Fundação **Casa de Rui Barbosa**

Iniciativas em Humanidades Digitais

Renato Rocha Souza
renato.souza@fgv.br

Using artificial intelligence to identify state secrets

FGV EMAp
ESCOLA DE MATEMÁTICA APLICADA

FGV CPDOC

A Twofold Mission

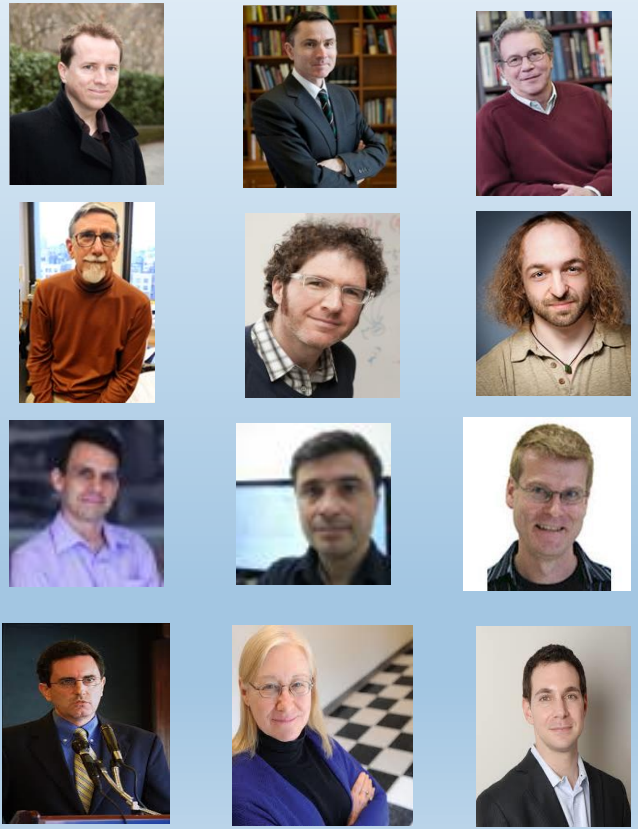
- Applied research to help government explore and manage sensitive information
- Tools to help people understand what information is released, and what needs to be protected

History as Data Science

We turn documents into data and develop tools to explore history.

A Coordinated Approach

A Research & Development Team of Data Scientists, Social Scientists, Engineers, and Web Developers, and Stakeholders



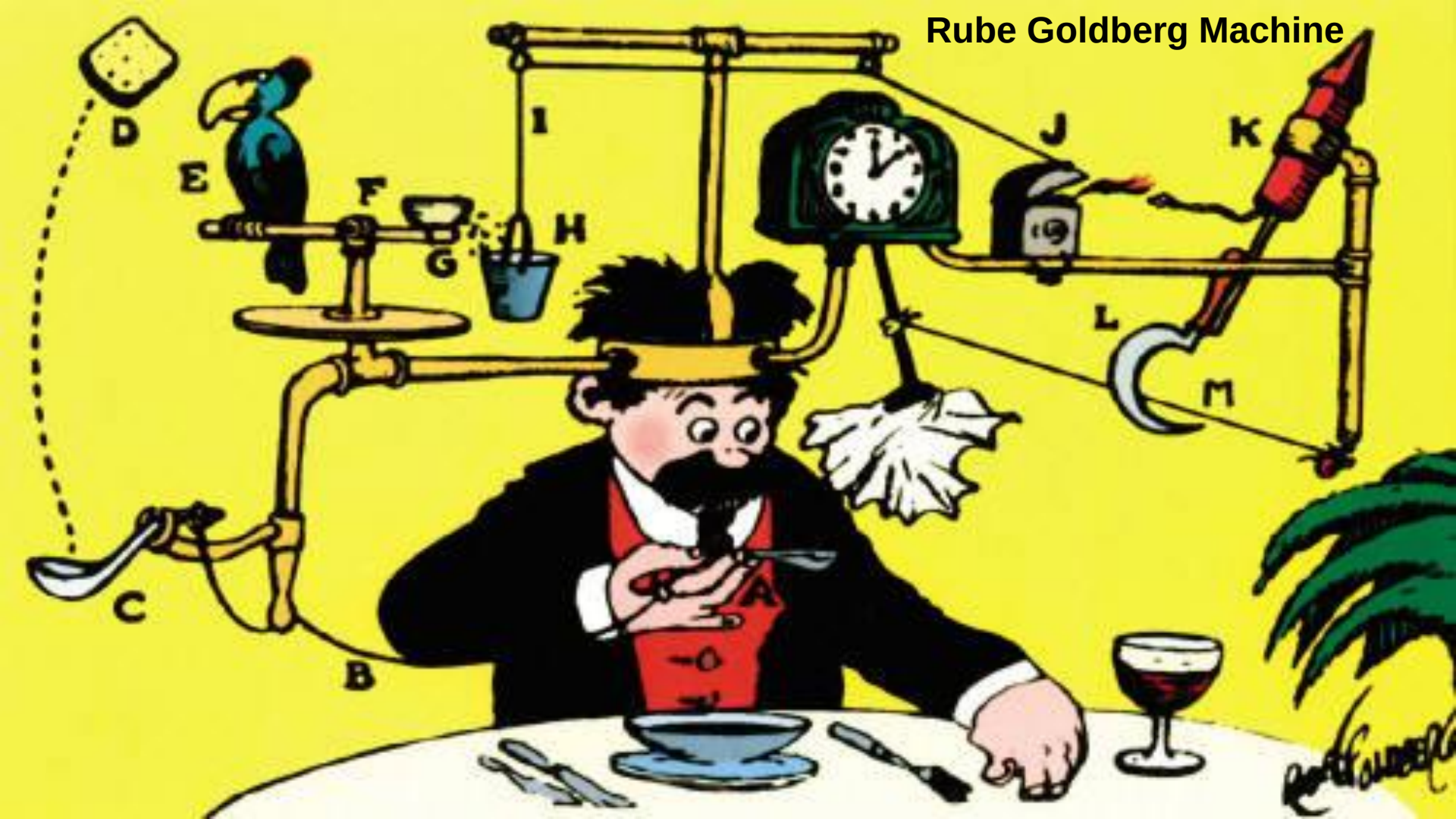
The Biggest Database of Declassified Documents

- **The Foreign Relations of the United States** (1945-1980). A curated collection of the ~80,000 most important declassified documents selected by State Department historians with access to every government department and agency.
- **The State Department Central Foreign Policy Files** (1973-1978). 1.7 million State Department Cables and metadata from ~500,000 more still classified cables and documents delivered by diplomatic pouch.
- **Henry Kissinger Telephone Conversations** (1973-1976)**.** 4.5 thousand transcripts of Kissinger Telephone Conversations during his tenure as Secretary of State.
- **The Hillary Clinton Emails** (2009-2012). As of November release, 16,246 email chains with a total of 40,737 individual messages
- **Other Collections To Come:**
  - President's Daily Briefs
  - FBI "Vault" of FOIA'ed documents
  - NATO Archives
  - Aramis (UK Foreign Office Cables 1992-2000)

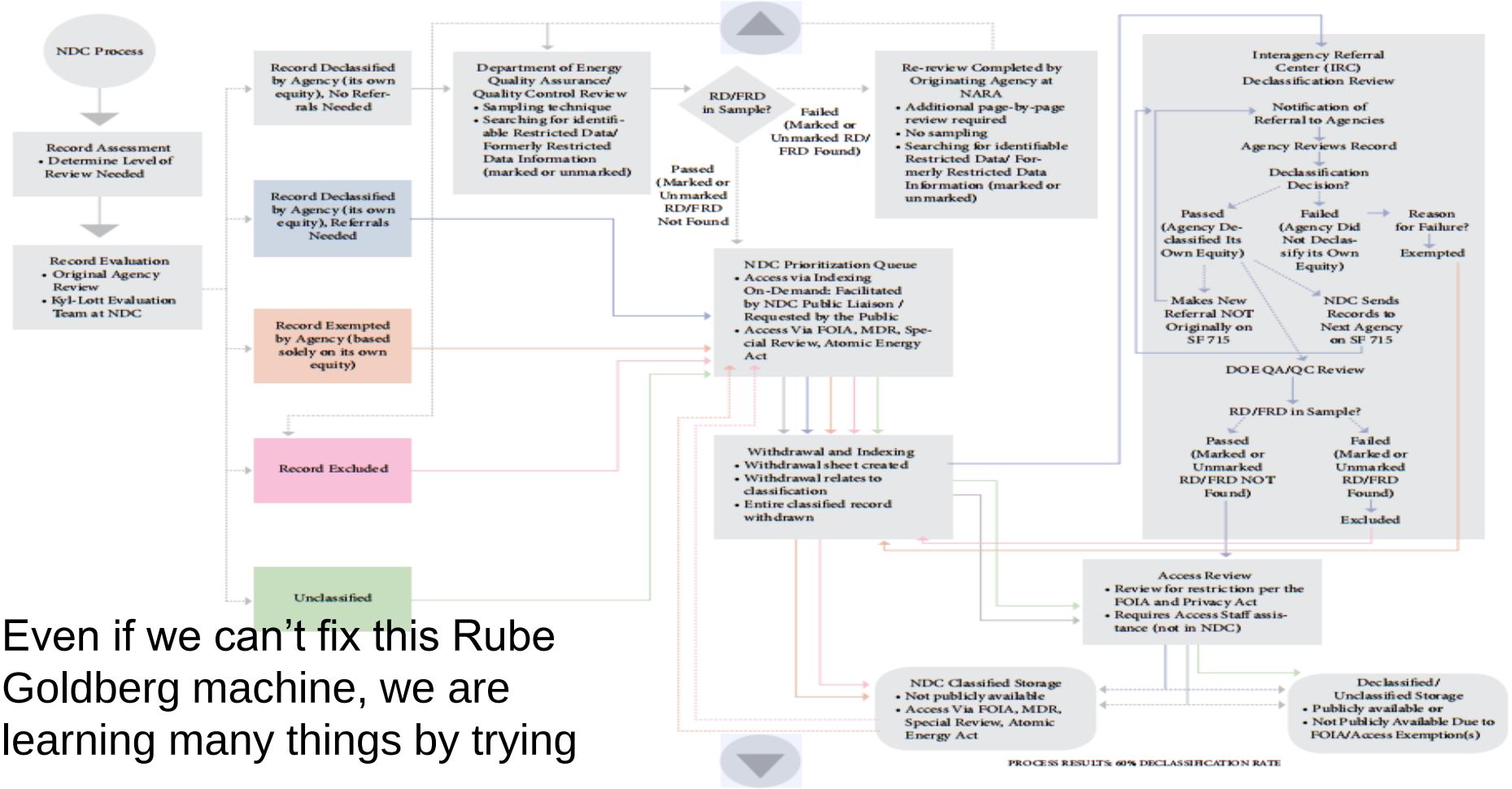
IV: Predicting Classification

- Data from State Dept. Cables collection, 1973-1978
- Use Machine Learning techniques to predict the classification levels, which are tagged in the documents

