ASTROSTATISTICS AND THE PATHWAY TO INTERDISCIPLINARITY

Rafael S.de Souza
Shanghai Astronomical Observatory
Chair: Cosmostatistics Initiative
Vice-President: International Astrostatistics Association

OUTLINE

- Generalized Linear Models
- Statistical Learning
- Discovering stellar clusters

COIN

Cosmostatistics Initiative

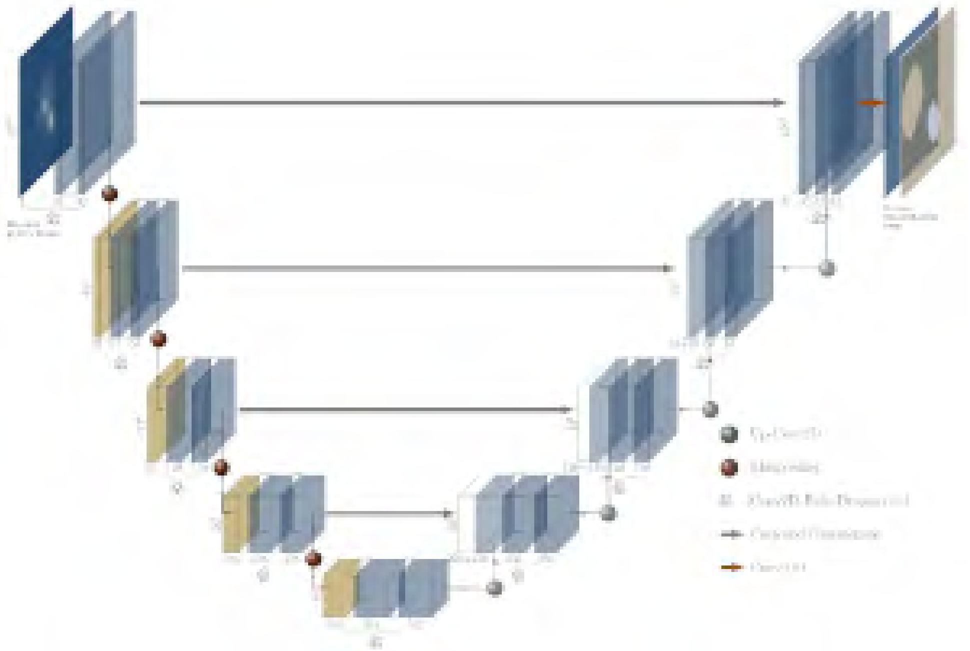
COIN

Cosmostatistics Initiative

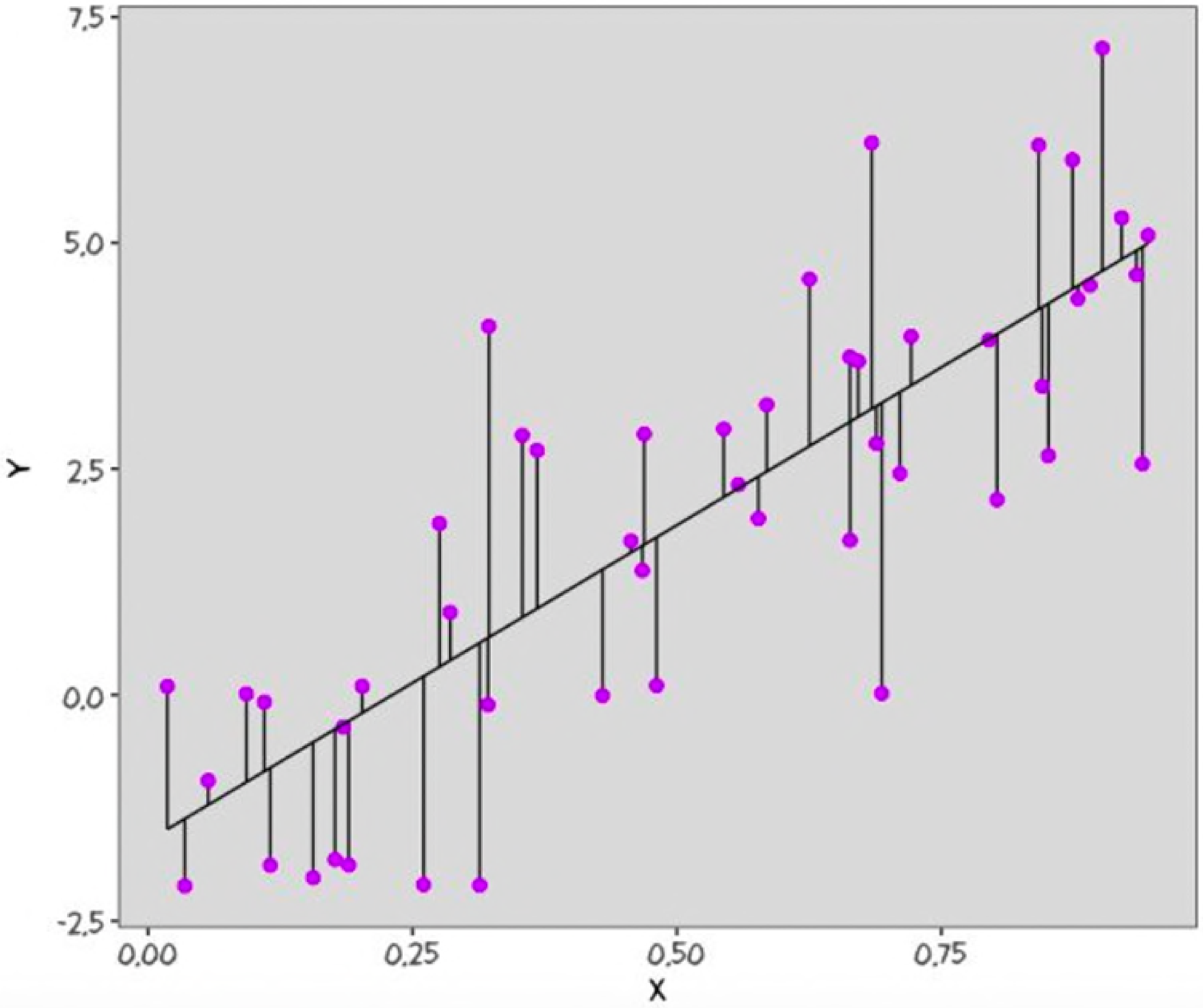
Interdisciplinary science development

70 researchers over 25 Countries

LSST DESC & COIN
RESSPECT
Recommendation System for Spectroscopic Follow-up



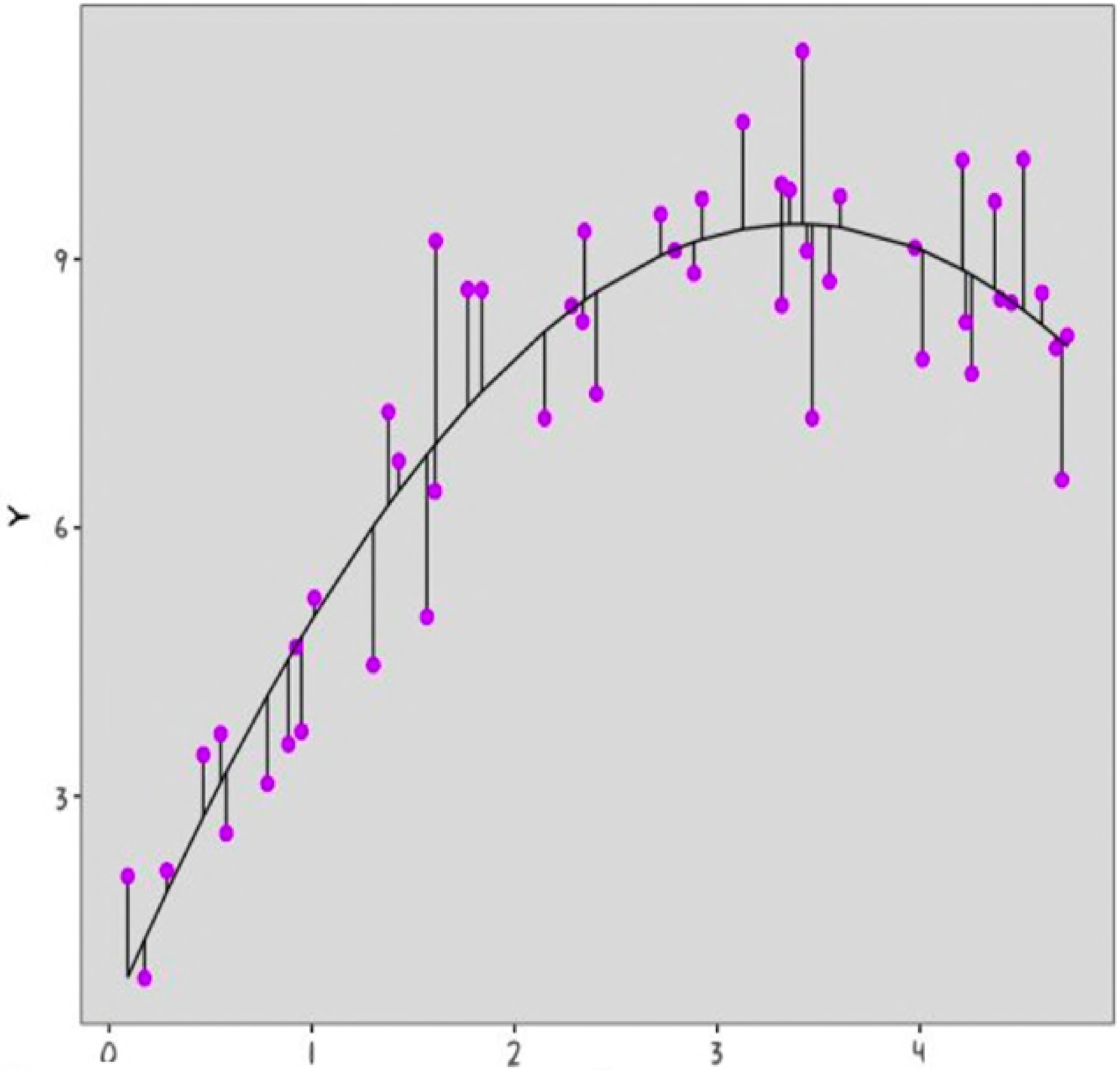
Normal Linear Models



$$y = ax + b + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

Key assumptions:
y is real and unbounded;
Homoscedastic variance

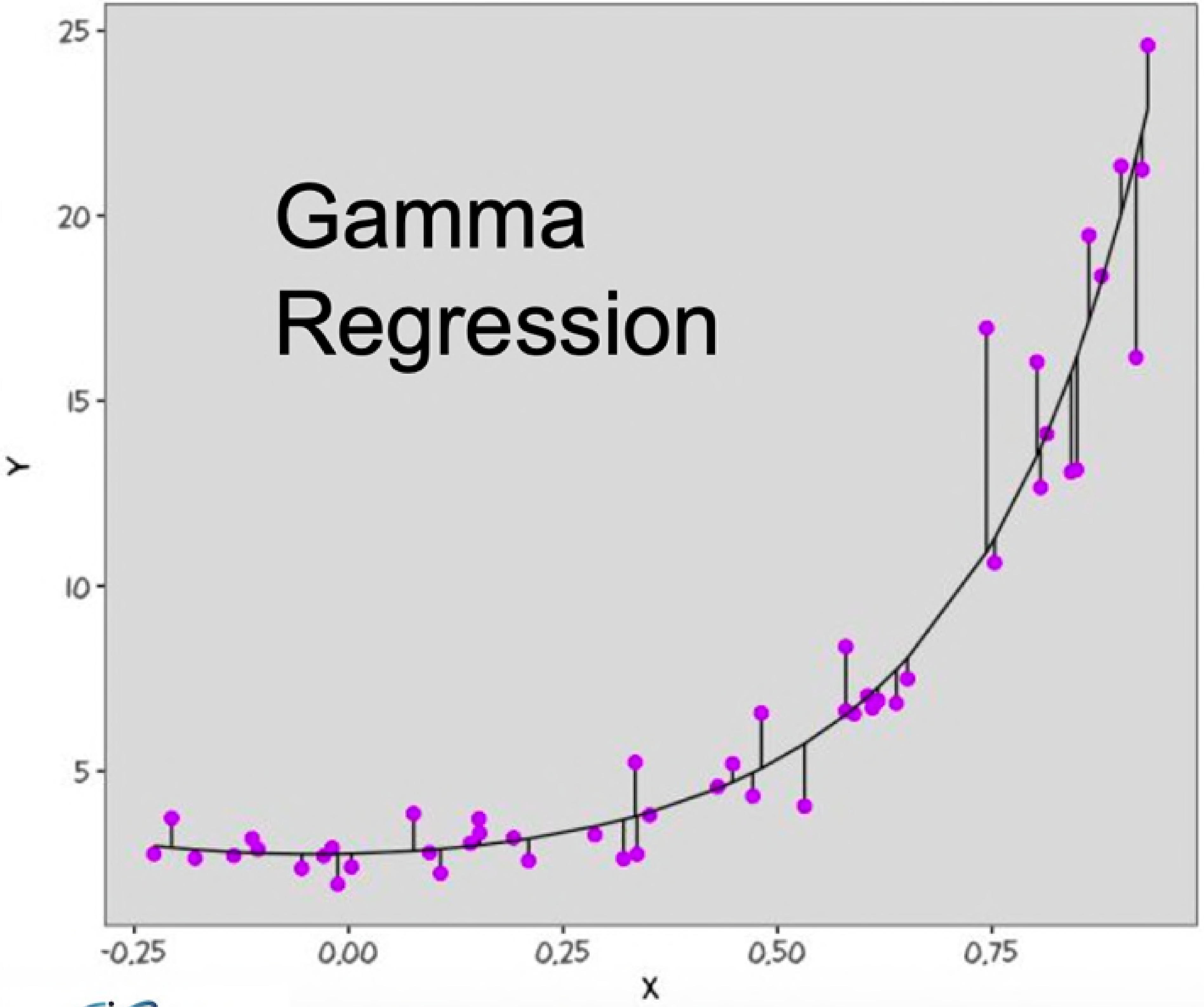
Normal (Gaussian) Linear Models



$$Y_i \sim \text{Normal}(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

Gaussian Models

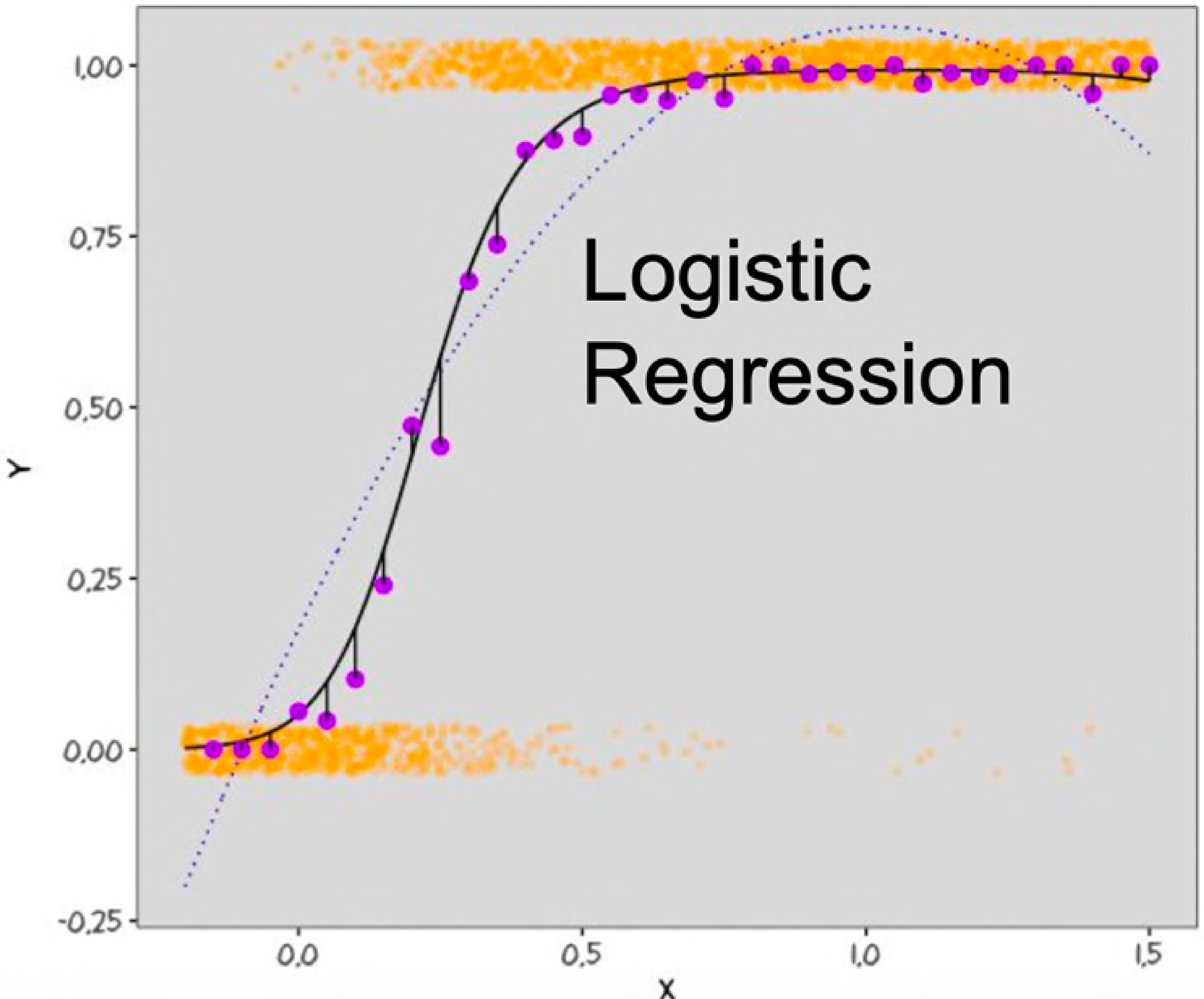
Limitations



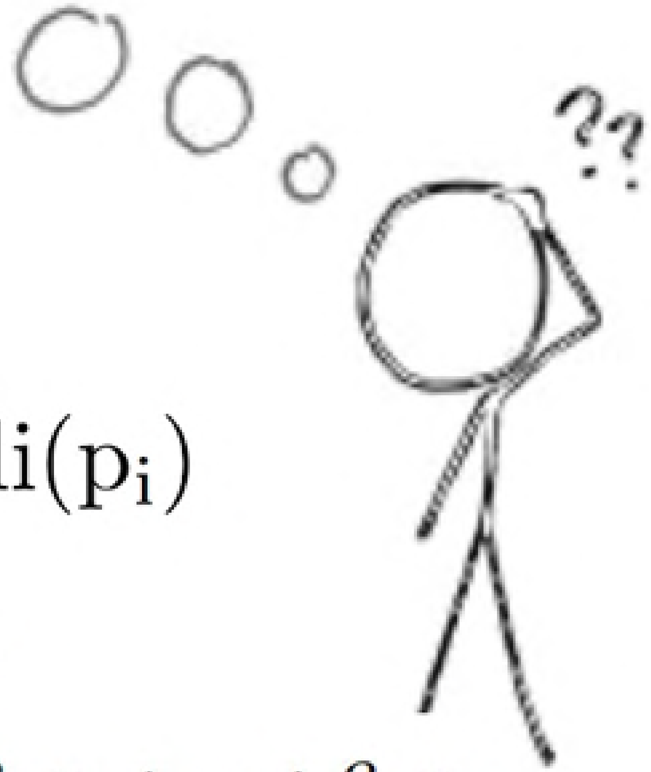
Non-fixed variance,
aka Heteroscedasticity

