Experimental design

Alex Malz
LINCC@CMU

DSFP Session 16

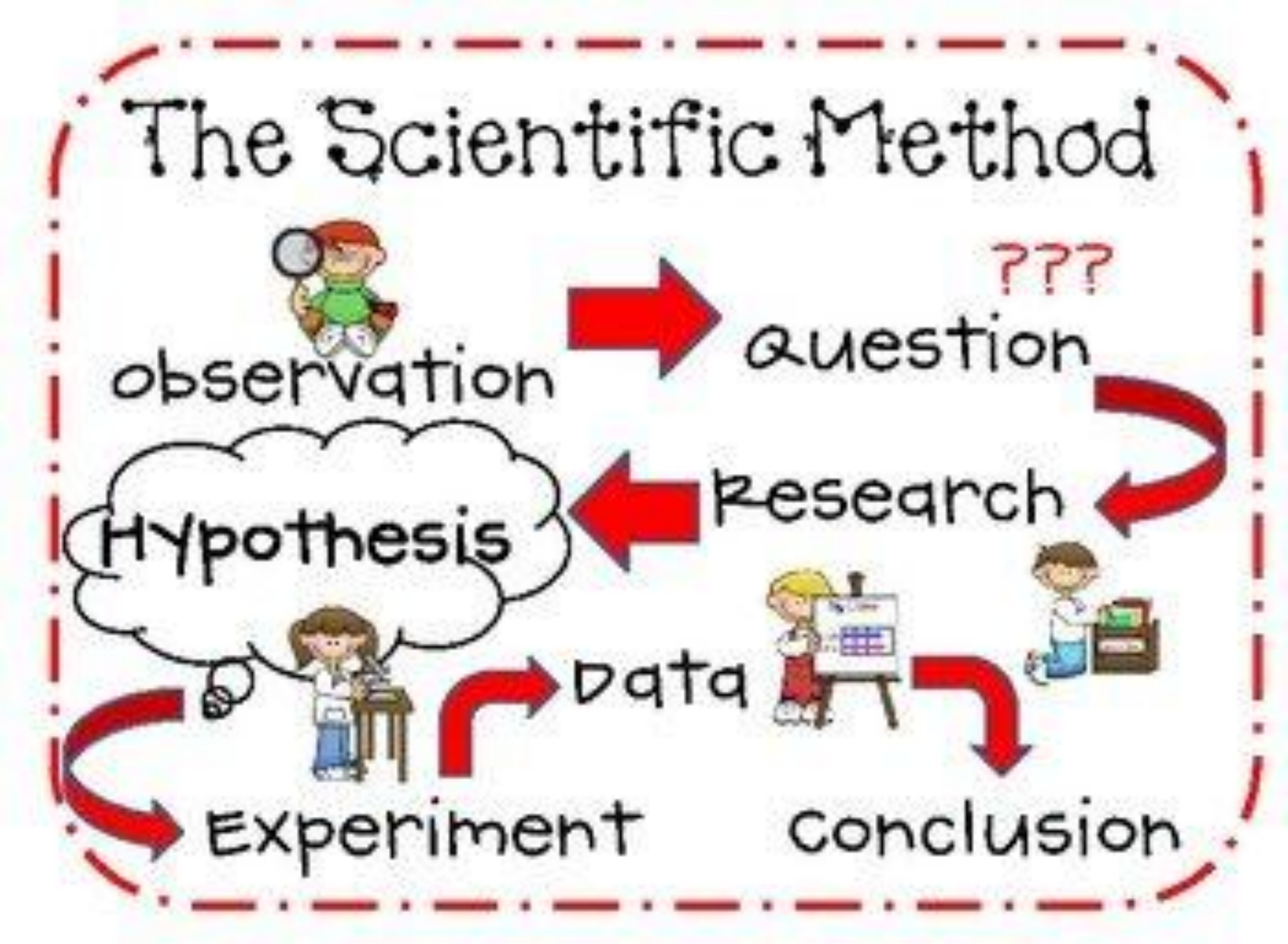
Overview

The purpose of this lecture is to introduce a sort of metascience:
How do we apply the scientific method to how we do science?

Probabilistic graphical models are pretty much my favorite thing!
But my secret agenda is for you to incorporate this way of thinking into your hacks, for the experiments in which you prove how great your hierarchical models are.

What does “experimental design” actually mean?

LSSTC DATA SCIENCE FELLOWSHIP PROGRAM



credit:N.Cunningham

How does observational astronomy compare with a physics lab experiment?

Lab physics vs. Observational astro

How does observational astronomy compare with a physics lab experiment?

Lab physics vs.

- Assume physical model *M* (and other assumptions *I*)
- Seek to constrain physical parameters θ
- Know controlled experimental conditions ϕ
- Make many independent observations $\{x_i\}$
- …

Estimate p(x | ϕ)

likelihood

Observational astro

- Assume physical model *M* (and other assumptions *I*)
- Seek to constrain physical parameters θ
- Learn unknown initial conditions ϕ
- Gather data *x* from only one observable universe
- …

Estimate p(ϕ | x)

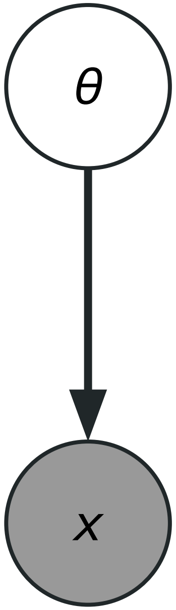
posterior

Conditional probability & forward models

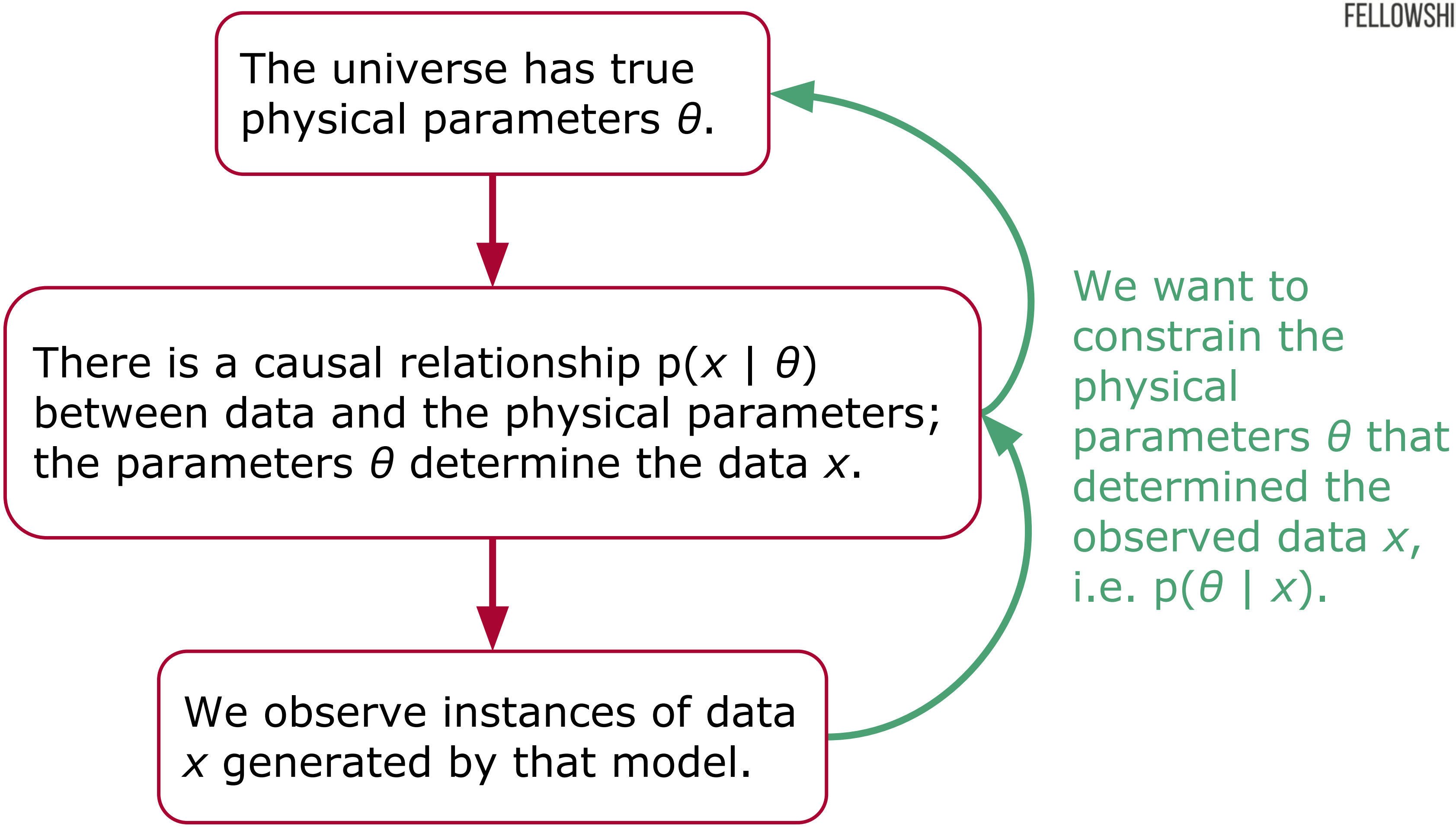
The universe has true physical parameters θ.

There is a causal relationship p(x | θ) between data and the physical parameters; the parameters θ determine the data x.

We observe instances of data x generated by that model.



Conditional probability & forward models



The universe has true physical parameters θ.

There is a causal relationship $p(x \mid \theta)$ between data and the physical parameters; the parameters θ determine the data x.

We observe instances of data x generated by that model.

We want to constrain the physical parameters θ that determined the observed data x, i.e. $p(\theta \mid x)$.

Review

Let’s be critical of what we saw ereyesterday!

What do these metrics miss?

