Data Intensive Astronomy Primer

Andreas Wicenec
WHY?
INTRODUCTION

SPACE IS BIG. REALLY BIG. YOU JUST WON'T BELIEVE HOW VASTLY, HUGELY, MIND-BOGGLINGLY BIG IT IS. I MEAN,
Douglas Adams was absolutely correct.

Space is really big.

Douglas Adams
1952 - 2001
What’s out there?

Adapted from Quinn
How Big?

How Far?

How Many?

How Old?

Adapted from Quinn
Radius

<table>
<thead>
<tr>
<th></th>
<th>Planet</th>
<th>Star</th>
<th>Galaxy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 km</td>
<td>4.2 km</td>
<td>10^6 km</td>
<td>10^{17} km</td>
</tr>
<tr>
<td>10^3 km</td>
<td></td>
<td>1</td>
<td>10^{+11}</td>
</tr>
<tr>
<td>10^{-3}</td>
<td></td>
<td>1</td>
<td>10^{+11}</td>
</tr>
</tbody>
</table>

Distance

<table>
<thead>
<tr>
<th></th>
<th>Sun</th>
<th>Galactic Centre</th>
<th>Nearby Galaxy</th>
<th>Nearby Star</th>
<th>Planet</th>
<th>Star</th>
<th>Galaxy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^8 km</td>
<td>10^2</td>
<td>10^{+17} km</td>
<td>10^{+17} km</td>
<td>10^{+13} km</td>
<td>10^{+2}</td>
<td>10^{+7}</td>
<td>10^{+11}</td>
</tr>
<tr>
<td>10^{13} km</td>
<td>10^{+13} km</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adapted from Quinn
Are we in a new era of astronomy?

- Amount and complexity of data has to be seen relative to the effort it takes to gather and process it.
- Compared to previous times it is obvious that gathering data nowadays is almost trivial. Too easy??
- Hipparchus and Tycho Brahe spent significant periods of their entire life to observe some 1,000 stars. Their achievements and conclusions are phenomenal! Probably only a few of us today would be able to derive similar knowledge from such data.
Are we in a new era of astronomy?

The instruments and methods invented and used by ancient and medieval astronomers had always been unique and cutting-edge. Very often stunningly big and complex.

The size of the instruments was always just necessary but never sufficient to derive the knowledge. In addition it always required the brightest minds.
Ancient and Medieval Instruments

Chankillo Astronomical Complex
~300 BC

Al-Khujandi’s Mural Sextant
~1000 AD

Warren Field Calendar
ca. 8000 BC

Armillary Sphere
~1500 AD

Photo credits:
- Chankillo Astronomical Complex: David Edgar
- Al-Khujandi’s Mural Sextant: Alexis
- Warren Field Calendar: Leoboudv
- Armillary Sphere: Leoboudv
Eyes on the Sky

Doubling time ~ 20 years
Gathering information

- 1598 Nearby stars
- 1845 Ink sketch of nearby galaxy
- 1880 First photographs
- 1969 Invention of CCD
Numbers per night

Year

Doubling time < 1 year
The First Revolution: Telescopes and Photographic Plates

f=45m, Keplerian telescope
Johannes Hevelius, 1673

Houghton Library at Harvard University

Replica of Newton’s second telescope

Photo credit: Andrew Dunn

Herschel 49” Telescope
1789

Mt. Palomar’s 200-inch Hale Telescope, pointing to the zenith, as seen from the east side.
The Second Revolution: CCDs

ICRAR Starfox: Wide-field, low surface brightness imaging FLI camera: 8173x6132 pix

GAIA focal plane: 106 CCDs, 0.5 m²

E2V QE plot

Image credit: Astrium

Image credit: NAOA

Graphics credit: nexyad.net

Graphics credit: Olympus Microscopy Research Centre

Image credit: ICRAR

Image credit: NAOA
The Third Revolution: Multi-wavelength and Satellites

Multiwavelength Whirlpool Galaxy

- **COLD GAS**: Radio waves reveal regions of gas cool enough for CO₂ molecules to exist.
- **COOL STARS**: Infrared shows smaller cool red stars that make up most of the galaxy.
- **SOLAR STARS**: Optical light comes from stars around the size of the Sun.
- **HOT STARS**: Ultraviolet shows the larger hot blue stars that are less frequent in galaxies.
- **HOT GAS**: X-rays are emitted from the hottest regions of gas where atoms are ionized.