Gaussian Models

Limitations



Binary data



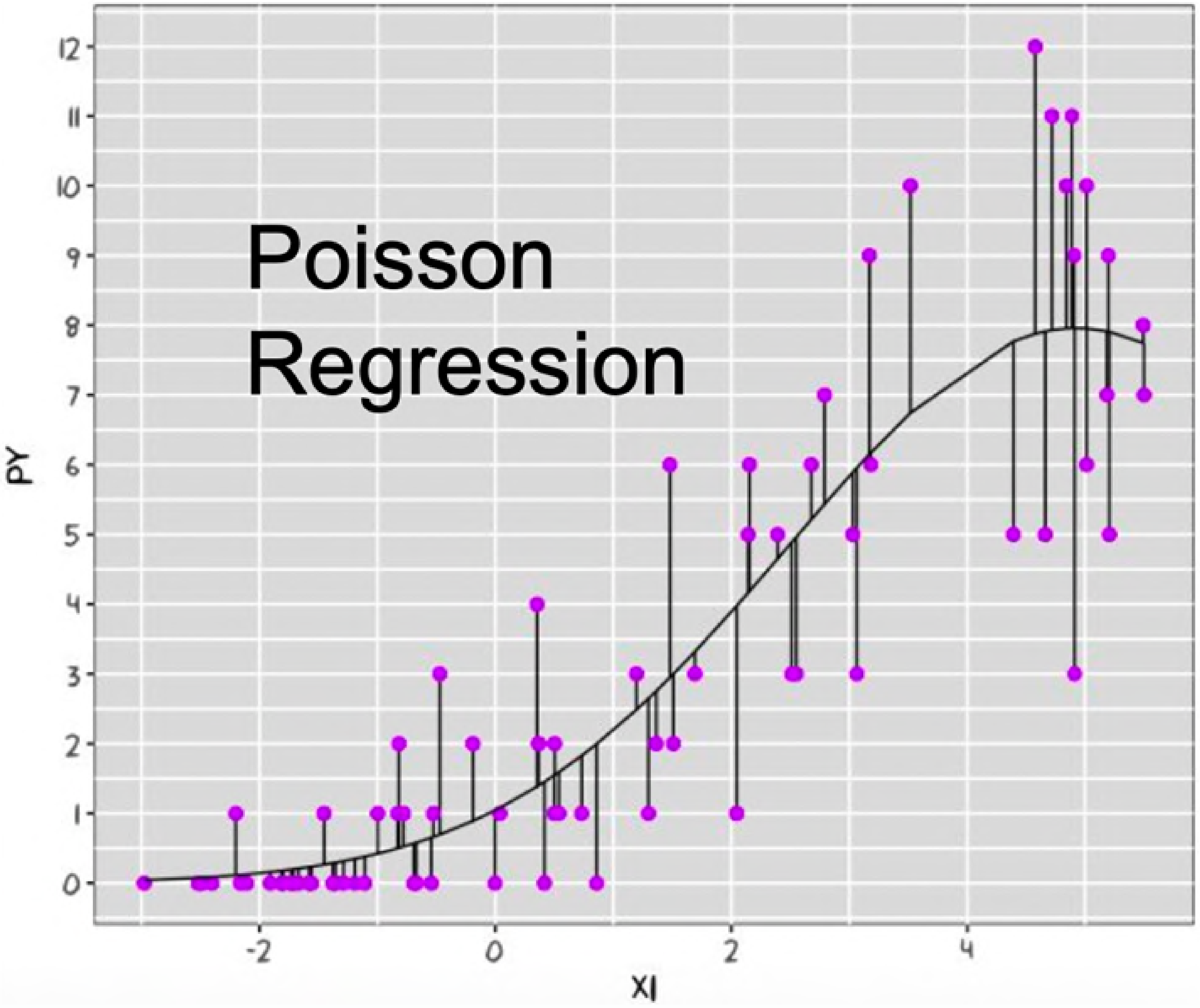
$$\mathrm{y_i} \sim \mathrm{Bernoulli}(\mathrm{p_i})$$

$$\mathrm{logit}(\mathrm{p_i}) = \eta_\mathrm{i}$$

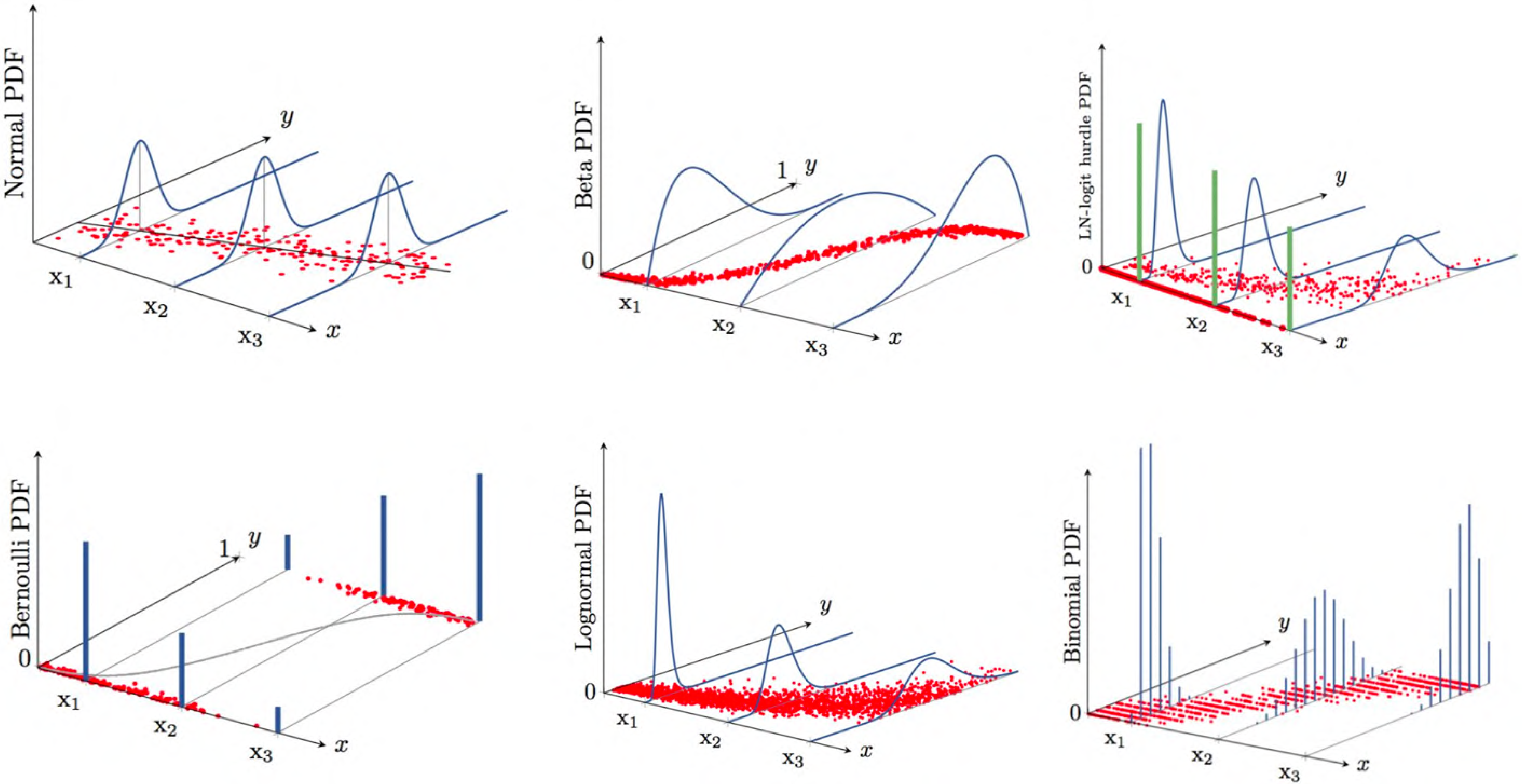
$$\eta_i \equiv \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

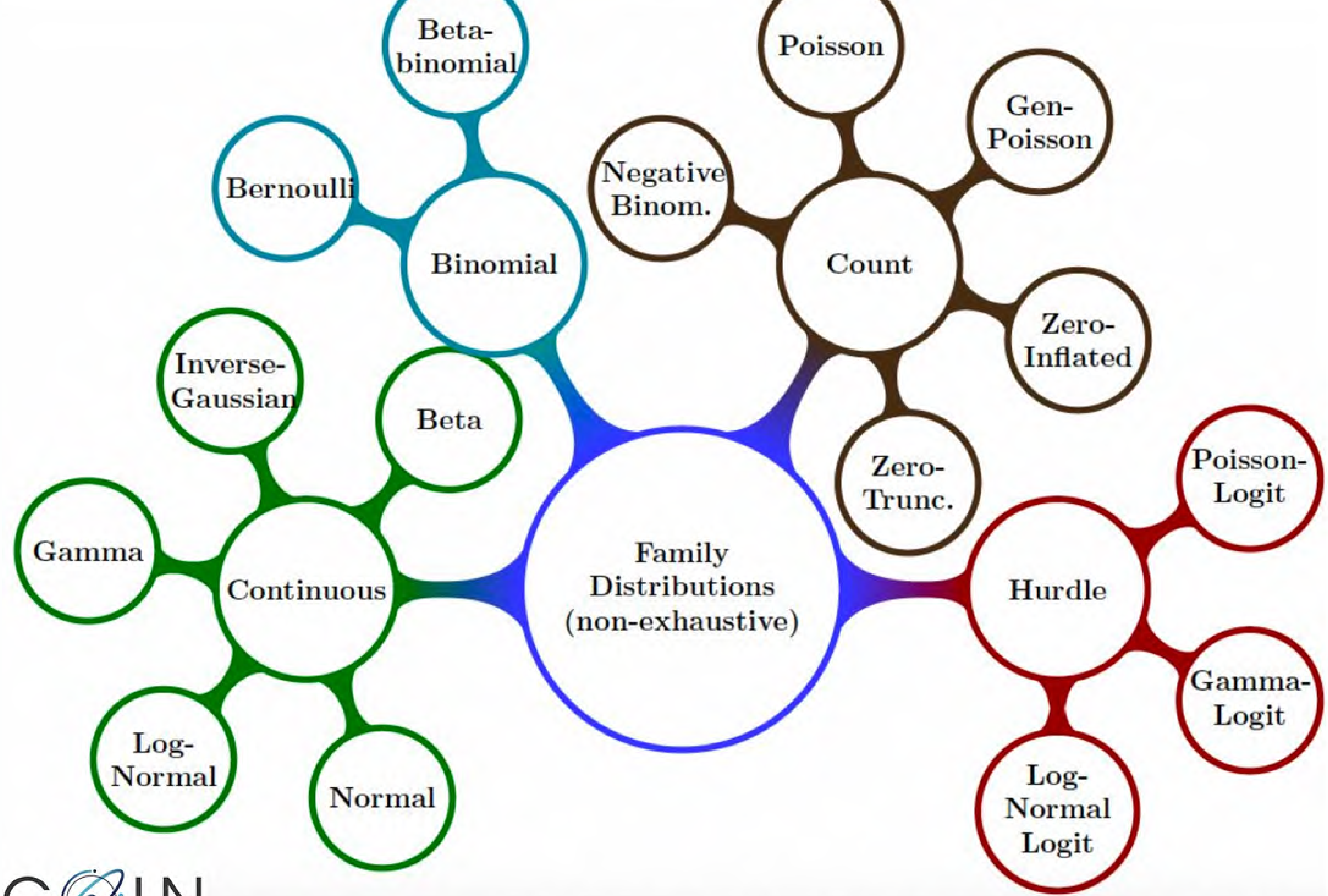
Gaussian Models

Limitations



Discrete data


Normal PDF
y
x_1
x_2
x_3
x
Beta PDF
1
y
0
x_1
x_2
x_3
x
LN-logit hurdle PDF
y
0
x_1
x_2
x_3
x
Bernoulli PDF
1
y
0
x_1
x_2
x_3
x
Lognormal PDF
y
0
x_1
x_2
x_3
x
Binomial PDF
y
0
x_1
x_2
x_3
x



COIN
Cosmostatistics Initiative

Figure from Bayesian Models for Astrophysical Data, CUP, 2017

$$
\boxed{\begin{aligned}
&Y_i \sim f(\mu_i, a(\phi)V(\mu_i)),\\
&g(\mu_i) = \eta_i,\\
&\eta_i \equiv \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.
\end{aligned}}
$$

Special Cases

Linear regression

$$Y_i \sim Normal(\mu_i, \sigma^2),$$

$$\mu_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Logistic regression

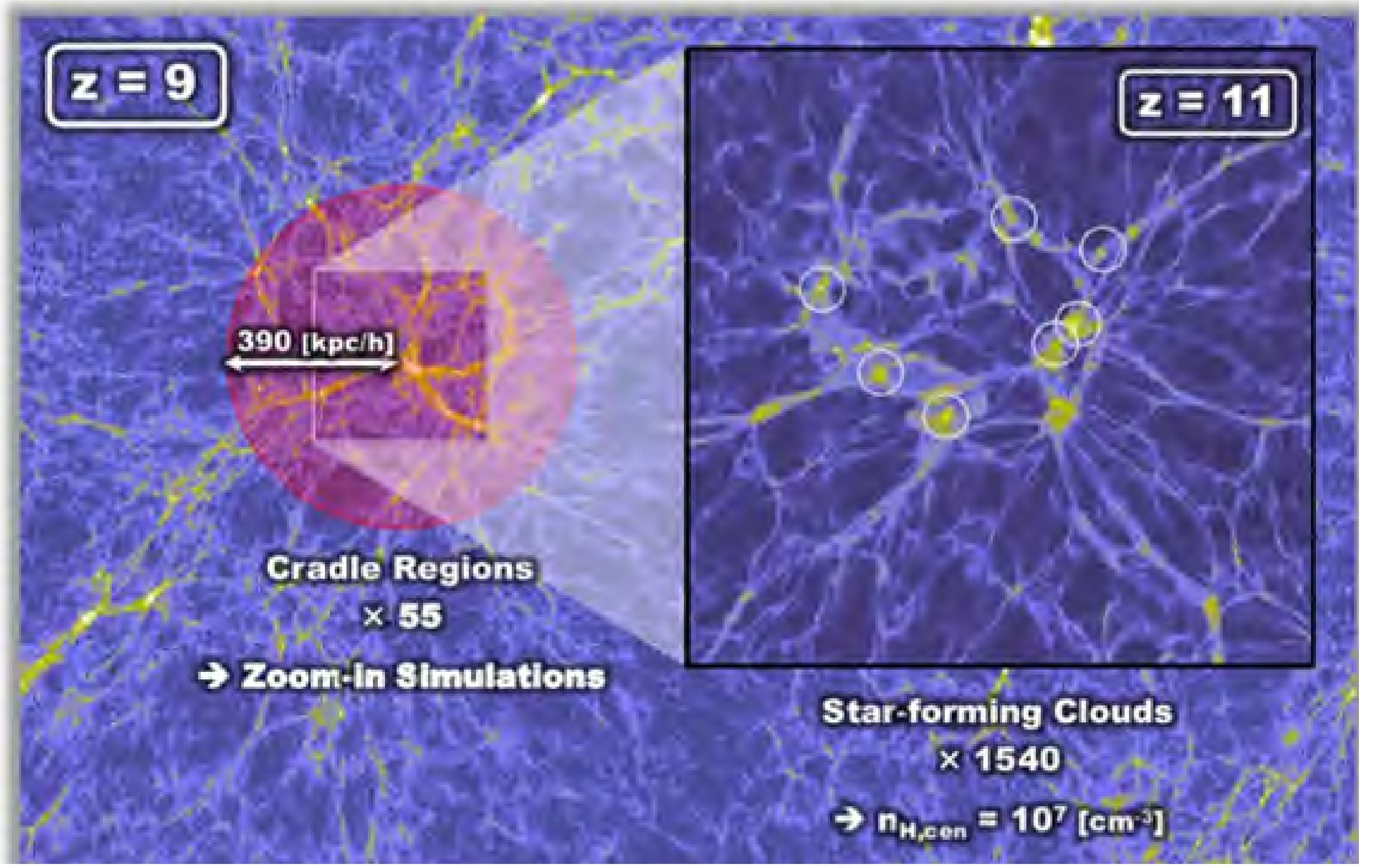
$$\mathrm{y_i} \sim \mathrm{Bernoulli}(\mathrm{p_i})$$

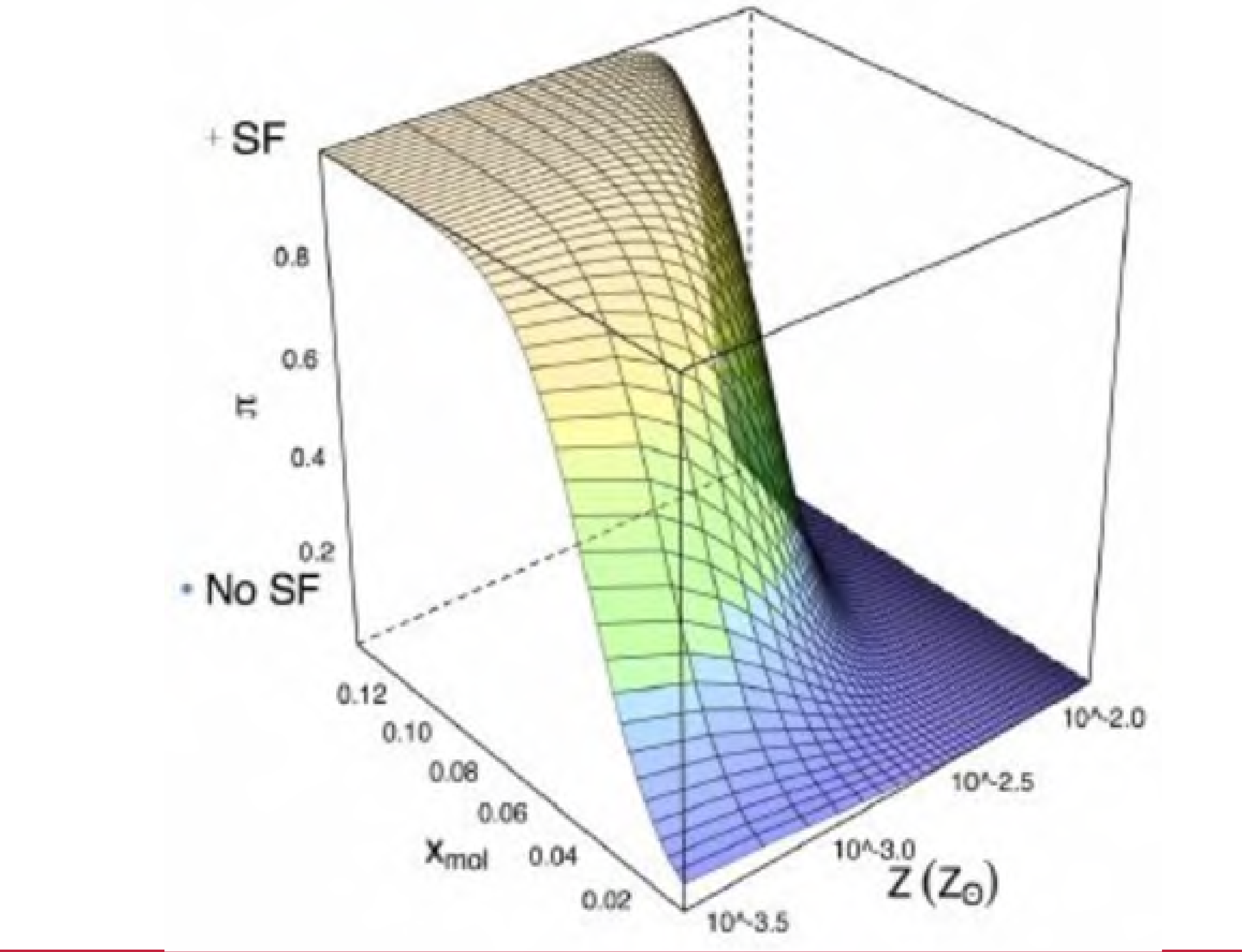
$$\mathrm{logit}(\mathrm{p_i}) = \eta_\mathrm{i}$$

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

Generalized Linear Models: Logistic regression

Modelling Star Formation activity as a function of molecular fraction and metallicity in primordial haloes simulations



