

# MACHINE LEARNING

with



## Introduction to Association Discovery

**BigML Inc.**  
2851 NW 9th,  
Corvallis, OR 97330, U.S.  
<https://bigml.com/>  
[info@bigml.com](mailto:info@bigml.com)



1. Prerequisites & Learning Goals	4
2. Introduction	5
3. Use Case Example & Data Source	7
4. Data Preparation	8
5. Analysis & Modeling	11
6. Results	18
7. Conclusion	24



## 1. PREREQUISITES & LEARNING GOALS

This guide is intended as an introductory material for beginners learning the ropes with the Association Discovery technique. It assumes no prior background in Machine Learning although basic understanding of the concepts would definitely help enhance the reader's level of comprehension.

If you would like to get a crash course on the general Machine Learning concepts and the accompanying process, we highly recommend that you download and read our [Machine Learning Primer](#) guide (coming soon).

The reader will also need a BigML account to practice with the use cases outlined. In order to help more people experience and get better with Machine Learning, BigML offers FREE subscriptions for Machine Learning tasks involving less than 16MB of data, which is enough for most of the exercises included in our guides. In addition, as a sign of support for creative new research and development in the fields of Machine Learning and Predictive Analytics, we offer FREE PRO subscriptions to educators and students. This offer requires that the user sign up with his/her .edu email account.

At the end of this guide, the reader should be able to upload data, create an Association Discovery ready dataset out of it, and finally be able to run and interpret his/her own Association Discovery tasks in an iterative manner all through the use BigML's browser-based user interface.



## 2. INTRODUCTION

Association Discovery (aka Association Rule Learning) is a Machine Learning technique used to identify statistically significant relations between data elements. It can yield sets of interesting rules that may lead to useful business insights locked up in large multidimensional datasets containing many variables each of which may entail many different values be they numeric, categorical or text values.

Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to market basket analysis based on Point of Sale data association rules are employed today in many application areas including Web usage mining, intrusion detection, continuous production, and bioinformatics. In contrast with sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.<sup>1</sup>

Before we proceed with our use case we would like to answer the frequently asked question “What makes BigML’s Association Rule Discovery different than other Association Discovery products?”

Originally developed by Professor Geoff Webb who heads the Centre for Data Science at Monash University of Melbourne, BigML’s implementation of this well researched data analysis technique sets itself apart because it

- Concentrates on discovering relationships between values rather than merely variables. This is a non-trivial distinction. If someone is told that there is an association between gender and some medical condition, they are likely to immediately wish to know which gender is positively associated with the condition and which is not. Association mining zeroes in on this critical question of interest. Furthermore, associations between values, rather than variables, can be more powerful (discover weaker relationships) when variables have more than two values.
- Strictly controls the risk of making false discoveries through spurious associations. A serious issue inherent in any attempt to identify associations with classical methods is

---

<sup>1</sup> [https://en.wikipedia.org/wiki/Association\\_rule\\_learning](https://en.wikipedia.org/wiki/Association_rule_learning)



an extreme risk of false discoveries. These are apparent associations that are in fact only artifacts of the specific sample of data that has been collected. As such, BigML offers the only commercial association discovery software to provide strict statistical control over the risk of making any such errors. In fact, BigML's Association Discovery lets the user specify a statistical significance level, which is set to 0.05 by default.

- Scales very effectively to high-dimensional data. The standard statistical approach to categorical association analysis (i.e. log-linear analysis) has complexity that is exponential with respect to the number of variables. In contrast, association mining techniques can typically handle many thousands of variables.
- Avoids the problems due to model selection. Most data mining techniques produce a single global model of the data. A problem with such a strategy is that there will often be many such models, all of which describe the available data equally well. Association mining can find all local models rather than a single global model. This empowers the user to select between alternative models on grounds that may be difficult to quantify for a typical statistical system to take into account.



### 3. USE CASE EXAMPLE & DATA SOURCE

To demonstrate how you can take advantage of BigML's proprietary Association Discovery feature, we will use the Home Theatre Info dataset containing movie metadata information on over 250,000 DVDs offered in North America. This dataset comes in multiple files in CSV, Excel and HTML formats including

- DVD List File
- Directors File
- Actors File

These files can be downloaded free of charge from <http://www.hometheaterinfo.com/dvdlist.htm>.

For the record, BigML has the ability to import

- CSV (Comma or Tab Separated)
- Excel
- Gzipped (.gz) or compressed (.bz2)
- ARFF
- Direct URLs
- Amazon S3 Buckets,
- Dropbox,
- Google Cloud or Google Storage
- Azure files.

For the sake of this exercise, we will utilize CSV formatted files from the aforementioned Home Theatre Info web site.



## 4. DATA PREPARATION

Data Preparation is a key part of any Machine Learning flow, but unfortunately it usually gets the short shrift due to its unsexy nature. In order to get the reader acquainted with this critical step, we will briefly explain the steps that we took to pre-condition our raw data files before we uploaded it to BigML.

