



XXXII Conferencia Latinoamericana de Estudios Informáticos

Santiago de Chile
2006

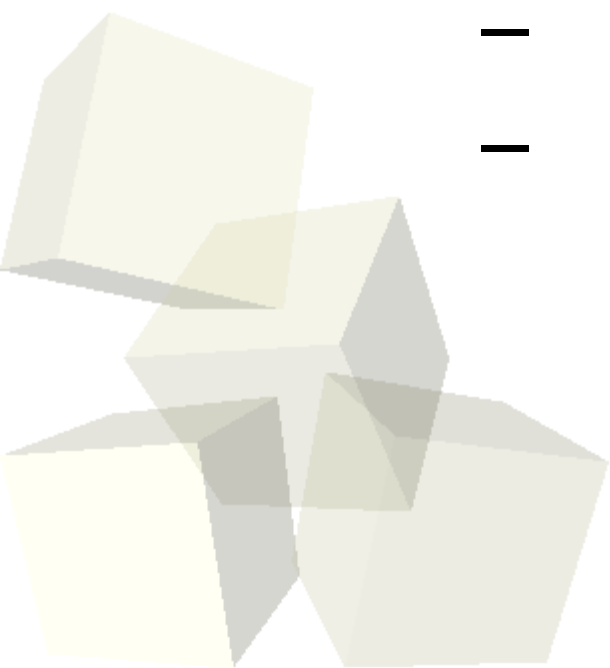
Análisis de Desempeño de EquipAsso:
un algoritmo para el cálculo de itemsets
frecuentes basado en operadores
algebraicos relacionales



Grupo de Investigación GRIAS – Línea KDD
Universidad de Nariño - Colombia

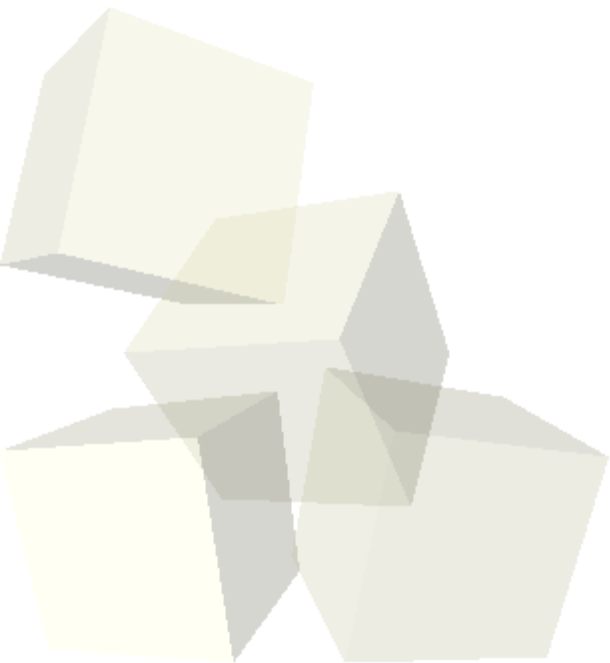


- Conceptos Preliminares.
- Algoritmo EquipAsso.
- Aspectos de la Implementación de EquipAsso en Tariy
- Evaluación de Rendimiento
- Conclusiones y Trabajos futuros





Conceptos Preliminares





Reglas de Asociación

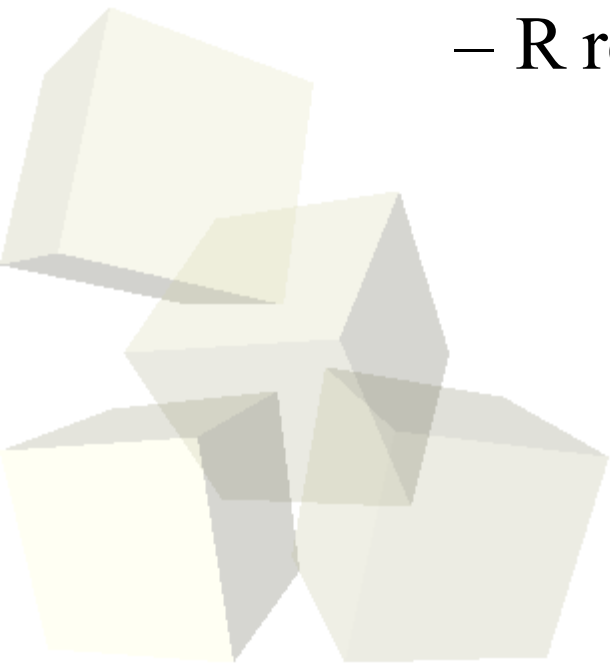
- El problema fue formulado por Agrawal et al (1993)
- Aplicación canasta de mercado(*market-basket*)
- Los resultados se presentan en formas de reglas de asociación que cumplen unas especificaciones mínimas dadas por el usuario, expresadas en forma de *soporte* y *confianza*.



- Associator (α)
- EquiKeep (χ)

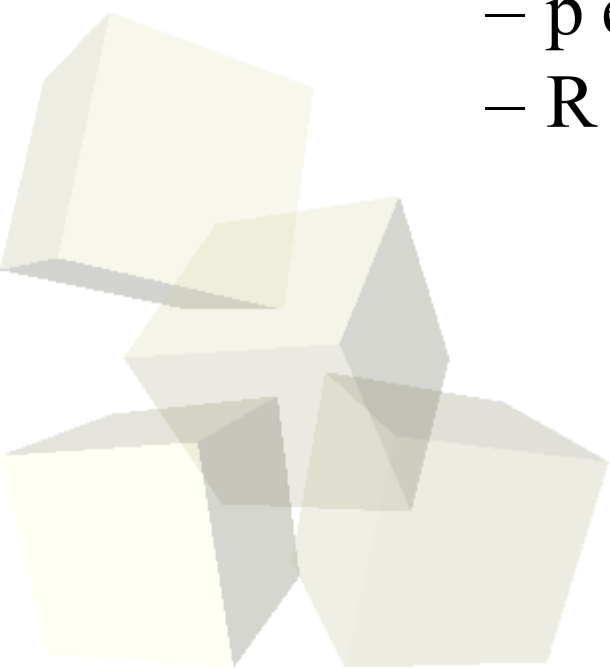


- Associator genera, a partir de cada tupla de una relación R , todos sus posibles subconjuntos (*itemsets*) de diferente tamaño.
- $\alpha_{IS,ES}(R)$
 - IS tamaño inicial *itemset*
 - ES tamaño final *itemset*
 - R relación



