GEORGIA COLLEGE

GEORGIA'S PUBLIC LIBERAL ARTS UNIVERSITY

Successfully Harnessing Data Science In IR

Eric Braun, resigned January 31, 2017

Georgia College and State University

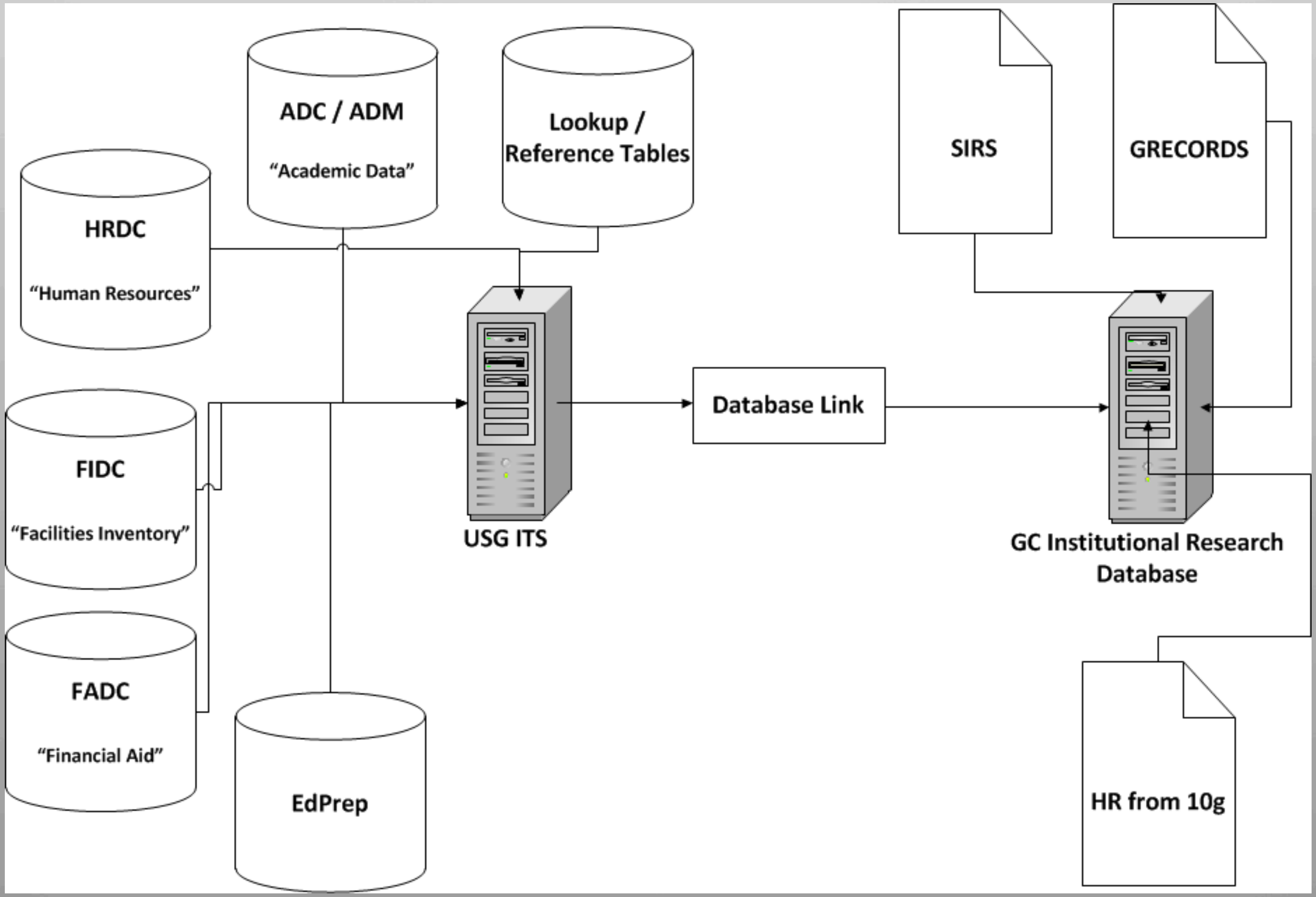
Presentation Objective: A brief primer

- What data infrastructure is required for data science?
- What is likely to be a successful data science implementation?
- A case study illustrating data science in action

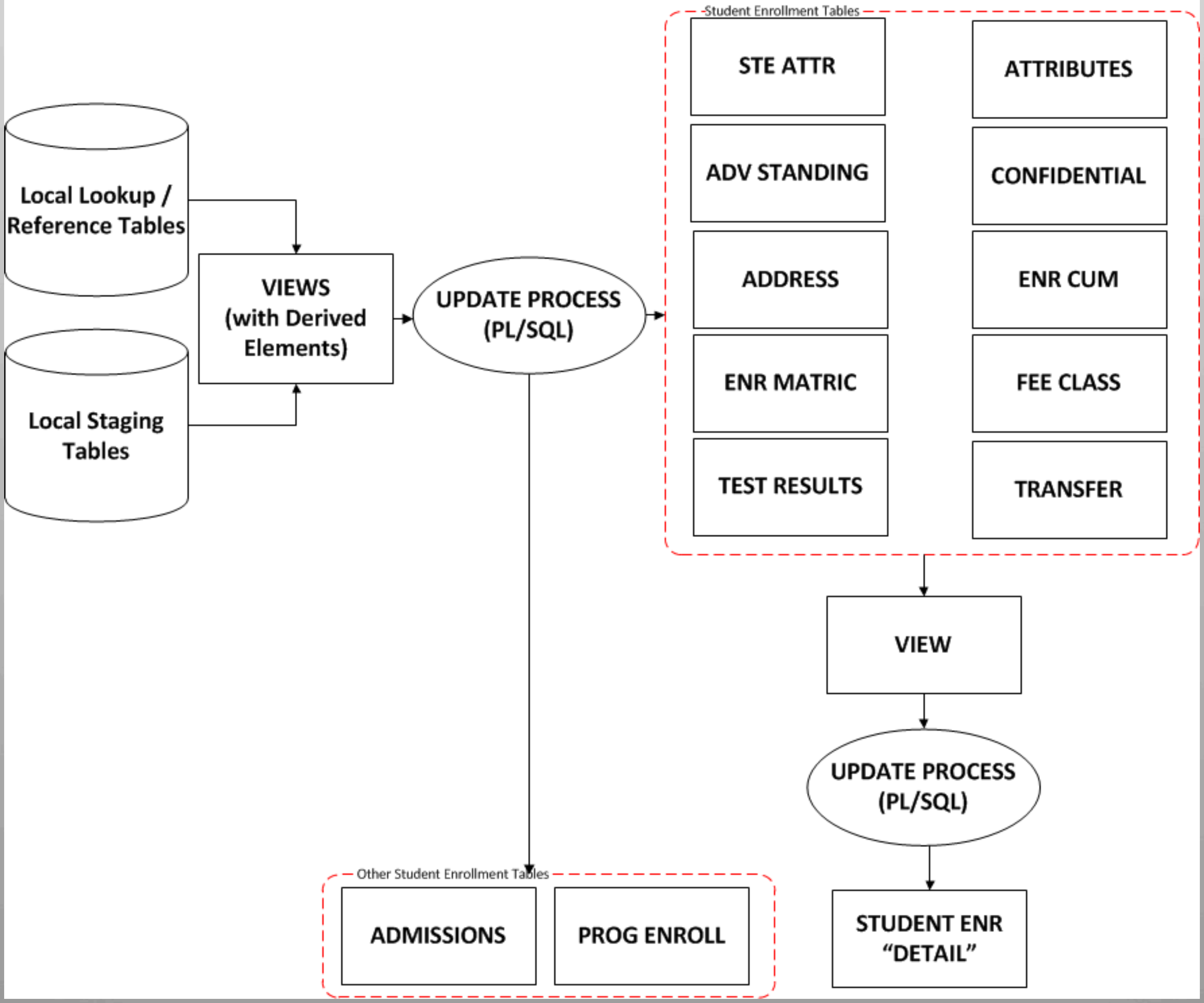
What is a Data Scientist?

