Schedule

- Thu 4/26 (5 teams)
  - (Sports) EBM: Examining Baseball's Metrics
  - (Misc.) PBC: Predicting Breast Cancer
  - (Misc.) DBH: Determining predictors of H-1B salary and approval
  - (Misc.) AFG: It's All Funds & Games - Predicting Kickstarter Success
  - (Misc.) MPT: Information Extraction from Text Data

- Tue 5/1 (5 teams)
  - (Sports) MLB: Predicting MLB Performance Based on Minor League Statistics
  - (Sports) MML: Making March Less Mad - Predicting the NCAA Men's Basketball Tournament
  - (Sports) POW: Predicting the Outcome of Week 1 Collegiate Football Games
  - (Movie) NPM: The Netflix Problem: Movie Clustering and Classification Based on Ratings
  - (Movie) ACC: Actor Clustering and Cast Significance

Schedule (cont.)

Presentation	QA	Teams (Apr 26, May 1)
2:01-2:12 pm	2:13-2:15 pm	EBM (baseball), MLB
2:16-2:27 pm	2:28-2:30 pm	PBC (cancer), MML (madness)
2:31-2:42 pm	2:43-2:45 pm	DBH (h1b), POW (week 1)
2:46-2:57 pm	2:58-3:00 pm	AFG (funds), NPM (netflix)
3:01-3:12 pm	3:13-3:15 pm	MPT (text), ACC (actor)

- **You must attend the presentation when your group is presenting.**
- **Please send me the slides after your presentation.**
- **For the students who will not be able to come to the class, you can fill in grading forms based on the slides, however, it will only take 50% weight when we average all the scores on each team.**

Example: If student A and B give 90 and 80 to group X (neither A or B is in X),
student C gives 85; A and B attended the class, C did not; then
*the score from students to the members in group X will be (90+80+85*0.5)/2.5 = 85.*

Dr. Taeho Jung

Data Security and Privacy Lab (DSP-Lab)
CSE 20110 Discrete Mathematics (Fall 2017)
CSE 40622 Cryptography (Spring 2018)

Grading Oral Presentation

Introduction:	15%	Provide context. What questions are being addressed?
Solution/Method:	30%	What did you do? Why did you choose this method? What tools and techniques did you use?
Data and Experiments:	10%	What data did you use? Are your experimental methods reliable?
Evaluation and Results:	30%	What evaluation did you do? Do your conclusions match your results?
Presentation Quality:	15%	Clarity of speaking (5%), organization (5%), and visuals (5%).

Grading Form

- Students (anonymized; skip your own team): 60%
- Invited faculty: 30%
- Instructor: 10%

	Intro (15)	Solution, method (30)	Data and experiments (10)	Evaluation, analysis, results (30)	Presentation quality (15)	Sum (100)
EBM						
PBC						
DBH						
AFG						
MPT						
MLB						
MML						
POW						
NPM						
ACC						

How to Have Grade A?

- Calculated score >= 93
  - **HW1*5% + HW2*5% + HW3*5% + HW4*5%**
  - **Mid exam*20%** (at most 100*20% though honor code bonus)
  - **Final exam*30%** (no honor code bonus)
  - Course project
    - **Proposal*(100/10)*3% + Milestone*(100/15)*4.5%**
    - Presentation (at most 100*7.5%, up to +20% for early-bird: Apr. 26)

      83.333 → 100 (may happen)
      - **Students*4.5%**
      - **Invited faculty*2.25%**
      - **Instructor*0.75%**
    - **Final project paper*7.5%**
      - Usually proportional to the presentation
    - **Code/data package*7.5%**

Letter Grades

- A: [93, 100]
- A-: [90, 93)
- B+: [87, 90)
- B: [84, 87)
- B-: [81, 84)
- C+: [78, 81)
- C: [75, 78)

Final Exam

- Time: May 8 (Tuesday) 10:30 am – 12:30 pm
- Location: 117 DeBartolo

- Write down your answers/solutions on the blue book.
- Return your exam paper after the exam.
- You can have a double-sided letter-size reference paper.

- You must bring a pen/pencil/writing tool.
- You had better bring a calculator.
- You are not allowed to use laptop/computer/cellphone!
- You are not allowed to bring text book.

