University of Southampton

COMP6237 Data Mining
Lecture 13: Revision

Zhiwu Huang

Zhiwu.Huang@soton.ac.uk

Lecturer (Assistant Professor) @ VLC of ECS

University of Southampton

Lecture slides available here:

http://comp6237.ecs.soton.ac.uk/zh.html

(Thanks to Prof. Jonathon Hare and Dr. Jo Grundy for providing the lecture materials used to develop the slides.)

Exam timetable: https://www.southampton.ac.uk/leaf_content/uploads/generic_semester_two_exam_timetable_202425.pdf

Image credit: FLUX

ECS Revision Week - Snacks and Safety Nets

I would like to take the opportunity to introduce a new initiative to aid ECS students which is to be run during revision week **12 – 16 May 2025**. During revision weeks, a stand will be set up in two of the designated student spaces (**Social Space Building 16 and Social Space Building 60)**. Beverages and snacks will be delivered twice daily. A member of ECS staff will be present throughout the day in one-hour increments to engage with students, offering refreshments and informal conversation about their revision progress. **This will run between 10 am – 4 pm daily during the week of 12 – 16 May.**

Exam Format

- MCQs + Computer-aided
  - Shoaib (40 marks) + Zhiwu (40 Marks) + Markus (20 Marks)

- My part (15 MCQs):
  - MCQs 1-12 : 2 marks each (easy/normal), one single answer each
  - MCQs 13-15: 5 or 6 marks each (hard)
    - Each MCQ contains 2 or 3 subquestions
    - Each subquestion has one single answer

MCQ Examples

More examples: https://comp6237.ecs.soton.ac.uk/lectures/pdf/13-Revision-MCQ_examples_ZH.pdf

Question 3: How does the DBSCAN algorithm handle outliers in the dataset?

A. Assigns outliers to the nearest cluster
B. Removes outliers from the dataset
C. Labels outliers as noise
D. Forms separate clusters for outliers

Question 4: Which clustering algorithm is particularly effective for identifying clusters with varying shapes and densities?

A. K-means
B. DBSCAN
C. Hierarchical clustering
D. None of above

MCQ Examples

Question 5:

	Feature 1	Feature 2	Feature 3
Instance1	xx	xx	xx
Instance2	xx	xx	xx
Instance3	xx	xx	xx

Assume user A wanted to compute Euclidean distance for each pair of instances. Which option is correct?

A. …
B. …
C. …
D. None of the above

MCQ Examples

Question 6: Given the 1D data points {1, 2, 3, 4, 6, 7, 8, 9, 10}, if k-means clustering is run with k=2 and initial cluster centers at C1=2 and C2=8 using Manhattan distance.

S1: Which points will be in the cluster with center 8 in the first iteration?

A. {4,6,7,8,9,10}
B. {6,7,8,9,10}
C. {7,8,9,10}
D. {8,9,10}

S2: xxxxxxxxxxxx:

A...

B....

C...

D...

	Distance to C1=2	Distance to C2=8	Assigned Center
1	|1-2|	|1-8|	C1
2	|2-2|	|2-8|	C1
3	|3-2|	|3-8|	C1
4	|4-2|	|4-8|	C1
6	|6-2|	|6-8|	C2
7	|7-2|	|7-8|	C2
8	|8-2|	|8-8|	C2
9	|9-2|	|9-8|	C2
10	|10-2|	|10-8|	C2

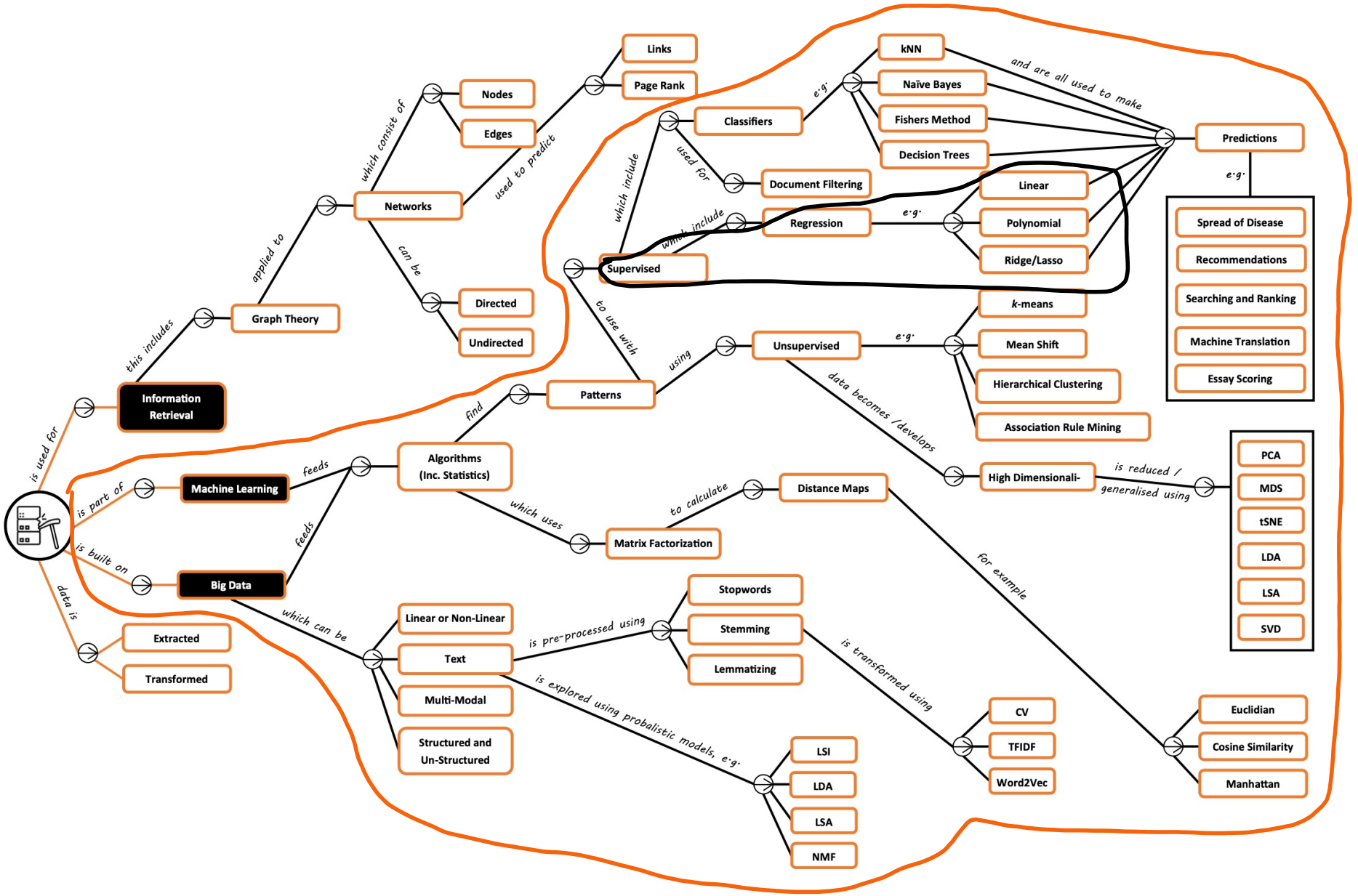
Exam Format

- MCQs + Computer-aided
  - Shoaib (40 marks) + Zhiwu (40 Marks) + Markus (20 Marks)

