Análise de Dados Simbólicos
Geração e Exportação de SOs no SODAS

Aluno:
Anderson Costa
afbfc@cin.ufpe.br

Introdução

- Diferentes abordagens para análise de dados
- Análise de dados estatística
  - Síntese de grande quantidades de dados
  - Processar diretamente os dados
- Outra abordagem
  - Transformar os dados detalhados em descrições sumarizadas
  - Analisar as descrições
- Análise de dados simbólicos
  - Descrever objetos de alto nível (*assertions*)
  - Descrição de grupos de indivíduos e as variações dentro dos grupos
  - Estender os métodos tradicionais de análise estatística

Terminologia

- Indivíduos
- Descrições
- Objetos
- Variáveis simbólicas
- Tabela de dados Simbólicos
  - Valores numéricos e/ou categóricos

Abordagem Simbólica

- **Objetos simbólicos** são dados agregados de acordo com um grupo de indivíduos que serão tratados como uma nova unidade estatística.

- Atributo de grupo: variáveis que **formam** o grupo

- Variáveis simbólicas: variáveis que **descrevem** o grupo

OS_1
OS_2
OS_3

- Estatística Descritiva
- Análise Fatorial
- Métodos de Classificação

...

Abordagem Simbólica

Matriz de Dados

id	Sex	Age	Activity	Sleep	Cooking
1	1	2	1	2	1
2	2	2	3	3	1
3	1	2	1	4	1
4	1	1	1	2	1
5	1	1	1	4	1
6	2	3	3	4	1
7	1	2	1	4	1
8	2	2	1	4	1
9	1	2	1	2	1
10	2	1	3	2	1

Agrupada por:
sexo, idade e atividade

Matriz de Dados Simbólicos

	Sleep	Cooking
M 35-59 E	Scarce Particip. (0.60), Medium Particip. (0.30), High Particip. (0.10)	No Particip. (0.74), High Particip. (0.01), Medium Particip. (0.06), Scarce Part
F 35-59 I	Scarce Particip. (0.45), Medium Particip. (0.43), High Particip. (0.11)	No Particip. (0.02), High Particip. (0.74), Medium Particip. (0.21), Scarce Part
M < 35 E	Scarce Particip. (0.62), Medium Particip. (0.27), High Particip. (0.11)	No Particip. (0.86), High Particip. (0.01), Medium Particip. (0.09), Scarce Part
F > 59 I	Scarce Particip. (0.19), Medium Particip. (0.37), High Particip. (0.44)	No Particip. (0.09), High Particip. (0.45), Medium Particip. (0.35), Scarce Part
F < 35 I	Scarce Particip. (0.64), Medium Particip. (0.30), High Particip. (0.06)	No Particip. (0.29), High Particip. (0.29), Medium Particip. (0.20), Scarce Part
M > 59 I	Scarce Particip. (0.16), Medium Particip. (0.26), High Particip. (0.58)	No Particip. (0.67), High Particip. (0.04), Medium Particip. (0.13), Scarce Part
F < 35 E	Scarce Particip. (0.45), Medium Particip. (0.37), High Particip. (0.18)	No Particip. (0.22), High Particip. (0.08), Medium Particip. (0.24), Scarce Part
F 35-59 E	Scarce Particip. (0.71), Medium Particip. (0.24), High Particip. (0.06)	No Particip. (0.09), High Particip. (0.15), Medium Particip. (0.41), Scarce Part

