www.csiro.au

The Australian SKA Pathfinder EMU survey: mining radio survey data for the unexpected

Ray Norris
Astroinformatics Redmond September 2012

CSIRO

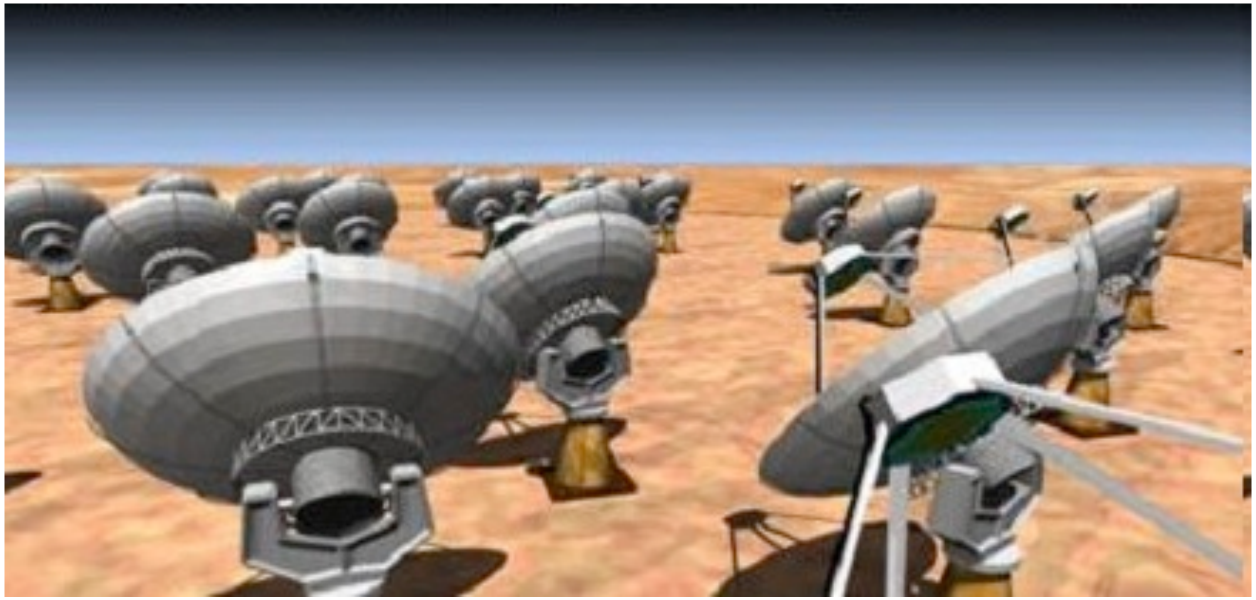
Overview

- **ASKAP Overview**
- **EMU overview**
- **EMU data challenges**
- **Mining the unexpected in large datasets**

- **Caveat:**
  - Everything in this talk is superficial
  - Each bullet point could be expanded to a 1-hr talk!

ASKAP=Australian SKA Pathfinder

- **A$170m (=US$170m) project now under construction in Western Australia**
- **Completion 2014?**
- **36*12m antennas**
- **Antennas have a 192-pixel phased array feed (PAF)**
- **30 sq. deg FOV!**



This is a … y migrated

ASKAP Design Specifications

• Number of antennas	36 (630 baselines)
• Antenna diameter	12 m (3 axis)
• Maximum baseline	6 km
• Cont. Angular resolution	10 arcsec
• Sensitivity	65 m^2/K
• Frequency range	700 – 1800 MHz
• Focal plane phased array	192 elements (96 dual pol)
• Field of view	30 deg^2
• Processed bandwidth	300 MHz
• Number of channels	16 384

Antennas

Antennas built by CETC54 (China)

