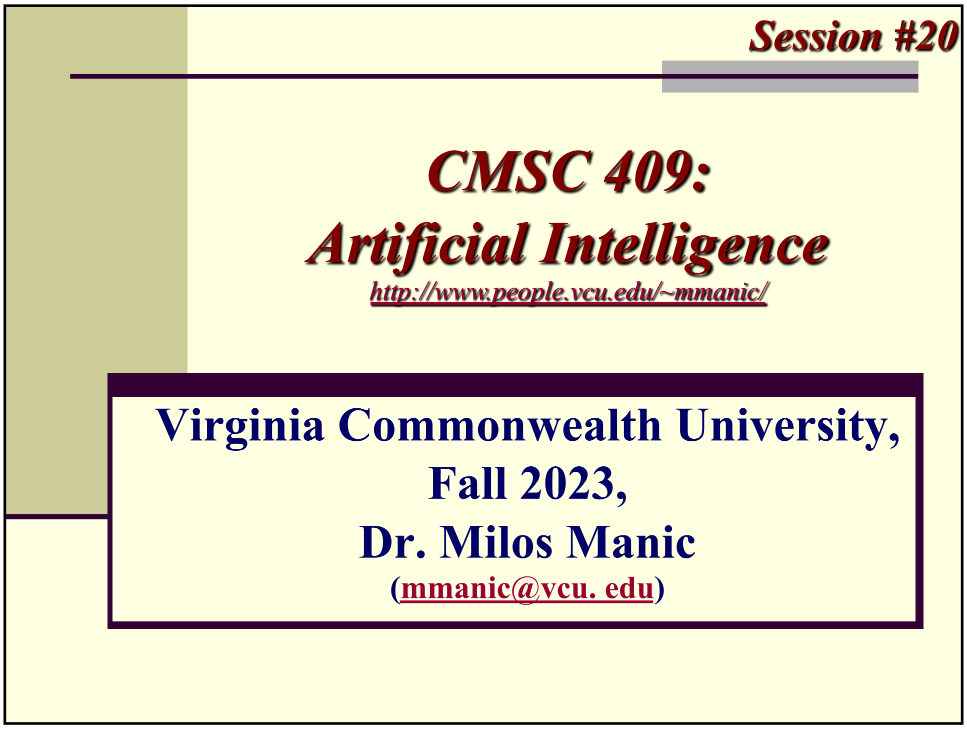
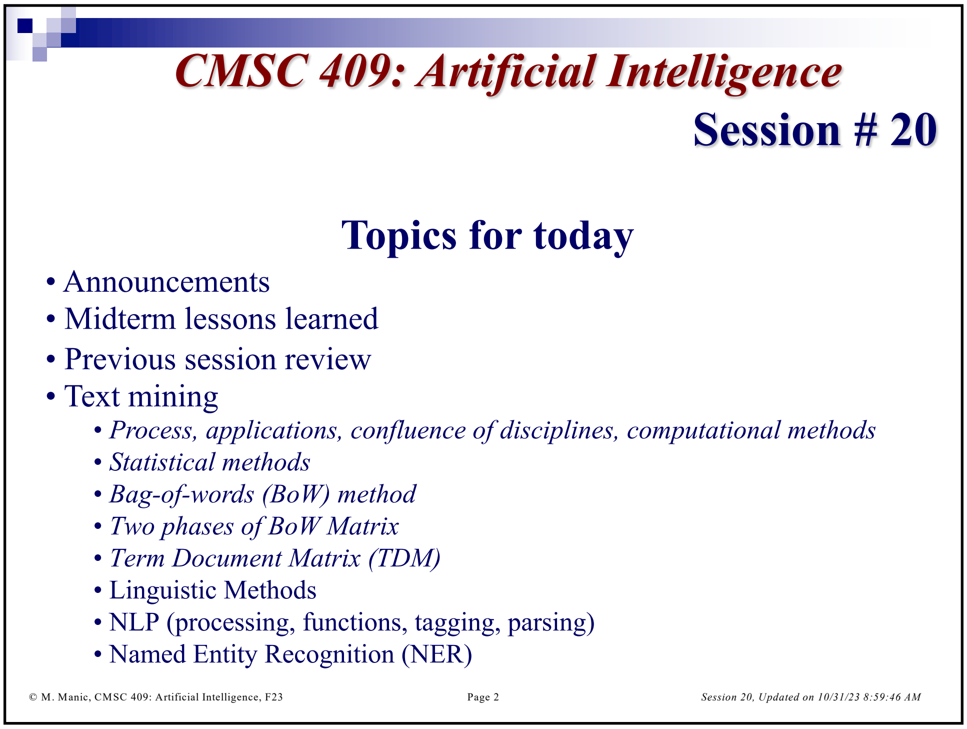


Session #20

CMSC 409: Artificial Intelligence

http://www.people.vcu.edu/~mmanic/

Virginia Commonwealth University,
Fall 2023,
Dr. Milos Manic
(mmanic@vcu. edu)

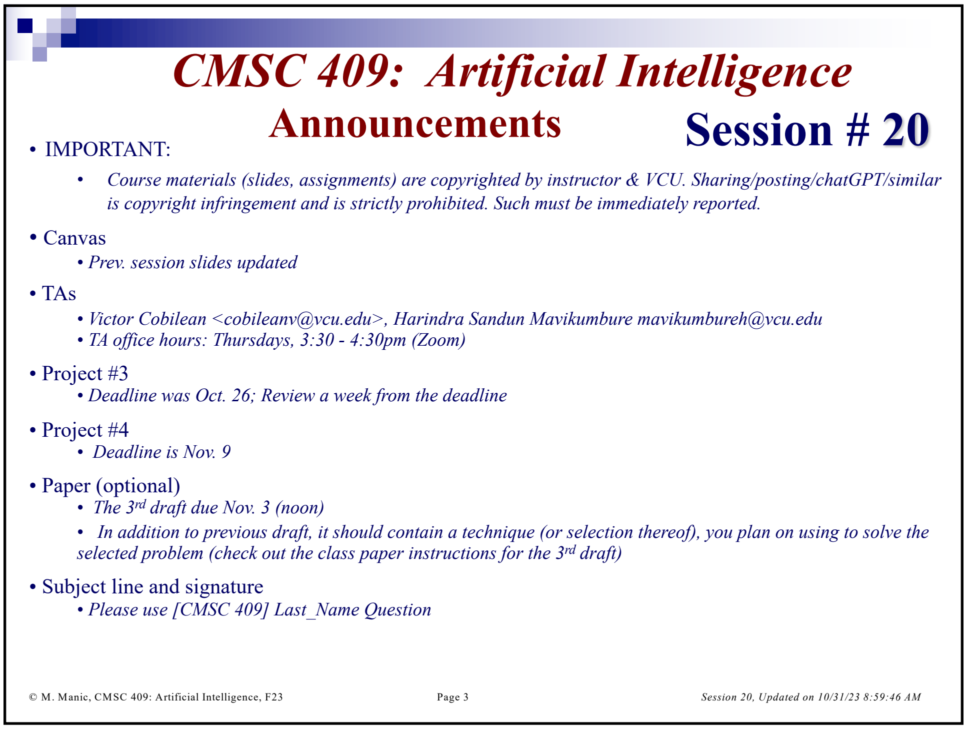


CMSC 409: Artificial Intelligence

Session # 20

Topics for today

- Announcements
- Midterm lessons learned
- Previous session review
- Text mining
  - *Process, applications, confluence of disciplines, computational methods*
  - *Statistical methods*
  - *Bag-of-words (BoW) method*
  - *Two phases of BoW Matrix*
  - *Term Document Matrix (TDM)*
  - Linguistic Methods
  - NLP (processing, functions, tagging, parsing)
  - Named Entity Recognition (NER)

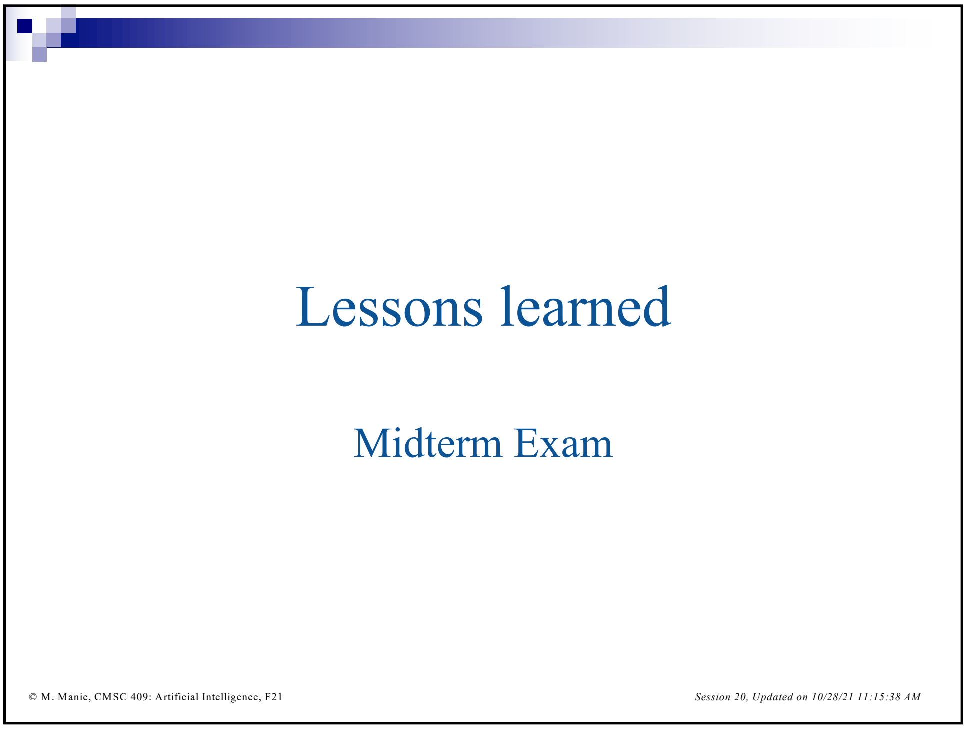


CMSC 409: Artificial Intelligence

Announcements

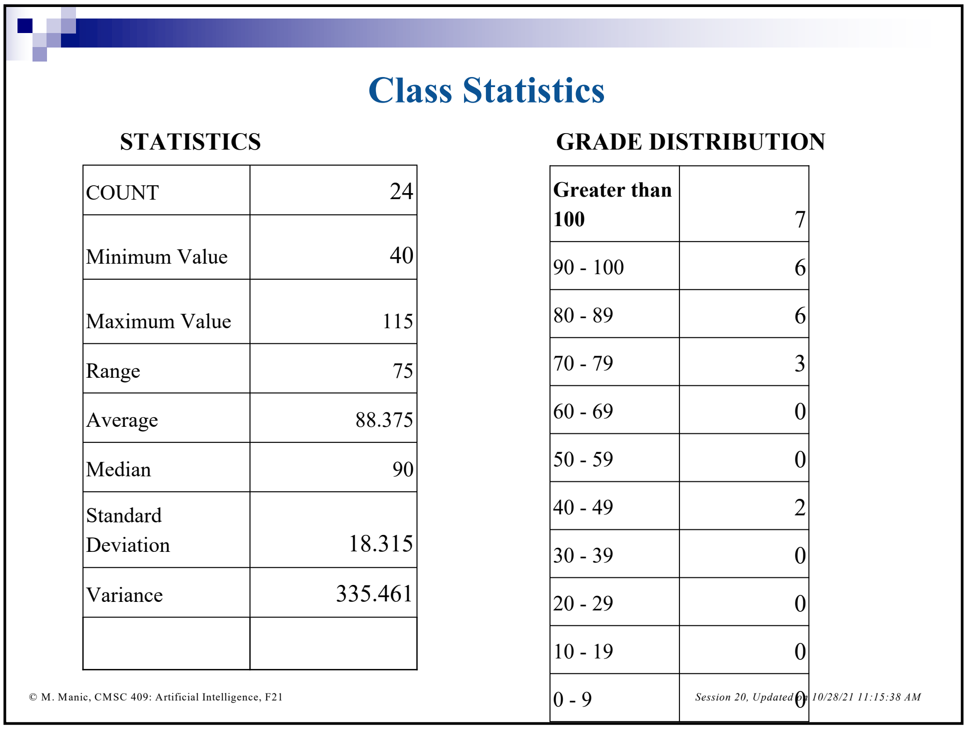
Session # 20

- IMPORTANT:
  - *Course materials (slides, assignments) are copyrighted by instructor & VCU. Sharing/posting/chatGPT/similar is copyright infringement and is strictly prohibited. Such must be immediately reported.*
- Canvas
  - *Prev. session slides updated*
- TAs
  - *Victor Cobilean <cobileanv@vcu.edu>, Harindra Sandun Mavikumbure mavikumbureh@vcu.edu*
  - *TA office hours: Thursdays, 3:30 - 4:30pm (Zoom)*
- Project #3
  - *Deadline was Oct. 26; Review a week from the deadline*
- Project #4
  - *Deadline is Nov. 9*
- Paper (optional)
  - *The 3rd draft due Nov. 3 (noon)*
  - *In addition to previous draft, it should contain a technique (or selection thereof), you plan on using to solve the selected problem (check out the class paper instructions for the 3rd draft)*
- Subject line and signature
  - *Please use [CMSC 409] Last_Name Question*



Lessons learned

Midterm Exam



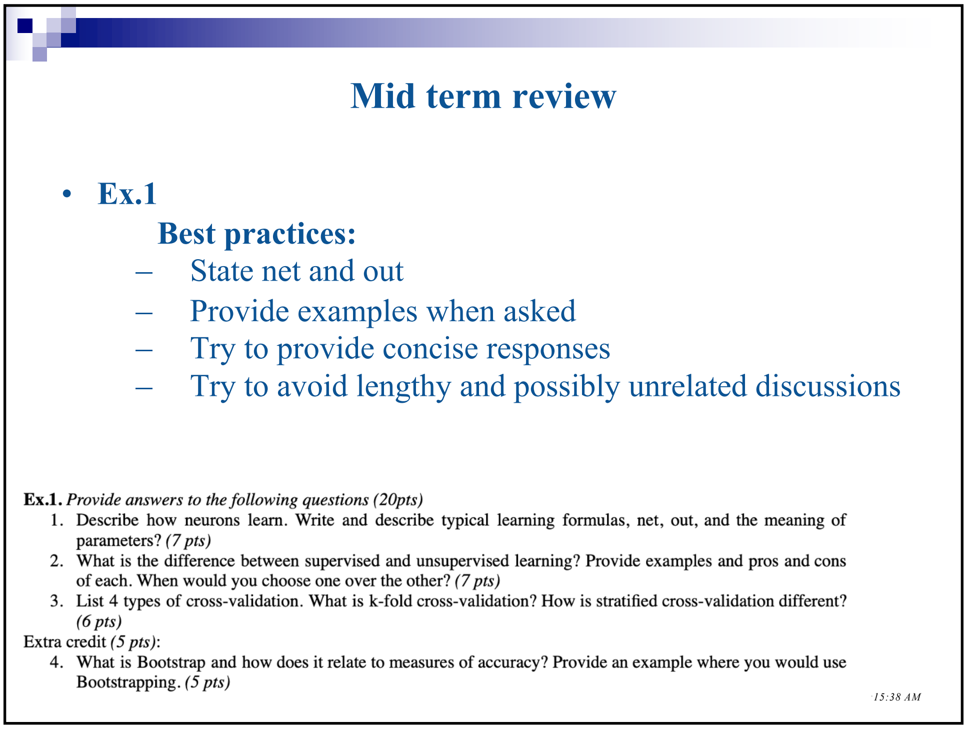
Class Statistics

STATISTICS

COUNT	24
Minimum Value	40
Maximum Value	115
Range	75
Average	88.375
Median	90
Standard Deviation	18.315
Variance	335.461

GRADE DISTRIBUTION

Greater than 100	7
90 - 100	6
80 - 89	6
70 - 79	3
60 - 69	0
50 - 59	0
40 - 49	2
30 - 39	0
20 - 29	0
10 - 19	0
0 - 9	0



Mid term review

- **Ex.1**

  Best practices:
  - State net and out
  - Provide examples when asked
  - Try to provide concise responses
  - Try to avoid lengthy and possibly unrelated discussions

