Big Data Project

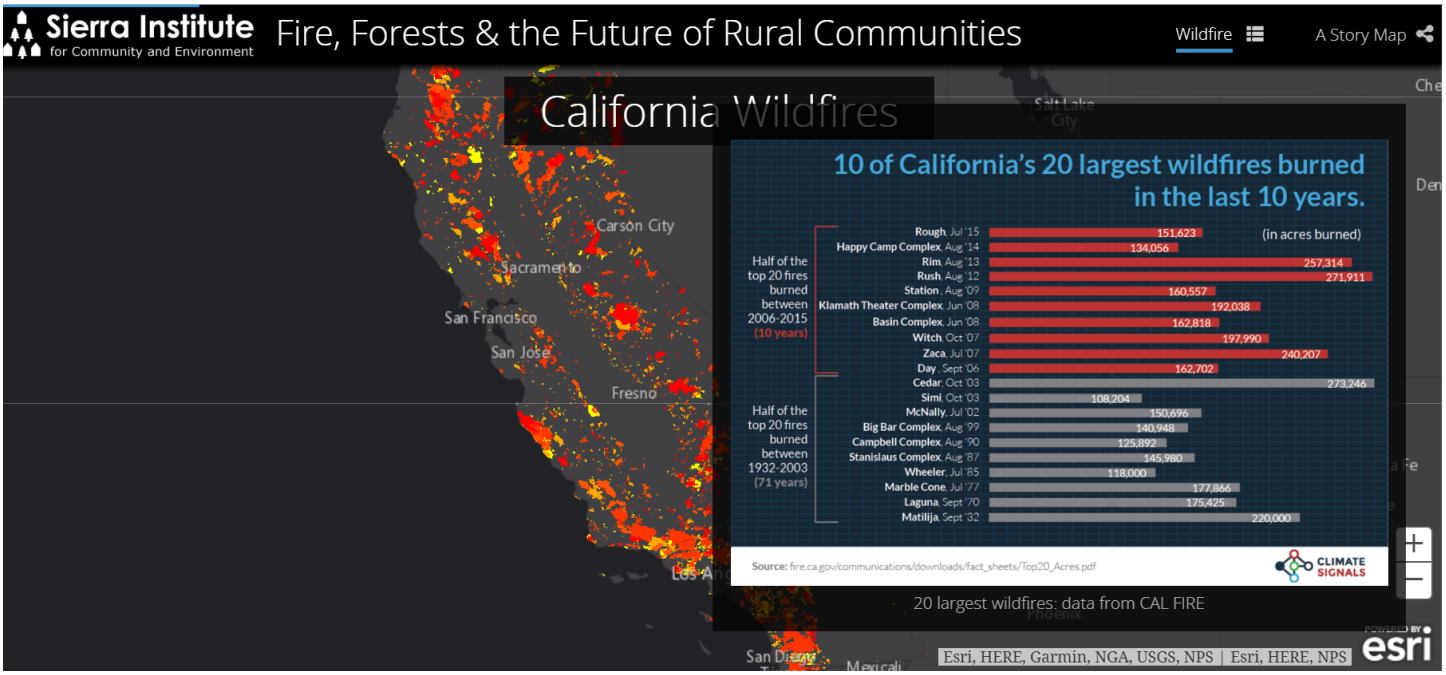
Objective: To collaborate with APCS Principles students, and to computationally analyze big data sets. The goal is to detect patterns, associations, and/or trends and present them in ways that people understand.

Background:
The term "big data" refers to data sets too large or complex for traditional data-processing applications. The collection of data has increased year-over-year into an explosion of storage and management. Over 2.5 quintillion bytes (2.5×10^{18} bytes) of data are created and stored each day (Fortune magazine, May 2018). According to a Reuters report, "big data" is growing 45% year-over-year. Who is storing and using this data?

