데이터 사이언스 교과과정에 대한 연구*

A Study on the Curriculums of Data Science

이 명 호 (Myongho Yi)**

초 록

본 연구는 국내외 데이터 사이언티스트(Data Scientist) 양성을 위한 데이터 사이언스(Data Science) 프로그램의 교과과정을 분석하였다. 이를 위해 국내 7개 대학교와 미국의 10개 대학교를 분석하였다. 14개의 데이터 사이언스 과정이 대학원 중심으로 운영되고 있는 것으로 나타났다. Conway의 데이터 사이언스 3대 영역 중 수학 및 통계 지식 영역에 국내는 10% 그리고 미국은 26%가 치중되어 있는 것으로 분석되었다. 강의계획서 분석에서 수업내용 및 평가 방법은 국내외가 유사한 것으로 나타났다. 본 연구 결과는 국내 데이터 사이언스 교과과정 개발에 기초 자료로 활용될 수 있을 것이다.

ABSTRACT

The purpose of this study is to compare seven data science programs in Korea and ten data science programs in the US. Results show that 14 data science programs are housed in graduate schools. 10% of data science courses in Korea and 26% in the US fall under the Math and Statistics Knowledge area, one of the three areas defined by Conway. The syllabus analysis does not show much differences in terms of class contents and grading. The results of this study can be used to design data science programs that are more effective and well-grounded.

키워드: 데이터 사이언티스트, 데이터 사이언스, 문헌정보학, 빅데이터
Data Scientist, Data Science, Library and Information Science, Big Data

---

* 본 논문은 2015년도 상명대학교 교내연구비를 지원받아 수행하였음.
** 상명대학교 문헌정보학과 조교수(josephlee@smu.ac.kr)
논문접수일자 : 2016년 2월 23일 논문심사일자 : 2016년 3월 17일 게재확정일자 : 2016년 3월 21일
한국비블리아학회지, 27(1): 263-290, 2016. [http://dx.doi.org/10.14699/kbiblia.2016.27.1.263]

1. 서 론

1.1 연구의 필요성 및 목적

우리는 많은 정보를 웹이라는 환경을 통해서 이용하고 있고 웹을 통해서 많은 데이터가 수집, 저장, 관리되고 있다(Loukides 2010). 2012년 디지털 데이터의 양이 아날로그 데이터양을 앞지르는 것을 시작으로 이제는 측정조차 불가능한 많은 양의 디지털 데이터가 만들어지고 있다. 디지털 데이터 생성의 예는 비행기나 감시 카메라 등 각종 센서에서 만들어지는 데이터, 소셜 미디어(Social Media)를 통해 만들어지는 데이터, 웹 데이터 등이 있다. 또한 데이터의 90%는 지난 2년간 만들어진 최근 데이터이다(SINTEF 2013). 이처럼 최근에 만들어진 많은 양의 다양한 데이터는 수집되고 분석이 되어져야 할 필요성을 인식하지만 기존의 방법으로는 데이터의 처리가 어려워지고 있다(Shi, Yu, Zhu, and Tian 2014). 이러한 데이터를 수집, 분석, 처리하는 학문 분야 중 하나가 데이터 사이언스(Data Science)이며, 이는 수학, 통계학, 경영학, 컴퓨터과학, 문헌정보학 등 다양한 학문의 융합이라고 정의되고 있다(Shi et al. 2014). 데이터 사이언스 과정을 통해서 데이터 사이언티스트가 배출이 되는데 데이터 사이언티스트(Data Scientist)에 대한 긍정적인 전망의 연구는(Davenport and Patil 2012; 이성신, 최재황, 이창수 2013) 데이터 사이언티스트에 대한 사회적 수요 증가로 이어지고 있다(Miller 2013; Provost and Fawcett 2013). 데이터 사이언티스트에 대한 수요가 급증하면서 국내외적으로 데이터 사이언스 프로그램을 제공하는 학교가 기학급수적으로 늘어나고 있으며(Lipman 2014; 김희섭, 남권희, 강보라 2013; 이성신, 최재황, 이창수 2013; 장덕현 2015; 장윤금 2014) 데이터 사이언스 프로그램을 운영하는 학과도 경영학, 통계학, 컴퓨터과학, 문헌정보학 등 매우 다양하다(Swanstrom 2015). 데이터 사이언스에 대한 관심과 데이터 사이언티스트에 대한 사회적 수요가 증가하고 있지만 데이터 사이언스 교과과정에 대한 연구는 부족하여 이에 대한 연구가 더 필요하다. 본 연구는 국내외에 증가하고 있는 데이터 사이언스 프로그램에 대한 교과과정을 살펴보고 데이터 사이언스 정의와 관련된 교과 과정이 개설되고 있는지 그리고 어느 교육단위에서 데이터 사이언스 학위를 제공 하고 있는지 살펴보고자 한다. 또한 국내와 미국의 데이터 사이언스 교과과정을 비교하여 국내 데이터 사이언스 교과과정에 대한 개선 방향을 제시하고자 한다.

1.2 연구방법

본 연구에서는 데이터 사이언티스트 양성을 위한 교육 및 교과과정의 개선방안을 제안하기 위해 국내 데이터 사이언스 학위 교과과정과 미국의 데이터 사이언스 학위 교과과정을 비교·분석한다. 이를 위해 문헌조사법과 비교분석법을 사용하고 있다.

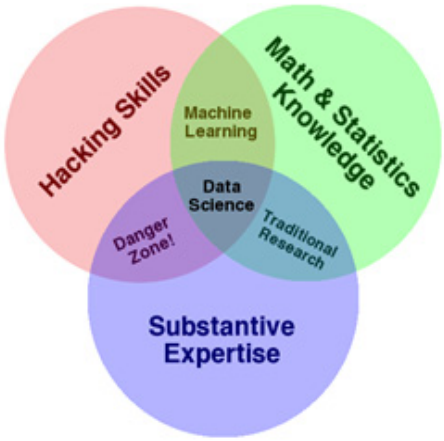
구체적인 분석 방안은 다음과 같다. 첫째, 데이터 사이언스 교과과정과 관련된 연구 논문들을 조사함으로써 교과과정과 관련된 연구들의 경향 및 내용들을 분석하였다. 둘째, 2016년 현재 데이터 사이언스(Data Science) 또는 빅데이터(Big Data) 과정을 개설하고 있는 504개

대학 중에서 미국의 10개 대학과 국내 7개 대학의 교과과정 및 학위현황 등을 비교・분석하였다. 셋째, 데이터 사이언스 교과과정을 Conway의 데이터 사이언스 정의를 기본으로 한 특정분야 지식, 수학 및 통계 지식, 해킹 지식 세 개 분야로 구분하여 각 학교의 영역별 분포도에 대하여 비교・분석하였다. 영역 범주에 포함시키기 어려운 일부 과목은 기타 영역에 포함시켰다. 넷째, 데이터 사이언스가 설치되어 있는 학과를 조사하여 국내외적인 흐름을 조사하였다.

