

Mineração de Dados

João Carlos Xavier Júnior

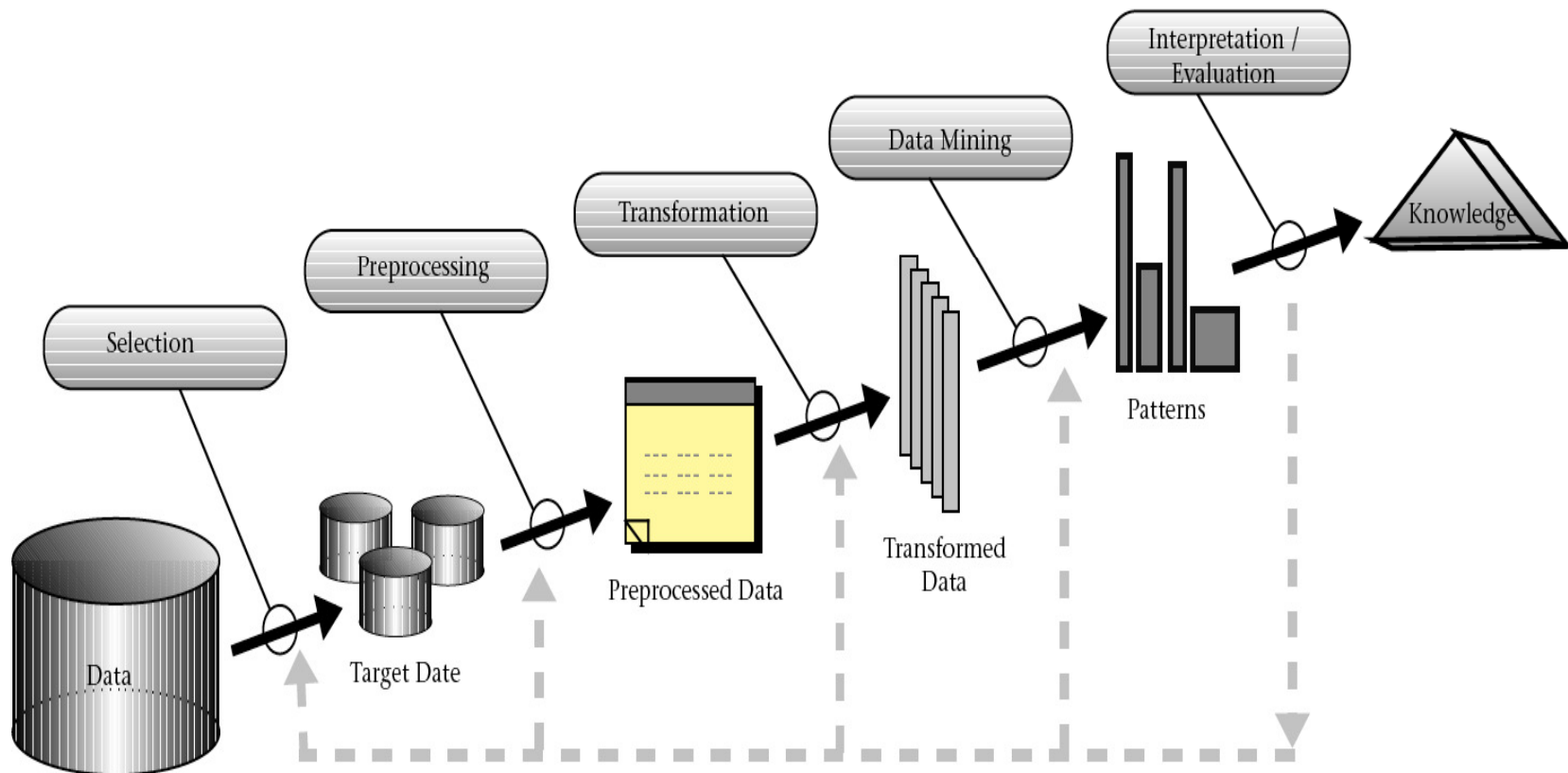
jcxavier@imd.ufn.br

Introdução

- ❑ Processo de **descoberta de conhecimento** em bases de dados (*Knowledge Discovery in Database – KDD*).
 - Encontrar padrões úteis embutidas em **grandes volumes** de dados.
 - A descoberta deve ser relevante, desconhecida e de máxima abrangência, ou seja, o “**ouro**”.
- ❑ Analogia com a **mineração**: grandes volumes de dados são “peneirados” na tentativa de se encontrar alguma coisa de valor.

KDD

□ Etapas do KDD:

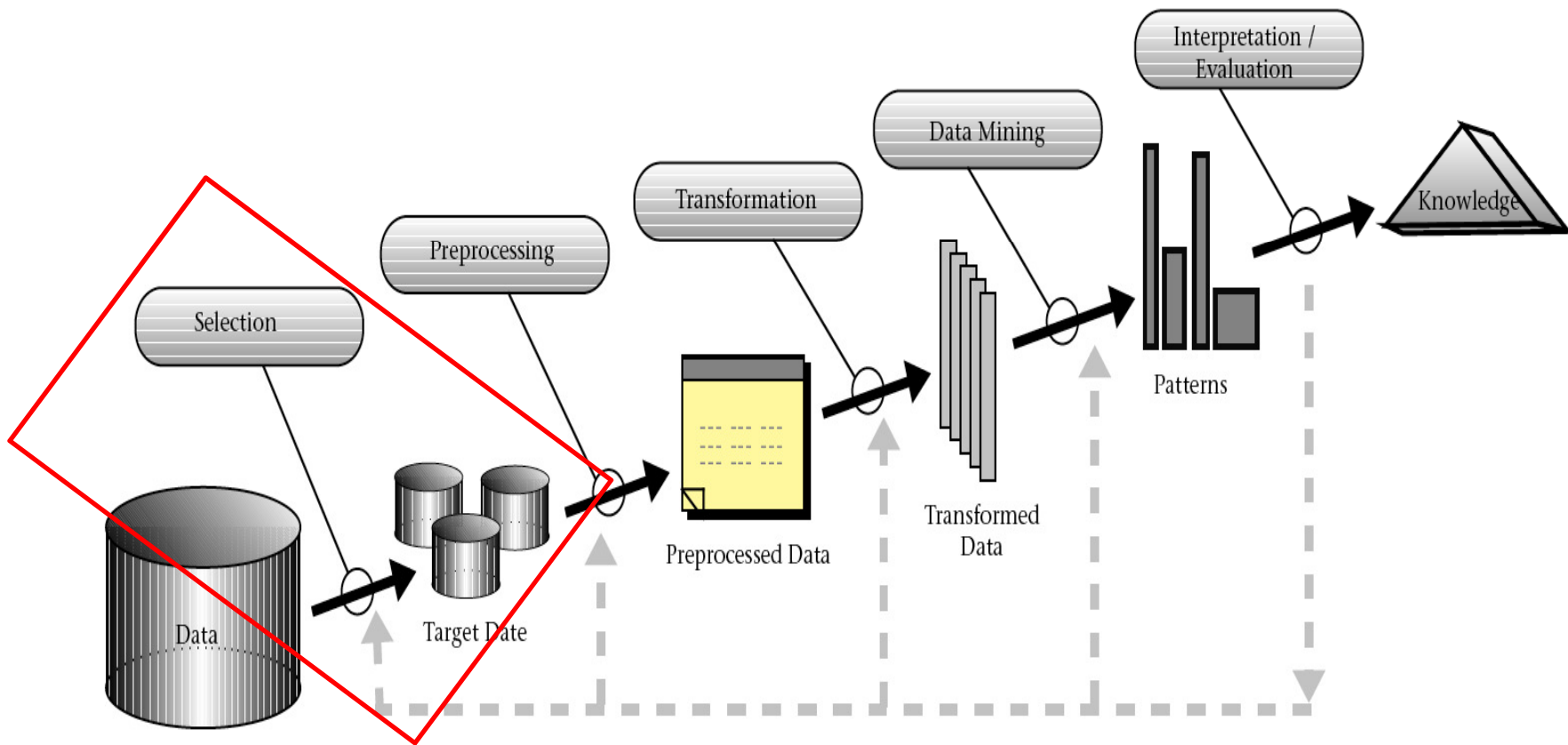


Mineração de Dados

- ❑ Termo em inglês:
 - ❖ *Data Mining*
- ❑ Também conhecida por **Aprendizado de Máquina**
 - ❖ *Machine Learning*
- ❑ Termos que se confundem, mas são similares.

