Bayes vs. ML in large-scale surveys

Demographics with eROSITA

Johannes Buchner

collaborators: Julien Wolf, Mara Salvato, eROAGN

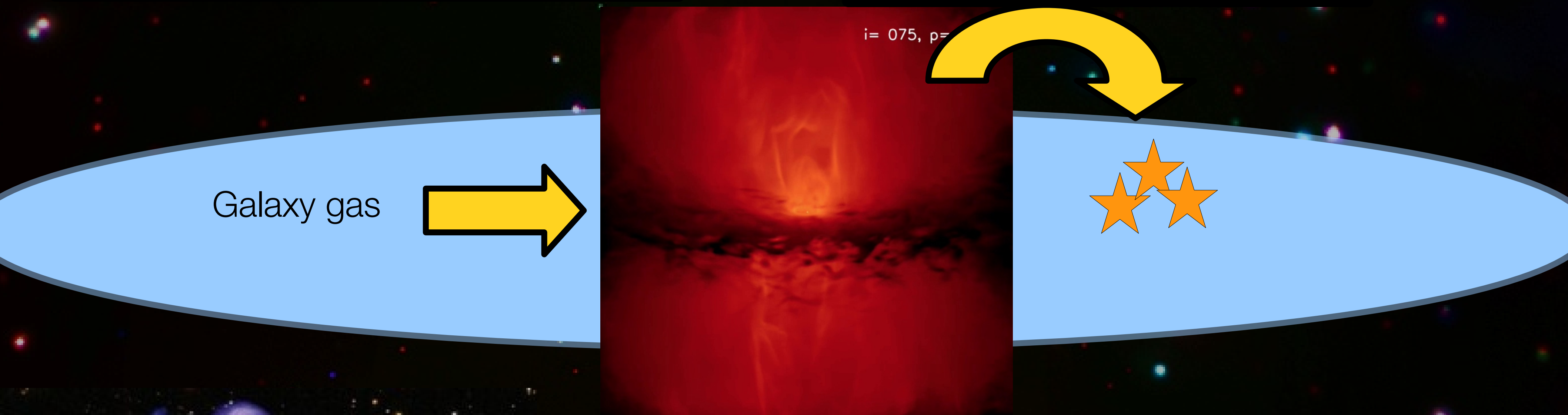
http://astrost.at/istics/

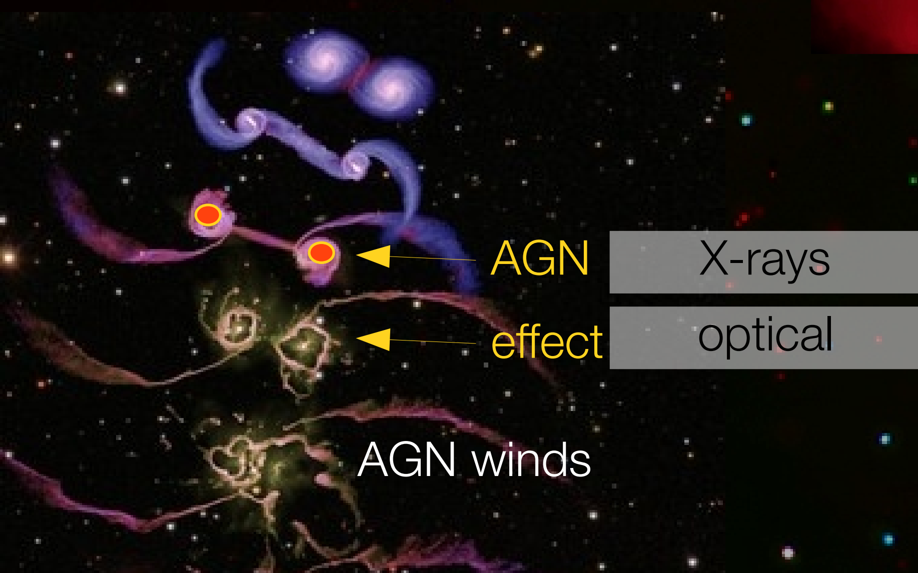
ML-IAP, Oct 2021

Galaxy-AGN interaction

Which galaxies activate AGN when?

Impact of AGN on host galaxy?