2. 이론적 배경

2.1 데이터 사이언스 정의 및 영역

Jagadish(2015)에 의하면 빅데이터는 데이터의 특징에서 나온 의미이고 데이터 사이언스는 데이터의 사용에서 시작 된 것으로 그 차이를 설명하고 있다. National Consortium for Data Science는 데이터 사이언스를 디지털 데이터에 대한 과학적 관찰, 이론 개발, 시스템적 분석, 가설 실험, 검증을 하는 분야라고 정의하고 있다(Jagadish 2015). 데이터 사이언스 용어는 2001년도에 Cleveland에 의해서 처음 사용이 되었다(Cleveland 2001). Cleveland(2001)의 연구에서 컴퓨터 사이언스와 통계학이 지니고 있는 각각의 한계를 극복하고 서로 시너지 효과를 낼 수 있도록 하자는 의미에서 데이터 사이언스를 제안했다. 데이터 사이언스 정의에 대한 구체적인 연구는 Conway(2010)의 벤 다이어그램이 많이 알려져 있다. Conway는 〈그림 1〉에서 보듯이 세 개의 다른 영역의 교집합 부분을 데이터 사이언스라고 정의하고 있다. 첫 번째 영역은 수학 및 통계 지식이고 두 번째는 해킹 기술이며 세 번째는 특정분야에 대한 지식이다. 세 영역이 겹치는 부분을 기계학습(Machine Learning), 전통적 연구(Traditional Research), 그리고 위험영역(Danger Zone)이라고 추가적인 정의를 하고 있다. 기계학습은 해킹 기술과 수학 및 통계 지식이 겹치는 부분으로 주로 기계에 의존하여 데이터를 분석하는 영역으로 정의하고 있다. 두 번째 전통적 연구 영역이다. 이는 특정분야 지식과 수학 및 통계 지식이 겹치는 영역으로 박사급의 연구자들이 특정 분야에 분석을 주로 통계학을 이용하는 부분으로 정의하고 있다. 마지막으로 위험영역은 특정 분야 지식과 해킹 기술이 만나는 부분으로 수학이나 통계에 대한 지식이 없이 데이터를 분석하는 것은 위험이 존재한다고 설명하고 있다.



〈그림 1〉 Conway 데이터 사이언스 밴다이어그램

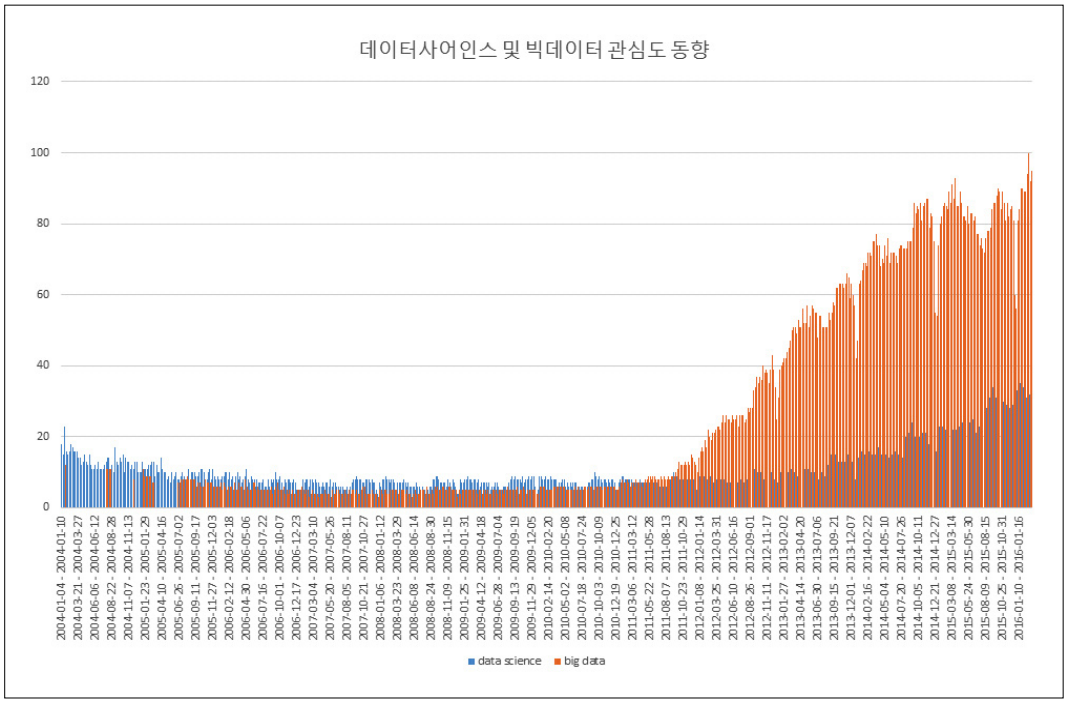
2.2 데이터 사이언스 동향

〈그림 2〉에서 보듯이 빅데이터라는 용어에 대한 전 세계적인 관심도는 2012년도에 급격히 증가하고 있다(Google Trends 2016).

2012년도 빅데이터에 대한 급격한 관심도의 증가는 여러 가지 이유가 있다. 세계적인 연구기관인 Gartner가 매년 발표하는 10개 기술 중 2012년 기술에 빅데이터가 선정되었다(Gartner 2011). 또한 세계경제포럼은 2012년 떠오르는 10대 기술 중 하나로 빅데이터 기술을 선정하였으며 국내에서도 지식경제부에서 2012년도에 10대 핵심기술 하나로 빅데이터를 선정하였다(최윤식 2012).

2.3 데이터 사이언스 교과과정 관련 연구

데이터 사이언스 프로그램의 수는 급속도로 증가하고 있으나 데이터 사이언스 교과과정에 대한 연구는 매우 부족한 실정이다. 다만 문헌정보학 또는 데이터 사이언스를 제공하는 주요 학과 중 하나인 School of Information 또는 iSchool에 대한 교육과정은 연구가 되고 있다(Chu 2012; Brynko 2012; 김희섭, 남권희, 강보라 2013; 이성신, 최재황, 이창수 2013; 장덕현 2015; 장윤금 2011, 2014). 장덕현(2015)은 세계적인 iSchool 운동에 대해서 국내 문헌정보학이 나아가야 할 방향에 대해서 제시하고 있다. Chu(2012)는 미국의 iSchools과 non-iSchools의 교과과정을 분석하였으며 iSchools은 non-



〈그림 2〉 데이터 사이언스, 빅데이터 용어 관심도

iSchools보다 월등히 많은 과목을 제공하고 있음을 밝혔다. 김희섭, 남권희, 강보라(2013)는 iSchool 대학들의 교육과정을 파악하였는데. iSchool 대학은 정보학, 실습연구, 경영 등의 영역이 높은 것으로 분석하였다. 이재윤(2015)은 데이터 사이언스를 문헌정보학과 도서관의 입장에서 기술하였다. 특히 데이터 리터러시를 문헌정보학 영역에서 수용할 수 있는 데이터 사이언스 대체 개념으로 제시하였다(이재윤 2015). 데이터 리터러시는 가공되지 않은 정보를 어떤 방식으로 처리할 수 있는 능력으로 정의하고 있다. Bussaban and Waraporn(2015)는 컴퓨터과학 교과목과 수학학과의 교과목을 일부 선택하여 데이터 사이언스 프로그램으로 제공한 결과 데이터 사이언스 프로그램에 등록하는 학생들의 수가 증가하고 교수들의 긍정적인 효과를 가져왔다고 연구 결과를 발표하였다. 〈표 1〉에서 보듯이 데이터 사이언스 영역을 5가지로 나누고 이에 관련된 교과 과목을 제시하였다(Bussaban and Waraporn 2015).

〈표 1〉 데이터 사이언스 교과 과정

영역	교과 과정
통계	시뮬레이션, 시각화, 구축
데이터 형식	텍스트, 데이터 정제
데이터 기술	데이터베이스, XML
프로그래밍	프로그래밍
비즈니스	웹 출판

3. 연구 설계

3.1 연구 대상 선정 및 자료수집

연구대상으로 학교 선정은 Data Science Community에 게시되어 있는 504개의(2016년 2월 현재) 데이터 사이언스 프로그램 제공 대학 중 미국의 Top 10 대학을 선정하였다. 미국의 Top 10 대학은 2015년 US News & World Report에서 제공하는 랭킹을 참조하였다. 국내의 경우는 데이터 사이언스 또는 빅데이터 관련 학위를 제공하는 7개 대학을 모두 선정하였다. 연구대상의 학과 홈페이지, 이메일 문의, 문헌조사 등을 통하여 교과과목을 조사하였다. 교과 과목은 학사, 석사, 박사 등 학위 별로 구분하여 자료를 수집하였다. 〈표 2〉는 국내 데이터 사이언스 프로그램을 가나다 순으로 정리한 표이다. 〈표 3〉은 미국의 데이터 사이언스 프로그램을 ABC 순으로 정리한 표이다. [부록 1]은 국내 대학별 그리고 교과과정을 가나나 순으로 정리한 표이다. [부록 2]는 미국 대학별 그리고 교과과정을 ABC 순으로 정리한 표이다. 데이터 사이언스 교과목에 대한 비교와 함께 연구대상의 과목 중 강의계획서 [부록 3]과 [부록 4]를 분석하였다.

