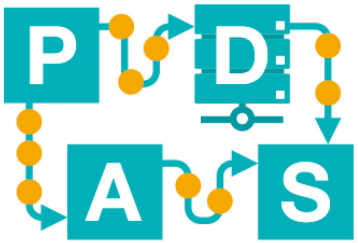
Association Rules

Lecture 12

IDS-L12



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Outline of Today's Lecture

- Recap: FP-growth
- Association rules
- Example: Iris data set
- Evaluating association rules
- Simpson's paradox

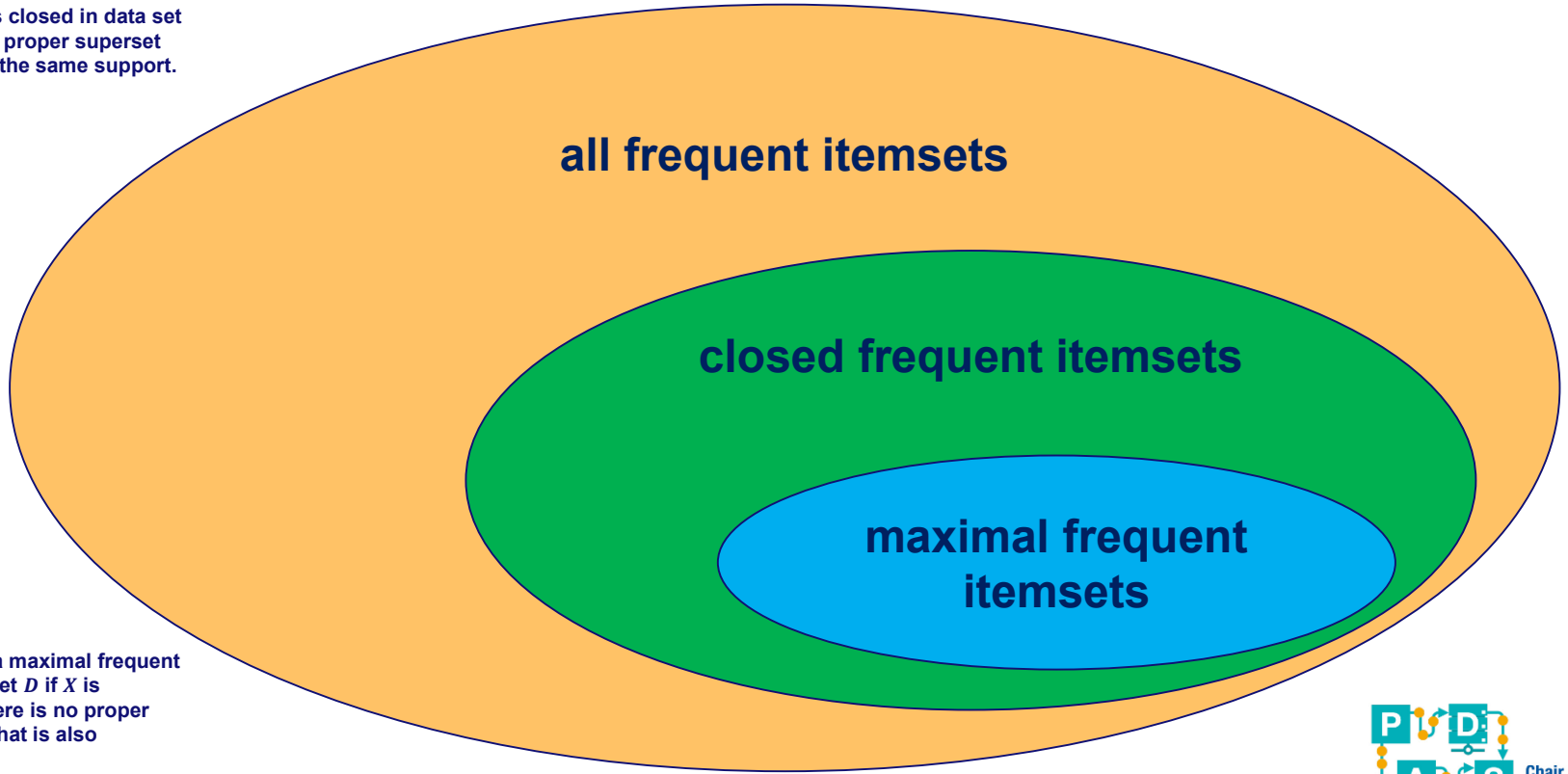
Recap: FP-growth

Frequent itemsets

- $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ is the set of all items.
- $D \in \mathfrak{B}\left(\mathcal{P}(\mathcal{I})\right)$ (such that $\emptyset \notin D$) is the multiset of transactions.
- $\{X \subseteq \mathcal{I} | support(X) \geq min\_sup\}$ are the frequent item sets.
- Example:
  - $D = [\{\text{milk}\}, \{\text{milk}\}, \{\text{bread}\}, \{\text{milk, bread}\}, \{\text{rice, bread}\}]$ is a dataset with 5 transactions.
  - $support(\{\text{milk}\}) = 3$, $support(\{\text{milk,bread}\}) = 1$, etc.

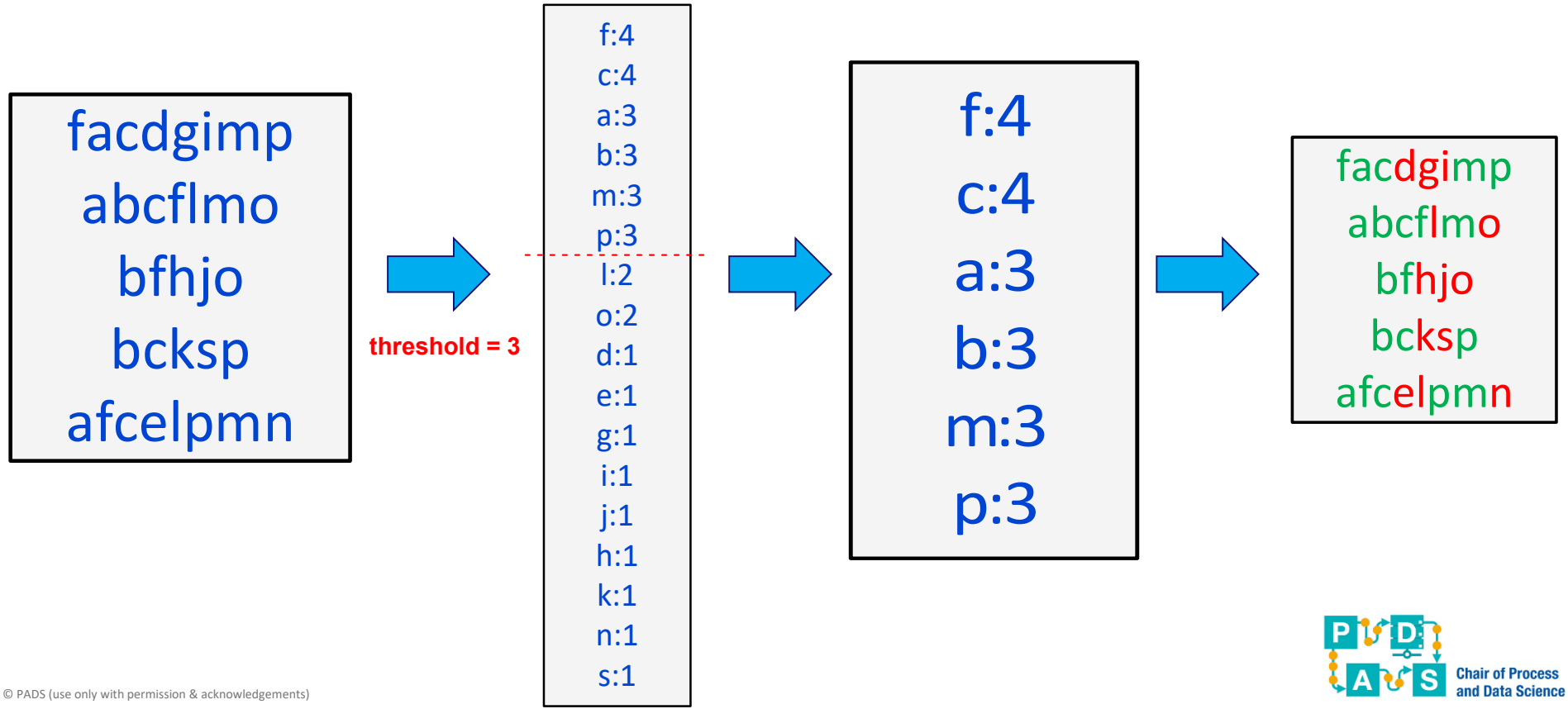
Relationships

An itemset X is closed in data set D if there is no proper superset $Y \supset X$ that has the same support.



