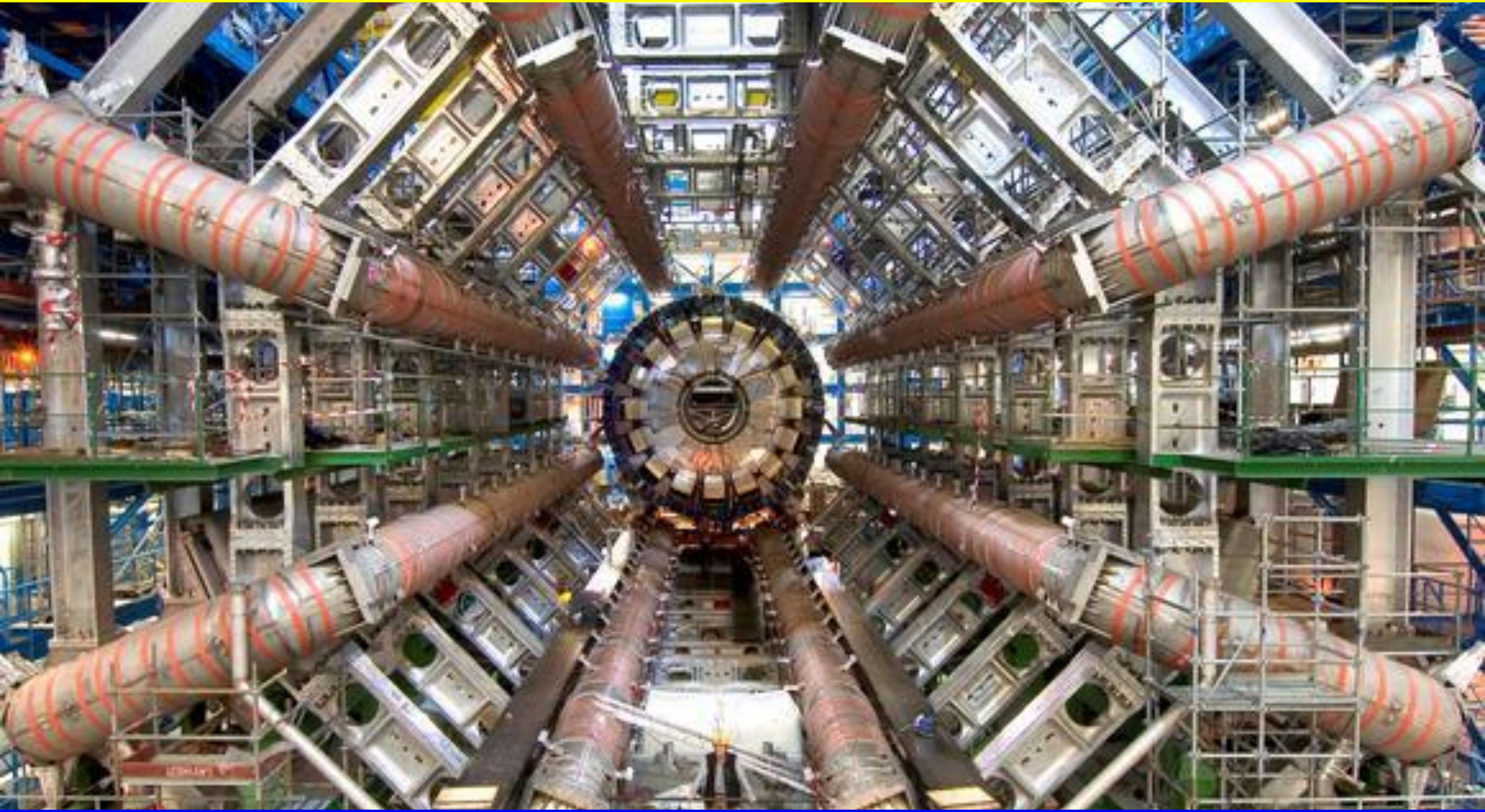
WTF? Discovering the Unexpected

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

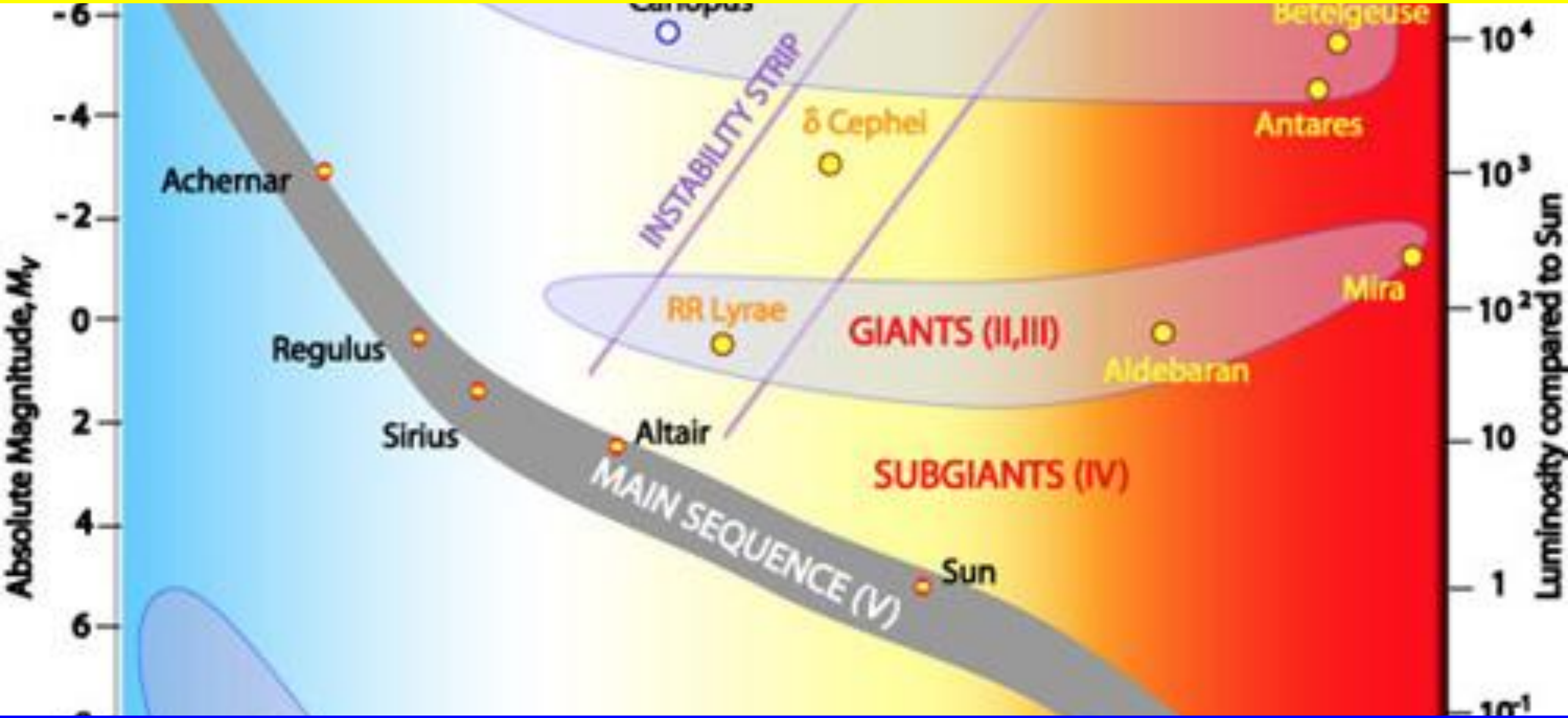
How does science work?



