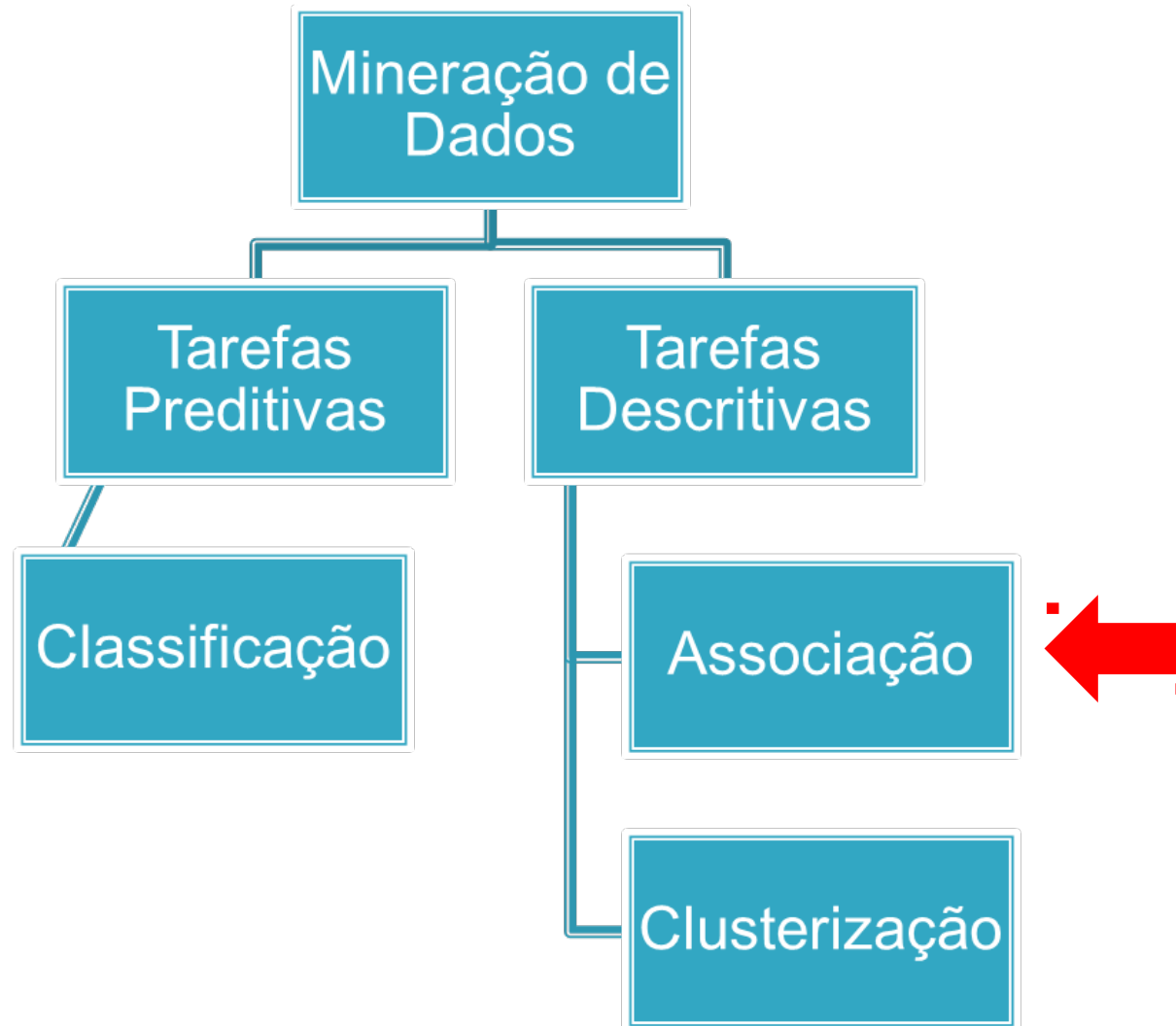