KDD

Selecção dos Dados:



Seleção dos Dados

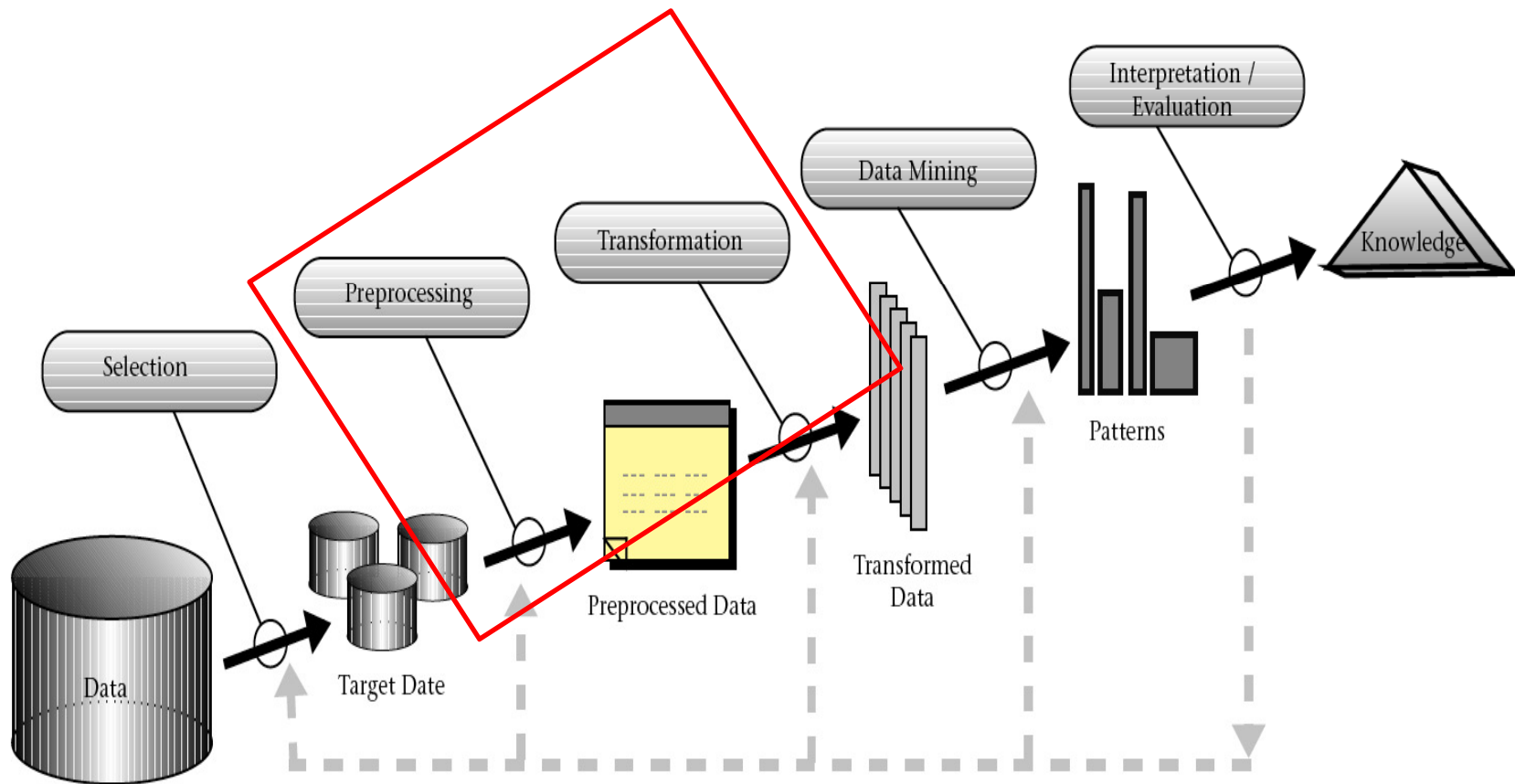
- ❑ O processo de seleção é **bastante complexo**;
- ❑ Os dados podem vir de uma série de **fontes diferentes**:
 - ❖ Data warehouses, planilhas, sistemas legados (SGBD).
- ❑ Podem possuir os mais **diversos formatos**;
- ❑ É comum ocorrer a necessidade de uma **ferramenta específica** para a carga dos dados.

Ferramentas ETL

- ❑ Ferramentas utilizadas para extrair, transformar e carregar (*Extract, Transform and Load*) dados;
- ❑ Diferentes fontes de dados (integração);
- ❑ Ferramentas para Integração de dados (*data integration*), segundo Gartner:
 - ❖ PowerCenter (Informatica);
 - ❖ InfoSphere (IBM);
 - ❖ SAP Data Services (SAP);
 - ❖ SAS Data Management (SAS) + Hadoop,
 - ❖ Oracle Data Integrator (Oracle);
 - ❖ Open Studio (Talend).

KDD

❑ Pré-Processamento:

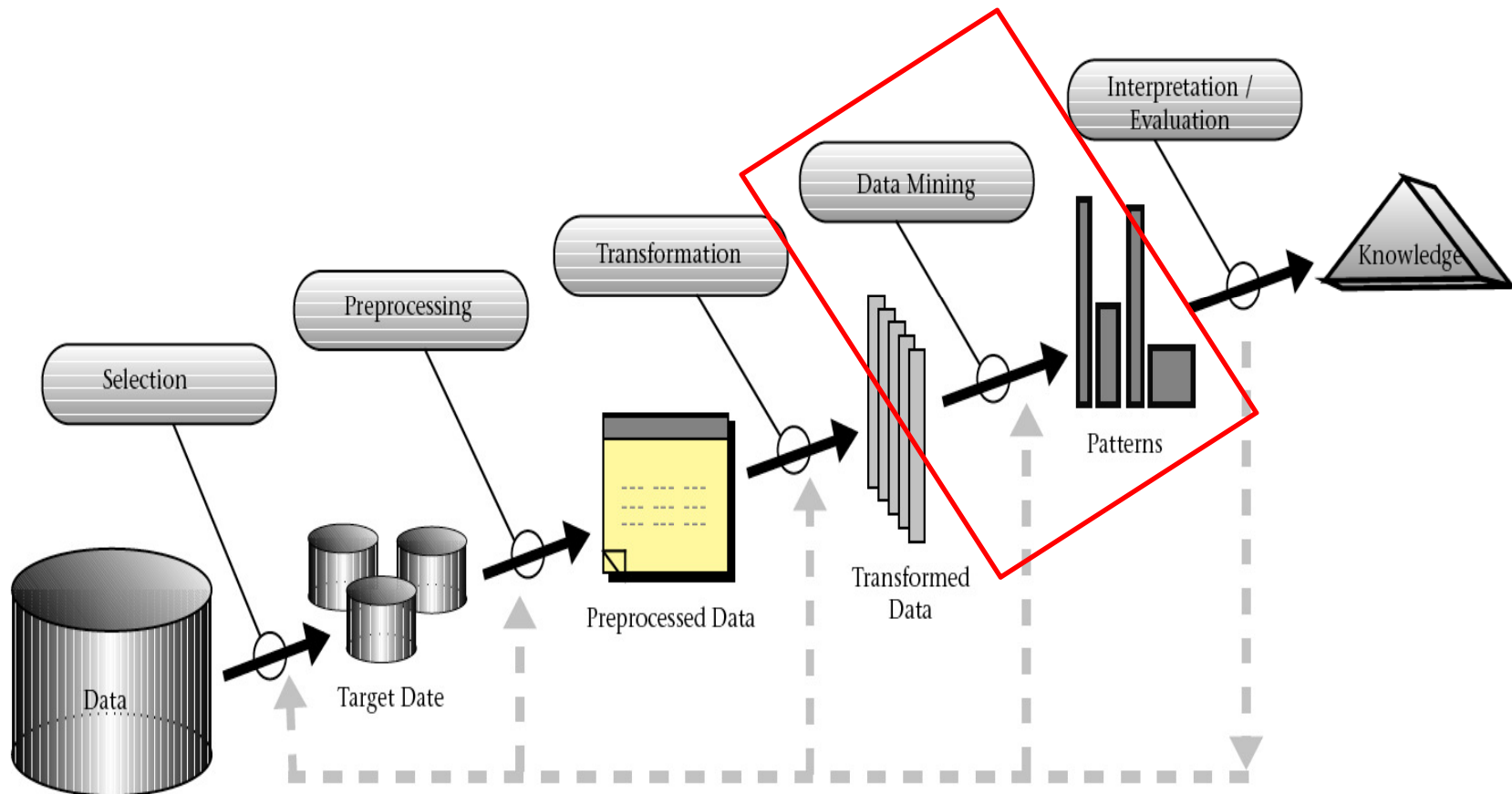


Pré-Processamento

- ❑ Pré-processamento: etapa onde são aplicados **métodos** para **limpeza, tratamento e redução** do volume de dados.
 - ❖ Dados possuem defeitos;
 - ❖ Atributos inadequados;
 - ❖ Quantidade muito grande de atributos (multi-dimensionalidade).
- ❑ Sem **pré-processamento** não existe Aprendizado de Máquina (algoritmos).

KDD

□ Aprendizado de Máquina:



O que é Aprendizado?

❑ O que significa aprender?

O que é Aprendizado?

❑ Aprender significa:

- ❖ “Alcançar ou conseguir conhecimento, cognição, educação ou especialidade através da experiência ou do estudo”;
- ❖ “Ficar-se competente ou apto em algo”;
- ❖ “Ficar-se eficiente ou capaz, em alguma coisa, de forma gradual”.

❑ Aprendizagem, **não é memorizar**. Qualquer computador pode memorizar, a dificuldade está em **generalizar** um comportamento para uma nova situação.

Aprendizado de Máquina

- ❑ **Objetivo principal:** construção de programas de computador que melhoram seu desempenho **por meio da experiência**.
 - ❖ Utilizar técnicas que tentam procurar **padrões** (conhecimento) nos dados disponíveis.
- ❑ São conhecidas como técnicas **orientadas a dados**.
 - ❖ Aprendem automaticamente **a partir dos dados**.
 - ❖ Geração de **hipóteses** (modelos) a partir dos dados.

Aprendizado de Máquina

- ❑ As tarefas de aprendizado podem ser divididas em:
 - ❖ Preditivas;
 - ❖ Descritivas.

- ❑ Também chamados de:
 - ❖ Aprendizado Supervisionado;
 - ❖ Aprendizado Não Supervisionado.

Aprendizado de Máquina

- As tarefas de aprendizado podem ser divididas em:

- ❖ Preditivas;

- ❖ Descritivas.

Class Labels

- Também chamados de:

- ❖ Aprendizado Supervisionado;

- ❖ Aprendizado Não Supervisionado.

Aprendizado de Máquina

- As tarefas de aprendizado podem ser divididas em:

- ❖ Preditivas;

- ❖ Descritivas.

Class Labels

- Também chamados de:

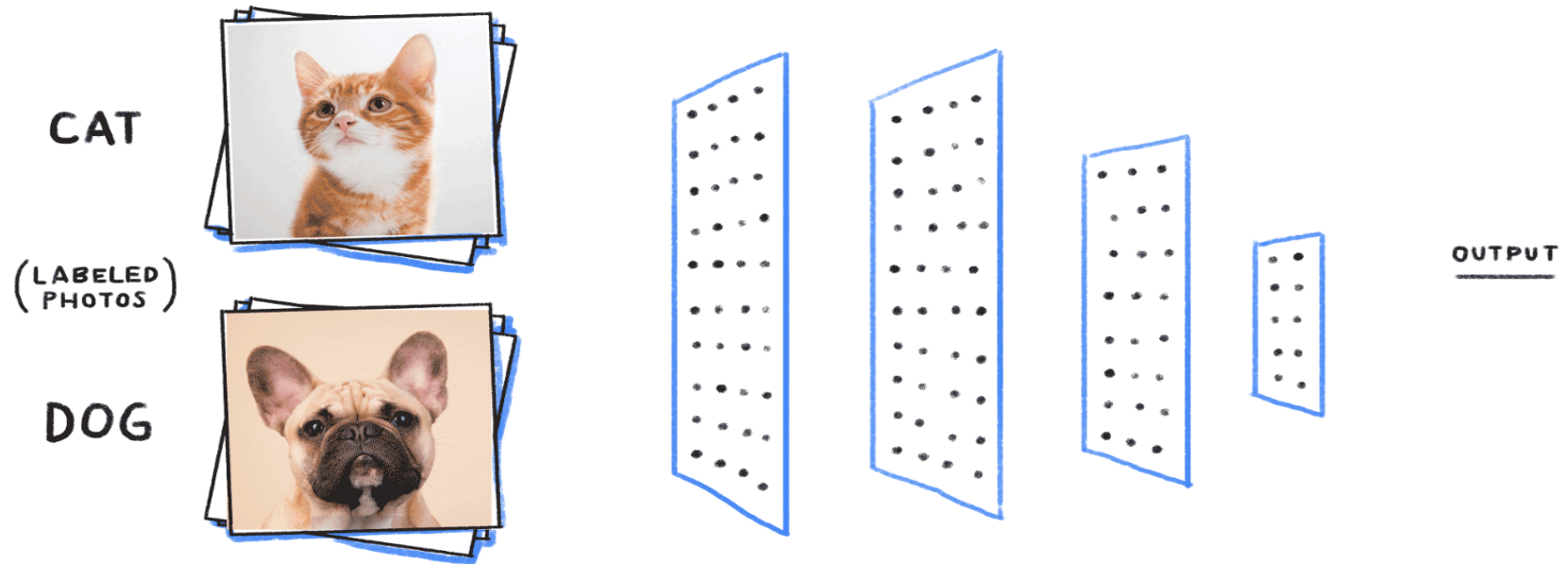
- ❖ Aprendizado Supervisionado;

- ❖ Aprendizado Não Supervisionado.

