Resit exam questions (2020-07-16)

Resit exam (July 2020)

- This was an oral resit exam
- Students had to answer ~20 questions in ~20-25 minutes
- Questions were chosen at random by throwing a dice, advancing the number shown on the dice, and asking that question, then throwing the dice again, and so on ...

TT02 Data preparation

TT02. Data preparation

How do you convert a numerical variable into a categorical variable?

TT02. Data preparation

What is the difference between equi-width and equi-depth binning?

TT02. Data preparation

Suppose the salaries in a company are
20k, 40k, 50k, 100k, 120k, 140k

Divide into three equi-width bins.

Divide into three equi-depth bins.

TT02. Data preparation

What does it mean to do **schema integration**?

TT02. Data preparation

After a dataset has passed a **data cleaning** process, what do we know about this dataset?

TT02. Data preparation

What are the two options that we have if in a record one or more values are missing?

TT02. Data preparation

Suppose in a database for traffic fines a record is missing the **model** of the car. What should we do with that record?

TT02. Data preparation

Suppose in a database for traffic fines a record is missing the **plate** of the car. What should we do with that record?

TT02. Data preparation

How do we obtain the **standardized** value for a variable?

TT02. Data preparation

We have a variable taking values {1, 2, 3, 4, 5}
μ=3.0, σ=1.41

Normalize by using standardization

TT02. Data preparation

How do we obtain the **min-max scaled** value for a variable?

TT02. Data preparation

We have a variable taking values {1, 2, 3, 4, 5}

Normalize by using min-max scaling

TT03 Similarity

TT03. Similarity

What is the similarity of an object to itself if the similarity is in a scale from 0.0 to 1.0?

What is the distance of an object to itself if the distance is in a scale from 0.0 to 1.0?

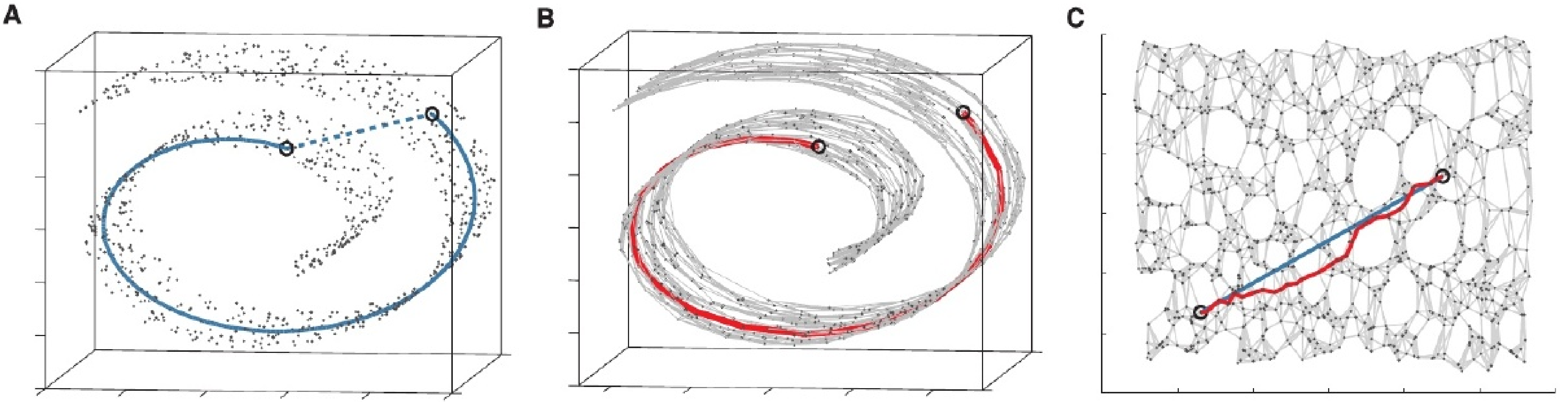
TT03. Similarity

What is the formula for the L_2 norm?

What is the formula for L_p norm?

TT03. Similarity

Explain how ISOMAP works



TT03. Similarity

Compute the **Jaccard similarity** between these two sets:

{orange, car, shoe}

{apple, car, shoe}

TT03. Similarity

Compute the **Jaccard distance** between these two vectors:

[0, 1, 0, 0, 1, 1]

[0, 1, 0, 0, 0, 0]

TT04 Near duplicates

TT04. Near duplicates

Suppose you have a dataset of N exams by students

What would be a naïve, brute force approach to detect if any of those students copied from another? How slow would be that method?