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)

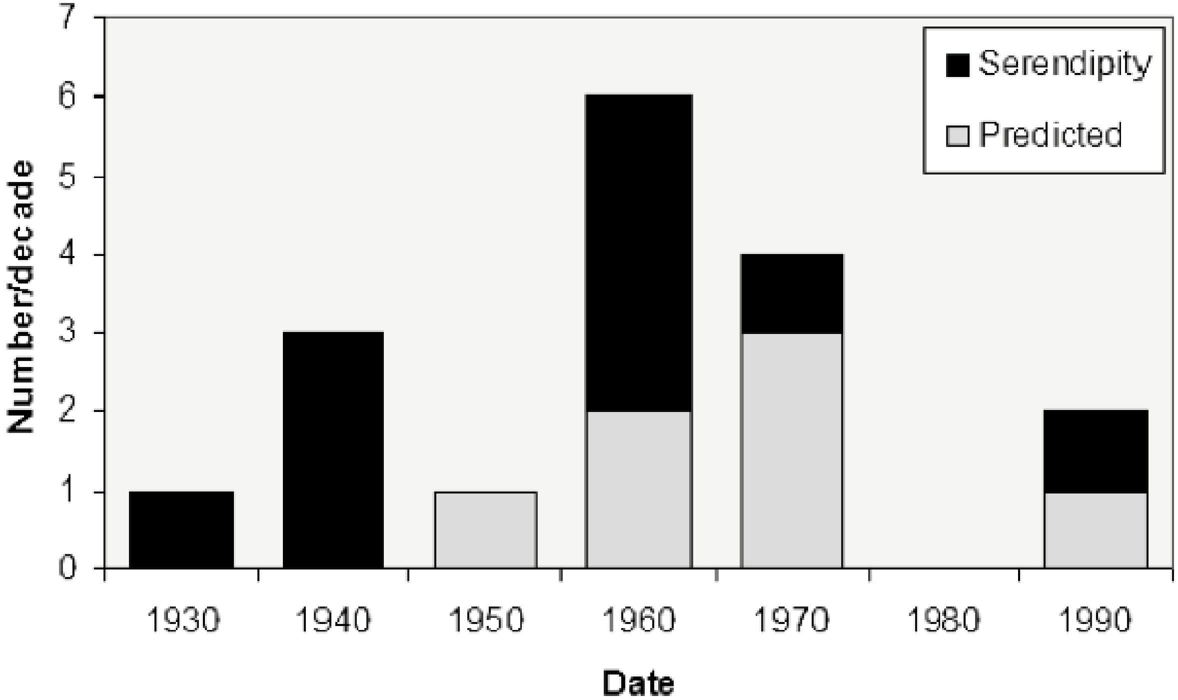


Astronomy usually works more in an “explorer” mode

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”?



(b) Predicted v Serendipity

Serendipity:10

Predicted: 7

From Ekers (2009) PoS(sps5)007

See also:

- *Harwit(1981), Cosmic Discovery*
- *Kellermann(2009) PoS(sps5), 44*
- *Wilkinson et al.(2004), New Astr. Rev., 48, 1551 45*
- *Wilkinson(2007) the Modern Radio Universe, 144*
- *Wilkinson(2015) (AASKA14), 65*

Discoveries with HST

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

from Norris et al. 2013: arXiv1210.7521

Discoveries with HST (see e.g. Lallo: *arXiv:1203.0002)*

Project	Key	Planned?	Nat.	Highly	Nobel prize?
Use Cepheids to					
study intergalac uv spectroscopy					
Medium-deep s					
Image quasar ho					
Measure SMBH					
Exoplanet atmos					
Planetary Nebul					
Discover Dark En					✓
Comet Shoemak					
Deep fields (HDF					
Proplyds in Orion			✓		
GRB Hosts			✓		

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

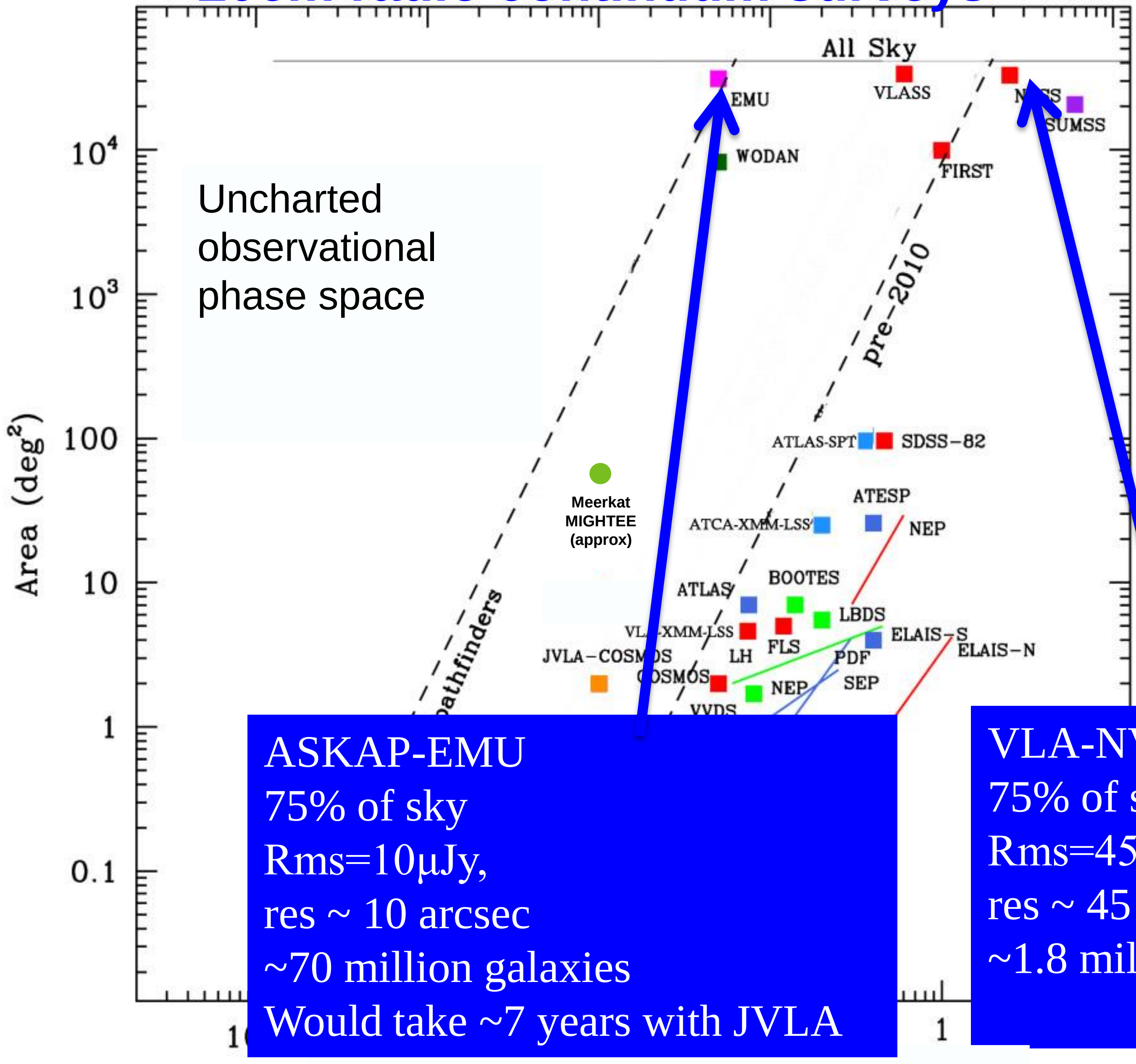
Jocelyn Bell:

- 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

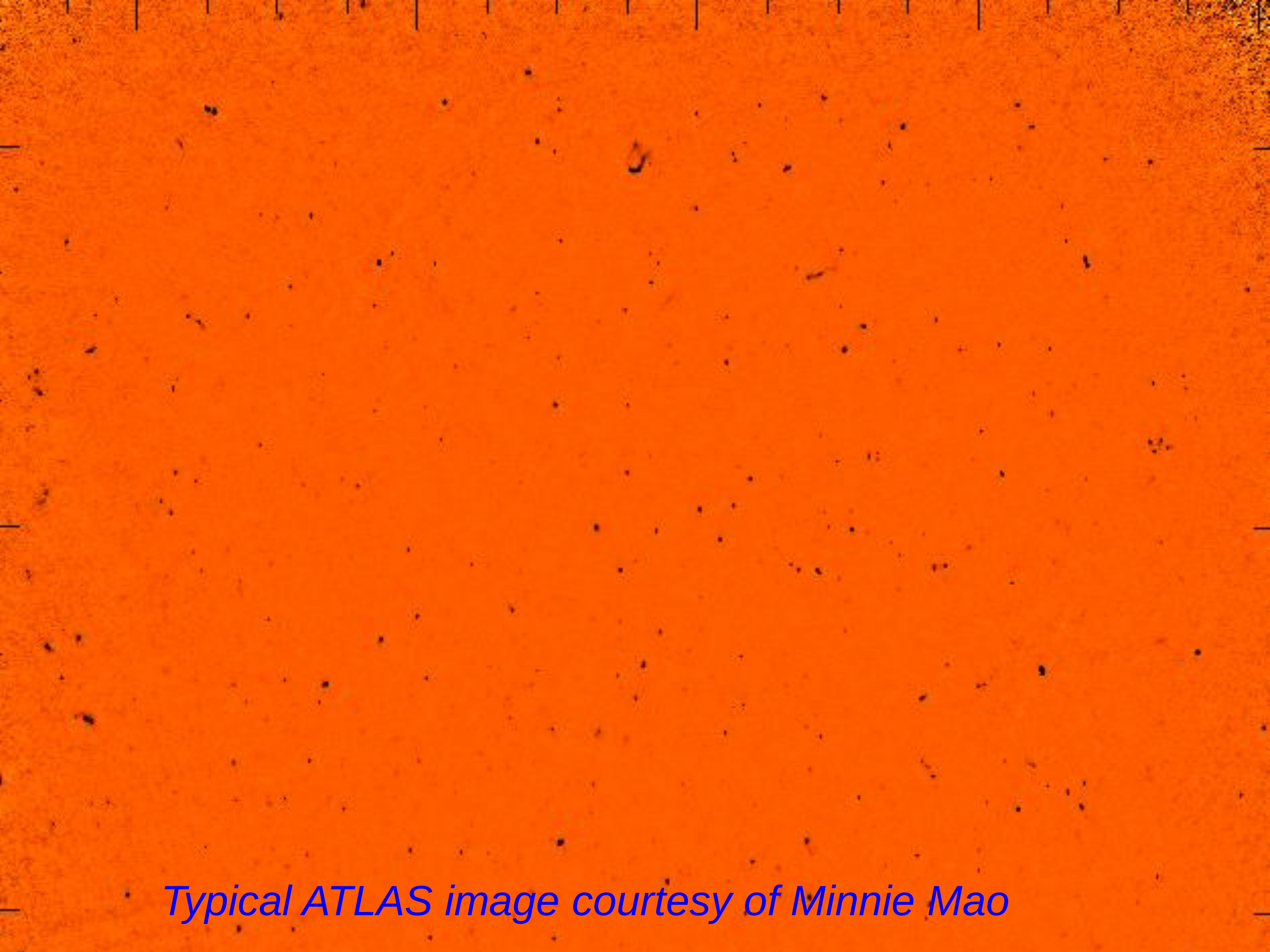


ASKAP-EMU
75% of sky
Rms=10µJy,
res ~ 10 arcsec
~70 million galaxies
Would take ~7 years with JVLA

