Introduction. My name is Jason Wei and I am interested in doing a PhD in natural language processing (NLP). I recently presented a paper on data augmentation in NLP as part of the main conference at EMNLP-IJCNLP 2019. Before that, my research as an undergraduate at Dartmouth focused on medical image analysis, for which I have published four papers in medical journals and one workshop paper at NeurIPS 2019. I aim to conduct creative research to move the needle in NLP.

My NLP research. My most recent research explored data augmentation techniques in NLP. While looking for methods to combat overfitting in models trained on small datasets, I realized that the research on text data augmentation was sparse compared with the techniques commonly used in computer vision, and I was inspired to address this gap in the field. In my paper, I presented four token perturbation operations for augmenting text data, which I called EDA (Easy Data Augmentation), and optimized their parameters. I tested EDA on five benchmark text classification datasets and found that, on average, training with EDA using only 50% of the training set achieved the same accuracy as normal training using all available data. Through this research, I learned to intentionally design studies, conduct rigorous evaluation, and write clearly - skills that will serve me well during my PhD. I presented my paper at the 2019 EMNLP-IJCNLP conference in Hong Kong and posted my code online, which has reached more than 300 people on Github so far.

NLP research interests. One topic that interests me is curriculum learning in the context of unsupervised pre-training. The current standard language representation model, BERT [Devlin et al, 2018], uses masked language model and next sentence prediction as pre-training tasks to achieve strong results. I think we can do better, however, using curriculum learning, a paradigm from computer vision in which models learn better features when gradually pre-trained on harder tasks. Imagine, for example, a sequence-to-sequence task in which two words in a sentence are swapped and a language model predicts the original sentence. We use this two-word-swap prediction task for pre-training, and once the validation loss for this task plateaus, we then increase the difficulty of the task by swapping three words and asking the model to predict the original sentence. After the model learns the three-word-swap task, we then swap four words, and so on, increasing the difficulty of the pre-training task until the model receives the words in a random order and learns to form the original sentence. Intuitively, this task (or other similar tasks) seems harder than the two pre-training tasks used in BERT, and so it would be interesting to see whether this curriculum learning idea can push the performance of BERT even higher.

Previous work in medical image analysis. Before my interest in NLP, I first became involved in machine learning through research in medical image analysis, for which I have published five papers. I worked on generative image translation for pathology images and presented my paper on the topic at this year's NeurIPS Machine Learning for Health Workshop. I have also published four medical journal papers on deep learning for histopathology image analysis. Working in the medical image domain, where datasets are often small, taught me to rigorously analyze data manually and to take creative approaches to preventing overfitting, skills that I hope to bring with me to the NLP domain.

Why I do research. My career goal is to move the needle in NLP by doing creative research that inspires others. I recently found out that the techniques from my paper on data augmentation in NLP were taught in Stanford's Natural Language Understanding course (CS 224U). In fact, a group of Stanford graduate students there had developed a similar technique for question answering and had cited me in their paper, saying that my techniques had inspired them. As someone who watched my fair share of Stanford lectures when I first began studying deep learning, it was motivating to see my research taught in a course and inspiring others in the field. Seeing how I, even as an undergraduate student, can have an impact in the field of NLP is what drives me to work harder, and I hope to continue working on projects that push the field forward and inspire new ways of thinking.

Why I want to do a PhD. I'm excited to focus on research and dive deep into NLP during my PhD. When I'm actively designing a study or running experiments, I often feel a state of flow, and sometimes, even when I'm doing other things like working out or trying to fall asleep, my mind can't seem to get away from the problems I'm working on or ideas that I've been interested in. I think that this is an indicator that I will enjoy and thrive during my PhD. The explosion of machine learning has excited and inspired me for much of my last two years; as a result, becoming an expert and moving the field forward through doing a PhD is an obvious choice for me.

Why I'm interested in CMU. Carnegie Mellon has one of the top machine learning programs in the world, and I want to join one of the largest communities of talented students and professors who are interested in the same research questions as I am. In particular, I'm interested in the Language Technologies Institute, where I would love to work with professors such as Jaime Carbonell and Edward Hovy on topics including text generation, text summarization, and question answering. Finally, after listening to Randy Pausch's Last Lecture and reading his book, I have been intrigued by the culture of CMU and would love to experience it firsthand.