@R_Trotta

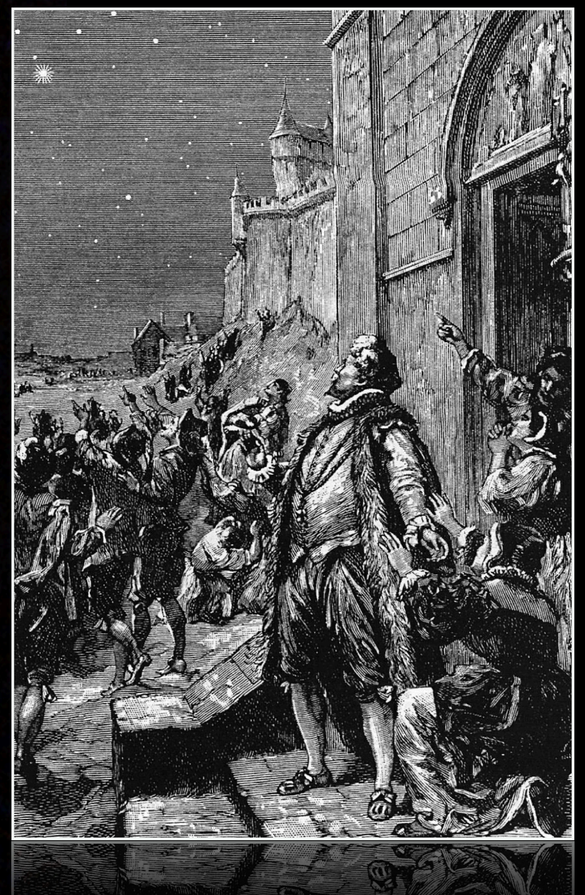
Supervised Learning with Biased Training Data and Applications to Supernova Type Ia Cosmology

Roberto Trotta

Imperial College London
&
International School for Advanced Study (SISSA)

CERN-EP/IT Data science seminar
9 Dec 2020

Camille Flammarion,
Astronomie populaire (1880)



Outline

Setting the scene:

Supernovae Type Ia and the accelerated expansion

The statistical problem:

Supervised learning in the presence of training set bias

Idea:

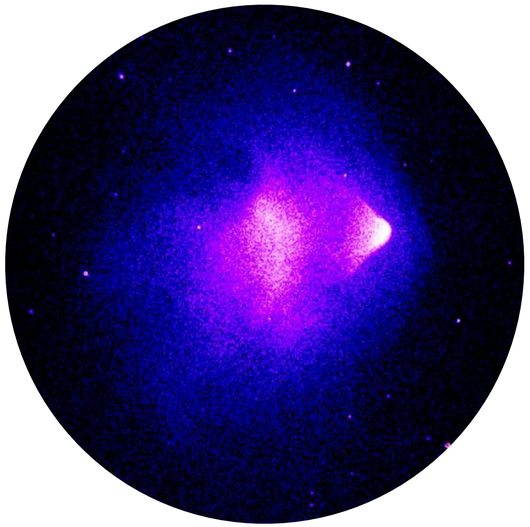
Propensity Score Stratification approach

“Stratified Learning” approach:

Performance in classification and regression problems

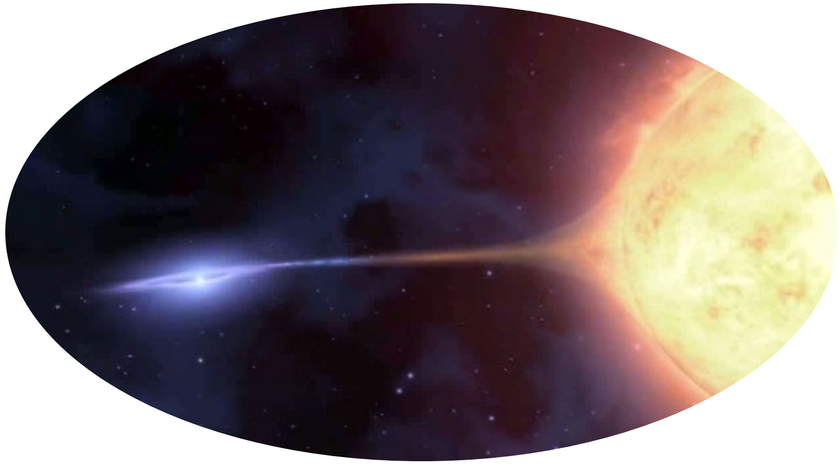
Conclusions

What is dark matter?



Credit: M. Markevitch

What makes the cosmos accelerate?



Credit: NASA

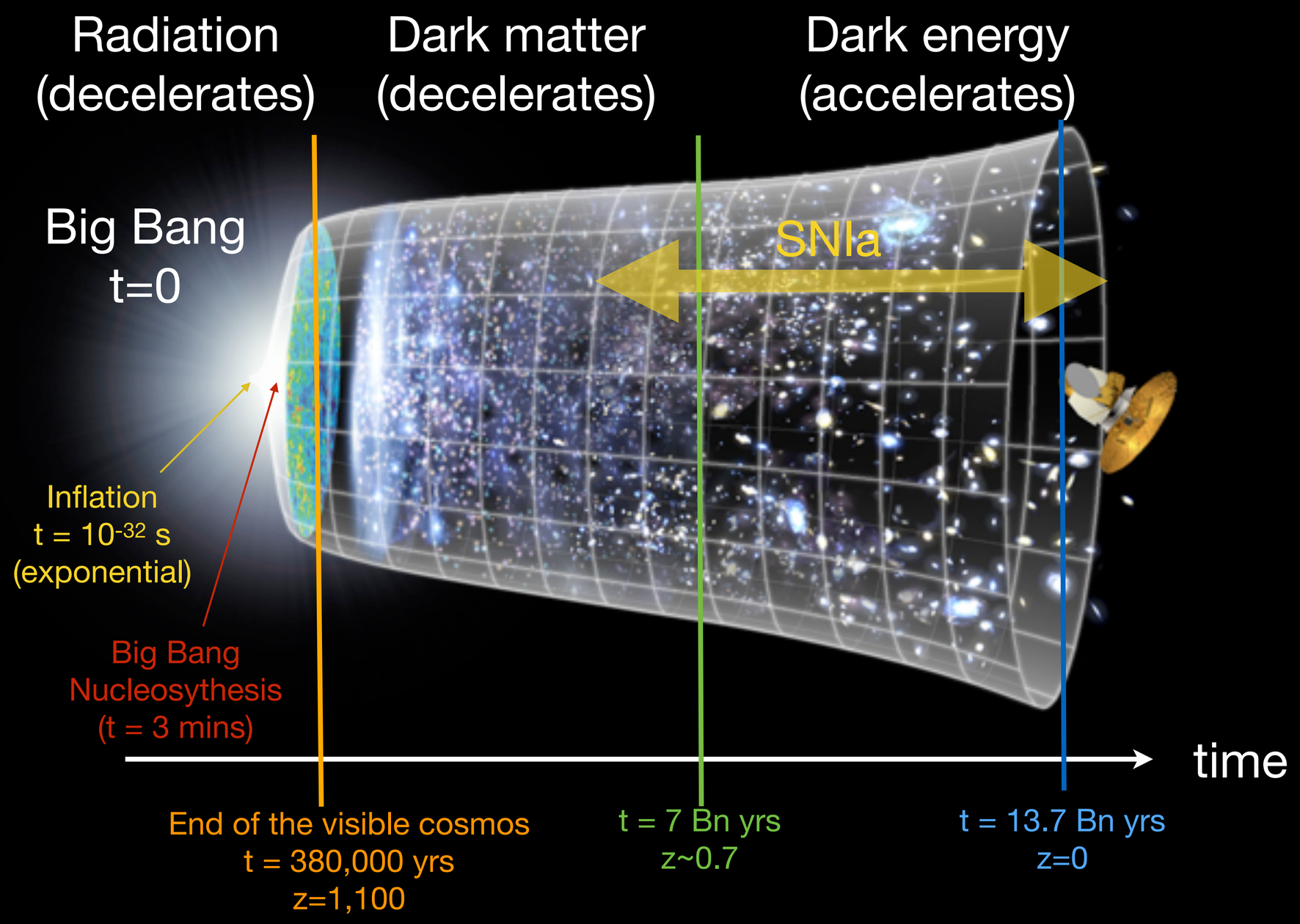
Is there life elsewhere?

