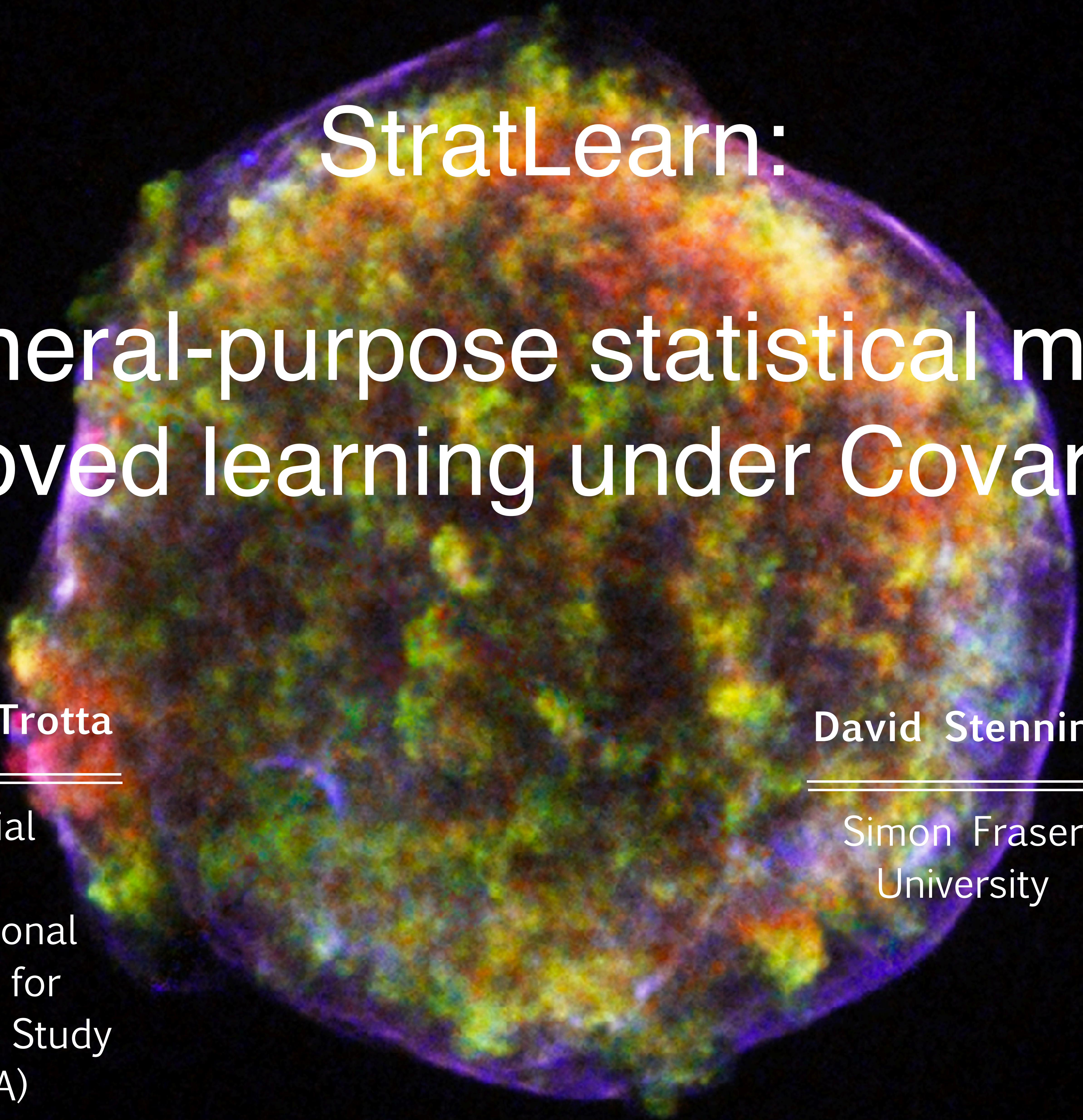
@R_Trotta



StratLearn:

A general-purpose statistical method for improved learning under Covariate Shift

Max Autenrieth

Imperial (Statistics Section)

Roberto Trotta

Imperial & International School for Advanced Study (SISSA)

David Stenning

Simon Fraser University

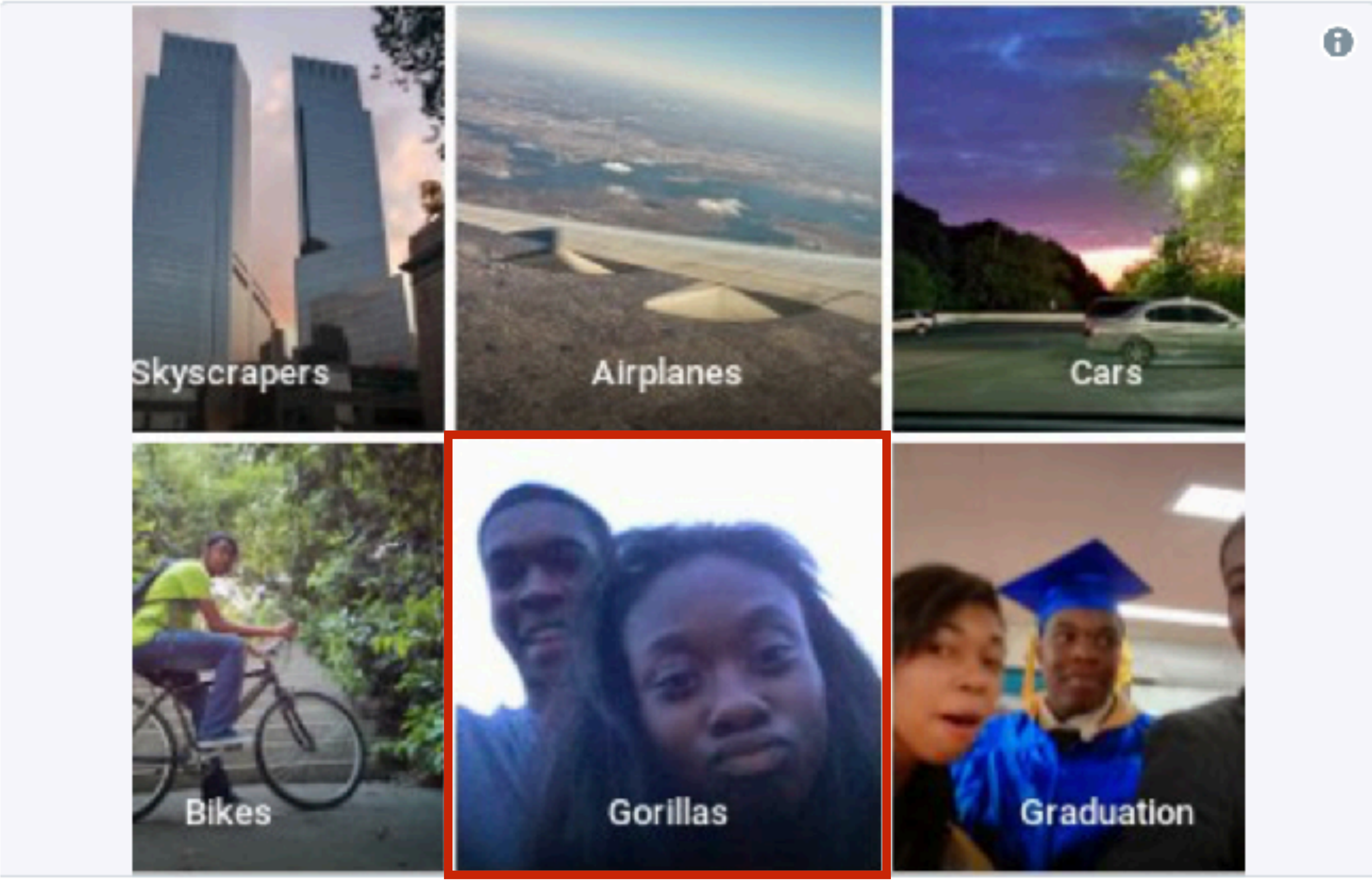
David van Dyk

Imperial (Statistics Section)

June 14th 2022

The problem: machine learning classifiers trained on non-representative data generalize poorly.

2015

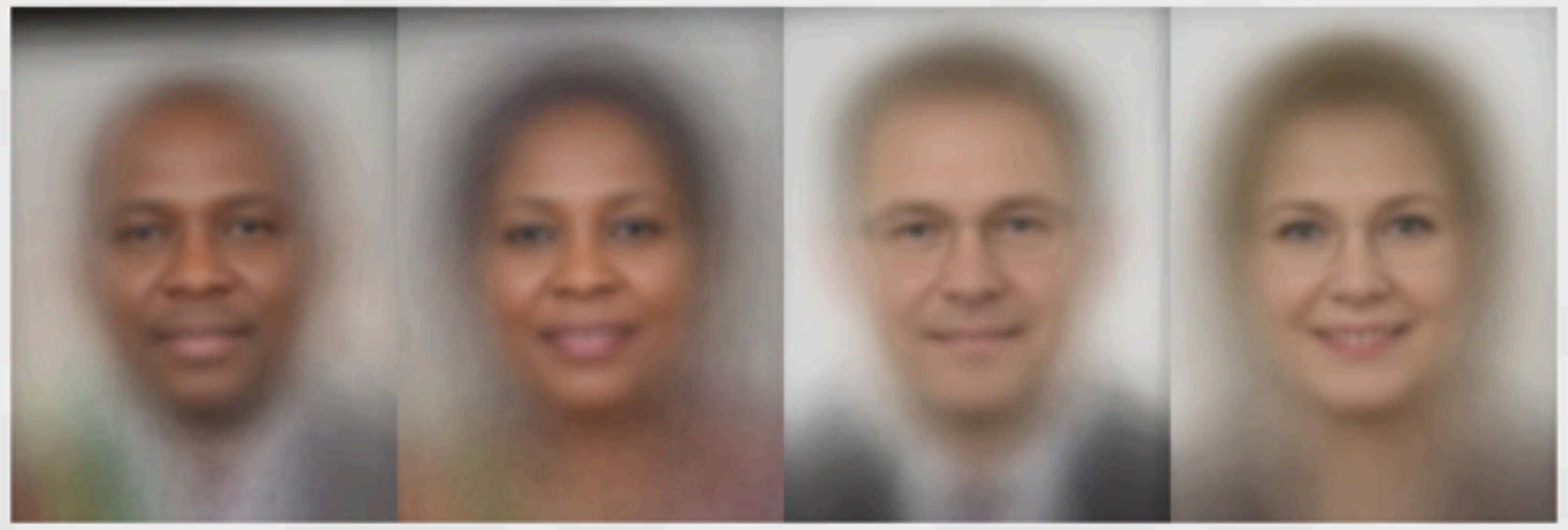


Jacky lives on @jalcine@playvicious.social now.
@jackyalcine

2019

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Gender Shades (MIT Media Lab, 2019)



Poor accuracy in facial recognition for dark skinned females

Cosmology

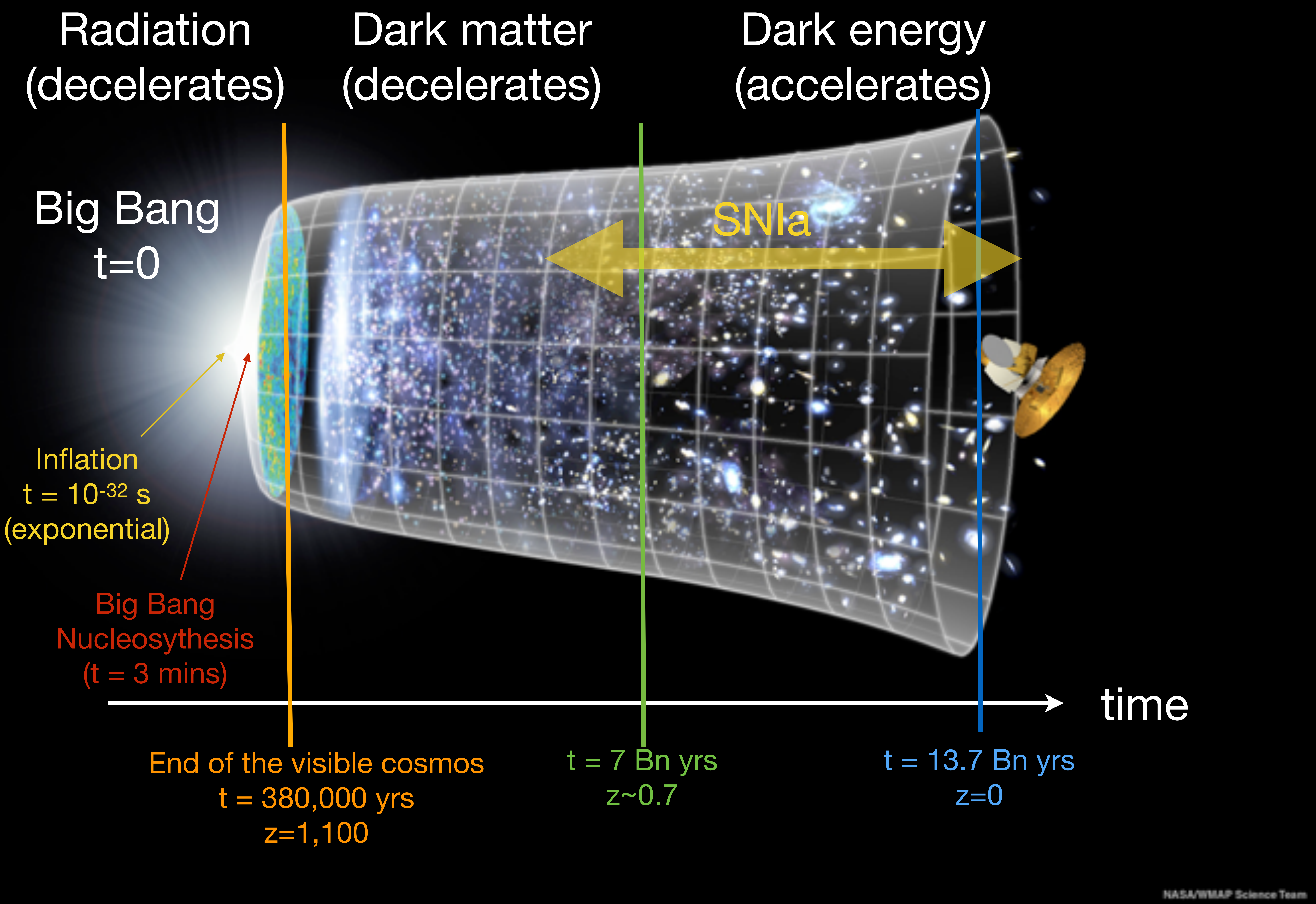
Incorrect classification of Type Ia vs non-Ia from photometric data leads to cosmological parameters systematic bias.

2018

TOM SIMONITE BUSINESS 01.11.2018 07:00 AM

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

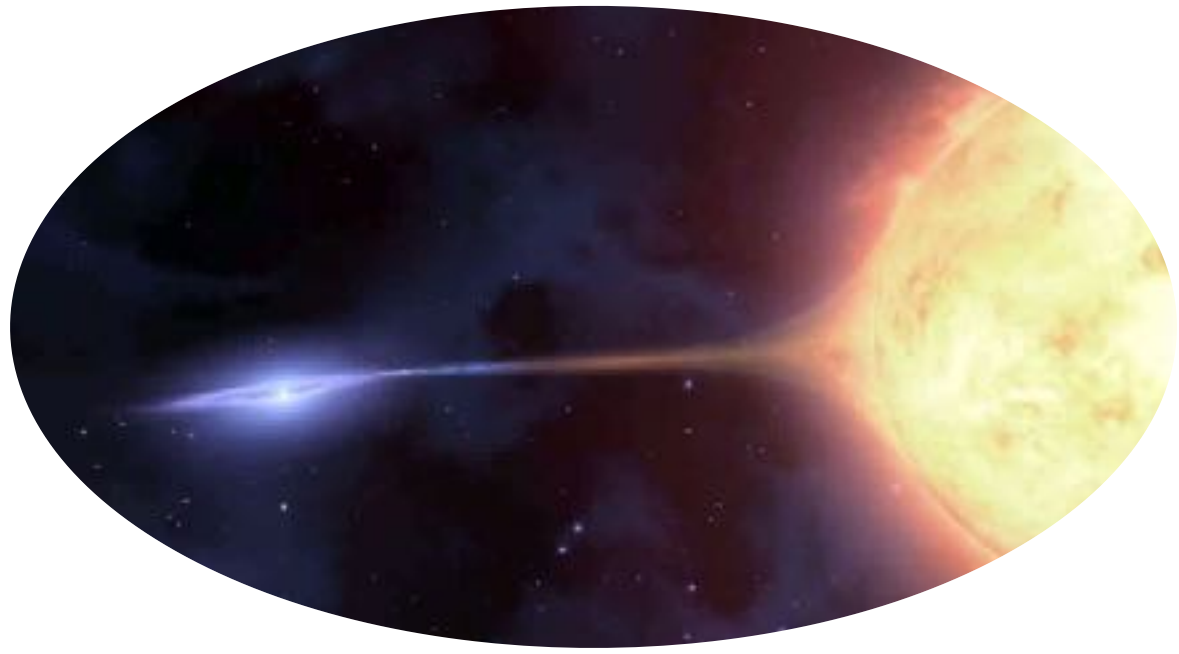

Radiation
(decelerates)
Dark matter
(decelerates)
Dark energy
(accelerates)
Big Bang
t=0
SNIa
Inflation
t = 10^{-32} s
(exponential)
Big Bang
Nucleosynthesis
(t = 3 mins)
time
End of the visible cosmos
t = 380,000 yrs
z=1,100
t = 7 Bn yrs
z~0.7
t = 13.7 Bn yrs
z=0
NASA/WMAP Science Team

Supernovae Type Ia: Origin

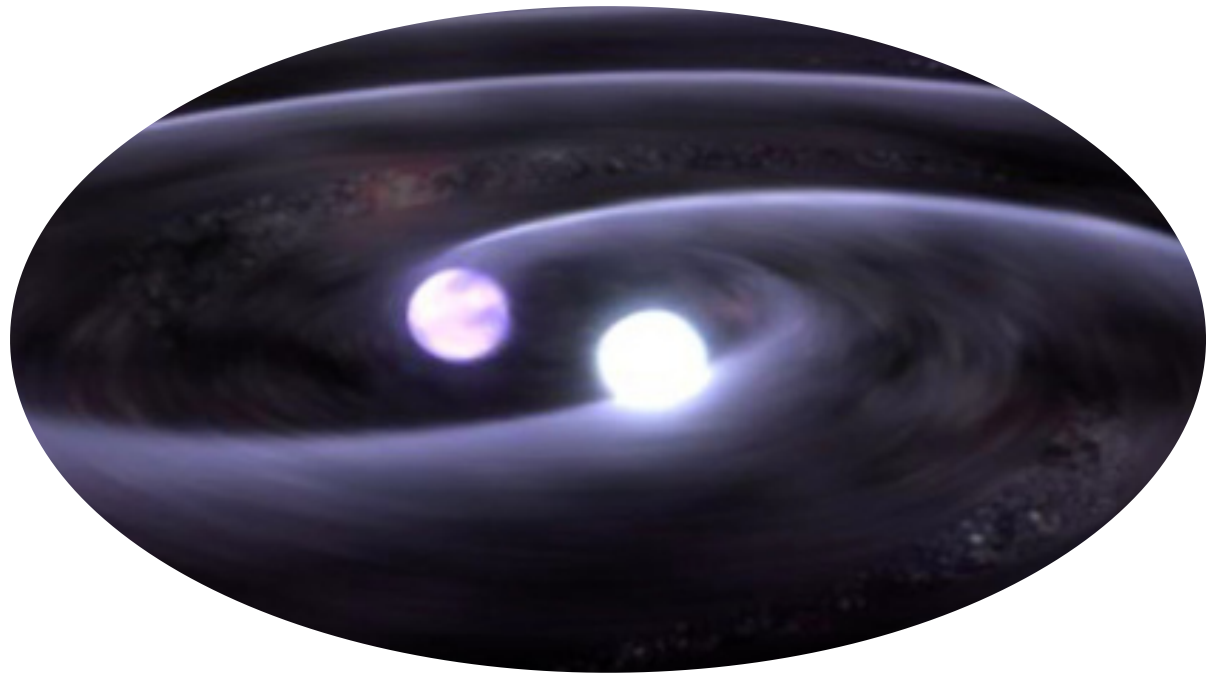
Imperial College London & SISSA DATA SCIENCE

PROGENITORS

CO white dwarf accreting mass.



