UNIVERSIDADE FEDERAL DO CEARÁ

Técnicas de Anonimização

Prof. Javam Machado

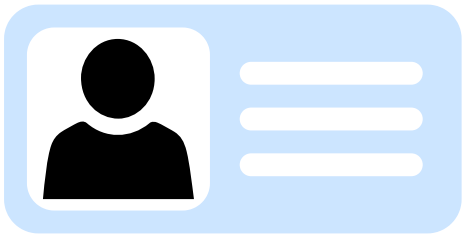
LSBD/DC/UFC

Agosto/2019

Na Aula Passada...

Anonimização

- Maneira mais promissora para a disponibilização de dados
- Conjunto de técnicas que modificam os dados originais
- Anônimo significa não ser caracterizado unicamente



?

Apenas Remover Identificadores...

... é suficiente para garantir a privacidade de indivíduos?

Nome	Idade	Gênero	Endereço	Telefone	Conta	Saldo (R$)
David C. F.	22	M	Av. L	98533 1234	2234-0	1.033,25
John T. A.	23	M	Av. K	98772 2531	7749-2	814,92
Helton B.	25	M	Av. K	98156 0092	8491-7	515,09
Maria L. Jr	32	F	Rua J	99913 9026	5723-1	2.194,79

Apenas Remover Identificadores...

... é suficiente para garantir a privacidade de indivíduos?

Idade	Gênero	Endereço	Telefone	Conta	Saldo (R$)
22	M	Av. L	98533 1234	2234-0	1.033,25
23	M	Av. K	98772 2531	7749-2	814,92
25	M	Av. K	98156 0092	8491-7	515,09
32	F	Rua J	99913 9026	5723-1	2.194,79

Apenas Remover Identificadores...

... é suficiente para garantir a privacidade de indivíduos?

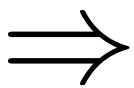
Idade	Gênero	Endereço	Telefone	Conta	Saldo (R$)
22	M	Av. L	98533 1234	2234-0	1.033,25
23	M	Av. K	98772 2531	7749-2	814,92
25	M	Av. K	98156 0092	8491-7	515,09
32	F	Rua J	99913 9026	5723-1	2.194,79

Apenas Remover Identificadores...

... é suficiente para garantir a privacidade de indivíduos?

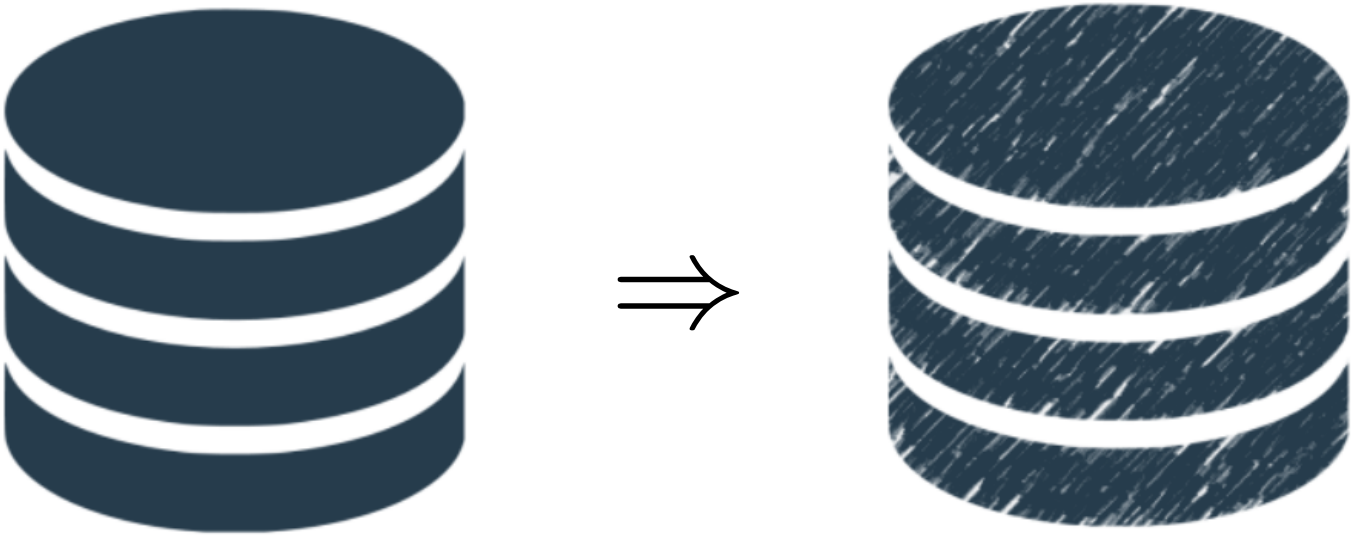
Idade	Gênero	Endereço	Telefone	Conta	Saldo (R$)
22	M	Av. L	98533 1234	2234-0	1.033,25
23	M	Av. K	98772 2531	7749-2	814,92
25	M	Av. K	98156 0092	8491-7	515,09
32	F	Rua J	99913 9026	5723-1	2.194,79

Idade	Gênero	Endereço	Telefone
22	M	Av. L	98533 1234



Nome
David C. F.