Credit: Ricardo Ramirez

5% Normal Matter

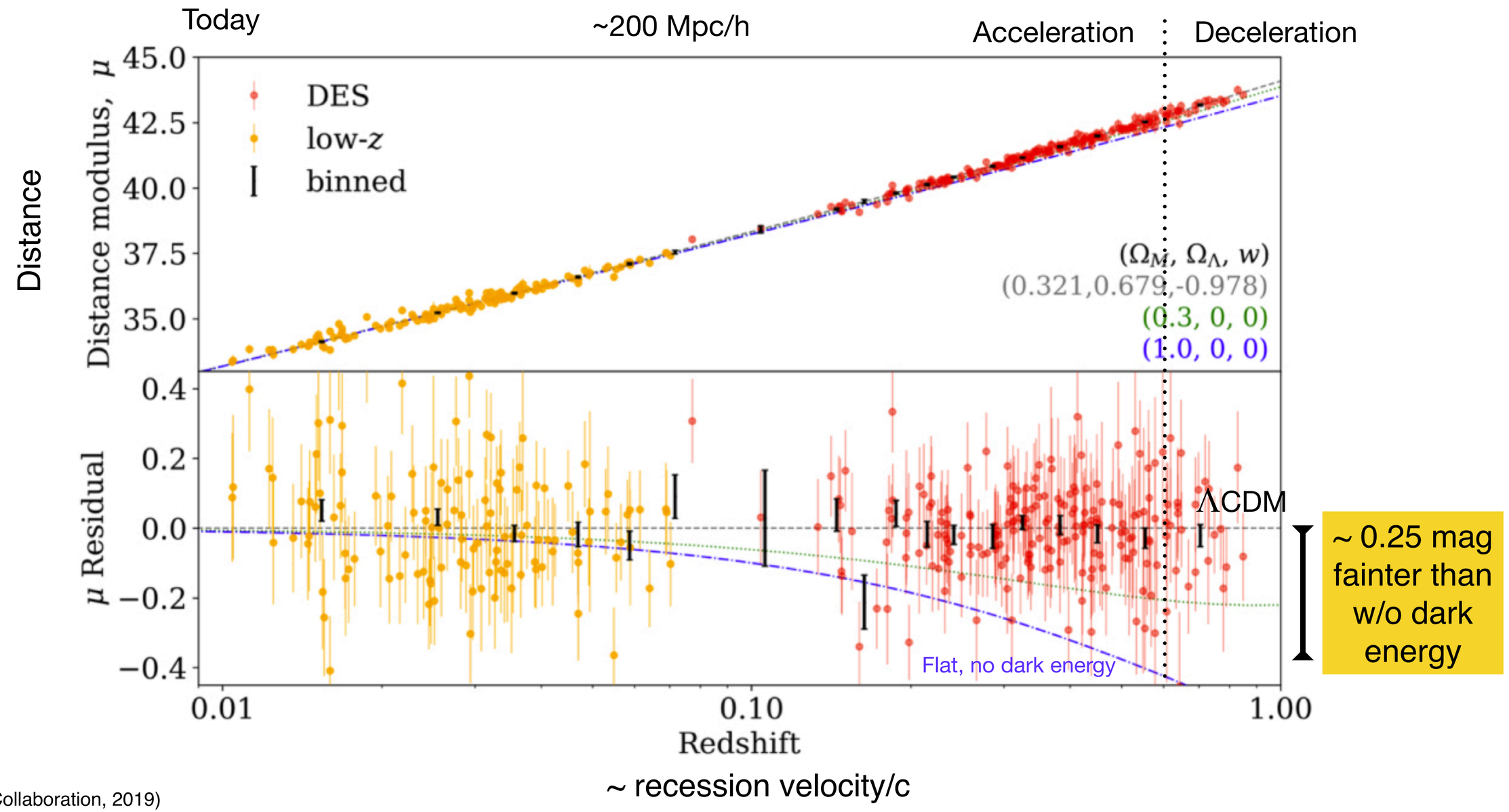
25% Dark Matter

70% Dark Energy



NASA/WMAP Science Team

Cosmic Acceleration

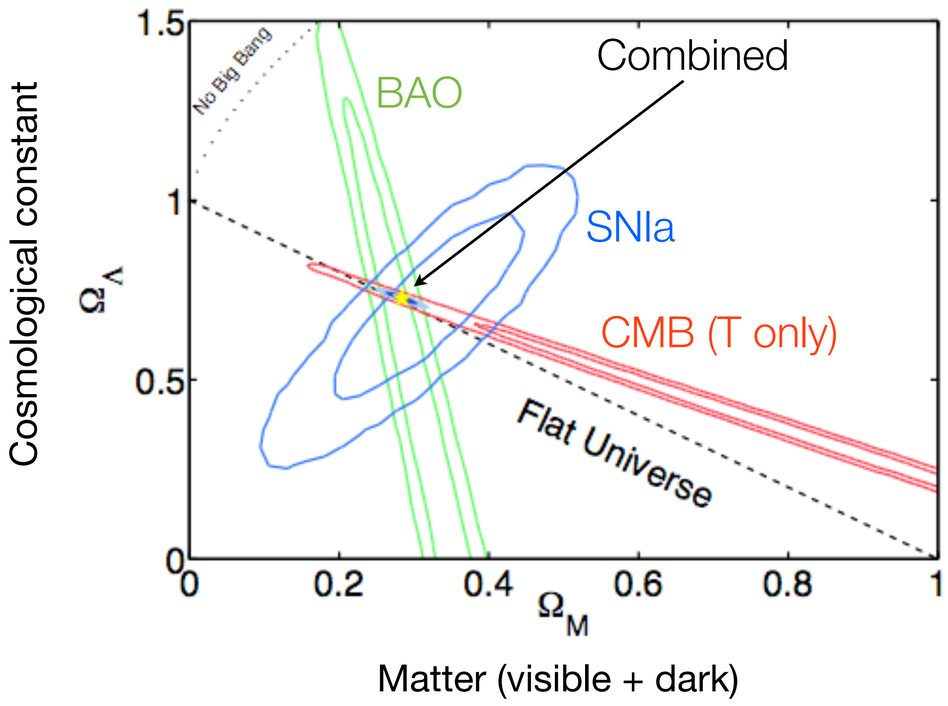


Abbott et al (DES Collaboration, 2019)

Cosmological Constraints

Combination of multiple probes **is required** to determine the cosmological parameters to high precision.

Baryonic Acoustic Oscillations (BAO) and Cosmic Microwave Background (CMB) are standard rulers.



March et al, MNRAS 418(4):2308-2329, (2011)

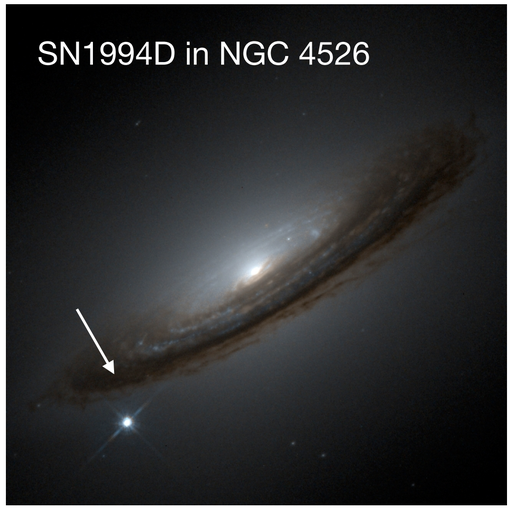
$$\Omega_m = 0.316 \pm 0.009$$
$$H_0 = 67.8 \pm 0.9 \text{ km/s/Mpc}$$
$$\Omega_\Lambda = 0.684 \pm 0.009$$
$$|\Omega_\kappa| < 0.005$$
$$\text{Age} = 13.796 \pm 0.029 \text{ Gyr}$$

Planck Collaboration (2015)

How accurate are these results?

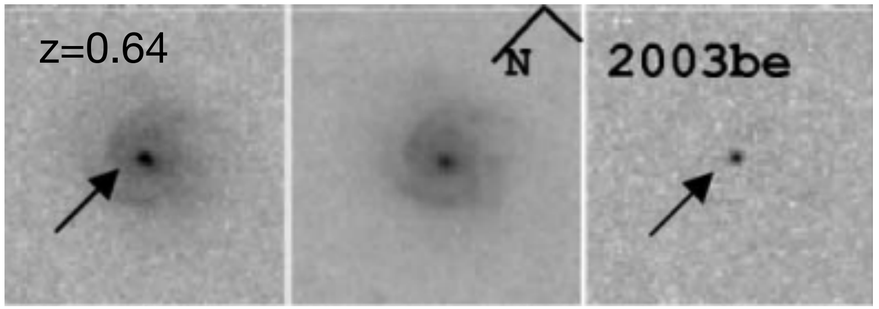
Supernovae Type Ia: Observations

DETECTION



Treffers et al (1994); imaged by HST

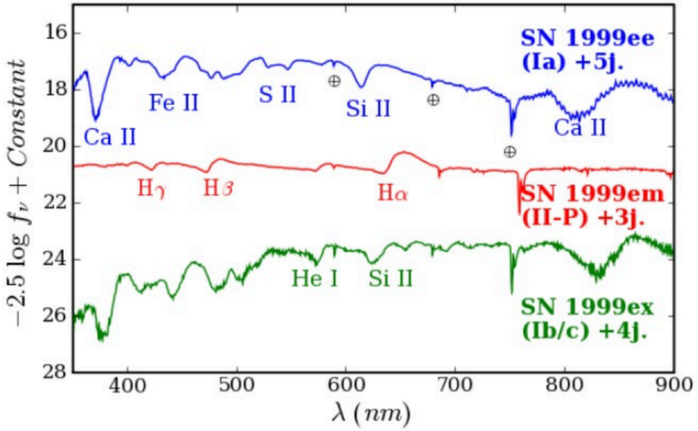
During After Subtracted



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

TYPING

Spectra: Type Ia have no H lines

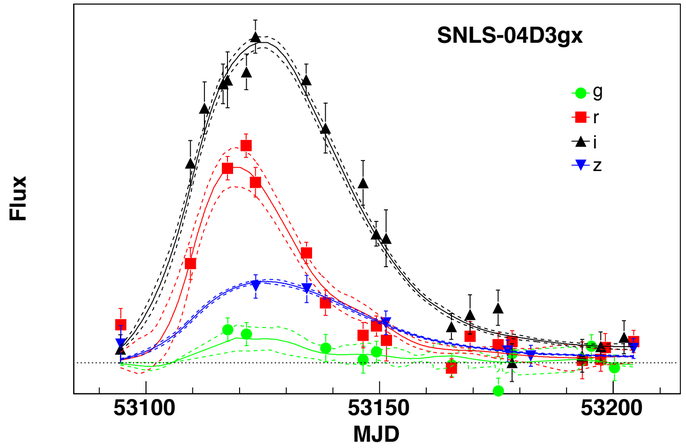


Credit: Julien Guy

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

FOLLOW-UP

~ 15 days ~ 20 days



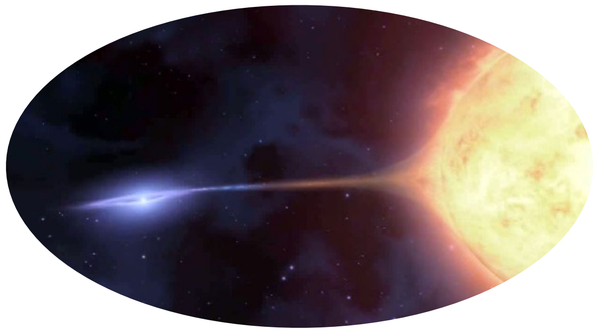
Guy et al (2007)

Lightcurves: time-evolution of brightness in several colour filters

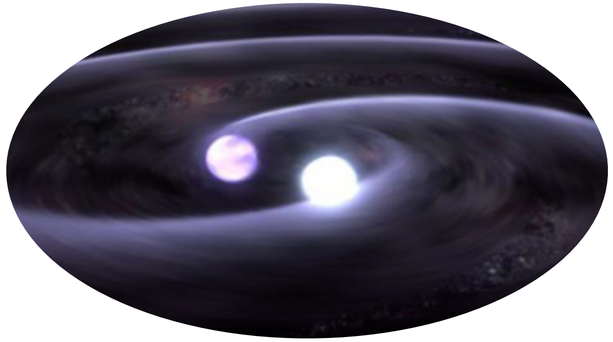
Supernovae Type Ia: Origin

PROGENITORS

CO white dwarf accreting mass.