〈표 2〉 국내 데이터 사이언스 프로그램

(가나다순)

학교명	학과명	학위과정
국민대학교	데이터 사이언스학과	석사
단국대학교	데이터 사이언스학과	석사
상명대학교	데이터 사이언스학과	학사
서울대학교	디지털정보융합학과	석사
서울산업대학교	데이터 사이언스학과	석사
성균관대학교	데이터 사이언스학과	학사
연세대학교	빅데이터	석사

〈표 3〉 미국 데이터 사이언스 프로그램

(ABC 순)

학교명	학과명	학위과정
Columbia University	Data Science Institute	석사
Indiana University	School of Informatics and Computing	석사
New York University (NYU)	Center for Data Science	석사
Rutgers University	Business School	석사
Southern Methodist University (SMU)	Humanities and Sciences, Engineering and Arts	석사
Stanford University	Statistics	석사
University of California - Berkeley	School of Information	석사
University of Michigan	Electrical Engineering and Computer Science	학사
University of Virginia	Data Science Institute	석사
University of Wisconsin - Madison	Statistics	석사

4. 결과 분석

국내의 데이터 사이언스 총 과목 수는 181개이며 가장 많은 교과목을 제공하는 학교는 34과목을 제공하는 서울대학교와 성균관대학교이다. 국내는 평균 25개의 데이터 사이언스 과목을 제공하고 있다. 미국의 데이터 사이언스 총 과목 수는 91개이며 가장 많은 교과목을 제공하는 학교는 18과목을 제공하는 Indiana 대학교이다. 미국은 평균 9개의 데이터 사이언스 과목을 제공하고 있다.

4.1 교과 과목 분석

조사된 272 개의 데이터 사이언스 과목을 분류하고자 Conway(2010)의 데이터 사이언스 정의에 따른 세 개의 영역으로 구분하였다. Conway(2010)에 의하면 데이터 사이언스는 수학 및 통계 지식, 해킹 기술, 그리고 특정분야에 대한 지식으로 분류가 된다. 본 연구에서 특정분야 지식은 데이터 사이언스의 세 개 영역 중 하나이지만 272개 교과과정에서는 분류하는데 어려움이 있다. 이는 데이터 사이언티스트의 다양한 특정분야에 대한 전문지식으로 이에 대한 별도 교과목은 데이터 사이언스 교과과정에서 제시를 하고 있지 않기 때문이다. 〈표 4〉에서와 같이 두 개 영역으로 분류한 결과 국내 데이터 사이언스 전체 과목 중 해킹 기술이 차지하는 비율은 6.1%이고 수학 및 통계 과목은 10%이다. 해킹 기술, 수학 및 통계 지식의 영역이 아닌 기타 영역은 83.9 %이다. 〈표 5〉에서와 같이 미국 데이터 사이언스 91개 과목 중 해킹 기술은 7.6%이며 수학 및 통계 분야는 26%이다. 기타 영역은 66.4%이다. 국내의 데이터 사이언스 교과과정과 미국의 데이터 사이언스 교과과정 모두 수학 및 통계 분야에 치중된 것을 볼 수 있다. 또한

〈표 4〉 국내 데이터 사이언스 교과 과목 비율

데이터 사이언스 영역	과목 수	비율
해킹기술	11	6.1%
수학 및 통계 지식	18	10%
기타 영역	152	83.9%

〈표 5〉 미국 데이터 사이언스 교과 과목 비율

데이터 사이언스 영역	과목 수	비율
해킹기술	7	7.6 %
수학 및 통계 지식	24	26 %
기타 영역	60	66.4 %

국내의 기타 영역은 83.9%로 미국의 기타 영역 66.4%보다 높은 것으로 분석되었다.

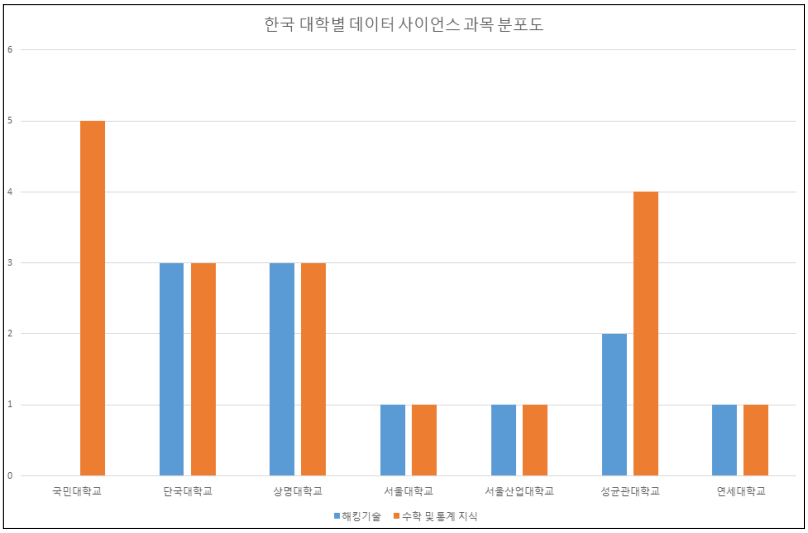
4.2 강의계획서 분석

국내외 대학에서 제공하는 있는 데이터 사이언스 개론 수업 중 홈페이지에 공개되어 있는 강의계획서 [부록 3]과 [부록 4] 내의 수업내용, 수업방법 및 평가방법을 분석하였다. 국내 데이터 사이언스 개론 수업 내용은 데이터 사이언스에 대한 정의, 데이터 시각화, 데이터 보안, 데이터 마이닝 등 다양한 내용이 기술되어 있는 것으로 나타났다. 수업 방법은 강의, 토론, 팀프로젝트 등이 사용되고 있는 것을 알 수 있었다. 평가 방법은 수업 참여, 시험, 팀프로젝트 등이 적용되고 있었다. 미국 데이터 사이언스 개론 수업 방법은 토론, 팀프로젝트 등이 사용되고 있다. 평가 방법은 과제, 시험, 팀프로젝트, 수업 참여 등이 적용되고 있었다. 데이터 사이언스 개론에 대한 수업방법 및 평가방법은 국내외간 큰 차이가 없는 것으로 분석되었다.