Rube Goldberg Machine

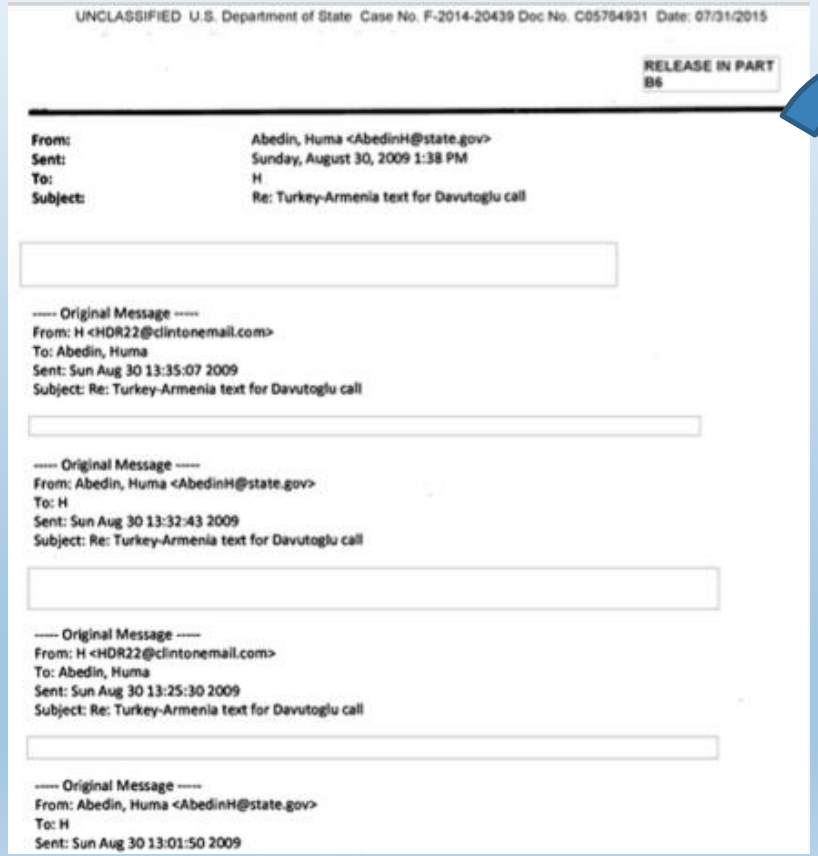


Our Focus: Tools to Manage Classification and Declassification Risks



Even if we can't fix this Rube Goldberg machine, we are learning many things by trying

From documents to Formatted Data



UNCLASSIFIED U.S. Department of State Case No. F-2014-20439 Doc No. C05764931 Date: 07/31/2015

RELEASE IN PART
B6

From: Abedin, Huma <AbedinH@state.gov>
Sent: Sunday, August 30, 2009 1:38 PM
To: H
Subject: Re: Turkey-Armenia text for Davutoglu call

----- Original Message -----
From: H <HDR22@clintonemail.com>
To: Abedin, Huma
Sent: Sun Aug 30 13:35:07 2009
Subject: Re: Turkey-Armenia text for Davutoglu call

----- Original Message -----
From: Abedin, Huma <AbedinH@state.gov>
To: H
Sent: Sun Aug 30 13:32:43 2009
Subject: Re: Turkey-Armenia text for Davutoglu call

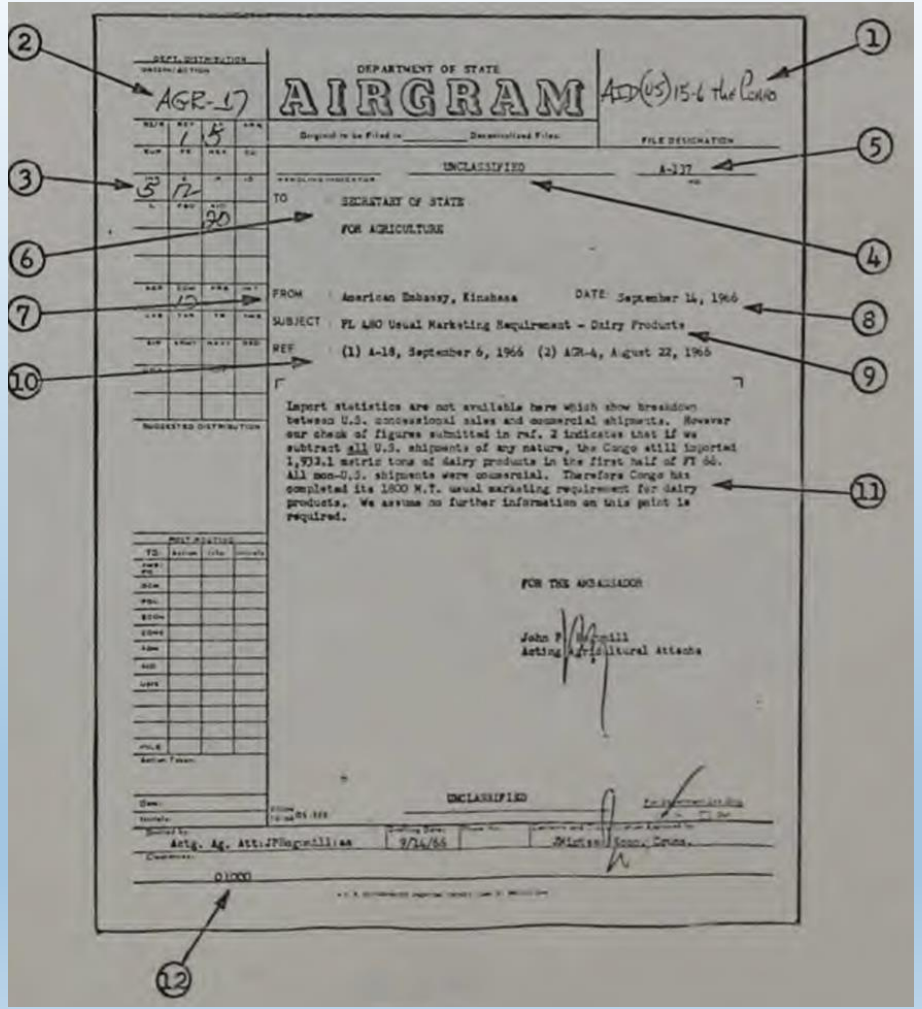
----- Original Message -----
From: H <HDR22@clintonemail.com>
To: Abedin, Huma
Sent: Sun Aug 30 13:25:30 2009
Subject: Re: Turkey-Armenia text for Davutoglu call

----- Original Message -----
From: Abedin, Huma <AbedinH@state.gov>
To: H
Sent: Sun Aug 30 13:01:50 2009

	id	origclass	body	concepts	subject	from	to	date	office	channel	type	orighand	origphand	pha
0	1974DARES00479	LIMITED OFFICIAL USE	president nyerere accepting long-standing invi...	president arts chief of state visits exhibits	nyerere visits residence	daressalaam	state	1974-02-14	action af	n/a	TE	n/a	n/a	n/a
1	1974SOFIA01580	UNCLASSIFIED	usia for iee ics in letter to ambassador mr ri...	speakers visits	possible voluntary speaker richard friedman us...	sofia	usia	1974-09-05	action cu	n/a	TE	n/a	n/a	n/a
2	1974STATE155445	UNCLASSIFIED	annette michelson ms annette michelson associa...	arts magazines editors visits	travel of asst editor art forum magazine annet...	state	moscow	1974-07-18	origin eur	n/a	TE	n/a	n/a	n/a
3	1975BANGKO24812	UNCLASSIFIED	as part of bicentennial program post will give...	arts cultural presentations	cu/arts american specialist grantee as part of...	bangkok	state	1975-11-25	action cu	n/a	TE	n/a	n/a	n/a
			charles		potential gift									

Identifying features

1. File Designation
2. Action Distribution
3. Information Distribution
4. Security Classification
5. Airgram Number
6. Addressees
7. Originating Post
8. Preparation Date
9. Subject
10. References
11. Text



Used Fields	Description
origclass	The original class of the cable (the classification target)
body	Full text of the cable
subject	Keywords of subjects dealt with in the document.
concepts	Concepts attributed to the document
TAGS	Traffic Analysis by Geography and Subject
from	Who/where sent the document.
to	Who/where received the document.
office	Which State Department office or bureau sent the document.
date	Document creation date