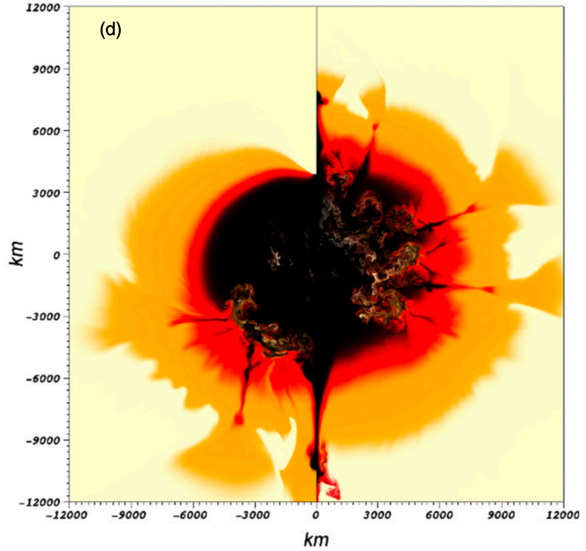
Single-degenerate
(e.g. Hosseinzaden et al, 2017)



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

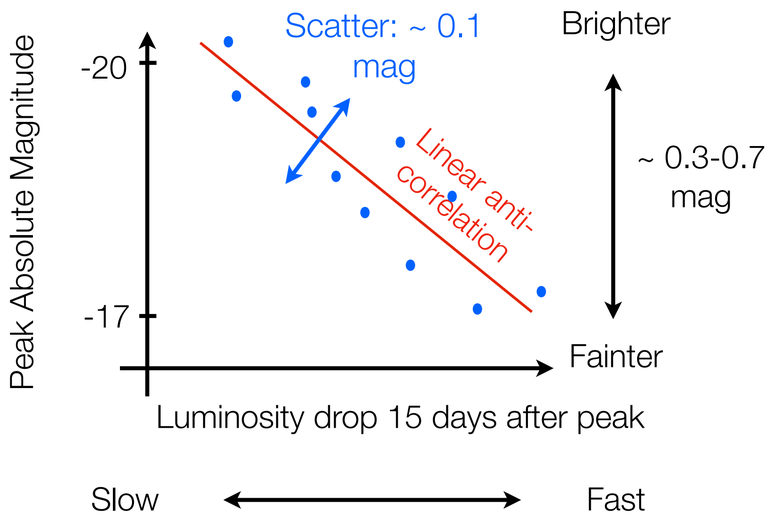
EXPLOSION

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

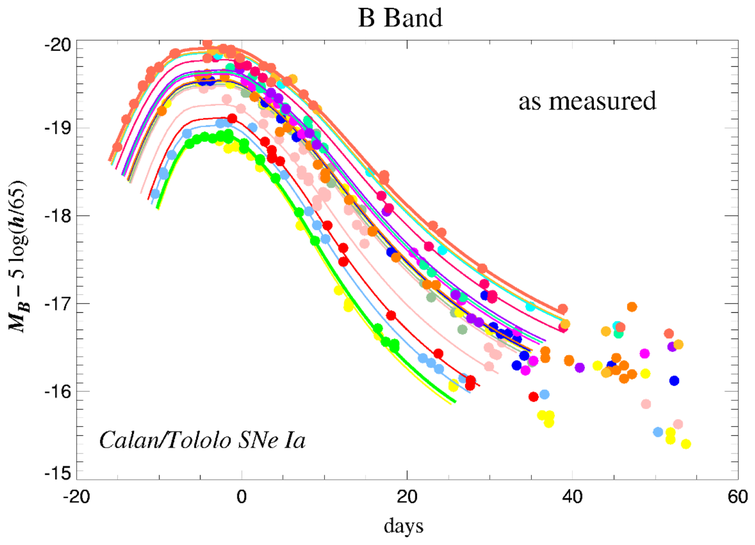
"Brighter SNIa are slow decliners"

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

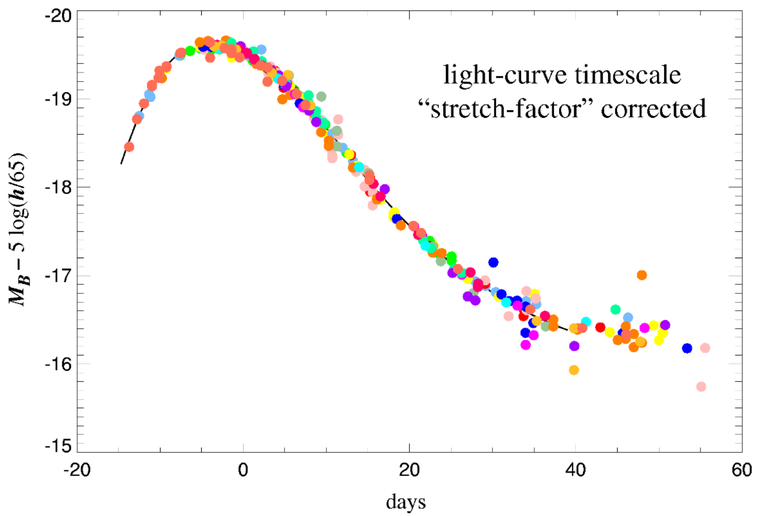
BEFORE CORRECTION

BRIGHTER

FAINTER



AFTER CORRECTION



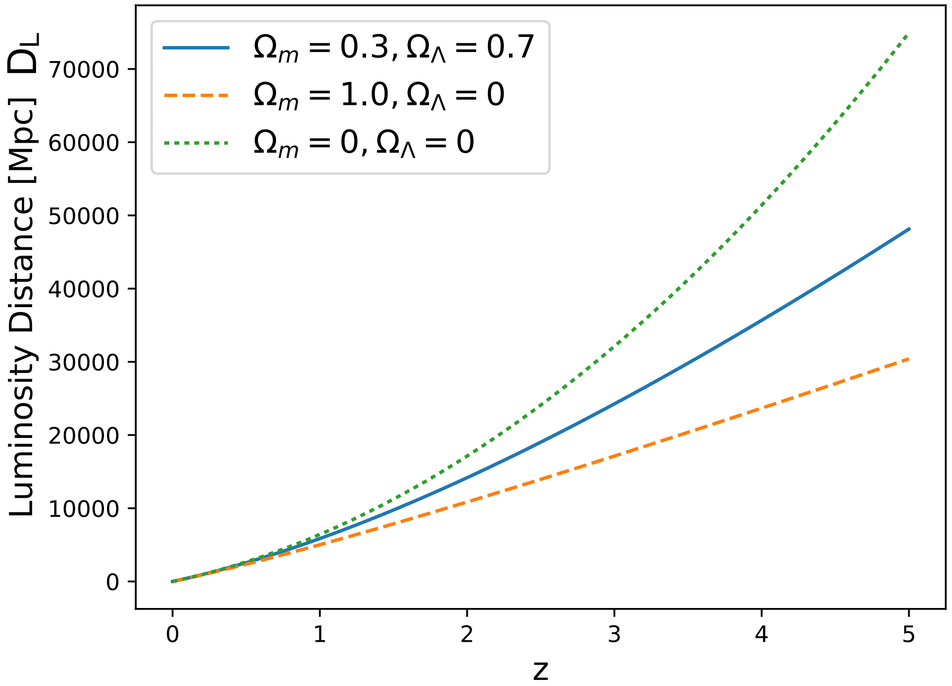
Kim et al (2007)

Constraining Cosmological Parameters

DISTANCE-REDSHIFT RELATION

Measure redshift and distance modulus

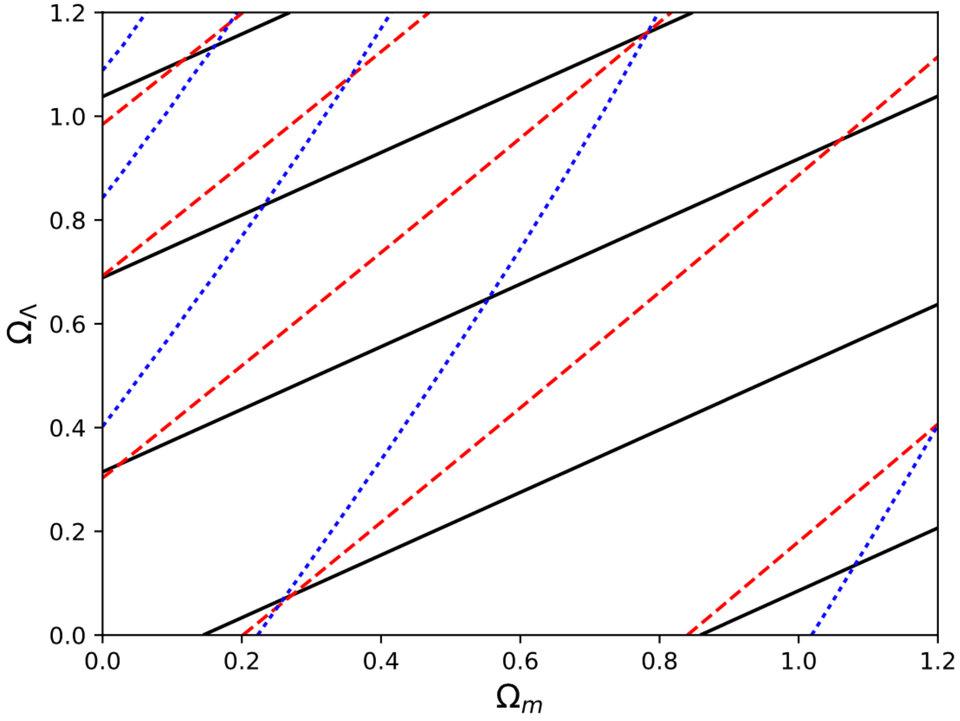
$$\mu = m_B - M + \alpha x_1 - \beta c$$