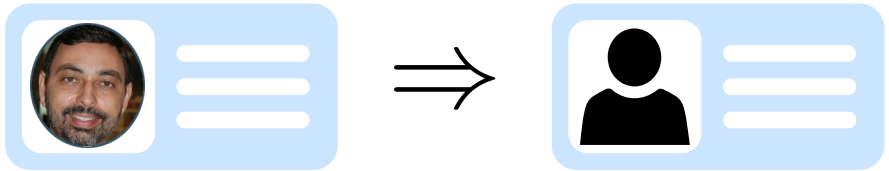
Objetivo da Anonimização



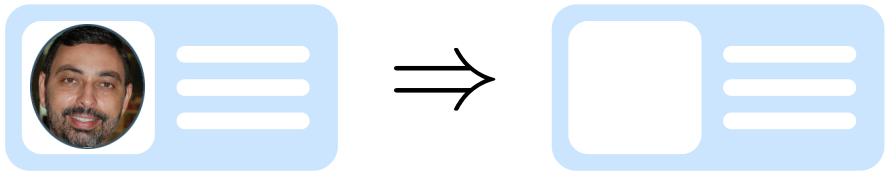
D D'

Técnicas de Anonimização

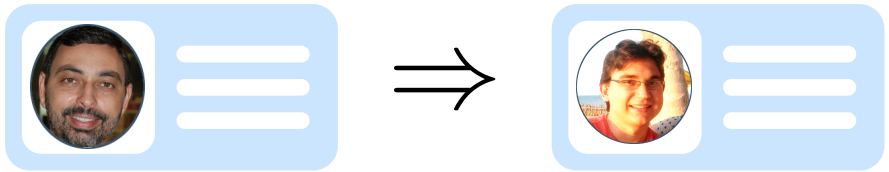
- Generalização



- Supressão



- Perturbação



Generalização

Substituição de valores dos atributos **semi-identificadores** por valores semanticamente semelhantes, porém menos específicos

Idade
22
23
25
32

$\Rightarrow$

Idade
[22-24]
[20-24]
[25-29]
[30-34]

Nascimento
13/08/1994
07/12/1993
20/01/1992
02/03/1985

$\Rightarrow$

Nascimento
08/1994
12/1993
01/1992
03/1985

Atributos e Suas Categorias

Atributos Numéricos

Podem ser dimensionados e representados pro números

- Idade
- Altura
- Peso
- Conta
- Salário

Atributos Categóricos

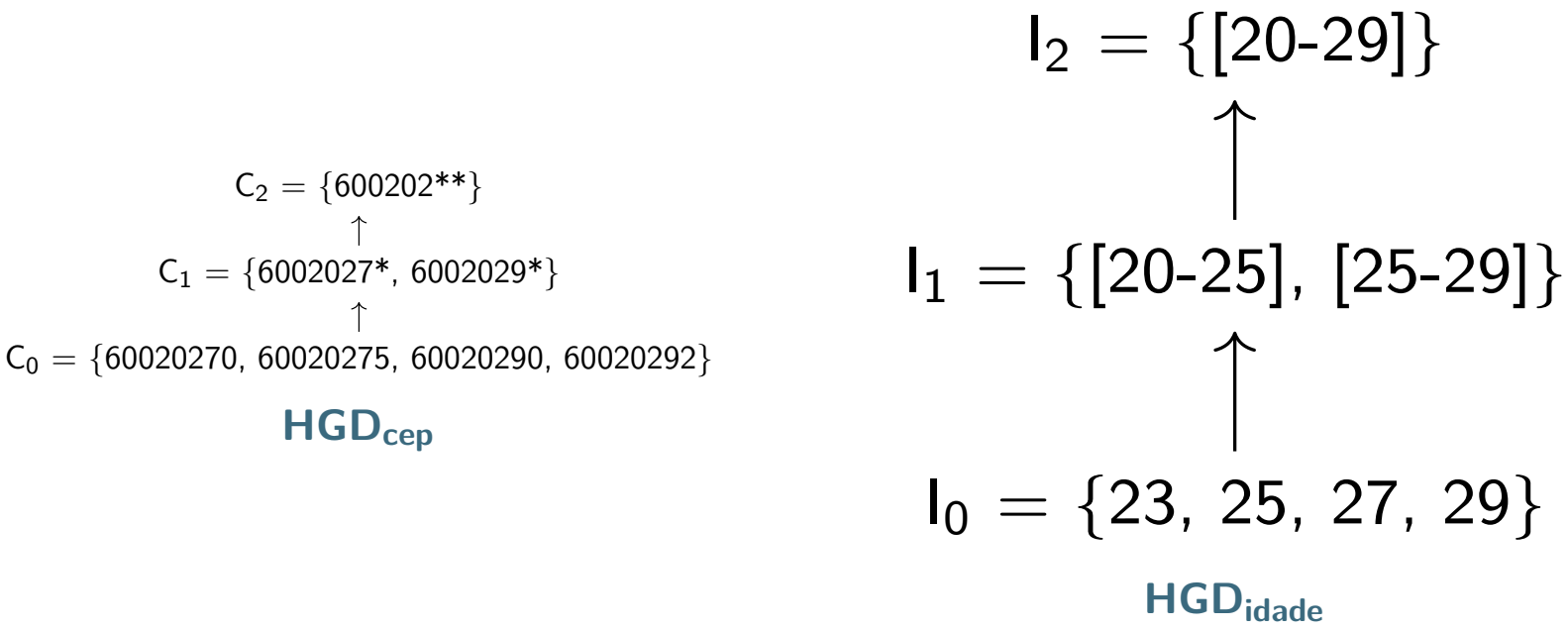
Podem assumir apenas uma quantidade finita de valores

- Gênero
- Profissão
- Nacionalidade
- Endereço

Hierarquia e Generalização

Representa a **semântica** dos atributos

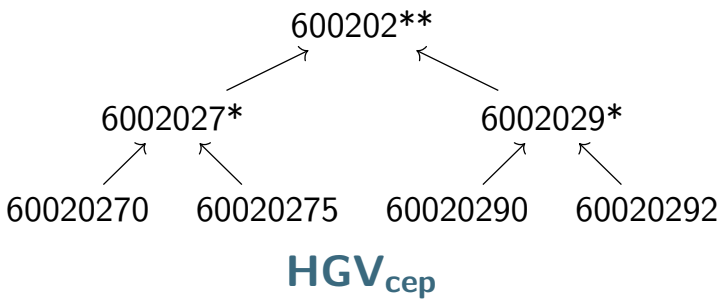
Hierarquia de Generalização de Domínio (HGD)



