Drag and Drop Machine Learning

By Source, Fair use, https://en.wikipedia.org/w/index.php?curid=25934588

About me

- 20 years in computer security
- Formerly: MOREnet, REN-ISAC
- Currently: Jack Henry and Associates
- Twitter: @bethayoung
- Currently: Syracuse University, Masters Applied Data Science

What is machine learning?

- Microsoft: Computing systems that improve with experience
- Others:
  - A type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed
  - Related to computational statistics, which focuses on prediction-making through the use of a computer

Traditional Programming

FIGURE 1-1 Traditional programming paradigm.

Machine Learning

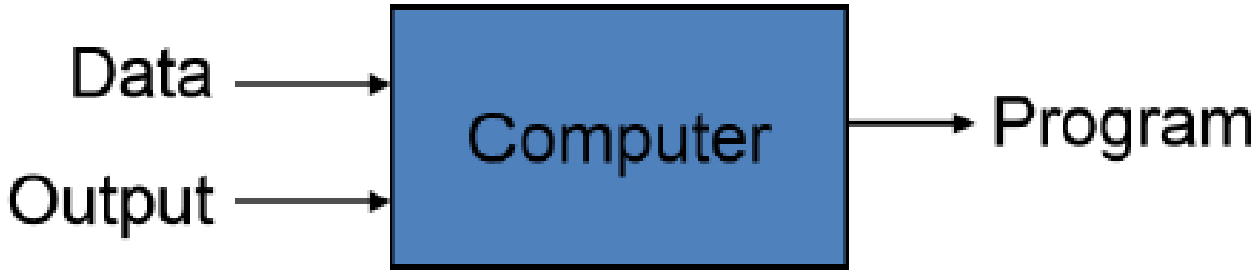


FIGURE 1-2 Machine learning programming paradigm.

Supervised vs Unsupervised

Supervised Learning – Training data includes the desired output

- Classification (buckets of data) – predictive responses fall in just a few known values
- Regression – continuous variables such as profit and loss

Unsupervised Learning – Training data is not include desired output

- The success of the predictive model relies on the ability on infer and identify patterns
- Cluster analysis is the most common form

Demo time!

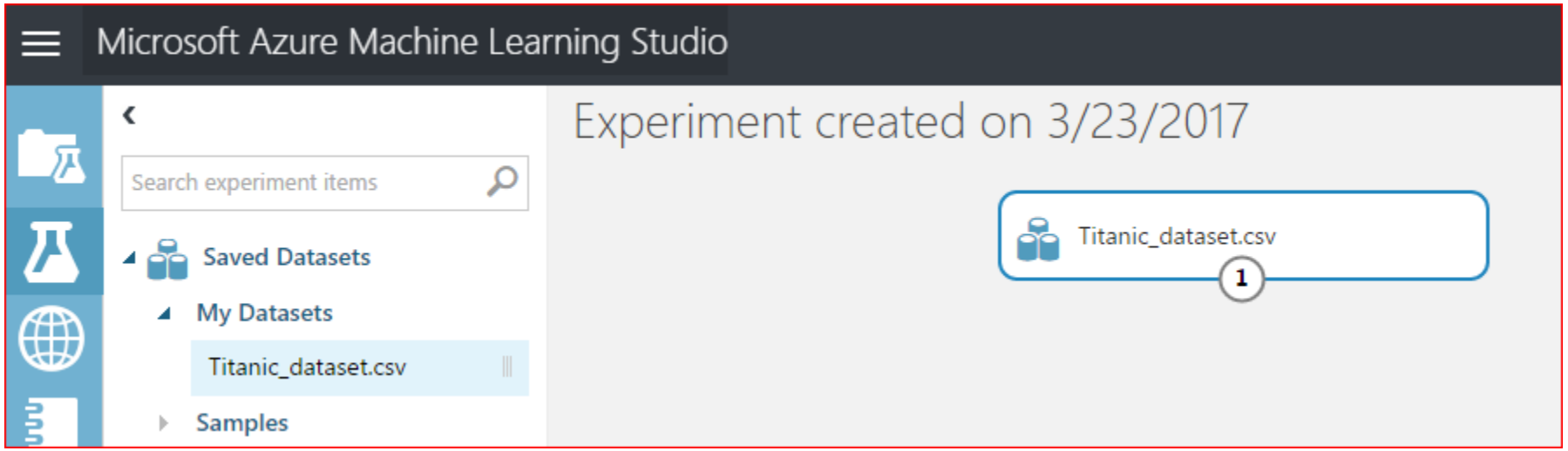
https://studio.azureml.net

We are going to use pre-classified data to train our model.

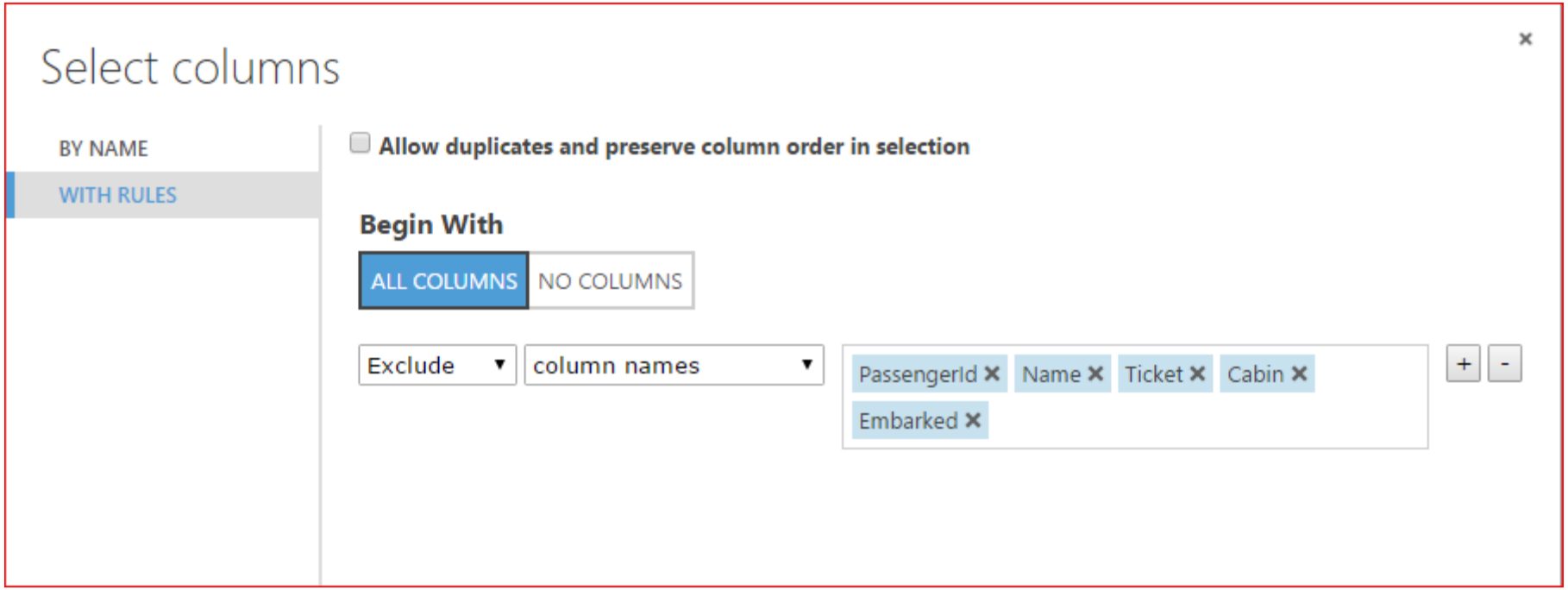
Dataset:
https://www.kaggle.com/c/titanic/data

Step 1: New Experiment

- Click the +NEW at bottom left of screen and select “new blank experiment”
- Next, we have to add new data to our experiment
- Click New again and this time select “Dataset” and add the Titanic data set.
- Drag the data set to the blank experiment.

