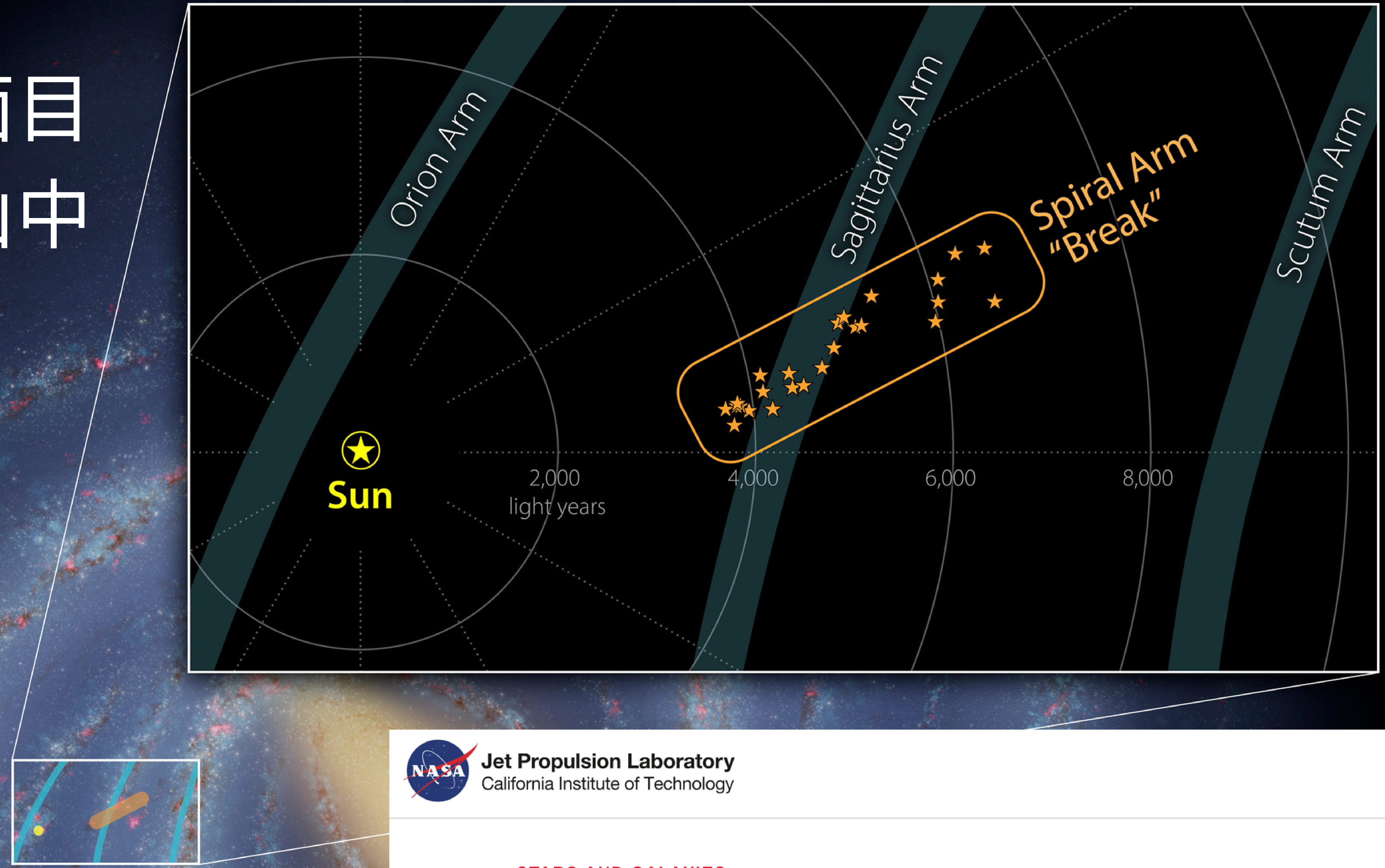
ASTROSTATS TIPS OF THE DAY

Rafael S.de Souza
Shanghai Astronomical Observatory
Chair: The Cosmostatistics Initiative

不识庐山真面目
只缘身在此山中



Jet Propulsion Laboratory
California Institute of Technology

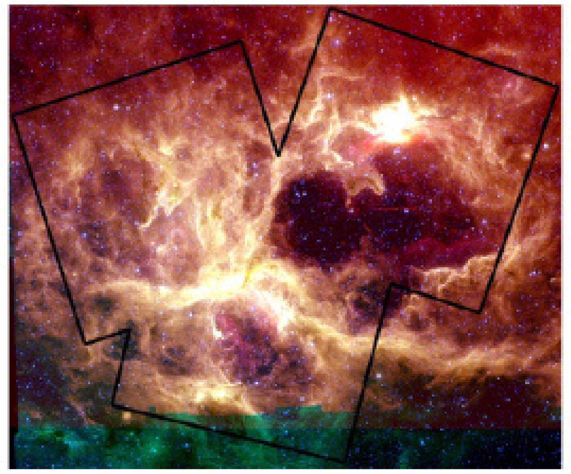
STARS AND GALAXIES

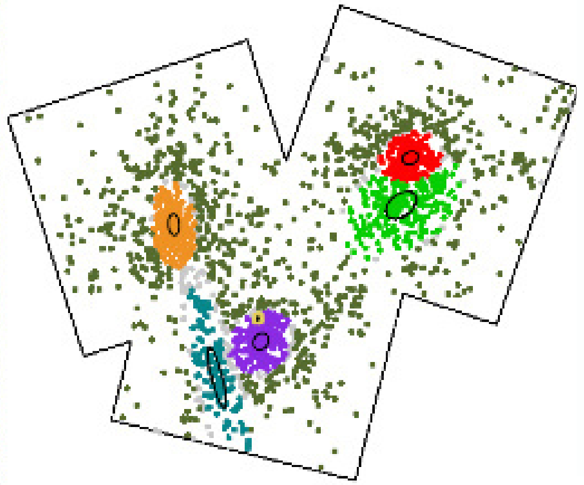
Astronomers Find a 'Break' in One of the Milky Way's Spiral Arms

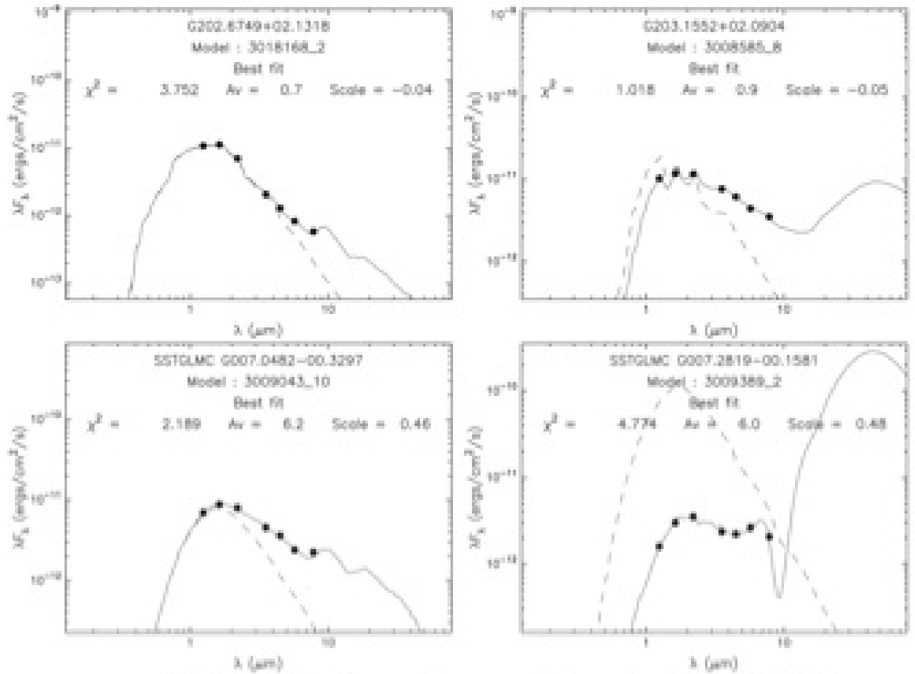
Spitzer's GLIMPSE survey
|b| < 1-2 deg

Benjamin+2003
Churchwell+2009

Massive Young Star-forming Complex Study in IR and X-ray
Feigelson+2013
Townsley+2014
Kuhn+2013ab,2014
Povich+2013
Broos+2013







SED fitting from Povich+2013

IRAC catalogs → Cross match (•2MASS •UKIDSS •VVV) → Color cuts → SED Fitting → Creation of training sets → Imputation → YSO Classifier

Gaia cross match → Remove duplicates → Cluster Analysis on Spatial Distributions → Validate groups using clustering in PM/Plx → Obtain mean astrometric properties of groups

THE ASTROPHYSICAL JOURNAL
SUPPLEMENT SERIES

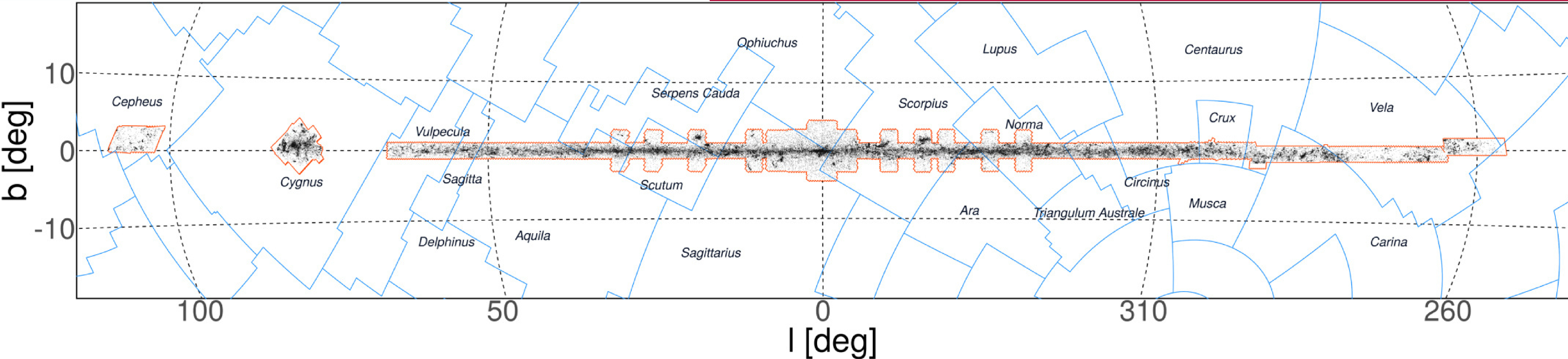
SPICY: The Spitzer/IRAC Candidate YSO Catalog for the Inner Galactic Midplane

Michael A. Kuhn[1], Rafael S. de Souza[2], Alberto Krone-Martins[3,4], Alfred Castro-Ginard[5], Emille E. O. Ishida[6], Matthew S. Povich[1,7], Lynne A. Hillenbrand[1], and for the COIN Collaboration

Published 2021 June 2 •
The Astrophysical Journal Supplement Series, Volume 254, Number 2
Citation Michael A. Kuhn *et al* 2021 *ApJS* **254** 33

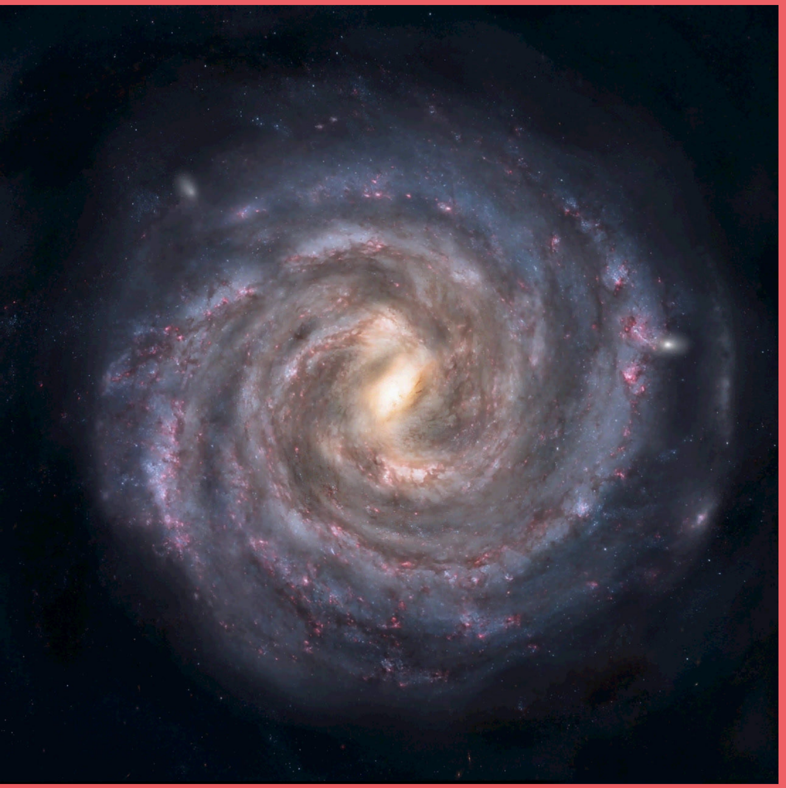
120,000 new YSOs

The SPICY catalog is the largest homogeneous sample of YSO candidates available to date for the inner regions of the Milky Way