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

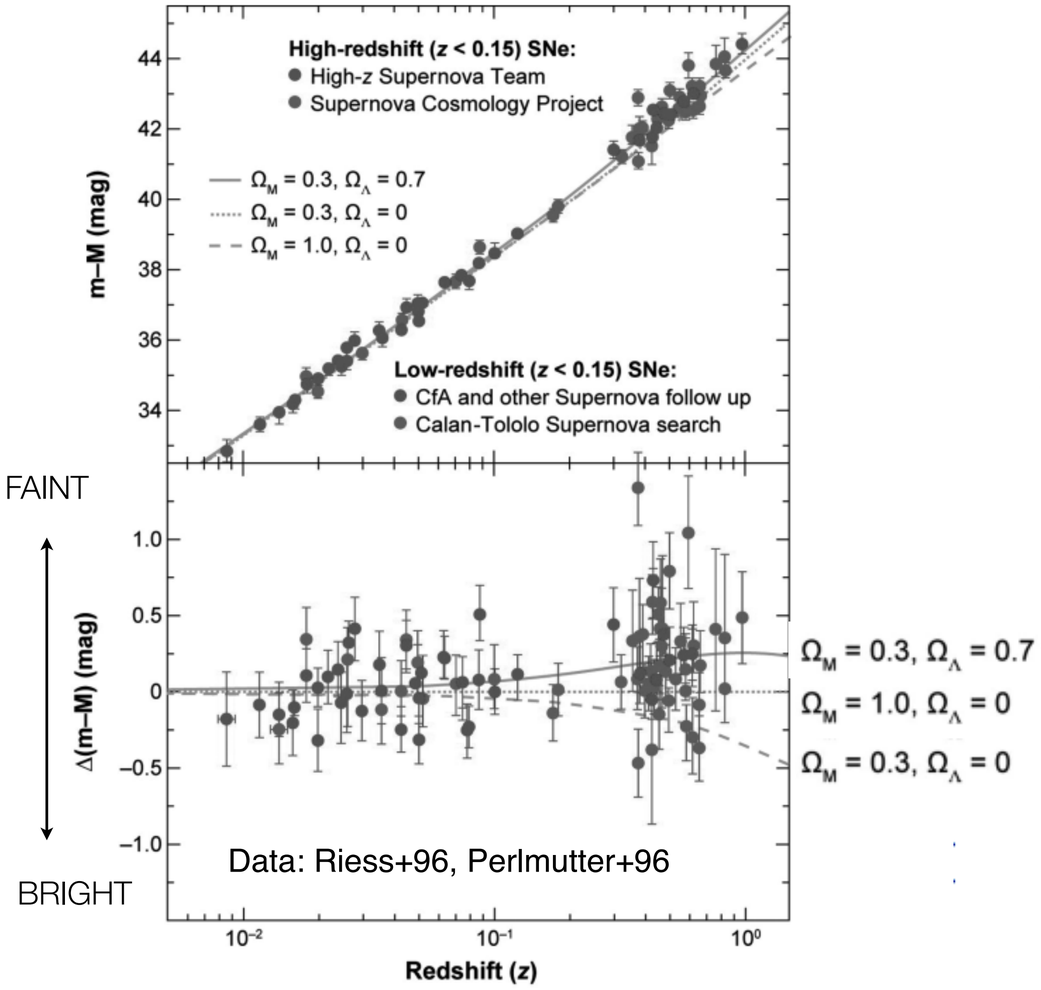
DIFFERENTIAL DISTANCE MEASUREMENT

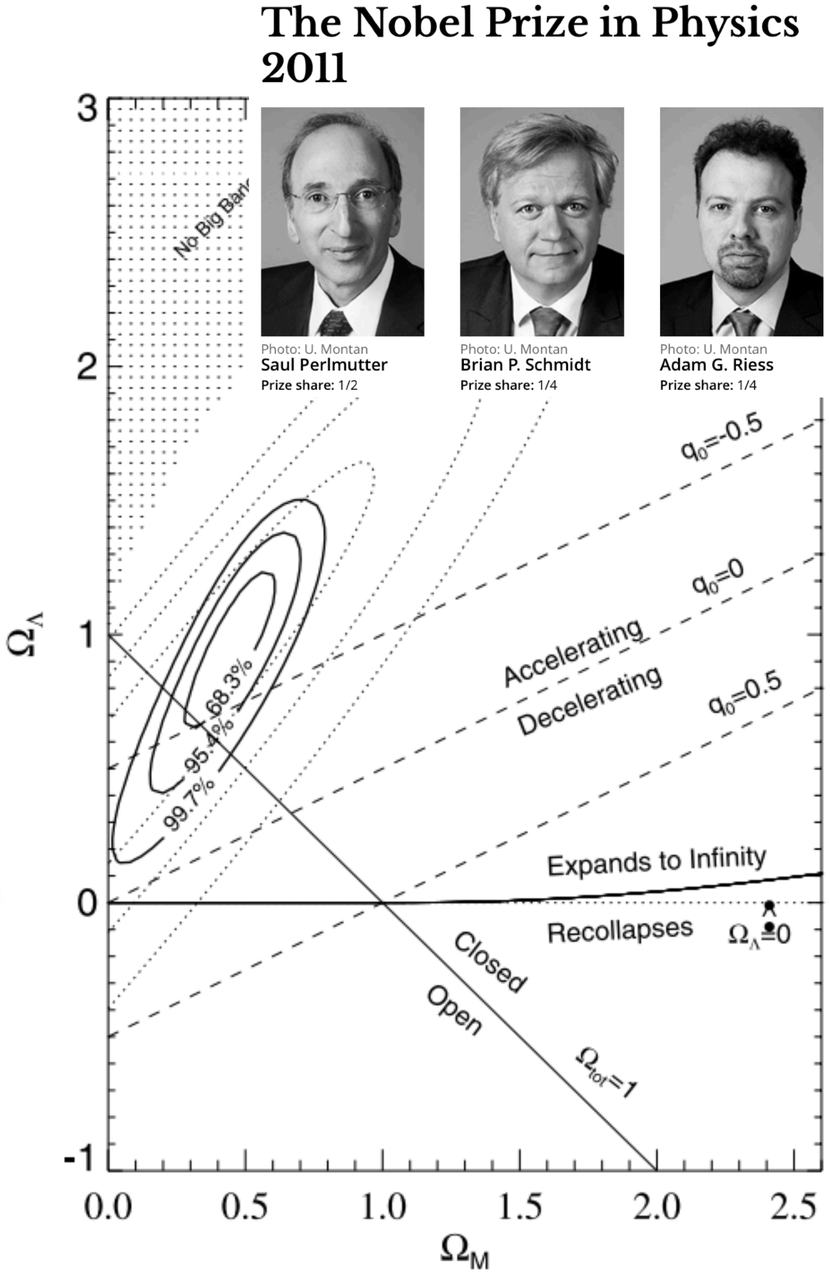
$z = 0.1$ $z = 0.5$ $z = 1.0$



Contours of constant D_L at various redshifts

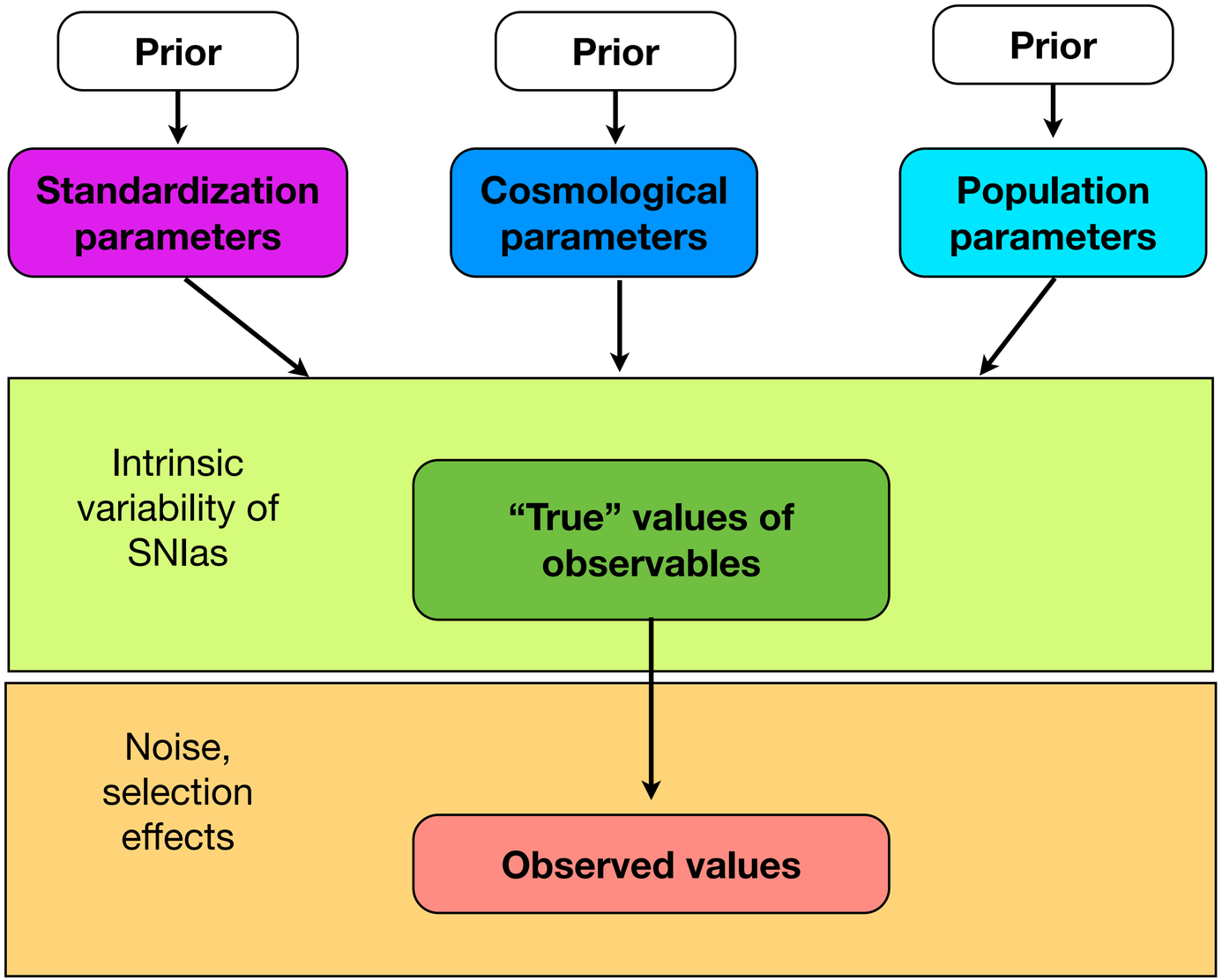
COLLECT THE NOBEL PRIZE





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

Bayesian Hierarchical Modelling



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

For more details, see:

BHM: March, RT et al (2011)

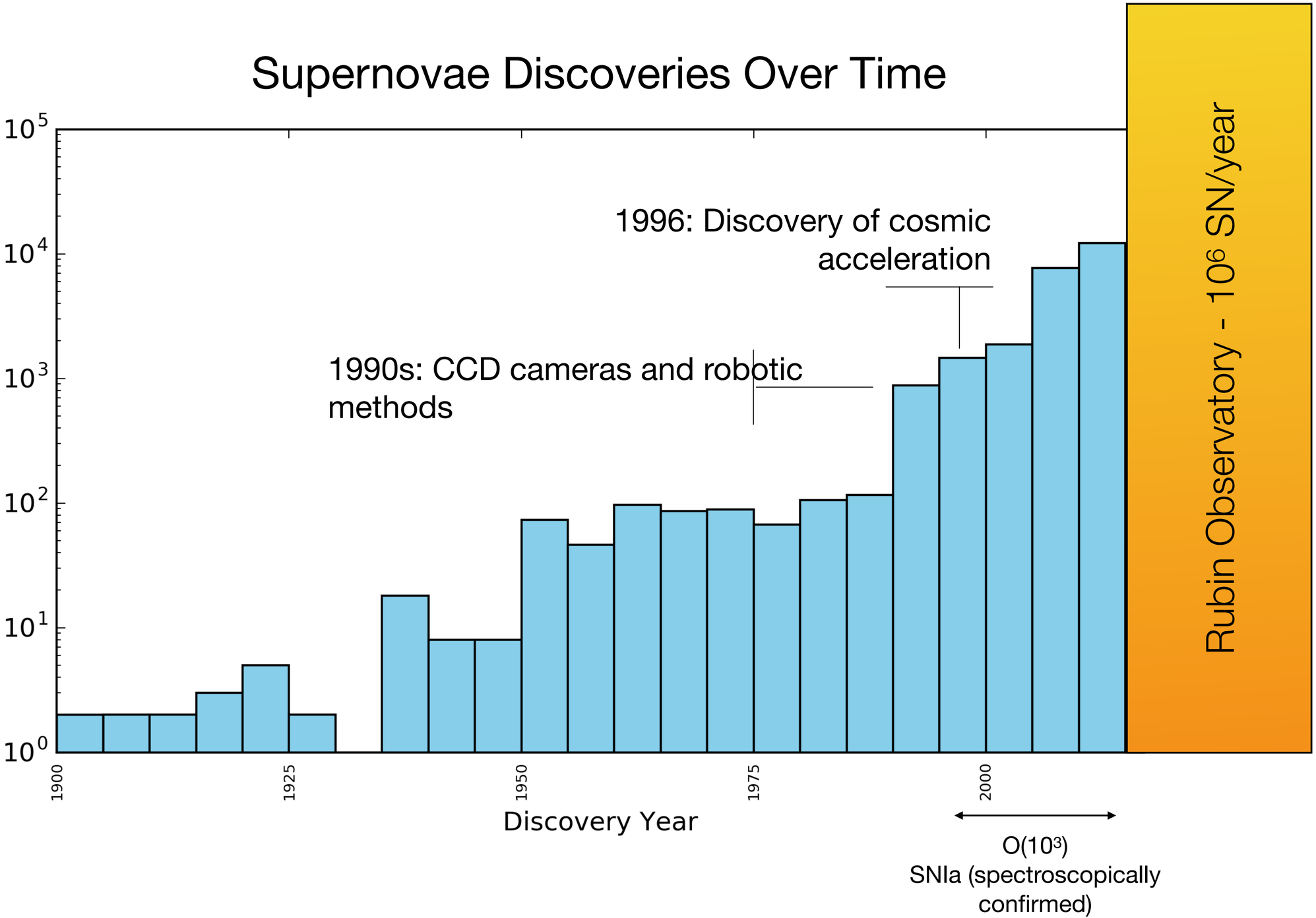
Unity: Rubin et al (2015)

BAHAMAS: Shariff, RT et al (2016)

Simple-BayeSN: Mandel et al (2017)

Steve: Hinton et al (2019)

