Interviewees: Marc Brocksmidth

- Marc is a Senior Research Manager at Microsoft Research at Cambridge, UK.

- His focus is on the intersection of Deep Learning and Software Engineering namely “Deep Program Understanding”.

- His Phd and post-doctorate research from RWTH Aachen was on automated termination analysis of heap-manipulation imperative programs with focus on Java Bytecode.

- He has been with Microsoft Research since he received his Phd being involved with Post-doc research since 2011.

    - **Microsoft Research (MSR)** is the research subsidiary of Microsoft. It was formed in 1991, with the intent to advance state-of-the-art computing and solve difficult world problems through technological innovation in collaboration with academic, government, and industry researchers.

Interviewees: Anmol Joshi

- Anmol is a Service Engineer at Boeing where he is working on Automation of Customer Service requests via Data Science Methods.

- He graduated from the University of California, San-Diego with a Bachelor's of Science in Mechanical and Aeronautical Engineering.

- He has completed coursework from Harvard University and the University of Toronto.

- He is an extremely passionate Deep Learning Hobbyist who has made it is career as well.

- He has been with Boeing since 2015.

    - **The Boeing Company** is an American multinational corporation that designs, manufactures, and
      sells airplanes, rotorcraft, rockets, satellites, telecommunications equipment, and missiles worldwide. The company also provides leasing and product support services. Boeing is among the largest global aerospace manufacturers; it is the second-largest defense contractor in the world based on 2018 revenue.

Evaluation of the Interview

- Your Personal Analysis of the Data Scientist Professional's Disposition

    - Both data scientists were highly passionate about their respective fields. Marc has a more managerial role now but is still actively involved in all the research. The fact that despite being incredibly qualified, he helped me answer all the questions I had with just incredible depth was very much appreciated!

    - Anmol is a more hands-on data scientist where he is leading multiple projects in the field of sequence models. His story about getting into the field of Data Science is inspiring: his studies are in engineering but always had a knack for statistics. His passion for statistics and Machine Learning got him interested in pursuing Deep Learning as a career.

- The interviewee's level of professionalism

    - Super professional data scientists who were in different stages of their careers but both exemplified a sincere passion in each of their fields. Each interviewee talks about the importance of data cleaning, research and well written code.

- Analysis about how desirable it may be to work as a data science professional

    - **Microsoft Research**: I work at Microsoft, however, Microsoft Research sounds like an incredibly ideal place to work! Autonomy and researching the bleeding edge of technology are one of the few benefits. The pay is highly competitive and a considerable amount of equity is given.

    - **Boeing:** I am more inclined to work at a Software based firm. Boeing does seem like an awesome place to work but seems like the domain isn't my cup of tea. Although, there is a lot to be learnt from some of their projects. The pay is competitive, however, the recent climate and issues has resulted in a bit of turbulence in terms of compensation. He does get full autonomy based on the stakeholders over all vision.

Evaluation of the Interview

- Ideas that might improve or enhance the interviewee's flow of information in the field of data science

  - For Marc, the ideas are permeated in the field through constant publishing of articles and further research into the bleeding edge of Deep Programming Learning. He mentioned that there is a constant influx of collaborative innovations that cause the field to progress. Another aspect of the flow of information would be to encourage Open Source Software Development through Github.

  - For Anmol, the ideas he gets are from other senior data scientists and research papers as well. He recently attended "Deep Learning Summer School" in Canada that provided him with considerable amount of insight.

Lessons Learnt

1. There is no silver bullet in model creation. The evaluation depends on the data.
2. Keep the code simple and highly decoupled from the data.
3. Treat the code that generates the initial representations of your data with as much importance as your modeling code.
4. The code that can be used for initial representation can be made decoupled from our modeling code if the pipeline is setup correctly.
5. Experiment as much as you can! You will never get the right answer right away as there is no right answer. It's only improvement of representation.
6. Keep researching and reading white papers to learn ideas from across other subfields.

Appendix: Questions

I asked as many questions from the list provided. However, here are the specifics to the project:

Initial Node Representation

- Could you please elaborate on the meaning of representing the type information of the variable as the element-wise maximum? Does this mean it is always the leaves of the inheritance trees are used here? For example List<Animal> is replaced by List<Dog> and never List<Animal>?
- How were the strings from the subtokens / type information mapped to a vector of features? Was transfer learning at all used here in this stage?

Evaluation

- Was there any correlation between the accuracy of the results and the size (in terms of lines of code) of the project?
- Based on the Ablation studies was there any pattern recognized based on the types of edge enabled and the length / category of project?
- Does the code to run the AvgBiRNN for the baseline models exist somewhere on Github? I'll be super curious to run the model and test out with different hyperparameters. I'd also be curious to see how a plain-vanilla RNN unit would do here.
- The examples mentioned in the paper for the VARMISUSE task are all pertinent to C# repositories. Was there any reason why python modules weren't used to generate the graphs using the ast module?

General Questions

- Has further work been done to swap out the Gated Recurrent units with "fancier" units such as LSTM units?
- What are the natural successors to the GGNN, if any, in terms of improved models? Are there any papers you can recommend?
- As I am familiarizing myself more with the codebase, is there way I can contribute? I didn't see any open issues that demand any immediate attention. It'll be super awesome for my understanding to help in any way with the tf-gnn-samples or the dpu-utils.