



KE5205 TEXT MINING 2017

TEXT CATEGORIZATION

Leong Mun Kew
Institute of Systems Science
National University of Singapore

email: munkew@nus.edu.sg

© 2017 National University of Singapore. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS, other than for the purpose for which it has been supplied.



Objectives of this module

At the end of this module, you can:

- **Describe what is text categorization and how text categorization systems work**
- **Evaluate a text categorization system with respect to a business scenario**
- **Understand how supervised and unsupervised text categorization works**
- **Understand what is topic modeling**



Outline for this module

- **What is text categorization?**
- **How does supervised text categorization work?**
 - Document data set
 - Building a classifier
 - Evaluation
 - Running the classifier
- **Text categorization application examples**
- **Unsupervised text categorization**
 - Document clustering
- **Topic Modeling**



WHAT IS TEXT CATEGORIZATION?



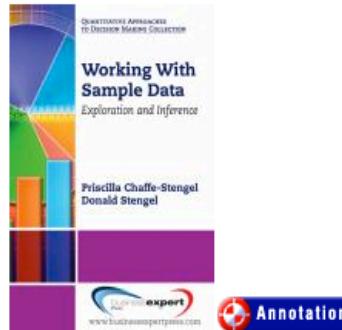
Contrast with a library catalog

- Example to right
 - Subject: Statistics
- Assigned by a cataloger
- Slow, tedious
- May be inconsistent

Record 1 of 381 in National Library Board
Search was: Statistics

Search type: Search by Subjects

▶ Next



Title [Working with sample data](#) : exploration and inference / Priscilla Chaffe-Stengel, Donald N. Stengel.

Author [Chaffe-Stengel, Priscilla M.](#)

Publisher New York : Business Expert Press, c2012.

Physical 151 p. : ill. ; 23 cm.

Description

Notes Includes index.

"The quantitative approaches to decision making collection"--Cover.
Originally published in 2011.

Other [Stengel, Donald N.](#)

Contributors

Search by [Commercial statistics](#)

Subjects

[Statistics](#).



MESH index of a single journal paper

Below is an example of a **complete reference** in Medline (OvidSP) showing the journal article details and the list of MeSH headings (some with subheadings) assigned to it by the NLM Indexers:

Unique Identifier	20980007
Record Owner	From MEDLINE, a database of the U.S. National Library of Medicine.
Status	MEDLINE
Authors	Asejczyk-Widlicka M , Sroda W , Schachar RA , Pierscionek BK .
Authors Full Name	Asejczyk-Widlicka, M. Sroda, W. Schachar, R A. Pierscionek, B K.
Institution	Institute of Physics, Wroclaw University of Technology, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland.
Title	Material properties of the cornea and sclera: a modelling approach to test experimental analysis
Source	Journal of Biomechanics. 44(3):543-6, 2011 Feb 3.
Abbreviated Source	J Biomech. 44(3):543-6, 2011 Feb 3.
NLM Journal Name	Journal of biomechanics
Publishing Model	Journal available in: Print-Electronic Citation processed from: Internet
NLM Journal Code	0157375, hjf
Country of Publication	United States
MeSH Subject Headings	Computer Simulation *Cornea / ph [Physiology] Finite Element Analysis Humans *Intraocular Pressure / ph [Physiology] Muscle Rigidity *Sclera / ph [Physiology] Visual Acuity / ph [Physiology]
Abstract	<p>The indexers have assigned the subheading Physiology to this MeSH descriptor</p> <p>Material properties of cornea and sclera are important for maintaining the shape of the eye and the requisite surface curvatures for optics. They also need to withstand the forces of external and internal musculature and fluctuations in intraocular pressure (IOP). These properties are difficult to measure accurately. Variable results have been reported. A previously published experimental procedure, involving the measurement of the material properties of the eyeball coats were obtained, has been modelled in this study using Finite Element Analysis, in order to test the accuracy of the experiment. Material properties were calculated from the model and the resulting relationships between stress and strain were compared to their experimentally obtained counterparts. The agreement between model and experiment was close for the sclera but more varied for the cornea.</p> <p>The pressure vessel model can be applied for measuring the material properties of the sclera but is less accurate for the cornea. Copyright Copyright 2010 Elsevier Ltd. All rights reserved.</p>

MeSH descriptors
assigned by indexers-
taken from the list of
preferred terms used to
describe topics

The indexers have
assigned the
subheading
Physiology to this
MeSH descriptor



Automatic text categorization (also known as “classification”)

- **The process of assigning text documents uniquely into two or more categories (a document cannot be in more than one category)**
 - E.g., spam filtering – binary decision: “spam” or “not spam”
- **The process of assigning one or more category labels to a text document (a document may have more than one category)**
 - E.g., news filtering – which category to assign to news articles:
 - Sports, Olympics, Football (natural class)
 - Political, Business, Home,... (news sections)
 - Asian, Europe, Middle-East, ... (geographical)



Some Examples of Text Classification

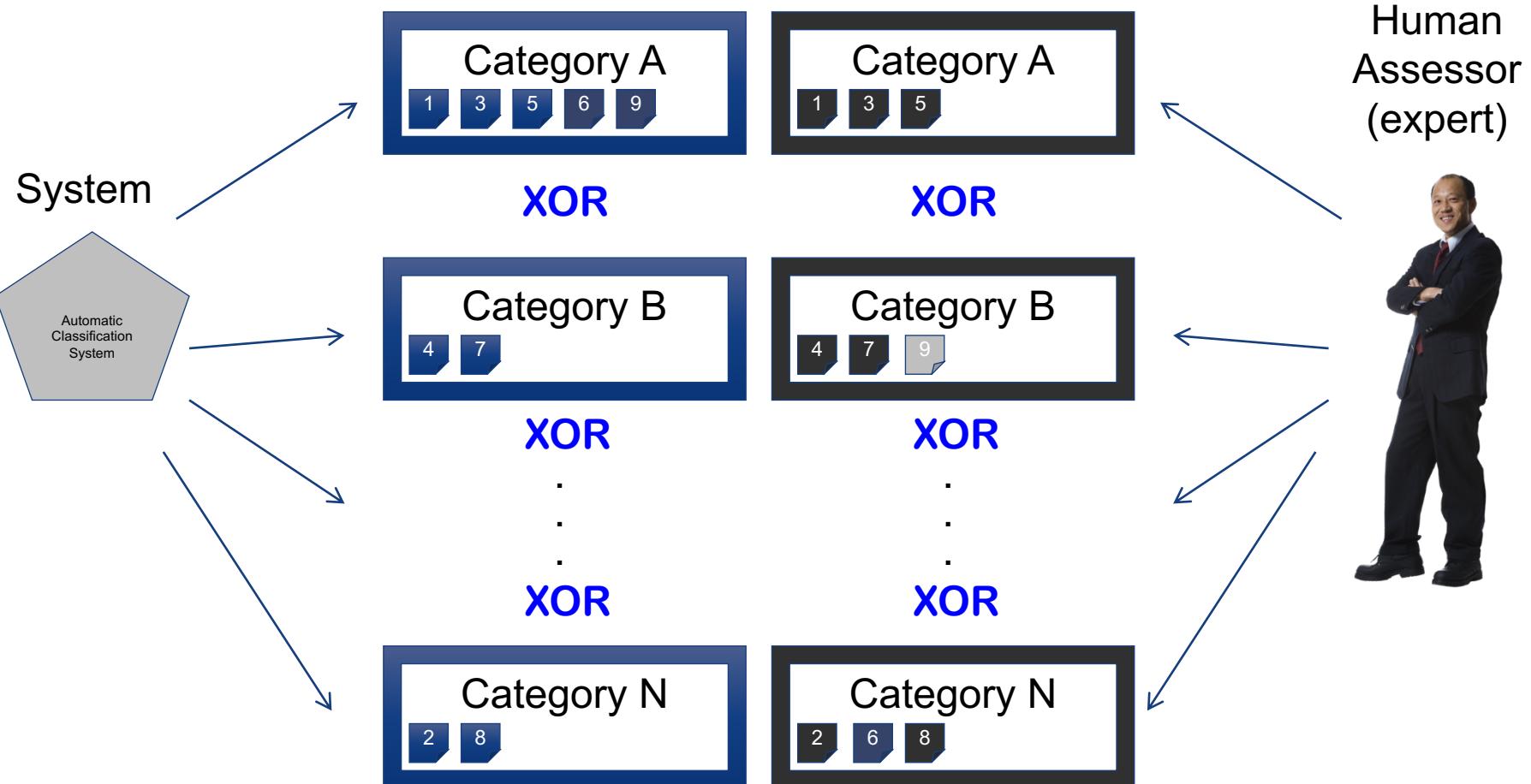
- **Assigning subject categories to documents**
- **Email spam detection**
- **Medical diagnosis**
- **Identifying a language (before further processing)**
- **Identifying fraud (anomaly detection)**
- **Sentiment analysis (e.g., positive/negative reviews)**
- **Monitoring a news feed (e.g., news about Pope Francis)**
- **Etc.**



ACCURACY



What is “accuracy”?

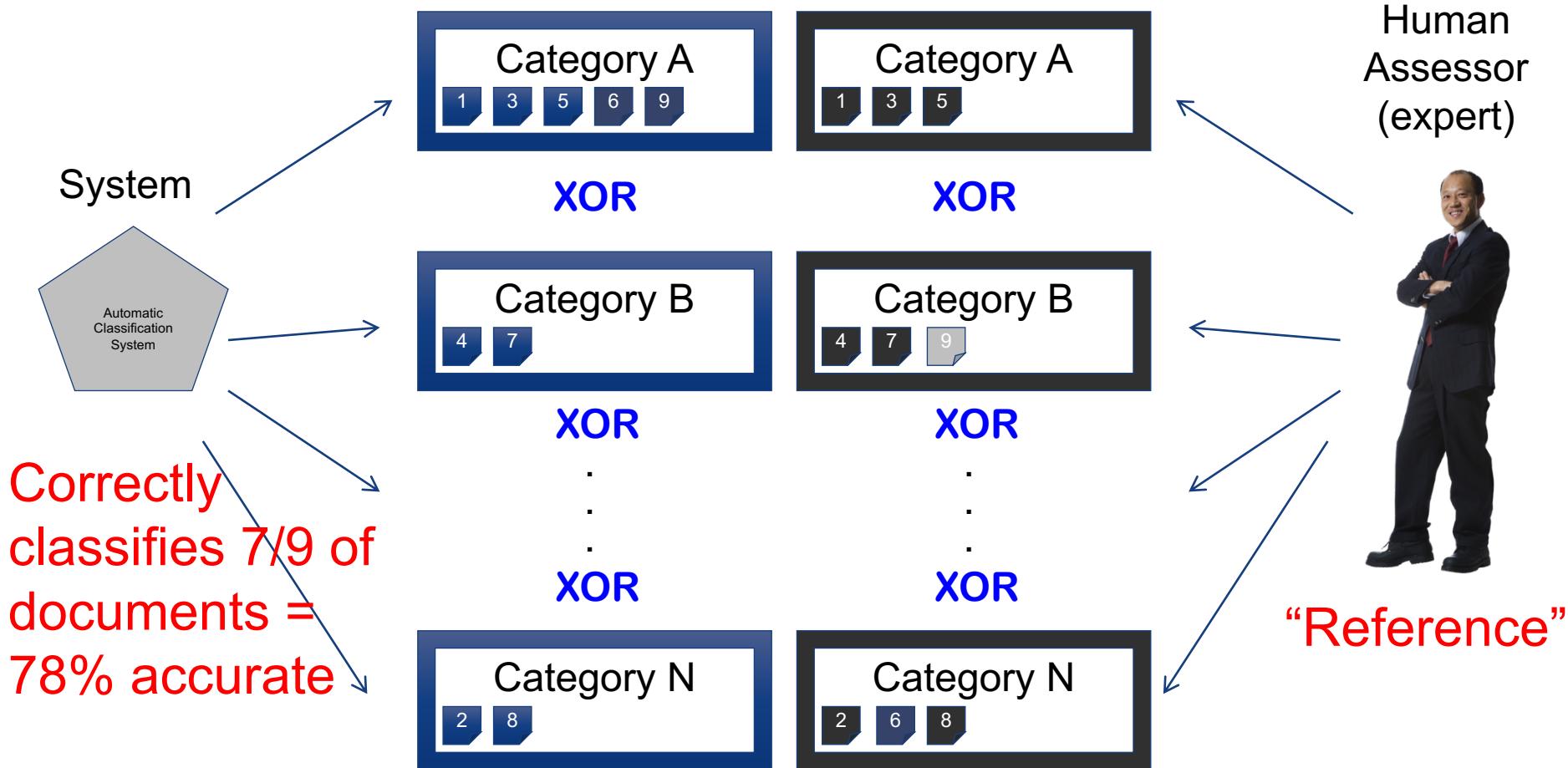




What do we mean by “Accuracy”

