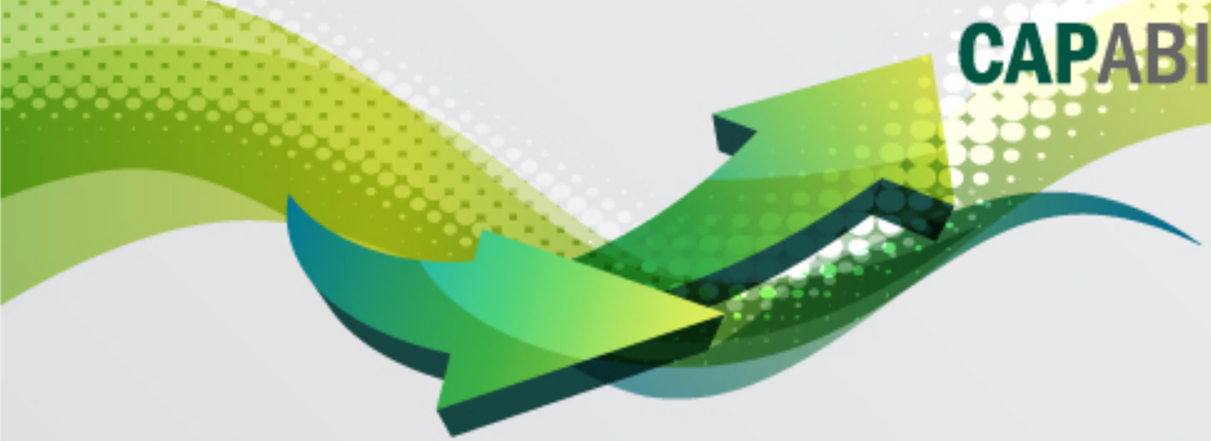


CAPABILITY DEVELOPMENT
ST LOUIS REGION

Daugherty
BUSINESS SOLUTIONS

Machine Learning,
by Rommel Villagomez

October 28, 2017 St. Louis, MO

Instructor Bio

- **Have been playing with Machine Learning for a few years now.**
- **More recently completed several courses in Coursera around Machine Learning and Data Science.**
- **Currently, part of the Mentor Academy for Applied Data Science, where I collaborate with Dr. Christopher Brooks from the University of Michigan, the instructor for the Applied Data Science specialization at Coursera.**

What is Machine Learning?

- Is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world.

Types of machine learning

- Supervised learning
- Unsupervised learning

Supervised Learning

- **We know what we want to predict**: e.g. predict sales for next year, price of a certain stock in the future, inventory needs for next month, etc.

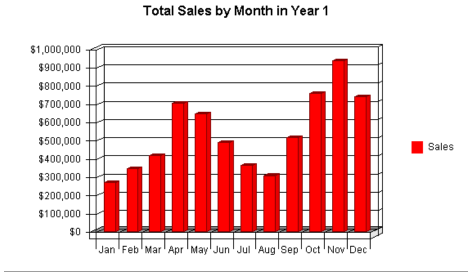
- We have two types:
  - Classification Problems
  - Regression Problems

Supervised Learning/Classification Problems

- We are trying to predict results in a discrete category (e.g. Yes/No)
  - Given information about a patient, we are trying to predict if he will develop diabetes in his lifetime (Yes/No)
  - Given a handwritten number, we are trying to determine what digit was written (0-9)
  - Given the history of customer, we are trying to predict if she will pay her credit card bills or not (Yes/No)

Supervised Learning/Regression Problems

- We are trying to predict results within a continuous output
  - Given data about a house, e.g. location, year that was built, number of bedrooms, etc. we are trying to predict the selling price of the house (0-10 million)
  - Given a picture of a person we are trying to determine the age of that person (0-100)
  - Given data about past sales we are trying to predict future sales (10M – 20 M)



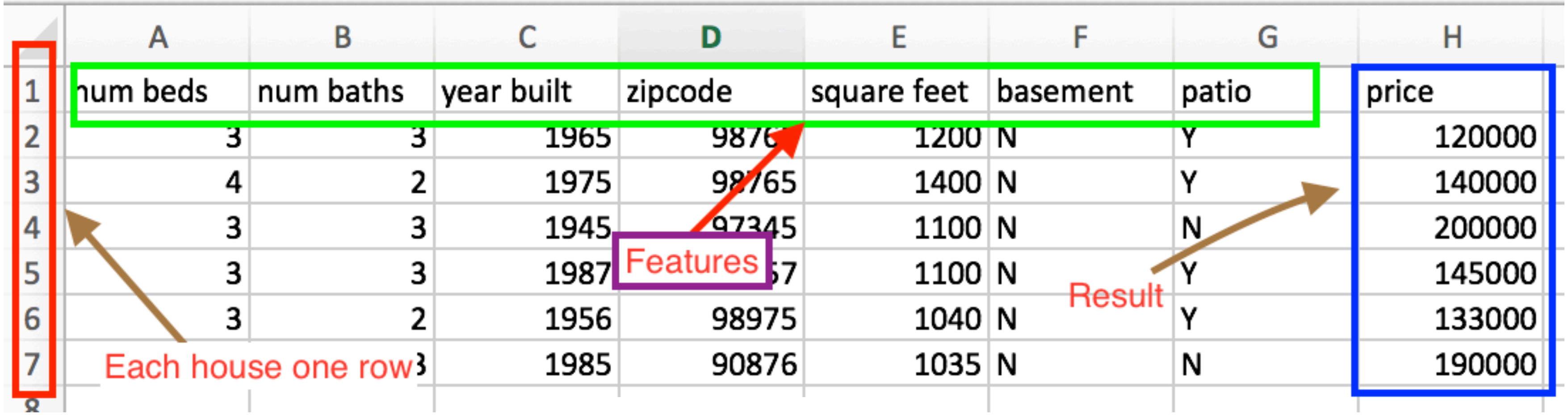
Unsupervised learning

- Allow us to approach problems with little or no idea of what our results are going to look like.
- We cluster the data into different groups.