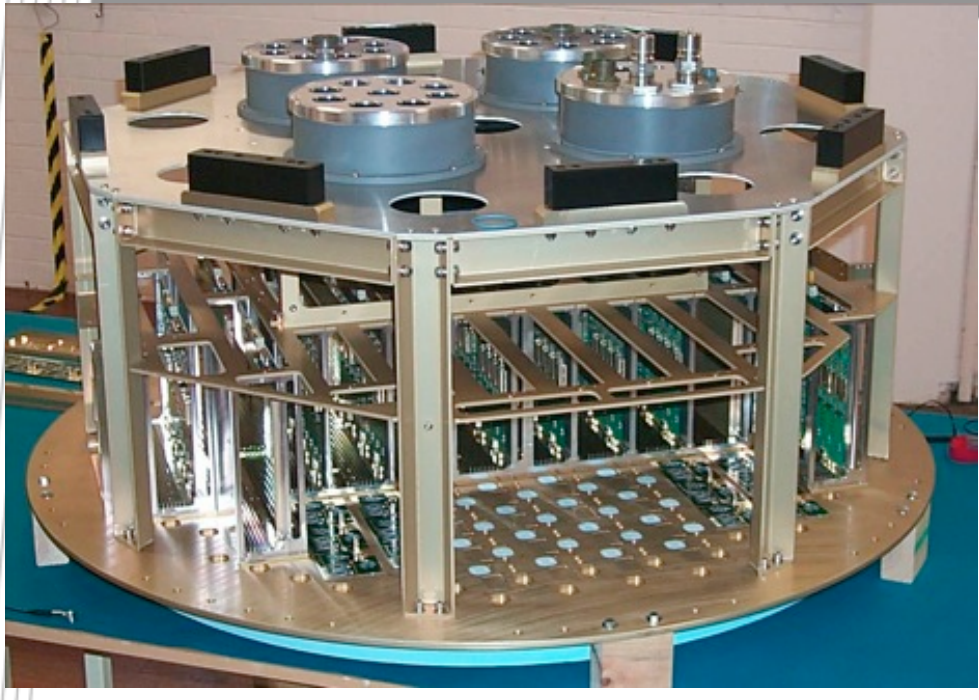
Delivered and assembled on site

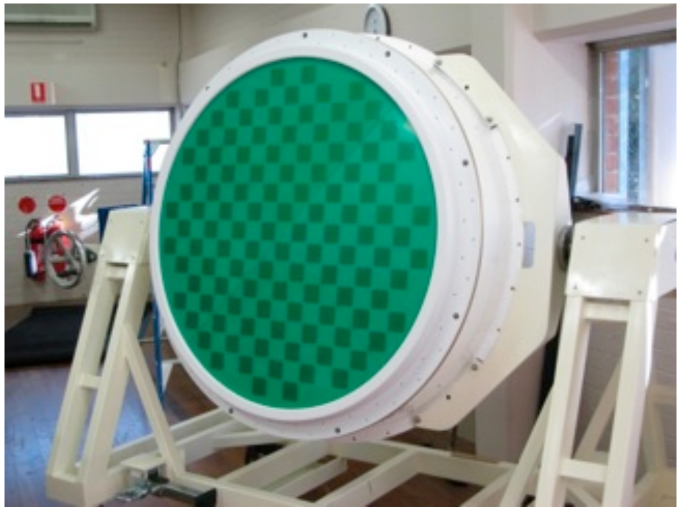
Antenna 1 delivered late 2009

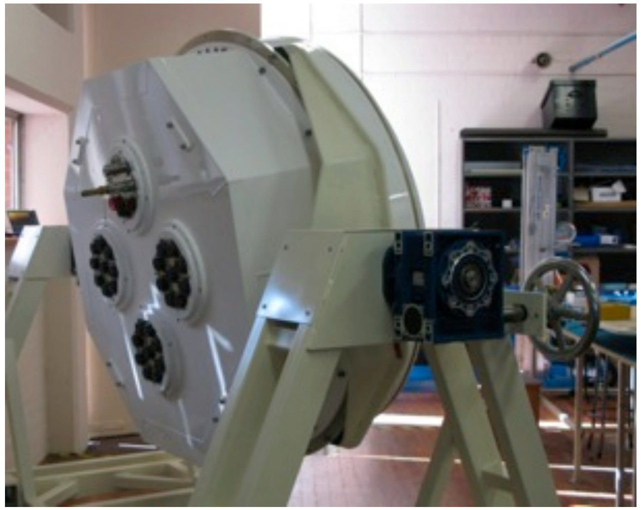
Antenna 36 delivered early 2012

surface rms < 0.5mm

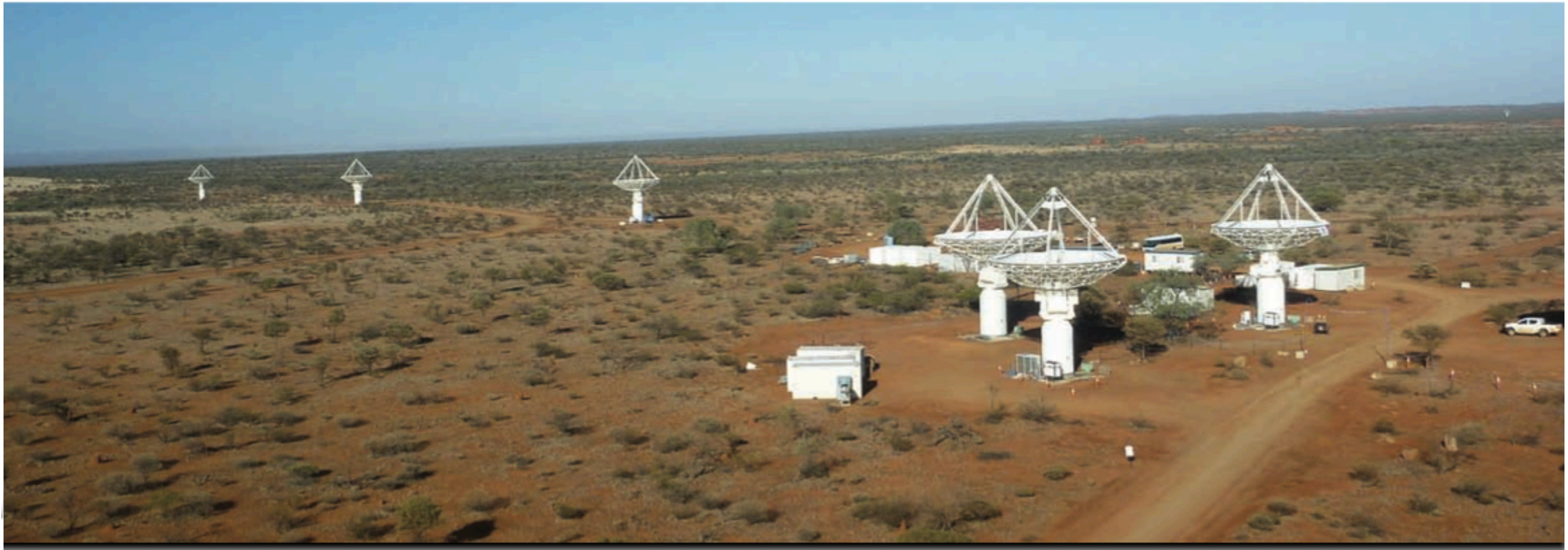
Phased-Array Feeds (PAFs)



