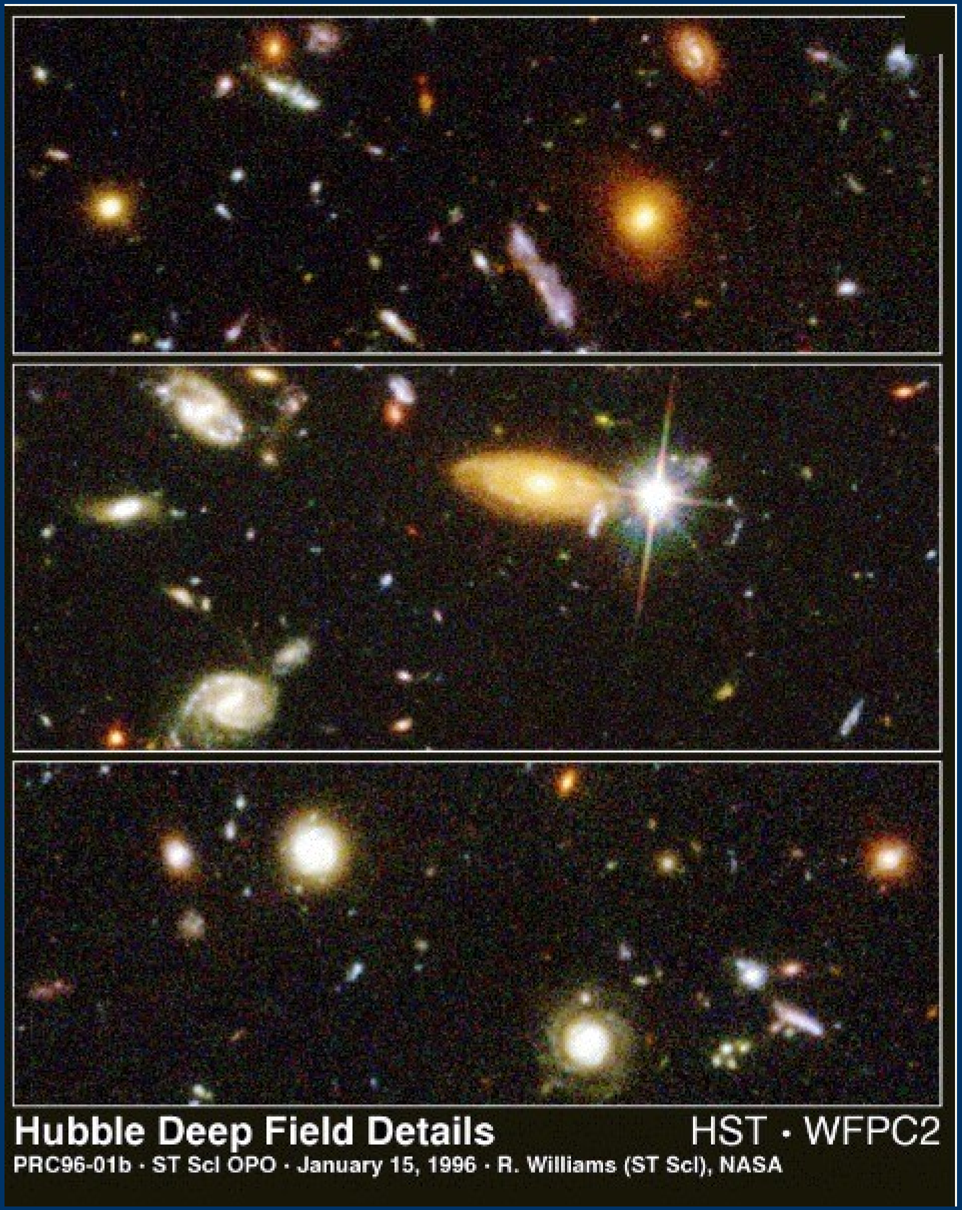
Space in Cyberspace: hidden patterns in astrophysical datasets

Aleksandra Solarz
National Centre for Nuclear Research, Poland
with M. Bilicki, M. Gromadzki, A. Pollo

30.06.2017
EWASS, Prague

Digital Sky Surveys

- As large and as deep as possible
- Sky surveys designed to provide statistical samples of celestial objects.
- Spatial overview, completeness, homogeneous datasets;
- Base for general conclusions about objects;
- Rare and/or unusual objects;

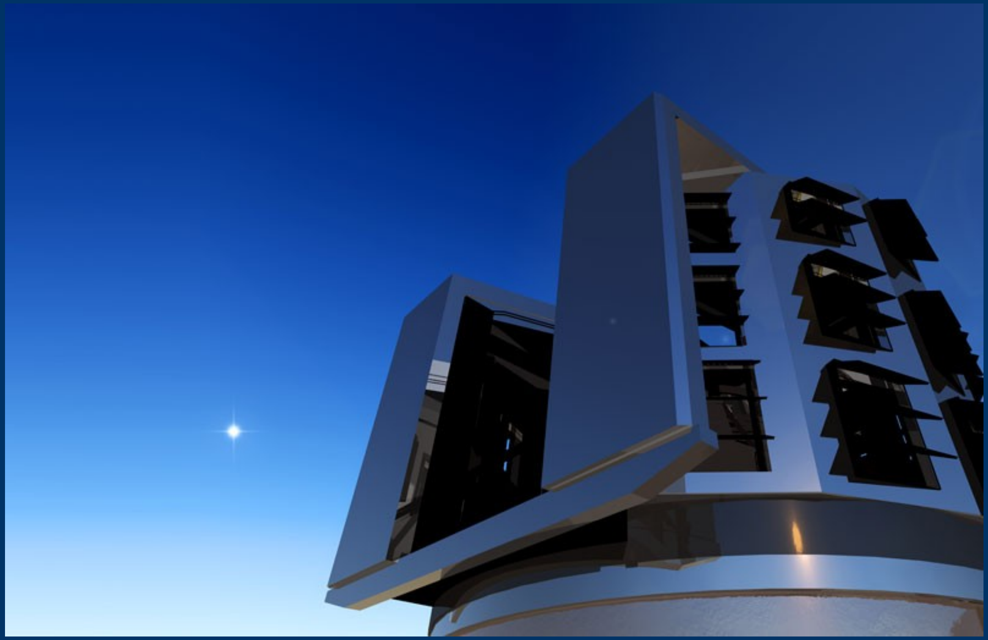


Data avalanche

- SDSS: ~115 TB in total
- Zwicky Transient Facility (ZTF; start 2017)

  1 PB of image data ~1 billion objects
- Large Synoptic Survey Telescope (LSST; first light ~2020); 30 TB PER NIGHT
- The Square Kilometer Array (SKA) ~4.6 Zetabytes

- **Need of automated tools to detect, characterize and classify gathered information**



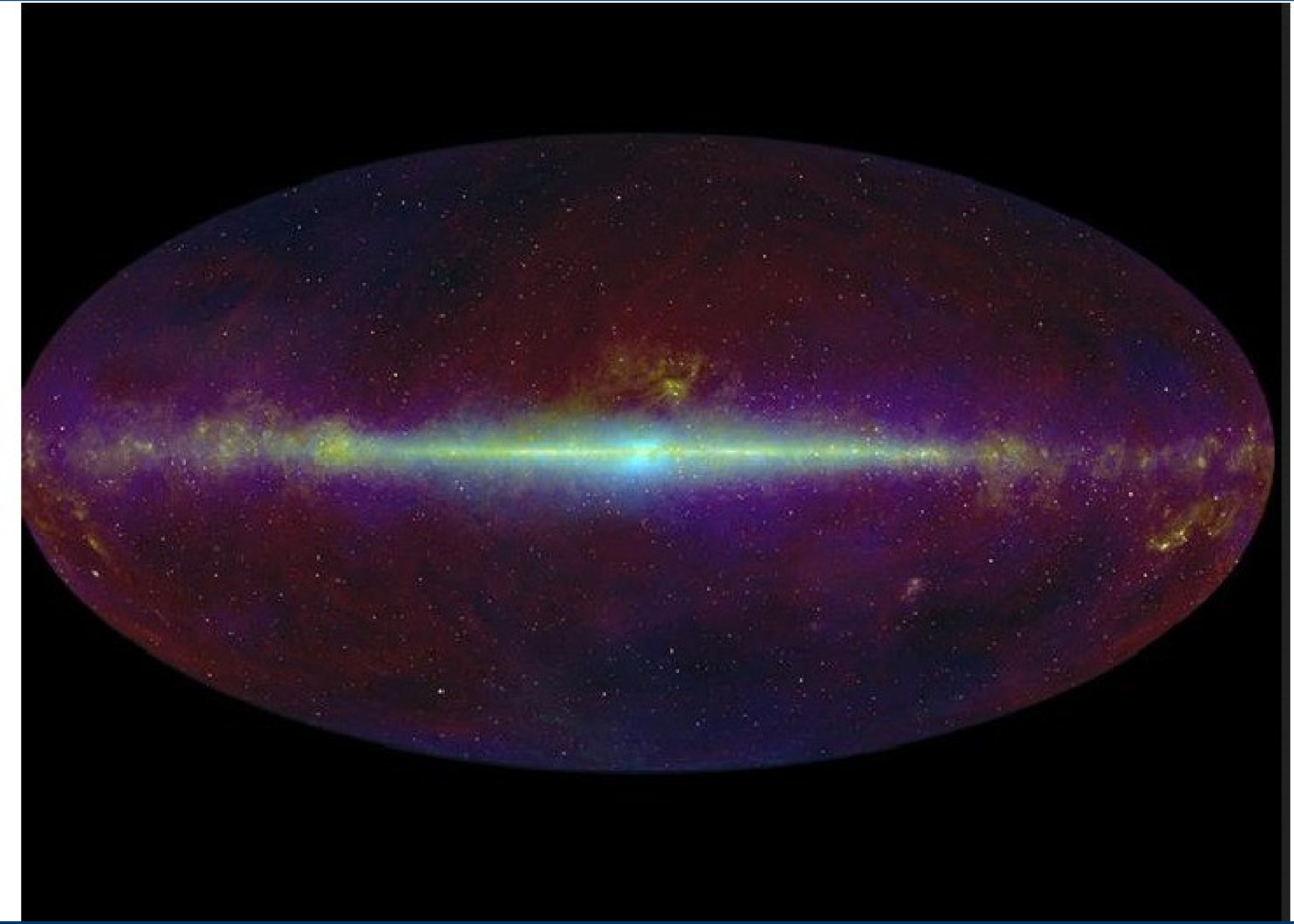
https://www.lsst.org/lsst

SKA; South Africa

The deepest and widest so far: Wide-field Infrared Survey Explorer (WISE)

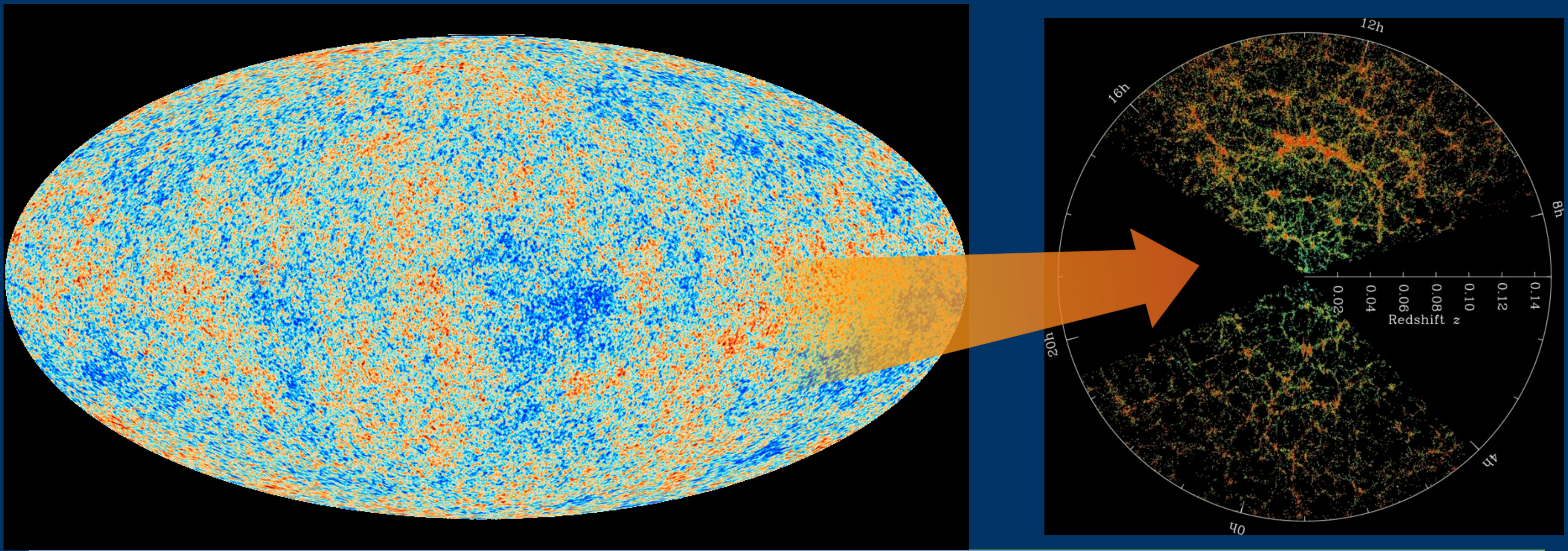
- All-Sky survey in IR
- Detected over 747 mln sources

