

Knowledge Distillation From Big Administrative Data

Ayshah Chan, Marco Körner

Background

Administrative data offers a *huge repository of data* that can be used to *further data-driven modelling* but struggles with issues with *data alignment* and *harmonization*.

Foreseen Challenges

Existing models rely on sequential sentences and paragraphs whereas administrative data contain *tabularized short phrases and numbers*. Significant computation requirements call for *more efficient training methods*.

Expected Outcome

A *foundation model* trained on *administrative data* that will enable the *creation of a wide variety of training data* and *further applications* in different domains.