Unsupervised Learning/ Examples

- Google News
- Galaxy Classification
- Finding genes responsible for certain diseases

Price of a house / how to organize your data



	A	B	C	D	E	F	G	H
1	num beds	num baths	year built	zipcode	square feet	basement	patio	price
2	3	3	1965	9876	1200	N	Y	120000
3	4	2	1975	98765	1400	N	Y	140000
4	3	3	1945	97345	1100	N	N	200000
5	3	3	1987	[illegible]	1100	N	Y	145000
6	3	2	1956	98975	1040	N	Y	133000
7	3	3	1985	90876	1035	N	N	190000

Features

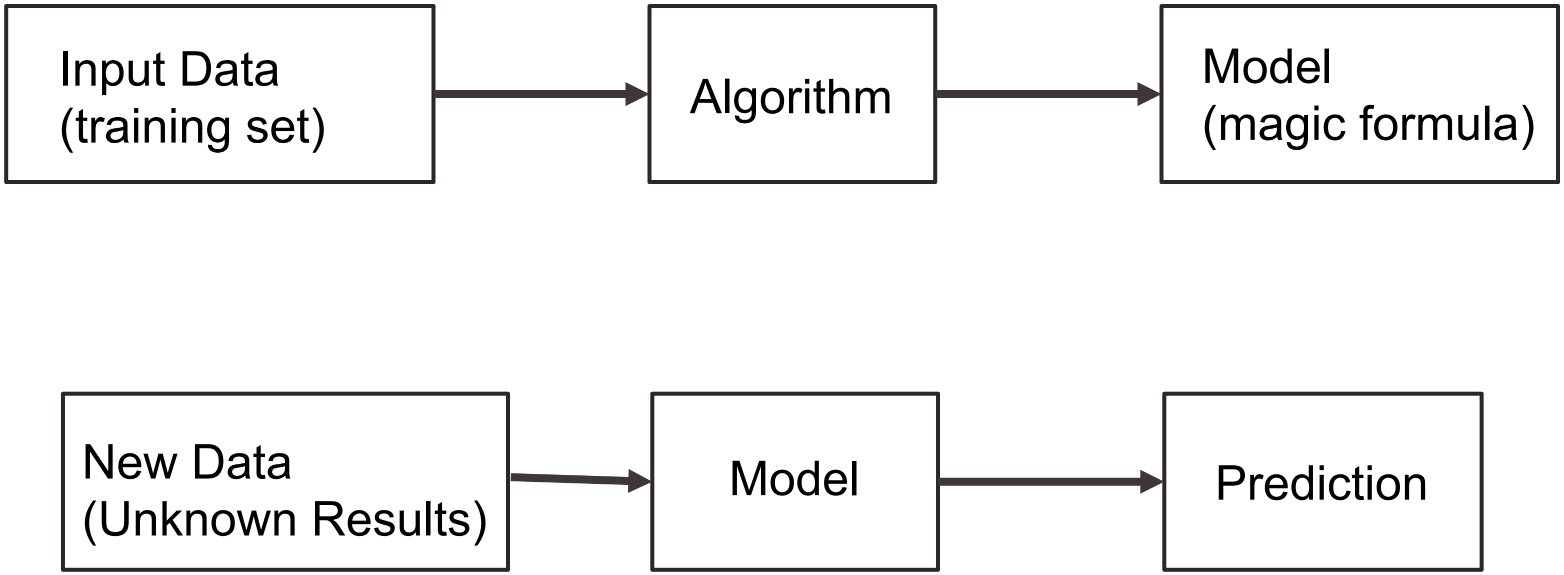
Result

Each house one row

Supervised Learning Classification Problem

- Our goal is to build a spam filter that will classify emails as either spam or non spam.
  - Sort out through thousands of log files and determining if they contain useful information (e.g. intruder attacks) so we can keep them, or if it's ok to delete them.
  - Sorting out through thousands of emails or other documents determining if they are useful for an investigation or not.

Supervised Learning Workflow



Input Data (training set) → Algorithm → Model (magic formula)

New Data (Unknown Results) → Model → Prediction

Process we will follow to get our Input:

- We select 5000 emails, about 2/3 of them are non-spam emails, 1/3 are spam
- We will normalize all the emails.
- We will select only the words that appear 100 or more times in those 5000 emails.
- We will organize the data in a similar fashion as what we did before.

Normalize each email

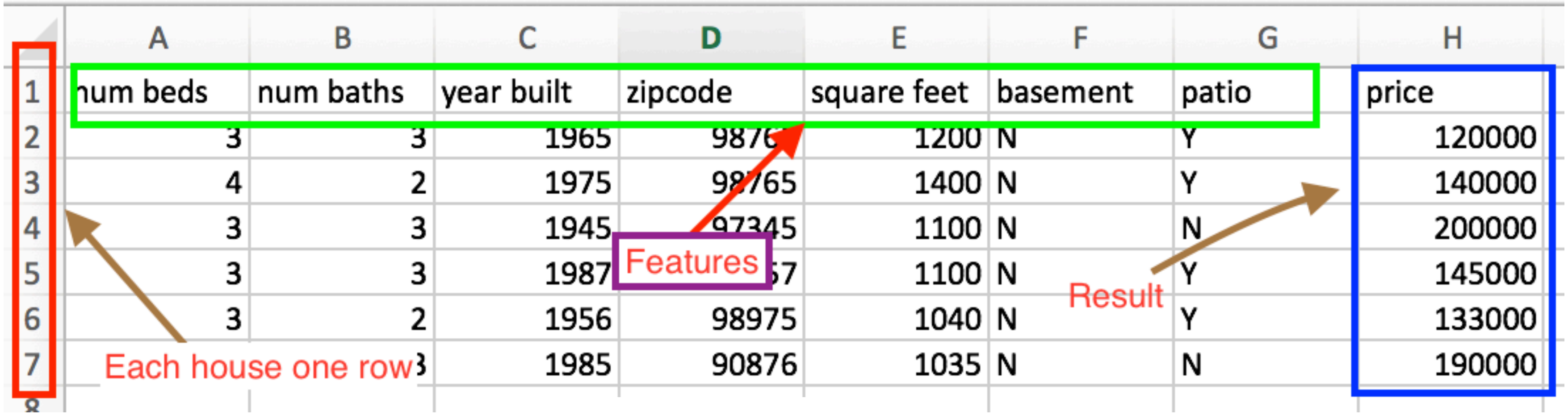
- Lowercase all words
- Strip html tags, and change urls to ‘httpaddr’
- All email addresses are changed to ‘emailaddr’
- All numbers are replaced with the word ‘number’
- All dollar signs are replaced with the word ’dollar’
- Words are reduced to their stemmed word. E.g. including, included, includes, include are all replaced with ‘includ’
- Punctuation, spaces and single letter words are removed