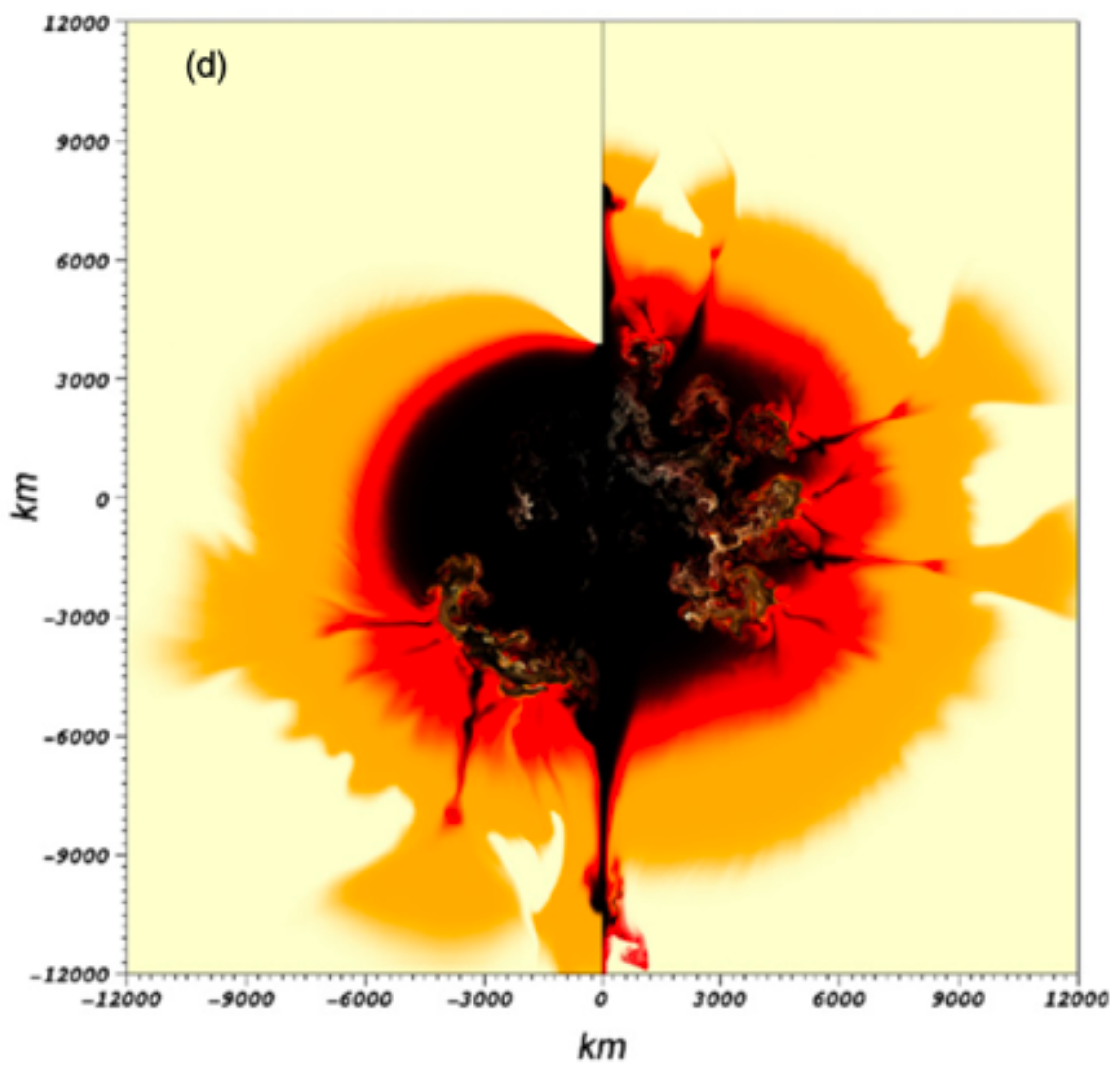
Single-degenerate
(e.g. Hosseinzaden et al, 2017)



Double-degenerate
(e.g. Roche & Garnavich, 2020)

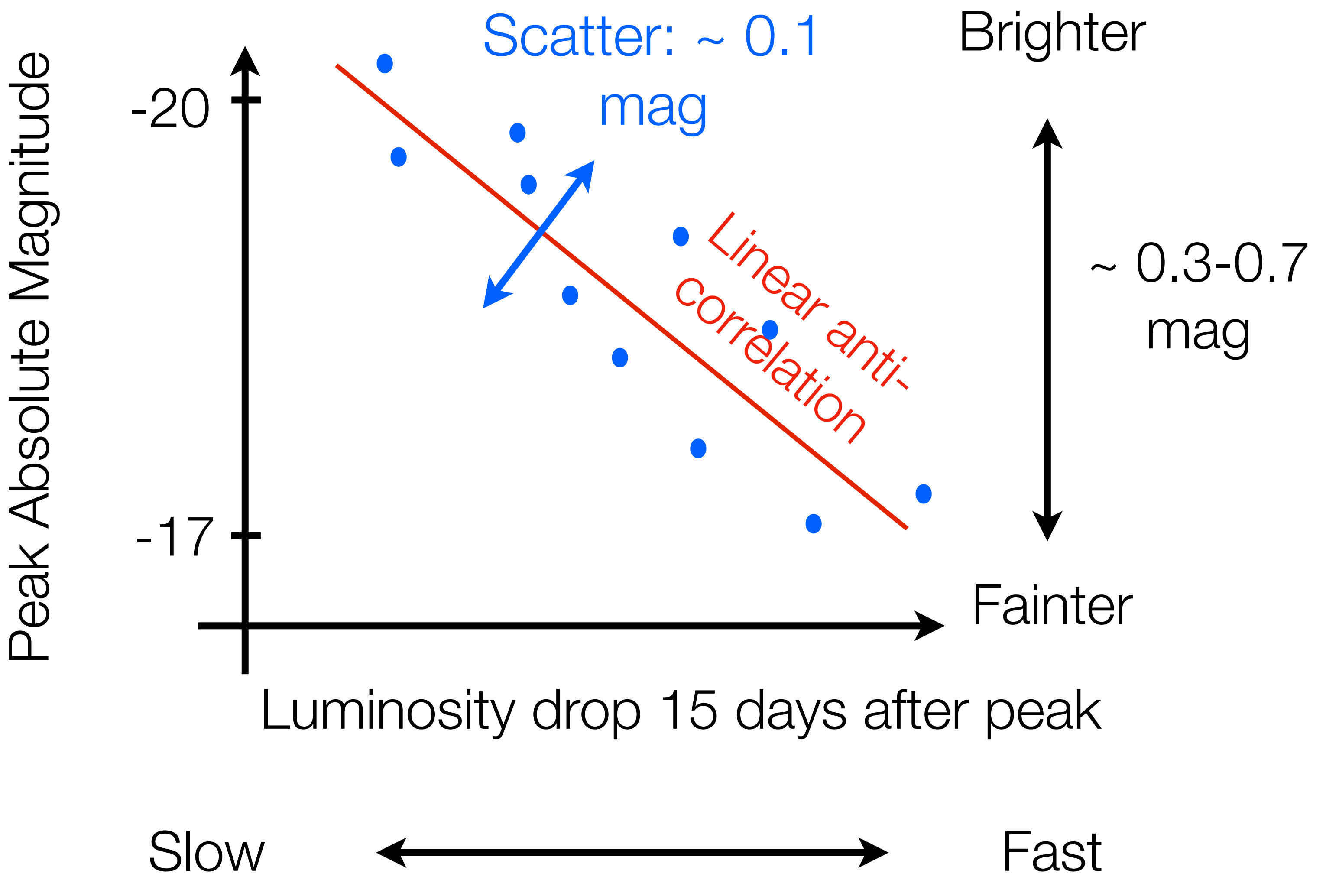
EXPLOSION

C detonation creates thermonuclear explosion.



Kruger et al (2012)

LUMINOSITY



Lightcurve powered by radioactive decay of ^{56}Ni.

Higher core density ⇨
Larger mass of ^{56}Ni & IGEs ⇨
Higher luminosity & opacity ⇨
SNIa brighter, slower to fade

Phillips, ApJ 413 (1993) L105-L108

Standardization of SNIas

"Brighter SNIa are slow decliners"

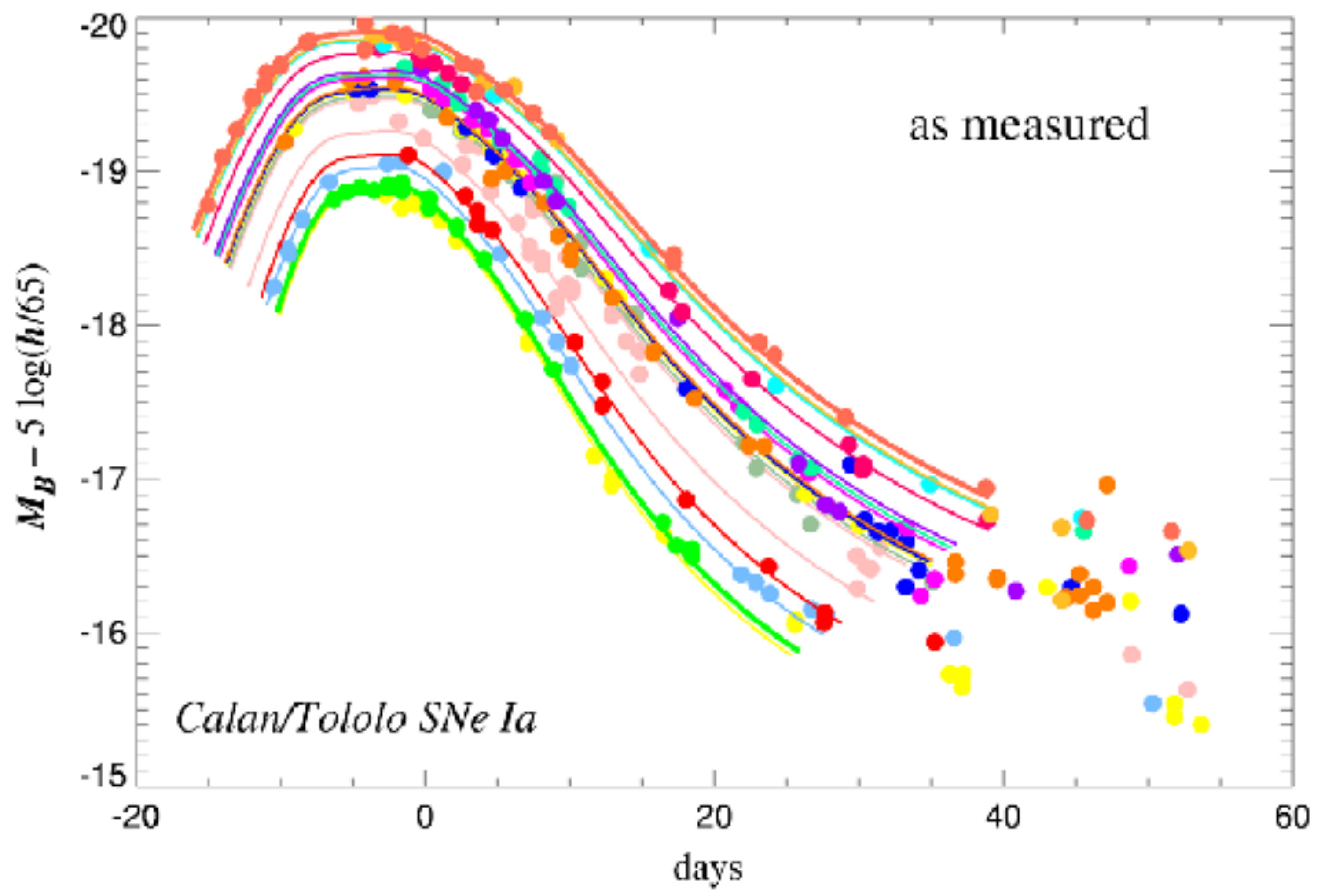
Use the empirical 2D linear correlations between absolute magnitude and “stretch” and colour to standardise SNIas to within ~0.1 mag residual dispersion at peak

BEFORE CORRECTION

Absolute B-band magnitude

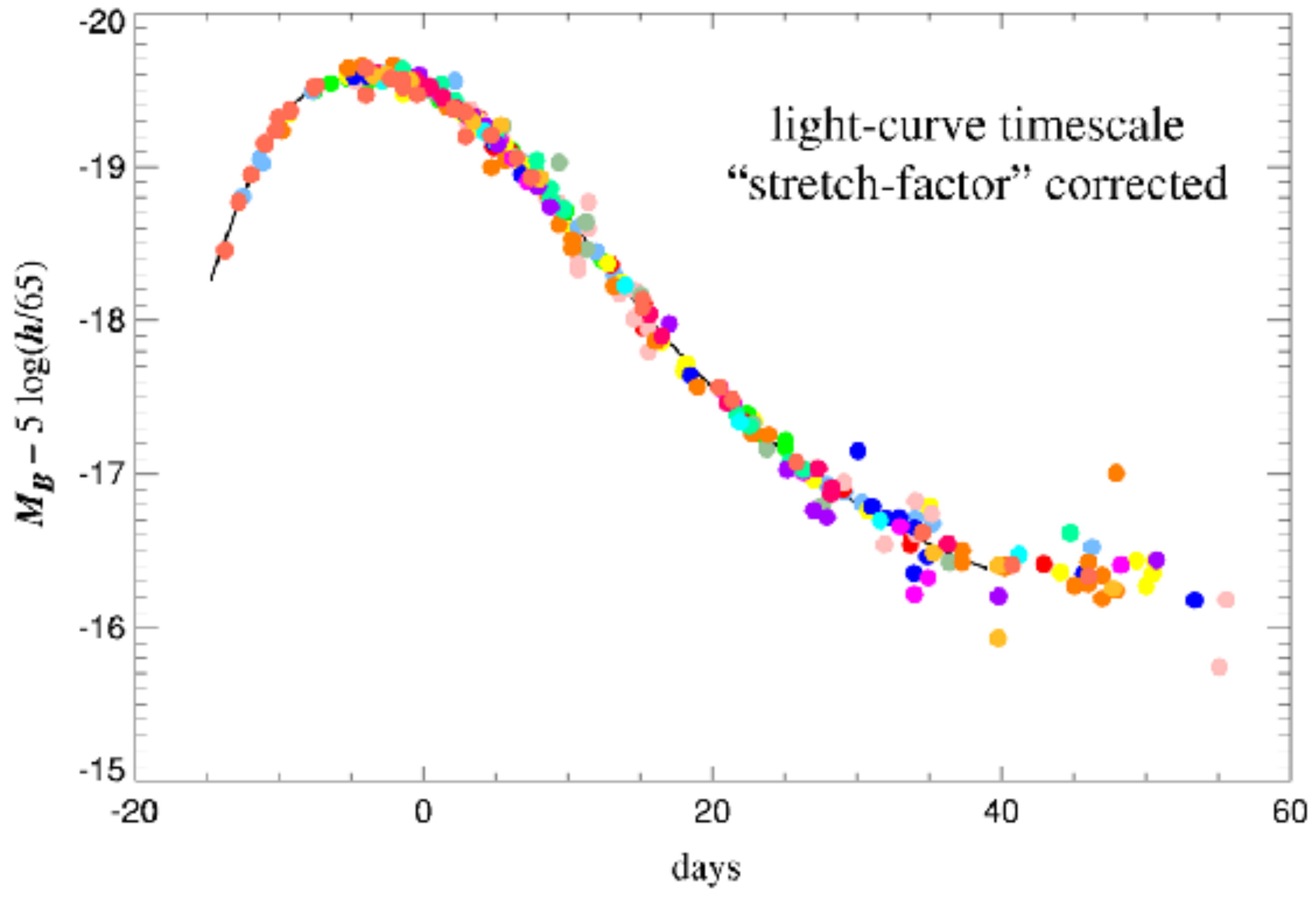
BRIGHTER

FAINTER



AFTER CORRECTION

Absolute B-band magnitude



Kim et al (2007)

Constraining Cosmological Parameters

DISTANCE-REDSHIFT RELATION

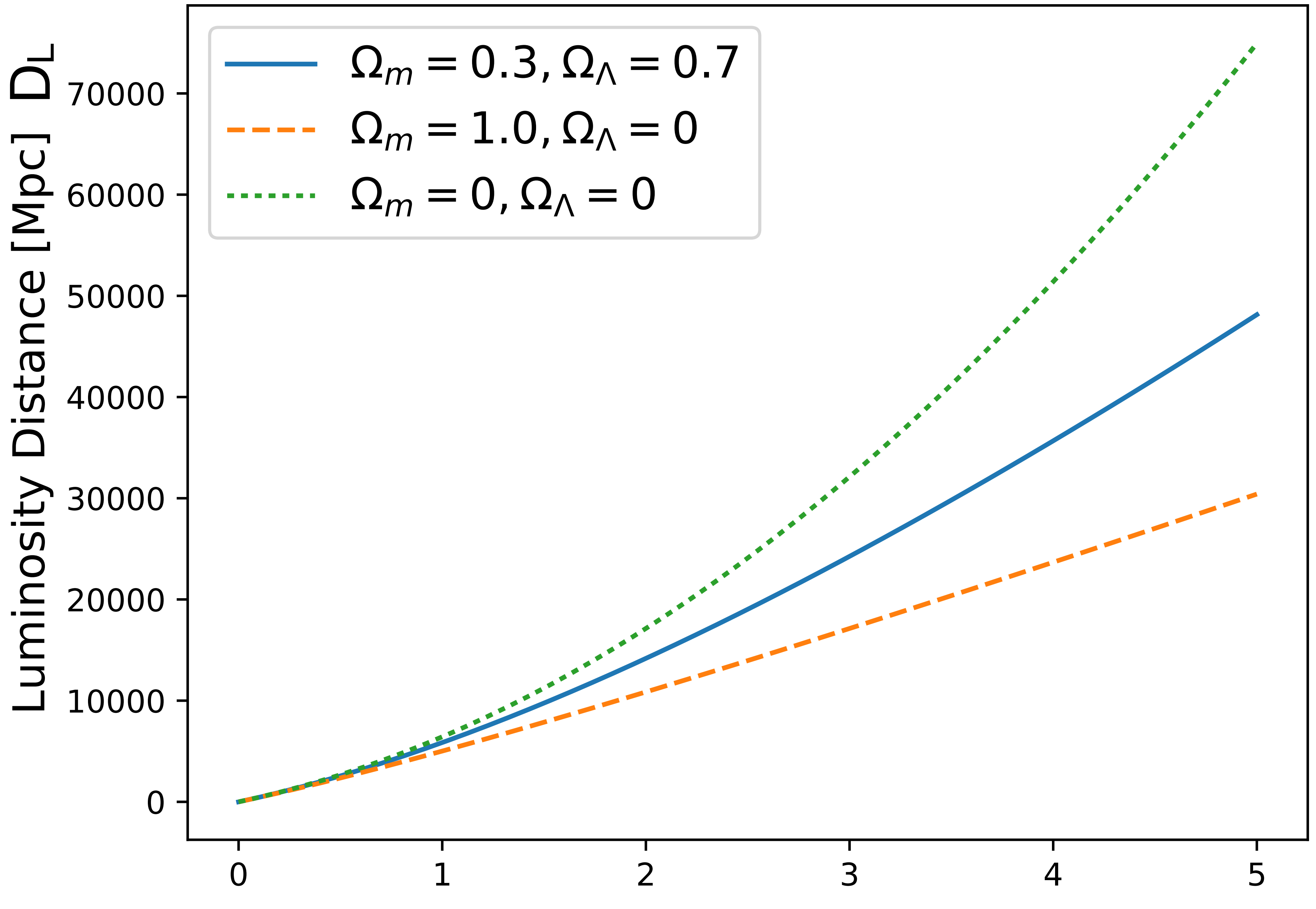
Measure redshift (z) and distance modulus, μ:

$$\mu = m_B - M + \alpha x_1 - \beta c$$

Apparent magnitude

Absolute magnitude

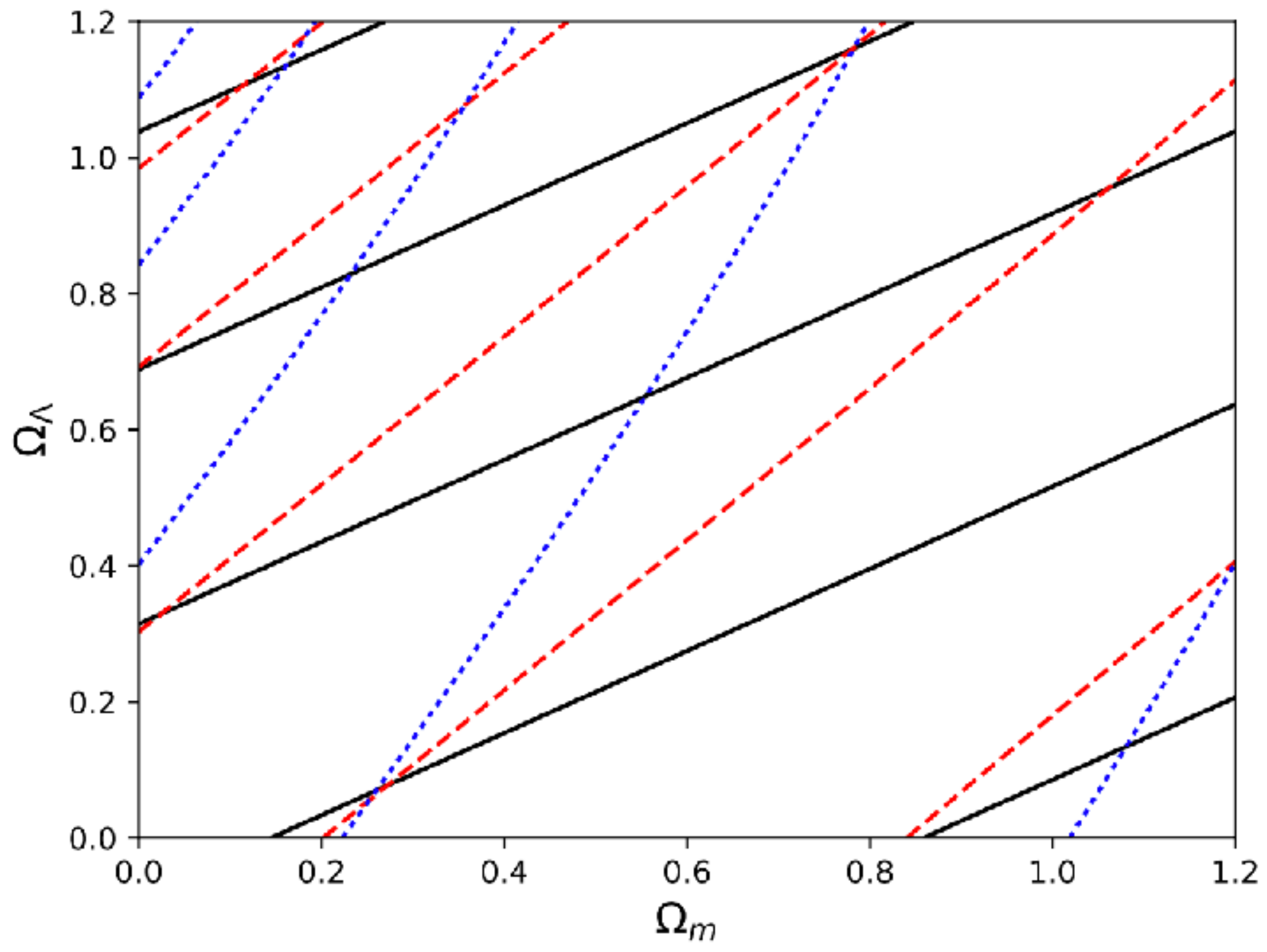
Linear standardization



$$\mu = 5\log_{10}\frac{D_L}{\text{Mpc}} + 25$$

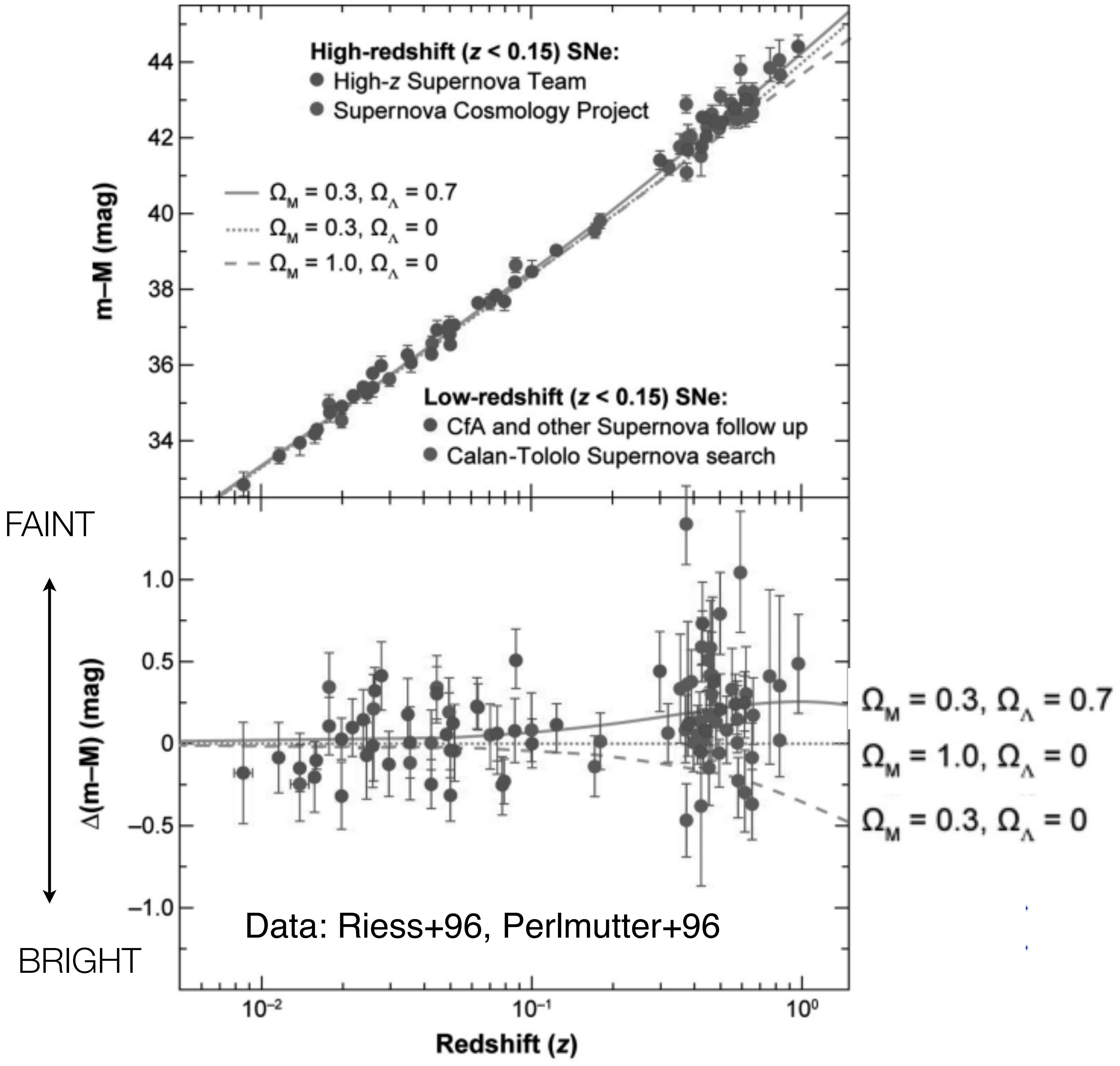
DIFFERENTIAL DISTANCE MEASUREMENT

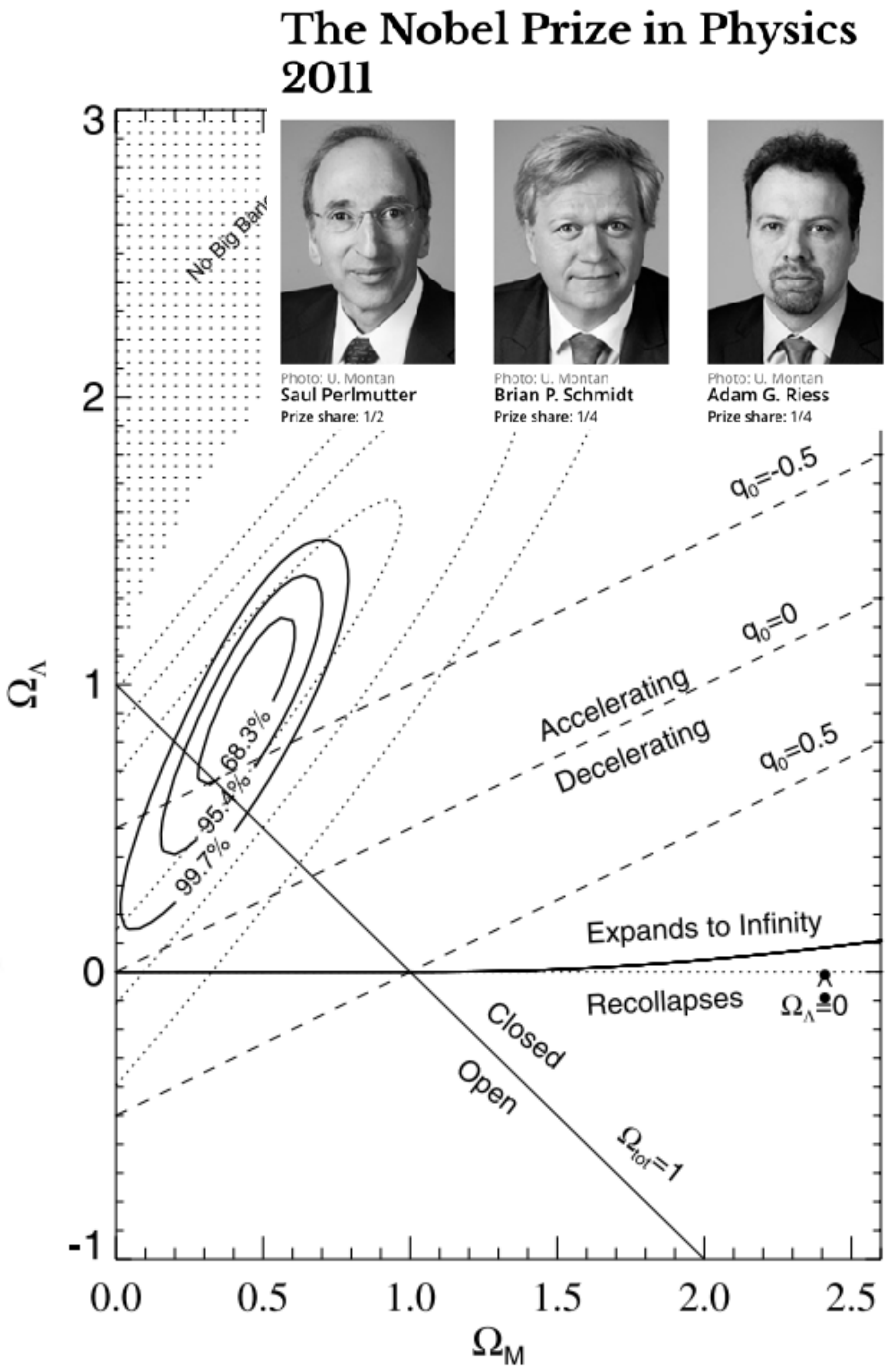
$z = 0.1$ $z = 0.5$ $z = 1.0$



Contours of constant D_L at various redshifts

COLLECT THE NOBEL PRIZE

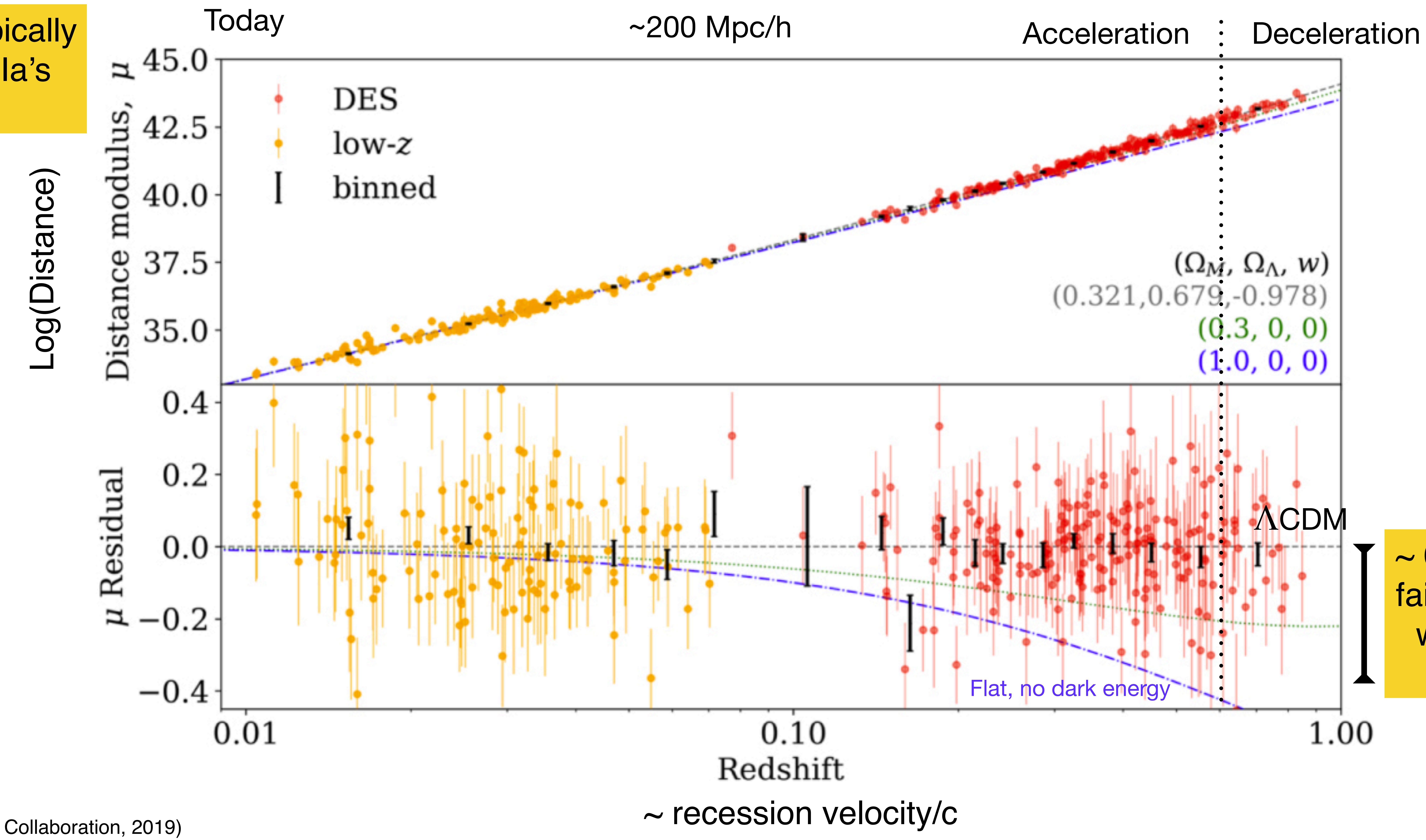




Riess et al, ApJ, 607:665-687 (2004)

Distance-Redshift Relation Measurement

Spectroscopically confirmed Ia's only



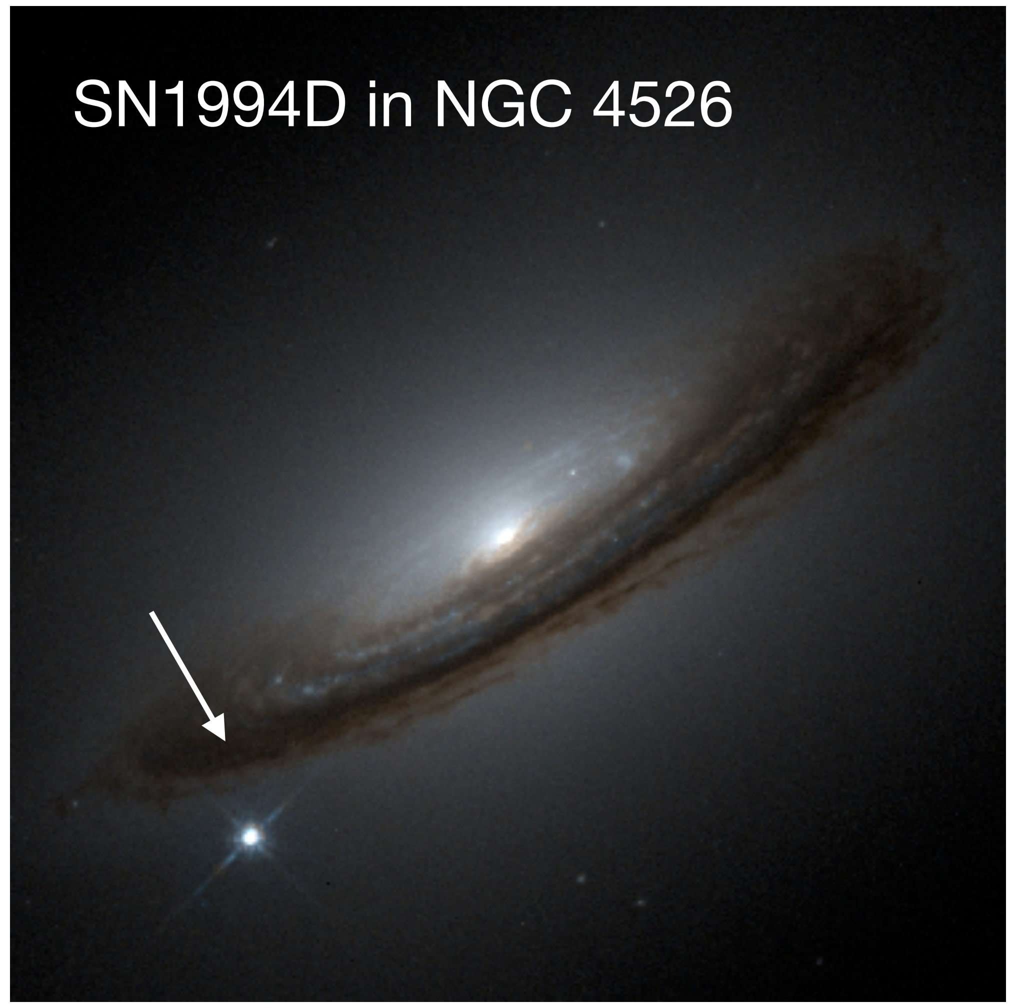
~ 0.25 mag fainter than w/o dark energy

Abbott et al (DES Collaboration, 2019)

Supernovae Type Ia: Observations

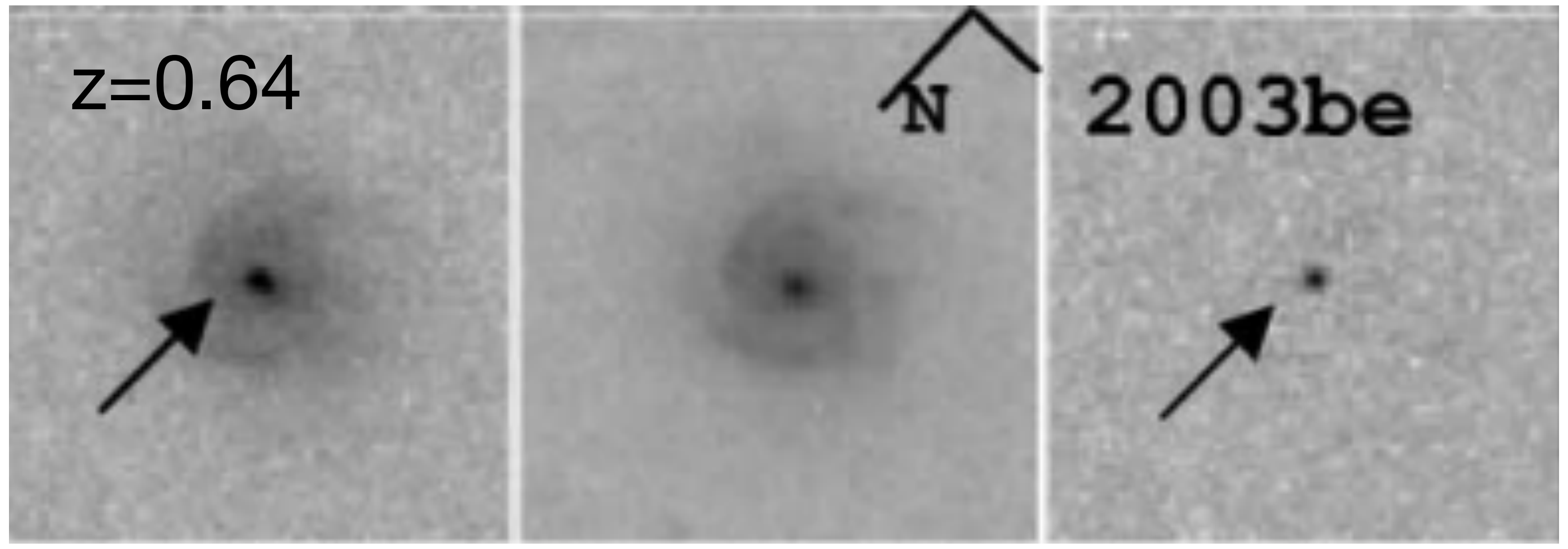
Imperial College London & SISSA DATASCIENCE

DETECTION



Treffers et al (1994); imaged by HST

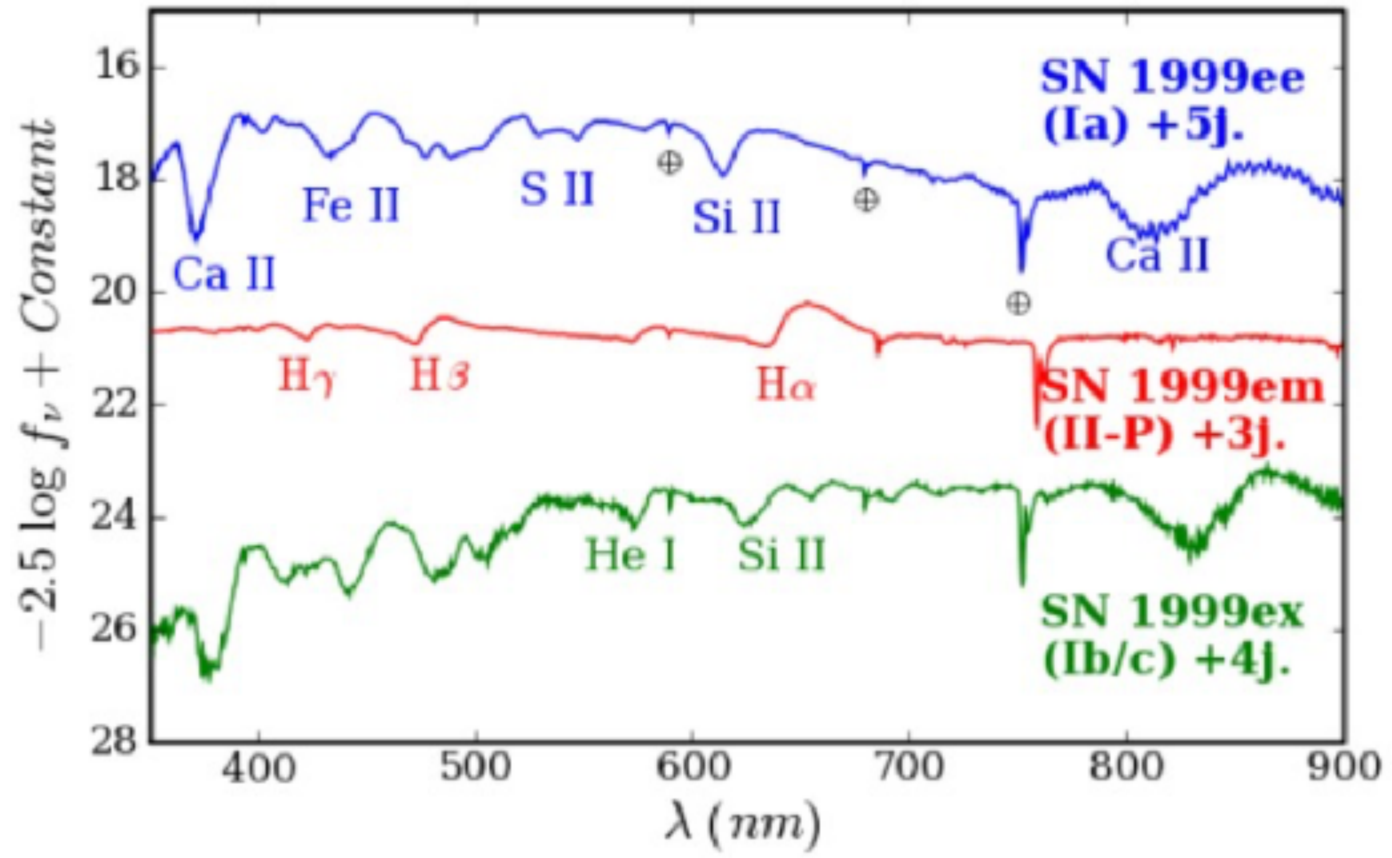
During After Subtracted



Riess et al, ApJ, 607:665-687 (2004)