- My part (15 MCQs):
  - MCQs 1-12 : 2 marks each (easy), one single answer each
  - MCQs 13-15: 5 or 6 marks each (hard)
    - Each MCQ contains 2 or 3 subquestions
    - Each subquestion has one single answer

- How to prepare?
  - Understand concepts and ideas for easy/normal MCQs
    - Cover all the content from lecture slides except for appendix ones
    - Check out overview slides + learning-outcome slides for big pictures
  - Do some calculations on learned algorithms for hard MCQs

Roadmap (My Part)



Lecture 1 – Recommendation

- **LO1:** Mastering fundamental concepts and mathematical calculations of content-based and user-based/item-based collaborative filtering approaches, such as (exam)
  - Measuring distances/similarities between users or items
  - Predicting the missing rating using content-based approach
  - Calculating the predicted rating with user-based collaborative filtering approach
- **LO2**: Implement basic algorithms using Python (coursework)

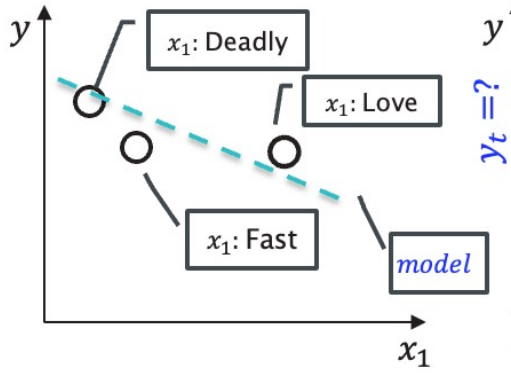
Content-based

$$\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} (\theta^T X_i - y)^2$$

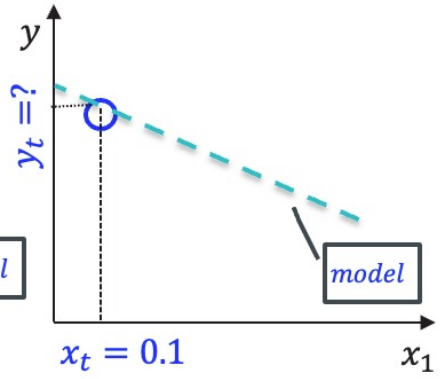
$x(\text{Love}) = (x_1, x_2) = (1, 0.1)$

Film	Dave (y)	x_1 romance	x_2 action
Love Really	4	1	0.1
Deadly Weapon	5	0.1	1
Fast and Cross	4	0.2	0.9
Star Battles	? (y_t =?)	0.1	1

1. Fitting User Behavior



2. Prediction



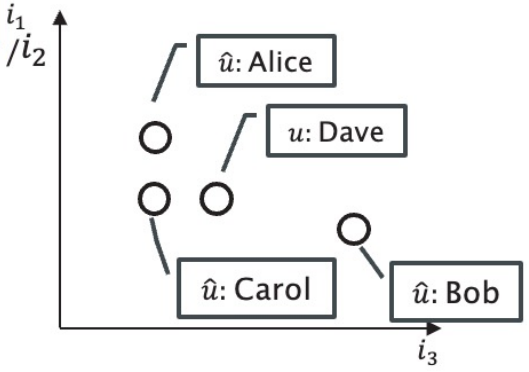
User-based Collaborative

$$r_{u,i} = \frac{\sum_{\hat{u} \in U} sim(u, \hat{u}) r_{\hat{u},i}}{\sum_{\hat{u} \in U} |sim(u, \hat{u})|}$$

$\hat{u}(\text{Alice}) = (i_1, i_3) = (4, 5)$

	Film	Alice	Bob	Carol	Dave
i_1	Love Really	4	1		4
i_2	Deadly Weapon		1	4	5
i_3	Fast and Cross	5		5	4
	Star Battles	1	5	2	? ($r_{u,t}$ =?)

1. Finding Similar Users



2. Prediction

Sim(Dave, Alice) = 0.8
Sim(Dave, Carol) = 0.6

$$r_{u,t} = \frac{0.8 \times 1 + 0.6 \times 2}{0.8 + 0.6}$$

e.g., Cosine Similarity:

$$cos(\theta) = \frac{\boldsymbol{p} \cdot \boldsymbol{q}}{|\boldsymbol{p}||\boldsymbol{q}|} = \frac{\sum_{i=1}^{N} p_i q_i}{\sqrt{\sum_{i=1}^{N} p_i^2} \sqrt{\sum_{i=1}^{N} q_i^2}}$$

Note: The similarities (0.8, 0.6) above are merely used for demonstration. With Cosine similarity, they'd be 0.99 and 0.98 if we assume the missing ratings can be replaced with the user's average on other films excluding 'Star Battles' (e.g., "Deadly Weapon" with Alice is 4.5).

Item-based?