- A statistician in the age of Big Data
- Uses algorithms and statistics to inform decision making

Foundation for Data Science: A Data Warehouse



Data Warehouse: Student Enrollment Tables



Data Science Checklist for Success

- Is there a resource allocation problem at hand?
- Are decision makers adjudicating between multiple solutions?
- Are there data available to conduct a proper statistical analysis of decision outcomes?
- Is there a working relationship between the data scientist and the decision makers?

Case Study:

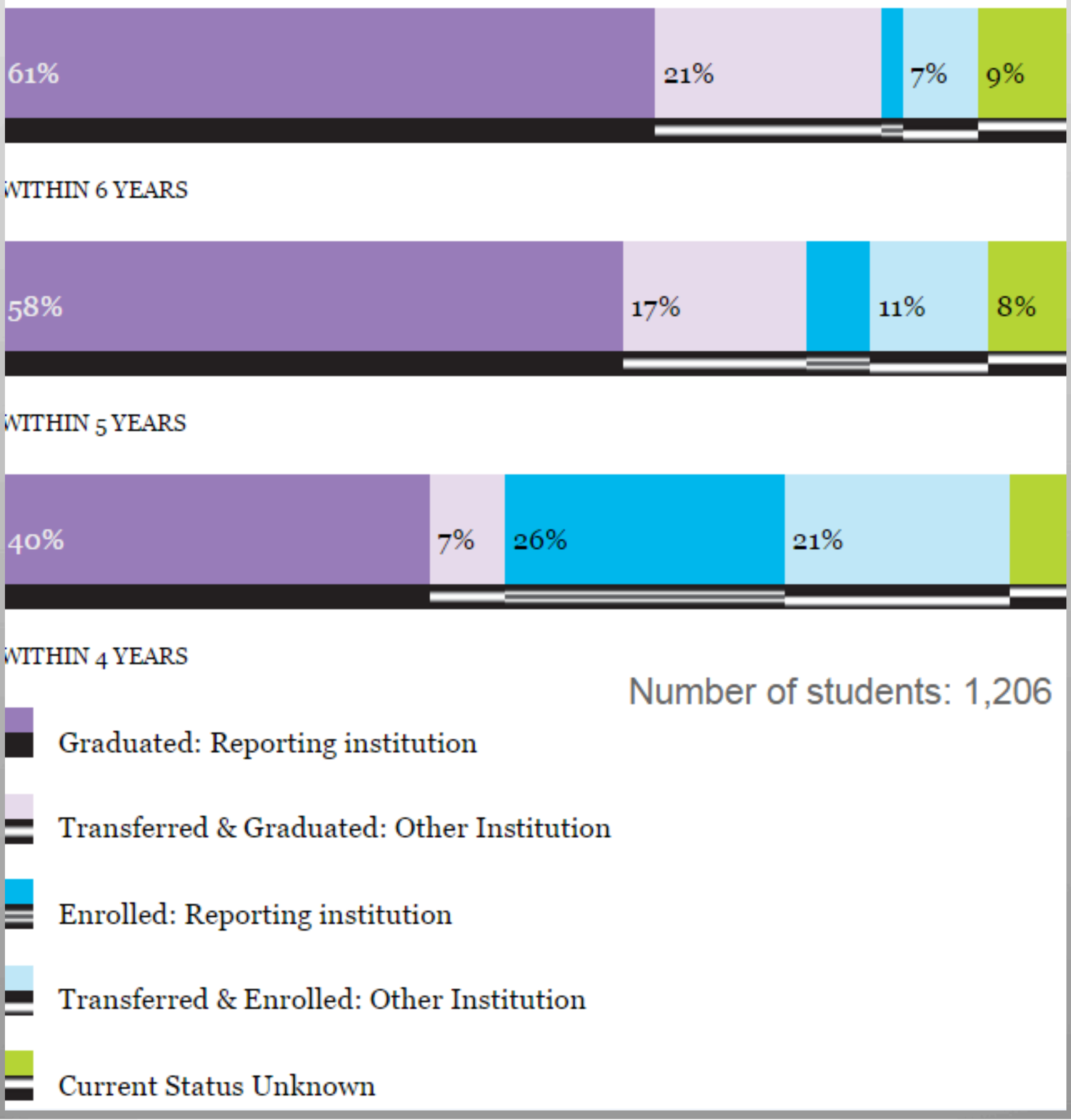
How to allocate student retention resources?

- An important metric for the success of an institution is retention rate
- There are limited resources to implement programs to enhance retention
- What subpopulations should be targeted and with what programs?

Georgia College: Transfer Risk

- An institutional goal at GCSU is to increase retention
- Approximately 1/3 of recent FTF cohorts eventually transfers to other 4 year institutions
- Question: What individuals and subpopulations are most at risk for transfer?

First-Time Full-Time Students Starting Fall 2009



Number of students: 1,206

Solution: Event History Analyses

- Question takes the form:

  ‘what factors affect whether’ or *‘what is the chance that’* an event will subsequently happen?

- Causal inference is not a goal

- Data are longitudinal

Common Regression Methods Will Be Biased

- Must simultaneously take into account:

1. Whether an event has occurred
2. Length of period at risk for the event to occur

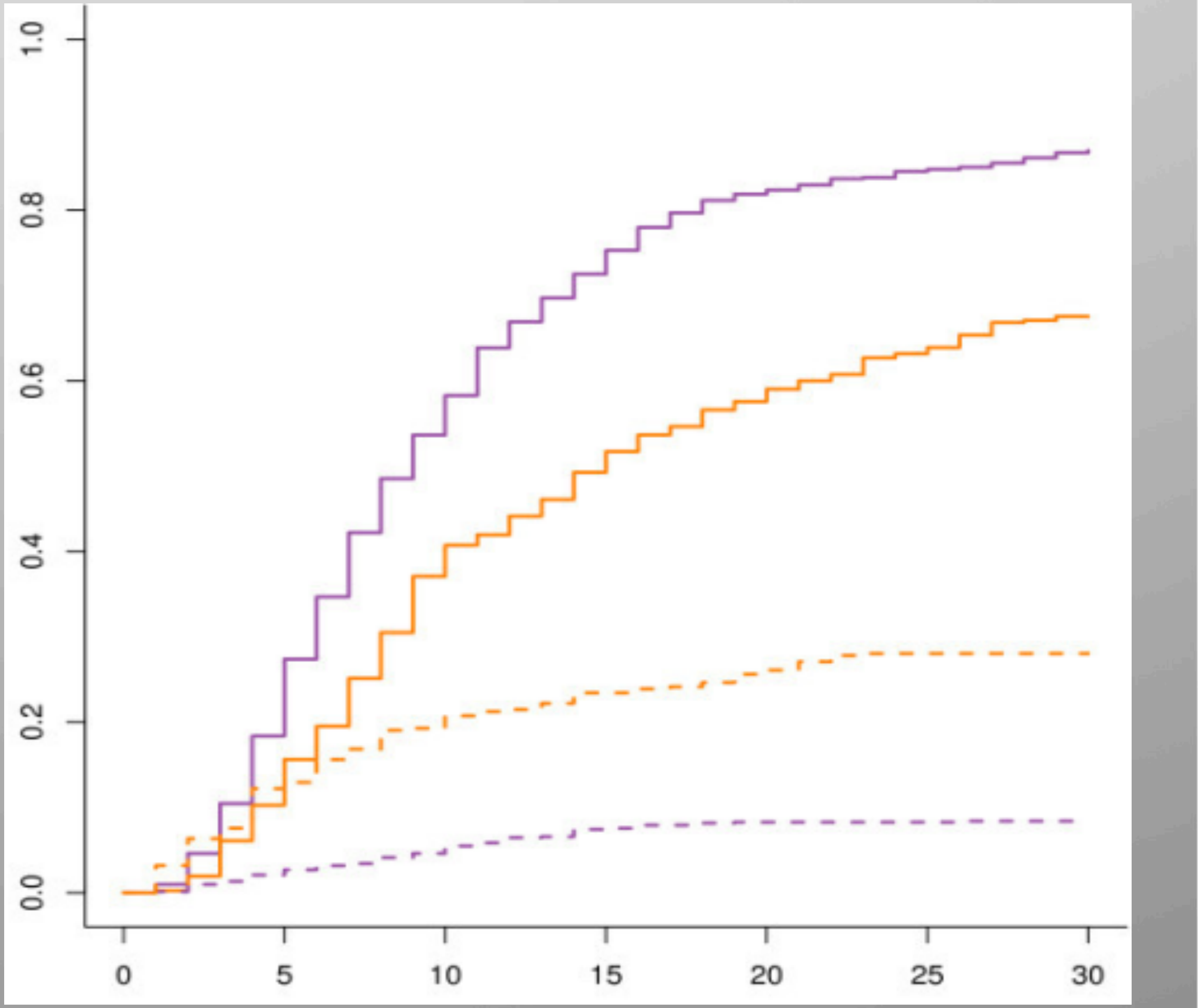
- Normal and logistic regression mathematical assumptions are not compatible

