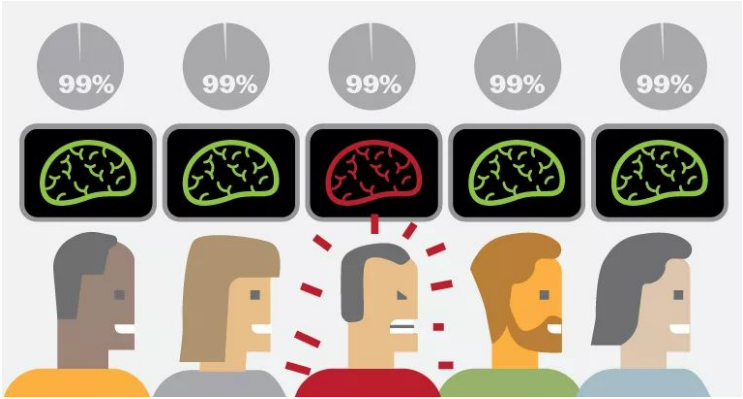
Mass Shootings in the US and the Factors that Cause them

Big Data Science - Spring 2018

Herbert Li
Daniel Rivera
Ross Abramson

Abstract: The Core Plan

- There is a lacking efficiency in assessing the likelihood of mass shooters.
- Mass shooters tend to behave sporadically, but there are recurring patterns.
- Through a mixture of criminology consulting and data processing, we will develop a ranking system to give insight into the likelihood of a mass shooting based on several internal (ex: psyche) and external features (location, time, gun sales).



Business Understanding

- **Problem? The current system.**
- **Objectives? Identify key features.**
- **Benefits?**
- **Who would want this data?**

Data Sources

- **Firearms Data** - ATF (The Bureau of Alcohol, Tobacco, Firearms, and Explosives) Listing of Federal Firearms Licensees and Listing of Firearms Manufacturers
- **Crime Data** - datasets containing all recorded mass shooting from 1982-2012 and Gun violence database
- **Health Data** - Suicide Death Rate per State (2014-2016), National Mental Health Services Survey (2016)
- **Education Data** - State Dropout and Completion Data (2016), School Survey on Crime and Safety (2009-10)

Project Architecture

- Since we're taking in a lot of data from a lot of different sources, we'd like to use **Hadoop** and **Spark** for this project, and use **RapidMiner** for fast prototyping
- Programming Languages: we plan to use mainly the **Java** and **Python** APIs
- Two scenarios:
  - **Clustering Task**: use clustering to see whether there is a common thread between mass shootings
  - **Ranking Task**: Train a machine learning model so that if it is given a set of test features for a particular zip-code, it would be able to predict how likely a mass shooting will happen in this area
- Data → Hadoop HDFS → Feature extraction using spark.mlib → spark.ml (for preliminary tests) → Feature Selection → spark.ml (final models) → visualizations (using dimensionality reduction techniques) and spark.ml for evaluation and results on test data

Past Research

- **Why gun control can help prevent gun violence? (Loria, 2018)**
- Firearm regulations in the U.S.: trend of gun violence. (Bouchet, 2017)
- US Mass Shootings Analysis 1966-2017: do demographics correlate to shooters in mass shootings? (Smith, 2018)
- A Guide to Mass Shootings in America. (Follman et al., 2018)
- **Mental Illness, Mass Shootings, and the Politics of American Firearms. (Metzl et al., 2015)**
- **The impact of the Orlando mass shooting on fear of victimization and gun-purchasing intentions. (Stroebe et al., 2017)**
- Socioeconomic factors and mass shootings in the US. (Kwon et al., 2017)
- Columbine Revisited: Myths and Realities About the Bullying–School Shootings Connection (Mears et al., 2017)

Next Steps...

- Come up with the list of data sources that we consider relevant to the task at hand: demographics, gun-control laws, mental illness, income inequality, bullying, etc.
- Analyze the data to extract the most relevant features for our hypothesis i.e., information that can help us quantify the likelihood of a mass shooting.
- Decide what the output of our model should be:
    - Numerical value? (e.g. probability, score.)
    - Categorical value? (define categories and thresholds).
- Come up with a way of quantifying the performance of our model: will traditional metrics like accuracy be adequate, or will we need to use something else?
- Once the model has been run, an iterative improvement process can be executed:
    - Can we extract new features from the data that might improve our model performance?
    - Rank features according to their predictive power and select the optimal set.