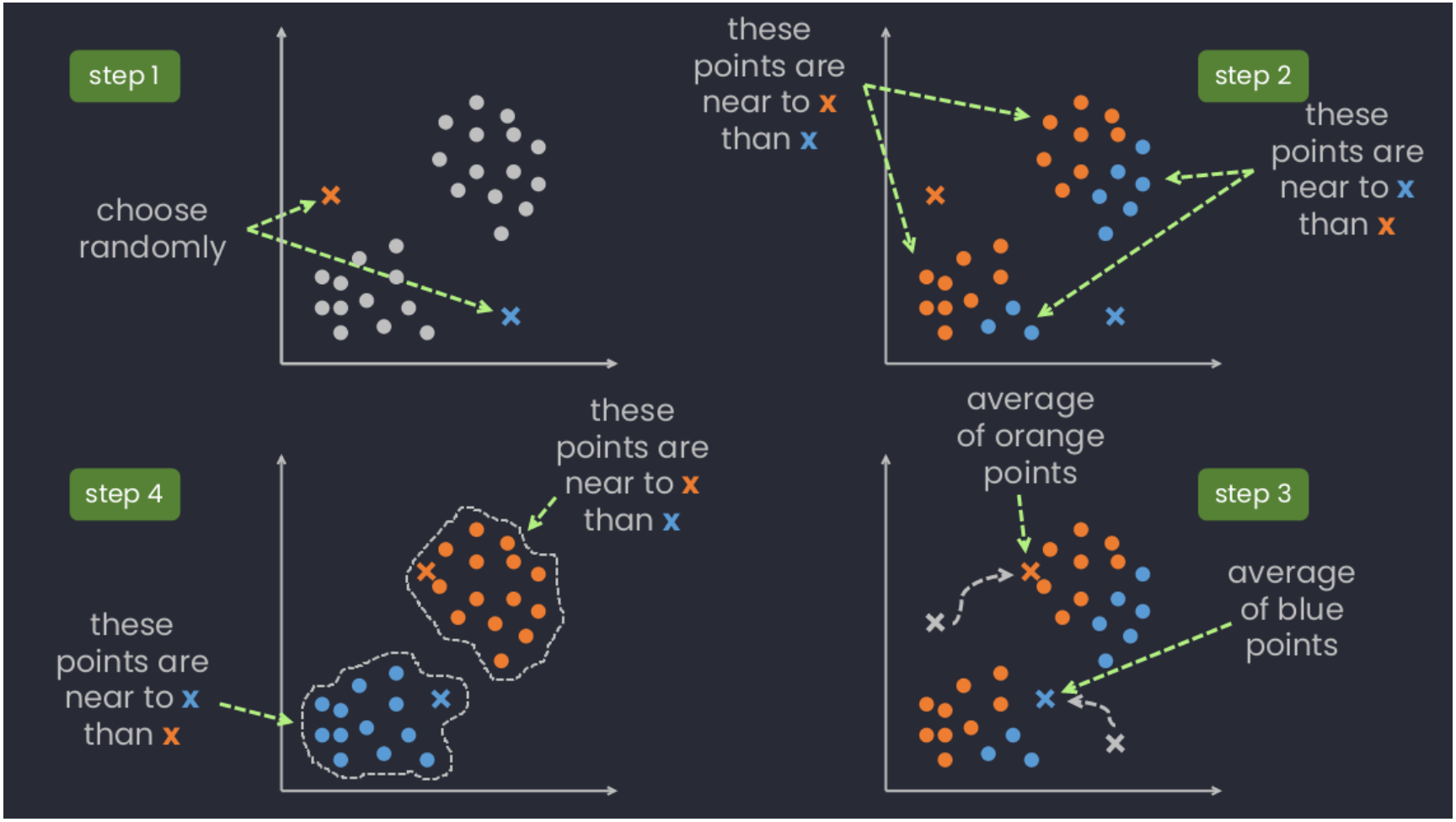
$$\text{sim}(i, j) = \frac{\sum_u r_{u,i} \cdot r_{u,j}}{\sqrt{\sum_u r_{u,i}^2} \cdot \sqrt{\sum_u r_{u,j}^2}}$$

- $r_{u,i}$ and $r_{u,j}$ are the ratings given by user u to items i and j.

Lecture 2 - Clustering

- **LO1**: Comprehend the key ideas and the essential mathematical formulations employed in clustering methods (exam).
  - E.g., how is sum of squared error (SSE) defined?
  - E.g., understand the pros and cons of the learned algorithms
- **LO2**: Compute the fundamental stages of learned clustering approaches (exam).
  - E.g., given a dataset and a distance metric, be prepared to follow the selected clustering algorithm to cluster the instances in the dataset

Algorithm – K-means



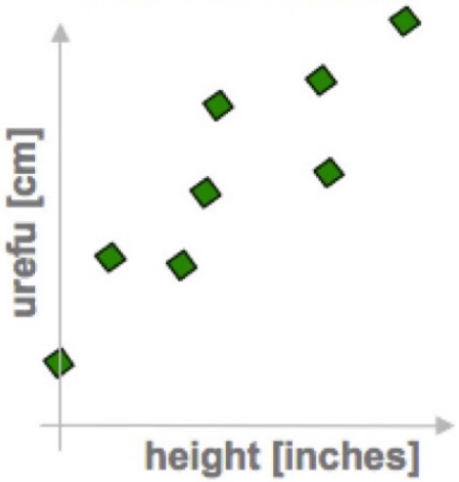
Credit: Pratik Thorat

Lecture 3 – Statistics & PCA

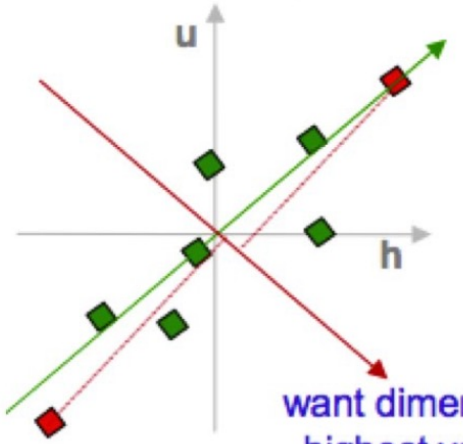
- **LO1**: Compute basic statistics such as variance and covariance over the given data (exam)
- **LO2**: Understand the key ideas and essential steps of PCA using EVD, SVD and truncated SVD (exam).

PCA in a nutshell

1. correlated hi-d data
("urefu" means "height" in Swahili)



2. center the points



want dimension of highest variance

3. compute covariance matrix

$$\begin{array}{c} \\ h \\ u \end{array}\begin{array}{c} \begin{array}{cc} h & u \end{array} \\ \begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \end{array} \rightarrow \mathrm{cov}(h,u) = \frac{1}{n}\sum_{i=1}^{n} h_i u_i$$

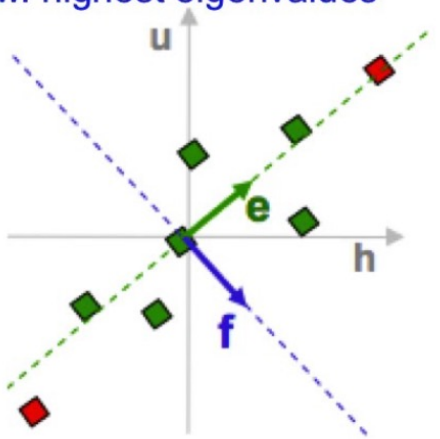
4. eigenvectors + eigenvalues

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} e_h \\ e_u \end{pmatrix} = \lambda_e \begin{pmatrix} e_h \\ e_u \end{pmatrix}$$

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} f_h \\ f_u \end{pmatrix} = \lambda_f \begin{pmatrix} f_h \\ f_u \end{pmatrix}$$