This data comes in separate files that we upload to a MySQL database for some data wrangling. The structure of each file is seen below:

### 1. DVD List Table

- ID
- DVD Title
- Studio
- Released
- Status
- Sound
- Versions
- Price
- Rating
- Year
- Genre
- Aspect
- UPC
- DVD Release Date





- Time Stamp

## 2. Actors Index Table

- ID
- Actor ID
- DVD List ID

## 3. Directors Index Table

- ID
- Director ID
- DVD List ID

## 4. Actors Table

- Actor ID
- Actor Name (Name (Last Name, First Name))

## 5. Directors Table

- Director ID
- Director Name (Last Name, First Name)

As usual with a Machine Learning project, we had to do some data wrangling using MySQL in order to transform those tables to a Machine Learning ready format. This involved a combination of data audits, feature selection, de-normalization and feature engineering.<sup>2</sup>

---

<sup>2</sup> For more on what these terms mean we suggest beginners download our Machine Learning Primer guide.)



At the end of our data wrangling joining the three original tables we arrive at the following data table after eliminating less promising data fields such as DVD Aspect Ratio based on our subjective opinion:

#### DVDs Table

- ID
- Studio
- Rating
- Year
- Genre
- Director ID
- Actor IDs (Data field containing the IDs for up to 10 actors taking part in the movie separated by semicolons e.g. 123; 345; 555; 678; etc.)

Please note that the decision to combine all Actor IDs in a single field was not arbitrary, but was a conscious choice as it allows for less association rules with either the Antecedent or the Consequent conditions involving missing values e.g. Rating = Missing -> Genre = Missing etc.

We also generated a simpler and more streamlined table of Director ID and Actor ID pairs as follows by combining Actor and Director Index tables:

#### Actor Director Table

- Director ID
- Actor ID

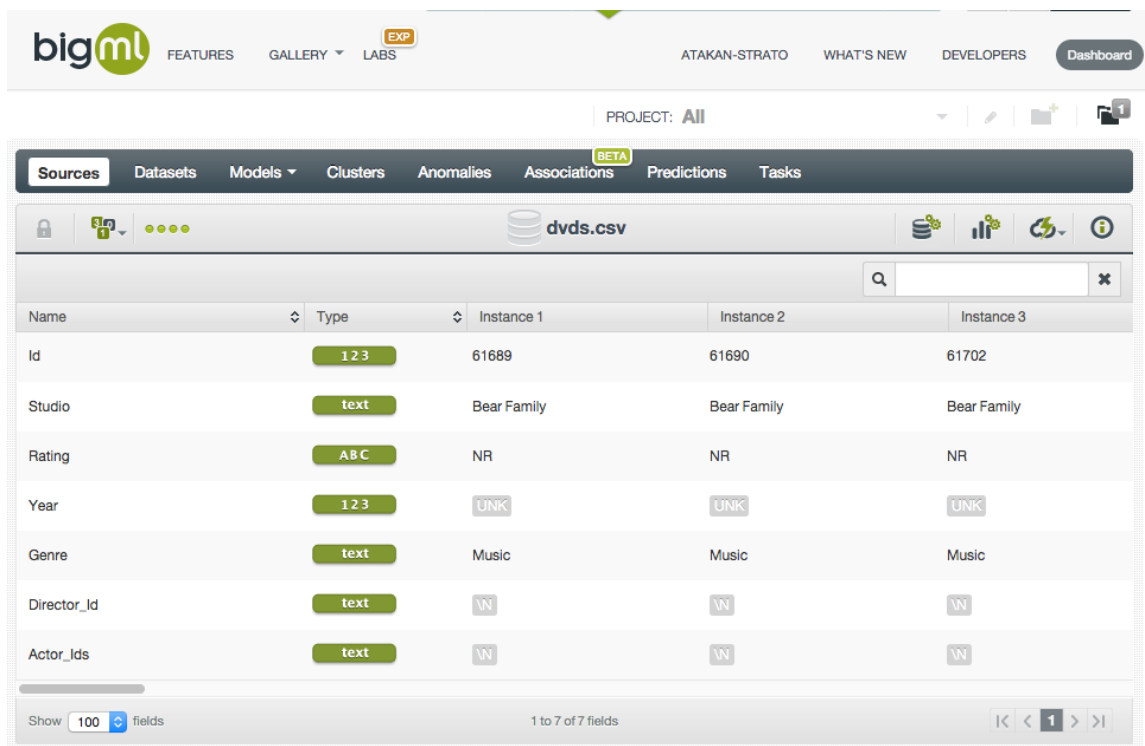


## 5. ANALYSIS: CONFIGURING AND RUNNING ASSOCIATION DISCOVERY TASK

Our analysis will follow the usual Machine Learning process evident in the way main menu tabs are presented on BigML's web interface from left to right starting with the Source tab.