An itemset X is a maximal frequent itemset in data set D if X is frequent, and there is no proper superset $Y \supset X$ that is also frequent.

FP-growth example



Filter and reorder based on frequency

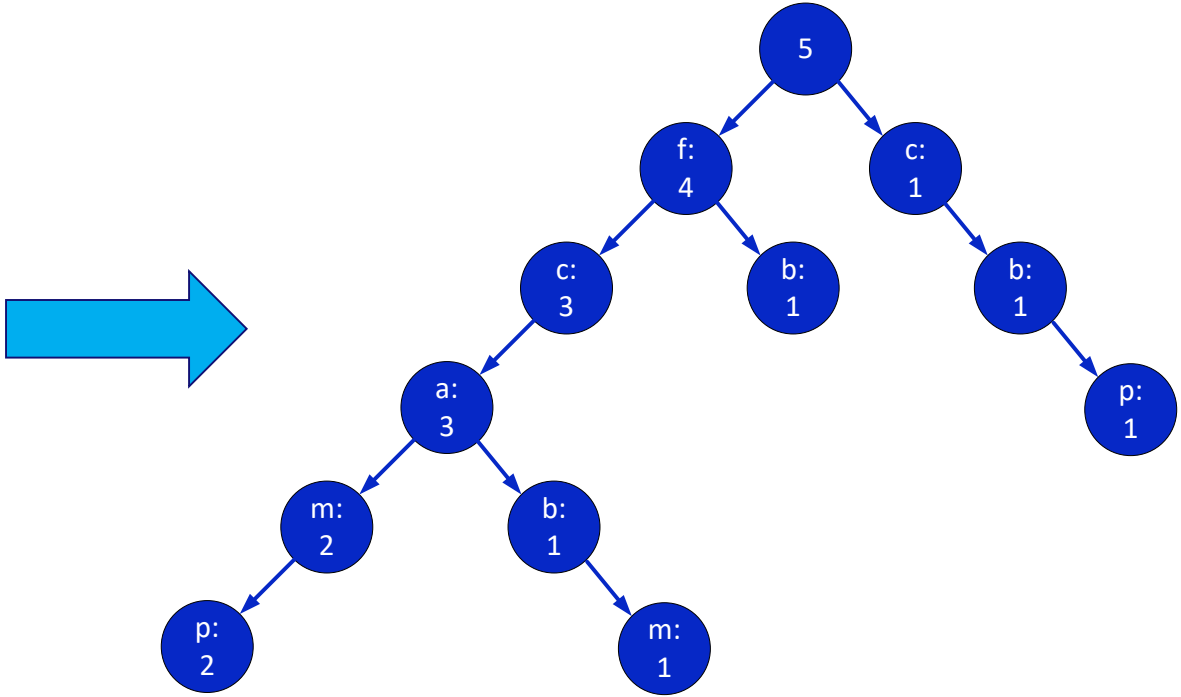
facdgimp
abcflmo
bfhjo
bcksp
afcelpmn

f:4
c:4
a:3
b:3
m:3
p:3

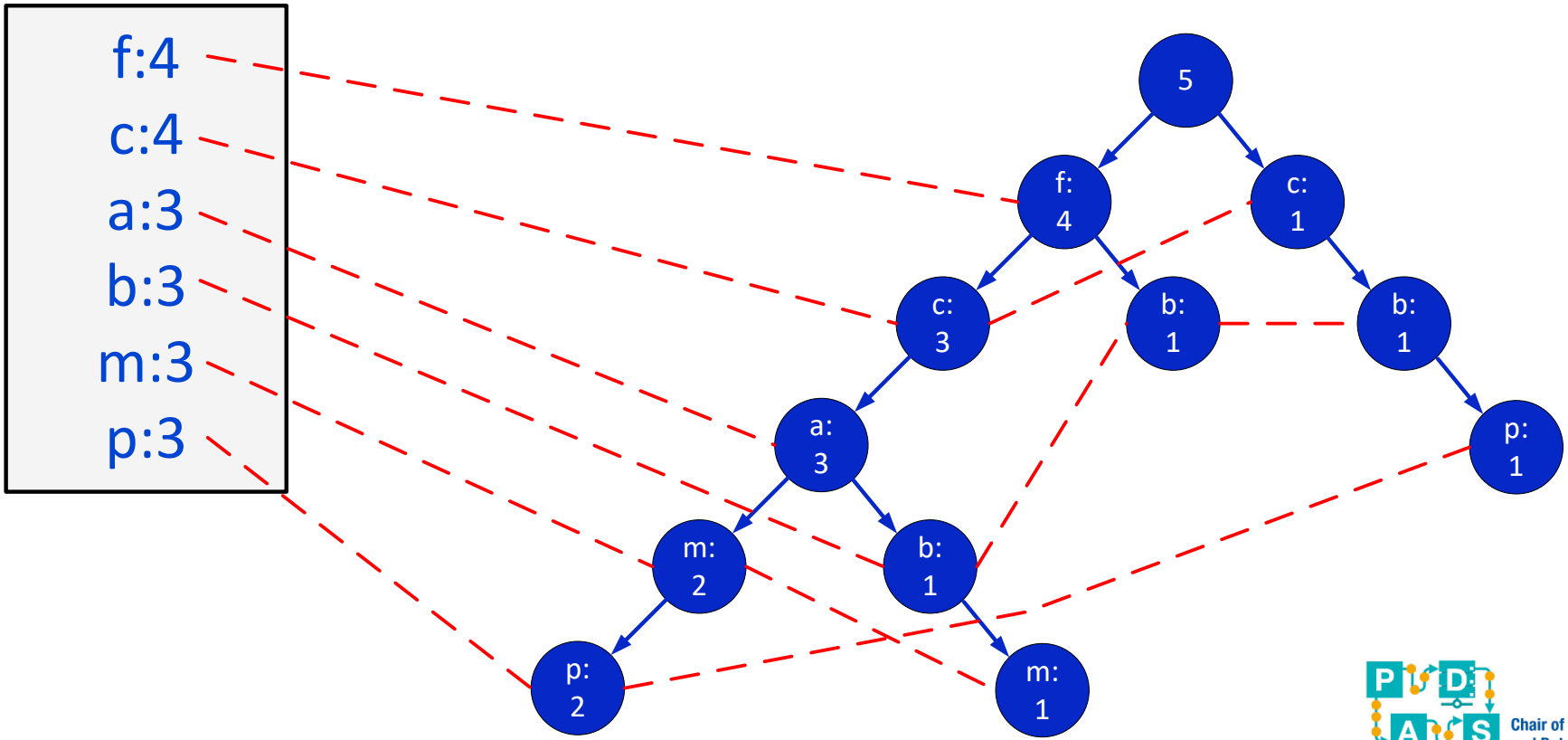
fcamp
fcabm
fb
cbp
fcamp

Build FP-tree

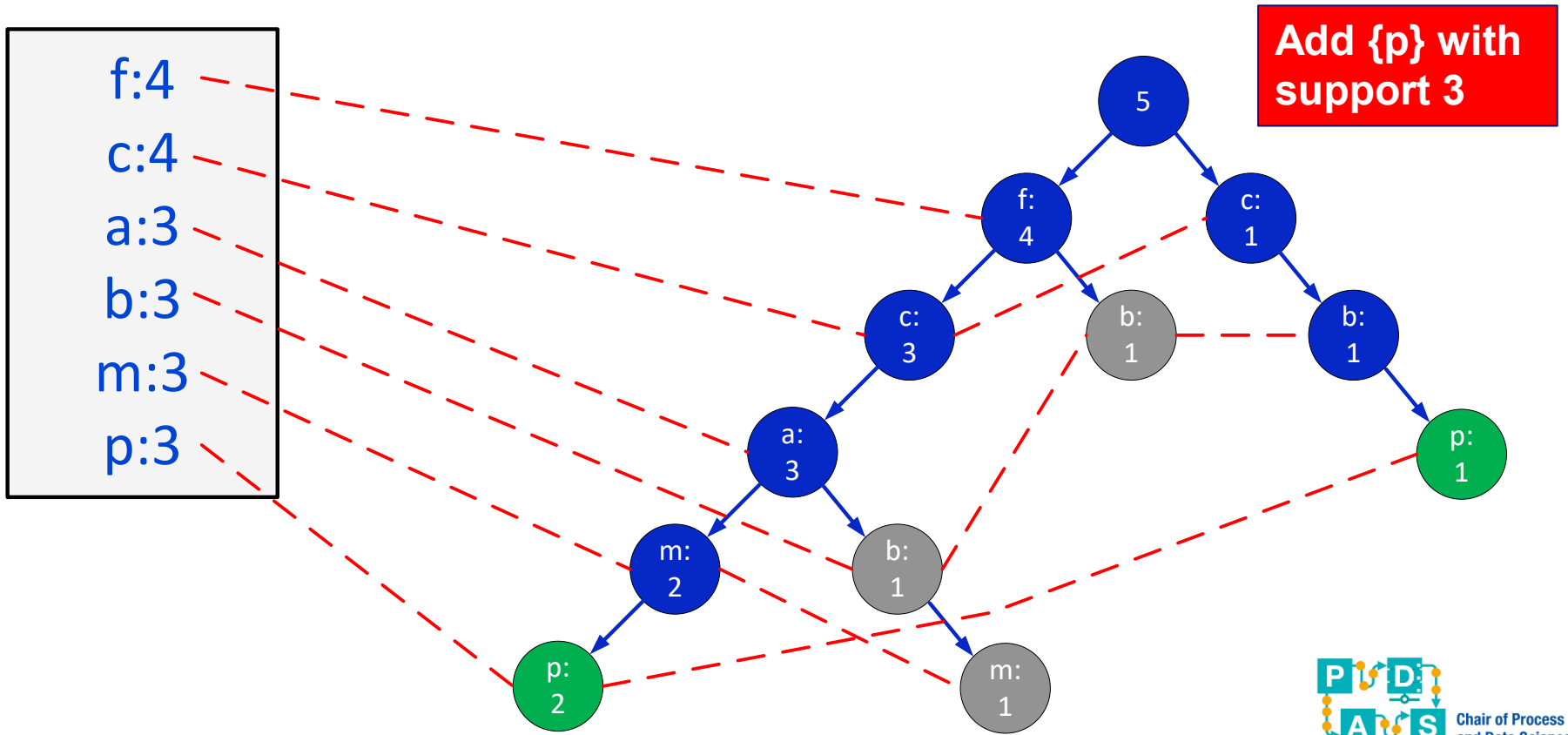
fcamp
fcabm
fb
cbp
fcamp



Node links

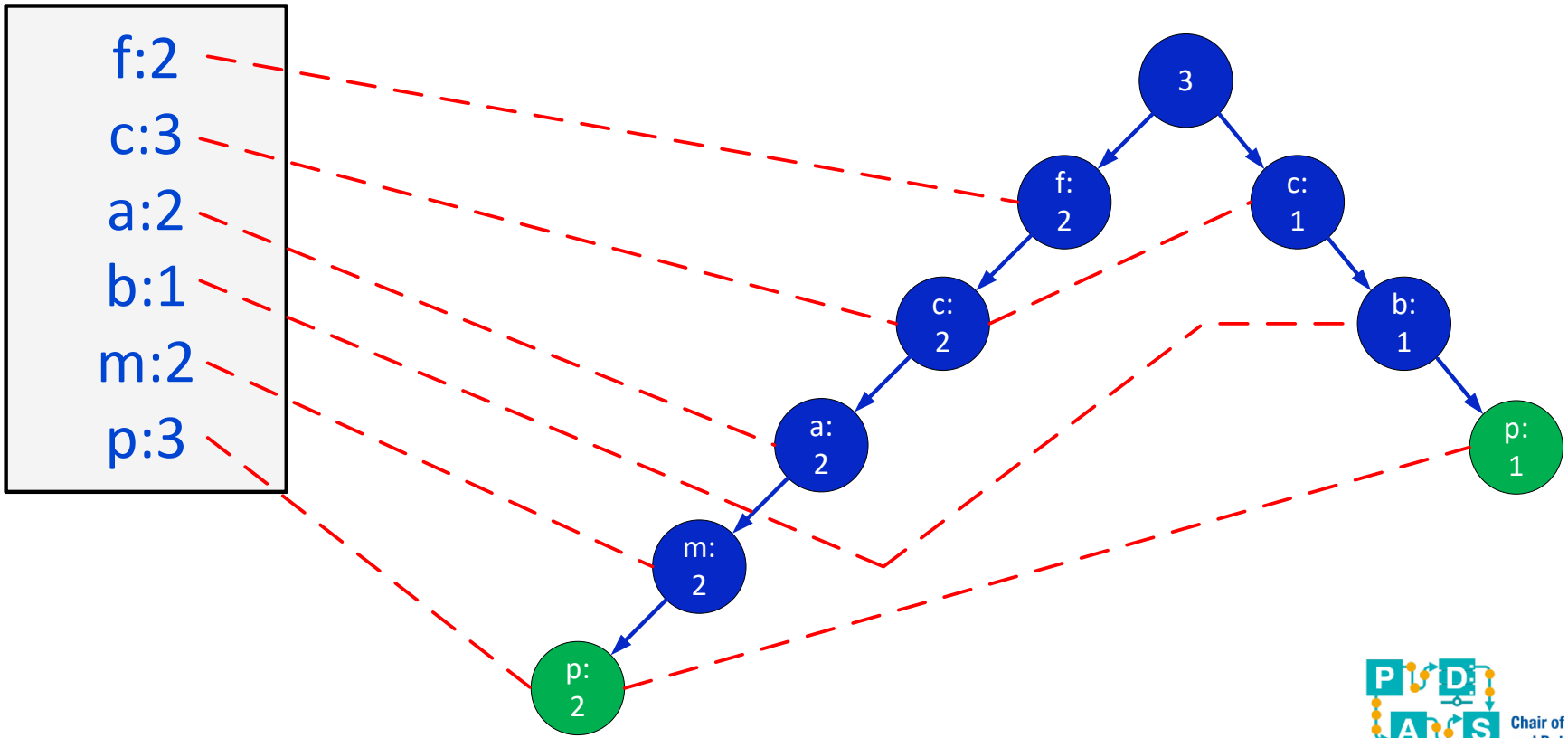


Consider postfix p



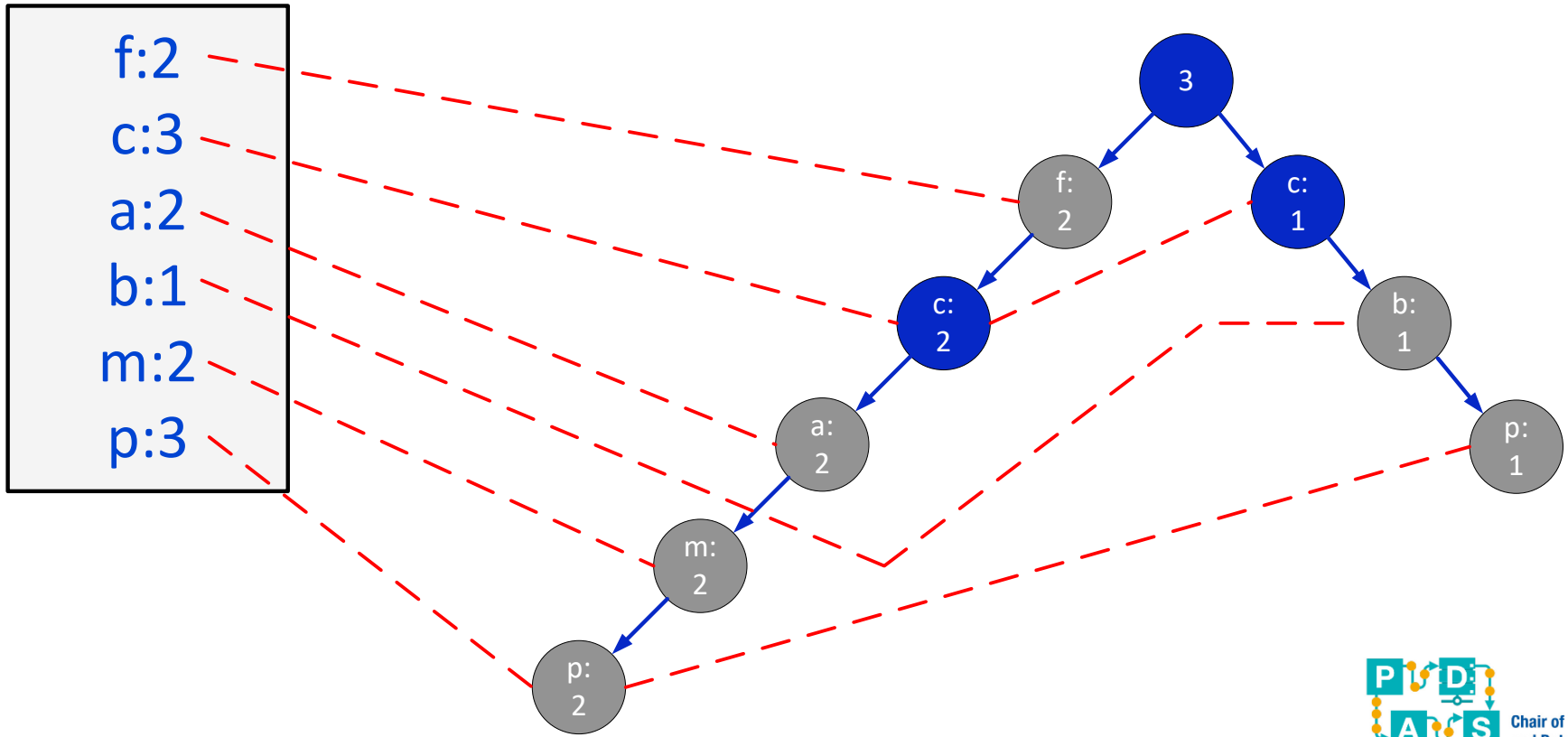
Chair of Process and Data Science

Conditional pattern-base for postfix p

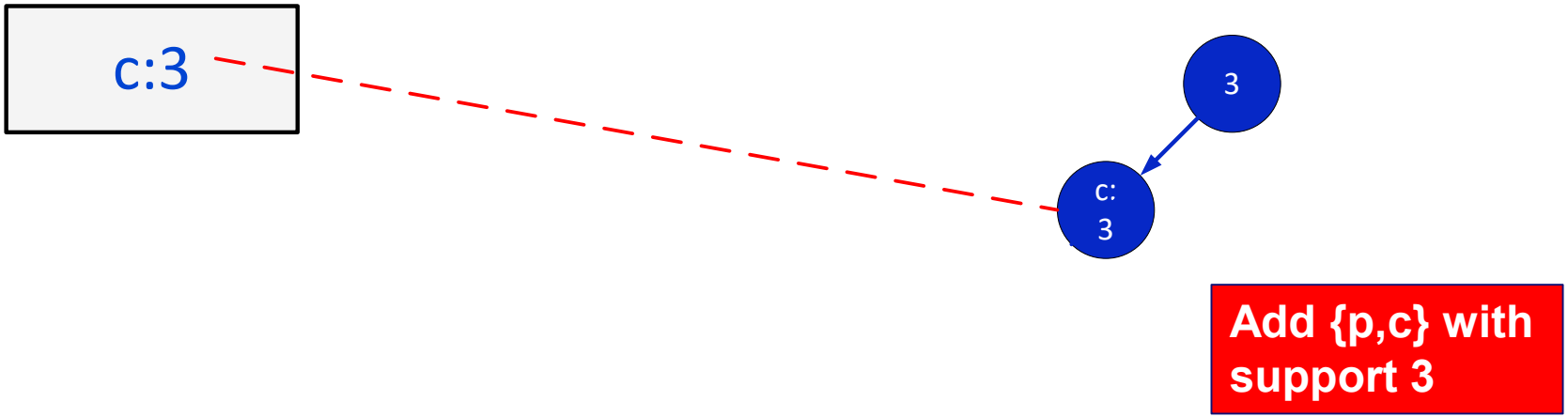


Chair of Process and Data Science

Towards the conditional FP-tree for postfix p

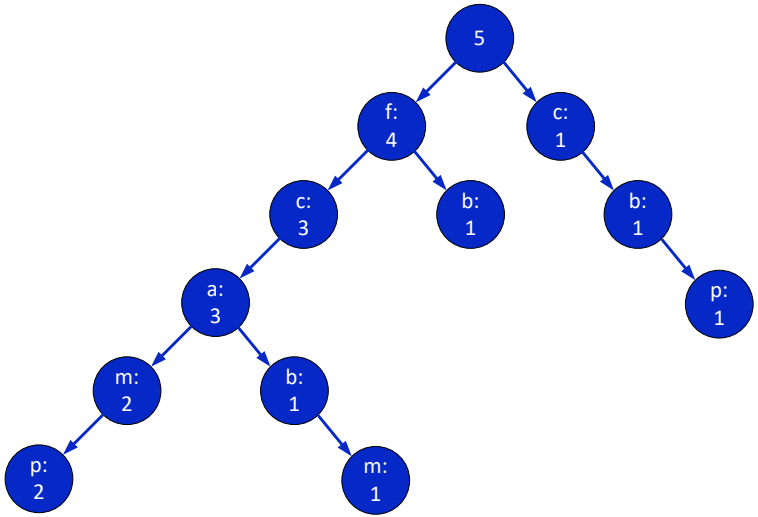


Conditional FP-tree for postfix p



Repeat and recurse …

All frequent item sets generated



Item sets	support
{f},{c}	4
{a}, {b}, {m}, {p}, {p,c}, {m,a,c,f}, {m,c,f}, {m,a,f}, {m,a,c}, {m,a}, {m,c}, {m,f}, {a,c,f}, {a,c}, {a,f}, {c,f}	3

Challenges related to frequent itemsets

- Combinatorial explosion.
- Many frequent itemsets (or suddenly very few).
- How to find the interesting ones?
- How to turn itemsets into rules?

Association rules

Association rules

- $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ is the set of all items.