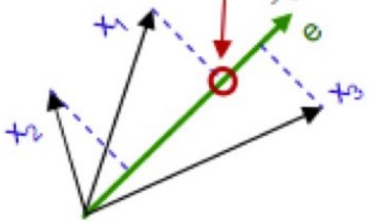
```
eig(cov(data))
```

5. pick m<d eigenvectors
w. highest eigenvalues

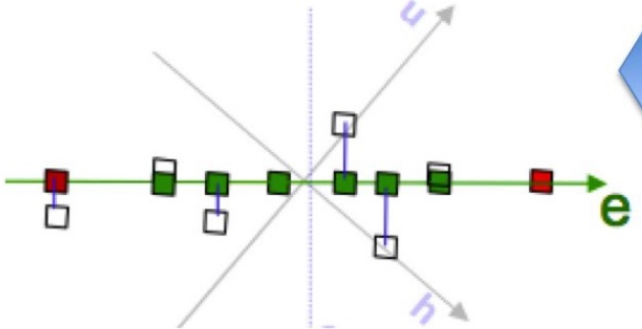


6. project data points to those eigenvectors

$$x_e' = x^T e = \sum_{j=1}^{d} x_{ij} e_j$$



7. uncorrelated low-d data



Source: https://biddypuffdiddy.wordpress.com

- **LO1**: understand the basic idea of all the learned dimensionality reduction methods (exam)
  - E.g., describe the basic idea of the suggested algorithm
  - E.g., understand the pros and cons of the learned algorithms
  - E.g., given a dataset, be prepared to follow the selected algorithm to calculate its key steps on the given data

Lecture 4- Embedding Data

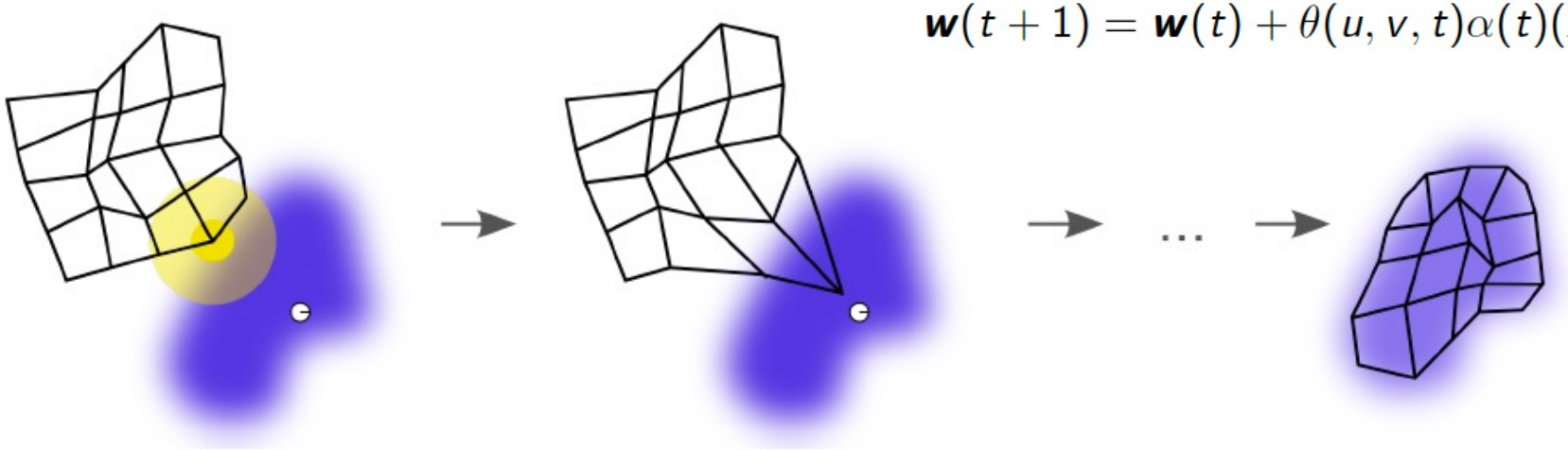
SOMs: two phases

- training
- mapping

To start training, the set of nodes each has a random starting position defined in the feature space.

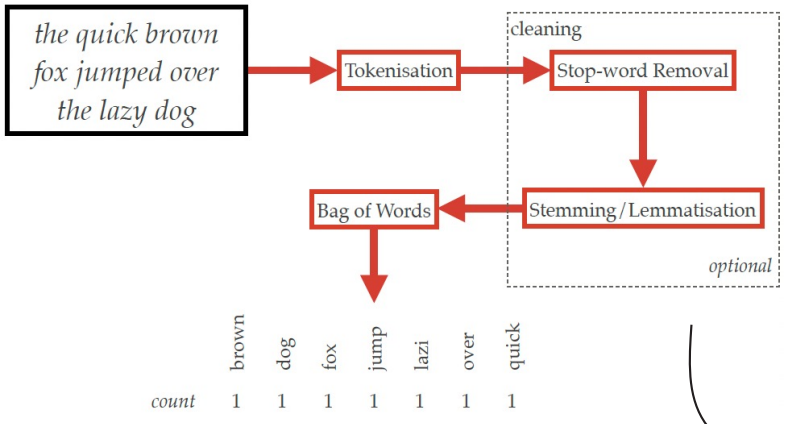
This is then updated by taking one feature vector, finding which unit is the *best matching unit* (BMU) then moving that unit and, to a lesser extent, its neighbours, closer to that data point

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \theta(u, v, t)\alpha(t)(x_i - \boldsymbol{w}(t))$$

