Arya Farahi
Department of Statistics and Data Science

Toward Trustworthy Probabilistic Machine Learning

Bayesian Deep Learning for
Cosmology and time-domain astronomy
June, 2022

The University of Texas at Austin
Department of Statistics and Data Sciences
College of Natural Sciences

What have we covered so far?

Computational models to estimate the posterior distributions?

Novel model classes.

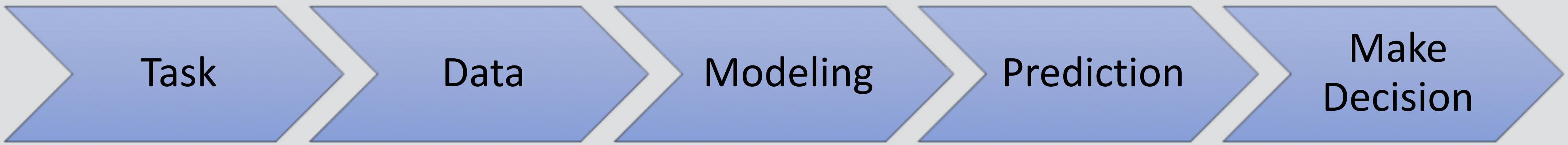
Applications of (Bayesian? and non-Bayesian) Deep Learning models to time-domain astronomy and cosmology?

How to define and estimating the uncertainties?

In this talk, I will discuss a computational framework that evaluates the trustworthiness of a probabilistic model.

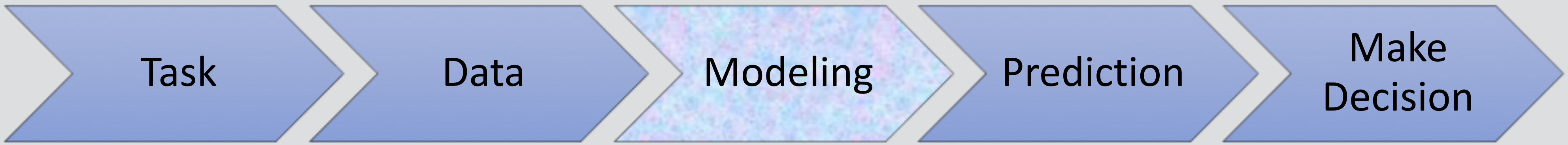
A Key Challenge

Is our AI-system fair?



A Key Challenge

Is our AI-system fair?



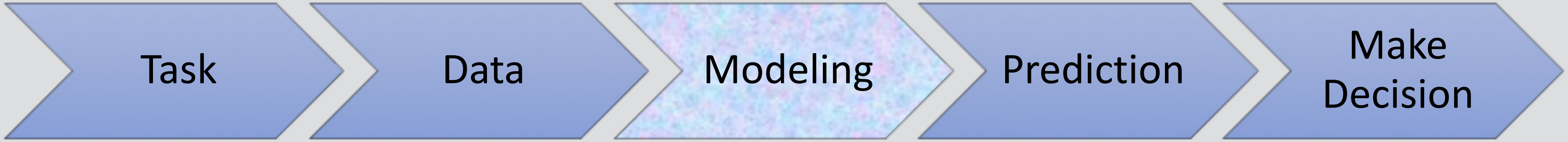
What are the potential sources of bias?

A decision-maker or a scientist makes decision and acts based on **the information** provided by a model. It is the job of the modeler to guaranteed the trustworthiness of the provided information.

A Key Challenge

unbiased (inference) and
optimal (decision making)

Is our AI-system ~~fair~~?



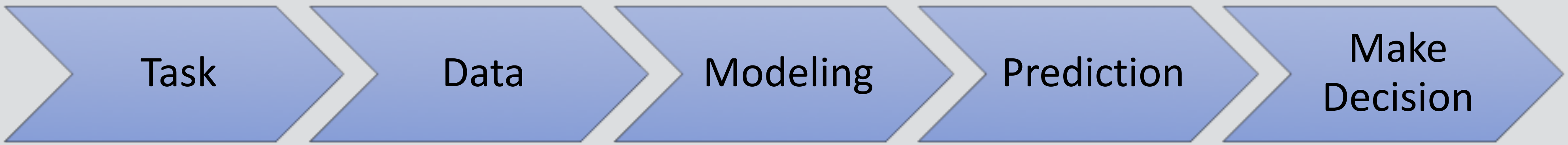
What are the potential sources of bias?

A decision-maker or a scientist makes decision and acts based on **the information** provided by a model. It is the job of the modeler to guaranteed the trustworthiness of the provided information.

A Key Challenge

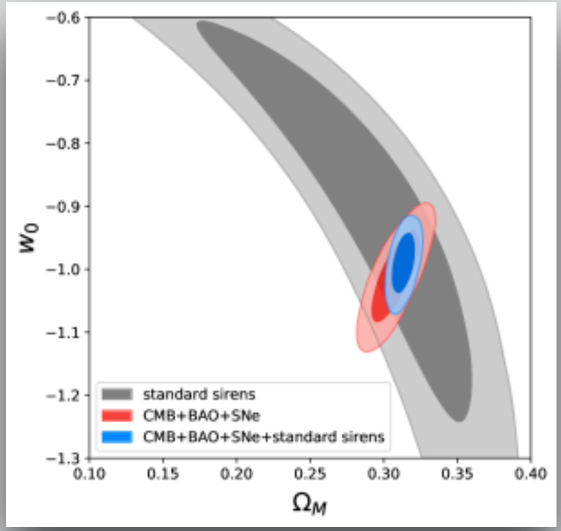
unbiased (inference) and
optimal (decision making)

Is our AI-system ~~fair~~?



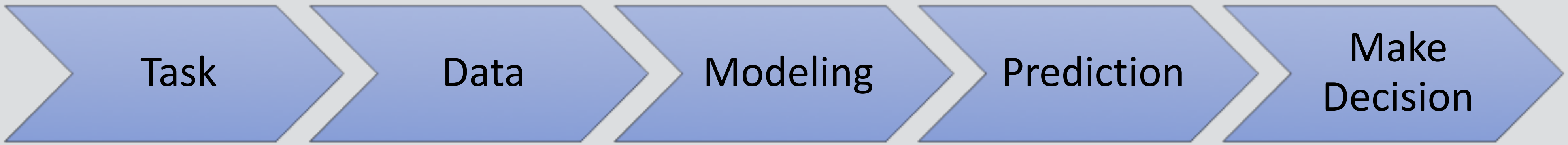
Selecting a sample of transient events for a follow-up study, given the fact that follow-up resources are limited?

$$\min_{a} \mathbb{E}[\mathrm{FOM}(a)]$$



A Key Challenge

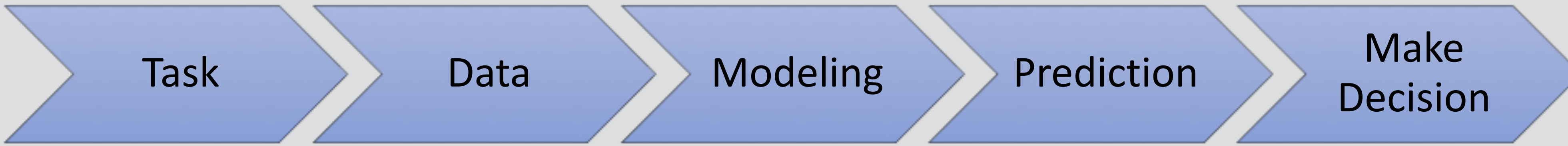
How to evaluate trustworthiness of a probabilistic classifier?



Accuracy, Precision,
FPR, FNR,
AUC, Brier Score
Log-loss, Entropy

A Key Challenge

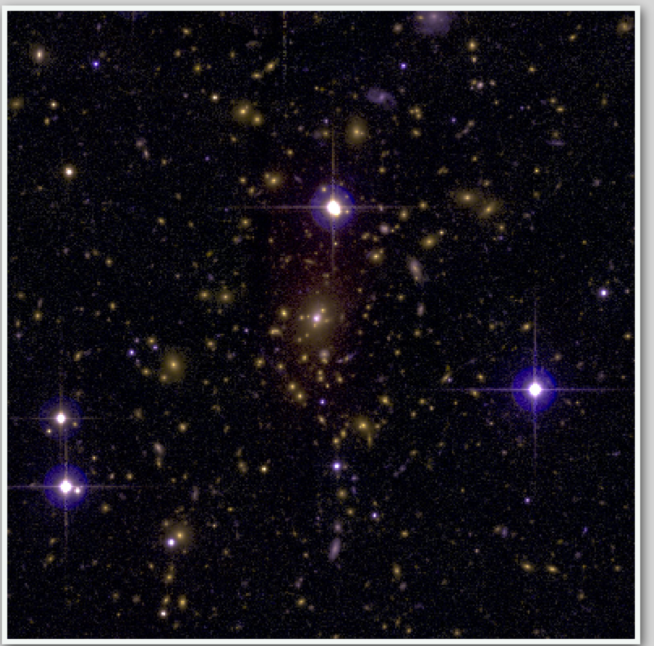
How to evaluate trustworthiness of a probabilistic classifier?



Trustworthiness evaluation ≡ Goodness-of-fit evaluation

~~Accuracy, Precision,~~
~~FPR, FNR,~~
~~AUC, Brier Score~~
~~Log-loss, Entropy~~

A recipe for probing properties of dark matter and dark energy with galaxy clusters



Abell 1835, Credit: Allen et al. (2011)