Before and after

- > Anyone knows how much it costs to host a web portal ?
  >
  Well, it depends on how many visitors youre expecting. This can be anywhere from less than 10 bucks a month to a couple of $100. You should checkout http://www.rackspace.com/ or perhaps Amazon EC2 if youre running something big.. To unsubscribe yourself from this mailing list, send an email to: groupname-unsubscribe@egroups.com

anyon know how much it cost to host a web portal well it depend on how many visitor your expect thi can be anywher from less than number buck month to coupl of dollarnumb you should checkout httpaddr or perhap amazon ecnumb if your run someth big to unsubscrib yourself from thi mail list send an email to emailaddr

Price of a house



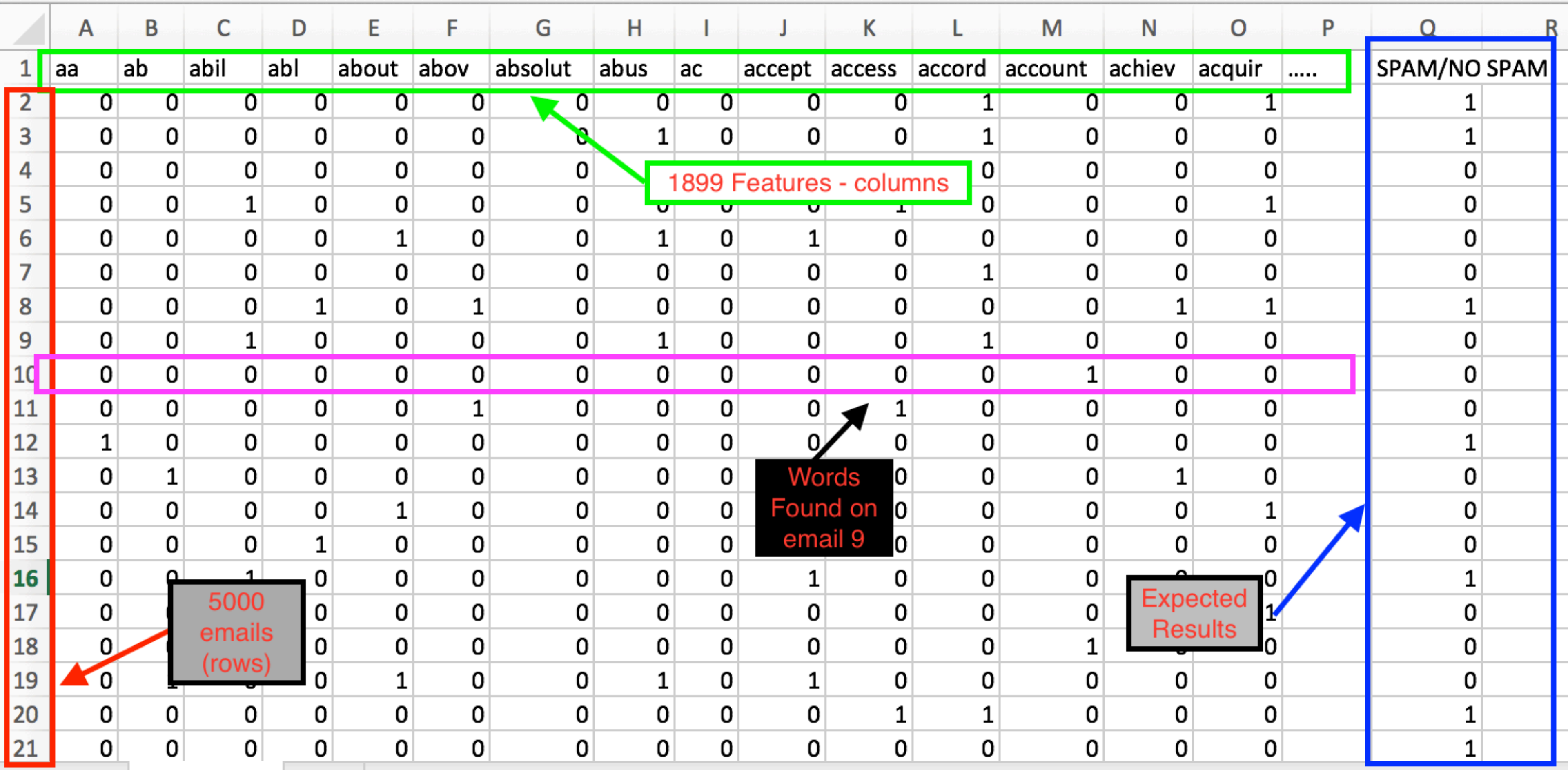
	A	B	C	D	E	F	G	H
1	num beds	num baths	year built	zipcode	square feet	basement	patio	price
2	3	3	1965	9876	1200	N	Y	120000
3	4	2	1975	98765	1400	N	Y	140000
4	3	3	1945	97345	1100	N	N	200000
5	3	3	1987	57	1100	N	Y	145000
6	3	2	1956	98975	1040	N	Y	133000
7	[illegible]	3	1985	90876	1035	N	N	190000