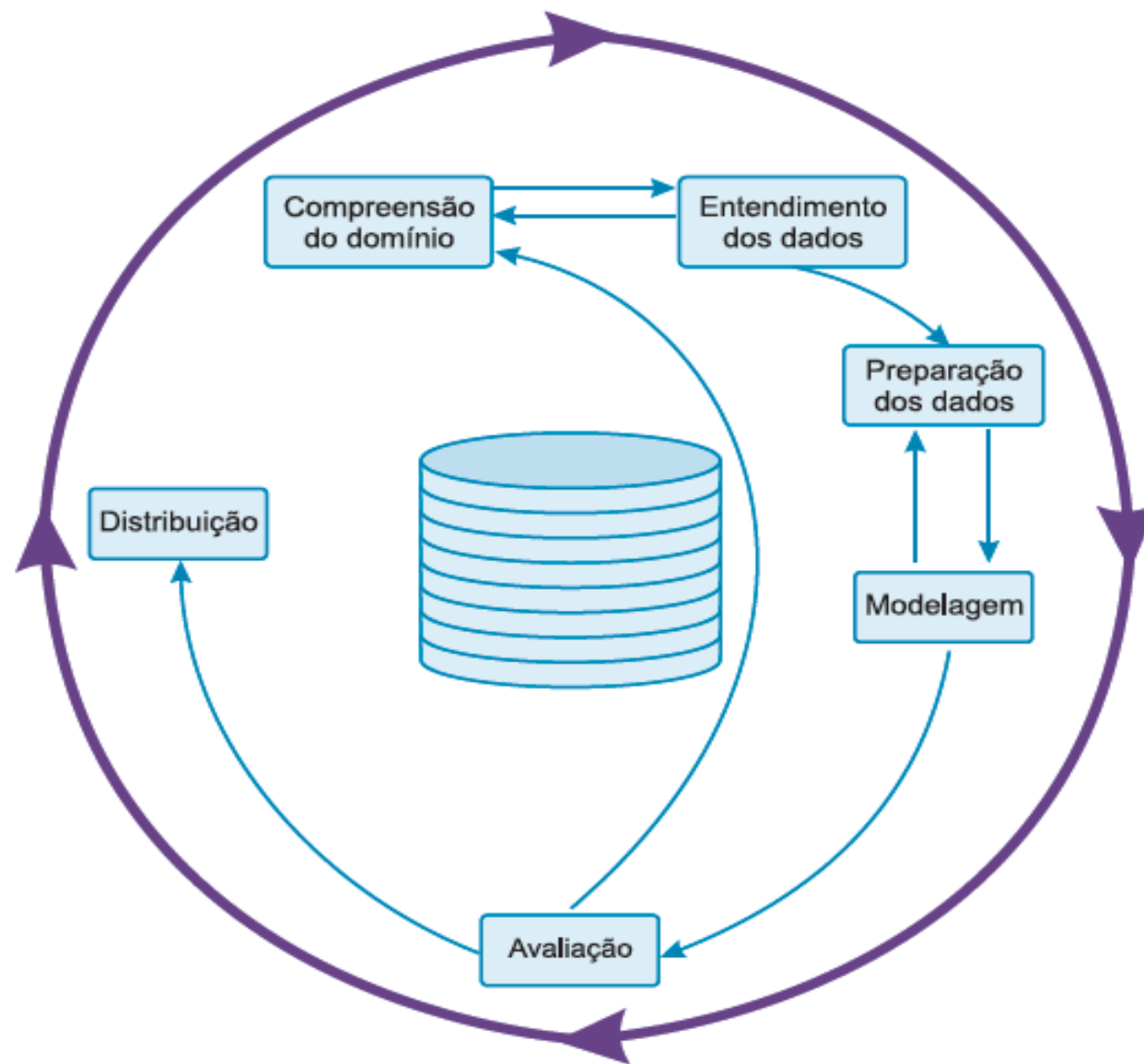


