Proposal for New Concentration in Big Data & Machine Learning (BDML)  
Master of Science in Analytics

Institution: Georgia State University

College: College of Arts and Sciences

Departments: Computer Science; Mathematics and Statistics

Contact name and email/phone:

1) Dr. Rafal Angryk  
Professor of Computer Science  
E-mail: rangryk@gsu.edu  
Phone: (404) 413-5729

2) Dr. Xiaojun (Matt) Cao  
Professor of Computer Science  
E-mail: cao@gsu.edu  
Phone: (404) 413-5732

3) Dr. Yichuan Zhao  
Professor of Mathematics and Statistics  
Email: yichuan@gsu.edu  
Phone: (404) 413-6403

Date of submission: September 18, 2018

Implementation term: Fall 2019 Semester

Degree / major program: Master of Science in Analytics (MSSA)

Concentration title: Big Data & Machine Learning (BDML)
• Program description and objectives

The Department of Computer Science within the College of Arts and Sciences proposes adding a Big Data & Machine Learning (BDML) concentration to the already existing Master of Science in Analytics (MSA) program. Our proposal responds to the growing need for highly qualified professionals who can transform Big Data into practical and valuable insights. In general terms, Data Scientists extract meaning from data. The simple definition contrasts with the highly technical procedures required to extract meaning from Big Data. The proposed BDML program will allow GSU students to gain the technical skills that industry increasingly expects from Data Scientists. Big Data comes from the Internet of Things (IoT), robotics, autonomous vehicles, and other IT-related fields such as scientific labs working with medical or remote-sensing data, companies specializing in big data processing and analysis, cloud storage and computing services. These companies aggressively seek graduate-level professionals who can (1) collect, clean, manage, analyze and interpret big data, (2) derive new knowledge from big data, (3) make sure these discoveries are transferred to the form of actionable items for upper administrators, and (4) clearly communicate to stakeholders through sophisticated but human-friendly computer visualization techniques.

The proposed BDML concentration for the already existing MSA degree will be offered as a full-time, traditional Master of Science program with STEM designation and 27 credit hours of required courses, 6-8 credit hours of elective courses, and 1 credit hour of a BDML capstone project. Depending on the choice of electives, the entire degree will span from 34 to 36 credit hours.

About ¾ of all courses will be taught within the Department of Computer Science. The remaining ¼ of students’ course load will come from the Department of Mathematics and Statistics. A dedicated Program Director in Computer Science department will oversee students enrolled in the BDML concentration. She/he will provide guidance and support to help students progress through the degree, choose a capstone project, and find internships or appropriate sprints during the summer term. The overall length of the program will be 16 months or three semesters. Instruction will be fully residential.

In the summer, our students will either (1) participate in data science internship(s) (evaluated by surveys completed by employers), (2) work under faculty supervision on their own big data-driven machine learning project(s), and/or (3) participate in data science sprint(s) (evaluated by sprint instructors). The sprints will be offered through interdisciplinary collaborations between Computer Science Department and other departments in our university, or industry. We are going to have sprints that are driven by data domains, letting our students improve their analytical skills in this big data sector, which they would like to specialize in during their future professional careers. Some examples: (1) *spatiotemporal data mining* for students interested in a growing GIS industry, autonomous vehicles, traffic monitoring, and multiple sensors/IoT devices (e.g., water quality and leakage sensor deployment and analytics of real-time data) (2) *monochromatic computer vision* for students interested in robotics, astrophotgraphy, solar physics, space weather and medical image analysis, (3) *sequential data analysis* for students wanting to specialize in mining time-series and bioinformatics, (4) *polychromatic object recognition* for students interested in automated identification of patterns in digital photography and video (e.g., security cameras) data, and (5) *graph data mining* for students interested
in social sciences and FMRI (functional magnetic resonance imaging) data analyses. We believe choosing data formats as central points of our summer BDML activities will help students broaden job opportunities, and will provide other academic units at GSU with novel and interdisciplinary collaboration opportunities. Financial support for the highly specialized data sprints will come from the extra program fees charged by all MS in DSA programs on our campus.