Que danado significa “Class Labels”????

Aprendizado de Máquina

□ Class Labels ...



<https://becominghuman.ai/building-an-image-classifier-using-deep-learning-in-python-totally-from-a-beginners-perspective-be8dbaf22dd8>

Aprendizado de Máquina

- ❑ Tarefas Preditivas: objetivo é encontrar uma **função** (hipótese), que a partir dos dados de treinamento possa ser utilizada para **prever** um novo valor.
 - ❖ Classificação;
 - ❖ Regressão.

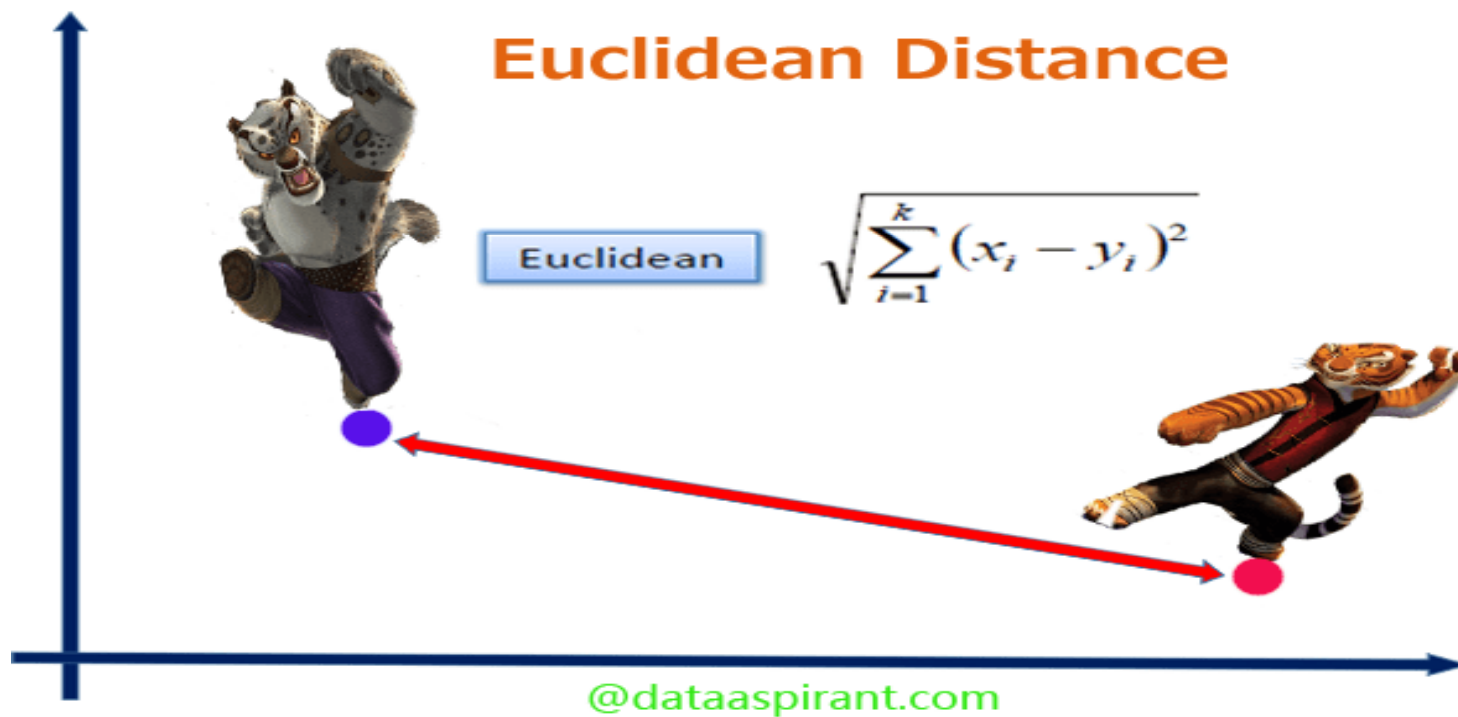
- ❑ Qual a diferença de classificação e regressão?
 - ❖ Conceitualmente são similares.
 - ❖ A principal diferença é que o atributo a ser predito é contínuo (regressão), em vez de discreto (classificação).
 - ❖ Exemplo: **estatura** e **número de filhos**.

Aprendizado de Máquina

- Tarefas Descritivas: a meta é **explorar** ou **descrever** um conjunto de dados, sem fazer uso do **atributo classe**.
 - ❖ **Associação**: são usadas para descobrir elementos que ocorrem em comum dentro de um determinado conjunto de dados.
 - $\{\text{leite, pão}\} \rightarrow \{\text{manteiga}\}$.
 - ❖ **Clustering** ou **agrupamento**: encontra similaridade entre os objetos (instâncias ou registros) através de uma medida (geralmente, distância).

Aprendizado de Máquina

□ Distância:



<http://dataaspirant.com/2015/04/11/five-most-popular-similarity-measures-implementation-in-python/>

Aprendizado de Máquina

❑ Algoritmos de Classificação (Preditivos):

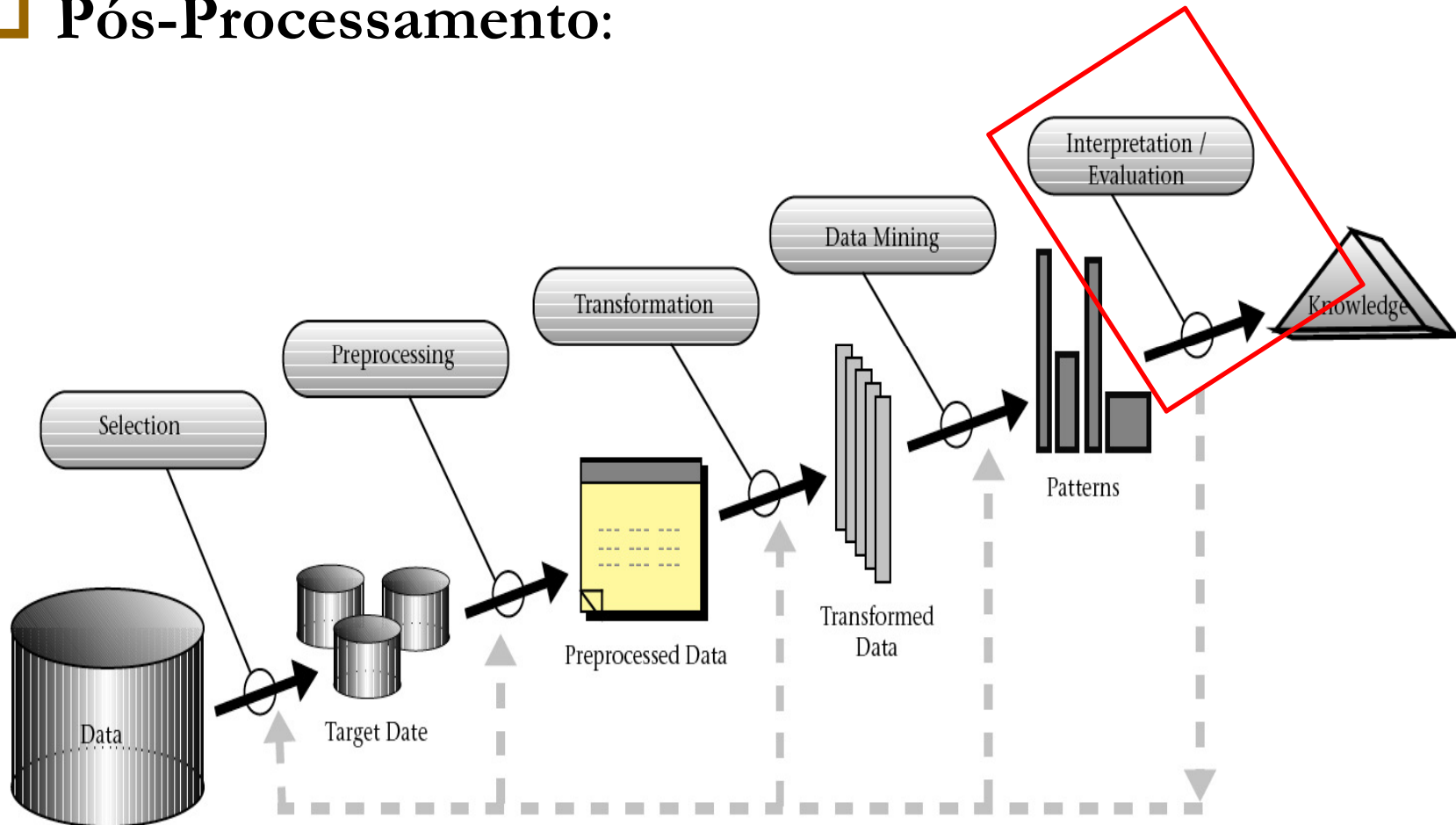
- ❖ k-NN;
- ❖ Naive Bayes;
- ❖ Árvores de Decisão (J48);
- ❖ Redes Neurais Artificiais (MLP);
- ❖ Máquina de Vetores de Suporte (SMO).