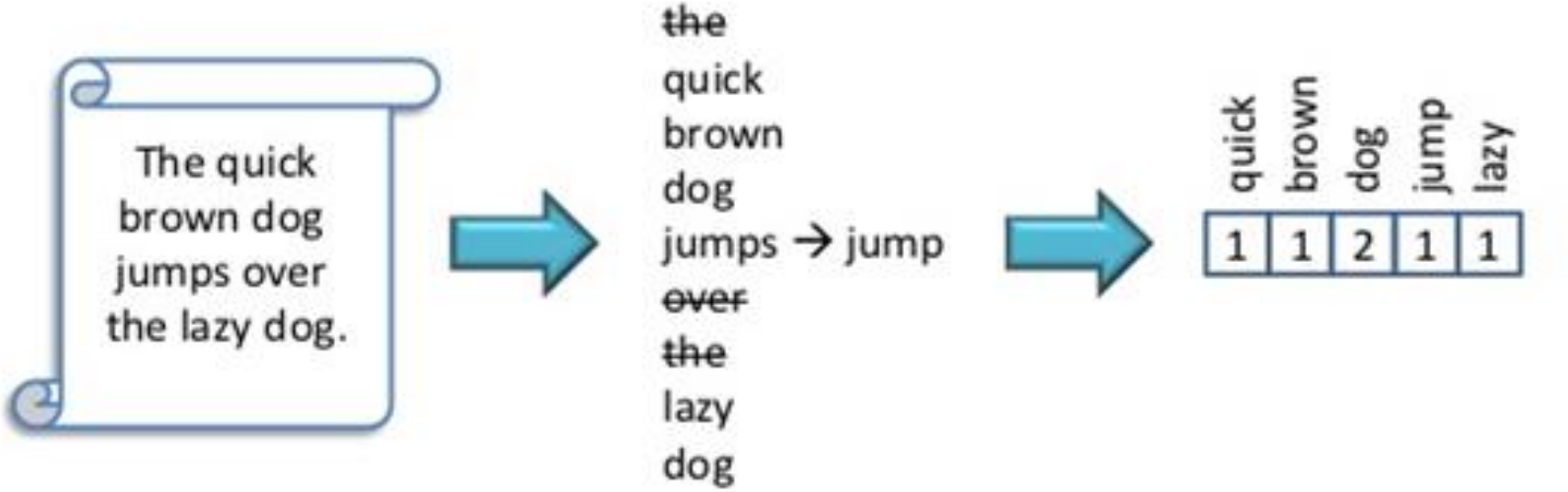
Situation	Total in Database	Unclassified	Limited Official Use	Confidential	Secret
declassified cables	**1.758.279**	**876.797**	**411.973**	**375.690**	**93.635**
Error messages for *body*	119.744	53.935	21.744	25.233	18.832
blank *body*	8282	2.726	1.645	1.924	1.987
blank or n/a *concepts*	634.967	445.300	114.507	65.502	9.658
blank or n/a *subject*	26.109	16.490	5.820	2.914	885
blank or n/a *from*	17	7	6	3	1
blank or n/a *to*	9.740	6.027	1.572	1.698	443
Used for classifier	**981.083**	**368.043**	**280.251**	**270.477**	**62.312**

Feature Engineering

- Hyphenation was eliminated from textual fields, as there were garbage from the original printed versions that were scraped from the web;
- Compound names of places in textual fields were aggregated, enabling them to be treated as a single token (i. e. NEW YORK was transformed to NEWYORK). They were present in all textual features, but that step was specially important in the case of the *from* and *to* fields, which represent the names of the embassies. These fields were aggregated under a new field *embassy*, for the vectorization process;
- Tokenization was made and all the trailing punctuation and words with length of 1 were eliminated. Underscores and hyphens in the middle of words were maintained;
- Stopwords were removed, using NLTK english stopwords list;
- Tests were made using stemmed forms of words, but it didn't enhance the performance and the stemming was discarded.
- The field *date* was transformed in a boolean field *weekday* - indicating whether the date fell in a weekend or not; and another field *year+month*, used to test hypothesis on the temporal series of cables regarding to classification windows. That doesn't prove useful, though, for the whole span, although could be promising for small periods of time.
- The fields *body*, *subject*, *concepts*, *tags*, *embassy* and *office* were added altogether in a new field/feature *all_text*, which was tested as an alternative to combining all the other features by concatenating those vectors, with very similar results.

