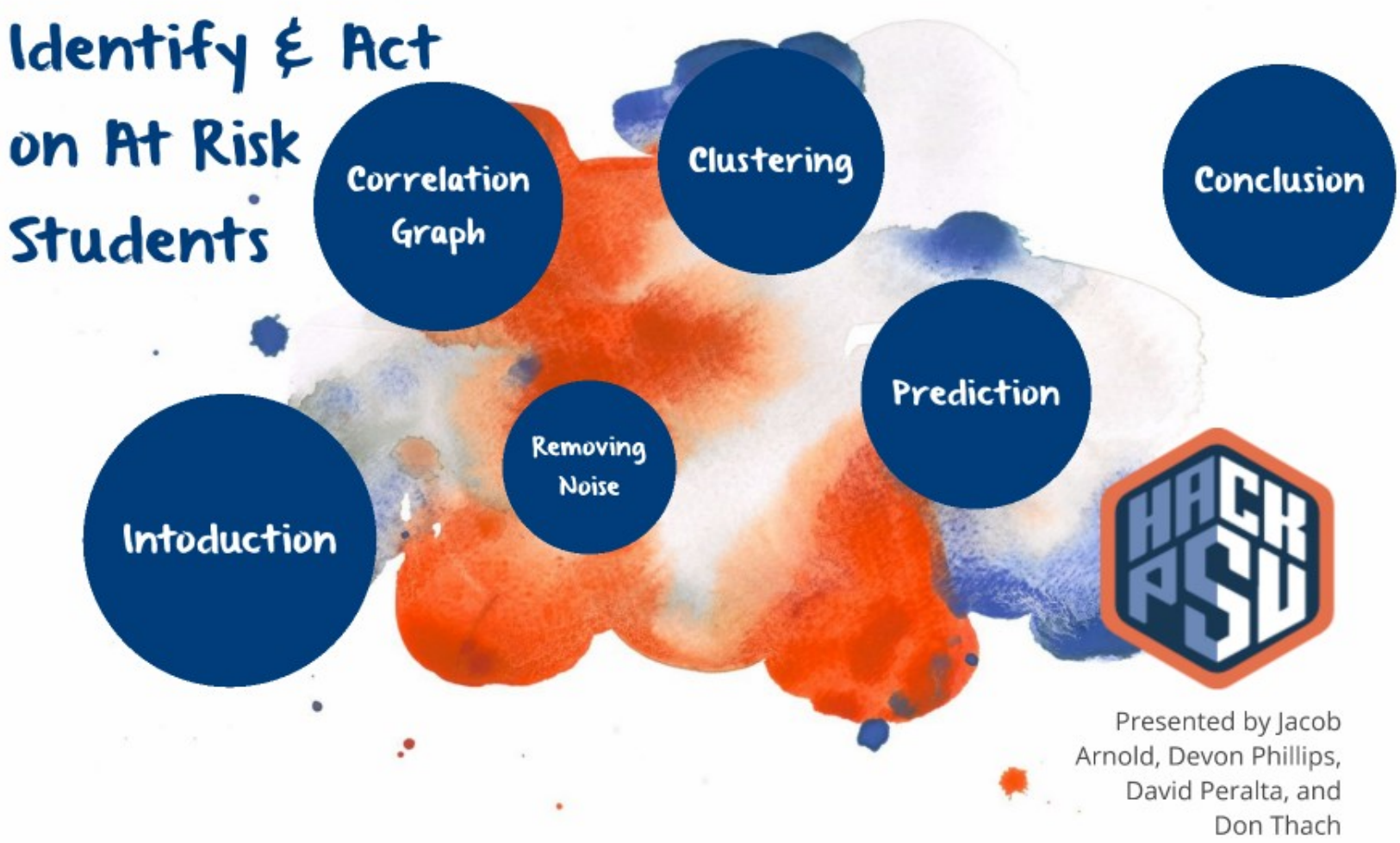
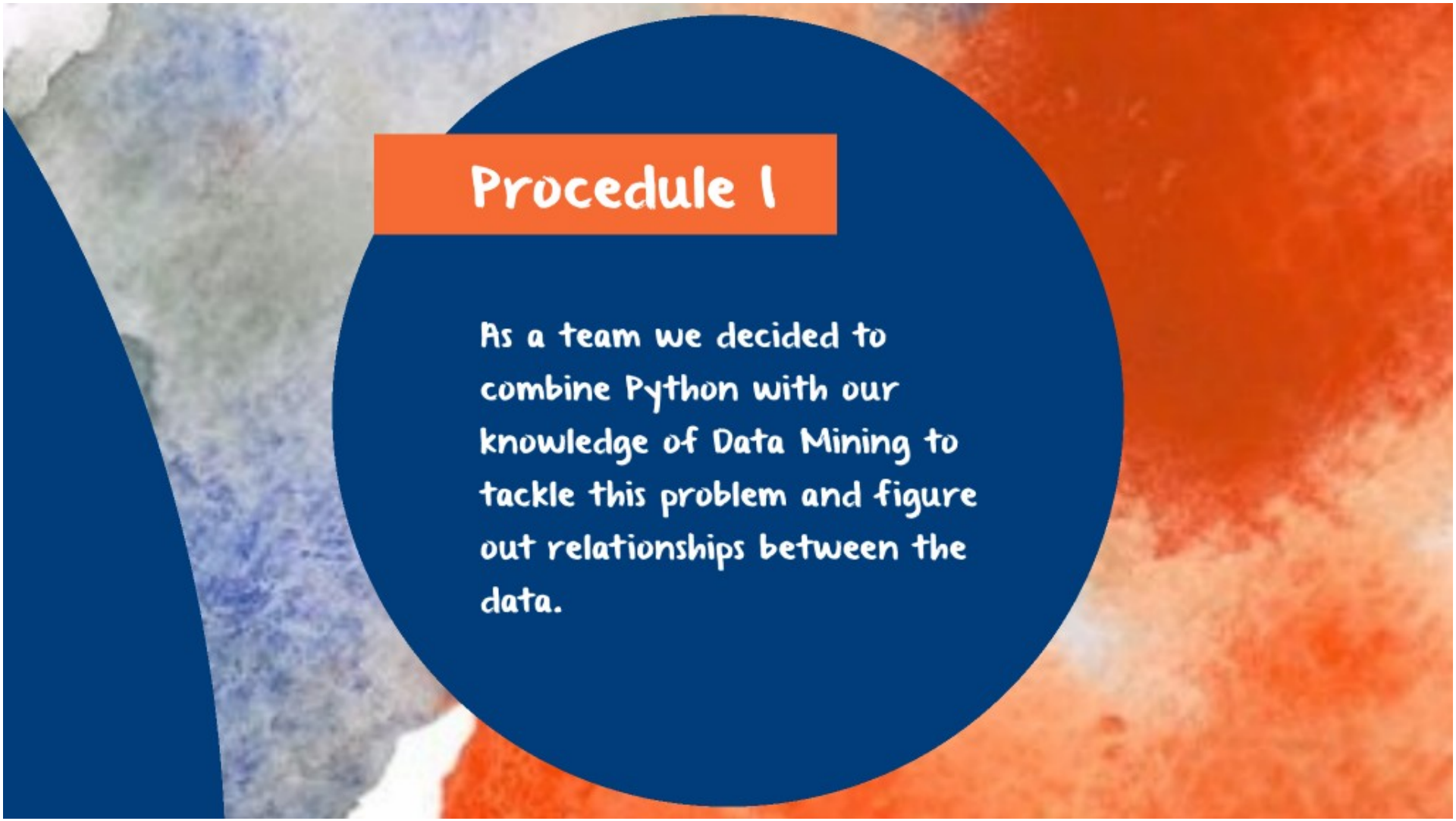