TYPING

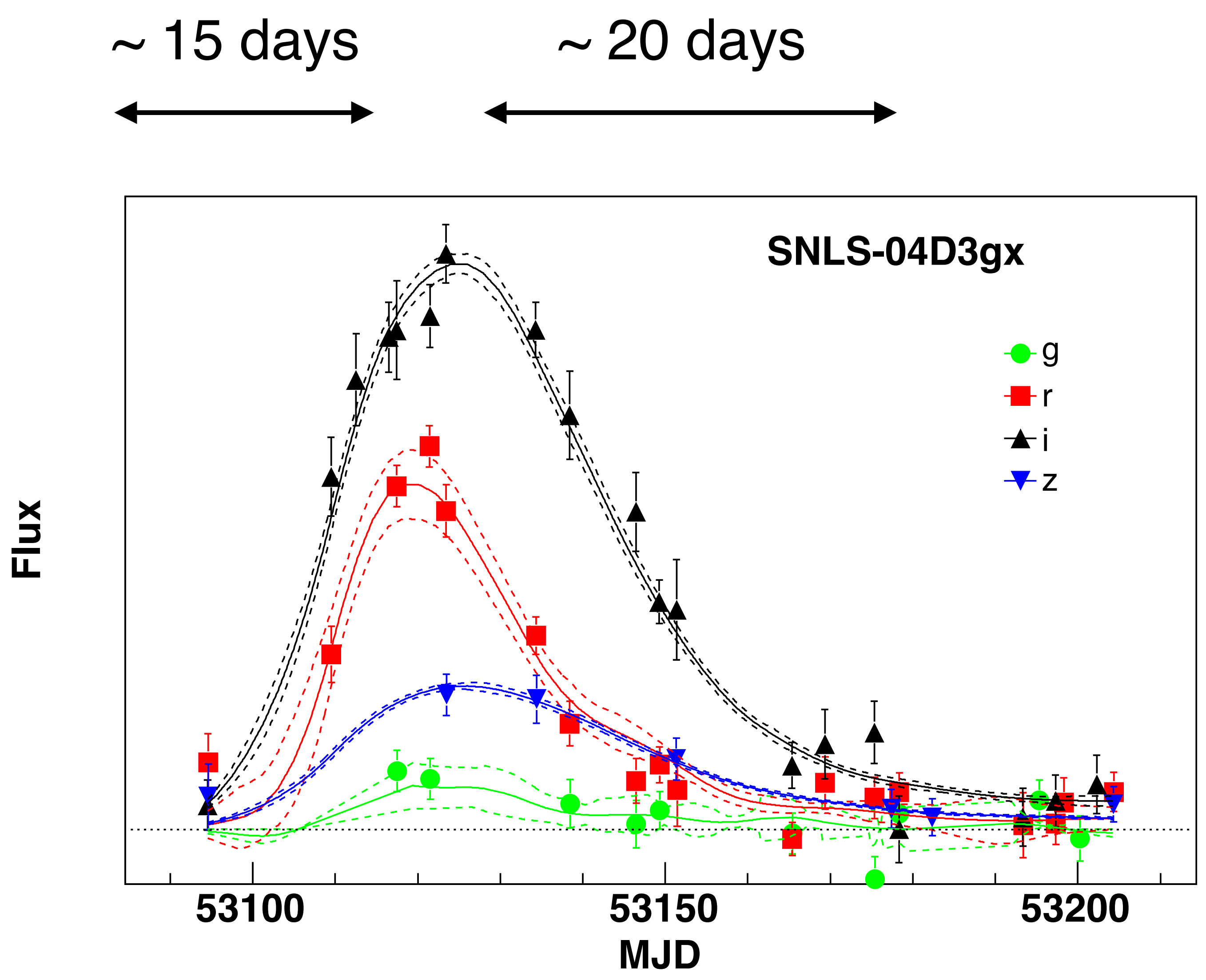
Spectra: Type Ia have no H lines



Credit: Julien Guy

Confirming Ia is easy with spectra.
Much harder with just photometry (i.e., low-res spectral information): only probabilistic classification

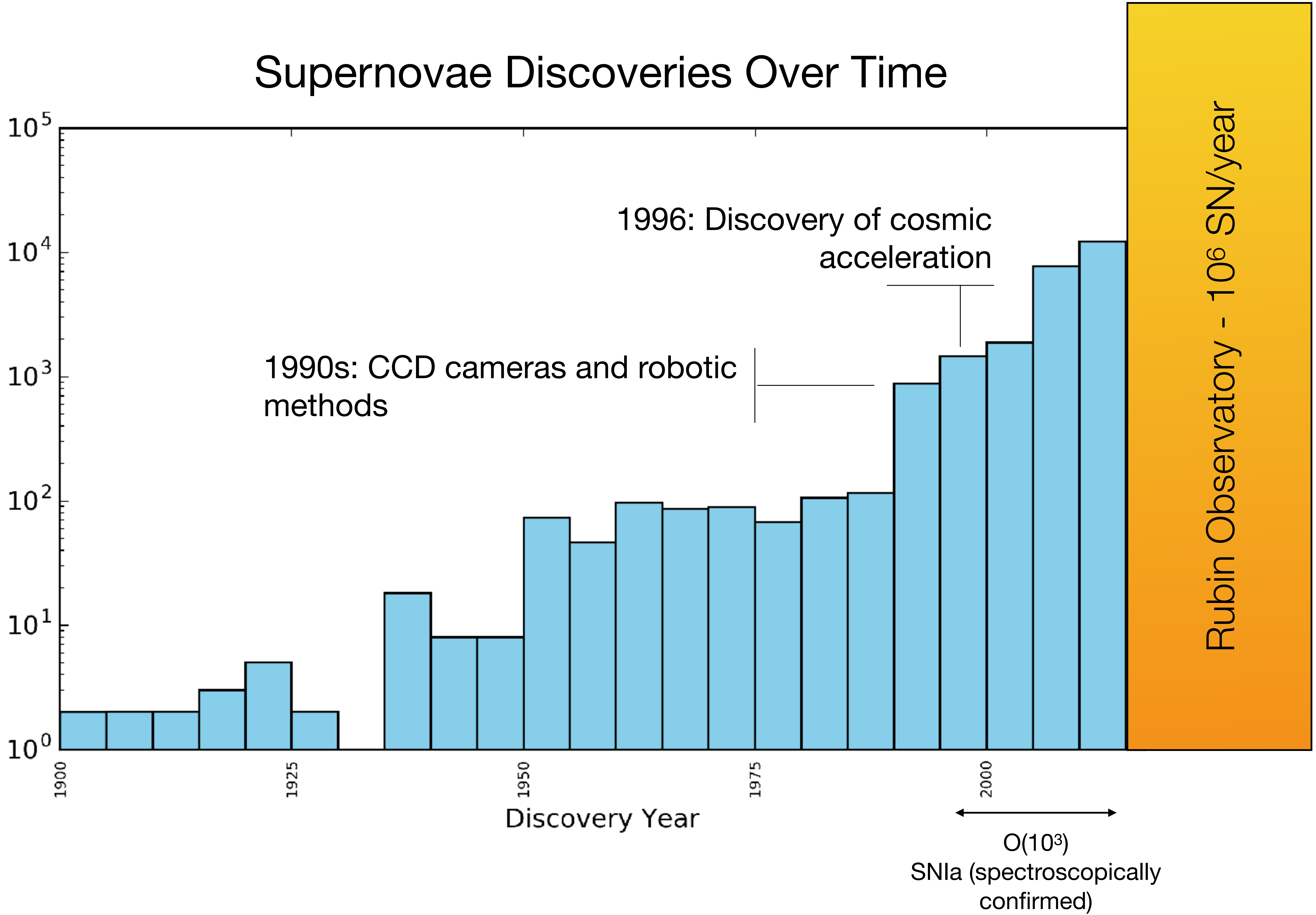
FOLLOW-UP



Guy et al (2007)

Lightcurves: time-evolution of brightness in several colour filters

Used for standardization and cosmological inference


Supernovae Discoveries Over Time
10^5
10^4
10^3
10^2
10^1
10^0
1996: Discovery of cosmic acceleration
1990s: CCD cameras and robotic methods
Rubin Observatory - 10^6 SN/year
1900
1925
1950
1975
2000
Discovery Year
O(10^3)
SNIa (spectroscopically confirmed)

Bayesian Hierarchical Modelling of SNIa data

For more details, see:

BHM: March, RT et al (2011)

Unity: Rubin et al (2015)

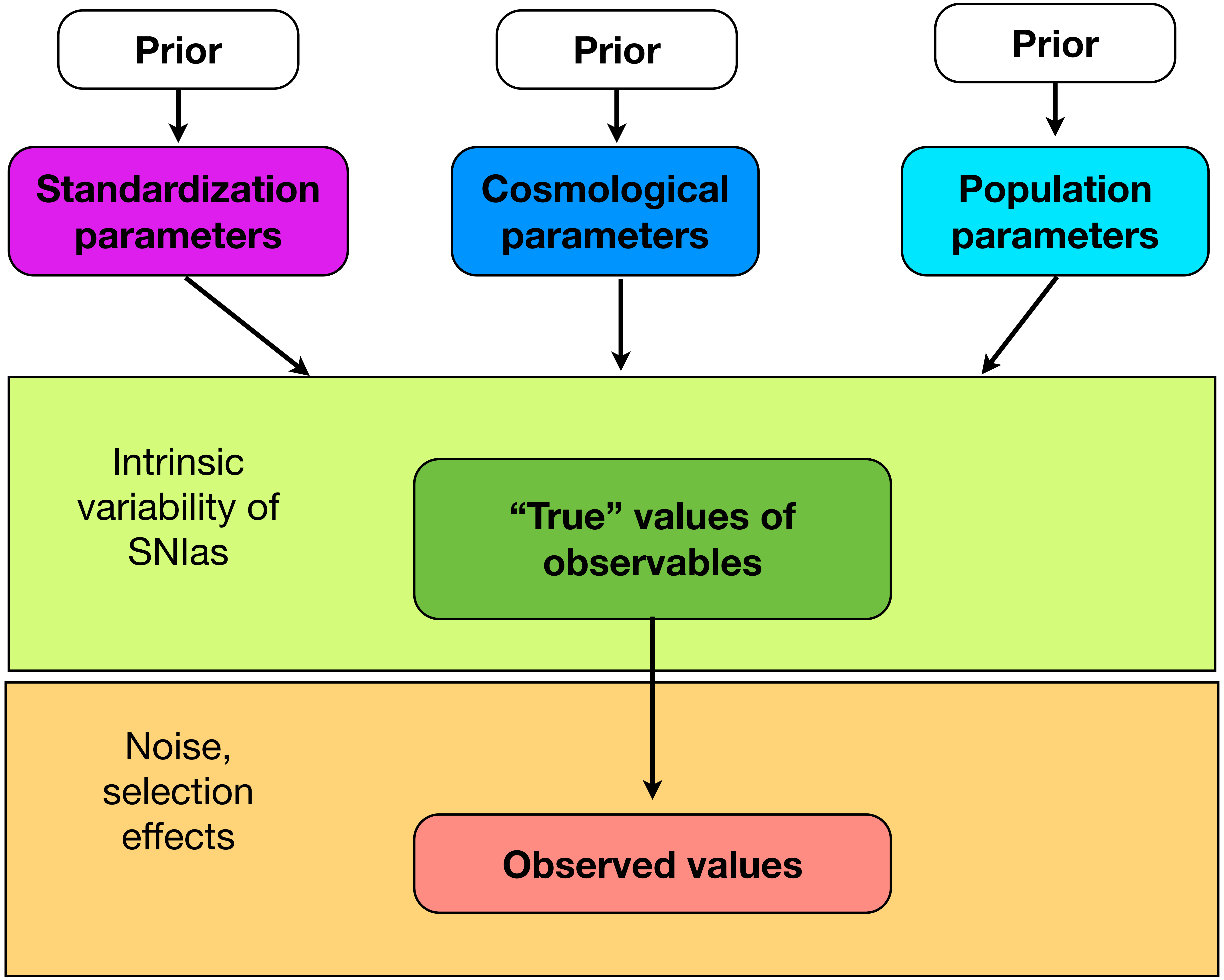
BAHAMAS: Shariff, RT et al (2016); Rahman, RT et al (2022)

Simple-BayeSN: Mandel et al (2017)

Steve: Hinton et al (2019)

Upcoming:

MALFOI: Karchev, RT & Weniger (soon)



SNIa population distributions
Environmental properties (age, metallicity, SFR, host type)

Latent variables:

unknown and unobserved, are integrated out during the inference

Noisy data subject to truncation

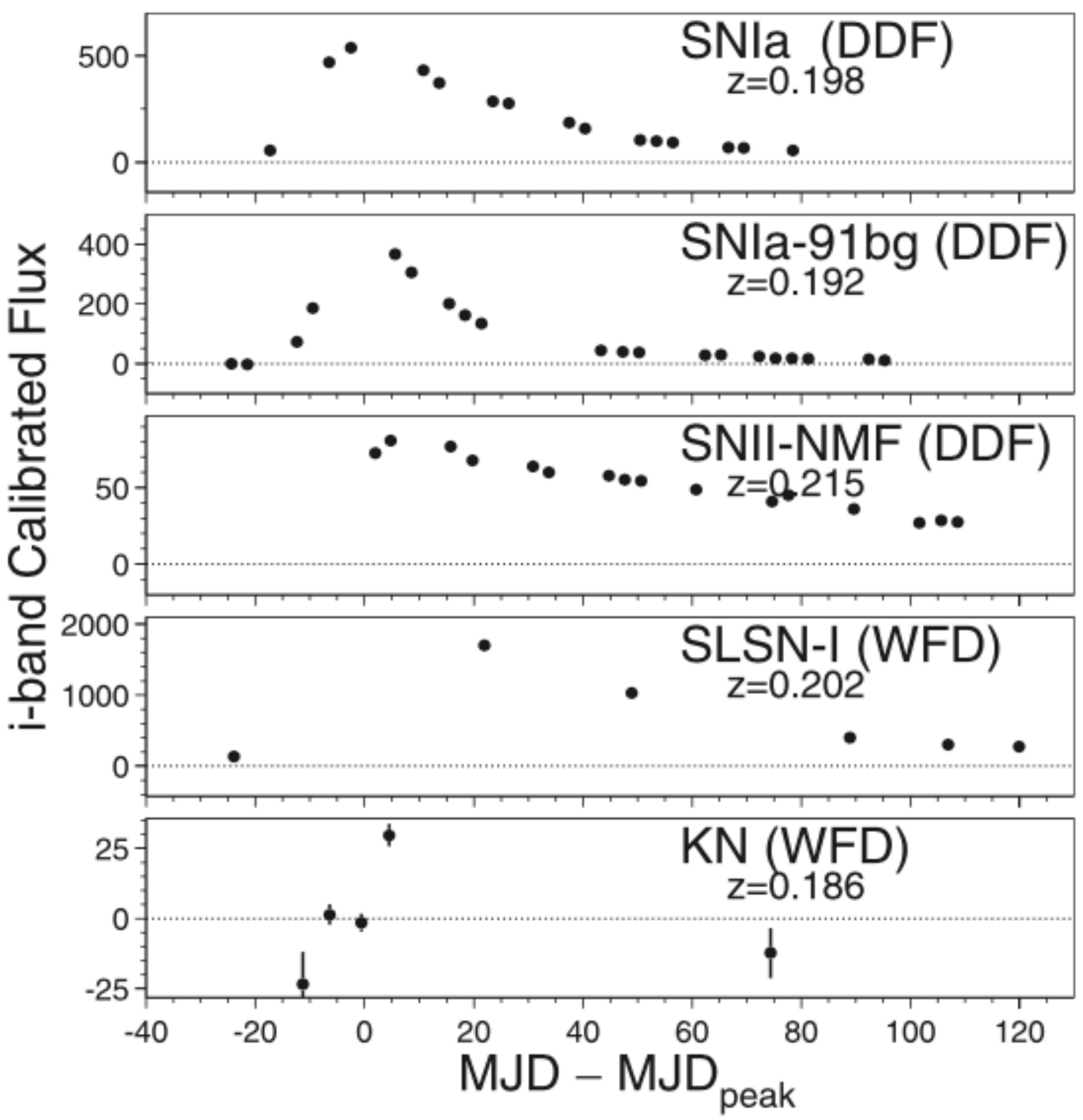
The Problem:

We want to classify Type Ia vs non-Ia **reliably** and **efficiently** from light-curve data alone.

BUT:

Spectroscopic training set is non-representative.

Simulated light-curves



Kessler et al (2019)

Classification challenges:

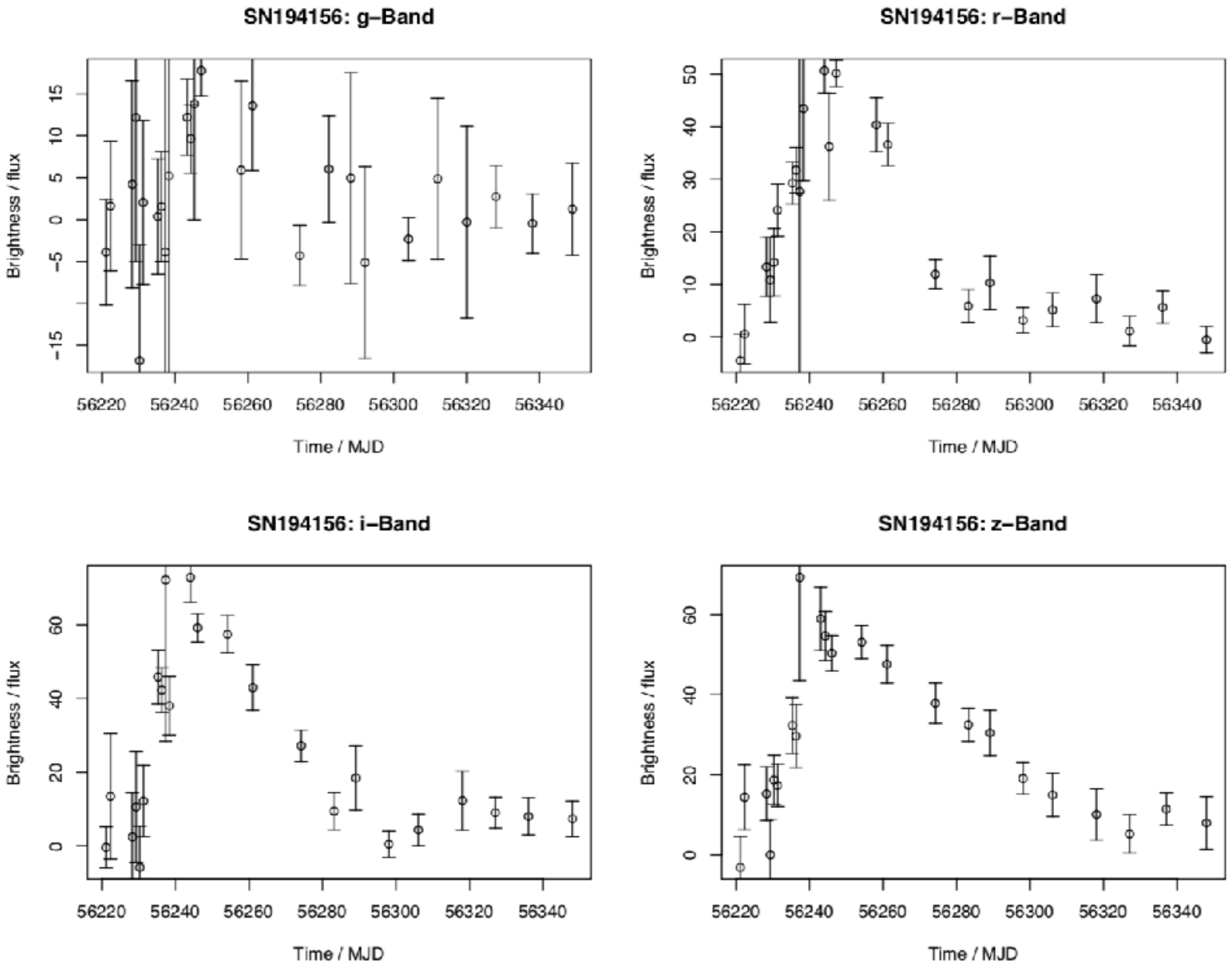
The Photometric LSST Astronomical Time-series Classification Challenge PLAsTiCC (Kessler et al, 2019)

Supernova Photometric Classification Challenge (Kessler et al, 2010)

The Future: Photometric SNIa Cosmology

- SNIa identification relies on observationally expensive spectroscopy
- In the future, we won't have spectra for all SNIa candidates (DES: 3,000 SNIa over 5 yrs; LSST: 10,000 SNIa/yr)
- SNIa classification needed from multi-band imaging alone
  - "SN Classification Challenge" (Kessler+10)
  - more recently: LSST Kaggle comp (2018)