Intrinsic scatter

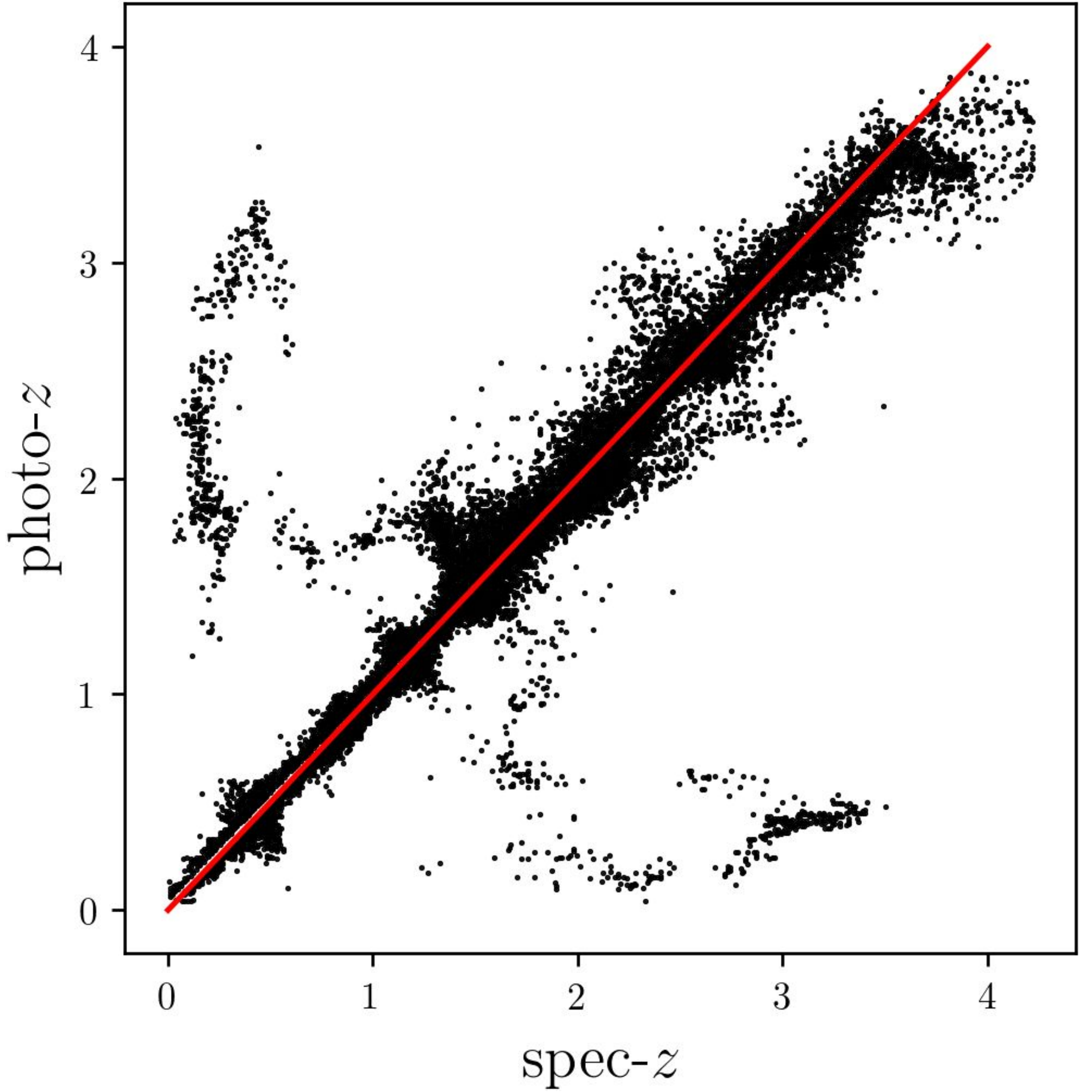
$$\sigma_z \ < \ 0.02(1+z)$$

Bias

$$\langle|z-\hat{z}|\rangle \ < \ 0.003(1+z)$$

Catastrophic outlier rate

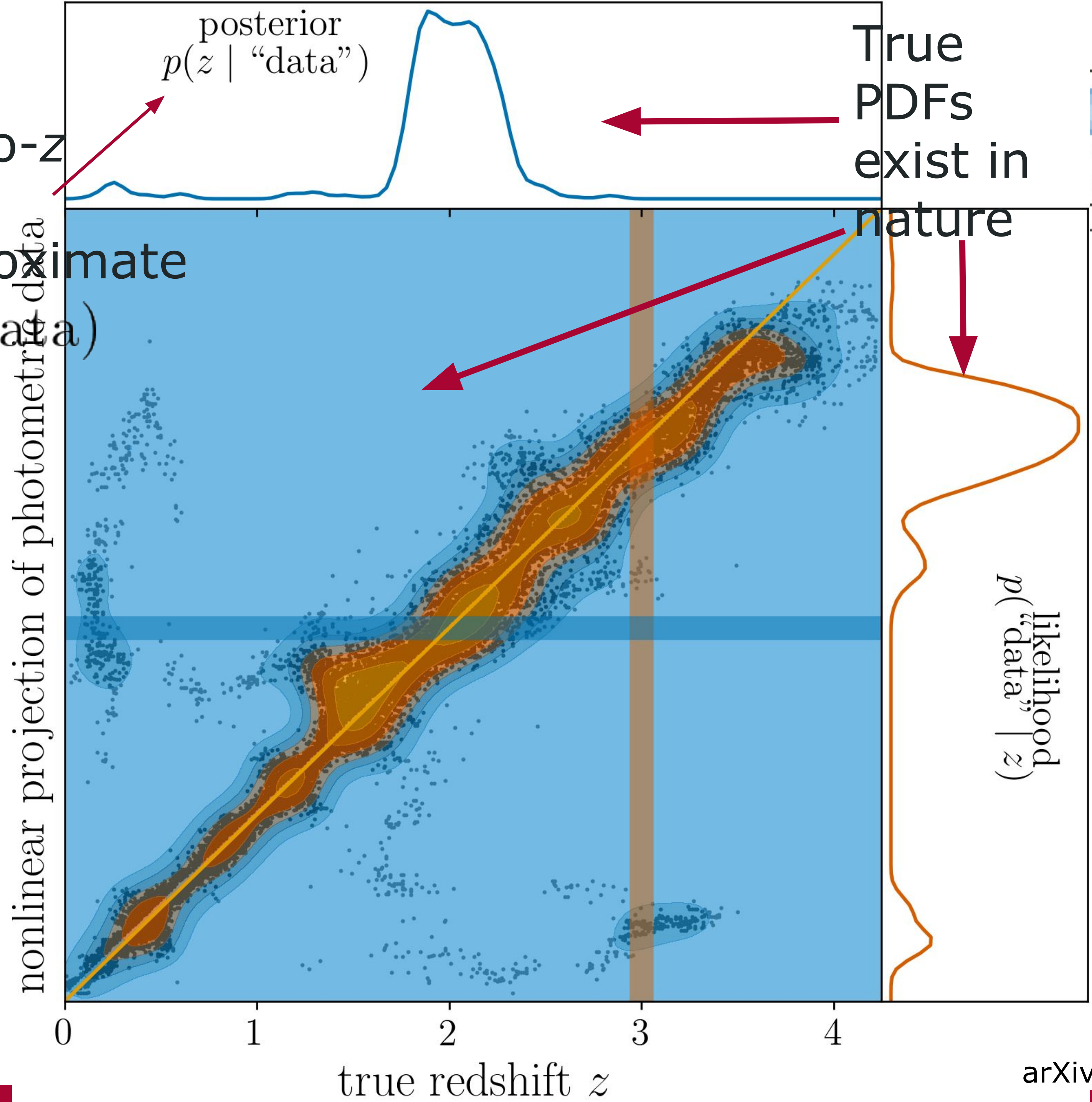
$$\mathrm{N}_{|z-\hat{z}|>3\sigma_z} \ < \ 0.1\mathrm{N}_{\mathrm{LSST}}$$



arXiv:1501.07897

Photo-z PDFs approximate $\hat{p}(z|\text{data})$

True PDFs exist in nature



arXiv:2007.12178

The LSST-DESC PZ DC1 experiment

Motivation: identify the best photo-z posterior code for LSST-DESC

Data: cosmological redshifts & photometry catalog painted on N-body simulation

Control: idealized, shared prior information

The LSST-DESC PZ DC1 experiment

Motivation: identify the best photo-*z* posterior code for LSST-DESC

Data: cosmological redshifts & photometry catalog painted on N-body simulation

Control: idealized, shared prior information

Quantitative metrics of 1D PDF ensembles

Root-mean-square Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\int (p_{\mathrm{true}}(z) - \hat{p}_{\mathrm{est}}(z))^2 dz}$$

Kullback-Leibler Divergence (KLD)

$$\mathrm{KLD}[\hat{p}_{\mathrm{est}}(z); p_{\mathrm{true}}(z)] = \int_{-\infty}^{\infty} p_{\mathrm{true}}(z) \log\left[\frac{p_{\mathrm{true}}(z)}{\hat{p}_{\mathrm{est}}(z)}\right] dz$$

No true posteriors available!

Cumulative Distribution Function (CDF)

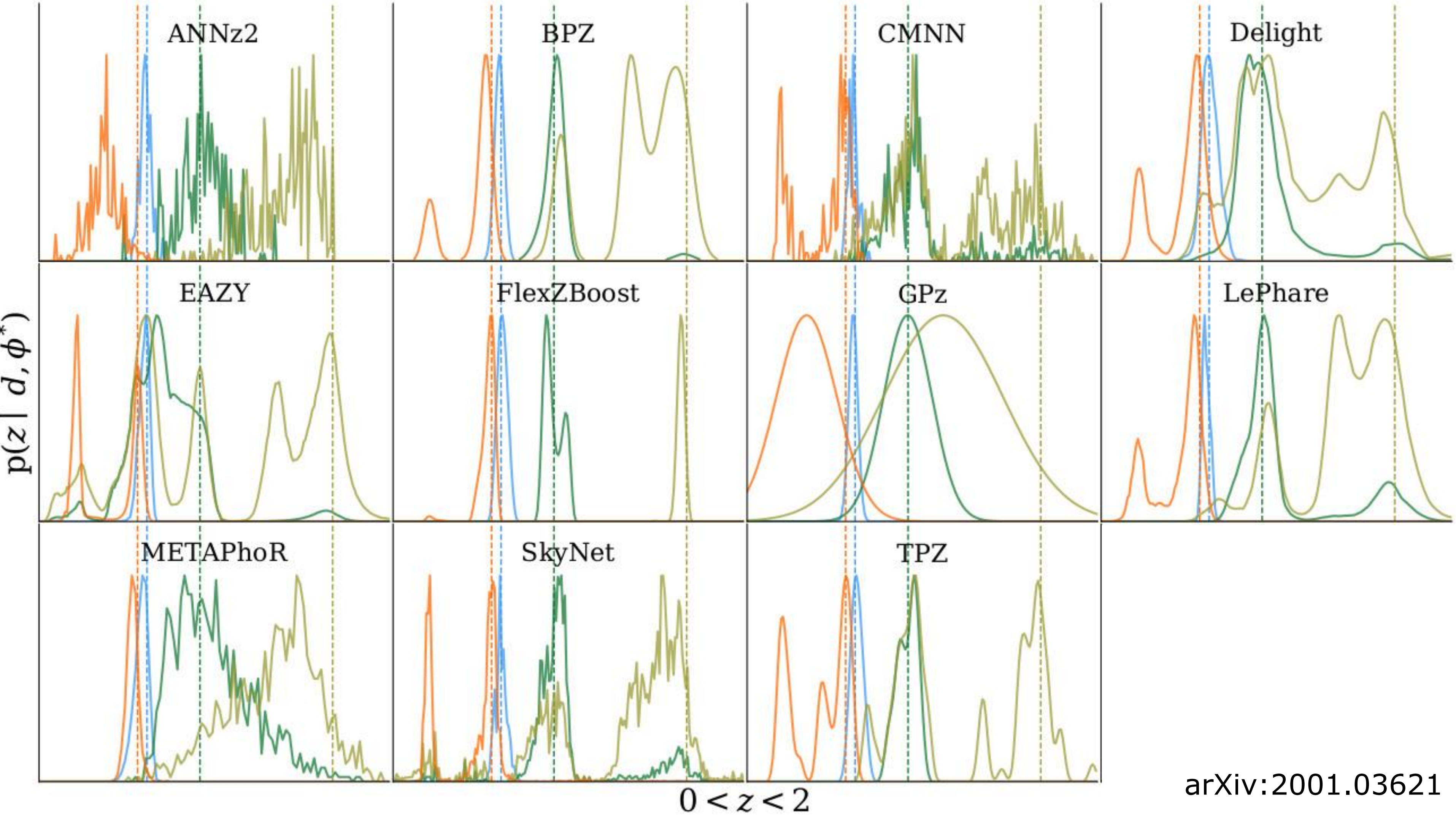
$$\mathrm{CDF}[\hat{p}, z'] \equiv \int_{-\infty}^{z'} \hat{p}(z) dz$$

Probability Integral Transform (PIT)

$$P(\mathrm{PIT} \equiv \mathrm{CDF}[\hat{p}, z_{true}])$$

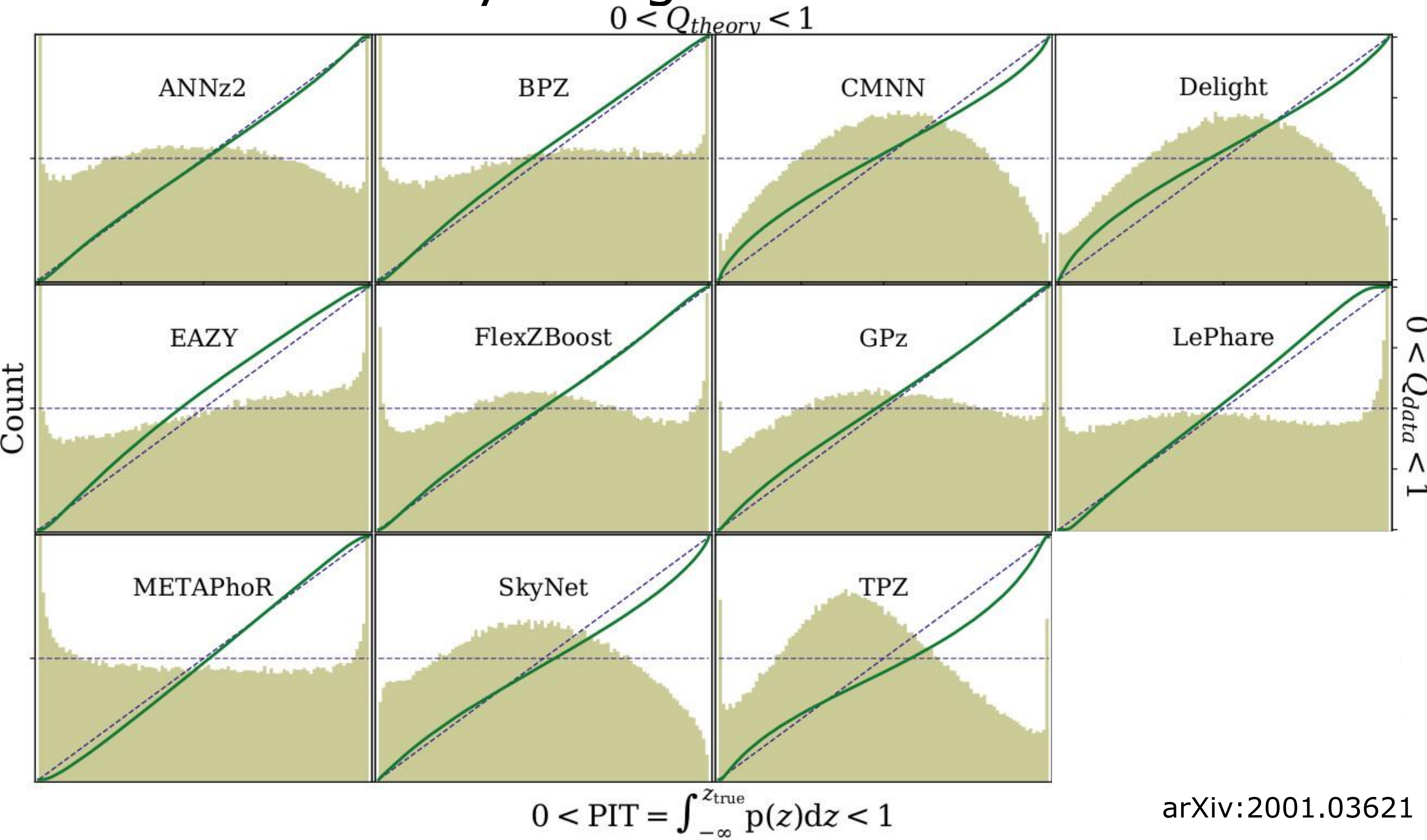
Quantile-quantile (QQ) Plot $\sim \int p(PIT) dPIT$

Can we think of an estimation model. . .



arXiv:2001.03621

. . . that does well by the metrics, without actually doing what we want?



arXiv:2001.03621

The LSST-DESC PZ DC1 experiment

Wait, what do we want again?

