WTF?
Building Machine Astronomers to Discover the Unexpected in Astronomical Surveys

Ray Norris, Western Sydney University & CSIRO Astronomy & Space Science,
Overview

1. ASKAP/EMU
2. The process of astronomical discovery
3. WTF: discovering the unexpected in radio surveys
4. SETI
ASKAP: Australian SKA Pathfinder
A major new survey telescope in its own right
PAF = Phased Array Feed

Data Rate to correlator = 100 Tbit/s
= 3000 Blu-ray disks/second
= 62km tall stack of disks per day
= world internet bandwidth in June 2012

Processed data volume = 70 PB/year
Cray XC30 Series Supercomputer

472 Compute Nodes:
- 2 x 3.0 GHz Intel Xeon CPUs
- 10 Cores per CPU
- 64 GB DDR3-1866
- Cray Aries Interconnect
- Cray Dragonfly Topology
- 200 TeraFLOPS

1.4 Petabytes Lustre Data Storage
Recent image by Ian Heywood of NGC7232 region
• 12 antennas
• Exposure: about 20 hours
• Bandwidth: 48 MHz
• Area: 30 sq. deg.
• 2 pointings = 72 beams
• 150 uJy/beam RMS
• 2310 components > 5sigma
Deep radio image of 75% of the sky (to declination +30°)
Frequency range: 1100-1400 MHz

40 x deeper than NVSS (the largest existing radio survey)
- 10 µJy rms across the sky

5 x better resolution than NVSS (10 arcsec)
Better sensitivity to extended structures than NVSS
Will detect and image ~70 million galaxies at 20cm
c.f. 2.5 million detected over the entire history of radio-astronomy so far

All data to be processed in pipeline
Images, catalogues, cross-IDs, to be placed in public domain
## EMU Key Science Projects

<table>
<thead>
<tr>
<th>Key project</th>
<th>Title</th>
<th>Project Leader</th>
</tr>
</thead>
<tbody>
<tr>
<td>KP1.</td>
<td>EMU Value-Added Catalogue</td>
<td>Nick Seymour</td>
</tr>
<tr>
<td>KP2.</td>
<td>Characterising the Radio Sky</td>
<td>Ian Heywood</td>
</tr>
<tr>
<td>KP3.</td>
<td>EMU Cosmology</td>
<td>David Parkinson</td>
</tr>
<tr>
<td>KP4.</td>
<td>Cosmic Web</td>
<td>Shea Brown</td>
</tr>
<tr>
<td>KP5.</td>
<td>Clusters of Galaxies</td>
<td>Melanie Johnston-Hollitt</td>
</tr>
<tr>
<td>KP6.</td>
<td>cosmic star formation history</td>
<td>Andrew Hopkins</td>
</tr>
<tr>
<td>KP7.</td>
<td>Evolution of radio-loud AGN</td>
<td>Anna Kapinska</td>
</tr>
<tr>
<td>KP8.</td>
<td>Radio AGN in the EoR</td>
<td>Jose Afonso</td>
</tr>
<tr>
<td>KP9.</td>
<td>Radio-quiet AGN</td>
<td>Isabella Prandoni</td>
</tr>
<tr>
<td>KP10.</td>
<td>Binary super-massive black holes</td>
<td>Roger Deane</td>
</tr>
<tr>
<td>KP11.</td>
<td>Local Universe</td>
<td>Josh Marvil</td>
</tr>
<tr>
<td>KP12.</td>
<td>The Galactic Plane</td>
<td>Roland Kothes</td>
</tr>
<tr>
<td>KP13.</td>
<td>SCORPIO: Cataloguing the Radio Stars in our Galaxy</td>
<td>Grazia Umana</td>
</tr>
<tr>
<td>KP14.</td>
<td>WTF: Discovering the Unexpected</td>
<td>Ray Norris</td>
</tr>
<tr>
<td>KP15.</td>
<td>The Magellanic Clouds</td>
<td>Miroslav Filipovic</td>
</tr>
</tbody>
</table>
Size of radio continuum surveys over time

ASKAP Radio Continuum survey: EMU = 70 million

NVSS = 1.8 million

current total = 2.5 million
What fraction of discoveries in astronomy were “Popperian”?

Serendipity: 10
Predicted: 7

See also:
- Harwit (1981), Cosmic Discovery
- Wilkinson (2007) the Modern Radio Universe, 144
- Wilkinson (2015) (AASKA14), 65

From Ekers (2009) PoS(sps5)007
## Discoveries with HST

<table>
<thead>
<tr>
<th>Project</th>
<th>Key project</th>
<th>Planned?</th>
<th>Nat. Geo. top ten?</th>
<th>Highly cited?</th>
<th>Nobel prize?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Cepheids to improve value of H0</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>study intergalactic medium with uv spectroscopy</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium-deep survey</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image quasar host galaxies</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measure SMBH masses</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exoplanet atmospheres</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planetary Nebulae</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discover Dark Energy</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Comet Shoemaker-Levy</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep fields (HDF, HDFS, UDF, FF, etc)</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Proplyds in Orion</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRB Hosts</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*from Norris arXiv:1611.05570*
<table>
<thead>
<tr>
<th>Project</th>
<th>Key</th>
<th>Planned?</th>
<th>Nat.</th>
<th>Highly cited?</th>
<th>Nobel prize?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Cepheids to improve value of $H_0$</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>study intergalactic medium with uv spectroscopy</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium-deep survey</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image quasar host galaxies</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measure SMBH masses</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exoplanet atmospheres</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planetary Nebula discoveries</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discover Dark Energy</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Comet Shoemaker</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep fields (HDF, HDFS, UDF, FF, etc)</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proplyds in Orion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRB Hosts</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
</tr>
</tbody>
</table>

