How machine learning is accelerating data integration?

Abhishek Tiwari

Citation: *A. Tiwari*, "How machine learning is accelerating data integration?", Abhishek Tiwari, 2017. doi:10.59350/26y6f-65m22

Published on: December 24, 2017

Data integration generally requires in-depth domain knowledge, a strong understanding of data schemas and underlying relationships. This can be time-consuming and bit challenging if you are dealing with hundreds of data sources and thousands of event types (see my recent article on ELT architecture). Various data integration solution providers are trying to capitalize on this gap by offering various machine learning based features to overcome these challenges.

Automated entity relationship modeling

Data integration platforms such as Panoply, Informatica, and Tamr have applied machine learning techniques to automate the schema modeling process. These platforms are not only leveraging machine learning but also natural language processing to discover the underlying entities, and model the entity relationships with minimal human intervention.

The machine learning process starts with the discovery of domain data types for attributes (aka columns or fields). Then machine learning engine assembles these individual attributes into higher-level business entities. Next, the formation of entity relationships takes place.

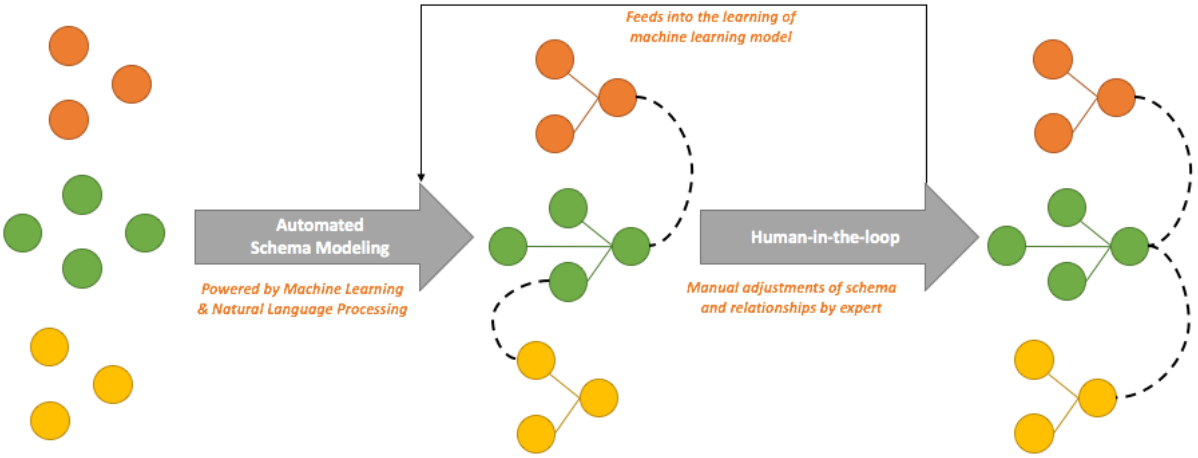

Figure 1: Bottom-up, machine learning and natural language processing based approach for schema or entity relationships modeling with the human-in-the-loop.

As you expect machine learning can perform 90% of work required for the entity discovery and entity relationship mapping associated with the development of data warehouse, data lake, and search indexes. Rest 10% manual adjustments and customizations can be performed by a human. Machine learning models can construct entities and associated relationships purely based on the data and signals embedded deep inside the data, or alter the existing entities and relationships in real-time to adopt against the addition of a new data source all while utilizing a probabilistic/statistical model.

The discovery of domains and data type for attributes is a classification problem. For example, you can classify attributes as email, zip code, street, state, country, first name, last name, price, etc with a very high accuracy. Similarly, clustering is used to group the individual attributes. So first name and last name can be grouped as customer entity purely by computing data similarity using the Bray-Curtis and Jaccard coefficients

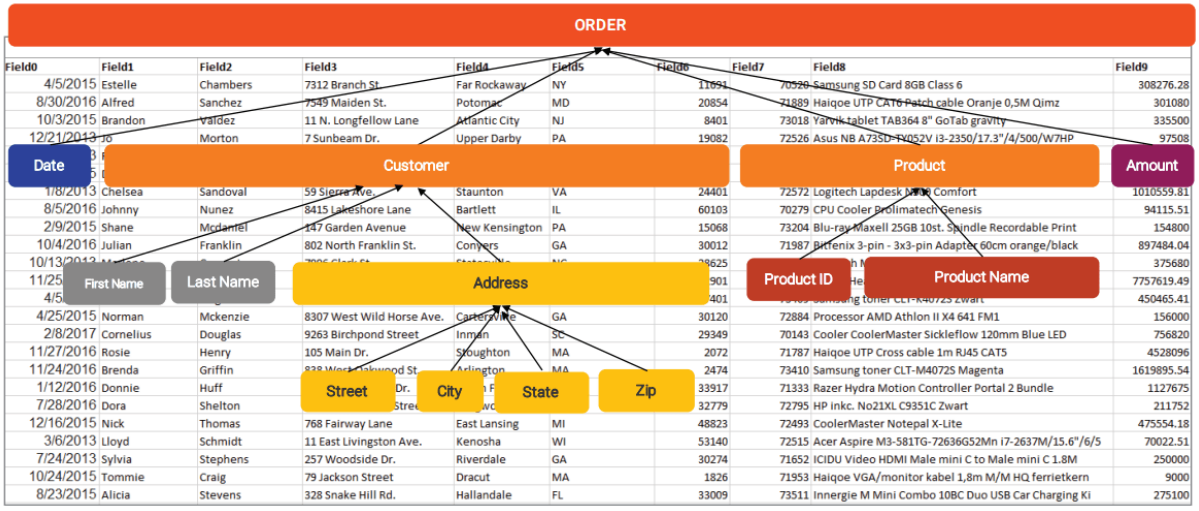

Figure 2: Automated entity discovery relies on accurate data type classification for attributes. Image credits Informatica.

In a nutshell, machine learning greatly accelerates schema modeling and remodeling required for the data integration to an extent that schema can be optimized based on query or data access pattern.

Intelligent data recommendations

Machine learning techniques have been also applied to intelligent data recommendations. Platforms like SnapLogic and Informatica can recommend next-best-action or suggest datasets, transforms, and rules. It is possible to make recommendations on potential transformations like - aggregations, joins and unions. Snaplogic reports up to 80-90% accuracy on intelligent data recommendations which is expected to improve with feedback loop provided by manual adjustments and customizations performed by a human.

Intelligent data recommendations are also quintessential to support the self-service data access model i.e. direct, easy and timely access to data to anyone. For instance, data scientists can get recommendations on which data sets to use for their projects or suggestions on additional data sets that may complement their existing data sets. With these recommendations, data scientists can now develop a strong understanding of business and domain data more quickly than ever.

Tip of the iceberg

We are aware that data integration providers are already applying machine learning to solve problems like anomaly detection and structure discovery, but the holy grail remains fully-automated data integration. What we have seen so far is a combination of supervised and unsupervised learning applied to solve specific high-impact problems with a feedback mechanism using human-in-the-loop.