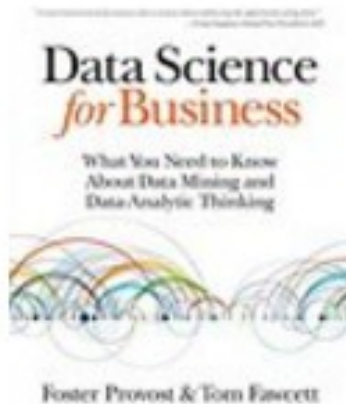
# Tarefas de Mineração de Dados



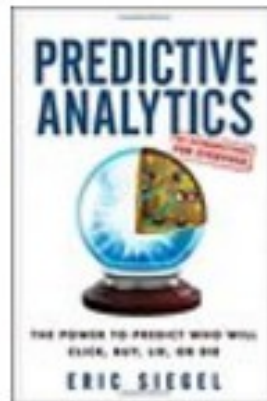


# Motivação

## Frequently Bought Together



+



+



Total price: **CDN\$ 102.41**

Add all three to Cart

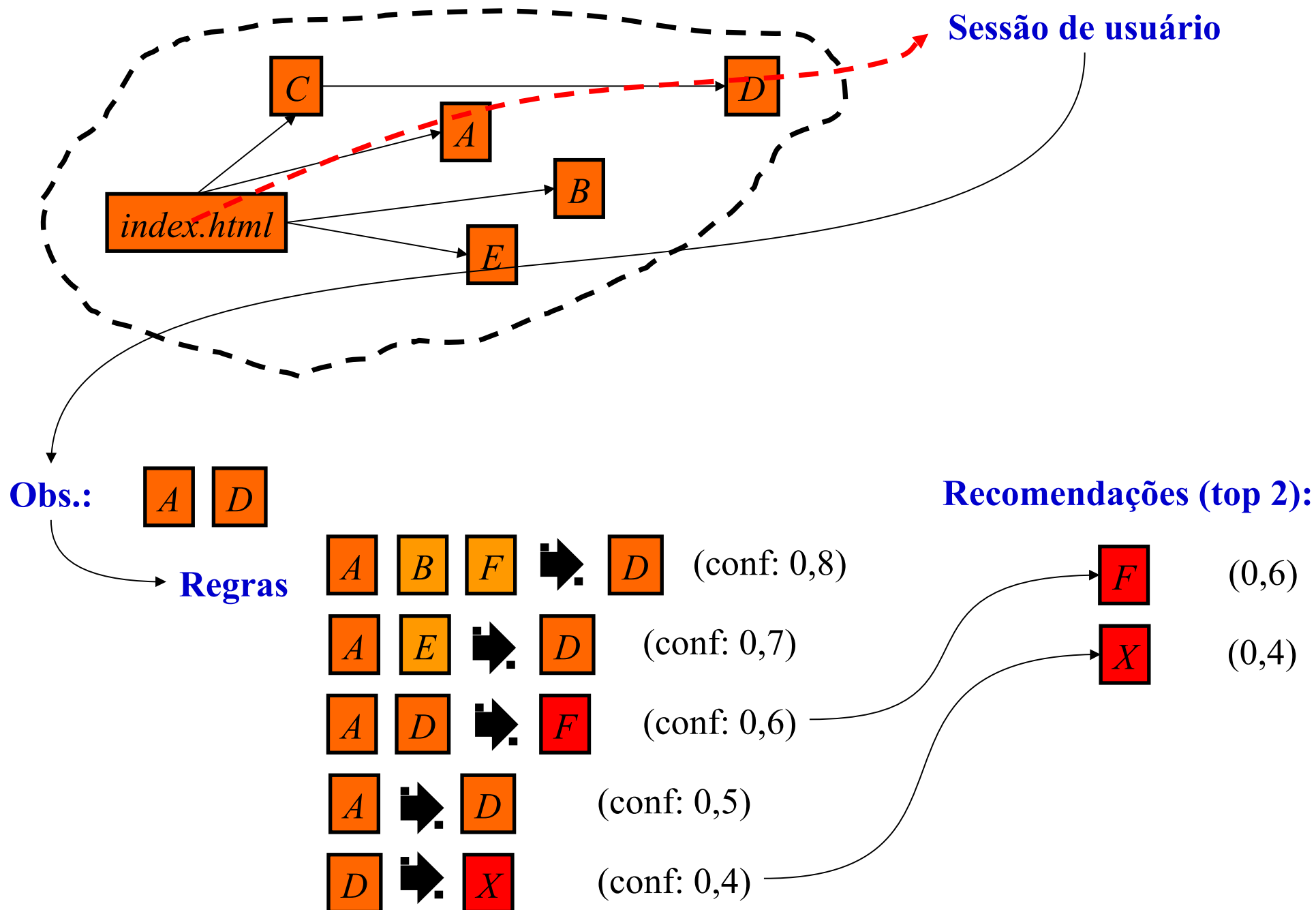
# Motivação: Youtube

## 2.2 Related Videos

One of the building blocks of the recommendation system is the construction of a mapping from a video  $v_i$  to a set of similar or *related* videos  $R_i$ . In this context, we define similar videos as those that a user is likely to watch after having watched the given *seed video*  $v$ . In order to compute the mapping we make use of a well-known technique known as association rule mining [1] or co-visitation counts. Consider sessions of user watch activities on the site. For a given time period (usually 24 hours), we count for each pair of videos  $(v_i, v_j)$  how often they were co-watched within sessions. Denoting this co-visitation count by  $c_{ij}$ , we define the *relatedness score* of video  $v_j$  to base video  $v_i$  as:

The YouTube Video Recommendation System