Identify & Act on At Risk Students
Correlation Graph
Clustering
Conclusion
Prediction
Removing Noise
Intoduction
HACK PSU
Presented by Jacob Arnold, Devon Phillips, David Peralta, and Don Thach

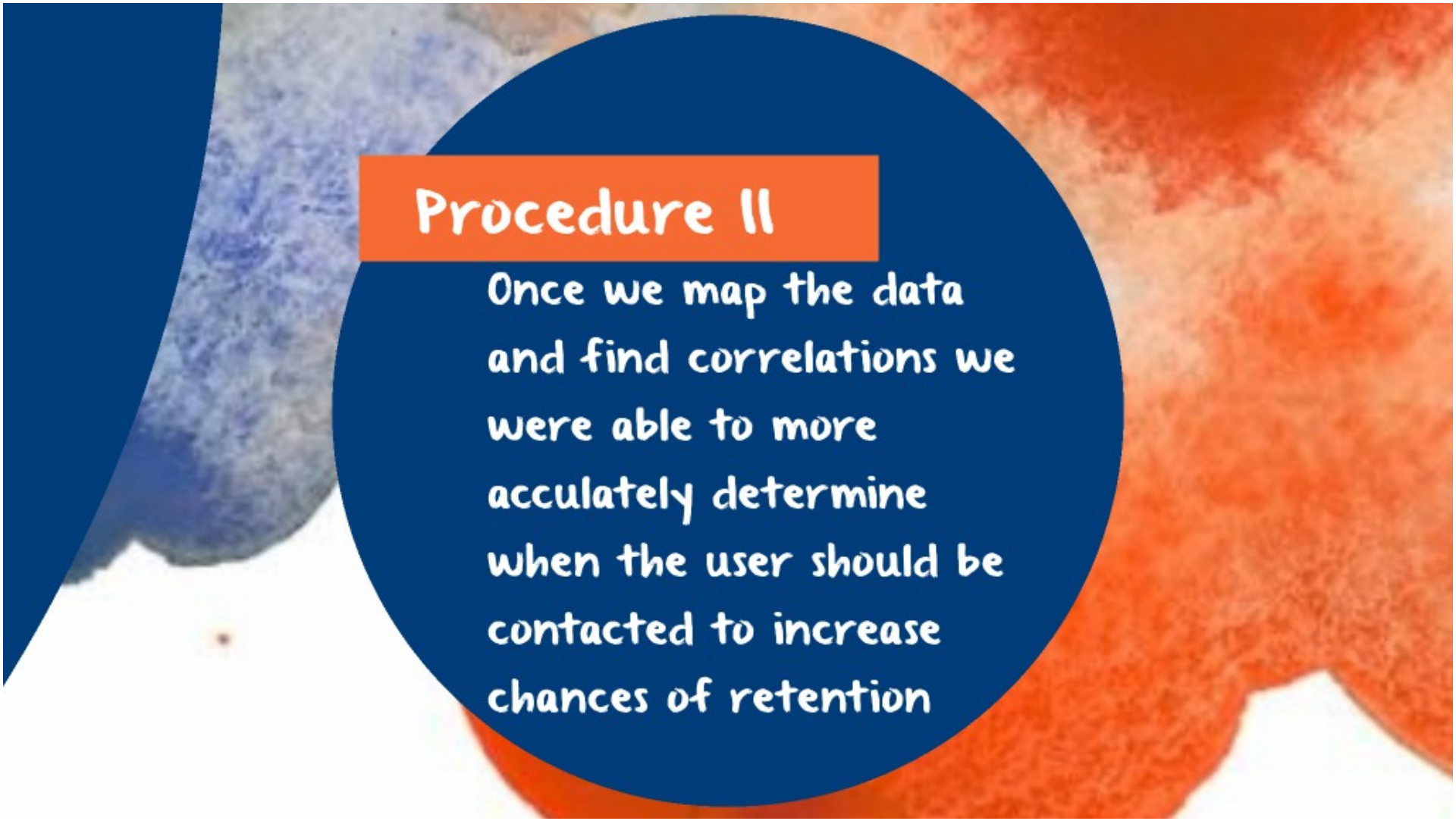
Intoduction

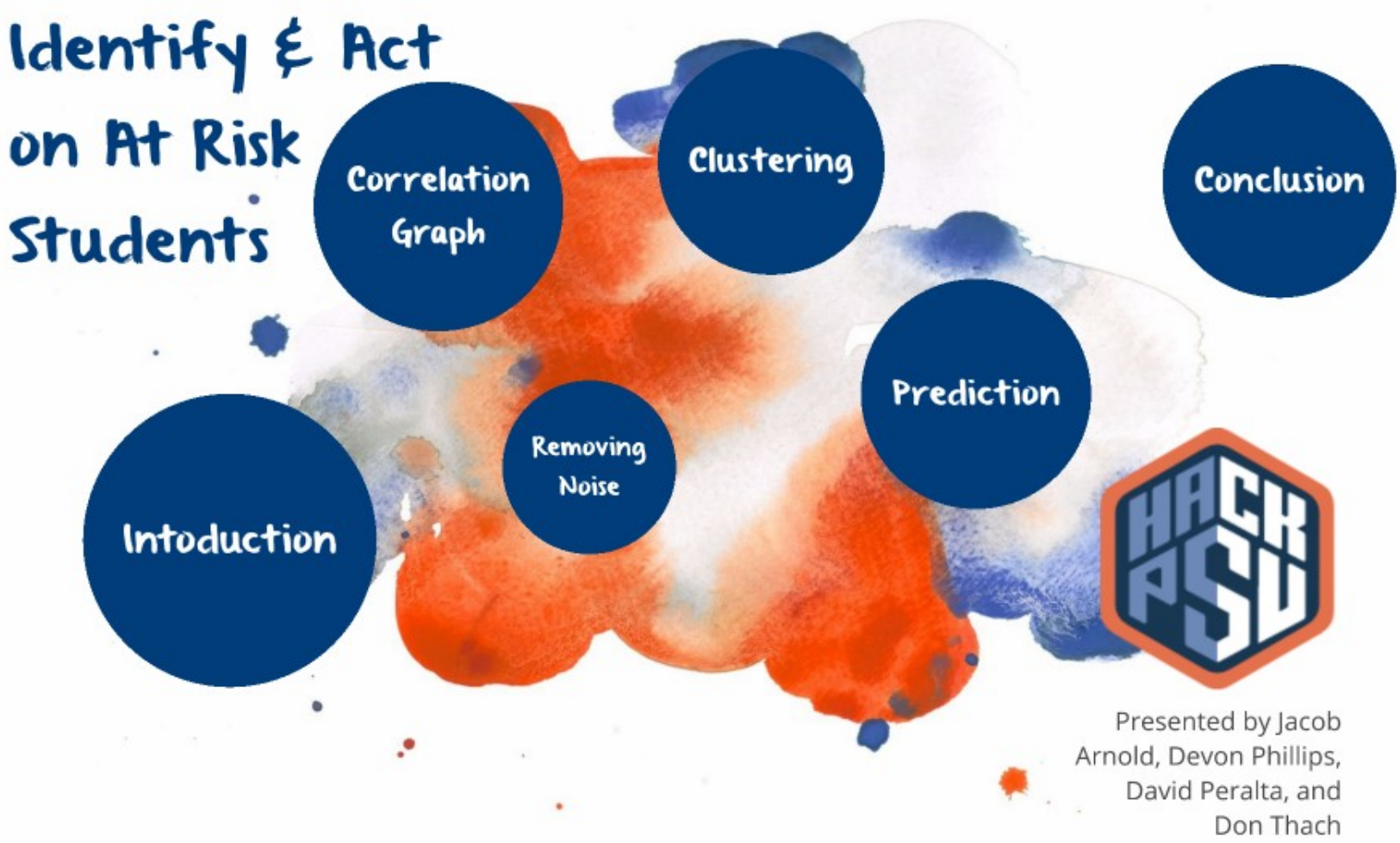
Our goal in this challange was to analyze the data of over 5 million students and figure out which students were academically at risk and deciding the best time to contact the student to avoid the student failing.

