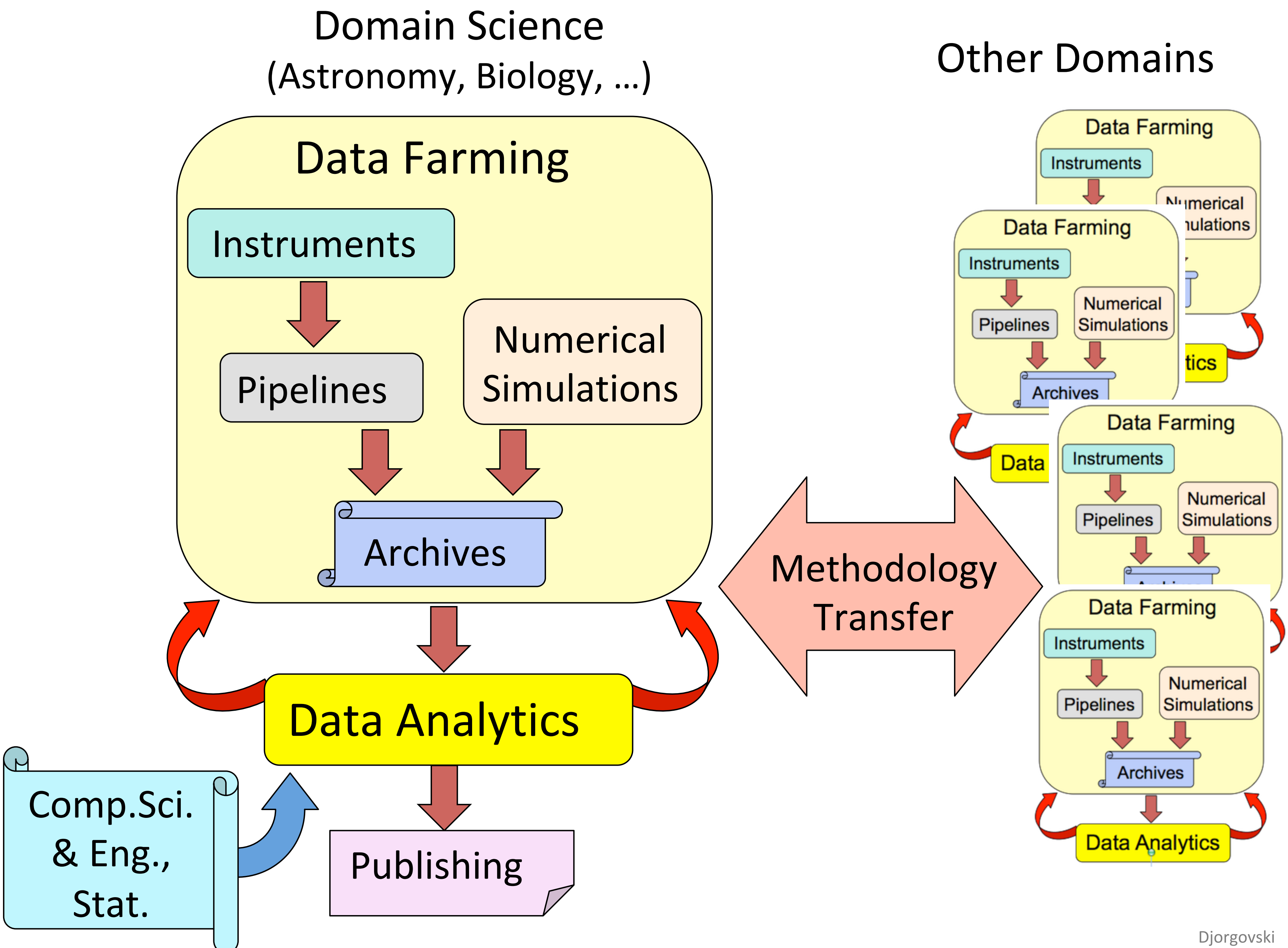
# The Uses of Machine Intelligence: Science on the Carbon-Silicon Interface