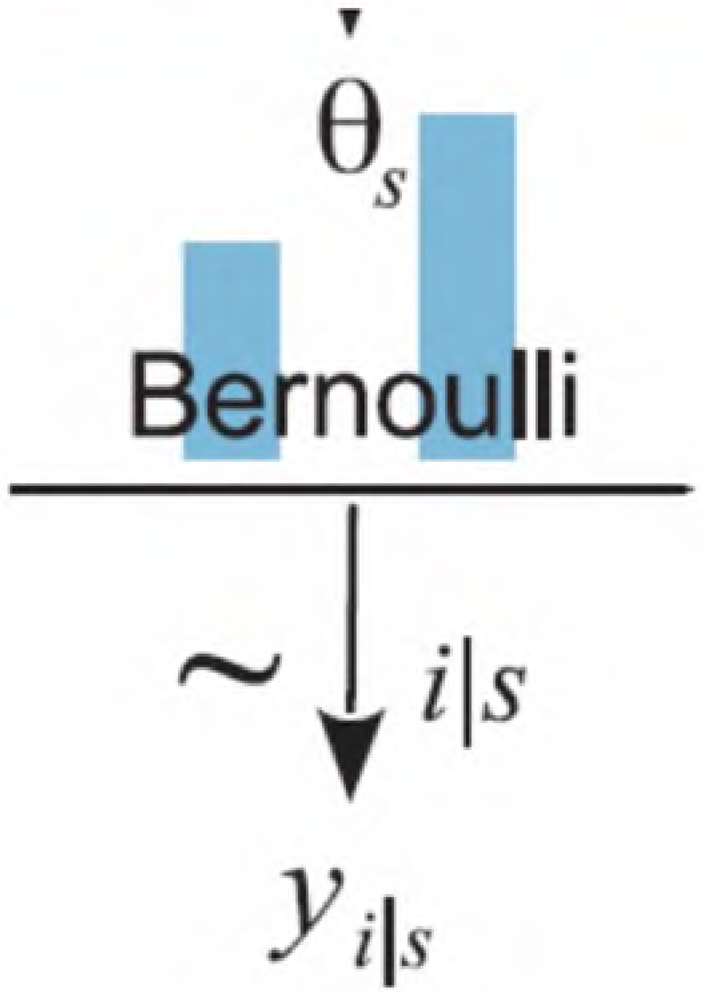


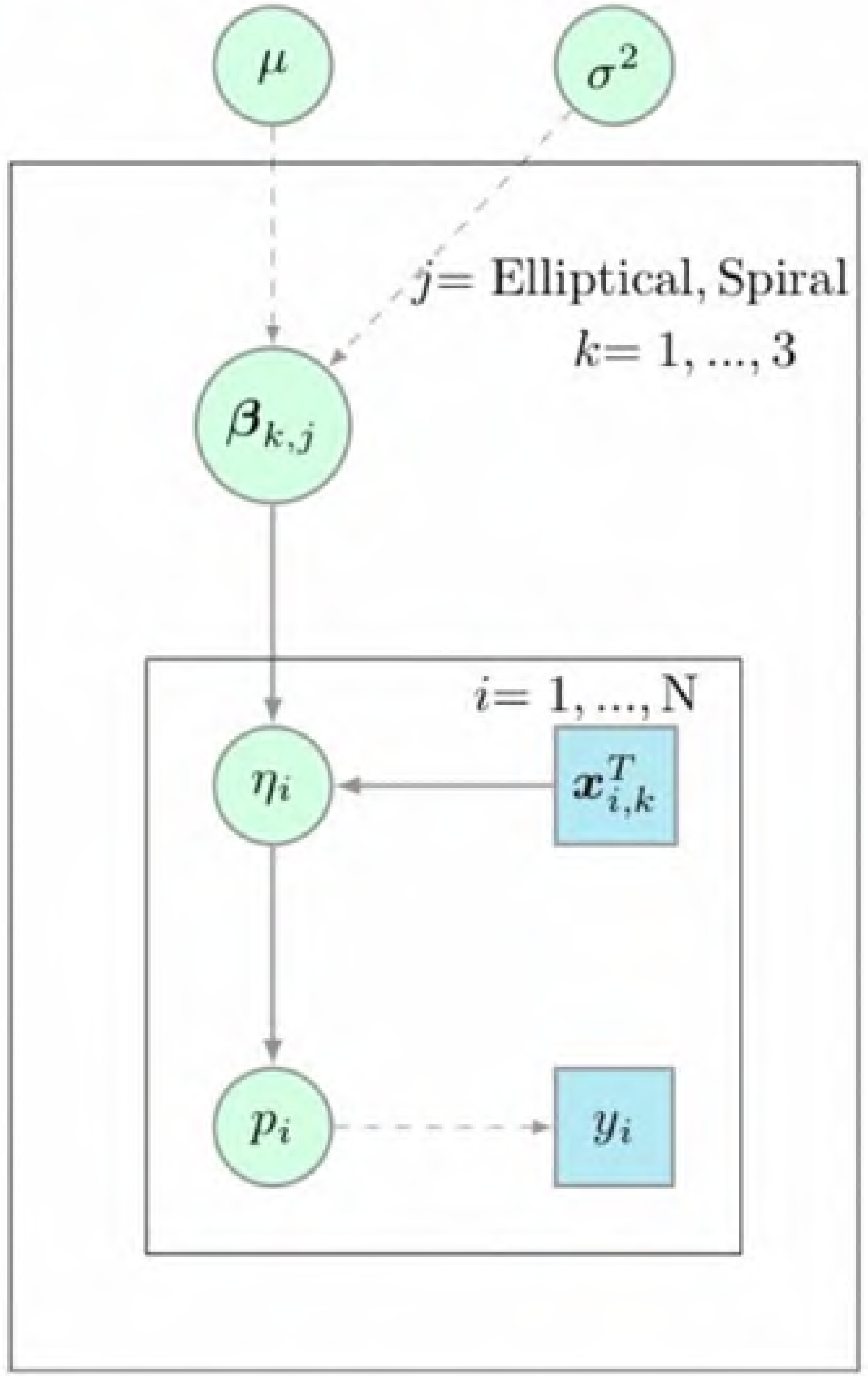
Is the cluster environment quenching the Seyfert activity in elliptical and spiral galaxies?

R. S. de Souza ✉, M. L. L. Dantas ✉, A. Krone-Martins, E. Cameron, P. Coelho, M. W. Hattab, M. de Val-Borro, J. M. Hilbe, J. Elliott, A. Hagen ... Show more

Monthly Notices of the Royal Astronomical Society, Volume 461, Issue 2, 11 September 2016, Pages 2115–2125, https://doi.org/10.1093/mnras/stw1459

Bayesian Hierarchical
Logistic Regression

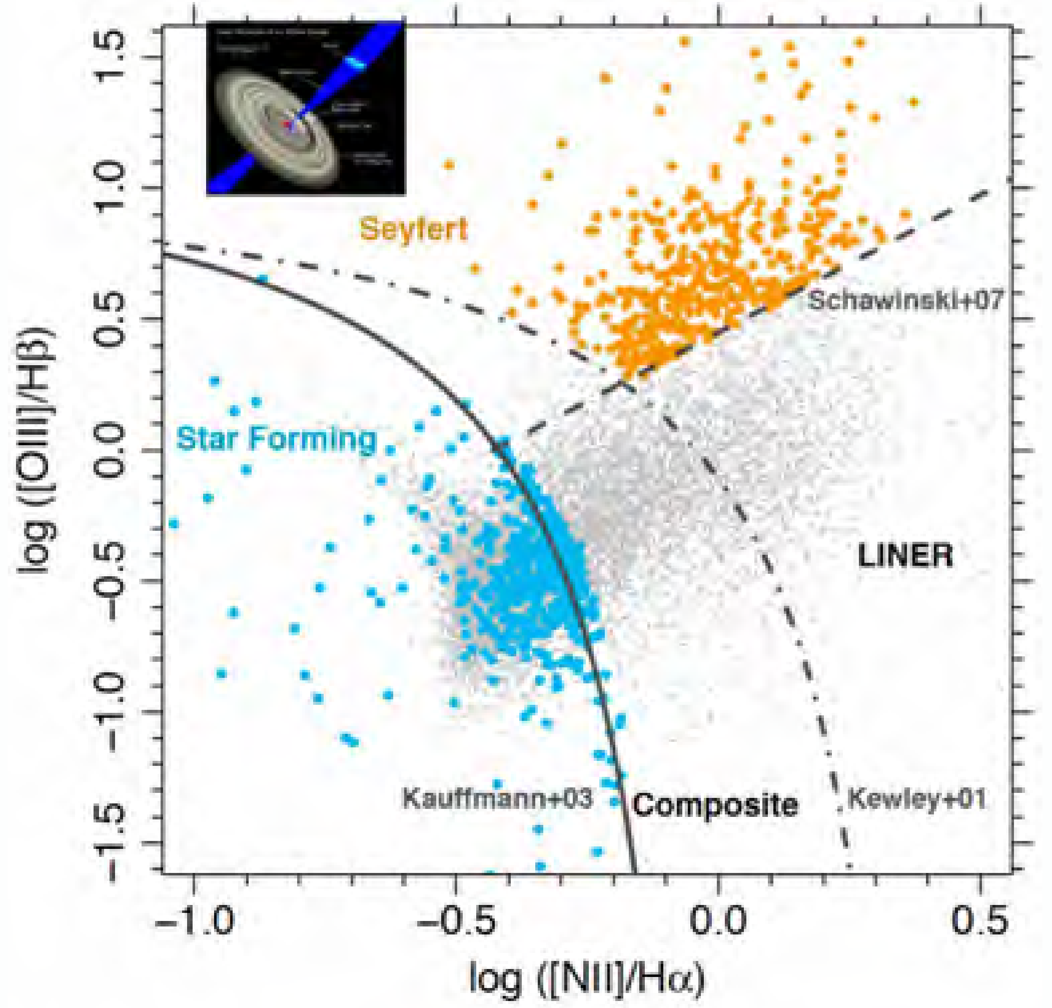


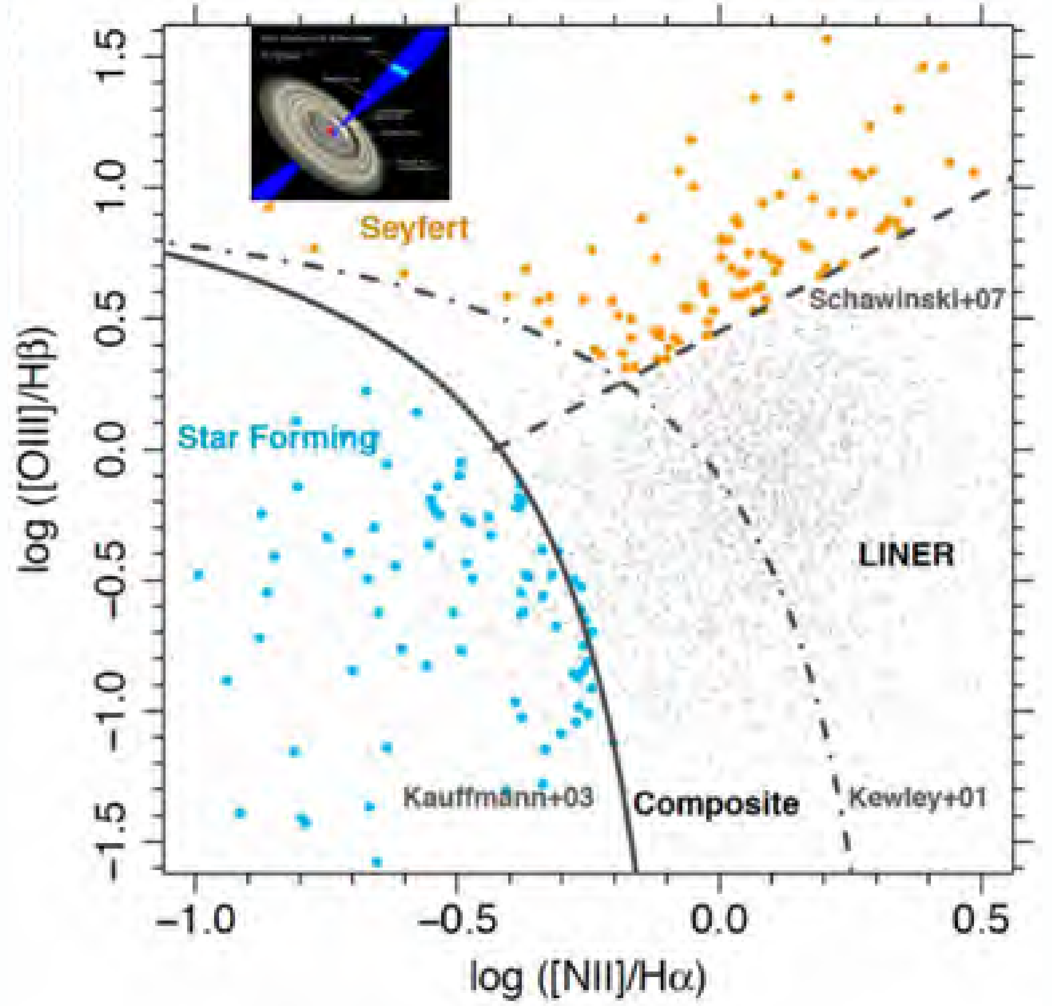


$$y_i \sim \mathrm{Bernoulli}(p_i)$$
$$\mathrm{logit}(p_i) = \eta_i$$
$$\eta_i = \boldsymbol{x}^T_{i,k}\beta_{k,j}$$
$$\boldsymbol{x}^T_{i,k} = \begin{pmatrix} 1 & (\log M_{200})_1 & \left(\frac{r}{r_{200}}\right)_1 \\ \vdots & \vdots & \vdots \\ 1 & (\log M_{200})_N & \left(\frac{r}{r_{200}}\right)_N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

Generalized Linear Models

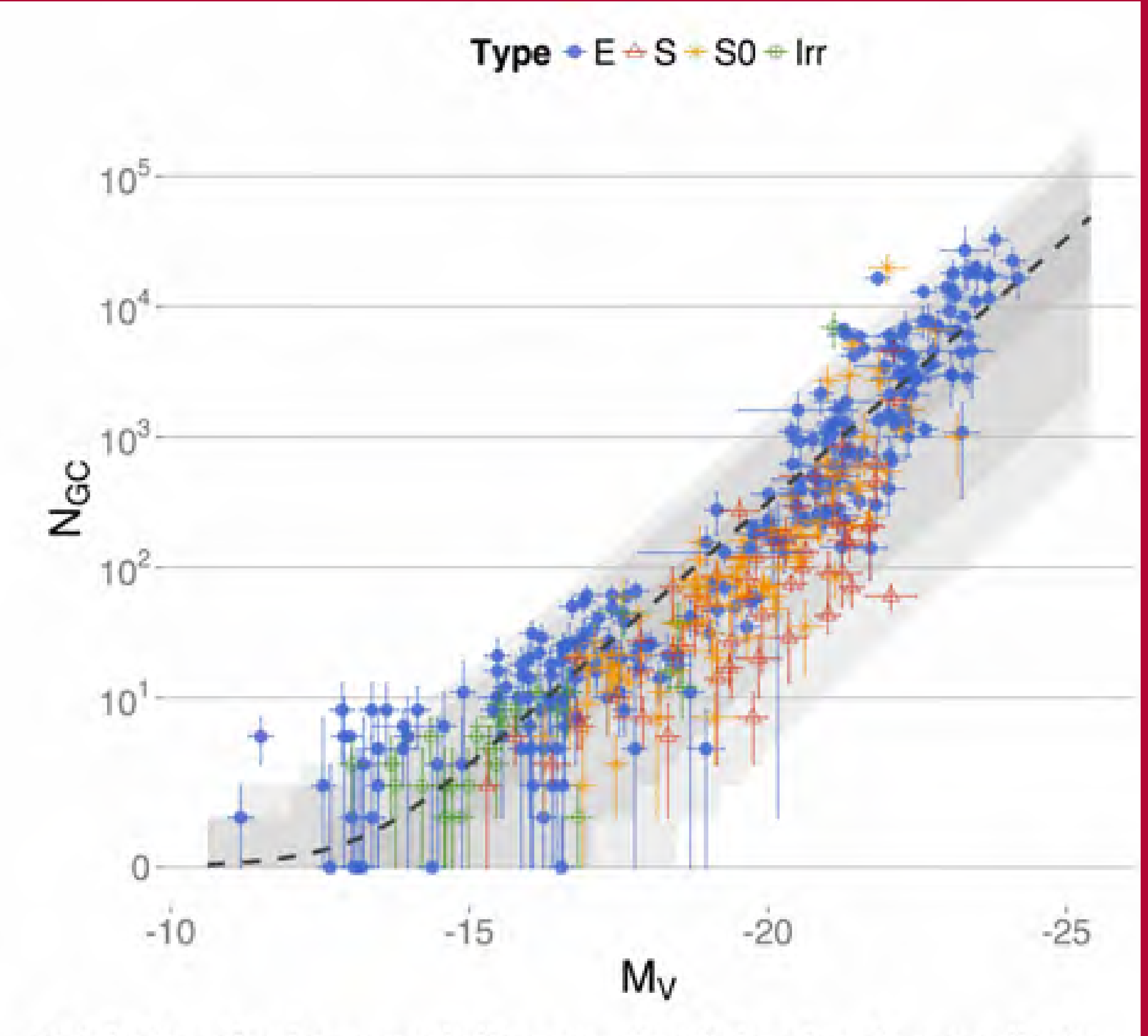
Seyfert prevalance as function of environment and galaxy morphology



