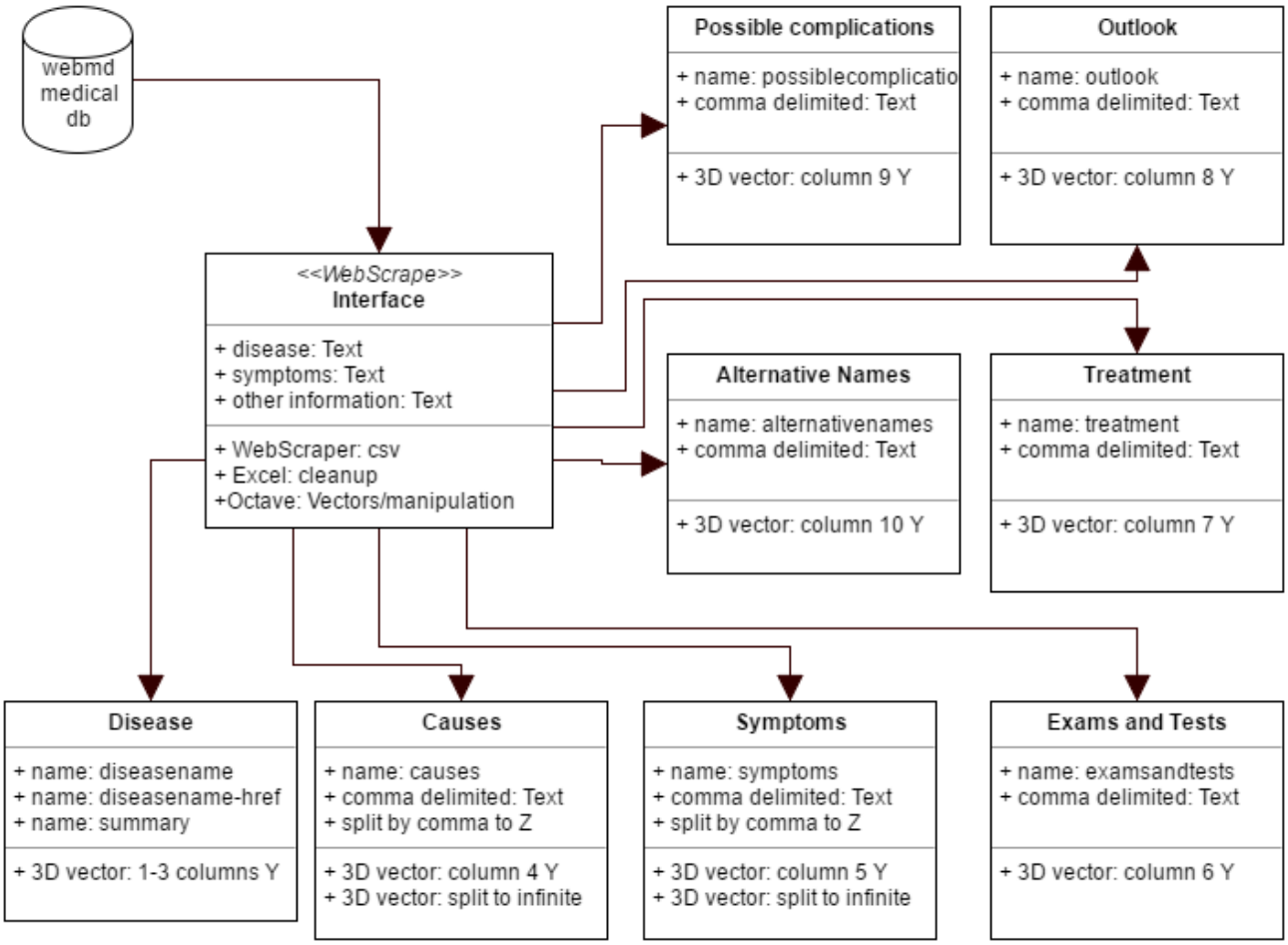
MOTIVATION and Goal

Machine Learning in general is very interesting topic and I wanted to give my best in given time frame two weeks. Prior to project selection at Moodle I wondered how come there is no training for Octave or examples from old projects. Moodle's training videos etc. had basically calculus and other ML related theories, which needed to be converted into algorithms in Octave. Time to deliver project after watching training videos and passing quizzes was about two weeks: last quiz was opened 30.11.2016. This forces me to use tools I handle best.