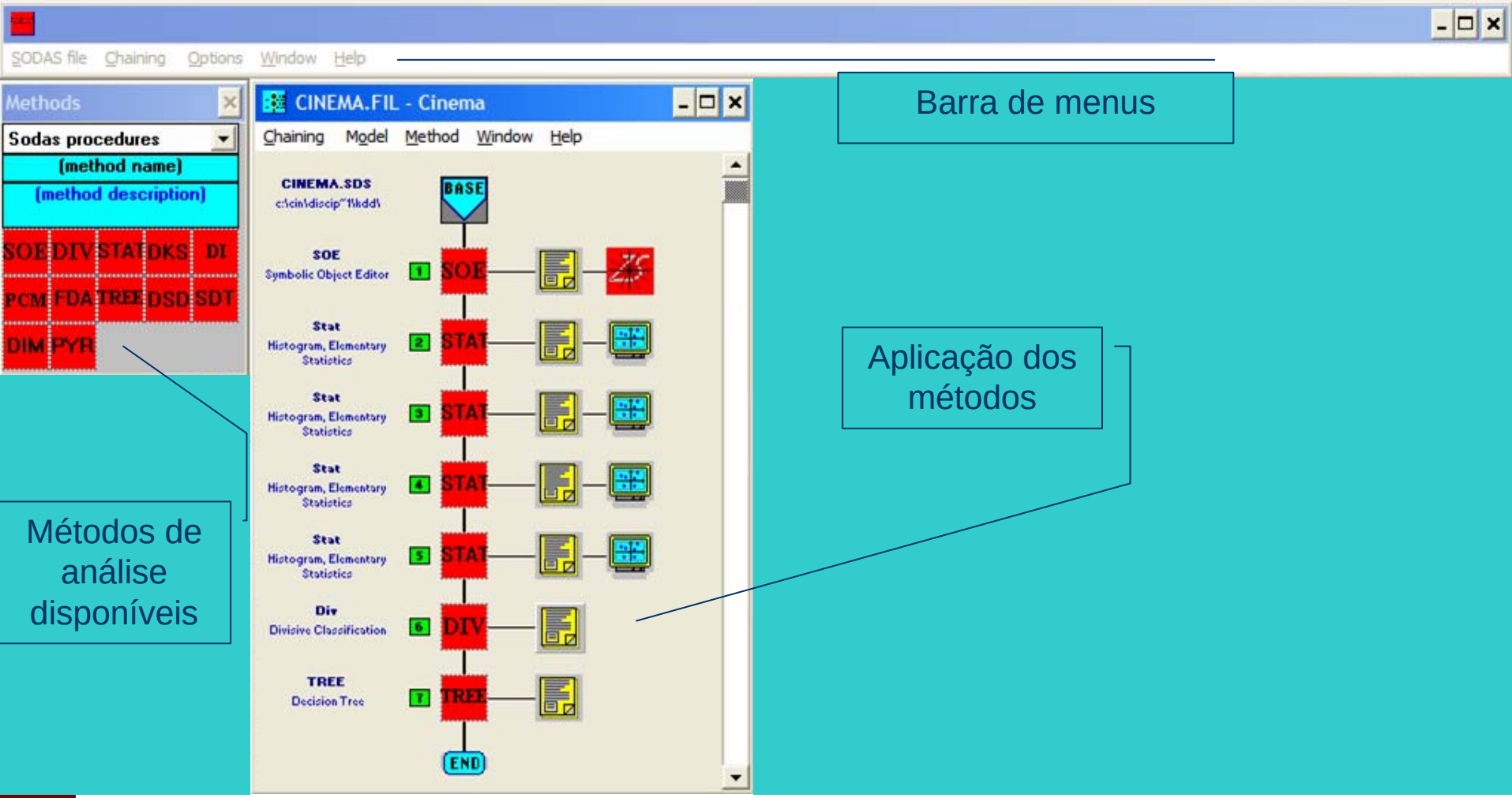
Fontes de Dados Simbólicos

- A partir de variáveis categóricas
  - Como "*tipo do empregado*"
  - Obtido por clusterização
- **De bancos de dados**
  - Consultas originando novas variáveis
- Do conhecimento do especialista
- De dados confidenciais
  - Para esconder informações privadas. Ex: IBGE
- De dados estocásticos
  - Distribuição de probabilidade
- De séries temporais
  - Descrevendo intervalos de tempo

SODAS – Symbolic Official Data Analysis System

- Protótipo disponível gratuitamente
  - http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm
- Funcionalidades
  - Construção de tabelas de dados simbólicos a partir de BD's tradicionais
  - Descrição de regras e hierarquias
  - Análise dos através de métodos de análise de dados simbólicos
    - Estatística descritiva
    - Análise Fatorial
    - Agrupamento
    - Árvore de Decisão
    - ...