Each house one row

Features

Result

Spam filter Input



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	aa	ab	abil	abl	about	abov	absolut	abus	ac	accept	access	accord	account	achiev	acquir	.....	SPAM/NO SPAM	
2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1		1	
3	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0		1	
4	0	0	0	0	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]	0	0	0	0		0	
5	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1		0	
6	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0		0	
7	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	
8	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1		1	
9	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0		0	
10	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		0	
11	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0		0	
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0		1	
13	0	1	0	0	0	0	0	0	0	[illegible]	0	0	0	1	0		0	
14	0	0	0	0	1	0	0	0	0	[illegible]	0	0	0	0	1		0	
15	0	0	0	1	0	0	0	0	0	[illegible]	0	0	0	0	0		0	
16	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0		1	
17	0	0	[illegible]	0	0	0	0	0	0	0	0	0	0	[illegible]	1		0	
18	0	0	[illegible]	0	0	0	0	0	0	0	0	0	1	0	0		0	
19	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0		0	
20	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0		1	
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		1	

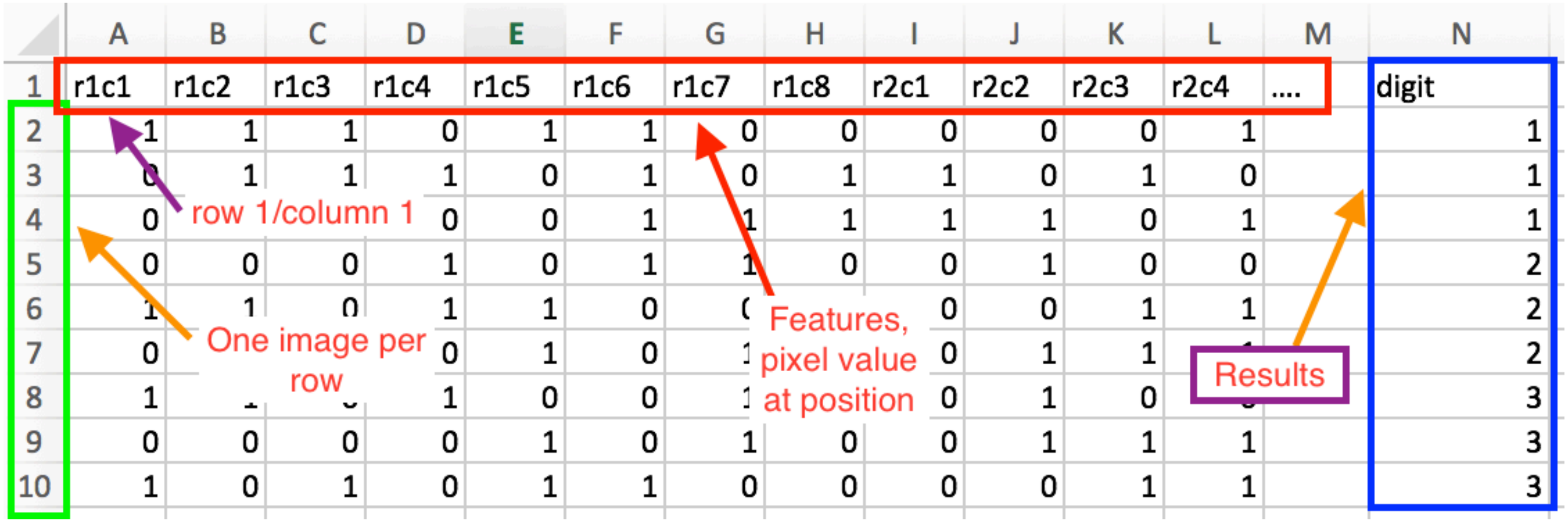
1899 Features - columns

Words Found on email 9

5000 emails (rows)

Expected Results

Hand-printed digits



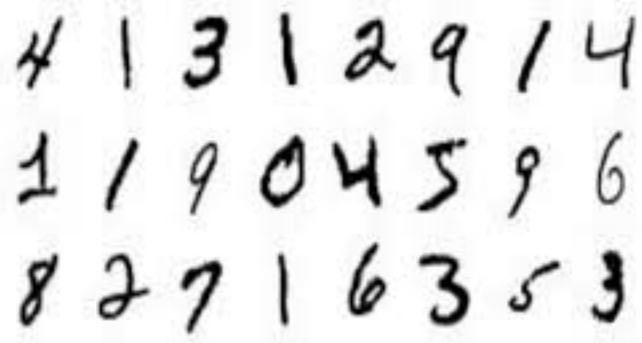
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	r1c1	r1c2	r1c3	r1c4	r1c5	r1c6	r1c7	r1c8	r2c1	r2c2	r2c3	r2c4	....	digit
2	1	1	1	0	1	1	0	0	0	0	0	1		1
3	0	1	1	1	0	1	0	1	1	0	1	0		1
4	0	[illegible]	[illegible]	0	0	1	1	1	1	1	0	1		1
5	0	0	0	1	0	1	1	0	0	1	0	0		2
6	1	1	0	1	1	0	0	[illegible]	0	0	1	1		2
7	0	[illegible]	[illegible]	0	1	0	1	[illegible]	0	1	1	1		2
8	1	1	0	1	0	0	1	[illegible]	0	1	0	0		3
9	0	0	0	0	1	0	1	0	0	1	1	1		3
10	1	0	1	0	1	1	0	0	0	0	1	1		3