APPROACH

My plan was to download diseases-symptoms database, either directly or by scraping, clean the data and make an application with Octave. Most important is to get data in, then focus to other phases of the project.

For the ML standpoint, I need to have data which can be classified. Most likely the csv-format will be best for Octave. Data should be cleaned with Excel. Cleaning should be done so that outcome classification be produced from the data. Overall plan would be as follows:



When data is in the csv-format, I need to learn how to use Octave and make asked two or more models by using Octave code. My plan is to look from Google examples and use existing samples as basis of my code.

Compared to Java or VBA language in Octave does not look so complex, but documentation seems to be like in the Unix in general, not very descriptive or intuitive.

DISCUSSION

Project description was at the beginning pointing to Mayo Clinic web site, but it was not clear shall project cover all listed diseases and symptoms found from the Mayo Clinic. It was changed some day to Medlineplus.

I planned to try to split symptoms to their own words and then create vectors from symptom to disease. I guessed best way is to split to 3D array/vector, where depth Z contains split words from symptoms. I believed, that they can be used for weighting the match and for regressions and other ML algorithms.

First issue was that the Octave cannot do web scraping, therefore I had to study which tools works best in this case. Then I ended up to Web scraping issues:

- There are many web scraping utilities available, commercial and free applications. I decided to test out htttrack[1]. It took about 18 hours to download Mayo Clinic website to the PC over 100 Mbit/s connection, but those files are having unique html-file names and therefore it is complex to make a script, which organizes them so that they can be stripped out from the html-code and imported to the Excel. Data size was 1,97 gigabytes holding 61 614 files (just html, no pictures or videos etc media). I consider this as a show stopper for the httrack: not able to categorize nor import data to Octave or to Excel. I had to look for other tools.
- I found a Web Scraper[2] plugin for the Chrome shall be easiest and free tool. This point I found that the project description had changed pointing to Medlineplus, not Mayo Clinic.
- Try #1: problem with multi lines (list items) items are scraped so that they are all together without separator. I found multi-item working somewhat promising, but for some reason full scale scrape gave nothing. Test run took 20 hours. Then I did again small scale testing and finally got promising data. However, full scale scraping failed again.

1. I decided to look other tools for scraping. Biggest issue with Web Scraper was that regexr (regular expression) can't replace list separators (html <li>) into other characters and all items is scraped into on long word. Therefore, I tried following tools:
    a. Data Miner – need to pay and complex.
    b. iRobot: too complex and very old.

Spent a lot of time with these tools, but they are either complex to use or commercial and rather expensive.

Finally, I got an idea to go back to Web Scraper plugin and scrape symptoms etc. in a html format and then later in Excel clean out html codes and replace list item (balloons) with a dot (separation character).

PC to scraped again some 20 hours. This method worked fine. Importing csv into Excel works like a charm and found 4396 diseases. Data looked promising overall. I cleaned the data by:

- replaced with "." -> overcomes the list problem. Now each line ends to "." which could be used as a tag for symptom etc. separation later in Octave or VBA code.
- removed html codes with replacing <*> to nothing.
- Excel diseases-db.xslx size is 4084 (4 megs) and in csv format size is 10143 (10 megs).

Phase 1 was completed fine. Data is usable.

Then I run into show stopper with Octave. I needed to make an 3D Vector/Array for the Octave, where we later load the csv file. Found, that Octave do not support 3-dimensional (z,y,z) Vectors/Arrays at all. I planned to make two separate Vectors and build some sort of relationship for split words.
Next I hit into problems with Octave version. In the lecture videos, download link pointed to version 3.2.4, which is very old and IO is very limited. I uninstalled old and installed version 4.2, but for some reason pkg load io behaved strange in my primary PC. I installed Octave into another PC, and it seemed to work. Decided to clean original PC installation, and after reinstallation the IO module worked. However, importing csv file with strings remained still big issue. After testing many methods, I got this method working for import:

```
[DisName,DisHref,Summary,Causes,Symptoms,ExmTests,Treatment,Outlook,Poscomp,AltNames] =
textread('disdb.csv','%s %s %s %s %s %s %s %s %s %s','delimeter',',',1) ….
```

[1] Description at http://www.httrack.com
[2] Free from Chrome Web Store

Resulting vector was ugly, not usable at all. I googled around and realized, that %s should be replaced with %q, but “%q” option (shall read whole text row) do not work with textread! This means, that importing with textread is out of question. I tried Octave’s xlsread from xlsx files. This is not working either for strings. I tried saving csv as a text etc. Googling around for a few days. I did not find any help nor examples. Felt pressure to deliver project and considered this as a show stopper for full scale Octave project.

Main goal was to make a Virtual Doctor. I needed rethink with which tool I can categorize strings and import them into Octave and then deliver project in given timeframe. Started building Excel with export sheet OctaveLink, but unfortunately, Excel do not have Octave API. Therefore, I had to complete whole project with Excel VBA and understand, that I have to do this in a “poor man’s way” with those constraints which Excel has.

My new plan was to build a word list of symptoms and then find full and partial matches in relation to disease for classification and pass classification to SVM etc. modules. Theoretical point of view, I end up to SVM Classification method. A linear SVM requires solving a quadratic program with several linear constraints. Linear classification is possible to do with Excel, but classifier margins I was not able to resolve.

I also investigated what is ML and Excel situation now and it seems that Microsoft bought Revolution Analytics (R-focuses) and ML is in fact available at Azure and DataScope. I do have an Azure account, but sharing Azure API key with peers reviewing my work is not tempting. However, it maybe so that ML capabilities will be added to future Excel-versions and therefore this exercise should be fine.

Virtual Doctor operation

This is just one Excel file holding data, which is imported from csv. Data can be updated any time. Design looks for the column name and adjust VBA code based on the column names. You can access the VBA code by pressing Alt + F11. The VBA code is commented.

User-sheet contains macros which triggers subroutines based on which cell value is changed. As an example, if user types in symptom 1, macro “readisease” at Module1 is executed.

First sheet is a User input and output page. User do not need to use other sheets, but I left them open and usable so that you can review them.

Red squared cells are cells user is supposed to type in (they are unlocked in case you would like to protect the sheets). Sheet has 6 input fields for symptoms, which we will use as vectors. When user input symptom, Excel will go thru the database and return first and last name of the disease where disease is found. Excel also counts how many diseases matches to the symptom.

Search method is xlPart, which looks partial matches. All matching diseases are collected to OctaveLink page (this was my first priority) and then each repeating, same name disease from other vectors/symptoms are calculated. If the same disease is repeating with other search words, then we give more weight on the disease by counting them. Last we check duplicates with pivot table (because there could be more than just one disease with same weight) and bring our best preidction to User sheet.

In the middle of the User sheet we will show how many diseases and words the database has. We also list total diseases matching search criteria and how many times they are repeating. With this information we also calculate simple probability % and how many % of maximum repeats disease hits.

System can be learned so that if preidcted disease is wrong, user can select Y to “consider this result as false positive” and then select new disease, which should match to search words / vectors. If displayed disease is learned, background color will be changed to turquoise and “THIS DISEASE IS LEARNED” is showed in the User sheet. Example:

	A	B	C	D	E
1	**Virtual Doctor by Jari Hiltunen 2016**		Words in symptoms	Count words	Scroll down for more details
2	Diseases in the database	4396	92197		
	Sheet will update by pressing enter	Found from	Normal		
3	Or by pressing Ctrl+r	# of diseases	Propability density	First disease matching search criteria	Last disease matching search criteria
4	Type symptoms 1-6 below (vector x1-6) to red circled cells.				
5	Fever	270		ABO incompatibility	Zika virus disease
6	Headache	168		Acetaminophen dosing for children	Zika virus disease
7	Vomiting	202		Abdominal aortic aneurysm	Zollinger-Ellison syndrome
8	Yellow skin	25		ABO incompatibility	Wilson disease
9	Dehydration	23		24-hour urine protein	VIPoma
10	Pain	698		A guide to help children understand cancer	Zollinger-Ellison syndrome
11	**Total listed diseases matching**	1386		**5 are repeating maximums of total inputted symps**	6
12	Results from default method #1 (~ linear regression model)				THIS DISEASE IS LEARNED!
13	**Predicted disease**	Abdominal aortic aneurysm			
14	**Disease reference**	https://medlineplus.gov/ency/article/000162.htm			
15	**Summary**	The aorta is the main blood vessel that supplies blood to the abdomen, pelvis, and legs. An abdominal aortic aneurysm occurs when an area of the aorta t			
16	**Causes**	The exact cause of an aneurysm is unknown. It occurs due to weakness in the wall of the artery. Factors that can increase your risk of having the proble			
17	**Full list of symptoms**	Pain in the abdomen or back. The pain may be severe, sudden, persistent, or constant. It may spread to the groin, buttocks, or legs.Passing out.Clammy s			
18	**Exams and tests**	Your health care provider will examine your abdomen and feel the pulses in your legs. The provider may find:A lump (mass) in the abdomen.Pulsating se			
19	**Treatment**	If you have bleeding inside your body from an aortic aneurysm, you will need surgery right away.If the aneurysm is small and there are no symptoms:Surg			
20	**Prognosis**	The outcome is often good if you have surgery to repair the aneurysm before it ruptures.			
21	**Possible complications**	null			
22	**Alternative names**	Aneurysm - aortic; AAA			
23	Propability % of prediction by repeats	**91,67 %**	of total maximums	83,33 %	<-- how many % out of vectors are maximum repeat
24	Consider this result as false positive?	N	What disease is right?		
25	**Disease reference**				
26	**Summary**				
27	**Causes**				
28	**Full list of symptoms**				
29	**Exams and tests**				

Normal operation User sheet looks like this:

	A	B	C	D	E
1	**Virtual Doctor by Jari Hiltunen 2016**		Words in symptoms	Count words	Scroll down for more details
2	Diseases in the database	4396	92197		
	Sheet will update by pressing enter	Found from	Normal		
3	Or by pressing Ctrl+r	# of diseases	Propability density	First disease matching search criteria	Last disease matching search criteria
4	Type symptoms 1-6 below (vector x1-6) to red circled cells.				
5	Chest pain	91		Achalasia	Wolff-Parkinson-White syndrome (WPW)
6	Loss of appetite	66		Acute adrenal crisis	Yellow fever
7	wheezing	36		Absent pulmonary valve	Visceral larva migrans
8	fatigue	175		Acromegaly	Waldenstrom macroglobulinemia
9	Cough that does not go away	2		Lung cancer	Lung cancer - non-small cell
10	blood	351		24-hour urinary aldosterone excretion test	Zollinger-Ellison syndrome
11	**Total listed diseases matching**	721		**6 are repeating maximums of total inputted symps**	6
12	Results from default method #1 (~ linear regression model)				
13	**Predicted disease**	Lung cancer			
14	**Disease reference**	https://medlineplus.gov/ency/article/007270.htm			
15	**Summary**	Lung cancer is cancer that starts in the lungs.The lungs are located in the chest. When you breathe, air goes through your nose, down your windpipe (t			
16	**Causes**	Lung cancer is the deadliest type of cancer for both men and women. Each year, more people die of lung cancer than of breast, colon, and prostate canc			
17	**Full list of symptoms**	Chest pain.Cough that does not go away.Coughing up blood.Fatigue.Losing weight without trying.Loss of appetite.Shortness of breath.Wheezing.			
18	**Exams and tests**	Lung cancer is often found when an x-ray or CT scan is done for another reason.If lung cancer is suspected, the doctor will perform a physical exam an			
19	**Treatment**	Treatment for lung cancer depends on the type of cancer, how advanced it is, and how healthy you are:Surgery to remove the tumor may be done whe			
20	**Prognosis**	You can ease the stress of illness by joining a cancer support group. Sharing with others who have common experiences and problems can help you not			
21	**Possible complications**	null			
22	**Alternative names**	Cancer - lung			
23	Propability % of prediction by repeats	**85,71 %**	of total maximums	100,00 %	<-- how many % out of vectors are maximum repeat
24	Consider this result as false positive?	N	What disease is right?		

