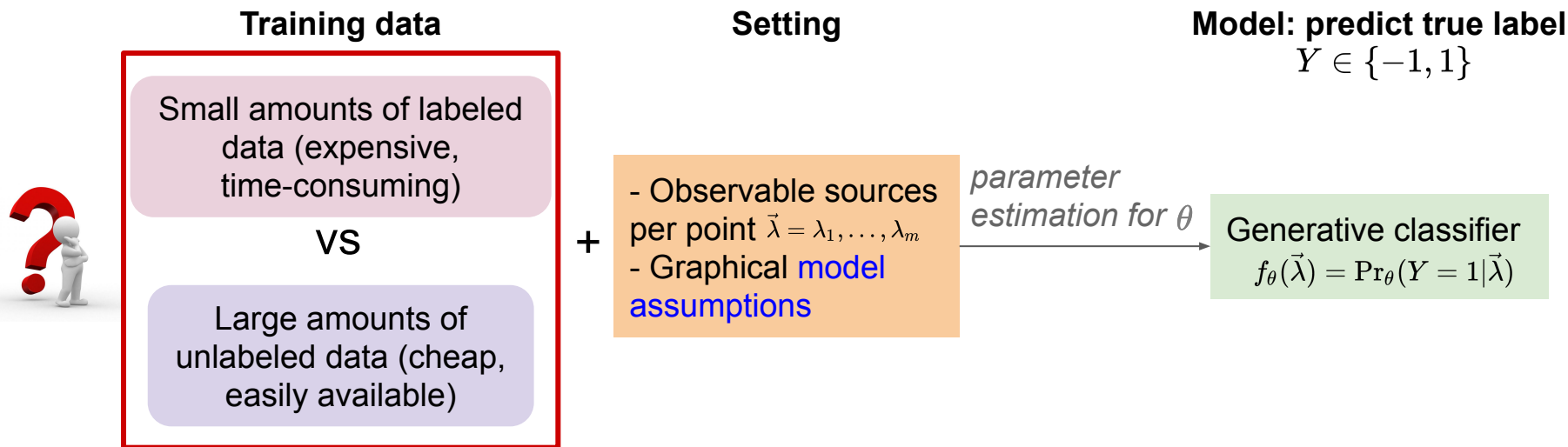


The value of labeled vs unlabeled data in latent variable graphical models

Mayee Chen, Ben Cohen-Wang, Steve Mussmann, Fred Sala, Chris Ré

The tradeoff between labeled and unlabeled data

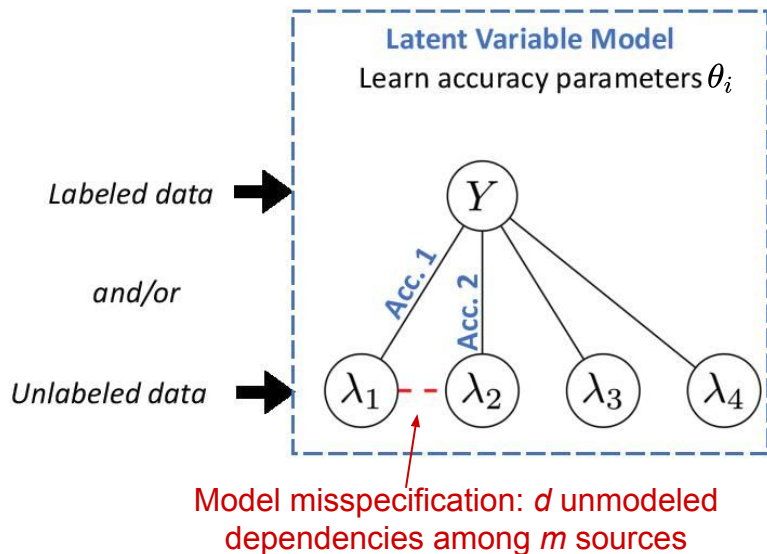


Our results:

A theoretical decision framework based on comparing generalization error of $f_{\theta}(\vec{\lambda})$ to understand the tradeoffs between using labeled or unlabeled data.