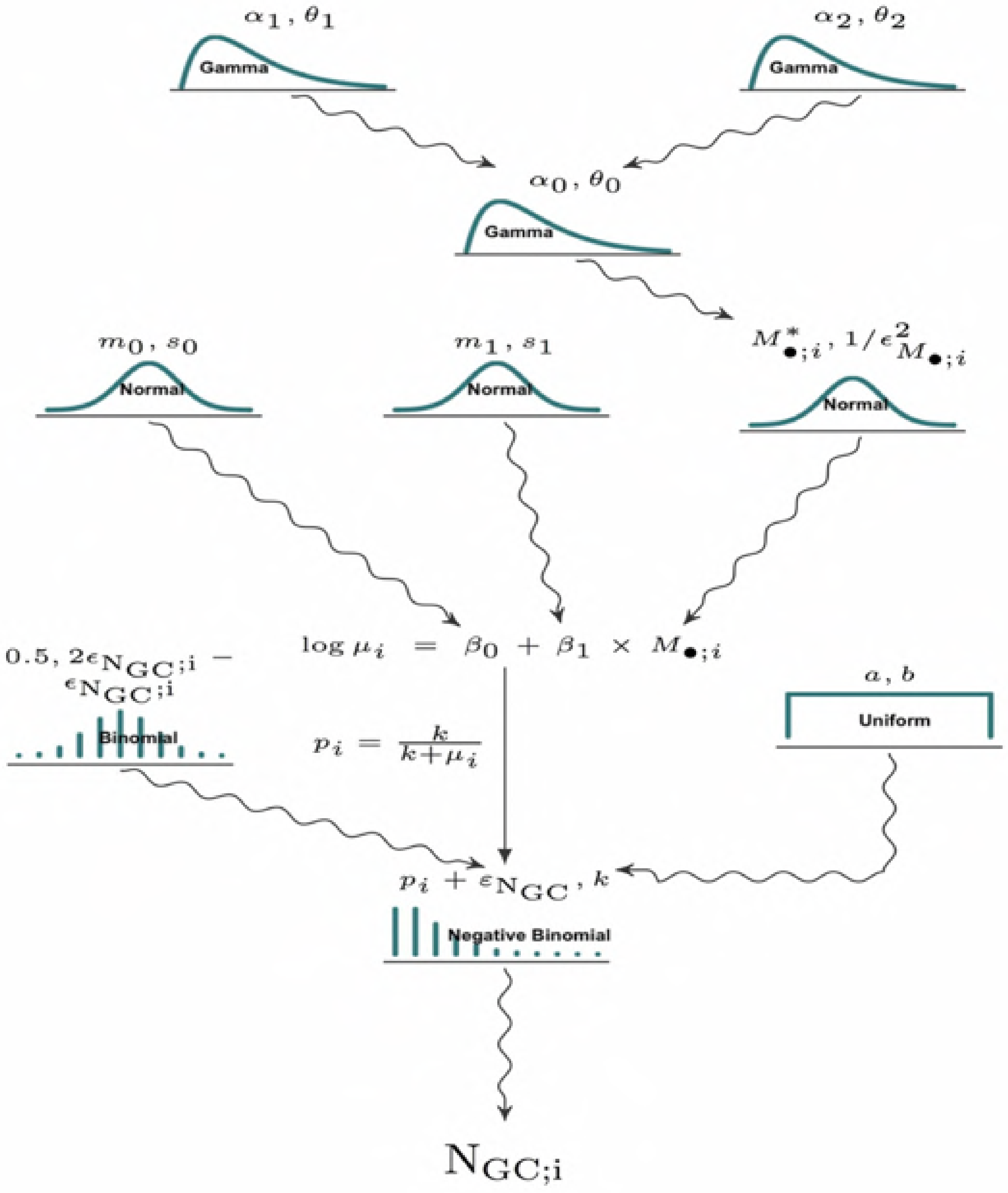


Generalized Linear Models: Negative Binomial Regression

Globular Cluster Counts as function of galaxy properties



Globular cluster population vs absolute visual magnitude



Generalized Linear Models

$$Y_i \sim f(\mu_i, a(\phi)V(\mu_i)),$$

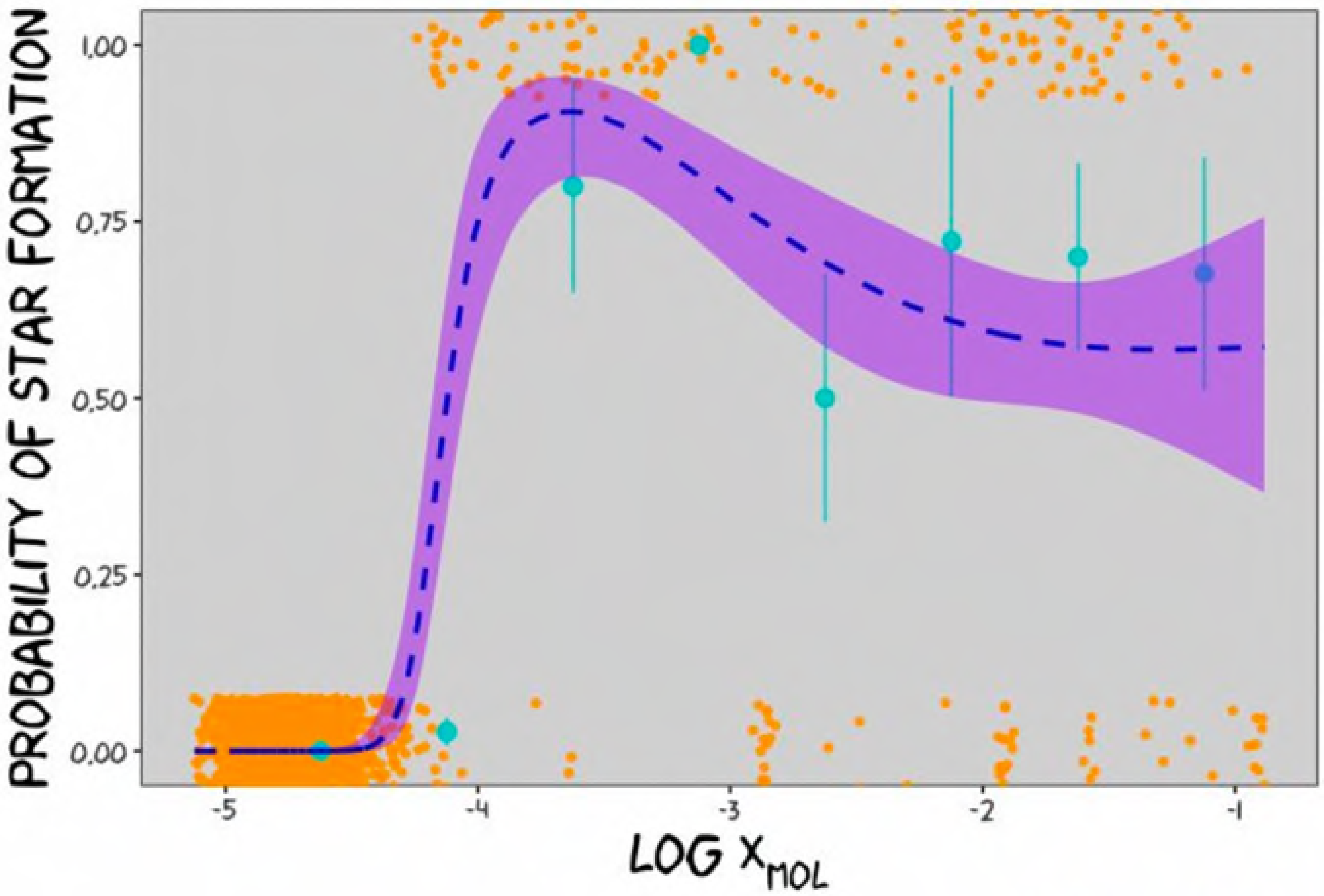
$$g(\mu_i) = \eta_i,$$

$$\boxed{\eta_i \equiv \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.}$$

Natural GLM extension

Generalized Additive Models

$$g(y) = f(x_1) + f(x_2) + \cdots + f(x_D)$$



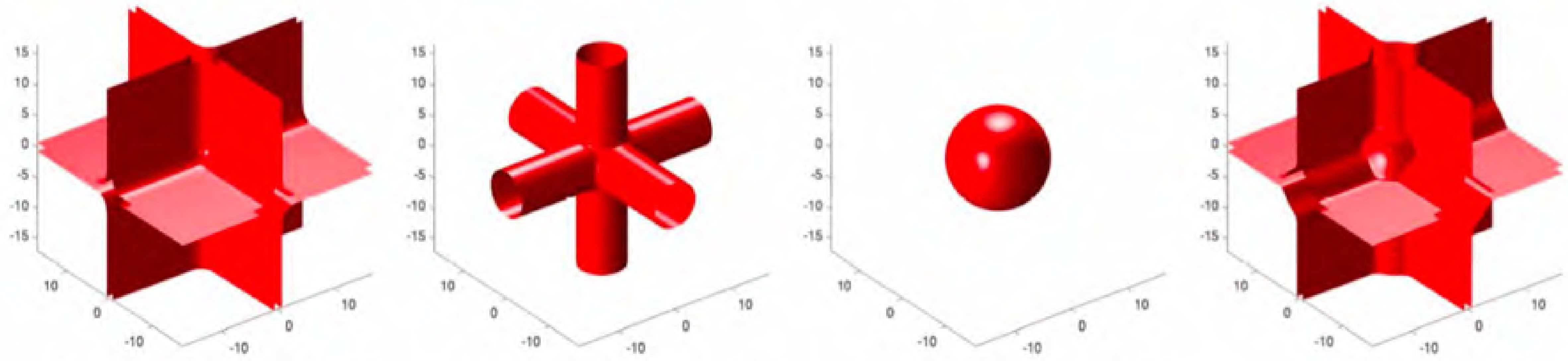
COIN
Cosmostatistics Initiative

de Souza, R. S, et al Astronomy and Computing, Volume 12, p. 21-32.

Geometric Interpretation

Changes how you perceive the data

Non-local Interactions | Local interactions | Hybrid

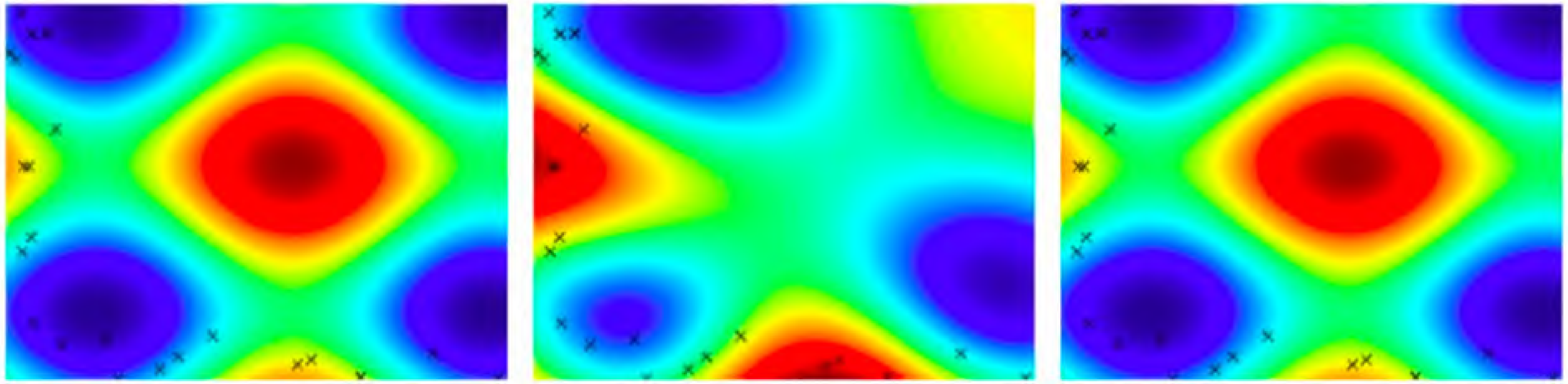


1st order interactions $k_1 + k_2 + k_3$

2nd order interactions $k_1 k_2 + k_2 k_3 + k_1 k_3$

3rd order interactions $k_1 k_2 k_3$ (Squared-exp kernel)

All interactions (Additive kernel)



True Function & data locations

Squared-exp GP posterior mean

Additive GP posterior mean

OUTLINE

- Generalized Linear Models
- Statistical Learning
- Discovering stellar clusters

Supervised ML model

data	training, target
χ	set of all samples, x
Y	set of possible labels, y
h_{train}	learner: $y_{est;i} = h_{train}(x_i)$
L	Loss function

Hypothesis: Training is representative of target

Data generation model:

$x_i \sim P_{\chi}$

$f \rightarrow$ true labeling function, $y_i = f(x_i)$

$L_{data,f}(h) \equiv P_{x\sim data}(h_{train}(x) \neq f(x))$

Shai and Shai, *Understanding ML: From Theory to Algorithms*, 2014, CUP

Supervised ML model

data training, target

Machine Learning algorithm

$\mathcal{X}$ set of all samples, x

Y set of possible labels, y

Hypothesis: training is representative of target

h_{train} learner: $y_{est;i} = h_{train}(x_i)$

L Loss function

Data generation model:

$x_i \sim P_{\mathcal{X}}$

$f \rightarrow$ true labeling function, $y_i = f(x_i)$

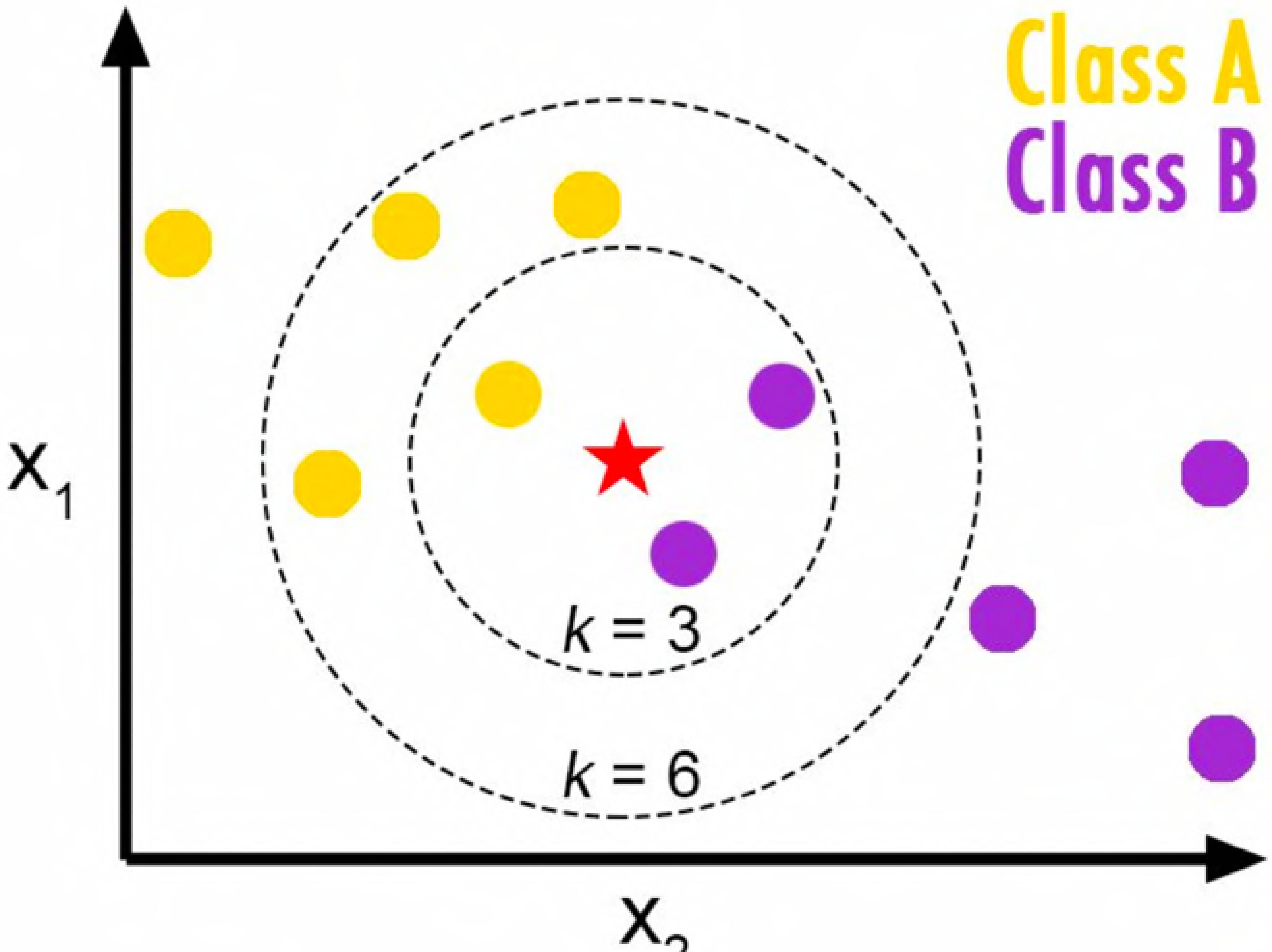
$L_{data,f}(h) \equiv P_{x \sim data}(h_{train}(x) \neq f(x))$

Shai and Shai, *Understanding ML: From Theory to Algorithms*, 2014, CUP

Example of supervised ML algorithm for classification

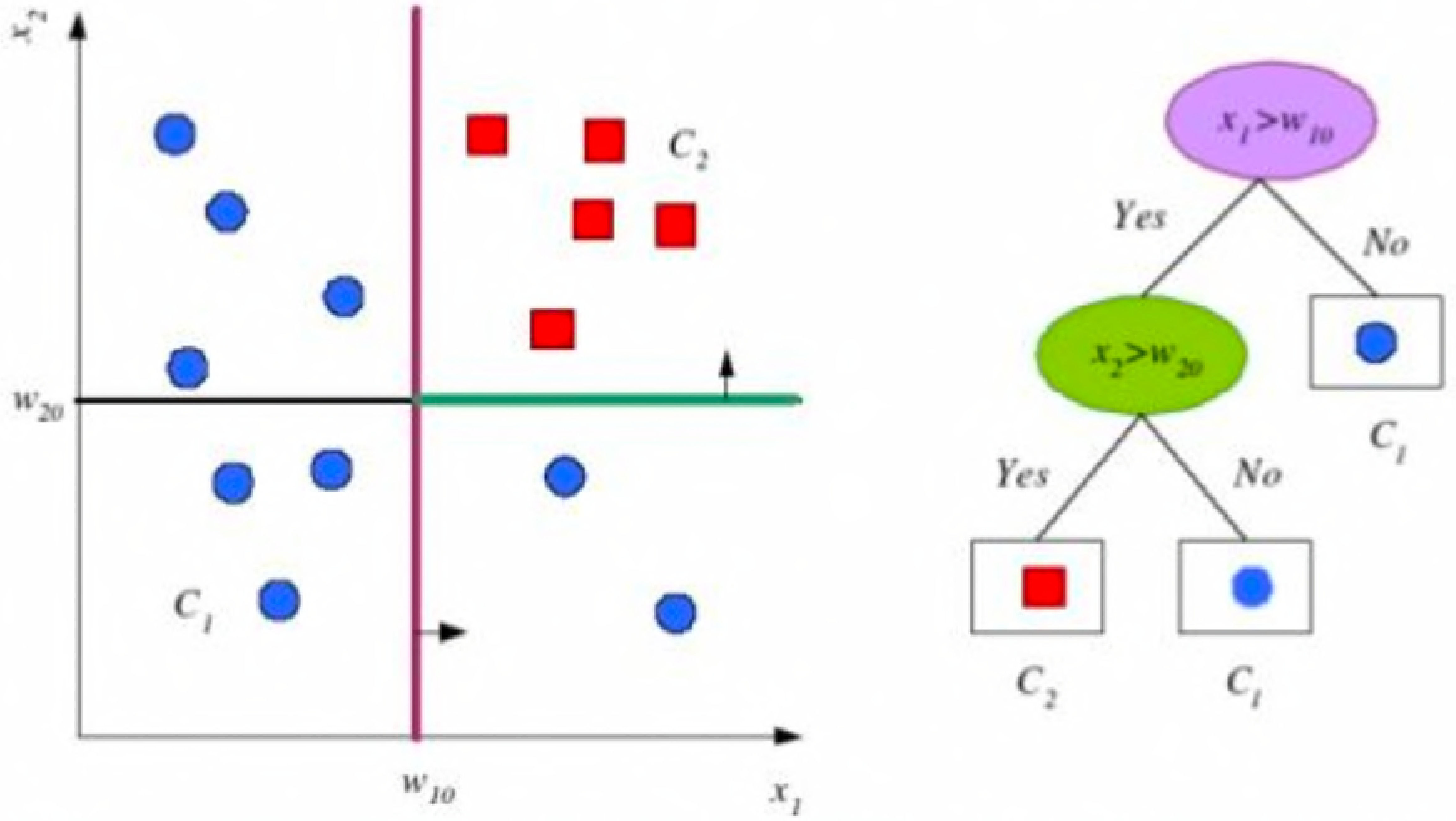
k-Nearest Neighbor (kNN)