- $D \in \mathfrak{B}\big(\mathcal{P}(\mathcal{I})\big)$ (such that $\emptyset \notin D$) is the multiset of transactions.
- $A \Rightarrow B$ with $A \subseteq \mathcal{I}$, $B \subseteq \mathcal{I}$, and $A \cap B = \emptyset$ is an association rule.
- For example, $\{\text{bread}\} \Rightarrow \{\text{butter}, \text{cheese}\}$.

Examples

- **{Bitburg, Pumpernickel} ⇒ {Merlot}**
  (people that buy Bitburg pilsner and Pumpernickel bread also tend to buy Merlot wine)
- **{Bitburg} ⇒ {Heineken, Palm}**
  (people that buy Bitburg pilsner tend to buy both Heineken and Palm pilsner)
- **{Carbonara, Margherita } ⇒ {Espresso, Tiramisu}**
  (people that buy Bitburg pilsner and Pumpernickel bread, also tend to buy Merlot wine)
- **{BPI, IDS} ⇒ {APM}**
  (students that take the BPI and IDS courses also tend to take the APM course)
- **{part-245, part-345, part-456} ⇒ {part-372}**
  (when parts 245, 345, and 456 are replaced, then often also part 372 is replaced)

Support and Confidence

- $support(A \Rightarrow B) = support(A \cup B) = \frac{support_{count}(A \cup B)}{support_{count}(\emptyset)}$

Fraction of instances containing at least the items in $A \cup B$.

- $confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)} = \frac{support_{count}(A \cup B)}{support_{count}(A)}$