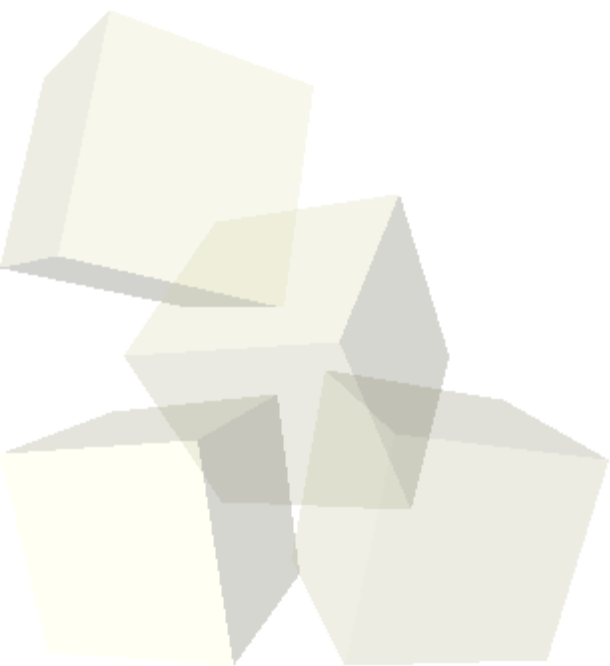


- *EquiKeep* restringe los valores de los atributos de cada una de las tuplas de la relación R a únicamente los valores de los atributos que satisfacen una expresión lógica.
- $\chi_p(R)$
 - p expresión lógica
 - R relación





- *Associator Range*
- *EquiKeep On*





Associator Range

```
SELECT <ListaAtributosTablaDatos> [INTO  
                                <NombreTablaAssociator>]  
FROM <NombreTablaDatos>  
WHERE <CláusulaWhere>  
ASSOCIATOR RANGE <numero1> UNTIL <numero2>  
GROUP BY <ListaAtributosTablaDatos>
```

<ListaAtributosTablaDatos>:=<Atributo> | <ListaAtributos>
<numero1>:= 1,2,3,4 ...
<numero2>:= 1,2,3,4 ...

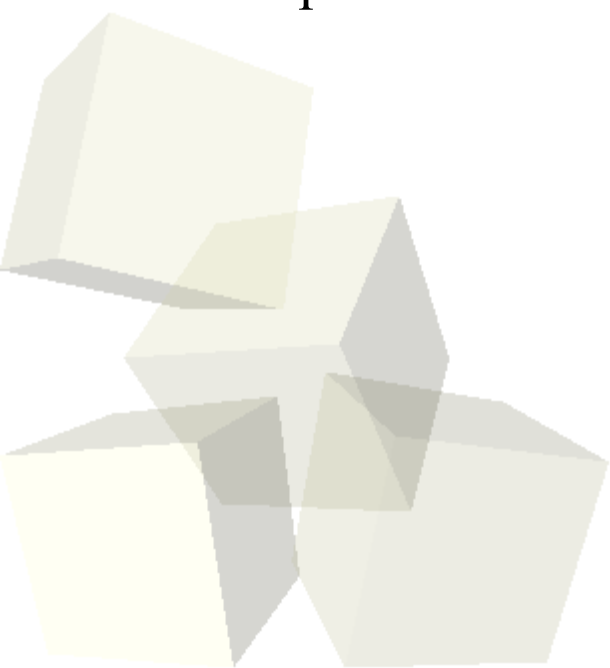


```
SELECT <ListaAtributosTablaDatos> [INTO  
                                   NombreTablaEquiKeep>]  
FROM <NombreTablaDatos>  
WHERE <CláusulaWhere>  
EQUIKEEP ON <CondiciónValoresAtributos >
```

<ListaAtributosTablaDatos>:=<Atributo> | <ListaAtributos>

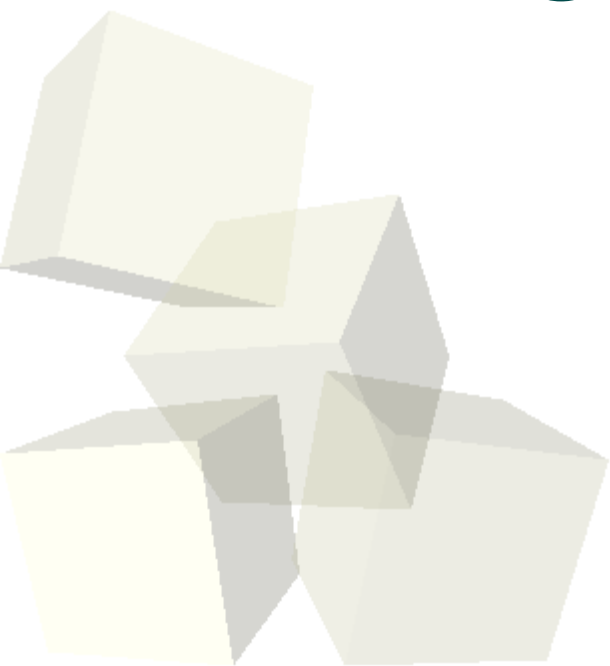
<CondiciónValoresAtributos>:=<Atributo=Valor><operador><Atributo=Valor>