- You measure an automatic categorization system by:
 - How well it classifies a set of documents against a “reference”
 - This “reference” is normally a human expert
- Reference
 - Gold standard – accepted as being the best available
 - May not be perfect, e.g., tumour board for oncology
 - Good enough
 - Human expert(s), typically 80% agreement, good methodology
 - Better than nothing
 - “your boss tells you to do this, so you recruit your friends, family,...”
- Most of the time, no such thing as “absolute truth”

One measure of accuracy





Discussion

- **Weather prediction (system predicts one week in advance)**

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
System	Sunny	Drizzle	Rain	Sunny	Cloudy	Thunderstorms	Sunny
Actual	Sunny	Rain	Cloudy	Sunny	Drizzle	Thunderstorms	Cloudy

- **Questions:**
 - How “accurate” is the weather prediction system?
 - Do you have a tolerance for error? +/- margin of error?
 - How about outcomes – will I get wet if I use the system to decide whether or not to carry an umbrella? Will I get angry?



Discussion

- There is a desert in USA where it only rains 10 days in a year. But when it rains, there are flash floods which kill an average of 20 people per year.
 - You build a weather prediction system. How would you measure accuracy?
 - If you said no rain every day, you would automatically be right $355/365$ days in a year = 97% “accuracy”
 - If you said rain every day, you would be right only $10/365$ = 3% of the time, but you would save 20 lives. Would this be true?
 - You are right half the time = 50%, so you would save 10 lives, but ONLY if people listen to your predictions.
 - What would you propose?
-
-



HOW DOES AUTOMATIC TEXT CATEGORIZATION WORK?



Text Categorization Phases

Two Phases

1. Training – creating the text “classifier” (automatic categorization engine)

- You need a set of documents, already categorized
- Divide the set into training (typically 70%) and testing (30%)
- Build your classifier such that it's able to accurately classify the training set of documents to your level of comfort
 - “level of comfort” depends on how hard is the task! ☺
- Evaluate your classifier on the test set; ensure sufficient accuracy

2. Running – using your classifier on new sets of documents

- You will not know how well it performs
- Need to “audit” the results occasionally (use an assessor)
 - Assess random sample of the documents against the predicted categories



DOCUMENT DATA SET



Movie reviews classified as “good” and “bad”

POSITIVE POLARITY (GOOD)

- a mesmerizing cinematic poem from the first frame to the last .
- a well-put-together piece of urban satire .
- one can't deny its seriousness and quality .
- hard to resist .
- a naturally funny film , home movie makes you crave chris smith's next movie .
- a true-blue delight .
- a fun ride .
- a surprisingly funny movie .
- the script is smart and dark - hallelujah for small favors .
- a flick about our infantilized culture that isn't entirely infantile .

-
- unfortunately the story and the actors are served with a hack script .
 - too slow for a younger crowd , too shallow for an older one .
 - terminally brain dead production .
 - one lousy movie .
 - this movie . . . doesn't deserve the energy it takes to describe how bad it is .
 - a cleverly crafted but ultimately hollow mockumentary .
 - it's an 88-minute highlight reel that's 86 minutes too long .
 - the whole affair is as predictable as can be .

NEGATIVE POLARITY (BAD)

From: <http://karpathy.ca/mlsite/lecture2.php>



Movie reviews classified as “good” and “bad”

POSITIVE POLARITY (GOOD)

- a mesmerizing cinematic poem from the first frame to the last .
- a well-put-together piece of urban satire .
- one can't deny its seriousness and quality .
- hard to resist .
- a naturally funny film . home movie makes you crave chris smith's next movie .
- a true-blue delight .
- a fun ride .
- a surprisingly funny movie .
- the script is smart and dark . hallelujah for small favors .
- a flick about our infantilized culture that isn't entirely infantile .

5000 reviews

Training Set

70%

30%

Test Set

- unfortunately the story and the actors are served with a hack script .
- too slow for a younger crowd , too shallow for an older one .
- terminally brain dead production .
- one lousy movie .
- this movie . . . doesn't deserve the energy it takes to describe how bad it is .
- a cleverly crafted but ultimately hollow mockumentary .
- it's an 88-minute highlight reel that's 86 minutes too long .
- the whole affair is as predictable as can be .

5000 reviews

NEGATIVE POLARITY (BAD)

From: <http://karpathy.ca/mlsite/lecture2.php>



BUILDING A CLASSIFIER

JUST SOME EXAMPLES (NOT EXHAUSTIVE)



Creating classifiers

- **Hand-coded classifiers (the “good old days!”)**
 - If <conditions> then <category> else NOT<category>, where conditions are normally in disjunctive normal form

If	((wheat & farm)	or
	(wheat & commodity)	or
	(bushels & export)	or
	(wheat & tonnes)	or
	(wheat & winter & ¬soft))	then WHEAT else ¬ WHEAT

From: F. Aiolli, *Text Categorization*, <http://www.math.unipd.it/~aiolli/corsi/SI-0607/Lez09.251006.pdf>



Inducing classifiers (1)

- **Probabilistic Classifiers**

- Represent the probability that a document d_i belongs to category c_j by

$$P(c_j|d_i) = P(c_j)P(d_i|c_j)/P(d_i)$$

where $P(c_j)$ is the probability that a randomly selected document belongs to category c_j , $P(d_i)$ is the probability that a randomly selected document has d_i as a vector representation and $P(d_i|c_j)$ is the probability that category c_j contains d_i .

- **Decision Tree Classifiers**

- Uses symbolic representation in a tree network
- Internal nodes are terms, edges are tests on the weights, and leaf nodes are categories
- The classifier works by traversing a path to the appropriate leaf node



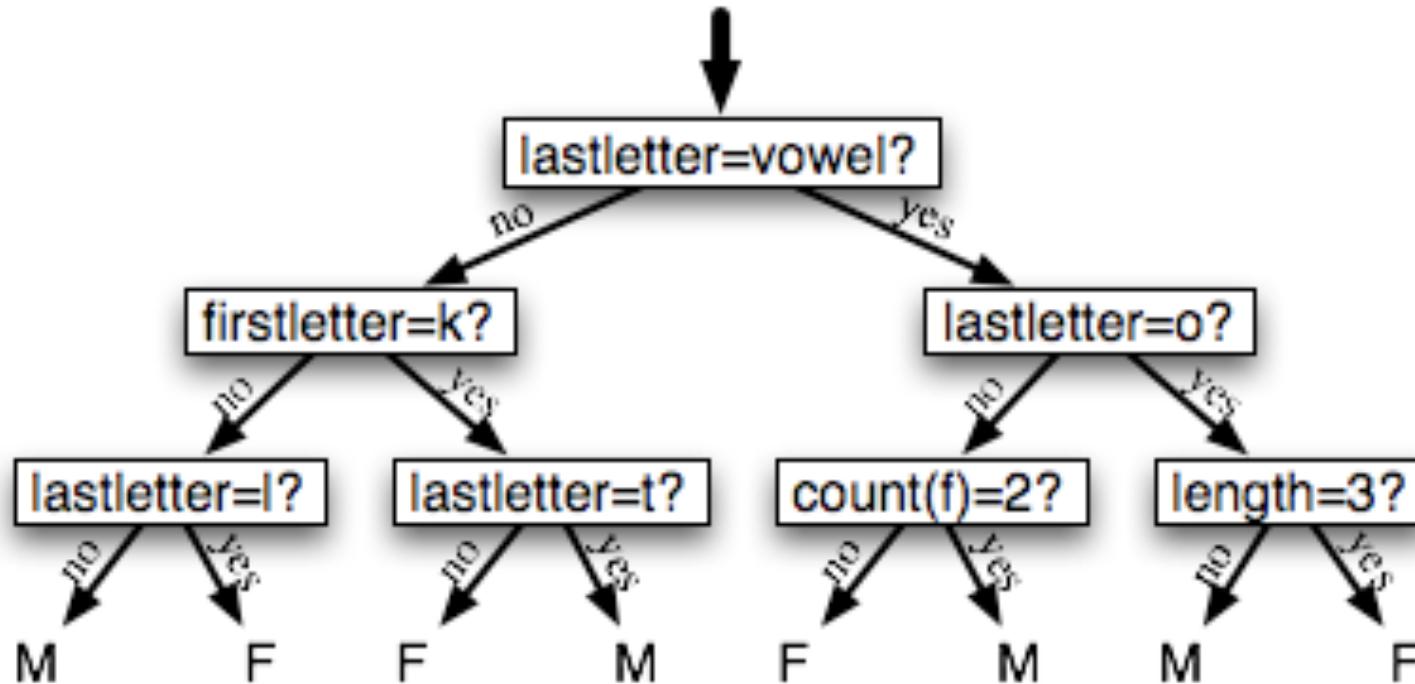
Exercise

- **List of boys names**
 - Alan
 - Barry
 - Colin
 - Dexter
 - Edward
 - Frederick
 - Howard
 -
- **List of girls names**
 - Anna
 - Betty
 - Chelsea
 - Doris
 - Elizabeth
 - Fanny
 - Hortense
 -

Given a list of boys' names and a list of girls' names, build a simple classifier to distinguish boys' names from girls' names



Example of a decision tree to decide if a name is male or female

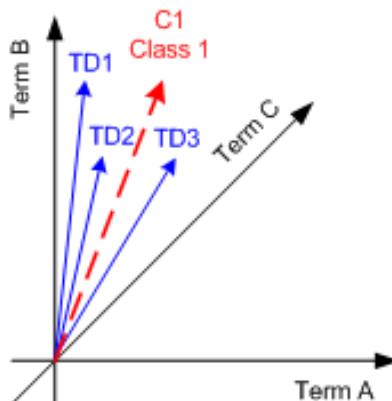


From: <http://nltk.googlecode.com/svn/trunk/doc/book/ch06.html>

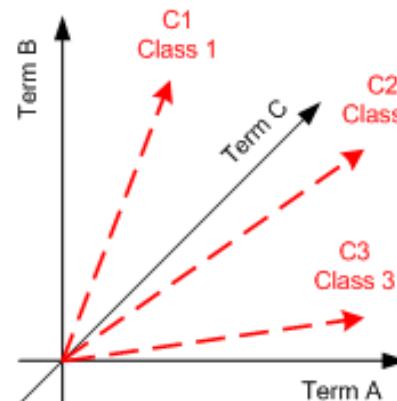
Inducing classifiers (2)