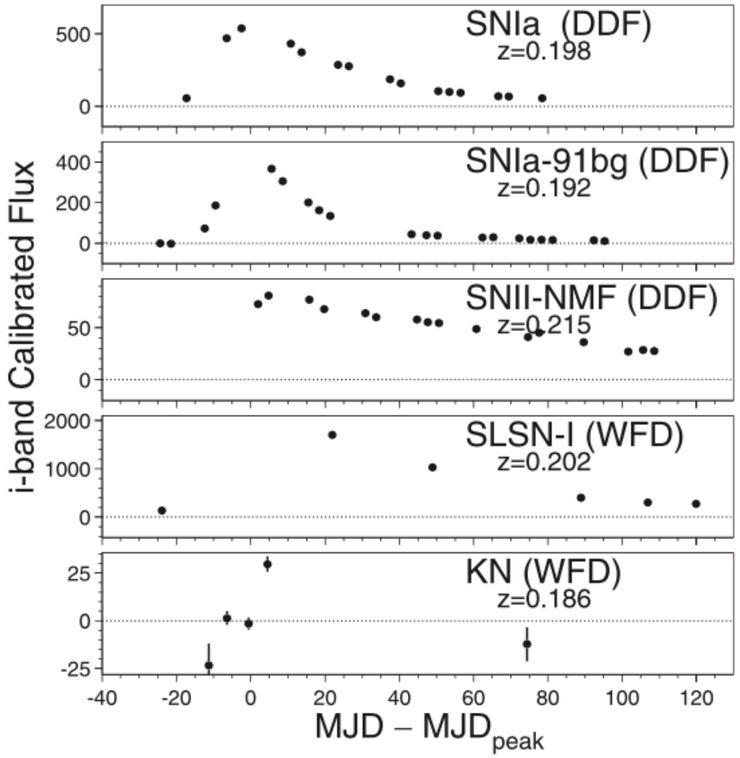
Supernovae Discoveries Over Time



Statistical Challenge

To classify Type Ia vs non-Ia **reliably** and **efficiently** from light-curve data alone

Simulated light-curves



Kessler et al (2019)

<u>Classification challenges:</u>

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

Supernova Photometric Classification Challenge (Kessler et al, 2010)

Covariate Shift, or Biased Training Set

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

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

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

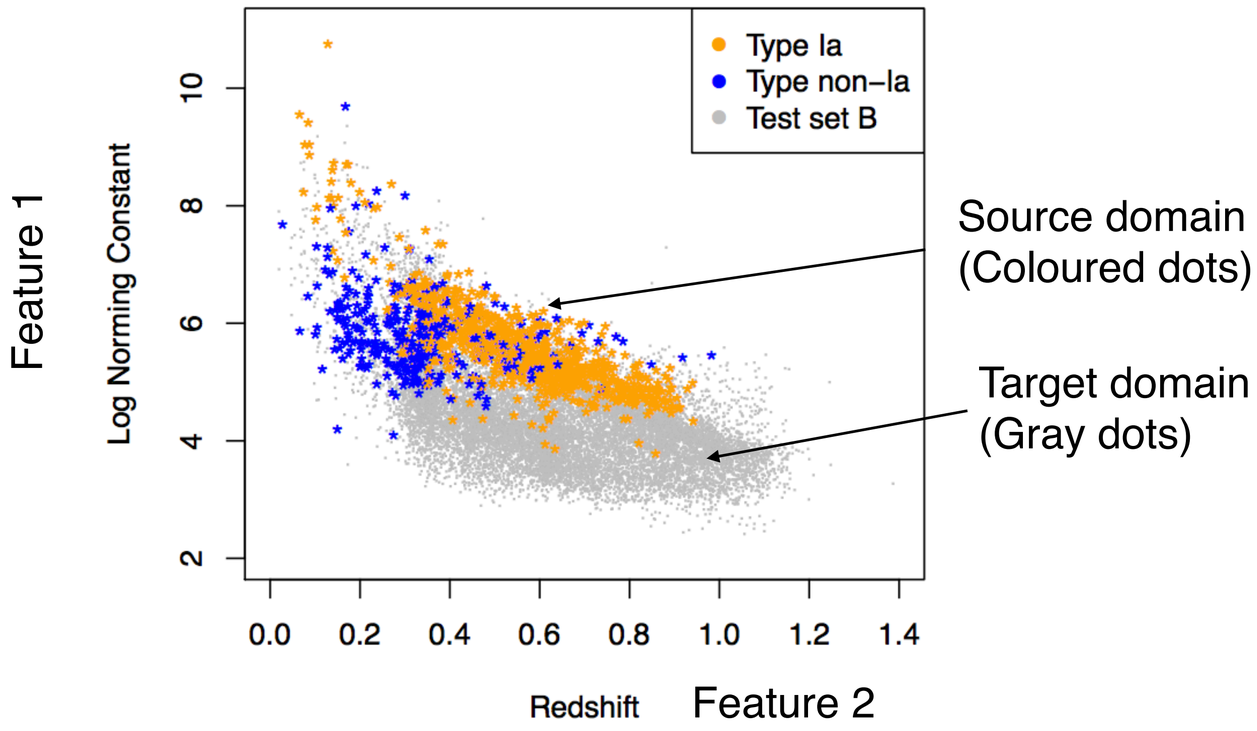
Task: predict $\{y_i^t\}$

Covariate shift occurs when:

$$p_s(y \mid x) = p_t(y \mid x)$$

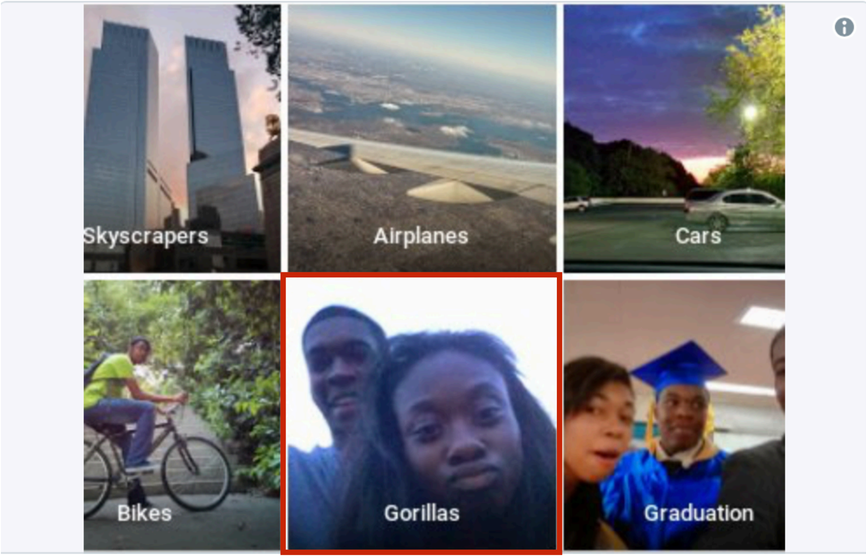
and $p_s(x) \neq p_t(x)$

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



Revsbech, RT, van Dyk (2018)

2015

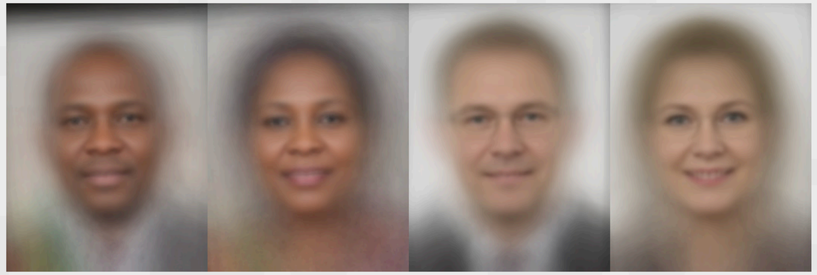


Jacky lives on @jalcine@playvicious.social now.
@jackyalcine

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

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.

Cosmology

Incorrect classification of Type Ia vs non-Ia would lead to cosmological parameters systematic bias.

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:

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

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)

Previous work (STACCATO) by Esben Revsbech (Revsbech, RT, van Dyk, 2018)

Propensity scores

e(x$_i$) = probability for object *i* to be selected into the source domain:

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

Idea:

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

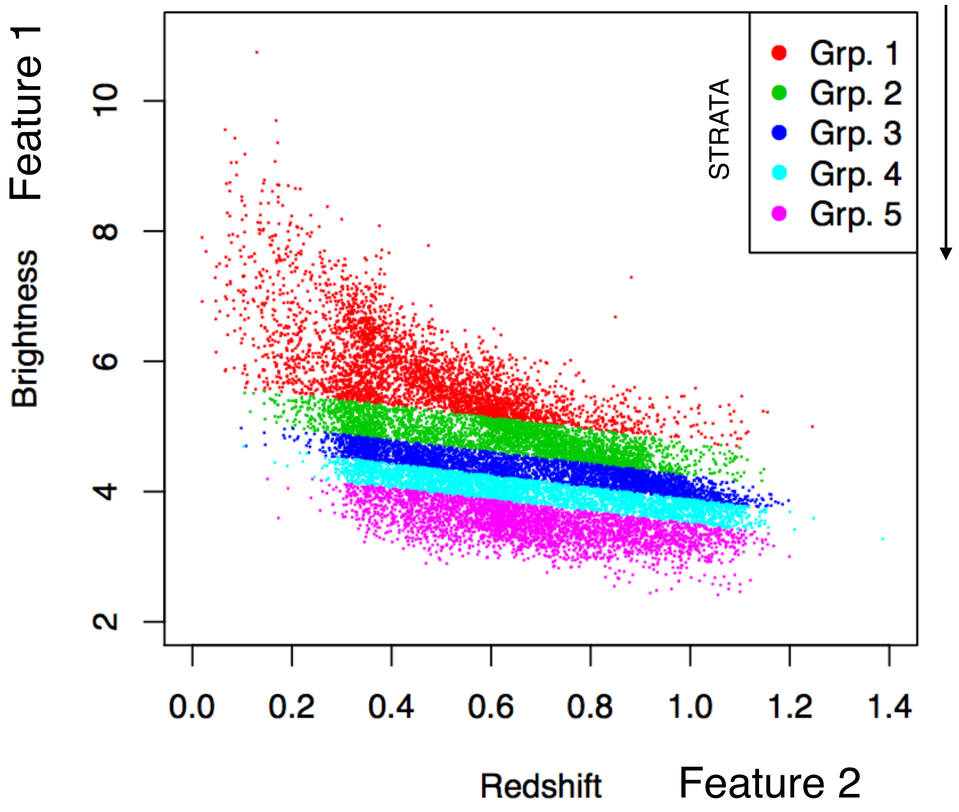
$$p_{s_j}(x, y) \approx p_{t_j}(x, y) \text{ for } j = 1, \ldots, k$$

STACCATO

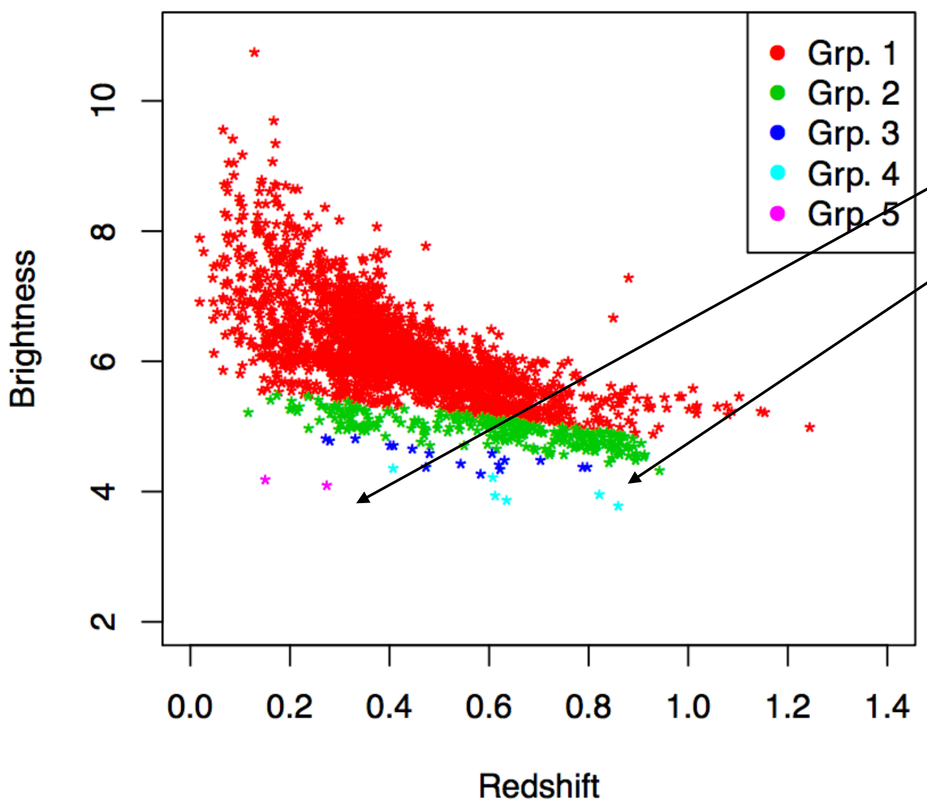
Previous work: **S**yn**T**hetically **A**ugmented Light **C**urve **C**lassific**AT**i**O**n (STACCATO), Revsbech et al (2018)

Propensity score partitioning of target domain (test data)

Decreasing probability of being in training set



Propensity score partitioning of source domain (training data)



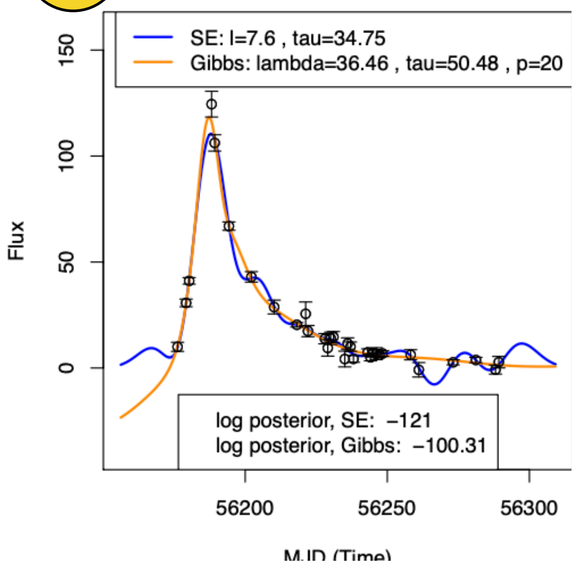
Issue: sparsity in lowest propensity score grouping requires augmentation via simulation in the source domain.

