## 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

ORDER								
Field0	Field1	Field2	Field3	Fielda	Fields	Fields F	Field7 Field8	Field9
4/5/201	5 Estelle	Chambers	7312 Branch St	Far Rockaway	NY	11091	70520 Samsung SD Card 8GB Class 6	308276.28
8/30/201	6 Alfred	Sanchez	7549 Maiden St.	Potomac	MD	20854	71889 Haigoe UTP CAT6 Patch cable Oranje 0,5M Qimz	301080
10/3/201	5 Brandon	Valdez	11 N. Longfellow Lane	Atlantic City	NJ	8401	73018 Yarvik tablet TAB364 8" GoTab gravity	335500
12/21/201	or <del>C</del>	Morton	7 Sunbeam Dr.	Upper Darby	PA	19082	72526 Asus NB A73SD-TX052V 13-2350/17.3"/4/500/W7HP	97508
Date	3 F 5 C		Customer				Product	Amount
1/8/201	3 Chelsea	Sandoval	59 Sierre Ave.	Staunton	VA	24401	72572 Logitech Lapdesk 🖉 Comfort	1010559.81
8/5/201	6 Johnny	Nunez	8415 Lakeshore Lane	Bartlett	IL.	60103	70279 CPU Cooler Prolimatech Genesis	94115.51
2/9/201	5 Shane	Mcdamel	147 Garden Avenue	New Kensington	PA	15068	73204 Blu-ray Maxell 25GB 10st. Spindle Recordable Print	154800
10/4/201	6 Julian	Franklin	802 North Franklin St.	Convers	GA	30012	71987 Bittenix 3-pin - 3x3-pin Adapter 60cm orange/black	897484.04
10/13/201	2		700C Clark Ct	States We	NC	<sup>28</sup> 625	h N	375680
11/25 F	First Name	Last Name	-	Address		901 401		7757619.49 450465.41
4/25/201	5 Norman	Mckenzie	8307 West Wild Horse Ave.	Cartersville	GA	30120	72884 Processor AMD Athlon II X4 641 FM1	156000
2/8/201	7 Cornelius	Douglas	9263 Birchpond Street	Inman	SC	29349	70143 Cooler CoolerMaster Sickleflow 120mm Blue LED	756820
11/27/201	6 Rosie	Henry	105 Main Dr.	Stoughton	MA	2072	71787 Haigoe UTP Cross cable 1m RJ45 CAT5	4528096
11/24/201	6 Brenda	Griffin	838 Wost Dakwood St.	Arlington	МΔ	2474	73410 Samsung toner CLT-M4072S Magenta	1619895.54
1/12/201	6 Donnie	Huff	Street Dr. C	itv <sup>16</sup> Sta	te	Zip 33917	71333 Razer Hydra Motion Controller Portal 2 Bundle	1127675
7/28/201	6 Dora	Shelton	Stree	NC		32779	72795 HP inkc. No21XL C9351C Zwart	211752
12/16/201	5 Nick	Thomas	768 Fairway Lane	East Lansing	MI	48823	72493 CoolerMaster Notepal X-Lite	475554.18
3/6/201	3 Lloyd	Schmidt	11 East Livingston Ave.	Kenosha	WI	53140	72515 Acer Aspire M3-581TG-72636G52Mn i7-2637M/15.6"/6/5	70022.51
7/24/201	3 Sylvia	Stephens	257 Woodside Dr.	Riverdale	GA	30274	71652 ICIDU Video HDMI Male mini C to Male mini C 1.8M	250000
10/24/201	5 Tommie	Craig	79 Jackson Street	Dracut	MA	1826	71953 Haiqoe VGA/monitor kabel 1,8m M/M HQ ferrietkern	9000
8/23/201	5 Alicia	Stevens	328 Snake Hill Rd.	Hallandale	FL	33009	73511 Innergie M Mini Combo 10BC Duo USB Car Charging Ki	275100

**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.