Term Information

Effective Term	Autumn 2021

General Information

Course Bulletin Listing/Subject Area	Physics
Fiscal Unit/Academic Org	Physics - D0684
College/Academic Group	Arts and Sciences
Level/Career	Graduate, Undergraduate
Course Number/Catalog	5680
Course Title	Big Data Analytics in Physics
Transcript Abbreviation	BigDataAnalytics
Course Description	Provides an introduction to machine learning and advanced algorithms, with an emphasis on practical physics-based applications, using publicly available data sets. The goal is to provide an introduction to Data Science for students who may want to pursue this as a career option and/or apply these techniques in a research environment.
Semester Credit Hours/Units	Fixed: 3

Offering Information

Length Of Course	14 Week
Flexibly Scheduled Course	Never
Does any section of this course have a distance education component?	No
Grading Basis	Letter Grade
Repeatable	No
Course Components	Lecture
Grade Roster Component	Lecture
Credit Available by Exam	No
Admission Condition Course	No
Off Campus	Never
Campus of Offering	Columbus

Prerequisites and Exclusions

Prerequisites/Corequisites	Enrollment in the Physics or Engineering Physics major; C- or higher in CSE 1222, CSE 1223, Engineering 1281H, or Astronomy 1221; C+ or higher in Physics 1251. Or instructor permission.
Exclusions	
Electronically Enforced	Yes

Cross-Listings

Cross-Listings	

Subject/CIP Code

Subject/CIP Code	40.0801
Subsidy Level	Doctoral Course
Intended Rank	Junior, Senior, Masters, Doctoral

Requirement/Elective Designation

The course is an elective (for this or other units) or is a service course for other units

Course Details

Course goals or learning objectives/outcomes

- Understand how to process, clean, and prepare data for further analysis.
- Understand how to visualize data in order to gain insights regarding feature importance.
- Understand and be capable to apply a wide variety of machine learning tools in regression, classification, and clustering, with an emphasis on applications in Physics.
- Design their own machine learning solution to a Physics problem from start to finish.

Content Topic List

- Introduction to Python, manipulating data files, linear regression, classification using support vector machines, the confusion matrix, multi-class classification, decision trees and random forests, logistic regression, using the Ohio SuperComputer;
- Neural networks, multi-layer perceptrons, Siamese Networks, and the iPhone Face Recognition Algorithm.

Sought Concurrence

Yes

Attachments

- Concurrence_Form_Final.pdf
  (Concurrence. Owner: Thaler,Lindsey Nicole)
- Sample_Plan.pdf: Sample 4-year plan
  (Other Supporting Documentation. Owner: Thaler,Lindsey Nicole)
- Curriculum_Map.pdf: Curriculum Map
  (Other Supporting Documentation. Owner: Thaler,Lindsey Nicole)
- BigDataNewSyllabus_New.pdf
  (Syllabus. Owner: Thaler,Lindsey Nicole)
- projectInfo.pdf: Final project information. This is from AU20.
  (Other Supporting Documentation. Owner: Thaler,Lindsey Nicole)

Comments

- Please see Panel feedback email sent 04/06/21. *(by Hilty,Michael on 04/06/2021 09:29 AM)*
- Please upload the updated curriculum map. *(by Humanic,Thomas John on 02/23/2021 09:34 AM)*
- Please upload your updated curriculum map for the BS showing how the new course fulfills what program goal and at what level. *(by Vankeerbergen,Bernadette Chantal on 02/17/2021 05:21 PM)*

Workflow Information

Status	User(s)	Date/Time	Step
Submitted	Thaler,Lindsey Nicole	02/16/2021 09:47 AM	Submitted for Approval
Approved	Humanic,Thomas John	02/16/2021 11:59 AM	Unit Approval
Revision Requested	Vankeerbergen,Bernadette Chantal	02/17/2021 05:21 PM	College Approval
Submitted	Humanic,Thomas John	02/17/2021 05:48 PM	Submitted for Approval
Approved	Humanic,Thomas John	02/17/2021 06:00 PM	Unit Approval
Revision Requested	Vankeerbergen,Bernadette Chantal	02/17/2021 09:21 PM	College Approval
Submitted	Humanic,Thomas John	02/18/2021 06:00 AM	Submitted for Approval
Approved	Humanic,Thomas John	02/18/2021 06:02 AM	Unit Approval
Revision Requested	Vankeerbergen,Bernadette Chantal	02/18/2021 10:32 AM	College Approval
Submitted	Humanic,Thomas John	02/18/2021 12:00 PM	Submitted for Approval
Revision Requested	Humanic,Thomas John	02/23/2021 09:34 AM	Unit Approval
Submitted	Thaler,Lindsey Nicole	02/23/2021 09:36 AM	Submitted for Approval
Approved	Humanic,Thomas John	02/23/2021 09:38 AM	Unit Approval
Approved	Vankeerbergen,Bernadette Chantal	02/23/2021 10:40 AM	College Approval
Revision Requested	Hilty,Michael	03/10/2021 03:10 PM	ASCCAO Approval
Submitted	Thaler,Lindsey Nicole	03/12/2021 09:27 AM	Submitted for Approval
Approved	Humanic,Thomas John	03/12/2021 10:49 AM	Unit Approval
Approved	Vankeerbergen,Bernadette Chantal	04/01/2021 05:33 PM	College Approval
Revision Requested	Hilty,Michael	04/06/2021 09:29 AM	ASCCAO Approval
Submitted	Thaler,Lindsey Nicole	04/28/2021 09:16 AM	Submitted for Approval
Approved	Humanic,Thomas John	04/28/2021 09:21 AM	Unit Approval
Approved	Vankeerbergen,Bernadette Chantal	04/28/2021 12:53 PM	College Approval
Pending Approval	Jenkins,Mary Ellen Bigler Hanlin,Deborah Kay Oldroyd,Shelby Quinn Hilty,Michael Vankeerbergen,Bernadette Chantal	04/28/2021 12:53 PM	ASCCAO Approval

Physics 5680
Big Data Analytics in Physics

Course Information

- **Course times:** Lectures: Tuesdays and Thursdays from 12:40p.m.-2:45 p.m.
- **Credit hours:** 3

Instructor

- **Name:** Richard E. Hughes
- **Email:** hughes.319.@osu.edu
- **Office location:** 3140 PRB
- **Office hours:** By Appointment
- **Preferred means of communication:**
  - My preferred method of communication for questions is **email.**
  - My class-wide communications will be sent through the Announcements tool in CarmenCanvas. Please check your notification preferences (go.osu.edu/canvas-notifications) to be sure you receive these messages.

Course Prerequisites

Enrollment in the Physics or Engineering Physics major; C- or higher in CSE 1222, CSE 1223, Engineering 1281H, or Astronomy 1221; C+ or higher in Physics 1251. Or instructor permission.

Course Description

This course provides an introduction to machine learning and advanced algorithms, with an emphasis on practical physics-based applications, using publicly available data sets. The goal is to provide an introduction to Data Science for students who may want to pursue this as a career option and/or apply these techniques in a research environment.

Learning Outcomes

By the end of this course, students should successfully be able to do the following:

- Understand how to process, clean, and prepare data for further analysis.
- Understand how to visualize data in order to gain insights regarding feature importance.
- Understand and be capable to apply a wide variety of machine learning tools in regression, classification, and clustering, with an emphasis on applications in Physics.
- Design their own machine learning solution to a Physics problem from start to finish.

