精簡型樣探勘與軟體操作

國立宜蘭大學資訊工程系
吳政瑋 助理教授
wucw@niu.edu.tw

本堂教學重點

- 傳統頻繁項目集探勘缺點
- 精簡型樣探勘目的
- 精簡型樣類型
- 精簡型樣探勘技術
- 精簡型樣探勘軟體操作

傳統頻繁項目集探勘缺點 (Cont. 1/5)

- 傳統的頻繁項目集探勘演算法，如：Apriori [1]、DHP [2]、H-mine [3]、FP-Growth [4]、Eclat [5]，從**稀疏型資料集(Sparse Dataset)**中挖掘頻繁項目集(Frequent Itemset；FI)時，往往會有不錯的執行效率。

傳統頻繁項目集探勘缺點 (Cont. 2/5)

- **稀疏型資料集(Sparse Dataset)**
  - 項目之間的關聯性較低，且交易長度較短。
  - 由於此特性，使得傳統 FI Mining 演算法得以有效利用**向下封閉性質(Downward Closure Property)**縮減搜尋空間。
  - 從Sparse Dataset中找出的頻繁項目集通常**個數較少**且**長度較短**。

傳統頻繁項目集探勘缺點 (Cont. 3/5)

- **密集型資料集(Dense Dataset)**
  - 項目之間的關聯性較強，且交易長度較長。
  - 由於此特性，傳統 FI Mining 演算法無法有效利用 **Downward Closure Property** 縮減搜尋空間。
  - 從Dense Dataset中找出的 FI 通常**個數較多**且**長度較長**。

傳統頻繁項目集探勘缺點 (Cont. 3/5)

- 先前研究普遍認為交易資料、網頁瀏覽紀錄屬於Spare Dataset。
- 然而，在現實生活中有許多資料屬於Dense Dataset，如：蘑菇特性統計資料集(Mushrooms)、西洋棋下棋紀錄(Chess)等。

傳統頻繁項目集探勘缺點 (Cont. 5/5)

- 傳統頻繁項目集探勘的一項缺點是**產生過多頻繁項目集**
  - 探勘速度緩慢
  - 記憶體使用量過高
  - 太多項目集不易應用與理解

精簡型樣探勘

- 精簡型樣探勘
  - 藉由僅探勘**具代表性型樣(Representative Patterns)**，以提高**探勘效率**

- 由於所需挖掘的型樣個數減少，帶來下列好處
  - 探勘速度較快
  - 記憶體使用量較少
  - 型樣少則容易應用與理解

常見的精簡型樣類型

- 精簡型樣又稱為型樣的**精簡表示法(Concise Representation、Condensed Representation)**

- 頻繁項目集常見的精簡表示法包括：
  - 最大頻繁項目集(Maximal Frequent Itemset；MFI)
  - 頻繁封閉項目集(Frequent Closed Itemset；FCI)
  - Top-*k* 頻繁項目集(Top-*k* Frequent Itemset)

最大頻繁項目集 (Maximal Frequent Itemset；MFI)

- 若一個頻繁項目集 X **不存在**任何 Proper Superset Y 也是頻繁項目集，則稱 X 為**頻繁最大項目集**。

- 舉例說明
  - 若$\theta = 3$ ，{ABC}：4 為 FI，且它不存在任何Proper Superset 也是 FI，則{ABC}為 MFI。

最大頻繁項目集的特性

- 若 *X* 為 MFI，則它所有 Subset 皆為 FI。

- 舉例說明
  - 若{ABC}為MFI，則它所有 Subset 皆為 FI， 如：{A}、{B}、{C}、{AB}、{AC}、{BC}、{ABC} 皆為 FI。

常見的最大頻繁項目集探勘演算法

- Conventional FI Mining + Post Process
- Max-Miner [5]
- GenMax [6]
- Mafia [7]
- FPMax [8]

最大頻繁項目集探勘缺點

- 最大頻繁項目集探勘缺點
  - 僅探勘 MFI 雖然不會遺失所有的FI，但是會**遺失所有 FI 的支持數**。
  - 找出所有 MFI 後，若要取得所有FI之支持數，可先還原出所有 FI，再掃描原始資料庫一次，計算每個 FI 的支持數。

頻繁封閉項目集 (Frequent Closed Itemset；FCI)

- 若一個頻繁項目集 X **不存在**任何 Proper Superset Y，使得$SC(X) = SC(Y)$，則稱 X 為**頻繁封閉項目集**。

- 舉例說明
  - 若θ=3，{ABC}：4 為 FI，且它不存在任何 Proper Superset 與其支持數相同，則{ABC} 為 FCI。
  - 承上，{AB}：4 為 FI，但 $SC(\{AB\}) = SC(\{ABC\})$，故{AB} 不為 FCI。

動腦時間 1 (3 mins)

- 請問下列何者為頻繁封閉項目集？

{ABC}:4

{AB}:5　　{AC}:4　　{BC}:4

{A}:7　　{B}:5　　{C}:4　　{D}:4

動腦時間 1 (解說)

- 頻繁封閉項目集以黃底顯示

{ABC}:4

{AB}:5　　{AC}:　　{BC}:

{A}:7　　{B}:　　{C}:　　{D}:4

頻繁封閉項目集的特性

- FCI 是 FI 的**精簡不失真表示法(Concise and Lossless Representation)**，令 C 為 FCI 的完整集合，F為FI的完整集合，則
  - $C \subseteq F$，C為F的子集合
  - $|C| \leq |F|$，FCI 的個數通常小於 FI 的個數
  - 僅探勘 FCI 不會遺失任何 FI 及其支持數的資訊

Closure Operation

- **Closure Operation**
  - *tidset*({A}) = {#1, #3, #4, #5}
  - *Closure*({A}) = {ACTW} ∩ {ACTW} ∩ {ACDW} ∩ {ACDTW} = {ACW}

Database	
TID	Transactions
#1	ACTW
#2	CDW
#3	ACTW
#4	ACDW
#5	ACDTW
#6	CDT

Closure Operation

- **Closed Itemset**
  - *Closure*(*X*) = *X*, *X* is called ***closed itemset*** ;
    Otherwise, *X* is called ***non-closed itemset***
  - *SC*(*X*) = *SC*(*Closure*(*X*))

- **Frequent Closed Itemset**
  - *X* is a FI
  - *C* is *closed* (*X* is a *closed itemset*)

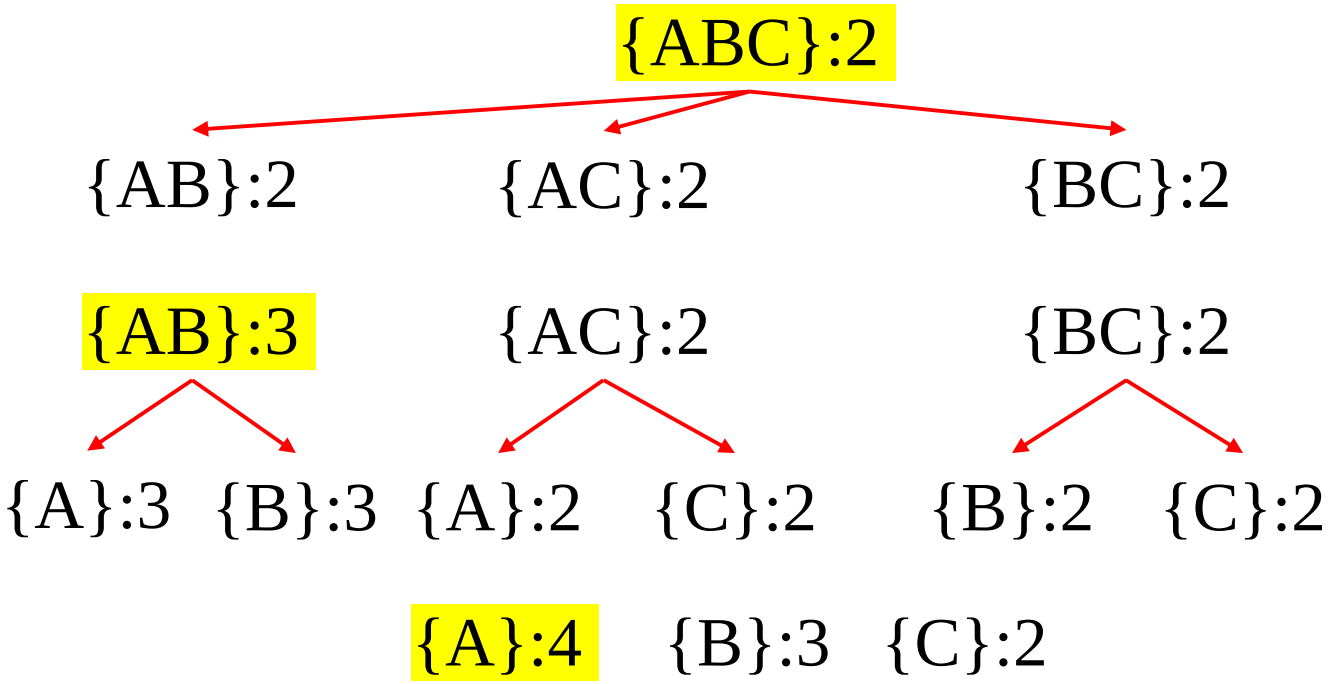
Database	
TID	Transactions
#1	ACTW
#2	CDW
#3	ACTW
#4	ACDW
#5	ACDTW
#6	CDT

從 FCI 還原出 FI 完整資訊的演算法

- 從 FCI 還原出 FI 完整資訊的演算法
  - LevelWise [9]
  - DFI-Growth [10]
  - DFI-List [11]

舉例說明
LevelWise 演算法

FCI	SC
{ABC}	2
{AB}	3
{A}	4



FI 的完整集合 F = {{A} : 4 , {AB} : 3 , {ABC} : 2 , {AC} : 3
{B} : 3 ,{BC}: 2, {C}: 2}

常見的頻繁封閉項目集探勘演算法

- Conventional FI Mining + Post Process
- A-Close [12]
- Charm [13]
- DCI_Close [14]
- FPClose [15]

Charm 演算法簡介

- Charm 為 **C**losed **A**ssociation **R**ule **M**ining的縮寫，其中 H 為虛字。
- 其採用**垂直式資料庫(Vertical Database)**。
- 其改良**Eclat演算法[5]**，使其可用 FCI Mining。
- 其採用的技巧包含：深度優先搜尋(Depth First Search)，分而治之，遞迴、雜湊。

