CIS 550 Project - Milestone 2

Group 15
Hu Chenchen 11021747
Xiang Chen 66770481
Zhong Wen 54397279
Dhwani Kapoor 84152362

Motivation for the idea

The project is based on developing a prediction model for predicting the probability of a country to win a medal in the next Olympic Games. The existence of data in the form of time series offers flexibility of using machine learning algorithms for prediction.

A nation's performance in previous Olympic games, past performance of athletes, geography, economic factors and population are some of the important prediction variables that can decide the preparation of the nation for upcoming Olympic Games.

- Economic factors – People of affluent nations have affinity towards spending more time on recreation and sports than poorer nations
- Geography – Location(latitude) offers flexibility to a nation for practicing in a particular sport discipline(s).
- Population – Separates larger nations of the world from many small nations, both in terms of number and genetic diversity
- Athletes' performance – An athlete's' performance is related with a variety of other entities (such as the country, the year, other competitions) which supplements the development of a diverse schema

Features that will definitely be implemented in the application

1. General statistics of the Olympic games with respect to national ranking, best players, composition of medals as well as the changes between different years (d3.js interactive charts)

2. Build models to predict the number of medals, the distribution of medals, winner and performance of certain nationality/gender/athlete

Features that might implemented in the application

1. Show the influence of historical events/macroeconomics/technology development on the Olympics

2. Show the social profile of a certain athlete (social networks, world ranking, news etc.)

Technology and tools to be used

1. Node.js, also some modules in Node.js like Express.
2. D3.js
3. AngularJS
4. HTML/CSS/jQuery
5. Facebook, Twitter, Google API
6. AWS
7. Relational Databases Service (SQL)
8. DynamoDB(NoSQL)

Complementary data sources

1. Olympics records after 2008

2. Macroeconomic features

3. Other national competition records, athletes' ranking

Hockey:
http://www.opensourcesports.com/hockey/
http://www.flyershistory.com/cgi-bin/hsp-inventory.cgi

Baseball:
http://www.baseball-databank.org/
http://seanlahman.com/baseball-archive/statistics/

Swimming:
https://www.swimrankings.net/index.php?page=meetSelect&selectPage=BYTYPE&meetType=8
http://www.itftennis.com/seniors/rankings/rankings-list/players.aspx?Gender=F&AgeGroup=V35&Nation=Any&From=0&To=25&Name=&Type=S

4. Athletes' profile
http://www.sports-reference.com/olympics/athletes/
(includes biography information like weights, height, nationality, age and sex)

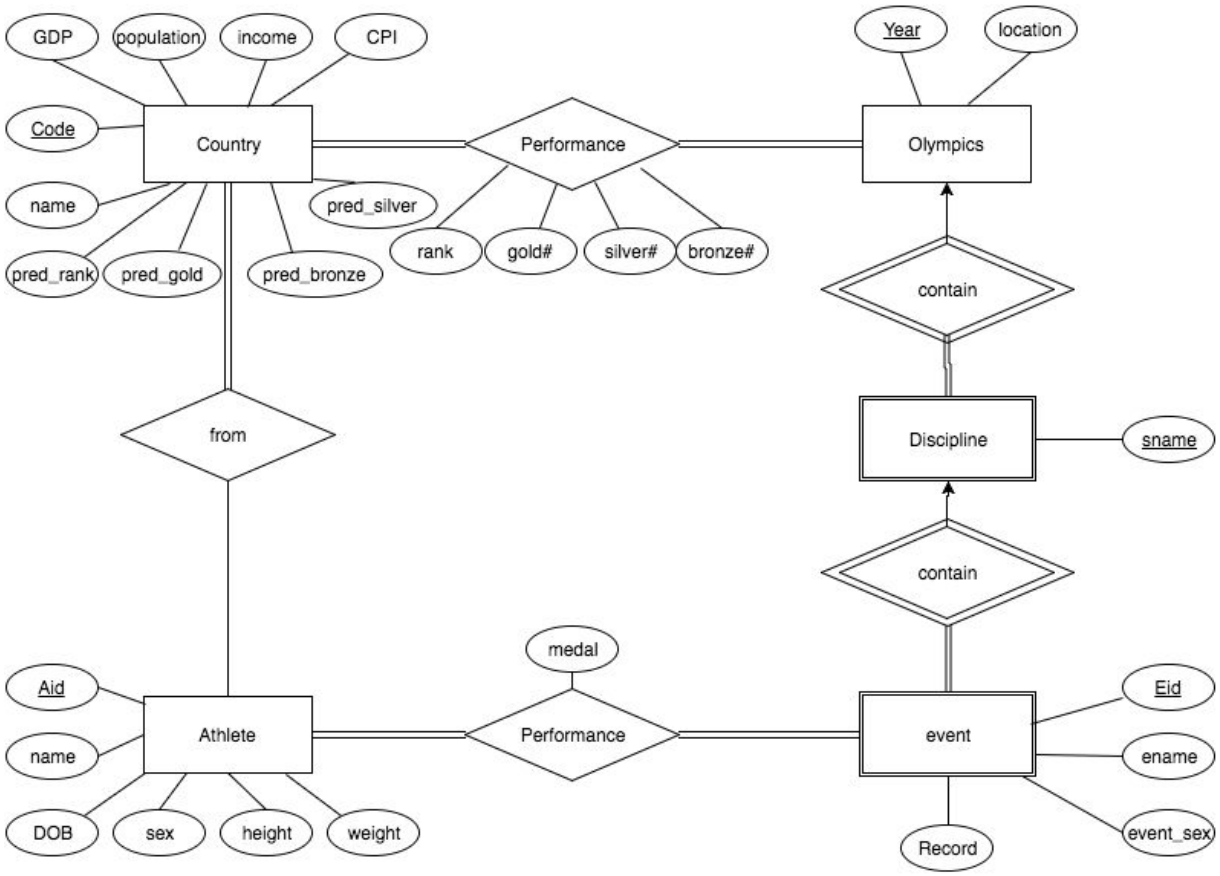
http://fanpagelist.com/category/athletes/olympics/view/list/sort/fans/page1 (olympics athelete's fan page ranking, facebook link, twitter link, # of fans on facebook, # of followers on twitter, total)