  (15 PB of data; tables + images)
- Publicly available (position, photometry in 4 bands (3.6-22 um))
- Low angular resolution (~6")
- No redshift information so far



(http://wise2.ipac.caltech.edu/docs/release/allsky/)

Objectives



- Create as complete and as deep catalogues of stars, galaxies and quasars as possible (with as little effort as possible) to get a better understanding of the formation and evolution of the Universe

- **WISE: largest and deepest → perfect for testing efficient methods of fast and effective catalogue creation for further studies**

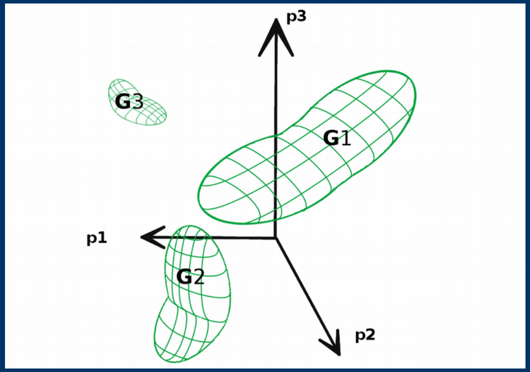
Exploration of parameter spaces

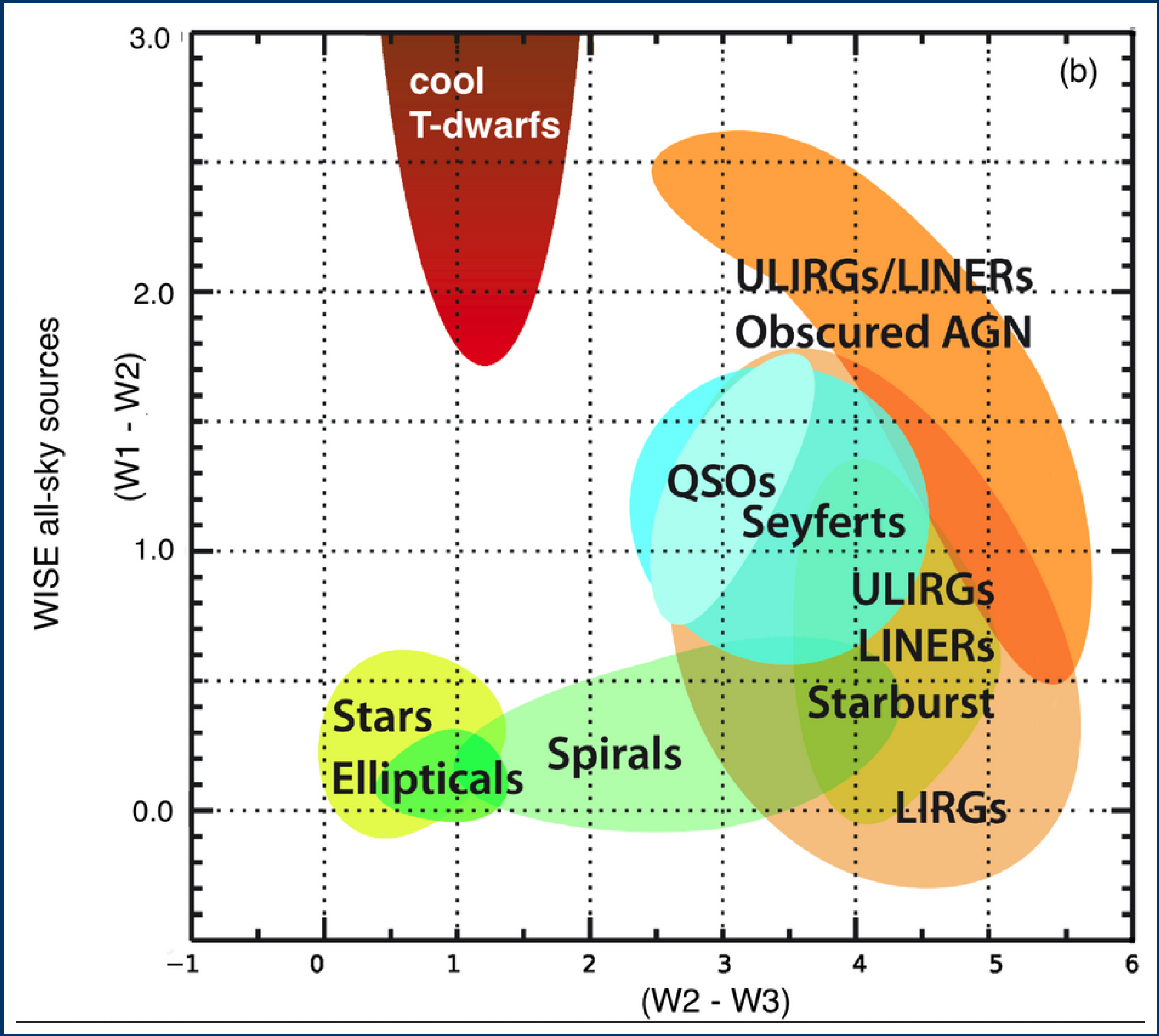
The usual approach to selection of desired sources: CC diagrams

BUT! **With simple approach much information is lost/unseen by human eye**

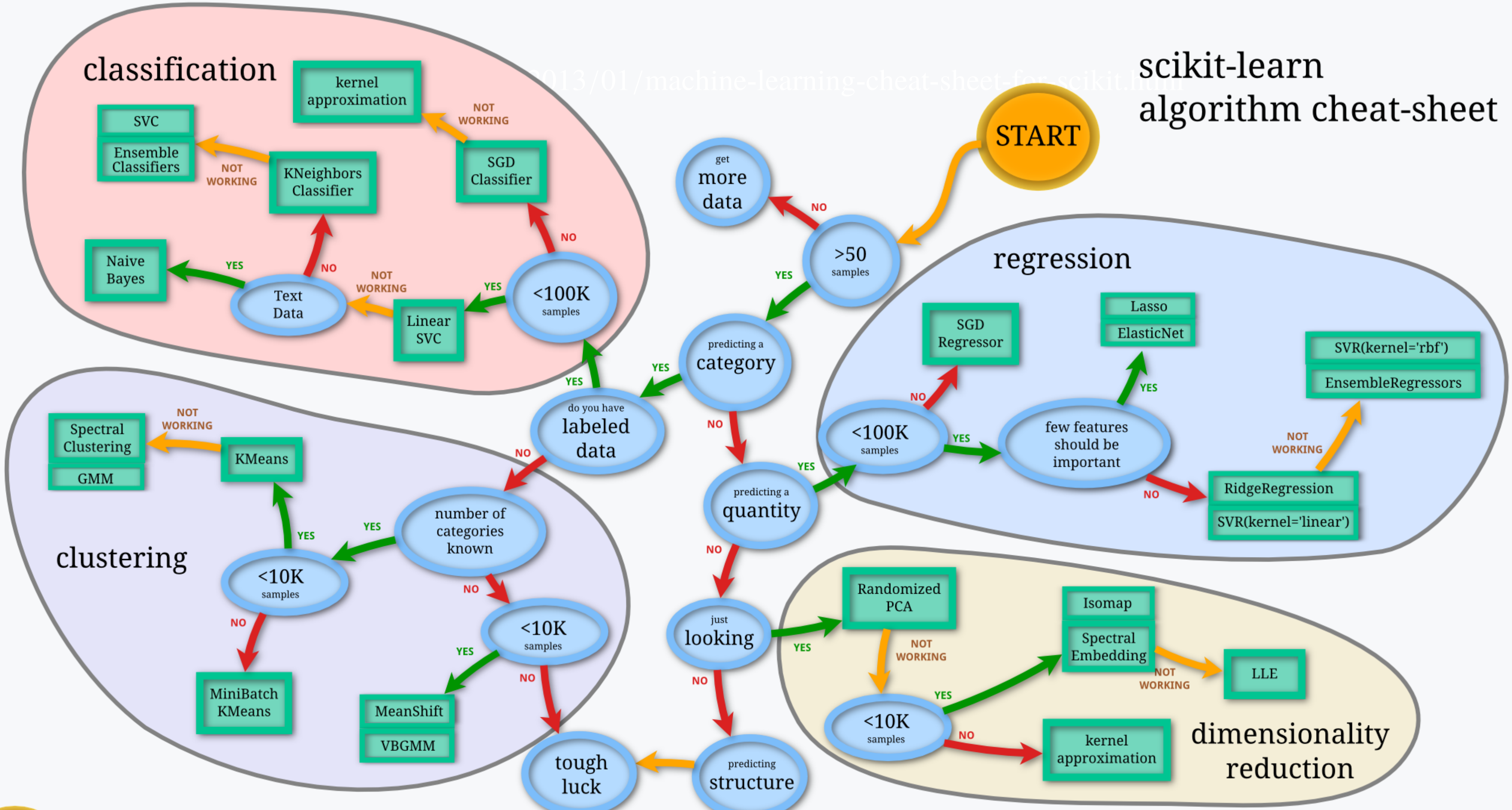
- **A computer can be more precise and deal with a lot of data at once; not restricted to three dimensions**

→ Machine learning!





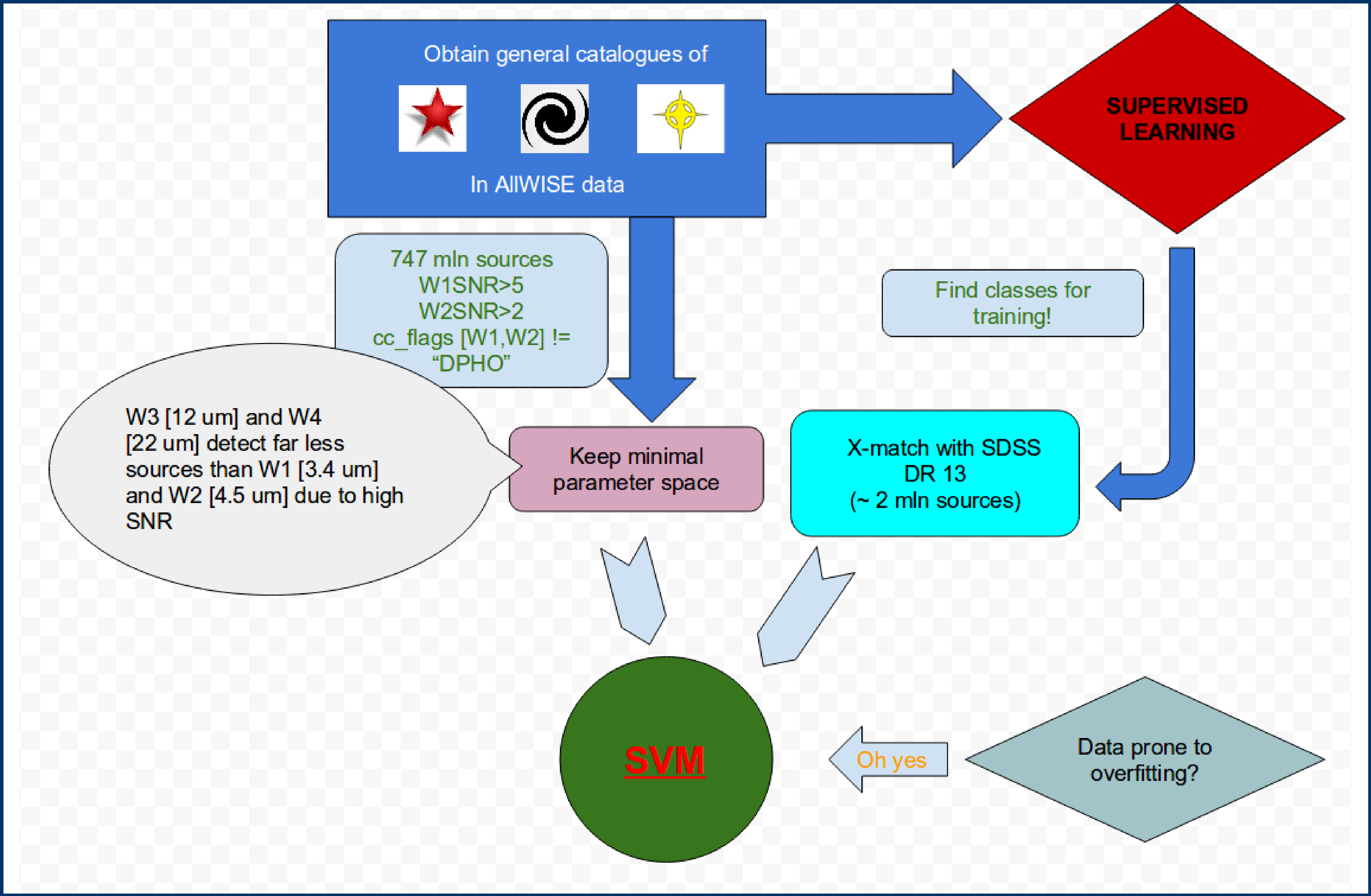
Best algorithm?



