**Option 1: Crashes in SF** (time / streets)

**The Problem**: Data from prior vehicle crashes in San Francisco can be used to predict when and where crashes are likely to occur in San Francisco and the severity of them based on the location and time of day.

**The Data**: (find the data that pertains to SF only)

* Scrape data to get up-to-date accident reports (article: <http://archives.sfexaminer.com/sanfrancisco/sf-is-receiving-vehicle-collision-data-faster-with-new-database/Content?oid=2650160> ):
  + <http://accidentdatacenter.com/us/california/san-francisco-oakland-san-jose-ca/san-francisco>
  + <http://accidentdatacenter.com/us/california/san-francisco-oakland-san-jose-ca>
* OTS Ratings (2008-2013): <http://www.ots.ca.gov/Media_and_Research/Rankings/default.asp>
* Speed compliance data in SF: <https://data.sfgov.org/Public-Safety/Speed-Limit-Compliance-San-Francisco-CA/9tap-3c96>
* 2011 US Fatal Car Crashes: <https://datamarket.azure.com/dataset/bigml/carcrashusa2011>
* NHTSA (national highway traffic safety administration) 1994-2014: <http://www-fars.nhtsa.dot.gov/Main/index.aspx>
* Crash Stats (2004-2016): <http://www-nrd.nhtsa.dot.gov/cats/listpublications.aspx?Id=F&ShowBy=DocType>

**Hypothesis:** Prior SF crash / accident data (geographic data – when and where there was a crash; severity of the crash) used to predict crashes in SF based on time and location

**Option 2: Salaries; Rent / Mortgage; LA vs. SF**

**The Problem**:

**The Data**:

* SF DATA: <https://data.sfgov.org/>
* LA DATA: <https://data.lacounty.gov/>
* SF Salaries:
  + <https://www.kaggle.com/kaggle/sf-salaries>
  + <https://www.kaggle.com/kaggle/sf-salaries/scripts>
  + <https://data.sfgov.org/data?search=salaries>
    - Salary Ranges by Job Classification: <https://data.sfgov.org/City-Management-and-Ethics/Salary-Ranges-by-Job-Classification/7h4w-reyq>
    - Employee Compensation (City employees 2013): <https://data.sfgov.org/City-Management-and-Ethics/Employee-Compensation/88g8-5mnd>
    - Fulltime Employees by Job Classification: <https://data.sfgov.org/City-Management-and-Ethics/Full-Time-Employees-By-Job-Classification/uk3k-9ai5>
    - Job Titles By Classification: <https://data.sfgov.org/City-Management-and-Ethics/Job-Titles-by-Classification/58un-vqfs>
  + Payscale - Average Salaries in SF by job description: <http://www.payscale.com/research/US/Location=San-Francisco-CA/Salary>
* LA Salaries:
  + LA County Employee Salaries: <https://data.lacounty.gov/Operations/LA-County-Employee-Salaries/8rdv-6nb6>
  + Average Salary and Benefit Earnings by Position: <https://data.lacounty.gov/Operations/Average-Salary-and-Benefit-Earnings-by-Position/6635-a2qt>
  + Payscale - Average Salaries in LA by job description: <http://www.payscale.com/research/US/Location=Los-Angeles-CA/Salary>
* SF Rent Prices:
  + Your Home Suite.com: <https://www.yourhomesuite.com/welcome-to-san-francisco-bay-area?gclid=Cj0KEQjw94-6BRDkk568hcyg3-YBEiQAnmuwknA1EvoOd1bZwabSzxSn7jd-LS8PEInrgWEMQ17_YuUaApcV8P8HAQ>
  + Zumper: <https://www.zumper.com/apartments-for-rent/san-francisco-ca>
  + Rent Jungle: <https://www.rentjungle.com/san-francisco-apartments-and-houses-for-rent/>
* LA Rent Prices:
  + Your Home Suite: <https://www.yourhomesuite.com/furnished-rentals/los-angeles-ca-united-states?data%5Bis_home_page_search%5D=1&data%5Bautocomplete%5D=Los+Angeles%2C+CA%2C+United+States&data%5Blat%5D=34.0522342&data%5Blng%5D=-118.2436849&data%5Bpostal_code%5D=&data%5Bformatted_address%5D=Los+Angeles%2C+CA%2C+USA&data%5Barrival_date%5D=&data%5Bbedrooms%5D>=
  + Zumper: <https://www.zumper.com/apartments-for-rent/los-angeles-ca>
  + Rent Jungle: <https://www.rentjungle.com/los-angeles-apartments-and-houses-for-rent/>
* LA Data:

<http://www.numbeo.com/cost-of-living/compare_cities.jsp?country1=United+States&country2=United+States&city1=Los+Angeles%2C+CA&city2=San+Francisco%2C+CA&tracking=getDispatchComparison>

* SF Data:

<http://www.numbeo.com/cost-of-living/compare_cities.jsp?country1=United+States&country2=United+States&city1=San+Francisco%2C+CA&city2=San+Francisco%2C+CA&tracking=getDispatchComparison>

* SF Mortgages:
* LA Mortgages:
* *Comparing info w/ similar house / apartment size*
* *\*Stick w/ Rent*