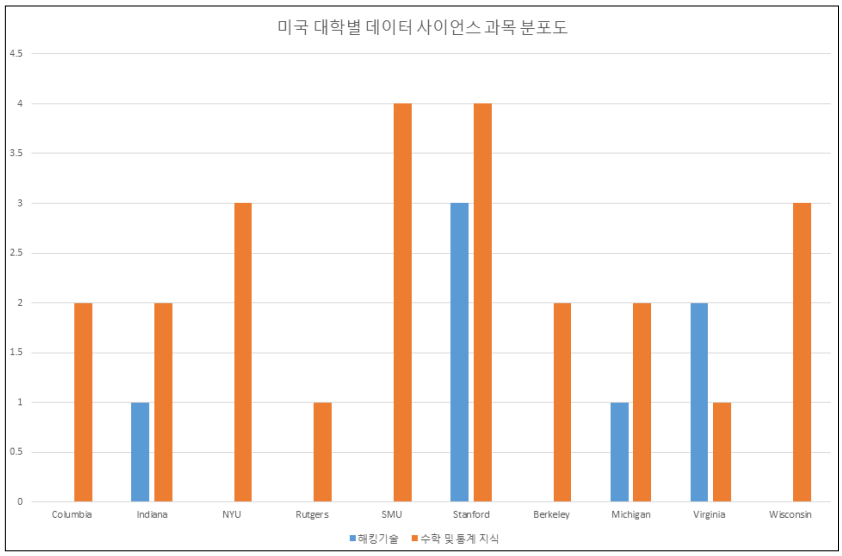
4.3 대학별 교과과정 영역 분석

국내 7개 대학의 데이터 사이언스 정의에 따른 해킹기술과 수학 및 통계 지식 과목에 대한 분포도를 조사하면 〈그림 3〉과 같다. 국민대학교는 수학 및 통계지식 영역에 집중되어 있고 다른 6개 대학은 두 영역을 모두 제공하고 있다.

미국 10개 대학의 데이터 사이언스 정의에 따른 해킹기술과 수학 및 통계 지식 과목에 대한 분포도를 조사하면 〈그림 4〉와 같다. Columbia,



〈그림 3〉 국내 대학별 데이터 사이언스 과목 분포도



〈그림 4〉 미국 대학별 데이터 사이언스 과목 분포도

NYU, Rutgers, SMU, Berkeley, Wisconsin 대학은 수학 및 통계지식 영역에 집중되어 있고 다른 4개 대학은 두 영역을 모두 제공하고 있다. Virginia 대학만 제외하고 수학 및 통계지식 영역에 집중하고 있음을 볼 수 있다.

4.4 데이터 사이언스 학과 및 학위과정 분석

국내와 미국 데이터 사이언스 학위 및 학교 비교는 〈표 6〉과 같다. 17개 국내외 데이터 사이언스 학위과정 중 학부에 개설된 데이터 사이언스 과정은 국내에 2개, 미국에 1개이다. 82%의 데이터 사이언스 과정이 대학원 중심으로 운영되고 있음을 보여주고 있다.

데이터 사이언스가 제공되고 있는 학위과정 비교는 〈표 7〉과 같다. 국내는 7개 대학 중 5개 대학이 데이터 사이언스학과로 학위과정이 명시되어 있으며 미국도 4개 대학이 Data Science로 학위과정이 명시되고 있다. 또한 미국은 3개 대학이 통계학과에 학위과정이 개설되어 있다.

〈표 6〉 국내와 미국 데이터 사이언스 학위 비교

국가	학위	학교
국내	학사	상명대학교 성균관대학교
	석사	국민대학교 단국대학교 서울대학교 서울산업대학교 연세대학교
미국	학사	University of Michigan
	석사	Columbia University Indiana University New York University Rutgers University Southern Methodist University Stanford University UC Berkeley University of Virginia University of Wisconsin - Madison

〈표 7〉 국내와 미국 데이터 사이언스 학위과정 비교

국가	학위과정	학교
국내	데이터 사이언스학과	상명대학교 성균관대학교 국민대학교 단국대학교 서울산업대학교
	디지털정보융합학과	서울대학교
	빅데이터학과	연세대학교
미국	Information and Data Science	UC Berkeley
	Master of Information	Rutgers University
	Informatics	Indiana University
	Data Science	Columbia University
		New York University
		University of Virginia
		University of Michigan
	Statistics: Data Science	Stanford University
		University of Wisconsin - Madison
		Southern Methodist University (SMU)

5. 결론 및 제언

데이터 사이언스에 대한 국내외적 수요가 기하급수적으로 증가하고 있다. 이러한 가운데 데이터 사이언티스트를 양성하는 프로그램이 늘어나고 있다. 본 연구는 이러한 데이터 사이언스 프로그램의 교과과목에 대한 국내외 현황을 살펴보았다. 2016년 현재 국내에는 총 7개의 대학/대학원 과정이 있으며 과목 수는 181개의 과정을 제공하고 있다. 본 연구를 위해 선정된 미국의 10개 데이터 사이언스 프로그램은 총 91개의 교과과정을 개설하고 있다. Conway의 데이터 사이언스 정의에 의한 세 개의 영역별로 272개 과목을 분류한 결과 국내, 미국 모두 수학 및 통계 지식 영역 과목이 각각 10%, 26%로 가장 많은 교과과정 영역을 차지하고 있다. 국내외 대학별 교과과정을 분석한 결과 국내에서는 국민대학교가 수학 및 통계지식 영역에 집중되어 있고 다른 6개 대학은 두 영역을 모두 제공하고 있다. 미국 대학 중 Columbia, NYU, Rutgers, SMU, Berkeley, Wisconsin 대학은 수학 및 통계지식 영역에 집중되어 있고 다른 4개 대학은 두 영역을 모두 제공하고 있다. 국내의 기타 영역은 83.9%로 미국의 기타 영역 66.4%보다 높은 것으로 분석 되었다. 조사된 17개의 데이터 사어언스 프로그램 중 3개 대학이 학부과정에 데이터 사이언스 프로그램이 개설 되어 있고 14개 대학은 대학원 과정에 데이터 사이언스 프로그램이 개설되어 있다. 82%의 데이터 사이언스 과정이 대학원 중심으로 운영되고 있는 것으로 분석이 되었다. 국내 데이터 사이언스 개론 수업에 대한 강의계획서 분석 결과 미

국 데이터 사이언스 개론 수업과 수업방법, 평가방법에서 큰 차이점은 없는 것으로 분석되었다. 다만 국내외적으로 데이터 사이언티스트에 대한 수요가 급증하고 있는 상황에 비추었을 때 국내 데이터 사이언스 프로그램은 미국과 비교하면 양적으로 많이 부족한 상황이다. 국내 데이터 사이언스 프로그램은 Conway의 데이터 사이언스 3대 영역에 골고루 개설이 되어야 할 것이다. 3대 영역 중 특정분야 지식 영역은 데이터 사이언스 과정을 학부에서 제공하는 대학은 해킹 기술 또는 수학 및 통계 분야에 대한 교육뿐만 아니라 다전공 또는 부전공의 형태로 특정분야에 대한 지식도 같이 교육이 되도록 교과과정을 운영해야 할 것이다. 대학원에서 데이터 사이언스를 제공하는 대학은 특정분야 지식 영역을 가진 학생들이 대학원에서 해킹 기술과 수학 및 통계 지식에 대한 교과과정을 이수함으로써 3대 영역을 다양하게 갖춘 데이터 사이언티스트가 양성이 되도록 교과과정을 운영해야 할 것이다.

참 고 문 헌

김희섭, 남권희, 강보라. 2013. 북미지역 iSchool 대학과 L-School의 교육과정 비교분석.『한국문헌정보학회지』, 47(4): 295-314.

이성신, 최재황, 이창수. 2013. 한국형 iSchool 탐색.『한국도서관·정보학회지』, 44(4): 277-294.

이재윤. 2015.『데이터 사이언스와 데이터 리터러시』. Paper presented at the 제22회 한국정보관리학회 학술대회, 중앙대학교, 서울.

장덕현. 2015. iSchool 논의에 대한 비판적 담론분석.『한국도서관·정보학회지』, 46(1): 135-154.

장윤금. 2011. 문헌정보학 교육의 변화에 관한 국가 간 비교 연구.『한국비블리아학회지』, 22(4): 317-340.

장윤금. 2014. iSchool 대학의 발전, 교육 및 연구 동향 분석.『한국문헌정보학회지』, 48(1): 369-386.