Procedule I

Procedure II


Procedule 1
As a team we decided to combine Python with our knowledge of Data Mining to tackle this problem and figure out relationships between the data.

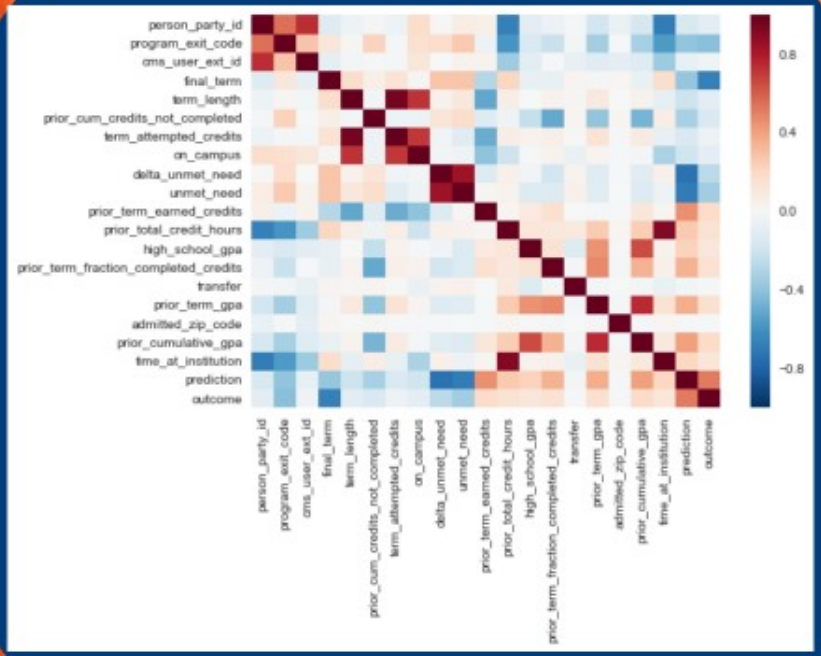

Procedure II
Once we map the data and find correlations we were able to more acculately determine when the user should be contacted to increase chances of retention