- The Rocchio Classifiers

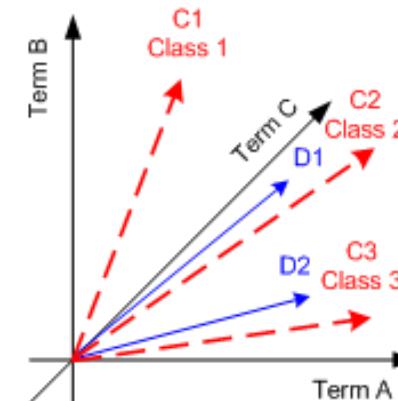
- Each category is represented by a prototypical document, i.e., profile vector
- Documents are classified by similarity to the profile vector



A) Training Document Representations in Vector Space define Class Representation (Centroid)



B) Class Representation by Class Vectors (Centroids)



C) Classification of Documents by similarity between Document Vector and Class Vector

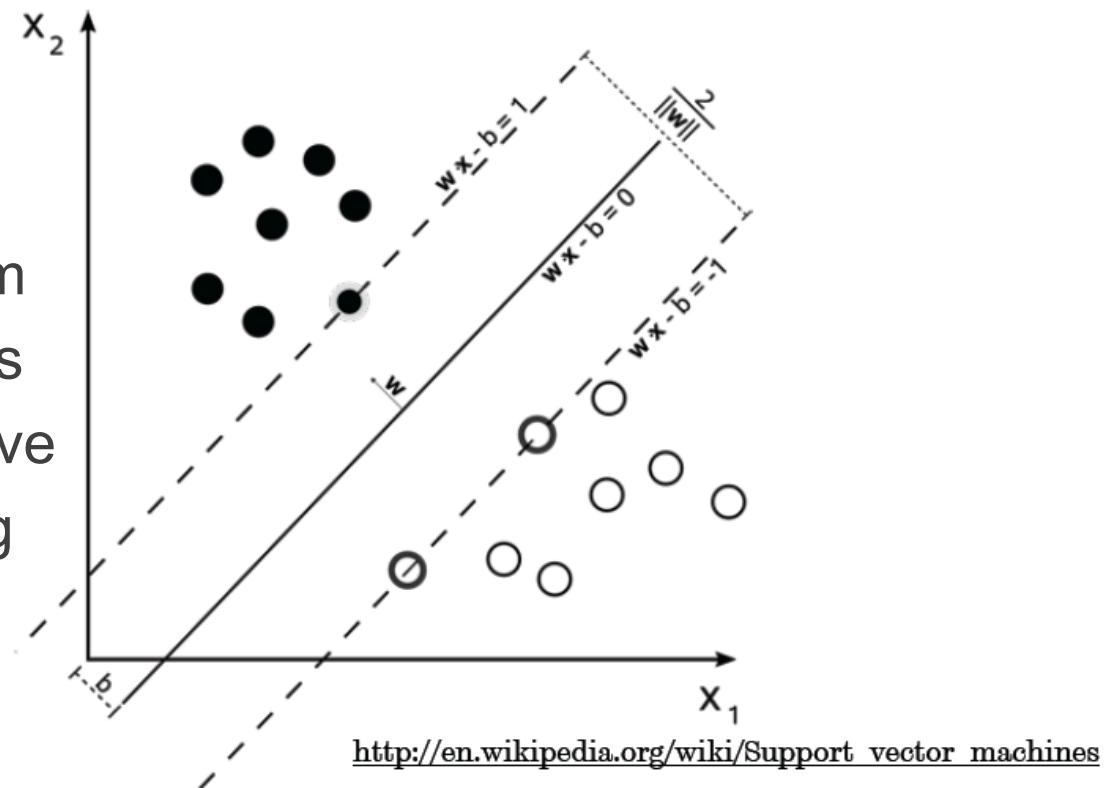
From: http://www.iicm.tugraz.at/about/Homepages/cguetl/courses/isr/opt/classification/Vector_Space_Model.html



Inducing classifiers (3)

- **Support Vector Machines (SVMs)**

- SVMs divide the term space in hyperplanes separating the positive and negative training samples.
- The surface that provides the widest separation between the support surfaces is selected



[http://en.wikipedia.org/wiki/Support vector machines](http://en.wikipedia.org/wiki/Support_vector_machines)



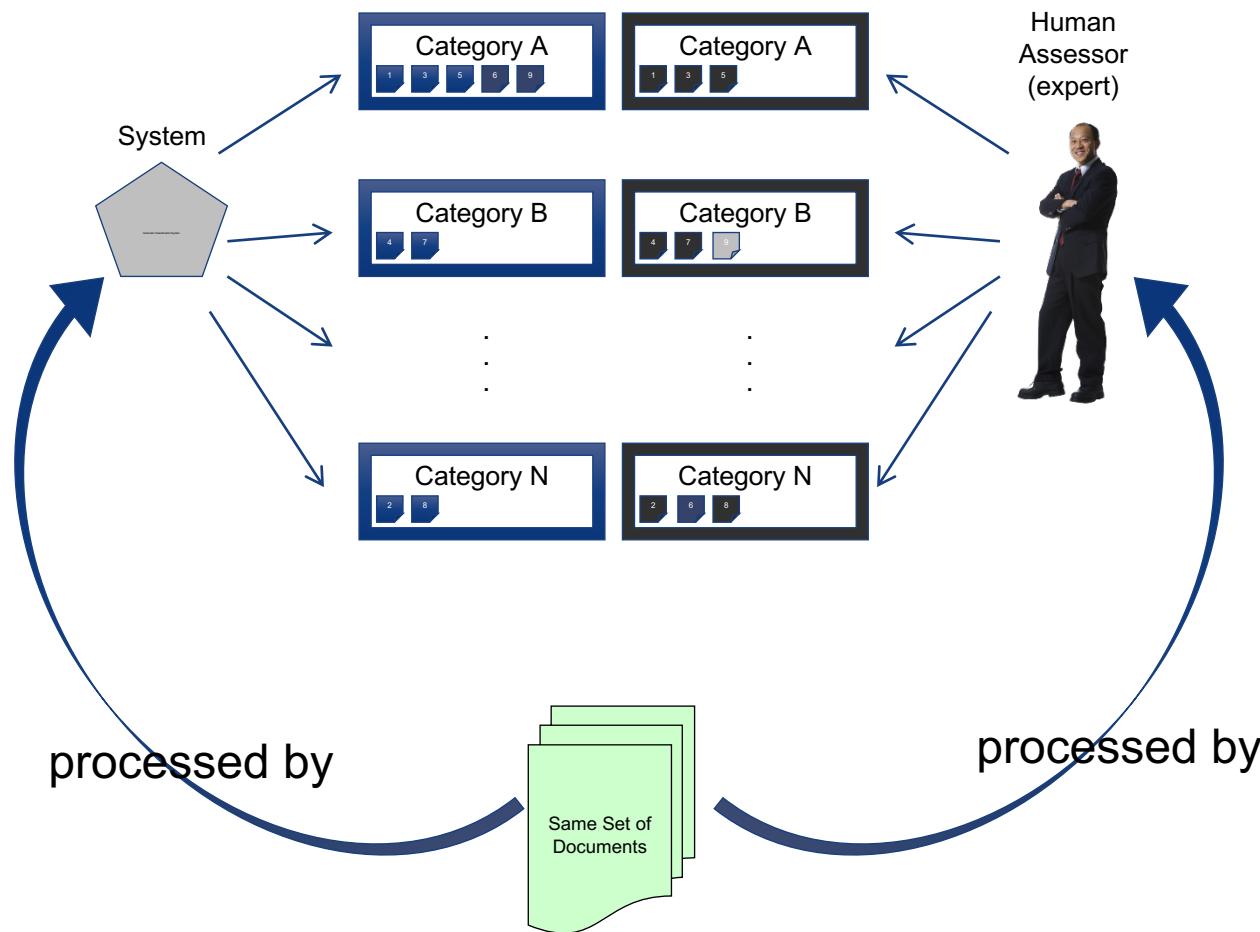
EVALUATION

COMPARING DIFFERENT CLASSIFIERS FOR THE SAME CATEGORIZATION TASK



Remember this?

Predicted Categories Actual Categories



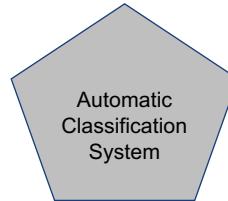


Confusion Matrix

		Predicted Categories						
		A	B	C			...	N
Actual Categories	A							
	B							
	C							
	⋮							
	N							



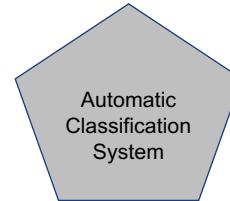
Example (using %)



		Predicted Categories						
		A	B	C			...	N
Actual Categories	A	87%	2%	5%			...	1% = 100%
	B	6%	90%	0%			...	2% = 100%
	C	12%	2%	77%			...	4% = 100%
	⋮						⋮	
	N	21%	0%	4%			...	65% = 100%



Example (using #)



		Predicted Categories						
		A	B	C			...	N
Actual Categories	A	143	34	17			...	2
	B	67	1289	44			...	239
	C	980	234	3454			...	88
	⋮						⋮	
	N	87	24	63			...	650
								= Tot(N) docs



Consider the simple 2x2 matrix (2000 documents were classified)

