



Licap

Formação Cientista de Dados



PUC Minas

Formação do Cientista de Dados

Montagem do conjunto de dados – Módulo Básico

Luis Enrique Zárate

Conteúdo do Curso



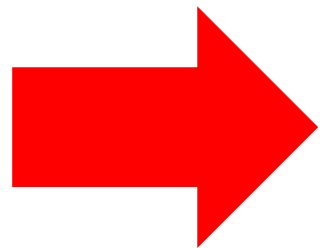
- Montagem da base de dados
 1. Fonte e origem dos dados
 2. Tipos de bases de dados: Estáticas, Temporais e Longitudinais
 3. Pivotagem reversa
 4. Enriquecimento dos dados
 5. Melhoramento dos dados
 6. Granulidade dos dados
 7. Realce de características
 8. Avaliação da representatividade da base de dados
 9. Análise de domínio dos atributos
 10. Incremento exponencial da amostra
 11. Expansão exponencial da amostra
 12. Distanciamento entre instâncias
 13. Problemas de indução
 14. Fluxo para a tomada de decisão

Montagem da Base de Dados



- ❑ O cientista de dados deve buscar em bases de dados disponíveis e fontes externas dados para os atributos vinculados ao domínio do problema considerado.

Algumas restrições de acesso a Dados



- ✓ Fatores Legais (ex. dados sensíveis)
- ✓ Fatores Departamentais (ex. setor financeiro)
- ✓ Razões Políticas (ex. restrições a dados públicos)
- ✓ Formato dos Dados (ex. junção de dados de bases distintas)
- ✓ Conectividade (ex. Bloqueio para uso de crawlers)
- ✓ Arquiteturas das Bases de Dados (ex. Junção de tecnologias antigas)

Alguns destes problemas são fortes restrições. Estas restrições devem ser consideradas para o prosseguimento do projeto de Ciência dos Dados.

Base de dados: Estáticas



$$[X] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix}$$

N: representa o número de instâncias, registros ou exemplos.

M: representa o número de atributos ou variáveis (numéricas ou categóricas, Ex. Sexo, idade, peso, estado civil,...).

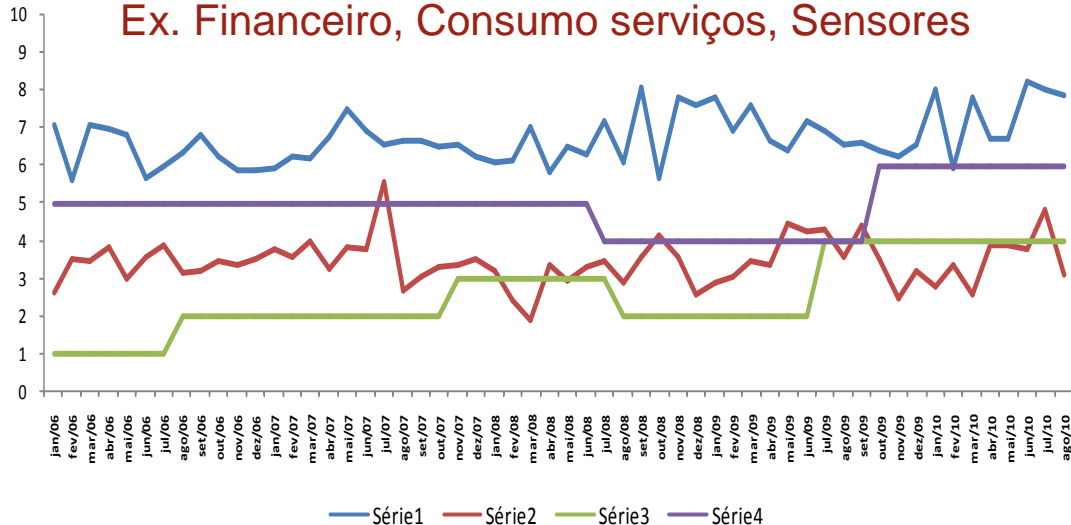
OBS: Qualquer base considerada Estática é intrinsecamente Temporal

Base de dados: Temporais



$$[Z] = \begin{bmatrix} Z_{t^{11}} & Z_{t^{12}} & \cdots & Z_{t^{1M}} \\ Z_{t^{21}} & Z_{t^{22}} & \cdots & Z_{t^{2M}} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{t^{N1}} & Z_{t^{N2}} & \cdots & Z_{t^{NM}} \end{bmatrix}$$

Ex. Financeiro, Consumo serviços, Sensores



N: representa o número de instâncias, registros ou exemplos.

M: representa o número de atributos ou variáveis (numéricas ou categóricas).

Cada elemento **Z_{tij}** corresponde a uma observação do registro **i** e atributo **j** .

Cada valor **$t=1...T$** corresponde ao período de observação na série

Base de dados: Temporais

$$[Z] = \begin{bmatrix} Z_{t^{11}} & Z_{t^{12}} & \dots & Z_{t^{1M}} \\ Z_{t^{21}} & Z_{t^{22}} & \dots & Z_{t^{2M}} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{t^{N1}} & Z_{t^{N2}} & \dots & Z_{t^{NM}} \end{bmatrix}$$



Feature Extraction

$$Z^*_{ijh} = [\bar{Z}_{ijh}, \hat{T}_{ijh}] \rightarrow [Z^*] =$$



Média e Tendência

$$[Z^*] = \begin{bmatrix} Z^*_{111} & Z^*_{121} & \dots & Z^*_{1M1} \\ Z^*_{211} & Z^*_{221} & \dots & Z^*_{2M1} \\ \vdots & \vdots & \ddots & \vdots \\ Z^*_{N11} & Z^*_{N21} & \dots & Z^*_{NM1} \\ Z^*_{112} & Z^*_{122} & \dots & Z^*_{1M2} \\ Z^*_{212} & Z^*_{222} & \dots & Z^*_{2M2} \\ \vdots & \vdots & \ddots & \vdots \\ Z^*_{N12} & Z^*_{N22} & \dots & Z^*_{NM2} \\ \ddots & \ddots & \ddots & \ddots \\ Z^*_{11H} & Z^*_{12H} & \dots & Z^*_{1MH} \\ Z^*_{21H} & Z^*_{22H} & \dots & Z^*_{2MH} \\ \vdots & \vdots & \ddots & \vdots \\ Z^*_{N1H} & Z^*_{N2H} & \dots & Z^*_{NMH} \end{bmatrix}$$

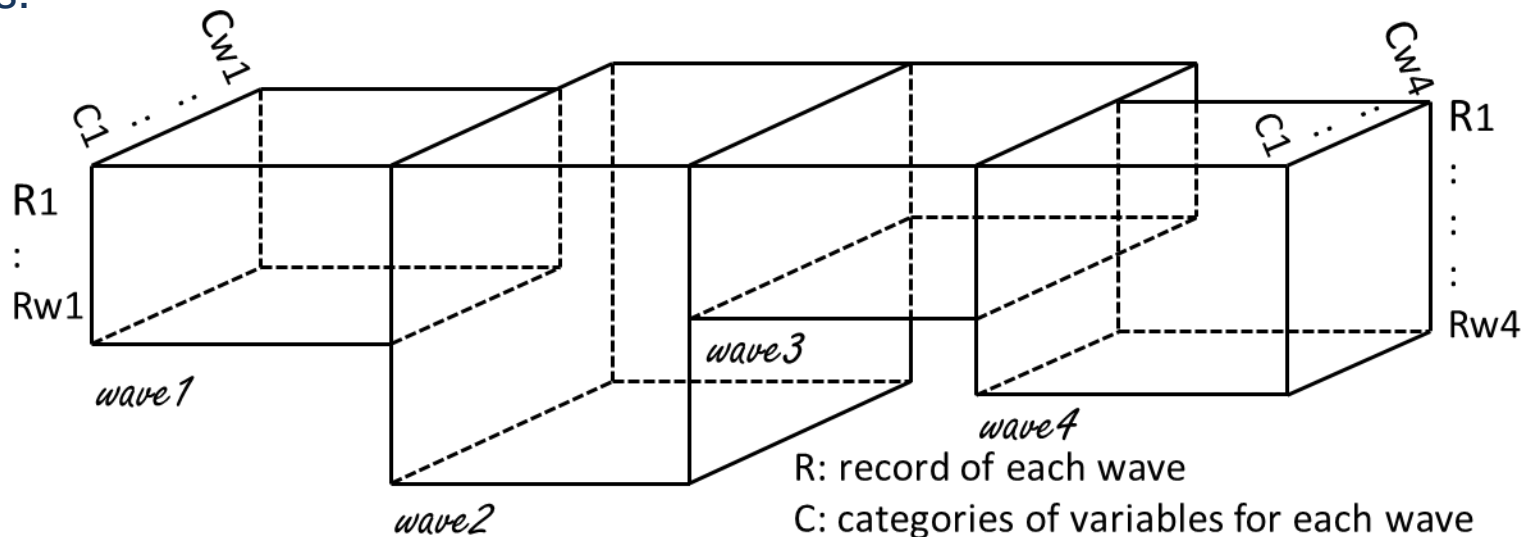
Jan1

.....

JanH

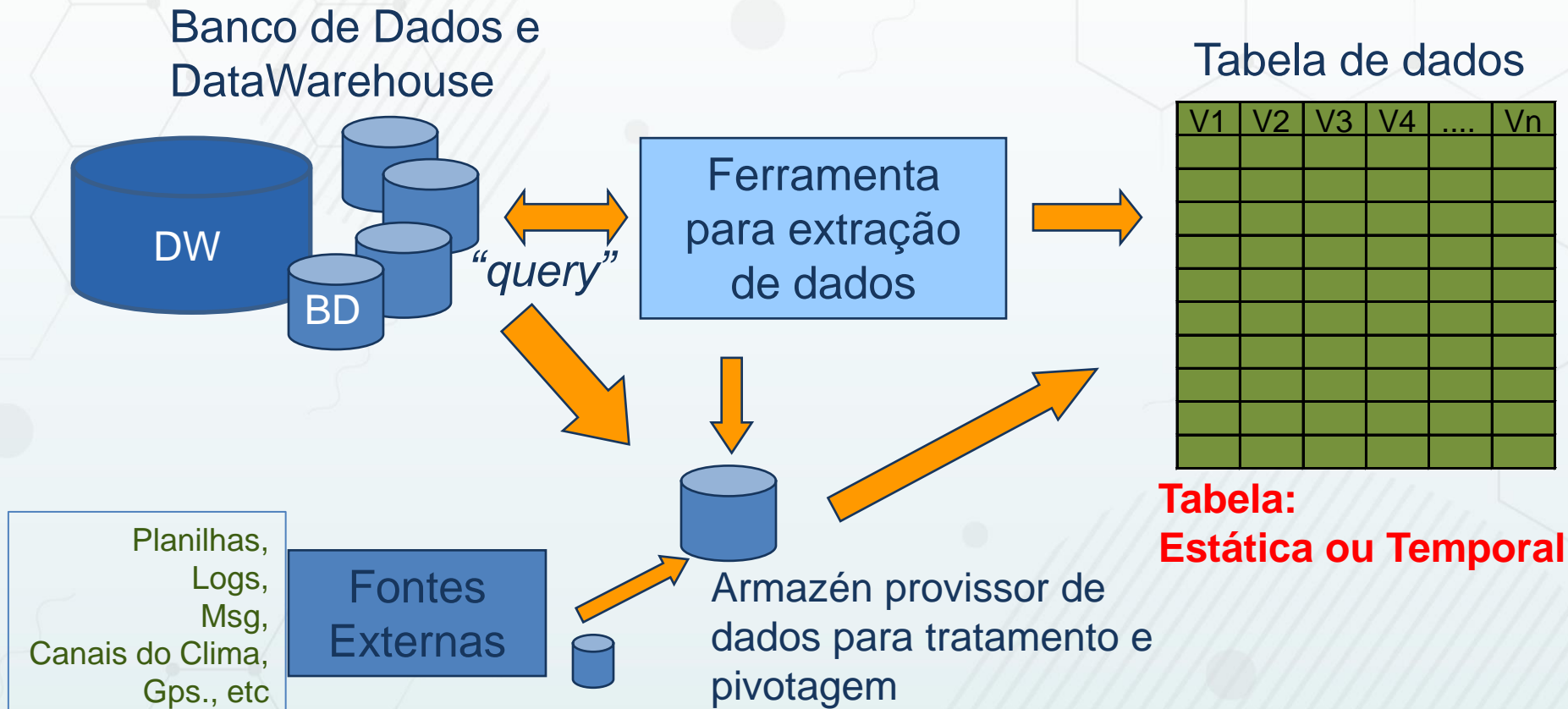
Base de dados: Longitudinais

Dados longitudinais é uma forma de dado temporal na qual a mesma amostra de registros é observada repetidamente em diferentes pontos de tempo chamadas ondas.



Ex. Estudos de longevidade, de doenças, sociais

Fonte e Origem dos Dados



Montagem da base de dados - Pivotagem reversa



- Para montar um conjunto de dados para o projeto possivelmente será necessário aplicar um processo de pivotagem.
- A pivotagem é necessária para construir um conjunto de dados consistente.

PIVOTAGEM REVERSA

Registro de transações de Supermercado => Registro de consumo

Registro de transações

| ID Trans | Data | ID Prod | Quant. |
|----------|------|---------|--------|
| X | | | |
| X | | | |
| Y | | | |
| X | | | |

Cliente X
Cliente Y
Cliente Z

| | | | | | | | | | |
|--|--|--|--|--|--|--|--|--|--|
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |




| | | | |
|----|----|----|----|
| T1 | T2 | T3 | T4 |
| Q1 | Q2 | Q3 | Q4 |

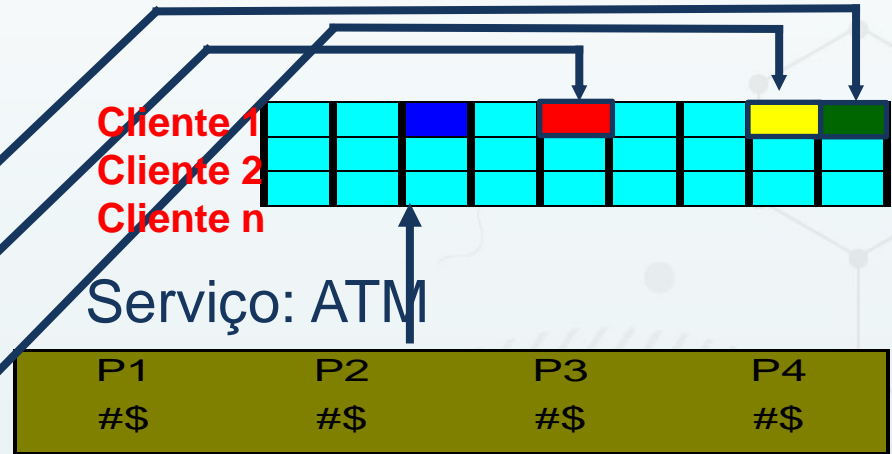
T: tipo de produto; Q: Número de itens

PIVOTAGEM REVERSA

Registro de transações bancárias => Registros por clientes

Registro de transações diárias

| Data | Cta. | Ag. | Saldo | Serviço |
|------|------|-----|-------|---|
| | 1 | 01 | |  |
| | 1 | 01 | |  |
| | 1 | 01 | |  |
| | 2 | 02 | | |



P: período do dia; #: número de transações; \$: volume movimentado no período

PIVOTAGEM REVERSA

Registro de transações Cartão de Crédito => Registro por clientes

Registro de transações mensais

| ID | Data | Hora | Tipo Loja | Valor |
|----|------|------|-----------|-------|
| 1 | | | | |
| 1 | | | | |
| 1 | | | | |
| 2 | | | | |

Cliente 1
Cliente 2
Cliente n

| | | | | | | | | |
|--|--|--|--|--|--|--|--|--|
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |

| | | | |
|------|------|------|------|
| L1 | L2 | L3 | L4 |
| :%\$ | :%\$ | :%\$ | :%\$ |

L: loja; %:tipo de loja; \$:volume da compra

É o processo de inserir dados de fontes externas ao conjunto de dados.

Exemplo 1: o perfil dos grupos pode não ser suficiente para decidir a liberação de crédito de uma pessoa. Pode ser necessário inserir o histórico de crédito e/ou de consumo.

Exemplo 2: a valorização de um imóvel somente pelas suas característica intrínsecas pode não ser suficiente para decidir seu valor. É necessário inserir informações sobre dados relativos ao lazer, índice de criminalidade, projetos de expansão futura, etc.

MELHORAMENTO DOS DADOS



É o processo de realçar características dos dados sem adição de fontes externas.

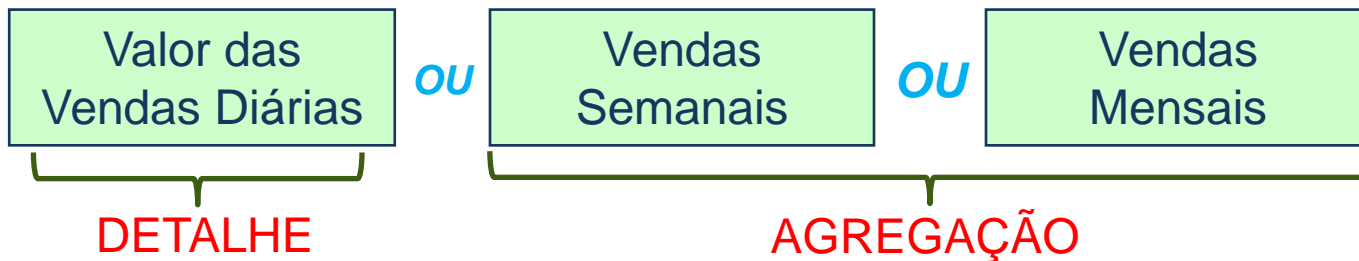
Exemplo 1: do campo “observações clínicas” (campo textual) podem ser *extraídas características* adicionais para definir melhor o perfil de cada paciente.

Exemplo 2: em processos físicos, quando a variabilidade de um parâmetro é grande (por exemplo, temperatura) pode ser necessário medir esse parâmetro com períodos de amostragem menor. Isso realça as características do parâmetro.

Granularidade = nível (detalhes/Agregação)

Dados detalhados pode ser preferível a dados agregados

Para um determinado produto



➔ REALCE DE CARACTERÍSTICAS

Ex. Registro de Medições de variáveis físicas - Clima

Registro de transações horárias: O importante é a quantidade de mudanças significativas nas medidas.

| Data | Hora | Pressão atmosférica | Velocidade vento | Humidade relativa | Radiação |
|-------|-------|---------------------|------------------|-------------------|----------|
| 01-04 | 06:00 | | | | |
| 01-04 | 07:00 | | | | |
| 01-04 | 08:00 | | | | |
| | | | | | |



06:00
06:15
06:30
06:45
07:00

Realce

➔ Avaliação da Representatividade



- ❑ A partir da análise dos domínios dos atributos deve ser feita uma avaliação da representatividade da base de dados resultante da etapa de montagem. Entende-se por representatividade, conter dados suficientes para descrever o domínio de problema.
- ❑ Caso a base de dados resultante não seja representativa o suficiente para a descoberta de conhecimento, o cientista de dados pode decidir por prosseguir, voltar à etapa anterior ou impor restrições ao conhecimento extraído.

➡ Analisando o domínio das variáveis Licap

Por exemplo 1: Pontos por multa de trânsito = $\{7, 5, 4, 3, 0\}$. Para traçar o perfil dos motoristas é necessário ter uma representatividade equilibrada entre as combinações desses valores. A soma de pontos com valor elevado pode representar a existência de outliers, o que obrigaria a segmentar o estudo e colocar restrições aos resultados alcançados.

Por exemplo 2: Estado civil de pessoas = $\{S, C, V, D\}$. A falta de registros ou o desequilíbrio destes em relação ao estado civil pode levar a restrições nos resultados, dependendo do domínio de problema sendo tratado.

➡ Analisando o domínio das variáveis Licap

Uma reflexão:

Se existir mais de 50 valores distintos num atributo discreta. Então, uma amostra de dados não pode conter menos de 50 instâncias observadas.

Caso existam mais valores que instâncias observadas a amostra não está completa e deverá ser coletada uma amostra maior. É importante contar com uma base de dados suficientemente grande e representativa.

Por exemplo: Considerando pontos por infração de trânsito [3,4,5,7]

Quantos pontos acumulados na carteira do motorista, até 3 multas, podem existir?

| 1 | 3-- | 4-- | 5-- | 7-- |
|----|-----|-----|-----|-----|
| 2 | 33- | 43- | 53- | 73- |
| 3 | 34- | 44- | 54- | 74- |
| 4 | 35- | 45- | 55- | 75- |
| 5 | 37- | 47- | 57- | 77- |
| 6 | 333 | 433 | 533 | 733 |
| 7 | 334 | 434 | 534 | 734 |
| 8 | 335 | 435 | 535 | 735 |
| 9 | 337 | 437 | 537 | 737 |
| 10 | 343 | 443 | 543 | 743 |
| 11 | 344 | 444 | 544 | 744 |
| 12 | 345 | 445 | 545 | 745 |
| 13 | 347 | 447 | 547 | 747 |
| 14 | 353 | 453 | 553 | 753 |
| 15 | 354 | 454 | 554 | 754 |
| 16 | 355 | 455 | 555 | 755 |
| 17 | 357 | 457 | 557 | 757 |
| 18 | 373 | 473 | 573 | 773 |
| 19 | 374 | 474 | 574 | 774 |
| 20 | 375 | 475 | 575 | 775 |
| 21 | 377 | 477 | 577 | 777 |

Por exemplo: Pontos por infração de trânsito [3,4,5,7]

Pontos na carteira do motorista até 3 multas:

**Caso exista interesse no aspecto temporal
no cometimento das multas (Mineração de
Eventos Complexos)**

$$\sum_{r=1}^{multas} n^r = 4^1 + 4^2 + 4^3 = 84$$

| 1 | 3-- | 4-- | 5-- | 7-- |
|----|-----|-----|-----|-----|
| 2 | 33- | 44- | 55- | 77- |
| 3 | 34- | 45- | 57- | |
| 4 | 35- | 47- | | |
| 5 | 37- | | | |
| 6 | 333 | 444 | 555 | 777 |
| 7 | 334 | 445 | 557 | |
| 8 | 335 | 447 | 577 | |
| 9 | 337 | 455 | | |
| 10 | 344 | 457 | | |
| 11 | 345 | 477 | | |
| 12 | 347 | | | |
| 13 | 355 | | | |
| 14 | 357 | | | |
| 15 | 377 | | | |

Por exemplo: Pontos por infração de trânsito [3,4,5,7]

Pontos na carteira do motorista até 3 multas:

Caso não exista interesse na ordem no cometimento das multas:

$$casos = 4 + 10 + 20 = 34$$

Incremento exponencial da amostra Licap

- Entendamos o problema do incremento exponencial da amostra:

| V1 | V2 | V3 | V4 | ... | Vn |
|----|----|----|----|-----|----|
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |

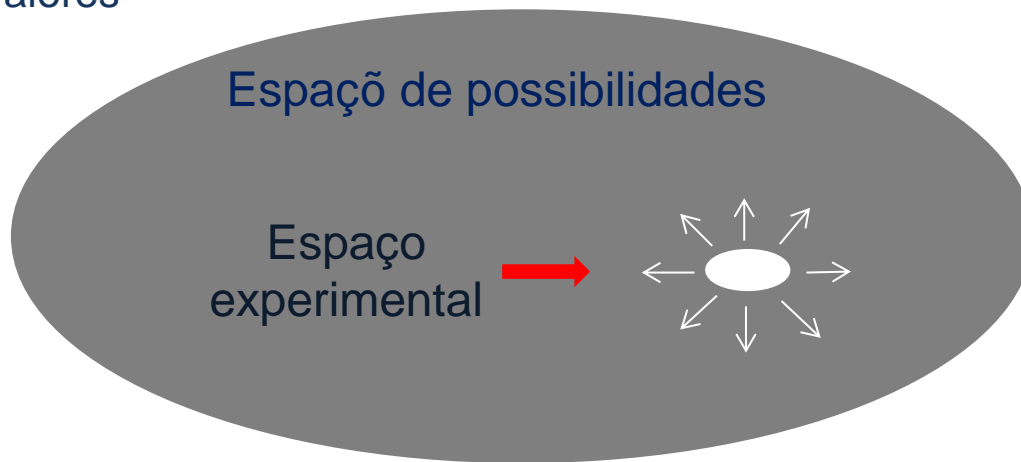
Consideremos
que cada
variável possui
 $k=4$ possíveis
valores

Consideremos que nosso
conjunto de dados possui
 $n=15$ variáveis (atributos)

Logo:

A representatividade completa deve possuir:

$$Tamanho = k^n = 4^{15} = 1.073.741.824 \text{ instâncias}$$



**Nossos
modelos são
sempre
imperfeitos e
com
restrições.**

Expansão exponencial da amostra



Para capturar uma uniforme e pequena porção de dados num espaço de alta-dimensão é necessária uma grande vizinhança. O cálculo desta expansão é dado por:

$$\text{expansão}(p, d) = p^{1/d}$$

Exemplo:

Se a porção da amostras em relação à população é $p = 10\%$

para 1-dim: expansão(0,1;1)=0,1

para 2-dim: expansão(0,1;2)=0,32

para 3-dim: expansão(0,1;3)=0,46

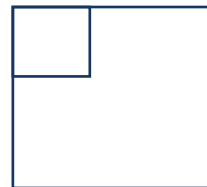
para 10-dim: expansão(0,1;3)=0,80

para 1-dim:
expansão(0,1;1)=0,1



0,10

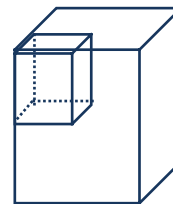
para 2-dim:
expansão(0,1;2)=0,32



0,32

para 3-dim: expansão(0,1;3)=0,46

0,46



Distanciamento entre instâncias

Cada instância deve estar mais próximo da fronteira do sub-hipercubo, que de outra instância da amostra. A distância esperada “D” entre objetos num espaço d-dim é dado por:

$$D(d, n) = \frac{1}{2} \left(\frac{1}{n} \right)^{1/d}$$

Exemplo:

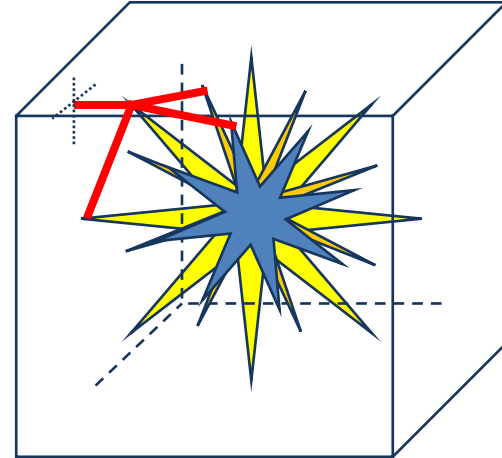
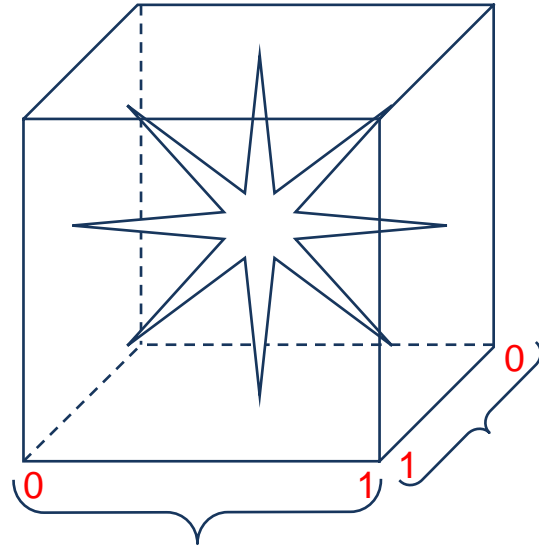
Se $n=10000$

para 1-dim: $D(1, 10000)=0,00005$

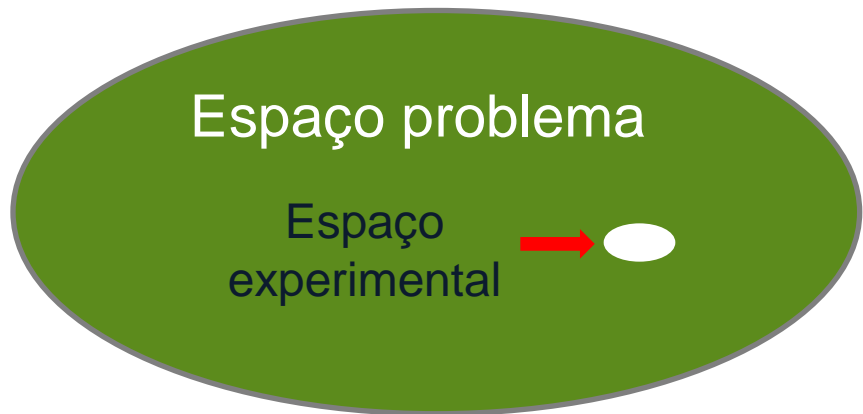
para 2-dim: $D(2, 10000)=0,005$

para 3-dim: $D(3, 10000)=0,023$

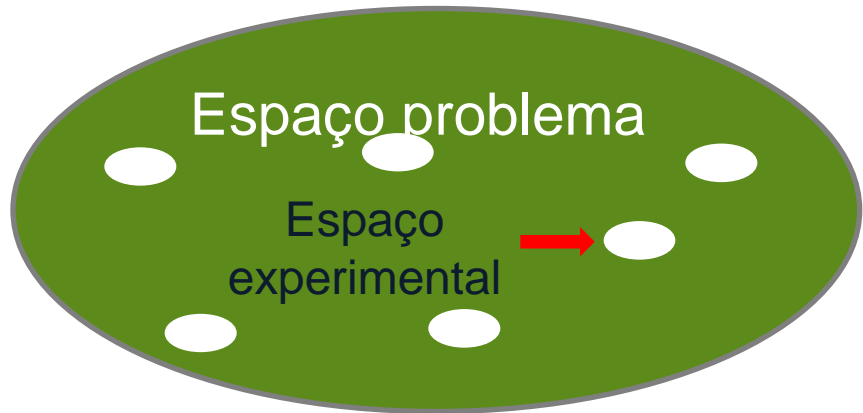
para 10-dim: $D(10, 10000)=0,20$



→ Nossos Modelos são Representativos o Suficiente? Licap



**Modelos restritos
sem capacidade de
generalização global**



**Modelos com maior
capacidade de
generalização**

➡ Problema de Indução de Hume



1711 — 1776

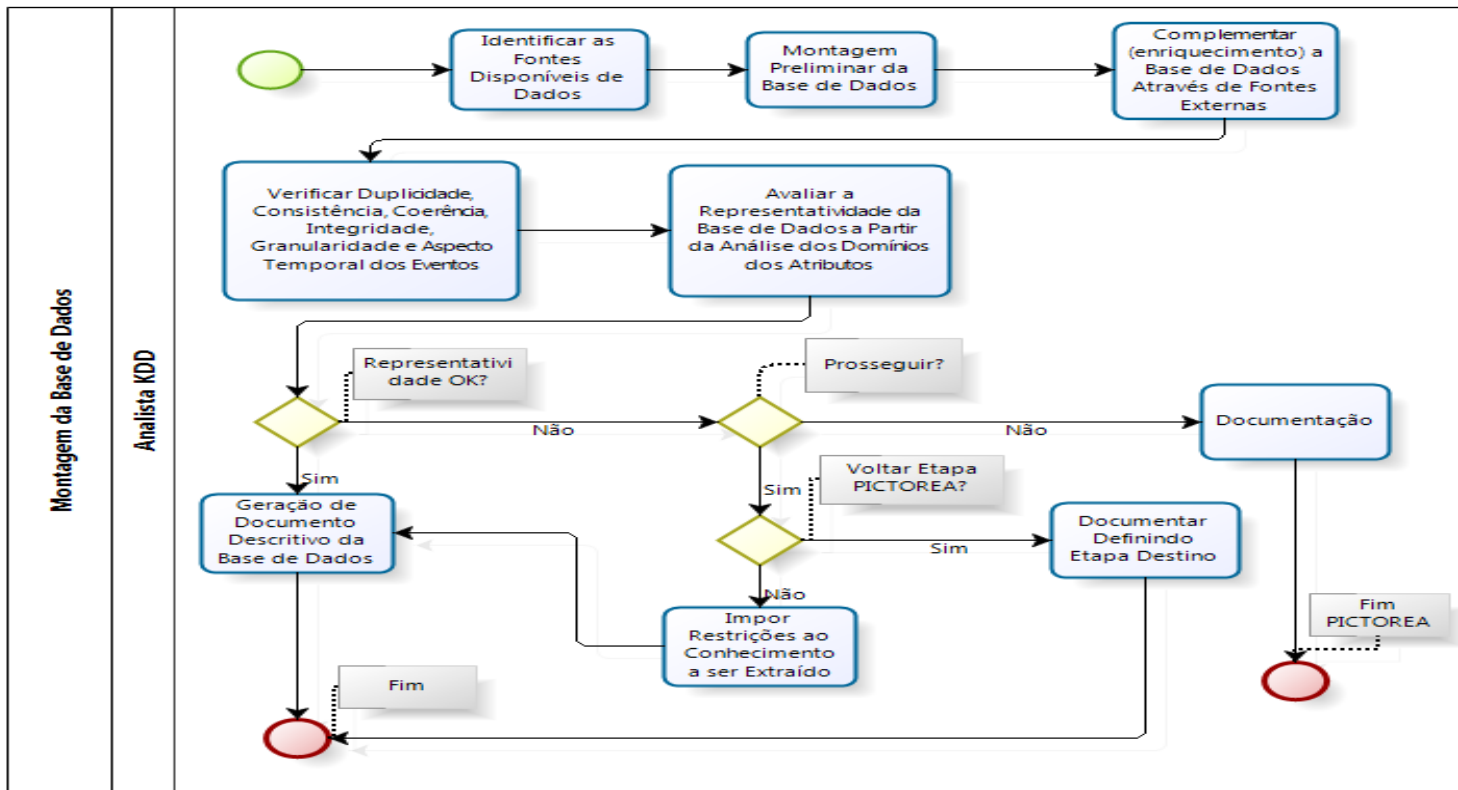


Não é possível generalizar a partir de observações persistentes

Fluxo para tomada de decisão



- Caso o Cientista de Dados opte por não prosseguir, os motivos são documentados e o processo de descoberta de conhecimento é cancelado.





Licap

Prática 5 – Para o problema identificado na etapa anterior, caracterize o domínio de problema por meio de atributos.



PUC Minas



Licap

Formação Cientista de Dados

Obrigado!



PUC Minas