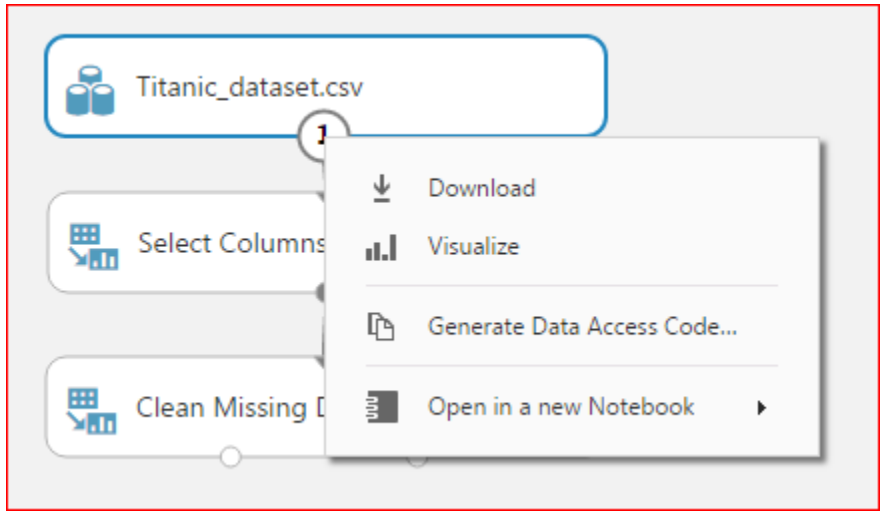


Selecting columns



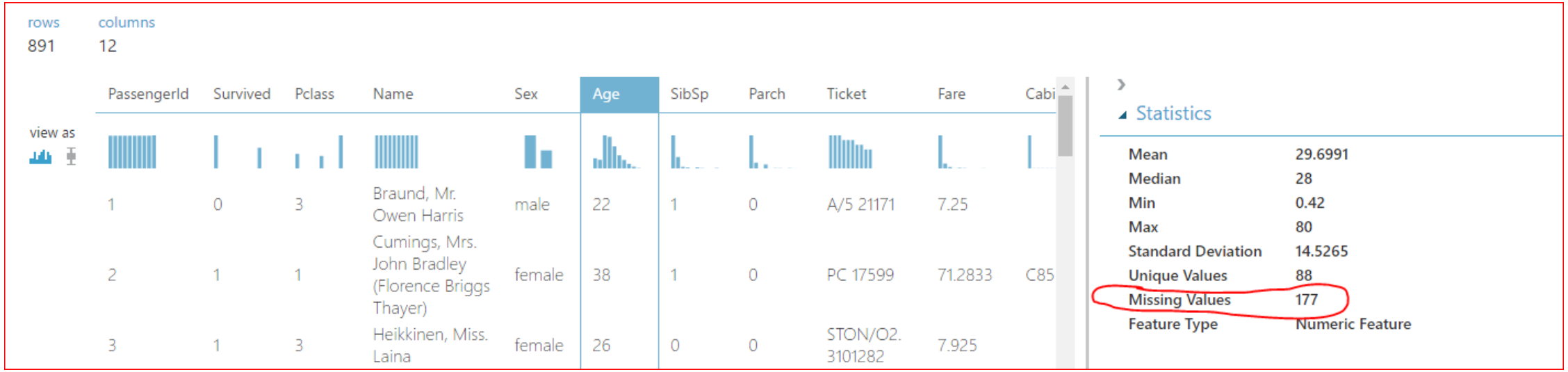
Clean Missing data – Part 1

- Sometimes columns have missing data. To check to see if your data has any missing values, right click on the bottom connector and select Visualize data.



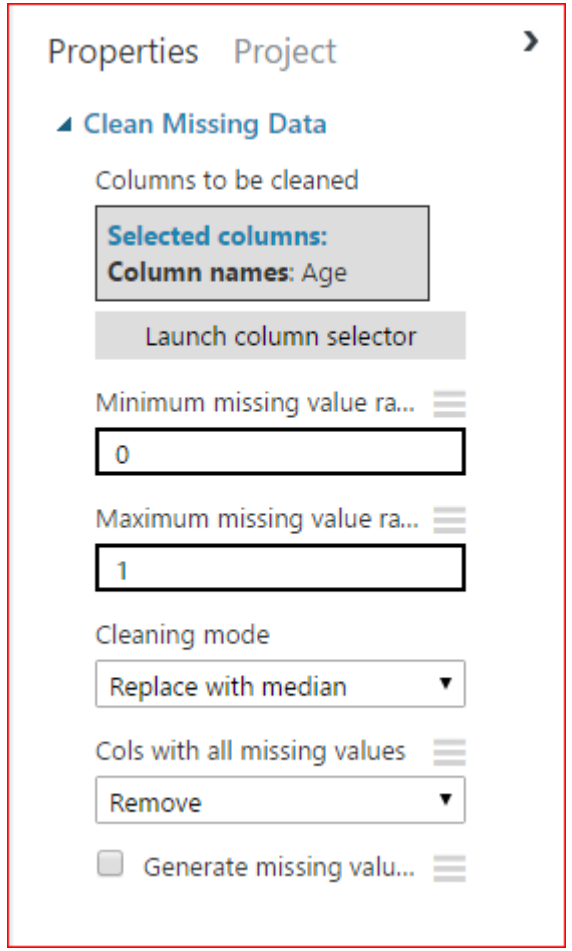
Clean Missing Data – Part 2

- Select a column and review the information in the Statistics window on the right. Decide what you are going to do about the missing data



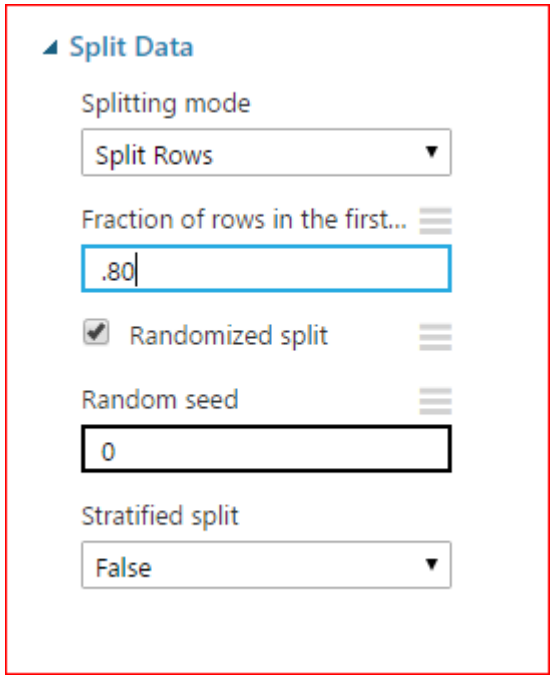
Clean missing data – Part 3

- In our case, the only column with missing values is “Age” and we don’t want to use 0 for an age and we don’t want to exclude about a quarter of our data set. We will replace the missing data with the median age, which is automatically calculated for us.



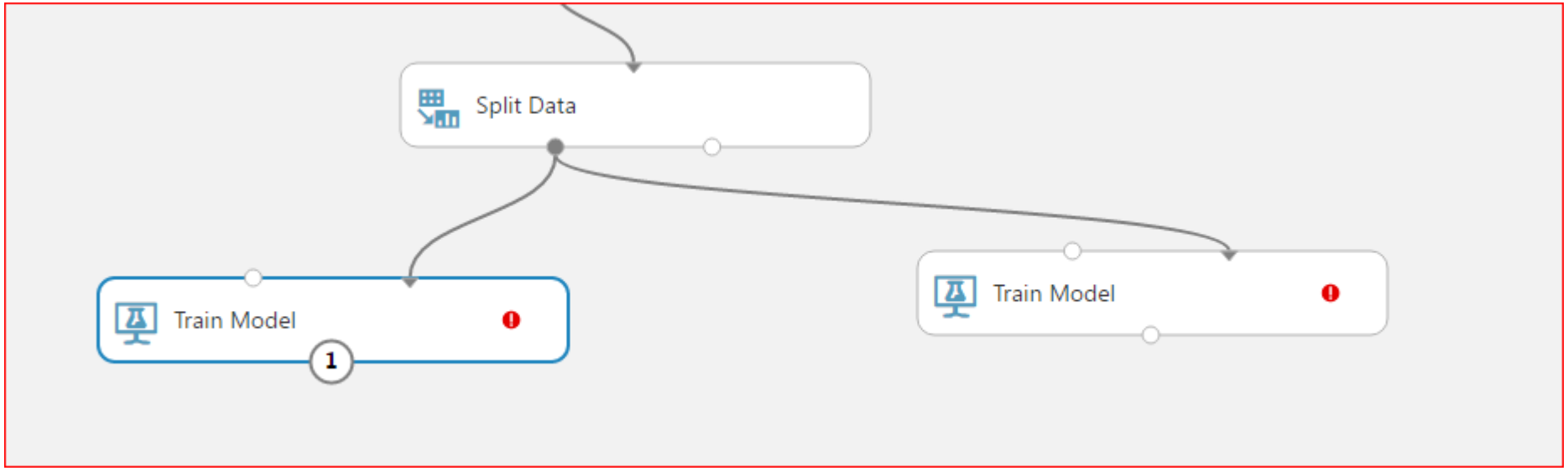
Split the data

- Academics like 80/20 but you can split it any way you like. This is to divide your data set into "learning" and "verify." The learning portion will be to train the model and the verify portion is used to see how well the model learned.



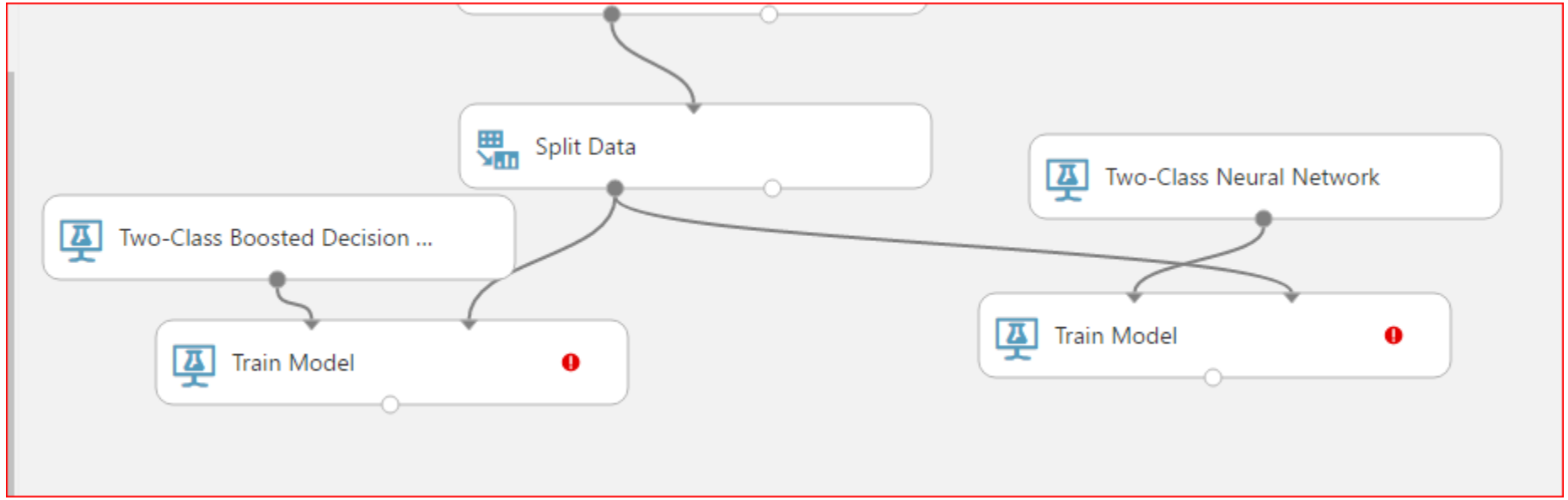
Train the model

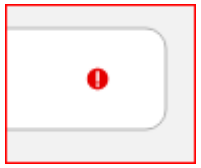
- You can do multiple models at the same time to decide which one is best. We are going to start with two.