**Hypothesis:** Places with higher rents/ mortgages – focusing on San Francisco – offer salaries at a higher rate for the same jobs than in places with lower rents/mortgages

**Option 3:** Infrequently Reported Notifiable Diseases in the US (*possible focus on West Nile*) & Possible Correlations for why they go frequently go unreported

OR Individuals who are more susceptible to certain diseases; where in the US it’s likely to occur; when (month)

**The Problem**: Diseases / infections not properly diagnosed by doctors / rate of people not properly reporting

* Encephalitis

CDC Article re. West Nile in NY 1999: <http://www.cdc.gov/mmwr/pdf/wk/mm4838.pdf>

CDC article re. West Nile in US: <http://www.cdc.gov/westnile/resources/pdfs/wnvguidelines.pdf>

Article 2010 (says places where it’s more likely to occur in US)” <http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6030a1.htm?s_cid=mm6030a1_w>

**The Data**:

* NNDSS – data for notifiable diseases that go frequently unreported 2016: <https://catalog.data.gov/dataset/nndss-table-i-infrequently-reported-notifiable-diseases>
* 2014: <https://catalog.data.gov/dataset/nndss-table-i-infrequently-reported-notifiable-diseases-d8610>
* 2015: <https://catalog.data.gov/dataset/nndss-table-i-infrequently-reported-notifiable-diseases-07903>
  + <http://www.healthdata.gov/dataset/nndss-table-ii-west-nile-virus-disease>
* West Nile Virus Cases (2006-Present): <https://catalog.data.gov/dataset/west-nile-virus-cases-2006-present-1bb7e>
  + Listed West Nile as a notifiable disease that frequently goes un-reported
* CDC: US Mortality Cases 1999-2009: <https://catalog.data.gov/dataset/cdc-wonder-mortality-multiple-cause-of-death>
* CDC: US Mortality Causes of Death 1979-present: <https://catalog.data.gov/dataset/cdc-wonder-mortality-underlying-cause-of-death>
* Infectious Disease Cases in CA 2001-2014: <https://catalog.data.gov/dataset/infectious-disease-cases-by-county-year-and-sex-2001-2014>
* West Nile Virus – mosquito tests: <http://www.healthdata.gov/dataset/west-nile-virus-wnv-mosquito-test-results>
* COMPARE TO POSSIBLE FACTORS: <http://wwwn.cdc.gov/nchs/nhanes/search/default.aspx>

**Hypothesis:**

SF Giants – chances of strike outs/home runs…etc.

* ***The Problem***: What's the background and scope of the project idea? What problem are you attempting to address or solve? Who may it matter to?
* ***The Data***: What data exists to help solve this problem? Where is it coming from? What does the data look like? What is the observation?

***Your Hypotheses***: Given the problem and data you're aware of, what do you believe is the solution? What does success look like?

assumptions about the data would be about their type and distribution,

\*CORRELATION W/ THE OUTCOME = HYPOTHESIS

* High level data points can help show the "improvement" scope that's available.

Cite similar/previous work that will better inform your strategy that can be taken.

Examples:

### EX: Read and evaluate the following problem statement:

Using Planet Express customer data from January 3001-3005, determine how likely previous customers are to request a repeat delivery using demographic information (profession, company size, location) and previous delivery data (days since last delivery, number of total deliveries).

#### 1. What is the outcome?

Answer: Return customer indicator (yes/no).

#### 2. What are the predictors/covariates?

Answer: Age, gender, location, date of first deliveries, profession, days since last delivery, and number of total deliveries.

#### 3. What timeframe is this data relevent for?

Answer: January 3001-3005.

#### 4. What is the hypothesis?

Answer: Demographic and previous delivery information will allow us to predict if a customer will be a repeat customer.

# EX:

### ****Read & Evaluate the following problem statement:****

Determine which free-tier customers will covert to paying customers, using demographic data collected at signup (age, gender, location, and profession) and customer useage data (days since last log in, and activity score 1 = active user, 0 = inactive user) based on Hooli data from January - April 2015.

#### **1. What is the outcome?**

Answer: Return free-tier customer indicator (yes/no)

#### **2. What are the predictors/covariates?**

Answer: age, gender, location, profession, days since last log in, activity score (active vs. inactive user)

#### **3. What timeframe is this data relevent for?**

Answer: January - April 2015

#### **4. What is the hypothesis?**

Answer: demographic and customer usage data will allow us to predict which free-tier customers will convert to paying customers

#### **2. What is the outcome?**

Answer: return applicants acceptance into UCLA graduate program (y/n)

#### **3. What are the predictors/covariates?**

Answer: admittance (y/n), GRE score, GPA, prestige (1-4)

#### **4. What timeframe is this data relevent for?**

Answer: UCLA graduate school applicants

#### **4. What is the hypothesis?**

Answer: Using previous UCLA applicant data (admittance (1=y / 0=n), GRE score, GPA, and prestige), one can predict if an applicant will be accepted into UCLA's graduate program.

Using the above information, write a well-formed problem statement.

Using UCLA graduate application data from previous applicants, one can determine the factors that influence UCLA admission rate and predict if a student will be accepted or not based on their GRE score, GPA and level of prestige.