Curtin University

Predicting Students Success with Leganto,a proof of concept machine learning project

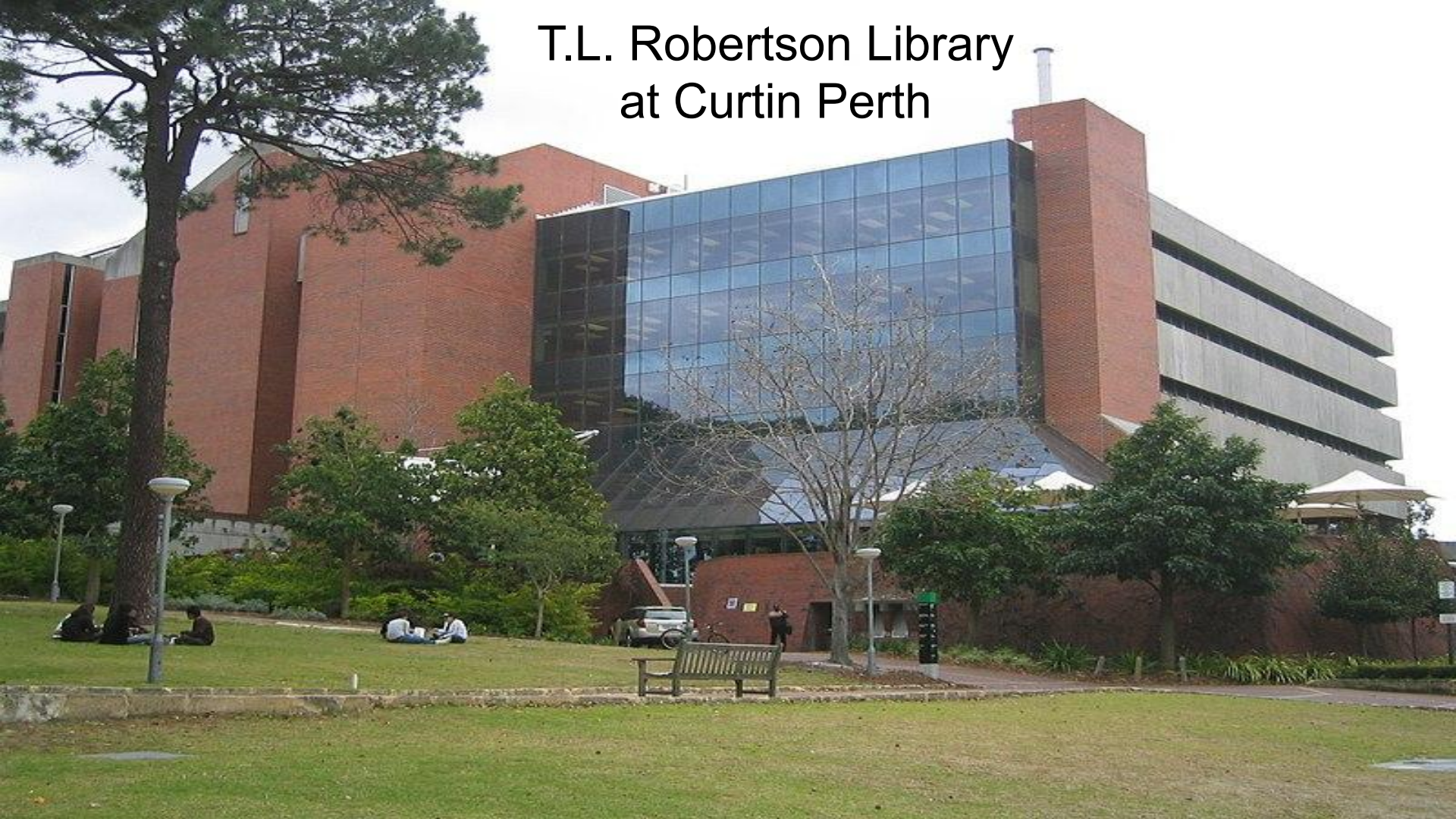
Mrs Linda Sheedy
August 2019

With thanks to

Mr Peter Green, Mr Gal Darom,
Mr Tomer Katz, and Mr David Lewis

For their contributions to this project and this presentation