The Master of Science in Analytics (MS A) degree, with proposed concentration on Big Data & Machine Learning (BDML), will be earned after all courses are successfully completed, together with 45-minutes-long presentation of the capstone project, defended in front of 2-3 faculty members, followed by an approved written report on the project. Each project will need to encompass three fundamental stages of a data mining project that includes (1) data collection/preparation, (2) machine or statistical learning, and (3) evaluation/assessment followed by the interpretation of discovered new knowledge.

The new BDML concentration of MSA program will meet the high demand for data-savvy professionals in the IT-related job market and attract more Computer Engineering, Electrical Engineering, Statistics and Applied Mathematics majors to our graduate program. Upon completion of our concentration of the MSSA degree, all students will possess the following data science skills and abilities:

1. Understand the principles of Data Science (DS), enhance DS knowledge and be able to apply DS theory for practical DS tasks.
2. Collect, store, search, mine and visualize big data. Learn how to transform raw data into tangible value and evaluate data in terms of volume, variety, velocity, source, etc.
3. Identify the DS tasks of an organization and design corresponding solutions. Learn how to evaluate DS options and limitations which could possibly influence organizational needs.
4. Analyze and evaluate multiple DS models. Understand the strength and limitation of diverse DS models in terms of various DS tasks and be able to select appropriate ones to solve the given DS task(s).
5. Interpret outcomes from employed DS model(s). Transform the observations from data resources and outcomes from DS models into actionable business strategies, and persuade decision makers about practical benefits coming from these discoveries.
6. Have the leadership skills and knowledge to lead a team of DS professionals, and to provide long-term, high-quality DS services to their employers.

Objectives 1, 3 and 4 will be introduced in Csc 6780 and STAT 8670 and reinforced in Csc 6740, Csc 6850 and one of the following STAT classes: 8561, 8700, 8310. They will be further reinforced and mastered in the two 8000-level electives taken in the last semester of proposed concentration. Objective 2 will be introduced in Csc 6780 and then reinforced in Csc 6710, to be finally mastered in Csc 6760. Some of the 8000-level electives offered in the last semester provide additional opportunities to master the objective 2. Objective 5 will be introduced in Csc 6780 and STAT 8670 and reinforced in Csc 6740, Csc 6850, and then mastered through the two 8000-level electives and final BDML project. Objective 6 will be introduced in Csc 6780 and Csc 8902 and reinforced through the completion of two collaborative, group research projects which are already offered as parts of the Csc 6740 and Csc 6850 courses. It will be mastered through summer activities (i.e., data science internships or sprints) and completion and presentation of their capstone project.
Taking two 8000-level electives will provide students with additional opportunity to master objectives 1, and 3 through 6. Moreover, participating in summer activities (internships, projects or sprints), funded by the additional program fees) will give students an opportunity to reinforce their mastership of objectives 2 through 6. Finally, the BDML Capstone Project will serve as reinforcement of students’ mastership of objectives 1 through 5.

<table>
<thead>
<tr>
<th>Objectives:</th>
<th>Courses:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Understand the principles of Data Science (DS)</td>
<td>CSc 6780 Found. of Data Science</td>
</tr>
<tr>
<td>2. Collect, store, search, mine and visualize big data</td>
<td>CSc 6710 Database Systems</td>
</tr>
<tr>
<td>3. Identify the DS tasks of an organization and design solutions</td>
<td>STAT 8670 Comp. Methods in Stat.</td>
</tr>
<tr>
<td>4. Analyze and evaluate multiple DS models</td>
<td>CSc 8902 Ethics</td>
</tr>
<tr>
<td>5. Interpret outcomes from employed DS model(s)</td>
<td>CSc 6740 Intro. to Machine Learning</td>
</tr>
<tr>
<td>6. Have DS leadership skills</td>
<td>CSc 8650 Intro. to Machine Learning</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mastered</th>
<th>Mastered</th>
<th>Mastered</th>
</tr>
</thead>
</table>

- Complete program curriculum (i.e., catalog copy)

### Master of Science in Data Science and Analytics (STEM Degree)

**Proposed Concentration (from Fall 2019): Big Data & Machine Learning (BDML)**

**TOTAL: 34-36 credits**

### Admission Requirements:

- Programming
- Data Structures / Algorithms
- Probability
- Statistics
- Calculus
- Linear Algebra

### 1st Semester Required Courses (12 credits)

- CSc 6780 Found of Data Science (4 credits)
- CSc 6710 Database Systems (4 credits)
- STAT 8670 Computational Methods in Statistics (3 credits)
- CSc 8902 Ethics (1 credit) - New class offered in the form of online modules, extended with case studies provided by our faculty (through seminars and lectures)

### Other Required Courses (15 credits)

- CSc 6760 Big Data Programming (4 credits)
- CSc 6740 Data Mining (4 credits)
- CSc 6850 Introduction to Machine Learning (4 credits)
- STAT 8561 Linear Statistical Analysis I (3 credits) OR
- STAT 8700 Categorical Data Analysis (3 credits) OR
- STAT 8310 Applied Bayesian Statistics (3 credits)
## Two 8000-level Elective Courses (6-8 credits)

- courses below can be counted as electives only if they were not taken earlier by a student
  - CSc 8530 Parallel Algorithms (4 credits)
  - CSc 8710 Deductive Databases and Logic Programming (4 credits)
  - CSc 8711 Databases and the Web (4 credits)
  - CSc 8712 Advanced Database Systems (4 credits)
  - CSc 8713 Spatial and Scientific Databases (4 credits)
  - CSc 8740 Advanced Data Mining (4 credits)
  - CSc 8741 Graph Mining (4 credits)
  - CSc 8810 Computational Intelligence (4 credits)
  - CSc 8850 Advanced Machine Learning (4 credits)
  - CSc 8851 Deep Learning (4 credits)
  - CSc 8910 Data Dissemination in Online Social Networks (4 credits)
  - STAT 8090 Applied Multivariate Statistics (3 credits)
  - STAT 8561 Linear Statistical Analysis I (3 credits)
  - STAT 8610 Time Series Analysis (3 credits)
  - STAT 8674 Monte Carlo Methods (3 credits)

## BDML Capstone Project (1 credit)

- CSc/STAT 8930 MS Project (1 credit) – covering our students’ Capstone Presentation.
  Students complete an independent multifaceted culminating project that is an extension of a class project to include concepts learned throughout their graduate program.
  Students are required to conduct a 45-minute presentation defense of their project to two faculty members, as well as submit a final written report.

Besides the curriculum provided above, the program will offer professional development workshops to better prepare the student to enter the work field. This includes teaching students about the dynamics of receiving and accepting offers, understanding compensation packages and expectations. The Industry Liaison will reach out to alumni that have experience with internships for mentoring students in the right direction, as well as maintain a recruitment list of companies that hired students from within the program and facilitate company visits to the campus for holding workshops.

- Evidence of the need for and interest in this program, including projected enrollments.

According to a ranking at glassdoor.com (see 50 Best Jobs in America, https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm), Data Scientist and Data Engineer careers are the top 1 and top 3 best jobs in the United States, respectively, and have remained at the top for the past few years. This is truly remarkable as the job description for a data scientist or data engineer did not even exist ten years ago.
**Data Scientist:** The most advanced analytics professional with sophisticated computer, statistical and mathematical skills

$120K
Median Base Salary

293%
Job Posting Growth (2013 H2-2016 H2)

---

Fig. 1. **Left:** Buoyant Job Market for Data Analytics Specialists, copied from the source: EAB-Data-Analytics-Study 2017, page 17. **Right:** Historical Employer Demand for Master’s-Level Data Science and All Master’s-Level Professionals (July 2013-December 2017, National Data), copied from the source: EAB - Employer Demand for GSU Master’s-Level Data Science Professionals.

Recent reports highlight the fact that a shortage of qualified workers with technical skills is a challenge for organizations across industries and geographical boundaries. Back in 2012, the research firm *Gartner* predicted that there would be a shortage of 100,000 data scientists in the United States by 2020. A year earlier, *McKinsey* put the national gap in data scientists and others with deep analytical expertise at 140,000 to 190,000 people by 2017. This estimate suggests a demand that is 60% greater than the current supply. In 2014, the consulting firm *Accenture* found that more than 90% of its clients planned to hire people with data science expertise, but more than 40% cited a lack of talent as the number one problem. In the most recent study done by the Education Advisory Board (i.e. EAB, Data Analytics Study 2017, [https://www.eab.com/research-and-insights/continuing-and-online-education-forum/studies/2017/data-analytics-capitalizing-on-creative-disruption](https://www.eab.com/research-and-insights/continuing-and-online-education-forum/studies/2017/data-analytics-capitalizing-on-creative-disruption)) Data Analytics professionals are called “Masters of the Universe” with 293% job posting growth registered during the 2013-2016 period, a median base salary of $120,000 reported for regular employees, and $254,000 for data science managers.

The BDML concentration is the College of Arts and Science’s response to the obvious market needs. The concentration will leverage faculty strengths, thereby growing and strengthening the local economy. In addition, the proposed program is consistent with national trends in the discipline. Many schools already offer general and/or specialized DS programs, whether in person or online. Companies such as telecom, utilities, fintech, consumer goods, and Internet industries are all excited about the potential for extracting insights from their data, but they cannot find qualified applications who can deal with the large amounts of complex data they are generating.

The BDML concentration will be structured to best fit the needs of students who seek a foundation in data science to further their employment opportunities but who do not want to specialize in algorithmic (i.e., pure computer science) or business aspects of data science, per se. Courses will be developed and taught by expert faculty from both the CS and Math & Statistics departments.
keeping with best pedagogical practices, these courses will be interactive. Students will engage in their learning and will have opportunities to practice what they have learned. We have consulted with industry practitioners to make sure that our learning objectives are current and valuable to employers.

The technical demands of data science support locating the BDML concentration within the Department of Computer Science at GSU’s College of Art and Science. Although successful data scientists require a combination of domain knowledge and technical skills, computational advances explain data science’s rise to the top of job rankings in recent years. Uniquely at GSU, the technically sophisticated BDML concentration will prepare students for careers within IT-related companies (Big Data, Cloud Storage and Computing Services), within companies developing remote sensing technologies (Robotics, Internet of Things), and the emerging Autonomous Vehicles industry.

In summary, the demand for and shortage of data science professionals place the CS department in an excellent position. We expect to graduate skilled new data scientists who have industry-specific knowledge fused with IT acumen and will thus be in high demand across industries at the local, regional, and national levels.

To project our enrollments, we investigated MS in Data Science enrollments coming from four comparable academic institution. The results are presented in the table below. Based on these statistics alone, we expect 20 students in the first year. This will be the first cohort of the program. With hiring new teaching/clinical faculty (1 in CS and 1 in Math/Stat) and 1 additional stuff member, we will be able to grow the program by additional 30 students. If the program continues to attract more applicants, additional hires of faculty and stuff may be necessary.

<table>
<thead>
<tr>
<th>Institution</th>
<th>MS Data Science Enrollment (2017)</th>
</tr>
</thead>
<tbody>
<tr>
<td>George Mason University</td>
<td>67</td>
</tr>
<tr>
<td>University of Virginia</td>
<td>82</td>
</tr>
<tr>
<td>Indiana University – Bloomington</td>
<td>116</td>
</tr>
<tr>
<td>North Carolina State University</td>
<td>118</td>
</tr>
</tbody>
</table>

The proposed Big Data & Machine Learning (BDML) concentration of the Master of Science in Analytics degree aligns perfectly with Georgia State University’s vision and mission to “become a leading public research university addressing the most challenging issues of the 21st century (Goal #3)”. As noted above, one of the biggest challenges of the 21st century is the growing gap between our ability to generate Big Data and our ability to analyze it.

The proposed BDML concentration also aligns equally well with the strategic plan of College of Art and Sciences (CAS). Goal #2 of this strategic plan is for our College to “grow and strengthen graduate programs”. In addition, the proposed MS in DS will also significantly advance CAS’ Initiative 2: “design new graduate programs to address market needs”.

- **Anticipated impact on other programs within the offering department, the college, or the university.**

We do not expect to be taking students from other programs within our university. The only two programs, somehow related to the proposed concentration, are (1) our own MS in
Computer Science program, and (2) the original Business Analytics concentration of campus-wide MS in Analytics degree.

With our ambition of attracting students primarily from the areas of Computer Engineering and Electrical Engineering, we will not be competing with MSA’s concentration in Business Analytics, which has been originally established for students with backgrounds in business and actuarial science. Each year, Department of Computer Science receives quite a significant number of applications from students with bachelor’s degrees in Computer Engineering and Electrical Engineering, who have desire to be accepted to our MS in Computer Science program. However, due to a large number of deficiency courses in computer science area, a large number of applicants we receive, and a high selectivity of our MS in CS program admission process, we usually are not able to accept these students to the CS Department. Establishing the new BDML concentration in the Department of Computer Science will enable us to bring these students (typically with very strong academic backgrounds in STEM disciplines) to our campus.

- Additional resource requirements, if any, and budget implications (e.g., personnel costs, library acquisitions, computing/equipment costs, facilities and other operating costs, graduate student support). Intended method of funding additional costs if any.

For the first 20 students, we will require: 4 GTAs and, one staff member (the Program Administrator), but no tenure-track faculty. The Program Administrator will provide logistical and procedural support for the program. The 4 GTAs will grade homework assignments, track participation, manage interactive assignments, and etc. Additional resources may be required to mount this program, dependent on growth.

If second-year enrollment exceeds 30 students, we will need additional 4 GTAs, 2 instructional faculty (1 from Math/Stat and 1 from CS), and 1 staff person serving as Industry Liaison.

We are requesting seed money for a (1) Program Administrator and (2) Industry Liaison, who will, respectively, (1) oversee the application process and student progression in the degree, and (2) facilitate connections with industry. We expect this support to be needed during early years of the program only, and will be self-sustained from the program fees for future years.

Students working on capstone projects will require some supercomputing processing. For the first cohort, the collaboration with GSU Research Solutions will provide the necessary computing infrastructure. After year 1 of the program, we will need to build our own infrastructure and a System Administrator will be needed. We are requesting funds for Big Data processing servers, as well as for a System Administrator line.

- Administration of the program.

The governing body for the program will consist of a dedicated Program Director in Computer Science department and one designated faculty from Statistics. The directors of Graduate studies in both CS and Statistics will also facilitate the program. The administration team will meet at least once per semester with faculty participants to troubleshoot the curriculum issues and discuss
the improvement to the program as needed. Admissions and academic advisement related to the program will be administered through the CAS Office of Academic Assistance. Program-specific issues and resolution of formal student complaints will be resolved by the Directors of Graduate Studies and Chairs for Computer Science, and Mathematics and Statistics.

- **Process for admitting students to the program.**

The admission process will be similar to the MS program in Computer Science. The aforementioned administration team will manage the admission to the program and the certification of its successful completion. The program will admit new students only in Fall semester.

Applicants will be required to submit (1) transcripts from all previous enrollments in higher education, (2) three letters of recommendation, and (3) a brief personal essay explaining the applicant’s interest in the program. International students from non-English-speaking countries will be required to submit TOEFL scores. The admission criteria include:

- Completion of a baccalaureate degree (in engineering, computer science, natural and physical sciences, mathematics, statistics, etc.).
- A minimum GPA of 3.0.
- GRE scores (the GRE requirement may be waived for applicants with prior GPA of 3.5 or higher).
- Completion of the following foundation coursework or equivalent, with a grade no lower than a B:
  - Programming
  - Data Structures / Algorithms
  - Probability
  - Statistics
  - Calculus
  - Linear Algebra

- **Advisement process and resources for students in the program.**

The BDML directors from two departments will serve as the academic advisors for students in the program. They will act in consultation with the department chair and the Graduate Office in College of Arts and Sciences. A tenured professor from Computer Science department will serve as BDML Program Director.