- **Data processing:**
  - Automated object / event classification, pattern recognition
  - Automated data quality control (anomaly/fault detection and repair)
- **Data mining, analysis, and understanding:**
  - Clustering, classification, outlier / anomaly detection
  - Pattern recognition, hidden correlation search
  - Assisted dimensionality reduction for visualization
  - Workflow control in Grid- or Cloud-based apps
- **Data farming and data discovery:** semantic web, etc.
- **Code design and implementation:** from art to science?

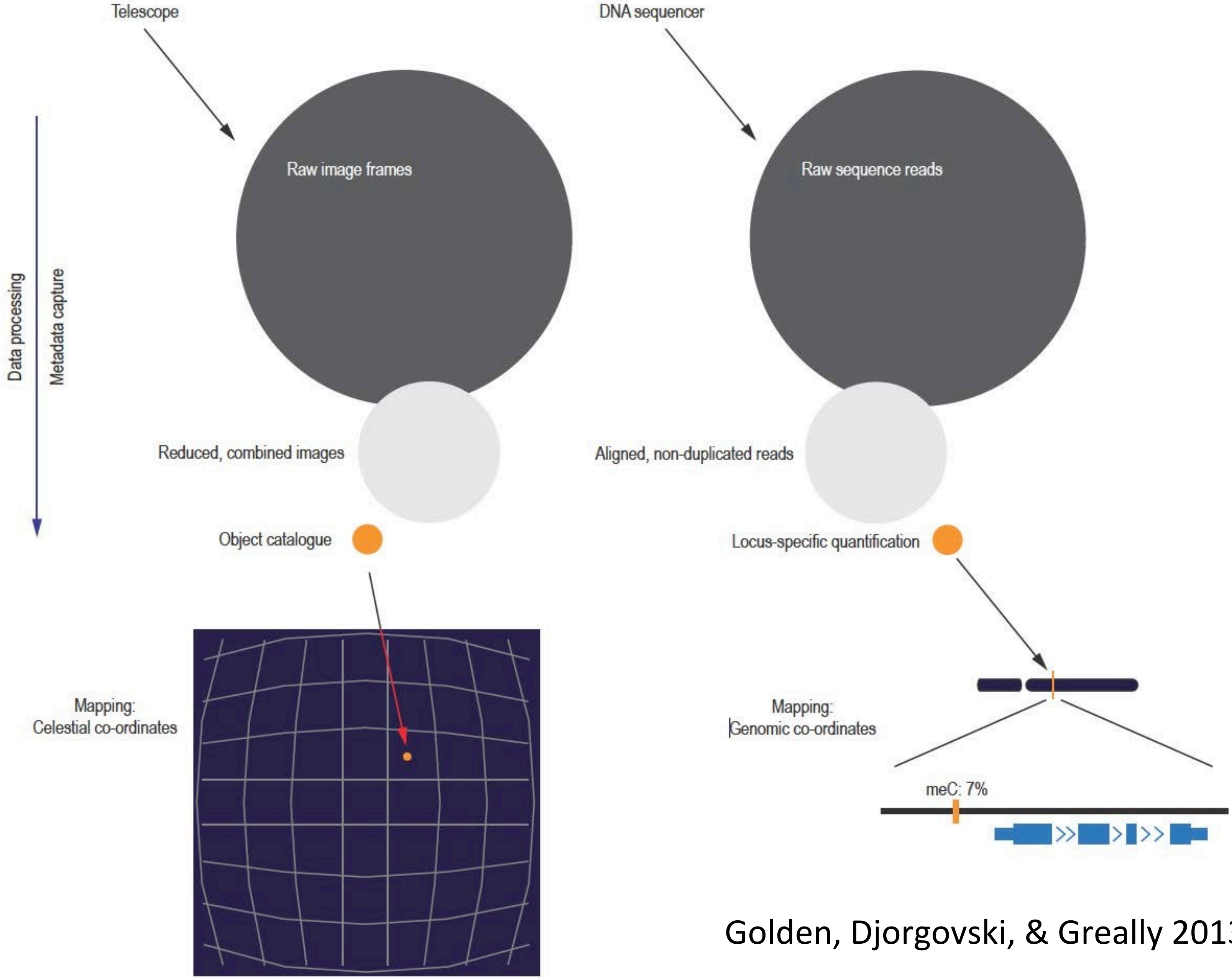
# Data Science Methodology Transfer

There are common challenges and a common underlying methodology to much of the data science (computing, IT, ML, statistics…)

How can we transfer the cyberinfrastructure developments, experience, and solutions from one scientific domain to others?


Domain Science
(Astronomy, Biology, …)
Other Domains
Data Farming
Instruments
Pipelines
Numerical Simulations
Archives
Data Analytics
Publishing
Comp.Sci. & Eng., Stat.
Methodology Transfer
Data Farming
Instruments
Pipelines
Numerical Simulations
Archives
Data Analytics
Djorgovski

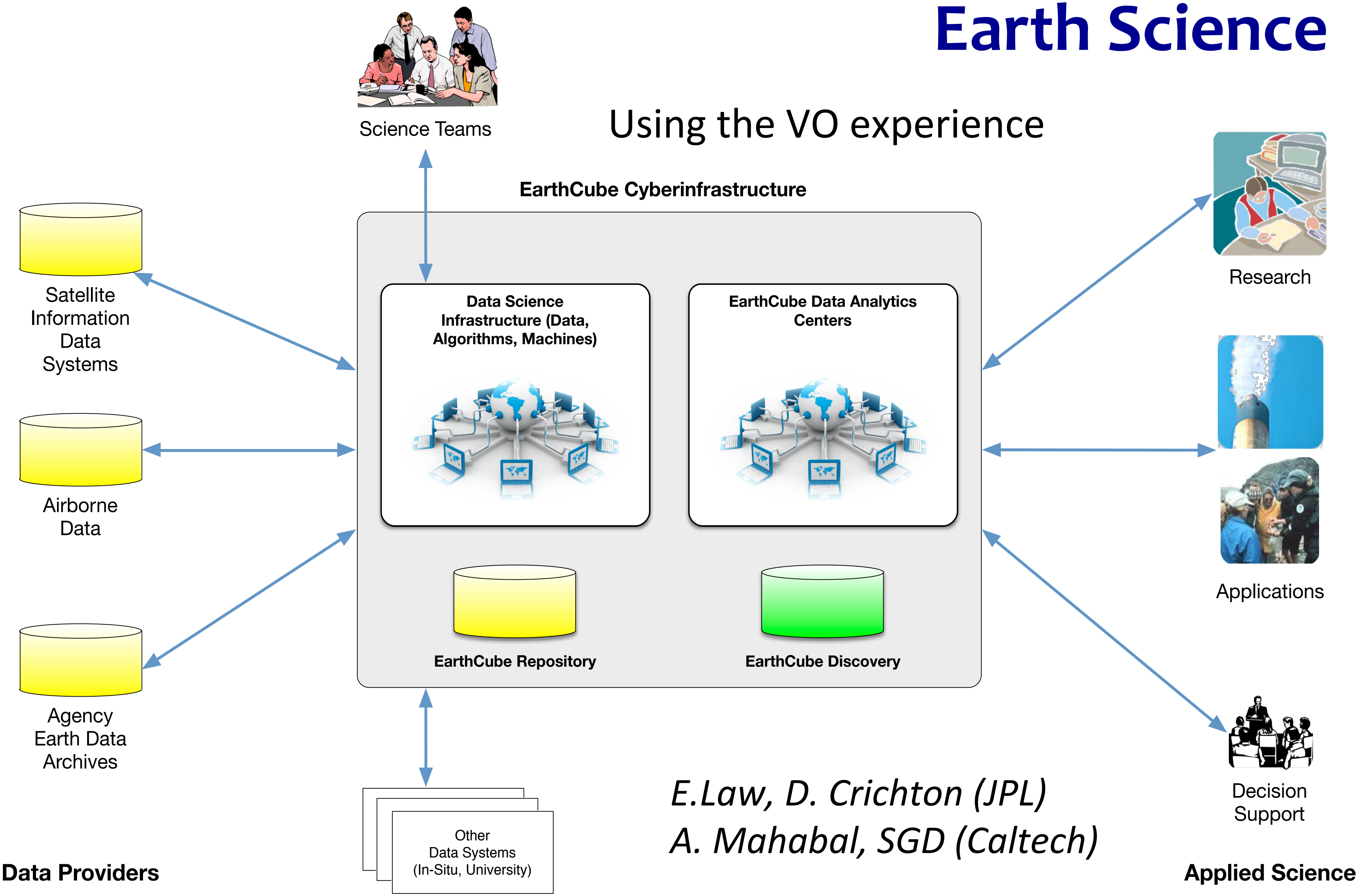
# AstroGenomics?