최윤식. 2012.『미래학자의 통찰법』. 서울: 김영사.

Ahn, I. J. 2012. "The Specification of Science Education Programs in the Local Public Library: Focusing on the Programs In G-city." *International Journal of Knowledge Content Development & Technology*, 2(1): 17-35.

Brynko, B. 2012. "iSchools: shaping the information landscape." *Information Today*, 29(8): 1.

Bussaban, K. and P. Waraporn. 2015. "Preparing Undergraduate Students Majoring in Computer Science and Mathematics with Data Science Perspectives and Awareness in the Age of Big Data." *Procedia - Social and Behavioral Sciences*, 197: 1443-1446.

Chu, H. 2012. "iSchools and Non-iSchools in the USA: An Examination of Their Master's

Programs." *Education for information*, 29(1): 1-17.

Cleveland, W. S. 2001. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." *International Statistical Review*, 69(1).

Conway, D. 2010. The Data Science Venn Diagram [online]. [cited 2016.1.12]. 〈http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram〉.

Davenport, T. H. and D. J. Patil. 2012. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review*, 90(10): 70-76.

Gartner. 2011. Gartner Identifies the Top 10 Strategic Technologies for 2012 [online]. [cited 2016.3.10]. 〈http://www.gartner.com/newsroom/id/1826214〉.

Google Trends. 2016. Data Science and Big Data Trends [online]. [cited 2016.1.5]. 〈https://www.google.com/trends/explore#q=data%20science%2C%20Big%20Data&cmpt=q&tz=Etc%2FGMT-9〉.

Han, M. K. 2011. "A Study on the Suggestions and Analysis on the Education Programs in the Presidential Libraries." *International Journal of Knowledge Content Development & Technology*, 1(1): 39-60.

Jagadish, H. V. 2015. "Big Data and Science: Myths and Reality." *Big Data Research*, 2(2): 49-52.

Lipman, B. 2014. Universities Increasing Programs for Data Scientists [online]. [cited 2016.1.3]. 〈http://www.wallstreetandtech.com/careers/universities-increasing-programs-for-data-scientists/d/d-id/1318139〉.

Loukides, M. 2010. What is Data Science? [online]. [cited 2016.1.12]. 〈http://dmlab.uos.ac.kr/html/lecture/Database(2011-2)/strata2011_what-is-data-science_pdf.pdf〉.

Miller, C. C. 2013. "Data Science: The Numbers of Our Lives." *The New York Times*.

Provost, F. and T. Fawcett. 2013. "Data Science and its Relationship to Big Data and Data-Driven Decision Making." *Big Data*, 1(1): 51-59.

Shi, Y., P. S. Yu, Y. Zhu, and Y. Tian. 2014. "Explore New Field of Data Science Under Big Data Era: Preface for ICDS 2014." *Procedia Computer Science*, 30: 1-3.

SINTEF. 2013. Big Data, for better or worse: 90% of world's data generated over last two years [online]. [cited 2015012.12]. 〈https://www.sciencedaily.com/releases/2013/05/130522085217.htm〉.

Swanstrom, R. 2015. College and University Data Science Degrees [online]. [cited 2015.12.27]. 〈http://datascience.community/colleges〉.

• **국문 참고자료의 영어 표기**
(English translation / romanization of references originally written in Korean)

Chang, Yunkeum. 2011. "Comparative Study of the Changes in LIS Education in Korea, U.S.A. and Australia." *Journal of Korea Biblia Society for Library and Information Science*, 22(4): 317-340.

Chang, Yunkeum. 2014. "An Analysis of the Growth, Education and Academic Research Trends of iSchools." *Journal of Korean Society for Library and Information Science*, 48(1): 369-386.

Choi, Yoon Sik. 2014. *Insight into Futures*. Seoul: Gimmyoung.

Jang, Durk-Hyun. 2015. "iSchool Movement: A Critical Discourse Analysis." *Journal of Korea Library and Information Science Society*, 46(1): 135-154.

Kim, Heesop, Kwon Hee Nam, and Bora Kang. 2013. "A Comparative Analysis on Curriculum of iSchools and L-School in North America." *Journal of the Korean Society for Library and Information Science*, 47(4): 295-314.

Lee, Jae Yun. 2015. *Data science and data literacy*. Paper presented at the 22nd Korea Society for Information Management, Seoul.

Lee, Seongsin, Jae Hwang Choi, and Changsoo Lee. 2013. "An Exploratory Study On Korean iSchool." *Journal of Korea Library and Information Science Society*, 44(4): 277-294.

[부록 1] 국내 데이터 사이언스 교과과정 (가나다순)

대학교(교과목수)	교과과정
국민대학교 (27)	경영통계(Business Statistics) 데이터사이언스개론(Introduction to Data Science) 데이터베이스관리론(Database Management) R-프로그래밍(R-programming) 다변량통계분석(Multivariate Statistical Analysis) 데이터마이닝(Data Mining) 비즈니스서비스분석(Business Service Analysis) 빅데이터분산처리론(Big Data Distributed Processing) EDA와빅데이터시각화(EDA and Big Data Visualization) 선형통계분석론(Linear Statistical Analysis) 소셜네트워크분석(Social Network Analysis) 텍스트마이닝과 소셜애널리틱스(Text Mining and Social Analytics) 빅데이터통합과모델링(Big Data Integration and Modeling) 경영최적화와시뮬레이션(Business Optimization and Simulation) 비즈니스인사이트와의사결정(Business Insight and Decision Making) 연구방법론(Research Methods for Business) 빅데이터프로젝트(Big Data Project) 파이낸스어날리틱스(Finance Analytics) 경영정보시스템(Management Information System) e-비즈니스개론(Introduction to e-Business) 연구윤리와논문연구(Research Ethics & Thesis Study) 오퍼레이션애널리틱스(Operation Analytics) 마케팅애널리틱스(Marketing Analytics) 비즈니스모델및전략(Business Model & Strategy) 정보기술최신동향(Current Trends in Information Technology) 통계자료처리론(Analysis of Statistic Data) 데이터사이언스세미나(Research Seminar in Data Science)
단국대학교 (32)	경영정보시스템(Management Information System) 전사적 자원관리론(Enterprise Resource Planning) 경영의사결정론(Business Decision Making) 컴퓨터 프로그래밍(Computer Programming) 데이터베이스(Databases) 유닉스/리눅스(Unix/Linux) 선형통계분석(Linear Statistical Analysis) 데이터베이스 관리(Database Management) 분산 시스템(Distributed Systems) 정보 검색(Information Retrieval) 클라우드 컴퓨팅(Cloud Computing) 빅데이터 처리(Big Data Processing) 하둡 프로그래밍(Hadoop Programming) 자연어 처리(Natural Language Processing) 지식 표현과 모델링(Knowledge Representation and Modeling) 인공 지능 응용(Artificial Intelligence Application)