Ex.1. *Provide answers to the following questions (20pts)*

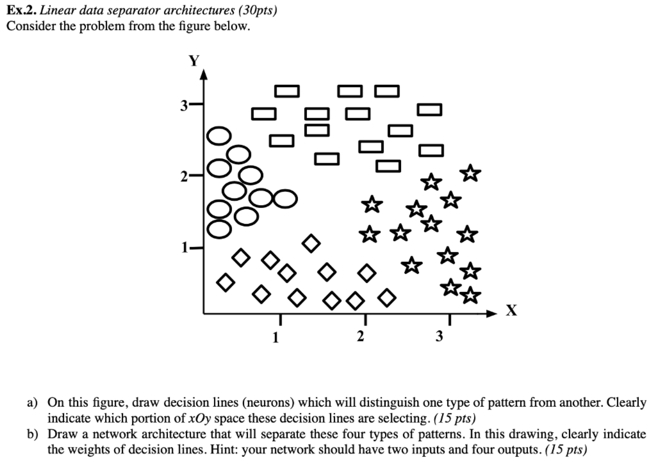
1. Describe how neurons learn. Write and describe typical learning formulas, net, out, and the meaning of parameters? *(7 pts)*
2. What is the difference between supervised and unsupervised learning? Provide examples and pros and cons of each. When would you choose one over the other? *(7 pts)*
3. List 4 types of cross-validation. What is k-fold cross-validation? How is stratified cross-validation different? *(6 pts)*

Extra credit *(5 pts)*:

4. What is Bootstrap and how does it relate to measures of accuracy? Provide an example where you would use Bootstrapping. *(5 pts)*

Mid term review

- **Ex. 2**



Ex.2. *Linear data separator architectures (30pts)*
Consider the problem from the figure below.

a) On this figure, draw decision lines (neurons) which will distinguish one type of pattern from another. Clearly indicate which portion of *xOy* space these decision lines are selecting. *(15 pts)*

b) Draw a network architecture that will separate these four types of patterns. In this drawing, clearly indicate the weights of decision lines. Hint: your network should have two inputs and four outputs. *(15 pts)*

Mid term review

- **Ex. 2**

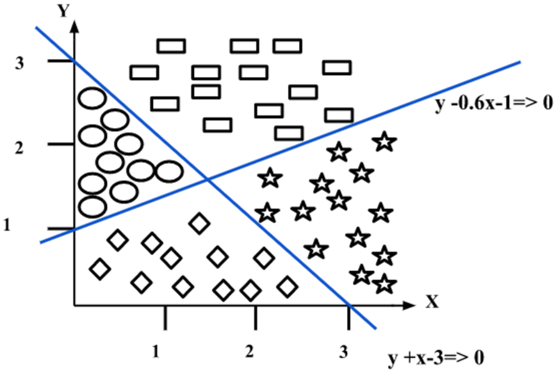
  Best practices:

  - State input weights/threshold (use decision line coefficients): $w_{x*}x+w_{y*}y+w_c>=0$
  - Clearly indicate the portion of the xOy space selected (or deselected) by decision line ("small arrow")
  - Verify architecture; network architecture should start with two inputs, end with four outputs (for the 4 types of patterns to be distinguished)
  - To state functionality of the output neuron and output intended to recognize (rectangle, diamond, etc.)

Mid term review

- Ex. 2 a)

Solution

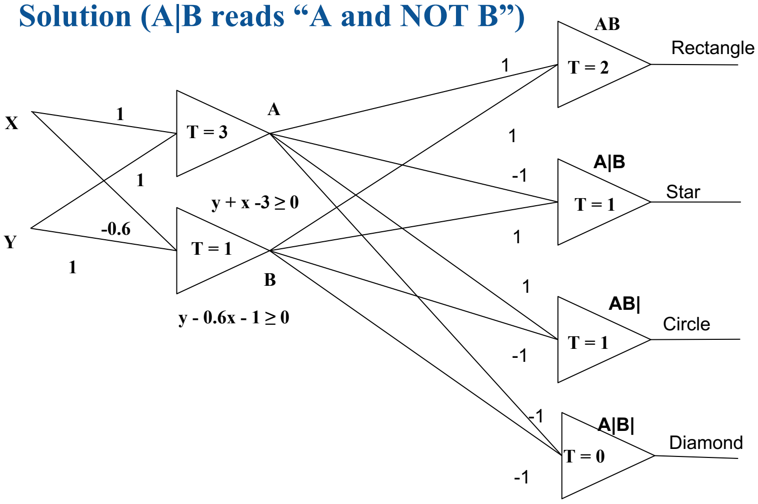


a) On this figure, draw decision lines (neurons) which will distinguish one type of pattern from another. Clearly indicate which portion of *xOy* space these decision lines are selecting. *(15 pts)*

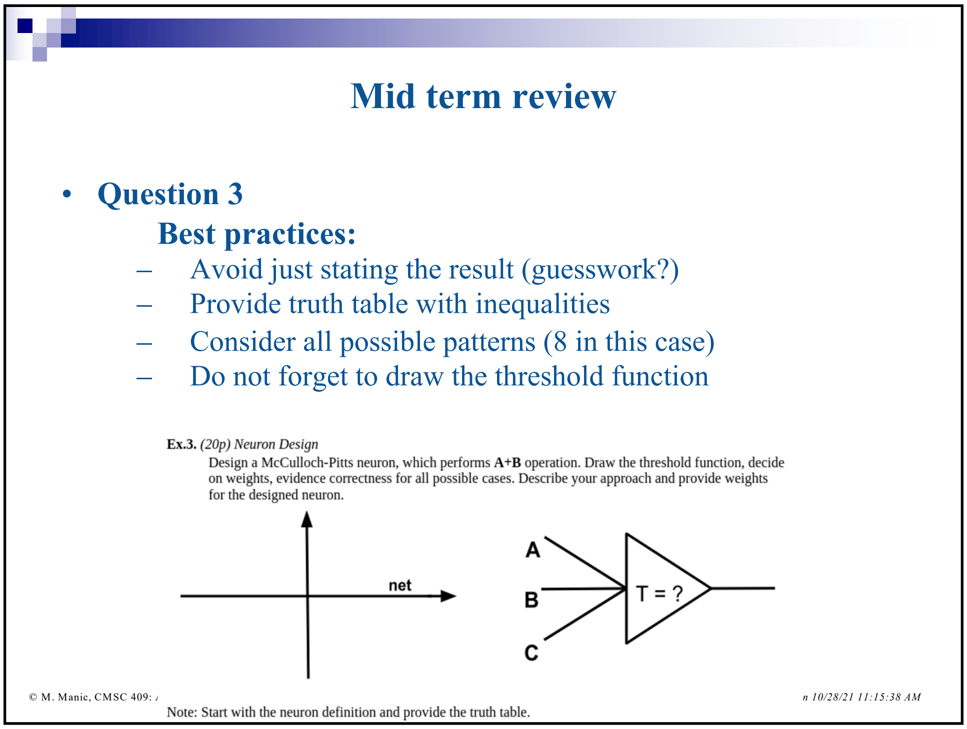
Mid term review

- Ex. 2 b)

Solution (A|B reads "A and NOT B")



b) Draw a network architecture that will separate these four types of patterns. In this drawing, clearly indicate the weights of decision lines. Hint: your network should have two inputs and four outputs. *(15 pts)*



Mid term review

- **Question 3**

  Best practices:
  - Avoid just stating the result (guesswork?)
  - Provide truth table with inequalities
  - Consider all possible patterns (8 in this case)
  - Do not forget to draw the threshold function

Ex.3. *(20p) Neuron Design*

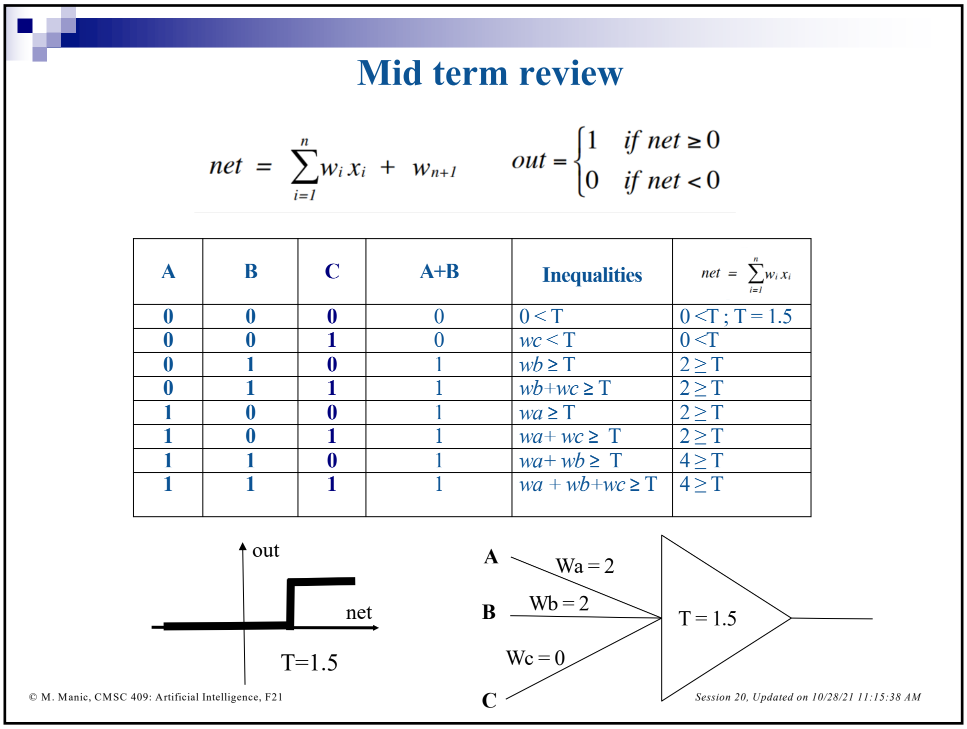
Design a McCulloch-Pitts neuron, which performs **A+B** operation. Draw the threshold function, decide on weights, evidence correctness for all possible cases. Describe your approach and provide weights for the designed neuron.