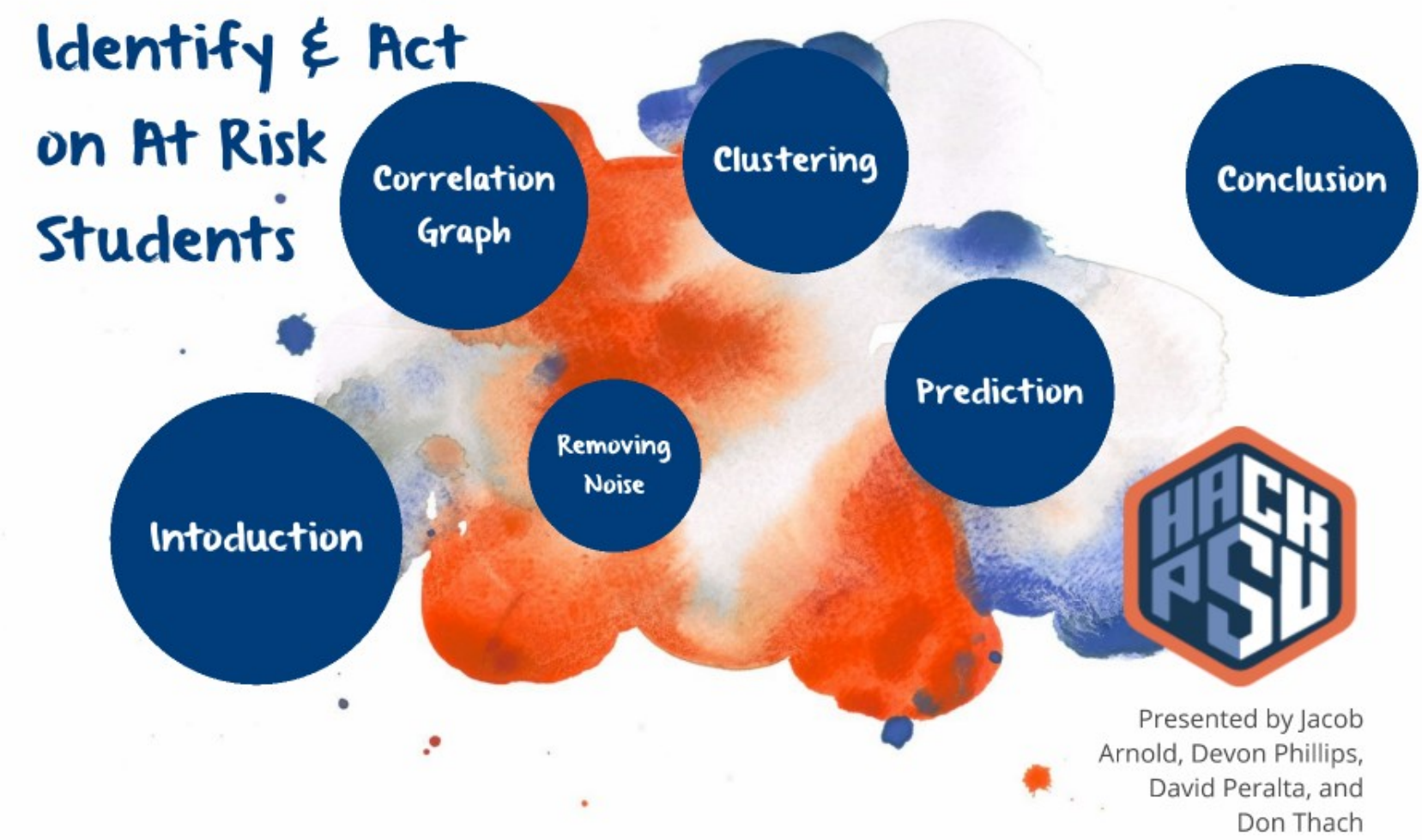

Identify & Act on At Risk Students
Correlation Graph
Clustering
Conclusion
Prediction
Removing Noise
Intoduction
HACK PSU
Presented by Jacob Arnold, Devon Phillips, David Peralta, and Don Thach