Fraction of instances containing at least the items in $A \cup B$ of the instances that contain at least the items in A.

Support and Confidence: Example

- $support(\{\text{bread}\} \Rightarrow \{\text{butter}, \text{cheese}\}) =$
  $support(\{\text{bread}, \text{butter}, \text{cheese}\}) =$
  $$\frac{support_{count}(\{\text{bread}, \text{butter}, \text{cheese}\})}{support_{count}(\emptyset)}$$
- $confidence(\{\text{bread}\} \Rightarrow \{\text{butter}, \text{cheese}\}) =$
  $$\frac{support(\{\text{bread}, \text{butter}, \text{cheese}\})}{support(\{\text{bread}\})} =$$
  $$\frac{support_{count}(\{\text{bread}, \text{butter}, \text{cheese}\})}{support_{count}(\{\text{bread}\})}$$

Support and Confidence: Example

- $support(\{butter, bread\} \Rightarrow \{cheese\}) =$
  $support(\{bread, butter, cheese\}) =$
  $$\frac{support_{count}(\{bread, butter, cheese\})}{support_{count}(\emptyset)}$$
  unchanged
- $confidence(\{butter, bread\} \Rightarrow \{cheese\}) =$
  $$\frac{support(\{butter, bread, cheese\})}{support(\{butter, bread\})} =$$
  $$\frac{support_{count}(\{butter, bread, cheese\})}{support_{count}(\{butter, bread\})}$$
  higher
  (divide by a smaller number)

Probabilistic interpretation

- $\boldsymbol{support(A \Rightarrow B) = support(A \cup B) = P(A \cup B)}$

Probability that an instance contains at least $A \cup B$.

