JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Volume 15, Issue 3
October 2022
ISSN: 1946-1836

Special Issue: Data and Business Analytics

In this issue:

iscap
Information Systems & Computing Academic Professionals

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is three to four issues a year. The first date of publication was December 1, 2008.

JISAR is published online (https://jisar.org) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (https://conisar.org)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the target acceptance rate for the journal is under 38%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org. Special thanks to members of ISCAP who perform the editorial and review processes for JISAR.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Editors

2022 JISAR Editorial Board

Using Analytics to understand Performance and Wellness for a Women's College Soccer Team

Christopher Njunge
cnjunge@callutheran.edu

Paul D. Witman
pwitman@callutheran.edu
California Lutheran University
Thousand Oaks, CA

Patrick Holmberg
Phomey12@gmail.com

Joel Canacoo
jcanacoo@gmail.com

Abstract

This study used analytics to examine the effect of Home Field Advantage (HFA) on the Internal Load (Session Rate of Perceived Exertion (sRPE)), External Load (from GPS trackers) and wellness (based on surveys) for a Division III women's soccer team in Home and Away matches. First the home advantage (HFA) in the Southern California Intercollegiate Athletic Conference (SCIAC) that the team plays in was quantified using conference games only for all nine teams for three seasons. A multiple regression analysis was used to analyze the relationship between Goal Difference and sRPE, External Load variables and Wellness variables at the team, position, and athlete level based on 12 athletes. The results showed the league had an adjusted HFA of 57%. The analysis showed that Defenders were impacted more by away matches based on the medium effect sizes of the difference between their home and away measures for sRPE and sleep quality, and small effect sizes for Distance in high-speed zones and stress. One athlete, a forward, had different mean sRPE in home and away matches. A regression mediation analysis was carried out to test the mediating effect of game location on the relationship between internal load and goal difference. Results showed that game location does mediate this relationship. These findings validated the existence of HFA in sports and the findings of differences in the impact of HFA to athletes in specific positions can be used as a guide in analyzing and acting on the performance and well-being of players at each position. The study demonstrated the value of analytics in gaining insights about players' performance and well-being.

Keywords: Internal Load, Sports Analytics, Soccer, Home Field Advantage, Performance

1. INTRODUCTION

Home Field Advantage (HFA) and its impact on player performance continue to be an area of interest for fans and practitioners of sports. This paper sought to use analytics to analyze the performance of soccer players in home versus away matches to quantify the differences in their performance and increase the understanding of these differences. Soccer is a sport involving eleven players on each side and featuring one goalkeeper and 10 outfield players. It is played with a round ball and players' positions are often

divided into defenders, midfielders and forwards. Goals are often few, typically less than three over the two 45-minute halves played.

One framework seeking to explain HFA has five major components: game location, game location factors, critical psychological states, critical behavior states and performance outcome (Carron et al.,2005). This framework suggests that multiple factors account for the variation in players' performance and well-being in home and away matches. Some of the non-physical factors potentially influencing subjective well-being that Abbott et al. (2018) discussed include match location, quality of match opposition and match result; collectively called situational match variables. Bailey (2013) has proposed regular testing and single-subject design when taking part in athlete monitoring and tracking.

Novelty and Practical Relevance – Use of analytics in analysis of HFA in sports using team, position and individual match variables can extend the understanding of home advantage in sports in general and soccer in particular.

Theoretical Contribution – This paper uses analytics to highlight the differences in the way athletes and players in different positions, even on the same team, are impacted by home field advantage in soccer. These findings can be extended to other sports in order to improve preparation of athletes and teams and improve their performance and well-being.

2. LITERATURE REVIEW

A review of the literature was carried out using the search terms 'Home Field Advantage', 'Internal Load', 'External Load' and 'Wellness'. Out of the 254 articles identified, 76 were selected for further review based on relevance of the title as determined by the authors. Of the 76 selected, further analysis was carried out by reviewing the abstract and high-level reviews of the methodology and findings.

A further review of the literature was carried out with the following search terms: Sport, Match Play, GPS, Level, Gender and Region. The articles were categorized by sport and whether or not they featured match play, meaning individuals or teams competing. 46 of the 76 articles related to soccer and four of them featured more than one sport. 50 of the 76 articles featured match play and 36 used GPS devices in their study. 30 of the studies were conducted using professional players as research subjects, 21 featured amateurs and 13 identified the subjects as elite. Most of the articles, 61, researched males, eight both male and female athletes, while only five specifically carried out research on female athletes. More than forty of the articles featured research carried out in Europe or the UK and only 11 of the studies were in the US.

Gaps

A review of the literature shows a gap in the study of soccer athletes in the US and particularly female athletes. The use of GPS devices in quantifying external load was quite prevalent as was the use of self-reported RPE to identify internal player load and wellness surveys to track the wellbeing of players. Only one study aimed to use machine learning algorithms to predict the internal load based on external load measures, and this was for professional soccer players in Europe. The authors are unaware of any study where analysis of internal load, external load and wellness of soccer players is carried out at a US Division-III college level (D-III is an indication of relatively small school size and does not permit athletic scholarships). This study seeks to begin to close those gaps while answering new research questions.

3. RESEARCH QUESTIONS

This paper seeks to investigate the following research questions:

- Does homefield advantage exist for a DIII women's soccer team and league?
- Are changes in players' wellness and external load measures reflected in the outcome of a match at the team and position level?
- Do the players' internal load, external load and wellness measures differ for home versus away matches?
- Does game location mediate the relationship between internal load (sRPE) and goal difference?

4. METHODS

In selecting the athletes to include in the study several factors were considered. First, three seasons in which the GPS tracking, internal load and wellness data was available for the entire season were selected. During those seasons 50 players had appeared for the team. Given that only 10 outfield players are used in a match, there was significant variation in the data

available for the 50 players. The players were therefore ranked based on their mean scores on the following variables: External Load from GPS Trackers (Load, Duration and Distance) as well as Total Games Played. Their ranking on these measures was then added together and divided by the number of measures to find the average ranking. The top 12 players were then selected, as they had all played more than 20 matches in the time frame.

For the internal and external load analysis, three seasons' worth of data were collected for the 12 athletes in the study. The categories of data collected and their respective sources were: External Load (PlayerTek - https://www.playertek.com/us/), sRPE/Internal Load (FitFor90 - https://www.fitfor90.com/) and Wellness (FitFor90). The external load data in Playertek included metrics captured by Catapult GPS trackers during matches. The data show players' total distance covered, duration, high speed running, accelerations and decelerations. The Internal Load data in FitFor90 included responses to surveys that the players filled out after matches indicating the difficulty of the session (rating), and the duration. Session rating is multiplied by duration to arrive at RPE (Rate of Perceived Exertion). The wellness data in FitFor90 included responses to surveys that players filled out daily related to their wellbeing. Measures include: Fatigue, Mood, Soreness, Stress, Sleep Quality and Sleep Hours. A total of 27 different variables were captured, of which a small subset was later found to explain variation in the regression model.

Multiple regression analysis using the enter method was used to analyze the relationship between the Goal Difference (dependent variable) and the sRPE (internal load), external load variables, and wellness variables.

For the regression mediation with moderation, the 'mediate' package in R was used to carry out the analysis. Internal Load, Stress, and Goal difference were used as the X, Y and M variables respectively. The indirect effect was calculated using the method proposed by (Sobel, 1982). Nonparametric Bootstrapping with the Percentile Method were used to test the significance of the indirect effect of the mediating variable, Stress. Analysis of the moderating effect of game location on goal difference was carried out.

The data was pulled by exporting CSV files with player-level data from the FitFor90 and Playertek sites, as entered by the players. The data was then cleaned, aggregated, and analyzed using R-Studio. Once imported and cleaned, the data was joined to enable the multiple regression analysis to be carried out. sRPE was calculated by multiplying the athlete rating of a session (game), by the duration of the session.

5. RESULTS

HFA was quantified per team, per season, using the widely accepted approach of defining the number of points gained at home as a percentage of all points earned (Pollard, 1986). However, this method has been shown to be impacted by the team's ability and the overall aggregated home advantage of the league during that season. As such, both of these factors need to be accounted for when comparing home advantage from different seasons. Pollard and Gomez (2009) have described a process where team ability and league HFA for a season can be calculated.

An HFA of over 50% indicates a positive home field advantage. The resulting HFA for the league was 60%, 54% and 58% for the 2017, 2018 and 2019 seasons respectively and 57% over the three seasons, indicating the existence of HFA in the league. The research subject team's HFA for the 2017, 2018 and 2019 seasons was 54%, 57% and 62% respectively and 58% on average for the three seasons. Only regular season matches (40 in total) for the three seasons were included in the HFA analysis in order to have a comparable number of home and away matches.

In response to research question one, the results show that home field advantage does exist both for the team and the league.

Regression Analysis Results

A multiple linear regression analysis was carried out at the team and position level, and results are shown in Table 1. At the team level, the regression results showed an R Square value of 80.4% and an Adjusted R Square value of 49.2%, indicating that the independent variables (sRPE, Distance in Speed Zone 5, Distance in Deceleration Zone 1, Stress, and Sleep Quality) indeed explain nearly half of the variation in the dependent variable of Goal Difference and thus the outcome of a match. Table 2 shows the differences in the means of these variables in home and away matches. The 27 independent variables initially included consisted of several categories as used in similar analyses (Jaspers, 2018) and as shown in Table 3.

Table 1 shows the five independent variables that had a positive correlation with Goal Difference and were significant at the 95% level. No variables were significant for midfielders only, and the share of variability of the dependent variable Goal Difference explained by the independent variables for forwards was very low (.027). These results show that defenders' independent variables have the strongest correlation with the dependent variable, Goal Difference.

In response to the second research question, 'are changes in players' wellness and external load measures reflected in the outcome of a match at the team and position level?', the results show that changes to these measures are statistically significant at a position level but not at a team level. See Figure 2 for the differences in mean player load by position.

T-Test Comparing Variables and Matched Pairs

The variables that were significant were therefore selected for further analysis in comparing them for home and away matches. This was carried out to establish whether their means were statistically different during home matches versus away matches. None of the five significant variables had p-values of less than the alpha value of 0.05, indicating that the means were not statistically different for home and away matches based on the sample. However, the effect sizes were still included. A paired samples t-test was used, and the Cohen D statistic was used to measure the effect size of those differences as shown in Table 2 below. These results show that at the team level the mean of the five variables that correlated with Goal Difference, sRPE, Distance in Speed Zone 5, Distance in Deceleration Zone 1, Stress and Sleep Quality were not statistically different in home versus away matches as reflected by their low p-values. Despite the lack of significant difference in this data set sRPE and Sleep Quality had medium effect sizes and warrant further investigation.

Similar to the team-level analysis, the means of the five variables selected based on the regression analysis were not statistically different based on the matched pair t-test. However, their Cohen's D effect sizes are included in Figure 1. Additionally, a player-level analysis was carried out to establish whether the mean of the five variables above was statistically different in their home versus away matches. Figure 2 shows the internal load of players in three positions. For one of the players, whose position was forward, the sRPE (internal load) had a p-value of .043 indicating that the mean of their sRPE is statistically different between home and away matches. This suggests that there are likely differences in the way individual players experience and are impacted by home and away matches. A follow up study such as that proposed by Baily (2013) that studies single subjects may shed more light on the impact of home and away matches on individual athletes based on additional measures.

In response to the third research question, "do the players' internal load, external load and wellness measures differ for home versus away matches?", the results show that these measures are not statistically different for home versus away matches for players on this team over the three seasons. The differences in the means of the variables that were statistically significant in the regression analysis were not statistically significant at a team or positional level. However, for one of the players, a forward, the mean of sRPE in home versus away matches was statistically different.

6. DISCUSSION

This study used analytics to evaluate performance and wellness of female college soccer athletes on one team over three seasons and found that HFA exists in the SCIAC conference at 57%, and the team whose athletes were selected for analysis had a home advantage of 58%. Five variables, sRPE, Distance in Speed Zone 5, Distance in Deceleration Zone 1, Stress, and Sleep Quality were found to be significantly correlated with Goal Difference. It is notable that of the five variables, one related to internal load, two related to external load and two to wellness. Although the team-level and position-level means of these variables was found to not be statistically different given the sample size, one of the athletes had significantly different mean sRPE between home and away matches. Given that only 12 out of 51 athletes were selected for the analysis (limited by their tenure with the team across all three seasons of study), it is likely more players have internal load, external load and wellness variables that are different between their home and away matches. If coaches and trainers can identify the players with significant differences in these variables for home and away matches, they can intervene or modify their performance to improve individual, position, and team performance.

Regression Mediation Analysis
Since the relationship between internal load (sRPE) and goal difference was validated in the previous regression analysis, additional regression analysis was carried out to investigate whether stress mediated the relationship between internal load and goal difference. The approach proposed by(Barron & Kenny, 1986) was followed.

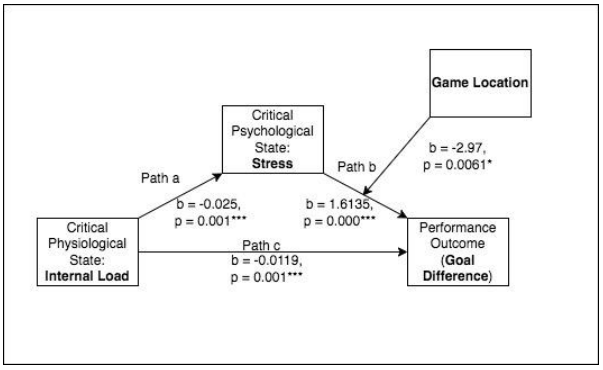


Figure 3: Linear regression results for the mediation model with moderation

Mediation is a causal chain showing how one variable impacts another which subsequently impacts another. In this analysis, the Home Field Advantage model (Carron et al.,2005) was used to investigate whether game location mediated the relationship between internal load (sRPE) and goal difference. The mediation analysis was carried out in R-Studio using the 'mediation' library. The analysis was carried out in three steps, in line with existing literature (Barron & Kenny, 1986).

First the relationship between internal load and goal difference was tested. The total effect of Load on Goal Difference is negative and significant as shown in Figure 3, 'path c' (b = -0.0119, p <0.001***). Next the relationship between internal load and stress was tested, Figure 3, 'path a' (b = -0.0025, p <0.001***). Next the relationship between stress and goal difference Figure 3, 'path b' (b = 1.6135, p <0.000***). The indirect effect of the moderating variable, stress, on Goal difference was then calculated by multiplying the coefficients of 'path a' and 'path b' in Figure 3 above (a= -.0025 * b = 0.9023), yielding -0.0022. This indirect effect is lower than the 'path c' beta value of 0.0119, meaning that stress partially mediates the relationship between internal load and goal difference. Finally, the effect of venue on the relationship between stress and goal difference was calculated and was positive and significant. Thus validating that match location impacts the success of a team as measured by goal difference and evidences the existence of home field advantage. Readiness, a measure of players' wellness and based on a questionnaire the athletes take, was analyzed but its relationship to goal difference was not significant. Given the similarity of readiness to self-efficacy, further analysis is encouraged to establish ways that players' psychological states impact their performance and whether this is mediated by game location, thus causing a home field advantage.

Three of the four research questions were validated through the research. The study showed that home field advantage does exist at the league and team level over the three years. The study also showed that changes to players' external, internal and wellness load measures impact the outcome of a game, as measured by goal difference, particularly at the position level. For the third research question, the study showed that the internal, external and wellness measures were not statistically different in home versus away matches.

Limitations and Future Work – Given the duration of the study was three years, it limited the number of athletes that could be included in the study, which in turn limits the generalizability of the results. Broader (more seasons) and deeper (more athletes) analyses can be carried out to validate the findings of the study. Additional match variables such as wellness or injuries can be added to the study to better understand impacts to players beyond the day of the match. Longitudinal data to explore causal relationships is not available and would add to the research. Also, providing insights to athletes and coaches in a short period of time, during or soon after the game, can help teams make adjustments within or between matches and improve their performance.

7. CONCLUSION

Given the increased importance of sports in society in general, and of college sports in particular, understanding team, position, and individual athlete experiences of matches both at home and away is crucial. Since a minority of the athletes are likely to become professional players, understanding how they are impacted by the sport and keeping track of various match variables in home and away matches can improve their experience as students, workers, and family members and improve their individual well-being as well as that of society.