### a. Upload Source Files

As a warm up exercise we will first feed the Director and Actor pairs data file mentioned in the previous section of this document. There are a total of 1,004,177 rows of Director Actor ID pairs in this source. Next up we upload the main Home Theatre Info source file, which contains 282,092 observations (rows). In each instance it takes less than 10 seconds for each upload to finish. Figures 1 and 2 show the resulting source upload details for the DVD.csv file. As seen below, instances are vertically presented (think of it as the transposed version of the typical Excel spreadsheet view). This gives us the chance to quickly peruse the file for potential problems with the upload. It is also worth noting that BigML automatically assigns data types to each variable (numeric, text, datetime or categorical). The user can manually override these choice as necessary.



Name	Type	Instance 1	Instance 2	Instance 3
Id	1 2 3	61689	61690	61702
Studio	text	Bear Family	Bear Family	Bear Family
Rating	A B C	NR	NR	NR
Year	1 2 3	UNK	UNK	UNK
Genre	text	Music	Music	Music
Director_Id	text	W	W	W
Actor_Id	text	W	W	W

**Figure 1** – Header plus instances 1 through 3

Instance 9	Instance 10	Instance 11
220757	22659	276862
Zeitgeist	BMG Music	Indican Pictures
NR	NR	NR
2010	UNK	2013
Documentary	Music	Horror
0015625	UN	0026488
166525 ;;;;;;;;;	UN	18913.0;19612.0;27498.0;128088.;160989.;204932.;204933.;204934.;204935.;204936.

**Figure 2** – Instances 9 through 11

## b. Create Datasets

After uploading our two source files, it is time to create datasets from those. Since data sources come in many shapes and formats, BigML relies on datasets instead as the standardized building blocks for iterative analysis and modeling on the platform. This is precisely why you want to spend some quality time on making sure that you create the dataset in the best possible way to lend itself best to subsequent analysis steps. Otherwise you will be faced with the familiar GIGO (Garbage In Garbage Out) situation. In our case (Figure 3), we chose to add “UNK” and “” (blank) as additional null value identifiers. We also enabled text analytics with the ‘Full-term’ setting and semicolon as value separators. This latter configuration is to take advantage of the Actor IDs contained in our last data field. Notably, full term tokenization omits tokenizing all terms, which would break down a Studio name such as Warner Brothers into ‘Warner’ and ‘Brothers’ resulting in potentially uninteresting associations involving partial terms.