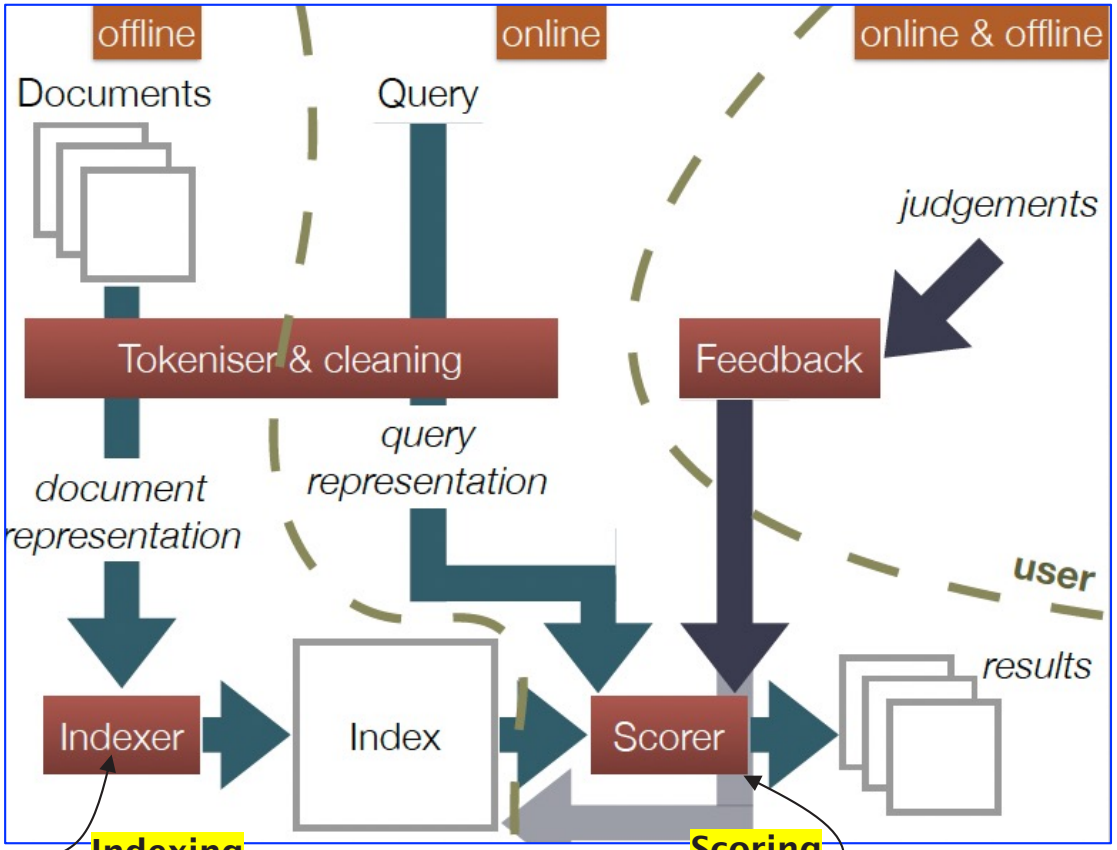


Lecture 5 – Search & Rank

- **LO1**: Demonstrate an understanding of the fundamental concepts and approaches for search and ranking, such as: (exam)
  - Understanding the basic pipeline of searching and raking
  - Encoding document/query using the vector space modeling methods
  - Mastering the indexing and matching methods for search

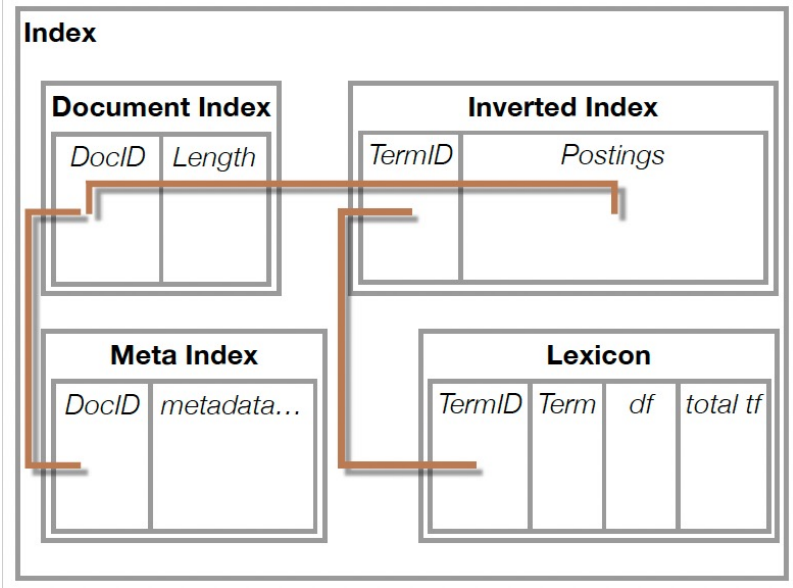


Encoding
(vector space modeling)



Indexing
(sorting)

Scoring
(matching)



Number of times *w* appears in *d*

Total #docs in collection

$$f(\mathbf{q},\mathbf{d}) = \sum_{i=1}^{N} q_i y_i = \sum_{w \in q \cap d} c(w,q)c(w,d)log\frac{M+1}{df(w)}$$

Number of times *w* appears in *q*

document frequency (number of docs containing *w*)

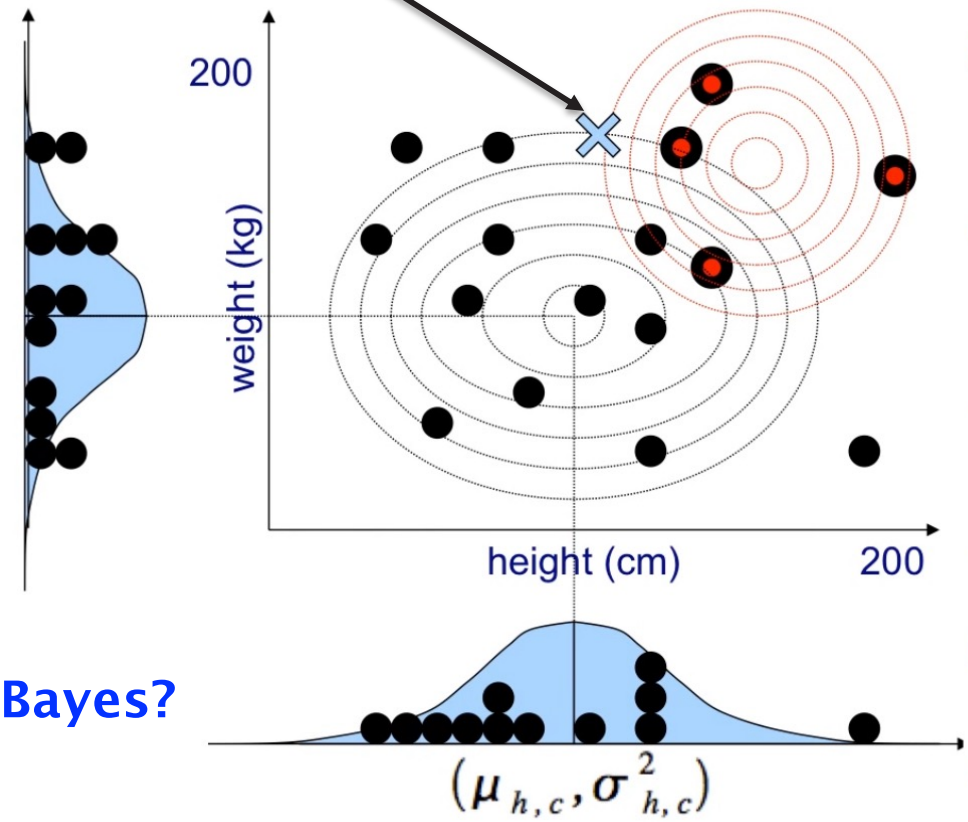
- **LO1**: Demonstrate an understanding of the fundamental concepts and approaches for document filtering, such as: (exam)
  - Understanding the basic ideas behind Naive Bayes and Fisher's methods
  - Applying Naïve Bayes and Fisher's Methods to classify documents as spam or not spam
  - Discussing the pros and cons of using Naive Bayes vs. Fisher's methods for document filtering