대학교(교과목수)	교과과정
	개인정보 보호(Information Privacy) 센서 네트워크(Sensor Networks) 이미지 처리와 인식(Image Processing and Recognition) 인간-컴퓨터 상호작용(Human-Computer Interaction) 경영최적화이론(Business Optimization Theory) 데이터 분석 및 기획(Data Analysis and Planning) 데이터 마이닝(Data Mining) 소셜 네트워크 마이닝(Social Network Mining) 기계 학습(Machine Learning) 다변량통계분석(Multivariate Statistical Analysis) 빅데이터 마케팅(Big Data Marketing) R 프로그래밍(R Programming) Hana 기술 응용(Application of Hana Technology) 시각화 기법(Visualization Techniques) 데이터 분석 사례 연구(Data Analysis Case Study) 데이터 분석 프로젝트(Data Analysis Project)
상명대학교 (23)	데이터사이언스개론(Introduction to Data Science) R기반 빅데이터 분석(Big Data Analytics using R) 빅데이터분석실습(Practice in Big Data Analytics) 자료구조(Data Structure) 운영체제(Operating System) 데이터베이스(Database) 인공지능(Artificial Intelligence) 데이터 모델링과 마이닝(Data Modeling and Mining) 프로그래밍 I(Programming I) 웹·UNIX기초(Introduction to Web and UNIX) 자바프로그래밍(Java Programming) 통계(Statistics) 웹프로그래밍 I(Web Programming 1) 정보통신기술론(Introduction to Information Technology) 빅데이터보안(Big Data Security) 경영통계(Business Statistics) 경영정보시스템(Management Information Systems) 경영빅데이터분석(Bigdata Analysis for Management) ICT융합과 시장창조(ICT Convergence & Market Development) 경영소프트웨어활용(Usage of Management Software) ICT활용 비즈니스 혁신(Business Innovation by Using ICT) e거버먼트와 정보보안(Theory of Electronic Government and Information Security) 조사방법과 데이터분석(Research Method and Data Analysis)
서울산업기술대학 (14)	데이터 사이언스 개론(Introduction to Data Science) 데이터 처리 언어(Analytical Languages: R & Python) 데이터 마이닝(Data Mining) 다변량 통계분석(Multivariate Statistics) 비정형 데이터 분석(Unstructured Data Analysis) 심화 기계 학습(Advanced Machine Learning) 비즈니스 어낼리틱스(Business Analytics)

대학교(교과목수)	교과과정
	데이터 시각화(Data Visualization) 사회연결망 분석(Social Network Analysis) 분산 데이터베이스 시스템(Distributed Database System) 데이터 사이언스 특론(Advanced Topics in Data Science) 빅데이터 보안(Security in Big Data) 제조업 사례연구(Case Study on Manufacturing) 서비스 사례연구(Case Study on Service)
서울대학교 (34)	현대정보검색 개론(Introduction to Modern Information Retrieval) 정보추구행동론(Human Information Behavior) 컴퓨터프로그래밍 개론(Introduction to Computer Programming) 웹 응용 시스템(Web Application System) 온톨로지기술 입문(Introduction to Ontology Technology) 지식정보처리론(Knowledge Information Processing) 복잡계 이론 및 디지털 정보 분석(Digital data analysis in complex systems) 정보행동 연구조사 방법론(Research method in information user analysis) 비주얼라이제이션(Visualization) 프로젝트 기획과 실제(Project Management) 게임의 이해(Understanding Games) 디지털스토리텔링(Digital Storytelling) 정보정책(Information Policy) 산학연구 프로젝트(Field Project) 음악정보검색 입문(Introduction to Music Information Retrieval) 모바일 응용 프로그래밍(Mobile Application Programming) 과학기술사(History of Science and Technology) 인터랙티브 사운드 디자인(Interactive Sound Design) 음악정보검색 시스템(Music Information Retrieval Systems) 기업가적 사고방식(Entrepreneurial Mindset) 대학원논문연구(Dissertation Research) 데이터 분석 개론(Introduction to Data Analysis) 정보융합 기계학습(Machine Learning for Information Studies) 소셜컴퓨팅(Social Computing) 정보융합 세미나(Seminar in Information Studies) 인간컴퓨터상호작용 연구특강(Research Topics in Human Computer Interaction) 정보융합 통계분석(Statistical Analysis for Information Studies) 인간컴퓨터상호작용연구(Human Computer Interaction Research) 게임과 콘텐츠 특강(Topics in Game and Contents) 정보융합 특강(Topics in Information Studies) 융합과학기술개론(Introduction to Convergence Science and Technology) 융합 지식의 실무 응용 1(Field Applications of Convergence Knowledge 1) 융합 지식의 실무 응용 2(Field Applications of Convergence Knowledge 2) 융합 프로젝트 설계(Interdisciplinary Project Design)
성균관대학교 (34)	데이터사이언스개론(Introduction to Data Science) 데이터베이스설계론(Designing Database Systems) 디지털 인문학(Digital Humanities) 프로그래밍언어론(Programming Languages) 통계자료관리(Statistical Data Management)

대학교(교과목수)	교과과정
	컴퓨터네트워크(Computer Networks) 정보행위론(Information Behavior) 정보학개론(Introduction to Information Science) 정보조직론(Introduction to Information Organization) 정보분류론(Introduction to Information Classification) 정보검색론(Information Retrieval) 소비자가족질적방법(Qualitative Method for Consumer) 소비자가족양적방법(Quantitative Method for Consumer) 메타데이터론(Principles of Metadata) 마케팅관리(Marketing Management) 독어학개론II(Introduction to German Linguistics II) 경영정보시스템(Management Information Systems) 통계적추론입문(Introduction to Statistical Infererence) 통계적데이터마이닝(Statistical Data Mining) 컴퓨터그래픽스(Computer Graphics) 정보분석평가론(Information Analysis and Evaluation) 인터넷프로그래밍 실습(Principles and Practice in Internet Programming) 인터넷서버구축론(Internet Servers Administrations 인터넷과 경영(Internet and Management) 의학정보개론및처리론(Introduction to Health Science) 응용독어학(Applied German Linguistics) 시맨틱웹시스템구축실습(Designing Semantic Web System) 소셜데이터분석실습(Social Data Analytics) 소비자조사연구(Consumer Survey & Analysis) 소비자니즈분석(Analysis of Consumer Needs) 비주얼프로그래밍(Visual Programming) 데이터시각화실습(Practice in Data Visualization) 다변량통계분석입문(Introduction to Multivariate Statistics) 건강정보시스템실습(Practice in Health Information)
연세대학교 (17)	소셜미디어분석론 비즈니스환경과 소셜네트워크 IT 비즈니스 다이나믹스 디지털 비즈니스 전략 빅데이터 관련 법과 제도 빅데이터컴퓨팅 대용량 멀티미디어 자료처리 데이터 프로그래밍 실무 빅데이터 처리 표준 및 제도 비즈니스환경에서의 통계적 분석론 R-데이터마이닝 소셜애널리틱스 데이터최적화와 시뮬레이션 빅데이터 분석 및 응용 연구조사방법론 산업별 빅데이터 응용 빅데이터 의사결정모형

[부록 2] 미국 데이터 사이언스 교과과정 (ABC 순)

대학교(교과목수)	교과과정
Columbia University (7)	Probability Algorithms for Data Science Statistical Inference and Modeling Computer System for Data Science Machine Learning for Data Science Exploratory Data Analysis and Visualization Data Science Capstone and Ethics
Indiana University, Bloomington (18)	Big Data Applications and Analytics Management, Access, and Use of Big and Complex Data Introduction to Statistics Database Theory and System Design Information Visualization Scientific Data Management and Preservation Data Curation Organizational Informatics and Economics of Security Big Data Open Source Software and Projects Analysis of Algorithm Advanced Database Concepts Introduction to Statistics Cloud Computing Search Machine Learning Data Mining Security for Networked Systems Information Visualization
New York University (6)	Intro to Data Science Statistical and Mathematical Methods for Data Science Machine Learning and Computational Statistics Big Data Inference and Representation Capstone Project in Data Science
Rutgers University (5)	Fundamental of Analytics Advanced Analytics & Practicum Regression Analysis Database Design and Management Cloud Computing & Big Data
Southern Methodist University (10)	Experimental Statistics I Doing Data Science Experimental Statistics II File Organization & Database Management Data and Network Security Visualization of Information Data Mining Statistical Sampling Quantifying the World Immersion