Distance based



Example of supervised ML algorithm for classification

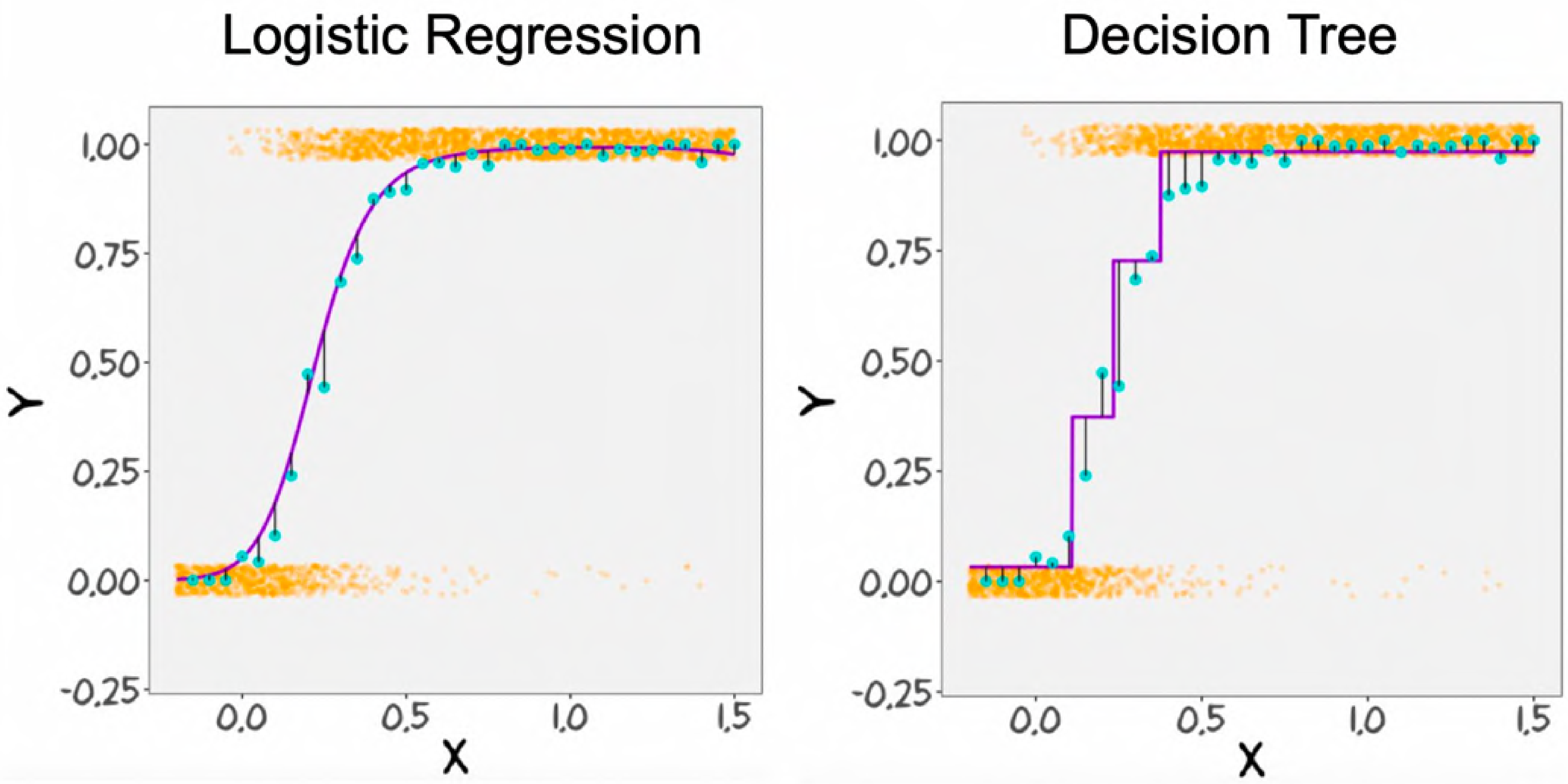
Decision Trees



http://www.lewisgavin.co.uk/Machine-Learning-Decision-Tree/

Logistic regression "vs" Decision Tree

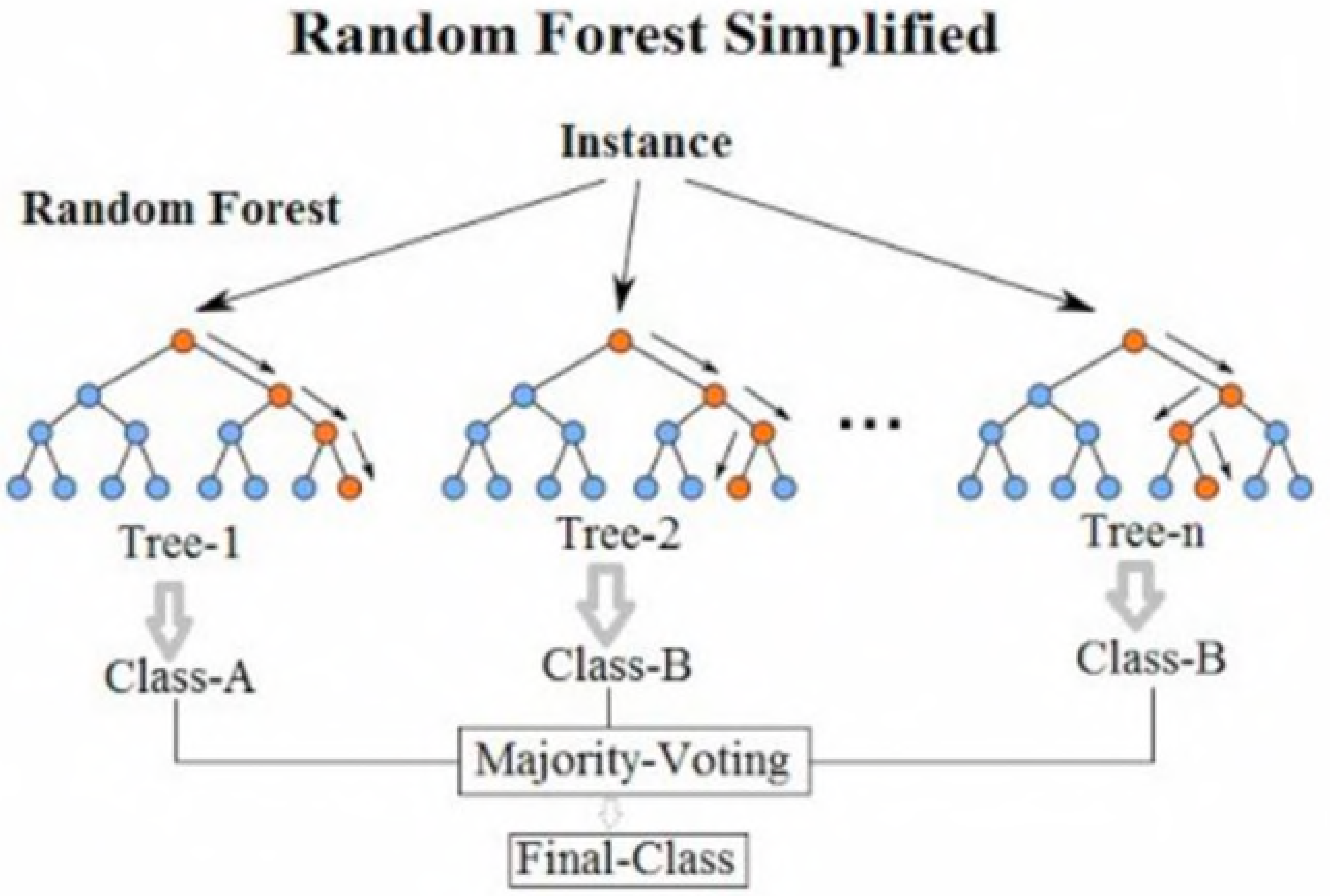
Caveat: You get what you ask for



Example of supervised ML algorithm for classification

Random Forests

Ensemble method

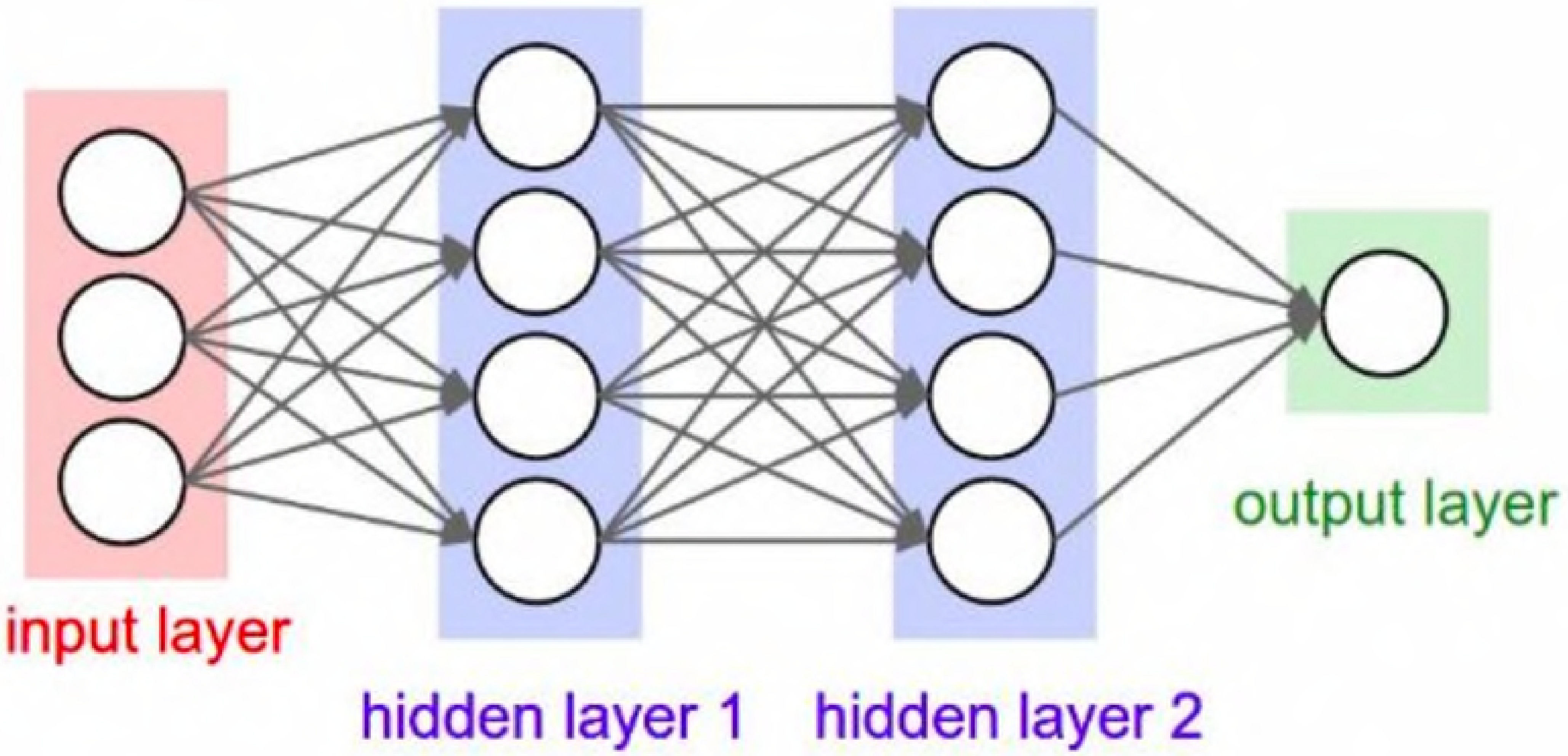


https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

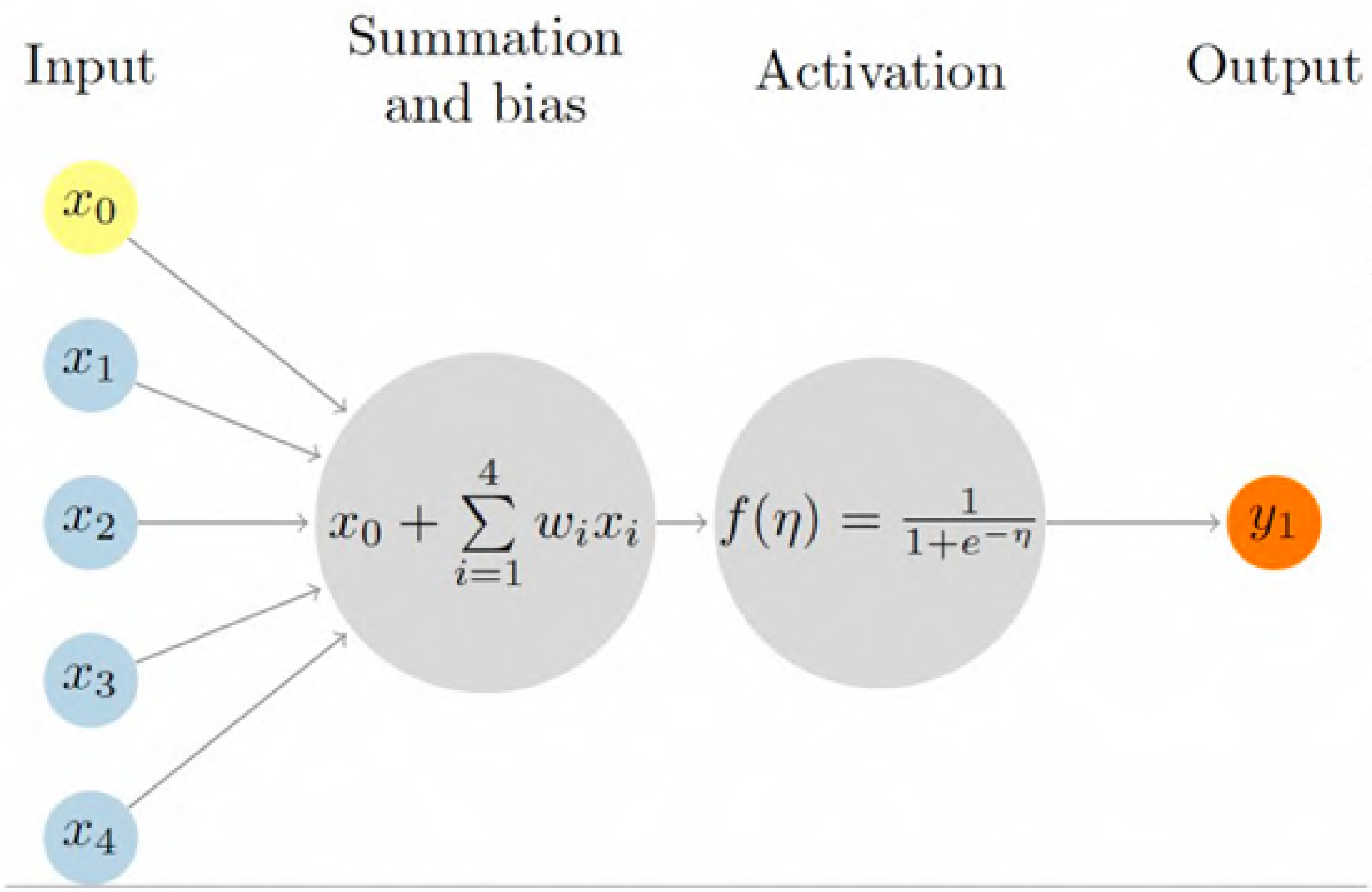
Example of supervised ML algorithm:

Deep Neural Network

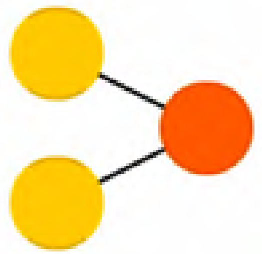
All layers internal to the network (not input or output layer) are considered hidden layers.



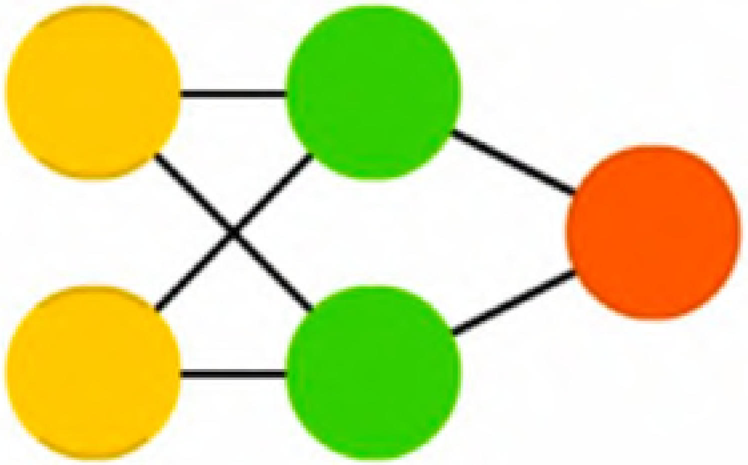
Single-Layer Perceptron $\equiv$ Logistic Regression



Good old days. Pretty much it, gets data, sums data, transforms data (i.e. sigmoid, logit, ...), outputs data.



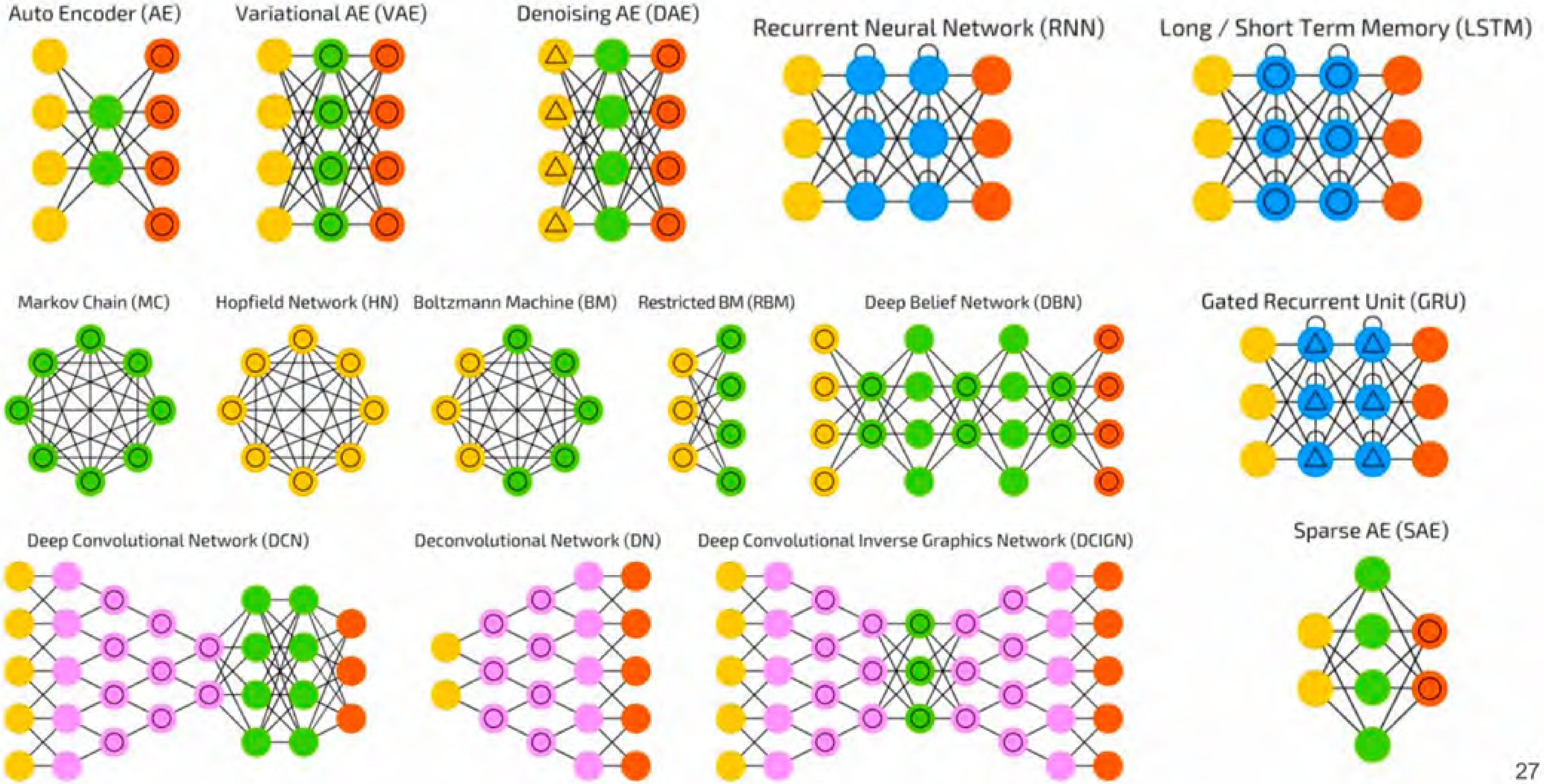
Multi-Layer Perceptron



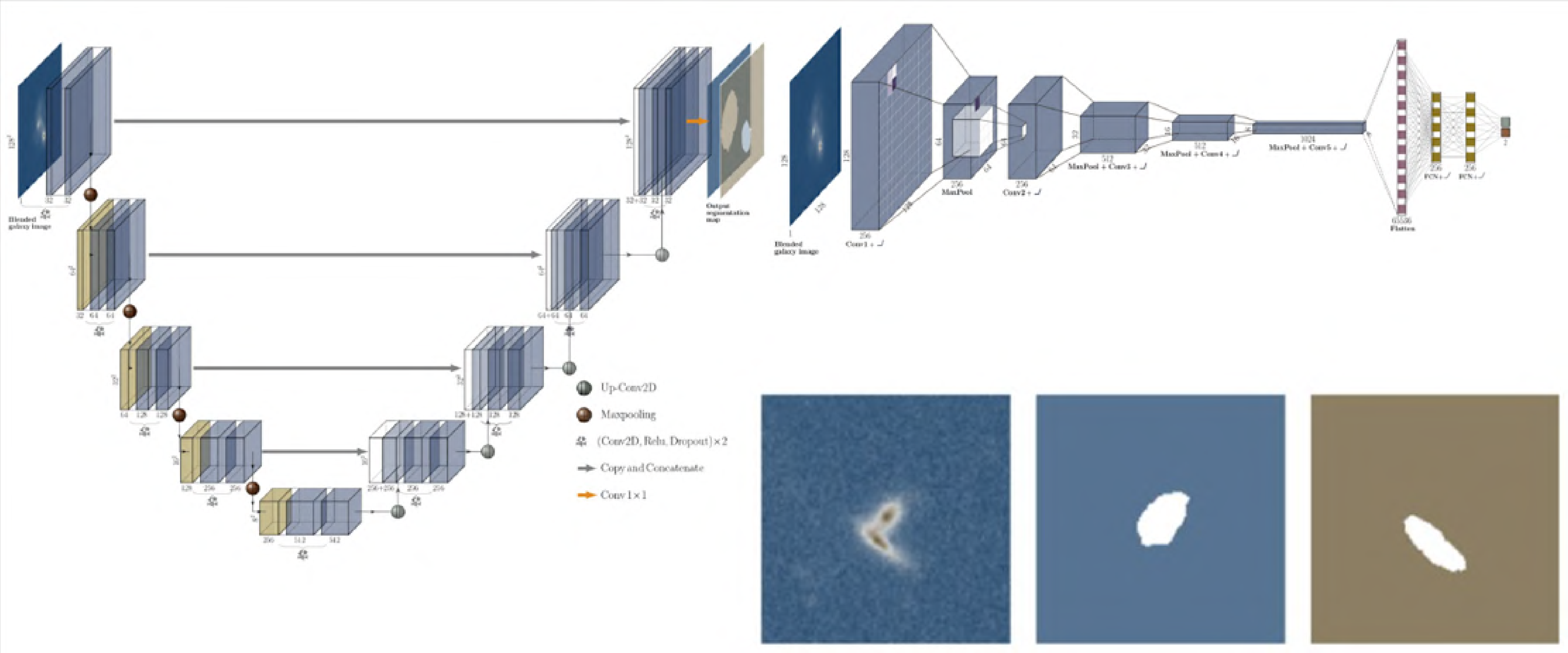
The 50's. Let's add some extra layer between input and output ("hidden layer").