Finding a set of clusters

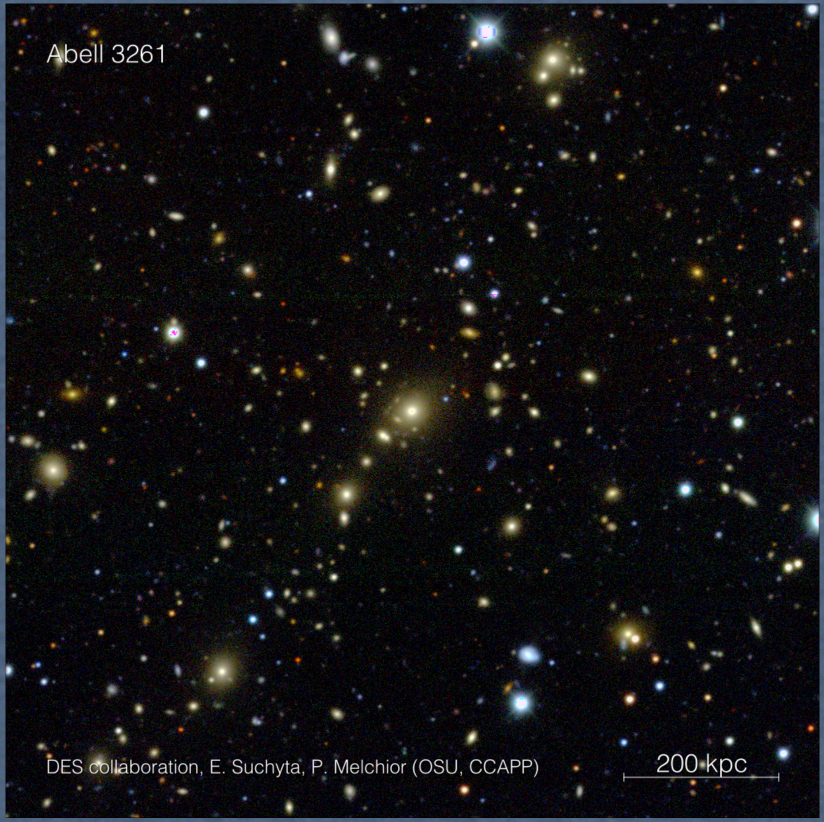
Measuring their observable quantities

Mapping observables to the host halo mass

Cluster Finding Algorithm

redMaPPer cluster finding algorithm

Overdensity of red galaxies on the sky



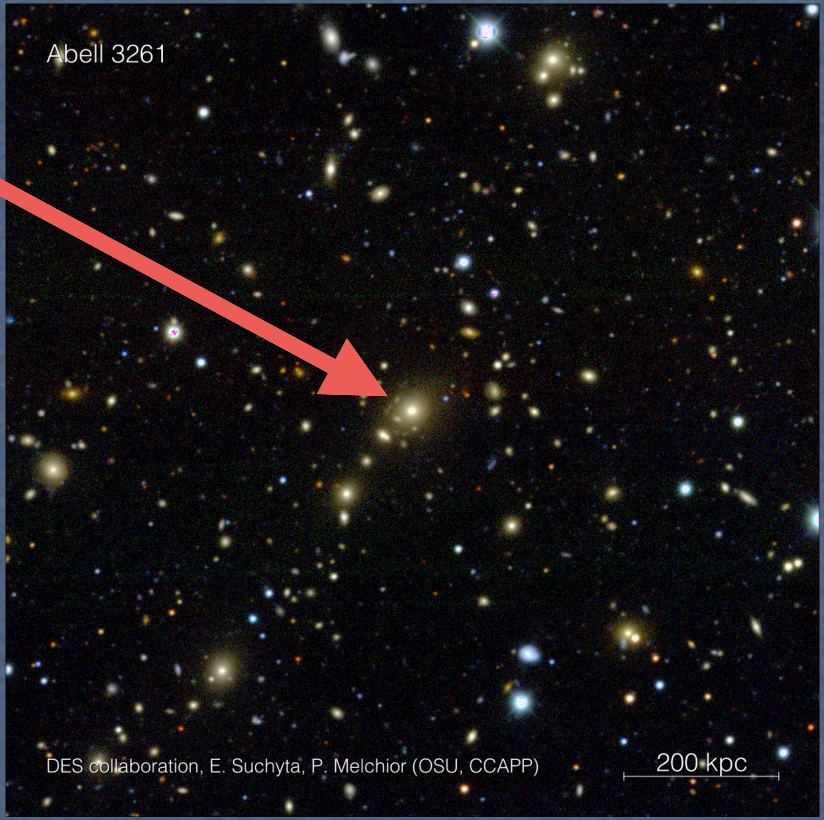
Rykoff et al., (ApJ, 2014), McClintock et al. (MNRAS, 2019), Farahi et al. (MNRAS, 2019)

Cluster Finding Algorithm

redMaPPer cluster finding algorithm

Overdensity of red galaxies on the sky

Find a candidate central galaxy



Rykoff et al., (ApJ, 2014), McClintock et al. (MNRAS, 2019), Farahi et al. (MNRAS, 2019)

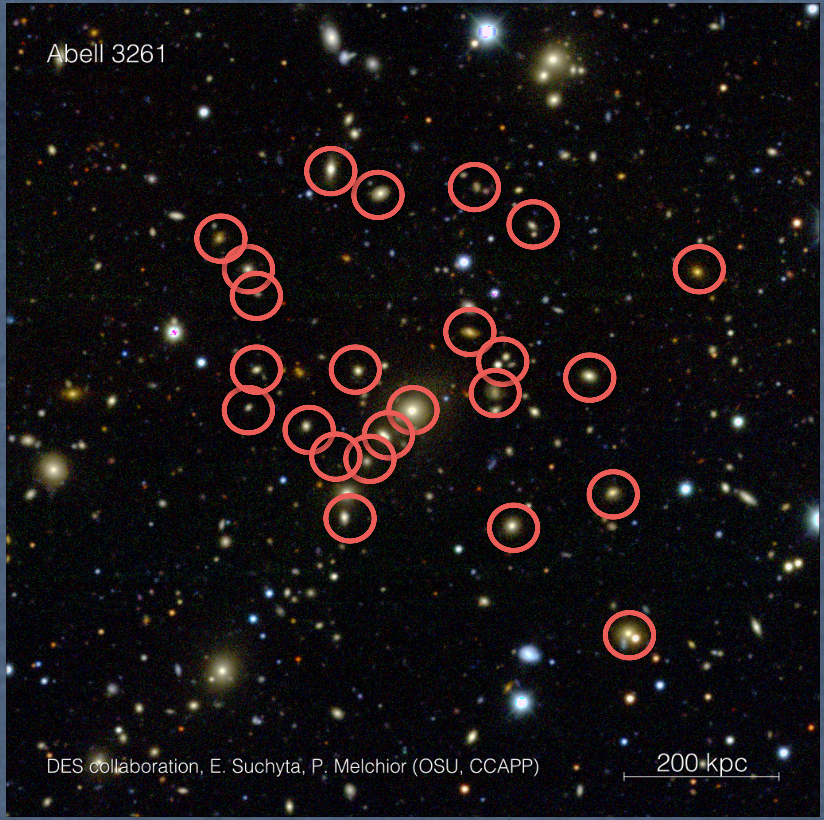
Cluster Finding Algorithm

redMaPPer cluster finding algorithm

Overdensity of red galaxies on the sky

Find a candidate central galaxy

Assign a membership probably
to each galaxy



Rykoff et al., (ApJ, 2014), McClintock et al. (MNRAS, 2019), Farahi et al. (MNRAS, 2019)

Cluster Finding Algorithm

redMaPPer cluster finding algorithm

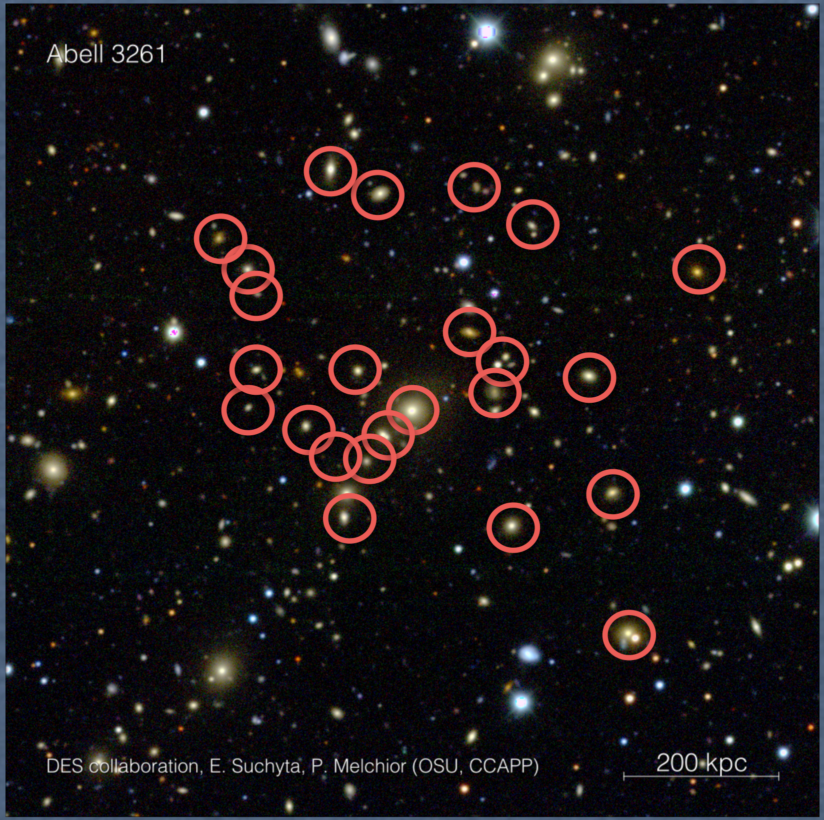
Overdensity of red galaxies on the sky

Find a candidate central galaxy

Assign a membership probably
to each galaxy

Estimate the number of red galaxies

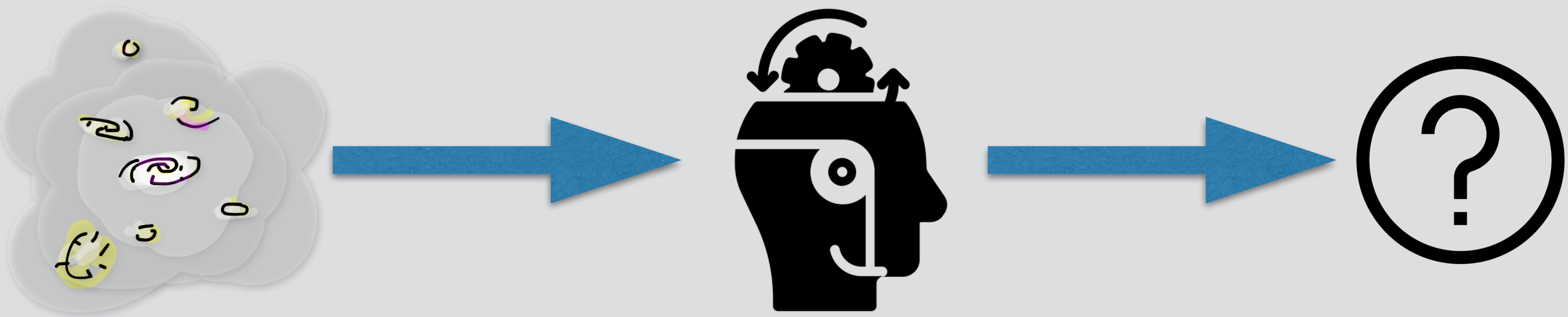
$$\lambda_{\mathrm{RM}} = \sum p_{mem}$$



Rykoff et al., (ApJ, 2014), McClintock et al. (MNRAS, 2019), Farahi et al. (MNRAS, 2019)

Trustworthy Classifier

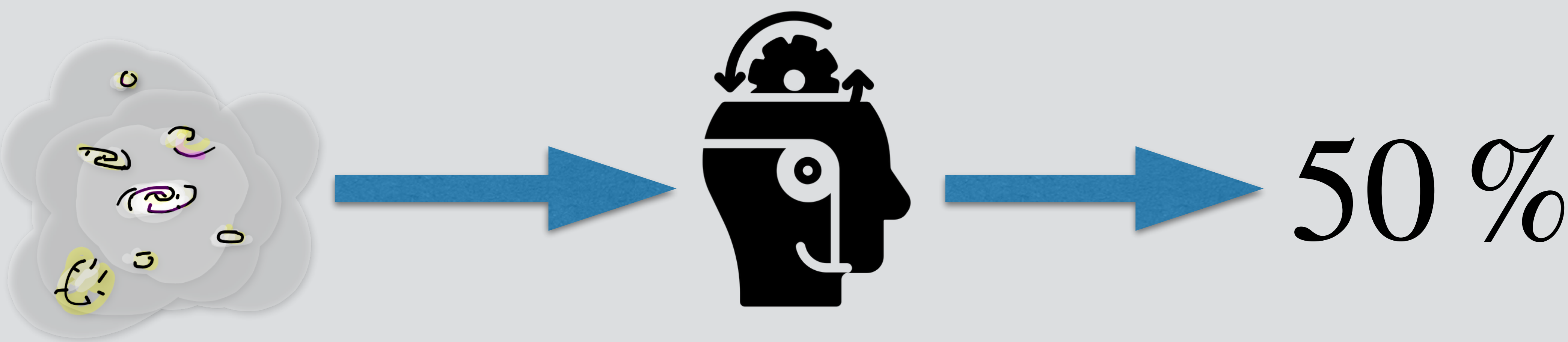
Example. Consider a model predicts the probability of membership for a set of galaxies given a BCG.