8. ACKNOWLEDGEMENTS

I would like to acknowledge my co-authors for their tireless work on the paper, and thank the reviewers and conference chairs for their constructive feedback and support.

9. REFERENCES

Abbott, W, Brownlee, T, Harper, LD, Naughton, RJ and Clifford, T (2018) The independent effects of match location, match result and the quality of opposition on subjective wellbeing in under 23 soccer players: a case study. Research in Sports Medicine. ISSN 1543-8627

Armatas, V. & Pollard, R. (2014) Home advantage in Greek football, European Journal of Sports Science, 14(2): 116-122

Arne Jaspers, Tim Op De Beéck, Michel S. Brink, Wouter G.P. Frencken, Filip Staes, Jesse J. Davis, and Werner F. Helsen (2018) Relationships Between the External and Internal Training Load in Professional Soccer: What Can We Learn From Machine Learning?, International Journal of Sports Physiology and Performance, 2018, 13, 625-630, DOI: https://doi.org/10.1123/ijspp.2017-0299

Bailey, C. Longitudinal Monitoring of Athletes: Statistical Issues and Best Practices. J. of SCI. IN SPORT AND EXERCISE 1, 217–227 (2019). https://doi.org/10.1007/s42978-019-00042-4

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51(6), 1173-1182.

Carron, A., Loughhead, T. M & Bray, S. R. (2005), The home advantage in sport competitions: Courneya and Carron (1992) conceptual framework a decade later, Journal of Sports Sciences, 23:4, 395-407, DOI: 10.1080/02640410400021542

Davenport, T. H. 2014. "Analytics in Sports: The New Science of Winning," International Institute for Analytics (2014:2), pp.1-28.

Hayes, A. F. (2013). Introduction to mediation, moderation, and conditional process analysis. New York: The Guilford Press.

Peter Fowler, Rob Duffield, Adam Waterson & Joanna Vaile (2015) Effects of Regular Away Travel on Training Loads, Recovery, and Injury Rates in Professional Australian Soccer Players, International Journal of Sports Physiology and Performance, 2015, 10, 546 -552, DOI: https://doi.org/10.1123/ijspp.2014-0266

Pollard, R. (1986) Home advantage in soccer: A retrospective analysis, Journal of Sports Sciences, 4(3): 237-248.

Pollard, R. & Pollard, G. (2005) Long-term trends in home advantage in professional team sports in North America and England (1876 – 2003), Journal of Sports Sciences, 23(4): 337-350.

Sams, ML, Wagle JP, Sato K, DeWeese BH, Sayers AL, Stone MH. Using the Session Rating of Perceived Exertion to Quantify Training Load in a Men's College Soccer Team. J Strength Cond Res. 2020 Oct;34(10):2793-2799. doi: 10.1519/JSC.0000000000003793. PMID: 32868677.

Scheck, Andrew W., M.S. Evaluation of the Validity of the Fit For 90 Subjective Training Load and Wellness Measures. (2017)

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological Methodology 1982* (pp. 290-312). Washington DC: American Sociological Association.

Ravé Guillaume, Granacher Urs, Boullosa Daniel, Hackney Anthony C. & Zouhal Hassane (2020) How to Use Global Positioning Systems (GPS) Data to Monitor Training Load in the "Real World" of Elite Soccer, Frontiers in Physiology https://www.frontiersin.org/article/10.3389/fphys.2020.00944 11 ,1664-042X

Taisuke Kinugasa (2013) The Application of Single-Case Research Designs to Study Elite Athletes' Conditioning: An Update, Journal of Applied Sport Psychology, 25:1, 157-166, DOI: 10.1080/10413200.2012.709578

Vasilis Armatas & Richard Pollard (2014) Home advantage in Greek football, European Journal of Sport Science, 14:2, 116-122, DOI: 10.1080/17461391.2012.736537

Appendices

Category	Variable	Team (Adj R^2 = .421) Beta Sig.		Defenders (Adj R^2 = .455) Beta Sig.		Midfielders (Adj R^2 = .252) Beta Sig.		Forwards (Adj R^2 = .027) Beta Sig.	
		Beta	p-value	Beta	p-value	Beta	p-value	Beta	p-value
Internal Load	sRPE	-.602**	.026	-.762**	.013	-.261	.406	-1.49**	.030
External Load	Distance in Speed Zone 5	.491**	.042	.834**	.020	.138	.601	-.057	.855
External Load	Distance in Deceleration Zone 1	2.161	.233	.462	.851	2.39	.198	5.55**	.026
Wellness	Stress	.398**	.011	.500**	.004	.093	.647	.101	.712
Wellness	Sleep Quality	.516**	.032	.398**	.016	-.172	548	.185	.612

Table 1 - Regression Analysis Results

	Home		**Away**		**Difference**		**Effects**	
Variable	Mean	SD	Mean	SD	Mean	SD	Cohen's D	Effect Size
sRPE	486.41	131.10	523.42	95.58	(37.01)	35.52	(0.30)	M
Distance in Speed Zone 5	0.01	0.01	0.01	0.01	0.00	0.00	0.12	L
Distance in Deceleration Zone 1	0.53	0.18	0.48	0.13	0.04	0.04	N/A	N/A
Stress	1.17	0.55	1.23	0.46	(0.07)	0.10	(0.01)	L
Sleep Quality	1.57	0.41	1.50	0.56	0.07	(0.15)	0.31	M

Table 2 - T-Test Values and Effect Sizes

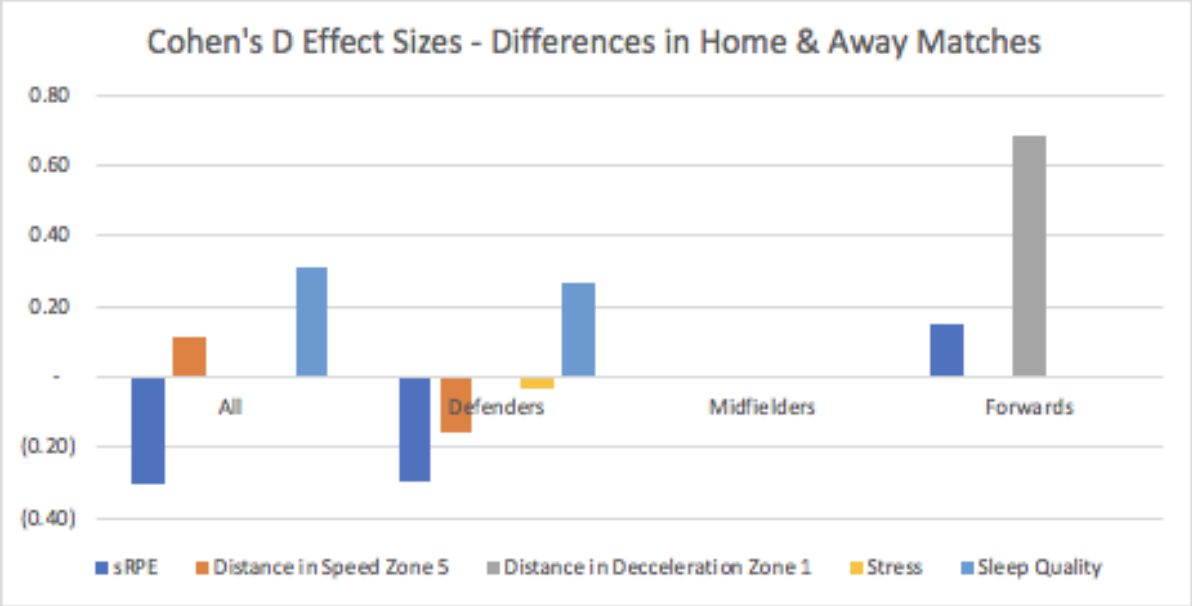


Figure 1 - Effect Sizes of Differences in Means at Position Level

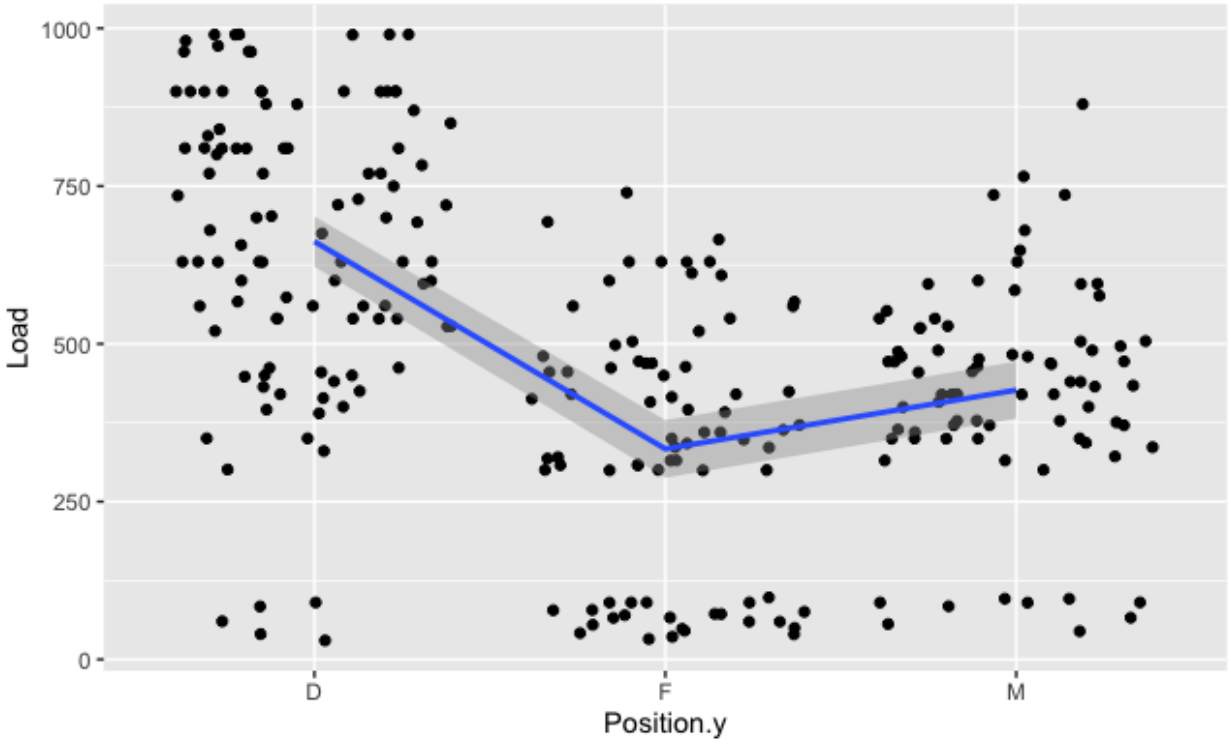


Figure 2 – Player Load by position

No.	Variable	Category
1	Load (sRPE)	Internal Load
2	Duration	External Load
3	Distance_(miles)	External Load
4	Sprint_Distance_(yards)	External Load
5	Top_Speed_(mph)	External Load
6	Distance_Per_Min_(yd/min)	External Load
7	Distance_in_Speed_Zone_1__(miles)	External Load
8	Distance_in_Speed_Zone_2__(miles)	External Load
9	Distance_in_Speed_Zone_3__(miles)	External Load
10	Distance_in_Speed_Zone_4__(miles)	External Load
11	Distance_in_Speed_Zone_5__(miles)	External Load
12	Distance_in_Deceleration_Zones:_0_-_1_m/s/s__(miles)	External Load
13	Distance_in_Deceleration_Zones:_1_-_2_m/s/s__(miles)	External Load
14	Distance_in_Deceleration_Zones:_2_-_3_m/s/s__(miles)	External Load
15	Distance_in_Deceleration_Zones:_3_-_4_m/s/s__(miles)	External Load
16	Distance_in_Deceleration_Zones:_>_4_m/s/s__(miles)	External Load
17	Distance_in_Acceleration_Zones:_0_-_1_m/s/s__(miles)	External Load
18	Distance_in_Acceleration_Zones:_1_-_2_m/s/s__(miles)	External Load
19	Distance_in_Acceleration_Zones:_2_-_3_m/s/s__(miles)	External Load
20	Distance_in_Acceleration_Zones:_3_-_4_m/s/s__(miles)	External Load
21	Distance_in_Acceleration_Zones:_>_4_m/s/s__(miles)	External Load
22	Fatigue	Wellness
23	Mood	Wellness
24	Soreness	Wellness
25	Stress	Wellness
26	SleepQuality	Wellness
27	SleepHours	Wellness

Table 3 – List of Independent Variables

This study examines privacy, confidentiality and security with EHR systems and investigates patient's perceived security of online medical records, particularly of international patients. The format of this study is as follows. First is a discussion of a relevant literature followed by methodology discussion and test results. The manuscript concludes with results, limitations, and future research.

Classification of Hunting-Stressed Wolf Populations Using Machine Learning

John C. Stewart
stewartj@rmu.edu

G. Alan Davis
davis@rmu.edu

Diane Igoche
igoche@rmu.edu

Department of Computer Information Systems
Robert Morris University
Moon Township, PA 15108 USA

Abstract

The preservation of Wolf populations in North America has been controversial for hundreds of years. The preservation of ecosystems or the reintroduction of wolf populations in areas to redress the ecological balance has taken place in recent decades. In other areas, wolves are hunted in an effort to manage them. Previous studies have identified physiological characteristics as an indicator of higher stress levels in individual wolf subjects in heavily hunted populations. This stress impacts reproduction, social structure and pack dynamics. The current study supports a prior study that used statistics to show elevated stress levels in hunted wolf populations. Using machine learning (k-nearest neighbor), we were able to classify individual wolf subjects as belonging to hunting-based stressed populations based on physiological data with high accuracy.

Keywords: Machine Learning, data mining, k-NN, physiological indicators, classification

1. INTRODUCTION/LITERATURE REVIEW

Human originated mortality of predator populations (i.e., hunting) has been documented as having a myriad of additional negative impacts on the affected population (Coltman, 2003, Darimont 2009). Hunting traditionally has the goal of selecting the strongest and fittest, thus impacting reproduction by reducing breeding of the healthiest members of the populations. Studies of trophy hunting of rams determined the effects of selection, placed an emphasis on harvesting of trophy rams of heavier weight and larger horn size (Festa-Bianchet, et al. 2004, Coltman, et al., 2003). Rams with higher value in terms of breeding were found to be shot at a lower age, eliminating their reproductive value to the populations (Coltman, et al., 2002).

The complicated social structure of wolf populations makes them extremely vulnerable to elevated mortality and a disruption of behavior dynamics that would occur from human intervention (Haber, 1996). While wolves can recover from a moderate decrease in population, ongoing pressures can affect behavior, the components of social structures, and genetic factors. This combination of factors can have potential long-term impacts on group and pack recovery (Rausch 1967; Haber 1996;

Jezdrzejewski et al. 2005; Sidorovich et al. 2007; Rutledge et al.2010, 2012).
Wolf populations that are heavily impacted by hunting predictably produce more female offspring (Sidorovich, et al, 2007). In addition, genetic diversity in wolf populations is affected by intense hunting (Jezdrzejewski et al. 2005). As an example, researchers have found that harvesting of wolves outside protected areas can impact the social dynamics of neighboring populations (Rutledge, et al.,2010). Further, and not surprisingly, wolf pup mortality is a critical factor in the rate of population growth (Rausch, 1967).

While changes in population numbers are easily measured, physiological impacts of hunting have only been documented in a very limited number of studies. Elevated levels of hormones like cortisol are an indicator of increased stress in hunted individual subjects (Bateson and Bradshaw 2007). Additionally, stress can negatively affect the social behavior of the target species population (Gobish, et al. 2008).

Testosterone is vital to male reproduction capability but is also an indicator of behavior. Within the social structure of the population, testosterone may be found to increase when there is an imbalance in that component (Oliveria, 2004).

Several studies have found elevated levels of the hormones cortisol, testosterone, and progesterone in pregnant females, giving an indication of the reproductive activity in the population (Foley, et al. 2001). A few studies have proposed a relationship between female testosterone levels and the social structure of the populations (Albert, et al. 1991 and Bryan, et al., 2013).

All of the negative consequences of hunting leads to the following research question: How does human caused mortality affect wolf populations on the physiological level? Only one study has evaluated hormone levels in wolf populations to determine how human-caused mortality may impact group behavior, reproduction, and social dynamics. (Bryan, et al., 2015). Additional research is needed to accurately assess the effects of hunting on wolf physiology.