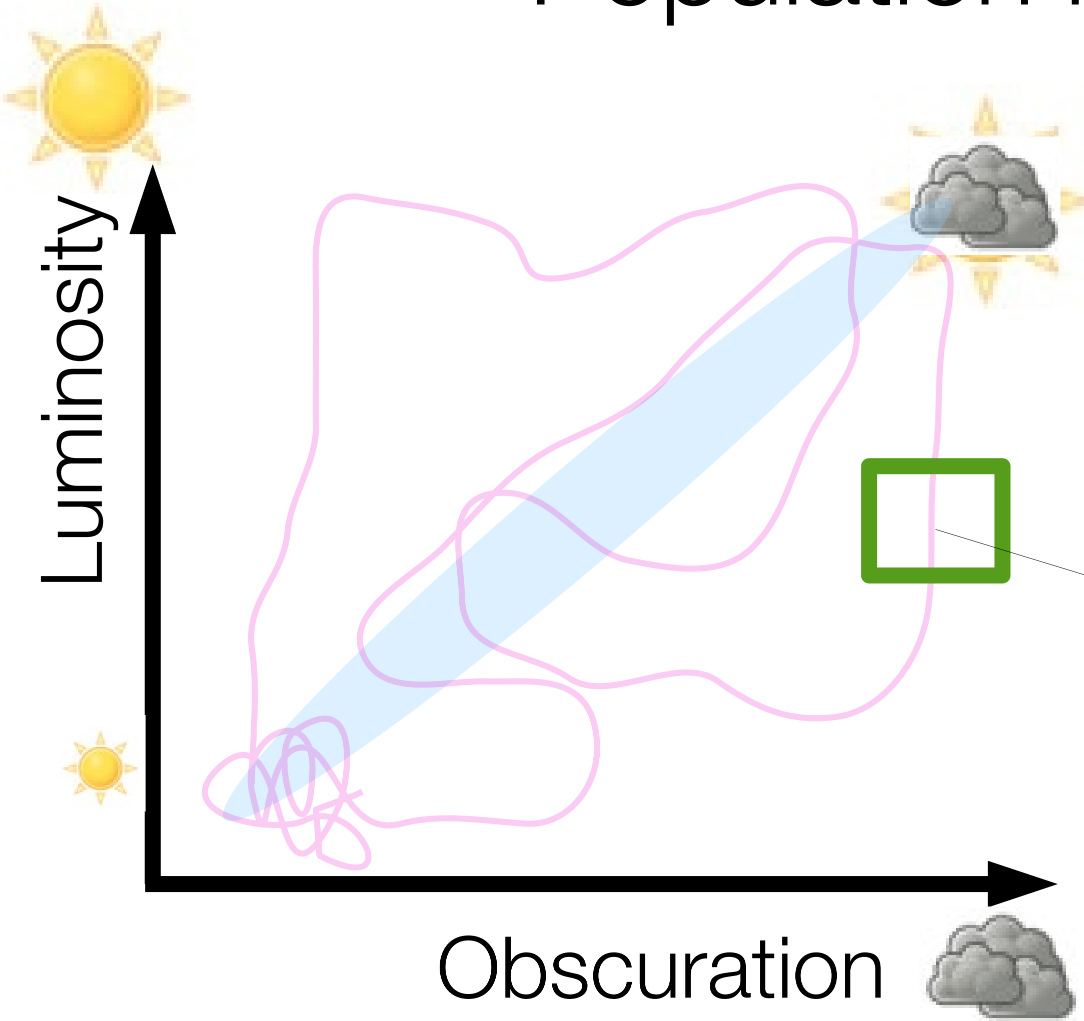


Highly stochastic, multi-scale process

multi-wavelength needed

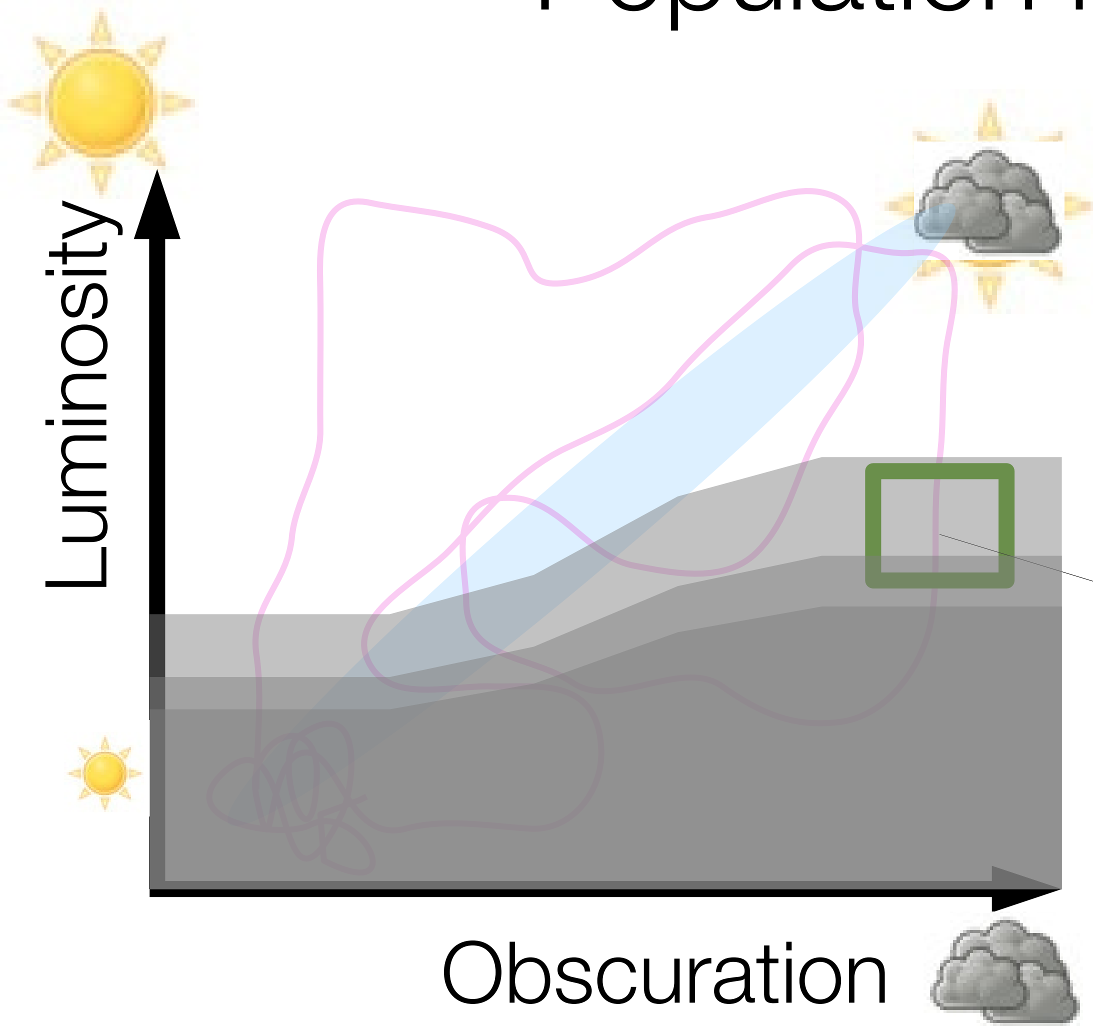
biased samples

Population inference



$\Phi(L, N_{\mathrm{H}})$ #/Mpc³

Population inference



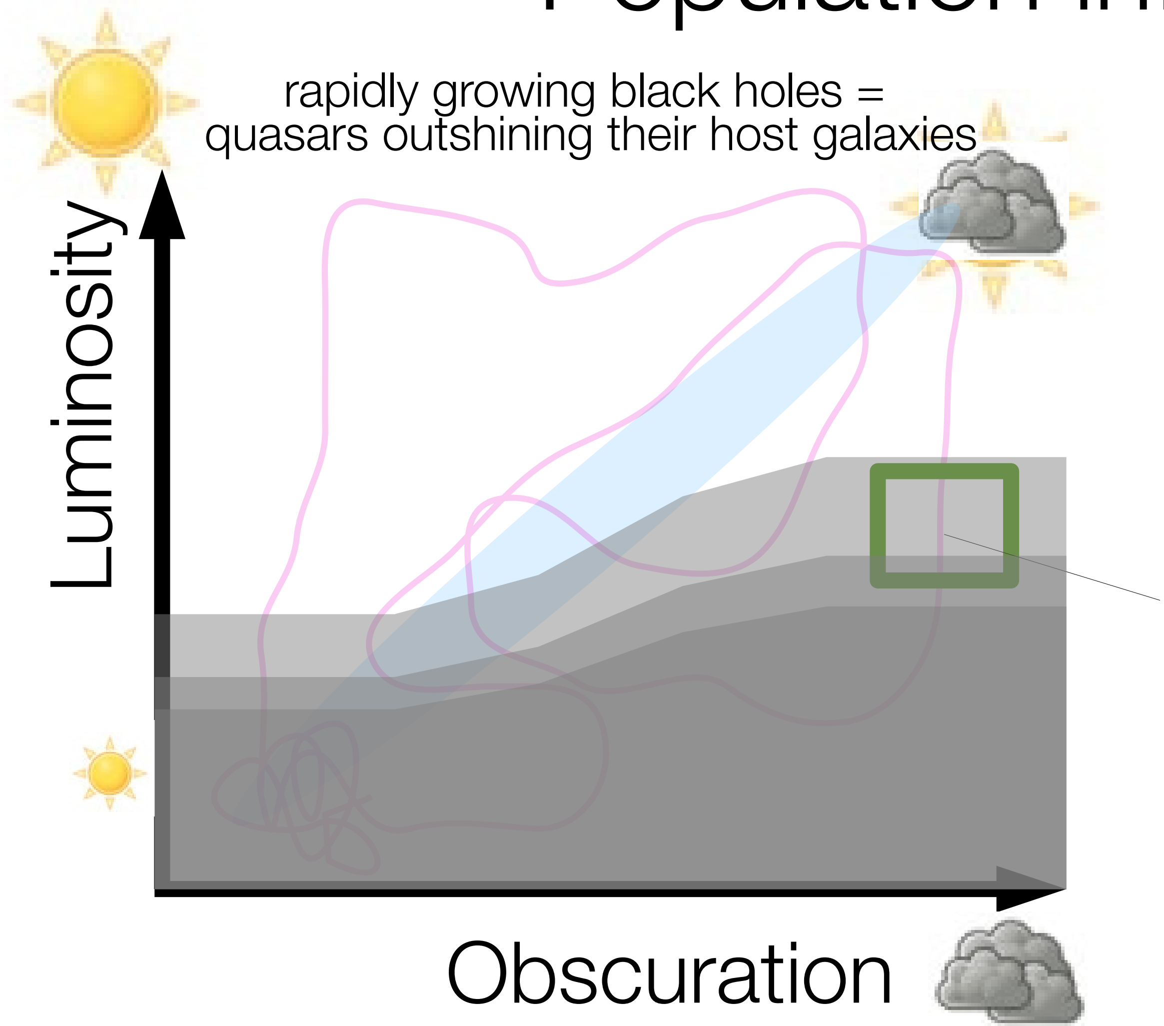
$\Phi(L, N_{\mathrm{H}})$ #/Mpc³

Addressing selection bias:

- X-ray selection
- Accounting for incompleteness

→ demographics of the underlying population

Population inference



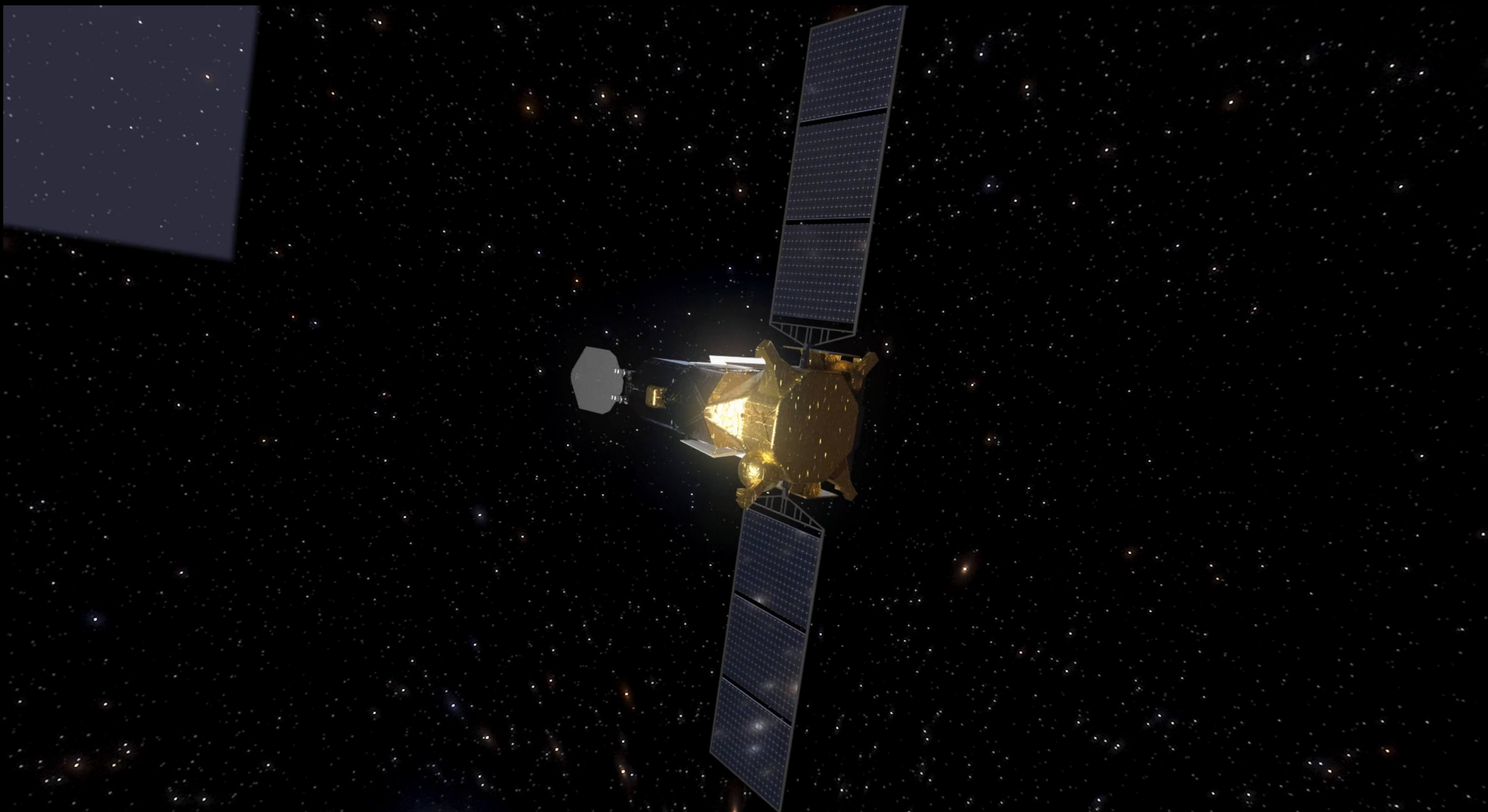
$\Phi(L, N_{\mathrm{H}})$ #/Mpc³

Addressing selection bias:

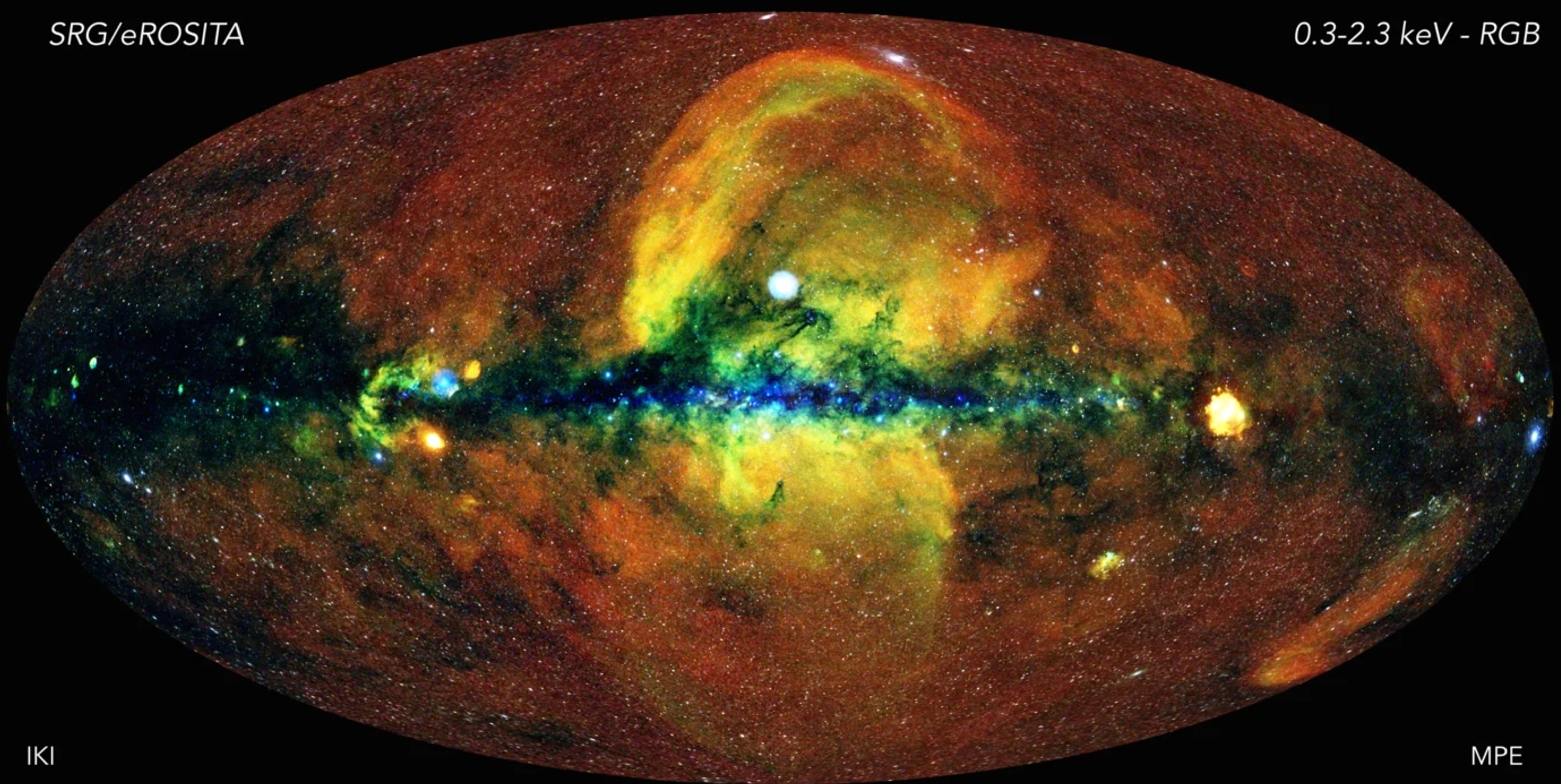
- X-ray selection
- Accounting for incompleteness

