Science Drivers for Big Data
Joseph Lazio
SKA Program Development Office &
Jet Propulsion Laboratory, California
Institute of Technology

© 2010 California Institute of Technology.
Government sponsorship acknowledged.

Data Intensive Astronomy

• “There is nothing new under the Sun ....”
• Case studies
  1. Cosmology and imaging surveys
  2. Fundamental physics from pulsar observations and pulsar surveys
  3. Identifying interference

Exploring the Universe with the world’s largest radio telescope
Data-Intensive Astronomy

Data Volumes

ιππαρχος (Hipparchus)
• ca. 135 BCE
• 850 entry stellar catalog

Computational Limitations

Harvard computers
• Production of stellar plates and spectra (“data rate”) was increasing enormously
• Examined and classified telescope output
• Forerunners of human mathematical computers

Expanding the Universe with the world’s largest radio telescope

Key Science for the SKA
(a.k.a. m- and cm-λ astronomy)

Emerging from the Dark Ages & the Epoch of Reionization

Strong-field Tests of Gravity with Pulsars and Black Holes

Galaxy Evolution, Cosmology, & Dark Energy

The Cradle of Life & Astrobiology

Origin & Evolution of Cosmic Magnetism

Expanding the Universe with the world’s largest radio telescope
21st Century Astrophysics

Cosmology & Fundamental Physics
- Gravity
  - Can we observe strong gravity in action?
  - What is dark matter and dark energy? (dark energy and BAOs with H I galaxies)
- Magnetism
- Strong force
  - Nuclear equation of state

Galaxies Across Cosmic Time, The Galactic Neighborhood, Stellar and Planetary Formation
- Galaxies and the Universe
  - How did the Universe emerge from its Dark Ages?
  - How did the structure of the cosmic web evolve?
  - Where are most of the metals throughout cosmic time?
  - How were galaxies assembled?
- Stars, Planets, and Life
  - How do planetary systems form and evolve?
  - What is the life-cycle of the interstellar medium and stars? (biomolecules)
  - Is there evidence for life on exoplanets? (SETI)

Cosmology

Origin and Fate of the Universe
- Era of “precision cosmology” … or precision ignorance
- Need to sample a substantial volume of the Universe

Composition of the Universe
Cosmology and Sky Surveys

Volume \sim D^2 \Delta D \Omega

- D – distance; \Omega – solid angle
- Surveying to larger D is difficult \implies need larger telescopes
  “square kilometre” of SKA
- Surveying larger sky areas \Omega just requires more observing time

• Image the sky, locating galaxies
  Analysis of locations compared with cosmological models to constrain parameters

• Two broad classes of surveys
  – Continuum: e.g., NVSS, FIRST, ASKAP/EMU, WSRT/APERTIF/WODAN
  – Spectroscopic: SDSS, Arecibo ALFALFA, ASKAP/WALLABY, SKA H i survey
  Spectroscopic surveys locate in \textbf{3-D space}! very powerful

• Ultimate goal: spectroscopic survey of 1 billion galaxies
**Imaging with Arrays**

Fourier transform plane

- $N_{\text{data}} \sim N_{\text{antenna}}^2 N_{\text{frequency}} N_{\text{time}}$

Image plane

- Fourier Transform $\leftrightarrow$

**Imaging Surveys**

Requirements
- Many antennas in order to provide sensitivity and image quality
  - $N_{\text{antenna}}$
- Spectral resolution because of wide-field effects, line emission from galaxies, or both
  - $N_{\text{frequency}}$
- Long integrations in order to obtain adequate signal-to-noise ratio
  - $N_{\text{time}}$ e.g., 500 hr at 1 s sampling?