2. RESEARCH METHODOLOGY

The current research seeks to determine whether individual wolves can be classified as belonging to a heavily-stressed population due to hunting, or as a member of a population with lower hunting pressure. The criteria for measuring stress will be via the measurement of hormones and reproductive steroids in the wolf's fur. More specifically, this study evaluates the hormone levels of two separate wolf populations in Northern Canada that were originally studied by Bryan, et al. in 2015.

The distinction between these two wolf populations is marked by differences in the level of hunting and the percentage reduction of the population. Wolves in the tundra-taiga area were heavily hunted using snowmobiles and firearms. Taiga is characterized by dense conifers, like spruce and pine. Conversely, tundra regions lack any tree cover. Wolves in the second area (i.e., boreal forests) had a lower level of mortality and were killed predominately by trapping. Boreal forests consist of deciduous and conifer trees, and experience wide-ranging temperatures from lows in winter to highs in summer (Musiani, M. & Paquet, P.C., 2004).

Bryan, et al., (2015), predicted that there would be elevated levels of stress and signs of increased reproduction activity in the heavily hunted tundra-taiga wolves, as evidenced by high rates of hormone production (testosterone, progesterone, and cortisol). The researchers in the 2015 study compared the tundra-taiga wolves to wolves in areas of lower hunting pressure, such as those in the boreal forest (Packard & Mech 1980, 1983; Packard, Mech & Seal 1983; Haber 1996, Bryan, et al., 2015).

Sampling Method
The samples (n=152) were collected in a prior study in Nunavut, Northwest Territories and Alberta, Canada (Musiani, et al., 2007). The samples (See Appendix, populations 1 and 2) consisted of wolf hair samples collected during the winter months. The process of extracting the hormones from the wolf hair, including quality control methodologies, is outlined in the Bryan, et al. study (2015).

Bryan, et al., (2015) used predominantly statistical analysis methods in attempting to differentiate the tundra-taiga wolves from the boreal forest wolves. The researchers used ANOVA and Welch's t-tests to compare the two wolf groups, concluding that wolves from the more heavily hunted populations had increased levels of reproduction and stress related hormones. They also determined that these physiological characteristics are in response to environmental factors, including human-induced mortality (Bryan, et al., 2015). The researchers

did list confounding factors, such as ecological and genetic-based differences that could explain hormonal discrepancies. Also, the higher levels of cortisol in the tundra-taiga wolves could be attributed to extended low levels in the food supply in summer, when wolves must travel farther to catch up with migrating caribou. Finally, the massing of tundra-taiga wolf populations near caribou in summer may cause a mingling of wolves and the inevitable interactions among members of different groups (which could also explain the elevated levels of testosterone). The boreal wolves, conversely, have more traditional territories and stability, leading to fewer intergroup interactions (Walton, et al., 2001, Musiani, et al., 2007).

In order to mitigate the impact of confounding factors, the researchers used a control group of wolves (n=30) from a heavily-hunted population in a boreal forest region (See Appendix, population 3). The hormone samples in the control group showed higher levels of cortisol than in boreal forest populations. The wolves in the control group also had similar levels of cortisol as wolves in the heavily hunted northern tundra-taiga region. Therefore, the study concluded that higher cortisol levels are the result of increased mortality rates, possibly coupled with some habitat related factors (Bryan, et al., 2015).

There are several implications revealed by the differences in hormone levels in the Bryan, et al. study. First, reproduction rates are altered (and the social structure, along with the reproduction rates) when there is no longer a dominant pair (i.e., pack hierarchy), and other pack members are not prevented from breeding. The stability of the social group, characterized by a single litter per pack each year, is threatened (Haber, 1996). Second, physiological effects of the disruption in the social framework, like increased cortisol levels, can enhance wolf musculature and release stored energy (Saplosky, 1993). Lastly, high levels of testosterone aid in any challenges an individual wolf may have within the social structure, where strength and dominance of the situation are necessary (Wingfield, et al., 2001).

The current research centers upon the following research questions: Are human-exploited wolf populations more heavily impacted physiologically? Are hormone levels affected to a larger extent in exploited wolf populations, as opposed to those in less stressed populations? And finally, can the type of population an individual wolf may inhabit be identified based upon the measurement of hormone levels as indicators of stress?

Hypotheses Tested

The research hypotheses to be tested in this study are as follows:

H1: Individual wolves can be classified as belonging to a heavily exploited population based on hormone levels.

H2: Machine learning classification can be used to support the results obtained by Bryan, et al. (2015) that human-caused mortality may impact group behavior, reproduction, and social dynamics, and populations as determined by the hormone levels in affected wolves. That is, wolves can accurately be classified into one of two groups: those with high levels of hunting-induced stress, and those with less stress.

The objective of the current study is to determine whether the physiological consequences of hunting (as determined by levels of stress and reproductive hormones in hair, an indicator of elevated endocrine activity), can be used to classify wolves as belonging to a highly-stressed group or a less-stressed group.

To test these hypotheses the current study used data previously analyzed by Bryan, et al. (2015) and k-Nearest Neighbor as the classification methodology to determine wolf membership in heavily stressed versus low stressed populations, based on hormone levels. The 2015 dataset included subject wolves from two separate areas and environments. The dataset contained 45 wolves from a lightly-hunted group in a northern boreal forest, and 103 wolves from a heavily-hunted Tundra-taiga forest area.

All samples were taken as part of a prior study (Musiani, et al., 2007). The samples consisted of hair from the wolf subjects. Cortisol, testosterone, and progesterone (females) levels were measured in each hair sample. The data, listing area, gender, and levels of the three hormones can be found in the Appendix.

Machine Learning Algorithm Used

k-Nearest Neighbor (k-NN) was used to compare cortisol and testosterone levels in the different populations and to determine the accuracy in predicting each population, based on its hormone levels. Bryan et al., (2015) determined that higher levels of cortisol and testosterone were found in the tundra-taiga wolves and concluded that this higher level may be an indicator of social instability. The current

study also used k-NN to compare progesterone levels in the female wolves in the two populations.

Due to the lower numbers of northern boreal forest wolves, stratified sampling was used. In addition, the data were partitioned into 70% for training and 30% for testing. A k-NN algorithm was applied to the data using the knn() function from the class package in R and RStudio. The confusionMatrix() function from the caret package was used to determine accuracy of the classification and sensitivity, specificity, Kappa, and the No Information Rate.

3. RESULTS AND DISCUSSION

The highest accuracy in predicting group membership of the wolves was 86.96% with k=3 (as shown in Table 1). The true positive rate was 100% and the false positive rate was 83%, which supports the validity of the model. The No Information Rate is higher than desired, at an elevated 78%. The Kappa statistic (a measurement of the agreement between accuracy and random chance) was 68%, which indicates moderate agreement.

Table 1. Results of Classification of Wolf Subjects based on Cortisol and Testosterone Levels (k=3)

Measurement	Value
Accuracy	.8697
Sensitivity	1.00
Specificity	0.833
Kappa	0.6849
No Information Rate	0.7826

The current study also measured the difference in progesterone levels between the female wolves in the taiga-tundra and the northern boreal forest and classified them using k-NN (Table 2). Along with k-NN, a similar sampling and data partitioning method was used to preprocess the data. After preprocessing, it was determined that k=14 had the highest accuracy in predicting classification at 0.8333, and with a sensitivity and specificity at 0.7143 and 0.8824, respectively. The No Information Rate was 0.7083, indicating the model has some validity in classification. The Kappa was 0.5966 showing, on the low end, moderate agreement between random and model accuracy.

Table 2. Results of Classification of Female Wolves Based on Progesterone Levels (k=14)

Measurement	Value
Accuracy	.8333
Sensitivity	0.7143
Specificity	0.8824
Kappa	0.5966
No Information Rate	0.7083

4. CONCLUSIONS

Past research on this topic has proposed that elevated levels of the hormones cortisol, testosterone, and progesterone in taiga-tundra wolves are explained by the synergistic effects of hunting pressures, the habitat, or sampling (Bryan, et al., 2015). In the Bryan, et al., study, the researchers compared cortisol levels in the taiga-tundra wolves to those of a control group of 30 wolf subjects (i.e., Little Smokey wolves) in a heavily hunted boreal forest area in an effort to explain the differences in habitat and ecosystem characteristics. The results of this study showed statistically higher cortisol levels in both the Little Smokey and taiga-tundra wolves, compared to the northern boreal forest wolves.

The current study used the k-NN classification algorithm to show that individual wolves can be classified as belonging to heavily hunting-pressured groups based on cortisol and testosterone levels. This classification was also shown to be at a highly-accurate level. The current study also concluded that classification of female wolves (using the k-NN classifier) is possible with a favorable accuracy, based on the females' levels of progesterone. Our results support the findings of Bryan, et al., (2015) that showed statistically-significant differences in hormone levels between taiga-tundra and boreal forest wolf populations (i.e., heavily hunted vs. lightly trapped populations). Our findings support our hypothesis that individual wolves can be classified as belonging to a heavily exploited population based on hormone levels. Additionally, k-nn, a machine learning methodology can be used as a classification mechanism for this purpose.

Prior studies have concluded that the potential ramifications of heavy human-caused mortality

in wolves are substantive chronic stress, and negative alterations in reproduction and breeding practices. These negative effects on breeding, compared with non-distressed populations are not known. However, predictable genetic outcomes like in-breeding, lack of diversity, increased disease, as well as an elevated danger of population extinction are potential long-term effects of heavy hunting (Leonard, et al., 2005).

If a link exists between stress levels in wolf populations and human-based hunting, then aside from the impact on wolf populations, the effects on entire ecosystems can be influenced. Wolves are recognized as a keystone species in their natural habitat (Boyce, 2018; Ripple and Beschta, 2012). Therefore, their absence or minimization can have far reaching impacts on entire ecosystems.

Limitations of Study

In this study, we did not account for differences in male and female subjects in the analysis of cortisol and testosterone levels. This difference in levels between wolf sexes can be evaluated in a later study. It should also be noted that the sample size in this study was relatively small, particularly with the northern boreal forest wolves (i.e., n = 45). However, the research was unfortunately limited by the amount of available data. Additionally, only one machine learning algorithm for classification (i.e., the k-NN classifier). Various machine-learning techniques and models could be employed in future studies. These additional techniques could be used to determine whether wolves can be more accurately classified based hormone levels as indicators of human-caused stress.

5. REFERENCES

Albert, D., Jonik, R. & Walsh, M. (1992) Hormone-dependent aggression in male and female rats: experiential, hormonal, and neural foundations. Neuroscience & Biobehavioral Reviews, 16, 177–192.

Bateson, P. & Bradshaw, E.L. (2007) Physiological effects of hunting red deer (Cervus elaphus). Proceedings of the Royal Society of London. Series B: Biological Sciences, 264, 1707–1714.

Boyce, Mark S. (2018) Wolves for Yellowstone: dynamics in time and space, Journal of Mammalogy, 99, 5, 1021–1031.

Bryan, Heather M. et al. (2015), Heavily hunted wolves have higher stress and reproductive steroids than wolves with lower hunting pressure, Functional Ecology, Article-journal, https://doi.org/10.1111/1365-2435.12354

Bryan, H.M., Darimont, C.T., Paquet, P.C., Wynne-Edwards, K.E. & Smits, J.E. (2013b) Stress and reproductive hormones in grizzly bears reflect nutritional benefits and social consequences of a salmon foraging niche. PLoS ONE, 8, e80537.

Coltman, D.W., O'Donoghue, P., Jorgenson, J.T., Hogg, J.T., Strobeck, C. & Festa-Bianchet, M. (2003) Undesirable evolutionary consequences of trophy hunting. Nature, 426, 655–658.

Coltman, D. W., Festa-Bianchet, M., Jorgenson, J. T. & Strobeck, C. (2002) Age-dependent sexual selection in bighorn rams. Proc. R. Soc. Lond. B 269, 165–172.

Darimont, C.T., Carlson, S.M., Kinnison, M.T., Paquet, P.C., Reimchen, T.E. & Wilmers, C.C. (2009) Human predators outpace other agents of trait change in the wild. Proceedings of the National Academy of Sciences, 106, 952–954.

Festa-Bianchet, M., Coltman, D. W., Turelli, L. & Jorgenson, J. T. Relative allocation to horn and body growth in bighorn rams varies with resource availability. Behav. Ecol. (in the press)

Foley, C.A.H., Papageorge, S. & Wasser, S.K. (2001) Noninvasive stress and reproductive measures of social and ecological pressures in free ranging African elephants. Conservation Biology. 15.1134-1142.

Gobush, K.S., Mutayoba, B.M. & Wasser, S.K. (2008) Long term impacts of poaching on relatedness, stress physiology, and reproductive output of adult female African elephants. Conservation Biology, 22, 1590–1599.

Haber, G.C. (1996) Biological, conservation, and ethical implications of exploiting and controlling wolves. Conservation Biology, 10, 1068–1081.

Jezdrzejewski, W., Branicki, W., Veit, C., Me€ Augorac, I., Pilot, M.g., Bunevich, A. et al. (2005) Genetic diversity and relatedness within packs in an intensely hunted

population of wolves Canis lupus. Acta Theriologica, 50, 3–22.

Leonard, J.A., Vilà, C. & Wayne, R.K. (2005) Legacy lost: genetic variability and population size of extirpated US grey wolves (Canis lupus). Molecular Ecology, 14, 9– 17.

Musiani, M. & Paquet, P.C. (2004) The practices of wolf persecution, protection, and restoration in Canada and the United States. BioScience, 54, 50–60.

Musiani, M., Leonard, J.A., Cluff, H.D., Gates, C.C., Mariani, S., Paquet, P.C. et al. (2007) Differentiation of tundra/taiga and boreal coniferous forest wolves: genetics, coat colour and association with migratory caribou. Molecular Ecology, 16, 4149–4170.

Oliveira, R.F. (2004) Social modulation of androgens in vertebrates: mechanisms and function. Advances in the Study of Behavior, 34, 165–239.

Packard, J.M. & Mech, L.D. (1980) Population regulation in wolves. Biosocial Mechanisms of Population Regulation (eds M.N. Cohen, R.S. Malpass & H.G. Klein), pp. 135–150. Yale University Press, New Haven, Connecticut, USA.

Packard, J.M. & Mech, L.D. (1983) Population regulation in wolves. Symposiumon Natural Regulation of Wildlife Populations. Proceedings 14 (eds F.L. Bunnell, D.S. Eastman & J.M. Peek), pp. 151–174. University of Idaho, Moscow, Idaho, USA.

Packard, J.M., Mech, L.D. & Seal, U.S. (1983) Social influences on reproduction in wolves. Wolves of Canada and Alaska (ed. L.N. Carbyn), pp. 78–85. Canadian Wildlife Service Report Series 45, Ottawa, Ontario, Canada.

R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rausch, R.A. (1967) Some aspects of the population ecology of wolves, Alaska. American Zoologist, 7, 253–265.

Ripple, William & Beschta, Robert. (2012). Trophic cascades in Yellowstone: The first 15 years after wolf reintroduction. Biological Conservation. 145. 205–213.

Russell, E., Koren, G., Rieder, M. & Van Uum, S.H. (2014) The detection of cortisol in human sweat: implications for measurement of cortisol in hair. Therapeutic Drug Monitoring, 36, 30–34.

Rutledge, L.Y., Patterson, B.R., Mills, K.J., Loveless, K.M., Murray, D.L. & White, B.N. (2010) Protection from harvesting restores the natural social structure of eastern wolf packs. Biological Conservation, 143, 332–339.

Rutledge, L.Y., White, B.N., Row, J.R. & Patterson, B.R. (2012) Intense harvesting of eastern wolves facilitated hybridization with coyotes. Ecology and Evolution, 2, 19–33.

Sapolsky, R.M. (1993) The physiology of dominance in stable versus unstable social hierarchies. *Primate Social Conflict* (eds W.A. Mason & S.P. Mendoza), pp. 171– 204. State University of New York Press, Albany, New York, USA.

Sidorovich, V., Stolyarov, V., Vorobei, N., Ivanova, N. & Jedrzejewska, B. (2007) Litter size, sex ratio, and age structure of gray wolves, Canis lupus, in relation to population fluctuations in northern Belarus. Canadian Journal of Zoology, 85, 295–300.