# Exemplo de um Sistema de Recomendação usando Associação



# Estrutura dos dados

transaction ID	items
1	milk, bread
2	bread, butter
3	beer
4	milk, bread, butter
5	bread, butter

		items			
		milk	bread	butter	beer
transactions	1	1	1	0	0
	2	0	1	1	0
	3	0	0	0	1
	4	1	1	1	0
	5	0	1	1	0

# Associação

- ▶ Uma regra de associação é uma implicação da forma  $(X \rightarrow Y)$ , onde  $X$  e  $Y$  são conjuntos de itens e  $X \cap Y = \emptyset$ .


TID	Lista de Itens
T1	praga_colmo, praga_raizes
T2	praga_colmo, produção, cachaça, logística
T3	praga_raizes, produção, cachaça, etanol
T4	praga_colmo, praga_raizes, produção, cachaça
T5	praga_colmo, praga_raizes, produção, etanol

- R1: {cachaça}  $\rightarrow$  {produção}
- R2: {cachaça, praga\_colmo}  $\rightarrow$  {praga\_raizes}

# Ilustrando o Princípio Apriori

Item	Frequência
Praga_colmo	4
Praga_raizes	2
Cachaça	4
Produção	3
Logística	4
Açúcar	1

Itens (1-itemset)




Item	Frequência
{Praga_colmo, Cachaça}	3
{Praga_colmo, Produção}	2
{Praga_colmo, Logística}	3
{Cachaça, Produção}	2
{Cachaça, Logística}	3
{Produção, Logística}	3

Pares (2-itemsets)

Não há necessidade de gerar candidatos que contêm Praga\_raízes ou Açúcar.