Lecture 6 - Doc Filtering

1. Priors

$$P(a)=\frac{4}{4+12}=0.25\ ;\ P(c)=0.75$$

2. Likelihoods



$$p(h_x|c)=\frac{1}{\sqrt{2\pi\,\sigma^2_{h,c}}}\exp-\frac{1}{2}\left(\frac{(h_x-\mu_{h,c})^2}{\sigma^2_{h,c}}\right)$$

$$p(w_x|c)=\frac{1}{\sqrt{2\pi\,\sigma^2_{w,c}}}\exp-\frac{1}{2}\left(\frac{(w_x-\mu_{w,c})^2}{\sigma^2_{w,c}}\right)$$

$$p(h_x|a)=\frac{1}{\sqrt{2\pi\,\sigma^2_{h,a}}}\exp-\frac{1}{2}\left(\frac{(h_x-\mu_{h,a})^2}{\sigma^2_{h,a}}\right)$$

$$p(w_x|a)=\frac{1}{\sqrt{2\pi\,\sigma^2_{w,a}}}\exp-\frac{1}{2}\left(\frac{(w_x-\mu_{w,a})^2}{\sigma^2_{w,a}}\right)$$

Feature independence

$$P(x|a)=p(h_x|a)\,p(w_x|a)$$

$$P(x|c)=p(h_x|c)\,p(w_x|c)$$

Naïve Bayes?

3. Posteriors

$$P(a|x)=\frac{P(x|a)P(a)}{P(x|a)P(a)+P(x|c)P(c)}$$

$$p(c|x)=\frac{p(x|c)p(c)}{p(x|a)p(a)+p(x|c)p(c)}$$

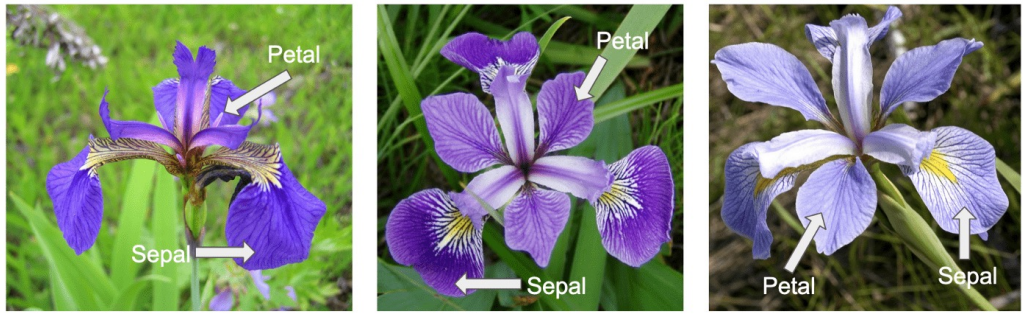
Fisher method? $P(a|x)=P(a|h_x)P(a|h_y)$?

Credit: Victor Lavrenko

Lecture 7- Decision Trees

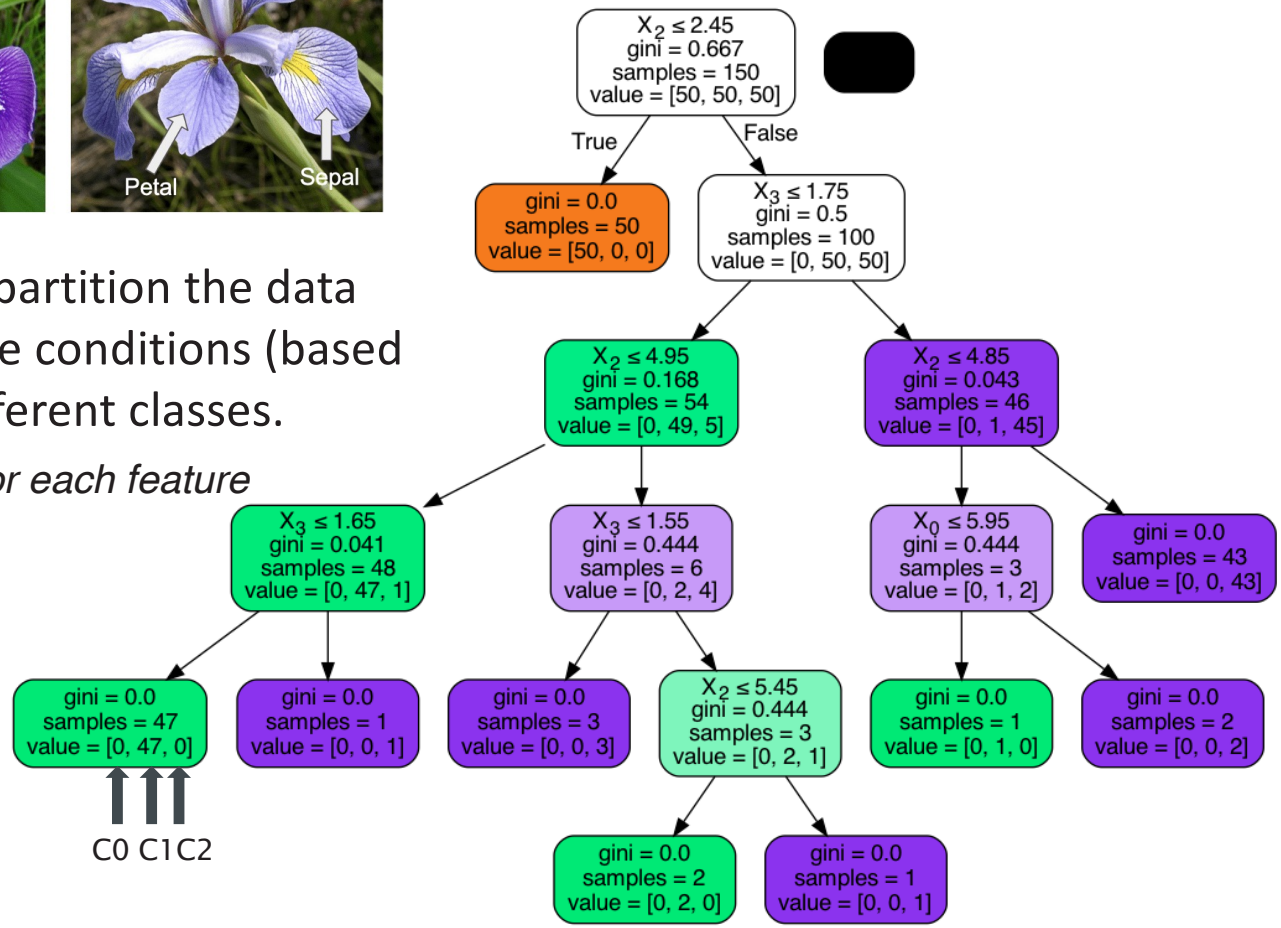
- **LO1**: Demonstrate an understanding of decision tree fundamentals, such as (exam)
  - Calculating impurity and constructing a decision tree using algorithms like ID3, C4.5, etc, given a dataset and distance metric
  - Addressing overfitting in decision trees like pruning
  - Understanding ensemble methods using decision trees like bagging, boosting, random forests

Iris setosa *Iris versicolor* *Iris virginica*



The objective is to efficiently partition the data based on the most informative conditions (based on features) that separate different classes.