Walton, L.R., Cluff, H.D., Paquet, P.C. & Ramsay, M.A. (2001) Movement patterns of barren-ground wolves in the central Canadian Arctic. Journal of Mammalogy, 82, 867–8

Wingfield, J.C., Lynn, S. & Soma, K.K. (2001) Avoiding the 'costs' of testosterone: ecological bases of hormone-behavior interactions. *Brain Behavior and Evolution*, **57**, 239– 251.

APPENDICES

Appendix A: Wolf Hair Sample Data Collected during Musiani, et al. Study (2007)

Individual	Sex	Population	Colour	Cpgmg	Tpgmg	Ppgmg
1	M	2	W	15.86	5.32	NA
2	F	1	D	20.02	3.71	14.37622
3	F	2	W	9.95	5.3	21.65902
4	F	1	D	25.22	3.71	13.42507
5	M	2	D	21.13	5.34	NA
6	M	2	W	12.48	4.6	NA
7	M	1	W	26.78	4.58	NA
8	M	1	D	15.41	9.27	NA
9	F	1	D	33.87	4.81	19.9127
10	F	2	W	17.29	5.07	34.59806
11	F	1	W	9.43	4.47	25.88548
12	F	1	W	8.84	3.75	15.86882
13	F	1	D	34	4.76	33.08362
14	F	1	D	14.3	6.06	24.82876
15	M	1	D	12.16	5.75	NA
16	M	1	D	22.43	6.15	NA
17	F	2	W	26.26	4.93	25.00037
18	M	2	W	15.8	5.24	NA
19	M	1	W	7.93	4.14	NA
20	M	1	D	4.75	3.34	NA
21	M	2	W	9.17	4.02	NA
22	M	2	W	21.52	4.91	NA
23	M	1	W	10.79	3.91	NA
24	F	2	W	22.69	6.47	21.50033
25	F	2	W	22.17	4.28	31.8274
26	F	2	W	15.34	5.53	34.0765
27	F	1	W	20.48	5.06	20.21606
28	F	1	W	16.19	4.79	18.29115
29	F	1	W	24.05	3.7	21.29735
30	M	2	W	16.45	6.09	NA
31	F	2	W	21.91	4.19	36.40797
32	F	2	W	32.24	6.94	40.92793
33	F	2	W	23.99	5.97	45.9136
34	F	2	W	27.82	7.76	47.2674
35	F	2	W	19.83	6.55	40.93838
36	F	2	W	12.16	4.34	26.65583
37	F	2	W	19.05	6.34	23.90413
38	F	2	D	13.91	4.72	26.36326
39	F	2	D	17.16	9.25	34.64966
40	F	1	W	30.16	6.8	19.61885
41	F	2	W	24.38	5.49	28.12497

42	F	2	D	10.14	3.81	NA
43	M	2	W	18.4	4.98	NA
44	M	2	W	15.21	7.17	NA
45	M	2	W	24.64	15.13	NA
46	M	2	W	22.49	14.45	NA
47	M	2	W	17.42	5.36	NA
48	M	2	W	29.51	9.12	NA
49	M	2	W	27.3	10.75	NA
50	M	2	W	14.04	7.19	NA
51	M	2	W	11.77	5.17	NA
52	M	2	W	23.6	6.97	NA
53	M	2	W	18.14	5.7	NA
54	M	2	W	11.25	4.4	NA
55	F	1	W	14.82	10.81	NA
56	F	2	W	26.39	6.47	24.46521
57	M	2	W	15.15	4.52	NA
58	M	2	W	14.04	6.01	NA
59	M	2	W	21.39	7.36	NA
60	F	2	W	20.02	5.19	31.40929
61	M	2	W	24.64	14.08	NA
62	M	2	W	13.46	4.09	NA
63	M	2	W	18.79	9.74	NA
64	F	2	W	11.77	4.95	21.01472
65	F	2	W	19.96	7.62	28.06955
66	F	2	W	12.68	3.82	27.90797
67	F	2	W	19.76	5.26	27.37918
68	M	2	D	20.35	14.98	NA
69	F	2	W	17.68	5.97	53.28191
70	F	2	W	23.66	6.13	48.53432
71	F	2	W	17.23	7.24	NA
72	F	2	W	25.74	4.88	37.65696
73	F	2	W	19.89	6.35	31.90467
74	F	1	D	14.24	3.95	28.87637
75	M	2	W	17.55	5.02	NA
76	M	2	W	16.32	5.86	NA
77	M	2	W	15.34	5.78	NA
78	F	2	W	11.64	4.87	22.87393
79	M	2	W	13.65	5.04	NA
80	M	2	W	11.57	5.24	NA
81	M	2	W	20.35	5.98	NA
82	M	2	W	8.91	4.58	NA
83	M	2	W	9.1	4.4	NA
84	M	2	D	21.65	7.81	NA
85	M	1	D	10.6	3.65	NA
86	M	1	D	12.35	9.57	NA
87	F	1	D	7.93	3.83	16.77475

88	F	1	D	8	4.26	19.49892
89	F	1	D	7.61	4.24	22.56011
90	M	1	W	11.96	5.62	NA
91	M	1	D	14.82	5.35	NA
92	F	1	W	14.43	5.08	34.81566
93	F	1	D	19.57	6.81	16.67624
94	F	1	W	12.55	3.25	13.19328
95	F	1	D	12.61	3.54	13.62372
96	F	1	D	10.21	4.49	18.52082
97	M	1	D	15.99	5.82	NA
98	F	1	D	32.24	4.8	25.20981
99	M	1	D	15.41	5.68	NA
100	M	1	D	13.98	5.45	NA
101	M	1	D	16.32	6.65	NA
102	M	1	D	6.37	3.31	NA
103	M	1	W	8.19	3.81	NA
104	M	1	W	12.29	3.95	NA
105	F	2	W	12.16	4.37	13.17322
106	F	2	W	16.19	4.43	26.32807
107	F	2	W	11.83	3.48	16.40101
108	F	2	W	10.47	3.9	17.56024
109	F	2	W	21.13	5.09	29.29508
110	F	2	W	18.59	4.49	21.51784
111	F	2	W	12.09	3.96	28.49073
112	F	2	W	13	3.83	30.98607
113	F	2	W	12.09	4.65	28.62749
114	F	2	W	13.26	4.48	25.66584
115	F	2	W	12.03	4.32	19.28812
116	F	2	W	17.36	5.01	30.00925
117	F	2	W	18.14	3.56	12.7591
118	F	2	W	15.93	4.65	22.72246
119	F	2	W	12.29	5.01	23.24402
120	F	2	W	17.42	4.38	18.35924
121	F	2	W	13.2	5.3	18.88097
122	F	2	W	14.5	5.01	21.06504
123	F	2	D	11.44	4.04	16.154
124	M	2	D	11.57	5.68	NA
125	M	2	W	15.28	3.9	NA
126	M	2	W	13.46	5.1	NA
127	M	2	W	13.2	4.76	NA
128	M	2	W	11.25	4.89	NA
129	M	2	W	16.58	7.54	NA
130	M	2	W	13.2	5.07	NA
131	M	2	W	14.04	5.65	NA
132	M	2	W	17.03	5.81	NA
133	M	2	W	17.81	4.88	NA

134	M	2	W	12.48	4.86	NA
135	M	2	W	11.44	4.34	NA
136	M	2	W	40.43	9.13	NA
137	M	2	D	14.3	4.53	NA
138	M	2	W	14.89	4.32	NA
139	M	2	W	16.77	4.4	NA
140	M	2	D	9.95	4.31	NA
141	M	2	W	10.34	4.36	NA
142	M	2	W	20.54	8.06	NA
143	F	1	W	12.81	6.25	26.73429
144	F	1	W	16.51	4.62	28.10653
145	M	1	D	11.12	6.71	NA
146	M	1	D	11.64	4.51	NA
147	M	1	W	18.92	7.57	NA
148	M	2	W	19.89	5.35	NA
149	U	3		9.69	4.23	NA
150	U	3		19.37	4.26	NA
151	U	3		19.76	4.56	NA
152	U	3		11.31	7.73	NA
153	U	3		11.25	3.81	NA
154	U	3		13.85	4.28	NA
155	U	3		17.62	4.54	NA
156	U	3		22.82	4.34	NA
157	U	3		18.14	10.33	NA
158	U	3		13.52	8.12	NA
159	U	3		21.58	5.79	NA
160	U	3		8.91	29.74	NA
161	U	3		9.17	3.14	NA
162	U	3		14.17	10.32	NA
163	U	3		12.09	6.7	NA
164	U	3		54.47	61.79	NA
165	U	3		10.4	4.2	NA
166	U	3		50.31	5.48	NA
167	U	3		33.74	9.61	NA
168	U	3		14.76	8.94	NA
169	U	3		22.3	6.16	NA
170	U	3		23.21	10.59	NA
171	U	3		19.24	5.66	NA
172	U	3		13.07	4.4	NA
173	U	3		49.14	6.21	NA
174	U	3		73.19	6.41	NA
175	U	3		37.05	4.75	NA
176	U	3		16.45	7.29	NA
177	U	3		43.81	6.09	NA
178	U	3		14.89	3.53	NA

Appendix B: k-NN Algorithm and Resulting Confusion Matrix Coded in R

```
set.seed(123)
index <- initial_split(WolfData, prop = 0.7, strata =
"Population")

# index <- sample(2, nrow(WolfData), replace=TRUE,
prob=c(0.90, 0.10))

index

trainData <- WolfData[index==1,]

testData <- WolfData[index==2,]

trainData1 <- trainData[-1]
testData1 <- testData[-1]

trainDataLabels <- trainData [, 1]
testDataLabels <- testData [ ,1]


install.packages ("class")    # if necessary
library(class)

set.seed(13876)
WolfDataPred<- knn(train = trainData1, test = testData1, cl =
trainDataLabels, k=3)

##  Evaluating model performance ----

# load the "gmodels" library
install.packages('gmodels')
library(gmodels)

# Create the cross tabulation of predicted vs. actual

CrossTable(x = testDataLabels, y = WolfDataPred,
         prop.chisq=FALSE)

dim(testDataLabels)
dim(WolfDataPred)

install.packages('caret')

#Import required library
library(caret)

confusionMatrix(testDataLabels,WolfDataPred)


i=1                      # declaration to initiate for loop
k.optm=1                 # declaration to initiate for loop
for (i in 1:28){
   knn.mod <-  knn(train=trainData1, test=testData1,
cl=trainDataLabels, k=i)
   k.optm[i] <- 100 * sum(testDataLabels ==
knn.mod)/NROW(testDataLabels)
   k=i
   cat(k,'=',k.optm[i],'\n')       # to print % accuracy
}
```

A Cloud-based System for Scraping Data From Amazon Product Reviews at Scale

Ryan Woodall
Jrw5074@uncw.edu

Douglas Kline
klined@uncw.edu

Ron Vetter
vetterr@uncw.edu
Department of Computer Science

Minoo Modaresnezhad
modarsm@uncw.edu

Congdon School of Supply Chain, Business Analytics,
and Information Systems
University of North Carolina Wilmington
Wilmington, NC 28403

Abstract

Amazon product reviews can provide a rich source of data for natural language processing research. To support a related research project, we built a custom cloud-based system for obtaining Amazon product reviews. A third-party cloud-based scraping service automatically retrieved scraping jobs, then notified Azure Data Factory through an Azure Function. Raw scraping data was then transferred in batches to Azure Data Lake Storage, then custom SQL transformed the data for convenient query in an Azure SQL database. The system was used to obtain 17,962 product reviews and produce data sets in several formats. This paper fully describes the system, and offers lessons learned from the experience.

Keywords: Data Pipeline, Cloud, Amazon Reviews, Big Data, Azure.

1. INTRODUCTION

Modern companies struggle with big data collection and processing, and it has become best practice to accomplish this with data pipelines in the cloud. These systems need to be maintainable, adaptable, repeatable, and scalable. To explore modern cloud-based data-oriented system development, we created a system to gather large numbers of Amazon product reviews.

Amazon reviews are important for researchers exploring natural language processing. Each product has a distinctive dictionary, i.e., the words used in reviews change for each product category and each product. Reviews offer the ability for researchers to evaluate methods across challenging situations such as dictionary,

data quantity, data quality, themes, sentiment, etc.

Several Amazon Review data sets are freely available online. The data set used for McAuley, et al. (2015) and He & McAuley (2016) was primarily gathered for research relating the review text to product images. This data set is quite large covering many categories and products. However, it is somewhat dated (May 1996 - July 2014), and does not include some of the review quality features implemented by Amazon, e.g., verified purchases. A subsequent dataset was collected by Ni, et al. (2019) providing more recent reviews (through 2018), but still lacks the newer review quality indicators, as well as details about the reviewer that can be used to filter fake reviews.

Amazon (2019) offers their own large review data set, which includes whether the review was tied to a verified purchase. This still leaves out many data items that can be used to evaluate the quality and reliability of a review such as reviewer name, reviewer rating, reviewer social media names, etc.

Web scraping of publicly accessible web content has been contested in recent years. The data analytics company HiQ filed a complaint against LinkedIn's practice of restricting access to users' profiles (Katrix & Schaul 2019). HiQ argued that the practice was anti-competitive and violated state and federal laws. LinkedIn argued that web scraping violates the Computer Fraud and Abuse Act (CFAA), which prohibits accessing a protected computer without authorization. The court favored HiQ stating that users, rather than LinkedIn, held copyright of their own data and that users clearly intended for their data to be publicly accessible. Furthermore, giving companies exclusive rights to users' data would create "information monopolies" that harm public interest. This ruling was upheld in 2022 (Tse & Brian 2022).

A few for-pay services exist providing functionality similar to that created in this project. However, the actual data pipeline is opaque with no visibility of the transformations occurring to the data. Amazon itself has recently begun offering a for-pay API for accessing their data. However, API access tokens are given only to developers from vetted companies, and we did not explore this avenue.

Amazon reviews can be used for many natural language processing (NLP) research purposes such as sentiment analysis, bot detection, theme analysis, summarization, and recommender systems. Furthermore, reviews are the primary method for consumers to evaluate products for purchases.

For a related research project (Gokce, et al 2021) we decided to collect a custom data set that would be more current and include the missing items identified above. We used this opportunity to review modern methods for large scale collection of data in the cloud. Rather than filtering the existing massive data sets for the data needed, we would create smaller targeted sets created for our tasks. Another concern we had was that the actual processing steps performed on the raw reviews is unclear, and perhaps not repeatable. Ni, et al. (2019) offers their data in several forms with different levels of "aggressive" removal of reviews for various reasons. We took this opportunity to curate our own data set in a fully auditable, repeatable manner, where every data modification was explicitly described by code.

To fully explore modern data pipeline methods, we established the following goals:

- Cloud-based – the system should entirely reside in the cloud
- Automated – the system should orchestrate tasks without in-progress intervention
- Scalable – the system should be scalable
- Formats – the system should offer the data set(s) in several formats

2. TECHNOLOGIES

In this section we describe the technologies that we used and why they were chosen. Many vendors offer cloud-based services. To limit the overwhelming options and to align with our existing technology skillsets, we chose to use Azure cloud services as much as possible. Generally equivalent services exist on most cloud provider platforms.

We used a third-party cloud-based web scraping service called WebScraper (2020) because of its convenient low-code nature and our prior experience with its desktop-product. Scrape job definitions can be authored in the Chrome web browser and exported in JSON format. Scrape jobs can traverse paginated web pages, drill down and up through pages, and gather related data entities such as product, review, reviewer, etc. Bot-detection-avoidance features are

included such as pauses between page requests, and the use of multiple IP addresses for requests. In the cloud-based service, a full API is provided for programmatic definition of jobs.

Azure Data Lake Storage Gen2 was used to store raw files, and also as the storage area for the Azure SQL instance. This service offers a hierarchical name space, high scalability, metered fees, and can support map-reduce style operations.

Azure Data Factory (ADF) integrates seamlessly with Data Lake Storage and can be used to orchestrate data workflows among various locations and services in the cloud. Our initial intent was to use ADF for the bulk of the data operations, but ultimately, it was mainly used to move data from one location to another.

An Azure SQL relational database instance was used to deliver final data in an easily query-able format. In addition, relational models offer a quite compact representation of the data, reducing storage costs, and improving performance.