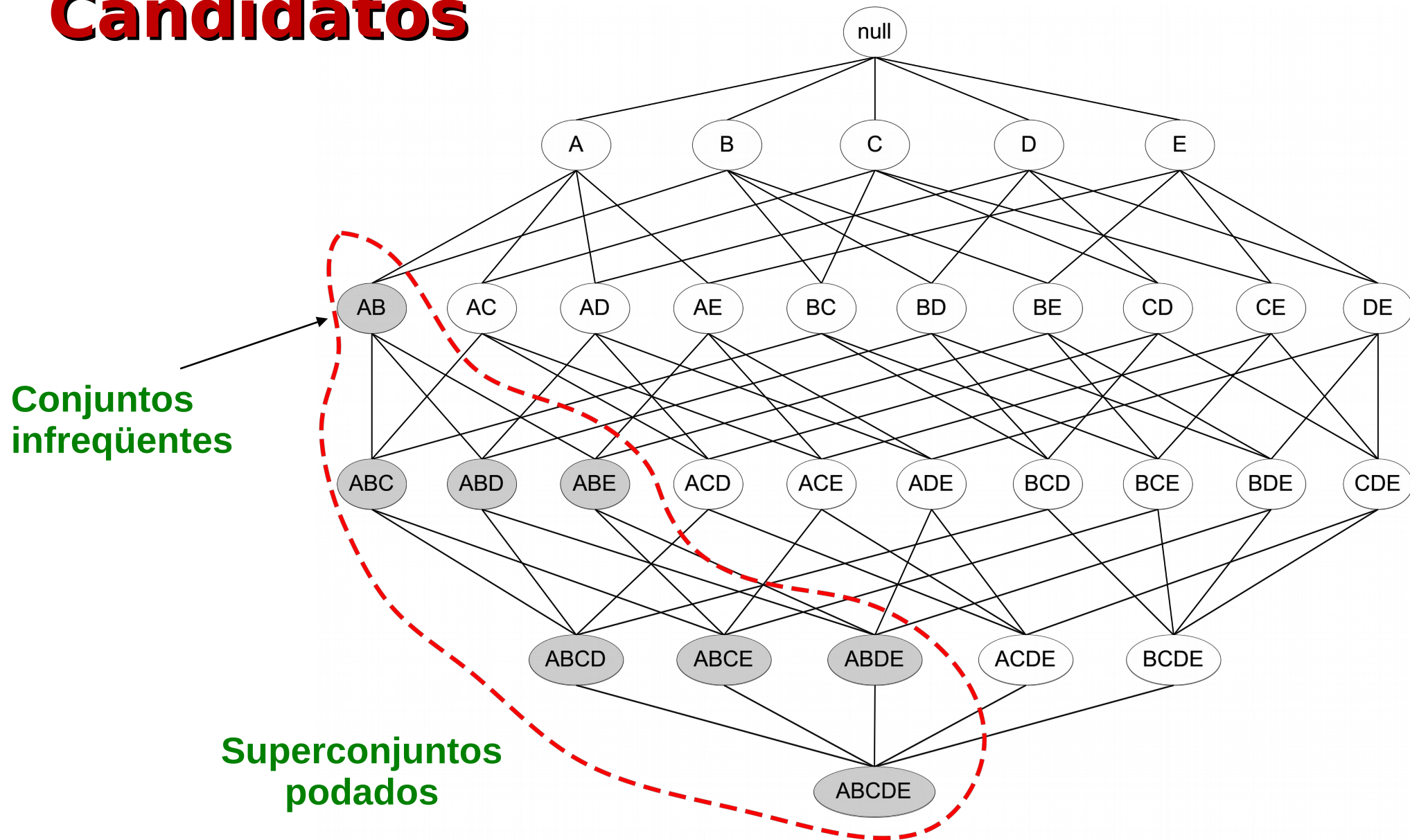
Mínimo Suporte = 3



Item	Frequência
{Praga_colmo, Cachaça, Logística}	3



# Reduzindo o Número de Candidatos



# Métricas

## ➤ Suporte (s):

\* Fração das transações que contém ambos X e Y.

$$* \text{Sup}(X \rightarrow Y) = P(X \text{ e } Y)$$

## ➤ Confiança (c):

\* Mede a frequência dos itens em Y que aparece nas transações em que contem X.

$$* \text{Conf}(X \rightarrow Y) = P(Y|X).$$

$$* \text{Conf}(X \rightarrow Y) = \text{Sup}(X \text{ e } Y) / \text{Sup}(X)$$

TID	Lista de Itens
T1	praga_colmo, praga_raizes
T2	praga_colmo, produção, cachaça, logística
T3	praga_raizes, produção, cachaça, etanol
T4	praga_colmo, praga_raizes, produção, cachaça
T5	praga_colmo, praga_raizes, produção, etanol

## ➤ Exemplo:

$\{\text{cachaça}\} \rightarrow \{\text{produção}\}$

$$\text{Sup} = \frac{\text{Freq}(\text{cachaça}, \text{produção})}{|T|} = 3/5$$

$$\text{Conf} = \frac{\text{Freq}(\text{cachaça}, \text{produção})}{\text{Freq}(\text{cachaça})} = 3/3$$

# Big Data

Apriori-Map/Reduce Algorithm