Trustworthy Classifier

Example. Consider a model predicts the probability of membership for a set of galaxies given a BCG.

→ for a subset of galaxies RM(x) = 50%



Trustworthy Classifier

Example. Consider a model predicts the probability of membership for a set of galaxies given a BCG.

→ for a subset of galaxies RM(x) = 50%

→ this subset consists of

- 10% star forming galaxies (group A)
- 90% quenched galaxies (group B)

Trustworthy Classifier

Example. Consider a model predicts the probability of membership for a set of galaxies given a BCG.

→ for a subset of galaxies RM(x) = 50%

→ this subset consists of

- 10% star forming galaxies (group A)
- 90% quenched galaxies (group B)

→ An observational study finds that the frequency of being gravitationally bound is

- 95% for group A
- 45% for group B

Trustworthy Classifier

→ for a subset of patients RM(x) = 50%

→ this subset consists of

- 10% star forming galaxies (group A)
- 90% quenched galaxies (group B)

→ An observational study finds that the frequency of being gravitationally bound is

- 95% for group A
- 45% for group B

→ **On average**, the probability of 0.5, on average, is calibrated

$$\frac{10 \times 95\% + 90 \times 45\,\%}{100} = 50\,\%$$

Trustworthy Classifier

→ for a subset of patients RM(x) = 50%

→ An observational study finds that the frequency of being gravitationally bound is

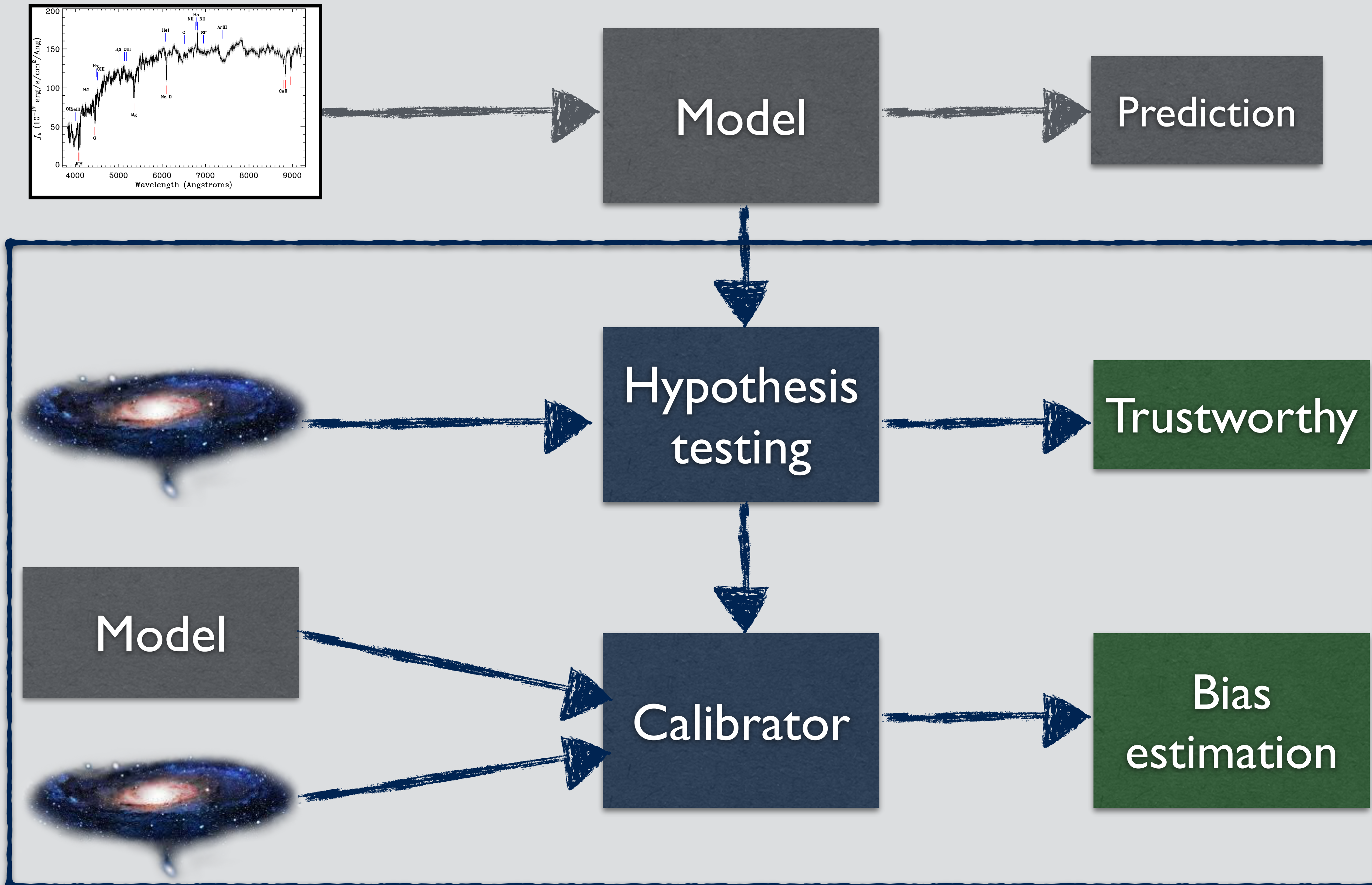
- 95% for group A
- 45% for group B

→ **On average**, the probability of 0.5 is properly calibrated

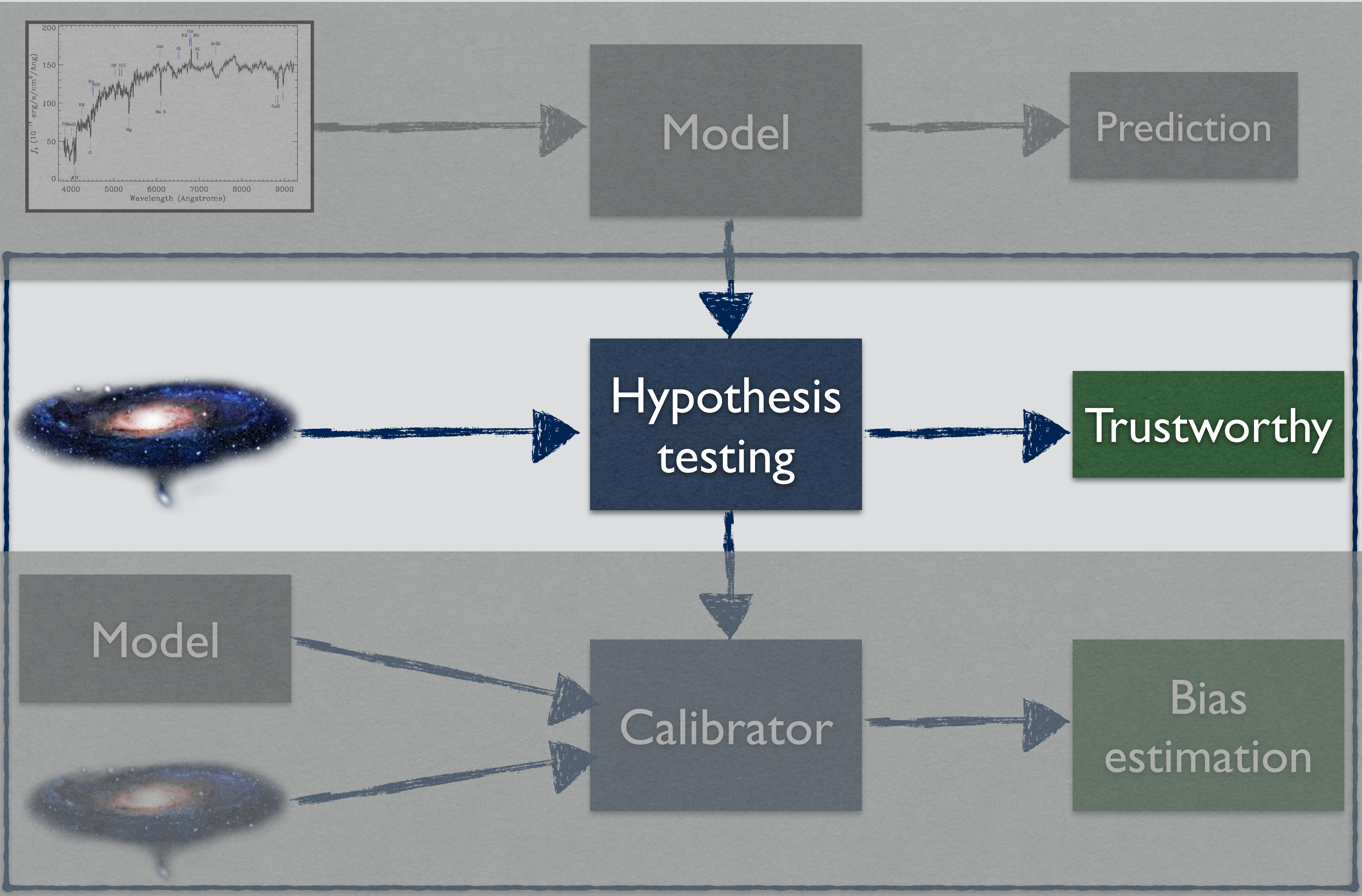
$$\frac{10 \times 95\% + 90 \times 45\,\%}{100} = 50\,\%$$

→ However, the model is not conditionally calibrated.