**Source preview**

**Locale**  
English (United States)

**Separator**  
, (comma)

**Quote**  
" (double quote)

**Missing tokens**  
", NaN, NULL, N/A, \N, null, -, UNK, #REF!, #VALUE!, ?, #NULL!, #N/A

**Header**  
ml a,b,c

**Expand date-time fields**  
DISABLED ☒ ENABLED

**TEXT ANALYSIS**  
DISABLED ☒ ENABLED

**Language**  
Auto detect

**Tokenize**  
Full terms ☒ ☐ does runs ☐ A/a

**Items separator**  
; (semicolon)

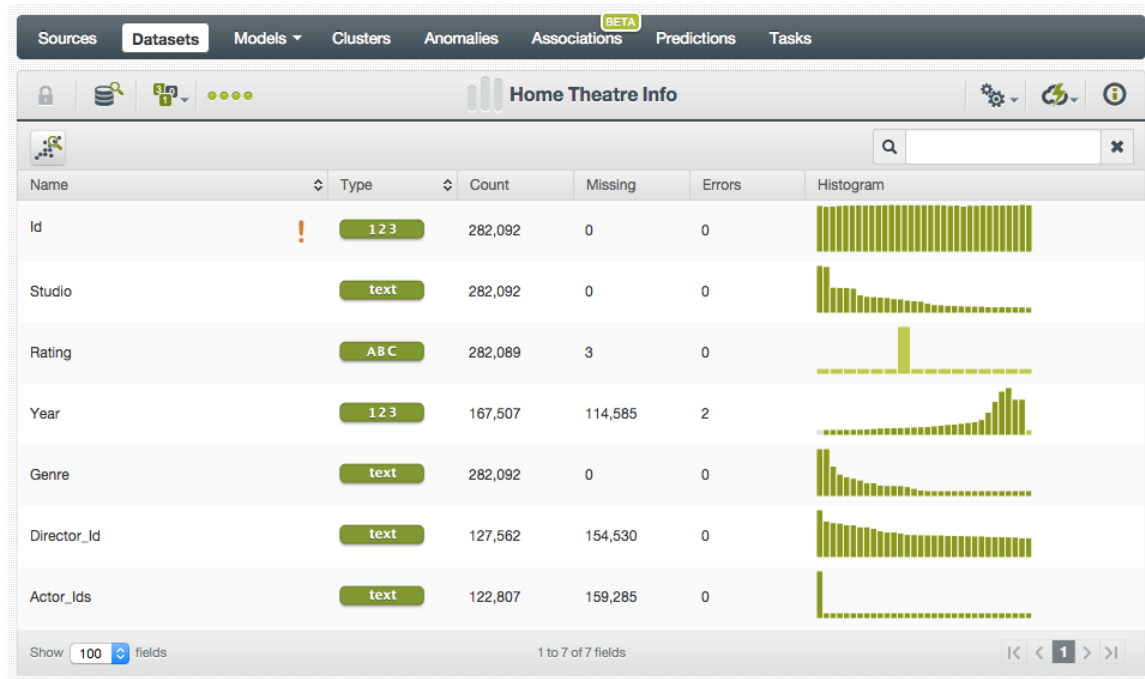
**Reset** **Update**

Name	Type	Instance 1	Instance 2	Instance 3
Id	Numeric	61689	61690	61702
Studio	Text	Bear Family	Bear Family	Bear Family
Rating	Categorical	NR	NR	NR
Year	Numeric	UNK	UNK	UNK
Genre	Text	Music	Music	Music
Director_Id	Text	\N	\N	\N
Actor_Ids	Text	\N	\N	\N

Show 100 fields 1 to 7 of 7 fields

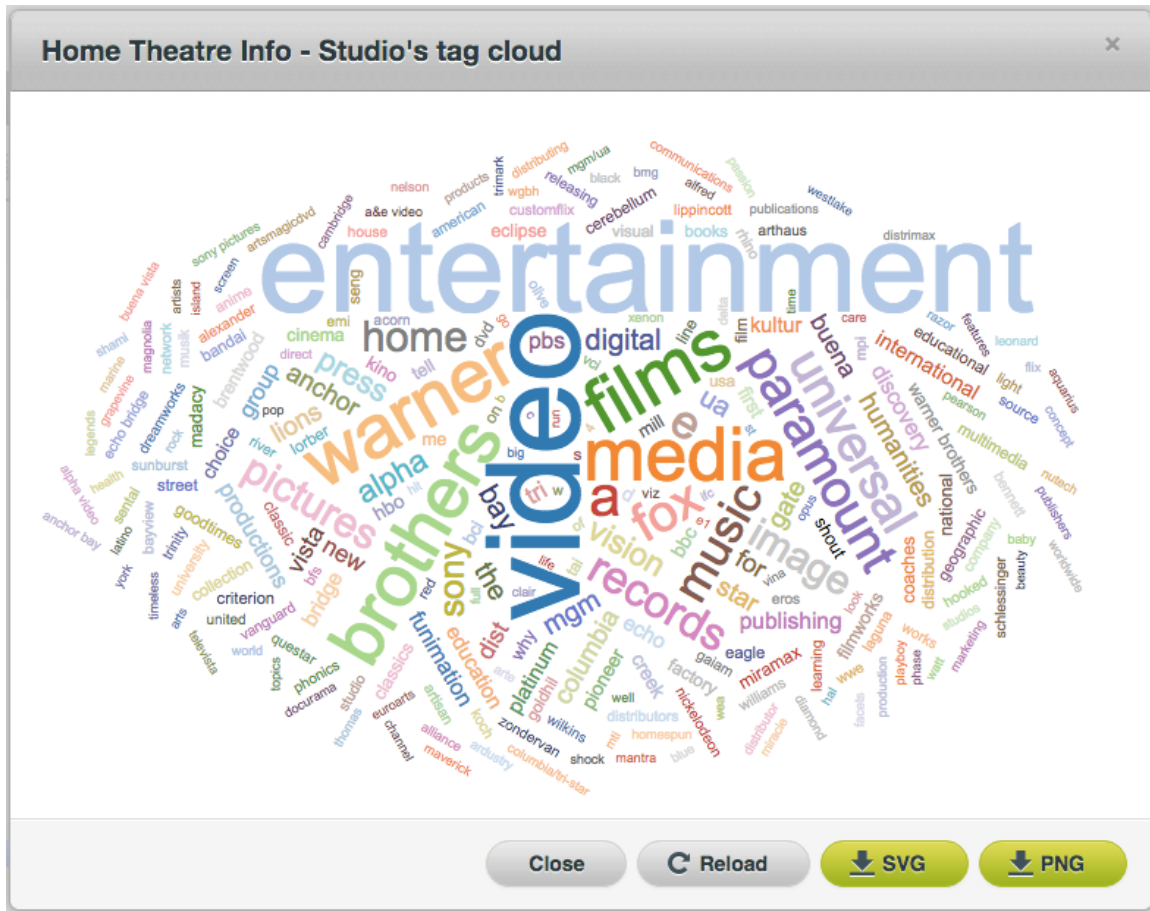
Figure 3 – Dataset configuration

Besides format standardization, dataset creation automatically presents the user with key summary statistics on every data field in the source file turned into a dataset. As seen in Figure 4, this includes missing value and error counts as well as the distribution of values for each variable. This saves the user the time and effort to run many queries in parallel as would be the case in a traditional database environment.