Motivation: identify the best photo-*z* posterior code for LSST-DESC

Data: cosmological redshifts & photometry catalog painted on N-body simulation

Control: idealized, shared prior information

The LSST-DESC PZ DC1 experiment

LSSTC DATA SCIENCE FELLOWSHIP PROGRAM

Wait, what do we want again?

"The best" is whatever the metric says it is!

Motivation: identify the best photo-*z* posterior code for LSST-DESC

Data: cosmological redshifts & photometry catalog painted on N-body simulation

Control: idealized, shared prior information

The LSST-DESC PZ DC1 experiment

Wait, what do we want again?

“The best” is whatever the metric says it is!

But what do we really want?

Motivation: identify the best photo-*z* posterior code for LSST-DESC

Data: cosmological redshifts & photometry catalog painted on N-body simulation

Control: idealized, shared prior information

The LSST-DESC PZ DC1 experiment

Wait, what do we want again?

“The best” is whatever the metric says it is!

But what do we really want?

Here, we want an estimator that extracts redshift information from photometry.

Motivation: identify the best photo-*z* posterior code for LSST-DESC

Data: cosmological redshifts & photometry catalog painted on N-body simulation

Control: idealized, shared prior information

Quantitative metrics of 1D PDF ensembles

Root-mean-square Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\int (p_{\mathrm{true}}(z) - \hat{p}_{\mathrm{est}}(z))^2 dz}$$

Kullback-Leibler Divergence (KLD)

$$\mathrm{KLD}[\hat{p}_{\mathrm{est}}(z); p_{\mathrm{true}}(z)] = \int_{-\infty}^{\infty} p_{\mathrm{true}}(z) \ \log\left[\frac{p_{\mathrm{true}}(z)}{\hat{p}_{\mathrm{est}}(z)}\right] \ dz$$

No true posteriors available!

Cumulative Distribution Function (CDF)

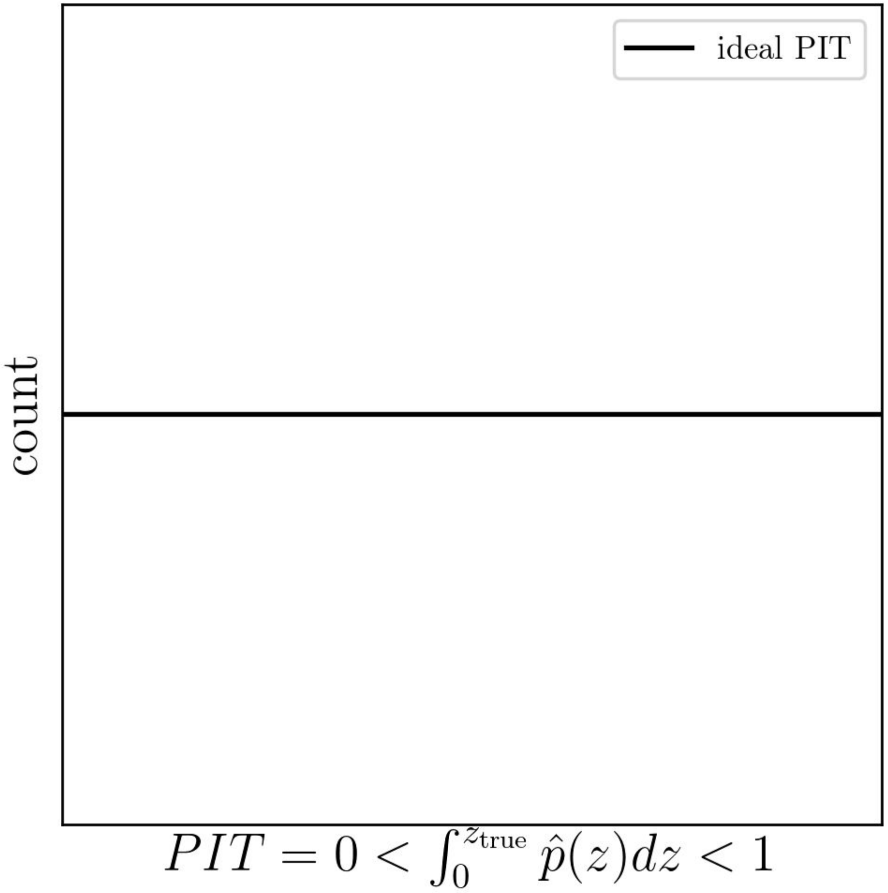
$$\mathrm{CDF}[\hat{p}, z'] \equiv \int_{-\infty}^{z'} \hat{p}(z) dz$$

Probability Integral Transform (PIT)

$$P(\mathrm{PIT} \equiv \mathrm{CDF}[\hat{p}, z_{true}])$$

Quantile-quantile (QQ) Plot $\sim \int p(PIT) dPIT$

Sketch the PIT histogram of an ideal photo-*z* posterior estimator.



$$\mathrm{CDF}[\hat{p}, z'] \equiv \int_{-\infty}^{z'} \hat{p}(z)dz$$

$$P(\mathrm{PIT} \equiv \mathrm{CDF}[\hat{p}, z_{true}])$$

Is the metric inappropriate?

Think *adversarially* to trick the metric into rewarding a wrong answer.

How can you satisfy the metric’s criteria without satisfying the criteria you as an experimenter actually care about?

Identify *pathological* cases that fool the metric but are obviously wrong.

Can you think of a null test in the form of a trivially bad answer that performs well by the metric?

The worst photo-*z* posterior estimator

The worst photo-*z* posterior estimator

1. Make a histogram of training set galaxy redshifts.

The worst photo-*z* posterior estimator

1. Make a histogram of training set galaxy redshifts.
2. Read in the test set photometry.

The worst photo-*z* posterior estimator

1. Make a histogram of training set galaxy redshifts.
2. Read in the test set photometry.
3. Count the number of test set galaxies.

The worst photo-*z* posterior estimator

1. Make a histogram of training set galaxy redshifts.
2. Read in the test set photometry.
3. Count the number of test set galaxies.
4. Discard the test set photometry.

The worst photo-*z* posterior estimator

1. Make a histogram of training set galaxy redshifts.
2. Read in the test set photometry.
3. Count the number of test set galaxies.
4. Discard the test set photometry.
5. Return the histogram of training set galaxy redshifts as the photo-*z* posterior of every test set galaxy.

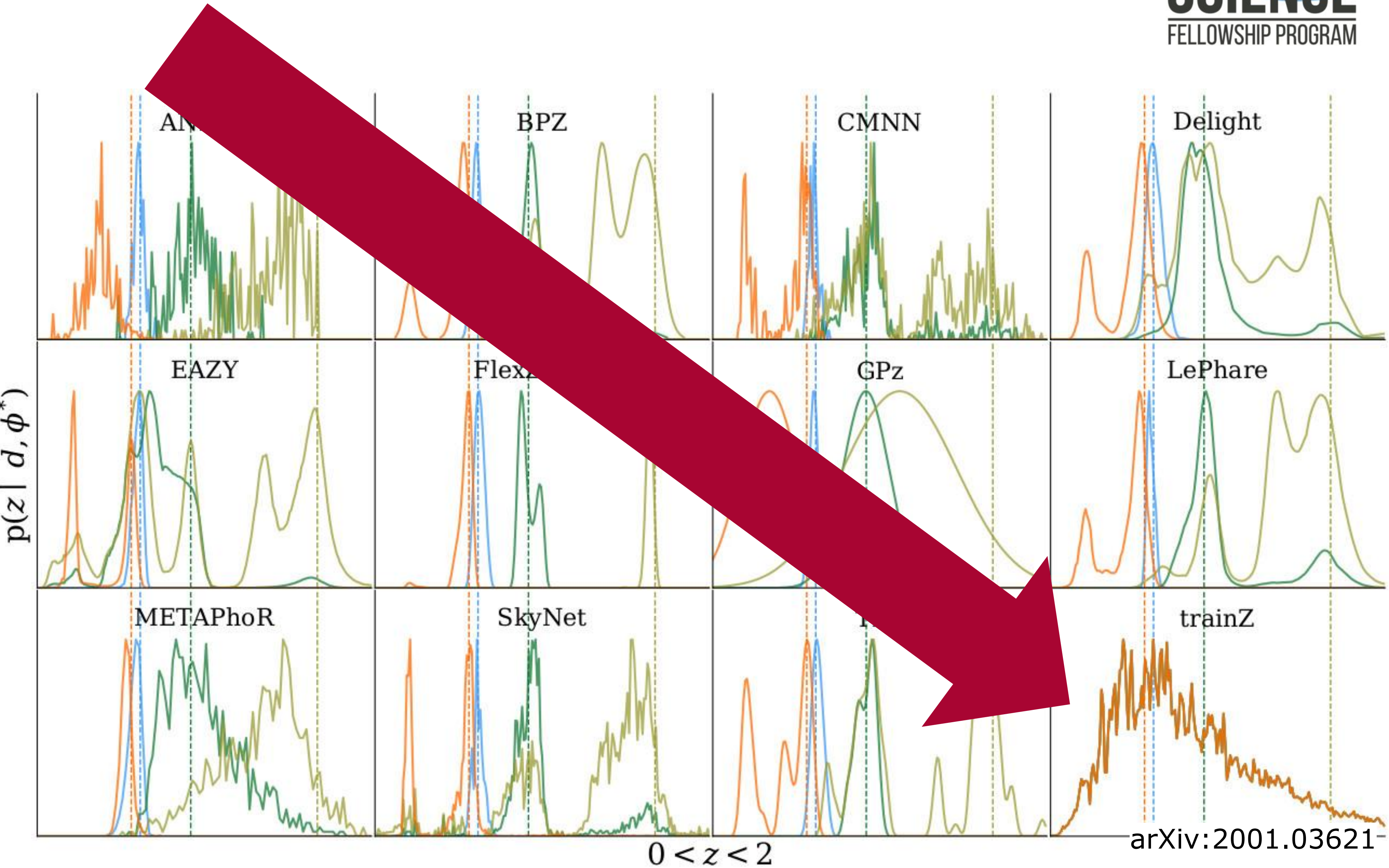
The worst photo-*z* posterior estimator

1. Make a histogram of training set galaxy redshifts.
2. Read in the test set photometry.
3. Count the number of test set galaxies.
4. Discard the test set photometry.
5. Return the histogram of training set galaxy redshifts as the photo-*z* posterior of every test set galaxy.
6. …

The worst photo-*z* posterior estimator

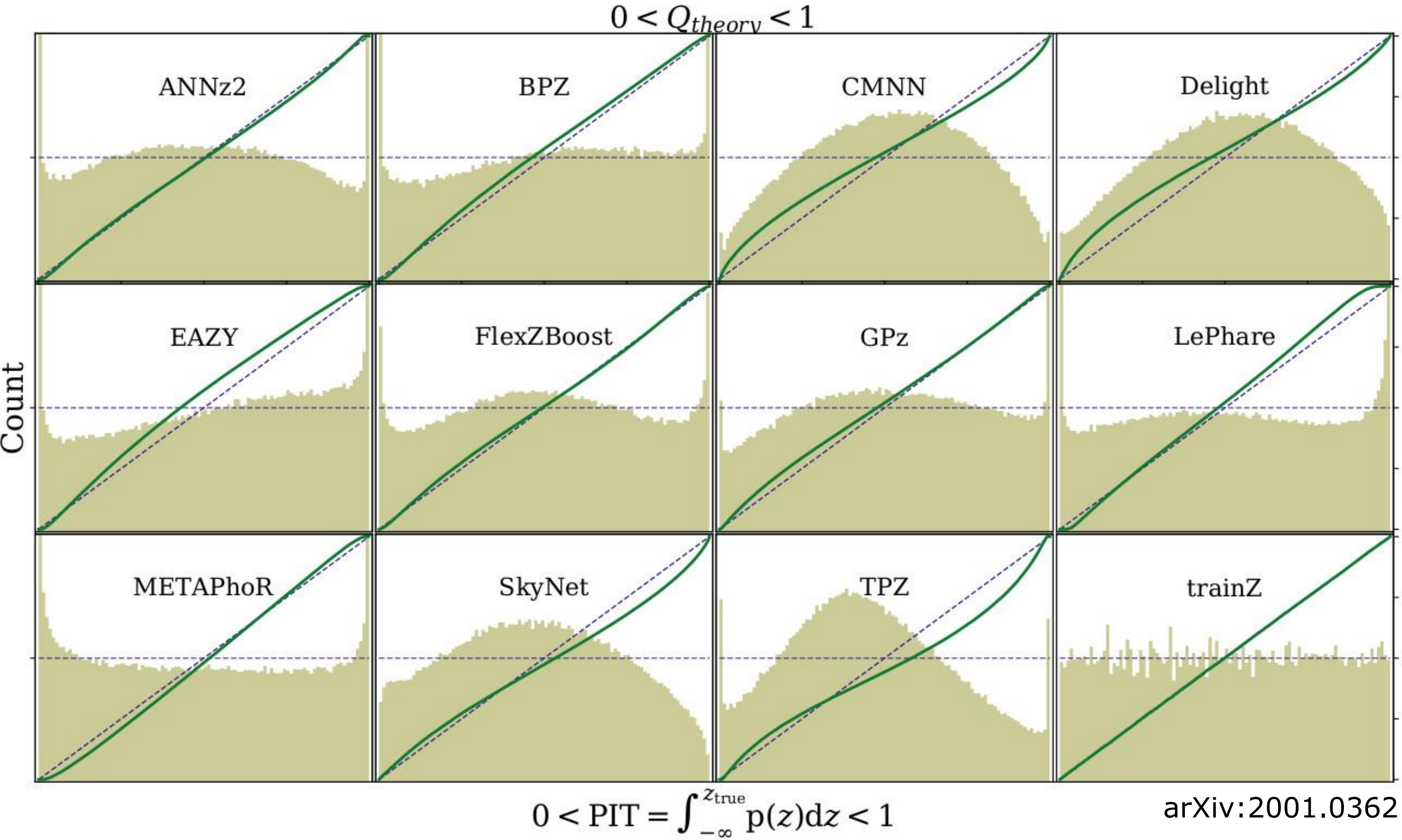
1. Make a histogram of training set galaxy redshifts.
2. Read in the test set photometry.
3. Count the number of test set galaxies.
4. Discard the test set photometry.
5. Return the histogram of training set galaxy redshifts as the photo-*z* posterior of every test set galaxy.
6. …
7. Profit!

`trainZ` is a “control” case



arXiv:2001.03621

`trainZ` has nearly perfect PIT & QQ.

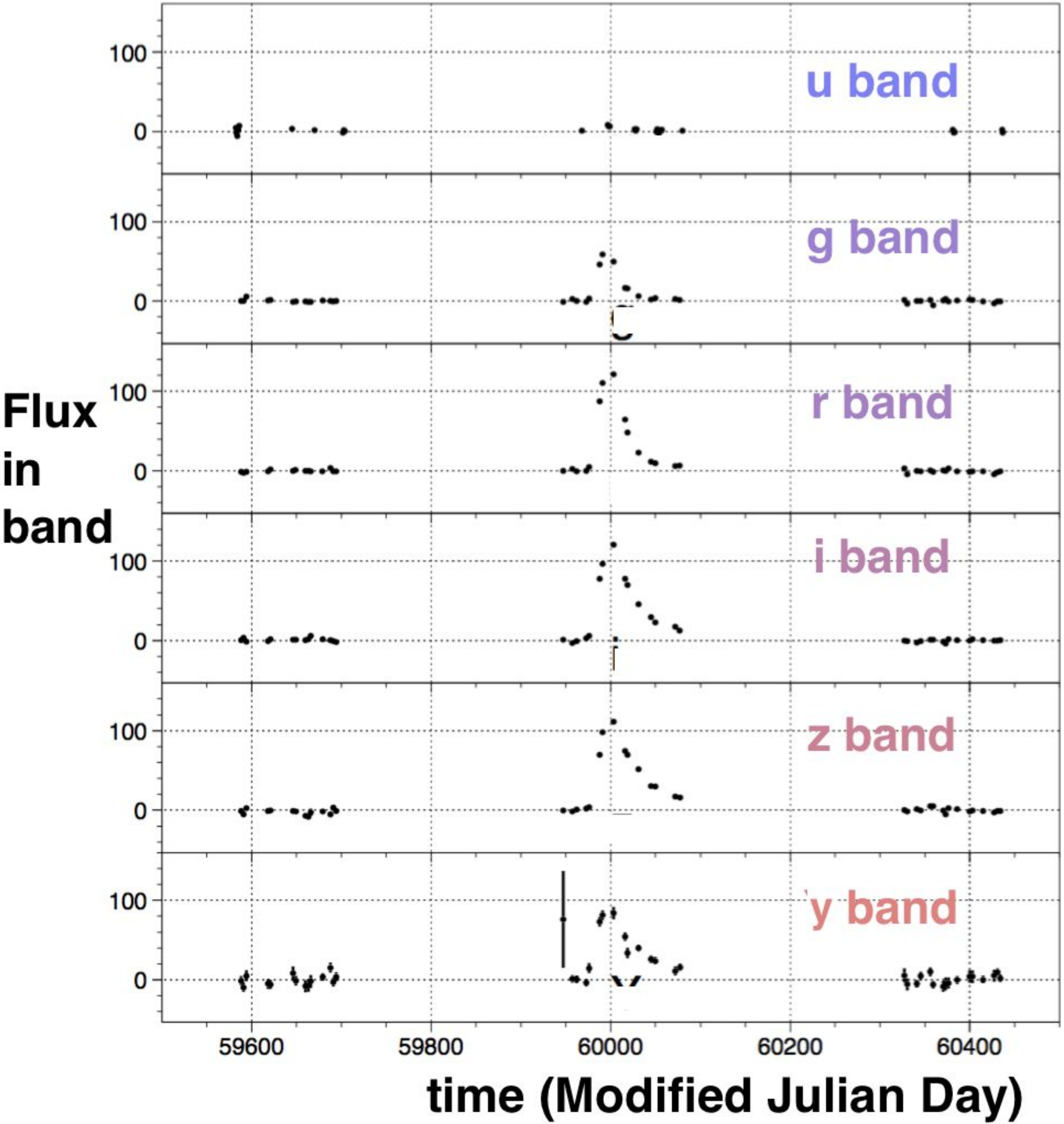


arXiv:2001.03621

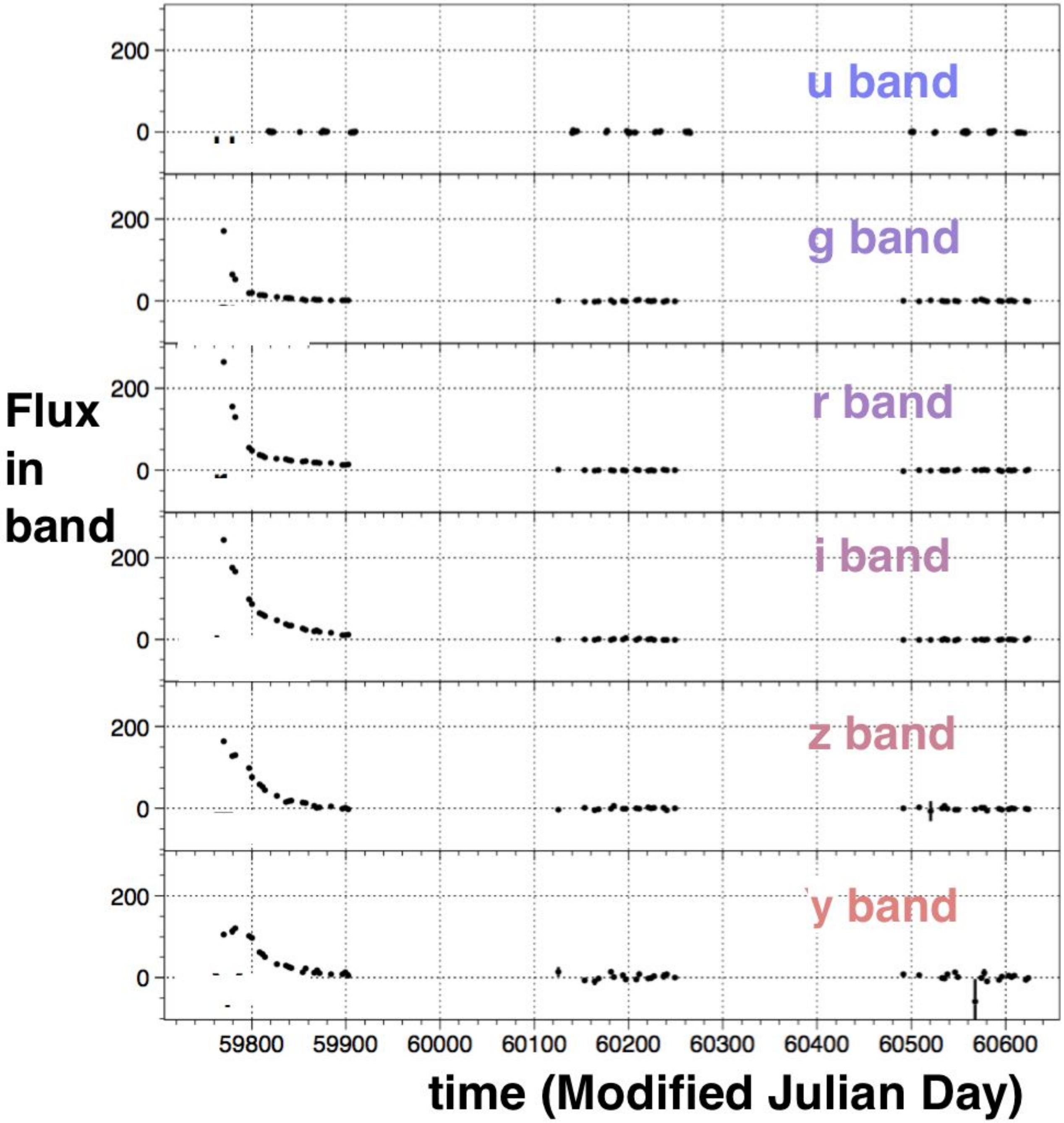
How does one create a specialized metric for a specific goal?

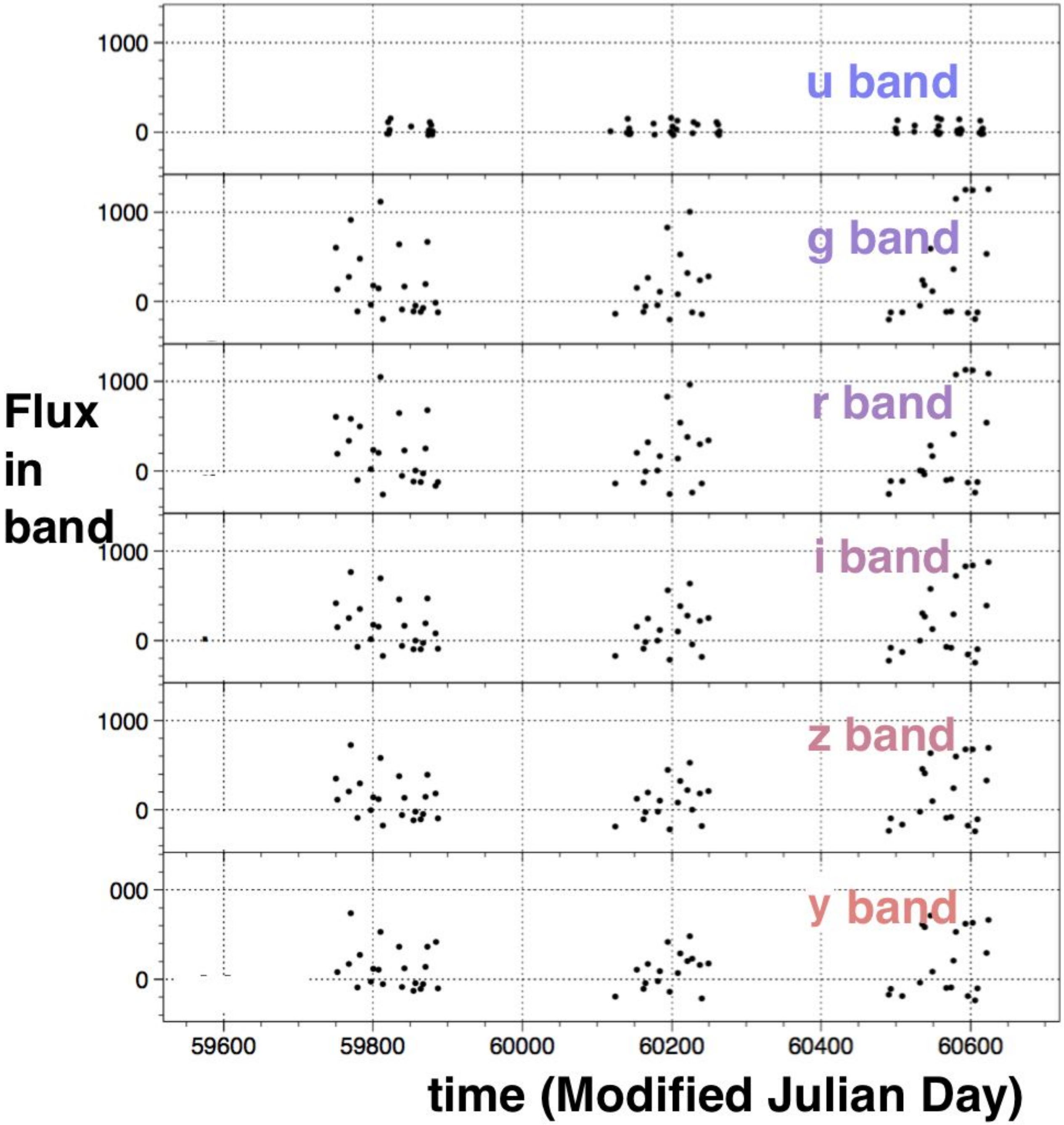
Context: The Photometric LSST AStronomical TIme-series Classification Challenge (PLAsTiCC)