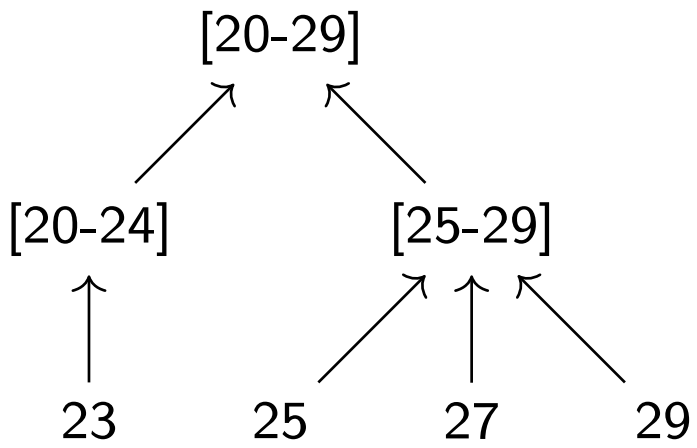
Hierarquia e Generalização

Representa a **semântica** dos atributos

Hierarquia de Generalização de Valor (HGV)

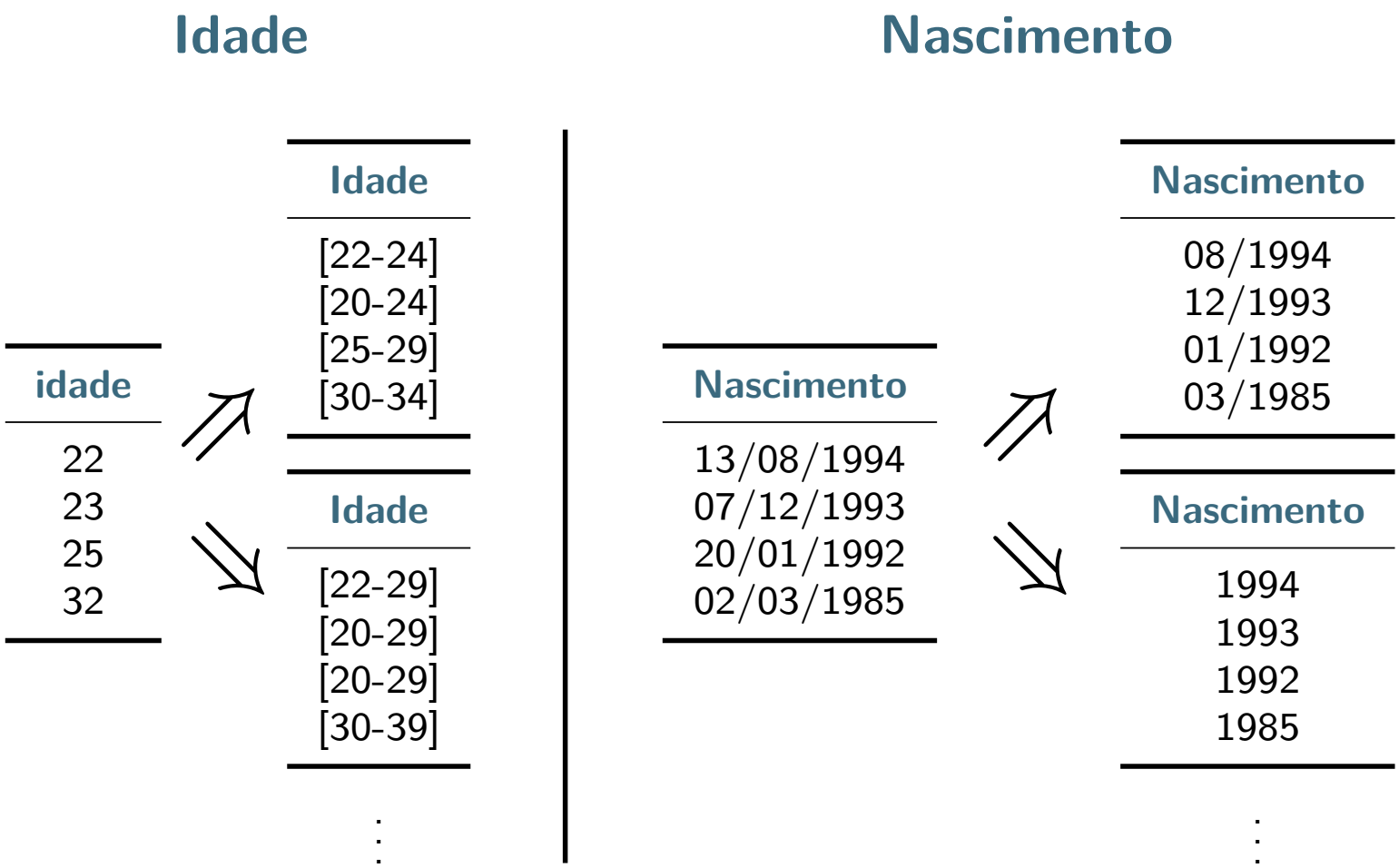


$\mathbf{HGV_{cep}}$



$\mathbf{HGV_{idade}}$

Qual a Melhor Generalização Possível?