Azure Functions were used to trigger events across distributed components of the system. With Azure Functions, a secure HTTPS endpoint could be called with proper credentials, triggering events in a remote system and/or passing data between systems.

We planned to use Azure Key Vault to manage the secrets needed for secure communication between distributed system components. Secrets we planned to keep in the Azure Key Vault included connection strings, credentials (username/password), and API tokens. Ultimately, Azure Key Vault proved unwieldy for our relatively small project and required a full Azure Active Directory Domain and high-level domain credentials. For expedience without sacrificing security, secrets were stored in each linked service defined in Azure Data Factory.

The many cloud services were declaratively defined in Azure Resource Manager (ARM) templates. ARM templates define the desired end state of a collection of services and manage the connections and security concerns among the services. ARM templates are text-based and declarative. With a system's ARM template, a complete replica of a complex distributed set of cloud services can be perfectly replicated. Furthermore, the entire system can be versioned in source control.

Even with the cloud-based services used, there were places where programming code was required. Transact SQL, Microsoft's procedural scripting language, was used to transform data from a staging table to relational tables in the database. Python code was used in the Azure function.

3. SYSTEM DESCRIPTION

In this section, we describe the system as it was ultimately built. The System Overview diagram in Appendix A gives a high-level view of the major components. The Webscraper component was the only non-Azure part. All other components are Azure Data Factory workflows and were housed in a single Azure Resource Group from which an ARM template could be generated, completely defining all components and their interactions.

The system operated as a set of Azure Data Factory workflows:

- Start Jobs (periodic)
- Scrape Data (external)
- Record Completion (episodic)
- Retrieve Reviews (periodic)
- Create Tabular Data (periodic)
- Create Flat Data (periodic)

Each of the above workflows is decoupled from the others, so that work is "pulled" through the system rather than "pushed". The workflows marked as periodic are implemented via timers. Each periodic workflow wakes up at defined intervals and completes any waiting work. Waiting work is recorded in the ScrapeJob and ScrapeJob_Status tables (see the relational schema in Appendix B). The only episodic workflow, Record Completion, is implemented as an Azure Function. This Azure Function is called by the external Webscaper service to indicate that a scraping job has completed. Azure Data Factory was mainly used to move data, record job state, and orchestrate the process.

The Start Jobs workflow looks for new product review scraping jobs and calls the external scraping service to start them. New scrape jobs are entered (by humans or otherwise) in the ProductURLs.csv file in Azure Storage. A scrape job is defined by the url of a product on Amazon. The same scrape definition is used on all products, and can be found in Appendix C. The Start Jobs component is implemented as a periodic (every 15 minutes) azure data factory pipeline. For every URL in the file, a POST is made to the WebScraper service API using secure credentials. Job status information is

stored in the Azure SQL tables and various files in Azure Storage. A single record is inserted into the Product table. Appendix D shows the actual ADF pipeline used for this component. This is a good example of a simple ADF workflow. Similar simple workflows were used throughout the system.

The Scrape Data operation is performed over time by the WebScraper third party service. When it is complete, the WebScraper service POSTs to the Record Completion implemented as an Azure Function exposed as an HTTP endpoint. This message does not transfer the data, but merely indicates that the scrape job has been completed. The Azure Function is implemented in Python, and simply records completion of the task in the SQL ScrapeJob_Status table in the SQL database.

The Retrieve Reviews component is implemented as a periodic azure data factory pipeline similar to the Start Jobs component. For each completed scrape job, it makes an API call to the WebScraper API for the resulting data and metadata and moves the data to Azure Storage. A single product produces a single scrape job, which produces a single set of reviews, in a single file. This component places a potentially large JSON file in Azure Storage and creates related records in the ScrapeJob_status table.

The Create Tabular Data takes as input the JSON file from Azure Storage and inserts many records into the Review and Reviewer tables. The ADF activity diagram can be seen in Appendix E. This component moves a file of reviews into a staging table in the SQL database. The data is inserted into the staging table via an efficient bulk-load operation with no integrity checking. The data is unmodified from the WebScraper component and enters the staging table with all character-based data types. A stored procedure written in Transact-SQL transforms the data and inserts it into the Product, Review, and Reviewer relational tables. Notably, the star-rating is transformed from prose ("one star out of five") to an integer data type.

The Create Tabular Data component was initially implemented using Azure Data Factory. We found this to be slow, inefficient, and costly. The ADF processes appear to be implemented as record-at-a-time operations. ADF was still used to orchestrate the component, with stored procedures used strategically where efficient set-at-a-time operations were possible.

Finally, the Create Flat Data component executes straightforward SQL queries to produce a CSV flat file, as well as a hierarchical JSON file. This is for the convenience of users. Data Analytics professionals normally consume data in these formats. Because the data is in a relational schema with a database, any subset of the data can be readily produced in any format using straightforward SQL.

4. RESULTS & DISCUSSION

To test the system reviews for 15 products across 9 subcategories in the Audio Books & Originals category were collected. The subcategories were:

- Bios & Memoires
- Self-Development
- Literature & Fiction
- Business & Careers
- Science Fiction & Fantasy
- Teen & Young Adult
- Health & Wellness
- Computers & Technology
- Kids

In total 17,962 reviews were collected and processed through the pipeline, producing data in CSV and JSON formats, as well as recording all entity instances in the relational database. Each form of the data was retained, providing a full audit-trail of all changes to the data.

The Webscraper service made requests in three parallel threads emanating from separate IP addresses, with adjustable delays between page requests. The number of reviews per product ranged from 293 to 2690, and scrape times per product ranged from 1.5 to 9.75 hours with a combined scraping time of 67.5 hours. The number of records scraped per hour average approximately 266.

In general, we accomplished the goals set out at the beginning of the project. The system can produce large amounts of Amazon reviews in a variety of formats in an automated manner.

By creating the system, we learned much and will relate some opinions and advice for developers implementing similar systems with these technologies.

Azure Data Factory & Azure Functions

It became clear that Azure Data Factory (ADF) and Azure Functions were not fully mature products. In some cases, the products were changing during the development of this system. Specifically, overnight changes in features broke

the system multiple times during development, requiring rewrites of code. Second, integrations with products and languages were not complete. For example, when using Python to write Azure Functions, there was no direct access to Azure SQL, while there was direct access to CosmosDB.

We found Azure Data Factory jobs difficult and costly to debug, as well as expensive and inefficient. Processing appears to be row-by-agonizing-row, which was unnecessary in our situation. Ultimately, some parts that we intended to write in Azure Data Factory were implemented as set-at-a-time operations in SQL. We found this to be more transparent, easier to write and debug, and ultimately much faster than the ADF pipelines.

Third Party Cost

We chose to use the cloud-based WebScraper service for its low-code approach and our familiarity with the desktop product. However, this service was the majority of the cost for producing reviews. The cost structure does not scale to the level we want. This is not critical of the service; it worked well, as advertised, and required minimal effort.

To continue collecting reviews at scale with reasonable costs, we anticipate custom coding the review collection component. There are several libraries available, including python libraries Beautiful Soup (2021) and Scrapy (2021). We would not have to write a general purpose full-featured cloud-based web scraping tool with all features of WebScraper, but could write code specific to our needs. It appears feasible to implement a custom scraper in one or more virtual machines or containers, or even in an Azure Function. A less elegant approach would involve running the desktop browser-based free version of WebScraper in virtual desktops and collecting results as they are produced.

Cloud-based Considerations

The cloud-based distributed architecture comes with benefits and challenges. A distributed architecture necessitates decoupling and defining explicit interfaces between components, which generally produces a very clear transparent system. However, secure communication becomes onerous compared to a monolithic application. In systems with more components and services, the need for secrets management would be required through a product like Azure Key Vault.

ARM templates provide the ability to persist the exact definition of the web of services in a complex system in a text-based, version-able form. This opens the door to team development and change management that didn't exist pre-cloud. However, each service we used was quite complex in its own right; each imposing its own learning curve. Ultimately, the logic of the entire system was spread across many places. In this cloud-based environment, it is very important to have clear roles for each component, and explicit interface contracts to manage interactions.

5. ACKNOWLEDGEMENTS

We wish to recognize the support of the Congdon School of Supply Chain, Business Analytics, and Information Systems at the University of North Carolina Wilmington.

6. REFERENCES

Amazon Web Services Open Data, 2019. Multilingual Amazon Reviews Corpus https://registry.opendata.aws/amazon-reviews.

Beautiful Soup (2021) https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

Gokce, Y., Kline, D, Vetter, R., Cummings, J. (2021) Automated Text Reduction: Comparison of Reduced Reading List Creation Methods. Annals of the Master of Science in Computer Science and Information Systems at UNC Wilmington, 15(1) paper 2. http://csbapp.uncw.edu/data/mscsis/full.aspx.

He, R., & McAuley, J. (2016, April). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In proceedings of the 25th international conference on world wide web (pp. 507-517).

Katrix, Basileios "Bill", & Schaul, Robert J (2019) Data Scraping Survives! (At Least for Now) Key Takeaways from 9th Circuit Ruling on the HiQ vs LinkedIn case. The National Law Review, 30 September 2019.

McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015, August). Image-based recommendations on styles and substitutes. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval (pp. 43-52).

Ni, J., Li, J., & McAuley, J. (2019, November). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 188-197).

Scrapy (2021) https://scrapy.org/.

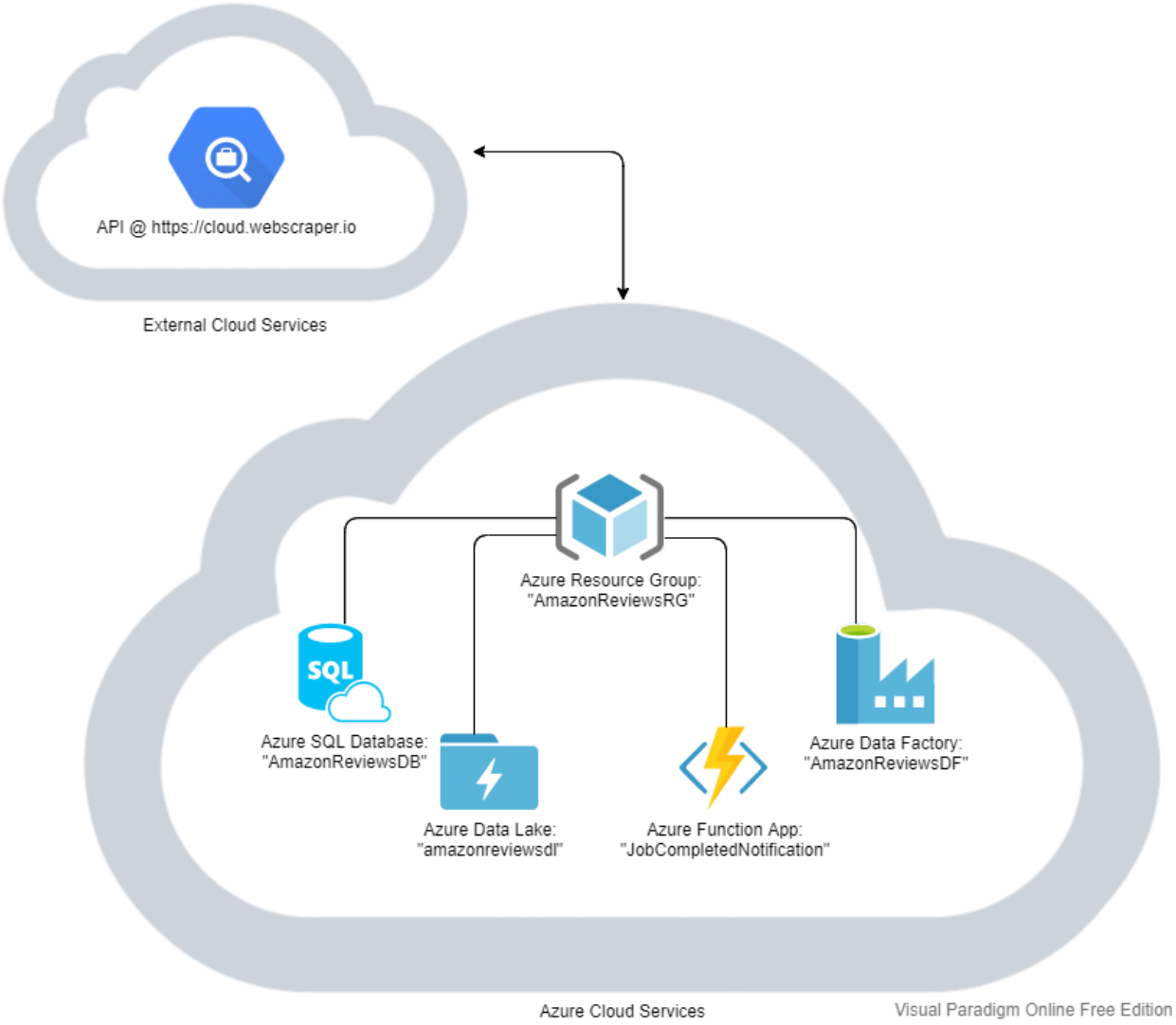
Tse, Shing, & Brian, Kristin (2022) HiQ vs LinkedIn. The National Law Review, 19 April 2022.

Webscraper Documentation. (2020). Retrieved September 5, 2020, from https://webscraper.io/documentation

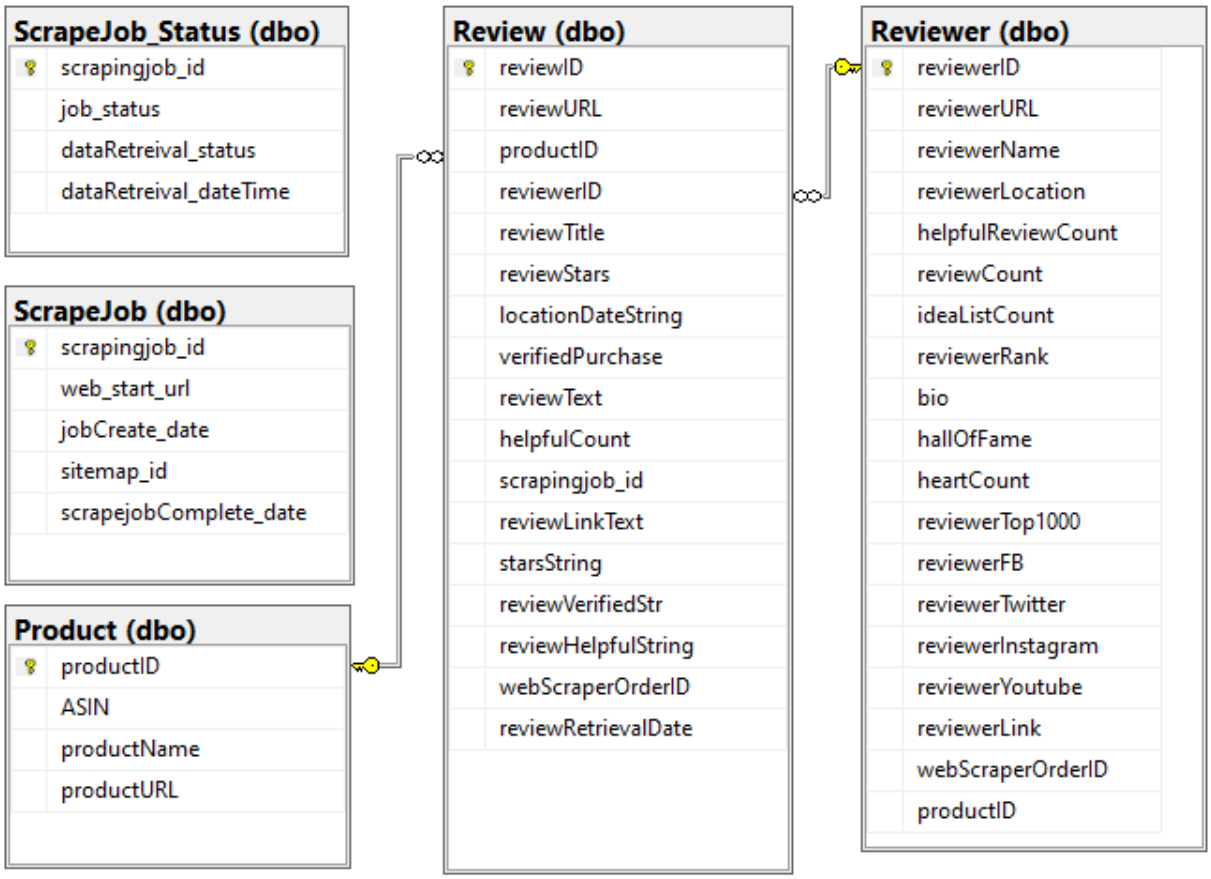
Woodall, R., Kline, D, Vetter, R., Modaresnezhad, M. (2021) A Data Pipeline for Amazon Review Collection and Preparation. Annals of the Master of Science in Computer Science and Information Systems at UNC Wilmington, 15(1) paper 3. http://csbapp.uncw.edu/data/mscsis/full.aspx.