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

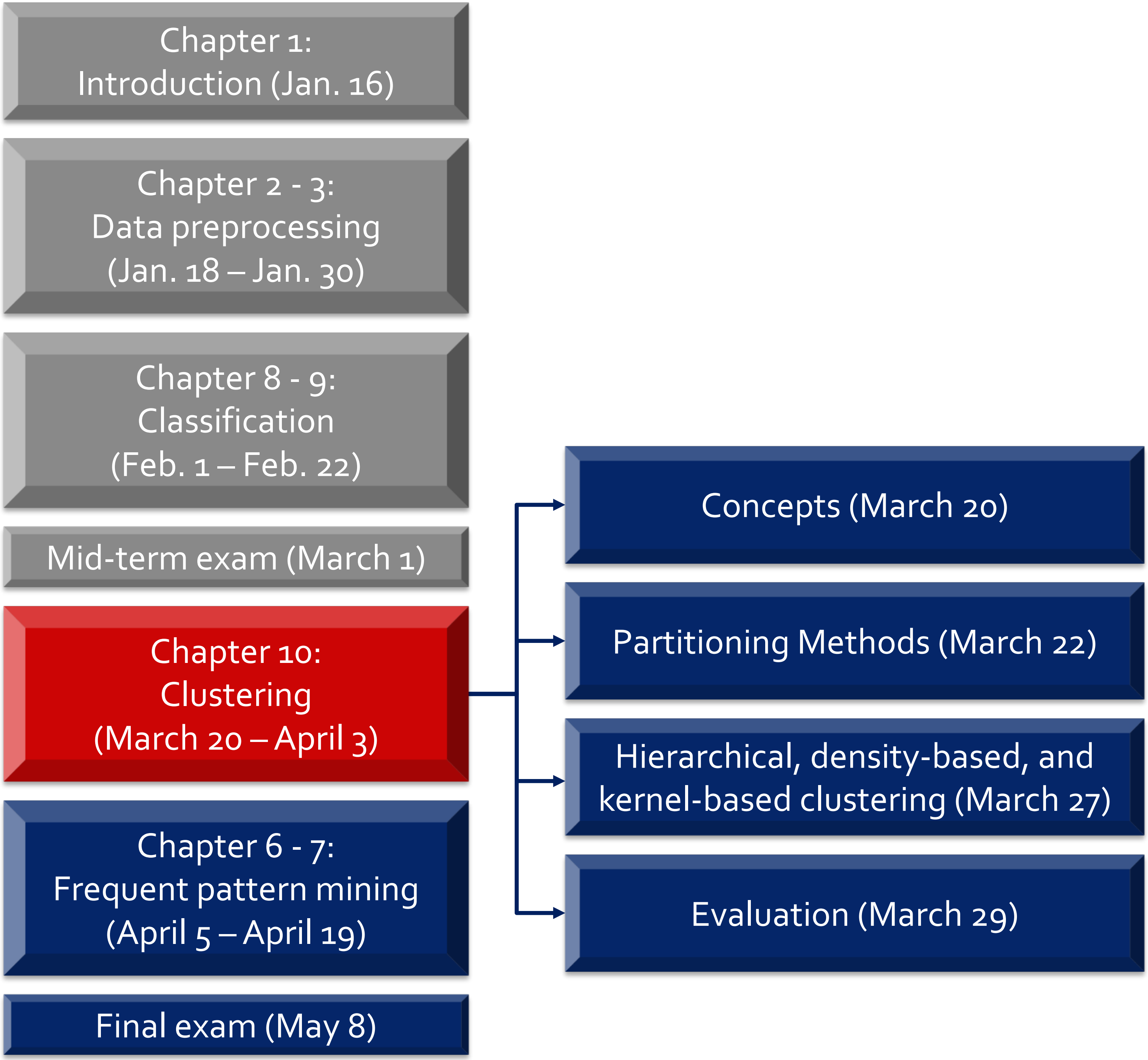
Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)



Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

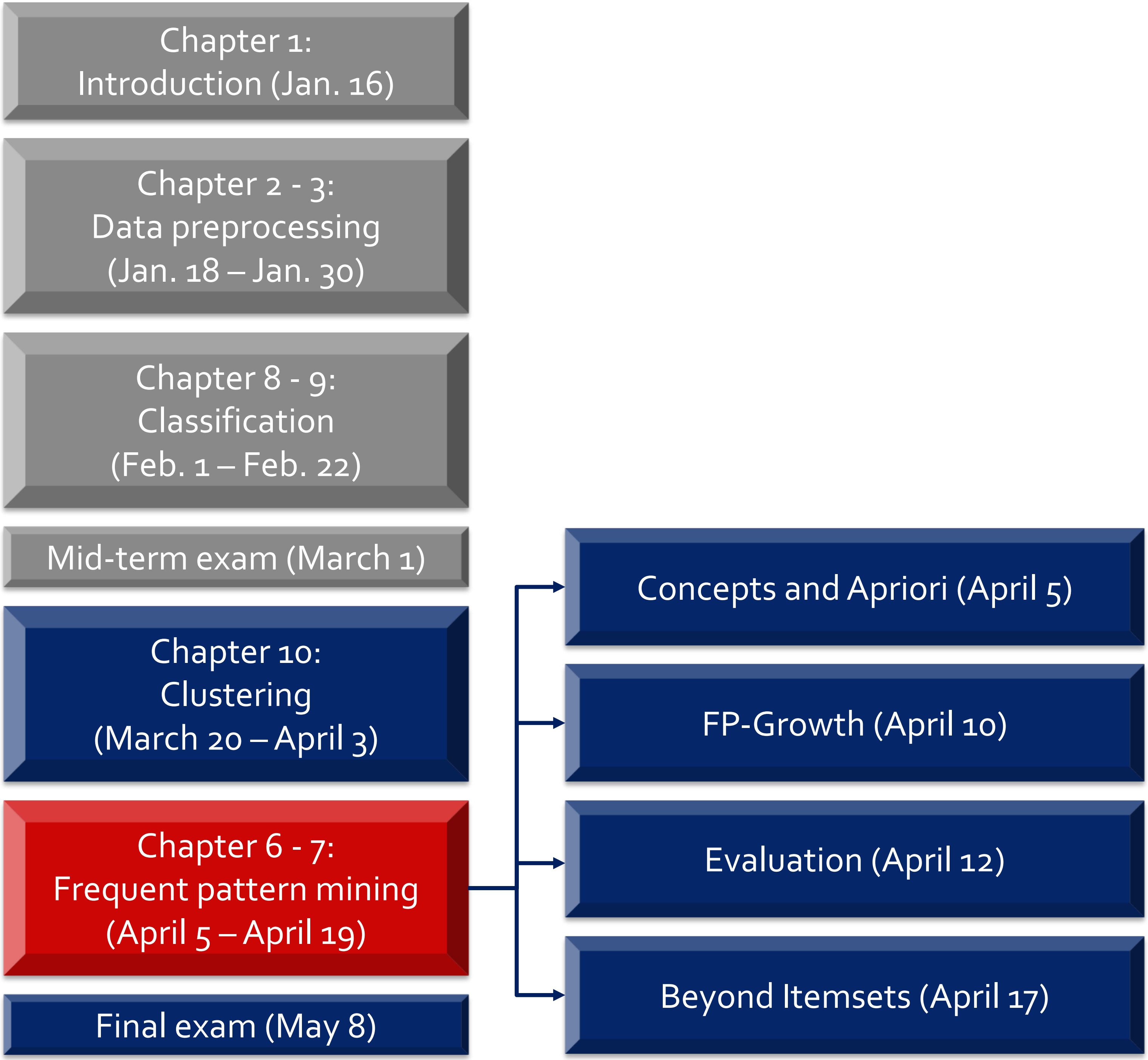
Final exam (May 8)

Concepts (March 20)

Partitioning Methods (March 22)

Hierarchical, density-based, and kernel-based clustering (March 27)

Evaluation (March 29)



Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Concepts and Apriori (April 5)

FP-Growth (April 10)

Evaluation (April 12)

Beyond Itemsets (April 17)

Review

Everybody 100 points if I got in Playoffs…

questions to answer…

Are you ready?

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q1: Who is the instructor of Data Science Spring'18?

A)

B)

C)

D)

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q2:

What is cluster?
What is cluster analysis/clustering?
What is the difference between ***classification*** and ***clustering***?

What are the two **properties** of a good cluster?

List at least three **applications** of cluster analysis.

List at least four types of **data sets** for cluster analysis.

What are the three pairs of clustering **task types** (e.g., partitional vs hierarchical)?

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q3:

What is the **objective function** of K partitioning methods?