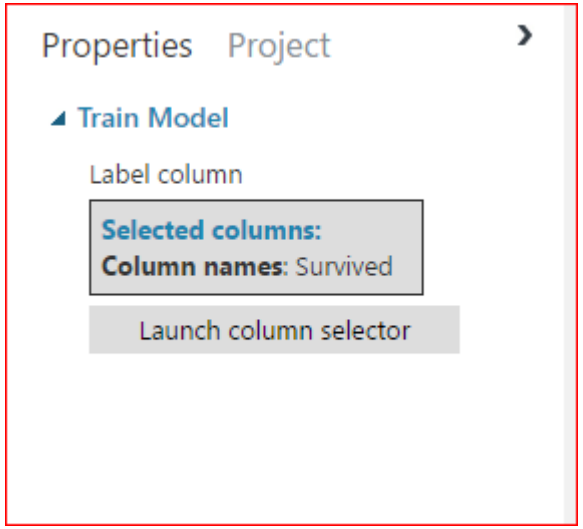
We also need to pick the Classification – what mathematical model do we want to use?

- If the prediction is one of two values (survived or died, yes or no), then pick a two-class model. The red ! means there is something else still needed.



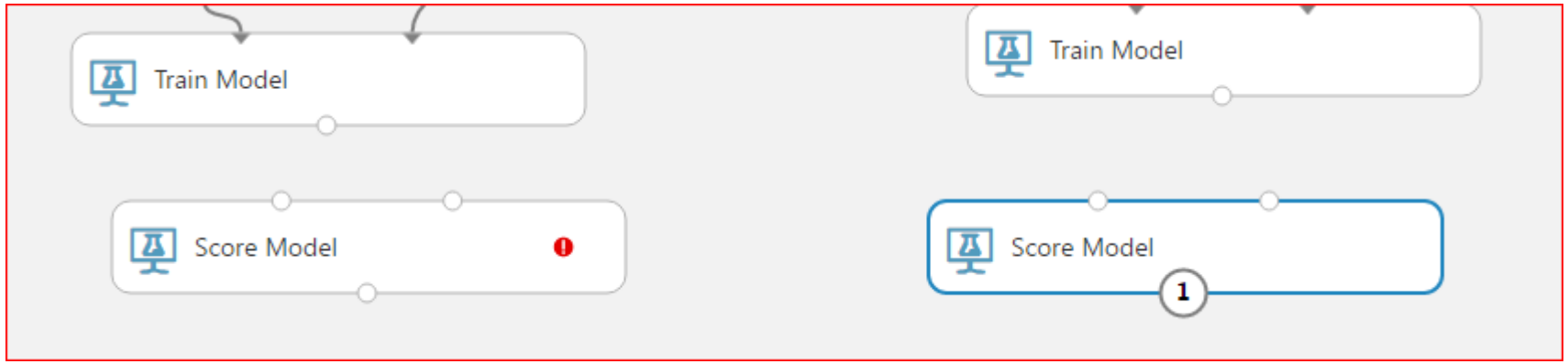
Fixing the 

- Click on it and it will tell you what is missing – in this case, we need to tell the model which column has the result information (whether a person survived or not)

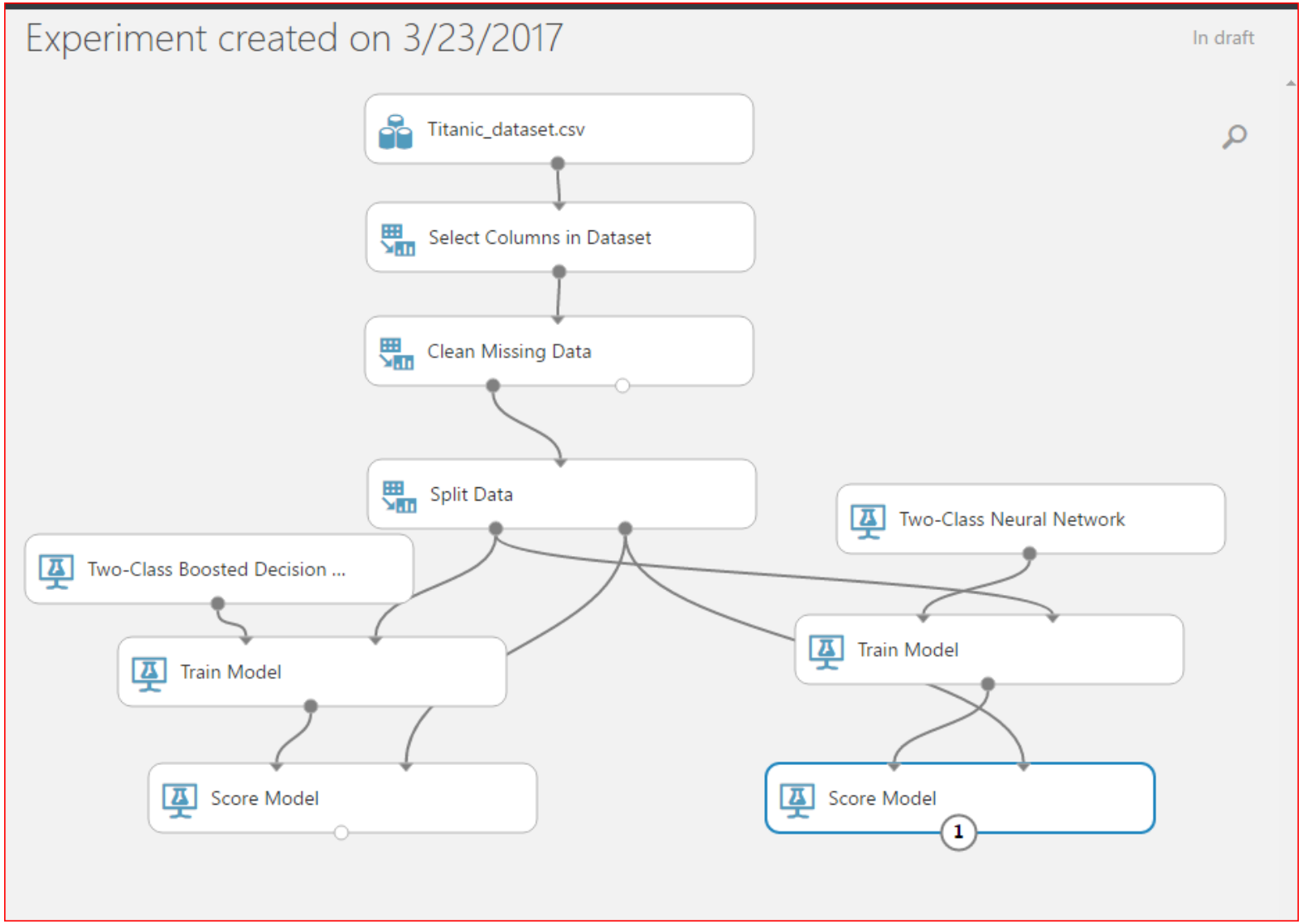


Scoring the module

- Let the process run and generate scoring results (how accurate was the model, what was the false negative or false positive rate, etc). The number of connects indicates how many connection it needs. Since I need to score two training modules, I need two scoring modules since I can’t connect more than one thing to each. Notice that the scoring module needs two inputs though.

