CS 458/658: Introduction to Data Mining

Data Mining Course Project

Instructor: Lei Yang

Department of Computer Science and Engineering, UNR – Fall 2019

PE 105, TuTh 12:00PM – 1:15PM

Project

- **One key goal** of this course is to take advantage of your intelligence and (limited) experience (so you're audacious and creative) to expand your knowledge in creating something useful and interesting
- Group project
  - Groups (3 students per group)
    - 26 undergraduates
    - 10 graduates
    - **Email me your group members by 9/24; otherwise I will randomly put your name in a group.**
  - You can apply whatever techniques you learnt from data mining course and other sources

Tasks

- Tasks
  - Task 1: Document Classification
  - Task 2: Exploring Environmental Data at NRDC
  - Task 3: Exploring Used Auto Purchase Dataset
- Bonus task (10%)
  - Task 4: Wildfire Smoke Detection

Undergraduates

- Tasks:
  - Task 1
  - Pick one of the following:
    - Task 2
    - Task 3

Graduates

- Tasks
  - Task 1
  - Task 2
  - Task 3

Evaluation

- Final report (due **Dec 12, 2019** in **Webcampus**) (**35%**)
  - Each member need to submit your own report and indicate your contribution in %
- Class presentation and/or demo (**5%**)
  - Each group will present their work. Each member needs to present.
  - Your presentation will be evaluated by the other groups using an evaluation form.
  - Each presentation is 20 mins with 5 mins for Q&A.
    - Nov. 26, 2019
    - Dec. 3, 2019
    - Dec. 5, 2019
    - Dec. 10, 2019

Task Description

Task 1: Classification

- Provided data
  - The training set and its label information
  - The testing set
- Hidden data
  - The label information of the testing data
  - The data will be used for the purpose of evaluation

Data Format

- The training set
  - training.txt
  - The first column is the information ID
  - The second column is the feature ID
  - The third column is the value of the feature
  - The default values of features are zeros

```
1 16 1
1 23 4
1 27 1
1 29 8
1 30 2
1 33 3
1 42 1
1 54 1
1 72 1
1 81 1
```

Data Format

- The label information of the training set
  - label_training.txt
  - Each row represents a data point in the training set
  - 1 is true information while -1 is misinformation

```
1
-1
1
-1
1
1
1
-1
-1
```

Data Format

- The testing set
  - testing.txt
  - It has the same format as the training set

```
1 16 1
1 23 1
1 27 1
1 29 2
1 50 1
1 245 1
1 340 1
1 388 1
1 589 1
1 638 1
1 764 1
1 902 1
1 905 1
1 2774 1
1 8066 1
1 10762 2
```

Model Challenge from Model Selection

- There are so many classifiers
  - Which one is better?

- There may be parameters in classifiers
  - How to determine the optimal values?

Evaluation

- Classification accuracy will be used to evaluate the quality of the predicted labels

  - Comparing the hidden labels with your predicted labels

- Your final grades will strongly depend on the rankings of the quality of the predicted labels you provide