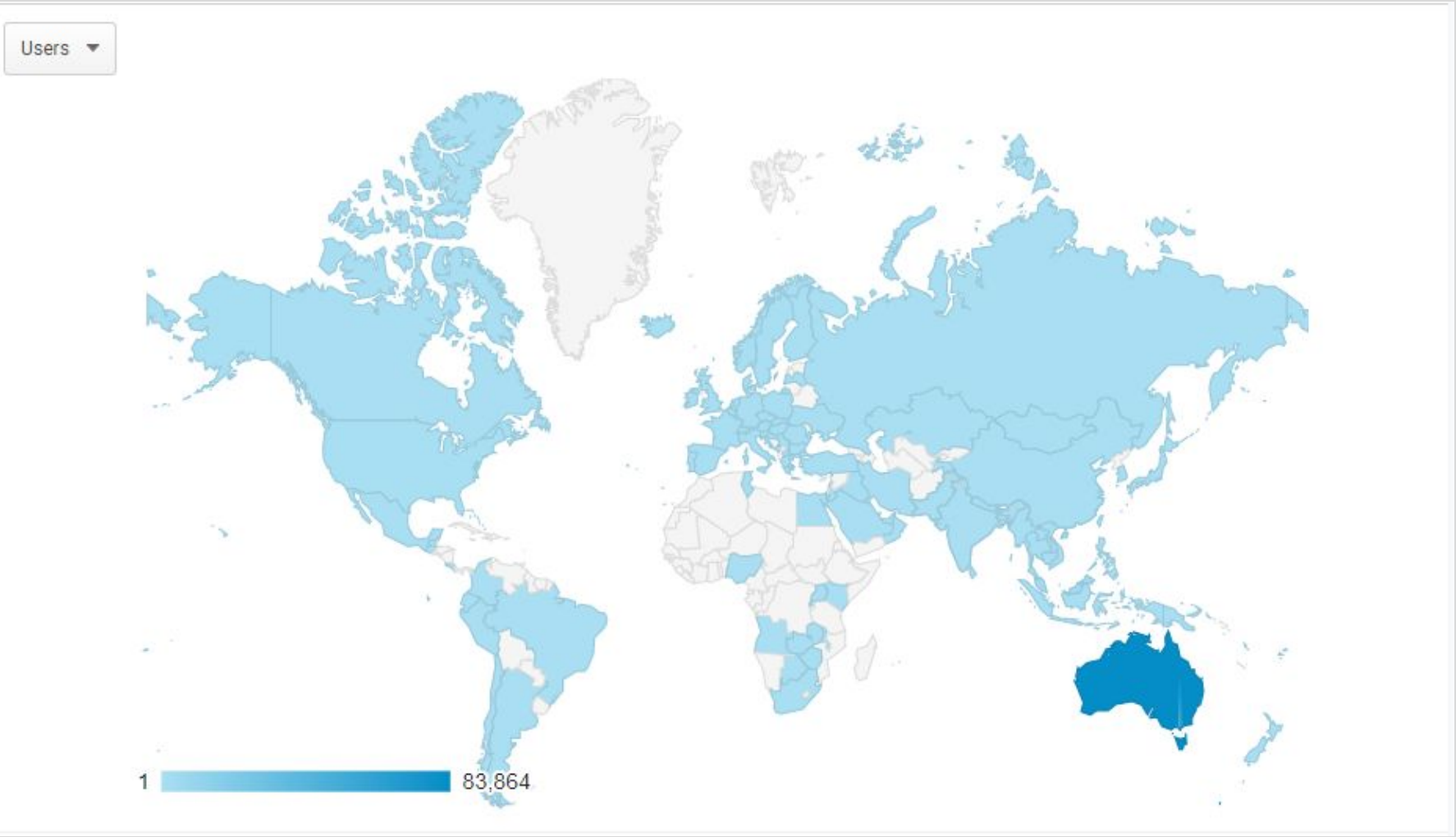
T.L. Robertson Library at Curtin Perth



Curtin University

Learning and Teaching being delivered from

Perth Western Australia
(including Bentley, and Perth City)
Kalgoorlie Western Australia
Malaysia Campus
Mauritius Campus
Singapore Campus
Dubai