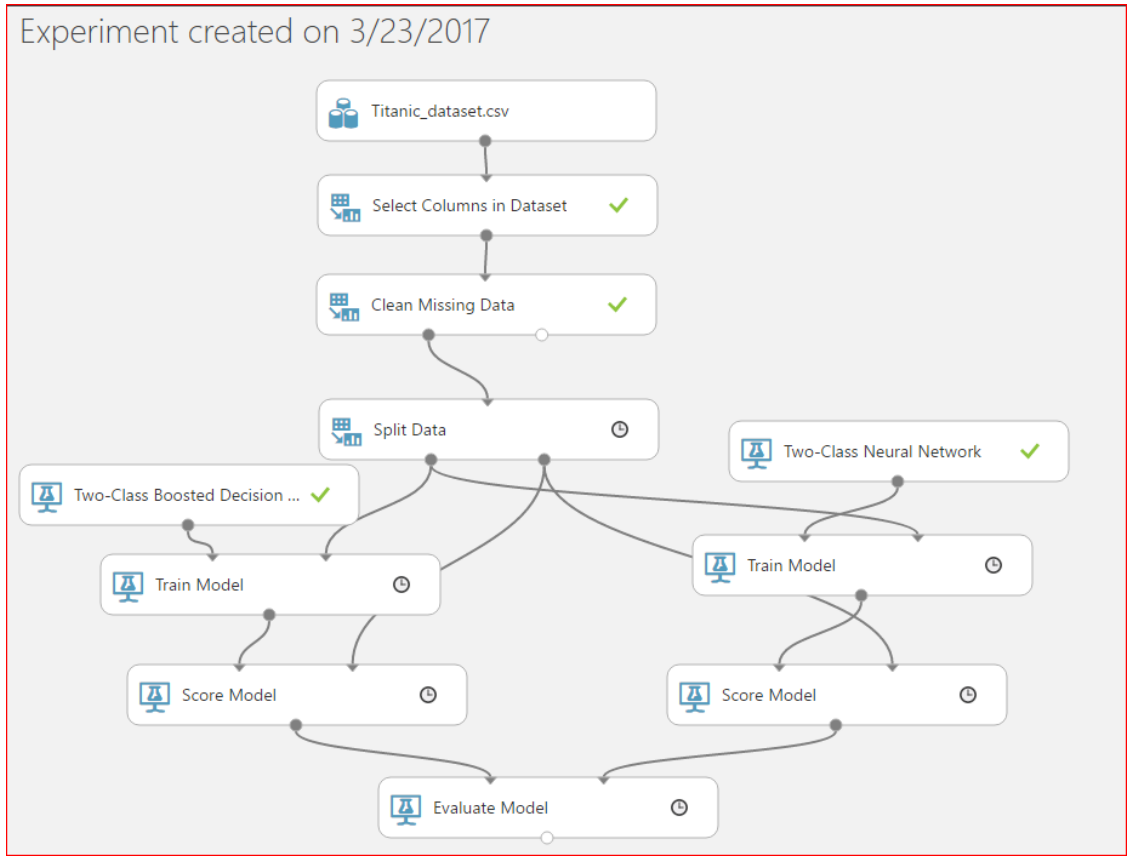


Connect the other part of the data set!



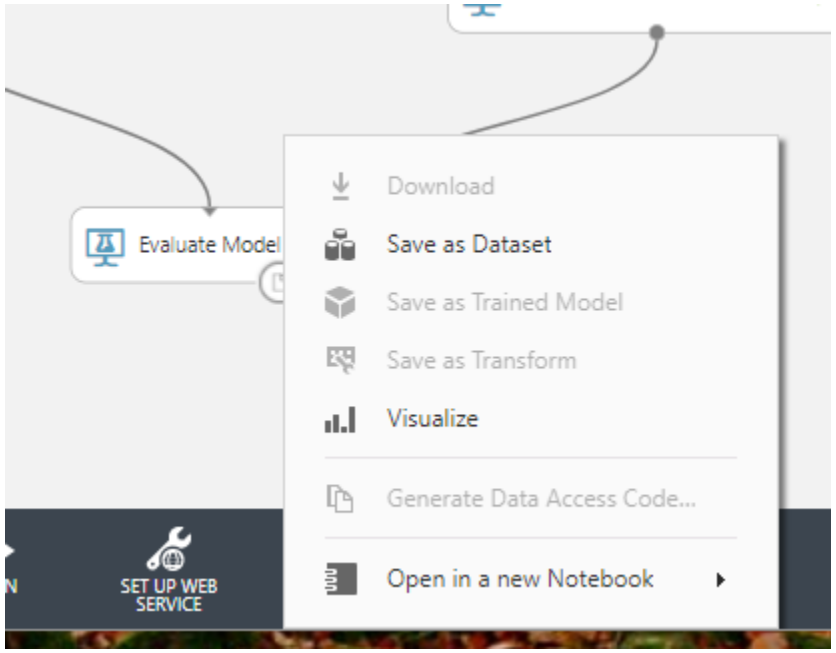
Now to evaluate the modules – which one was best?

- Add an Evaluate Module and then select run. As the individual pieces finish, you will see a green checkmark.