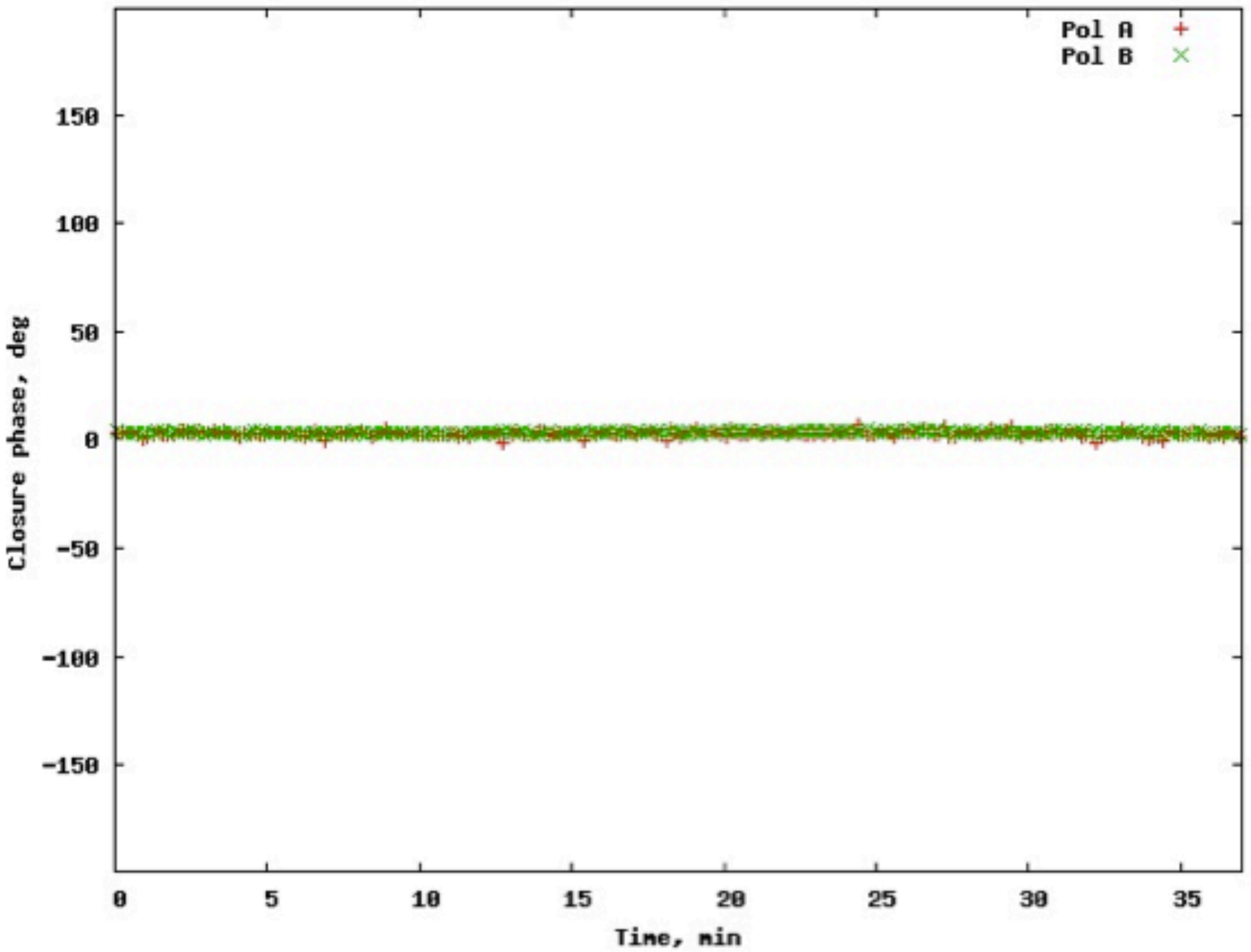




ASKAP in early August 2012 – all 36 antennas completed



PAF Closure phase achieved – 20 August 2012



ASKAP Science

38 proposals submitted to ASKAP

2 selected as key projects

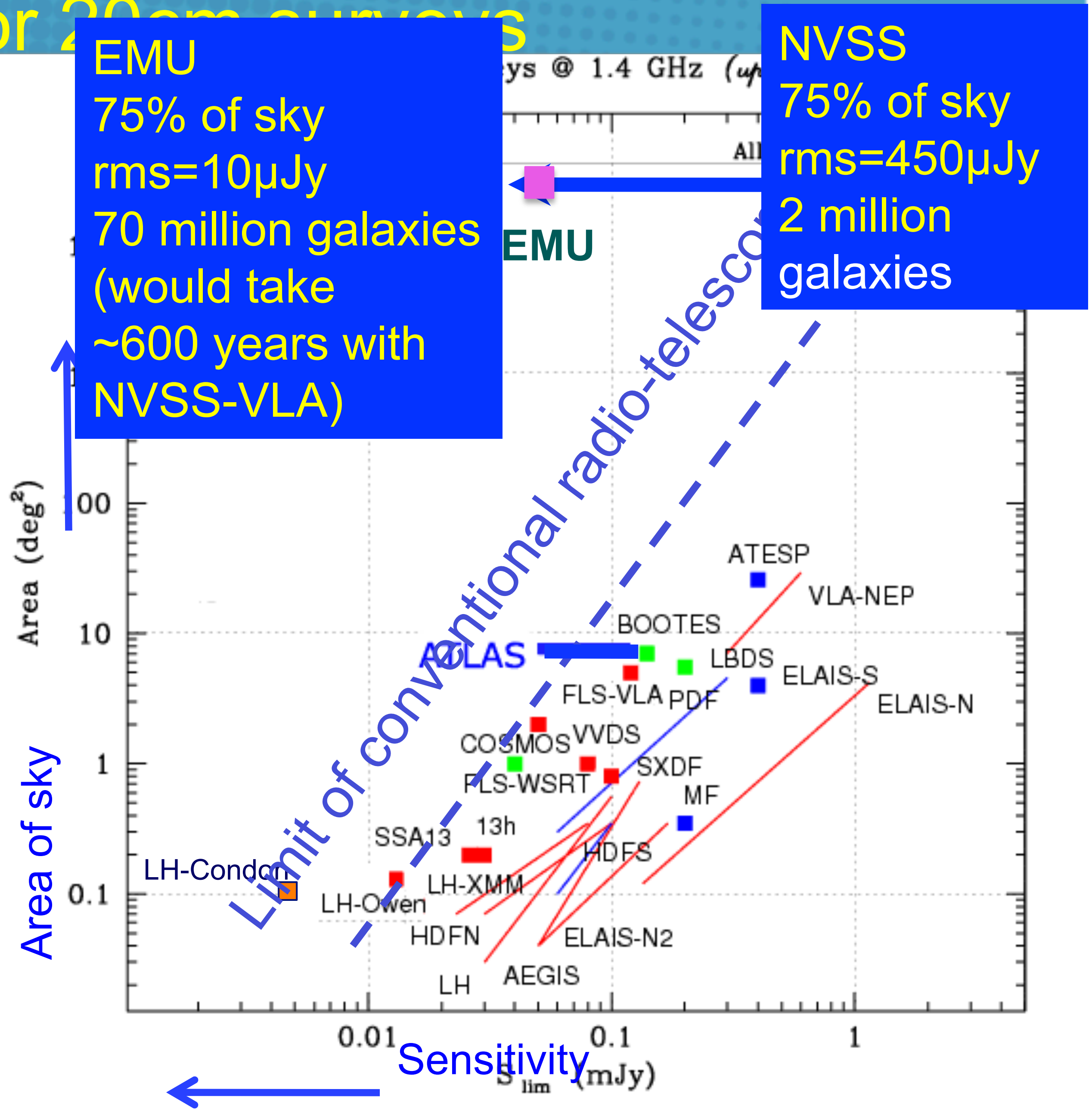
- EMU all-sky continuum (PI Norris)
- WALLABY all-sky HI (PI Koribalski & Staveley-Smith)

8 others approved at lower priority

- COAST pulsars etc
- CRAFT fast variability
- DINGO deep HI
- FLASH HI absorption
- GASKAP Galactic
- POSSUM polarisation
- VAST slow variability
- VLBI