TT04. Near duplicates

How many **different** 2-word-gram shingles are contained in the phrase “to be or not to be”?

TT04. Near duplicates

Permutation π

Rows=Shingles, Columns=Documents

	D1	D2	D3	D4
2	1	0	0	0
3	1	0	0	1
1	0	1	0	1
6	0	1	0	1
4	0	0	0	1
5	1	0	1	0

Compute the signature vector under π

D1	D2	D3	D4

TT04. Near duplicates

What is the similarity between each pair of documents, if this is the signature matrix?

	D1	D2	D3	D4
π_1	1	1	4	5
π_2	3	3	3	2
π_3	2	5	2	2

TT05 Itemsets

TT05. Itemsets

What is a transaction?

What is an itemset?

TT05. Itemsets

What is the support of an itemset?

TT05. Itemsets

Indicate the support of an itemset here:

tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Highlighter, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT05. Itemsets

What is a frequent itemset?

TT05. Itemsets

Indicate frequent itemsets if minsup=0.4

tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Highlighter, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT05. Itemsets

Indicate why the monotonicity property holds:

$$J \subseteq I \Rightarrow sup(J) \geq sup(I)$$

TT05. Itemsets

What is a closed itemset?

TT05. Itemsets

What is a closed itemset in this database?

What is a non closed itemset in this database?

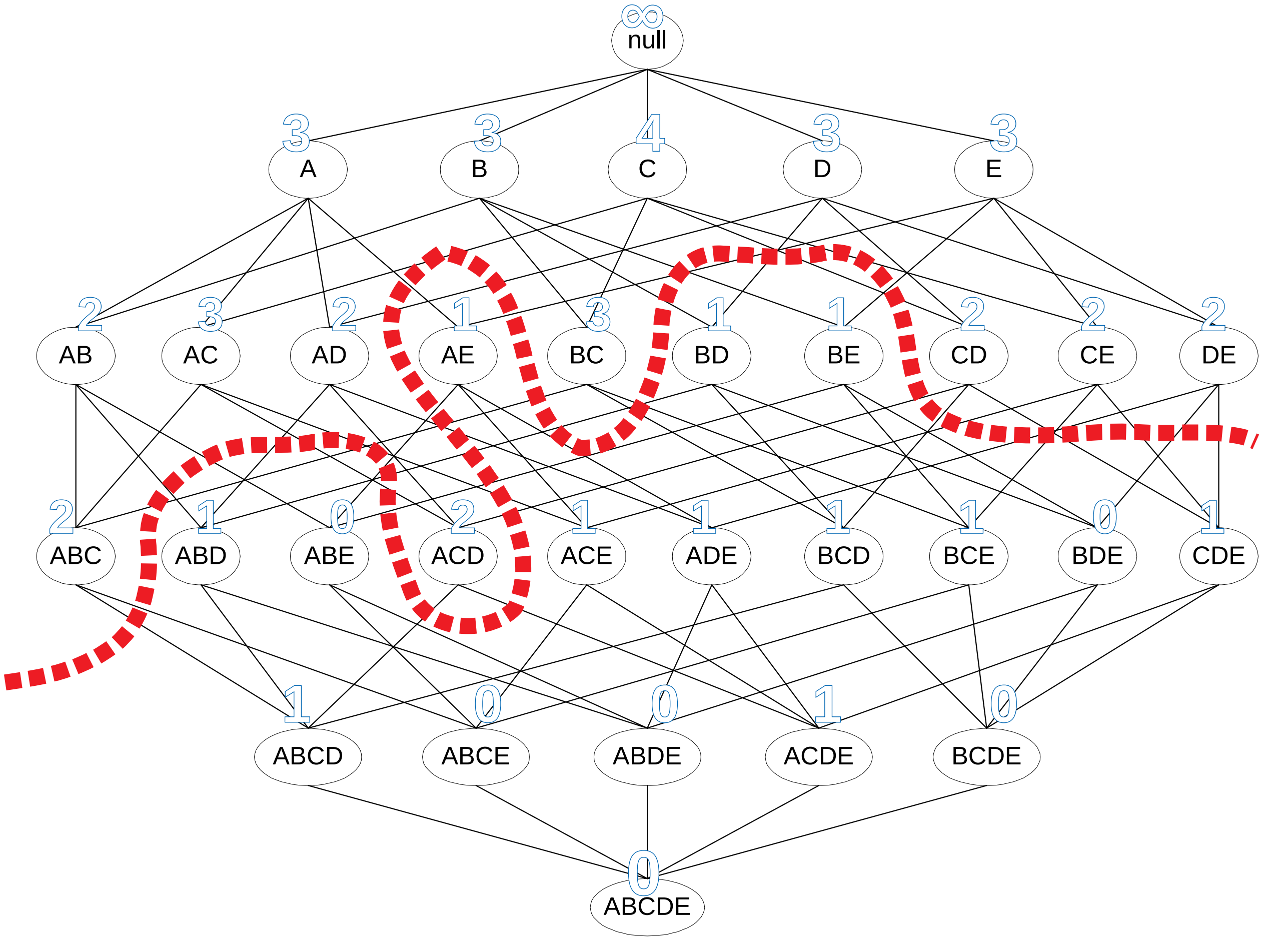
tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Highlighter, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT05. Itemsets

Numbers indicate itemset frequencies

Indicate what is the red line

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



TT05. Itemsets

What is the confidence on a rule?

What is the formula of the confidence of a rule?

TT05. Itemsets

Indicate the confidence of the rule
{Pencil} => {Eraser}

tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Highlighter, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT06 Association rule mining

TT06. Association rule mining

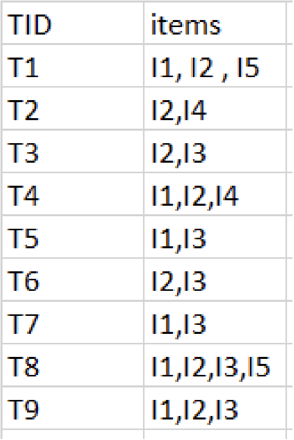
Explain the apriori algorithm on this dataset, with minsup=2 (minsup=0.4).

Tip: first write a table with itemsets of size 1 (itemset, support)

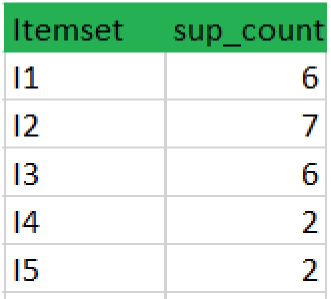
tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Highlighter, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT06. Association rule mining

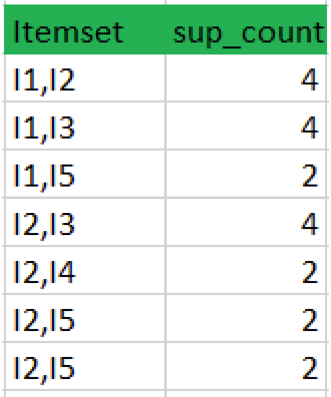
Obtain one rule of the form {a,b} ⇒ {c} that has confidence 50% from these itemsets:



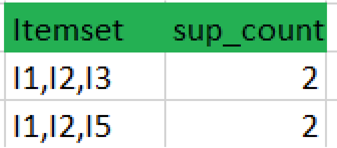
TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3



Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2



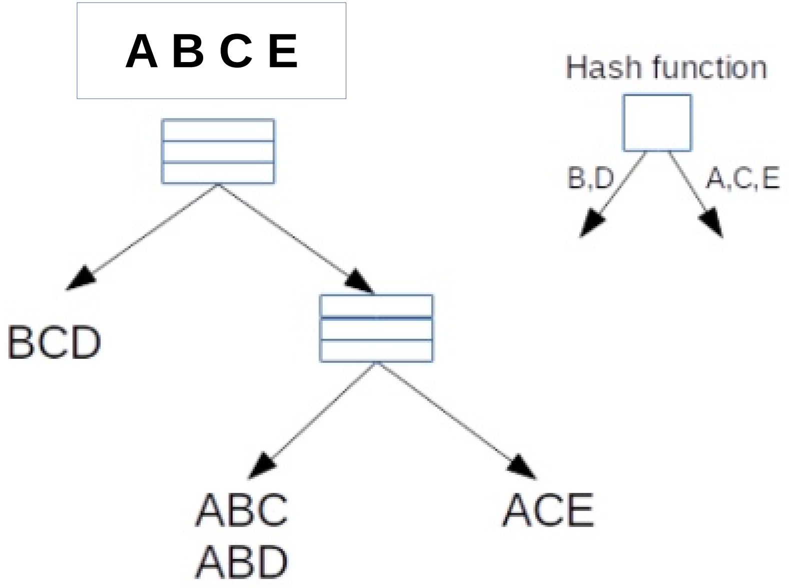
Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I2,I5	2



Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

TT06. Association rule mining

Indicate in this hash tree which candidates are visited if we are looking for itemsets contained in
{ A, B, C, E}



TT08 Recommender systems

TT08. Recommender systems

What is a recommender system?

TT08. Recommender systems

What is the cold-start problem in recommender systems?

TT08. Recommender systems

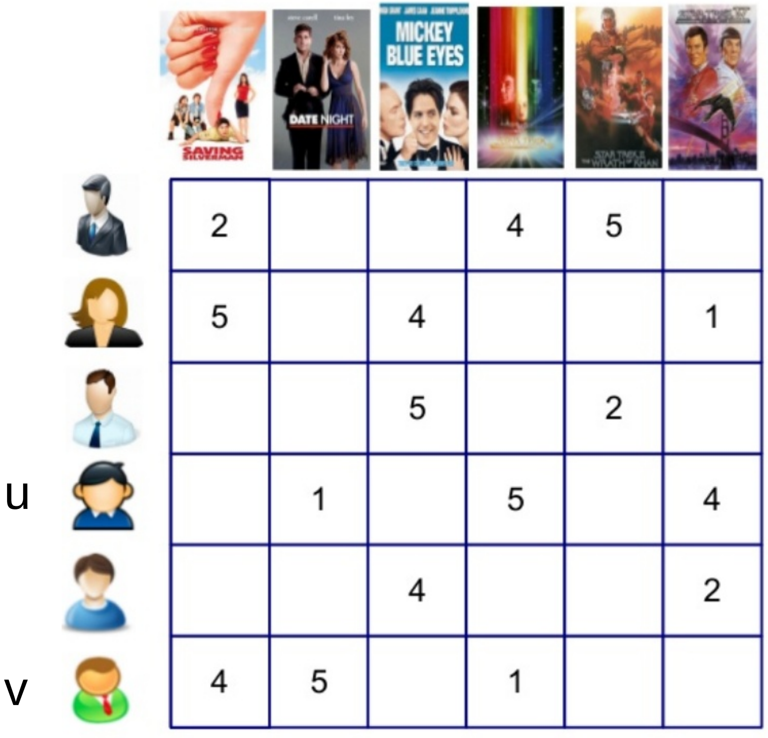
What is an utility matrix in recommender systems?

In real recommender systems, is the utility matrix completely known or partially known? Why?

TT08. Recommender systems

Compute the similarity between users u and v in this dataset

$$\mathrm{sim}(u,v) = \frac{\sum_{i \in I_{u,v}} (u_i - \hat{u}) \cdot (v_i - \hat{v})}{\sqrt{\sum_{i \in I_{u,v}} (u_i - \hat{u})^2 \cdot \sum_{i \in I_{u,v}} (v_i - \hat{v})^2}}$$


