Cyberspace, the Endless Frontier*

*With the apologies to Vannevar Bush and Mr. Spock

S. George Djorgovski
Astronomy Dept. and Center for Data-Driven Discovery, Caltech

PSIDA meeting, St. Louis, April 2018
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 21\textsuperscript{st} 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 \(\rightarrow\) new knowledge
  – These challenges are universal

• A great synergy of the computationally enabled science, and the science-driven IT
Exponential Growth of Data Volumes

... and Complexity on Moore’s law time scales

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

Understanding of complex phenomena requires complex data!
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
Hypothesis-driven science

Hypothesis/theory

Experiment

Data analysis

Understanding

Data-driven science

Data sets and streams

Data exploration, Pattern discovery

Hypothesis/theory

Data analysis

Understanding

The two approaches are complementary
A Modern Scientific Discovery Process

Data Gathering (finstruments, sensor networks, their pipelines...)

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

Key Technical Challenges

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

Data Understanding

New Knowledge

+feedback
Astronomy Has Become Very Data-Rich

• Typical digital sky surveys generate ~ 100 TB each, plus a comparable amount of derived data products
  – PB-scale data sets are imminent
• 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 (?)
... 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
There Are Lots Of Stars In The Sky...

Modern sky surveys obtain $\sim 10^{15} - 10^{16}$ bytes of images, catalog $\sim 10^8 - 10^9$ objects (stars, galaxies, etc.), and measure $\sim 10^2 - 10^3$ numbers for each.
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...
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

$\text{t}_2 \sim 1.5 \text{ yrs}$

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

MB GB TB PB EB

CCDs Surveys VO AstroInfo
Image Proc. Pipelines Databases

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

Djorgovski
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

- Integrated SZ
- Grav. Lensing
- Sachs-Wolfe
- CMB Signal
- Galactic Thermal
- Gal. Nonthermal
- Radio Sources
- Galaxies (SF)
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
## 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.

<table>
<thead>
<tr>
<th>Applications (in alphabetical order)</th>
<th>Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aladin</td>
<td>Search for Images: Aladin, Datascope, SkyView, VODESKTOP, Data Discovery Tool</td>
</tr>
<tr>
<td>AppLauncher</td>
<td>Search for Spectra: Aladin, CASSIS, Datascope, SPLAT, Specview, VOServices, VOSpec, Data Discovery Tool</td>
</tr>
<tr>
<td>CASSIS</td>
<td>Search for Catalogues: Aladin, Datascope, TOPCAT, VODESKTOP, Data Discovery Tool</td>
</tr>
<tr>
<td>CDS Xmatch Service</td>
<td>Search for Time Series</td>
</tr>
<tr>
<td>Data Discovery Tool</td>
<td></td>
</tr>
<tr>
<td>Filter Profile Service</td>
<td></td>
</tr>
<tr>
<td>Iris</td>
<td></td>
</tr>
<tr>
<td>Montage</td>
<td></td>
</tr>
<tr>
<td>Octet</td>
<td></td>
</tr>
<tr>
<td>SkyView</td>
<td></td>
</tr>
<tr>
<td>Specview</td>
<td></td>
</tr>
<tr>
<td>SPLAT</td>
<td></td>
</tr>
<tr>
<td>TAPHandle</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VO-compliant Tools &amp; Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS9: Image visualisation</td>
</tr>
<tr>
<td>GOSSIP: SED fitting</td>
</tr>
<tr>
<td>VirGO: Search for Images and Spectra</td>
</tr>
<tr>
<td>IRAF: Image Reduction &amp; Analysis</td>
</tr>
<tr>
<td>World Wide Telescope</td>
</tr>
<tr>
<td>Gaia - Graphical</td>
</tr>
<tr>
<td>Astronomy and Image Analysis</td>
</tr>
<tr>
<td>SIMBAD</td>
</tr>
<tr>
<td>TESELA</td>
</tr>
<tr>
<td>VizieR</td>
</tr>
</tbody>
</table>
What has the IVOA achieved?
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
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
About the Cost Efficiency...

- F-35: $122M
- USS Zumwalt: $4.5B
- M-1: $9M
- Delta IV Heavy: $350M
- Cassini: $3.9B

-$1M? - Data Science?
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 \(\Rightarrow\) Need agency support, mandates
     2. Results \(\Rightarrow\) *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*
Beyond VO: AstroInformatics

- VO became a global data grid of astronomy, but it produced very few **knowledge discovery tools**
- AstroInformatics is a **mechanism for a broader community inclusion** (both as contributors and as consumers)
- It is a bridge between astronomy and computer science, information technology, statistics, etc. (“data science”), and for the methodology sharing with other fields.
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
Pattern or structure (Correlations, Clustering, Outliers, etc.) Discovery in High-Dimensional Parameter Spaces

D >> 3 parameter space hypercube

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

But in some corner of some sub-D projection of this data space, there is something ≠ noise

Djorgovski
Machine Discovery of Analytical Expressions ("Laws of Nature")

- Symbolic regression finds best-fitting mathematical description of a sample of data via evolutionary algorithm
- Cast binary classification as:
  \[ \text{class} = g(f(x_1, x_2, x_3, \ldots, x_n)) \]
- \( f(x) \) is equation of discriminating hyperplane
- Tests in astronomy (Graham et al. 2015):

Classification of variable stars ->

\[ (2.2 - \left( \frac{X}{11} \right)) + (7 \times \cos(Y)) \]

<- Fundamental Plane discovery
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.
Science Originates on Interfaces

... between human minds, data, and other informational constructs
The Revolution in Scholarly Publishing

- Increasing complexity and diversity of scientific data and results
  - Data, archives, metadata, virtual data, simulations, algorithms, blogs, wikis, multimedia...
  - *From static to dynamic*: evolving and growing data sets
  - *From print-oriented to web-oriented*

- Institutional, cultural, and technical challenges:
  - Curation by domain experts
  - Effective peer review and quality control
  - Persistency and integrity of data and pointers
  - Interoperability and metadata standards
  - *From the ownership of storage media to the ownership of access to the bits*

*As the science evolves, so does its publishing*
Effective visualization is the bridge between quantitative information and human intuition.

Man cannot understand without images
Aristotle, De Memoria et Reminiscentia

You can observe a lot just by watching
Yogi Berra, an American philosopher
A Key Challenge: Visualizing Complexity

• Hyperdimensional structures (clusters, correlations, etc.) are likely present in many complex data sets, whose dimensionality is commonly in the range of $D \sim 10^2$–$10^4$, and will surely grow.

• It is not only the matter of data understanding, but also of choosing the appropriate data mining algorithms, and interpreting the results.

• We are biologically limited to perceiving $\sim 3$–$12$ (?) dimensions.

What good are the data if we cannot effectively extract knowledge from them?

“A man has got to know his limitations”

Dirty Harry, another American philosopher.
Traditional Data Visualization

An example from astronomy: a subset of data on quasar properties, from a 6-dimensional data space.

These are 2 out of the 15 possible 2-D plots, but even then relationships involving >2 variables are lost.
Diving Into the 6-Dimensional Data Space in Virtual Reality
Users interacting with the data, machine intelligence, and each other in a shared virtual space.
Keck Institute for Space Studies Symposium on Virtual and Augmented Reality for Space Science and Exploration

Caltech, Jan. 30, 2018

Co-Organizers: E. Law, S. Davidoff, SGD

Videos: www.kiss.caltech.edu/symposia/space_science
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.
From the Information Technology to the Cognition Technology: Towards a Human-Computer Collaborative Discovery

Vannevar Bush (1945)

Man-Computer Symbiosis

J.C.R. Licklider (1960)
The Rise of the Machines

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 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?
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?
AstroGenomics?

Golden, Djorgovski, & Greally 2013
EarthCube: Software Architecture for Earth Science

Using the VO experience

EarthCube Cyberinfrastructure

Data Science Infrastructure (Data, Algorithms, Machines)

EarthCube Data Analytics Centers

EarthCube Repository

EarthCube Discovery

Satellite Information Data Systems

Airborne Data

Agency Earth Data Archives

Other Data Systems (In-Situ, University)

Data Providers

Research

Applications

Decision Support

Applied Science

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:
  - Interferometry (1999)
  - Cancer Research (2001)
  - 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)
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

Event

Detection

Classification

Decision making

Follow-up

Lake Castaic M4.2 Jan 4 2015 Heatmap

Communications of the ACM

Your Phone as Quake Detector

Djorgovski
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
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

- Computational science ≠ Computer 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
Science in Cyberspace

Theory and Simulations

Visual Displays and Linking of Data and Knowledge

Published Literature

Data Archives

Semantic Web

Virtual Observatory

Djorgovski
The Key Points

- **Cyberspace** is the new arena where humans interact with each other, and with the world of information.

- **Science** in the 21st 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