Big Data Meets Learning Science

Apache Spark Summit East 2017

Alfred Essa
VP, Research and Data Science
McGraw-Hill Education
@malpaso

1	Innovation Pipeline
2	McGraw-Hill Learning Science
3	Spark, DataBricks

Mc Graw Hill Education

Speed of innovation, not data, is the differentiator.

Spark Factor

Technology

Apache Spark

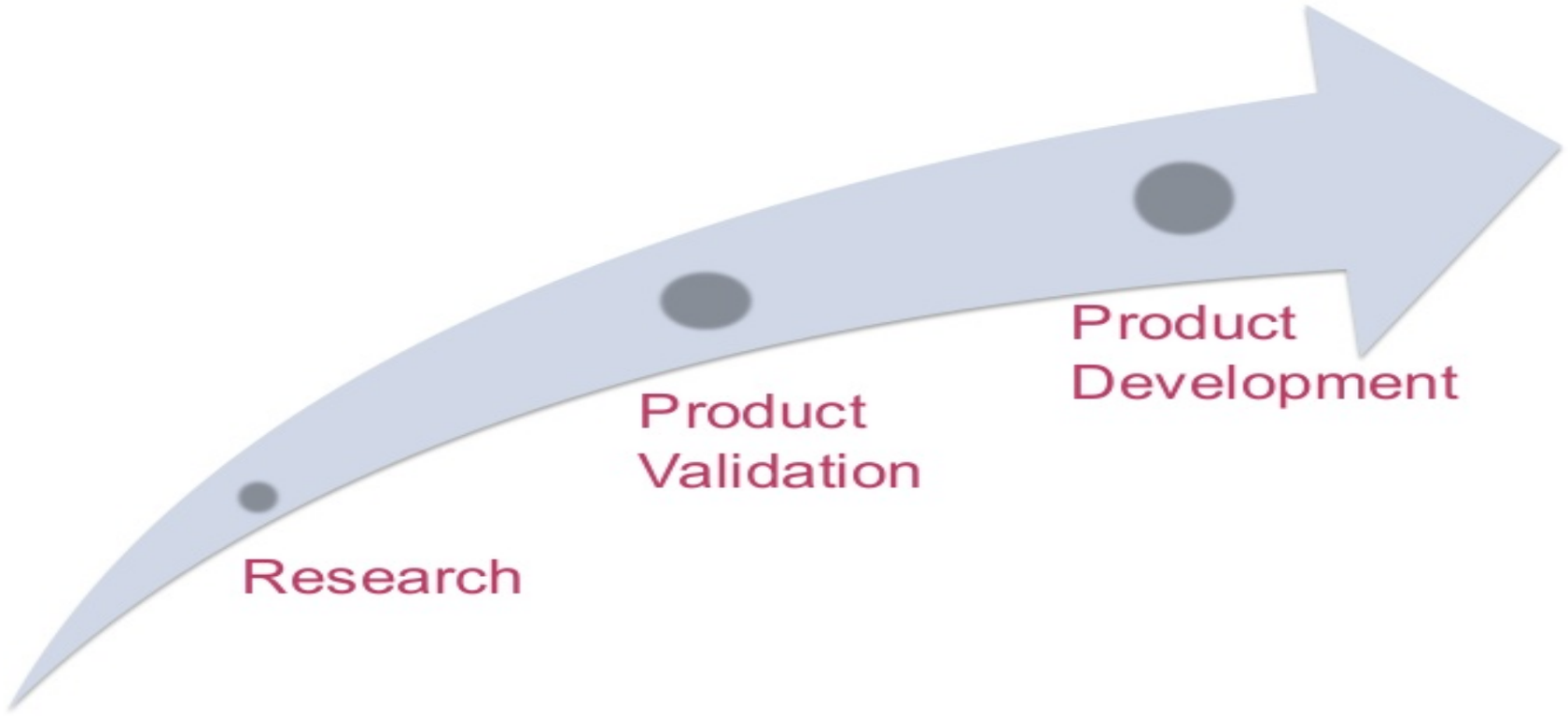
Time to Market

People

Process

DataBricks

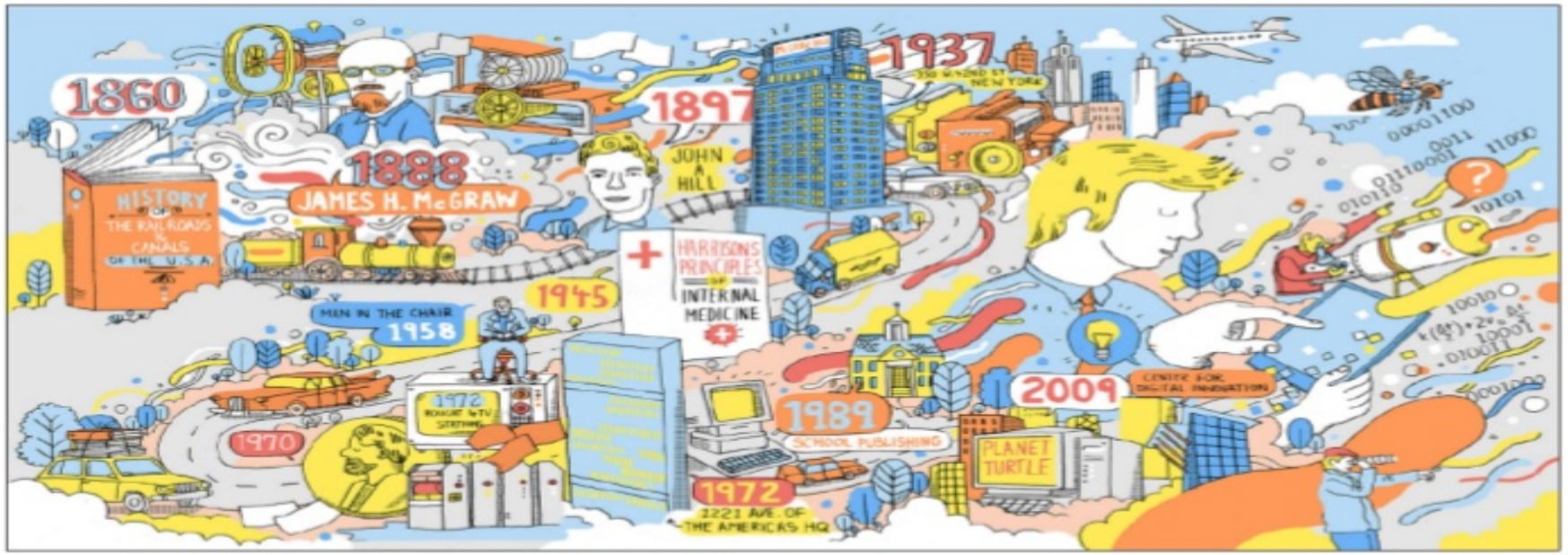
Innovation Pipeline



Databricks underpins our innovation pipeline and workflow.

2 McGraw-Hill Learning Science

From Print to Digital: 128-year Journey



K-12, Higher Ed & Professional businesses

~4,800 employees

Adaptive Platform Leverages MHE Reach and Scale

May 2013
Introduction of **SmartBook**

Now
1,500+ adaptive products available

~4,000
Authors trained to use MHE Adaptive

~5,500,000
Learners who have used MHE Adaptive

~10,000,000,000
Student interactions

Research Phase

Learning Tool for Optimizing Acquisition and Recall

1

Learning Science Principles

- Effortful Recall
- Spaced Practice
- Interleaving

2

Cognitive Science Model

- Stacked Algorithm

3

Mobile App

databricks

StudyWise_Dashboard_Notebook (Python)

Import Notebook

StudyWise Analtytics Data

Anonymized user interaction data is send in JSON format from a users mobile device to an S3 bucket.

This s3 bucket is mounted in the DBFS file system.