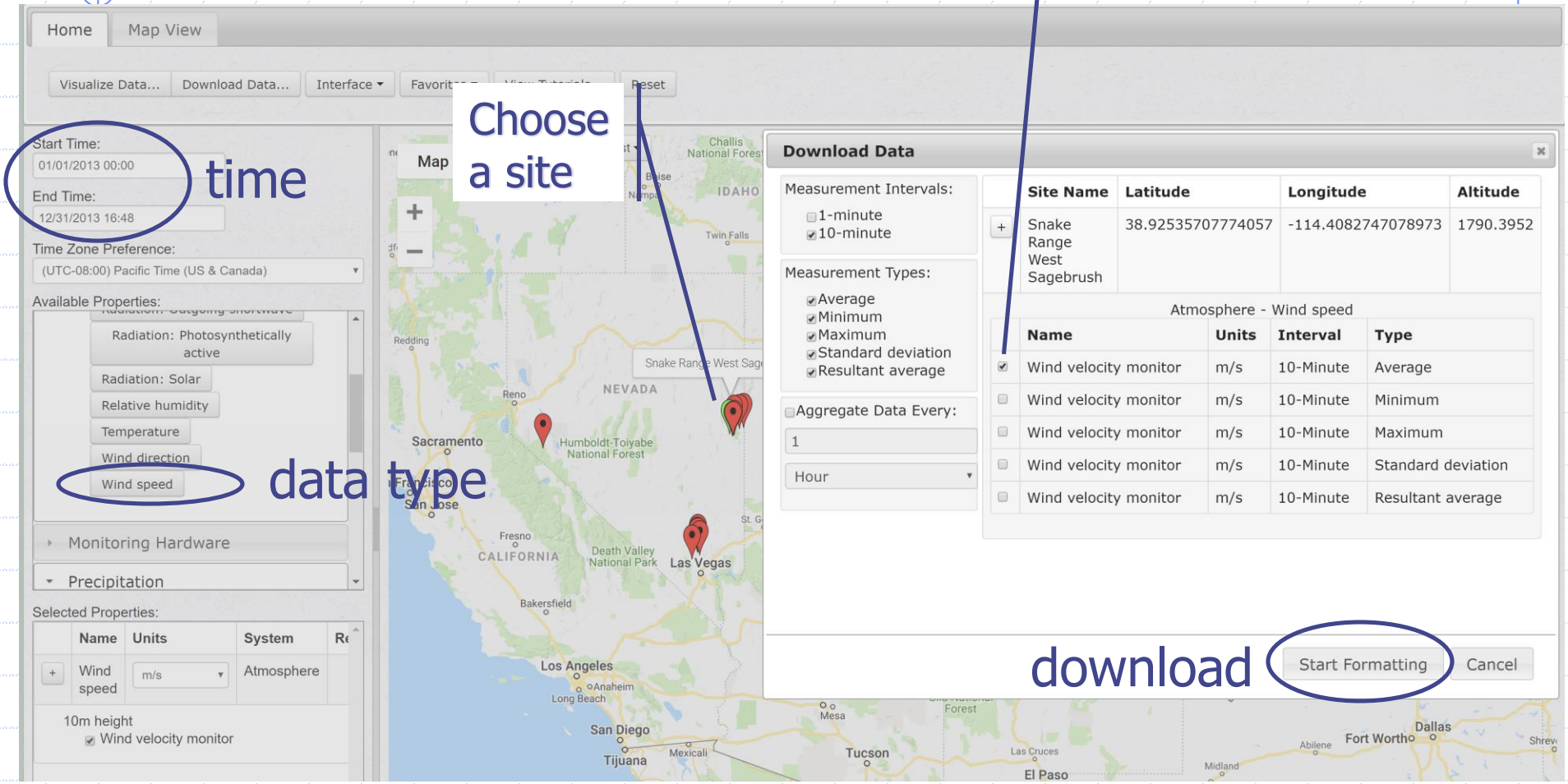
Task 2: Exploring Environmental Data at NRDC

Sample project

- Wind speed prediction
  - Predict the quantitative wind speed at different sites using the historical information in the same sites as well as data in other neighbor sites.
- The data is available from http://sensor.nevada.edu/SENSORDataSearch/
  - Snake Range East Sagebrush (EB)
  - Snake Range East Subalpine (EA)
  - Snake Range East Salt Desert Shrub (ED)
  - Snake Range West Subalpine (WA)
  - Snake Range West Montane(WM)
  - Snake Range West Sagebrush (WB)
  - Snake Range West Pinyon-Juniper(WP)

10 minutes wind data

Download data



Example data

```
Site Name:,Snake Range West Sagebrush
Deployment:,Wind velocity monitor
Monitored System:,Atmosphere
Measured Property:,Wind speed
Vertical Offset from Surface:,10m height
Units:,m/s
Measurement Type:,Average
Measurement interval:,00:10:00
Time Stamp ((UTC-08:00) Pacific Time (US & Canada))