Appendix A

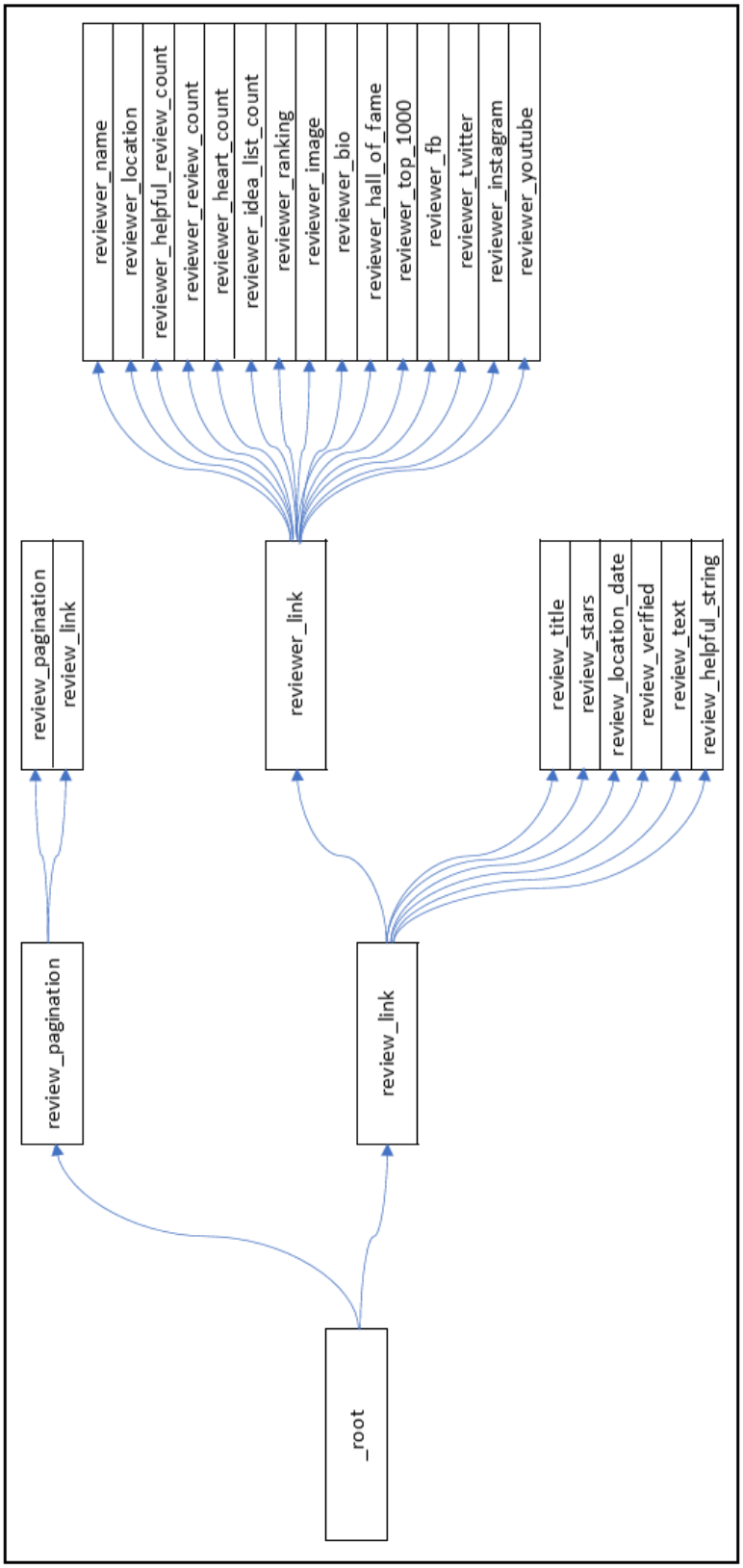
System Overview



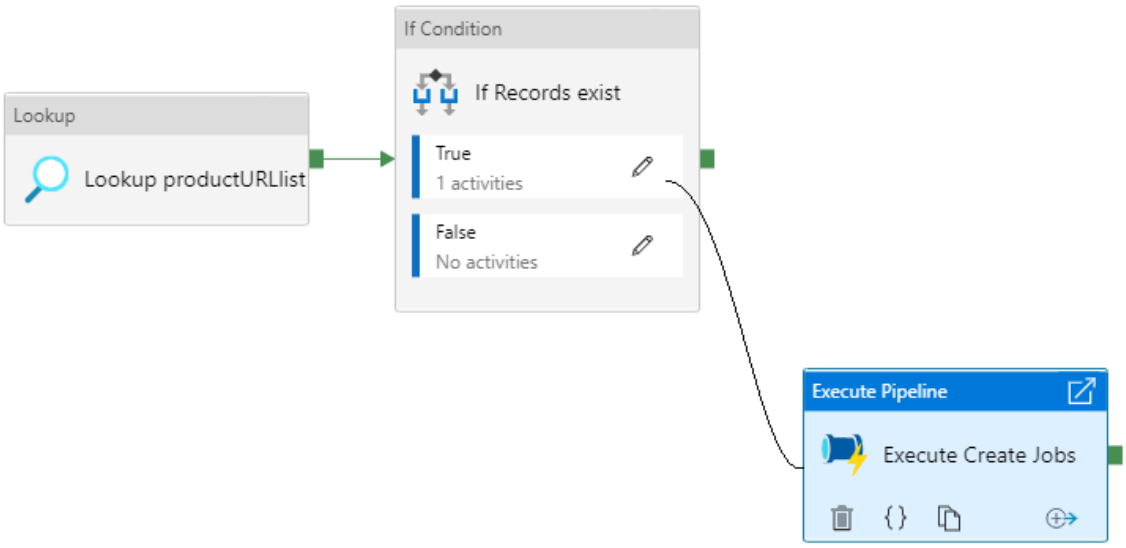
Appendix B


ScrapeJob_Status (dbo)
scrapingjob_id
job_status
dataRetreival_status
dataRetreival_dateTime
ScrapeJob (dbo)
scrapingjob_id
web_start_url
jobCreate_date
sitemap_id
scrapejobComplete_date
Product (dbo)
productID
ASIN
productName
productURL
Review (dbo)
reviewID
reviewURL
productID
reviewerID
reviewTitle
reviewStars
locationDateString
verifiedPurchase
reviewText
helpfulCount
scrapingjob_id
reviewLinkText
starsString
reviewVerifiedStr
reviewHelpfulString
webScraperOrderID
reviewRetrievalDate
Reviewer (dbo)
reviewerID
reviewerURL
reviewerName
reviewerLocation
helpfulReviewCount
reviewCount
ideaListCount
reviewerRank
bio
hallOfFame
heartCount
reviewerTop1000
reviewerFB
reviewerTwitter
reviewerInstagram
reviewerYoutube
reviewerLink
webScraperOrderID
productID

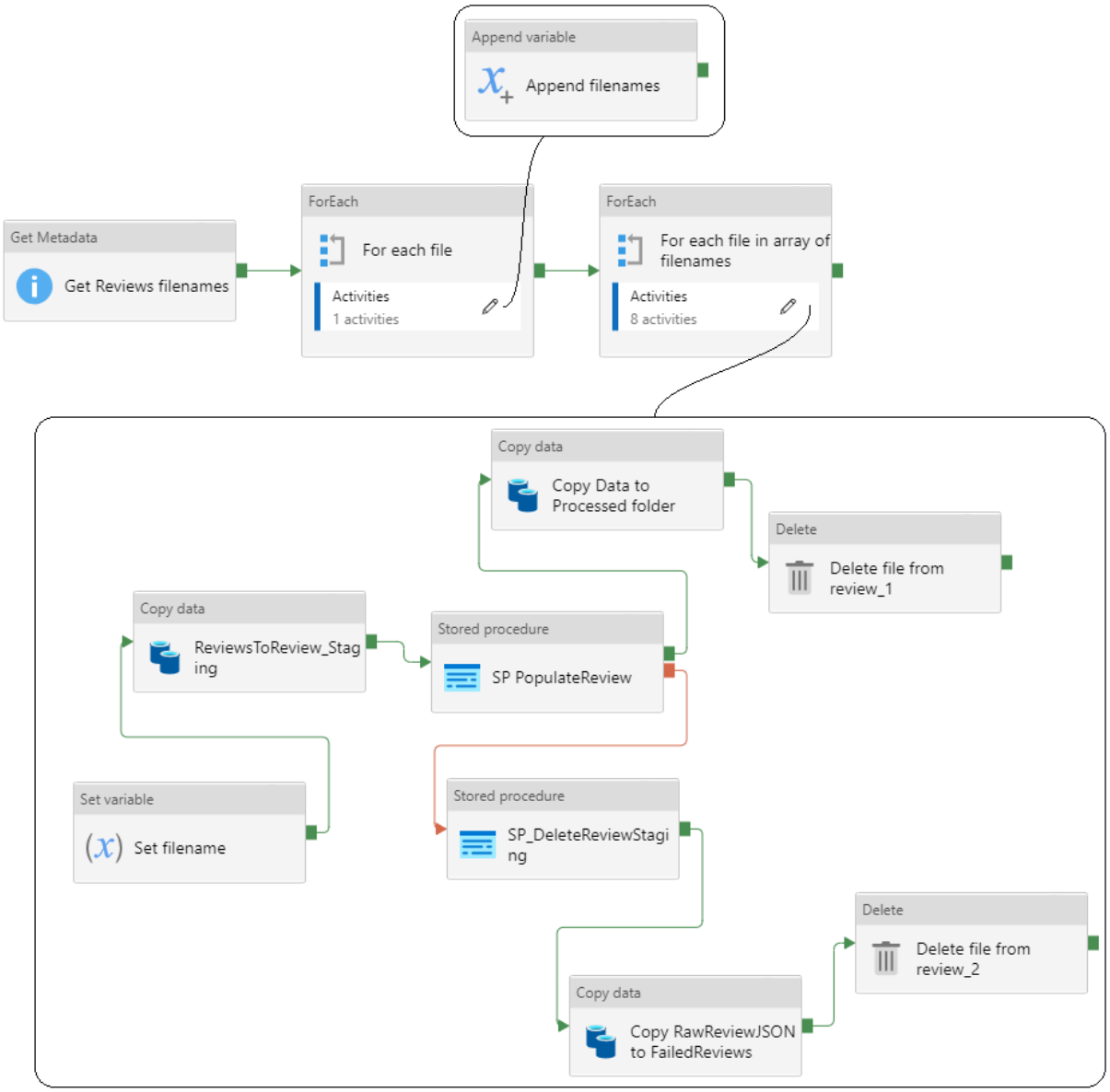
Appendix C


reviewer_name
reviewer_location
reviewer_helpful_review_count
reviewer_review_count
reviewer_heart_count
reviewer_idea_list_count
reviewer_ranking
reviewer_image
reviewer_bio
reviewer_hall_of_fame
reviewer_top_1000
reviewer_fb
reviewer_twitter
reviewer_instagram
reviewer_youtube
review_pagination
review_link
reviewer_link
review_title
review_stars
review_location_date
review_verified
review_text
review_helpful_string
review_pagination
review_link
_root

Appendix D


Lookup
Lookup productURLlist
If Condition
If Records exist
True
1 activities
False
No activities
Execute Pipeline
Execute Create Jobs

Appendix E


Append variable
Append filenames
Get Metadata
Get Reviews filenames
ForEach
For each file
Activities
1 activities
ForEach
For each file in array of filenames
Activities
8 activities
Copy data
Copy Data to Processed folder
Delete
Delete file from review_1
Copy data
ReviewsToReview_Staging
Stored procedure
SP PopulateReview
Set variable
Set filename
Stored procedure
SP_DeleteReviewStaging
Delete
Delete file from review_2
Copy data
Copy RawReviewJSON to FailedReviews

Grounded Theory Investigation into Cognitive Outcomes with Project-Based Learning

Biswadip Ghosh
bghosh@msudenver.edu
Dept. of Computer Information Systems,
Metropolitan State University of Denver,
Denver, Colorado, USA

Abstract

There is increasing use of business analytics (BA) systems in industry to support decision making and process improvement. BA systems provide specialized functions for data collection, cleaning, analysis, query, and reporting. The need for BA skills in the workplace is driving the growth of graduate and undergraduate programs. However, such curriculum presents pedagogical challenges due to the interdisciplinary nature of enterprise BA work and the demand for a broader range of skills by the industry. BA courses need to go beyond emphasizing tool procedural skills and quantitative statistical knowledge. Project based learning (PBL) refers to pedagogy that engages students in educational content that is based on standards and practical business use cases and supports building higher level competencies such as problem solving, critical thinking, collaboration, communication, and innovation. Incorporating PBL in a BA course allows students to experience real world BA projects by working with business end-users. This study collects interview data from the students and participating business users and explores how PBL leverages real-world situational conditions, and group interactions to increase higher level cognitive learning outcomes. The research uses grounded theory to identify relationships among PBL, and group and individual factors on cognitive outcomes.

Keywords: Cognitive outcomes, enterprise business analytics, project-based learning, grounded theory.

1. INTRODUCTION

Organizations are collecting large amounts of data in enterprise applications and then deploying business analytics (BA) systems to utilize these data sets to improve their cross-functional business processes (Elbashir, Collier and Davern, 2008). Enterprise business analytics refers to the use of BA tools that leverage enterprise systems to create and deploy models that span multiple functions (Davenport and Harris, 2007). This is accomplished by utilizing BA applications that can aggregate cross functional datasets extracted from systems such as ERP to create new organization wide capabilities. As opposed to departmental or function-based analytics applications, enterprise analytics has several advantages, such as broad impact across the organization and the ability to yield "one version of the truth" information. The growing adoption of enterprise analytics is also creating an increasing need for BA skills in the workplace and driving the growth of graduate and undergraduate coursework and educational programs at universities (Mills, Chudoba, and Olsen 2016). However, the curriculum of such programs presents pedagogical challenges due to the demand for a broader range of skills by industry (Radovilsky and Hedge, 2022). Paul and MacDonald (2020) compiled and classified a list of six groups of skills that include business knowledge, technical coding and programming, data modeling, and problem solving, in addition to typical quantitative knowledge like data mining and statistical methods. A Delphi study and survey with an industry panel by Cegielski and Jones-Farmer (2016), along with job

content analysis revealed that a business education, together with problem solving and communications skills were in greater demand by industry than mere quantitative knowledge. Yet the primary focus of many current BA educational programs continues to be the coverage of quantitative skills and building BA tool procedural knowledge. Current BA pedagogy remains mostly "hands-on" skill-based, highly procedural, and narrow in scope, and does not allow the typical student to grasp the tight data integration among business functions and the inter-disciplinary nature of BA jobs in industry. This is resulting in a mismatch ("gaps") of skills generated by educational institutions and skills demanded by employers. To resolve these skills gaps, Markov, Braaganza, Taska, Miller and Hughes (2017). recommends the creation of new learning pathways and programs that concentrate on emphasizing business domain knowledge, BA industry practices and processes such as CRISP-DM, and managerial and communications skills.

Radovilsky and Hedge (2022) documents a wide diversity in course content and pedagogy in BA educational programs and finds no consistency in the coverage of the four sets of skills – technical, analytical, business and communications. Their analysis of 121 course syllabi, which was taught over four academic years from 2016 to 2020, shows limited consistency in the courses with regard to pedagogy and content covered. Courses in business analytics continue to emphasize quantitative theory and quantitative methodologies and BA tool procedural skills. However, the feedback from industry suggests that only learning the mechanics of a BA tool in conjunction with quantitative statistical methods is insufficient for students preparing for enterprise BA jobs. It is imperative for educators to expand their approach and integrate these theoretical curricula with project-based assignments to broaden student learning outcomes, particularly higher-level, "real-world" cognitive outcomes such as judgment, critical analysis, confidence and application of BA systems to practical scenarios. Yap and Drye (2018) describes the successful application of practice-oriented projects to introduce theoretical BA content to students in a practical way. Their approach emphasizes the use of real-world data sets and application of relevant technology and methodology to create useable products for end users.