What is the **centroid** of a group of data points?

What is the **medoid** of the group?

What is the major difference between ***centroid*** and ***medoid***?

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q4: **K-Means Clustering**

Given *K*, the number of clusters, the *K-Means* clustering algorithm is outlined as follows

```
Select K points as initial centroids
Repeat
    Form K clusters by assigning each data object to its nearest centroid using a distance metric
    Move each centroid to the mean of its assigned data objects (i.e., re-compute the centroid of each cluster)
Until convergence
    Change in cluster assignment less than a threshold
```

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q5: Pros and Cons of K-Means Clustering

Pro:

What is the complexity?

Cons:

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q5: Pros and Cons of K-Means Clustering

Pro:

What is the complexity?

Cons:

Specify K: run a range of values and select the best (min SSE); use rule of thumb or "elbow" method

Local optimum - sensitive to initialization: heuristics to choose initialization, for example, the farthest points

Sensitive to noise and outliers: use K-Medoids or K-Medians

Only applicable for numerical data: use K-Modes for categorical data

Unable to discover clusters with non-convex shapes: use density-based clustering (DBSCAN) or Kernel K-Means

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q6: **Kernel K-Means**

What is the objective function?

What is Kernel Matrix?

List two common kernel functions.

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q6: **Kernel K-Means**

What is the objective function?

$$\underset{\mathcal{J}_1,\ldots,\mathcal{J}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j\in\mathcal{J}_i} \left\| \mathbf{a}_j - \frac{1}{|\mathcal{J}_i|}\sum_{l\in\mathcal{J}_i}\mathbf{a}_l \right\|_2^2 .$$

$$\Downarrow$$

$$\underset{\mathcal{J}_1,\ldots,\mathcal{J}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j\in\mathcal{J}_i} \left\| \phi(\mathbf{a}_j) - \frac{1}{|\mathcal{J}_i|}\sum_{l\in\mathcal{J}_i}\phi(\mathbf{a}_l) \right\|_2^2 .$$

What is Kernel Matrix?

$$\kappa(\mathbf{a}_i,\, \mathbf{a}_j) = \langle \phi(\mathbf{a}_i),\, \phi(\mathbf{a}_j) \rangle .$$

List two common kernel functions.

Polynomial kernel: $K(x_i,x_j) = (x_i^T x_j + c)^d$

RBF kernel: $K(x_i,x_j) = \exp(-\gamma|x_i-x_j|^2)$

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q7: **DBSCAN**

```
Specify $\varepsilon$ and MinPts
Arbitrarily select a point $p$
Retrieve all points density-reachable from $p$
    If $p$ is a core point, a cluster is formed
    If $p$ is a border point, no points are density-reachable from $p$, and DBSCAN visits the next point of the database
Continue until all of the points have been processed
```

What are the Pros and Cons of DBSCAN?

Q7: **DBSCAN**

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

```
Specify $\varepsilon$ and MinPts
Arbitrarily select a point $p$
Retrieve all points density-reachable from $p$
    If $p$ is a core point, a cluster is formed
    If $p$ is a border point, no points are
    density-reachable from $p$, and DBSCAN
    visits the next point of the database
Continue until all of the points have been
processed
```

What are the Pros and Cons of DBSCAN?

Pro: Non-convex shape; partial clustering (outliers not in clusters); not have to specify K; O(n logn)

Con: Sensitive to the two parameters

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q8: **External Evaluation for Clustering**

Matching-based

- Purity (matching)
- Purity (maximum matching)
- Matching-based Precision, Recall, F1

Pairwise

- Confusion matrix (pairwise TP/FN/FP/TN)
- Jaccard coefficient
- Rand Statistic
- Pairwise precision, recall, and
- Fowlkes-Mallow Measure

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q9: **BetaCV Internal Evaluation for Clustering**

$$BetaCV = \frac{W_{in} / N_{in}}{W_{out} / N_{out}}$$

- The smaller, the better the clustering, when the weight is distance
- The bigger, the better the clustering, when the weight is similarity

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q10: Concepts

What is *k*-itemset?
What is *absolute support*?
What is *relative support*?
What is minimum support *min_sup*? And what is frequent itemset?

For an association rule X→Y,
what is *support*? Is it relative or absolute?
What is *confidence*?
Think about Y→X,
is *support* symmetric? Is *confidence* symmetric?

