WTF?
Discovering the Unexpected

Ray Norris, Western Sydney University &
CSIRO Astronomy & Space Science,
How does science work?

Karl Popper: experiments test theory!
• e.g. High energy physics, LHC, Higgs Boson
• Falsifiable predictions remain the “gold standard” of good
Kuhn et al. showed Popperian science is not the only mode (e.g. exploration, understanding, insight)

Astronomy usually works more in an “explorer” mode
Ron and I often used to discuss this...
The great thing about working with Ron is that we were both excellent communicators.
What fraction of discoveries in astronomy were “Popperian”?

Serendipity: 10
Predicted: 7

From Ekers (2009) PoS(sps5)007
See also:
- Harwit (1981), Cosmic Discovery
- Wilkinson (2007) the Modern Radio Universe, 144
- Wilkinson (2015) (AASKA14), 65

From Ekers (2009) PoS(sps5)007
<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 et al. 2013: arXiv1210.7521*
## Discoveries with HST

(see e.g. Lallo: arXiv:1203.0002)

<table>
<thead>
<tr>
<th>Project</th>
<th>Key</th>
<th>Planned?</th>
<th>Nat.</th>
<th>Highly</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 mass</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>

### 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)
The process of astronomical discovery
The discovery of pulsars

Jocelyn Bell:
• explored a new area of observational phase space
• knew the instrument well enough to distinguish interference from signal
• observant enough to recognise a sidereal signature
• open minded – prepared for discovery
• within a supportive environment
• persistent

See Bell-Burnell (2009) PoS(sps5)014 for a personal perspective
Could Jocelyn Bell make that discovery with next-generation surveys (e.g. ASKAP-EMU)?
20cm radio continuum surveys

Uncharted observational phase space

ASKAP-EMU
75% of sky
Rms=10μJy,
res ~ 10 arcsec
~70 million galaxies
Would take ~7 years with JVLA

VLA-NVSS
75% of sky
Rms=450μJy,
res ~ 45 arcsec
~1.8 million galaxies

5σ Sensitivity (mJy)
Typical ATLAS image courtesy of Minnie Mao
PAFs => Big Data

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
ASKAP Science Data Processor Platform

- The *galaxy* system at Pawsey
- 472 x Cray XC30 Compute Nodes
  - 200 TFlop/s Peak
- Cray Aries (Dragonfly topology)
- Cray Sonexion Lustre Storage
  - 1.4 PB usable
  - 480 x 4TB Disk Drives, RAID 6 + Hot Spares
  - Peak I/O performance: 30 GByte/s
Could Jocelyn Bell Discover the Unexpected in ASKAP data?

- 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
What does ASKAP need to do to discover the unexpected?

• **Maximise the volume of new phase space**
  • E.g. all-sky survey, extend parameter range, or very deep

• **Retain flexibility**
  • don’t optimise the telescope ONLY for science goals

• **Develop data mining software to search for the unexpected**
  • This will be an important part of data-intensive research
mining radio survey data for the unexpected

WTF = Widefield ouTlier Finder
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)
E.g. support vector machine approach:

- Finding unexpected objects
  - finding classes of unclassified objects
  - finding anomalous objects
- p measurables (E.g. colours/spectral indices/morphologies)
- set up a training set of known types of object.
- Arrange in a phase space
- Are there parts of the phase space which are observable but don’t contain known objects?
- Represent each object by a vector with p components
- What line/hyperplane most clearly bounds the known objects?
- Or, equivalently, what line/hyperplane maximally separates known objects from unknown objects?
Self-organised maps
courtesy Kai Polsterer & Enno Middelberg
WTF Phase 1 (2015-early 2016)

• Received a grant from Amazon Web Services to develop WTF on the AWS cloud platform

• Goals:
  • Implement WTF, initially as an open challenge (c.f. Kaggle)
  • Evaluate AWS platform as a collaborative research environment

• Approach
  • Set up challenges consisting of data (images and tables) with embedded “EMU eggs”
  • Data include both simulations and real data
  • Invite ML and other algorithm groups to discover the EMU eggs
  • Develop visualisation tools to understand the process and data
Built Data Challenges, invited ML groups to find buried “EMU eggs”

Results:

(a) Some people solved the challenge using innovative ways round our process

(b) Others found them too hard – the problem was too loosely specified (e.g. “WTF am I supposed to do with this?”)
The Ekers criterion: If you don’t have the occasional failure then you’re not being sufficiently ambitious

✔ Ekers criterion
× Perhaps a little over-ambitious
WTF Phase 1 outcomes: A learning experience!