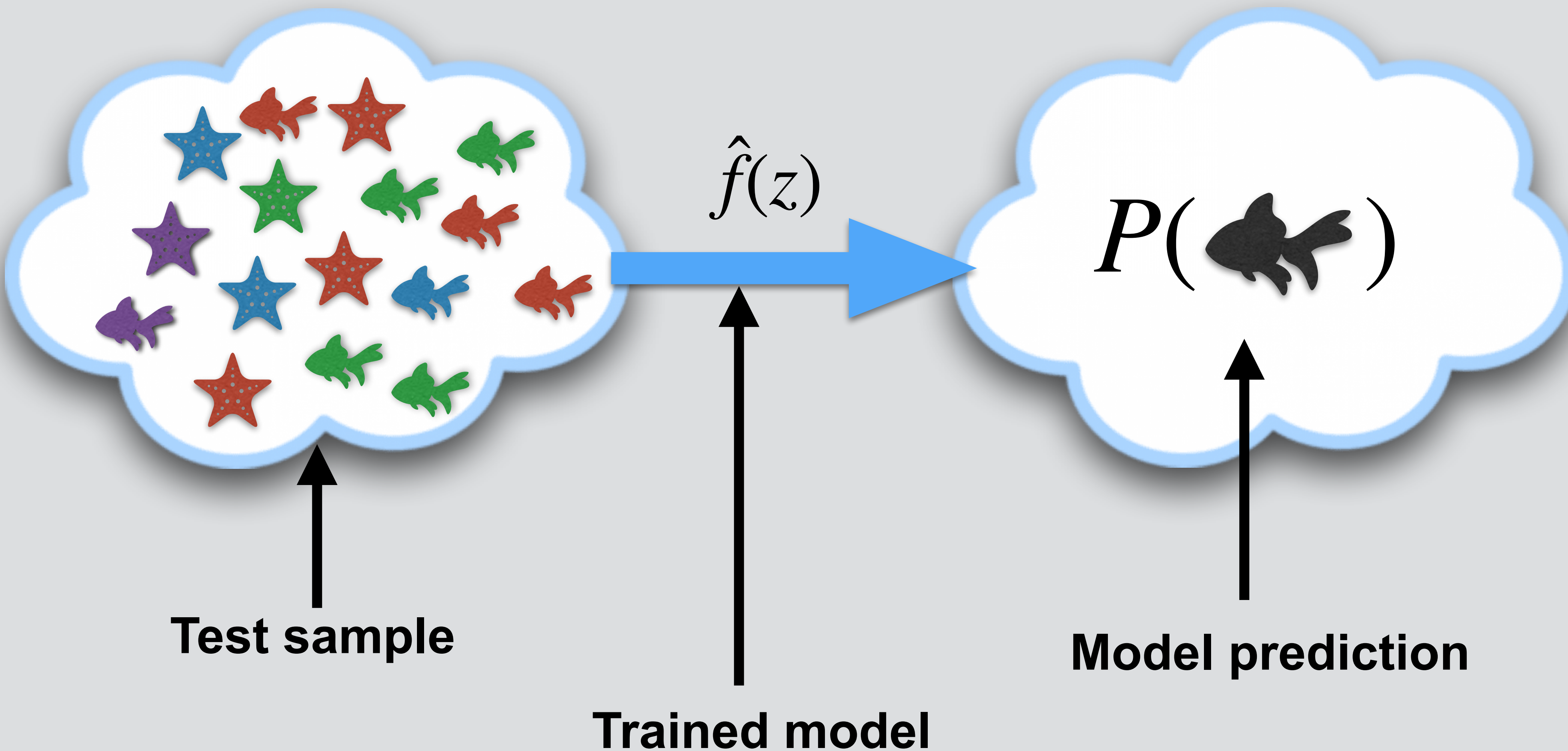
KiTE: open-sourced solution for trustworthiness quantification



KiTE: open-sourced solution for trustworthiness quantification



Problem setup and a test statistic



Definitions

Definition [conditional calibration]. Model $\hat{f}$ is conditionally calibrated if and only if

$$p(y = 1 \mid x, \hat{f}(x) = \alpha) = \alpha \qquad \text{for all } x \in \mathcal{X} \text{ and } \alpha \in [0,1].$$

group feature

model prediction probability

frequency of observation

* Group-wise calibration, local calibration, and conditional calibration are used interchangeably.

Definitions

Definition [conditional calibration]. Model $\hat{f}$ is conditionally calibrated if and only if

$$p(y = 1 \mid x, \hat{f}(x) = \alpha) = \alpha \qquad \text{for all } x \in \mathcal{X} \text{ and } \alpha \in [0,1].$$

group feature

model prediction probability

frequency of observation

Definition [marginal calibration]. Model $\hat{f}$ is marginally calibrated if and only if

$$\int_{x \in \mathcal{X}} p(y = 1 \mid x, \hat{f}(x) = \alpha)\mathrm{d}x = \alpha \qquad \text{for all } \alpha \in [0,1].$$

* Group-wise calibration, local calibration, and conditional calibration are used interchangeably.

** Global and marginal calibration are used interchangeably.

Theoretical Consequences

Definition [conditional calibration]. Model $\hat{f}$ is conditionally calibrated if and only if

$$p(y = 1 \mid x, \hat{f}(x) = \alpha) = \alpha \qquad \text{for all } x \in \mathscr{X} \text{ and } \alpha \in [0,1].$$

- **Uniqueness.** A conditionally calibrated model is equivalent to the true a-posteriori distribution $p(y \mid x)$. [Cohen & Goldszmidt, PKDD, 2004]
- **Optimality.** A conditionally calibrated model is the optimal classifier (minimizes the Bayes error). [Cohen & Goldszmidt, PKDD, 2004]
- **Goodness-of-fit.** A miscalibrated model is not a good fit to data.

Test statistic (Expected Local Calibration Error)

Theorem. Model $\hat{f}$ is conditionally calibrated if and only if

$$\mathrm{ELCE}^2[k, \hat{f}, p] = 0$$

where

$$\mathrm{ELCE}^2[k, \hat{f}, p] := \mathbb{E}\left[(Y - \hat{f}(x))^\top k(\{x, f(x)\}, \{x', f(x')\})(Y' - \hat{f}(x'))\right].$$

rate of error

Our proposed test statistic

A kernel function

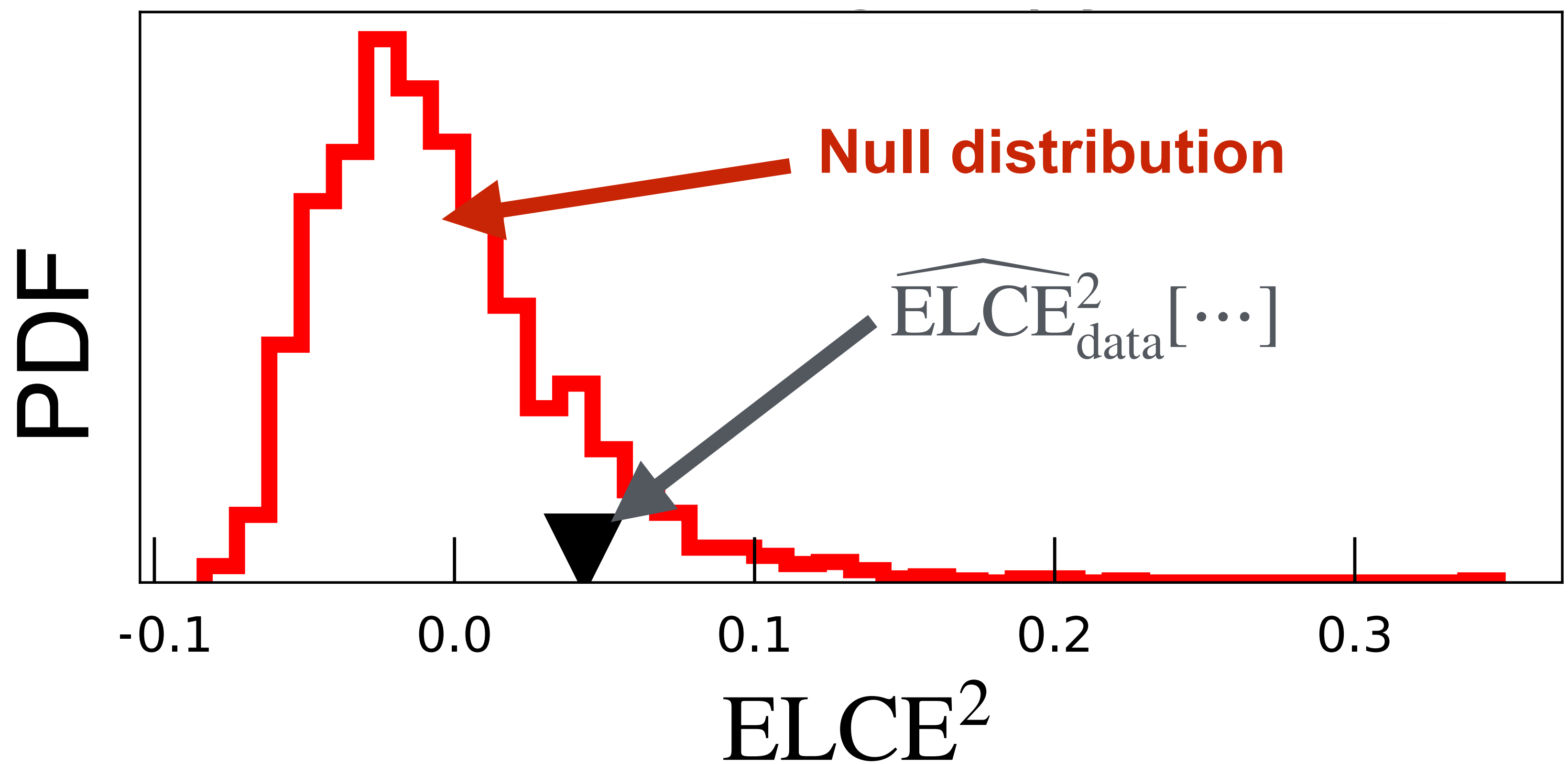
Null Hypothesis: Model $\hat{f}(z)$ is conditionally calibrated on x.

Farahi&Koutra (under review), motivated by Gretton et al. (JMLR, 2012)

Hypothesis testing in a finite sample setting

p-value is the probability that the observed $\widehat{\mathrm{ELCE}}$ is larger than the null distribution.

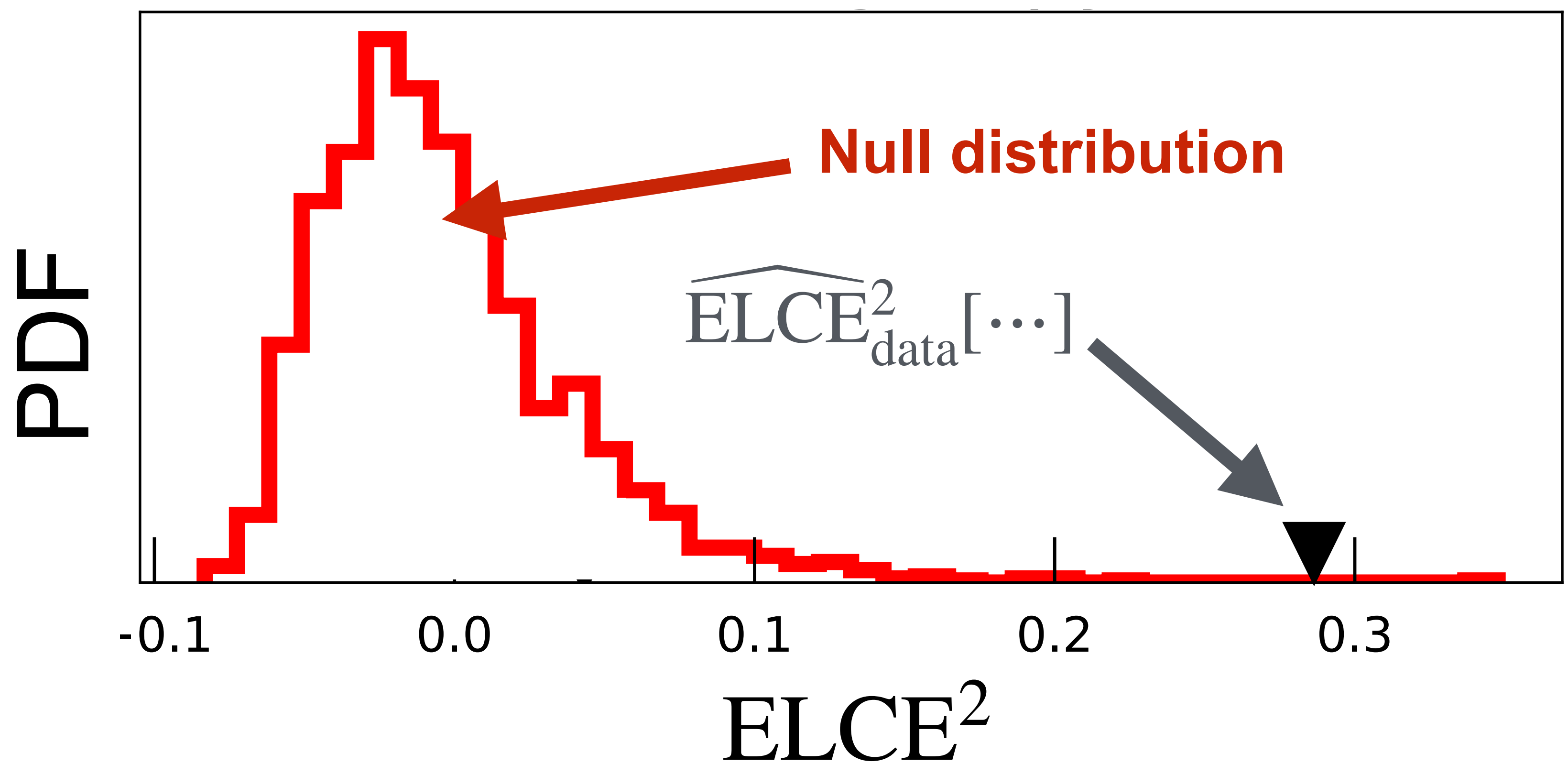
$$1 - p = \Pr(\widehat{\mathrm{ELCE}}^2_{\mathrm{null}}[\cdots] < \widehat{\mathrm{ELCE}}^2_{\mathrm{data}}[\cdots])\,.$$