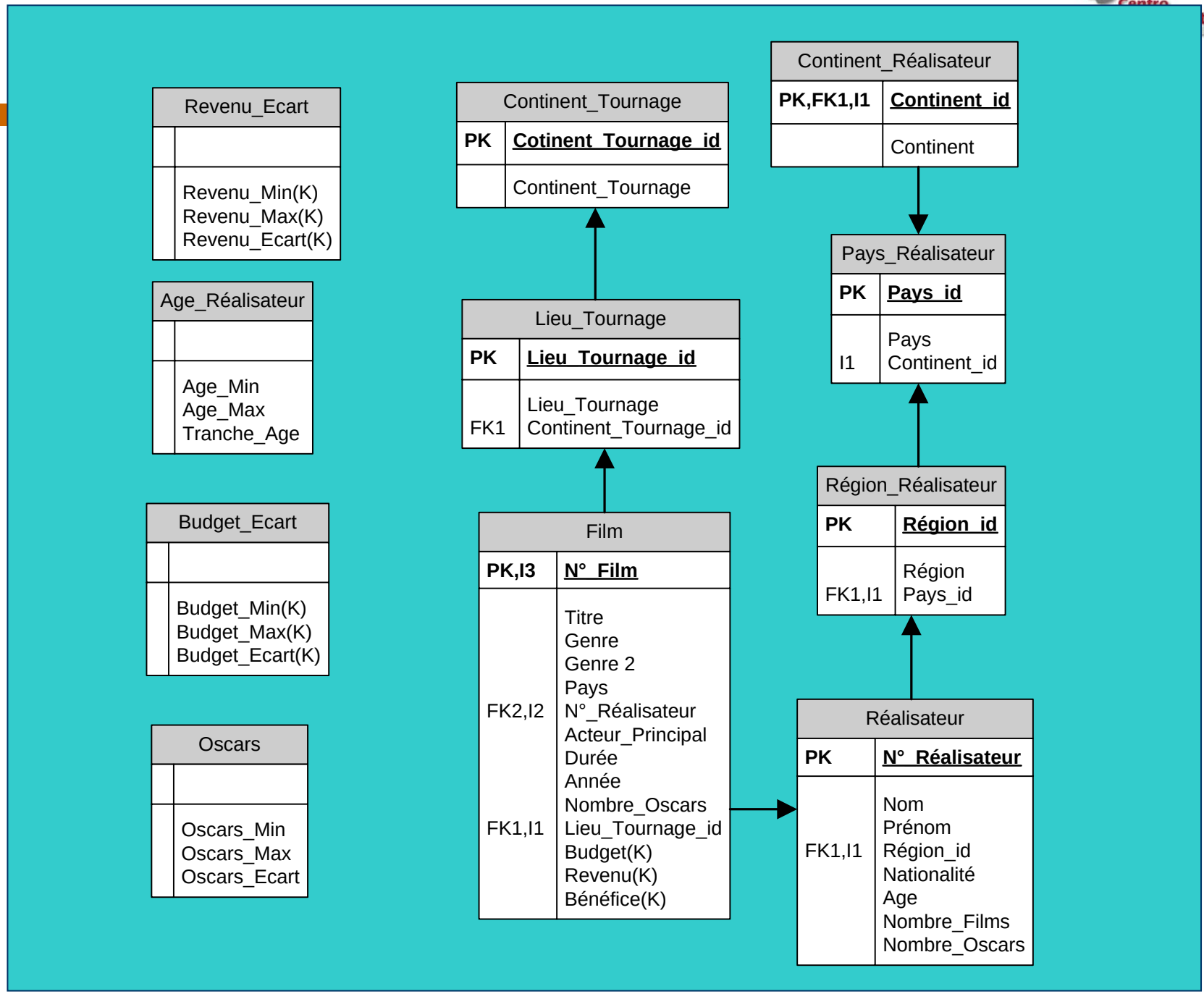
Interface do SODAS



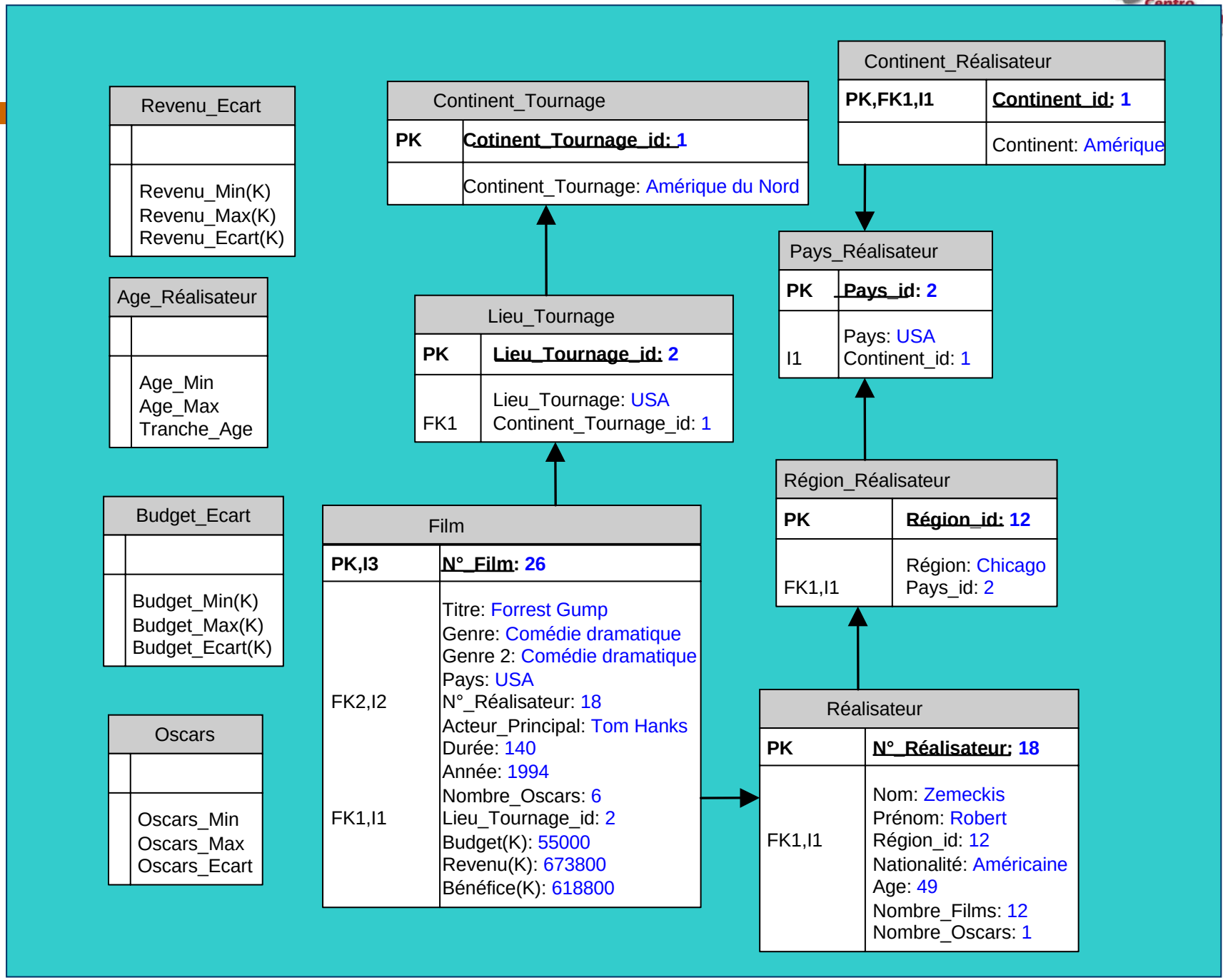
Análise de dados com o SODAS

- 1° Passo:
  - Possuir os dados em um BD relacional
  - Exemplo:
    - Dados sobre os filmes indicados ao Oscar.
      - Nome do filme, ator principal, diretor, ano, local de produção....

