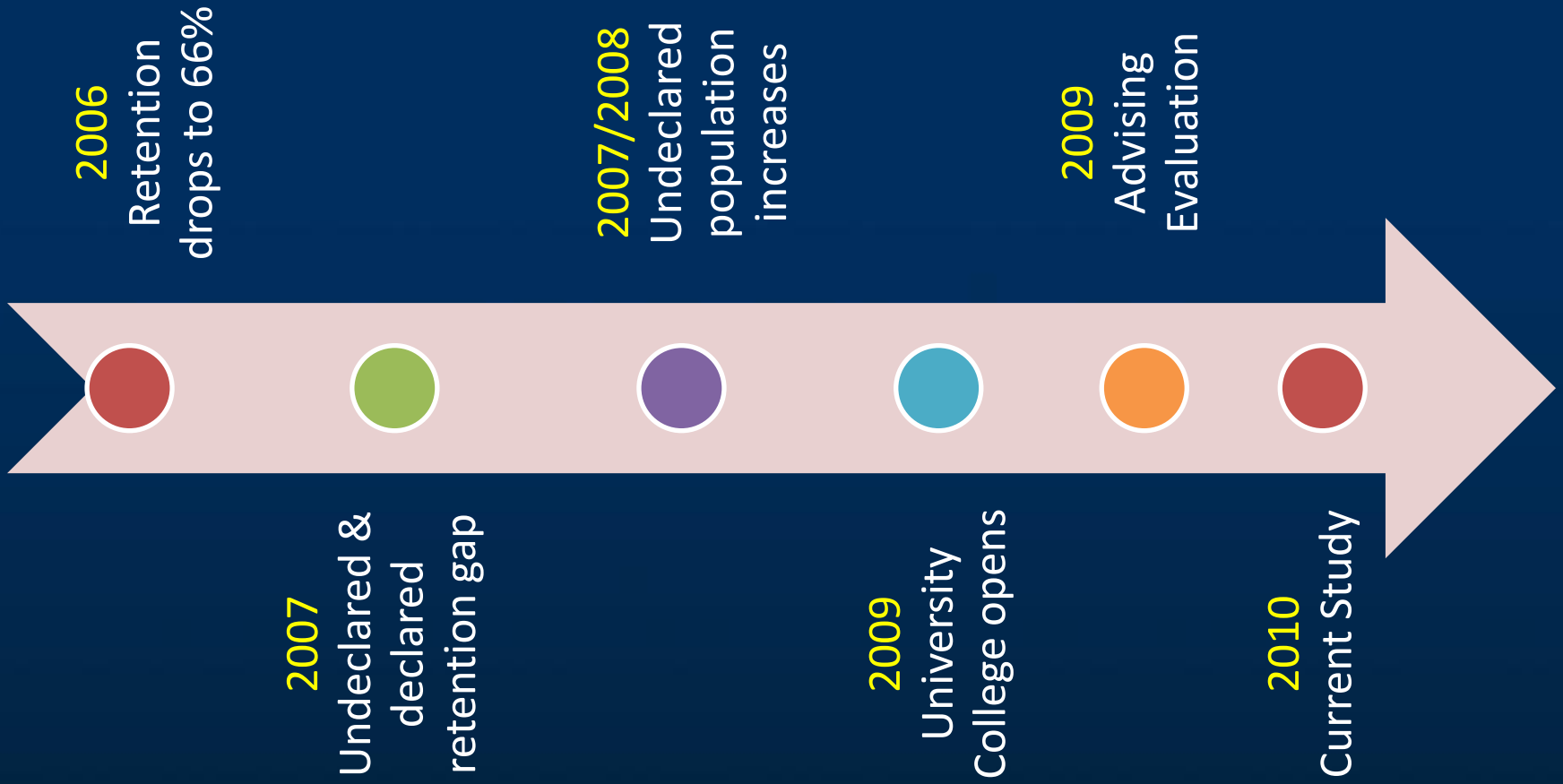
Using Data Mining to Predict Persistence among Undecided First-Year Students: Combining Institutional, CIRP Survey, and National Clearinghouse Data

Kim Black, Ph.D.
Karen Raymond, Ph.D.
Stephanie Torrez, M.A.