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

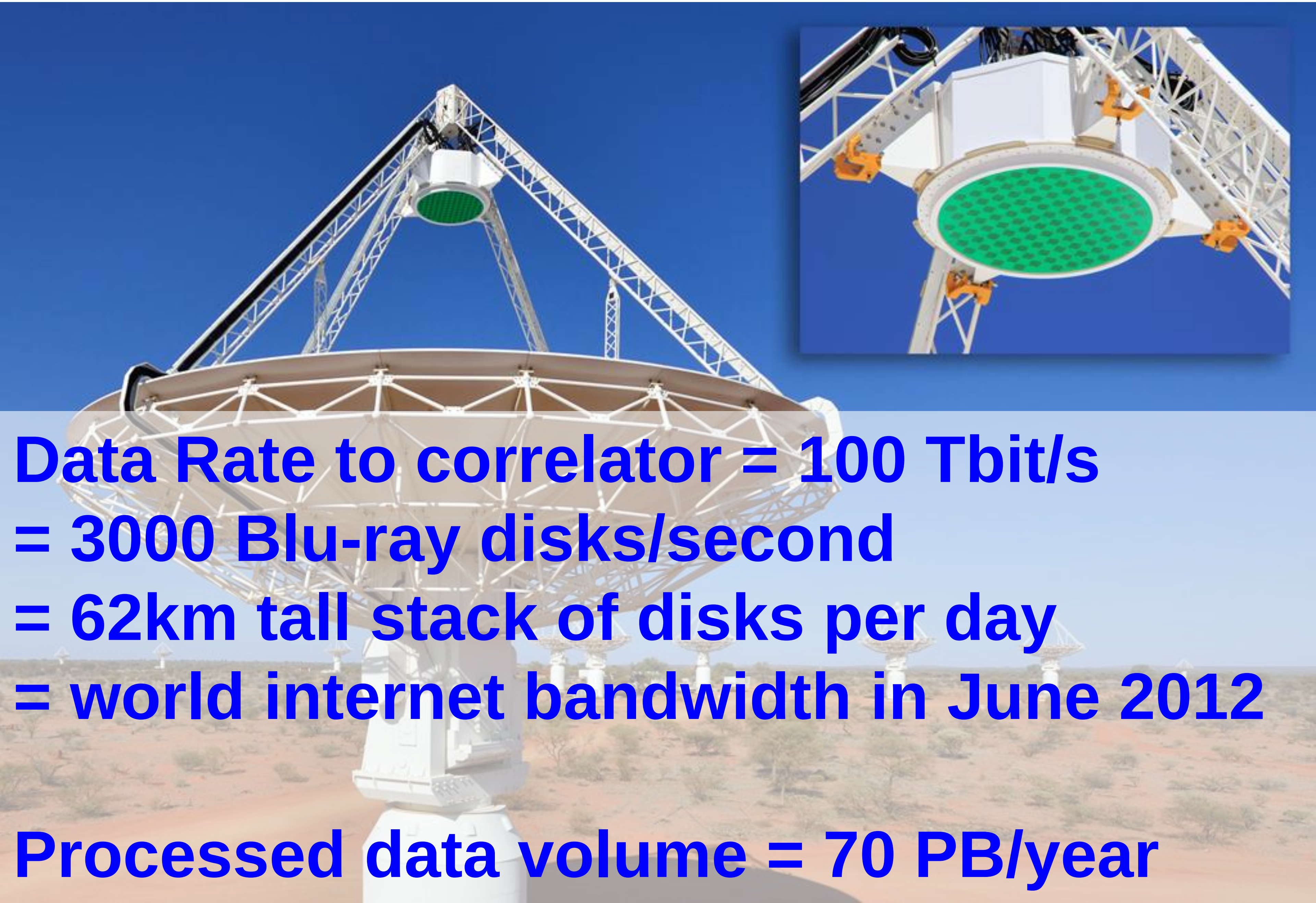
5σ Sensitivity (mJy)

CSIRO



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

CSIRO

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

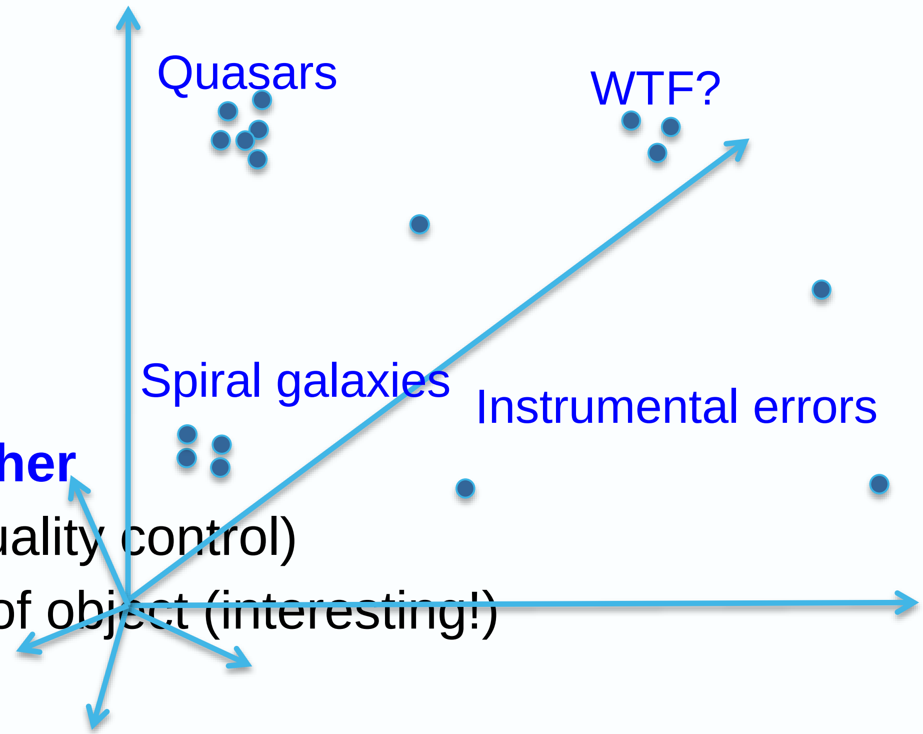
WTF?

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

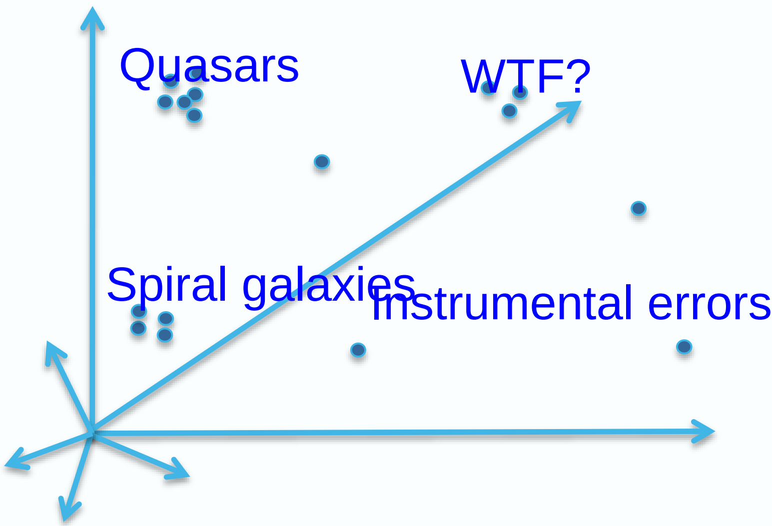


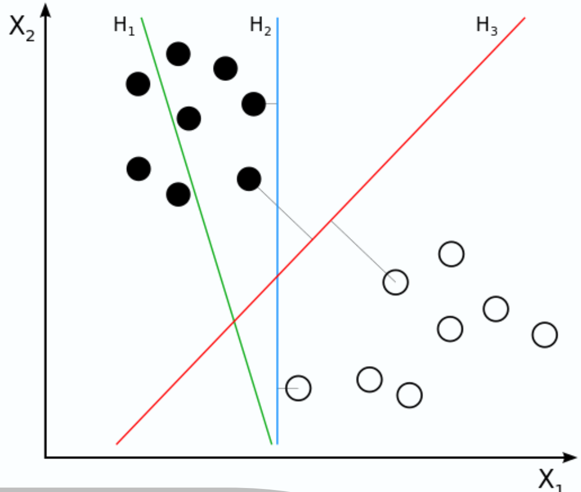
Identified objects/regions will be either

- 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:

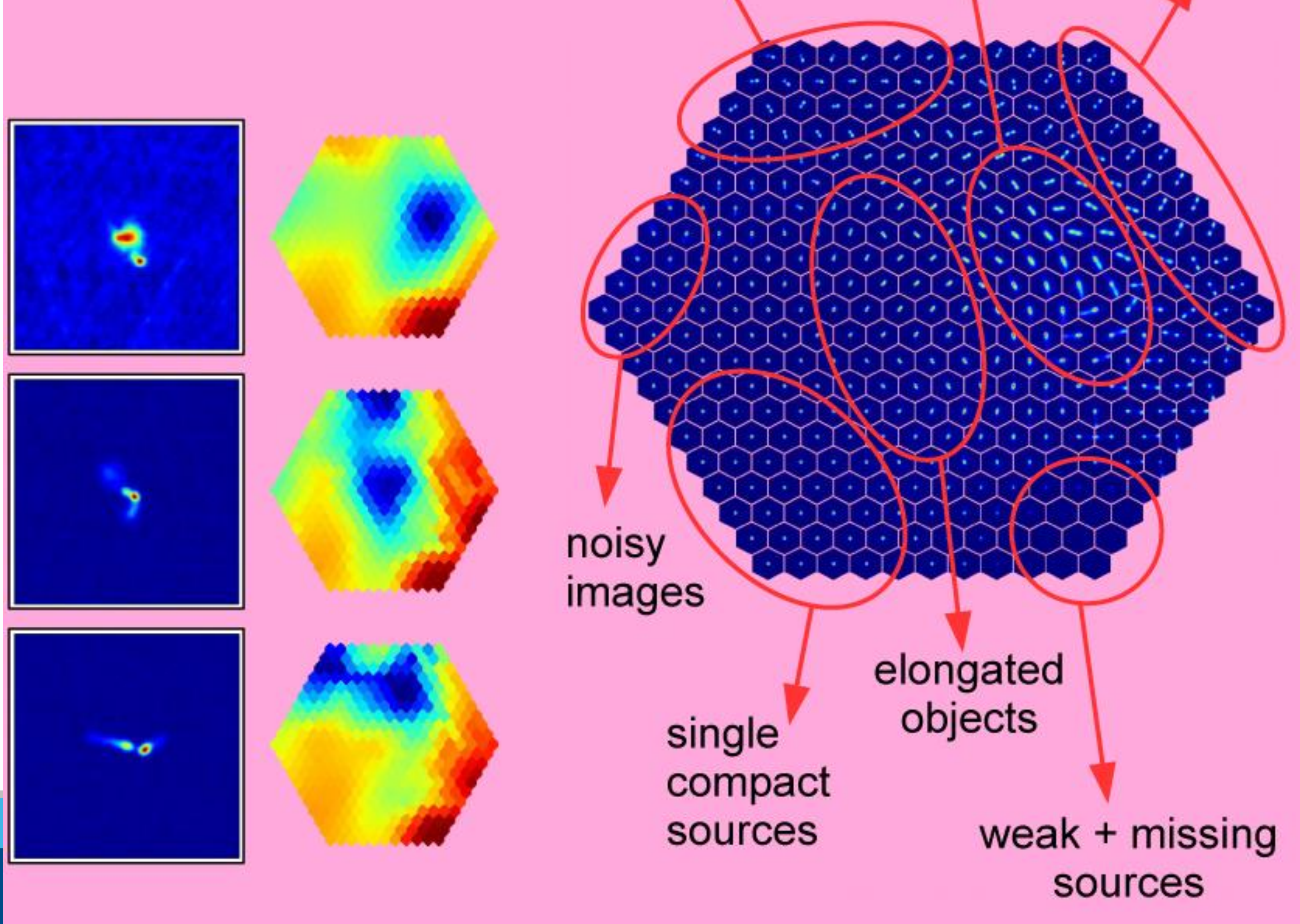
- Finding unexpected objects
  - = finding classes of unclassified objects
  - = finding anomalous objects
- p measurables (E.g. colours/spectral indices/morphologies)
- set up a training set of known types of object.
- Arrange in a phase space
- Are there parts of the phase space which are observable 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?")

Challenge Data

a are hosted on Amazon s3.

unmodified data sets

ATLAS CDFS DR1

- Data description
- FITS image (~3 sq deg, 46 MB)
- component table (Table 4),
- source table (Table6),
- thumbnail images as a .tgz or individu

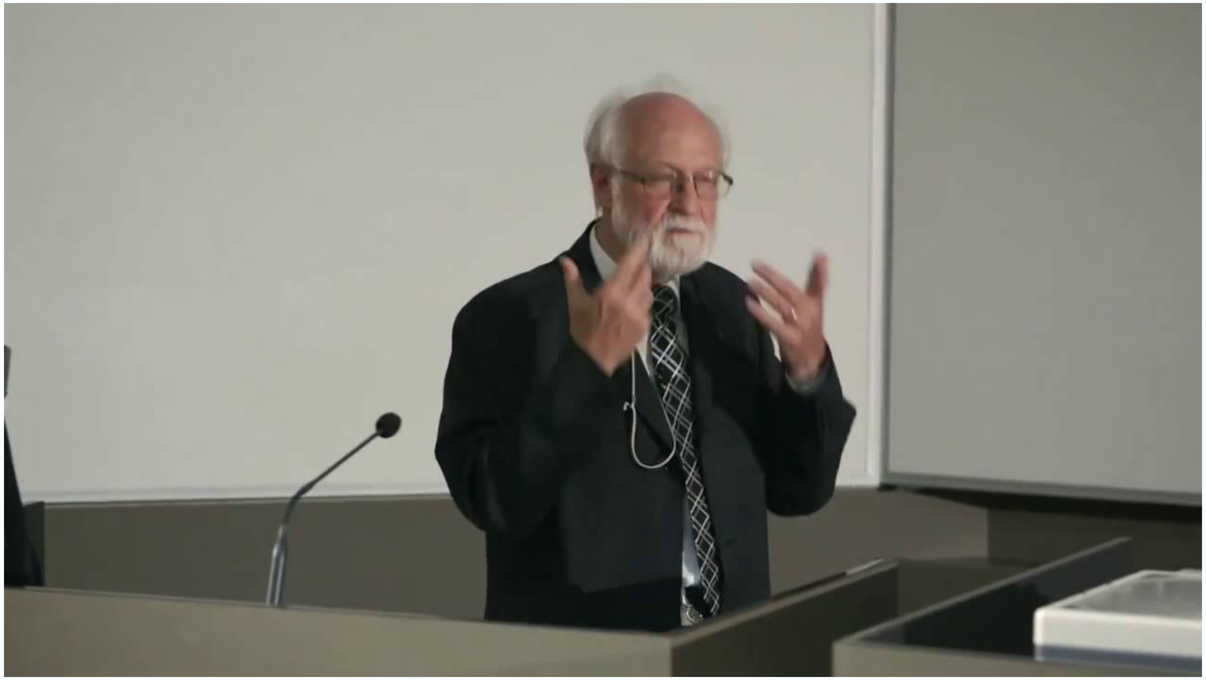
ATLAS ELAIS-S1 DR1

- Data description:
- FITS image (TBD),
- component table (Table 4) (TBD),
- source table (Table6)(TBD),
- thumbnail images as a .tgz or individu

STRIPE 82 in radio (VLA-S82/FIRST/NVSS)

- Data descriptions: VLA-S82 (1.8"), FIRS
- FITS images for all available via Skyvie
- Catalogues: VLA-S82, FIRST, NVSS (FIR
- Thumbnail images: N/A

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.