Back

scikit learn

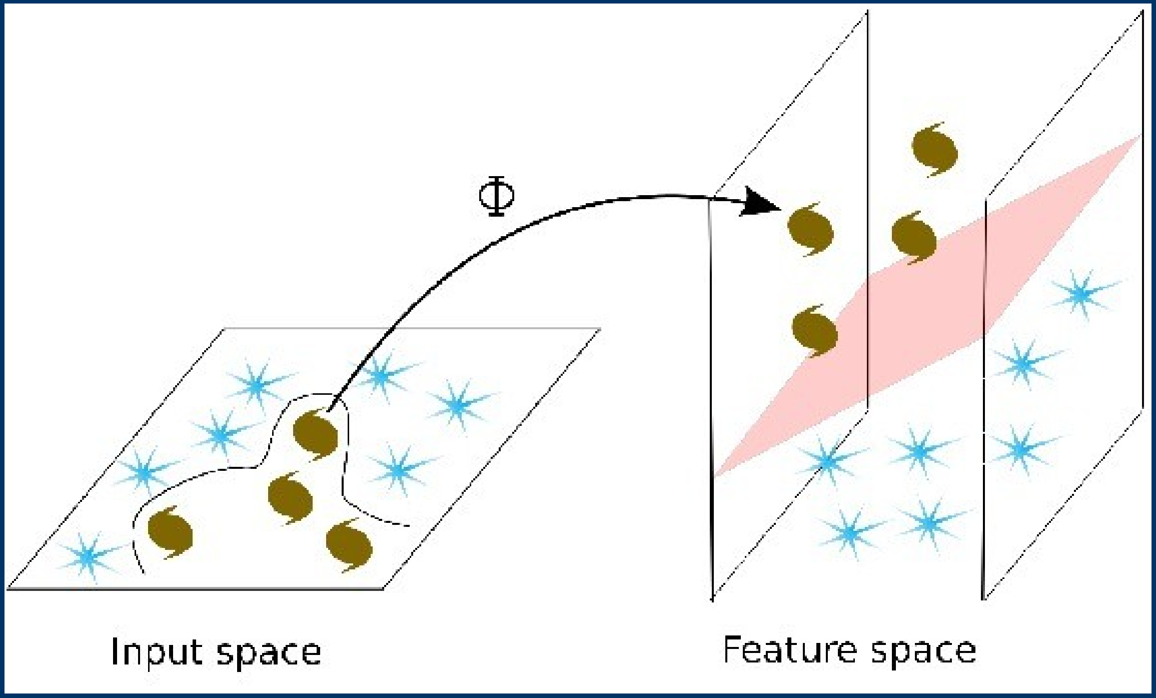
Best algorithm for WISE?

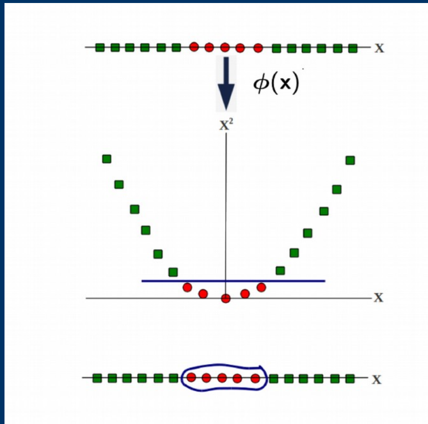


Support Vector Machines (SVM): a supervised approach

SVM: segregate data into 2 (or more) categories based on training examples

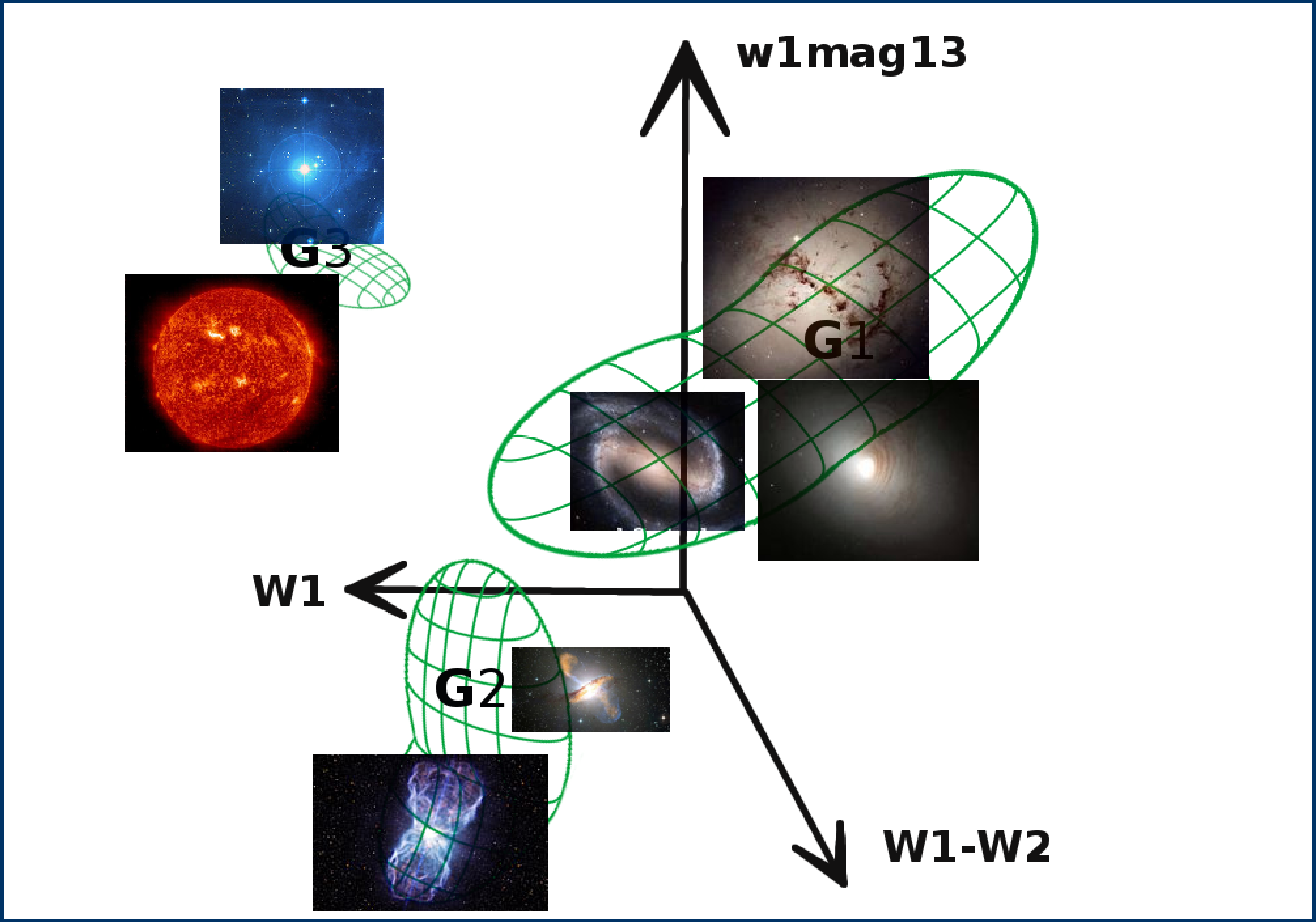
- Use kernel functions to map input data into higher dimensional feature space
- Find a hyperplane separating two classes in the feature space
- New data: class assigned based on their relative position from the boundary

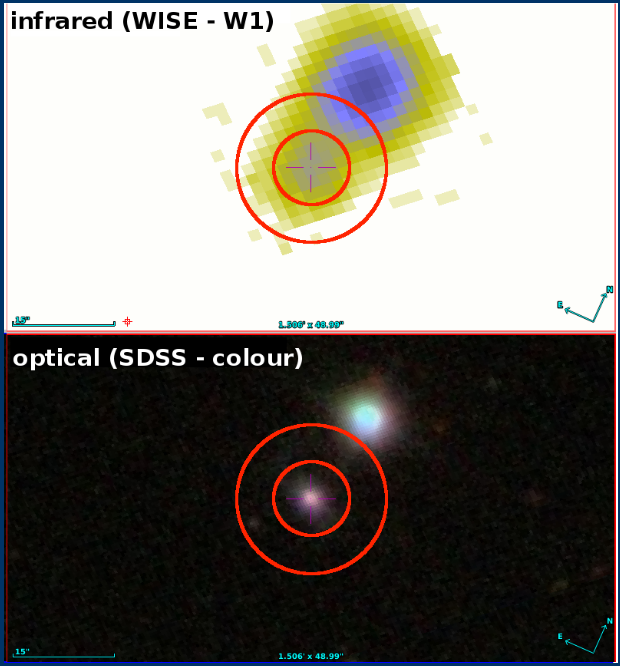




WISE: first attempt at source classification

AllWISE x SDSS (α,δ) parameter space: W1, W1-W2, w1mag13



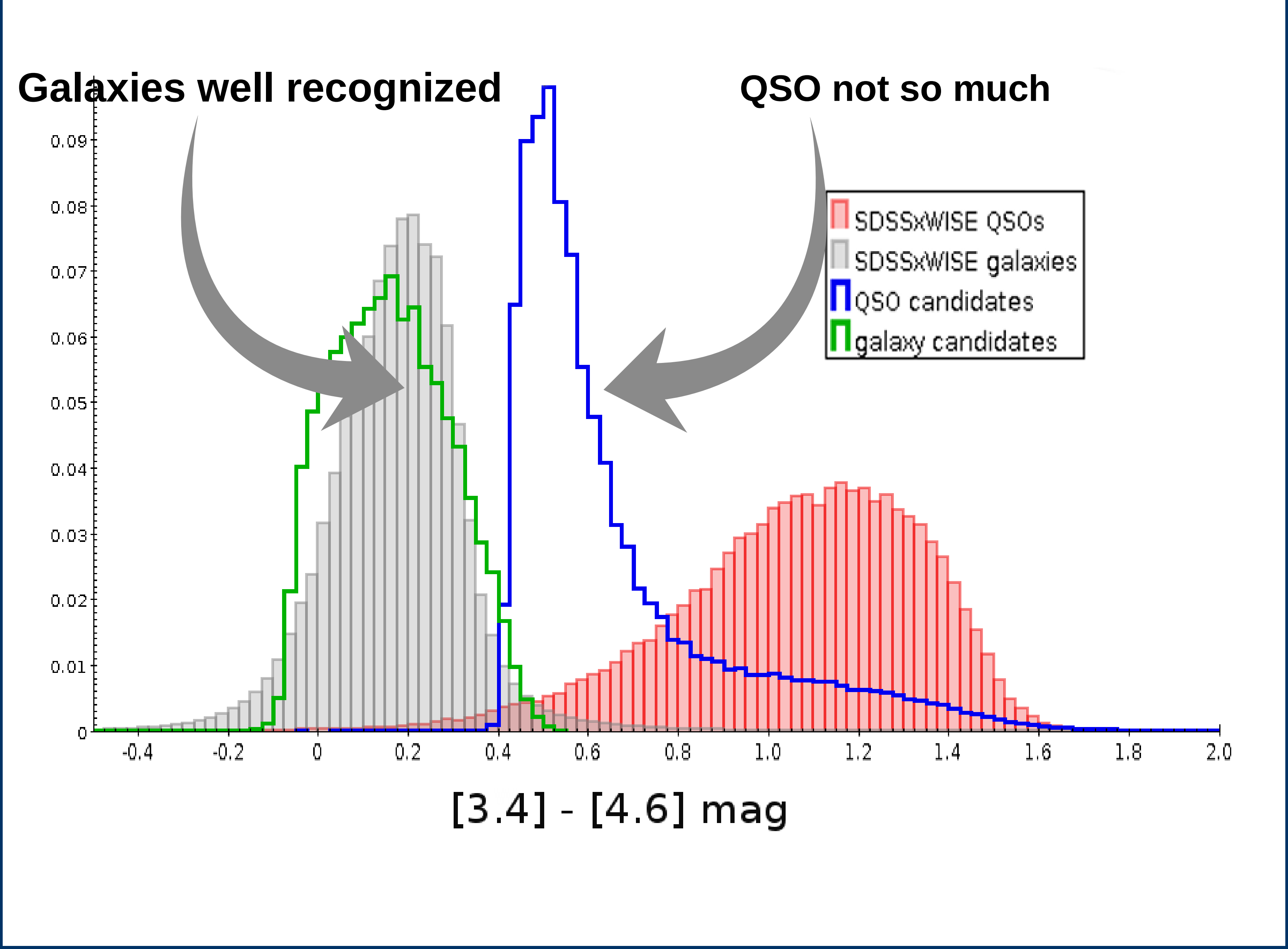


W1mag13 =

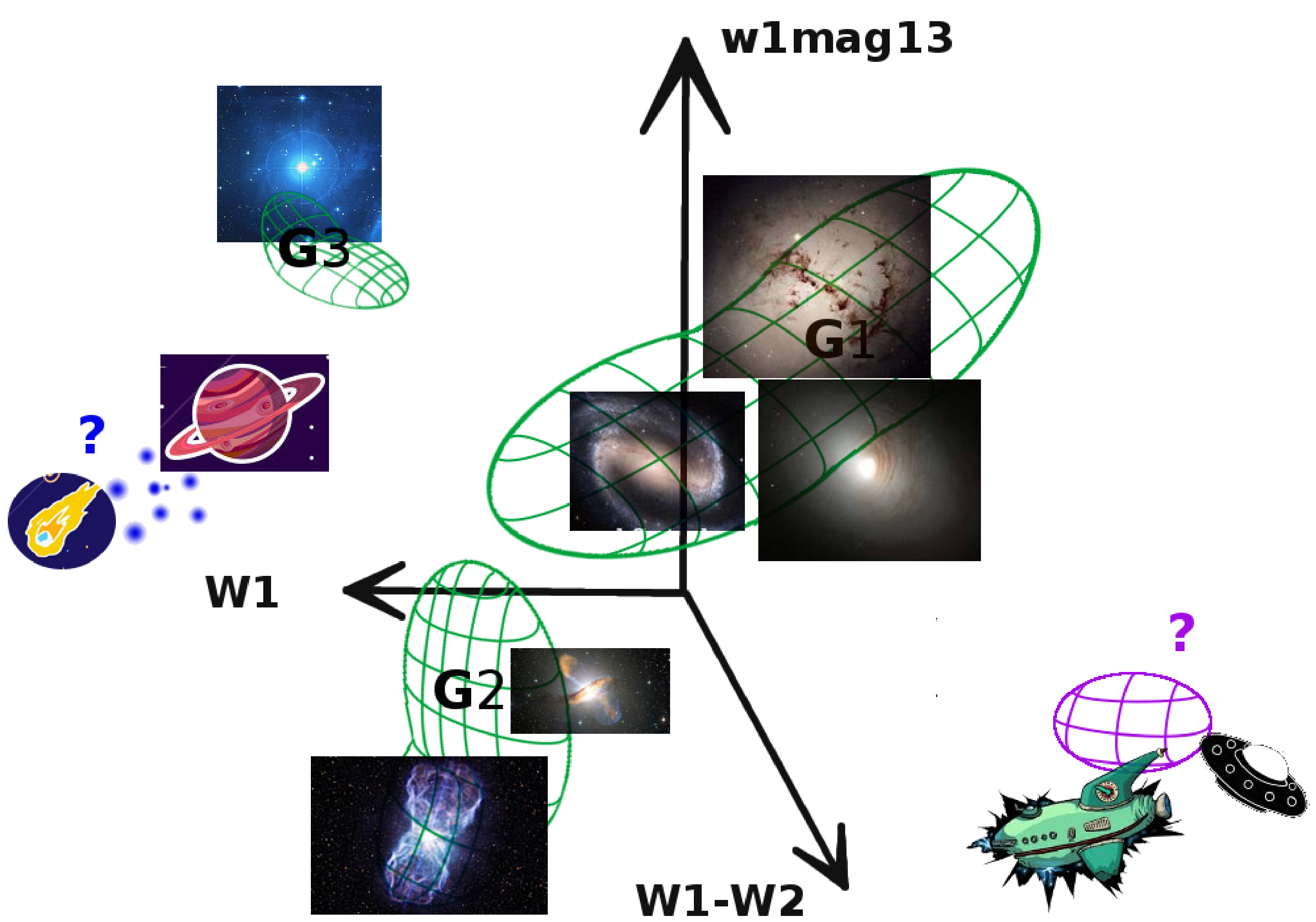
= w1mpro(5”) -w1mpro(11”)

Compactness parameter

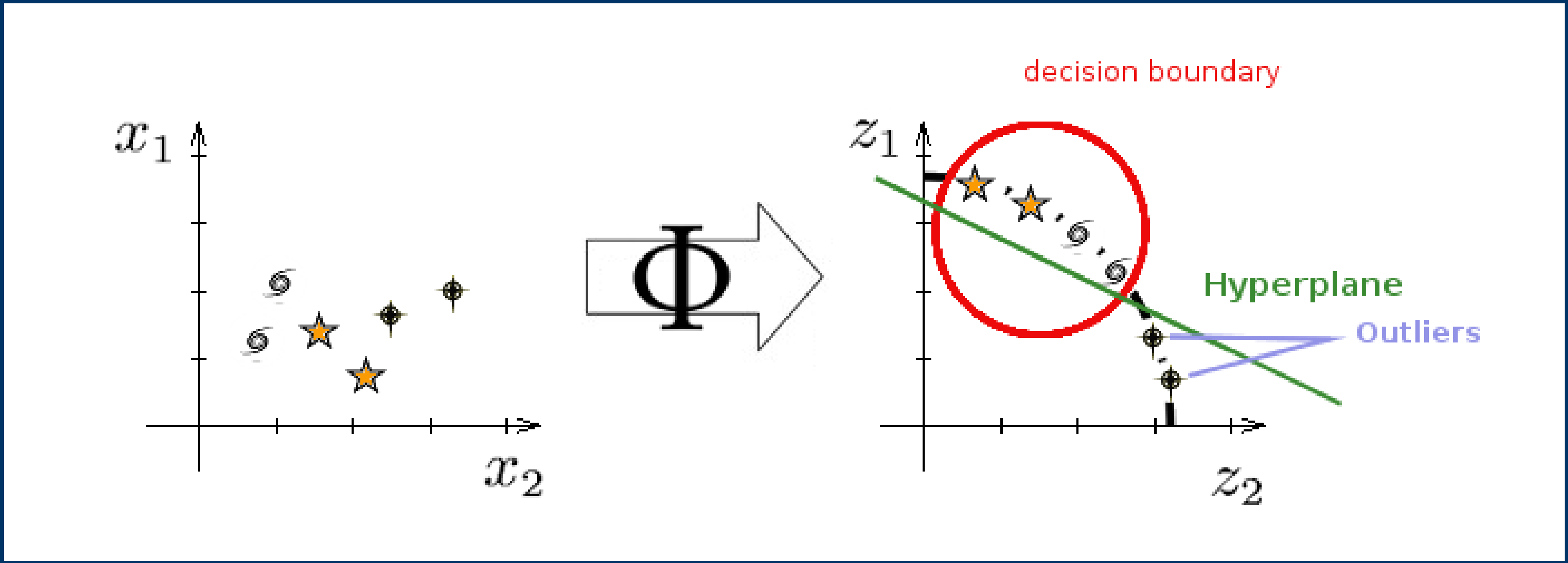
WISE: first attempt at source classification



WISE: what caused the algorithm to fail

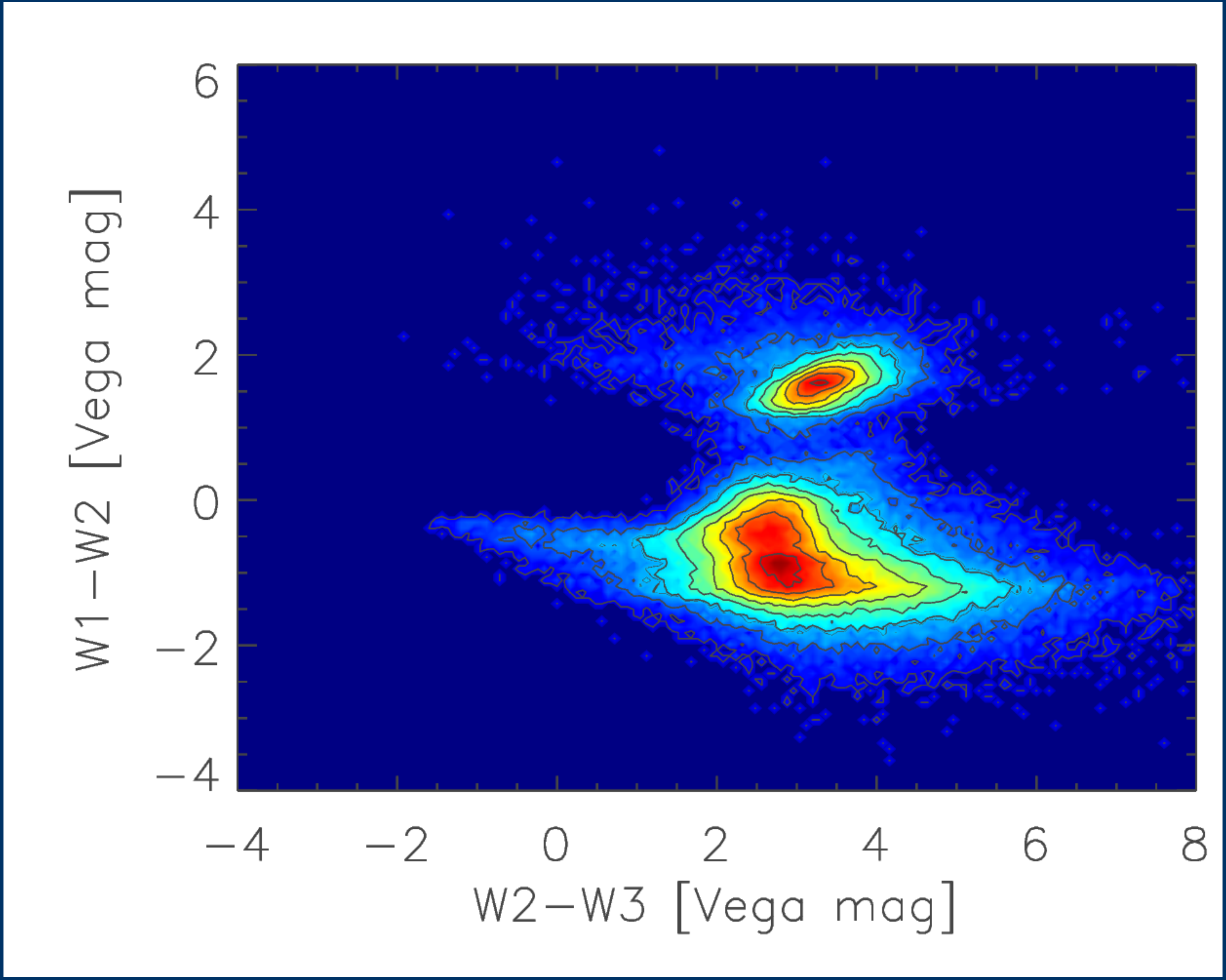


One-class SVM enhancement



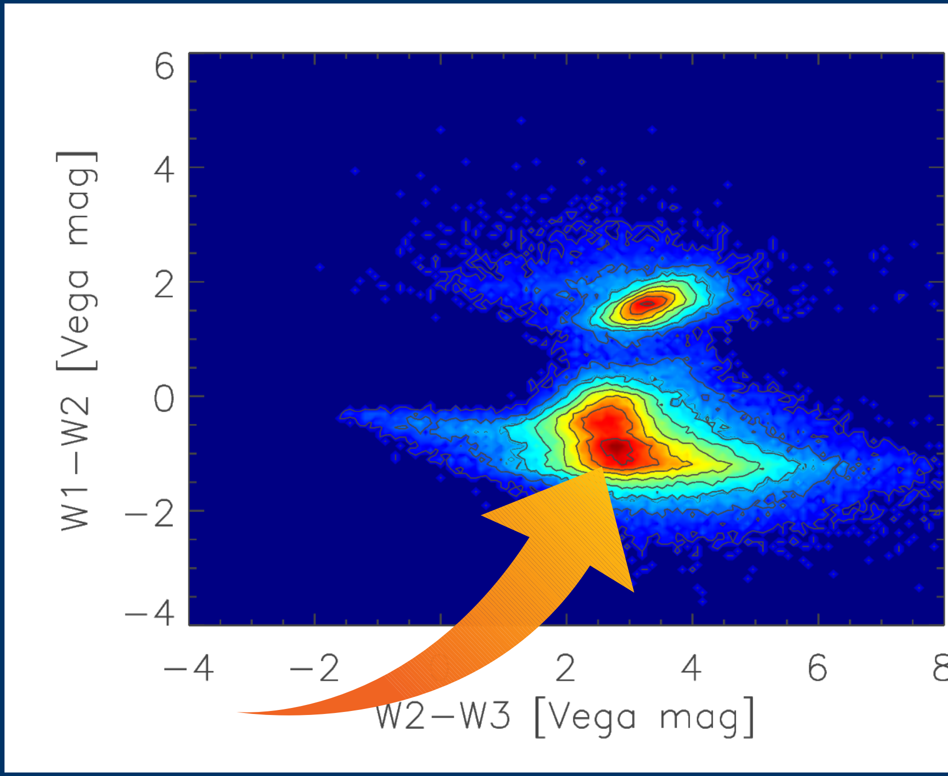
- Create one ‘known’ class (mix of AllWISE x SDSS galaxies, stars, QSOs)
- Hypersurface hugging the expected sources
- Anything with ‘unknown’ patterns falls outside the hypersurface => anomalies

Results
~650,000 anomalous sources

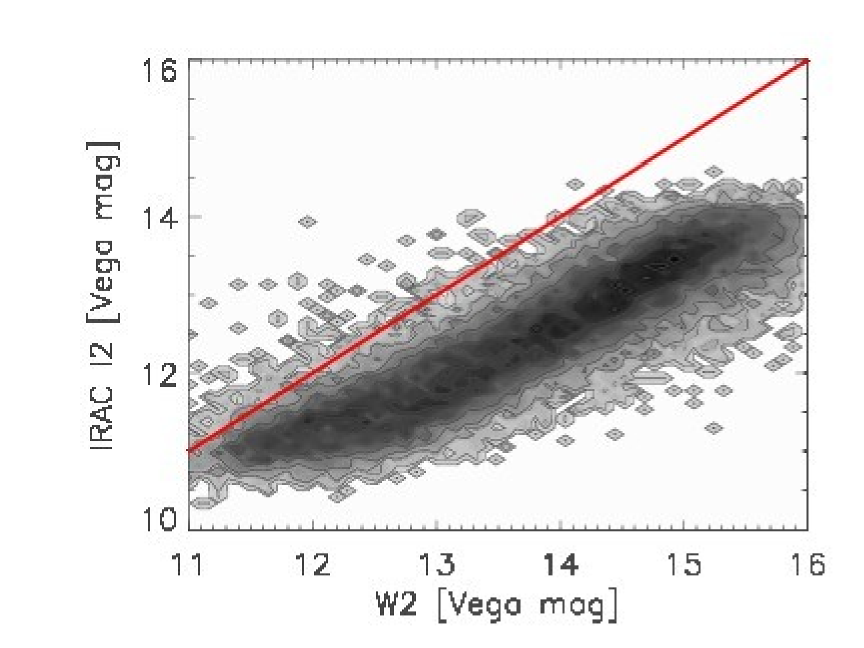


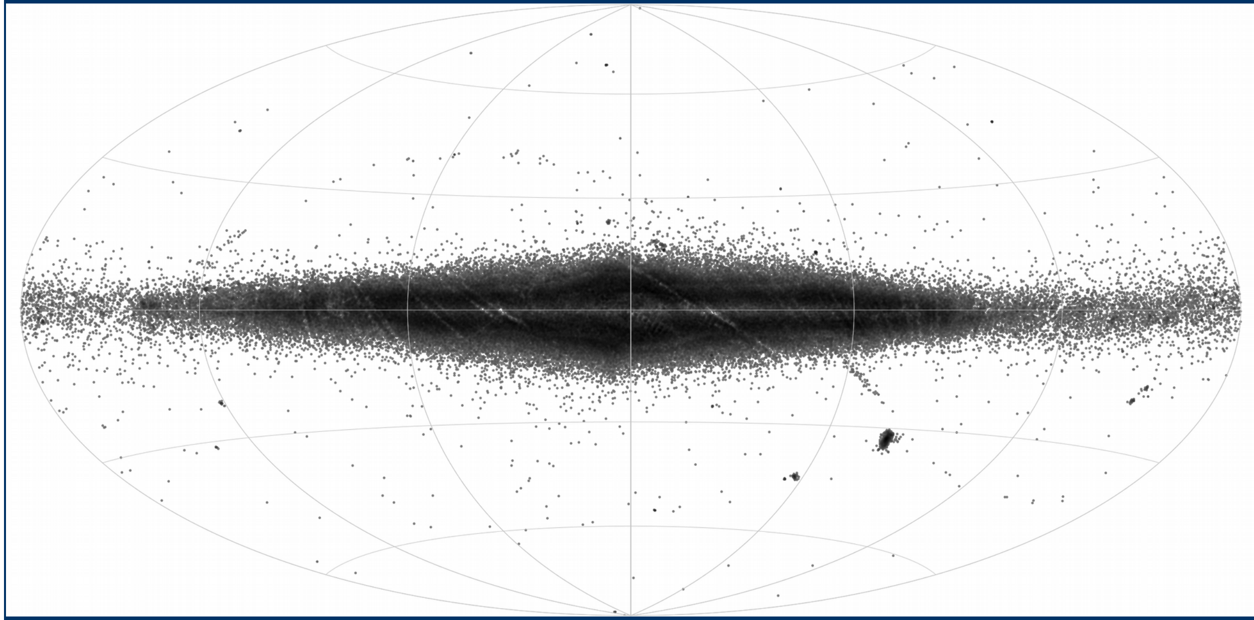
Spurious sources

- W1-W2 ~ -1 ; 80%
- Spitzer GLIMPSE:
  IRAC I1 [3.6 um], IRAC I2 [4.5 um]
- Low WISE resolution (6”)
  in crowded fields => blends
- OCSVM: good tool for selecting hidden artefacts

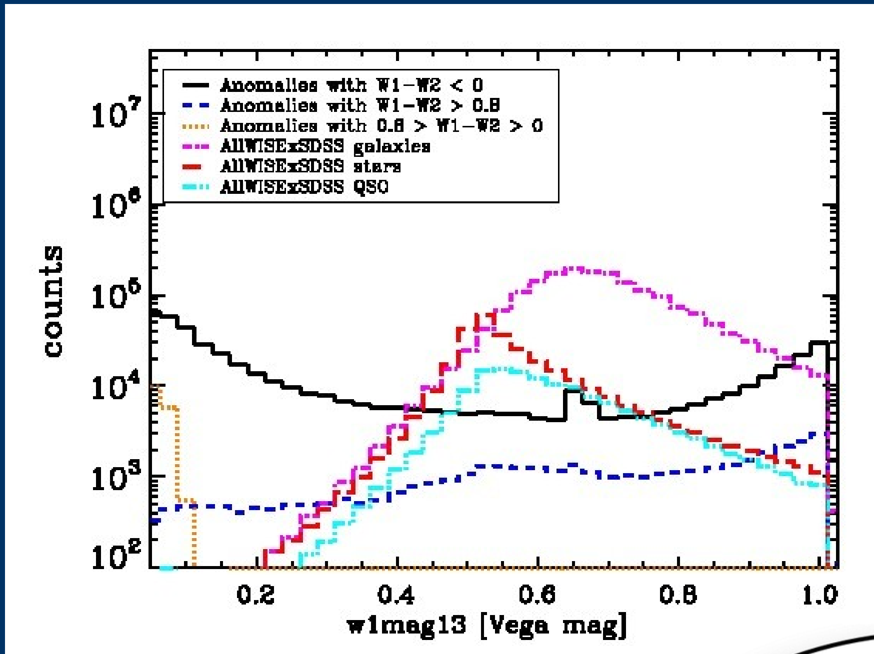


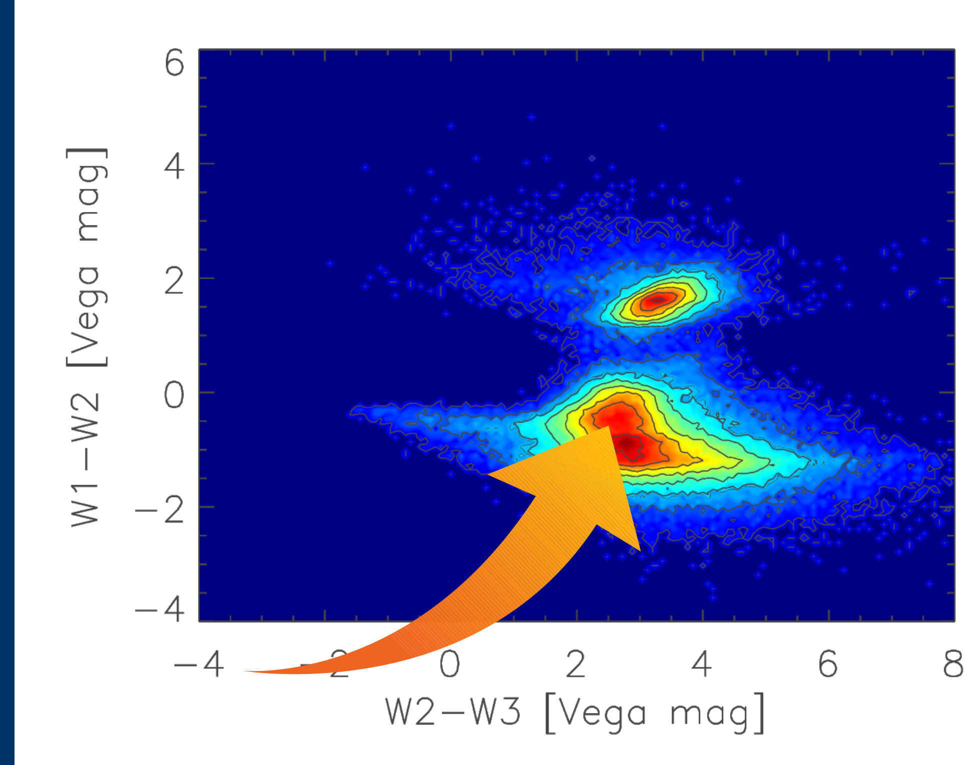
Solarz et al. 2017



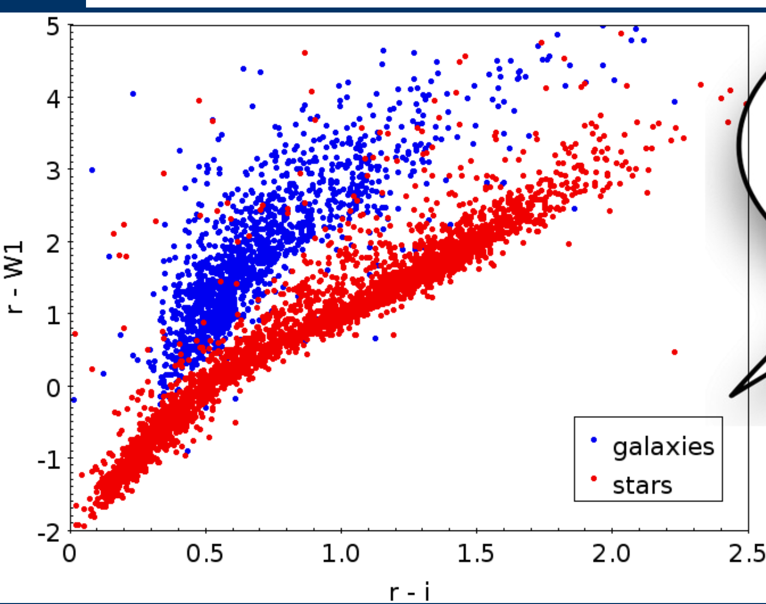


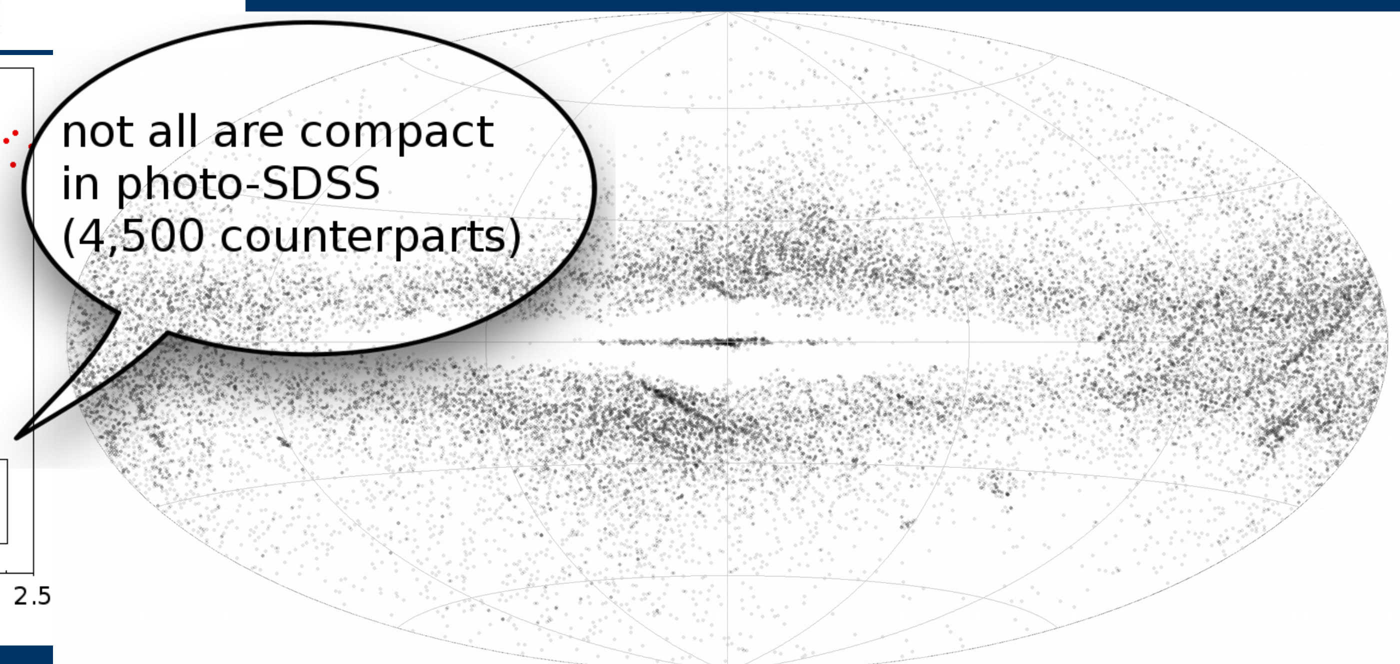
Mix of galaxies and stars?