Graduate students will work with faculty advisors to finish their capstone projects, which will explore in depth a single data science task, chosen through consultation between the student and advisor. The project will be accomplished by gathering, preprocessing and cleaning the relevant data, exploring alternative data science approaches, performing quantitative comparison of multiple data science models, and evaluating and interpreting discovered new knowledge. The student will write a report on the project, prepare an oral presentation, and pass an oral final examination given by an ad hoc faculty committee headed by the capstone project advisor.

- **Measures to assess the effectiveness of the program.**
Program Outcome

1. Recruit and retain 20 students per year, including those from underrepresented populations in the STEM fields
   • Program enrollment
   • Student demographics (collected from application data)

2. Within 2 years, at least 85% of students will graduate successfully from the program
   • Graduation rate data

3. Graduating students will find positions in the field of Data Science, Big Data, and Machine Learning
   • Job placement data
   • Survey(s) collected from alumni

Student Outcome
As a result of completing the MS in DSA with concentration in BDML, students will demonstrate...

<table>
<thead>
<tr>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ...understanding of DS principles and be able to apply DS theory for practical DS tasks</td>
</tr>
<tr>
<td>2. ... ability to collect, store, search, mine and visualize data</td>
</tr>
<tr>
<td>3. ... ability to identify the DS tasks of an organization and design corresponding solutions</td>
</tr>
<tr>
<td>4. ...ability to analyze, compare and evaluate multiple DS models</td>
</tr>
<tr>
<td>5. ...ability to interpret outcomes from employed DS model(s)</td>
</tr>
<tr>
<td>6. ...leadership skills and knowledge to lead a team of DS professionals</td>
</tr>
</tbody>
</table>

- Offices, departments, committees, and individuals consulted during the development of the proposal.

During the process of developing this proposal, the following GSU units were consulted.

- Robinson College of Business
- Graduate office in College of Arts & Sciences
- Office of Associate Provost for Graduate Programs

- Approval path for program proposal, noting all formal department- or college-level votes.

Approval Process
The approval process for proposals consists of the steps below.

1. The proposal was approved by Computer Science and Mathematics and Statistics on 9/18/2018.
2. The proposal was sent to CAS graduate council for approval on 9/26/2018.
3. Proposals require the approval of the dean of the college responsible for the administration of the new program. Individual colleges may elect to require that proposals first be reviewed or formally considered by college faculty, a college undergraduate or graduate committee, or some other college-level body. Approval at this stage, including any formal vote, should be noted as specified in the program proposal.

4. The dean should send approved proposals to the Vice Provost and to the chair of the University Senate Committee on Academic Programs. The vice provost will review for compliance with Georgia Board of Regents (BOR) policies and will ask the Associate Provost for Institutional Effectiveness to do the same for Southern Association of Colleges and Schools (SACS) requirements. Depending on the nature of the proposed program, the vice provost may ask others to review the proposal as well. For example, if the program involves collaboration with an international partner institution, the Associate Provost for International Affairs will also be asked to review the proposal.

5. If the proposal is found to be in good order, the Vice Provost will notify the chair of CAP that it is ready for committee consideration. If problems are found with the proposal, the Vice Provost will contact the college dean to resolve the issues before initiating University Senate deliberations.

6. Within CAP, the proposal initially will be deliberated on by a subcommittee, most typically the Undergraduate Council or Graduate Council (depending on the level of the program being considered). The subcommittee may elect to invite the proposing parties to attend a meeting at which the proposal is discussed. The subcommittee will then make a recommendation to CAP. At a meeting to which the proposing parties will be invited, CAP will deliberate and vote on the proposal. At both the subcommittee and full committee levels, requests may be made to the proposing parties for changes to be made to the proposal.

7. The chair of CAP will notify the Vice Provost and the Associate Provost for Institutional Effectiveness in writing of the recommendation of CAP.

8. If any aspect of the proposed program constitutes a substantive change by BOR or SACS standards, the university may be required to submit additional notifications or seek approval from either or both of these bodies. In such cases, the proposing units may be required to provide additional program information.

9. The provost, on behalf of the university president, will make the final decision on implementation of the new program.

10. Normally, the program can be officially added to the university curriculum (i.e., added to the record of official programs in the Banner system) after final university approval.