***Key idea**: Greedily find best split for each feature*



C0: 5
C1: 5

Non-homogeneous,

High degree of impurity

C0: 9
C1: 1

C0 C1C2

Homogeneous,

Low degree of impurity

Four features, only three are used here, and one is only used once.

Lecture 8 - KNN

- **LO1**: Demonstrate an understanding of the fundamentals of nearest neighbor classifiers, such as: (exam)
  - Understanding the k-NN algorithm, including key steps like distance calculations, voting, etc
  - Applying weighted k-NN and its use of weighted distance calculations
  - Discussing advantages and disadvantages of k-NN models

Assign class of unknown point based on **majority class** of *closest* ***K*** **neighbours** in feature space

Lecture 9- Market Basket

- **LO1**: Demonstrate an understanding of market basket analysis concepts and techniques, such as: (exam)
  - ❖ Calculating support and confidence for itemsets
  - ❖ Understanding the key steps of the Apriori algorithm for association rule mining
  - ❖ Using the Apriori algorithm to generate association rules from transaction data

EXAMPLE OF ASSOCIATION RULES

Assume there are 5 customers

3 of them bought milk, 2 bought potato chip and 2 bought both of them

Transaction 1:	Frozen pizza, cola, milk
Transaction 2:	Milk, potato chips
Transaction 3:	Cola, frozen pizza
Transaction 4:	Milk, potato chips
Transaction 5:	Cola, pretzels

milk → potato chip

How about Potato chip → Milk ?

support milk = P(milk) = 3/5 = 0.6

support potato chip = P(potato chip) = 2/5 = 0.4

support = P(milk & potato chip) = 2/5 = 0.4

confidence

= support (milk & potato chip)/support(milk)

= 0.4/0.6

= 0.67

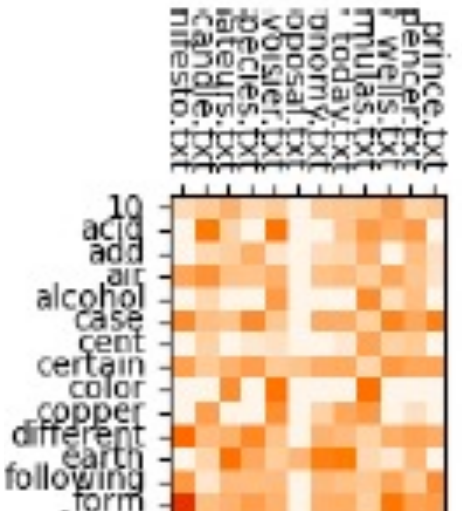
CONFIDENCE= P(Milk & potato chip) /P(Milk)

https://www.analyticsvidhya.com

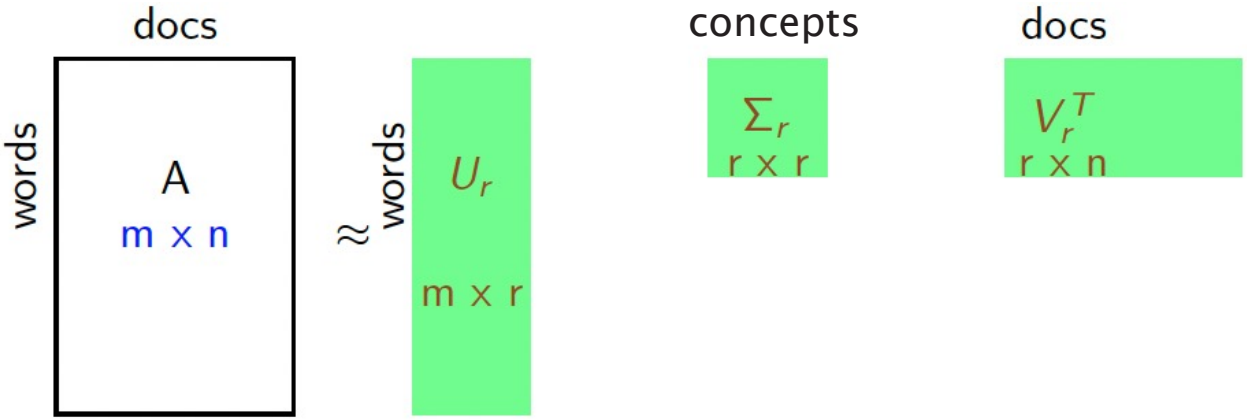
Lecture 10 - LSA

- **LO1**: Demonstrate an understanding of techniques for finding independent semantic features, such as: (exam)
  - ❖ Comprehending the core concepts of Latent Semantic Analysis (LSA) and apply LSA on a dataset
  - ❖ Discussing the advantages and disadvantages of algorithm

Latent Semantic Analysis (LSA) using Bag of Words (BoW) & truncated SVD



Bag of Words (BoW)



Each row of V_r corresponds to an eigenvector of $A^T A$

- This means it is proportional to the covariance or correlation between the documents
- These are the *concepts*

Features

Each row of U_r describes a term as a vector of weights with respect to r *concepts*

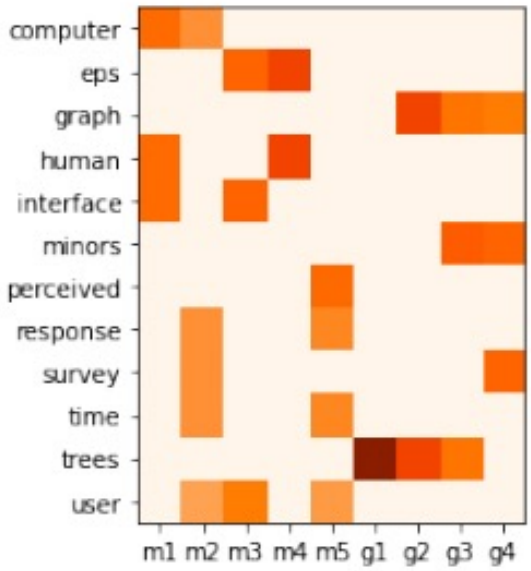
Each column of V_r describes a document as a vector of weights with respect to r concepts