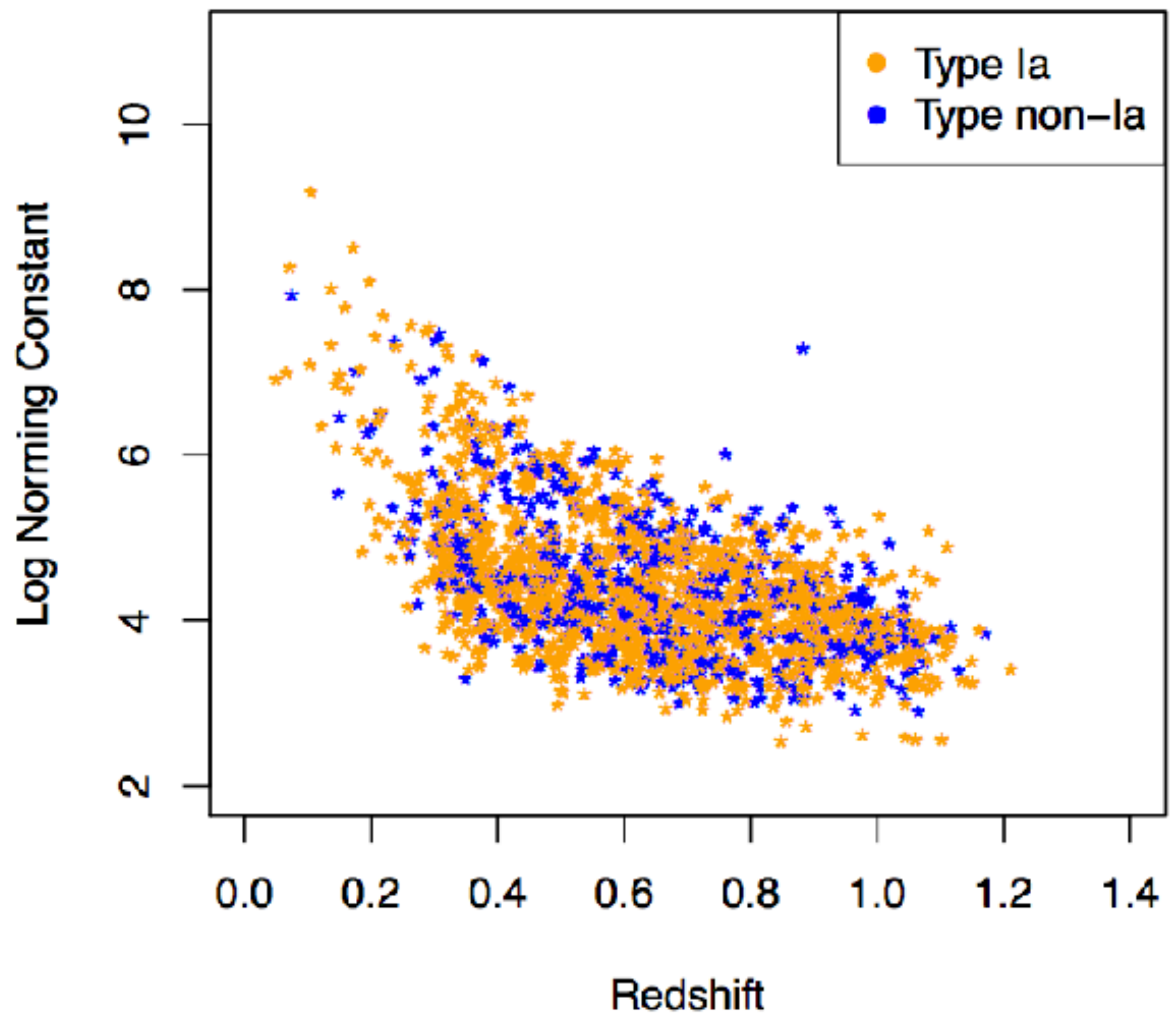
Example of simulated LC data



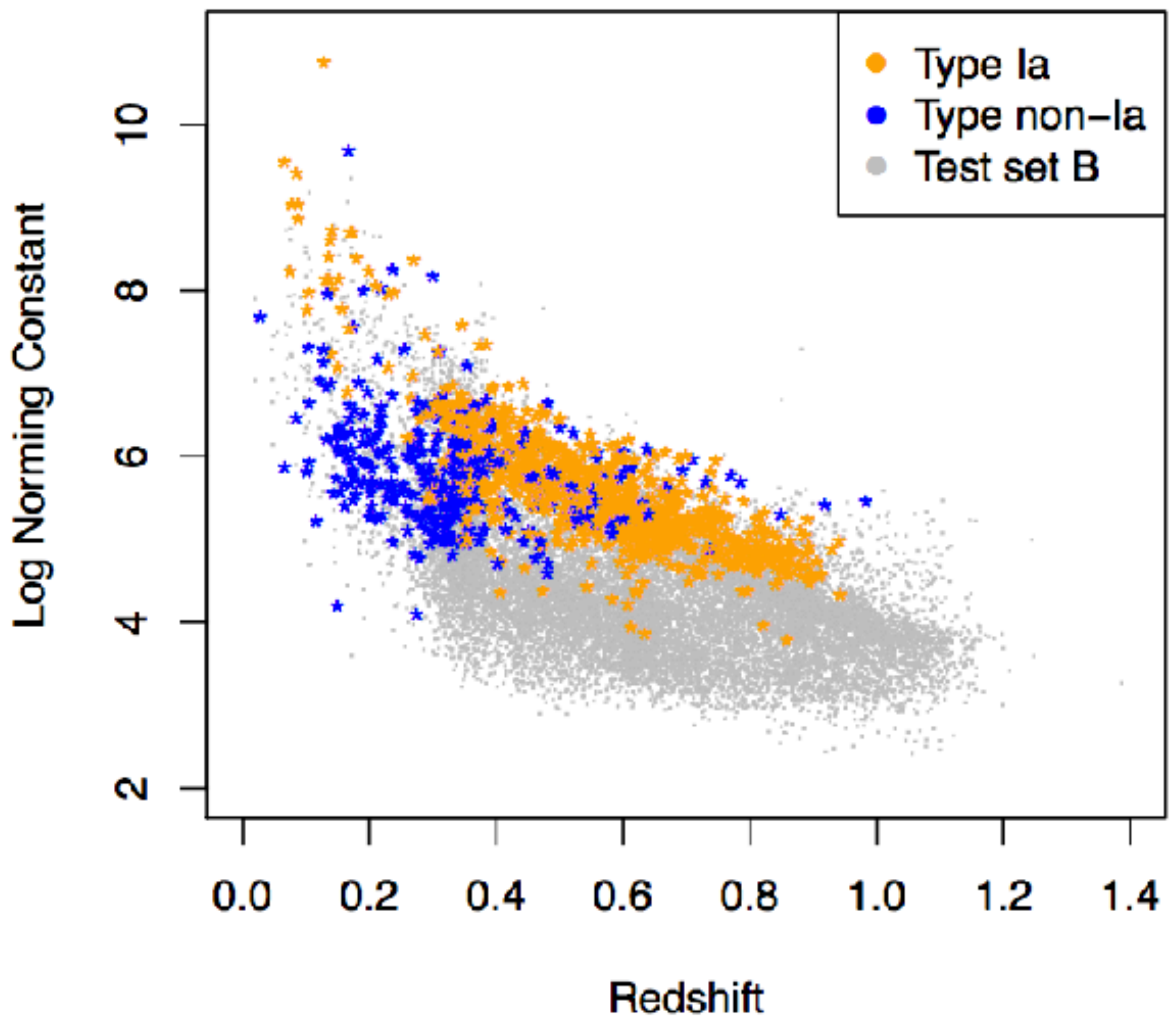
The Future: Photometric SNIa Cosmology

- **Problem:** Training set is biased. Especially at high *z*, more SNIa's than in the population, hence non representative

Random training data (unavailable)

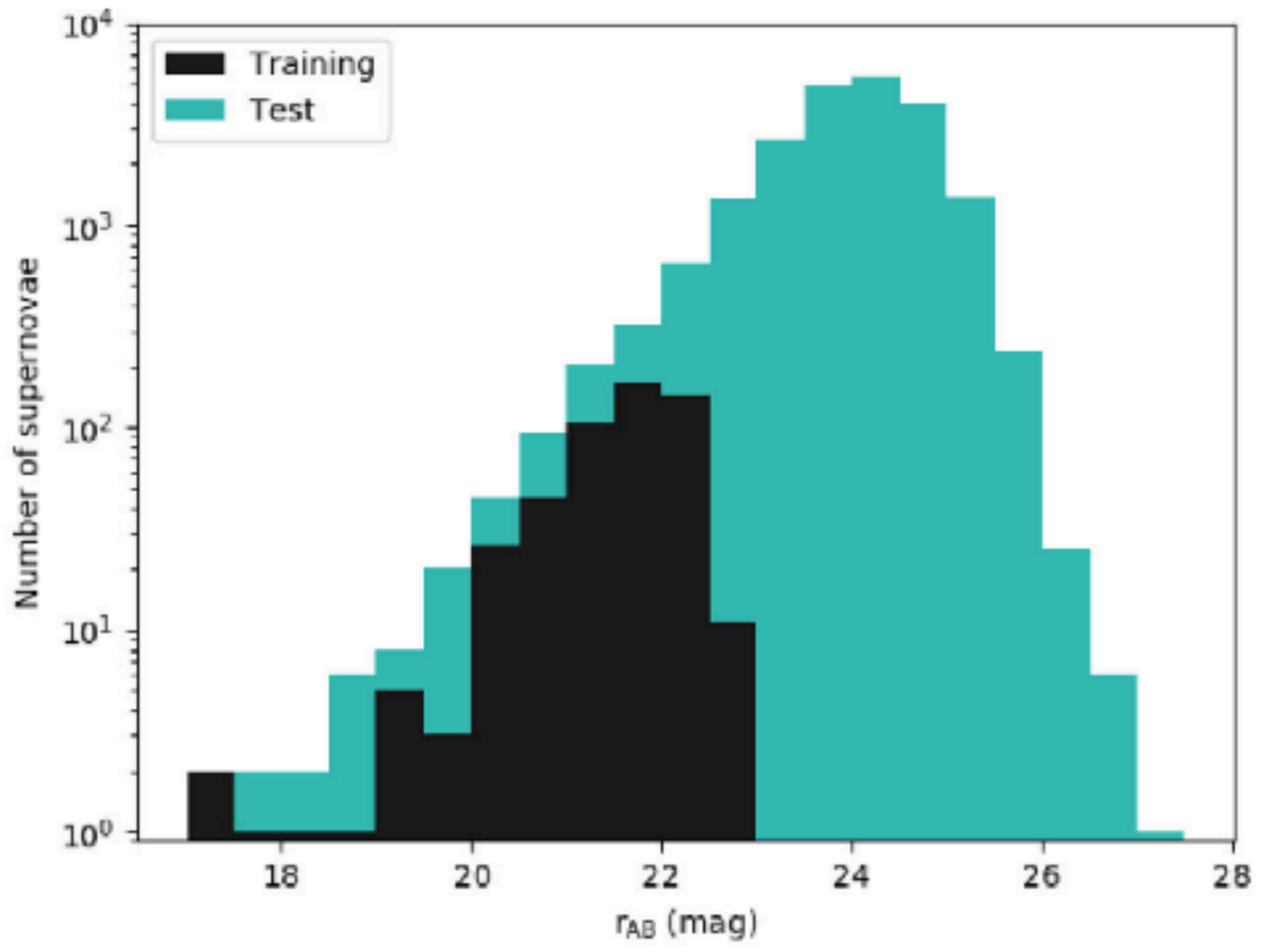


Biased training data (colour) vs Test Set (gray)

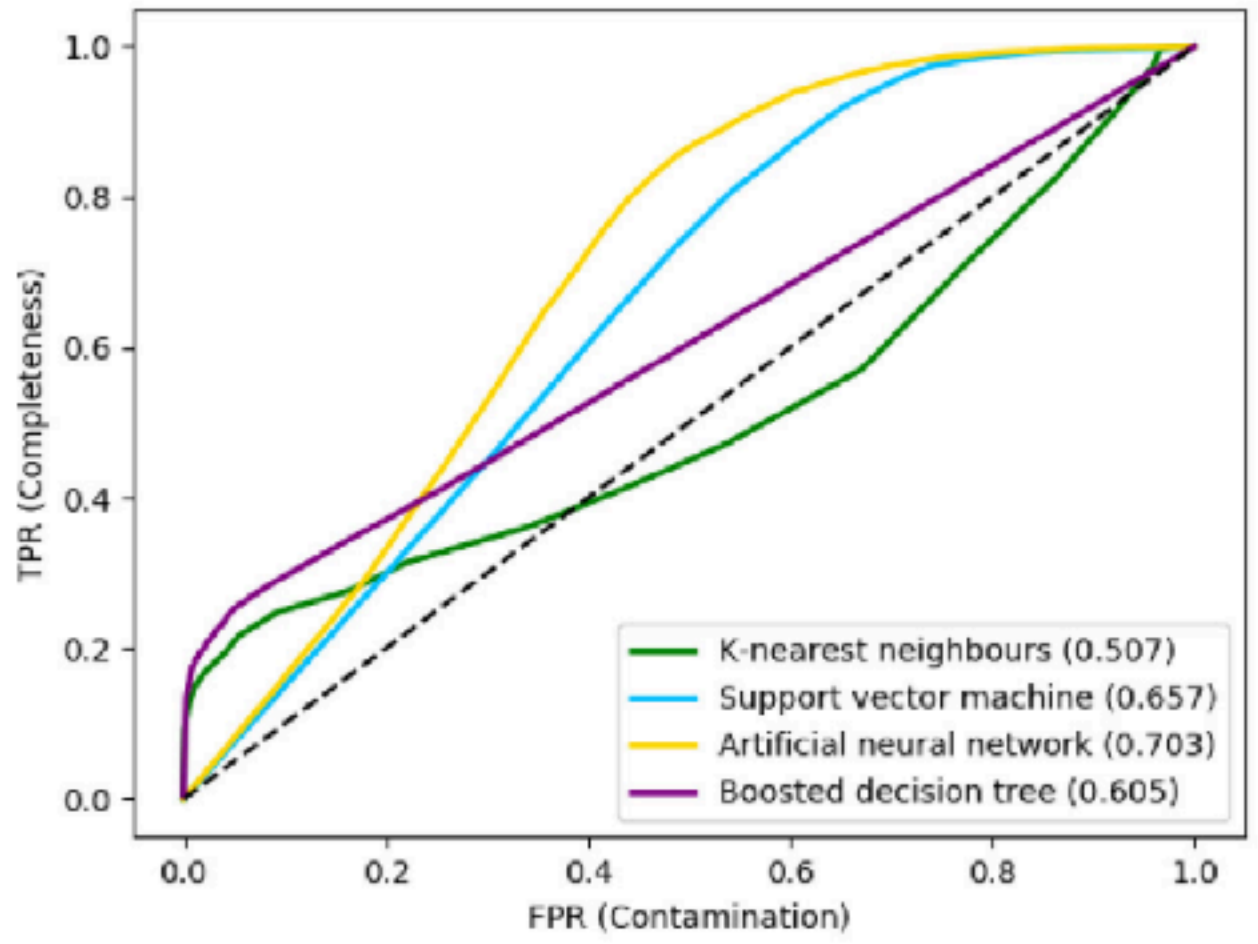


Solution 1: data augmentation

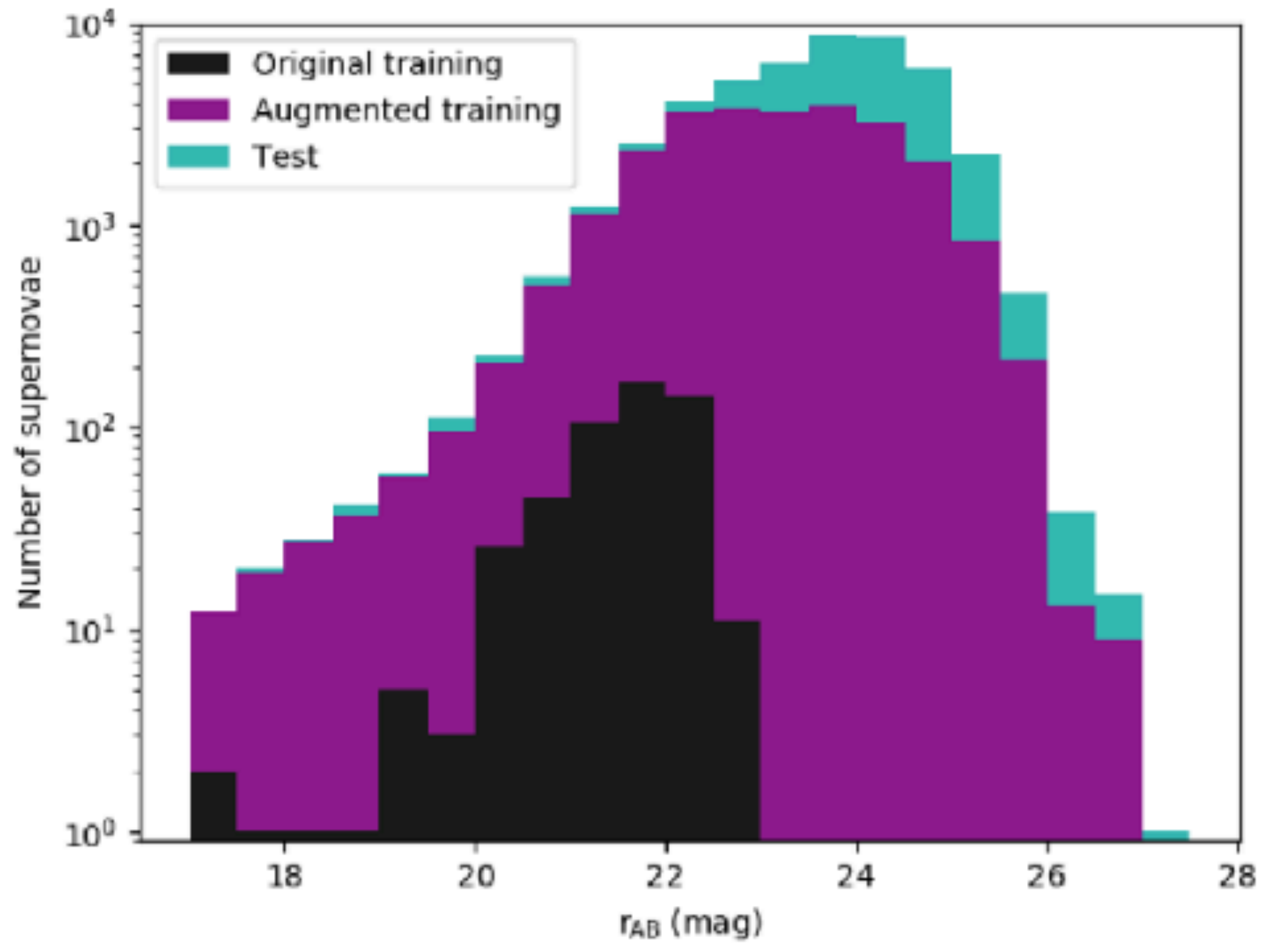
Unrepresentative spectroscopic training set leads to poor classification performance



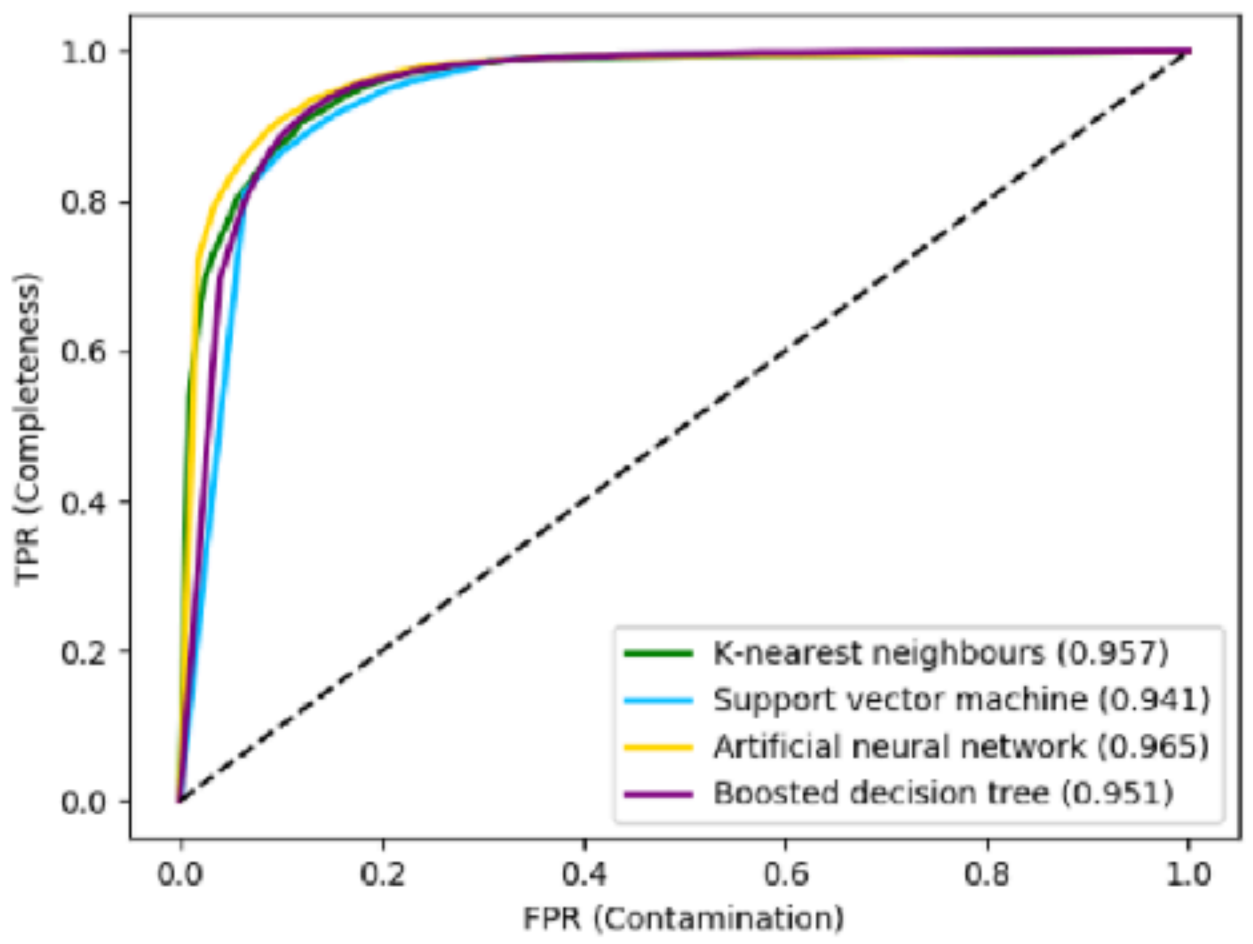
(a)



+ 50-fold data augmentation improves massively the AUC



(a)



However:

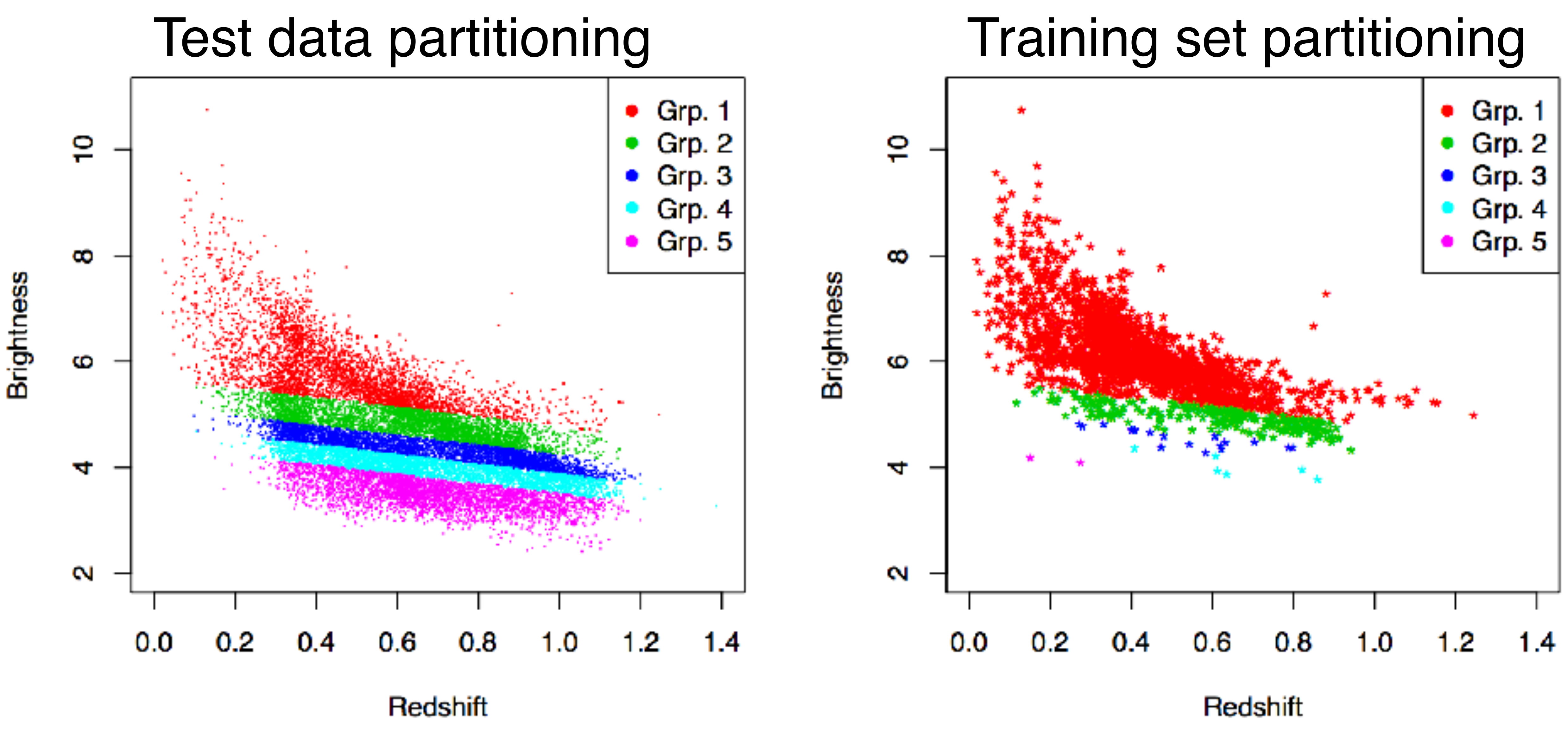
Succesfull augmentation relies on a good light-curve model (e.g. GPs) AND the assumption of no astrophysical evolution with *z*