垂直式資料表示法 (Vertical Data Representation)

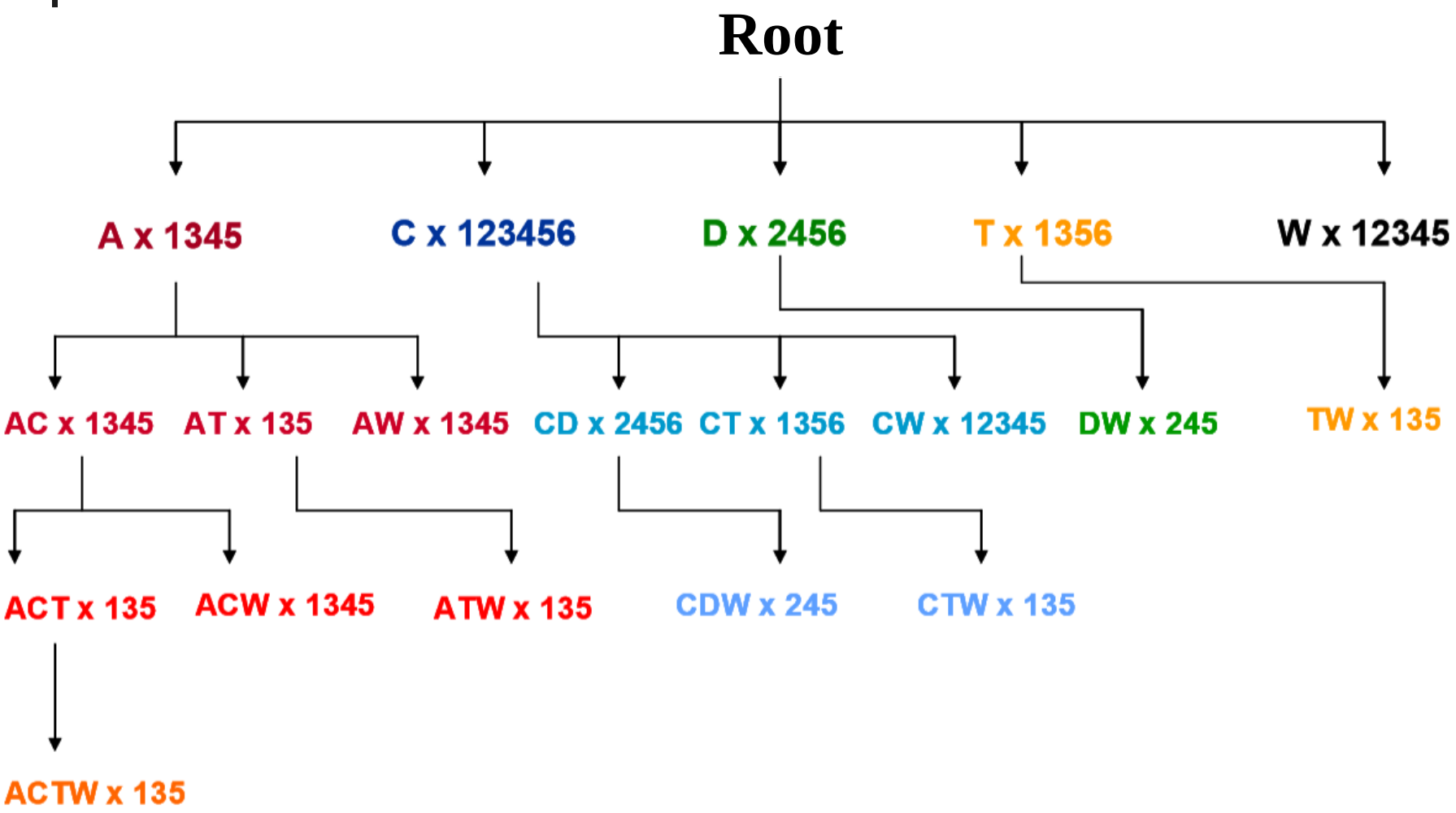
Horizontal Database

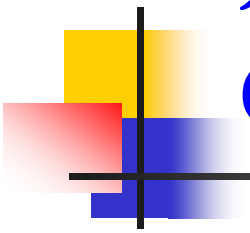
Tid	Item
T_1	ABCE
T_2	BDE
T_3	ACE
T_4	ADE

Vertical Database

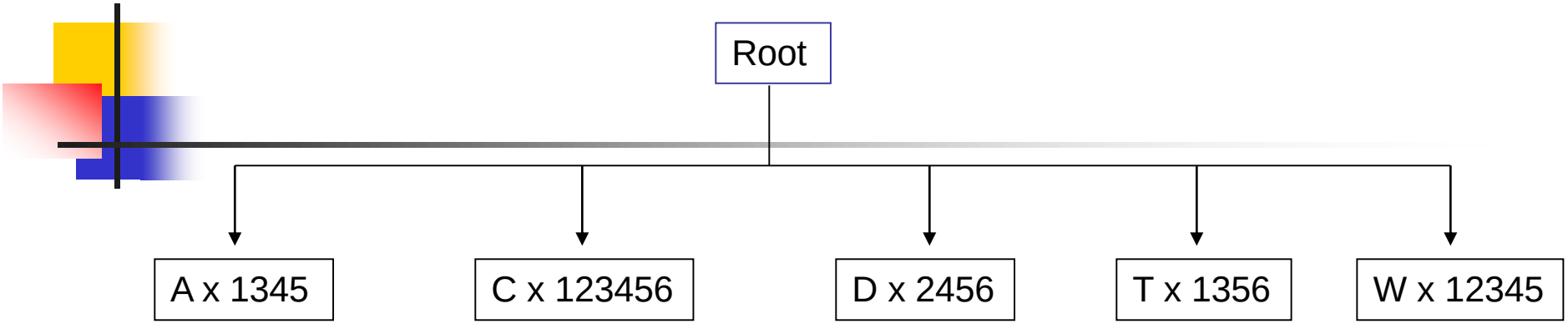
Item	Tidset
A	1 3 4
B	1 2
C	1 3
D	2 4
E	1 2 3 4

舉例說明
Eclat 演算法[11]

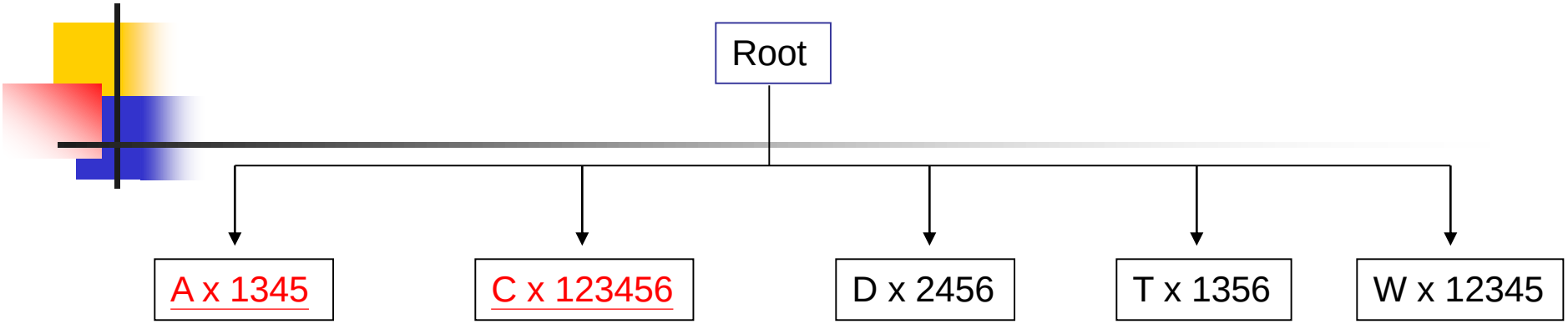




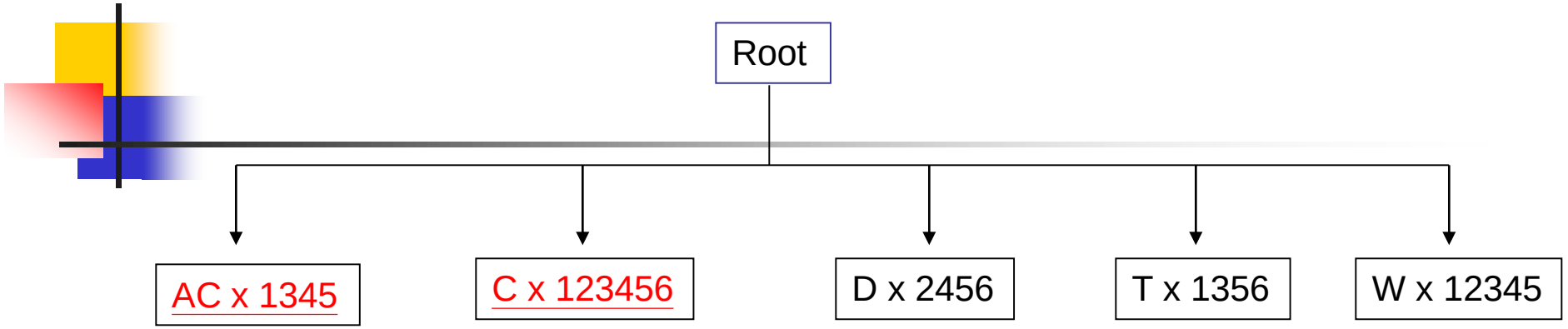
舉例說明
Charm 演算法[8]