UNIVERSITY of NORTHERN COLORADO

Bringing education to life.

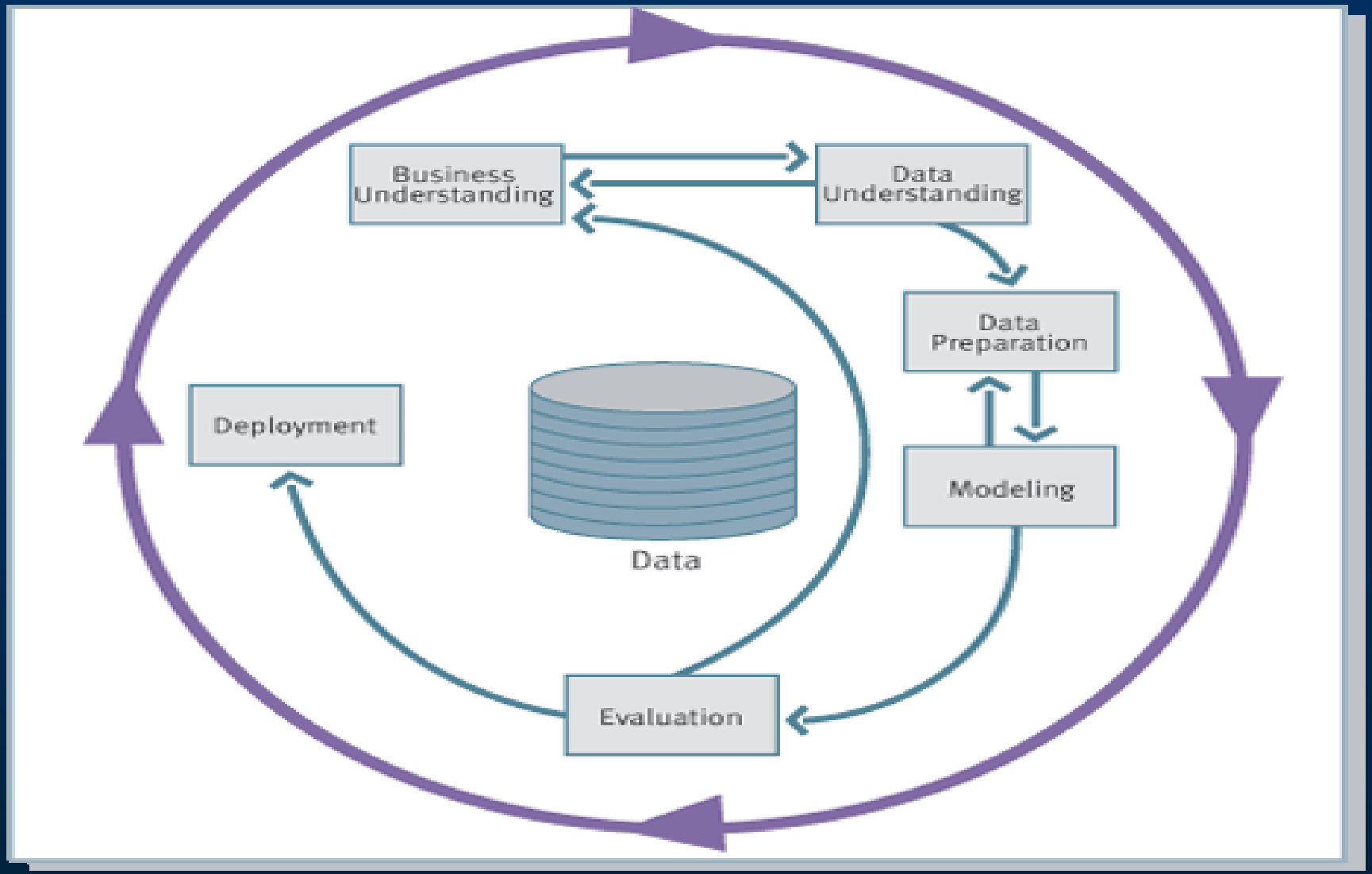
The Story of Our Study



Institutional Data Mining Capacity

- Implementation of Banner SIS Fall 2006
- Operational Data Store and Enterprise Data Warehouse
- External data sources:
  - CIRP (Since 2007)
  - Clearinghouse
  - MapWorks (Since 2007)
  - Admitted Student Survey
- PASW Modeler 13 (Spring 2009)
- Enrollment probability modeling

CRISP-DM Process



Source: http://www.crisp-dm.org/

Data Mining Project Goals

- To differentiate types of undeclared students
- To assess the utility of these types for customizing advising needs and prioritizing resources
- To evaluate the relationship between the types of undeclared students who left and where they transferred to
- To identify attributes of undeclared students that predict first year persistence

Data Understanding

ALL INCOMING
First Time Full Time

Attribute	2007	2008
Residents	90%	86%
First Generation	27%	31%
Minority	17%	17%
Male	40%	40%
FAFSA filer	74%	77%
Avg ACT	22 (3)	22 (3)
Avg Index	104 (12)	104 (12)
EFC_{MED}	12,065	12,624
CIRP NOT MATCHED	44%	41%

ASA ASSIGNED

Attribute	2007	2008
Residents	91%	87%
First Generation	27%	32%
Minority	17%	15%
Male	55%	54%
FAFSA filer	69%	73%
Avg ACT	21 (3)	21 (3)
Avg Index	99 (10)	99 (10)
EFC_{MED}	10,986	10,923
CIRP NOT MATCHED	49%	43%

Outcome Comparisons

ALL INCOMING

Term Outcome	2007	2008
Term GPA 1st Fall	2.65 (1.02)	2.73 (1.08)
Term GPA 1st Spr	2.50 (1.13)	2.22 (1.31)
Retained Spring	88.8%	86.6%
Retained 2nd Fall	70.8%	68.6%

ASA ASSIGNED

Term Outcome	2007	2008
Term GPA 1st Fall	2.43 (1.06)	2.28 (1.09)
Term GPA 1st Spr	2.26 (1.15)	1.98 (1.30)
Retained Spring	87.9%	84.7%
Retained 2nd Fall	66.0%	63.4%

Data Preparation

Pre-College

ADMISSIONS APPLICATION

* Race/Ethnicity
* Gender
* Parent education
* Geographic information
* Test scores, HSGPA, etc.
* Age
* SSN
* Etc.

CIRP-Freshmen Survey

*College Choice
* Past and Future Behaviors
* Habits of Mind
* Academic and Social Self Concept
• Social Agency
• Financial Concerns
• Self Ratings, etc.

Enrollment and Financial Aid Data

ENROLLMENT

* Major
* Program
* Time-status
* Total credits

TERM OUTCOMES

* Term GPA
* CUM_GPA
* Credits attempted, earned
* Withdrawal information
* Degree awarded