row 1/column 1

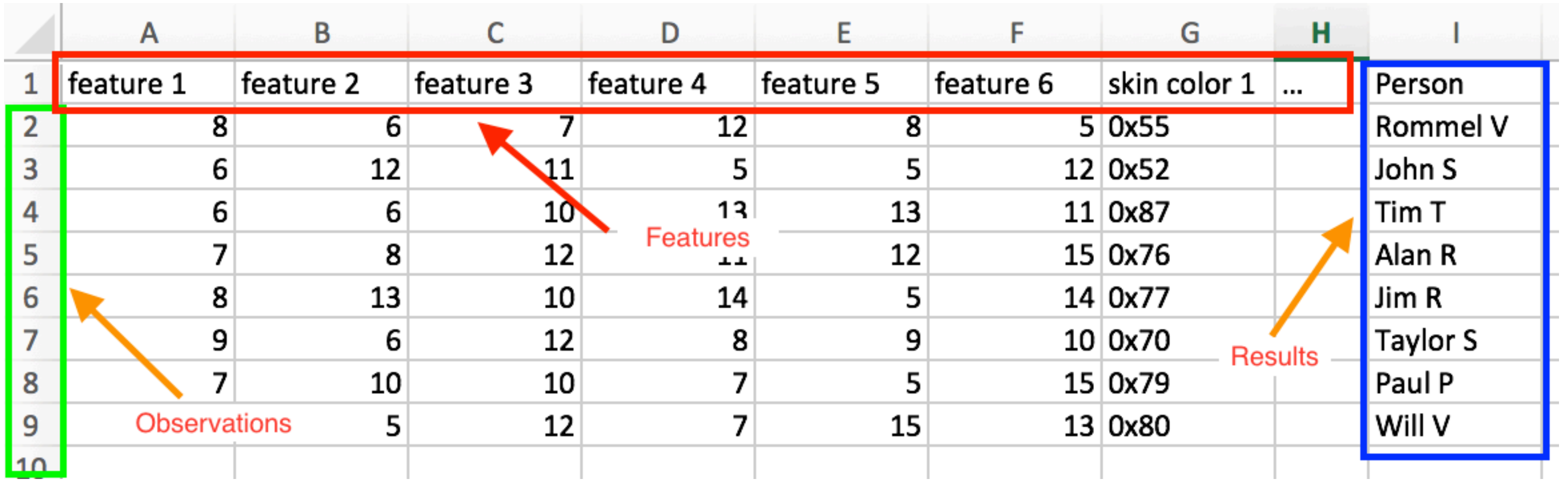
One image per row

Features, pixel value at position

Results



Face recognition

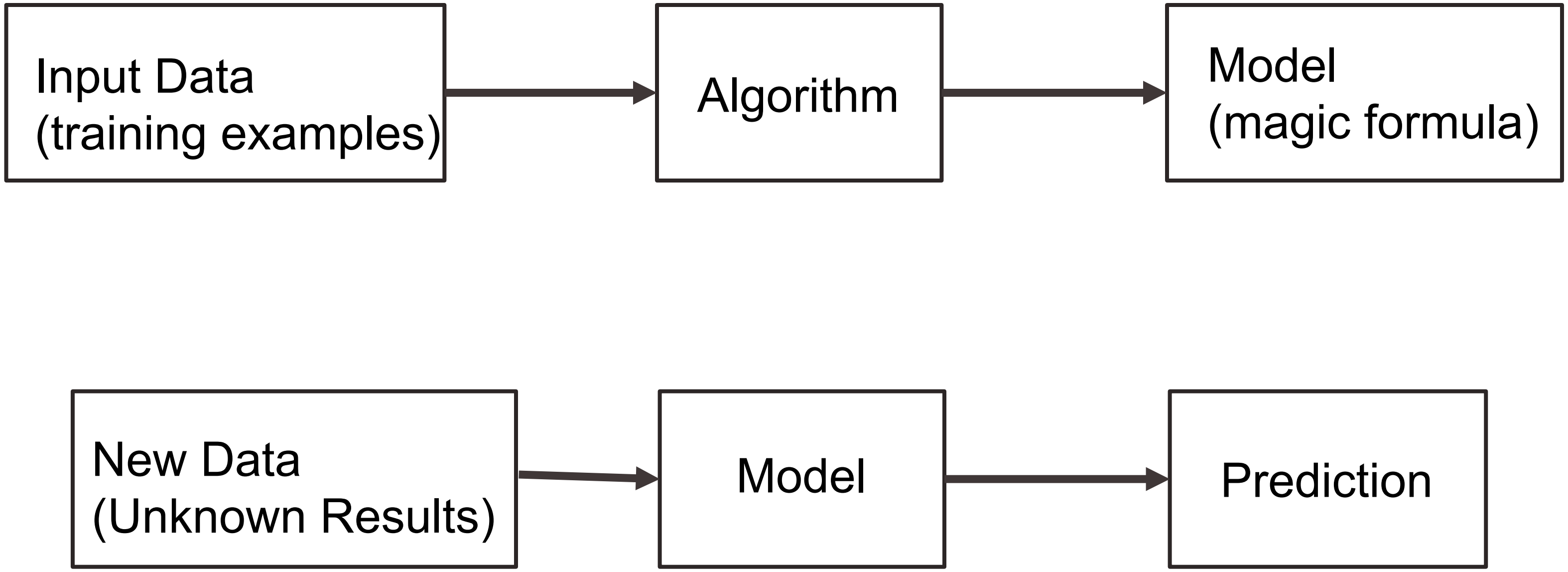


	A	B	C	D	E	F	G	H	I
1	feature 1	feature 2	feature 3	feature 4	feature 5	feature 6	skin color 1	...	Person
2	8	6	7	12	8	5	0x55		Rommel V
3	6	12	11	5	5	12	0x52		John S
4	6	6	10	13	13	11	0x87		Tim T
5	7	8	12	11	12	15	0x76		Alan R
6	8	13	10	14	5	14	0x77		Jim R
7	9	6	12	8	9	10	0x70		Taylor S
8	7	10	10	7	5	15	0x79		Paul P
9	[illegible]	5	12	7	15	13	0x80		Will V
10									

Key concept

- If you want to solve any Supervised Learning problem (regression or classification) your input is going to be a table where each row is an observation. The columns are your features, and the last column is the value that you want to determine or predict.