→ demographics of the underlying population

eROSITA



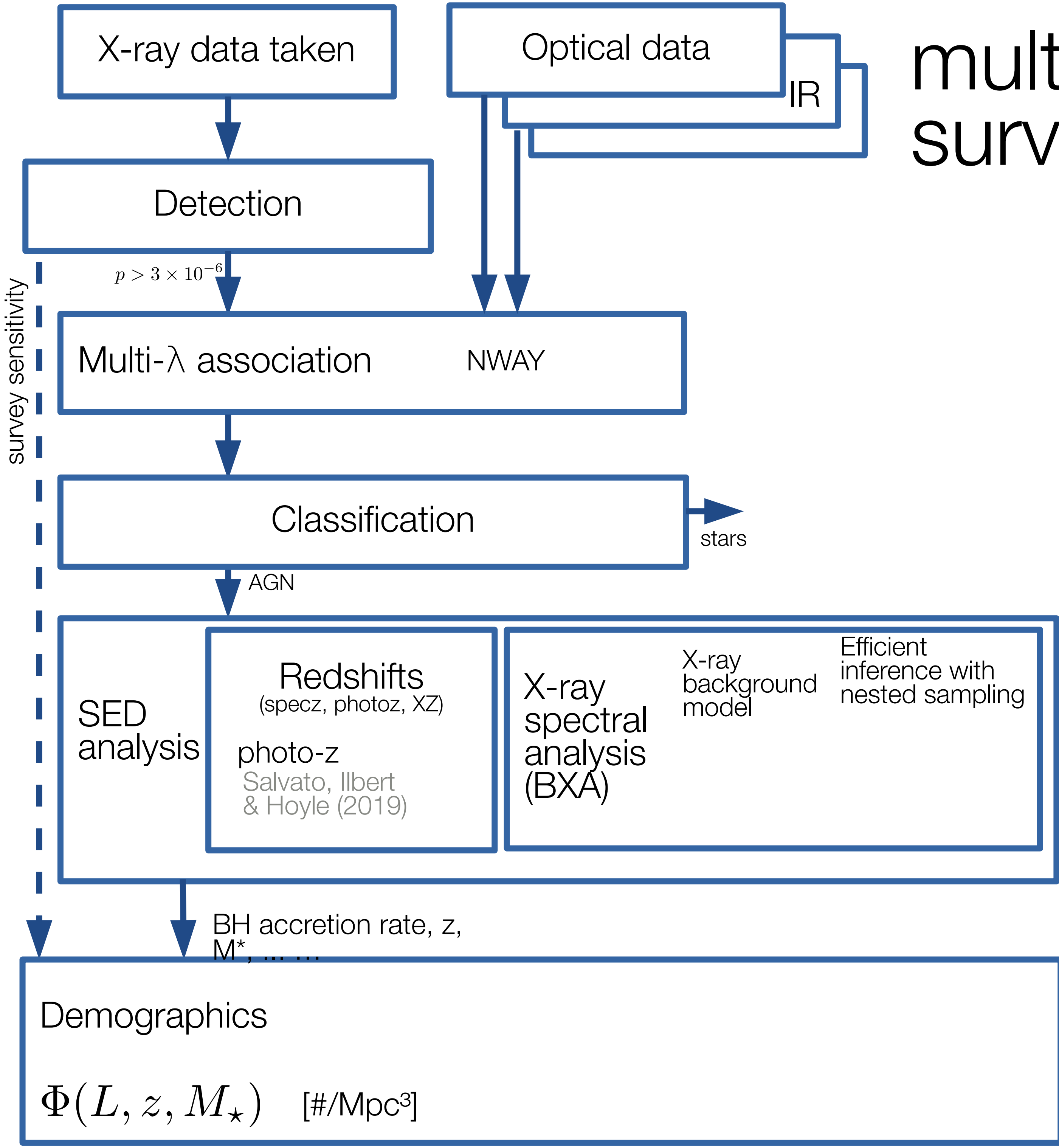
eROSITA

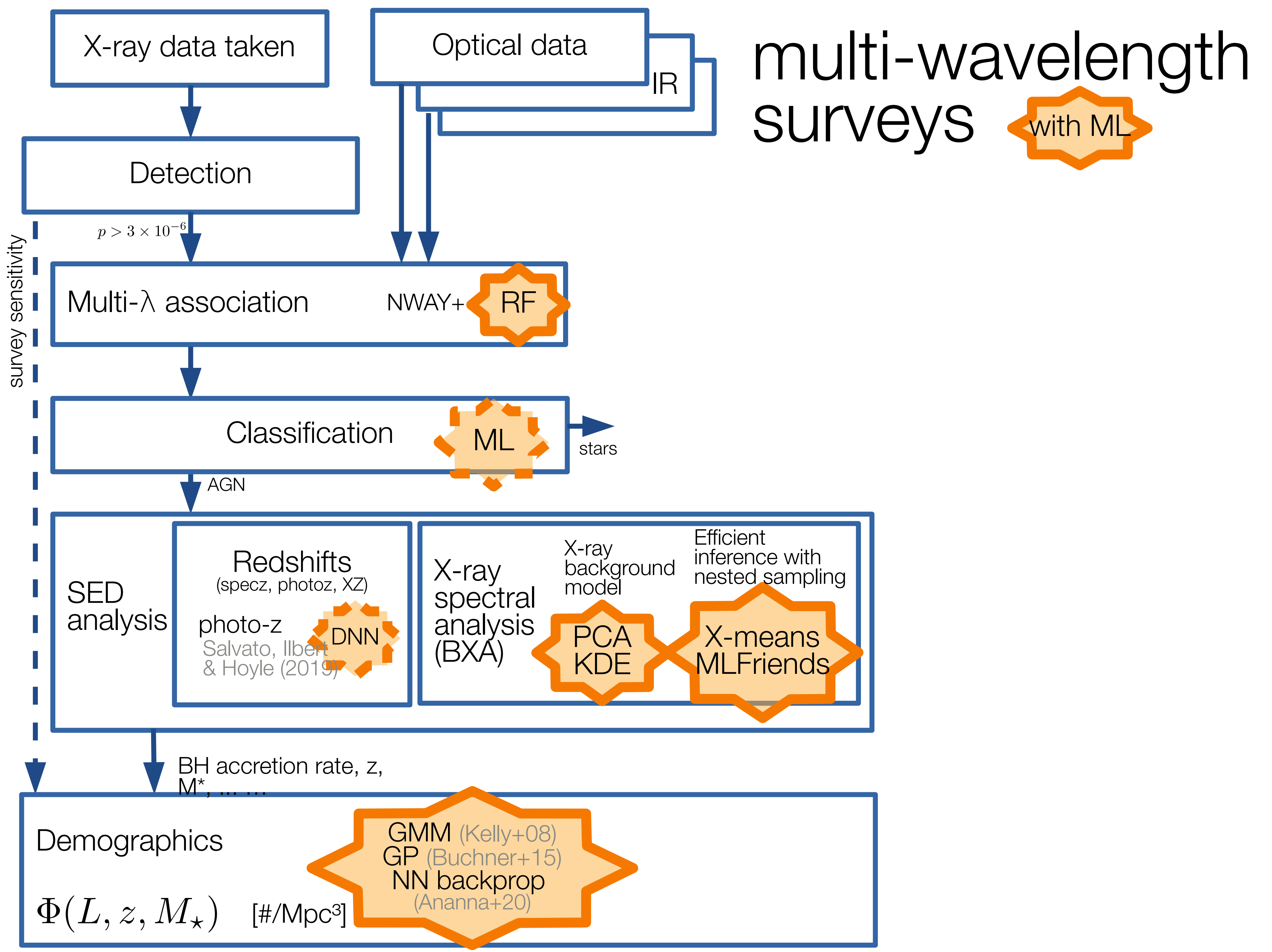


~1 million AGN in the first all-sky survey

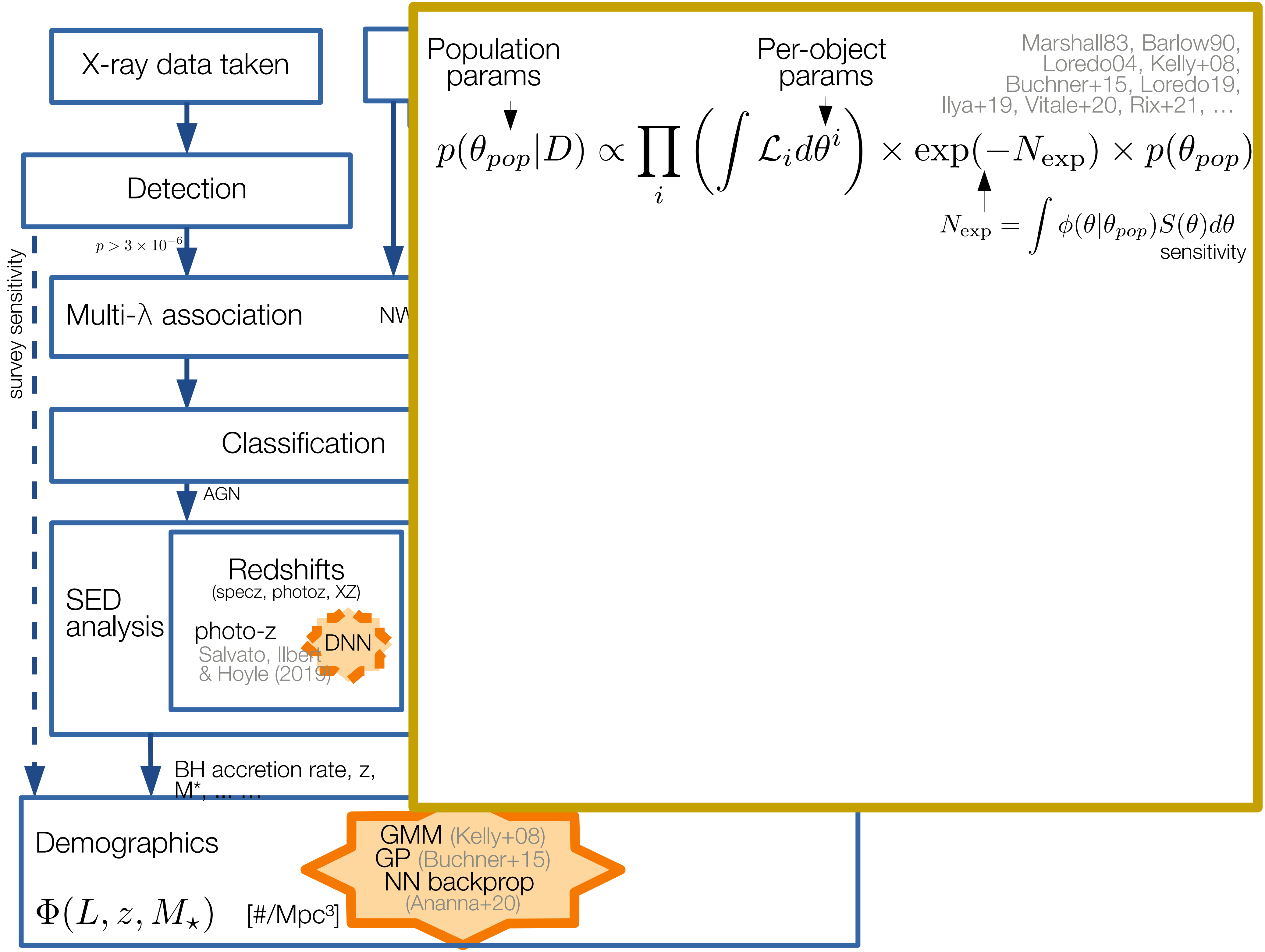
8 all-sky surveys, one every 6 months
AGN variability, physics, demographics, …

