**INFO 6105: DATA SCIENCE ENGINEERING AND TOOLS**

**About this course**

The goal of this online course is to introduce the fundamentals of machine learning (ML) and data science. This course is aimed at students who are looking to learn and apply state-of-the-art ML algorithms to real-world datasets.

There are two unique and foundational pillars of this course— (1) We use real-life situational analogies to intuitively understand the algorithmic aspects of data science and ML, and (2) We use real-life datasets from Kaggle, and use powerful machine learning libraries in Python, to put these algorithms to practice.

Students enrolled in this course can expect to learn both theory and practice of modern data science in a very interactive and intuitive manner. After completing this course, participants will be able to build effective ML pipelines within minutes.

*The programming language we will use for our course is Python*.

**Important Note**

The first few modules will introduce more of “how to think” about machine learning, and we will look only at simple ML algorithms. But somewhere along the way, starting module IV, we will progress steeply to some of the state-of-the art algorithms used in the industry**.**

**Are there any pre-requisites to take this course?**

There are very few pre-requisite for this course:

* 1. Some familiarity with programming. It can be in any language.
  2. Your own Laptop/Desktop

**What will you learn in this course?**

1. Understand the central principles of machine learning from a conceptual, mathematical, and programming point of view.
2. Understand the why and how of data pre-processing for machine learning.
3. How to build, refine, and measure the performance of a machine learning model using real-life datasets.
4. Explore the commonalities and differences between some of the state-of-the-art machine learning algorithms including deep learning.
5. How to do all of the above in a cloud computing framework (Google Collab, a free cloud compute platform)
6. Machine learning best practices for the industry as it’s currently being practiced.
7. Tips and tricks to troubleshoot problems in real-life data sets. E.g. How to tackle datasets containing predominantly one class, i.e unbalanced datasets

**Meet the instructor**



Ramkumar Hariharan is currently lead data scientist with Macro-Eyes, Seattle , where he drives diverse projects in the Artificial Intelligence & Healthcare space. Previously, he has led multiple high-impact data-driven projects at some of the leading institutes in Seattle. These include Fred Hutch, University of Washington (UW), and the Institute for Systems Biology. His areas of focus include data analyses, data visualization, and predictive analytics of both structured and unstructured data.

Ram has a 15-year history of developing and delivering more than 20 computational, biomedical, and data science courses at a variety of levels. His courses, lectures, online teaching, and motivational talks have been overwhelmingly well-received in Seattle, the UK, Japan and in India. Ram serves as affiliate faculty at Northeastern University, affiliate of UW e-sciences institute, bootcamp leader at General Assembly, and mentor with Springboard. He has also led education and training programs for Fred Hutch. He specializes in using powerful, everyday analogies to explain seemingly complex computational and data science concepts and math.

Ram’s teaching philosophy is grounded in one strong belief: there is no one size fits all approach to teach, or to learn a new concept. Each concept/algorithm for this course will be presented in at least 4 different ways — visual, abstract (math), code, and sometimes excel too! In short, it’s a four dimensional, immersive course.

**How is the course going to be delivered?**

This course is organized as a series of modules. Each module will contain

1. 2—4 video or audio lectures (at least 1 hr)
2. Slides from the video lecture
3. Jupyter notebooks with Python code shown in the videos
4. Data for running the examples, assignments and final project
5. Links to great resources
6. Text books suggestions

**How are you going to be graded?**

There will be three assignments in total, one going out every fourth week. Each assignment will have questions that either require you to write a response, select a response, or more likely, write Python code to solve a problem. Your performance on the assignments will contribute 60% towards your final grade.

You are required to complete one final project beginning at week 4. Project will be decided with students who most likely will work in teams of 3 or 4. Scores on your project will contribute 40% towards your final grade for this course.

**Project Details**

You can either pick a project from a list of project this course will offer OR, come up with your own project. You get to decide. E-mail support/ one on one /one on group video call interaction will further bolster the project updates. In this course, you will have all the code to do almost any ML project. So if you find it troublesome to write your own code, you can use the code in the notebooks, and massage it to run your project !!!

**How to ask for help and other benefits**

Ram will be available by email throughout the duration of this course, and will gladly help out students with INFO 6105. If there’s need, we can also do a few live stream interactions.

*Perks: for active data science job seekers, Ram will be happy to leverage his professional network to pass along CVs of students! This has resulted in some of his previous students landing jobs, or sometimes getting interviews from companies!*