1/1/2013 12:00:00 AM,0.5136489600000000
1/1/2013 12:10:00 AM,0.073761600000000000
1/1/2013 12:20:00 AM,0.348691200000000000
1/1/2013 12:30:00 AM,0.291023040000000000
1/1/2013 12:40:00 AM,0.476544640000000000
1/1/2013 12:50:00 AM,0.435864000000000000
1/1/2013 1:00:00 AM,1.001369600000000000
1/1/2013 1:10:00 AM,0.899444480000000000
1/1/2013 1:20:00 AM,0.206979520000000000
1/1/2013 1:30:00 AM,0.604398080000000000
1/1/2013 1:40:00 AM,0.710793600000000000
1/1/2013 1:50:00 AM,0.430052480000000000
1/1/2013 2:00:00 AM,0.198485760000000000
1/1/2013 2:10:00 AM,0.175239680000000000
1/1/2013 2:20:00 AM,0.598139520000000000
1/1/2013 2:30:00 AM,1.322791360000000000
1/1/2013 2:40:00 AM,0.473415360000000000
1/1/2013 2:50:00 AM,0.105948480000000000
1/1/2013 3:00:00 AM,0.760862080000000000
```

Challenges

- Data preprocessing
  - Missing data
- What methods to use?
  for regression?
- How to tune parameters?
- ...