❑ Algoritmos de Clustering (Descritivos):

- ❖ k -Means
- ❖ Hierárquico aglomerativo;
- ❖ Expectation–Maximization (EM)

KDD

❑ Pós-Processamento:

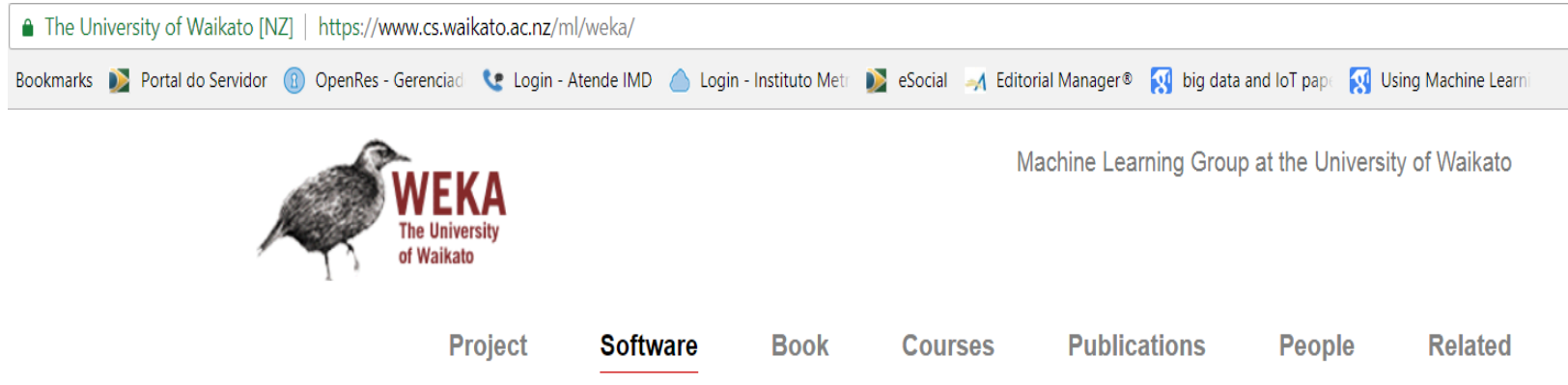


Pós-Processamento

- ❑ Pós-processamento:
 - ❖ Obtenção do conhecimento.
 - ❖ O que fazer com o conhecimento adquirido?
 - ❖ Que benefícios isto pode me trazer?

- ❑ Medidas de Avaliação:
 - ❖ Medidas de desempenho (classificação);
 - ❖ Medidas de qualidade (clustering).

Plataformas de ML



Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like **this**, and the bird sounds like **this**.

Weka is open source software issued under the **GNU General Public License**.

We have put together several free online courses that teach machine learning and data mining using Weka. Check out the **website for the courses** for details on when and how to enrol. The videos for the courses are available **on Youtube**.

Yes, it is possible to apply Weka to process **big data** and perform **deep learning**!

Plataformas de ML

🔒 Não seguro | scikit-learn.org/stable/

Bookmarks Portal do Servidor OpenRes - Gerenciad Login - Atende IMD Login - Instituto Metr eSocial Editorial Manager® big data and IoT pap Using Machine Learn

 [Home](#) [Installation](#) [Documentation](#) [Examples](#)



scikit-learn
Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.

Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dúvidas ...





Conceitos Relevantes



Conjunto de dados


- ❑ Composto por objetos (**instancias** ou **padrões**) que representam um objeto (físico ou abstrato);
- ❑ Cada objeto: vetor de características (**atributos**) onde cada um está associado a uma propriedade do objeto;
- ❑ Formalmente representado por:

$$X_{n \bullet d}$$

Onde ***n*** é o número de objetos e ***d*** é a dimensionalidade do espaço de objetos.

Conjunto de dados

❑ Exemplo (*weather dataset*):



Instância	Outlook	Temperature	Humidity	Wind	Play
D1	sunny	hot	high	weak	no
D2	sunny	hot	high	strong	no
D3	overcast	hot	high	weak	yes
D4	rain	mild	high	weak	yes
D5	rain	cool	normal	weak	yes
D6	rain	cool	normal	strong	no
D7	overcast	cool	normal	strong	yes
D8	sunny	mild	high	weak	no
D9	sunny	cool	normal	weak	yes
D10	rain	mild	normal	weak	yes
D11	sunny	mild	normal	strong	yes
D12	overcast	mild	high	strong	yes
D13	overcast	hot	normal	weak	yes
D14	rain	mild	high	strong	no

Definição dos atributos

□ Tipos:

- ❖ **Atributos nominais:** são também conhecidos como categóricos.
- ❖ **Atributos ordinais:** semelhantes aos categóricos, porém seus valores podem ser ordenados.
- ❖ **Atributos intervalares:** apresentam valores numéricos com ordem entre eles. Também existe diferença entre eles.
 - Obs: Zero (0) é relativo (temperatura ...)

Atributos Nominais

- ❑ Os valores são **não numéricos** e são **não ordenados**.
- ❑ Exemplos: cor (carro), modelo, marca, etc.
 - ❖ **Obs:** duas instâncias apresentam ou não o mesmo valor;
 - ❖ Não existe relação entre os valores nominais (sem ordem);
 - ❖ Somente testes de igualdade podem ser realizados.
- ❑ **Problema:** como calcular a distância entre duas instâncias?

Atributos Ordinais

- ❑ Impõe uma ordem nos valores.
- ❑ Porém: não existe distancia nos valores predefinidos.
- ❑ Exemplo: atributo “**temperature**” da base *weather*.
 - ❖ Valores: “hot” > ”mild” > “cool”
- ❑ Atributos ordinais podem ter seus valores substituídos diretamente por números, como forma de visualizar melhor a ordem entre eles.
 - ❖ Valores: 1.0 > 0.5 > 0.0

Atributos Intervalares

- ❑ Os intervalos são ordenados e medidos em unidades fixas e iguais.
- ❑ Exemplo: atributo “ano”
 - ❖ Intervalo: de 1900 – 2014.
 - ❖ A diferença entre 2 valores faz sentido?
 - ❖ Normalizar é preciso em alguns casos.