Neural Networks

Architecture is the key



Galaxy Deblending via Deep Learning



Supervised ML model

data — training, target

χ — set of all samples, x

Y — set of possible labels, y

h_{train} — learner: $y_{est;i} = h_{train}(x_i)$

L — Loss function

Hypothesis: Training is representative of target

Data generation model:

$x_i \sim P_{\chi}$

$f \;\rightarrow$ true labeling function, $y_i = f(x_i)$

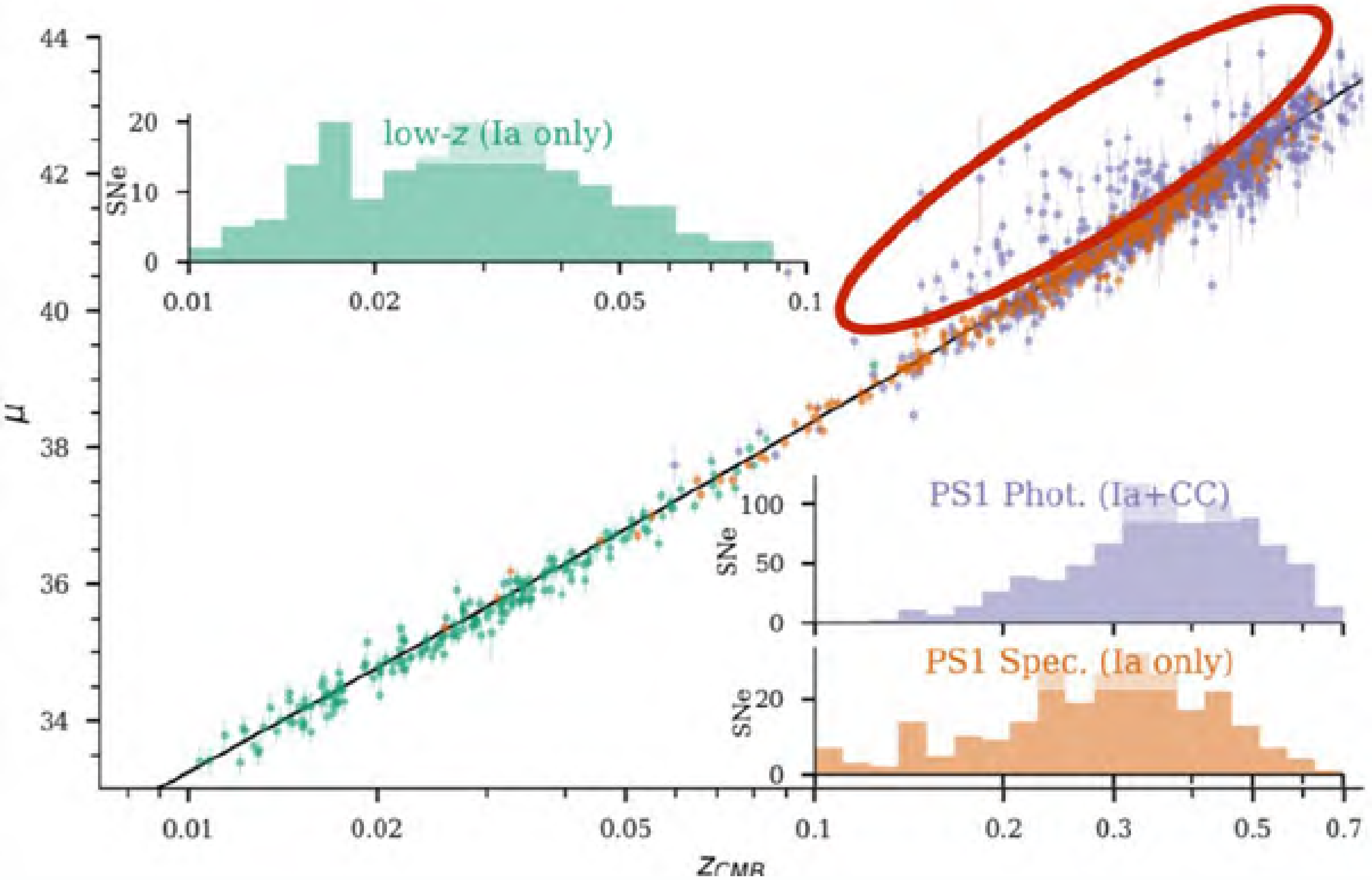
$L_{data,f}(h) \equiv P_{x\sim data}(h_{train}(x) \neq f(x))$

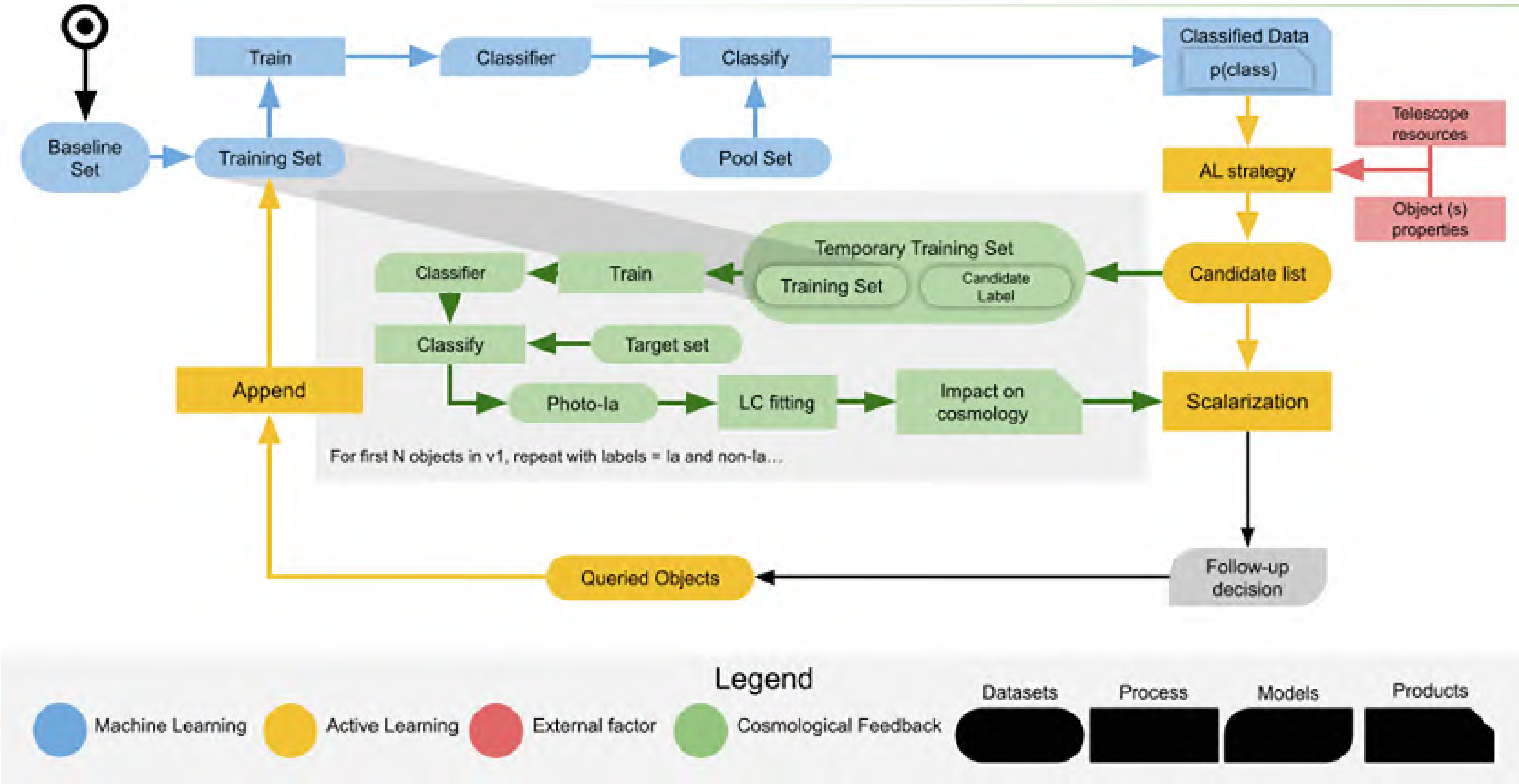
Shai and Shai, *Understanding ML: From Theory to Algorithms*, 2014, CUP

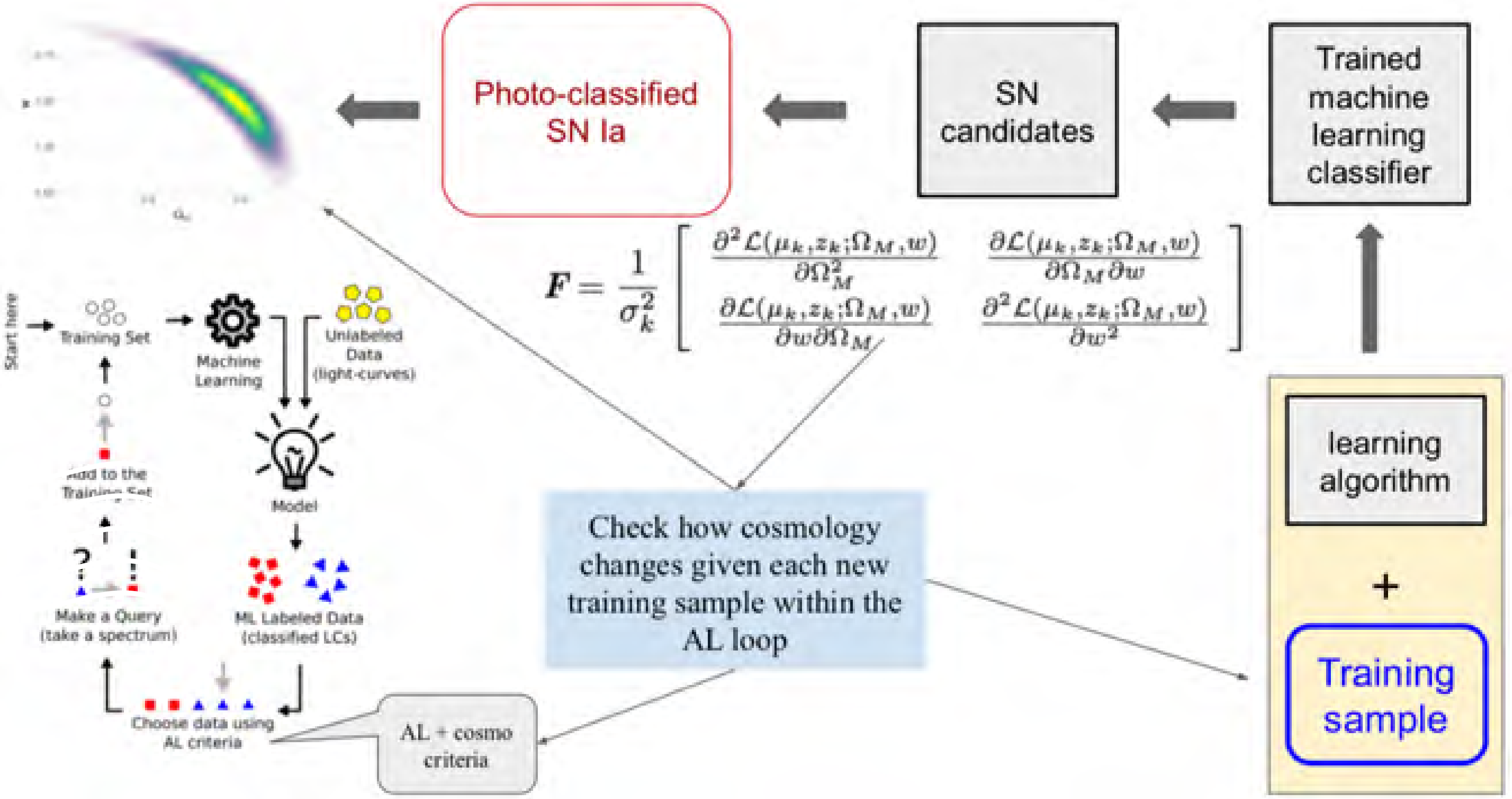
LSST DESC & COIN
RESSPECT
Recommendation System for Spectroscopic Follow-up

The REcommendation System for SPECTroscopic follow-up (RESSPECT) is a collaboration between COIN and LSST-DESC which aims to adapt active learning strategies for the construction of optimized training samples for supernova photometric classifcation in the context of LSST.

Supernova Cosmology photometric data







Resource allocation for extragalactic Transients

Challenges:

- Feature extraction of unevenly, noisy, incomplete multivariate time-series
- Online learning
- Scalable uncertainty quantification
- Domain-specific knapsack constraints, e.g. telescope time allocation, cosmology informed loss function

COIN
Cosmostatistics Initiative

The Cosmostatistics Initiative

The Cosmostatistics Initiative (COIN) is an international network which aims to create an interdisciplinary environment where collaborations between astronomers, statisticians and machine learning experts can flourish. The group utilizes a management model which can find parallel in technological start-ups: based on a dynamic, non-hierarchical and people-centric approach.

The LSST Dark Energy Science Collaboration

The LSST Dark Energy Science Collaboration (DESC) is an international collaboration preparing for a variety of cosmological analyses with the Large Synoptic Survey Telescope (LSST) data. In advance of LSST's first observations, DESC will help prepare for LSST science analysis, make synergistic connections with ongoing cosmological surveys and provide the dark energy community with state of the art analysis tools.

LSST DESC Dark Energy Science Collaboration

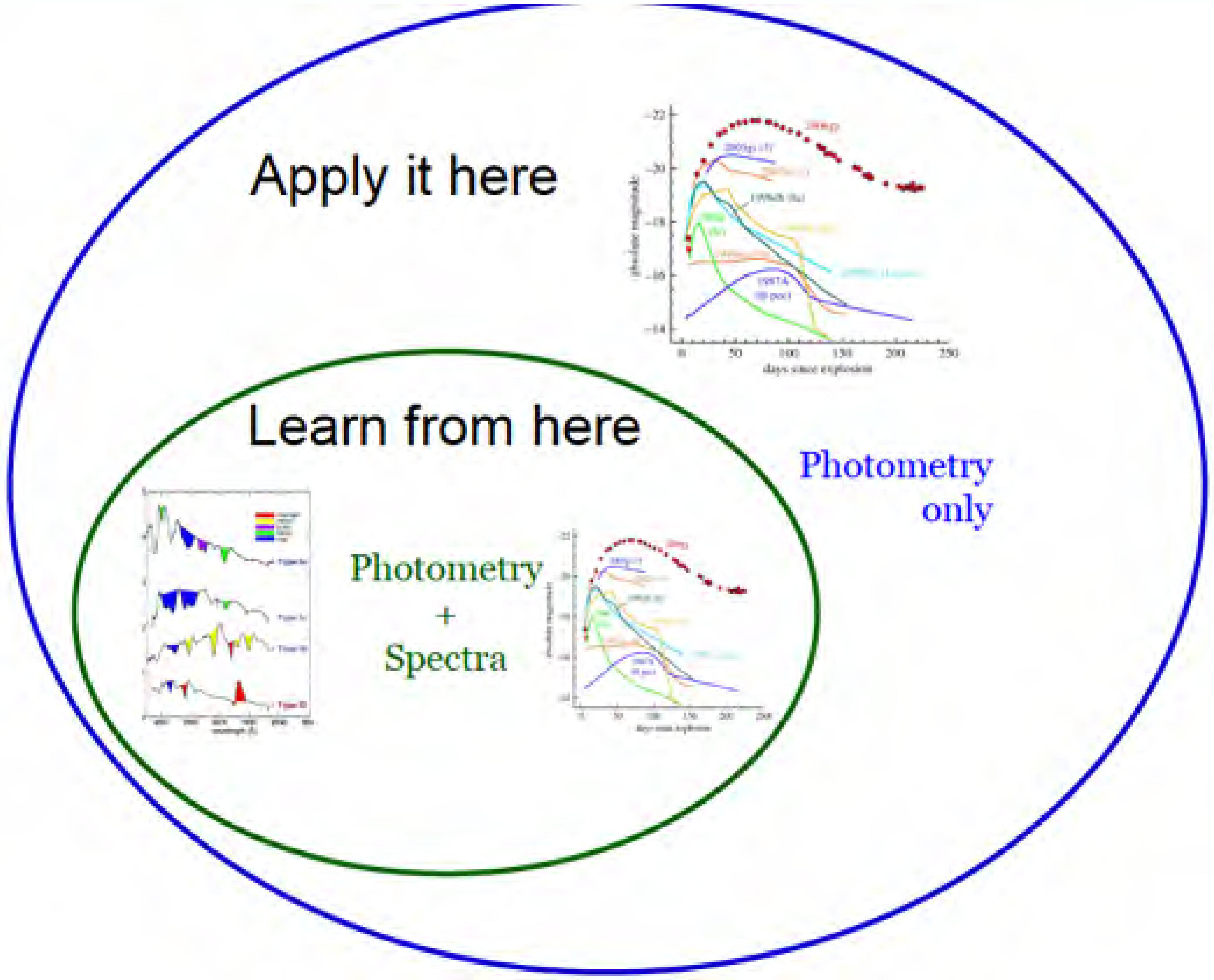
RESSPECT

LSST DESC & COIN RESSPECT

The REcommendation System for SPECTroscopic follow-up (RESSPECT) is a collaboration between COIN and LSST-DESC which aims to adapt active learning strategies for the construction of optimized training samples for supernova photometric classifcation in the context of LSST.

The team is formed by researchers from both collaborations who are working together in the development of a recommendation system which will enable informed decisions regarding the allocation of spectroscopic follow-up resources and consequent optimized scientific results from purely photometric samples.