FINANCIAL AID

* FAFSA Filer Y1 and Y2
* EFC Y1 and Y2
* Merit and Need-Based Award Levels

Fall to Fall Outcomes

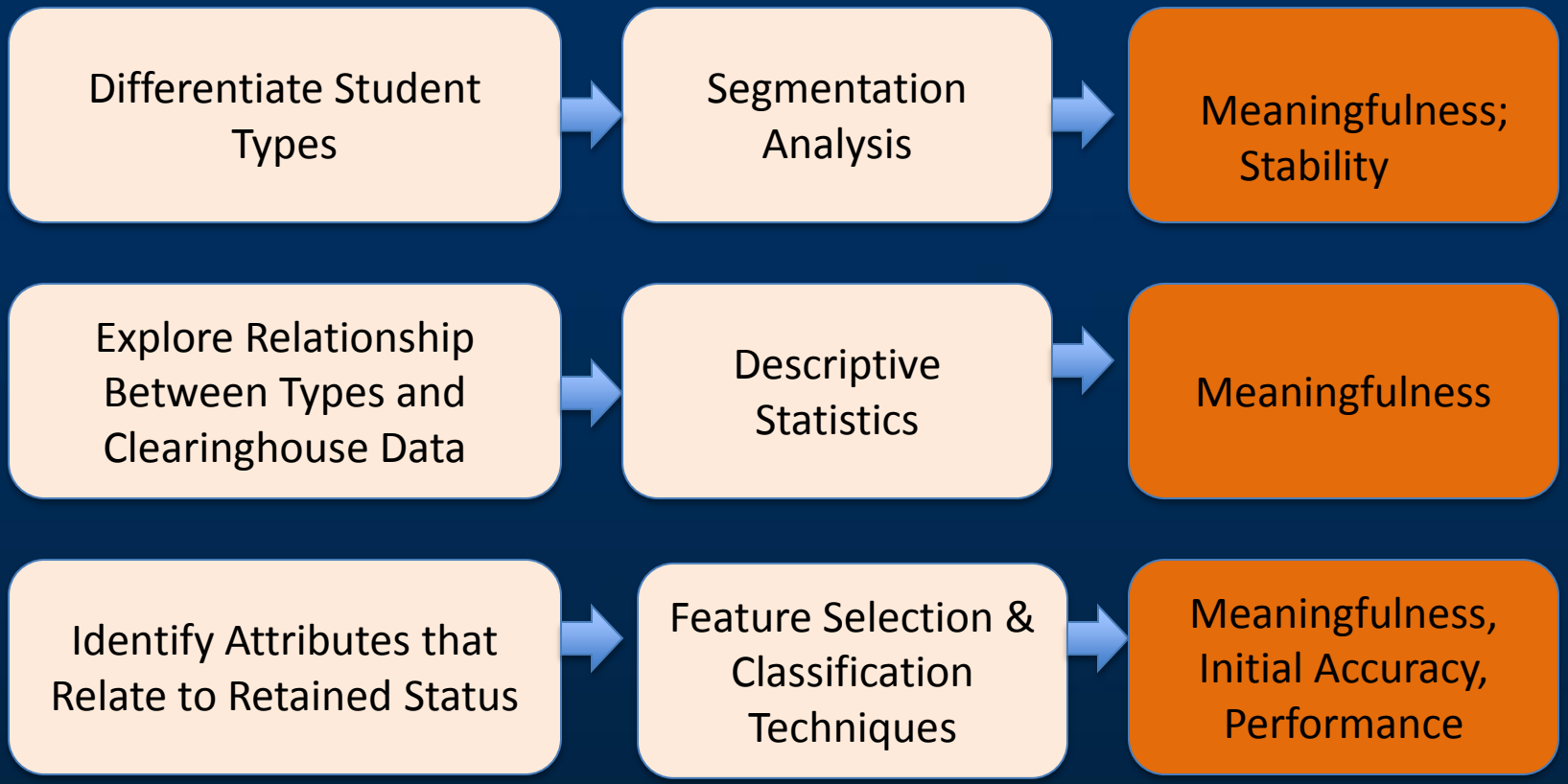
ENROLLMENT

* Retained into First Spring
• Retained into Second Fall

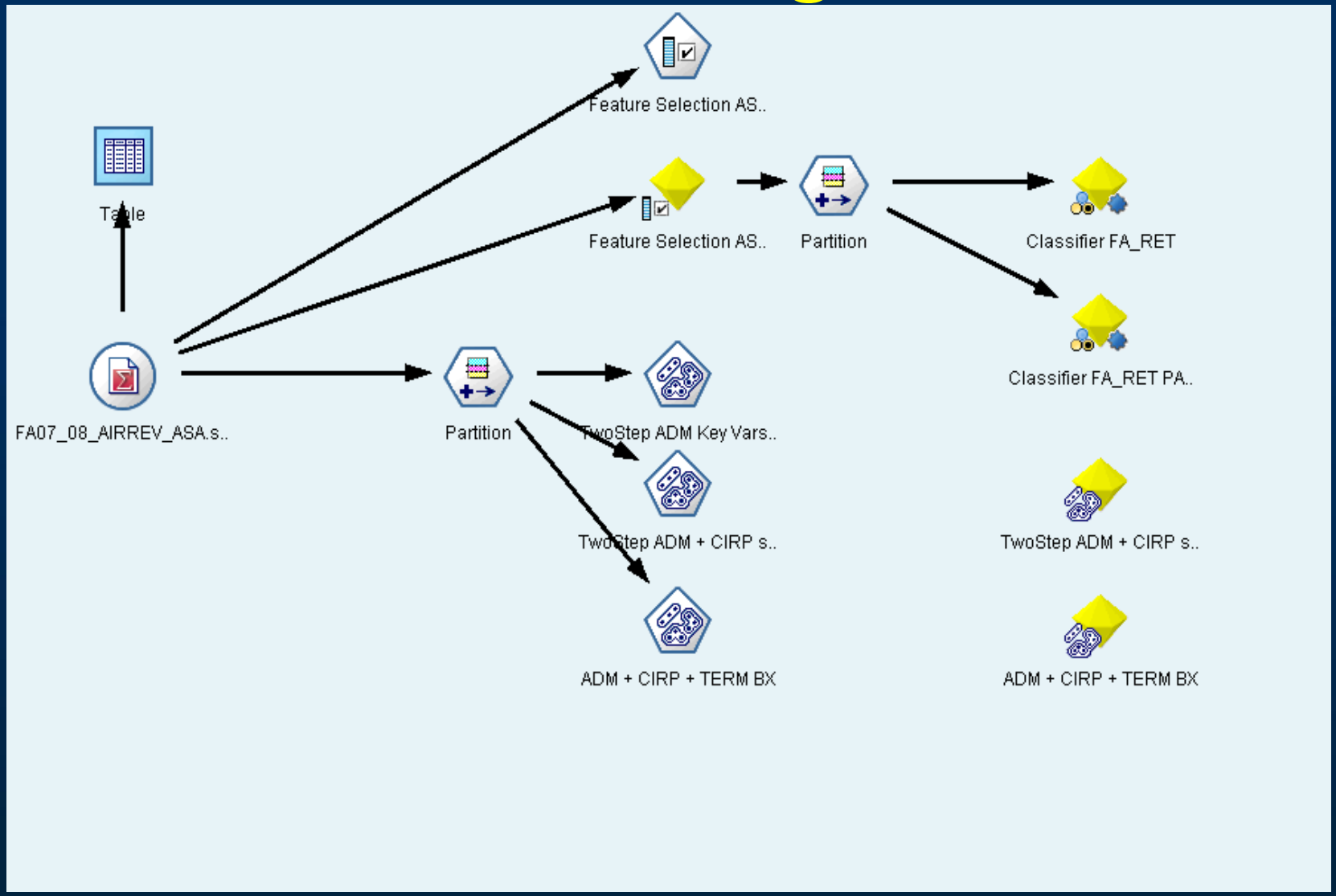
CLEARINGHOUSE:

* Re-enrolled at 2 Year
* Re-enrolled at 4 Year
* Not Found (No evidence of enrollment anywhere)
* Graduated

Modeling and Evaluation



PASW Modeling Stream



FEATURE-SELECTION

	Rank	Field	Type	Importance	Va
☑	1	SP1F_TERMGPA	Range	Important	1.0
☑	2	FA1F_TERMGPA	Range	Important	1.0
☑	3	FAFSA_FILERY2	Set	Important	1.0
☑	4	FA1F_TOTCREDITS	Range	Important	1.0
☑	5	SCHOOL_GPA	Range	Important	1.0
☑	6	IND_SCR	Range	Important	1.0
☑	7	PERCENTILE	Range	Important	1.0
☑	8	FUTACT11	Range	Important	0.998
☑	9	ACT05	Range	Important	0.998
☑	10	FUTACT04	Range	Important	0.998
☑	11	HPW07	Range	Important	0.997
☑	12	FUTACT05	Range	Important	0.994
☑	13	FEMALE	Set	Important	0.991
☑	14	ACT07	Range	Important	0.991
☑	15	ACT03	Range	Important	0.991
☑	16	RATE07	Range	Important	0.986
☑	17	GOAL08	Range	Important	0.985
☑	18	ACT16	Range	Important	0.983
☑	19	AID3	Range	Important	0.98
☑	20	CLUSTER_FINAL2	Set	Important	0.979
☑	21	FINCON	Range	Important	0.974
☑	22	CHOOSE11	Range	Important	0.967
☑	23	ACT02	Range	Important	0.966