Carrick et al (2021MNRAS.508….1C)

A first solution: STACCATO

- Our first solution (Revsbech, RT, van Dyk, MNRAS, 473, 3, 3969 (2018), arxiv:1706.03811): **S**yn**T**hetically **A**ugmented Light **C**urve **C**lassific**AT**i**O**n (STACCATO) approach
  - Fit light curve with Gaussian Process (GP)
  - Compute Diffusion Map (to quantify similarities between LCs), Richards+12
  - Perform Random Forest Classification on diffusion coordinates
  - **New:** Group SNs according to **Propensity Score** (probability of belonging to the training set): bias within groups reduced by 90% (Rosenbaum & Rubin, 84)

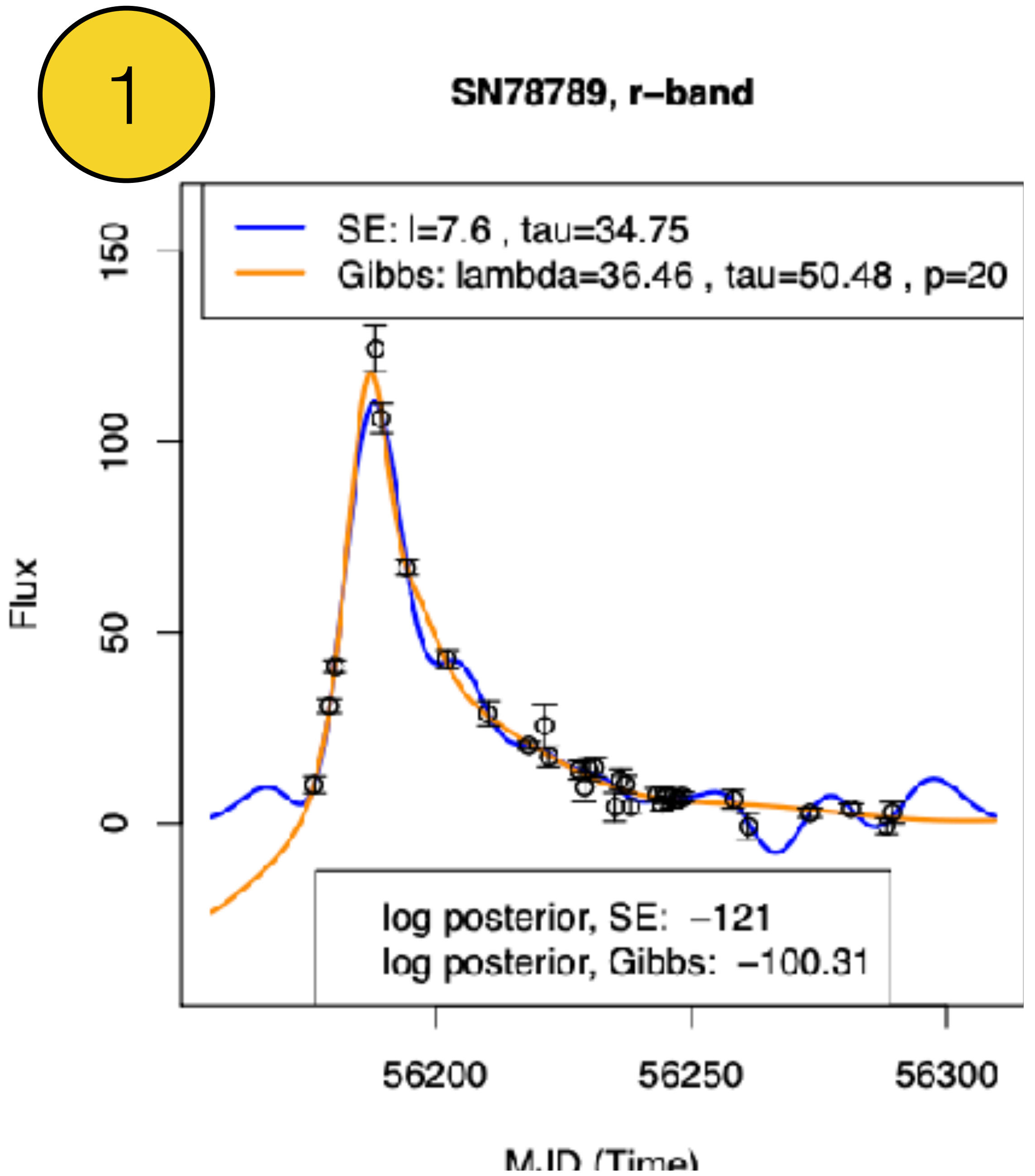
Propensity Score Grouping:

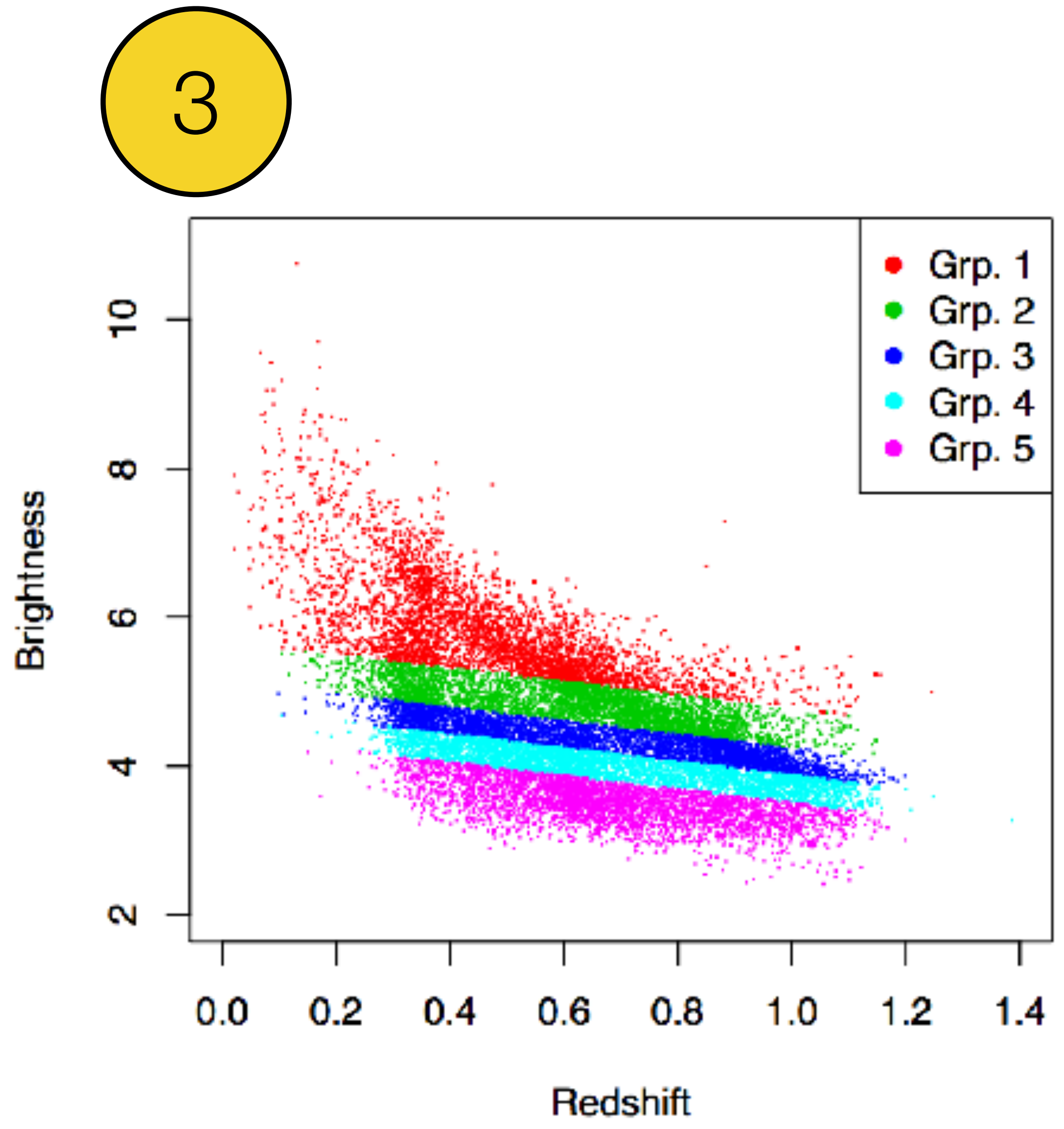


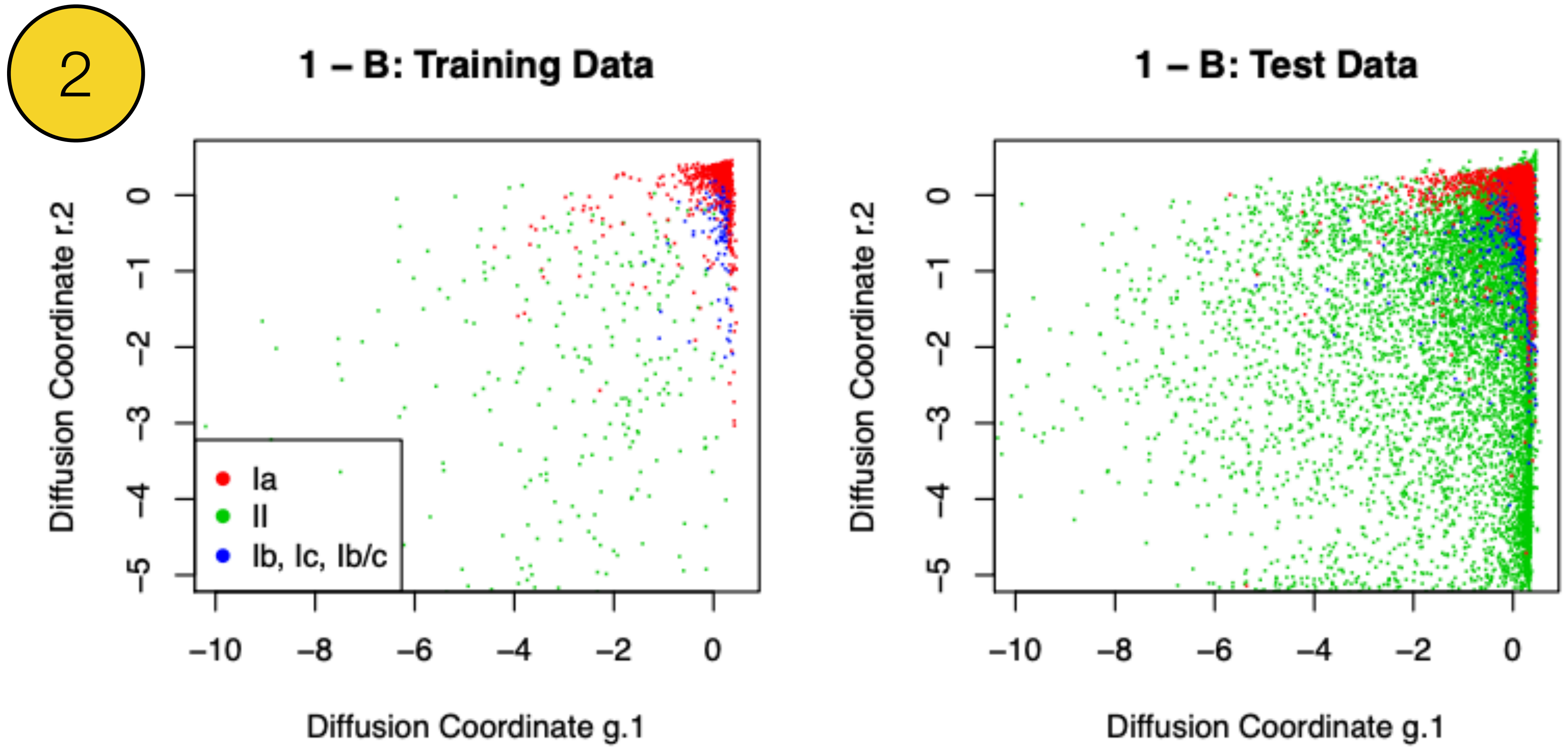
STACCATO: Analysis Details

Imperial College London & SISSA DATA SCIENCE

1. Fit Gaussian Process (GP) to light curve data
2. Apply diffusion map technique to map the fitted light curve into a covariate space of dim~O(100), following Richards et al (2012)
3. Stratify light curve data according to propensity scores quantiles
4. Sample new synthetic light curves from fitted GPs in groups with sparse training data (requires validation set to optimise augmentation scheme)
5. Use Random Forest on the stratified diffusion coordinates (group-by-group) for classification



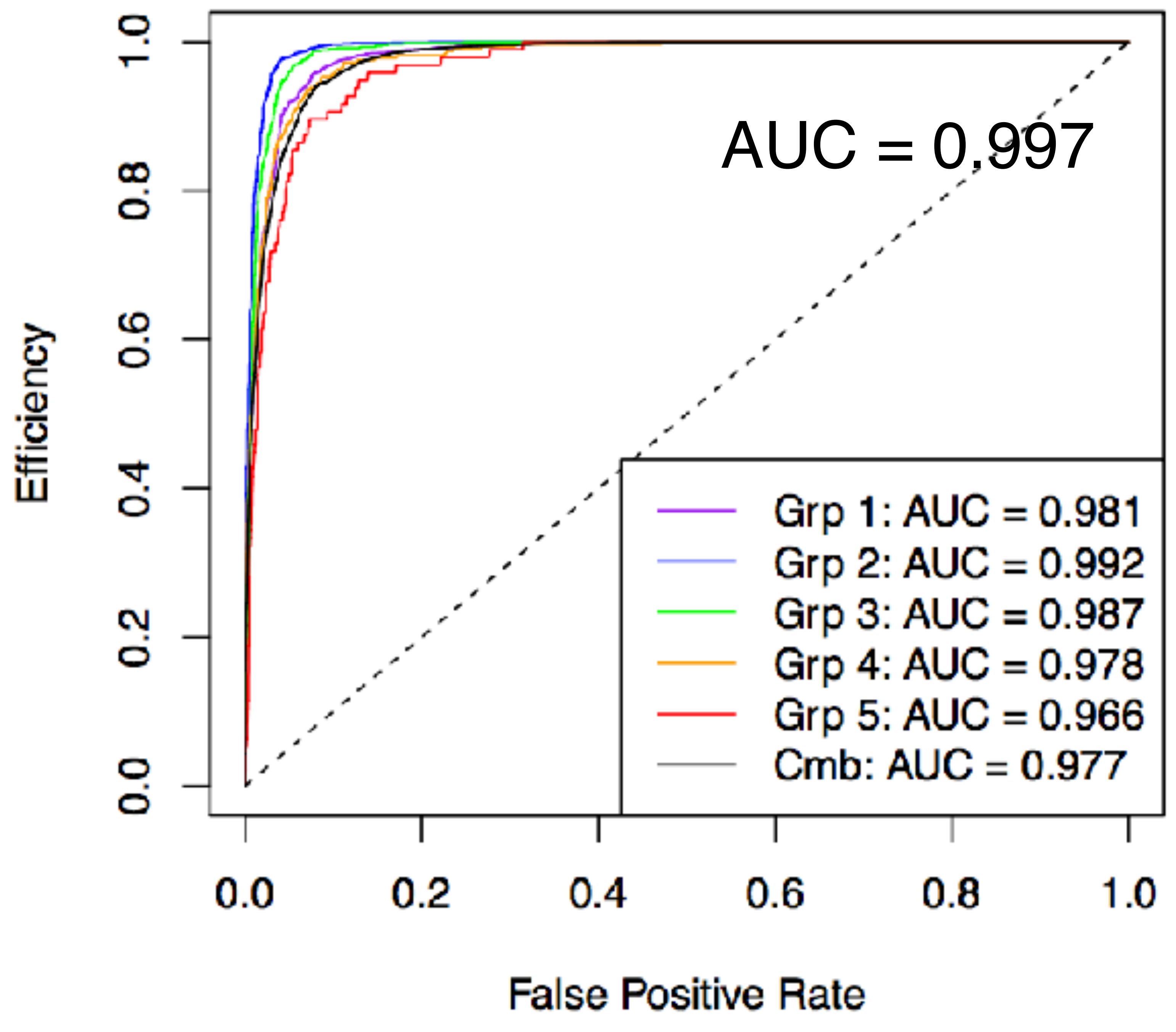




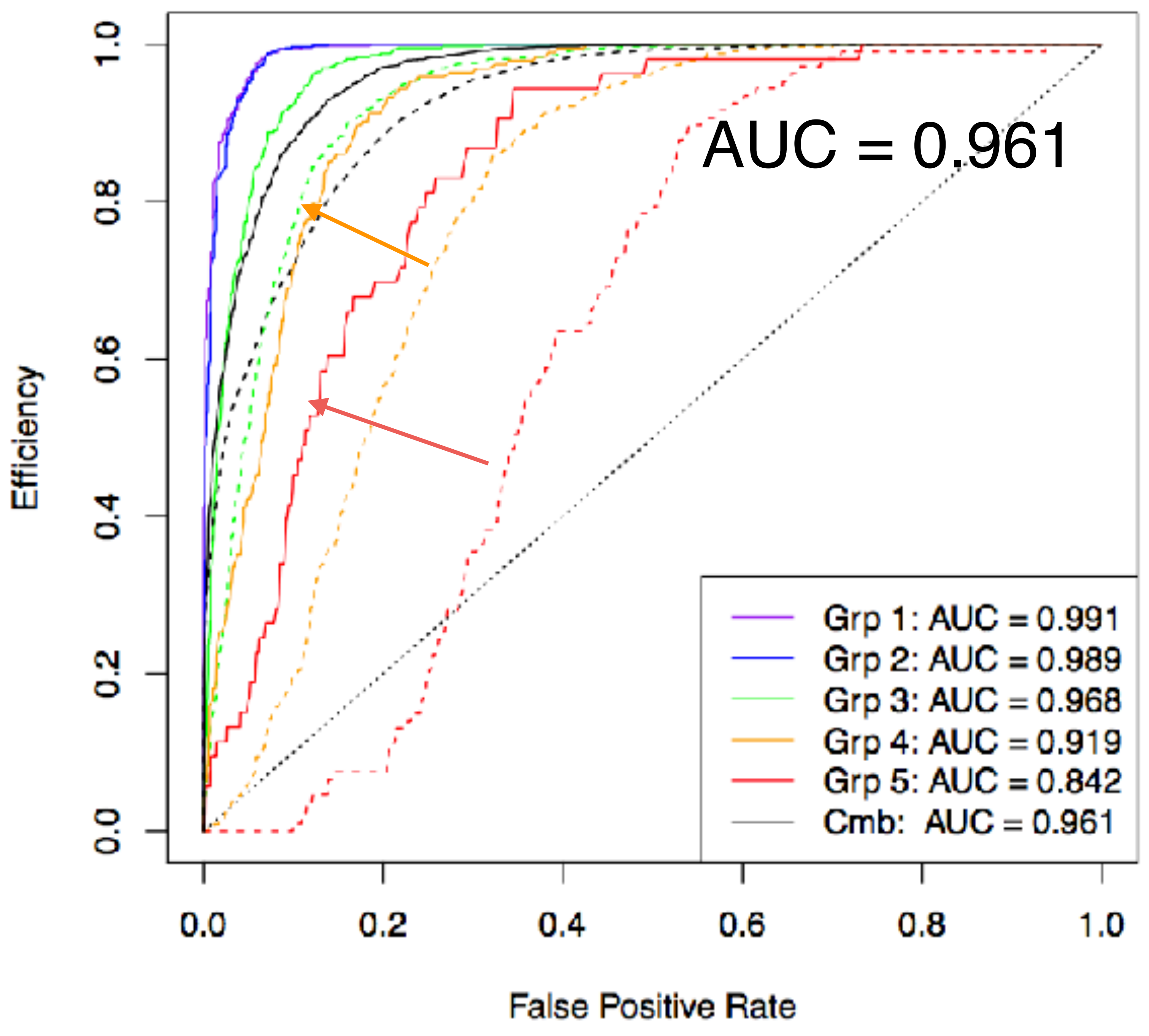
Augmenting LCs via GP Resampling

- STACCATO augments the training set by synthetically generating LCs from the GP **according to the Propensity Scores -> corrects under-sampling where it matters**
- Evaluated using Area under the ROC Curve (AUC)

Gold Standard
(unbiased training set)



STACCATO
dashed (solid) w/o (with) augmentation



Ideally we would like:

A better solution that does NOT rely on data augmentation and its assumptions.

Covariate Shift, or Biased Training Set

Light-curve data ... Type Ia or not

Given a feature space, X, and a label space, $Y \, (K > 1$ classes/dependent variables)

Spectroscopic training set

we have n_s labelled samples $\{x_i^s, y_i^s\}$ from the source domain

Photometric light-curve only

n_t unlabelled samples from the target domain, $\{x_i^t\}$.

Is it a Ia?