Desired positive prediction

		Predicted	
Actual	Predicted	Yes	No
	Yes	1350 90%	150 10%
	No	100 20%	400 80%

False negative

False positive

Desired negative prediction



EVALUATING MULTIPLE CLASSIFIERS



Discussion

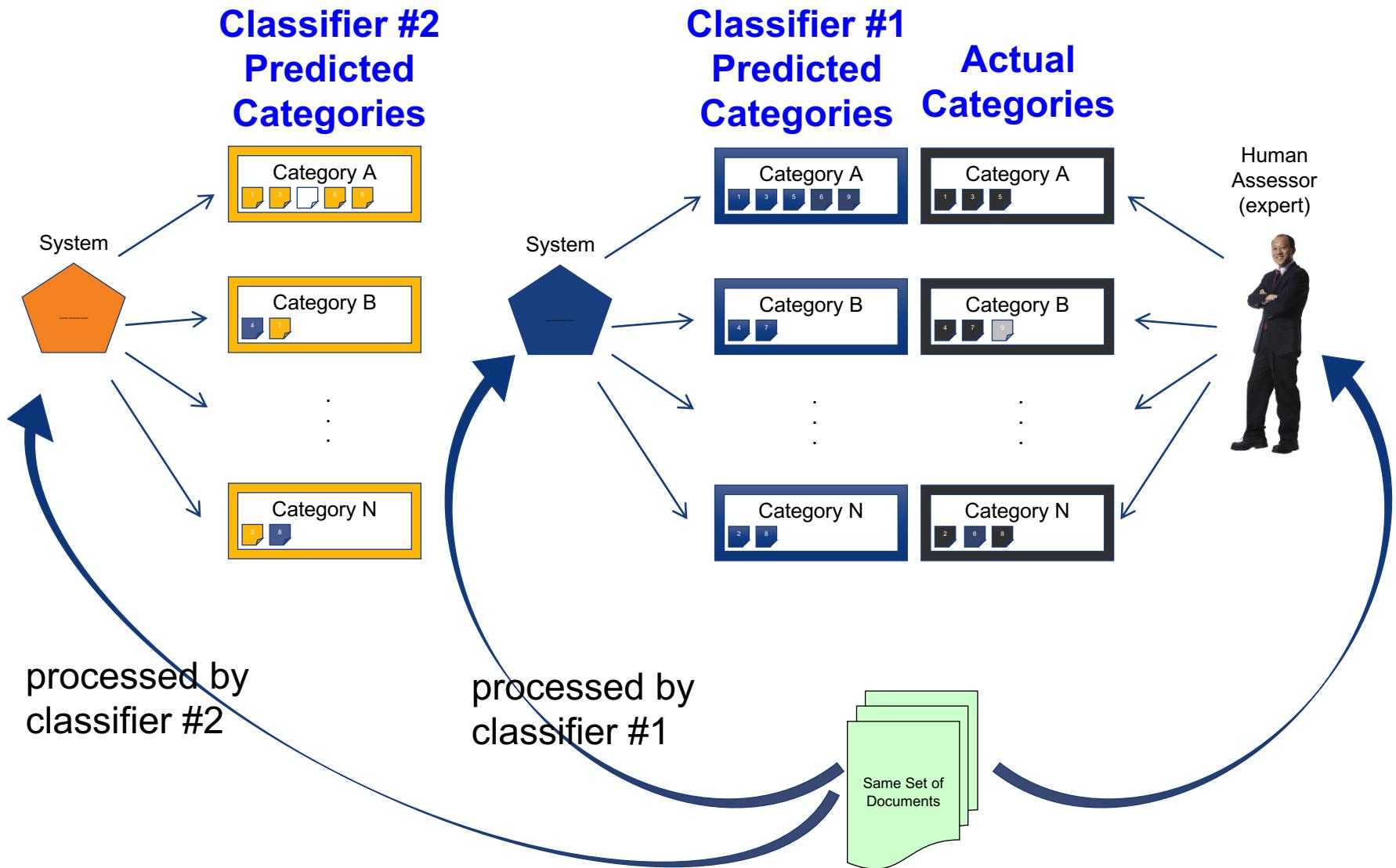
- **Weather prediction. You ask two people to predict whether it will rain or not in the coming week:**

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
ZZ	Yes	No	Yes	No	Yes	No	No
MK	No	No	No	Yes	No	No	Yes
Actual	Yes	Yes	Yes	No	Yes	Yes	No

- **Questions:**
 - Who is more “accurate”? ZZ is right 5/7 times. MK is right 0/7 times.
 - Who should you ask in future if you don’t want to get wet?
 - Who is the better “classifier”?



What happens with 2 classifiers?





Comparing models: e.g. 2x2 matrix (actual numbers of documents)

Classifier #1



		Predicted	
		Y	N
Actual	Y	900	100
	N	40	410

Seems quite good
for both predictions

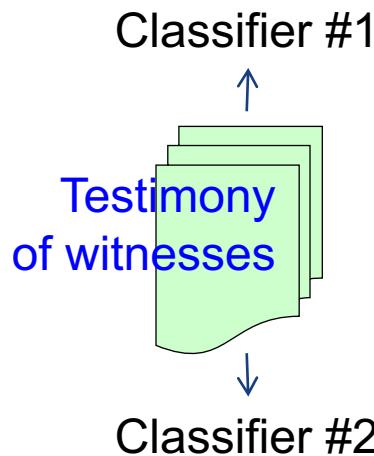
Classifier #2

		Predicted	
		Y	N
Actual	Y	700	300
	N	2	448

Reduced the false positives
but false negatives increased

? **Which classifier is better?**

Adding the semantics – the courtroom



		Predicted	
		Guilty	Innocent
Actual	Guilty	900	100
	Innocent	40	410

Let 100 guilty go free
Convict 40 innocent persons

		Predicted	
		Guilty	Innocent
Actual	Guilty	700	300
	Innocent	2	448

Let 300 guilty go free
Convict 2 innocent persons

?

Which classifier is better?

Adding a cost function – fraud investigation

The diagram illustrates the flow of data through two classifiers. On the left, a stack of green rectangles labeled "Insurance Claim Statements" is shown. An arrow points from this stack up to "Classifier #1". Another arrow points down from "Classifier #1" to "Classifier #2".

		Predicted	
		Genuine	Fraud
Actual	Genuine	900	100
	Fraud	40	410

		Predicted	
		Genuine	Fraud
Actual	Genuine	700	300
	Fraud	2	448

Company loses \$80k in fraud
Company pays \$255k in costs

? **Which classifier is better?**

Company loses \$4k in fraud
Company pays \$374k in costs

The average fraud costs the company \$2000
It costs the company \$500 to investigate each suspected fraud

Adding a cost function – fraud investigation

- **Consider Doing nothing (don't act to identify fraud):**
 - Predicted fraud = 0 cases @ \$500 per case costs \$0k for investigation.
 - Undetected fraud is 450 cases @\$2k/fraud loses \$900k.
 - Overall -\$0k -\$900k = -\$900k

		Predicted	
		Genuine	Fraud
Actual	Genuine	900	100
	Fraud	40	410

		Predicted	
		Genuine	Fraud
Actual	Genuine	700	300
	Fraud	2	448

The average fraud costs the company \$2000
It costs the company \$500 to investigate each su

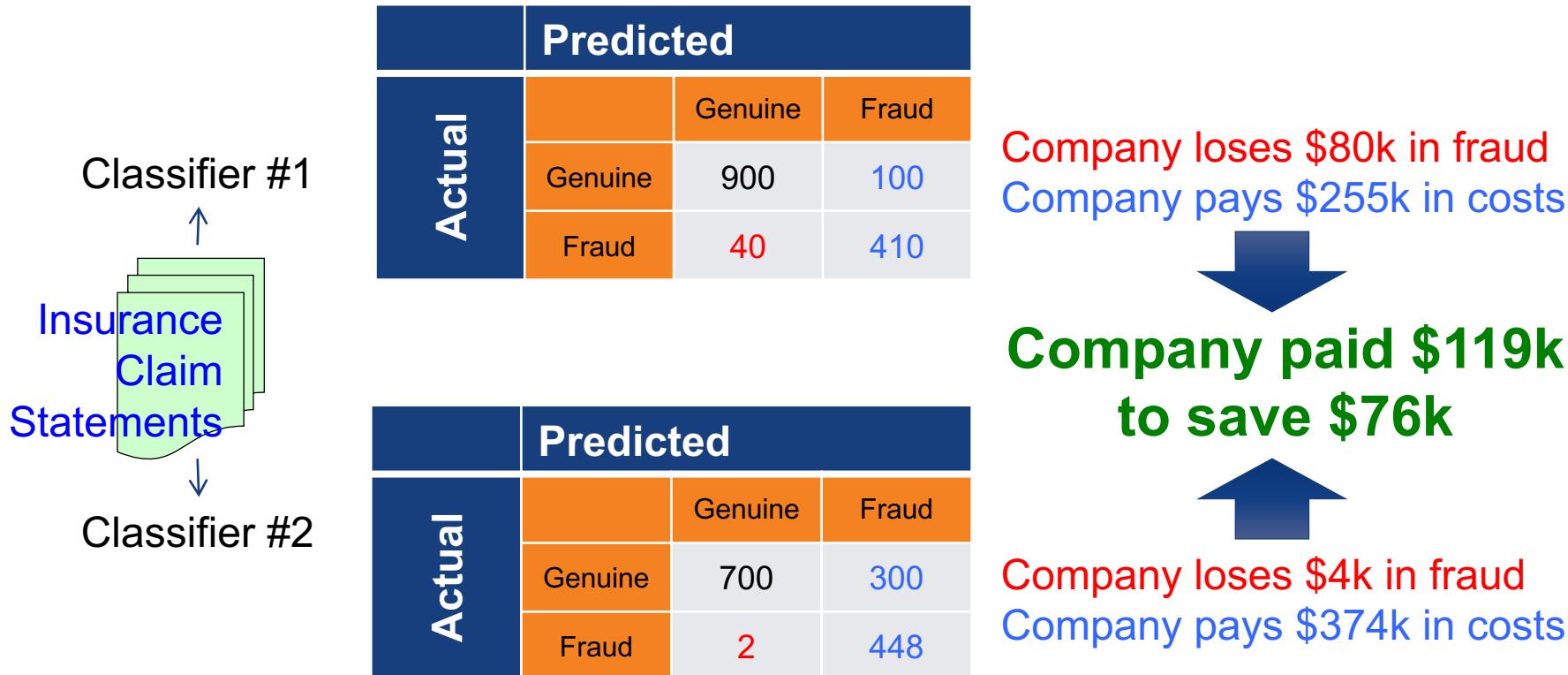
Analysis for classifier #1:

- Predicted fraud = 510 cases @ \$500 per case costs \$255k for investigation.
- Undetected fraud is 40 cases @\$2k/fraud loses \$80k.
- Overall -\$255k -\$80k = -\$335k

Analysis for classifier #2:

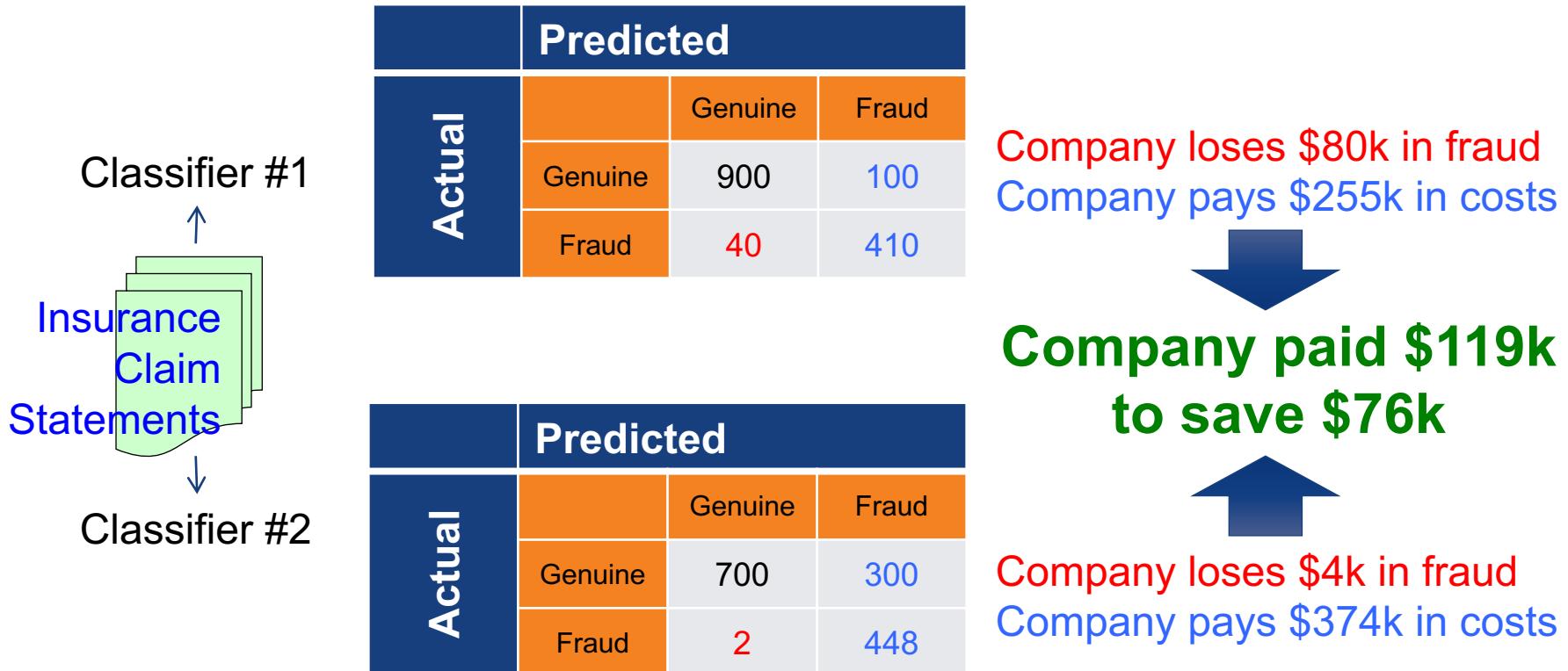
- Predicted fraud = 748 cases @ \$500 per case costs \$374k for investigation.
- Undetected fraud is 2 cases @\$2k/fraud loses \$4k.
- Overall -\$374k -\$4k = -\$378k

Adding a cost function – fraud investigation



The average fraud costs the company \$2000
It costs the company \$500 to investigate each suspected fraud

Consider the branding?