What is *closed pattern*? Is it lossless? (What does "lossless" mean?

What is *max pattern*? Is it lossless?

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

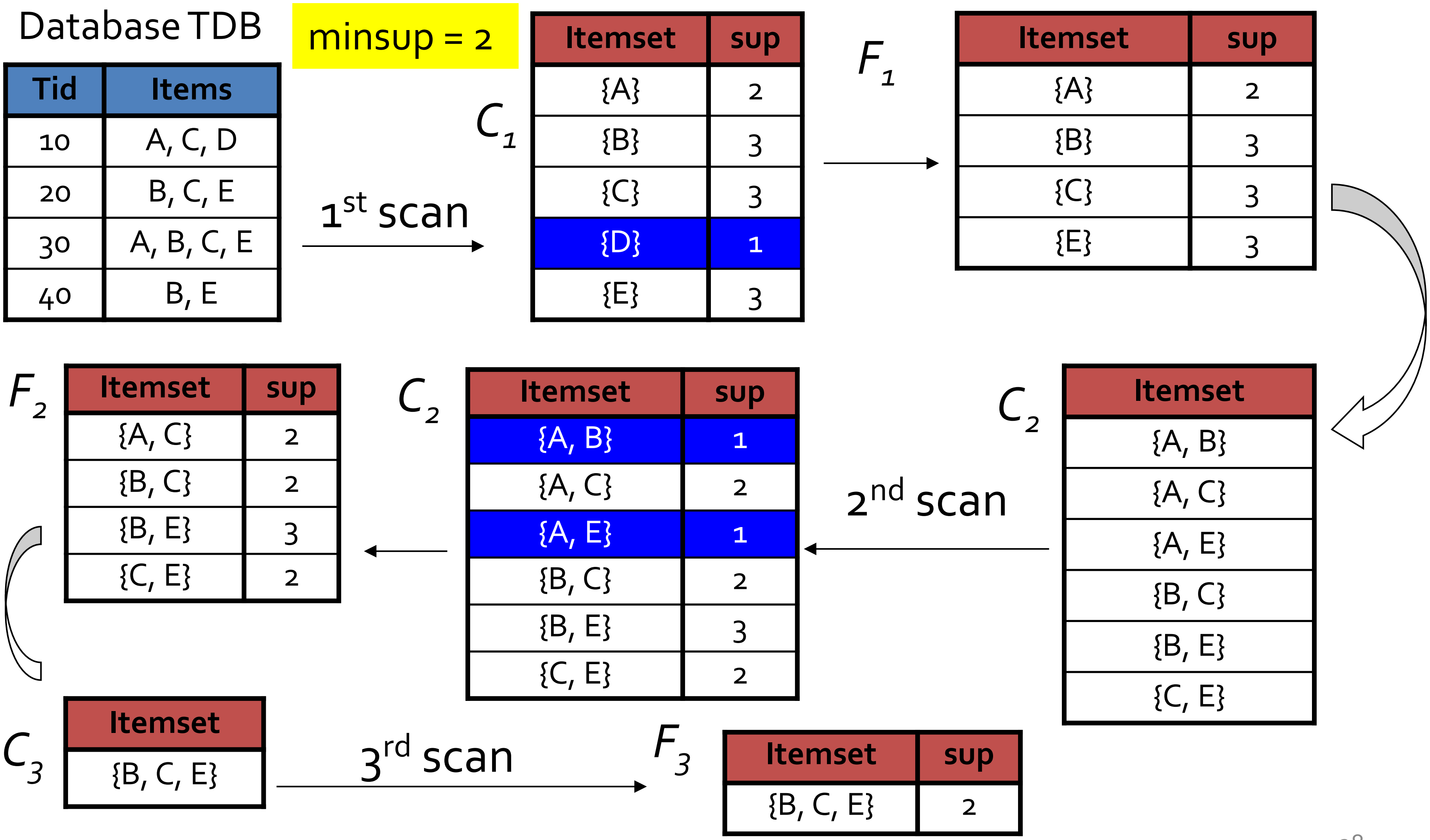
Q11: **Apriori**

What is Apriori property (or called the Downward Closure Property)?