© M. Manic, CMSC 409: ,

n 10/28/21 11:15:38 AM

Note: Start with the neuron definition and provide the truth table.

11



Mid term review

$$net = \sum_{i=1}^{n} w_i x_i + w_{n+1} \qquad out = \begin{cases} 1 & if\ net \geq 0 \\ 0 & if\ net < 0 \end{cases}$$

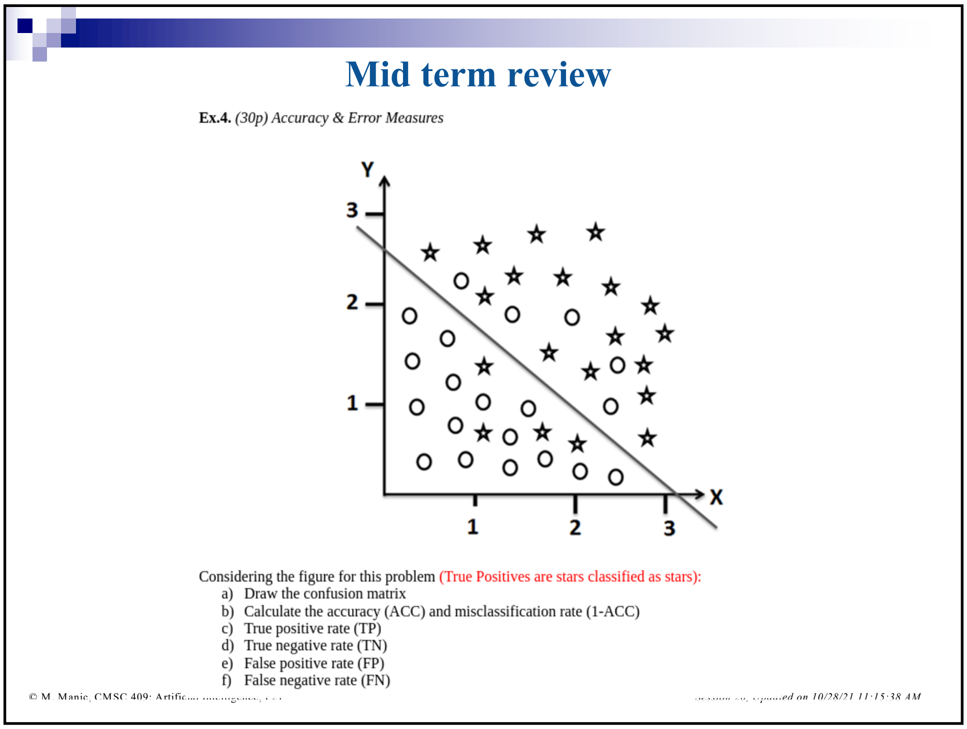
A	B	C	A+B	Inequalities	$net = \sum_{i=1}^{n} w_i x_i$
0	0	0	0	0 < T	0 <T ; T = 1.5
0	0	1	0	*wc* < T	0 <T
0	1	0	1	*wb* ≥ T	2 ≥ T
0	1	1	1	*wb+wc* ≥ T	2 ≥ T
1	0	0	1	*wa* ≥ T	2 ≥ T
1	0	1	1	*wa+ wc* ≥ T	2 ≥ T
1	1	0	1	*wa+ wb* ≥ T	4 ≥ T
1	1	1	1	*wa + wb+wc* ≥ T	4 ≥ T

out
net
T=1.5
A
B
C
Wa = 2
Wb = 2
Wc = 0
T = 1.5

© M. Manic, CMSC 409: Artificial Intelligence, F21

Session 20, Updated on 10/28/21 11:15:38 AM

12

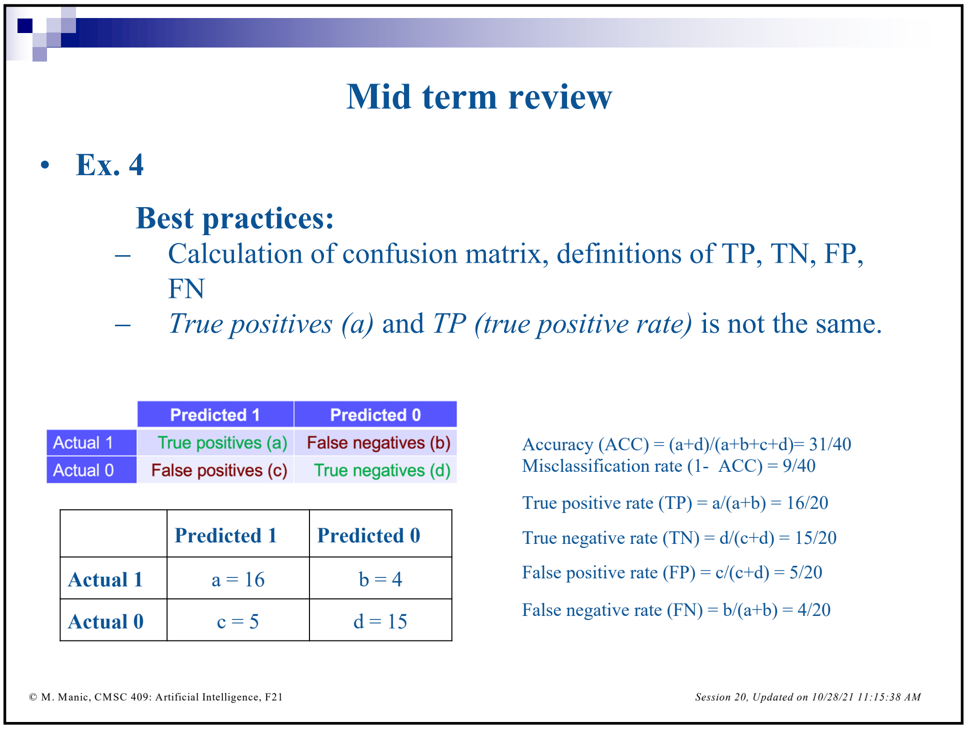


Mid term review

Ex.4. *(30p) Accuracy & Error Measures*

Considering the figure for this problem (True Positives are stars classified as stars):

a) Draw the confusion matrix
b) Calculate the accuracy (ACC) and misclassification rate (1-ACC)
c) True positive rate (TP)
d) True negative rate (TN)
e) False positive rate (FP)
f) False negative rate (FN)



Mid term review

- **Ex. 4**

Best practices:

– Calculation of confusion matrix, definitions of TP, TN, FP, FN
– *True positives (a)* and *TP (true positive rate)* is not the same.

	Predicted 1	Predicted 0
Actual 1	True positives (a)	False negatives (b)
Actual 0	False positives (c)	True negatives (d)

	Predicted 1	**Predicted 0**
Actual 1	a = 16	b = 4
Actual 0	c = 5	d = 15

Accuracy (ACC) = (a+d)/(a+b+c+d)= 31/40
Misclassification rate (1- ACC) = 9/40

True positive rate (TP) = a/(a+b) = 16/20

True negative rate (TN) = d/(c+d) = 15/20

False positive rate (FP) = c/(c+d) = 5/20

False negative rate (FN) = b/(a+b) = 4/20

Mid term review

Ex.5. Extra Credit Reinforcement and Q-Learning
Solve and explain the Family Crisis problem. (15 pts)

These family members should cross to the other side of the bridge.
Each person (A, B, C, D) crosses the bridge at different speeds: 1 second (person A), 3 seconds (person B), 6 seconds (person C) and 10 seconds (person D). The bridge can hold a maximum of 2 persons at a time. A pair must walk together at the rate of the slower person. The lamp will last 20 seconds only. Write down the steps to solve this problem.