**Summary:**

Of the “top ten” HST discoveries:

- 1 was a key project
- 4 were planned by astronomers but were not key projects
- 5 were totally unexpected (e.g. dark energy)
How to make unexpected discoveries with next-generation surveys (e.g. ASKAP-EMU)?
Uncharted observational phase space

20cm radio continuum surveys

Area (deg^2)

5\sigma Sensitivity (mJy)
Radio Continuum Surveys: area vs. sensitivity

EMU
Apertif

Log(area/degrees)

Log(Limiting sensitivity/mJy)

Single Dish
Interferometer Array
Single-pixel synthesis array
Phased array
PAF synthesis array
Cylinder synthesis array
Other

MIGHTEE
COSMOS
FLS-WSRT
SWIRE
Chilesconpol
eMERGE
Frontier Fields

ATLAS
LH-W

ASKAP-BETA
LotSS-HETDEX

ASKAP-ESP3

Stripes82
ATLAS-SPT

GLASS

VLASS
LOFAR
FIRST
SUMSS
WENSS

NVSS

TGSS
Can ASKAP Discover the Unexpected?

• Data volumes are huge – cannot sift by eye
• Instrument is complex – no single individual will be familiar with all possible artifacts
• ASKAP will be superb at answering well-defined questions (the “known unknowns”)
• Humans won’t be able to find the “unknown unknowns”
• Can we mine data for the unexpected, by rejecting the expected?

If not, ASKAP will not reach its full potential i.e. it will not deliver value for money
mining radio survey data for the unexpected

WTF?

WTF = Widefield outlier Finder
An example of what a WTF/SETI detection might look like

- Spectral index
- Optical luminosity
- Variability
- Polarisation

Planetary nebulae
WTF? SETI?
IFRS
Radio galaxies
Blazars
Hi-z USS
Optical luminosity
Mining large data sets for the unexpected

WTF will work by searching the n-dimensional (large n) phase space of observables, using techniques (both supervised and unsupervised) such as

- kNN (k-nearest-neighbours)
- Neural nets/deep learning
- Self-organised maps
- Support vector machine
- Random forest

Identified objects/regions will be either

- processing artifacts (important for quality control)
- statistical outliers of known classes of object (interesting!)
- New classes of object (WTF)

Earlier public challenge less successful than hoped

- Partly because of difficulty of providing a good WTF training set (e.g. face)
- But attracted collaborators from computer science community
Type 1 discoveries: Unexpected objects (e.g. pulsars, quasars) Simple anomaly detection – right?
Type 2 discoveries: unexpected phenomena (e.g. Hubble expansion, dark energy, dark matter)
“Machine learning in astronomy” collaboration

• An open collaboration – see mlprojects.pbworks.com

• Participants from Astronomy, Maths, & Engineering Depts. at several Australian Universities + CSIRO, EMU project, etc.

• Projects include
  • Ray Norris+: building training/test sets (placed in public domain)
  • Laurence Park+: radio source classification techniques
  • Gary Segal (PhD)+: Anomaly detection for WTF
  • Pero Manojlovic (PhD)+: Finding bent-tail galaxies
  • Kieran Luken (M. Res) et al.: photometric redshifts
  • Nick Ralph (M. Res)+: learning from ASKAP monitoring data
  • Katherine James (vac. stud.)+: radio source classification with CNN

• See also
  • Baron+Poznanski 2016, The weirdest SDSS galaxies: results from an outlier detection algorithm, arXiv:1611.07526
  • Aniyan+Thorat 2017 Classifying Radio Galaxies with Convolutional Neural Network arXiv:1705.03413
We have found that much of our ML work is limited by the lack of good training sets.

We are making the following training sets publicly available on mlprojects.pbworks.com:

Available now/very soon:

- **ATLAS DR3 raw training set**
  5000 radio sources with IR cross-IDs, labels, 90% simple, 10% double/triple/etc

- **ATLAS DR3 enhanced training set**
  4500 simple radio sources, 4500 complex with IR cross-IDs, labels, etc

- **Synthetic training set**
  20,000 simple, 20,000 complex (tricky – beware of Russian tanks!)

Available in future:

- Bent tails
- **WTF/SETI** (tricky – limited by our imaginations)
- Redshifts (SEDs + spec-z’s)
Self-organised maps
courtesy Kai Polsterer & Enno Middelberg
Proposal:
- ~30 dual-polarisation tied-array beams, within 30 sq deg field
- each within one of the PAF beams
- 300 MHz bandwidth, divided into 18 kHz channels
- Piggyback mode: within 5 deg of pointing centre
- Dedicated SETI mode: 30 independently steerable beams
WTF is a SETI Detection?

• If we get a SETI detection, it will typically be ~2.8 Gyr older than us
• So we have as much in common with ET as we do to a slime-mould
• We are unlikely to be able to predict (or even recognise) a SETI signal
• So the best strategy is to search for something which is not astrophysical
• So the SETI problem is very similar to WTF problem.
• Very amenable to machine learning algorithms (neural nets, deep learning, self-organised maps, spatial local outlier, etc.)
• Likely SETI discriminators include circ. polarisation, bandwidth, etc.
• Challenge: how do we build good training sets?
• Instead, use anomaly detection techniques (e.g. Baron+Poznanski)
We acknowledge the Wajarri Yamaji people as the traditional owners of the ASKAP site.

See arXiv:1611.05570