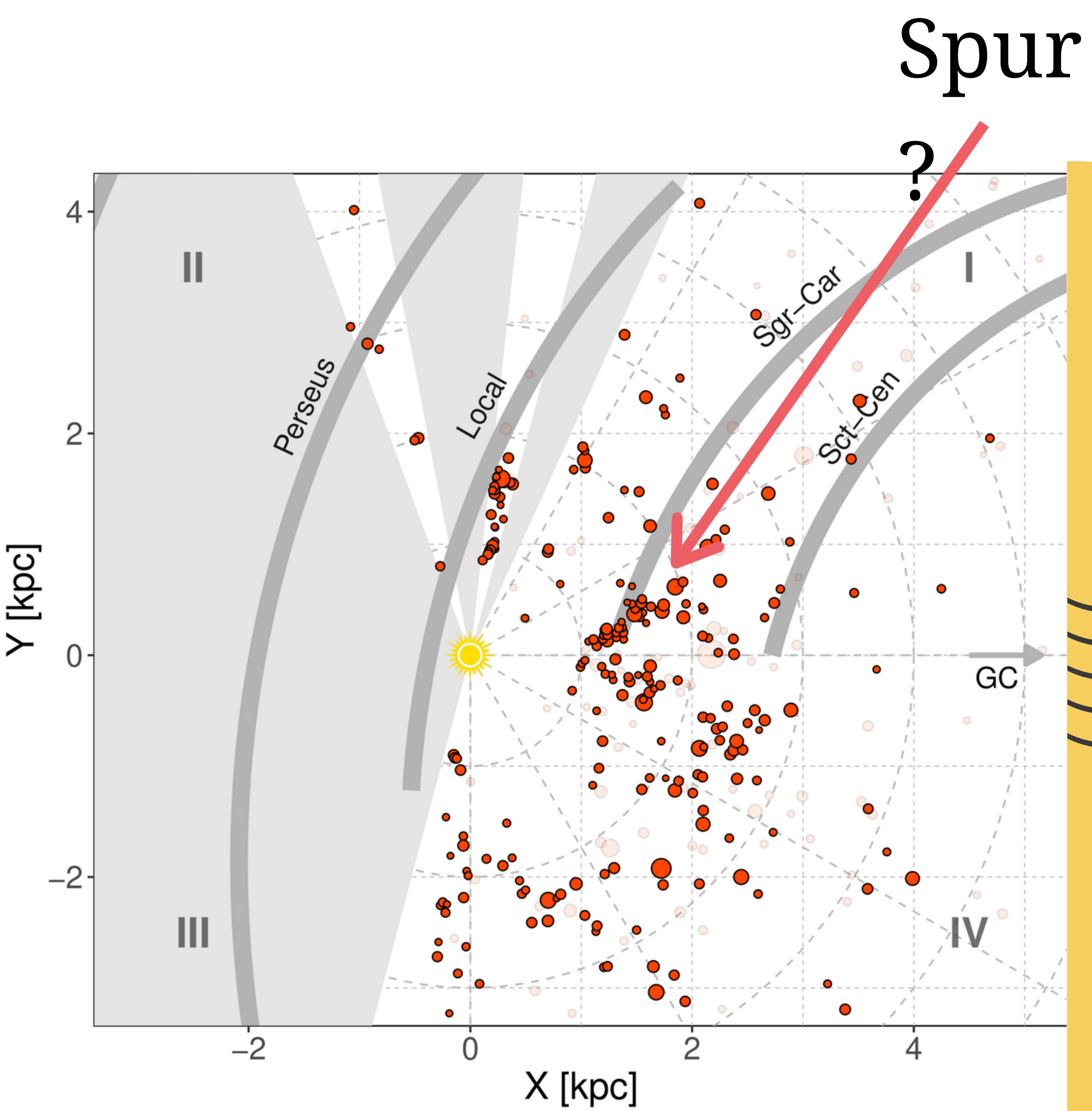


Spatial distribution of YSO groups

Good tracers of star forming regions and galactic structure



Independent probe of spiral arms structure



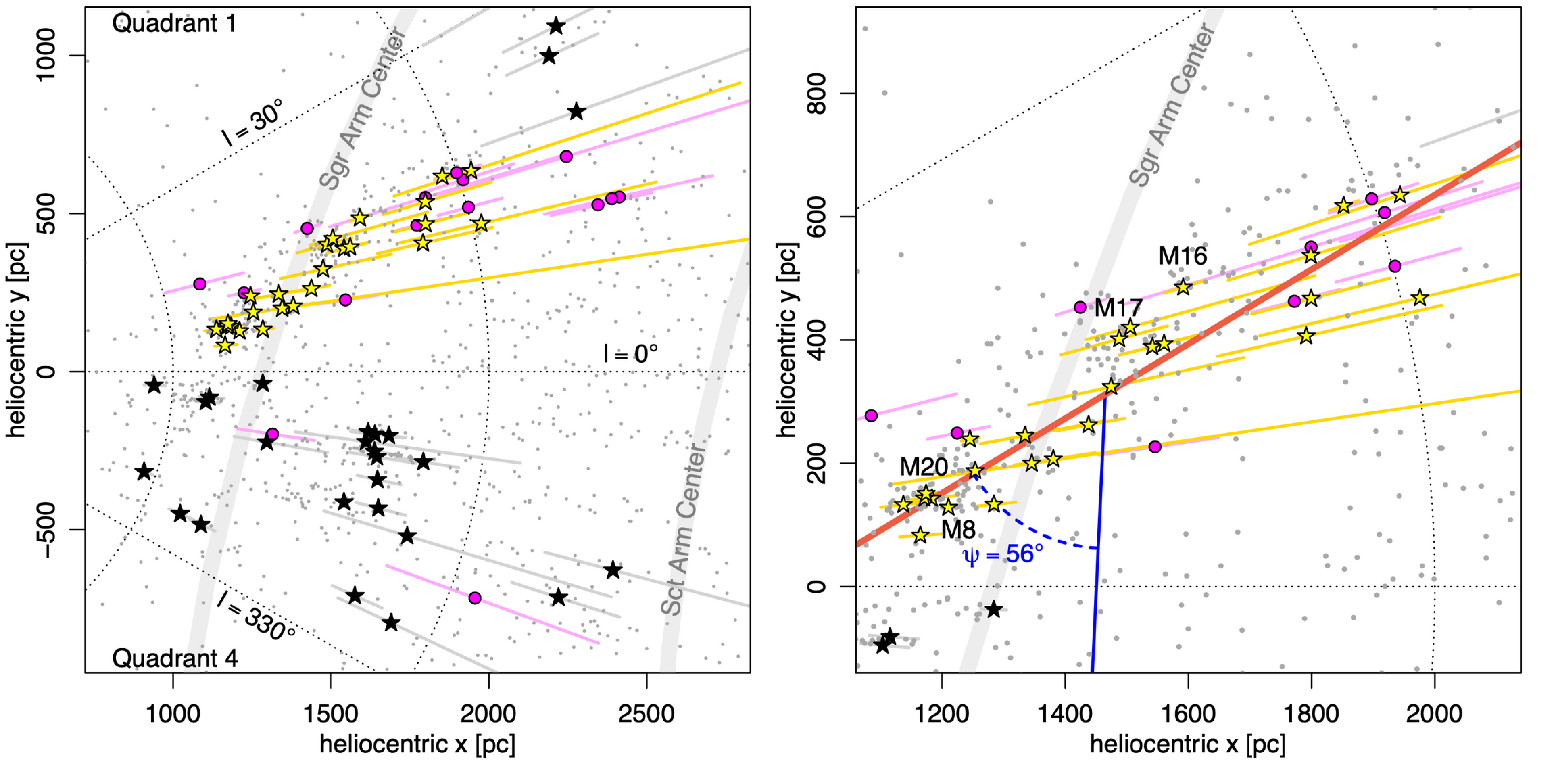
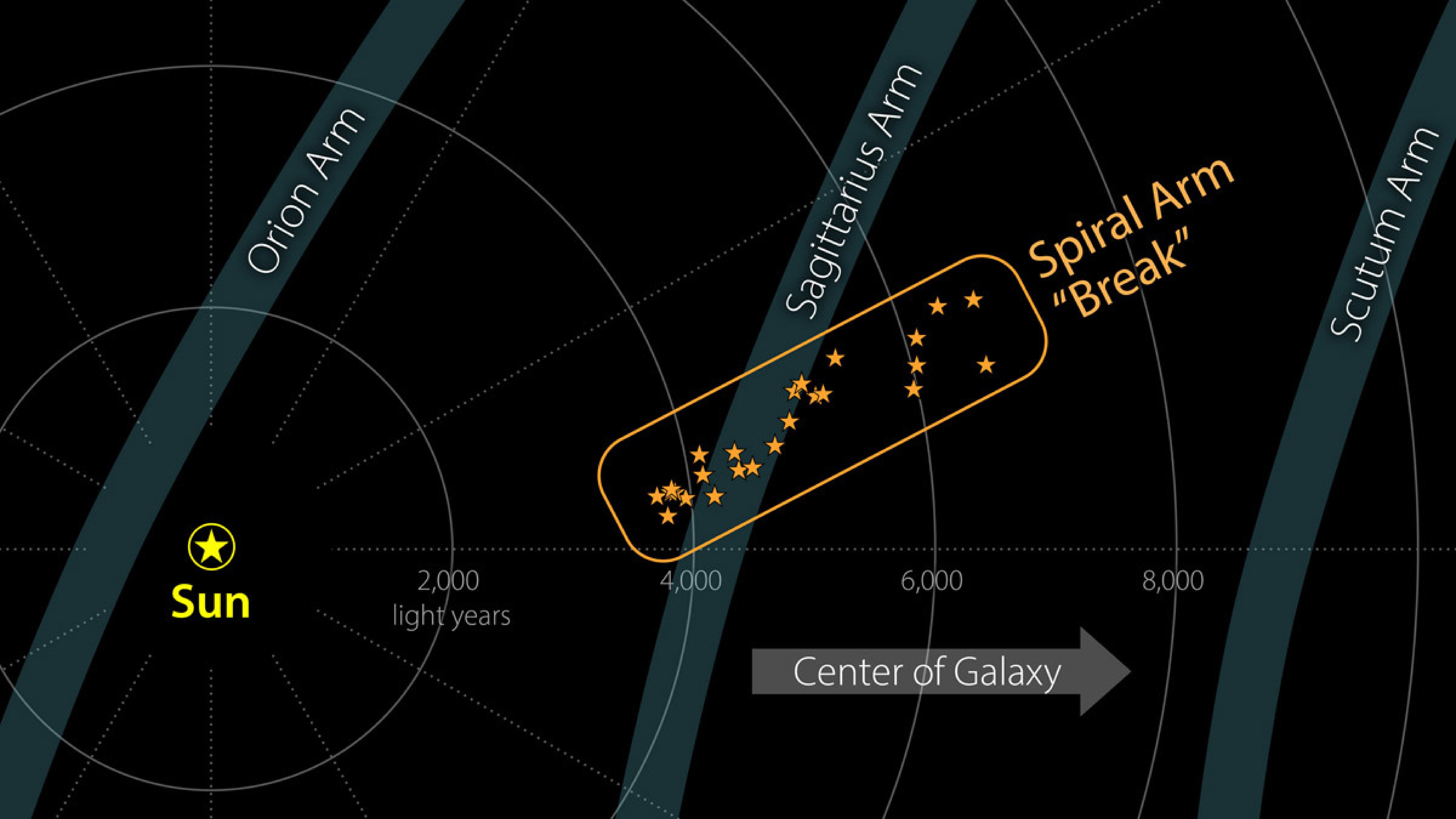
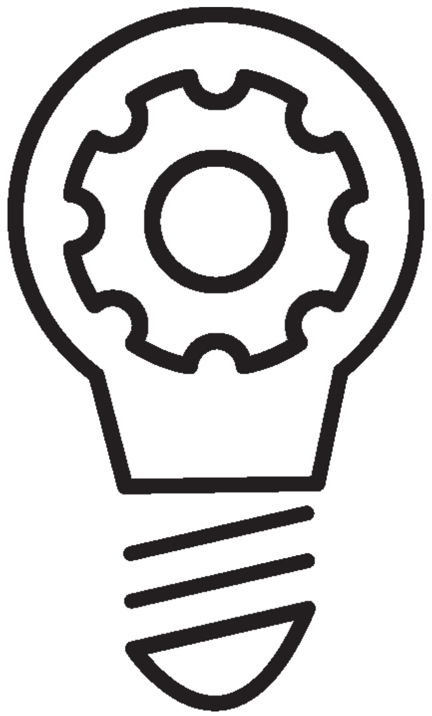


Fig. 3. Galactic map of YSO groups (star symbols), masers (magenta circles), and non-clustered SPICY YSO candidates (gray points) in heliocentric *xy* coordinates. The right panel shows a zoomed-in view. Groups associated with the structure are color-coded yellow, while others are black. The spiral-arm centers defined by Reid et al. (2019) are indicated by the grey bands. The red line indicates the major axis of the feature identified here with its 56° pitch angle illustrated in blue.


Orion Arm
Sagittarius Arm
Scutum Arm
Spiral Arm
"Break"
Sun
2,000
light years
4,000
6,000
8,000
Center of Galaxy

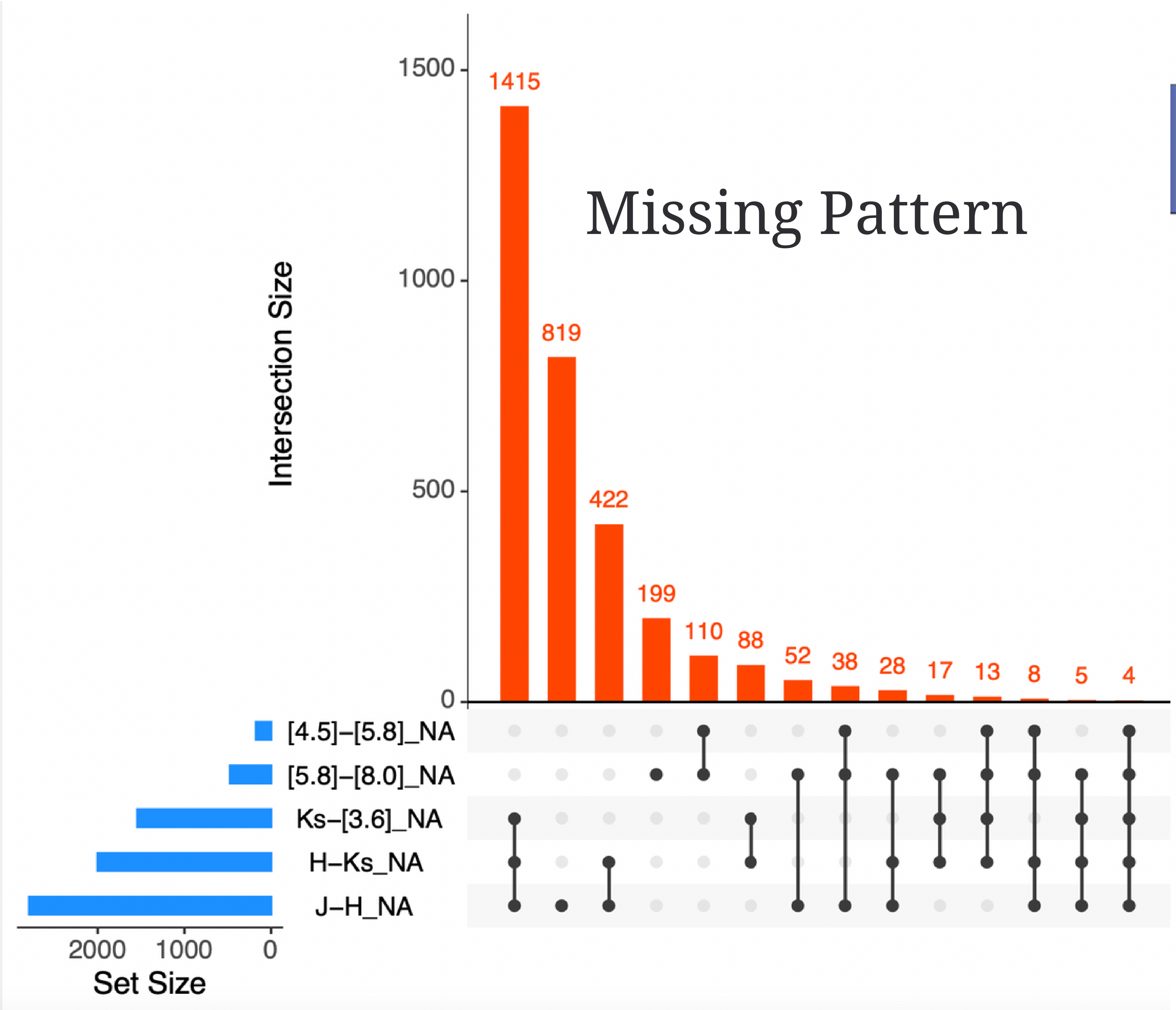
Some common data issues and possible approaches

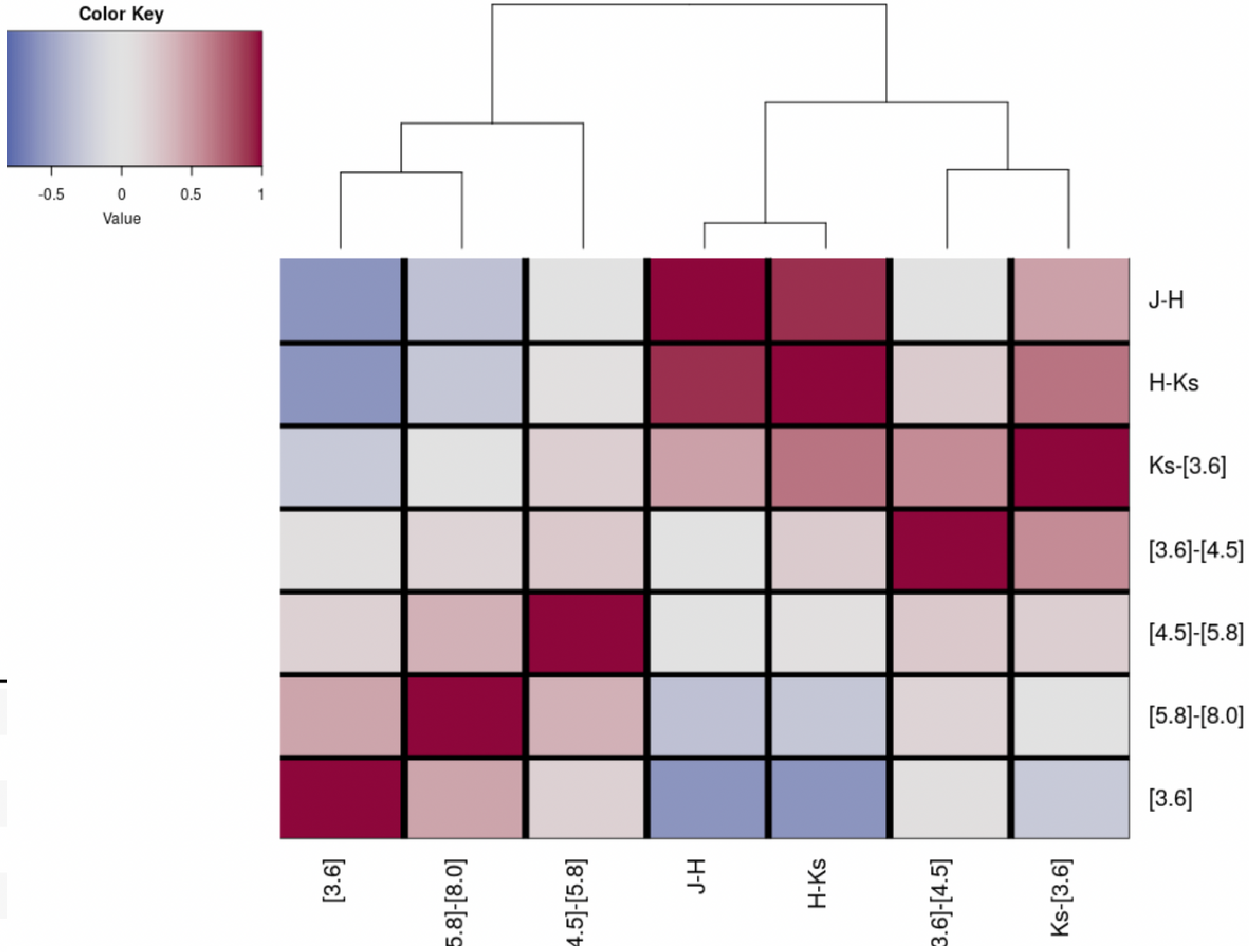


- Missing data - Imputation techiques
- Measurement Errors - Hierarchical Bayesian
- High-dimensionality - Principal Component Analysis
- Non-linear maps - Regression Tree models, etc.
- Non-Gaussianity - Generalized Linear Models