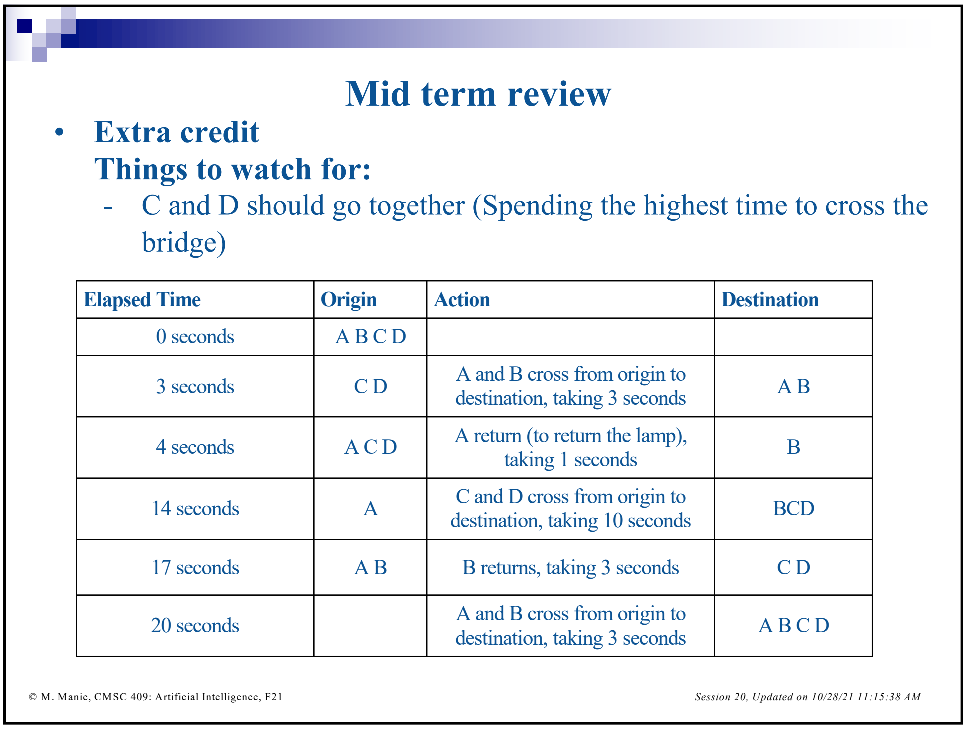
Note: Please note that it is night, so they must have a lamp every time they are crossing the bridge.

Starting position:

Mid term review

- **Extra credit**

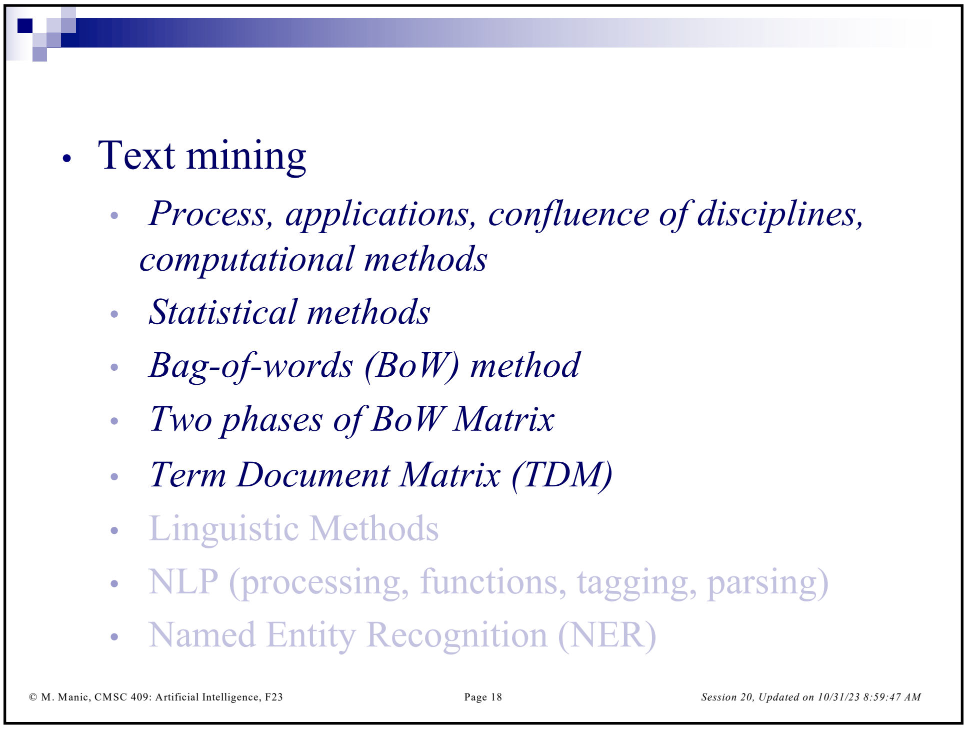
  Best practices:
  - Present the steps clearly
  - Assign positive reward for person A, B, C, D crossing the bridge within 20 sec. (reward could be +30)
  - Assign slightly negative reward for each time step takes to cross all 4 persons. (reward could be -1)
  - Assign large negative reward when lamp goes off, meaning exceeds 20 seconds. (reward could be -20)



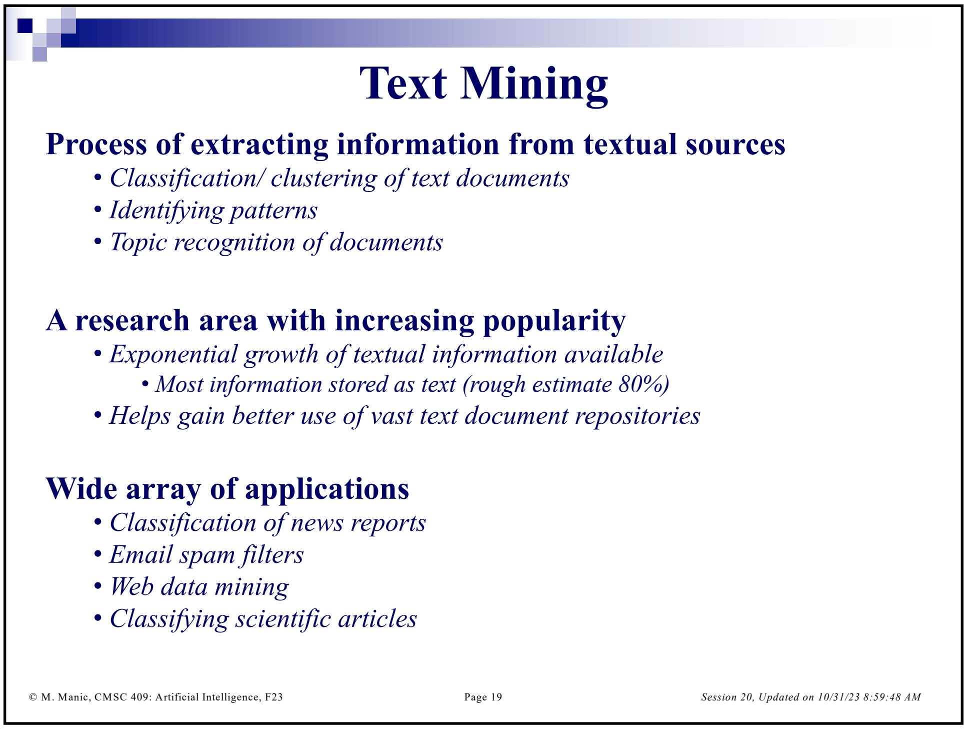
Mid term review

- **Extra credit**
  Things to watch for:
  - C and D should go together (Spending the highest time to cross the bridge)

Elapsed Time	**Origin**	**Action**	**Destination**
0 seconds	A B C D		
3 seconds	C D	A and B cross from origin to destination, taking 3 seconds	A B
4 seconds	A C D	A return (to return the lamp), taking 1 seconds	B
14 seconds	A	C and D cross from origin to destination, taking 10 seconds	BCD
17 seconds	A B	B returns, taking 3 seconds	C D
20 seconds		A and B cross from origin to destination, taking 3 seconds	A B C D