To infer the JSON schema for this data, read one record:

```
> df_studywise_one = sqlContext.read.json("dbfs:/mnt/r_dvtl-prod.mheducation.com/dvtl-document-api/prod/data/2017/02/01/20/dvtl-document-api-firehose-prod-1-2017-02-01-20-31-15-49a79219-8ebe-4e1b-8a0a-8575538c9c12")
```

databricks StudyWise_Dashboard_Notebook (Python)

The five StudyWise apps were released in the Apple App Store on Jan. 31, 2017. Here we read all of the data from Feb. 1 through Feb. 7, 2017.

The command below reads this data into a Spark DataFrame.

```
> df_studywise = sqlContext.read.schema(schema_one).json("dbfs:/mnt/dvtl-document-api.dvtl-prod.mheducation.com/prod/data/2017/02/*/*/*")
```

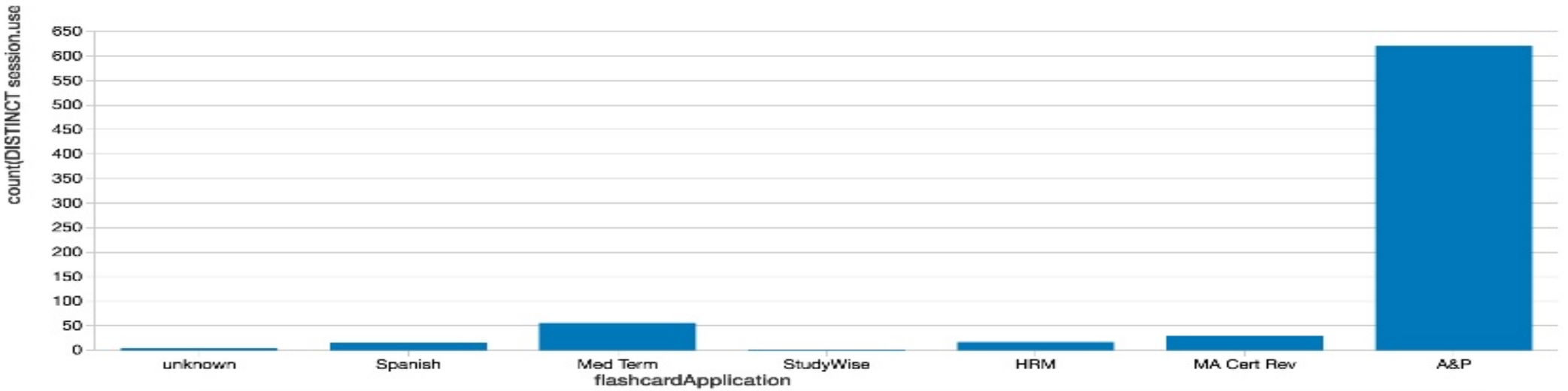
To be able to run straight Spark SQL on this data, we load it into a Temporary View:

```
> df_studywise.createOrReplaceTempView("studywise")
```

Now do a SQL query to see how many questions have been answered per app in this data:

```
> %sql select session.flashcardApplication, count(*) from studywise group by session.flashcardApplication
```