multi-wavelength surveys

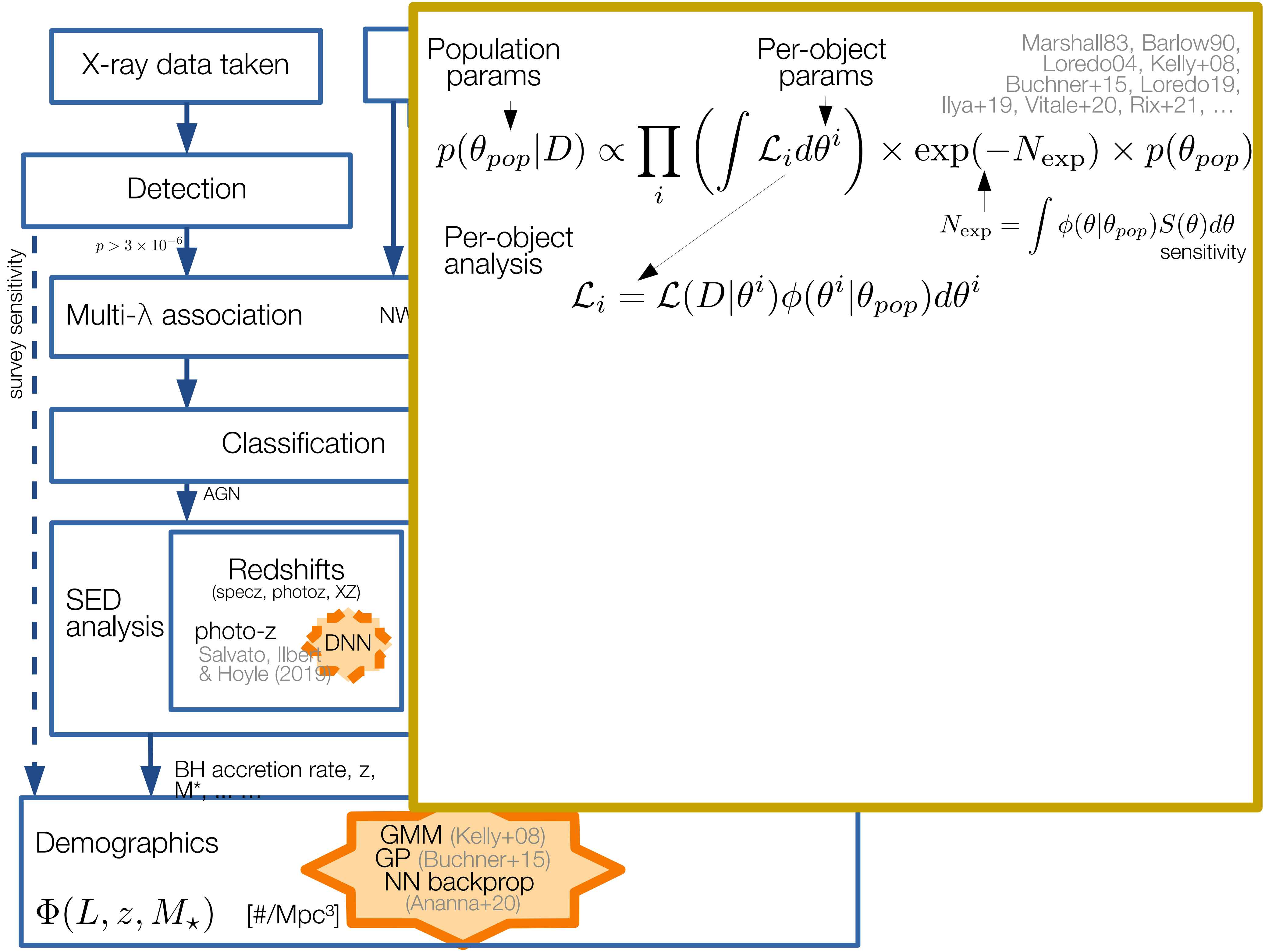




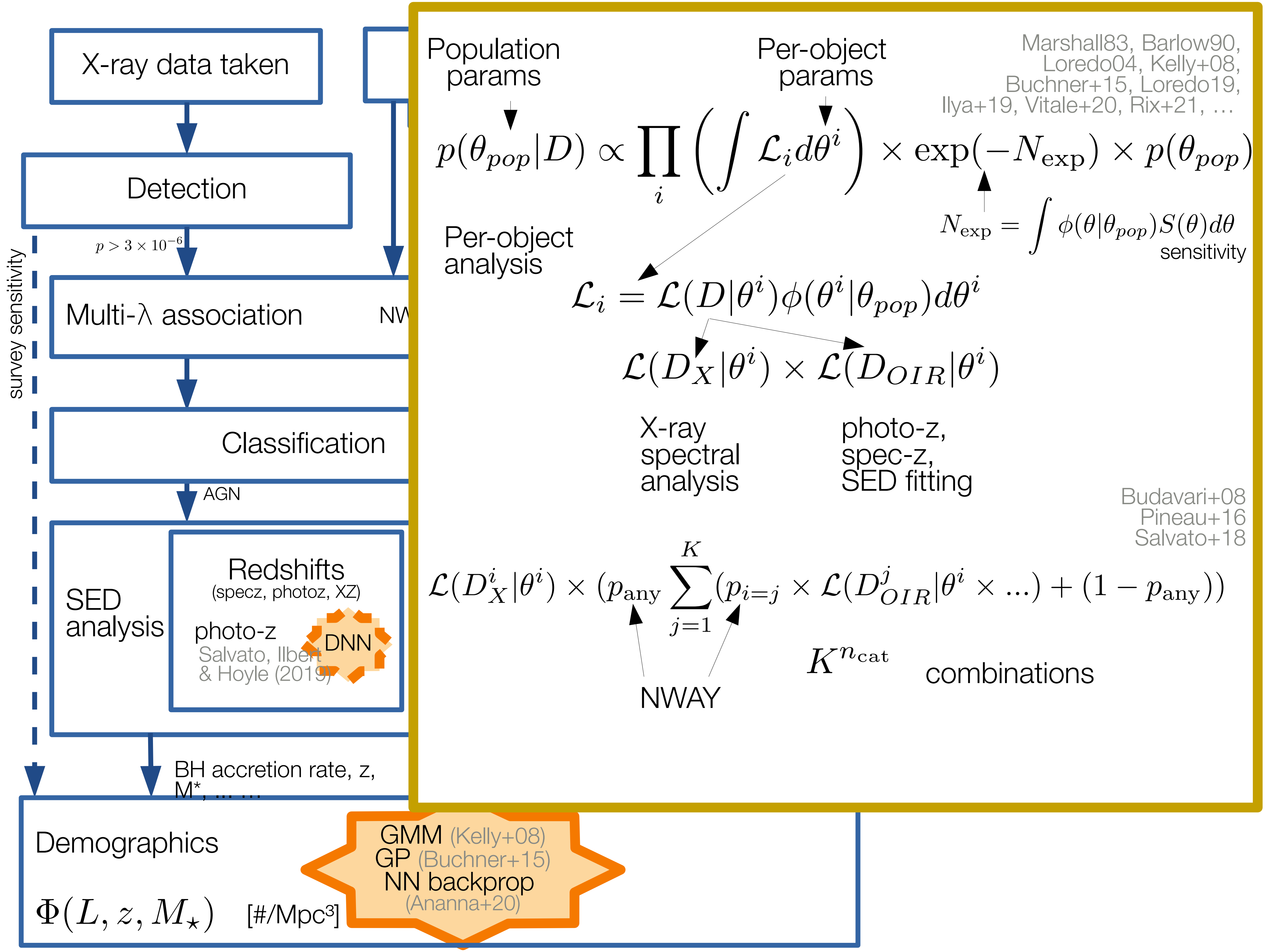
multi-wavelength surveys
with ML
X-ray data taken
Optical data
IR
Detection
$p > 3 \times 10^{-6}$
survey sensitivity
Multi-λ association
NWAY+
RF
Classification
ML
stars
AGN
SED analysis
Redshifts
(specz, photoz, XZ)
photo-z
DNN
Salvato, Ilbert & Hoyle (2019)
X-ray spectral analysis (BXA)
X-ray background model
PCA
KDE
Efficient inference with nested sampling
X-means
MLFriends
BH accretion rate, z, M*, ...
Demographics
$\Phi(L, z, M_\star)$
[#/Mpc³]
GMM (Kelly+08)
GP (Buchner+15)
NN backprop
(Ananna+20)