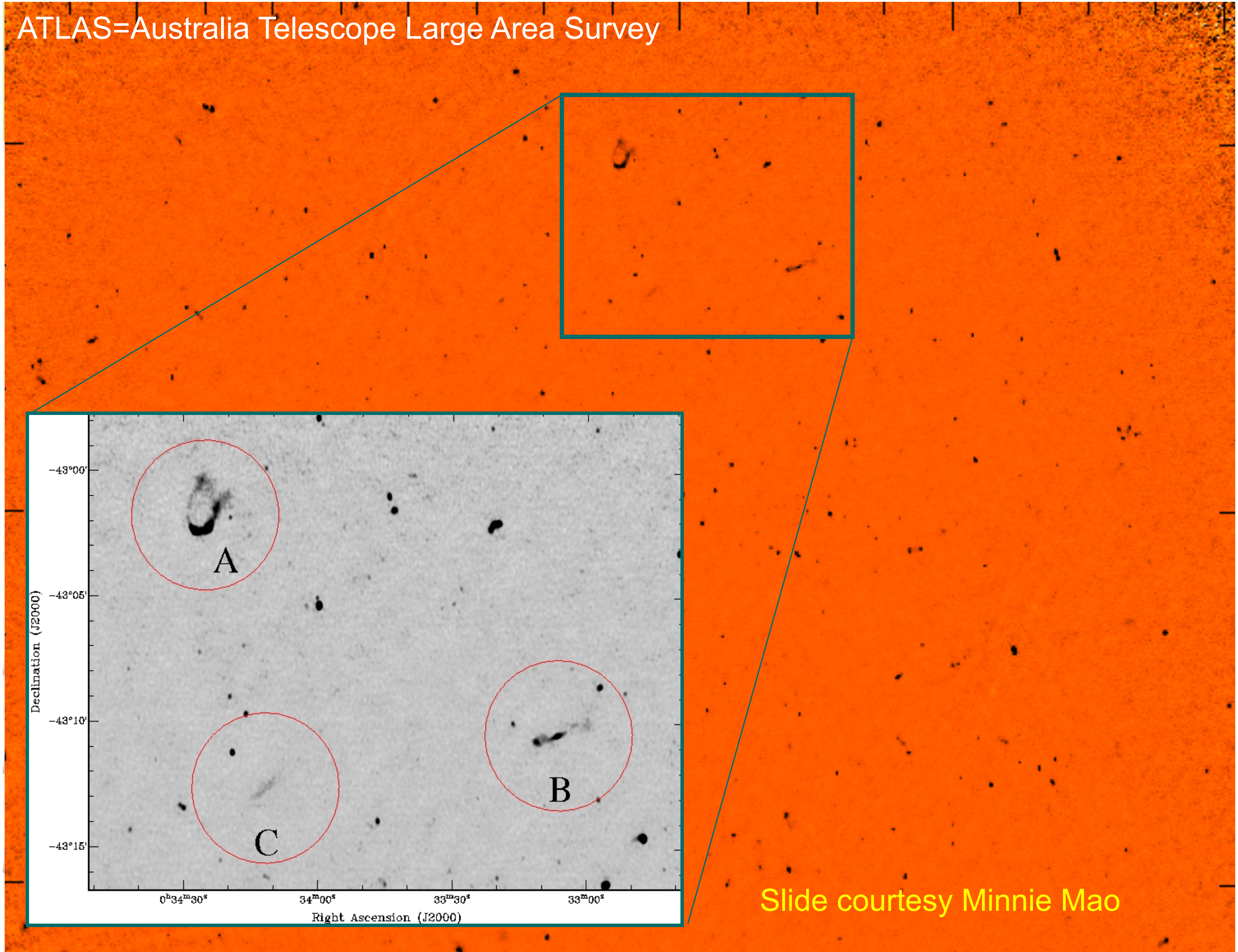
- Deep radio image of 75% of the sky (to declination +30°)
- Frequency range: 1100-1400 MHz
- 45 x deeper than NVSS
  - 10 μJy rms across the sky
- 5 x better resolution than NVSS (10 arcsec)
- Better sensitivity to extended structures than NVSS
- Will detect and image ~70 million galaxies at 20cm
- All data to be processed in pipeline
- Images, catalogues to be placed in public domain
- Survey starts 2014(?)
- Total integration time: ~1.5 years ?
- Project includes cross-ID's and redshifts

Observational phase space: Current major 20cm surveys



ASKAP/EMU data challenge

- **Raw data from antennas: 9 Tbits/s**
- **Processed by correlator and beamformer at site (10 MW power)**
- **10 Gb/s shipped to processor at Perth**
- **All processing (selfcal, fft, deconvolution, source extraction) done in Perth at Pawsey HPC centre**
- **Required uv data storage 70 Pbyte/yr**
- **Can only afford to store 4 Pbyte/yr**
- **EMU images: 100 Tbyte**
- **Source extraction -> EMU catalogs: 30 Gbyte (public domain)**
- **Cross-ids, redshifts -> Value-added catalog: 50 Gbyte**



ATLAS=Australia Telescope Large Area Survey
A
B
C
Declination (J2000)
-43°00'
-43°05'
-43°10'
-43°15'
Right Ascension (J2000)
0h34m30s
34m00s
33m30s
33m00s
Slide courtesy Minnie Mao

ASKAP EMU

Evolutionary Map of the Universe

Science Goals

How did galaxies form and evolve?

Science Goals

1) Evolution of SF from z=2 to the present day,
- using a wavelength unbiased by dust or molecular emission.

2) Evolution of massive black holes
- how come they arrived so early? How do binary MBH merge?
- what is their relationship to star-formation?

3) Explore the large-scale structure and cosmological parameters of the Universe.
- E.g. Independent tests of dark energy models

4) Explore an uncharted region of observational parameter space
- almost certainly finding new classes of object.

5) Explore Clusters & Diffuse low surface brightness radio objects

6) Generate an Atlas of the Galactic Plane

7) Create a legacy for surveys at all wavelengths (Herschel, JWST, ALMA, etc)

Technical Challenges

- **Survey Strategy**
- **Performance of PAF**
  - uniformity, poilarisation, sidelobes, etc.
- **Image Processing**
  - Dynamic range, calibration, sensitivity as function of scale size, etc.
- **Source Extraction**
- **Cross-identification**
- **Redshifts**
- **Data delivery (Value-added catalogue/VO)**

Source Extraction
(WG chair: Andrew Hopkins, AAO)