Task: predict $\{y_i^t\}$

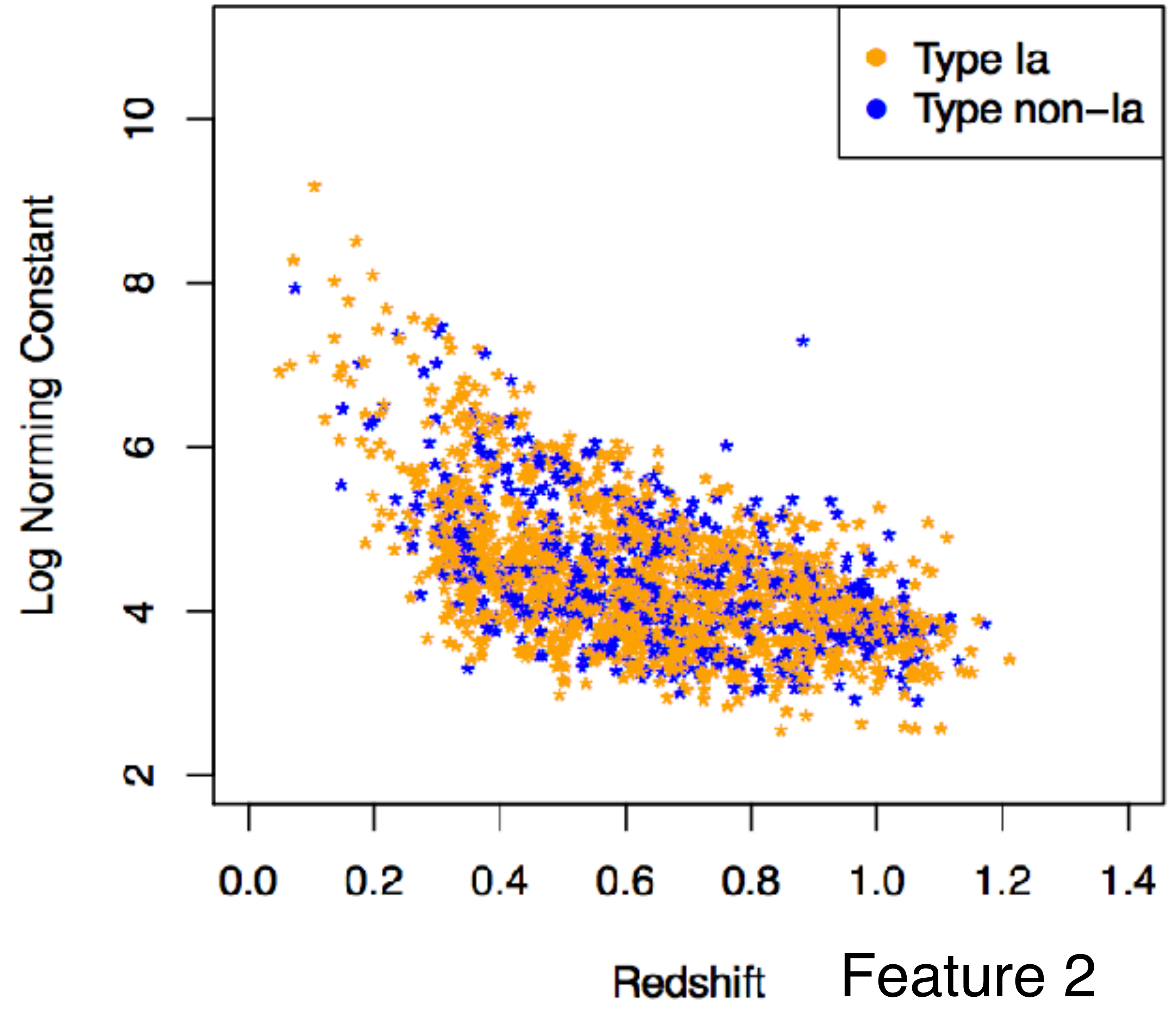
Features: redshift & apparent mag
Label: Ia or non-Ia

Covariate shift occurs when:

$$p_s(y\,|\,x) = p_t(y\,|\,x)$$

and $p_s(x) \neq p_t(x)$

I.e., the training set is non-representative of the test set.



Feature 1

Feature 2

Source domain (Coloured dots)

Target domain (Gray dots)

Revsbech, RT, van Dyk (2018)

Target Risk Estimation

Weighted ML estimation of risk

Approach: choose the classification/ regression function *f(x)* so as to to minimise the risk (= expected loss) over the target domain.

Shimoidara (2000) showed:

$$E_{(x,y)\sim D_t}[\ell(f(x),y)] = E_{(x,y)\sim D_s}\left[\frac{p_t(x)}{p_s(x)}\ell(f(x),y)\right]$$

Target (i.e. test data)

Source (i.e. training data)

The ratio of densities (weights) can be difficult to estimate reliably.

Bias correction approach

Let *s* be a binary indicator variable controlling training set selection (s=1).

Zadrozny (2004) showed:

$$E_{(x,y)\sim D}[\ell(f(x),y)] = E_{(x,y)\sim \tilde{D}}\left[\ell(f(x),y \,|\, s=1)\right]$$

$$\tilde{D} = \frac{P(s=1)}{P(s=1\,|\,x)}D$$

Estimate $p(s=1\,|\,x)$ via e.g. logistic regression, then draw samples from $\tilde{D}$.

Our Approach: Propensity Score Stratification

Work by **Max Autenrieth** (Stats PhD student), in collaboration with David van Dyk (Imperial) & David Stenning (Simon Fraser U.)

Improving on our previous work ("STACCATO"), Revsbech, RT, van Dyk (2018): StratLearn, Autenrieth et al (2021), arXiv 2106.11211

Propensity scores

$e(x_i)$ = probability for object *i* to be selected into the source domain, using the *whole features* set:

$$e(x_i) \equiv P(s_i = 1 \,|\, x_s, x_t)$$

Key idea (StratLearn):

subdivide ("stratify") target and source data in *k* subgroups according to quantiles of their propensity scores. Then supervised learning in each stratum ("stratified learner")

Propensity scores as balancing scores

Rosenbaum & Rubin (1983, 1984) show that, conditional on their propensity scores, the *k* subgroups ("strata") have approximately balanced covariate distribution, i.e.

$$p_{s_j}(x) \approx p_{t_j}(x) \text{ for } j = 1,\ldots,k$$

Since $p_s(y\,|\,x) = p_t(y\,|\,x)$, it follows that

$$p_{s_j}(x, y) \approx p_{t_j}(x, y) \text{ for } j = 1,\ldots,k$$

Hence covariate shift approximately disappears.

Toy Example (illustration of StratLearn)

1D simulation (regression):

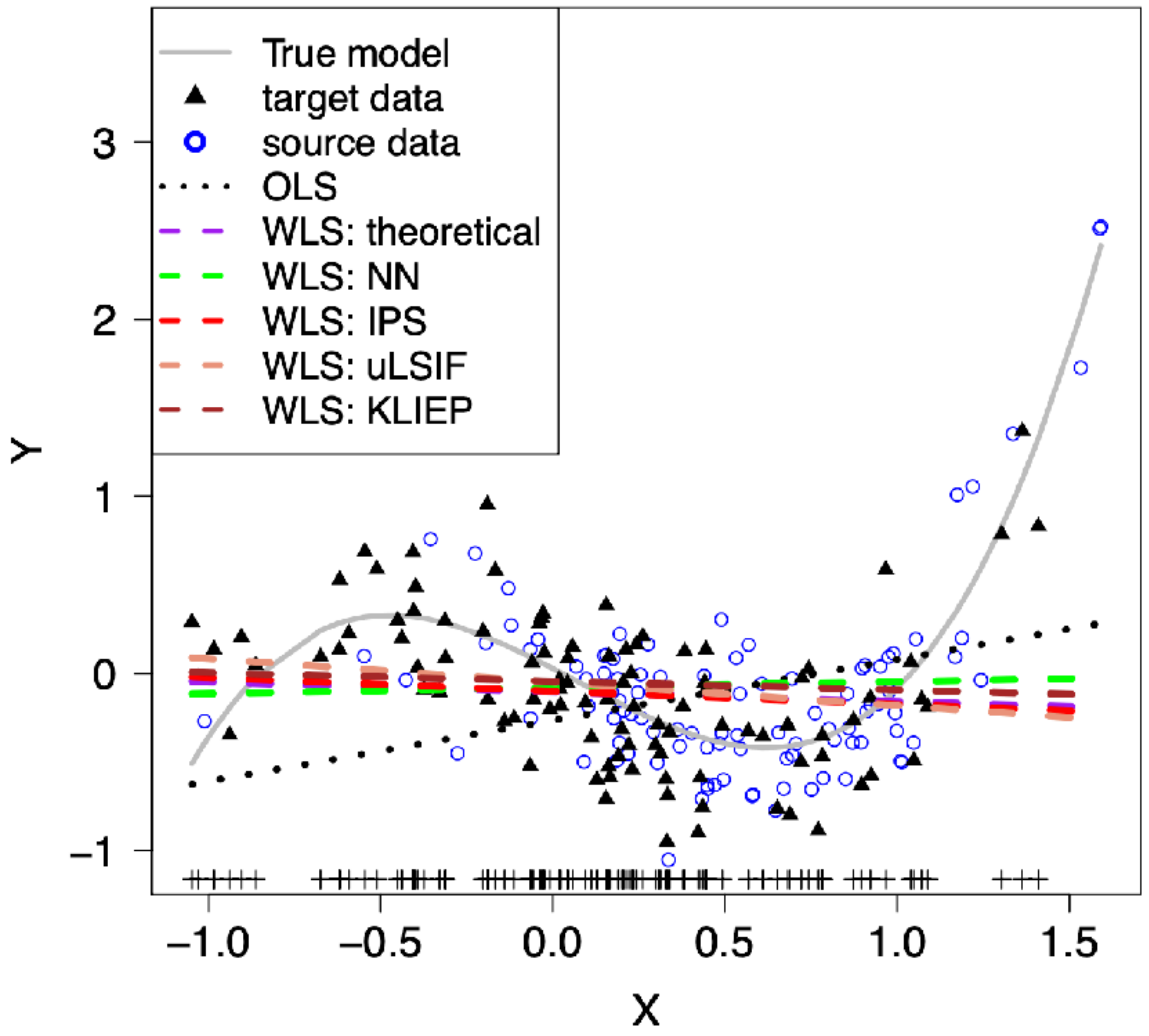
Source: $x_s \sim N(0.5, 0.5^2)$
Target: $x_t \sim N(0.2, 0.5^2)$

Outcome: $y_s = -x + x^3 + \epsilon$
Where $\epsilon \sim N(0, 0.3^2)$

Fit with misspecified (ie wrong) model:

Ordinary least square regression fit

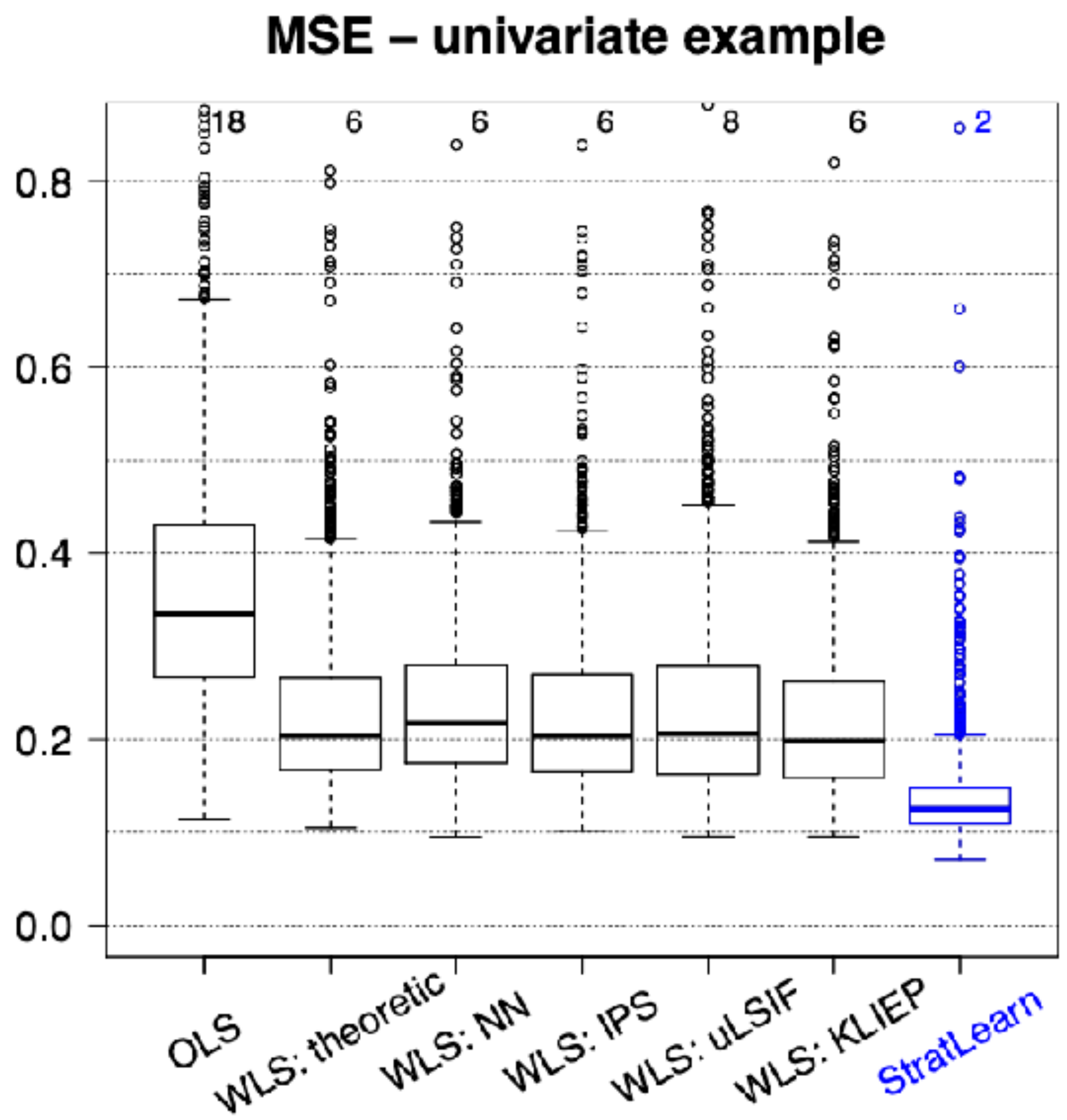
(For the experts: importance weighting does not work in this case)

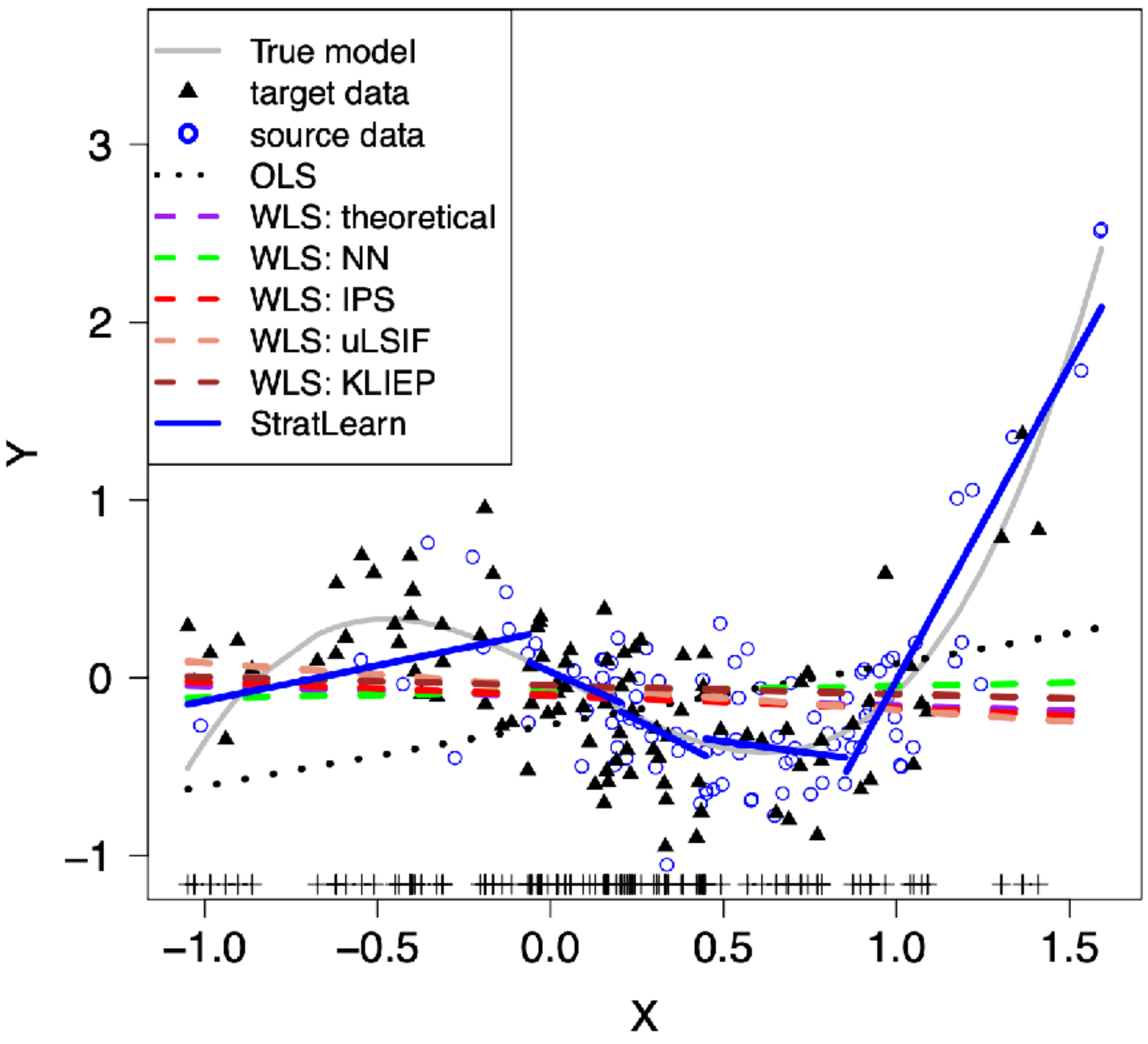


Toy Example

StratLearn solution:

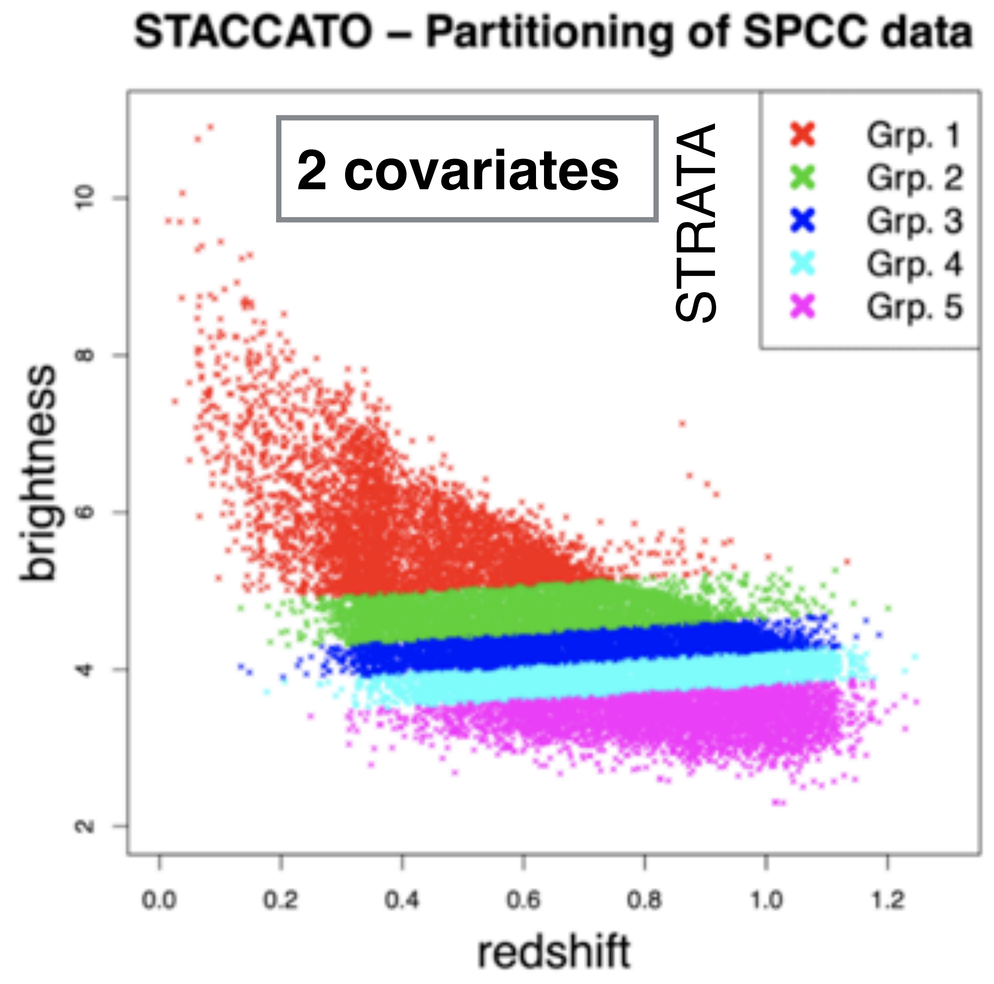
Subdivide the covariate space (ie. x axis) according to quintiles of propensity scores and fit the (wrong) linear model in each



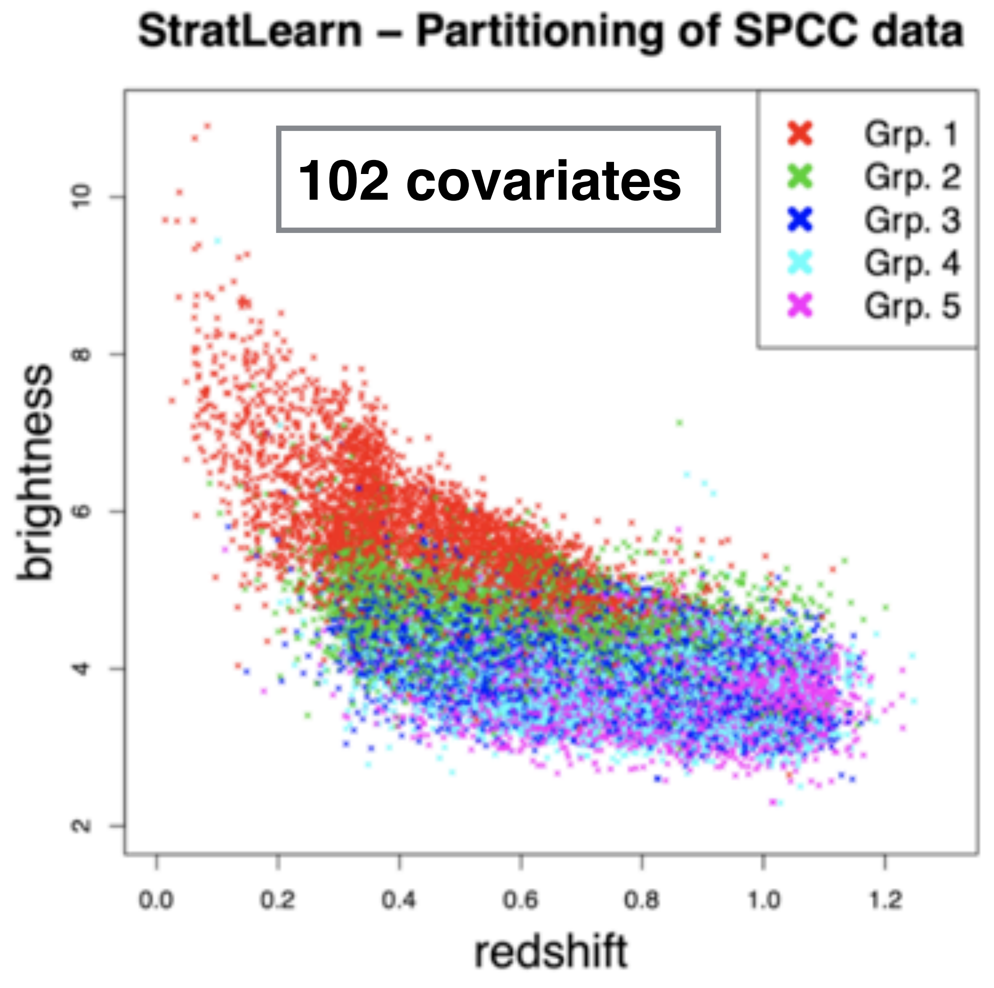


StratLearn on SNIa data

Propensity score partitioning of target domain (test data):



Decreasing probability of being in training set



Conditional on the propensity scores (i.e., within each stratum), the source and target outcomes are approximately the same.