대학교(교과목수)	교과과정
Stanford University (15)	Linear Algebra Numerical Optimization Discrete Mathematics Advanced Programming for Scientists and Engineering Software Design in Modern Fortran for Scientists and Engineering Computer Organization and Systems Large Scale Software Development Introduction to Parallel Computing using MPI, openMP and CUDA Introduction to Statistical Inference Regression Models / Statistical Modeling Modern Applied Statistics: Learning Modern Applied Statistics: Data Mining Parallel and Distributed Data Management Social and Information Network Analysis
University of California, Berkeley (11)	Research Design and Applications for Data and Analysis Exploring and Analyzing Data Storing and Retrieving Data Applied Machine Learning Data Visualization and Communication Capstone Behind the Data: Humans and Values Experiments and Causal Inference Scaling Up! Really Big Data Machine Learning at Scale Applied Regression and Time Series Analysis
University of Michigan (5)	Discrete Mathematics Programming and Elementary Data Structures Data Structures and Algorithms Introduction to Probability and Statistics Applied Statistics I
University of Virginia (9)	Programming and Systems for Data Science Statistical Computing for Data Science Linear Models for Data Science Foundations of Computer Science Data Mining Practice and Application of Data Science Ethics of Big Data Capstone Project Machine Learning
University of Wisconsin - Madison (6)	Data Practicum Interdisciplinary Literacy/Professional Competencies Introduction to Statistical Inference Mathematical Statistics I Methods I and II Professional Development as a Statistician

[부록 3] 국내 데이터 사이언스 교과과정 강의계획서

Course Information

Professor:
Class Website:
Email:
Phone:
Office:
Office hours: by Appointment

Course Hours/Location

Course Description

This course serves as an introduction to the interdisciplinary and emerging field of big data. Students will learn to use concepts, tools, and techniques from business, computer science, media software, and library and information science to solve problems using data. Topics and tools will include RelFinder, Open Refine, Data Visualization, Data Quality Controls, and various case studies.

Course Objectives

Learn about what it's like to be a Big Data Scientist
Be able to do some of what a Big Data Scientist does
To enable students to learn and apply basic concepts of data collect, refine, and visualize using various tools.

Course Textbook

Data Science for Business
Foster Provost and Tom Fawcett
ISBN-13: 978-1449361327

An Introduction to Data Science
Jeffrey Stanton
You can download here.
https://itunes.apple.com/us/book/introduction-to-data-science/id529088127?mt=11

Course Grading

Assignments	Percentages	Due Date
Individual Project	15%	
Group Project	35%	
Midterm Exam	20%	
Final Exam	20%	
Class participation	10%	Through-Out

Course Assignments

Individual Project (More detail will be provided)
Group Project (More detail will be provided)

Course Policy

Attendance Policy
Consistent and attentive attendance is vital to academic success, and is expected of all students. Grades are determined by academic performance, and instructors may give students written notice that attendance related to specific classroom activities is required and will constitute a specific percentage of students' grades.

Instructors are strongly encouraged to keep a record of student attendance. They should note absences due to documented student illness, serious illness or death in the student's immediate family, official school activity, nation-recognized religious holiday, active military service that is of a reasonable brief duration, or other verified absences deemed appropriate by the instructor. Students must consult with instructors regarding the completion of make-up work.

Absences do not exempt students from academic requirements. Excessive absences, even if documented, may result in a student failing the course. An incomplete may be granted if the student has a passing grade, but only if the instructor determines that it is feasible for the student to successfully complete remaining assignments after the semester.

Late Assignments
Students are responsible for submitting assignments on the date due. Late assignments will be reduced one letter grade for each module missed. If you miss the due date, the highest

grade you can achieve is a 79, and so on.

Communication Policy

Given the important role of communication in instructional quality, I will respond to course-related voice mail or email within 4 business days and will return or give an update on course assignments within 3 working weeks of the due date. Please use the Questions forum for questions when at all possible, especially when other students might benefit from the answer, and save email for private questions. However, when responding via email, please be sure and include all prior pertinent email information for my reference. Note: I do not monitor the discussion board or email 24/7 or weekends; therefore, do not wait until the last minute to email me a question and then expect an immediate response. Please include "**Course Title**" **and** "**your full name**" **on your e-mail subject to bypass my spam filtering rules.**

Academic Integrity

Honesty in completing assignments is essential to the mission of the University and to the development of the personal integrity of students. Cheating, plagiarism, fabrication or other kinds of academic dishonesty will not be tolerated and will result in appropriate sanctions that may include failing an assignment, failing the class, or being suspended or expelled.

Course Calendar

Week	Tuesday	Thursday
1	Getting Acquainted Course Introduction - Syllabus	J-Ch 0. Data Science: Many Skills
2	O-Ch 1. What is Data Science?	O-Ch 1. What is Data Science?
3	J-Ch1. About Data: Unstructured vs. Structured Data	F-Ch1. Data-Analytic Thinking
4	J-Ch 2. Identifying Data Problems	F-Ch 2. Business Problem and Data Science Solutions
5	J-Ch 8. Big Data? Big Deal!	J-Ch 8. Big Data? Big Deal!
6	O-Ch 7. Extracting Meaning from Data	J-Ch 7. Extracting Meaning from Data
7	O-Ch 9. Data Visualization and Fraud Detection	O-Ch 10. Social Networks and Data Journalism
8	Miderm	
9	Field Trip: NIA Big Data Center	Field Trip: NIA Big Data Center
10	Special Topic: Linked Data	Special Topic: Linked Data
11	O-Ch 10. Social Networks and Data Journalism	Special Topic: Social Media Analysis and Marketing
12	O-Ch 14. Data Engineering	F-Ch 13. Data Science and Business Strategy
13	Special Topic: Big Data Security	Special Topic: Big Data Security
14	Special Topic: Data Mining	Special Topic: Big Data Standard
15	O-Ch 16. Next Generation Data Scientists, Hubris, and Ethics	O-Ch 16. Next Generation Data Scientists, Hubris, and Ethics
16	Final Exam	

[부록 4] 미국 데이터 사이언스 교과과정 강의계획서

Professor	
Office: Hours	
Email	
Telephone	
Classroom	
Class time	
First/Last Class	
Final Quiz	
Course Assistants CA Office Hours	

1. Course Overview

This course will change the way you think about data and its role in business.

Businesses, governments, and individuals create massive collections of data as a byproduct of their activity. Increasingly, decision-makers and systems rely on intelligent technology to analyze data systematically to improve decision-making. In many cases automating analytical and decision-making processes is necessary because of the volume of data and the speed with which new data are generated.

We will examine how data analysis technologies can be used to improve decision making. We will study the fundamental principles and techniques of data mining, and we will examine real-world examples and cases to place data-mining techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science. In addition, we will work "hands-on" with data mining software.

After taking this course you should:

1. Approach business problems data-analytically. Think carefully & systematically about whether & how data can improve business performance, to make better-informed decisions for management, marketing, investment, etc.

2. Be able to interact competently on the topic of data mining for business intelligence. Know the fundamental principles of data mining, that are the basis for data mining processes, algorithms, & systems. Understand these well enough to interact with CTOs, expert data miners, consultants, etc. Envision opportunities.

3. Have had hands-on experience mining data. Be prepared to follow up on ideas or opportunities that present themselves, e.g., by performing pilot studies.

2. Focus and interaction

The course will explain through lectures and real-world examples the fundamental principles, uses, and some technical details of data mining and data science. The emphasis primarily is on understanding the fundamental concepts of data science and business applications of data mining. We will discuss the mechanics of how the methods work as is necessary to understand and illustrate the fundamental concepts and business applications. This is not an algorithms course. However, many techniques are the embodiment of one or more of the fundamental principles.

I will expect you to be prepared for class discussions by having satisfied yourself that you understand what we have done in the prior classes. The assigned readings will cover the fundamental material. The class meetings will be a combination of lectures/discussions on the fundamental material, discussions of business applications of the ideas and techniques, case discussions, and student exercises.