Golden, Djorgovski, & Greally 2013

# EarthCube: Software Architecture for Earth Science

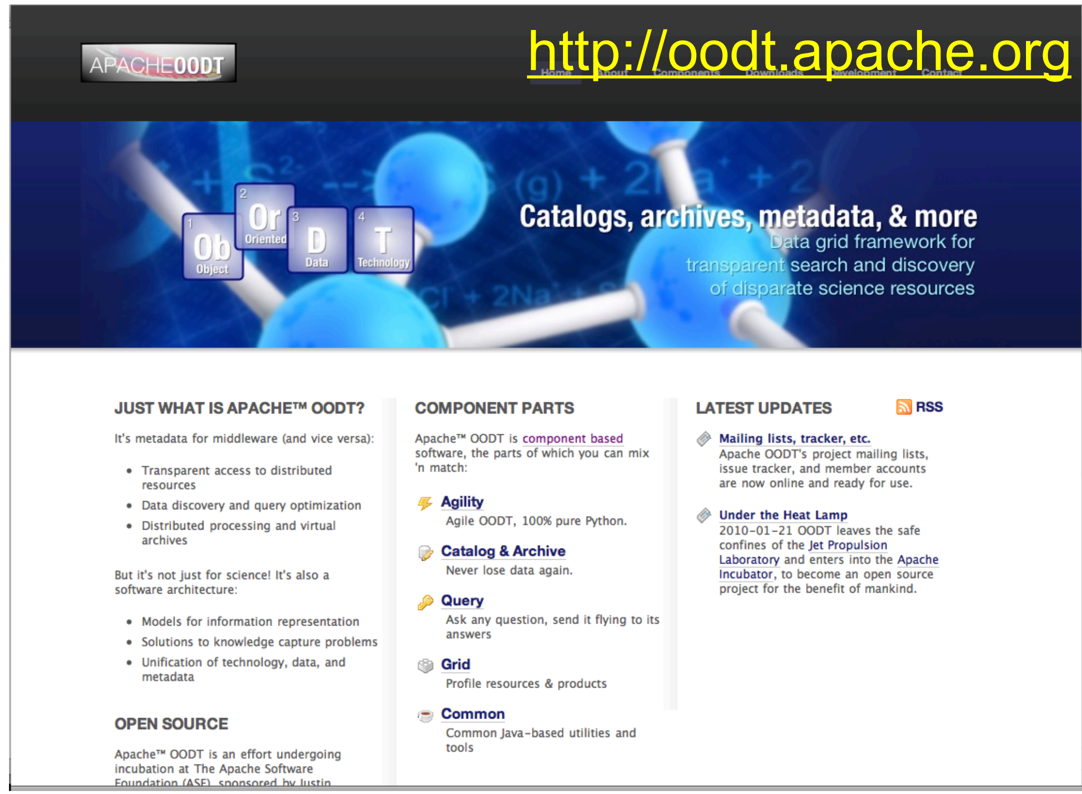
Using the VO experience




*E.Law, D. Crichton (JPL)*
*A. Mahabal, SGD (Caltech)*

# OODT: An Apache Open Source Framework for Building Distributed Data Intensive Systems

- An architectural style and framework for capture and sharing of distributed repositories
- Funded by NASA in 1998
- Applications to:
  - Planetary Science (1999)
  - Interferometry (1999)
  - Cancer Research (2001)
  - Earth Science (2002)
  - Medicine (2003)
  - Climate Research (2008)
  - Radio Astronomy (2010)
  - DARPA (2012)