Challenge Data

a are hosted on Amazon s3.

unmodified data sets

ATLAS CDFS DR1

- Data description
- FITS image (~3 sq deg, 46 MB)
- component table (Table 4),
- source table (Table6),
- thumbnail images as a .tgz or individual files

ATLAS ELAIS-S1 DR1

- Data description:
- FITS image (TBD),
- component table (Table 4) (TBD),
- source table (Table6)(TBD),
- thumbnail images as a .tgz or individual files

STRIPE 82 in radio (VLA-S82/FIRST/NVSS)

- Data descriptions: VLA-S82 (1.8"), FIRST (5"), NVSS (45")
- FITS images for all available via Skyview
- Catalogues: VLA-S82, FIRST, NVSS (FIRST and NVSS also via Vizier)
- Thumbnail images: N/A
- Note: these data may be useful for angular resolution comparison stu

etary unmodified data sets

are made kindly available pre-publication for use for WTF data challenge s
ners, and any publication may be subject to conditions (e.g. authorship req
g on this data for their PhDs. Please don't!

ATLAS-SPT

- Data Description
- 9 large fits image tiles (9 files, each ~1.4 GB)
- 32x32 tile HiPS survey,
- 64x64 tile HiPS survey

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)

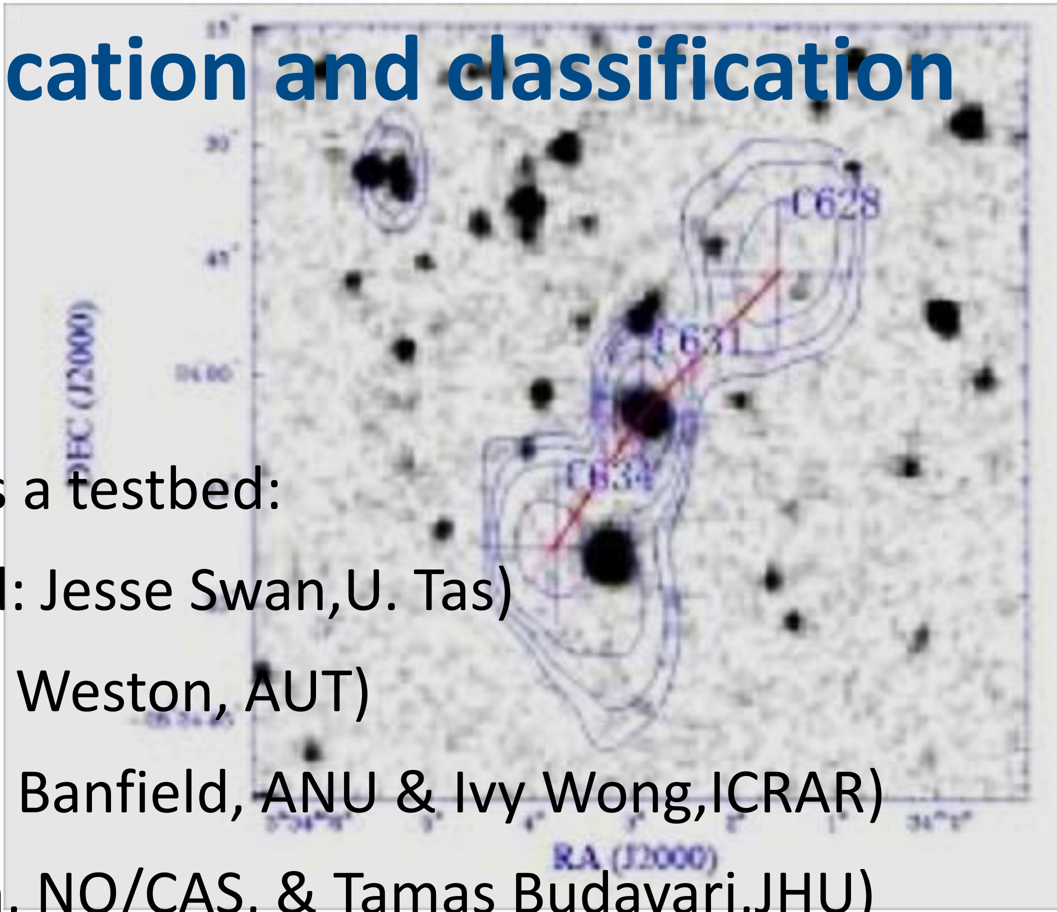
1) Build up group with local expertise

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%

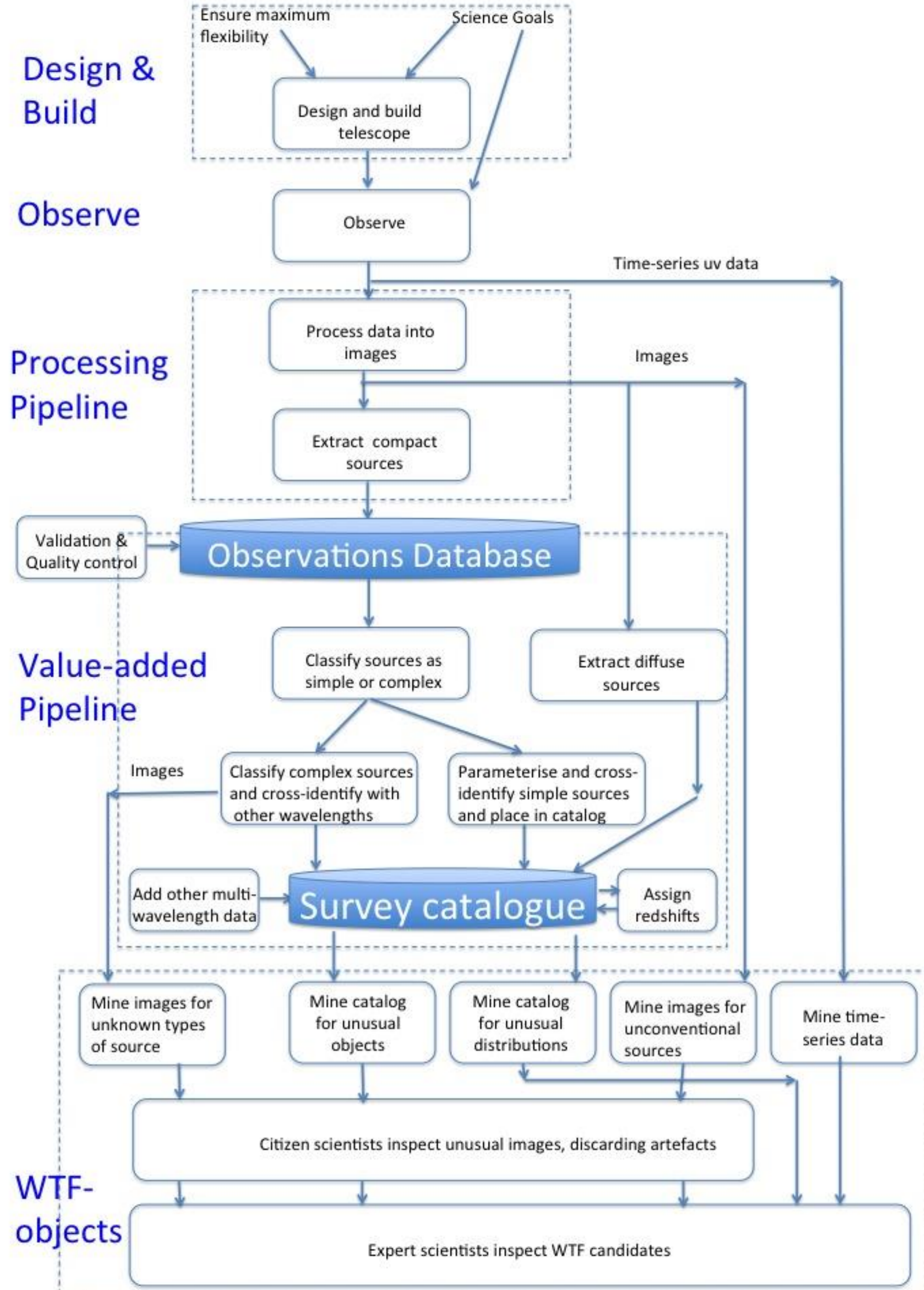
Current projects using ATLAS as a testbed:

- 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

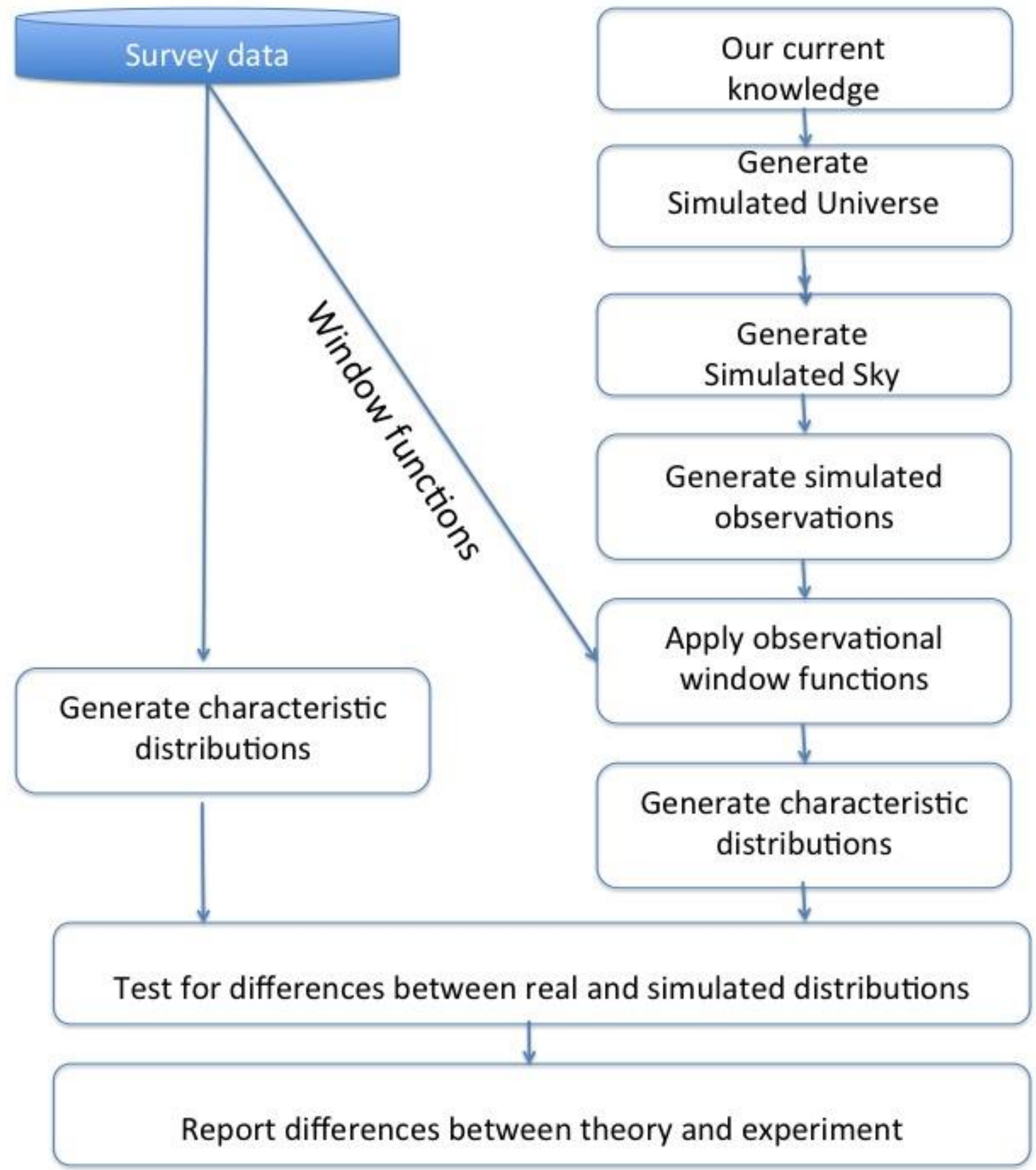
WTF Phase 2 (2016-7)