Supervised Learning Workflow



Input Data (training examples) → Algorithm → Model (magic formula)

New Data (Unknown Results) → Model → Prediction

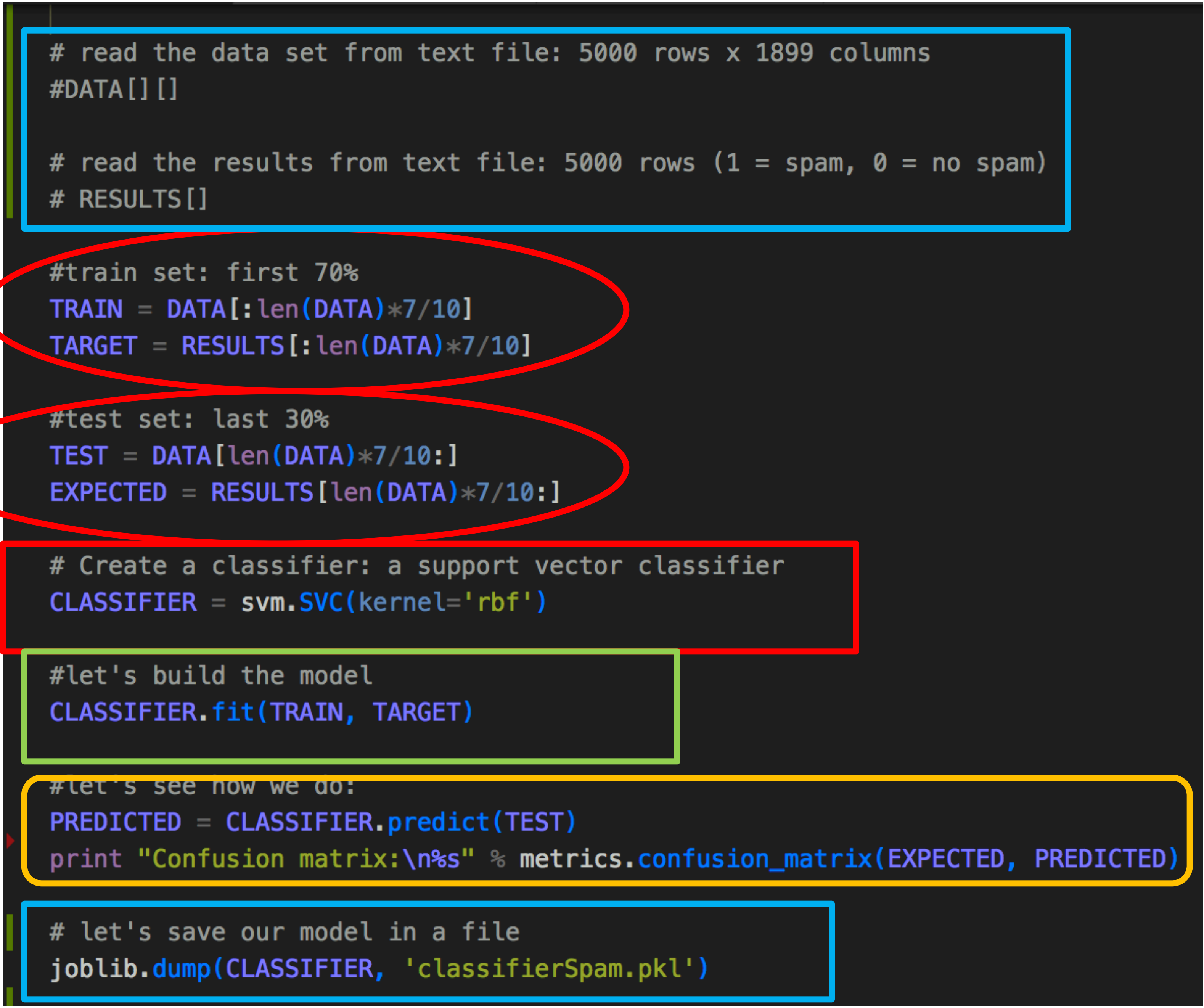
How to I get my model?

- So far what we’ve done is put our input into a table.
- Load them in Apache Spark, or any tool that can come up with a model.
- Build your own regression classifier in your favorite language and calculate the model that way
- Use Neural Networks/Deep Learning to find your model (overkill in this spam example)
- Use Support Vector Machines (powerful optimized algorithms) to calculate your model
- Use other mechanisms to output a model

Considerations

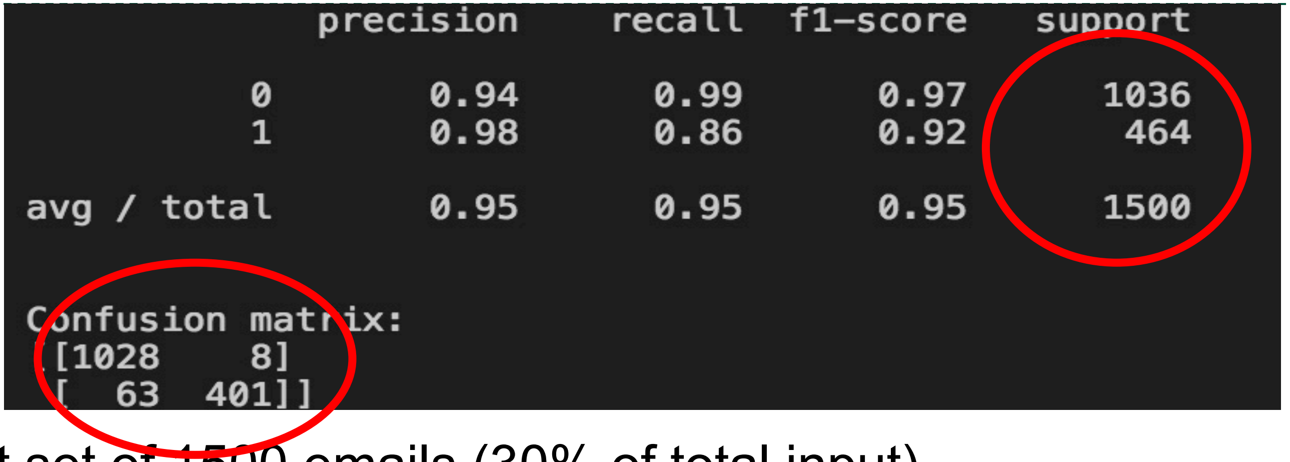
- If it’s going to run in production, should it be done in Matlab, Python or R for example?
- Can the model be generated in one language and exported to another?
- Is the model going to be run on the server side (e.g. Java) or client side (e.g JavaScript)
- Security considerations when pulling the data from the db into a different system for developing the model