- $\boldsymbol{confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)} = P(B|A)}$

Conditional probability that an instance contains the items in B given that it contains the items in A.

Take "probability" with a grain of salt: We are only considering a sample.

Generating association rules from frequent itemsets

Given $D \in \mathfrak{B}(\mathcal{P}(\mathcal{I}))$, min_sup, and min_conf?

How to generate all association rules that have a minimal support min_sup and a minimal confidence min_conf?

$$support(A \Rightarrow B) = support(A \cup B) \geq min\_sup$$

$$confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)} \geq min\_conf$$

Ensuring $support(A \Rightarrow B) \geq min\_sup$

How to ensure $support(A \Rightarrow B) = support(A \cup B) \geq min\_sup$?

✓**Easy!**

- **Consider frequent itemsets as a basis.**
- **We know that there has to be some $C = A \cup B$ such that $|C| \geq 2$ and $support(C) \geq min\_sup$, i.e., apply Apriori or FP-growth to generate all such frequent C item sets.**

Split frequent item sets

- Consider all itemsets C such that $|C| \geq 2$ and $support(C) \geq min\_sup$.
- Consider all splits $C = A \cup B$ such that A and B are disjoint and non-empty.
- This way we can generate all candidate rules $A \Rightarrow B$.

If $|C| = n$, how many candidate rules $A \Rightarrow B$ are there?

- Answer: $2^n - 2$ (distribute the elements, both sides should be non-empty)

Ensuring $confidence(A \Rightarrow B) \geq min\_conf$

- How to ensure $confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)} \geq min\_conf$?
- Note that itemsets $A \cup B$ and A are frequent. Hence, theirs supports have already been computed when using Apriori or FP-growth.
- Therefore, we can simply test every candidate rule $A \Rightarrow B$ and only return the ones where $\frac{support(A \cup B)}{support(A)} \geq min\_conf$.

Example (from book)

Transactional Data for an *AllElectronics* Branch

TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

$\{I1, I2\} \Rightarrow I5$ $\quad confidence = 2/4 = 50\%$

$\{I1, I5\} \Rightarrow I2,$ $\quad confidence = 2/2 = 100\%$

$\{I2, I5\} \Rightarrow I1,$ $\quad confidence = 2/2 = 100\%$

$I1 \Rightarrow \{I2, I5\},$ $\quad confidence = 2/6 = 33\%$

$I2 \Rightarrow \{I1, I5\},$ $\quad confidence = 2/7 = 29\%$

$I5 \Rightarrow \{I1, I2\},$ $\quad confidence = 2/2 = 100\%$

Summary

- Once we have determined the frequent itemsets, we can return all association rules that have a minimal support min_sup and a minimal confidence min_conf.
- However:
  - There may be many!
  - Most are not interesting!

Confidence-based pruning

- **Consider association rule $A \Rightarrow B$, and itemsets X and Y, such that $(X \cup Y) \cap (A \cup B) = \emptyset$.**
  - $confidence(A \cup X \Rightarrow B) \geq confidence(A \Rightarrow B)$
  - $confidence(A \Rightarrow B \cup Y) \leq confidence(A \Rightarrow B)$
  - $confidence(A \cup X \Rightarrow B) \geq confidence(A \Rightarrow B \cup Y)$
- **Hence, if $confidence(A \cup X \Rightarrow B) < min\_conf$, then $confidence(A \Rightarrow B \cup Y) < min\_conf$.**

Removing redundant rules

- Consider two different association rules $A \Rightarrow B$ and $A' \Rightarrow B'$ having an identical support and confidence, i.e., $support(A \Rightarrow B) = support(A' \Rightarrow B')$ and $confidence(A \Rightarrow B) = confidence(A' \Rightarrow B')$.
- $A' \Rightarrow B'$ is redundant if $A' \subseteq A$ and $B' \subseteq B$.

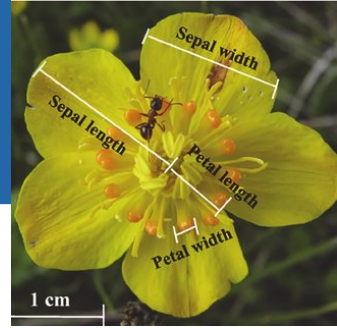
- Using only closed frequent itemsets will avoid generating redundant rules. (Recall: An itemset X is closed if there is no proper superset $Y \supset X$ that has the same support.)

Using only closed frequent itemsets will avoid the generation of redundant rules

- Assume $A' \Rightarrow B'$ is redundant, i.e., there is another rule $A \Rightarrow B$ such that $support(A \Rightarrow B) = support(A' \Rightarrow B')$, $confidence(A \Rightarrow B) = confidence(A' \Rightarrow B')$, and $A' \subseteq A$ and $B' \subseteq B$.
- Also assume $C' = A' \cup B'$ is closed, i.e., is no proper superset $C \supset C'$ that has the same support.
- We find a contradiction:
  - $C' = A' \cup B' \subset A \cup B = C$ (because the rules are different).
  - Hence, $support(A \Rightarrow B) = support(C) < support(C') = support(A' \Rightarrow B')$ (contradiction first assumption).

Example
Iris data set

Iris flower data set



- A multivariate data set introduced Ronald Fisher in 1936.
- 150 records with five attributes: sepal length (a1), sepal width (a2), petal length (a3), petal width (a4) and species.
- Three species: Iris setosa, Iris virginica and Iris versicolor.

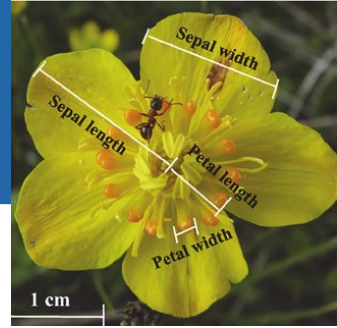
Iris setosa

Iris versicolor

Iris virginica

Chair of Process and Data Science

Iris flower data set



150 instances with five attributes: sepal length (a1), sepal width (a2), petal length (a3), petal width (a4), and species.

sepal length	sepal width	petal length	petal width	iris
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
5	3,6	1,4	0,2	Iris-setosa
5,4	3,9	1,7	0,4	Iris-setosa
4,6	3,4	1,4	0,3	Iris-setosa
5	3,4	1,5	0,2	Iris-setosa
4,4	2,9	1,4	0,2	Iris-setosa
4,9	3,1	1,5	0,1	Iris-setosa
5,4	3,7	1,5	0,2	Iris-setosa
…	…	…	…	…

Iris setosa

Iris versicolor

Iris virginica

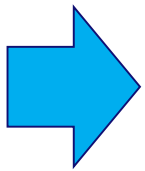
Chair of Process and Data Science

Transform into itemsets

(actually a supervised approach would provide more insights)

- **20 different items, e.g., 5 items for sepal length (a1)**

sepal length
5,1
4,9
4,7
4,6
5
5,4
4,6
5
4,4
4,9
5,4
4,8



Chair of Process and Data Science

Itemset data

label ↑	a1 = range1 [-∞ - 5.050]	a1 = range2 [5.050 - 5.650]	a1 = range3 [5.650 - 6.150]	a1 = range4 [6.150 - 6...	a1 = range5 [6.550 - ∞]	a2 = range1 [-∞ - 2.750]
Iris-setosa	false	true	false	false	false	false
Iris-setosa	true	false	false	false	false	false
Iris-setosa	true	false	false	false	false	false
Iris-setosa	true	false	false	false	false	false
Iris-setosa	true	false	false	false	false	false
Iris-setosa	false	true	false	false	false	false
Iris-setosa	true	false	false	false	false	false
Iris-setosa	true	false	false	false	false	false
Iris-setosa	true	false	false	false	false	false
Iris-setosa	true	false	false	false	false	false
Iris-setosa	false	true	false	false	false	false
Iris-setosa	true	false	false	false	false	false
Iris-setosa	true	false	false	false	false	false
Iris-setosa	true	false	false	false	false	false

Apply FP-growth (min_sup=0.1)

No. of Sets: 32
Total Max. Size: 3

Min. Size: 1
Max. Size: 3
Contains Item:
Update View

Size	Support	Item 1	Item 2	Item 3
1	0.333	a2 = range2 [2.750 - 3.050]		
1	0.253	a4 = range3 [1.150 - 1.550]		
1	0.247	a3 = range1 [-∞ - 1.550]		
1	0.227	a4 = range1 [-∞ - 0.250]		
1	0.220	a1 = range2 [5.050 - 5.650]		
1	0.220	a2 = range1 [-∞ - 2.750]		
1	0.213	a1 = range1 [-∞ - 5.050]		
1	0.207	a2 = range4 [3.150 - 3.450]		
1	0.200	a1 = range3 [5.650 - 6.150]		
1	0.200	a1 = range5 [6.550 - ∞]		
1	0.200	a3 = range4 [4.650 - 5.350]		
1	0.200	a3 = range5 [5.350 - ∞]		
1	0.193	a3 = range3 [3.950 - 4.650]		
1	0.193	a4 = range5 [1.950 - ∞]		
1	0.173	a4 = range2 [0.250 - 1.150]		
1	0.167	a1 = range4 [6.150 - 6.550]		
1	0.160	a2 = range5 [3.450 - ∞]		
1	0.160	a3 = range2 [1.550 - 3.950]		
1	0.			
2	0.			
2	0.			
2	0.			