- 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



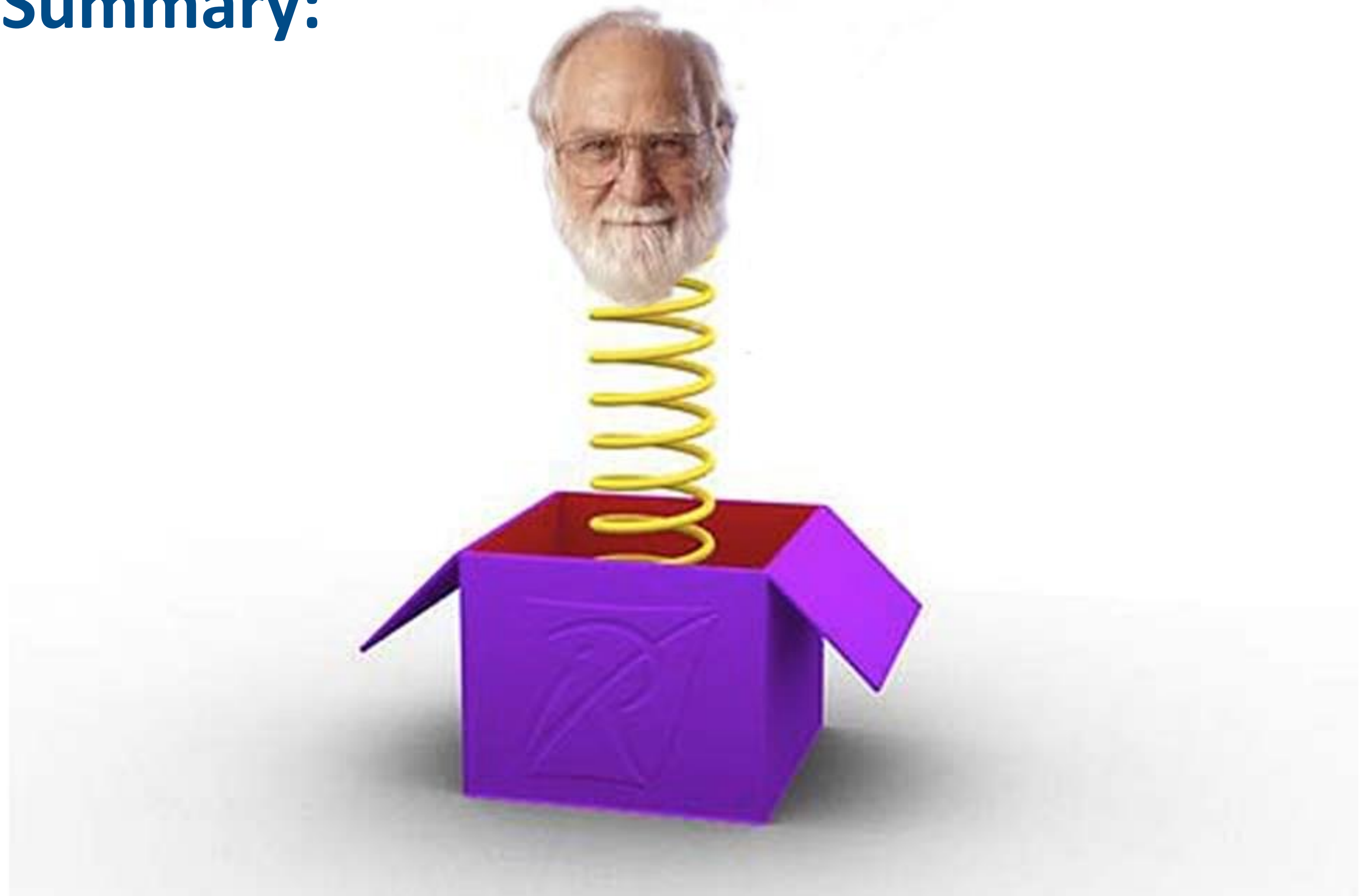
Flowchart for discovering unexpected phenomena



WTF Phase 3: Re-start WTF challenge

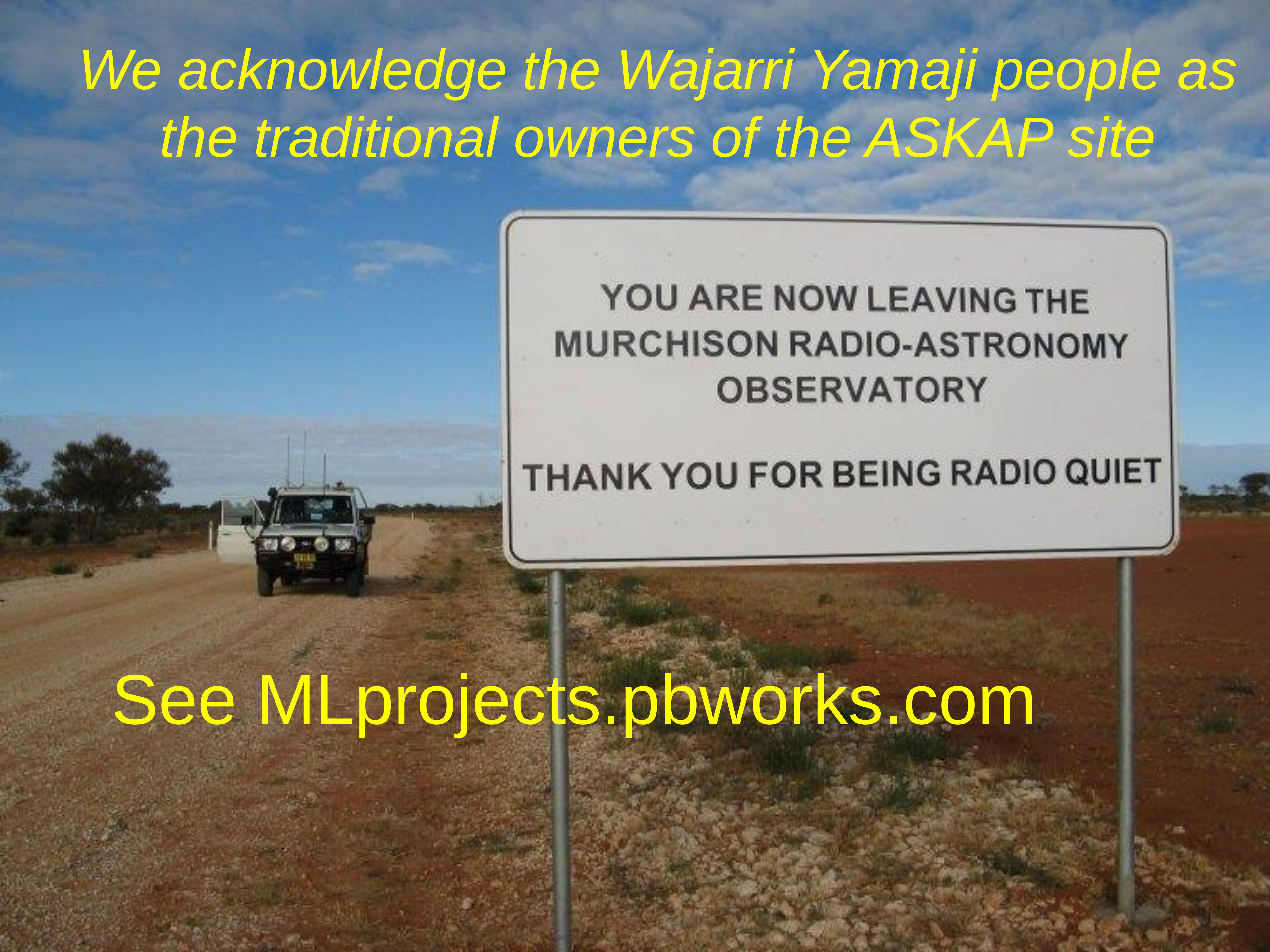
- Set up data sets for challenge using EMU data
- Include both image and tables, including multiwavelength data
- Include well-documented:
  - Training sets
  - Simulated discovery sets
  - Real EMU data
- Focus on in-house research
- Also invite other groups to beat us

Summary:



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



See MLprojects.pbworks.com