Machine Estimation of Exposure

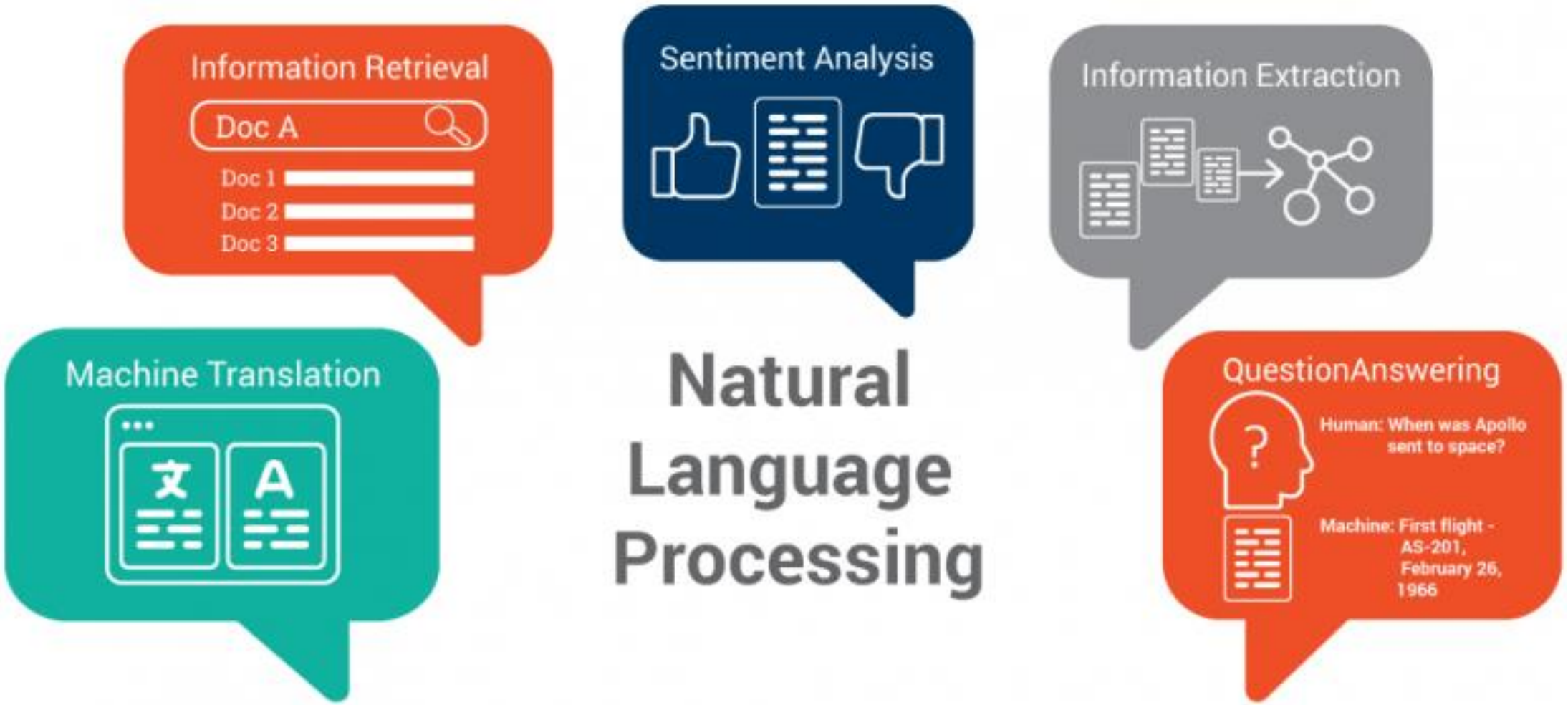
Graduate Qualifying Project - Fall 2018

Huanhan Liu | Rushikesh Naidu | Yi Pan | Yun Yue

Mentors: Mark Baldi | Matthew Fitzpatrick

Advisors: Fatemah Emdad | Chun Kit Ngan

MassDEP WPI



https://medium.com/greyatom/introduction-to-natural-language-processing-78baac3c602b