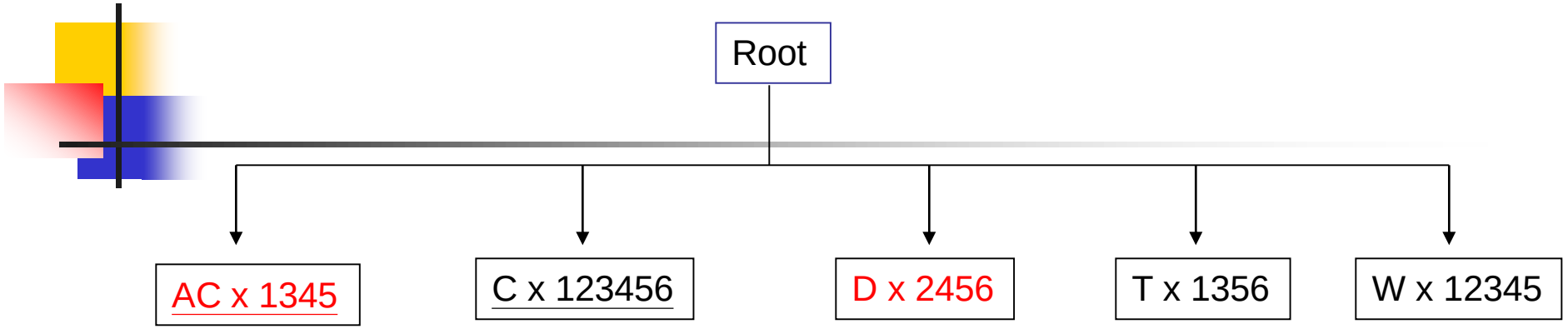
Key	FCI	N



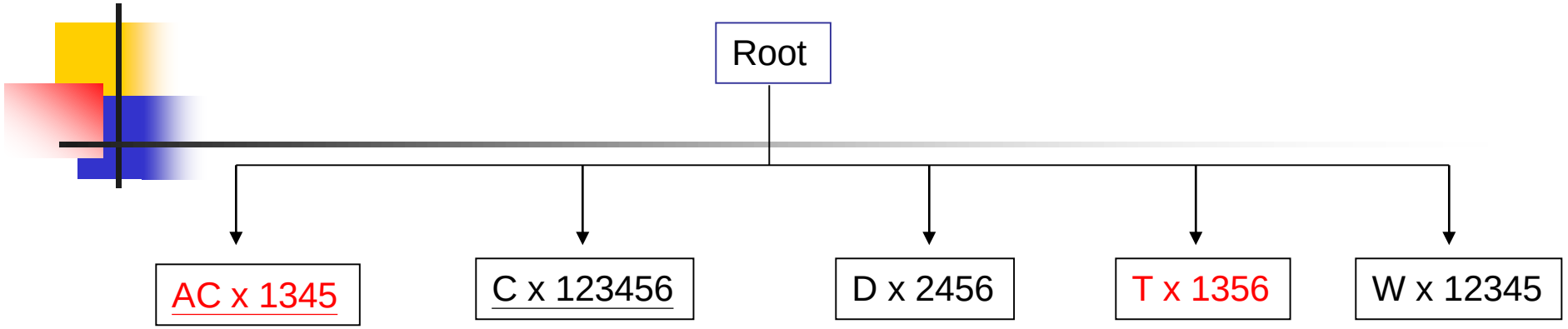
Key	FCI	N



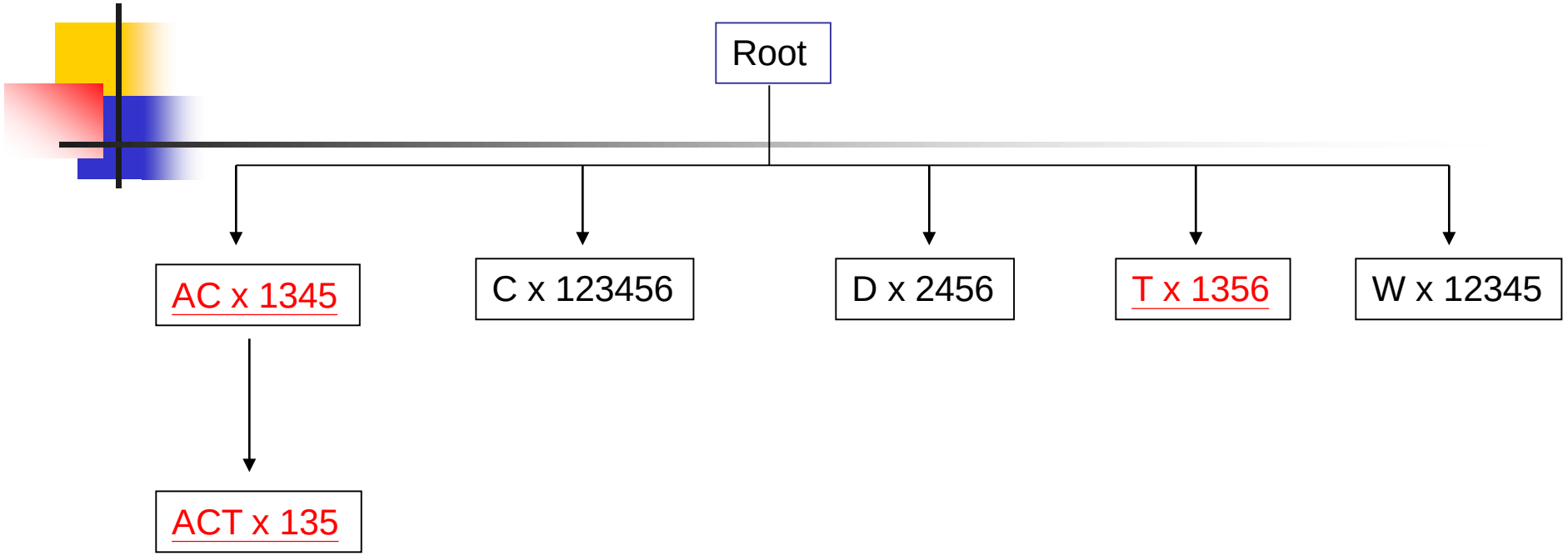
Key	FCI	N



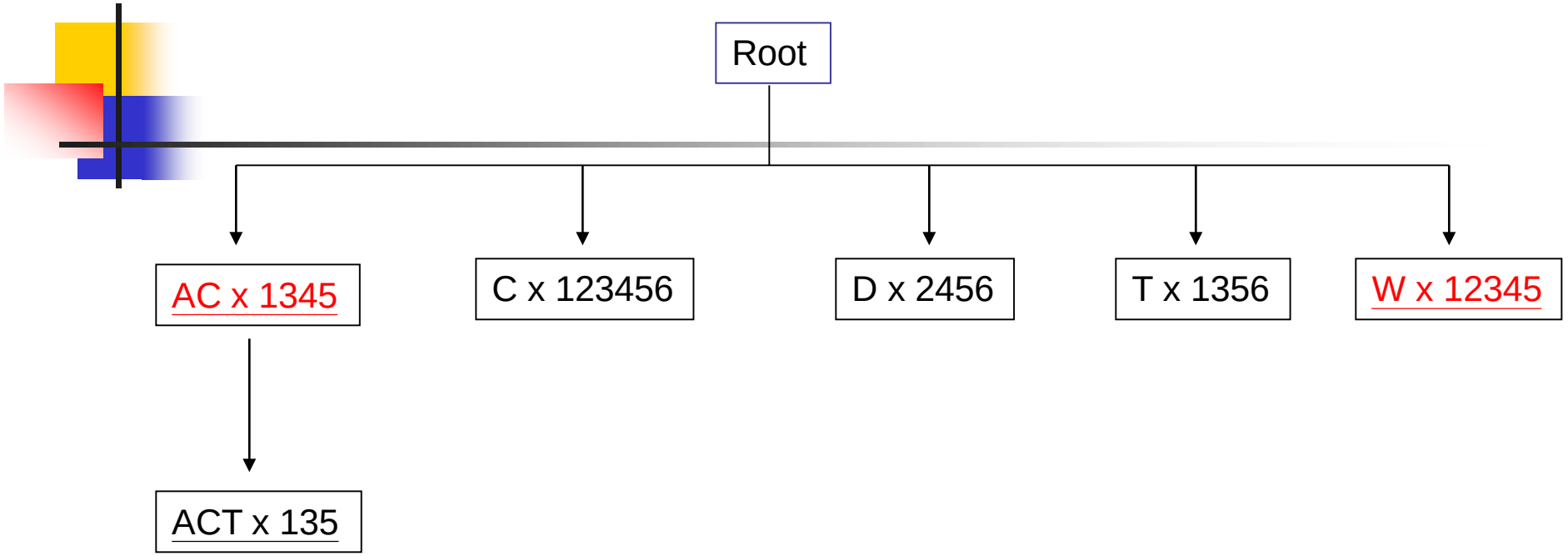
Key	FCI	N



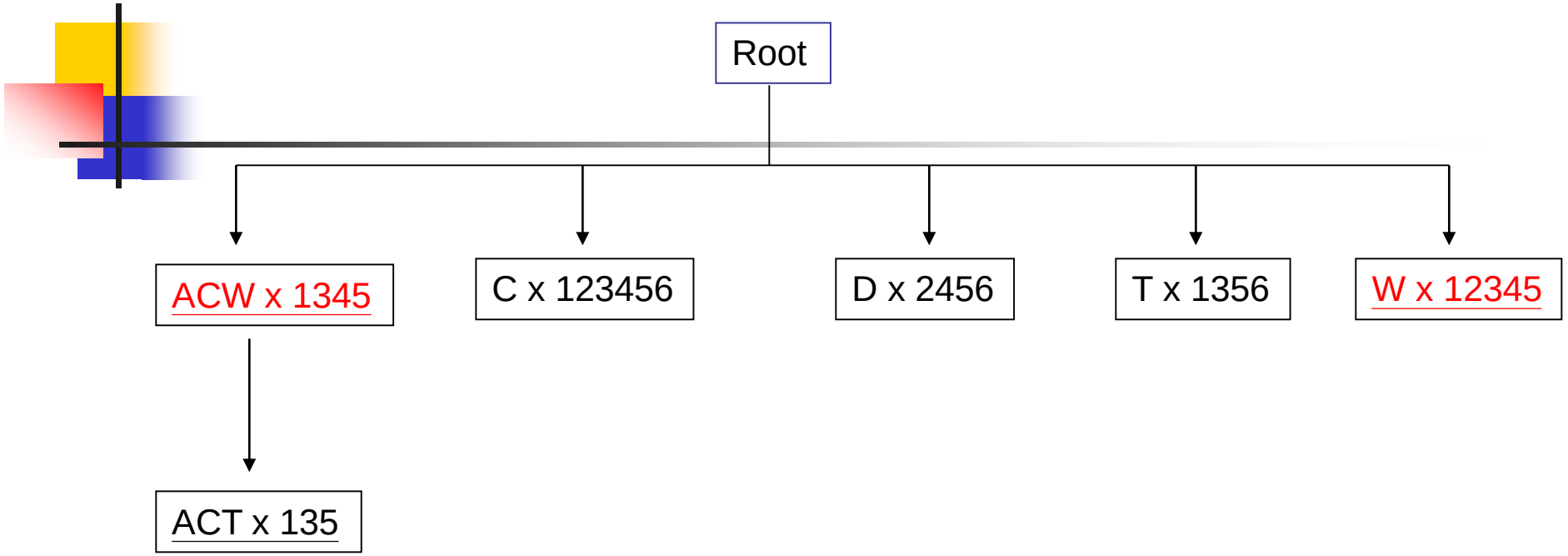
Key	FCI	N



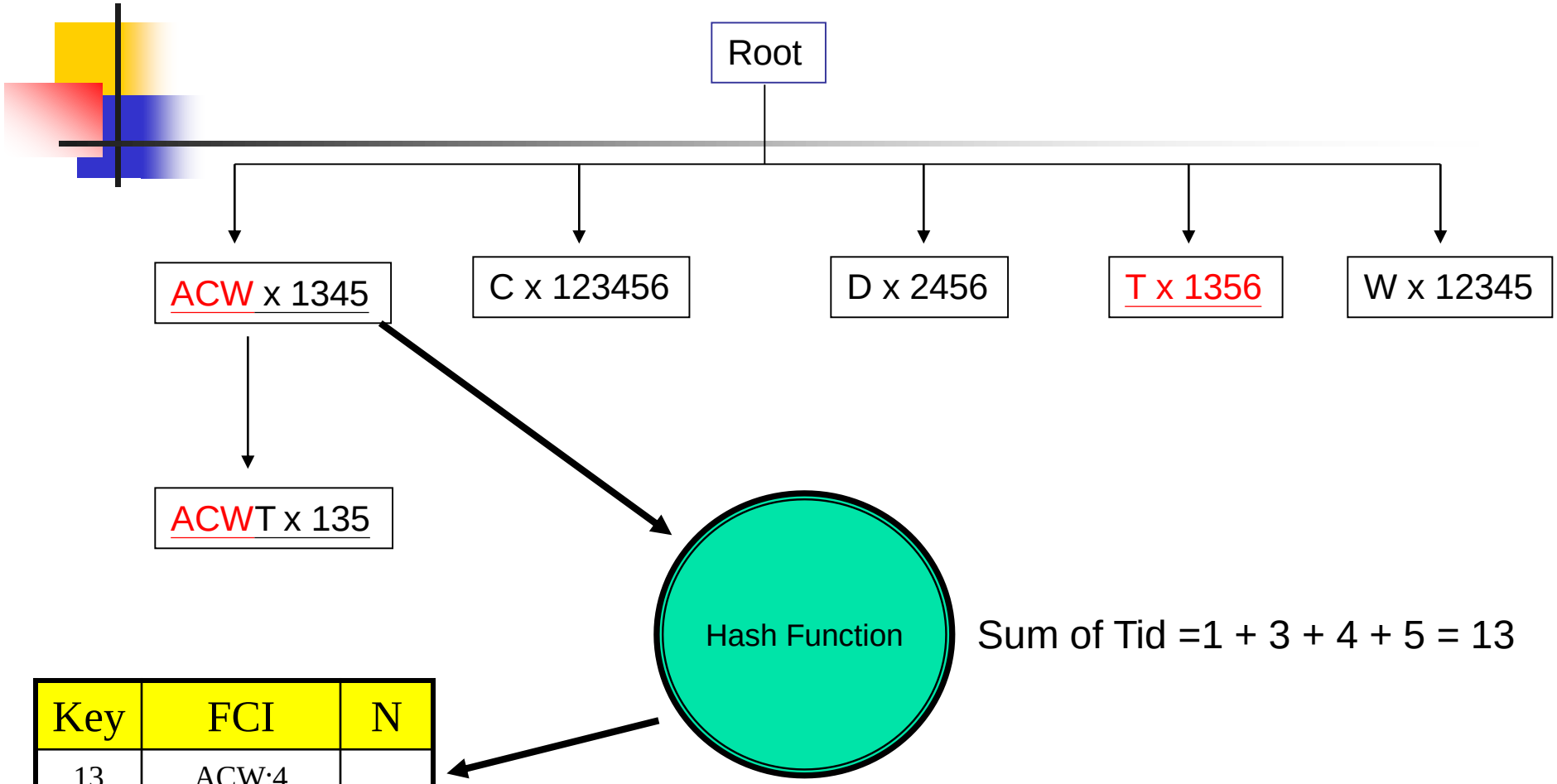
Key	FCI	N



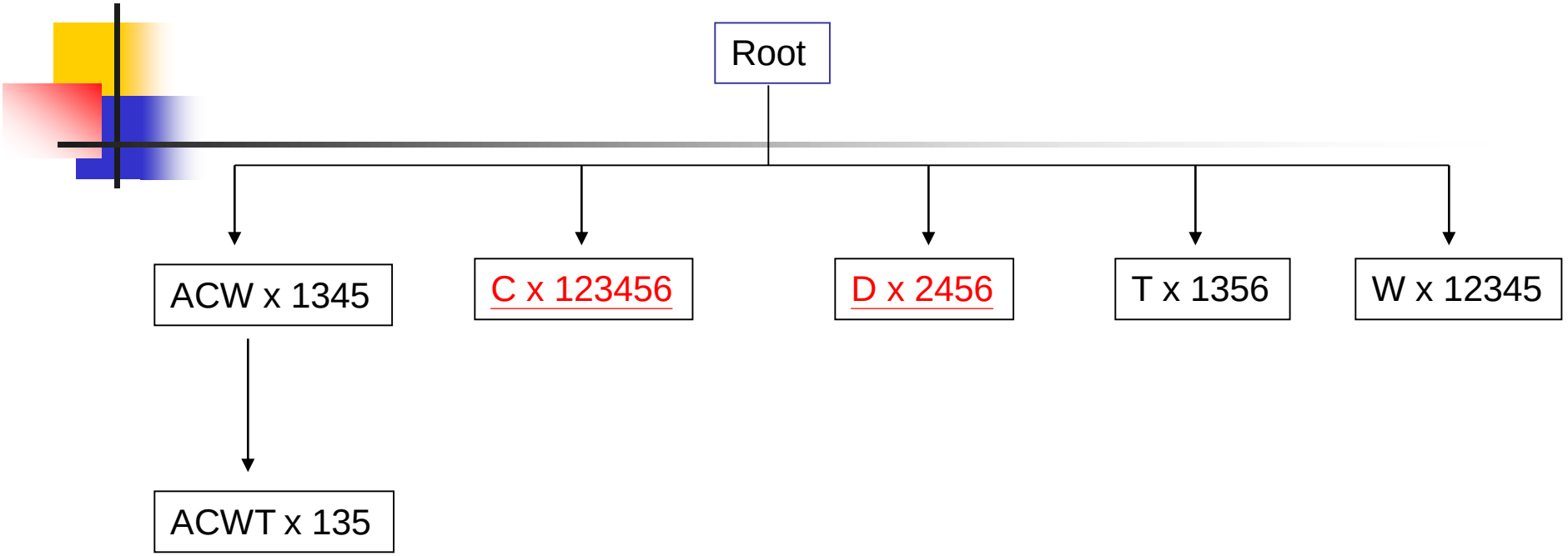
Key	FCI	N



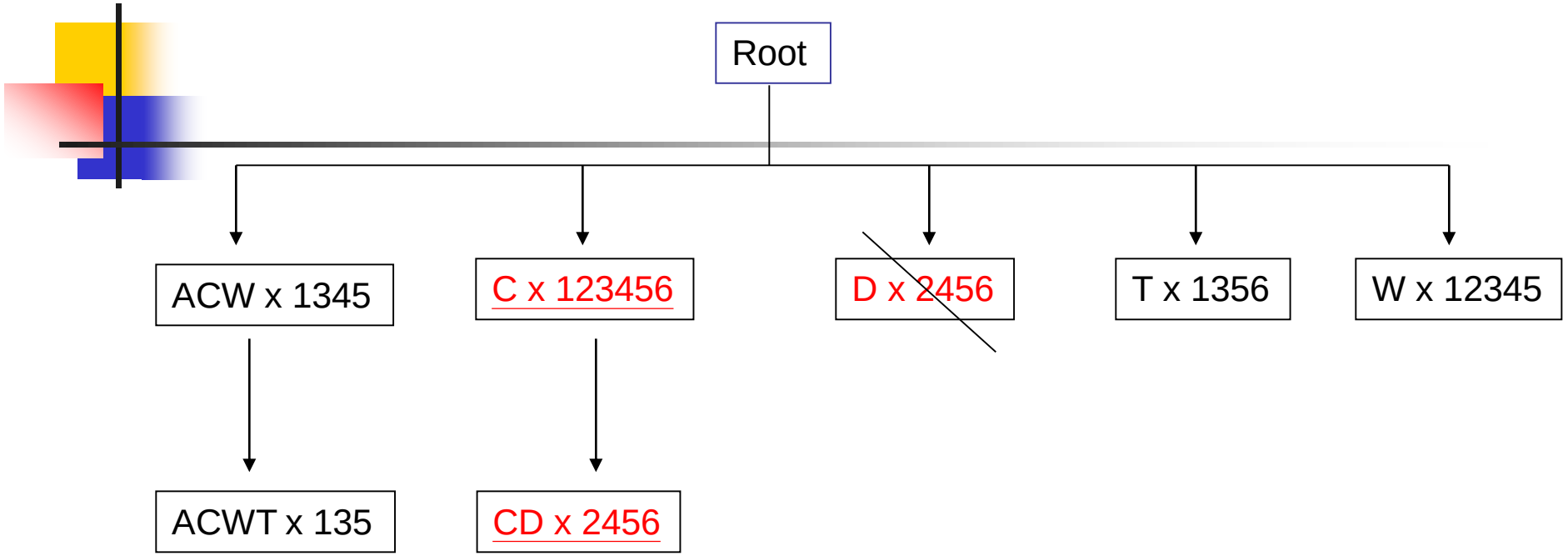
Key	FCI	N



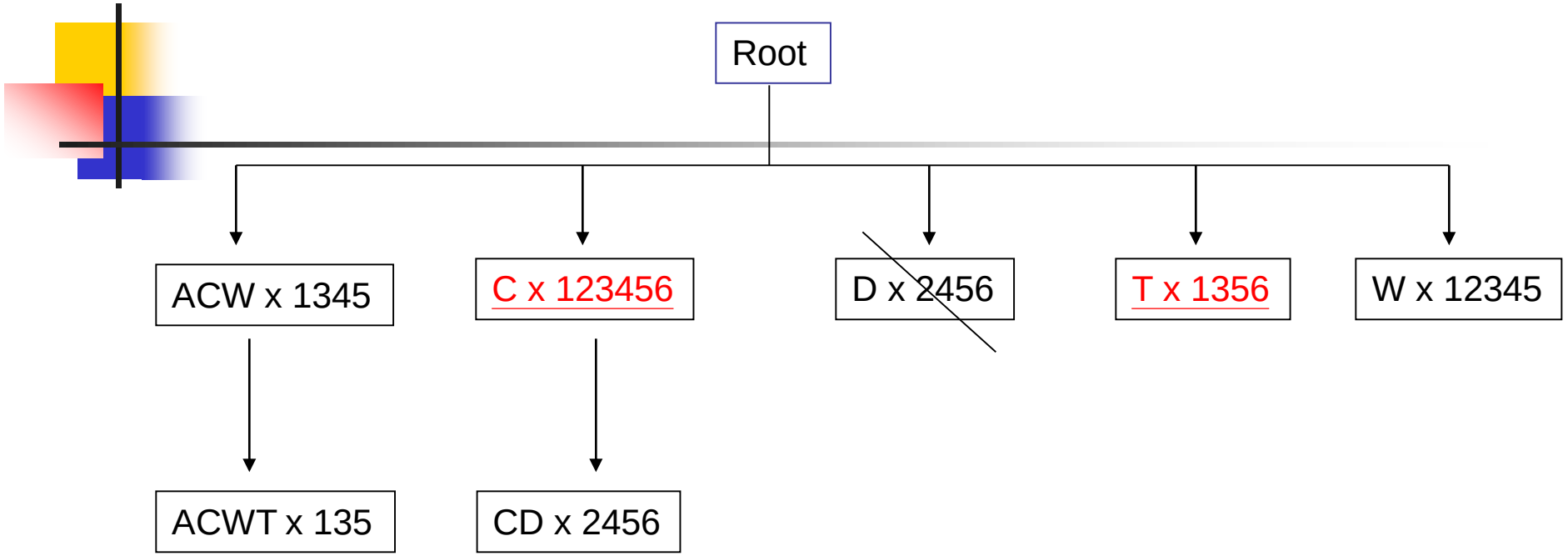
Key	FCI	N
13	ACW:4	
9	ACWT:3	



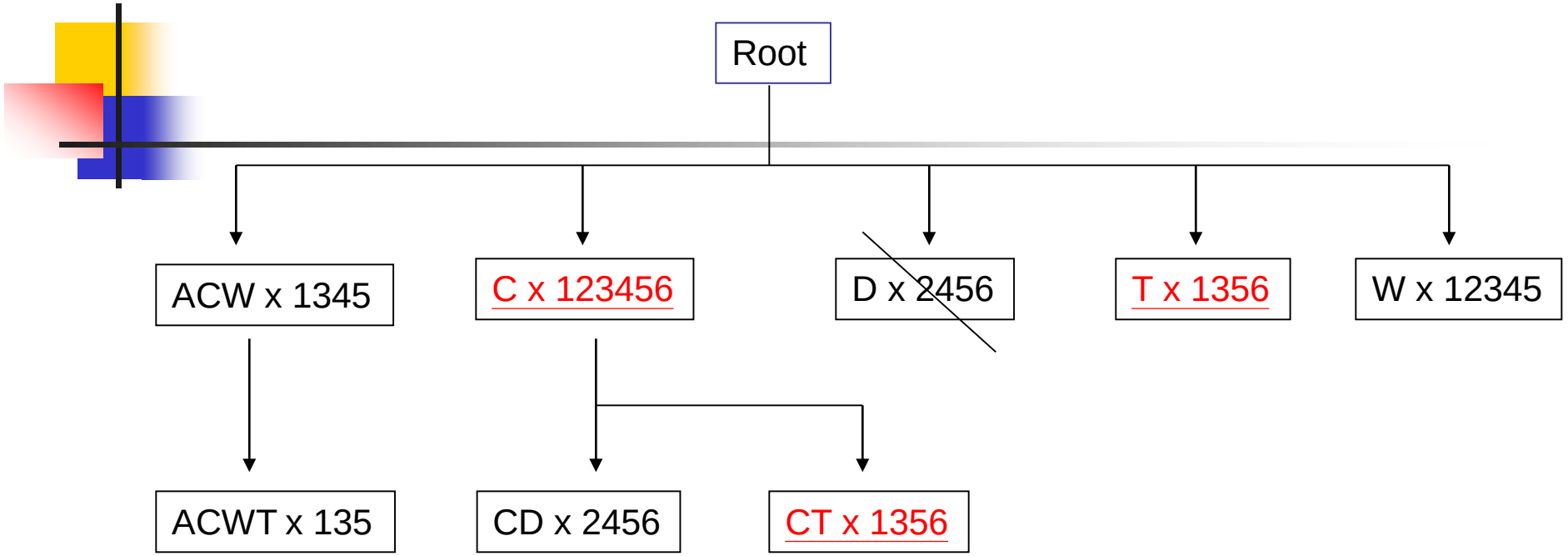
Key	FCI	N
13	ACW:4	
9	ACWT:3	



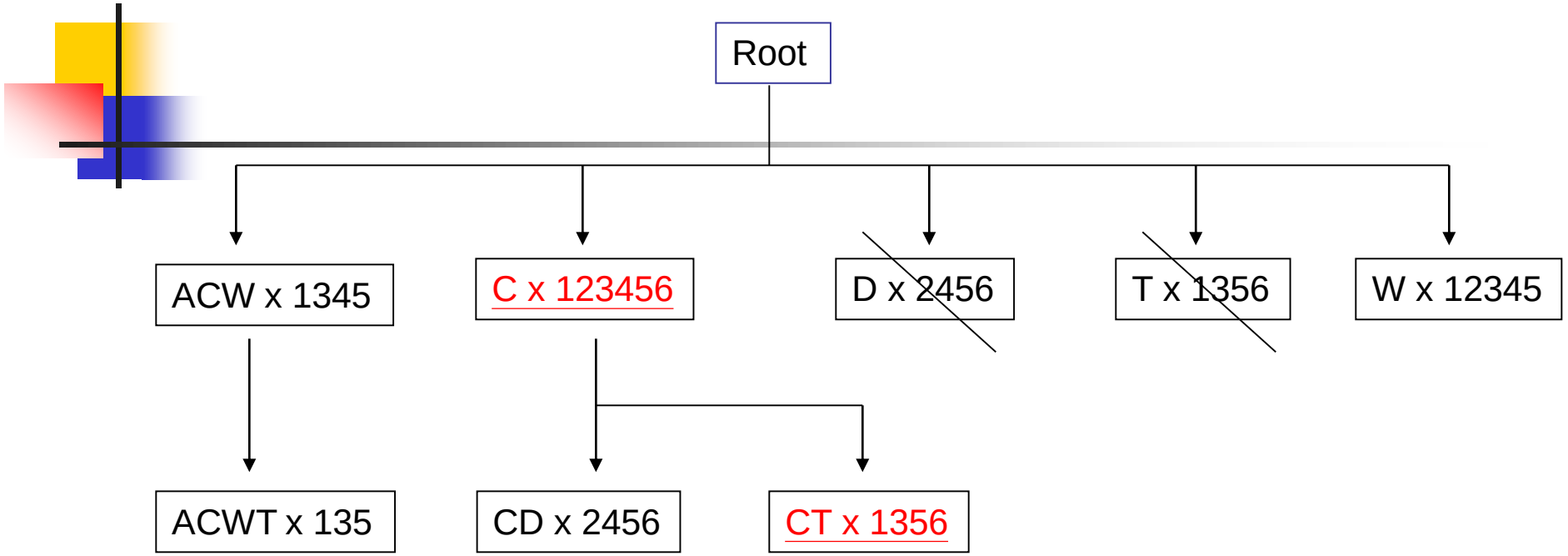
Key	FCI	N
13	ACW:4	
9	ACWT:3	