- **EMU source extraction team currently exploring available source finders (sExtractor, sfind, DuChamp, etc).**
- **None are yet optimum**
- **Will incorporate optimum algorithm into ASKAP processing pipeline**
- **See (e.g.)**
  - Compact continuum source finding for next generation radio surveys (Hancock, P.J., Murphy, T., Gaensler, B.M., Hopkins, A., & Curran, J.R. 2012, mnras, 422, 1812)
  - The completeness and reliability of threshold and false-discovery-rate source extraction algorithms for compact continuum sources (Huynh, M., Hopkins, A., Norris, R., et al. 2011, arXiv:1112.1168)
  - BLOBCAT: Software to Catalogue Flood-Filled Blobs in Radio Images of Total Intensity and Linear Polarization (Hales, C.A., Murphy, T., Curran, J.R., et al. 2012, arXiv:1205.5313)

Cross-Identification for EMU
(WG chair: Loretta Dunne, Canterbury Uni)

- **We plan to develop a pipeline to automate cross-IDS**
  - using intelligent criteria
  - not simple nearest-neighbour
  - working closely with other survey groups
  - use all available information (probably Bayesian algorithm)
- **Expect to be able to cross-ID 70% of the 70 million objects**
- **20% won't have optical/IR ID's**
- **What about the remaining 10% (7 million galaxies)?**


What
GALA
Home
How To Take
the
ct

Redshifts (WG chair: Nick Seymour, CSIRO)

- **Only ~1% of EMU sources will have spectroscopic redshifts (most from WALLABY)**
- **Generating photometric redshifts for AGNs is notoriously unreliable**
- **EMU redshift group (Seymour, Salvato, Zinn, et al) exploring a number of different approaches:**
  - template fitting
  - kNN algorithms
  - SoM algorithms
  - etc

Warning: paradigm shift approaching!

For many questions addressed by large surveys, the properties of the individual objects are less important than the properties of ssmples of the population.

E.g. For a cosmological test, you don't care about the z of an individual galaxy – what is the ISW of the population at z=0.1 c.f. those at z=0.5. This is a much easier question.

Examples:

1) Polarisation

- mean redshift of polarised sources ~1.9
- mean redshift of unpolarised sources ~1.1

2) Spectral index

- Steep spectrum sources have a higher redshift than moderate spectrum sources

3) Radio-k relation

- High values of $S_{20cm}/S_{2.2\mu m}$ have high z
- even a non-detection is useful

Combine all such indicators (+others) to assign a probabilistic redshift distribution (=> statistical redshifts)