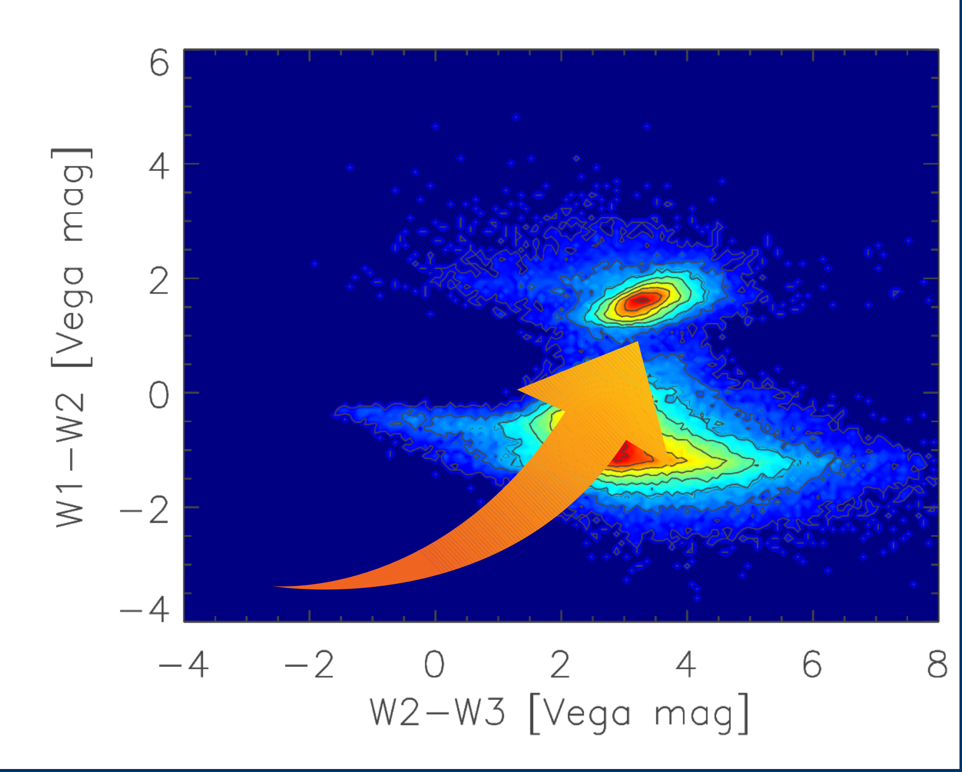
Solarz et al. 2017

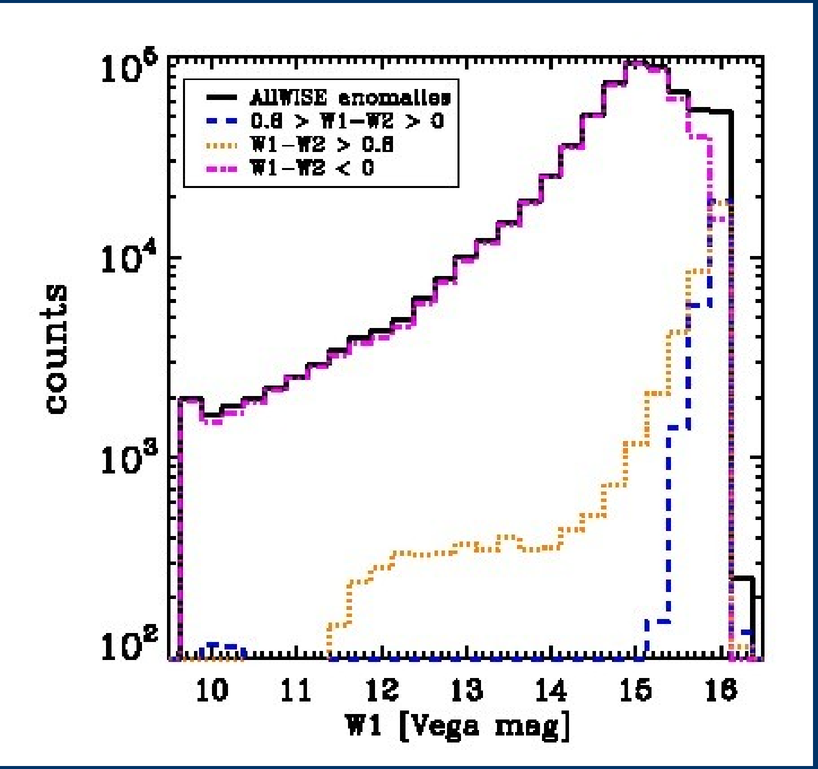




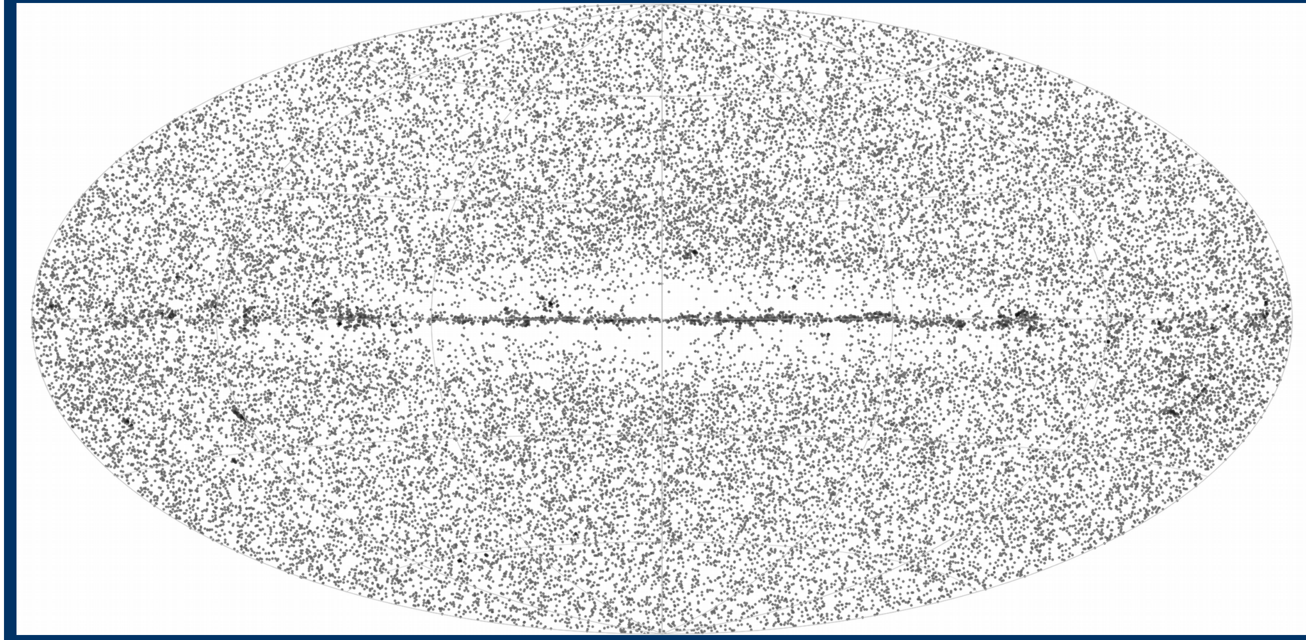
AGN candidates?

- 40,000 sources
- W1 ~ 16 [Vega mag], W3 [12 um] ~ 10 [Vega mag]
- no starlight can be redshifted to this channel
- Warm dust emission/PAH emission lines
- From theoretical predictions: AGN colours (Jarrett et al. 2011)
- Galactic Plane: mostly blends;





Solarz et al. 2017

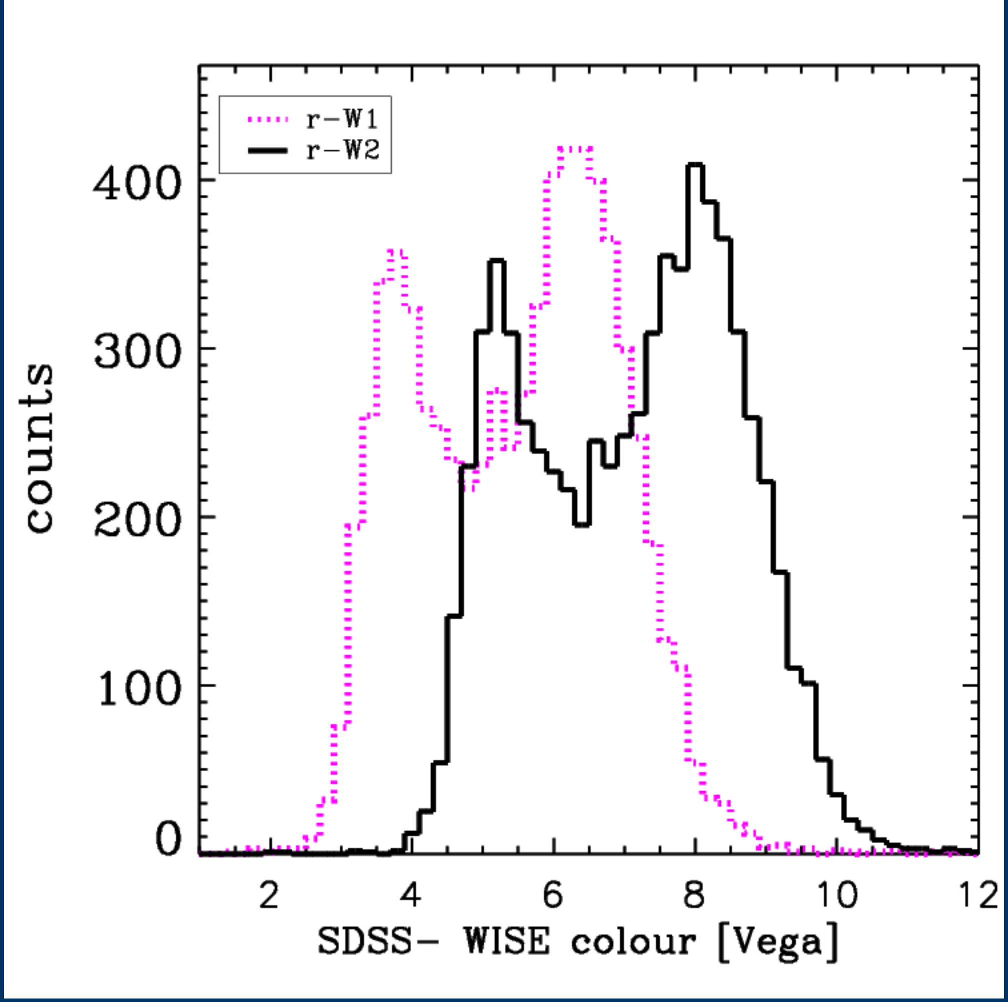


Obscured/Unobscured AGNs

- 7000 found in photometric SDSS, but no spectrum
- => all sky extrapolation (to full depth of WISE): 40% with no optical counterpart
- Two populations of AGNs: obscured and unobscured
- No other counterparts in any publicly available catalogues

To confirm:

- Follow-up optical photometry needed (future SDSS releases?)
- Spectroscopy would be best