<operador>:=AND,OR,NOT, IN





Algoritmo EquipAsso



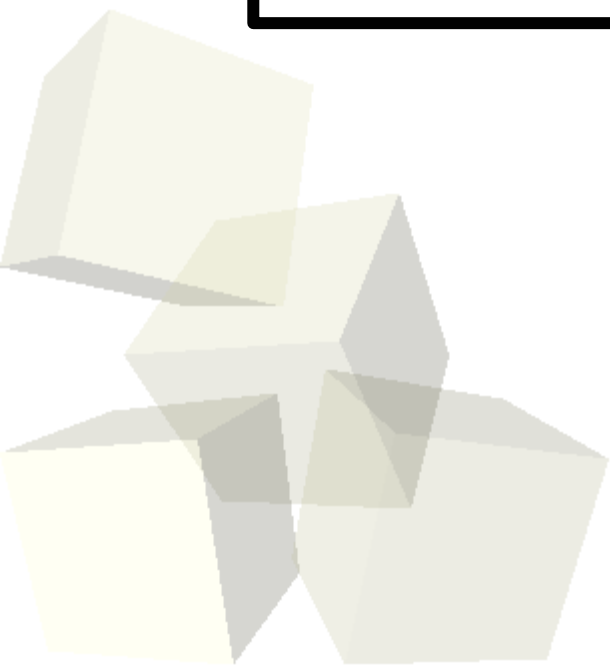
Descripción Algoritmo EquipAsso

```
Leer  $g$  // tamaño de la regla
 $L1 = \{1 - \text{conjuntos de ítems frecuentes}\}$ 
Forall transacciones  $t \in D$  do begin
    // se aplica el operador EquipKeep
     $R = \chi_{L1}(D)$ 
     $k = 2$ 
    // genera todos los conjuntos de ítems posibles
     $R' = \alpha_{k,q}(R) = \{ \bigcup_{all} X_i \mid X_i \subseteq t_i \}$ 
End
    // conjuntos de ítems frecuentes
     $L = \{ \text{count}(R') \mid \text{count} \leq \text{minsup} \}$ 
```



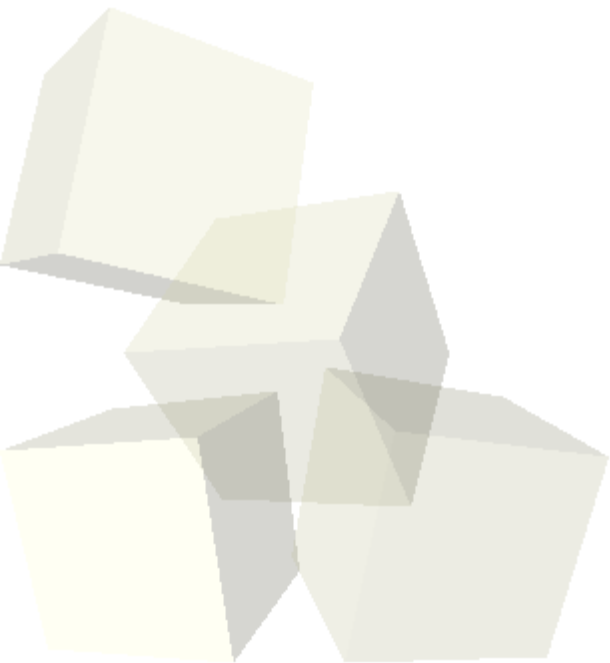
Algoritmo SQL EquipAsso

```
SELECT <ListaAtributosTablaDatos>, count(*) AS soporte  
INTO R1  
FROM D  
EQUIKEEP ON L1  
ASSOCIATOR RANGE k UNTIL g  
GROUP BY <ListaAtributosTablaDatos>  
HAVING count(*) >= minsup
```





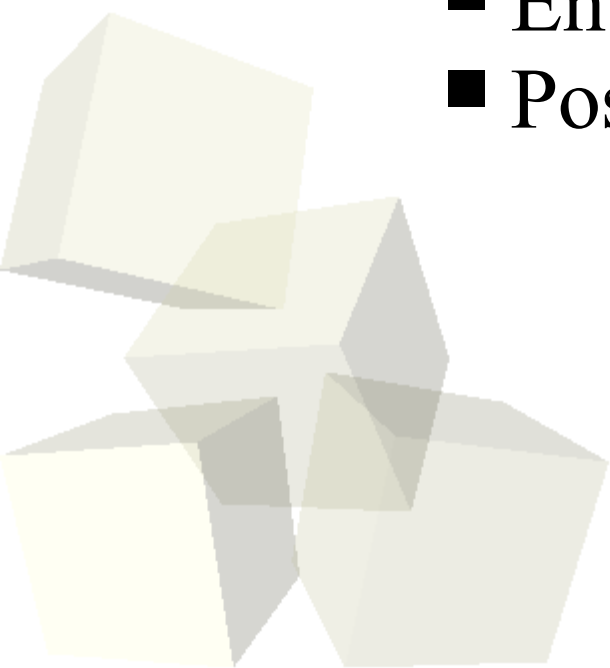
Aspectos de la Implementación de EquipAsso en Tariy