LSSTC DATA SCIENCE FELLOWSHIP PROGRAM



arXiv:1810.00001


u band
g band
r band
i band
z band
y band
Flux in band
200
0
59800
59900
60000
60100
60200
60300
60400
60500
60600
time (Modified Julian Day)



LSSTC
DATA
SCIENCE
ELLOWSHIP PROGRAM

The challenge of the PLAsTiCC metric

What made PLAsTiCC challenging

Noisy, incomplete LCs
Many classes, subclasses
Diverse science applications

Why the metric required critical thinking

Classification PMFs vs. labels
Kaggle requires single scalar

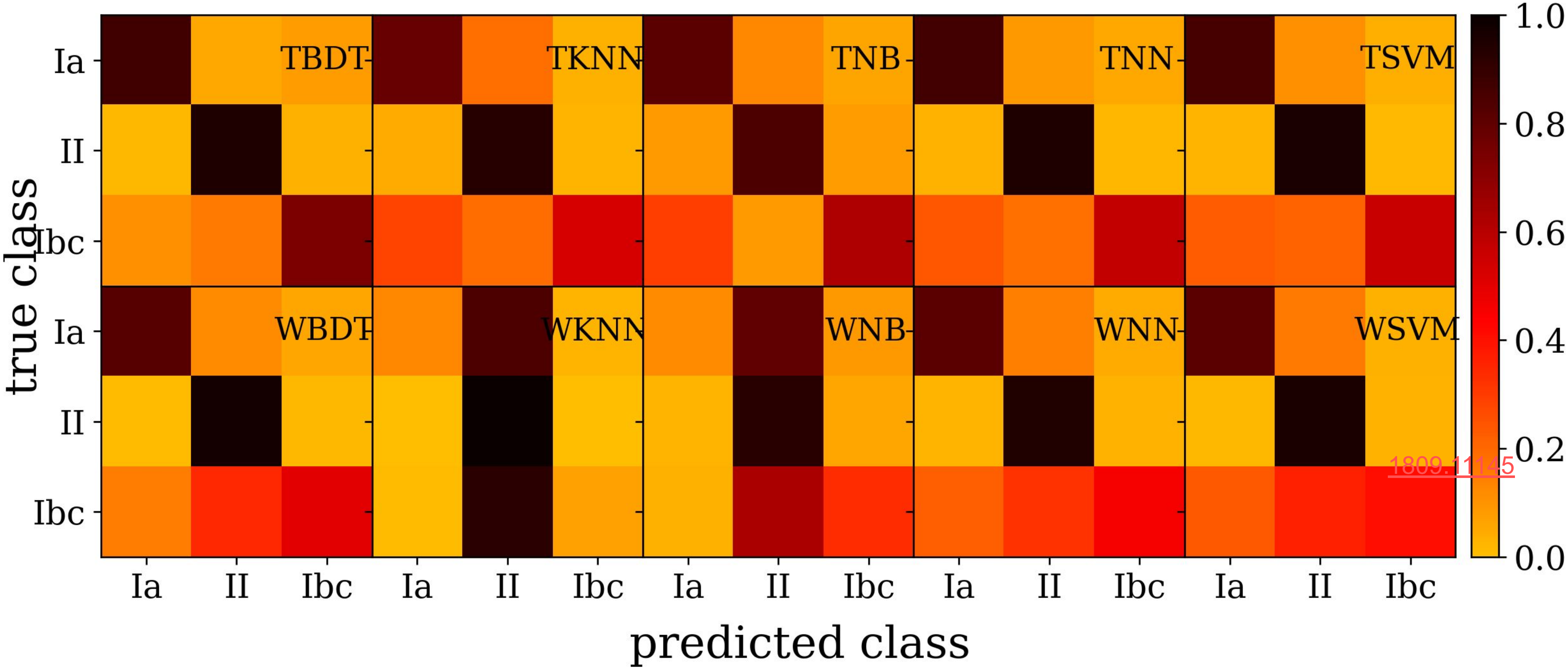
Confusion matrix & T/F P/N rates

	Actual true	Actual false
Predicted true	True positives (TP)	False positives (FP)
Predicted false	False negatives (FN)	True negatives (TN)

Purity = Precision = TP / (TP + FP)

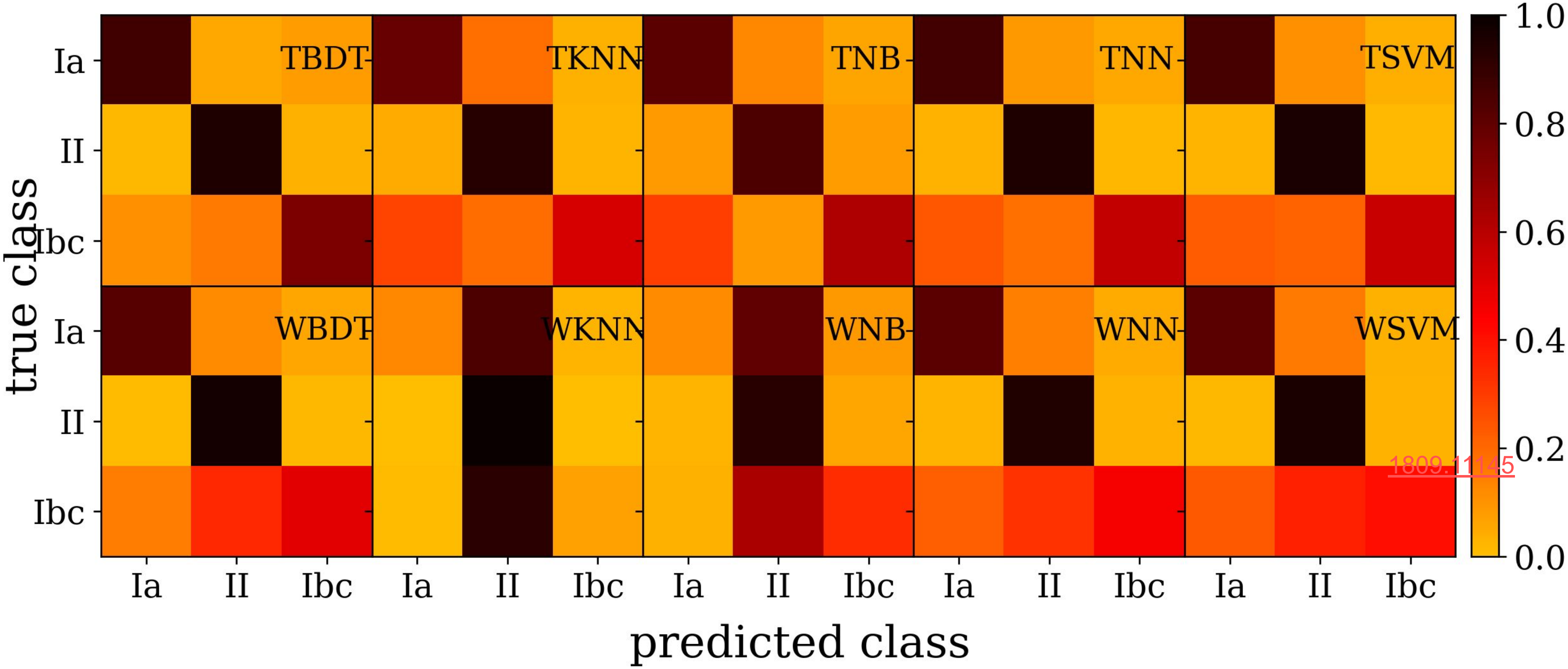
Efficiency = Recall = TP / (TP + FN)

Confusion matrix for SN classification



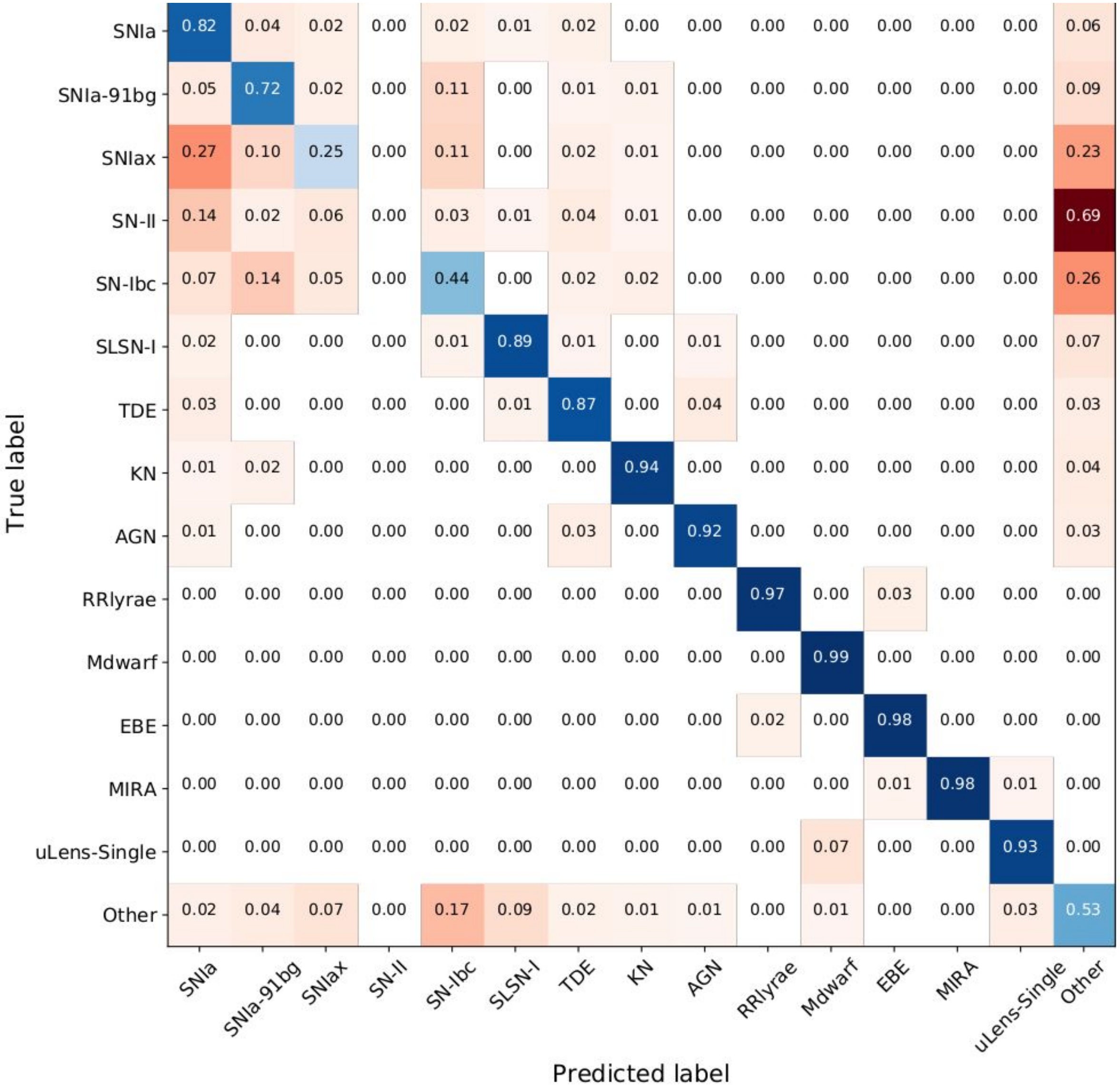
arXiv:1809.11145

What do we want to reward vs. punish?