Methodology

Understanding Data

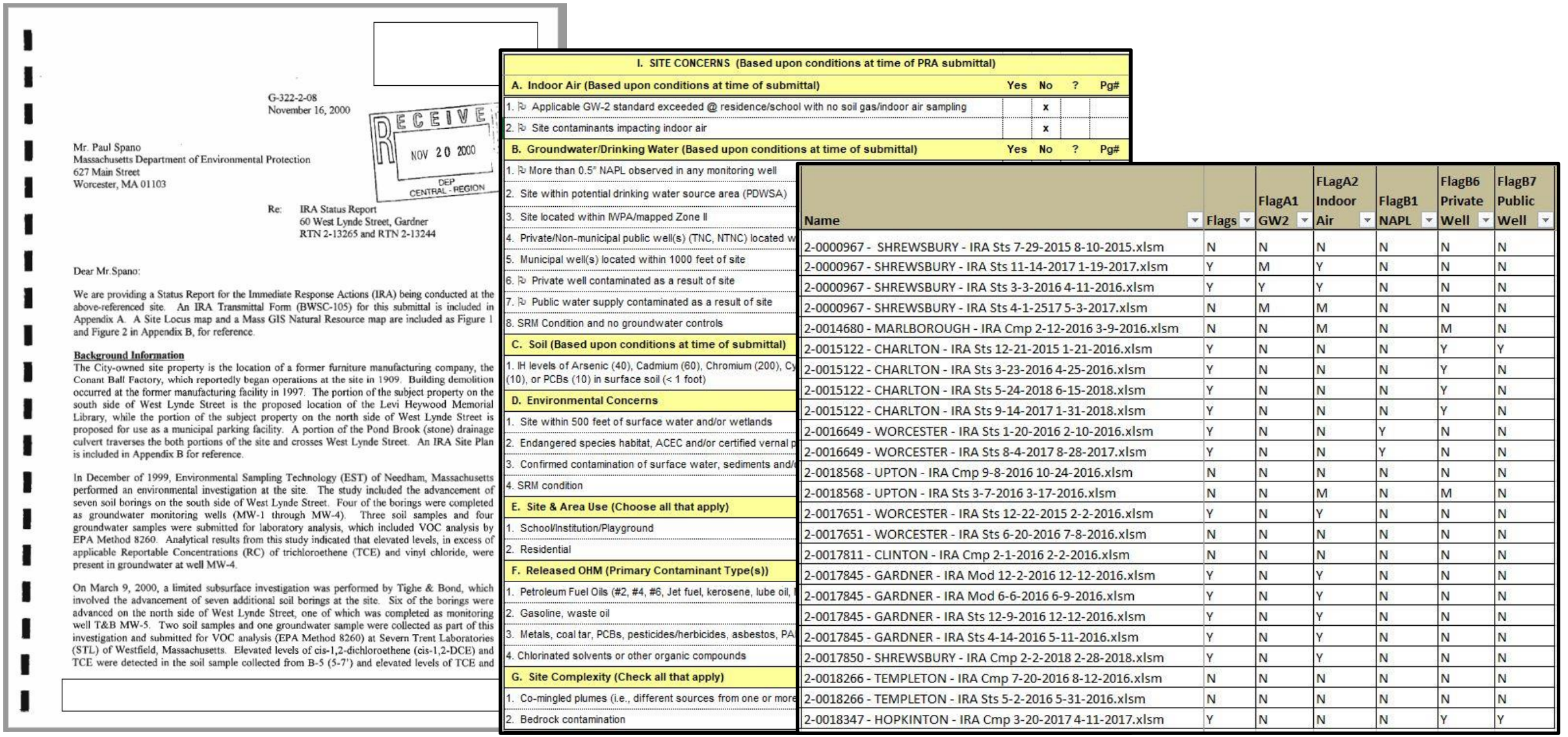
Data Cleaning

Natural Language Processing

Neural Network

MassDEP WPI

Understanding the Data



G-322-2-08
November 16, 2000

RECEIVED
NOV 20 2000
DEP
CENTRAL - REGION

Mr. Paul Spano
Massachusetts Department of Environmental Protection
627 Main Street
Worcester, MA 01103

Re: IRA Status Report
60 West Lynde Street, Gardner
RTN 2-13265 and RTN 2-13244

Dear Mr.Spano:

We are providing a Status Report for the Immediate Response Actions (IRA) being conducted at the above-referenced site. An IRA Transmittal Form (BWSC-105) for this submittal is included in Appendix A. A Site Locus map and a Mass GIS Natural Resource map are included as Figure 1 and Figure 2 in Appendix B, for reference.

Background Information
The City-owned site property is the location of a former furniture manufacturing company, the Conant Ball Factory, which reportedly began operations at the site in 1909. Building demolition occurred at the former manufacturing facility in 1997. The portion of the subject property on the south side of West Lynde Street is the proposed location of the Levi Heywood Memorial Library, while the portion of the subject property on the north side of West Lynde Street is proposed for use as a municipal parking facility. A portion of the Pond Brook (stone) drainage culvert traverses the both portions of the site and crosses West Lynde Street. An IRA Site Plan is included in Appendix B for reference.

In December of 1999, Environmental Sampling Technology (EST) of Needham, Massachusetts performed an environmental investigation at the site. The study included the advancement of seven soil borings on the south side of West Lynde Street. Four of the borings were completed as groundwater monitoring wells (MW-1 through MW-4). Three soil samples and four groundwater samples were submitted for laboratory analysis, which included VOC analysis by EPA Method 8260. Analytical results from this study indicated that elevated levels, in excess of applicable Reportable Concentrations (RC) of trichloroethene (TCE) and vinyl chloride, were present in groundwater at well MW-4.

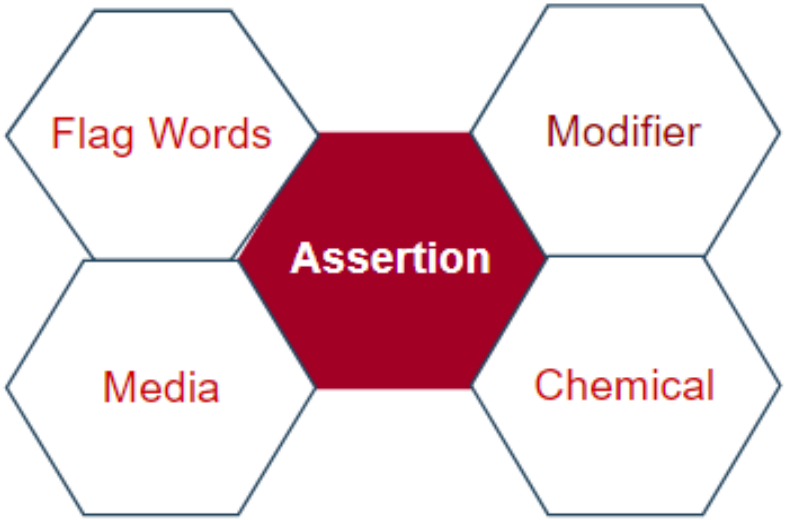
On March 9, 2000, a limited subsurface investigation was performed by Tighe & Bond, which involved the advancement of seven additional soil borings at the site. Six of the borings were advanced on the north side of West Lynde Street, one of which was completed as monitoring well T&B MW-5. Two soil samples and one groundwater sample were collected as part of this investigation and submitted for VOC analysis (EPA Method 8260) at Severn Trent Laboratories (STL) of Westfield, Massachusetts. Elevated levels of cis-1,2-dichloroethene (cis-1,2-DCE) and TCE were detected in the soil sample collected from B-5 (5-7') and elevated levels of TCE and