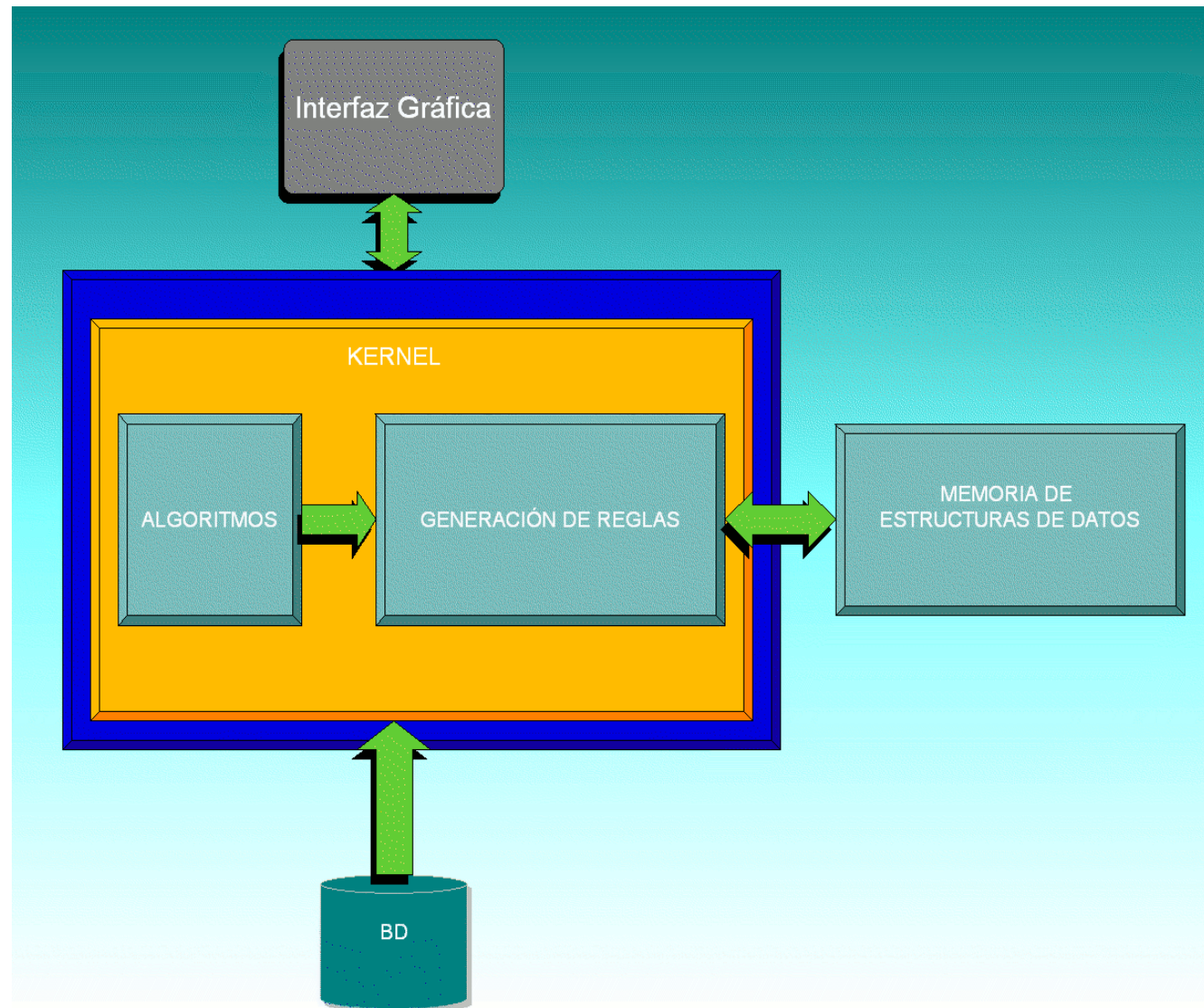
Equipo y Software

- Procesador AMD 64 bits a 2 Ghz
- 1 GB de RAM
- DD ATA 80 GB a 150 MB/sg
- Sistema operativo Fedora Core
- Lenguaje Java 1.5.0_06
- Entorno de desarrollo NetBeans 5.0
- PostgreSQL versión 7.3.4.



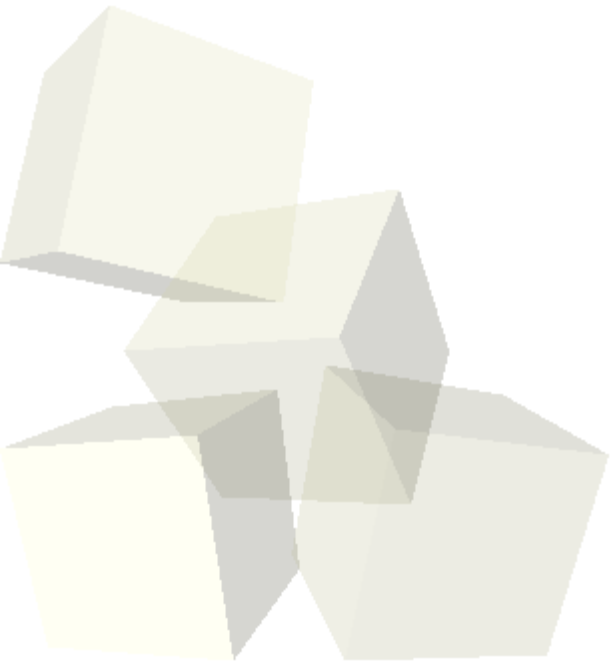


Arquitectura Tariy





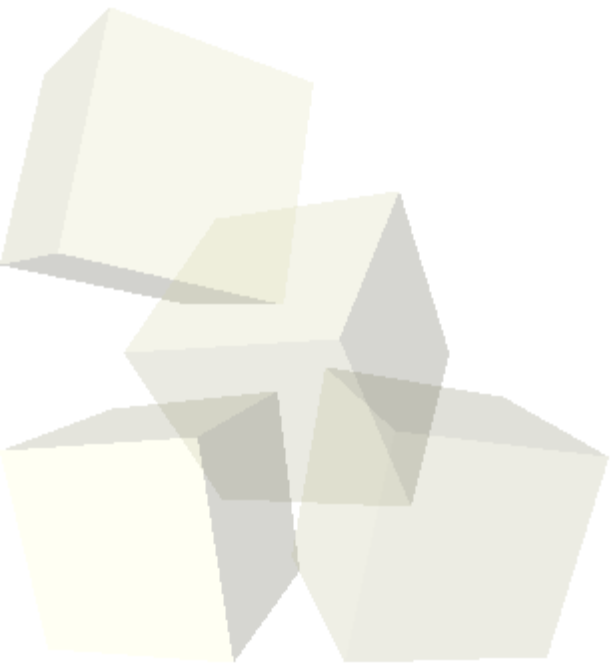
Evaluación de Rendimiento





Algoritmos Evaluados

- A priori [Agrawal et al. 1994]
- FP-Growth [Han et al. 2000]
- EquipAsso [Timarán et al. 2005]





Conjuntos de datos utilizados

Nomenclatura	Numero de Registros	Numero de Transacciones	Promedio items por Transacción
BD85KT7	555.123	85.692	7
BD40KT5	194.337	40.256	5
BD10KT10	97.824	10.731	10



