

UNIVERSITY SENATE
UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

Introduced by: Graduate Academic Council
University Planning & Policy Council

Date: April 17, 2017

PROPOSAL TO ESTABLISH A
MASTER OF SCIENCE (M.S.) PROGRAM IN DATA SCIENCE

IT IS HEREBY PROPOSED THAT THE FOLLOWING BE ADOPTED:

1. That the University Senate approves the attached proposal to establish a M.S. program in Data Science as approved by the Graduate Academic Council (4/17/17) and the University Planning & Policy Council (3/2/16).
2. That this proposal be forwarded to the President for approval.



New Program Proposal: Graduate Degree Program

Form 2B

Version 2016-9-13

This form should be used to seek SUNY’s approval and New York State Education Department’s (SED) registration of a proposed new academic program leading to master’s or doctoral degree. Approval and registration are both required before a proposed program can be promoted or advertised, or can enroll students. The campus Chief Executive or Chief Academic Officer should send a signed cover letter and this completed form (unless a different form applies¹), which should include appended items that may be required for Sections 1 through 6, 9 and 10 and MPA-1 of this form, to the SUNY Provost at program.review@suny.edu. The completed form and appended items should be sent as a single, continuously paginated document.² If Sections 7 and 8 of this form apply, External Evaluation Reports and a single Institutional Response should also be sent, but in a separate electronic document. Guidance on academic program planning is available [here](#).

Table of Contents

NOTE: Please update this Table of Contents automatically after the form has been completed. To do this, put the cursor anywhere over the Table of Contents, right click, and, on the pop-up menus, select “Update Field” and then “Update Page Numbers Only.” The last item in the Table of Contents is the List of Appended and/or Accompanying Items, but the actual appended items should continue the pagination.

Section 1. General Information	2
Section 2. Program Information	4
2.1. Program Format	4
2.2. Related Degree Program	4
2.3. Program Description, Purposes and Planning	4
2.4. Admissions	9
2.5. Academic and Other Support Services	9
2.6. Prior Learning Assessment	10
2.7. Program Assessment and Improvement	10
Section 3. Program Schedule and Curriculum	10
Section 4. Faculty	12
Section 5. Financial Resources and Instructional Facilities	13
Section 6. Library Resources	14
Section 7. External Evaluation	14
Section 8. Institutional Response to External Evaluator Reports	14
Section 9. SUNY Undergraduate Transfer	14
Section 10. Application for Distance Education	14
Section MPA-1. Need for Master Plan Amendment and/or Degree Authorization	15
List of Appended Items	16

¹Use a different form if the proposed new program will lead to a graduate degree or any credit-bearing certificate; be a combination of existing registered programs (i.e. for a multi-award or multi-institution program); be a breakout of a registered track or option in an existing registered program; or **lead to certification as a classroom teacher, school or district leader, or pupil personnel services professional** (e.g., school counselor).

²This email address limits attachments to 25 MB. If a file with the proposal and appended materials exceeds that limit, it should be emailed in parts.

Section 1. General Information

a) Institutional Information	Date of Proposal:	10/17/16
	Institution's 6-digit SED Code :	210500
	Institution's Name:	University at Albany
	Address:	1400 Washington Ave., Albany, NY 12222
	Dept of Labor/ Regent's Region :	Capital District
b) Program Locations	List each campus where the entire program will be offered (with each institutional or branch campus 6-digit SED Code): University at Albany, 210500.	
	List the name and address of off-campus locations (i.e., extension sites or extension centers) where courses will offered, or check here [<input type="checkbox"/>] if not applicable :	
c) Proposed Program Information	Program Title:	Data Science
	Award(s) (e.g., M.A., Ph.D.):	M.S.
	Number of Required Credits:	Minimum [36] If tracks or options, largest minimum [<input type="text"/>]
	Proposed HEGIS Code :	1703.00
	Proposed 6-digit CIP 2010 Code :	27.0301
	If the program will be accredited, list the accrediting agency and expected date of accreditation:	
	If applicable, list the SED professional licensure title(s) ³ to which the program leads:	
d) Campus Contact	Name and title: Mark Steinberger, Director of Graduate Studies, Department of Math. and Stat. Telephone: 518-257-0951 E-mail: mark@albany.edu , nyimed@gmail.com	
e) Chief Executive or Chief Academic Officer Approval	Signature affirms that the proposal has met all applicable campus administrative and shared governance procedures for consultation, and the institution's commitment to support the proposed program. E-signatures are acceptable. Name and title: Darrell Wheeler, Interim Provost and Senior Vice President for Academic Affairs	
	If the program will be registered jointly⁴ with one or more other institutions, provide the following information for <u>each</u> institution: Partner institution's name and 6-digit SED Code : Name, title, and signature of partner institution's CEO (or append a signed letter indicating approval of this proposal):	

³ If the proposed program leads to a professional license, a [specialized form for the specific profession](#) may need to accompany this proposal.

⁴ If the partner institution is non-degree-granting, see SED's [CEO Memo 94-04](#).

Attestation and Assurances

On behalf of the institution, I hereby attest to the following:

That all educational activities offered as part of this proposed curriculum are aligned with the institutions' goals and objectives and meet all statutory and regulatory requirements, including but not limited to Parts 50, 52, 53 and 54 of the Rules of the Board of Regents and the following specific requirements:

That credit for study in the proposed program will be granted consistent with the requirements in §50.1(o).

That, consistent with §52.1(b)(3), a reviewing system has been devised to estimate the success of students and faculty in achieving the goals and objectives of the program, including the use of data to inform program improvements.⁵

That, consistent with §52.2(a), the institution possesses the financial resources necessary to accomplish its mission and the purposes of each registered program, provides classrooms and other necessary facilities and equipment as described in §52.2(a)(2) and (3), sufficient for the programs dependent on their use, and provides libraries and library resources and maintains collections sufficient to support the institution and each registered curriculum as provided in §52.2(a)(4), including for the program proposed in this application.

That, consistent with 52.2(b), the information provided in this application demonstrates that the institution is in compliance with the requirements of §52.2(b), relating to faculty.

That all curriculum and courses are offered and all credits are awarded, consistent with the requirements of §52.2(c).

That admissions decisions are made consistent with the requirements of §52.2(d)(1) and (2) of the Regulations of the Commissioner of Education.

That, consistent with §52.2(e) of the Regulations of the Commissioner of Education: overall educational policy and its implementation are the responsibility of the institution's faculty and academic officers, that the institution establishes, publishes and enforces explicit policies as required by §52.2(e)(3), that academic policies applicable to each course as required by §52.2(e)(4), including learning objectives and methods of assessing student achievement, are made explicit by the instructor at the beginning of each term; that the institution provides academic advice to students as required by §52.2(e)(5), that the institution maintains and provides student records as required by §52.2(e)(6).

That, consistent with §52.2(f)(2) of the Regulations of the Commissioner of Education, the institution provides adequate academic support services and that all educational activities offered as part of a registered curriculum meet the requirements established by state, the Rules of the Board of Regents and Part 52 of the Commissioner's regulations.

CHIEF ADMINISTRATIVE or ACADEMIC OFFICER/ PROVOST	
Signature	Date
Type or print the name and title of signatory	Phone Number

⁵ The NY State Education Department reserves the right to request this data at any time and to use such data as part of its evaluation of future program registration applications submitted by the institution.

Section 2. Program Information

2.1. Program Format

Check all SED-defined [formats, mode and other program features](#) that apply to the **entire program**.

a) **Format(s):** Day Evening Weekend Evening/Weekend Not Full-Time

b) **Modes:** Standard Independent Study External Accelerated Distance Education
NOTE: If the program is designed to enable students to complete 50% or more of the course requirements through distance education, check Distance Education, see Section 10, and append a [Distance Education Format Proposal](#).

c) **Other:** Bilingual Language Other Than English Upper Division Cooperative 4.5 year 5 year

2.2. Related Degree Program

NOTE: This section is not applicable to a program leading to a graduate degree.

2.3. Program Description, Purposes and Planning

a) What is the description of the program as it will appear in the institution's catalog?

Master of Science in Data Science:

This program is designed to provide students with the foundations in all three major fields of Data Science: Topological Data Analysis, Machine Learning and Statistical Methods. Students will also develop practical working skills in at least one of them.

The program requires a minimum of 36 credits with an average grade of B or better. There is no foreign language requirement.

Requirements for Admission

In addition to the general University requirements for admission to graduate studies, an applicant's undergraduate major preferably should have been in the mathematical sciences. Students who are deficient in their mathematical preparation must make up such deficiencies before being formally admitted into the program.

Program description:

The program is divided into three clusters: Topological Data Analysis, Machine Learning and Statistical Methods. Students are also expected to take a course in computer methods, AMAT502 (Modern Computing for Mathematicians).

Required courses for the degree consist of the following:

1. AMAT502: Modern Computing for Mathematicians.
2. Two basic courses in the Topological Data Analysis cluster: Topological Data Analysis I (AMAT583) and Topological Data Analysis II (AMAT584).
3. Three basic courses in the Machine Learning cluster: Function Theory and Functional Analysis for Applications (AMAT590), Optimization Methods and Nonlinear Programming (AMAT591), and Methods of Machine Learning (AMAT592).
4. Three basic courses in the Statistics cluster: Introduction to Theory of Statistics I (AMAT554), Applied Statistics (AMAT565), and Nonparametric Methods of Statistical Inference (AMAT581).
5. Each of the three clusters includes a “practicum” course: Practical Methods in Topological Data Analysis (AMAT585); Practical Methods in Machine Learning (AMAT593); Computational Methods in Statistical Inference (AMAT582) Students are expected to take at least one of these.
6. Students must choose two of the following electives: Advanced Linear Algebra (AMAT524); Introduction to Stochastic Processes (AMAT560); additional practicum courses from part 5, above.

University capstone requirement

The practicum course serves as the capstone experience for this Master’s degree. The practicums require comprehensive analysis of data sets along with oral presentations or poster presentations of the results. Students must complete one of the practicum courses with a grade of B or better.

- b) What are the program’s educational and, if appropriate, career objectives, and the program’s primary student learning outcomes (SLOs)? **NOTE:** *SLOs are defined by the Middle States Commission on Higher Education in the [Characteristics of Excellence in Higher Education](#) (2006) as “clearly articulated written statements, expressed in observable terms, of key learning outcomes: the knowledge, skills and competencies that students are expected to exhibit upon completion of the program.”*

Graduates of this program will have attained the following educational and career objectives:

- They will understand the fundamental principles and theories of topological data analysis, machine learning and statistical methods. These three areas represent the major components of Data Science.
- They will have the ability to critically analyze data in practice using these three components.
- They will display the knowledge and skills sufficient to establish a career as a Data Scientist, and to further broaden and develop their capabilities.
-

Student learning outcomes:

- Knowledge: through coursework, students will acquire foundational knowledge of the disciplines of topological data analysis, machine learning, and statistical methods in Data Science.
- Skills: through coursework and practicum, students will learn to manipulate data sets with appropriate software, to ask appropriate questions about the data, and to follow up their initial analyses with further questions and investigations.
- Competencies: through coursework and practicum, students will demonstrate the ability to critically analyze specific data sets.

c) How does the program relate to the institution's and SUNY's mission and strategic goals and priorities? What is the program's importance to the institution, and its relationship to existing and/or projected programs and its expected impact on them? As applicable, how does the program reflect diversity and/or international perspectives? For doctoral programs, what is this program's potential to achieve national and/or international prominence and distinction?

This program will help attract and retain outstanding graduate students and prepare them for being able to tackle complicated problems in Data Analysis using the most contemporary mathematical methods. It also addresses the need for more STEM workers.

Powerful new mathematical methods of data analysis have been developed over the last 10 years, incorporating algebraic, topological and function analytical tools (arising from corresponding areas of "pure" mathematics) along with traditional statistical methods. The resulting new fields of Topological Data Analysis, Computational Algebraic Geometry, and Machine Learning have developed techniques to reveal the internal structure of large data sets, resulting in striking new findings such as distinguishing between different types of breast cancer.

This program offers training in these new methods, as well as rigorous training in Statistical Methods. As such it fills a gap in the University's offerings in meeting the increasing demand for Data Scientists and complements the offerings of the College of Emergency Preparedness, Homeland Security and Cybersecurity, and of the College of Engineering and Applied Sciences.

In general, we are unaware of the existence of similar programs, though one such program is currently being built in Germany. Moreover, only three mathematics departments in the country, Bowling Green State University, South Dakota State University and University Nebraska at Omaha currently have Data Science Master's degrees. None of them incorporate either the topological methods or the reproducing kernel techniques that have become valuable in machine learning. This program positions us to become a nationwide leader in providing training in the mathematical foundations in Data Science.

d) How were faculty involved in the program's design? Describe input by external partners, if any (e.g., employers and institutions offering further education?)

Faculty committees within the department were involved at every step in the development of this program. We also consulted with employers in the biomedical industry, state government and in actuarial and data consulting:

- We received a letter of support (attached) from Richard Scanu, the CFO and COO of Community Care Physicians, PC, Latham, NY, including strong interest in providing internships for our students.
- We received a letter of support (attached) from Dr. Mycroft Sowizral, Research Scientist and BRFS Coordinator at the New York State Department of Health – Bureau of Chronic Disease Evaluation and Research. He believes internships might be possible. He has also offered to help with outreach to other branches of the Department of Health.
- We have consulted with Drs. Kevin Madigan, Irina Moskalenko and Anand Rao of PwC Corporation. Dr. Rao is the head of their data analysis section, and offered to provide data sets for our students to analyze.

e) How did input, if any, from external partners (e.g., educational institutions and employers) or standards influence the program's design? If the program is designed to meet specialized accreditation or other external standards, such as the educational requirements in [Commissioner's Regulations for the profession](#), append a side-by-side chart to show how the program's components meet those external standards. If SED's Office of the Professions requires a [specialized form](#) for the profession to which the proposed program leads, append a completed form at the end of this document.

f) Enter anticipated enrollments for Years 1 through 5 in the table below. How were they determined, and what assumptions were used? What contingencies exist if anticipated enrollments are not achieved?

Year	Anticipated Headcount Enrollment			Estimated FTE
	Full-time	Part-time	Total	
1	15	5	20	
2	25	10	35	
3	35	10	45	
4	45	10	55	
5	50	10	60	

These were determined in part in consultation with Shantou University in China. They expect to be sending us a significant number of Master’s students. In addition, a number of our current Master’s students and applicants have expressed a strong interest in Data Science. With good advertising, we should be able to attract a good cadre of students. We will work hard to attract these students, both through advertising and direct outreach. In particular, we intend to sponsor an annual symposium in Data Science for undergraduates within traveling distance of Albany. We will also work with faculty we know, including a number of our own Ph.D.s, at both local schools and nationally.

- g) Outline all curricular requirements for the proposed program, including prerequisite, core, specialization (track, concentration), internship, capstone, and any other relevant component requirements, but do not list each General Education course.

Course Title	Credits	Course Title	Credits
Core:		Capstone: Choose one of the following:	
Modern Computing for Mathematicians	3	Practical Methods in Topological Data Analysis	3
Topological Data Analysis I	3	Practical Methods in Machine Learning	3
Topological Data Analysis II	3	Computational Methods in Statistical Inference	3
Function Theory and Functional Analysis for	3		
Optimization Methods and Nonlinear	3	Elective: Choose two of the following:	
Methods of Machine Learning	3		
Introduction to the Theory of Statistics	3	Advanced Linear Algebra	3
Applied Statistics	3	Introduction to Stochastic Processes	3
Nonparametric Methods of Statistical Analysis	3	Additional capstone courses	3
Total required credits: 36			

h) Program Impact on SUNY and New York State

- h)(1) **Need:** What is the need for the proposed program in terms of the clientele it will serve and the educational and/or economic needs of the area and New York State? How was need determined? Why are similar programs, if any, not meeting the need?

There are no similar programs either in New York State or nationally. We are providing a unique training program here. It gives training in new topological and analytic techniques in data science.

We consulted Dr. Lisa Montiel, Assistant Provost, Office of Institutional Research & Data Analytics, Research and Policy Analysis Unit, State University of New York. There is currently no specific information about Data Science employment. She provided information about related fields. The results are below.

From our own discussions with employers, we are confident the training we will provide our graduates will allow them to find positions in the actuarial field, the insurance industry, state government, biomedical research, etc.

Occupation Snapshot of Occupations Related to Data Science in New York

		Current						Historical				Forecast							
		Four Quarters Ending with 2016q2			2016q2			Total Change over the Last 5 Years		Avg Ann % Chg in Empl 2011q2-2016q2		US		Over the Next 10 Years					
SOI	Title	Empl	Avg. Annual Wage	Local Quotl	Unempl	Unempl Rat	Empl	New %	New %	US	Current Online Job Ad	Total Open	Total Demand	Total Demand	Total Demand	Avg. Annual Growth Percn			
15-2011	Actuaries	1,501	\$129,900	1.05	2	0.2%	-28	-0.4%	-0.4%	1.2%	36	645	426	219	1.4%				
15-2021	Mathematicians	159	\$121,400	0.81	1	0.5%	2	0.2%	0.2%	0.8%	0	66	27	39	2.2%				
15-2031	Operations Research Analysts	6,012	\$101,500	1.04	125	2.2%	456	1.6%	1.6%	2.1%	7	3,024	1,358	1,666	2.5%				
15-2041	Statisticians	2,038	\$79,800	1.12	22	1.1%	188	2.0%	2.0%	1.6%	21	1,048	404	644	2.8%				
15-2091	Mathematical Technicians	77	\$66,000	1.14	0	n/a	5	1.3%	1.3%	1.6%	0	13	11	2	0.2%				
15-2099	Mathematical Science Occupations, All Other	76	\$84,400	0.78	0	n/a	4	1.1%	1.1%	1.8%	0	26	17	9	1.1%				
15-2000	Mathematical Science Occupations	9,864	\$101,500	1.05	150	1.6%	626	1.3%	1.3%	1.8%	64	4,823	2,244	2,579	2.3%				
11-9121	Natural Sciences Managers	3,110	\$142,800	0.92	78	2.5%	137	0.9%	0.9%	1.1%	24	765	684	81	0.3%				
25-1022	Mathematical Science Teachers, Postsecondary	4,302	\$91,200	1.28	99	2.5%	335	1.6%	1.6%	0.4%	0	1,532	785	747	1.6%				

Source: JobsEQ*
 Data as of 2016Q2 unless noted otherwise
 Note: Figures may not sum due to rounding.
 1. Occupation wages are as of 2015 and should be taken as the average for all Covered Employment
 Exported on: Friday, August 26, 2016 2:32 PM

h)(2) Employment: For programs designed to prepare graduates for immediate employment, use the table below to list potential employers of graduates that have requested establishment of the program and state their specific number of positions needed. If letters from employers support the program, they may be **appended** at the end of this form.

Employer	Need: Projected positions	
	In initial year	In fifth year

h)(3) Similar Programs: Use the table below to list similar programs at other institutions, public and independent, in the service area, region and state, as appropriate. Expand the table as needed. **NOTE:** Detailed program-level information for SUNY institutions is available in the [Academic Program Enterprise System \(APES\)](#) or [Academic Program Dashboards](#). Institutional research and information security officers at your campus should be able to help provide access to these password-protected sites. For non-SUNY programs, program titles and degree information – but no enrollment data – is available from [SED's Inventory of Registered Programs](#).

While there are many programs in Data Science offered by business schools, Data Science Centers and

Computer Science departments, all of them concentrate on Statistical Inference and/or Machine Learning. None of them offer Topological Data Analysis. Moreover, none of them offer the depth and breadth of this program. In particular, we are offering something unique here.

- h)(4) *Collaboration:*** Did this program's design benefit from consultation with other SUNY campuses? If so, what was that consultation and its result?

N/A.

- h)(5) *Concerns or Objections:*** If concerns and/or objections were raised by other SUNY campuses, how were they resolved?

No concerns or objections were raised to our Letter of Intent.

2.4. Admissions

- a)** What are all admission requirements for students in this program? Please note those that differ from the institution's minimum admissions requirements and explain why they differ.

In addition to the general University requirements for admission to graduate studies, an applicant's undergraduate major preferably should have been in the mathematical sciences. Students who are deficient in their mathematical preparation must make up such deficiencies before being formally admitted into the program.

- b)** What is the process for evaluating exceptions to those requirements?

Review by the Graduate Committee of our department, which is composed of full-time faculty.

- c)** How will the institution encourage enrollment in this program by persons from groups historically underrepresented in the institution, discipline or occupation?

We will advertise in multiple ways, including posters, advertisements and recruiting talks. We are also currently investigating additional methods of outreach to members of underrepresented groups. With the assistance of Dr. Tamra Minor, the Assistant Vice President for the Office of Diversity and Inclusion, we are working on developing contacts in Mathematical Science Departments at Historically Black Colleges and Universities.

- d)** What is the expected student body in terms of geographic origins (i.e., same county, same Regents Region, New York State, and out-of-state); academic origins; proportions of women and minority group members; and students for whom English is a second language?

We will work to recruit students from colleges in this region, New York State and nationally, and are also working to establish connections internationally. We will recruit worldwide.

2.5. Academic and Other Support Services

- a)** Summarize the academic advising and support services available to help students succeed in the program.

Three faculty members will devote 100% of their advising/mentoring duties to the students in this program, together with their Ph.D. advising/mentoring. The Director of Graduate Studies will give these students special attention. We will seek out additional forms of support.

- b) Describe types, amounts and sources of student financial support anticipated. Indicate the proportion of the student body receiving each type of support, including those receiving no support.

The students in this program will be primarily self-funded. We will seek out additional funding to help defray costs for students. Some students may be supported by scholarships from private sources, foundations or agencies.

2.6. Prior Learning Assessment

If this program will grant credit based on Prior Learning Assessment, describe the methods of evaluating the learning and the maximum number of credits allowed, **or check here [x] if not applicable.**

2.7. Program Assessment and Improvement

Describe how this program's achievement of its objectives will be assessed, in accordance with [SUNY policy](#), including the date of the program's initial assessment and the length (in years) of the assessment cycle. Explain plans for assessing achievement of students learning outcomes during the program and success after completion of the program. **Append** at the end of this form, **a plan or curriculum map** showing the courses in which the program's educational and, if appropriate, career objectives – from Item 2.3(b) of this form – will be taught and assessed.

NOTE: *The University Faculty Senate's [Guide for the Evaluation of Undergraduate Programs](#) is a helpful reference.*

This program's assessment will be performed along with the assessments of the other graduate programs in the Department of Mathematics and Statistics. In addition, the department will start tracking the employment record of the graduates as soon as the program is established. Consistent with University policy, the Department of Mathematics and Statistics maintains a 7-year assessment cycle for its programs. The Department will apply the same methodology to the assessment of the MS in Data Science that it performs in the assessment of all its programs. This will include direct assessment of student work in core courses, indirect assessment through student surveys, and indirect assessment through student focus groups. The assessment methods will identify successes and deficiencies in the program, and we will use assessment results to address deficiencies and build and maintain program strength. As part of the 7-year assessment cycle, the Department conducts yearly assessments of its programs and courses to determine whether learning objectives are being met.

Section 3. Program Schedule and Curriculum

Complete the **SUNY Graduate Program Schedule** to show how a typical student may progress through the program. This is the registered curriculum, so please be precise. Enter required courses where applicable, and enter generic course types for electives or options. Either complete the blank Schedule that appears in this section, or complete an Excel equivalent that computes all sums for you, found [here](#). Rows for terms that are not required can be deleted.

NOTES: *The **Graduate Schedule** must include all curriculum requirements and demonstrate that expectations from in Regulation 52.2 <http://www.highered.nysed.gov/ocue/lrp/rules.htm> are met.*

Special Cases for the Program Schedules:

- *For a program with multiple tracks, or with multiple schedule options (such as full-time and part-time options), use one Program Schedule for each track or schedule option. Note that licensure qualifying and non-licensure qualifying options cannot be tracks; they must be separate programs.*
- *When this form is used for a multi-award and/or multi-institution program that is **not** based entirely on existing programs, use the schedule to show how a sample student can complete the proposed program. **NOTE:** Form 3A, [Changes to an Existing Program](#), should be used for new multi-award and/or multi-institution programs that are based entirely on existing programs. [SUNY policy](#) governs the awarding of two degrees at the same level. **a)** If the program will be offered through a nontraditional schedule (i.e., not on a semester calendar), what is the schedule and how does it impact financial aid eligibility? **NOTE:** *Consult**

with your campus financial aid administrator for information about nontraditional schedules and financial aid eligibility.

- b) For each existing course that is part of the proposed graduate program, **append** a catalog description at the end of this document.
 - c) For each new course in the graduate program, **append** a syllabus at the end of this document. **NOTE:** *Syllabi for all courses should be available upon request. Each syllabus should show that all work for credit is graduate level and of the appropriate rigor. Syllabi generally include a course description, prerequisites and corequisites, the number of lecture and/or other contact hours per week, credits allocated (consistent with [SUNY policy on credit/contact hours](#)), general course requirements, and expected student learning outcomes.*
 - d) If the program requires external instruction, such as clinical or field experience, agency placement, an internship, fieldwork, or cooperative education, **append** a completed [External Instruction](#) form at the end of this document.
- SUNY Graduate Program Schedule**

Program/Track Title and Award: M.S. in Data Science

- a) Indicate **academic calendar** type: [x] Semester [] Quarter [] Trimester [] Other (describe):
- b) **Label each term in sequence**, consistent with the institution’s academic calendar (e.g., Fall 1, Spring 1, Fall 2)
- c) Use the table to show **how a typical student may progress through the program**; copy/expand the table as needed.
- d) Complete the last row to show program totals and comprehensive, culminating elements. **Complete all columns that apply to a course.**

Term 1

Course number and title	New	Credits	Co/Prerequisites
AMAT 502 Modern Computing for Mathematicians		3	
AMAT 583 Topological Data Analysis I	x	3	
AMAT 590 Function Theory and Functional Analysis for Applications	x	3	
AMAT 554 Introduction to Theory of Statistics I		3	
Term credit total		12	

Term 2

AMAT 584 Topological Data Analysis II	x	3	AMAT 583
AMAT 591 Optimization Methods and Nonlinear Programming	x	3	AMAT 590
AMAT 565 Applied Statistics		3	AMAT 554
Term credit total		9	

Term 3

AMAT 592 Methods of Machine Learning	x	3	AMAT 591
AMAT 581 Nonparametric Methods of Statistical Inference	x	3	AMAT 555
AMAT 585 Practical Methods in Topological Data Analysis	x	3	AMAT 584
Term credit total		9	

Term 4

AMAT 582 Computational Methods in Statistical Inference	x	3	AMAT 581
AMAT 593 Practical Methods in Machine Learning	x	3	AMAT 592
Term credit total		6	

Total credits 36

Culminating elements: AMAT 585, 582 or 593.

Section 4. Faculty

- a) Complete the **SUNY Faculty Table** on the next page to describe current faculty and to-be-hired (TBH) faculty.
- b) **Append** at the end of this document position descriptions or announcements for each to-be-hired faculty member.

NOTE: CVs for all faculty should be available upon request. Faculty CVs should include rank and employment status, educational and employment background, professional affiliations and activities, important awards and recognition, publications (noting refereed journal articles), and brief descriptions of research and other externally funded projects. New York State’s requirements for faculty qualifications are in in Regulation 52.2
<http://www.highered.nysed.gov/ocue/lrp/rules.htm>

- c) What is the institution’s definition of “full-time” faculty?

All faculty in this program are full-time tenure track faculty who, in addition to maintaining an active research program and advising doctoral students, teach 2 courses each semester.

SUNY Faculty Table

Provide information on current and prospective faculty members (identifying those at off-campus locations) who will be expected to teach any course in the graduate program. Expand the table as needed. Use a separate Faculty Table for each institution if the program is a multi-institution program.

Faculty Member Name and Title/Rank (Include and identify Program Director with an asterisk)	% of Time Dedicated to This Program	Program Courses Which May Be Taught (Number and Title)	Highest and Other Applicable Earned Degrees (include College or University)	Discipline(s) of Highest and Other Applicable Earned Degrees	Additional Qualifications: List related certifications, licenses and professional experience in field
PART 1. Full-Time Faculty					
Boris Goldfarb, Assoc. Prof.	25%	583, 584, 585	Ph.D. Cornell	Mathematics	
Martin Hildebrand, Prof.	25%	554, 555, 560,	Ph.D. Harvard	Mathematics	
Elizabeth Munch, Asst. Prof.	50%	502, 583, 584, 585	Ph.D. Duke	Mathematics	
Karin Reinhold, Assoc. Prof.	25%	554, 555, 560, 581,	Ph.D. Ohio State	Mathematics	
Malcolm Sherman, Assoc. Prof.	25%	554, 555, 560, 581,	Ph.D. UC Berkeley	Mathematics	
Mark Steinberger*, Assoc. Prof.	25%	591	Ph.D. U of Chicago	Mathematics	
Michael Stessin, Prof.	25%	590	Ph.D. Moscow State U	Mathematics	
Marco Varisco, Asst. Prof.	25%	502, 583, 584	Ph.D. Muenster	Mathematics	
Yiming Ying, Assoc. Prof.	50%	592, 593	Ph.D. Zhejiang	Mathematics	
Part 2. Part-Time Faculty					
Part 3. Faculty To-Be-Hired (List as					
GBH1 Asst. Prof. 9/17	50%	590, 591, 592, 593,			
GBH2 Asst. Prof. 9/18	50%	583, 584, 585, 554,			
GBH3 Asst. Prof. 9/18	50%	583, 584, 585, 581,			

Note: The percentages in the above table are approximate. Our teaching load consists of four courses per year, at least two of which should be undergraduate courses. The faculty members in question will also be expected to teach core and advanced graduate courses.

Section 5. Financial Resources and Instructional Facilities

a) What is the resource plan for ensuring the success of the proposed program over time? Summarize the instructional facilities and equipment committed to ensure the success of the program. Please explain new and/or reallocated resources over the first five years for operations, including faculty and other personnel, the library, equipment, laboratories, and supplies. Also include resources for capital projects and other expenses.

We will need three new faculty members to cover the courses. The library will be able to give us adequate coverage for the needs of the program, as will the IT department (see attached letters – Doc 9, p40, Doc 8, p41). Most of the program can be accommodated in ordinary classrooms, with some classes held in the existing computer labs.

b) Complete the five-year SUNY Program Expenses Table, below, consistent with the resource plan summary. Enter the anticipated academic years in the top row of this table. List all resources that will be engaged specifically as a result of the proposed program (e.g., a new faculty position or additional library resources). If they represent a continuing cost, new resources for a given year should be included in the subsequent year(s), with adjustments for inflation or negotiated compensation. Include explanatory notes as needed.

- (a) *Personnel*
(including faculty and all others)
- (b) *Library*
- (c) *Equipment*
- (d) *Laboratories*
- (e) *Supplies*
- (f) *Capital Expenses*
- (g) *Other (Specify):*
- (h) **Sum of Rows Above**

Program Expense Categories	Expenses (in dollars)					
	Before Start	Academic Year 1:	Academic Year 2:	Academic Year 3	Academic Year 4:	Academic Year 5:
(a) Personnel (including faculty and all others)		80,000	240,000	240,000	240,000	240,000
(b) Library						
(c) Equipment*		4,000	8,000			
(d) Laboratories						
(e) Supplies						
(f) Capital Expenses						
(g) Other: Start Up Costs**	70,000	20,000	60,000	40,000		
(h)Sum of Rows Above		104,000	308,000	280,000	240,000	240,000
*These are one time expenses for computers.						
** start up funds for the new faculty: \$20,000 per year for two years for each new faculty member. \$70,000 course dev.						

Section 6. Library Resources

- a) Summarize the analysis of library collection resources and needs *for this program* by the collection librarian and program faculty. Include an assessment of existing library resources and accessibility to those resources for students enrolled in the program in all formats, including the institution's implementation of SUNY Connect, the SUNY-wide electronic library program.

See attached letter from Michael Knee, the library's bibliographer for our department (Doc 9-p41) . His summary is as follows:

The University Libraries have been committed to build and maintain collections in support of mathematics and statistics. In FY2015, the University Libraries spend nearly \$226,000 on materials for mathematics and statistics programs. The budget for mathematics and statistics should be able to accommodate required books and reference resources for data science. If additional journals or databases are required, funding will be needed. Materials the University Libraries does not own or provide access to can be obtained using interlibrary loan services.

- b) Describe the institution's response to identified collection needs and its plan for library development.

We believe we can operate this program using the available journals and databases.

Section 7. External Evaluation

SUNY and SED require external evaluation of all proposed graduate degree programs. List below all SUNY-approved evaluators who conducted evaluations (adding rows as needed), and **append at the end of this document** each original, signed [External Evaluation Report](#). **NOTE:** *To select external evaluators, a campus sends 3-5 proposed evaluators' names, titles and CVs to the assigned SUNY Program Reviewer, expresses its preferences and requests approval.*

Evaluator #1

Name: Gunnar Carlsson
Title: Professor of Mathematics
Institution: Stanford University

Evaluator #2

Name: Sayan Mukherjee
Title: Professor of Stat. Sci., Comp. Sci. and Math.
Institution: Duke University

Section 8. Institutional Response to External Evaluator Reports

Append at the end of this document a single *Institutional Response* to all *External Evaluation Reports*.

Section 9. SUNY Undergraduate Transfer

NOTE: *SUNY Undergraduate Transfer policy does not apply to graduate programs.*

Section 10. Application for Distance Education

- a) Does the program's design enable students to complete 50% or more of the course requirements through distance education? [x] No [] Yes. If yes, **append** a completed *SUNY Distance Education Format Proposal* at the end of this proposal to apply for the program to be registered for the distance education format.
- b) Does the program's design enable students to complete 100% of the course requirements through distance

education? No Yes

Section MPA-1. Need for Master Plan Amendment and/or Degree Authorization

a) Based on guidance on [Master Plan Amendments](#), please indicate if this proposal requires a Master Plan Amendment.
 No Yes, a completed [Master Plan Amendment Form](#) is **appended** at the end of this proposal.

b) Based on *SUNY Guidance on Degree Authorizations* (below), please indicate if this proposal requires degree authorization.

No Yes, once the program is approved by the SUNY Provost, the campus will work with its Campus Reviewer to draft a resolution that the SUNY Chancellor will recommend to the SUNY Board of Trustees.

SUNY Guidance on Degree Authorization. Degree authorization is required when a proposed program will lead to a [new degree](#) (e.g., B.F.A., M.P.H.) at an existing level of study (i.e., associate, baccalaureate, first-professional, master's, and doctoral) in an existing disciplinary area at an institution. Disciplinary areas are defined by the [New York State Taxonomy of Academic Programs](#). Degree authorization requires approval by the SUNY Provost, the SUNY Board of Trustees and the Board of Regents.

List of Appended Items

Appended Items: Materials required in selected items in Sections 1 through 10 and MPA-1 of this form should be appended after this page, with continued pagination. In the first column of the chart below, please number the appended items, and append them in number order.

Number	Appended Items	Reference Items
	<i>For multi-institution programs, a letter of approval from partner institution(s)</i>	Section 1, Item (e)
	<i>For programs leading to professional licensure, a side-by-side chart showing how the program's components meet the requirements of specialized accreditation, Commissioner's Regulations for the Profession, or other applicable external standards</i>	Section 2.3, Item (e)
	<i>For programs leading to licensure in selected professions for which the SED Office of Professions (OP) requires a specialized form, a completed version of that form</i>	Section 2.3, Item (e)
Doc 1, p17 Doc 2, p18	<i>OPTIONAL: For programs leading directly to employment, letters of support from employers, if available</i>	Section 2, Item 2.3 (h)(2)
Doc 3, p19	<i>For all programs, a plan or curriculum map showing the courses in which the program's educational and (if appropriate) career objectives will be taught and assessed</i>	Section 2, Item 7
Doc 4, p20	<i>For all programs, a catalog description for each existing course that is part of the proposed graduate major program</i>	Section 3, Item (b)
Doc 5, pp21-37	<i>For all programs with new courses, syllabi for all new courses in a proposed graduate program</i>	Section 3, Item (c)
	<i>For programs requiring external instruction, a completed External Instruction Form and documentation required on that form</i>	Section 3, Item (d)
Doc 6, p38 Doc 7, p39	<i>For programs that will depend on new faculty, position descriptions or announcements for faculty to-be-hired</i>	Section 4, Item (b)
	<i>For all programs, original, signed External Evaluation Reports from SUNY-approved evaluators</i>	Section 7
	<i>For all programs, a single Institutional Response to External Evaluators' Reports</i>	Section 8
	<i>For programs designed to enable students to complete at least 50% of the course requirements at a distance, a Distance Education Format Proposal</i>	Section 10
	<i>For programs requiring an MPA, a Master Plan Amendment form</i>	Section MPA-1

Also appended: Doc 8, Section 5, p 41; Doc 9, Section 6, p42

Document 1



Department of Health

ANDREW M. CUOMO
Governor

HOWARD A. ZUCKER, M.D., J.D.
Commissioner

SALLY DRESLIN, M.S., R.N.
Executive Deputy Commissioner

Dr. Steve Plotnick

Thank you for the opportunity to write a personal letter in support of the Math Department's plans to develop a Master's degree in Data Science. As you know, this is effectively the career path on which I have found myself during my 17 years (already?) employed by New York State (NYS) in a range of roles including data manager, data analyst, statistician, and researcher. My informal goal for these jobs has been to "translate data into information".

I am currently the coordinator of the NYS Behavioral Risk Factor Surveillance System (BRFSS), a large-scale (10,000+ respondents per year) phone survey of NYS adults on a wide range of health statuses, behaviors, and outcomes. (See <http://www.health.ny.gov/statistics/brfss/> and <http://www.cdc.gov/brfss/> for more information about the survey, its design, methodology, goals, and results at both the state and national levels). I find it to be an incredibly exciting, interesting, and challenging role within the NYSDOH. My training in the University at Albany's doctoral mathematics program has been critical in developing a successful, productive, and satisfying career. While the base skill set that I developed during my time in the University of Albany Mathematics program has been instrumental for my professional development, there are certainly opportunities for a M.A. or M.S. program in Data Science to focus key training for students interested in developing a career in this area.

There is great demand for the skill set that would result from achieving such a degree. A specific example would be the experience of management and analysis of large datasets. I would highly encourage the curriculum for this program to include some hands-on work with large (1,000,000 records or more) datasets. This will allow students to gain experience with the basic real-world challenges of managing this level of data, in addition to state-of-the-art analysis and modelling techniques. This would benefit the recipients of such a degree, since very few recent graduates have actually worked with datasets on this order of magnitude. (Whenever I interview a job applicant, one of my first questions is the size of the datasets with which the interviewee has worked.) Huge data systems have been implemented over the past few decades, and are in need of people trained in designing, developing, and managing them; extracting the necessary data; and analyzing and presenting the resulting information. Regardless of the ultimate goal of the analysis, valuable data analysis will originate from large datasets; it is invaluable to the analyst to understand the processes by which the source data were generated, transmitted, organized, as these factors will invariably impact on the data being considered.

Please keep me posted regarding the status of the math department's application. If there are ways in which I might help support the development of this program, please contact me. I specifically think there may be potential for internships with our program in the future. I also have some contacts in the Department of Health and may be able to facilitate linkages with other NYSDOH programs that could be potentially interested in offering internships.

Mycroft Sowizral, Ph.D.

Research Scientist / BRFSS Coordinator

New York State Department of Health - Bureau of Chronic Disease Evaluation and Research

(518) 473-0673

Cc: Ian Brissette
Mark Steinberger
Michael Stessin
Joan Mainwaring

Community Care Physicians, P.C.

Capital Region Health Park • 711 Troy-Schenectady Road, Suite 201 • Latham, NY 12110-2454
Main: (518) 783-3110 • Fax: (518) 782-3797



Document 2

To whom it may concern:

This is a letter expressing Community Care Physicians, P.C.'s (CCP) interest in establishing a connection with the Mathematics and Statistics Department at the University at Albany (SUNY) for the purpose of enhancing our understanding of data we have captured about our patient base.

CCP is a multispecialty medical group with headquarters in Latham, NY, with over 250 providers in 40 practices covering 18 clinical specialties. We see 240,000 unique patients every year, combining for more than 900,000 clinical encounters, making us one of the largest independent healthcare organizations in New York.

We are excited about the potential benefits offered from the use of Dr. Munch and Dr. Ying's data analysis techniques on our extensive data warehouse environment. To this end, we are eager to discuss the logistics of providing a graduate student with opportunities to work with our company as an intern as soon as this summer. We are also interested in discussing direct collaboration ventures with the faculty of the SUNY for joint-research projects at a future date.

We hope that this will mark the beginning of a fruitful relationship between CCP and SUNY. Please feel free contact me at (518) 782-3730 if you have any questions or need additional information.

Sincerely,

A handwritten signature in black ink, appearing to read "R. Scanu".

Richard Scanu, MBA

Chief Financial Officer, Chief Operating Officer

Serving Our Community For Over 25 Years

Internal Medicine • Family Medicine • Pediatric Medicine • Obstetrics • Gynecology • ImageCare Medical Imaging • Image Guided Radiation Therapy • Interventional Radiology • General Surgery
Urology, Physical Therapy • Urgent Care • Laboratory • Audiology • Podiatry • Dermatology • Diabetes Education and Nutrition • Occupational Medicine • Sports Medicine • Bariatric Medicine

Document 3

COMPUTING SKILLS	
Course	Objectives and Outcomes
AMAT 502: Modern Computing...	Background techniques

TOPOLOGICAL DATA ANALYSIS CLUSTER	
Course	Objectives and Outcomes
AMAT 583: Topological Data Analysis 1	(1), (a)
AMAT 584: Topological Data Analysis 2	(1), (2), (3), (a), (b), (c)
AMAT 585: Practical Methods...	(1), (2), (3), (a), (b), (c)

MACHINE LEARNING CLUSTER	
Course	Objectives and Outcomes
AMAT 590: Function Theory...	(1), (a)
AMAT 591: Optimization Methods...	(1), (a)
AMAT 592: Methods of Machine Learning	(1), (2), (3), (a), (b), (c)
AMAT 583: Practical Methods...	(1), (2), (3), (a), (b), (c)

CURRICULUM MAP Statistics Cluster	
Course	Objectives and Outcomes
AMAT 554: ... Theory of Statistics I	(1), (a)
AMAT 565: Applied Statistics	(1), (a)
AMAT 581: Nonparametric Methods...	(1), (2), (3), (a), (b), (c)
AMAT 582: Computational Methods...	(1), (2), (3), (a), (b), (c)
Electives	
Course	Objectives and Outcomes
AMAT 524: Advanced Linear Algebra	Background Techniques
AMAT 560: ... Stochastic Processes	Background Techniques

Document 4

Bulletin descriptions of existing courses to be used in the proposed M.S. in Data Science

Mat 502 Modern Computing for Mathematicians (3)

Introduction to (1) basic principles of computer algebra systems, (2) contemporary mathematical typesetting, and (3) mathematically literate techniques for disseminating mathematical content in both print and HTML-with-MathML forms from a single source. Several computer algebra systems will be examined with an eye toward understanding how to handle various mathematical objects and how to write procedures for tasks that are not handled natively. Written assignments will specify mathematical tasks and presentation standards. Prerequisite: Familiarity with undergraduate mathematics and some ability with computer code.

Mat 524 Advanced Linear Algebra (3)

Brief review of elementary linear algebra. Duality, quadratic forms, inner product spaces, and similarity theory of linear transformations. A term paper or other additional work is required. Prerequisite: Elementary linear algebra (Mat 220 or equivalent) and classical algebra (Mat 326 or equivalent).

Mat 554 (H Sta 554) Introduction to Theory of Statistics (3)

A mathematical treatment of principles of statistical inference. Topics include probability, random variables and random vectors, univariate and multivariate distributions and an introduction to estimation. Appropriate for graduate students in other disciplines and for preparation for the second actuarial examination. Prerequisite: Calculus or linear algebra.

Mat 560 (H Sta 560) Introduction to Stochastic Processes I (3)

An introduction to applied stochastic processes. Topics include Markov chains, queuing theory, renewal theory, Poisson processes and extensions, epidemic and disease models. Prerequisite: Mat 555 (H Sta 555) or an introductory probability course.

Mat 565 Applied Statistics (3)

A course in statistical methods for students with some knowledge of statistics. Topics include multiple regression, analysis of variance and nonparametric statistical techniques. Emphasis on data analysis and statistical methodology. May not be taken for credit by students with credit for Mat 465. Prerequisites: An introductory course in probability or statistics, and some experience with interpretation of data in a subject matter area.

Document 5

August 29, 2017

Mathematics 581 Nonparametric Statistics

Syllabus

Text: Myles Hollander et al: Nonparametric Statistical Models, 3rd ed. Wiley 2014.

Reference: P. Sprent & N.C. Smeeton: Applied Nonparametric Statistical Methods, 4th ed., Chapman & Hall 2014.

Prerequisites: Math 362 and 363 (or 367 and 467) and 214 or equivalent courses.

Math 581 is required for the MA in Data Science.

Instructor: Malcolm J. Sherman, Ph.D., Associate Professor
ES 114, 442-4628
MSherman@albany.edu

Office Hours: Tue: 11:45 - 12:45 (e.g., if the course meets Tu
Wed: 1:30 - 2:30 and Th at 1:15)
Thu: 2:45 - 3:45

Math 581 is a first year graduate course in nonparametric statistics; i.e., statistical methods applicable to data that do not come from a specific parametric family of distributions like the normal or binomial. Such tests may assume, for example, only that an underlying distribution is continuous. In a two sample test the null hypothesis may stipulate only that two populations have the same distribution without specifying a parametric family.

Nonparametric tests can be applied to nominal data on any scale whether ordered (e.g., high-low-medium, liberal-moderate-conservative) or purely nominal (e.g., male-female, or white-black-Hispanic-Asian). Standard parametric tests, in contrast, require interval or ratio scale data (e.g., temperature Fahrenheit, absolute temperature, weight, stature or distance). Standard tests for correlation coefficients are often not appropriate for non-normal data, so that non parametric measures of association, like Spearman's rho and Kendall's rank order correlation are needed. Non-parametric tests can also be utilized to investigate randomness or independence of data, including normal data.

The disadvantages of nonparametric tests are mainly that when standard parametric assumptions hold, the nonparametric tests waste information and require larger samples than applicable standard tests to achieve the same power. Significance levels for nonparametric tests may be computable by hand without tables for small samples, but exact tables for moderate or large samples may not be conveniently accessible, though some statistical packages (like R) can make these computations. Approximate significance levels for nonparametric tests can typically be calculated using normal, chi-square or F tables, though the proof of the asymptotic accuracy of such approximations is not elementary.

A one-semester nonparametric course must choose an appropriate balance between presenting a maximum number of tests and communicating how to recognize data sets whose analysis requires nonparametric methods. At some point during the course a decision will be made as to which computer statistical packages to use and how much time to devote to explaining the features of this package.

Math 581 will be taught as a shared resources course with Math 481, which is a new course that will replace Math 369. Some special assignments will be required only for students registered for 581. Such an assignment might be an application of nonparametric methods using a computer statistical package to a real data set.

A reasonable list of topics for Math 481/581 follows.

I. Introduction (1.5 weeks). Motivating examples of non-normal data (age at death, income distribution) and simulations to demonstrate the failure of standard parametric methods for such data which would also include a review of basic parametric methods). Review of medians and percentiles. Probability plots and the Kolmogorov–Smirnov test for whether data follow a specified distribution.

II. (1 week) Permutation tests for two sample problems (exact for small samples, descriptively for large ones).

III. 1.5 weeks. Sign test (both exact and with normal approximation to binomial) and Wilcoxon Signed rank test. Power functions. Comparison with t test.

IV. (3 weeks) Two sample tests (including the Mann-Whitney test for independent samples based on ranks), both with tables and normal approximations. Fisher exact test and chi-square tests. Two sample problems depend on whether the samples are independent or dependent (paired comparisons), an important distinction.

V. (1 week) Chi square tests, one way and two way contingency tables.

VI. (1 week) Kruskal Wallace (one-way ANOVA based on ranks).

VII. (1 week) Friedman test for comparison of rankings. (For example, analyzing the ordered brand preferences of a random group of individual consumers).

VIII. (1 week) Runs test for independence. (For example, testing whether a basketball team's ordered record of wins and losses can be modelled as a binomial population with a constant success probability.)

IX. (1 week). Order statistics. (For example, to test the lifetimes of light bulb, the data might consist of the time until burnout of the first 20 bulbs to fail, without waiting for the entire sample to fail.)

X. (1 week) Spearman's rho and Kendall's tau (non-parametric correlations).

XI. (1 week) brief account of Loess and robust regression - if time permits.

Homework will be assigned in class and is an essential part of the course. Class time will frequently be devoted to homework and exam questions will often be based upon homework

problems. While some homework will not be collected or graded, other assignments will be mandatory and will influence final grades.

Final grades will be based mainly upon total points received on the mid-term examination, the three highest scores on four half-period quizzes, the final examination and credit for mandatory homework. Since the lowest quiz grade will be dropped, a student can miss one quiz without penalty. Makeups will not be given for missed quizzes.

Class attendance is not required, though students must of course be present for tests and must hand in mandatory assignments on time whether or not they attend class. Students are expected to arrive on time and to remain until a class ends. In particular students are not free to leave and return to class (e.g., to take a cell phone call). Students who repeatedly arrive late will have their grades lowered in consequence. Students must of course turn off cell phone ringers in class. Students who know in advance they will need to leave before the end of a class should notify the instructor before the start of class and should sit near the door.

Calculators will be needed for homework problems and for quizzes and exams. Unless an announcement is made to the contrary, it should be assumed quizzes and examinations are open book; i.e., books, notes and calculators can be used during exams (though not shared among students). Students are not allowed to access the web during tests.

Students are invited to contact the instructor via e-mail, and can expect to receive prompt responses. Please use "Math581" as a subject when sending messages. E-mail is an additional option, not a substitute for office hours or other face-to-face contact.

Mathematics 582
Computational Methods in Statistical Inference
Syllabus

Course Description

Students will be required to do both an individual data science project and to participate in a joint project utilizing techniques studied in AMAT 581 (Nonparametric Statistics). Students will be required both to make oral presentations of their own work and to critique those of their fellow students.

In the first part of the course students will select or be assigned a data set to be analyzed using nonparametric techniques and to present the results orally. Class time will be devoted to discussions of these data sets. Students are also expected to have met individually with the instructor. Each student will give an oral 20 minute presentation of his or her analysis and conclusions.

In the second part of the course student will be divided into small groups (at most 3 or 4 students) and chose a topic or data set to investigate or analyze. Students will be expected to give brief oral progress reports to the full class.

In the last part of the course students will give oral presentations on their project and hand in their final (group) report.

Grading and Evaluation

Final grades will be based on a student's individual oral presentation (one-third), on his or her presentation of and participation in the group project (one-half) and on participation in class discussions (one-sixth).

Prerequisites

AMAT 581 (Nonparametric Statistics).

Boris Goldfarb • bgoldfarb@albany.edu

Office: ES 120B • Office hours: MWThF

This class is designed to prepare you for AMAT 584, Topological Data Analysis II, next semester. While I will mention data and indicate how the material in this class is relevant to data science, the emphasis here is on special topics in topology developed over the last 100 years. It is remarkable that the goals of topologists were usually within pure mathematics but the current product has a combination of properties that makes it uniquely fit for applications to data science. It is accepted that “data has shape” and so geometric study of data sets is fruitful and essential. What algebraic topology provides are algebraic, computable invariants of that geometry that are coordinate-independent and robust under perturbations and noise.

My goal in this class is to guide you through topics that have found the most successful applications to data science. They will serve as prerequisite knowledge for the second semester of the TDA sequence. This basically means that in 4 months we need to learn things that are usually properly taught in 4 semesters of algebraic topology. The trick is to be very selective and also set the goal of developing intuition behind concepts by working through examples and occasional proofs. Proofs in these course are better thought as mental exercises that develop intuition.

Textbook/notes: There is still no appropriate textbook available for a course of this kind. I have written notes which will be made available to you at no charge. Just a word of warning: the notes are a combination of material written specifically for this course and some material borrowed from other authors as needed—from books, articles and surveys. The result is a self-contained, readable resource. There are references to sources used but also suggestions for further reading. I want to think of the notes as a living document that will incorporate new material and information that appears as the course progresses. What is eventually in the notes depends on the background of students in this class, so your feedback is very much appreciated!

Pace: You will see that the notes have chapters roughly corresponding to the number of full weeks that the class meets during the semester. This is a good indication of the pace we will keep and what is covered each week.

Lectures: Our interaction will consist of lectures during the scheduled class times, Q/A during the lectures, submission/grading of homework assignments (see below), taking/grading in-class exams and take-home projects (see further below).

Homework: The homework assignments will be given out in the form of problem sets from the corresponding modules from the notes. I expect to collect and grade 8 homework problem sets.

Exams: There will be two one-hour in-class exams and two take-home longer projects. They will be spread evenly through the semester.

Grading: Overall course letter grades will be based on a curve obtained from test and homework scores combined according to the following scheme.

homework:	40% (5% each HW)
exams:	30% (15% each)
projects:	30% (15% each)

Prerequisites: This can be the first graduate class taken in mathematics. Linear algebra is a definite plus but general maturity and several undergraduate math classes are enough.

Topics: This list should give an overview of specific topics that correspond to chapters in the class notes.

Chapter 1, Elements of Graph Theory.

Graph theory is the most immediately applicable geometric area. We will see some famous theorems and applications of graphs. In this course, graphs are mainly one-dimensional examples of cellular and simplicial complexes.

Chapter 2, Higher-dimensional Simplicial Complexes.

Simplicial complexes are the most basic geometric input in algebraic topology. Many geometric spaces can be built as simplicial complexes. We will see lots of examples. One of the challenges is to develop some ability to visualize higher-dimensional complexes, at least on the intuitive level.

Chapter 3, Metric Spaces.

Data sets are studied after some sense of distance is introduced between the data points. This often reflects the intrinsic relations between the points. To have a consistent picture, one requires the distance function to satisfy certain properties called metric properties. We will learn elementary consequences of the metric properties. This chapter will also introduce basic notions of point-set topology, illustrated in metric spaces.

Chapter 4, Vietoris-Rips and Čech Complexes.

Given a metric space, one learns to build simplicial complexes generated in various ways from the metric. The Vietoris-Rips and Čech constructions are a good place to see this happen for the first time. They are the most basic but not the most efficient constructions.

Chapter 5, Convex Set Systems, Delaunay Complexes, Alpha Complexes.

These are some constructions which are fancier than those in Chapter 4. They have specific advantages and uses which we survey. The examples are also good practice in thinking about and visualizing concrete simplicial nerve complexes.

Chapter 6, Some General Topology.

This is where a student of topology traditionally starts. We have collected enough motivation for this general notion of a topological space. In order to explain the robustness of TDA applications and also understand why our future computations of homology give consistent results in “noisy” situations, we give an introduction to continuity.

Chapter 7, Homotopy, Homotopy Equivalence, Homotopy Invariance.

This continues the narrative in the preceding chapter. Homotopy invariance of the homology vector spaces and numerical data such as the Betti numbers and the Euler characteristic are the fundamental facts that make homology useful.

Chapter 8, Review of Linear Algebra.

The definition and computation of homology relies on matrix algebra. We will review only the most computationally relevant bits of linear algebra. These include the kernel of a matrix, the range of a matrix, the notion of the quotient vector space.

Chapter 9, More Linear Algebra.

The goal this week is to compute many examples. Some of these examples will be reused in homology computations.

Chapter 10, Simplicial Homology.

Here we define the homology vector spaces of a simplicial complex. We will do some simple computations with the main goal of understanding the relation between the algebra and the geometric meaning of the outcomes.

Chapter 11, More Simplicial Homology.

We will keep computing homology of increasingly interesting and meaningful complexes.

Chapter 12, Even More Simplicial Homology.

We will keep computing... Two points are going to be made clear at this stage. One is the abstract nature of homology and of the complexes that generate it. For example, we will examine nonorientable surfaces and other manifold phenomena that are present in high dimensions. Another point is the fact that topologists rarely use the definition of homology in computations. We will look at the Mayer-Vietoris sequence and learn what it does.

Chapter 13, Vista: Singular Homology, Cellular Homology.

Some more basic topology that will be essential for Morse theory and Reeb graphs and complexes next semester.

Chapter 14, Vista: Cohomology.

Some topics in algebraic topology that will let us tie loose ends from the last two chapters. If time permits, deeper look at the product structure in cohomology and duality.

Boris Goldfarb • bgoldfarb@albany.edu

Office: ES 120B • Office hours: MWThF

The basic template for Topological Data Analysis is to view a data set with an expert- defined distance function as a metric space. The practitioner then uses one of many TDA tools to generate (=compute) an algebraic signature for the set. This signature can be used for feature generation in machine learning. More visual signatures can be used by the data scientists in more immediate and ad-hoc ways, for example for discovering correlations between variables.

The most universal and well-known algebraic signature in TDA is persistent homology. Roughly half of the course will be devoted to defining this signature, computing it using various software implementations, and finally interpreting the results.

The other half of the course will teach another widely used tool, the Mapper. We will see the theory behind this algorithm which has connections to Reeb graphs and further topics in Morse theory. We will also see examples of its use in the machine learning context.

Prerequisites: It is essential that the student has taken the course AMAT 583, TDA I. The only kind of exception would be a very well-versed topology student who knows algebraic topology and some basic manifold theory.

Textbook/notes: There is still no appropriate textbook available for a course of this kind. I have written notes which will be made available to you at no charge. Just a word of warning: the notes are a combination of material written specifically for this course and some material borrowed from other authors as needed—from books, articles and surveys. The result is a self-contained, readable resource. There are references to sources used but also suggestions for further reading.

I want to think of the notes as a living document that will incorporate new material and information that appears as the course progresses.