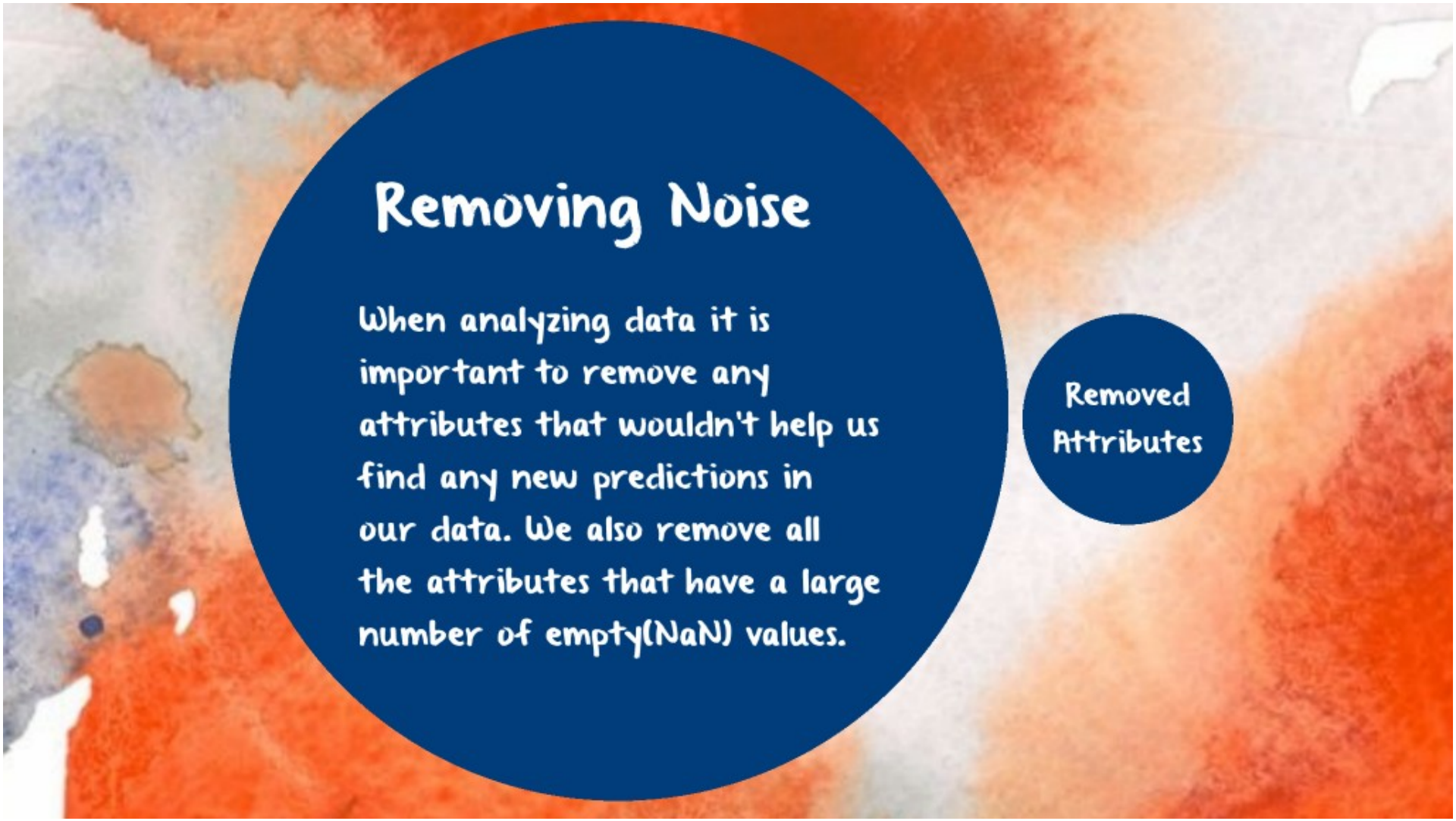
Correlation Graph

Our first step in using Dating Mining to see the hidden data in the database is to create a Correlation Graph which shows us which attributes are positively, neatrul, or negatively correlated.

Corr. Coeff.