- $N_{\text{data}} \sim N_{\text{antenna}}^2 N_{\text{frequency}} N_{\text{beams}} N_{\text{time}}$

<table>
<thead>
<tr>
<th></th>
<th>ASKAP</th>
<th>SKA Phase 1</th>
<th>SKA Phase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{\text{antenna}}$</td>
<td>30</td>
<td>250</td>
<td>1000</td>
</tr>
<tr>
<td>$N_{\text{beams}}$</td>
<td>30</td>
<td>1</td>
<td>1?</td>
</tr>
<tr>
<td>$N_{\text{frequency}}$</td>
<td>16k</td>
<td>16k?</td>
<td>16k?</td>
</tr>
</tbody>
</table>
Imaging Surveys II

<table>
<thead>
<tr>
<th>ASKAP</th>
<th>SKA Phase 1</th>
<th>SKA Phase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N_{\text{antenna}} = 30)</td>
<td>(N_{\text{antenna}} \sim 250)</td>
<td>(N_{\text{antenna}} \sim 1000)</td>
</tr>
<tr>
<td>(N_{\text{beams}} = 30)</td>
<td>(N_{\text{beams}} = 1)</td>
<td>(N_{\text{beams}} = 1?)</td>
</tr>
<tr>
<td>(N_{\text{frequency}} \sim 16k)</td>
<td>(N_{\text{frequency}} \sim 16k?)</td>
<td>(N_{\text{frequency}} \sim 16k?)</td>
</tr>
<tr>
<td>(N_{\text{data}} \sim 8 \times 10^{14})</td>
<td>(N_{\text{data}} \sim 2 \times 10^{15})</td>
<td>(N_{\text{data}} \sim 3 \times 10^{16})</td>
</tr>
<tr>
<td>(N_{\text{OPS}} \sim 8 \times 10^{15})</td>
<td>(N_{\text{OPS}} \sim 2 \times 10^{15})</td>
<td>(N_{\text{OPS}} \sim 3 \times 10^{20})</td>
</tr>
</tbody>
</table>

- Imaging is more than “just” an FFT. Gridding, deconvolution, wide-field corrections, self-calibration, …
- Community estimates are \(10^4\) to \(10^5\) ops per visibility datum(!)

Exploring the Universe with the world’s largest radio telescope

Fundamental Physics with Radio Pulsars

Arrival times of pulses from radio pulsars can be measured with phenomenal accuracy
- Better than 100 ns precision in best cases
- Enables high precision tests of fundamental physics
  - Theories of gravity, gravitational waves, nuclear equation of state
  - 1993 Nobel Prize in Physics
- Problem: Not all pulsars are equal!
  - Good “timers” < 10% of total population
- Need to find many more!
  - All-sky survey

Ultra-relativistic binaries & gravitational wave studies
Pulsar Surveys I

Requirements

• Large bandwidths because pulsars are faint
• Long integration times because pulsars are faint
• Rapid time sampling in order to resolve pulse profile
• Narrow frequency channelization in order to mitigate interstellar scattering

• For a “pixel” on the sky, accumulate data for a time $\Delta t$ over a bandwidth $\Delta \nu$
  
  Suppose $\Delta t = 20$ min., $\Delta \nu = 800$ MHz
• Time sampling $\delta t$ with frequency channelization $\delta \nu$
  
  For GBT GUPPI, $\delta t = 81.92$ $\mu$s, $\delta \nu = 24$ kHz
  
  60 GB data sets per pixel …

Pulsar Surveys II

For GBT

• At 800 MHz, “pixel” ~ 16' = 0.3°
• About 350 kpixels in the sky
• 20 PB data set

For SKA

• At 800 MHz, “pixel” = 1.2'
• About 76 Mpixels in the sky
• 4.6 EB data set
Interference Mitigation

• Most of the radio frequency (RF) spectrum is not reserved for use by radio astronomy
  In fact, very little is!
• Other passive users are fine
• Active users can be troublesome!
  – GPS, microwave ovens, cell phones, car ignitions, electric fences, …

Interference Mitigation II

• Radio flux density measured in Janskys
  1 Jy = 10^{-26} W/m^2/Hz
• 10 \mu Jy = EVLA, GBT, ASKAP, MeerKAT, … sensitivity
• 10 nJy = SKA aim
• 100 Jy ~ cell phone on Moon
Data Visualization

- Data acquired from an array is at least 4-D
  - Visibility (= antenna_i × antenna_j)
  - Frequency
  - Time
  - Polarization
  - (Beams? for a multi-beam system)
- Best results still obtained by hand

Summary – Data-Intensive Astronomy

- Exciting science!
- Leads to exciting data challenges
  - Data volume (Exabyte)
  - Processing requirements (Exo-flop)
  - Data visualization (HMI)
  - ...
- Answers on Thursday