Gonna take ~1500 words to get to a 10 minute video

Hello and welcome to my presentation on Weakly Supervised Person Name Transliteration using Twitter Data. My name is Greenland and this project was done in collaboration with Fivecast. My supervisors are Matt Lowry and Jason Signolet, my academic supervisor is Lingqiao Liu.

**Introduction**

Firstly, let me introduce what exactly is transliteration. Transliteration is the act of converting a word from one writing system to another. Transliterations express the word directly through sounds.

Note that this is different from translations, translations instead convert the meaning of a word from one language to another.

As an example going from Arabic to English, the Arabic word Allah transliterated would be Allah, we express the sounds of the Arabic word using English sounds. Allah translated would be God in English, here we express the meaning of the Arabic word.

**Motivation**

You might be wondering, why keep looking into transliterating? Surely the act of converting sounds of one language into another has already been implemented already. And you are correct, there exists standard transliteration systems for pretty much all language pairs. However, when choosing a name, not everyone is going to use a standard transliteration system, many will self-transliterate their name. In this project we are interested in capturing this informal transliteration process.

The produced system will provide an alternative transliteration to the standard transliteration systems.

In this project we are interested in Person Name transliterations, this is the act of transliterating a person’s name from one language to another. [probably remove]

A real-life application of transliterations is in the security vetting process of foreign nationals. To accurately determine the profile of someone given a single name, we would want to investigate documents in multiple languages. To identify the person of interest we need to search using the transliterated name.

**Data source**

Twitter provides the data that is driving this whole operation. Twitter is chosen specifically because: it has a global user space, users have to enter an English user name along with a screen name that can be in any language. Therefore, a user signing up to Twitter is forced to transliterate their screen name when entering their username. However, this is very noisy data as users do not have to provide a transliteration of their screen name. The solution to this problem will be explained in the cleansing section.

**Project Goal**

Now that we are properly motivated, we will have a look at the project goals. The overarching goal is to create a pipeline that takes in

**Methodology**

**Results**