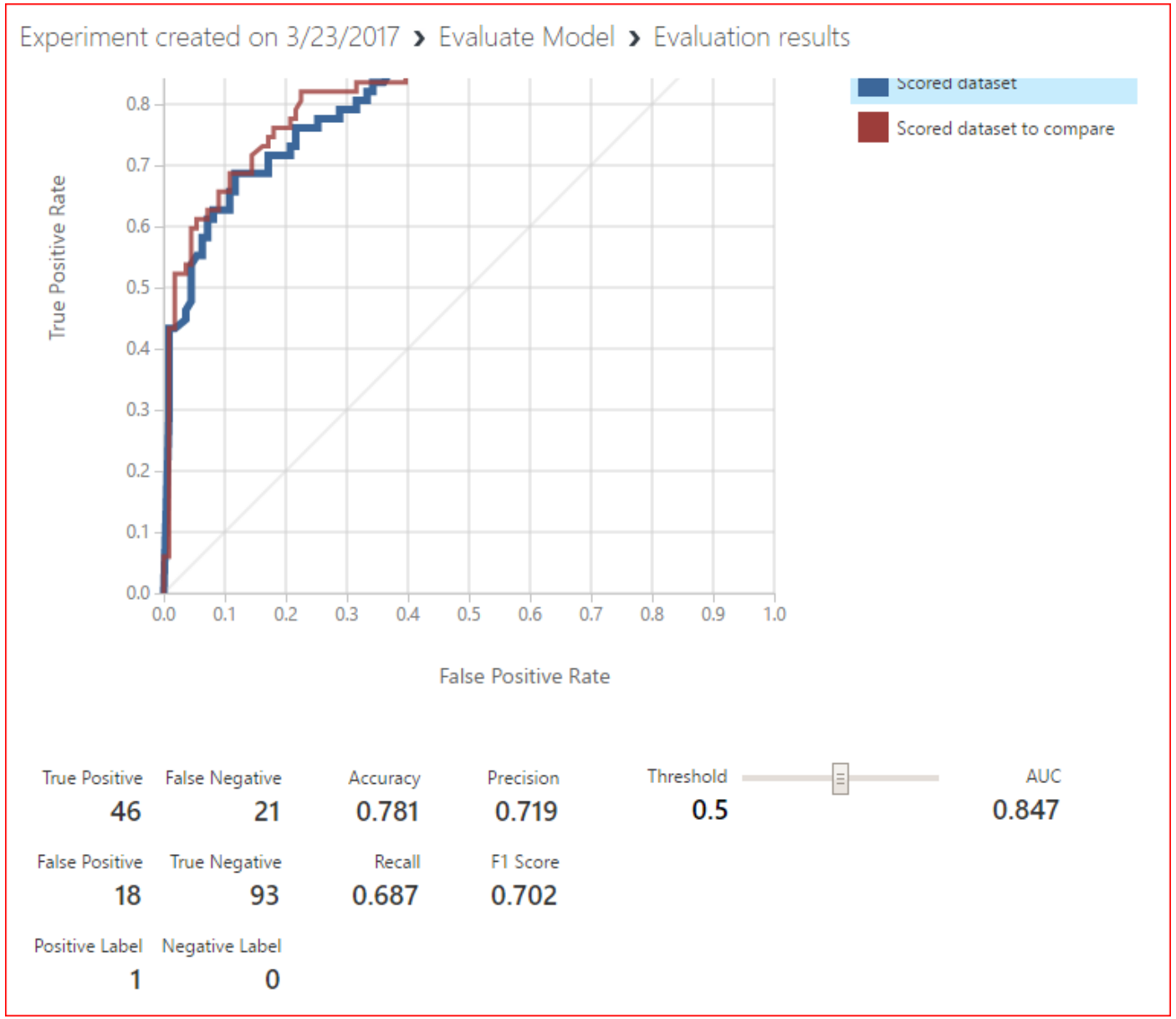
Finding the Evaluation data

- Right click on the evaluate model process connector and select Visualize



Evaluating the results

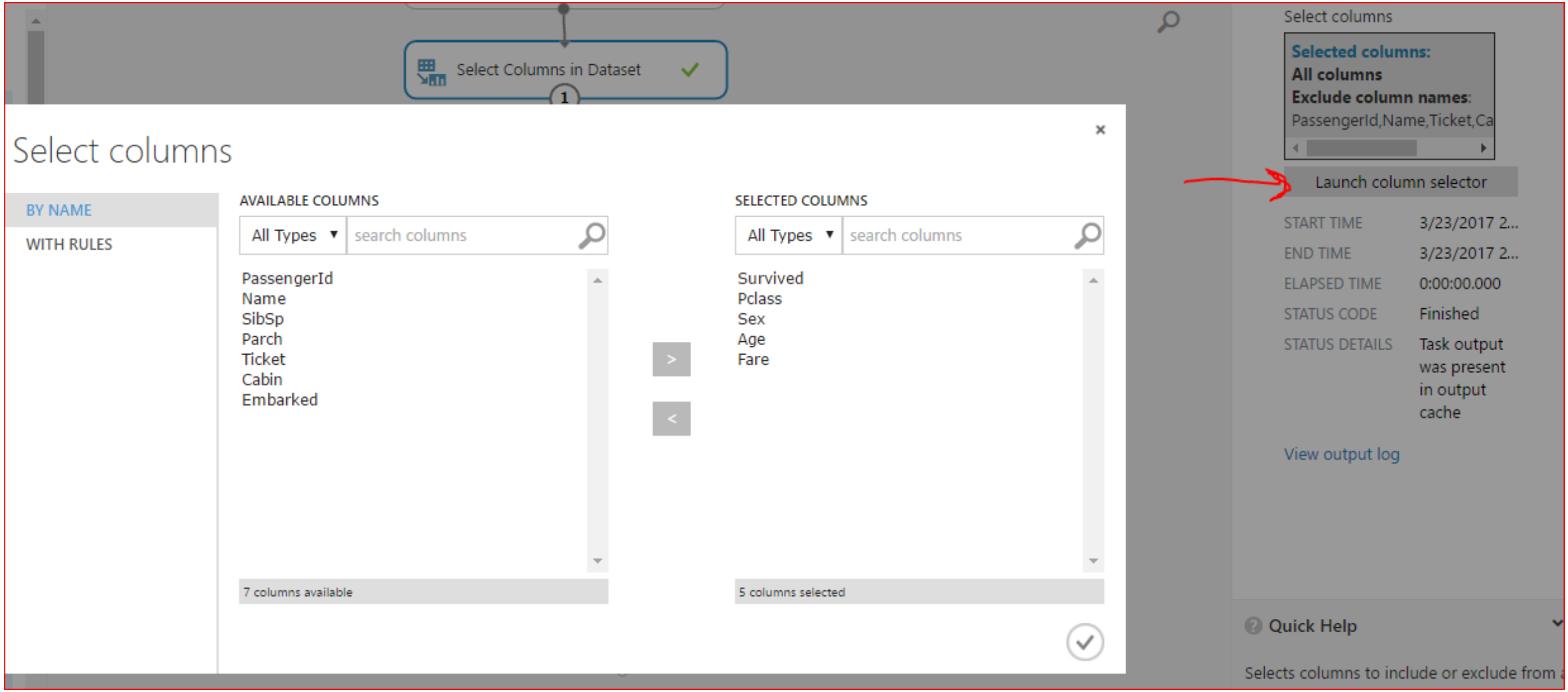
- There is a bunch of math behind the values shown. I can try to explain it…
- This one wasn't very good, adjusting column selection will help



Accuracy, Precision, Recall and F1 score

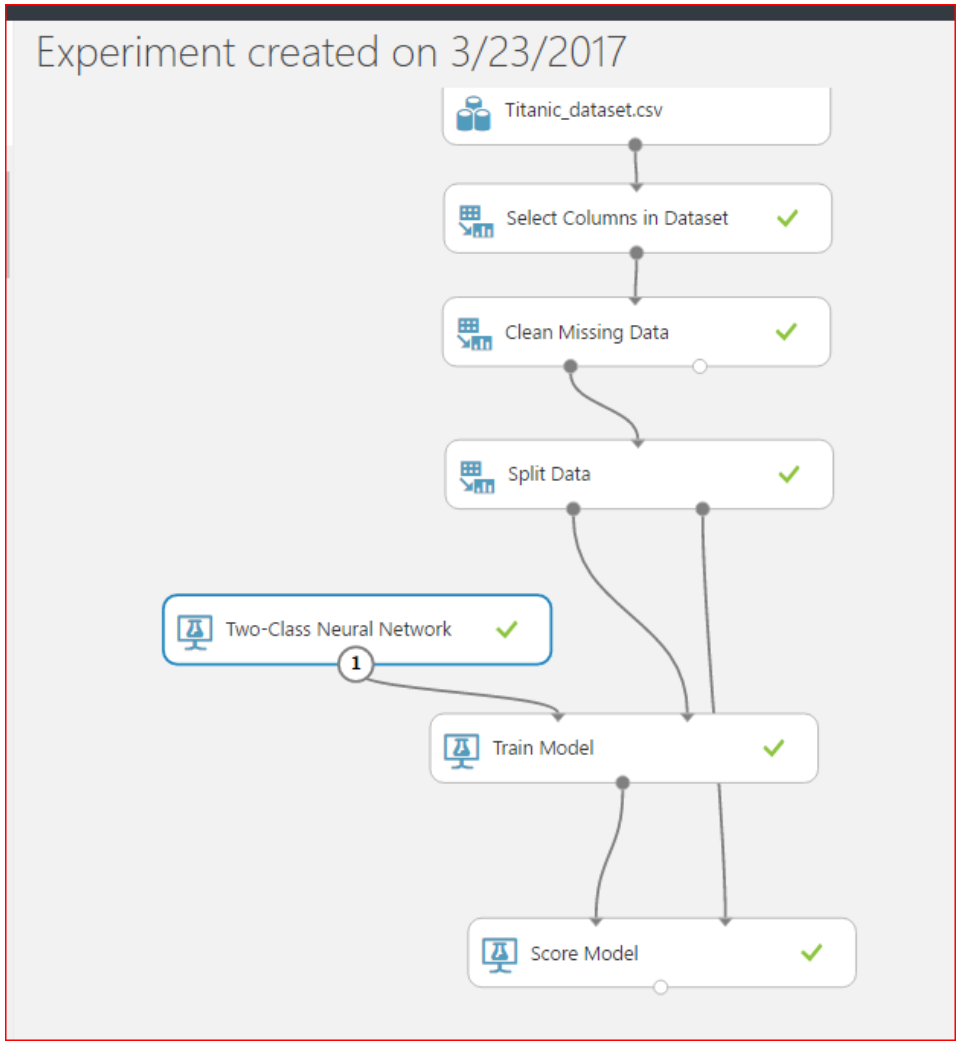
- Accuracy – the proportion of correct results that were achieved
- Precision – the fraction of relevant instances among the retrieved instances; high precision means substantially more relevant results were returned over irrelevant ones; TP/(TP+FP)
- Recall – fraction of relevant instances that have been retrieved over the total amount of relevant instances; high recall means results include the most relevant results; TP/(TP+FN)
- F1 Score – conveys the balance between precision and recall 2*((precision*recall)/(precision+recall))

Changing the dataset

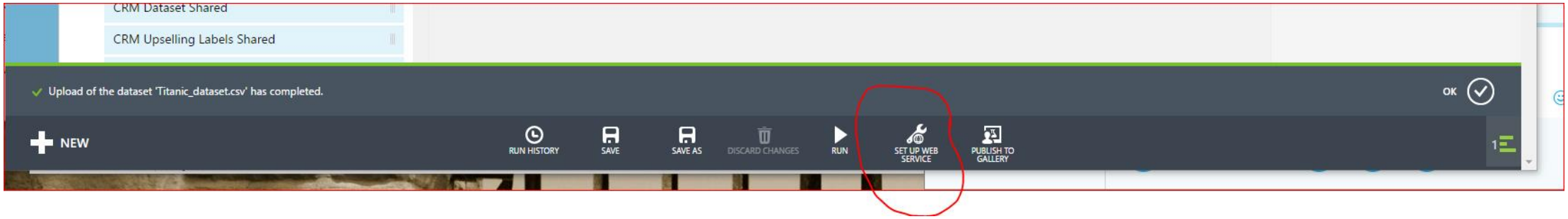


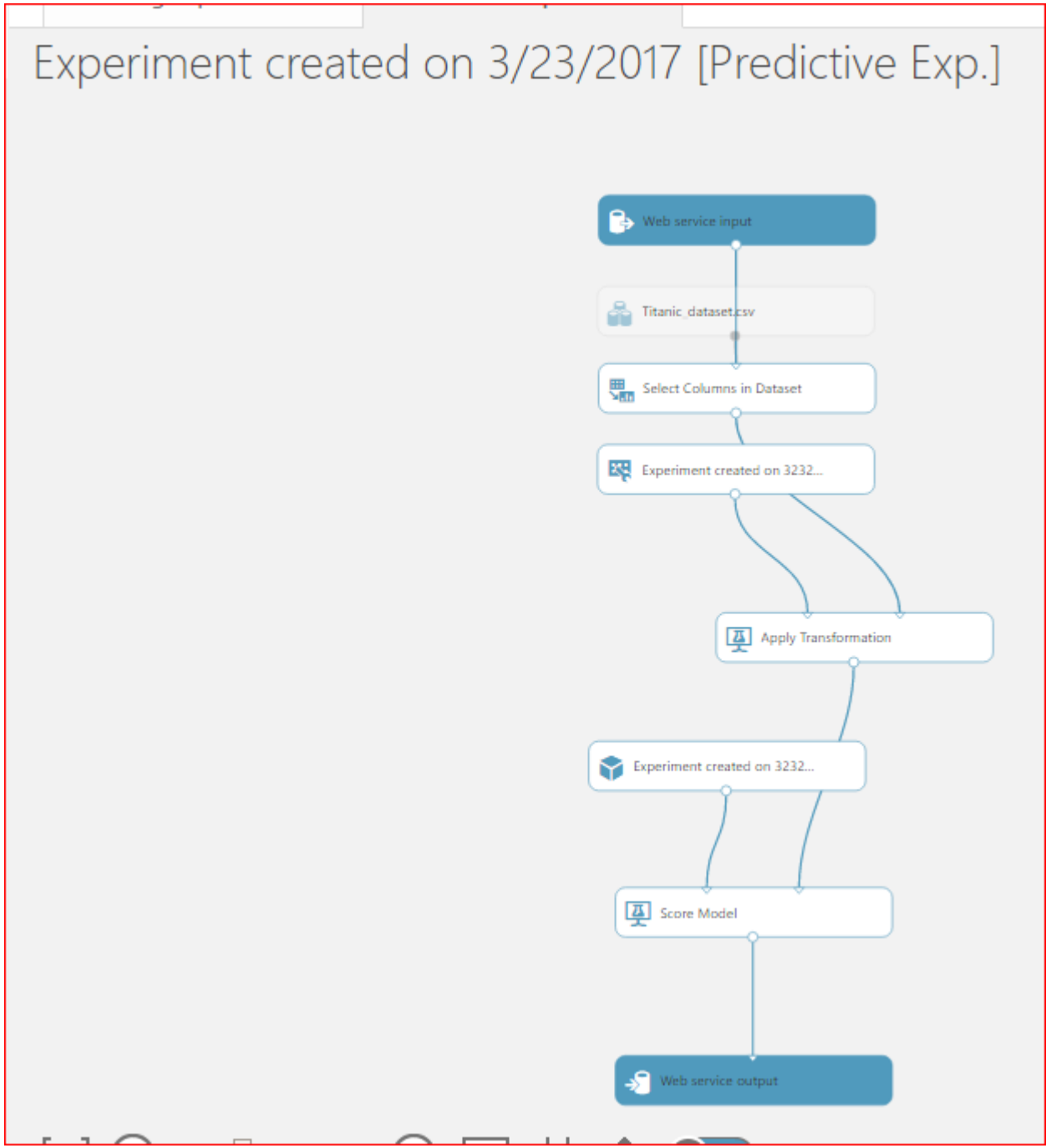
Once you are satisfied with the training, time to build an application around it