Conjunto de Generalizações

$C_2 = \{600202^{**}\}$

↑

$C_1 = \{6002027^*, 6002029^*\}$

↑

$C_0 = \{60020270, 60020275, 60020290, 60020292\}$

HGD_{cep}

$<C_0, S_0, I_0> = ???$

$S_1 = \{Todos\}$

↑

$S_0 = \{Masculino, Feminino\}$

HGD_{sexo}

$I_2 = \{[20\text{-}29]\}$

↑

$I_1 = \{[20\text{-}25], [25\text{-}29]\}$

↑

$I_0 = \{23, 25, 27, 29\}$

HGD_{idade}

Conjunto de Generalizações

$C_2 = \{600202^{**}\}$

↑

$C_1 = \{6002027^*, 6002029^*\}$

↑

$C_0 = \{60020270, 60020275, 60020290, 60020292\}$

HGD_{cep}

$S_1 = \{Todos\}$

↑

$S_0 = \{Masculino, Feminino\}$

HGD_{sexo}

$I_2 = \{[20\text{-}29]\}$

↑

$I_1 = \{[20\text{-}25], [25\text{-}29]\}$

↑

$I_0 = \{23, 25, 27, 29\}$

HGD_{idade}

$<C_0, S_0, I_0> = ???$

Sem generalização de dados
(dados originais)

Conjunto de Generalizações

$C_2 = \{600202^{**}\}$

↑

$C_1 = \{6002027^*, 6002029^*\}$

↑

$C_0 = \{60020270, 60020275, 60020290, 60020292\}$

HGD_{cep}

$S_1 = \{Todos\}$

↑

$S_0 = \{Masculino, Feminino\}$

HGD_{sexo}

$I_2 = \{[20\text{-}29]\}$

↑

$I_1 = \{[20\text{-}25], [25\text{-}29]\}$

↑

$I_0 = \{23, 25, 27, 29\}$

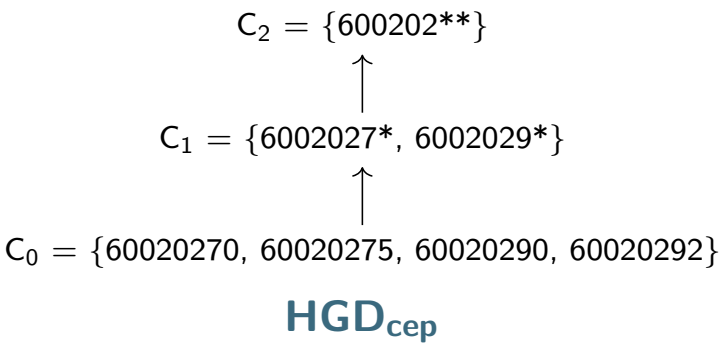
HGD_{idade}

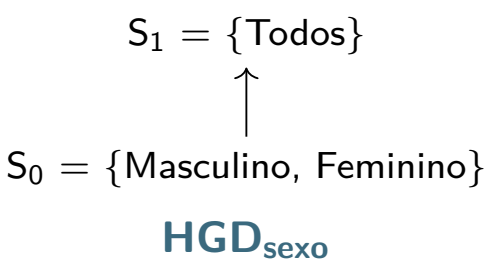
$<C_0,S_0,I_0> = ???$

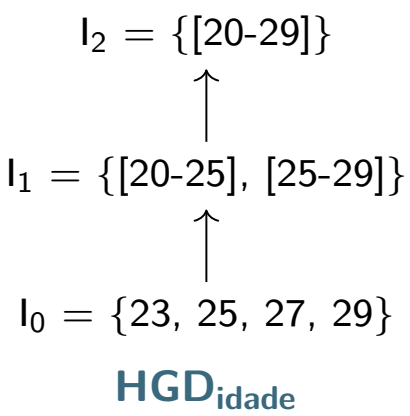
Sem generalização de dados
(dados originais)

