**Business Problem: Political Network and Influence Among Graduates of Top US Universities**

1. What are the connections between graduates of Top US Schools and US Politicians ?
   * How is this network affecting US Federal Government in recent elections?
2. Current election season: what is current real time social media sentiment?
3. How much money do graduates of top schools donate and to which politicians/parties?
   * What are the total dollars donated to politicians and how are dollars split by party?

## **Obtaining and Loading Data**

## **Challenge:** How to obtain a complete data set of LittleSis.org

## LittleSis makes its entire database freely available through API accessbile with a user key

* Calls are HTTP calls; limited to 10,000/day

**Plan** : Obtain data via loop calls to API or webcrawler

* Unsure of total DB size/time required

**Resolution**: Reach LittleSis support for direct transfer of complete DB dump

**Challenge**: SQL format of MySQL Littesis datadump incompatible with PostgresSQL.

* How to migrate, transform and load local 4 GB MySQL DB to EC2/Postgres DB

**🡺 Why Postgres as Datasource?**

Postgres is the database of choice for the following reasons:

* Recent coursework featured Postgres (familiarity)
* Open source , non-proprietary
* Simple table structure and types
* Pre-installed and configured on EC2 AMI for w205

**Plan:**

1. generate schema from MySQL
2. export MySQL tables to CSVs
3. migrate CSVs to EC2

**Challenge:** MySQL CSV export format **incompatible** with Postgres

* Entire wikis dedicated to converters; not straightforward process

**Resolution: Discovery of Pentaho converter**

* Pentaho takes a MySQL .sql and automaically generates :

1. Postgres schema and Postgres-compatible csvs
2. Entire Postgres .sql dump file including schema and data

**Resolution: Build LittleSis database stepwise/modularly:**

1) load schema

2) load data/CSVs generated by Pentaho

**Why build step-wise rather than running Postgres .SQL?**

* Devloper can view/modify schema code
* Understand grants in schema
* Modify schema/grants piece-wise while developing

**Working with LittleSis Tables:**

**Challenge:** tables are highly indexed and require multiple, iterative joins to achieve digestible content.

**Resolution:** had to create a number of intermediate tables to study data in a “digestible” format and understand how to satisfy the project proposal. All of these tables are not used by the serving layer but were necessary part of understanding the data locations/flow.

**Challenge**: dealing with big data, there’s an inherent tendency to “Boil the Ocean”:

* LittleSis is a vast trove of a political data / connections; desire was to use as much as possible
* Dilemma - too much interesting data/too little time
* Advanced SQL required: target information required advanced SQL
* Example: (this can be considered for next/future steps):
  + Lobby groups/PACs are not tagged to a political party in LittleSis
  + However, recursive joins of Lobbists/PACs connects may reveal a political association
  + Challenge: number of recursive calls varies; political association not guaranteed

***Endpoint decision to answer project question:***

*Identify all persons (not organizations) classified as ‘elected representative’ or ‘politician’ and query donations made to these people by graduates of Top Schools.*

**Choosing a Serving Layer**

**Why use a REST API as the Serving Layer?**

* REST API lends itself to “walking a graph”of relationships
  + Queries linked to results of previous query
* Similar in concept to following links through web pages
* Similar to LittleSis.org website

**Final Results using REST API request:**

curl http://<hostname>.compute-1.amazonaws.com:8080/topschools

curl http://<hostname>.compute-1.amazonaws.com:8080/donationsummaries/<school or ‘all’>/<year or ‘all’>

curl [http://<hostname>.compute-1.amazonaws.com:8080/donations/<school or all>](http://ec2-54-87-186-104.compute-1.amazonaws.com:8080/donations/%3cschool%20or%20all%3e)

curl [http://<hostname>.compute-1.amazonaws.com:8080/donations/<gradid](http://ec2-54-87-186-104.compute-1.amazonaws.com:8080/donations/%3cgradid)>

curl [http://<hostname>.compute-1.amazonaws.com:8080/connections/<gradid](http://ec2-54-87-186-104.compute-1.amazonaws.com:8080/connections/%3cgradid)>