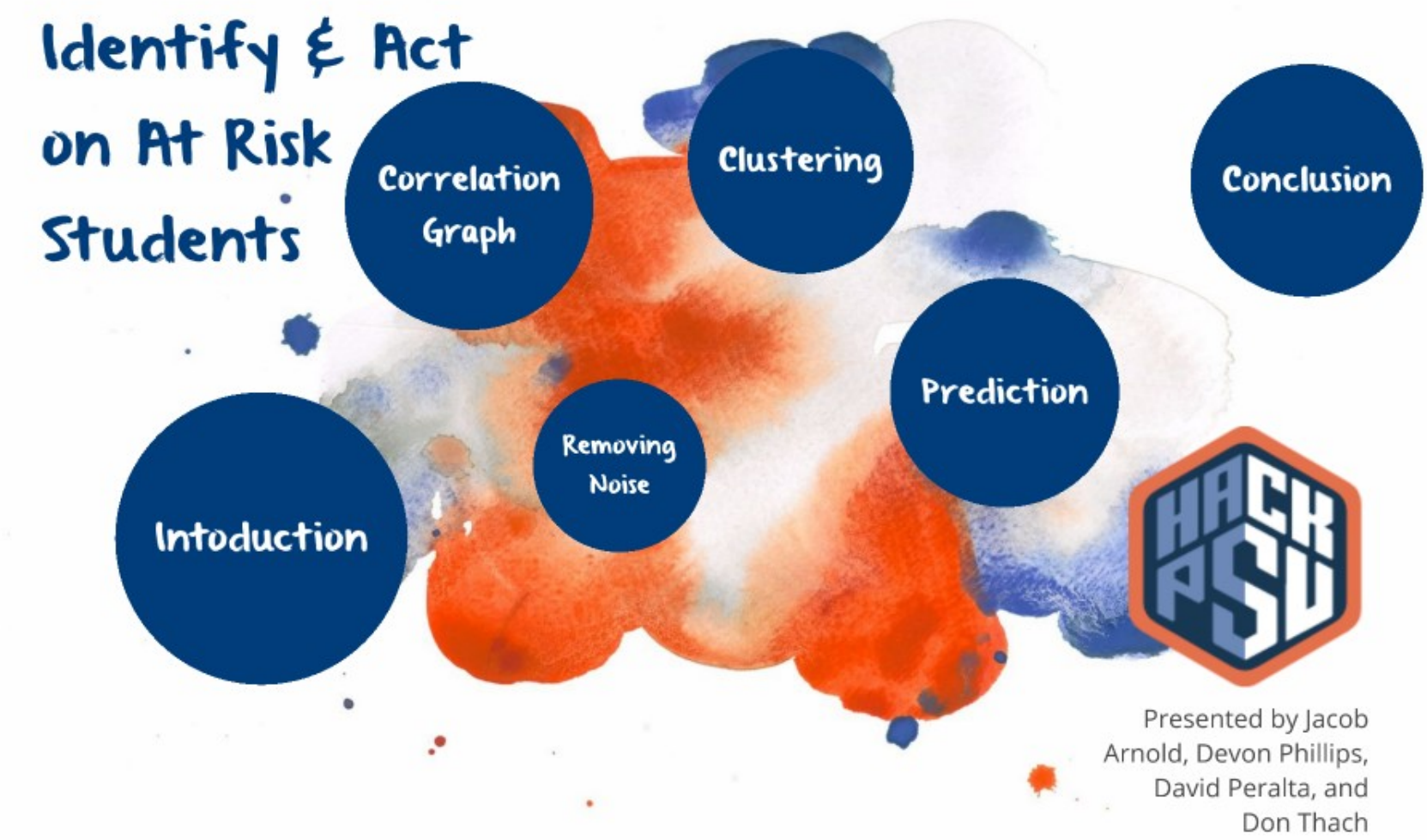

person_party_id
program_exit_code
cms_user_ext_id
final_term
term_length
prior_cum_credits_not_completed
term_attempted_credits
on_campus
delta_unmet_need
unmet_need
prior_term_earned_credits
prior_total_credit_hours
high_school_gpa
prior_term_fraction_completed_credits
transfer
prior_term_gpa
admitted_zip_code
prior_cumulative_gpa
time_at_institution
prediction
outcome
0.8
0.4
0.0
−0.4
−0.8


Identify & Act on At Risk Students
Correlation Graph
Clustering
Conclusion
Prediction
Removing Noise
Intoduction
HACK PSU
Presented by Jacob Arnold, Devon Phillips, David Peralta, and Don Thach


Removing Noise
When analyzing data it is important to remove any attributes that wouldn't help us find any new predictions in our data. We also remove all the attributes that have a large number of empty(NaN) values.
Removed Attributes

Removed Attributes

In this dataframe we removed credential, full_part_time, term_start_month, prior_term_start_month, prior_full_part_time, primary_major, admitted_zip_code, start_dt, end_dt, next_start_dt, program_start_date, program_end_date


Identify & Act on At Risk Students
Correlation Graph
Clustering
Conclusion
Prediction
Removing Noise
Intoduction
HACK PSU
Presented by Jacob Arnold, Devon Phillips, David Peralta, and Don Thach

Clustering

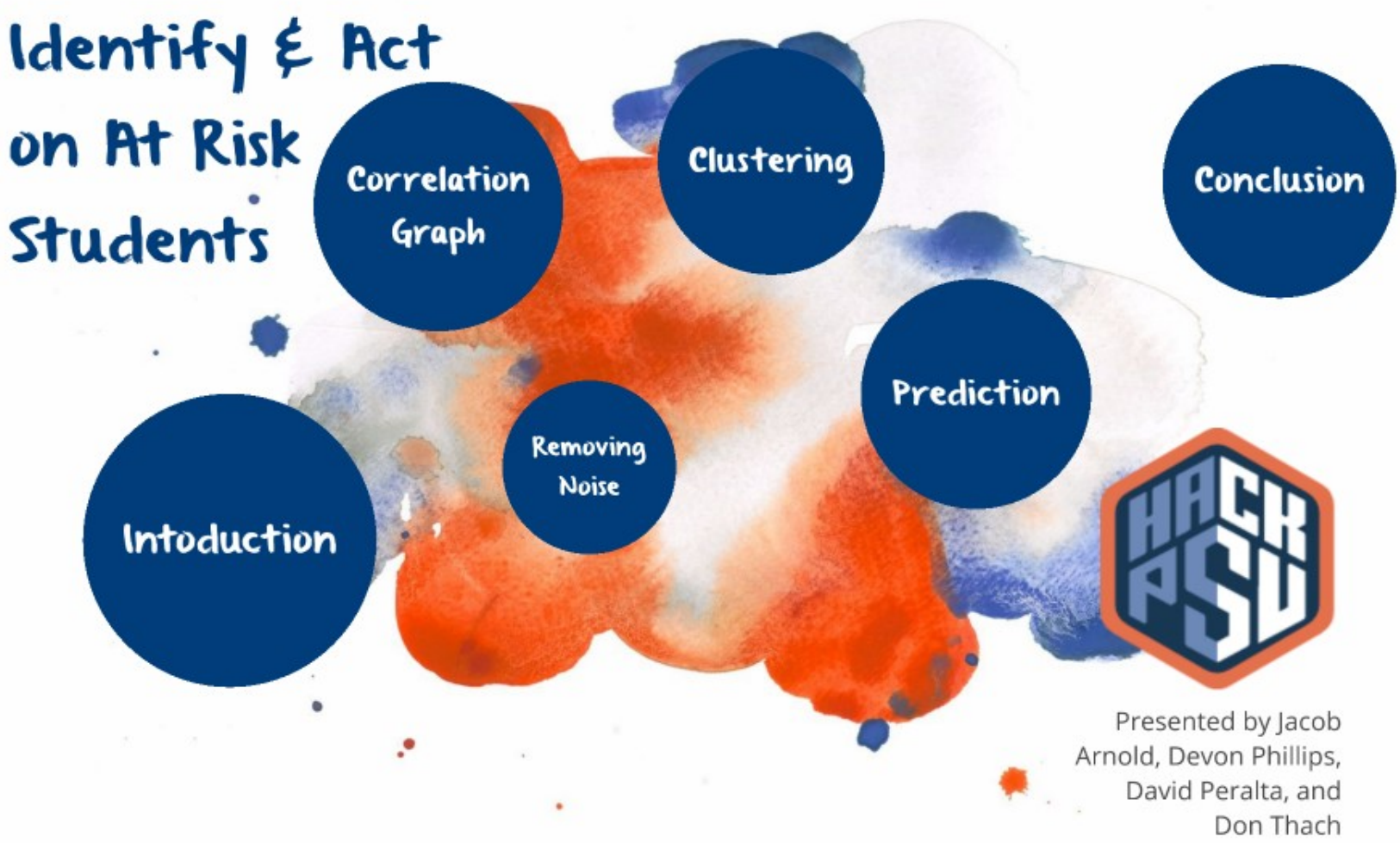
To analyze the data better we used data clustering to cluster data that has similar values. Using the clusters versus the inertia we calculated that we should split the data into 2 clusters. The first cluster has the students likely to retain. The second cluster has all the students that are not likely to retain.