$<C_2,S_1,I_2> = ???$

Conjunto de Generalizações







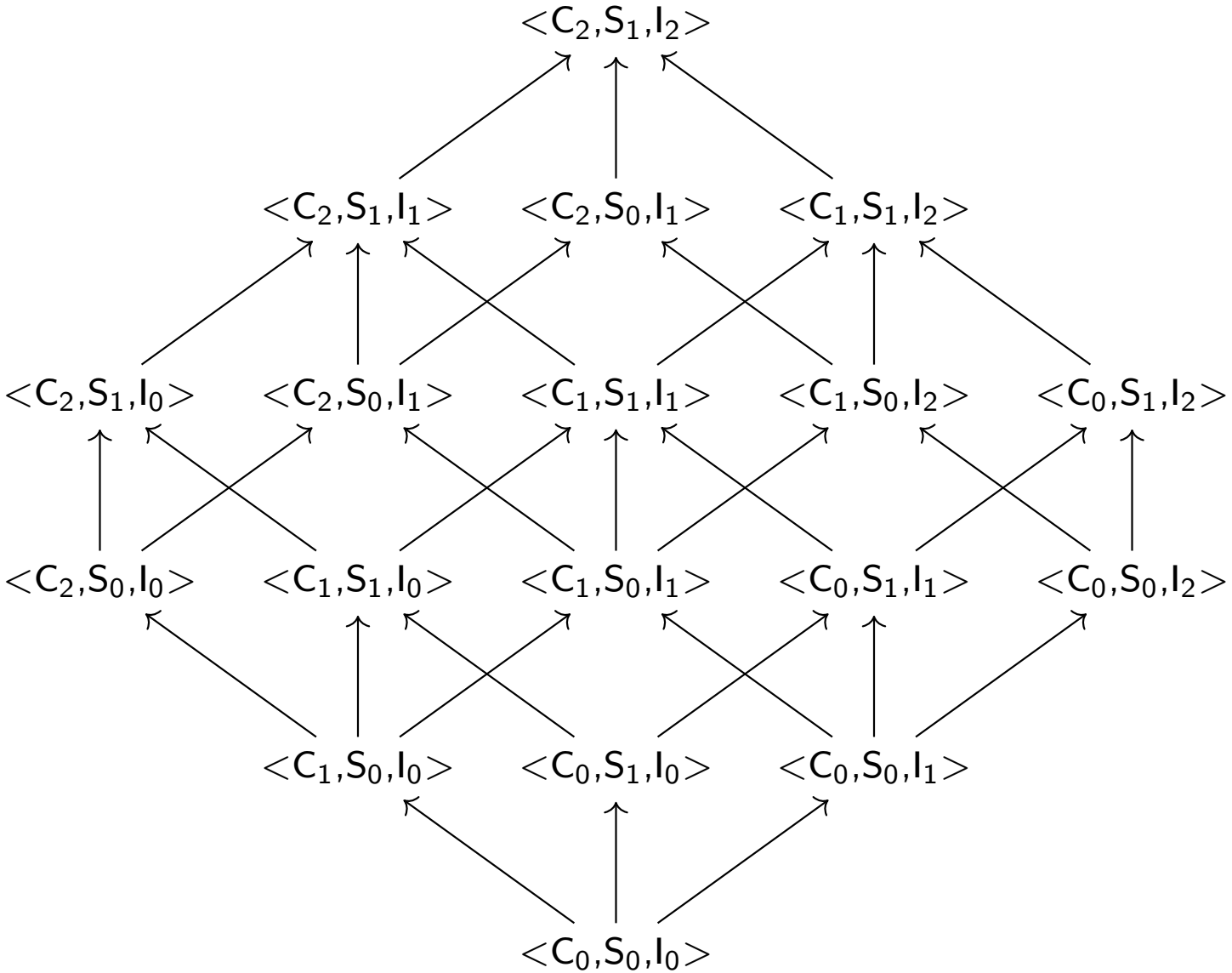
$<C_0,S_0,I_0> = ???$

Sem generalização de dados (dados originais)

$<C_2,S_1,I_2> = ???$

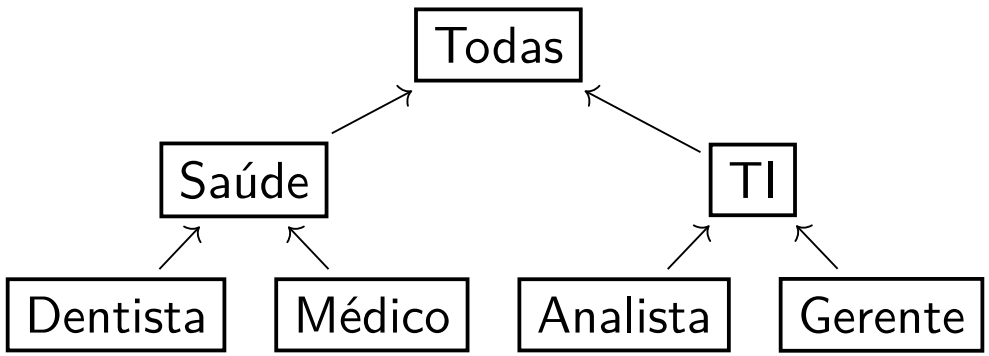
Maior generalização possível

Reticulado de Generalizações



Generalização de Domínio Completo

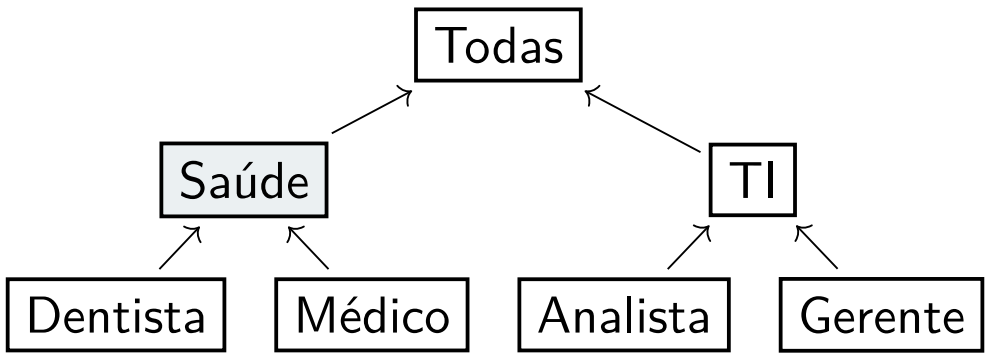
Todos os valores de um atributo **semi-identificador** são generalizados para o mesmo nível



Nome	Profissão
David	Analista
John	Analista
Bob	Médico
Maria	Dentista
Caio	Analista
Alan	Médico
Zeck	Gerente

Generalização de Domínio Completo

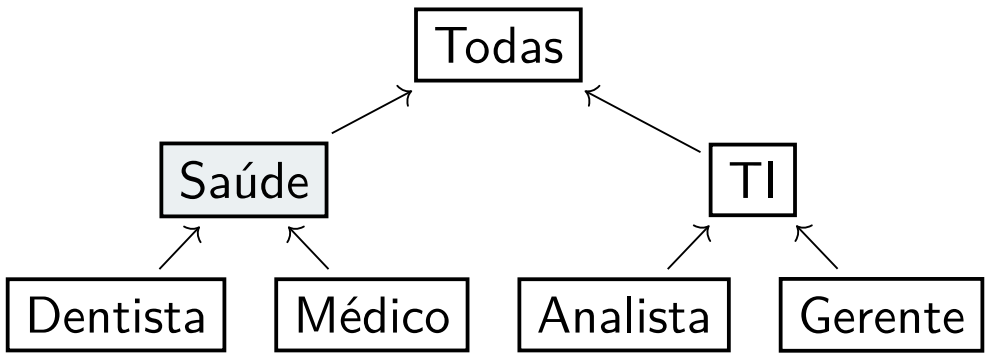
Todos os valores de um atributo **semi-identificador** são generalizados para o mesmo nível



Nome	Profissão
David	Analista
John	Analista
Bob	Médico
Maria	Dentista
Caio	Analista
Alan	Médico
Zeck	Gerente