- When **model assumptions** are incorrect, labeled data tends to be better → present a general way of correcting model misspecification to improve unlabeled data performance

Theoretical decision framework



Error Decomposition of Classifier:

Generalization error =

Irreducible error + sampling noise +
inference bias + **parameter estimation error**

Only term that differs in the two cases!

For labeled data: goes to 0

For unlabeled data: $\mathcal{O}(d/m)$

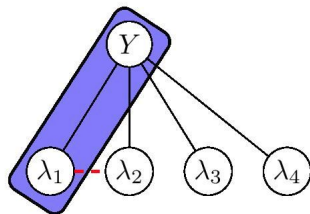
Labeled data is generally better to use when model misspecification is unaddressed.

Hope: fix model misspecification → can unlabeled data be better to use?

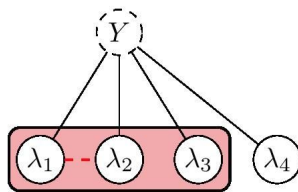
Correcting misspecification in unlabeled case: medians estimate

Example:
estimating θ_1

Labeled case

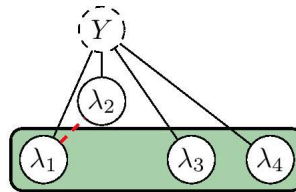


Unlabeled case



inconsistent: conditional independence violated

Corrected case



$\lambda_1, \lambda_2, \lambda_3 \rightarrow 0.73$ **inconsistent**

$\lambda_1, \lambda_3, \lambda_4 \rightarrow 0.78 = \hat{\theta}_1$

$\lambda_1, \lambda_2, \lambda_4 \rightarrow 0.81$ **inconsistent**

Select **median**
accuracy parameter

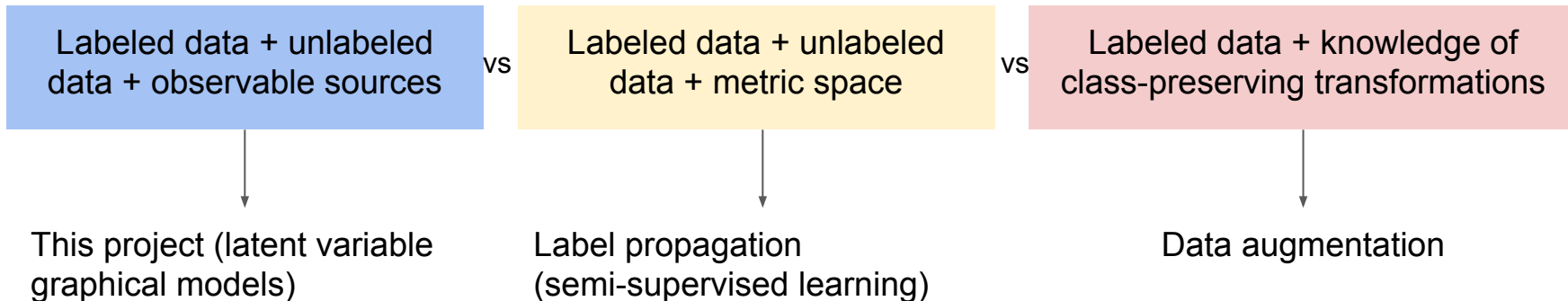
- When enough unlabeled data, removes $\mathcal{O}(d/m)$ asymptotic bias $\rightarrow$ unlabeled data more “valuable” now
- This is true for general method of moments estimators!

*Adding a little labeled data can further close the performance gap compared to a large labeled dataset

Application: Weak Supervision (40K points)	F1 score
Labeled	71.79
Unlabeled	64.81
Corrected unlabeled	68.12
Corrected unlabeled + 1% extra labeled points*	71.04

Future Work

More general frameworks for choosing between different **sources of supervision**



Theoretical guidance for practitioners building models by quantifying tradeoffs between number of samples and how much information sources provide - and how this changes when model assumptions are wrong.