Cluster Data

Cluster Data

We see the 2 clusters 0, and 1, Cluster 0 has 362 sets that need to get contacted as they are likely to not retain, and 4617 likely to retain. In cluster 1 we have 64 sets that are unlikely to retain and 2911 likely to retain. Looking at this we can see cluster 0 is the cluster that has the more users that will need to be contacted.

Index	0.0	1.0
0	362	4617
1	64	2911


Identify & Act on At Risk Students
Correlation Graph
Clustering
Conclusion
Prediction
Removing Noise
Intoduction
HACK PSU
Presented by Jacob Arnold, Devon Phillips, David Peralta, and Don Thach

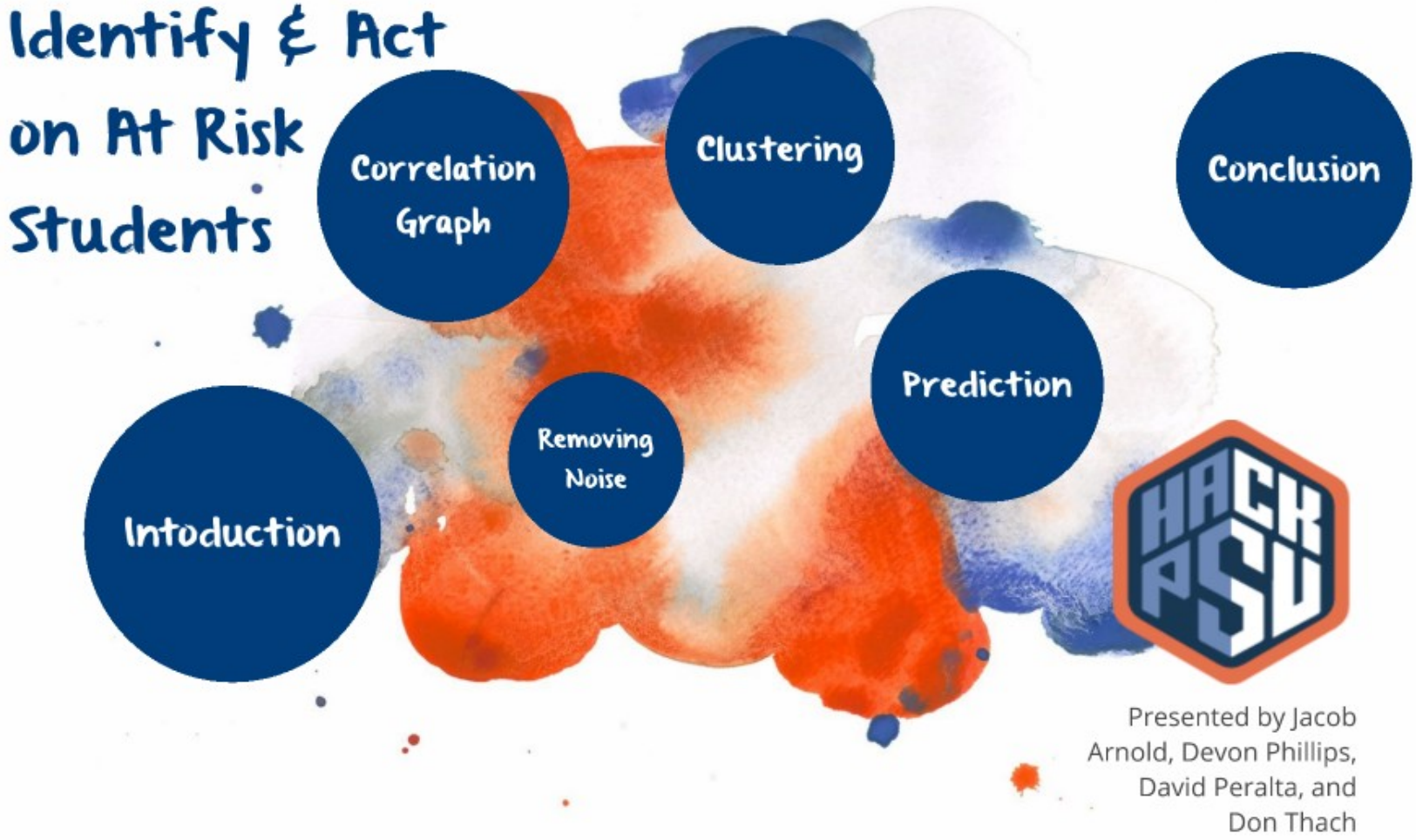
Prediction

We introduce a new student with fictional data to the dataset. To check the new students outcome we rerun the cluster. If the student is in the likely retention cluster then we will not contact the student, if the new student is in the unlikely retention cluster then the system will email the student.

New Student

New Student

We added a real example of a student with 'final_term': 0, 'term_length': 120, 'term_attempted_credits': 15, 'prior_term_earned_credits': 16, 'prior_total_credit_hours':58, 'high_school_gpa':3.8, 'prior_term_fraction_completed_credits':3.7, 'transfer':0, 'prior_cumulative_gpa':3.6, 'time_at_institution': 2, 'prediction':.75, 'gpa_change':0.2 and because of this data the student was placed in the unlikely to pass and the student will be contacted.