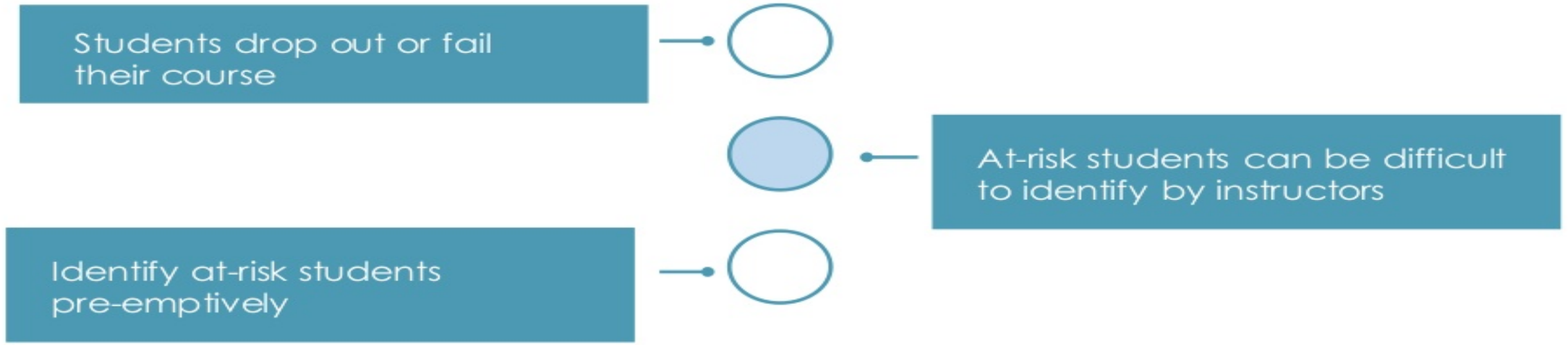
```
> %sql select  session.flashcardApplication, count(distinct session.userID) from studywise group by session.flashcardApplication
```



Mc Graw Hill Education

3 Spark, DataBricks

The Problem


Students drop out or fail their course
At-risk students can be difficult to identify by instructors
Identify at-risk students pre-emptively