arXiv:1809.11145

Many-class confusion matrix

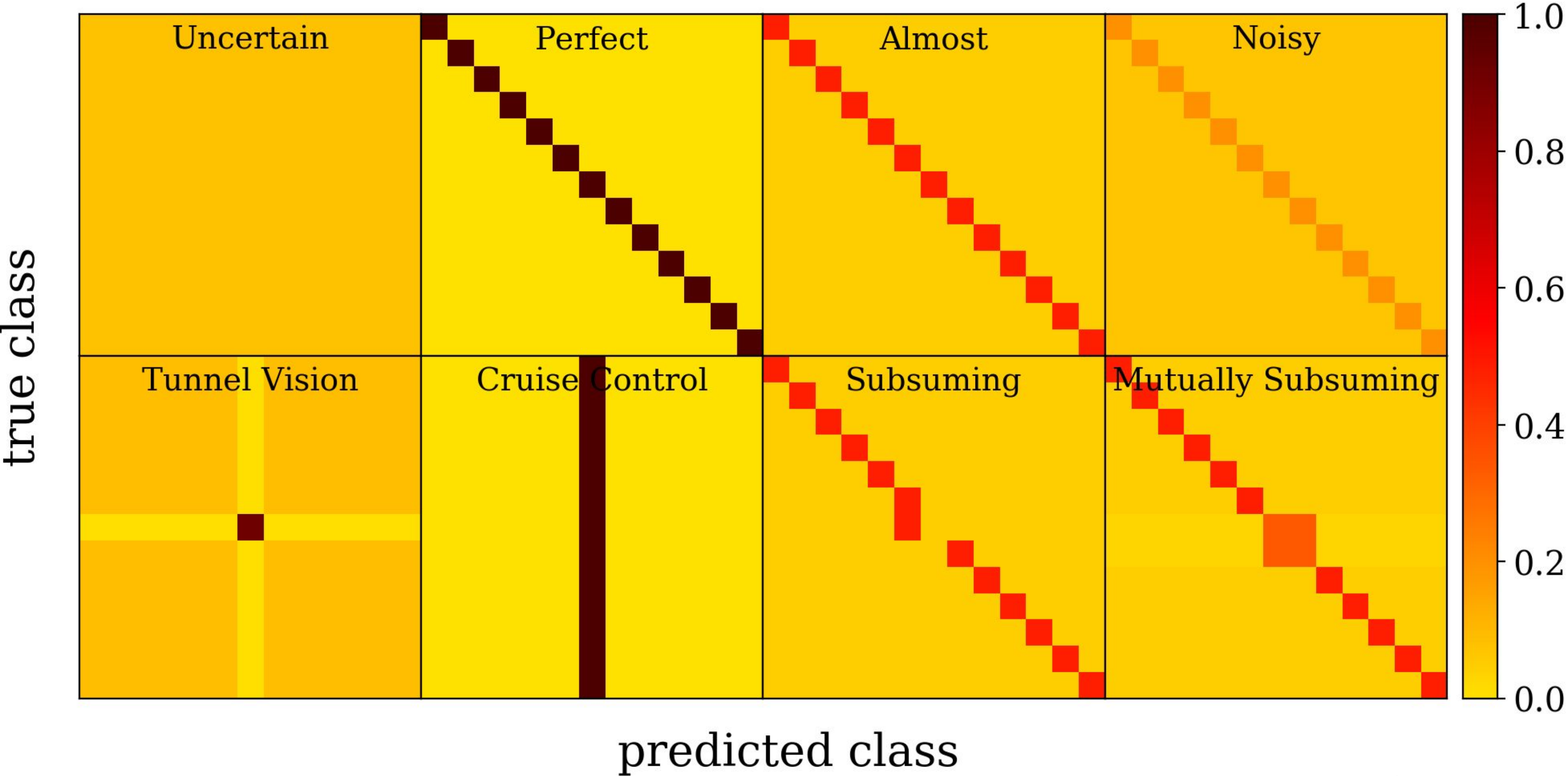


	SNIa	SNIa-91bg	SNIax	SN-II	SN-Ibc	SLSN-I	TDE	KN	AGN	RRlyrae	Mdwarf	EBE	MIRA	uLens-Single	Other
SNIa	0.82	0.04	0.02	0.00	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06
SNIa-91bg	0.05	0.72	0.02	0.00	0.11	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.09
SNIax	0.27	0.10	0.25	0.00	0.11	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.23
SN-II	0.14	0.02	0.06	0.00	0.03	0.01	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.69
SN-Ibc	0.07	0.14	0.05	0.00	0.44	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.26
SLSN-I	0.02	0.00	0.00	0.00	0.01	0.89	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.07
TDE	0.03	0.00	0.00	0.00	0.00	0.01	0.87	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.03
KN	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.04
AGN	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.03
RRlyrae	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.03	0.00	0.00	0.00
Mdwarf	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00
EBE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.98	0.00	0.00	0.00
MIRA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.98	0.01	0.00
uLens-Single	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.93	0.00
Other	0.02	0.04	0.07	0.00	0.17	0.09	0.02	0.01	0.01	0.00	0.01	0.00	0.00	0.03	0.53

arXiv:2012.12392

Think adversarially! Devise a pathological classifier and sketch its confusion matrix.

Think adversarially! Devise a pathological classifier and sketch its confusion matrix.



arXiv:1809.11145

Quantitative metrics of 1D PDFs

Root-mean-square Error

$$\mathrm{RMSE} = \sqrt{\int \left(P(z) - \hat{P}(z)\right)^2 dz}$$

Kullback-Leibler Divergence

$$\mathrm{KLD}\left[\hat{P}(z)|P(z)\right] = \int_{-\infty}^{\infty} P(z)\ \log\left[\frac{P(z)}{\hat{P}(z)}\right]\ dz$$

Quantitative metrics of categorical PMFs

Root-mean-square Error

Brier Score $B_{\mathrm{LC}\ n} \equiv \sum_{\mathrm{class}\ m=1}^{M} (\tau_{n,m} - \hat{p}(m|\mathrm{data}_n))^2$

Kullback-Leibler Divergence

Log-loss $L_{\mathrm{LC}\ n} \equiv - \sum_{\mathrm{class}\ m=1}^{M} \tau_{n,m} \ln[\hat{p}(m|\mathrm{data}_n)]$

What free parameters are chosen by the experimenter?

Root-mean-square Error

Brier Score $B_{\mathrm{LC}\ n} \equiv \sum_{\mathrm{class}\ m=1}^{M} (\tau_{n,m} - \hat{p}(m|\mathrm{data}_n))^2$

Kullback-Leibler Divergence

Log-loss $L_{\mathrm{LC}\ n} \equiv -\sum_{\mathrm{class}\ m=1}^{M} \tau_{n,m} \ln[\hat{p}(m|\mathrm{data}_n)]$

$$Q = \frac{1}{\sum_m W_m} \sum_{m=1}^{M} W_m \frac{1}{\sum_n w_{n,m}} \sum_{n=1}^{N} w_{n,m} Q_{n,m}$$

What free parameters are chosen by the experimenter?

Root-mean-square Error

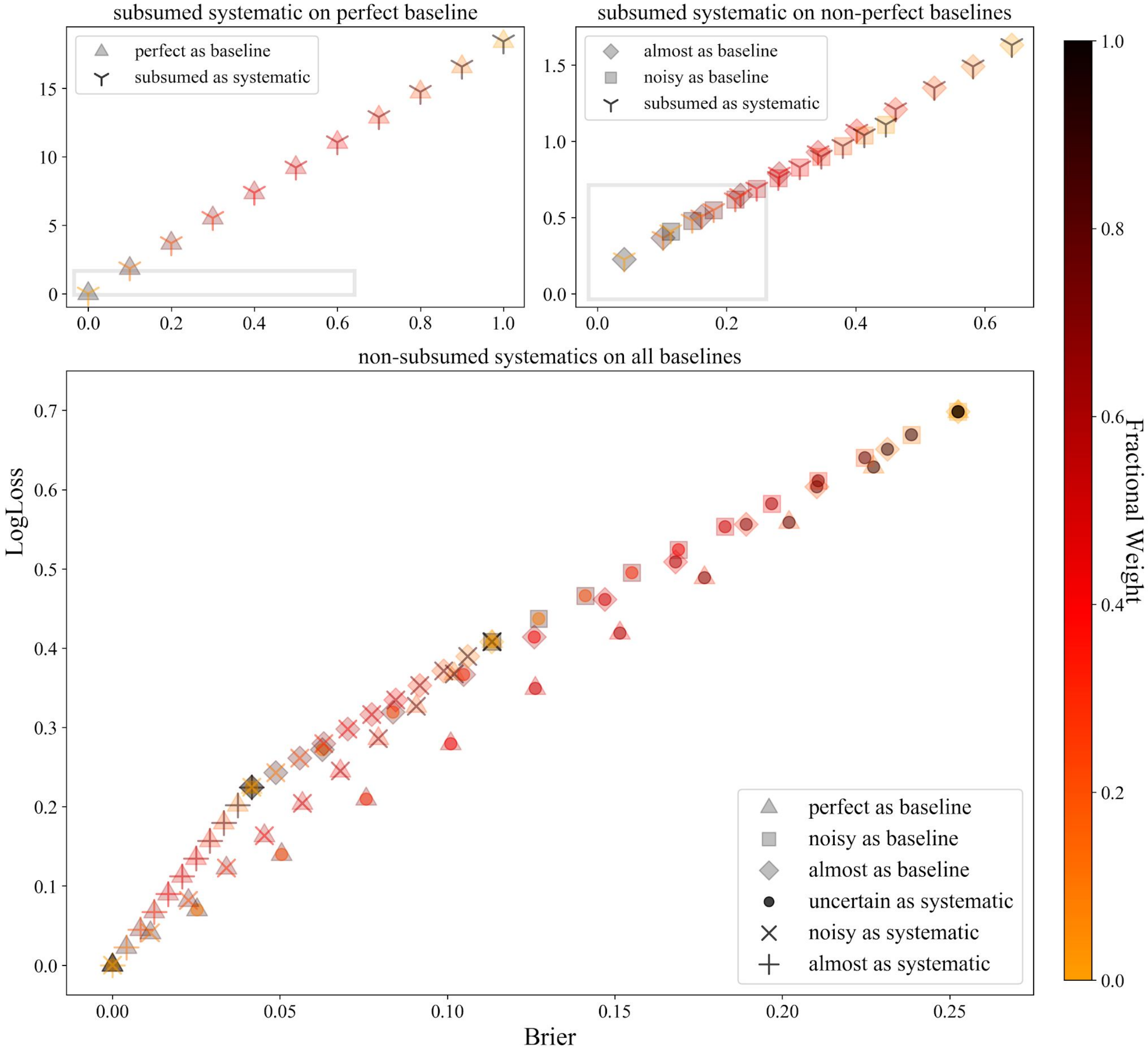
Brier Score $B_{\mathrm{LC}\ n} \equiv \sum_{\mathrm{class}\ m=1}^{M} (\tau_{n,m} - \hat{p}(m|\mathrm{data}_n))^2$

Kullback-Leibler Divergence

Log-loss $L_{\mathrm{LC}\ n} \equiv -\sum_{\mathrm{class}\ m=1}^{M} \tau_{n,m} \ln[\hat{p}(m|\mathrm{data}_n)]$

$$Q = \frac{1}{\sum_m W_m} \sum_{m=1}^{\boxed{M}} \boxed{W_m} \frac{1}{\sum_n w_{n,m}} \sum_{n=1}^{\boxed{N}} \boxed{w_{n,m}} Q_{n,m}$$