This means: inside each stratum, the imbalance has been redressed, i.e. source data are approximately representative

Important: the underlying theorem only valid *if all potential confounding covariates* (i.e., things the SNIa type could depend on) are included in the propensity score estimation!

Stratum	Set	Number of SNe	Number of SNIa	Prop. of SNIa
1	Source	958	518	0.54
	Target	3306	1790	0.54
2	Source	120	28	0.23
	Target	4144	927	0.22
3	Source	13	4	0.31
	Target	4250	540	0.13
4	Source	7	4	0.57
	Target	4257	610	0.14
5	Source	4	4	1
	Target	4259	662	0.16

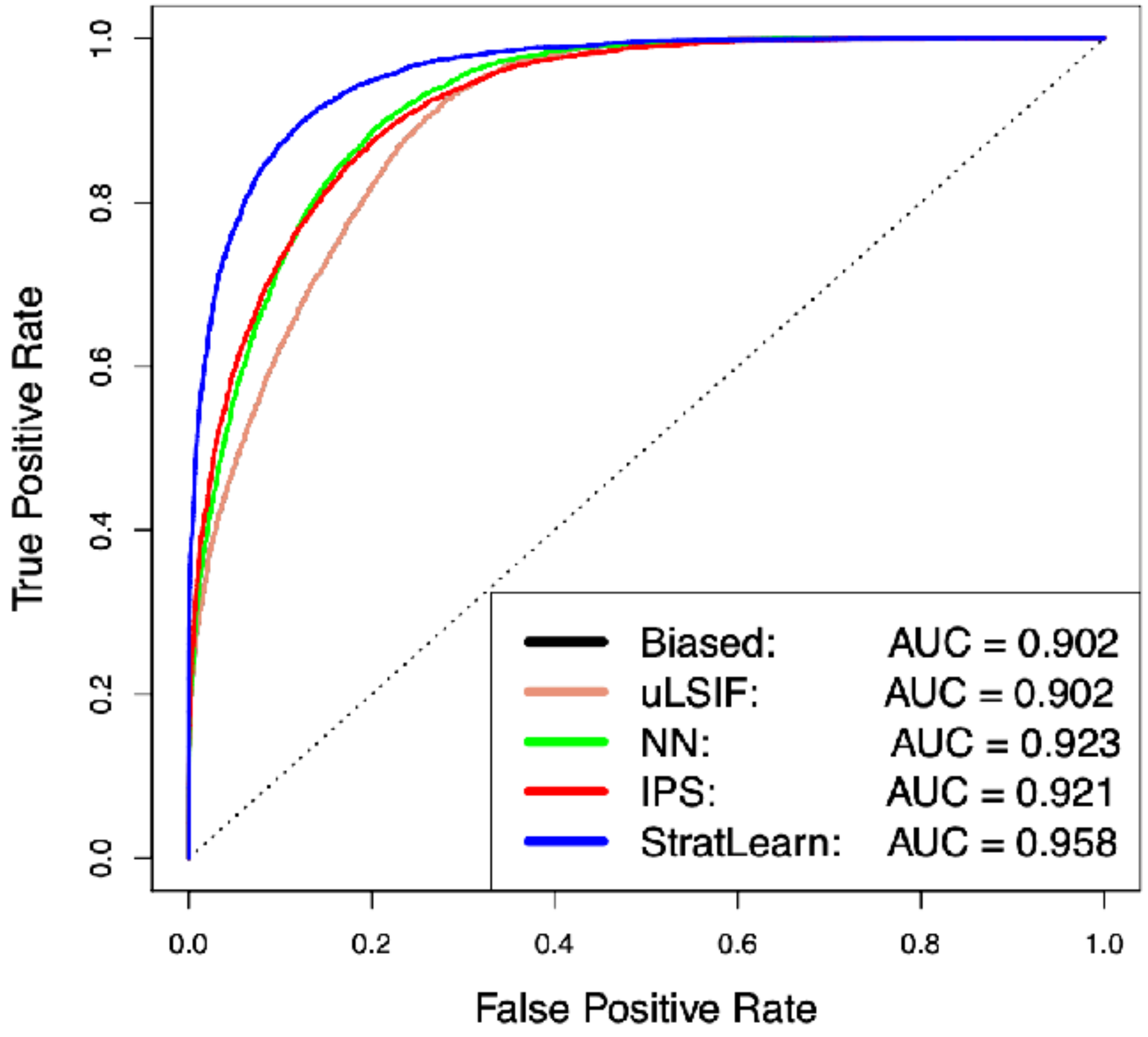
👍 Balanced proportions

👍 Balanced proportions

Classification Performance with StratLearn

Imperial College London & SISSA DATA SCIENCE

SPCC Challenge data (v2):



StratLearn performance close to "gold standard" of unbiased training set (AUC=0.977 vs 0.958) without any augmentation

Cf previous results:
Lochner et al (2016):
AUC= 0.855

Pasquet et al (2019):
AUC=0.939

Revsbech et al ("STACCATO", 2018):
AUC=0.94

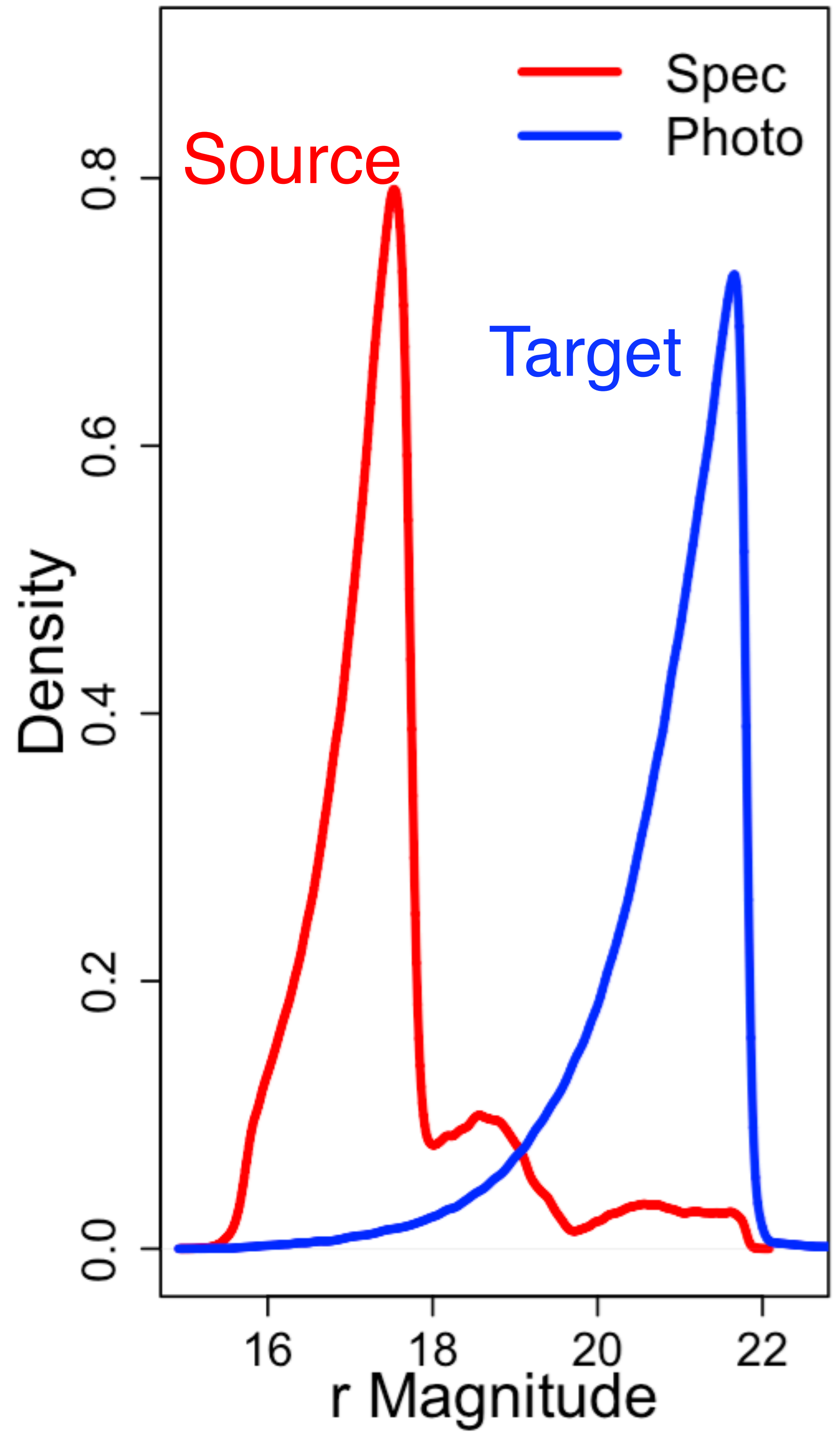
Note: AVOCADO (Boone, 2019), winner of the PLASTiCC challenge 2019, uses an extended version of STACCATO (incl. augmentation).

StratLearn for regression

We seek a principled framework for covariate shift adaptation via propensity score stratification that does not need augmentation:

1. Augmentation is problem-specific
2. Augmentation usually requires a validation set (not available)

Propensity score stratification leads to approximately balanced (i.e., unbiased) sub-groups, on which to perform supervised classification/regression.



Classification (2-way)

SNIa vs non-Ia, AUC:

Gold standard: 0.977
Best-in-class : 0.937
STACCATO : 0.961
StratLearning : **0.973**

Multivariate regression

Photo-z estimation
(~500,000 source/targets from SDSS DR 8)

Conditional Density Estimation Problem

The problem: estimate the conditional density of redshift, z, given the observed covariates (magnitudes in 5 different colour filters), in the presence of covariate shift.

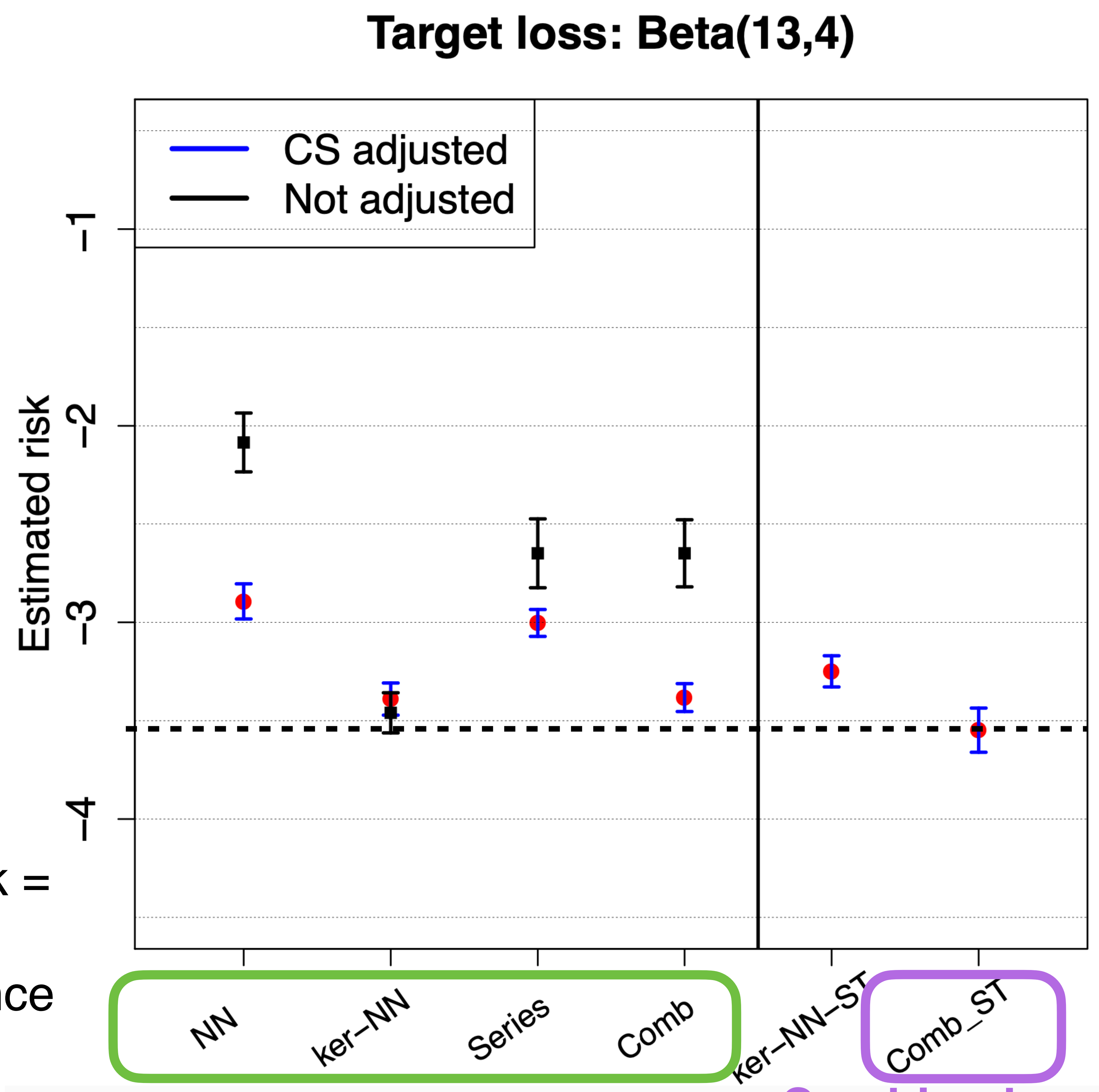
We compare our performance to the estimators in Izbicki et al (2017).

Approach:

1. "StratLearning" partitions the source and target domain according to propensity scores (**no augmentation needed**)
2. Within each group, we combine two conditional density estimation models from Izbicki et al (2017), *ker-NN* and *Series,* via a weighted average.
3. Weight is optimised by minimising the empirical loss on a validation set (a sub-set of the training set, **no test data needed**)

StratLearn: Photo-z Performance

Moderate shift
Low covariate dimensions

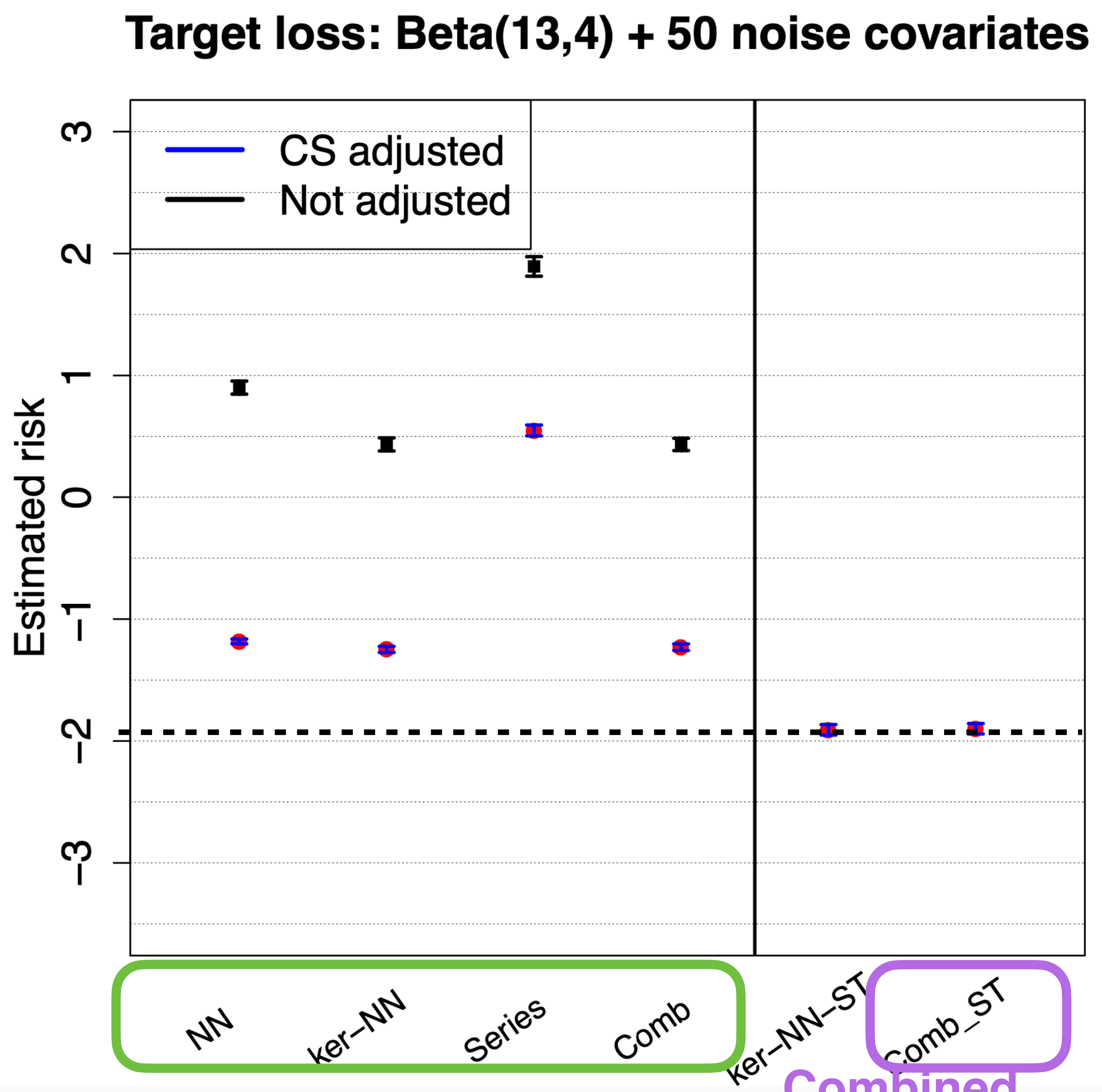


Lower risk = better performance

Previous methods

Combined StratLearning

Moderate shift
High covariate dimensions



Previous methods

Combined StratLearning

StratLearning outperforms previous methods for this problem.

Performance improvement is larger in the presence of high-D noisy covariate space (right panel).

Conclusions

1. Covariate shift is an important and recurrent phenomenon in supervised learning. In dark energy research, it will affect the next generation of large SNIa data.

2. We propose a general approach (StratLearn) based on stratifying source and target domain according to propensity scores (= probability of an object to be included in the source domain).

3. Within strata, source and target domains are better balanced: StratLearn shows improved performance in regression and classification tasks compared to best-in-class alternatives.

Thanks to my collaborators: Max Autenrieth (PhD student), David van Dyk (Imperial), David Stenning (Simon Fraser U.). Paper here: https://arxiv.org/abs/2106.11211

Thank you!

www.robertotrotta.com

@R_Trotta

Check out our new group:

datascience.sissa.it

References (Astro)

- E. A. Revsbech, R. Trotta, D. A. van Dyk, STACCATO: a novel solution to supernova photometric classification with biased training sets. *Monthly Notices of the Royal Astronomical Society* **473**, 3969-3986 (2018).
- R. Kessler *et al.*, Models and Simulations for the Photometric LSST Astronomical Time Series Classification Challenge (PLAsTiCC). *Publications of the Astronomical Society of the Pacific* **131**, 094501 (2019).1.
- Boone, K., Avocado: Photometric Classification of Astronomical Transients with Gaussian Process Augmentation. *The Astronomical Journal* **158**, 257 (2019).
- R. Kessler, et al (2010), Supernova Photometric Classification Challenge. ArXiv:1001.5210
- T. M. C. Abbott *et al.* (2019), First Cosmology Results using Type Ia Supernovae from the Dark Energy Survey: Constraints on Cosmological Parameters. *The Astrophysical Journal* **872**, L30.
- M. C. March, R. Trotta, P. Berkes, G. D. Starkman, P. M. Vaudrevange (2011), Improved constraints on cosmological parameters from Type Ia supernova data. *Monthly Notices of the Royal Astronomical Society* **418**, 2308-2329.
- S. R. Hinton *et al.* (2019), Steve: A Hierarchical Bayesian Model for Supernova Cosmology. *The Astrophysical Journal* **876**, 15.
- H. Shariff, X. Y. Jiao, R. Trotta, D. A. van Dyk (2016), BAHAMAS: New Analysis Of Type Ia Supernovae Reveals Inconsistencies With Standard Cosmology. *Astrophys. J.* **827**, 25.
- D. Rubin *et al.* (2015), UNITY: Confronting Supernova Cosmology's Statistical and Systematic Uncertainties in a Unified Bayesian Framework. *The Astrophysical Journal* **813**, 137 (2015).
- J. W. Richards, D. Homrighausen, P. E. Freeman, C. M. Schafer, D. Poznanski (2012), Semi-supervised learning for photometric supernova classification. *Monthly Notices of the Royal Astronomical Society* **419**, 1121.

References (Stats)

- Shimoidara (2000), Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference 90,* 2, 227–244.
- Zadrozny (2004), Learning and evaluating classifiers under sample selection bias. In Proceedings of the 21st international conference on Machine learning. ACM, 114.
- Rosenbaum, P. R. and Rubin, D. B. (1983), The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1, 41–55.
- Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association 79,* 387, 516–524.
- Chen, X., Monfort, M., Liu, A. & Ziebart, B.D.. (2016). Robust Covariate Shift Regression. Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, in PMLR 51:1270-1279