The Rise of Conversational Data Miners: From Conversational To Professional

Hanhan Wu
Computing Science
Simon Fraser University
Burnaby, Canada
hanhanw@sfu.ca

INTRODUCTION

In recent years, "Data Mining" has become a popular topic in almost all the research areas and industries [1]. Data Mining activities include extracting implicit, previously unknown, and potentially useful information from data. The extracted data is used to build computer programs that sift through data sources automatically, seeking regularities or patterns to help enhance business [2]. Knowledge such as applied statistics, machine learning, data engineering, etc. are required for professional data miners [3].

Every organization generates data, and experts use data mining technologies to deliver important and useful information to facilitate performance analysis of business organizations. More and more requirements of this process lead to a rapidly growing demand for data mining experts who unfortunately, are in short supply. Reports suggest that by 2018, a shortage of data scientists will emerge, ranging from 140,000 to 190,000 in the U.S. alone [4]. [5] also gives market projections for big data technology and services and envisages a growth from €1.8 billion ($2.0 billion) in 2013 to €5.3 billion in 2018 in Europe. To meet this continuously growing market demand, more schools and online classes are offering courses and programs to prepare people for the data mining industry [6].

In order to meet the demand of professional data miners, education programs are adding data mining courses or data science materials in other classes, such as Astronomy [7]. However, in modern professional workplace, a project requires the collaboration of experts from multiple roles. For example, a professional data miner needs to collaborate with business project managers, data engineers, domain experts and so on. [8] indicates that better collaboration comes from more grounding in the communication, which proves that, in order to work better for data mining projects, people in other roles should also have the basic ideas of data mining. This paper uses "conversational data miners" to describe people who need to understand data mining basic knowledge such as statistics methods, machine learning algorithms, data engineering and relative software development process, but they do not need to dive deeper into the data work.

This study focuses on a group of people who were working as conversational data miners but now they are learning data mining in order to become professional data miners. The study purpose is to understand the learning situation of this group, including: what motivated them to be professional data miners, what do they want to learn and what do they need to learn, how do they learn and their learning barriers.

In this study, I interviewed two groups of the Simon Fraser University (SFU) Big Data students. The first group has 5 newest cohort students, who worked as conversational data miners before they joined SFU. The second group has 5 SFU Big Data new graduates, who were working as conversational data miners before coming to SFU and will become entry-level professional data miners from 2017 May. While the first group can provide information about their current learning situation, the second group can reflect on their learning path for becoming professional data miners from the conversationals and provide suggestions for both conversational data miners and current data mining education programs.

The main contribution of this research work is to provide empirical insights of the learning situation of a group of conversational data miners who plan to become professional data miners. Based on all the findings, the discussion section will also provide practical suggestions for data mining education programs.

RELATED WORK

To contextualize the insights about conversational data miners, I draw up research from data mining education in multiple areas such as Statistics, Computer Science, Engineering, Science, Art and online education that serves for general population.

Statistics Students Focus

A majority research of the education in data mining comes from Statistics. In Stephanie's Guide to Teaching Data Science [9], they have shared the general principles and detailed guidance for teaching data mining courses, which derived from their successful experience. Baumer teaches students how to think with data to deal with real world data mining problems [10]. Other studies talk about how to train students with practical skills such as programming in order to prepare them better for the real data mining world [11, 12, 13, 14, 15]. However, all of their teaching targets are Statistical students, who have the knowledge background of statistics and machine learning required by data mining.

Data Mining Education for STEM Students

The second majority data mining teaching guidance comes from Computer Science, Engineering and Science. In Computer Science, data mining teaching tends to focus on a specific area, such as database design and management [16, 17, 18], data modeling [19, 20, 21], and teaching ethical considerations in data mining courses [22]. In Science field, such as Physics, Astronomy and Bioinformatics, data mining courses are no longer optional for their students, researchers are talking about how to help students gain practical skills in data mining, such as applying machine learning in Astronomy Physics class [23]. However, STEM students all have more trainings in mathematics, which will help them understand statistics easier.

