1. **Motivation for this project. What is the problem?**

As a budding data analyst or data scientist I wanted to know which is the best location in USA for me to settle based on my preference.

Currently there is no such system which predicts the location based on the needs.

1. **How to approach this problem?**

For building predicting system we will need job title, company names, years of experience, qualification, salary, preferred livability, weather and job opportunities. By using all these features, we will be predicting the location. For this we have collected data from multiple sources namely Levels FYI, weather data, tech employment count, livability scores. Based on the requirements of an individual we will predict and tell which will be the most preferred location for him/her to work.

1. **Data characteristics, and how to process data.**

The data for building this model is collected from different sources. The levels\_FYI dataset is collected from Kaggle 2021.

**Accuracy:** In our data some features are not accurate. We will fix in data engineering. Eg: levels - column

**Completeness:** Base salary, tags and education have missing values. We will use Nearest neighbor imputation (Base salary only 5% missing value, numerical value), mode imputation (tags and education).

**Reliability:** Taken from levels FYI website, so its reliable (tell about the website)

**Relevance:** As we are predicting the locations, we have collected all data relevant to the problem domain.

**timeliness**: As we are collecting most recent data, we will be able to verify if the model gives the correct prediction.

the weather data is collected from <https://www.currentresults.com/Weather/US/average-annual-state-temperatures.php>.

The tech employment count is collected from <https://www.vox.com/2017/9/26/16346960/software-industry-growth-kansas-indiana-idaho-north-carolina>.

The livability score data is collected from <https://wallethub.com/edu/best-states-to-live-in/62617>

**how to process data**

we will fix the inconsistencies in the levels column by comparing the level names of each company and divide them to 4 levels. The base salary column has 5% missing value which will be fixed using Nearest neighbor imputation algorithm. mode imputation (tags and education). We will split the location feature into city and state. We will drop cityid , dmaid, row number, timestamp and races as these features are not relevant to location. Then we will join the weather rank data by states.

The states are categorized good , average and bad based on the weather rank.

tech employment count (same as the weather). livability score(same as weather). Finally the categorical values are converted using 101 encoding.