Project based learning (PBL) is a pedagogical approach that successfully blends the formal and informal phases of learning new skills and emphasizes the casual transfer of knowledge among group members (Marcris, 2011; Leidner and Jarvenpaa, 1995). Gupta, Bostrom and Huber (2010) found four categories of pedagogical factors that impact learning outcomes: (i) technology characteristics, (ii) individual motivation, (iii) social influence, and (iv) situational constraints. Each of these are sufficiently represented in group project-based learning programs. Educational outcomes depend on the pedagogy used and the shared insight of the participating students and faculty, who are the stakeholders of the BA curriculum (Bose, 2009). PBL participants learn from each other as well as from the course content by executing the educational program in a practical setting, solving real world projects. Such group based educational programs are also more supportive of the cognitive outcomes necessary for individuals to become successful industry practitioners of BA systems.

Gupta, Bostrom and Huber (2010) also reported the difficulty to assess "real-world" cognitive outcomes during the learning period with existing assessment models, as such measures rely on future job performance. Published BA pedagogy research also does not report any suitable measurement models to make cognitive outcome assessments during the educational program. However, the authenticity of the learning environment created by PBL, which demands students execute genuine workplace tasks, supports the development of a measurement framework to allow self-assessment of learning outcomes, including cognitive outcomes, during the learning process.

Research Goals

The focus of this research is to study the impact of group project-based learning (PBL) on the cognitive outcomes of students of BA courses. PBL programs allow students to work in groups to learn and apply the theoretical concepts collectively with real world business end users. The project activities are supported with genuine real life project scenarios along with interactions with these business users. This study aims to contribute to the body of knowledge by researching an innovate project based learning program and proposing an assessment model to measure the effect of the PBL program on the cognitive outcomes of the participants.

2. GROUNDED THEORY

This study uses qualitative research with interpretative methods based on semi-structured

interviews. Interpretive research is inductive and does not rely on previous literature or prior empirical evidence (Eisenhardt, 1989, Strauss and Corbin, 1990). The objective of grounded theory is to generate constructs and discover relationships among the constructs using qualitative data. Rather than start with a pre-conceived research model and hypotheses to test, grounded theory uses an inductive approach, which is data driven, and through simultaneous data collection and analysis to discover patterns and concepts underlying the phenomena. This methodology places emphasis on abstracting participants' accounts of experiences and events and relating those to existing literature to explain the phenomena (Strauss and Corbin, 1990, Suddaby, 2006). In this approach data is analyzed by comparing incidents and connecting emerging concepts in concert with theoretical research. This recursive activity employs theoretical sampling whereby additional data collection builds around the occurring findings and narrowing the scope of the study until theoretical saturation is reached where no new data changes the emergent constructs. Moreover, this type of methodology explains process, `how' research questions, and context, and provides detailed information for deducing constructs for theory generation and elaboration.

Proposed PBL Pedagogy

The essential elements of the proposed PBL pedagogy are: (1) including significant content that is relevant and derived from standards and concepts at the heart of practical business use cases, (2) building higher-level competencies such as problem solving, critical thinking, collaboration, communication, and creativity/innovation and (3) engaging the students in an extended, rigorous process of asking questions, using resources, and developing answers. These characteristics of PBL are supported by providing open-ended project scenario(s) that students understand and find intriguing. These scenarios generate interest and curiosity among the students and produce a need to gain knowledge, understand concepts, and apply skills to create outcomes that are applicable to their jobs. Mimicking the real work environment is critical and is achieved by allowing the students to make choices about the BA information products to be created for their assigned course project. The PBL project allows them to give and receive feedback on the quality of their work, leading them to make revisions and motivating further inquiry.

The Project Based Learning (PBL) pedagogy used in this study also incorporates several learning elements, including: (1) the use of "messy" datasets, (2) interactions with actual client business users to allow the students to build systems to target these real organizational users, and (3) an iterative approach for the project development using periodic reviews with the business users. The PBL program was adopted inside a senior experience business analytics course for undergraduate IS majors.

	PBL Topics	**Practical Group Work**
1	Read/Analyze a BA project Case Study to discern the nuances of a "industrial" sized BA project.	Analyze/Discuss Case Study to identify the project stakeholders, phases, challenges faced, and strategies. Enumerate project activities and efforts needed in phases of a typical BA Project
2	CRISP-DM Methodology Data Visualization	Use Visualization Tool on a real-world data set to discover and understand data relationships.
3	Business Use Case & Systems Requirement Analysis; Scope definition of assigned project	Collect and analyze the project requirements and use cases from business user – create wireframe prototype of the application user interface
4	Learn Key Performance Indicators (KPI) Information data Lifecycle and Data Quality	BA Tool feature selection and learning (tool procedural) Identify cross functional KPI's for the end user use cases
5	Data Modelling Data prep and model creation	Build logic-based data model to support the end user reequipments and use cases collected
6	Data mgmt. and storage tools (ETL); Predictive Analytics & Data Relationship	Identify Input data and sources Data Storage Design Build and test prelim project & user reviews
7	BA project feasibility analysis; Unstructured data analysis	Add "what-if analysis" to BA project Feasibility Analysis, Build, test, deploy final project with Users
8	Project Reports, Presentations, Documentation	Project/User doc and Project Presentations, Project Retrospective

Table 1: PBL Weekly Topics & Assignments

The PBL program was administered as a practical summative term project assignment over the second 8 weeks of a 16-week semester. The detailed schedule of the PBL learning topics and group assignment is listed in Table 1.

The 8-week PBL assignment was embedded in a semester long face-to-face BA course with weekly lectures on theory and in class and outside class assignments with a leading ERP vendor's BA tools. The theory was delivered with lectures on various topics such as business analytics models and case studies, requirement gathering and documentation, dashboard design, data modelling data management, and project management. There were 11 students in the course, and they were provided 2.5 hours/week of instruction about analytics methods, principles, and case studies as part of the theoretical portion of the course. Students were divided into groups of 3 and given access to a BA industry consultant and business end users. PBL required the students interact with real business users to define the actual project assignment in detail, including user scenarios. A large data set was extracted from the client company's ERP system and provided to the students to work with. The data set contained financial, production, materials, human resources, and operational maintenance, training, and safety data. The students were required to learn and use the CRISP-DM (www.crisp-dm.eu) methodology to define and implement a business analytics project that the business end users would use.

3. DATA COLLECTION AND ANALYSIS

The 11 students were divided into four small groups (3 members in each, except one with 2 members) and each assigned to a business end user. Their first objective was to thoroughly understand, from a business perspective, what their assigned business end user really wanted to accomplish with the BA project. The participants documented the business use cases and made decisions on how to utilize the data set to support the KPI's deemed necessary by the business user. The groups then designed and built BA dashboards that displayed the functional variables and relationships (in the data). They designed quantitative KPI models to add "what-if" scenarios with the BA tools. Contacts in the client company and the BA consultant were available during the entire 8-week duration to answer questions and review project scope and designs.

As the objective of this research is to generate theory, which explains how higher-level cognitive outcomes are enhanced with PBL, a total of 16 interviews were conducted with multiple stakeholders after the 8th week of the PBL learning pedagogy (Table 2). A pilot interview was conducted with one of the business users and a student, followed by 3 subsequent stages of interviews. In all, eleven students, 4 business users and one IT industry consultant, were interviewed over four weeks. Concurrently, the relevant published literature was searched and analyzed. The generalizability of the findings of a qualitative study are strengthened by including more than one participant's perspective and incorporating theoretical perspectives at multiple levels of analysis into the discussion. A grounded theory model of measuring the impact of PBL on cognitive learning outcomes is a product of this research study. Although the interviews were open-ended, the following questions guided the theory building:

1. What types of challenges did you face in performing the project methodology including requirements analysis?
2. What knowledge needed to be shared to define the BA system with the business end-users (students)?
3. How were the project activities facilitated by group members and knowledge shared between students and business users?
4. What were the educational benefits and drawbacks of incorporating a practical project with "messy" data and interactions with real business users?

Interviewee's role	Number of interviews	Hours
Undergraduate Student of IS	11	5.5
Business User	4	4.0
BA Industry Consultant	1	1.0

Table 2. Interviewees' roles and numbers

Data Analysis

The interview scripts were coded using nVivo software. Each interview was transcribed to a separate document and the documents uploaded into the tool. This tool has a sophisticated search engine and features that enable saving search terms and outputting search results for specific terms. Coding in grounded theory has three

stages: open coding, selective coding and theoretical coding. In the open coding phase, the transcripts from the interviews were listed as quotes and analyzed line by line to identify concepts. The key concepts emerged from open coding, and a technique was used for categorizing interview data allowing the major concepts to be identified along with their properties (Table 3). Subsequent theoretical coding was used to relate concepts to other concepts, establishing a model of the perceived phenomena. Analysis continued until no further concepts emerged - the point at which theoretical saturation is reached.

4. RESULTS

The grounded theory approach culminated in a model that sheds light on a fresh theoretical perspective of enhancing the higher-level cognitive outcomes in a BA course with PBL pedagogy (Figure 1). The theoretical model relates the four concepts found from coding the interview data: PBL, cognitive outcomes (CO), individual factors (IF) and group interactions (GI) and is illustrated in Figure 1.

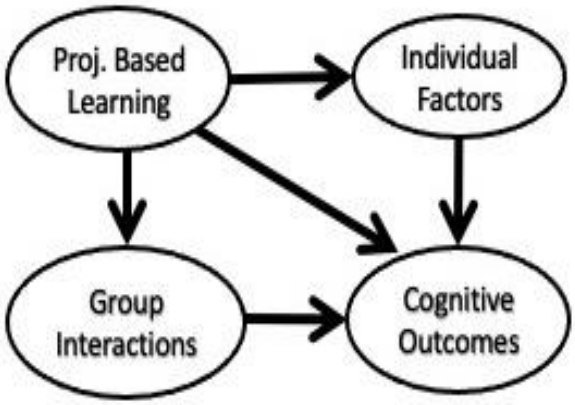


Figure 1: Grounded Theory Model

Project Based Learning (PBL)

The collaborative PBL projects are designed to require students to work in groups and learn the practical use of BA tools and methods by participating in genuine real-world experiences. Compeau, et.al. (1995) proposed a framework of key factors in the management of BA courses that highlights different phases of a BA project such as initiation, formal, informal, and continued learning and addresses the issue of transfer of learning to the workplace. The course structures together with the content impact the learning outcomes of the participants. PBL participants learn from each other as well as from the program content (Marcris, 2011; Leidner and Jarvenpaa, 1995) and execute the learning tasks in a genuine real world setting. PBL based pedagogy emphasizes these phases of learning and the casual transfer of knowledge among group members. Some of the beneficial use of PBL are identified by the business end users in interviews:

(1) *"Current business analytics work is complicated as it crosses several knowledge areas, and it is critical that students learn and use the standard methodology before they come to the workplace."*

IT tools to support group project-based learning includes collaboration systems (Microsoft Teams), descriptive content (lecture notes) and document management systems (Google Docs). Student engagement is also achieved with the help of discussion boards in the Canvas course management systems to allow rapid, real-time flow of information in response to student questions (Ghosh, Yoon &, Fustos, 2013). Students mentioned the benefits of the group projects:

(2) *"The project details were left up to the group and required working with the end users."*

(3) *"Group work was very helpful. We used Microsoft Teams to share with each other"*

The students learn from the knowledge of other group members, who come from different educational pathways to understand cross functional KPI's and build a logic driven BA model, using enterprise level data sets, that can be used to measure these KPI's. Such learning content also fosters joint work, the need for business problem solving and reflection and sharing of insights among the group members (Ryan and Deci, 2000). The BA Consultant says:

(4) *"They understand how real analytics projects are done. Students get job ready."*

There is considerable evidence to suggest that this peer support is also important to improve learning and course outcomes (Worrell, Gallagher & Mason, 2006; Volkoff, Elmes & Strong, 2004). Other students mentioned project characteristics such as: "*Needed more definition"* and the "*open nature of the project scope*" of what to accomplish made the project

"*interesting and challenging*."

PBL based pedagogy fosters the long-term success of educational programs in BA systems. Student quotes say:

(5) "*Most difficult part was dealing with the messy data and the project was frustrating at first, then we figured it out over time."*

(6) "*Working with a business user was difficult to coordinate and fully understand what they were asking for and why."*

(7) "*Working with real data and business users allowed me to learn ways of data collection, cleaning, aggregating and refreshing and then operationalizing analytics algorithms."*

Cognitive Outcomes (CO)

The three categories of learning outcomes are: (1) procedural goals such as the ability to use BA tools, (2) cognitive goals that focus on solving real business problems and (3) meta-cognitive goals that focus on the student's belief regarding their own abilities with the content (Gupta, et.al, 2010). A quote from the BA consultant notes,

(8) "*We have to do a better job to prepare students for work, where they work with incomplete pieces of information and be able to flush out the details in iterations".*

Cognitive learning outcomes (CO) include the mental awareness and judgments of the students and their ability to transfer their learning to new situations, such as applying the software application to solve a new problem different from what was used in the course project. Finally, meta- cognitive goals focus on enhancing the learner's ability to understand his/her own learning and information processing capabilities and confidence (Gupta, Bostrom and Huber, 2010). Business users quote says,

(9) *"It is crucial that the students get work experience during their college years. That is the only way they can succeed on the job after their degree".*

Higher level cognitive outcomes also include the growth of self-confidence to allow the transfer of the learning to new situations that require understanding the interactions of multiple parts of a complex scenario. When cognitive outcomes are emphasized in the learning program, the participants build the capability to apply their learning in real-world scenarios (Gupta, Bostrom and Huber, 2010). They grasp the path to apply the acquired knowledge of BA tools and methods, such as appropriate KPI's selection and implementation from organizational data. A student quote says:

(10) "*We could understand, from a business perspective, what the user really wanted to measure and accomplish from our project."*

This pedagogical approach also holds promise to address the difficulties of grasping the nuances of "real-world" BA methods without adversely impacting broader educational standards (Chang and Chou, 2011). Cognitive outcomes also include the growth of self-confidence to allow the transfer of the learning to new situations that require understanding the interactions of multiple parts of a complex scenario. To mimic real-world problems, which are typically ill-structured, the assigned PBL projects are loosely defined initially to require the groups to collaborate extensively to characterize the project scope. Student quotes say:

(11) *"Defining the project scope was important to be able to finish the work."*

The students proceed to identify diverse sources of data from different functional areas and design and create BA information products that span multiple business processes. Student interviews mention:

(12) *"We had so many questions that not all of them got answered."*

The learning outcomes for PBL is supported by four different sets of determinants: technology, individual difference, social influence, and situational constraints. PBL builds engagement among the students through trust (Gefen, 2002) and social integration during the learning process (Elbanna, 2003) and drives collaboration and knowledge sharing in the group. They share and combine their individual learning to support building a "big picture" and establishing their own collective group discourse (Wang and Ramiller, 2009). A student quote says:

(13) "*We had to make decisions and keep working on the project. Every week there were new items to work on and this rapid, flow of information in*

response to our questions helped guide our work."

Individual Factors (IF)

Individual factors include "mental states" and "learning traits". While "mental states" are general influences on performance that vary over time and include temporal factors such as motivation level and interest level, "traits" (such as preferred learning style are static aspects of individual factors, that affect a broad range of outcomes over time (Bostrom, Olfman and Sein, 1990). These factors play a role in how the PBL program can impact each students' learning process and outcome (Gupta, Bostrom and Huber, 2010). A student quote says:

(14) "*I had to work harder in some weeks to meet the deadlines with the business users. I did not want to be the slacking group member."*

Motivation to learn refers to the desire of an individual to learn knowledge and/or skills (adapted from Noe, 1986). Motivation to learn has been extensively studied in training literature and shown to be a key determinant of choices individuals make to engage in, attend to, and/or persist in learning activities (e.g., Klein et al., 2006; Facteau et al., 1995; Noe and Schmitt, 1986). Motivation theory explains individuals' learning behaviors (Van Der Heijden, 2004; Tharenou, 2001) and suggests that individual behavior is determined by two fundamental types of motivation: extrinsic (utilitarian) motivation and intrinsic (hedonistic) motivation (Alavi, Wheeler, and Valacich, 1995). As a student interview says:

(15) "It was good to work on a practical project that may benefit business people."