The average fraud costs the company \$2000
It costs the company \$500 to investigate each suspected fraud



Classifier evaluation

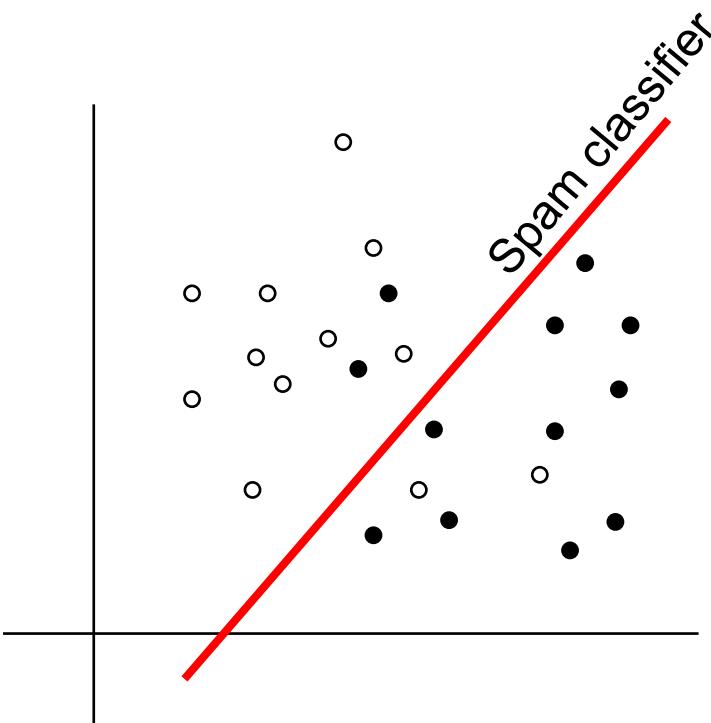
- Evaluation of classifiers is done with respect to a **business context**
- Evaluation of classifiers is normally done **empirically**
- Experimental evaluation focuses on **effectiveness**, i.e., the ability of the classifier to make the right classification decision
- Precision & Recall concepts as applied to (multi-class) categorization
 - Precision is the probability that if a random document d_i is categorized under category c_j , that decision is correct
 - Recall wrt c_j is the probability that if a random document d_i should be categorized under c_j , then the decision is taken



RUNNING THE CLASSIFIER

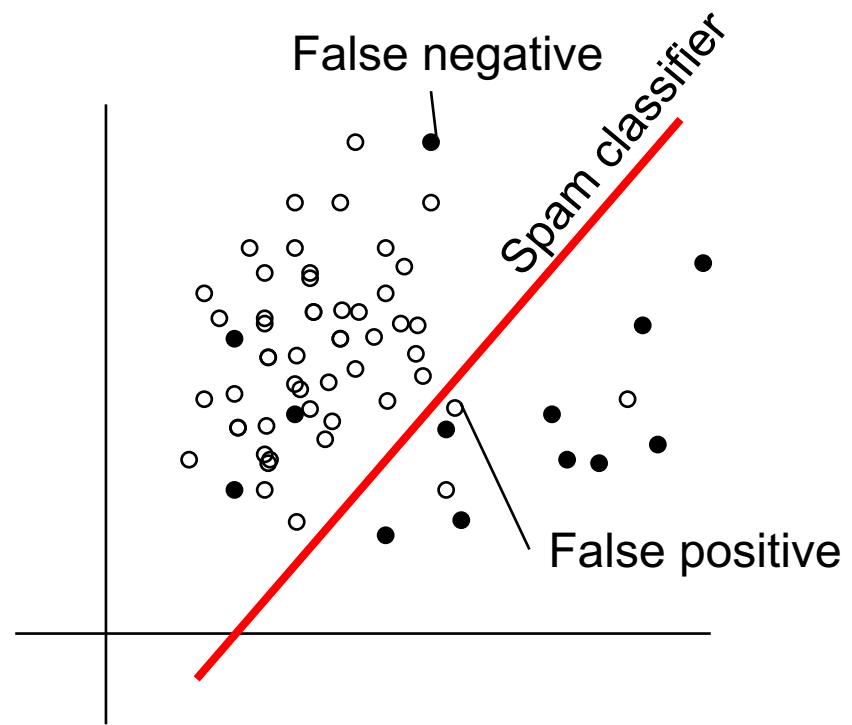


Expect False Results eg: spam filtering



- Email data – non-spam
- Email data – spam

Training Set

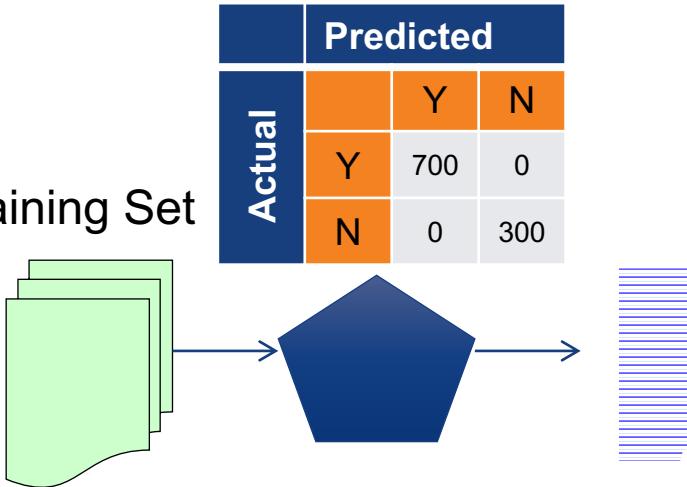


- Email non-spam
- Email spam

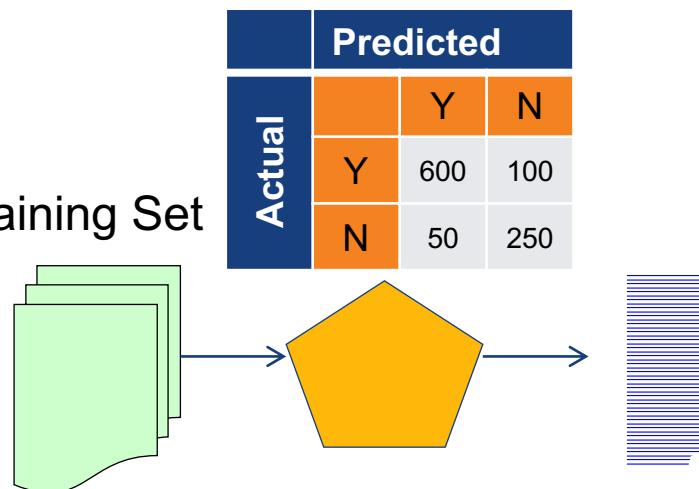
Real email stream

Overfitting the Training Set

Training Set



Training Set



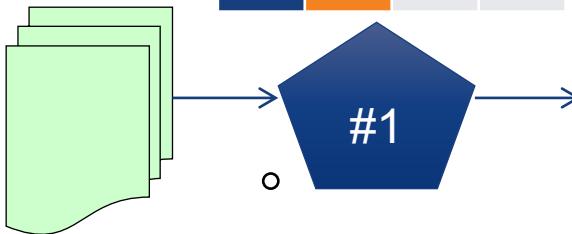
?

Which classifier is better?

Overfitting the Training Set

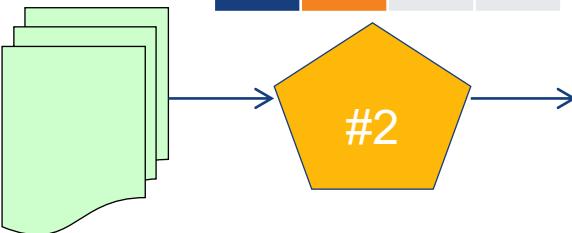
		Predicted	
		Y	N
Actual	Y	700	0
	N	0	300