Evaluation

- For each site, you need to provide prediction accuracy of your proposed approach based on the following measure
  - Mean absolute error (MAE)

$$MAE = \frac{1}{number\ of\ points} \sum |forcast - actual|$$

  - Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{number\ of\ points} \sum |forcast - actual|^2}$$

- Compare your approach with the following Benchmark:
  - Persistent forecast: predicted_wind(t)=actual_wind(t-1)

Task 3: Exploring Used Auto Purchase Dataset

Exploring Used Auto Purchase Dataset (1)

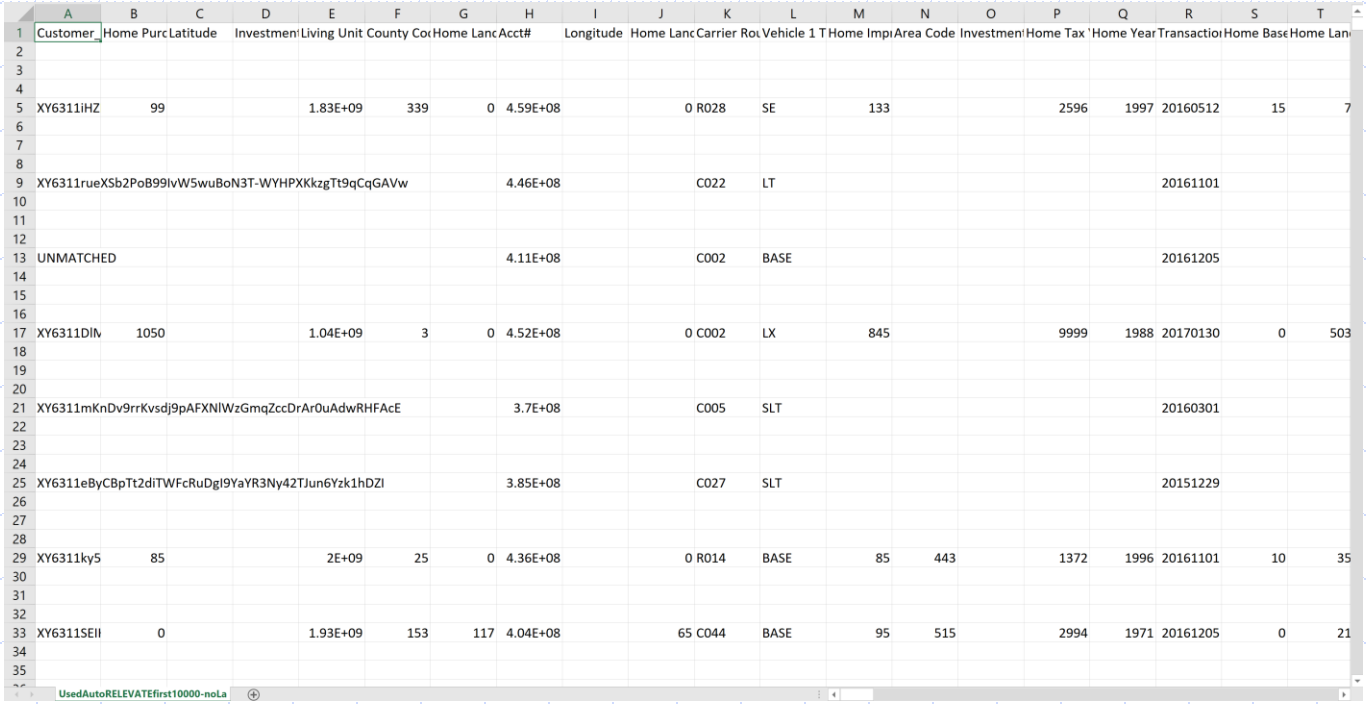
- Dataset: the set of all used auto purchases for the past 5 years in US
  - Number of attributes: 280
    - Vehicle Info (Model, Engine, Drive Type) Home Info (Purchase Price/Date, Value, Year), Address (State, County), Loan Info (Monthly Mortgage), Demographic (Ethnic, Number of Children), Behavior Info (Investment, Interest in Travel/Reading, Presence of Premium Credit Card) ……
    - No Personally identifiable information

Acknowledgement: Thanks Marketing Evolution for sharing this dataset.