Users
1
83,864

Leganto - Reading List Solution

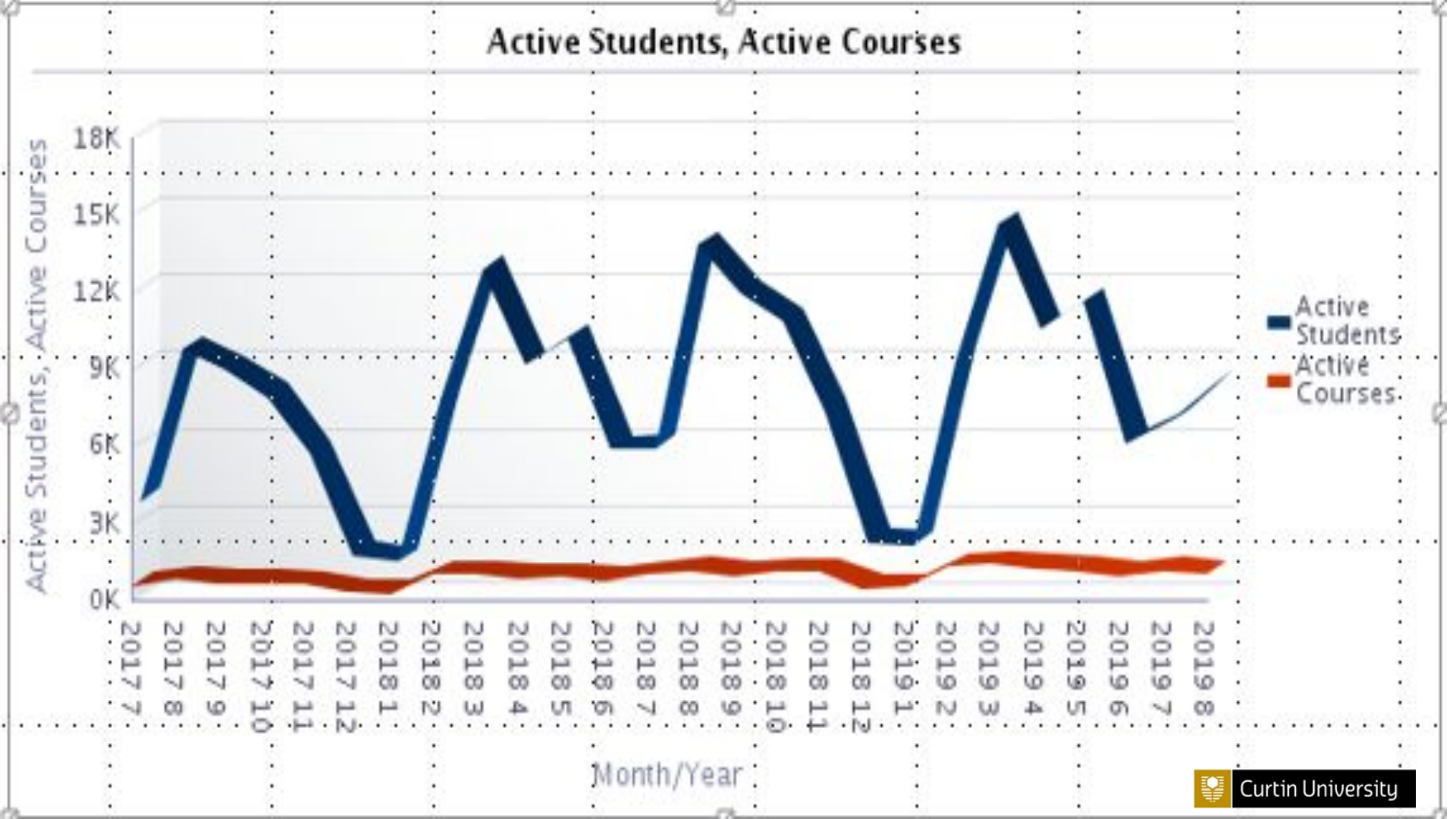
- In August 2015,

  Curtin Library chooses the Ex Libris Leganto as its new Reading List solution

- Curtin Library successfully implemented in the first Semester 2017

- Between 2016 and August 2019, teaching staff have

  - created 4,700+ Reading Lists (includes Reading Lists migrated from previous system)
  - added over 115, 300 citations which resulted in
  - 1.5 million (non-unique) full text views of resources by 42, 400+ unique active students