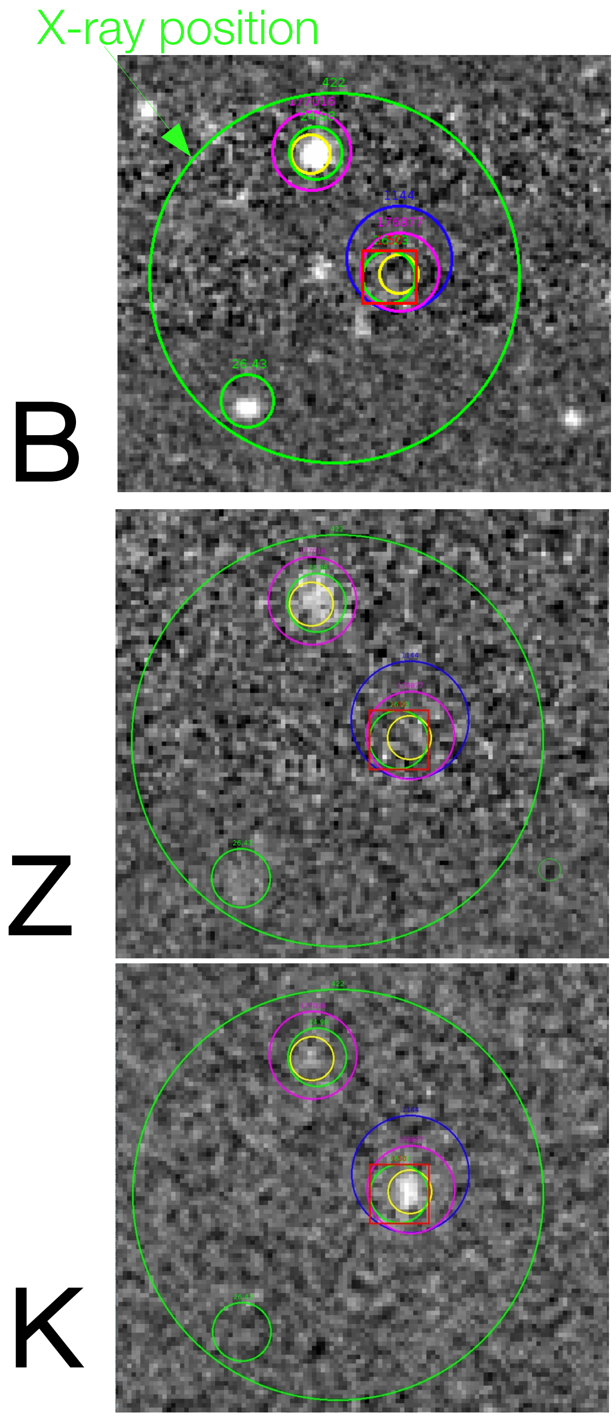
X-ray data taken
Detection
$p > 3 \times 10^{-6}$
survey sensitivity
Multi-λ association
NW
Classification
AGN
SED analysis
Redshifts
(specz, photoz, XZ)
photo-z
Salvato, Ilbert & Hoyle (2019)
DNN
BH accretion rate, z, M*, ...
Demographics
$\Phi(L, z, M_\star)$ [#/Mpc³]
GMM (Kelly+08)
GP (Buchner+15)
NN backprop
(Ananna+20)
Population params
Per-object params
Marshall83, Barlow90, Loredo04, Kelly+08, Buchner+15, Loredo19, Ilya+19, Vitale+20, Rix+21, ...
$p(\theta_{pop}|D) \propto \prod_i \left(\int \mathcal{L}_i d\theta^i \right) \times \exp(-N_{\exp}) \times p(\theta_{pop})$
$N_{\exp} = \int \phi(\theta|\theta_{pop}) S(\theta) d\theta$
sensitivity



X-ray data taken
Detection
$p > 3 \times 10^{-6}$
Multi-λ association
NW
Classification
AGN
SED analysis
Redshifts
(specz, photoz, XZ)
photo-z
Salvato, Ilbert & Hoyle (2019)
DNN
survey sensitivity
BH accretion rate, z, M*, ...
Demographics
$\Phi(L, z, M_\star)$ [#/Mpc³]
GMM (Kelly+08)
GP (Buchner+15)
NN backprop
(Ananna+20)
Population params
Per-object params
Marshall83, Barlow90, Loredo04, Kelly+08, Buchner+15, Loredo19, Ilya+19, Vitale+20, Rix+21, …
$p(\theta_{pop}|D) \propto \prod_i \left(\int \mathcal{L}_i d\theta^i \right) \times \exp(-N_{\exp}) \times p(\theta_{pop})$
$N_{\exp} = \int \phi(\theta|\theta_{pop}) S(\theta) d\theta$
sensitivity
Per-object analysis
$\mathcal{L}_i = \mathcal{L}(D|\theta^i)\phi(\theta^i|\theta_{pop})d\theta^i$


X-ray data taken
Detection
$p > 3 \times 10^{-6}$
survey sensitivity
Multi-λ association
NW
Classification
AGN
SED analysis
Redshifts
(specz, photoz, XZ)
photo-z
Salvato, Ilbert & Hoyle (2019)
DNN
BH accretion rate, z, M*, ...
Demographics
$\Phi(L, z, M_\star)$
[#/Mpc³]
GMM (Kelly+08)
GP (Buchner+15)
NN backprop
(Ananna+20)
Population params
Per-object params
Marshall83, Barlow90, Loredo04, Kelly+08, Buchner+15, Loredo19, Ilya+19, Vitale+20, Rix+21, ...
$p(\theta_{pop}|D) \propto \prod_i \left(\int \mathcal{L}_i d\theta^i\right) \times \exp(-N_{\exp}) \times p(\theta_{pop})$
$N_{\exp} = \int \phi(\theta|\theta_{pop}) S(\theta) d\theta$
sensitivity
Per-object analysis
$\mathcal{L}_i = \mathcal{L}(D|\theta^i)\phi(\theta^i|\theta_{pop})d\theta^i$
$\mathcal{L}(D_X|\theta^i) \times \mathcal{L}(D_{OIR}|\theta^i)$
X-ray spectral analysis
photo-z, spec-z, SED fitting
Budavari+08
Pineau+16
Salvato+18
$\mathcal{L}(D_X^i|\theta^i) \times (p_{\rm any} \sum_{j=1}^{K} (p_{i=j} \times \mathcal{L}(D_{OIR}^j|\theta^i \times ...) + (1 - p_{\rm any}))$
$K^{n_{\rm cat}}$ combinations
NWAY

