

# MODELAGEM E PREPARAÇÃO DE DADOS PARA APRENDIZADO DE MÁQUINA: Montagem da base de dados

Professor:  
Luis E. Zárate

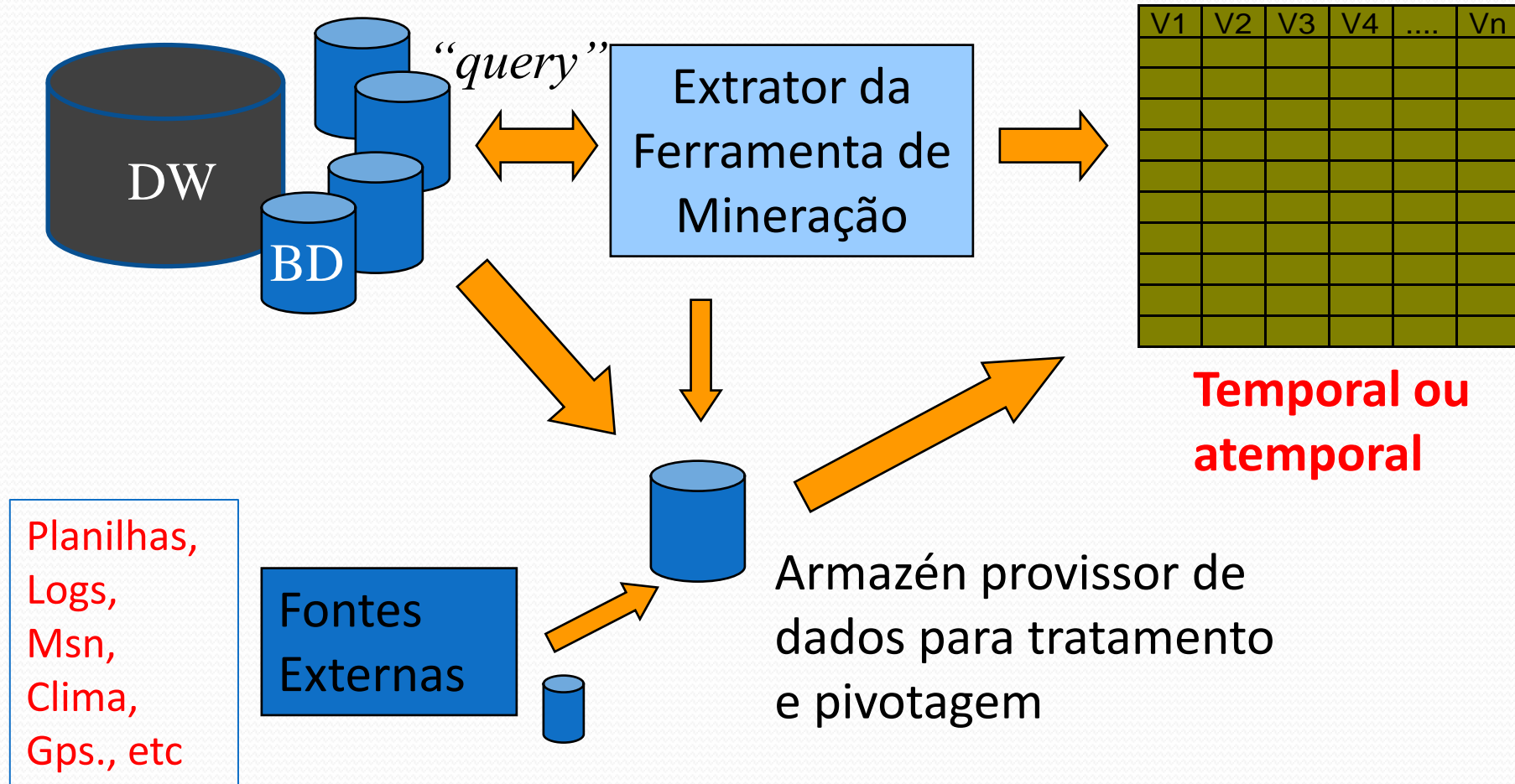
# Montagem da Base de Dados

- O cientista de dados busca em bases de dados disponíveis e fontes externas dados para os atributos essenciais vinculados ao domínio do problema identificados nas etapas anteriores.