Data Mining Education in Non-STEM Area

Besides Statistics and STEM areas, data mining is in fact playing a significant role in non-STEM areas, such as Art. Artists are applying machine learning to create music robots [24] or to find music patterns in order to generate more creative music [25]. Their major learning resources are textbooks and online resources, but Art education programs are lack of hands-on training [25].

Data Mining Education for General Population

In order to serve for larger population in learning data mining, many courses are becoming online to enable both professional and conversational data miners to enroll. The most popular online education method is massive open online courses (MOOCs). In Simon's survey about data mining in MOOCs [26], they indicates that data mining has already seamless embedded in MOOC ecosystem and the community helps learners to connect to each other, but its major education content is videos without considering different learning styles, it measures learners' achievements without adjusting the difficulty level for each individual learner. Meanwhile, other online courses are trying to ease the data mining learning experience, such as standardizing the system setup [27], or focusing on data preparation, processing and visualization, without requiring students have any data mining knowledge background [28]. But these courses are still struggling to balance the teaching materials for the population with a wide variety knowledge background.

To sum up, current education research has made a number of recommendations for data mining training. However, most of them are focusing on Statistics, STEM students who are getting education in statistics, machine learning, mathematics and programing skills required by data mining area. Non-STEM students who need to apply data mining are lack of in-person education but search for help from online resources, while MOOCs and other online data mining courses are having difficulties to address various individual learning requirements. In order to fulfill some education gap, this study focuses on providing empirical insights of the learning situation of conversational data miners who did not get the same data mining trainings as Statistics or STEM students but plan to become professional data miners.

METHOD

Research Site

This study interviewed students from SFU Big Data Program. SFU Big Data is a graduate program started from 2015 September, aiming at creating professional data miners for the industry. Students have to pass courses and projects in Data Mining, Machine Learning, 2 Big Data Science Lab Experiments, Algorithms, Operating Systems and 1 data mining related course such as Deep Learning or Crowd Sourcing. Meanwhile, they have to gain at least 4 months data mining co-op work experience.

In 2017, it is still a new data mining education program and the students background varies. 10% students have decades work experience in data engineering, project management and information engineering. 20% students are newly graduated CS students who had neither work experience nor data mining training. The rest 70% students had at least 2 years work experience in business analysis, information systems, software development but they had none or few data mining knowledge before Joining SFU Big Data.

The 10 interview participants of this study all worked as conversational data miners before joining SFU Big Data, they had none or less knowledge in statistics, machine learning, data engineering but had to work with professional data miners. They are all at the age between 20 and 35 years old, 3 female and 7 male.

Group A has 5 newest cohort students. 2 worked as Business Analysts, 2 worked as web & mobile developers and 1 worked as software developer. They all plan to become data scientists after school graduation.

Group B had 5 new graduates. Before joining SFU Big Data, 2 of them worked as Business Analysts while the other 3 worked as web & mobile developers. They all have done 8 months data science co-op and from 2017 May, 4 of them will work as full time data scientists and 1 will work as full time big data engineer.

Interview Design and Setup

All the questions are open ended. A question can be sightly changed based on the observations from previous interviews and also depends on the interviewee's background, such as work experience and academic knowledge.

A consent form is given to each interviewee to sign, helping them make sure that the interviews will not influence their academic results, and it will respect their data privacy as well as their choice to stop the interview at any time. Meanwhile, a meeting room near their lab at SFU is booked to make them feel comfortable as well as convenient. Moreover, the definitions of "data miner", "data mining", "conversational data miners" and "professional data miners" are explained in detail to guarantee everyone is having the same concepts in mind.

Questions for Group A - Conversational Data Miners

The 1, 2 questions are to understand their academic or work background, as well as their experience as conversational data miners, in order to help later questioning.

1. Could you briefly talk about your background before joining SFU Big Data?

2. How was your experience when you were working as a conversational data miner?

The 3 question is to understand their motivations to become professional data miners.

3. What made you decide to become a professional data miner?

The 4, 5, 6, 7 questions are designed to understand what do they want to learn, how do they learn and the learning barriers with conversational data miner background.

4. What do you want to learn in order to become professional data miners?

5. How do you plan to achieve your learning goals? Have you tried any learning materials or education program?

6. What are the most important skills you want to learn in data mining?

7. Do you have any barrier when you are learning?

The 8 is to learn whether they have any suggestion for data mining education programs.