NWAY – Bayesian association



- Automated association of N catalogs simultaneously

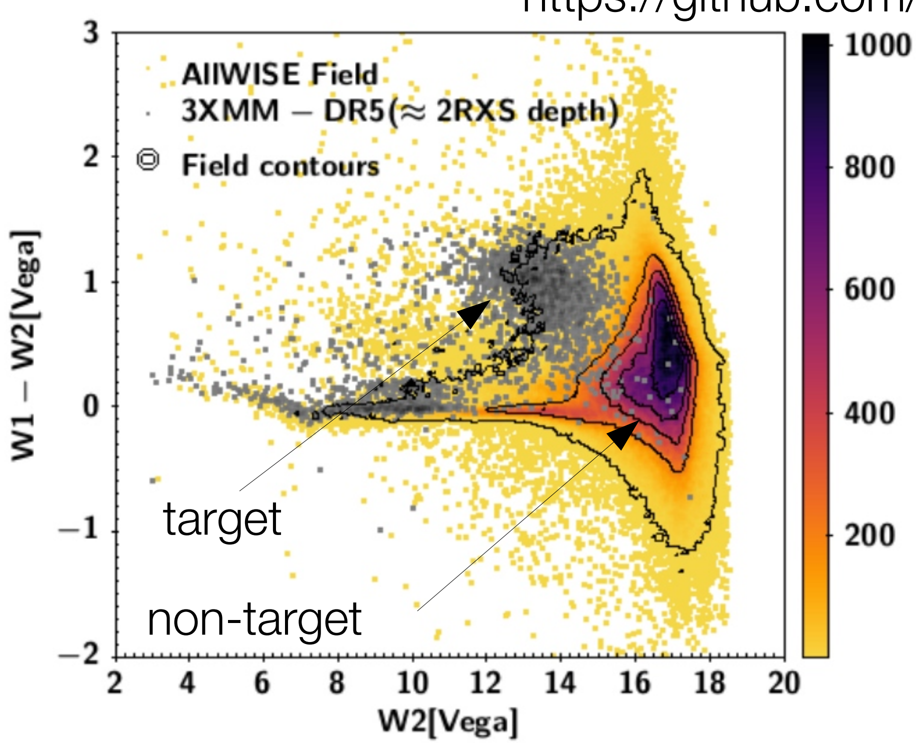
Salvato, Buchner+18

- Use color information to weigh alternatives in a consistent Bayesian framework

→ higher completeness and purity

→ becoming popular across fields

https://github.com/JohannesBuchner/nway

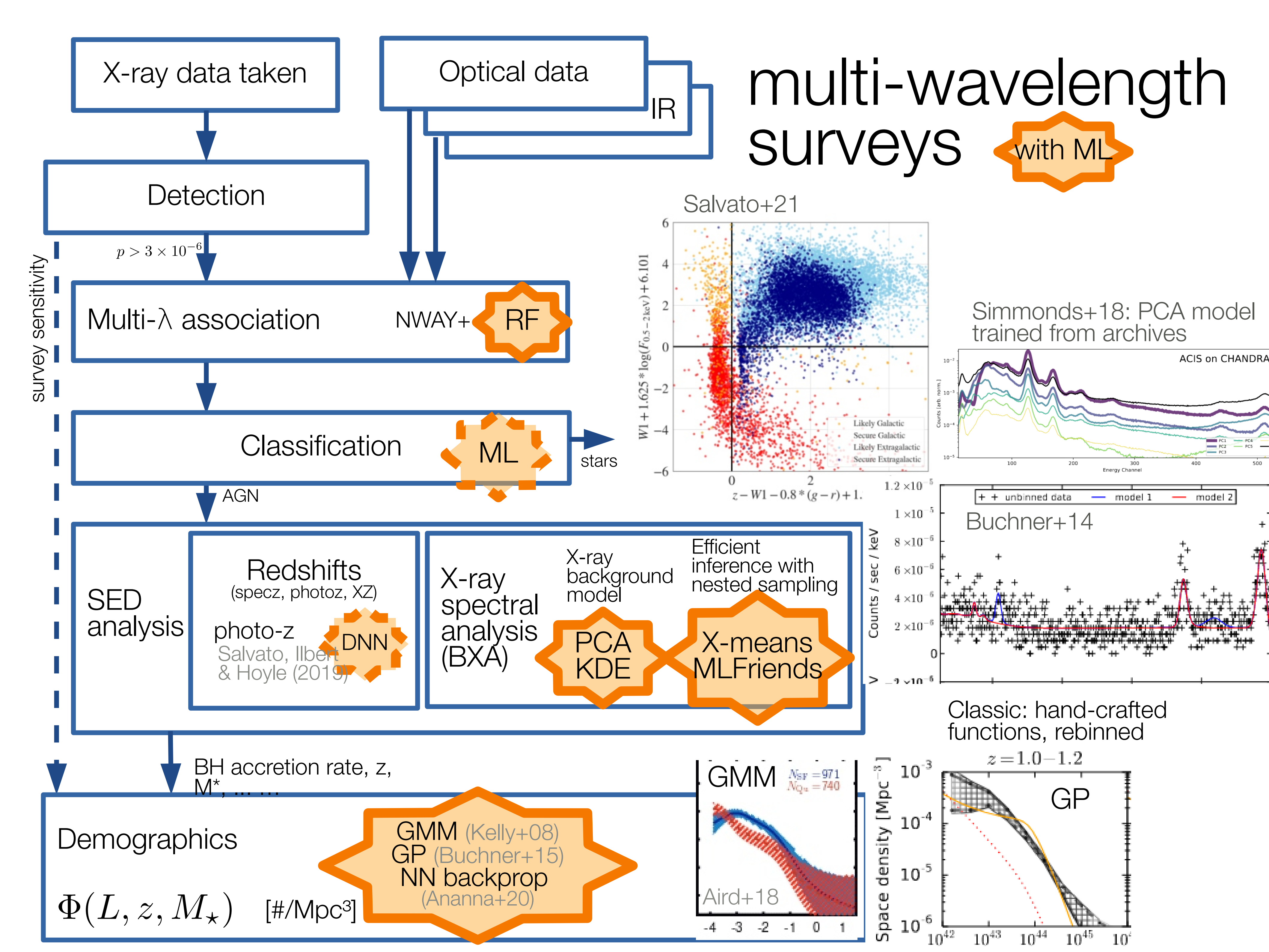


Automatically learn separations

Transfer learning from previous surveys

ML priors:
Random forests learn photometry of X-ray sources → judge probability of NWAY options (Julien Wolf)

