Self Generation of Policies for Training Data Curation

A. Abu Jabal (Purdue), **S. Witherspoon, I. Manotas (**IBM US**), E. Bertino (**Purdue**), S. Calo, S. Chakraborty, D. Verma (**IBM US**), G. Cirincione, A. Swami (**ARL**), G. De Mel (**IBM UK**), G. Pearson (Dstl**).

Demonstration Context

Building Machine Learning Models in a Coalition Environment:

- Data available from different partners.
- Not all partners are trusted equally.
- Untrusted partners may have useful data that can help make better models.
- Data acceptance policies determine which data to accept and which to reject.
- Writing policies manually is hard because it depends on the contents of the data.

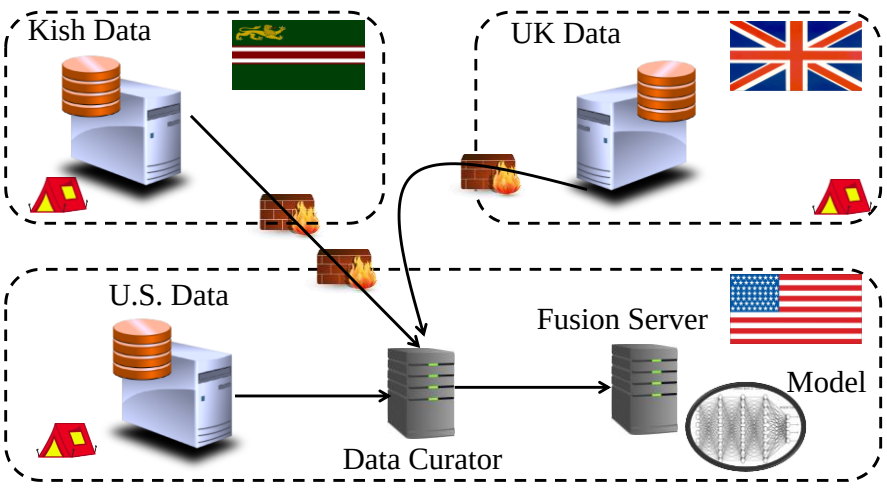

Figure 1: Data Sharing Scenario for Supporting Training of Machine Learning Models

Data Curator

A software that cleans, transforms and collects training data from partners.

Uses policies to guide its operation:

- Sequencing Policies.
- Data Acceptance Policies.

Manual Policies

Can only accept/reject data from a partner.

Needs to rely on preconceived notions of which partner is trusted.

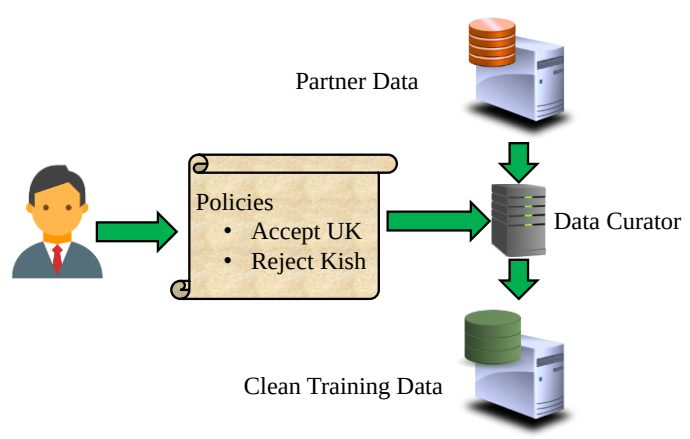

Figure 2: Manual Data Curation Policies

Self-Generated Policies

- Created after examining data from partners.
- Can decide on fine-granularity of acceptance.
- Can examine QoI and VoI of partner data.

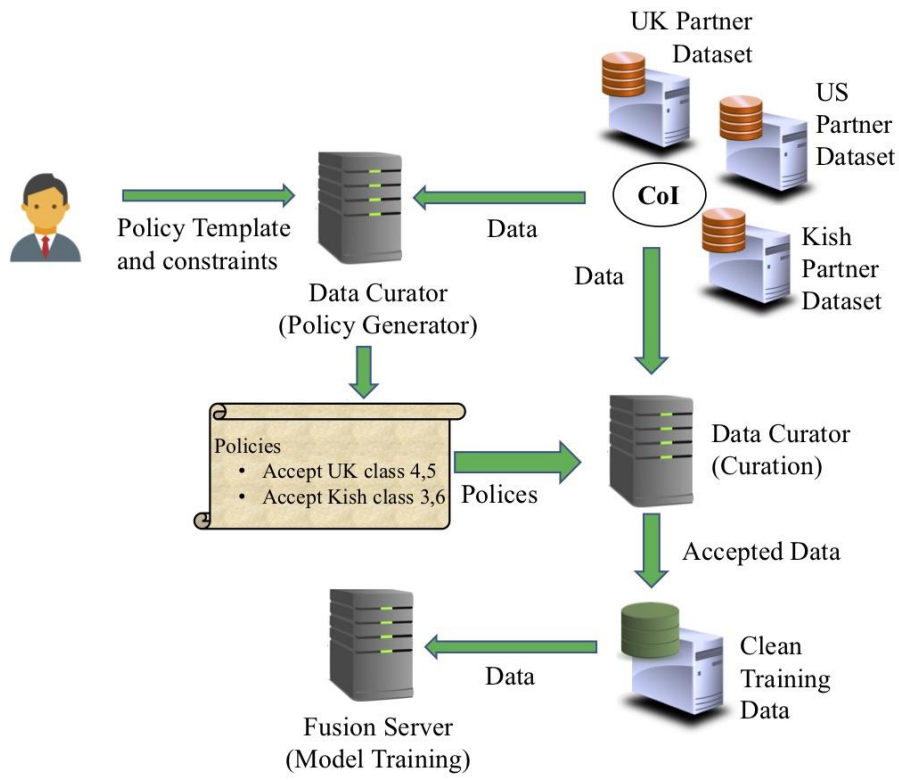

Figure 3: Self-Generated Policies

Self-generation of policies for data curation results in better training data and better model.