Expectations

Students are expected to come into the course with the following:

- A laptop/chromebook or similar. Mac/Linux/Windows/ChromeOS are all fine, since we will be using a browser-based environment (Jupiter notebooks) for all of our programming.
- Basic programming skills, which could be in any of a number of different languages, such as C++, java, python, etc.
- All course assignments will be done in python. Entering into the course after doing a simple online python tutorial would be a good idea!
- Basic knowledge of statistics and probability (such as would be obtained in Physics 3700 or a statistics course)
- Enthusiasm for learning and a desire to challenge oneself!

Course Materials, Fees and Technologies

Required Materials and/or Technologies

- Text: Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow, 2nd Edition, Aurelien Geron ; ISBN: 978-1-492-03264-9

Required Equipment

- **Computer:** A laptop/chromebook/PC (or similar) which provides internet access via a browser. Significant editing will be done in the browser, so a device with a large enough screen and a keyboard are strongly recommended. The operating system for this device is not as important since the coursework will all be accomplished using a browser-based environment (Jupiter notebooks and similar) for all of our programming.
- **Other:** a mobile device (smartphone or tablet) to use for BuckeyePass authentication

CarmenCanvas Access

You will need to use BuckeyePass (buckeyepass.osu.edu) multi-factor authentication to access your courses in Carmen. To ensure that you are able to connect to Carmen at all times, it is recommended that you do each of the following:

- Register multiple devices in case something happens to your primary device. Visit the BuckeyePass - Adding a Device (go.osu.edu/add-device) help article for step-by-step instructions.
- Request passcodes to keep as a backup authentication option. When you see the Duo login screen on your computer, click **Enter a Passcode** and then click the **Text me new codes** button that appears. This will text you ten passcodes good for 365 days that can each be used once.
- Install the Duo Mobile application (go.osu.edu/install-duo) on all of your registered devices for the ability to generate one-time codes in the event that you lose cell, data, or Wi-Fi service.

If none of these options will meet the needs of your situation, you can contact the IT Service Desk at 614-688-4357 (HELP) and IT support staff will work out a solution with you.

Technology Skills Needed for This Course

- Basic computer and web-browsing skills
- Navigating CarmenCanvas (go.osu.edu/canvasstudent)

Other Skills Needed for This Course

- Basic programming skills, which could be in any of a number of different languages, such as C++, java, python, etc.   All course assignments will be done in python. Entering into the course after doing a simple online python tutorial would be a good idea!

Technology Support

For help with your password, university email, CarmenCanvas, or any other technology issues, questions or requests, contact the IT Service Desk, which offers 24-hour support, seven days a week.

- **Self Service and Chat**: go.osu.edu/it
- **Phone:** 614-688-4357 (HELP)
- **Email:** servicedesk@osu.edu

Digital Flagship

Digital Flagship is a student success initiative aimed at helping you build digital skills for both college and career. This includes offering an engaging collection of digital tools and supportive learning experiences, university-wide opportunities to learn to code, and a Design Lab to explore digital design and app development. Digital Flagship resources available to help Ohio State students include on-demand tutorials, The Digital Flagship Handbook (your guide for all things tech-related), workshops and events, one-on-one tech consultations with a peer or Digital Flagship staff member, and more. To learn more about how Digital Flagship can help you use technology in your courses and grow your digital skills, visit go.osu.edu/dfresources.

Grading and Faculty Response

How Your Grade is Calculated

Assignment Category	Points
Class assignments	60%
Reading Quizzes	11%
Final Project	29%

See Course Schedule for due dates.

Descriptions of Major Course Assignments

- **Weekly assignments**
  There will be weekly assignments to allow you to practice machine learning concepts from reading and in class jupyter notebooks.

- **Reading Quizzes**
  There will be short online quizzes that will focus on readings and class assignments. If you have done the readings. And the assignments these should be extremely straightforward.

- **Final Project**
  1. Description: The project should be one in which machine learning/advanced algorithms can be brought to bear on an interesting problem, either to "solve" that problem, or at least deliver meaningful insights regarding the problem.
  2. The projects will all be single person projects.
  3. The projects should be science oriented.
  4. The deliverables do not include a final report but instead a working jupyter notebook as well as a single page “Poster” (in ppt/pdf form only).

Academic integrity and collaboration: Your submitted assignments should be your own original work. We do encourage students to help each other understand the material. We also anticipate and encourage students to use online resources such as stackoverflow.com to help them with implementation of parts of various algorithms for their assignments. However, the bulk of each assignment should be unambiguously each student's own work.

In the final project, you should follow [MLA/APA/Chicago etc.] style to cite the ideas and words of your research sources. You are encouraged to ask a trusted person to proofread your assignments before you turn them in but no one else should revise or rewrite your work.

Late Assignments

Late assignments will be accepted for reduced credit up to 1 week late, with gradually reduced credit. Quizzes will be available online for a specified window only - there are no makeups. Two quizzes may be dropped

Instructor Feedback and Response Time

I am providing the following list to give you an idea of my intended availability throughout the course. Remember that you can call 614-688-4357 (HELP) at any time if you have a technical problem.

- **Preferred contact method:** If you have a question, please contact me first through my Ohio State email address. I will reply to emails within **24 hours on days when class is in session at the university**.
- **Class announcements:** I will send all important class-wide messages through the Announcements tool in CarmenCanvas. Please check your notification preferences (go.osu.edu/canvas-notifications) to ensure you receive these messages.
- **Grading and feedback:** For large weekly assignments, you can generally expect feedback within **seven days**.

Grading Scale

93–100: A
90–92.9: A-
87–89.9: B+
83–86.9: B
80–82.9: B-
77–79.9: C+
73–76.9: C
70–72.9: C-
67–69.9: D+
60–66.9: D
Below 60: E

Course Schedule

The following is a preliminary schedule.   If adjustments are needed during the semester, as revised schedule will be posted to the Carmen page and a notice will be made using the Announcements tool in CarmenCanvas. Refer to the CarmenCanvas course for up-to-date due dates.

Week	Chapter/pages	Topic
1	Notes	Python Intro
		Manipulating data files;
		Visualization with matplotlib and plotly
2	Chapter 2, 5	Intro to linear regression; Dealing with missing data;
		test/train splits; feature scaling and categorical data
3	Chapter 3	Intro to Classification using support vector machines;
		The confusion matrix;
		ROC curves and AUC
4		Multi-class classification;
		k-fold validation
5	Chapter 6, 7	Decision Trees and Random Forests;
		Over- and Under-fitting, and the Bias-Variance Tradeoff
		Feature Importance
6	From Notes	Linear Regression and Gradient Descent; writing your own regressor from scratch
		Logistic Regression (write your own from scratch)
7	From Notes	Using the Ohio SuperComputer Center (OSC) batch system;
		Softmax Regression
		Neural Networks from scratch
8	Chapter 10	Introduction to Keras: The Industry Standard Neural Network Library
		Multi-Layer Perceptrons
9	Chapter 14	Convolutional Neural Networks
10	Chapter 17	Project Proposal Due
		Autoencoders; Stacked Autoencoders and Classification
11	Chapter 17	Visualization of learned Features in Neural Networks Adversarial Examples
		Variational Autoencoders
12	Notes, Chapter 16,17	1D Convolutional Neural Networks and Sequences Text Classification
		Project Progess Report Due
13	Notes	Siamese Networks and the iPhone Face Recognition
		Algorithm
14	Chapter 17	Possible Additional topics: Generative Adversarial Networks,
	Chapter 15	Recurrent Neural Networks
		During Exam Week: Project due (no final exam)

Other Course Policies

Discussion and Communication Guidelines

The following are my expectations for how we should communicate as a class. Above all, please remember to be respectful and thoughtful.