8. Do you have any suggestion for data mining education programs, either online or in-person program?

Questions for Group B - Professional Data Miners

The 1 question is designed to basically understand their professional data mining work experience to help further questioning.

1. Could you briefly talk about your recent 8 months work experience that related to data mining?

The 2, 3, 4 question is to understand their learning path from conversational data miners to professional data miners.

2. What are the knowledge and skills help your professional data mining work?

3. How did you learn those knowledge or skills?

4. What are the most important data mining knowledge or skills in your opinion?

The 5, 6 questions are to understand their suggestions for people who will share similar learning path and suggestions for education programs.

5. About learning, what do you recommend to those conversational data miners who also want to become professional data miners?

6. Do you have any suggestion for current data mining education programs, either online or in-person program?

ANALYSIS AND RESULTS

During the interview, I took detailed notes of what the participants talked. During the analysis stage, I coded each answer with "Plan", "Action", "Strategy" or "Result", then compared the code from similar interview questions and categorized the code.

The results presented below are the major patterns in answering the research questions and participants' suggestions to data mining education programs.

Motivations to be The Professionals

Among all the 5 conversational data miners, 3 of them chose to become professional data miners because they thought it would be easier to find a job and get higher payment. The other 2 chose this path because they realized that they love data work and plan to learn deeper in order to become data scientists.

What To Learn vs What Need To Learn

I categorized the code from conversational data miners to understand what do they want to learn, then compared with the code from professional data miners about what need to learn. Their answers are very similar and professional data miners added more practical insights.

A. From Conversational Data Miners - What To Learn

All the 5 conversational data miners mentioned that statistics and machine learning are the most important knowledge they need. Practical skills are also important, such as big data development with Spark, data analysis with Python are all what the industry needs, and therefore will help them find a better job.

1. *...I think deeper understanding of statistics and machine learning will help me become a data scientist. The reason I chose SFU Big Data to help me achieve my goal was not only it has machine learning, data mining courses, but also it has many hands-on lab experiments with popular tools that the industry want, such as Spark, Hadoop and Python… (C2)*

2. *...I really like the 2 lab courses, we are learning everything that the job market requires. We are also learning data mining, machine learning, these were the knowledge I was lack of in order to become a data scientist. Besides what we are learning, I also want to learn more about statistics… (C4)*

B. From Professional Data Miners - What Need To Learn

While all the 5 professional data miners also mentioned the importance of hands-on big data development, analysis skills such as Spark and Python, as well as the knowledge in statistics and machine learning, they also mentioned other skills such as business communication and data visualization:

1. *...During my 8 months data scientist co-op work, what we have learned in big data analysis with Spark, machine learning helped so much, I also tried to learn deeper about statistics which helped my real world data analysis a lot. Also I have found that in the real world workplace, we have to collaborate with people in different areas, learning how to communicate with them well and get clear business requirements are also important… (P2)*

2. *...The lab courses were helpful in my data analysis, I use Python everyday, machine learning and data mining courses are also good. I also need data visualization skills because those business people just understand visualization, how to make my model clear to them was very important… (P5)*

To summarize, they all think statistics, machine learning knowledge and hands-on data analysis skills are important to help a conversational data miner become a professional. With industry data mining experience, the professional data miners also added business communication and data visualization as what need to learn.

How Do They Learn

Both conversational and professional data miners shared their methods in learning knowledge and skills they need. The methods can be categorized into online learning and in-person learning.

Among 5 conversational data miners, SFU Big Data is the only in-person data mining education program they have attended. 3 of them tried to learn machine learning online courses and data mining online tutorials before they joined SFU. At SFU, 2 of of these 3 people tried online courses in statistics and the other one always practices through Kaggle data analysis competition. Other 2 people have never tried online learning materials.

For all the 5 professional data miners, SFU Big Data Program is also the only in-person data mining education program. They all frequently used online learning resources in order to become professional data miners.

A. Online Learning Methods

The online learning materials they use can be summarized into online courses, tutorials, data analysis competition. For all the 3 conversational data miners and 5 professional data miners who use online learning resources, they all emphasis on how did they learn from Coursera, EdX and Webinar.