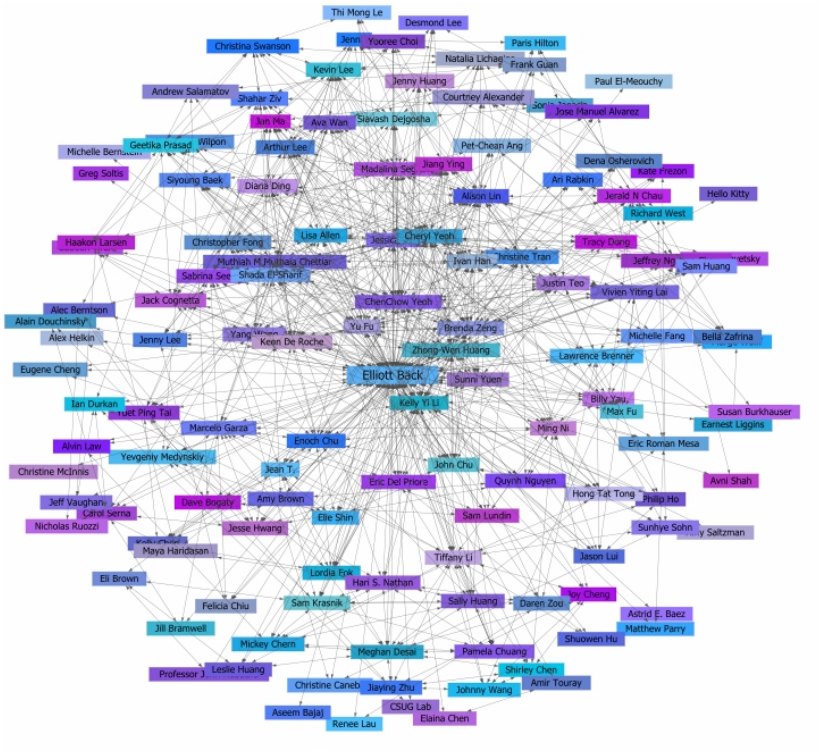
The two biggest collectors of data are government and business. The US Government collects huge volumes of data for a variety of useful purposes. For example, the Center for Disease Control (CDC) collects data on the spread of disease, the US Census Bureau tracks geographic population statistics, and the National Oceanic and Atmospheric Administration (NOAA) tracks weather patterns. Businesses collect data to market and sell services and products, or to improve efficiencies. For example, Amazon tracks customer searches to increase sales, Facebook collects personal information to sell to marketeers, and Pacific Gas and Electric (PG&E) monitors customer meters for electricity usage.

Add to this the explosive number of Internet Of Things (IoT). This is where simple, ubiquitous devices, like thermostats, security cameras, and wearable devices, collect and store data on the Internet. Once on the Internet, the data can be shared with people, businesses, government agencies, and other IoT devices to improve lives and make money.

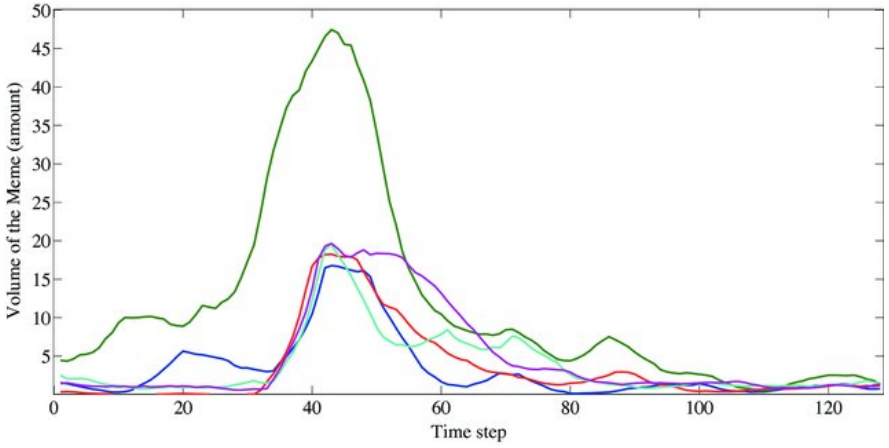
The most meaningful part of all this data is to display patterns, associations, and trends in a way people can understand. One method is the "story map" which uses a series of graphics to convey a story or to build an argument. An example of a story map is Sierra Institute's web site (below) on fire dangers in California. (https://www.arcgis.com/apps/Cascade/index.html?appid=2aa4e916e76943e2b5d84bd791be8fea)

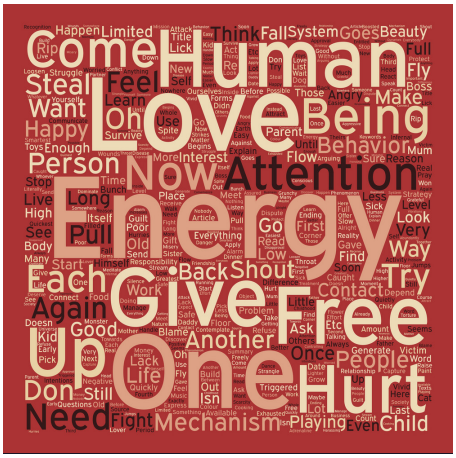


Another method is the "network map" which displays the interconnectivity and relationship between items. For example, the graphic below shows the relationships between people on a social network.