Resultados para BD85KT7

BD85KT7			
Soporte (%)	Tiempo (ms)		
	Apriori	FP-Growth	EquipAsso
4,15	750	166	85
4,75	362	162	82
5,35	365	164	83
5,95	365	162	83
6,55	120	159	80

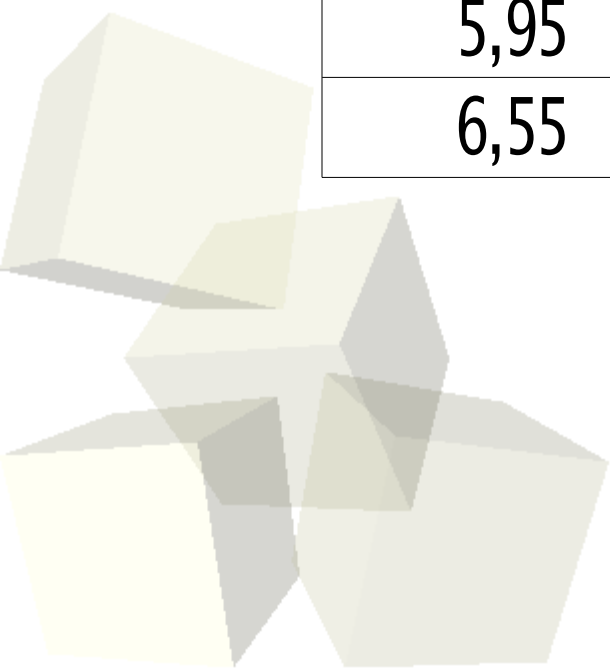
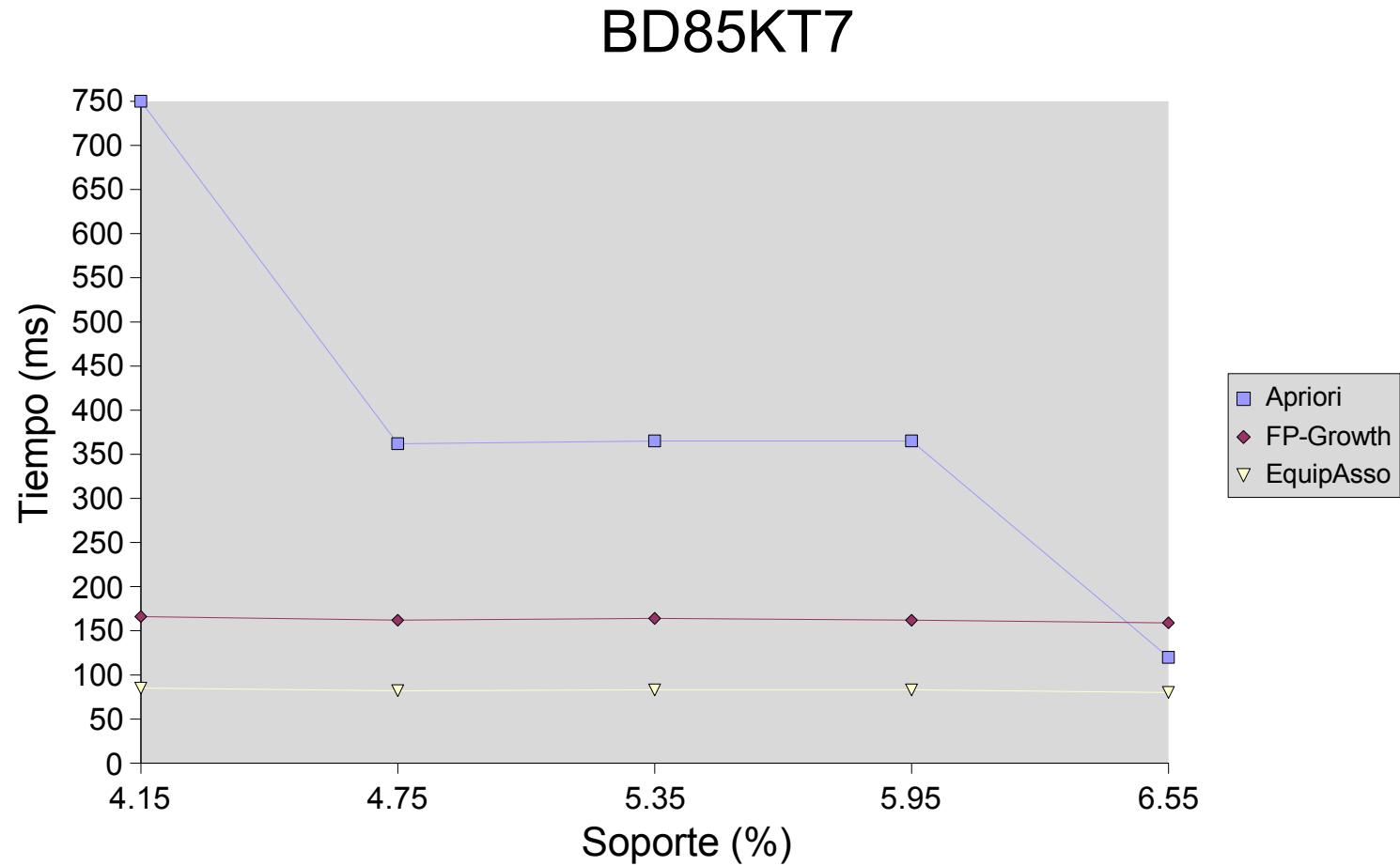