However, they have quite different online learning style.

Some prefer documentation format:

1. *... I hate video format and prefer reading documentation. Some videos are too fast that I could not follow, some are too slow, when I tried to skip a part I missed some important concepts. Reading documentation allows me to follow at my own pace… (C2)*

2. *...I prefer documentation more than videos, reading documentation is faster and it is more searchable… (P1)*

Some prefer video format:

3. *...I really love video format courses, they give more visual sense to me and they are very interactive, I can answer the questions after a certain section and it does help me understand deeper… (C4)*

4. *...I used online courses a lot, they have perfect UI… (P3)*

The 3 conversational data miners all have mentioned some common online courses such as Coursera Machine

Learning taught by Andrew Ng, but they are showing different understanding levels:

1. *...I tried Stanford Machine Learning through Coursera, I could follow everything he said and the quiz helped me understand those concepts deeper… (C4)*
2. *...I had difficulty to understand Andrew Ng's machine learning class, sometimes I had to pause the video to search for the meaning of a terminology, and sometimes, it's still difficult to understand the content… (C3)*

B. In-person Learning Methods

When talking about in-person learning program, they all emphasized on the importance of peer discussion in in-person learning. For example:

1. *…Sometimes, the class materials were difficult for me to understand, but later my classmates and I discussed a lot, which made a difference … (C1)*
2. *…When I was sitting in the lab, some experiments were difficult. Exchanging ideas with friends finally helped me finish all the work… (C5)*

To summarize how the 10 participants learn in order to become professional data miners from the conversationals. 8 of them use online learning materials, majorly are online courses and they are showing different learning style as well as understanding levels. SFU Big Data is the only in-person data mining program they attend and peer discussion plays an important role in their learning.

Learning Barriers

The learning barrier majorly come from online learning materials, and the participants' answers are all about finding the right learning resources. The cause of this difficulty varies, for example:

Some online content are not searchable.

1. *...Many video content are not searchable, websites only use the video title for people to search, it is difficult for me to find the right content… (C4)*

The terminology used in the learning material creates difficulty.

2. *...Different online papers and tutorials can use different terminology to express the same thing, it took me very long time to figure that out, sometimes I simple totally got confused… (C2)*

The content assumes the audience has relative knowledge background such as statistics, while in fact these conversational data miners do not:

3. *...I had difficulty in following some machine learning courses, because I could not understand the statistics behind. They simply use the formula and the theory directly without further explanation… (C1)*

Suggestions to Data Mining Education Programs

Both conversational and professional data miners talked about their suggestions for the improvement of data mining education programs.

A. Suggestions To Online Education Programs

8 participants who use online learning materials all hope the online content can do more to help them understand the content, such as describe the terminology or the statistics knowledge behind:

1. *... I hope they will have some documentation to describe all the terminology and the statistics theory used in the teaching materials, because sometime even when I tried to search for those things, I still could not understand the content well… (C2)*
2. *...Describe why do they use those statistics methods, such as why do they talk about using chi-square distribution in the analysis… (C3)*

3 participants mentioned the content should be separated into different levels based on students' knowledge background.

1. *...I didn't have too much statistics knowledge, but those video kept using the terminology. That works for statistics students, but not me… (C3)*

B. Suggestions To In-person Education Programs

All the 10 participants strongly recommend to add statistics in SFU Big Data program, meanwhile 3 of them also mentioned adding data visualization course.

Research On Online Courses

There is no study target conversational data miners' learning and the interview participants majorly have feedback on online learning resources. Therefore I have also researched all the 3 online course systems they mentioned and summarized the features as well as limitations.

A. Coursera

1. Interactive video format, with questions in the video.
2. Some video courses have document format.
3. Students have to pass assignments and projects in order to get the certificate.

B. EdX

1. It is using YouTube video, which is less interactive.
2. No document format.
3. It also gives certificate when students passed the assignments and projects.

C. Webniar

1. The video format is not interactive, but the content has real world projects for students to follow.
2. No document format.
3. It also has sample project code for downloading to help hands-on learning.

The limitations of all these 3 online course systems are similar:

1. They are ignoring the different learning style. Some students prefer video format while others prefer

documentation. But these online courses all use video format.

2. They are trying to use the same courses to serve for the general population, without documenting the terminology or theory used in the courses.

DISCUSSION

In this study, I have interviewed 5 participants who worked as conversational data miners and plan to become professional data miners, and the other 5 are entry-level professional data miners who were conversational data miners. The research questions focus on understanding conversational data miners, about their motivations to become professional data miners, what do they want to learn and what do they need to learn, their learning methods as well as learning barriers.

Findings of the research questions can be summarized as below:

1. The motivations are all job oriented. 60% think there are more job opportunities in data mining while 40% chose the career path because of the passion in data work.
2. 100% conversational data miners think statistics, machine learning and hands-on skills required by the industry are important for them to learn. Professional data miners also added data visualization and business communication as needed skills in the real world workplace.
3. The learning methods for conversational data miners who want to become professional data miners are majorly online materials, while SFU Big Data is the only in-person data mining education program they are attending. Both online and in-person learning content they use are similar, but they are showing quite different learning style, 50% only prefer video format while 50% only prefer documentation format. Moreover, they have different understanding levels toward the same courses too. Furthermore, peer discussion plays a significant role in their in-person learning.
4. The leaning barriers majorly come from searching for the right online learning resources, but reasons vary. It can be because the video content is not searchable or conversational data miners are lack of the understanding of the terminology or required knowledge in statistics.

Since there is no study helps conversational data miners' learning. After studying relative work which serves for general population through online courses, and after checking the major online courses used by conversational data miners. I have come up with the following suggestions to help improve online data mining education programs, in order to help conversational data miners' learning.

A. Improvement Based On Participants Feedback

1. Multi-format Learning Materials

- Provide multi-format for each online course, at least it should have both video format and documentation format, so that learners in different style can choose the one they prefer.

2. Multi-levels & Intelligent Recommendations

- For each course, separate the content into different levels, such as basic-level, mid-level and advanced-level. Before learners take a course, they can do a short simple test so that the system will help them decide which level of the course they should take.
- The course system can also recommend courses based on the courses taken by people in similar roles. For example, Alice is a Business Analysts, there are 10 Business Analysts in her city have taken Machine Learning course, the system could also recommend Alice to try Machine Learning.
- The course system can also recommend courses based on relative courses a learner has already finished. For example, if Alice has finished Statistics, the course system could recommend her Machine Learning, to help her learn more.

3. Annotate Video Format

- In order to make video learning resources more searchable based on their content. The online course system could add more annotations to the videos.

B. Crowd Sourcing

Crowd Sourcing can be powerful to improve online learning resources.

1. Crowd Sourcing Terminology Help

- As some interview participates mentioned, sometimes they struggled to understand the learning materials because of they didn't understand the terminology. Inspired by the idea of LemonAid [29], online course system could allow a learner to click on a certain word appeared in the video or documentation, then the crowd sourced description will appear.

2. Crowd Sourcing Video Annotation Help

- Users who took the video courses could also help annotate a video in order to make the content more searchable.

C. Limitations and Future Work

Despite the insights gained and suggestions provided through this study, there are some limitations.

1. The conversational data miners are all at graduate degree level, therefore may not be able to represent the whole population. The future work needs to include both participants with higher and lower education degrees.
2. The In-person program are majorly around SFU Big Data Program. In the future, more in-person programs need to be investigated.

REFERENCES

1. Thomas W. Dinsmore (2016). Disruptive Analytics Charting Your Strategy for Next-Generation Business Analytics. Apress

2. David Donoho (2015). 50 years of Data Science

3. Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal. 2017 Data mining: Practical machine learning tools and techniques (4th ed), Morgan Kaufmann

4. Manyika, James, et al [2011]. "Big data: The next frontier for innovation, competition, and productivity," report from McKinsey Global Institute. www.mckinsey.com/ insights/business_technology/ big_data_the_next_ frontier_for_innovation

5. Roger fang, Sama Tuladhar (2006). Teaching data warehousing and data mining in a graduate program of information technology. Journal of computing sciences in colleges. Volume 21, Issue 5. 137-144