32 frequent itemsets

2	0.113	a1 = range5 [6.550 - ∞]	a4 = range5 [1.950 - ∞]	
2	0.140	a3 = range5 [5.350 - ∞]	a4 = range5 [1.950 - ∞]	
3	0.107	a2 = range2 [2.750 - 3.050]	a4 = range3 [1.150 - 1.550]	a3 = range3 [3.950 - 4.650]
3	0.113	a3 = range1 [-∞ - 1.550]	a4 = range1 [-∞ - 0.250]	a1 = range1 [-∞ - 5.050]
3	0.100	a1 = range5 [6.550 - ∞]	a3 = range5 [5.350 - ∞]	a4 = range5 [1.950 - ∞]

Return association rules (min_conf=0.7)

Premises	Conclusion	Support	Confidence
a1 = range5 [6.550 - ∞], a3 = range5 [5.350 - ∞]	a4 = range5 [1.950 - ∞]	0.100	0.750
a2 = range2 [2.750 - 3.050], a4 = range3 [1.150 - 1...	a3 = range3 [3.950 - 4.650]	0.107	0.762
a4 = range1 [-∞ - 0.250], a1 = range1 [-∞ - 5.050]	a3 = range1 [-∞ - 1.550]	0.113	0.773
a4 = range1 [-∞ - 0.250]	a3 = range1 [-∞ - 1.550]	0.180	0.794
a3 = range1 [-∞ - 1.550], a1 = range1 [-∞ - 5.050]	a4 = range1 [-∞ - 0.250]	0.113	0.810
a3 = range3 [3.950 - 4.650]	a4 = range3 [1.150 - 1.550]	0.167	0.862
a1 = range5 [6.550 - ∞], a4 = range5 [1.950 - ∞]	a3 = range5 [5.350 - ∞]	0.100	0.882
a2 = range2 [2.750 - 3.050], a3 = range3 [3.950 - 4...	a4 = range3 [1.150 - 1.550]	0.107	1

8 rules

Apply FP-growth (min_sup=0.01) (was 0.1)

No. of Sets: 277

Total Max. Size: 4

Min. Size: 1

Max. Size: 4

Contains Item:

Update View

Size	Support	Item 1	Item 2
1	0.333	a2 = range2 [2.750 - 3.050]	
1	0.253	a4 = range3 [1.150 - 1.550]	
1	0.247	a3 = range1 [-∞ - 1.550]	
1	0.227	a4 = range1 [-∞ - 0.250]	
1	0.220	a1 = range2 [5.050 - 5.650]	
1	0.220	a2 = range1 [-∞ - 2.750]	
1	0.213	a1 = range1 [-∞ - 5.050]	
1	0.207	a2 = range4 [3.150 - 3.450]	
1	0.200	a1 = range3 [5.650 - 6.150]	

277 frequent itemsets (was 32)

Return association rules (min_conf=0.7)

Premises	Conclusion	Support	Confidence
a3 = range5 [5.350 - ∞]	a4 = range5 [1.950 - ∞]	0.140	0.700
a2 = range1 [-∞ - 2.750], a3 = range3 [3.950 - 4.650]	a4 = range3 [1.150 - 1.550]	0.047	0.700
a3 = range1 [-∞ - 1.550], a2 = range4 [3.150 - 3.450]	a1 = range1 [-∞ - 5.050]	0.047	0.700
a3 = range1 [-∞ - 1.550], a4 = range1 [-∞ - 0.250], ...	a1 = range2 [5.050 - 5.650]	0.047	0.700
a2 = range2 [2.750 - 3.050], a1 = range3 [5.650 - 6...	a4 = range3 [1.150 - 1.550]	0.067	0.714
a4 = range3 [1.150 - 1.550], a1 = range3 [5.650 - 6...	a2 = range2 [2.750 - 3.050]	0.067	0.714
a4 = range1 [-∞ - 0.250], a3 = range2 [1.550 - 3.950]	a1 = range1 [-∞ - 5.050]	0.033	0.714
a2 = range4 [3.150 - 3.450], a1 = range4 [6.150 - 6...	a4 = range5 [1.950 - ∞]	0.033	0.714
a3 = range5 [5.350 - ∞], a4 = range5 [1.950 - ∞]	a1 = range5 [6.550 - ∞]	0.100	0.714
a2 = range2 [2.750 - 3.050], a1 = range1 [-∞ - 5.050]	a3 = range1 [-∞ - 1.550], a4 = range1 [-∞ - 0.250]	0.033	0.714

131 rules (was 8)

Return association rules (min_conf=0.5)

Premises	Conclusion	Support	Confidence
a2 = range3 [3.050 - 3.150]	a1 = range5 [6.550 - ∞]	0.040	0.500
a4 = range3 [1.150 - 1.550], a1 = range4 [6.150 - 6...	a2 = range2 [2.750 - 3.050]	0.027	0.500
a1 = range3 [5.650 - 6.150], a3 = range4 [4.650 - 5...	a2 = range2 [2.750 - 3.050]	0.040	0.500
a2 = range2 [2.750 - 3.050], a3 = range3 [3.950 - 4...	a1 = range3 [5.650 - 6.150]	0.053	0.500
a2 = range2 [2.750 - 3.050], a4 = range5 [1.950 - ∞]	a1 = range5 [6.550 - ∞]	0.040	0.500
a3 = range4 [4.650 - 5.350], a4 = range5 [1.950 - ∞]	a2 = range2 [2.750 - 3.050]	0.027	0.500
a2 = range2 [2.750 - 3.050], a4 = range4 [1.550 - 1...	a3 = range4 [4.650 - 5.350]	0.033	0.500
a2 = range2 [2.750 - 3.050], a4 = range4 [1.550 - 1...	a3 = range5 [5.350 - ∞]	0.033	0.500
a3 = range3 [3.950 - 4.650], a1 = range4 [6.150 - 6...	a2 = range2 [2.750 - 3.050]	0.020	0.500

323 rules (was first 8, then 131)

Challenges

- **Many rules**
- **Confidence and support are only part of the story**

Evaluating association rules

Confusion matrix for association rules

Consider association rule $A \Rightarrow B$

$A \Rightarrow B$	B is included	B is not included	
A is included	#AB	#A~~B~~	#A
A is not included	#~~A~~B	#~~AB~~	#~~A~~
	#B	#~~B~~	#all

$$support(A \Rightarrow B) = {}^{\#AB}/_{\#all} \qquad confidence(A \Rightarrow B) = {}^{\#AB}/_{\#A}$$

Confusion matrix for association rules

Consider association rule $A \Rightarrow B$

$A \Rightarrow B$	B is included	B is not included	
A is included	#AB	#A~~B~~	#A
A is not included	#~~A~~B	#~~AB~~	#~~A~~
	#B	#~~B~~	#all

$$support(A \Rightarrow B) = {}^{\#AB}/_{\#all} \qquad confidence(A \Rightarrow B) = {}^{\#AB}/_{\#A}$$

Chair of Process and Data Science

High support, high confidence

$A \Rightarrow B$	B is included	B is not included	
A is included	100	0	100
A is not included	0	0	0
	100	0	100

$$support(A \Rightarrow B) = {}^{100}/_{100} \qquad confidence(A \Rightarrow B) = {}^{100}/_{100}$$

Low support, high confidence

$A \Rightarrow B$	B is included	B is not included	
A is included	10	0	10
A is not included	40	50	90
	50	50	100

$support(A \Rightarrow B) = {}^{10}/_{100}$

$confidence(A \Rightarrow B) = {}^{10}/_{10}$

Low support, low confidence

$A \Rightarrow B$	B is included	B is not included	
A is included	10	40	50
A is not included	25	25	50
	35	65	100

$$support(A \Rightarrow B) = {}^{10}/_{100} \qquad confidence(A \Rightarrow B) = {}^{10}/_{50}$$

How about this one?