Gráfico para BD85KT7



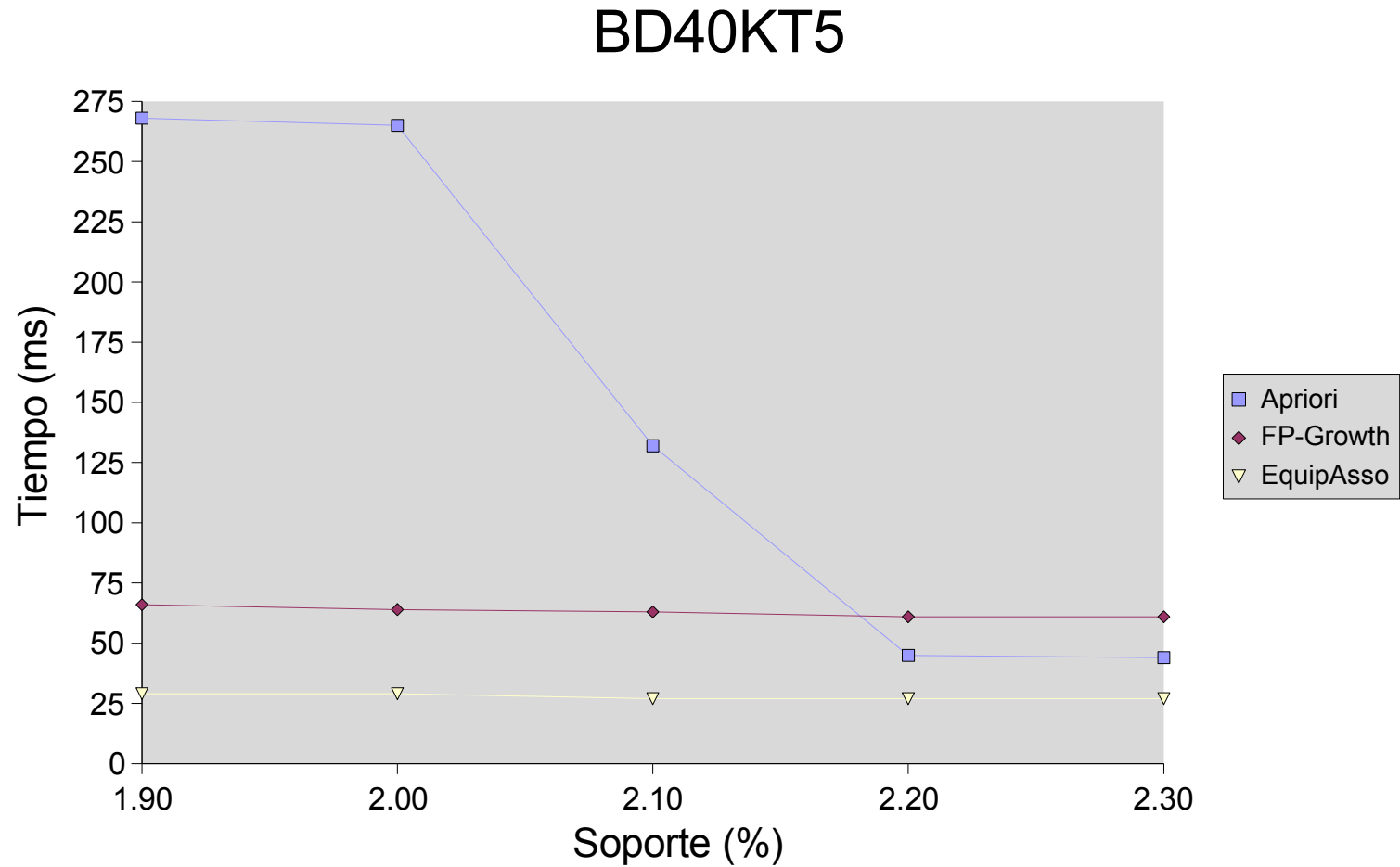


Resultados para BD40KT5

BD40KT5			
Soporte (%)	Tiempo (ms)		
	Apriori	FP-Growth	EquipAsso
1,90	268	66	29
2,00	265	64	29
2,10	132	63	27
2,20	45	61	27
2,30	44	61	27



Gráfica para BD40KT5





Resultados para BD10KT10

BD10KT10			
Soporte (%)	Tiempo (ms)		
	Apriori	FP-Growth	EquipAsso
3,00	525	28	15
4,00	182	26	14
5,00	105	25	13
6,00	53	24	13
7,00	52	24	13