Competing Risks Analysis

Fine-Gray regression is canonical

Example output:

Probability over time four different possible outcomes



Random Forests for Prediction

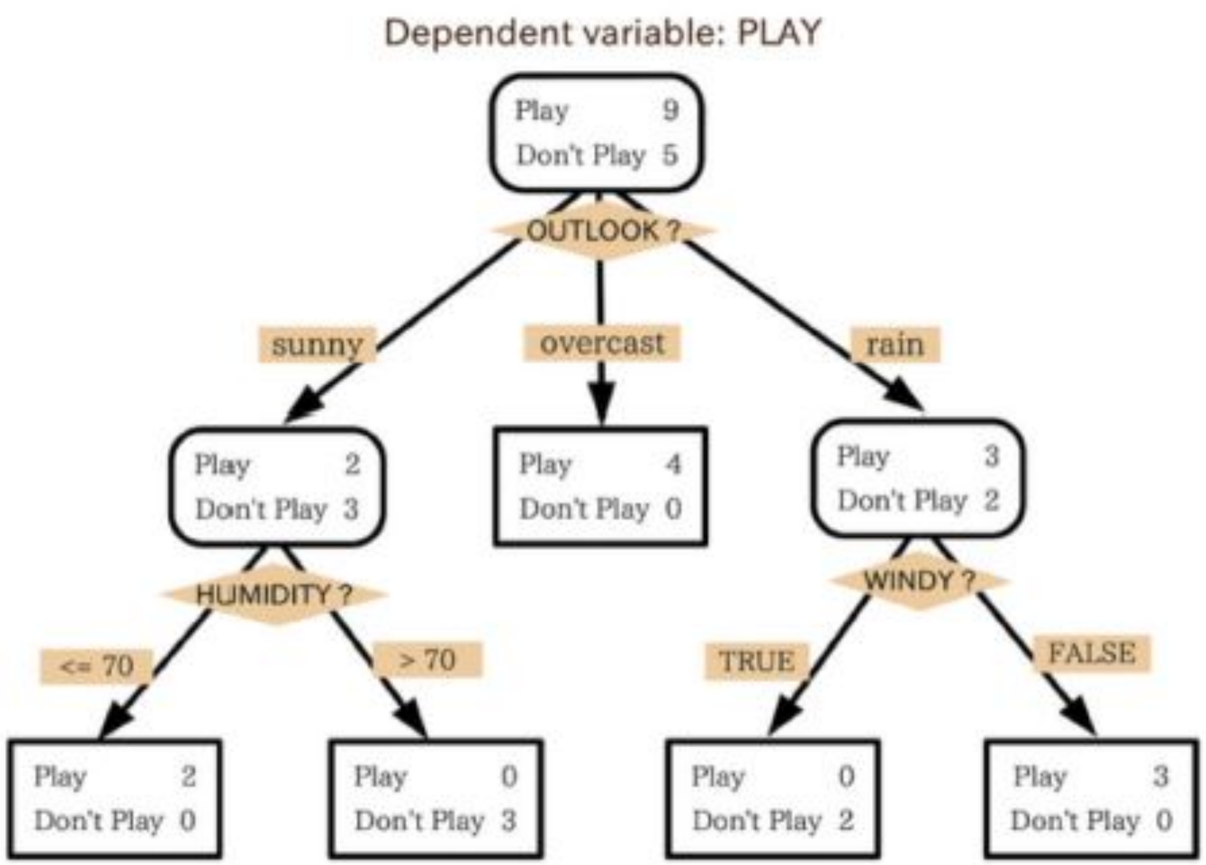
- Regression models are often fast to develop and straightforward to interpret
- Regression, however, usually relies on a set of assumptions that don't easily conform to real world noisy data
- Random forests are a machine learning technique that handles noisy data more robustly