**Figure 4 – Dataset statistics**

More interestingly, as visible in Figure 5, text fields are accompanied with a tag cloud that summarize token frequencies based on the tokenization settings specified during dataset configuration. The tag cloud is downloadable in PNG and SVG formats for inclusion in presentations and other types of documentation.



**Figure 5** – Tag cloud for the ‘Studio’ text field

### c. Feature Engineering

Feature engineering is a worthwhile endeavor that can increase the accuracy of classification models and do wonders in many scenarios. However, in this analysis we are more interested in exploring statistically significant links between the existing variables in our DVD database so we will hold off on generating new data fields as a result. However we do encourage you to give a read to our [Feature Engineering User Guide](#) for future reference.

#### d. Analysis & Modeling

Now that we have the right datasets, configuring an Association task is really easy. BigML offers two options here:

- 1-Click Association
- Configure Association

Choosing the 1-Click Association kicks off a new task right away using the entire dataset. The advantage to this route is its ease of use since we do not have to go into any configuration details. However, for larger datasets with text fields that tend to create many tokens (remember that each individual word is a token by default unless the dataset was configured differently) the resulting task may take a long time to finish. Therefore, at least for starters, it is important to quickly configure an Association Task with a smaller subset of our dataset by using the sampling configuration settings. We can then iterate on the association task with different configuration settings based on the results at hand.

To configure an Association task, simply click on the Gears icon on the top right of your dataset detail view followed by the Configure Association menu option. This will open up the configuration panel with options including:

- Maximum # of Associations (capped at 500 for non-Virtual Private Cloud version of BigML)
- Maximum # of Items in Antecedent
- Search Strategy
- No Complement Items

And many more as described in detail in the Association Discovery User Manual.

If you click on the 'Configure' sub-navigation bar within the configuration drawer and the 'Sampling' sub-navbar after that you will arrive at the area that let's you control the sample size by using a slider control. We will use a 5% sample of our dataset to speed things up at this first iteration. Given that we have many thousands of observations in each dataset, that should still yield a large enough sample to draw conclusions on.

Our task takes less than 30 seconds to bring back results. The results are presented in the List View by default, which gives us crucial statistics of each association rule ordered by Leverage in our case because we selected Leverage as our search field before we ran





this task. However, we can easily change the sort order by clicking on any of the field headers (e.g. Coverage, Confidence etc.), which let's us compare and contrast different associations multi-dimensionally. The Association Discovery User Manual goes into further detail in describing the formulae and the meaning behind each rule metric.

BigML's Association Discovery also includes a 'Visual View' to help the user better grasp the results. The advantage of the Visual View is hidden in the well-known adage "A picture is worth a 1000 words." Currently, this view shows the most relevant associations according to the 'Leverage' measure. The slider bar let's the user see how more or less number of significant associations are spatially related to each other.

#### e. Model Evaluation

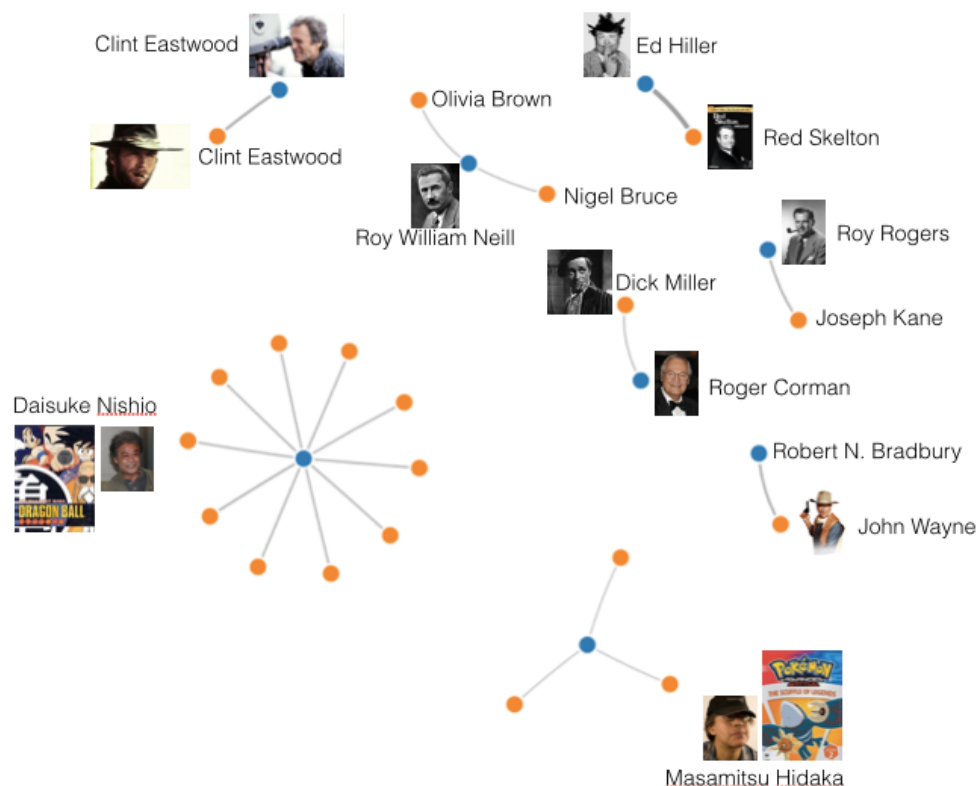
Since we were engaged in exploratory analysis on our datasets, we did not worry about model evaluation in this example case. Nevertheless, it would be prudent of us to observe whether our shortlisted associations of interest do hold in future data updates. After all, this data source is only a snapshot in time representation of what's going on in the movie industry. Thus it is likely to evolve further, which may introduce brand new associations while weakening previous ones. In an ideal production deployment setting, such incremental data would be periodically fed to BigML refreshing the association discovery task each time via a robust data pipeline powered by BigML's RESTful API and its its powerful native scripting language WhizzML. More to come on this with our impending WhizzML User Guide.



