

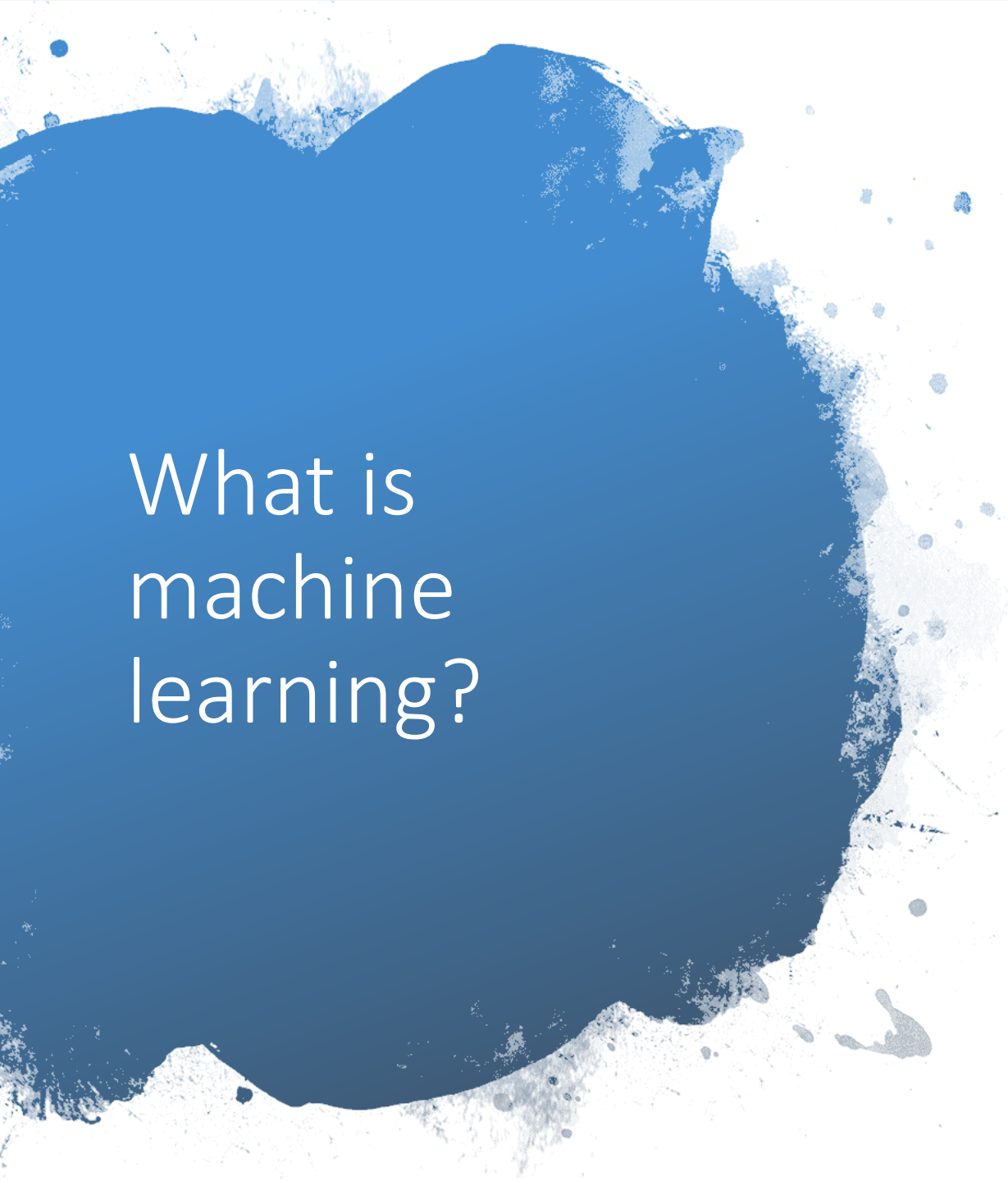
Drag and Drop Machine Learning





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

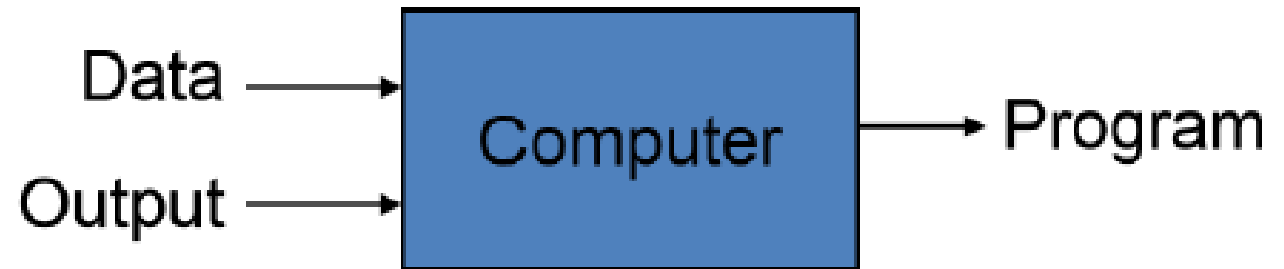


FIGURE 1-2 Machine learning programming paradigm.

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>

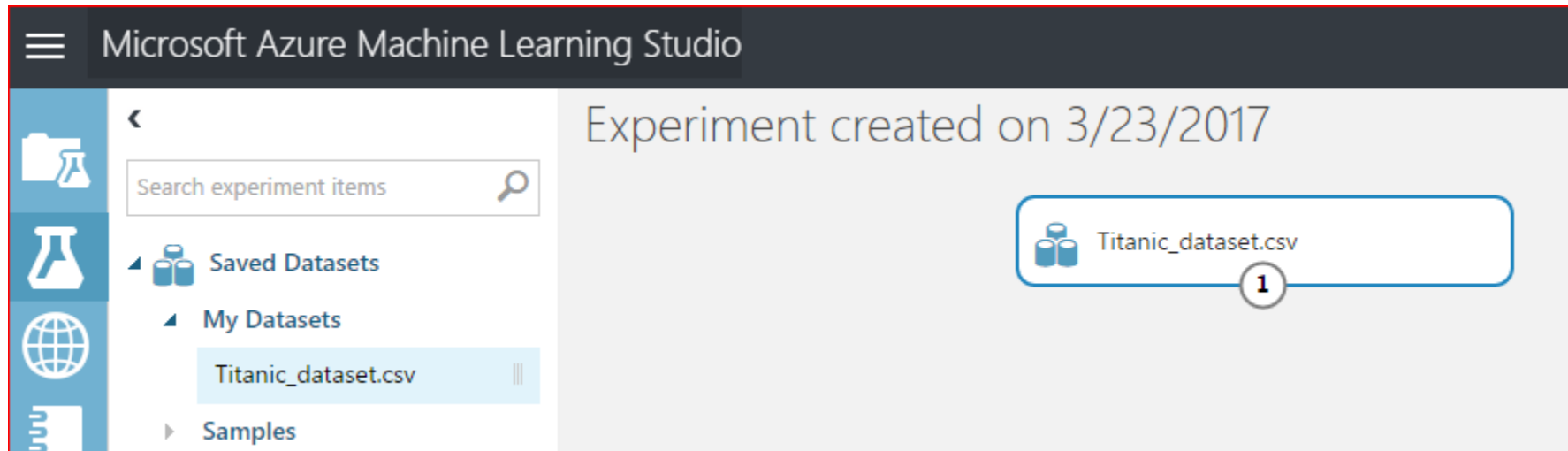
We are going to use pre-classified data to train our model.

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

Select columns

BY NAME

WITH RULES

☐ Allow duplicates and preserve column order in selection

Begin With

ALL COLUMNS

NO COLUMNS

Exclude

column names

PassengerId ✕

Name ✕

Ticket ✕

Cabin ✕

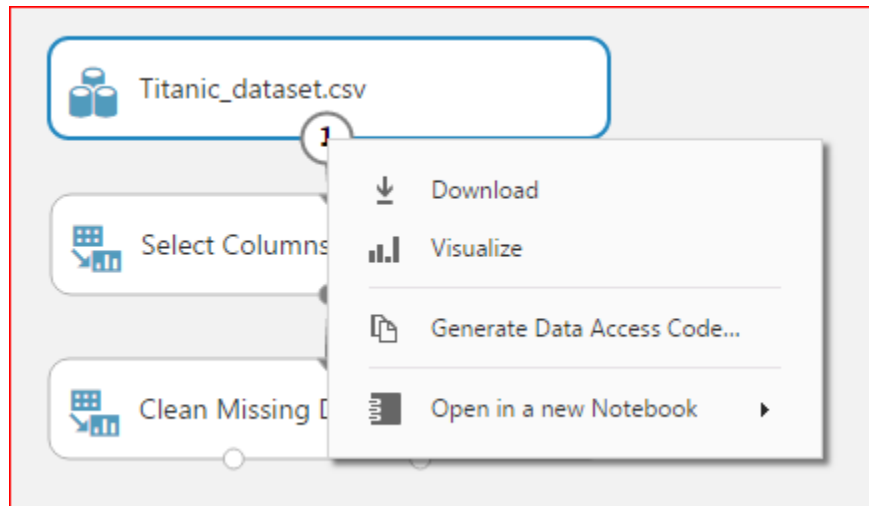
Embarked ✕

+

-

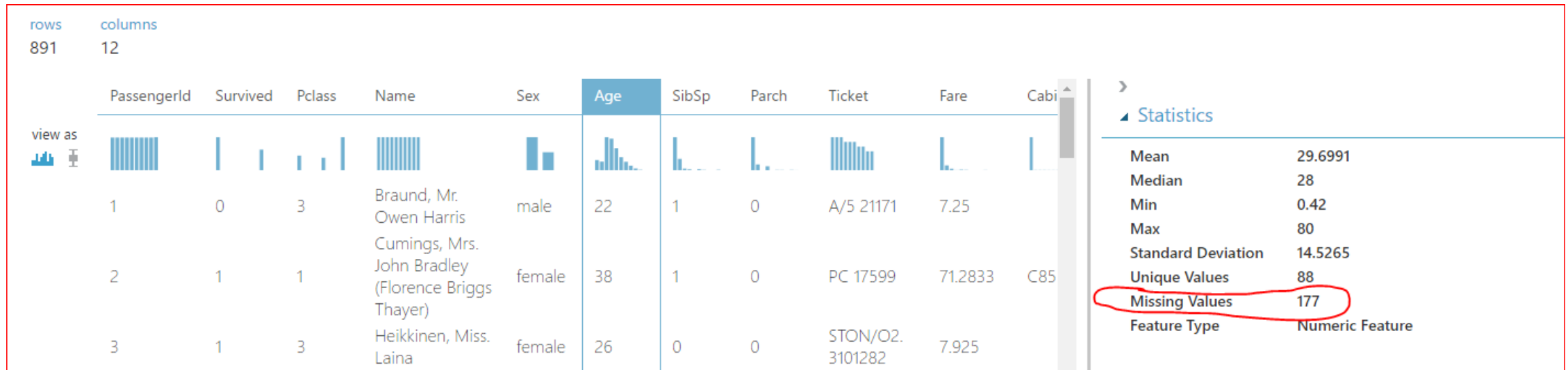
Clean Missing data – Part 1

- 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.

Properties Project >

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: Age

Launch column selector

Minimum missing value range

0

Maximum missing value range

1

Cleaning mode

Replace with median ▼

Cols with all missing values

Remove ▼

☐ Generate missing values

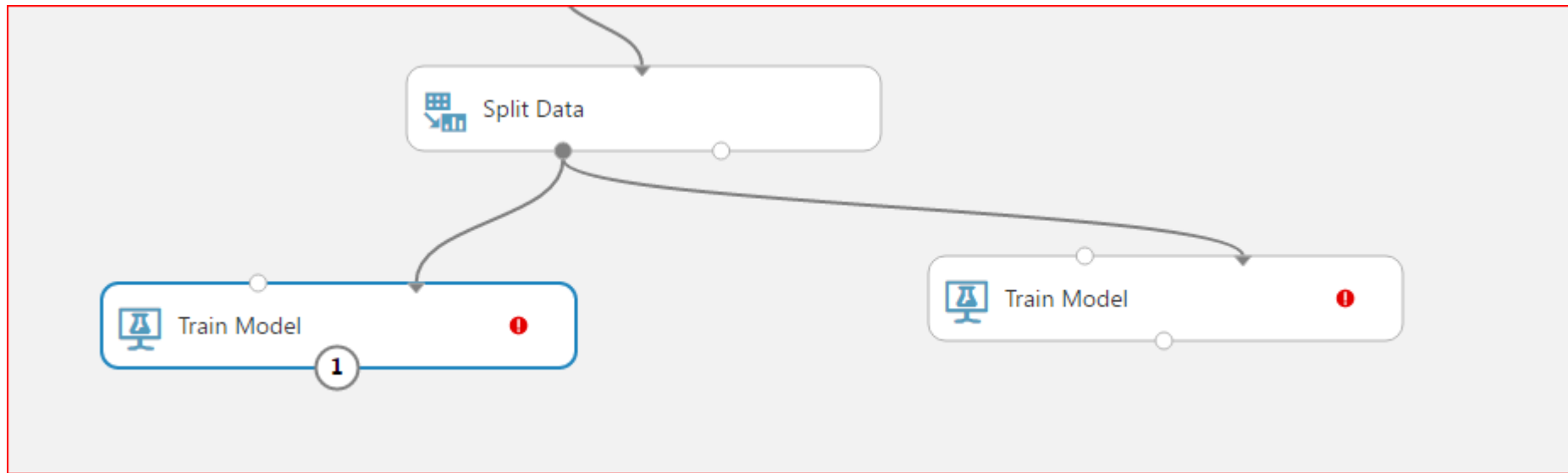
Split the data

- 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.

Split Data
Splitting mode
Split Rows
Fraction of rows in the first...
.80
☒ Randomized split
Random seed
0
Stratified split
False

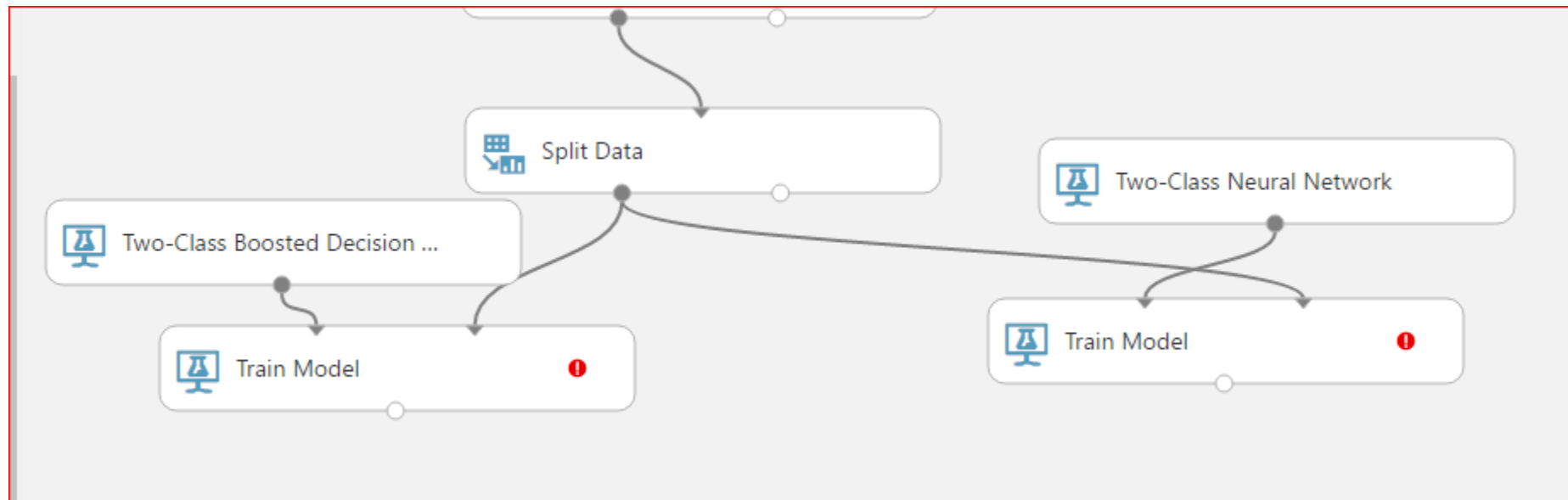
Train the model

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

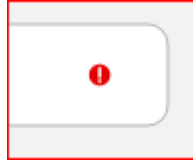


We also need to pick the Classification – what mathematical model do we want to use?

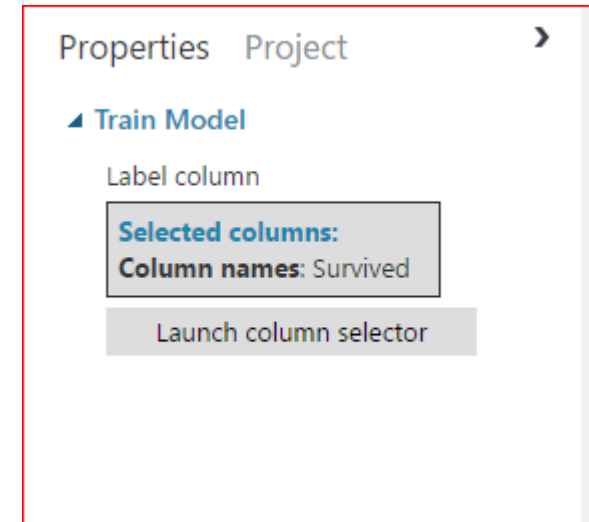
- If the prediction is one of two values (survived or died, yes or no), then pick a two-class model. The red ! means there is something else still needed.



Fixing the

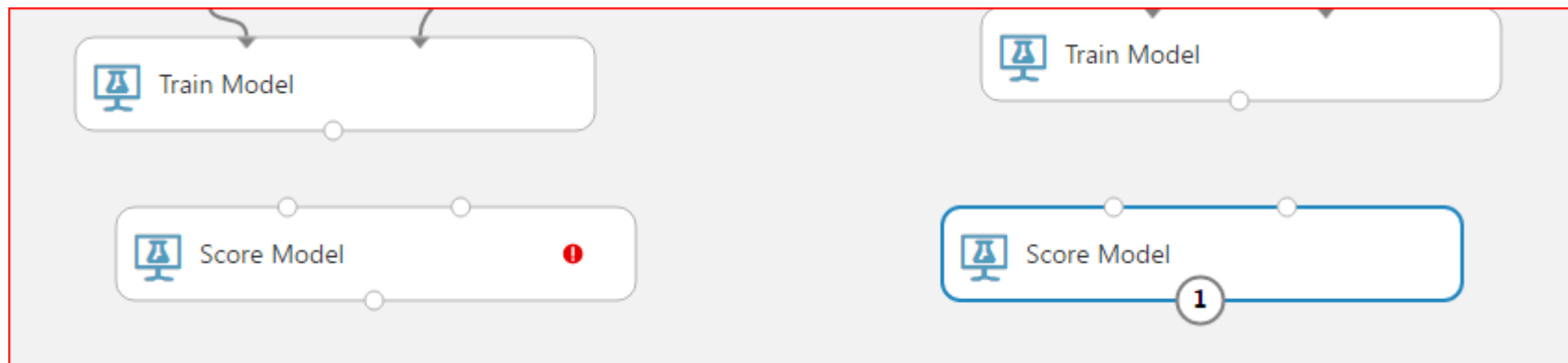


- Click on it and it will tell you what is missing – in this case, we need to tell the model which column has the result information (whether a person survived or not)

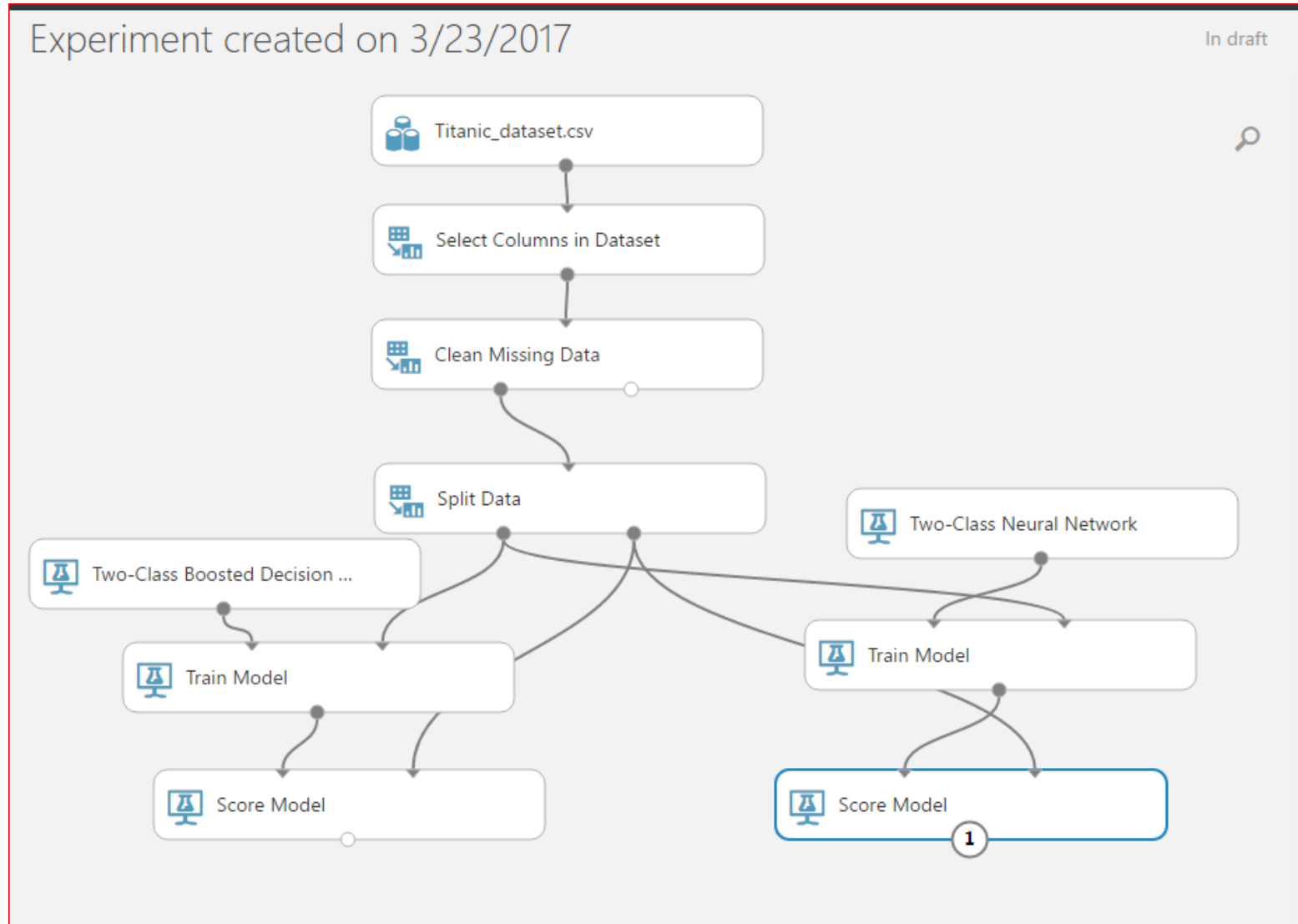


Scoring the module

- 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.

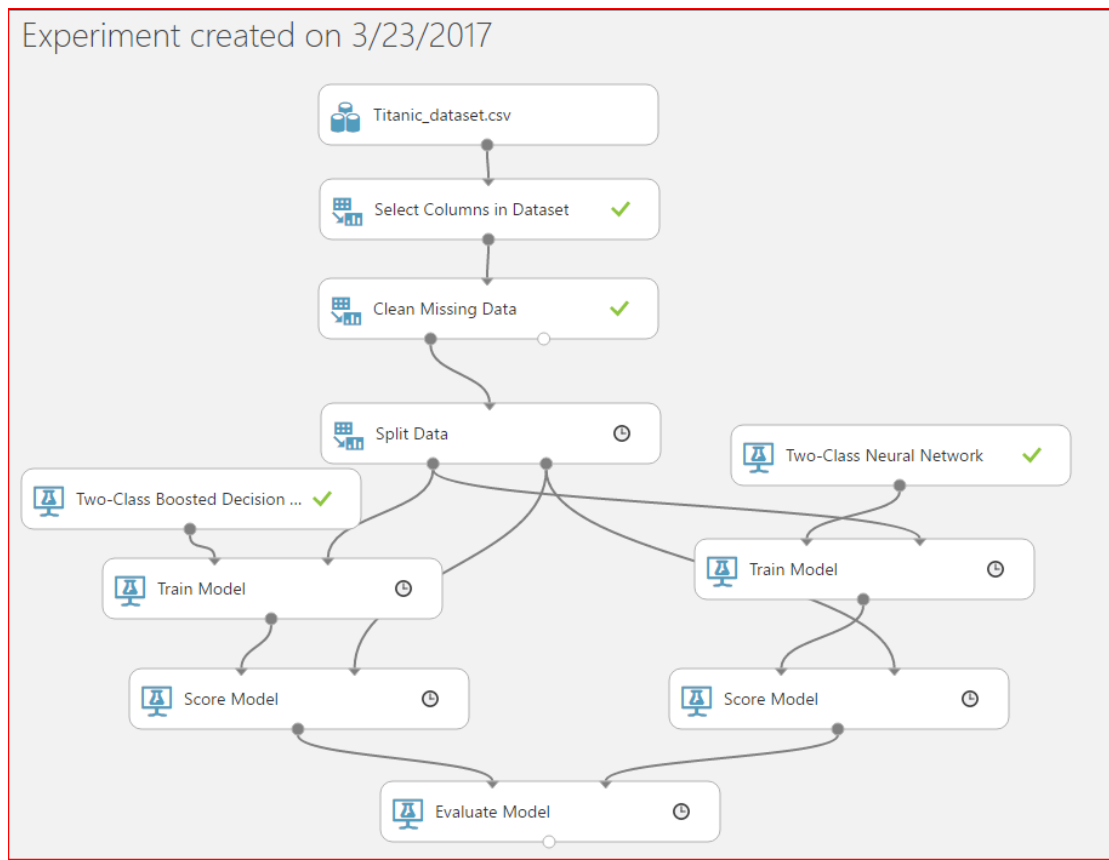


Connect the other part of the data set!



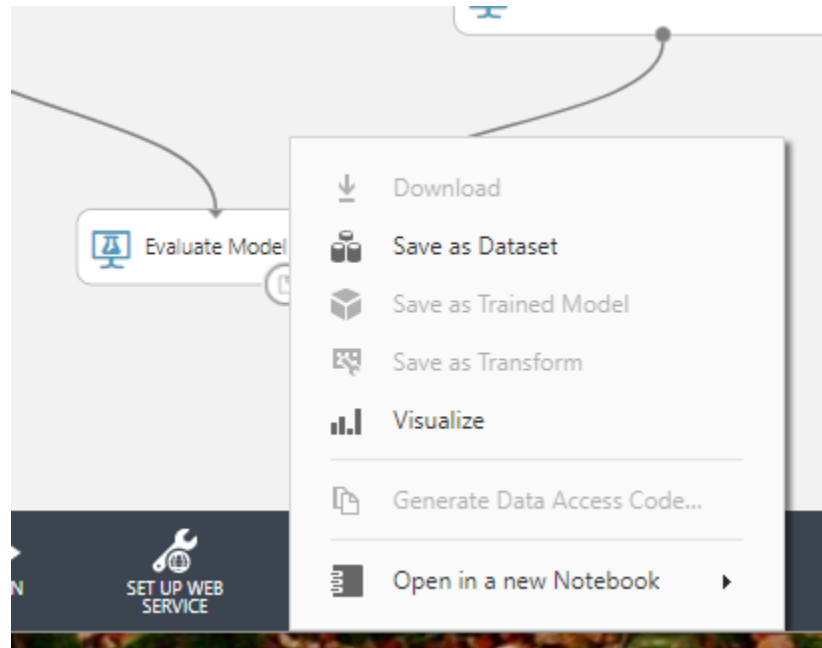
Now to evaluate the modules – which one was best?

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



Finding the Evaluation data

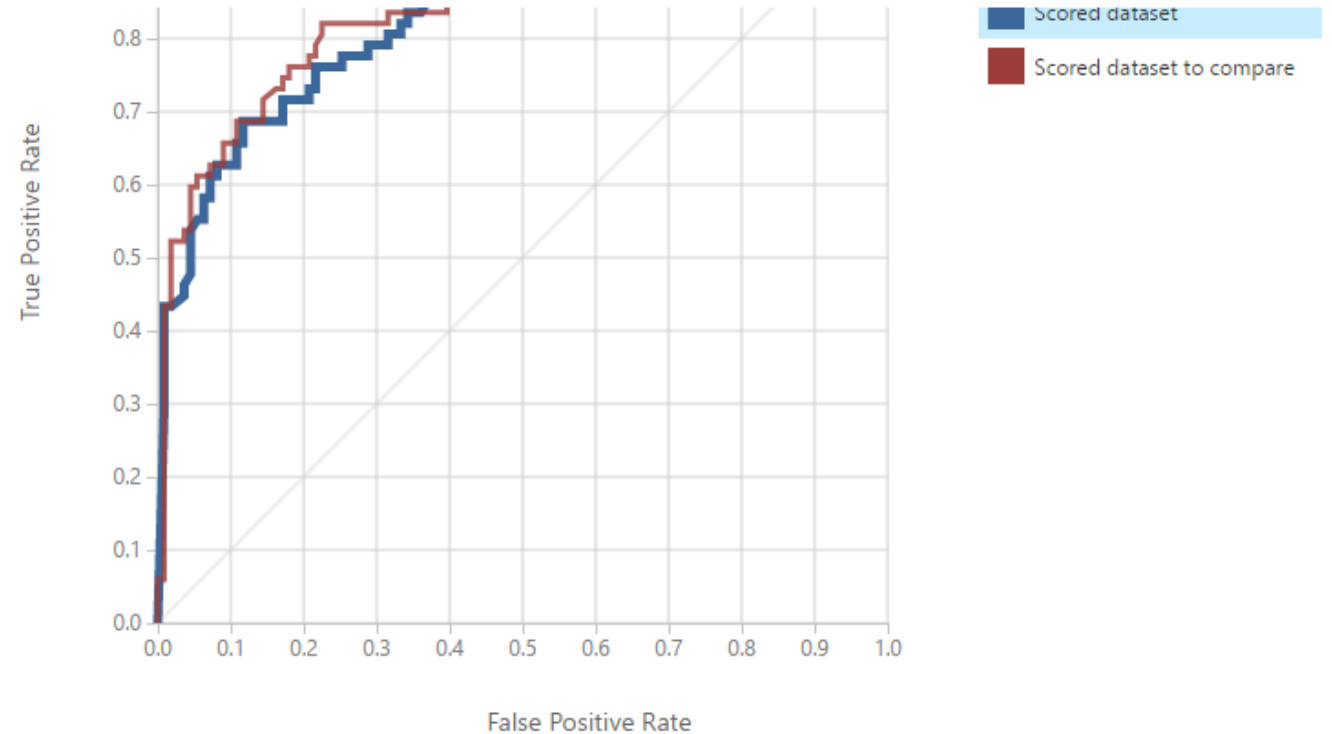
- Right click on the evaluate model process connector and select Visualize



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

Experiment created on 3/23/2017 > Evaluate Model > Evaluation results



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
46	21	0.781	0.719	0.5	0.847
False Positive	True Negative	Recall	F1 Score		
18	93	0.687	0.702		
Positive Label	Negative Label				
1	0				

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

The screenshot shows a 'Select columns' dialog box with two main panes: 'AVAILABLE COLUMNS' and 'SELECTED COLUMNS'. The 'AVAILABLE COLUMNS' pane lists 7 columns: PassengerId, Name, SibSp, Parch, Ticket, Cabin, and Embarked. The 'SELECTED COLUMNS' pane lists 5 columns: Survived, Pclass, Sex, Age, and Fare. A red arrow points to a 'Launch column selector' button in the sidebar. The sidebar also displays a summary of the selected columns and a 'View output log' link.

Select columns

Selected columns:
All columns
Exclude column names:
PassengerId, Name, Ticket, Ca

Launch column selector

START TIME 3/23/2017 2...

END TIME 3/23/2017 2...

ELAPSED TIME 0:00:00.000

STATUS CODE Finished

STATUS DETAILS Task output was present in output cache

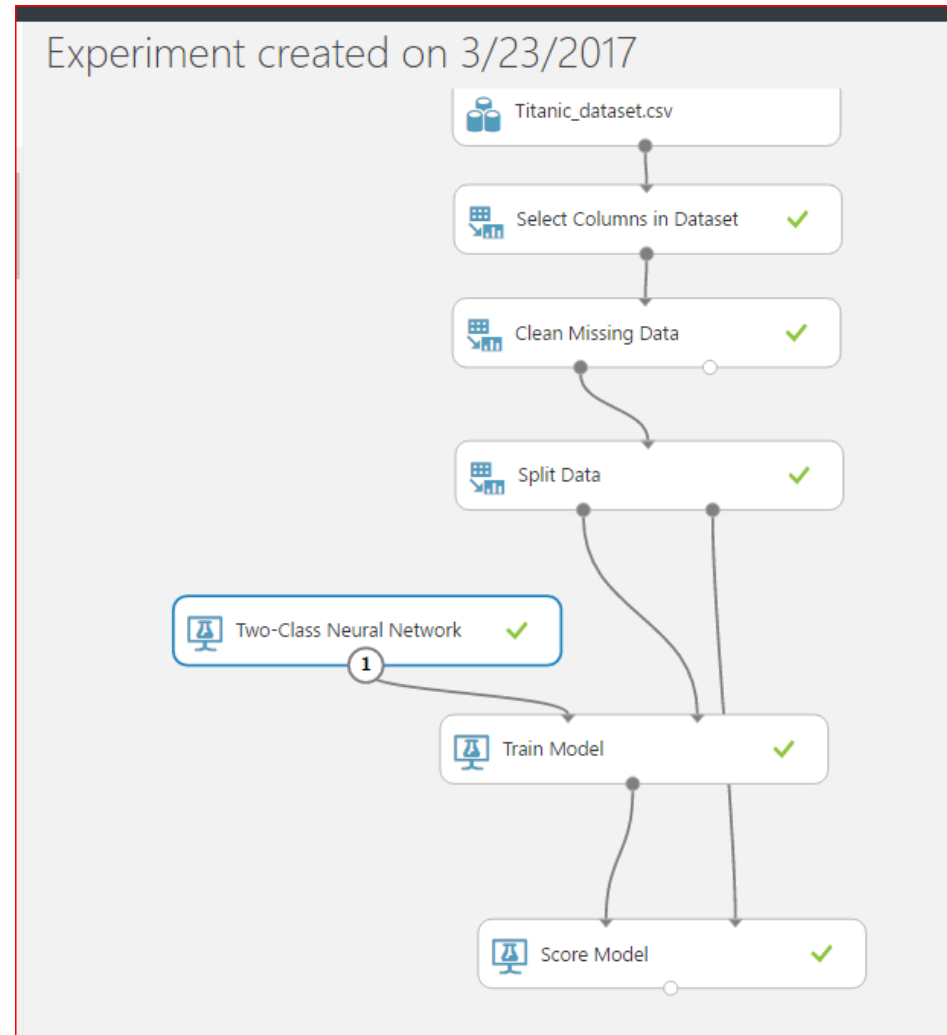
[View output log](#)

Quick Help

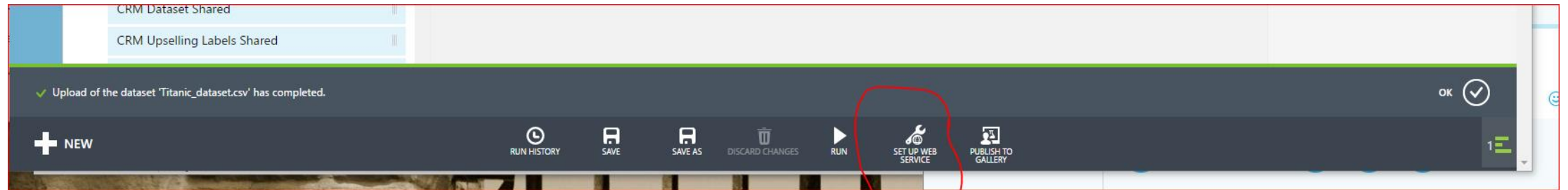
Selects columns to include or exclude from a

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

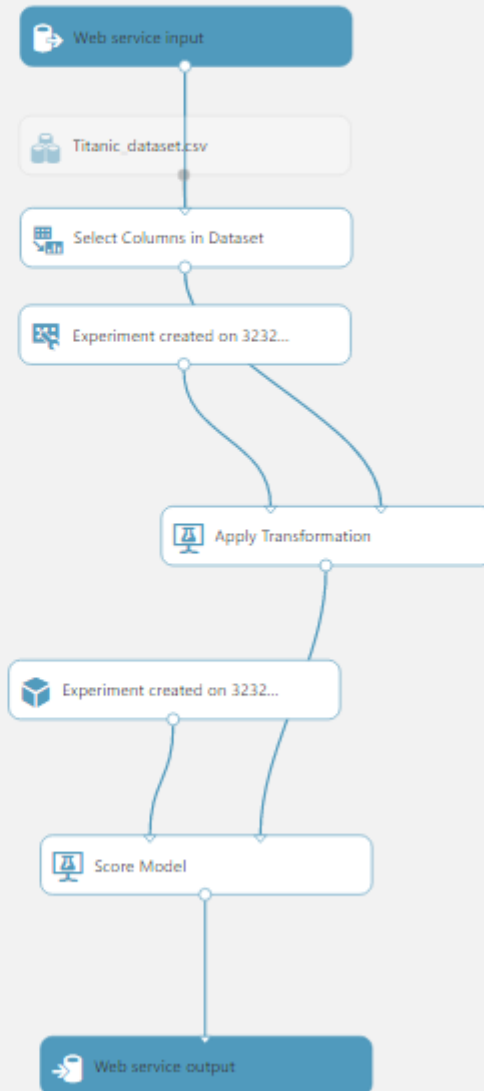
- First, remove the unused model and the evaluation module and click run to check for errors
- After it runs and you should have green checkmarks



Build the Web Service – Predictive Web service



Experiment created on 3/23/2017 [Predictive Exp.]



- 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

experiment created on 3/23/2017 [predictive exp.]

DASHBOARD CONFIGURATION

General [New Web Services Experience](#) preview

Published experiment

[View snapshot](#) [View latest](#)




Description

No description provided for this web service.

API key

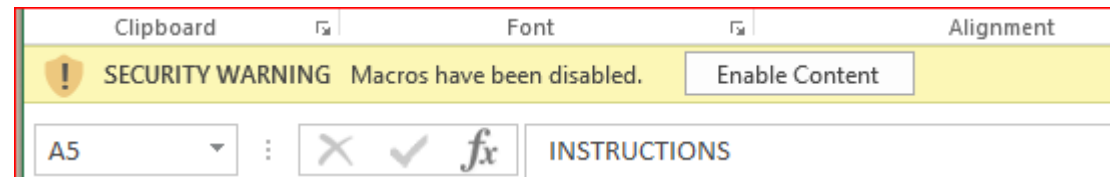
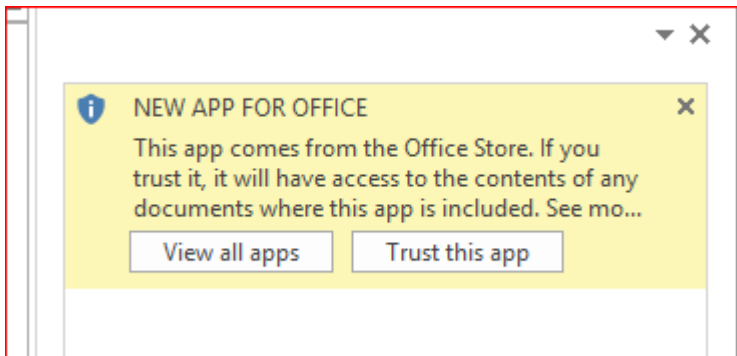
H0jTrseQcwNOJl/qfQKglNii0EGGrDG8SJ0vDY/fi0Lz10IBPRyYjAlUk71Q3U9yCUUIBbS0mcgMs0eXRIQadw==

Default Endpoint

API HELP PAGE	TEST	APPS	LAST UPDATED
REQUEST/RESPONSE	Test Test <small>preview</small>	 Excel 2013 or later  Excel 2010 or earlier workbook	3/23/2017 3:10:28 PM
BATCH EXECUTION	Test <small>preview</small>	 Excel 2013 or later workbook	3/23/2017 3:10:28 PM

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

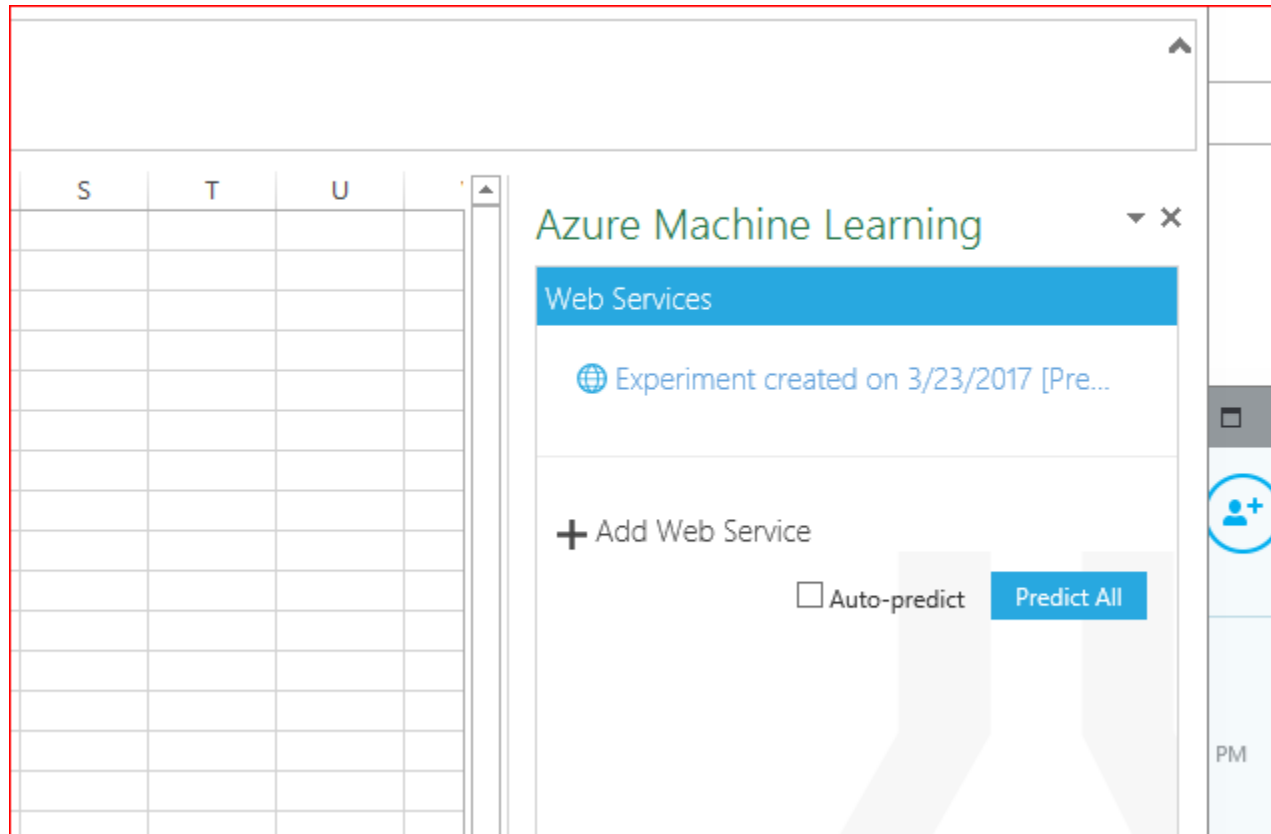
Once you have enabled macros and the table has been generated, please enter your input values in the **PARAMETERS** section. Once all parameters have been entered, **PREDICTED VALUES** will be automatically computed.

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 YrF	48	0	0	0	0	100	Cabin S		0	1	F	48	0	0	100	1	0.897295713

Page 10 of 10

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

The screenshot displays the Azure Machine Learning interface. On the left, a data table is visible with the following columns: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. The table contains five rows of sample data. On the right, the 'Azure Machine Learning' sidebar is open, showing the 'Experiment created on 3/23/2017 [Predictive...]' section. Under the '1. VIEW SCHEMA' tab, the 'Inputs' section is expanded, and the 'Use sample data' button is highlighted. The 'Output' section is also expanded, showing the 'Include headers' checkbox checked.

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr.	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mr	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Mr	female	26	0	0	STON/O2. 31	7.925		S
4	1	1	Futrelle, Mr	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. W	male	35	0	0	373450	8.05		S

Azure Machine Learning

← Experiment created on 3/23/2017 [Predictive...]

1. VIEW SCHEMA

- > Inputs
- > Outputs
- > Global Parameters

2. PREDICT

Input: input1

Type range or click button to select

☒ My data has headers

[Use sample data](#) ?

Output: output1

Enter output cell (e.g. A20)

☒ Include headers

Click the input range
and the output range.

Start playing!

Azure Machine Learning

← Experiment created on 3/23/2017 [Predictive...]

1. VIEW SCHEMA

2. PREDICT

▼ Input: input1

Range selected

☒ My data has headers

Use sample data ?

▼ Output: output1

A10

☒ Include headers

Predicting will override existing values.
This can't be undone.

Got it!

Predict ▼ ☒ Auto-predict

3. ERRORS

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

Machine learning

Supervised

Large sets of labeled data

Malware identification

Spam detection

Missing labeled data

Anomaly detection

Risk scoring

Unsupervised

Clustering

Entity classification

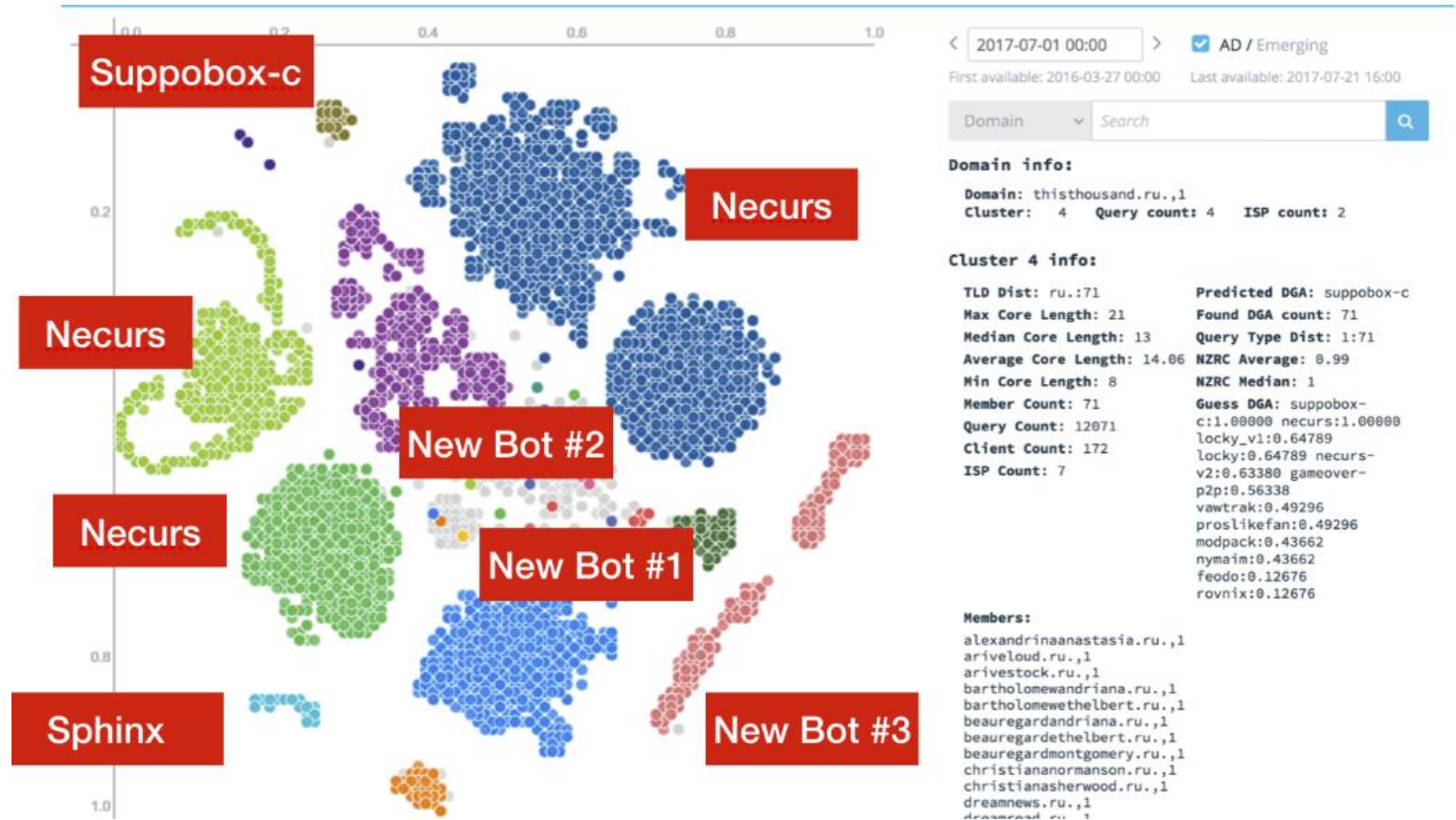
Association rule learning

Anomaly detection

Dimensionality reduction

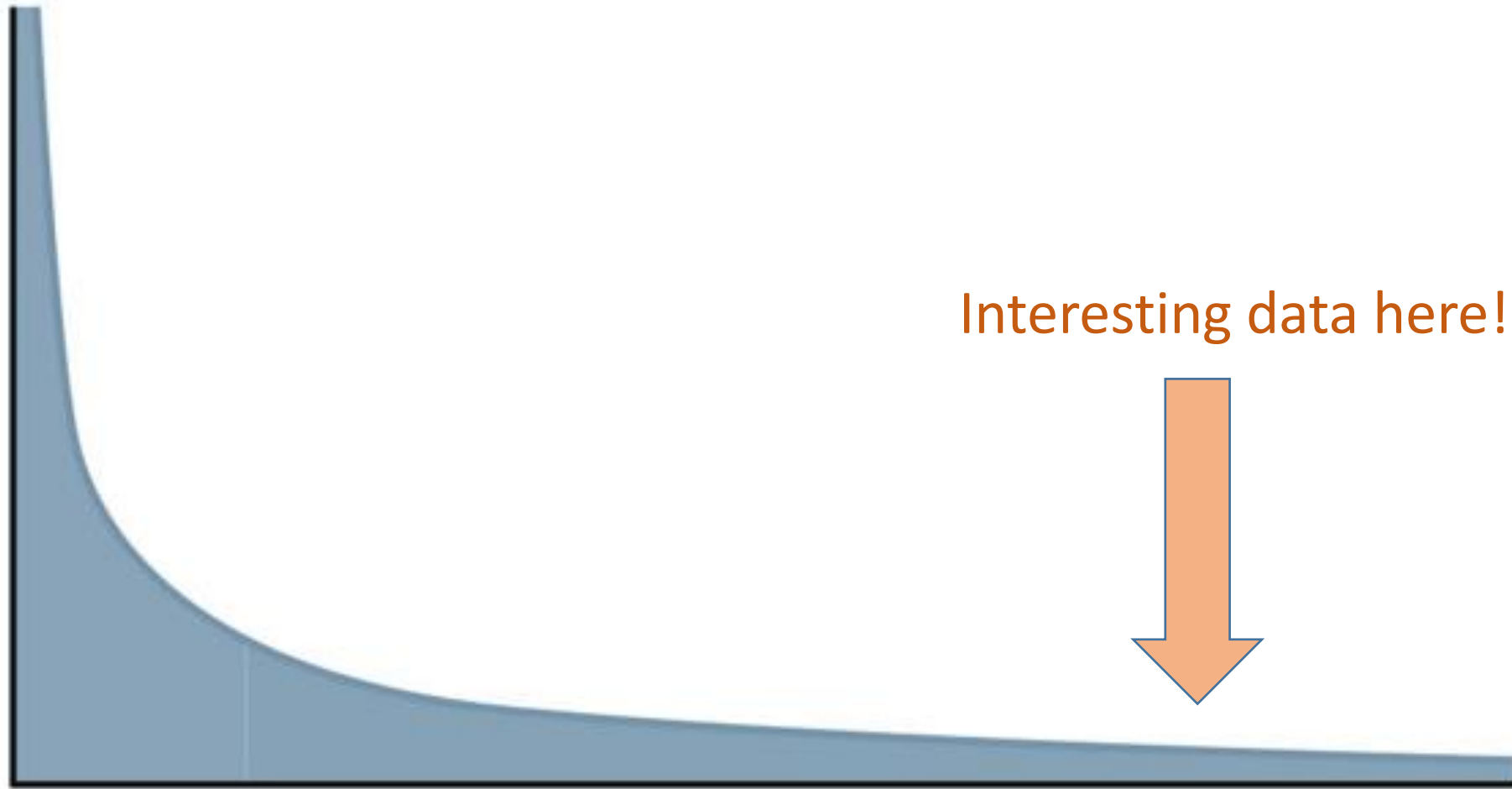
Data exploration

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