Active Students, Active Courses
Active Students, Active Courses
18K
15K
12K
9K
6K
3K
0K
2017 7
2017 8
2017 9
2017 10
2017 11
2017 12
2018 1
2018 2
2018 3
2018 4
2018 5
2018 6
2018 7
2018 8
2018 9
2018 10
2018 11
2018 12
2019 1
2019 2
2019 3
2019 4
2019 5
2019 6
2019 7
2019 8
Month/Year
Active Students
Active Courses

A “Proof of Concept” project

- Ex Libris approached Curtin with a proof of concept proposal in 2017
- To use machine learning to investigate the correlation between student success and activity within the Leganto Reading List
- Curtin has been using learning analytics to predict student success and to identify students at risk of failing to complete their studies
- Project would add to early intervention strategies
- Offer from Ex Libris to partner in a proof of concept project was accepted

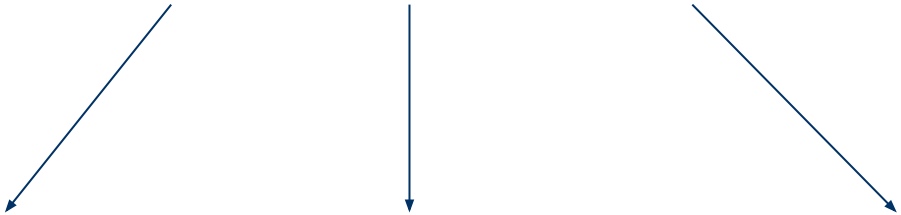
Early indication for students at risk

Advanced analytics

Gal Darom & Tomer Katz

Ex Libris
a ProQuest Company

- $$T = f(Data)$$



Business Need

Transformation

Various type of data (structured/unstructured)

$T = f(Data)$ - What is the Business Need We Seek?

T - Predicting students early in the semester that most likely will struggle with their course

$T = f(Data)$ – Creating the Dataset

$Data$ - Data is available from two sources:

- **Curtin University**
  - Students Profile
  - Students grades and academic Status
- **EX Libris Leganto**
  - Students engagement (usage) with the course resource list

Time and Data

- Significant time and effort to produce the Curtin data
- Negotiation with Student Services and Digital and Technology Solutions
- Leganto data needed to accumulate over multiple semesters
- Success data is reported after the event

$T = f(Data)$ – Creating the Dataset

- Matching Process to combine one dataset from the two data sources
- Developing meaningful features from the unified dataset to improve the model accuracy
- Examples:
  - Student engagement in relation to the average class engagement
  - Weighted student engagement per course

Complexity of Data

- People need to understand the data – conversations over the phone
- Course structure is complicated
- Student demographics complicated
- Language is not standardised – unit vs course, instructor vs coordinator
- Definition of ‘success’ matured over time

$T = f(Data)$ – Choosing the Algorithm

f - During the PoC several algorithm reviewed

1. Decision Tree (DT)
2. General Linear Model (GLM)
3. Naïve Baysian (NB)
4. Support vector machine (SVM)
5. Random Forest (RF)

$T = f(Data)$ – Choosing the Algorithm

We found that for the relevant dataset and business need the Random Forest (RF) algorithm was the best suited classifier

```
rf variable importance

                                Overall
data.GRADE_AVG                  100.000
data.STU_AVERAGE_USAGE_PAST      63.315
Usage_stu_course                 33.274
Age_numeric                      28.325
rel_w1                           24.942
```