# Fonte e Origem dos Dados

# Banco de Dados e DataWarehouse

## Tabela de 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,...).

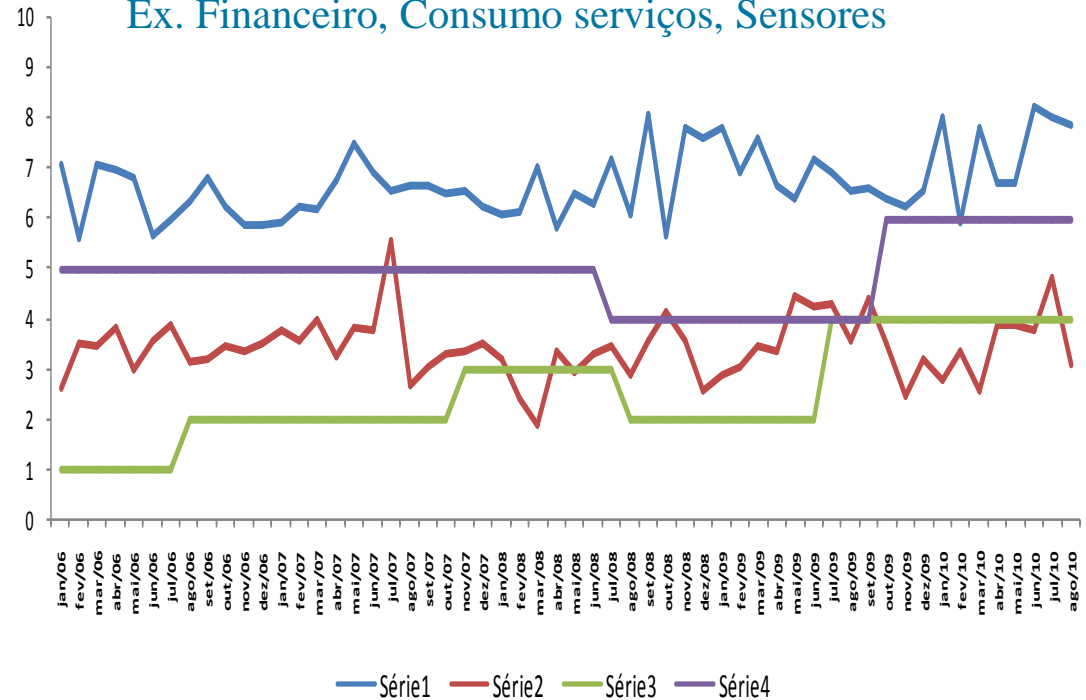
**OBS: Qualquer base considerada Estática é intrinsecamente Temporal**

# Base de dados Temporais

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

- $N$ : representa o número de instâncias, registros ou exemplos.
- $M$ : representa o número de atributos ou variáveis (numéricas ou categóricas).

Ex. Financeiro, Consumo serviços, Sensores



Cada elemento  $Z_{tij}$  corresponde a uma observação do registro  $i$  e atributo  $j$ .  
Cada valor  $t=1 \dots T$  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**