LSST DESC & COIN
RESSPECT
Recommendation System for Spectroscopic Follow-up



DARK ENERGY
SURVEY

LSST

Supernova Cosmology photometric data

~ 3,000 cosmological useful SNe

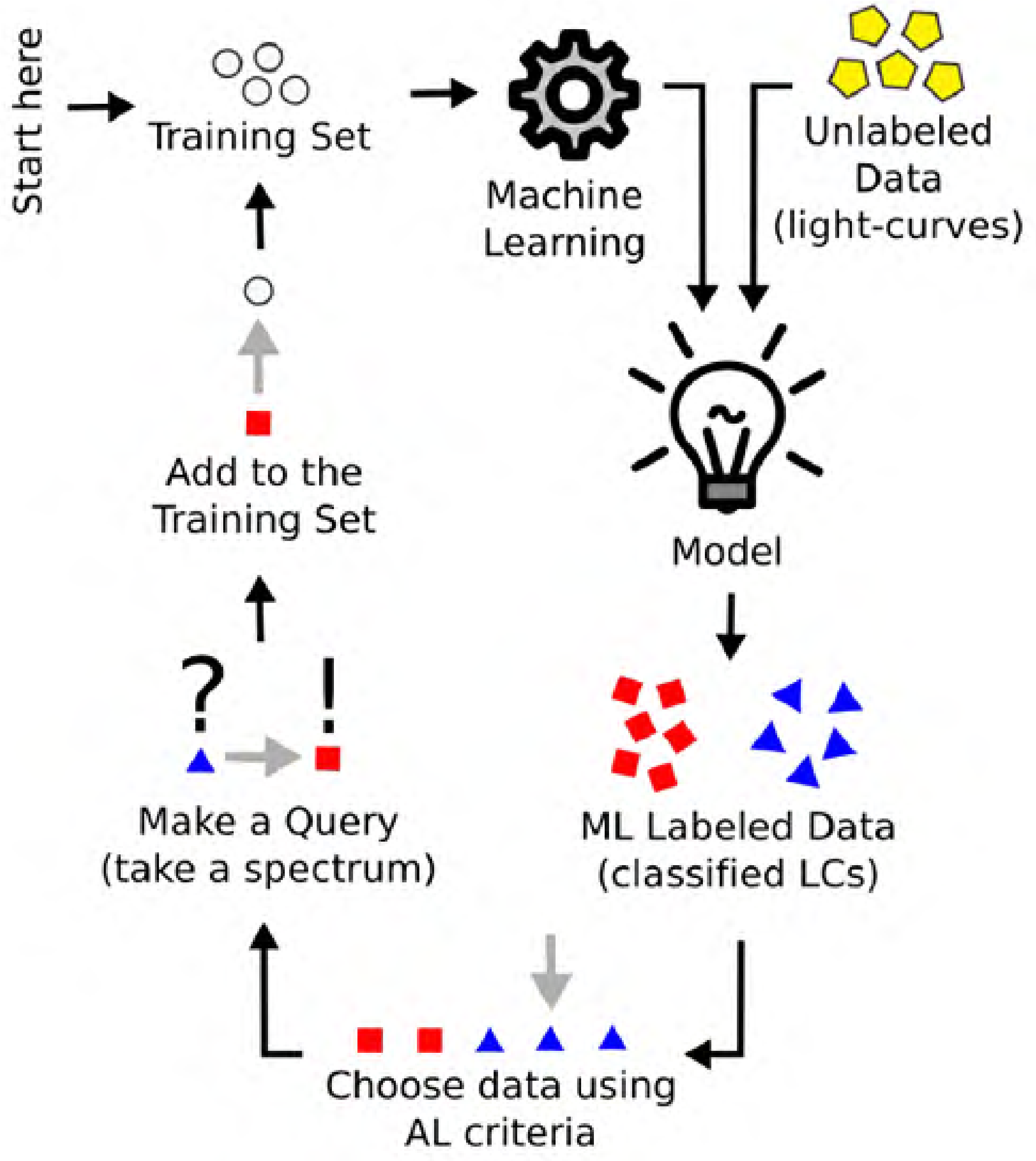
~ 100,000 cosmological useful SNe

LSST DESC & COIN
RESSPECT
Recommendation System for Spectroscopic Follow-up

Challenges:

- Window of Opportunity for Labelling
- Evolving Samples - We must make query decisions before we can observe the full LC
- Multiple Instruments
- Evolving Costs - Observing costs for a given object changes as it evolves.

Active Learning



OUTLINE

- Generalized Linear Models
- Statistical Learning
- Discovering stellar clusters

Star Clusters

The cluster members share common properties, like age, distance from the Sun, and velocity, span ages from ~ 1Myr to > 10 Gyrs.

Laboratories for Stellar Evolution Models!

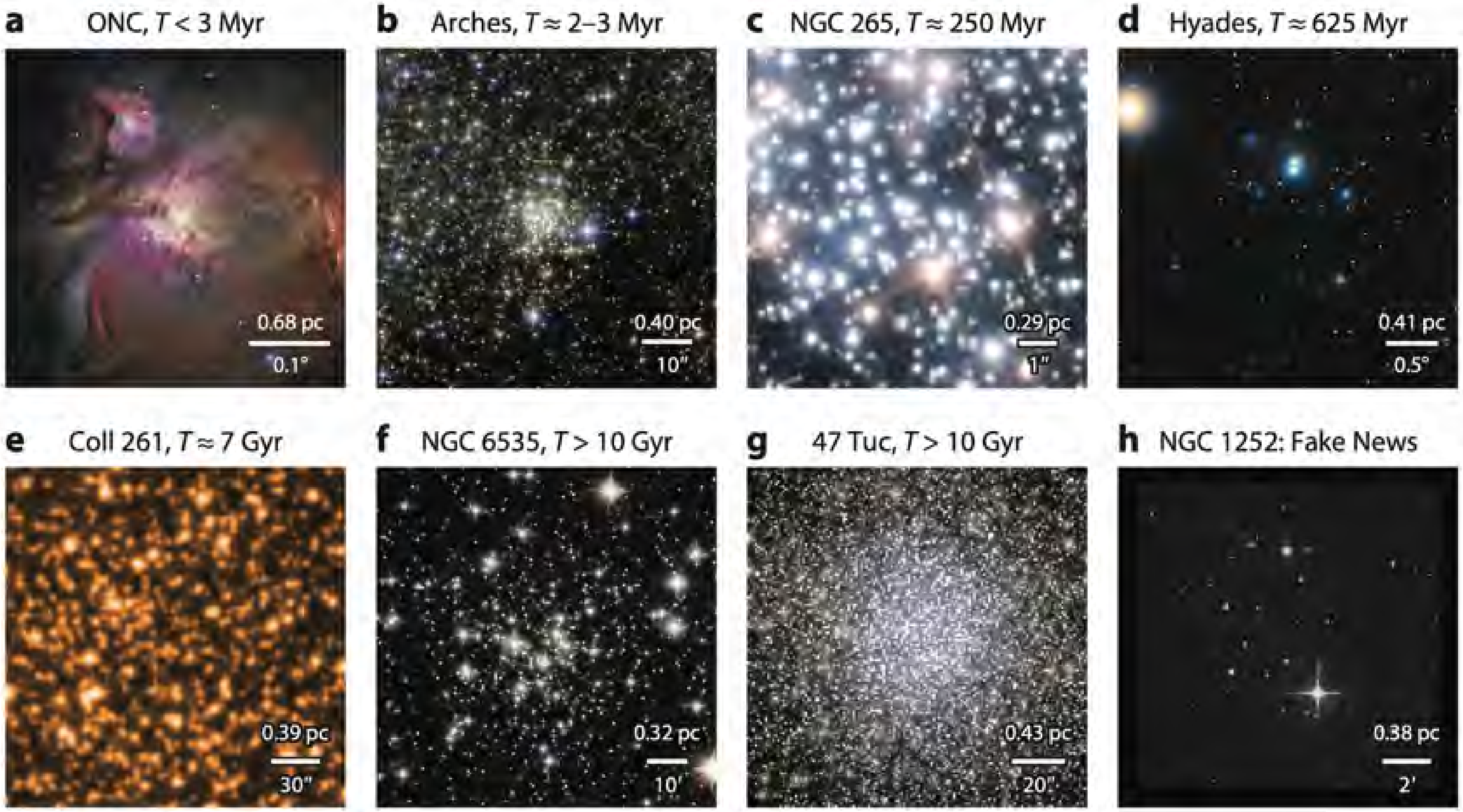
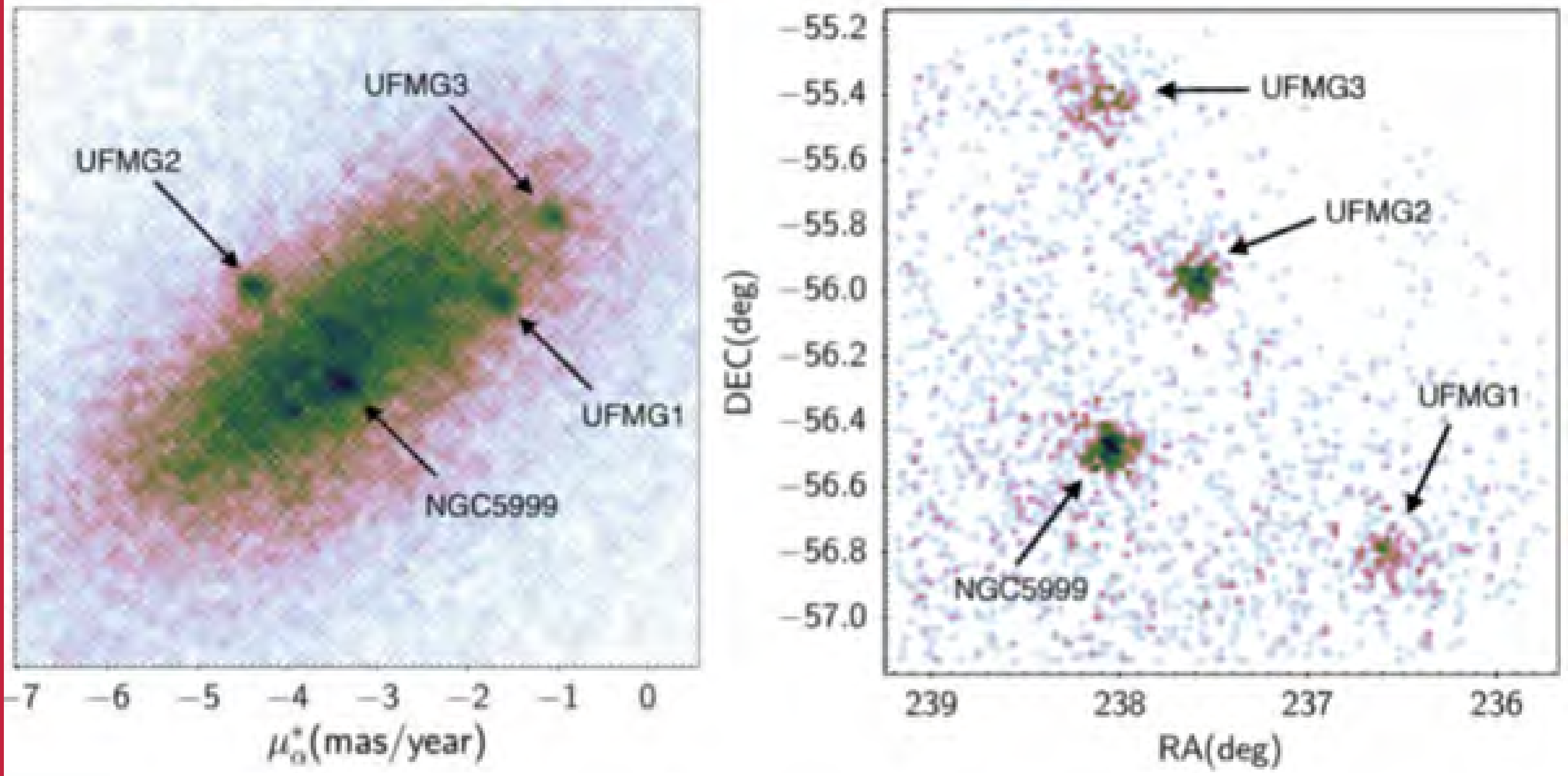


Figure from *Star Clusters Across Cosmic Time*
M.R. Krumholz, C. F. McKee, J. Bland-Hawthorn
Annual Review of Astronomy and Astrophysics 2019 57:1, 227-303

A needle in a haystack

There are still associations hidden in plain sight

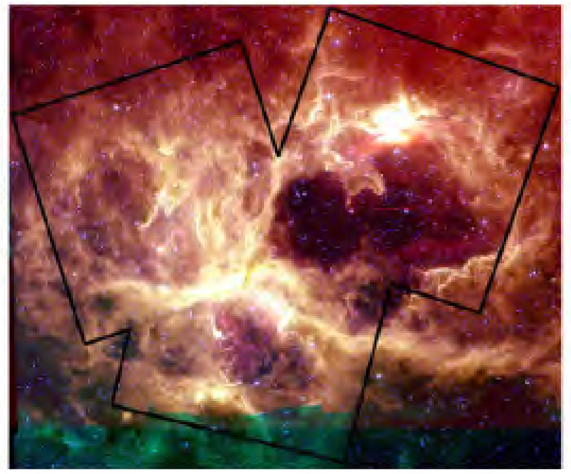


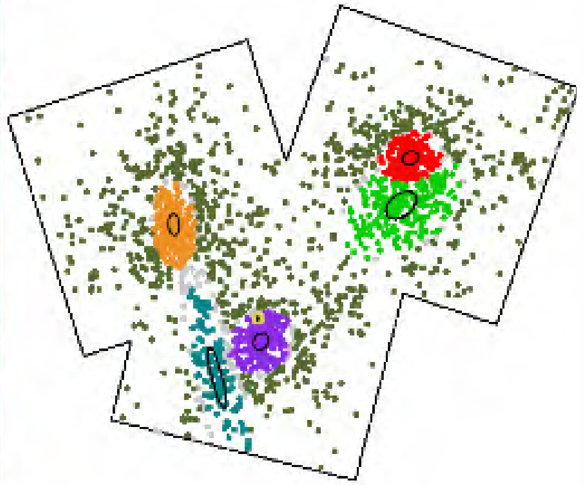
Ferreira, et. al. MNRAS, 2019, 483, 4 p. 5508