A small subset is used in this course project

Dataset(UsedAutoRELEVATEfirst10000-noLatLong.csv)

	Attribute 1	Attribute 2	…
Data entry 1			
Data entry 2			
……			

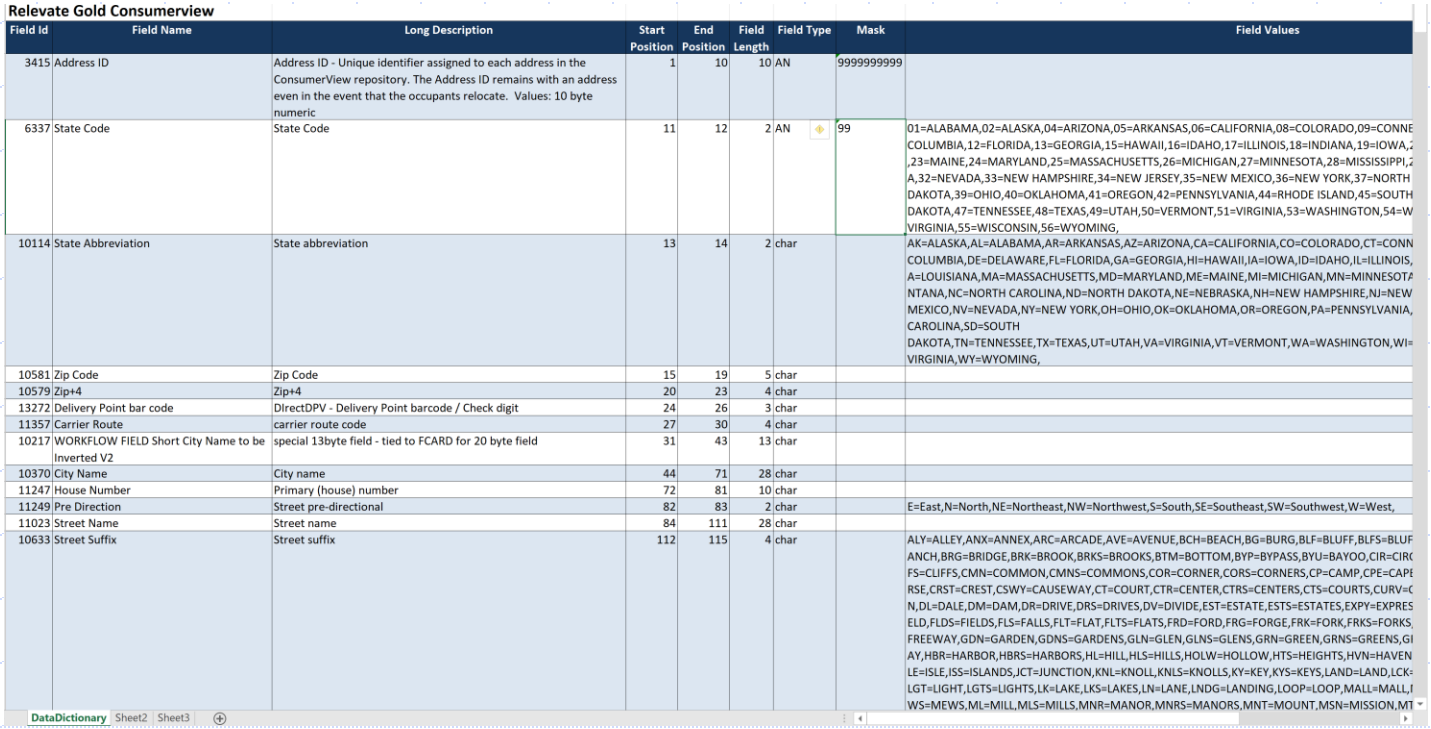


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Customer_	Home Purc	Latitude	Investmen	Living Unit	County Co	Home Lanc	Acct#	Longitude	Home Lanc	Carrier Ro	Vehicle 1 T	Home Impı	Area Code	Investmen	Home Tax	Home Year	Transactio	Home Base	Home Lan
2																				
3																				
4																				
5	XY6311iHZ	99			1.83E+09	339	0	4.59E+08		0	R028	SE	133			2596	1997	20160512	15	7
6																				
7																				
8																				
9	XY6311rueXSb2PoB99IvW5wuBoN3T-WYHPXKkzgTt9qCqGAVw							4.46E+08			C022	LT						20161101		
10																				
11																				
12																				
13	UNMATCHED							4.11E+08			C002	BASE						20161205		
14																				
15																				
16																				
17	XY6311DlN	1050			1.04E+09	3	0	4.52E+08		0	C002	LX	845			9999	1988	20170130	0	503
18																				
19																				
20																				
21	XY6311mKnDv9rrKvsdj9pAFXNlWzGmqZccDrAr0uAdwRHFAcE							3.7E+08			C005	SLT						20160301		
22																				
23																				
24																				
25	XY6311eByCBpTt2diTWFcRuDgI9YaYR3Ny42TJun6Yzk1hDZI							3.85E+08			C027	SLT						20151229		
26																				
27																				
28																				
29	XY6311ky5	85			2E+09	25	0	4.36E+08		0	R014	BASE	85	443		1372	1996	20161101	10	35
30																				
31																				
32																				
33	XY6311SEII	0			1.93E+09	153	117	4.04E+08		65	C044	BASE	95	515		2994	1971	20161205	0	21
34																				
35																				

UsedAutoRELEVATEfirst10000-noLa

Data Dictionaries (EXP REL Custom.xls)