Training Set

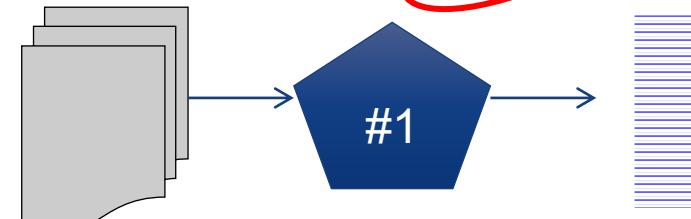


		Predicted	
		Y	N
Actual	Y	600	100
	N	50	250

Training Set

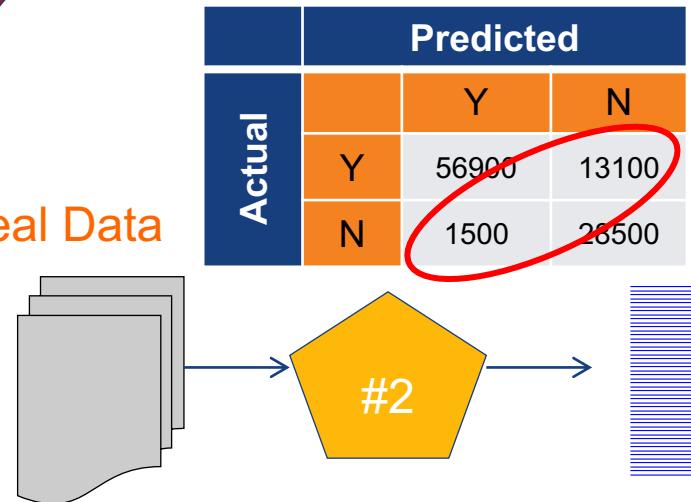


Real Data



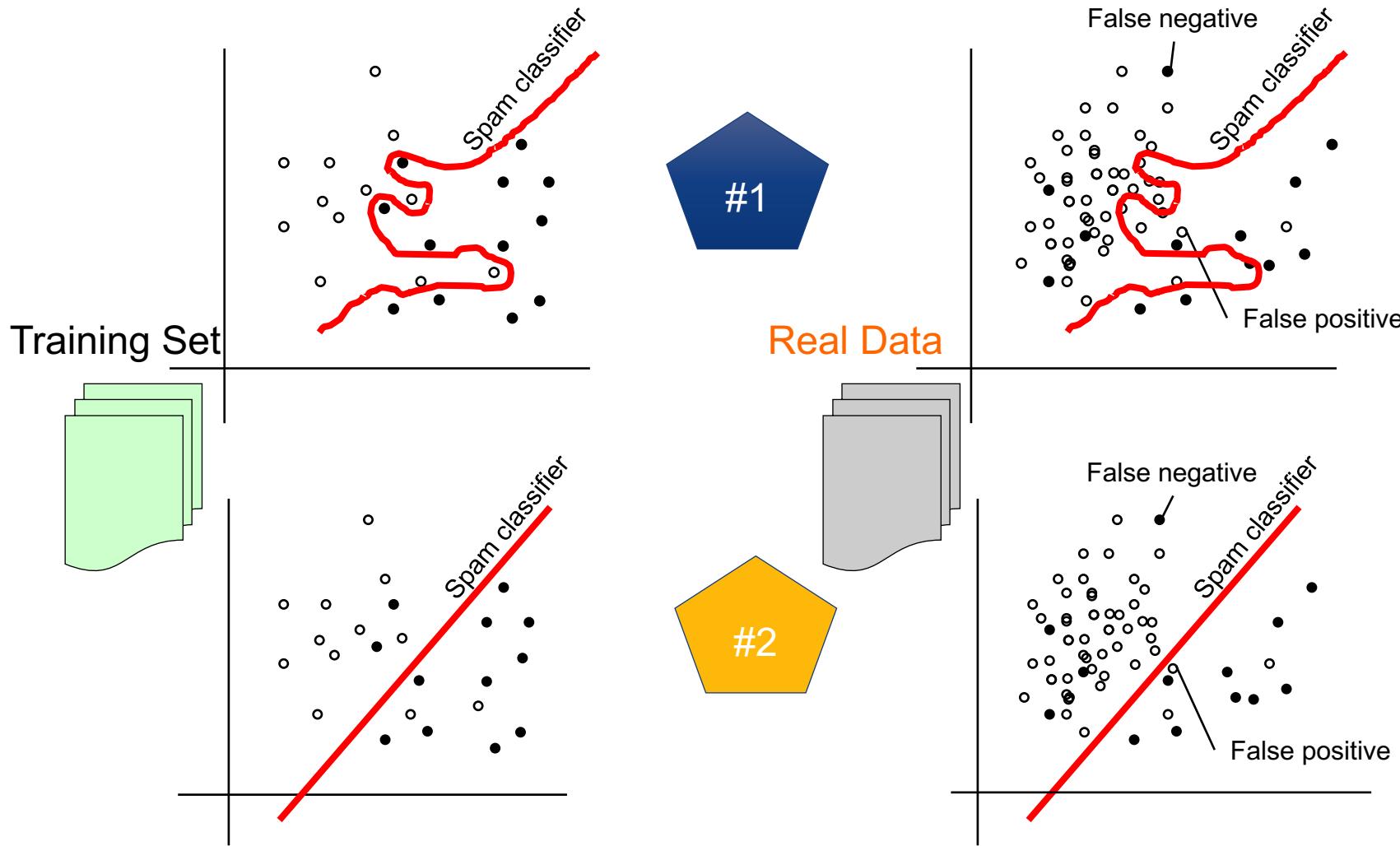
REAL
WORLD

Real Data





Overfitting the Training Set – what happened?



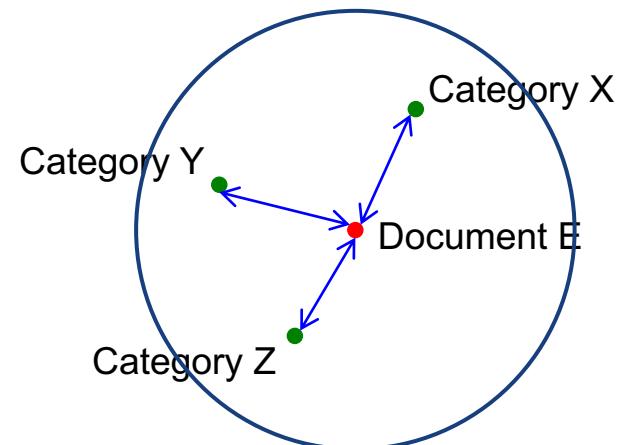
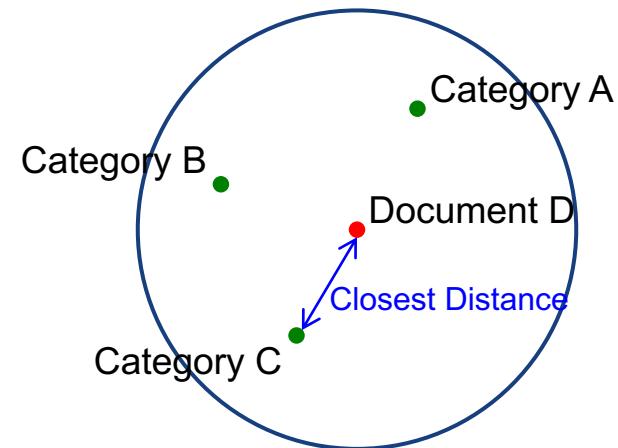


Hard and Soft Categorization

- Fully automated classifiers make “hard” binary decisions
 - In example to right, the document, D, is assigned to category C only.
- Semi-automated (interactive) classifiers instead are created by allowing “soft” real-value decisions
 - Rank the categories according to their measure of appropriateness for the document
 - In example to right, the document, E, is assigned 3 possible categories:

Rank	Category	Probability
1	Z	0.76
2	X	0.72
3	Y	0.54

- Used for computer assisted human decision making
 - For example, in critical applications such as medical diagnosis



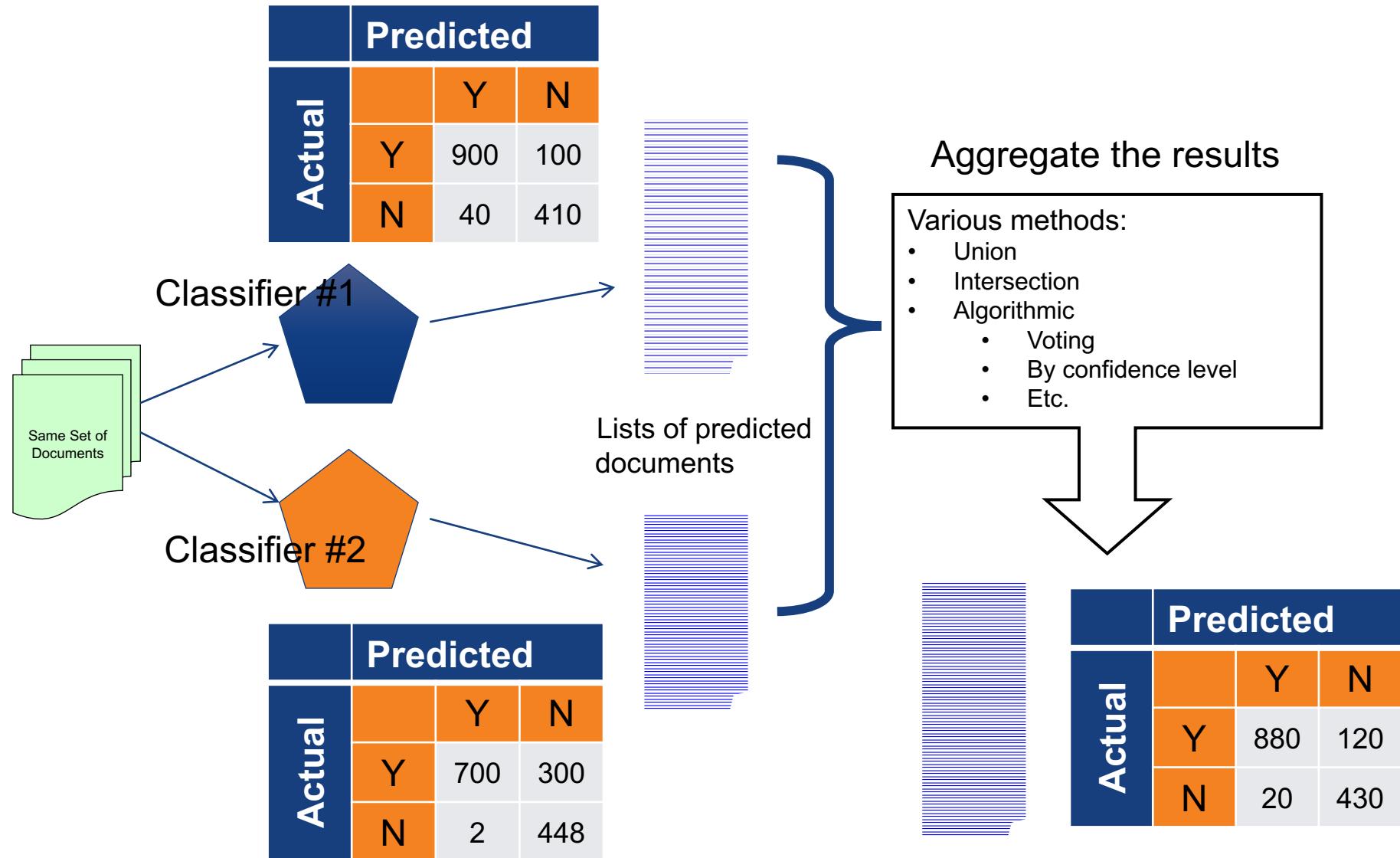


Aggregating multiple classifiers





Running with more than one classifier





TEXT CATEGORIZATION APPLICATION EXAMPLES

Boosting Identification of Fraudulent Claims

The video player shows a comparison between two classification matrices. The top matrix, titled 'Data: Classification matrix (fraud ...)', is for models without Text Mining results. It shows 538 observed Yes cases predicted as Yes, and 138 observed Yes cases predicted as No. The bottom matrix, titled 'Data: Classification matrix (fraud w...)', is for models utilizing Text Mining results. It shows 612 observed Yes cases predicted as Yes, and 61 observed Yes cases predicted as No. A red box highlights the bottom matrix.

Predicting Fraudulent Claims – Comparison

- Fraud = Yes
 - Without Text Mining results, we missed 138
 - With Text Mining results, we missed 64
- 74 additional fraudulent claims were detected by using Text Mining results
- This is over 10% of fraudulent claims in this small data set

4:41 / 6:01

Text Mining Series: Predicting Fraudulent Claims

StatSoft · 95 videos

1,715

1,375

Like 5 | Dislike 0

Uploaded on 15 Nov 2011

In this case study, fraud detection models are built using the structured variables and provide a good predictive model, finding fraudulent claims. Then with the aid of STATISTICA Text Miner,

From: <http://www.youtube.com/watch?v=OIQpm8qTog4>

Automatic Categorization of Documents

The screenshot shows a YouTube video player with a video thumbnail at the top. The thumbnail displays a 3D bar chart from STATISTICA software titled "Classification matrix 1 Dependent variable: Topic: Earnings? Options: Categorical response, Tree number 1, Test sample". The chart has three bars: one red bar near zero, one green bar around 400, and one large green bar reaching approximately 2000. Below the video player, the video's title and channel information are visible:

Text Mining Series - Automatically Classify Text Documents
StatSoft • 95 videos

The video has 3,022 views, 1,375 subscribers, and 12 likes. It was uploaded on 27 Oct 2011. The video content summary is: "In this case study, there is a need to automatically classify text documents based on their content. Currently, the text articles are manually read and acted upon. Our goal is to automate as much as