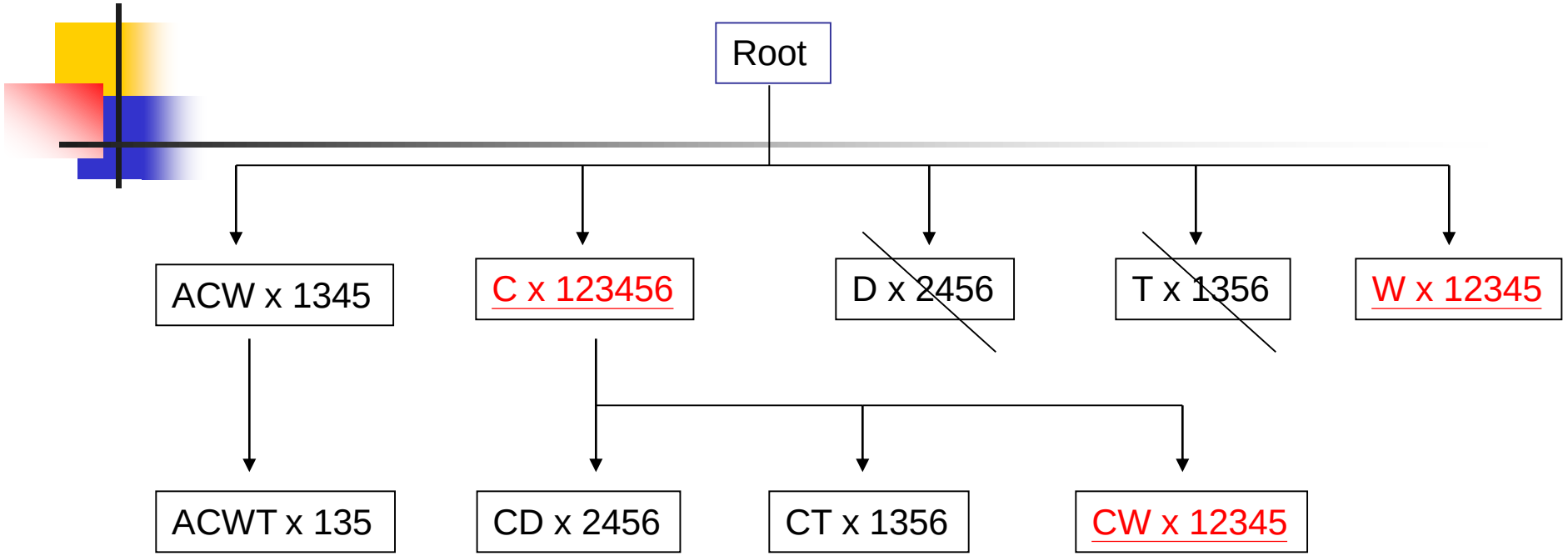
Key	FCI	N
13	ACW:4	
9	ACWT:3	



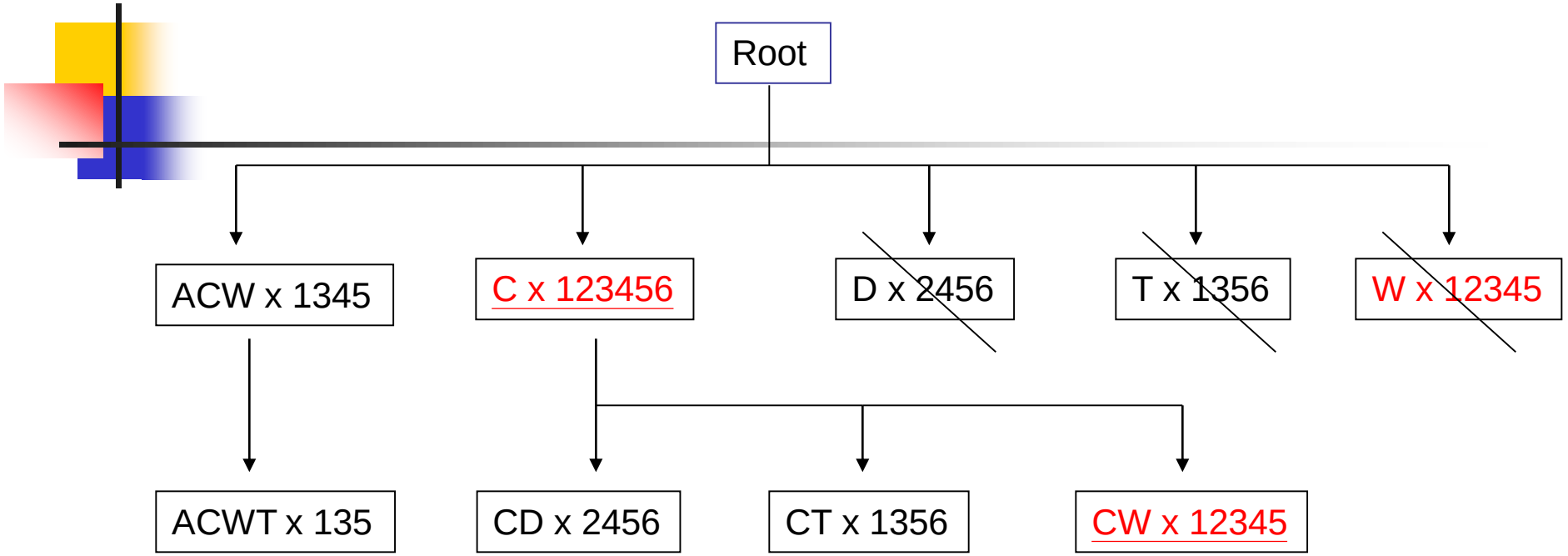
Key	FCI	N
13	ACW:4	
9	ACWT:3	



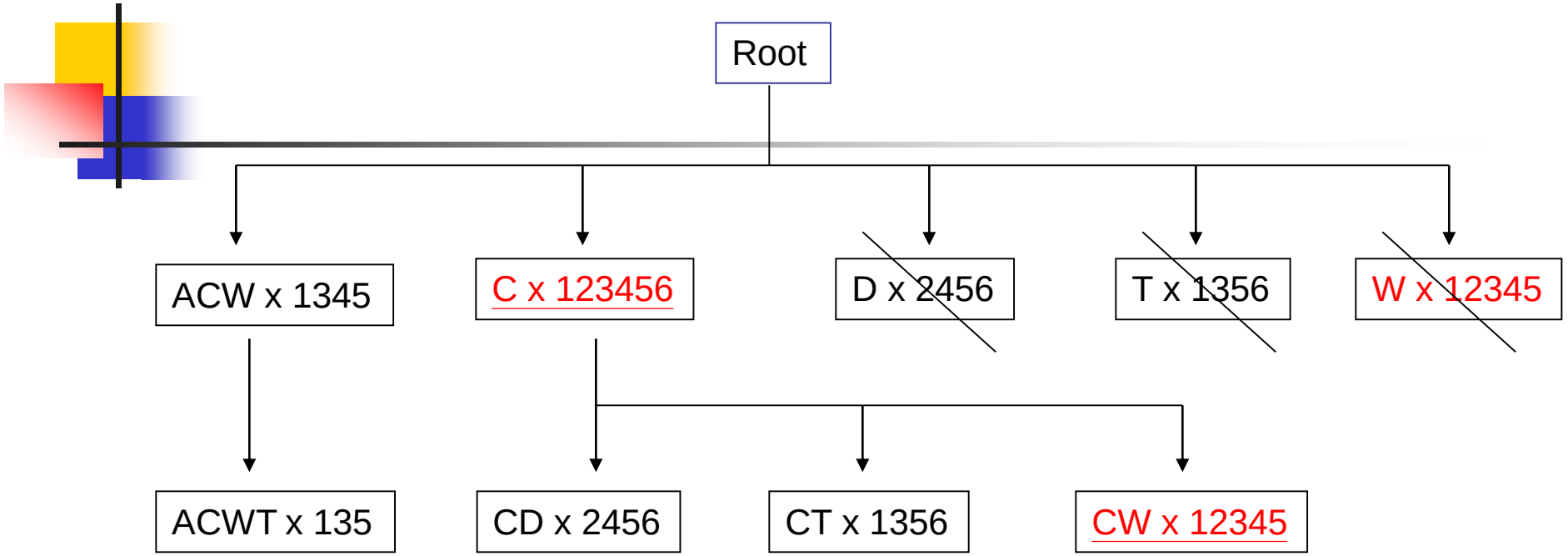
Key	FCI	N
13	ACW:4	
9	ACWT:3	



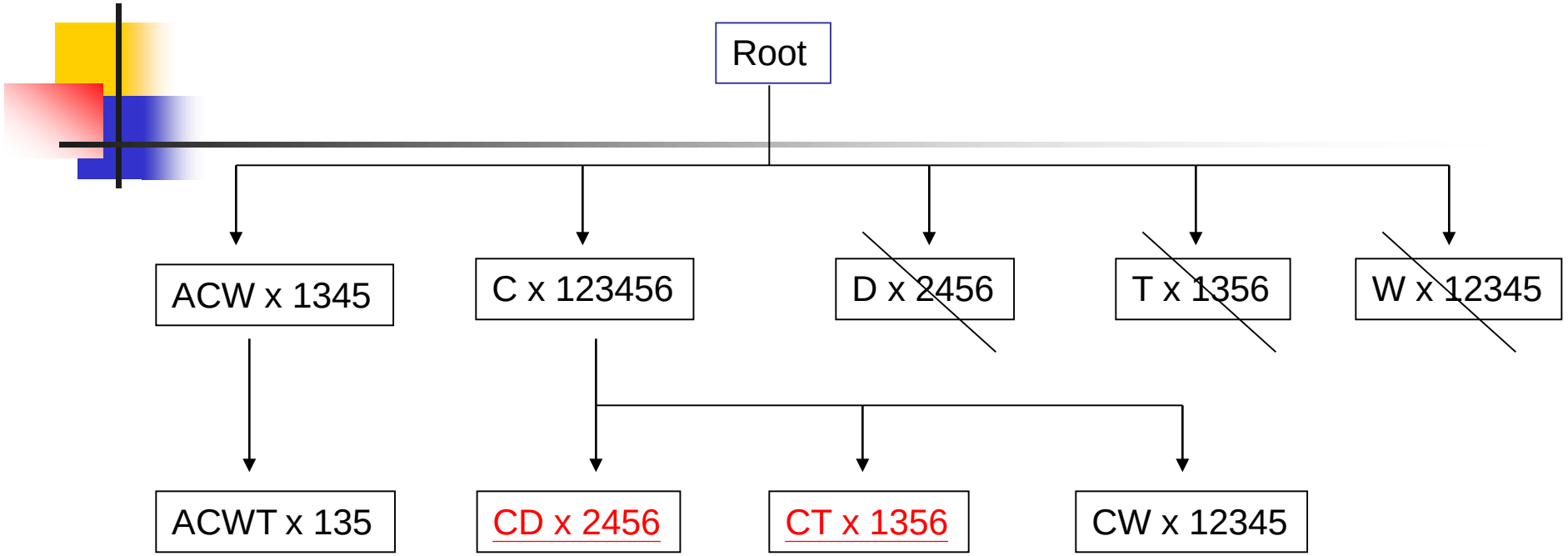
Key	FCI	N
13	ACW:4	
9	ACWT:3	



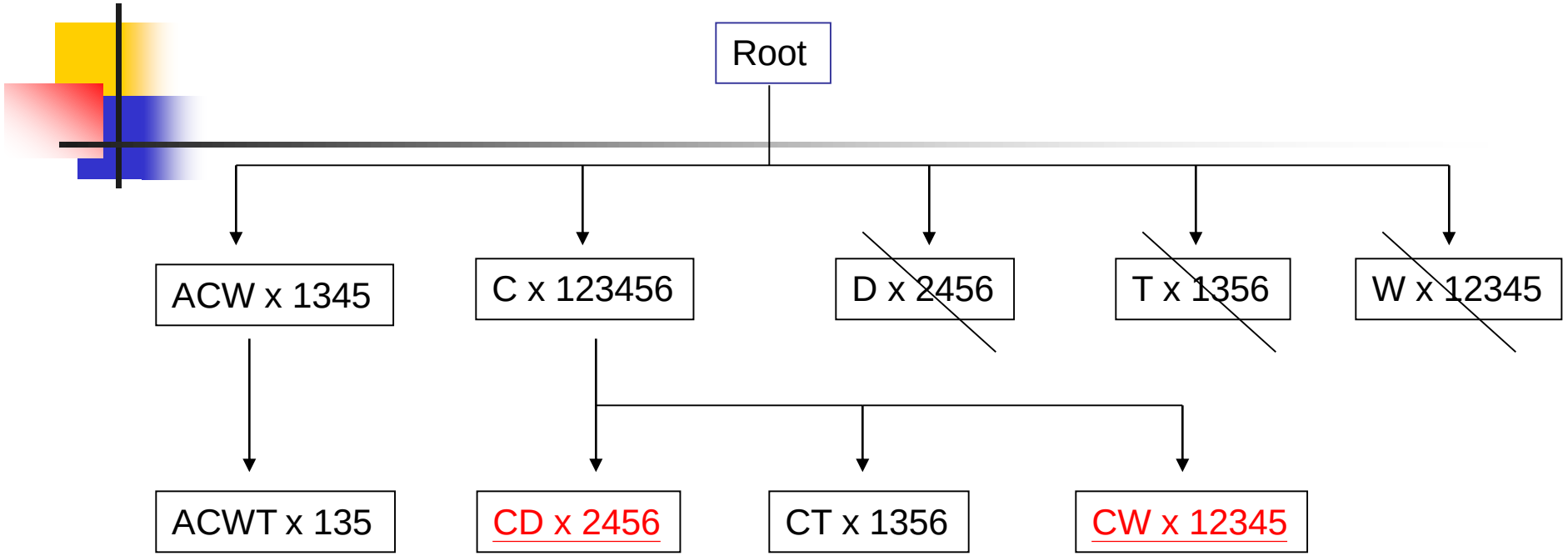
Key	FCI	N
13	ACW:4	
9	ACWT:3	



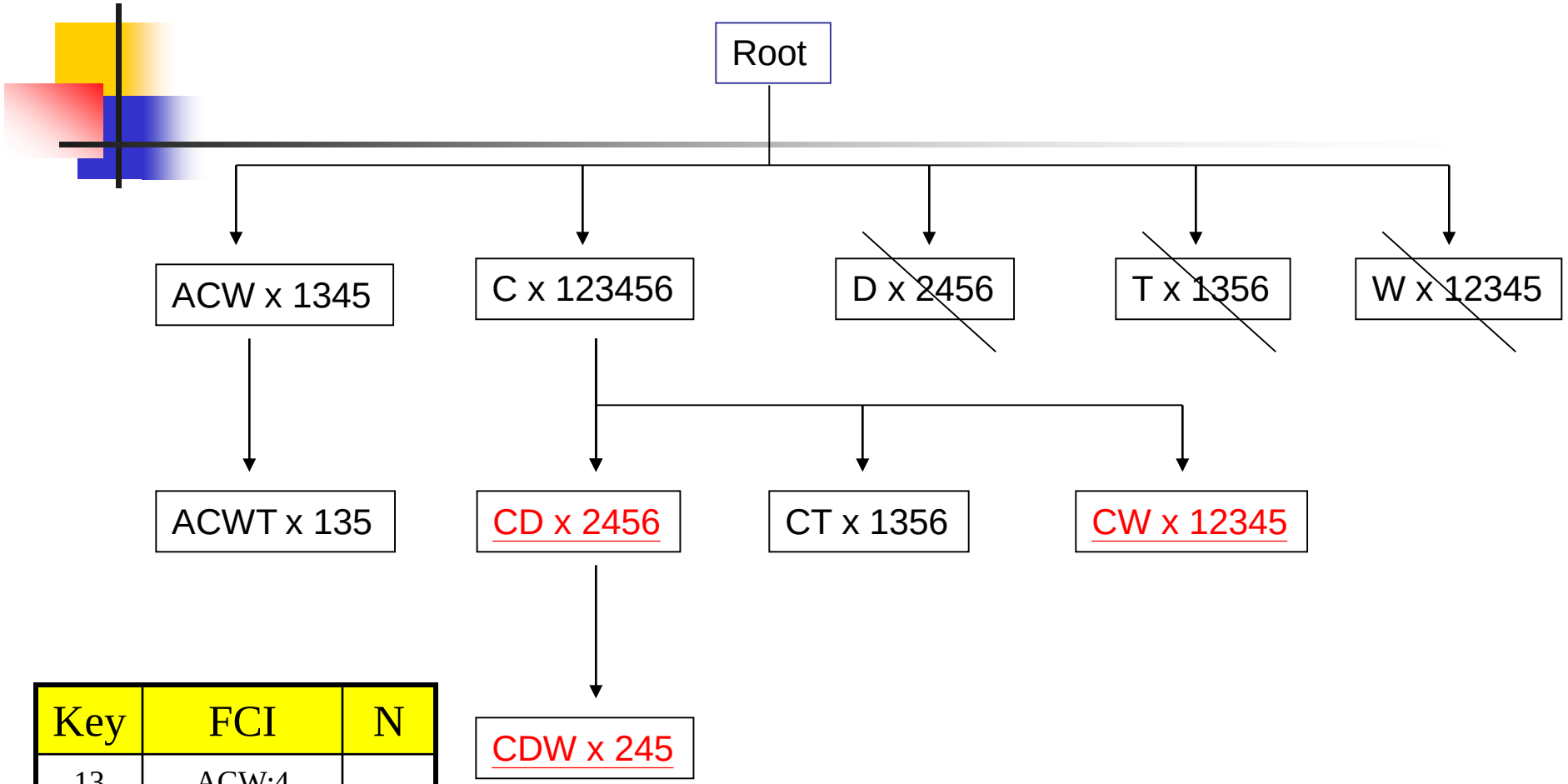
Key	FCI	N
13	ACW:4	
9	ACWT:3	
21	C:6	



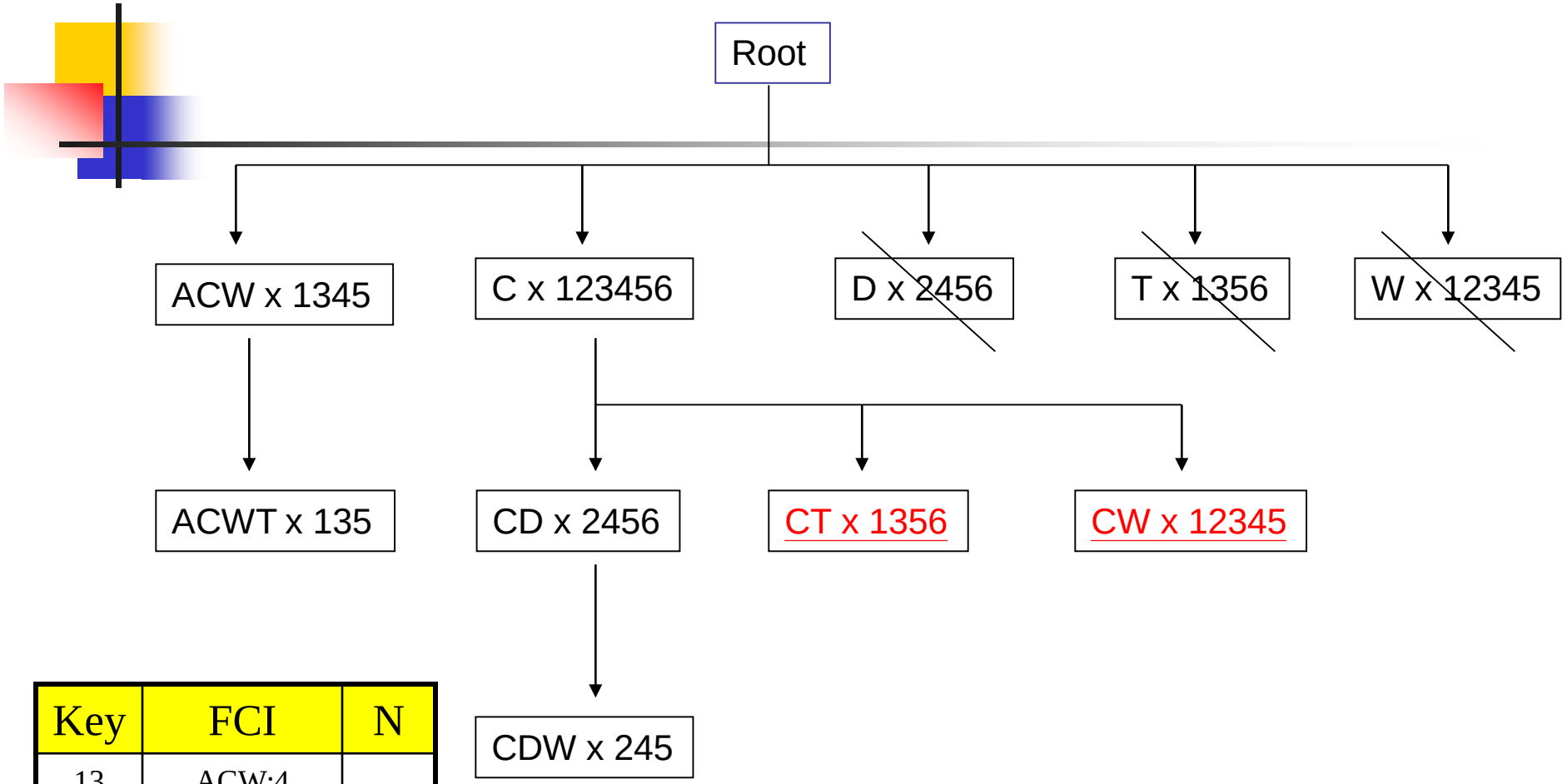
Key	FCI	N
13	ACW:4	
9	ACWT:3	
21	C:6	



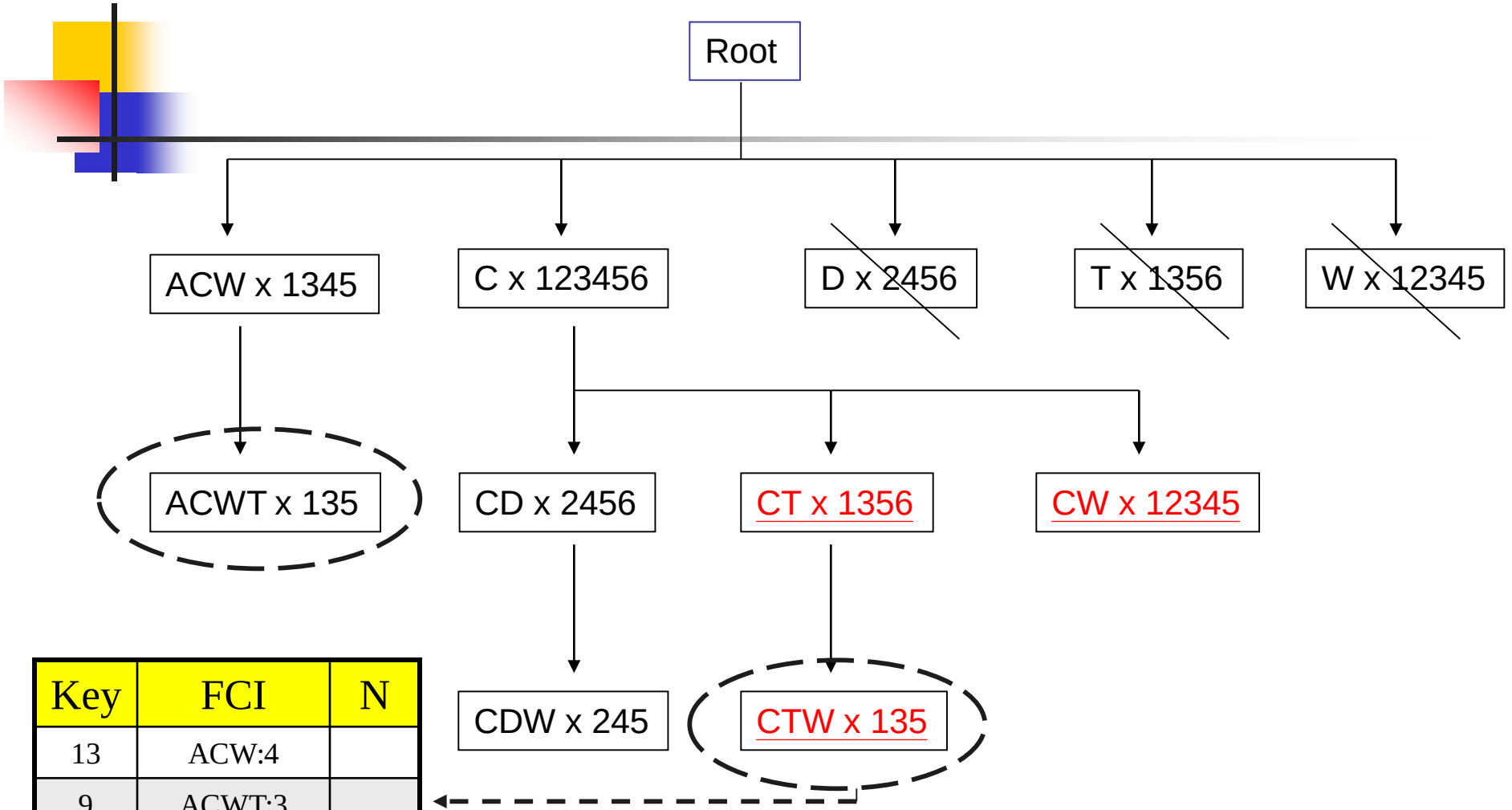
Key	FCI	N
13	ACW:4	
9	ACWT:3	
21	C:6	



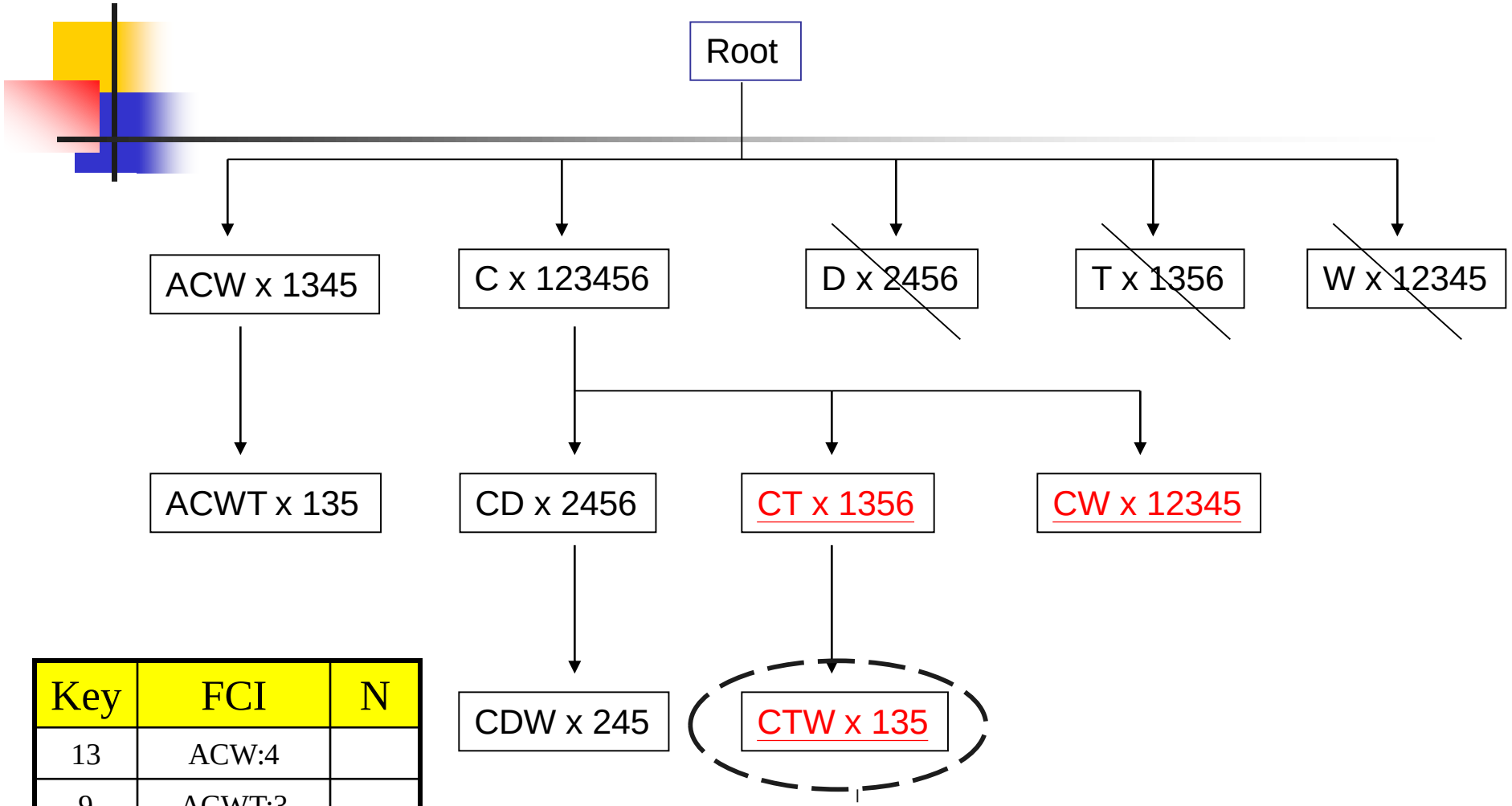
Key	FCI	N
13	ACW:4	
9	ACWT:3	
21	C:6	
17	CD:4	
11	CDW:3	



Key	FCI	N
13	ACW:4	
9	ACWT:3	
21	C:6	
17	CD:4	
11	CDW:3	



Key	FCI	N
13	ACW:4	
9	ACWT:3	
21	C:6	
17	CD:4	
11	CDW:3	



Key	FCI	N
13	ACW:4	
9	ACWT:3	
21	C:6	
17	CD:4	
11	CDW:3	
15	CT:4	

CW:5	

結論

- 傳統頻繁項目集探勘缺點
- 精簡型樣探勘目的
- 精簡型樣類型
- 精簡型樣探勘技術
- 精簡型樣探勘軟體操作

Reference

- [1] Bayardo Jr, R. J. (1998, June). Efficiently mining long patterns from databases. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, pp. 85-93.
- [2] Gouda, K., & Zaki, M. J. (2005). Genmax: An efficient algorithm for mining maximal frequent itemsets. Data Mining and Knowledge Discovery, 11(3), pp.223-242.
- [3]Burdick, D., Calimlim, M., Flannick, J., Gehrke, J., & Yiu, T. (2005). MAFIA: A maximal frequent itemset algorithm. IEEE transactions on knowledge and data engineering, 17(11), pp.1490-1504.
- [4]Grahne, G., & Zhu, J. (2003, May). High performance mining of maximal frequent itemsets. In 6th International Workshop on High Performance Data Mining (Vol. 16, p. 34).
- [5]Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999). Efficient mining of association rules using closed itemset lattices. Information systems, 24(1), pp.25-46.
- [6]Huang, J., Lai, Y. P., Lo, C., & Wu, C. W. (2019, July). An Efficient Algorithm for Deriving Frequent Itemsets from Lossless Condensed Representation. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 216-229). Springer, Cham.