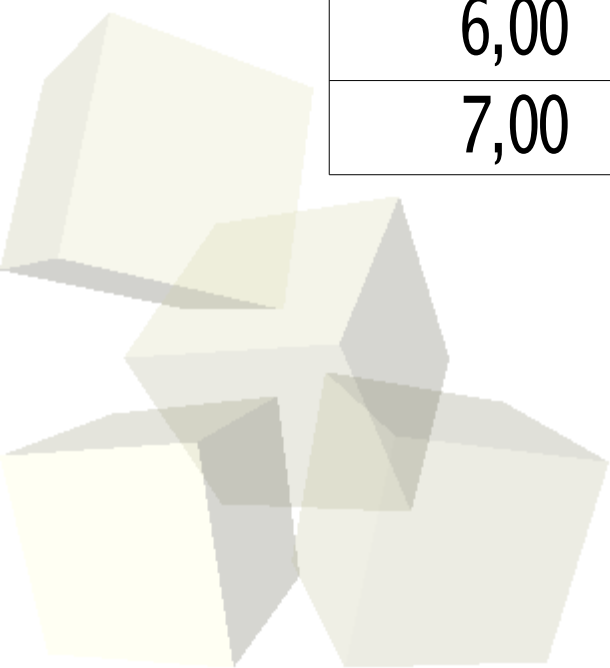
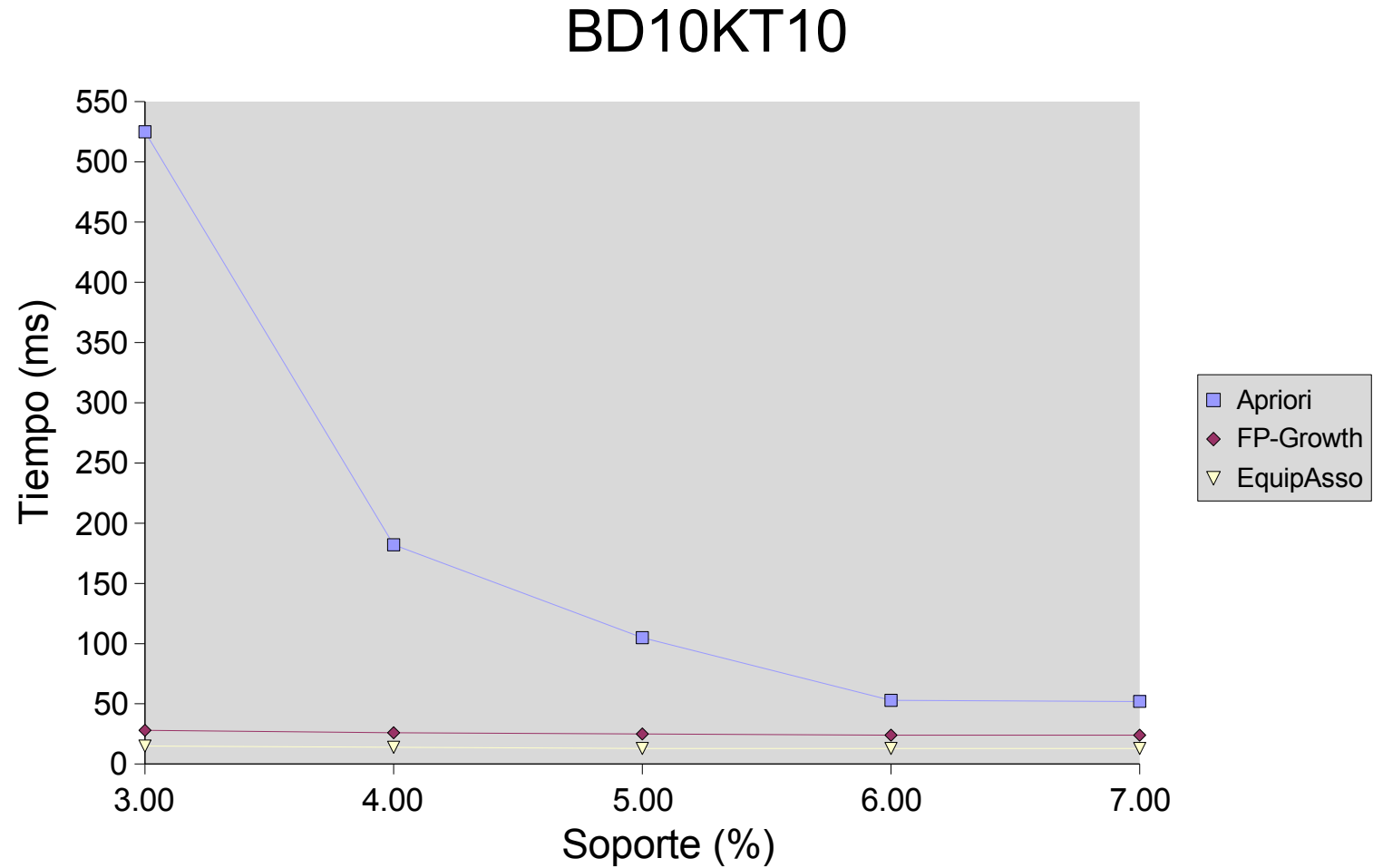