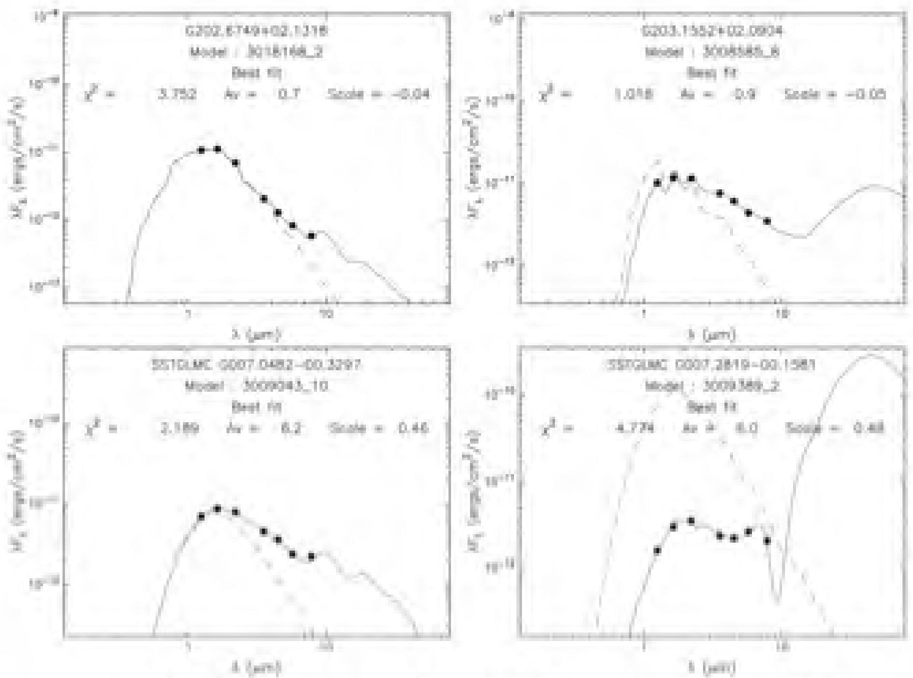
Spitzer's GLIMPSE survey
|b| < 1-2 deg

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

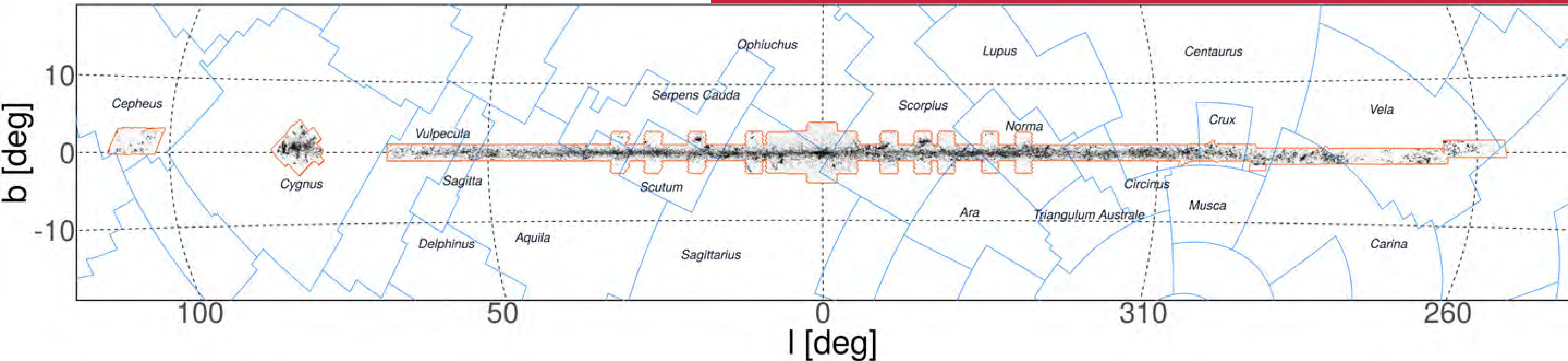
Published 2021 June 2 •

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

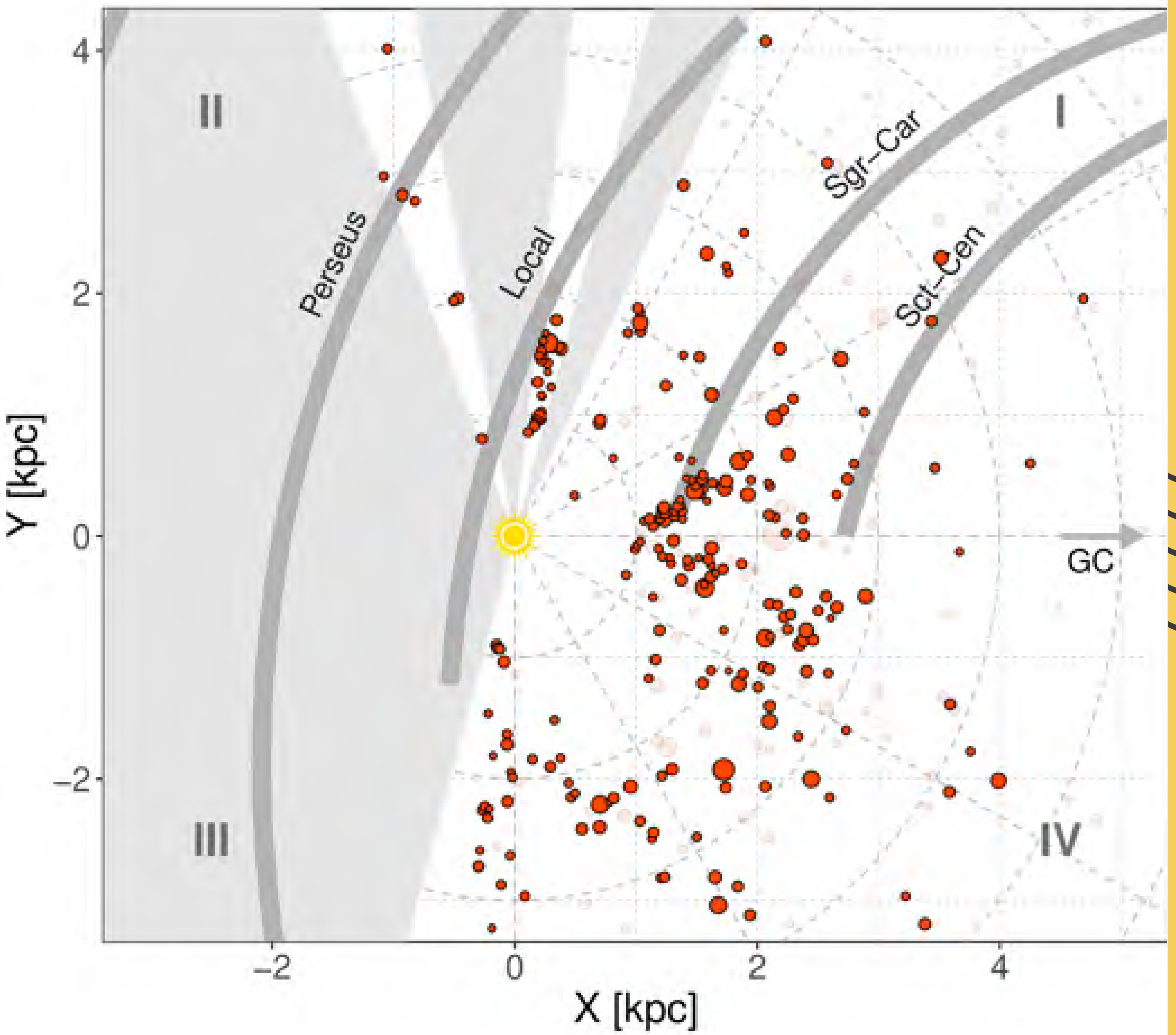


SPICY - 120,000 YSOs discovered in the Galaxy

ASTRO-AWARE STATISTICAL LEARNING

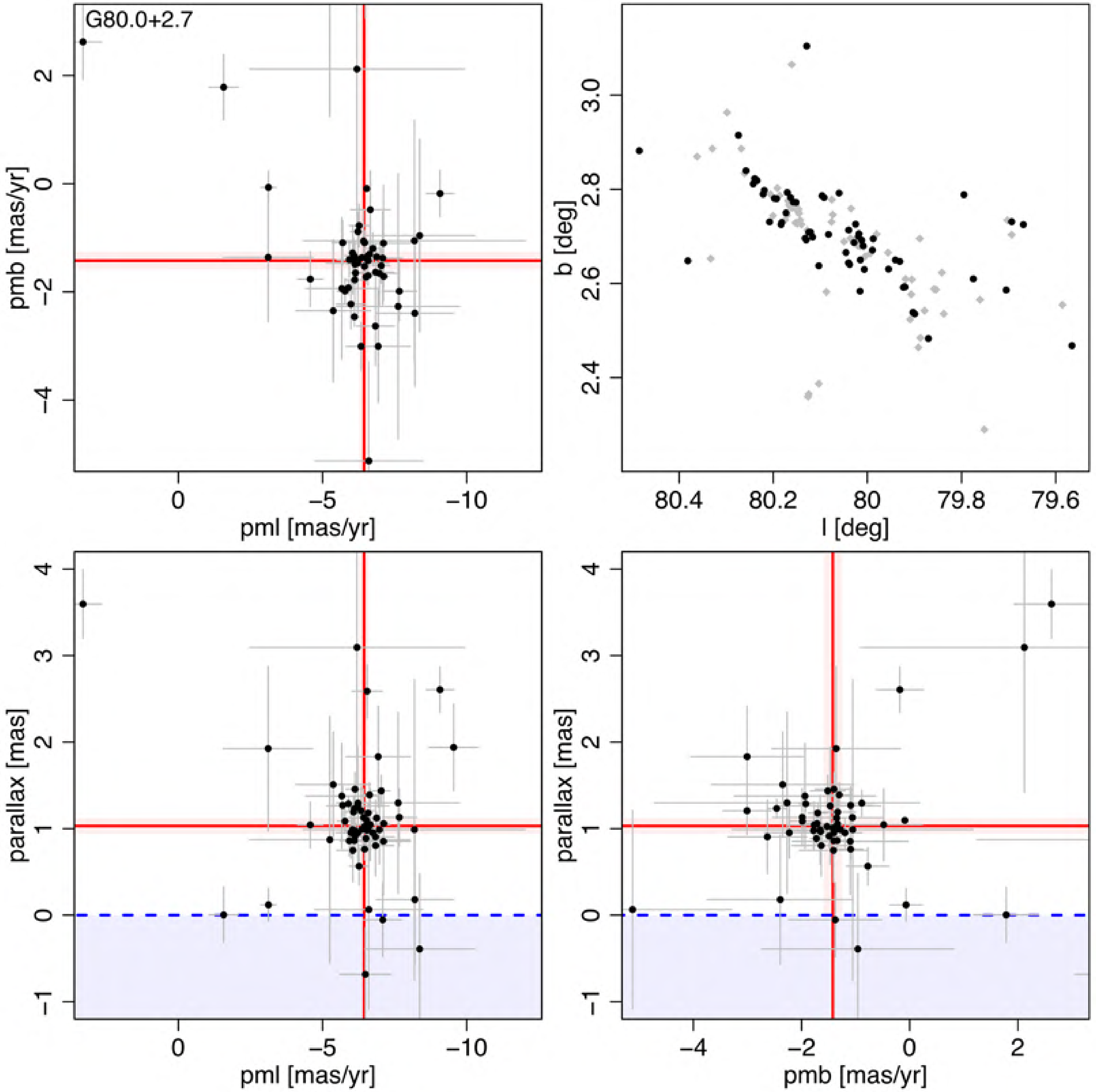
Challlenges:

- Domain knowledge regularization - statistical clusters vs Astronomical ones
- Heteroscesdastic uncertainties with known variance, covariance
- Selection effects, missing not at random

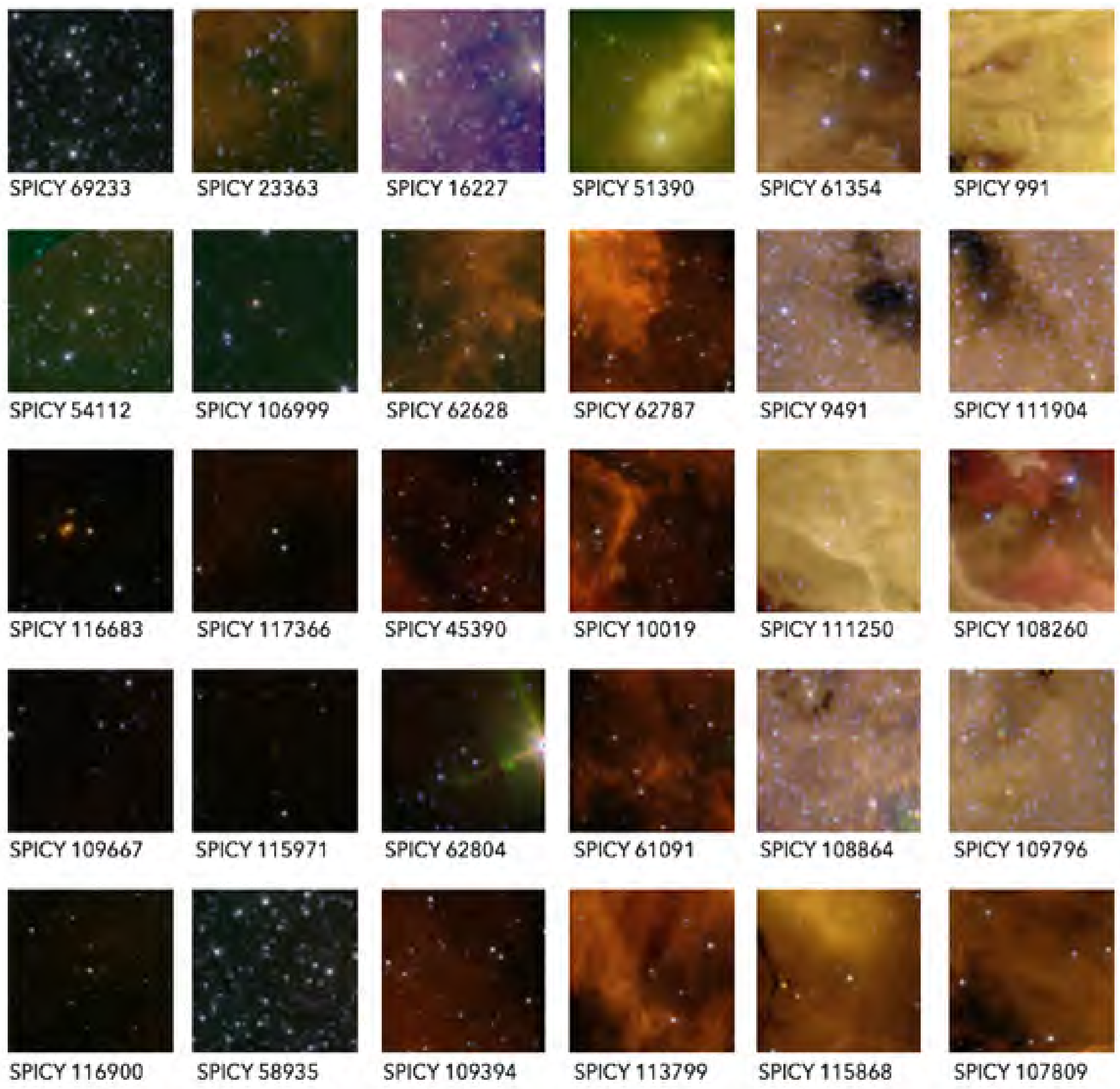


Hierarchical Bayesian Models

ASTROMETRIC PROPERTIES OF THE STELLAR GROUPS



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



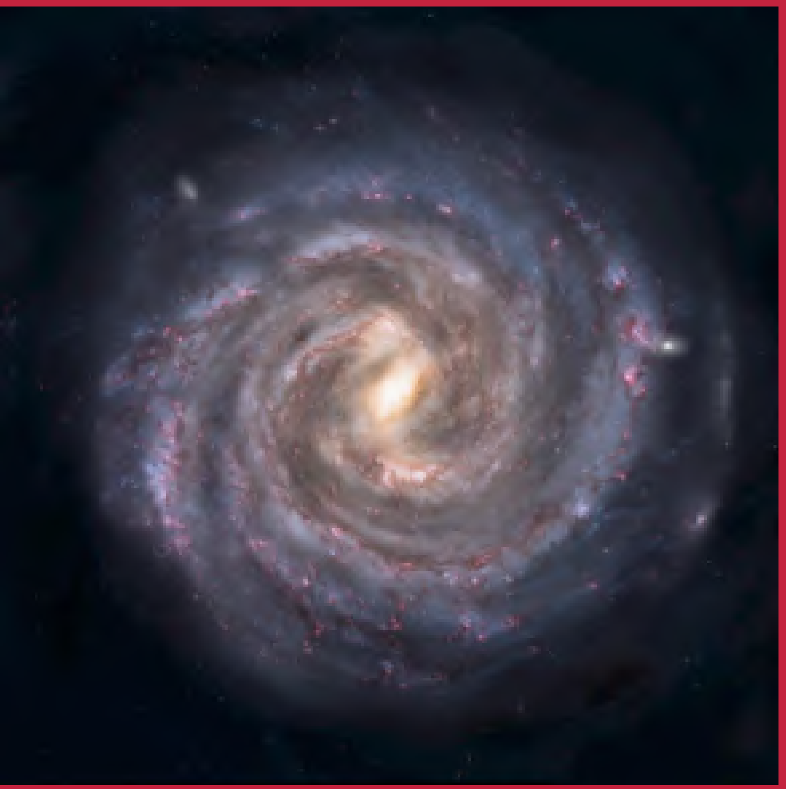
117,224 PNG stamps

251GB album built via PostgreSQL

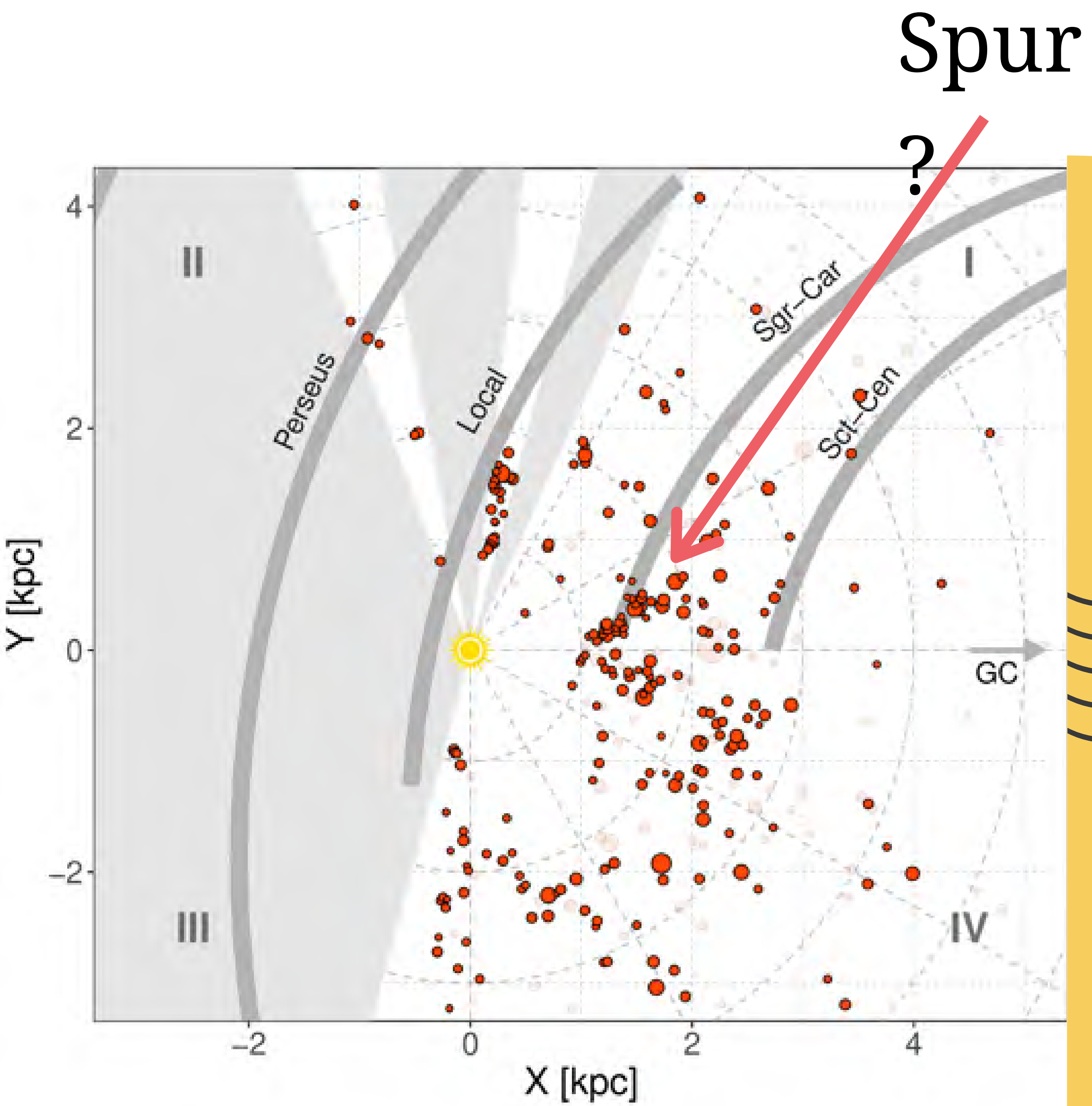
Potential testbed for computer vision, texture analysis, novelty detection, feature extration, etc.

Spatial distribution of YSO groups

Good tracers of star forming regions and galactic structure



Independent probe of spiral arms structure



Spiral arms are not smooth,
continuous features.

This substructure appears associated with
much of the star formation in the arm.

NGC 5236 (Larsen & Richtler 1999)

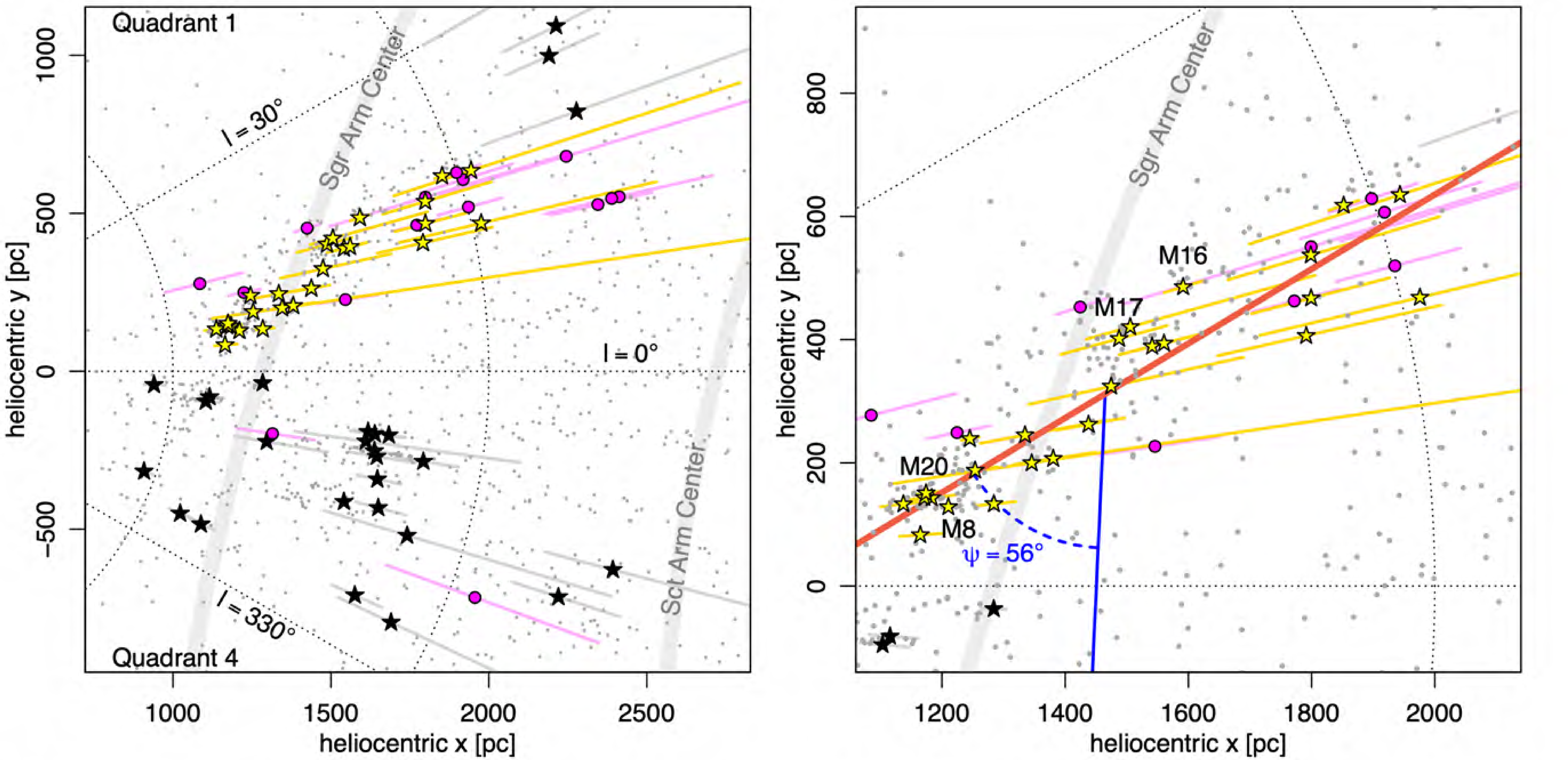


Fig. 3. Galactic map of YSO groups (star symbols), masers (magenta circles), and non-clustered SPICY YSO candidates (gray points) in heliocentric *xy* coordinates. The right panel shows a zoomed-in view. Groups associated with the structure are color-coded yellow, while others are black. The spiral-arm 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.

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