LSST DESC & COIN

RESSPECT

Recommendation System for Spectroscopic Follow-up

Project update

Emille E. O. Ishida on behalf of the RESSPECT team

LSST France, 4 November 2020

Fundamental truth about supervised learning:

The performance of any supervised learning algorithm is dependent on the degree of **representativeness** between training and target samples.

Fundamental truth about supervised learning:

The performance of any supervised learning algorithm is dependent on the degree of **representativeness** between training and target samples.

Fundamental truth about SN cosmology with LSST:

Machine learning for photometric classification is unavoidable.

The goal of RESSPECT:

To build a **recommendation system** for the construction of an **optimal training sample** given available spectroscopic resources.

The goal of RESSPECT:

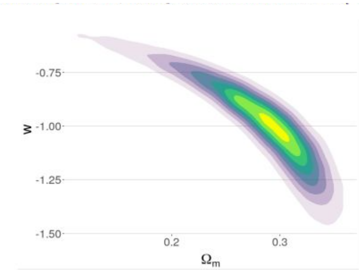
To build a **recommendation system** for the construction of an **optimal training sample** given available spectroscopic resources.

*It is **NOT** the goal of RESSPECT:*

- Build a better classifier
- Maximize the number of spectroscopically confirmed SN Ia
- Test a complete cosmology pipeline
- ...

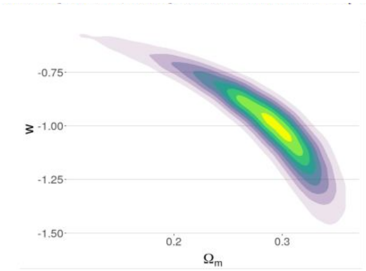
The SN Ia photometric cosmology pipeline

Cosmology results from photometrically classified SN IA



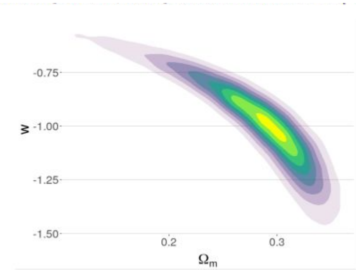
The SN Ia photometric cosmology pipeline

Cosmology results from photometrically classified SN IA



The SN Ia photometric cosmology pipeline

Cosmology results from photometrically classified SN IA



←

Photo-classified SN Ia

←

SN candidates

←

Trained machine learning classifier

Context

The SN Ia photometric cosmology pipeline

Cosmology results from photometrically classified SN IA

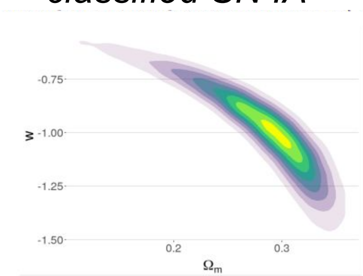


Photo-classified SN Ia

SN candidates

Trained machine learning classifier

learning algorithm

+

Training sample

Context

The SN Ia photometric cosmology pipeline

Cosmology results from photometrically classified SN IA

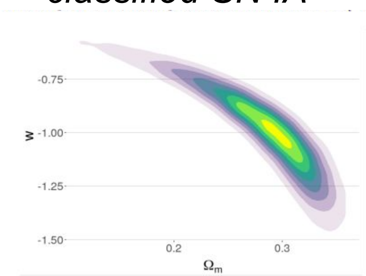


Photo-classified SN Ia

SN candidates

Trained machine learning classifier

learning algorithm

+

Training sample

`Optimize this!`

The SN Ia photometric cosmology pipeline

Cosmology results from photometrically classified SN IA

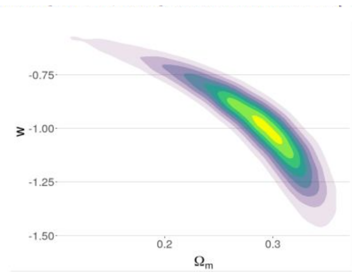


Photo-classified SN Ia

SN candidates

Trained machine learning classifier

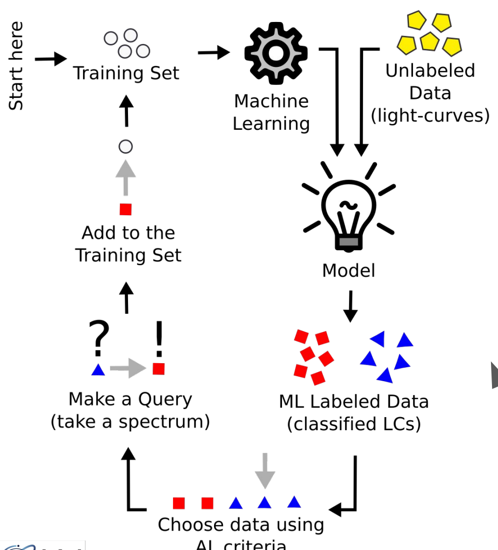
learning algorithm

+

Training sample

Optimize this!