ID	Field Name	Description	…
Data entry 1			
Data entry 2			
……			



Relevate Gold Consumerview

Field Id	Field Name	Long Description	Start Position	End Position	Field Length	Field Type	Mask	Field Values
3415	Address ID	Address ID - Unique identifier assigned to each address in the ConsumerView repository. The Address ID remains with an address even in the event that the occupants relocate. Values: 10 byte numeric	1	10	10	AN	9999999999	
6337	State Code	State Code	11	12	2	AN	99	01=ALABAMA,02=ALASKA,04=ARIZONA,05=ARKANSAS,06=CALIFORNIA,08=COLORADO,09=CONNE COLUMBIA,12=FLORIDA,13=GEORGIA,15=HAWAII,16=IDAHO,17=ILLINOIS,18=INDIANA,19=IOWA,2 ,23=MAINE,24=MARYLAND,25=MASSACHUSETTS,26=MICHIGAN,27=MINNESOTA,28=MISSISSIPPI,2 A,32=NEVADA,33=NEW HAMPSHIRE,34=NEW JERSEY,35=NEW MEXICO,36=NEW YORK,37=NORTH DAKOTA,39=OHIO,40=OKLAHOMA,41=OREGON,42=PENNSYLVANIA,44=RHODE ISLAND,45=SOUTH DAKOTA,47=TENNESSEE,48=TEXAS,49=UTAH,50=VERMONT,51=VIRGINIA,53=WASHINGTON,54=W VIRGINIA,55=WISCONSIN,56=WYOMING,
10114	State Abbreviation	State abbreviation	13	14	2	char		AK=ALASKA,AL=ALABAMA,AR=ARKANSAS,AZ=ARIZONA,CA=CALIFORNIA,CO=COLORADO,CT=CONN COLUMBIA,DE=DELAWARE,FL=FLORIDA,GA=GEORGIA,HI=HAWAII,IA=IOWA,ID=IDAHO,IL=ILLINOIS, A=LOUISIANA,MA=MASSACHUSETTS,MD=MARYLAND,ME=MAINE,MI=MICHIGAN,MN=MINNESOTA NTANA,NC=NORTH CAROLINA,ND=NORTH DAKOTA,NE=NEBRASKA,NH=NEW HAMPSHIRE,NJ=NEW MEXICO,NV=NEVADA,NY=NEW YORK,OH=OHIO,OK=OKLAHOMA,OR=OREGON,PA=PENNSYLVANIA, CAROLINA,SD=SOUTH DAKOTA,TN=TENNESSEE,TX=TEXAS,UT=UTAH,VA=VIRGINIA,VT=VERMONT,WA=WASHINGTON,WI= VIRGINIA,WY=WYOMING,
10581	Zip Code	Zip Code	15	19	5	char		
10579	Zip+4	Zip+4	20	23	4	char		
13272	Delivery Point bar code	DirectDPV - Delivery Point barcode / Check digit	24	26	3	char		
11357	Carrier Route	carrier route code	27	30	4	char		
10217	WORKFLOW FIELD Short City Name to be Inverted V2	special 13byte field - tied to FCARD for 20 byte field	31	43	13	char		
10370	City Name	City name	44	71	28	char		
11247	House Number	Primary (house) number	72	81	10	char		
11249	Pre Direction	Street pre-directional	82	83	2	char		E=East,N=North,NE=Northeast,NW=Northwest,S=South,SE=Southeast,SW=Southwest,W=West,
11023	Street Name	Street name	84	111	28	char		
10633	Street Suffix	Street suffix	112	115	4	char		ALY=ALLEY,ANX=ANNEX,ARC=ARCADE,AVE=AVENUE,BCH=BEACH,BG=BURG,BLF=BLUFF,BLFS=BLUF ANCH,BRG=BRIDGE,BRK=BROOK,BRKS=BROOKS,BTM=BOTTOM,BYP=BYPASS,BYU=BAYOO,CIR=CIR FS=CLIFFS,CMN=COMMON,CMNS=COMMONS,COR=CORNER,CORS=CORNERS,CP=CAMP,CPE=CAPE RSE,CRST=CREST,CSWY=CAUSEWAY,CT=COURT,CTR=CENTER,CTRS=CENTERS,CTS=COURTS,CURV=C N,DL=DALE,DM=DAM,DR=DRIVE,DRS=DRIVES,DV=DIVIDE,EST=ESTATE,ESTS=ESTATES,EXPY=EXPRES ELD,FLDS=FIELDS,FLS=FALLS,FLT=FLAT,FLTS=FLATS,FRD=FORD,FRG=FORGE,FRK=FORK,FRKS=FORKS, FREEWAY,GDN=GARDEN,GDNS=GARDENS,GLN=GLEN,GLNS=GLENS,GRN=GREEN,GRNS=GREENS,GI AY,HBR=HARBOR,HBRS=HARBORS,HL=HILL,HLS=HILLS,HOLW=HOLLOW,HTS=HEIGHTS,HVN=HAVEN LE=ISLE,ISS=ISLANDS,JCT=JUNCTION,KNL=KNOLL,KNLS=KNOLLS,KY=KEY,KYS=KEYS,LAND=LAND,LCK= LGT=LIGHT,LGTS=LIGHTS,LK=LAKE,LKS=LAKES,LN=LANE,LNDG=LANDING,LOOP=LOOP,MALL=MALL,I WS=MEWS,ML=MILL,MLS=MILLS,MNR=MANOR,MNRS=MANORS,MNT=MOUNT,MSN=MISSION,MT