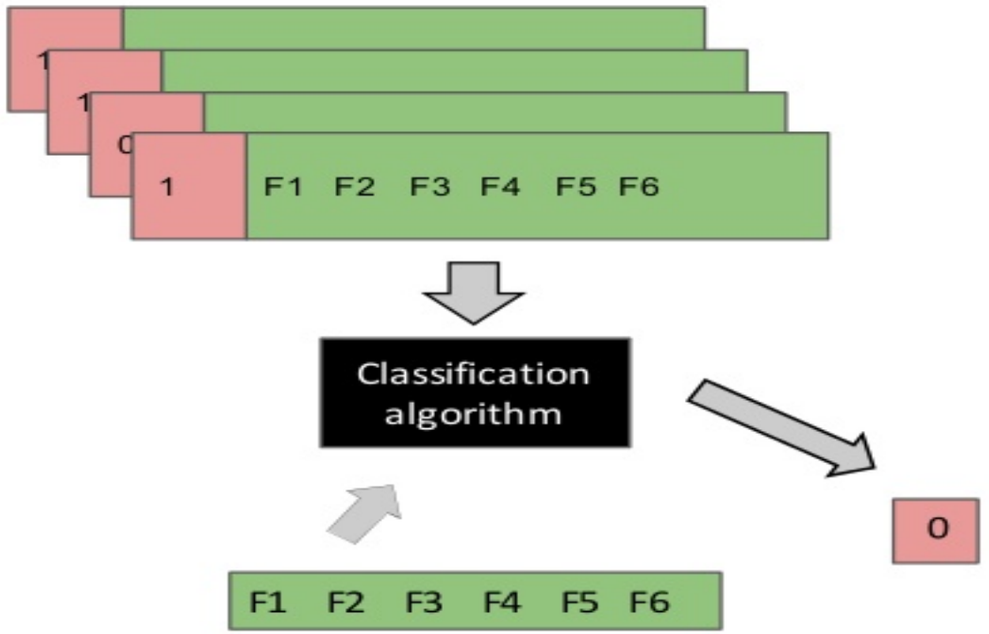
The Solution

A classifier to predict abandonment

Jacqueline Feild
Data Scientist

Nicholas Lewkow
Data Scientist

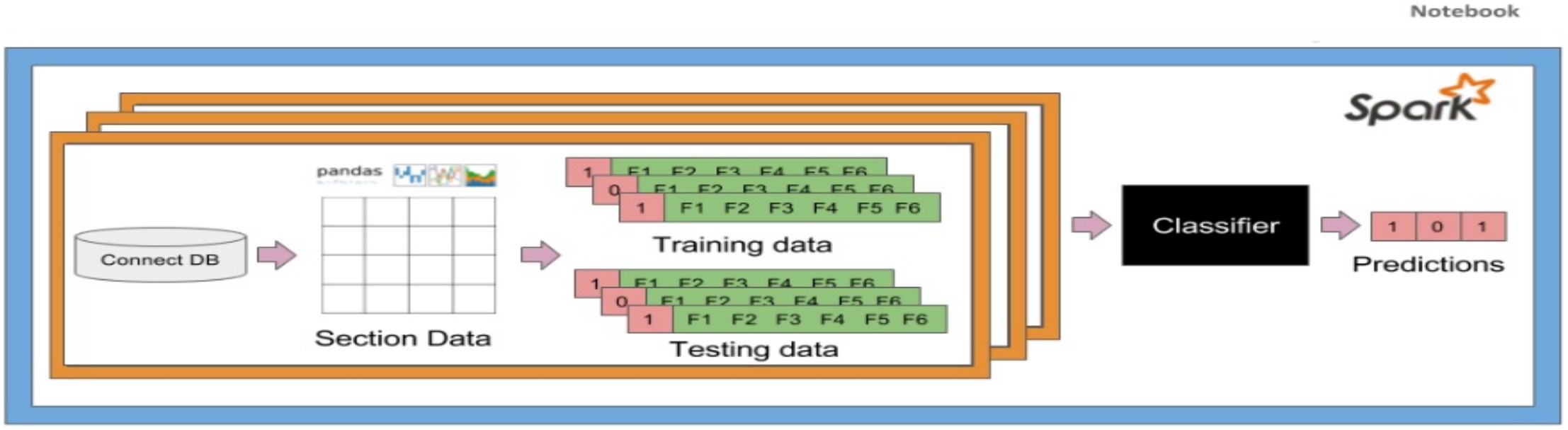
Solution: A Classifier to Predict Abandonment



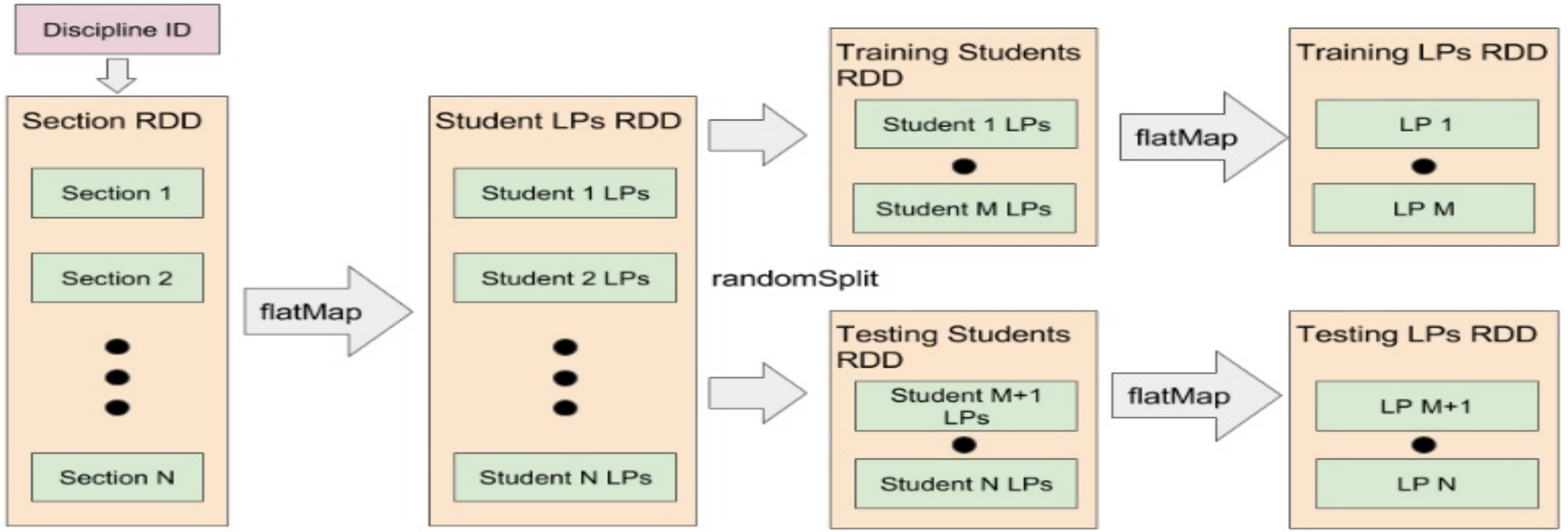
- Logistic Regression used for initial classification algorithm
  - Simple algorithm to interpret
  - Provides probability estimates instead of hard classification label
  - Allows for simple interpretation of feature importance
- One classifier works for all disciplines

