Astronomy in the Era of Big Data: From Virtual Observatory to Astroinformatics and beyond

Prof. S. George Djorgovski

Astronomy Department and Center for Data-Driven Discovery, Caltech

TIARA Summer School on Astrostatistics and Big Data
Taipei, Taiwan, September 2017
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
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!

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

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
Supercomputers vs. Server Farms (Cloud)

They support different kinds of computing
Hypothesis-driven science

Hypothesis/theory

→ Experiment

→ Data analysis

→ Understanding

Data-driven science

Data sets

Data streams

→ Data exploration, Pattern discovery

→ Hypothesis/theory

→ 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

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

**Data Understanding**

Key Technical Challenges

+feedback

New Knowledge

Djorgovski
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)
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.

A small portion of the Galactic bulge
The Panchromatic Universe

Near IR starlight

Far IR warm dust

Hα ionized gas

X-Ray accretion
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

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 Machine Learning AI

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 Became Real

http://us-vo.org

http://ivoa.net

http://www.euro-vo.org
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>
<th>VO-compliant Tools &amp; Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aladin</td>
<td>Search for Images: Aladin, DataScope, SkyView, VODesktop, Data Discovery Tool</td>
<td>DS9: Image visualisation</td>
</tr>
<tr>
<td>AppLauncher</td>
<td>Search for Spectra: Aladin, CASSIS, DataScope, SPLAT, Specview, VOServices, VOSpec, Data Discovery Tool</td>
<td>GOSSIP: SED fitting</td>
</tr>
<tr>
<td>CASSIS</td>
<td>Search for Catalogues: Aladin, DataScope, TOPCAT, VODesktop, Data Discovery Tool</td>
<td>VirGO: Search for Images and Spectra</td>
</tr>
<tr>
<td>CDS Xmatch Service</td>
<td>Search for Time Series</td>
<td>IRAF: Image Reduction &amp; Analysis</td>
</tr>
<tr>
<td>Data Discovery Tool</td>
<td></td>
<td>World Wide Telescope</td>
</tr>
<tr>
<td>Filter Profile Service</td>
<td></td>
<td>Gaia - Graphical</td>
</tr>
<tr>
<td>Iris</td>
<td></td>
<td>Astronomy and Image Analysis</td>
</tr>
<tr>
<td>Montage</td>
<td></td>
<td>SIMBAD</td>
</tr>
<tr>
<td>Octet</td>
<td></td>
<td>TESELA</td>
</tr>
<tr>
<td>SkyView</td>
<td></td>
<td>VizieR</td>
</tr>
<tr>
<td>Specview</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPLAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAPHandle</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A compilation of tools and services

IVOA is now mainly a standards coordination body
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 Bgd. and its Foregrounds

- Integrated SZ
- Grav. Lensing
- Sachs-Wolfe
- CMB Signal
- Galactic Thermal
- Gal. Nonthermal
- Radio Sources
- Galaxies (SF)
The Web has a truly transformative potential for education

- 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
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
Virtual Observatory Eulogy

*(de mortibus nil nisi bonum)*

The Good:

- Progress on interoperability, standards, etc.
- A global *data grid of astronomy*
- Empowering a broad community
- Some useful web services
- Community training, outreach
- Better than most other fields (yes!)

The Not So Good:

- Data exploration and mining tools
  
  That is where the science comes from!
  
  Thus, little VO-enabled science
  
  Thus, a slow community buy-in

Better: **AstroInformatics!**
From Virtual Observatory to Astroinformatics

• A bridge field connecting astronomy with computer science and engineering, statistics

• A mechanism for a broader community inclusion, both as contributors and as consumers

• A mechanism for an interdisciplinary data science methodological sharing with other fields
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 dwarfs
Systematic Exploration of the Observable Parameter Space (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

SDSS

Dimensionality $\leq$ the number of observed quantities

Physical Parameter Space

Fundamental Plane of hot stellar systems

E

dSp

GC

Both are populated by objects or events
• Not filled uniformly: clustering indicates different families
• Clustering + dimensionality reduction = correlations
• High dimensionality poses analysis challenges
Exploration of Parameter Spaces is a Central Problem