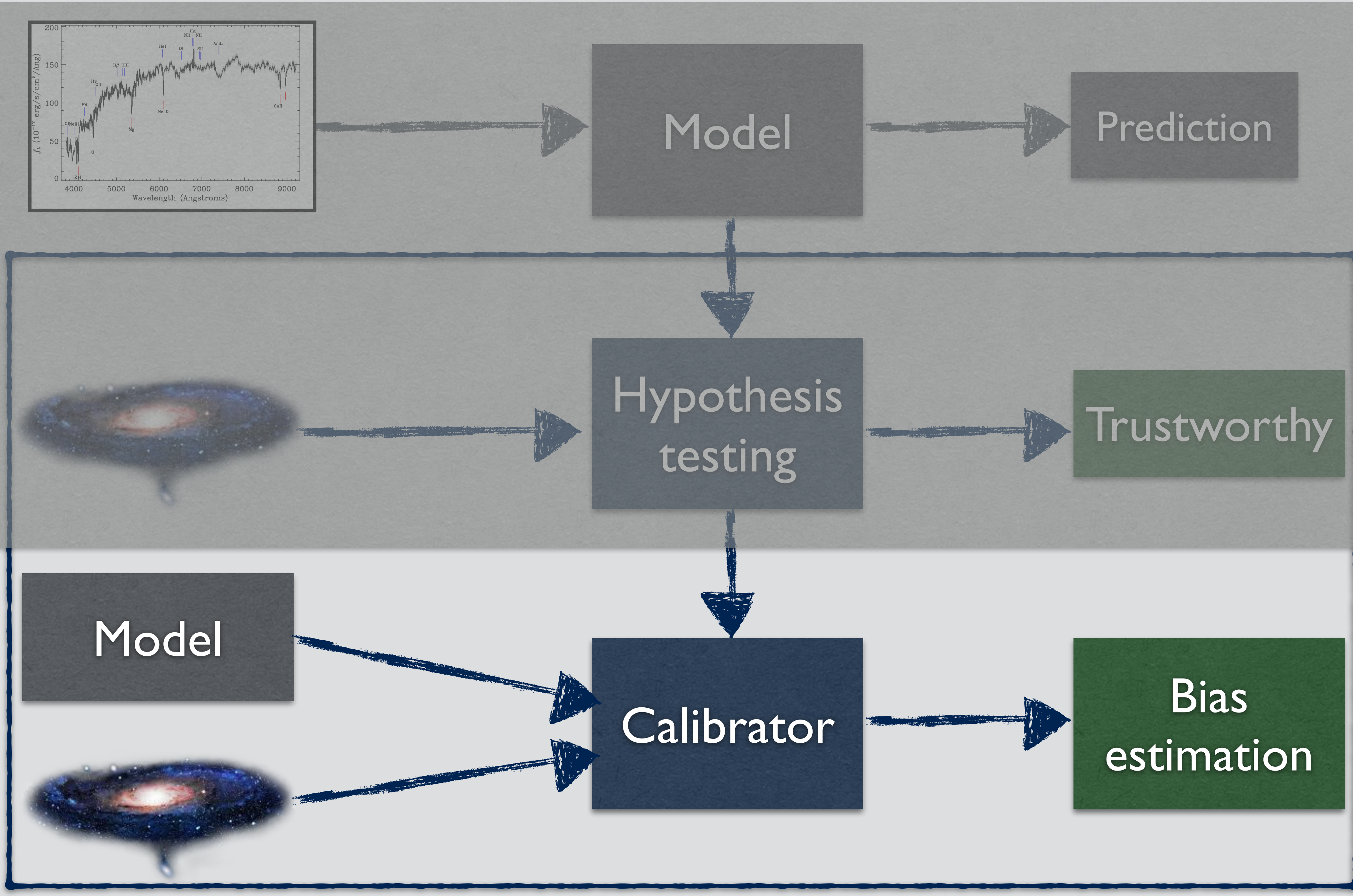
Hypothesis testing in a finite sample setting

p-value is the probability that the observed $\widehat{\mathrm{ELCE}}$ is larger than the null distribution.

$$1 - p = \Pr(\widehat{\mathrm{ELCE}}^2_{\mathrm{null}}[\cdots] < \widehat{\mathrm{ELCE}}^2_{\mathrm{data}}[\cdots])\,.$$



KiTE: open-sourced solution for trustworthiness quantification



A locally-aware calibration method

Our model is:

$$\hat{f}_c(x) = \hat{f}(x) + b(x)$$

additive bias

Trained model

Calibrated model

Our goal is to estimate additive bias by exploiting information provided in the calibration sample.

An estimator of calibration bias

Suppose $a = [(y_1 - \hat{f}_1), \cdots, (y_n - \hat{f}_n)]$,

$\kappa(x) = [k(x_1, x), \cdots, k(x_n, x)]$, and $K_{ij} = k(x_i, x_j)$.

where n is the calibration sample size and x is a new data point.

Now we can estimate individual level and group level bias:

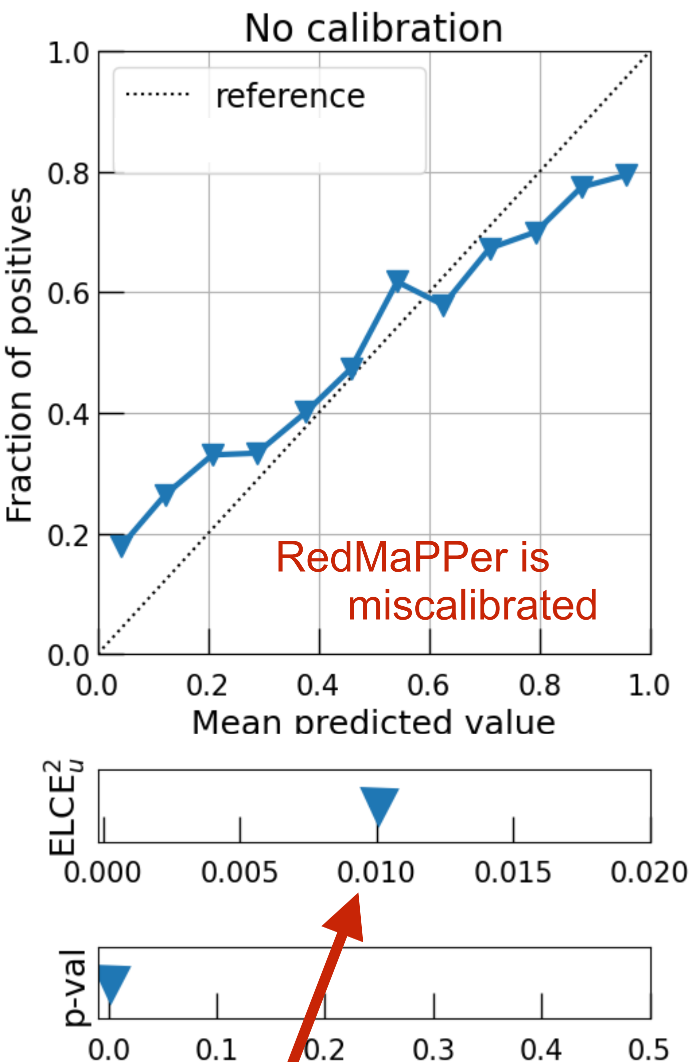
individual level — $\widehat{b}(x) = a(K + \lambda \mathbb{I})^{-1}\kappa(x)$

Farahi, Esteves, Koutra (under review)

In collaboration with Danai Koutra (EECS, UoM)


Credit: NASA/JPL-Caltech/GSFC/SDSS

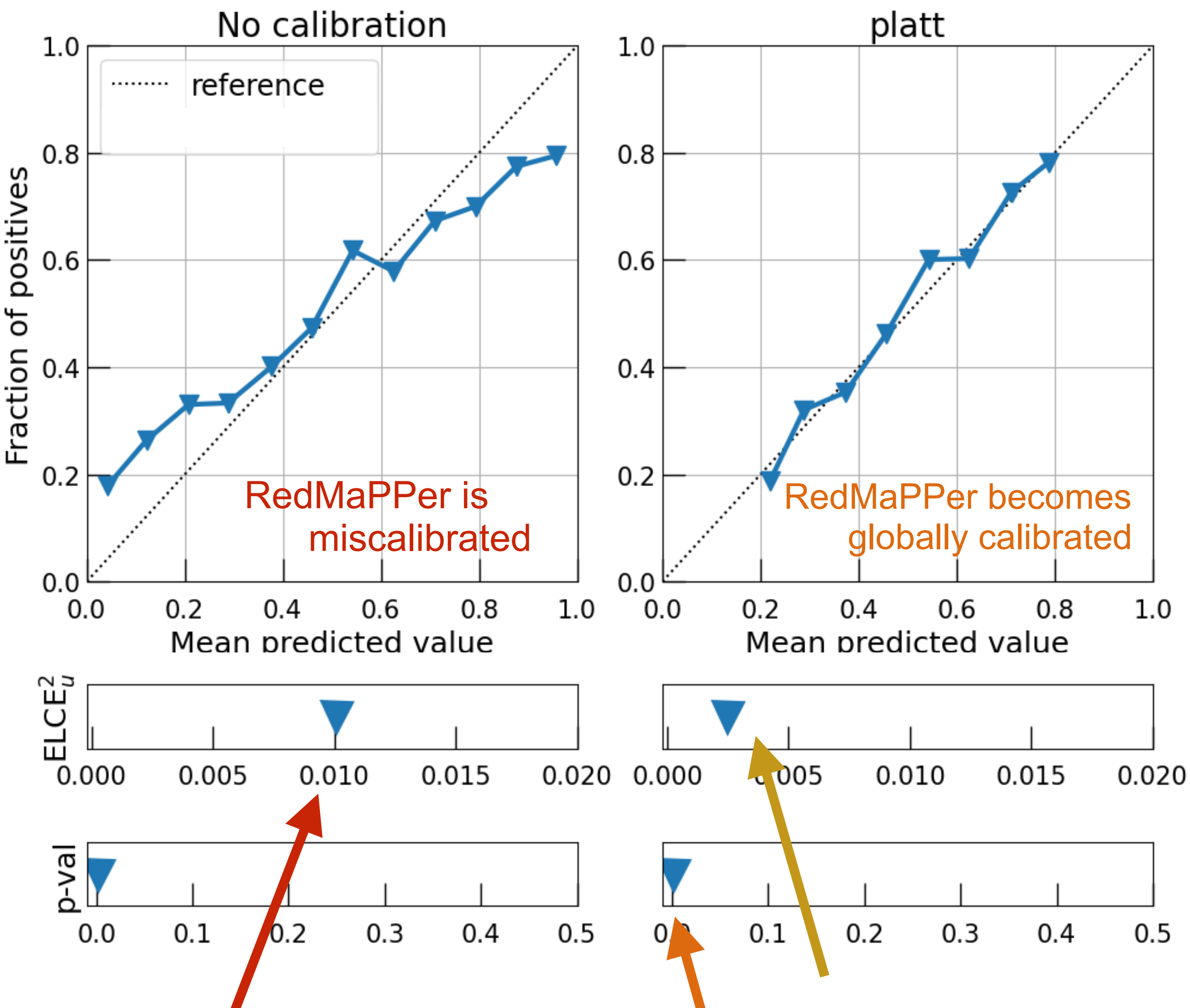
Attempts to Calibrate Cluster Finding Algorithms