tl;dr both metrics considered were robust to pathological classifiers



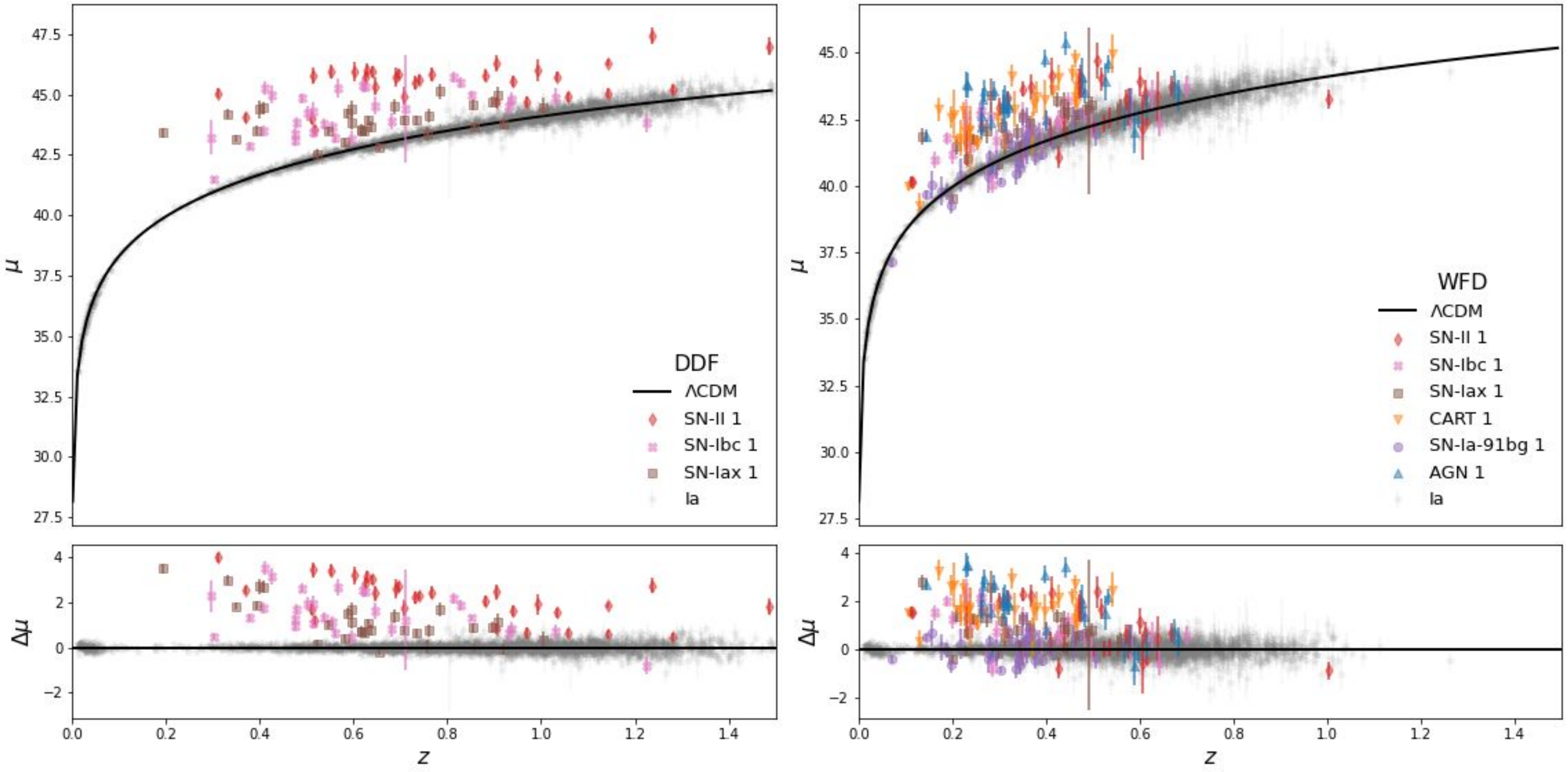
arXiv:1809.11145

Another astrophysical example: transient classification for SNIa cosmology

How can we choose classification metrics that reward what we really want?

LSSTC DATA SCIENCE FELLOWSHIP PROGRAM

SN Ia cosmology with contaminants



SN Ia cosmology with contaminants

Goal: build spectroscopic training set (prior) for classifier (model) to get best constraints on SNIa cosmology given limited follow-up resources

Wait, what do we want again?

"The best" is whatever the metric says it is!

But what do we really want?

Here, we want a decision metric that identifies most cosmologically impactful light curves for follow-up

Classification metrics alone are degenerate

- The *accuracy* is defined as

$$\mathcal{A} = \frac{TP + TN}{N}, \tag{1}$$

where a value closer to unity is more accurate.

- The *purity* (also known as *precision*) is defined as

$$\mathcal{P} = \frac{TP}{TP + FP}, \tag{2}$$

where a value closer to unity is more pure.

- The *efficiency* (also known as *recall*) is defined as

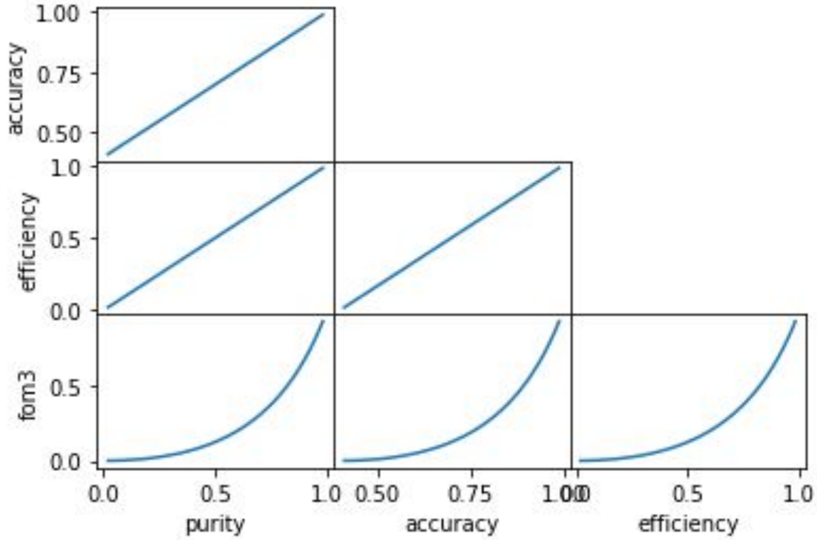
$$\mathcal{R} = \frac{TP}{TP + FN}, \tag{3}$$

where a value closer to unity is more efficient.

- The SNPhotCC defined a *Figure of Merit (FoM)*,

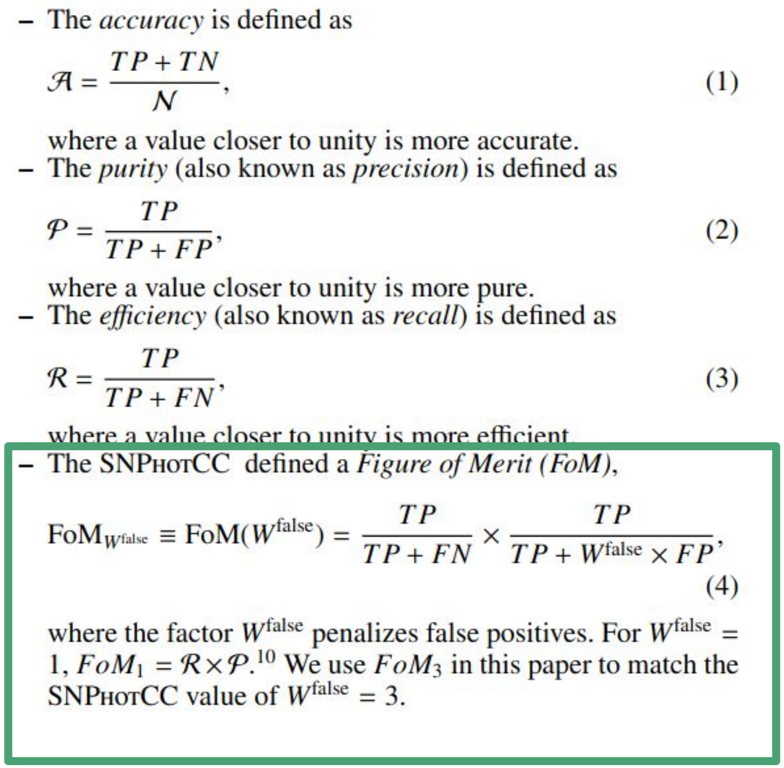
$$\mathrm{FoM}_{W^{\mathrm{false}}} \equiv \mathrm{FoM}(W^{\mathrm{false}}) = \frac{TP}{TP + FN} \times \frac{TP}{TP + W^{\mathrm{false}} \times FP}, \tag{4}$$

where the factor W^{false} penalizes false positives. For $W^{\mathrm{false}} = 1$, $FoM_1 = \mathcal{R} \times \mathcal{P}$.[10] We use FoM_3 in this paper to match the SNPhotCC value of $W^{\mathrm{false}} = 3$.



and don’t relate directly to cosmology.
Will using them achieve our goals?

Classification metrics alone are degenerate



- The *accuracy* is defined as

$$\mathcal{A} = \frac{TP + TN}{N}, \tag{1}$$

where a value closer to unity is more accurate.

- The *purity* (also known as *precision*) is defined as

$$\mathcal{P} = \frac{TP}{TP + FP}, \tag{2}$$

where a value closer to unity is more pure.

- The *efficiency* (also known as *recall*) is defined as

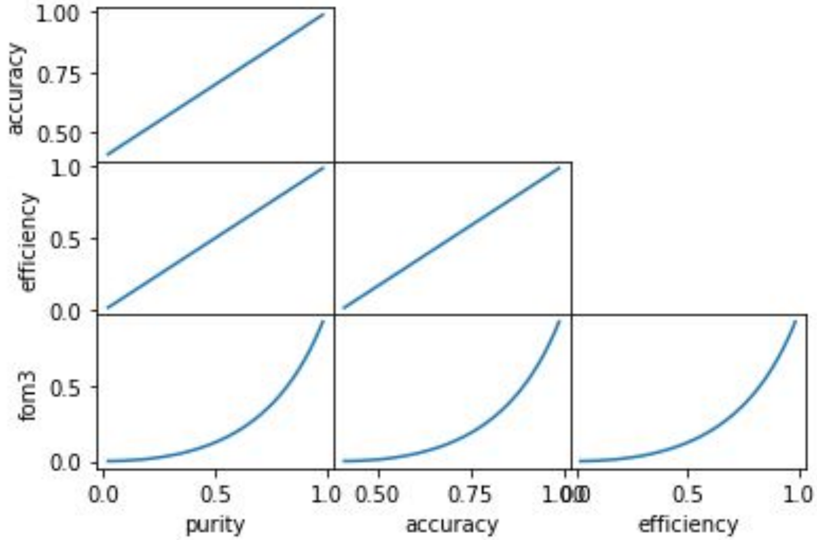
$$\mathcal{R} = \frac{TP}{TP + FN}, \tag{3}$$

where a value closer to unity is more efficient.

- The SNPhotCC defined a *Figure of Merit (FoM)*,

$$\mathrm{FoM}_{W^{\mathrm{false}}} \equiv \mathrm{FoM}(W^{\mathrm{false}}) = \frac{TP}{TP + FN} \times \frac{TP}{TP + W^{\mathrm{false}} \times FP}, \tag{4}$$

where the factor W^{false} penalizes false positives. For $W^{\mathrm{false}} = 1$, $FoM_1 = \mathcal{R} \times \mathcal{P}$.[10] We use FoM_3 in this paper to match the SNPhotCC value of $W^{\mathrm{false}} = 3$.



and don’t relate directly to cosmology.
Will using them achieve our goals?

Is the metric appropriate?

Think *adversarially* to trick the metric into rewarding a wrong answer.

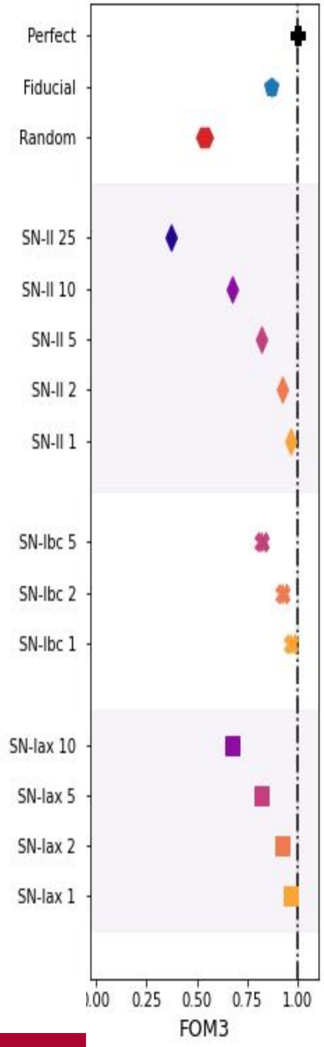
How can you satisfy the metric's criteria without satisfying the criteria you as an experimenter actually care about?

Identify *pathological* cases that fool the metric but are obviously wrong.

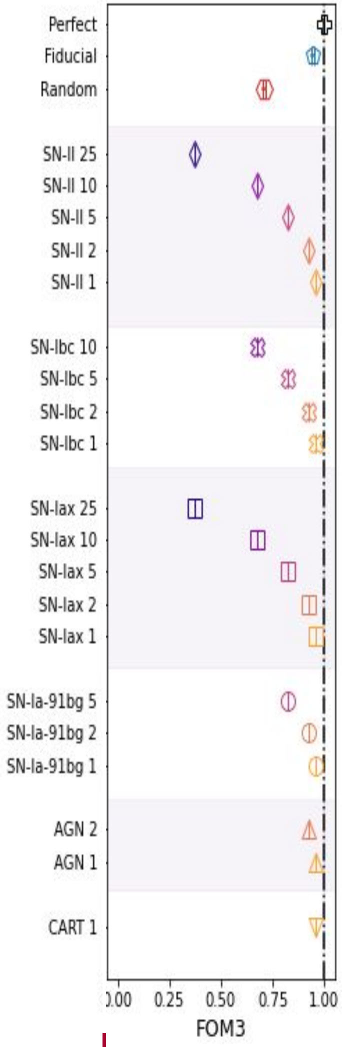
Can you think of a null test in the form of a trivially bad answer that performs well by the metric?