- Text mining
  - *Process, applications, confluence of disciplines, computational methods*
  - *Statistical methods*
  - *Bag-of-words (BoW) method*
  - *Two phases of BoW Matrix*
  - *Term Document Matrix (TDM)*
  - Linguistic Methods
  - NLP (processing, functions, tagging, parsing)
  - Named Entity Recognition (NER)



Text Mining

Process of extracting information from textual sources

- *Classification/ clustering of text documents*
- *Identifying patterns*
- *Topic recognition of documents*

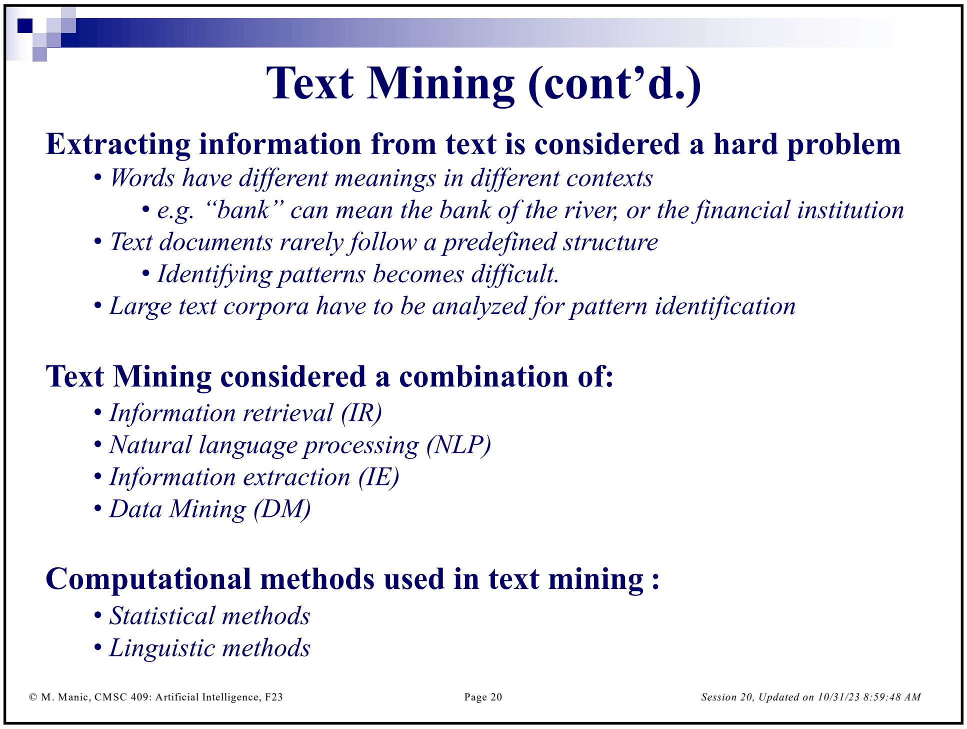
A research area with increasing popularity

- *Exponential growth of textual information available*
  - *Most information stored as text (rough estimate 80%)*
- *Helps gain better use of vast text document repositories*

Wide array of applications

- *Classification of news reports*
- *Email spam filters*
- *Web data mining*
- *Classifying scientific articles*

19



Text Mining (cont'd.)

Extracting information from text is considered a hard problem

- *Words have different meanings in different contexts*
  - *e.g. "bank" can mean the bank of the river, or the financial institution*
- *Text documents rarely follow a predefined structure*
  - *Identifying patterns becomes difficult.*
- *Large text corpora have to be analyzed for pattern identification*

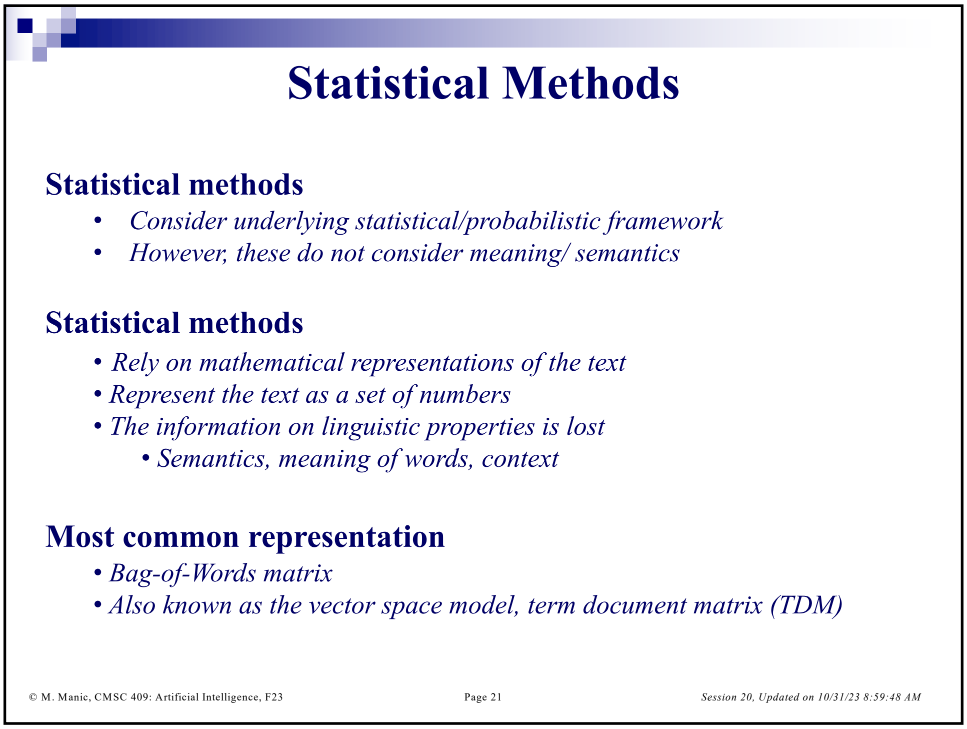
Text Mining considered a combination of:

- *Information retrieval (IR)*
- *Natural language processing (NLP)*
- *Information extraction (IE)*
- *Data Mining (DM)*

Computational methods used in text mining :

- *Statistical methods*
- *Linguistic methods*

20



Statistical Methods

Statistical methods

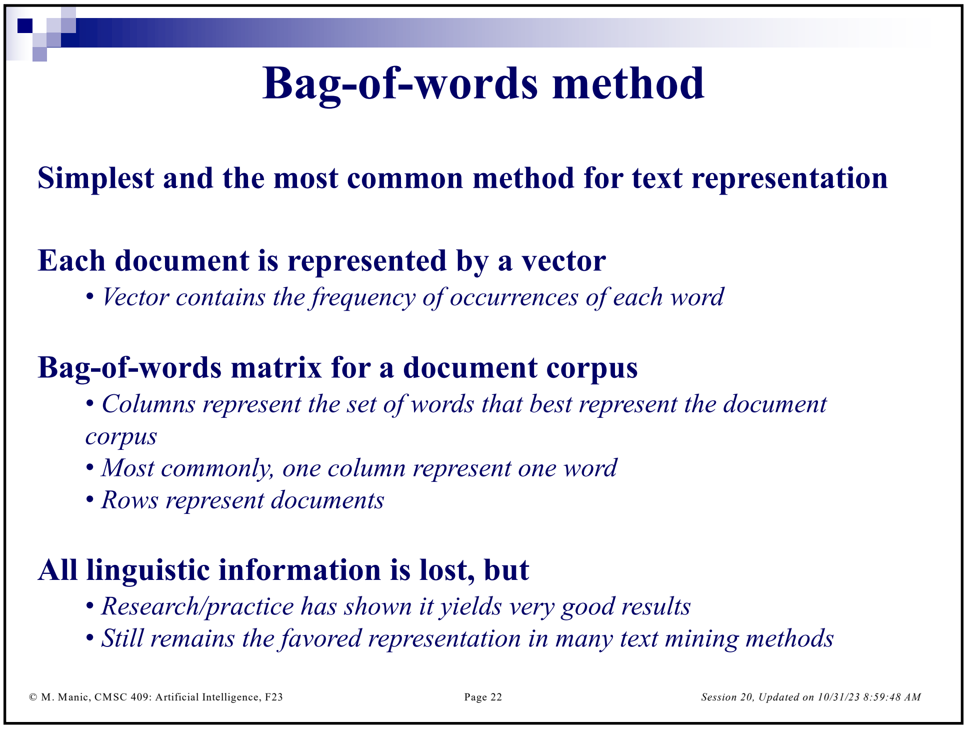
- *Consider underlying statistical/probabilistic framework*
- *However, these do not consider meaning/ semantics*