redMapper is
miscalibrated

Johnny Esteves
(Physics, U-Michigan)

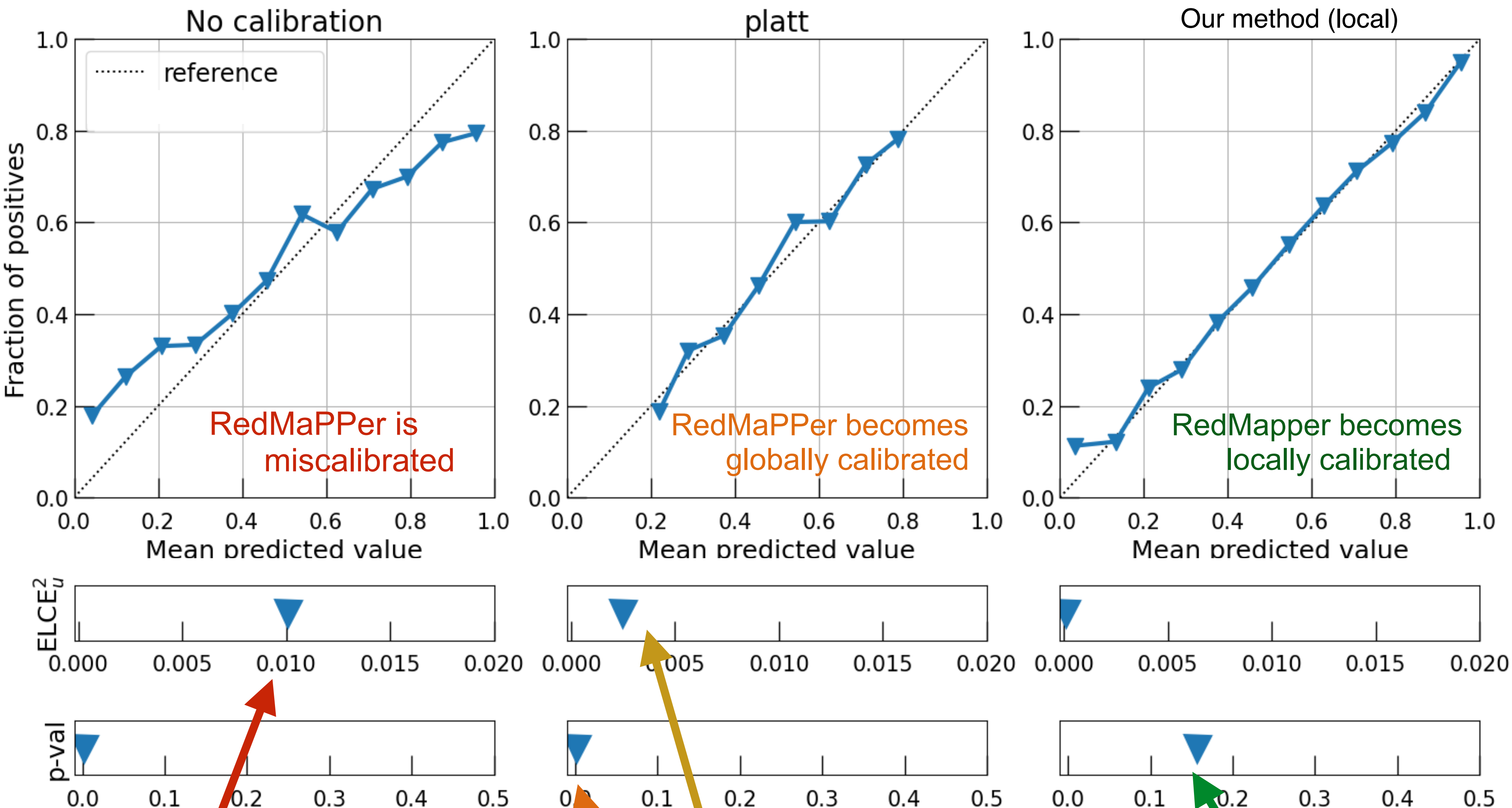
Attempts to Calibrate Cluster Finding Algorithms



Even after Platt's scaling correction
redMaPPer remains miscalibrated
but less severe.

redMapper is
miscalibrated

Attempts to Calibrate Cluster Finding Algorithms



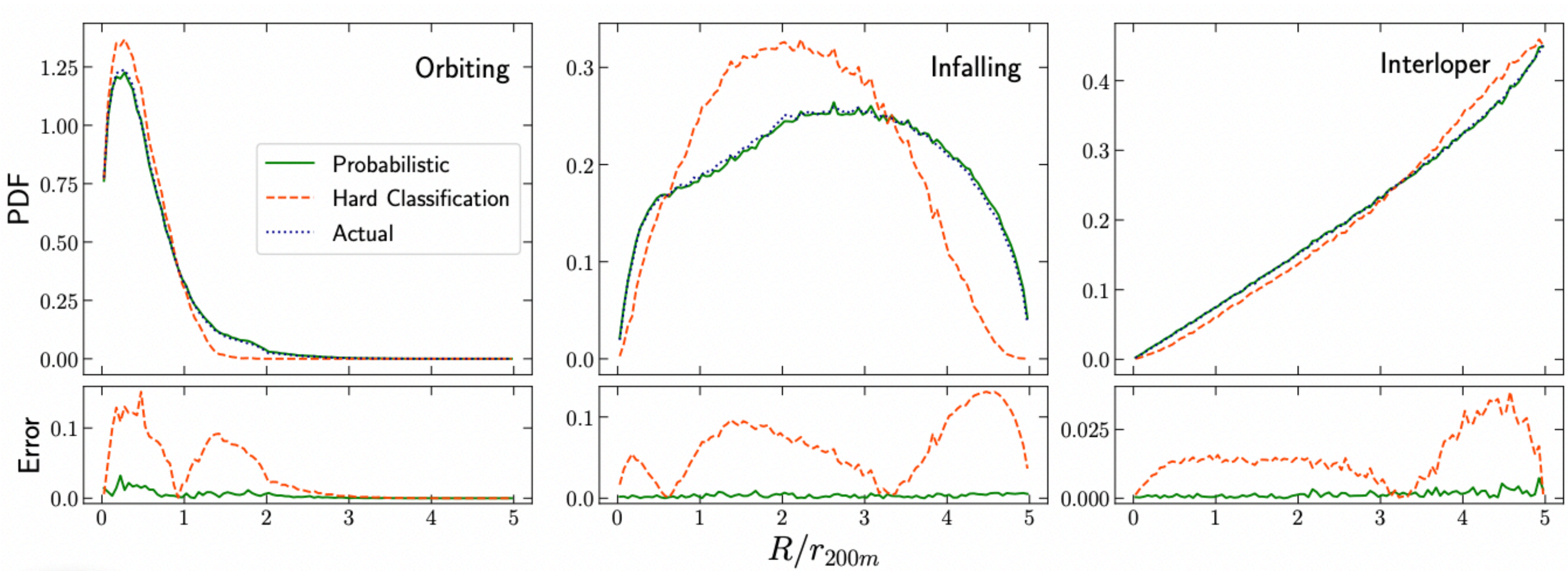
redMapper is
miscalibrated

Even after Platt's scaling correction
redMaPPer remains miscalibrated
but less severe.

RedMapper becomes calibrated!

Multi-classification tasks

Classifying galaxies into orbiting, infall and interloper.

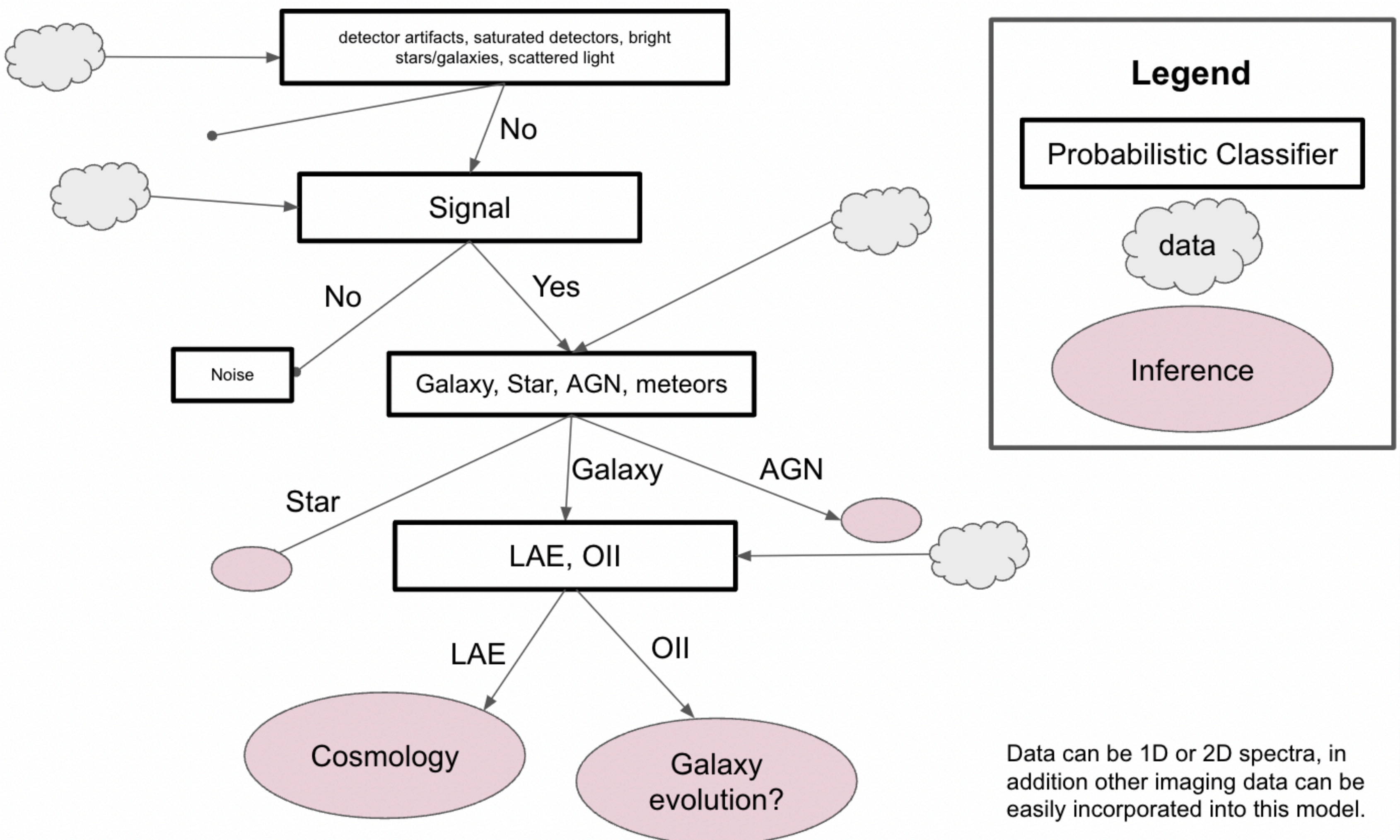


Danny Farid
(Math, Undergraduate at Yale)

Farid, et al. arXiv:2205.01700

HETDEX Illuminating the Darkness
Hobby-Eberly Telescope Dark Energy Experiment

Hierarchical classification tasks



Data can be 1D or 2D spectra, in addition other imaging data can be easily incorporated into this model.

Conclusion

Task → Data → Modeling → Prediction → Make Decision

The are lessons to be learned from other communities. We do not need to reinvent the wheel.

(e.g., ML interpretability, bias quantification, uncertainty modeling)

Establishing **trustworthiness** of prediction models utilized in decision-making and inference pipelines is a necessary step to achieve unbiased inference and optimal decision-making.

KiTE is a tool for trustworthiness quantification and calibration of probabilistic classifiers.

https://afarahi.github.io

Statistics and Data Science (SDS)
arya.farahi@austin.utexas.edu

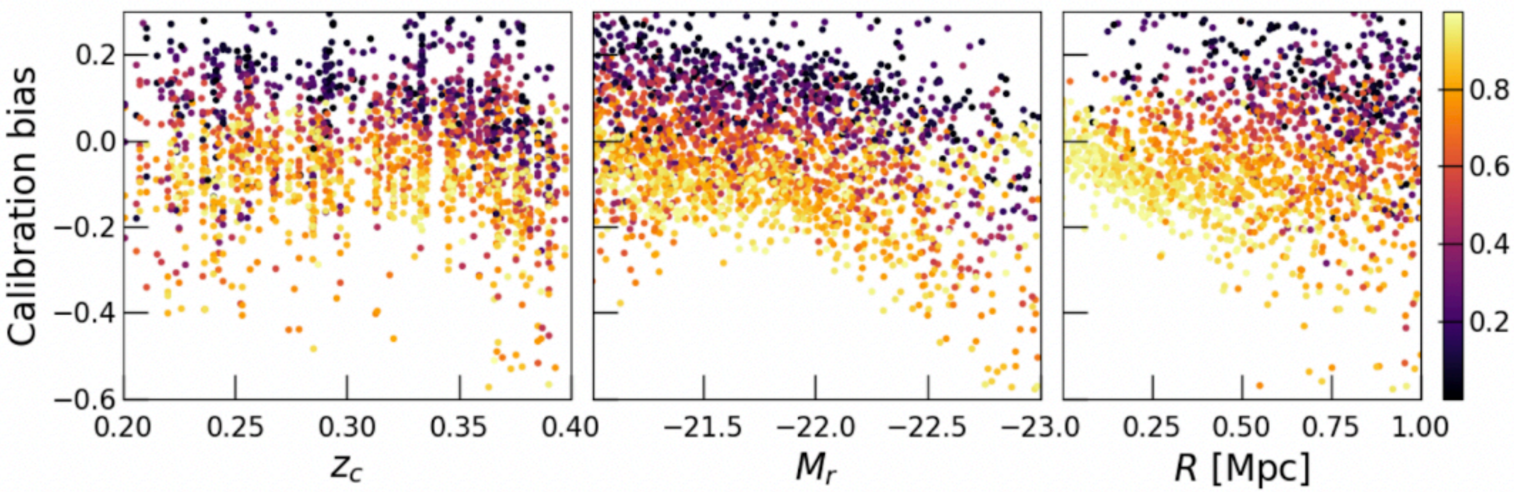
$$\Pr(\text{member}) = \frac{n_{\text{cls}} p(R_i, z_i, \mathbf{c})}{n_{\text{cls}} p(R_i, z_i, \mathbf{c}_i) + n_{\text{bkg}} p_{\text{bkg}}(z_i, \mathbf{c}_i)}$$

$$p(R, z, \mathbf{c}) = p(\mathbf{c} \mid z, R) p(z \mid R) p(R)$$

Galaxy red-sequence model

Photo-z model

NFW model



Hypothesis testing in a finite sample setting

Corollary 1 [Convergence Bound]. Suppose $0 \leq k(\,.\,,\,.\,) \leq K$ then the estimator is bounded under the null hypothesis. The bound is

$$\Pr\left(\widehat{\mathrm{ELCE}}^2_u(k, \{x, y, z\}, \hat{f}) > \epsilon \mid H_0\right) < \exp\left(-\frac{\epsilon^2 n}{8K^2}\right).$$

Corollary 2 [Convergence Rate]. A hypothesis test of level α_p for the null hypothesis has the acceptance region

$$\widehat{\mathrm{ELCE}}^2_u(k, \{x, y, z\}, \hat{f}) < \frac{\sqrt{8}K}{\sqrt{n}}\sqrt{\alpha_p^{-1}}$$

thus, the estimator has a convergence rate of $n^{-\frac{1}{2}}$.

Theoretical Consequences

Definition [group-wise (local) calibration]. Model $\hat{f}$ is locally calibrated if and only if

$p(y = 1 \mid x, \hat{f}(z) = \alpha) = \alpha$ for all $x \in \mathscr{X}$ and $\alpha \in [0,1]$.

A feature set that on which an auditor wants to evaluate the trustworthiness of forecaster.
{age, gender, race, income-level}

Input of the model.
{age, genomic expression, gender}

Model $\hat{f}(z)$.
Predicts the likelihood of developing cancer.

There can be overlap between x and z.

Simulated experiment

Generative model:

$$x_1 \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

$$x_2 \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

$$y \sim \text{Bernoulli}(\bar{p} = \text{sigmoid}(x_1 + x_2))$$

Classifiers:

$$f = \text{sigmoid}(x_1 + x_2)$$

$$\hat{f}_1 = \text{sigmoid}(x_1)$$

$$\hat{f}_2 = \text{sigmoid}(0.5 + 1.3x_1)$$

$$\widehat{\text{ELCE}}^2(\hat{f}_1) < \widehat{\text{ELCE}}^2(\hat{f}_2)$$

thus, model $\hat{f}_1$ is closer to the true model.

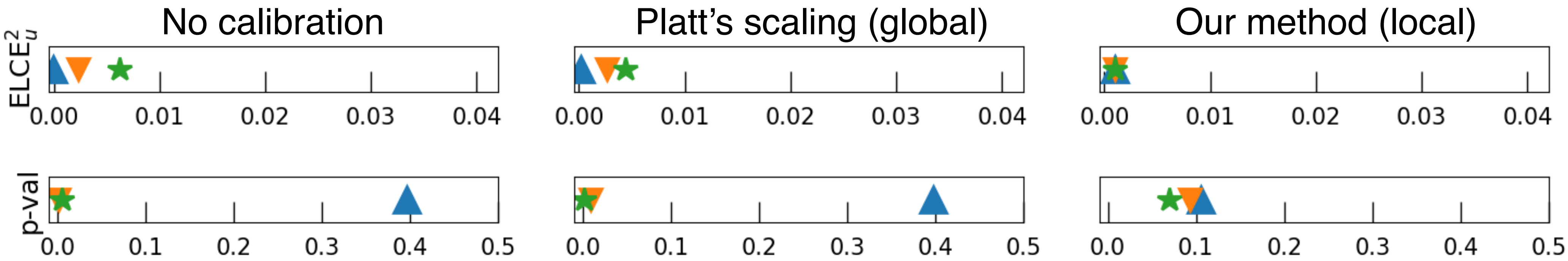
Achieving local calibration

Classifiers:

$f = \text{sigmoid}(x_1 + x_2)$ [generative model — Bayes Classifier]

$\hat{f}_1 = \text{sigmoid}(x_1)$ [conditionally miscalibrated]

$\hat{f}_2 = \text{sigmoid}(0.5 + 1.3x_1)$ [marginally miscalibrated]



Platt et al. (ALMC, 1999)

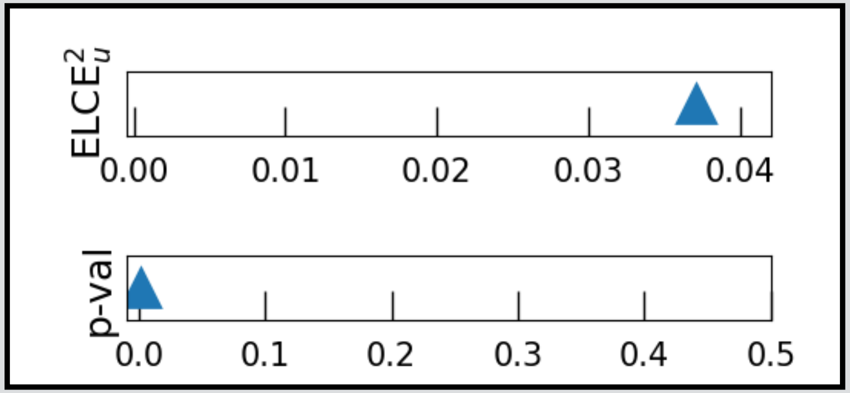
Auditing predictive models with KiTE

1. COMPAS recidivism data set

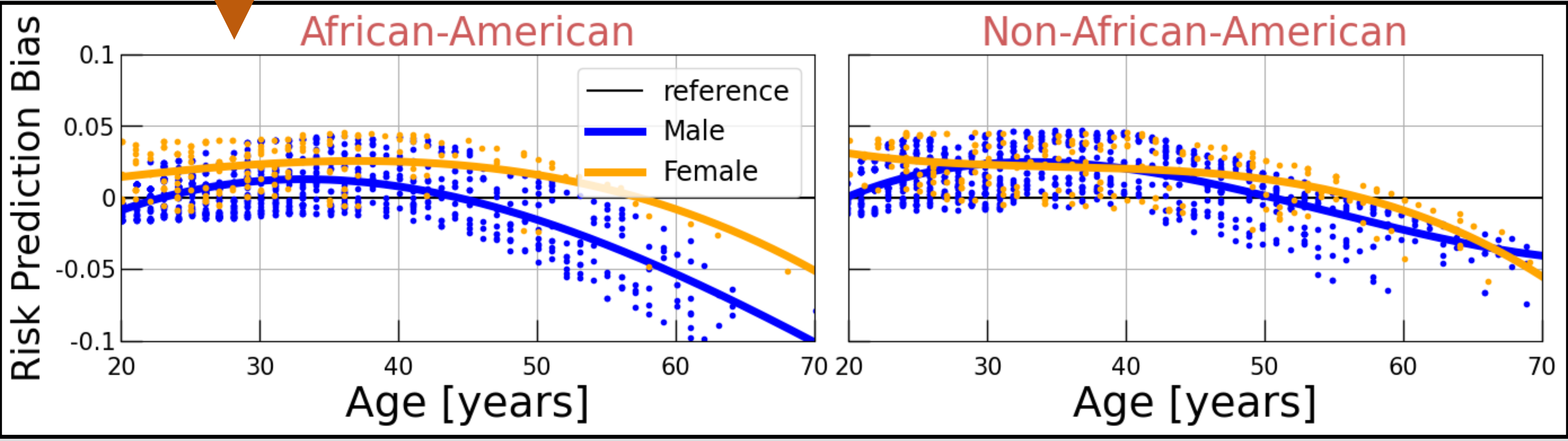
2. Train a Random Forest classifier

3. Perform hypothesis testing

    x = {race, age, gender}



4. Estimate calibration bias



An estimator of calibration bias

Suppose $a = [(y_1 - \hat{f}_1), \cdots, (y_n - \hat{f}_n)]$,

$\kappa(x) = [k(x_1, x), \cdots, k(x_n, x)]$, and $K_{ij} = k(x_i, x_j)$.

where n is the calibration sample size and x is a new data point.

Now we can estimate individual level and group level bias:

individual level — $\widehat{b}(x) = a(K + \lambda \mathbb{I})^{-1}\kappa(x)$

group level — $\widehat{b}(x) = \int_{x \in X \subset \mathcal{X}} a(K + \lambda \mathbb{I})^{-1}\kappa(x)\, \mathrm{d}x$

In collaboration with Danai Koutra (EECS, UoM)

Literature Review and Challenges

→ A key goal of calibration is to ensure the information provided by a model is trustworthy. e.g., Miller (1962); Murphy (1972;1973); Gneiting & Raftery (2005).

→ Calibration problem is known as one of the pillars of algorithmic fairness. e.g., Pleiss, Raghavan, et al., (NeurIPS, 2017), Kleinberg, et al., (ITCS, 2017).

→ **Challenge 1.** Hypothesis testing is a missing key. Vaicenavicius, et al., (AISTATS, 2019).

→ **Challenge 2.** Quantifying group-wise prediction bias is challenging, particularly in a high dimensional setting. e.g., Zhang, et al., (KDD, 2017), Hebert-Johnson et al. (ICML, 2018).

6 Conclusion **Vaicenavicius et al. (AISTATS, 2019)**

Evaluation of model calibration is about checking whether probabilities predicted by a model match the distribution of realized outcomes. In this article, we built on existing calibration evaluation approaches and proposed a general mathematical framework for evaluating model calibration, or a chosen aspect of it, in classification problems. We showed that empirical estimates of intuitive miscalibration measures should not be used in a naive way to compare probabilistic classifiers but instead can be employed in hypothesis tests for testing model reliability. We hope our developments and attempts in rigorous model calibration evaluation will encourage other researchers to study this essential topic further.

Our contribution

→ **Contribution 1.** Hypothesis testing.

Testing whether a model is group-wise calibrated, as oppose to be population level calibrated.

(e.g., Widmann et al. (NeurIPS, 2019)).

→ **Contribution 2.** Group-wise calibration.

Perform group-wise calibration as oppose to population level calibration.

(e.g., Chakravarti, (MOR,1989), Platt et al. (ALMC, 1999), Zadrozny & Elkan (ICML, 2001), Zadrozny & Elkan (KDD, 2002), Naeini et al., (AAAI, 2015), Guo et al., (JMLR, 2017)).

Theoretical Consequences

Definition [group-wise (local) calibration]. Model $\hat{f}$ is locally calibrated if and only if

$p(y=1 \mid x, \hat{f}(z)=\alpha)=\alpha$ for all $x \in \mathscr{X}$ and $\alpha \in [0,1]$.

if $x=z$, then

- **Uniqueness.** locally calibration model equivalent to the true a-posteriori distribution $p(y \mid x)$. [Cohen & Goldszmidt, PKDD, 2004]
- **Optimality.** A locally calibrated model is the optimal classifier (minimizes the Bayes error). [Cohen & Goldszmidt, PKDD, 2004]
- **Covariate invariant.** A locally calibrated model remains locally calibrated if the covariate's distribution changes $p(x) \to q(x)$.

Test statistic (Expected Local Calibration Error)

Theorem. Model $\hat{f}$ is locally calibrated if and only if $\mathrm{ELCE}^2[k, \hat{f}] = 0$ where

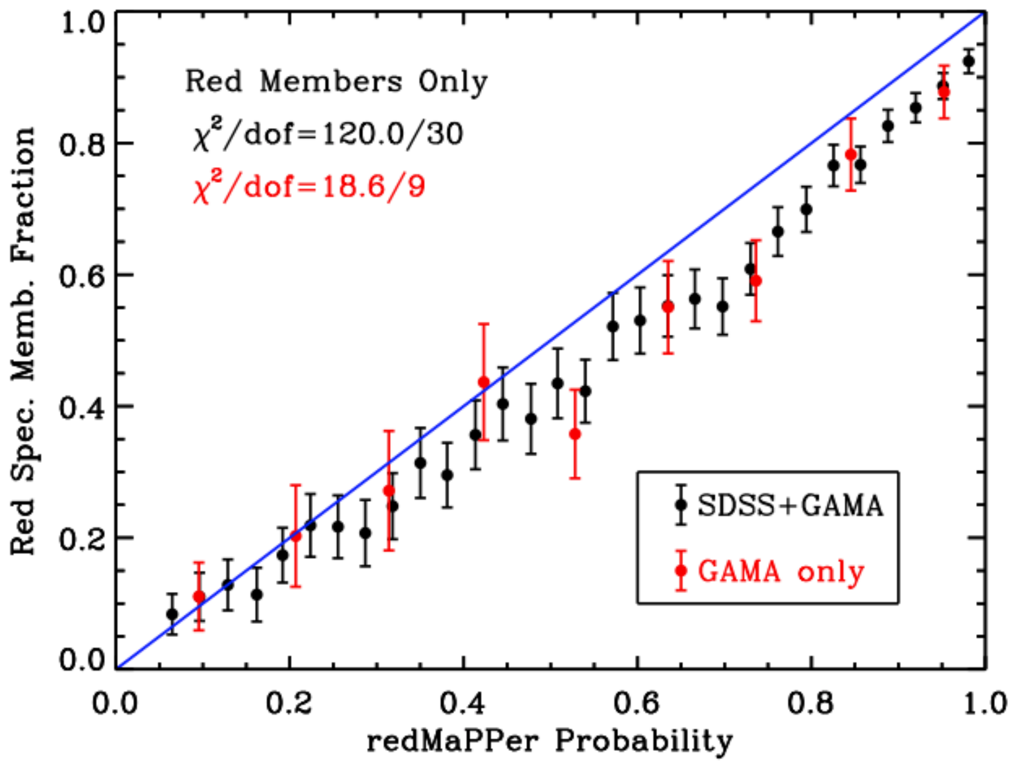
$$\mathrm{ELCE}^2[k, \hat{f}] := \mathbb{E}\left[(Y - \hat{f}(x))^\top k(x, x')(Y' - \hat{f}(x'))\right].$$

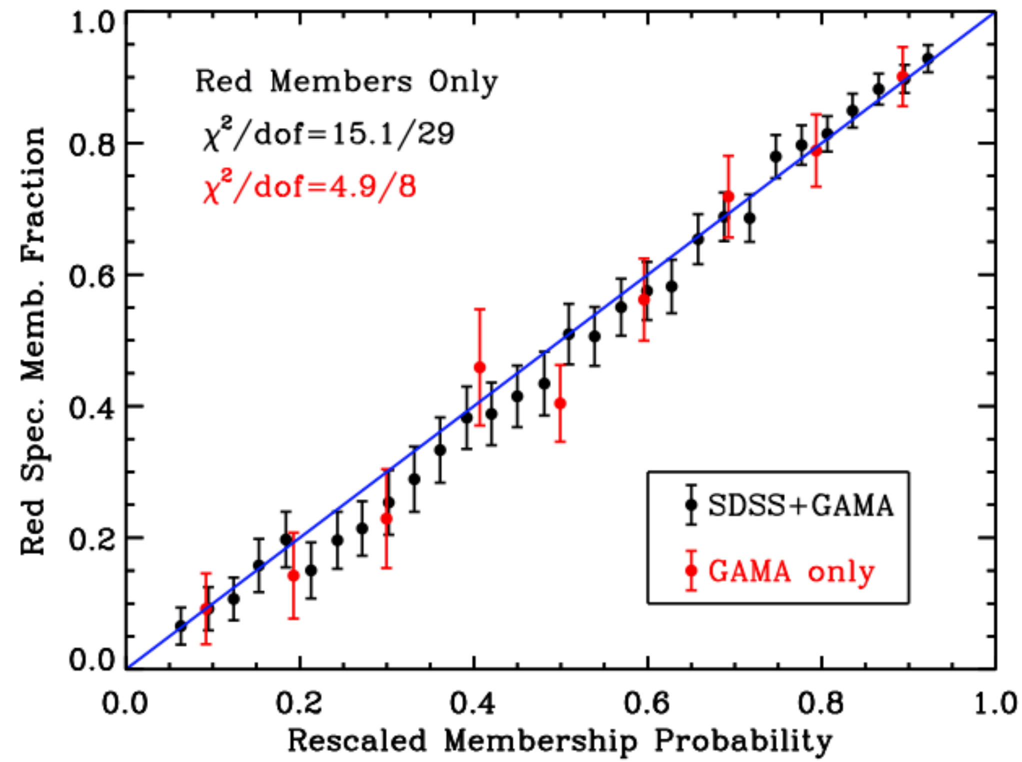
Corollary: ELCE test statistic is a metric.

Thus, it may be employed in performing model comparison and model selection.

Since ELCE quantifies the prediction of which model is closer to the actual class probability. A model with smaller ELCE can be considered as a less unfair model.

Attempts to Calibrate Cluster Finding Algorithms





Rozo et al., (MNRAS, 2015)
Farahi et al., (MNRAS, 2016)

In collaboration with
August Evrard
(Physics, U. Michigan)

Eduardo Rozo
(Physics, U. Arizona)

Eli Rykoff
(Physics, Stanford)

A model is miscalibrated, now what?

→ **Challenge:** ML models are often miscalibrated. Thus, we need to develop a method to calibrate an untrustworthy classifier.

→ **Literature:** Proposed calibration methods are generally concerned about global calibration

(e.g., Chakravarti, (MOR,1989), Platt et al. (ALMC, 1999), Zadrozny & Elkan (ICML, 2001), Zadrozny & Elkan (KDD, 2002), Naeini et al., (AAAI, 2015), Guo et al., (JMLR, 2017)).

→ **Our contribution:** A method of local calibration.