A Random Forest Of Decision Trees

A set of features and rules are used to predict the category of each X

Overtraining avoided through randomization

Example: A Decision Tree

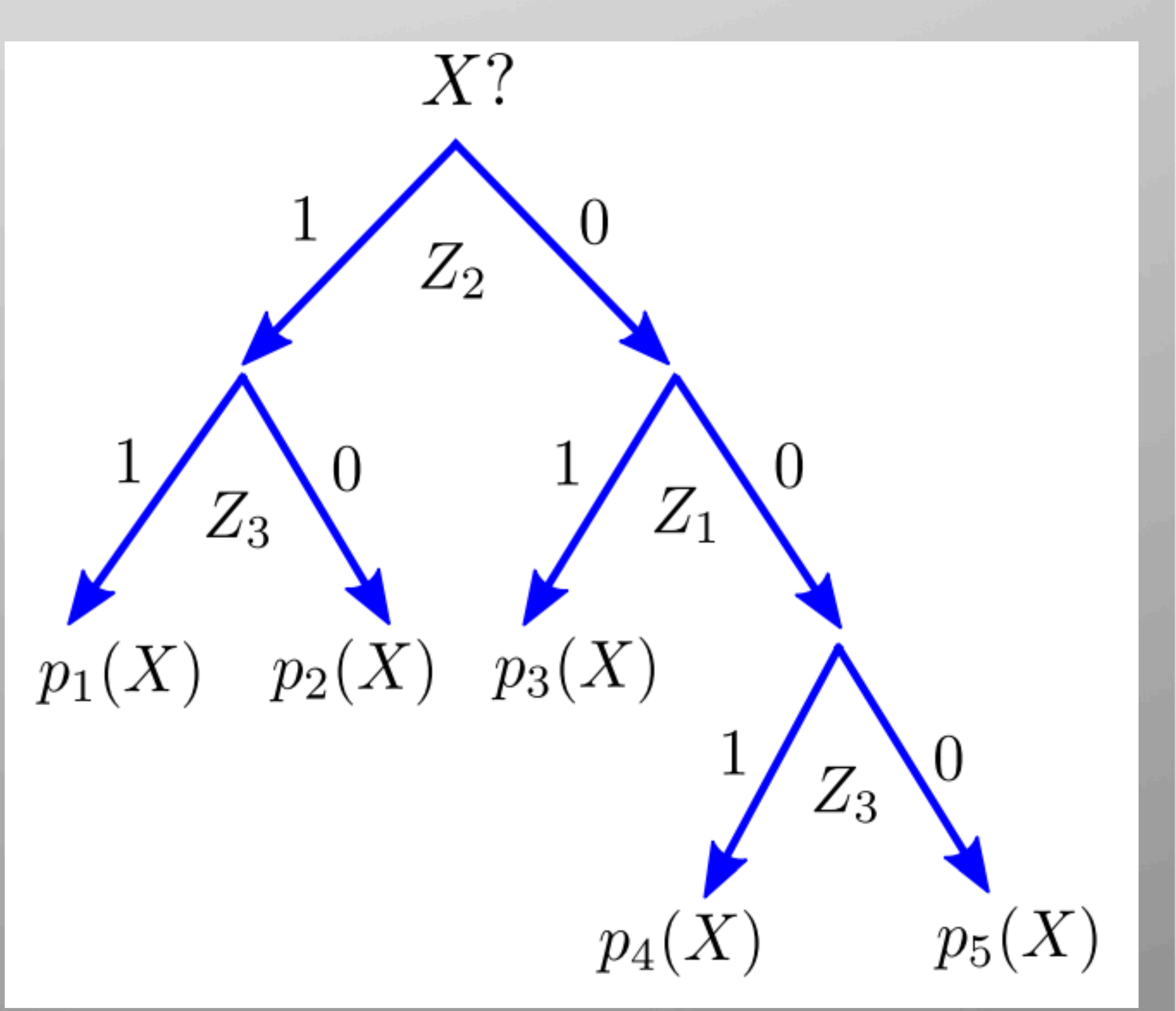


CitizenNet, 2012

Solution: Randomize and Average

Estimate many trees with a random set of features and observations

Asymptotically, the estimate will be more robust



Difficulties with Random Forest

- The method require statistical knowledge to appropriately construct and error check
- Random forests are computationally expensive with very large data sets
- Relevant data must have been already aggregated

Data Set Contents

9,945 FTF students enrolled between Fall 2007 and Fall 2015

- Oracle Databases: personal academic, demographic and financial records
- Internal Flat Files: program participation
- External Flat Files: US Census, National Student Clearinghouse

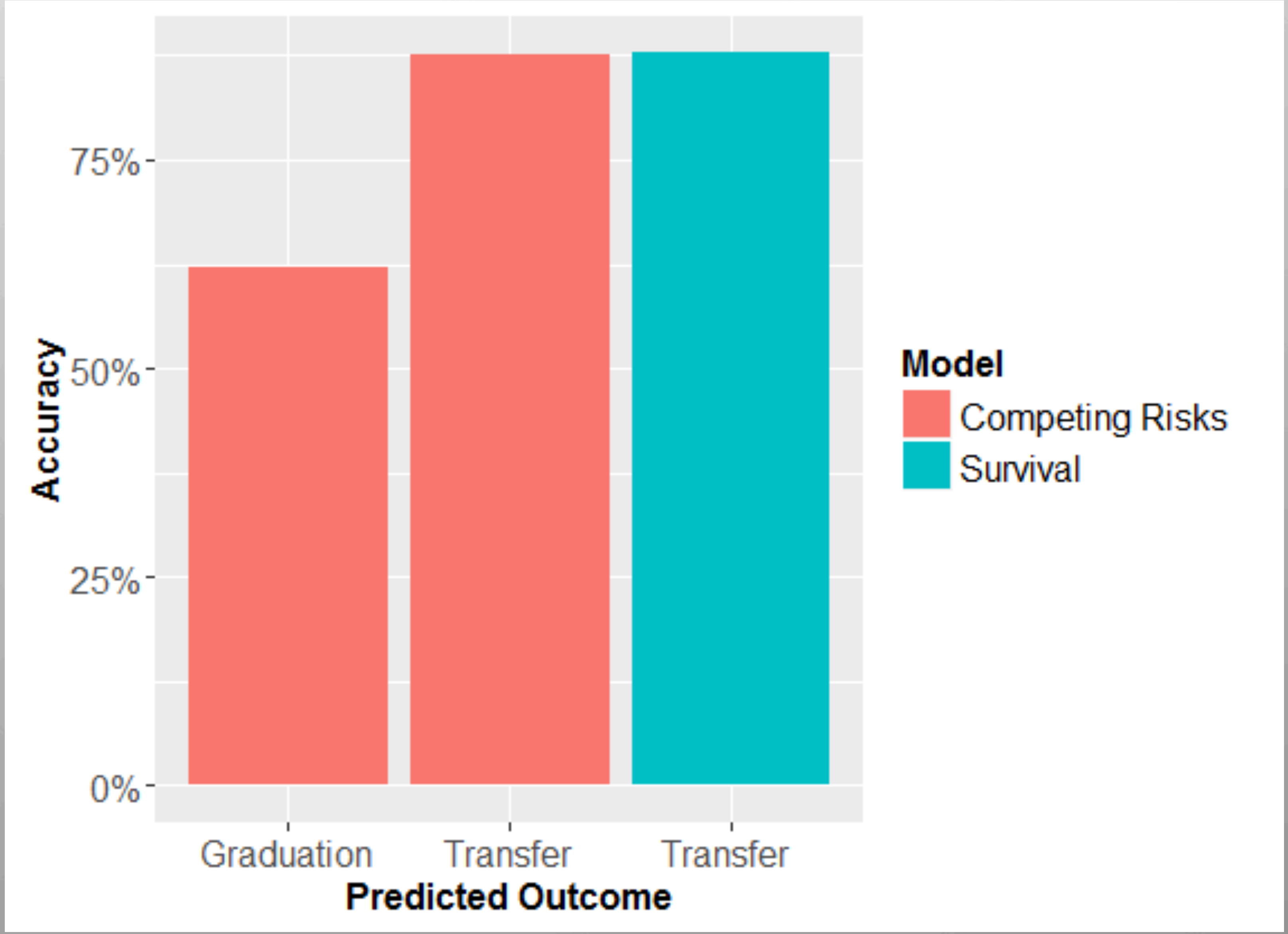
Second Step: Model Development

- Statistical Software: R (open source)
- A high level programming language and development environment with extensive statistical libraries
- R random forest package used: randomForestSRC

Second Step: Model Development

Prediction more accurate for transferring: 88% vs 68%

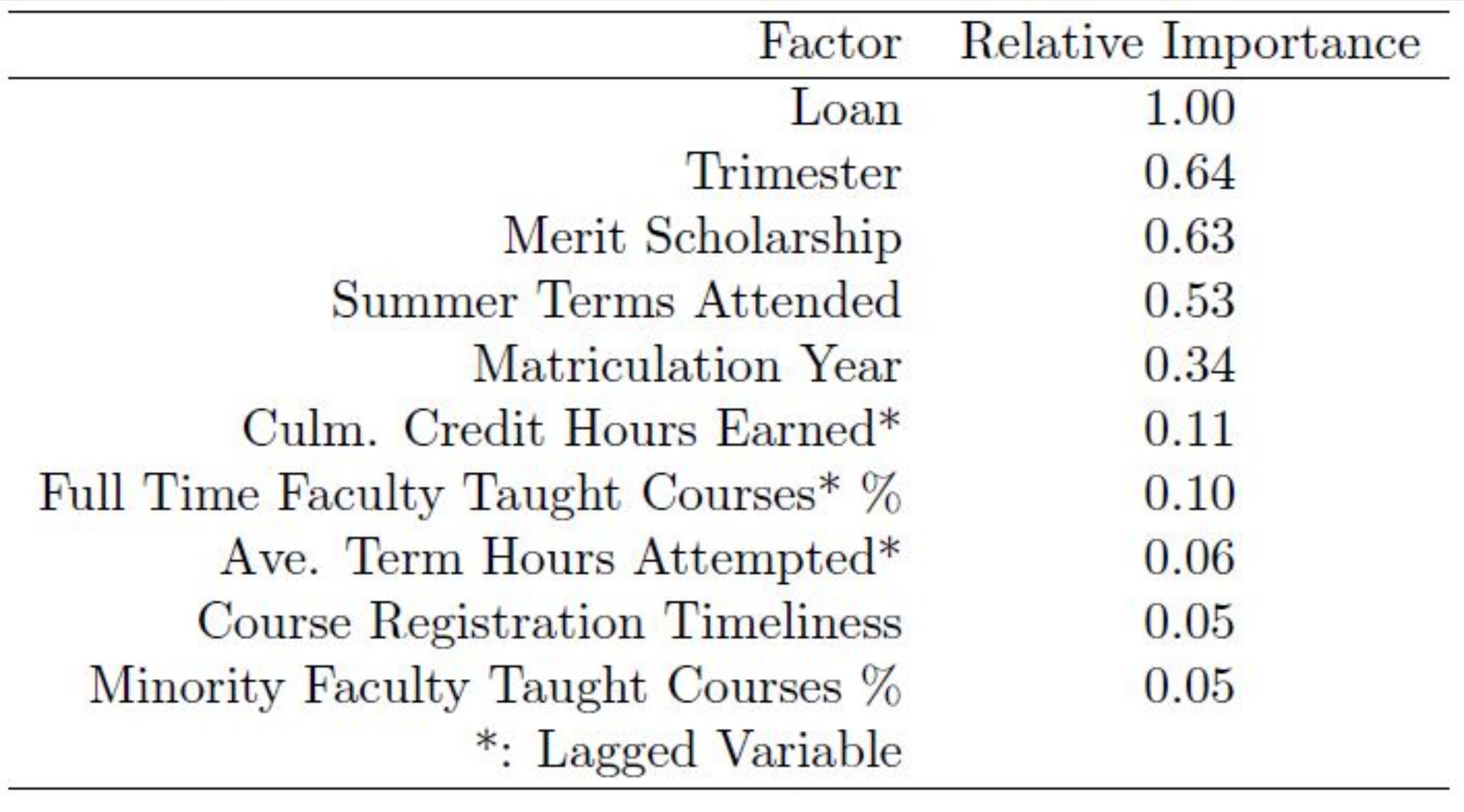
Predicting graduation requires excluding transfer