Gráfico para BD10KT10

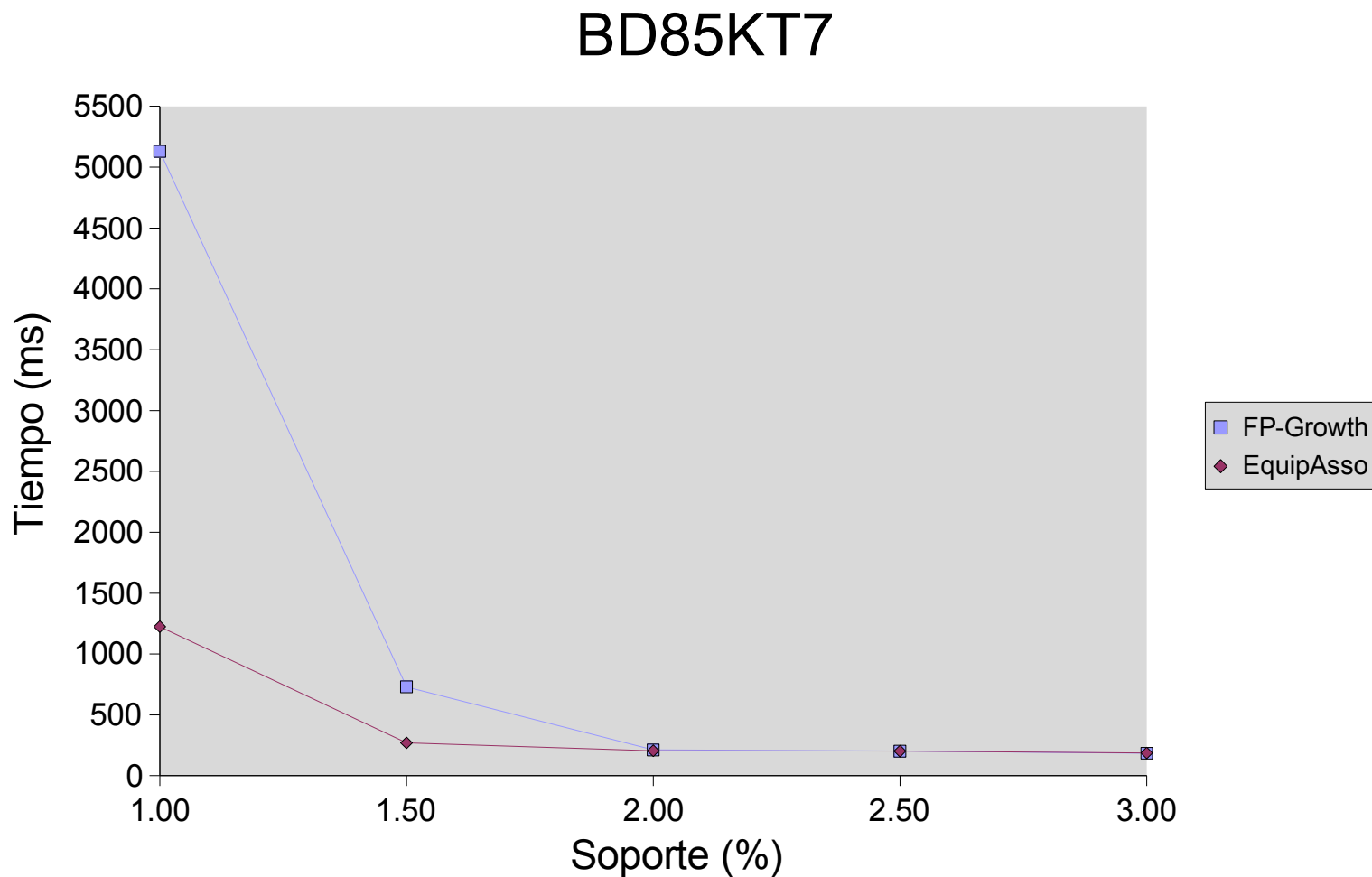


Resultados para BD85KT7 soportes bajos

BD85KT7			
Soporte (%)	Tiempo (ms)		
	Apriori	FP-Growth	EquipAsso
1,00	-	5130	1225
1,50	-	730	270
2,00	-	212	205
2,50	-	202	202
3,00	-	185	187

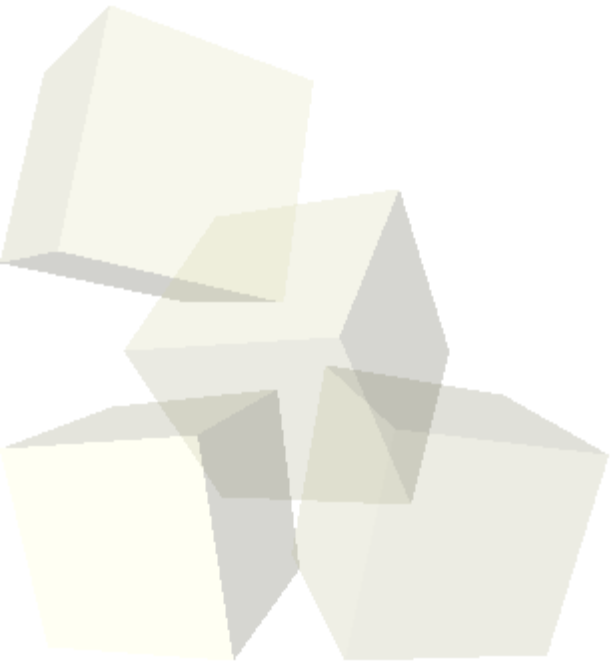


Gráfico para BD85KT7 soportes bajos



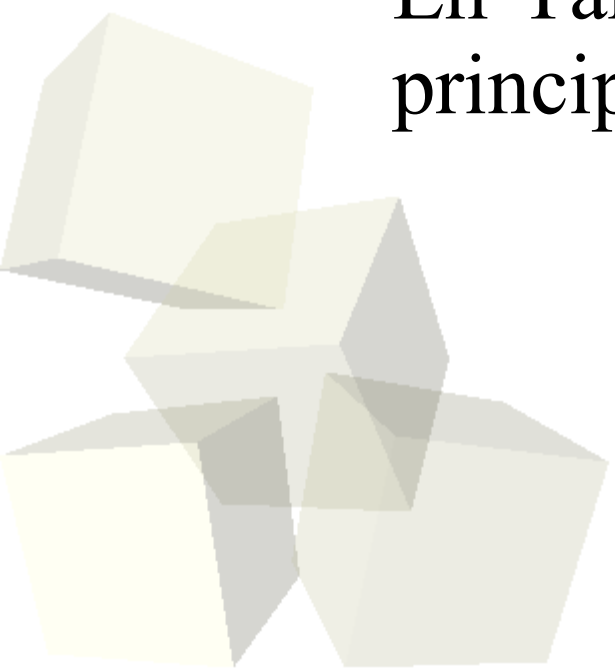


Conclusiones y Trabajos Futuros



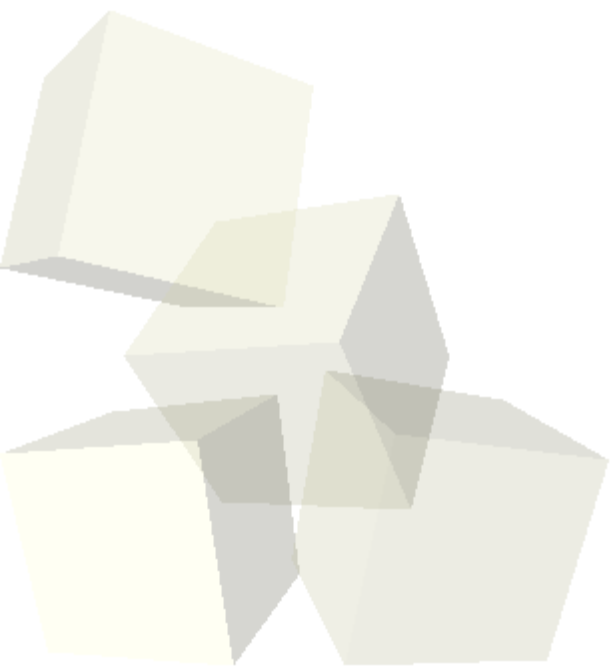


- Para soportes altos, *EquipAsso* y *FP-Growth* tienen tiempos de ejecución semejantes.
- Para soportes bajos, *Equipasso* tuvo un mejor desempeño que *FP-Growth*.
- Bajo rendimiento de *Apriori* con respecto a *EquipAsso* y *FP-Growth*
- En *Tariy*, *Equipasso*, consume más memoria principal que *FP-Growth*





- Complementar pruebas con otros conjuntos de datos reales.
- Dotar a Tariy de:
 - Una interfaz gráfica.
 - Nuevos módulos que soporten otras tareas y algoritmos de minería de datos.



Gracias



Ricardo Timarán Pereira, Ph.D.

Andrés O. Calderón Romero

Iván Ramírez Freyre

Juan Carlos Alvarado

Fernando Guevara

Grupo de Investigacion GRIAS – Línea KDD
Universidad de Nariño - Colombia