Reference

- [7] Cheng, H., Philip, S. Y., & Han, J. (2006, December). Ac-close: Efficiently mining approximate closed itemsets by core pattern recovery. In Sixth International Conference on Data Mining (ICDM'06) (pp. 839-844). IEEE.
- [8] Zaki, M. J., & Hsiao, C. J. (2002, April). CHARM: An efficient algorithm for closed itemset mining. In Proceedings of the 2002 SIAM international conference on data mining (pp. 457-473). Society for Industrial and Applied Mathematics.
- [9] Grahne, G., & Zhu, J. (2005). Fast algorithms for frequent itemset mining using fp-trees. IEEE transactions on knowledge and data engineering, 17(10), 1347-1362.
- [10]Grahne, G., & Zhu, J. (2003, May). High performance mining of maximal frequent itemsets. In 6th International Workshop on High Performance Data Mining (Vol. 16, p. 34).
- [11] Zaki, M. J. (2000). Scalable algorithms for association mining. IEEE transactions on knowledge and data engineering, 12(3), pp.372-390.

教師資訊

- 姓名：吳政瑋 (小吳老師)
- 現職：宜大資工助理教授
- 學歷：成功大學資工博士
- 研究興趣：資料探勘、人工智慧、AIoT應用
- 通訊方式
  - 電子信箱：wucw@niu.edu.tw
  - 校內電話: (03)9317331
  - Line: silvemoonfox
  - Office: 格致大樓E405室
  - 數位學習園區
- 實驗室：AI與資料科學實驗室
  - https://sites.google.com/view/cwwuadslab/

意見交流

歡迎提供意見與指導!!

您的寶貴意見將使本系更進步!!