6. Davenport, Thomas H., and D.J. Patil [2011]. "Data Scientist: The Sexiest Job of the 21st Century," Harvard Business Review. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

7. Robert J. Brunner and Edward J. Kim, Teaching Data Science (2016)

8. Clark, H.H. & S.E. Brennan, Grounding in Communication excerpt: / from Perspectives on socially shared cognition / edited by Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley SOURCE: Washington; American Psychological Association; c1991; pp.127-149

9. Stephanie C. Hicks, Rafael A. Irizarry, A Guide to Teaching Data Science (2016)

10. Baumer, B.. 2015. "A Data Science Course for Undergraduates: Thinking With Data." The American Statistician 69 (4): 334–42. doi: 10.1080/00031305.2015.1081105.

11. Hesterberg, T. C. 2015. "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum." The American Statistician 69 (4): 371–86.

12. Horton, N. J. 2015. "Challenges and Opportunities for Statistics and Statistical Education: Looking Back, Looking Forward." The American Statistician 69 (2): 138–45. doi:10.1080/00031305.2015.1032435.

13. Horton, N. J., Benjamin S. Baumer, and H. Wickham. 2015. "Setting the Stage for Data Science: Integration of Data Management Skills in Introductory and Second Courses in Statistics." Chance 28: 40–50. http://chance.amstat.org/2015/04/setting-the-stage/.

14. Nolan, D., and Temple Lang, D. 2015. "Explorations in Statistics Research: An Approach to Expose Undergraduates to Authentic Data Ana

15. Johanna Hardin, Roger Hoerl, Nicholas J. Horton et al., Data Science in Statistics Curricula: Preparing Students to "Think with Data" (2014)

16. Murray, M. & Guimaraes, M. (2009). „Animated Courseware Support for Teaching Database Design", Issues in Informing Science and Information Technology Volume 6, 201-211.

17. Barnes, S. & Kuzma, J. (2009). „Empirical study in teaching first-year database students", 7th International Workshop on Teaching, Learning and Assessment of Databases, 6 July 2009, University of Birmingham.

18. Philip, G. C. (2007). „Teaching Database Modelling and Design: Areas of Confusion and Helpful Hints", Journal of Information Technology Education, Volume 6, 481-497.

19. Chilton, M. A., McHaney, R. & Chae, B. (2006). „Data modelling education: The changing technology", Journal of Information Systems Education, 17(1), 17-20.

20. Czenky, M. & Kormos, J. (2014). „Concept systematization with concept maps in data modelling", Teaching Mathematics and Computer Science, 12/2, 149–166.

21. M'arta Czenky, An Examination of the Effectiveness of Teaching Data Modelling Concepts (2015)

22. Emanuelle Burton, Judy Goldsmith, Sven Koenig et al., Ethical Considerations in Artificial Intelligence Courses (2017)

23. Marcelo Ponce, Erik Spence, Daniel Gruner et al., Scientific Computing, High-Performance Computing and Data Science in Higher Education (2016)

24. Bevilacqua, F., Schnell, N., Rasamimanana, N., Zamborlin, B., and Gu´edy, F. (2011). Online gesture analysis and control of audio processing. In Musical Robots and Interactive Multimodal Systems, pages 127–142. Springer

25. Rebecca Fiebrink, Baptiste Caramiaux, The Machine Learning Algorithm as Creative Musical Tool (2016)

26. Simon Fauvel and Han Yu, A Survey on Artificial Intelligence and Data Mining for MOOCs (2016)

27. Chris Holdgraf, Aaron Culich, Ariel Rokem, Fatma Deniz, Maryana Alegro, Dani Ushizima, Portable learning environments for hands-on computational instruction: Using container- and cloud-based technology to teach data science (2017)

28. Robert J. Brunner and Edward J. Kim, Teaching Data Science (2016)

29. Chilana, P., Ko, A.J., and Wobbrock, J.O. (2012) LemonAid: SelectionEBased Crowdsourced Contextual Help for Web Applications. *Proc ACM Conference on Human Factors in Computing Systems*. New York: ACM Press, pp. 1549E1558.