Pré-Processamento



Por que fazer o pré-processamento?

- ❑ Os dados possuem defeitos:
 - ❖ Incompletos:
 - Ausência de valores (missing values).
 - ❖ Ruidosos:
 - Valores discrepantes (outliers).
 - ❖ Inconsistentes:
 - Valores com erro ou fora da realidade.
- ❑ Sem dados de boa qualidade o resultado do método de AM é **pobre**.

Etapas do pré-processamento

□ Limpeza dos dados:

- ❖ Preencher dados ausentes, “alisar” ruído, identificar e/ou remover valores aberrantes, resolver inconsistências.

□ Transformação de dados:

- ❖ Transformação de tipos;
- ❖ Normalização.

□ Redução de Dados:

- ❖ Redução no volume de dados (instâncias e atributos).

Limpeza dos Dados

- ❑ Valores perdidos (*missing values*):
 - ❖ Valor de atributo de uma instância não está disponível.
- ❑ Pode ser consequência de:
 - ❖ Inconsistência com outros dados gravados.
 - ❖ Não informação de dados.
- ❑ É importante diferenciar dados perdidos de não aplicáveis:
 - ❖ Exemplo: o valor de um atributo “anos-de-casado” seria inexistente (não aplicável) no caso de instâncias que representem adultos que nunca foram casados.

Limpeza dos Dados - Valores perdidos

- ❑ Tratamento 1: desconsiderar os valores perdidos.
 - ❖ Incluir na base de dados apenas aquelas instâncias com dados completos, excluindo as outras.

Algum problema???

Limpeza dos Dados - Valores perdidos

- Tratamento 2: definir um limiar de valores perdidos aceitável.
 - Defina uma proporção (limiar) de valores perdidos aceitável e, baseado nisto, são excluídos as instâncias e/ou atributos com níveis excessivos.

Algum problema???

Limpeza dos Dados - Valores perdidos

- ❑ Tratamento 3: atribuir um valor ao que está faltando.
 - ❖ Certo, mas como fazer????



Limpeza dos Dados - Valores perdidos

□ Tratamento 3: método de atribuição.

- ❖ Estimar valores perdidos com base em valores válidos do mesmo atributo.
 - Substituição pela média;
 - Substituição pela mediana; ou
 - Substituição pela moda.
- ❖ Válido apenas para atributos numéricos e categóricos.
- ❖ Mas, cuidado
 - Por exemplo, como substituir o valor perdido para um atributo nominal ou categórico como “sexo”?

Limpeza dos Dados - Valores perdidos

❑ Substituição com Média e Mediana:

	Instâncias => $\lfloor n/2 \rfloor + \lfloor (n/2)+1 \rfloor$									
-	1	2	3	4	5	6	7	8	Mediana	Média
Desordenados	15	3	0	6	7	1	10	12	-	-
Ordenados	0	1	3	6	7	10	12	15	6,5	6,75

	Instâncias => $\lfloor n/2 \rfloor + 1$									
	1	2	3	4	5	6	7	8	Mediana	Média
Desordenados	15	3	0	6	7	1	10		-	-
Ordenados	0	1	3	6	7	10	15		6	6

Limpeza dos Dados - Valores perdidos

❑ Substituição com Moda:

	idade	educacao	altura	estado_civil	raca	sexo			Moda_EC	Moda_R	
1	39	Bacharelado	1,51	Solteiro	Branco	Masculino	Solteiro		4	10	Branco
2	50	Bacharelado	1,65	Casado	Branco	Masculino	Casado		8	6	Negro
3	38	Mestrado	1,92	Divorciado	Branco	Masculino	Divorciado		3	1	Indio
4	53	Licenciatura	1,52	Casado	Negro	Masculino	Viuvo		2		
5	28	Bacharelado	1,65	Casado	Negro	Feminino					
6	37	Mestrado	1,47	Casado	Branco	Feminino					
7	49	Licenciatura	1,54	Viuvo	Negro	Feminino					
8	52	Mestrado	1,65	Casado	?	Masculino					
9	31	Mestrado	1,92	Solteiro	?	Feminino					
10	142	Bacharelado	1,59	Casado	Branco	Masculino					
11	37	Licenciatura	1,65	?	Negro	Masculino					
12	30	Tecnico	1,87	Casado	Indio	Masculino					
13	23	Bacharelado	1,67	Solteiro	Branco	Feminino					
14	32	Doutorado	1,81	?	Negro	Masculino					
15	25	Licenciatura	3,57	Viuvo	Branco	Masculino					
16	32	Mestrado	3,56	Solteiro	Branco	Masculino					
17	38	Tecnico	1,49	Casado	?	Masculino					
18	43	Mestrado	1,98	Divorciado	Branco	Feminino					
19	40	Doutorado	2,35	?	Branco	Masculino					
20	54	Licenciatura	1,87	Divorciado	Negro	Feminino					

Limpeza dos Dados: ruído e/ou valores aberrantes

- Técnicas para identificar valores ruidosos ou aberrantes:
 - ❖ Variância;
 - ❖ Clustering;
 - ❖ Regressão linear.

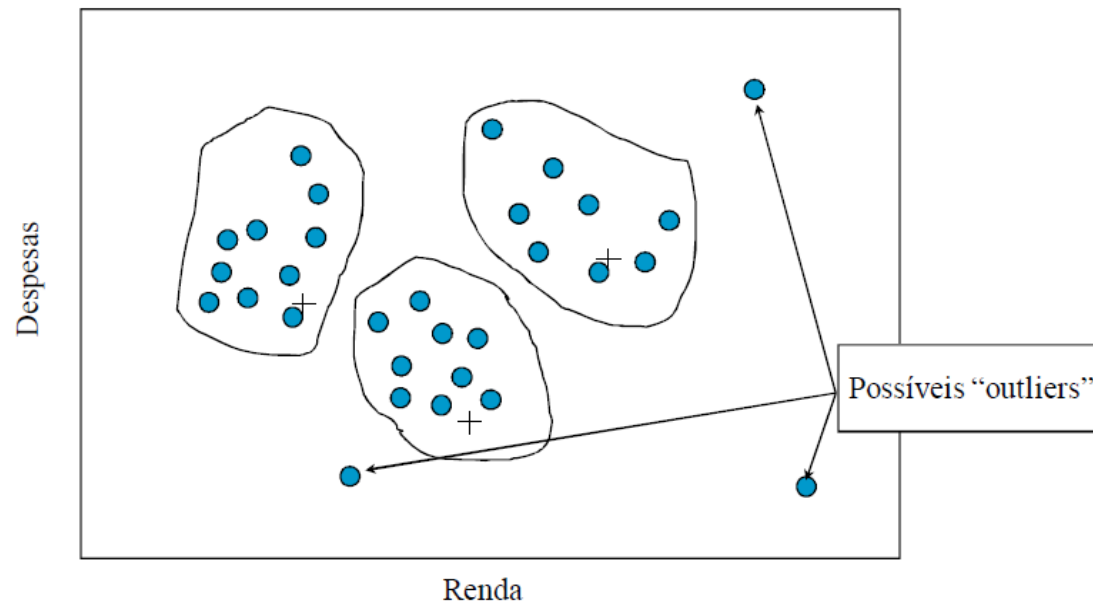
Limpeza dos Dados: valores aberrantes

Calculando a Variância:

Variância							Média						
	idade	educacao	altura	estado_civil	raca	sexo		altura	1,9874				
1	39	Bacharelado	1,51	Solteiro	Branco	Masculino	1	1,51	-0,4774	27	4,12	2,1326	
2	50	Bacharelado	1,65	Casado	Branco	Masculino	2	1,65	-0,3374	17	3,57	1,5826	
3	38	Mestrado	1,92	Divorciado	Branco	Masculino	3	1,92	-0,0674	18	3,56	1,5726	
4	53	Licenciatura	1,52	Casado	Negro	Masculino	4	1,52	-0,4674	25	2,45	0,4626	
5	28	Bacharelado	1,65	Casado	Negro	Feminino	5	1,65	-0,3374	21	2,35	0,3626	
6	37	Mestrado	1,47	Casado	Branco	Feminino	6	1,47	-0,5174	20	1,98	-0,0074	
7	49	Licenciatura	1,54	Viuvo	Negro	Feminino	7	1,54	-0,4474	3	1,92	-0,0674	
8	52	Mestrado	1,65	Casado	Branco	Masculino	8	1,65	-0,3374	9	1,92	-0,0674	
9	31	Mestrado	1,92	Solteiro	Indio	Feminino	9	1,92	-0,0674	12	1,87	-0,1174	
10	142	Bacharelado	1,59	Casado	Branco	Masculino	10	1,59	-0,3974	22	1,87	-0,1174	
11	37	Licenciatura	1,65	Casado	Negro	Masculino	11	1,65	-0,3374	29	1,82	-0,1674	
12	30	Tecnico	1,87	Casado	Indio	Masculino	12	1,87	-0,1174	14	1,81	-0,1774	
13	23	Bacharelado	1,67	Solteiro	Branco	Feminino	13	1,67	-0,3174	24	1,81	-0,1774	
14	32	Doutorado	1,81	Solteiro	Negro	Masculino	14	1,81	-0,1774	28	1,79	-0,1974	
15	40	Doutorado	?	Casado	Indio	Masculino	15		-1,9874	26	1,76	-0,2274	
16	34	Mestrado	?	Casado	Indio	Masculino	16		-1,9874	13	1,67	-0,3174	
17	25	Licenciatura	3,57	Viuvo	Branco	Masculino	17	3,57	1,5826	23	1,67	-0,3174	
18	32	Mestrado	3,56	Solteiro	Branco	Masculino	18	3,56	1,5726	2	1,65	-0,3374	
19	38	Tecnico	1,49	Casado	Branco	Masculino	19	1,49	-0,4974	5	1,65	-0,3374	
20	43	Mestrado	1,98	Divorciado	Branco	Feminino	20	1,98	-0,0074	8	1,65	-0,3374	
21	40	Doutorado	2,35	Casado	Branco	Masculino	21	2,35	0,3626	11	1,65	-0,3374	
22	54	Licenciatura	1,87	Divorciado	Negro	Feminino	22	1,87	-0,1174	10	1,59	-0,3974	
23	35	Mestrado	1,67	Casado	Negro	Masculino	23	1,67	-0,3174	7	1,54	-0,4474	
24	43	Licenciatura	1,81	Casado	Branco	Masculino	24	1,81	-0,1774	4	1,52	-0,4674	
25	59	Doutorado	2,45	Divorciado	Branco	Feminino	25	2,45	0,4626	1	1,51	-0,4774	
26	156	?	1,76	Casado	Branco	Masculino	26	1,76	-0,2274	19	1,49	-0,4974	
27	19	?	4,12	Solteiro	Branco	Masculino	27	4,12	2,1326	6	1,47	-0,5174	
28	54	Licenciatura	1,79	Casado	Indio	Masculino	28	1,79	-0,1974	15		-1,9874	
29	39	Bacharelado	1,82	Divorciado	Branco	Masculino	29	1,82	-0,1674	16		-1,9874	
30	49	Tecnico	?	Casado	Branco	Masculino	30		-1,9874	30		-1,9874	

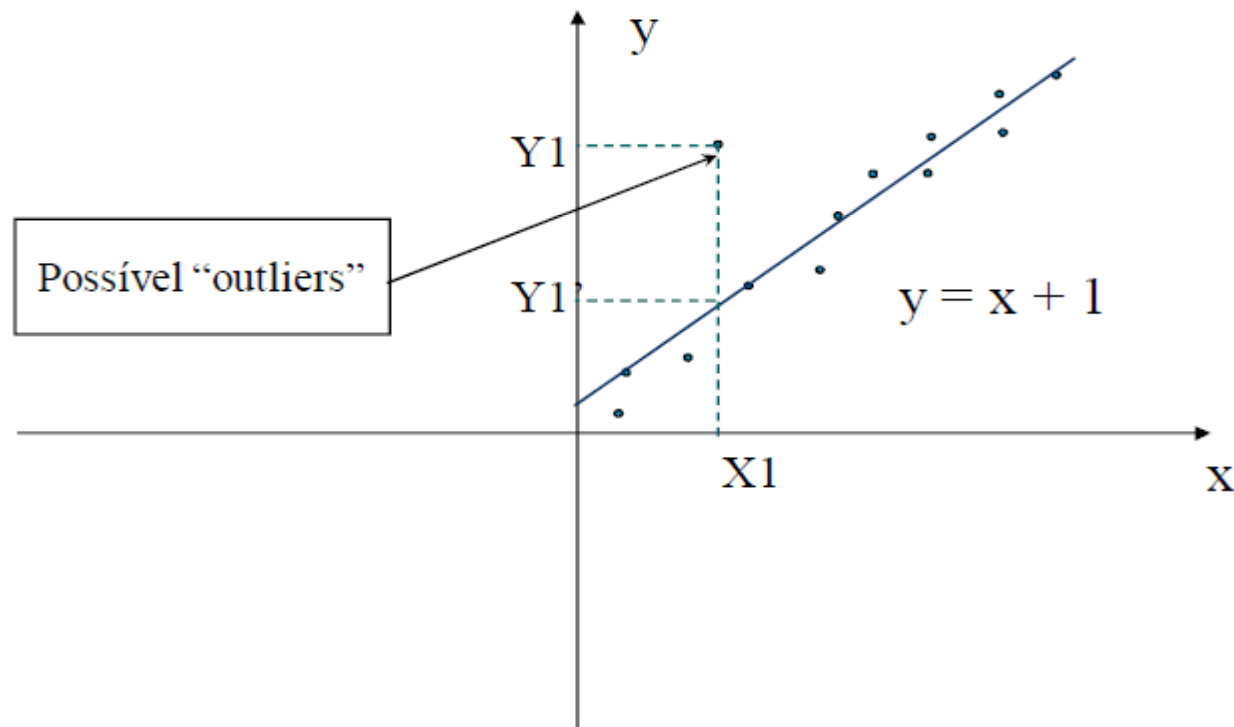
Limpeza dos Dados: valores aberrantes

- ❑ Análise de Cluster: detecção de valores aberrantes.
 - ❖ Os valores são organizados em grupos;
 - ❖ Os valores isolados podem ser considerados aberrantes.