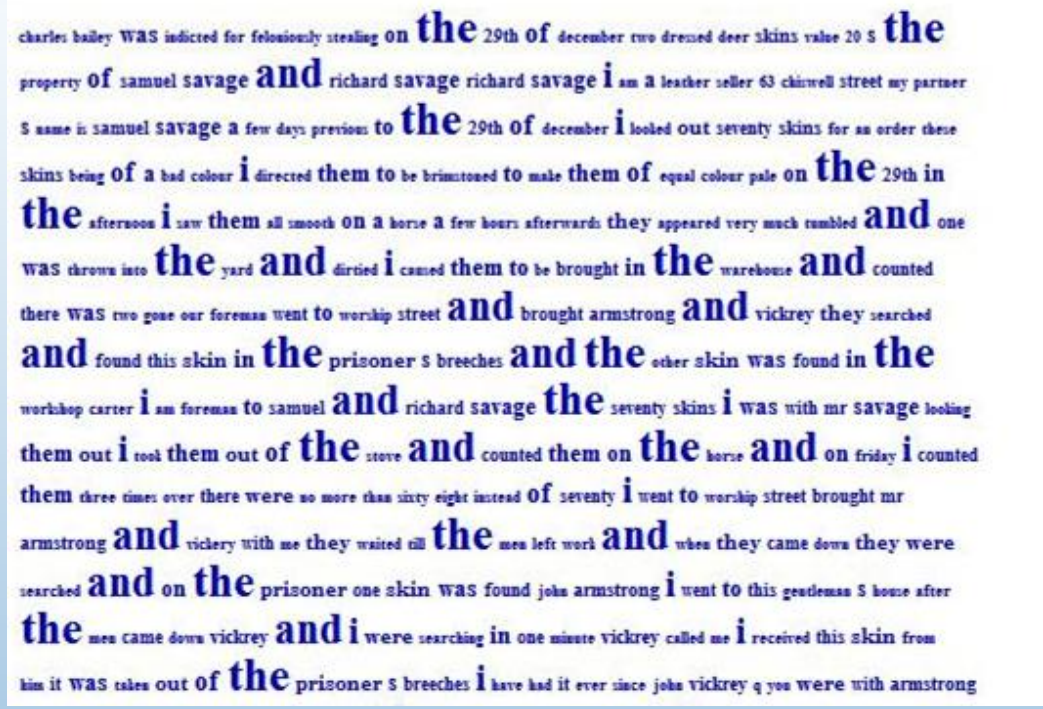
Bags of words

- Tokenize
- Remove stop words
- Lemmatize
- Compute weights

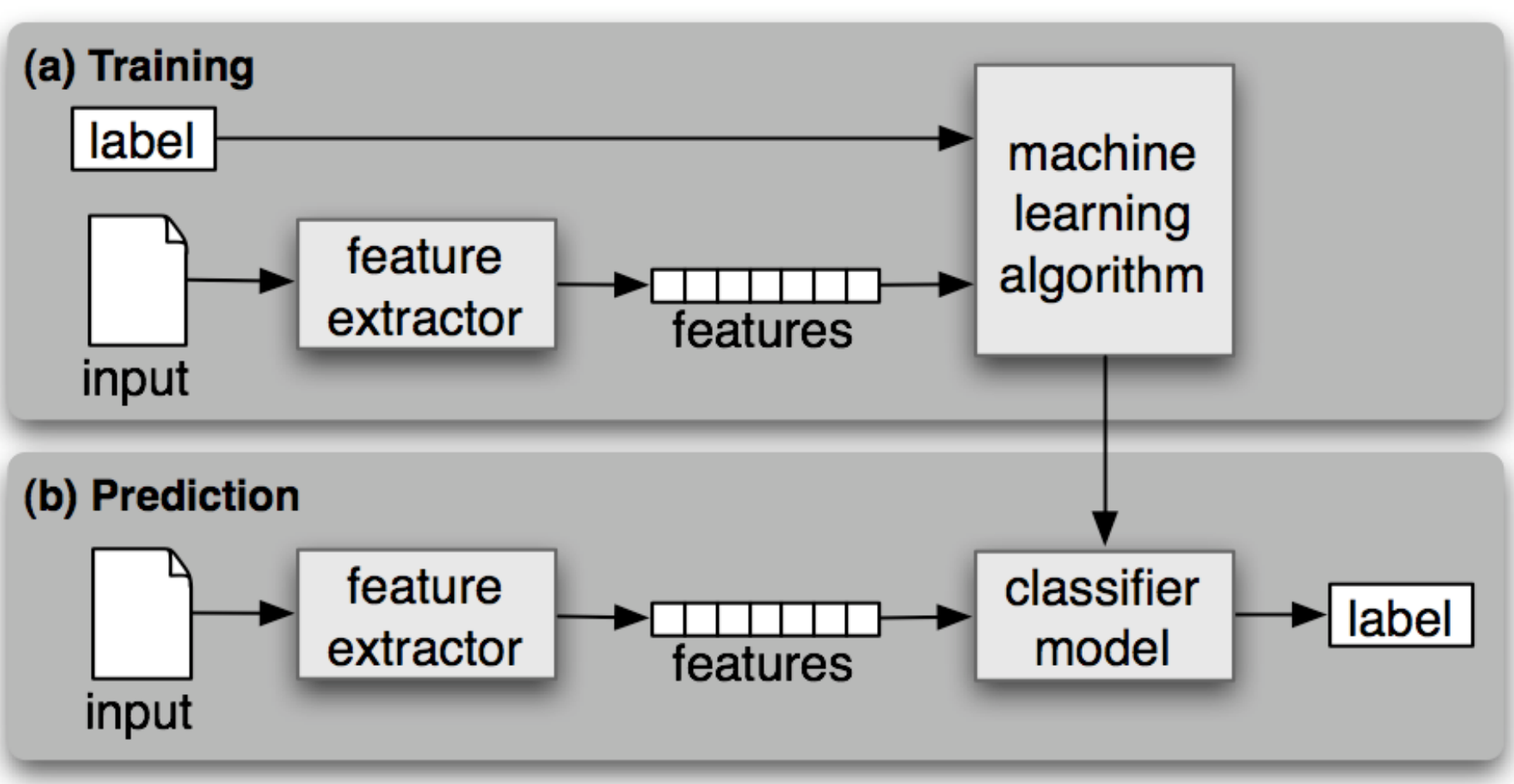


http://pt.slideshare.net/mgrcar/text-and-text-stream-mining-tutorial-15137759

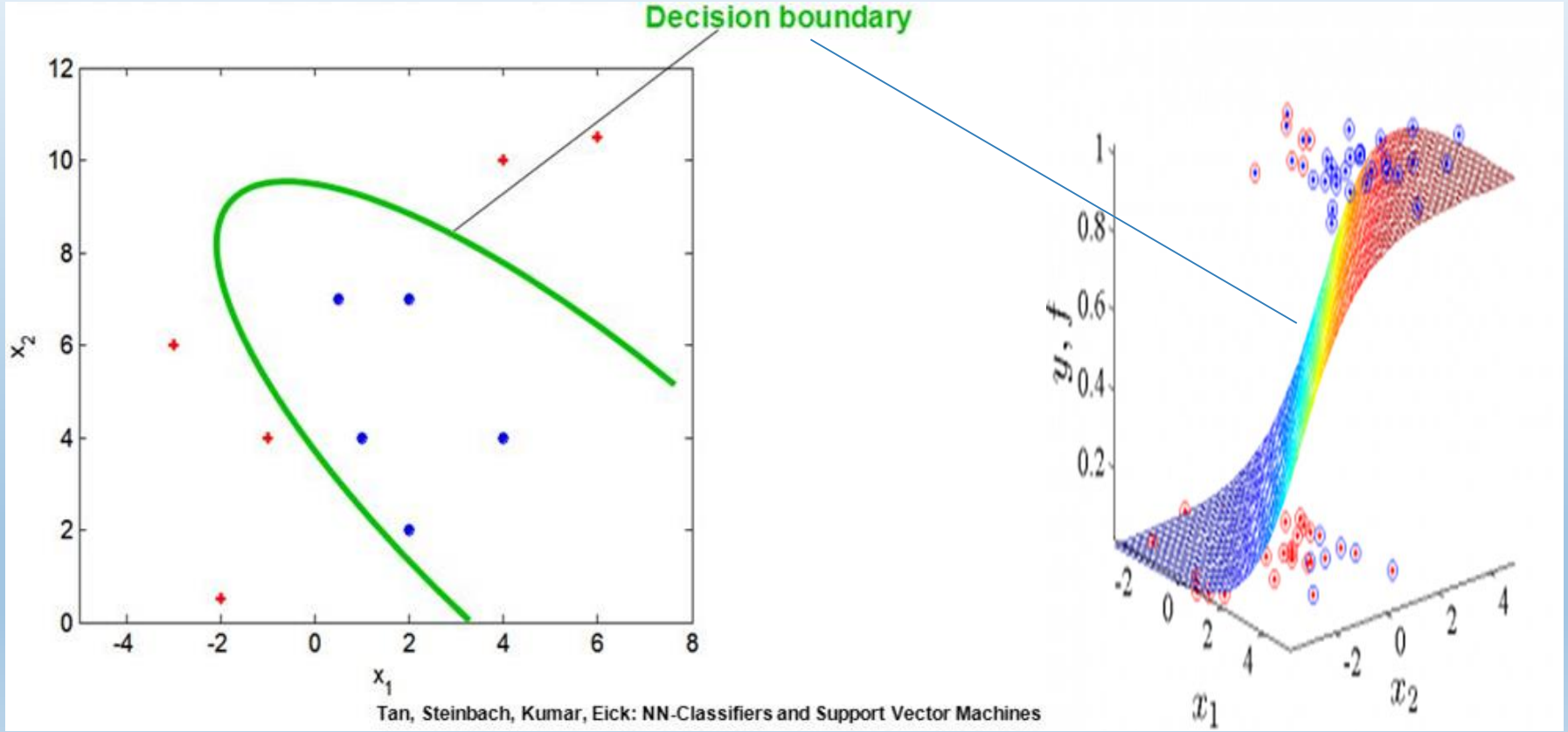
We have tested many alternate weighting schemes, as TfIDf

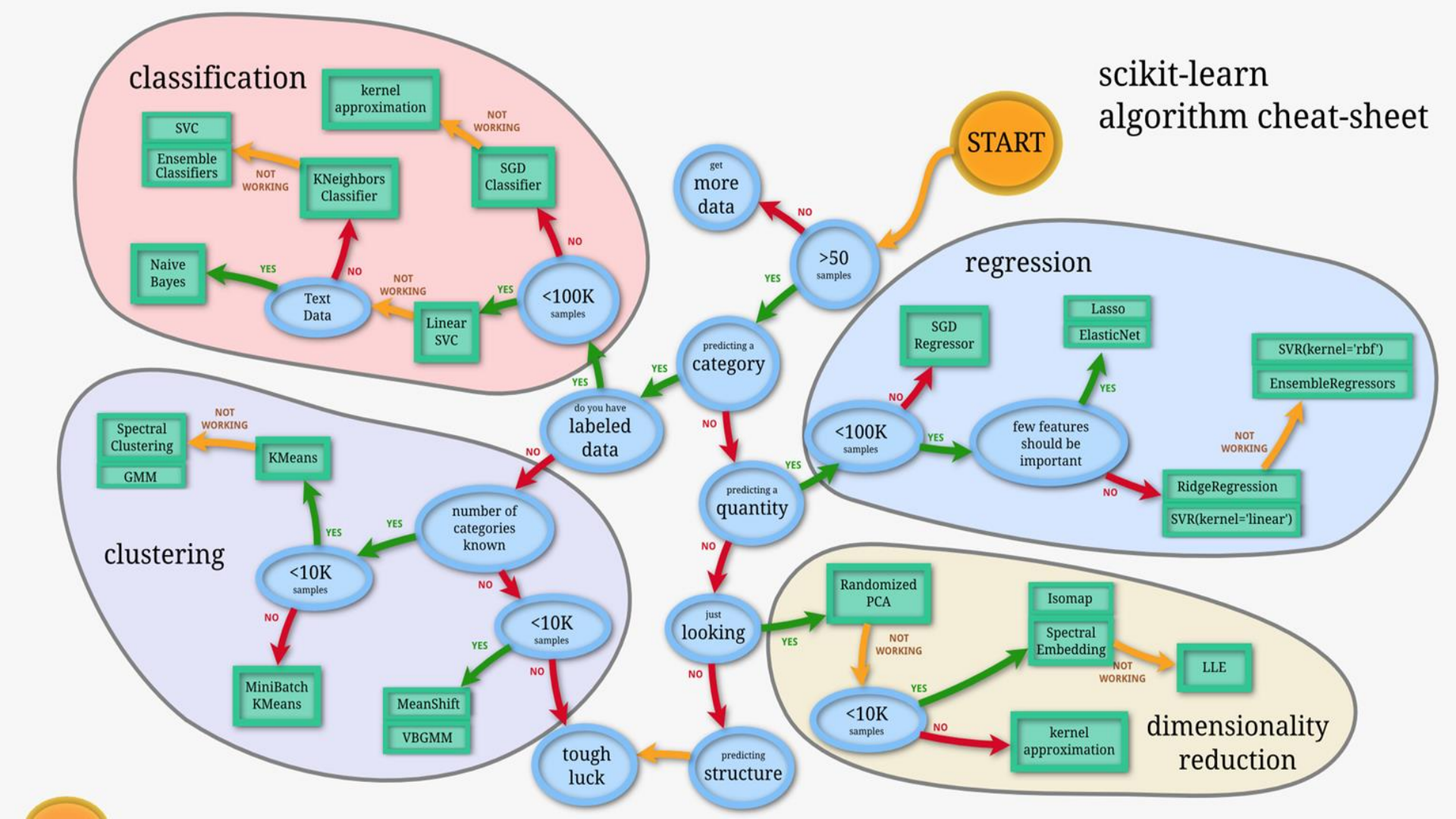


$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$


(a) Training
label
input
feature extractor
features
machine learning algorithm
(b) Prediction
input
feature extractor
features
classifier model
label

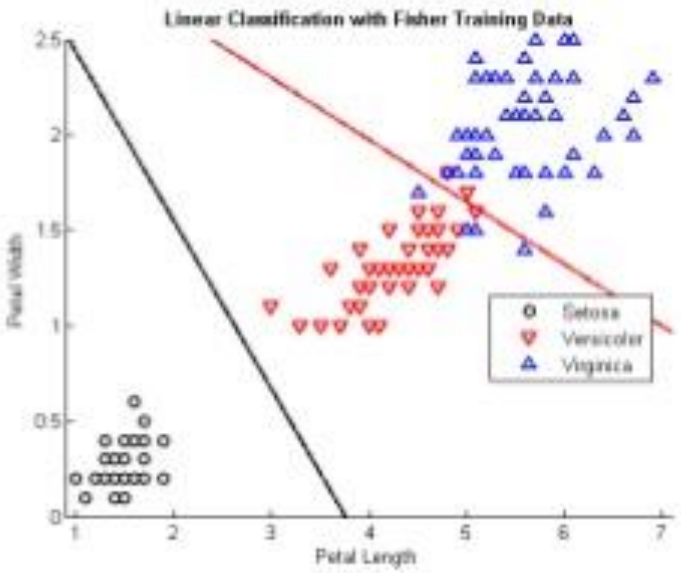
Classification in AI can be seem as analogous to “learning good decision boundaries” that separate the examples belonging to diferente classes in the data set



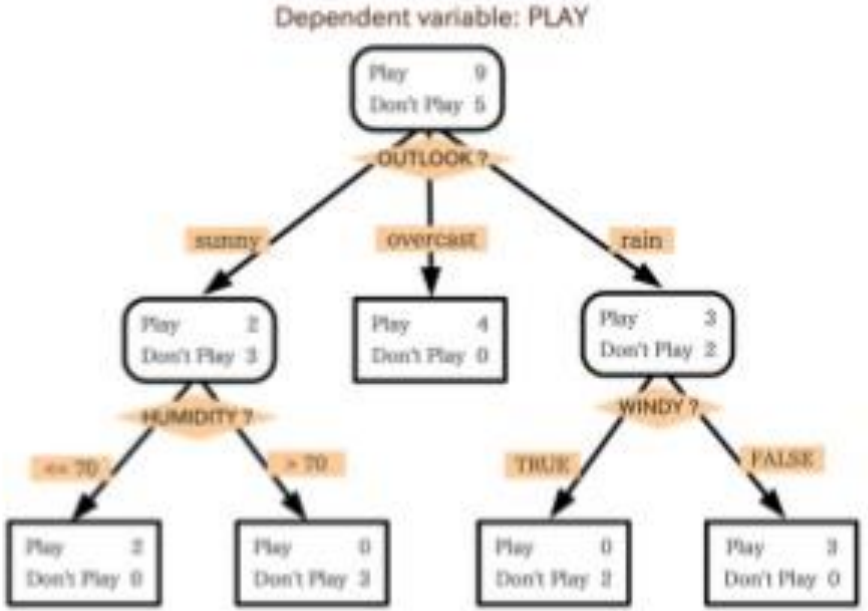

scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
YES
Linear SVC
NOT WORKING
Text Data
NO
Naive Bayes
YES
KNeighbors Classifier
NO
Ensemble Classifiers
NOT WORKING
SVC
SGD Classifier
NO
kernel approximation
NOT WORKING
classification
NO
number of categories known
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering
GMM
NO
<10K samples
NO
MiniBatch KMeans
YES
<10K samples
YES
MeanShift
VBGMM
NO
clustering
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck

Algorithms

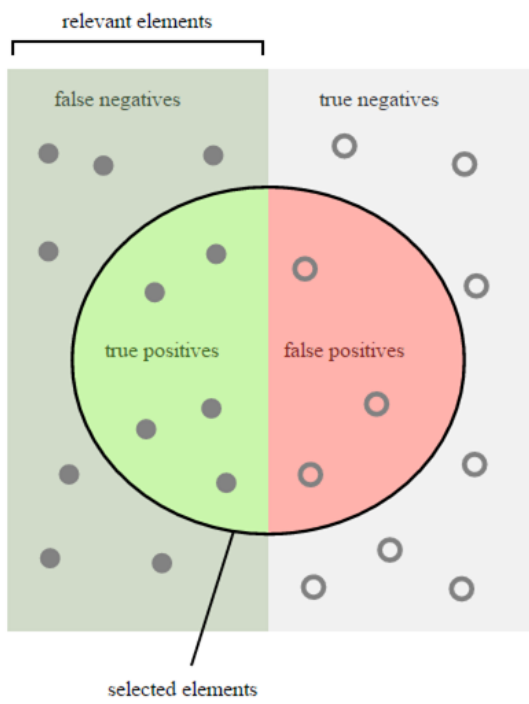
Linear Models



Decision Trees

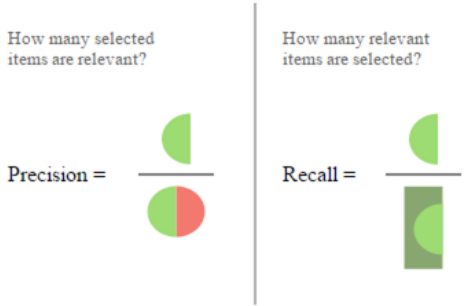


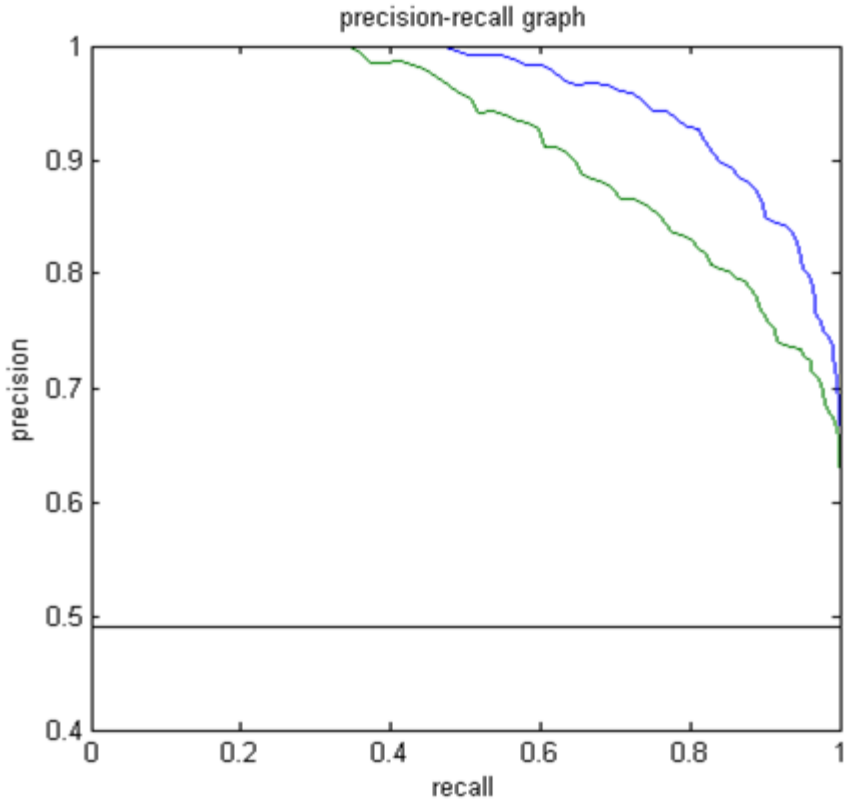
		Predicted condition	
	Total population	Predicted Condition positive	Predicted Condition negative
True condition	condition positive	**True positive**	**False Negative** (Type II error)
	condition negative	**False Positive** (Type I error)	**True negative**
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$
		False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$