I. SITE CONCERNS (Based upon conditions at time of PRA submittal)				
A. Indoor Air (Based upon conditions at time of submittal)	**Yes**	**No**	**?**	**Pg#**
1. Applicable GW-2 standard exceeded @ residence/school with no soil gas/indoor air sampling		x		
2. Site contaminants impacting indoor air		x		
B. Groundwater/Drinking Water (Based upon conditions at time of submittal)	**Yes**	**No**	**?**	**Pg#**
1. More than 0.5" NAPL observed in any monitoring well				
2. Site within potential drinking water source area (PDWSA)				
3. Site located within IWPA/mapped Zone II				
4. Private/Non-municipal public well(s) (TNC, NTNC) located w				
5. Municipal well(s) located within 1000 feet of site				
6. Private well contaminated as a result of site				
7. Public water supply contaminated as a result of site				
8. SRM Condition and no groundwater controls				
C. Soil (Based upon conditions at time of submittal)				
1. IH levels of Arsenic (40), Cadmium (60), Chromium (200), Cy (10), or PCBs (10) in surface soil (< 1 foot)				
D. Environmental Concerns				
1. Site within 500 feet of surface water and/or wetlands				
2. Endangered species habitat, ACEC and/or certified vernal p				
3. Confirmed contamination of surface water, sediments and/				
4. SRM condition				
E. Site & Area Use (Choose all that apply)				
1. School/Institution/Playground				
2. Residential				
F. Released OHM (Primary Contaminant Type(s))				
1. Petroleum Fuel Oils (#2, #4, #6, Jet fuel, kerosene, lube oil,				
2. Gasoline, waste oil				
3. Metals, coal tar, PCBs, pesticides/herbicides, asbestos, PA				
4. Chlorinated solvents or other organic compounds				
G. Site Complexity (Check all that apply)				
1. Co-mingled plumes (i.e., different sources from one or more				
2. Bedrock contamination				

Name	Flags	FlagA1 GW2	FLagA2 Indoor Air	FlagB1 NAPL	FlagB6 Private Well	FlagB7 Public Well
2-0000967 - SHREWSBURY - IRA Sts 7-29-2015 8-10-2015.xlsm	N	N	N	N	N	N
2-0000967 - SHREWSBURY - IRA Sts 11-14-2017 1-19-2017.xlsm	Y	M	Y	N	N	N
2-0000967 - SHREWSBURY - IRA Sts 3-3-2016 4-11-2016.xlsm	Y	Y	Y	N	N	N
2-0000967 - SHREWSBURY - IRA Sts 4-1-2517 5-3-2017.xlsm	N	M	M	N	N	N
2-0014680 - MARLBOROUGH - IRA Cmp 2-12-2016 3-9-2016.xlsm	N	N	M	N	M	N
2-0015122 - CHARLTON - IRA Sts 12-21-2015 1-21-2016.xlsm	Y	N	N	N	Y	Y
2-0015122 - CHARLTON - IRA Sts 3-23-2016 4-25-2016.xlsm	Y	N	N	N	Y	N
2-0015122 - CHARLTON - IRA Sts 5-24-2018 6-15-2018.xlsm	Y	N	N	N	Y	N
2-0015122 - CHARLTON - IRA Sts 9-14-2017 1-31-2018.xlsm	Y	N	N	N	Y	N
2-0016649 - WORCESTER - IRA Sts 1-20-2016 2-10-2016.xlsm	Y	N	N	Y	N	N
2-0016649 - WORCESTER - IRA Sts 8-4-2017 8-28-2017.xlsm	Y	N	N	Y	N	N
2-0018568 - UPTON - IRA Cmp 9-8-2016 10-24-2016.xlsm	N	N	N	N	N	N
2-0018568 - UPTON - IRA Sts 3-7-2016 3-17-2016.xlsm	N	N	M	N	M	N
2-0017651 - WORCESTER - IRA Sts 12-22-2015 2-2-2016.xlsm	Y	N	Y	N	N	N
2-0017651 - WORCESTER - IRA Sts 6-20-2016 7-8-2016.xlsm	N	N	N	N	N	N
2-0017811 - CLINTON - IRA Cmp 2-1-2016 2-2-2016.xlsm	N	N	N	N	N	N
2-0017845 - GARDNER - IRA Mod 12-2-2016 12-12-2016.xlsm	Y	N	Y	N	N	N
2-0017845 - GARDNER - IRA Mod 6-6-2016 6-9-2016.xlsm	Y	N	Y	N	N	N
2-0017845 - GARDNER - IRA Sts 12-9-2016 12-12-2016.xlsm	Y	N	Y	N	N	N
2-0017845 - GARDNER - IRA Sts 4-14-2016 5-11-2016.xlsm	Y	N	Y	N	N	N
2-0017850 - SHREWSBURY - IRA Cmp 2-2-2018 2-28-2018.xlsm	Y	N	Y	N	N	N
2-0018266 - TEMPLETON - IRA Cmp 7-20-2016 8-12-2016.xlsm	N	N	N	N	N	N
2-0018266 - TEMPLETON - IRA Sts 5-2-2016 5-31-2016.xlsm	N	N	N	N	N	N
2-0018347 - HOPKINTON - IRA Cmp 3-20-2017 4-11-2017.xlsm	Y	N	N	N	Y	Y