Outline of Apriori (level-wise, candidate generation and test)

```
    Initially, scan DB once to get frequent 1-itemset
    Repeat
        Generate length-(k+1) candidate itemsets from length-k frequent itemsets
        Test the candidates against DB to find frequent (k+1)-itemsets
        Set k := k +1
    Until no frequent or candidate set can be generated
    Return all the frequent itemsets derived
```

Apriori



Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

minsup = 2

1st scan

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

F_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

F_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan

F_3

Itemset	sup
{B, C, E}	2

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q12: Discussion on Apriori

What is the biggest weak point of the Apriori algorithm? Is it efficient?

Outline of Apriori (level-wise, candidate generation and test)

```
    Initially, scan DB once to get frequent 1-itemset
    Repeat
        Generate length-(k+1) candidate itemsets from length-k frequent itemsets
        Test the candidates against DB to find frequent (k+1)-itemsets
        Set k := k +1
    Until no frequent or candidate set can be generated
    Return all the frequent itemsets derived
```

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

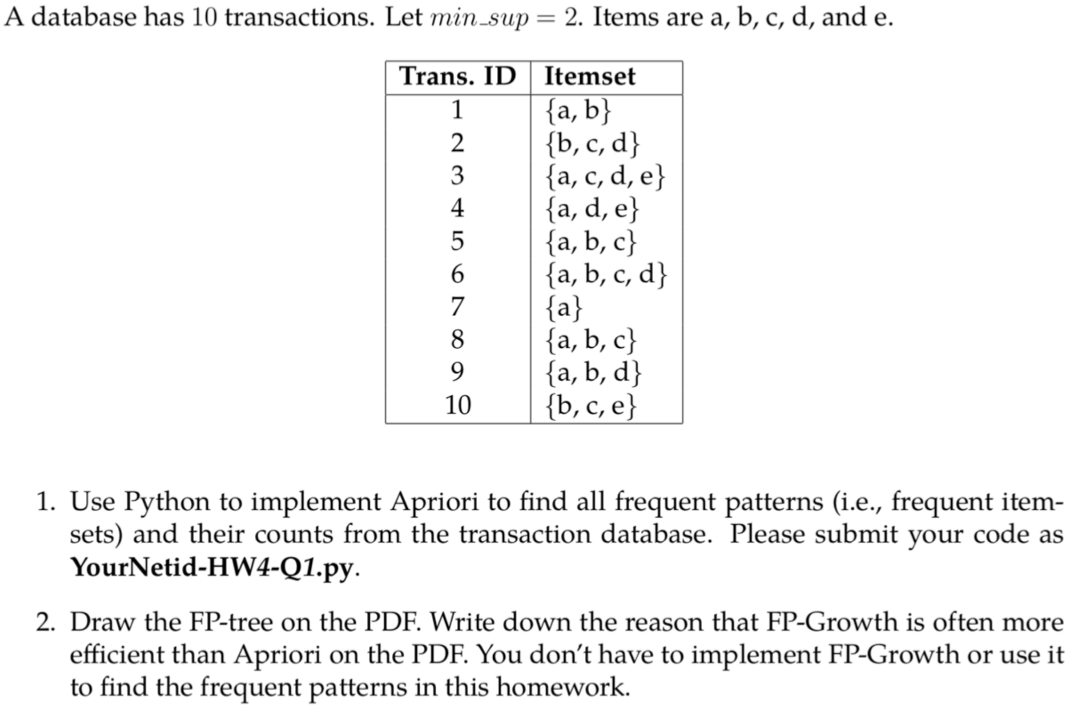
Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q13: **FP-Growth**

- Find frequent single items and partition the database based on each such item
- Recursively grow frequent patterns by doing the above for each partitioned database (also called *conditional database*)
- To facilitate efficient processing, an efficient data structure, FP-tree, can be constructed



A database has 10 transactions. Let $min\_sup = 2$. Items are a, b, c, d, and e.

Trans. ID	Itemset
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

1. Use Python to implement Apriori to find all frequent patterns (i.e., frequent itemsets) and their counts from the transaction database. Please submit your code as **YourNetid-HW4-Q1.py**.
2. Draw the FP-tree on the PDF. Write down the reason that FP-Growth is often more efficient than Apriori on the PDF. You don't have to implement FP-Growth or use it to find the frequent patterns in this homework.

Find frequent patterns and closed patterns

Trans. ID	Items bought
1	ACFG
2	ABCF
3	ABCDF
4	BDE

If min_sup = 2, are they closed patterns?