The graph below left is a time plot of social memes. The graph below right is big text analysis.





In this project, you will be provided some big data sets that have factual data. This work will be broken down into steps including:

- Preparing a set of data for analysis by parsing
- Preparing a set of data for analysis by building a representation
- Designing algorithm(s) to synthesize meanings from large data sets

Some outputs of this work, in order to answer questions, will be:

- Geographic "story" maps
- Large text analysis displays (aka word clouds and chi-squared analysis of words of interest)
- Meme / Social Data displays (network maps, time plots, or geographic maps)
- Statistical analysis of data sets

Collaboration:
APCS A and APCS Principles students will be working together in teams to accomplish this project. There will be a division of labor according to the following guidelines.

APCS Principles students will be in charge of:

(a) A **need statement** defining the targeted question(s) of this study. This should go with hypotheses (null and alternate) that occur for each smaller piece of analysis (see below). Strong need statements are insightful, not obvious, so may include citation with respect to background research.

(b) Specifying one or a maximum of three **additional data sets** beyond the three data sets provided (and including citation to access this data in Excel, Google Sheets, or CSV format). These should be at the same geographic scale as provided in the three aforementioned sets and directly tied to the need statement AND require use of the other three data sets.

(c) Providing **specifications for one or more representations** needed to address questions raised by the need statement. This should go with a loose flow diagram for an algorithm that would manage data to build this new representation.

(d) **Hypothesis and null hypothesis** for each data analysis conducted, including work done by the APCSA contributor. This should guide data inclusion in the representation (c, above) and algorithm design.

(e) **Geographic "story" map**, which can give a wide granularity perspective about two or more variables (more than one map is an option) that could be explored in more detail/from a modified perspective elsewhere.

(f) **Large text data analysis**, which can give a wide granularity perspective about two or more variables (more than one pair of word clouds, chi-squared analysis is an option), what could be explored in more detail/from a modified perspective elsewhere. Alternatively, network, time plot, or geographic maps of memes could be used.

(g) Completing **statistical analysis** on large data sets, based on new representations and algorithm work. For instance, if the representations of "book worms" and "sporty folks" are complete, one could acquire mean hours spent outside, number of friends, and so on, using a chi-squared analysis. This work could also be built into the algorithm(s) if there is time, so this work can be delegated as group dynamics permit.

APCS A students will be in charge of:

(a) **Parsing and merging data sets** to allow them to be processed to build "story" maps, used to build representation(s), and used during mathematical processing.

(b) **Building representation(s)** meeting specs by the team.

(c) **Designing and building algorithm(s)** meeting specs by the team.

(d) Meeting regularly with team members and making sure those **meetings are recorded in Google Docs**, with all docs related to the product hyperlinked to those meeting "minutes."

(e) **Build graphical output** after analysis, if team wants a different output than what APCS Principle classes use.

Regular meetings to check on progress with the project. All meetings will be documented through Google Docs and all produced documents hyperlinked in these "minutes." The week prior to the first meeting, the teacher will provide this document and orient the students with looking at the data sets to get a feel for what they contain and (for APCS Principles) generating need statements. Here is a loose framework for the meetings:

March 12: Team introductions, need statement presented. APCS A assigned with parsing three data sets. APCS Principles will work on finding another data set, drafting specs and hypotheses.

March 15: Check in about process. Parsing the three sets should be well on its way to being done. New data set(s) shared. Discussion about specs and hypotheses (drafts revised). APCS Principles will finalize specs and hypotheses. APCS A will finish parsing three sets and start working with new set(s).

It's expected for groups to be sharing information back and forth between March 15 & 22.

March 22: Check in about process. 3 parsed sets should be available to all, now. Final specs and hypotheses shared. APCS Principles will share plans/results for work: either (e) and (f). APCS A team members will finish parsing new set(s) and begin working to build new representation and algorithm.

Last week of March: Check in about process. Both teams will share remaining work and determine who/how to complete (g).

April 5: Last meeting - everything should be finished.