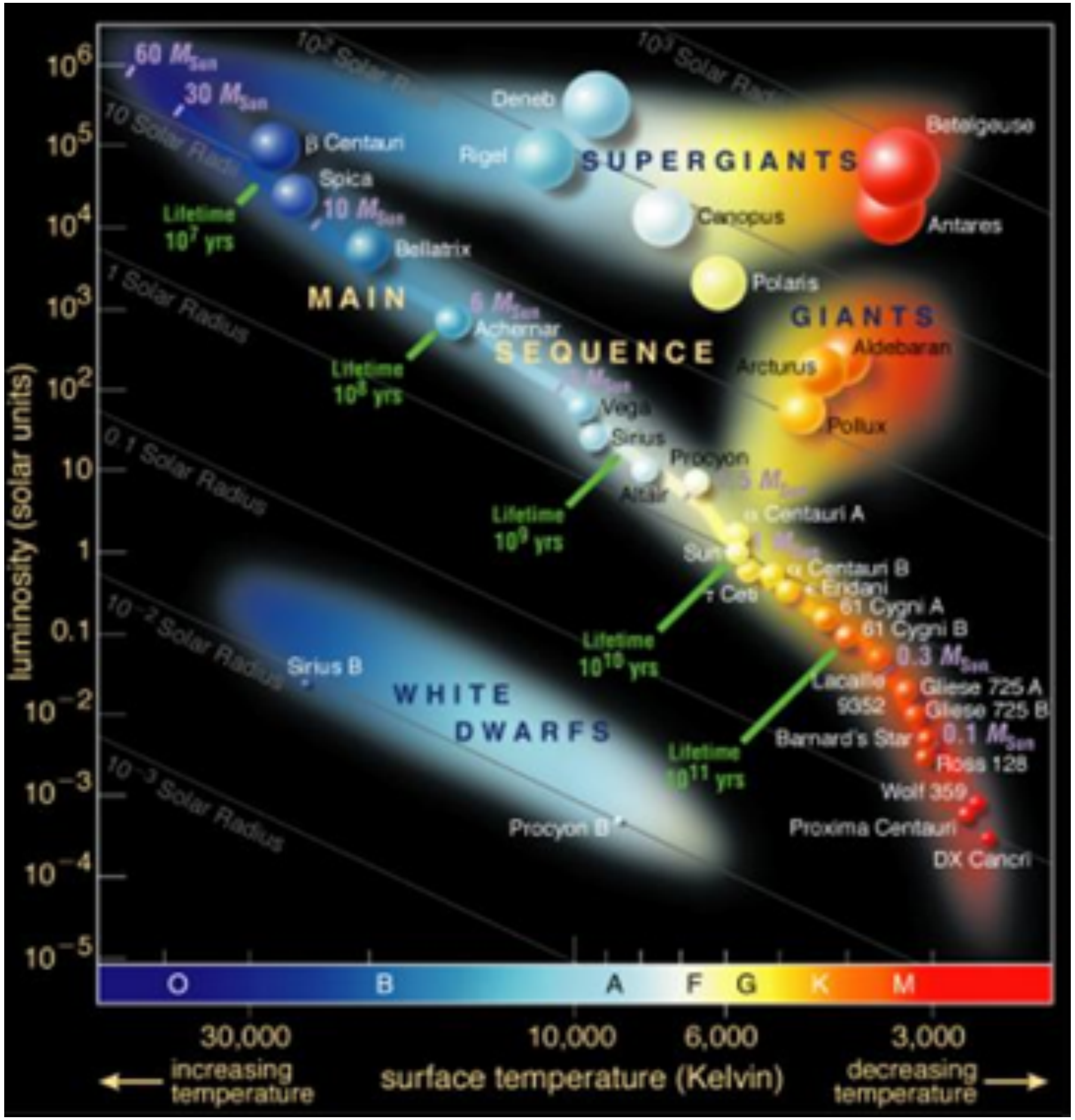

mining radio survey data
for the unexpected

Science Goal 4: Discovering the Unexpected

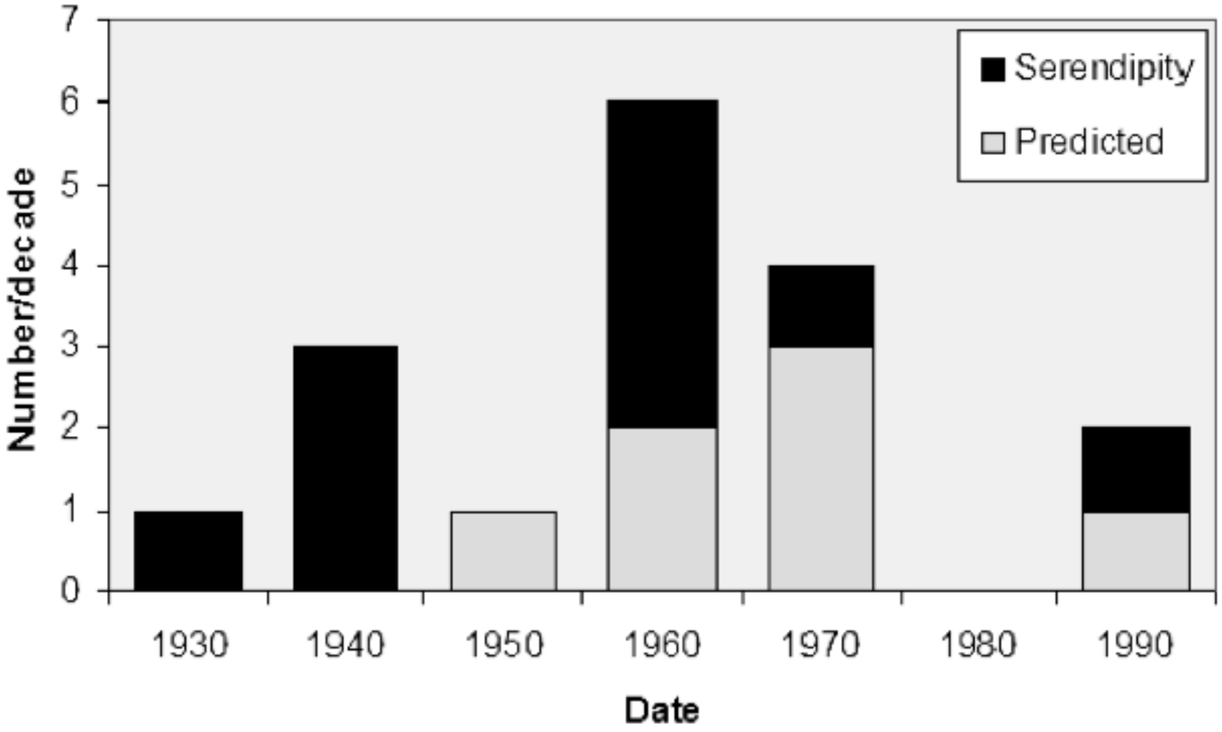
WTF?

WTF = Widefield ouTlier Finder

Astronomy usually works in an "explorer" mode, rather than testing hypotheses



What fraction of recent major discoveries in astronomy were "Popperian"?



(b) Predicted v Serendipity

+1 for dark energy (2012)

Serendipity:11
Predicted: 7

From Ekers (2009) PoS(sps5)007

Example: The discovery of pulsars

Jocelyn Bell:

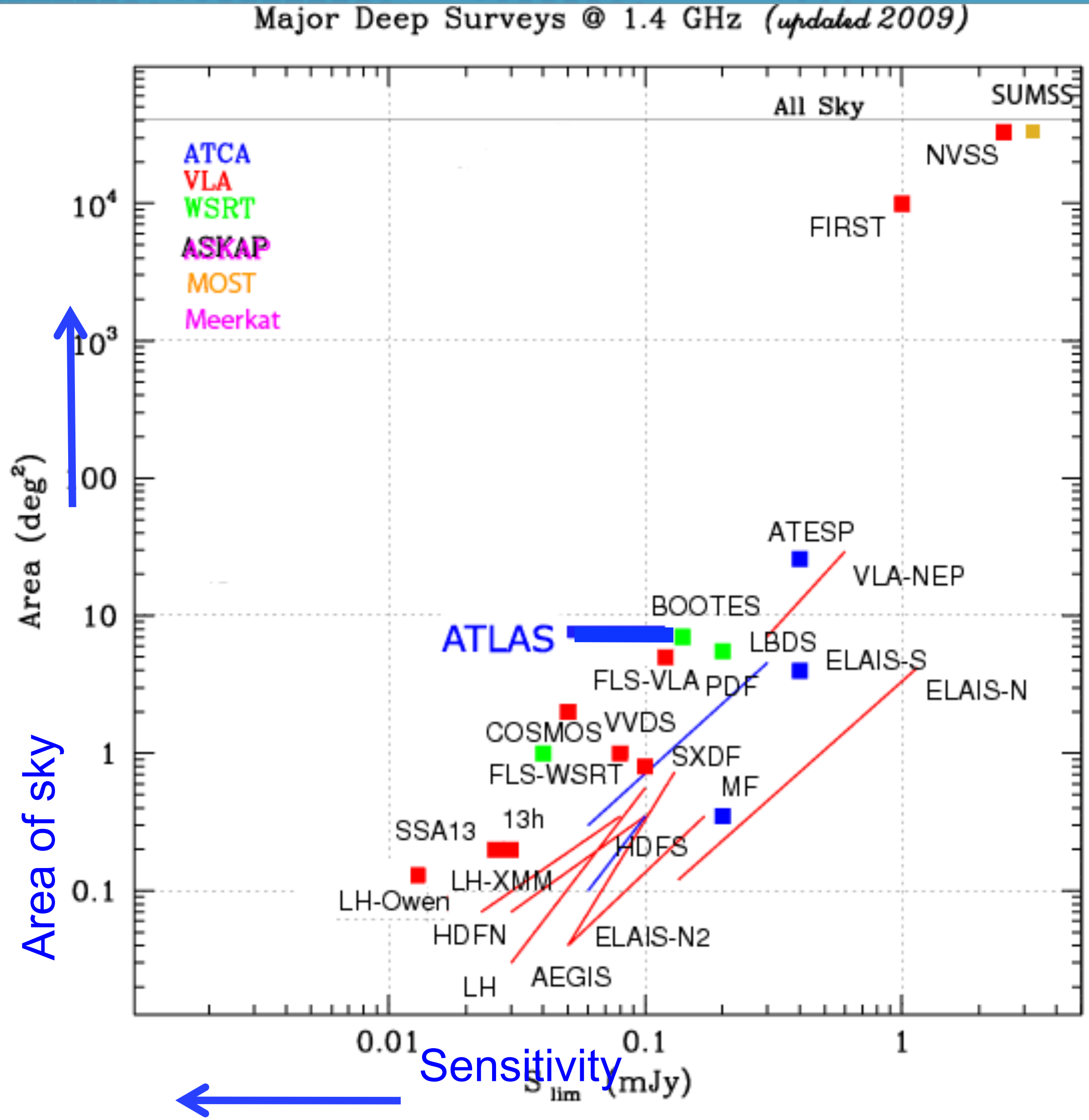
- **explored a new area of observational phase space**
- **knew the instrument sufficiently well to distinguish interference from signal**
- **observant enough to recognise a sidereal signature**
- **open minded – prepared for discovery**
- **within a supportive environment**
- **persistent**

See Bell-Burnell (2009) PoS(sps5)014 for a personal perspective

Current major 20cm surveys

Are the discoveries distributed uniformly across this diagram?

Is the difficulty of finding them spread uniformly across this diagram?



Discovering the Unexpected?

- Certainly we're sampling new parameter space
- But our data volumes will be huge
- Nobody will have sufficient familiarity with the data or with the instrument to be a "Jocelyn Bell"
- Instead we will find (or not find) what we are looking for.
- We won't find things we are not looking for (the "unknown unknowns")
- Can we mine data for the unexpected, by rejecting the expected?

Should we try?

- **EMU will discover 70 million radio objects, most of them previously unknown**
  - <z>= 1.8 for AGN, <z>=1.1 for SF galaxies
- **Experience suggests that there are new discoveries to be made in this dataset**
- **If we don’t tackle this problem, then we are failing to extract the maximum science from the data**

Discovering the Unexpected

(

Unlikely to stumble across new types of object,

Instead, systematically mine the EMU database,

- discarding objects that already fit known classes of object

Identified objects/regions will be either

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

How to find the unexpected?

- **Decision tree approach: Attempt to classify all objects with optical IDs etc, using all available properties, and flag those with good data that cannot be classified**
- **Zoo approach: put all "odd" sources on RadioZoo, and see if anybody spots something odd.**
- **Cluster analysis: assemble all n properties of data in an n-dimensional space. Most will cluster. Flag those that don't.**
- **kFN: opposite of kNN approach (similar to cluster analysis?)**
- **SoM: self-organised maps**
- **Bayesian approach (aka infinite improbability drive): given our knowledge of physics/telescope, how likely is this data?**
- **Ensemble approach – use all the above. And what else?**

EMU is an open project

- WTF currently in formative stage - collaborators invited
- If successful, approach should be applicable to other large surveys

- If you have good ideas on any of the above, we'd love to work with you!

- Data challenge to be issued late 2012 using ATLAS data
- Initially focussed on EMU


YOU ARE NOW LEAVING THE
MURCHISON RADIO-ASTRONOMY
OBSERVATORY
THANK YOU FOR BEING RADIO QUIET
For more info: arXiv **1106.3219**