Generalização de Domínio Completo

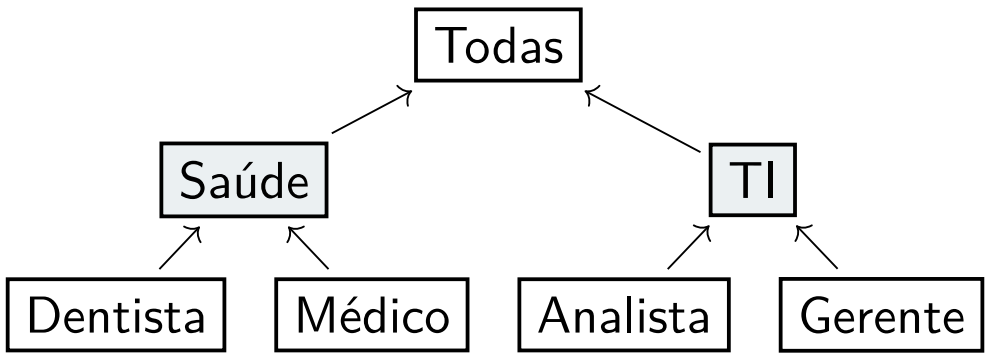
Todos os valores de um atributo **semi-identificador** são generalizados para o mesmo nível



Nome	Profissão
David	Analista
John	Analista
Bob	Saúde
Maria	Saúde
Caio	Analista
Alan	Saúde
Zeck	Gerente

Generalização de Domínio Completo

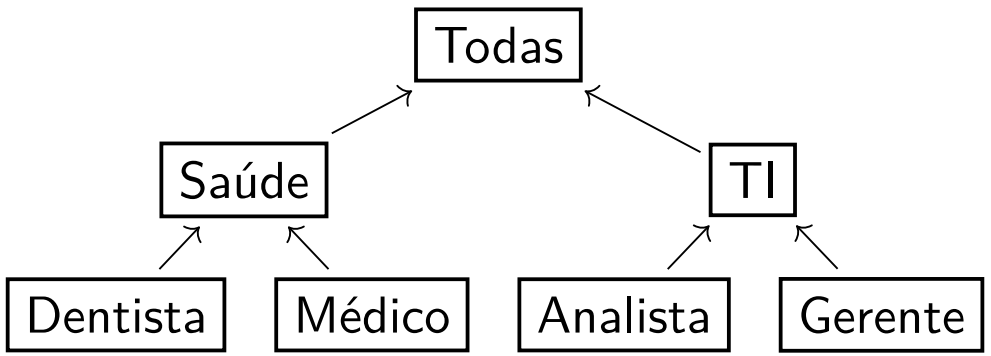
Todos os valores de um atributo **semi-identificador** são generalizados para o mesmo nível



Nome	Profissão
David	TI
John	TI
Bob	Saúde
Maria	Saúde
Caio	TI
Alan	Saúde
Zeck	TI

Generalização de Sub-árvore

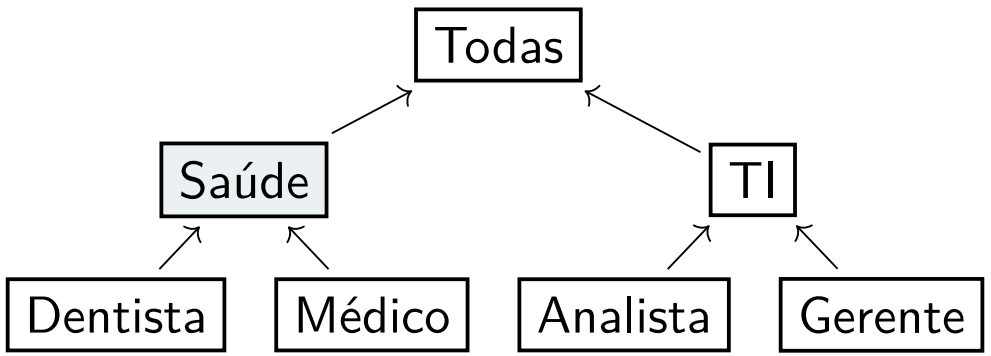
Todos os valores de um atributo **semi-identificador** de uma **sub-árvore** são generalizados para o mesmo nível



Nome	Profissão
David	Analista
John	Analista
Bob	Médico
Maria	Dentista
Caio	Analista
Alan	Médico
Zeck	Gerente

Generalização de Sub-árvore

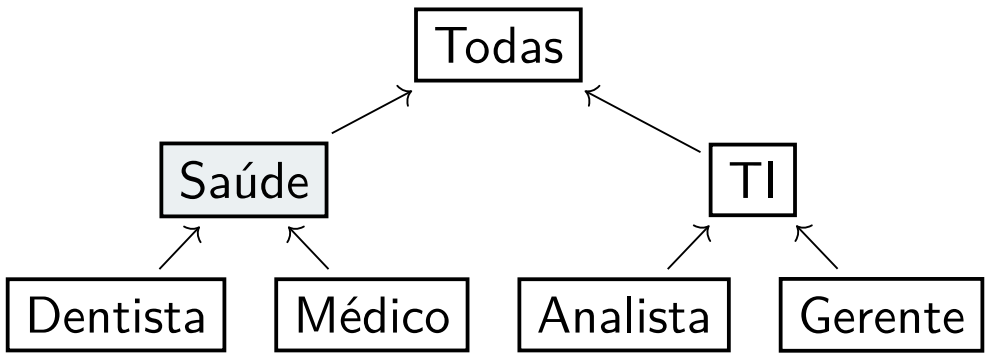
Todos os valores de um atributo **semi-identificador** de uma **sub-árvore** são generalizados para o mesmo nível



Nome	Profissão
David	Analista
John	Analista
Bob	Médico
Maria	Dentista
Caio	Analista
Alan	Médico
Zeck	Gerente