If search word is not found from the database, user will be informed and background color of the symptom turns to white. I have done macros so, that they could be ended and then macro needs to be restarted with ctrl+r.

If user knows, that this is false positive, user can input new disease related to symptom combination.

	A	B	C	D	E
5	Chest pain	91		Achalasia	Wolff-Parkinson-White syndrome (WPW)
6	Loss of appetite	66		Acute adrenal crisis	Yellow fever
7	wheezing	36		Absent pulmonary valve	Visceral larva migrans
8	fatigue	175		Acromegaly	Waldenstrom macroglobulinemia
9	Cough that does not go away	2		Lung cancer	Lung cancer - non-small cell
10	blood	351		24-hour urinary aldosterone excretion test	Zollinger-Ellison syndrome
11	**Total listed diseases matching**	721		**6 are repeating maximums of total inputted symps**	6
12	Results from default method #1 (~ linear regression model)				
13	**Predicted disease**	Lung cancer			
14	**Disease reference**	https://medlineplus.gov/ency/article/007270.htm			
15	**Summary**	Lung cancer is cancer that starts in the lungs.The lungs are located in the chest. When you breathe, air goes through your nose, down your windpipe (tra			
16	**Causes**	Lung cancer is the deadliest type of cancer for both men and women. Each year, more people die of lung cancer than of breast, colon, and prostate cance			
17	**Full list of symptoms**	Chest pain.Cough that does not go away.Coughing up blood.Fatigue.Losing weight without trying.Loss of appetite.Shortness of breath.Wheezing.			
18	**Exams and tests**	Lung cancer is often found when an x-ray or CT scan is done for another reason.If lung cancer is suspected, the doctor will perform a physical exam and			
19	**Treatment**	Treatment for lung cancer depends on the type of cancer, how advanced it is, and how healthy you are:Surgery to remove the tumor may be done when			
20	**Prognosis**	You can ease the stress of illness by joining a cancer support group. Sharing with others who have common experiences and problems can help you not f			
21	**Possible complications**	null			
22	**Alternative names**	Cancer - lung			
23	Propability % of prediction by repeats	**85,71 %**	of total maximums	100,00 %	<-- how many % out of vectors are maximum repeat
24	Consider this result as false positive?	Y	What disease is right?	Leg	Bowlegs
25	**Disease reference**	https://medlineplus.gov/ency/article/001585.htm			
26	**Summary**	Bowlegs is a condition in which the knees stay wide apart when a person stands with the feet a			
27	**Causes**	Infants are born bowlegged because of their folded position in the mother's womb. Bowed legs			
28	**Full list of symptoms**	Knees that do not touch when standing with feet together (ankles touching).Bowing of legs is sa			
29	**Exams and tests**	A health care provider can often diagnose bowlegs by looking at the child. The distance betwee			
30	**Treatment**	No treatment is recommended for bowlegs unless the condition is extreme. The child should be seen by the provider at least every 6 months.Special sho			
31	**Prognosis**	In many cases the outcome is good, and there is most often no problem walking.			
32	**Possible complications**	null			
33	**Alternative names**	Genu varum			
34	Are you sure you want to add this?	N	If you select Y, then search vectors and new disease will be added to the Learning sheet and in later searches corrected.		