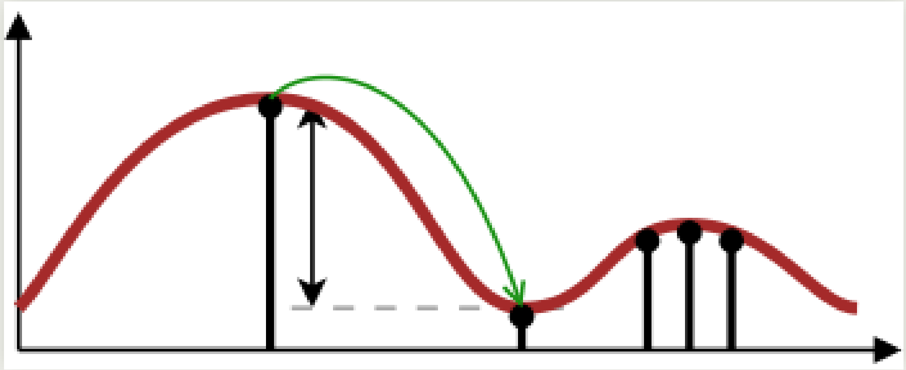
Salvato, Wolf+21

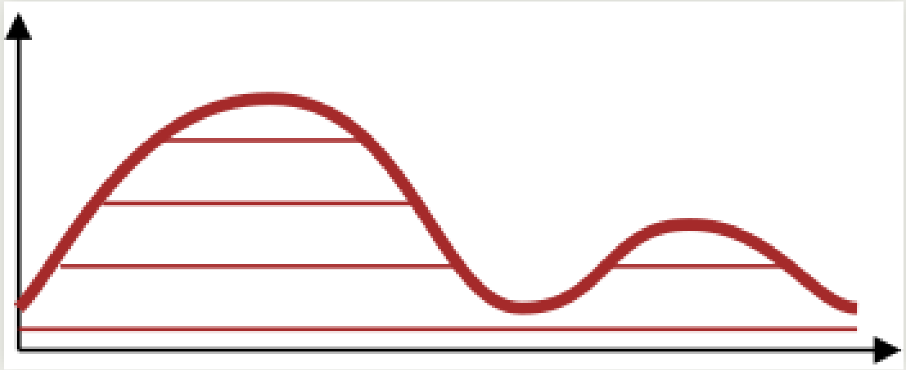


multi-wavelength surveys
with ML
X-ray data taken
Optical data
IR
Detection
$p > 3 \times 10^{-6}$
survey sensitivity
Multi-λ association
NWAY+
RF
Classification
ML
stars
AGN
SED analysis
Redshifts
(specz, photoz, XZ)
photo-z
DNN
Salvato, Ilbert & Hoyle (2019)
X-ray spectral analysis (BXA)
X-ray background model
PCA KDE
Efficient inference with nested sampling
X-means MLFriends
BH accretion rate, z, M*, ...
Demographics
$\Phi(L, z, M_\star)$
[#/Mpc³]
GMM (Kelly+08)
GP (Buchner+15)
NN backprop
(Ananna+20)
Salvato+21
Likely Galactic
Secure Galactic
Likely Extragalactic
Secure Extragalactic
Simmonds+18: PCA model trained from archives
ACIS on CHANDRA
Energy Channel
Buchner+14
unbinned data
model 1
model 2
Counts / sec / keV
Classic: hand-crafted functions, rebinned
GMM
Aird+18
$z = 1.0 - 1.2$
GP
Space density [Mpc^{-3}]

Nested Sampling idea

- MCMC: only consider likelihood ratios. Integration by vertical slices
- nested sampling: compute geometric size at various likelihood thresholds
- orthogonal, unique re-ordering of volume by likelihood

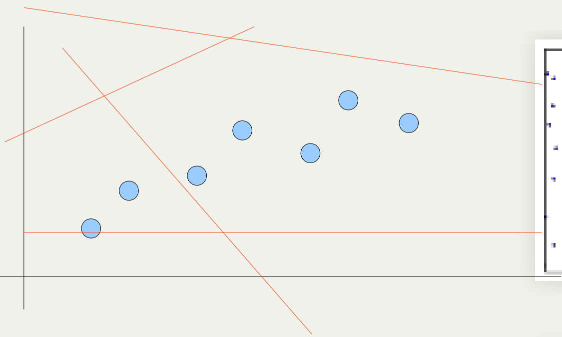


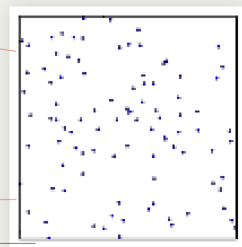


$$\sum \underbrace{\text{Shrinkage} \times \text{Likelihood}}_{\text{Importance of shell}} = Z$$

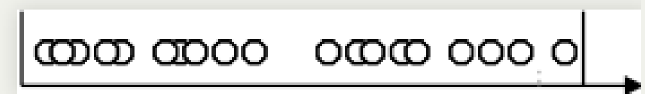
→ track volume shrinkage as likelihood increases

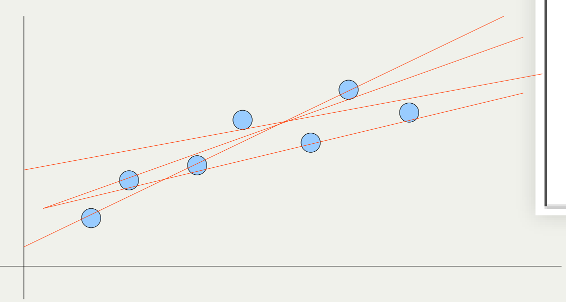
Nested Sampling algorithm

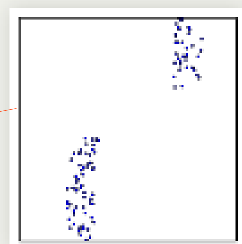




- Start with volume 1, draw randomly uniformly 200 points
- remove one, volume shrinks by 1/200.







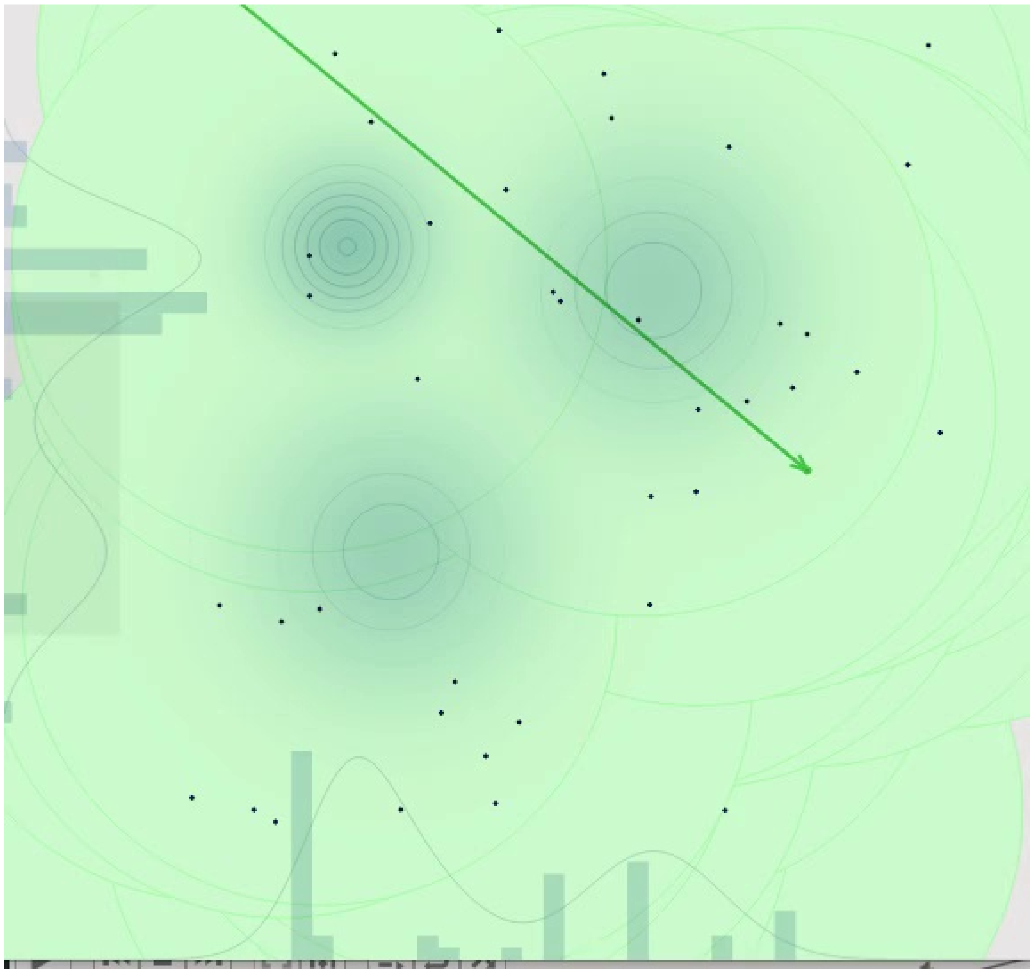
- draw a new one excluding the removed volume
- Unique ordering of space required: via likelihood

draw a new uniformly random point, with higher likelihood

(the crux of nested sampling)

- Scanning up vertically, done at some point
- converges (flat at highest likelihood)

L-restricted prior sampling



https://johannesbuchner.github.io/UltraNest/method.html

Find live point neighborhood

Sample from there → efficient!

padding to be safe:
super-set of unknown contour

General solutions exist!

X-means clustering with fudge parameters for padding → MultiNest (Shaw+07, Feroz&Hobson08)

Learn padding by train/test bootstrap split → MLFriends (Buchner16,19,21)

Non-volume preserving flows ease parameter space exploration (Moss+2019)

Bayes+ML in surveys

Consistent Bayesian inference framework

Model selection effects, go from sample to underlying demographics

- Physically meaningful parameters
- Physical models where we trust them and care
- Probabilistic machine learning where we do not

Good practices:
ablation studies,
understand how model regularizes,
test on simulate data,
separate observing instrumentation effects from physical process

NWAY, BXA, UltraNest, PyMultiNest

https://astrost.at/istics