$A \Rightarrow B$	B is included	B is not included	
A is included	80	10	90
A is not included	0	10	10
	80	20	100

$$support(A \Rightarrow B) = {}^{80}/_{100} \qquad confidence(A \Rightarrow B) = {}^{80}/_{90}$$

Seems to be a good rule because if A is not included, B is also never included.

And how about this one?

$A \Rightarrow B$	B is included	B is not included	
A is included	80	10	90
A is not included	10	0	10
	90	10	100

$$support(A \Rightarrow B) = {}^{80}/_{100} \qquad confidence(A \Rightarrow B) = {}^{80}/_{90}$$

Same support and confidence. However, it seems a poor rule because if A is not included, B is always included.

Need lift

$$lift(A \Rightarrow B) = \frac{support(A \cup B)}{support(A) \cdot support(b)} = \frac{P(A \cup B)}{P(A) \cdot P(B)}$$

If $lift(A \Rightarrow B) \approx 1$ then A and B are independent

If $lift(A \Rightarrow B) \ll 1$ then A and B are negatively correlated

If $lift(A \Rightarrow B) \gg 1$ then A and B are positively correlated

Is the rule surprising?

$A \Rightarrow B$	B is included	B is not included	
A is included	9	1	10
A is not included	81	9	90
	90	10	100

$$confidence(A \Rightarrow B) = {}^{9}/_{10}$$

$$lift(A \Rightarrow B) = \frac{{}^{9}/_{100}}{{}^{10}/_{100}\,{}^{90}/_{100}} = 1$$

No!

Is the rule surprising?

$A \Rightarrow B$	B is included	B is not included	
A is included	9	1	10
A is not included	0	90	90
	9	91	100

$$confidence(A \Rightarrow B) = {}^{9}/_{10}$$

$$lift(A \Rightarrow B) = \frac{{}^{9}/_{100}}{{}^{9}/_{100}\,{}^{10}/_{100}} = 10$$

Yes!

Chair of Process and Data Science

Is the rule surprising?

$A \Rightarrow B$	B is included	B is not included	
A is included	9	1	10
A is not included	90	0	90
	99	1	100

$$confidence(A \Rightarrow B) = {}^{9}/_{10}$$

$$lift(A \Rightarrow B) = \frac{{}^{9}/_{100}}{{}^{99}/_{100}\ {}^{10}/_{100}} = \frac{10}{11}$$

A little bit, but …

Chair of Process and Data Science

Simpson's paradox

Simpson’s paradox

- Simpson's paradox is the phenomenon that a trend appears in several different groups of data but disappears or reverses when these groups are combined.
- First described by Edward Simpson in 1951.
- Nice example of “How to lie with statistics?”.
- The paradox refers to problems often encountered in social-science and medical-science.

Simpson's paradox

$A \Rightarrow B$	B is included	B is not included	
A is included	a+p	(b-a) + (q-p)	b+q
A is not included	c+r	(d-c) + (s-r)	d+s
	a+c+p+r	b+d+q+s - (a+c+p+r)	b+d+q+s

$$confidence(A \Rightarrow B) = {a+p}/{b+q}$$

$$lift(A \Rightarrow B) = \frac{{a+p}/{b+d+q+s}}{{b+q}/{b+d+q+s} \quad {a+c+p+r}/{b+d+q+s}}$$

Simpson's paradox

$A \Rightarrow B$	B is included	B is not included	
A is included	a+p	(b-a) + (q-p)	b+q
A is not included	c+r	(d-c) + (s-r)	d+s
	a+c+p+r	b+d+q+s - (a+c+p+r)	b+d+q+s

$$confidence(A \Rightarrow B) = {}^{a+p}/_{b+q} \qquad lift(A \Rightarrow B) = \frac{(a+p)(b+d+q+s)}{(b+q)(a+c+p+r)}$$

Is this possible?

$A \Rightarrow B$	B is included	B is not included	
A is included	a+p	(b-a) + (q-p)	b+q
A is not included	c+r	(d-c) + (s-r)	d+s
	a+c+p+r	b+d+q+s - (a+c+p+r)	b+d+q+s

$$confidence(A \Rightarrow B) = {}^{a+p}/_{b+q} > confidence(notA \Rightarrow B) = {}^{c+r}/_{d+s}$$

$$confidence(A \Rightarrow B) = {}^{a}/_{b} < confidence(notA \Rightarrow B) = {}^{c}/_{d}$$

$$confidence(A \Rightarrow B) = {}^{p}/_{q} < confidence(notA \Rightarrow B) = {}^{r}/_{s}$$

Is this possible?

$A \Rightarrow B$	B is included	B is not included	
A is included	a+p	(b-a) + (q-p)	b+q
A is not included	c+r	(d-c) + (s-r)	d+s
	a+c+p+r	b+d+q+s - (a+c+p+r)	b+d+q+s

$$\frac{a+p}{b+q} > \frac{c+r}{d+s} \qquad \frac{a}{b} < \frac{c}{d} \qquad \frac{p}{q} < \frac{r}{s}$$

Is this possible? Yes!

$$\frac{a+p}{b+q} > \frac{c+r}{d+s} \qquad \frac{a}{b} < \frac{c}{d} \qquad \frac{p}{q} < \frac{r}{s}$$

a = 1
b = 3
c = 34
d = 100
p = 66
q = 100
r = 2
s = 3

$$\frac{1+66}{3+100} > \frac{34+2}{100+3} \qquad \frac{1}{3} < \frac{34}{100} \qquad \frac{66}{100} < \frac{2}{3}$$

$$\frac{67}{103} > \frac{36}{103}$$

Simpson's paradox

$A \Rightarrow B$	B is included	B is not included	
A is included	1+66	(3-1) + (100-66)	3+100
A is not included	34+2	(100-34) + (3-2)	100+3
	1+34+66+2	3+100+100+3 - (1+34+66+2)	3+100+ 100+3

$$confidence(A \Rightarrow B) = {}^{1+66}/_{3+100} > confidence(notA \Rightarrow B) = {}^{34+2}/_{100+3}$$

$$confidence(A \Rightarrow B) = {}^{1}/_{3} < confidence(notA \Rightarrow B) = {}^{34}/_{100}$$

$$confidence(A \Rightarrow B) = {}^{66}/_{100} < confidence(notA \Rightarrow B) = {}^{2}/_{3}$$

a = 1
b = 3
c = 34
d = 100
p = 66
q = 100
r = 2
s = 3

Chair of Process and Data Science

Simpson's paradox

$A \Rightarrow B$	B is included	B is not included	
A is included	1+66	2+34	3+100
A is not included	34+2	66+1	100+3
	35+68	68+35	103+103

$$confidence(A \Rightarrow B) = {}^{67}/_{103} > confidence(notA \Rightarrow B) = {}^{36}/_{103}$$

$$confidence(A \Rightarrow B) = {}^{1}/_{3} < confidence(notA \Rightarrow B) = {}^{34}/_{100}$$

$$confidence(A \Rightarrow B) = {}^{66}/_{100} < confidence(notA \Rightarrow B) = {}^{2}/_{3}$$

Chair of Process and Data Science

Simpson's paradox

$A \Rightarrow B$	B is included	B is not included	
A is included	1+66	2+34	3+100
A is not included	34+2	66+1	100+3
	35+68	68+35	103+103

The presence of A has a positive effect on the occurrence of B in the overall set (supporting the rule). However, the effect cannot be seen in the subsets!