Generalização de Sub-árvore

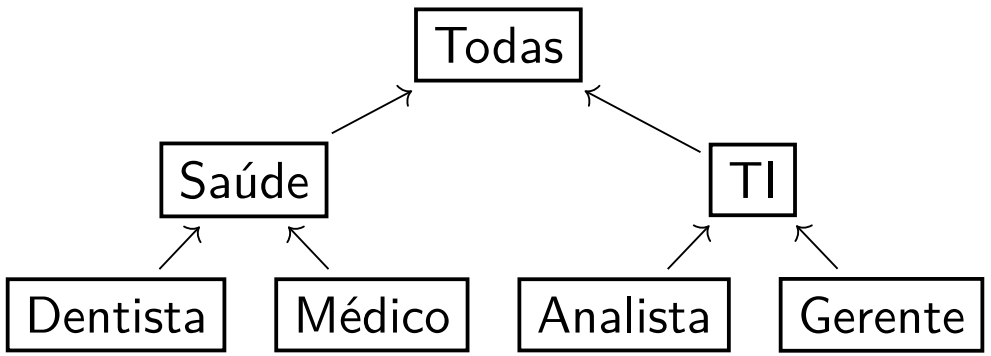
Todos os valores de um atributo **semi-identificador** de uma **sub-árvore** são generalizados para o mesmo nível



Nome	Profissão
David	Analista
John	Analista
Bob	Saúde
Maria	Saúde
Caio	Analista
Alan	Saúde
Zeck	Gerente

Generalização de Irmãos (*sibling*)

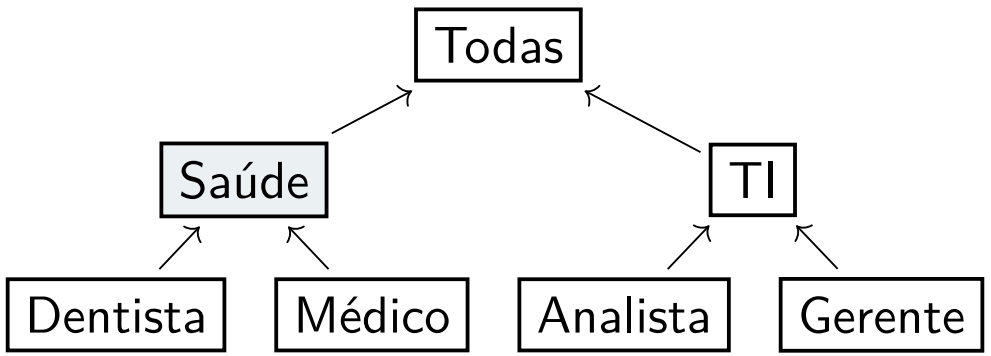
Não exige que todos os valores de **irmãos** nas folhas de uma sub-árvore sejam generalizados



Nome	Profissão
David	Analista
John	Analista
Bob	Médico
Maria	Dentista
Caio	Analista
Alan	Médico
Zeck	Gerente

Generalização de Irmãos (*sibling*)

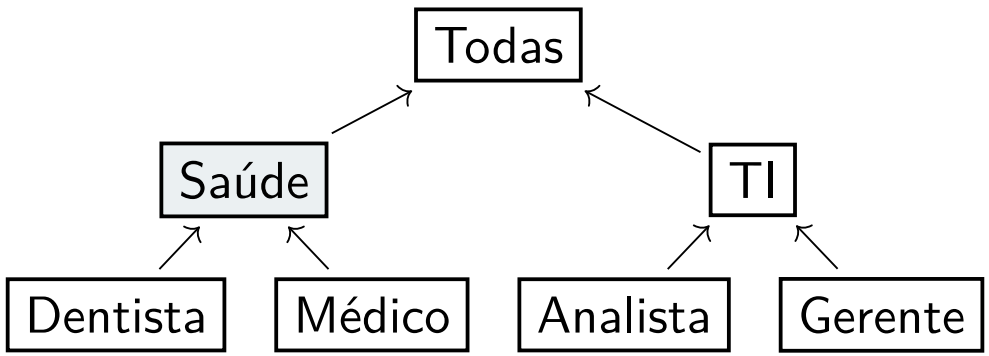
Não exige que todos os valores de **irmãos** nas folhas de uma sub-árvore sejam generalizados



Nome	Profissão
David	Analista
John	Analista
Bob	Médico
Maria	Dentista
Caio	Analista
Alan	Médico
Zeck	Gerente

Generalização de Irmãos (*sibling*)

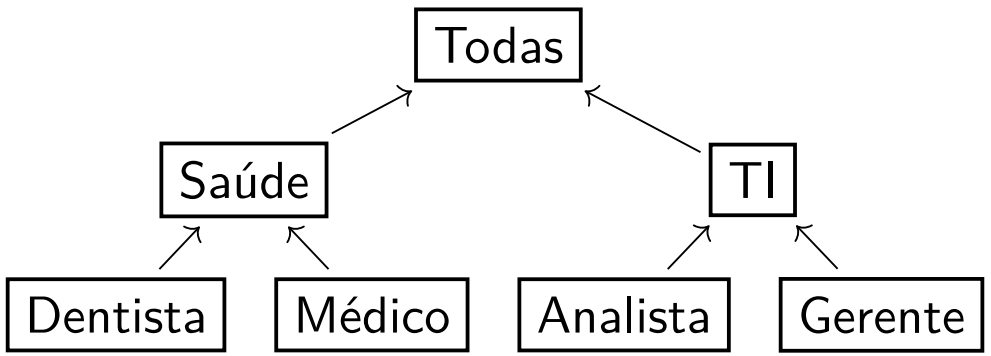
Não exige que todos os valores de **irmãos** nas folhas de uma sub-árvore sejam generalizados



Nome	Profissão
David	Analista
John	Analista
Bob	Saúde
Maria	Dentista
Caio	Analista
Alan	Saúde
Zeck	Gerente