- **Writing style**: While there is no need to participate in class discussions as if you were writing a research paper, you should remember to write using good grammar, spelling, and punctuation. A more conversational tone is fine for non-academic topics.
- **Tone and civility**: Let's maintain a supportive learning community where everyone feels safe and where people can disagree amicably. Remember that sarcasm doesn't always come across online.
- **Citing your sources**: When we have academic discussions, please cite your sources to back up what you say. For the textbook or other course materials, list at least the title and page numbers. For online sources, include a link.
- **Backing up your work**: Consider composing your academic posts in a word processor, where you can save your work, and then copying into the Carmen discussion.

Academic Integrity Policy

See Descriptions of Major Course Assignments for specific guidelines about collaboration and academic integrity in the context of this online class.

Ohio State's Academic Integrity Policy

Academic integrity is essential to maintaining an environment that fosters excellence in teaching, research, and other educational and scholarly activities. Thus, The Ohio State University and the Committee on Academic Misconduct (COAM) expect that all students have read and understand the university's Code of Student Conduct (studentconduct.osu.edu), and that all students will complete all academic and scholarly assignments with fairness and honesty. Students must recognize that failure to follow the rules and guidelines established in the university's *Code of Student Conduct* and this syllabus may constitute "Academic Misconduct."

The Ohio State University's *Code of Student Conduct* (Section 3335-23-04) defines academic misconduct as: "Any activity that tends to compromise the academic integrity of the university or subvert the educational process." Examples of academic misconduct include (but are not limited to) plagiarism, collusion (unauthorized collaboration), copying the work of another student, and possession of unauthorized materials during an examination. Ignorance of the university's *Code of Student Conduct* is never considered an excuse for academic misconduct, so I recommend that you review the *Code of Student Conduct* and, specifically, the sections dealing with academic misconduct.

If I suspect that a student has committed academic misconduct in this course, I am obligated by university rules to report my suspicions to the Committee on Academic Misconduct. If COAM determines that you have violated the university's Code of Student Conduct (i.e., committed academic misconduct), the sanctions for the misconduct could include a failing grade in this course and suspension or dismissal from the university.
If you have any questions about the above policy or what constitutes academic misconduct in this course, please contact me.

Other sources of information on academic misconduct (integrity) to which you can refer include:

- Committee on Academic Misconduct (go.osu.edu/coam)
- Ten Suggestions for Preserving Academic Integrity (go.osu.edu/ten-suggestions)
- Eight Cardinal Rules of Academic Integrity (go.osu.edu/cardinal-rules)

Copyright for Instructional Materials

The materials used in connection with this course may be subject to copyright protection and are only for the use of students officially enrolled in the course for the educational purposes associated with the course. Copyright law must be considered before copying, retaining, or disseminating materials outside of the course.

Statement on Title IX

All students and employees at Ohio State have the right to work and learn in an environment free from harassment and discrimination based on sex or gender, and the university can arrange interim measures, provide support resources, and explain investigation options, including referral to confidential resources.

If you or someone you know has been harassed or discriminated against based on your sex or gender, including sexual harassment, sexual assault, relationship violence, stalking, or sexual exploitation, you may find information about your rights and options on Ohio State's Title IX website (titleix.osu.edu) or by contacting the Ohio State Title IX Coordinator at titleix@osu.edu. Title IX is part of the Office of Institutional Equity (OIE) at Ohio State, which responds to all bias-motivated incidents of harassment and discrimination, such as race, religion, national origin and disability. For more information, visit the OIE website (equity.osu.edu) or email equity@osu.edu.

Commitment to a Diverse and Inclusive Learning Environment

The Ohio State University affirms the importance and value of diversity in the student body. Our programs and curricula reflect our multicultural society and global economy and seek to provide opportunities for students to learn more about persons who are different from them. We are committed to maintaining a community that recognizes and values the inherent worth and dignity of every person; fosters sensitivity, understanding, and mutual respect among each member of our community; and encourages each individual to strive to reach his or her own potential. Discrimination against any individual based upon protected status, which is defined as age, color, disability, gender identity or expression, national origin, race, religion, sex, sexual orientation, or veteran status, is prohibited.

Your Mental Health

As a student you may experience a range of issues that can cause barriers to learning, such as strained relationships, increased anxiety, alcohol/drug problems, feeling down, difficulty concentrating and/or lack of motivation. These mental health concerns or stressful events may lead to diminished academic performance or reduce a student's ability to participate in daily activities. No matter where you are engaged in distance learning, The Ohio State University's Student Life Counseling and Consultation Service (CCS) is here to support you. If you find yourself feeling isolated, anxious or overwhelmed, on-demand mental health resources (go.osu.edu/ccsondemand) are available. You can reach an on-call counselor when CCS is closed at 614- 292-5766**. 24-hour emergency help** is available through the National Suicide Prevention Lifeline website (suicidepreventionlifeline.org) or by calling 1-800-273-8255(TALK). The Ohio State Wellness app (go.osu.edu/wellnessapp) is also a great resource.

Accessibility Accommodations for Students with Disabilities

Requesting Accommodations

The university strives to make all learning experiences as accessible as possible. If you anticipate or experience academic barriers based on your disability including mental health, chronic or temporary medical conditions, please let me know immediately so that we can privately discuss options. To establish reasonable accommodations, I may request that you register with Student Life Disability Services (SLDS). After registration, make arrangements with me as soon as possible to discuss your accommodations so that they may be implemented in a timely fashion.

Disability Services Contact Information

- Phone: 614-292-3307
- Website: slds.osu.edu
- Email: slds@osu.edu
- In person: Baker Hall 098, 113 W. 12th Avenue

Accessibility of Course Technology

This online course requires use of CarmenCanvas (Ohio State's learning management system) and other online communication and multimedia tools. If you need additional services to use these technologies, please request accommodations with your instructor.

- CarmenCanvas accessibility (go.osu.edu/canvas-accessibility)
- Streaming audio and video
- CarmenZoom accessibility (go.osu.edu/zoom-accessibility)
- Collaborative course tools
- As you begin your college career, which might eventually involve asking professors to recommend you for graduate programs, jobs, or internships, please be aware that professors are not obligated to write references for any student who asks us. I don’t write a reference for a student unless I can write a very positive and specific one. Therefore, your job as a college student is to become the kind of student professors can rave about in recommendations — hardworking, collegial, and intellectually inquisitive and honest. Consider maintaining relationships over time with professors, so that they know you well enough to write for you. Many juniors and seniors tell me they wish they had thought about this during their first year.

Policy on Recommendation Letters

As you begin your college career, which might eventually involve asking professors to recommend you for graduate programs, jobs, or internships, please be aware that professors are not obligated to write references for any student who asks us. I don’t write a reference for a student unless I can write a very positive and specific one. Therefore, your job as a college student is to become the kind of student professors can rave about in recommendations — hardworking, collegial, and intellectually inquisitive and honest. Consider maintaining relationships over time with professors, so that they know you well enough to write for you. Many juniors and seniors tell me they wish they had thought about this during their first year.

Physics Major (Non-Honors Advanced Physics Option)

Year	Autumn Semester	Credit hours	Comment		Spring Semester	Credit Hours	Comment
1	Physics 1250	5	Intro Physics I		Physics 1251	5	Intro Physics I
	Math 1151	5	Calculus I		Math 1152	5	Calculus II
	ASC 1100	1	Survey		CSE 1222°	3	C++ Programming
	Foreign Lang. 1	4			Foreign Lang. 2	4	
	Total Hours	**15**			**Total Hours**	**17**	
2	Physics 2300	4	Mechanics I		Physics 2301	4	Mechanics II
	Physics 2095	1	Seminar		Physics 3700	3	Data Ana. Lab
	Math 2153	4	Calculus III		Math 2415†	3	Diff. Equations
	Foreign Lang. 3	4			Gen Ed	3	
	Gen Ed	3			Gen Ed	3	
	Total Hours	**16**			**Total Hours**	**16**	
3	Physics 5500	4	Quantum I		Physics 5501	4	Quantum II
	Physics 5680**	3	Big Data Analytics		Physics 5400	4	E&M
	Gen Ed	3			Gen Ed	3	
	Gen Ed	3			Gen Ed	3	
	Free Elective◊	3					
	Total Hours	**16**			**Total Hours**	**14**	
4	Physics 5600	4	Stat. Mech.		Physics 5300	4	Theoretical Mech.
	Physics 5700	3	Senior Lab		Physics Elective*	4	
	Gen Ed	3			Gen Ed	3	
	Free Elective◊	3			Free Elective◊	3	
	Total Hours	**13**			**Total Hours**	**14**	

Courses in yellow are only offered in the term shown

Enrollment information can be found at physics.osu.edu/controlled-access-courses

† Math 5520H can be taken in place of Math 2415 and 2568.

** or Physics 6810 (Computational Physics) or Physics 3201H (Holography) or Physics 4700 (Electonics Lab)

* Acceptable Physics Elective include Physics 3470 (optics) or any of the Physics 68xx courses

° or CSE 1223 or Astronomy 1221

◊ Free electives are only required if a student needs to take extra courses in order to reach the minimum 121 credit hour requirement set by the College of Arts and Sciences.

The Ohio State University
College of the Arts and Sciences Concurrence Form

The purpose of this form is to provide a simple system of obtaining departmental reactions to course requests. **An e-mail may be substituted for this form.**

An academic unit initiating a request should complete Section A of this form and send a copy of the form, course request, and syllabus to each of the academic units that might have related interests in the course. Units should be allowed two weeks to respond to requests for concurrence.

Academic units receiving this form should respond to Section B and return the form to the initiating unit. Overlap of course content and other problems should be resolved by the academic units before this form and all other accompanying documentation may be forwarded to the Office of Academic Affairs.

A. Proposal to review

Physics 5680: Big Data Analytics in Physics

Initiating Academic Unit Course Number Course Title

New 2/13/2020

Type of Proposal (New, Change, Withdrawal, or other) Date request sent

Computer Science & Engineering 2/28/2020

Academic Unit Asked to Review Date response needed

B. Response from the Academic Unit reviewing

Response: include a reaction to the proposal, including a statement of support or non-support (continued on the back of this form or a separate sheet, if necessary).

On behalf of the CSE Curriculum Committee I am happy to offer concurrence for Physics 5680, "Big Data Analytics in Physics."

Signatures

Name	Position	Unit	Date
1. Teodorescu	Chair, Curriculum Comm	CSE	02/17/20
2.			
3.			

Revised 5/27/14

Big Data Final Project

FALL, 2020

R. HUGHES

Course Final Project

My hope for this course is that it will give you practical data analytics tools to use, if and when you need them, whether in a future research project or as a data scientist in industry. The Final Project is a chance for you to demonstrate that you have learned enough to handle a machine learning task - on your own!

Goals:

1. The project should be one in which machine learning/advanced algorithms can be brought to bear on an interesting problem, either to "solve" that problem, or at least deliver meaningful insights regarding the problem.
2. The projects will all be single person projects.
3. The projects should be science oriented.
4. The deliverables do not include a final report but instead a single page “Poster” (in ppt/pdf form only) and accompanying 3-minute video.

Project Choices:

You have the choice of attempting a predefined project or coming up with your own. In either case, **it will be necessary that the project has some reasonable connection to science.** I am comfortable with the possibility that this connection may not be obvious, but if so you will need to make the case in your project proposal that such a connection with science exists.

1. **Already existing project:** For these you can use the resources of the Kaggle website (or other similar online resources). Kaggle hosts open machine learning competitions for both commercial and non-profit organizations. They provide data, background information, an evaluation metric, and a forum for each competition. There are (as of Oct 20, 2020) 20 open competitions and hundreds completed competitions. You can choose an open OR an already completed competition. I anticipate that you may need to substantially modify the goals as stated on the Kaggle website in order to be able to complete your project in the allotted time.

2. **You choose the project:** Here you are free to choose something from current research that you are doing, or perhaps something you have been interested in learning about. You are also allowed to combine this project with a project from another class, though you must fulfill all requirements of this project. More on this below in the FAQ.

Machine Learning Project Checklist

1. Define the problem: what exactly are you trying to do?
2. Get the data. (Sometimes this means creating your data!)
3. Explore the data with an eye to learning how you might use the data to solve your problem.
4. Prepare the data: this may include normalizing, augmenting, modifying (cropping, centering, etc) the data.
5. Investigate a number of different models and choose 2 or 3 that seem particularly good.
6. Fine tune the models and either choose the best or combine them.
7. Present your solution to the problem using your final model.
8. Launch your system with your final model and monitor results.

Project Timeline

1. **Proposal:** This describes the project that you will do, as well as outlining your proposed approach. This approach may change as you learn more! Submitted via Carmen:
   - **Due date: 5pm Friday October 30, 2020**
2. **Mid-term "Report":** This is a checklist of things you have managed to complete, as well as a preliminary version of your poster (see below). The primary purpose of this is to make sure you are on track to complete the project. Submitted via Carmen:
   - **Due Date: 5pm Thursday November 12, 2020**
3. **Completed project:** A completed project will be comprised of a final poster in pdf/ppt form (submitted via Carmen) as well as an accompanying video (submission via carmen or youtube).
   - **Due Date: 5pm Thursday December 10, 2020**

Project Proposal

This describes the project you will do, as well as outlining your proposed approach. This approach may change as you learn more! Sources for possible projects are listed.

* Due date: **5pm Friday October 30, 2020**

* Deliverables: Completed proposal form; see the file "Project Proposal Template.doc" in the module finalProject/docs folder. Submitted via Carmen.

* Grading: This is a completion based grade, worth **2 out of 58 (3.4%) of the project points**, if all parts are done and it is turned in on time. You will lose 25% (of the 10%) for each day it is turned in late. I encourage you to share with me your proposed project ahead of the due date if you have any concerns that it is not appropriate.

Project Proposal

[Name]
[Date]

Project Proposal: [Name of Project]

Project Source

[Kaggle/UCI/Other Online/Research]

Background

[Why is the project being undertaken? Describe an opportunity or problem that the project is to address.]

Objectives/Goals

- [specific & measurable objective 1]
- [specific & measurable objective 2]
- [specific & measurable objective 3]

Proposed Timeframe

Description of Work	Start Date	End Date
Research on problem/previous approaches	5PM, October 30, 2020	
Obtain Data	…	
Visualize/Explore data		
Mid-project Checklist Due	5PM, November 12, 2020	
Begin Model Testing		
Fine-tune Model		
Prepare Poster		
Project Due	5PM, December 10, 2020	

Signatures

[Student Name] [Signature]

Mid-term "Report"

This is checklist of things you have managed to complete, as well as a preliminary version of your poster (see below). The primary purpose of this is to make sure you are on track to complete the project.

Due Date: **5pm Thursday November 12, 2020**

Deliverables: Completed checklist (see finalProject/docs folder); associated items highlighted in checklist; current version of your poster (see finalProject/poster folder). It is ok if most of your poster is placeholders at this point.

Grading: This is a completion based grade, worth **4 out of 58 (6.8%) of the project points**. Completion means:

- the entire checklist is filled out
- At least 3 of the item in the checklist are marked as completed
- the associated jupyter notebook(s) and/or python script(s) and checklist are submitted via Carmen
- You will lose 25% (of the 10%) for each day it is turned in late.
- Contact me with questions!

Mid-term "Report"

Mid-Project Report and Checklist

Name:	
Date:	
Project Title:	

Yes	**No**	**Date Complete**	**Task**
			Data has been transferred to the appropriate computing platform (OSC/research group/personal laptop)
			If not complete explain:
			Background research has been completed (and section for poster summerizing this is done).
			If not complete explain:
			Preliminary visualization of data have been performed. Jupyter notebook available in project folder
			If not complete explain:
			Timeline for attaining remaining project goals has been sketched out. Text doc available in project folder
			If not complete explain:
			Project poster has been started. Version available in project folder.
			If not complete explain:
			Initial/test model has been chosen and initial pass with this model has been attempted.
			If not complete explain:

Completed project: Due Date: 5pm Thursday December 10, 2019

A completed project will be comprised of the following, all submitted via carmen:

1. A final poster (in ppt *and* pdf form only - there is no formal poster session). The final poster should be in a form such that you could give a copy to an interested party (future employer or research advisor) to give them insight into the kind of work you are capable of.
2. A 3-minute video (see next page).
3. Grading: See rubric.

Project Video

NOTE: ”Stolen” from Berkeley’s CS289A Course!

1. The video should be clear and understandable, describing everything you think is important about your project (motivation, description, techniques, results, etc.).
2. The video needs to be self-contained: any 6820 student should be able to understand what you did (at least at a high level) without consulting any other materials.
3. You can make the video as simple as slides with a voice overlay, or as fancy as you want.
4. As long as it is clear and understandable, you will not be graded on the fanciness of the video. Content is what will matter. (Fanciness might be fun, though.). Imagine talking about your project to friends over lunch.
5. You must upload the video on YouTube and provide us with the link. You may choose to keep the video private (i.e., only those with the link can view it), in which case only the instructors will view it. You can make the video public if you want to.
6. **Important:** The video can be at most 3 minutes long. This is a very strict requirement; a video of length 3 minutes and 1 second does not count. The length is counted as whatever YouTube says it is. In case you are worried about how you will fit an entire class project into three minutes, take a look at these videos, which fit an entire Ph.D. thesis into three minutes.

Project Rubric

Physics 6820: Big Data Analytics in Physics: Final Project Rubric

Project Points Breakdown

Item	Points	Explanation
Project Proposal	/2	
Project Mid-term Report	/4	
Is the poster a good mix of text and visuals	/5	
Does poster explain problem and importance	/8	
Does poster explain algorithm used and why	/8	
Does poster explain results	/8	
Does poster indicate future improvements	/8	
If video <= 3 minutes:	If >3 minutes next 3 lines ZERO	
Video: description of problem	/5	
Video: description of approach	/5	
Video: description of results and future	/5	
Late Penalty	10% per day	
TOTAL	/58	

Resources

Here are some online sources:

1. The Kaggle competition [website](https://www.kaggle.com/competitions).
2. The UCI machine learning repository [website](http://archive.ics.uci.edu/ml/index.php)
3. The Berkeley CS229 project website: https://people.eecs.berkeley.edu/~jrs/189s19/project.html
4. The Stanford CS229 project [website](http://cs229.stanford.edu/projects.html). The section titled "Previous projects" has hundreds of example projects from past years. Note that posters only appear to be available for years 2015-present, and reports (something I am not asking you to do) are available for all years. Many (though not all) of these projects are also based on Kaggle competitions. Some particularly good examples of posters are:
   - Large-scale Protein Atlas Compartmentalization Analysis]
   - A Data-Driven Approach for Predicting Elastic Properties of Inorganic Materials]
   - Galaxy Morphology
   - Modeling and Optimization of Optical Devices using a Variational Autoencoder

FAQ(1)

1. I found on the web a complete working solution to my project problem. Can I use it?

No. As mentioned above, each Kaggle competition has a forum of questions and answers devoted to that particular competition. We encourage you to consult that forum! However, as much as is practical, we strongly urge you to **not** consult any posted solutions. This will defeat the whole purpose of the project: The value the project has as a learning experience is **much** more important than its contribution to your final grade in this course!

2. Can I do this project as part of a team with another student?

No. However, we do anticipate that many students in the class may end up working on the same subset of projects. Consulting with these students about problems you run into and methods you are trying is fine. But the work you submit as your project must be **overwhelmingly** your own. You should be able to explain and justify every choice that is made in the course of completing your project.

3. Is it okay to use a dataset that is not public ?

Yes. We don't mind you using a dataset that is not public, as long as you have the required permissions to use it. We don't require you to share the dataset either as long as you can accurately describe it in the Final Report.

FAQ(2)

4. Are the final presentations public?

Yes. All projects will be viewable by fellow students via our classroom carmen. We will also host final posters (just like the CS229 website above). Let us know if this is not possible (because the data is private) or desired for your project. We want to advertise this course and its potential to future students!

5. Can I combine this project with another class project?

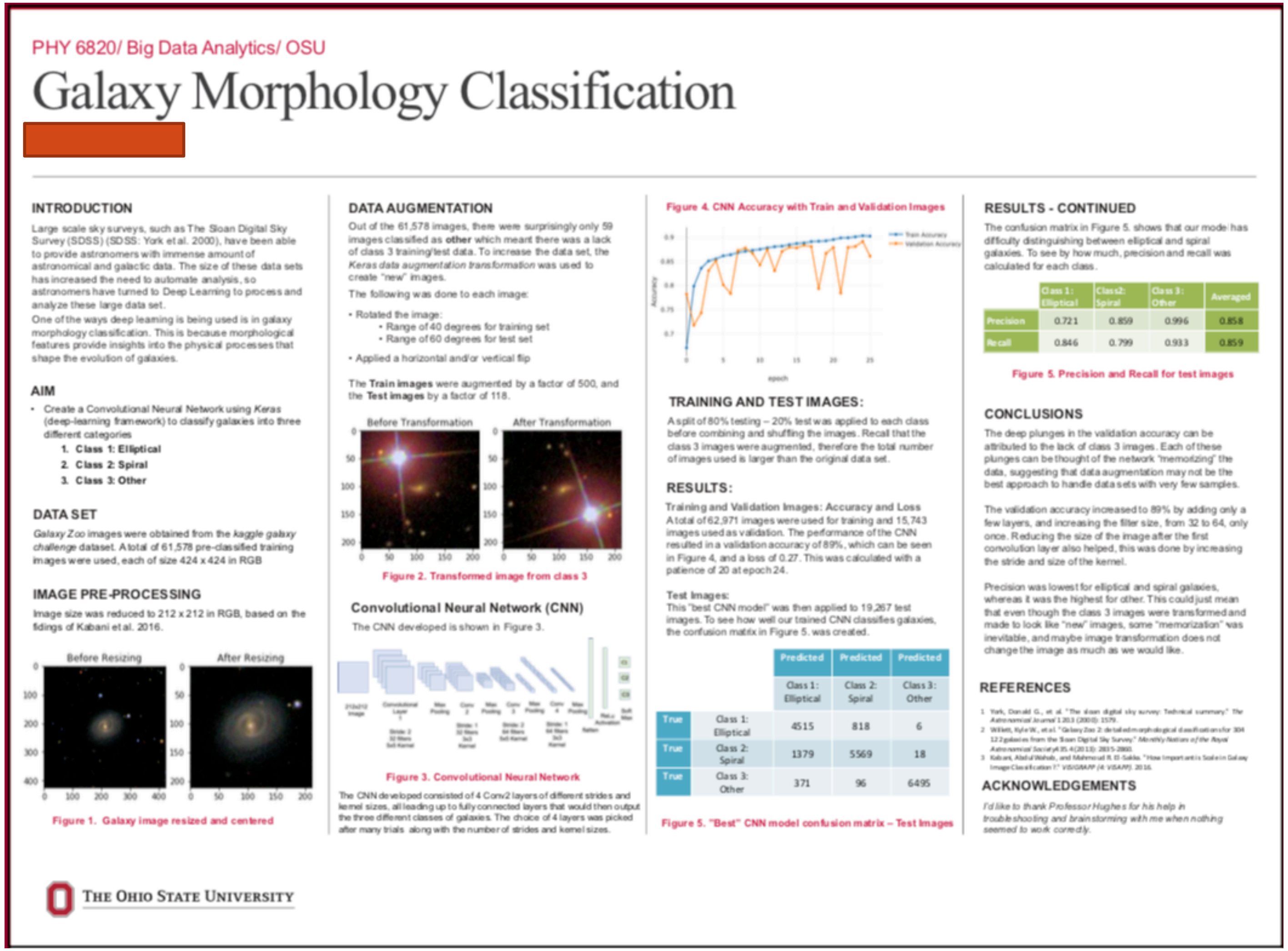
Yes. In general it is possible to combine your project for this course and another class, but with the following caveats:

* You should make sure that you follow all the guidelines and requirements for this project (in addition to the requirements of the other class). So, if you'd like to combine this project with a class X but class X's policies don't allow for it, you cannot do it.

* You cannot turn in an identical project for both classes, but you can share common infrastructure/code base/datasets across the two classes.

* Clearly indicate in your mid-project report as well as an addendum (text document) to your final poster, which part of the project is done for this class and which part is done for a class other than this one. For shared projects, we also require that you submit the final report from the class you're sharing the project with.

Example project from 2018



Example project from 2018

College of Arts & Sciences
Department of Physics

Higgs Search Using Machine Learning

Overview

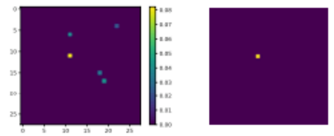
The application of machine learning techniques to particle physics analysis is an active and interesting area of research. Such techniques could obviate the need for lengthy data preparation while improving results.

Goals:

1. Train a five-layer Fully Connected Network (FCN) on all parameters and compare to results in [1]
2. Improve on previous results by means of a more complex network, acting on both low- and high-level parameters.
3. Examine the filters of the trained CNN to determine what features in the data for which its filters are searching.

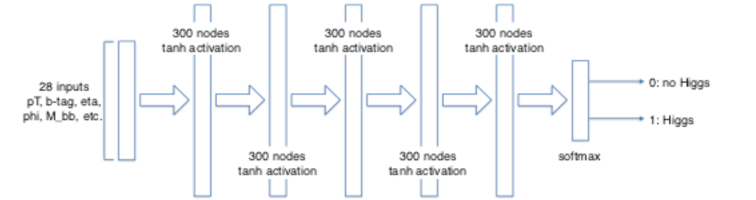
Data

- 1,100,000 Monte Carlo-simulated collider events
  - 500,000 used as test set
- 28 features
  - 21 are kinematic properties (detector outputs)
    - low-level parameters
    - E.g. pT, b-tag, eta
  - 7 are functions of low-level parameters
    - high-level parameters
    - invariant masses (M_bb, M_jj, etc.)
- Each event tagged as a "Higgs" or "no-Higgs" event
- Data organized into eta-phi map, with pT, b-tag and lepton data as layers (example below)

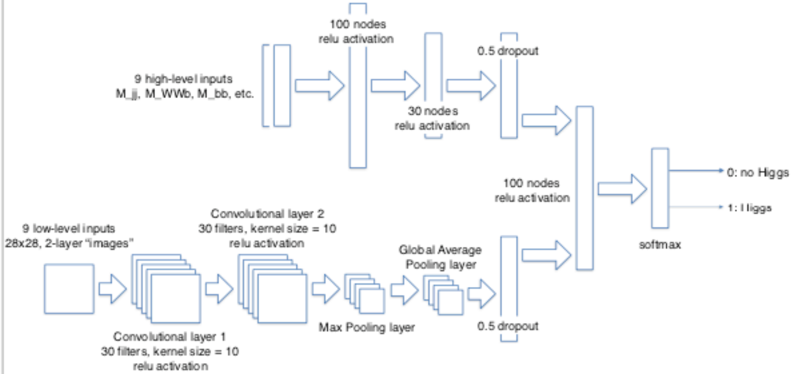


Models used

Two neural network architectures were explored:



- A fully connected 5-layer NN; all 28 parameters were used as inputs (460,802 trainable parameters)
- Matched architecture in [1]



- A two-headed network, consisting of a CNN for the low-level parameters and an FCN for the high-level parameters (106,392 trainable parameters)
- Fewer parameters = lower odds of overfitting

Model Results

Five-layer FCN (1,050,000 training events, 50,000 testing events

	Predicted Background	Predicted Signal
Background	167,780	60,405
Signal	67,695	204,102

- Precision = TP/(TP + FP) = 0.7716
- Recall = TP/(TP+FN) = 0.7509
- AUC = 0.8251

CNN (525,000 training events, 25,000 test events

	Predicted Background	Predicted Signal
Background	7508	2744
Signal	4202	10,546

- Precision = TP/(TP + FP) = 0.7935
- Recall = TP/(TP+FN) = 0.7151
- AUC = 0.7999

Conclusions

- Performance of the FCN was similar to that found in [1]
- The CNN produced similar, but still somewhat weaker results
  - Possible tweaks to structure could result in improved or even superior performance
- The CNN was trained on half of the data and a quarter of the trainable parameters to achieve the present results
  - Possible spatial dependence of the input parameters taken into account by the convolutional structure
  - Could point to underlying structure exploitable in future analysis

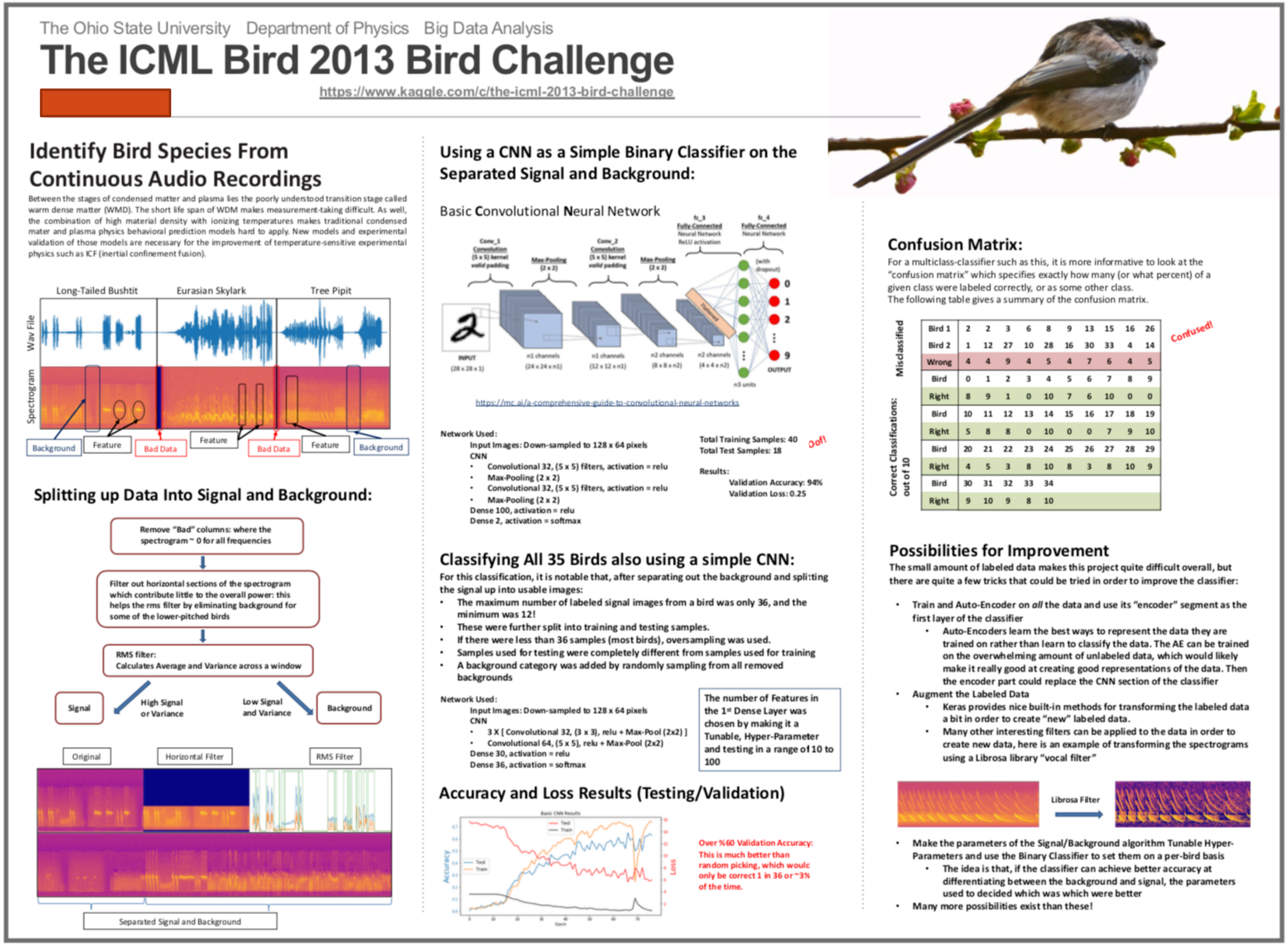
Citations

(1) Baldi, P. et al. Searching for exotic particles in high-energy physics with deep learning. Nat. Commun. 5:4308 doi: 10.1038/ncomms5308 (2014)

(2) Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

THE OHIO STATE UNIVERSITY

Example project from 2018



The Ohio State University Department of Physics Big Data Analysis

The ICML Bird 2013 Bird Challenge

https://www.kaggle.com/c/the-icml-2013-bird-challenge

Identify Bird Species From Continuous Audio Recordings

Between the stages of condensed matter and plasma lies the poorly understood transition stage called warm dense matter (WMD). The short life span of WDM makes measurement-taking difficult. As well, the combination of high material density with ionizing temperatures makes traditional condensed mater and plasma physics behavioral prediction models hard to apply. New models and experimental validation of those models are necessary for the improvement of temperature-sensitive experimental physics such as ICF (inertial confinement fusion).

Splitting up Data Into Signal and Background:

Using a CNN as a Simple Binary Classifier on the Separated Signal and Background:

Basic Convolutional Neural Network

https://mc.ai/a-comprehensive-guide-to-convolutional-neural-networks

Network Used:
Input Images: Down-sampled to 128 x 64 pixels
CNN
- Convolutional 32, (5 x 5) filters, activation = relu
- Max-Pooling (2 x 2)
- Convolutional 32, (5 x 5) filters, activation = relu
- Max-Pooling (2 x 2)

Dense 100, activation = relu
Dense 2, activation = softmax

Total Training Samples: 40
Total Test Samples: 18
Oof!

Results:
Validation Accuracy: 94%
Validation Loss: 0.25

Classifying All 35 Birds also using a simple CNN:

For this classification, it is notable that, after separating out the background and splitting the signal up into usable images:
- The maximum number of labeled signal images from a bird was only 36, and the minimum was 12!
- These were further split into training and testing samples.
- If there were less than 36 samples (most birds), oversampling was used.
- Samples used for testing were completely different from samples used for training
- A background category was added by randomly sampling from all removed backgrounds

Network Used:
Input Images: Down-sampled to 128 x 64 pixels
CNN
- 3 X [Convolutional 32, (3 x 3), relu + Max-Pool (2x2)]
- Convolutional 64, (5 x 5), relu + Max-Pool (2x2)

Dense 30, activation = relu
Dense 36, activation = softmax

The number of Features in the 1st Dense Layer was chosen by making it a Tunable, Hyper-Parameter and testing in a range of 10 to 100

Accuracy and Loss Results (Testing/Validation)

Over %60 Validation Accuracy:
This is much better than random picking, which would only be correct 1 in 36 or ~3% of the time.

Confusion Matrix:

For a multiclass-classifier such as this, it is more informative to look at the "confusion matrix" which specifies exactly how many (or what percent) of a given class were labeled correctly, or as some other class.
The following table gives a summary of the confusion matrix.