- D
- ABCF
- BF
- BD
- ACF

Use Apriori to find all frequent patterns

Use FP-Growth to find all frequent patterns

Write down all closed patterns and their support

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

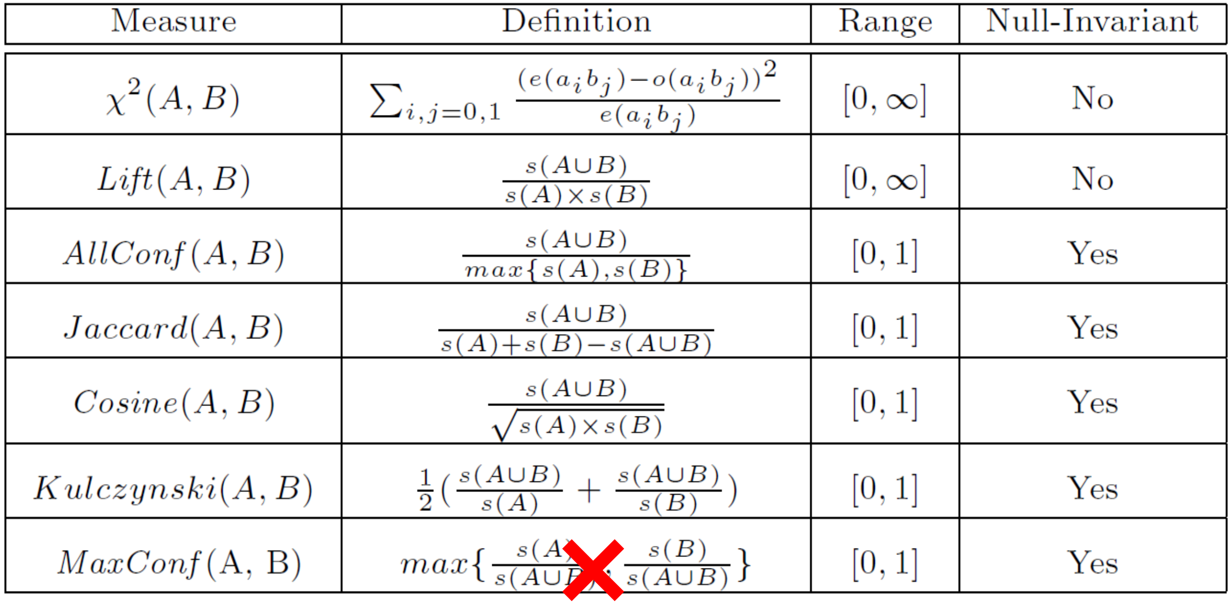
Final exam (May 8)

Q14: Association interestingness measures

What is null-invariance?

Give two null-variant measures.

Give five null-invariant measures and prove this property.



Measure	Definition	Range	Null-Invariant
$\chi^2(A, B)$	$\sum_{i,j=0,1} \frac{(e(a_i b_j) - o(a_i b_j))^2}{e(a_i b_j)}$	$[0, \infty]$	No
$Lift(A, B)$	$\frac{s(A \cup B)}{s(A) \times s(B)}$	$[0, \infty]$	No
$AllConf(A, B)$	$\frac{s(A \cup B)}{max\{s(A), s(B)\}}$	$[0, 1]$	Yes
$Jaccard(A, B)$	$\frac{s(A \cup B)}{s(A) + s(B) - s(A \cup B)}$	$[0, 1]$	Yes
$Cosine(A, B)$	$\frac{s(A \cup B)}{\sqrt{s(A) \times s(B)}}$	$[0, 1]$	Yes
$Kulczynski(A, B)$	$\frac{1}{2}(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)})$	$[0, 1]$	Yes
$MaxConf(A, B)$	$max\{\cancel{\frac{s(A)}{s(A \cup B)}, \frac{s(B)}{s(A \cup B)}}\}$	$[0, 1]$	Yes

max{ s(AUB) / s(A) , s(AUB) / s(B) }

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Q15: Sequential patterns

What is item, event, and sequence?
What is sequential pattern?

Seq. ID	Sequence
1	(AB)C(FG)G
2	(AD)CG(ABF)
3	AB(FG)

If min_sup = 2, are they sequential patterns?

- ACF
- (FG)B
- (FG)
- B(FG)
- GF

KEEP
CALM
AND
GOOD
LUCK!