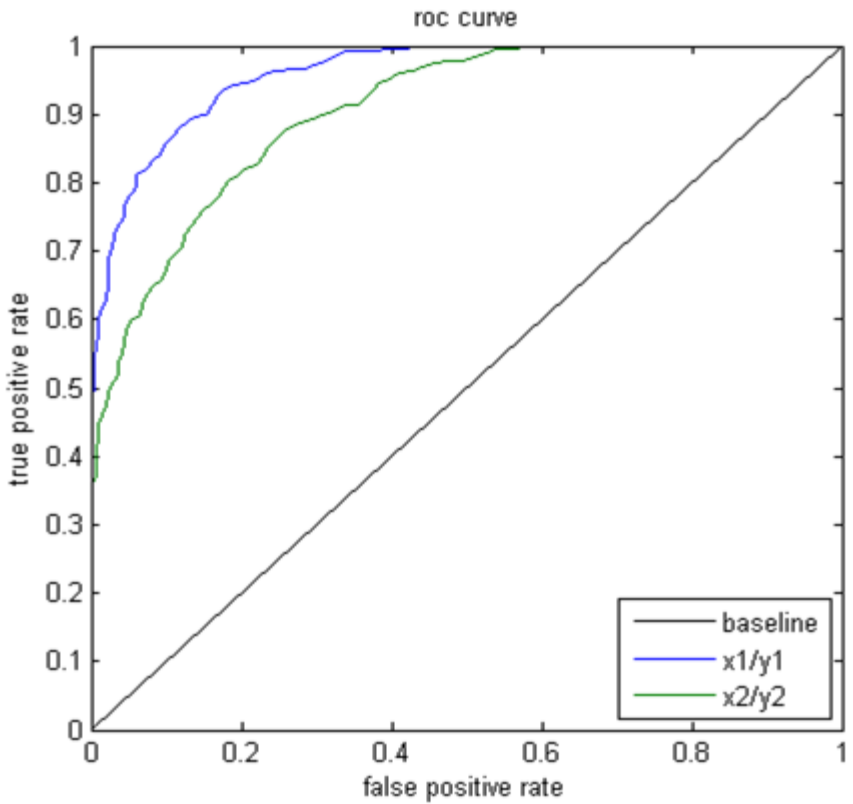


$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

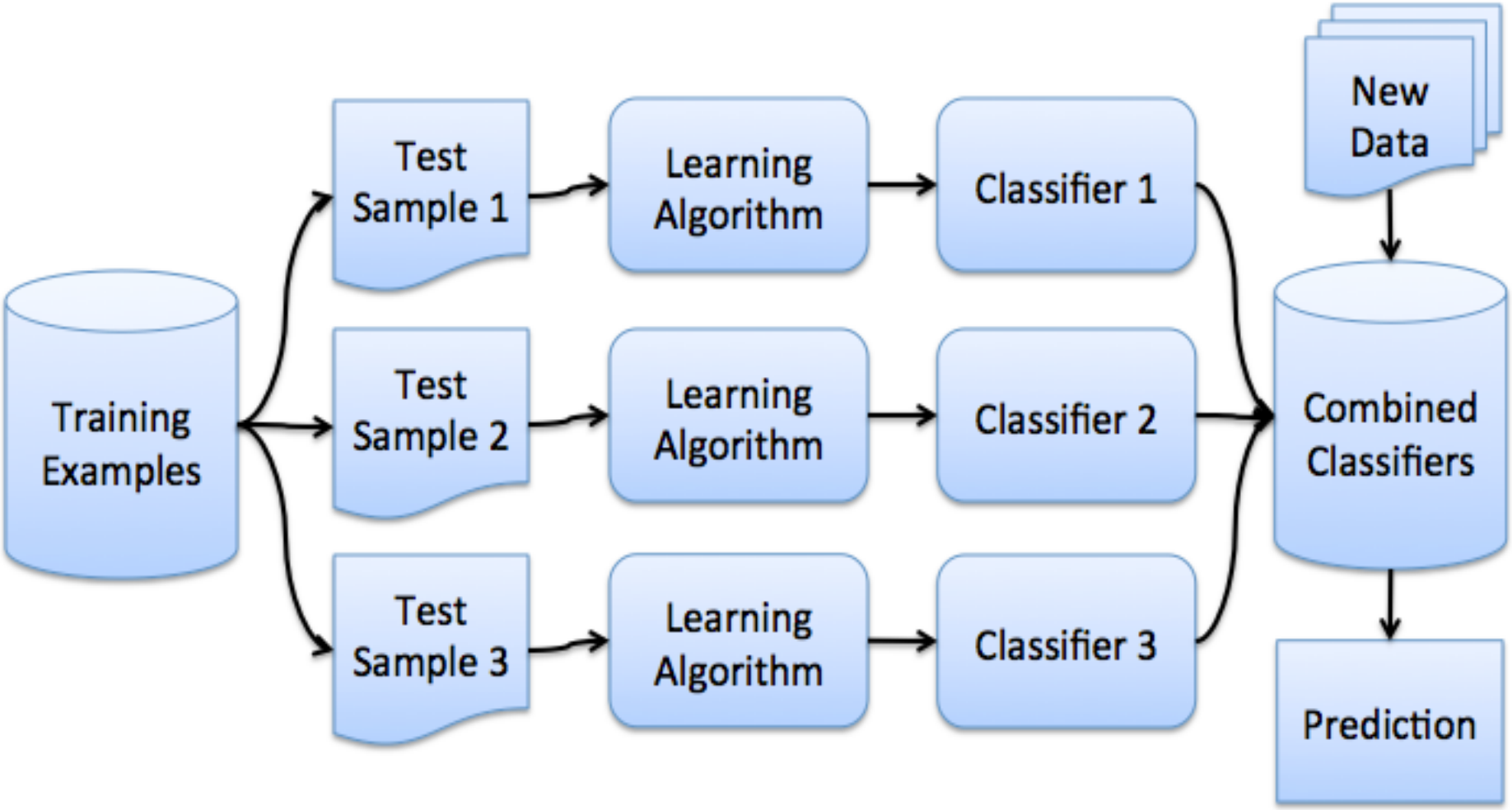


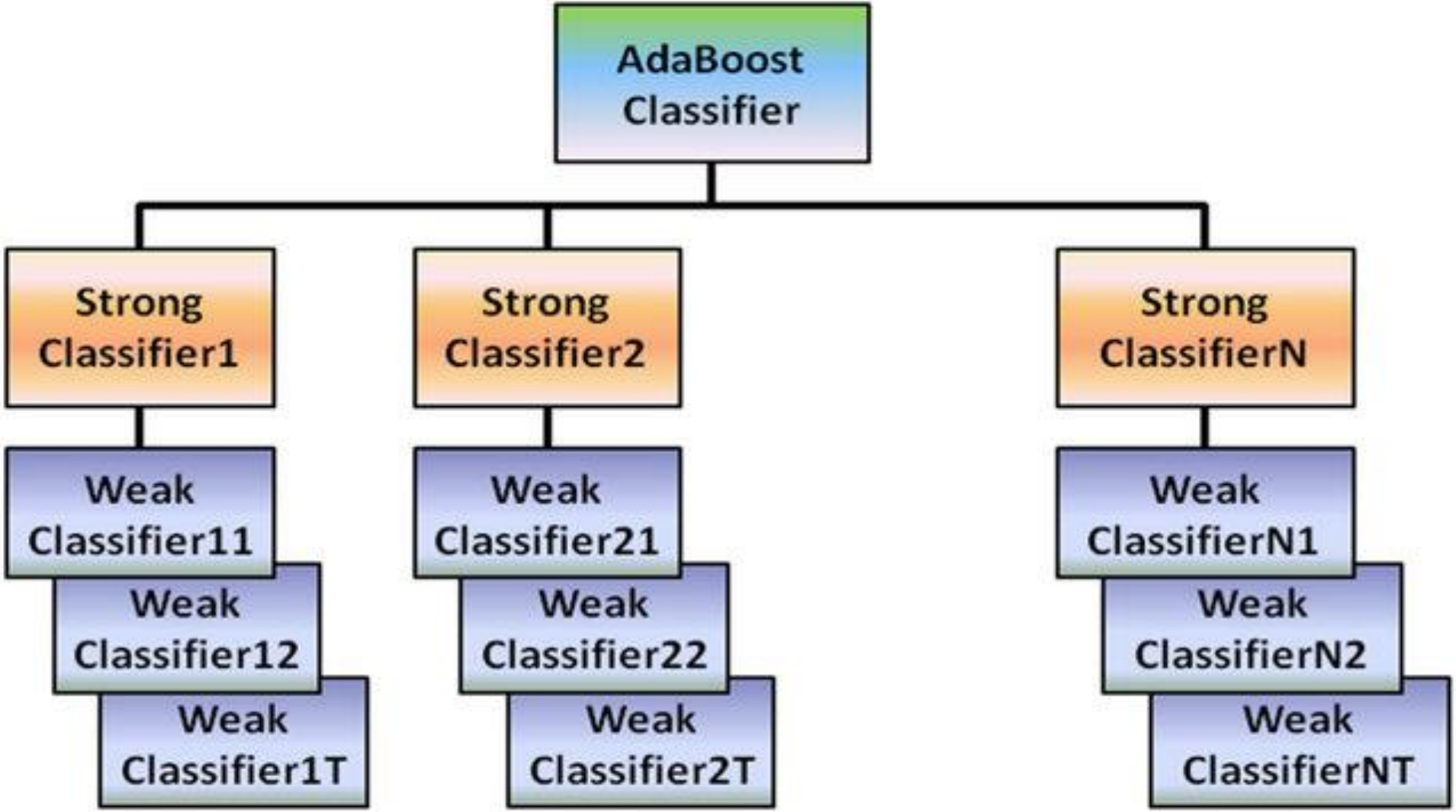

precision-recall graph
precision
recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0
0.2
0.4
0.6
0.8
1


roc curve
true positive rate
false positive rate
baseline
x1/y1
x2/y2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.2
0.4
0.6
0.8
1

Very ordinary results at first... ~0.75

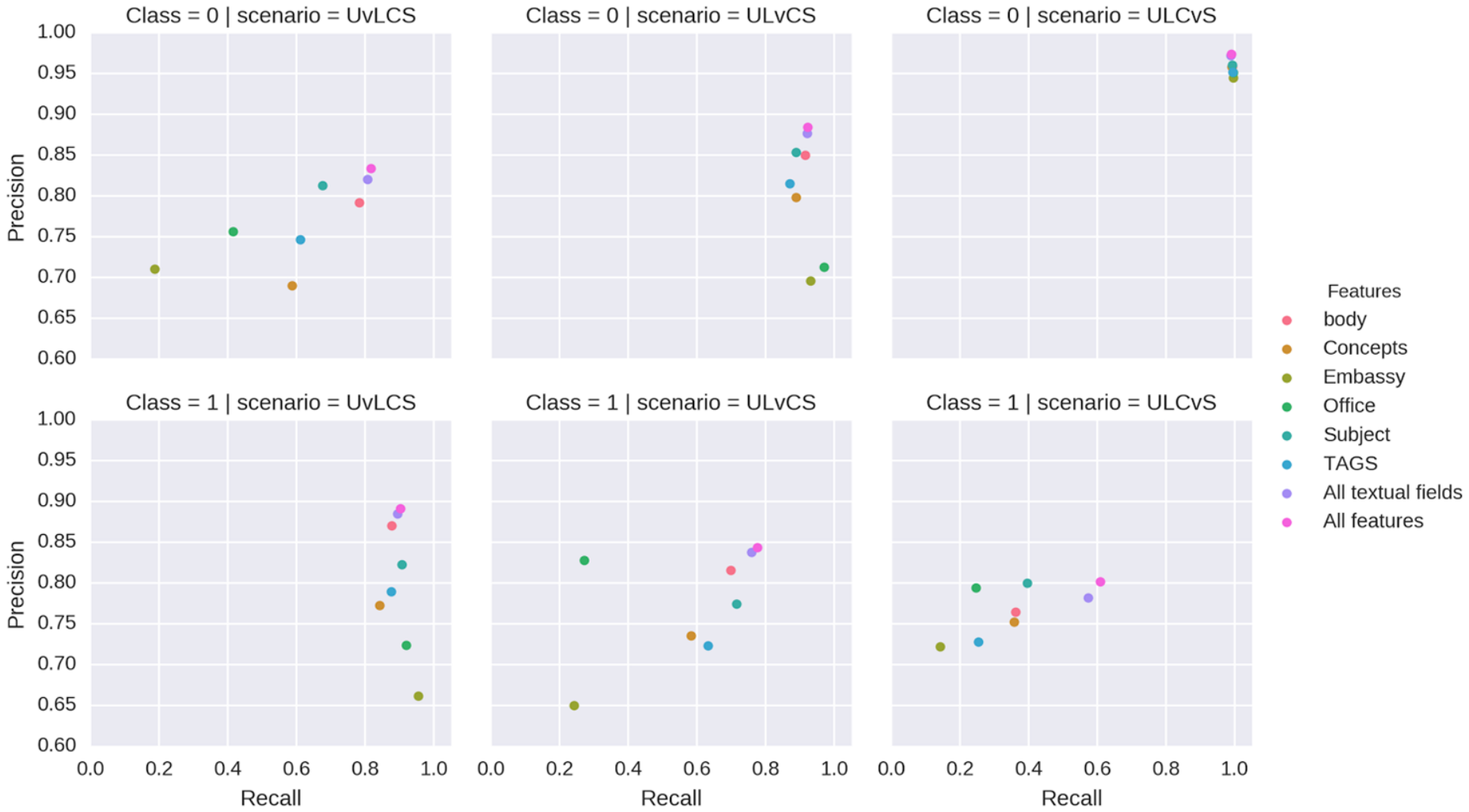
- Simple classifiers
- Scarce feature engineering
- Few data cleansing