Misclassified	Bird 1	2	2	3	6	8	9	13	15	16	26
	Bird 2	1	12	27	10	28	16	30	33	4	14
	Wrong	4	4	9	4	5	4	7	6	4	5
Correct Classifications: out of 10	Bird	0	1	2	3	4	5	6	7	8	9
	Right	8	9	1	0	10	7	6	10	0	0
	Bird	10	11	12	13	14	15	16	17	18	19
	Right	5	8	8	0	10	0	0	7	9	10
	Bird	20	21	22	23	24	25	26	27	28	29
	Right	4	5	3	8	10	8	3	8	10	9
	Bird	30	31	32	33	34					
	Right	9	10	9	8	10					

Confused!

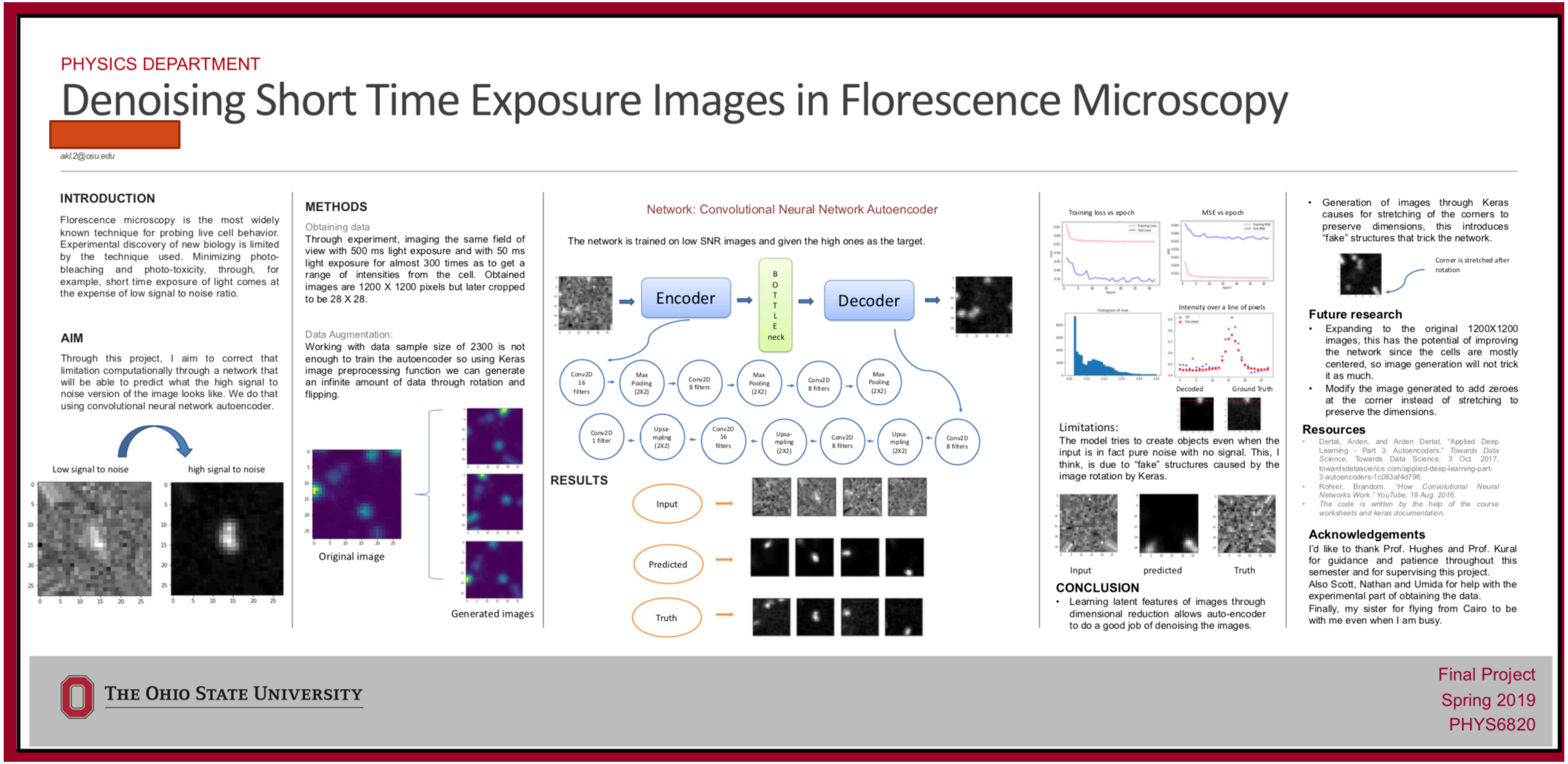
Possibilities for Improvement

The small amount of labeled data makes this project quite difficult overall, but there are quite a few tricks that could be tried in order to improve the classifier:

- Train and Auto-Encoder on *all* the data and use its "encoder" segment as the first layer of the classifier
  - Auto-Encoders learn the best ways to represent the data they are trained on rather than learn to classify the data. The AE can be trained on the overwhelming amount of unlabeled data, which would likely make it really good at creating good representations of the data. Then the encoder part could replace the CNN section of the classifier
- Augment the Labeled Data
  - Keras provides nice built-in methods for transforming the labeled data a bit in order to create "new" labeled data.
  - Many other interesting filters can be applied to the data in order to create new data, here is an example of transforming the spectrograms using a Librosa library "vocal filter"

- Make the parameters of the Signal/Background algorithm Tunable Hyper-Parameters and use the Binary Classifier to set them on a per-bird basis
  - The idea is that, if the classifier can achieve better accuracy at differentiating between the background and signal, the parameters used to decided which was which were better
- Many more possibilities exist than these!

Example project from 2018



		Physics Major Program Outcomes					
		Undergraduate Physics majors acquire a basic mastery of fundamental areas of physics, from classical mechanics, through electricity and magnetism, and finally to modern physics including quantum mechanics and relativity.	Undergraduate Physics majors develop powerful analytical and problem solving skills in areas involving both physics and mathematics.	Undergraduate Physics majors acquire a basic mastery of experimental physics.	Undergraduate Physics majors have acquired a basic mastery of data reduction and error analysis.	Undergraduate Physics majors effectively communicate their physical understanding both professionally and colloquially (orally and in writing).	Undergraduate majors are apprised of and encouraged to participate in academic research, industrial research and/or outreach activities which are consistent with their interest, ability and postgraduate plans.
required courses	Physics 2095: Physics Seminar						3
	Physics 2300: Mechanics I	3	3	1			
	Physics 2301: Mechanics II	3	3	1			
	Physics 3700: Data Analysis Lab	1	3	3	3	3	1
	Physics 5400: Electromagnetism	3	3				
	Physics 5500: Quantum Mechanics	3	3				2
	Physics 5700: Physics Senior Lab	2	3	3	3	3	
required 3rd lab (choose 1)	Physics 3201H: Holography	2	3	3		2	
	Physics 4700: Electonics Lab	2	3	3	2	3	2
	Physics 5680: Big Data Analytics	1	3	2	3	1	2
	Physics 6810: Computational Physics	1	3	2	2	2	2

Relationship: 1 light, 2 intermediate, 3 high