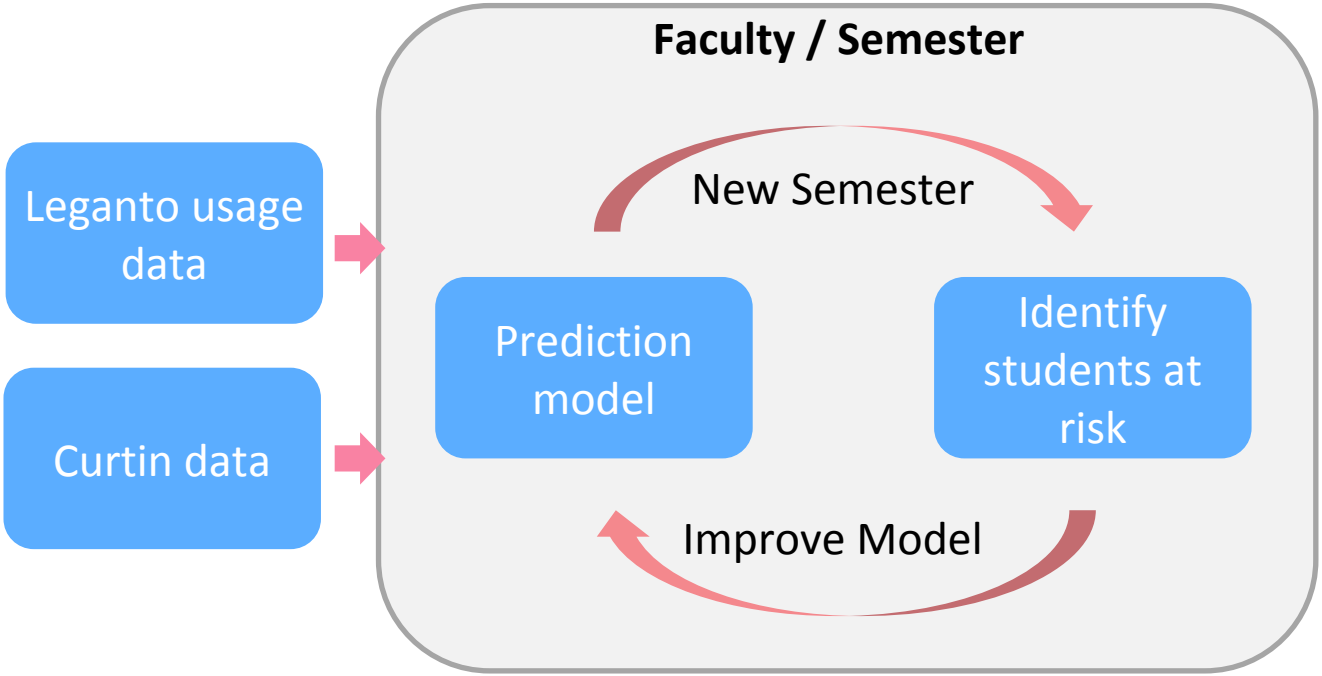
data.Grade_AVG	Student historical average grades
data.STU_AVEGRAGE_USAGE_PAST	Historical usage engineered feature
Usage_stu_course	Weighted student usage per course
Age_numeric	Student age
rel_w1	Student usage in week 1 in relation to class

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1325  108
         1   11   25

               Accuracy : 0.919
                 95% CI : (0.9038, 0.9324)
    No Information Rate : 0.9095
    P-Value [Acc > NIR] : 0.1085

                  Kappa : 0.2676
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9918
            Specificity : 0.1880
         Pos Pred Value : 0.9246
         Neg Pred Value : 0.6944
             Prevalence : 0.9095
         Detection Rate : 0.9020
   Detection Prevalence : 0.9755
      Balanced Accuracy : 0.5899

       'Positive' Class : 0
```

The model total accuracy is 91.9%

Recall: The model will catch 18.8% of students who are at risk (25 / (108+25))

Precision: Prediction of risk student is 69.44% (For 10 students predicted as at risk, 7 will be actually at risk) (25 / (11+25))



The Future

- Proof of Concept is ongoing as more time and data can improve the model
- Too early for Ex Libris to know if this might feature in Leganto in the future

Thank you

Make tomorrow better.

Images

A large blank world map

https://upload.wikimedia.org/wikipedia/commons/c/cf/A_large_blank_world_map_with_oceans_marked_in_blue.PNG (This file is licensed under the Creative Commons Attribution-Share Alike 2.5 Generic, 2.0 Generic and 1.0 Generic license.)