Creating a model using support vector machines



```
# read the data set from text file: 5000 rows x 1899 columns
#DATA[][]

# read the results from text file: 5000 rows (1 = spam, 0 = no spam)
# RESULTS[]

#train set: first 70%
TRAIN = DATA[:len(DATA)*7/10]
TARGET = RESULTS[:len(DATA)*7/10]

#test set: last 30%
TEST = DATA[len(DATA)*7/10:]
EXPECTED = RESULTS[len(DATA)*7/10:]

# Create a classifier: a support vector classifier
CLASSIFIER = svm.SVC(kernel='rbf')

#let's build the model
CLASSIFIER.fit(TRAIN, TARGET)

#let's see how we do:
PREDICTED = CLASSIFIER.predict(TEST)
print "Confusion matrix:\n%s" % metrics.confusion_matrix(EXPECTED, PREDICTED)

# let's save our model in a file
joblib.dump(CLASSIFIER, 'classifierSpam.pkl')
```

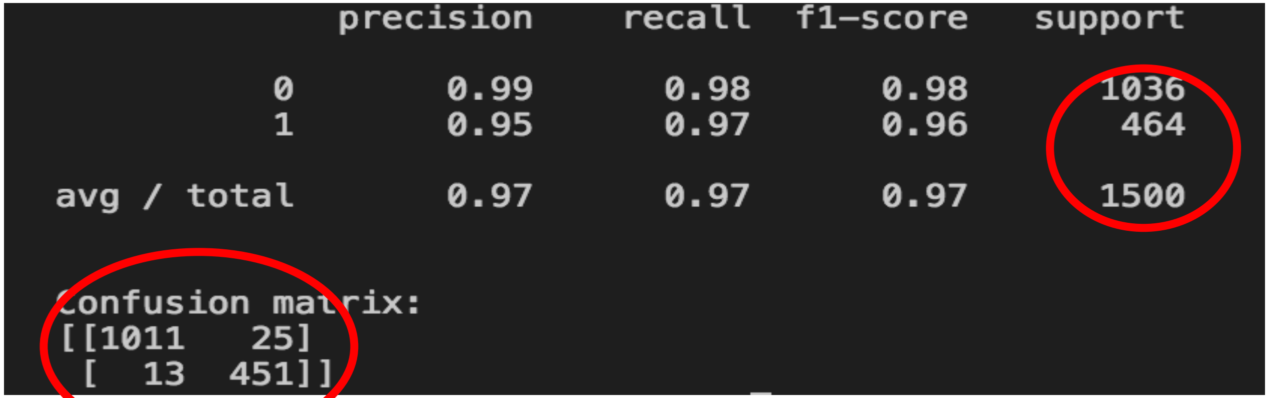
Results for model using RBF kernel



```
             precision    recall  f1-score   support

          0       0.94      0.99      0.97      1036
          1       0.98      0.86      0.92       464

avg / total       0.95      0.95      0.95      1500


Confusion matrix:
[[1028    8]
 [  63  401]]
```

- Test set of 1500 emails (30% of total input)
- 1036 were non-spam
- 464 were spam
- 1028 correctly id as non-spam (99%)
- 401 correctly id as spam (86%).
- 8 non-spam thrown into the spam folder (emails you will never see unless you look in the spam folder)
- 63 spam emails thrown into your Inbox (annoying)

Results for model using Linear kernel



```
             precision    recall  f1-score   support

          0       0.99      0.98      0.98      1036
          1       0.95      0.97      0.96       464

avg / total       0.97      0.97      0.97      1500


Confusion matrix:
[[1011   25]
 [  13  451]]
```

- Test set of 1500 emails (30% of total input)
- 1036 were non-spam
- 464 were spam
- 1011 correctly id as non-spam (98%).
- 451 correctly id as spam (97%).
- 25 non-spam thrown into the spam folder (emails you will never see unless you look in the spam folder)
- 13 spam emails thrown into your Inbox (annoying)

Spam detection: minimize false positives

- If detecting spam you want to make sure you don't send emails that are not spam to your spam folder, because likely you will not see those ever again.
- If emails with spam are positives, and non-spam emails are negative, then:
- You want to make sure your false positive is zero (non spam emails ending up if your spam folder) even at the expense of having a large number of false negatives (spam emails ending up in your inbox)

Detecting Fraud: minimize false negatives

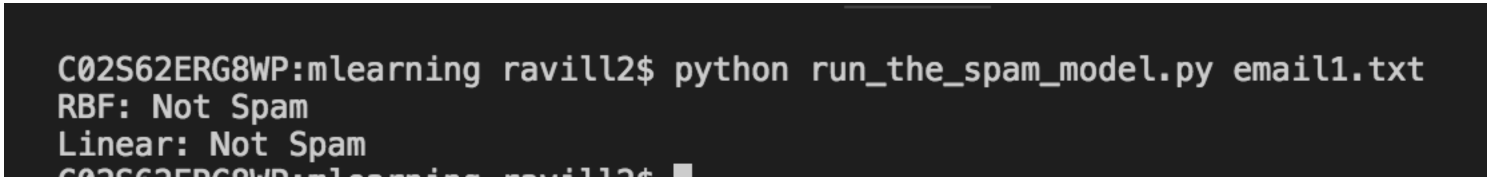
- If detecting fraud you want to make sure you detect all fraudulent transactions.
- If fraud are positive transactions and non-fraud are negative transactions, then:
- You want to make sure your false negative is zero (fraudulent transactions missed) even at the expense of having a large number of false positives (non fraudulent transactions flagged as fraud)
- This because you can always apologize to the clients and for example give them a gift certificate for the inconvenience.
- A fraudulent transaction not detected is more costly than dozens of gift certificates

Email1, let’s see what the models say

Anyone knows how much it costs to host a web portal ?

Well, it depends on how many visitors you're expecting. This can be anywhere from less than 10 bucks a month to a couple of $100. You should checkout http://www.rackspace.com/ or perhaps Amazon EC2 if you’re running something big.