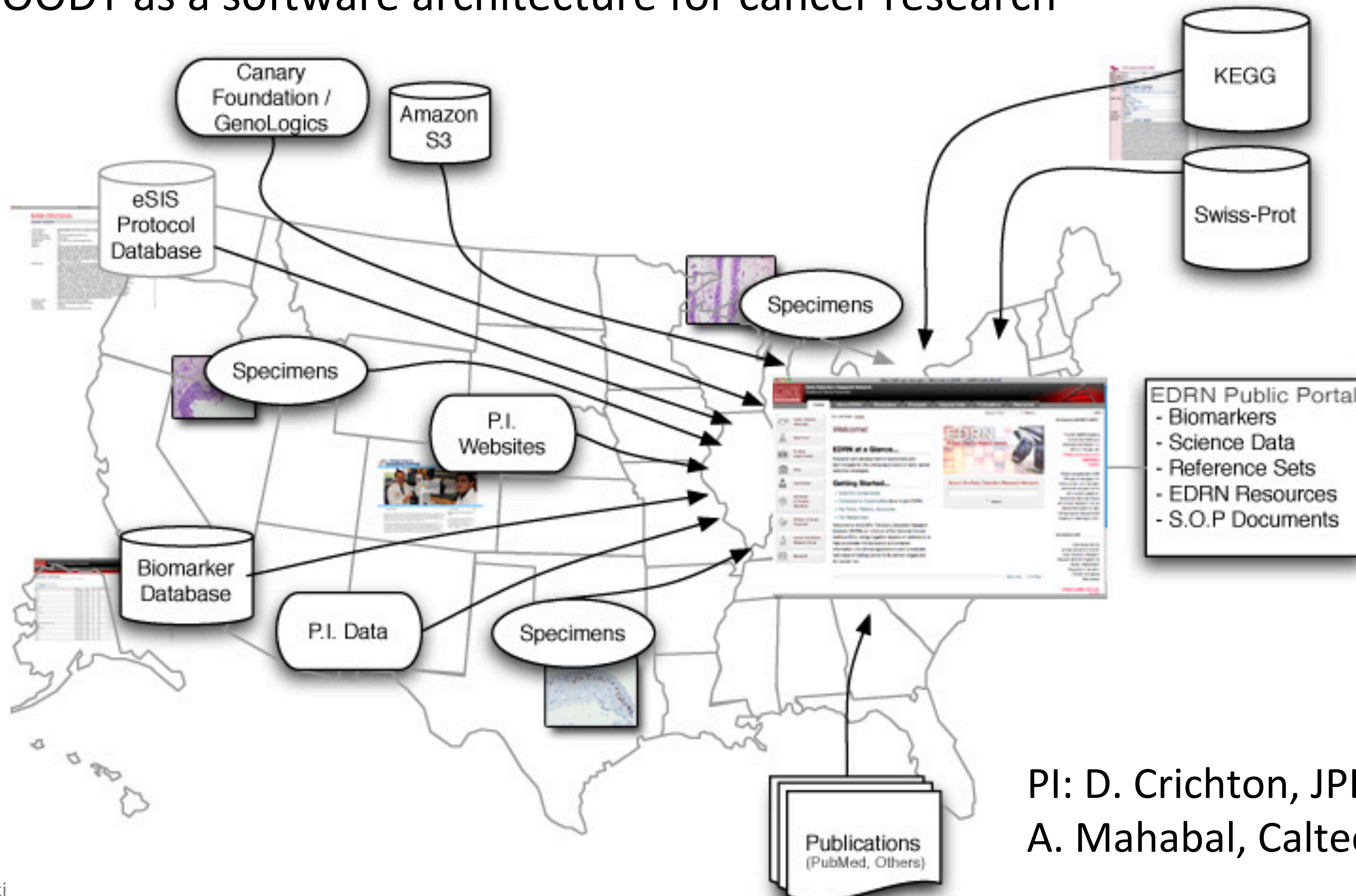



- Runner-up NASA Software of the Year, 2003
  - ✧ First NASA ASF open source project
- Top level project at Apache Software Foundation (2011)