Solarz et al. 2017

Summary

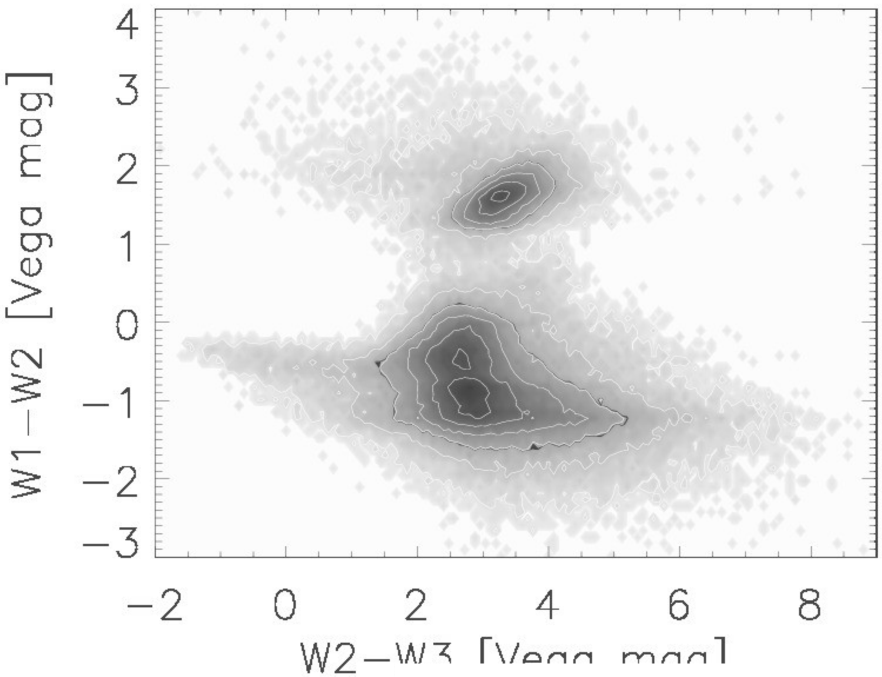
- We need to deal with unusual patterns in the data
- i.e. search for 'unknown unknowns'
- Anomaly detection:
- OCSVM => efficient selection of interesting and previously unclassified objects

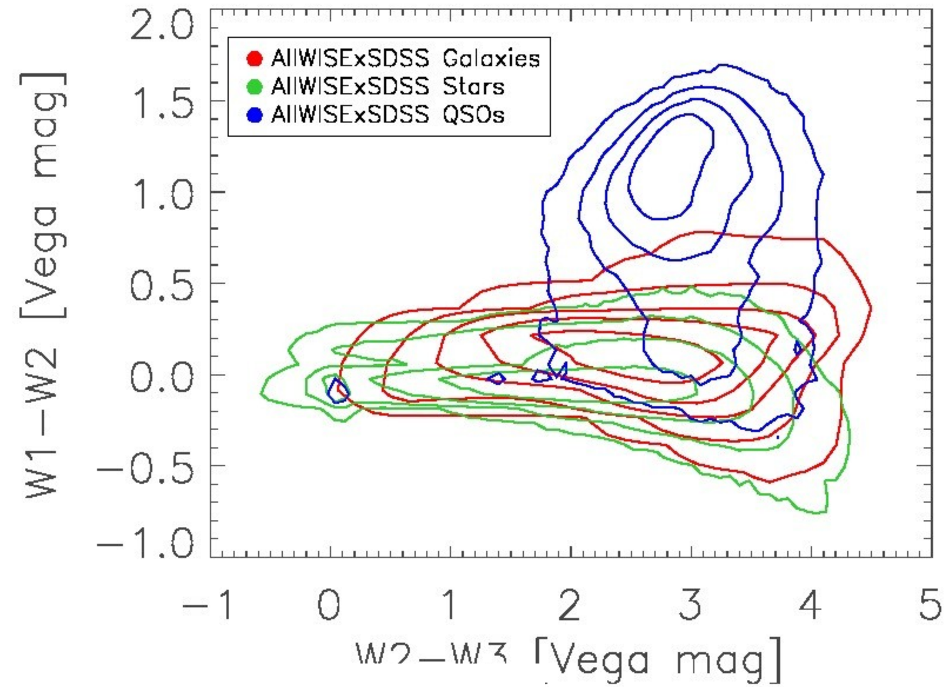
  → cleaning the data of unexpected/unaccounted for artifacts
- Verify nature of selected AGN candidates + correlation function calculations

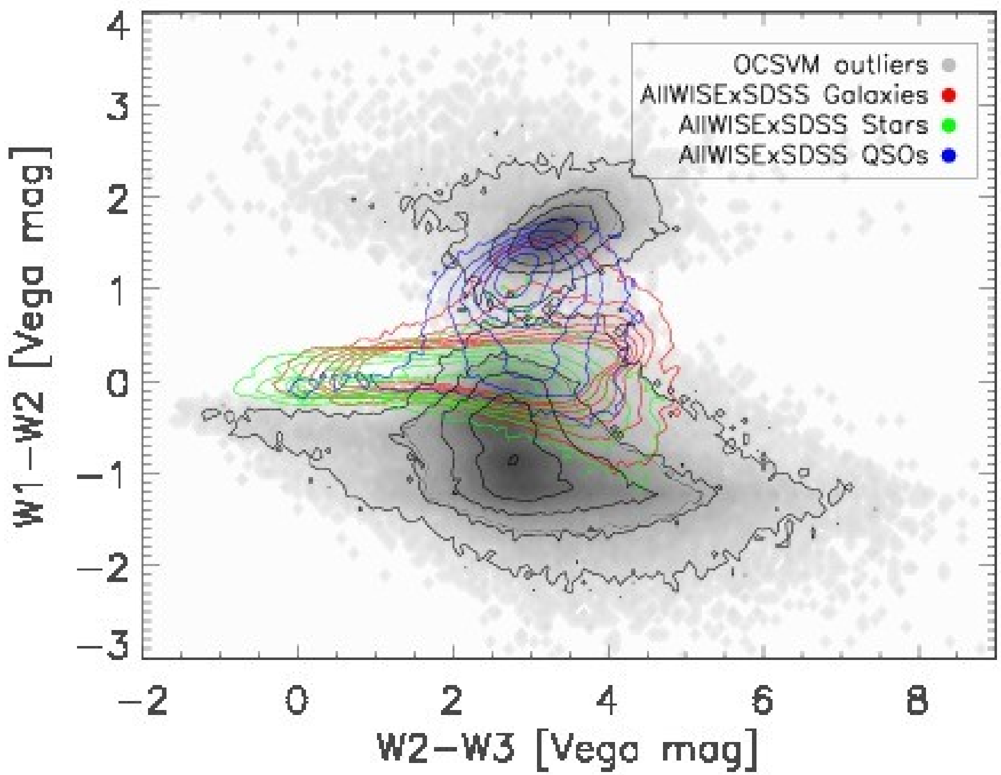
- http://www.R-project.org
- https://cran.r-project.org/web/packages/doParallel/index.html
- https://cran.r-project.org/web/packages/caret/index.html

Special thanks to Mark Taylor for the TOPCAT (Taylor 2005) and STILTS (Taylor 2006) software

Backup slides


W1−W2 [Vega mag]
W2−W3 [Vega mag]
4
3
2
1
0
−1
−2
−3
−2
0
2
4
6
8


AllWISExSDSS Galaxies
AllWISExSDSS Stars
AllWISExSDSS QSOs
W1−W2 [Vega mag]
W2−W3 [Vega mag]
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1
0
1
2
3
4
5


OCSVM outliers
AllWISExSDSS Galaxies
AllWISExSDSS Stars
AllWISExSDSS QSOs
W1−W2 [Vega mag]
W2−W3 [Vega mag]
4
3
2
1
0
−1
−2
−3
−2
0
2
4
6
8

Why (OC) SVM?

- Domain-based methods: location of the novelty boundary based on nearest points

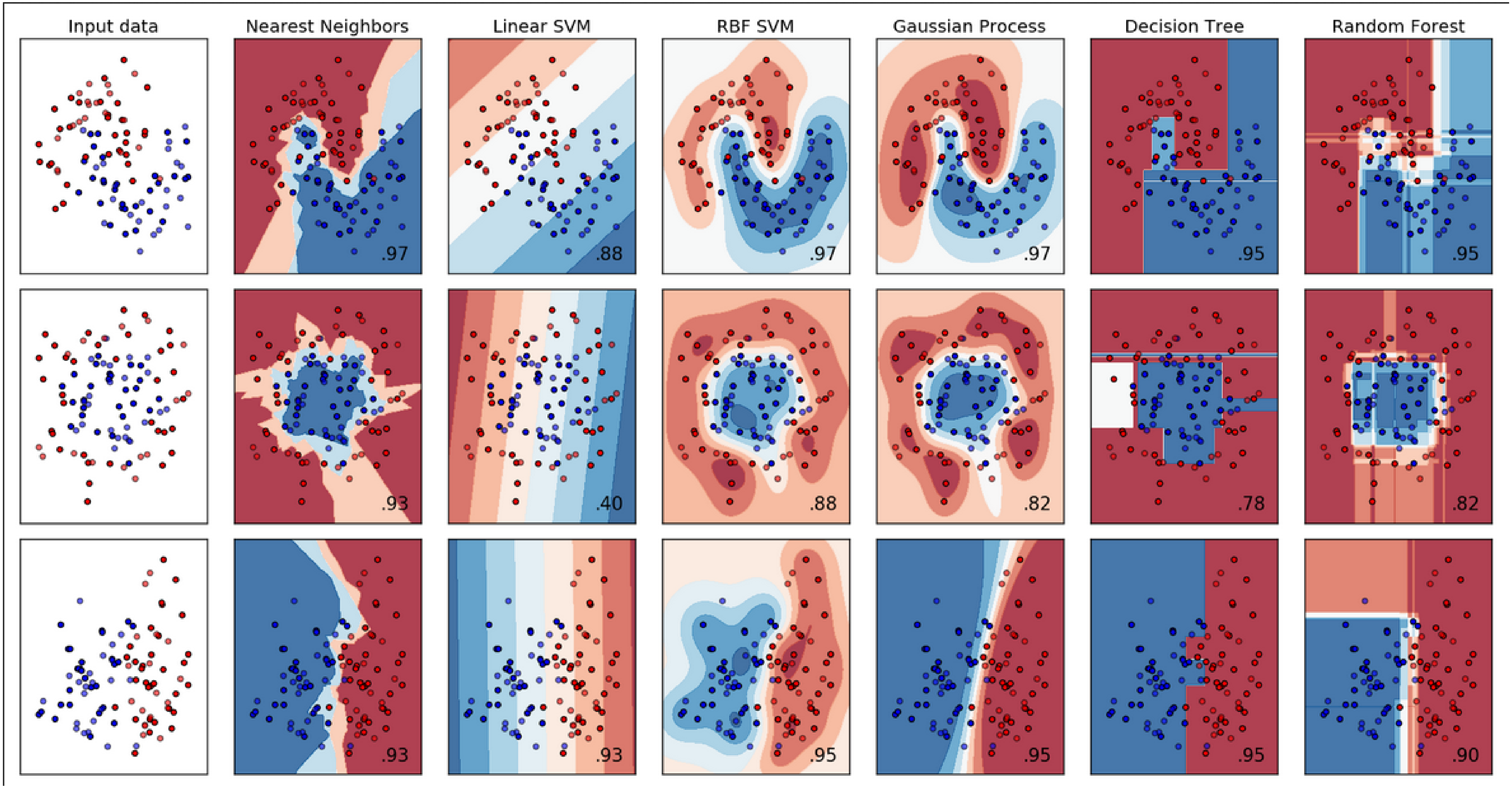
- Do not make any assumptions about the data distribution

- distance-based methods, e.g. NN; clustering: require definition of the distance metrics, distance measures in many dimensions lose ability to differentiate between normal and outlying data points; lack the flexibility of parameter tuning => unsuitable for full automatisation

- Great review of different anomaly dtection schemes:

  Pimentel et al. 2014

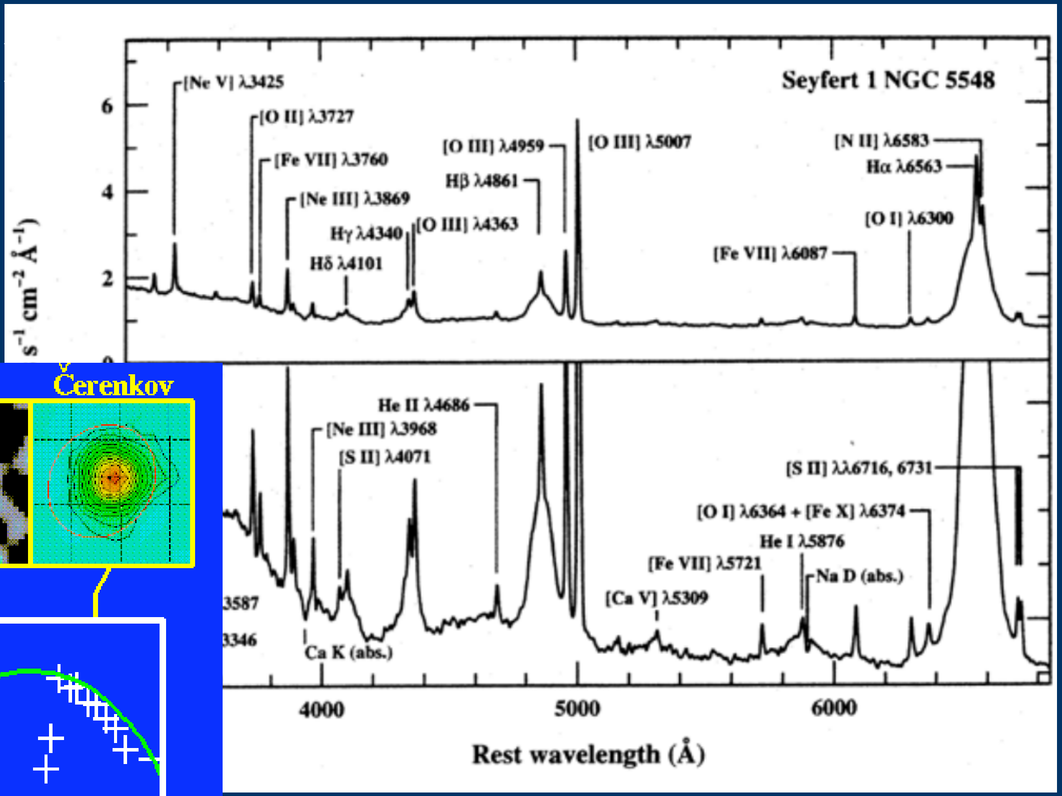
Best algorithm?