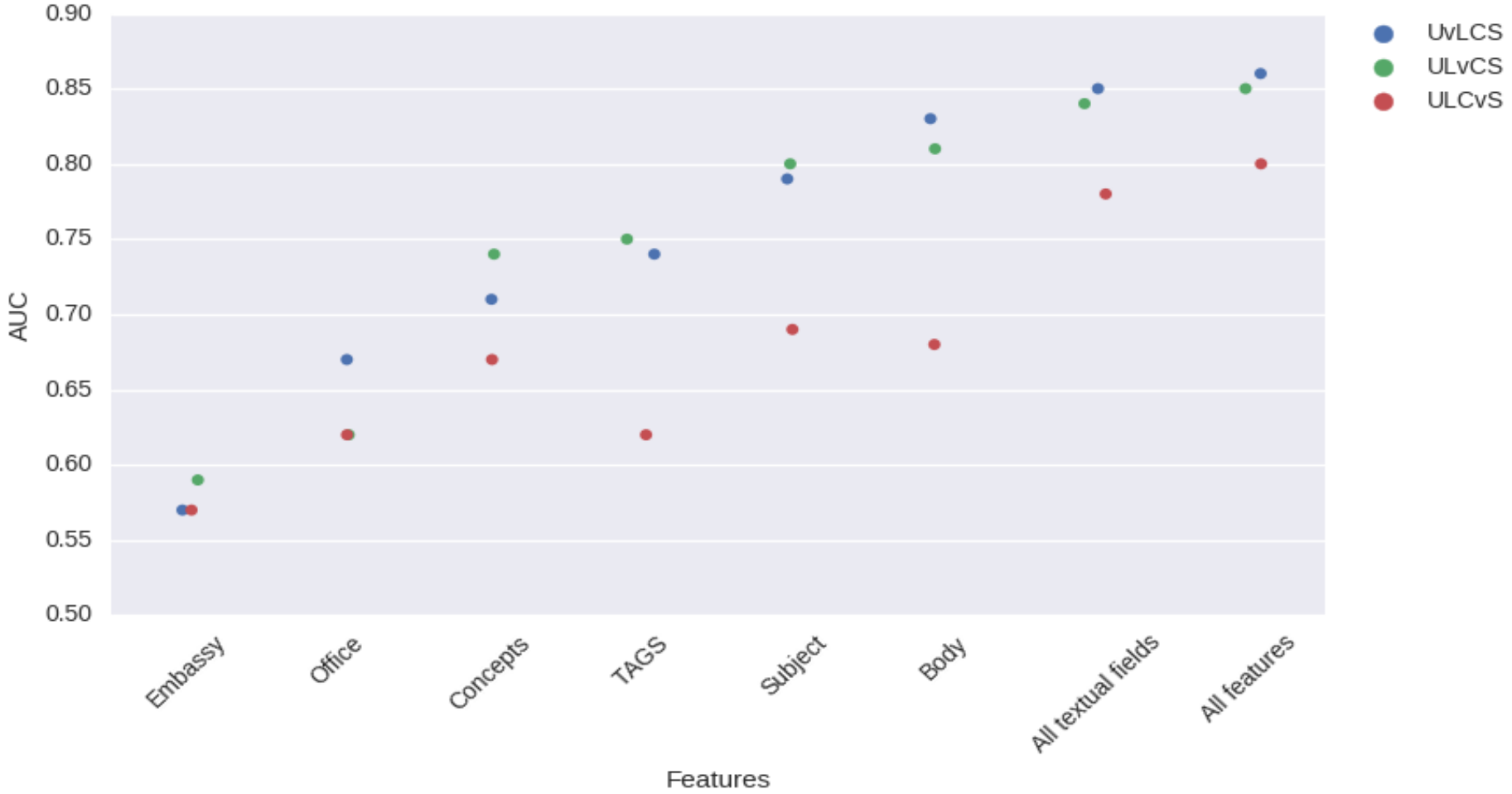

Training Examples
Test Sample 1
Test Sample 2
Test Sample 3
Learning Algorithm
Learning Algorithm
Learning Algorithm
Classifier 1
Classifier 2
Classifier 3
New Data
Combined Classifiers
Prediction


AdaBoost Classifier
Strong Classifier1
Strong Classifier2
Strong ClassifierN
Weak Classifier11
Weak Classifier12
Weak Classifier1T
Weak Classifier21
Weak Classifier22
Weak Classifier2T
Weak ClassifierN1
Weak ClassifierN2
Weak ClassifierNT

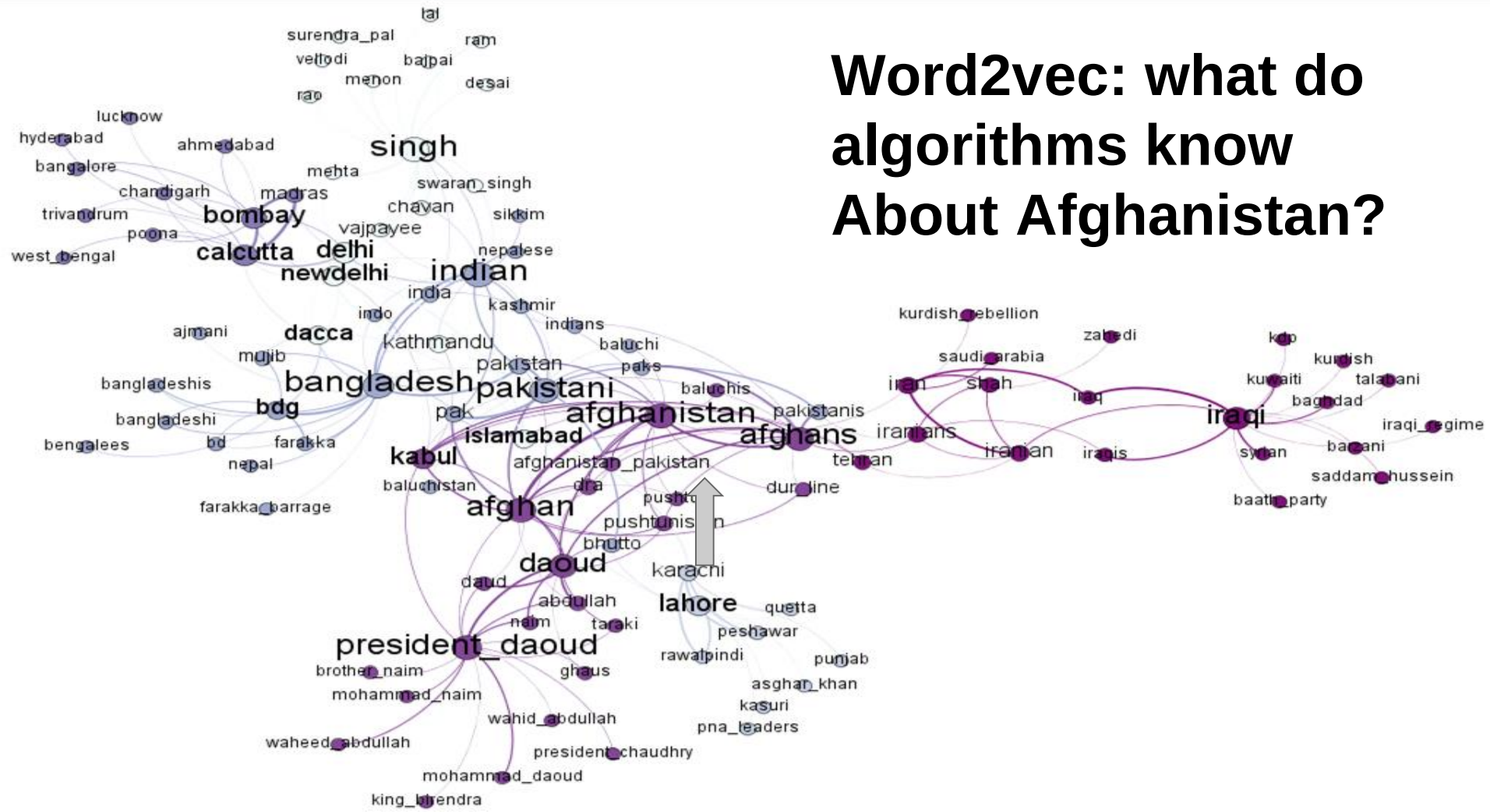
Classifier	ROC/AUC Score	Accuracy Score	Precision (class 0/1)	Recall (class 0/1)	f1-score (class 0/1)
Stochastic Gradient Descent	0.8462/ 0.8338	0.8576/ 0.8512	(0.82/0.88)/ (0.83/0.86)	(0.80/0.89)/ (0.76/0.90)	(0.81/0.89)/ (0.79/0.88)
Logistic Regression	0.8457/ 0.8434	0.8569/ 0.8574	(0.81/0.88)/ (0.82/0.88)	(0.80/0.89)/ (0.79/0.90)	(0.81/0.89)/ (0.81/0.89)
Linear SVM*	0.8454/ 0.8452	0.8563/ 0.8583	(0.81/0.88)/ (0.82/0.88)	(0.80/0.89)/ (0.79/0.90)	(0.81/0.89)/ (0.81/0.89)
Ridge	0.8261/ 0.8373	0.8448/ 0.8546	(0.82/0.86)/ (0.83/0.87)	(0.75/0.90)/ (0.77/0.91)	(0.78/0.88)/ (0.80/0.89)
Bagging (w/ Dec. Tree)	0.8048/ 0.8049	0.8172/ 0.8173	(0.76/0.85)/ (0.76/0.85)	(0.76/0.85)/ (0.76/0.85)	(0.76/0.85)/ (0.76/0.85)
Extremely Randomized Trees	0.8036/ 0.7938	0.8365/ 0.83	(0.86/0.83)/ (0.86/0.82)	(0.67/0.94)/ (0.65/0.94)	(0.76/0.88)/ (0.74/0.87)
AdaBoost (w/ Random F.)	0.8031/ 0.8072	0.8190/ 0.8222	(0.77/0.85)/ (0.77/0.85)	(0.74/0.87)/ (0.75/0.87)	(0.75/0.86)/ (0.76/0.86)
Random Forest	0.7964/ 0.7994	0.8310/ 0.8316	(0.86/0.82)/ (0.85/0.82)	(0.66/0.94)/ (0.67/0.93)	(0.74/0.87)/ (0.75/0.87)
Perceptron*	0.7856/ 0.7963	0.8138/ 0.8112	(0.80/0.82)/ (0.75/0.84)	(0.67/0.90)/ (0.74/0.86)	(0.73/0.86)/ (0.75/0.85)
Passive Aggressive*	0.7745/ 0.8226	0.8095/ 0.837	(0.82/0.81)/ (0.79/0.86)	(0.63/0.91)/ (0.76/0.88)	(0.71/0.86)/ (0.78/0.87)
Multinomial Naive Bayes	0.7735/ 0.7828	0.7614/ 0.7992	(0.64/0.87)/ (0.74/0.83)	(0.82/0.73)/ (0.72/0.85)	(0.72/0.79)/ (0.73/0.84)
Bernoulli Naive Bayes	0.6885/ 0.6885	0.6538/ 0.6538	(0.52/0.84)/ (0.52/0.84)	(0.83/0.55)/ (0.83/0.55)	(0.64/0.66)/ (0.64/0.66)