Solution: optimize GP-based augmentation via validation set (usually not available).

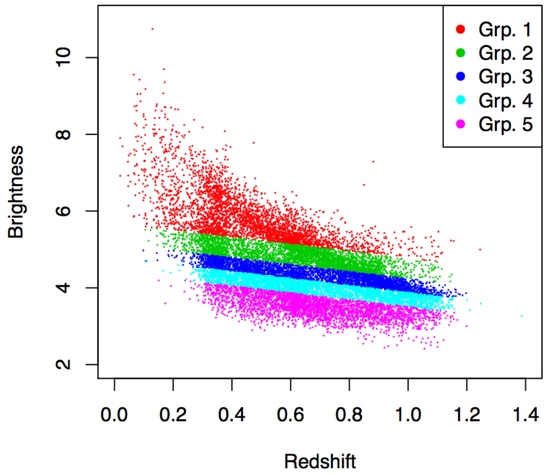
STACCATO: Analysis Details

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

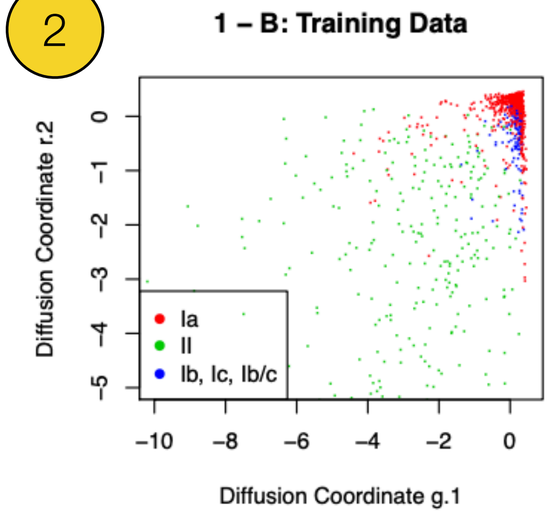
1

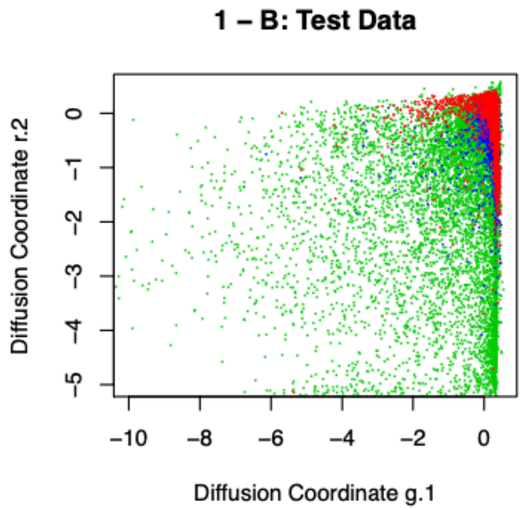


3



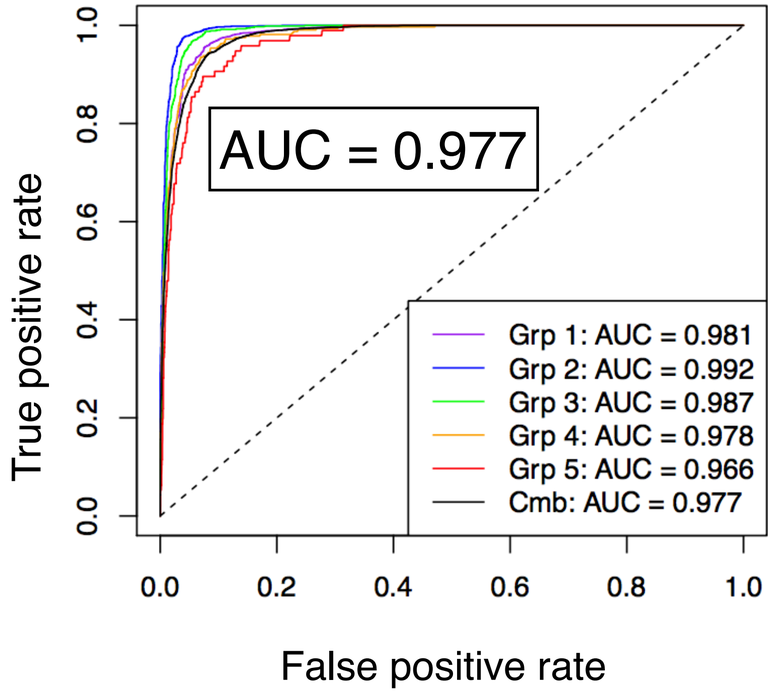
2



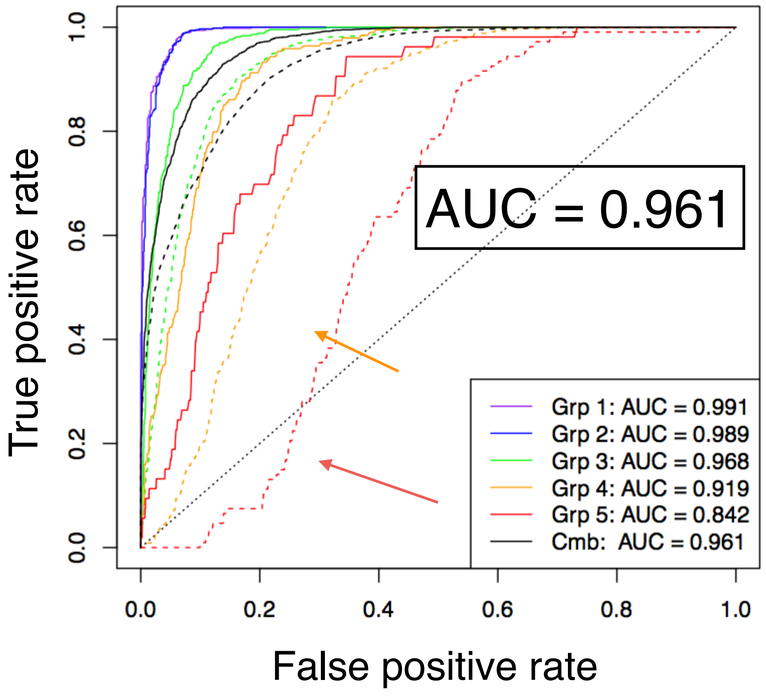


Classification Performance with Augmentation

Representative source data (unbiased)



STACCATO on biased source data



Dashed (solid) lines are before (after) augmentation; colours are for individual strata; black combined.

STACCATO performance close to “gold standard” of unbiased training set (left)

Later, the STACCATO framework was improved and extended (w/o using validation) in AVOCADO (Boone, 2019), winner of the PLASTiCC challenge 2019.

Current Work: Stratified Learning

Work by **Max Autenrieth** (Stats PhD student), in collaboration with David van Dyk (Imperial)

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.

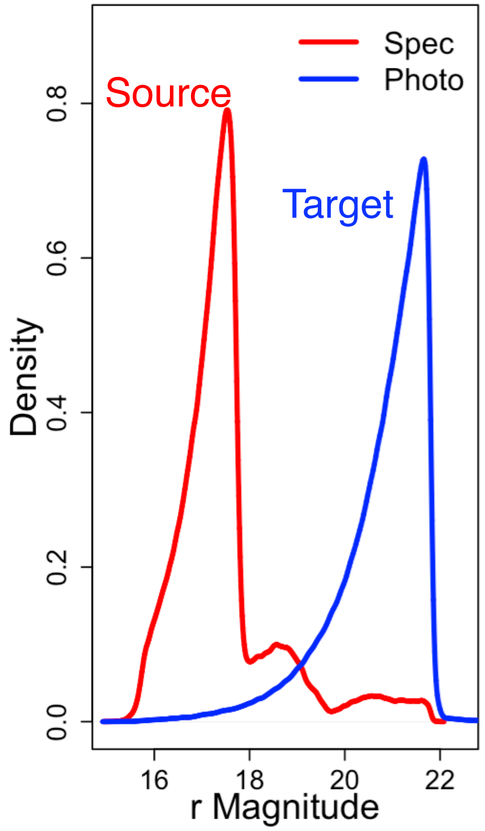
Classification (2-way)

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)



Stratified Covariate Balancing

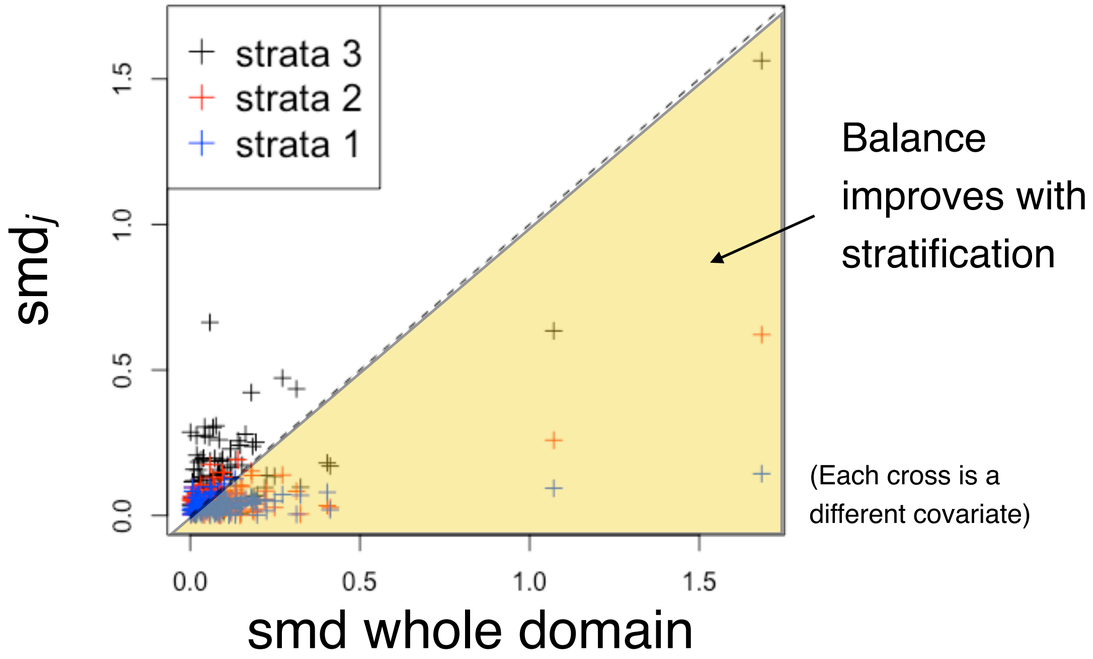
The improvement in covariate balancing between source and target domain is assessed via absolute standardised mean differences (smd) in each "stratum" *j* and covariate x_j:

$$\text{smd}_j \equiv \left| \frac{\bar{x}_{s,j} - \bar{x}_{t,j}}{\sqrt{\frac{\sigma^2_{s,j} + \sigma^2_{t,j}}{2}}} \right|$$

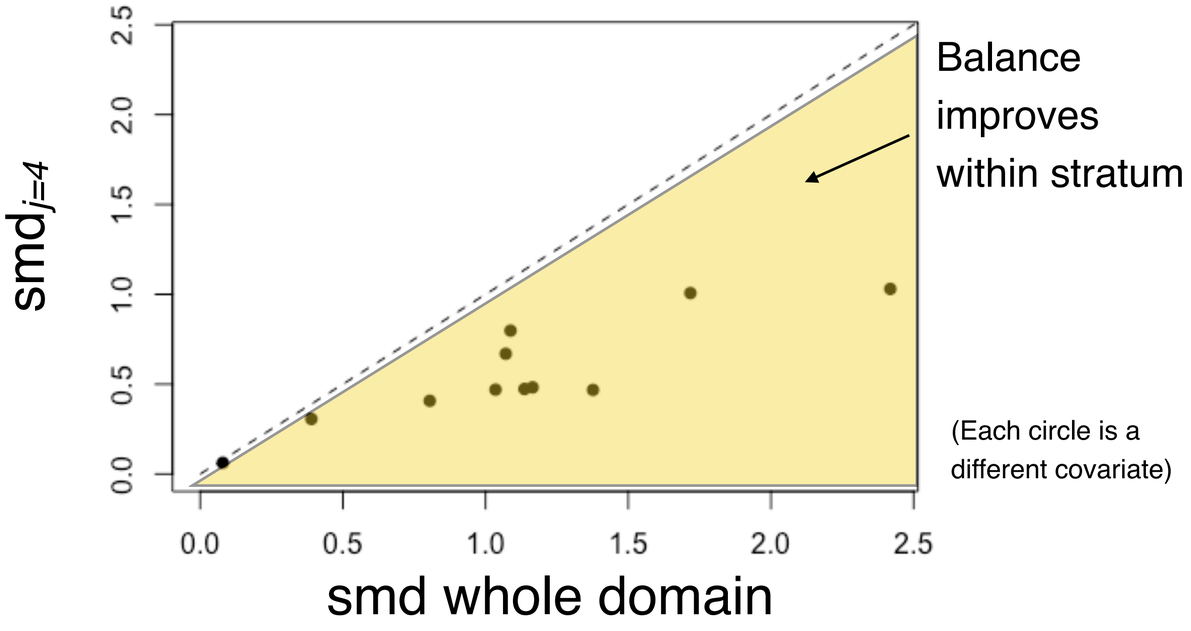
$\bar{x}_{\tau,j}$ = sample mean for $\tau = s, t$

$\sigma_{\tau,j}$ = sample std for $\tau = s, t$

SNIa SPCC Challenge data



UCI Wine data (from Chen et al, 2016)



Conditional Density Estimation Problem

Work by **Max Autenrieth** (Stats PhD student), in collaboration with David van Dyk (Imperial)

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.

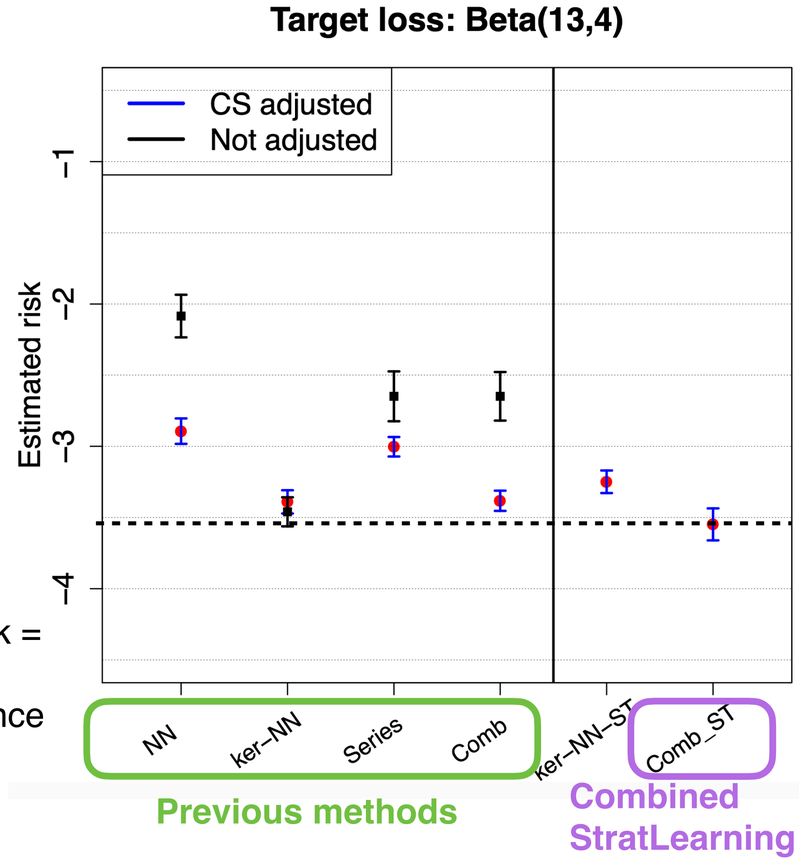
We compare our performance to the estimators in Izbicki et al (2017).

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**)

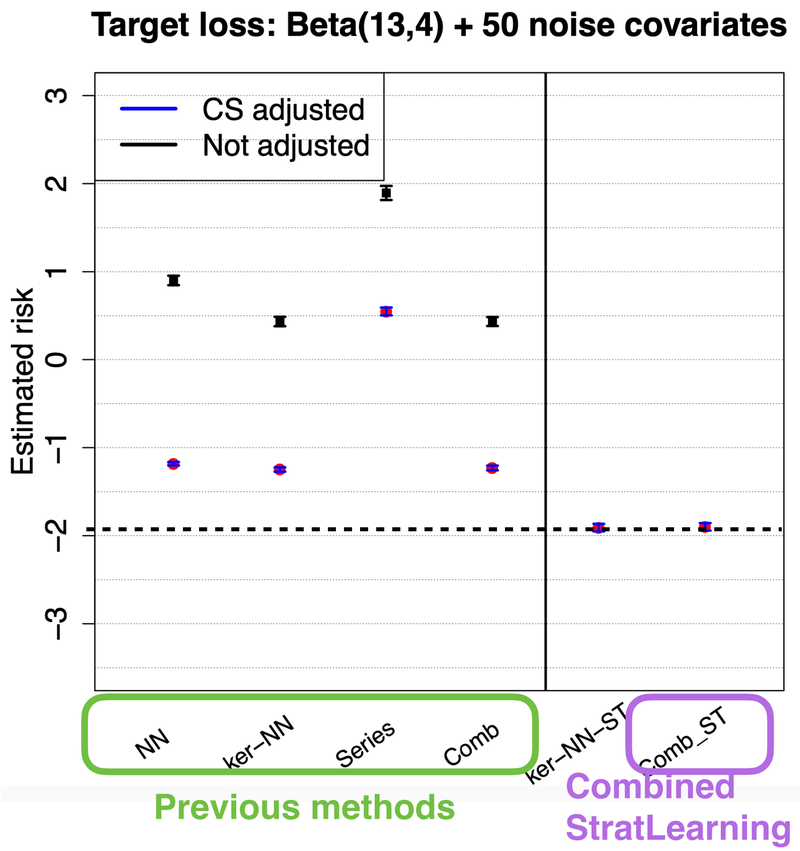
“StratLearning”: Photo-z Performance

Moderate shift
Low covariate dimensions



Lower risk = better performance

Moderate shift
High covariate dimensions



StratLearning outperforms previous methods for this problem.

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

Work by **Max Autenrieth** (Stats PhD student), in collaboration with David van Dyk (Imperial). Paper upcoming.

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 based on stratifying source and target domain according to propensity scores (= probability of an object to be included in the source domain). Methods paper upcoming.

3. Within strata, source and target domains are better balanced: “Stratified Learning” 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.)

Thank you!

www.robertotrotta.com

@R_Trotta

Currently on leave of absence in
Trieste, Italy

Starting up a new
Data Science Initiative

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