Understanding the Data

I. SITE CONCERNS (Based upon conditions at time of PRA submittal)
A. Indoor Air (Based upon conditions at time of submittal)
1. Applicable GW-2 standard exceeded @ residence/school with no soil gas/indoor air sampling
2. Site contaminants impacting indoor air



- Flag Words – impact, affect, contaminate…
- Media – soil, groundwater, indoor air…
- Modifier – greater than, less than…
- Chemical – CVOCs, gasoline, petroleum, TCE…

MassDEP WPI

Data Cleaning

- Compile reports/tech screen scores
- Unlock reports
- Extract PDF reports to text
- Identify/aggregate keyword/flag word sentences
- Images/Tables?
- Eliminate non-essential numeric characters
- Annotate extracted sentences

https://www.invensis.net/blog/data-processing/5-advantages-of-data-cleansing/

Natural Language Processing

	royalty	femininity	intelligence
king	0.9	-0.9	0.5
queen	0.9	0.9	0.5
man	0.1	-0.9	0.5
woman	0.1	0.9	0.5
smart	0.5	0	0.87
intelligent	0.5	0	0.9

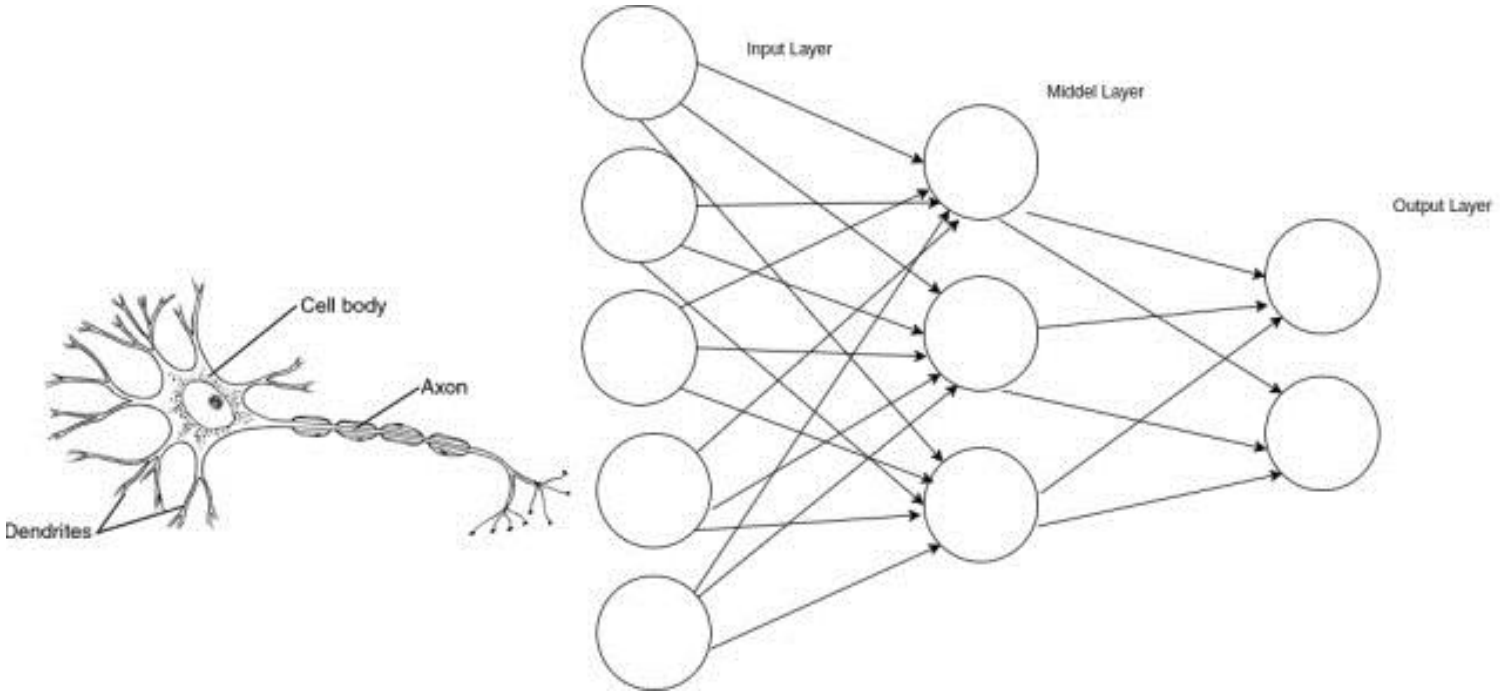
Word To Vector

- Term Frequency - Inverse Document Frequency
- Skip-Gram / Neighbor words prediction

https://primer.ai/blog/Chinese-Word-Vectors/

Neural Network

Artificial neural networks are computing systems vaguely inspired by the biological neural networks that constitute animal brains.
(https://en.wikipedia.org/wiki/Artificial_neural_network)



https://dzone.com/articles/an-introduction-to-the-artificial-neural-network

MassDEP WPI

Neural Network

Long Short Term Memory

Long short-term memory model is a recurrent neural network composed of units/cells with an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. *(https://en.wikipedia.org/wiki/Long_short-term_memory)*

For Example: You remember you eat lunch, how to eat lunch, what you like for lunch, what you had for lunch, and you try new foods for lunch that you may or may not like

Convolutional Neural Network

A Convolutional Neural Network (CNN) is comprised of one or more convolutional layers and then followed by one or more fully connected layers. The architecture of a CNN is designed to take advantage of the input feature local connections and tied weights followed by some form of pooling which results in translation of invariant features.
(http://ufldl.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork/)

For Example: If you look at a small portion of a picture of a cat you may only see fur, as you move your view frame over the cat picture you see more cat features, cat ears, cat mouth and cat eyes until you finally realize you are looking at a picture of a cat.