Limpeza dos Dados: valores aberrantes

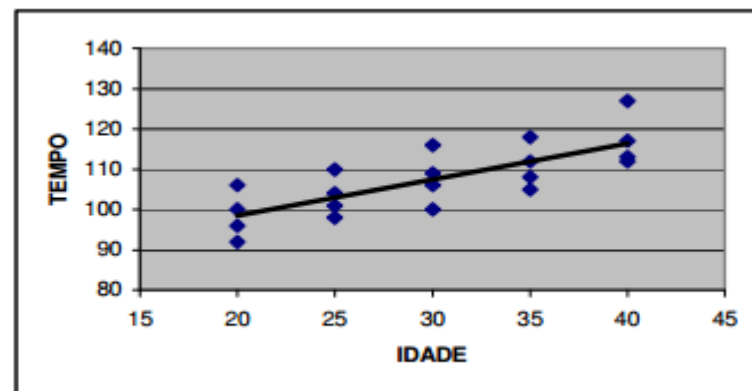
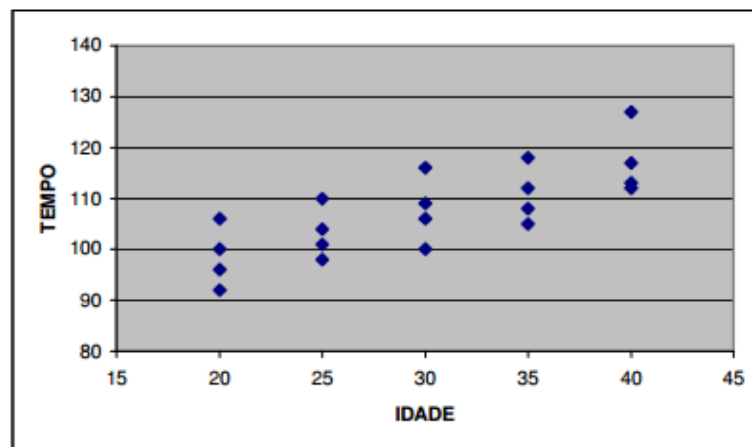
□ Regressão linear:



Limpeza dos Dados: valores aberrantes

□ Regressão linear: exemplo

Y - Tempo de reação (segundos)	X - Idade (em anos)
96	20
92	20
106	20
100	20
98	25
104	25
110	25
101	25
116	30
106	30
109	30
100	30
112	35
105	35
118	35
108	35
113	40
112	40
127	40
117	40



$$Y = a + bX$$

$$b = \frac{\sum x_i y_i - n \bar{y} \bar{x}}{\sum x_i^2 - n \bar{x}^2}$$

$$a = \bar{y} - b \bar{x}$$

Limpeza dos Dados: valores aberrantes

□ Regressão linear: exemplo

n	yi	xi	My	Mx			b	a	Y = a + b*X	(Y-yi)^2
1	96	20	107,5	30	1920	400			98,5	6,25
2	92	20			1840	400			98,5	42,25
3	106	20			2120	400			98,5	56,25
4	100	20			2000	400			98,5	2,25
5	98	25			2450	625			103,0	
6	104	25			2600	625			103,0	
7	110	25			2750	625			103,0	
8	101	25			2525	625			103,0	
9	116	30			3480	900			107,5	
10	106	30			3180	900			107,5	
11	109	30			3270	900			107,5	
12	100	30			3000	900			107,5	
13	112	35			3920	1225			112,0	
14	105	35			3675	1225			112,0	
15	118	35			4130	1225			112,0	
16	108	35			3780	1225			112,0	
17	113	40			4520	1600			116,5	
18	112	40			4480	1600			116,5	
19	127	40			5080	1600			116,5	
20	117	40			4680	1600			116,5	
					900	1000	0,9	80,5		

Limpeza dos Dados: ruído e/ou valores aberrantes

- Técnicas para remover ruídos:
 - Alisamento:
 - Mediana;
 - Média.

Transformação de dados

- ❑ Necessário obter os dados em uma forma apropriada para a mineração.
- ❑ Tipos:
 - Numérico \rightarrow Numérico (Normalização);
 - Numérico \rightarrow Nominal (Discretização);
 - Nominal \rightarrow Numérico;
 - Nominal \rightarrow Binário;
 - Ordinal \rightarrow Numérico;
 - Numérico \rightarrow Ordinal.

Transformação de dados

□ Normalização ou mudança de escala:

- ❖ Propósito da normalização: minimizar os problemas oriundos do **uso de unidades** e **dispersões distintas** entre as variáveis.
- ❖ As variáveis podem ser normalizadas segundo a amplitude ou segundo a distribuição.

$$Att_1 = \left(\frac{x_i - \min}{\max - \min} \right)$$



Ano	Vlr (norm.)
1900	0,0
1914	0,1
1950	0,4
1981	0,7
1999	0,9
2005	0,9
2014	1,0

Transformação de dados

□ Transformação de Numérico para Nominal:

- ❖ Dois motivos principais:

- Alguns algoritmos de **AM** manipulam melhor atributos com valores nominais.
- Eficiência: redução do numero de valores.

- ❖ Os atributos numéricos são “**discretizados**” em um pequeno número de intervalos distintos.

- ❖ Muito importante quando se trabalha com **árvores de decisão**.

Transformação de dados

□ Transformação de Nominal para Numérico:

- ❖ Algumas técnicas de **AM** manipulam melhor valores numéricos.
 - Exemplos: redes neurais (MLP), k-NN.
- ❖ Os valores de atributos nominais precisam ser transformados em valores numéricos.
- ❖ Existem estratégias diferentes para atributos com valores **binário, ordinal e nominal multi-valorado**.

Transformação de dados

❑ Transformação de Nominal para Numérico:

❖ Atributos binários:

- Exemplo: Sexo = Masculino ou Feminino.

❖ Converta para o Atributo_0_1 com os valores 0,1.

- Sexo = M \Rightarrow Sexo_0_1 = 1
- Sexo = F \Rightarrow Sexo_0_1 = 0

sex	Male	Male	Male	Male	Female	Female	Female	Male	Female	Male	Male	Male	Female	Male
sex_0_1	1	1	1	1	0	0	0	1	0	1	1	1	0	1

Transformação de dados

❑ Transformação de Nominal para Numérico:

❖ *Atributos ordinais* (e.g., grau_de_satisfação com um produto) podem ser convertidos para números preservando a ordem natural.

- Muito Satisfeito \Rightarrow 0.8
- Satisfeito \Rightarrow 0.6
- Pouco Satisfeito \Rightarrow 0.4
- Insatisfeito \Rightarrow 0.2

❖ Por que é importante preservar a ordem natural?

- Para permitir comparações que façam sentido:
`grau_de_satisfação > 0.4`

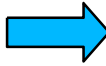
Transformação de dados

❑ Transformação de Nominal para Numérico:

❖ Atributos nominais multi-valorados um número pequeno de possíveis valores (e.g., < 20).

- Religião = {Católica, Protestante, Budista, ..., etc}
- Para cada valor v de Religião, crie um atributo binário R_v , que será 1 se Religião = v , 0 caso contrário.

ID	Religião	...
Pessoa 1	Budista	
Pessoa 2	Católico	



ID	$R_{Católica}$	$R_{Protestante}$	$R_{Budista}$...
Pessoa 1	0	0	1	
Pessoa 2	1	0	0	

Transformação de dados

❑ Transformação de Nominal com muitos Valores:

❖ Exemplos:

- Código Postal (CEP) de uma cidade;
- Profissão.

❖ Agrupe valores “naturalmente”:

- 150 bairros (CEP) de Recife \Rightarrow 3 ou 5 regiões.
- Profissões: selecione a mais frequentes e agrupe o resto.

❖ Crie atributos binários para os valores selecionados.

❑ Ignore atributos cujos valores são únicos para cada instância: RG, CPF, matrícula SIAPE, ...

Dúvidas ...