BD de Films: Esquema Relacional



Exemplo de indivíduo: Forrest Gump

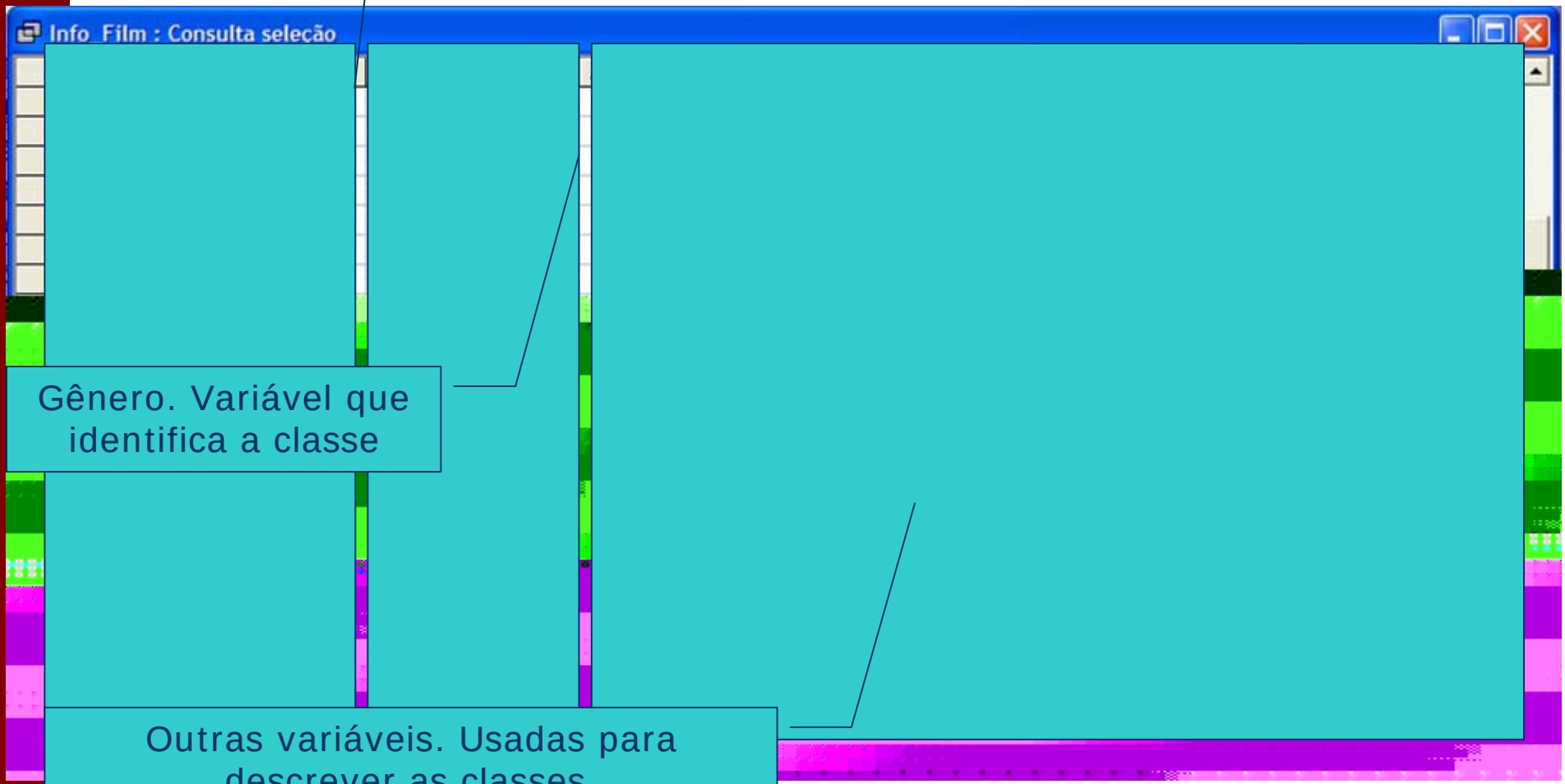


Definição do contexto

- Indivíduos que deseja-se estudar:
  - Categorias de filmes
    - Cada gênero de filme representa uma categoria:
      - Drama, Ficção científica, Ação, Western, ...
  - A descrição de cada classe será obtida através da generalização de seus membros de 1a. Ordem

Generalização dos dados

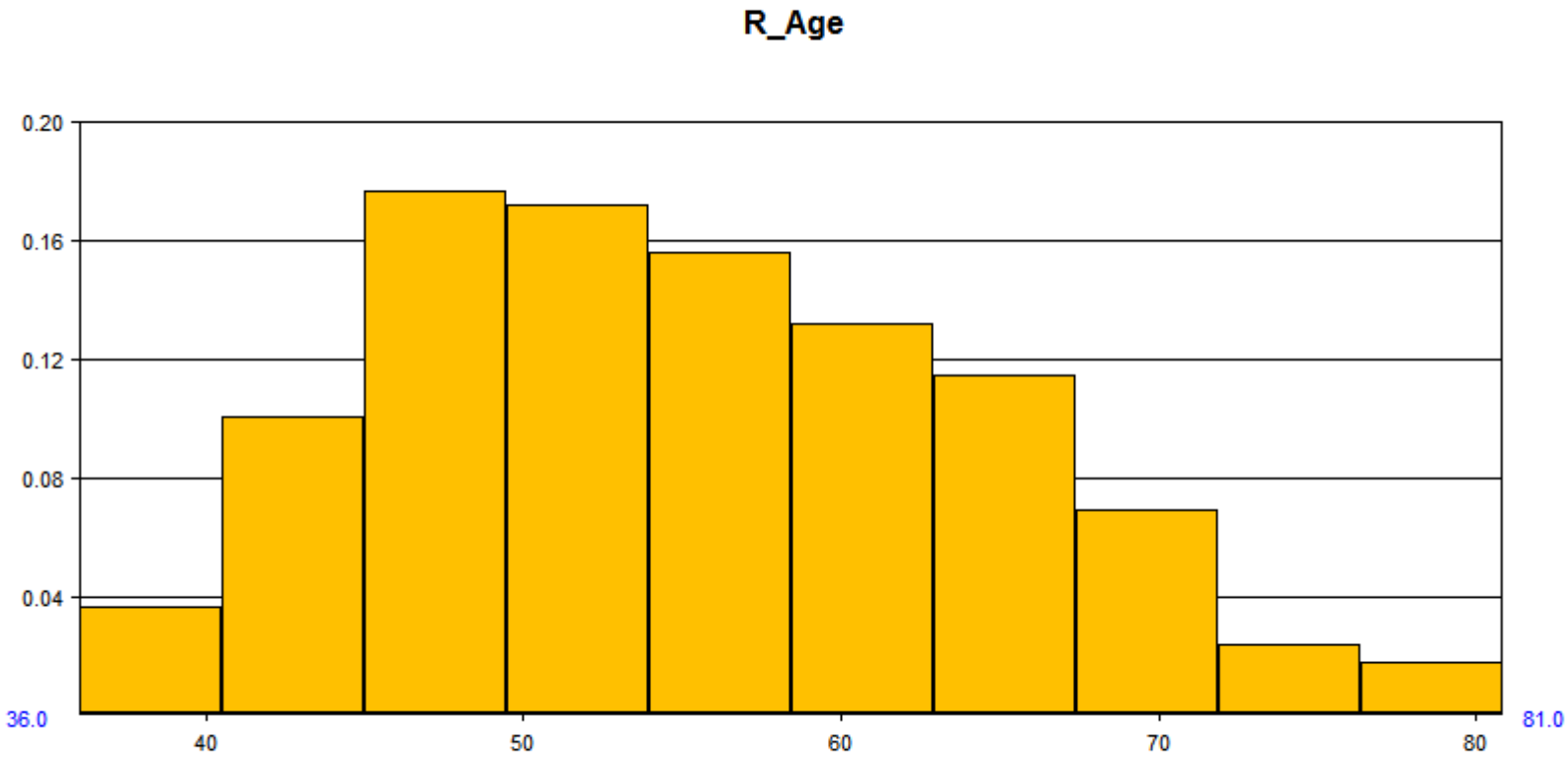
Nome do filme.
Será desprezado



Gênero. Variável que identifica a classe

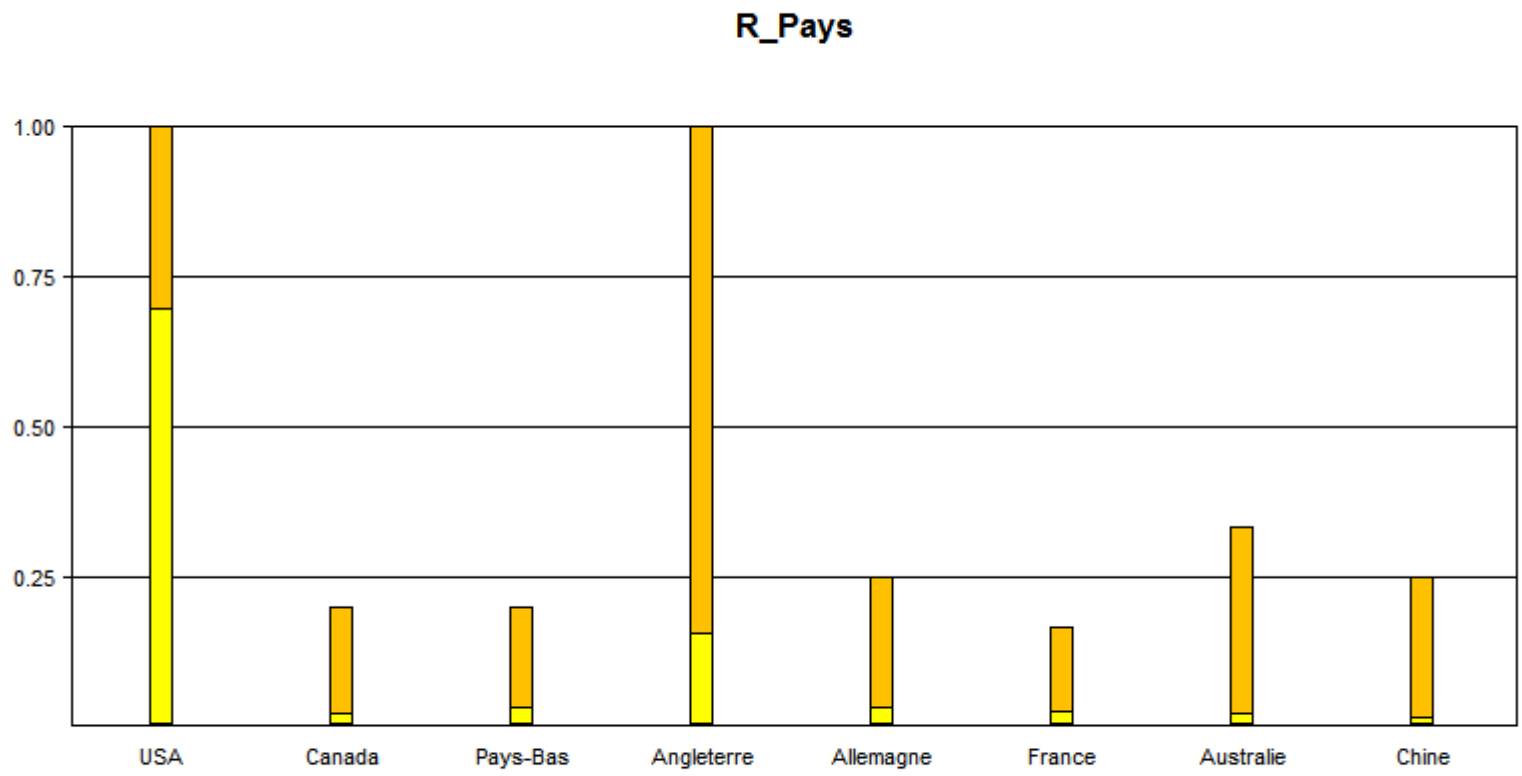
Outras variáveis. Usadas para descrever as classes.