SNPhotCC FoM can't distinguish contaminant class

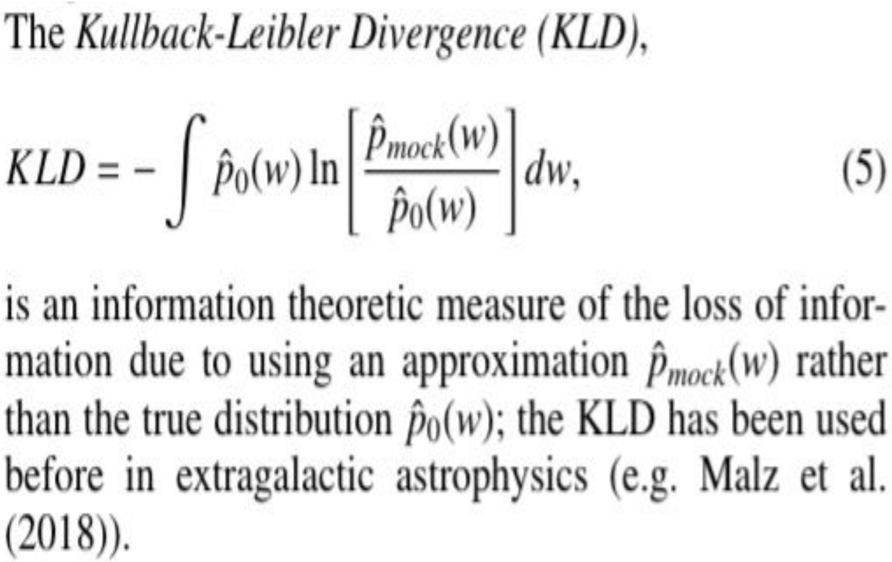
DDF



WFD



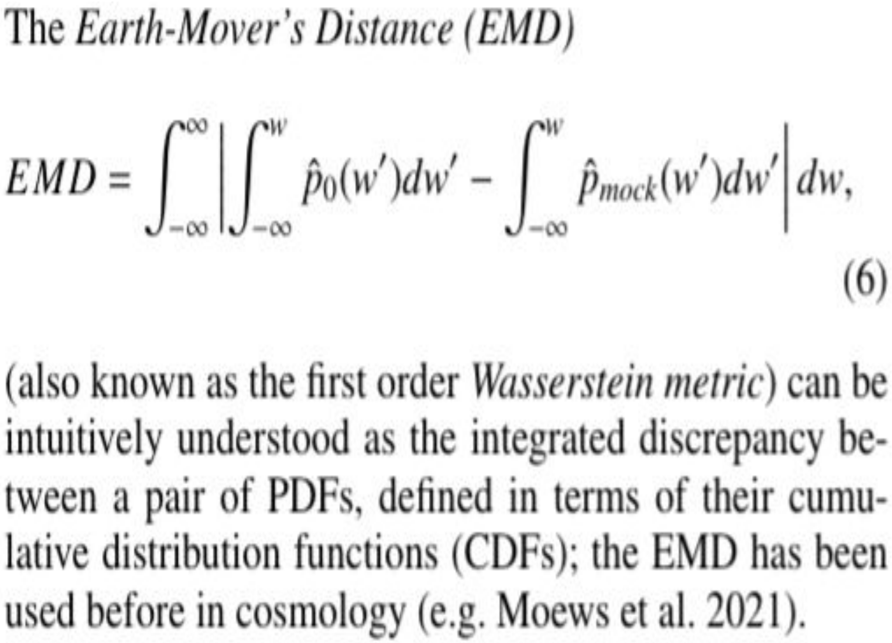
Quantitative metrics of posterior samples $\{ w_i \}$



The *Kullback-Leibler Divergence (KLD)*,

$$KLD = -\int \hat{p}_0(w) \ln \left[\frac{\hat{p}_{mock}(w)}{\hat{p}_0(w)} \right] dw, \qquad (5)$$

is an information theoretic measure of the loss of information due to using an approximation $\hat{p}_{mock}(w)$ rather than the true distribution $\hat{p}_0(w)$; the KLD has been used before in extragalactic astrophysics (e.g. Malz et al. (2018)).



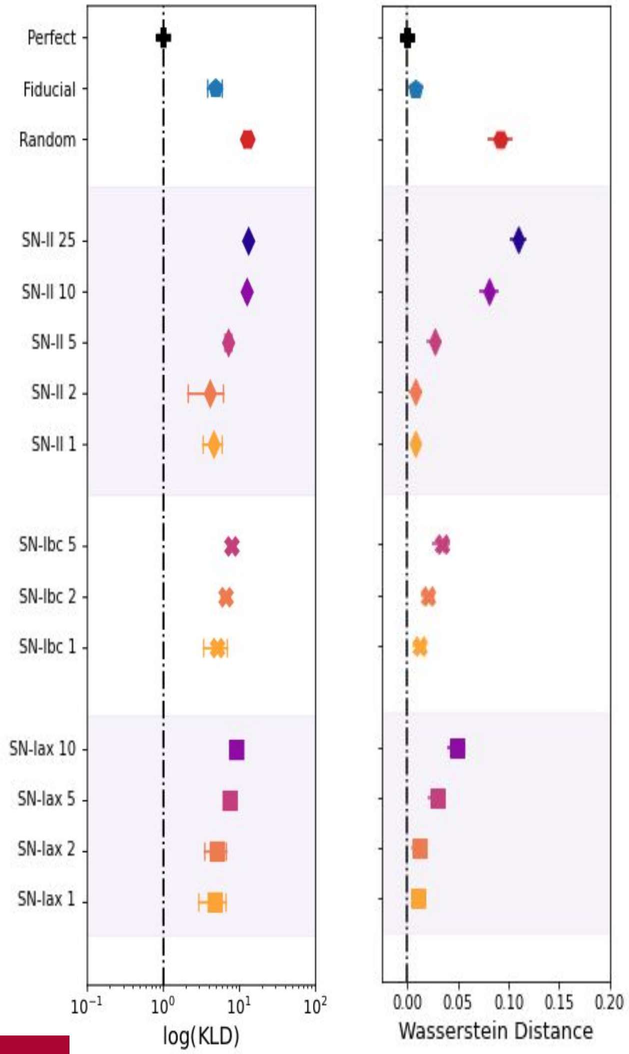
The *Earth-Mover's Distance (EMD)*

$$EMD = \int_{-\infty}^{\infty} \left| \int_{-\infty}^{w} \hat{p}_0(w')dw' - \int_{-\infty}^{w} \hat{p}_{mock}(w')dw' \right| dw, \qquad (6)$$

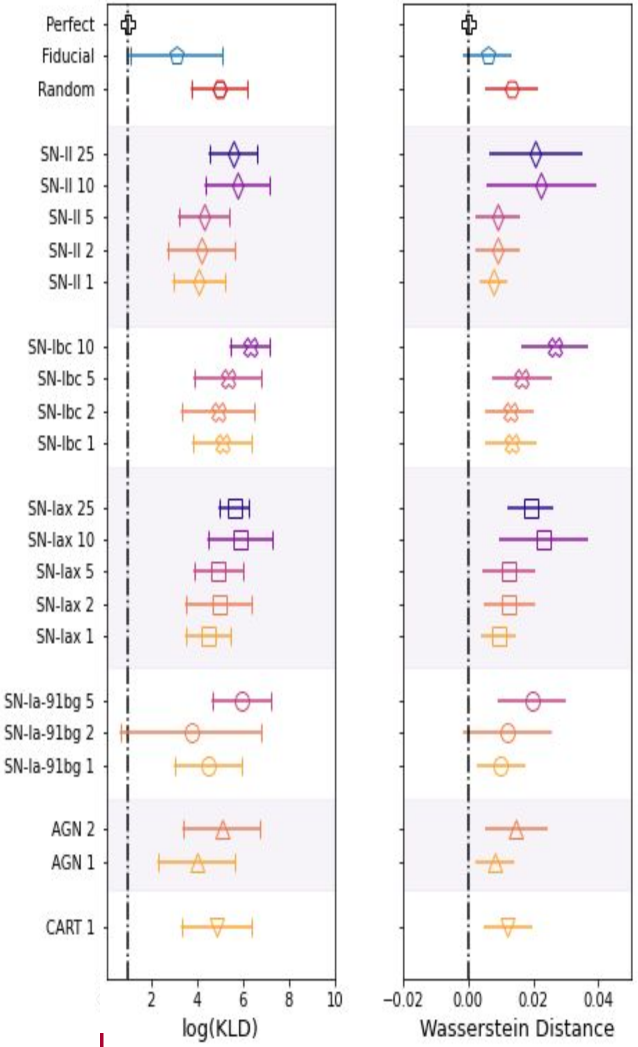
(also known as the first order *Wasserstein metric*) can be intuitively understood as the integrated discrepancy between a pair of PDFs, defined in terms of their cumulative distribution functions (CDFs); the EMD has been used before in cosmology (e.g. Moews et al. 2021).

Posterior samples' KLD and EMD are sensitive to contaminant class

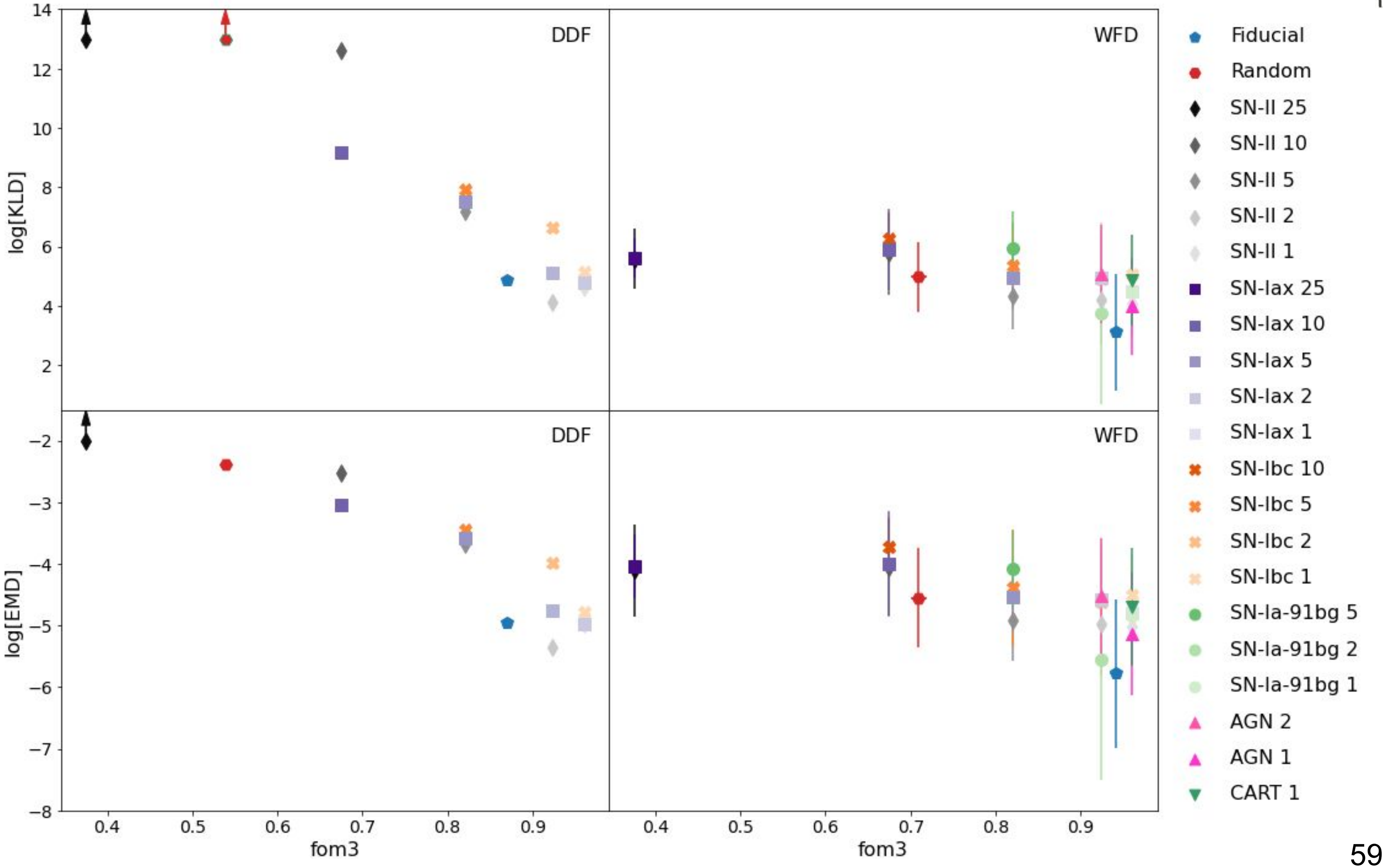
DDF



WFD



Contaminant type matters to cosmology



Closing thoughts

Probabilistic graphical models are pretty much my favorite thing!
But my secret agenda is for you to incorporate this way of thinking into your hacks, for the experiments in which you prove how great your hierarchical models are.