You are expected to attend every class session, to arrive prior to the starting time, to remain for the entire class, and to follow basic classroom etiquette, including having all electronic devices turned off and put away for the duration of the class (this is Stern policy, see below) and refraining from chatting or doing other work or reading during class. In general, we will follow Stern default policies unless I state otherwise. I will assume that you have read them and agree to abide by them: http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511

The Blackboard site for this course will contain lecture notes, reading materials, assignments, and late-breaking news. You should check the Blackboard site daily, and I will assume that you have read all announcements and class discussion.
If you have questions about class material that you do not want to ask in class, or that would take us well off topic, please detain me after class, come to office hours to see me or the TAs, or ask on the discussion board. The discussion board is much better than sending me email, which frankly I have a hard time keeping up with. Also, if you have the question, someone else may too and everyone may benefit from the answers being available on Blackboard. Also,

please try to answer your classmates' questions. In grading your class participation I will include your contributions to the discussion board. You will not be penalized for being wrong in trying to participate on the discussion board (or in class).

Worth repetition: It is your responsibility to check Blackboard (and your email) at least once a day during the week (M-F), and you will be expected to be aware of any announcements within 24 hours of the time the message was sent.
I will check my email at least once a day during the week (M-F). Your email will get my priority if you include the special tag [DM Grad] in the email subject header. I use this tag to make sure to process class email first. If you do not include the special tag, I may not read the email for a while (maybe a long while). If you forget and send without the tag and then remember, just send it again including the tag.

3. Lecture Notes and Readings

Book: The textbook for the class will be:
Data Science for Business: Fundamental principles of data mining and data analytic thinking Provost & Fawcett (O'Reilly, 2013).
I will give you a copy of this book in class. I wrote it over the past couple years, in response to feedback from this course—in particular, that the available books were not adequate. This book covers the fundamental material that will provide the basis for you to think and communicate about data mining for business analytics. We will complement the book with discussions of applications, cases, and demonstrations.

Lecture notes: For many classes I will hand out lecture notes. I intend that the notes on the fundamental material will follow the book very closely. I expect you to ask questions about any material in the notes that is unclear after our class discussion and reading the book. I wrote the book to follow the class closely, and to free us up for more discussion of applications, etc.—so many of your questions may be answered in the book. If not, please let me know! Depending on the direction our class discussion takes, we may not cover all material in the class notes for any particular session. If the notes and the book are not adequate to explain a topic we skip, you should ask about it on the discussion board. I will be happy to follow up.

I will hand out or post some additional required readings as we go along. Note that some of

these readings may be accessible for free only from an NYU computer. If you can't access a link from home, please try it from school.
For those interested in going further, these following supplemental books give alternative perspectives on and additional details about the topics we cover. These are completely optional; you will not be required to know anything in these readings that are not in the primary materials or lectures. I have many other books that I can recommend, for example if you want a reference to a more mathematical treatment of the topics. Please don't hesitate to come and talk to me about what supplemental material might be best for you, if you want to go further.

- Supplemental readings: posted to blackboard or distributed in class; labeled as supplemental/optional

- Supplemental book (optional):
  Data Mining Techniques, Second Edition
  by Michael Berry and Gordon Linoff, Wiley, 2004
  ISBN: 0-471-47064-3
- available as ebook for free: http://site.ebrary.com/lib/nyulibrary
- Many students find this book to be an excellent supplemental resource
- The Third Edition just came out in the past year. I have not read it yet. Berry says it has been improved substantially. I have a copy in my office if you want to talk a look at it before buying it.
- available from Amazon

- "Weka Book" (optional):
  Data Mining: Practical Machine Learning Tools and Techniques, Third Edition
  by Ian Witten, Eibe Frank, Mark Hall
  ISBN-10: 0123748569
  o available from Amazon
  o This book provides much more technical details of the data mining techniques and is a very nice supplement for the student who wants to dig more deeply into the technical details. It also provides a comprehensive introduction to the Weka toolkit.

4. Requirements and Grading

The grade breakdown is as follows:

1. Homeworks: 20%
2. Term Project: 30%
3. Participation & Class Contribution: 20%
4. Final Quiz: 30%

At NYU Stern we seek to teach challenging courses that allow students to demonstrate differential mastery of the subject matter. Assigning grades that reward excellence and reflect differences in performance is important to ensuring the integrity of our curriculum. In my experience, students generally become engaged with this course and do excellent or very good work, receiving As and Bs, and only one or two perform only adequately or below and receive C's or lower. Note that the actual distribution for this course and your own grade will depend upon how well each of you actually perform this particular semester.

Homework Assignments

The homework assignments are listed (by due date) in the class schedule below. Each homework comprises questions to be answered and/or hands-on tasks. Except as explicitly noted otherwise, you are expected to complete your assignments on your own－without interacting with others.

Completed assignments must be handed on blackboard at least one hour prior to the start of class on the due date (that is, by 5pm), unless otherwise indicated.

Assignments will be graded and returned promptly. Answers to homework questions should be well thought out and communicated precisely, avoiding sloppy language, poor diagrams, and irrelevant discussion.

The hands-on tasks will be based on data that we will provide. You will mine the data to get hands-on experience in formulating problems and using the various techniques discussed in class. You will use these data to build and evaluate predictive models.

For the hands-on assignments you will use the (award-winning) toolkit Weka, part of the Pentaho open source business intelligence suite:

http://www.cs.waikato.ac.nz/ml/weka/ download the "latest stable" version (3.6.9) (which is

the version associated with the 3rd edition of the Weka Book) http://www.pentaho.com

IMPORTANT: In order to use Weka you must have access to a computer on which you can install software. If you do not have such a computer, please see me immediately so we can make alternative arrangements. You should bring your computer to the second class. During the class we will have a "lab session" during which we will install and configure the software, get it running, and dealing with the inevitable glitches that a few of you might experience. If you need additional help with using the data mining software, please see the Course Assistant. Generally the Course Assistant should be the first point of contact for questions about and issues with the homeworks. If they cannot help you to your satisfaction, please do not hesitate to come see me.

Late Assignments

As stated above, assignments are to be submitted on Blackboard at least one hour prior to the start of the class on the due date. Assignments up to 24 hours late will have their grade reduced by 25%; assignments up to one week late will have their grade reduced by 50%. After one week, late assignments will receive no credit. Please turn in your assignment early if there is any uncertainty about your ability to turn it in on time.

Term Project

A term project report will be prepared by student teams. We will give you the instructions on how to form your teams. Teams are encouraged to interact with the instructor and TA electronically or face-to-face in developing their project reports. You will submit a proposal for your project about half way through the course. Each team will present its project at the end of the semester. We will discuss the project requirements and presentations in class.

Final Quiz

The final quiz will be a take-home to be completed during the week following the last class. The subject matter covered and the exact dates will be discussed in class.

Participation/Contribution/Attendance/Punctuality

Please see Section 2.

Regrading

If you feel that a calculation, factual, or judgment error has been made in the grading of an assignment or exam, please write a formal memo to me describing the error, within one week after the class date on which that assignment was returned. Include documentation (e.g., pages in the book, a copy of class notes, etc.). I will make a decision and get back to you as soon as I can. Please remember that grading any assignment requires the grader to make many judgments as to how well you have answered the question. Inevitably, some of these go "in your favor" and possibly some go against. In fairness to all students, the entire assignment or exam will be regraded.

FOR STUDENTS WITH DISABILITIES: If you have a qualified disability and will require academic accommodation during this course, please contact the Moses Center for Students with Disabilities (CSD, 998-4980) and provide me with a letter from them verifying your registration and outlining the accommodations they recommend. If you will need to take an exam at the CSD, you must submit a completed Exam Accommodations Form to them at least one week prior to the scheduled exam time to be guaranteed accommodation.