Pace: The notes have chapters roughly corresponding to 3-5 class meetings during the semester. This is a good indication of the pace we will keep and what is covered each week.

Lectures: Our interaction will consist of lectures during the scheduled class times, Q/A during the lectures, submission/grading of homework assignments (see below), taking/grading in-class exams and take-home projects (see further below).

Homework: The homework assignments will be given out in the form of problem sets from the corresponding modules from the notes. I expect to collect and grade 6 homework problem sets.

Exams: There will be two one-hour in-class exams and two take-home longer projects.

Grading: Overall course letter grades will be based on a curve obtained from test and homework scores combined according to the following scheme.

homework:	30% (5% each HW)
exams:	30% (15% each)
projects:	40% (20% each)

Topics: This list should give an overview of specific topics that correspond to chapters in the class notes.

Chapter 1, Introduction to Persistence.

This will start the introduction to the concept of persistence in general and persistent homology in particular. The main idea is that the persistent (=significant) features in data sets correspond to the features of a family of parametrized simplicial complexes with longer survival times.

Chapter 2, Persistent Homology.

We will organize the persistent homology computations in two ways used in TDA: in the form of barcodes and the persistence diagrams. I will demonstrate several implementations of the standard algorithm based on the Smith form. We will also discuss improvements in the algorithms that improve the complexity of computations to $O(n^3)$, where n is the size of the set.

Chapter 3, Filtration-based Persistence.

We will reinterpret the various nerve constructions that lead to persistence in terms of choices of filtration functions. This will give a unified framework for facts about persistent homology.

Chapter 4, A Survey of Software Implementations.

We will demonstrate how one works with various software implementations of the original persistent homology algorithms (e.g. javaPlex) and the recently developed refinements (e.g. Ripser). Some other implementations such as Dionysius which exploit discrete Morse theoretic manipulations will be postponed till later in the semester.

Chapter 5, Two Case Studies.

In preparation for the first project, we will inspect the case studies which apply persistent homology to two data sets of different kinds. We will finish the week with Q/A discussion of expectations and other guidelines. One case study is an application of persistent homology to analyze aspects of performance of professional sports teams. The second case study uses persistent homology in wages analysis to extract the significant predictors.

Chapter 6, Morse Theory and Discrete Morse Theory.

Classical Morse theory examines the relationship between the topology of a manifold M and real-valued functions defined on M using the level cuts of a function. We will give a survey of this idea, then show how this leads to two developments in TDA. One development is based on discrete Morse theory of Robin Forman. It allows one to simplify the simplicial complexes that appear in TDA before applying the standard TDA pipeline.

Chapter 7, Reeb Graphs; Clustering Algorithms.

This week prepares the foundations for the Mapper algorithm and the clustering algorithms that need to be performed as part of the Mapper. The theoretical background is in Reeb graphs and ultimately Morse theory that we have reviewed.

Chapter 8, Introduction to the Mapper.

The Mapper is a TDA tool which provides a 1-dimensional summary of the data set in terms of a colored graph. It accomplishes the dimension reduction of the information contained in the multi-dimensional data set, creating a visual summary at the same time. The summary can be immediately exploited by the data scientist or fed through additional automated tools.

Chapter 9, Two Case Studies.

There will be at least two case studies presented which should illustrate the use of the Mapper in a routine analysis of the data set. One study is an application to marketing research. The other is to the standard logistic regression on a wage dataset. We will also attempt to analyze a data set of interest to students, live in class.

Chapter 10, Vista: Coverage in Sensor Networks via Persistent Homology.

As the students work on the second, last project in the final week of the course, I will present one of the latest applications of a deep topological subject, the theory of sheaves, to sensor networks. At the moment, the plan is to explain the background and then read through the paper Coverage in sensor networks via persistent homology by Vin de Silva and Robert Ghrist (highlighted as one of the 50 most important scientific developments in 2007 by Scientific American).

AMAT 585 Practical Methods in Topological Data Analysis Spring 2018

Boris Goldfarb • bgoldfarb@albany.edu

Office: ES 120B • Office hours: MWThF

This class is the practicum, the final installment in the Topological Data Analysis sequence (AMAT 583-584-585). The goal of the course, for each participant, is to choose a data science problem whose solution will employ the use of TDA, design the project and successfully complete it.

In terms of the timeline, the course will be divided into roughly 3 equal parts. In the first third of the course, the class will discuss possible topics and problems, choose individual problems to solve in the course of the semester. Meanwhile the instructor will present an example of a successful project, walking the class through the stages of the process.

The second third of the course will be occupied by the analysis of the individual problems and the decision process on the methods to use. In class, the time will be spent on presentations by students explaining their analysis of their problems. Each student will discuss the presentation with the instructor beforehand. I expect half of the class be involved in the presentations in this time period.

The last third of the course will consist of presentations of completed (or almost completed) projects and writing up reports. The other half of the class will present their results.

Grading: Overall course letter grades will be based on two components according to the following scheme.

presentation:	40%
project:	60%

Prerequisites: This course requires the student to have been through the first two semesters of this TDA sequence (AMAT 583-584). Exceptions would require exceptional circumstances such as expertise in TDA.

AMAT-590 "FUNCTION THEORY AND FUNCTIONAL ANALYSIS FOR APPLICATIONS"

SYLLABUS

Function analytical methods are among traditional mathematical tools applied in science and engineering. They play a very important role in numerical analysis and computational mathematics in general. Recent development of Data Science, and, in particular, machine learning further emphasized the importance of these methods for applications. The proposed course is one of the basic courses of our new two year Master's Program in Data Science. It is a prerequisite for subsequent courses in machine learning. The goal of this course is to give an exposure to the background material necessary for successful applications of function analytical tools. The main challenge here arises from the fact that none of the basic topics is even slightly touched in a typical undergraduate curriculum. Here we included only the most essential notions, but even with such an arrangement the amount of material is quite substantial. Unfortunately a two year program does not allow more than one course of this nature, so such important topics as numerical differentiation and integration, gradient methods, Tikhonov's regularization, etc. have been omitted. They might be covered in an elective courses in optimization and numerical analysis.

1. Main topics

1). Lebesgue measure and integral

Outer measure, Lebesgue measurable sets, measurable functions, Littlewood principles. Lebesgue integral, Fatou's Lemma, Monotone convergence and Lebesgue convergence theorems.

2). Spaces of Lebesgue integrable functions

The L_p spaces, Young's, Minkowski, Holder's, and Jensen's inequalities. Convergence in L_p . Bounded linear functionals on L_p .

3). Banach spaces

Definition and simple properties of norm spaces. Completeness and Banach spaces. Examples: l_p spaces and Sobolev spaces W^α , Sobolev norm. Hahn-Banach theorem, separation theorems.

4). Duality

Bounded linear functionals, dual spaces, reflexive spaces. Duality in L_p and l_p spaces. Support functions.

5). Bounded linear operators

Boundedness and continuity, principle of uniform boundedness, Open mapping and closed graph theorems. Integral operators.

6). Hilbert space

Scalar products, Schwarz inequality. Examples: spaces ℓ_2 and L_2 , Hilbert norms on spaces of matrices. The Riesz-Frechet representation theorem. Orthogonality, orthonormal bases, Gram-Schmidt process.

7). Reproducing kernel Hilbert spaces

Definition of Reproducing Kernel Hilbert space, positive definite symmetric kernels, uniqueness and other properties of RKHS. Examples of kernels and corresponding spaces: Heat kernel, connection to a Gaussian process, Laplace kernel, polynomial kernel, negative definite symmetric kernel. Connection to support vector machines.

8). Non-linear analysis in Banach spaces

Gâteaux and Fréchet derivatives of operators. Banach space version of Rolle's theorem. Higher order derivatives and sufficient condition of extremum. Banach space version of Lagrange's principle.

2. Course schedule

The following timetable is proposed for the material coverage;

Lebesgue measure and integral	- 2 weeks;
Spaces of Lebesgue integrable functions	- 1 week;
Banach spaces	- 3 weeks;
Duality	- 1 week;
Bounded linear operators	- 2 weeks;
Hilbert space	- 2 weeks;
Reproducing kernel Hilbert spaces	- 3 weeks;
Non-linear analysis in Banach spaces	- 2 weeks.

It is recommended that there are 3 exams given during this course and a cumulative final. The topics covered by these exams are as follows:

- Exam 1 - Lebesgue measure and integral, Spaces of Lebesgue measurable functions;
- Exam 2 - Banach spaces, Duality, bounded maps, general properties of Hilbert space;
- Exam 3 - Reproducing kernel Hilbert spaces, Non-linear analysis.

The course material contains a number of different topics covered by a number of texts. It is difficult to find a single textbook presenting all the topics in a way suitable for a Master's student. Instead the instructor will provide lecture notes for the course.

References

- [1] N.Dunford and J.T.Schwartz, Linear Operators, part I: General Theory, Interscience Publishers, Inc., New York, 1967.
- [2] E. Kreyszig, Introductory Functional Analysis with Applications, [3] P.D.Lax, Functional Analysis, John Wiley & Sons, 2002.
- [4] M.Mohri, A.Rostamizadeh and A.Talwaker, Foundation of machine Learning, MIT Press, 2012.
- [5] H.L.Royden and P.M.Fitzpatrick, Real Analysis, Fourth Edition, Prentice Hall, 1988.

Math 591 Syllabus, Spring 2017

Mark Steinberger

MAT 591 Optimization methods and nonlinear programming
Class time MWF 1:40–2:35
Prerequisites Basic linear algebra and calculus of several variables
Instructor Mark Steinberger
Title Associate Professor
My office ES 132I
Office hours MWF 12:35–1:30 and by arrangement
Email mark@albany.edu
Please include Math 591 in the subject line.
Text Stephen Boyd and Lieven Vandenberghe,
Convex Optimization. Cambridge University Press.
My home page <http://math.albany.edu/~mark>
Course home page <http://math.albany.edu/~mark/classes/591/>

The book is available online for free via <http://stanford.edu/~boyd/cvxbook/>. (An additional book you might find useful is Dimitri Bertsekas, Nonlinear Programming, Athena Scientific, but it is neither required nor necessary.)

We will cover material from Chapters 2–5 and 9–11 from Boyd and Vandenberghe. The topics of interest include the basics of convex sets and convex functions, followed by specific techniques to optimize convex functions, e.g., Newton's method, gradient descent, linear programming, quadratic optimization, and semidefinite programming.

The basic context is as follows. We are given a real-valued function $f(x_1, \dots, x_n)$ of several variables, and we wish to find a point $x = (x_1, \dots, x_n)$ at which f attains its minimum value. In practice, the points x which are feasible for practical problems are often subject to constraints. Conceptually, this says that we should regard f as having a fixed domain $C \subset \mathbb{R}^n$, i.e., f is a function $f: C \rightarrow \mathbb{R}$.

We wish to find the points at which f attains its minimum value, and what that value is.

In practice, this is an impossible problem, so we need to place additional conditions on C and on f . In most of the cases we consider here, we shall insist that both C and f be convex.

The convexity of C means that for any pair of points $x, y \in C$, the line segment between them lies in C , i.e.,

- (1) $(1-t)x + ty \in C$ whenever $x, y \in C$ and $t \in [0, 1]$. The convexity of f means that
(2) $f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$ whenever $x, y \in C$ and $t \in [0, 1]$.

When working with only one variable, i.e., when $n = 1$, this can be seen in terms of familiar ideas from calculus. (1) means that C is an interval, and (2) means f is “concave up”. If f is twice differentiable, then one method of solution is a second derivative test.

We will examine a variety of issues in the n -dimensional case, including methods of finding approximate solutions, and contexts in which convex optimization can be used to solve important applied problems.

There will be three in-class exams for this course, along with group work done either in-class or as homework. The point count for the final grade is determined as follows:

Group work 13%
Each in-class exam 29%

Class attendance is absolutely essential. If for some reason you need to miss class, it is imperative that you get notes from someone. And finding someone who takes good notes isn't always easy. :-) Also, it is usually easier to digest the material if you see and hear it presented. In any case, you are expected to

attend class. The university's medical excuse policy is available at http://www.albany.edu/health_center/medicaexcuse.shtml.

You are strongly encouraged to discuss this material with each other and with me, both in office hours and in class. Verbalizing mathematical questions is a very useful step toward understanding them. Classroom discussion is strongly encouraged. Please ask questions! If there is something you don't understand or can't follow, there will be a number of other people in the class in the same boat. So a number of people will benefit if you ask.

It is very important to stay current with the material. If you fall behind, it will be hard to catch up. And if you are having trouble, please do come to office hours early on. If you leave it until the last minute, you probably won't be able to learn it in time.

But office hours are not only for those who have fallen behind. Office hours are extremely helpful for learning and I seriously enjoy discussing the material with students and helping them learn. It is especially useful to work with a group of students. The synergy really helps everyone learn.

Other than during exams, you are strongly encouraged to work with other students and with me. The in-class work will be mainly in groups. During exams, you may ask me questions but should not communicate with anyone else. Use of phones during exams is prohibited. The university's academic integrity policy is available via http://www.albany.edu/undergraduate_bulletin/regulations.html

Syllabus for AMAT592: Machine Learning

August 23, 2016

Course Description

The primary goal is to provide students with the tools and principles needed to solve both the traditional and modern data science problems encountered in practice. The course requires the basic knowledge of function analysis and optimization courses that the students learn in the first year.

In particular, the course covers a wide variety of topics in machine learning. It introduces the key terms, concepts and methods in machine learning, with an emphasis on developing critical analytical skills through hands-on exercises of actual data analysis tasks. At the same time, it will cover modern machine learning topics such as Boosting and online learning for large-scale data analysis. In addition, the students will practice basic programming skills to use software tools in machine learning. The main programming language in this course will be MATLAB/OCTAVE which is one of the most widely used languages for data analysis.

Goals: At the completion of the course the student will:

- be able to define and use key terms, concepts and methods in machine learning;
- be able to apply a number of complex and advanced mathematical and numerical techniques to a wide range of problems and domains.
- be able to identify the compromises and trade-offs which must be made when translating theory into practice;
- be able to utilize machine learning models and computation tools to perform basic data analysis on data sets from practical problems;
- be able to critically interpret and summarize data analysis results;

Main Topics

1. Introduction

What is machine learning category of machine learning tasks, generalization, cross validation

2. Description of data

Representation of data matrix, description of data (correlation, variance, mean, and standardization)

3. Probabilistic formulation of classification

Linear prediction problems, perceptron algorithm, risk bounds, misclassification error, convex loss function, generalization error

4. Support Vector Machine (SVM)

Large margin classifier, recap of reproducing kernel Hilbert space (RKHS), representer theorem, dual formulation, quadratic programming

5. Statistical learning theory

Rademacher average, VC dimension, Concentration inequalities, generalization bounds

6. Boosting

AdaBoost algorithm, coordinate descent formulation of AdaBoost, convergence analysis

7. Regression

Least square regression, sparse learning with L1-regularization (LASSO), Bayesian formulation, kernel ridge regression, maximum likelihood (ML), maximum a posteriori (MAP)

8. Unsupervised Learning

Principal component analysis (PCA), K-means clustering, Gaussian mixture model (GMM), expectation maximization algorithm (EM)

9. Online learning algorithms

Halving algorithm, exponential weight/weighted majority, online convex optimization, online gradient descent

10. Advanced topics

Multiple kernel learning (MKL) for data integration, distance metric learning

Course Schedule

A tentative timetable is proposed for the materials covered in the course:

1. Introduction and description of data: 2 weeks
2. Probabilistic formulation of classification: 1.5 week
3. Support Vector Machines: 2 weeks
4. Statistical learning theory: 2 weeks
5. Boosting: 1 week
6. Regression: 2 weeks
7. Unsupervised learning: 2.5 weeks
8. Online Learning: 2 weeks
9. Advanced topics: 1 week

There will be a lab practice fortnightly. The lecturer will be in the lab to guide the students to practice data analysis tasks and answer related programming questions which are assigned in the homework.

Grading and Evaluation

There will be four homework and one final exam. The final exam consists of 20% of the course grade and each homework gives 20%. Specifically, the final exam are on written questions, i.e. there will be no questions on practical programming. In each homework, there will be theoretical machine learning questions and real data analysis tasks which require practical programming as well as critical interpretation of data analysis results. In particular, the topics covered by homework are as follows:

1. Homework 1: Data Description and Probabilistic Formulation of Classification
2. Homework 2: SVM and Statistical Learning Theory
3. Homework 3: Boosting and Regression
4. Homework 4: Unsupervised Learning and Online Learning

Syllabus for AMAT593: Practical Methods in Machine Learning

Course Description

In this practical course, each student will have chances to conduct a data science project using machine learning knowledge that they have learned in the previous semester, and write a critical report about the obtained results. In particular, the course will mainly consist of three important units, which are explained as follows.

- In the first part of the course, students are divided into groups, and each group needs to select one paper for presentation from a given list of scientific papers which are already published in machine learning conferences such as ICML and NIPS etc. The instructor will give an example presentation of a paper and meet regularly with the individual group. In the presentation, each group needs to produce a PPT presentation, explaining its motivation, method used and the main results of the paper. This part is tentatively scheduled to be between weeks 1-5.
- In the second third of the course, the class will discuss possible machine learning topics or data analysis problems, and then select a particular data science project to work on. The student will need to select machine learning algorithms to solve this practical data analysis problem. During the class, most of the time will be spent on presentations by students explaining their progress of analyzing their problems. This part takes most of the time which is scheduled between weeks 6-12.
- The last part of this course will consist of the final presentation of the individual projects and writing up the final reports. The other half of the class will present their results. This will happen during weeks 13-16.

Grading and Evaluation

The final grade is based on the evaluation on one group-based presentation, one final individual- project presentation and one final report. The group presentation will consist of 30% of the final grade, and the presentation for the individual project is 20%. The final project will contribute 50% to the final grade.

Prerequisites

This course requires the student to have taken the machine learning course (AMAT 592) or equivalent.

Document 6

Department of Mathematics and Statistics

University at Albany, SUNY

Tenure-track position

The Department of Mathematics and Statistics at the University at Albany, State University of New York, invites applications for a tenure-track assistant professor position in the area of "machine learning", to start in Fall 2017.

We are looking for candidates who will significantly contribute to the department's research, closely collaborate with existing members of the department, and enhance our undergraduate and graduate programs.

Candidates should possess excellent research credentials as demonstrated by their PhD dissertation, publications, external funding, and as supported by letters of recommendation from experts in the field. Also of great importance are teaching credentials demonstrated by student evaluations and/or teaching awards and supported by letters of recommendation.

Candidates are required to have a PhD or an equivalent doctoral degree in Mathematics from a university accredited by the U.S. Department of Education or an internationally recognized accrediting organization. Postdoctoral experience and a successful record of external funding are highly desirable. All candidates must address in their applications their ability to work with a culturally diverse population and should provide statements on teaching and research.

Candidates should apply using the University employment portal <http://albany.interviewexchange.com> and have at least four letters of recommendation sent to the Chair, Department of Mathematics and Statistics, University at Albany, Albany, NY 12222. At least two letters should address the candidate's research and at least one should address the candidate's teaching. These letters can also be emailed to mstessin@albany.edu. The deadline for applications is January 1, 2017.

The University at Albany is an EO/AA/IRCA/ADA Employer.

Document 7

Department of Mathematics and Statistics

University at Albany, SUNY

Tenure-track position

The Department of Mathematics and Statistics at the University at Albany, State University of New York, invites applications for a tenure-track assistant professor position in the area of "computational topology/computational algebra", to start in Fall 2018.

We are looking for candidates who will significantly contribute to the department's research, closely collaborate with existing members of the department, and enhance our undergraduate and graduate programs.

Candidates should possess excellent research credentials as demonstrated by their PhD dissertation, publications, external funding, and as supported by letters of recommendation from experts in the field. Also of great importance are teaching credentials demonstrated by student evaluations and/or teaching awards and supported by letters of recommendation.

Candidates are required to have a PhD or an equivalent doctoral degree in Mathematics from a university accredited by the U.S. Department of Education or an internationally recognized accrediting organization. Postdoctoral experience and a successful record of external funding are highly desirable. All candidates must address in their applications their ability to work with a culturally diverse population and should provide statements on teaching and research.

Candidates should apply using the University employment portal <http://albany.interviewexchange.com> and have at least four letters of recommendation sent to the Chair, Department of Mathematics and Statistics, University at Albany, Albany, NY 12222. At least two letters should address the candidate's research and at least one should address the candidate's teaching. These letters can also be emailed to mstessin@albany.edu. The deadline for applications is January 1, 2017.

The University at Albany is an EO/AA/IRCA/ADA Employer.

Document 8



Office of the Chief Information Officer

January 25, 2016

Professor Steinberger
Mathematics Statistics
University at Albany, SUNY

Dear Professor Steinberger:

I am writing to indicate that after reviewing your proposal and emails, I have determined that the new Masters in Data Science program will not have any fiscal impact on Information Technology Services (ITS).

As the DATA SUNY program gears up we can talk about how that offering will provide the needed support for the practicum proposed for second year students that you anticipate will be needed in fall 2018.

This sounds like an exciting program, and we in ITS are excited to see the new offering started here at the University. If I can be of any additional assistance please contact me.

Sincerely,

A handwritten signature in cursive script that reads "Carole Sweeton".

Carole Sweeton

Interim CIO
Information Technology Services

Information Technology Building
1400 Washington Avenue, Albany, NY 12222
PH: 518-956-8080
www.albany.edu

Document 9

An Evaluation of the Resources of the University at Albany Libraries in Support of a Master's program in Data Science

Introduction

The University Libraries collect, house, and provide access to all types of published materials in support of the research and teaching of the schools, colleges, and academic departments of the University. This evaluation considers those portions of the libraries' collections and services that would support a Master's program in Data Science in the Department of Mathematics and Statistics. Currently, the University Libraries support undergraduate, masters, and doctoral studies and research, as well as faculty research in the Department of Mathematics and Statistics.

Library Collections

The University Libraries are among the top 115 research libraries in the country. The University Library, the Science Library, and the Dewey Graduate Library contain more than two million volumes and over 2.9 million microforms. The Libraries provide access to more than 97,000 online serials and over 340,000 online books. Whenever possible, current subscriptions are available online. Additionally, the Libraries serve as a selective depository for U.S. Government publications and house collections of software and media.

Books

The University Libraries routinely acquires books to support the teaching and research of the Department of Mathematics and Statistics. Books in print and ebook formats are acquired on approval, via standing order for series, and firm order. Many of the books acquired will support a Master's program in Data Science. It should be noted that the Department of Computer Science also has an emphasis in data science, and books are acquired to support that interest. During FY2015, the University Libraries spend over \$28,000 on books to support the Department of Mathematics and Statistics programs. If additional books are needed to support a Data Science program, the Subject Librarian for Mathematics and Statistics should be able to acquire them by firm order.

Journals

Currently, the University Libraries subscribe to 156 mathematics and statistics journals. There are an additional 136 journals tagged as "mathematics" (including statistics journals) in Elsevier's ScienceDirect full-text database. Additional mathematics and statistics journals are found in these full-text sources: Academic Search Complete, Applied Science and Technology Source, and SIAM (Society for Industrial and Applied Mathematics) Online Journals. Although mostly oriented toward computer science and engineering, the ACM Digital Library and the IEEE / IET Electronic Library also contain journal and magazine articles on data science. In FY2015, the University Libraries spend nearly \$187,000 on journals to support programs in the Department of Mathematics Statistics.

Ulrich's International Periodicals Directory lists seven general data science journals. Three of the journals are open access, and are currently available: *Data Science Journal*, *EJP Data Science*, and *Oxford Journal of Intelligent Decision and Data Science*. Another journal, *Journal of Data Science*, is available through our subscription to the Academic Search Complete full-text database. The remaining four journals are not available. They are:

Advances in Data Science and Adaptive Analysis (2017 cost: \$449)

Annals of Data Science (2017 cost: \$384)

International Journal of Data Science (2016 cost: \$665)

International Journal of Data Science and Analytics (2017 cost: \$471).

If any of these journals or other journals are required, funding would be needed.

Reference Collection

The reference section of the Science Library houses a collection of resources in support of the mathematics and statistics programs. Some of the resources are available in the Science Library and some are available online. There are several reference books related to data science, including these titles:

Cambridge Dictionary of Statistics
Concise Encyclopedia of Statistics
Encyclopedia of Computer Science
Encyclopedia of Data Warehousing and Mining
International Encyclopedia of Statistical Science
Oxford Dictionary of Statistics
The Princeton Companion to Applied Mathematics.

If additional reference resources are needed, the Subject Librarian for Science Reference should be able to acquire them by firm order.

Databases and Digital Collections

The University Libraries currently subscribe to several databases and digital collections important to mathematics and statistics. Those databases are listed and described below.

Comprehensive Databases

MathSciNet - Based on *Mathematical Reviews* and *Current Mathematical Publications*, *MathSciNet* provides comprehensive coverage of pure and applied mathematics. Entries include abstracts, author summaries, or in many cases, signed critical reviews. It covers journal articles, books, and conference proceedings back to 1940.

The Jahrbuch Project: Electronic Research Archive for Mathematics - The Jahrbuch Project complements *MathSciNet*. It is based on *Jahrbuch über die Fortschritte der Mathematic*, and serves as an index to important mathematical publications for the time period of 1868 - 1942. Links are provided to publications that have been retrodigitized.

Current Index to Statistics (CIS) - A joint venture of the American Statistical Association and the Institute of Mathematical Statistics, *CIS* is a bibliographic index to publications in statistics and related fields. It currently indexes the entire contents of more than 162 core journals, selectively indexes an additional 1,200 journals, and indexes some 11,000 mathematical statistics books and conference proceedings. *CIS* is international in scope.

Digital Collections/Full-Text Databases

SIAM Journals Online – Provides access to the thirteen journals published by the Society for Industrial and Applied Mathematics. Online access is available from 1997 forward for all of the journals, except for those that began publication after 1997.

ScienceDirect - Contains the full-text of over 1,100 Elsevier journals. Backfile coverage starts with 1995.

Academic Search Complete - It is a scholarly, multidisciplinary, database, that contains more than 7,000 full-text periodicals (6,000+ are peer reviewed) in social sciences, humanities, mathematics, science, and technology.

Applied Science and Technology Source - Providing access to the full-text from more than 1,400 journals and magazines, this database covers academic journals, trade journals, professional society journals, conference

proceedings, and other resources. Subjects covered include applied mathematics, artificial intelligence, chemistry, computing, information technology, energy, engineering, materials, robotics, solid state technologies, and textiles.

Related Databases

Web of Science - *WoS* indexes the core journals for all science and technology subjects, including mathematics and statistics. Besides keyword and author searching, one of its key features is the ability to track an author's citation and determine who has cited that work.

Scopus – *Scopus* indexes and abstracts the contents of 21,500+ journal titles from over 5,000 publishers. It also covers conference proceedings, book series, and scientific Web pages and patents. All subjects are covered. It tracks an author's work and determines who has cited that work.

These databases should be able to support the proposed Data Science program. If additional databases are required, funding would be needed.

Interlibrary Loan and Delivery Services

The University Libraries' Interlibrary Loan (ILL) Department borrows books and microforms, and obtains digital copies of journal articles and other materials not owned by the Libraries from sources locally, state-wide, nationally, and internationally. ILL services are available at no cost to the user for faculty, staff, and students currently enrolled at the University at Albany. Users can manage their requests through the use of ILLiad, the University Libraries' automated interlibrary loan system, which is available through a Web interface at <https://illiad.albany.edu/>.

The University Libraries also provide delivery services for books and articles housed in any of the three libraries. Books can be delivered to one of the libraries or for faculty, to departmental addresses. Articles are scanned and delivered electronically via email. The Libraries also provide free delivery services to the home addresses of online learners and people with disabilities. Delivery services are managed through ILLiad as well.

Summary

The University Libraries have been committed to build and maintain collections in support of mathematics and statistics. In FY2015, the University Libraries spend nearly \$226,000 on materials for mathematics and statistics programs. The budget for mathematics and statistics should be able to accommodate required books and reference resources for data science. If additional journals or databases are required, funding will be needed. Materials the University Libraries does not own or provide access to can be obtained using interlibrary loan services.

Michael Knee
Subject Librarian for Mathematics and Statistics
September 2, 2016



External Evaluation Report

Form 21D

Version 201-08-02

The External Evaluation Report is an important component of a new academic program proposal. The external evaluator's task is to examine the program proposal and related materials, visit the campus to discuss the proposal with faculty and review related instructional resources and facilities, respond to the questions in this Report form, and submit to the institution a signed report that speaks to the quality of, and need for, the proposed program. The report should aim for completeness, accuracy and objectivity.

The institution is expected to review each External Evaluation Report it receives, prepare a single institutional response to all reports, and, as appropriate, make changes to its program proposal and plan. Each separate External Evaluation Report and the Institutional Response become part of the full program proposal that the institution submits to SUNY for approval. If an external evaluation of the proposed program is required by the New York State Education Department (SED), SUNY includes the External Evaluation Reports and Institutional Response in the full proposal that it submits to SED for registration.

Institution: University at Albany

Evaluator Name (Please print.): Gunnar Carlsson

Evaluator Title and Institution: Professor of Mathematics, Emeritus, Stanford University and President and Founder, Ayasdi Inc.

Evaluator Signature:

Proposed Program Title: Data Science

Degree: Master of Science

Date of evaluation: On site visit February 21-22, 2017. Report dated March 22, 2017

I. Program

1. Assess the program's **purpose, structure, and requirements** as well as formal mechanisms for program **administration and evaluation**. Address the program's academic rigor and intellectual coherence.

The purpose of the program is to provide training for prospective data scientists, that includes the latest developments within the subject. A student graduating from the program should have the necessary skills to be employed as a data scientist in private and public sector organizations, and would have access to the most recent developments in the subject.

The structure of the program makes a great deal of sense. It incorporates the key machine learning as well as statistical methods that are crucial to success as a data scientist, and gives an introduction to new machine learning methods under the heading of "topological data analysis". In addition, it requires several "practicum" courses, which if I understand correctly will focus on the day to day, low level tasks required for data science. Finally, there are some options for advanced work, consisting of a course in stochastic processes or in linear algebra, or an additional practicum course. Taken together, this set of requirements is coherent and quite rigorous.

In terms of administration, the program will be formally administered by Mark Steinberger, the Director of Graduate Study in the department, who will also be teaching a course in the program. From my discussions with the faculty who will be teaching in the program, it appears that there is a great deal of commitment to it, and that they will function well together as a group in running it.

One comment I would add is that I think it is imperative that every student who goes through the program achieve some training in the “nuts and bolts” of preparing data for analysis. This aspect of the subject is referred to as “extract, transform, and load” (ETL) in industry. This means exposure to the languages Python and R, to methodologies such as Microsoft’s pivot tables, and to ways of dealing with some common unstructured data types, such as plain text, images, databases of molecules, etc. A student who does not have this experience will not be able to land in a position and become productive immediately. As the program develops I would be happy to answer questions about the desirable scope of this kind of experience.

2. Comment on the **special focus** of this program, if any, as it relates to the discipline.

The special focus of this program is clearly something very new in the development of mathematics as a discipline. In my view, the development of Data Science as a focus in mathematics is a very desirable thing, both from the point of view of the discipline as well as from the point of view of the entities that will be hiring the graduates of the program. For mathematics, it gives a whole new class of problems to be worked on, that have very significant consequences. This is important in that it will improve the quality of the theory, as well as make it more relevant. From the point of view of the users of the discipline outside of academia, it is my view that in working on problems related to data, there are at least two components, one being *what* to compute, and the other being *how* to compute it. In my view the second of these problems is ideally dealt with by computer scientists, although mathematicians can certainly have very useful input to the algorithm development. On the other hand, the first problem about what to compute is most naturally the province of mathematicians, statisticians, and domain specialists. Furthermore, it is in most cases the most critical and difficult one, since the computational around size and capacity have been worked on for a substantial period of time.

3. Comment on the plans and expectations for **self-assessment and continuous improvement**.

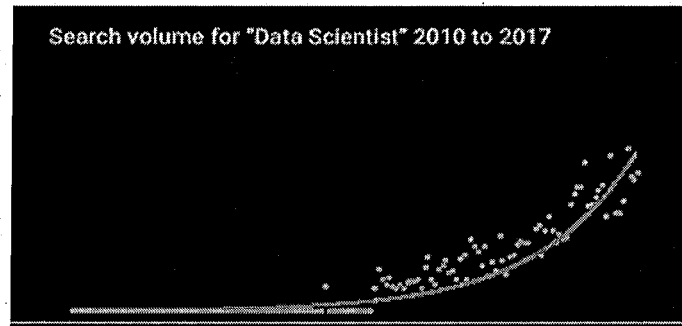
The plans are to follow the usual protocols for assessment of graduate programs, and they will be tracking employment numbers for the graduates once the program is up and running. This is likely sufficient, but I would suggest additionally that for the first few years, feedback from graduates about their experiences after entering the workforce would be extremely useful in tuning the curriculum. Similarly, feedback from employers of graduates would also be useful. There appears to be a lot of positive reaction to the plan from potential employers described in the program description, and it should be checked that the program indeed lives up to the expectations.

4. Discuss **the relationship** of this program to other programs of the institution and collaboration with other institutions, and assess available support from related programs.

In other academic institutions, the typical stakeholders in a program such as are computer science, statistics and sometimes mathematics, or a computational mathematics program. Given that statistics is a part of mathematics at Albany, many of the questions about relationships are moot. However, as the program develops, it may very well be that finding some way to work with computer science on some aspects of the program, particularly the practicum components above, and specifically with the ETL aspects discussed in point 1. CS may be interested in developing courses in software used, as well as in techniques for transforming data. This should in no way be allowed to interfere with the strong focus of the program on the things which are best handled in mathematics and statistics.

5. What is the evidence of **need and demand** for the program locally, in the State, and in the field at large? What is the extent of occupational demand for graduates? What is the evidence that demand will continue?

The need for data scientists is growing at an enormous rate, and it is likely the most rapidly growing category of positions. This can be verified by various statistics, gathered by LinkedIn and others. As an indication, the graph below shows the growth of search volume for the term "Data Scientist" over times.



It is anecdotally the case that most companies report large amounts of data that they have gathered is going unused due to the lack of analysts for it. It is a known fact that the gathering of data is growing enormously. It is therefore very clear that companies see the need for people to analyze the data they have gathered. In my view, the problem of extracting information and understanding from big and complex data is the most challenging problem of our time, both because of its intellectual appeal as well as the importance of the solutions to all aspects of society. The need for this kind of work can only grow going forward.

II. Faculty

6. **Evaluate the faculty**, individually and collectively, with regard to training, experience, research and publication, professional service, and recognition in the field.

The thrust of the program is in three directions, namely machine learning, topological data analysis, and statistical methods. The different faculty members cover the three different directions as follows. (I have understood that there is a strong likelihood that Michael Lesnick will be coming to Albany, and so am including him)

Goldfarb: topological data analysis

Hildebrand: probability/statistics

Lesnick: topological data analysis, machine learning, statistics

Munch: topological data analysis, machine learning, statistics

Reinhold: probability/statistics

Sherman: probability/statistics

Steinberger: machine learning, topological data analysis

Stessin: probability/statistics, machine learning

Varisco: topological data analysis

Ying: machine learning.

I fully expect that the members of the group will broaden and eventually be able to take on teaching beyond the boundaries I describe above. That said, in future hiring, it might be useful to consider one additional person whose primary focus is machine learning. It is encouraging that over 2/3 of the faculty have topics related to the program as part of their primary research focus.

7. **Assess the faculty in terms of number and qualifications and plans for future staffing.** Evaluate faculty responsibilities for the proposed program, taking into account their other institutional and programmatic commitments. Evaluate faculty activity in generating funds for research, training, facilities, equipment, etc. Discuss any critical gaps and plans for addressing them.

There is a plan to add three new faculty, which together with the existing faculty should be enough to cover the program. I would recommend that one of these faculty members come with machine learning as his/her scholarly focus. It appears that with these new faculty members, the courses will be sustainable. The advising and administrative portions of the program should be manageable. It is explicitly stated that three existing faculty members will devote 100% of their advising/mentoring time to the program, which is quite solid for the first 2 years. If the projected enrollment numbers materialize, that may have to be revisited, and the department may need to devote additional advising resources to this group by the time the 4th or 5th year is reached.

8. Evaluate credentials and involvement of **adjunct faculty and support personnel.**

N/A, as far as I can tell.

III. Students

9. Comment on the **student population the program seeks to serve**, and assess plans and projections for student recruitment and enrollment.

The population being served consists of undergraduates with a background in a mathematical discipline, such as mathematics, statistics, or computer science who want specialized training that will permit them to be hired as a data scientist in a public or private entity. Because of geography, one expects that there will be heavy

representation from New York State, but there are plans to recruit students from China, as well. Further, the program has a unique focus which could be expected to attract students from other parts of the United States. The plan is to recruit world wide, but it would appear that one could do a very effective job recruiting in state By having faculty members do personal recruiting for campuses in the SUNY system. The subject of data science is very attractive to many students, and lends itself to expositions of things that are of a great deal of interest. I would believe that such in person recruiting efforts would be most effective. I also think that several of the faculty would be very adept at such presentations.

10. What are the prospects that recruitment efforts and admissions criteria will supply a sufficient pool of highly qualified applicants and enrollees?

Given the (valid) perception that data science is experiencing a great demand for people, and the unique nature of the program being offered, I expect that there will be a sufficient pool to fill the numbers mentioned in the projected enrollment tables. There are already arrangements being worked out for recruiting from China, and my guess is that personal recruitment within the SUNY system will yield good candidates.

11. Comment on provisions for encouraging participation of persons from underrepresented groups. Is there adequate attention to the needs of part-time, minority, or disadvantaged students?

The plan is to work with the University's Office of Diversity and Inclusion to obtain contacts within the historically Black colleges and universities. Working with the Diversity office is likely the best way to obtain such contacts, and also to obtain feedback on the Department's efforts.

12. Assess the system for monitoring students' progress and performance and for advising students regarding academic and career matters.

The monitoring of student's progress will be done in the usual way, through the advising system. It will be particularly important to establish early that the advising should not be pro forma but should involve systematic and regular meetings with students to assess their performance, both from their own point of view as well as the departments. It will fall on the advisors to be well aware of employment opportunities, perhaps more so than might be the practice in other programs. I would also recommend that advisors be aware of internship opportunities, and work to generate them.

13. Discuss prospects for graduates' post-completion success, whether employment, job advancement, future study, or other outcomes related to the program's goals.

The prospects seem to be very strong, given the high demand for people with the expertise provided by the program.

IV. Resources

14. Comment on the adequacy of physical resources and facilities, e.g., library, computer, and laboratory facilities; practica and internship sites or other experiential learning opportunities, such as co-ops or service learning; and support services for the program, including use of resources outside the institution.

Physical resources do not appear to be an issue here.

15. What is the **institution's commitment** to the program as demonstrated by the operating budget, faculty salaries, the number of faculty lines relative to student numbers and workload, and discussions about administrative support with faculty and administrators?

The Department's commitment is clearly strong, evidence for which is my discussions with the department members at the on site visit. The University's commitment is also strong, and I base that statement on my discussions with the relevant Dean and other administrators.

V. Summary Comments and Additional Observations

16. Summarize the **major strengths and weaknesses** of the program as proposed with particular attention to feasibility of implementation and appropriateness of objectives for the degree offered.

The major strength is the mathematical sophistication of the program, and its focused nature on Data Science, which is an area for which demand is growing rapidly. I do not see major weaknesses, but would recommend that the department be very honest in self assessing the program, making sure that it handles not only the theory but also the "nuts and bolts" aspects of data science. This is not difficult to do, but is extremely important for the success of the students. I would also make sure to provide exposure of the students to people from industry, so that they can see what is really critical in an industrial or government setting.

17. If applicable, particularly for graduate programs, comment on the ways that this program will make a **unique contribution** to the field, and its likelihood of achieving State, regional and/or national **prominence**.

The program has a very unique focus on the mathematical side. This is very important, since the role of mathematics is often to decide what to compute, as opposed to how to compute it. Mathematicians, in collaboration with subject matter experts, can make good decisions on the "what" issue, and produced good ways of modeling data.

18. Include any **further observations** important to the evaluation of this program proposal and provide any **recommendations** for the proposed program.

Recommendations are as follows.

- Make sure to have adequate coverage of the ETL (Extract, transform, and load) capabilities. This means familiarity with Python and R, familiarity with transform and aggregation techniques, feature engineering, and specific examples of working with complex unstructured data, such as free text, images, molecules, etc.
- At least in the initial few years, have frequent departmental discussions of the nature of the program. You will likely find that some things need to be added and others can be deleted, and this kind of rapid evaluation can iron out kinks before they become serious.
- Take very seriously the creation of a good pipeline of internship programs for students to participate in.
- Bring in industrial and public sector people who work with or use data to present in a colloquium or lecture series, to acquaint students with the problems arising outside academia.
- Do intense personal recruiting from the SUNY system.



The State University
of New York

External Reviewer Conflict of Interest Statement

I am providing an external review of the application submitted to the State University of New York by:

University at Albany

(Name of Institution or Applicant)

The application is for (circle A or B below)

A) New Degree Authority

B) Registration of a new academic program by an existing institution of higher education:

Master of Science in Engineering.

(Title of Proposed Program)

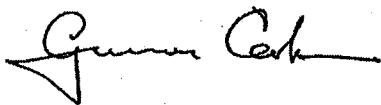
I affirm that I:

1. am not a present or former employee, student, member of the governing board, owner or shareholder of, or consultant to the institution that is seeking approval for the proposed program or the entity seeking approval for new degree authority, and that I did not consult on, or help to develop, the application;
2. am not a spouse, parent, child, or sibling of any of the individuals listed above;
3. am not seeking or being sought for employment or other relationship with the institution/entity submitting the application?
4. do not have now, nor have had in the past, a relationship with the institution/entity submitting the application that might compromise my objectivity.

Name of External Reviewer (please print):

Gunnar Carlsson

Signature:





External Evaluation Report

Form 210

Version 201-08-02

The External Evaluation Report is an important component of a new academic program proposal. The external evaluator's task is to examine the program proposal and related materials, visit the campus to discuss the proposal with faculty and review related instructional resources and facilities, respond to the questions in this Report form, and submit to the institution a signed report that speaks to the quality of, and need for, the proposed program. The report should aim for completeness, accuracy and objectivity.

The institution is expected to review each External Evaluation Report it receives, prepare a single institutional response to all reports, and, as appropriate, make changes to its program proposal and plan. Each separate External Evaluation Report and the Institutional Response become part of the full program proposal that the institution submits to SUNY for approval. If an external evaluation of the proposed program is required by the New York State Education Department (SED), SUNY includes the External Evaluation Reports and Institutional Response in the full proposal that it submits to SED for registration.

Institution:

Evaluator Name (Please print.): Sayan Mukherjee

Evaluator Title and Institution: Professor at Duke University in Statistical Science, Mathematics, and Computer Science

Evaluator Signature:

Proposed Program Title: Masters in Data Science

Degree: MS

Date of evaluation: February 5, 2017

I. Program

1. Assess the program's **purpose, structure, and requirements** as well as formal mechanisms for program **administration and evaluation**. Address the program's academic rigor and intellectual coherence.

The program is a MS in Data Science that will reside in a mathematics department. The idea is to develop the theory or foundations of data science with a focus on mathematics and a bridge between theoretical mathematics and real data. The program will have as foci a track in machine learning (ML), a track in topological data analysis (TDA), and also optimization. Computation is ubiquitous in the program. I am not concerned that statistics and computer science are not involved in the program as all the faculty I spoke with said computing is central and will be in all courses, some of the faculty have strong numerical analysis and machine learning backgrounds. I am also not concerned with the lack of statistics as some faculty that will teach in ML and TDA are very aware and concerned with uncertainty, reproducibility, and formulating stochastic models.

2. Comment on the **special focus** of this program, if any, as it relates to the discipline.

The special focus of this program is to develop a Data Science Masters that builds a theoretical foundation based on mathematics. For example students will learn functional analysis and algebraic topology in core classes for the ML and TDA track, respectively. The unique idea is quickly moving from foundational ideas in

pure mathematics to applications of these ideas to data. This can be very attractive to mathematical students that want to engage real world problems.

3. Comment on the plans and expectations for **self-assessment and continuous improvement**.
The plans and expectations for self-assessment and continuous improvement are not well developed. The department should work with the administration to develop such a plan. The chair of the mathematics department has ideas and the proposal has some plans but it needs to be further developed.

4. Discuss **the relationship** of this program to other programs of the institution and collaboration with other institutions, and assess available support from related programs.
This program is currently not linked to other programs at U Albany. As the program develops it should make links with other programs at U Albany that are interested in data analysis, for example business analytics programs. However, this should only happen after the program has formed it's own structure and culture.

5. What is the evidence of **need and demand** for the program locally, in the State, and in the field at large? What is the extent of occupational demand for graduates? What is the evidence that demand will continue?
There is a great deal of need for data scientists in Albany and New York. The job growth in data science will continue to increase and there will be demand for graduates. With the increase in technology and computational power to collect and process data there is strong evidence that there will be a continued need for data science. If the program can bridge students from pure mathematics to applied data analysis it will flourish.

II. Faculty

6. **Evaluate the faculty**, individually and collectively, with regard to training, experience, research and publication, professional service, and recognition in the field.

The mathematics faculty at Albany has a real strength in pure mathematics, especially topology and algebra. There are two faculty members that have experience with data analysis: Elizabeth Munch has a strong record in Topological Data Analysis (TDA) and Yeming Ying has a strong record in Machine Learning (ML). TDA and ML will form two of the three tracks in the MS program. The department needs more faculty with the research program of Elizabeth and Yeming for the MS in Data Science to flourish. Boris Goldfarb is also very involved in the MS program and he is an excellent blend of theory with an appreciation for applications. He has knowledge in computing which also helps. These are the three key faculty in making the MS program run. The program will require more faculty support to flourish. In keeping with the tradition of the department there is a need for faculty with very solid theoretical foundations in mathematics with applied data analysis experience.

7. **Assess the faculty in terms of number and qualifications and plans for future staffing. Evaluate faculty responsibilities** for the proposed program, taking into account their other institutional and programmatic commitments. Evaluate faculty **activity in generating funds** for research, training, facilities, equipment, etc. Discuss any **critical gaps and plans for addressing them**.

I did not talk in detail about future staffing although it was made clear that there is a need for more faculty. There is also a clear need for faculty that have as part of their research program dealing with data in a broad sense—either developing methods for data analysis or analyzing methods for data analysis are sufficient. The biggest gap I would say is the lack of faculty that can interpolate between theory and data science. Also faculty that can interpolate between data science and theory will be better able to bring in funding which can help with the Data Science MS.

8. Evaluate credentials and involvement of **adjunct faculty and support personnel**.

I did not meet adjunct faculty nor did I meet support personnel for the program. We did discuss a bit about the program requiring some administrative staff once the program was in operation.

III. Students

9. Comment on the **student population the program seeks to serve**, and assess plans and projections for student recruitment and enrollment.

The program will serve two student populations: (1) students that come from a pure mathematics background and want to move towards real world applications using some ideas from pure mathematics and (2) students from an applied mathematics or theoretical engineering background that would like to blend pure and applied mathematics for applications. Initially recruiting 5-10 students and then increasing the numbers is reasonable. The projections provided seem adequate. There will need to be advertising, the University can help with this. Some videos of applications or data analysis examples on the departmental website would be very helpful. Also some of these videos should include women. In addition, a clear program that links the students with coop or internship connections will help the program immensely.

10. What are the prospects that recruitment efforts and admissions criteria will supply a **sufficient pool of highly qualified applicants and enrollees**?

The chances are reasonably high. There is a great demand for data science, there almost no data science masters programs in Mathematics departments, so there is little competition. With some visibility and a robust internship/industry interaction program the program can thrive. There is a need however to have more applied faculty involved. As things are currently there are 2 faculty that have extensive familiarity with data with a third faculty that has some experience.

11. Comment on provisions for encouraging participation of **persons from underrepresented groups**. Is there adequate attention to the needs of part-time, minority, or disadvantaged students?

We did discuss representation of women in mathematics. There were not specific plans about underrepresented groups, however I did discuss with the chair and some faculty about using the program to bring in underrepresented groups into the PhD program. It would be good for this aspect to be spelled out further by the department.

12. Assess the system for monitoring **students' progress and performance** and for **advising students** regarding academic and career matters.

There needs to be a more detailed assessment plan for the program. We did talk about exit interviews, we also talked about having a profession seminar, about bringing in industry partners. It would be good to have the discussions formalized into an assessment document. There are capstone classes for the different tracks which seem to be useful. Again one concern is the lack of data oriented faculty in the Mathematics department to advise the students.

13. Discuss prospects for graduates' post-completion success, whether **employment, job advancement, future study, or other outcomes related to the program's goals**.

If the students obtain some real data science and data analysis experience they will do well in terms of employment and positive job outcomes. The ability to do real data analysis with a solid mathematical foundation and some practical internship experience will provide a good basis to build a career on.

IV. Resources

14. Comment on the adequacy of physical **resources and facilities**, e.g., library, computer, and laboratory facilities; practica and internship sites or other experiential learning opportunities, such as co-ops or service learning; and support services for the program, including use of resources outside the institution.

There is a need for computing infrastructure. This does not mean computing labs. What is needed is working with the University IT to set up virtual machines that have the software and computing power for students to use for their capstone projects as well as for the data analysis and computing that permeates all of the classes and tracks. Digital lockers and virtual machines may be good options. It is very important that an adaptive and flexible infrastructure is set up with robust wireless in all the rooms in the mathematics department.

15. What is the **institution's commitment** to the program as demonstrated by the operating budget, faculty salaries, the number of faculty lines relative to student numbers and workload, and discussions about administrative support with faculty and administrators?

The administrators and Deans I met with were very supportive of the program and want to see it succeed. They also were willing to make an initial commitment to the program in terms of teaching, computing resources.

V. Summary Comments and Additional Observations

16. Summarize the **major strengths and weaknesses** of the program as proposed with particular attention to feasibility of implementation and appropriateness of objectives for the degree offered.

The major weaknesses are

- a) The lack of infrastructure for data analysis in terms of computing;
- b) The lack of faculty with familiarity with data analysis (2.5 out of the entire faculty);
- c) The lack of a clearly stated internship/professional outreach program. It is possible that the department with help from the administration can set this up.

The major strengths are

- a) A unique opportunity to build a Data Science program in Mathematics;

- b) A small but strong group of three faculty members with excellent theoretical ability as well as a clear connection and familiarity with data science and data analysis;
- c) A commitment from the administration to build up the program and provide resources such as linking the program with internship and coop possibilities.

17. If applicable, particularly for graduate programs, comment on the ways that this program will make a **unique contribution** to the field, and its likelihood of achieving State, regional and/or national **prominence**.

Having a Data Science Masters in a Mathematics Department and especially a department with a history of pure mathematics is unique. With the advent of some data analysis methods based on pure math including Topological Data Analysis and Algebraic Statistics if the department can link strongly with real applications this program can flourish.

18. Include any **further observations** important to the evaluation of this program proposal and provide any **recommendations** for the proposed program.



The State University
of New York

External Reviewer Conflict of Interest Statement

I am providing an external review of the application submitted to the State University of New York by:

t

(Name of Institution or Applicant)

The application is for (circle A or B below)

A) New Degree Authority

B) Registration of a new academic program by an existing institution of higher education:

Masters in Data Science administered by the mathematics department

(Title of Proposed Program)

I affirm that I: Sayan Mukherjee

1. am not a present or former employee, student, member of the governing board, owner or shareholder of, or consultant to the institution that is seeking approval for the proposed program or the entity seeking approval for new degree authority, and that I did not consult on, or help to develop, the application;
2. am not a spouse, parent, child, or sibling of any of the individuals listed above;

3. am not seeking or being sought for employment or other relationship with the institution/entity submitting the application?
4. do not have now, nor have had in the past, a relationship with the institution/entity submitting the application that might compromise my objectivity.

Name of External Reviewer (please print):

Sayan Mukherjee

Signature:

Department of Mathematics and Statistics' response to external evaluations for the Master of Science Program in Data Science

The Department is grateful to both distinguished evaluators Professor Carlsson and Professor Mukherjee for their very helpful comments. These comments are in line with our own vision of the development of the program and we find the suggestions made to be very specific and constructive.

A few remarks below provide an additional information on topics discussed by the evaluators.

1. **Professor Mukherjee writes:** "The plans and expectations for self-assessment and continuous improvement are not well developed. The department should work with the administration to develop such a plan. The chair of the mathematics department has ideas and the proposal has some plans but it needs to be further developed."

In this respect we would like to mention that the College of Arts and Sciences as well as the University in general has a rigorous and well-established procedure for self-assessment of courses and programs. We've been doing self-assessment of our courses for years and are going to follow these well-established and accepted guidelines. In addition for the first 3-5 years we will track the job placement of our graduates and keep in touch with them collecting their opinions on the material not included in the program that from their point of view based on their practical experience would be worth covering.

2. **Evaluating our faculty' potential Professor Mukherjee states:** "The mathematics faculty at Albany has a real strength in pure mathematics, especially topology and algebra. There are two faculty members that have experience with data analysis: Elizabeth Munch has a strong record in Topological Data Analysis (TDA) and Yeming Ying has a strong record in Machine Learning (ML). TDA and ML will form two of the three tracks in the MS program. The department needs more faculty with the research program of Elizabeth and Yeming for the MS in Data Science to flourish. Boris Goldfarb is also very involved in the MS program and he is an excellent blend of theory with an appreciation for applications. He has knowledge in computing which also helps. These are the three key faculty in making the MS program run. The program will require more faculty support to flourish. In keeping with the tradition of the department there is a need for faculty with very solid theoretical foundations in mathematics with applied data analysis experience."

We would like to emphasize that our department has strong internationally recognized groups in analysis, algebra and topology. As the other evaluator Professor G. Carlsson mentioned we currently have a big group of faculty capable of covering a substantial number of courses in all three clusters of the program including statistical inference, background courses in functional analysis and algebraic topology. Also, among the faculty whose current primary interests lie in "pure" mathematics some have previous research experiences in applied areas such as approximation theory, applied combinatorics, optimization, modelling, and numerical analysis. These faculty are also capable and willing to contribute to the program. Yet, indeed, we do need

more faculty specializing in machine learning, optimization and applied topology to successfully perform interdisciplinary research that includes data analysis. This year hiring partially filled this shortcoming: we hired one person in the area of machine learning and two in the area of applied topology. This is especially important for the MS program under review, since on top of capstone practicum courses and internship we would like to encourage our student to participate in research performed by the faculty, thus getting an additional exposure to real applications of the methods of data analysis they learn in the classroom.

3 Commenting on the participation of people from underrepresented groups Professor Mukherjee remarks: "We did discuss representation of women in mathematics. There were not specific plans about underrepresented groups, however I did discuss with the chair and some faculty about using the program to bring in underrepresented groups into the PhD program. It would be good for this aspect to be spelled out further by the department."

We would like to mention that bringing in women, people of color and other underrepresented groups is always on the top of the departmental agenda. This is especially important in the area of mathematics where the representation of these groups is traditionally much below the average. Our commitment to diversity and inclusion in both graduate student population and faculty profile is universal and goes way beyond this particular program.

4. Discussing the composition of the program Professor Carlsson states: "One comment I would add is that I think it is imperative that every student who goes through the program achieve some training in the "nuts and bolts" of preparing data for analysis. This aspect of the subject is referred to as "extract, transform, and load" (ETL) in industry. This means exposure to the languages Python and R, to methodologies such as Microsoft's pivot tables, and to ways of dealing with some common unstructured data types, such as plain text, images, databases of molecules, etc. A student who does not have this experience will not be able to land in a position and become productive immediately. As the program develops I would be happy to answer questions about the desirable scope of this kind of experience. "

We completely agree with this assessment and take this point very seriously. The exposure to Python and R is incorporated in our courses. Based on our first experience we might develop a short (a quarter length) course in these languages and require students to take this course (possibly in the summer).

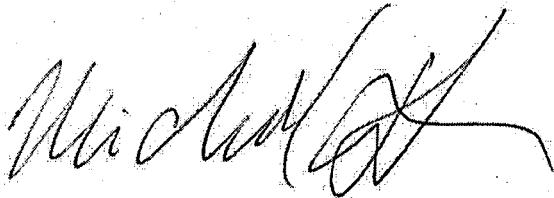
5. Both evaluators mentioned importance of an established partnership with potential employers in government and industry. In particular, they mention the internship program for the students enrolled in the program.

We fully agree with these comments and think that the issue of internship, or even broader, of partnership is extremely important. We have already started the process. A number of our alumni work for the New York State government, and deal with big data on a day-to-day business. They are very enthusiastic about this new program and express their willingness to help with creating internship slots to train our students. We also think about a possibility of their participation in the program as instructors for practicums, thus bringing a real life experience into a classroom. We also currently talking to the staff of the Machine Learning Laboratory in GE regarding a possibility of internship for our students. We are also in contact with City-Internship,

a New York City company that specializes on arranging internship for students. We will continue to establish connections with other possible partners as the program develops.

6. Professor Mukherjee writes in his comments "There is a need for computing infrastructure. This does not mean computing labs. What is needed is working with the University IT to set up virtual machines that have the software and computing power for students to use for their capstone projects as well as for the data analysis and computing that permeates all of the classes and tracks. Digital lockers and virtual machines may be good options. It is very important that an adaptive and flexible infrastructure is set up with robust wireless in all the rooms in the mathematics department.

We agree that the issue of computational infrastructure should be addressed since computational methods are incorporated in practically all courses listed in the program. We think that he proposed an effective way of addressing the problem and intend to contact IT to discuss the details of the implementation of such program.

A handwritten signature in black ink, appearing to read "Michael Stessin". The signature is fluid and cursive, with a long horizontal stroke at the end.

Michael Stessin, Chair