



CELENA TOON

PYTHON FOR DATA SCIENCE, CSCI E-29

HARVARD UNIVERSITY

---

# AN ATTEMPT TO SOLVE CRIME WITH SUPERVISED LEARNING

# CSCI E-29 2018 STAFF

---



Nenad Svrzikapa  
Instructor



Lena Hajjar  
Teaching Fellow



Philip Lodine  
Teaching Fellow



Kaleigh Douglas  
Teaching Fellow



Alan Xie  
Teaching Fellow



Joe Palin  
Teaching Fellow

# ABSTRACT

---

MAX = 10

- ▶ Dataset: U.S. Homicide Reports from 1980 – 2014
- ▶ Approx. 30% of these cases are unsolved
- ▶ Can we use machine learning algorithms to predict information about the perpetrator in homicide cases?

# REQUIRED LIBRARIES AND FILES

---

- ▶ PANDAS

```
conda install -c anaconda pandas=0.19.2
```

- ▶ Numpy

```
conda install -c anaconda numpy=1.13.3
```

- ▶ Scikit-learn

```
conda install -c anaconda numpy=0.19.1
```

- ▶ Homicide Reports, 1980-2014 Dataset

```
https://www.kaggle.com/murderaccountability/homicide-reports
```

URL

- ▶ [https://github.com/seatuna/ml-homicide-project/blob/master/knn\\_homicide.ipynb](https://github.com/seatuna/ml-homicide-project/blob/master/knn_homicide.ipynb)

# K-NEAREST NEIGHBORS

---

- ▶ Trained 2 KNN algorithms ( $k = 5$ )
  - ▶ 1. Age left as continuous data
  - ▶ 2. Encoded Age columns into categorical data
- ▶ Target: Victim age, sex, ethnicity, weapon
- ▶ Labels: Perpetrator age, sex, ethnicity, relationship
- ▶ Label Encoding and One Hot Encoding for categorical data

# RESULTS

---

- ▶ Accuracy
  - ▶ KNN with continuous ages: 13.47%
  - ▶ KNN with encoded (categorical) ages: 4.3%
- ▶ Increasing k decreased the algorithm's performance
- ▶ Possible conclusions:
  - ▶ Too many uncorrelated columns?
  - ▶ Too many labels to predict?

# REFERENCES

---

- ▶ Standardized Survey Classifications

- ▶ <http://www.pgagroup.com/standardized-survey-classifications.html>

- ▶ Introduction to Machine Learning with Python

- ▶ <https://www.amazon.com/Introduction-Machine-Learning-Python-Scientists/dp/1449369413>

- ▶ Data Analytics Takes on Serial Killers:

- ▶ <https://www.forbes.com/sites/metabrown/2017/12/09/data-analytics-takes-on-serial-killers-aspiring-data-scientists-try-this-at-home/#384d58704795>

# ACKNOWLEDGMENTS

---

- ▶ Our wonderful instructors and TAs of CSCI E-29!