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
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.
A Key Challenge: Visualising Multidimensional Data Spaces

• Hyperdimensional structures (clusters, correlations, etc.) may be present in many complex data sets, whose dimensionality may be $D \sim 10^2 - 10^4$, or higher

• It is a matter of *data understanding*, choosing the right data mining algorithms, and interpreting the results

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

What good are the data if we cannot effectively extract knowledge from them?
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.
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?
From the Information Technology to the Cognition Technology: Towards a Human-Computer Collaborative Discovery

**AS WE MAY THINK**
A TOP U.S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLD IN WHICH MAN-MADE MACHINES WILL START TO THINK
by VANNEVAR BUSH

Vannevar Bush (1945)

**Man-Computer Symbiosis**
J.C.R. Licklider (1960)
Machine Discovery Using Eureqa
Lipman et al., http://creativemachines.cornell.edu/eureqa

- Employs symbolic regression to determine best-fitting functional form to data and its parameters simultaneously
- Specify the building blocks to be used: algebraic operators, analytical functions, constants
- See Graham et al. (2013)
Human-annotated images (crowdsourced?)

- Semantic descriptors
  - Machine processing
  - Evolving novel algorithms

Challenges: Optimizing for different levels of user expertise; optimal input averaging; encoding contextual information; etc.

(Lead: M. Graham)
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?
Center for Data-Driven Discovery

- A new research center at Caltech
  - Serves the research efforts Institute-wide
- A part of the Caltech-JPL joint initiative for data science and technology
- The goals are to assist faculty in formulation and execution of data-intensive projects, and facilitate interdisciplinary sharing of methods, ideas, novel projects, etc.
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:
  - 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 ACM

Your Phone as Quake Detector
From Sky Surveys to Neurobiology

- Using the data analytics tools based on ML, developed for the analysis of sky surveys, to design a better diagnostics for autism
- Feature importance using random forests =>
  - Next: correlate with MRI scans

(with R. Adolphs et al.)
From Sky Surveys to Neurobiology

• 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

• Dependent features:
  – “I find it easy to put myself in somebody else’s shoes”
  – “I can tell if someone is masking their true emotion”
  – “I feel at upset”

• Accuracies of ~90% but small sample data set and feature degeneracy

M. Graham, CD³
From Sky Surveys to Neurobiology

- Feature importance => 6-dimensional parameter space

Mixed <= => Control

Cylinders = Autistic, Cubes = Control
Stripped = Male, Solid = Female
Innovative Data Visualization

Using the emerging technologies of virtual reality, commodity hardware and software, for an immersive, interactive, collaborative visual data analytics and exploration

C. Donalek, SGD (CD³)
S. Davidoff (JPL)
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
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*
Science Originates on Interfaces

… between human minds, data, and other informational constructs.
Technology changes how we communicate and convey information.

What comes after the Web and the Internet?

Increasing immediacy, increasing fidelity.
Science in Cyberspace

Theory and Simulations

Visual Displays and Linking of Data and Knowledge

Published Literature

Data Archives

Semantic Web

Virtual Observatory

Djorgovski
Training the Next Generation

• We are developing a new curriculum about the tools and methods for computational, data-intensive research in the 21st century

• An example: **the first virtual summer school on big data analytics**, a joint venture with JPL:

![JPL-Caltech Virtual Summer School](image)

Big Data Analytics

September 2 – 12, 2014

• Making use of on-line educational technologies (e.g., MOOCs) to maximize our impact:

https://www.coursera.org/learn/bigdataanalytics/

(re-opening soon!)
The Key Points

• Science is being transformed by the computing technology and the resulting exponential growth of data
  – Data-driven science is quantitatively and qualitatively different from the traditional model

• Many challenges are common; their solutions constitute a rise of the new scientific methodology
  … and methodology transfer can and should be done

• Machine learning and computational statistics tools are essential, and many challenges remain
  – Uses of machine intelligence will lead to a collaborative human-computer discovery, and a cognition technology

• Multidimensional data visualization is a key bottleneck
  – Virtual reality will be a powerful platform for both data visualization and scientific collaboration
“If everything is under control, you are just not driving fast enough!”

Stirling Moss, Formula 1 driver

“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