- First, remove the unused model and the evaluation module and click run to check for errors
- After it runs and you should have green checkmarks



Build the Web Service – Predictive Web service

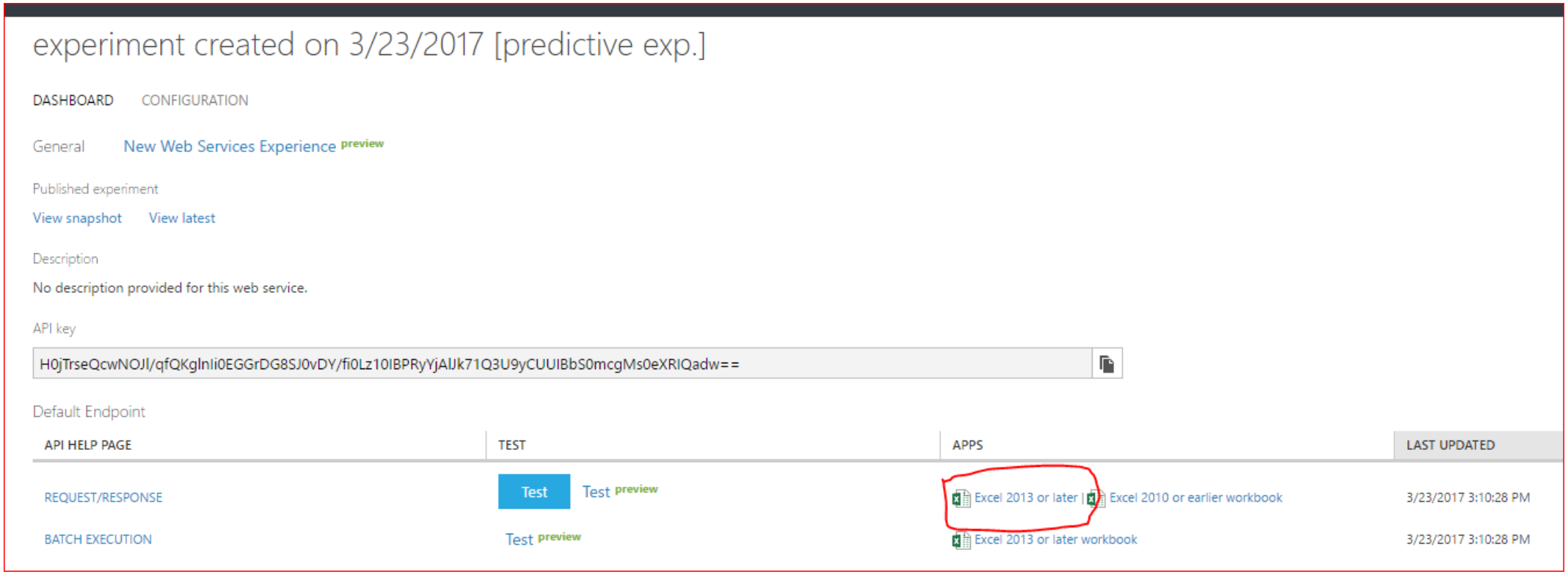




- Once you have this, click “run” again, so it will check for errors, again…
- Once the check is completed, select the “Deploy as web service”

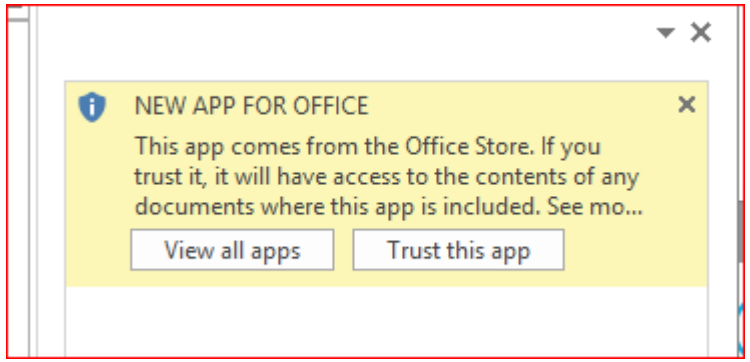
Now we can build an application around our model

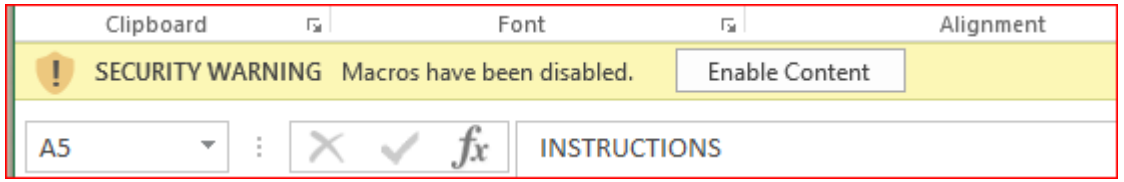
- Notice that there is an API key
- We can also do tests from this screen



Now it gets scary…

- An Excel document will be downloaded to your machine.
- Open it and then do what every security professional will tell you not to do – Trust this app (or Enable Content, depending on which version of Excel) (are you scared yet?)





In Excel 2010, it does this

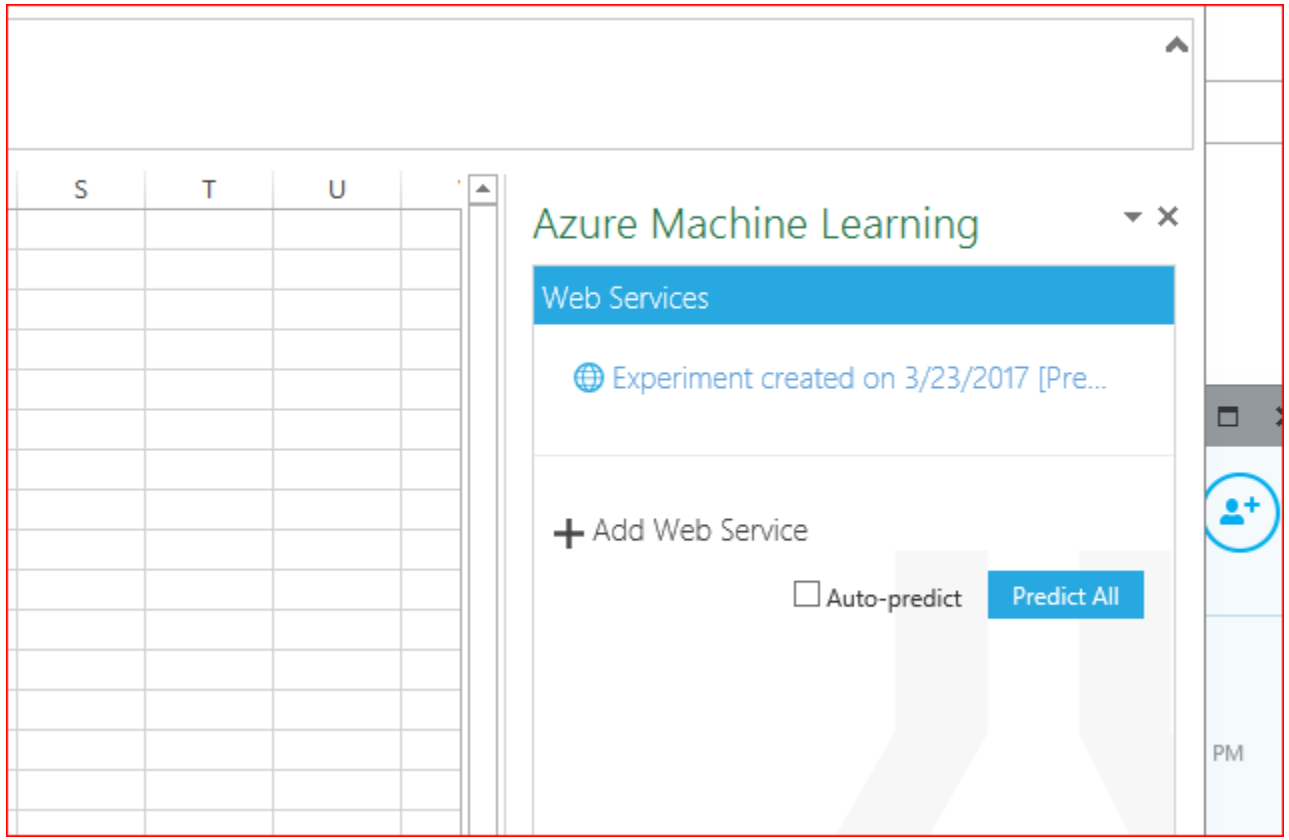
INSTRUCTIONS

Once you have enabled macros and the table has been generated, please enter your input values in the **PARAMETERS** section. Once all parameters have been entered, **PREDICTED VALUES** will be automatically computed.