Selected fields:26 Total fields available:205

AUTOMATIC CLASSIFIER

- Specify number of models to be created and compared
- Use partitioned data; models ranked by accuracy
- Final model: C.5 Classification tree, 86% accuracy
- Attributes important in classifying students:
  - Spring Census Time Status (WD, LHT)
  - Spring Term Final GPA
  - Fall Term Final GPA
  - HS GPA
  - Diversity of neighborhood where they grew up
  - FAFSA Filer Y2

Cluster Results: Preferred Model

Attribute	HIGH ACHIEVING UNDECLARED	UNDECLARED FEMALE	UNDECLARED MALE	UNDECLARED TRIO ELIGIBLE
Index Score	117 (11)	100 (11)	96 (6)	96 (7)
ACT	27 (2)	21 (2)	20 (2)	19 (2)
% Residents	93%	95%	97%	100%
% Minority	12%	0%	0%	100%
% Female	44%	100%	0%	46%
% First Gen	19%	34%	34%	61%
EFC_{MED}	$13,060	$10,818	$11,948	$5,431
FALL_F GPA	2.69 (1.05)	2.66 (.90)	2.18 (1.00)	2.20 (1.02)
SPR RET	93%	88%	87%	89%
SPR_F GPA	2.59 (1.21)	2.49 (1.10)	1.81 (1.21)	1.85 (1.17)
FALL RET	67%	73%	61%	67%
FALL LOSS	Not Found/4YR	4 YR*	NOT FOUND	2 YR*

What's Next?

Summary of Findings

- Identification of preliminary types of undeclared students
- Information about attributes related to fall to fall persistence
- Understanding of where different types of students went after they left
- Delivery of a comprehensive and integrated dataset from of Institutional, CIRP, and Clearinghouse data for future studies

Lessons Learned/Future Actions

- Data mining does require commitment from analysts and practitioners alike
- Data mining can yield to new understanding of a stable problem in higher education (retention)
- Data mining is not a “one shot deal” and may not be appropriate for all types of problems
- Data mining may require a mind shift in how institutional research and analysis is done

Contact Information

- Kim Black, Director of Assessment
  - kim.black@unco.edu
- Karen Raymond, Senior Research Analyst
  - karen.raymond@unco.edu
- Stephanie Torrez, Executive Director of Academic Support and Advising
  - Stephanie.torrez@unco.edu

Procedure /Logic

- Build models based on the following test design:
  - Explore important variables in the CIRP for all incoming freshmen
  - Focus on the target population only
  - Begin models with the *"cheap and easy"*
  - Gradually integrate relevant CIRP variables into models
  - Decide which one makes most sense

Building Clusters

- Model 1: ALL Incoming CIRP Only
- Model 2: ASA Only ADM APP ONLY (Index only)
- Model 3: ASA ADM APP (Add in ACT scores)
- Model 4: ASA ADM APP, CIRP scales, Ed Asps, Financial Concerns
- Model 5: ASA ADM APP, CIRP scales, Probable Major/career, Ed Asps, Future Acts
- Model 6: ADM APP and Almost Everything Else

Cluster CIRP Scale Means

Attribute	HIGH ACHIEVING UNDECLARED	UNDECLARED FEMALE	UNDECLARED MALES	UNDECLARED TRIO ELIGIBLE
Habits of Mind	2.43 (.32)	2.34 (.30)	2.23 (.31)	2.31 (.31)
Academic Self	3.90 (.47)	3.32 (.58)	3.47 (.53)	3.54 (.5)
Social Self	2.65 (.53)	2.57 (.54)	2.66 (.56)	2.78 (.54)
Social Agency	2.21 (.66)	2.30 (.55)	2.20 (.64)	2.42 (.67)
College Rep	1.95 (.56)	2.26 (.55)	2.12 (.52)	2.32 (.50)
College Involve	3.48 (.67)	3.62 (.68)	3.19 (.69)	3.48 (.67)