Missing data

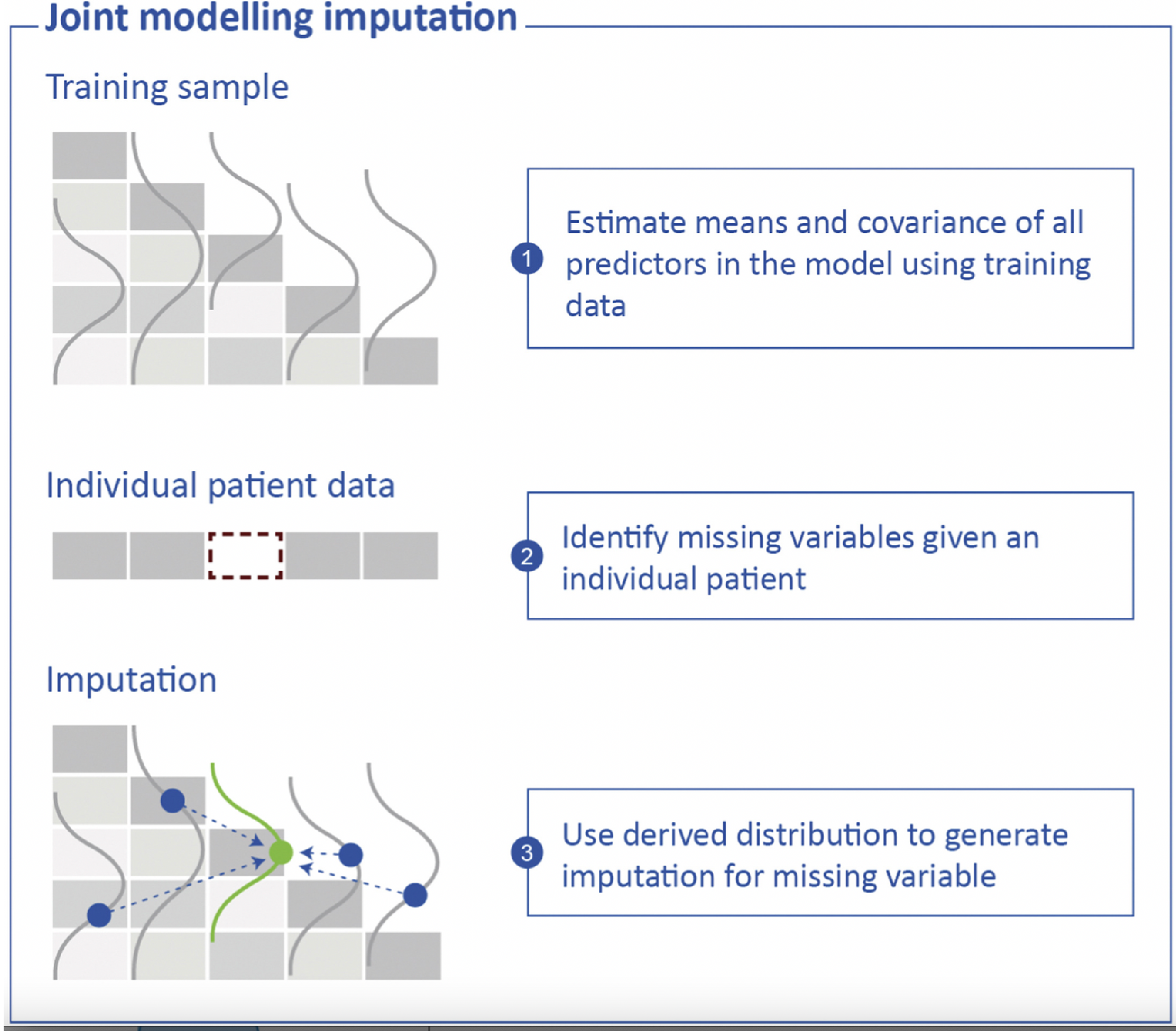
Data is correlated -> Find a joint distribution





Traditional approaches relying upon finding a joint distribution and sampling missing data from it.

Hard to find for high-dimensional cases, Multivariate assumption may not hold.

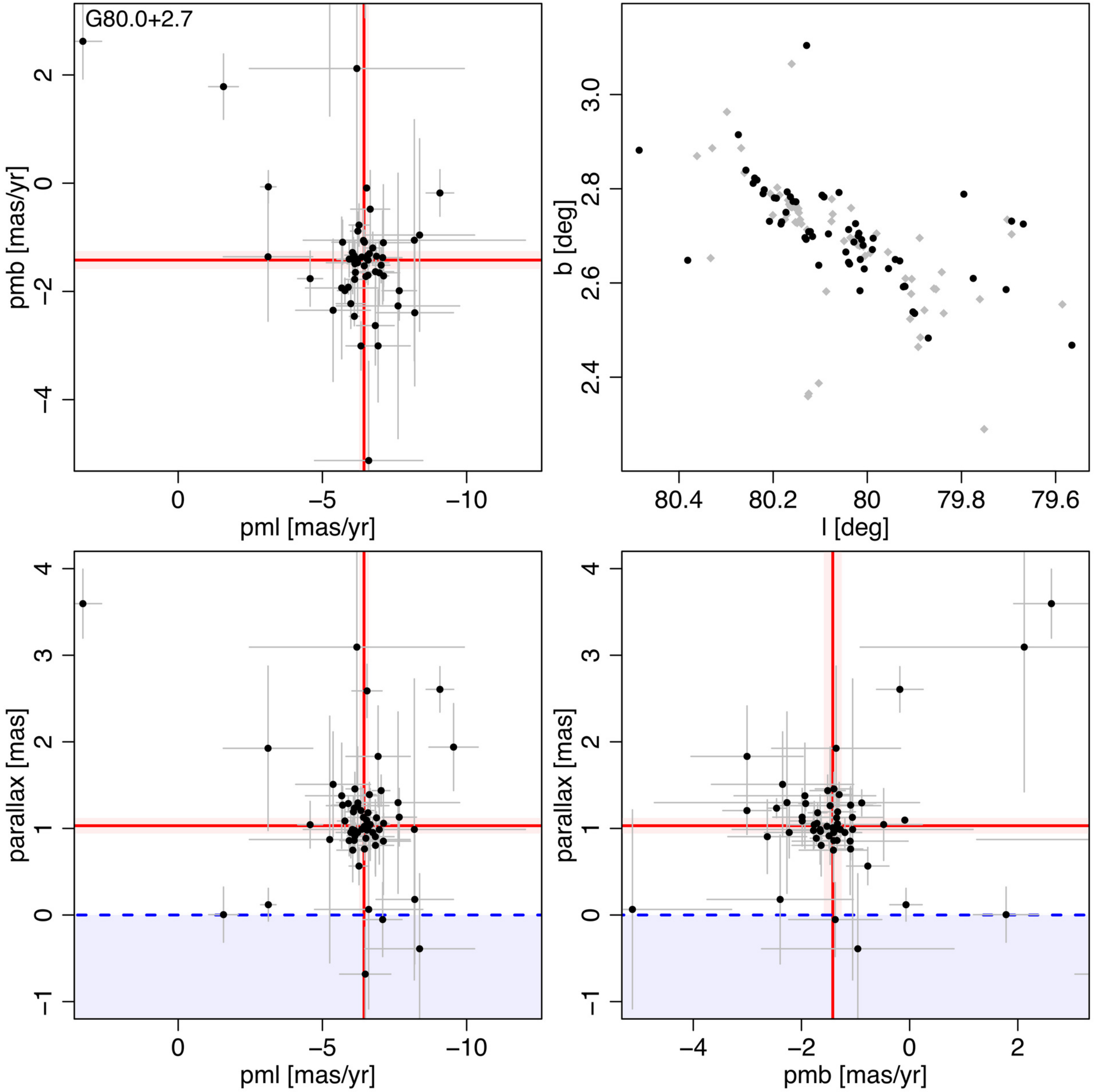


Hierarchical Bayesian Models

ASTROMETRIC PROPERTIES OF THE STELLAR GROUPS

$$\varpi_\mathrm{i} \sim \mathscr{T}(1/\mathrm{d}_\odot, \sigma^2_{\varpi_\mathrm{i}}, \nu),$$
$$\nu \sim \Gamma(2, 0.1),$$
$$d_\odot \sim \mathrm{Uniform}(0, 25),$$
$$i = 1 \ldots n_\mathrm{Gaia}$$

- Heteroscedastic measurement errors, outliers, non-normality, etc.
- Principled statistics still needed

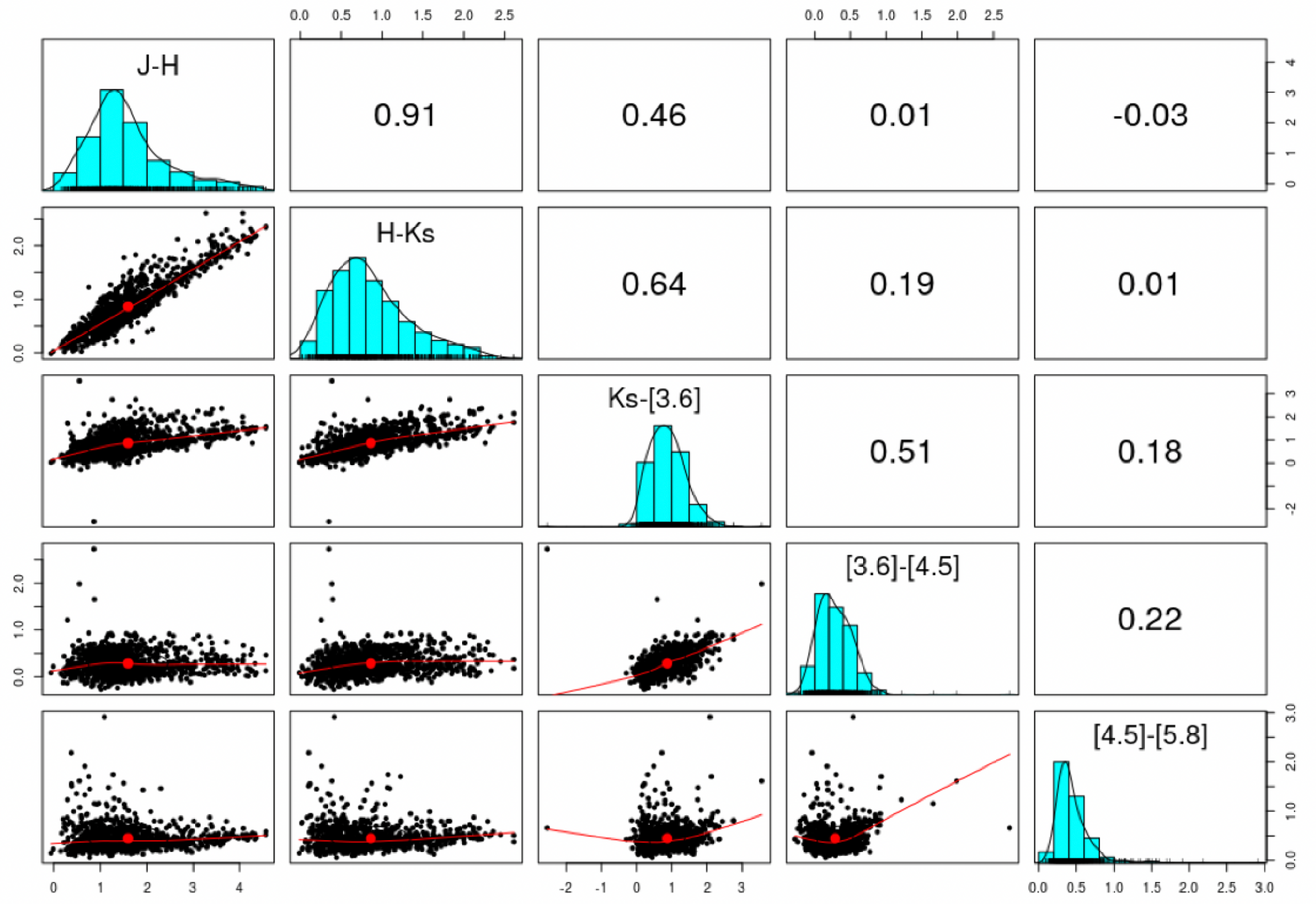


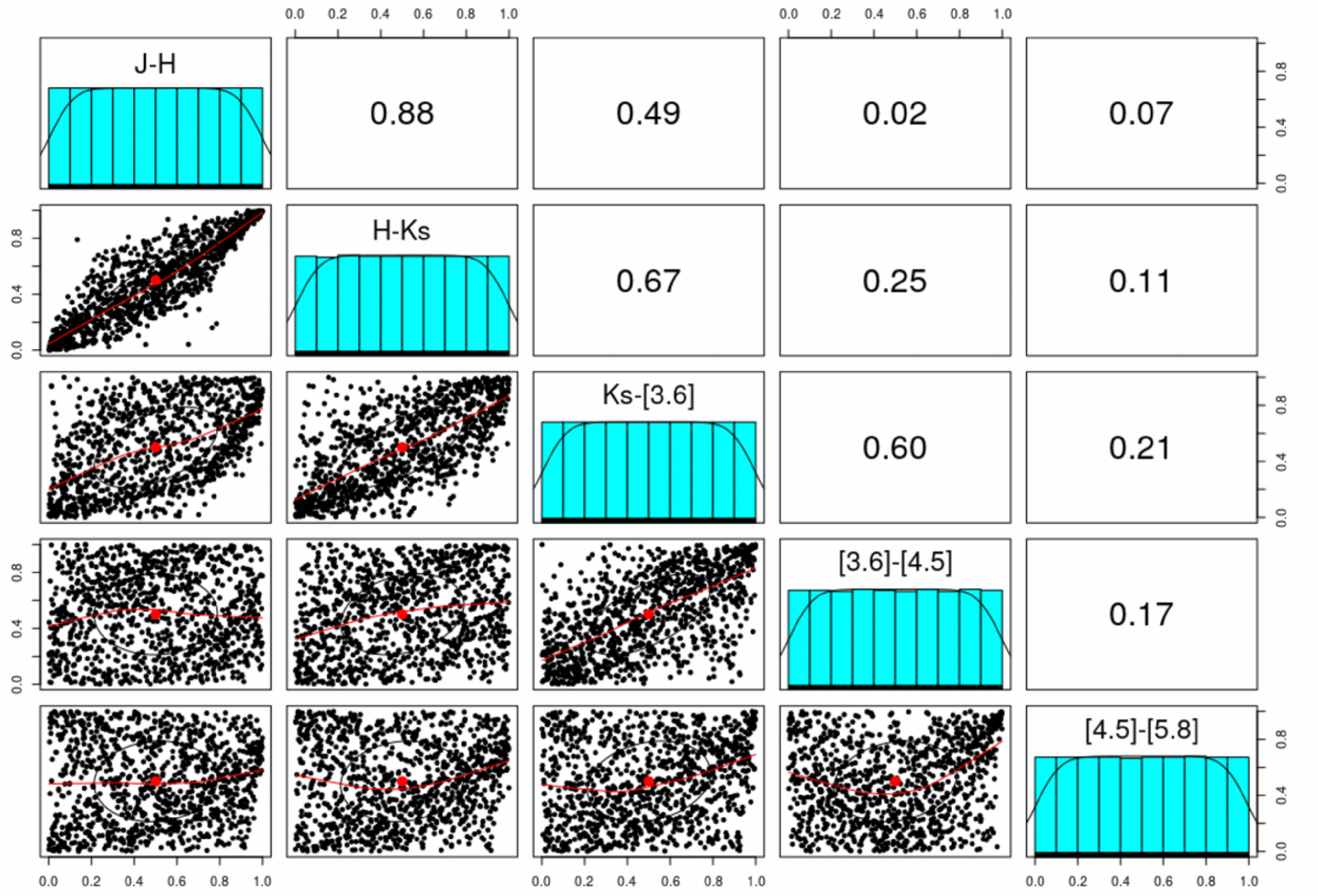
Multiple Imputation via Gaussian Copulas

Sklar's Theorem: *Let F be a p-dimensional joint distribution function with marginals F1, . . . , Fp. Then there exists a copula C with uniform marginals such that F(x1,...,xp) = C(F1(x1),...,Fp(xp)) ()*

Ann. Appl. Stat. 1(1): 265-283 (June 2007).
DOI: 10.1214/07-AOAS107

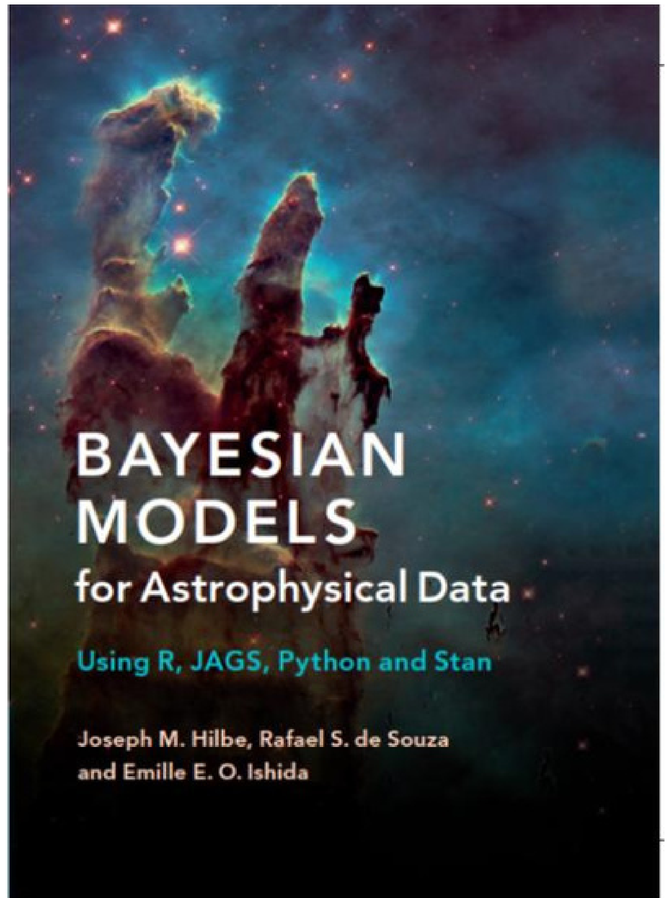
$$f(\mathbf{x}) = c(F_1(\mathbf{x}^1), \ldots, F_p(\mathbf{x}^p))\ \Pi_{j=1}^{p} f_j(\mathbf{x}^j)$$





Domain Specific Languages for Bayesian Analysis

- JAGS
- Stan
- Edward
- TensorFlow Probability
- Greta, ….




```
# Bayesian_plx_to_d
# Transform parallax into distance (kpc)
# INPUT:
# w - parallax (mas)
# errw - associated uncertainty (mas)
#
# Return heliocentric distance (kpc)
require(R2jags)

Bayesian_plx_to_d <- function(w,errw,nobs){
nobs = nobs
model.data <- list(w = w,                       # Parallax
                   errw = errw,                 # Error in Parallax
                   N = nobs)                    # Sample size
NORM <- "
model{
# Likelihood
for (i in 1:N){
 w[i] ~ dt(1/d,pow(errw[i],-2),nu)
 }
#nu <- nuMinusOne + 1
#numMinusOne ~ dexp(1/29)
nu ~ dgamma(2,0.1)

# Weakly uniform prior for distance
d ~ dunif(0,25)

plx <- 1/d

            }"
params <- c("d","plx")
```

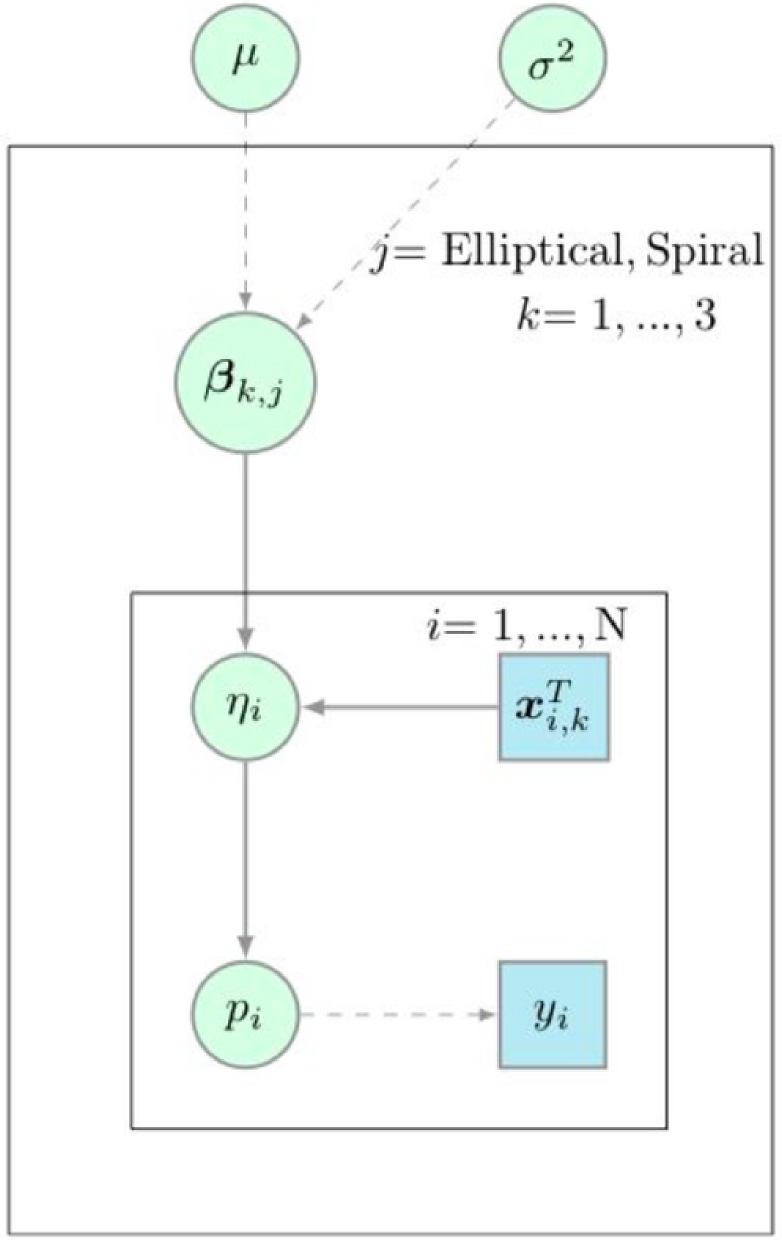
JAGS (Just Another Gibbs Sample) language

Enables to create Hierarchical Bayesian Models for general regression purposed quite fast.

Is the cluster environment quenching the Seyfert activity in elliptical and spiral galaxies?

R. S. de Souza[1★], M. L. L. Dantas[2], A. Krone-Martins[3], E. Cameron[4], P. Coelho[2], M. W. Hattab[5], M. de Val-Borro[6], J. M. Hilbe[7], J. Elliott[8] and A. Hagen[9], for the COIN Collaboration

```
#Model
jags_model<-"model{
#Shared hyperpriors for beta
tau ~ dgamma(1e-3,1e-3) #Precision
mu  ~ dnorm(0,1e-3)      #mean
#Diffuse prior for beta
for(j in 1:2){
for(k in 1:3){
beta[k,j]~dnorm(mu,tau)
}}
# Likelihood
for(i in 1:N){
Y[i] ~ dbern(pi[i])
logit(pi[i]) <- eta[i]
eta[i] <- beta[1,gal[i]]*X[i,1]+
beta[2,gal[i]]*X[i,2]+
beta[3,gal[i]]*X[i,3]
}}"
```

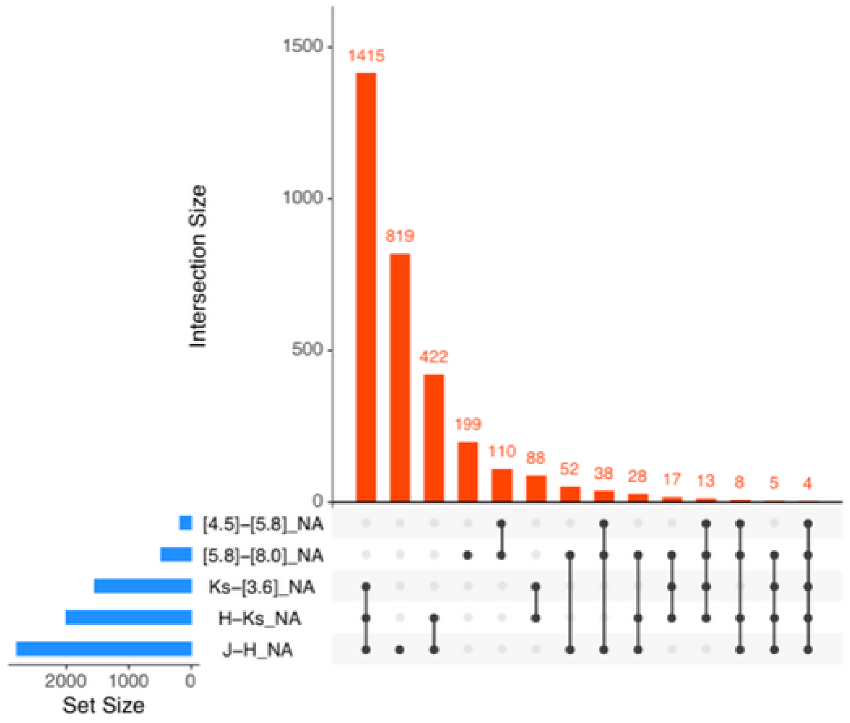


$$\mathrm{y_i} \sim \mathrm{Bernoulli}(\mathrm{p_i})$$
$$\mathrm{logit}(\mathrm{p_i}) = \eta_\mathrm{i}$$
$$\eta_i = \boldsymbol{x}^T_{i,k}\beta_{k,j}$$
$$\boldsymbol{x}^T_{i,k} = \begin{pmatrix} 1 & (\log M_{200})_1 & \left(\frac{r}{r_{200}}\right)_1 \\ \vdots & \vdots & \vdots \\ 1 & (\log M_{200})_N & \left(\frac{r}{r_{200}}\right)_N \end{pmatrix}$$
$$\beta_{k,j} \sim \mathrm{Normal}(\mu, \sigma^2)$$
$$\mu \sim \mathrm{Normal}(0, 10^3)$$
$$\tau \sim \mathrm{Gamma}(10^{-3}, 10^{-3})$$
$$\sigma^2 = 1/\tau$$
$$j = \mathrm{Elliptical, Spiral}$$
$$k = 1, \ldots, 3$$
$$i = 1, \ldots, N$$

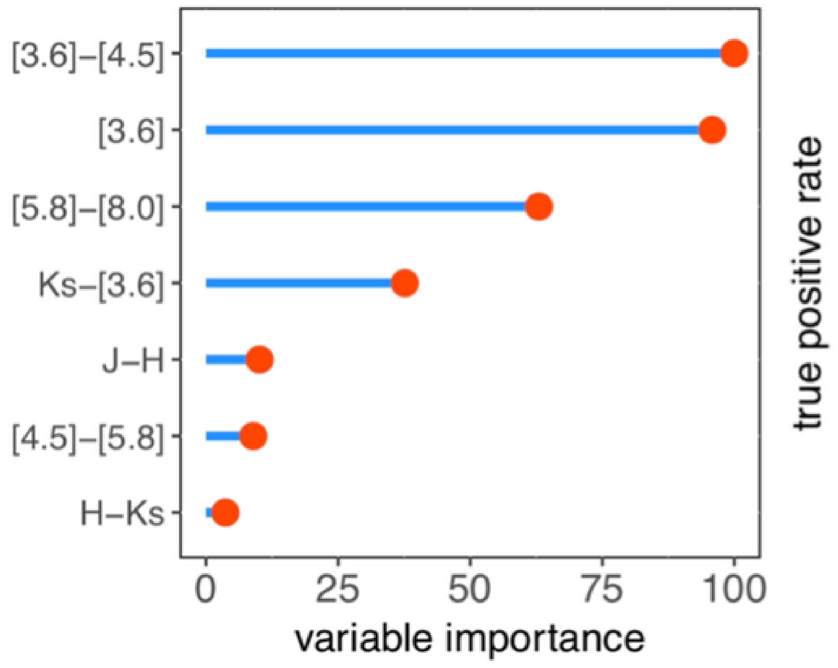
Classification scheme

"traditional" SED fitting to weed out reddened stars → generate YSO and non-YSO training sets by cross-matching with previous studies → copula imputation of missing data; MCMC using SBGCOP (Hoff 2018) → Random forest

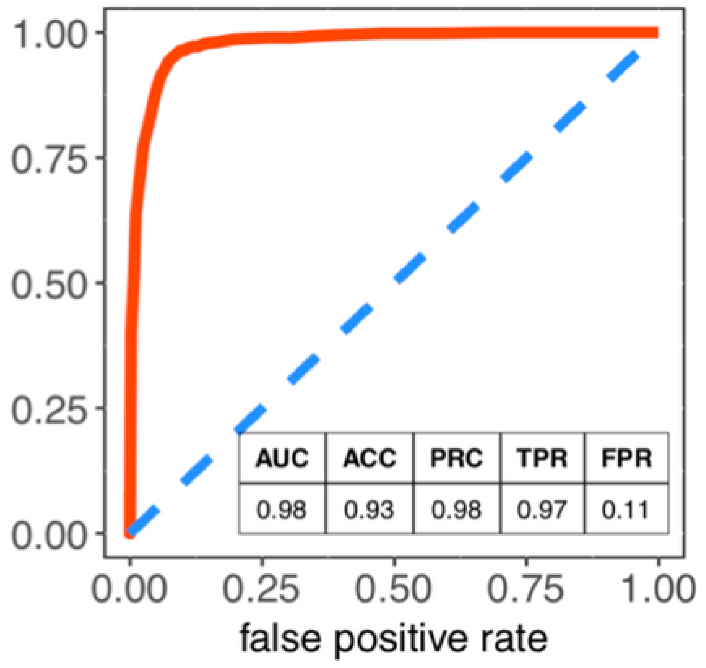
50 million mid-IR sources

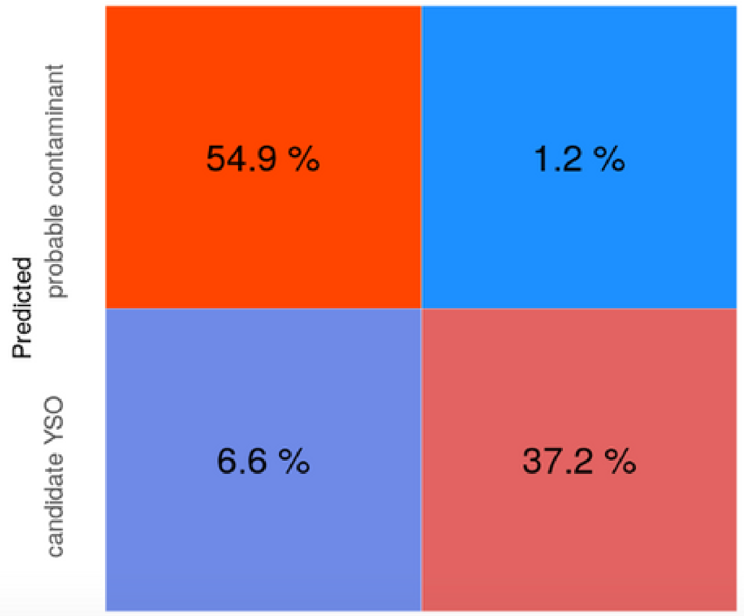


117,446 YSO candidates



Classifier performance





THE ASTROPHYSICAL JOURNAL
SUPPLEMENT SERIES

SPICY: The Spitzer/IRAC Candidate YSO Catalog for the Inner Galactic Midplane

Michael A. Kuhn[1], Rafael S. de Souza[2], Alberto Krone-Martins[3,4], Alfred Castro-Ginard[5], Emille E. O. Ishida[6], Matthew S. Povich[1,7], Lynne A. Hillenbrand[1], and for the COIN Collaboration

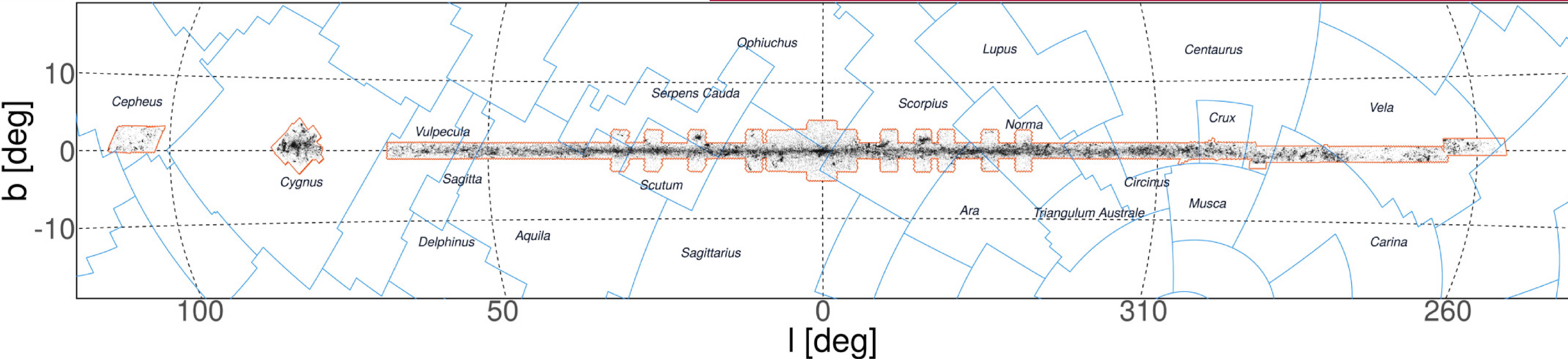
Published 2021 June 2 •

The Astrophysical Journal Supplement Series, Volume 254, Number 2

Citation Michael A. Kuhn *et al* 2021 *ApJS* **254** 33

120,000 new YSOs

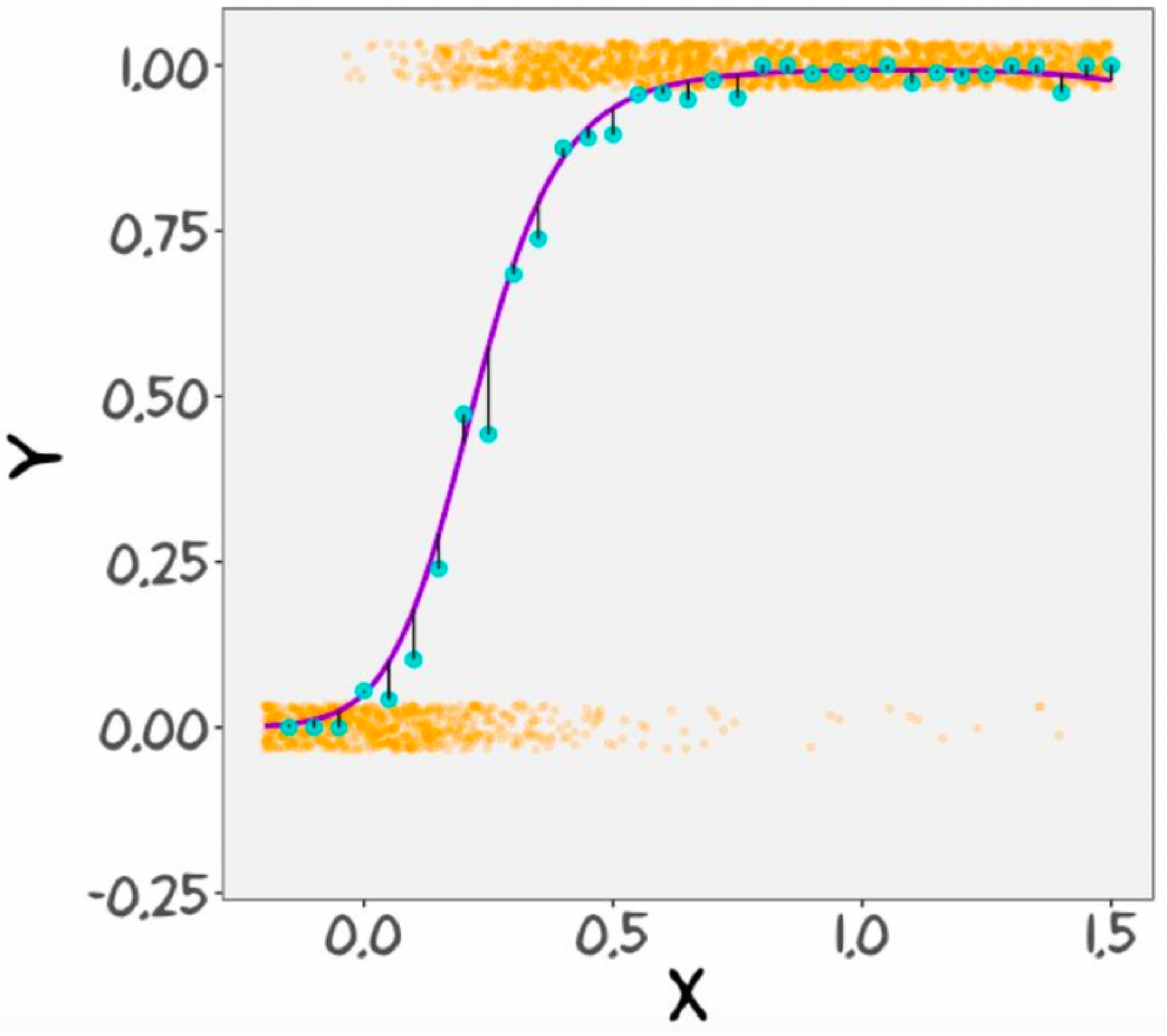
The SPICY catalog is the largest homogeneous sample of YSO candidates available to date for the inner regions of the Milky Way



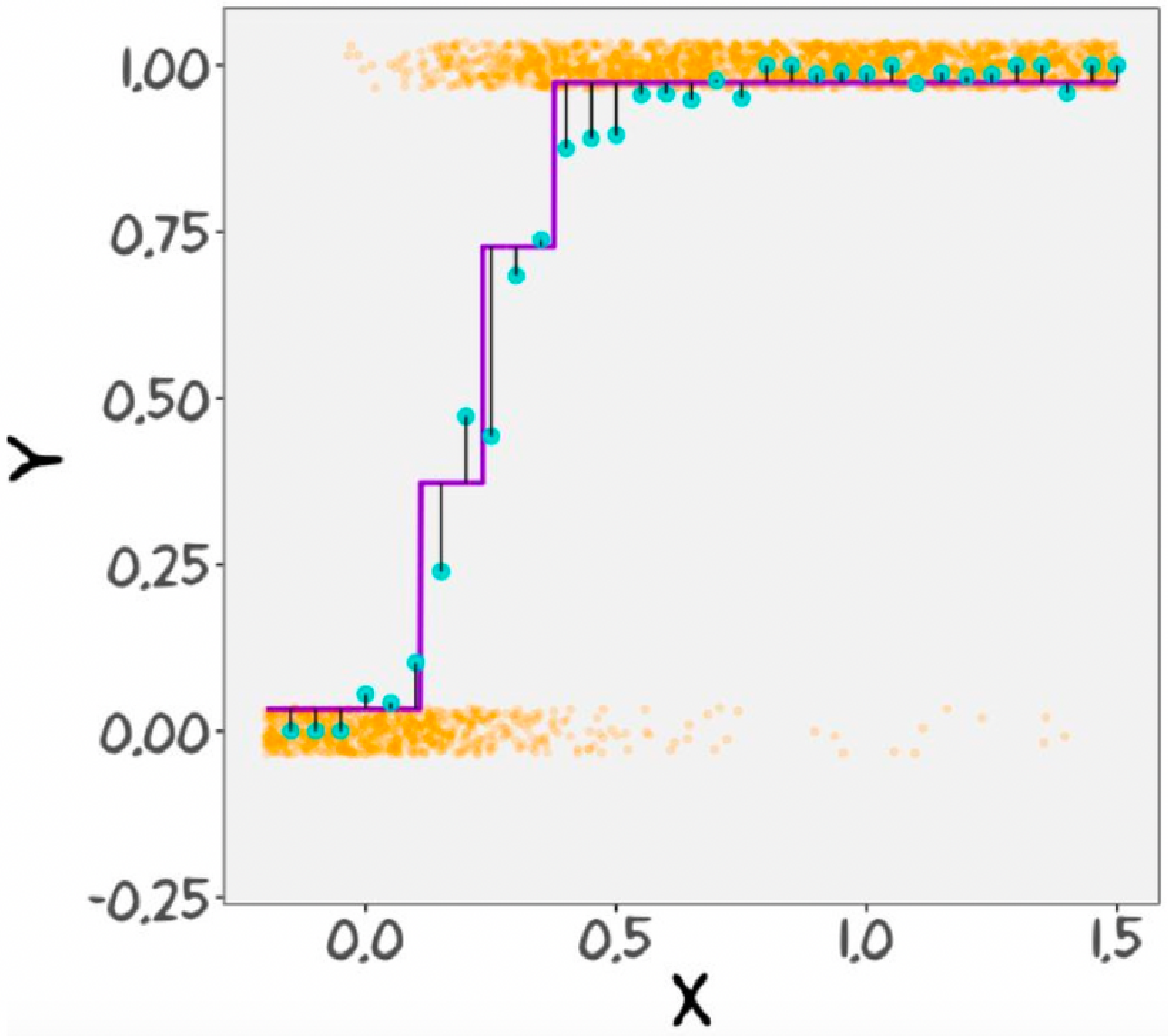
Logistic regression "vs" Decision Tree

Caveat: You get what you ask for

Logistic Regression

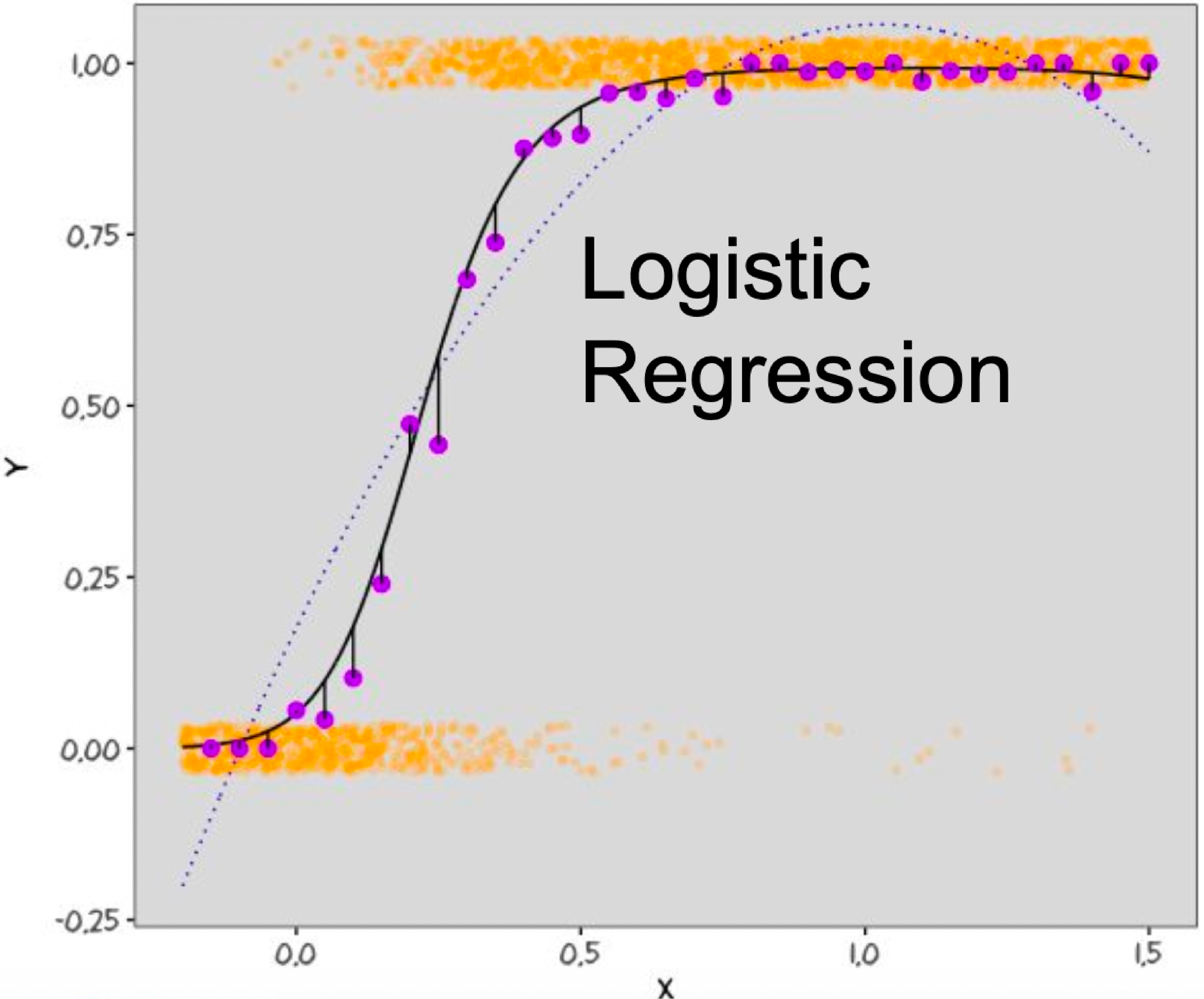


Decision Tree

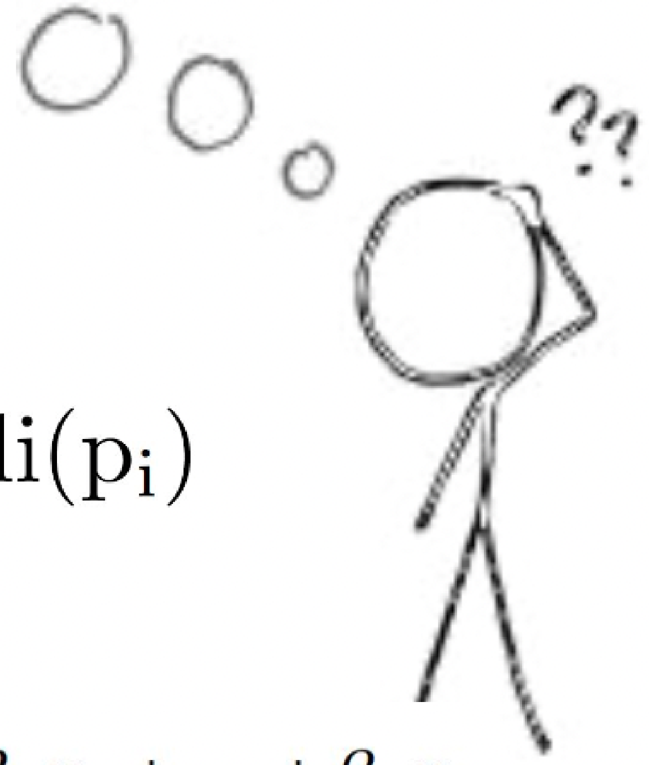


Gaussian Models

Limitations



Binary data



$$y_i \sim \mathrm{Bernoulli}(p_i)$$

$$\mathrm{logit}(p_i) = \eta_i$$

$$\eta_i \equiv \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

ACCEPTED MANUSCRIPT

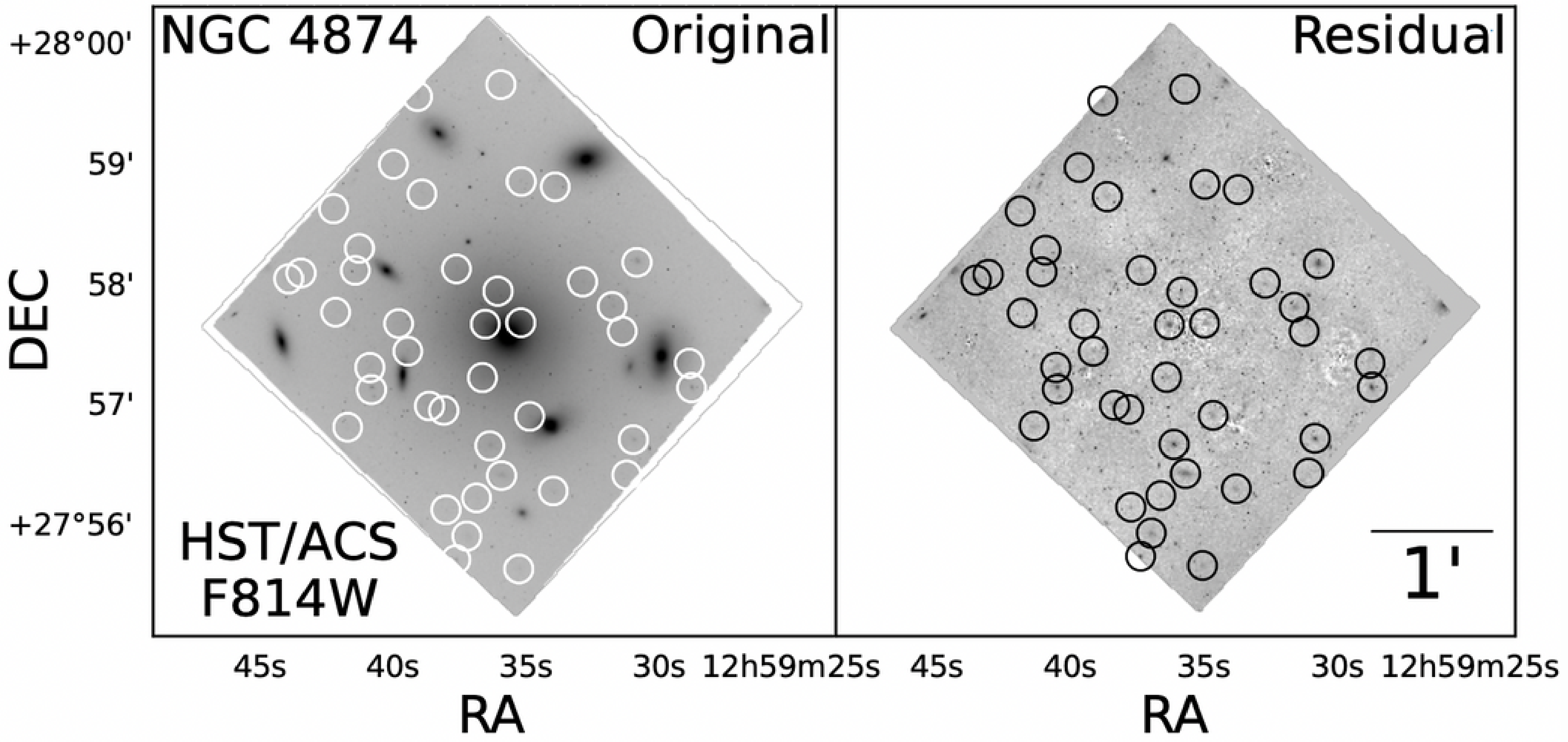
A high occurrence of nuclear star clusters in faint Coma galaxies, and the roles of mass and environment

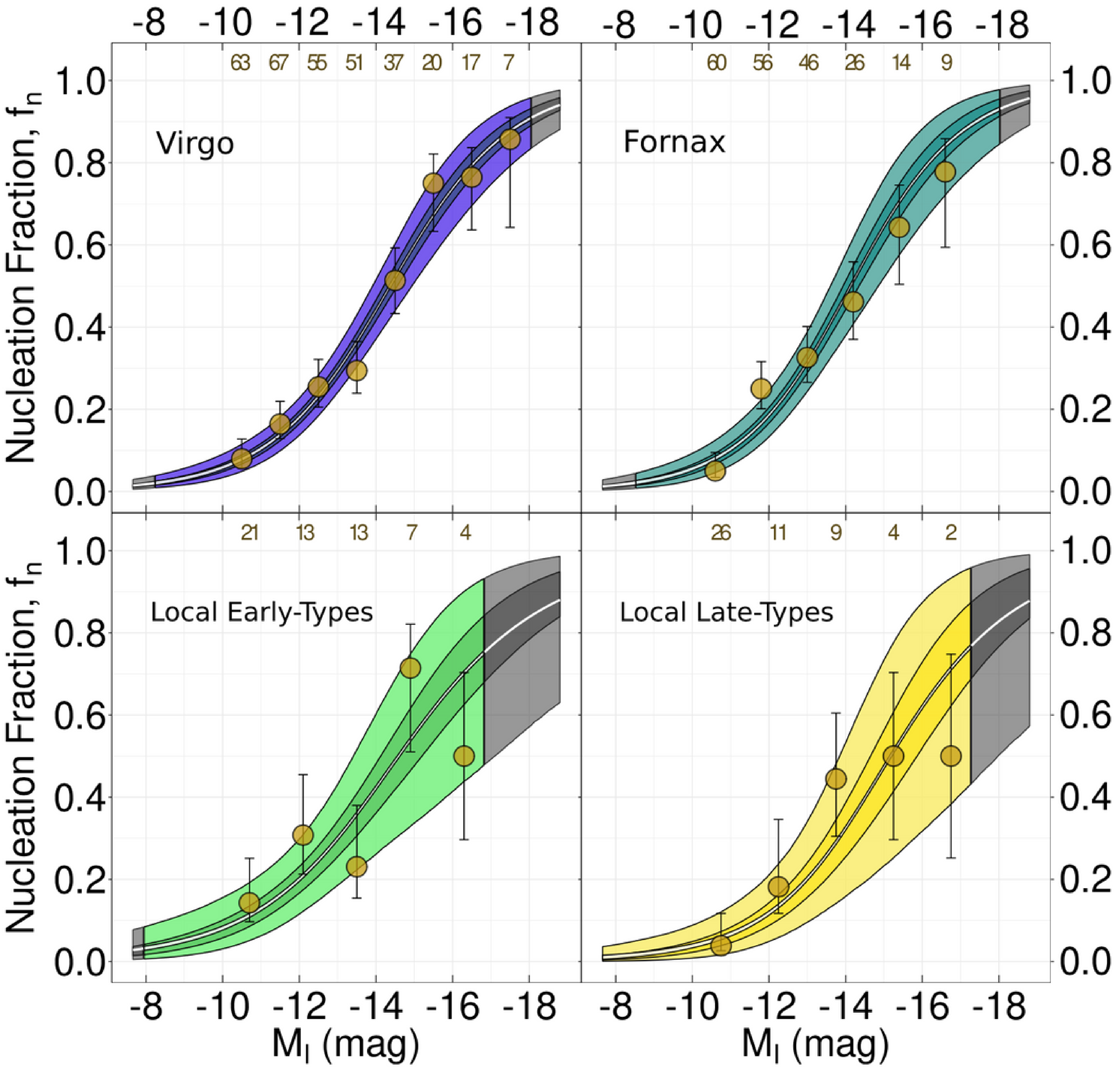
Emílio Zanatta ✉, Rubén Sánchez-Janssen, Ana L Chies-Santos,
Rafael S de Souza, John P Blakeslee

Monthly Notices of the Royal Astronomical Society, stab2348,
https://doi.org/10.1093/mnras/stab2348
Published: 23 August 2021

Logistic Regression

NSCs are the densest known star clusters in the Universe. With apparent magnitudes between -14 and -10 mag in the infrared, they are on average 40 times brighter than globular clusters, although their effective radii are not larger than 2 to 5 parsecs.





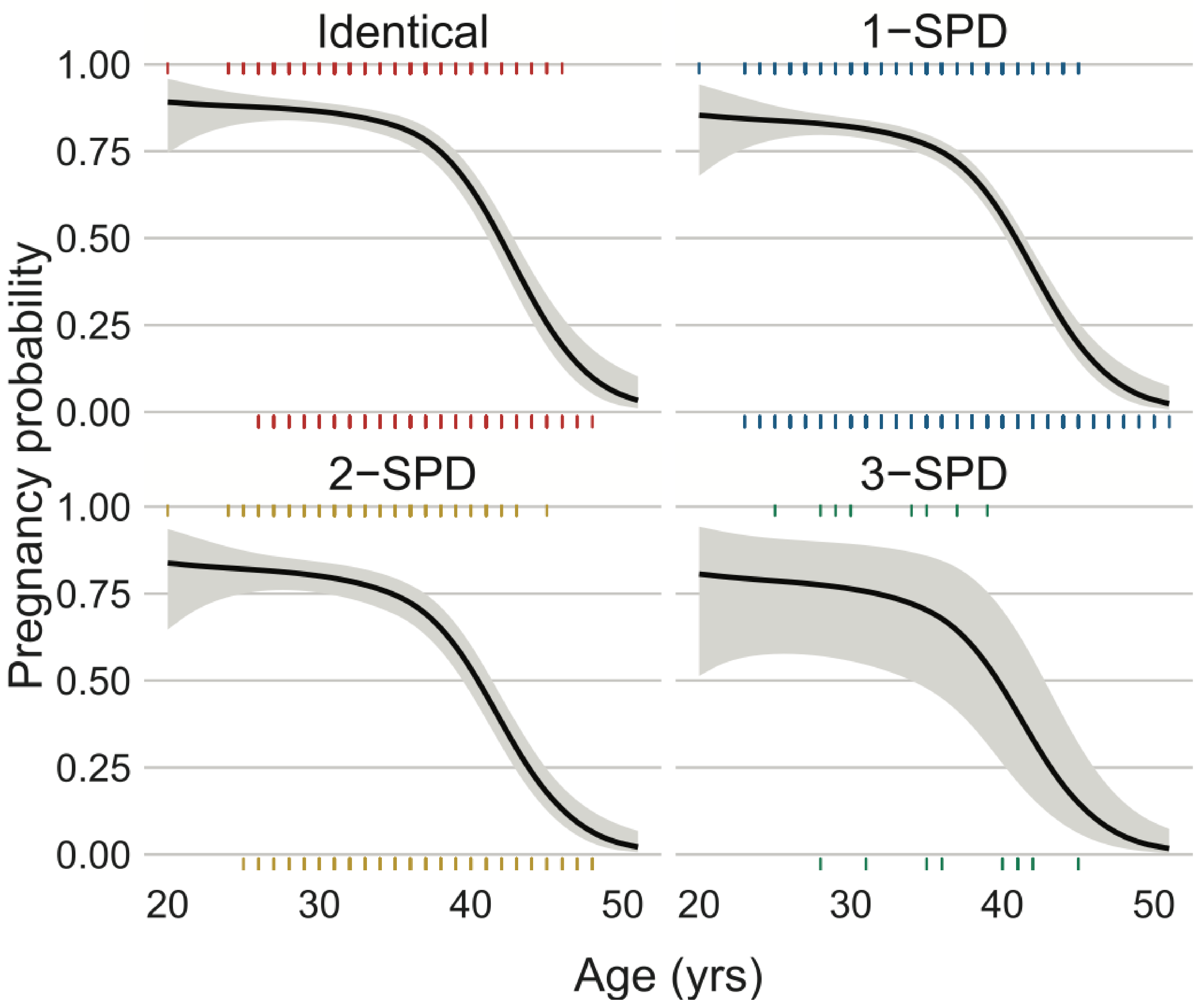
Statistical Methods in Medical Research

Fallopian tube anatomy predicts pregnancy and pregnancy outcomes after tubal reversal surgery

Rafael S de Souza, Gary S Berger

First Published July 7, 2021 | Research Article | Find in PubMed | Check for updates
https://doi.org/10.1177/09622802211023543

Article information

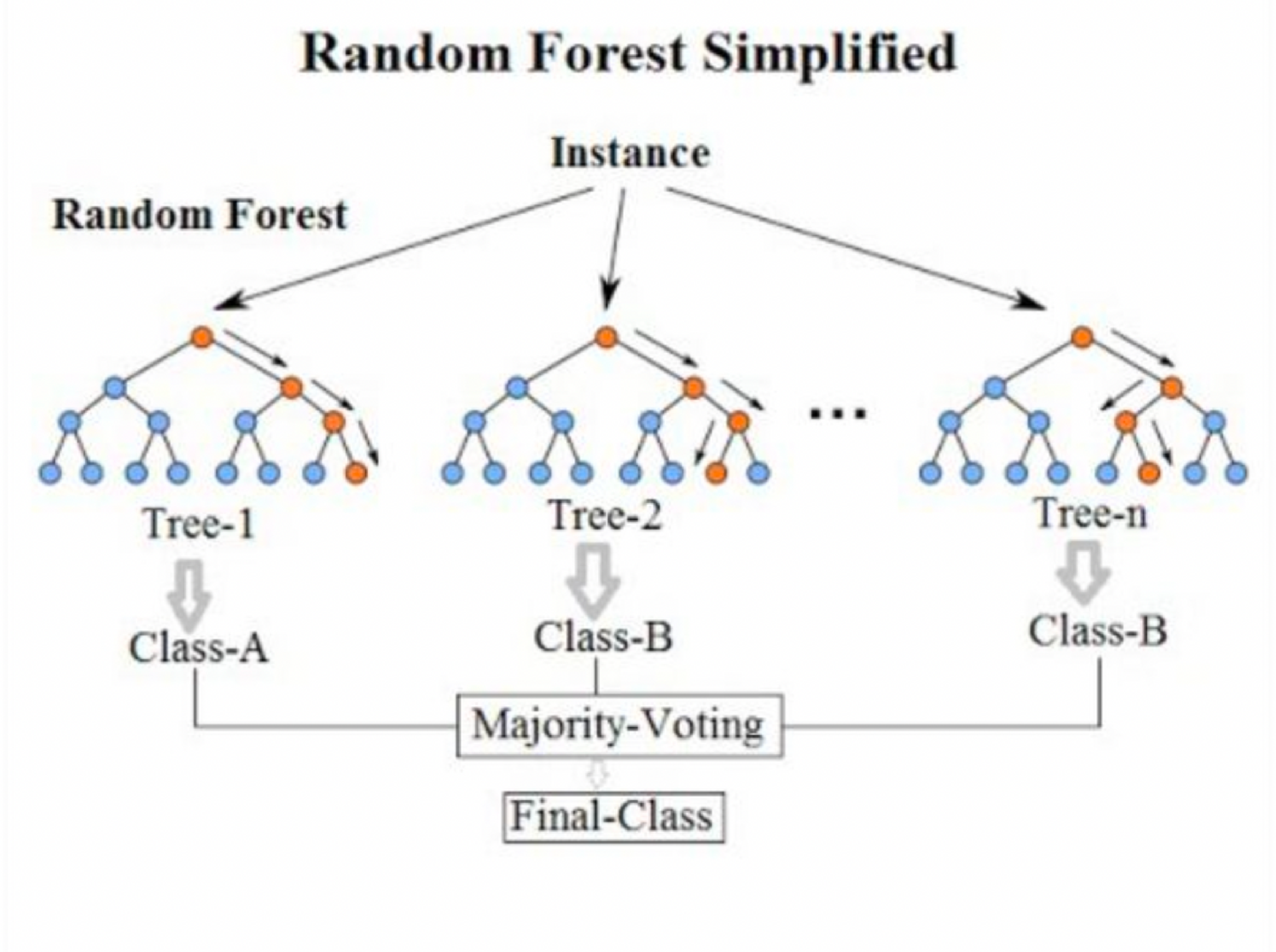


Logistic Regression again

Example of supervised ML algorithm for classification

Random Forests

Ensemble method

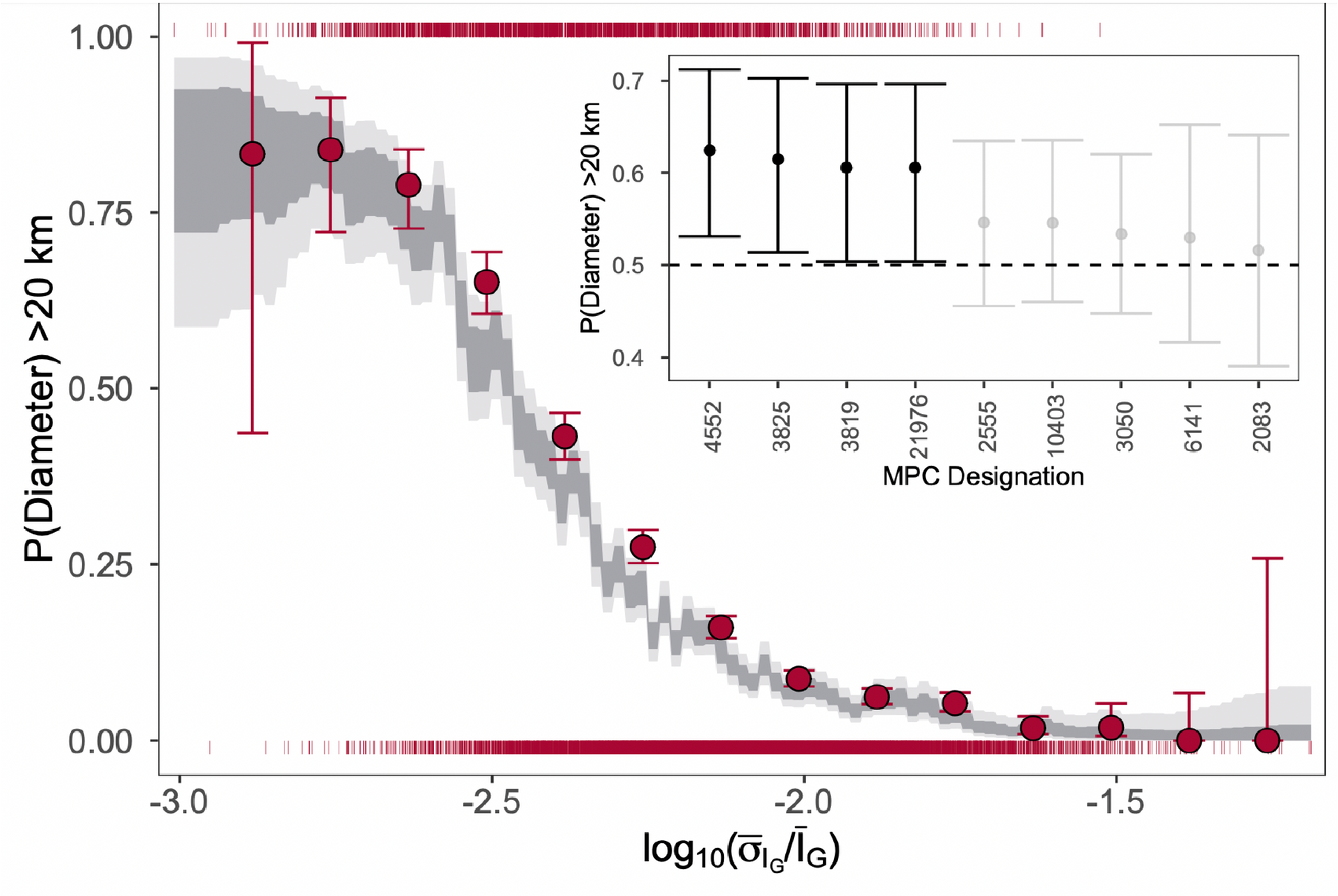


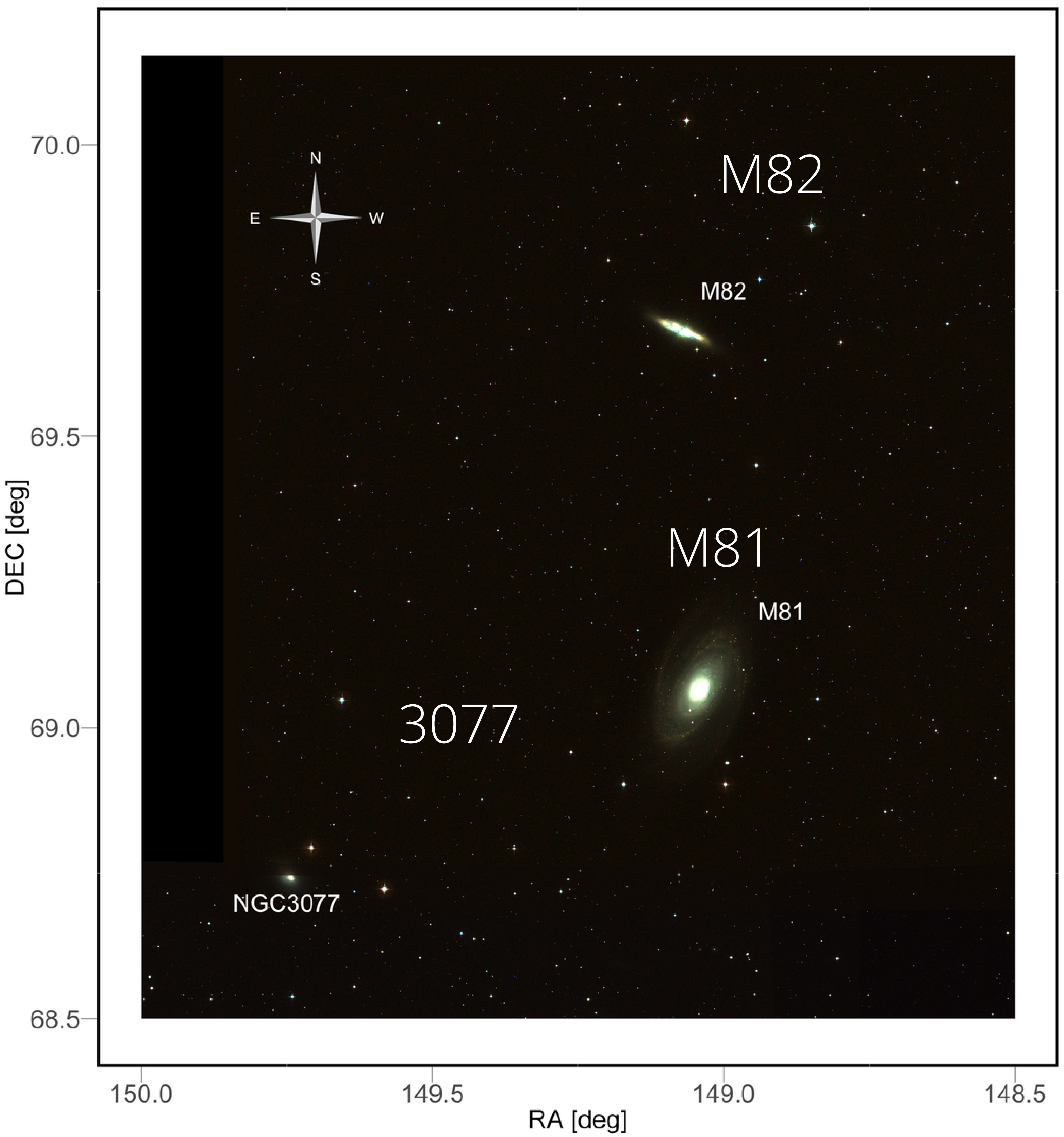
https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

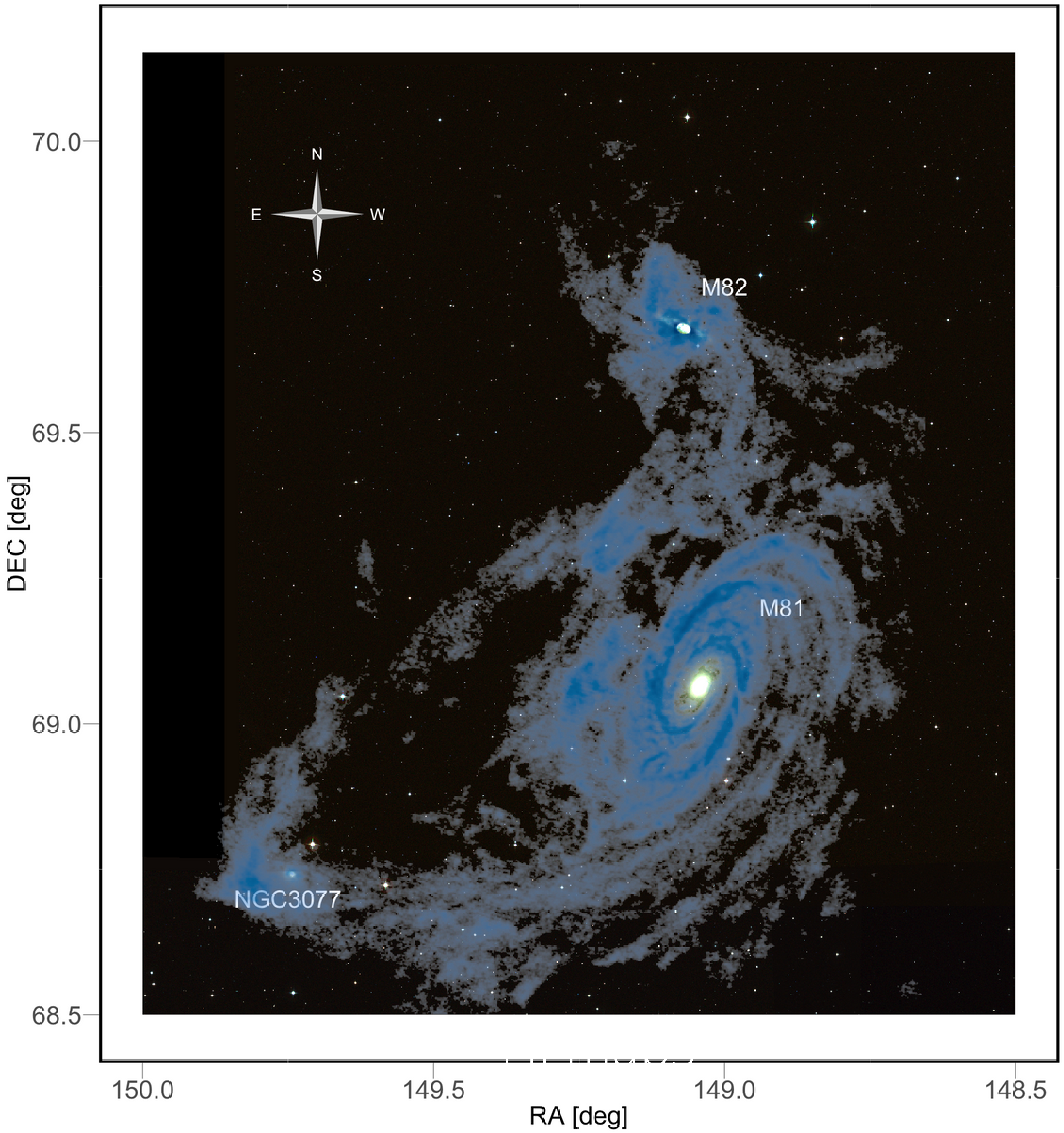
Probabilistic modeling of asteroid diameters from Gaia DR2 errors

Rafael S. de Souza [1] Alberto Krone-Martins [2,3] Valerio Carruba [4] Rita de Cassia Domingos [5] E. E. O. Ishida [6] Safwan Alijbaae [7] Mariela Huaman Espinoza [8] and William Barletta [4]

Logistic Bayesian Additive Regression Trees

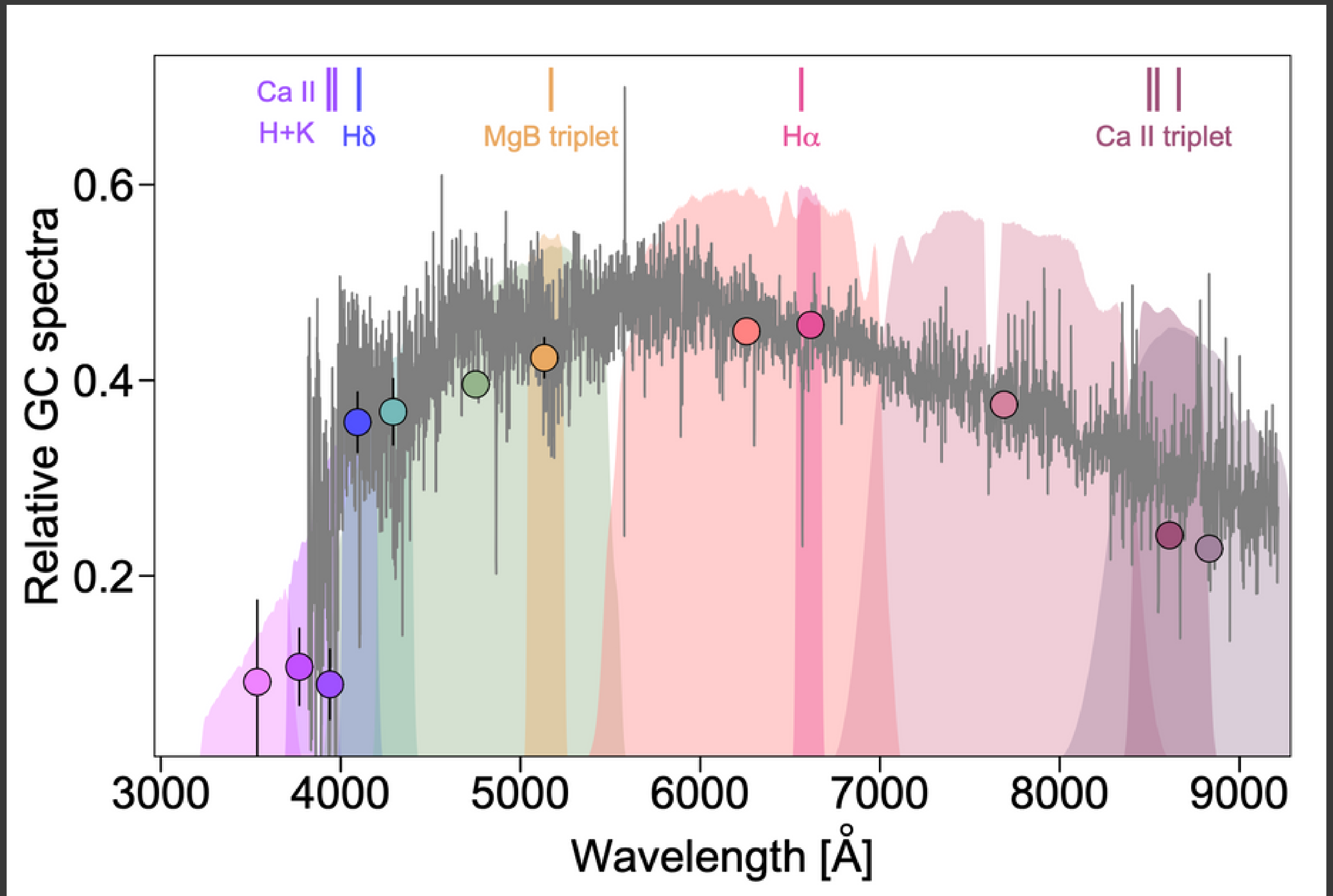


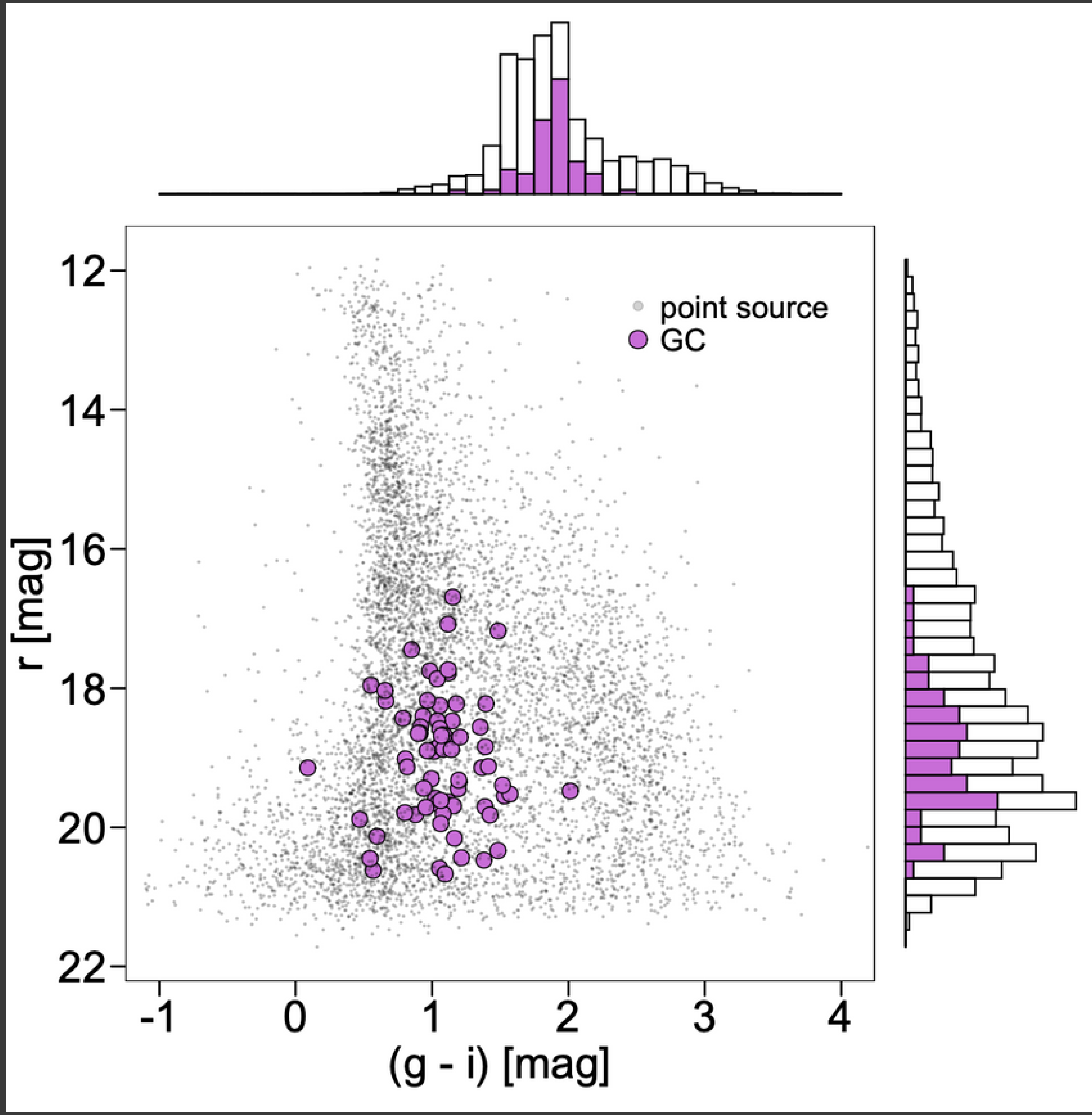

M82
M82
M81
M81
3077
NGC3077
N
E
W
S
DEC [deg]
RA [deg]
70.0
69.5
69.0
68.5
150.0
149.5
149.0
148.5


M82
M81
NGC3077
N
E
W
S
DEC [deg]
RA [deg]
70.0
69.5
69.0
68.5
150.0
149.5
149.0
148.5