**Top 10 Schools (can be swapped out when reproduced) :**

Harvard

Yale  
Princeton

UPenn

UChicago

MIT

Duke

Columbia

Cal Tech

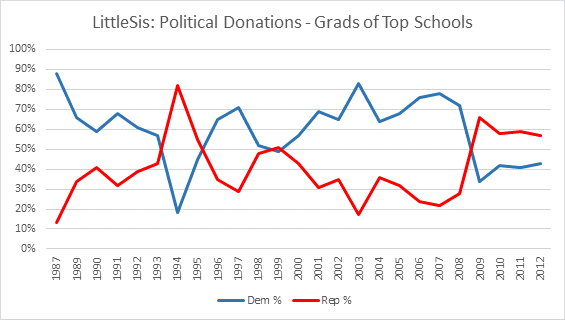
Stanford

**Lessons Learned:**

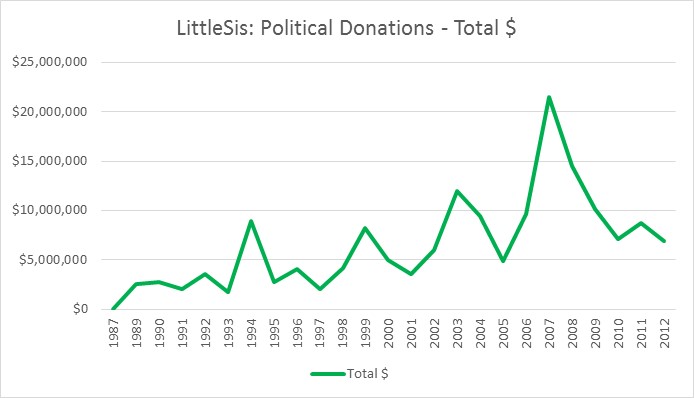
* SQL and CSV formats between databases can be utterly incompatible
* Without 3rd party tool, nearly impossible to migrate
  + Manually changing csv/schema formats for a large DB (even using a shell script) too tedious
* Working with highly indexed data requires an intermediate/personal ‘read’ dictionary
* LittleSis publishes API documentation, but is far from complete; unanswered questions :
* Why do years for donation records cease at 2012?
* What is CRUP field in FEC\_filing table
* If start date/end date spans several years but # of filings == 1, how should donation amount be understood? (To resolve this I only used start date, not end date)
* Why are some elected\_representatives (60 of them) not in the political\_candidate table?
* REST API is an elegant tool for walking a connections graph; difficult to debug via HTTP calls
* Overall LittleSis is fascinating database and a great organization to work with. I look forward to working more with Littlesis.org in the future !!!!!

**Findings and Discussion (p1. of 3):**

Results for all graduates across all top schools:



* 1994 saw a record breaking Republican sweep of federal government positions
* Post 9/11 and entry into the Iraq war, donations trended higher to Democratic candidates /politicians (2001-2008)
* This changed after Obama became president, after which Republican donations surpassed Democratic donations.
* The record-breaking Repbulican sweep of 2010 elections were a direct result of this support.
* Dollar spike in 2007-2008 (seen above) driven by the Obama Presidency campaign.



**Findings and Discussion (p2. of 3):**

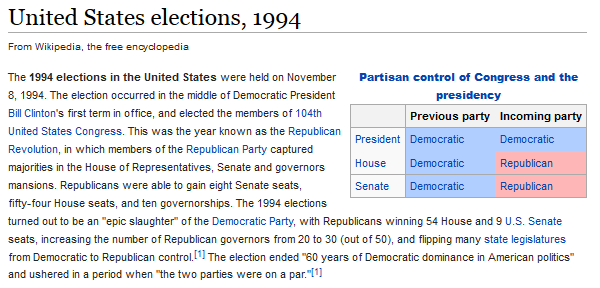
Using the REST API, one can generate all the summary views in the discussion below using the call:

*<hostname>/donationsummaries/<string:topschool>/<string:year>, where year == “all”.*

The REST API can also be used to show top 10 donations for a top school or for a single grad, top 10 connections for a grad, and additional summary data for donation patterns of graduates of top schools along party lines.

Results show that donation patterns at the top schools ***have direct impact on national elections***. While UPenn, UChicago, Princeton, Columbia show consistent democtratic support, other schools see flips between democrat and republican, that mirror the winning parties for those years. In the charts below, there are visible spikes in donations to Republicans during 1994 and 2010 election seasons when republicans had record breaking sweeps in the federal government.

From Wikipedia, record breaking republican election sweeps in 1994 and 2010:





**Findings and Discussion (p3. of 3):**