Generalização de Células

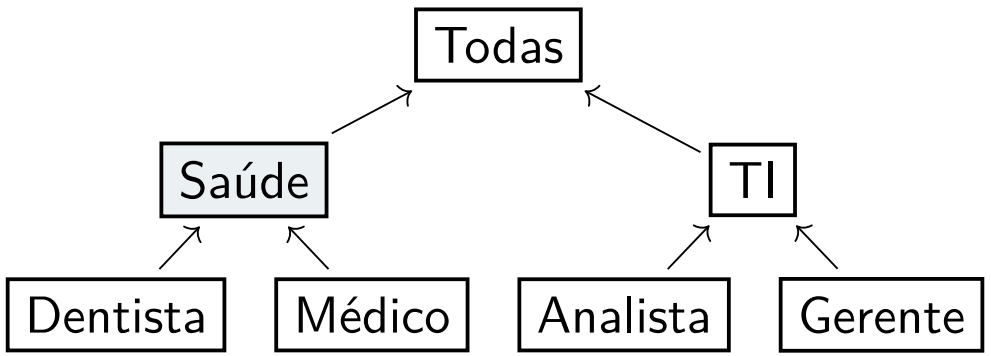
Apenas algumas instâncias de um atributo **semi-identificador** são generalizadas



Nome	Profissão
David	Analista
John	Analista
Bob	Médico
Maria	Dentista
Caio	Analista
Alan	Médico
Zeck	Gerente

Generalização de Células

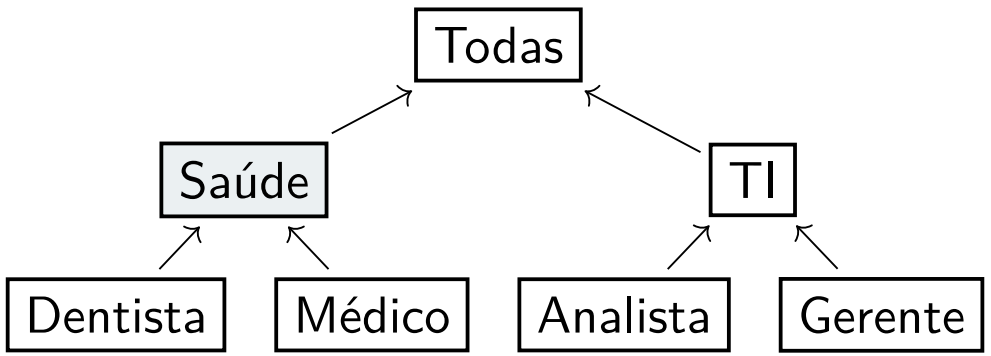
Apenas algumas instâncias de um atributo **semi-identificador** são generalizadas



Nome	Profissão
David	Analista
John	Analista
Bob	Médico
Maria	Dentista
Caio	Analista
Alan	Médico
Zeck	Gerente

Generalização de Células

Apenas algumas instâncias de um atributo **semi-identificador** são generalizadas



Nome	Profissão
David	Analista
John	Analista
Bob	Saúde
Maria	Dentista
Caio	Analista
Alan	Médico
Zeck	Gerente

Supressão

Um ou mais valores em um conjunto de dados são removidos ou substituídos por algum valor especial

Idade
22
23
25
32

⇒

Idade

Nascimento
13/08/1994
07/12/1993
20/01/1992
02/03/1985

⇒

Nascimento
*
*
*
*

Supressão de Registro

Um registro é removido inteiramente do conjunto de dados

Profissão	Salário (R$)
Analista	9.000,00
Analista	9.500,00
Dentista	16.500,00
Analista	7.500,00
Médico	23.000,00
Gerente	17.200,00

$\Rightarrow$

Profissão	Salário (R$)
*	*
*	*
Dentista	16.500,00
*	*
Médico	23.000,00
Gerente	17.200,00

Supressão de Valor

Remoção de todas as instâncias de um determinado valor (ou intervalo) de um atributo

Ex. Remoção Salário $<$ R$20.000,00

Profissão	Salário (R$)
Analista	9.000,00
Analista	9.500,00
Dentista	16.500,00
Analista	7.500,00
Médico	23.000,00
Gerente	17.200,00

$\Rightarrow$

Profissão	Salário (R$)
Analista	*
Analista	*
Dentista	*
Analista	*
Médico	23.000,00
Gerente	*

Supressão de Células

Apenas algumas instâncias de valores de um atributo são removidas

Profissão	Salário (R$)
Analista	9.000,00
Analista	9.500,00
Dentista	16.500,00
Analista	7.500,00
Médico	23.000,00
Gerente	17.200,00

$\Rightarrow$

Profissão	Salário (R$)
Analista	9.000,00
Analista	9.500,00
Dentista	16.500,00
Analista	*
Médico	23.000,00
Gerente	17.200,00

Perturbação

- Substituição de valores de atributos semi-identificadores originais por valores fictícios
- Informações estatísticas calculadas a partir dos dados originais não diferenciam significativamente de informações estatísticas calculadas anteriormente
- Não preserva a veracidade dos dados

Adição de Ruído

Consiste em substituir um valor original de atributo "v" por "v+r", onde "r" é um valor, denominado **ruído**

Profissão	Salário (R$)
Analista	9.000,00
Analista	9.500,00
Dentista	16.500,00
Analista	7.500,00
Médico	23.000,00
Gerente	17.200,00

$\Rightarrow$

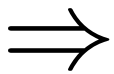
Profissão	Salário (R$)
Analista	9.100,00
Analista	9.600,00
Dentista	16.600,00
Analista	7.600,00
Médico	23.100,00
Gerente	17.300,00

r = R$100,00

Permutação de Dados

Consiste em permutar dois valores do mesmo atributo de registros diferentes

Profissão	Salário (R$)
Analista	9.000,00
Analista	9.500,00
Dentista	16.500,00
Analista	7.500,00
Médico	23.000,00
Gerente	17.200,00



Profissão	Salário (R$)
Analista	9.000,00
Analista	9.500,00
Dentista	17.200,00
Analista	7.500,00
Médico	23.000,00
Gerente	16.500,00

Geração de Dados Sintéticos

Consiste em duas etapas...

1. Gerar um modelo estatístico a partir do conjunto de dados
2. Gerar dados sintéticos a partir do modelo estatístico