*(PI: D. Crichton, JPL)*

# EDRN: A Virtual, National Integration Cancer Biomarkers Knowledge System

OODT as a software architecture for cancer research



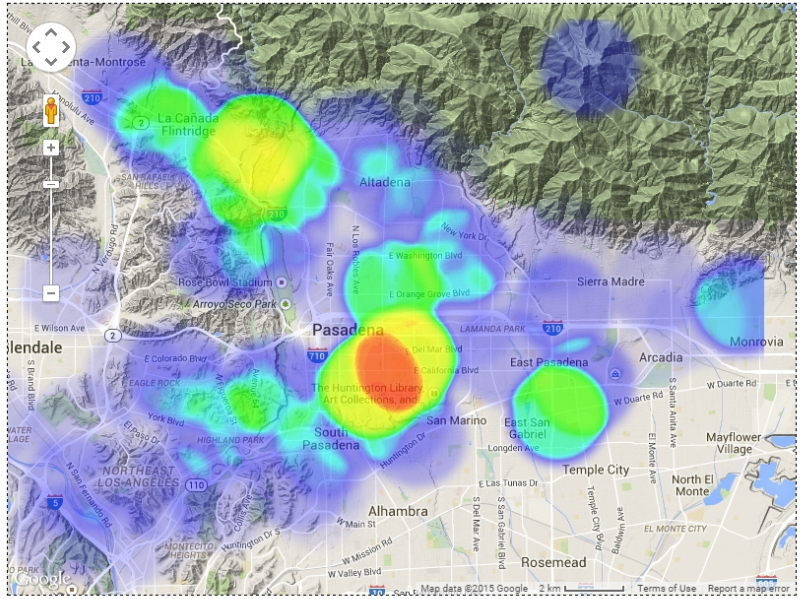

PI: D. Crichton, JPL
A. Mahabal, Caltech

# Real Time Classification and Response

Seismology:
Cell phones as a
sensor network

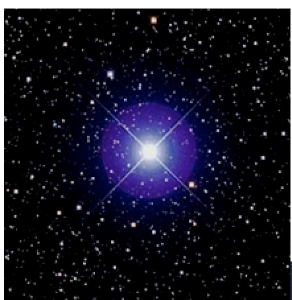
Time domain
astronomy

Lake Castaic M4.2 Jan 4 2015 Heatmap

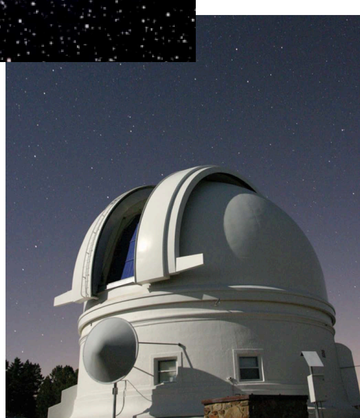




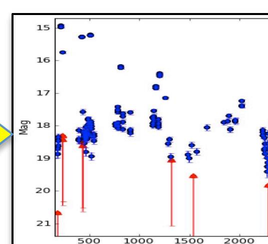
Event



Detection



Classification

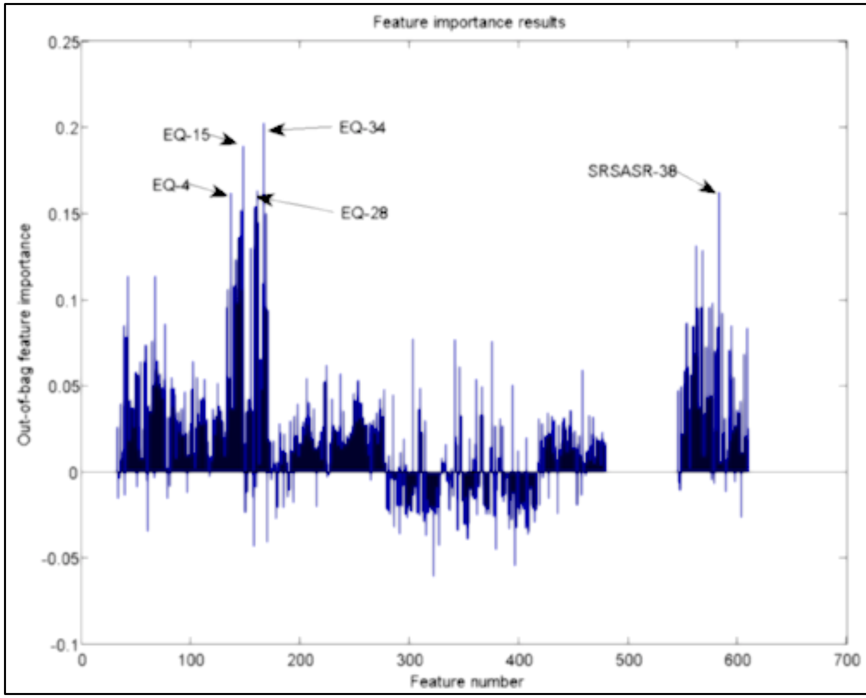


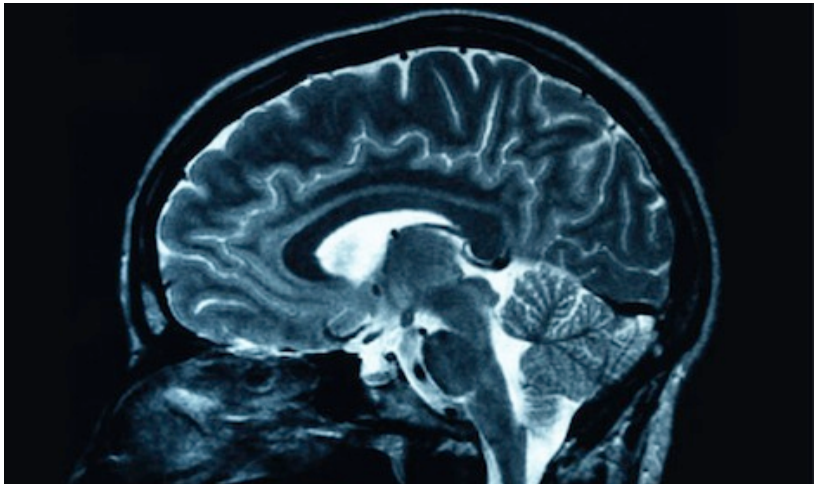

Decision
making

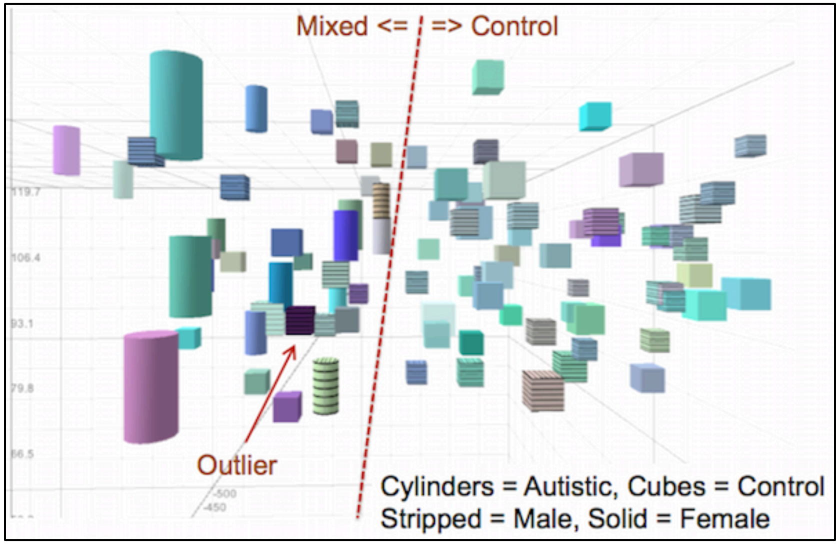
Follow-up

# From Sky Surveys to Neurobiology

- Using the data analytics tools based on Machine Learning, developed for the analysis of sky surveys, to design a better diagnostics for autism
- Next: analysis of brain MRI data






Djorgovski