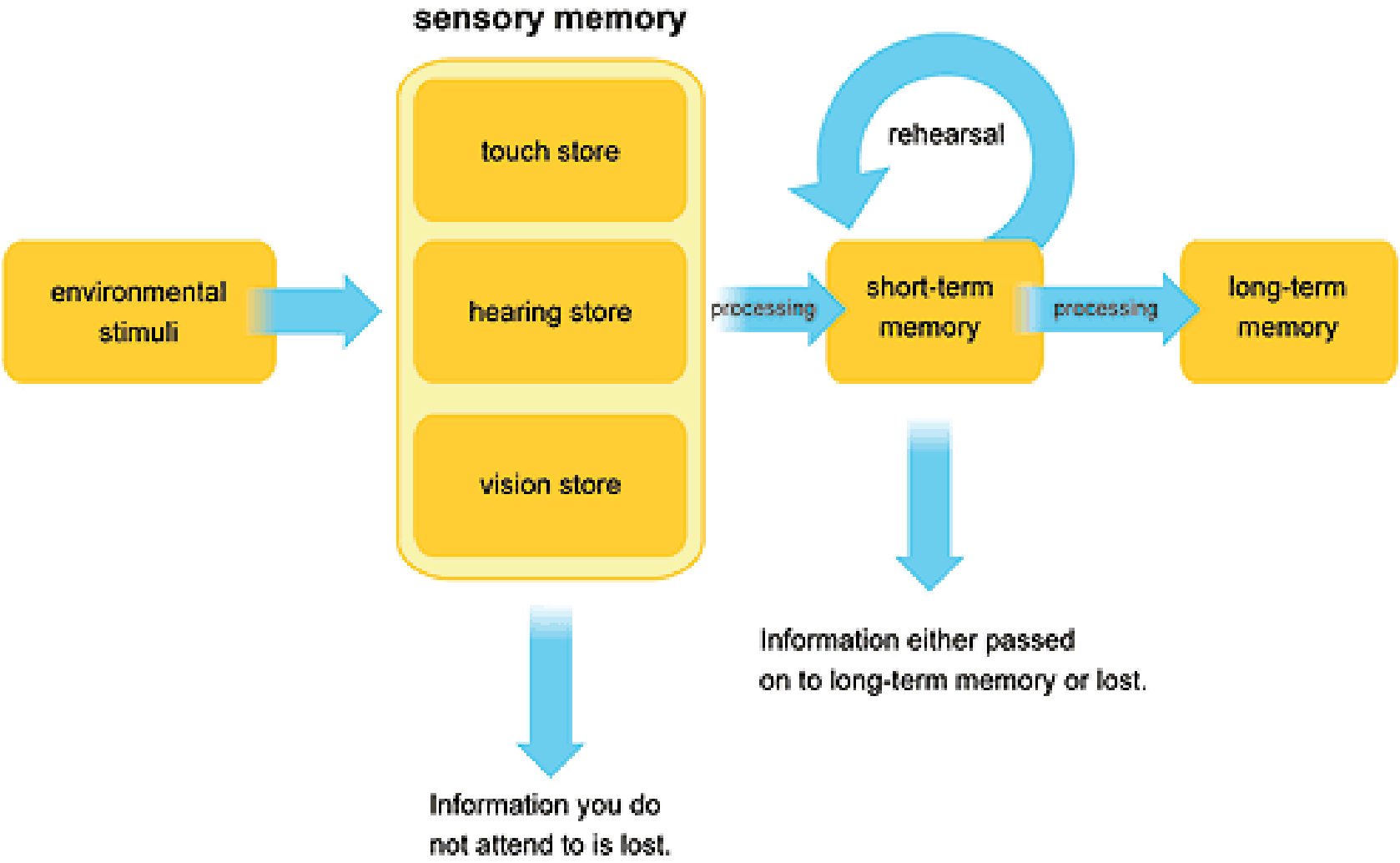
MassDEP WPI

PDF Sentence Extraction

[' CMG supervised excavation of another approximately 475 cubic yards of petroleum-impacted soil from the Property in 2012, and there were also minor impacted soil removals in 2005, 2012, 2013, and 2014 (primarily drill cuttings from advancement of soil borings for groundwater monitoring wells).',
 ' 2 fuel oil) in former Site groundwater monitoring well MW-3D on June 28, 2000.',
 ' We installed replacement monitoring wells in January 2012, including MW-103M (replacement to MW-3D) and MW-103D (replacement to MW-3R).',
 ' Subsequent IRA activities identified measurable LNAPL in Site monitoring wells MW-13 & MW-103M and ISCO injection wells INJ-02, INJ-03, INJ-04, INJ-4R (replacement to INJ-04), and INJ-09 through INJ-12.',
 ' We also identified >¾" LNAPL in monitoring well MW-103M (6") and injection well INJ-11 (¾") on that date, along with ¾" LNAPL in injection wells INJ-02 and INJ-10.',
 '" The OMM plan also calls for periodic analysis of groundwater samples from certain monitoring wells at the Property.',
 "10' of LNAPL in monitoring well MW-13 and 0.",
 ' CMG collected groundwater samples from monitoring wells MW-13 and MW-103M on September 8, 2017 using low-flow methodology.',
 ' CMG also accurately gauged the depth to groundwater in 15 Property monitoring wells and 3 ISCO injection wells on September 8, 2017.',
 '0425(3)(c)] CMG placed LNAPL and groundwater recovered from Site monitoring wells in September 2017 (estimated at 7 gallons, or 26 liters) into the 55-gallon accumulation drum securely stored inside the barn building on the Property.',
 ' We performed the following scope of services between April 2017 and September 2017: \uf0b7 Gauged select Site monitoring and injection wells, and recovered LNAPL from monitoring wells MW-13 & MW-103M and ISCO injection well INJ-4R on September 8, 2017; \uf0b7 Collected low-flow groundwater samples from monitoring wells MW-103M & MW-13 on September 8, 2017 and submitted these for laboratory analyses; \uf0b7 Informed the Property owners of the results of laboratory analyses; \uf0b7 Tabulated & graphed LNAPL accumulation data; \uf0b7 Prepared this OMM & IRA Status Report; and \uf0b7 Prepared Comprehensive Response Action and IRA transmittal forms for Admirals Bank's electronic certification and submittal using eDEP.',
 ' Increasing exploration (such as placement of test pits, completion of additional soil borings with subsequent collection of soil samples for laboratory analysis, installation of additional groundwater monitoring wells with subsequent collection of groundwater samples for laboratory analysis, and conducting surface geophysical survey techniques) may better delineate subsurface conditions.',
 ' CMG collected groundwater samples from two monitoring wells at your property on September 8, 2017 (MW-13 & MW-103M).']