Member responsibility for project components

Responsibility:
Data extraction and populating the database (data cleaning, formatting and entity resolution) - Hu Chenchen
Schema design - Xiang Chen
Web design - Zhong Wen
Prediction model - Dhwani

Experimental validation:
Latency - Dhwani, Xiang Chen
Concurrent requests - Hu Chenchen, Zhong Wen

Relational schema



Remarks:
1. sport and discipline are merged together
2. team events are not separated from individual event (we will save all the athletes' name instead of the country name of a football match)

DDL:

```
CREATE TABLE Country(
        code CHAR(3),
        name VARCHAR(50),
        GDP FLOAT,
        population FLOAT,
        income FLOAT,
        CPI FLOAT,
        pred_rank INTEGER,
        pred_gold INTEGER,
        pred_silver INTEGER,
        pred_bronze INTEGER,
        PRIMARY KEY (code)
);

CREATE TABLE Olympics(
        year YEAR,
        location VARCHAR(50),
        PRIMARY KEY (year)
);

CREATE TABLE PerformanceOfCountries(
        code CHAR(3),
        year YEAR,
        rank INTEGER,
        num_of_gold INTEGER,
        num_of_silver INTEGER,
        num_of_bronze INTEGER,
        PRIMARY KEY (code, year),
        FOREIGN KEY (code) REFERENCES Country,
        FOREIGN KEY (year) REFERENCES Olympics
);

CREATE TABLE hasDiscipline(
        year YEAR,
        dname VARCHAR(50),
        PRIMARY KEY (year, dname),
        FOREIGN KEY (year) REFERENCES Olympics
);

CREATE TABLE hasEvents(
        year YEAR,
        dname VARCHAR(50),
        eid INTEGER,
```

```
        ename VARCHAR(50),
        event_gender VARCHAR(1),
        record FLOAT,
        PRIMARY KEY (year, dname, eid),
        FOREIGN KEY (year) REFERENCES Olympics,
        FOREIGN KEY (dname) REFERENCES hasDiscipline,
);

CREATE TABLE Athletes(
        aid INTEGER,
        name VARCHAR(255),
        DOB DATE,
        gender VARCHAR(5),
        height FLOAT,
        weight FLOAT,
        PRIMARY KEY (aid)
);

CREATE TABLE PerformanceOfAthletes(
        year YEAR,
        aid INTEGER,
        medal VARCHAR(6),
        eid INTEGER,
        dname VARCHAR(50),
        PRIMARY KEY (aid, dname, eid, year),
        FOREIGN KEY (aid) REFERENCES Athletes,
        FOREIGN KEY (eid) REFERENCES hasEvents,
        FOREIGN KEY (dname) REFERENCES hasDiscipline,
        FOREIGN KEY (year) REFERENCES Olympics,
        CHECK ( medal IN ('gold', 'silver', 'bronze') )
);

CREATE TABLE Represents(
        aid INTEGER,
        code CHAR(3),
        PRIMARY KEY (aid, code),
        FOREIGN KEY (aid) REFERENCES Athletes,
        FOREIGN KEY (code) REFERENCES Country
);
```

Remarks:

1. do we need to separate location into city and host?
2. Is it better to do prediction by discipline?

3. event name and event ID can be merged together?

NoSQL description

athlete{
DOB,
Name:{first, middle, last,},
Nationality:{current; past},
Sex,
Weight,
Height,
Records: {[Date, location, team/nation, discipline, event, records, medal]}
Rank{[year; rank]}
Social network: {[facebook, twitter, wikilink, fan number, follower number]}
}