# The Fourth Paradigm Redux

- The information content of modern data sets is so high as to enable profitable data mining

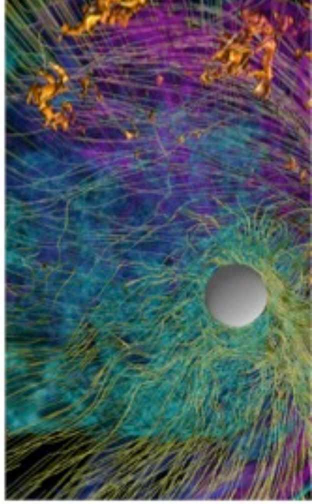
- Data fusion reveals new knowledge which was not recognizable in the individual data sets

- Data complexity requires machine intelligence to assist a human comprehension and understanding

**The Fourth Paradigm =**
**Data Fusion + Data Mining + Machine Learning**

# Some Thoughts About Data Science

- Comput***ational*** science $\neq$ Comput***er*** science
- Data-driven science is *not* about data, it is about ***knowledge extraction*** (the data are incidental to our real mission)
- Information and data are (relatively) cheap, but the expertise is expensive
  - Just like the hardware/software situation
- Data science as the "new mathematics"
  - It plays the role in relation to other sciences which mathematics did in ~ 17th - 20th century
- Computation: an interdisciplinary glue/lubricant
  - Many important problems (e.g., climate change) are inherently inter/multi-disciplinary



Djorgovski

# The Key Points

- **Cyberspace** is the new arena where humans interact with each other, and with the world of information
- **Science** in the 21$^{st}$ century is increasingly data-rich and computationally enabled, driven by the evolution of technology; thus, **the scientific method evolves**
  - New fields (X-Informatics), new (and perishable) types of scientific institutions, new publishing modalities…
  - Astronomy success(?) story: VO, Astroinformatics
  - *It is not all about data; the real focus is on the shared **knowledge discovery methodologies***
  - Important well beyond science: enabling new science-technology-commerce **synergies**

**“May all of your problems be technological”**

*Jim Gray*

**“If you don’t like change, you’re going to like irrelevance even less”**

*General Eric Shinseki*

**“Science progresses through funerals”**

*Max Planck*

**“If everything is under control, you are just not driving fast enough!”**

*Stirling Moss, Formula 1 driver*