DataDictionary | Sheet2 | Sheet3

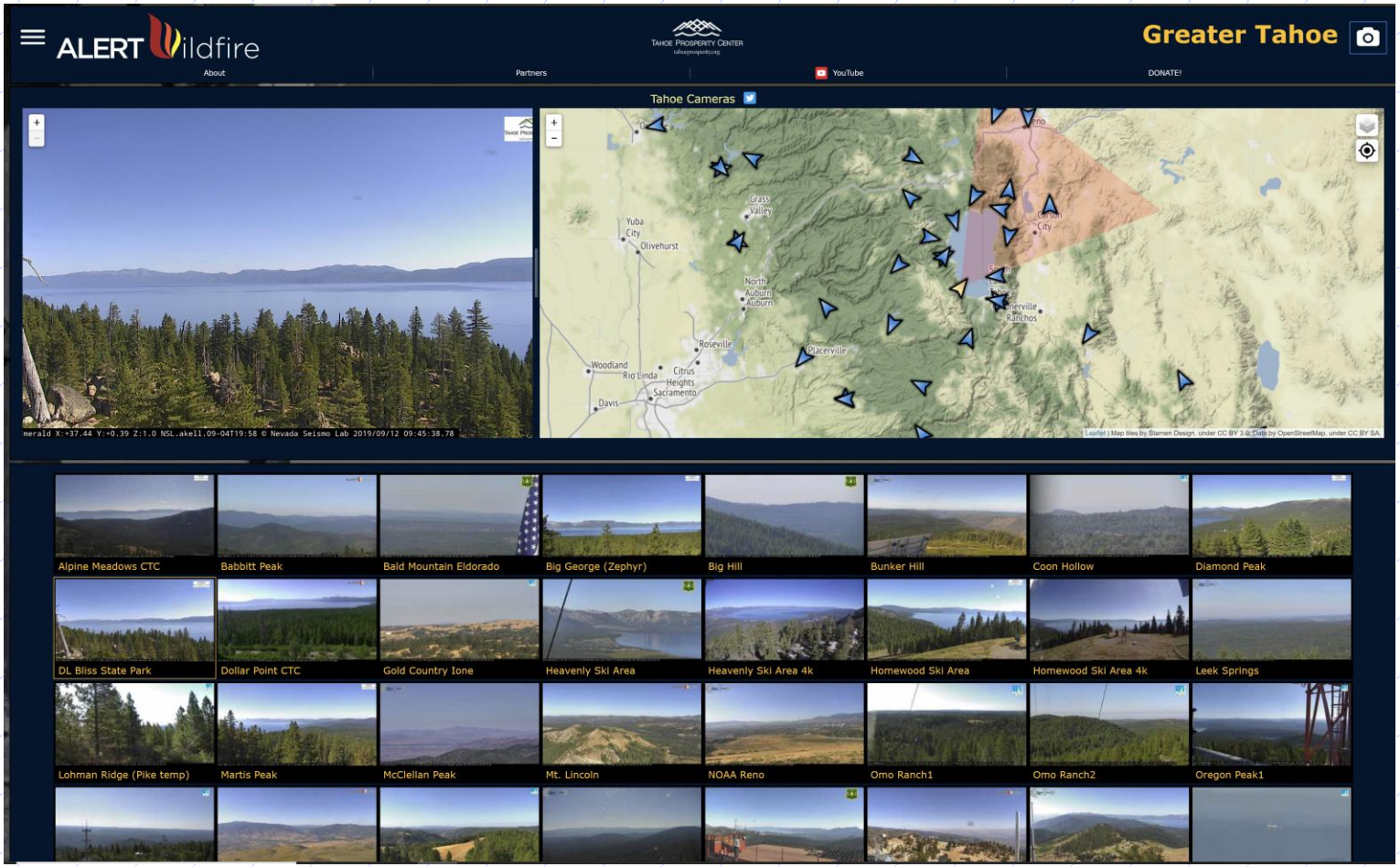
Exploring Used Auto Purchase Dataset (2)

- Project description:
  - Selection: Due to the size/heterogeneity of the original data, we need to select a target data.
  - Preprocessing: Data exist in many types (continuous, nominal) and forms, and may have missing values.
  - Transformation: To better extract useful patterns from dataset.
  - Data mining: Explore different data mining algorithms
  - Interpretation/Evaluation

- Goal: extraction of useful patterns from dataset
  - What car type will be purchased, given customer's info?
  - What customer's type, given a car purchased?
  - How to divide a market into distinct subsets of customers?