## 6. RESULTS

Interpreting Machine Learning results can be a tall order at times due to the complexity of the underlying methods of analysis. BigML's aim is to accelerate the diffusion of Machine Learning in organizations by carefully packaging algorithms and analytic approaches that lend themselves well to interpretation by subject matter experts (SMEs) and the developers working closely with them in implementing smart applications. Oftentimes these smart apps are expected to go beyond merely gaining insights from data by deploying similar insights and decision aids to the end consumer.

BigML's Association Discovery fits this mold really well as the learning curve is much smoother due to a combination of intelligent defaults and its ability to scale to large datasets. In our example, the Association Discovery task we ran on the Director Actor pairs found 100 rules involving 69 different items (variables and their corresponding values e.g. Actor ID = 72342 etc.).<sup>3</sup>



**Figure 6** – Actor and Director associations

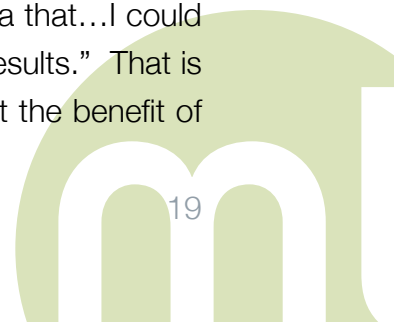
<sup>3</sup> CAUTION: These associations are not necessarily pointing out to the existence of any causality between the antecedent and consequent pairs identified.

Looking at the visual results view in Figure 6, we can see the top associations between Actors and/or Directors. Even though the dataset had separate fields for Actor and Director IDs the Association Discovery algorithm was able to identify different combinations of rules involving such as Actor vs. Actor, Actor vs. Director, and Director vs. Director. In this visual (which we manually augmented with some movie images) the Actors are marked in orange and directors in blue. Even if you may not be the biggest movie buff, the system has definitely stumbled upon some interesting World Cinema associations without any supervision. Examples include:

- The most prolific network of collaboration is in the bottom left, which points out to Japanese Director Daisuke Nishio's Dragonball Z creative cabal including many voice over artists bringing his anime characters to life.
- Next up, we see a similar scenario play out albeit at a smaller scale with the Japanese director Masamitsu Hidaka of Pokemon fame.
- Then there is the director Robert N. Bradbury and the Western king John Wayne partnership.
- Perhaps a lesser-known relationship is that of the middleweight boxer turned low budget film actor Dick Miller and the prolific engineer turned independent film producer/director Roger Corman known for his horror flicks. IMDB cites "He settled in Los Angeles in the mid-1950s, where he was noticed by producer/director Roger Corman, who cast him in most of his low-budget films, usually playing unlikeable sorts, such as a vacuum-cleaner salesman in Not of This Earth (1957)."
- A unique self-referential relationship exists between actor/director Clint Eastwood and himself. Of course, he needs no introduction. Talk about a Do-It-Yourself guy!

Just like that the history of chaotic moviemaking dating back to the silent era comes into focus pointing out to some of the strongest collaborations from its past that have soundly "beaten the odds" of randomness.

Those of you that are more experienced Machine Learning practitioners and maybe non-Hollywood junkies may be thinking "So what, we had a bunch of paired data that...I could just run a 'group by' SQL query against that same table to arrive at similar results." That is indeed true, but with BigML's Association Discovery implementation you get the benefit of



added association strength measures and a network visualization view that may not be apparent at a glance by looking at a long list database query results.