If the web service you consume is hosted in a **Free Workspace** you may experience delay due to throttling. Upgrade to **a Standard Workspace** to have higher performance.

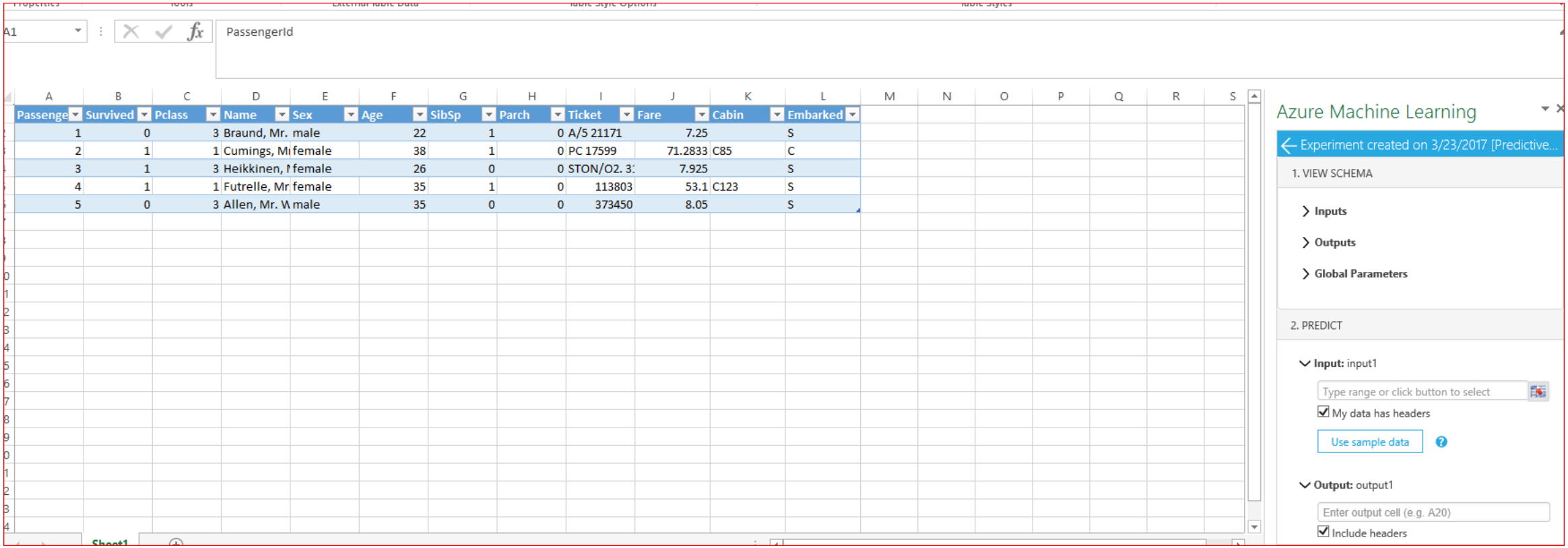
PARAMETERS												PREDICTED VALUES								
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	ScoredLabels	ScoredProbabilities
32	0	1	Beth Y	F	48	0	0	0	100	Cabin	S	0	1	F	48	0	0	100	1	0.897295713

There are a few more steps in 2013



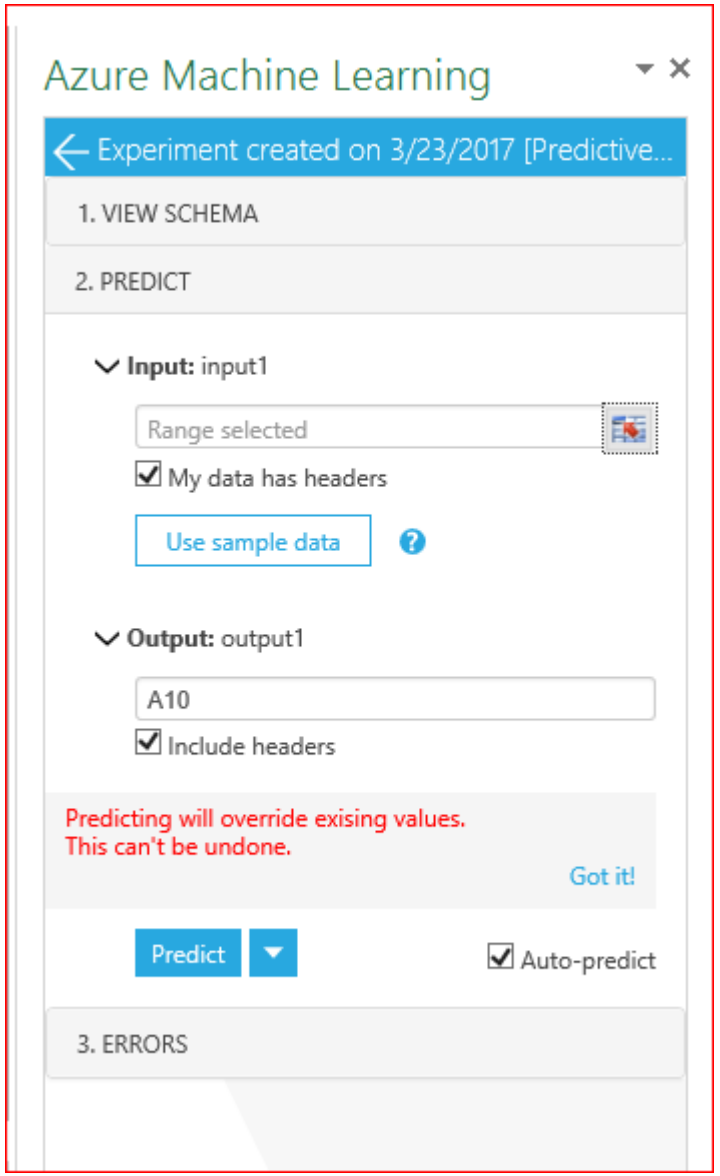
- Select the Experiment
- And click the “auto-predict” option

Click Use sample data to auto-populated the column names and some sample data



Click the input range and the output range.

Start playing!



What could we do with our Titanic model?

http://demos.datasciencedojo.com/demo/titanic/

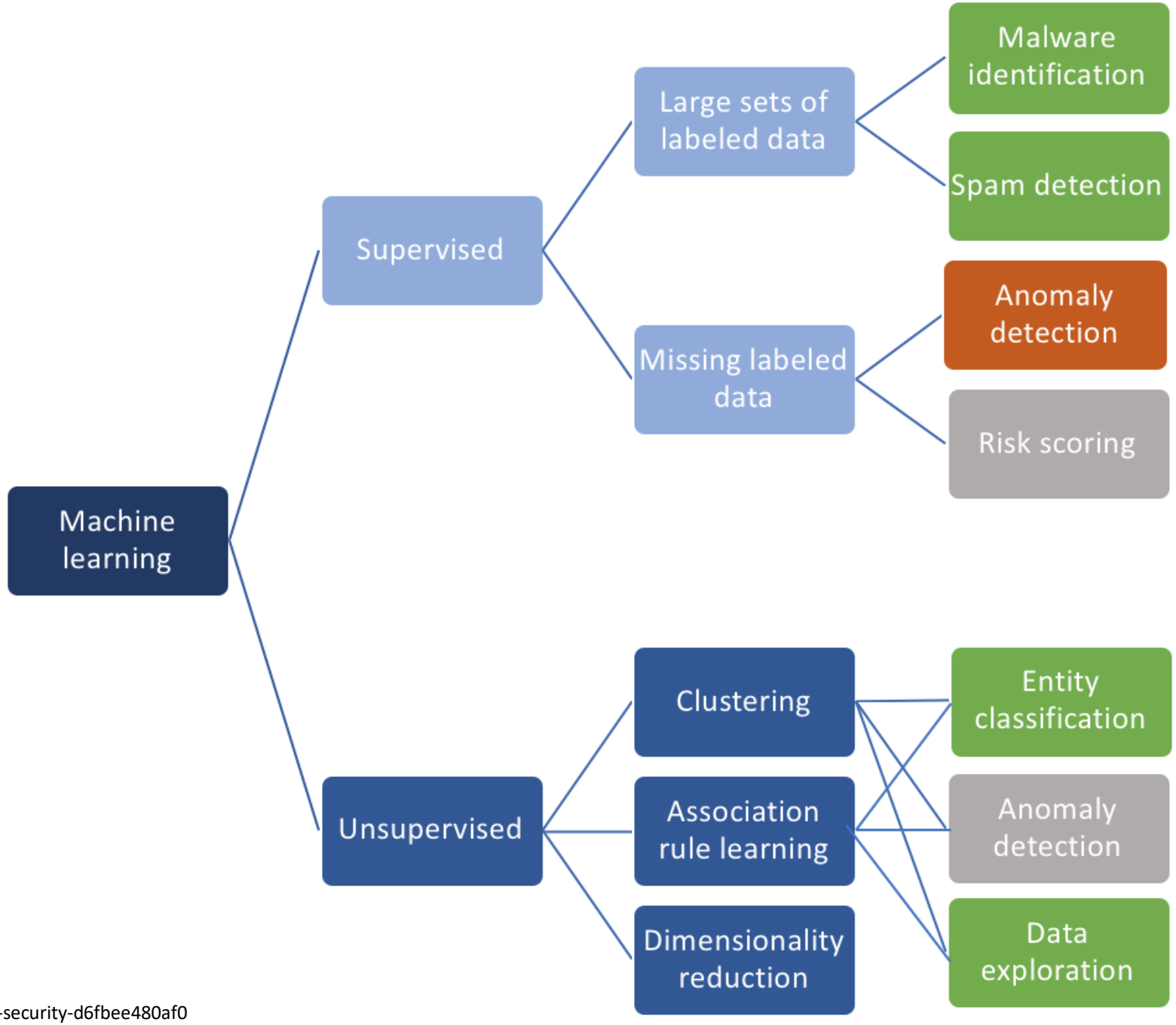
Yeah, so? That was just if I would live or die…

- It is used in a lot of places
  - Amazon's "you may also like"
  - Netflix Recommendation engine
  - Google's self-driving car
  - LinkedIn and Facebook "people you may know"
  - Banks use it to decide whether to offer a house loan or what the interest rate maybe
  - Police departments are using it to judge whether a person might re-offend
  - Courts are using it in sentencing phases
  - Schools are using it to determine teacher effectiveness

And on and on.