This suggests that motivation to learn can influence an individual's behavior (e.g., Kontoghiorghes, 2002; Colquitt et al., 2000; Noe, 1986). Compelling messages received from group members in support of the application of BA are also likely to influence individual factors about the expected outcomes of the curriculum. A student quote says:

(16) "*My group helped me understand better."*

The level of interaction within the project group facilitates individual engagement with the learning program. In group-based learning, team members work together and influence each other's motivation by voicing demands for contributions. Group projects require individuals to cooperate and work together but have significant learning benefits of efficiency and productivity (Baskin, Barker and Woods, 2005). As a student says:

(17) "I had to stay on schedule to work successfully with my group members."

Motivation is influenced by various factors, such as peer support (Facteau et al., 1995; Baldwin & Ford, 1988), and situational constraints (e.g., lack of time or resources) (Klein et al., 2006). In addition, motivation to learn is influenced by individual characteristics such as self-efficacy (Al-Eisa et al., 2009; Colquitt et al., 2000) and perceived benefits (Noe & Wilk, 1993). The motivation literature suggests that motivation can impel action and act as an inducement to action. According to Locke and Latham (2004), motivation can affect three aspects of action: direction (choice), intensity (effort), and duration (persistence). An business user quote says:

(18) "*The students were interested to learn about our business and address our needs."*

In addition, training literature suggests that motivation to learn can influence behavioral intention (e.g., Tharenou, 2001; Noe & Wilk, 1993). For example, according to Al-Eisa et al. (2009), motivation to learn was found to influence learning skill transfer intention, which refers to a commitment to apply newly acquired knowledge or skills to the work setting. A student quote says:

(19) "*My interest about business analytics jobs grew from doing this course."*

Group Interactions (GI)

Group interactions comprise factors such as if team members shared diverse viewpoints and if such interactions were valued as well as the nature of cooperation and the level of dialog achieved within the team. Project based learning (PBL) that uses authentic, complex scenarios creates an impetus for group dialog to apply that knowledge to solve the problem assigned (Uribe, Klein and Sullivan, 2003).

Shared cognition theory focuses on individual learning within a social situation, allowing for social interactions that support the individual's cognitive development with help from more capable group members. Based on shared

cognition theory, project-based learning (PBL) allows participants to engage in learning activities by working in groups to investigate and respond to a complex question, problem, or challenge (Marcris, 2011; Alavi, Wheeler and Valacich, 1995; Leidner and Jarvenpaa, 1995). A business user quote says:

(20) "Our dialog with the students was beneficial to all of us. They got some work experience and we got new ideas."

The level of interaction within the project group facilitates individual engagement with the learning program. In group-based learning, team members work together and influence each other's motivation by voicing demands for contributions. A student quote says:

(21) "We supported each other in our group as the project was challenging and was it was necessary to divide up the work."

PBL supports collaborative group learning and the sharing of knowledge among team members. The PBL group creates, and shares goals and learns together by working jointly and solving the problems posed by the project. The group interactions play a critical role in the learning environment through the size and heterogeneity of the team. Group interactions impact learning outcomes by developing diverse knowledge and building broader perspectives that span business functions (Seethamraju, 2008). As a student says:

(22) "Group members were helpful to understand the project tasks as well as how to do the project."

Students of BA must grasp and integrate cross-disciplinary knowledge so they can communicate and work cooperatively with others (Wang and Ramiller, 2009). Based on situated learning theory, effective group learning programs must require that group members reflect upon their learning and contribute their experiences, observations, and insights back into the group's collective discourse in a group-based collaborative setting (Wang and Ramiller, 2009). As important referents communicate in the PBL setting, an individual may incorporate the opinions of peers as a part of her own belief structure (Fulk 1993; Lewis, Agarwal and Sambamurthy, 2003). As a student says:

(23) "I liked the ideas shared by my group members as I never thought of them before."

Group theories suggest that many factors can influence the outcomes of group-based learning (Sharda, Romano, Lucca, Weiser, Scheets, Chung and Sleezer, 2004). This includes group characteristics, such as composition (level of homogeneity and heterogeneity), amount of group cooperation and the nature of group communications. Group influence has been found to emanate from a variety of sources (Lewis, Agarwal and Sambamurthy, 2003). Each participant brings their own experience and expertise to share their knowledge with the group. There is a constant interaction and collaboration among participants that allows everyone to develop more improved skills in solving problems, than if they worked alone (Sharda, et.al., 2004). The joint experience allows each participant to explore the project from other user's perspectives and helps them to bridge "gaps" in understanding, create new meanings and explanations through shared understanding and practical use to perform specific tasks (Chang and Chou, 2011).

	Property	Quote
PBL	Cross functional, group problem solving approach	23
	Interactions with real world business users and "messy" data	2
	"Fuzzy" details to be worked out using iterative methodology	11
CO	Mastery of BA methodology and industry practices	19
	Self-confidence to execute BA project (beyond Tech credentials)	10
	Demonstration of BA project skills thru adaptability and application	1
IF	Motivation- intrinsic	14
	Motivation- extrinsic	15
	Learning style	Bostrom ,et.al. (1990)
GI	Support and Teamwork	3
	Knowledge Sharing and Cognition	22

Table 3. Concept Development and Coding

4. CONCLUSIONS

Business analytics (BA) courses are growing in university curricula as students seek to build BA skills and knowledge in response to employment demand from industry. Commercial organizations are increasingly adopting BA

systems to facilitate data driven decision making by allowing easier data manipulation, visualization, and processing. However, the complexity and diversity of BA systems and their inter-disciplinary nature make their pedagogy difficult at the curriculum level. Many institutions find that emphasizing quantitative knowledge and building BA tool procedural skills fall short of what is demanded by industry. Authentic real-world project-based learning (PBL) requires that students work with "messy" data with incompatibilities, select and apply complicated algorithms to process the data and the engage with actual business users to learn to manage their involvement with project tasks while learning to use the BA methodologies and tools. Data pre-processing is often not covered in traditional BA courses but is a key learning outcome of the PBL pedagogy in a BA courses. The use of practical projects with real world business end users allowed students to better understand these aspects of practical BA systems.

The study develops an innovate project based learning (PBL) program for BA courses and proposes a model based on grounded theory. The PBL program allows participants to learn the concepts of BA collectively and is supported by a market leading vendor's BA tool. The unique features of the program are (1) use of actual real world client data and (2) availability of client business users to allow the participants to collect analytics business requirements, (3) the educational diversity of PBL group members and (4) the iterative approach to the project development using periodic reviews. The study found that PBL is a viable pedagogical approach to support higher cognitive outcomes of BA courses. PBL increases interactions among students working in project groups that provide a higher cognitive level of learning. The interactions of the students with the business end users were essential for the reliability of the dashboards and reports and their use for decision making.

This study meets the criteria of applicability in grounded theory. It fits the substantive area of study, and it is understandable to the practitioner, and it provides potential control for the action and conditions to which it applies. The results of the study finds evidence to support the notion that project based BA learning programs promote strong group interactions that drive to increase student motivation. The contents of the learning program, such as the use of authentic real-world scenarios, the involvement of external business end users and the diversity in the student backgrounds support building higher cognitive outcomes of the participants.

Implications

This study supports the findings from prior research in the context of BA course curriculum that four categories of individual factors: technology characteristics, motivation, social influence and situational constraints have a critical impact in BA learning outcomes. These factors are all sufficiently represented in the proposed group PBL pedagogy.

Based on interview data collected among students, end user and an industry consultant, this paper finds support for an empirical model that shows a relationship between PBL and cognitive outcomes. Additionally, relationships between group interactions and individual motivation to learn on cognitive outcomes was modelled. The following points follow:

1. BA curriculum must be guided by inter-disciplinary knowledge and skills and go beyond quantitative skills to include real-world experiences to build cognitive outcomes.
2. Business Analytics systems differ from other IS implementations by crossing functional boundaries and do not fit well with many current BA courses in current educational curricula.
3. Learning programs that emphasize practical projects and experimentation can allow participants to have greater motivation to learn and lead to higher levels of cognitive outcomes.
4. The group project-based learning (PBL) approach also supports group interactions that benefit students and the business end users.

5. REFERENCES

Alavi, M., Wheeler, B.C., Valacich, J.S. (1995). Using IT to Reengineer Business Education: An Exploratory Investigation of Collaborative Tele Learning. *MIS Quarterly 19(3)*, 293-312.

Al-Eisa, A. S., Furayyan, M. A., Alhemoud, A. M. (2009). An empirical examination of the effects of self-efficacy, supervisor support and motivation to learn on transfer intention. *Management Decision 47(8),* 1221-1244.

Baldwin, T. T., Ford, J. K. (1994). Transfer of training: A review and directions for future

research. The training and development sourcebook, 180.

Baskin, C., Barker, M. Woods, P. (2005). When group work leaves the classroom does group skills development also go out the window*? British Journal of Educational Technology 36 (1),* 29-31.

Bose, R. (2009). Advanced Analytics: Opportunities and Challenges. *Industrial Management & Data Systems 109(2*), 155-172.

Bostrom, R.P., Olfman, L., Sein, M.K. (1990). The Importance of Learning Style in End-User Training. *MIS Quarterly 14(1),* 101-119.

Cegielski, C.G. & Jones-Farmer, L.A. (2016). Knowledge, Skills and Abilities for Entry-Level Business Analytics Positions: A Multi-Method Study. *Decision Sciences Journal of Innovative Education, 14(1),* 91-118.

Chang, H.H., Chou, H.W. (2011). Drivers and effects of Enterprise Resource planning post-implementation learning. *Behaviour and Information Technology 30 (2),* 251-259.

Colquitt, J. A., LePine, J. A., Noe, R. A. (2000). Toward an integrative theory of training motivation: a meta-analytic path analysis of 20 years of research*. Journal of Applied Psychology, 85(5)*, 678-707.

Compeau, D., Olfman, L., Sein, M., Webster, J. (1995). End-user training and learning. *Communications of the ACM 39 (7),* 24-26.

Davenport, T.E. and Harris J.G. (2007). Competing on Analytics: The New Science of Winning. Harvard Business School Press.

Eisenhardt, K.M. (1989). Building Theories from case study research. *Academy of Management Review, 14(4),* 532-550.

Elbanna, A. (2003). Achieving social integration to implement ERP systems. In: 11th European Conference on Information Systems, paper 33

Elbashir, M.Z., Collier, P.A., Davern, M.J. (2008). Measuring the Effects of Business Intelligence Systems: The Relationship between Business Process and Organizational Performance. *International Journal of Accounting Information Systems, 9(3),* 135-153.

Facteau, J. D., Dobbins, G. H., Russell, J. E., Ladd, R. T., Kudisch, J. D. (1995). The influence of general perceptions of the training environment on pretraining motivation and perceived training transfer. *Journal of management 21(1),* 1-25.

Fulk, J. (1993). Social construction of communication technology. Academy of *Management journal 36(5),* 921-950.

Gefen, D. (2002). Nurturing client's trust to encourage engagement success during the customization of ERP systems. *Omega 30(4),* 287-299.

Gefen, D., Karahanna, E. and Straub, D.W. (2003). Trust and TAM in Online Shopping: An Integrated Model, *MIS Quarterly 27(1)*, 51-90.

Ghosh, B., Yoon, T. E., Fustos, J. T. (2013). Enhancing Functional Fit with Continuous Training during the ERP Post-Implementation Phase. *International Journal of Information Systems in Service Sector 5(2),* 30-45.

Gupta, S., Bostrom, R.P., Huber, M. (2010). End-User Training Methods: What we know, Need to Know. *The DATABASE for Advances in Information Systems 41(4),* pp. 9-39.

Kontoghiorghes, C. (2002). Predicting motivation to learn and motivation to transfer learning back to the job in a service organization: A new systemic model for training effectiveness. *Performance Improvement Quarterly*, *15*(3), 114-129.

Klein, H. J., Noe, R. A., Wang, C. (2006). Motivation to learn and course outcomes: The impact of delivery mode, learning goal orientation, and perceived barriers and enablers. *Personnel Psychology (59*), 665-702.

Leidner, D.E. Jarvenpaa, S.L. (1995). The use of Information Technology to enhance Management School education: A theoretical view. *MIS Quarterly 19(3*), 265-291.

Lewis, W., Agarwal, R., Sambamurthy, V. (2003). Sources of Influence on Beliefs About Information Technology Use: An Empirical Study of Knowledge Workers. *MIS Quarterly 27(4)*, 657-678.

Locke, E. A., & Latham, G. P. (2004). What should we do about motivation theory? Six recommendations for the twenty-first century. *Academy of management review, 29(3),* 388-403.

Markov, W., Braaganza, S., Taska, B., Miller, S.M. & Hughes, D. (2017). The Quant Crunch: How the Demand for Data Science

Skills is Disrupting the Job Market. Retrieved August 1, 2022 from https://www.ibm.com/downloads/cas/3RL3VXGA

Marcris, A.M. (2011). Enhancing Enterprise Resource Planning users' understanding through ontology-based training. Computers in Human Behavior. 27(4), 1450-1459.

Paul, J.A. & MacDonald, L. (2020). Analytics Curriculum for Undergraduate and Graduate Students. *Decision Sciences Journal of Innovative Education, 18(1),* 22-58.

Mills, R.J., Chudoba, K.M., & Olsen, D.H. (2016). 15 Programs Responding to Industry Demand for Data Scientists: A Comparison Between 2011-2016. *Journal of Information Systems Education*, *27(2),* 131-140.

Noe, R. A. (1986). Trainees' attributes and attitudes: Neglected influences on training effectiveness. *Academy of management review 11(4),* 736-749.

Noe, R. A., Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model*. Personnel Psychology 39(3*), 497-523.

Noe, R. A., Wilk, S. L. (1993). Investigation of the factors that influence employees' participation in development activities. *Journal of applied psychology, 78(2),* 291-302.

Radovilsky, Z. & Hegde, V. (2022). Contents and Skills of Data Mining Courses in Analytics Programs. *Journal of Information Systems Education, 33(2),* 182-194.

Ryan, R.M., Deci, E.L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions, *Contemporary Educational Psychology 25,* 54-67.

Seethamraju, R. (2008). Enhancing Student Learning through ERP Business Simulation Game. In: Proceedings of the AIS SIG-ED IAIM Conference.

.

Sharda, R., Romano Jr., N.C., Lucca, J.A., Weiser, M., Scheets, G., Chung, J.M., Sleezer, C.M. (2004). Foundation for the Study of Computer-Supported Collaborative Learning Requiring Immersive Presence. *Journal of Management Information Systems 20(4),* 31-63.

Strauss, A. and Corbin, J. (1990). Basics of Qualitative Research: Grounded Theory. Procedures and Techniques Sage.

Suddaby, R. (2006). From the Editors: What Grounded Theory is Not. *Academy of Management Journal 49(4),* 633-642.

Tharenou, P. (2001). The Relationship of Training Motivation to Participation in Training and Development. *Journal of Occupational and Organizational Psychology 74(5),* 599-621.

Uribe, D., Klein, J.D., Sullivan, H. (2003). The Effect of Computer Mediated Collaborative Learning on Solving ill-Defined Problems. *ETR&D 51(1),* 5-19.

Van Der Heijden, H. (2004). User Acceptance of Hedonic Information Systems. MIS Quarterly 28, 695-704.

Volkoff, O., Elmes, M., Strong, D. (2004). Enterprise systems, knowledge transfer and power users. *Journal of Strategic Information Systems 13(4*), 279-304.

Wang, P., Ramiller, N.C. (2009). Community Learning in Information Technology Innovation. *MIS Quarterly 33(4),* 709-734.

Worrell, J., Gallagher, K., Mason, R. (2006). Explaining the structure of post-implementation ERP teams*.* In: 12th Annual Americas Conference on Information Systems, Acapulco, MX.

Yap, A. Y., Drye, S. (2018). The Challenges of Teaching Business Analytics: Finding Real Big Data for Business Students. *Information Systems Education Journal,* 16(1) pp 41-50