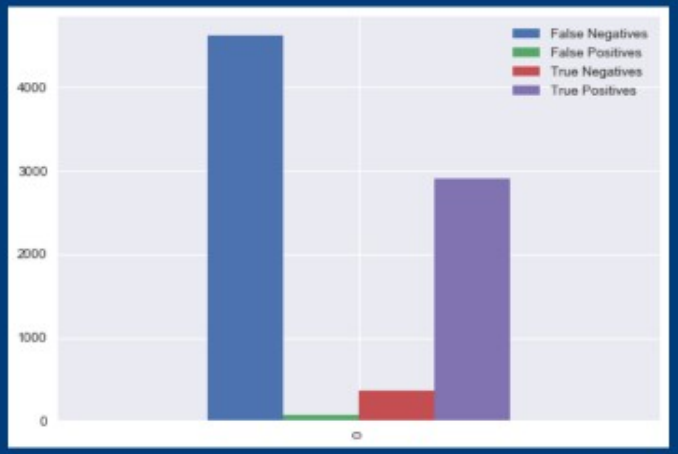

Identify & Act on At Risk Students
Correlation Graph
Clustering
Conclusion
Prediction
Removing Noise
Intoduction
HACK PSU
Presented by Jacob Arnold, Devon Phillips, David Peralta, and Don Thach

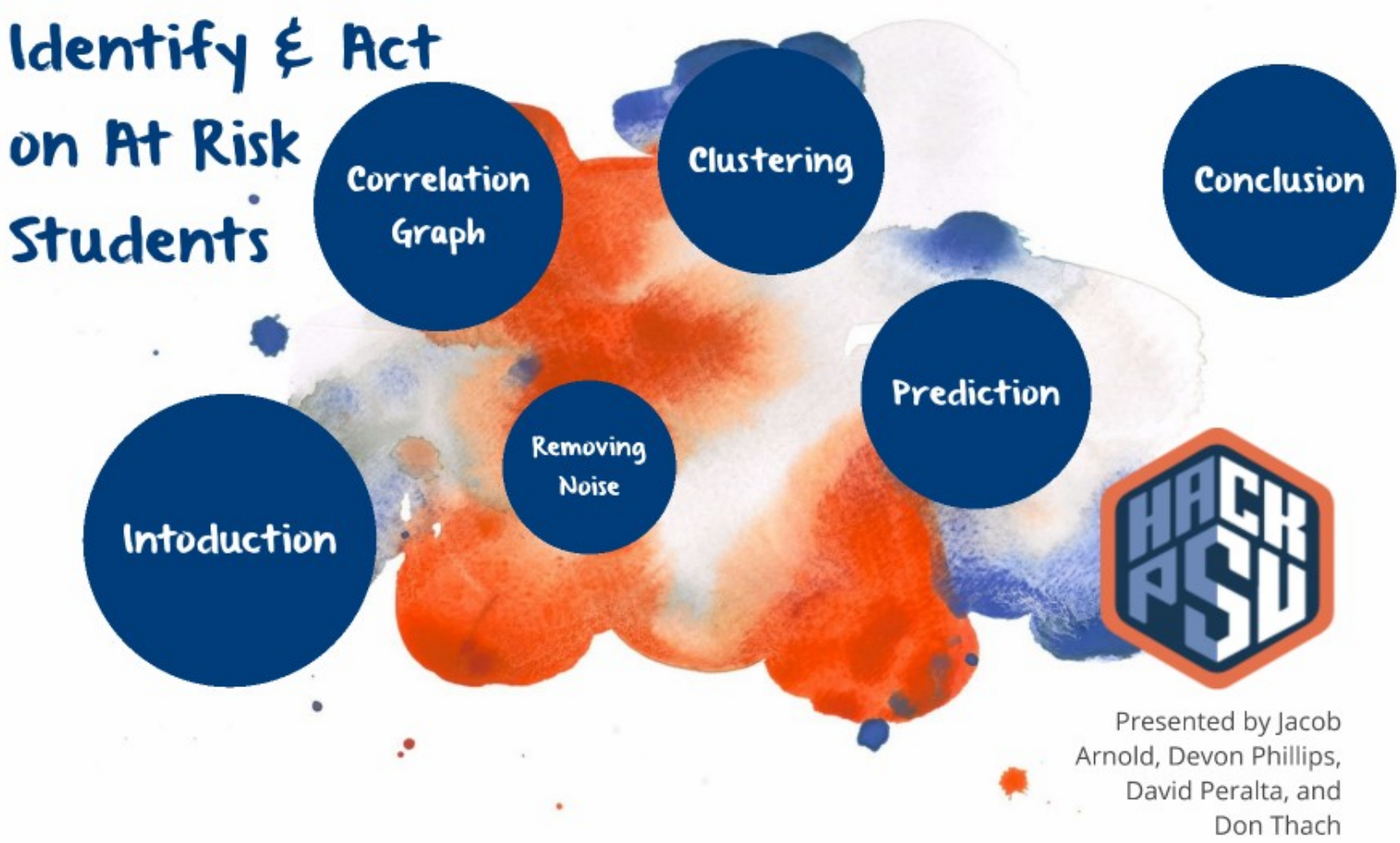
Conclusion

During this project we used our knowledge of data mining to find new data relationships among the millions of students. We tested our data by using new test students.

Accuracy

Accuracy


False Negatives
False Positives
True Negatives
True Positives
4000
3000
2000
1000
0


Identify & Act on At Risk Students
Correlation Graph
Clustering
Conclusion
Prediction
Removing Noise
Intoduction
HACK PSU
Presented by Jacob Arnold, Devon Phillips, David Peralta, and Don Thach