Simpson's paradox

$A \Rightarrow B$	B is included	B is not included	
A is included	1+66	2+34	3+100
A is not included	34+2	66+1	100+3
	35+68	68+35	103+103

$$\text{lift}(A \Rightarrow B) = \frac{103 \cdot 1}{35 \cdot 3} = 0.981$$

$$\text{lift}(A \Rightarrow B) = \frac{103 \cdot 66}{100 \cdot 68} = 0.999$$

$$\text{lift}(A \Rightarrow B) = \frac{206 \cdot 67}{103 \cdot 103} = 1.307$$

Chair of Process and Data Science

Simpson's paradox

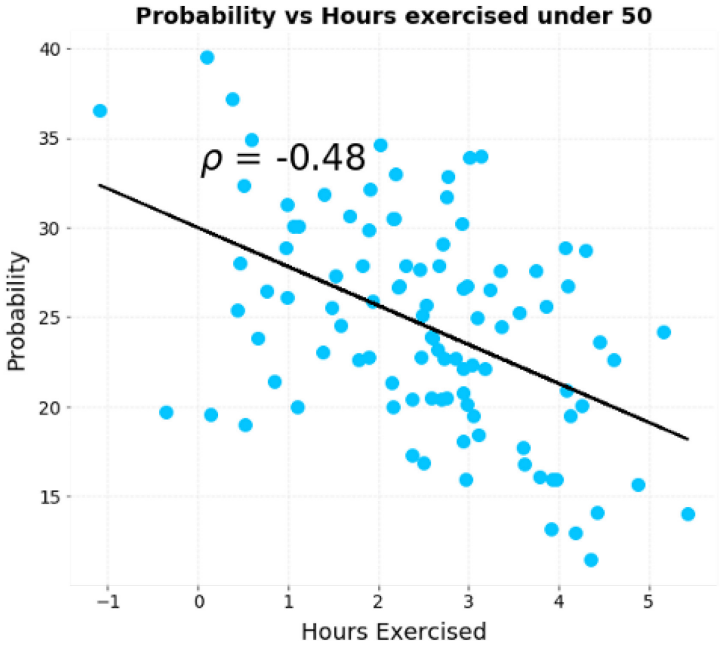
	Recommend Sophia's	Recommend Carlo's
Male	$\frac{50}{150} = 30\%$	$\frac{180}{360} = 50\%$
Female	$\frac{200}{250} = 80\%$	$\frac{36}{40} = 90\%$
Combined	$\frac{250}{400} = 62.5\%$	$\frac{216}{400} = 54\%$

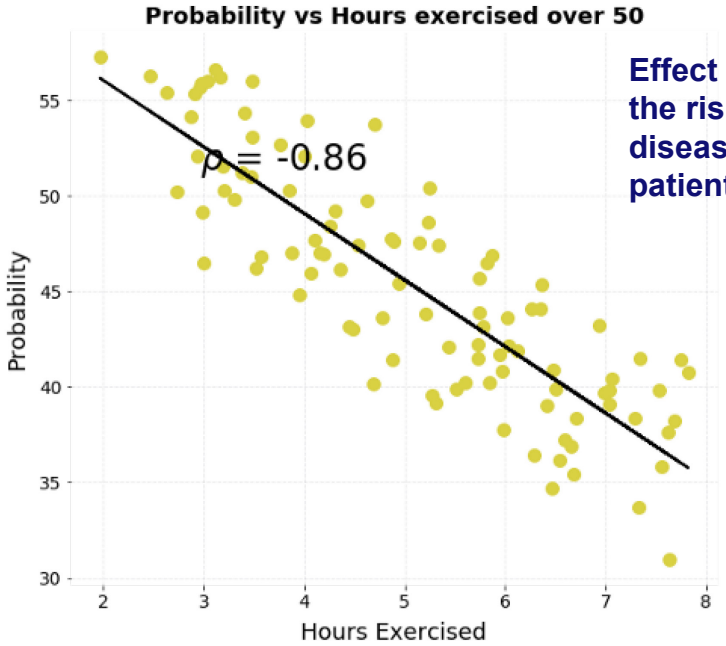
Each fraction shows the number of users who would recommend the restaurant out of the number asked. Carlo's has far more responses from men than from women while the reverse is true for Sophia's.

The data clearly show that Carlo's is preferred when the data are separated, but Sophia's is preferred when the data are combined!

Simpson's Paradox: How to Prove Opposite Arguments with the Same Data by William Koehrsen towardsdatascience.com

Simpson's paradox (it is good to exercise)



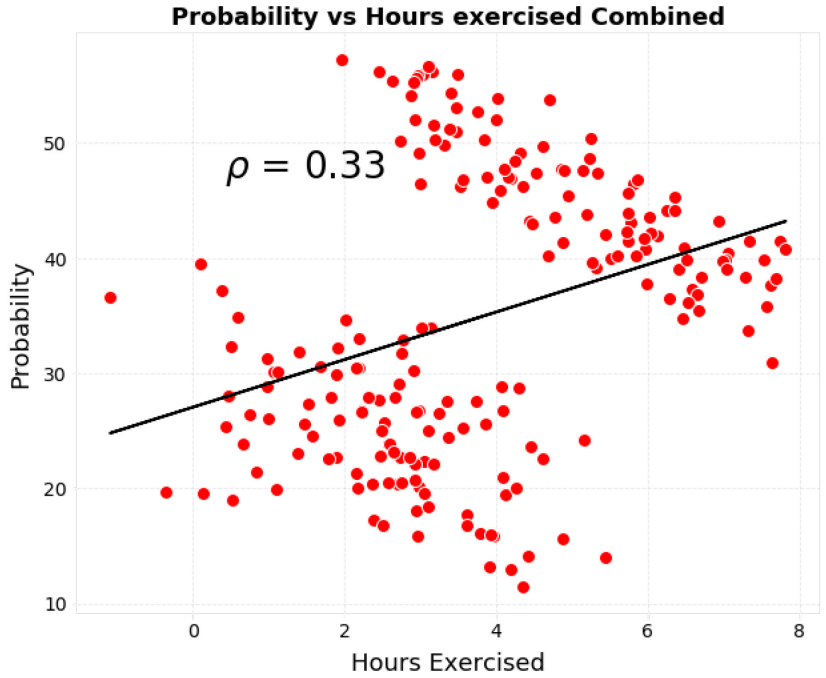


Effect of exercising on the risk of developing a disease for two sets of patients (below/over 50).

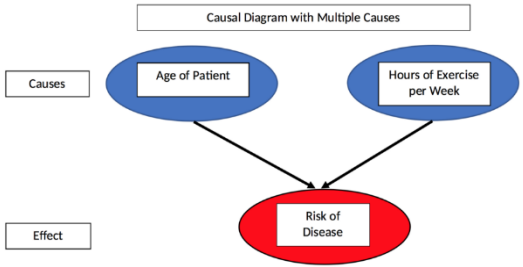
Simpson's Paradox: How to Prove Opposite Arguments with the Same Data by William Koehrsen towardsdatascience.com

Chair of Process and Data Science

Simpson's paradox (exercising will kill you)



Effect of exercising on the risk of developing a disease for the whole group.



Simpson's Paradox: How to Prove Opposite Arguments with the Same Data by William Koehrsen towardsdatascience.com

Chair of Process and Data Science

Conclusion

Short summary of lecture

- Frequent itemsets can be used to build rules.
- Quality notions: support, confidence, and lift.
- Be careful interpreting measures without understanding the data (e.g., Simpson's paradox).

#	Lecture	date	day
Lecture 1	Introduction	10/10/2018	Wednesday
Lecture 2	Crash Course in Python	11/10/2018	Thursday
Instruction 1	*Python*	*12/10/2018*	*Friday*
Lecture 3	Basic data visualisation/exploration	17/10/2018	Wednesday
Lecture 4	Decision trees	18/10/2018	Thursday
Instruction 2	*Decision trees and data visualization/exploration*	*19/10/2018*	*Friday*
Lecture 5	Regression	24/10/2018	Wednesday
Lecture 6	Support vector machines	25/10/2018	Thursday
Instruction 3	*Regression and support vector machines*	*26/10/2018*	*Friday*
Lecture 7	Neural networks (1/2)	31/10/2018	Wednesday
Instruction 4	*Neural networks and supervised learning*	*02/11/2018*	*Friday*
Lecture 8	Neural networks (2/2)	07/11/2018	Wednesday
Lecture 9	Evaluation of supervised learning problems	08/11/2018	Thursday
Instruction 5	*Neural networks and supervised learning*	*09/11/2018*	*Friday*
Lecture 10	Clustering	14/11/2018	Wednesday
Lecture 11	Frequent items sets	15/11/2018	Thursday
Lecture 12	Association rules	21/11/2018	Wednesday
Lecture 13	Sequence mining	22/11/2018	Thursday
Instruction 6	*Clustering, frequent items sets, association rules*	*23/11/2018*	*Friday*
Lecture 14	Process mining (unsupervised)	28/11/2018	Wednesday
Lecture 15	Process mining (supervised)	29/11/2018	Thursday
Instruction 7	*Process mining and sequence mining*	*30/11/2018*	*Friday*
backup		31/01/2019	Thursday
extra	*Question hour*	*01/02/2019*	*Friday*