MassDEP WPI

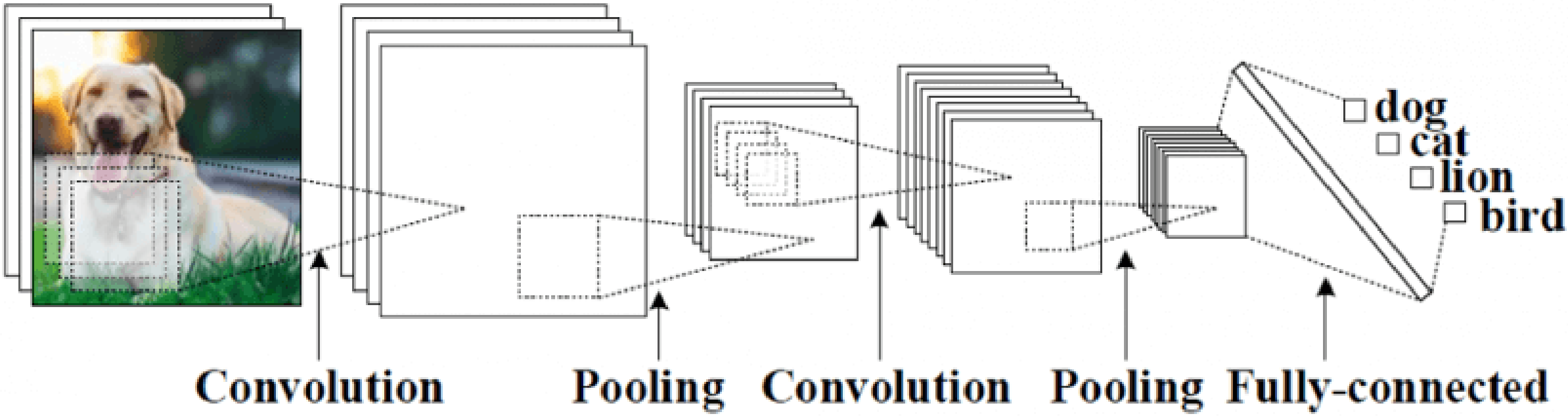
Long Short Term Memory



MassDEP WPI

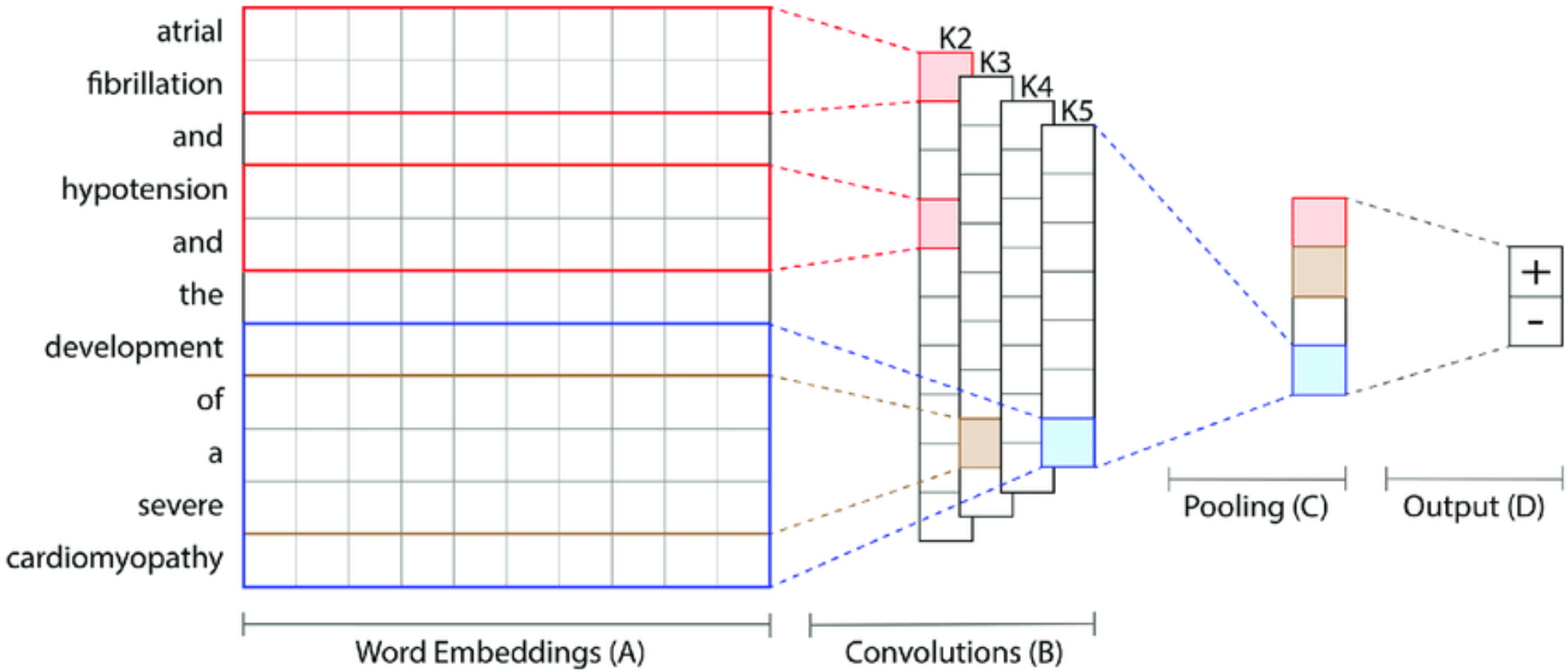
http://www.bbc.co.uk/schools/gcsebitesize/science/add_ocr_21c/brain_mind/complexrev3.shtml

Convolutional Neural Network



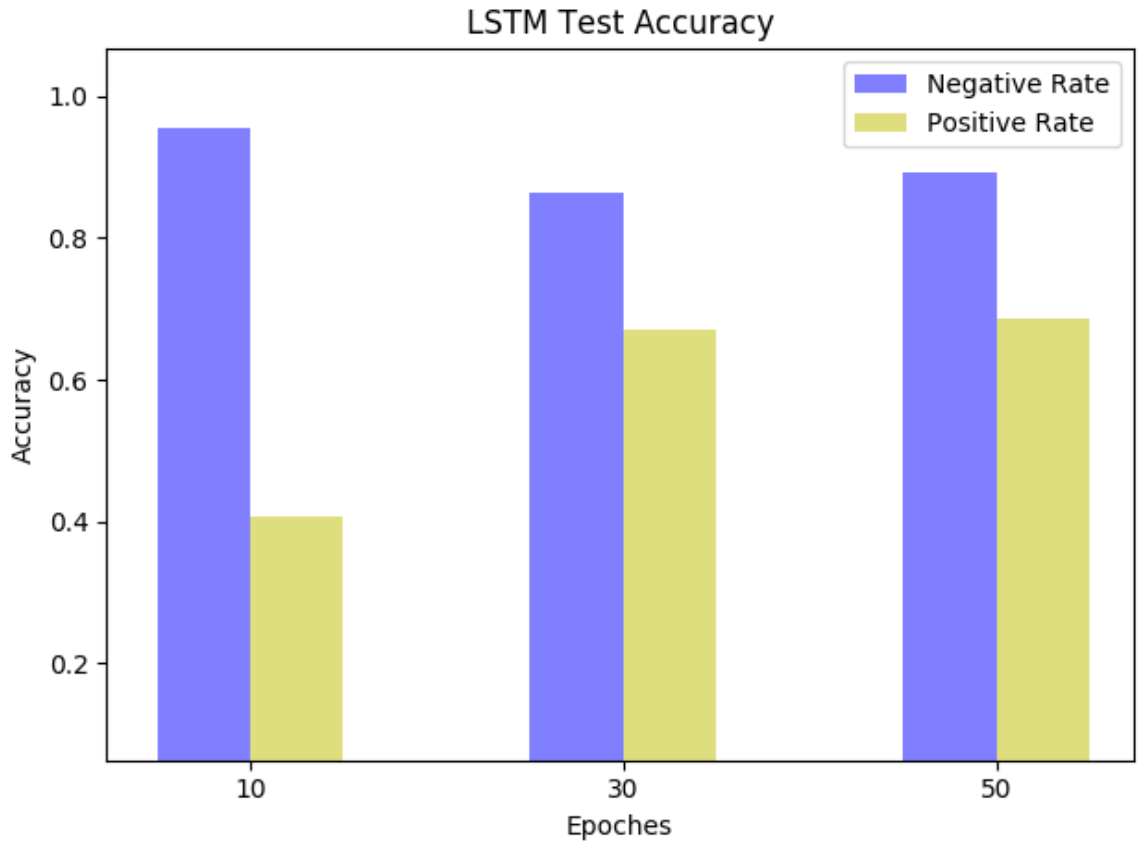
https://www.ayasdi.com/blog/artificial-intelligence/using-topological-data-analysis-understand-behavior-convolutional-neural-networks/

Convolutional Neural Network



https://www.researchgate.net/figure/Overview-of-the-basic-CNN-architecture-A-Each-word-within-a-discharge-note-is_fig1_323213106