Graphics credit: University of Chicago
Astronomical Data Collections in the Past...
Photographic Plates
In July 1850, daguerrotype photographer John Adams Whipple captured the first-ever picture of a star, Vega, using the observatory’s mahogany-and-brass Great Refractor. By the late 1880s, observatory director Edward Charles Pickering had endeavoured to photograph the entire sky, collecting photographs from the northern and southern hemispheres. The college shipped a 24-inch telescope to Arequipa, Peru, in 1896, and followed with telescopes in South Africa and other locations. During the next three decades, astronomers slipped glass plates into the observing tubes of their telescopes, making exposures of the entire sky, and then gathered them and shipped them back to Cambridge, Mass.

Result: ~525,000 plates!!

Credit: Rebecca Boyle, Popular Science

The North American Astronomical Photographic Plate Center

Credit: Hubble/ESA
Thousands of plates taken with Palomar Schmidt, UK Schmidt and ESO Schmidt telescopes between 1949 and 1990s to cover the whole sky in multiple bands.

Distribution to observatories in glass, film and paper copies.

Scanning started at STScI in 1986, published as DSS in 1994 and then turned into several catalogues as well.
Hubble Space Telescope
The Space Telescope European Coordination Facility was born out of a MoU in 1977 between NASA and ESA to provide the FOC, the solar arrays and 15 staff members in exchange of 15% observing time.

Community realised the need for more local support and expertise for observation planning and data reduction as well as a more direct access to HST data.
ECF was established with ESO in 1983.

Lots of cross-fertilisation of people, software, hardware and archiving technologies between ECF and ESO.

HST data analysis workshops merged with ADASS mid 1990s.

Advanced methods developed by ECF. Truly multi-national collaboration spanning many organisations.

Officially closed end of 2010.
...NOW

ALMA
...NOW

Murchison Widefield Array
MWA
Population density: 0.002/km$^2$
Overall MWA dataflow

**Tier 0 (MRO, Western Australia)**
- Online archive
- Online processing
- Correlator
- PFB
- Receiver
- Tile

400 MB/s

**Tier 1 (Perth, Western Australia)**
- MWA Buffer
- Science Archive
- NGAS Client
- Data Capture
- Vis + Voltage
- QA
- MS & Images
- RV
- pipeslines
- Galaxy / Magnus
- Pawsey, Perth (Disk + Tape)
- CRAR

**Tier 2 (Asia Pacific & North America)**
- Mirrored Archive
- USA
- Mirrored Archive
- India
- Mirrored Archive
- New Zealand
- Mirror Archive

**NGAS**

**MWA Buffer**

**Science Archive**

**Online archive**

**Online processing**

**Correlator**

**PFB**

**Receiver**

**Tile**

**Vis + Voltage**

**QA**

**MS & Images**

**RV**

**pipeslines**

**Galaxy / Magnus**

**Pawsey, Perth (Disk + Tape)**

**CRAR**

**Mirror Archive**

**NGAS Client**

**Web**

**UI**

**astropy**

**VO**

**aarnet**

**400 MB/s**

**Beam former**

**Tile**

**Beam former**
Very steady growth rate of ~410 TB/month\(^1\).

This converts to 166 MB/s sustained average for 39 months on a 750 km network link!

166 MB/s == 42% of MWA max. data rate.

Max. recorded data rate on link > 1.2 GB/s.

Slight drop recently in fact is a dramatic increase of a factor of 4, counteracted by a specially developed visibility(!) compression algorithm.

[1] About 2X the yearly rate suggested in M. Lacy, D.Halstead, ALMA NAASC memo 110

As of 1 Dec 2017
Long Haul Data Transport

20,662 Kilometers
350 MBytes/s
Global Data Transfer

From Perth to Boston (20,000 Kilometers)

➢ Almost real-time data transfer of full MWA data rate over 20,000 km!
➢ Not using any dedicated link or arrangement.
➢ Was one of the highest volume and throughput, long-term sustained, scientific data transfers using just public network.
Usage of MWA Archive

**Total: 24197 TB**

Not including automatic mirroring of data to partner institutions.

As of 06 December 2017
MWA Buffer

➢ 10 high density storage servers.
➢ total of ~3 PB usable capacity.
➢ running 2 NGAS servers each in cache mode.
Petascale Data Flow

- ASKAP 70 Tb/s, MWA 320 Gb/s, SKA
- ASKAP 1.5 PF/s, SKA >1 EF/s
- ASKAP 5 GB/s, MWA 8 GB/s, SKA >300 GB/s
- ASKAP 13 PB, SKA 800..3000 PB
- ASKAP 100 TF/s, SKA 30..100 PF/s
- ASKAP & MWA 6 TB/day, SKA
- ASKAP 100 TF/s, SKA 30 PF/s
- ASKAP 3 PB/year, SKA 18 PB/year

Adapted from Cornwell
Why do we think this is possible?
Projected Performance

Source: top500.org
# Data Rates and Volume

<table>
<thead>
<tr>
<th>Telescope</th>
<th>Raw Data Rate</th>
<th>Archive Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>MWA</td>
<td>1.4 TB/hour</td>
<td>5 PB/year</td>
</tr>
<tr>
<td>LSST</td>
<td>1.5 TB/hour</td>
<td>6 PB/year</td>
</tr>
<tr>
<td>ASKAP</td>
<td>9 TB/hour</td>
<td>5.5 PB/year</td>
</tr>
<tr>
<td>SKA1-LOW</td>
<td>1,400 TB/hour</td>
<td>150 PB/year</td>
</tr>
</tbody>
</table>

Note: A typical laptop has 1 TB of disk space. You can store all 7 Harry Potter books (~3,500 pages) 150,000 times on 1 TB!
SDP Key Performance Requirements  -- SKA Phase 1

SDP Local Monitoring & Control

- High Performance
  - ~100 PetaFLOPS

- Data Intensive
  - ~100 PetaBytes/observation (job)

- Partially real-time
  - ~10s response time

- Partially iterative
  - ~10 iterations/job (~6hour)

Telescope Manager

SDP Local Monitoring & Control

Data Processor

- High Performance
- Data Intensive
- Partially real-time
- Partially iterative

Data Preservation

- High Volume & High Growth Rate
  - ~100 PetaByte/year

- Infrequent Access
  - ~few times/year max

Delivery System

- Data Distribution
  - ~100 PetaByte/year from Cape Town & Perth to rest of World

- Data Discovery
  - Visualisation of 100k by 100k by 100k voxel cubes

Science Data Processor

~1 Gbytes⁻¹ (TBC)

~1 Tbytes⁻¹

~20 Gbytes⁻¹

~10 Gbytes⁻¹
Algorithms
what's the best algorithm to get the desired answer

Pipeline Logic
reduction components and sequence

Component Parameters
default parameter values of components

Data Parallelisation
hints about the potential of parallelism

Parallel Execution
what is executed where

Parallel Coding
writing parallel code

Code Optimisation
optimise parallel code

I/O Optimisation
optimise I/O on hardware

OS and hardware co-design
optimise hardware for code to be run
Logical Graph Template

Logical Graph

Physical Graph Template

Physical Graph

Algorithms
what’s the best algorithm to get the desired answer

Pipeline Logic
reduction components and sequence

Component Parameters
default parameter values of components

Data Parallelisation
hints about the potential of parallelism

Parallel Execution
what is executed where

Parallel Coding
writing parallel code

Code Optimisation
optimise parallel code

I/O Optimisation
optimise I/O on hardware

OS and hardware co-design
optimise hardware for code to be run
Bird’s eye view

scatter

corner turning

6 groups
(channels)

3 gathers (2 channels per subband)

3 AWImager iterations for each subband
SKA = SKAO + SRC

SKA Observatory (SKAO)

Joined SKAO-SRC functions
- User support for SKAO data products
- User support for SKAO provided software and tools
- Distribution of SKA data packs to users

SKA Regional Centers (SRC)

Essential SRC functions
- Development and provision of long-term SKA Science Archive
- Provision and management of computational resources for post-processing and analysis
- Provide platforms for continued development of software (pipelines and tools)
Global Network of SRCs

Multiple regional SRCs, locally resourced, heterogeneous in architecture, fully interoperable.
Thank You!