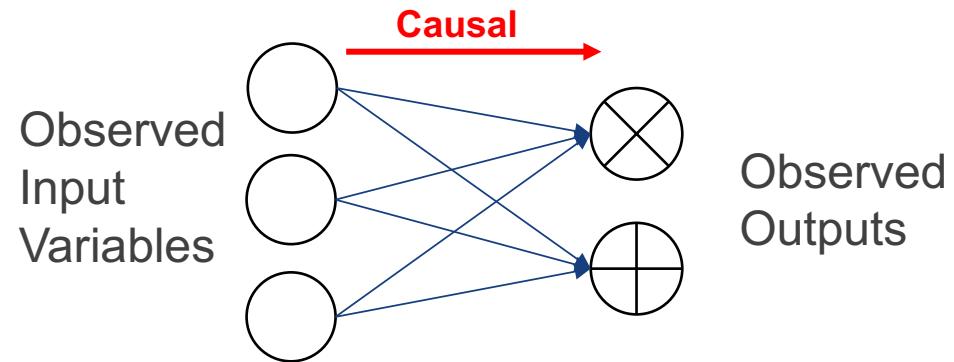
From: <http://www.youtube.com/watch?v=Q5K3gyQJkC0>



UNSUPERVISED TEXT CATEGORIZATION

Supervised vs Unsupervised

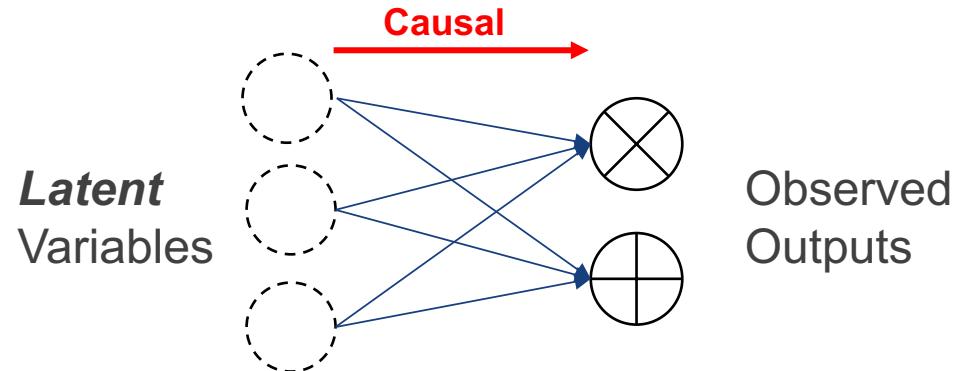
- **Supervised categorization**



Essentially similar models

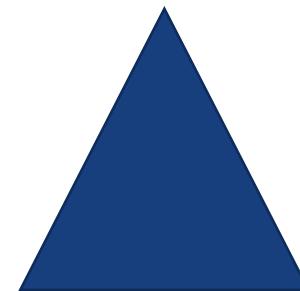
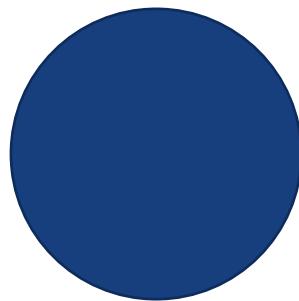
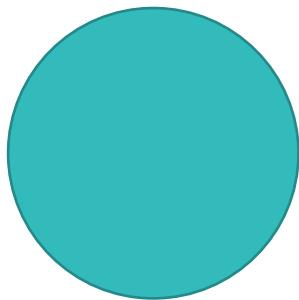
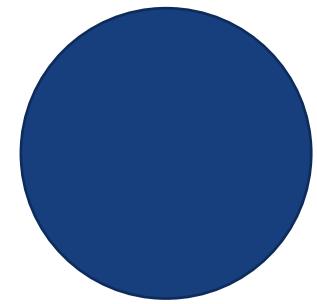
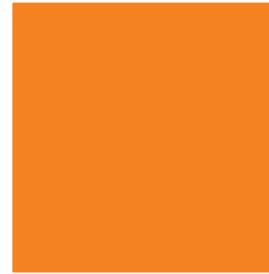
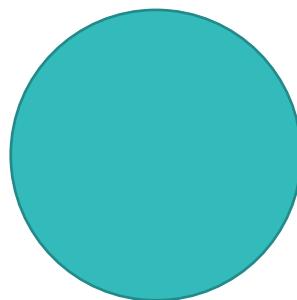
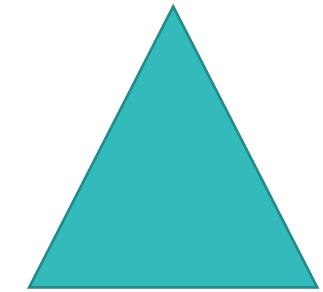
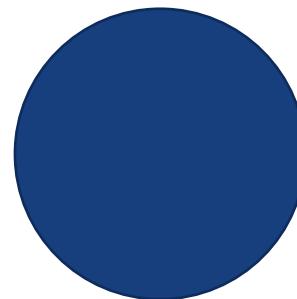
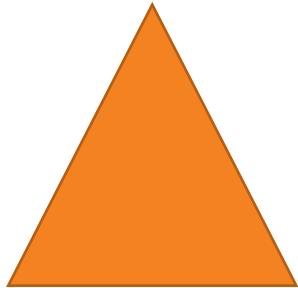
- Inputs causally effects the outputs
- Input variables may not be observable

- **Unsupervised categorization**



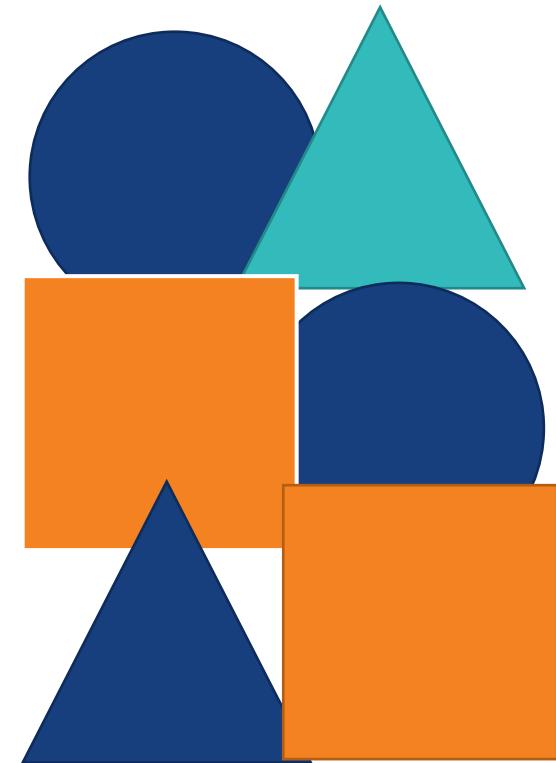
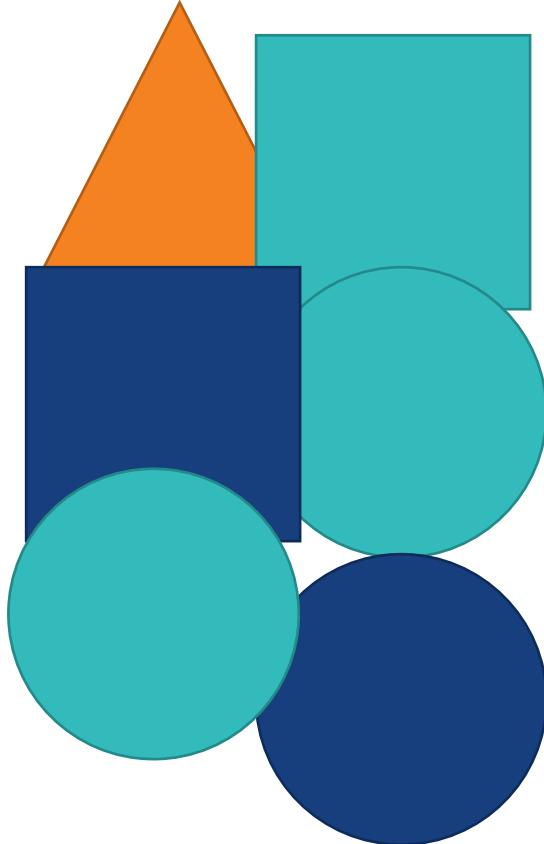


Example





What's the latent variable?





What's the latent variable?





Text Example

- **What's unusual?**
 - Anomaly detection



Date	Amount	Location
2 Mar	\$40	Penang
5 Mar	\$20	KL
17 Mar	\$30	KL
4 Apr	\$80	Ipooh
9 Apr	\$30	KL
14 Apr	\$70	KL
20 May	\$100	Johor
25 May	\$20	KL
31 May	\$3	Kiev
4 Jun	\$40	KL
23 Jun	\$50	KL
30 Jun	\$30	KL
16 Jul	\$70	Ipooh
16 Jul	\$50	Ipooh



DOCUMENT CLUSTERING



What is text clustering?

- **Clustering is the task of grouping a set of documents in such a way that the documents in each group are more “similar” to each other than to documents in other groups.**
- **Clustering lets you explore your data**
 - Many tools are interactive
- **You can understand your data better, e.g.:**
 - What groupings exist in your data?
 - How many are there? How big is each group?
 - What are the common terms?
 - Are there anomalies?



Clustering Example

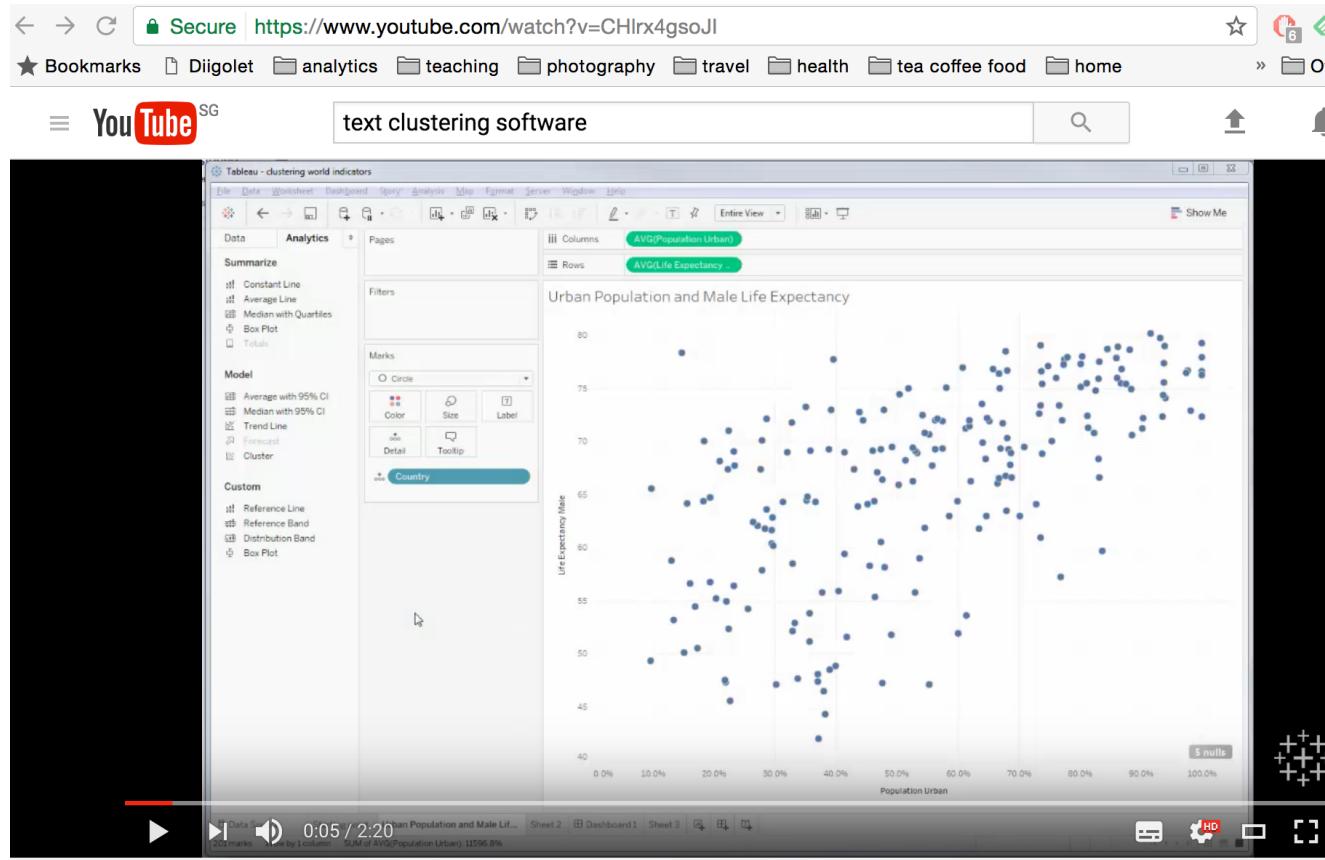


Tableau 10 Clustering Demo



Up next



Autoplay

From: <https://www.youtube.com/watch?v=CHlx4gsoJl>



Patent Clustering

Secure | <https://www.youtube.com/watch?v=Z-4S7kloHa8>

Bookmarks Diigolet analytics teaching photography travel health tea coffee food home art culture misc Privacy