Third Step: Assist Decision Makers

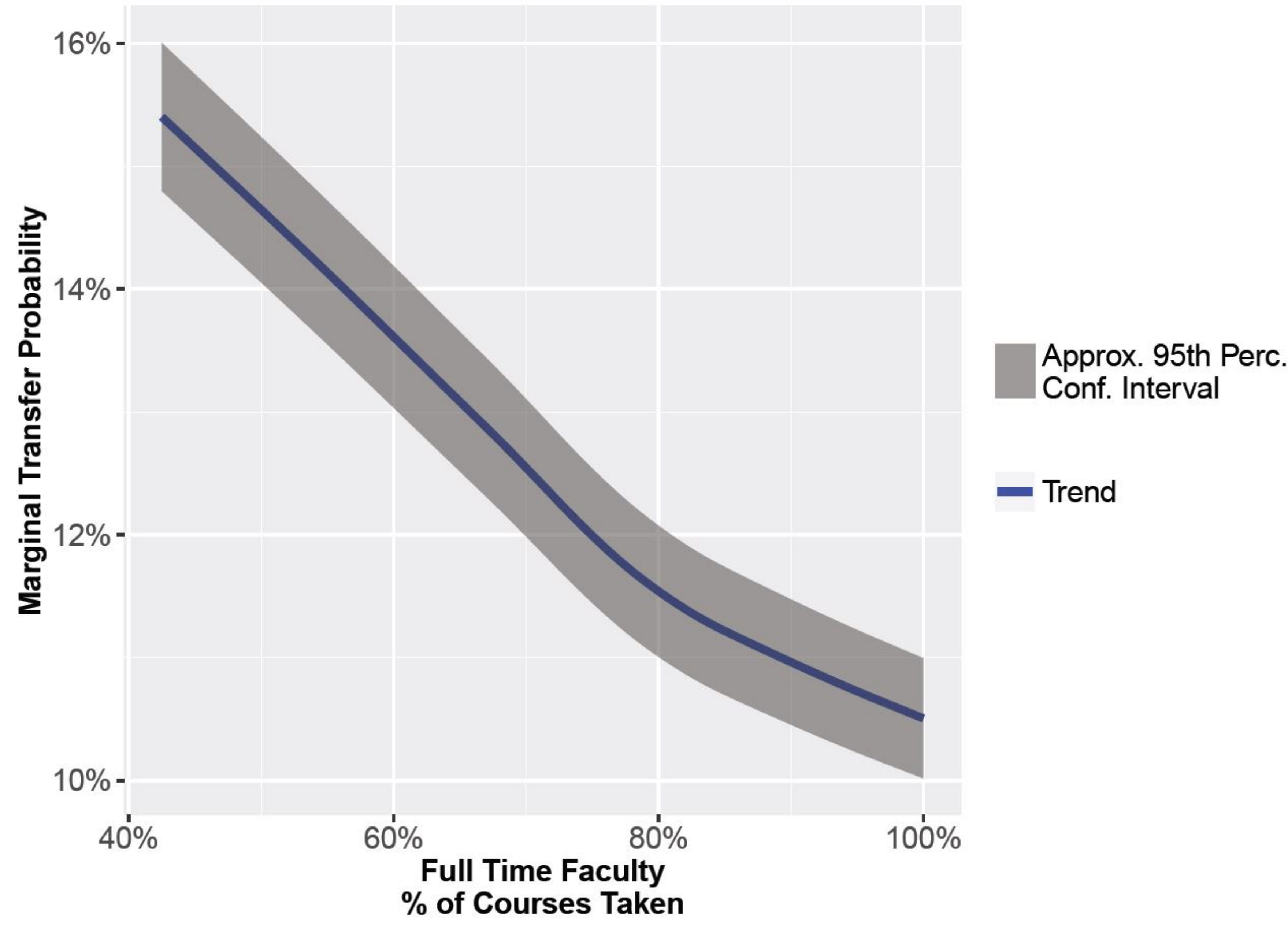
- Work with program directors to target subpopulation they can assist
- Create a dashboard that provides individual or aggregated risk predictions
- Write an executive summary that puts results in non-technical language with recommendations

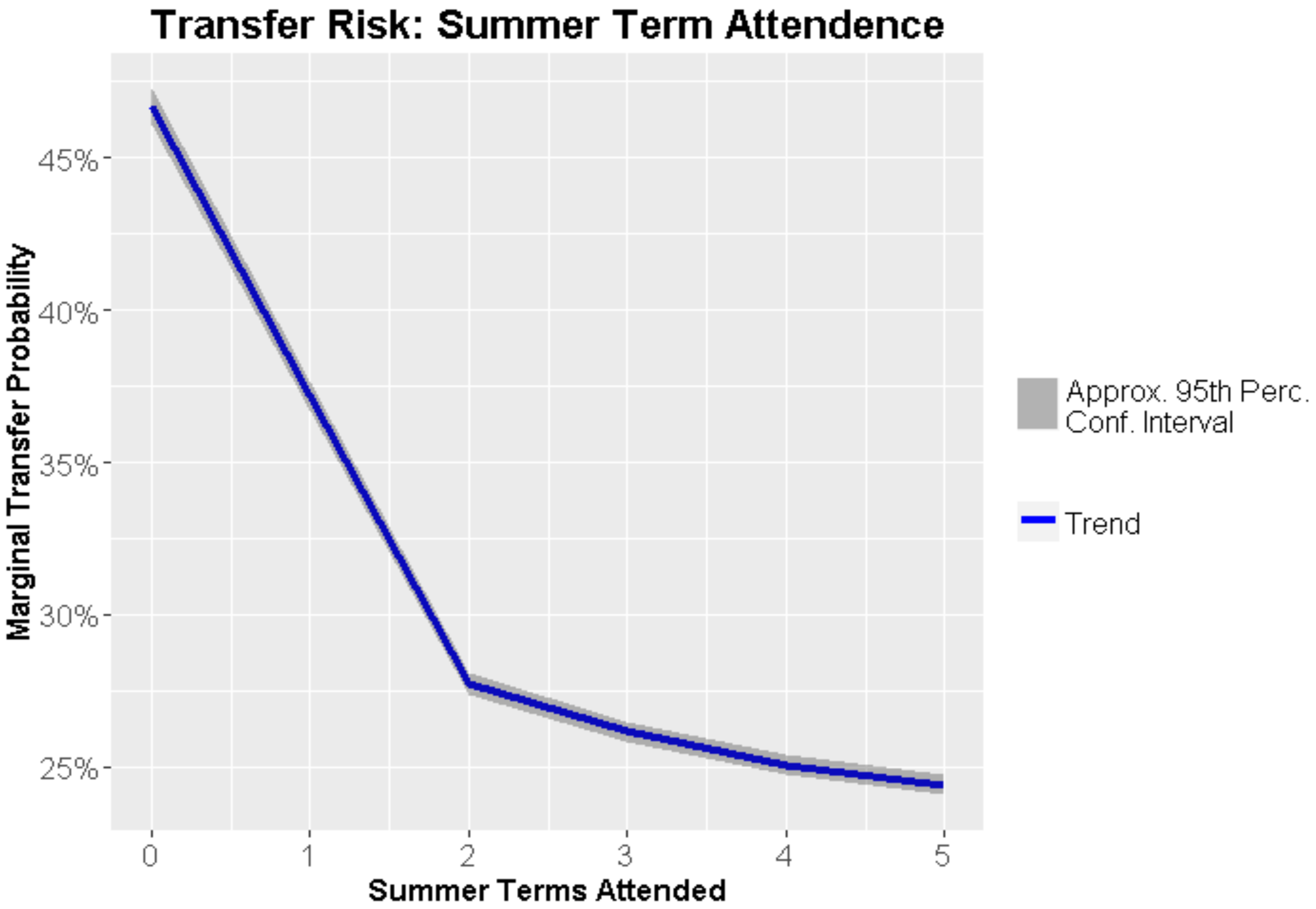
Top 10 Predictors of Transfer Risk

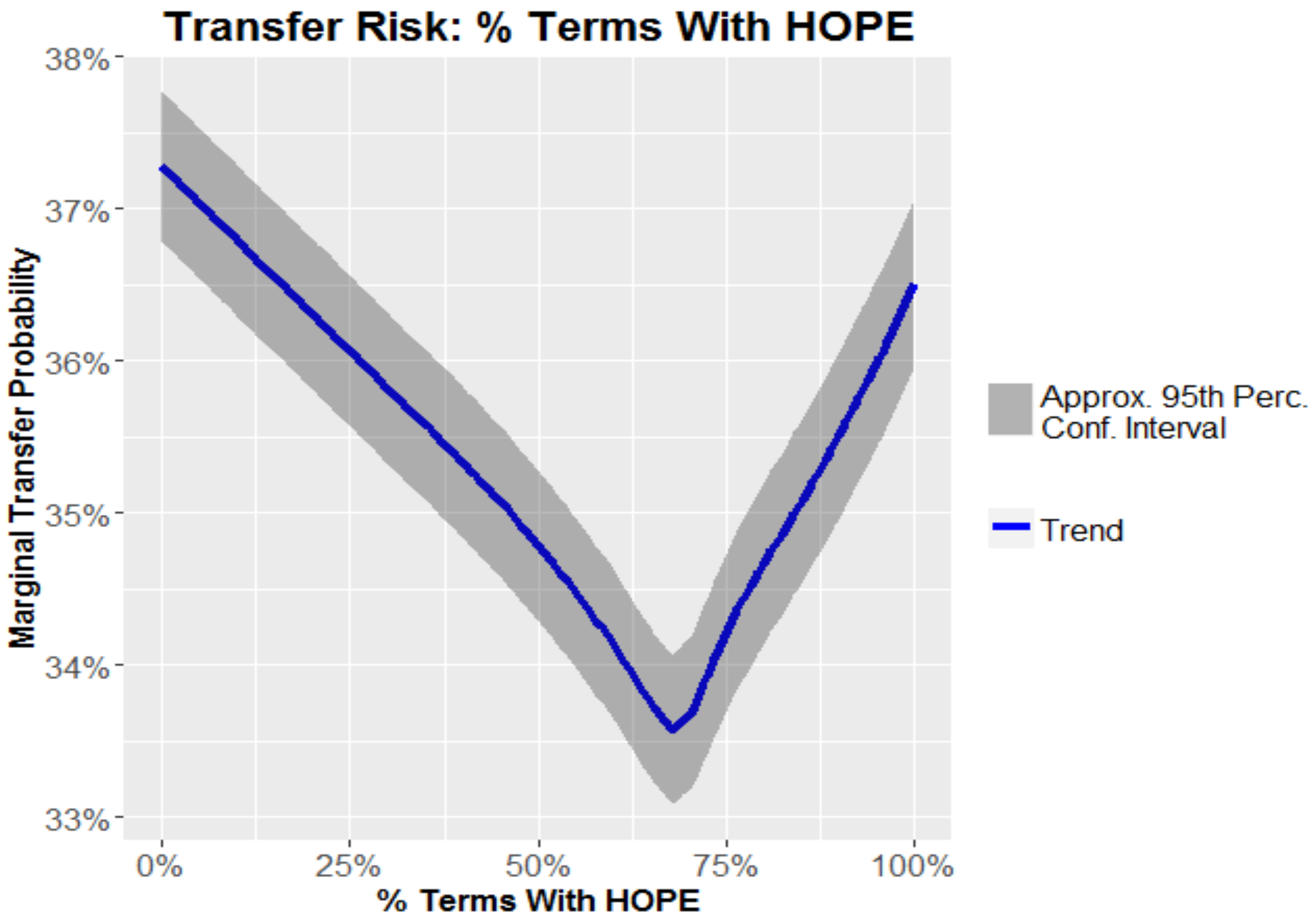


Factor	Relative Importance
Loan	1.00
Trimester	0.64
Merit Scholarship	0.63
Summer Terms Attended	0.53
Matriculation Year	0.34
Culm. Credit Hours Earned*	0.11
Full Time Faculty Taught Courses* %	0.10
Ave. Term Hours Attempted*	0.06
Course Registration Timeliness	0.05
Minority Faculty Taught Courses %	0.05
*: Lagged Variable	

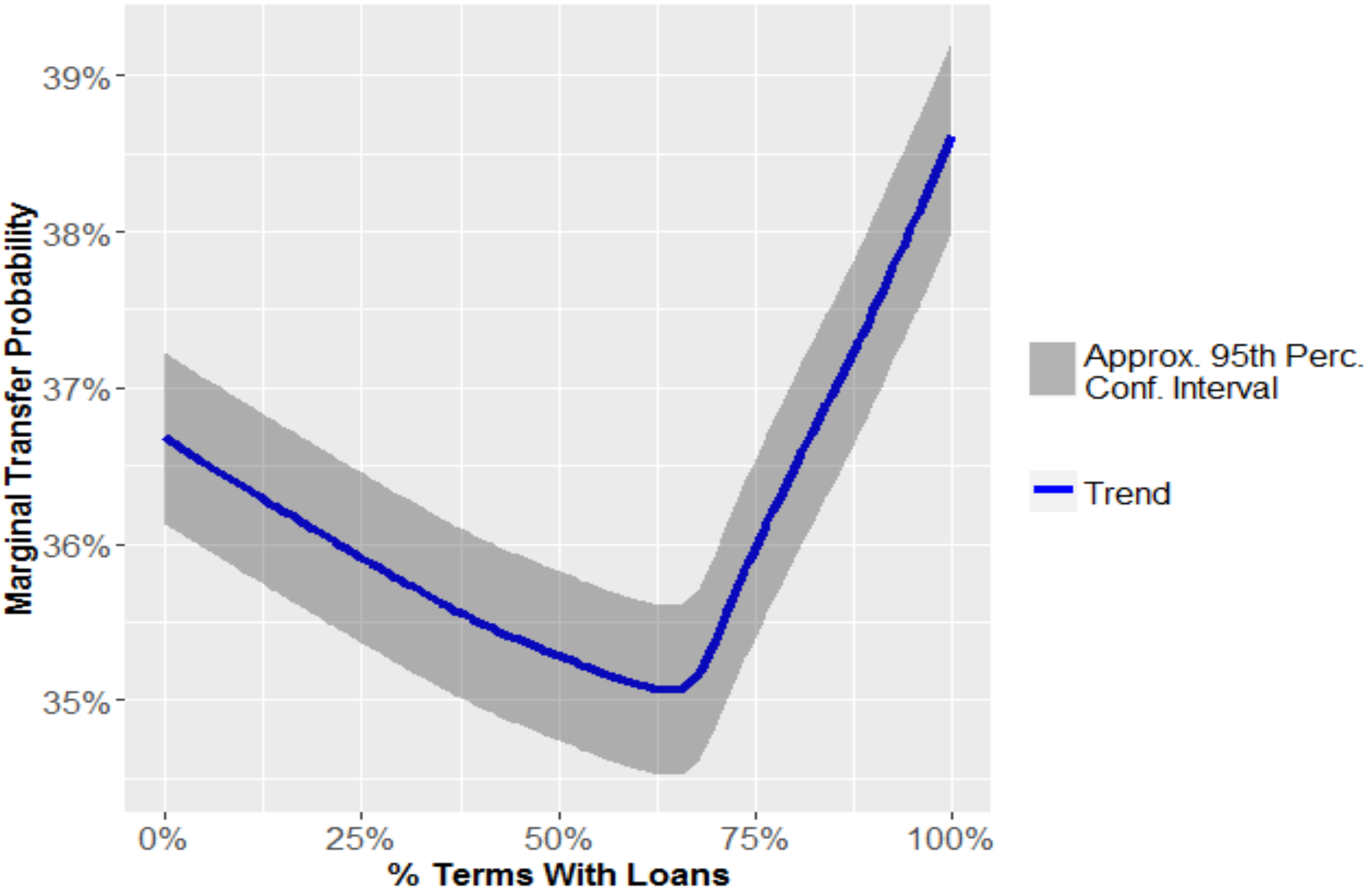
Effect of Full Time Faculty on Transfer Risk



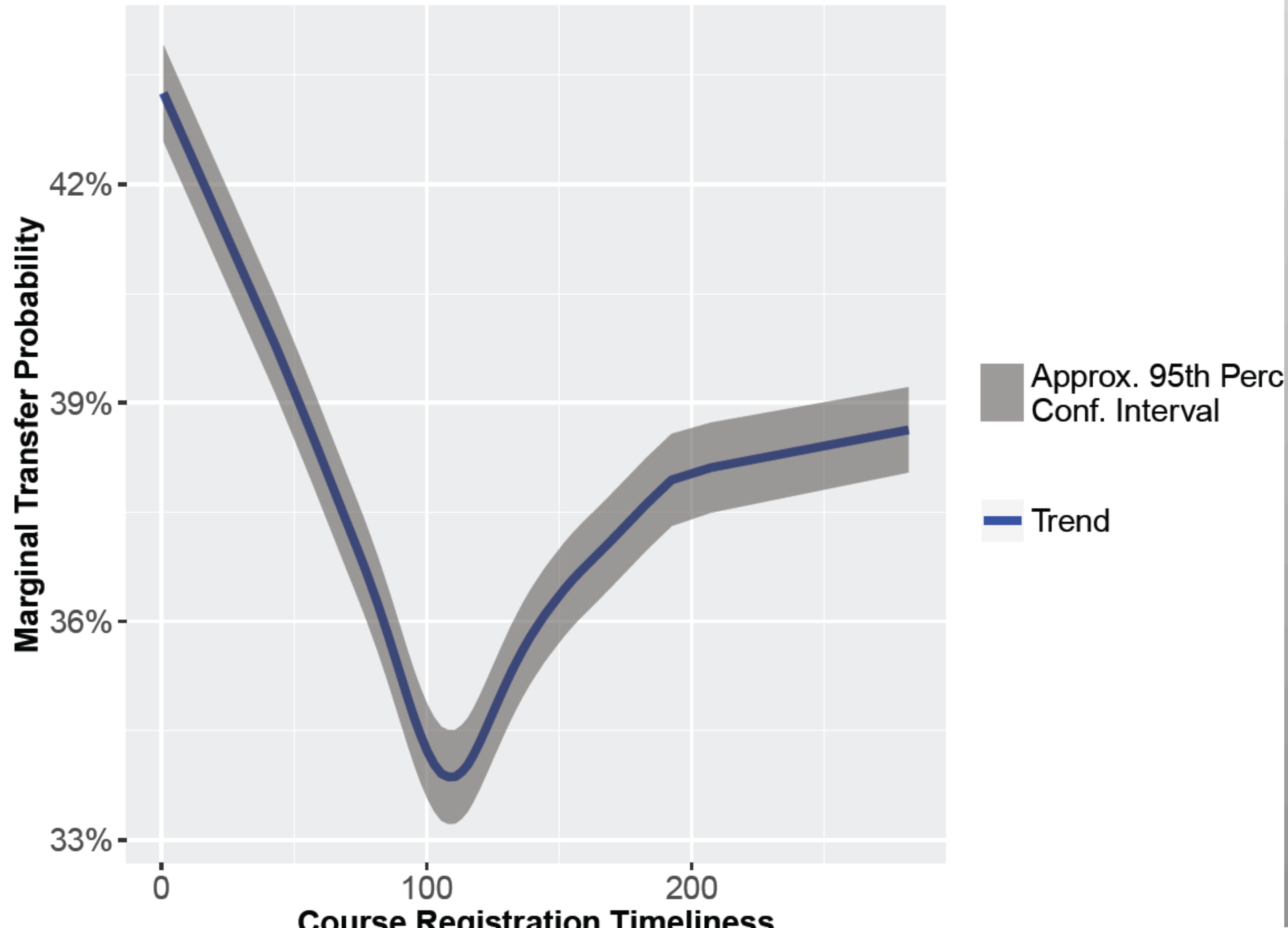

Transfer Risk: Summer Term Attendence
Marginal Transfer Probability
45%
40%
35%
30%
25%
0
1
2
3
4
5
Summer Terms Attended
Approx. 95th Perc. Conf. Interval
Trend


Transfer Risk: % Terms With HOPE
Marginal Transfer Probability
38%
37%
36%
35%
34%
33%
0%
25%
50%
75%
100%
% Terms With HOPE
Approx. 95th Perc. Conf. Interval
Trend

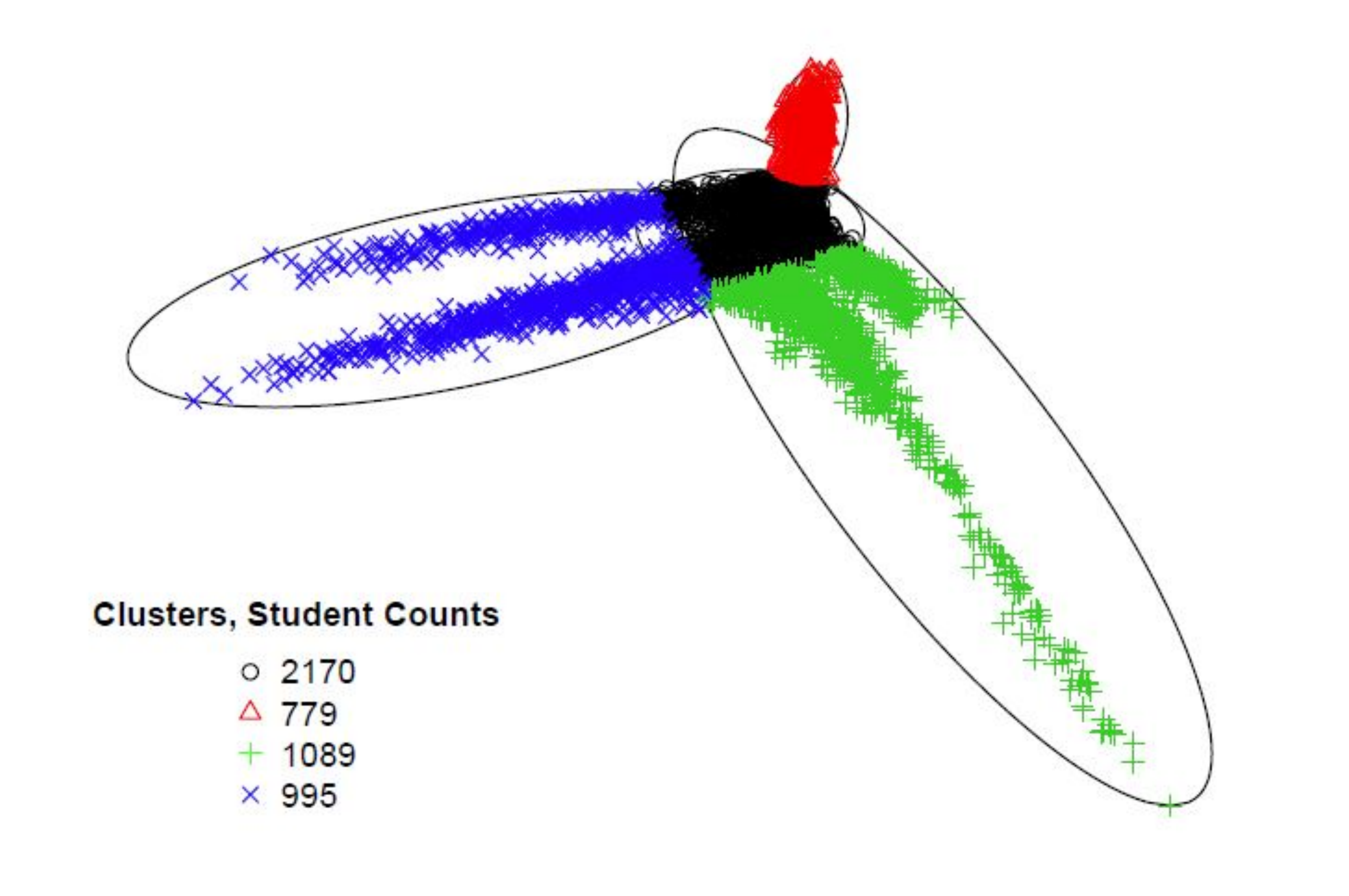
Transfer Risk: % Terms With Loans



Transfer Risk: Course Registration Timeliness



Student Cohort Characterization: Post-Hoc K-Means Clustering



Institutional Users of Model Results

- Advising

Student level predictions for at-risk student targeting

- Enrollment Management

Student and cohort level analyses for class size management

- Provost

High level policy considerations

How Can We Support Your Decision Making?

- Greatest marginal impact can be had on decisions involving resource allocation or policy making
- What decisions do you make in regards to students? Faculty? Courses?
- We hope to form a working group to develop a solution that will suit your needs

Contact Information:

Eric Braun

Office of Institutional Research and Effectiveness

Georgia College and State University

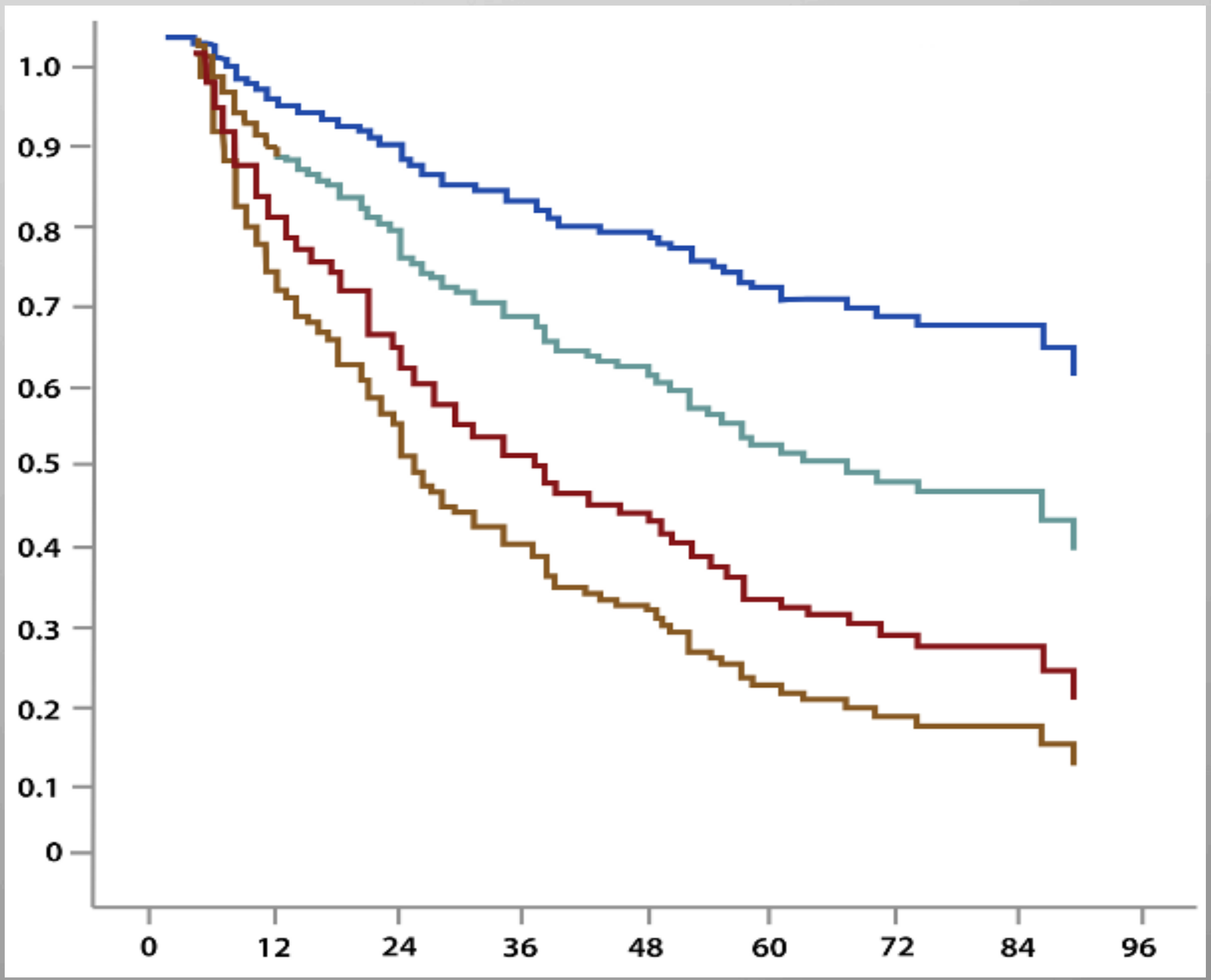
eric.braun@gcsu.edu

Survival Analysis

Cox regression is canonical

Example output:

Probability of event over time for several groups



Competing Risks Analysis

- Question takes the form:

  ‘*what factors affect whether*’ or ‘*what is the chance that*’ one of several competing events occurs?

- For example, a student may drop out, graduate or transfer from an institution

Role for Competing Risks Regression

- Theorize a set of factors predictive of an outcome
- Develop an understanding of the factors relationships through model diagnostics
- Estimate marginal effects of the factors on the probability of the outcome

Modeling to assist institutional policy

- A limited amount of resources exists to target various groups at risk of transferring
- Can students with the most potential of transferring be targeted with a reasonable degree of accuracy?
- Random forests can both provide insight