J-PLUS PHOTOMETRY

For 7.2K point sources plus 73 confirmed GCs





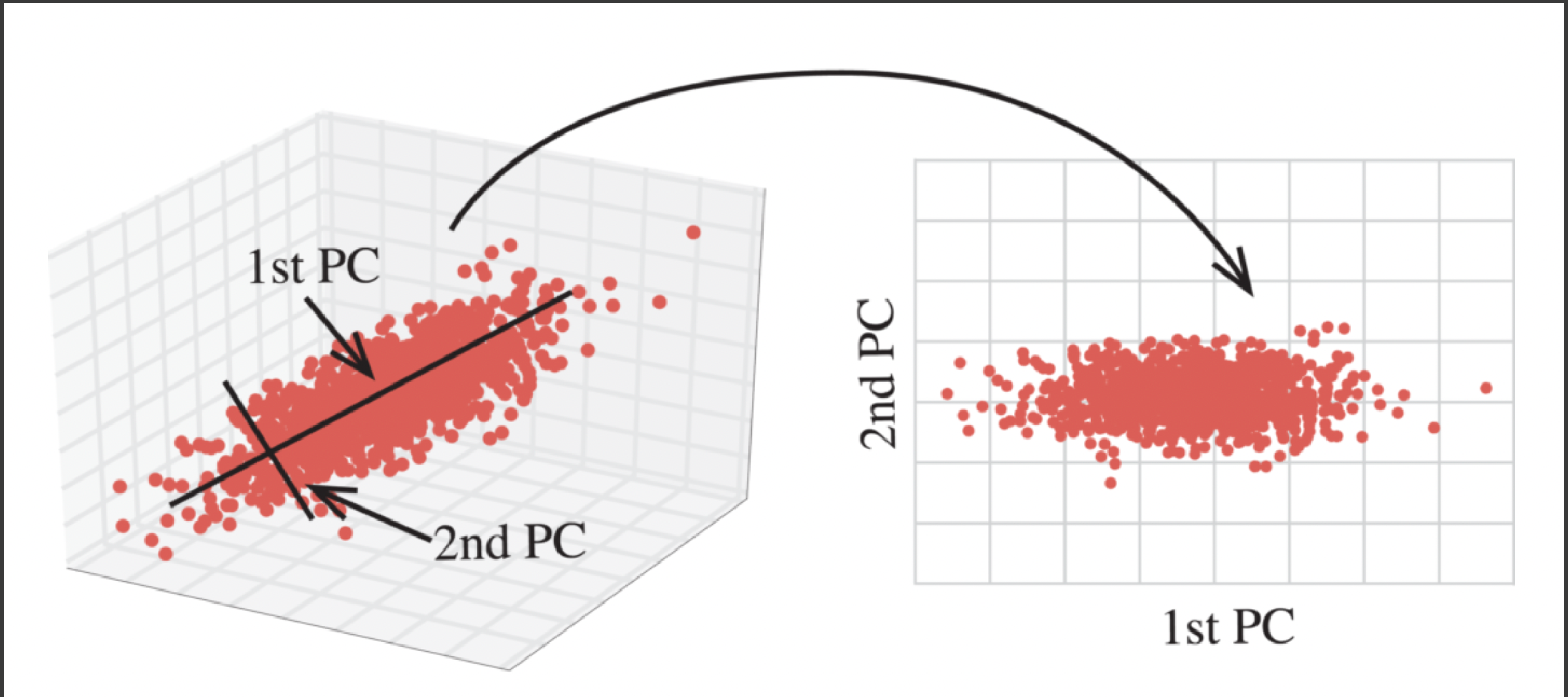
General Statistical pipeline

Copula Imputation for missing data

Uncertainty aware PCA

Search for GC twins via Propensity Score matching

Quick reminder



2.25 Other Topics in Soft-Modeling: Maximum Likelihood-Based Soft-Modeling Methods

P. D. Wentzell, Dalhousie University, Halifax, NS, Canada

Uncertainty Aware Principal Components Analysis

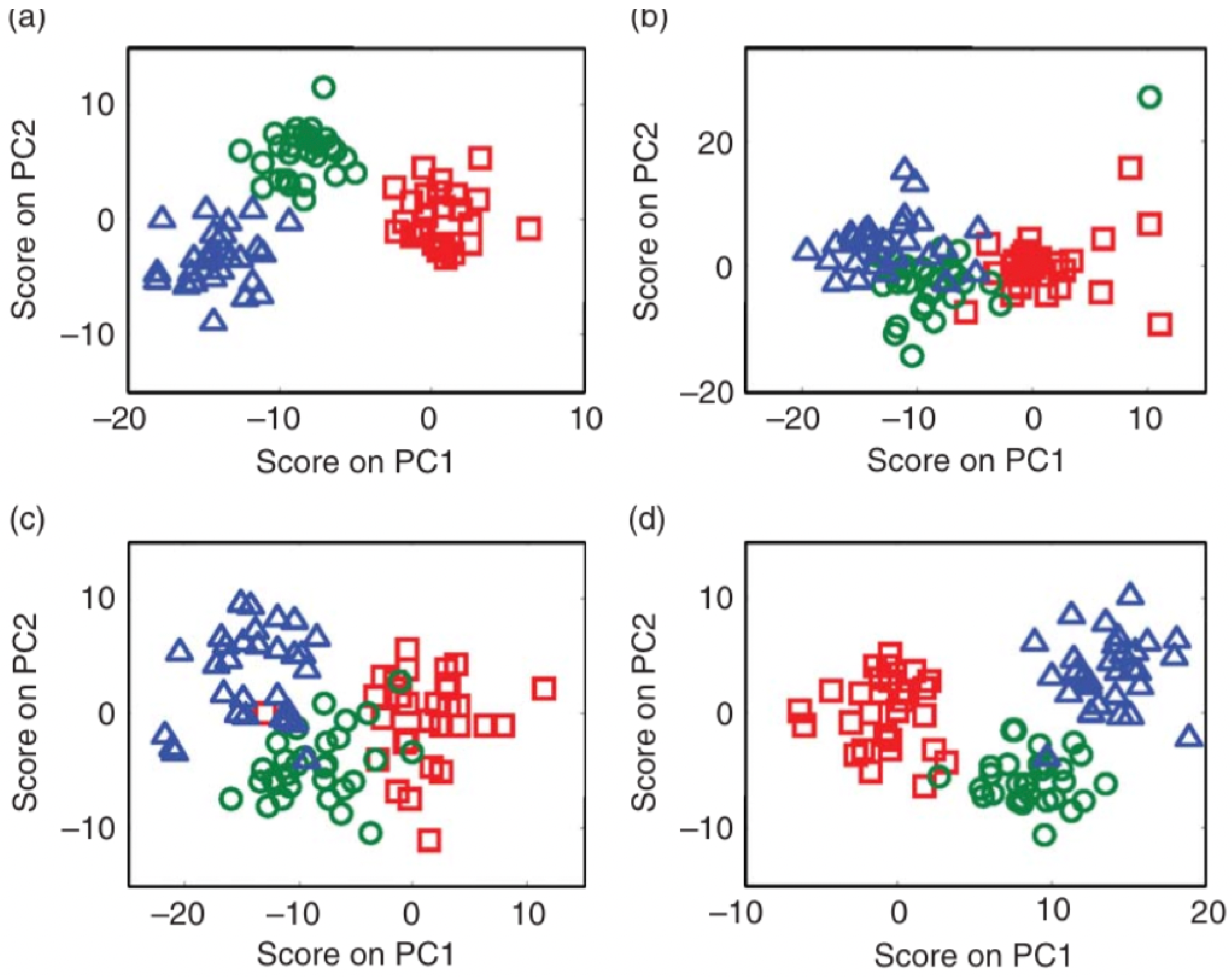
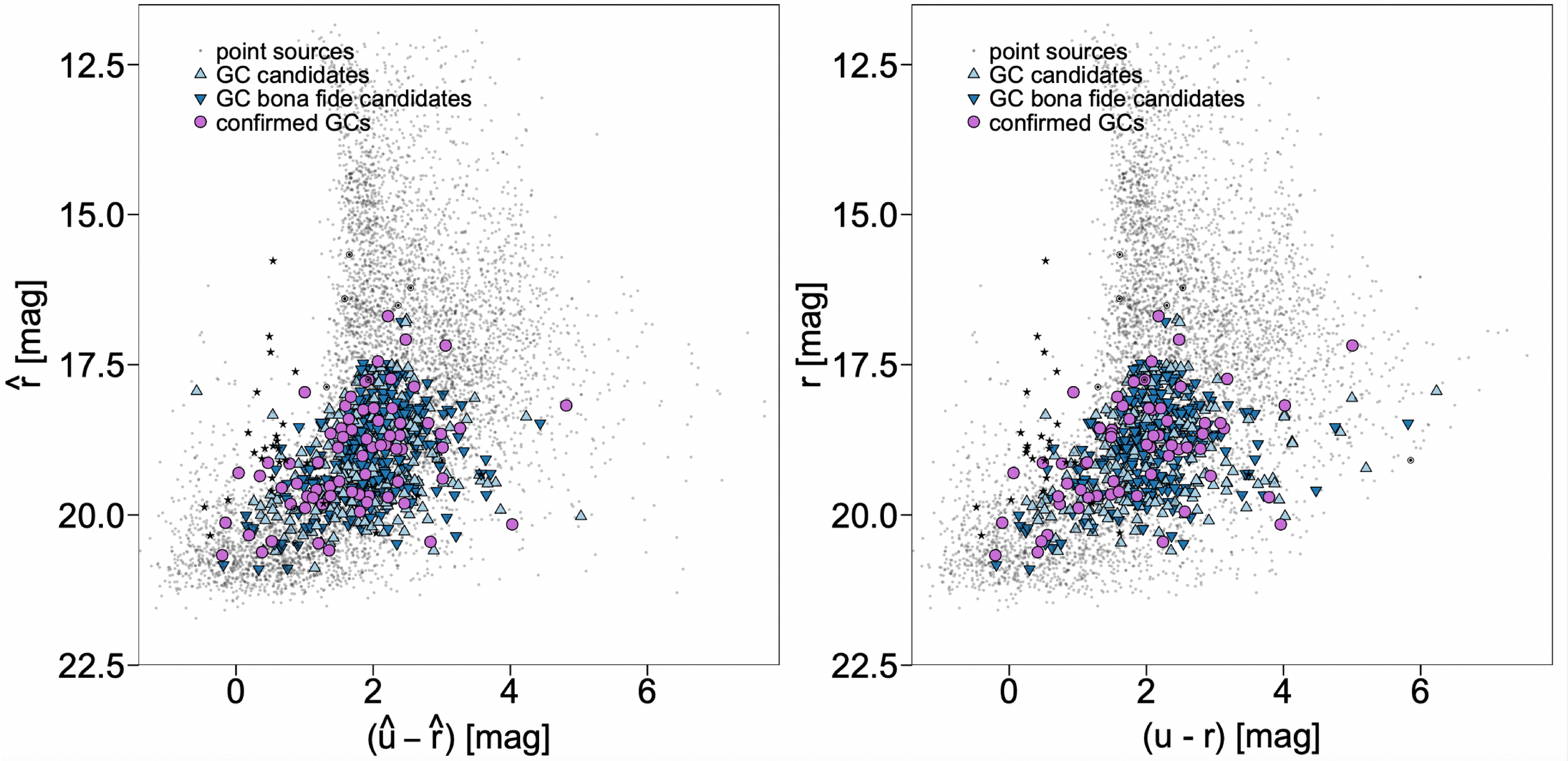


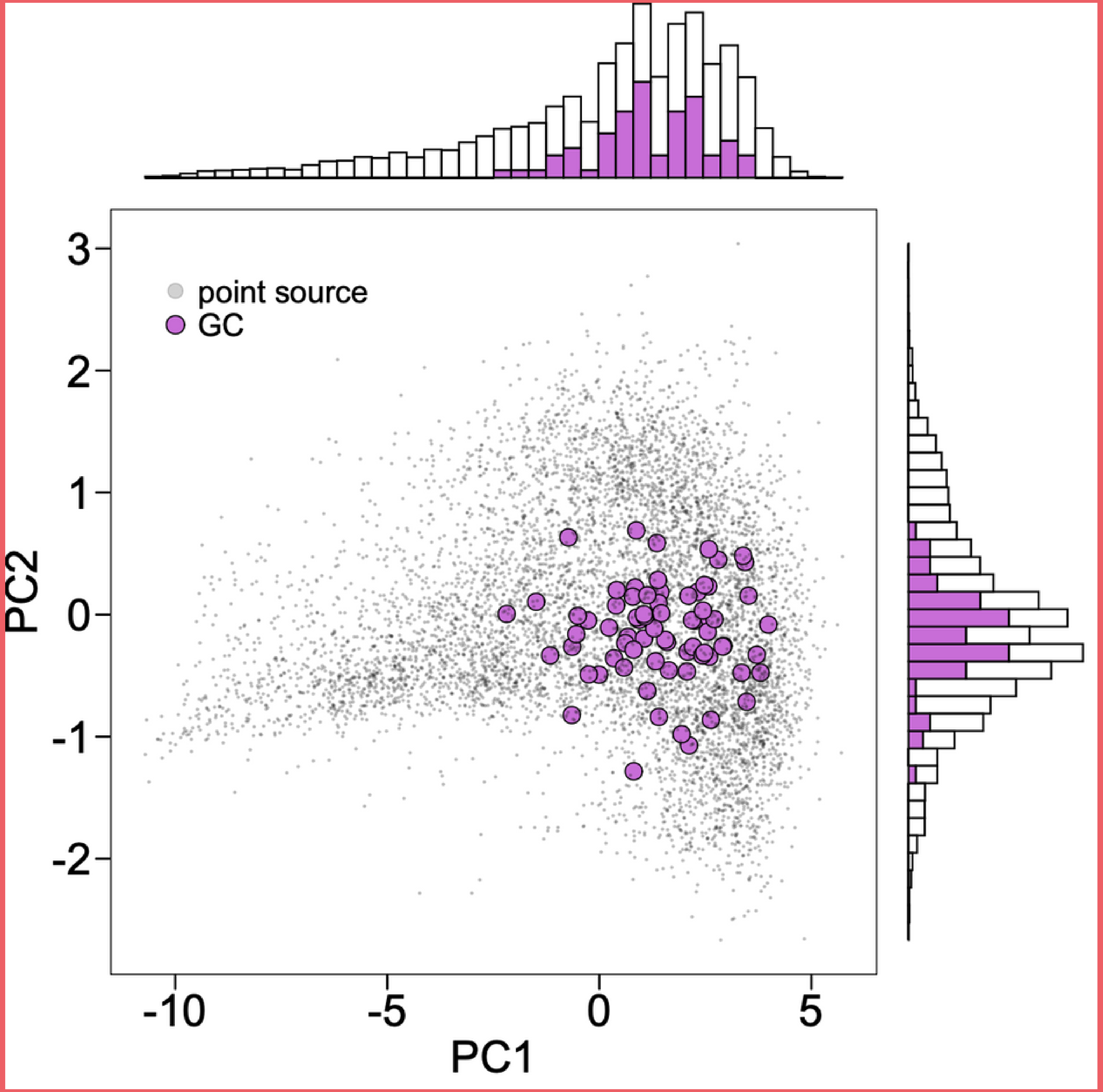
Figure 16 Application of PCA and MLPCA to simulated clustered data with heteroscedastic noise: (a) PCA results for noise-free measurements, (b) PCA results for noisy measurements, (c) MLPCA results for noisy measurements with true measurement variances, and (d) MLPCA results with 'buffered' measurement variances.

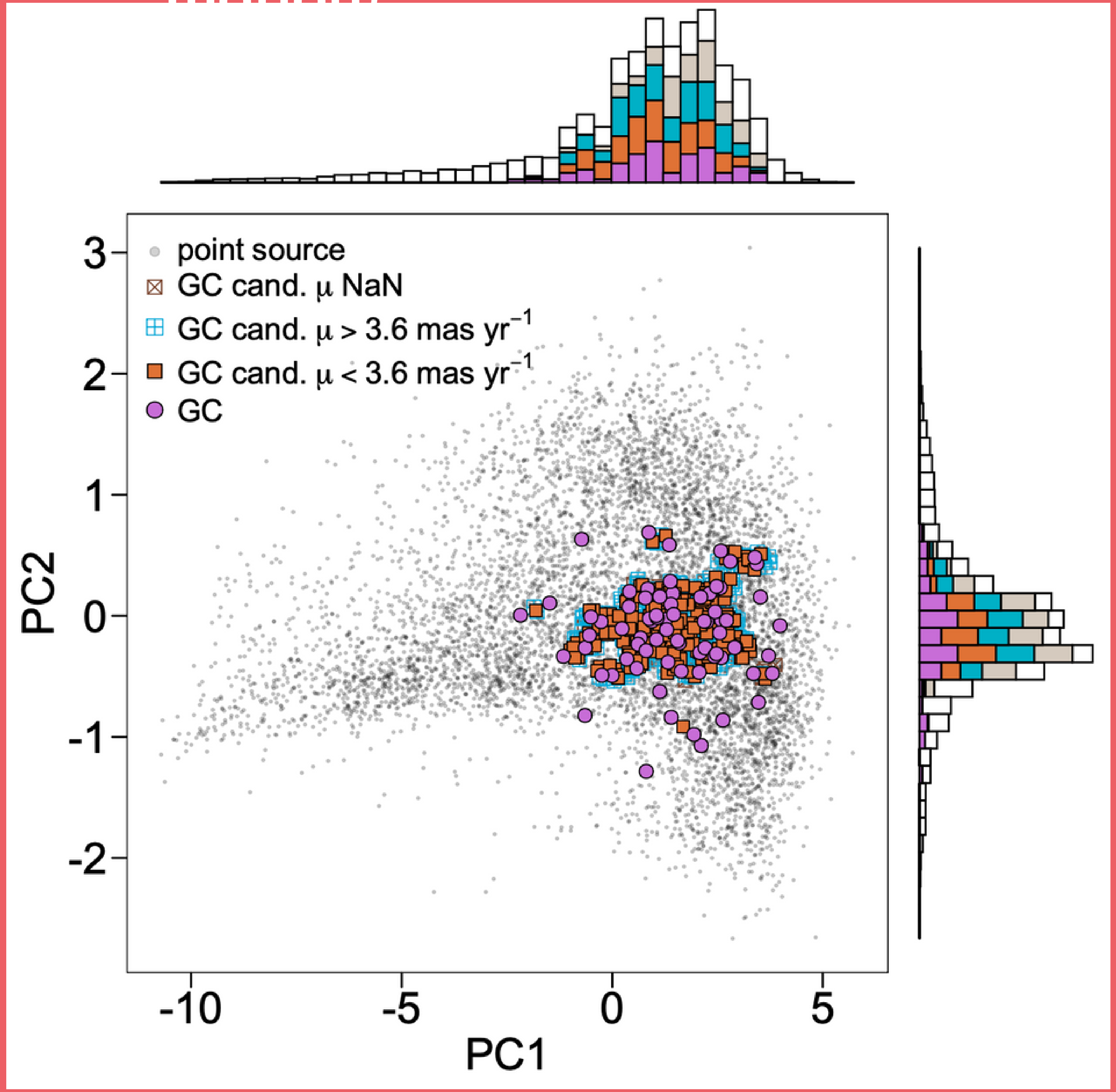


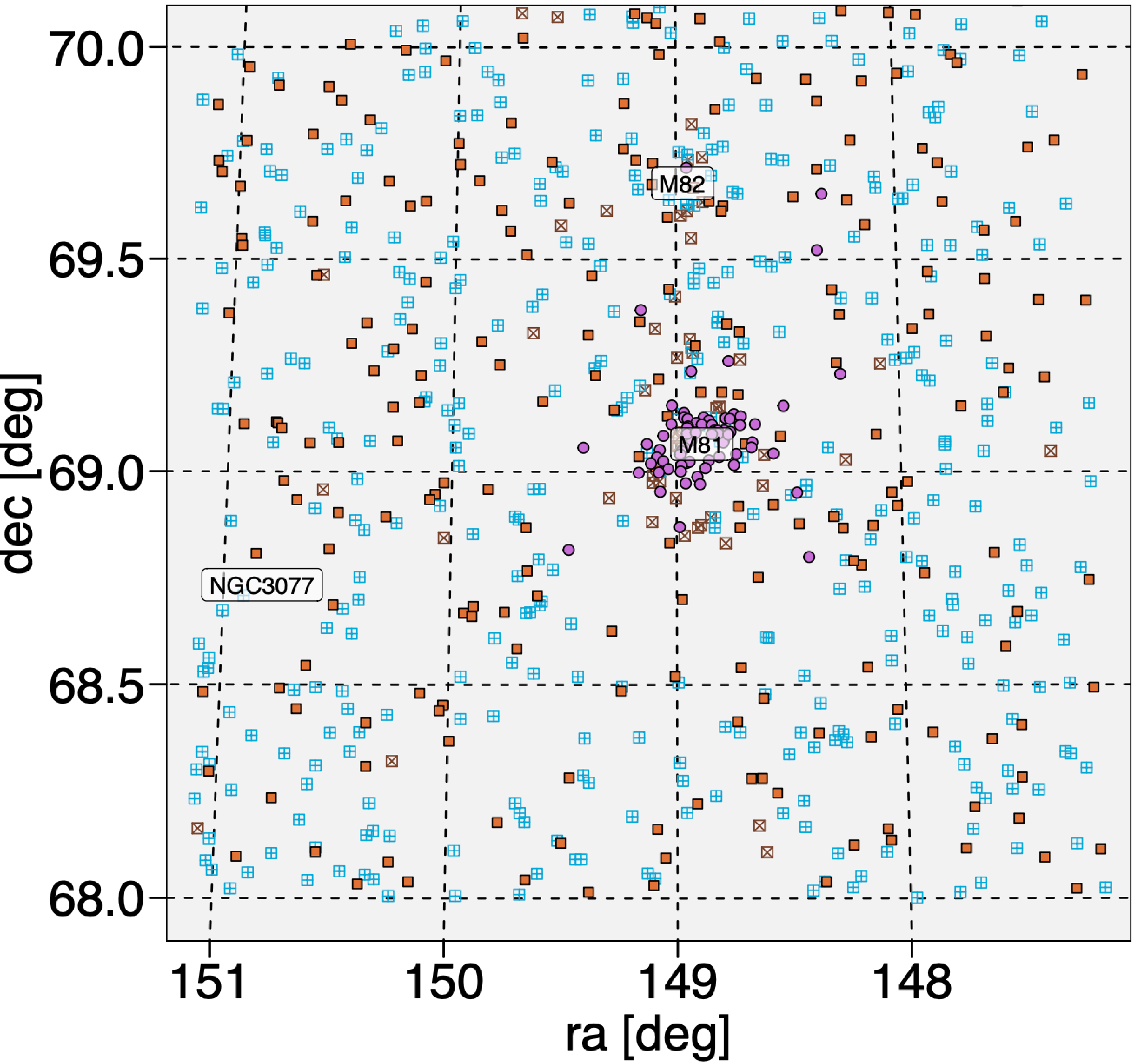
point sources
GC candidates
GC bona fide candidates
confirmed GCs
$\hat{r}$ [mag]
12.5
15.0
17.5
20.0
22.5
0
2
4
6
$(\hat{u} - \hat{r})$ [mag]
point sources
GC candidates
GC bona fide candidates
confirmed GCs
r [mag]
12.5
15.0
17.5
20.0
22.5
0
2
4
6
(u - r) [mag]

640 GC candidates

310 bona fide (i.e. low proper motion)







Spatial distribution of
640 candidates
310 bonafide

The largest list of GCs around the triplet to date. The next step is to expand the search, and get spectra.

Sit down before fact as a little child, be prepared to give up every preconceived notion, follow humbly wherever and to whatever abysses nature leads...

Thomas Huxley

References

- Copulas:
- Uncertainty Aware PCA:
- Hierarchical Bayesian Models: