DataPalooza:
November 9th, 2018
Ian Linville

Main Summary:

The day was segmented up into a series of seminars and sessions: the Fireside Chat/Introduction, the Research Highlights seminars, a skills learning session, and then finally the keynote speaker. I'll describe each in its individual segment!

Fireside Chat/Introduction:

Philip Bourne, the director of the DIS came out and gave a talk about data science and our environment. He wanted people to observe the things around them, and be aware of how we began to mathematically model our environment. He emphasized Jeffersonian principles saying that we should share data openly, since everyone is working together towards the same goal of sustaining humanity. However, in actuality personal goals can cloud these overarching goals, resulting in the need for a cultural change and renewal of the mindset about data science.

Then Jim Ryan came out to discuss the future of data science at UVA. He emphasized his desire to make UVA the flagship university of the country, not just VA, and believes that Data science is a large step towards that. His desire to inspire community, discovery, and service is based in the sharing and collaboration of information and ideas. He's aware that we need to use our own data to improve people's experiences, and that we should be data driven yourself. Both Philip Bourne and Jim Ryan believe that education is a key step in increase data science driven innovation, claiming that likely in the future everyone will be required to know some sort of data processing capabilities.

Research Highlights:

We decided to go to the research highlight session about machine learning, in hopes of further education ourselves about this specific topic. The highlight session was set up so four individual speakers talked for about 15 minutes each on their projects. My favorite two included a second year grad students report about quantifying hypertrophic cardiomyopathy automatically, and how she classifies different classes of hearts based on MRI images. Her goal was to create a CNN image classification and automatic segmentation process. She went into a lot of detail about the processing of the images. Essentially, the process included extracting the 4 chamber long axis end diastole images for each person, then using the CNN trained on previous images, to process and classify the heart. She had some problems using SVM classification at first only getting 68% success, but then when she used a new CNN trained on septal data using a keras library in Tensorflow she got results that showed 78% accuracy. Her level of knowledge of image processing was incredible, and we actually sat with her during the Keynote speaker and discussed our capstone projects with her and she was able to offer some insight to what we could do.

My second favorite talk was about DeepRacing AI, a group trying to teach autonomous vehicles to handle obscure cases in traffic. They trained the (min) cars under similar simulations to F1 racers, showing how they cut corners, and how cars pass each other either around or between the side of the road and the other car. They emphasized using localization and mapping to create a scene understanding. The car must know where it is at all time in relation to everything else around it, and see the signs and be able to identify what is a car vs person vs stoplight etc. The idea they have is that safety of automatic cars should be based on their ability to react and their agility. If a car is able to register a swerve, and react with agile behavior, it will be much more safe than a car trained simply to stop. They collect data through thousands of trials, and even have a competition every year to see how well people can create these AI F1 cars.

Skills Sessions:

I signed up for the Machine Learning session, based in Python. We created an algorithm that took flight data and was able to predict the flight delay of any given flight. The overall process of machine learning is: Ingest, Process, Predict, and Visualize. The first two points are the typical process of data engineering, whereas the second half is more of the machine learning aspect. The official summary of what we accomplished was: Create a data science VM in Azure, import the data into VM using curl, create a jupyter notebook in the VM, use Pandas to clean and prepare data, use Scikit-learn to build a machine learning model, and then finally use Matplotlib to visualize the results.

One thing that I thought was really cool in this session was learning how to use these virtual notebooks such as Azure and google collab to create code in an online IDE environment. I actually had never even heard of either of these IDE models and it was a really interesting change than the standard IDE's I've always used. It was also really nice to get to use Python again after not having used it since my second year.

Keynote Presentation:

Presented by Robin Thottungal, CTO and Chief Data scientist at the National Gallery of Arts, this was my favorite session of the day. He discussed so many outside political perspectives that I have always been concerned about when it comes to AI. While he discussed many aspects of it, he kept bring it back to the main concern of people's involvement and trust in AI. He was insistent that data science should have empathy, because we must be concerned with the inherent bias in training data used for the AI. "Predictive policing" is the idea that people are screened through a model, and then judged to see if they would commit a future crime based on the AI model. This system was shown to be incredibly racist, and enabled decision makers to make decisions that affected people's live under extremely biased "advice" from the algorithm.

He also mentioned a lot about how we need to understand the audience and stakeholders involved in AI. Things like automatic trucks, or burger flipping AI, would displace thousands to millions of industry workers, who then will need to find new jobs or placements in society. AmazonGo puts retail workers and cashiers out of business, AI vehicles could put out truck drivers, the second most common industry in America among men.

The emphasis on using data science to benefit society as a whole rather than just produce information really stuck out to me. He believes that we could use data science in a good way to show the public impacts of things such as climate change, and the larger problems that normally people are not able to understand. You can even use data science for impact analysis of first responders, trying to understand who should respond where under given environmental disasters such as hurricanes.

I feel like I always here about how amazing Data Science is, and while I agree, I personally think it's extremely important to be aware of the negative effects it could possibly bring as well. AI and machine learning has the capability to give extreme benefits, but also extreme suffering in the terms of jobs.

Conclusion:

I learned a ton during Datapalooza! It gave me so much hope for the future of data science, seeing how incredible and smart some of the projects people are working on are. I actually feel really grateful for having the opportunity to experience it. I learned about both smaller projects, and the larger grand scheme in the data science community. I left with some major questions, such as how do we remove this bias in AI learning databases. Will this even be possible? For example, there was an article about how a facial recognition software was much more accurate on white people than minorities due to the database of images being of majority white people. This was so bad that google's facial recognition actually recognized black people as gorillas. Is there a way that we can tell what a database will be biased on before we let the AI learn? Or is it simply a trial and error process until we create a non-biased AI. These are the things I want to keep looking into, and a field I had not thought about much until Datapalooza.