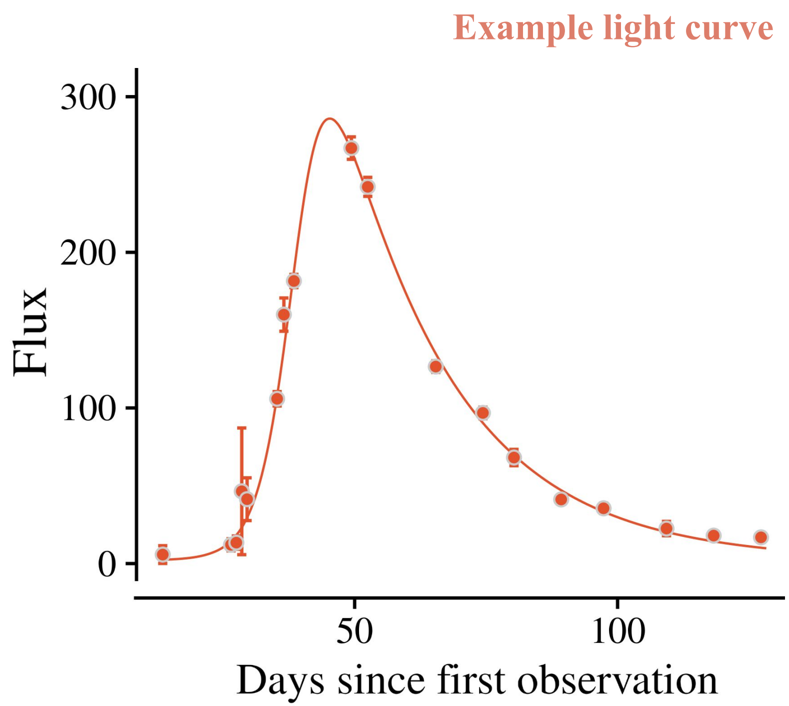
Using this!



Ishida et al., 2019, MNRAS from CRP #4

Take into account observational caveats

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



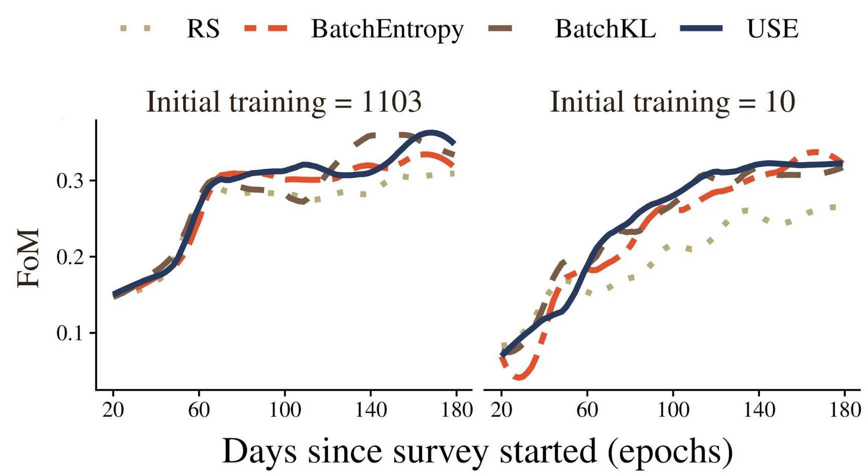
[Kennamer et al., 2020 - arXiv:astro-ph/2010.05941](https://arxiv.org/abs/2010.05941)

Conclusion of phase 1

It is advisable to start from scratch

- Supernova Photometric Classification Challenge data (SNPCC)
- Data separated into four groups
  - Original training set - 1,103
  - 18,216 in pool set
  - 1,000 objects each in validation and test sets
- Assumed access to an 8m and 4m telescope for labeling
  - 6 hours per telescope on each night

- Pre-processed data with parametric fits (Bazin *et al.* 2009)
- Observing Costs calculated from brightness estimates of each objects and telescope properties



Kennamer et al., 2020 - arXiv:astro-ph/2010.05941

Active Learning details

- Ensemble of Random Forest Classifiers for query decisions
- **Four** Active Learning Strategies under **knapsack constraints**:
  - Random Sampling
  - Uncertainty Sampling
    - Entropy used to measure uncertainty
  - Batch Entropy
    - Measures a joint entropy over batches
    - Takes advantage of submodular properties of entropy
  - Batch KL-Divergence
    - Measures a Joint KL-Divergence/Mutual Information, equivalent to BatchBALD
    - Takes advantage of submodular properties of the KL-Divergence/Mutual Information