YouTube SG patent clustering

Noogle

"Internet of Things" OR IoT

Cloud Data (88)
Switch Module (84)
Connecting Plate (75)
Lock Controlling (75)
Management platform Server (75)
Smart home based on Internet of Things (75)

MULTI-FUNCTIONAL SMART LAWNMOWER BASED ON INTERNET OF THINGS
A multi-functional smart lawnmower based on the Internet of Things, comprising a machine frame (1), an operating controller, an air water supply pipe (2), an operating URL: https://worldwide.espacenet.com/publicationDetails?id=WO2017008623

METHOD AND APPARATUS FOR GENERATING PACKET DATA NETWORK CONN
The present disclosure relates to a communication method and system for converging a 5th-Generation (5G) communication system and a system for supporting a higher data rate beyond a 4th-Generation URL: https://worldwide.espacenet.com/publicationDetails?id=CN104201703

METHOD AND APPARATUS FOR GENERATING PACKET DATA NETWORK CONN
The present disclosure relates to a communication method and system for converging a 5th-Generation (5G) communication system for supporting higher data rates beyond a 4th-Generation URL: https://worldwide.espacenet.com/publicationDetails?id=CC10581720

METHOD AND APPARATUS FOR PERFORMING COMMUNICATION IN WIRELESS
The present disclosure relates to a communication scheme and system for converging a 5th-generation (5G) communication system for supporting a higher data rate beyond a 4-th generation URL: https://worldwide.espacenet.com/publicationDetails?id=CN10581720

METHOD AND APPARATUS FOR PERFORMING COMMUNICATION IN WIRELESS
The present disclosure relates to a communication scheme and system for converging a 5-th generation (5G) communication system for supporting a higher data rate beyond a 4-th generation URL: https://worldwide.espacenet.com/publicationDetails?id=CN10581720

METHOD AND SYSTEM FOR SYNCHRONIZING COMMUNICATION BETWEEN N
The present disclosure relates to a sensor network, Machine Type Communication (MTC), Machine-to-Machine (M2M) communication, and technology for Internet of Things (IoT). The URL: https://worldwide.espacenet.com/publicationDetails?id=CC10581640

SMART VEHICLE
A vehicular gesture control system includes a plurality of cameras mounted in a cabin to detect edges of an object; and a processor to translate the edges as mouse movement and mouse clicks URL: https://worldwide.espacenet.com/publicationDetails?id=CN10581640

SELF-SERVICE PASSENGER SERVICE SYSTEM BASED ON INTERNET OF THINGS
A self-service passenger service system based on Internet of Things comprises a ticket checking system (1), a central control system (2), a wireless communications apparatus (3), and a smart seat URL: https://worldwide.espacenet.com/publicationDetails?id=CN10581642

Method for performing a cognitive clustering of patent documents

Cloud Data (88)
Switch Module (84)
Connecting Plate (75)
Lock Controlling (75)
Management platform Server (75)
Smart home based on Internet of Things (75)

Internet of Things Nodes
Smart Terminal
Electronic Product
Locations Inside Production
Information of Product
Card Number
Intelligent Electronic
Application Layer
Controlling Electronic Devices
Machine Interface
Monitor a Water
Wireless Switch
Production Cost
Management platform Server
Switching power Supply
Timing Switch
Base Plate
Electronical Structure
Type Layer
Layer of Internet of Things
Connecting Layer
Network Node
Communication Node
Vehicle Terminal
Network Switch
Intelligent Lock
Noogles
Smart home based on Internet of Things

How to use cognitive clustering in 100 seconds

noggle.online

Subscribe 4

49 views

Add to Share More

Like 0 Dislike 0

From: <https://www.youtube.com/watch?v=Z-4S7kloHa8>



Notes about Clustering

- **Your clusters may surprise you**
 - Documents tend to fall into natural classes (clusters)
 - There will be some surprising ones (worth drilling down!)
- **You can control the number of clusters (depends on the algorithm)**
 - You don't want too many clusters (overfit!)
 - You don't want too few clusters (meaningless)
 - **Clusters should lead to fulfilling business outcomes**
- **You don't need training phase to create clusters**
 - Clustering can be language independent (but monolingual)



TOPIC MODELING



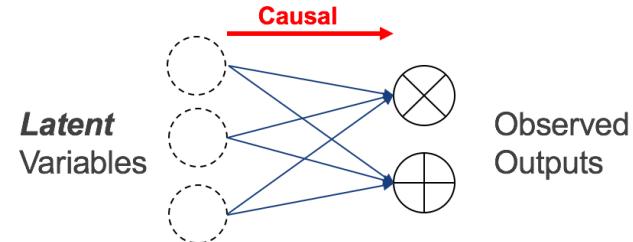
What is topic modeling?

- **Recall...**

- Latent (or hidden) variables?
- Output are collections of documents

- **“Topic” modeling**

- Can we figure out what discourses (==latent variables) would generate the collection of documents?
- These discourses are just bunches of words
 - If done well, the bunches of words would seem naturally to be together, e.g.,
 - “wag”, “bark”, “bone”, “bite”, “dog”
 - “pilot”, “plane”, “wing”, “flight”
- These bunches of words constitute **topics**

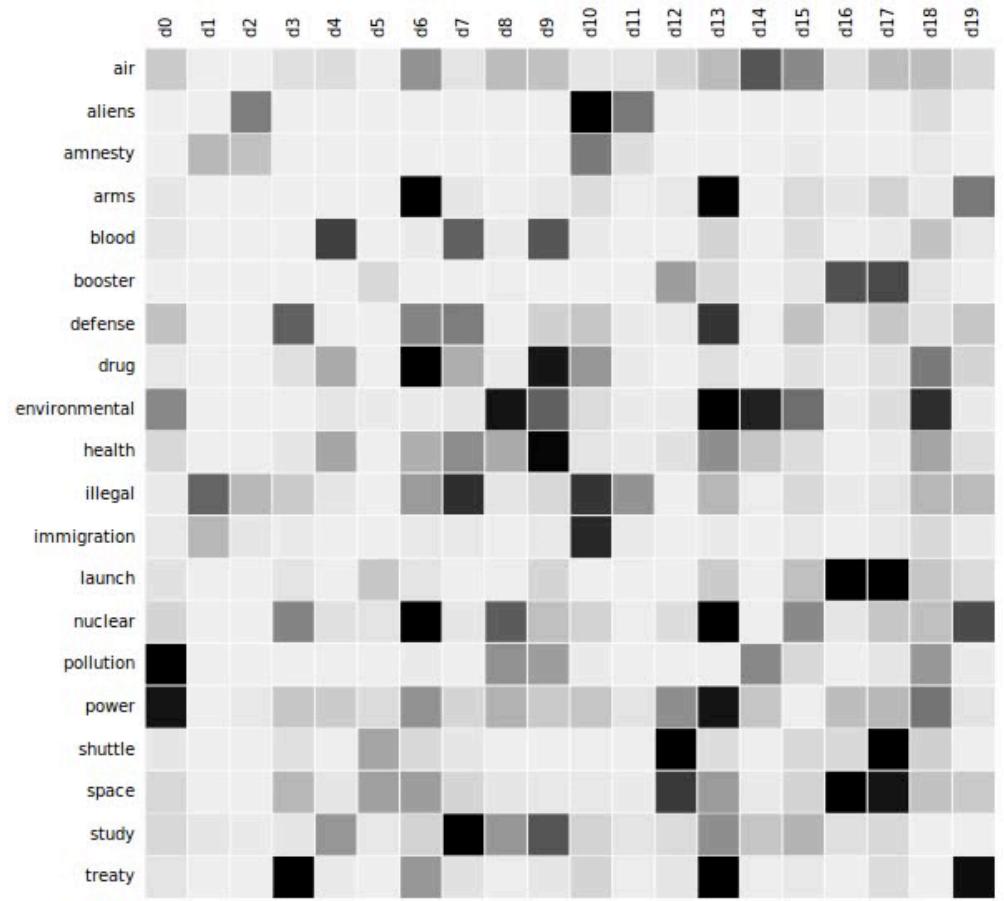


Animation of topic modeling

- **Columns = documents**
- **Rows = words**
- **Squares = frequency**
- **Darker = higher frequency**

- **Group:**
 - Documents using similar words
 - Words which occur in similar documents

***Resulting set of words
are “topics”***



From: http://topicmodels.west.uni-koblenz.de/ckling/tmt/svd_ap.html



LDA Topic Model Explanation

Secure <https://www.youtube.com/watch?v=3mHy4OSyRf0>

Bookmarks Diigolet analytics teaching photography travel health tea coffee food home art culture misc

YouTube SG topic modeling lda

“*all models are wrong, but some are useful*”

George E. P. Box

▶ ▶ ⏪ 18:17 / 20:36

LDA Topic Models



Andrius Knispelis

Subscribe 253

15,120 views

389 4

+ Add to Share More

From: <https://www.youtube.com/watch?v=3mHy4OSyRf0>



More examples of applications

- **Analysis of text, e.g.,**
 - Diachronic analysis:
 - Speeches during election campaign
 - Economy, abortion, build wall, reduce taxes,...
 - Speeches after taking office
 - Reduce taxes, create jobs, immigration, China,...
 - Contrast analysis:
 - Different candidates positions and issues
 - Characteristics of various media publications
 - Evolution of discourse over time and space
 - Word usage and diffusion
 - Engagement patterns



Reference & Resources

- **Fabrizio Sebastiani, *A Tutorial on Automated Text Categorization*,**
web.iit.ac.in/~jawahar/PRA-03/textCat.pdf
- **F. Aiolli, *Text Categorization*, downloaded from**
<http://www.math.unipd.it/~aiolli/corsi/SI-0607/Lez09.251006.pdf>
- **John Elder, Gary Miner, Bob Nisbet. *Practical Text Mining and Statistical Analysis for non-Structured Text Data Applications*, Academic Press, 2012**
- **Chris Manning & Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999**
- **Scott Weingart, *Topic Modeling for Humanists: A Guided Tour*, downloaded from**
<http://www.scottbot.net/HIAL/index.html@p=19113.html>
- **Ted Underwood, *Topic Modeling made just simple enough*, downloaded from**
<https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/>
- **NLP resources:** <http://nlp.stanford.edu/links/statnlp.html>