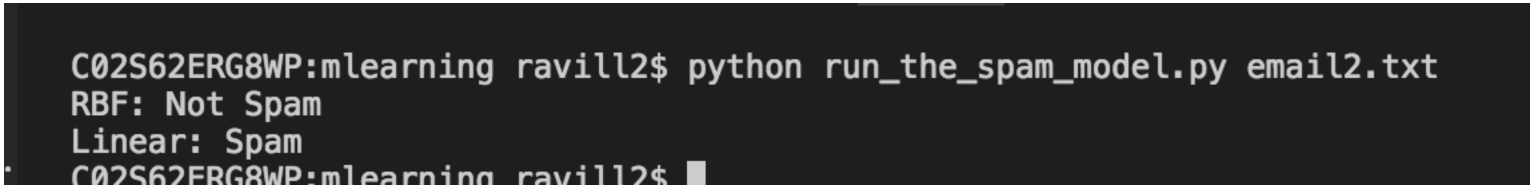
To unsubscribe yourself from this mailing list, send an email to:groupname-unsubscribe@egroups.com



```
C02S62ERG8WP:mlearning ravill2$ python run_the_spam_model.py email1.txt
RBF: Not Spam
Linear: Not Spam
C02S62ERG8WP:mlearning ravill2$ █
```

Email 2

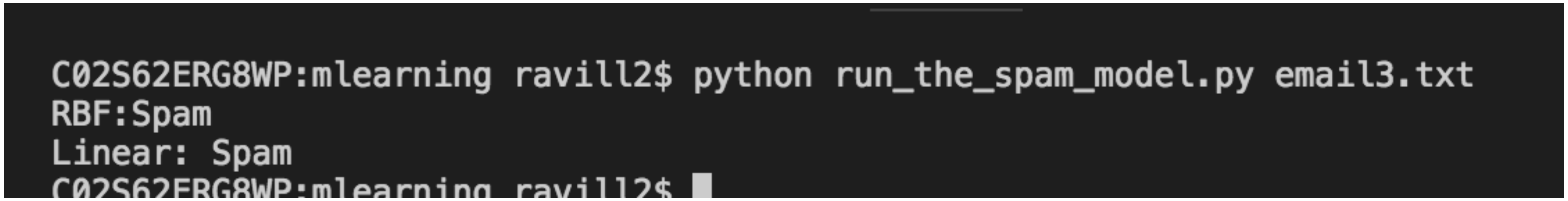
- Folks, my first time posting - have a bit of Unix experience, but am new to Linux. Just got a new PC at home - Dell box with Windows XP. Added a second hard disk for Linux. Partitioned the disk and have installed Suse 7.2 from CD, which went fine except it didn't pick up my monitor. I have a Dell branded E151FPp 15" LCD flat panel monitor and a nVidia GeForce4Ti4200 video card, both of which are probably too new to feature in Suse's default set. I downloaded a driver from the nVidia website and installed it using RPM. Then I ran Sax2 (as was recommended in some postings I found on the net), but it still doesn't feature my video card in the available list. What next? Another problem. I have a Dell branded keyboard and if I hit Caps-Lock twice, the whole machine crashes (in Linux, not Windows) - even the on/off switch is inactive, leaving me to reach for the power cable instead. If anyone can help me in any way with these probs., I'd be really grateful -I've searched the 'net but have run out of ideas. Or should I be going for a different version of Linux such as RedHat? Opinions welcome. Thanks a lot, Peter



```
C02S62ERG8WP:mlearning ravill2$ python run_the_spam_model.py email2.txt
RBF: Not Spam
Linear: Spam
C02S62ERG8WP:mlearning ravill2$
```

Email 3

- Do You Want To Make $1000 Or More Per Week? If you are a motivated and qualified individual - I will personally demonstrate to you a system that will make you $1,000 per week or more! This is NOT mlm. Call our 24 hour pre-recorded number to get the details.   000-456-789 I need people who want to make serious money. Make the call and get the facts. Invest 2 minutes in yourself now! 000-456-789



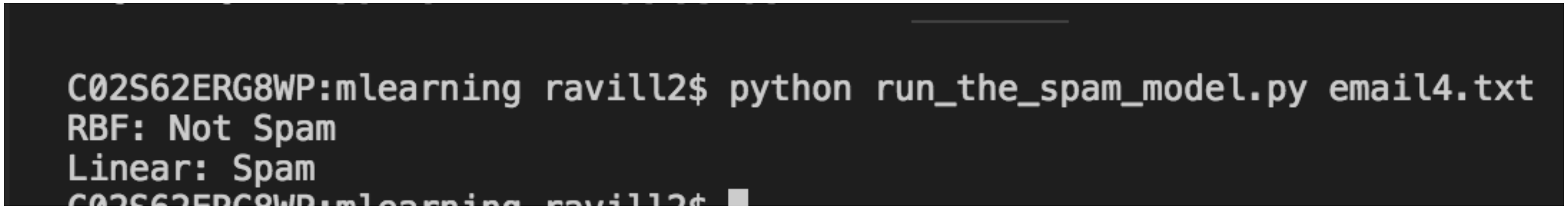
```
C02S62ERG8WP:mlearning ravill2$ python run_the_spam_model.py email3.txt
RBF:Spam
Linear: Spam
C02S62ERG8WP:mlearning ravill2$ 
```

Email 4

Best Buy Viagra Generic Online

Viagra 100mg x 60 Pills $125, Free Pills & Reorder Discount, Top Selling 100% Quality & Satisfaction guaranteed! We accept VISA, Master & E-Check Payments, 90000+ Satisfied Customers!

http://medphysitcstech.ru



```
C02S62ERG8WP:mlearning ravill2$ python run_the_spam_model.py email4.txt
RBF: Not Spam
Linear: Spam
C02S62ERG8WP:mlearning ravill2$ █
```

Final Thoughts

- Machine Learning is an always expanding field, specially with the advent of Deep Learning. There will be some exciting developments in the future that we want to be part of

Thank you, thank you….

References

- https://en.wikipedia.org/wiki/Machine_learning
- https://www.coursera.org/learn/machine-learning/home/week/1
- https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/
- Spam dataset from Coursera, Machine Learning training materials.
- Pictures from all over the Internet