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



$$[Z^*] =$$



Média e  
Tendência

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

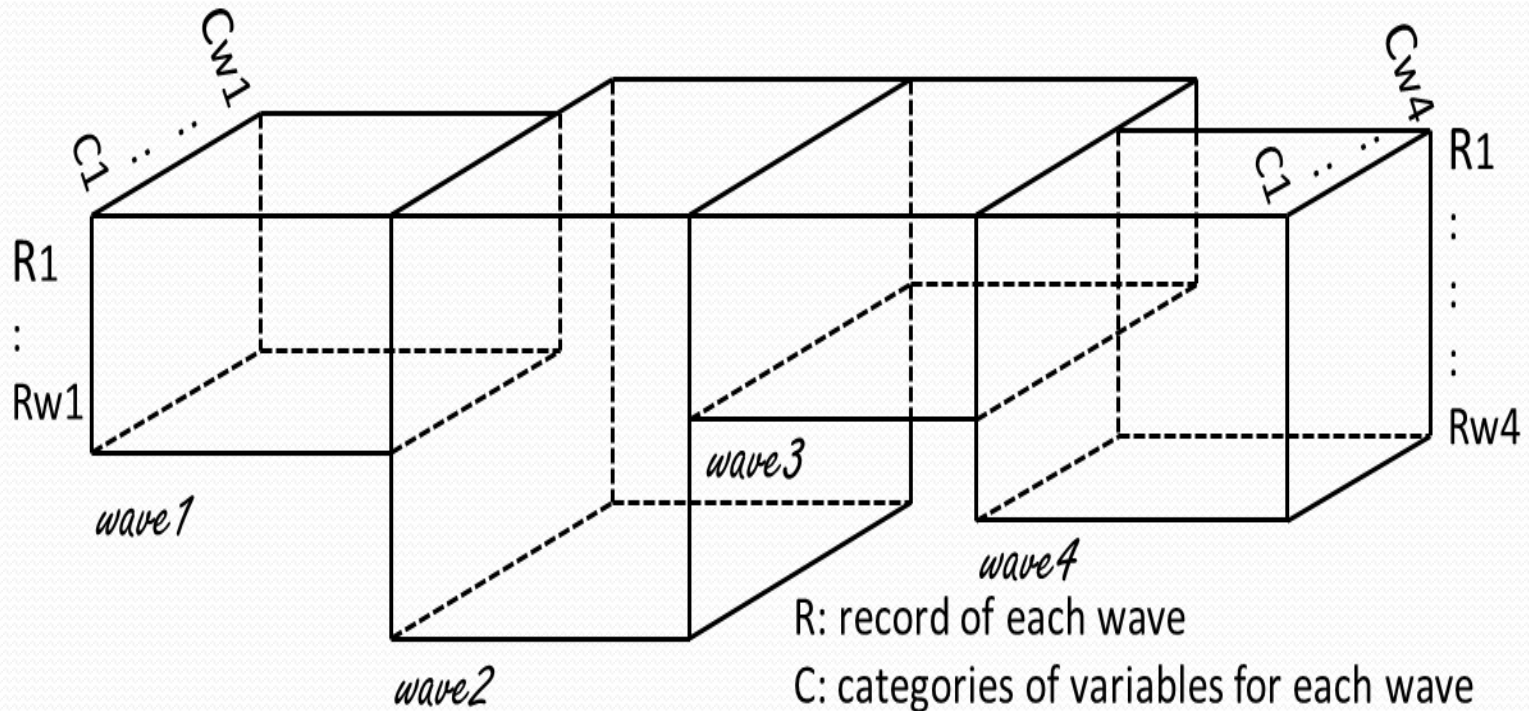
Jan1

...