Long Short Term Memory Result



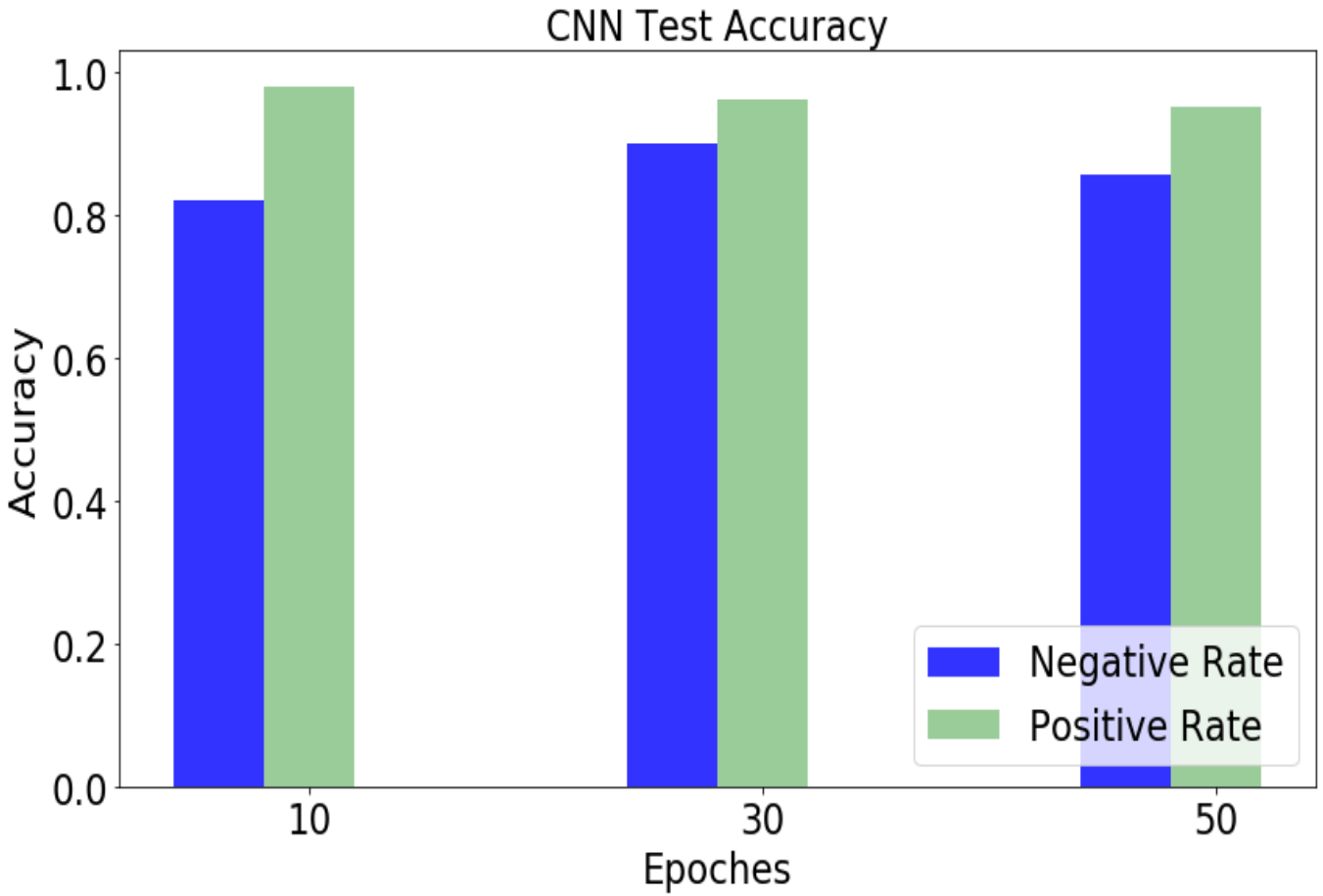
Summary

1. Final test accuracy of model
   - Positive flag prediction accuracy: 70%
   - Negative flag prediction accuracy: 90%

2. More training steps increase largely on positive flag prediction accuracy, with a trade off of slight decrease on negative accuracy

MassDEP WPI

Convolutional Neural Network Result



Summary

1. Final test accuracy of model
   - Positive flag prediction accuracy: 96%
   - Negative flag prediction accuracy: 85%

2. Add punishment when model predict negative but the real situation is positive. Model has a better positive accuracy than negative accuracy.

MassDEP WPI

Summary and Conclusion

- NLP with deep learning methods (CNN and LSTM-RNN) provides a feasible solution for flag condition prediction of text based IRA reports. In both the CNN and LSTM model, prediction performance shows promising results on correctly identifying positive flag conditions based on the collected test reports.

- Further data cleaning, more balanced data sampling, and a more comprehensive model will increase the accuracy on flag condition predictions.

Project Mentors

- Mark E. Baldi, Deputy Regional Director, BWSC
- Matthew Fitzpatrick, BWSC Data Management Coordinator

Faculty Advisors

- Elke A. Rundensteiner, Data Science Director, WPI
- Fatemeh Emdad, Data Science Professor, WPI
- Chun-Kit Ngan, Data Science Professor, WPI

MassDEP WPI

GQP MassDEP Fall 2018 Team

- Huanhan Liu, MS Data Science, WPI, hliu7@wpi.edu
- Rushikesh Naidu, MS Data Science, WPI, ranaidu@wpi.edu
- Yi Pan, MS Data Science, WPI, ypan@wpi.edu
- Yun Yue, MS Data Science, WPI, yyue@wpi.edu

MassDEP WPI