Y.C. C.S. Summer 2021: Turbo Charge Your C.S. Education

- **Description:** This special course has been designed to turbo-charge the C.S. education of ambitious first and second year Y.C. C.S. majors. Students will learn aspects of either Data Science or Distributed Systems (depending on the team they choose / are assigned to) that they are normally are only introduced to in their 3rd or 4th year in the major, and apply them to build a real, working system. Guided by accomplished industry mentors, and using professional-grade tools and methodologies, students will first learn important prerequisite tools and systems, and then execute a large C.S. project in teams of four students and one mentor. The course will culminate with each team preparing a video demonstration and public repository of their project.

- **Course Number:** COM 4570H
- **Credits:** 3
- **Course Name:** Industrial Software Development
- **Prerequisites:**
  - Data Structures (COM 1320)
  - Mathematics for Computer Science (COM 1310) OR Discrete Structures (COM 1504)
- **Schedule:** Double Summer session, June 2 – August 5, full time Monday through Thursday and half a day on Friday. Each team will decide exactly which hours daily to meet and work (e.g. 9am-5pm, 11am-7pm, etc.)
- **Student Eligibility:**
  - Students must be honors (or honors-eligible) Y.C. C.S. majors completing either their first or second year in the major this Spring (i.e. taking COM 1320 or COM 2546 this Spring semester.)
  - Students must be registered in a four year, B.S. track in the Y.C. C.S. major – either Distributed Systems or Data Science
- **Admissions into the Course:** Because there are a limited number of spots based on the number of mentors we have, students must apply and receive permission from the Y.C. C.S. department chair to register. Once registration for the course opens, we will try to make admissions decisions quickly so the students know where they stand for the Summer.
- **Requirements Fulfilled:** this course does not fulfill any requirements in the Y.C. C.S. major, but is an honors course.

**Stage 1: Orientation and Software Engineering On-Ramp**
The Summer will start with a crash course in a number of needed technologies, including: version control and collaboration with Github, Amazon Web Services (EC2, S3, Lambda, SNS, RDS), Linux command line basics, web technologies and networking, web programming (HTML, css, javascript.)

**Stage 2: Projects (In order of last name of mentor)**
We will do our best to match students with their preferred track and project, but can’t guarantee placement into a specific track or project.

**Reliable Chat**

**C.S. Track:** Distributed Systems

**Number of Students:** 4

**Mentor:** Henrique Andrade, Senior Site Reliability Engineer, Security Engineering, Bloomberg

**Project Description:** WhatsApp, SMS, and other messaging systems have become significant parts of daily life. Students will build a chat service, starting by researching and using network programming techniques such as sockets and/or gRPC/Thrift. As the project progresses, core distributed systems features, such as scalability and fault tolerance, will be investigated and addressed.
Patent Prior Art Finder

**C.S. Track:** Data Science  
**Number of Students:** 4  
**Mentor:** Dave Feltenberger, Senior Staff Software Engineer, Google  
**Project Description:** Students will use the [Google Public Patents](https://www.google.com/patents) dataset and learn to access, download, clean, and explore the data. They will then use various machine learning techniques to find related patents to a given blurb of text. This project will require students to learn about large datasets, explore approaches to text similarity, and build a simple tool to allow an imaginary patent filer to search for existing patents on a given topic.

Stock Analysis and Selection

**C.S. Track:** Data Science  
**Number of Students:** 4  
**Mentor:** Aliza Heching, Researcher, IBM TJ Watson Research Lab  
**Project Description:** Students will build a tool for stock analysis and selection using publicly available data sets such as 13F filings, Yahoo Finance, and SEC filings. The project will require students to gather, clean, and combine data from the multiple data sources. The consolidated data will then be analyzed to provide the end user with insight to support stock selection decision-making. The tool should be designed so that data and analysis can be easily updated by the end user as data refreshes become available.

Deep Learning in Computer Vision

**C.S. Track:** Data Science  
**Number of Students:** 4  
**Mentor:** Gershom Kutliroff, Chief Science Officer at Taranis  
**Project Description:** Advances in deep learning have revolutionized the field of computer vision and are driving innovations in many fast-growing industries, notably driverless cars, autonomous robots, and AR, to name a few. In this project, we will begin with a quick introduction to convolutional neural networks and learn to work with Tensorflow, Google’s framework for developing and training ML models. Depending on students’ interests and ambitions, together we will select a common problem in computer vision, such as classification, detection, semantic segmentation, or pose estimation, and an appropriate dataset. Students will learn to preprocess data, train deep neural networks, techniques to validate results, and how to run inference on the generated model. Because the frameworks and libraries we will be using are in Python, comfort working in Python is strongly recommended.

Multi-Tier System from Scratch: Customer Relationship Management (CRM)

**C.S. Track:** Distributed Systems  
**Number of Students:** 8  
**Mentor:** Alex Porcelain, Security Architect, Goldman Sachs  
**Project Description:** We will build a specialized customer relationship management (CRM) system from scratch. CRM is one of the largest and fastest-growing categories of software, and it is projected to grow as more data is collected about customers and prospects. The teams will be responsible for designing and implementing a full solution including data stores, middleware, infrastructure and a web application.

Smart City Reporting System

**C.S. Track:** Distributed Systems  
**Number of Students:** 4  
**Mentor:** David Rosenstark, Senior Machine Learning Engineer, Intel Corp.
**Project Description:** Smart Cities is fast becoming a reality. One of the aspects of a smart city is a system to identify problems that occur in the streets of our city, and ensure they will be taken care of. We will build a cloud based system to enable receiving reports of issues found in our city (potholes, fallen street signs, etc.), routing them to the correct departments to handle them and monitoring they are taken care of. We will research ways to enable validation of pictures sent to us of the problems reported and possible triage of different reports.

**Distributed Datastore**  
**C.S. Track:** Distributed Systems  
**Number of Students:** 4  
**Mentor:** Yoel Spotts, Principal Software Engineer at Elastic

**Project Description:** The sales team is in a panic: Our document datastore product has become so popular that customers are clamoring for bigger and better. They are storing so much information, they want the ability to store their data across multiple servers - we need to build a distributed datastore! We will learn about the challenges of distributed data, the pros and cons of different approaches and the tradeoffs we need to make. What should we prioritize? Consistency of data? Availability? Fault tolerance? Can we have all three? This will be an exciting journey into the world of distributed systems.