Cinéma



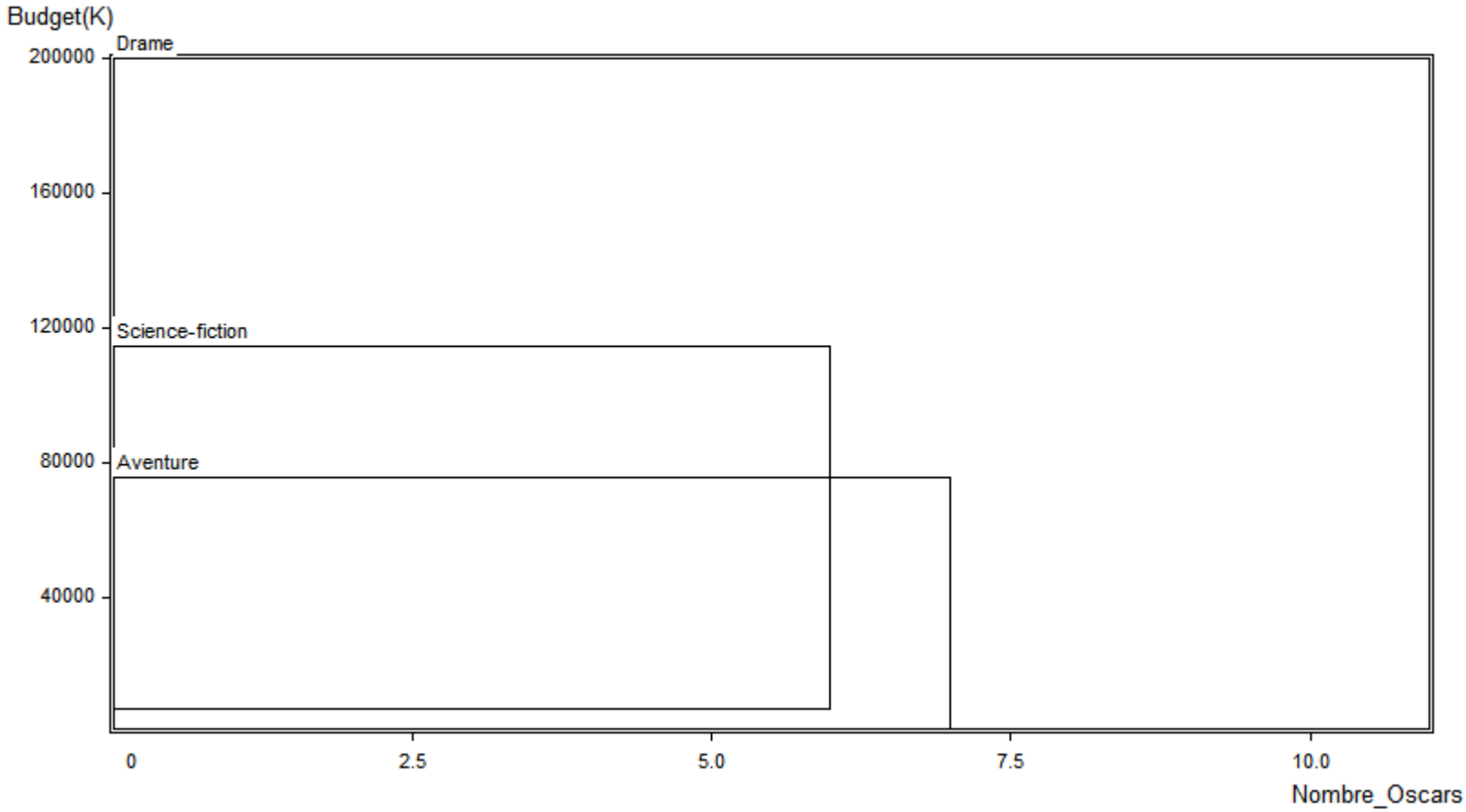
Relative frequencies (interval) - file CINEMA.SDS - variable 'R_Age' - 10 classes - class width: 4.5

Cinéma



Min / mean / max - file CINEMA.SDS - variable 'R_Pays'

Cinéma


Budget(K)
200000
160000
120000
80000
40000
Drame
Science-fiction
Aventure
0
2.5
5.0
7.5
10.0
Nombre_Oscars

DIV – Divisive Clustering

- Método de agrupamento hierárquico
- Divisivo
- Usuário informa número de classes desejado

PARTITION IN 5 CLUSTERS :
-------------------------:

Cluster 1 (n=2) :
"Western" "Espionnage"

Cluster 2 (n=4) :
"Science-fiction" "Fantastique"
"Comédie fantastique" "Film de guerre"

Cluster 3 (n=1) :
"Catastrophe"

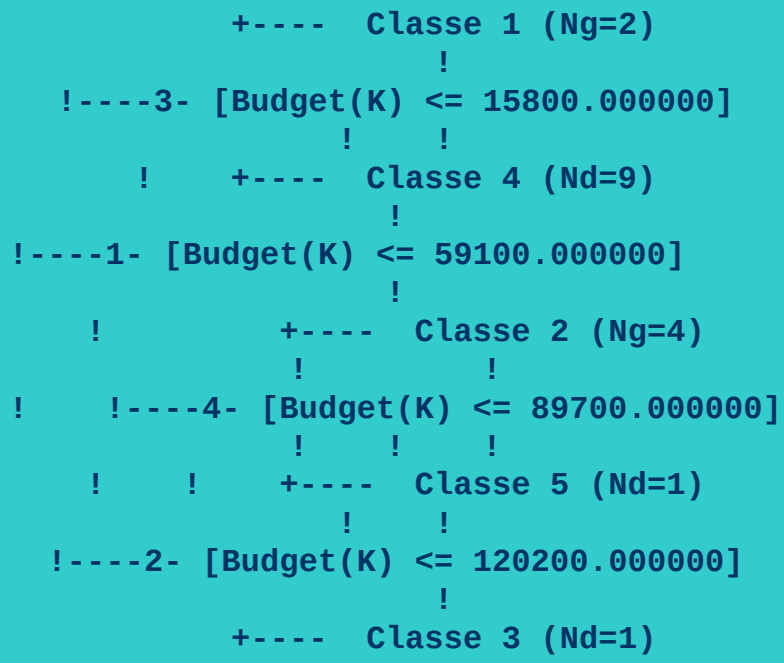
Cluster 4 (n=9) :
"Suspense" "Comédie dramatique"
"Film historique" "Policier"
"Action" "Animation"
"Comédie" "Aventure" "Comédie musicale"

Cluster 5 (n=1) :
"Drame"

Explicated inertia : 93.120786

THE CLUSTERING TREE :

- the number noted at each node indicates the order of the divisions
- Ng <-> yes and Nd <-> no



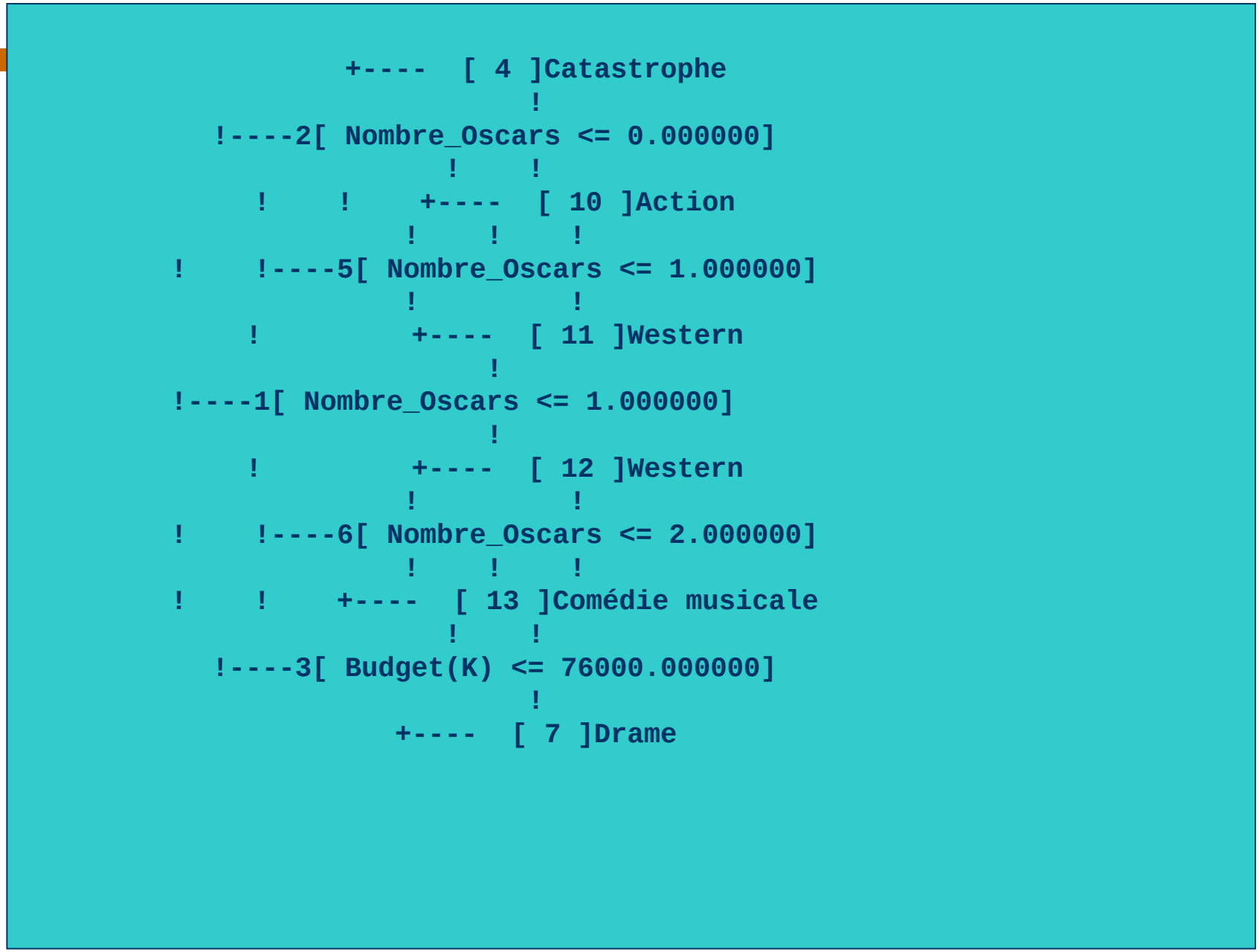
TREE – Árvores de decisão

- Fuzzy ou “Hard”
- Variável identificadora de classe
  - Nominal
- Variáveis descritoras
  - Todos os tipos

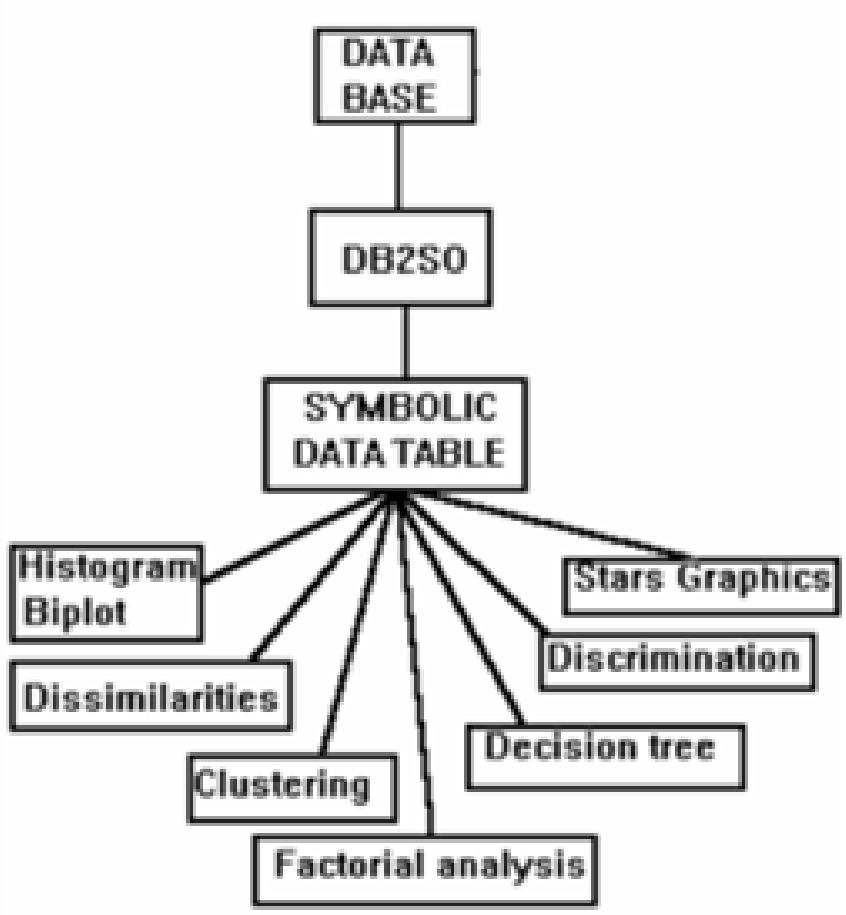
```
==================================
| EDITION OF DECISION TREE       |
==================================


              PARAMETERS :
        Learning Set         :    17
        Number of variables  :     2
        Max. number of nodes:    11
     Soft Assign           : ( 1 ) FUZZY
 Criterion coding     : ( 3 ) LOG-LIKELIHOOD
Min. number of object by node    :        5
Min. size of no-majority classes :        2
Min. size of descendant nodes    :     1.00
Frequency of test set            :     0.00

       + --- IF ASSERTION IS TRUE  (up)
                     !
           --- x [ ASSERTION ]
                     !
       + --- IF ASSERTION IS FALSE (down)
```


+---- [4]Catastrophe
!----2[Nombre_Oscars <= 0.000000]
+---- [10]Action
!----5[Nombre_Oscars <= 1.000000]
+---- [11]Western
!----1[Nombre_Oscars <= 1.000000]
+---- [12]Western
!----6[Nombre_Oscars <= 2.000000]
+---- [13]Comédie musicale
!----3[Budget(K) <= 76000.000000]
+---- [7]Drame

SODAS: Resumo



Análise de Dados Simbólicos
Geração e Exportação de SOs no SODAS

Aluno:
Anderson Costa
afbfc@cin.ufpe.br