JanH

# Base de dados Longitudinais

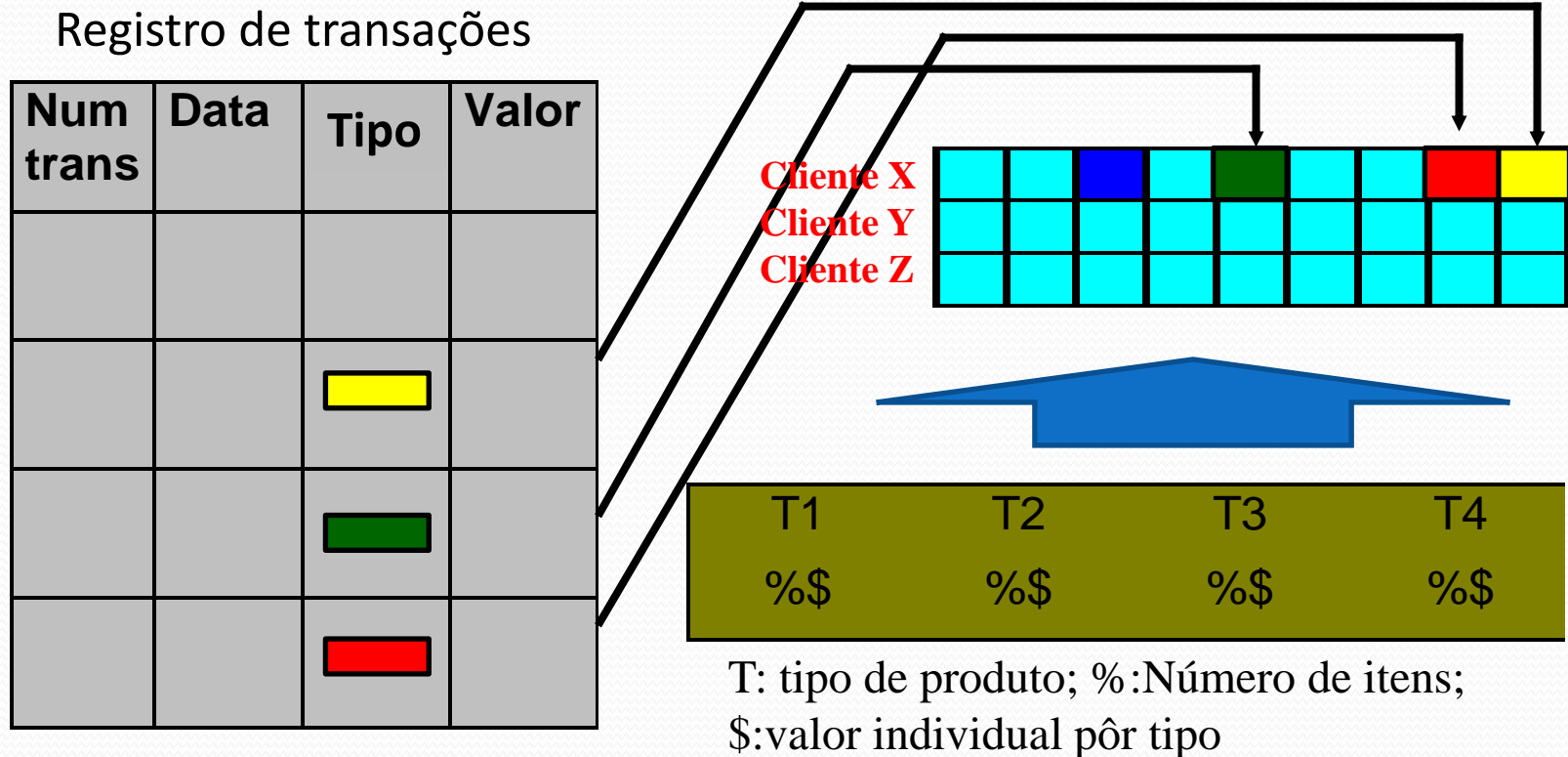
Dados longitudinais é uma forma de dado temporal na qual a mesma amostra de registros é observada repetidamente em diferentes pontos de tempo chamadas ondas.



Ex. Estudos de longevidade, de doenças, sociais

# Montagem da base de dados – Pivotagem reversa

## Registro de transações de Supermercado => Registro de consumo






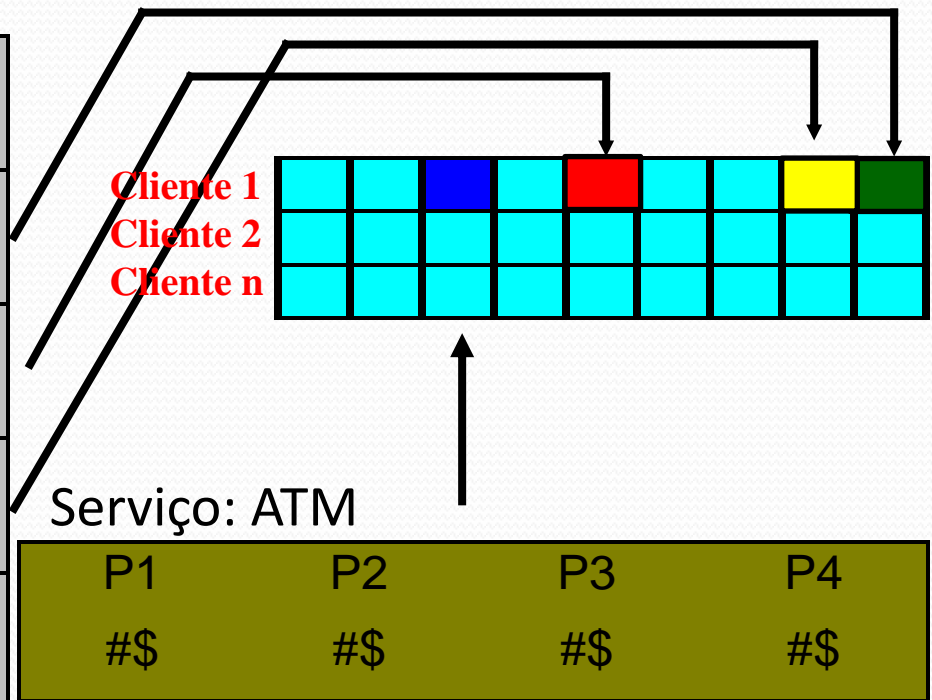


# Montagem da base de dados – Pivotagem reversa

**Registro de transações bancárias => Registro de clientes**

Registro de transações diárias

Data	Cta.	Ag.	Saldo	Serviço
				
				
				



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

# Avaliação da Representatividade

- Entendamos o problema da representatividade:

V1	V2	V3	V4	...	Vn

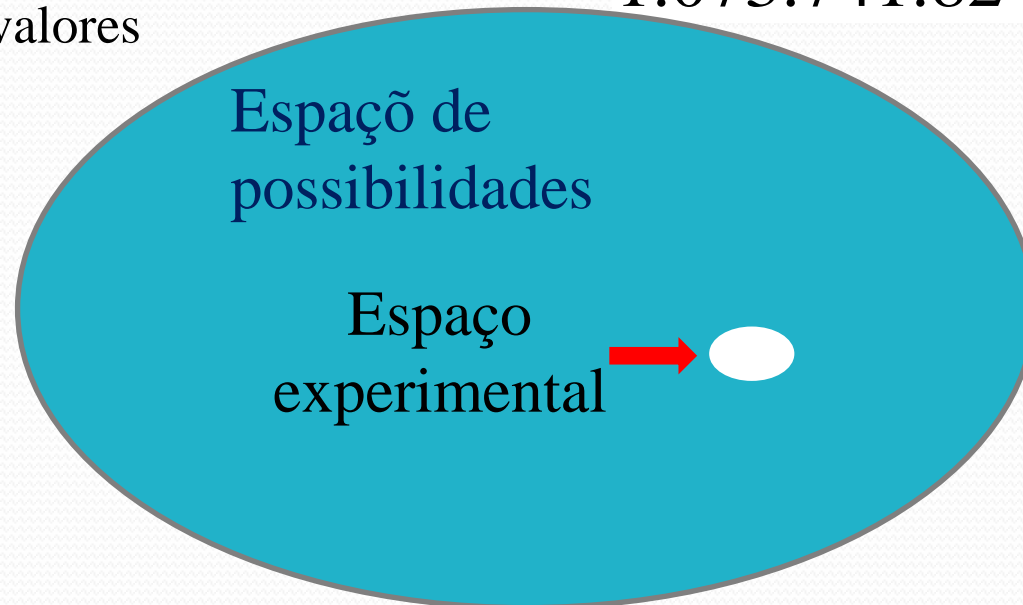
Consideremos que cada variável possui  $k=4$  possíveis valores

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

**Logo:**

A representatividade completa deve possuir:

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



**Nossos modelos são sempre imperfeitos.**

# Analizando o domínio das variáveis

- **Por exemplo 1:** Pontos por multa de transito = {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.

# Avaliação da Representatividade

- Deve ser realizada uma avaliação da representatividade da base de dados criada a partir da análise dos domínios das variáveis (entende-se por representatividade conter dados suficientes para descrever o domínio de problema).
- Caso o banco de dados resultante não seja representativo o suficiente para a descoberta de conhecimento, o cientista de dados pode decidir por prosseguir, pode voltar a alguma etapa anterior ou impor restrições ao conhecimento a ser extraído.
- Caso o Cientista de Dados opte por não prosseguir, os motivos são documentados e o processo de descoberta de conhecimento é cancelado.

# Obrigado

Professor:  
Luis E. Zárate