Parallel Pipeline for Creating Classifier

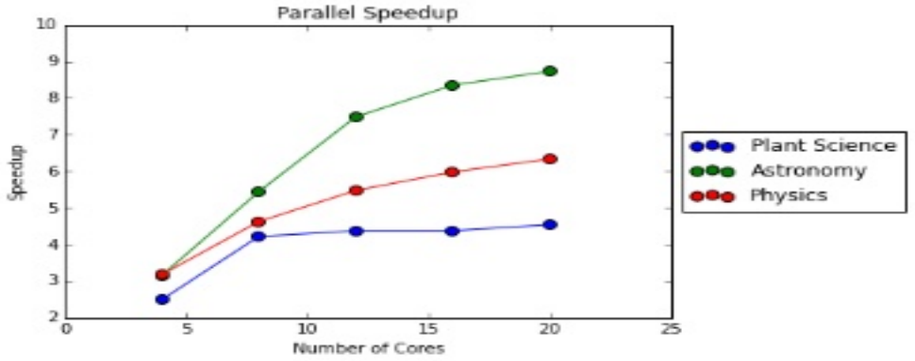
The Spark Pipeline



Spark Transformation



Speedup with Spark

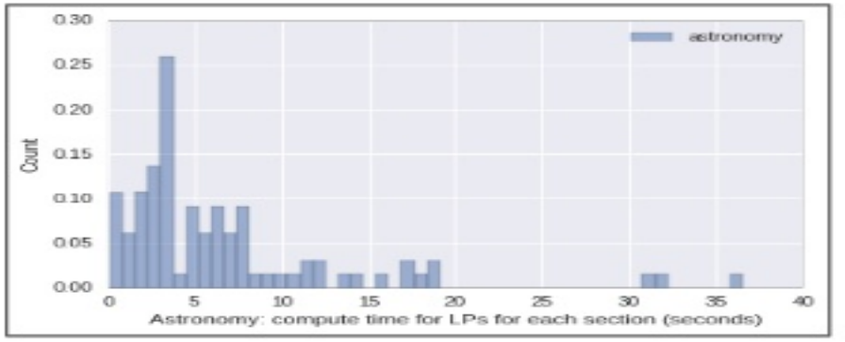


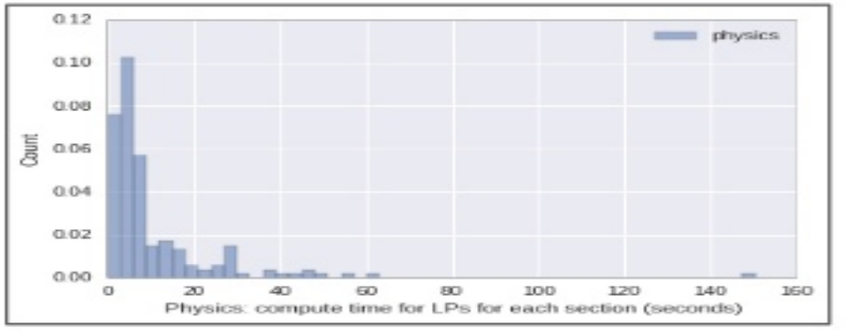
$$S_n = \frac{t_1}{t_n}$$

S_n: Speedup from n cores

t_1: Time to run on 1 core

t_n: Time to run on n cores





Evaluate Model Accuracy

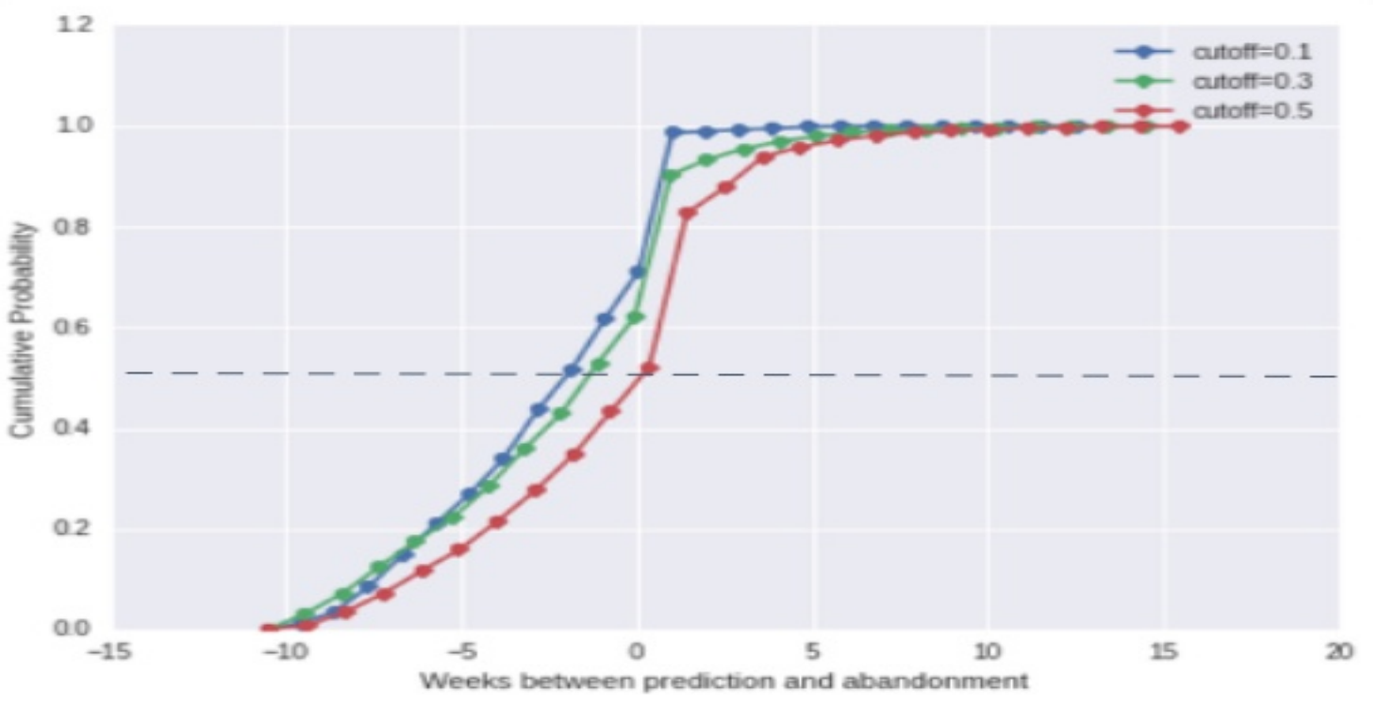
- Use area under the receiver operating characteristic curve (AUC-ROC) as another measure of model accuracy
  - 0.9 - 1.0 = excellent
  - 0.8 - 0.9 = good
  - 0.7 - 0.8 = fair
  - 0.6 - 0.7 = poor
  - 0.5 - 0.6 = fail

- Look at how the AUC-ROC for a model changes throughout the semester

Evaluate Intervention Window

Intervention Window:

How much time in advance can we provide for an intervention to occur prior to abandonment?



Conclusions

Technology is important, but build an agile innovation workflow with Databricks.