Lecture 11- NMF & LDA

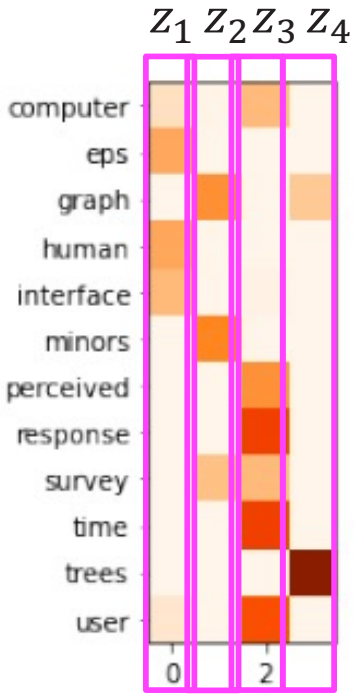
- **LO1**: Demonstrate an understanding of techniques for finding independent features for topic modeling, such as: (exam)
  - ❖ Comprehending the core concepts of NMF and apply NMF on a dataset
  - ❖ Understanding the key idea and steps of probabilistic models like PLSA and LDA
  - ❖ Discussing the advantages and disadvantages of the learned algorithms

Non-negative Matrix Factorisation (NMF)

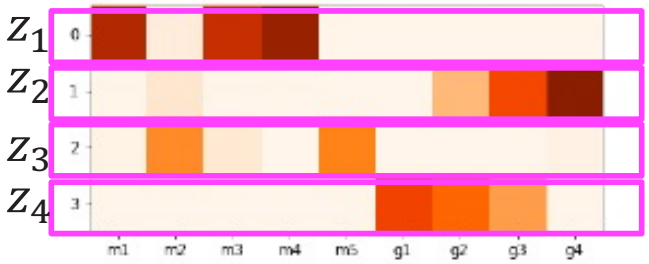
z_i: i-th topic



A



W

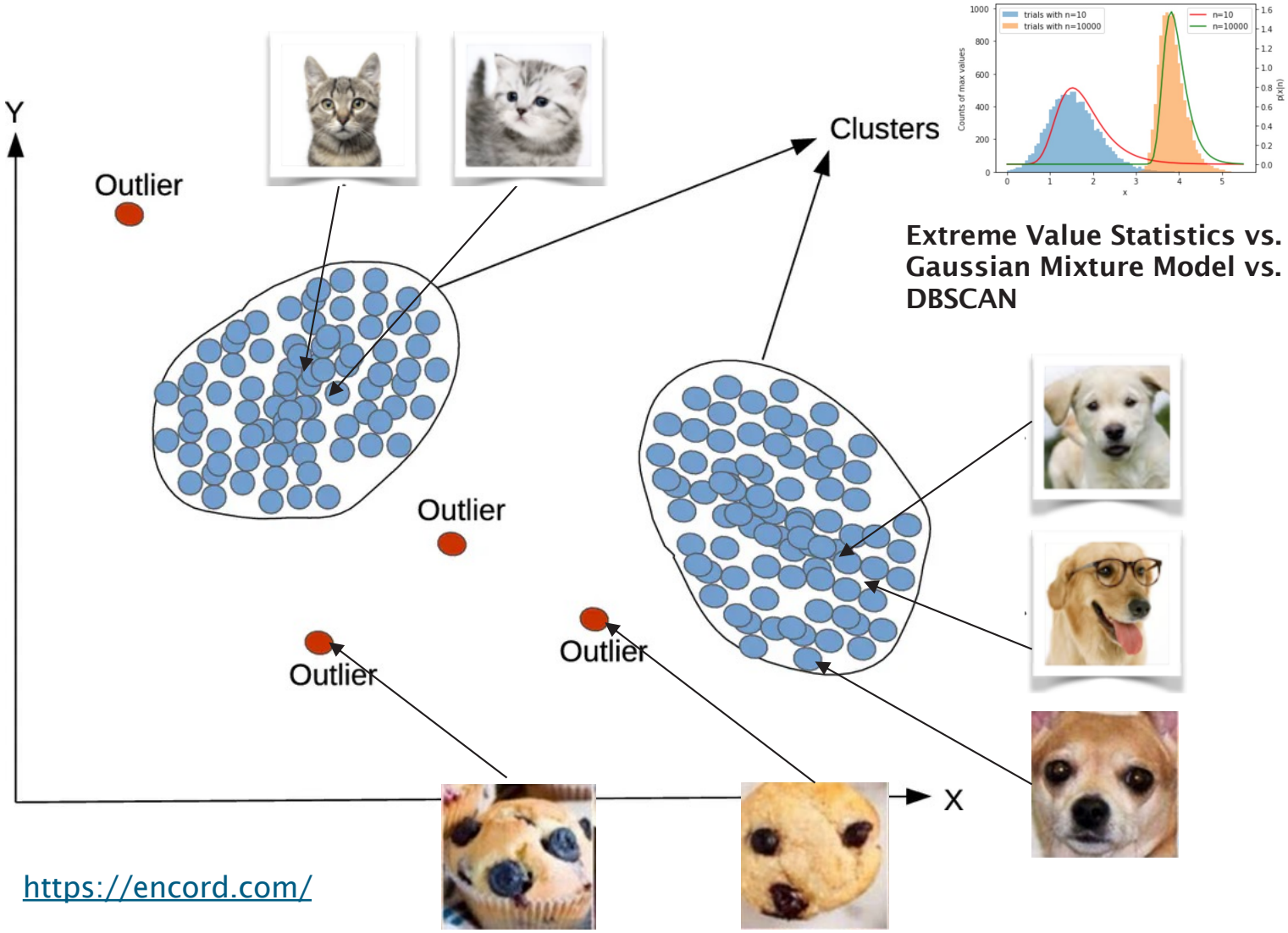


H

W has the *basis vectors*, showing how the words are clustered
H has the topic memberships for the documents.

Lecture 12 - Outlier Det

- **LO1**: Demonstrate an understanding of techniques for outlier detection, such as: (exam)
  - ❖ Applying extreme value analysis method
  - ❖ Understanding the key idea and steps of the learned clustering methods for outlier detection
  - ❖ Discussing the advantages and disadvantages of the learned outlier detection approaches



Extreme Value Statistics vs. Gaussian Mixture Model vs. DBSCAN

https://encord.com/

Revision Techniques and Exam Prep

Planets in the solar system

Mnemonic means that you use a memorable phrase to help you with memorising facts

My very eager mother just served us nachos

Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune