Random Notes

**Google Cloud Platform** (cloud.google.com)

Scalable computing power as needed for our work

Google Compute Engine (GCE)

* Similar to Amazon Elastic Compute Cloud EC2
* Rent computers as needed
* IaaS (Infrastructure as a service)

Google App Engine (GAE)

* Fully managed environment for running applications
* Easy deployment
* PaaS (Platform as a service)

BigQuery

* Analysis for BigData (lots of data) in the cloud
* Results can be saved as tables or comma separated values

Storage

* Cloud SQL (relational MySQL database)
  + comma/tab separated values can be stored as tables
  + BigQuery output
  + Suitable for our work
* Cloud Datastore
  + No fixed table format
  + NoSQL schema-less database
  + Probably not needed
* Cloud Storage
  + General storage for anything

Services

* Prediction API
  + Machine Learning Algorithms from Google!
  + Could be very useful for our work

**Software Setup**

* Install Google Cloud SDK
  + On Mac do:

curl https://sdk.cloud.google.com | bash

* + Other platforms look at

https://developers.google.com/cloud/sdk/

* Authenticate to Google cloud platform
  + From terminal run ‘gcloud auth login’
* Download Eclipse IDE from eclipse.org
* Download Google Plugin for Eclipse for easily building cloud based application from within Eclipse
  + Help->Install New Software

https://dl.google.com/eclipse/plugin/4.3

**Project Setup**

* Login to Google
* Go to cloud.google.com and click “Go to my console”
* Create Project
  + I have created a project called called **PredictGenocide** with project ID **optimus-ml**
  + Requires credit card number

**Data**

* **GDELT**
  + [**http://gdeltproject.org/**](http://gdeltproject.org/)
  + quarter-billion-record GDELT Event Database of human activities
  + available in Google BigQuery
* **Parus Analytics**
  + <http://eventdata.parusanalytics.com/>
  + Country and region specific data
  + GDELT probably has everything
  + Books and papers
    - http://eventdata.parusanalytics.com/books.html

**Genocide Prediction**

* Training data and test data
  + Train/tune model using training data and evaluate performance against test data. For example can model predict genocide in Darfur (2003) based on all past events upto 2001/2002?
* Prediction level
  + Country Level
    - too generic for complex big countries like India, China
    - Almost all available genocide prediction work at country level
  + State/province level
  + District/county level

|  |  |  |
| --- | --- | --- |
| Genocide Predicted | Genocide Occurred |  |
| Yes | Yes | True Positive (TP) |
| Yes | No | False Positive (FP)  (Type I error) |
| No | No | True Negative (TN) |
| No | Yes | False Negative (FN)  (Type II error) |

* Precision
  + TP / (TP + FP)
  + Fraction of predicted genocide instances that actually occurred
  + Also called positive predictive value
* Recall
  + TP / (TP + FN)
  + Fraction of actual genocide occurrence that are predicted
  + Also knows as ***sensitivity***
* Specificity
  + TN / (FP + TN)
  + Also known as true negative rate
* ROC Curve
  + Plot of Rate of False Positives vs True Positives
  + The best classification has the largest area under curve
* A good model should strive to
  + Eliminate or Minimize False Negative
  + Minimize False Positive (although false positives could result from genocide prevention)
  + High precision and high recall
* Forecasting
  + The sample data must include all combinations of politically stable/unstable and genocide/no-genocide cases to create models with forecasting power (5 year forecasting)
  + For genocide to occur a country need not be experiencing political instability. Similarly a country experiencing political instability need not have occurrence of genocide
  + Real time forecasting
  + Forecasting for all countries in all years and out of sample
    - model does not ‘see’ the data it is trying to predict
* Forecasting Levels
  + Structural
    - Predict the cases (regions or countries) most likely to experience conflict
  + Dynamic
    - Predict a probability of conflict breaking out at a known point in future
  + Counter Factual
    - Predict how the change in some policy (introduction of aid or peacekeepers) will affect the likelihood or magnitude of conflict
* Complex models are not necessarily better
* Variables
  + Infant Mortality rate
  + Literacy
  + GDP
  + Religious mix of population
  + Gender mix
  + Conditions in neighboring countries
  + Government (Authoritative, Democratic, military)
  + Inequality
    - Income inequality
    - Rights inequality
  + Refugees from abroad
  + Counter-insurgency campaigns (to eradicate guerrilla forces)
  + Past conflict history
  + Political instability
  + Economic conditions
  + Elections
  + Assassinations
  + Condition of women
  + Access to healthcare
  + Access to electricity
  + Access to water
  + Access to toilet
* Many variables are correlated but correlation is not causation
  + Focus on forecasting power or determining causal relationship?
  + Correlated variables are basically risk indicators like symptom of medical condition
    - High blood pressure is associated with high risk of heart disease and it is also a cause (Causation)
    - Chest pain is a predictor of heart trouble although it is not a cause (Correlation)
    - Presence of peacekeeping force tends to precede genocide onset in some cases (Correlation)
  + For actionable policy ‘targets’ we need to determine causal factors
  + For forecasting using risk indicators we need correlation
* Monitoring Agencies
  + The Satellite Sentinel Project (<http://satsentinel.org/>)
    - Limited scope (monitors only Sudan)
  + Genocide Watch
  + The International Crisis Group (ICG)
* Asia
  + Has experienced decline in the incidence of genocide and politicide from the beginning of the 1980s.
  + No recorded genocide since 1992 except Sri Lanka in 2009
* Structural indicators such as GDP, infant mortality, past or adjacent conflicts change too slowly. They nonetheless affect the overall probability
* Social media indicators change too quickly
* Newsworthy events are “just right”