Feature	Class Combination	ROC/AUC Score	Accuracy Score	Precision (class 0/1)	Recall (class 0/1)	Average f1-score
Subject	(U vs L,C,S)	0.79	0.82	0.81/0.82	0.68/0.91	0.74/0.86
	(U,L vs C,S)	0.80	0.83	0.85/0.77	0.89/0.72	0.87/0.74
	(U,L,C vs S)	0.70	0.96	0.99/0.80	0.99/0.40	0.98/0.53
Concepts	(U vs L,C,S)	0.72	0.75	0.69/0.77	0.59/0.84	0.63/0.81
	(U,L vs C,S)	0.74	0.78	0.80/0.74	0.89/0.58	0.84/0.65
	(U,L,C vs S)	0.68	0.91	0.96/0.75	0.99/0.36	0.97/0.48
Body	(U vs L,C,S)	0.83	0.84	0.79/0.87	0.78/0.88	0.79/0.87
	(U,L vs C,S)	0.81	0.84	0.85/0.82	0.92/0.70	0.88/0.75
	(U,L,C vs S)	0.68	0.95	0.96/0.76	0.99/0.36	0.98/0.49
TAGS	(U vs L,C,S)	0.74	0.78	0.75/0.79	0.61/0.88	0.67/0.83
	(U,L vs C,S)	0.75	0.79	0.82/0.72	0.87/0.63	0.84/0.67
	(U,L,C vs S)	0.62	0.95	0.95/0.73	0.99/0.25	0.97/0.38
Embassies (From/To)	(U vs L,C,S)	0.57	0.67	0.71/0.66	0.19/0.95	0.30/0.78
	(U,L vs C,S)	0.59	0.69	0.70/0.65	0.93/0.24	0.80/0.35
	(U,L,C vs S)	0.57	0.94	0.94/0.72	1.00/0.14	0.97/0.24
Office	(U vs L,C,S)	0.67	0.73	0.76/0.72	0.42/0.92	0.54/0.81
	(U,L vs C,S)	0.62	0.73	0.71/0.83	0.97/0.27	0.82/0.41
	(U,L,C vs S)	0.62	0.95	0.95/0.79	1.00/0.25	0.97/0.38
All_Text	(U vs L,C,S)	0.85	0.86	0.82/0.88	0.81/0.89	0.81/0.89
	(U,L vs C,S)	0.84	0.87	0.88/0.84	0.92/0.76	0.90/0.80
	(U,L,C vs S)	0.78	0.92	0.97/0.78	0.99/0.57	0.98/0.66
All Features (independent vectors)	(U vs L,C,S)	0.86	0.87	0.83/0.89	0.81/0.90	0.82/0.90
	(U,L vs C,S)	0.85	0.87	0.88/0.84	0.92/0.78	0.90/0.81
	(U,L,C vs S)	0.81	0.97	0.97/0.80	0.99/0.61	0.98/0.69
	(U vs C, S)	0.93	0.93	0.93/0.93	0.94/0.92	0.93/0.93


Class = 0 | scenario = UvLCS
Class = 0 | scenario = ULvCS
Class = 0 | scenario = ULCvS
Class = 1 | scenario = UvLCS
Class = 1 | scenario = ULvCS
Class = 1 | scenario = ULCvS
Precision
Recall
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.0
0.2
0.4
0.6
0.8
1.0
Features
body
Concepts
Embassy
Office
Subject
TAGS
All textual fields
All features


0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
AUC
Embassy
Office
Concepts
TAGS
Subject
Body
All textual fields
All features
Features
UvLCS
ULvCS
ULCvS

Word2vec: what do algorithms know About Afghanistan?



Enriching feature-set with semantic vectors

A extended feature-set including features derived from word2vec Analysis, can improve the preformance of the classifier.

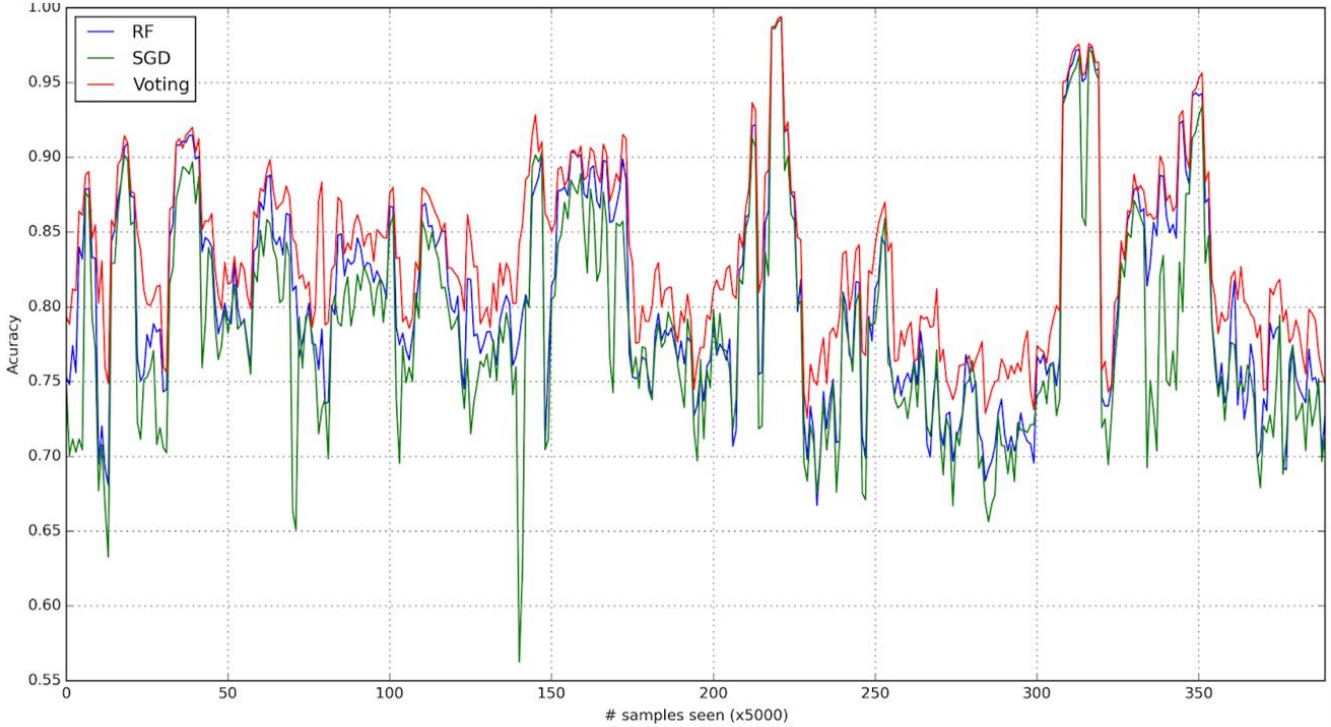
Semantic feature vectors:

$$\overrightarrow{d} = \frac{\sum_i tfidf_{w_i,d} \times \overrightarrow{w_i}}{\sum_i tfidf_{w_i,d}}$$

Exploring Temporal Evolution

Experiment 1:

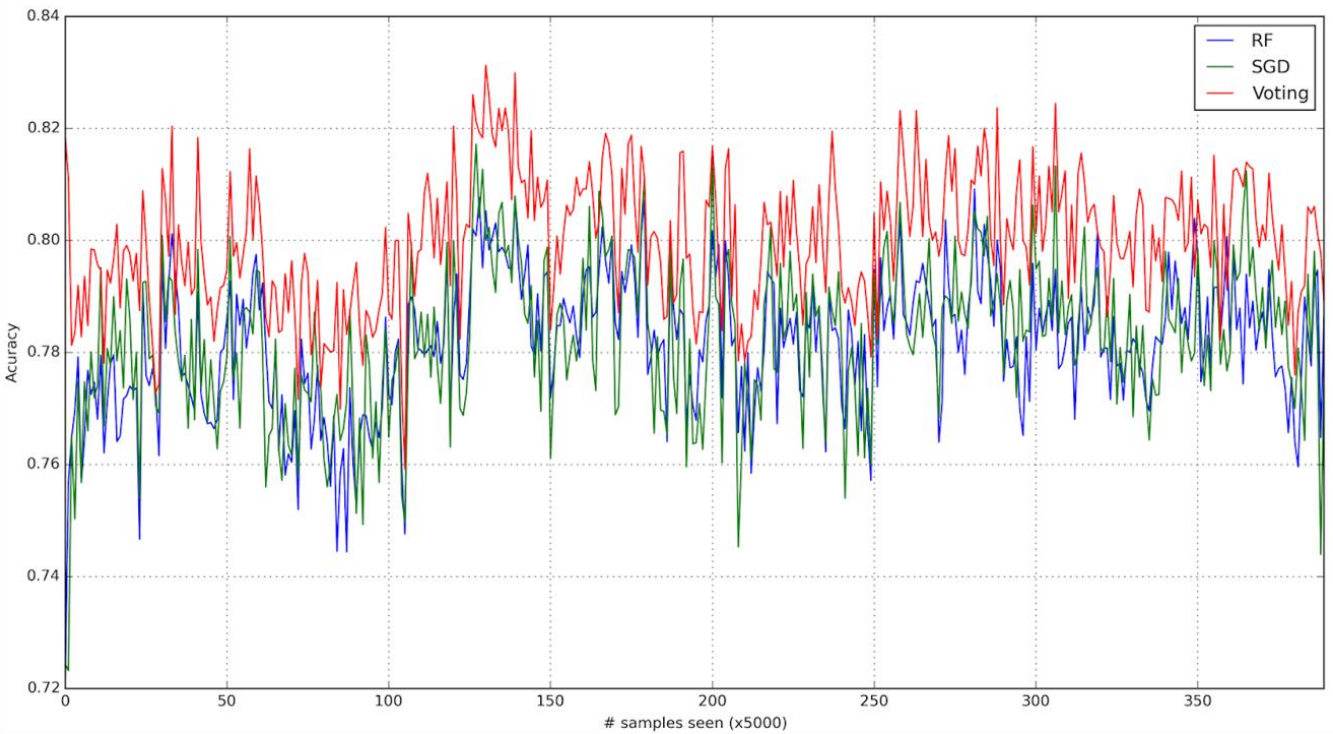
- Split up cable collection if chunks of 5K cables
- Randomize the order
- Run a batch training on this set of batches



Exploring Temporal Evolution

Experiment 2:

- Sort cables in ascending dates
- Split the collection in 5K chunks
- Train the classifier on these chunks from the older ones to the newer.



Next Steps

- Analyze Brazilian and US political reverberations, exploring FGV CPDOC archives diplomatic documents;
- Analyze temporal issues regarding cables
- Dig deeper in the misclassified cables, and identify the misleading features (or the main sources of human errors)
- Identify documents' authorship using ML techniques;
- Build better interfaces for computer aided human classification tasks.