Now, let us turn our attention to our main dataset, which combines data fields that this warm up exercise left out. We follow a similar process to configure a new association task for this one with a 5% sampling rate. After less than 10 seconds of waiting, the analysis comes up with the following top associations for the enhanced DVD metadata dataset.

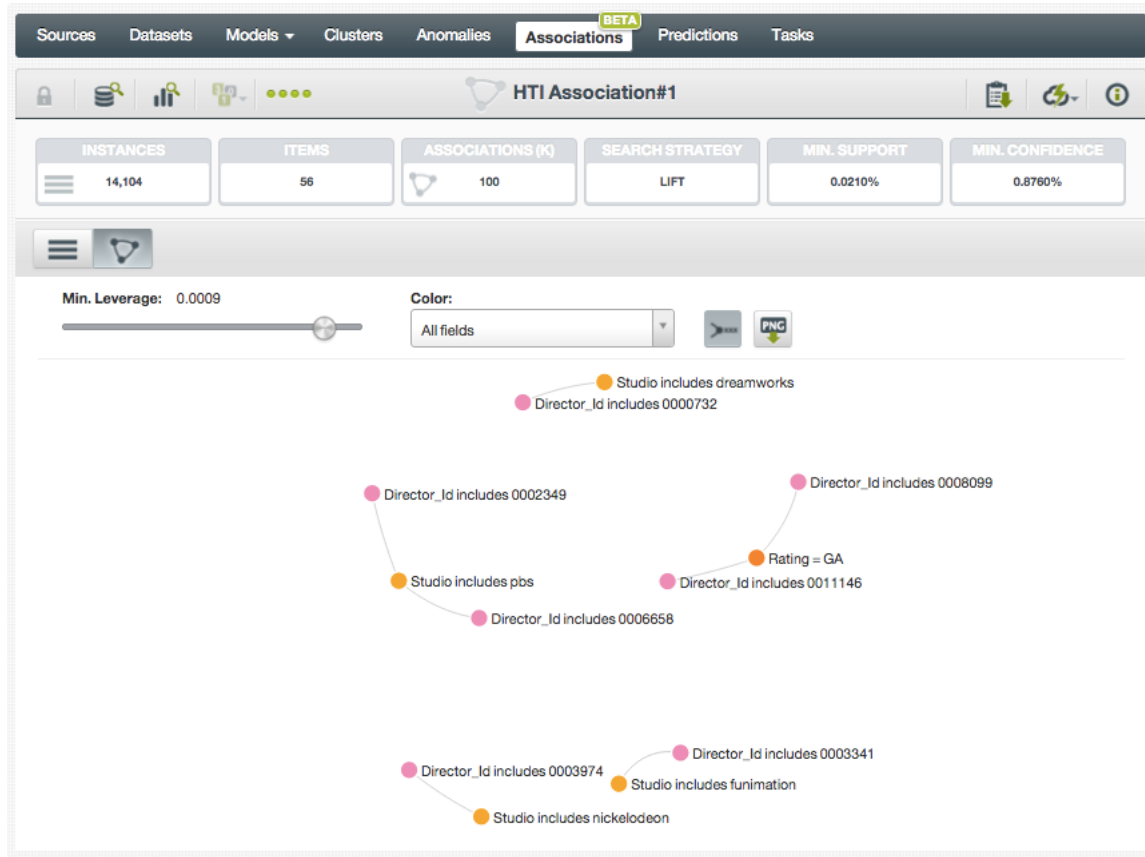
HTI Association#1							
INSTANCES		ITEMS	ASSOCIATIONS (K)	SEARCH STRATEGY	MIN. SUPPORT	MIN. CONFIDENCE	
14,104		56	100	LIFT	0.0210%	0.8760%	
Antecedent	Consequent	Coverage	Support	Confidence	Leverage	Lift	
Genre includes anime	Rating = MA13	3.6300%	1.8150%	50.0000%	0.0175	26.2156	
Rating = MA13	Genre includes anime	1.9070%	1.8150%	95.1670%	0.0175	26.2156	
Rating = UR	Genre includes late night	2.0700%	1.4530%	70.2050%	0.0141	34.8654	
Genre includes late night	Rating = UR	2.0140%	1.4530%	72.1830%	0.0141	34.8654	
Genre includes anime	Studio includes funimation	3.6300%	0.6880%	18.9450%	0.0066	24.9724	
Studio includes funimation	Genre includes anime	0.7590%	0.6880%	90.6540%	0.0066	24.9724	
Genre includes anime	Rating = MA17	3.6300%	0.6520%	17.9690%	0.0063	24.8462	
Rating = MA17	Genre includes anime	0.7230%	0.6520%	90.1960%	0.0063	24.8462	
Genre includes anime	Rating = MA15	3.6300%	0.5670%	15.6250%	0.0054	25.6250	
Rating = MA15	Genre includes anime	0.6100%	0.5670%	93.0230%	0.0054	25.6250	
Rating = MA13	Studio includes funimation	1.9070%	0.4330%	22.6770%	0.0042	29.8907	
Rating = VAR	Genre includes var	1.7730%	0.4330%	24.4000%	0.0042	26.2701	
Genre includes var	Rating = VAR	0.9290%	0.4330%	46.5650%	0.0042	26.2701	