Lessons learned:

- Challenges of using AWS
- Preparing the data is a major task and takes far more time and thought than expected.
- Tests for evaluating algorithms is non-trivial. The obvious tests often get it wrong.
- Difficult to design algorithms to discover the unexpected when you don’t yet have algorithms to discover the expected!
- Decided to re-think process and walk before we run.
The WSU Astrophysical machine learning group

• Still ramping up
  • Staff from Astronomy, Maths, Engineering
  • Collaborators from ANU, U. Herts, CSIRO-CASS & CSIRO-Data61
  • 4 graduate students potentially starting early 2017 (2 PhD, 2 Masters)

1) Build up group with local expertise

2) Work on well-defined EMU problems (known-unknowns), such as
  • Radio source classification and cross-identification (lead: Ray Norris WSU/CSIRO)
  • Photometric & Statistical redshifts (lead: Kieran Luken, WSU, & Chris Wolf, ANU)
  • Detection of SETI signals (lead: Ray Norris & Ain de Horta, WSU)
  • Detection of time-varying sources (lead: Martin Bell, CSIRO)
  • Intelligent ASKAP monitoring (Nic Ralph, Malte Marquarding, Craig Haskins)
  • Image error recognition and artefact removal (TBD)
  • RFI Mitigation (TBD)

3) Eventually extend techniques to the much harder WTF problem
EMU Source identification and classification

Best expert reliability:
- NVSS 90%
- ATLAS 99%

Current projects using ATLAS as a testbed:
- Expert manual cross-ID (lead: Jesse Swan, U. Tas)
- Likelihood ratio (lead: Stuart Weston, AUT)
- Radio Galaxy Zoo (lead: Julie Banfield, ANU & Ivy Wong, ICRAR)
- Bayesian (lead: Dongwei Fan, NO/CAS, & Tamas Budavari, JHU)
- Machine Learning 1 (lead: Ray Norris, WSU/CSIRO)
- Machine Learning 2 (lead: Julie Banfield, ANU)
- Comparison of Techniques
WTF Phase 2 (2016-7)

- Start developing modules which will become the elements of the WTF machine
- Test data for WTF at each stage
- Includes source classification, cross-ID, artefact removal, etc
- Test on EMU Early science
Flowchart for discovering unexpected objects
Flowchart for discovering unexpected phenomena

1. Survey data
2. Generate characteristic distributions
3. Test for differences between real and simulated distributions
4. Report differences between theory and experiment

Flowchart:
- **Survey data**
  - Our current knowledge
    - Generate Simulated Universe
    - Generate Simulated Sky
    - Generate simulated observations
    - Apply observational window functions
    - Generate characteristic distributions
    - Test for differences between real and simulated distributions
    - Report differences between theory and experiment

Window functions
WTF Phase 3: Re-start WTF challenge

- Set up data sets for challenge using EMU data
- Include both image and tables, including multiwavelength data
- Include well-documented:
  - Training sets
  - Simulated discovery sets
  - Real EMU data
- Focus on in-house research
- Also invite other groups to beat us
Can we create a machine that replicates Ron’s brain, thinking outside the box?
We acknowledge the Wajarri Yamaji people as the traditional owners of the ASKAP site.

YOU ARE NOW LEAVING THE MURCHISON RADIO-ASTRONOMY OBSERVATORY
THANK YOU FOR BEING RADIO QUIET

See MLprojects.pbworks.com