You still haven’t told me the security application…

Security application



https://towardsdatascience.com/ai-and-machine-learning-in-cyber-security-d6fbee480af0

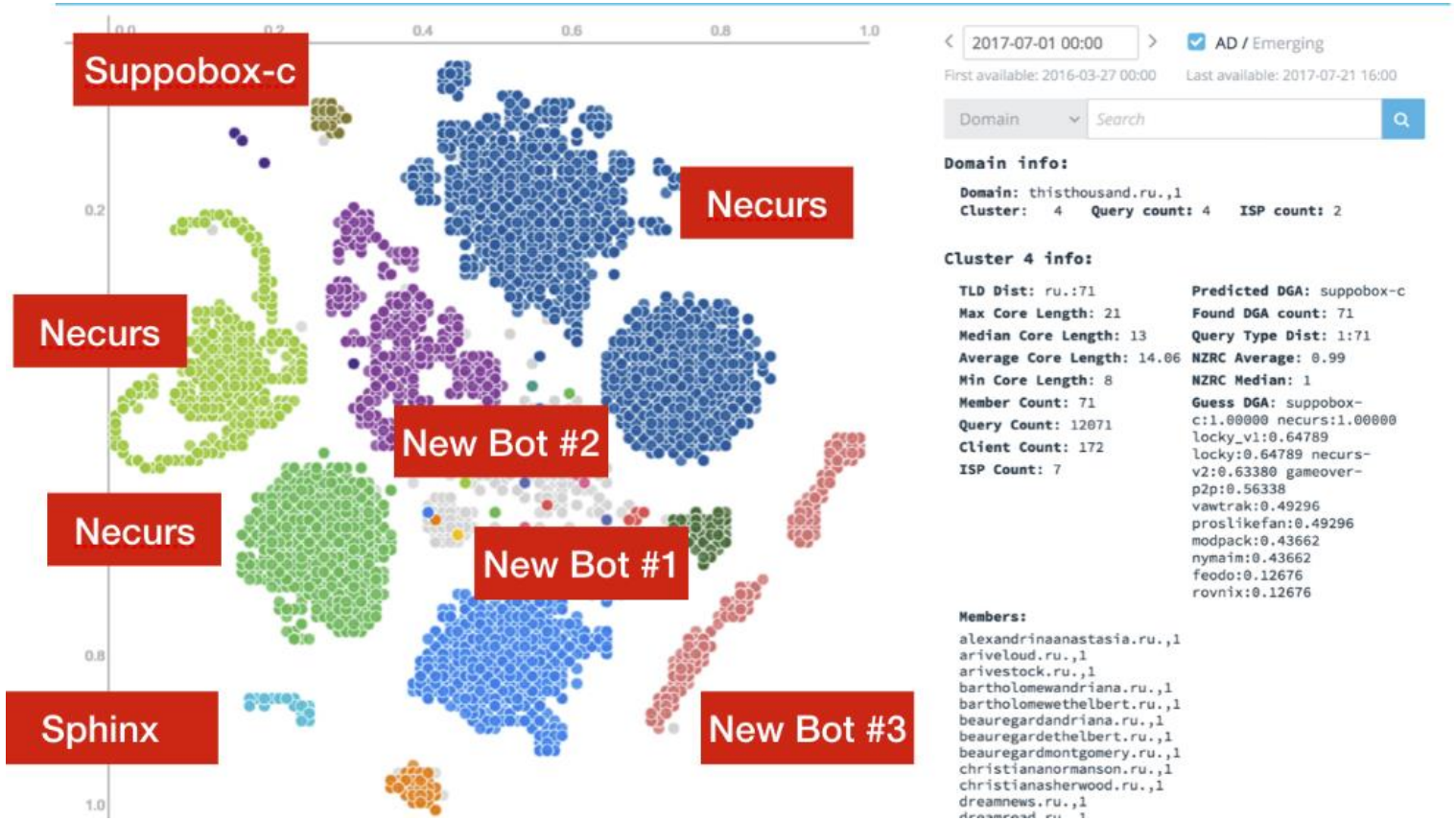
Spam or ham?

Good email

spam

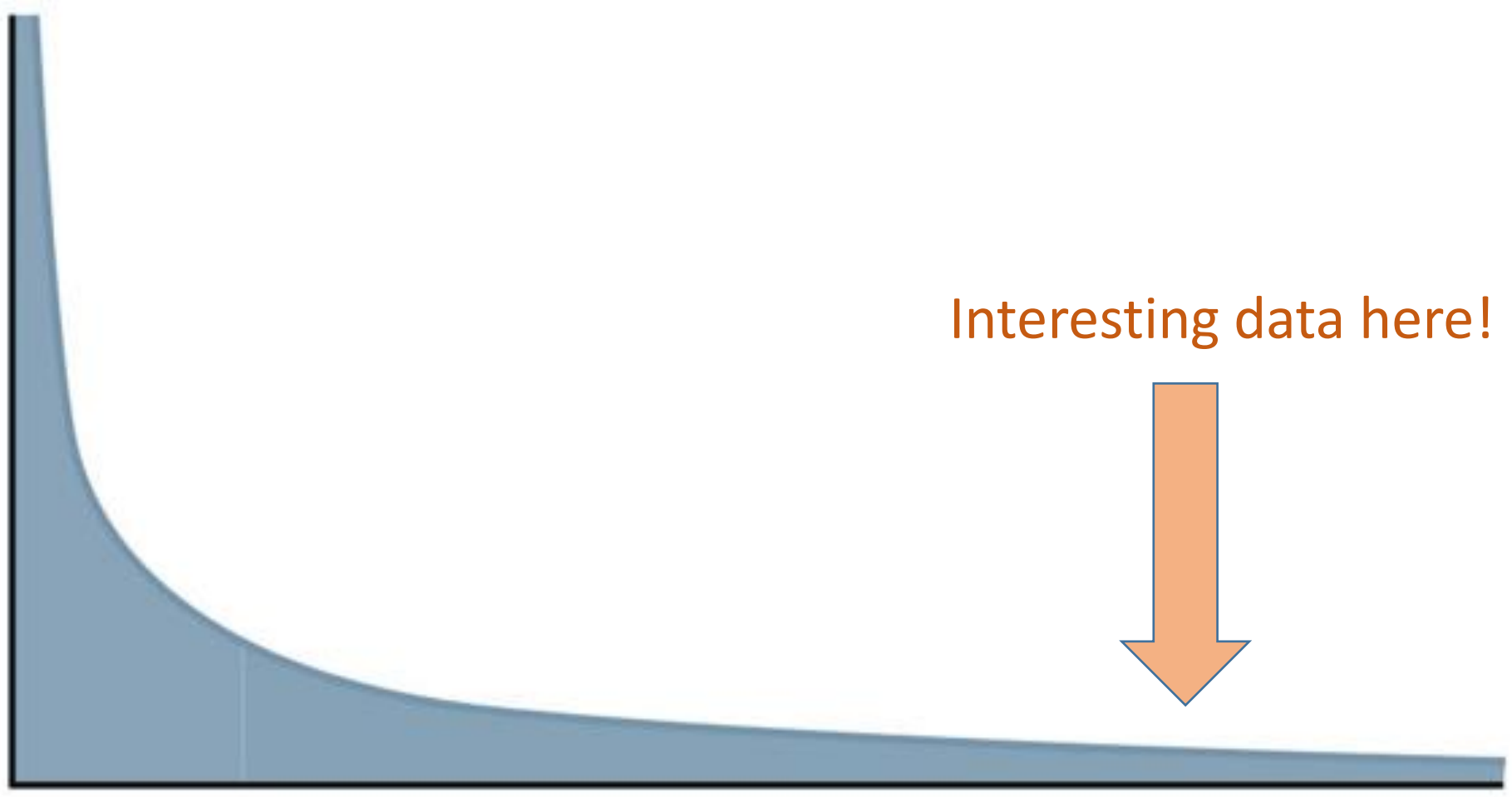
http://beta.cambridgespark.com/courses/jpm/01-module.html

Domain Clustering Algorithm to find DGA domains



https://blogs.akamai.com/2018/01/a-death-match-of-domain-generation-algorithms.html

Hunt in the long tail



Thanks!

Beth Young

@bethayoung

young.beth.a@gmail.com