Figure 7 – World cinema associations list

As you can see, the highest leverage association rules that float up to the top consistently involve Genre, Studio and Rating fields. Given that leverage scores range between 0 and 1 we can conclude these are notable, but not the strongest possible associations given all other observations in the dataset totaling some 282,000 rows representing individual DVDs. However, these rules show significant “Lift” that justifies that we take them seriously. For example, the first rule stating that whenever the Genre is “Anime” the Rating is “MA13” has a lift score of approximately 26. This means the association between these two values of these two different variables is 26 times more likely in our dataset than what would be expected from simple coincidence. It is also worth noticing that the second rule in the result set is the reverse of the first rule with the same Leverage and Lift values as expected. Yet the Coverage, Support and Confidence scores are different because those are calculated with respect to the “Antecedent”, which is different in each case.

It is not surprising that the main variables involved (i.e. Genre, Studio, Rating) are those data fields containing smaller variety of possible values in the dataset as opposed to other variables such as Director ID and Actor IDs for each DVD. Less values result in less potential combinations of those values, which in turn is more likely to yield higher coverage pairings and association rules. The higher coverage meaning more instances likely has a favorable bias towards more statistically significant findings making it past the 0.05 default significance threshold.





**Figure 8** – Word cinema associations visual

Once again, as we click on the visual results view button, our rules come alive as shown in Figure 8. There are some real gems in this case. For example,

- Director ID 732 associated with the movie studio Dreamworks is non other than Mr. Steven Spielberg.
- Director ID 3974 associated with Nickelodeon is Chris Gifford – the creator of the hugely popular children’s cartoon series ‘Dora the Explorer’.
- Daisuke Nishio and comes up again, this time linked to ‘Funimation’, which is the leading studio for anime and foreign entertainment licensing and distribution in North America.

- PBS (Public Broadcasting Service) is strongly associated with the directorial godfather to many in the documentary genre, Ken Burns as well as Stephen Ives (known for The American Experience series among others).
- It is also very plausible that the Pokemon creator Masamitsu Hidaka and Dave Filoni of Star Wars anime fame are tied with the 'General Audiences' movie rating.
- We did not include those in the graphic to avoid crowding out but playing with the 'Leverage' slider one can also reveal
  - Stanley Donan (Singin in the Rain) -> Musical
  - And the one we were curious about from the get go Alfred Hitchcock -> Mystery/Suspense

Depending on the viewpoint of the user, some of these findings may be hardly groundbreaking, but we can confidently say they are all valid short of data quality issues that must customarily be addressed at the beginning of the flow as we did. The key idea here is that Association Discovery is your best friend when you have the challenge of finding non-spurious relationships in heaps of data that include many variables each involving many values of their own. Our example in Figure 8 went beyond the Director – Actor associations from Figure 6 and successfully mixed in new variables such as Genre, Rating, Studio and their specific relevant values.

This last point also raises the challenge of identifying what makes an “interesting” association. There are many academic papers written on this seemingly simple question so it is more complex than it appears. For the sake of our exercise, we would leave this as a judgment the subject matter expert has to make. If it is useful in his/her setting then it is good to go. If not, just skip to the next association.<sup>4</sup>

---

<sup>4</sup> NOTE: In future releases, we will provide more filtering capabilities to aid the user to narrow down the list results. If you have ideas of your own be sure to let us know at [info@bigml.com](mailto:info@bigml.com).



## 7. CONCLUSION

Association Discovery is a powerful technique applicable to many different use cases in order to unearth hidden relations between numerous variables and their numeric, text or categorical values. In this user guide, we analyzed the relationships between films, stars, directors, genres and similar topics extracted from a publicly available DVD metadata database. We found some rather obvious results on how those variables are related, while also stumbling upon some fresh ideas not so obvious at first glance.

All in all, using a straightforward example, this guide merely scratches the surface in terms of what you can achieve by employing Association Rule Discovery as part of your Machine Learning projects. In future user guides we will explore more complex scenarios involving Association Discovery as one of multiple BigML components employed to solve a Machine Learning challenge.

Built in statistical tests inherently performed by BigML's Association Discovery implementation coupled with its ease of use, configurability and interpretability makes for an exciting contribution to the toolsets of Machine Learning Specialists, Developers and Business Analysts alike. We hope to learn from your experience and to further improve our offering, so be sure to send us a note at [info@bigml.com](mailto:info@bigml.com) with your ideas and feedback.