Statistical methods

- *Rely on mathematical representations of the text*
- *Represent the text as a set of numbers*
- *The information on linguistic properties is lost*
  - *Semantics, meaning of words, context*

Most common representation

- *Bag-of-Words matrix*
- *Also known as the vector space model, term document matrix (TDM)*



Bag-of-words method

Simplest and the most common method for text representation

Each document is represented by a vector

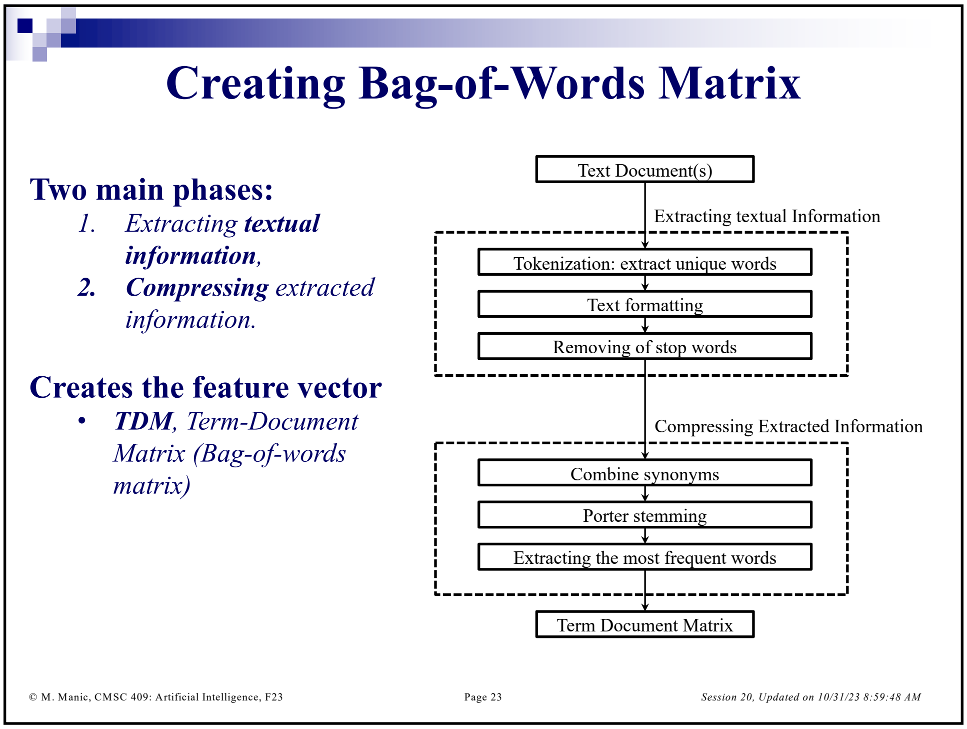
- *Vector contains the frequency of occurrences of each word*

Bag-of-words matrix for a document corpus

- *Columns represent the set of words that best represent the document corpus*
- *Most commonly, one column represent one word*
- *Rows represent documents*

All linguistic information is lost, but

- *Research/practice has shown it yields very good results*
- *Still remains the favored representation in many text mining methods*



Creating Bag-of-Words Matrix

Two main phases:

1. *Extracting **textual information**,*
2. ***Compressing** extracted information.*

Creates the feature vector

- ***TDM**, Term-Document Matrix (Bag-of-words matrix)*

Text Document(s)

Extracting textual Information

Tokenization: extract unique words

Text formatting

Removing of stop words

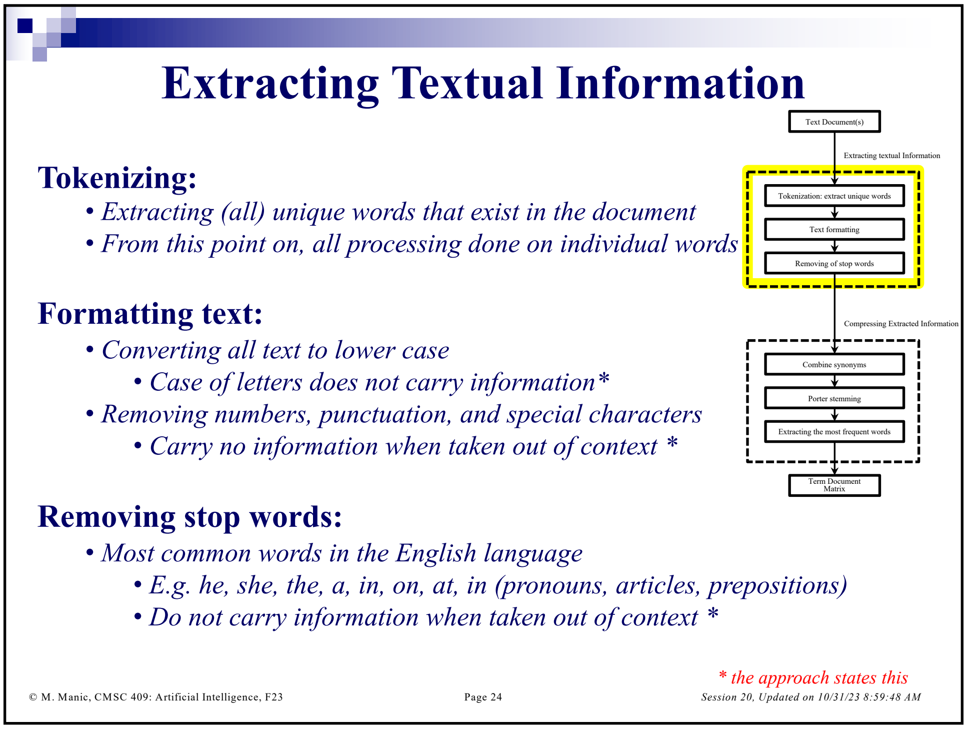
Compressing Extracted Information

Combine synonyms

Porter stemming

Extracting the most frequent words

Term Document Matrix



Extracting Textual Information

Text Document(s)

Extracting textual Information

Tokenization: extract unique words

Text formatting

Removing of stop words

Compressing Extracted Information

Combine synonyms

Porter stemming

Extracting the most frequent words

Term Document Matrix

Tokenizing:

- *Extracting (all) unique words that exist in the document*
- *From this point on, all processing done on individual words*

Formatting text:

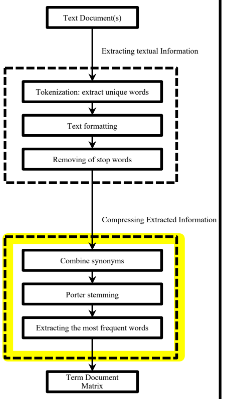
- *Converting all text to lower case*
  - *Case of letters does not carry information**
- *Removing numbers, punctuation, and special characters*
  - *Carry no information when taken out of context **

Removing stop words:

- *Most common words in the English language*
  - *E.g. he, she, the, a, in, on, at, in (pronouns, articles, prepositions)*
  - *Do not carry information when taken out of context **

** the approach states this*

Compressing Extracted Information


Text Document(s)
Extracting textual Information
Tokenization: extract unique words
Text formatting
Removing of stop words
Compressing Extracted Information
Combine synonyms
Porter stemming
Extracting the most frequent words
Term Document Matrix

After textual information extraction

- *What remains is: unique keywords existing in the document which carry some information**
- *Compressing selects the subset (of extracted words) carrying the most information*

Combining synonyms

- *Combining words with the similar meaning*
- *Reduces the number of words used*

Stemming

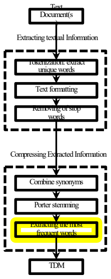
- *Reducing words to their word stem, or base*
- *Converting the word to its basic form and combining words*
  - *E.g. - **listening, listened** get stemmed to **listen***
  - *E.g. - coming, came, come => come (basal form)*
  - *E.g. - eating, eats, eaten => eat (basal form)*

** the approach states this*

Compressing Extracted Information contd.