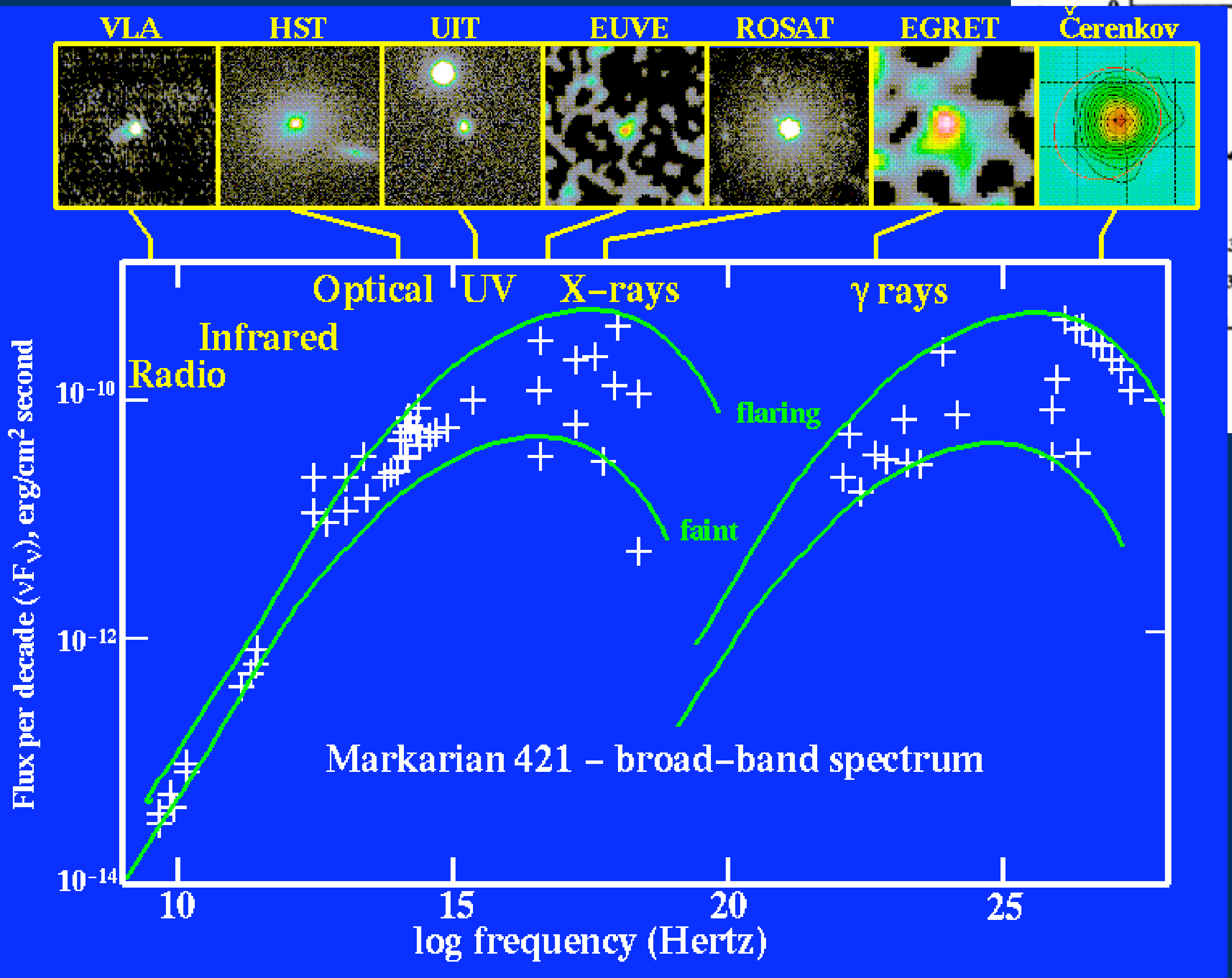


Credit: http://scikit-learn.org

Variety of AGN sources

- Seyfert galaxies (spirals)
- Quasars (Nuclear emission dominates
- Blazars (violetnly variable,
- Radio galaxies (ellipticals)
- Etc.



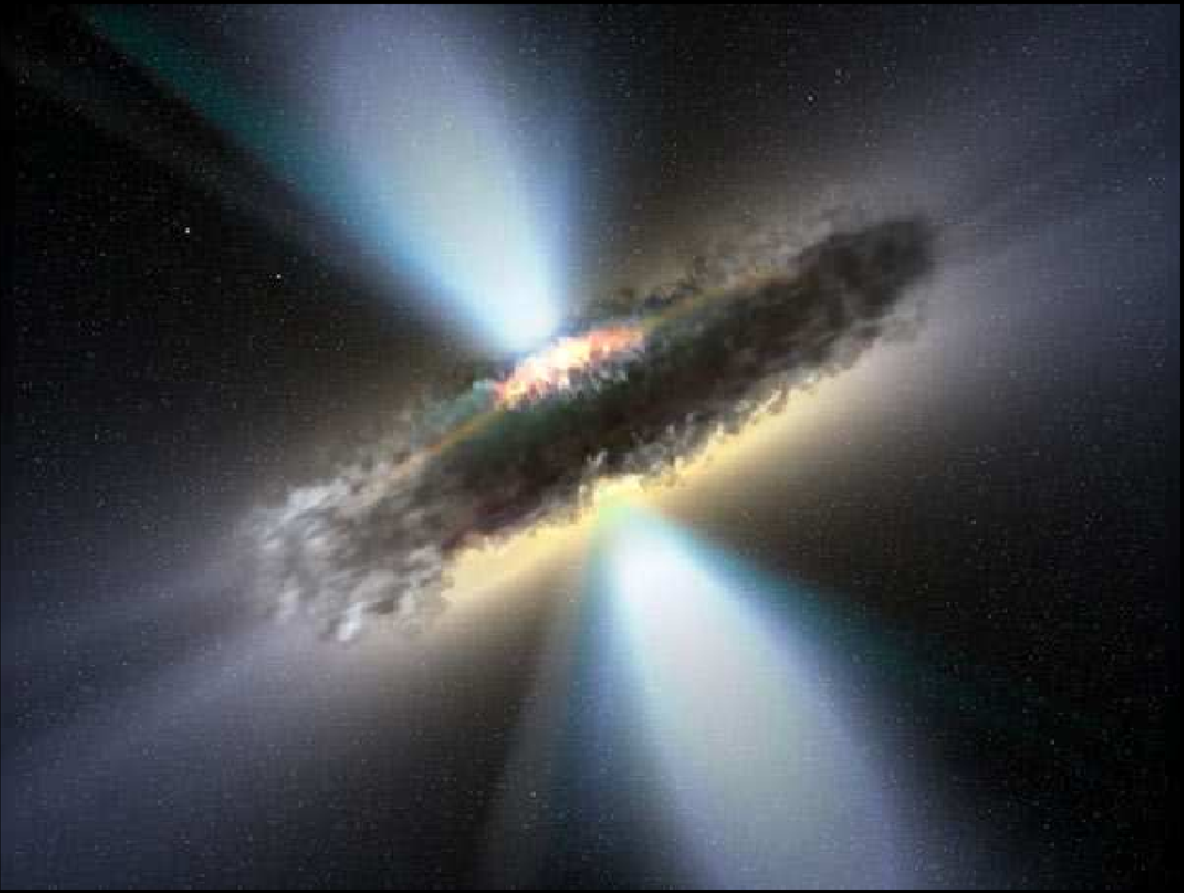


Obscured/unobscured AGNs: unified model

why such variety of observed phenomena → different objects?

Unified models:

- different classes of AGN => different orientations of intrinsically similar systems to the observer's line of sight.

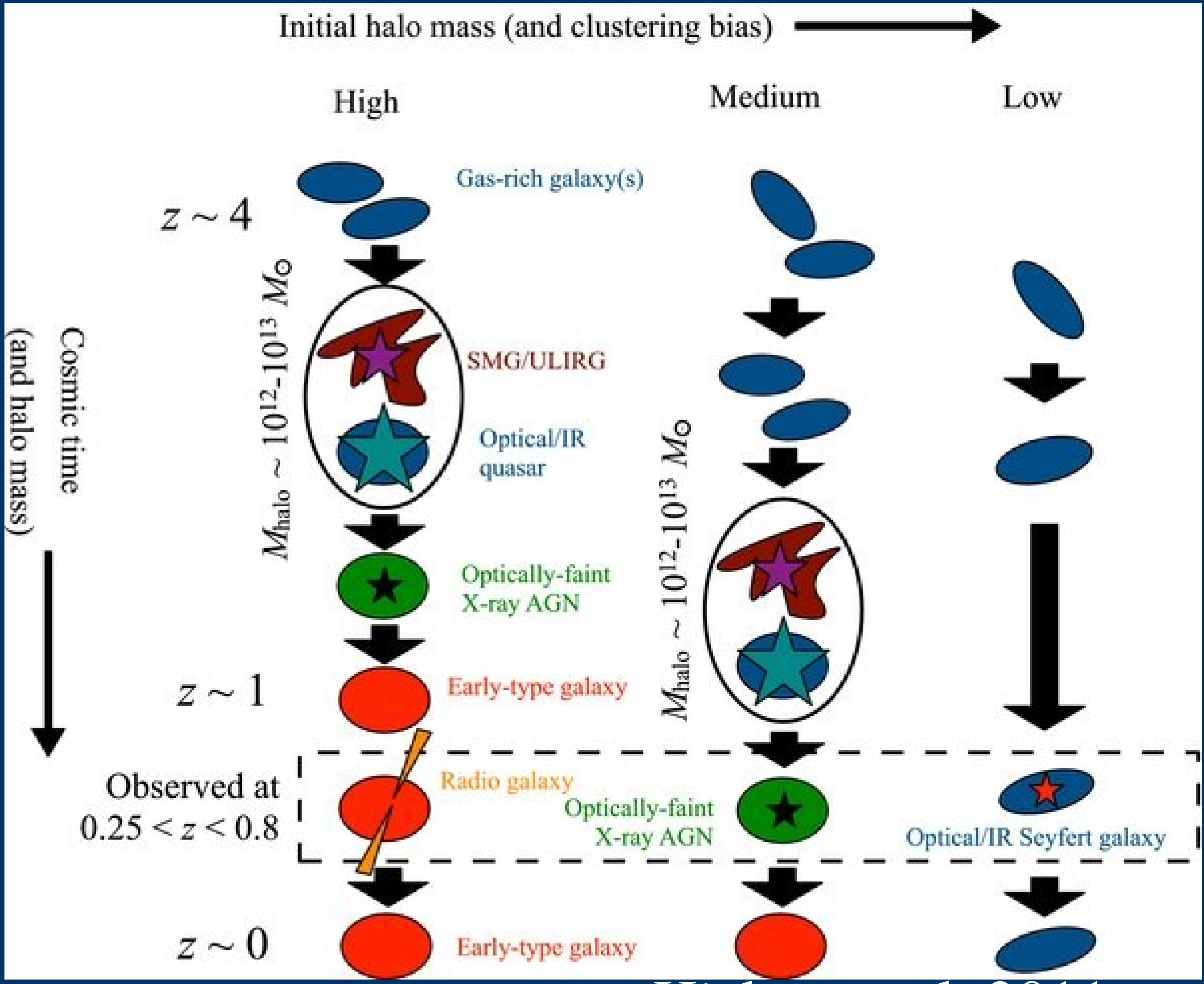


ESA/V. Beckmann

Obscured/unobscured AGNs: evolutionary differences

According to clustering measurements obscured/unobscured AGNs => separate populations evolving in different way

➔ If we get photo/spec redshifts we can weigh in on this discussion



Hickox et al. 2011