Task 4: Wildfire Smoke Detection

ALERTWildfire



http://www.alertwildfire.org/blmnv/index.html

Wildfire Smoke Dataset in NevadaBox

https://nevada.box.com/s/zh4zpwzxpvg5lftqvqsv3oc42gkc6ulo

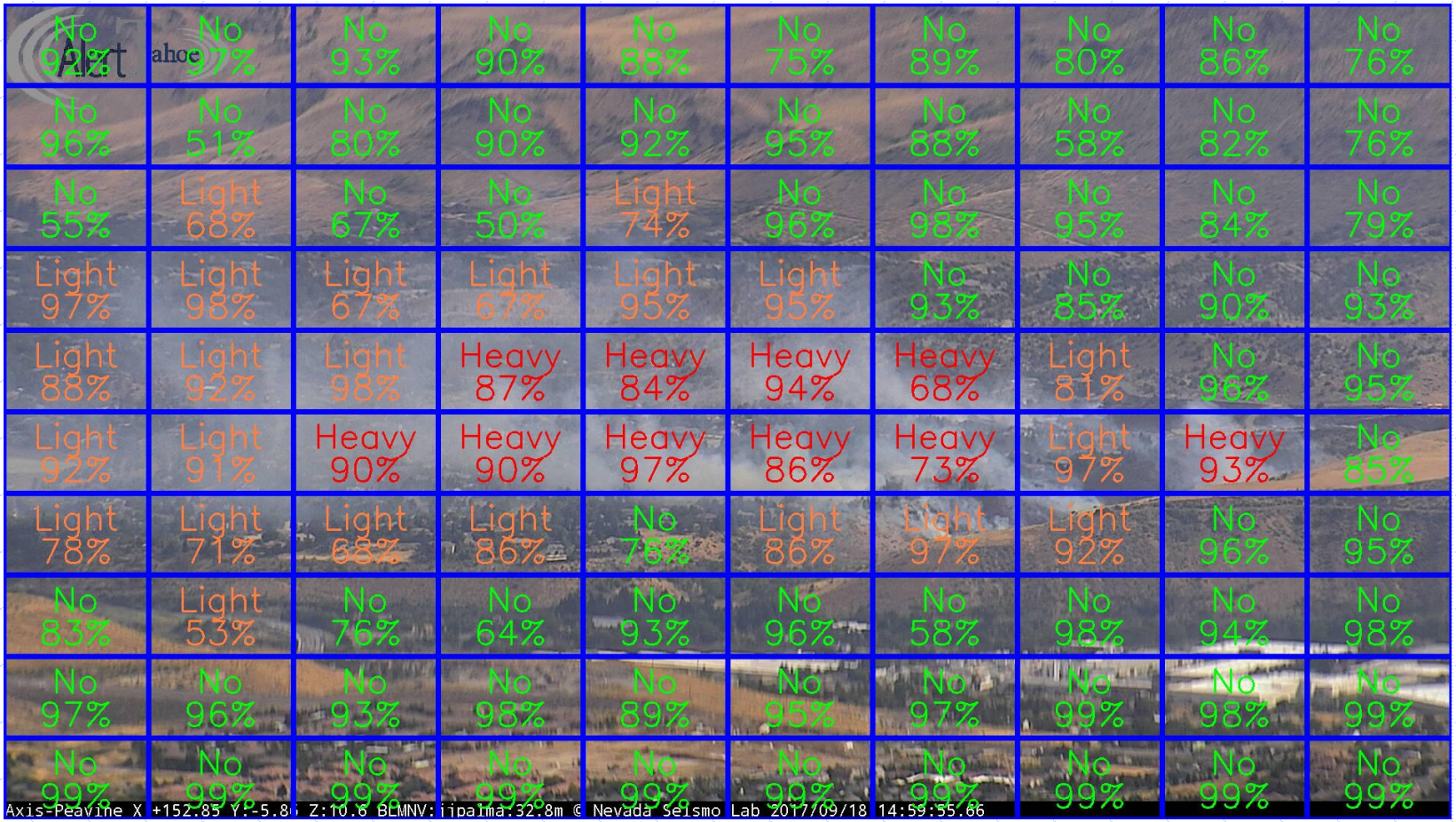
- 962 video files
- Each video is about 1 minute

- Acknowledgement: the dataset is from ALERTWildfire

Wildfire Smoke Detection

- Project description
  - Data Preprocessing: Label the data!
  - Data mining: Explore different data mining algorithms
    - Deep learning
  - Challenges:
    - weather and night conditions
      - E.g., Cloud looks very similar to smoke in images
- Goal
  - Develop a model that can accurately detect wildfire smoke under different conditions
  - Estimate the smoke density

Smoke Detection Example



Deliverables

- Datasets with labels
  - For each video, extract each frame of the video and provide the corresponding labels
    - E.g., you can submit a csv file
- Codes for smoke detection

Project Report

Report

- All the reports should be in the form of python notebook, i.e., report.ipynb
- All your codes should be runnable in python notebook in Google Colab

Report Format

- Cover Section
  - Team members and their contribution in %
- Introduction
- Literature review for each task
- Task 1
  - Your approach (e.g., Preprocessing, Model selection, Parameter selection, Your solution)
  - Your conclusion
- Task 2
- Task 3
- Task 4
- List of documents/codes you submitted

Report requirements

- The report should be as concise as possible while providing all necessary information required to replicate your plots.
- In literature review, you need to show your understanding of the literature by reading and comparing the existing work.
  - Cite your references properly. You can use google scholar to download citation.
- Your submitted code should have proper comments.