SAVING SILVERMAN
DATE NIGHT
MICKEY BLUE EYES
STAR TREK II THE WRATH OF KHAN

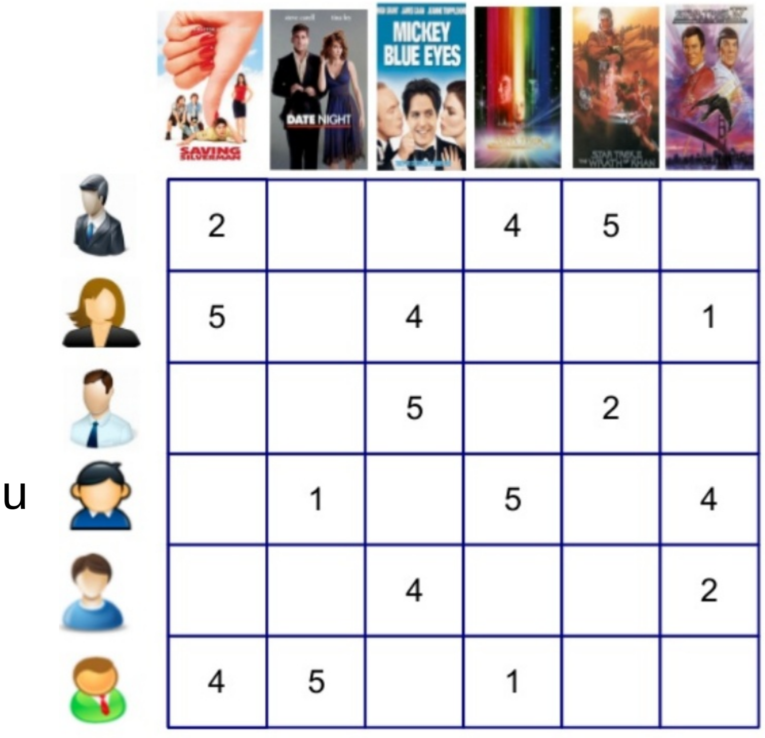
	2			4	5	
	5		4			1
			5		2	
u		1		5		4
			4			2
v	4	5		1		

TT08. Recommender systems

Suppose you have computed all similarities of users to u.

Explain how do you recommend movies to user u using the formula below

$$\text{score}(u,i) = \hat{u} + \frac{\sum_{v: v_i \neq \text{NULL}} \text{sim}(v,u) \cdot (v_i - \hat{v})}{\sum_{v: I_{u,v} \neq \emptyset} |\,\text{sim}(v,u)|}$$



	2			4	5	
	5		4			1
			5		2	
u		1		5		4
			4			2
	4	5		1		

TT09 Outlier detection

TT09. Outlier detection

What is an outlier?

TT09. Outlier detection

How do you find outliers using extreme value analysis?

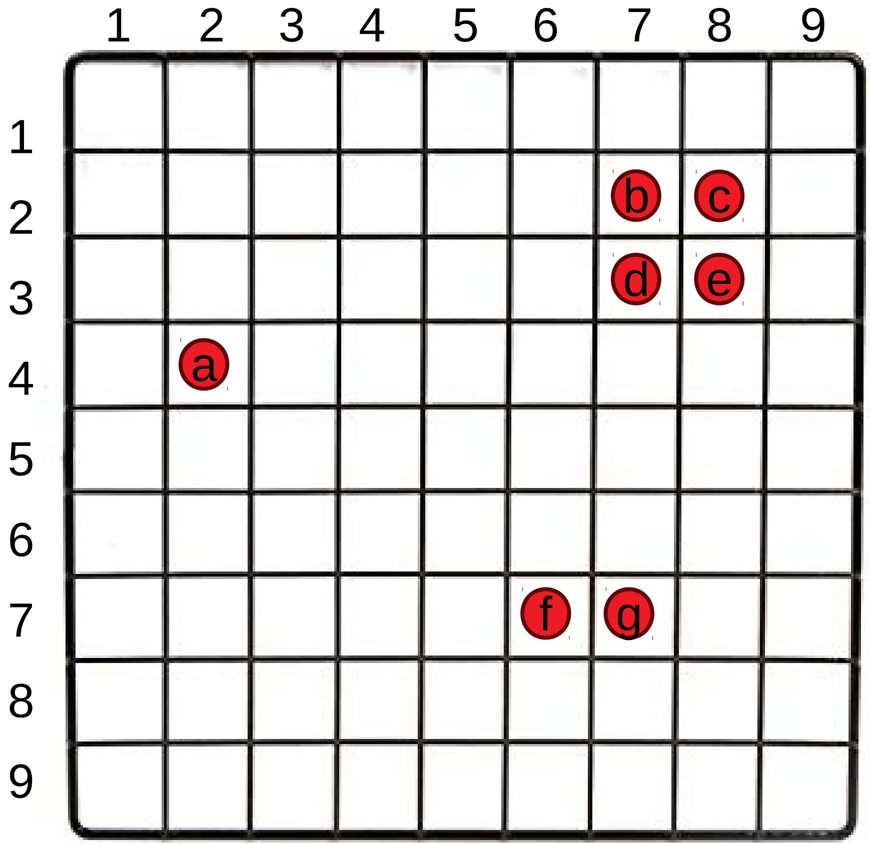
TT09. Outlier detection

Describe one situation in which extreme value analysis is inappropriate for finding outliers

TT09. Outlier detection

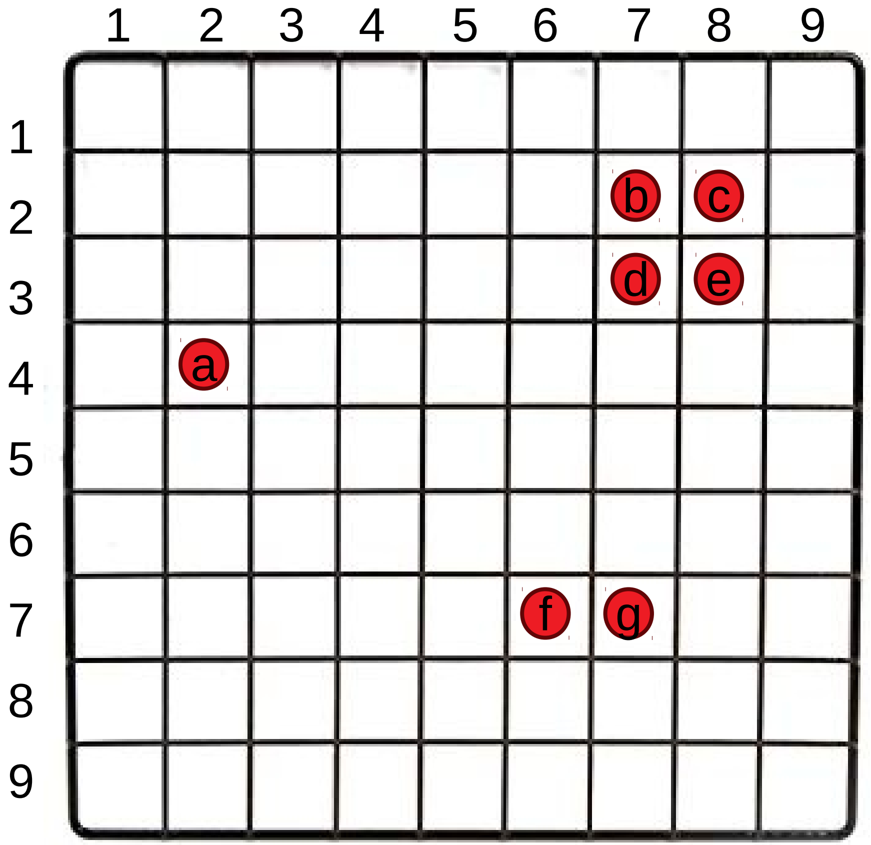
Indicate how do you create an isolation forest over the graph on the right

Explain what the outlier score for a point depends on (no need to give a formula)



TT09. Outlier detection

Indicate how a grid-based method would work to find outliers in this dataset



	1	2	3	4	5	6	7	8	9
1									
2							b	c	
3							d	e	
4		a							
5									
6									
7						f	g		
8									
9									

TT10 Streams

TT10. Streams

What is a data stream?

TT10. Streams

Suppose we have a stream of the type (u, v) indicating that user u watched video v

Indicate how to sample 1% of the users and the videos they have watched from this stream

TT10. Streams

Suppose we have a stream of photos from a photo sharing site

Indicate how to sample 100 photos from this stream **uniformly at random**

TT10. Streams

Explain how a Bloom filter works

TT10. Streams

Imagine you have an abacus of only one line, and 6 discs on that line

Indicate how to count to one million with this abacus using probabilistic counting

Indicate what is the maximum error you could make

TT10. Streams

Imagine you have an abacus of only one line, and 6 discs on that line

Indicate how to count to one million with this abacus using probabilistic counting

Indicate what is the maximum error you could make

TT11 Time series

TT11. Time series

In a time series:

What is a contextual attribute?

What is a behavioral attribute?

TT11. Time series

Interpolate the following time series using **linear interpolation** to obtain the values on Monday at midnight and Tuesday at midnight

Monday 12:00 – 33^{o}C

Tuesday 00:00 – ???

Tuesday 06:00 – 30^{o}C

Wednesday 00:00 – ???

Wednesday 18:00 – 36^{o}C

TT11. Time series

Compute a moving average with k=2 in the following series:

t	1	2	3	4	5	6	7	8	9	10
y_t	3	9	5	3	2	-4	0	12	4	6
y_t^{MA2}										

TT11. Time series

Explain how dynamic time warping works and indicate what is it can be used for