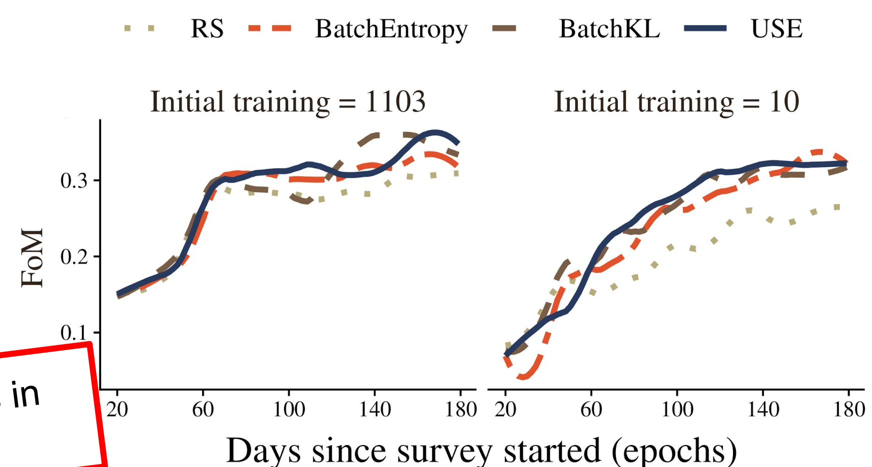
Proved equivalence between KL-Divergence and Bayesian Active Learning by Disagreement (BALD) - check the Appendix!

Kennamer et al., 2020 - arXiv:astro-ph/2010.05941

It is advisable to start from scratch

- Supernova Photometric Classification Challenge data (SNPCC)
- Data separated into four groups
  - Original training set - 1,103
  - 18,216 in pool set
  - 1,000 objects each in validation and test sets
- Assumed access to an 8m and 4m telescope for labeling
  - 6 hours per telescope on each night

- Pre-processed data with parametric fits (Bazin *et al.* 2009)
- Observing Costs calculated from brightness estimates of each objects and telescope properties



Accepted for oral presentation at IEEE Symposium Series in Computational Intelligence, 2020

arXiv:astro-ph/2010.05941

Active Learning part is closed

- Start from scratch
- Simple AL strategies are the best we can do
- Further improvements will require theoretical development

This is only half of the story ...

The cosmology part ...

Cosmology results from photometrically classified SN IA

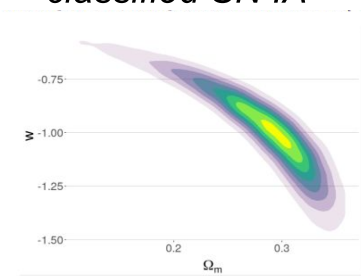


Photo-classified SN Ia

SN candidates

Trained machine learning classifier

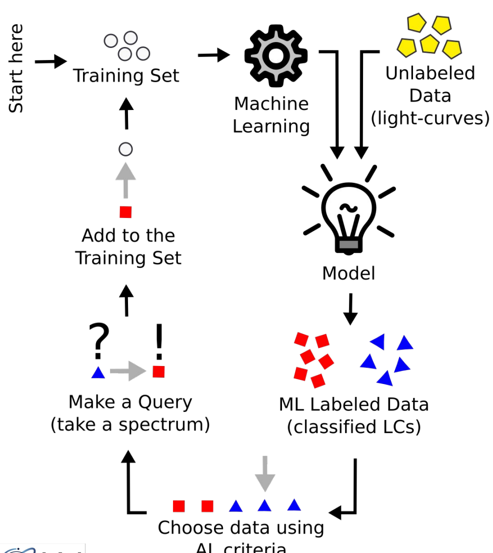
learning algorithm

+

Training sample

This also changes!

Every time this changes ...



COIN Cosmostatistics Initiative

Ishida et al., 2019, MNRAS from CRP #4

This is only half of the story ...

The cosmology part ...

Cosmology results from photometrically classified SN IA

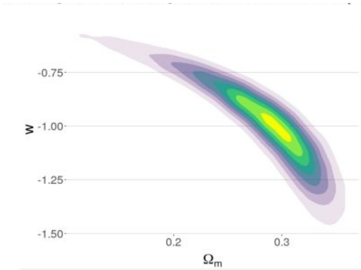


Photo-classified SN Ia

SN candidates

Trained machine learning classifier

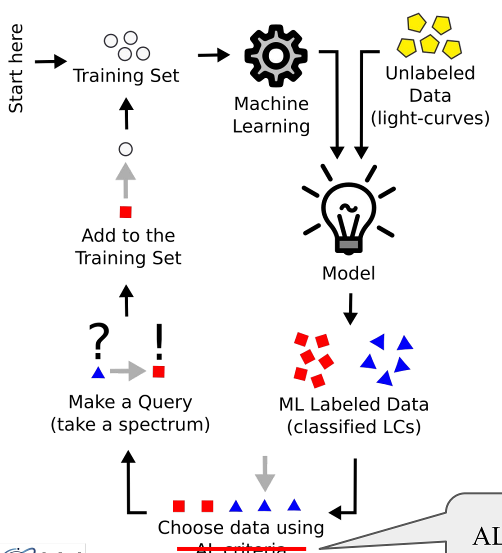
learning algorithm

+

Training sample

This also changes!

Every time this changes ...

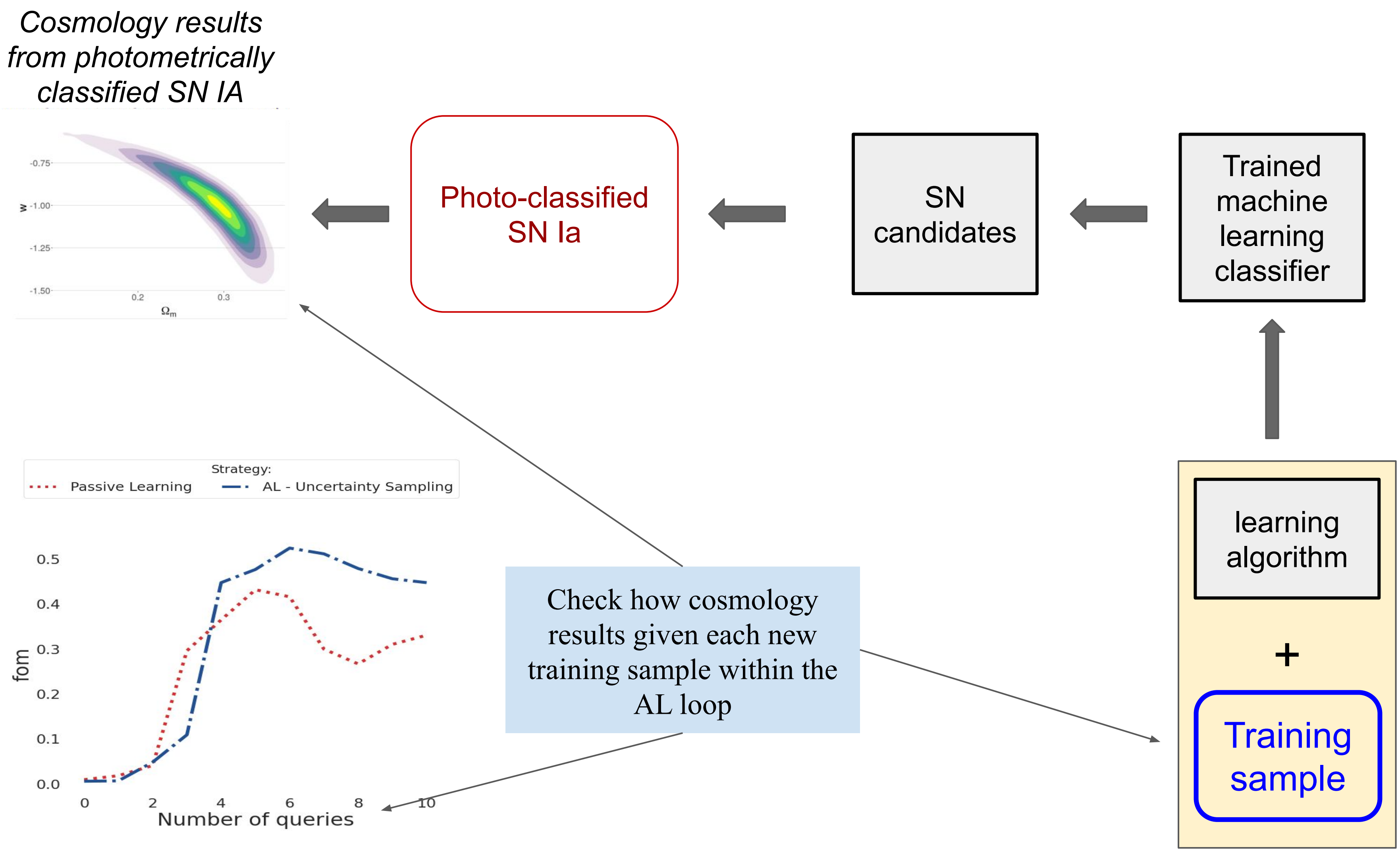


AL + cosmo criteria

Ishida et al., 2019, MNRAS from CRP #4

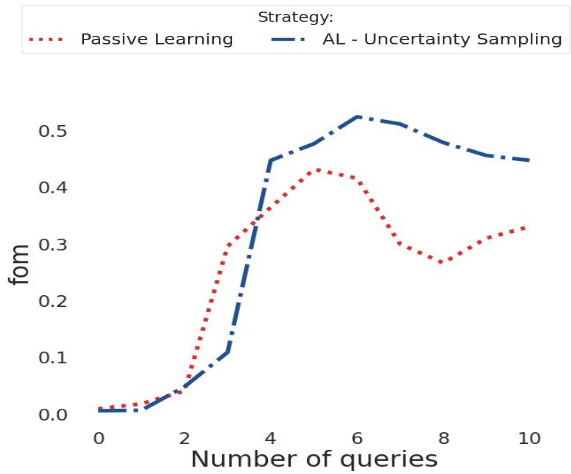
Hypothesis

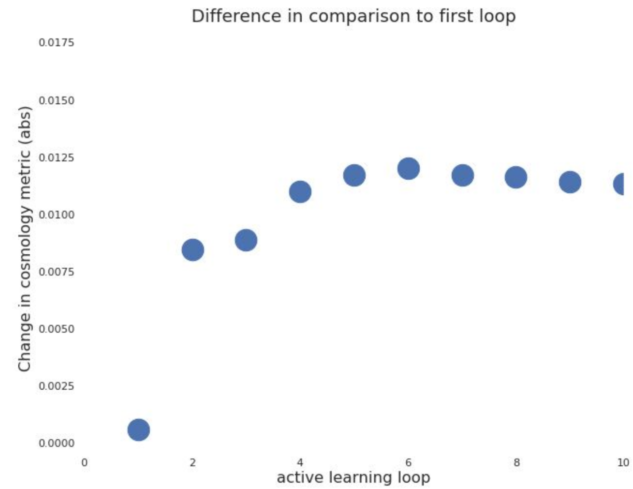
Better classifier leads to a better cosmology



Hypothesis

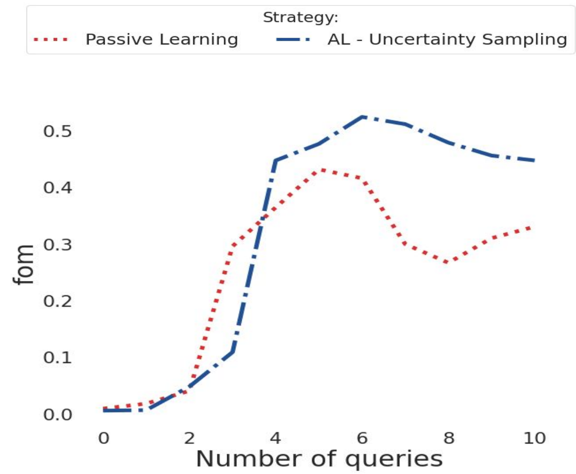
Better classifier leads to a better cosmology

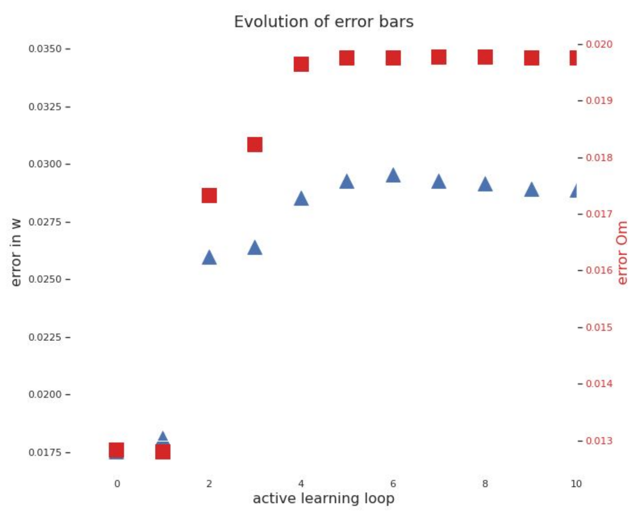




Hypothesis

Better classifier leads to a better cosmology





`Goal:` use the cosmology metric to choose between different **possible batches** per night, which would have the **same effect** from the classifier point of view

Summary and next steps

- Active learning stage is well developed with important theoretical contribution for computer science community
- Currently running the pipeline on PLAsTiCC data for astronomy paper
- Implement the choices between multiple batches in the same run

- Cosmology metric is under development
- Develop a mathematically coherent procedure to combine active learning and cosmology metric

Negotiating extension of ICA until DESC 2021 Summer meeting...

Summary and next steps

- Active learning stage is well developed with important demonstration for computer science community
- Currently running the pipeline on PLAsTiCC data for astronomy paper
- Implement the choices between multiple batches in the same run

- Cosmology metric is under development
- Develop a mathematically coherent procedure to combine active learning and cosmology metric

Interdisciplinarity at its best!

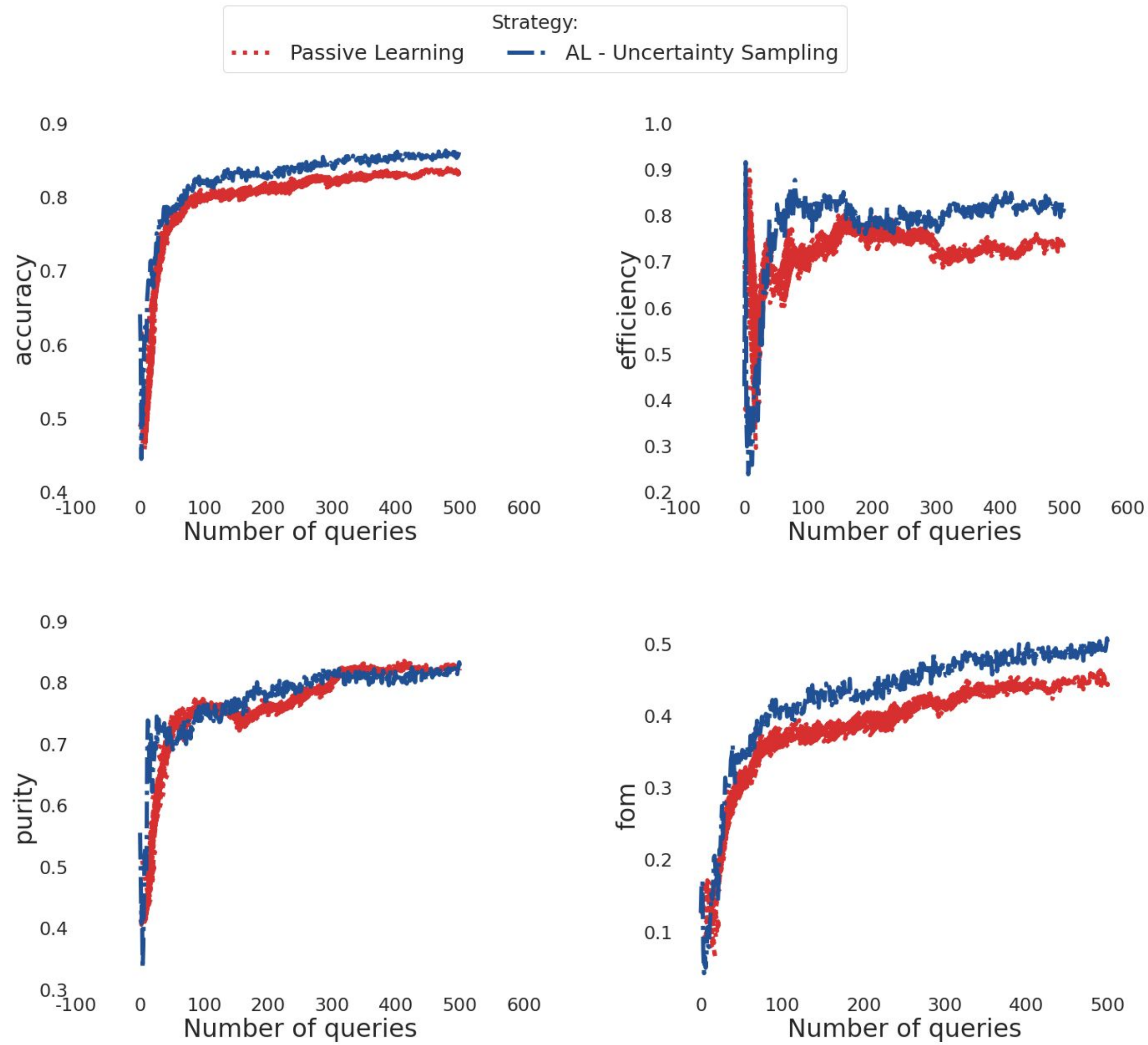
COIN Cosmostatistics Initiative

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

LSST DESC Dark Energy Science Collaboration

Extra slides

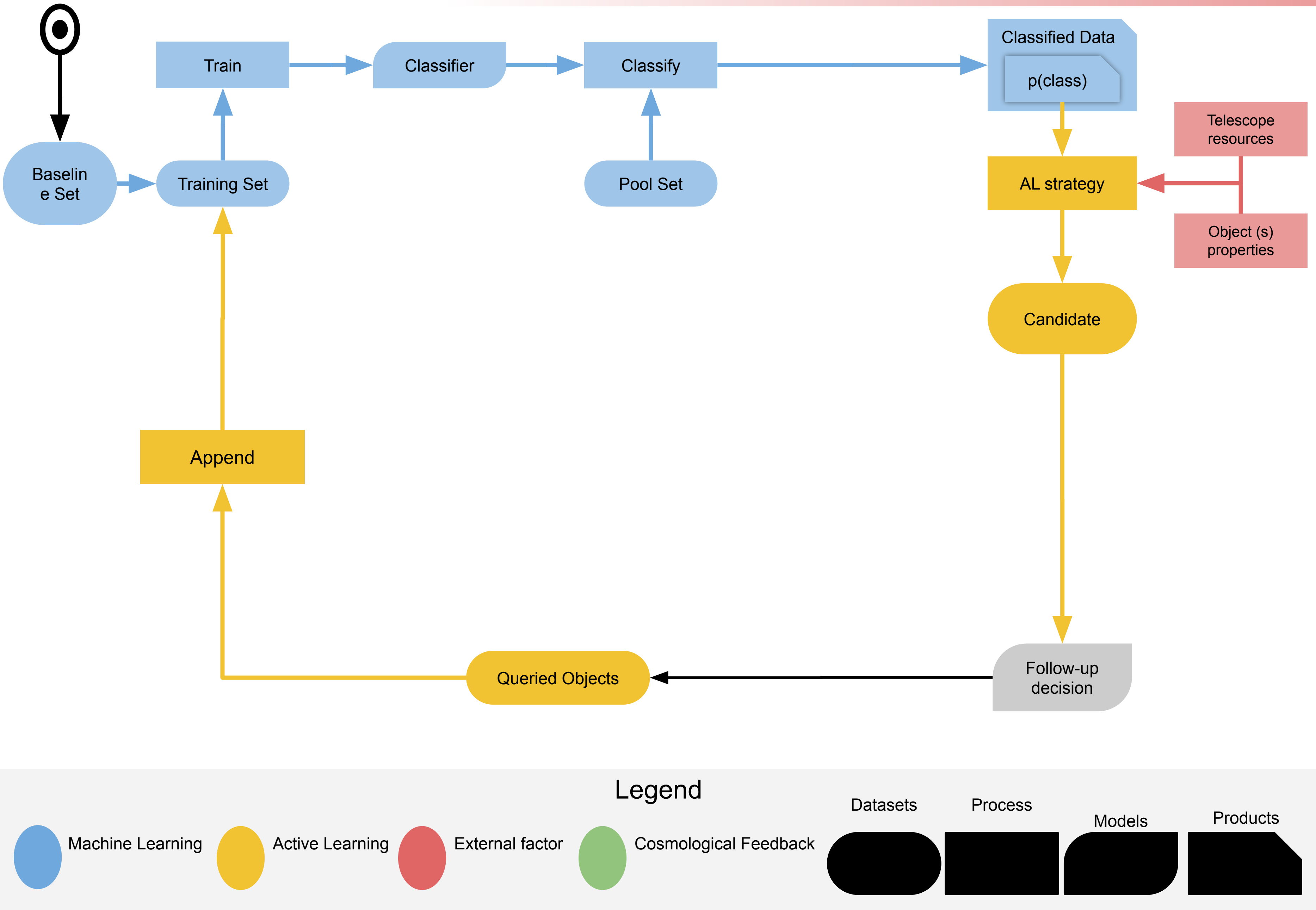
PLAsTiCC results are virtually the same as SNPCC



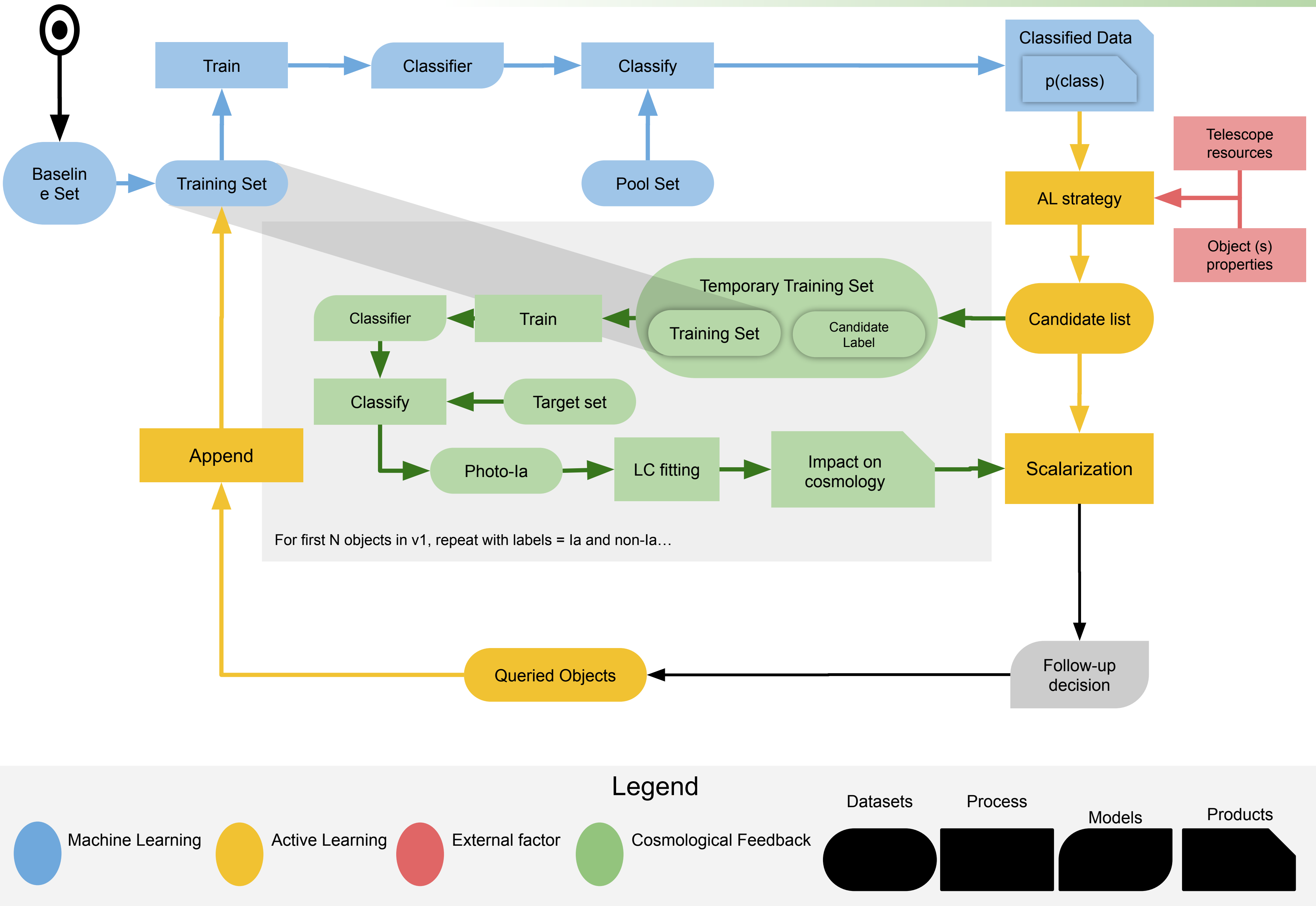
- Initial training of 10 objs
- Validation of 8k objs
- Full LC
- No obs effects considered

Feature extraction takes a while ...

External Factors



Cosmological Feedback



Measuring Disagreement / Query by Committee

Train an ensemble of models on available labeled data

Vote Entropy

$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_y \frac{\text{vote}_{\mathcal{C}}(y,x)}{|\mathcal{C}|} \log \frac{\text{vote}_{\mathcal{C}}(y,x)}{|\mathcal{C}|}$$

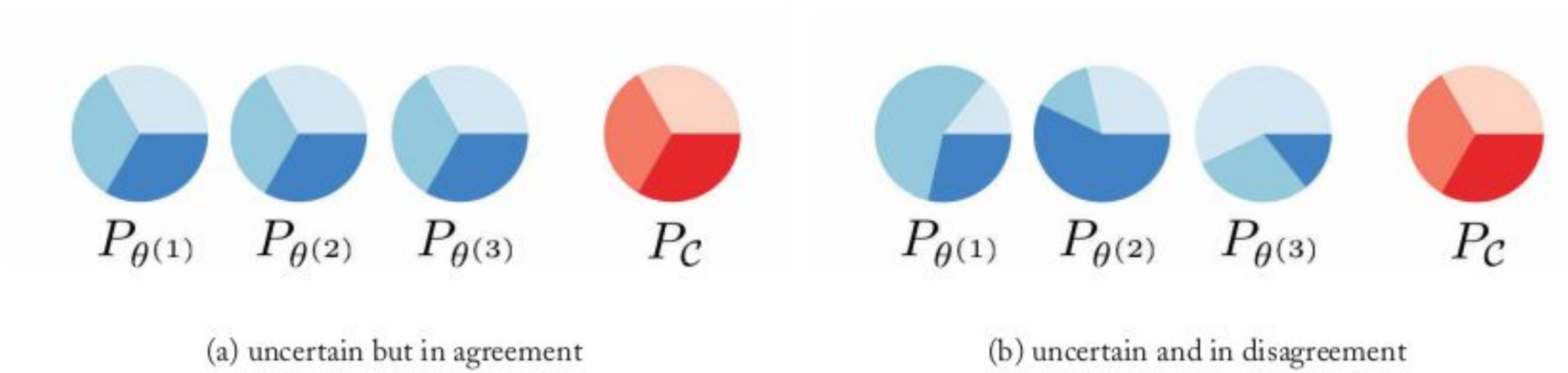
Soft Vote Entropy

$$x^*_{SVE} = \underset{x}{\operatorname{argmax}} - \sum_y P_{\mathcal{C}}(y|x) \log P_{\mathcal{C}}(y|x),$$

Kullback-Leibler divergence

$$x^*_{KL} = \underset{x}{\operatorname{argmax}} \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} KL\big(P_\theta(Y|x) \parallel P_{\mathcal{C}}(Y|x) \big)$$

Entropy vs KL Divergence



- Equal Entropy
- Low KL

- Equal Entropy
- High KL

Image Source: Settles 2012 Claypool Publishers