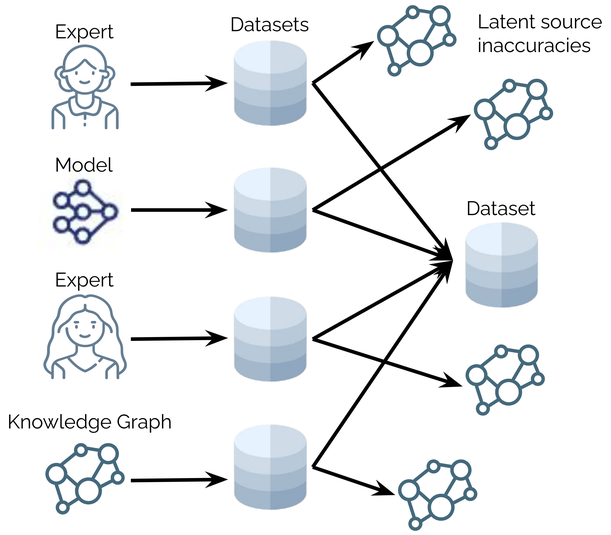
Building Blocks for Synthesizing From Multiple Sources

Putra Manggala (supervised by Holger H. Hoos and Eric Nalisnick)

Goal: Synthesize from multiple data-generating processes/datasets to solve downstream decision-making tasks

Aggregate: create a new dataset that take into account source inaccuracies

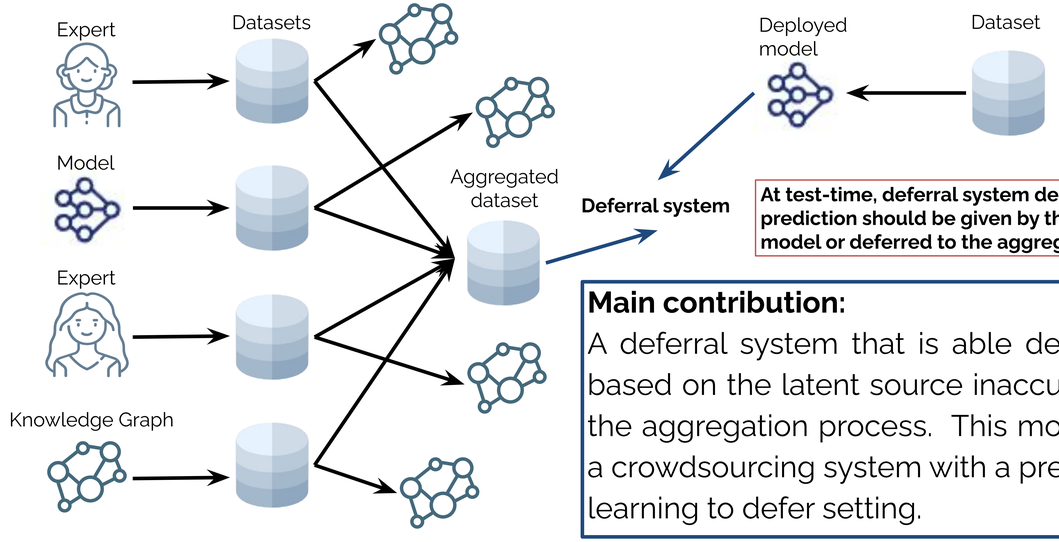


Main contribution:
A generative model that infers the latent sources' inaccuracies and ground truth from probabilistic datasets using optimal transport.

Probabilistic dataset: each instance is labeled probabilistically (soft label).

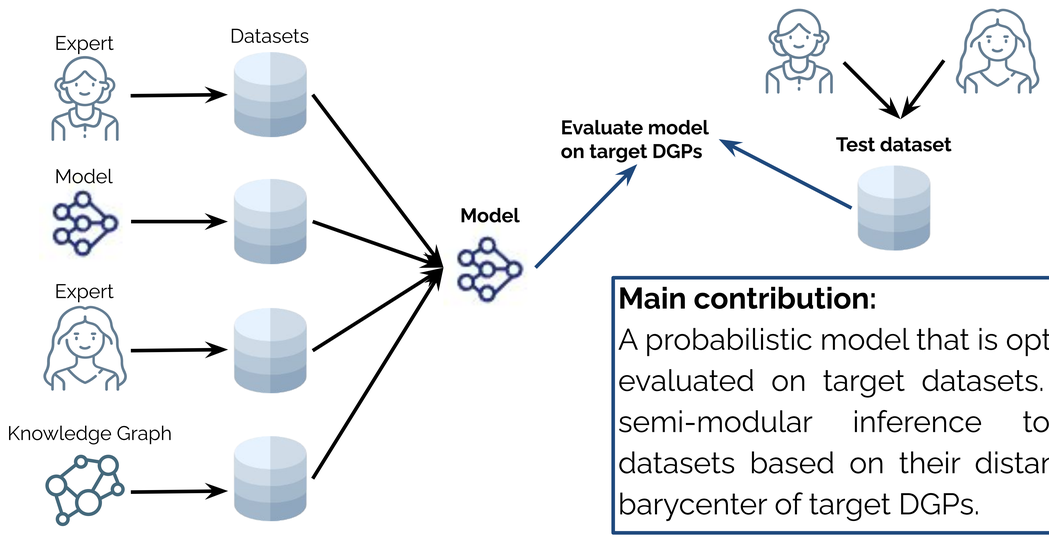
Preliminary works presented in TPM 2022 and HMCaT 2022.

Defer: create a model that is capable of deferring to aggregated dataset



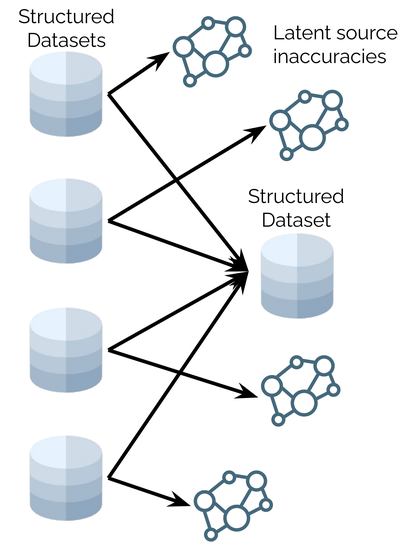
Main contribution:
A deferral system that is able decide on deferrals based on the latent source inaccuracies inferred by the aggregation process. This modularly integrates a crowdsourcing system with a predictive model in a learning to defer setting.

Learn: create a model that is optimal for target data-generating processes



Main contribution:
A probabilistic model that is optimal when evaluated on target datasets. We use semi-modular inference to temper datasets based on their distance to the barycenter of target DGPs.

Aggregate various modalities



Main contribution:
(An extension of upper-left quadrant)

We extend the underlying optimal transport approach to handle various label modalities, such as ranking, text and graph..

This requires a substantial change in the representation of latent source inaccuracies. While in the categorical case, a confusion/cost matrix is typically used to model a source's inaccuracies, another data structure is needed to represent inaccuracies for structured data.