Bowlegs
CT angiography - arms and legs
College students and the flu
Doppler ultrasound exam of an arm or leg
Foot, leg, and ankle swelling
Healthy food trends -- beans and legumes
Leg CT scan
Leg MRI scan

By selecting Y information will be added to Learning sheet:

	A	B	C	D	E	F	G	H	I	J	K
1	**This sheet are used for learning. Below list will be learning which output is false positives and which input should be treated as best match.**										
2	Automagically filled by the VBA from row 10 onwards.										
3	After item is listed here, we will do search again from the beginning (from module 1).										
4											
5											
6											
7											
8	Following rows are for false positives = supervised learning. Return correct disease								12	<-- Last used row	
9	Wrong Disease	Vector1	Vector2	Vector3	Vector4	Vector5	Vector6	Propability %	Max %	Item #	Learned/right disease
10	ABO incompatibility	Severe itching	Bleeding	Wrist pain	Yellow skin	Pain	Fever	0,348856209	0,5	1	Zika virus disease
11	Acute adrenal crisis	Fever	Headache	Vomiting	Yellow skin	Pain	Dehydration	0,916666667	0,833	2	Abdominal aortic aneurysm
12	Lung cancer	Chest pain	Loss of appetite	wheezing	fatigue	Cough that does no	blood	0,857142857	1	3	Bowlegs

Next times same search words and predicted disease matches, corrected disease with explanations will be displayed with warning that disease is learned, not predicted.

The OctaveLink sheet was initially designed for linking information into Octave, but I had to skip that option due to lack of API. Information is used for prediction.

	A	B	C	D	E	F	G	H
1	**This page is inteded for linking information into Octave and pulling calculated data back! Automatically filled from 10'th row. (obsolete due to lack of API) - now using for collecting information**							
2	**Based on symptom #1 x^(i)**		**Based on symptom #2 x^(i)**		**Based on symptom #3 x^(i)**		**Based on symptom #4 x^(i)**	
3	Listed symptoms	91	Listed symptoms	66	Listed symptoms	36	Listed symptoms	175
4	Maximum address	B57	Maximum address	D54	Maximum address	F34	Maximum address	H108
5	Maximum repeats of maximum	6	Maximum repeats of maximum	6	Maximum repeats of maximum	6	Maximum repeats of maximum	6
6	Lung cancer	Max name	Lung cancer	Max name	Lung cancer	Max name	Lung cancer	Max name
7	# other diseases having same maximum	1	# other diseases having same maximum	1	# other diseases having same maximum	1	# other diseases having same maximum	1
8	Probability factor / activator max	0,857142857	Probability factor / activator max	0,857142857	Probability factor / activator max	0,857142857	Probability factor / activator max	0,857142857
9	Disease names	# same diseases	Disease names	# same diseases	Disease names	# same diseases	Disease names	# same diseases
10	Achalasia	1	Acute adrenal crisis	3	Absent pulmonary valve	1	Acromegaly	1
11	Angioplasty and stent placement - heart	2	Acute cytomegalovirus (CMV) infection	2	Allergies - overview	1	Acute adrenal crisis	3
12	Aortic insufficiency	2	Acute lymphoblastic leukemia (ALL)	1	Alpha-1 antitrypsin deficiency	1	Acute cytomegalovirus (CMV) infection	2
13	Aortic stenosis	2	Acute mountain sickness	2	Anaphylaxis	1	Acute kidney failure	2
14	Aortic valve surgery - minimally invasive	1	Acute myeloid leukemia (AML) - children	1	Ascariasis	2	Acute mountain sickness	2
15	Aortic valve surgery - open	2	Alcoholic ketoacidosis	2	Aspergillosis	2	Acute myeloid leukemia	1
16	Arterial insufficiency	2	Amebic liver abscess	1	Aspiration pneumonia	4	Addison disease	1
17	Asbestosis	1	Autoimmune hepatitis	2	Asthma	1	Alcohol withdrawal	1

I wish you like this outcome. It does not fully cover project requests, but I did not have any other options in this short timeframe. I wish this version works for all Excel versions.