Extracting most frequent word

- *Words that do not appear often do not carry significant information about the document corpus*
- ***Most frequent in a whole corpus of documents***

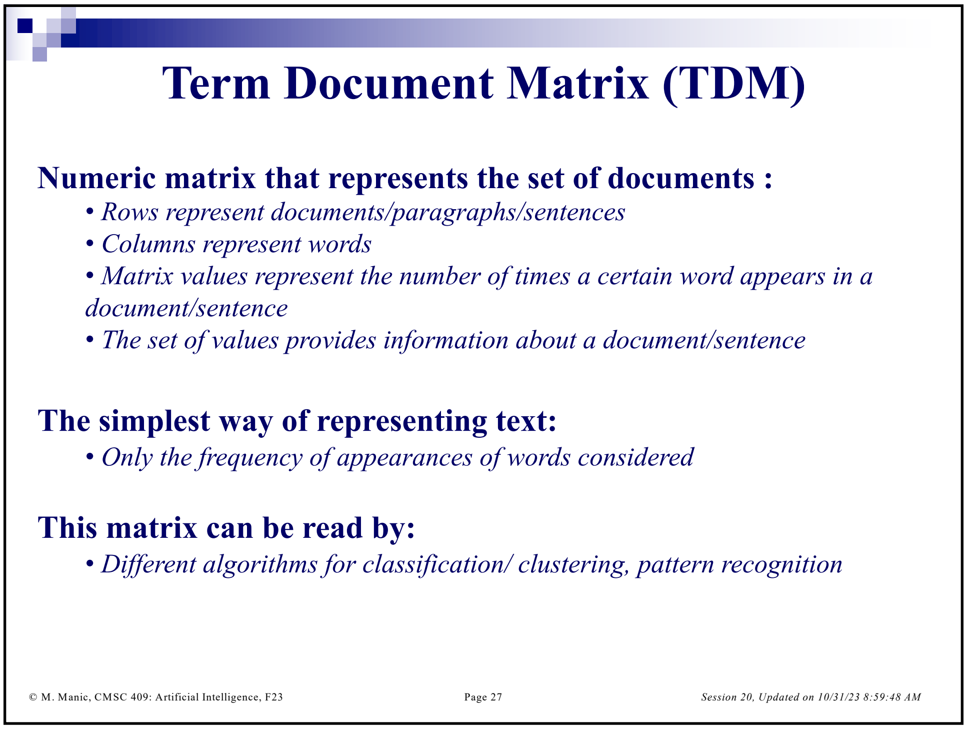

Text Document(s
Extracting textual Information
Tokenization: extract unique words
Text formatting
Removing of stop words
Compressing Extracted Information
Combine synonyms
Porter stemming
Extracting the most frequent words
TDM

Remaining set of words

- *Does not contain special characters, numbers and punctuations*
- *Does not contain stop words*
- *Does not contain words with redundant meaning*
- *Does not contain words below a certain frequency of appearance*

Thus

- *The generalized set of words that best represent the document(s) is remaining*



Term Document Matrix (TDM)

Numeric matrix that represents the set of documents :

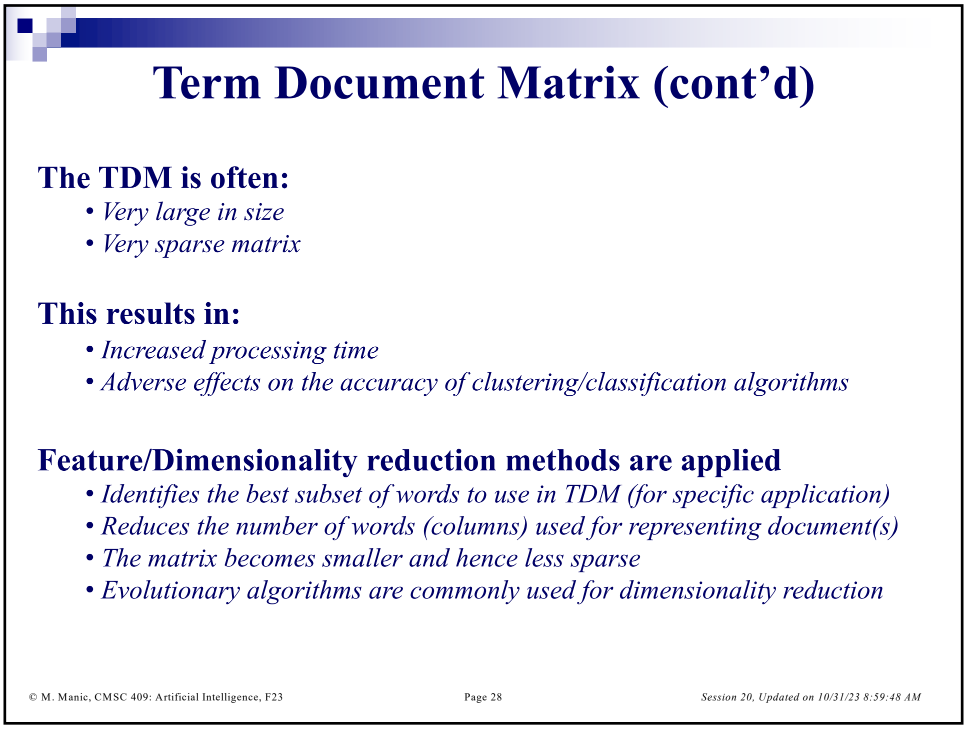
- *Rows represent documents/paragraphs/sentences*
- *Columns represent words*
- *Matrix values represent the number of times a certain word appears in a document/sentence*
- *The set of values provides information about a document/sentence*

The simplest way of representing text:

- *Only the frequency of appearances of words considered*

This matrix can be read by:

- *Different algorithms for classification/ clustering, pattern recognition*



Term Document Matrix (cont'd)

The TDM is often:

- *Very large in size*
- *Very sparse matrix*

This results in:

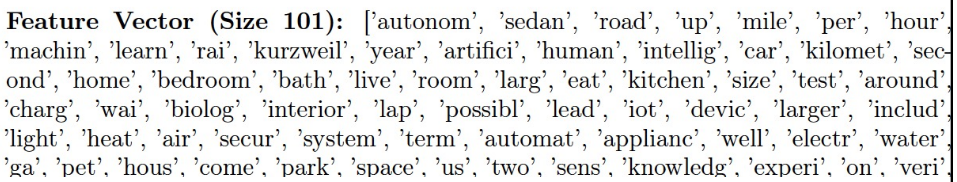
- *Increased processing time*
- *Adverse effects on the accuracy of clustering/classification algorithms*

Feature/Dimensionality reduction methods are applied

- *Identifies the best subset of words to use in TDM (for specific application)*
- *Reduces the number of words (columns) used for representing document(s)*
- *The matrix becomes smaller and hence less sparse*
- *Evolutionary algorithms are commonly used for dimensionality reduction*

Term Document Matrix (cont’d)

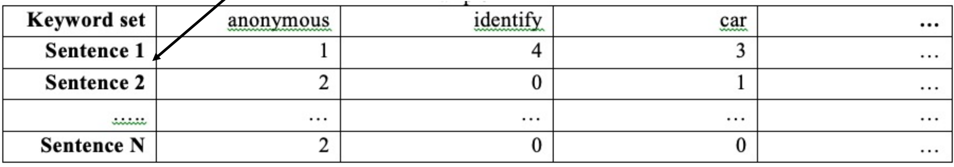
Feature Vector, example:



Feature Vector (Size 101): [’autonom’, ’sedan’, ’road’, ’up’, ’mile’, ’per’, ’hour’, ’machin’, ’learn’, ’rai’, ’kurzweil’, ’year’, ’artifici’, ’human’, ’intellig’, ’car’, ’kilomet’, ’second’, ’home’, ’bedroom’, ’bath’, ’live’, ’room’, ’larg’, ’eat’, ’kitchen’, ’size’, ’test’, ’around’, ’charg’, ’wai’, ’biolog’, ’interior’, ’lap’, ’possibl’, ’lead’, ’iot’, ’devic’, ’larger’, ’includ’, ’light’, ’heat’, ’air’, ’secur’, ’system’, ’term’, ’automat’, ’applianc’, ’well’, ’electr’, ’water’, ’ga’, ’pet’, ’hous’, ’come’, ’park’, ’space’, ’us’, ’two’, ’sens’, ’knowledg’, ’experi’, ’on’, ’veri’,

what we are comparing (could be a document, paragraph, sentence)

TDM, example:



Keyword set	anonymous	identify	car	…
Sentence 1	1	4	3	…
Sentence 2	2	0	1	…
……	…	…	…	…
Sentence N	2	0	0	…

Things to remember…

- **Feature vector can be long…**
  - *Remember, there is not “right” or “wrong” number (dimensionality)*
  - *The process is in a way “unsupervised”*
  - *But, the threshold for word occurrence frequency will affect it greatly - that's your "knob" to control it.*

- **Our clustering algorithms …**
  - *Can handle large dimensionality, but..*
  - *Dimensionality reduction may help...*